Sergie Andrei Gerrits Arruda

EXPLORING DATA LITERACY:
CONCEPTS AND DETERMINANTS FOR DATA SKILLS DEVELOPMENT

Dissertação no âmbito do Mestrado em Gestão orientada pelo Professor Doutor Manuel Paulo Albuquerque Melo e apresentada à Faculdade de Economia da Universidade de Coimbra.

Julho de 2020

1 2 9 0 FACULDADE DE ECONOMIA UNIVERSIDADE D COIMBRA

Sergie Andrei Gerrits Arruda

Exploring data literacy: concepts and determinants for data skills development

Dissertação de Mestrado em Gestão orientada pelo Professor Doutror Manual Paulo Albuquerque Melo e apresentada à Faculdade de Economia da Universidade de Coimbra para obtenção do grau de Mestre

Coimbra, 2020

ACKNOWLEDGEMENTS

I believe that no great outcome originates exclusively from individual effort. Every achievement and realisation derives from a plethora of factors that permeate our lives, especially the people who help to shape who we are. Each presence, each act of kindness, each aid was a fundamental part of every step taken on my journey so far, and every person that I've met along the way deserves acknowledgement. They are part of it and live in every step I take.

With that said, there would be too many names to mention, so I will endeavour to speak with my actions throughout my life to show how grateful I am. I thank God for allowing me to experience everything that has passed. I thank my parents for being a core pillar that supported my journey, one that was fundamental to me. I thank all my family for standing with me on the road to this point, especially my grandparents, sister, cousins, uncles and aunts and those who became part of the family.

To all the friends who supported my journey, especially the ones who concretely helped with the dissertation. This is to both the old ones, separated physically by our choices of paths and the new ones met in this new chapter, not to forget the digital ones, especially the Discord community. Thank you for everything you have done. My thanks also to JEEFEUC and its members. This journey would not have been the same without this fantastic experience and the people involved.

My sincerest thanks to Professor Paulo Melo for accepting me as his student and advising in the development of this research, and to all other professors and professionals from the University of Coimbra. The work of an educator is a long and arduous journey, but bear in mind your teachings live through every soul you have touched in the classroom, and I am no different. May one day I give back directly or to society with the knowledge you have given me.

My thanks also to every individual who inspired me to achieve more, even those I have not met in person. May technology continue to be used to bring people together and develop us evermore. And as one study journey ends, may these experiences enlighten the upcoming paths and may the memories we made ever be our strength.

My deepest thanks to all of you. And thank you, Coimbra.

Technology is a useful servant but a dangerous master.
Christian Lous Lange

Remembering that you are going to die is the best way I know to avoid the trap of thinking you have something to lose.
You are already naked. There is no reason not to follow your heart.
Steve Jobs

Abstract

As technology progresses and it becomes increasingly easy to harness the power of data for a multitude of purposes, the need to understand all the processes involving data around us grows as well. In that sense, data literacy also gains more relevance, which the knowledge about data and the many skills involved, such as reading, collecting, interpreting, visualizing, manipulating, managing, and deciding based on data. Nevertheless, the concept of data literacy is relatively new and is still under development. Moreover, it has several variations in the academic literature, where each author defines it based on their specific contexts. Still, there is a core of common elements and competencies commonly mentioned in the literature which can be a guiding point to study the theme. In the end, this is a complex concept which involves a range of skills from the technological manipulation of data to the critical interpretation of its analysis. Despite the many existing levels of expertise of this skill, virtually any person can benefit from having at least a basic data literacy. However, there is an apparent mismatch between what is expected from new professionals and their abilities. On one side, large organisations leverage technology and optimize workflow based on gigantic datasets, thus shaping reality with data-driven decisions and they expect new professionals to have the necessary knowledge to do the same. On the other side, most people lack the skills to correctly assess how data is being used daily in activities that affect them, from political decisions to marketing advertisements. This research aims to analyse what are the elements involved in this concept and then design a generic theoretical data literacy framework that enables the understanding of what is involved for the average individual and what factors determine the development of this skill. While it is not yet possible to provide a final answer, evidence suggests data literacy is still concentrated in specific people with higher or specific educational backgrounds, but mostly in those with a positive attitude (or simply an interest) towards data. Notwithstanding, evidence also shows that some people are simply unaware of the many implications of this skill and can be led to learn given due stimuli.

Keywords: Data Literacy, Education, Business, Technology, Skills

Resumo

Conforme a tecnologia progride e torna-se cada vez mais fácil utilizar o poder dos dados para diversos propósitos, cresce também a necessidade de entender todos os processos que envolvem dados à nossa volta. Nesse sentido, a literacia em dados também ganha mais relevância, que é o conhecimento sobre os dados e as várias habilidades envolvidas, como ler, coletar, interpretar, visualizar, manipular, gerenciar e decidir com base nos dados. No entanto, o conceito de alfabetização de dados é relativamente novo e ainda está em desenvolvimento. Além disso, ele possui diversas variações na literatura acadêmica, onde cada autor a define com base em seus contextos específicos. Ainda assim, há um núcleo de elementos e competências comuns comumente mencionados na literatura que podem ser um ponto de referência para estudar o tema. No final, este é um complexo conceito que envolve uma gama de habilidades, desde a manipulação tecnológica de dados até a interpretação crítica de sua análise. Apesar dos muitos níveis de conhecimento existentes dessa habilidade, praticamente qualquer pessoa pode se beneficiar de ter pelo menos um conhecimento básico de dados. Entretanto, há um aparente descompasso entre o que se espera de novos profissionais e suas habilidades. Por um lado, as grandes organizações aproveitam a tecnologia e otimizam o fluxo de trabalho com base em conjuntos de dados gigantescos, moldando a realidade com decisões orientadas por dados e elas esperam que novos profissionais tenham o conhecimento necessário para fazer o mesmo. Por outro lado, a maioria das pessoas não possui as habilidades necessárias para avaliar corretamente como os dados estão sendo usados diariamente em atividades que os afetam, desde decisões políticas a anúncios de marketing. Esta pesquisa tem como objetivo analisar quais são os elementos envolvidos nesse conceito e, em seguida, projetar uma estrutura teórica genérica de literacia em dados que permita entender o que está envolvido para o indivíduo médio e quais fatores determinam o desenvolvimento dessa habilidade. Embora ainda não seja possível fornecer uma resposta final, as evidências sugerem que a literacia em dados ainda está concentrada em pessoas específicas com formação educacional mais alta ou específica, mas principalmente naquelas com uma atitude positiva (ou simplesmente um interesse) em relação aos dados. Não obstante, as evidências também mostram que algumas pessoas simplesmente desconhecem as muitas implicações dessa habilidade e podem ser levadas a aprender dados os devidos estímulos.

Palavras-chave: Literacia em dados, Educação, Negócios, Tecnologia, Habilidades

List of figures

List of tables

List of Acronyms

AISeL – Association for Information Systems eLibrary

CCSS – Common Core State Standards

ICT – Information, Communication and Technology

DLFT – Data literacy For Teachers

DLSP – Data Literacy for Safety Professionals

EBSCO – Elton B. Stephens Co.

IEEEXPLORE – Institute of Electrical and Electronics Engineers Xplore

JSTOR – Journal Storage

MBA – Master of Business Administration

MoRM – Museum of Random Memory

NGSS – Next Generation Science Standards

SRD – Safety-Related Data

SPSS – Statistical Package for the Social Sciences

STEM – Science, Technology, Engineering and Mathematics

INDEX

1. Introduction

As technology becomes increasingly accessible, data use has become present everywhere. From being used to identify digital footprints of an internet user to providing large amounts of materials for academic or business purposes, people are often interacting with data in some way, be it reading, generating, analysing, interpreting or making a decision.

Researches incorporate the use of technology and generate a large amount of data in a new paradigm of e-research (Koltay, 2015). Data is often tracked and used by companies everyday activities, such as credit card transactions, analytics from websites or social media activity (Pothier & Condon, 2019), travel apps, smart meters or sharing economy (Wolff, Wermelinger, & Petre, 2019), to name a few. Authors such as D'Ignazio (2017) claim that data is now faced as a currency of power. Automated systems process data and guide decision-making regarding a range of activities from a company's marketing to government acts.

The new scenario has been enabled by the improvement in ICT (Information and Communication Technology) infrastructure, coupled with high bandwidth networks (Koltay, 2017). Following the comments from Hautea, Dasgupta, and Hill (2017), it is important to outline that a study from 2013 has estimated that the world was producing 2.6 quintillion bytes of data per day and with that, over 90 per cent of the data in the world had been produced in the course of the previous 10 years. Most of this derives from basic interactions on the web, digital footprints or digital traces, which is the data that originates from the human interaction with computing systems.

Despite the substantial stimulus for the development of skills to handle data, the reality of most professionals does not seem to match the expectations of organisations, be it businesses in general, educational institutions, the academic community and a variety of other stakeholders. For example, schools are expected to make data-based decisions, when in reality educators do not feel prepared to do this (van Geel, Keuning, Visscher, & Fox, 2017). Companies expect new business professionals to have the data skills needed both in jobs that are and that are not data-centred, yet new professionals are not prepared accordingly. Big businesses spearhead and leverage the use of data, but most people have their data collected and processed without even realising it or understanding it. This kind of situation creates a context of discrepancy and imbalance (D'Ignazio, 2017). This is where data literacy, the set of knowledge related to data-skills, comes in.

This relatively new term has seen considerable development in the academic literature, however, despite the large adoption of data-use by organizations and their expectations for professionals to be able to do it, the literature shows there is a lack of preparedness in such skills. In addition to this, people use data-products daily without understanding it and how this lack of knowledge can also affect the effectuation of basic citizen needs like understanding how to make political choices or comprehending arguments by advertisers. How can there be such asymmetry in the knowledge of data? Moreover, what determines that someone is data literate and what can affect it? With these questions in mind, I aim to research data literacy by exploring the literature and testing a model that may help to clear some of the doubts in this field.

The initial part of this work is based on a systematic literature review, which is detailed later in chapter 4. At first, I provide a brief exploration of the literature by outlining how relevant data has become in society and why it warrants an investigation of data literacy. Sequentially, I present the state of the art of data literacy and what authors argue about it. The next chapter dives deeper into the concept by showing the different facets of data literacy presented in specific contexts. Chapter 4 presents the methodology for the systematic literature review I developed and showcases the key takeaways from the literature about what is involved in becoming data literate.

Next, I build on the knowledge gathered up to that point by structuring a proposal of a data literacy framework which shows the major dimensions involved in the concept as well as what elements can affect an individual's knowledge on the topic. The goal of this framework is to create a way of visualizing aspects that are involved in data literacy for an average person, without going deeper into skills that may be too specific to certain fields of study, such as advanced statistics, data management techniques and others. As such, the framework should be capable of being used to understand what is involved in the data literacy skills of an average citizen who can, for example, use data to apply the knowledge in everyday situations, such as understanding data presented in advertisements, processing raw data (the basics), making decisions and critically questioning data.

Then, I explain the survey and quiz that were used as a means to test the framework, later presenting the results obtained. Finally, the discussion on the findings, limitations, potential guidelines for future research and conclusion are presented. I stress that by no means this work intends to propose a final answer or an exhaustive framework, but rather highlight some of the most important findings, present an idea of what a sample of the people's data literacy may look like and provide insights so that more studies can be furthered in this domain. Data literacy is still a concept that is being developed and my contribution here is to present an alternative broad view of it and understand what is involved in it.

2. A Glimpse at the Development of Data in Society

This initial view for the following chapters aims to elucidate what data literacy is, and it was written based on the systematic literature review that I performed in this research, which is detailed further ahead in chapter 4.

To understand data literacy, we need first to comprehend a few premises, that is, its context, why data has had an increasingly bigger role in society, and where data literacy fits in that scenario. To start, it is important to outline that data is a broad term that can encompass any kind of information that can be stored in a digital form, whether it is a text, image, audio, video, software, animation, number or any other type of media (Koltay, 2017). Kristin Fontichiaro and Oehrli (2016) define that data encompasses: 1) the raw numeric representation of information, such as percentages and averages; 2) Information that can be used algorithmically, such as in the case of many common internet services or smartphone applications; 3) The visual representation of numerical information, considering tables, charts and similar means.

Studying data is relevant since its use has become growingly present in everyday life and applied in many different areas. According to Ebbeler, Poortman, Schildkamp, and Pieters (2017), the use of data in schools in the USA begun as early as in the 1990s and led to the existence of information systems capable of collecting, processing and storing data on a large scale. As technology develops and using data becomes easier and easier, data-related topics trend more, including the spark of new terms such as Big Data. Big Data is not only characterized by its size, but there is another element that also makes it especially attractive, which is the capacity to aggregate, search and cross-reference large data sets and all the applicability in different areas (Koltay, 2017). Several fields of study and work now have the means to optimize their work in myriads of ways with unprecedented ease. To Wolff, Kortuem, and Cavero (2015), big data is also defined by the rapid manner in which it is updated and achieves high volume, leading to considerable variations across time.

Miller (2016), a specialist from the company IBM, believes data is now all around us and impacting all people. It has become a driver for innovation. This new scenario has not only created several new professions such as the data scientist, but it has also brought extra dimensions to several professions, normally involving analytics and data analysis. As such, education must follow this trend. What is more, innovative companies such as Uber can be seen as a service that in its core leverages data, making use of digital maps, reviews, routes, images, profiles, locations, etc.

For businesses, the applications of harnessing data are a core part of any function or activity. Business decisions are often made by analysing existing data of a company, which led to a new field of acting often called business intelligence (Wolff et al., 2019). This field deals with the investigation and application of both human capabilities and technology to solve business problems by providing analysis to support decision-making and management (Ranjan, 2005). According to Pothier and Condon (2019), data is also used across organizations as a very powerful asset, since it can track their operations, customers, competitors, and the market, making it easier to make decisions based on it. For this reason, companies are now aiming to become data-centric, managing data in an effective way that can prevent the loss of resources. Data literacy lies at the heart of the skills required to instigate such change.

It is observed that the usage of data is not anymore secluded to a few select technical fields that may need it, but rather, it is increasingly becoming popular among varied segments of people. As such, data analysis is now often used to yields results for many different types of end-users, like teachers, students and administrators. Also, information must be treated in a way that can be easily communicated to others, often visualized through dashboards and charts, providing insights that can be actioned by decision-makers. In any case, for both the production and receiving sides of the analysis, data literacy is a necessary skill so that efficient communication may occur, and thus, the goals of the data analysis may be fulfilled (Wolff, Moore, Zdrahal, Hlosta, & Kuzilek, 2016).

In a scenario that values the knowledge of data considerably, it is expected that an overall ability to use it would be highly valued, which is where the idea of data literacy comes in. Gray, Gerlitz, and Bounegru (2018) suggest that data literacy in this century will become the most important new skill, enabling people to make the most of data and leverage capacities and technologies, which aids companies, states, and citizens. In a complementary manner, Rolf, Knutsson, and Ramberg (2019) quote that information and data literacy is a term considered one of the 5 areas that constitute digital competence in a framework designed by The European Digital Competence Framework for Citizens, alongside Communication and collaboration, digital content creation, safety, and problem-solving. In another example that highlights this competence, dealing with data has been pointed as one of 5 technical skills which information professionals should learn (Robinson & Bawden, 2017).

Despite all these considerations, data literacy is considered a new term that still needs development. Pothier and Condon (2019) affirm that the characterisation of data literacy has largely come from the studies of information science, as it is complimentary to information literacy. To Wang, Wu, and Huang (2019), the concept of data literacy is still emergent, having been proposed at the beginning of the XXI Century and supported by several organizations, such as the International Association for Social Science Information Services and Technology, the Association of Public Data Users, and the Interuniversity Consortium for Political and Social Research. However, they also mention that several authors have since then created diverse definitions for the term, such as the ability to comprehend, use and manage data or the ability to use data effectively to inform decisions. I will further explore definitions later on.

According to Gray et al. (2018), the United Nations affirms that data literacy can be a catalyst to promote change and make progress towards the future. Moreover, data is a valuable resource that enables the extraction of value in a multitude of contexts, such as economic, technological, social, democratic, etc.

Despite there not being a singular definition for this term, Wang et al. (2019) elucidate the following points in common: 1) Data literacy focuses in solving problems, formulating and answering questions through the collection and application of data; 2) Another notable point is the focus on the processing of data, such as the understanding, use, and management of data in accordance to the specific goals that require it; 3) The focus also lies in decision making, as a skill that allows decision-makers to transform data into information and ultimately into actionable knowledge and insights to guide and support the decision-making process; 4) The professional skills and learning skills of the person contribute to how data literacy affect him; 5) data literacy depends on critical thinking, influencing directly the way how the selection, evaluation, and analysis of data leads to decision-making.

Authors such as Dichev and Dicheva (2017) have taken a slightly different approach, promoting the idea of data science literacy, which entails some higher degree of emphasis on the computational, statistical and scientific aspect of the literacy, which comprehends elements like data management and processing, data modelling, machine learning, and visualization tools. Data Science can be understood as a multidisciplinary field which aims to analyse data and extract insights with the use of scientific processes, or it can also be defined as the set of components (theories, concepts, tools, technologies, processes) that enable raw data to be reviewed and produce analyses and information. In any case, data literacy is important as a means of acquiring the competencies necessary for the extraction of knowledge (Dichev & Dicheva, 2017).

For the above authors, this is a fundamental literacy that should be included in the early levels of education, because some degree of knowledge in this field will be important for several professions such as marketing, finance, politics, journalism, and others. Even for students who aim to pursue a career in areas not related to statistics and computers, data science literacy plays a role in helping us understand society and the environment better. We will later see that the existence of different “data literacies” is not uncommon, each one incorporating and putting emphasis on the specificities of a certain domain of study and their specific needs for data.

In this context, it is also worth mentioning that other types of literacies are connected to this topic, and, therefore, are important for the concept of data literacy. To (Markham, 2020), a literacy in something means a level of understanding and critical awareness that leads us to keep asking questions. Its requirements are curiosity, critical orientation and enough skills to start, and an information background that allows you to verify whether your curiosity is warranted.

Another two related terms, Information literacy and statistical literacy are often mixed with data literacy because of the topics involved, however, they bear some differences. Wang et al. (2019) argue that Data literacy also has to do with information-related matters, but it focuses on the functional ability regarding the collection, processing, management, evaluation, and application of data. They explain that some authors affirm that information literacy requires both data literacy and information management, which implies that if a person possesses good information literacy, it would also possess good data literacy. However, in the era of Big Data, several researchers, practitioners, and institutions have considered data literacy as an extension and expansion of information literacy instead of a smaller part. I will further address this topic later.

With all these considerations made, we can have a glimpse of the breadth of what the studies of data can entail by taking a look at how a University in London decided to structure a course on the subject. According to Robinson and Bawden (2017), the data course they structured has a starting point at the modern phenomenon of data deluge caused by the development of computer systems, which later leads to a need to understand the technology involved such as internet protocols, file formats, programming and query languages, contextual application, accessing databases, understanding the evolutionary trends of artificial intelligence, among other topics.

With this background in mind, it is assumable that the role data is playing is society has grown considerably. Changes in society require changes in our practices, as the rules of how things work are altered, so must the societal actors. This "datafication" process can be addressed in different manners, but a common answer is often data literacy (Gray et al., 2018). It is now needed to explain what is (or are) the actual meaning(s) of data literacy.

2.1 State of the Art: What It Means to Be Data Literate

The key takeaways from this introductory approach that contextualizes the need to develop data-related skills and how recent data literacy leads us to the idea that the concept is still under development, there being different definitions and skills which have not yet been standardized (Pothier & Condon, 2019). Despite the lack of standardization, it is possible to identify core elements that involve the idea of data literacy in the literature, which was achieved here by a systematic literature review in which details are displayed further in this work. Before beginning this topic, I highlight that Mandinach and Gummer are two key authors in the field of education whose contributions are widely quoted in other articles. Their research and gatherings with other scholars have yielded significant development in the field.

With the premises that revolve around the notion of data literacy in mind, we can outline a few core definitions. In the field of education, Mandinach and Gummer (2015) say that data literacy can be perceived as "the collection, examination, analysis, and interpretation of data to inform some sort of decision in an educational setting". This definition can serve as a starting point from where I can develop the concept. However, Koltay (2017), in his research, stresses several different definitions which involve those concepts, but also add other capabilities such as being able to summarize and prioritize data, developing hypotheses, identifying problems, manage, assess, handle and ethically use data, processing and filtering it, knowing how to search and store it and many other processes involving the general use of data, which means to us that it is important to not take the existing definitions as a given final answer, but rather interpret all of them under a common light.

When it comes to the skills involved, IBM and Oceans of Data's list of competencies mentions knowing how to define problems, wrangle data, self-manage it, choose methods and tools, analyse the data, communicate findings and engage in lifelong learning (Miller, 2016). This last item provides an interesting insight as to the impact that evolving technology and the development of the concept have since technology progression can outdate previously existing knowledge.

Mandinach and Gummer (2015) also mention five recommendations that can constitute a roadmap of the components that are part of the implementation of data use in educational institutions: "(a) create a cycle of inquiry; (b) make students their own data-driven decision-makers; (c) develop an explicit vision for data use; (d) enculturate data use through the provision of necessary supports; and (e) provide a data system". Other important points these authors mention are leadership, appropriate technology, and ongoing professional development. Others such as Duffner-Ylvestedt and Rayner (2016) defend that students should be capable of explaining about open data and how it influences science, finding relevant data for their field, apply critical thinking to data and understand the challenges of data use, among other aspects.

Referring to the business context, Pothier and Condon (2019) identify seven data literacy competencies, which are 1) data organization and storage, which is necessary because data inside companies are handled by a wide diversity of staff and departments, and because of that, organization and clear processes are vital for efficiency; 2) understanding data used in business contexts, which is a contextual application of data and it involves being able to understand the usefulness of data, its origin, appropriateness and general aspects that would be important to drive decision-making within the organisation; 3) evaluating the quality of data sources, a competency that enables professionals to assess the quality of the data, allowing the subsequent interpretation and decision-making to have a better foundation; 4) interpreting data, a skill that prepares business professionals to take actions based on the analysis of data; 5) data-driven decision making, an item which is responsible for converting data into actionable information and implementing solutions while weighing the positive and negative points found; 6) communicating and presenting effectively data, a competency that relies on the business professionals to convey complex ideas and create a coherent narrative according to the audience's level of familiarity with the topic; 7) data ethics and security, a skill that addresses current concerns with how to handle people's data, an important concern for companies as their reputations can be severely affected when the of individuals is mishandled.

D’Ignazio (2017) states that data literacy involves skills such as reading, working with, analysing, and arguing with data, all of which are part of a wider process of inquiring into the world. She says that while there is a lack of consistent approaches when it comes to helping new people to acquire data literacy, many choose a technically centred approach, relying on technical skills and statistical skills. However, the author believes this should not be the case, as there is a need to connect the skills with concepts of citizenship and empowerment.

These arguments are not distant from each other, which is why the authors presented so far define a more or less cohesive view on data literacy with varying aspects. Dai (2020) summarizes this state of the art by saying there are 3 different approaches to data literacy: 1) data literacy can be considered as the application of critical thinking to the use of data, coupled with statistical knowledge; 2) the second approach builds from information literacy and provides more focus to databases, data management, documentation and standards, preservation and others, aspects which can be linked to the topic of librarians presented later; 3) the last approach deals with the data lifecycle, from the collection to the use of data, being closer to the ideas of Mandinach and Gummer. The fact that there are varied ways of categorizing data literacy, each with a different focus, is particularly relevant to the arguments that will later be further developed.

Finally, I emphasize how authors often mention other ideas associated with data literacy depending on the specific field of focus they research. Koltay (2015), for example, presents a definition of data literacy in a higher level of complexity and often with elements that prevail in the fields of research and librarianship, aspects that are akin to Dai's (2020) second approach. Grillenberger and Romeike (2018), in their article about a theoretically founded data literacy competency, derive from the perspective of computer science and data science courses, including skillsets such as modelling, partitioning, design principles and others which are rather limited to a select population but do not represent the broader scope of the literacy components on a foundational level which most people need to know. Halliday (2019) says data literacy requires skills related to data management, visualization, and computation of quantitative results to generate information with the data acquired. Authors such as D’Ignazio (2017) and Markham (2020) tackle a wider scope and more accessible version of data literacy based on critical reasoning, which could be linked to Dai's (2020) first approach. It can be observed that there are both universal competencies and others which are broad but can also be customized in different contexts, despite this not being always easy (Pothier & Condon, 2019). Each approach presents itself as valid and warrants deeper investigation.

One can notice that there are related topics in each definition, as well as some which are exclusive to some definitions and absent in others. More important than summarizing the concept of data literacy in a set of words is the perception of what it actually entails, and in which ways it can be used, as well as other underlying "why's and how's" involved in it. Considering the many sources that can be observed when building a framework for Data Literacy, I aim to further explore what is encompassed in the idea of what it means to be data literate for most people and what it takes. The main conclusion to be drawn from these observations is that the debate on the meaning of data literacy is still developing, despite the existence of common grounds. However, as a complex term with many facts, limiting the analysis to these conceptualizations will not suffice for a more comprehensive approach that also addresses some of the peculiarities mentioned by authors. Next, I try to take a step further by understanding how data literacy compares to similar terms in the literature.

2.2 Overlapping Domains of Data, Information and Statistical Literacies

Designing a framework for data literacy requires addressing the overlapping topics this field has with information literacy and some other fields. Also here there are several visions about this, however, informational and statistical literacy are the most often discussed literacies that are highly connected to data literacy.

One vision is that statistical literacy involves components such as knowing the ways to calculate averages, differentiate correlations and causation, comprehending margins of errors and biases presented in data (Kirstin Fontichiaro & Oehrli, 2016). Robert Gould (2017) emphasizes that when it comes to statistics, it is necessary to differentiate the needs of what consumers and producers need to know. For him, there is a technical dimension of understanding statistics per se, as well as overall knowledge of how they are applied. Citizenship, for example, is mentioned as a reason to develop these skills, since the debates in society often involve this kind of knowledge. The associated skills, therefore, can involve knowing who are the major actors involved in the process of data collection and how and why they do it, comprehending matters regarding storage, privacy, modelling, and origin, as well as being able to process data and understand different the different ways it can happen. He also argues that ongoing definitions of data literacy seem to incorporate statistical literacy and go beyond that, addressing additional matters. He believes a possible reason for this difference has to do with the background of those who are spearheading the development of the concept, which, as we will further see, has gained a lot of contribution from information literacy and the studies of librarians.

Koltay (2016) says information literacy is a field which researchers and librarians have dealt with for a longer time and its concept has been more widely accepted, which implies in recognizing the need for information, knowing how to solve problems with information, critically assessing data and its sources, data management processes, comprehending socio-cultural elements in information and the context where it is applied, etc.

In addition to that, Pothier and Condon (2019) state that the literature also shows variations in the term data literacy, taking into account the focus each author had in their respective fields of study, giving birth to concepts such as data information literacy, research data literacy, science data literacy, and others. Each term reflects its own different approach, that adds emphasis to elements that are more relevant in each specific context. Authors such as Duffner-Ylvestedt and Rayner (2016) quote the belief that data literacy is (an integral) part of information literacy, while others such as Kjelvik and Schultheis (2019) and Gray et al. (2018) place data literacy at the intersection point of a range of skills. In the case of the latter authors, data literacy is at the intersection of statistical literacy, information literacy, and technical skills.

Based on Koltay (2017), it is reasonable to argue that attempting to provide a strict differentiation among each concept adds little to this work. Literacies are naturally multifaceted and making a statement that one is part of the other does not change their functionality. The very boundaries between each literacy have not been disclosed by academic literature. Instead, emphasis should be on the fact that information literacy (an older concept) by itself does not suffice to deal with the myriad of situations involving data previously described.

These general components of data literacy involving skills such as collection, analysis, interpretation, data-driven decision-making, and others will be further explored throughout this work. Nonetheless, my research found that many authors (which I will gradually present) place critical reasoning as a fundamental component, being the reason why it is important to stress this at an early stage of this study. Among the many possible examples provide, when Gray et al. (2018) propose their notion of “data infrastructure literacy” this becomes evident. For them, data literacy is very often presented as a combination of technical capacities (some of which I already argued), but they propose instead an expanded concept that also includes a component related to critical thinking, which deals with understanding and using the infrastructure in which data is produced, stored, used, analysed and shared, also promoting inquiry, imagination and intervention. This means understanding that humans are behind data and what we see is not a clear representation of nature and things as they are, but rather the final output of a decision process made by people. The author calls this idea as data infrastructure literacy, as it relies on understanding the functioning of the many infrastructures where we use data. I will revisit this concept later when focusing on critical reasoning.

Once again, we see that data literacy is described in very similar ways by authors, but a few new elements are cited in different researches, depending on the specific field of study in which we are analysing data literacy. This is a complex term, with varied interpretations, different lists of skills, and a lack of standardized definition (Klenke, Schultz, Tokarz, & Azadbakht, 2020). Nevertheless, there is a core of general skills that can be moulded to the needs of the context and how the skill has to be applied. With that said, a generic view can only outline the broader concept without addressing important elements that only exist in specific contexts and these elements help to understand some elements of data literacy that only appear when analysing them under a contextual light. Considering the extant literature, I believe that diving into a few of the most relevant domains of study sheds an important light of how the generic view can unfold when contextually applied. Thus, in the next topic, I will cover some of the most common fields of applications I found through the systematic literature review to explore this theme on a deeper level.

3. A Further Dive on the Concepts of Data Literacy and Its Applications Across Different Fields

I chose to structure the main fields here with a section regarding the general public and 3 topics dedicated to specific professional applications. While the topics for “business activities” and “researchers and librarians” could potentially be capable of addressing most fields in a generic approach, I created a topic for “education” due to how relevant it was in the literature research, how it relates to both academic and professional backgrounds and how important it is for most citizens when considering that this generic term covers from basic education to university level.

3.1 Education

As stated, the field of education receives considerable attention in the literature about data literacy and it is not hard to understand why it has such relevance. As Kirstin Fontichiaro and Oehrli (2016) mention, students notice numbers very early and have to deal with them in their first projects, be it related to populations, consumerism and many activities, which leads them to know that the data presented in numbers is a powerful mean to convey information. While the usage of data in education is not a new phenomenon, there is a considerable deficit in its application to improve students’ learning, sometimes because data collection is not done systematically or even because teachers may sometimes not have the necessary knowledge (Phanchalaem, Sujiva, & Tangdhanakanond, 2016).

The importance of data literacy for education can be considered clear, as educators make decisions daily, choosing how to guide students, how to adjust their practices, understanding how students are impacted, among other decisions. For this reason, Kippers, Poortman, Schildkamp, and Visscher (2018) say educators can benefit from using data-based decision making (DBDM), which has the potential to create high-quality decisions based on data, such as the identification of students strengths and weaknesses.

In general, the last decade has shown a considerable amount of debate regarding data-driven decision making and evidence-based practices in education, despite being a relatively new topic. According to Mandinach and Gummer (2015), while there is a lack of concrete evidence about whether data-based decisions positively affect teachers’ practice and students’ performance, the existing researches can provide many insights. Their research on the theme across several years led to the premises that it is unquestionable that educators must be armed with data. Also, there is a problematic conflation between assessment literacy and data literacy, which are often perceived as the same, since some teachers believe data literacy is only about assessments, despite it not being the case. The authors believe assessment literacy is a smaller element, which is included in the larger picture of data literacy. Moreover, they understand that teacher preparation for data use does not suffice for proper data literacy. Thus, their idea of Data Literacy For Teachers (DLFT) comprehends both the data skills and content knowledge and pedagogical content knowledge. The latter two are more related to the application of data literacy in their specific context, something I explore later.

Mandinach and Gummer (2016) also sought to analyse how American institutions handle the matter of Data literacy. The definition of a data literate educator, in accordance to experts who participated in the Data Quality Campaign, is of a professional that continuously, effectively and ethically accesses, interprets, actions, and communicates multiple types of data from variable sources to improve students' results. The Data Quality Campaign is a bipartisan advocacy organization which has been supporting the usage of data by teachers, awareness of data literacy and differentiation between assessment literacy and data literacy.

Regarding the role of North American states in addressing the topic, the authors mention that, according to the North Carolina Department of Public Instruction, data literacy can be defined as the level someone has regarding how to find, evaluate and use data to inform instructions. This concept is applied in a simple way to instruct educators on the purpose of data literacy. Moreover, it is observed by the authors that the state of North Carolina is the only in the USA concerned with data literacy to the point of creating a webpage to elucidate such concept, but it is not the only one to address the concept. The state of Virginia, for example, seeks to promote data literacy skills. Moreover, in general, US states appear to give more attention to assessment literacy.

Educational organizations like the Council of Chief State School Officers, the Council for Accreditation of Educator Preparation and the National Board of Professional Teaching Standards all have advocated in favour of data literacy as an important component of teacher's education, which also includes knowing how to analyse student learning needs and adjust teaching in accordance to existing data. And while data literacy is widely considered important, this difference in the levels of presence and how concerned policies are about this also happens across different countries. The authors state that the UK and Poland, for example, have reported having data skills, while Germany, Lithuania and the Netherlands appear to lack it, but in either case, none of these countries shows deep and rooted evidence of data literacy and the many aspects that revolve around it.

Piro, Dunlap, and Shutt (2014) emphasize how the in the United States there are data systems implemented which been able to track students' knowledge in a given moment and make recommendations based on what would be needed to improve. They also report that there has been a positive link between the use of data in instructional decisions and improvement in students' achievements.

Kirstin Fontichiaro and Oehrli (2016) affirm that students are expected to be able to use data fluently, involving the collection, analysis, use of tables and figures and overall representation of data in text and visual representation, an expectation that can be partially attributed to how states are moving forward with standards defined by educational institutions, such as the College, Career, and Civic Life (C3) Framework for Social Studies State Standards, Common Core State Standards (CCSS), and Next Generation Science Standards (NGSS). However, they argue that students often cannot make a strong sense of the meaning of the numbers and they treat it as an objective representation of reality.

Chin, Blair, and Schwartz (2016) highlight a few domains within some of those standards which involve data literacy skills. Considering the CCSS, there is "Math Practice: Model with mathematics", which consists of being able to deal with quantities in practical situations and mapping their relationships with the usage of tools, tables, graphs and others. Furthermore, there is "English Language Arts: reading science and technical subjects", encompassing the translation of quantitative and technical information to different forms, such as text, visual or mathematical. When it comes to the NGSS, there are 8 practices which according to their website[1] are descriptions of behaviours applied by scientists in their practices. The authors emphasized "Analysing and interpreting data", "Use mathematical and computational thinking" and "Obtaining, evaluating and communicating information". This seems directly related to how some authors elaborate on their framework of data literacy, which we will come back to later.

The content presented up to this point would emphasize the importance of data literacy in the educational background, but empirical studies seem to point to a lack thereof.

[1] It can be consulted here: https://ngss.nsta.org/PracticesFull.aspx

For instance, teachers have an understanding of how to use data to a degree but do not have specific skills that would promote effectiveness. One finding shows that the existence of data teams in which teachers can participate facilitates handling of data by combining both a data specialist and teachers as the professionals who need a better application of data (Mandinach & Gummer, 2015). Asides from that, they point out that some elements related to the profession, namely content knowledge and pedagogical content knowledge affect directly the capacity to use data for teachers and in the classroom. Horizontal and vertical expertise on the field of data also presents as a contributor to a teacher's performance in data using, which supports the idea that other stakeholders around the teacher influence his levels of data literacy. In a similar understanding, Ebbeler et al. (2017) mention that despite many schools having a considerable amount of quantitative and qualitative data, such as voice recordings regarding students' attention to homework or instruction quality, educators use mostly summative data about students evaluations performance, which by itself provides limited understanding and insights. It is now common practice that summative data is being used for accountability of educational institutions, and it is not limited to the United States (where data systems have been largely implemented to track the progress of schools) but also in a range of other countries, many of them in Europe (Piro et al., 2014).

With all these considerations in their studies, Mandinach and Gummer (2016), joined by other educators in the analysis of the scientific literature and government documents regarding data literacy, concluded that the framework of DLFT is comprised of the following domains: 1) identify problems/frame questions; 2) use data; 3) transform data into information; 4) transform information into decisions; 5) evaluate outcomes. This is a cyclical process which is often mentioned in the literature, albeit with a few variations. I will further refer to it as the data cycle.

Furthermore, the framework is also influenced by elements related to the area of teaching which are added to the framework to also incorporate other forms of knowledge which are broadly accepted in the field of education as essential to good teaching, totalizing this 7 final elements: 1) content knowledge; 2) general pedagogical knowledge; 3) curriculum knowledge; 4) pedagogical content knowledge; 5) knowledge of learners and their characteristics; 6) knowledge of educational contexts; 7) knowledge of educational ends, purposes and values. The image the authors present to depict the conceptual framework can be found in Figure 1:

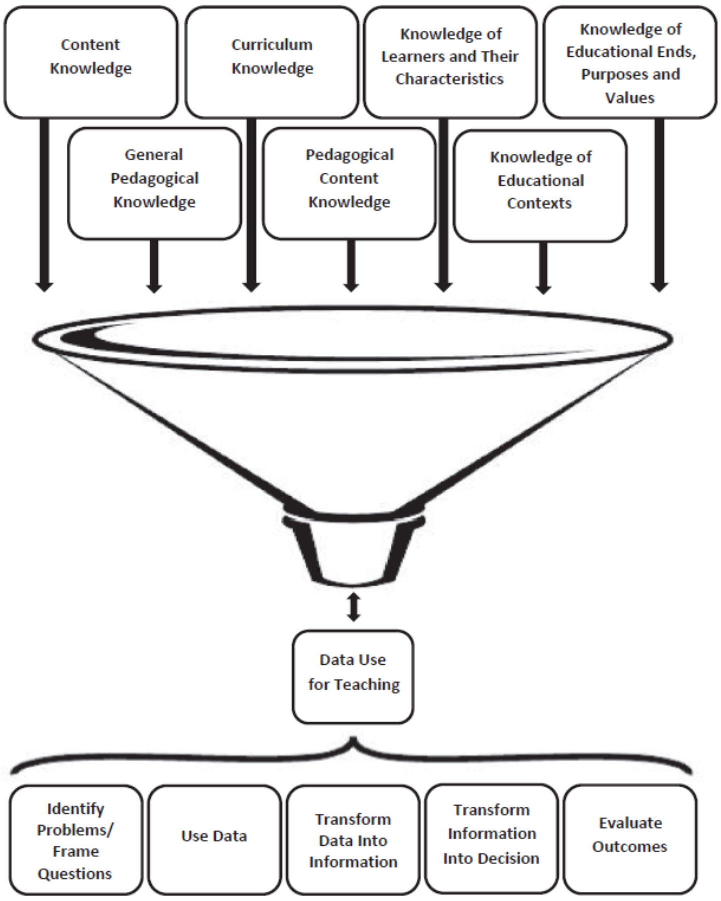

Content Knowledge
Curriculum Knowledge
Knowledge of Learners and Their Characteristics
Knowledge of Educational Ends, Purposes and Values
General Pedagogical Knowledge
Pedagogical Content Knowledge
Knowledge of Educational Contexts
Data Use for Teaching
Identify Problems/ Frame Questions
Use Data
Transform Data Into Information
Transform Information Into Decision
Evaluate Outcomes

Figure 1: Conceptual DLFT Framework proposed by Mandinach and Gummer (2016)

It is worth explicating that the bidirectional arrows below the funnel in Figure 1 indicate that data skills influence and are influenced by the seven forms of knowledge. This model places some emphasis on how the specific elements of the teaching profession matter for data literacy, meaning that generic data knowledge is limited without knowing how the profession works. The description of each of the 5 points of data knowledge will be further explored in the model proposed by this work further (Chapter: 4.2.1 components of data literacy).

With that said, we can summarize this by saying that Mandinach and Gummer (2016) consider that data literacy for education/teaching is the ability to transform information into actionable instructional knowledge and practices, which can be done by the collection, analysis and interpretation of all types of data to help us make instructional steps, combining data understanding with patterns, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge and an understanding of how children learn. They also believe data use must be introduced early in teacher’s education, as it may be late to do so when they are already practising. This introduction must also preferably not occur through standalone courses but rather integrated into their regular preparation, as this is part of the training of the profession and not an extra separated item. This statement is aligned with the studies from Dunlap and Piro (2016), who claim that teachers in their research feel unprepared when they first begin practising and express discomfort with the use of data. They also affirmed that as teachers' confidence and sense of self-efficacy regarding data grows, the more likely it is for them to use data in their practices.

Another framework that supports a similar vision as the ones presented so far is the one proposed by Kippers et al. (2018), in which they define the components of data literacy as 1) setting a purpose; 2) collecting data; 3) analysing data; 4) interpreting data; 5) and tracking instructional action. All of them also constitute a cycle. Each component is involved with at least one of 8 steps, which range from the definition of the problem to the implementation of measures and evaluation. Their model can be found in Figure 2: Data literacy framework proposed by *Kippers et al. (2018)*:

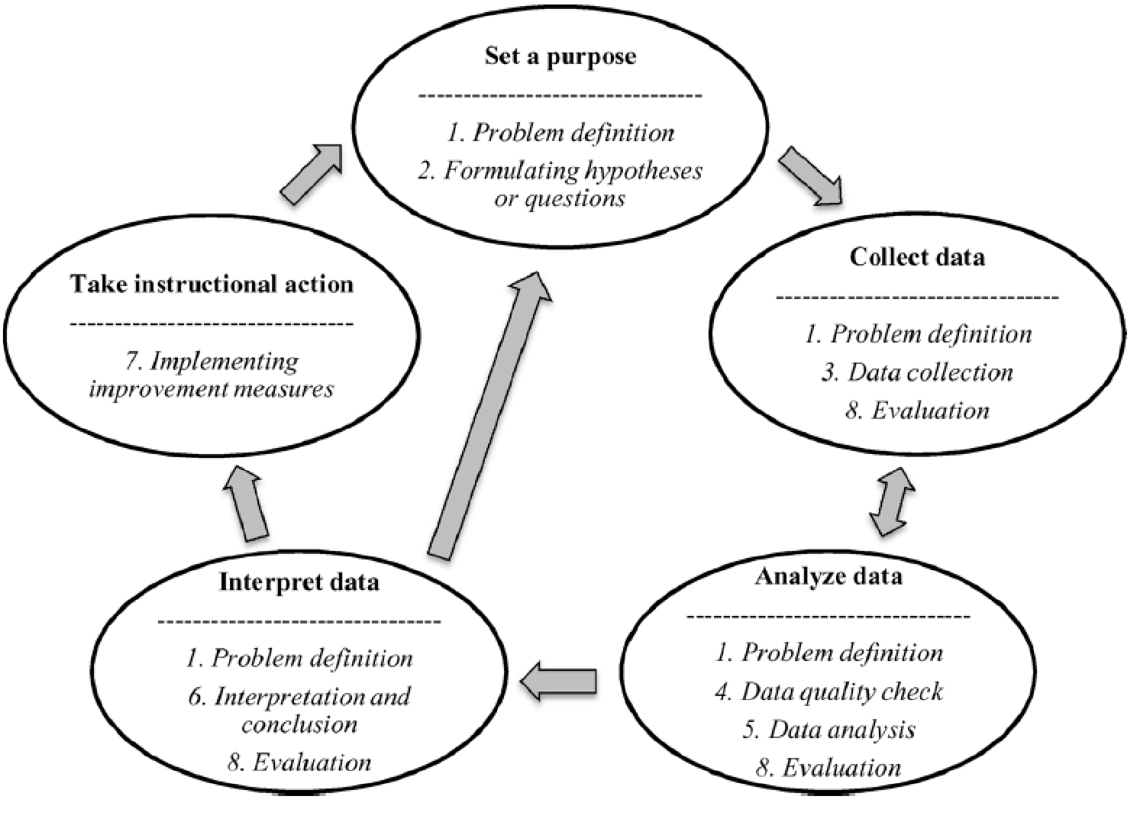
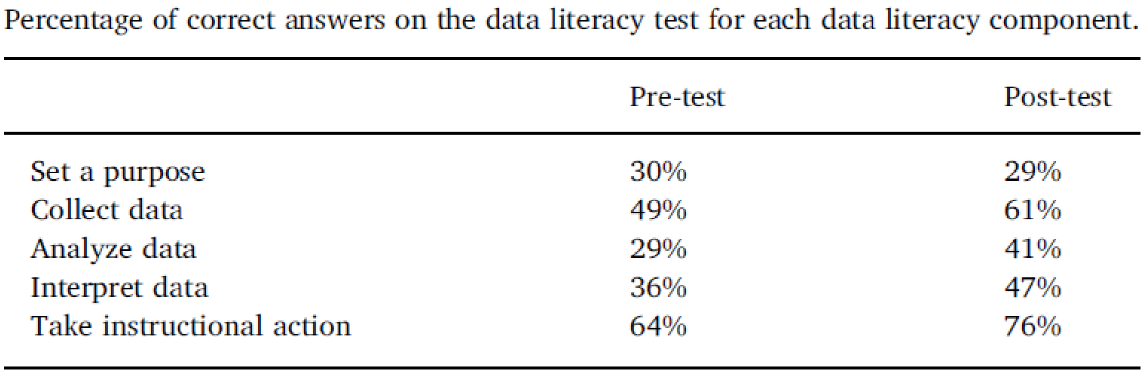


Figure 2: Data literacy framework proposed by *Kippers et al. (2018)*

This framework was used by the authors in a data use intervention in 6 secondary Dutch schools to assess how teachers coped with data and how their results were affected after the intervention. With the aid of specialists, teachers have done a pre-test on their knowledge regarding data literacy, followed by training across the period of a year and finally a post-test. The results of the test are shown in Table 1:

Table 1: Results from the data intervention in the studies by Kippers et al. (2018)



Percentage of correct answers on the data literacy test for each data literacy component.

	Pre-test	Post-test
Set a purpose	30%	29%
Collect data	49%	61%
Analyze data	29%	41%
Interpret data	36%	47%
Take instructional action	64%	76%

It is noticeable that there was an overall improvement in each of the analysed areas after the training, however, the authors believe the results could still improve a lot. The educators seemed to struggle notably with the formulation of adequate hypotheses and questions and set purposes, while on the other hand, using technology, such as Microsoft Excel, to create sheets and graphs and perform statistical analyses was an improvement, as some could not do it before. Curiously, this result is somewhat aligned with the research made by Phanchalaem et al. (2016) in Thailand, which concluded that teachers needed improvements in data analysis but not as much in the use of technology to analyse and store data.

One of the reasons for the lack of data literacy for educators is that some educational systems emphasize the theoretical component over the practical, coupled with the lack of use of data in courses and of teacher preparing. New educators are expected to be able to use student data as a means to improve effectiveness, which is why it is necessary to train teachers to develop data skills (Piro et al., 2014). Teal et al. (2015) have argued that in many cases teachers claim that one of the most fundamental problems to developing data literacy lies at the amount of training that is dedicated to it, not to mention that institutions are usually with their curriculum full and with priorities fully allocated towards other areas, becoming difficult to spare resources for data literacy.

Taking an approach from the studies of life science, Gibson and Mourad (2018) present a similar view of data literacy. For them, the foundation of this skill lies in the use of skills such as mathematic functions, modelling and data presentation, but they also constitute a broader field, including: 1) knowing what tools to use; 2) understanding their applications in the biological context; 3) interpreting the data based on the underlying question or hypothesis; 4) communicating results across varied platforms.

On a concluding remark, Kippers et al. (2018). suggest, among other things, that potential improvements could be obtained by bringing the tasks of data interventions more closely to teacher's daily practice on a micro-level, which reinforces the contextual aspect of learning about data. Moreover, Mandinach and Gummer's approach to introducing data early in teacher education should contribute to better-developed data literacy.

3.2 Business Activities

As it has been cited before, the applications of data literacy across several business activities can be easily identified, as data analysis can be used in marketing, finance, sales and a range of other professions, not to mention data-centred positions such as analysts and data scientists. I chose to talk about these last two jobs inside this topic and in different places across this work instead of a dedicated section due to the interconnectedness of their work and the many points I make along the dissertation.

Despite the importance placed on data literacy in a business context, in the systematic literature review made for this work, which consisted on a total of 138 academic articles, while this was mentioned in a few articles, surprisingly not many addressed the topic directly and going into further details. In one of the few articles found that had this focus, Pothier and Condon (2019) present several arguments to stress the importance of data literacy for businesses nowadays. Companies are in a strategic position in this big data scenario as they are massively involved in the production of data, which can be seen in card transactions, social media and website analytics, all of which are examples of data that can be leveraged and translated into understanding markets and consumers, increasing efficiency, making decisions, decreasing expenses and increasing market share. And this phenomenon is not restricted to large companies, as even the smaller ones generate data and have access to a range of tools that can be used.

Pothier and Condon (2019) adopt the position that being data-centric means how companies invest to capitalise on the value of data. A data-centric organisation can be determined by its technological, financial, and human resources. But while researches indicate that companies intend to become “extremely” data-centric, many of them believe they are currently at much lower levels than that. The authors indicate one of the reasons being that many companies focus on the technology needed to handle data, since having infrastructure is necessary, but to achieve a status of extremely data-centric this would not suffice. The part that pertains to human resources needs to be coupled with technology.

Work positions for data analysts and data management are in high demand, as they are key to data-centric operations, however, this need for data literacy does not lie exclusively on the roles of data scientists, business intelligence or data analysis, but rather, they are important for employees in all departments of a company, albeit on a smaller scale, which is why it should be a relevant skill for business students who want to meet modern workforce demands. Notwithstanding the abovementioned importance of data literacy for organisations and professionals, the authors noted that reality seems to differ in regards to the actual skills people have, as the new professionals are not being properly prepared for the types positions companies are offering on that sense. While companies are making a move towards data, candidates on the job market do not have the data skills required, to the point that poor data literacy and lack of relevant skills or staff are often quoted as major challenges in organisations, an occurrence which is not restricted to data-related jobs, as several types of professionals like the ones in the business field need a foundational understanding of how data is used in the business context, such as in marketing, human resources, finance, etc (Pothier & Condon, 2019).

Moreover, the demand for data-related jobs is not being met by the available supply, both at early career and senior levels, and this is slowing the progress of companies in areas which deal directly with data and those that do not. The expectations of employers for new graduates regarding skills and preparedness is not accurately matched by the actual skills students are developing, an argument which the authors further support by quoting a survey in which among 63000 managers surveyed, 36% claimed there was a lack in data analysis skill of new graduates, namely in the use of tools such as Excel, Tableau and R. This fact calls for the insertion of data skills in educational curriculums, however, the authors cite surveys which indicate that, while most MBA applicants expect to learn a considerable amount of data analytics skills in their programs, the top MBA programs in the United States and the United Kingdom do not offer a corresponding amount of data topics. Pothier and Condon (2019) then affirm that data literacy competencies have been mostly discussed with a focus on sciences, while some academic and professional areas have not received clear detailing, as the literature still needs development. All of this is evidence that the literature needs attention in regards to data literacy in a business context.

Another article found in the systematic literature review that deals extensively with a business context but specifically in the field of safety was developed by Wang et al. (2019), where it is possible to observe some similar occurrences when it comes to the characteristics needed for a safety professional to be considered data literate. To them, safety management has become a data-oriented field having as one of its main resources the Safety-Related Data (SRD). This area has been directly affected by the addition of technological improvements, the growing production of data by companies, the construction of databases full of data, the usage of systems, government regulations and a whole set of factors that contribute to the Big Data in the field of safety and at the same time reflects the need reflect the need for the adoption of sophisticated methods to deal with all the existing data.

SRD’s have shown to be particularly important because, when they are appropriate, they possess an invaluable capacity to improve performance in safety, preventing, identifying and controlling risks to safety, unsafe conditions and variable other factors that would previously be hard or impossible to do. On the other hand, SRD’s that are not actionable are of little value, leading to a potential loss of resources. Therefore, data-driven safety management presents itself as a supporting mechanism in the process of safety decision-making by the effective management of SRD’s.

However, the vast and growing amount of available data also leads to hardships for effective management, interpretation and adequate use of information. In that context, despite the advances in the discussions regarding safety-data management, for these authors, little is mentioned about Data Literacy for Safety Professionals (DLSP), a skill so fundamental for a safety professional that some companies consider it a hiring criterion that safety professionals be capable of collecting and applying SRD’s to promote safety solutions (Wang et al., 2019).

As such, the authors consider that the current scenario for safety management presents the following tendencies and challenges: 1) the wide usage of technology; 2) the fast development of SRD systems; 3) the huge amount of SRD generated; 4) the growing importance of SRD in safety management decisions; 5) the ascension of the implementation of data-driven safety management directly related to activities concerning SRD's; 6) the increasing attention organizations are giving to the development of DLSP; 7) the growing importance of DLSP; 8) The lack of deep theoretical studies about DLSP. Notwithstanding the lack of specific literature, much can be learned with the data-related tendencies from other professional fields in which it has become clearer what data literacy means, such as medicine, education and business.

DLSP derives of a combination of SRD knowledge and using skills that involve them. The definition of DLSP would thus derive from the very definition of data literacy, being it the ability to collect and analyse SRD's, using them as evidence for safety management. Or even, based on the framework of data literacy for educators, DLSP is a mix of skills, knowledge and dispositions that safety professionals must possess to be able to collect, evaluate, analyse, interpret and utilize SRD's effectively and responsibly in safety management (Wang et al., 2019). To sum up, the process of the application of SRD's starts by the identification of the problem or inquiry about a safety management problem, and then comes the SRD collection, its quality evaluation, followed by the analysis of the data, their interpretation and the process of decision-making, implementing improvement measures and finally, the evaluation, which is once again followed by the identification of the problem because this is a cyclical process.

In the field of safety, Wang et al. (2019) present 5 main factors a safety company should look for: 1) The collection of variable SRD, namely the ones related to the allocation of results that aim to improve safety, the process that is related to SRD, the results originated from the safety measures and overall satisfaction with these practices; 2) The presence of SRD's without a refinement that can lead to actionable insights is not useful, and therein lies the need for the analysis of the data for that purpose; 3) The manipulation process of SRD's must occur with an orientation that they have to be used to meet the safety management strategies and improve the effective, critical and ethic use of data-based safety; 4) SRD's can lead to new management strategies, such as evaluating the enforcement of regulations, determining new types of required training, prevention mechanisms, and others; 5) Safety-related decisions typically involve SRD's in two ways. Either they are needed to identify or clarify situations (identify risks, establish goals), or they are used to act (implement new policies, acquiring new equipment, reallocate resources).

Following that reasoning, the authors define that DLSP is defined by 4 main components: 1) General data knowledge and skills, involving statistics, software and hardware, concepts and methods related to this knowledge etc.; 2) Overall knowledge and skills related to SRD's, such as its concept, types, functionalities, applications, management capacity and others; 3) Attitudes, beliefs and awareness of SRD's, mainly focusing on the notion that they are useful for safety management and how they can be applied; 4) Use of SRD for safety management, showing a mastery of data skills to generate concrete results which can be applied in the domain of safety. Based on that, they allege that companies must develop DLSP training and education with the two main goals. The first, to make safety professionals data literate so that they can become effective safety managers in times of big data. The second is that they become professionals who can manage SRD's. In any case, the training must enable them in all processes that involve operating data so that they can be effective data-driven safety managers. Based on these points, the authors add that the content of DLSP, from a macro perspective, includes knowledge in the area of data science and safety science or management.

The main observable points are that, in comparison to the domains of science, not much is developed in the literature that addresses the mismatch between the needs of businesses for data literate professionals and what is developed in educational institutions. The work made by Wang et al. (2019) was a notable exception on the existing literature and it seems that encouragement to produce more of such work would be beneficial. Moreover, data is argued to be necessary not only in data-centric jobs but in virtually any work position in an organisation. This apparent gap in the literature will be one of the guidelines for the posterior research done here.

3.3 Researchers and Librarians

As Gibson and Mourad (2018) affirm, science is a data-driven process. Moreover, with the increase in the number of tools and overall means to collect, analyse and share data, scientists now require a growing familiarity with topics related to data science. In the field of biology, for example, genomic and geospatial data are some of its applications. Nonetheless, the literature also points to a lack of skills here, despite the discussions being more present.

Data used to be a resource for researchers mostly used by computer scientists, but it is now used across varied fields of study, developing what some consider a new research paradigm of data-intense scientific discovery (Grillenberger & Romeike, 2018). As digital research gains more attention, researchers across several areas, such as engineering, social sciences and humanities, start to interest themselves more about data, since the current technological context allows for new questions to be made and new ways to research. This process requires tools, infrastructure, processes and skilled personnel (Koltay, 2015).

In a University context, Duffner-Ylvestedt and Rayner (2016) affirm that students are potential future researchers, and therefore, they need to become data literate early. This involves being able to explain open data, finding data to their field of study and explaining its connection in the process of publication (contextual application), apply critical thinking and understand the challenges of reusing data. Additionally, after surveying 73 professors at the Uppsala University about the most important concepts of data literacy that students should be taught, it resulted that 21% chose “finding data repositories within the field” (data collection), 19% mentioned “data citing” (the practical application in the research domain), 16% chose “Ethical skills”, further 16% chose “Presenting tools for data visualization” and 11% said “How to evaluate data sets from repositories” (evaluation of the data), with the rest referring to other topics.

The role of librarians and the library has been quoted often in the systematic literature review, but most especially when connected to researches. Just like a library traditionally facilitates access to documents, access to data can now also be facilitated (Koltay, 2015). In this scenario, the role of libraries and librarians (especially academic ones) gains an extra layer of importance as authors such as Koltay (2017) argue. To him, researchers have not received adequate training regarding management and curation of data and have to learn on the job as they need it. On the other hand, libraries have shown they possess the capability to be more than repositories of reading materials but also include data services for research.

Koltay (2015) mentions the Association of Research Libraries stresses the position of libraries as relevant actors when it comes to providing data services and having experience with it. Librarians are trained to be familiar with the research data needs of researchers and can, therefore, have a key supporting role in research. “Research Data Services” is a term that includes both data management and data curation, which while different, do not have a clear boundary between them. Data curation, specifically, has been identified as one of the top trends for academic libraries more than once. This link is not a surprise, since, as Pothier and Condon (2019) say, information literacy (which as we asserted, is highly related to data literacy) has been associated with academic librarians. Not only librarians hold this position, but also, as Verbakel and Grootveld (2016) name it, all types of “data supporters” in general have a role in this whole process, which includes other professions such as research support officers, data stewards and others.

Similarly, Kirstin Fontichiaro and Oehrli (2016) state that school librarians are unique cross-disciplinary pollinators which can aid students with the understanding of data-related matters, pointing out 6 major themes which can be explored to improve their role in building data literacy, which are: 1) statistical literacy, comprehending both methods and practical applications; 2) data visualization, involving both understanding and making representations of data; 3) data in arguments, which reflects how we cope with arguments based on data; 4) big data and citizen science, helping people to understand the growing amount of data collected from us, often without our awareness; 5) personal data management, teaching students how to navigate the web with the proper notions of how platforms like Google and Facebook use data, such as showing ads and recommendation; 6) ethical data use, an item which takes into account how data can be used in misleading and unethical purposes. Verbakel and Grootveld (2016) also cite skills such as knowledge of technology, being able to cooperate, see the bigger context, engage in discussion with researchers, knowing different procedures, consulting knowledge etc.

But naturally, such changes warrant some necessary adaptations from the librarian’s side, which seem to be already happening, albeit slowly. To Koltay (2017), librarians will now need skills involving database design, content management, data mining, programming, among others, which are fundamental for this support role for researchers. In this context, new names for information specialists have been arising, such as data consultant, data librarian, data curator, data officer and even data librarianship. Accordingly, job advertisements for this profession have been demanding knowledge of data management and curation.

In this seemingly fruitful scenario for libraries, contrasting remarks are observable. For example, at Purdue University West Lafayette, in the USA, a program that intended to consolidate library services for students (promoting awareness, making workshops and other activities) has managed to grow its scope and show positive results, but on the other hand, levels of data literacy of students, even those in the same context (such as same course and degree), varies a lot, which requires considerable adaptation in their workshop to adequate the instruction provided to the public. Moreover, in other research, it was previously shown that, while students value instruction regarding the library, attendance to workshops on the topic can be lower than expected (Johnson & Zwicky, 2017). This field as well still has many coming challenges to find which ways it can better leverage the growth of data.

3.4 General Public

Another relevant matter is that not only some selected professional areas may benefit from data literacy, but rather, every person can. The underlying reason for this is that our globalized world generates the presence of data in people's everyday life, data which can be transformed into useful information when used correctly.

An example of an experiment done with the public was the Museum of Random memory (MoRM), in Denmark. The experiment realized by Markham (2020) had the goal of stimulating the curiosity of people regarding the constant production of data produced by people daily, making them think about all the pictures taken by phones, interactions in social media, online and card purchases, search patterns on the internet, amongst other aspects. MoRM aimed to generate curiosity and questions on the mind of people about how all of these processes work, what they mean, which implications they have etc. The experiment occurred in several editions, in the form of an exposition involving interactive technology, data analysis and art. In some cases, objects of little value were collected from people willing to discard them, and then placed into exhibitions. In other situations, people could "donate" a digital photo and state to which degree they would like to remember or forget the photo through an interactive panel which diminished the image's opacity the more the person opted to forget it.

An important point of the experiment was the presence of professionals along with the expositions, who were oriented to draw the attention of people to each exposition by talking about what they were showcasing or with appealing questions such as how companies are managing our images on the internet or questions related to how our internet activities are tracked. The visitors have displayed varying levels of interest, but in general, showed some curiosity about it.

As Mallavarapu et al. (2019) affirm, museums are informal learning environments, which usually aim to provide a balance between enjoyment and learning as a means to optimize the visitor's experience. The adoption of immersive open-ended exhibits and promotion of ludic engagement can be valuable tools to draw people to it. However, according to them, it has been found that often visitors cannot comprehend ideas that go beyond the simple concepts involving data. In line with this, in their research involving a museum which provided an interactive display of biomes in which visitors could manage resources such as plants and water usage, one group that did not receive feedback based on the data of their interaction could not come to more sophisticated conclusions as opposed to the group that received it.

Therefore, mere curiosity is not enough to provide a better understanding of how data works. Hautea et al. (2017) emphasize how the youth, which is often using online systems for diverse purposes, are having their activities registered and analysed in ways not previously possible. While this allows a better understanding of how they interact with the web, it becomes important for them to be aware of what happens in each activity they do.

Another example of how technology and available data can be used by potentially any person includes the case of humanitarian mapping. According to Quill (2018), it has become increasingly easy to use geospatial resources and use spatial data and develop data literacy skills. This has made it easier for open maps communities to develop and it also led to the further use of “mapathons”, which consists on intensive sessions aiming to use geospatial data to develop maps with several purposes, some of which have often aided in disaster responses, as in the Nepal earthquake in 2015 and Hurricane Maria in 2017.

Similarly, Wolff et al. (2015) mention their attempts to leverage big data in smart cities as they try in the city of Milton Keynes, in the United Kingdom. They have been developing “Urban data games” such as “Appathons” (design or produce apps to address an urban challenge) and eco-puzzles (puzzles in which users must gather and analyse the given urban data to identify a recent disaster). However, among the challenges faced, they report that citizens do not have enough data literacy, there being a recent survey which stated that 4 in 5 adults in the UK have low levels of numeracy (the ability to reason numerical concepts). One potential reason for this they cite is a disparity in the curricular math taught and schools and reality.

In a practical research, Hautea et al. (2017) analysed interactions from young users at the Scratch online community, which is a platform designed for people aged 8-16 to teach programming visually and interactively. Some projects on the site allowed users of the platform to see the title and statistics of their first projects on the platform. Some users were not aware of how the platform retained information for so long and therefore commented they found it both interesting and “scary”. In other cases, as users explored the increasing features of the platform they have expressed concerns about how invasive it can be considering the number of activities stored in the system.

With these general ideas in mind, I develop in the next subtopics a few of the major data-related issues for society as a whole. I also use in the first two topics the tool "Google Trends[2]" to emphasize the relevance of some of the topics at the present times. Where it says "Note" on these charts, it is a note from Google that at that current points that an improvement on their data collection system was implemented. In all cases, the range was made from the earliest point possible (January 1st, 2004) to the day of the consultation (June 6th, 2020).

3.4.1 the growing spread of fake news

A relevant phenomenon which any person is subjected to is the growing amount of fake news being spread, something which was also facilitated with the access to technology and popularization of social media and smartphones. As Shreiner (2018) mentions, a simple search for the term on Google yields over 33 million hits and studies from the University of Stanford point points that even digital natives can often have trouble distinguishing adverts and news or spotting biases and motivations behind information found on social media. This phenomenon has become a major concern since many are the examples that highlight a notable inability from most people to tell fake news apart from real news. The author also mentions that a notable case is the 2016 US presidential campaign.

When it comes to Google trends, the charts below provide relevant insights into how often the terms were searched in the United States and worldwide.

[2] This tool developed by Google enables to user to see how often certain terms were searched across time and space on Google It displays charts to help visualizing the trends for the search queries the user may want to know about. It can be consulted here: https://trends.google.com/trends/.

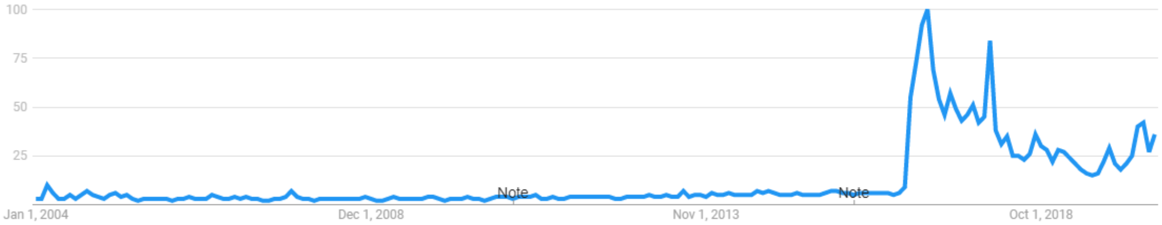


Figure 3: Trend of the term “Fake News” on Google in the United States over time[3]

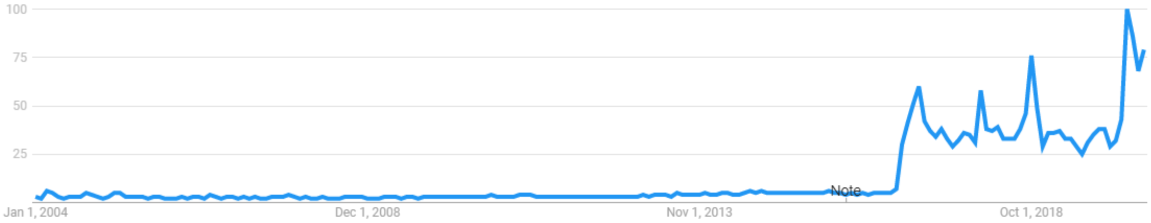


Figure 4: Trend of the term “Fake News” on Google worldwide over time[4]

In both cases, the term was not relevant on Google until it started to rise in October 2016 and peaking in February 2017, a period after which it has become relatively present on Google. Worldwide, a notable spike started on January and peaked in March of 2020. When analysing these, we should bear in mind that late 2016 corresponds to the last presidential elections in the United States (which occurred in early November of that year) and that the beginning of 2020 corresponds to the surge of the COVID-19 pandemic, two very relevant dates in modern history which spurred considerable online discussion.

Both theoretical literature and practical researches point to the addition of fake news to society’s vocabulary, and as such, data literacy plays a role here enabling people to cope with it.

3.4.2 privacy concerns and data safety

Discussions regarding personal data and privacy often permeate the topic of data. In some of the most impactful cases, we have Facebook’s unauthorized disclosure and

[3] The result and additional informations obtained can be reproduced following this link: https://trends.google.com/trends/explore?date=all&geo=US&q=fake%20news&hl=en

[4] These results can be found here: https://trends.google.com/trends/explore?date=all&q=fake%20news&hl=en

analysis of data as an example (Grillenberger & Romeike, 2018). When it comes to studying teenagers, Chi, Jeng, Acker, and Bowler (2018) show that they are growing in a world of technology and generating a lot of data, including personal information in social media. However, many of them do not understand underlying privacy issues and how they generate data.

On Google trends, some of the key terms actually show a decrease in this term despite it retaining some relevance, a possible sign that it is a topic of concern since earlier times of the internet. Two key terms that showed an increase were related to Google and to social media, which can be seen below:

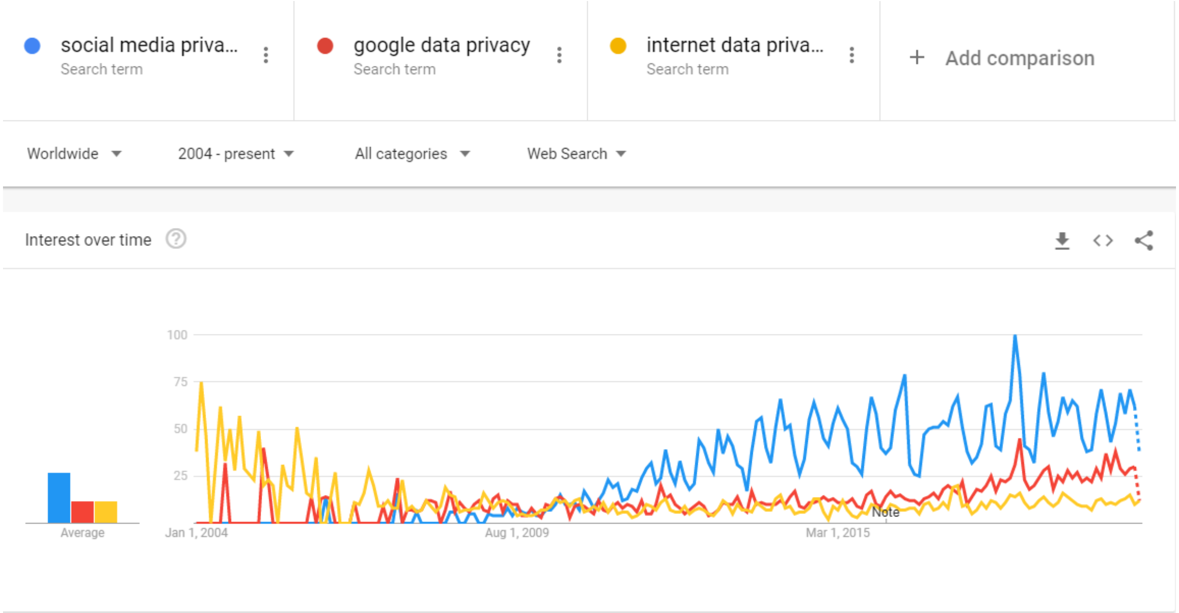
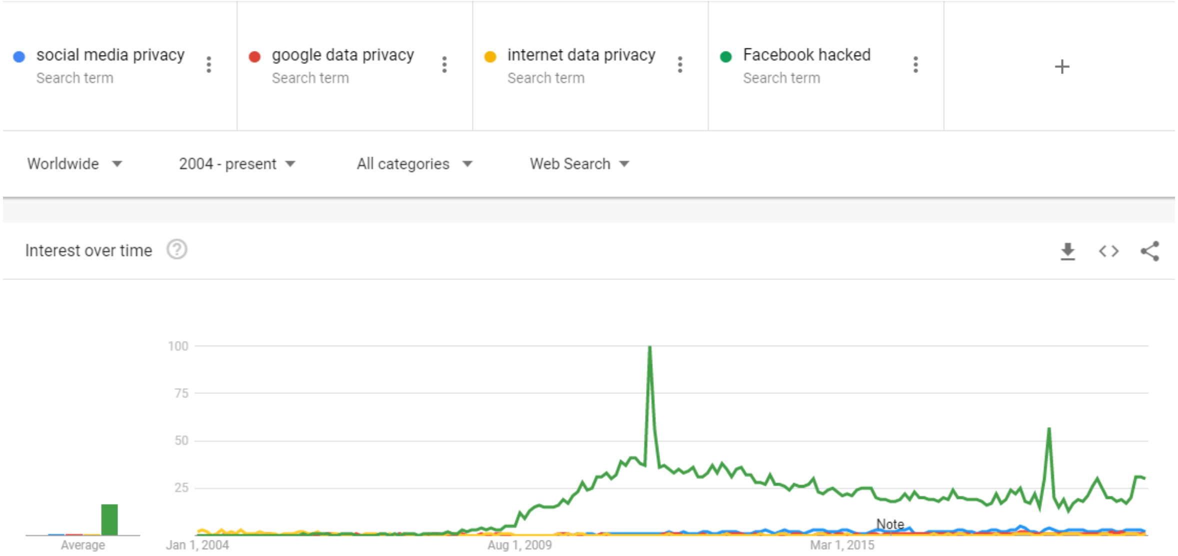


Figure 5: Trend of the privacy-related terms on Google worldwide over time[5]

The increased presence of Google and social media on everyday life seems to be backed by the fact that privacy within these two tools has been searched on Google in a trending manner, despite the volatility of the phenomenon. However, despite it not being the main goal of this work, it could be estimated that other combinations of search terms would yield more results to explore. As a test, the image below shows that more practical searches such as “Facebook hacked” easily make the same previous results less relevant:

[5] The result can be reproduced here: https://trends.google.com/trends/explore?date=all&q=social%20media%20privacy,google%20data%20privacy,internet%20data%20privacy&hl=en



● social media privacy
Search term

● google data privacy
Search term

● internet data privacy
Search term

● Facebook hacked
Search term

Worldwide ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Interest over time

Figure 6: Trend of some modern privacy-related terms on Google worldwide over time

3.4.3 inequality and the power of data

Another data-related issue our society faces is addressed by D'Ignazio (2017). According to her, there is a large disparity between those who possess the ability and means to collect, store and analyse data (usually States and corporations) and those who do not possess those, which applies to the vast majority of people. A few select specialists can harness data effectively, while the rest are more likely to be the subjects of data studies rather than using it for their own ends. She believes that rather than only teaching technical skills, such as reading charts, individuals have to learn how to use that chart to make the world a fairer place, that is, to connect technical skills with the broader scope of citizenship.

Similarly, Shreiner (2018) argues that data is used often to persuade people in political matters or to promote consumption, and while people often believe that claims backed by data are more persuasive, there is a low number of people who can indeed comprehend them, another example of the gap between those who make data and those who are targeted by it. The literature shows that civic empowerment has also been researched as a focal point of the applicability of data-skills for communities, as there are many ways in which curiosity and knowledge can be spurred in this field (Wolff et al., 2019).

While there may be other topics regarding data literacy, the examples provided so far suffice to promote awareness of the disparity between the current progress of data literacy and individuals' knowledge on data-related topics. We have also determined so far that data literacy comprises, in summary, one's capacity to use data from the starting point of defining a problem or asking a question, then proceeding to collect data, evaluate the data, analyse it, interpret it, draw conclusions and finally obtain actionable information that will be used to improve a situation. It is a cyclical process that needs re-evaluation along the way so that it may be adjusted and reapplied as deemed fit. This generic definition cannot easily be broken down into tangible steps and components without diving deeper into each case. Uses of data will vary according to the domain where it is being applied, that is, according to the particular conditions involved in the case, what goals are there, which resources are available etc.

Notwithstanding, there is still much to be known about the current scenario of data literacy, which raises the question of which societal structures are currently built in a way that promotes the development of data literacy? That is to say, in what way particular paths of education and work can affect an individual's data literacy? Are schools and workplaces lacking this development as it has been argued? We have observed some segments of society that benefit from data knowledge while others are lacking it, but even in contexts such as the business environment we still see considerable room for improvement. Therefore, with these questions in mind and considering the current practices in our society, such as the way education is structured, trending professions, demographics and other factors, I aim to develop research to uncover some information about individuals' relations with data, while also promoting some data literacy to potential participants by informing them.

4. Understanding and Researching a Data Literacy Framework

So far, I have identified different proposals of what data literacy means and which factors can influence it. Although they present a core of common elements, some authors have cited and elaborated on items which others do not mention, largely due to the different approaches and focuses each author had bearing their own specific context in mind.

Nevertheless, there are attempts to provide generic versions of a framework that could potentially be applied to any person. Despite all the extant research, authors such as Pothier and Condon (2019) mention how little surveys and reports are addressing the lack of data skills in today's professionals, which can be further confirmed by how this research found few articles regarding it through the systematic literature review. Moreover, I have established that many people are lacking data literacy, a skill which any person can benefit from. But companies are becoming growingly data-driven and individuals are adopting the use of technology often without understanding the underlying technologies and its implications, such as data retention and privacy issues. In an era which is often characterized by the reduction of asymmetry between people and companies, this appears to be an outlying factor.

Thus, with this research, we aim to understand which kind of people tend to develop higher data literacy, what contributes to it, how and why this development takes place, as well as promote the relevance of this skill to people who will take part on the survey.

4.1 Methodology of the Systematic Literature Review

The present research started with a general study of main data topics that concerned today's society and that could address the author's perceived need for a better understanding of data before arriving at data literacy as its driving topic. After defining it as the main subject of study that would fill this purpose of addressing the lack of data-skills in society, the English language was chosen for the work as it could have a higher potential to reach more people. It was then decided that the first step would be to start a systematic literature review to understand which matters are mainly being tackled by scholars.

Following that, the months of December of 2019 and January and February of 2020 were dedicated to collecting the potential bibliography to be used for the systematic literature review, a process which included the cleaning and managing of the obtained data and the evaluation of the relevance of the obtained materials. Thus, initially, it was discussed in which academic platforms the research would take place so that a reasonable sample of bibliography could be acquired for this analysis.

Then, a test was made by searching for the term “data literacy” in the academic platforms AISeL, JSTOR, EBSCO, Web of Science, Science direct, IEEEXPLORE and Google Scholar. We also tested the combination ‘“data literacy” AND survey’ and ‘“data literacy” AND review’, also applying, when present, the operator feature in the platforms for the terms “AND”. With this result, it was verified, as expected, that Google Scholar would be difficult to filter considering a large number of results, while the remaining platforms have shown results varying between some dozens e some hundreds. Based on a brief analysis of the materials, we decided to also include the terms “ICT literacy” and “Digital literacy” as keywords that could potentially retrieve related results.

Based on the data obtained, it was decided to try searching for the 3 types of literacy in variable combinations with the terms “survey” and “review”, alongside the operators “AND”, “NEAR” (this one only in the cases where it existed in the platform) and “OR” (for example, ‘"data literacy" OR "ICT literacy" OR "Digital Literacy"’, which would eliminate duplicates that could appear when searching 3 times for each of the terms). Some filters were also applied in the research, namely time filters (for results only between 2010 and the present date, which was the 20th of January 2020), language (to limit the results to English texts), peer-reviewed only (in the cases this filtering option existed) and also filters that would restrict the keywords to appear only in the title, abstract and keywords. During this stage, it was observed a large number of results in the searches involving the operator “NEAR” and the term “Digital Literacy”. These numbers can be observed in Annexe 1 of the present work.

Following that, a new search was made with the aforementioned filters, but this time limiting only to “Data literacy” and “ICT literacy”, once again using the combinations with the words “survey” and “review”, while also testing in how many cases these words would appear in the title of the materials found, of which the detailed results are in Annexe 1. The totals table from this research resulted in the following data:

Table 2: Results retrieved from the systematic literature review

FILTER	**"DATA LITERACY"** OR **"ICT LITERACY"**	**"DATA LITERACY"** OR **"ICT LITERACY"** AND **SURVEY**	**"DATA LITERACY"** OR **"ICT LITERACY"** AND **REVIEW**
ABSTRACT ONLY	1365	621	643
ABSTRACT ONLY WITH SURVEY/REVIEW IN THE TITLE	N/A	242	248

At the same time, the collection of the bibliography was done, which is analysed in the next step. For this, in every website used, the terms '"data literacy" OR "ICT literacy"' were employed and then the "Mendeley Web Importer" extension for the Google Chrome browser was utilized, an extension which scans the age and adds all found references to the reference manager "Mendeley. Each site was imported to a different folder in Mendeley. This procedure was executed only in the platforms Web of Science, Science Direct and EBSCO Discovery, since the results obtained here were of a higher volume.

Considering the importing process did not occur perfectly in all websites, manual verification of all files not retrieved was made so that they could be manually imported by using the tools and options presented by the websites themselves. Afterwards, those files were manually added to Mendeley. Once in Mendeley, its tool that eliminates duplicates was used in all cases the software would identify them. In a few cases, despite the filters applied, the site would expressly indicate that an article was not peer-reviewed, in which case it was eliminated.

Afterwards, Mendeley was used to export all references in a .bib format in each of the 3 folders. Then, the software "jabref" was used to convert all 3 .bib files into .csv format. From here on, the software "Microsoft Excel" was applied in order to open all 3 files and generate tables with the detailing of the references. In the cases of the websites AIESel, JSTOR and IEEEXPLORE, since the results were few, those were manually inserted into the file. A final table compiling all results was thus created. In a few cases, there were still duplicates and abnormal entries which Mendeley Web Importer generated, but all of these were manually eliminated in Excel.

Based on the numbers obtained, for the next part, it was decided to analyse the bibliography collected based on the search “data literacy” OR “ICT literacy”. The usage of both terms was to avoid being too restrictive initially. However, the number of results for the search of data literacy was considered enough and many results for ICT literacy were not directly related with the object of this research, so it was decided to start with only the results of the former, postponing the latter for the case the amount of materials did not suffice, which was not needed in the end.

Finally, I arrived at a total of 138 materials. Each material was categorized into 3 different groups: 1) Main, which refers to the most important articles, defined by having data literacy at the core of its topic and dedicating substantial content to it; 2) Secondary, a group with materials that either had data literacy as a secondary topic (with another data-related topic often being the main topic) or that, despite focusing on data literacy, the size of the article resulted in a shallower contribution; 3) Incidental, one last category for materials that had minimal or no content related to data literacy. Some articles, despite mentioning data literacy with some frequency, were still categorized as either secondary or incidental if they did not go deeper into explaining the concept but rather limited themselves to mentioning it loosely.

Additionally, I grouped some of the most often discussed topics to enable an overall view of which subjects are most discussed when it comes to data literacy, the results of which are argued in the next topic of this work. All of these categorisations were gradually reviewed as more time was dedicated to each material and the final results can be seen in the files presented on Annexe 1. It is to be acknowledged that these categories I created and how each material was distributed in it lacks a deep methodological rigour and may be considered somewhat arbitrary, however, the goal of doing so is merely to identify insights from where the theory studied here can further develop. The categories are by no means objectively right and strictly defined and nor are they exhaustive, but rather they aim at providing starting points from where we can explore the extant literature.

4.2 The Variables Involved in Data Literacy

Having collected the base material to explore the literature on data literacy, I now aimed to start designing a data literacy framework based on the aspects the literature found most relevant. But in other to begin that, I would first need to understand which topics predominate in this concept, be it related to what data literacy is, what skills are involved, what can affect data literacy, as well as any point found to be relevant.

After scouring the literature, I found 5 major groups that were subdivided into a total of 25 main topics related to data literacy. It is important to highlight that my choice for deciding to group topics the way I did does not aim to be a recommendation of a model nor an exhaustive list, but rather, they represent topics I considered that have been given considerable relevance by the literature. While this process is subjective, it aims to solely provide guidelines for the next steps while exploring the considerations made by several authors, not being a part of the framework.

Some of the major groups were inspired by the existing literature. Environmental factors are not often mentioned in the literature, but valid arguments were made in the case they are present and no objections were found so far. For this reason, the part regarding environmental factors is directly based on the points made by Wang et al. (2019). Attitude is sometimes directly or sometimes indirectly addressed, but only one article found, made by Chi et al. (2018), had this topic at its core.

In addition to that, the central elements that constitute the very concept of data literacy were inspired by the model proposed by Mandinach and Gummer (2016) and all of the other parts are also inspired on them but with a few differences. Their version of the framework is the result of a gathering with several scholars who finished a list of 59 different skills, many of which overlap each other. Some of these skills I grouped into a single element but there is room for alternative interpretation, not to mention that very often some kind of skill can easily fit several different concepts. One example is when students learn to take into account purchasing power parity when analysing the comparison between exchange rates of currencies, considering local prices need to be taken into account in the comparison of countries (Halliday, 2019). This is a case which revolves around, without excluding other possible interpretations, the transformation of data into information, contextual application and critical reasoning. The previously mentioned Science and Engineering Practices of the NGSS are also used to support the items presented here. Regardless, I reemphasize that this first moment aims to shed a light on individualized aspects of data literacy, but only the framework later proposed represents the model I propose for this research.

Besides, because the components are not strictly separated from each other, but are rather a set of intertwined capabilities, several examples illustrated often reflect varied components simultaneously, and these are not repeated in each section to avoid unnecessary repetition and also bearing in mind the idea that any division established here has the main purpose of illustrating specific topics within the domain of data literacy, but in practice, they are connected in ways that often make them hard to distinguish.

After reviewing all 138 articles, I arrived at the following conclusion on how they were presented in each article, which is shown in the annexes on Table 8 (absolute frequency per category) and Table 9 (frequency in the most important materials found, relative frequency and tendency of growth). These tables show a higher concentration of the first variables presented and a lower one and the end of the tables. Nonetheless, the variables that comprise attitude and environmental factors (lower end) appeared with a reduced frequency not because the literature argues in a way that excludes them from the concept of data literacy, but rather, they are not mentioned in most articles. Or in the case of attitude, this word by itself can appear somewhat often in articles but no further clarification is given on the concept and its role. On the other hand, in the cases where these variables appeared, they were developed in a way to validate their relevance, which is why I believe it is important to emphasize these points here and present them as relevant elements to the framework I will later propose.

4.2.1 components of data literacy

The first part has to do directly with what authors say that data literacy is and what it involves. As mentioned, it was mainly based on the propositions of Mandinach and Gummer (2016) and I chose to present it here in parts, from the beginning of the data cycle to its end. Additionally, since these concepts were previously introduced and explored, I will not dedicate lengthier explanations in this section, but rather focus on highlighting each component now isolated as a proper topic to facilitate visualization of what is involved in it.

4.2.1.1 identification of problems and framing questions

According to Kippers et al. (2018), this initial step aims to set a purpose for the use of data and involves defining problems and formulating hypotheses or questions. It has to do with setting the starting point for the use of data, namely identifying a problem and framing a question, which can potentially be solved with the use of data. It can comprehend the articulation of a problem of practice about a topic, recurring issues in a specific field, understanding how certain contexts work and the many variables involved, involving several stakeholders for a broader perspective (Mandinach & Gummer, 2016). To Schildkamp (2019), it can also happen that questions and problems may arise from the collection of data, as the data will show elements that may not have been known before. It is also relevant to mention that "asking questions and defining problems" is the first of the science and engineering practices mentioned in the NGSS.

Wolff et al. (2019) emphasize that while data literacy needs technical skills, it is also essential to pursue a data-driven inquiry because students often fail to understand the general context of the application of data. As such, I also consider this skill involves both: 1) framing a question that can be later analysed with the use of data; 2) Coming across data and being able to frame a question based on it.

4.2.1.2 data collection and reading

One of the first components of data literacy and one the most frequently discussed in the academic texts studied was the ability to collect and read data. This is a task which involves having the components that will later be analysed, transformed into information and interpreted, and that requires knowing different data systems, and how to locate, navigate and access them. Moreover, instead of relying on ready data, which does not always exist, certain professionals have the conditions to create data using their working environment. Teachers for example, by designing adequate assessments and implementing them in the classroom can generate data for posterior study (Mandinach & Gummer, 2016).

Data collection is directly connected to the problem that was defined before and which the collection aims to solve. Also, students must understand the reason to collect data and how it will lead to solutions to the questions being addressed (Gibson & Mourad, 2018). Collecting data, both quantitative and qualitative leads to a better understanding of the scope of the problem and determination of goals and has to be directly connected to the purpose set beforehand. Therefore, the choice of how to choose to collect data derives from the type of problem defined previously and what hypotheses and questions were formulated (Kippers et al., 2018).

To Gibson and Mourad (2018), talking about data literacy specifically in the domain of life sciences, basic knowledge on the collection of data involves knowing how to use instruments and technology to collect data, while intermediary knowledge comprehends identifying appropriate data for a biological question and hypothesis and also inserting data into spreadsheets or databases. Finally, advanced knowledge can imply a more rigorous methodology when collecting or sampling and also understanding the process of storing, managing, manipulating or querying databases. These remarks also stress that tools are relevant in the context of data literacy and directly connected to practical applications.

The second practice of the NGSS is relevant here, which is "planning and carrying out investigations" and it involves the clarification of what can be counted as data and the identification of variables or parameters. As such, collecting, recording, reading, finding, accessing, retrieving, creating and mining data are common verbs associated with this type of component.

4.2.1.3 cleaning or evaluating data quality

An important skill for anyone working with data is being able to assess the quality of data, choosing what is useful and what can be discarded, which is often referred to as cleaning the data. That includes the elimination of data that does not make sense (a score of 110 when the maximum is 100) or misleading information, as well as filtering, organising, managing and storing into databases (Mandinach & Gummer, 2016). It also involves removing blank and duplicate rows, uniformizing formats, fixing discontinuities or any other activities which will improve the quality and usability of the dataset, all to improve its quality (Erwin, 2015).

D'Ignazio (2017) mentions that cleaning data can often take 80% of the time of those who work with data, not to mention that tidy data must meet certain standards. Therefore, in a larger organizational environment, specialized people such as data scientists or analysts would be the ones to use this task while other stakeholders would touch only the ready results since this process involves a certain level of technical knowledge. This separation of a technical (as I am referring to the components which involve tools, statistics and data processing) and non-technical sides (such as reading, interpreting and decision-making of data literacy will be important for my framework later on.

The concept of data curation is seen in different ways. To Kjelvik and Schultheis (2019), curation is related to handling data, considering the range of processes dedicated to tidying up data, the general process of cleaning and preparing data sets. On the other hand, to Koltay (2017), curation involves knowing the ownership of data, which data should be retained and how, in which ways risk should be managed, what are the costs involved, what are the options to manage data, how is it accessed and how open it should be, as well as involve activities such as the creation of curation policies, procedures and practices, select documents for long-term preservation, monitor obsolescence of files, software and hardware etc. At any rate, the concept varies more on the depth of knowledge but still belongs to this category.

4.2.1.4 data manipulation, analysis and interpretation

Once the data is ready for analysis, the processes of manipulation, analysis, processing, handling and interpreting can go into place. This is a core element for professionals dealing with data, as the raw data normally cannot be actioned, but rather it needs processing to later become information that leads to actionable insights. Thus, it is needed to know how to observe patterns, interpret results obtained, synthesize data, articulate inferences and conclusions, summarise, generate connections (Mandinach & Gummer, 2016).

In a similar sense, Koltay (2016), when referring to the processes involved in controlling data, talks about the term data governance, which according to him, it can be considered to be “the exercise of decision making and authority that comprises a system of decision rights and accountabilities that is based on agreed-upon models, which describe who can take what actions, when and under what circumstances, using what methods”. Once again drawing from the notion that information literacy is highly relevant to data literacy, the author proceeds to emphasize the importance of the quality of data and that, considering how important and how big the available data is, there is a need to properly manage it and ensure a set of rules will be applied.

Interpretation involves the identification of key takeaways from the data and understanding the meaning of the results after processing it (Pothier & Condon, 2019). When describing analysis and interpretation, Gibson and Mourad (2018) emphasize a lot the statistical part and analytical reasoning, which shows that this component can delve into a more technical side of knowledge. For them, basic literacy encompasses being able to describe patterns in data and describe data with statistics, while intermediate knowledge involves analysing and interpreting data with statistics as well as interpreting results of statistical tests regarding the question or hypothesis that originated it. Advanced knowledge involves the incorporation of data analysis and statistical methods into experimental design, understanding assumptions, and being capable of comparing results.

4.2.1.5 extract insights, communicate data and transform it into actionable information

Finally, at the end of the data cycle, the information obtained has to lead to a decision to improve on the situation. This entails assessing where action needs to take place, what needs to be done, determining next steps, making necessary adjustments (Mandinach & Gummer, 2016). Another relevant information is that the NGSS science and engineering practices include "constructing explanations and designing solutions", "engaging in argument from evidence" and the aforementioned "obtaining, evaluating, and communicating information".

It is worth noting that, because the process is cyclical, after this part is finished professionals are advised to retake the step regarding the identification of problems and framing questions in order to verify whether the problem has been solved, how the outcomes have turned out to be, monitor changes, consider new making new decisions, among other considerations (Mandinach & Gummer, 2016).

Once again drawing from the considerations presented by Gibson and Mourad (2018), communication can involve, from basic data literacy, through intermediate, to advanced: 1) using technology for the construction of tables and figures and describing graphical and tabular presentations of data; 2) explaining relationships in data and understanding the use of data and analyses to argue based on evidence; 3) Evaluating strengths and limitations in data and understanding relationships between data and other issues.

4.2.2 variables associated with the development of data skills

Asides from the concept itself, some authors delved into specificities that are part of data literacy, some of which were found across several different texts. These are variables which relate to the previously mentioned components. For example, knowledge of statistics aids the process of manipulating data and data visualization skills is part of the process of both understanding and communicating data. For this reason, I decided to track some of the variables which were mentioned to aid with the process of identification of the main components that influence data literacy.

4.2.2.1 general data understanding and analytical thinking

This first item has to do with the analytical reasoning required to understand data concepts, as well as a general knowledge of what data is and how it can be used. Kjelvik and Schultheis (2019) believe data literacy lies at the intersection of quantitative reasoning, data science and authentic context, an idea I will further explore later. For this topic, what matters the most is the notion that quantitative reasoning, that is, being able to apply mathematical principles to solve problems by using logic and critical thinking and understanding numerical information in diverse representations, is one way to represent this idea. Not to mention that “using mathematics and computational thinking” is also one of the practices of the NGSS.

One example that illustrates it is elaborated by Dichev and Dicheva (2017), in which they attempted to stimulate data science literacy in students of diverse backgrounds through a short course. The content of the course was tailored to focus on a minimal core from the many components of the subject, resulting in a list of four skills, namely: formulating productive questions, thinking computationally, thinking analytically and visualizing and reporting summary data.

4.2.2.2 statistical knowledge

Statistics are currently present everywhere in daily life, such as in social media, journals and news (Kirstin Fontichiaro & Oehrli, 2016). While expertise is not a requirement to become data literate, some knowledge of statistics can improve the results that can be obtained from data. For example, according to Mandinach and Gummer (2016), for teachers, usually simple statistics like central tendency and dispersion would suffice, while more advanced techniques like regression and ANOVA are not necessary. The authors also recommend understanding psychometrics, such as the concepts of reliability, validity and error of measurement. Some examples were presented on the topic regarding manipulation, analysis and interpretation of data since there is a strong connection between statistics and that part of data literacy.

In a general way, it is also relevant that people who are handling data can also comprehend some of the implications involved. As Halliday (2019) exemplifies, if a country has a certain per capita GDP, it does not mean that it is reasonable to assume each person has that amount available or even that the average (in this case median) consumer in that economy has that amount as disposable income. Such a measure does not show the distribution of income or its skewness.

4.2.2.3 tool knowledge

The growing amount of data produced in society has become impossible for humans to handle without technology, which was further proven by the rise of big data. If we focus, for example, on data science literacy, the computational aspect gains even more relevance, as the core competencies involve, according to Dichev and Dicheva (2017), computational methods, data collection, processing and modelling, statistics and visual communication.

Besides, “developing and using models” is the second science and engineering practice of the NGSS. Therefore, knowing how to use technology is a must for anyone who wants higher performance dealing with data. This comprehends understanding data warehouses, spreadsheets (such as computer programs Microsoft Excel or Google Sheets), apps, dashboards, or on more advanced levels statistical software (SPSS, SAS), Business intelligence and data visualization software (Tableau, PowerBI), Query and Programming languages (SQL, Python, R etc.).

Programs such as Excel and SPSS tend to be closer to most students and help to build their preparedness for college and careers (Erwin, 2015) and also, authors such as Gibson and Mourad (2018) believe that conducting mathematical calculations and using spreadsheets and software for this purpose are elements at the basic level of data literacy. Additionally, knowledge of spreadsheets has been claimed by some students to be an important hiring factor and one of the most useful skills learned at educational institutions and how it translates into relevance in the workplace (Slayter & Higgins, 2018). Tableau, a data visualization software, is one of the examples given by Pothier and Condon (2019) of tools which the managers whished new hires had more knowledge. The field of Business intelligence also leverages a range of tools to perform their activities (Ranjan, 2005).

Following the ideas of Robinson and Bawden (2017), since the use of databases is relevant, it is natural that an understanding of query languages such as SQL would be mentioned. They also emphasize the role of coding, for example, in library contexts, enabling professionals to modify records, enrich metadata, convert file formats and other activities. Moreover, authors such as Dichev and Dicheva (2017) and Teal et al. (2015) quote Python and R as programming languages of choice for data analysis and useful ways of getting more out of spreadsheets.

4.2.2.4 contextual applicability

Understanding the whole context can lead to a better understanding of how variables may be influenced by other aspects, knowing what the existing connections between different factors are, and thus, it enables better use of data. In the case of librarians, for example, Robinson and Bawden (2017) argue that technical skills do not suffice, as they must be complemented with the knowledge of social and ethical implications as well as the cultural and political environment.

Additionally, according to Kjelvik and Schultheis (2019), contextualized authentic data facilitates the learning process, as students find it more engaging and interesting and that it makes more sense. In a similar note, D’Ignazio (2017) asserts that choosing a dataset that is relevant to the community, a topic which they care about and have a direct interest would benefit the learning process, while working with data that is not relevant for the learner would be alienating. In the same way, Teal et al. (2015) apply their data literacy workshops for teachers under the light of the specific domain of the teacher, aiming to address both the contextual applicability and enhancement of the learning process. Therefore, data applied into context facilitates learning both in motivational terms and in the learning itself, by adding a layer of real-world application that further enables individuals to act on data.

Contextual applicability is developed so naturally in many domains that, throughout this work, several other examples outside this section will be presented that reinforce its importance.

4.2.2.5 critical reasoning

Critical reasoning was an aspect that not only appeared often in the systematic literature review, but it was given special importance by several authors, which is why it receives extra attention here. A data literate professional must be capable of selecting and synthesizing the correct data while evaluating when data is being mishandled and presented in misleading ways (Koltay, 2017). To Hautea et al. (2017), critically assessing data is central to the definition of data literacy. For these authors, following the vision of Paulo Freire, the word critical has to do with the perception of how things exist in the world as part of the learning process. Being able to question data is important because individuals often see datasets they come across as true, without questioning it. As an example, Google searches yield ready results, devoid of context, such as the reasons why it was collected and by who, its limitations, etc (D'Ignazio, 2017).

When we consider the boundaries that data literacy share with information literacy, it becomes a relevant indicator of the importance of critical reasoning, considering the latter involves the capacity to think critically about concepts, claims and arguments (Koltay, 2017). That is, the very concept of information literacy (once again, whose domain of study is shared with data literacy) revolves around critical thinking.

The research made by Hautea et al. (2017) in the Scratch community highlights some of the topics involved in critical data literacy. First, data collection and retention imply in dealing with privacy. People's data can be retained in the net for long periods, stored and subjected to analyses which the users of a service may very often not be aware of. Second, scepticism and interpretation need to be applied when dealing with data, as data can be easily misused. Third, data is presented with some underlying assumptions and hidden decisions, since there is always a decision behind the production of data, that is to say, data is not truly raw since its production is often done with a specific purpose, meaning we frequently see what was intended to see, which sometimes excludes data which was not selected, not to mention the cases in which vague criteria that justify some decisions are applied. Fourth, algorithms that are data-driven can cause exclusion in some cases, which was illustrated with the possibility of how in Scratch some users can be excluded from using some programs unless they meet some criteria or reedit the constraints (only possible because Scratch's code is open and is still a challenge for new users). Fifth and last, awareness (or lack thereof) can alter the very way in which a certain platform is used. In Scratch's case, the new features which added more capacity to access and use user's data raised concerns about a shift of focus from making creative creations and sharing to obtaining followers and meeting some popularity goals. All of these notions are connected to the idea that data is a creation and a product of specific selections and transformation processes which someone made with a certain purpose in mind (Sorapure, 2019).

The case of Scratch can be directly inserted in the concept of data infrastructure literacy by Gray et al. (2018). The key takeaway from this theory is that despite the term raw being often used to refer to data, the authors argue that actual raw data is inexistent, as the moment an individual examines the real world, selects a slice of it to be collected and go through the entire data cycle, there was a decision motivated by specific purposes, with certain goals and tools, with designed goals and constraints, among other things. There is a collection of historical, social, political and cultural contextualization behind each dataset, and such contexts are often not shown along with the dataset, otherwise, data would easily be questioned more often, such as in the case of biased actors promoting an idea.

Online platforms, such as social media, are built with a set of rules of what is possible to do, such as liking, commenting, sharing, following etc, which means that data generated there respect the rules of the platform. Therefore, all its existing data are subject to the rules that govern the platform, so individuals cannot produce data in total freedom but only according to what the platform enables, and this must be taken into account when questioning data. The authors exemplify this with the case of Facebook, in which until 2015 an individual had more limited options to interact with posts, but then the "like" button developed to include five different expressions of emotions towards the content.

As we can see, several authors include some sort of critical reasoning component when referring to data literacy, and some take a step further and place this skill at the heart of the concept. For these reasons, I believe critical reasoning plays a key role in data literacy, which is why it was given special attention in this work for a proposal of a framework for data literacy.

4.2.2.6 data visualization

The ability to read graphs and charts correctly constitutes the set of abilities of a data literate professional (Koltay, 2017). One of the many components involved in business intelligence has to do with creating visualizations that will facilitate communicating data and making it easier to understand. Comprehending and creating mapped data, graphs, pie charts and various manners of data visualization is a great aid to students (Kirstin Fontichiaro & Oehrli, 2016). Not to mention that data has become multimodal, being presented not only in text form but often in visual images, graphs, paintings and a range of options (Shreiner, 2018).

There are dedicated studies just for the field of visualization of data. Taking a step further, Sorapure (2019) mentions information visualization, also known as InfoVis, as a field that is expanding into everyday life. While information graphics normally display static data to convey it in a more appealing way to communicate it, InfoVis goes further by providing dynamic representations of data. This is often done with the use of computers that can provide interactive displays of data in a visual manner coupled with options such as zooming, filtering and searching, as well as the use of text, creating a combination that enables a more powerful way to extract insights from data with a clear goal of facilitating communication of data. Projects based on this concept such as Dollar Street[6], rely on the notion that, in general, numbers and statistics turn people down, so lively representations of data can often attract the public with more ease.

However, the author argues that designers of such visualizations often have to make choices about how the visualization is going to be made and what kind of interactions are enabled. This further connects with the notion of data infrastructure literacy, as InfoVis is the product of a specific set of rules its designer intended to make. The author discovered that in Dollar Street, for example, there is a document which explains how some of the

[6] This website aims to display visual representations (with pictures) of some aspects of the lives of families across the world, such as their respective incomes, possessions, dreams etc. The goal of the site is to show that people from different cultures have a lot more in common with us than expected. The project can be consulted here: https://www.gapminder.org/dollar-street/

calculations were made, and since there were several complications in bringing different variables of earnings and cost from different places around the world to a common dollar base, several numbers the platform provides had to be "guesstimated", as the document says. This also happens to variations in the level of services provided by the government, such as healthcare, considerations of payment made in goods rather than money and other factors. In the end, the visualization is not entirely accurate but not all visitors would chase this document before taking the data as true.

The lack of development of data skills is perceivable in some contexts. Shreiner (2018) has shown that students often have trouble drawing conclusions and insights from data visualizations in their textbooks and also that they ignore them often under the assumption that no additional knowledge to what was already in the text would be provided. In another study, Shreiner (2019) elucidates how problematic this phenomenon is considering that data use is becoming increasingly present in a range of domains, incorporating the social sciences and humanities at a growing pace. In the fields of history, visual displays of data are used, for example, to show historical occupations in maps and development of relations, as well as a range of other valuable information which students often not only fail to understand but also to critically assess it.

4.2.2.7 data processing techniques

Very often it was noticed that the ability and knowledge on how to process data were mentioned, sometimes related or not to tools or statistics, knowing how to aggregate or disaggregate data, drill-down and others (Mandinach & Gummer, 2016).

Another example is brought by Gibson and Mourad (2018), to whom an intermediate data literacy involves understanding how to apply mathematical tools and technology to conduct calculations and knowing the relationships between them and a biological question (biological is mentioned since it is their specific domain of study, making it also an example of contextual application). For them, advanced data literacy would go a step further and include knowing how to choose proper tools for the desired studies and understanding how data is used in developing quantitative biological models. These examples also outline the importance of tools, analytical thinking and statistics, once again emphasizing how intertwined each component is. All in all, this item is highly connected to the process of manipulation and analysis of data, as well as statistical knowledge.

4.2.2.8 data authenticity

Though not discussed as often as some of the other items in this list, data authenticity was discussed across several sources. Authentic data influence data quality as it can represent properly the phenomenon studied, guaranteeing accuracy, validity and reliability (Koltay, 2017). That means the ability to find and select authentic data counts towards being a skilled data professional

Just like in the case of contextualized data, Kjelvik and Schultheis (2019) also claim that authentic data aids students' learning, as it encourages connections between data and the real world. To Roy Gould, Sunbury, and Dussault (2014), in the real world, data is often messy and not perfectly outlined in charts and tables in which students often deal with. For this reason, the authors believe that dealing with messy data enables students to think deeper about data and learn more, while being exposed to the notion that there is often no right answer but rather results from which conclusions and further studies can be derived from. D'Ignazio (2017) agrees with this reasoning by emphasizing that data in the real world is often messy and this helps learners to use critical thinking skills which are needed to handle data.

This corroborates the ideas from Erwin (2015), who states that students can develop data literacy by participating in tasks which involve authentic data analysis in a context of project-based learning. Authentic learning tasks are focused on solutions to related to real cases, involving cases, role-playing, problem-based activities and other means, having more value to a student than grades. He bases his research on previous experiments with separate groups in which it was concluded that, in those situations, when comparing students taught via data-centric methods as opposed to a control group of students, the former using real-world data sets showed significant learning advantage. There also reports of increased motivation and commitment when a task is perceived to be authentic.

4.2.2.9 data sources

Data can assume many shapes and sizes. Typically for the study a phenomenon there are multiple sources which professionals must know which one to turn to according to the case faced. Directly related to the collection of data, this would involve knowing how the difference of utility and use of each data, being able to use multiple sources, knowing the existence of multiple sources and how to access them (Mandinach & Gummer, 2016). Drawing from the concept of information literacy, Womack (2015) mentions one of the competencies involved is to evaluate information and its sources critically so that selected information can be incorporated into knowledge. This item appears to be more cited in the context of academic research.

4.2.3 attitude

One key factor that influences the development of data literacy is the individual's attitude towards it. Mandinach and Gummer (2016), for example, mention "dispositions, habits of mind, or factors that influence data use" as an extra component which, while not part of their framework, influences teaching in general and would thus impact the usage of data for classrooms. Kjelvik and Schultheis (2019) argue that quantitative reasoning, one of the components of data literacy, also involves an emotional response regarding the learner's attitudes, interest and beliefs.

In the educational context, Ebbeler et al. (2017) mention that a positive attitude towards data, which includes a belief that it can be used to improve problem-solving, is necessary for an effective data-use in schools. For them, the very development of data literacy relies not only on the development of the skills required but also on the attitude towards data. These notions are directly connected with some of the external factors we present further ahead.

Some degree of relevance was given to attitude in works such as these, but most did not seek to explore the topic further. However, one article found in the systematic literature review had at its core the topic of attitude towards data, providing better guidelines to address what it means. According to Chi, Jeng, Acker, and Bowler (2018), the ABC model of attitude, in the domain of psychology, divides it into three components, which are: affect, behaviour and cognition. These deal with how people feel, think and interact with a certain subject. Despite being different dimensions, they are connected. This model and the examples provided in their research, that consisted of interviews with 22 teenagers, will be the starting point for the development of this section.

4.2.3.1 affective state

The first item, affect, has to do with the feelings and emotions of an individual. In the research made by Chi et al. (2018), 42 quotes by teenagers were made in total, most of which were positive (20 quotes), citing often confidence, interest and curiosity towards data. Negative states appeared 12 times, mentioning anger, sadness and fear, usually in the context that there was not a lot of control on how data is created online and how it is diffused. Neutral states appeared 10 times, such as indifference or a mix of negativity and positivity.

Particularly present across my literature research was the notion of interest, as several cases dealt with people who are interested in data or, in some cases, interest and curiosity were developed after a given stimulus. This can be exemplified with the introductory course to data science literacy proposed by Dichev and Dicheva (2017), which had students from several different fields. All students were tested at the beginning and end of the course regarding their knowledge and attitude towards the topic. At first, in the pre-test, the students in general already showed a positive attitude regarding the worth of data science and how it is present in everyday life, a result which, in general, increased during the post-test. In the case of the Museum of Random Memory, strategically placed people and the questions they posed have caught the attention and picked the interest and curiosity of the nearby people.

4.2.3.2 cognitive state

To Chi et al. (2018), cognitive states address the knowledge, beliefs and thought processes. A recurring example of this topic that I found on the literature was about data interventions in educational establishments, generally done with data professionals promoting awareness of the many ways data is present and can be used on their activities and instructing teachers about how to use it. In the case of Dunlap and Piro (2016), they mention how educators are expected to use data and determine which actions can be taken to improve, not simply guessing what is needed, but rather understanding students. However, before the intervention, participant educators were showed unawareness of what data was, how they could collect it in the classroom, why it was relevant, which uses it had (such as prediction). After interventions participants claimed that data is a valuable resource, containing useful insights that could lead to adaptations in their practices. Other intervention cases such as Ebbeler et al. (2017) report gains in data literacy skills and attitude.

4.2.3.3 behavioural state

Finally, Chi et al. (2018) describe the behavioural aspect of attitude regards the practices or decisions associated with individuals' affective and cognitive states. Their research has shown that since a particular concern of teenagers was directed towards privacy, less confident and more fearful teens would go longer length in adopting safety measures in social media and personal devices usage.

Another illustration of how stimuli can benefit data literacy can be drawn from the experience made by Piro et al. (2014). In their case, an instructional intervention aimed at developing competencies in the understanding, interpretation and use of data led to a reported significant increase in the attitude of the participants towards data. Whereas before the intervention they reported high levels of discomfort and lack of confidence, after it the vast majority of participants claimed to now be in the opposite situation, with considerable confidence. As established before, confidence is relevant for consistent and proper usage of data in the work, but this work adds to that saying that working with data also improved self-efficacy, findings which the US Department of Education also corroborates. These gains after data interventions are connected to all 3 dimensions, since they are not strictly separated components, but are rather connected. Chi et al. (2018), for example, derive their study of behaviour as a consequence of individuals' affective and cognitive states.

4.2.4 environmental factors

The elements presented here are based on the work of Wang et al. (2019), who present a list of 8 components that can affect data literacy for safety professionals. These items were also found in other articles, albeit sometimes indirectly, and I decided to label them as environmental factors since they are all related to external factors which can affect the individual. However, I did not adopt this model in its entirety. For example, they include a separate component called SRD teams (teams comprised of safety and data professionals) and SRD coaches as key factors for the efficient use of data in organizations, as this way, some specialized people can conduct the practices involved in this field. I decided to not include a separate field for this as it can overlap with several of the items below, but it is agreeable that data teams and data coaches can be a valuable resource for organizations.

4.2.4.1 data culture

This item is related to the culture adopted inside the company related to data. A company that has a culture strongly related to data, promoting beliefs, attitudes and values related to data collection and its use in the organization's management will positively affect the worker's data literacy (Wang et al., 2019). Lack of collaboration and a data culture has also been associated with inabilities to plan and act on data (Wilson, 2016) and in the case of developing countries, hindering the development of data literacy (Schildkamp, Poortman, & Sahlberg, 2019).

The elements of attitude previously shown apply directly in this case, with the difference that here the focus lies not on the individual, but on the organisation, the context in which the individual is placed.

4.2.4.2 business culture

Wang et al. (2019) speaking about the field of safety, talk about safety culture as another influent factor. Also related to organizational culture, this is about the values, attitudes, perceptions and competencies that revolve around safety. If the company has a strong culture related to their work, data literacy for safety professionals becomes easier. Following the authors' ideas, I assume that data literacy for several professions is not an extra but rather a core element that has been gaining increased relevance due to the alterations in global technology. Considering I also found other examples of how strong work culture can aid the data literacy aspect, I decided to use the general “business culture” term to encompass these situations. As such, it is to be understood that an organization which has a culture that aims to understand well its work and present a good performance, it is a natural consequence that data be adopted across some sectors and a primary requirement for improved performance.

4.2.4.3 data leadership

In the case of Wang et al. (2019) about the field of safety, they mention that for an organization, is it important that those in safety roles make a positive influence to increase the use of data in the workplace, and in the highest management level is where the strategic use of data is defined. Schildkamp (2019) Also emphasizes the role of leaders, as they are responsible for determining goals and balancing the interests of stakeholders, the organisational culture and the mission, vision and values of the organisation while making sure goals are fulfilled. In a similar argument, LaPointe-McEwan, DeLuca, and Klinger (2017) point to an experience in Canada with middle leaders in schools (facilitators who usually used to be class teachers) have a positive effect on promoting data use among teachers. Gummer (2013) has reported that it has been widely mentioned in the literature how in a school context the school leaders have an essential role in the adoption of data by teachers. Therefore, in an organisational context, we cannot ignore the influence leaders have in the adoption of data use.

4.2.4.4 vision/awareness

In the 8 items presented by Wang et al. (2019) we have awareness of DLSP and a vision for SRD use in safety management. For the authors, awareness determines behaviours, and being aware of DLSP affects the way professionals seek to improve their performance by directing some attention to the study of this topic. Congruently with this line of thought, a vision for the use of SRD's in safety management means that safety managers and professionals must have a clear vision about the importance of SRD's inside the organization and why they should be used, suggesting that norms and expectations for the use of data be created inside the company. I decided to group both items under this part since I believe they are directly related.

4.2.4.5 data infrastructure

This topic relates to the necessary infrastructure itself that is required in organizations to have the means to analyse data. To Wang et al. (2019), adequate infrastructure related to SRD's is fundamental, as any company that relies on data for its actions will need the necessary means to handle data efficiently. This would involve hardware, software, networks and any other tool necessary for the practice, and of course, the company needs to provide the necessary training to guarantee proper results;

Schildkamp, Poortman, and Sahlberg (2019) provide important arguments on that regard. For these authors, most of the literature about data literacy stems from developed countries, but a shift towards analysing countries that are still developing may highlight other obstacles on the process of becoming data literate. In their research, it has been found that there are cases where a lack of infrastructure is as problematic or worse than lack of training. They specifically quote cases of unreliable information systems in Indonesia, lack of access to technology in Kenya, among others.

4.2.4.6 individual attributes

Wang et al. (2019) mention how the individual knowledge of members in the organization along with their respective backgrounds and its peculiarities will influence their relationship with DLSP. Similarly, Mandinach and Gummer (2016) claim that collaboration is valued for the development of data-related activities in an organisation, which is why the knowledge of others would matter. In a different work, they mention that teachers working in data teams compensate for their lack of data knowledge (Mandinach & Gummer, 2015). Therefore, the literature points to how the knowledge of others in the organisation can influence data-related activities.

4.2.5 other elements

In general, all elements of data literacy found across the literature could be placed inside the categories displayed above. However, two other topics related to the theme stood out considering the number of times they were cited by authors. While they are not part of the model, it is relevant to mention their presence in academic texts.

4.2.5.1 role of libraries and librarians

As mentioned previously, libraries and librarians possess an important role in addressing people’s need for data literacy. While we addressed this directly in a separate topic, bringing this here again has the purpose of stressing that libraries and librarians as powerful actors whose role can potentially benefit the promotion of data literacy. The articles cited previously deal directly with how this can be done, but since it is not directly related to the main goal of this research, we do not incorporate the role of the librarian in our framework or survey and research, but it is relevant to keep this topic as one important driver of data literacy.

4.2.5.2 ethical use of data

Topics related to the ethical use of data, privacy and security were regularly debated on some articles found through the systematic literature review. As the use of data surges, so do the possibilities it can be used in negative ways. Pothier and Condon (2019) mention that data breaches and data security, privacy and questionable use of data are all common topics in that sense. Matters such as intellectual property rights, confidentiality, individual privacy, appropriate attribution and citation are all relevant topics in this regard.

While the ethical use of data is an important aspect that has to be taken into account, I decided to not incorporate it into my framework as a separate item, as I believe this concern is not unique to the domain of data literacy but it also permeates several other domains of study. With that said, I believe ethical use of data is connected to understanding data as one of its smaller components, just like understanding the functioning of the internet and data production, since I follow the understanding from Pothier and Condon (2019), to whom the essence of this competency is connected to understanding the role of data in society and how and if we can use it.

4.3 The Research and the Major Dimensions of Data Literacy

It cannot be ignored that ongoing literature has shown to not only lack standardization of the concept of data literacy but also that authors adjust proposals according to the specific needs and peculiarities of their respective fields. Data literacy is so vast and widely applicable that it can be hard to elaborate on a model that does not take into account the many existing nuances. Another relevant conclusion so far is that the general knowledge people have about data literacy is considered to be quite lacking in comparison to how present the need for data skills is. My goal is to outline characteristics of data literacy which a wider public can bear in mind, for this reason, all models shown before must be taken with the required caution that accounts for which elements can be considered for people in general and which are specifically designed for a certain approach and must, therefore, be tailored for my proposition.

Thus, I now elaborate on the main contribution of this work, which consists of 3 major parts: 1) defining, in general, the major variables involved in data literacy and how they can comprehend all the elements mentioned up to this point, both regarding what constitutes this skill as well as what influences individuals on that regard; 2) Establish how these variables are linked, proposing a framework with a differentiated approach to data literacy that focuses on basic needs for most people, which is tested through a survey and quiz; 3) Identifying what educational and professional fields are more closely associated with data skills and which require further attention.

It is important to stress that under no circumstance I suggest a model that is "better" or capable of replacing others. Rather, I try to look at data literacy through an alternative angle that elucidates the core major components that people with different backgrounds can relate to, while also providing general guidelines of which skills and factors are involved in the process of becoming data literate. The scope and resources of this work by no means enables a thorough analysis, which is why it will only be able to begin to address the lack of existing academic work regarding the need for professionals to further develop in the studies of data. If results confirm the hypotheses later made, it should provide initial guidelines for professional and educational institutions to rethink their operations and strategies regarding data, while also providing educational insights for the promotion of a wider range of development in data skills.

4.3.1 elements involved in the development of data literacy

After going through the systematic literature review, I found the frequency in which each item was discussed in materials related to data literacy. Based on the idea that several of items cited before were highly intertwined, I decided to create bigger categories which could encompass several of them while trying to maintain some boundaries where skills start to differ. The theoretical justification for each item is also based on the arguments previously provided. Therefore, this section focuses on grouping the previously mentioned items into larger domains and explaining the underlying reasons for it.

With the foundational knowledge provided so far, I summarized all items into 4 main dimensions (related to the content of data literacy) plus 3 factors which affect it directly, as I believe they determine a person's data literacy. We can summarize the framework as an internal part with the four constituents of data literacy (which I labelled as understanding data, data manipulation and technical analysis, acting on data, and critical reasoning) and an external part which reflects the main influencers of an individual's data literacy level (demographic factors, attitude, and environmental factors are the determinants of data literacy).

A relevant reference that partially inspires this idea is the model shown by Gray et al. (2018), in Figure 7. It reflects an idea that data literacy is found at the intersection of information literacy, statistical literacy and technical skills. My proposition differs in the sense that I group statistical literacy and technical skills under the banner of "data manipulation and technical analysis", based on the fact that we have seen that only select portions of society are exposed to these types of knowledge depending on their educational and professional career. The part that pertains to information literacy is put inside "understanding data".

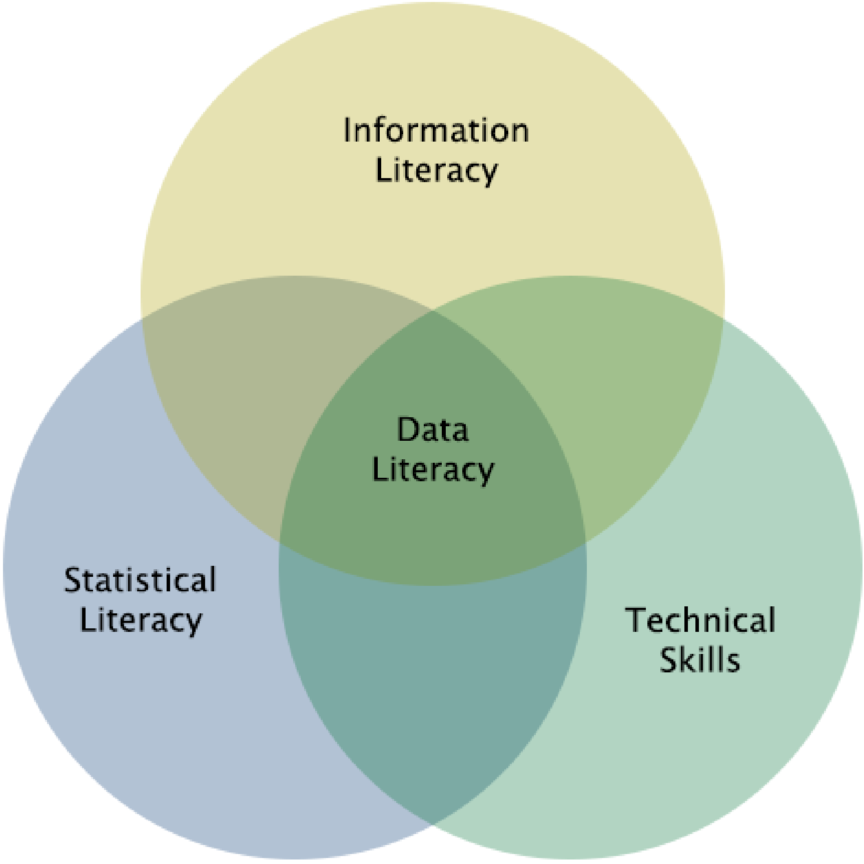


Figure 7: Adapted image from the works of Gray et al. (2018), which demonstrates the graphic from the UN Data Revolution website, available in the URL: https://www.undatarevolution.org/data-use-availability/

Another model of reference can be taken from Kjelvik and Schultheis (2019), where they present the idea depicted in Figure 8. Here the authors present quantitative reasoning, data science and authentic context as the constituents of data literacy. Before further explaining, it is important to mention that their model possesses a highly scientific approach, which is why it would not serve my purpose of a more everyday approach to data literacy, but their model can certainly provide valuable insights. As explained previously, their idea of quantitative reasoning regards the application of mathematical principles for problem-solving. With more recent development and goals similar to quantitative reasoning, data science is regarded as an interdisciplinary field that combines analytical programming to extract information from data, relying on math, statistics and computer knowledge. As their approach is a scientific one, I consider both these approaches to be part of the technical analysis and manipulation, with the consideration that part of quantitative reasoning can be linked to understanding data.

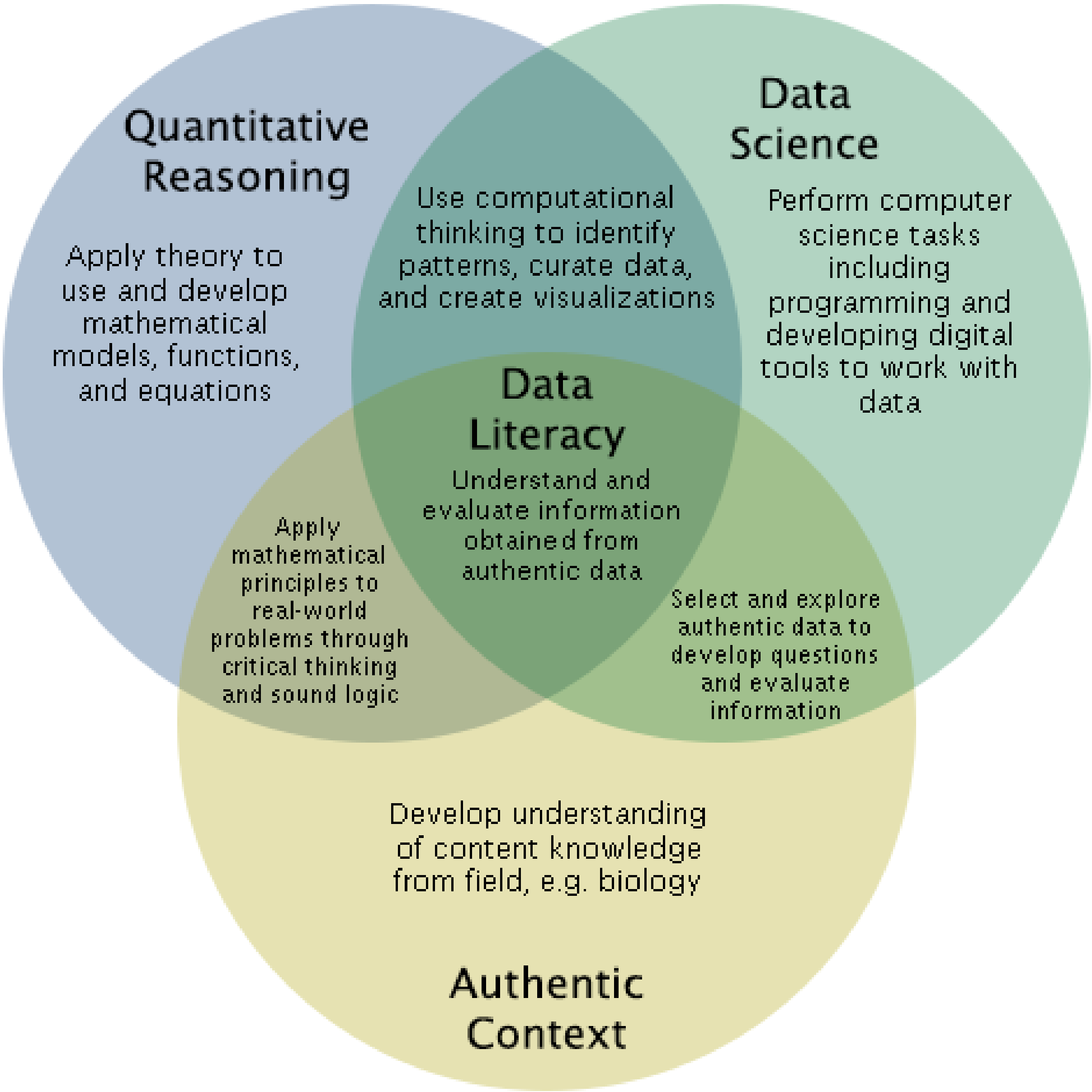

Quantitative Reasoning
Apply theory to use and develop mathematical models, functions, and equations
Data Science
Perform computer science tasks including programming and developing digital tools to work with data
Use computational thinking to identify patterns, curate data, and create visualizations
Data Literacy
Understand and evaluate information obtained from authentic data
Apply mathematical principles to real-world problems through critical thinking and sound logic
Select and explore authentic data to develop questions and evaluate information
Develop understanding of content knowledge from field, e.g. biology
Authentic Context

Figure 8: Data literacy framework proposed by Kjelvik and Schultheis (2019). Figure adapted from Kjelvik and Schultheis (2019)

My model also includes “acting on data” and “critical reasoning” as categories of their own due to the importance of extracting actionable insights and decision-making for the former and critical thinking being at the core of data literacy for the latter. The authentic context (which was briefly explained in my topic about contextual application) relies on a practical application, therefore it can also be interpreted as a part of both understanding and acting on data. Regardless, further elaboration and justification are presented in the next topics.

This relation is similar as the one described by Grillenberger and Romeike (2018) when trying to develop a data literacy competency model in the context of Computer Science and Data Science studies, in which case separating data gathering, modelling and cleansing was said to not be adequate. In their case, separating items into categories has proven difficult due to the large interconnectivity across different skills. In their model, the authors mention several competencies that are divided across major content areas and process areas, which are intertwined.

Once again, I emphasize that each dimension is not strictly separated from each other, especially considering that the following elements are present in more than one element: analytical reasoning; statistical knowledge; contextual application; data visualization; data authenticity; and data sources. In any case, the model will further be put to test in the following section of this work. Thus, my proposed data literacy framework consists of each of the elements presented next.

4.3.1.1 understanding data

The first element of my framework I decided to label as "understanding data", and it has to do with the overall notion of comprehending it, which later leads to the two next elements which require it as the base, as I will discuss later. This skill involves being able to read data, including tabular and visual forms (such as charts), as well as knowledge about different types of data, sources, how to tell if they are authentic, as well as understanding the process of data collection.

This involves mainly the aspects of reading, comprehending, collecting and interpreting data, while it also requires at least a base knowledge regarding analytical reasoning, data visualization, contextual application and some statistical knowledge. These are skills that according to Prado and Marzal (2013) as cited in Klenke et al., (2020), makes the individual capable of understanding the role of data in society, what data is, the forms data can take, how data can be interpreted, which sources of data exist.

4.3.1.2 data manipulation and technical analysis

This second category has to do with the more technical side of data, which involves the knowledge of tools and statistics, processing techniques, cleaning, evaluating and manipulating data, as well as any of the manners in which it can be done. This is often what technical professionals such as data scientists and statisticians do, but on a higher level in their case. Basic tasks, such as knowing basic functions in Excel and simple statistics are already examples of this skill. This is the skill which processes the raw data to turn it into information and all the technical processes involved. Therefore, it encompasses any process related to managing data, manipulating it, cleaning and evaluating quality, using tools and processing techniques or creating visualizations. As Pothier and Condon (2019) mention, evaluating data is directly related to statistical literacy, an argument which stresses the technicality in this task. Due to the number of situations where it will be relevant, I will later refer to it as DMTA to facilitate.

4.3.1.3 acting on data

The component we labelled as “acting on data” is related to the aspects of concrete actions derived from the data or the information obtained from it. Thus, it is related to the extraction of insights, the decision-making process, communicating findings and using visualizations to facilitate it, as well as knowing the application of data in its proper context. I also add interpretation here as I believe it is connected to the decision-making process and it occurs after the technical analyses have been performed. As Pothier and Condon (2019) explain, sometimes the professional responsible for the analysis of the data and the one responsible for its explanation is not the same, as different professions will see different value on what has been obtained.

This reasoning is also because there are professionals who focus on technicalities to provide deeper insights which then another person, the decision-maker, will act upon based on what the analyst provided. Therefore, knowing the process and knowing how to decide based on current information is connected but diverse.

As argued previously, most professionals do not need deep statistical knowledge, since those are often performed by specific roles. However, as Wilson (2016) states, many organizations, like educational ones, often want to but struggle to make data-informed decisions not due to a lack of data, but rather because of an inability to translate data into action, since professionals do not have enough data training. The author affirms that while institutions often believe on the skills educators possess, reality shows that teachers do not understand how to make data-based decisions since in many cases they do not see the usefulness and applicability of the data or how to frame the desired questions and decide on what should be invested. Van Geel et al. (2017) stress the importance that data has to lead to an effective application, instructional strategies and actionable knowledge that will address a certain problem.

With acting on data being a core element of data literacy, there is a notable need to drive change in this regard as well by highlighting it on this model, which not only comprises the final stages of the data cycle but the initial ones as well, since actions will derive and be built upon what was proposed in the earlier stages. Not knowing what has to be addressed will have lower chances of yielding a useful dataset. In the collection, analysis and actioning phases.

4.3.1.4 critical reasoning

As it was shown before, critical thinking skills are not simply auxiliary elements but rather a fundamental skill. Students need it to figure out what data means just like they need the ability to synthesize and evaluate data, assessing its quality and generating new information and knowledge (Duffner-Ylvestedt & Rayner, 2016). We have seen that data we see every day is not truly raw, but rather a selected slice of the universe, a simplification which the data creator chose to use. As Catherine D’Ignazio (2017) and Gray et al. (2018) mention, questions such as who collected the data, how, why, how it is being applied, which impacts and limitations there are, are all important matters behind any dataset but very often data is treated as a given truth. Subsection 5.2.2.5 has presented deeper elaborations on this matter, which is why I redirect the reader back there for further arguments in favour of the relevance of this topic.

4.3.1.5 demographics

Demographic personal data, such as age, educational level, profession, and others, are the starting point of what leads someone to data literacy. We have observed that data literacy is inserted in the context of some professions while lacking in others that could also benefit from it. Even in the contexts where data literacy is present, there is a disparity of knowledge in different scenarios, all of which can potentially be attributed to different levels of knowledge, experience etc.

Examples of this are the higher performance of CS/IT students in comparison to non-CS/IT in the introductory data science course made by Dichev and Dicheva (2017). In the data intervention made by Piro et al. (2014), lack of experience was frequently quoted by participants as one of the reasons they felt discomfort and avoided using data. Additionally, the fact that data literacy is a vast field and also a life-long learning process means that age and experience would play an important role. Similarly, the studies made by Klenke et al. (2020) studying the syllabi of 6 different University courses showed a larger presence of data literacy topics in the fields of Geology (69.7%) and Geography (53.2%) when compared to the others, Criminal Justice, Political Science, Journalism and Sociology, the latter with the least amount (2.1%).

4.3.1.6 attitude

My studies on subsection 4.2.3 attitude showed varied examples in which attitude can play a significant role in determining someone’s data literacy. In some cases, lack of awareness hinders one’s knowledge of data, but stimuli such as data interventions, for example, have shown to raise someone’s attitude towards data literacy. We consider this kind of intervention as attitude from environmental factors (the next item in this list) because the latter consists on recurring elements which are consistently present, such as the company’s culture and its infrastructure, while the former can be impacted by interventions since they are isolated events which happen on an occasion and spark higher interest towards the study of data, not changing the environment itself. The constitution of this element in my framework matches the ones presented before in topic 4.2.3.

4.3.1.7 environmental factors

While this item was not spoken about often, the few articles that mentioned it made a strong and logical point on how environmental factors can affect someone’s data literacy. Similar to topic 4.2.4, I consider this element of the framework to have the same characteristics presented there, with the addition that data culture and business culture could be merged into a single item. Regardless, I stress once again that rather than delving into specificities, the present study aims to outline a general framework that can be used to study data literacy.

5. Designing and Testing a Data Literacy Framework

Based on the concepts developed up to this point, I now aim to structure and elucidate how each of these elements is related and affects each other. I first present the proposed model of a framework for data literacy and then elaborate on the instruments that will put it to test.

5.1 Building a Data Literacy Framework

Based on the 7 elements presented in the previous topic, I now propose a generic data literacy framework which also comprehends both a general overview of the competencies involved as well as factors that influence a person's data skills. It is different from the previously presented models in the sense that it aims to present a concept of data literacy aimed at a larger public. It also differs on the structure, since it splits technical manipulation and actioning data as complementary but separated parts, all of which are grounded into an initial concept of understanding data, and all of them rely heavily on critical reasoning. Here follows the visual representation of the proposed model:

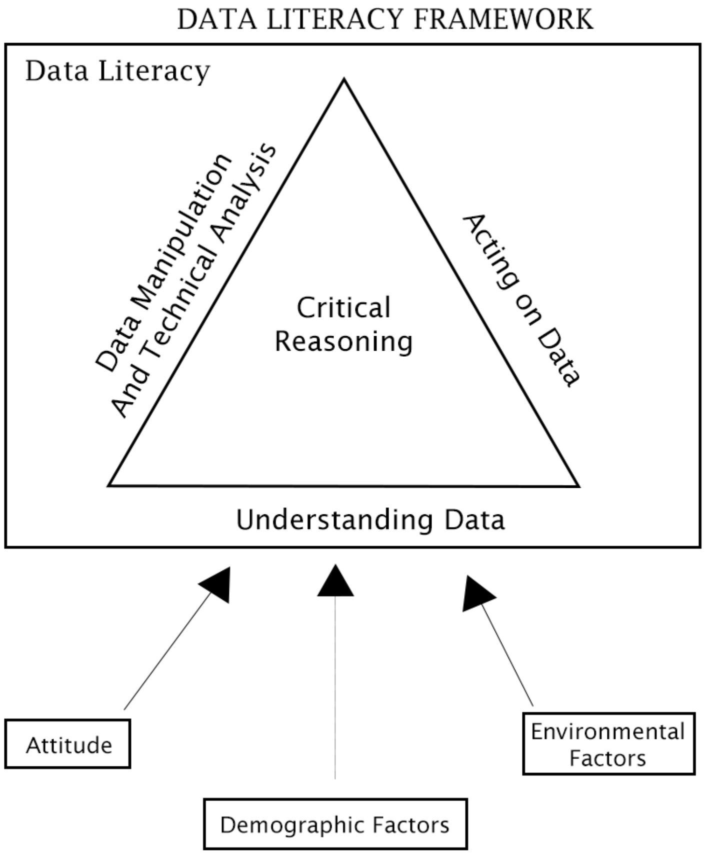


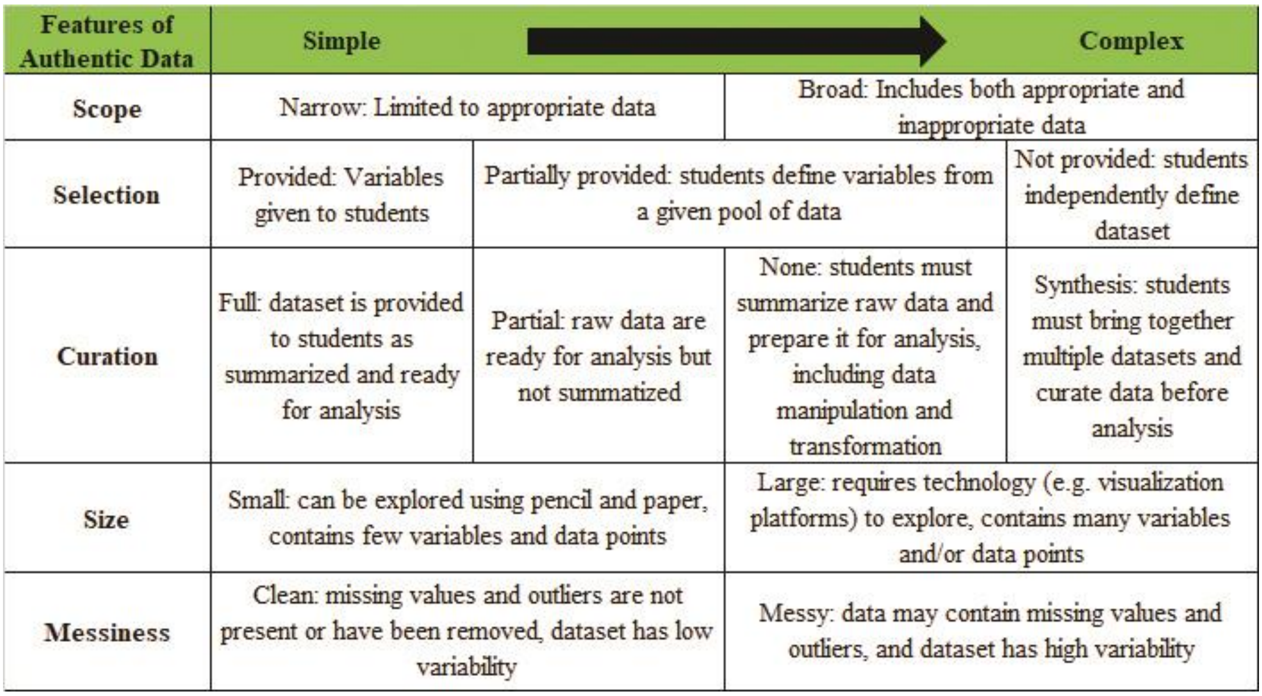
Figure 9: My proposal of a framework for data literacy

The top of the image depicts the concept of data literacy inside the rectangle, which encompasses a range of different topics. The centre of the image represents the core concept of data literacy, which is represented by a triangle in which each line represents one of the major dimensions of data literacy, which while they can be somewhat isolated, they are always connected. At the centre, we have “critical reasoning” as a core element which is present in any of the 3 other dimensions. The triangle has “understanding data” at the base since I have established it as the fundamental level from where the other concepts will develop. From there, the 2 lines that stem from it to the top represent the 2 major applications of data on a deeper level of literacy, which are “data manipulation and technical analysis” and “acting on data”. The reason why data literacy is represented by a wider rectangle in which the triangle lies at the middle is that some topics such as contextual application, authentic data and others do not belong to a single area, but rather to all of them. The bottom of the image depicts that the main factors that determine a person’s data literacy are their demographic factors, attitude and environmental factors.

As stated before, very often specific professionals are the ones who will handle technicalities involving software to process datasets, statistical analyses and building models, while a broader audience needs skills related to a general understanding of data so it can at least question data. On a professional context, very often employees will need this understanding coupled with the ability to make decisions based on the data they have. This is why the two lines that stem from the bottom of the triangle are different paths but are still connected since technical knowledge and decision-making are related but very often may be done by different people and different professionals will need varying degrees of knowledge. As an example, Pothier and Condon (2019) list seven business data literacy abilities which are fundamental for students to become data literate employees, which are: 1) organizing and storing data; 2) understanding data in the business contexts; 3) Evaluating the quality of data sources; 4) Interpretation of data; 5) Data-based decision making; 6) Communicating and presenting with data; 7) Data ethics and security. This is a view that emphasizes the understanding and acting sides of data knowledge.

Again, I emphasize that the goal for such a proposition of a data literacy framework is to elucidate a generic set of attributes which are part of data literacy and provide a visualization of how data literacy is developed in general, highlighting which types of qualifications are involved. While this does not aim to dive into deeper levels of knowledge, I do not ignore that each skillset can be honed into more specific skills that professional data users may have, but these cases are not entirely new skills outside of the framework, but rather a set of derived skills that are reached on higher levels of study. On that sense, Kjelvik and Schultheis (2019) argue that the studies of data can have various levels of complexity according to how each characteristic of the dataset is presented, which can be visualized in Table 3: how authentic datasets can become more complex depending on their characteristics. Adapted from Kjelvik and Schultheis (2019):

Table 3: how authentic datasets can become more complex depending on their characteristics. Adapted from Kjelvik and Schultheis (2019)



Features of Authentic Data	Simple			Complex
Scope	Narrow: Limited to appropriate data		Broad: Includes both appropriate and inappropriate data	
Selection	Provided: Variables given to students	Partially provided: students define variables from a given pool of data		Not provided: students independently define dataset
Curation	Full: dataset is provided to students as summarized and ready for analysis	Partial: raw data are ready for analysis but not summatized	None: students must summarize raw data and prepare it for analysis, including data manipulation and transformation	Synthesis: students must bring together multiple datasets and curate data before analysis
Size	Small: can be explored using pencil and paper, contains few variables and data points		Large: requires technology (e.g. visualization platforms) to explore, contains many variables and/or data points	
Messiness	Clean: missing values and outliers are not present or have been removed, dataset has low variability		Messy: data may contain missing values and outliers, and dataset has high variability	

On a similar line, Dichev and Dicheva (2017) and Erwin (2015), when explaining about learning with datasets, mention the theory behind the different levels of knowledge on Bloom's taxonomy, that is, a pedagogical theory that separates the knowledge of different contents under 6 levels, from least complex (remembering the content) to the most complex (using the knowledge to create new applications for it). That is to say, also in the field of data studies one concept can have different layers of knowledge, but it may still be one component. Translating that into my framework, DMTA, for example, can encompass the initial levels of simply adding data to a spreadsheet and go all the way into the construction of prediction models in Python. This is matter of degree of knowledge, but the broad topic remains the same. Having proposed the framework, it is now time to put it to test.

5.2 Research Tools

The literature review has shown that quite often a somewhat ludic approach to data literacy has managed to catch the attention of both people who have and do not have a previous interest in the topic. Some of the most notable examples mentioned here were the

Museum of Random memory (Markham, 2020) and the campus activities made by Dai (2020).

This goes in line with the studies from Chin et al. (2016). They claim that the growth of technology as a means to discover new knowledge has been driving the focus of assessments “from a retrospective, mastery model to a more prospective, process-based model”, an idea also reflected on the CCSS and NGSS in the United States. Besides, they say there is a mismatch between how assessments are done and what are the goals of preparing students to continue learning. Ebbeler et al. (2017) also agree that professional development requires time. In the case of data literacy, time is needed not only for the development of the skill itself but also for a shift in attitude. Such shift, therefore, requires a long-term approach that will over time enable professionals to incorporate data literacy into their skills. And in terms of attitude, Chi et al. (2018) emphasize in their research with teenagers that, while most of them exhibit confidence and some showed concerns about privacy or neutral attitudes, their actual knowledge was not assessed, not to mention that the authors also quote another work that has shown teenagers are unaware of some issues regarding personal data, which makes room for speculation about whether the studied sample had the knowledge to back up their feelings.

Chin et al. (2016) highlight informal learning experiences, which are often designed on the idea that students learn differently and aim to put people on a trajectory of life-long learning, which includes the development of skills and the accumulation of knowledge. In that context, they affirm that game-based technologies have proven to be useful tools to enhance both formal and informal learning. On a similar note, Wolff et al. (2015) affirm games are a learning resource that can motivate and support developing skills in the fields of data selection, cleaning, interpretation, analysis and visualization. This is the reason why they promote this skillset by utilising Urban Data Games, which are games designed to harness big data and contextually apply it in an urban environment to solve an urban problem.

Based on all the examples previously given about how data literacy is increasingly present in the life of any person and how there are notable gaps in the general public’s data skills, we make the following assumptions: 1) data literacy is a skill that is continuously developed over time; 2) every person can potentially benefit from data literacy; 3) ludic activities are useful to reach the general public and enhance the learning process.

With those premises, it was decided to create a research that is based not only on a survey but also a quiz, both of which are displayed in the online files in Annexe 1. The survey would convey questions related to how each person evaluates themselves regarding the different components of my data literacy framework, while also capturing demographic information. The quiz had the purpose to assess each of the participant's practical knowledge of data literacy via 15 questions, each with 4 alternatives and one right answer. It relies on a ludic slightly ludic format and provides feedback on the answers to raise awareness and promote learning in a few components of data literacy. This model is partially inspired by Dichev and Dicheva (2017), who conducted both knowledge tests and attitude assessments on their students on the studies of data.

Considering my background and that a considerable number of potential participants were native Portuguese speakers, both an English and a Portuguese version of the quiz were created. The quiz and the survey were both inserted into a single form on the web platform Google Forms and distributed digitally through social media. Additionally, two versions of the document were made, one which started with the survey first and then the quiz and another which follows the opposite rule, which was done as a way to identify whether someone's perception of their data skills would be affected after being trialled. Considering this applied to both English and Portuguese forms, 4 forms in total were created.

A small convenience test was performed on the survey to ensure quality, involving 5 individuals, 3 of which had a bachelor's degree in different areas, who provided more thorough feedback and suggestion of corrections. Table 4 will also show how the questions in the quiz were created considering the most relevant topics.

5.2.1 survey

The first of the instruments applied is a survey, which will be paramount to testing whether the proposed framework can be sustained. The survey applied counted with 5 questions for each of the 6 dimensions of data literacy outlined previously, plus one question asking about the respondent's knowledge on 5 different types of software tools commonly used to analyse data. Moreover, considering the preestablished relevance of demographics in our framework, the final part of the survey covered 12 questions.

Asides from the questions regarding demography and technology knowledge, the remaining had options in which respondents should choose in a 7-point Likert scale, that ranged from completely disagree to completely agree, their degree of agreeableness with select components of data literacy. The technology question included the types of software mentioned in topic 4.2.2.3 tool knowledge and its options ranged from 1 to 7, with the respective labels: 1) none; 2) vague notions; 3) basic knowledge; 4) standard knowledge; 5) Intermediate knowledge; 6) advanced knowledge; 7) professional knowledge. These labels were chosen by me as a way to provide some degree of guidance to those who would not manage to answer adequately without them. It is worth mentioning that the section regarding attitude had 2 questions towards affect, 2 regarding behaviour and only one for cognition because it can be considered that most of the other questions (Likert scales) in the survey regarded cognitive states.

The demography questions inquired about the participant’s age, gender, country of birth, country of residence, educational level, area of education, work status, average monthly income, type of function at work, work experience, the field of work and number of employees at the company. I expect these variables should provide a range of valuable information that can help to understand which kind of people are prone to developing data literacy. Additionally, it is relevant to say that participants had to option to not disclose specific information or provide a custom answer besides the options given. In the case of country of birth and nationality, the existing options were based on the closest nationalities to this researcher and were different in the English and Portuguese versions, but all of them had the custom option for the country to be typed.

5.2.2 quiz

Considering the knowledge gap aforementioned, and since raising awareness and promoting data literacy skills are all important factors for this quiz, the questions had a fairly easy level of difficulty so that potential respondents with a lower instructional level could also achieve a fair performance and not be pressured into quitting it. The campus activities made by Dai (2020) showing how data can be manipulated were also an example that inspired this activity. The order of the questions was established so that the first questions would involve little knowledge of data aspects as a way to try to minimise the number of people who would quit the survey.

Each question is aimed at training one or more of the aspects of the concept of data literacy, which means we could not assess attitude, environmental factors, and demographics here, which was already previously done. In Table 4, I show how each quiz question relates to the topics they are aimed at.

Table 4: Distribution of the quiz' questions across topics

Quiz distribution	Understanding data	Data manipulation and technical analysis	Acting on data	Critical reasoning
1	Yes			Yes
2	Yes			
3	Yes			
4	Yes	Yes		
5	Yes	Yes	Yes	Yes
6	Yes	Yes		Yes
7	Yes	Yes		Yes
8	Yes	Yes	Yes	Yes
9	Yes			Yes
10	Yes	Yes	Yes	
11	Yes		Yes	Yes
12	Yes	Yes	Yes	
13	Yes	Yes	Yes	Yes
14	Yes	Yes	Yes	Yes
15	Yes	Yes	Yes	Yes
Total	15	10	8	10

As we established that understanding data is the base from which other elements derive, it has been natural that it was in a way present in all questions. Moreover, how each question fits its category requires some interpretation and can be considered arbitrary by some, however, more important than a strict fit, the main goal here is to try to distribute topics as possible while maintaining a degree of accessibility for respondents. With that mentioned, I now present a brief justification of each question.

5.2.2.1 critical thinking applied to an everyday situation

The first question tackles an everyday situation regarding prudent forms of evaluating messages received on the application Whatsapp or in any other place, which reflects a common phenomenon in which fake news has been spreading at a fast pace and leading major platforms to give growing attention to the phenomenon.

As I had previously stated, fake news is one of the topics that affect society as a whole and the capacity to handle them properly has been outlined as one very desirable and useful use of data/information skills. Therefore, this question involves the understanding of how data is used online, interpretation skills, critical reasoning and questioning obtained information.

5.2.2.2 general knowledge of how data is used on the internet

The second question further addresses the knowledge of internet activity. As a lot of the data production occurs in the digital environment, we opted to verify people's understanding of everyday topics, such as how sites handle information, personal data tracking and recommendation algorithms.

5.2.2.3 general knowledge of how companies can handle data

Furthering on the topic of how our data is handled, the third questions seeks to explore how companies handle our data, dealing with topics related to purchases, prices and ads. As Kristin Fontichiaro and Oehrli (2016) affirm, a growing amount of data is collected by companies and in many cases, the citizens are not aware of it. This understanding is also in line with the analysis made by Hautea et al. (2017) about the "Scratch" platform previously mentioned.

5.2.2.4 the different averages and the US election

This question was directly inspired by the example provided by Cruz and Rubio, (2016). In this case, there are different ways to calculate an average, but some people may often not be aware of the different ways it can be assumed. In the example given, the mean would be too distant from most candidates and there is no mode, while the median would be a more representative number of the sample given. This question also follows Kirstin Fontichiaro and Oehrli (2016), who present the differentiation of mean, median and mode as one of the elements of statistical literacy.

Here, statistics notions, as well as the comprehension of how data is affecting something, are tested components.

5.2.2.5 poor formulation of sentences

This question was also inspired by Cruz and Rubio (2016), but this time following the "Rough Guide to Spotting Bad Science"[7] they present. The question presents questions with conflict of interest, unrepresentative samples, cherry-picked results as well as an option with broad and unprecise generalization. This can also be seen as an exemplification of the theory of data infrastructure literacy.

Therefore, this question deals with understanding aspects of data collection, differentiation of useless data, decisions based on the understanding connections, questioning information, inquire about external elements and general critical reasoning.

5.2.2.6 spurious correlations

Here, a case of spurious correlation[8] was presented. It tackles both statistical notions and the ability to read charts as well as the notions of correlation and causation (also mentioned in the Guide above cited). By using a real example that can be considered amusing, we aim to bring awareness of how correlation and causation are not necessarily intertwined. As argued before, being able to discern correlation from causation is a piece of important knowledge for those who seek to be statistical and literate (Kirstin Fontichiaro & Oehrli, 2016).

[7] For the direct material, refer to: http://www.compoundchem.com/wp-content/uploads/2014/04/Spotting-Bad-Science.pdf.

[8] The example was directly extracted from https://www.tylervigen.com/spurious-correlations, whose content is under a Creative Commons Attribution License and has captured some attention before as it has been quoted by news sites.

With these considerations in mind, this is an example that tests reading charts, interpreting data, statistics knowledge, understanding of patterns, questioning of information, pondering about external elements and elimination of noise in a dataset.

5.2.2.7 being able to question information

In this task, the respondent needs to be able to know how which questions could be effective to question an allegedly good product. Kirstin Fontichiaro and Oehrli (2016) cite a list of most and least healthy states made by the company behind the application "MyFitnessPal" and mention that questioning how the data was gathered (only to find it was based solely on users of the app) is an example of how a savvy librarian would act when confronted with potentially biased information.

The question deals, therefore, with knowing how the collection of data works, understanding how it can be transformed into information, questioning information, pondering about external elements and different scenarios.

5.2.2.8 surveys with inadequate samples

Here the case of inadequate samples was addressed directly. While a straightforward question, answering it requires some attention to how biased the sample was as well as a general academic understanding that very often it is not possible to survey all of the individuals that meet the criteria of the research.

For the reasons explained above, this question deals with an understanding of how data is collected and how they affect the work, notions of statistics, generating insights from data and understanding how they are connected, as well as comprehending how external elements affect the data obtained and having critical reasoning.

5.2.2.9 identifying misleading information

The ninth question brought a real case in the United Kingdom in which it was claimed that Colgate advertised the results of its research misleadingly. While it would be necessary to know what specific manipulation of information they applied, the options were created in a way that logic and attention could suffice.

The respondent is, therefore, led to use the abilities to understand how data is collected, questioning information and general critical reasoning.

5.2.2.10 line chart analysis and decision making

In this case, a line chart was presented, and the respondent is put in the role of a decision-maker with a pre-established strategy. Not only understanding the numbers presented was enough but also it required some degree of logic to reason that the number of students approved in the exam can never be higher than the number of students enrolled and as the lines are shown get closer, it means the percentage of approved students is increasing.

Thus, the question addresses reading charts, making sense of how the data is affecting the work, understanding patterns and the process of transformation into information, actioning information and decide to understand the connection of different components.

5.2.2.11 interpretation of acts in social media

Coming back to the topic of social media, question eleven based itself on a real Brazilian case of a reality show participant and Twitter. Similar to how it happened, the question embraces the need for thinking of how elements work and the ways they can be used before reaching a specific conclusion. This example is a direct illustration of the previously presented idea of data infrastructure literacy, as it tackles the underlying rules that govern a given social media, including what it can and cannot do, how it works and what has to be taken into account when analysing data (Gray et al., 2018).

This question trains the understanding of data production online, interpretation, communication of data, questioning information and thinking about alternative scenarios.

5.2.2.12 understanding tables and interpreting results

This case involves a simple representation of prices and the amount units sold, connecting them to how much input was needed for a product that had minimal sales but little costs of production, which is the key to understanding its worth and deciding about the worth of a product-related decision.

The components involved here are the interpretation of data, understanding the transformation of data into information, providing insights from the information, data-driven decision making, actions based on the understanding the connection between different components and communicating what has been found.

5.2.2.13 understanding bar charts and omission of data

For this case, a bar chart was supposed to show numbers from every year from 2001 to 2015 but omitted some numbers, which was the key to answer the question, requiring some attention to detail which could be encouraged by reading the correct option. The question provided a few insights so that fewer incorrect assumptions could be made about the missing years.

The components involved in this question were mainly the ability to read charts, statistics, deriving insights from information, actioning information and critical reasoning.

5.2.2.14 comparing results and manipulation of information

Question 14 focused on the analysis of a bar chart which the baseline had to be zero but instead was a high number, leading to a visual representation of the numbers which make their difference look considerably higher than the numbers show, a misleading representation. The already explored notion that visualizations convey what the creator wanted is emphasized here (deriving from the examples of data infrastructure literacy and InfoVis). Moreover, this is a direct example of Edward Tufte's Lie Factor, in which the full width of a chart is manipulated to create a distorted perception (Womack, 2015).

Therefore, here I involved the aspects related to reading charts, interpreting data, creating insights and actioning and questioning information.

5.2.2.15 correlations and logic in business

The final question presents a table with different data from a period in which there was an increase in ice cream sales. While the options provided make the question easier, the topics involved in it comprehend both the capacity to understand tables, correlations and basic business knowledge.

The skills involved in this question are understanding which data affects the work, interpreting data, transforming data into information, differentiating useless data, visualizing patterns, providing insights, making decisions based on how items are connected and eliminating noise data.

6. Results

I will now further refer to the questionnaires that had the survey part first (both English and Portuguese) as Type 1 and the ones that started with the quiz as Type 2 and the following information is summarised in Table 5.

Table 5: Overview of the questionnaire results

Valid Answers	Type 1		Type 2		Total:
	English	Portuguese	English	Portuguese	
	5	99	12	93	209
Invalid Answers	Duplicates		Not answered properly		Total:
	6		1		7

In total there were 104 valid answers in the Type 1 questionnaire, 5 being from the English version and 99 from the Portuguese one. Questionnaire Type 2 had 105 valid answers, 12 being from the English version and 93 from the Portuguese one, totalizing 209 answers (which is shown in the files of Annexe 1).

After gathering the answers, I merged all 4 versions of the questionnaire into just one, then translated all answers in Portuguese to English, cleaned the dataset and started the data analysis using Excel and SPSS. Asides from these answers, 7 were considered invalid and deleted, due to 1 appearing to be invalid (the individual answered 7 in all survey questions including the question about knowledge of tools, which is an unlikely scenario; yet, this same respondent got a score of 3 on the quiz, which likely means it was not a proper attempt) and 6 for being duplicates. The results below will show questionnaires Type 1 and 2 separate as well as merged into a single one.

6.1 Breakdown of the Demographics Sample

Before explaining the sample, a few notes must be made. Since some people opted to answer with the custom option instead of the given one, some adaptations where made to unify the results. Though limited to only a few results, the most discretionary adaptations were: researcher, legal, physiotherapist and marketeer to analyst/specialist; policeman, photographer, production assistant and public employee to operational; English teacher to professor. Moreover, in the educational level, the option “post-graduate” was added due to the number of people who typed in the custom field some sort of degree that fits in it; In the field of work, biotechnology was turned into “health” and “English” was turned into education. A few outliers in the field of work were turned into “other” due to not matching any of the given alternatives and being isolated cases. Additionally, the option “autonomous” on the dataset was changed to “self-employed”.

Also, an extra variable called “expatriates” was created by isolating people who were born in a country but reside in another, an attempt to see if there is a difference in this kind of population considering being an expatriate is a condition with several possible causes, such as need, educational and/or professional opportunities, resourcefulness etc. For posterior analysis, I will recode “Prefer not to disclose” as a missing value.

The findings were that Type 1 respondents are a little older, with a slightly higher educational level and income. Type 2 concentrates more of the population inclined towards STEM fields. All tables mentioned are presented in Annexe 3.

- AGE: The age of the sample was more inclined towards a younger population, with a slightly older presence in Type 1.

- GENDER: The gender was closely balanced, but with a slightly higher presence of males.

- COUNTRY OF BIRTH: By far, most people were born in Brazil, followed by Portugal and the rest was distributed across 12 countries. In total, 7 chose not to disclose. Because of these results, this variable will not be used in the analysis is there is not sufficient data for it.

- COUNTRY OF RESIDENCE: Again, Brazilian respondents top the table, but this time with a smaller percentage, as respondents living in Portugal and the United Kingdom see an increase in number. Here again, there is insufficient data to use this variable, so it will not be further considered.

- EXPATRIATES: In both cases, a little under 15% of the sample consisted of expatriates. Considering the total sample, the biggest group was of Brazilians living in Portugal, which amounts to a little over 40% of the total cases of expatriates.

- EDUCATIONAL LEVEL: Most respondents had at least a Bachelor's degree or equivalent education (or expected to finish within 12 months), followed by about 20% with a Master's degree.

- FIELD OF STUDY: Respondents came largely from the fields of social and behavioural sciences, followed by STEM fields. It is worth mentioning that Type 2 has a population more inclined towards STEM studies

- WORK STATUS: Most respondents were employees, followed by students.

- INCOME: The majority earned up to 2.5 times the minimum wage or equivalent, followed by a portion of people who preferred not to disclose the information. "Around minimum wage" or "no income" were the next positions, being not far behind.

- TYPE OF FUNCTION: The largest concentration of results was in the option "none", followed by analyst/specialist and then intern.

- EXPERIENCE: In line with what was shown so far, as a young sample, about 75% of the respondents had up to 7 years of experience, with 36,4% having up to 3 years.

- NUMBER OF EMPLOYEES AT ORGANIZATION: In both cases, the respondents chose the option "non-applicable", probably due to not being employed. In Type 1, the answer "over 500" was the second most chosen option, followed by "1 to 10" and "11 to 50". In Type 2, the second most chosen option was "1 to 10", followed by "11 to 50" and then "over 500", with also over 10% preferring not to disclose.

- FIELD OF WORK: Since the field of work question allowed for multiple answers, a separate dataset was created to allow the counting of all variables, which is why the total number surpasses 209. The sample was mainly comprised of people in the legal area, with administrative, education and technology having a considerable presence.

- QUIZ SCORE: In general, quiz scores had an average of over 60% right answers. In Type 1 the two questions with worst result were also the only ones below 50%, these being question 4 (basic concept of averages, with 23,08% score) and question 13 (with a bar chart with omitted results, where finding the right answer could be attributed to the respondent's attention to details. The average score was 25,96%). These were also the worst scoring questions in Type 2, but there 3 in other questions the accuracy fell between 40% and 50%.

6.2 Overall Results of Variables and Comparison of Questionnaires

Next, I present the averages given in each section of the survey, while also comparing these to the results in the Quiz. According to Aarts, Van Den Akker, and Winkens (2014) Cohen's d can be used to compare the mean value of numerical variables between two different groups, which in this case will be used to test if the means obtained for Type 1 and Type 2 questionnaires were significantly different. It is calculated by subtracting the two means and dividing by the average of their standard deviations. It is considered that values around 0,2 can represent a small effect, values around 0,5 are a medium effect and values around 0,8 are a large effect. Table 6 summarizes all values found:

Table 6: Summary of all values for Cohen's d

Cohen's d	Mean	Question 1	Question 2	Question 3	Question 4	Question 5
Quiz	0,13	N/A	N/A	N/A	N/A	N/A
Understanding data	0,20	-0,10	0,20	0,51	0,31	0,13
DMTA	0,05	-0,15	0,04	0,16	0,20	0,08
Software	-0,22	-0,19	-0,10	-0,23	-0,31	-0,28
Acting on Data	0,12	0,07	0,00	0,15	0,12	0,28
Critical Reasoning	0,14	0,18	0,10	0,14	0,09	0,19
Attitude	0,06	0,00	0,07	0,00	0,24	-0,03
Environmental Factors	-0,02	-0,01	0,05	0,02	-0,13	-0,04

In the end, only question 3 of "Understanding Data", which is about reading charts, had a Cohen's d that was not low. Though it was not a considerable difference, it is a slight indication that answering the survey after doing the quiz reduced respondents' confidence on that particular skill, but there is not enough evidence considering both the small differences in samples and the fact that the higher quiz score in sample 1 could indicate that this sample had indeed a higher data literacy. The details of each category are commented in the following topic and the respective tables are present in the files of Annexe 1 and tables in Annexe 4, in which case The dark blue column in these tables represent the mean of all 5 questions per section.

- UNDERSTANDING DATA

Most people, in general, answered with high values for their understanding of data, meaning they indicate they possess good knowledge on the topic, with question 2 (which regards collecting data) having the most outlying results slightly below the rest of the questions. This variation understandably had a lower result as it has some technical aspects in it. In this variable, Type 2 had results with a slightly higher variation than 1, though the difference is small. In all cases but one, Cohen's d was low, except for question 3, regarding the ability to read charts, with a medium value.

- DATA MANIPULATION, TECHNICAL ANALYSIS AND TOOLS

As can be expected, results for DMTA were lower than the results for other variables, though still with a mean above 4. The differences in mean values for Type 1 and 2 were also lower, as expressed by Cohen's d. It is worth pointing to the fact that the variations in the results in each question could be an indication that the questions measured items that were not as close to each other.

This section, as it can be expected, had the lowest average results. Curiously, Cohen's d was negative in all questions, which means respondents in Type 2 provided a higher evaluation than Type 1. This can be attributed to the fact that Type 2 has a higher presence of STEM professionals/students. In any case, the difference was not significative. Also, the very high kurtosis in some questions, especially in Type 1, and skewness display how the results were highly concentrated on the side of lack of knowledge.

- ACTING ON DATA

This section had the second-highest values for the averages a very small difference in Cohen's d, meaning there is barely any effect in the differences between versions. It also had the second smallest standard deviation, which alongside the skewness, shows how positive the results were. However, the values for kurtosis were not as high, so there was some spread among the high values.

- CRITICAL REASONING

Where “acting on data” had two “second places”, critical reasoning had the respective first places, having the highest averages and smallest standard deviation. In general, values were concentrated on the “upper end” of the scale, with a notable kurtosis in some questions in Type 1. It could be that the lower results in Type 2 and its dilution, as shown by the kurtosis, indicate the quiz had a difference, though Cohen’s d shows it was not significant.

- ATTITUDE

In this case, results were only slightly skewed towards the upper half, and the negative kurtosis shows the dilution of the results. A notable difference in type regards the mode for the third question, which is about believing that your current knowledge of data to be good. A significant number of respondents opted to answer 2, but the -1,06 kurtosis, the second most negative number across the sample, shows how diluted the results are, which is why the median and average are so different from the mode.

- ENVIRONMENTAL FACTORS

This section had results skewed to the upper side, but also the highest values for standard deviation, as well as only negative kurtoses and the most negative number obtained for kurtosis in question 4 of Type 1. Moreover, Cohen’s d value was very close to zero in all cases and it was also mostly negative. These results serve as an indication that Type 2, having a bigger STEM composition, believes to have had a background more closely related to data studies, but results, in general, were spread.

- QUIZ RESULTS

No major differences were observed across the different versions. Both versions had a median of 10 and mode of 11 and a mean close to 9,5. Values for standard deviation were slightly higher in version 2, and consequently, its kurtosis was negative, indicating this version had values slightly more spread. In either case, skewness was negative, corresponding to how most people were correct in over half the questions. Cohen's d was 0,13, which indicates it was quite low and there is not enough data to assume that making two versions had different results on the quiz, so when it comes to analysing the versions separately, neither the results were notably affected by the order nor was the opinion about the respondents' knowledge.

When comparing both Type 1 and 2, so far, in a few cases Cohen's d was even negative, which means there was an increase respondents' perception about their knowledge in the second version. Curiously, in several cases, Type 1 had higher values than Type 2, even though the former had more people in STEM fields. This smaller result in Type 2 can either be an effect from doing the quiz first or since STEM professionals know how deep and how much knowledge these fields can involve, they take a humbler approach when answering. Unfortunately, my current data is not enough for a final answer, so this observation can be considered for future research.

Based on these facts, the rest of the analysis will follow based on the total results. Nonetheless, the small hints at a difference could serve as potential future research, since there was some difference in each sample, albeit small.

6.3 Preparing the Dataset for Analysis

Before proceeding to explore the relations between the variables in the dataset, I now aim to analyse the data with a factor analysis and normality test, so that most adequate statistical procedures can ensue in the next chapter.

6.3.1 factor analysis – principal component analysis

First, to make sure that each group of 5 questions per dimension of data literacy is homogeneous, factor analysis was executed in SPSS with the principal component analysis extraction method. Osborne (2015) argues that oblique rotations are used when the factors are allowed to correlate, which usually is the case for Social Sciences (and naturally occurs here, but I show in Table 37 that most variables have correlations with statistical significance with varying coefficients), which is why I opted for an oblique rotation, more specifically, the Direct Oblimin (the full file can be found in Annexe 1).

Table 32 shows that the Kaiser-Meyer-Olkin Measure of Sampling Adequacy is close to 1, which means the factor analysis is useful for the data. Moreover, this is further confirmed in Bartlett's test of sphericity, which shows a significance value of 0,000. Table 33: total variance explained with Eigenvalues above 1 shows that a total of 7 factors were found with an Eigenvalue above 1 (the remaining were removed from the table but can be visualized in the original file). Figure 10 shows the Scree plot that illustrates these same results.

The results of the analysis are shown in matrices in Table 34 and Table 35. It can be noted that component 7 does not have significant results. However, every single other dimension of data literacy matches perfectly one component and with considerable values, except Data Manipulation and Technical Analysis, which was grouped with Attitude on component 1. Nevertheless, to avoid mixing these 2 variables (which would imply in some loss for the model, as attitude is an independent variable and DMTA a dependent one) I still considered them as two distinct groups. Moreover, the second question of DMTA had a value below 0,4 on its second question, which is why it will be removed from the subsequent analyses of DMTA. Nevertheless, this removed question regards statistical knowledge and is important, so I will include it in the model and test it separately.

To summarise, almost all items have passed the analysis and can be grouped into a dimension for subsequent analysis. Because averaging Likert scales can be deceiving, I use the SPSS function to round all numbers to the closest integer. For example, the 5 questions regarding understanding data are averaged (creating the "Understanding_averaged" variable) and then rounded (a new variable labelled as Understanding). This latter type variable is the one used in analyses. I will further refer to these groupings as "average variables".

Component 1 is mainly related to attitude and on a smaller scale (yet relevant) to DMTA. While it can be hypothesized that people with a better attitude towards data literacy are also more likely to know technical procedures, I cannot exclude DMTA from the analyses, but this analysis, when interpreted together with the correlation between attitude and DMTA (Table 36), begins to hint at a strong relation between them.

6.3.2 normality test

Next, to determine which statistical procedure will be most adequate for some data types, I perform a normality test on variables that regard score and the dimensions of data literacy.

Table 36 in the annexes shows that for all the variables in both Kolmogorov-Smirnov and Shapiro-Wilk tests the significance value for all variables was of 0,000, which means the null hypothesis is rejected and we can conclude that the data does not follow the normal distribution in all variables. Annexe 1 has all SPSS tests where it is also possible to see that the histograms, mostly due to their skewness, are not close to the normal distribution. Even when attempting a logarithmic or square transformation, the skewness is still considerable. Due to the absence of normality, the analysis was based on non-parametric tests, mainly Spearman correlations and logistic regressions.

6.4 Exploring the Relations Between Variables

With the preliminary tests done, I now attempt to understand the relations between the variables in the sample begin to try to test 2 generic hypotheses and answer the following underlying questions:

1) *A person's perception is not necessarily correlated with their knowledge*. The questions here are: is there a correlation between a person's perception of themselves and their results on the quiz? Are there areas of data literacy which are correlated? A positive attitude towards data literacy influences data knowledge? How do environmental factors act on someone's data literacy? This hypothesis comes from the fact that it has been shown in the literature that many individuals are unaware of the many aspects involved in data literacy. This "perception" is measured by the Likert scales, which are by nature attitude scales. However, in terms of attitude, the Likert scales were designed only to measure cognitive dimensions and not affective and behavioural dimensions. For this reason, it will be important to later test each dimension of the Attitude variable to see if the theory holds.

2) *Demographic factors, attitude and environmental factors determine a person's data literacy*. The main questions are aimed at understanding which aspects these are, such as which professional paths are more closely associated with a data literate person? Do people with specific backgrounds or certain types of attitude feel they have more or less data

literacy? Being older and having more general work experience translates into higher data literacy? These are 2 generic hypotheses that will be broken down to explore each existing dimension in the questionnaire.

For that purpose, I will use SPSS and execute the correlation for every demographic question, the means of each category of questions in the survey and the result of the quizzes. Where applicable, string variables were converted in SPSS to numeric order. As a separate dataset was created for the 2 questions regarding the type of work, they were calculated separately. Moreover, for this purpose, I created extra variables which are the average of each group of 5 questions from the survey, except for DMTA, in which case its second question (about statistics) was excluded as per the results of the factorial analysis, resulting in the average of the remaining 4 questions. However, I also consider the statistics question separately to test it, under the variable name of "Statistics knowledge".

6.4.1 performing a binary logistic regression in SPSS

Because of the type of data obtained in the dataset and its inherent problems (several variables with some not having enough answers, lack of normality), linear regression would not be ideal. Considering the dependent variables of the dataset are ordinal ("Score", "Understanding", "DMTA", "Software", "Acting" and "Critical"), they would ideally be adequate for ordinal logistic regression, however, the data fails the assumption of proportional odds, which is why it would not be possible. However, to still make use of the dataset and retrieve a few additional insights, I decided to recode the dependent variables so that values of 5, 6 and 7 (partially agree, agree and completely agree) would become a new variable. For example, those who have any value from 5 to 7 in "Understanding" would be a 1 in the new variable "Understanding_Literate", and any value below 5 would be a 0, thus creating the new binary dependent variables identified with the "literate" at the end of their names, which is fit for a binary/binomial logistic regression. While not an ideal solution, this should provide a few insights for future research.

However, the Score and Software variables are measured differently (the first ranges from 2 to 14 while the latter ranges from none to Professional knowledge). For the Score variable, I created 3 dummies named Score_High (12 or more) and Score_Positive (9 or more). For the Software variable, I created Software_High (5 or more) and Software_Positive (2 or more). This way it can be tested the groups of people with a high knowledge and groups that have at least a “positive” knowledge. The average quiz score ranged between 9 and 11 and over a quarter of respondents had a value below 2 in Software, which is why those are the thresholds.

Due to the complexity of the topic and considering how each software handles binary regression differently, I briefly elucidate here how the binary logistic regression was executed in SPSS. At any rate, all regression files are also available in Annexe 1. It is also worth mentioning that all the procedures that are shown in this subtopic follow the premium subscription content published by Laerd Statistics (2017)[9], to where I direct the reader for a deeper understanding of any of the procedures.

6.4.1.1 assumptions

The first assumption is that dependent variables must be dichotomous, which was achieved with the transformation mentioned before. The second assumption is that independent variables must be either categorical or continuous but ordinal variables can be used as long as they are treated as categorical, which is a condition we meet. Assumption 3 deals with the independence of observations, so that if one element fits into one category of the variable, it should not be possible to fit another, being therefore mutually exclusive, which is the case here.

Assumption 4 regards the sample size, which includes the fact that each category must also meet a certain size. Because several variables have options with a very low frequencies, I recoded the independent variables that will be used in the analysis according to how answers are distributed, which was the following: 1) Age became “Age_26_Plus” (which is roughly half of the sample); 2) Educational level became “University_Standard_Education” (for those with Bachelor Degrees) and “Post_graduation_education” (over half the sample has a Bachelor’s degree and very few have less than that, making this the best available cut point); 3) Field of education became “STEM_Education” and “Social_Education” (the two of the 3 options that are likely to handle data more often); 4) Work Status became “Employee” and “Entrepreneur” (the latter

[9] A free introduction to the assumptions, procedures, pseudo R Square, classification table and Exp(B) values can be found in: https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php

is for self-employed and business owners and this division makes that all non-working options become the 0 value); 5) Income became “Above_Min_Salary_Range” (considering all the people that did not disclose, this is the closest option to half of the sample); 6) Field of work became “specialists” (analyst/specialist, professors and managers/directors); 7) Work experience became “Experience_3_Plus” (roughly half of the sample); 8) Number of employees became “Company_11_Plus_Emp” (roughly half of the sample); 9) Attitude with values of 5 and more became “Attitude_Literate”; 10) Environment with values of 5 or more became “Environment_Literate”. The latter 2 were done similarly to the dependent variables.

Assumption 5 requires a linear relationship for the continuous variables, but my dataset this type of variable, so it is not a problem. The sixth assumption of the model is that there should be no multicollinearity (when 2 or more variables are highly correlated), which can happen when two or more independent variables are highly correlated. According to Laerd Statistics (2015), This test can be done in SPSS by making a linear regression and opting to see the collinearity diagnostics (I emphasize that country of birth and country of residence are not being used due to insufficient data and that fields of work will be treated separately because of the transformations that variable had to go through). The test involves using one dependent variable and dummy variables for all independent variables (which are the demographic, attitude and environment variables). Moreover, changing only the dependent variable in this test does not change the results. This test is shown in Table 39, where it can be seen that all variables have a tolerance value above 0,1 and thus, meet the assumption of no multicollinearity. Note that by default SPSS excludes a dummy from each group to avoid resulting in full collinearity.

The last assumption is that there should be a lack of significant outliers, high leverage points or highly influential points. This is shown by the values generated in the Casewise list for each regression. In the cases where entries in the dataset have a ZResid value above 2,5 or below -2,5, it means they are potential outliers, however, while it may flag the need for caution in the interpretation because this study aims to understand initial insights for data literacy, the few cases with outliers will not be removed but rather interpreted taking into account limitations.

6.4.1.2 executing the procedure and interpreting results

On SPSS, the Binary Logistic Regression was chosen and in each case one of the dependent variables was tested and the independent variables were all the same (Demographic variables, Attitude and Environment) and they were all previously classified as categorical variables for this. On the options tab, the changes were to show classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals and CI for exp(B): 95%. The display was changed to "at last step". The results will involve a series of tables with the following interpretation. The following tables are also backed by

The following tables are also based on the explanations provided by ReStore (2011) Starting from the section "Block 1: method = Enter", the Omnibus Tests of Model Coefficients tests how well the model predicts the categories in comparison to the lack of independent variables, in which case a statistically significant value (Sig. below 0,05) is desired, rejecting the null hypothesis. The Hosmer and Lemeshow Test aims to evaluate how poorly the model can predict, which is why here values above 0,05 are needed instead.

According to IBM (2016), in regression models that use a categorical dependent variable, it is not possible to obtain a value for R Square with all the characteristics from the R Square obtained in the linear regression, which is where Pseudo R Square comes in as an approximate solution. The table Model Summary contains the Cox & Snell R Square and Nagelkerke R Square, which are sometimes referred to as the pseudo-R Square just mentioned and have a lower value than in multiple regressions but can still be interpreted, albeit with more caution. To Laerd Statistics (2017), Nagelkerke is claimed to be usually preferred.

Classification table shows the percentage of cases that correctly fit into the expected result of the dependent variable assuming a standard cut value of 50%. Finally, the main results are shown in the table Variables in the Equation (which by default eliminates one variable per category to avoid redundancy, such as eliminating "gender male" since "gender female" answers the prediction), where it will be important for the result to be statistically significant. The most important value will be the Exp(B), which is what will explain how much a change in each variable will influence the outcome (fitting or not the condition of the dependent variable). Here, values above 1 show how more likely that category is to fit the dependent variable, and values below 1 mean less likelihood. With the basics explained, I now start the analysis.

6.4.1.3 general adequacy of the model

I will first present the general results for each binary combination to show the adequacy of the model, but their full interpretation will be addressed in topic 7. Discussion). All the following tables mentioned will be a compilation of the individual SPSS tables that exist for each test per variable. Every respective SPSS file is also available on Annexe 1, but the same tables on the annexe are also copied in the Excel file available.

All variables reject the null hypothesis in the Omnibus Tests of Model Coefficients and fail to reject it in the Hosmer and Lemeshow Test, meaning they pass both tests for adequacy. They also all present Exp(B) values capable of explaining the relations, but I will address this later. Table 40 shows the results for the variable Score_Positive and Table 41 for Score_High. Nagelkerke R Square indicates a respective prediction of 29,3% and 28,6% of cases. Classification is right in, respectively, 78,8% and 81,2% of cases. In each variable, only three cases have a slightly high ZResid, indicating they are potential outliers. Table 42 has the results for Understanding_Literate, with a Nagelkerke R Square explaining 40,7% of cases and classification table with an overall percentage of 83%. Casewise List shows 7 outliers, so some caution is needed. Table 43 is for DMTA_Literate and it explains the variance in 34,6% of cases with overall classification percentage of 75,2% and a Casewise list with 3 outliers with values only slightly above the threshold. Table 44 is for Statistics Literate and the R Square explaining 38,1% of cases and overall percentage in the classification table of 74,5%. Casewise List shows 4 but 3 are very close to the threshold.

Table 45 is for Software_Positive and Table 46 is for Software_High. They respectively present a Nagelkerke R Square for 47,7% and 40,2% of cases and classification table with an overall percentage of 77% and 90%. Regarding outliers, they respectively have 4 (2 close to the threshold) and 3 (1 close to the threshold). It should be noted that the Exp(B) value of STEM_Education for Software_Positive was abnormally high, but it was not statistically significant. While it may not be reasonable to adopt this number, it is safe to hold the assumption of a positive relationship because the dataset shows higher Software values for those in STEM.

Table 47 is for Acting_Literate with Nagelkerke R Square explaining 32,7% of cases and classification table with an overall of 82,4%. 6 outliers are present in the Casewise List, enticing some caution in the interpretation. Finally, Table 48 represents

Critical_Literate and explains 45,1% of the variance with the overall percentage in the classification table of 95,2%. 2 outliers are found. Table 49: summarizes all values of Exp(B) and their significance.

6.4.2 hypothesis 1: a person’s perception is not necessarily correlated with their knowledge

To start, I intend to compare the results of the Likert scales with the quiz scores via a correlation analysis so that it can be understood whether individuals are aware of their data skills. With that done, I executed the correlation analysis (Spearman correlation, due to the lack of normality in the data), which can be found in Table 37 in the annexes.

The results have indicated that shown that only the variables “Understanding”, “DMTA” and “Statistics Knowledge” are statistically significant, all of them at the 0,01 level. Moreover, these are also the ones with the highest correlation coefficient. All of these 3 were positively correlated, nevertheless, the highest coefficient was of 0,281 (for DMTA). That confirms the expected results that in the end, people with some technical knowledge (and therefore a more specific background) would score higher, however, the low correlation coefficient in all 3 cases (even lower for the remaining variables) shows that individuals do not tend to evaluate themselves well or have different perceptions as to what it means to have data knowledge. Considering DMTA had higher values, it can confirm the previous statement that people with a specific background know well how complex it can be and are therefore capable of more properly evaluating their results. The score variable will be further tested by checking whether people with higher or more specific educational levels have a specific subset of answers that corroborates these expectations.

6.4.3 hypothesis 2: demographic factors, attitude and environmental factors determine a person’s data literacy

To start, I used the same correlation analysis to test how attitude and environmental factors correlate with the remaining variables, which yielded positive results. Only the score variable, as mentioned above, did not correlate with these 2, but all the remaining ones had significant correlations, being 0,000, except for the correlation between “Environment” and “Critical” (0,004).

Regarding the correlations with the attitude variable, it can be seen that results ranged from a coefficient of 0,392 (for critical reasoning) up to 0,648 (DMTA), while statistics knowledge (0,542) and understanding data (0,59) were also notable. The high correlation between attitude and DMTA was expected as per the results of the factor analysis. I will further mention this in topic 7. Discussion, but for now, it is relevant to argue that a positive attitude towards data translates well into the several components of data literacy, especially for DMTA, which is a more technical variable, being restricted to a smaller universe of people.

When it comes to the "Environment" variable, the case of critical reasoning was also the weakest correlation (with a coefficient of 0,2 and significance of 0,004). All the other variables had a significance of 0,000 and a coefficient between 0,302 and 0,393, except for attitude, with a coefficient of 0,431. In summary, there is a moderately weak positive correlation between environmental factors and the other components of data literacy.

In the end, while the data suggests the importance of environmental factors to data literacy, attitude played a bigger role in these results. The regression results also confirm this but I will elaborate on it at the discussion. Also, based on both variables, there is some evidence that points to critical reasoning towards data not being highly linked to people with a more positive attitude or specific background (though the link still exists).

6.4.4 further exploring the case of attitude

I decided to further explore the attitude variable both due to the relevant results it yielded but also because, as argued, it is a variable that has the following 3 dimensions with its respective questions:

Affective
Question 1: I feel confident in my abilities handling data in general
Question 2: I am interested in the studies of data
Cognitive
Question 3: I believe my current knowledge of data is good
Behavioural
Question 4: I usually bear in mind how data works in my everyday activities that may involve it

Question 5: When using the internet or electronic devices, I make decisions taking into account how my data is used

Therefore, I performed a separate Spearman correlation (Table 38) with the individualized questions. In most cases the variation of the coefficient in each case was not considerable, however, there are a few points worth emphasizing.

First, correlating to the score, only question 2 had some significance and also the highest coefficient (though it was still low, being 0,168). Question 5 had the lowest coefficient in all variables, except for the score, while question 4 had the highest in 3 cases. The only cases where the coefficient was notable (at least 0,498) was the correlation between statistical knowledge ad Q1AT and Q3AT (0,5 and 0,583 respectively), understanding data and Q3AT and Q4AT (0,498 and 0,556 respectively) and DMTA with the first four questions (0,548, 0,546, 0,617, and 0,646, respectively).

6.4.5 other findings

The data analysis also yielded a few other results worth mentioning. When considering the Spearman correlation, most items had some statistical significance. But the ones that stand out the most (correlation coefficient above 0,47) are the correlations between statistical knowledge and understanding data (0,613), statistical knowledge and DMTA (0,701), statistical knowledge and Software knowledge (0,53), understanding data and DMTA (0,684), understanding data and acting on data (0,483), and software knowledge and statistical knowledge (0,550), DMTA and acting on data (0,478) DMTA and software knowledge were also something to consider, albeit with a slightly smaller correlation (0,446).

Despite statistical knowledge being left out of the DMTA components in the factorial analysis, it still retained a notable correlation, the same with software knowledge, since they are all part of the technical domain of data literacy. Moreover, my consideration that the elements of understanding data are a foundational aspect for DMTA and acting on data seems backed by the correlation found.

7. Discussion

It is fundamental to adopt a few premises before interpreting the results. 1) this work is exploring apparent gaps in the literature regarding general elements that can influence the average individual's data literacy, and as such, it needs more research and refinement because this work is limited to a preliminary observation that would warrant further confirmation in future research. Nonetheless, the theory developed here matches the general ideas of the literature and the data collected has shown to be adequate for analysis in general, but as it is usual with social sciences, a plethora of factors can affect the outcome of the research so the results should be taken with caution; 2) The lack of normality in the data implied the choice of certain statistical tools that required performing small transformations in the dataset. While there was no change to the values, the way each variable was categorized under one group instead of other can be done in different ways that only further experimentation could show more insights; 3) The sample used here was a convenience sample of 209 people most of which are known by the author of the work and this eliminated the potential diversification of results and even excluded the country variables; 4) Variables in Table 7 (which shows the summary tables of Exp(B) values and significances while also flagging all relevant results in terms of statistical significance, which will be the base for the discussion) that failed to achieve statistical significance will not be possible to be considered. All these premises mean that the data here should be interpreted bearing in mind that this is an indication of potential cause and effect relations but, under no circumstance a final model capable of being widely applied. It can, however, show existing patterns that lead to some insights, as the methods and variables applied here were broad in a way that intended to flag potential findings that can be further explored in future research.

Table 7: Summary table with Exp(B) value and significance in all regressions with flagged relevant results (statistical significance)

Summary of prediction - Value of Exp(B)									
Variables:	Score_ positive	Score_high	Understanding_ Literate	DMTA_ Literate	Statistics_ Literate	Software_ Positive	Software_ High	Acting_ Literate	Critical_ Literate
Gender(1)	2,527	1,074	0,959	1,839	2,010	2,773	0,490	1,057	0,761
Significance:	0,042	0,881	0,929	0,135	0,106	0,023	0,414	0,912	0,758
Expatriate(1)	1,051	1,642	0,998	0,774	1,493	3,893	0,462	0,479	0,125
Significance:	0,936	0,450	0,998	0,661	0,518	0,063	0,543	0,312	0,108
Age_26_Plus(1)	0,296	0,137	1,327	0,985	1,979	0,296	0,251	1,700	0,662
Significance:	0,023	0,001	0,587	0,975	0,182	0,038	0,196	0,367	0,694
University_Standard_ Education(1)	2,238	5,261	6,111	1,237	3,986	2,058	0,920	0,637	5,741
Significance:	0,226	0,157	0,022	0,742	0,097	0,305	0,955	0,534	0,216
Post_graduation_ education(1)	8,008	8,640	6,291	1,105	2,619	2,862	0,410	3,615	8,582
Significance:	0,015	0,076	0,039	0,892	0,282	0,190	0,574	0,214	0,156
STEM_Education(1)	1,265	2,015	1,310	0,857	1,705	#########	18,358	0,734	0,107
Significance:	0,706	0,334	0,726	0,802	0,390	0,998	0,030	0,694	0,272
Social_Education(1)	2,008	1,517	0,636	0,716	0,656	0,360	1,587	0,967	0,018
Significance:	0,148	0,447	0,386	0,468	0,390	0,031	0,724	0,951	0,029
Employee(1)	0,226	1,371	0,357	0,342	0,457	0,583	0,195	0,345	0,681
Significance:	0,018	0,649	0,137	0,078	0,232	0,407	0,172	0,118	0,788
Entrepreneur(1)	0,278	4,094	0,793	0,274	0,966	0,605	4,476	1,369	0,465
Significance:	0,083	0,065	0,741	0,061	0,962	0,500	0,314	0,692	0,521
Above_Min_ Salary_Range(1)	1,119	4,418	4,391	1,914	1,865	1,322	1,298	1,959	5,148
Significance:	0,811	0,007	0,004	0,134	0,198	0,549	0,781	0,189	0,107
Specialist(1)	3,776	0,790	1,102	1,447	0,984	0,735	0,957	2,421	0,242
Significance:	0,018	0,686	0,874	0,479	0,976	0,609	0,961	0,192	0,217
Experience_3_ Plus(1)	1,118	0,518	0,499	1,393	0,365	1,031	6,220	0,860	0,270
Significance:	0,841	0,272	0,232	0,525	0,090	0,958	0,102	0,799	0,182
Company_11_Plus_ Emp(1)	0,867	4,118	0,544	0,802	0,908	2,628	3,144	1,084	17,328
Significance:	0,767	0,007	0,215	0,615	0,844	0,049	0,263	0,875	0,037
Attitude_Literate(1)	0,734	1,476	10,255	6,194	8,764	3,152	15,593	7,859	10,865
Significance:	0,533	0,453	0,000	0,000	0,000	0,016	0,033	0,001	0,090
Environment_ Literate(1)	0,761	1,415	1,609	1,489	1,638	1,652	6,410	1,474	1,288
Significance:	0,565	0,484	0,356	0,362	0,288	0,272	0,144	0,473	0,832

- Predicting factors

I start by talking about all the predicting factors that were found in the analysis. First, I outline that the variables Expatriate, Entrepreneur and Experience_3_Plus did not have any statistical significance and had varying B values in their results. For all the cases that lacked statistical significance, it could be due to a methodology flaw or a peculiarity of the sample, so while they cannot be readily excluded from the theory, my model cannot support these variables.

Environmental Factors also lacked significance so it cannot be applied, however, it should be noted that in all cases except for the Score_Positive, the Exp(B) value was higher than one. Moreover, it has a weak but statistically significant positive correlation with all variables but Score. It suggests that it can be argued that environmental factors indeed can have the potential to affect data literacy, but my model did not manage to properly address it.

Gender had an Exp(B) value above 1 in most cases but it was significant only for Score_Positive and Software_Positive. Since male was coded as 1, then being male seems to have a higher likelihood of having software knowledge (2,7 times) and quiz score (2,5 times). A likely explanation could be linked to cultural aspects, as it is known that males predominate in math-related studies, even if the chosen field of study is not math-related. This thinking can be coupled with studying the fields of education we have in the sample. STEM_Education had an Exp(B) value above 1 in most cases, including one abnormally high value for Software_Positive, however, it had no statistical significance just like most other variables. Nevertheless, Software_High did have significance and its Exp(B) value indicates that STEM students are as much as 18 times more likely to have a high software knowledge, though this number should be read with caution since less than 20 people in the sample had such knowledge. Similarly, when analysing Social_Education, most values of Exp(B) were below 1 (inverse relation) and Software_Positive and Critical_Literate had statistical significance. The value for Software corroborates the idea that STEM students are naturally more linked to some aspects of data literacy, in this case, Social Sciences students are 2,7 times (1/0,360) less likely to have Software Knowledge than the rest of the sample. However, the surprising result for Critical Reasoning (being lower for students of social sciences) will be addressed later on its specific topic.

As it could be expected, University education had a positive result for most cases, though few retained its statistical significance. People in the category University_Standard_Education are 6,1 times more likely to have the qualities of the variable Understanding Data and people categorized as Post_graduation_education are 6,2 times more likely. The latter group is also 8 times more likely to have a positive score in the quiz.

Age_26_Plus had an Exp(B) value below 1 for Score_Positive, Score_High and Software_Positive. While somewhat surprising, considering data literacy is developed throughout the life, this can perhaps be attributed to newer generations being digital natives and as such, are quickly becoming adapted to specific trends that pertain to these times. While research (Hautea et al., 2017) shows young people do not always understand these processes, they have a wide knowledge of technology and its inherent trends, which makes it understandable that they would obtain good scores and know how to use software. However, it cannot be discarded that this here was a convenience sample and thus, such findings can only hint at a possible fact.

The variable Employee had low results in Score_Positive, but since the variable Entrepreneur did not have significance (and also low scores), it is hard to assume there is a category of work status that could reasonably be used in the model. Considering types of work, Specialists are 3,7 times more likely to have a positive score in the test. This could be linked to this group being composed of people with certain qualifications (similarly to the educational variables) that would make them more effective at understanding data-related problems, especially if considered that it is a natural expectation for companies that workers have some data skills (Pothier & Condon, 2019).

Moreover, Above_Min_Salary_Range had values above one for all variables but only Score_High and Understanding_Literate had significance, and people are approximately 4,4 times more likely to score high on the quiz or meeting the criteria of "Understanding Data". While this by itself cannot be an insight, when it is analysed together with educational levels it could hint the expected observation that education would play a role here, not to mention that because the relevance of the minimum salary varies across countries, there should be some caution interpreting this. Regarding the variable Company_11_Plus had notable Exp(B) values for Score_High (4,118), Software_Positive (2,628) and Critical_Literate (17,328). This is another variable that by itself hardly has meaning, but it can hint that people in larger companies tend to develop certain digital and problem-solving skills needed.

Finally, regarding attitude, it had the most positive results for this research. Except for Score_Positive, all variables had an Exp(B) value above 1, although this variable and Score_High and Critical_Literate did not have statistical significance. The results show that people who have a good attitude towards data tend to have considerably better odds at achieving good results with data literacy, more specifically with the variables

Understanding_Literate (10,255), DMTA_Literate (6,194), Statistics_Literate (8,764), Software_Positive (3,152), Software_High (15,593) and Acting_Literate (7,859).

- Quiz Score

The first important analysis is that, while it could be expected that people with higher data literacy would be scoring higher on the quiz, this was not necessarily true. Only 3 variables (Statistical Knowledge, Understanding and DMTA) had a statistically significant correlation with the Score Variable and even so, the highest coefficient (DMTA) was of 0,281, showing a weak correlation. Even attitude and environmental factors had both failed the correlation test but also the binary logistic regression, where they had no statistical significance.

For the variable Score_Positive, it is more likely to occur for males, people with less than 26 years, people with higher educational levels, people who work in a position that normally require qualified work (specialists) and in work positions different than employees, but this last one is difficult to interpret. Score_High is more likely to occur for people with less than 26 years, with a salary above the minimum range and who work in companies with more than 11 employees. I also note that Post_graduation_Education almost had statistical significance (0,076).

The quiz applied was not largely different from other types of quizzes and interpretation questions commonly seen nowadays and it also relied on statistical knowledge that is commonly seen in today's educational systems pre-university. It can be argued that for these reasons younger people would have a higher score and as mentioned before, since males predominate in math-related fields, this can potentially explain the gender gap. Educational levels would also be logically a predictor, but the salary, size of the company, type of work and work status do not present such ease, with the exception that in one way they can all symbolize types of positions that usually require a more developed skillset that enables them to achieve higher positions, be specialists in certain matters and have a better income. It should be noted that when I correlated each attitude question with the score, the one had that was statistically significant was about being interested in data, though the coefficient was low (0,168).

By filtering the Excel file, we can find a few additional insights. Those with a high school education had a mean score of 8,45, while those with a Bachelor's degree have a mean of 9,6 and those with an educational level above it have a mean of 10,28. People below

26 years had a mean of 10,05 and older people achieved a mean of 9,17. In the end, it stems from the development of education and constant contact with the digital world and its direct relations with data literacy, which in other words can be translated into practical experience handling data, but unfortunately, my model is not capable of providing a final answer to this.

- Understanding Data

People who are more likely to fit the category Understanding_Literate are the ones with a Bachelor’s degree (6,111) or higher (6,291), with an income above minimum salary (4,391) and with a good attitude towards data (10,255). As could be expected, educational level played an important role here while salary could be directly linked to it. The variable Attitude has shown to be the most relevant, so it shows that those who take an interest, who feel confident with data and bear in mind how it works on their daily affairs, will tend to believe they possess good capabilities with this skill. While it was expected that people who feel more inclined to data studies would likely have a higher level of data literacy, the numbers found for this variable show emphasize how big the gap is. It seems that data literacy is a skill that, though it is important for most people, it is highly concentrated in some people.

Moreover, the fact that Understanding Data correlates well with all other variables supports its foundational position in the framework.

- DMTA, Statistical Knowledge and Software knowledge

Now on the technical side, as expected, these variables are highly connected as seen on the correlations, which implies that people who tend to learn one also tend to develop one of these competencies tend to develop the others as well to some degree.

Because of the lack of statistical significance in most answers, DMTA_Literate and Staitistics_Literate could only be predicted by Attitude_Literate, with a notable result. People with a positive attitude towards data are 6,1 times more likely to fit the category DMTA_Literate and 8,7 times more likely to fit Statistics_Literate. Once again, attitude plays a relevant role in predicting variables related to data literacy and emphasizes how concentrated data literacy might be in some types of people.

As for software knowledge, the type of people expected to have some knowledge is males (2,7 times), people younger than 26 (3,3 times, which is 1/0,296), people outside of the field of social sciences (2,7 times, 1/0,360), workers on companies with more than 11 employees (2,6 times) and people with a positive attitude (3,1 times). High knowledge is expected to come from people in STEM education (18,3 times) and with a positive attitude towards data (15,5 times).

While I lack the full panorama of possible explaining factors, previous studies point in very similar ways to this scenario.

- Acting on Data

Only one variable was able to predict this variable. People with a positive attitude towards data seem 7,8 times more likely to fit the category Acting_Literate. It was expected that more factors could predict this variable, but the fact that, similarly to DMTA and Statistical Knowledge, only the variable attitude was a predicting factor. This seems to emphasize how DMTA and Acting on Data belong to a part of data literacy that is developed mostly by attitude than other factors. While Attitude also largely predicted Understanding Data, 4 other variables also did predict. Not discarding potentials flaws with my model, with the existing data it can be estimated that, from a certain point, attitude will be the determinant factor that leads to higher data literacy, taking people from the base of the triangle (understanding) to the upper parts (DMTA and acting).

- Critical Reasoning

This variable had a few different results. As mentioned previously, it had smaller correlations with other variables than the others. It also had the smallest coefficient in the correlations with Attitude and Environmental Factors. Also, Attitude_Literate did not have statistical significance here (0,090), but it should be noted that it was close and if it did have, its Exp(B) value was notable (10,865).

The only factors that did predict it were Company_11_Plus_Emp, with an Exp(B) value of 17,328, and Social_Education with a value of 0,018, meaning people outside this field are 55 times more likely to fit the variable Critical_Literate. As for the company size, once again I put forward the argument that maybe being part of a larger company could be associated with developing a skillset that leads people to believe they are more qualified in that regard. But as for the abnormal value regarding Social Science education, only further tests would be able to clearly explain this, however, it can be estimated people from this field, which often deals with unprecise terms and need critical reasoning, would judge themselves more moderately because they know better how much proficient they have to be with certain skills so that they would not easily choose a high number on the questionnaire. The score variable has shown that people fail to evaluate themselves well it also does not correlate well with critical reasoning, so it should be no surprise that by not knowing the many aspects involved in the studies of data, people would miscategorise themselves or have very different opinions about their skills.

- Overall analysis

Finally, based on all observations made, a few general points can be made. First, the fact that Understanding Data correlates well with all other variables and that it could be predicted by more factors than the rest seems to show how more accessible it is than, for instance, DMTA and Acting on data, which justifies the triangular framework proposed in this work.

Also, DMTA and related topics such as statistical and software knowledge were indeed linked to a more select range of people, namely those with interest in the area, those working in companies, those with a specific educational background in STEM or younger people. While it is natural for this specific set of skills to be more segmented, the fact that Software_Positive (which was based a mean result of 2-4 in a scale that goes up to 7 and where 2 means simply vague notions) highlights the gap in this field as well. Spreadsheet software such as Excel can aid in a multitude of tasks and it is still a notable gap in people's knowledge, especially considering that even if the individual answered "1" for every other software and "4" (standard knowledge) for Excel, the final mean would be 1,6, rounded to 2, meaning they would fit the category Software_Positive. But the sample shows that in the Software variable, 68 people had a value of up to 1,4 and 127 had a value of up to 2 and the total mean value was of 2,11, which demonstrates a great need for improvement in that regard.

While data literacy is a field of study that is present on people's lives, this study confirms previous studies that it is still more concentrated on people that are interested in the area, and therefore, further themselves in the studies of data. Still, the literature has shown cases such as museums (Markham, 2020) and courses (Dichev & Dicheva, 2017) that have opened the eyes of people to the benefits of data knowledge. Therefore, it is important to raise awareness of the benefits of data literacy and how present it is in people's affairs. One example that confirms such lack of awareness is that only about 47% of the people answered correctly the quiz question about how companies can handle data (question 4) and only 60% answered right the question which had a manipulated bar chart (question 14), as demonstrated in Table 23. When filtering the dataset, it is also possible to see that the younger population (up to 25 years old) answered these questions 55 and 74 times respectively, while the older one got it right 44 and 53 times, another argument for the younger population being more used to intricacies of the digital world.

Another argument for the need for such awareness is that most people cannot evaluate themselves properly in terms of data skills. The low correlations of the score variable and critical reasoning, as well as the specific predicting factors being unrelated to having a specific professional background. Moreover, the score was mainly determined by people who have more contact with academic studies and/or the digital world, whereas the belief that someone had good data capabilities translated poorly into the quiz score. This puts into perspective that Likert scales may not be the best option to evaluate a person's data literacy as they do not understand well the many aspects involved in this skill. As D'Ignazio (2017) mentions, there is a gap wherein on one side specialized organisations employ technology that captures, manages and uses the data of a multitude of people and on the other side, most people are unable to use data are instead the subject of data studies and fail to use data adequately for many daily activities that are directly linked to citizenship, such as understanding the advertisements they see every day or grasping political arguments. This gap is also true in the professional domain, considering the mismatch between what companies expect and what professionals can offer.

The model failed to completely prove the framework, but it allowed the conclusion that attitude is a key factor developing data literacy and that there is some concentration of knowledge in the hands of those more closely connected to technology and data studies, though I did not manage to measure demographics properly.

7.1 Limitations and Suggestions for Future Research

With all that has been argued, the limitations presented here cannot be ignored. First, the present work used a convenience sample which was limited to the people who are more accessible to the author, a restriction in both in the quantitative and qualitative (characteristics) terms. The sample was more inclined towards a younger population, more often based in Brazil or Portugal, many of which are either students or recent graduates and coming from social and behavioural sciences. Moreover, not all possibilities for the demographic variables could be tested and eventual problems with the questionnaire or any research methodology applied cannot be discarded.

This need for a more refined and precise model is particularly strong in the environmental factors, which failed to predict the variables and also had very modest values in the correlations. It should also be considered that this set of questions was aimed at the current "professional environment" of the individual, which could not be the prevalent environment for this person's professional path. On this sense, questions regarding the educational background could have had more to do with the environment, but the design of this research was different. Still, the statistical significance in the correlations existed in most cases and they all show a positive correlation, so this should be a starting point for future research to focus specifically on aspects of environmental factors to further refine this dimension.

Because the link between people's perceptions and their score is weak, future research needs to test data literacy in more practical sides, because there is evidence that Likert scales would not be the most adequate means of research, considering that either only more knowledgeable people can assess their knowledge correctly or there is a considerable variation in people's perception about their knowledge. Moreover, chapter 4.2 The Variables Involved in Data Literacy attempted to explore the key aspects which are discussed in the academic literature about data literacy, so hopefully, this can guide future research on the understanding of this complex concept.

This study aimed to explore several variables that could be linked to achieving data literacy, but only parts of it were successfully demonstrated. A model with so many variables would need more resources (time and a larger sample, for example) but it never intended to thoroughly explain the relations. Considering all the limitations, I believe it can shine a light for future studies that may want to focus on a specific part of data literacy and/or choose only a few predictors that they explore deeply.

8. Conclusions

Throughout the present research, I have arrived at a few conclusions. First, one of the consequences of technological advancement was enabling one to produce, collect, use, manipulate and analyse data in several different ways. It has become common for new products and services to be data-centric and for organisations to leverage this scenario to improve their results. In that same context, we observed that people, in general, are not prepared to handle data on daily activities and in the professional environment, a mismatch between institutions' expectations and practice. I also noticed that academic literature did not go into length about the professional aspects and which elements can affect someone's data literacy.

The systematic literature review showed that data literacy is a relatively recent term which has not been unified by scholars, as each of them present variable definitions which emphasize or present aspects depending on their specific context, be it data literacy for education, for safety, for business, for researchers etc. However, there are core common elements present in all cases, usually regarding the capacity to read, interpret, visualize, process, communicate and act on data. There is also a notable gap in the literature regarding the professional applications of data literacy and the need for its promotion because companies' expectations and the skills of most professionals are not accurately matched.

It should be noted that there is no uniform concept of data literacy and any work developed on that regard shines a different light that can be considered for further research. As for the predictors of data literacy, my proposed framework showed strong results regarding attitude but only partial results for demographic factors and it failed to produce notable results for environmental factors. Additionally, it seems reasonable to think that Understanding Data is the one aspect that is more easily accessible to most people (as the mean results, in general, were higher) while Critical Reasoning, Acting on Data and DMTA especially need some extra attention.

Still, the need for data literacy is present in daily activities for academic and corporate scenarios as well as civic activities. It should be efficient to promote basic knowledge and interest as soon as possible because academic knowledge and attitude have proven to be related to higher data literacy, and this is the same line of thinking that justifies being recommended for educators to learn about it early in their careers as a way to facilitate learning (Mandinach & Gummer, 2016) and data literacy is a life-long learning process, therefore introducing the foundation early should ease the process of development of the skill later. The status quo is that the skill is still very restricted to certain types of people who are connected to the digital world or particularly interested in the field, however, the literature has shown cases where the interest for this area can be sparked in a broader sample. Interest should be a key aspect in developing a positive attitude and this by its turn should lead to data literacy. Therefore, it could be the case for educational institutions to promote more capacitation in this field for a wide audience, explaining its relevance, but without going deeply into technicalities which some people may be averse to. This omission of technicalities may be especially justified, considering Mandinach and Gummer (2016) has argued that there is no need for it even for educators because they will mostly need only a part of these skills. I believe that the aspects of Understanding Data and Critical Reasoning should be the main focus while aspects of DMTA and Acting on Data should be offered on a basic level for a wide audience and more deeply for those whose interests have been further sparked, but such a proposal would have to be tested to see if such efforts would provide positive results.

Still, it is undeniable that the need is there. Google Trends has shown some data-related topics which are becoming increasingly common and the Quiz Scores show that there is a lot of room for improvement. While surely most people answered correctly over half the questions, the level of the quiz was not very high, and the sample analysed was relatively biased towards people who have some access to an academic degree. In other words, a broader and more diverse sample would likely result in lower scores and overall lower data literacy.

Finally, it should be noted that the existing gaps in the literature are concerning. There is a very strong need for data literacy in civic activities (D’Ignazio, 2017) and business needs (Pothier & Condon, 2019) but still most work, understandably, is focused on the academic side. Also, most data interventions, such as Kippers et al. (2018) and Dunlap and Piro (2016) were aimed at educators so that they can assess results in their practice but not necessarily pass on the knowledge to students, as they themselves are not always qualified in data skills. However, these interventions are an attempt to fix the lack of early introduction to data literacy. Instead, I believe that trying to capacitate people in data skills from early education would be more efficient. Some educational measures such as the North American NGSS and the results of this research are evidence that such a link between younger populations and data literacy may already be existing, albeit in need of development. While this term is relatively new and with theory under development, its importance is already present in society and this has consequences. There is a strong need for people to become data literate.

References

Aarts, S., Van Den Akker, M., & Winkens, B. (2014). The importance of effect sizes. *European Journal of General Practice*, *20*(1), 61–64. https://doi.org/10.3109/13814788.2013.818655

Chi, Y., Jeng, W., Acker, A., & Bowler, L. (2018). Affective, behavioral, and cognitive aspects of teen perspectives on personal data in social media: A model of youth data literacy. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10766 LNCS*, 442–452. https://doi.org/10.1007/978-3-319-78105-1_49

Chin, D. B., Blair, K. P., & Schwartz, D. L. (2016). Got Game? A Choice-Based Learning Assessment of Data Literacy and Visualization Skills. *Technology, Knowledge and Learning*, *21*(2), 195–210. https://doi.org/10.1007/s10758-016-9279-7

Cruz, T., & Rubio, M. (2016). *Data Literacy Strategies to Bolster Student Election Understanding*.

D'Ignazio, C. (2017). *Creative data literacy*. *23*(1), 6–18.

Dai, Y. (2020). How many ways can we teach data literacy? *IASSIST Quarterly*, *43*(4), 1–11. https://doi.org/10.29173/iq963

Dichev, C., & Dicheva, D. (2017). Towards Data Science Literacy. *Procedia Computer Science*, *108*, 2151–2160. https://doi.org/10.1016/j.procs.2017.05.240

Duffner-Ylvestedt, N., & Rayner, J. (2016). Hooking Up Data with Literacy: Creating an Educational Framework for Uppsala University Library. *Nordic Journal of Information Literacy in Higher Education*, *8*(1), 38–44. https://doi.org/10.15845/noril.v8i1.261

Dunlap, K., & Piro, J. S. (2016). Diving into data: Developing the capacity for data literacy in teacher education. *Cogent Education*, *3*(1). https://doi.org/10.1080/2331186X.2015.1132526

Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (2017). The effects of a data use intervention on educators' satisfaction and data literacy. *Educational Assessment, Evaluation and Accountability*, *29*(1), 83–105. https://doi.org/10.1007/s11092-016-9251-z

Erwin, R. W. (2015). Data Literacy: Real-World Learning Through Problem-Solving With Data Sets. *American Secondary Education*, *43*(2), 18–26. Retrieved from www.gapminder.org.

Fontichiaro, Kirstin, & Oehrli, J. (2016). Why Data Literacy Matters. *Knowledge Quest*, *44*(5), 21–27.

Fontichiaro, Kristin, & Oehrli, J. A. (2016). *2016 - Fontichiaro & Oehrli - Why Data Literacy Matters*. *44*(5), 21–27.

Gibson, J. P., & Mourad, T. (2018). The growing importance of data literacy in life science education. *American Journal of Botany, 105*(12), 1953–1956. https://doi.org/10.1002/ajb2.1195

Gould, Robert. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25.

Gould, Roy, Sunbury, S., & Dussault, M. (2014). *In praise of messy data: lessons from the search for alien worlds*. (November), 31–36.

Gray, J., Gerlitz, C., & Bounegru, L. (2018). *Data infrastructure literacy*. *5*(2), 205395171878631. https://doi.org/10.1177/2053951718786316

Grillenberger, A., & Romeike, R. (2018). Developing a theoretically founded data literacy competency model. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3265757.3265766

Gummer, E. (2013). *Factors Influencing the Functioning of Data Teams Kim*. 1–31.

Halliday, S. D. (2019). Data literacy in economic development. *Journal of Economic Education*, *50*(3), 284–298. https://doi.org/10.1080/00220485.2019.1618762

Hautea, S., Dasgupta, S., & Hill, B. M. (2017). *Youth Perspectives on Critical Data Literacies*. 919–930.

IBM. (2016). Evaluating the Model. Retrieved July 29, 2020, from https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/tutorials/plum_germcr_rsquare.html

Johnson, N. E., & Zwicky, D. A. (2017). GRIP: A university's program develops tracks to bridges, a professional development opportunity. *Proceedings - Frontiers in Education Conference, FIE, 2017-Octob*(1993), 1–3. https://doi.org/10.1109/FIE.2017.8190690

Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention? *Studies in Educational Evaluation*, *56*, 21–31. https://doi.org/10.1016/j.stueduc.2017.11.001

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting Messy with Authentic Data: Exploring the Potential of Using Data from Scientific Research to Support Student Data Literacy. *CBE - Life Sciences Education*, *18*(2). https://doi.org/10.1187/cbe.18-02-0023

Klenke, C. M., Schultz, T. A., Tokarz, R. E., & Azadbakht, E. (2020). Curriculum Data Deep

Dive : Identifying Data Literacies in the Disciplines Let us know how access to this document benefits you . Full-Length Paper Curriculum Data Dive : Identifying Data Literacies in the Disciplines. *Journal of EScience Librarianship*, *9*(1), 0–16.

Koltay, T. (2015). Data literacy: in search of a name and identity. *Journal of Documentation*, *71*(2), 401–415. https://doi.org/10.1108/JD-02-2014-0026

Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA Journal*, *42*(4), 303–312. https://doi.org/10.1177/0340035216672238

Koltay, T. (2017). Data literacy for researchers and data librarians. *Journal of Librarianship and Information Science*, *49*(1), 3–14. https://doi.org/10.1177/0961000615616450

Laerd Statistics. (2015). Ordinal logistic regression using SPSS Statistics. Retrieved July 24, 2020, from https://statistics.laerd.com/

Laerd Statistics. (2017). Binomial logistic regression using SPSS. Retrieved July 23, 2020, from https://statistics.laerd.com/

LaPointe-McEwan, D., DeLuca, C., & Klinger, D. A. (2017). Supporting evidence use in networked professional learning: the role of the middle leader. *Educational Research*, *59*(2), 136–153. https://doi.org/10.1080/00131881.2017.1304346

Mallavarapu, A., Lyons, L., Uzzo, S., Thompson, W., Levy-Cohen, R., & Slattery, B. (2019). Connect-to-connected worlds. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300237

Mandinach, E. B., & Gummer, E. S. (2015). Data-Driven Decision Making: Components of the Enculturation of Data Use in Education. *Teachers College Record*, *117*(4), 1–8. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=102822237&site=eds-live

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*, 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Markham, A. N. (2020). Taking Data Literacy to the Streets: Critical Pedagogy in the Public Sphere. *Qualitative Inquiry*, *26*(2), 227–237. https://doi.org/10.1177/1077800419859024

Miller, S. (2016). *Preparing the next generation for the cognitive era*. *36*, 23–25. https://doi.org/10.3233/ISU-160804

Osborne, J. W. (2015). What is rotating in exploratory factor analysis? *Practical Assessment,* *Research and Evaluation*, *20*(2), 1–7.

Phanchalaem, K., Sujiva, S., & Tangdhanakanond, K. (2016). The State of Teachers' Educational Data Use in Thailand. *Procedia - Social and Behavioral Sciences*, *217*, 638–642. https://doi.org/10.1016/j.sbspro.2016.02.084

Piro, J. S., Dunlap, K., & Shutt, T. (2014). A collaborative Data Chat: Teaching summative assessment data use in pre-service teacher education. *Cogent Education*, *1*(1), 1–24. https://doi.org/10.1080/2331186X.2014.968409

Pothier, W. G., & Condon, P. B. (2019). Towards data literacy competencies: Business students, workforce needs, and the role of the librarian. *Journal of Business and Finance Librarianship*. https://doi.org/10.1080/08963568.2019.1680189

Quill, T. M. (2018). Humanitarian Mapping as Library Outreach: A Case for Community-Oriented Mapathons. *Journal of Web Librarianship*, *12*(3), 160–168. https://doi.org/10.1080/19322909.2018.1463585

Ranjan, J. (2005). Journal of Theoretical and Applied Information Technology. *Www.Jatit.Org*, *60*(1), 60–70. Retrieved from http://www.jatit.org/volumes/research-papers/Vol9No1/9Vol9No1.pdf

ReStore. (2011). *The SPSS Logistic Regression Output*. Retrieved from https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod4/12/index.html#:~:text=The Omnibus Tests of Model,model and the new model.

Robinson, L., & Bawden, D. (2017). "The story of data": A socio-technical approach to education for the data librarian role in the CityLIS library school at City, University of London. *Library Management*, *38*(6–7), 312–322. https://doi.org/10.1108/LM-01-2017-0009

Rolf, E., Knutsson, O., & Ramberg, R. (2019). *An analysis of digital competence as expressed in design patterns for technology use in teaching*. *50*(6). https://doi.org/10.1111/bjet.12739

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Schildkamp, K., Poortman, C. L., & Sahlberg, P. (2019). Data-based decision making in developing countries: balancing accountability measures and improvement efforts. *Journal of Professional Capital and Community*, *4*(3), 166–171.

https://doi.org/10.1108/JPCC-07-2019-037

Shreiner, T. L. (2018). Data Literacy for Social Studies: Examining the Role of Data Visualizations in K–12 Textbooks. *Theory and Research in Social Education*, *46*(2), 194–231. https://doi.org/10.1080/00933104.2017.1400483

Shreiner, T. L. (2019). *Students′ use of data visualizations in historical reasoning: A think-aloud investigation with elementary, middle, and high school students*. *43*(4), 389–404. https://doi.org/10.1016/j.jssr.2018.11.001

Slayter, E., & Higgins, L. M. (2018, January 2). Hands-On Learning: A Problem-Based Approach to Teaching Microsoft Excel. *College Teaching*, Vol. 66, pp. 31–33. https://doi.org/10.1080/87567555.2017.1385585

Sorapure, M. (2019). Text, Image, Data, Interaction: Understanding Information Visualization. *Computers and Composition*, *54*, 102519. https://doi.org/10.1016/j.compcom.2019.102519

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., & Pawlik, A. (2015). Data Carpentry: Workshops to Increase Data Literacy for Researchers. *International Journal of Digital Curation*, *10*(1), 135–143. https://doi.org/10.2218/ijdc.v10i1.351

van Geel, M., Keuning, T., Visscher, A., & Fox, J. P. (2017). Changes in educators' data literacy during a data-based decision making intervention. *Teaching and Teacher Education*, *64*, 187–198. https://doi.org/10.1016/j.tate.2017.02.015

Verbakel, E., & Grootveld, M. (2016). 'Essentials 4 Data Support': Five years' experience with data management training. *IFLA Journal*, *42*(4), 278–283. https://doi.org/10.1177/0340035216674027

Wang, B., Wu, C., & Huang, L. (2019). Data literacy for safety professionals in safety management: A theoretical perspective on basic questions and answers. *Safety Science*, *117*, 15–22. https://doi.org/10.1016/j.ssci.2019.04.002

Wilson, M. (2016). Becoming Data and Information Rich in Education. *BU Journal of Graduate Studies in Education*, *8*(1), 5–9. Retrieved from https://eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=EJ1230534

Wolff, A., Kortuem, G., & Cavero, J. (2015). Urban data games: Creating smart citizens for smart cities. *Proceedings - IEEE 15th International Conference on Advanced Learning Technologies: Advanced Technologies for Supporting Open Access to Formal and Informal Learning, ICALT 2015*, 164–165. https://doi.org/10.1109/ICALT.2015.44

Wolff, A., Moore, J., Zdrahal, Z., Hlosta, M., & Kuzilek, J. (2016). Data literacy for learning analytics. *ACM International Conference Proceeding Series*, *25-29-Apri*, 500–501. https://doi.org/10.1145/2883851.2883864

Wolff, A., Wermelinger, M., & Petre, M. (2019). Exploring design principles for data literacy activities to support children's inquiries from complex data. *International Journal of Human Computer Studies, 129*, 41–54. https://doi.org/10.1016/j.ijhcs.2019.03.006

Womack, R. (2015). Data Visualization and Information Literacy. *IASSIST Quarterly, 38*(1), 12. https://doi.org/10.29173/iq619

Annexes

Annexe 1: Excel, SPSS and questionnaire files

Link to Excel files used, including the research in academic platforms, results of the systematic literature review, the templates of the questionnaires used, answers to the questionnaire with data analysis and SPSS files[10]:

https://drive.google.com/drive/folders/1GQreN-badszNvmQI0H2wueAVC4mbyj5M?usp=sharing

[10] If no longer available, please contact the author for the files.

Annexe 2: major topics of the systematic literature review

Table 8: Distribution of topics across articles per classification

Values	**Incidental**	**Main**	**Secondary**	**Grand Total**
Count of Type	42	49	47	138
Count of Identify problems / questions	3	32	15	50
Count of Collect / read data	10	48	40	98
Count of Clean / evaluate data quality	4	36	17	57
Count of Interpretation, analysis and manipulation	11	49	40	100
Count of Actionable information / Insight extraction / data communication	8	46	25	79
Count of General data understanding / analytical reasoning	3	34	18	55
Count of Statistical understanding	3	30	14	47
Count of Tools	4	30	18	52
Count of Contextual applicability	2	40	20	62
Count of Critical reasoning	3	34	12	49
Count of Data visualization	9	27	20	56
Count of Data processing techniques	1	18	12	31
Count of Data authenticity		7	5	12
Count of Data sources	1	15	6	22
Count of Affective		9	8	17
Count of Behavioural	1	5	3	9
Count of Cognitive	3	14	8	25
Count of Data culture	1	9	8	18
Count of Strong organisational culture		5	3	8
Count of Data leadership	2	6	5	13
Count of Vision / awareness	1	9	2	12
Count of Infrastructure		7	5	12
Count of Individual attributes		7	2	9
Count of Libraries	4	9	4	17
Count of Ethical use	1	16	6	23

Table 9: Frequency of topics in Main and secondary results and increase tendency

		Main and secondary results		Main results		Tendency	
Topic	% Total	Quantity	%	Quantity	%	Total to M/S	M/S to main
Identify problems / questions	36,23%	47	48,96%	32	65,31%	12,73%	16,35%
Collect / read data	71,01%	88	91,67%	48	97,96%	20,65%	6,29%
Clean / evaluate data quality	41,30%	53	55,21%	36	73,47%	13,90%	18,26%
Interpretation, analysis and manipulation	72,46%	89	92,71%	49	100,00%	20,24%	7,29%
Actionable information / Insight extraction / data communication	57,25%	71	73,96%	46	93,88%	16,71%	19,92%
General data understanding / analytical reasoning	39,86%	52	54,17%	34	69,39%	14,31%	15,22%
Statistical understanding	34,06%	44	45,83%	30	61,22%	11,78%	15,39%
Tools	37,68%	48	50,00%	30	61,22%	12,32%	11,22%
Contextual applicability	44,93%	60	62,50%	40	81,63%	17,57%	19,13%
Critical reasoning	35,51%	46	47,92%	34	69,39%	12,41%	21,47%
Data visualization	40,58%	47	48,96%	27	55,10%	8,38%	6,14%
Data processing techniques	22,46%	30	31,25%	18	36,73%	8,79%	5,48%
Data authenticity	8,70%	12	12,50%	7	14,29%	3,80%	1,79%
Count of Data sources	15,94%	21	21,88%	15	30,61%	5,93%	8,74%
Count of Affective	12,32%	17	17,71%	9	18,37%	5,39%	0,66%
Count of Behavioural	6,52%	8	8,33%	5	10,20%	1,81%	1,87%
Count of Cognitive	18,12%	22	22,92%	14	28,57%	4,80%	5,65%
Count of Data culture	13,04%	17	17,71%	9	18,37%	4,66%	0,66%
Count of Strong organisational culture	5,80%	8	8,33%	5	10,20%	2,54%	1,87%
Count of Data leadership	9,42%	11	11,46%	6	12,24%	2,04%	0,79%
Count of Vision / awareness	8,70%	11	11,46%	9	18,37%	2,76%	6,91%
Count of Infrastructure	8,70%	12	12,50%	7	14,29%	3,80%	1,79%
Count of Individual attributes	6,52%	9	9,38%	7	14,29%	2,85%	4,91%
Count of Libraries	12,32%	13	13,54%	9	18,37%	1,22%	4,83%
Count of Ethical use	16,67%	22	22,92%	16	32,65%	6,25%	9,74%

Annexe 3: sample results

- AGE:

Table 10: Age sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	0	0,00%	1	0,95%	1	0,48%
Less than 18 years	0	0,00%	0	0,00%	0	0,00%
18 - 25 years	55	52,88%	58	55,24%	113	54,07%
26 - 40 years	37	35,58%	39	37,14%	76	36,36%
41 - 64 years	11	10,58%	7	6,67%	18	8,61%
65+ years	1	0,96%	0	0,00%	1	0,48%
Total:	104	100,00%	105	100,00%	209	100,00%

- GENDER:

Table 11: Gender sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	0	0,00%	1	0,95%	1	0,48%
Female	52	50,00%	44	41,90%	96	45,93%
Male	52	50,00%	60	57,14%	112	53,59%
Total:	104	100,00%	105	100,00%	209	100,00%

- COUNTRY OF BIRTH:

Table 12: Country of birth sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	2	1,92%	5	4,76%	7	3,35%
Angola	1	0,96%	0	0,00%	1	0,48%
Brazil	89	85,58%	73	69,52%	162	77,51%
Colombia	1	0,96%	1	0,95%	2	0,96%
Czech Republic	1	0,96%	0	0,00%	1	0,48%
Germany	1	0,96%	0	0,00%	1	0,48%
Ireland	0	0,00%	1	0,95%	1	0,48%
Kuwait	0	0,00%	1	0,95%	1	0,48%
Netherlands	2	1,92%	1	0,95%	3	1,44%
Philippines	0	0,00%	1	0,95%	1	0,48%
Portugal	7	6,73%	18	17,14%	25	11,96%
Spain	0	0,00%	1	0,95%	1	0,48%
United Kingdom	0	0,00%	2	1,90%	2	0,96%
USA	0	0,00%	1	0,95%	1	0,48%
Total:	104	100,00%	105	100,00%	209	100,00%

- COUNTRY OF RESIDENCE:

Table 13: Country of residence sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	2	1,92%	4	3,81%	6	2,87%
Austria	1	0,96%	0	0,00%	1	0,48%
Brazil	82	78,85%	63	60,00%	145	69,38%
Colombia	0	0,00%	1	0,95%	1	0,48%
Czech Republic	1	0,96%	0	0,00%	1	0,48%
Germany	0	0,00%	2	1,90%	2	0,96%
Kuwait	0	0,00%	1	0,95%	1	0,48%
Netherlands	0	0,00%	1	0,95%	1	0,48%
Peru	1	0,96%	0	0,00%	1	0,48%
Portugal	13	12,50%	26	24,76%	39	18,66%
Spain	0	0,00%	1	0,95%	1	0,48%
United Kingdom	4	3,85%	6	5,71%	10	4,78%
Total:	104	100,00%	105	100,00%	209	100,00%

- EXPATRIATES:

Table 14: Expatriates sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
FALSE	90	86,54%	90	85,71%	180	86,12%
TRUE	14	13,46%	15	14,29%	29	13,88%
Total:	104	100,00%	105	100,00%	209	100,00%
Brazil to Portugal	6	42,86%	6	40,00%	12	41,38%

- EDUCATIONAL LEVEL:

Table 15: Educational level sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	1	0,96%	2	1,90%	3	1,44%
Middle school or equivalent	0	0,00%	0	0,00%	0	0,00%
High school or equivalent	10	9,62%	10	9,52%	20	9,57%
Bachelor's degree / Undergraduate degree or similar / First cycle degree program	58	55,77%	68	64,76%	126	60,29%
Post-Graduate Degree	6	5,77%	5	4,76%	11	5,26%
Master's degree	27	25,96%	18	17,14%	45	21,53%
Doctor's degree	2	1,92%	2	1,90%	4	1,91%
Total:	104	100,00%	105	100,00%	209	100,00%

- FIELD OF STUDY:

Table 16: Field of study sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	1	0,96%	5	4,76%	6	2,87%
Science, technology, engineering or math	17	16,35%	29	27,62%	46	22,01%
Arts and humanities	13	12,50%	20	19,05%	33	15,79%
Social and behavioural sciences	56	53,85%	40	38,10%	96	45,93%
Not applicable / Other	17	16,35%	11	10,48%	28	13,40%
Total:	104	100,00%	105	100,00%	209	100,00%

- WORK STATUS:

Table 17: Work status sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	1	0,96%	2	1,90%	3	1,44%
Student	39	37,50%	32	30,48%	71	33,97%
Unemployed (looking for work)	8	7,69%	9	8,57%	17	8,13%
Unemployed (not looking for work)	2	1,92%	4	3,81%	6	2,87%
Employee	37	35,58%	43	40,95%	80	38,28%
Self-employed	14	13,46%	11	10,48%	25	11,96%
Business owner	2	1,92%	4	3,81%	6	2,87%
Retired	1	0,96%	0	0,00%	1	0,48%
Total:	104	100,00%	105	100,00%	209	100,00%

- INCOME

Table 18: Income sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	10	9,62%	22	20,95%	32	15,31%
No income	13	12,50%	17	16,19%	30	14,35%
Below the minimum wage or equivalent	7	6,73%	2	1,90%	9	4,31%
Around the minimum wage or equivalent	12	11,54%	18	17,14%	30	14,35%
Up to 2.5 times the minimum wage or equivalent	18	17,31%	19	18,10%	37	17,70%
2.5 to 5 times the minimum wage or equivalent	18	17,31%	10	9,52%	28	13,40%
5 to 10 times the minimum wage or equivalent	12	11,54%	8	7,62%	20	9,57%
Over 10 times the minimum wage or equivalent	14	13,46%	9	8,57%	23	11,00%
Total:	104	100,00%	105	100,00%	209	100,00%

- TYPE OF FUNCTION AT WORK:

Table 19: Type of function sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	4	3,85%	11	10,48%	15	7,18%
None	22	21,15%	18	17,14%	40	19,14%
Voluntary	1	0,96%	4	3,81%	5	2,39%
Intern	15	14,42%	13	12,38%	28	13,40%
Operational	6	5,77%	9	8,57%	15	7,18%
Administrative	3	2,88%	4	3,81%	7	3,35%
Analyst/Specialist	21	20,19%	16	15,24%	37	17,70%
Professor	7	6,73%	8	7,62%	15	7,18%
Management/Direction	9	8,65%	6	5,71%	15	7,18%
Self-employed	12	11,54%	14	13,33%	26	12,44%
Business owner	4	3,85%	2	1,90%	6	2,87%
Total:	104	100,00%	105	100,00%	209	100,00%

- EXPERIENCE:

Table 20: Work experience sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	3	2,88%	5	4,76%	8	3,83%
No experience	3	2,88%	5	4,76%	8	3,83%
Less than 1 year	9	8,65%	7	6,67%	16	7,66%
1 to 3 years	38	36,54%	38	36,19%	76	36,36%
4 to 7 years	30	28,85%	29	27,62%	59	28,23%
8 to 10 years	5	4,81%	11	10,48%	16	7,66%
11 to 20 years	10	9,62%	4	3,81%	14	6,70%
Over 20 years	6	5,77%	6	5,71%	12	5,74%
Total:	104	100,00%	105	100,00%	209	100,00%

- NUMBER OF EMPLOYEES AT ORGANIZATION:

Table 21: Number of employees at the organisation sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	5	4,81%	11	10,48%	16	7,66%
1 to 10	18	17,31%	22	20,95%	40	19,14%
11 to 50	15	14,42%	17	16,19%	32	15,31%
51 to 100	7	6,73%	6	5,71%	13	6,22%
101 to 300	4	3,85%	8	7,62%	12	5,74%
301 to 500	2	1,92%	2	1,90%	4	1,91%
Over 500	22	21,15%	15	14,29%	37	17,70%
Not applicable	31	29,81%	24	22,86%	55	26,32%
Total:	104	100,00%	105	100,00%	209	100,00%

- FIELD OF WORK:

Table 22: Field of work sample

GROUPS	TYPE 1		TYPE 2		TOTAL	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Prefer not to disclose	4	3,15%	7	4,76%	11	4,01%
Administrative	13	10,24%	10	6,80%	23	8,39%
Finance/acounting	6	4,72%	11	7,48%	17	6,20%
Legal	53	41,73%	39	26,53%	92	33,58%
Human Resources	6	4,72%	8	5,44%	14	5,11%
Marketing/strategy/ advertisement	0	0,00%	0	0,00%	0	0,00%
Sales/commercial	2	1,57%	10	6,80%	12	4,38%
Arts/entertainment	2	1,57%	11	7,48%	13	4,74%
Design/media	1	0,79%	6	4,08%	7	2,55%
Architecture/ engineering/ construction	0	0,00%	0	0,00%	0	0,00%
Technology	9	7,09%	14	9,52%	23	8,39%
Education	10	7,87%	14	9,52%	24	8,76%
Safety or security	3	2,36%	2	1,36%	5	1,82%
Travel, tourism or hotels	5	3,94%	3	2,04%	8	2,92%
Food industry	0	0,00%	5	3,40%	5	1,82%
Transportation/ logistics	0	0,00%	0	0,00%	0	0,00%
Health	9	7,09%	6	4,08%	15	5,47%
Beauty and aesthetic	1	0,79%	0	0,00%	1	0,36%
Other	3	2,36%	1	0,68%	4	1,46%
Total:	127	100,00%	147	100,00%	274	100,00%

- QUIZ SCORES:

Table 23: Percentage of right answers on the quiz

	Correct Answers		
Question	Type 1	Type 2	Total
Q1	96,15%	94,29%	95,22%
Q2	72,12%	74,29%	73,21%
Q3	53,85%	40,95%	47,37%
Q4	23,08%	20,95%	22,01%
Q5	85,58%	86,67%	86,12%
Q6	81,73%	75,24%	78,47%
Q7	59,62%	43,81%	51,67%
Q8	74,04%	76,19%	75,12%
Q9	86,54%	84,76%	85,65%
Q10	50,96%	42,86%	46,89%
Q11	63,46%	64,76%	64,11%
Q12	63,46%	68,57%	66,03%
Q13	25,96%	26,67%	26,32%
Q14	56,73%	64,76%	60,77%
Q15	85,58%	82,86%	84,21%

Annexe 4: descriptive statistics regarding central tendency

Descriptive statistics found in each question of the questionnaire with a brief description of what each question was about.

I.e. Q1U is the first question from Understanding data and it had a mean of 5,29 on Type 1. This question is about understanding how one produces data.

- UNDERSTANDING DATA

Q1U – I am well aware of how I produce data on electronic devices and how data is very present in my daily life in different ways

Q2U – Finding data/information to solve a certain problem is one of my strong skills

Q3U – I am capable of reading charts and understand the information presented

Q4U – When I read data about something, I can have a good idea of how they are connected to what is happening

Q5U – I feel I can interpret well the meaning of numbers used to explain something

Table 24: Averages for the variable "Understanding Data"

UNDERSTANDING DATA		Quiz Score	MEAN	Q1U	Q2U	Q3U	Q4U	Q5U
Type 1	Mean	9,79	5,32	5,29	4,76	5,69	5,56	5,30
	Median	10,00	5,80	6,00	5,00	6,00	6,00	6,00
	Mode	11,00	6,00	6,00	6,00	6,00	6,00	6,00
	St. Dev.	2,28	1,37	1,58	1,69	1,21	1,03	1,36
	Skewness	-0,79	-0,89	-0,90	-0,71	-1,27	-0,54	-1,03
	Kurtosis	0,54	0,56	-0,27	-0,38	2,31	0,07	1,05
Type 2	Mean	9,48	5,04	5,45	4,42	5,05	5,18	5,12
	Median	10,00	5,20	6,00	5,00	5,00	5,00	5,00
	Mode	11,00	6,00	6,00	6,00	6,00	6,00	6,00
	St. Dev.	2,65	1,42	1,52	1,65	1,34	1,36	1,25
	Skewness	-0,53	-0,80	-1,06	-0,38	-0,77	-1,05	-0,73
	Kurtosis	-0,30	0,18	0,50	-0,72	0,19	0,96	-0,03
Total	Mean	9,63	5,18	5,37	4,59	5,37	5,37	5,21
	Median	10,00	5,60	6,00	5,00	6,00	6,00	5,00
	Mode	11,00	6,00	6,00	6,00	6,00	6,00	6,00
	St. Dev.	2,47	1,41	1,55	1,67	1,31	1,22	1,31
	Skewness	-0,65	-0,87	-0,97	-0,53	-0,96	-1,01	-0,87
	Kurtosis	0,04	0,41	0,06	-0,62	0,81	1,28	0,52
			Cohen's d					
			0,20	-0,10	0,20	0,51	0,31	0,13

- DATA MANIPULATION AND TECHNICAL ANALYSIS

Q1M – I know how to use programs, applications and other tools to transform data into information

Q2M – I have good skills with statistics

Q3M – I can put data together and come to different possible conclusions

Q4M – When I see several data regarding something, I can have an idea which are the most and least important

Q5M – I tend to see patterns in data or wonder if things are connected

Table 25: Averages for the variable "Data Manipulation and Technical Analysis"

Manipul. & Technical		Quiz Score	MEAN	Q1M	Q2M	Q3M	Q4M	Q5M
Type 1	Mean	9,79	4,42	3,64	3,84	4,83	4,82	4,95
	Median	10,00	4,60	4,00	4,00	5,00	5,00	5,00
	Mode	11,00	5,00	5,00	4,00	5,00	5,00	6,00
	St. Dev.	2,28	1,60	1,70	1,68	1,47	1,49	1,65
	Skewness	-0,79	-0,42	-0,03	-0,07	-0,70	-0,53	-0,77
	Kurtosis	0,54	-0,37	-1,03	-0,67	0,09	-0,06	-0,17
Type 2	Mean	9,48	4,33	3,91	3,77	4,59	4,53	4,83
	Median	10,00	4,60	4,00	4,00	5,00	5,00	5,00
	Mode	11,00	5,20	5,00	5,00	5,00	5,00	6,00
	St. Dev.	2,65	1,62	1,90	1,69	1,57	1,40	1,56
	Skewness	-0,53	-0,45	-0,05	0,01	-0,76	-0,74	-0,70
	Kurtosis	-0,30	-0,55	-1,17	-0,94	-0,16	-0,17	-0,32
Total	Mean	9,63	4,37	3,78	3,80	4,71	4,67	4,89
	Median	10,00	4,60	4,00	4,00	5,00	5,00	5,00
	Mode	11,00	5,20	5,00	5,00	5,00	5,00	6,00
	St. Dev.	2,47	1,61	1,80	1,68	1,52	1,45	1,60
	Skewness	-0,65	-0,42	-0,01	-0,03	-0,74	-0,60	-0,72
	Kurtosis	0,04	-0,46	-1,09	-0,82	-0,03	-0,10	-0,26
			Cohen's d					
			0,05	-0,15	0,04	0,16	0,20	0,08

- SOFTWARE KNOWLEDGE

Q1S – Spreadsheet software (Microsoft Excel and similar)

Q2S – Statistical software (SPSS, SAS etc)

Q3S – Business Intelligence and data visualization (Tableau, PowerBI etc.)

Q4S – Query Languages (SQL, MariaDB etc.)

Q5S – Programming languages optimized for data analysis (R, Python etc)

Table 26: Averages for the variable "Software Knowledge"

SOFTWARE		Quiz Score	MEAN	Q1S	Q2S	Q3S	Q4S	Q5S
Type 1	Mean	9,79	2,11	3,88	1,96	1,55	1,53	1,62
	Median	10,00	1,60	4,00	1,00	1,00	1,00	1,00
	Mode	11,00	1,40	3,00	1,00	1,00	1,00	1,00
	St. Dev.	2,28	1,35	1,60	1,54	1,16	1,14	1,29
	Skewness	-0,79	1,85	0,30	1,50	2,53	2,57	2,37
	Kurtosis	0,54	3,72	-0,73	1,04	6,39	6,79	5,10
Type 2	Mean	9,48	2,43	4,18	2,12	1,86	1,97	2,02
	Median	10,00	1,60	4,00	1,00	1,00	1,00	1,00
	Mode	11,00	1,60	4,00	1,00	1,00	1,00	1,00
	St. Dev.	2,65	1,63	1,65	1,64	1,57	1,71	1,58
	Skewness	-0,53	1,27	-0,15	1,29	1,85	1,74	1,63
	Kurtosis	-0,30	1,10	-0,85	0,42	2,29	1,88	1,75
Total	Mean	9,63	2,27	4,03	2,04	1,70	1,75	1,82
	Median	10,00	1,60	4,00	1,00	1,00	1,00	1,00
	Mode	11,00	1,40	3,00	1,00	1,00	1,00	1,00
	St. Dev.	2,47	1,50	1,63	1,59	1,38	1,47	1,46
	Skewness	-0,65	1,53	0,07	1,38	2,15	2,12	1,93
	Kurtosis	0,04	2,05	-0,87	0,66	3,83	3,73	2,92
			Cohen's d					
			-0,22	-0,19	-0,10	-0,23	-0,31	-0,28

- ACTING ON DATA

Q1AC – When I have some of the information needed, I have several ideas of actions that could be done to improve the situation

Q2AC – When I have the correct information about a situation, I know how to make a decision about it

Q3AC – I am capable of distinguishing how each decision may have different effects

Q4AC – In general, I normally understand well how things are related and how one change in a place can affect others

Q5AC – With the right data, I find it easy to explain to someone the reason why a certain action should be chosen

Table 27: Averages for the variable "Acting on Data"

ACTING ON DATA		Quiz Score	MEAN	Q1AC	Q2AC	Q3AC	Q4AC	Q5AC
Type 1	Mean	9,79	5,46	5,07	5,49	5,47	5,49	5,77
	Median	10,00	5,80	5,00	6,00	6,00	6,00	6,00
	Mode	11,00	6,00	6,00	6,00	6,00	6,00	6,00
	St. Dev.	2,28	1,24	1,37	1,21	1,23	1,17	1,21
	Skewness	-0,79	-0,81	-0,79	-0,62	-1,00	-0,78	-0,86
	Kurtosis	0,54	0,23	0,35	-0,57	1,10	0,44	-0,18
Type 2	Mean	9,48	5,31	4,98	5,49	5,29	5,35	5,44
	Median	10,00	5,60	5,00	6,00	5,00	6,00	6,00
	Mode	11,00	5,80	5,00	6,00	6,00	6,00	6,00
	St. Dev.	2,65	1,18	1,25	1,23	1,19	1,06	1,18
	Skewness	-0,53	-0,68	-0,48	-0,75	-0,64	-0,80	-0,73
	Kurtosis	-0,30	0,02	-0,27	0,00	0,04	0,03	0,31
Total	Mean	9,63	5,38	5,02	5,49	5,38	5,42	5,60
	Median	10,00	5,80	5,00	6,00	6,00	6,00	6,00
	Mode	11,00	6,00	6,00	6,00	6,00	6,00	6,00
	St. Dev.	2,47	1,21	1,31	1,22	1,21	1,12	1,21
	Skewness	-0,65	-0,73	-0,64	-0,68	-0,81	-0,76	-0,76
	Kurtosis	0,04	0,10	0,06	-0,30	0,49	0,25	-0,02
			Cohen's d					
			0,12	0,07	0,00	0,15	0,12	0,28

- CRITICAL REASONING

Q1C – I often question the information that I receive if I feel it may not tell the whole truth

Q2C – When someone tells me the cause of something, I tend to wonder what other possible causes may be involved

Q3C – In general, I consider I have a good critical thinking

Q4C – I often wonder how different a situation would be if certain things changed

Q5C – I usually organize information to filter what is actually important

Table 28: Averages for the variable Critical Reasoning

CRITICAL REASONING		Quiz Score	MEAN	Q1C	Q2C	Q3C	Q4C	Q5C
Type 1	Mean	9,79	5,96	6,11	5,91	5,92	6,01	5,83
	Median	10,00	6,00	6,00	6,00	6,00	6,00	6,00
	Mode	11,00	6,80	7,00	7,00	6,00	7,00	7,00
	St. Dev.	2,28	1,07	0,97	1,18	1,01	1,08	1,09
	Skewness	-0,79	-1,16	-1,37	-1,51	-1,45	-0,95	-0,51
	Kurtosis	0,54	1,95	2,69	3,04	4,54	0,17	-0,68
Type 2	Mean	9,48	5,79	5,90	5,79	5,77	5,90	5,60
	Median	10,00	6,00	6,00	6,00	6,00	6,00	6,00
	Mode	11,00	6,20	7,00	6,00	6,00	6,00	6,00
	St. Dev.	2,65	1,21	1,30	1,19	1,16	1,14	1,27
	Skewness	-0,53	-1,44	-1,68	-1,43	-1,30	-1,44	-1,34
	Kurtosis	-0,30	2,49	3,08	2,95	1,97	2,48	1,95
Total	Mean	9,63	5,87	6,00	5,85	5,85	5,96	5,71
	Median	10,00	6,00	6,00	6,00	6,00	6,00	6,00
	Mode	11,00	6,40	7,00	6,00	6,00	7,00	6,00
	St. Dev.	2,47	1,14	1,15	1,19	1,09	1,11	1,19
	Skewness	-0,65	-1,35	-1,68	-1,46	-1,38	-1,21	-1,05
	Kurtosis	0,04	2,42	3,57	2,87	2,94	1,45	1,28
			Cohen's d					
			0,14	0,18	0,10	0,14	0,09	0,19

- ATTITUDE

Q1AT – I feel confident in my abilities handling data in general

Q2AT – I am interested in the studies of data

Q3AT – I believe my current knowledge of data is good

Q4AT – I usually bear in mind how data works in my everyday activities that may involve it

Q5AT – When using the internet or electronic devices, I make decisions taking into account how my data is used

Table 29: Averages for the variable “Attitude”

ATTITUDE		Quiz Score	MEAN	Q1AT	Q2AT	Q3AT	Q4AT	Q5AT
Type 1	Mean	9,79	4,47	4,56	4,70	3,76	4,79	4,56
	Median	10,00	4,80	5,00	5,00	4,00	5,00	5,00
	Mode	11,00	5,20	5,00	6,00	5,00	5,00	5,00
	St. Dev.	2,28	1,68	1,42	1,76	1,66	1,79	1,77
	Skewness	-0,79	-0,46	-0,50	-0,45	-0,09	-0,67	-0,59
	Kurtosis	0,54	-0,58	-0,27	-0,75	-0,90	-0,49	-0,51
Type 2	Mean	9,48	4,38	4,56	4,58	3,76	4,37	4,61
	Median	10,00	4,80	5,00	5,00	4,00	5,00	5,00
	Mode	11,00	4,80	6,00	5,00	2,00	5,00	6,00
	St. Dev.	2,65	1,65	1,55	1,62	1,68	1,66	1,75
	Skewness	-0,53	-0,38	-0,62	-0,26	0,09	-0,43	-0,66
	Kurtosis	-0,30	-0,74	-0,51	-0,75	-1,06	-0,77	-0,63
Total	Mean	9,63	4,42	4,56	4,64	3,76	4,58	4,58
	Median	10,00	4,80	5,00	5,00	4,00	5,00	5,00
	Mode	11,00	5,40	5,00	6,00	5,00	5,00	6,00
	St. Dev.	2,47	1,67	1,48	1,69	1,66	1,74	1,76
	Skewness	-0,65	-0,41	-0,57	-0,35	0,00	-0,52	-0,62
	Kurtosis	0,04	-0,68	-0,41	-0,76	-0,99	-0,68	-0,58
			Cohen's d					
			0,06	0,00	0,07	0,00	0,24	-0,03

- ENVIRONMENTAL FACTORS

Q1E – My professional environments have a strong culture of using data for work. Me and my colleagues often take decisions based on it

Q2E – The managers, leading people or professors base a lot of their decisions on data and positively influence me to do the same

Q3E – My professional environments have infrastructure (hardware/machines and software/programs/applications) that allow me to handle data if needed (those can be computers, programs that handle information about the company, Microsoft Excel, SPSS, SAS, Tableau, PowerBI, Hadoop or others).

Q4E – I feel my professional environments make me want to know more about how data is used.

Q5E – I believe my professional path so far has stimulated the study of data

Table 30: Averages for the variable "Environmental Factors"

ENVIRONMENT		Quiz Score	MEAN	Q1E	Q2E	Q3E	Q4E	Q5E
Type 1	Mean	9,79	4,56	4,77	4,80	4,54	4,29	4,40
	Median	10,00	5,00	5,00	5,00	5,00	5,00	5,00
	Mode	11,00	5,80	6,00	6,00	6,00	5,00	6,00
	St. Dev.	2,28	1,93	1,99	1,81	1,91	1,99	1,94
	Skewness	-0,79	-0,49	-0,72	-0,68	-0,44	-0,24	-0,39
	Kurtosis	0,54	-0,88	-0,68	-0,54	-0,93	-1,21	-1,05
Type 2	Mean	9,48	4,60	4,78	4,70	4,50	4,54	4,48
	Median	10,00	5,00	5,00	5,00	5,00	5,00	5,00
	Mode	11,00	5,00	6,00	4,00	5,00	5,00	5,00
	St. Dev.	2,65	1,87	1,85	1,80	1,90	1,94	1,87
	Skewness	-0,53	-0,45	-0,55	-0,50	-0,41	-0,44	-0,35
	Kurtosis	-0,30	-0,79	-0,73	-0,54	-0,89	-0,93	-0,87
Total	Mean	9,63	4,58	4,78	4,75	4,52	4,42	4,44
	Median	10,00	5,00	5,00	5,00	5,00	5,00	5,00
	Mode	11,00	5,40	6,00	6,00	5,00	5,00	5,00
	St. Dev.	2,47	1,90	1,91	1,80	1,90	1,97	1,90
	Skewness	-0,65	-0,47	-0,64	-0,58	-0,42	-0,33	-0,37
	Kurtosis	0,04	-0,85	-0,70	-0,57	-0,92	-1,09	-0,96
			Cohen's d					
			-0,02	-0,01	0,05	0,02	-0,13	-0,04

- QUIZ RESULTS

Table 31: Averages for the quiz score

		Quiz Score	Cohen's d
Type 1	Mean	9,79	0,13
	Median	10,00	
	Mode	11,00	
	St. Dev.	2,28	
	Skewness	-0,79	
	Kurtosis	0,54	
Type 2	Mean	9,48	
	Median	10,00	
	Mode	11,00	
	St. Dev.	2,65	
	Skewness	-0,53	
	Kurtosis	-0,30	
Total	Mean	9,63	
	Median	10,00	
	Mode	11,00	
	St. Dev.	2,47	
	Skewness	-0,65	
	Kurtosis	0,04	

Annexe 5: factor analysis – principal component analysis

Table 32: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,892
Bartlett's Test of Sphericity	Approx. Chi-Square	4689,395
	df	595
	Sig.	,000

Table 33: total variance explained with Eigenvalues above 1

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	11,536	32,960	32,960	11,536	32,960	32,960	8,260
2	4,028	11,509	44,469	4,028	11,509	44,469	4,435
3	2,344	6,699	51,168	2,344	6,699	51,168	5,296
4	1,967	5,619	56,787	1,967	5,619	56,787	5,018
5	1,683	4,810	61,596	1,683	4,810	61,596	6,467
6	1,229	3,511	65,107	1,229	3,511	65,107	6,283
7	1,073	3,064	68,172	1,073	3,064	68,172	1,585

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

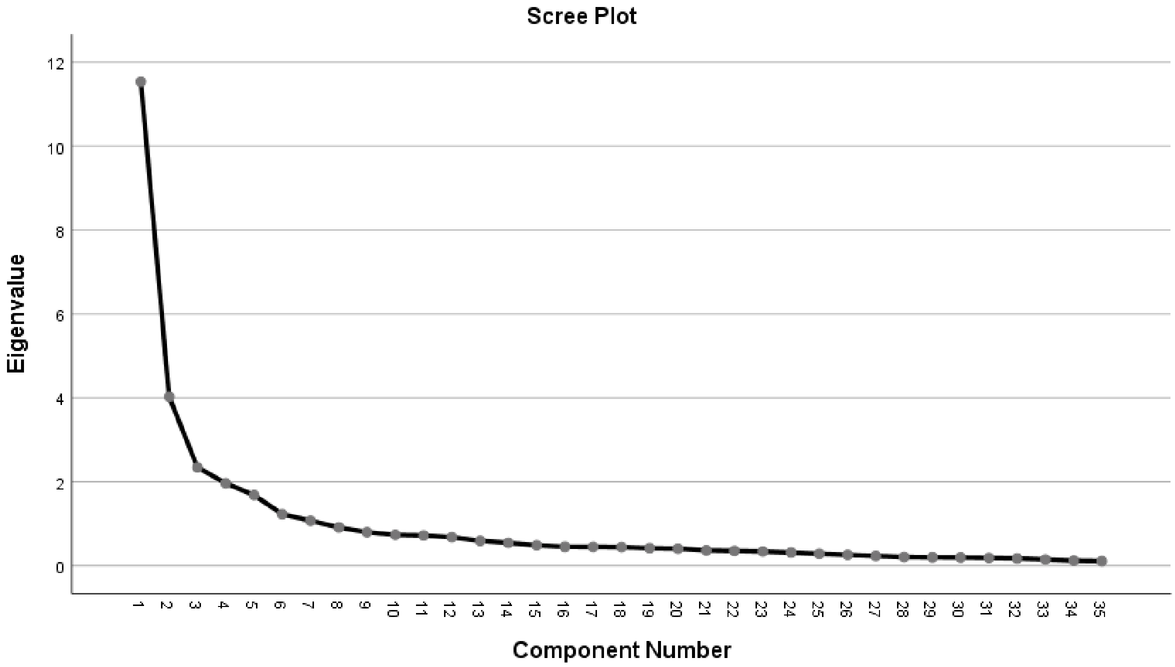


Figure 10: Scree plot depicting eigenvalues for each component

Table 34: Pattern Matrix

Pattern Matrix[a]

	Component						
	1	2	3	4	5	6	7
Q1U	0,096	-0,033	0,061	0,125	-0,054	0,507	0,378
Q2U	0,335	-0,015	0,185	0,001	-0,048	0,430	-0,084
Q3U	0,037	-0,146	-0,021	-0,020	-0,044	0,699	-0,207
Q4U	-0,015	-0,026	0,071	-0,052	-0,028	0,857	0,063
Q5U	-0,011	-0,121	0,034	0,031	-0,155	0,747	0,069
Q1M	0,403	0,091	0,078	0,290	0,010	0,188	-0,356
Q2M	0,284	0,037	-0,003	0,242	-0,113	0,352	-0,420
Q3M	0,571	0,023	0,064	-0,073	-0,081	0,330	-0,219
Q4M	0,544	0,081	-0,035	-0,036	-0,180	0,331	-0,060
Q5M	0,576	0,060	-0,064	-0,008	-0,199	0,249	-0,055
Q1S	0,111	0,120	0,156	0,540	0,044	0,086	-0,293
Q2S	0,033	0,194	0,018	0,645	0,002	-0,036	-0,339
Q3S	-0,015	-0,001	0,003	0,864	-0,073	-0,012	-0,019
Q4S	-0,014	-0,081	-0,008	0,910	-0,079	-0,012	0,129
Q5S	0,034	-0,028	0,003	0,888	0,016	-0,015	0,165
Q1AC	0,216	-0,019	0,060	0,080	-0,649	-0,123	0,064
Q2AC	-0,066	0,053	0,059	0,050	-0,873	0,015	-0,007
Q3AC	-0,005	-0,008	0,025	-0,065	-0,847	-0,016	-0,005
Q4AC	-0,052	-0,039	-0,116	0,077	-0,873	0,041	-0,001
Q5AC	-0,033	-0,131	0,052	-0,053	-0,636	0,164	-0,023
Q1C	0,066	-0,741	-0,049	-0,015	0,013	0,161	0,128
Q2C	0,068	-0,795	-0,027	-0,009	-0,026	0,064	0,026
Q3C	0,018	-0,766	-0,013	0,036	-0,020	0,140	-0,069
Q4C	0,043	-0,788	0,046	-0,052	-0,034	-0,126	0,059
Q5C	-0,023	-0,585	0,093	0,018	-0,165	-0,062	-0,389
Q1AT	0,535	-0,231	0,082	0,045	-0,051	-0,005	-0,206
Q2AT	0,711	-0,152	-0,026	0,122	0,043	0,069	0,153
Q3AT	0,714	-0,082	0,045	0,154	0,072	0,021	-0,144
Q4AT	0,824	-0,016	0,088	0,002	-0,040	0,013	0,105
Q5AT	0,649	-0,059	0,090	-0,009	-0,095	-0,140	0,082
Q1E	0,069	-0,001	0,842	-0,080	-0,105	-0,031	-0,078
Q2E	-0,035	0,026	0,841	-0,160	-0,032	0,041	-0,115
Q3E	-0,189	-0,134	0,791	0,229	0,096	0,030	-0,037
Q4E	0,135	0,042	0,776	0,090	-0,065	0,002	0,219
Q5E	0,277	0,092	0,688	0,068	-0,011	0,060	0,226

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.[a]
a. Rotation converged in 13 iterations.

Table 35: Structure matrix

Pattern Matrix[a]

	Component						
	1	2	3	4	5	6	7
Q1U	0,096	-0,033	0,061	0,125	-0,054	0,507	0,378
Q2U	0,335	-0,015	0,185	0,001	-0,048	0,430	-0,084
Q3U	0,037	-0,146	-0,021	-0,020	-0,044	0,699	-0,207
Q4U	-0,015	-0,026	0,071	-0,052	-0,028	0,857	0,063
Q5U	-0,011	-0,121	0,034	0,031	-0,155	0,747	0,069
Q1M	0,403	0,091	0,078	0,290	0,010	0,188	-0,356
Q2M	0,284	0,037	-0,003	0,242	-0,113	0,352	-0,420
Q3M	0,571	0,023	0,064	-0,073	-0,081	0,330	-0,219
Q4M	0,544	0,081	-0,035	-0,036	-0,180	0,331	-0,060
Q5M	0,576	0,060	-0,064	-0,008	-0,199	0,249	-0,055
Q1S	0,111	0,120	0,156	0,540	0,044	0,086	-0,293
Q2S	0,033	0,194	0,018	0,645	0,002	-0,036	-0,339
Q3S	-0,015	-0,001	0,003	0,864	-0,073	-0,012	-0,019
Q4S	-0,014	-0,081	-0,008	0,910	-0,079	-0,012	0,129
Q5S	0,034	-0,028	0,003	0,888	0,016	-0,015	0,165
Q1AC	0,216	-0,019	0,060	0,080	-0,649	-0,123	0,064
Q2AC	-0,066	0,053	0,059	0,050	-0,873	0,015	-0,007
Q3AC	-0,005	-0,008	0,025	-0,065	-0,847	-0,016	-0,005
Q4AC	-0,052	-0,039	-0,116	0,077	-0,873	0,041	-0,001
Q5AC	-0,033	-0,131	0,052	-0,053	-0,636	0,164	-0,023
Q1C	0,066	-0,741	-0,049	-0,015	0,013	0,161	0,128
Q2C	0,068	-0,795	-0,027	-0,009	-0,026	0,064	0,026
Q3C	0,018	-0,766	-0,013	0,036	-0,020	0,140	-0,069
Q4C	0,043	-0,788	0,046	-0,052	-0,034	-0,126	0,059
Q5C	-0,023	-0,585	0,093	0,018	-0,165	-0,062	-0,389
Q1AT	0,535	-0,231	0,082	0,045	-0,051	-0,005	-0,206
Q2AT	0,711	-0,152	-0,026	0,122	0,043	0,069	0,153
Q3AT	0,714	-0,082	0,045	0,154	0,072	0,021	-0,144
Q4AT	0,824	-0,016	0,088	0,002	-0,040	0,013	0,105
Q5AT	0,649	-0,059	0,090	-0,009	-0,095	-0,140	0,082
Q1E	0,069	-0,001	0,842	-0,080	-0,105	-0,031	-0,078
Q2E	-0,035	0,026	0,841	-0,160	-0,032	0,041	-0,115
Q3E	-0,189	-0,134	0,791	0,229	0,096	0,030	-0,037
Q4E	0,135	0,042	0,776	0,090	-0,065	0,002	0,219
Q5E	0,277	0,092	0,688	0,068	-0,011	0,060	0,226

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.[a]
a. Rotation converged in 13 iterations.

Annexe 6: normality test

Table 36: Normality tests with both Kolmogorov-Smirnov and Shapiro-Wilk tests

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Score	0,138	209	0,000	0,952	209	0,000
Q1U	0,247	209	0,000	0,853	209	0,000
Q2U	0,186	209	0,000	0,919	209	0,000
Q3U	0,211	209	0,000	0,877	209	0,000
Q4U	0,214	209	0,000	0,876	209	0,000
Q5U	0,215	209	0,000	0,893	209	0,000
Q1M	0,167	209	0,000	0,930	209	0,000
Q2M	0,135	209	0,000	0,944	209	0,000
Q3M	0,241	209	0,000	0,901	209	0,000
Q4M	0,220	209	0,000	0,922	209	0,000
Q5M	0,216	209	0,000	0,901	209	0,000
Q1S	0,148	209	0,000	0,946	209	0,000
Q2S	0,347	209	0,000	0,696	209	0,000
Q3S	0,398	209	0,000	0,576	209	0,000
Q4S	0,404	209	0,000	0,583	209	0,000
Q5S	0,373	209	0,000	0,628	209	0,000
Q1AC	0,188	209	0,000	0,916	209	0,000
Q2AC	0,247	209	0,000	0,881	209	0,000
Q3AC	0,232	209	0,000	0,892	209	0,000
Q4AC	0,253	209	0,000	0,881	209	0,000
Q5AC	0,237	209	0,000	0,879	209	0,000
Q1C	0,259	209	0,000	0,775	209	0,000
Q2C	0,244	209	0,000	0,814	209	0,000
Q3C	0,254	209	0,000	0,823	209	0,000
Q4C	0,248	209	0,000	0,819	209	0,000
Q5C	0,232	209	0,000	0,860	209	0,000
Q1AT	0,215	209	0,000	0,918	209	0,000
Q2AT	0,153	209	0,000	0,932	209	0,000
Q3AT	0,145	209	0,000	0,939	209	0,000
Q4AT	0,203	209	0,000	0,915	209	0,000
Q5AT	0,192	209	0,000	0,904	209	0,000
Q1E	0,189	209	0,000	0,887	209	0,000
Q2E	0,172	209	0,000	0,906	209	0,000
Q3E	0,179	209	0,000	0,912	209	0,000
Q4E	0,162	209	0,000	0,910	209	0,000
Q5E	0,157	209	0,000	0,916	209	0,000

a. Lilliefors Significance Correction

Annexe 7: correlations

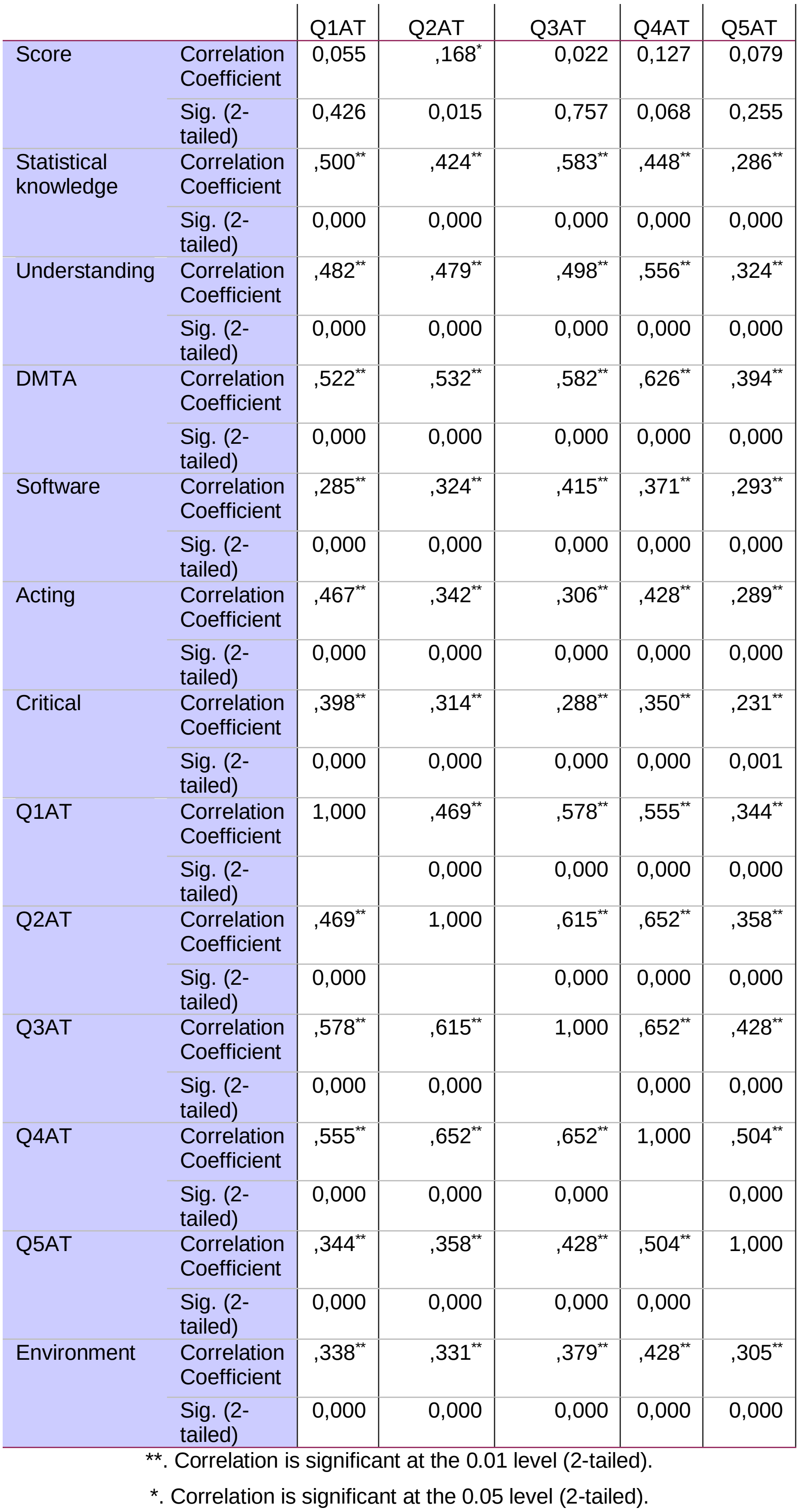
Table 37: Spearman correlation of score and Likert scale variables

Correlations

		Score	Statistical knowledge	Understanding	DMTA	Software	Acting	Critical	Attitude	Environment
Score	Correlation Coefficient	1,000	,184**	,224**	,281**	0,105	0,103	0,123	0,132	0,057
	Sig. (2-tailed)		0,008	0,001	0,000	0,131	0,138	0,075	0,056	0,415
Statistical knowledge	Correlation Coefficient	,184**	1,000	,613**	,701**	,530**	,437**	,211**	,542**	,302**
	Sig. (2-tailed)	0,008		0,000	0,000	0,000	0,000	0,002	0,000	0,000
Understanding	Correlation Coefficient	,224**	,613**	1,000	,684**	,280**	,483**	,344**	,590**	,358**
	Sig. (2-tailed)	0,001	0,000		0,000	0,000	0,000	0,000	0,000	0,000
DMTA	Correlation Coefficient	,281**	,701**	,684**	1,000	,446**	,478**	,285**	,648**	,387**
	Sig. (2-tailed)	0,000	0,000	0,000		0,000	0,000	0,000	0,000	0,000
Software	Correlation Coefficient	0,105	,530**	,280**	,446**	1,000	,219**	-0,004	,407**	,393**
	Sig. (2-tailed)	0,131	0,000	0,000	0,000		0,001	0,955	0,000	0,000
Acting	Correlation Coefficient	0,103	,437**	,483**	,478**	,219**	1,000	,410**	,451**	,333**
	Sig. (2-tailed)	0,138	0,000	0,000	0,000	0,001		0,000	0,000	0,000
Critical	Correlation Coefficient	0,123	,211**	,344**	,285**	-0,004	,410**	1,000	,392**	,200**
	Sig. (2-tailed)	0,075	0,002	0,000	0,000	0,955	0,000		0,000	0,004
Attitude	Correlation Coefficient	0,132	,542**	,590**	,648**	,407**	,451**	,392**	1,000	,431**
	Sig. (2-tailed)	0,056	0,000	0,000	0,000	0,000	0,000	0,000		0,000
Environment	Correlation Coefficient	0,057	,302**	,358**	,387**	,393**	,333**	,200**	,431**	1,000
	Sig. (2-tailed)	0,415	0,000	0,000	0,000	0,000	0,000	0,004	0,000	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 38: Spearman correlation with separate attitude questions



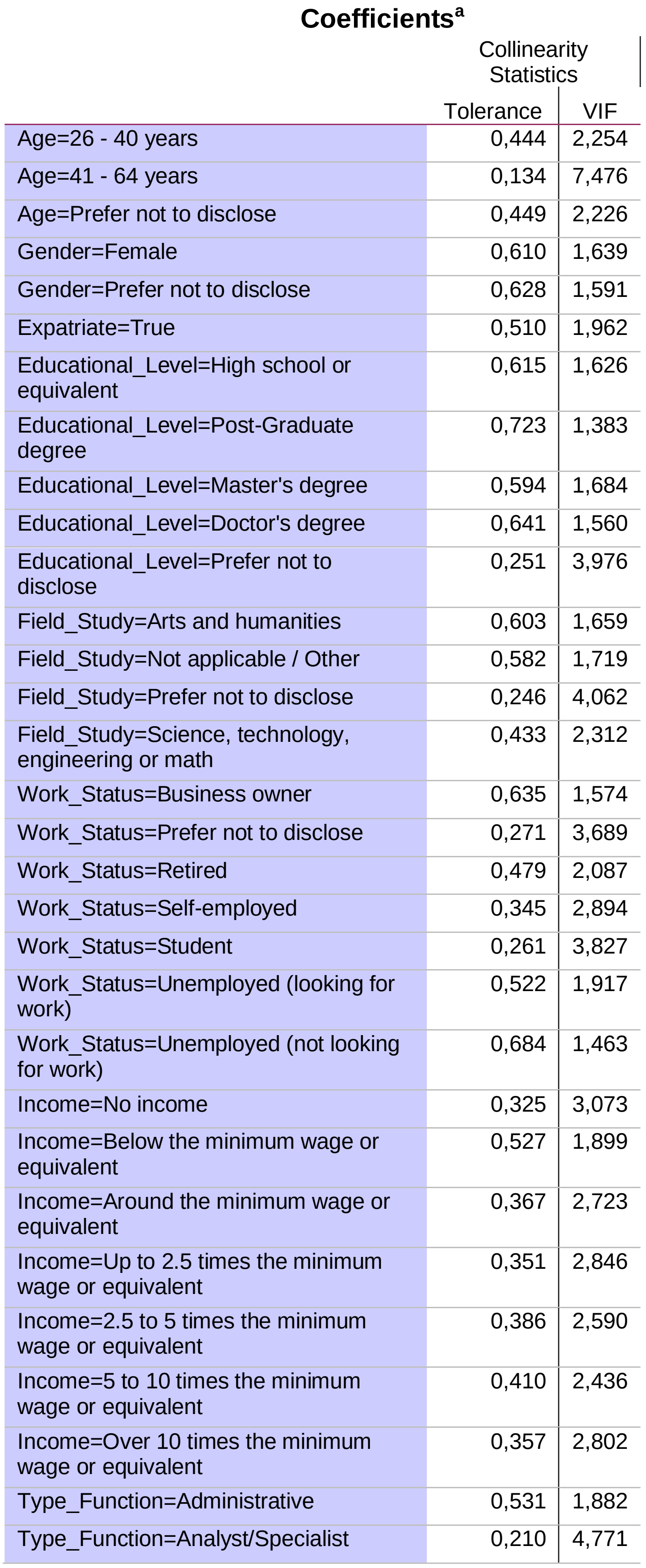
		Q1AT	Q2AT	Q3AT	Q4AT	Q5AT
Score	Correlation Coefficient	0,055	,168*	0,022	0,127	0,079
	Sig. (2-tailed)	0,426	0,015	0,757	0,068	0,255
Statistical knowledge	Correlation Coefficient	,500**	,424**	,583**	,448**	,286**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000
Understanding	Correlation Coefficient	,482**	,479**	,498**	,556**	,324**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000
DMTA	Correlation Coefficient	,522**	,532**	,582**	,626**	,394**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000
Software	Correlation Coefficient	,285**	,324**	,415**	,371**	,293**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000
Acting	Correlation Coefficient	,467**	,342**	,306**	,428**	,289**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000
Critical	Correlation Coefficient	,398**	,314**	,288**	,350**	,231**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,001
Q1AT	Correlation Coefficient	1,000	,469**	,578**	,555**	,344**
	Sig. (2-tailed)		0,000	0,000	0,000	0,000
Q2AT	Correlation Coefficient	,469**	1,000	,615**	,652**	,358**
	Sig. (2-tailed)	0,000		0,000	0,000	0,000
Q3AT	Correlation Coefficient	,578**	,615**	1,000	,652**	,428**
	Sig. (2-tailed)	0,000	0,000		0,000	0,000
Q4AT	Correlation Coefficient	,555**	,652**	,652**	1,000	,504**
	Sig. (2-tailed)	0,000	0,000	0,000		0,000
Q5AT	Correlation Coefficient	,344**	,358**	,428**	,504**	1,000
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	
Environment	Correlation Coefficient	,338**	,331**	,379**	,428**	,305**
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Annexe 8: binary logistic regressions

Table 39: Test for multicollinearity



Coefficients[a]

	Collinearity Statistics	
	Tolerance	VIF
Age=26 - 40 years	0,444	2,254
Age=41 - 64 years	0,134	7,476
Age=Prefer not to disclose	0,449	2,226
Gender=Female	0,610	1,639
Gender=Prefer not to disclose	0,628	1,591
Expatriate=True	0,510	1,962
Educational_Level=High school or equivalent	0,615	1,626
Educational_Level=Post-Graduate degree	0,723	1,383
Educational_Level=Master's degree	0,594	1,684
Educational_Level=Doctor's degree	0,641	1,560
Educational_Level=Prefer not to disclose	0,251	3,976
Field_Study=Arts and humanities	0,603	1,659
Field_Study=Not applicable / Other	0,582	1,719
Field_Study=Prefer not to disclose	0,246	4,062
Field_Study=Science, technology, engineering or math	0,433	2,312
Work_Status=Business owner	0,635	1,574
Work_Status=Prefer not to disclose	0,271	3,689
Work_Status=Retired	0,479	2,087
Work_Status=Self-employed	0,345	2,894
Work_Status=Student	0,261	3,827
Work_Status=Unemployed (looking for work)	0,522	1,917
Work_Status=Unemployed (not looking for work)	0,684	1,463
Income=No income	0,325	3,073
Income=Below the minimum wage or equivalent	0,527	1,899
Income=Around the minimum wage or equivalent	0,367	2,723
Income=Up to 2.5 times the minimum wage or equivalent	0,351	2,846
Income=2.5 to 5 times the minimum wage or equivalent	0,386	2,590
Income=5 to 10 times the minimum wage or equivalent	0,410	2,436
Income=Over 10 times the minimum wage or equivalent	0,357	2,802
Type_Function=Administrative	0,531	1,882
Type_Function=Analyst/Specialist	0,210	4,771

Type_Function=Business owner	0,460	2,174
Type_Function=Intern	0,398	2,514
Type_Function=Management/Direction	0,337	2,969
Type_Function=Operational	0,400	2,502
Type_Function=Prefer not to disclose	0,299	3,344
Type_Function=Professor	0,320	3,129
Type_Function=Self-employed	0,267	3,746
Type_Function=Voluntary	0,582	1,718
Work_experience=No experience	0,627	1,596
Work_experience=Less than 1 year	0,568	1,760
Work_experience=4 to 7 years	0,442	2,261
Work_experience=8 to 10 years	0,539	1,854
Work_experience=11 to 20 years	0,342	2,923
Work_experience=Over 20 years	0,145	6,908
Work_experience=Prefer not to disclose	0,364	2,750
Number_employees=1 to 10	0,442	2,263
Number_employees=11 to 50	0,522	1,916
Number_employees=51 to 100	0,571	1,752
Number_employees=101 to 300	0,543	1,842
Number_employees=301 to 500	0,763	1,311
Number_employees=Over 500	0,361	2,773
Number_employees=Prefer not to disclose	0,384	2,602
Attitude=1.0	0,774	1,292
Attitude=2.0	0,522	1,916
Attitude=3.0	0,522	1,917
Attitude=4.0	0,527	1,896
Attitude=6.0	0,480	2,081
Attitude=7.0	0,569	1,756
Environment=1.0	0,603	1,659
Environment=2.0	0,477	2,095
Environment=3.0	0,537	1,862
Environment=4.0	0,439	2,279
Environment=5.0	0,423	2,365
Environment=7.0	0,586	1,707

a. Dependent Variable: Score

Table 40: Binary Logistic Regression for the variable Score_Positive

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	36,874	15	0,001
	Block	36,874	15	0,001
	Model	36,874	15	0,001

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13,836	8	0,086

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	152,443[a]	0,200	0,293

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Score_Positive		Percentage Correct
Observed			,00	1,00	
Step 1	Score_Positive	,00	18	25	41,9
		1,00	10	112	91,8
	Overall Percentage				78,8

a. The cut value is ,500

Casewise List[b]

		Observed			Temporary Variable		
Case	Selected Status[a]	Score_Positive	Predicted	Predicted Group	Resid	ZResid	SResid
29	S	0**	0,953	1	-0,953	-4,495	-2,524
31	S	0**	0,854	1	-0,854	-2,416	-2,038
81	S	0**	0,875	1	-0,875	-2,647	-2,107
91	S	0**	0,856	1	-0,856	-2,436	-2,065
156	S	0**	0,885	1	-0,885	-2,771	-2,155
169	S	0**	0,847	1	-0,847	-2,354	-2,013

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender(1)	0,927	0,455	4,141	1	0,042	2,527	1,035	6,170
Expatriate(1)	0,050	0,624	0,006	1	0,936	1,051	0,309	3,571
Age_26_Plus(1)	-1,217	0,535	5,179	1	0,023	0,296	0,104	0,845
University_Standard_ Education(1)	0,805	0,665	1,467	1	0,226	2,238	0,608	8,238
Post_graduation_ education(1)	2,080	0,851	5,973	1	0,015	8,008	1,510	42,473
STEM_Education(1)	0,235	0,623	0,142	1	0,706	1,265	0,373	4,288
Social_Education(1)	0,697	0,482	2,095	1	0,148	2,008	0,781	5,160
Employee(1)	-1,488	0,627	5,632	1	0,018	0,226	0,066	0,772
Entrepreneur(1)	-1,281	0,738	3,013	1	0,083	0,278	0,065	1,180
Above_Min_Salary_ Range(1)	0,112	0,469	0,057	1	0,811	1,119	0,446	2,807
Specialist(1)	1,329	0,561	5,613	1	0,018	3,776	1,258	11,336
Experience_3_Plus(1)	0,112	0,557	0,040	1	0,841	1,118	0,375	3,332
Company_11_Plus_ Emp(1)	-0,143	0,482	0,088	1	0,767	0,867	0,337	2,230
Attitude_Literate(1)	-0,309	0,496	0,389	1	0,533	0,734	0,278	1,939
Environment_Literate(1)	-0,274	0,476	0,331	1	0,565	0,761	0,299	1,933
Constant	0,583	0,731	0,635	1	0,425	1,791		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 41: Binary Logistic Regression for the variable Score_High

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	34,536	15	0,003
	Block	34,536	15	0,003
	Model	34,536	15	0,003

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4,998	8	0,758

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	143,546[a]	0,189	0,286

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Score_High		Percentage Correct
Observed			,00	1,00	
Step 1	Score_High	,00	121	6	95,3
		1,00	25	13	34,2
	Overall Percentage				81,2

a. The cut value is ,500

Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable		
		Score_High			Resid	ZResid	SResid
9	S	1**	0,035	0	0,965	5,285	2,675
90	S	1**	0,126	0	0,874	2,636	2,138
96	S	1**	0,144	0	0,856	2,441	2,063
161	S	1**	0,153	0	0,847	2,356	2,029
193	S	1**	0,102	0	0,898	2,964	2,201

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender(1)	0,071	0,476	0,022	1	0,881	1,074	0,423	2,728
Expatriate(1)	0,496	0,657	0,570	1	0,450	1,642	0,453	5,955
Age_26_Plus(1)	-1,990	0,579	11,823	1	0,001	0,137	0,044	0,425
University_Standard_Education(1)	1,660	1,173	2,005	1	0,157	5,261	0,528	52,378
Post_graduation_education(1)	2,156	1,216	3,145	1	0,076	8,640	0,797	93,647
STEM_Education(1)	0,701	0,726	0,932	1	0,334	2,015	0,486	8,359
Social_Education(1)	0,417	0,549	0,577	1	0,447	1,517	0,518	4,449
Employee(1)	0,315	0,694	0,207	1	0,649	1,371	0,352	5,337
Entrepreneur(1)	1,410	0,764	3,407	1	0,065	4,094	0,916	18,289
Above_Min_Salary_Range(1)	1,486	0,550	7,307	1	0,007	4,418	1,504	12,975
Specialist(1)	-0,236	0,584	0,163	1	0,686	0,790	0,252	2,480
Experience_3_Plus(1)	-0,658	0,599	1,206	1	0,272	0,518	0,160	1,676
Company_11_Plus_Emp(1)	1,415	0,521	7,368	1	0,007	4,118	1,482	11,444
Attitude_Literate(1)	0,390	0,519	0,563	1	0,453	1,476	0,534	4,083
Environment_Literate(1)	0,347	0,497	0,489	1	0,484	1,415	0,535	3,746
Constant	-4,767	1,323	12,988	1	0,000	0,009		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 42: Binary Logistic Regression for the variable Understanding_Literate

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	54,693	15	0,000
	Block	54,693	15	0,000
	Model	54,693	15	0,000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12,043	8	0,149

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	140,602[a]	0,282	0,407

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Understanding_ Literate		
Observed			,00	1,00	Percentage Correct
Step 1	Understanding_ Literate	,00	29	17	63,0
		1,00	11	108	90,8
	Overall Percentage				83,0

a. The cut value is ,500

Casewise List[b]

		Observed			Temporary Variable		
Case	Selected Status[a]	Understanding_ Literate	Predicted	Predicted Group	Resid	ZResid	SResid
4	S	0**	0,912	1	-0,912	-3,215	-2,279
21	S	0**	0,896	1	-0,896	-2,935	-2,192
78	S	0**	0,959	1	-0,959	-4,853	-2,592
80	S	0**	0,855	1	-0,855	-2,424	-2,033
83	S	0**	0,884	1	-0,884	-2,762	-2,131
146	S	0**	0,893	1	-0,893	-2,893	-2,168
154	S	0**	0,945	1	-0,945	-4,156	-2,445
160	S	0**	0,842	1	-0,842	-2,309	-2,031
163	S	0**	0,897	1	-0,897	-2,953	-2,230

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Gender(1)	-0,042	0,468	0,008	1	0,929	0,959	0,383	2,399
Expatriate(1)	-0,002	0,701	0,000	1	0,998	0,998	0,253	3,940
Age_26_Plus(1)	0,283	0,522	0,294	1	0,587	1,327	0,477	3,692
University_Standard_ Education(1)	1,810	0,789	5,269	1	0,022	6,111	1,303	28,665
Post_graduation_ education(1)	1,839	0,890	4,270	1	0,039	6,291	1,099	35,997
STEM_Education(1)	0,270	0,772	0,122	1	0,726	1,310	0,288	5,955
Social_Education(1)	-0,453	0,522	0,751	1	0,386	0,636	0,228	1,770
Employee(1)	-1,029	0,693	2,207	1	0,137	0,357	0,092	1,389
Entrepreneur(1)	-0,232	0,702	0,109	1	0,741	0,793	0,200	3,140
Above_Min_Salary_ Range(1)	1,480	0,509	8,459	1	0,004	4,391	1,620	11,903
Specialist(1)	0,097	0,609	0,025	1	0,874	1,102	0,334	3,635
Experience_3_Plus(1)	-0,695	0,582	1,428	1	0,232	0,499	0,160	1,560
Company_11_Plus_ Emp(1)	-0,610	0,492	1,534	1	0,215	0,544	0,207	1,426
Attitude_Literate(1)	2,328	0,531	19,251	1	0,000	10,255	3,625	29,009
Environment_Literate(1)	0,475	0,515	0,852	1	0,356	1,609	0,586	4,415
Constant	-1,519	0,895	2,881	1	0,090	0,219		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 43: Binary Logistic Regression for the variable DMTA_Literate

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	49,040	15	0,000
	Block	49,040	15	0,000
	Model	49,040	15	0,000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,802	8	0,359

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	175,261[a]	0,257	0,346

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

Classification Table[a]

Observed			Predicted		
			DMTA_Literate		Percentage Correct
			,00	1,00	
Step 1	DMTA_Literate	,00	49	20	71,0
		1,00	21	75	78,1
	Overall Percentage				75,2

a. The cut value is ,500

Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable		
		DMTA_Literate			Resid	ZResid	SResid
37	S	0**	0,860	1	-0,860	-2,477	-2,027
96	S	1**	0,152	0	0,848	2,359	2,020
105	S	0**	0,869	1	-0,869	-2,576	-2,075
140	S	0**	0,876	1	-0,876	-2,656	-2,086
154	S	0**	0,880	1	-0,880	-2,710	-2,100
199	S	0**	0,860	1	-0,860	-2,477	-2,023

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Gender(1)	0,609	0,407	2,237	1	0,135	1,839	0,828	4,086
Expatriate(1)	-0,256	0,583	0,192	1	0,661	0,774	0,247	2,430
Age_26_Plus(1)	-0,015	0,490	0,001	1	0,975	0,985	0,377	2,575
University_Standard_ Education(1)	0,213	0,644	0,109	1	0,742	1,237	0,350	4,373
Post_graduation_ education(1)	0,100	0,737	0,018	1	0,892	1,105	0,260	4,689
STEM_Education(1)	-0,155	0,618	0,063	1	0,802	0,857	0,255	2,875
Social_Education(1)	-0,334	0,460	0,527	1	0,468	0,716	0,291	1,764
Employee(1)	-1,074	0,610	3,098	1	0,078	0,342	0,103	1,130
Entrepreneur(1)	-1,294	0,690	3,511	1	0,061	0,274	0,071	1,061
Above_Min_Salary_ Range(1)	0,649	0,433	2,244	1	0,134	1,914	0,819	4,473
Specialist(1)	0,369	0,522	0,501	1	0,479	1,447	0,520	4,024
Experience_3_Plus(1)	0,331	0,522	0,403	1	0,525	1,393	0,501	3,874
Company_11_Plus_ Emp(1)	-0,221	0,439	0,252	1	0,615	0,802	0,339	1,896
Attitude_Literate(1)	1,824	0,427	18,210	1	0,000	6,194	2,681	14,314
Environment_Literate(1)	0,398	0,437	0,831	1	0,362	1,489	0,632	3,506
Constant	-0,935	0,739	1,602	1	0,206	0,393		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 44: Binary Logistic Regression for the variable Statistics_Literate

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	54,027	15	0,000
	Block	54,027	15	0,000
	Model	54,027	15	0,000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7,933	8	0,440

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	163,375[a]	0,279	0,381

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Classification Table[a]

Observed			Predicted		
			Statistics_Literate		Percentage Correct
			,00	1,00	
Step 1	Statistics_Literate	,00	86	18	82,7
		1,00	24	37	60,7
	Overall Percentage				74,5

a. The cut value is ,500

Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable		
		Statistics_ Literate			Resid	ZResid	SResid
20	S	1**	0,102	0	0,898	2,965	2,219
27	S	0**	0,833	1	-0,833	-2,236	-2,013
28	S	1**	0,137	0	0,863	2,510	2,061
52	S	1**	0,136	0	0,864	2,524	2,076
87	S	1**	0,129	0	0,871	2,597	2,092
140	S	0**	0,857	1	-0,857	-2,448	-2,019

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Gender(1)	0,698	0,432	2,609	1	0,106	2,010	0,862	4,689
Expatriate(1)	0,401	0,619	0,418	1	0,518	1,493	0,443	5,024
Age_26_Plus(1)	0,683	0,511	1,783	1	0,182	1,979	0,727	5,390
University_Standard_ Education(1)	1,383	0,832	2,762	1	0,097	3,986	0,780	20,358
Post_graduation_ education(1)	0,963	0,894	1,159	1	0,282	2,619	0,454	15,111
STEM_Education(1)	0,534	0,621	0,739	1	0,390	1,705	0,505	5,758
Social_Education(1)	-0,422	0,491	0,738	1	0,390	0,656	0,250	1,718
Employee(1)	-0,784	0,656	1,426	1	0,232	0,457	0,126	1,653
Entrepreneur(1)	-0,034	0,717	0,002	1	0,962	0,966	0,237	3,942
Above_Min_Salary_ Range(1)	0,623	0,484	1,655	1	0,198	1,865	0,722	4,819
Specialist(1)	-0,016	0,538	0,001	1	0,976	0,984	0,343	2,823
Experience_3_Plus(1)	-1,007	0,593	2,883	1	0,090	0,365	0,114	1,168
Company_11_Plus_ Emp(1)	-0,096	0,489	0,039	1	0,844	0,908	0,348	2,366
Attitude_Literate(1)	2,171	0,498	19,027	1	0,000	8,764	3,305	23,241
Environment_Literate(1)	0,493	0,465	1,127	1	0,288	1,638	0,659	4,072
Constant	-3,488	0,947	13,562	1	0,000	0,031		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 45: Binary Logistic Regression for the variable Software_Positive

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	70,592	15	0,000
	Block	70,592	15	0,000
	Model	70,592	15	0,000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,068	8	0,427

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	145,717[a]	0,348	0,477

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

Observed			Predicted		
			Software_Positive		Percentage Correct
			,00	1,00	
Step 1	Software_Positive	,00	37	23	61,7
		1,00	15	90	85,7
	Overall Percentage				77,0

a. The cut value is ,500

Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable		
		Software_ Positive			Resid	ZResid	SResid
34	S	1**	0,138	0	0,862	2,504	2,071
49	S	1**	0,106	0	0,894	2,899	2,180
90	S	1**	0,134	0	0,866	2,539	2,096
174	S	1**	0,111	0	0,889	2,834	2,217

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Gender(1)	1,020	0,449	5,165	1	0,023	2,773	1,151	6,682
Expatriate(1)	1,359	0,732	3,447	1	0,063	3,893	0,927	16,351
Age_26_Plus(1)	-1,219	0,589	4,289	1	0,038	0,296	0,093	0,937
University_Standard_ Education(1)	0,722	0,704	1,051	1	0,305	2,058	0,518	8,176
Post_graduation_ education(1)	1,052	0,801	1,721	1	0,190	2,862	0,595	13,769
STEM_Education(1)	20,259	6641,520	0,000	1	0,998	628583196,368	0,000	
Social_Education(1)	-1,022	0,474	4,640	1	0,031	0,360	0,142	0,912
Employee(1)	-0,540	0,652	0,687	1	0,407	0,583	0,162	2,090
Entrepreneur(1)	-0,502	0,743	0,456	1	0,500	0,605	0,141	2,600
Above_Min_Salary_ Range(1)	0,279	0,466	0,358	1	0,549	1,322	0,530	3,295
Specialist(1)	-0,308	0,604	0,261	1	0,609	0,735	0,225	2,398
Experience_3_Plus(1)	0,031	0,588	0,003	1	0,958	1,031	0,326	3,262
Company_11_Plus_ Emp(1)	0,966	0,490	3,883	1	0,049	2,628	1,005	6,869
Attitude_Literate(1)	1,148	0,476	5,813	1	0,016	3,152	1,240	8,015
Environment_Literate(1)	0,502	0,457	1,205	1	0,272	1,652	0,674	4,050
Constant	-1,093	0,814	1,800	1	0,180	0,335		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 46: Binary Logistic Regression for the variable Software_High

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	30,790	15	0,009
	Block	30,790	15	0,009
	Model	30,790	15	0,009

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1,432	7	0,985

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	60,224[a]	0,170	0,402

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Software_High		Percentage Correct
Observed			,00	1,00	
Step 1	Software_High	,00	148	4	97,4
		1,00	11	2	15,4
	Overall Percentage				90,9

a. The cut value is ,500

Casewise List[b]

		Observed			Temporary Variable		
Case	Selected Status[a]	Software_ High	Predicted	Predicted Group	Resid	ZResid	SResid
25	S	1**	0,200	0	0,800	1,998	2,293
55	S	1**	0,061	0	0,939	3,915	2,488
124	S	1**	0,041	0	0,959	4,838	2,608
138	S	1**	0,130	0	0,870	2,582	2,343

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Gender(1)	-0,714	0,873	0,668	1	0,414	0,490	0,088	2,711
Expatriate(1)	-0,773	1,272	0,369	1	0,543	0,462	0,038	5,583
Age_26_Plus(1)	-1,383	1,071	1,669	1	0,196	0,251	0,031	2,045
University_Standard_Education(1)	-0,083	1,493	0,003	1	0,955	0,920	0,049	17,161
Post_graduation_education(1)	-0,892	1,584	0,317	1	0,574	0,410	0,018	9,149
STEM_Education(1)	2,910	1,343	4,692	1	0,030	18,358	1,319	255,503
Social_Education(1)	0,462	1,308	0,125	1	0,724	1,587	0,122	20,584
Employee(1)	-1,636	1,196	1,869	1	0,172	0,195	0,019	2,032
Entrepreneur(1)	1,499	1,489	1,013	1	0,314	4,476	0,242	82,870
Above_Min_Salary_Range(1)	0,261	0,940	0,077	1	0,781	1,298	0,206	8,196
Specialist(1)	-0,044	0,893	0,002	1	0,961	0,957	0,166	5,512
Experience_3_Plus(1)	1,828	1,119	2,668	1	0,102	6,220	0,694	55,766
Company_11_Plus_Emp(1)	1,146	1,024	1,251	1	0,263	3,144	0,422	23,410
Attitude_Literate(1)	2,747	1,291	4,528	1	0,033	15,593	1,242	195,754
Environment_Literate(1)	1,858	1,273	2,131	1	0,144	6,410	0,529	77,642
Constant	-7,422	2,494	8,858	1	0,003	0,001		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 47: Binary Logistic Regression for the variable Acting_Literate

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	37,377	15	0,001
	Block	37,377	15	0,001
	Model	37,377	15	0,001

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3,454	8	0,903

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	122,056[a]	0,203	0,327

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Classification Table[a]

Observed			Predicted		
			Acting_Literate		Percentage Correct
			,00	1,00	
Step 1	Acting_Literate	,00	8	23	25,8
		1,00	6	128	95,5
	Overall Percentage				82,4

a. The cut value is ,500

Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable		
		Acting_ Literate			Resid	ZResid	SResid
27	S	0**	0,815	1	-0,815	-2,098	-2,046
52	S	0**	0,898	1	-0,898	-2,973	-2,254
54	S	0**	0,917	1	-0,917	-3,328	-2,317
83	S	0**	0,921	1	-0,921	-3,403	-2,305
134	S	0**	0,893	1	-0,893	-2,892	-2,221
140	S	0**	0,895	1	-0,895	-2,927	-2,189
154	S	0**	0,957	1	-0,957	-4,717	-2,544

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Gender(1)	0,055	0,500	0,012	1	0,912	1,057	0,396	2,816
Expatriate(1)	-0,736	0,728	1,023	1	0,312	0,479	0,115	1,995
Age_26_Plus(1)	0,531	0,589	0,813	1	0,367	1,700	0,536	5,391
University_Standard_ Education(1)	-0,451	0,725	0,388	1	0,534	0,637	0,154	2,635
Post_graduation_ education(1)	1,285	1,035	1,542	1	0,214	3,615	0,476	27,477
STEM_Education(1)	-0,310	0,788	0,155	1	0,694	0,734	0,157	3,436
Social_Education(1)	-0,034	0,546	0,004	1	0,951	0,967	0,332	2,818
Employee(1)	-1,063	0,680	2,449	1	0,118	0,345	0,091	1,308
Entrepreneur(1)	0,314	0,792	0,157	1	0,692	1,369	0,290	6,466
Above_Min_Salary_ Range(1)	0,673	0,512	1,727	1	0,189	1,959	0,719	5,342
Specialist(1)	0,884	0,678	1,703	1	0,192	2,421	0,642	9,137
Experience_3_Plus(1)	-0,150	0,591	0,065	1	0,799	0,860	0,270	2,742
Company_11_Plus_ Emp(1)	0,080	0,509	0,025	1	0,875	1,084	0,400	2,937
Attitude_Literate(1)	2,062	0,616	11,190	1	0,001	7,859	2,348	26,302
Environment_Literate(1)	0,388	0,540	0,516	1	0,473	1,474	0,512	4,246
Constant	0,404	0,861	0,220	1	0,639	1,498		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 48: Binary Logistic Regression for the variable Critical_Literate

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	29,852	15	0,012
	Block	29,852	15	0,012
	Model	29,852	15	0,012

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7,789	8	0,454

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	45,596[a]	0,166	0,451

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Critical_Literate		Percentage Correct
Observed			,00	1,00	
Step 1	Critical_ Literate	,00	3	7	30,0
		1,00	1	154	99,4
	Overall Percentage				95,2

a. The cut value is ,500

Casewise List[b]

		Observed			Temporary Variable		
Case	Selected Status[a]	Critical_ Literate	Predicted	Predicted Group	Resid	ZResid	SResid
6	S	0**	0,990	1	-0,990	-10,056	-3,091
64	S	0**	0,962	1	-0,962	-5,029	-2,764
156	S	0**	0,802	1	-0,802	-2,012	-2,161

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2,000 are listed.

Variables in the Equation

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Gender(1)	-0,274	0,887	0,095	1	0,758	0,761	0,134	4,328
Expatriate(1)	-2,079	1,293	2,585	1	0,108	0,125	0,010	1,577
Age_26_Plus(1)	-0,413	1,051	0,154	1	0,694	0,662	0,084	5,195
University_Standard_ Education(1)	1,748	1,412	1,531	1	0,216	5,741	0,360	91,429
Post_graduation_ education(1)	2,150	1,514	2,015	1	0,156	8,582	0,441	166,932
STEM_Education(1)	-2,232	2,031	1,207	1	0,272	0,107	0,002	5,753
Social_Education(1)	-4,035	1,848	4,767	1	0,029	0,018	0,000	0,662
Employee(1)	-0,384	1,430	0,072	1	0,788	0,681	0,041	11,227
Entrepreneur(1)	-0,765	1,194	0,411	1	0,521	0,465	0,045	4,827
Above_Min_Salary_ Range(1)	1,639	1,016	2,601	1	0,107	5,148	0,703	37,715
Specialist(1)	-1,420	1,150	1,525	1	0,217	0,242	0,025	2,302
Experience_3_Plus(1)	-1,308	0,981	1,778	1	0,182	0,270	0,040	1,848
Company_11_Plus_ Emp(1)	2,852	1,367	4,354	1	0,037	17,328	1,189	252,494
Attitude_Literate(1)	2,386	1,405	2,882	1	0,090	10,865	0,692	170,696
Environment_Literate(1)	0,253	1,192	0,045	1	0,832	1,288	0,124	13,328
Constant	3,852	2,089	3,401	1	0,065	47,077		

a. Variable(s) entered on step 1: Gender, Expatriate, Age_26_Plus, University_Standard_Education, Post_graduation_education, STEM_Education, Social_Education, Employee, Entrepreneur, Above_Min_Salary_Range, Specialist, Experience_3_Plus, Company_11_Plus_Emp, Attitude_Literate, Environment_Literate.

Table 49: Summary table with Exp(B) value and significance in all regressions

Summary of prediction - Value of Exp(B)									
Variables:	Score_ positive	Score_high	Understanding_ Literate	DMTA_ Literate	Statistics_ Literate	Software_ Positive	Software_ High	Acting_ Literate	Critical_ Literate
Gender(1)	2,527	1,074	0,959	1,839	2,010	2,773	0,490	1,057	0,761
Significance:	0,042	0,881	0,929	0,135	0,106	0,023	0,414	0,912	0,758
Expatriate(1)	1,051	1,642	0,998	0,774	1,493	3,893	0,462	0,479	0,125
Significance:	0,936	0,450	0,998	0,661	0,518	0,063	0,543	0,312	0,108
Age_26_Plus(1)	0,296	0,137	1,327	0,985	1,979	0,296	0,251	1,700	0,662
Significance:	0,023	0,001	0,587	0,975	0,182	0,038	0,196	0,367	0,694
University_Standard_ Education(1)	2,238	5,261	6,111	1,237	3,986	2,058	0,920	0,637	5,741
Significance:	0,226	0,157	0,022	0,742	0,097	0,305	0,955	0,534	0,216
Post_graduation_ education(1)	8,008	8,640	6,291	1,105	2,619	2,862	0,410	3,615	8,582
Significance:	0,015	0,076	0,039	0,892	0,282	0,190	0,574	0,214	0,156
STEM_Education(1)	1,265	2,015	1,310	0,857	1,705	628583196,368	18,358	0,734	0,107
Significance:	0,706	0,334	0,726	0,802	0,390	0,998	0,030	0,694	0,272
Social_Education(1)	2,008	1,517	0,636	0,716	0,656	0,360	1,587	0,967	0,018
Significance:	0,148	0,447	0,386	0,468	0,390	0,031	0,724	0,951	0,029
Employee(1)	0,226	1,371	0,357	0,342	0,457	0,583	0,195	0,345	0,681
Significance:	0,018	0,649	0,137	0,078	0,232	0,407	0,172	0,118	0,788
Entrepreneur(1)	0,278	4,094	0,793	0,274	0,966	0,605	4,476	1,369	0,465
Significance:	0,083	0,065	0,741	0,061	0,962	0,500	0,314	0,692	0,521
Above_Min_ Salary_Range(1)	1,119	4,418	4,391	1,914	1,865	1,322	1,298	1,959	5,148
Significance:	0,811	0,007	0,004	0,134	0,198	0,549	0,781	0,189	0,107
Specialist(1)	3,776	0,790	1,102	1,447	0,984	0,735	0,957	2,421	0,242
Significance:	0,018	0,686	0,874	0,479	0,976	0,609	0,961	0,192	0,217
Experience_3_ Plus(1)	1,118	0,518	0,499	1,393	0,365	1,031	6,220	0,860	0,270
Significance:	0,841	0,272	0,232	0,525	0,090	0,958	0,102	0,799	0,182
Company_11_Plus_ Emp(1)	0,867	4,118	0,544	0,802	0,908	2,628	3,144	1,084	17,328
Significance:	0,767	0,007	0,215	0,615	0,844	0,049	0,263	0,875	0,037
Attitude_Literate(1)	0,734	1,476	10,255	6,194	8,764	3,152	15,593	7,859	10,865
Significance:	0,533	0,453	0,000	0,000	0,000	0,016	0,033	0,001	0,090
Environment_ Literate(1)	0,761	1,415	1,609	1,489	1,638	1,652	6,410	1,474	1,288
Significance:	0,565	0,484	0,356	0,362	0,288	0,272	0,144	0,473	0,832