

Aos meus pais

Agradecimentos

Gostaria de iniciar este trabalho agradecendo ao Professor Doutor Paulo Melo, meu mentor nesta jornada, pela grande disponibilidade sempre manifestada e motivação dada ao longo destes dois anos que me permitiram tomar as melhores opções e fomentar o meu crescimento pessoal e profissional da melhor forma.

Para além disso, a todos os professores que participaram da minha formação, na impossibilidade de os aqui mencionar na totalidade, um agradecimento pelo empenho e dedicação a todos os alunos.

Uma menção particularmente especial à Redcorp, na pessoa do seu fundador Ronald Reich, pela liberdade, confiança e apoio depositado, de forma altruísta e desinteressada, mas sem o qual este trabalho nunca teria sido possível. Para além disso, ao resto da equipa, pela amizade e afabilidade com que me receberam.

Também à minha antes de tudo amiga e além disso namorada Joana, que a muito me atura ao longo destes anos, mas em quem encontro sempre o suporte, motivação e irreverência necessária para descobrir novos limites e a jovialidade para acreditar num futuro feliz.

Finalmente, aos meus pais, irmã e família por me terem apoiado sempre, acompanhando-me e aparando os tombos pelo caminho, e por me proporcionarem todas as oportunidades de crescimento que tive, dando-me espaço para crescer, mas também a educação e bases morais que procuro seguir e tenho a certeza são feitos os grandes homens. A vocês, a mais que ninguém, procuro orgulhar e moldar-me à imagem que em mim reconhecem.

Resumo

A Internet tornou-se nas últimas décadas uma das mais poderosas e incontornáveis ferramentas de comunicação em todo o mundo, representando um dos mais importantes ecossistemas para a promoção das organizações e realização de transacções a nível global. Por conseguinte, a mensuração de resultados e do retorno no investimento feito em conteúdos digitais assume crescentemente importância para profissionais cujo papel é gerir o conhecimento e desempenho das organizações. Neste contexto, os *web analytics* são uma ferramenta indispensável para a contínua avaliação dos principais indicadores de desempenho do negócio, focando sobretudo o *website* como componente agregador da estratégia digital. A recolha e análise de dados na *web* aponta assim, em última análise, à optimização de conteúdos, do design e do modelo de negócio, através de mudanças fundamentadas na análise de métricas e nos factos transmitidos pelos dados, por oposição a simples inclinações pessoais do decisor.

De forma a explorar a aplicação destas técnicas em ambiente prático, recorreremos assim à ferramenta do Google Analytics para a análise de um caso de estudo, recorrendo à análise de um *website ecommerce* no ramo dos componentes informáticos, com empresa sediada na Bélgica. Trata-se assim de um ambiente B2B, explorando de forma extensiva neste trabalho os principais indicadores, através da análise individual dos relatórios providenciados por esta ferramenta e o seu contributo para a compreensão da evolução do negócio. Definimos para além disso numa fase inicial, o âmbito de aplicação e as tecnologias utilizadas, bem como os conceitos-chave associados a estas ferramentas. Para além disso, procuramos também aqui integrar os dados recolhidos com outras aplicações software, agilizando o tratamento e visualização para além da interface de utilizador.

Palavras-Chave: Web Analytics, Google Analytics, Ecommerce, Indicadores de performance, Experiência do utilizador.

Abstract

The web has become one of most powerful tools of communication in the world today, representing one of the most important environments for the promotion of organizations and the realization of transactions worldwide. Because of that, measuring the results and the return on the investment made on digital materials is increasingly important for professionals, whose job is to monitor knowledge and performance. In this context, web analytics applications are a valuable tool for continuously assess these indicators performance, focusing on the organizations' website as the core component for most digital strategies. The collection and analysis of web data ultimately aims at content, design and business optimization, based on educated premises supported on figures and facts, as opposed to decision processes based solely on personal inclination from decision makers.

In order to explore the application of these techniques in a business environment, we resort to Google Analytics for the analysis of a case study of a website from an ecommerce IT retailer based in Belgium, working in a B2B environment. This research extensively covers the main indicators available, individually assessing each report's contribution for the comprehension of business evolution. In addition, we start by defining the ambit of application, the technologies used, as well as the main concepts associated with this kind of tools. Moreover, we also look into the integration of web data with other software applications, for an agile visualization and treatment of the data.

Keywords: Web Analytics, Google Analytics, Ecommerce, Performance Indicators, User Experience.

Contents

<i>Agradecimientos</i>	<i>iii</i>
<i>Resumo</i>	<i>iv</i>
<i>Abstract</i>	<i>v</i>
<i>Contents</i>	<i>vi</i>
Symbols and Acronyms	ix
Figures	x
Tables	xii
Formulas.....	xiii
Models.....	xiii
1 Introduction	1
1.1 Why we need web analytics	2
1.1.1 Levels of analysis	4
1.2 Data-driven organizations.....	5
1.3 Data collection methodologies	7
1.3.1 Log files.....	8
1.3.2 Page Tagging.....	9
1.4 Privacy issues	11
1.4.1 “User ID” dimension	15
2 Google Analytics as Software as a Service	17
2.1 Core concepts and metrics.....	20
2.2 Defining indicators	26
2.3 Meeting Objectives and Indicators	30
2.4 Google Analytics reporting API	33
2.4.1 Integration of data with other applications	35
2.4.2 Statistical procedures in web analytics	37

3 Case study: Redcorp.....	41
3.1 Methodology.....	43
3.2 Previous Research	43
4 Google Analytics interface	47
4.1 Intelligence Events	48
4.1.1 Definition	48
4.1.2 Analysis.....	50
4.2 Audience.....	50
4.2.1 Definition	50
4.2.2 Analysis.....	51
4.2.3 Summary.....	58
4.2.4 Period Comparison	60
4.3 Acquisition.....	62
4.3.1 Definition	62
4.3.2 Analysis.....	65
4.3.3 Summary.....	74
4.3.4 Period Comparison	75
4.4 Behavior	78
4.4.1 Definition	78
4.4.2 Analysis.....	81
4.4.3 Summary.....	87
4.4.4 Period Comparison	88
4.5 Conversions.....	89
4.5.1 Definition	89
4.5.2 Analysis.....	91
4.5.3 Summary.....	103

4.5.6 Period Comparison	105
<i>5 Statistical Procedures</i>	<i>108</i>
5.1 Modeling with R	109
5.1.1 Session Dimensions	109
5.1.2 Channel Dimensions	116
<i>6 Concluding Remarks</i>	<i>121</i>
<i>References</i>	<i>123</i>
<i>Appendix</i>	<i>129</i>
Channel Dimensions – Models and Diagnostics	129
Baseline Model	129
Extended Model	130
Selected Model.....	131

Symbols and Acronyms

API – Application Programming Interface

CRAN – Comprehensive R Archive Network

CRM – Customer Relationship Management

ERP – Enterprise Resource Planning

GA – Google Analytics

GATC – Google Analytics Tracking Code

ICT – Information and Communication Technologies

KPI – Key Performance Indicator

MCF – Multi-Channel Funnel

OKR – Objectives and Key Results

PII – Personally Identifiable Information

PaaS – Platform as a Service

ROI – Return on Investment

SEO – Search Engine Optimization

SaaS – Software as a Service

Figures

Figure 1 – Usage of web analytics in global % (W3Techs Inc., 2013)	10
Figure 2 - Google Analytics platform components (Google Inc., 2014b)	11
Figure 3 - Google Analytics Mobile Application.....	20
Figure 4 - Administrator view - Goal setting.....	22
Figure 5 – Types of customer life cycle funnel (Waisberg & Kaushik, 2009).....	25
Figure 6 - Tatvic Excel dashboard	36
Figure 7 - Basic R environment and R Studio.....	37
Figure 8 - Supervised Learning for predictive models	39
Figure 9 - Levels of access in GA	47
Figure 10 - Sessions per day and basic indicators	48
Figure 11 – Customized alerts	49
Figure 12 – Alert for an increase in traffic with visitor type and source	50
Figure 13 – Distribution and interquartile range of transactions by country (using R and the API)	52
Figure 14 – Returning (blue) and New (orange) users per Number of sessions; % of engaged visitors (Page views >10); and Daily revenue	54
Figure 15 –Poor performing CPC campaign: 98.5% drop offs before the first interaction	56
Figure 16 – Path from organic to internal search on the 1 ST interaction	57
Figure 17 - Analytics keywords report and Google trends for the term “Redcorp” (12 months).....	68
Figure 18 - Weekly % of new visits, from 66% to 82%; and Indicators for organic new and returning visitors.....	70
Figure 19 -Facebook Ad Manager metrics.....	71
Figure 20 - Page speed suggestions for the default.aspx page	82
Figure 21 - Page loading time for non-mobile (compared to 3.21 average).....	83
Figure 22 – Page Analytics extension – Click rate for the “Monitors and Displays” section.....	87

Figure 23 - Percentage of cumulative revenue by the value of each transaction, with observation 2084 (of 2789 – 3 rd quartile) at only 25% cumulative value (data from the API)	92
Figure 24 - Goal conversion rate for All Goals	93
Figure 25 - Conversion funnel for the Order process flow	95
Figure 26 - Conversions and % value for top conversion paths	98
Figure 27 – Configuration of custom attribution model	99
Figure 28 - Distribution of transaction value for the 1 st and 2 nd periods (one and two).....	105
Figure 29 - T-test for Log transformed transaction value for the two periods	106
Figure 30 – Distribution of engagement and value variabes.....	110
Figure 31 - Q-Q plot for the residuals of Model 1 and 2	113
Figure 32 – Logarithmic variables distribution	114
Figure 33 - Model 4 diagnostic plots	115
Figure 34 - Diagnostic plots for Model 4	118
Figure 35 - Distribution of differences between predicted and actual value in absolute and % difference	119

Tables

Table 1 - Google Analytics Tracking cookies	15
Table 2 - Custom User ID dimension for test website using universal analytics	17
Table 3 – Automatic intelligence alerts	49
Table 4 - Indicators for the 3 main revenue-generating countries	51
Table 5 – Top ten cities outside Belgium	53
Table 6 – Correlation matrix for the effect of returning visits on the nr of visits, goal 8 (page views >10 per session) conversion and revenue	55
Table 7 – Revenue and sessions for the “Universite Catholique de Louvain” for the two main Operating systems	58
Table 8 - Top 3 countries between the two periods	60
Table 9 – Sessions for both periods in the top 3 countries	61
Table 10 -Traffic sources per number of pages per session	69
Table 11 - Indicators for the linkedin.com source	71
Table 12 - Acquisition source per generated revenue	73
Table 13 - Page views and average load times by country and device	84
Table 14 - Site search usage per source	86
Table 15 – Reverse path for order placements	93
Table 16 - Assisted conversions report for ecommerce transactions	97
Table 17 - Model comparison tool by mediums	100
Table 18 - unique transaction revenue and quantity per item	103
Table 19 – Correlation table between value and engagement variables	109
Table 20 - Correlation of variables for users' buying sessions	114

Formulas

Formula 1 – Statistical test for comparing proportions	60
--	----

Models

Model 1 - Coefficients for Linear Regression on session value	111
Model 2 - Linear model for Session value w/ transformed response variable	112
Model 3 - Linear Regression and diagnostic plots for users' buying sessions..	114
Model 4 – Linear model for channel revenue using the train subset	117

1 Introduction

Web analytics is often defined as the simultaneous combination of science and art of improving the performance of websites (Waisberg & Kaushik, 2009a). That is because while it is true that statistics and data mining techniques are used to explore the pallet of multiple data sources, it is also required to have deep levels of understanding and creativity, in order to not only interpret the data but provide the appropriate responses and drawing meaningful insights from the data. Furthermore, in order to develop users' online experience, we also have to deal with different stakeholders inside or outside the company, from designers, IT technicians, managers and, of course, our visitors and customers. Meeting all expectations is in this way a challenging experience, given the multiple parties involved in the website and content design and utilization. The job of a web analyst is however also to motivate the change and emphasize the contribution of each for the improvement of user experience (Kaushik, 2010b).

Therefore, web analytics is nowadays an essential monitoring tool, given the increasing importance of companies being online. This allows for global access to different publics, but also bears great impact on their image. The development of online content must therefore be carefully considered, following clear strategies and goals. Otherwise, inadequate content and campaigns can quickly disseminate a poor image of our company and affect other areas of the business beyond the digital world. In this work we are going to be looking into all the variables that help us assess website and business performance, starting by defining the ambit of web analytics, the technologies required, as well as discussing basic concepts. We then move into the analysis of a case study, going through the reports, dimensions and metrics. For each section we consider an introductory explanation of the reports, followed by an application of the terms in an analysis for the period between the 13th of January and the 30th of March, closing with a summary of the conclusions for that period. We then look to corroborate some of those assumptions by looking into a second period, from the 31st of March to the 29th of June, assessing the significance in the differences between proportions (*e.g.* conversion rates). Lastly, some exercises with R present us

with regressions exploring the role of different dimensions and metrics, and their contribution for explaining revenue.

1.1 Why we need web analytics

Web analytics are one of the most important marketing monitoring tools for companies that have online presence. In order to comprehend their visitors' experience and the way they navigate through web pages, it is fundamental that we have techniques for data collection and methodologies for its analysis. Only then we can get insight into customers' experiences and assess the relevance of our strategies for the business. Clifton (2012), in this sense recalls the premise of the XIX century scientist Lord Kelvin, who stated that only by measuring we can improve. This stream of thought thus reflects the spirit of web analytics and the advocacy of a scientific approach towards data. Data collection is therefore only the first step for obtaining insights, with a distinction between data and information. Data thus needs to be structured and interpreted, according to proven methodologies. There is in this sense a subjacent process, with goal in the enhancing of the organization's competitiveness, transforming knowledge into actions and giving organizations the tools to respond to environmental changes (Delen, 2013).

Because of this, many organizations are already putting web analytics to use, whether we are talking about public or private companies, governments, NGO's or personal pages. This is an increasingly common procedure across the web, with the scope of analysis varying from organization to organization, depending on the objectives of each web site and page. In this way, monitoring can vary from the visualization of simple metrics (*e.g.* number of daily visits), to a more profound and complex analysis which seek to understand more specific behavioral patterns (*e.g.* the reason why some ecommerce visitors fill their e-shopping carts, but never really purchase any products) (Pakkala, Presser, & Christensen, 2012). However, in order for the web analytics process to be effective, we first got to define the objectives for the website, its sections and the type of interactions we want our visitors to engage. In other words, we justify the existence of each digital material, defining what success

looks like in each scenario. Analytics then helps monitoring and assessing, aside from in-site sessions, the effectiveness of campaigns, sources of traffic, social and mobile interactions or changes made to the website.

This is in fact a technology of great potential in different areas, not only in marketing, sales and advertising, but also for managers, public relations or communication professionals, planners and strategists, assisting the development and follow up of contents. In this sense, the utilization of analytics is not merely destined to the measurement of commercial success, where sales are invariably the main goal (Kent, Carr, Husted, & Pop, 2011). Contrariwise, these packages offer the possibility of measuring a wide range of behaviors, suiting the needs of different organizations. Goals can thus be defined according to any metrics available, which aim to reflect different types of involvement with the website. Contrary to traditional marketing, it is now possible to accurately know the number of visits one ad generated, the time people spend reading an article or the main landing and exit pages. These examples, when contextualized, reflect people's responses to the pages, helping us improve and meeting the expectations of visitors. Because of that, this is an extremely promising tool, to easily gather great amounts of data and a great variety of indicators beyond the results of sales, without having to resort to time and resource consuming techniques such as large scale surveys.

According to some studies, in the USA for example the rate of ecommerce conversion for most websites oscillates only between 1% and 3%, which reflects the relatively small proportion of sessions in which transactions actually happen (Clifton, 2012a). What this tells us is that the reasons for accessing websites may vary widely, and is now up to us to adopt a proactive and critical attitude, seeking to interpret the data and the impact of our own actions – online or offline – in our business.

Traditional marketing and web analytics are thus part of the same continuum, permanently influencing each other. The advent of internet and web 2.0 thus contributed to the profound change of dynamics in the interaction between people, not only with companies, but especially between themselves (Balamurugan, Vasuki, Angayarkanni, & Aurchana, 2013). Hence, organizations now have to attract the interest of online communities, through the utilization of different, yet consistent,

digital strategies and materials. In this way, the digital environment is fertile ground for the emergence of new business and communication strategies, such as search engine optimization (SEO), blogging, social media, news feeds (RSS) and others (Miletsky, 2010). There are in this way many different forms of interaction with customers and stakeholders, which highlight the importance of the creation of meaningful contents, associated with ease of access, navigation and speed of connection.

The analysis of web trends may also assume two different perspectives, including off-site and on-site analysis. Off-site analysis in this way refers to the investigation and data collection across the Internet, regardless of the property of domains. Here we aim to collect relevant data for our organization transmitting us information about the size of our potential audience, visibility and share of voice of our organization or the buzz generated around a specific theme, product or action (Balamurugan et al., 2013; Clifton, 2012a). On the other hand, the utilization of on-site tools refers to the behavioral analysis of visitors within the boundaries of our own domain. This is the main scope of this work, consisting in the first-party collection, treatment and interpretation of data. With this methodology, we aim to evaluate the utilization that is given to our website, as well as answer questions related to the strategy and effectiveness of contents and campaigns (Balamurugan et al., 2013; Kent et al., 2011; Pakkala et al., 2012). Some of the most common questions are :

- Where do our visitors come from, what are the main paths they follow and how do they exit our website?
- Which are the contents our visitors are most interested in and are they finding what they are looking for?
- Do visitors find our contents relevant?
- Are we acquiring and engaging visitors?
- Who are our users' and how do they access our contents?

1.1.1 Levels of analysis

Different service providers and vendors may offer different functionalities, with different analysis techniques adapting to each organization. On-site analytics are,

however, considered as the more appropriate and ethical source of information while preserving anonymity and still contributing to the improvement of websites towards customers' expectations and the fulfilment of the organization's goals.

Delen & Demirkan (2013) in this context highlight the existence of three main analytics categories, including **descriptive**, **predictive** and **prescriptive** analysis. From this point of view descriptive reporting represents the starting point for any analysis, using data and reports to identify latent problems and opportunities. This is, as the name suggests, a descriptive phase in which we try to answer to the question of "what is happening?". In this phase, the analysis is mainly based on reports, dashboards, scorecards and other types of structured data. The main goal in this phase is to systematically define business problems and faults, as well as to identify latent opportunities where they may be margin for improvement.

On the other hand, predictive analytics step up the complexity of analytics by using mathematical techniques, such as statistics, to identify relationships and patterns between the variables. In this phase, we try to evaluate the impact of one variable over the other and the occurrence of future events. Hence, different conditions might be hypothesized to the explanation of a given outcome. Techniques such as data mining are therefore one of the main enablers of predictive analytics, which aim to anticipate the impact of different scenarios.

Lastly, prescriptive analytics are the natural consequence of the analytics process, culminating in the prescription of the best possible solution for a given problem. This category also relies on modeling techniques, the combination of data and expert knowledge in order to provide decision makers with the richest possible information for them to take the best course of action.

1.2 Data-driven organizations

The notion of data-driven organizations is a concept that goes beyond the sole use of web analytics. Avinash Kaushik, Google's evangelist and web expert, many times refers to the importance of people over tools, which reflects the role of knowledge and creativity as key components for interpreting and overcoming challenges. Therefore,

having a powerful web analytics system will only be helpful if the basic pre-requisite, of having a skillful and motivated team is met. Burby & Atchison (2007) identify a role of characteristics successful data-driven companies systematically present:

Firstly, companies with a strong analytical culture drive their decisions in accordance to business goals. Their numbers must therefore be interpreted under the light of a context, aiming at specific objectives. So defining what are the relevant metrics and with whom must they be shared with is one of the first steps for defining an adequate strategy. In this context some of the relevant metrics for each type of business model are discussed ahead in this work. However, even within the same company, different types of indicators might be relevant for different people or departments. A communication strategy must therefore be defined, for information to be pertinent and actionable for those who receive it.

Furthermore, data-driven organizations also base their decisions on educated premises and facts, rather than on feelings. In this way, experience is always important, but tradition should not be pretext for ignoring the analytical point of view. On the contrary, these should complement each other and organizations which can make the most out of both will certainly gain a competitive advantage. It is consequently important to acknowledge that experience is a subjective concept and that different people have different opinions and skills. By resorting to web analytics, organizations can objectively assess the critical areas of success for their online business and justify investment in the right goals at the right time. For this, after the definition of key areas, we should then try to define the key metrics to evaluate them, tying indicators to specific outcomes. In that way, when variations occur, we know what reactions to expect, the areas to invest in and the consequences that can be expected. Changes should thus aim to improve conversion rates and maximizing the ROI of our initiatives.

Another highlighted characteristic of data-driven organizations is the set-up of teams to operate under the same system of indicators. For this, it is first necessary to have a global set of goals, which will then successively boil down into the whole company. That way, every employee can have the same frame of reference, knowing that they are individually driving global success. Additionally, if we target specific

indicators for our teams, we should also segment audiences, customizing variables and adapting webpages to different needs. Looking merely at aggregate data is in that sense often misleading, as often niches are of major importance for the business, exhibiting different behaviors from the bulk of sessions. Managing expectations thus drives our conversions in a much more fruitful way for both parties. This helps us focus on objectives and improve on relevant areas.

The web analytics process therefore responds to a methodology of implementation, where we define the ambit of our work with tangible benchmarks. This process is widely addressed in literature (Burby & Atchison, 2007; Clifton, 2012a; Kaushik, 2010b), raising fundamental issues organizations should be aware of when developing strategies. With the multiplicity of variables and the amounts of data, our problem is nowadays to select the most relevant sources, being able to discard superfluous information.

1.3 Data collection methodologies

Methodologies for collecting visitors' data may also vary in extent, complexity of implementation and the vendor who provides the service. Different companies have different needs and must ponder between the existing options. These are systems which require commitment and investment, representing a consistent practice over time with properly defined strategies and indicators. Changing service providers will thus result in changes to the whole structure of analysis and to the process as a whole (Kaushik, 2010b). That is because different methodologies offer different features, specific to each technology. With few exceptions, data is generally not interchangeable between providers, leading to loss of (all) information when services are replaced. Among the different methodologies, there are however two which stand out as the most popular, which will be analyzed in this work.

In this sense, the first method is the analysis of *Log Files*, which refer to files of information automatically stored in web servers. These register events related to our visitors activity and are frequently referred to as a server-side operation. This was, technologically speaking, the first method to appear. On the other hand, *Page Tagging*

is nowadays the most common method for gathering information on the web, typically associated with Software as a Service (SaaS) and client-side storage of information. With this type of model, software and data are remotely accessed via a web browser, with no locally-stored information (Pakkala et al., 2012). Software is usually hosted in the cloud by a service provider, which guarantees a remote easy access to all its users.

Both methodologies present different advantages and disadvantages, requiring different levels of involvement and resources. Nonetheless, we today have a number of free options on the market, which guarantee access to web analytics tools for all organizations. Google Analytics is the best example of this type of services, with a freely available set of tools and a large community of active users contributing and discussing common problems and solutions. It is however necessary to commit to these tools, in order to fully understand their potential for improving the organization's performance.

1.3.1 Log files

The analysis of log files is a method which allows the collection of data without the need for an external service provider. This was the first and more traditional method for analyzing online users' behavior, since all web servers keep track of users' records by default. Moreover, whichever might be the visitors' browser or add-ins used, data will be collected and stored on the same network as the web server. This is an automatic process and no changes are needed to the web pages. Because of that, this is designated as a server-side method, since the entire process is solely depending on the server's gathering and storage of data. With this methodology, information is stored in the format of text files, which implies the development of mechanisms of analysis, as highlighted by Kaushik (2010).

One of the main obstacles is therefore the requirement for specific IT knowledge in order to develop appropriate systems of analysis and for guaranteeing we are able to obtain insights from the data. These however are valuable resources not all companies can afford to commit to this task. It is also necessary to have physical devices for data storage, which in the case of some websites might pose a significant

challenge, given the great amounts of inbound traffic. Paradoxically, this is also one of the main advantages of this technique, since the permanent existence of raw data allows for it to be reprocessed and reanalyzed at any time by different systems.

Nonetheless, Clifton (2012) emphasizes the limitations of *log analysis*, with cached pages by search engines affecting this method's precision. These prevent the direct interaction of visitors with the original website server, deflating the real number of interactions. Furthermore, since cached pages are associated with relevant content, missing observations from these pages might be of great importance. These visits are as a consequence excluded from our analysis, affecting its credibility. Contrariwise, bots such as search engine crawlers are also unrealistically counted as visitors, inflating the number of sessions.

Lastly, it is also impossible to track events which make use of interactive web languages (such as *flash*) using *log* files. This is because these do not generate page views, rather they refer to the use of interactive dynamic content. *Log* files in this context narrow our access to the users' experience, limiting our perspective of certain sections of the website.

1.3.2 Page Tagging

The Page Tagging methodology consists on the introduction of a *JavaScript* tracking code to a page in order to collect data about the activity on our website. Via the user's web browser, information is sent to remote data-collection servers provided by our web analytics vendor. This is therefore known as client-side data collection, since all the information is stored by the visitors' browsers in small text files known as *cookies*. These might also be characterized into different types, including session cookies, which are automatically deleted after the browser has been closed, or persistent cookies, used for later identification of the characteristics of each unique user. Apart from web analytics, cookies are also used to offer personalized services and pages. One example of that is the implementation of online shopping carts using (session) cookies (ICO, 2011).

For web analytics, the importance of cookies resides mainly in the anonymous identification of users, with best practices dictating the importance of using only first-party information. That is, information created and requested directly by the visitor to a particular domain. Another important characteristic of cookies is that they are harmless to the user and can be deleted anytime the user decides to do so.

Clifton (2012), in this sense mentions the popularity of *page tagging*, with about 90% of websites which collect data resorting to this methodology. Among the most popular services, Google Analytics stands out as the number one vendor, with a market share consistently over 80% among all the traffic analysis tools available for websites. That is about half of all websites online (W3Techs Inc., 2013). These are expressive numbers which reflect the popularity of Google Analytics and the comprehensive, agile functionalities offered which allow for an in-depth analysis of trends and variables, all free of charge. Furthermore, the ease of implementation and maintenance of such a powerful tool, associated with the Google brand, makes it the most recognized web analytics application. When compared to *log analysis*, the monetary advantages of *page tagging* also become evident, with constant updates developed by the vendor and not the company itself. Furthermore, as we will explore in this work, tools such as GA now possess powerful customization options, allowing for the creation of new variables, reports, dashboards or segments.



FIGURE 1 – USAGE OF WEB ANALYTICS IN GLOBAL % (W3Techs Inc., 2013)

All the information is remotely accessed via browser, while the data is stored and processed in the vendor's servers, excluding the need for the physical storage of

logs or the development of analysis systems. This would otherwise require a great deal of investment, proportional to the increasing in the number of visits. All monitoring and development costs of storage are thus eliminated, leaving only corporate teams the task of analysis. As we can see from the following scheme of the GA reporting structure, from the moment of collection to the reporting of information, the data goes through a process of four stages before reaching the end-user: **collection**, **processing** and **reporting**. All these happen remotely, without the interference of the server, being made available through the GA application, which in may be accessed either via the browser or the GA mobile application.

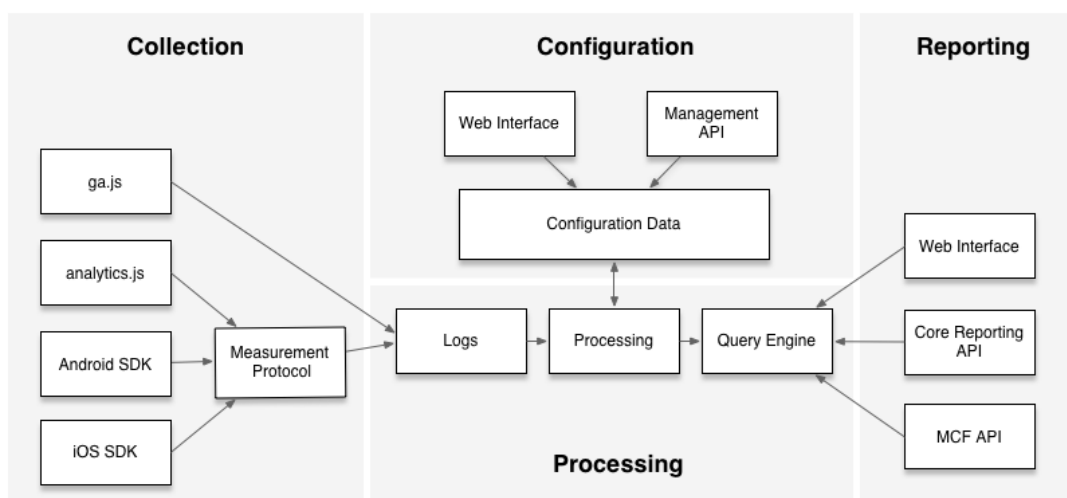


FIGURE 2 - GOOGLE ANALYTICS PLATFORM COMPONENTS (GOOGLE INC., 2014B)

1.4 Privacy issues

There are several issues arising from the utilization of internet services regarding the respect for private information. With the existence of just a few large corporations dominating a large share of services such as e-mail, search engines or social networks, it is clear that privacy becomes a more and more relevant matter of discussion. Companies such as Google, which possesses a great variety of products and services directly related to people's information, dominate the information economy and are increasingly involved into multiple dimensions of our lives. From our computer screens to phones and tablets, we become traceable at each step, with multiple aspects of our lives stored in the digital world (Ascensão, 2011). Never before have the real and the digital worlds been so interdependent, which raises the topic of the

importance of security and the extent to what information can be used against users. Recent events such as Snowden's case, make us question the ethics of the most powerful institutions in the world and people's right to privacy. There is therefore a discussion on the degree of exposure people are willing to commit, with the need for boundaries between quality of service and personal space.

In this sense, we are nowadays permanently connected to networks, emitting signs of our presence to the grid. In spite of conscious of this, we are getting used to customized services and real-time offers, with companies giving us the comfortable feeling of personalized care. Information is being shared but not only through the internet, but also mobile phones, GPS systems or even ATMs. It is thus possible for organizations to keep track of our records, being possible to link our identities to our actions. Guaranteeing the safety of people's information and right to privacy is therefore a major issue for governments nowadays, and companies must be obliged to guarantee a degree of anonymity in data, while still being able assure the quality of their services. Calabrese (2013) for example refers the utilization of anonymized mobile phone data for the improvement of public transportation in Abidjan, Ivory Coast, where about 70% of its 4.5 million inhabitants own a mobile phone. Spatiotemporal signals were in this case used to improve the routes of the city's overcrowded network of buses, with information favoring all citizens.

Databases are therefore an unavoidable part of the information world, present in a large part of our daily lives. This makes it necessary to guarantee that it is put to the people's service and not only for companies profit. Decuyper & Blondel (2013) in this sense discuss the paradox of information significantly improving our quality of live, while making it very easy to be followed or deceived, emphasizing the importance of data confidentiality.

The use of analytical tools on the web is in this sense also a powerful way of collecting data, which can be used to link a specific person to an equipment (such as a mobile phone), and explore their behavior. It is therefore necessary to establish appropriate rules of conduct and boundaries for these experiments. With its emergence, people are also starting to restrict webmasters' access to their information, by deleting cookies, using firewalls or downloading browser plugins,

preventing scripts from running. Different browsers, such as Mozilla Firefox or Google Chrome, now have a vast community of developers creating their applications sharing them through the browser's platform. Just a few examples are *AdBlock*, which identifies and prevents unsolicited publicity and ads from being loaded, *NoScript*, which allows us to restrain and select the scripts we want loaded, or *Block Yourself From Analytics*, which targets GA preventing the execution of the tracking code.

This can bring serious consequences to companies, including to traditional business models. Blocking advertising and access to certain kinds of information may affect the performance of a website in the long-term, the attainment of business goals (*e.g.* revenue generated by publicity) and consequently the own user's experience and satisfaction. For a long time the collection of data has been lacking specific regulation, but since 2011 new laws came into effect in the European Union, which went beyond the previous document, the Privacy and Electronic Communications Regulations of 2003, clearly defining the ambit of data collection and cookie utilization. These new rules aim to protect users' privacy, primarily targeting websites sharing third party information, identifying anonymous visitors or collecting data in spite of the visitors' consent. This law refers not only to cookies, but also to the use of other non-transparent tracking systems.

In this sense, the use of cookies for tracking behavior now requires consent from visitors for all the websites within the EU, which foresees that information is clearly and comprehensively provided about the purposes of storing data. However, there is a distinction between implicit and explicit consent and in the type of cookies being used. Only in certain circumstances is it necessary to obtain explicit consent, when more sensitive information is involved. Whichever the case, it is now mandatory that all websites in the EU inform visitors about the methodologies in use and the treatment being given to the information, with few exceptions to this rule. Some categories of cookies are also considered more benign, such as anonym session cookies or for determining the user's preferences (*e.g.* language settings), for shopping or to improve the user experience. First party cookies are in this context the only acceptable source of data, assuring users the anonymity of information (ICO, 2011)

The way of obtaining consent is however still a dubious question, since it is not mandatory to obtain explicit consent. Nevertheless, this information must be at the disposal of visitors, in the form of an easily accessible web page. Even so, browser configurations can also be perceived as an implicit sign of the user's will, since it provide the option of restricting cookies and the degree the user is willing to share information. The ICO also considers that in the case of analytics cookies, implied consent "might be the most practical and user-friendly option" (ICO, 2011, p. 9).

Regulation over the internet have always been problematic and in this case we must also have to take in account the fact that most organizations use analytics not with the purpose of following users, but to improve business performance and the online experience of users. Furthermore, some of these rules might also require interpretation, leaving room for misconceptions. As it is mentioned in the ICO document, information may only be collected if "strictly necessary", unless consent is provided. However, Clifton (2012) argues that many times users don't completely understand these mechanisms, and if explicit consent is necessary, they will simply deny access to the data, since it is the easiest, safest thing to do with no immediate repercussions. This may however bear severe consequences for businesses and affect users as well in the mid-term. The internet is in itself a fast environment, where people will not bother reading complicated regulations or thinking about their impact on business. Even so, monitoring variables has always been an essential component of any corporation, long before internet. Even governments and non-profit organizations need to be accountable and need to have their own metrics to respond to the changes in their environments. Therefore, collecting data is not a new activity and the balance must be found between the right for people's privacy and the necessary tools for organizations to continue improving their services.

In the scope of this work, it is also worth mentioning Google's policy aims to preserve visitor's privacy. This means only first-party information is used in the processing of data and no external sources of information will be considered for any of the metrics. Furthermore, there is no collection of personally identifiable information (PII), which means all data remains anonymous. The value of metrics thus comes in a somewhat aggregate form, with no directly attributable action to any of the site's

visitors. In order to drill down into the data in GA, one must use different dimensions in order to create segments, with no metrics are strictly associated with any particular visitor. The protection of personal privacy is in this sense one of the main concerns of GA's terms and conditions (Clifton, 2012a; Google Inc., 2014).

First-party cookies are used to distinguish unique visitors, domains, determine the start and end of a session and remember variable values from previous visits. Most of these are persistent cookies, meaning they endure beyond the duration of a session and are updated every time data is exchanged with GA. The Google Analytics Tracking Code (GATC) might also be customized, in order to define a domain name, campaign or set expiration limits for the acquisition of users. The default configurations for classic analytics (*ga.js*) are as follows (Google Inc., 2013):

Cookie	Expiration	Usage
_utma	2 years	Distinguishes users and sessions;
_utmb	30 mins	Determines the beginning of new sessions;
_utmc	End of session	Used in interaction with _utmb ;
_utmz	6 months	User's campaign and source information;
_utmv	2 years	Custom-variable data;

TABLE 1 - GOOGLE ANALYTICS TRACKING COOKIES

1.4.1 "User ID" dimension

One of the most relevant issues raised by the collection of non-PII information only on an aggregate form is that it hinders the possibility of companies knowing their visitors at the user level. In this way, the great majority of dimensions give us access only to the overall picture through the mean, absolute or percent values for segments of users according to dimensions related to time, actions, marketing channels, types of user or other aggregate data. At the user level it is very difficult to extract some individualized insight, seeming at first that this would be against Google's user policy.

However, GA data has for a long time been used for integration with other third party applications, such as CRM systems (Clifton, 2012), where PII is available for analysis. Most of the times we are not however interested in one particular user, but to understand the interaction of visitors with our website, regardless of their identity. While the existing dimensions can give us access to the overall picture, the fact is that extreme behaviors and particularities are often lost due to the inexistence of a user dimension where we identify unique visitors (in *ga.js*). Correia (2010) in this sense programmed a PHP class which can be used by *ga.js* (classic GA) to extract human-readable information from cookie data in order to integrate it with third-party proprietary systems, such as CRM or ERP.

With the launch of a new GA version in late 2013 (*analytics.js*), a new User ID feature was launched with it, which gives as much more accurate insight into each user's experience across multiple platforms. This feature is intended primarily for integration with the websites' authentication system, enabling us to differentiate between the users which log in to the site and those which don't. This is in this way a very useful feature since these are two very different groups of users.

It also opens the possibilities for the development of even greater functionalities, such as the customization of non-PII user ID dimensions given by Simpson (2014). This is an example presented as the new analytics version was being rolled out, illustrating the potentialities that some users have been trying to develop themselves. In this way, this developer uses the "custom dimensions" feature in order to create a new dimension to store a randomly generated code, attributed to each visitor. The scope of this dimension is thus defined by "User", in order for a code to be attributed to each unique visitor. This will result in the creation of a dimension for each unique visitor, without however retaining PII. Simpson (2014) however also created a chrome extension which enables the integration of PII in this dimension, using third-party data, which we will however not explore here.

The following chart is an example configured for a personal experimental website using universal analytics (*analytics.js*), which instead of log-in identification uses random but unique cookie ID. However, for websites with heavy traffic it would

be advisable to use users' accounts, otherwise we would have an extremely high amount of dimensions, resulting in un-actionable information.

Custom User ID ?	City ?	Sessions ? ↓	Pageviews ?	Total Events ?
1. 1836331545.1405589418	Castelo Branco Municipality	3 (20.00%)	3 (2.31%)	2 (1.45%)
2. 1102026378.1405798569	Lisbon	2 (13.33%)	2 (1.54%)	0 (0.00%)
3. 1836331545.1405589418	Coimbra Municipality	2 (13.33%)	5 (3.85%)	10 (7.25%)
4. 1696797312.1406584787	Covilha	1 (6.67%)	1 (0.77%)	2 (1.45%)
5. 485046386.1406818379	Lisbon	1 (6.67%)	2 (1.54%)	6 (4.35%)

TABLE 2 - CUSTOM USER ID DIMENSION FOR TEST WEBSITE USING UNIVERSAL ANALYTICS

2 Google Analytics as Software as a Service

As we have been discussing, *page tagging* is the most important web analytics methodology, with the underlying component of a service provided by a vendor. Much of this popularity derives from the comprehensive offer by Google Analytics. Through this type of service, Google provides a free web analytics service, which depending on the user's experience and needs, might be configured to attend specific issues through the customization of the tracking code, campaigns, reports and other elements. Beyond that, there is also the additional option of using the API for integration with other software, retrieving the metrics' values by using queries. Further ahead in this work, we will be using the Core Reporting API for automatically exporting values to the statistical software R.

The utilization of page tagging methodologies therefore eliminates the need for possessing local hard drives for storing the data or purchasing, developing and updating software for managing the retrieved information. All of these tasks are on the contrary entirely assumed by the web analytics platform, where all the data is collected, processed and virtually delivered via web browser. Services such as Google Analytics are thus related to the concept of cloud computing and the utilization of remote services and virtual environments as a platform for the delivering of software. Cloud applications are thus subjacent to a model of computation, of storage and communication for the data collected. Scaling and availability are two of the main

benefits associated with cloud services, providing the automatic allocation and management of great volumes of data.

Sultan (2013) therefore emphasizes the importance of business model of cloud services as a “pay-as-you-go” structure of pricing, which represents an advantage when compared to the traditional model of software distribution. In this sense, large sums of investment were traditionally associated with most business applications, not only in their installation, but also the maintaining and upgrading of features. Cloud computing thus provides companies with the opportunity of taking advantage of continuous upgrades to their systems, without the need of such investments. Capacity on demand also provides smaller businesses with the opportunity of adapting budgets to their needs, in a cost advantageous business model for organizations which now can take advantage of new technologies at more affordable costs. Armbrust et al. (2010) also point out the flexibility of this system, offering us the possibility to pay for the utilization of short-term services, such as the utilization of greater storage capacity during periods of higher necessity. This contributes for the delineation of cost-efficient strategies while eliminating the risks associated with the commitment required by traditional platforms. Among the main vendors which currently provide cloud-based services for business are Google, Microsoft, IBM, SAP or Salesforce, including solutions for different areas such as CRM, ERP, HR or other information systems.

Within the scope of cloud computing, we can however find different definitions for addressing different purposes. Clouds in this sense comprehend two major components, which consist of the data center software and the hardware. Software as a Service (SaaS) in this context refers to the providing of software from remote sources and its major contributions derive from both the reduction of costs with IT, but also accessibility from multiple mobile devices or via web browser. Different vendors and researchers however refer to other services, contextualizing the ambit of areas such as IaaS and PaaS – Infrastructure and Platform as a Service. In the first case (IaaS), we talk about the storage and processing of information, when remotely provided from the vendor's data centers. On the other hand, PaaS include the offering of development tools which allow users to create their own applications on top of pre-existing layers of software and hardware, according to their requirements and eliminating the need for

maintaining the entire system in which it is based (Armbrust et al., 2010; Sultan, 2013). Villegas et al. (2012) thus describes the conceptual architecture of the cloud as a layered model where IaaS represents the base of this structure, followed by PaaS and finally SaaS, which is where the requests to the bottom levels take place.

On the other hand, traditional software and hardware systems are much more inflexible in this sense and require much more commitment and pondering before being acquired. Because of this, there are three major situations pointed out by Armbrust et al. (2010) which clearly illustrate the usefulness and the unquestionable logic of cloud-based services:

First of all, there are many occasions when demand may vary over time for a service. This situation may lead to under or over-usage of traditional data centers. Resorting to on-demand services we are therefore able to gain access to more flexible computing resources for specific periods of time. Secondly, it may not be easy or clear for companies to anticipate their own demand properly. For example, if a new product is created or a new web page put online, there may be initial periods of great activity online, which can be reduced over time. Cloud computing allow us to respond to this initial demand, without us having to support these additional costs when larger capacity is no longer needed. Finally, expenses that would be made in a specific occasion can be distributed over time, as we use different services, allowing us to make options and redirect capital to other areas of the business when needed. We therefore overcome risks of over and under provisioning, keeping a lean structure of costs.

This is, as Sultan (2013) points out, a new kind of disruptive technology, which is already changing the way organizations are allowed to store, process and access information. Through the existence of public (the Internet) and private networks, a wide range of services can now be delivered and tailored to the users' needs, without the need of installing software or maintaining databases. Google Analytics may therefore be considered a SaaS application, where all software and associated data is hosted remotely. This is as we already know, an application accessed via browser or by mobile, which grants real time access to all business information.



FIGURE 3 - GOOGLE ANALYTICS MOBILE APPLICATION

Cloud computing is therefore a growing opportunity for companies to develop their business, allowing them to be constantly up to date, with systems tailored to their requirements. Since the installation and maintenance of software and hardware systems is now at the responsibility of vendors, organizations are now only responsible for the selection of their provider. Marston, Li, Bandyopadhyay, Zhang, & Ghalsasi (2011) also point out the role of corporate users as key for defining the terms, features and the regulation of cloud based services. The evolution of these services is thus propelled by its increasing utilization, with providers assuming the introduction of new features.

2.1 Core concepts and metrics

In order for us to comprehend the ambit of web analytics, we need to define some of the basic concepts associated with the metrics and dimensions for analysis. These issues are here discussed in order to help us define a working framework, as well as understand the indicators contributing to our studies. Metrics are associated to dimensions, which allow for the segmentation of the public according to behavior, technology, demographics, date of visit, traffic sources, conversions and other parameters. Creating customized segments is also a powerful feature for the personalization of analysis, dividing visitors according to user-defined conditions.

In this sense, we start by exploring the concept of **visit**, which in web analytics refers to an access made by any kind of identifiable device to our website. Also commonly referred to as sessions, each visit consists on the number of requests made by the same identifiable user over a period of time. In this way, it is necessary to define the terms for an end of a session, with different services adopting different postures. Since many times visitors leave open browsers and tabs, in most cases the last page viewed by visitors is not considered for session length. Moreover, long inactivity periods dictate the automatic end of a session, which in the case of Google Analytics represents a default value of 30 minutes without any interaction (Kaushik, 2010). If no request is made during that time, the session will thus be automatically ended. Other important indicators associated with sessions include the **visit duration** or the number of **pages per visit**. These are some of the most commonly used metrics to help characterize the engagement of our audience. On the other hand, we can also look into the behavior regarding the analysis of each **page** of the website using the **time on page** metric as well as, its **number of visualizations**. Depending on the page's configuration, we can also have the associated interaction with other dynamic elements, such as Flash, videos or social media activity, which is generally linked to a triggered **event** (Fagan, 2013). Through this, we are not only aiming to track the volume of visits to our pages, but also the involvement of visitors with content.

One common mistake is, however, not to differentiate between the concepts of visits (sessions) and visitors (unique individual users). The number of **unique visitors** aims in this sense to quantify the approximated number of unique devices coming to a website, which then start accumulating visit counts. This is however a tricky issue, since page tagging methodologies depend on the user's persistent cookies for a clear identification of the device, as we had previously discussed. As a consequence, if our visitors are blocking their cookies, chances are the number of unique (new) visitors is being inflated, as opposed to the real number of **returning unique**. The interpretation of volume of new versus returning visitors is consequently biased by these limitations, as no indicator can confirm the veracity of the absolute number of each category of visitors. Rather than the absolute values, the percentage variation in relation to previous periods can therefore contribute to a more accurate interpretation of the

each channel’s evolution, revealing trends rather than facts (Clifton, 2012a; Gupta, Mehta, Bhavsar, & Joshi, 2013). The rate of evolution is therefore not only important in terms of the visitors’ type but also the rate of **goal conversions** when compared to different segments and fluctuations in total amount of traffic.

Goals are in this context user-defined within the GA environment, deriving from the organization’s digital strategy. Each conversion consists in the fulfilment of a set of behavioral criteria or the accomplishment of a specific action during a session. The most pragmatic example of conversion objectives is the occurrence of a sale, in relation to ecommerce websites, representing a highly measurable and useful indicator for assessing the ROI of campaigns. However, not all websites are dedicated to ecommerce, nor are all orders placed online. Non-transactional goals such as the engagement level are therefore key for measuring success, with different metrics and actions revealing the visitors’ level of involvement with online content. The duration of visits, visualization of a number of pages, downloading of materials, social media sharing or the subscription of newsletters are just a few of the actions which might manifest interest on the visitor’s part . Clifton (2012), also mentions the existence of negative goals, for which we want to minimize the rate of conversions. For example, if onsite search is an important part of website, we will want to minimize the number of null search results.

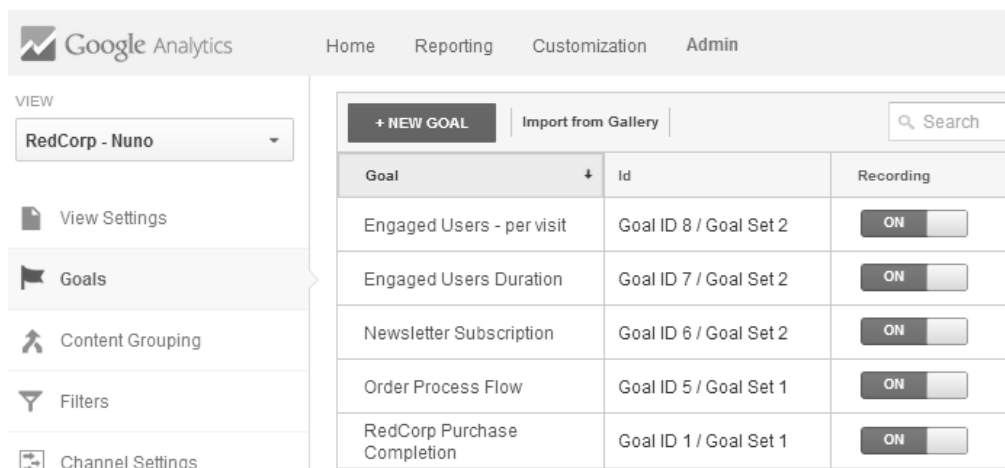


FIGURE 4 - ADMINISTRATOR VIEW - GOAL SETTING

Conversion rates are in this sense the main indicator of effectiveness for different campaigns and sources of traffic, in relation for example to the acquisition of

new qualified visits. In this context, we look to establish benchmarks for our goals, analyzing the precedence of visits, as well as the type of interaction with content. Lead generation is an important part of any acquisition strategy, with engagement levels providing an evaluation of adequacy of our content to the targeted audience. Furthermore, high engagement levels are generally a positive premonition for future sales in the case of ecommerce websites. With this, the relevance of **landing pages** might also dictate the difference between a future prospect and a **bounced visit** (single page visit).

In this sense, Studliffe (2012) and Allen (2012) illustrate the importance of page structure and design in order to attract the user's attention. Through the realization of eye tracking studies, these researchers were able to register behavioral differences between objective-oriented and browsing users. Moreover, the velocity of online environment often leads to short attention periods with certain types of elements frequently being ignored, such as large blocks of text. On the other hand, images, hyperlinks and different types of formatting attract users' attention and provide quick useful information, hierarchizing visual elements. However, it is not always possible to define a specific landing page, depending on the referral source, user's searching terms and other factors. Evaluating the main landing pages effectiveness according to each channel is however important, using indicators such as conversion and bouncing rates, required investment (if applicable) or associated revenue.

Different sources of traffic thus originate different types of visitors, from direct traffic users, who type the URL into their address bar or favorite pages, to organic results originated by search engines. There are also visits originated by paid advertising, of which Google AdWords is a paradigmatic example, as well as external sources from other websites, to which we call **referrals** (Kent et al., 2011). This last channel might in some cases be particularly important to work with because while paid advertising might contribute to a quick increase in the volume of visits, referrals often contribute with qualified traffic from websites on similar subjects and users on an ongoing search process. Moreover, link building is also one of the most important tasks in SEO, enabling us to work on the organic relevance of our website and page rank (Enge, Spencer, Stricchiola, & Fishkin, 2012).

In order to assess the effectiveness of each traffic channel as well as the content of our pages, Kaushik (2010) considers the **bouncing rate** as one of the “sexiest” indicators of performance. The reason for this is its straightforward interpretation, reflecting visitors’ lack of interest in the content displayed. The rate of bounced visits thus reveals our ability to retain visitors beyond the first interaction, an indication of engagement with content. In this way, it is expected that sources of traffic with higher percentage of returning visitors (such as the direct channel) also reveal lower bouncing rates. In this case, it may be appropriate to define segmented categories of analysis, according to objectives, level of investment and rate of conversions (*e.g.* isolating users from specific campaigns).

According to these perspectives, our pages’ **exit rate** also reveals the effectiveness of our strategy, with visitors’ exit pages contributing to the perception of users’ experience. The ending page of a visit might in this sense reflect the achievement or not of a conversion goal, with some pages associated to positive or negative exits. These are however relative values, depending on the context and combination with other metrics, such as time on page or number of pages per session. In the case of ecommerce websites for example, the main goal is often linked to transactions. An example of positive exit might in this case be an outgoing link to an electronic payment system, indicating a conversion for that visit. Contrariwise, if a high number of visitors is dropping-off along the conversion funnel that might indicate an inadequacy on any of the steps, needing to be reviewed. Besides ecommerce, non-transactional goals might also be associated with pages or actions, such as a download or a subscription.

In that sense, we therefore recognize the importance of analyzing **conversion paths**, which through the utilization of tools such as Google Analytics, allow us to assess the steps taken by visitors until they reach a conversion. This can also be referred to as the customer lifecycle funnel, since the number of visitors who initiates the process usually decreases at each step we get closer to conversion (Waisberg & Kaushik; Clifton, 2012).

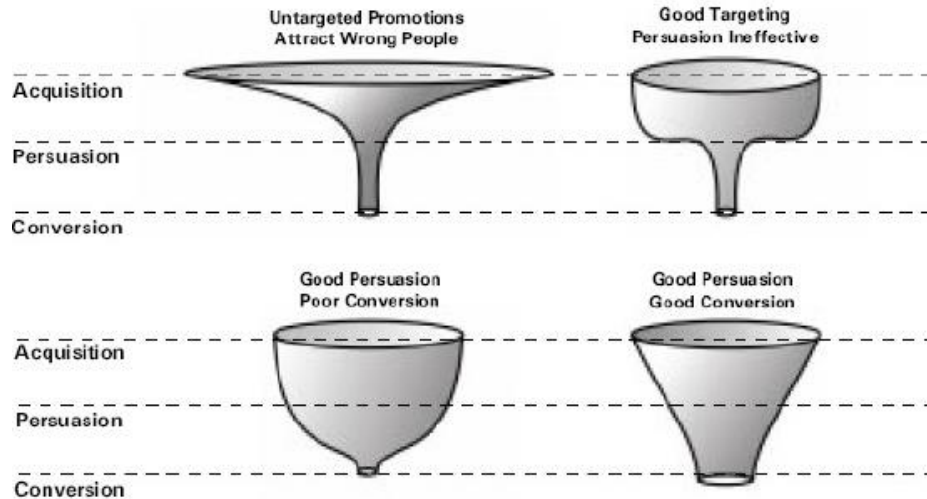


FIGURE 5 – TYPES OF CUSTOMER LIFE CYCLE FUNNELS (WAISBERG & KAUSHIK, 2009)

The main problem with attributing the source of conversion is however to ponder the contribution of all referrals in the process of conversion. Typically, most web analytics platforms only attributed credit to the last referral source, which is clearly a limitation and an over-simplification of reality. The theory behind it being that it may take various sessions for a visitor until they reach a conversion objective, for example to decide on a purchase. That does not mean however that only the last channel had influence on the user to make that purchase, but simply that the complexity of the whole process has been extremely reduced. The **multi-channel funnel** (MCF) analysis in this sense provides a much deeper understanding of the full referral path that led to a conversion, with reference to the various sources of online traffic. This analysis is however only possible through cookie identification, pondering each referrer's relative importance and the number of times a visitor accesses the website. These techniques also provide insight about most influential sources of information, helping us adapt our communication strategies and budget. Different reports and metrics in this sense contribute to the MCF analysis, among which we find the Assisted Conversions report and Top Conversion Paths in GA (Clifton, 2012), as well Visits and Days to Transaction in the case of ecommerce conversions.

Due to the high number of different metrics and the specificity of each business, the definition of indicators and segments must however be contextualized in the light of each organization. In order to help us in that task, GA interface offers over

one hundred default reports, combining metrics and dimensions, but also encouraging users to customize their profiles creating advanced reports, segments and dashboards. Hines (2013), in this sense highlights the existence of a community of Google Analytics users who actively contribute to the improvement of this tool, sharing knowledge and solving common issues. We can in that sense access Google's Analytics Solutions Gallery (google.com/analytics/gallery/), where we can find and share segments, reports and dashboards created by the Google team and worldwide contributors to help us improve our own analysis and solving common problems.

2.2 Defining indicators

The existence of large amounts of indicators paradoxically represents one of the most important challenges for managing information, due to the great variety of inputs and information. Because of this, Kaushik (2010) refers to the difference between reporting and analyzing and the maturation of the analytics process. Initial stages of analytics thus concentrate on **static reports**, drilling down and combining different basic metrics and dimensions, helping us identify the problem and providing an initial view on the situation. As our process becomes more complex, we then focus on identifying the **business main drivers**, as well as the impact of past and possible **future changes**, through the utilization of statistical methods, segmentation and sensitivity analysis. At this stage, more complex questions might be hypothesized in a set of what-if scenarios, looking to establish cause-effect relations. The last stages of analysis, focus on **optimizing** poor performing webpages and business areas, as well as predicting future evolution and impact of different indicators (Mohanty, Jagadeesh, & Srivatsa, 2013). This view is consistent with Delen & Demirkan's (2013) categories of analytics, which we explored in an earlier section, including the categories of descriptive, predictive and prescriptive analysis.

Moreover, beyond quantitative data we are interested in understanding the customer's experience. Kaushik (2010) in this sense refers to the importance of not only software, but especially investment in in-house intelligence. The author thus refers to the 10/90 rule, where only roughly 10% of the investment should be spent on

tools, while 90% on people, training and intelligence. These are the key determinants for success. The reason for that is the fact that every tool can provide us with a series of indicators and reports, which represent only the starting point. Understanding the impact of each variable and the evolution of business is however a much more complex issue, requiring proper strategy and the application of appropriate techniques for the transformation of data into knowledge.

Depending on our platform, different indicators might be provided, with some common characteristics between vendors. In this way, using the Google Analytics structure our data is processed according two different formats (Kutuçku, 2010): **metrics** and **dimensions**.

Metrics are here represented by a numeric value associated with a user behavior in our website, which can be calculated as an overall value or in segmented according to a dimension. Without segmentation, metrics provide aggregate and average values for the whole website, and are typically represented by columns of data. **Dimensions** on the other hand, correspond to the perspectives we want to adopt in order to explore the variations in metrics. These sets of criteria can not only correspond to our public, but also to some of the elements or sections on our website. These are typically represented as strings of data and tell us nothing without metrics. Besides default reporting, creating custom reports may also help us tailor the analysis, with GA allowing us to choose a combination of up to five dimensions and ten metrics per tab (and up to five tabs) for each custom report on its interface. Here we can also make a distinction between the user interface and the integration with other applications, which can be configured to run automatic analysis or for visualizing data, as we later discuss.

The definition of **Key Performance Indicators** (KPIs) is thus an essential starting point for the analysis, according to the online strategy. In this sense, KPIs are defined as the metrics which better help us understand the business evolution and the achievement of our goals. Kaushik (2010) refers to the critical few versus the insignificant many, as a common problem in the digital environment. Due to the great variety of indicators, the **critical few** are in this sense the metrics which have a direct impact on our business, with value variations tied to specific outcomes. KPIs thus

reflect important trends in decisive areas of our business, with significant oscillations motivating our immediate response (Waisberg & Kaushik, 2009).

In relation to this, Fagan (2013) citing Jansen (2009), points out the existence of different web categories, according to their business objectives. Because not all websites share the same business goals, different KPIs should be defined according to the business model for each website. In this context, some of the most common web categories include **Ecommerce**, **Content and media**, **Support and self-service**, and **Lead generation**. Many websites may also combine two or more categories, with different sections and purposes. An ecommerce website for example besides an online store, often includes a FAQ (frequently asked questions) support section.

According to this perspective, **ecommerce** webpages are mainly focused on the completion of transactions. This is one of the easiest categories to evaluate, since much of the website's success derives from revenue, a quantitative and straightforward approach to understand. Nevertheless, attention must be paid to different metrics, in order to extract insights about our visitors experience and the site's effectiveness. These indicators may thus include the average value of orders, the average value retained from each visit, bounce rate, conversion rate for different goals or customer loyalty (rate of returning versus new visits). However, no website is completely one-dimensional and different KPIs can help us put our efforts to perspective, identifying patterns through consistent methodologies and contextualizing results.

Kaushik (2010), also points out the importance of measuring the number of visits to purchase, which consists in the number of sessions it takes for a customer to place an order. This becomes more and more relevant as the items' price goes up, since customers tend to consider their alternatives with better care. Nonetheless, there are some strategies we can adopt to promote online sales, such as discounts exclusive to the online channel, which may help promoting this channel or induce a sense of urgency in buyers (Miletsky, 2010). Even so, there are many available variables for tracking the relevant phases of the transaction process, extending our knowledge about each of these stages. Some other more general indicators for ecommerce websites include the bouncing rates, associated with each page's lack of

relevancy, and conversion rates for different goals, associated to the completion of relevant stages in the conversion funnel. Apart from our web analytics platforms, we can also adopt active ways of gathering users' opinion, using tools such as surveys, in order to more deeply explore their perceptions and experiences.

Another web analytics category of analysis is the evaluation of **content**, which is many times reflected on the time visitors spend on our website and the number of interactions, as well as their proneness to return for another visit. Some of the indicators of interest thus include the evaluation of session depth and duration, each page's individual popularity, using metrics such as time on page, visitor loyalty (returning rate), recency or the acquisition of new visitors. A session long-lasting or with a higher page count is therefore connoted with higher engagement. This may sometimes also be reflected in the interaction with other elements, such as videos, comments or downloads. The completion of engagement goals is of course relevant, because it also contributes to other business objectives beyond sales. An example of that might be the acquisition of revenue through advertising, where having a bulky clickstream might ensure the sustainability of the model. While evaluating content, one of the most important factors is in this sense page popularity, given by the relation between a page's number of visualizations and the number of unique visitors (Fagan, 2013). In order to stimulate interest, besides relevant content, it is also necessary to be constantly updating and testing alternatives, keeping the website interesting in order to motivate returns (Burby & Atchison, 2007).

The existence of (self-)**support** content can also be analyzed using web metrics, with its effectiveness reflected by the satisfaction of our visitors with the information provided for solving their problems, as well as lower rates of direct contacts. This reduces the need for having direct support lines, with company representatives directly interacting with customers, with an effective web support section contributing for a lower structure of costs. In this sense, having low visit depth in sections associated with this, as well as low bouncing rates, is generally a positive sign of people finding meaningful content. Contrariwise, an intensive research process on these pages may reflect difficulty in finding information or a poor website architecture. The variation of average time on those pages may also be compared over time, as well as

the assessment of internal search terms and phrases. This helps us identify the pages' main problems, as well as the most common issues, contributing to the implementation of changes.

Lastly, online content might also aim to **generate leads** in order to collect information to develop advertising campaigns, create mailing lists or conducting market research. In this context, the conversion rates for specific goals, such as the number of newsletter subscriptions, might represent an accurate measure for determining the acquisition of new prospects and assess campaigns' effectiveness. As Fagan (2013) highlights, costs per lead are one of the main indicators for evaluating campaign relevance, as well as determining the most effective marketing channels in terms of ROI. In order to determine the better placement for an ad or a link, traffic concentration (visits to a page over the number of total visits) may also help identifying the pages with greater visibility on the website.

2.3 Meeting Objectives and Indicators

The definition of a digital strategy in a company may include many different materials and campaigns, which vary widely in terms of investment of time and money. Free solutions are nowadays increasingly common, illustrated nowadays by the importance of social networks. However, despite the free access to these tools, every action is a sign we emit about our activity as a company and must be duly justified with a well-defined strategy. Because of that, we need constant monitoring to evaluate the effectiveness of campaigns, in accordance to their stages of development, the business and available resources - monetary or human. Online environments are in this sense complex and diversified, requiring full time commitment, monitoring and consistent improvement. Not doing so, will result in the opposite effect, only hurting the organization's image.

A website must therefore be considered in the same way as any other extension of the organization, with great impact on its business. This requires the definition of goals, linking indicators to performance. Because of that, the strategy definition should therefore anticipate different scenarios, preventing the increase or

decrease on each indicator. Without this reference, the collection of data loses relevance, with no indication of business drivers or what we can do to improve. As we talked about (Mohanty et al., 2013), having a great number of variables can paradoxically become inoperable and therefore of no value. That is one of the main reasons we have the need for consistency, in order to obtain comparable data over time. That's why different researchers (Clifton, 2012a; Gupta et al., 2013; Kaushik, 2009) refer to analytics as a process, and not as an end in itself. Methodologies must thus be perceived as part of an enduring relationship, requiring commitment and constant monitoring and evaluation.

In this sense, different analytics platforms offer the option of customizing dashboards, reports and segments, in order to adapt the specificities of each organization and their need for analyzing different aspects of their online presence. The idea behind this (Kaushik, 2011) is that segmentation and customization of reports grants us a deeper comprehension about visitors' behavior, combining relevant metrics and dimensions. Only looking at aggregated data would on the contrary result in loss of information and under-specification of behavioral aspects. One of the main tasks of the analyst is therefore to comprehend the website's purpose and its main goals. For this case, Clifton (2012) defines the concept of **Objectives and Key Results** (OKRs), in strict connection to KPIs. According to this perspective, setting our OKRs is a four steps process:

First, we begin by mapping stakeholders, whether internal or external. In this sense, it may be relevant to talk to a representative of each department within the organization, trying to understand their strategic goals and online importance. Some of the most important stakeholders might include people with power to decide and make changes within the organization, with authority to allocate resources and decide on actions to prioritize. Conversely, external stakeholders should also be accounted for, such as consultancy agencies, which might need to have access to our analysis.

The second step is then to determine what the expectations of each stakeholder are. To do this, it might be necessary to arrange periodical meetings, hierarchizing priorities and evaluating the flow of operations. Different departments

have different goals, emphasizing the need for an agreement in favor of the business by cross-referencing data, contextualizing efforts and highlighting each contribution.

After obtaining consent from all teams, we thus reflect on the relation of each stakeholder's specific objectives with the project. In other words, we define success for each team, in relation to the online platforms. In this step we are therefore challenged to set measurable OKRs for each team, going beyond the macro picture. On the contrary, our goal here is to drill down and come up with relevant specific outcomes desirable for each stakeholder. The last step is therefore to evaluate the long list of objectives, distilling it down to an operable number of OKRs. In this sense, it is important to focus efforts in the most relevant objectives, each of them associated with a set of KPIs. In this phase is where it is important to be precise, in order to select the most crucial factors.

The definition of KPIs must therefore be an interpretation of these objectives, with the definition of targets for evaluating the business performance. This implies comprehensive contextualization of all the business variables, in an analysis of the environment conditionings. The goal here is to set challenging goals for each indicator, in order to improve performance, while still being able to realistically acknowledge strengths and weaknesses. Some analytics tools, such as GA, even provide us with the option of benchmarking our results against other companies within the same sector. Clifton (2012) however states that in spite of this is an interesting feature, it should not contribute decisively to our strategy. The fact is that each website has its own specificities and unique architecture. Besides, different companies have different ways to promote their businesses, targets and internal processes. Therefore, KPIs acquire meaning when interpreted at the light of their context. Comparing ourselves to other competitors may therefore be irrelevant or even misleading, since there is no context to these numbers. The author also points out that, even if we are talking direct competitors, it is almost impossible (and undesirable) for them to provide the same experience and share the same goals. Different websites will offer different experiences and benchmarking is much more important when done internally (over time), rather than in relation to our competitors.

In the same way, Kaushik (2011) also advises to drill down into our own data, resorting to statistical procedures and searching for variations in trends. The attribution of an economic value to conversions is also a feature worth exploring in web analytics, allowing us to integrate the online and offline marketing channels by monetizing experiences. The job of an analyst is therefore to anticipate the impact of visitors' actions in business performance, as well as the outcomes of our decisions. The main aspect is that the analysis should be capable of overcoming the visualization of data by studying scenarios from a holistic approach. Stakeholders thus expect an interpretation of the data, insights and solutions based on relevant indicators.

KPIs are in this way an effective measure of performance, allowing for the summarization and interpretation of data using other formats, such as spreadsheets or presentations. KPIs are however not a novelty in business, used with this or other designations to assess the organizations' performance. The necessity to monitor over time trends was always an underlying component of doing business, establishing watermarks and driving our actions (Clifton, 2012a).

2.4 Google Analytics reporting API

In computer sciences APIs are many times used by programmers in order to access information made available by a different platform. In this sense, APIs allow us to access the data contained in a database, according to a predetermined set of routines, classes and variables. These often come in the form of remotely accessed libraries from remote calls. It is thus necessary to know the specific rules and functions that accompany an API in order to specify the tasks to be run. While the API describes the expected behavior, a library is the implementation of these rules. In this sense, the utilization of APIs is usually useful for the automation of complex and time-consuming tasks, as well as the integration of information from multiple sources (Google, 2013). However, it is important to acknowledge the limitations of this tool, since data retrieved from APIs corresponds to a particular structure and only makes sense when integrated with data referring to the same unit of analysis.

With the proliferation of internet devices and the democratization of digital companies, the internet also became fragmented. Until only a few years ago, having a website with sporadically updated contents was sufficient for having a sufficient online presence. However, with the increasingly high complex networks of digital environments, website became insufficient and business solutions evolved with technology. An analogy can be made between the development of APIs and almost any other industry such as the automotive, in which evolution dictated the modulation and standardization of subsystems. The integration of optimized subsystems will in its part result in a great cost-to-performance relation (3ScaleNetworks, 2011), with major contributions deriving from the general public and the development of new features through the use of the API. In this sense, as long as the base structure of data is used, new functionalities can be added and updated according to the users' needs.

Many companies now take advantage of this by promoting the sense of community around their products through discussion, forums or galleries. This of course contributes for an improvement of product visibility through network effect, as well as increases the value of the platform itself. The major challenge with this is however being able to reach the critical mass to sustain a network of users and developers working in support of the platform. Once such is attained, it is thus much safer to guarantee the continuity of our user-base, since changes are often subject to losses and incompatibilities.

Google in this sense provides several APIs for Google Analytics, made accessible by programming languages such as Java, Python, JavaScript or PHP. In this work, we are going to be using the Core Reporting API, giving us access to most of the reports that can be consulted using the regular interface. While all of the data can be exported (up to 5000 rows of information at each time) using the interface, it has however to be done manually with only up to two dimensions. In this sense, if we are looking to explore the deep relations, this can be a time-consuming, not that agile methodology.

On the other hand, with access to the Core Reporting API this is a swifter process, with the automation of reporting tasks and integration of data with other applications. According to Google (2013) there are three fundamental concepts which underlie the utilization of the Core Reporting API: Firstly, we consider the request from

the application, specifying user credentials for the profile; Secondly, a query must indicate the values to be reported, including dimensions, metrics and the period relative to the analysis. Through this, we will segment our data according to the criteria defined in the dimensions; Lastly, the API returns a response in the form of a table, separated into rows (dimensions) and columns (metrics). Through this, we have access to the information directly on our statistical software.

The Google developers' web page in this context provides a wide array of tools to help users explore the potential of this API, including a reference guide for the query parameters, a dimension and metrics guide, a section for common queries and an automatic query explorer. Also, in the dimensions and metrics section, we can explore some of the valid combinations to the values that can be queried together, as not all the combinations are valid. GA has in this sense some limitations, with some combinations resulting in unintelligible data. This is because the data is mainly prepared to respond to the interface. However, it makes sense to drill into the opportunities presented by the API, to which we will use the software R. As we will see, this is a complete and agile statistical language, allowing us to integrate user-developed packages for innumerable functions and automatize the analysis process.

2.4.1 Integration of data with other applications

The main purpose for extracting data through the API is to better explore the relations between variables, create custom dashboards and to facilitate the visualization of behavior trends in our website. For this purpose, several solutions are available in the market by different companies, with tailored features for each company. Some examples of reporting tools include extensions such as Supermetrics, Next Analytics or Tatvic for MS Excel. These solutions generally enable their users to customize dashboards, integrating reports within the same sheet, a facilitator of the analysis process, since all the information is readily available, with the possibility of integrating multiple profiles. This is mainly a time-saving tool, which apart from GA can also integrate information from other Google sources, such as AdWords or Doubleclick.



FIGURE 6 - TATVIC EXCEL DASHBOARD

In this sense the utilization of the API allows for the utilization of data in other applications, which is a major asset for its community of users. The previous image depicts the utilization of the API for the development on a commercial application, reflecting the added value that can be drawn from this feature. In this case, several parties can contribute for value creation, taking benefit from new functionalities and building on the existing application.

In this work, we are going to use mainly R, an open-source statistical language for data analysis which is maintained and updated by developers worldwide, as part of the GNU – General Public License project. The main objective of this is to provide free software to anyone who wishes to use, share or modify it. Given its nature, it thus relies on a community developers for creating new packages in a dynamic and free environment (Coon, 1992). R is in this sense offers many options for exploring and transforming data, with features for modeling, analyzing, clustering, classifying, testing or graphically representing data, which can be extracted using the RGoogleAnalytics package.

In its most basic form, R has a very simple interface, with a console and a command line, which is our main tool for typing expressions. All commands are then interpreted by the software and result in a response (or error message). At first, it can

be a challenge to get acquainted to this interface, especially to users accustomed to friendly user interfaces, such as Excel or SPSS. Getting to know R's basic functions might initially be a time-consuming effort, but the existence of additional GUI's (Graphical User Interfaces) might help users of different levels of experience comprehend its resources. The wiki (rwiki.sciviews.org), is also a great starting point for new users, with first step indications and reference to introductory manuals. Such is the case of Coon (1992). The Comprehensive R Archive Network (CRAN), contains the main packages for users to download and is one of the most important resources for adding new features.

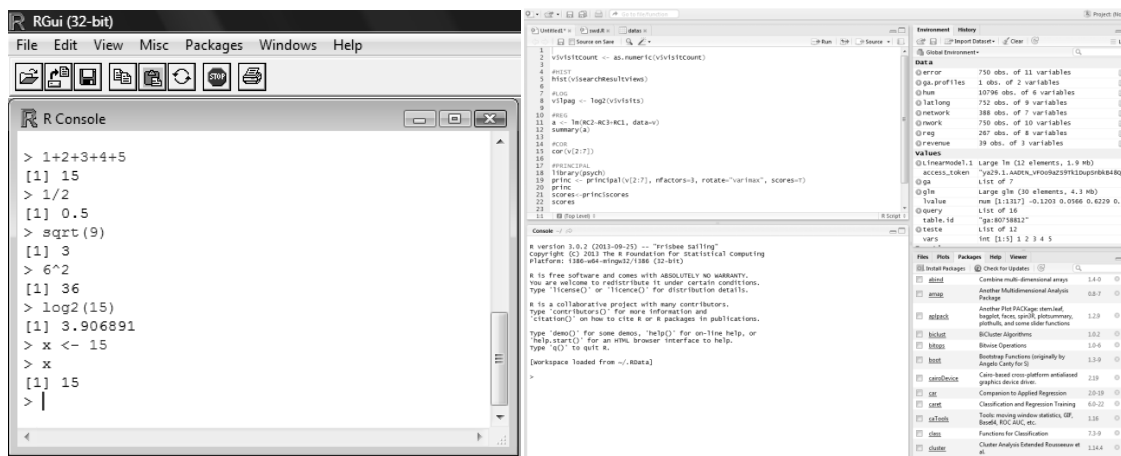


FIGURE 7 - BASIC R ENVIRONMENT AND R STUDIO

In this work, to initially explore the relations and tendencies in the data, apart from RStudio, in an initial phase we also used the Rcmdr and Rattle data mining GUIs, as proposed by Zhao (2013).

2.4.2 Statistical procedures in web analytics

Apart from the visualization and automation of reports, one of the major benefits from integrating R with GA is the fact that variables become easily accessible, allowing us to explore the existing relations of metrics and dimensions, using different procedures. There is in this sense an emerging discussion of whether the application of statistical techniques can be applied to web analytics data, such as predictive analytics.

In many areas, these are increasingly popular tasks, including marketing and CRM, since it allows for the anticipation of behaviors and events.

Kaushik (2007) however raises the discussion on whether web analytics data can provide appropriate information for proper predictive analytics, with actionable insights to the companies. In this way, this author rejects the idea of utilization of web analytics data in the ambit of predictive analytics, the main reason being the anonymous, incomplete and unstructured reality of web analytics data, which is subject to the sensibility of tagging methodologies and its imprecisions. This therefore hinders our chances of tying the behavior of people to expected outcomes. Furthermore, there are also a large number of variables which are not necessarily interconnected and many times have no relation to each other (or cannot be queried together).

Furthermore, users often exhibit a very heterogeneous behavior online, from where it is extremely difficult to deduce the primary purpose. The reason for that is because it takes very little effort for a user to click through multiple web pages. In that sense, determining the purpose of clickstream behavior for multiple people, because of aggregate values, can be highly inaccurate. Web analytics cannot also guarantee a holistic view of the customer across multiple touch points, platforms, devices or offline activity. Lastly, the pace of change on the often inhibits the credibility of results in predictive analytics, since predictions are to a great extent founded on the assumption of stability, contrary to the nature of the online ever-changing environment.

However, there are still a wide variety of metrics and dimensions which provide a window of opportunity for this work to explore the extent to which the utilization of GA variables can be employed. This is in fact demonstrated by (Araripe, Gondaliya, & Shah, 2013), where GA data is used to try to predict the probability of a customer returning a product. In this example however it is necessary to have in mind the need for attributing an ID to a particular customer and the supervised learning model followed. In this sense, existing data is used for estimating the weighing of variables and the development of a predictive model, using a machine learning algorithm for returning a probability value of return. This is called a **supervised method** because we use the training set (with a significant number of observations) to infer the function

used in new examples. From the train data, the algorithm is then used to predict the outcome of test observations. In this sense, an experiment might be conducted with our data, subdividing it into train and test sets (80%/20% of the data, respectively as suggested in the example). The second subset might then be compared to the real value, helping us evaluate the model performance and exploring the differences between actual and predicted values.

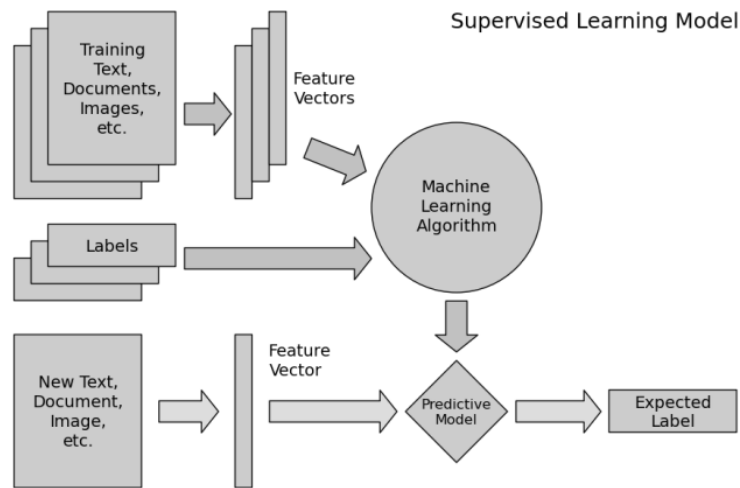


FIGURE 8 - SUPERVISED LEARNING FOR PREDICTIVE MODELS

Another example of the utilization of GA data to perform business forecasting is given by Wheble (2013) in which a regression analysis is used for estimating future traffic, based on the investment made on new campaigns. This is a very simple model, which however can give us the indication of the price we are willing to spend on advertising and online promotion. In this sense, by running a simple regression it is possible to try to predict how much traffic will be generated by these campaigns and relate that to the website’s goals. However, one of the major limitations of this example is the oversimplification of such a complex reality. Unless our website relies solely on the amount of visits for our business model (which is not a very good indicator of reliability), we have to take in account a much wider role of metrics such as session engagement, goal conversions, user precedence or transactions.

Furthermore, when taking in consideration this type of regressions it is important to maintain a critical stance since these explore solely linear relationships. James (2012) highlights several issues with linear regression models, with common

sense prevailing in the interpretation of outcomes. The increase in complexity of models often raises the rooted causation versus correlation issue, an especially relevant concern in the case of multiple regression models. In these cases, much information can be added without the consciousness however of where it is coming from. In the digital environment this is often the case, with huge amounts of information available, while much of it irrelevant for our purposes. Outliers are also a serious problem with internet data, both due to data collection methodologies and the users' erratic online behavior. The existence of software however makes it easier to run regression analysis, inspecting multiple different relationships.

Even so, James (2012) points out that the use of historical data might in some cases be inadequate, since business conditionings, especially in the case of newly formed or fast-moving areas of business, are continuously changing. In these contexts, comparing over time variations might be more adequate than forecasting. The reason for that being the inevitable fact that forecasting relies on past events in order to foresee future events. So, in areas where the business environment is still in development it is much more difficult to anticipate which will be the determinant variables for the organization's success. Furthermore, the more historical data we have, the better we know the variations of each variable and the better we can identify patterns, outliers and reduce error. This is a basic supposition in any application of statistics, where we assume that every measurement will be subject to some error. Theoretically, if we take enough observations of a given event, the random error (due to chance) will cancel itself out. We should thus attempt to collect a sufficient number of observations and use the most accurate forms of measurement available (Boslaugh & Watters, 2008).

3 Case study: Redcorp

For the realization of this work, we aimed to explore a real life situation in order for this analysis to go beyond the theoretical framework. Working with real data is in this sense extremely important, not only to better understand the practical applications of each tool, but because each case has its own particularities. Many factors can thus contribute to these differences, starting from the website's objectives and business model categories – which we explored earlier in the first section following Fagan's (2013) framework – but also the company's geographical location, its environment, the website's design, services provided and other factors. Because of each case's singularity, the definition of indicators for our digital marketing model is specific to each case, requiring a thorough evaluation of the current situation and the future strategy for the digital content of a company. Digital materials are nowadays an integral part of any company's image and the passive exhibitionism of websites no longer attracts users. On the contrary, they are now a synonym of fading brands.

Redcorp is in this sense one leading company of IT equipment and software in Europe, particularly in the Benelux region. This is a company dedicated exclusively to the B2B environment, selling and paying assistance to companies, professionals, public and private institutions through their website or via online and telephone conversations with sales and after-sales representatives, from their office based in Brussels. The company was created in 1989, looking to deliver an efficient and expeditious service to its worldwide customers, having a wide range of high tech products at competitive prices, but also providing a personalized assistance to its customers through direct contact to its representatives.

Because of that, the website is an essential aspect of the company's business and a fundamental contact point between the company, its customers and future prospects. Through the website users can not only consult and compare different products, their availability, prices and features, but also have direct contact with representatives of the company, place or track an order and manage their accounts and their history with the company. Furthermore, internal campaigns and promotions are available online, with this being a powerful relational tool with the customers. In terms of marketing and the management of customer life cycles and decision

processes the website is thus one of the most interesting areas to explore, due to the great number of tools available, the different ways customers reach it and what companies can do to maintain them. Customers nowadays assume a role in the conversation, with social media being a flagrant example of those dynamics. Multiple points of contact exist nowadays, with users opting-in and freely initiating the conversation.

In Belgium, similarly to the rest of Europe, the ICT industry and internet retailing are growing business opportunities, with new technological, distribution networks and a different cultural approach, more open and receptive to new technological-related services. Belgium is furthermore the center of the “Golden Banana” of Europe, which comprises the regions between the North of Italy to the UK, with access to a “one-day” market of 236 Million customers and a GDP of 1.5 trillion € (respectively 53% and 67% of the EU totals), in a 750 km radius. Furthermore, the country is also ranked number one in the more productive and globalized multi-lingual workforce (Deprest, 2012). Ecommerce in Belgium is also a consistently growing sector, where more than half of the population and 3/4 of internet users have made an online purchase, and not only youngsters. Trustworthiness is in this sense one of the main hurdles to avoid, with price, convenience and product range among the main reasons for consumers to opt for the online channel (Bloquiaux & Vuyst, 2013). The expectations are for sales to grow, with 2013 representing an increase of over 25%. Multimedia and hardware are among the most popular categories, next to clothing, home décor and appliances and toys (Hench, 2014).

Companies are because of that nowadays challenged to stay relevant and to attain the attention of the public eye, focusing on the interests of their target audience. Google Analytics is in this way one of the congregator tools for the monitoring and evaluation of the effectiveness of all channels, from external sources of traffic (advertisement, social media, referring websites...), to the internal behavior of users.

3.1 Methodology

In this work, we will focus on analytics indicators first exploring the application interface, all of its sections of reports and drilling down into the metrics. By combining different dimensions, we will conduct an exploratory and diagnostics study of the website's main trends. The objective of this is not only to evaluate the business in itself, but to also show the possibilities offered by GA and its interface. Each section will thus be divided into four subsections, with the first aiming at the **Definition** of each set of reports, and its primary aim; followed by an **Analysis** section in which we look into our case study, working the metrics and dimensions for each set of reports during the first period of time (13th of January to the 30th of March – 11 weeks). We then present a **Summary** of the main observations, followed by a **Period Comparison** section in a series of tests which aim to compare major changes in behavior for a second period (31st of March to the 29th of May – 13 weeks). Following that, we will also be using the API and R in order to run regression analysis on some of the most relevant metrics and dimensions to explain turnover, having in mind the structure and nature of the data. In these sections, we will use session dimensions, approaching the users perspective, as well as marketing channels for aggregate values.

3.2 Previous Research

Literature focusing on web analytics is not uncommon with many blogs, communities, tutorials, videos, books and content written on the subject. Google itself has a support section, online classes and solutions gallery for users to share their customizations and opinions. As we have seen, this is a very powerful tool, used by thousands of people worldwide, which additionally offers the opportunity to be adapted to the users' needs. Because of that, web analytics has also been theme to a number of papers and dissertations at an academic level, which are worth mentioning here, particularly in relation to case studies of organizations in other areas.

In this way, Fang (2007) for example used GA to track the users' behavior on the website of the Rutgers-Newark Law Library, aiming to understand the motivations behind searches and to evaluate the design and content of the site's pages. In this

work, site overlay, content by title, funnel navigation, visitor segments and summaries constituted the main information that was monitored, which resulted on design suggestions of improvement. Likewise, Lee (2011) also studies the behavior of users in a digital library environment, with the objective of inferring user satisfaction (tracking actions, user retention and triangulation of data with other sources – the 4Q online questionnaire), the impact and performance of the website (usage behaviors, user group, brand awareness and channel performance) and assisting decision making (on a User Interface level, content and levels of reporting). One of the aspects that is emphasized is the difference between the library environment and ecommerce websites and the issue of the definition of “success” for each of objective.

Still in the scope of the academic library environment, Fagan (2013) uses web analytics KPIs in order to assess the navigation of users and if they can find the appropriate databases for what they are looking for, as well as the returning rate for the website (loyalty). For this, metrics such as the number of page views, session time, depth, customer loyalty (unique visitors and return rate), and page popularity were considered for the research. We therefore see that this is an agile tool, which can be tailored to the necessities of any kind of company. Kent et al. (2011) on the other hand approach a broader application of web analytics, discussing its usefulness for communications, PR and information professionals, giving the example of four web sites. Among these we find an academy professor’s website, the site of the independent Institute for Policy Studies, the governmental City of Prague portal, and the professional information site PR Romania. In this work, the four case studies are briefly discussed, with highlight to the ease of comprehension of this tool, but also to the necessity of a conceptual framework which contextualizes our approach to data.

Through monitoring the ROI of marketing channels, the effectiveness of online campaigns and the improvement of the organization’s online presence, the consultancy agency Elisa DBI (2013) also helped one of the leading international health charities in the UK – Merlin – to increase its email registrations by 141%, reducing acquisition costs of subscribers and donations by 25%.

Another example is the research conducted by Pakkala et al. (2012) defining metrics for Food Composition websites from Denmark, Finland and Switzerland, and

the interaction of users with the content. These are websites containing information about different nutritional facts of food, for professionals or people interested in health and nutrition. This is a comparative study between the sites in which a framework containing common KPIs is defined for the three websites, and then compared in terms of user interaction and engagement. This research also aims to explain how the websites are found by users, the main drivers of traffic, user loyalty, content and main keywords, reflecting the main categories of interest for people who visit the website.

Kutuçku (2010) on the other hand studied the communication potential of the Middle East Technical University Institute website, which provides information to its current and potential students on courses, procedures and a point of contact with the academic environment of the institute. In this work, besides the data from six months of observations using web analytics, usability studies were also conducted with resort to the think-aloud methodology, where users were asked to perform tasks and assess the design, content, ease of use, brand recognition, self-efficacy and overall evaluation. GA also helped identify problematic pages, such as the landing pages, keyword clusters and the most relevant dimensions and metrics for the site, resulting in suggestions of improvement for content and interface design. The resort to usability tests in a broader sense, in which we can include web analytics or think-aloud studies, is for us one of the main interesting features in this work, bringing to attention that different levels of testing might be conducted. From controlled, laboratory situations where it is possible to obtain in-depth analysis and opinion from a limited number of subjects, to real-life situation from virtually all users who enter the site (web analytics). Of course there are some trade-offs between these approaches, with different levels of understanding about the “what”, “why” and “when”, confronting real life behaviors based on simple metrics with in-lab highly monitored experiments.

The collection of data for public institutions is also illustrated by Plaza (2011), having collected data for the Guggenheim Museum in Bilbao during 1092 days and 7561 visits. In this work, the author used GA’s export feature to read the data in a MS Excel format, having performed an analysis on the effect of the typology of visitors (new vs returning) on the number of pages seen per session, the precedence of users

(traffic channels) and their likelihood of returning to the site. Moreover, the effect of one variable over the other is also explored here by plotting the data, which enables us to observe the levels of correlation and the distribution of values. This is a descriptive work, being one of the few we could find which illustrates the use of other statistical packages to explore tendencies in the data using GA data export feature.

Paradoxically, in spite of the heavy use of GA in ecommerce websites, it is sometimes difficult to find academic research specifically in relation to this theme, as pointed out by Hasan, Morris, & Probets (2009). These authors investigated the use of web analytics in relation to three ecommerce websites and the extent this tool can be used to identify usability problems in specific sections of the website, making use of 13 indicators (Average page views per session, Session time and depth, Bouncing rate, Order conversion rate, Average search per visit, percent of visits with search, Search to exit ratio, Cart start rate, Cart completion rate, Checkout start rate, Checkout completion rate, Information find conversion rate). The authors thus argue that while web metrics are useful to identify general tendencies and specify problematic sections of the website, in a quick, easy and cheap way, in-depth knowledge about user navigation is often left unanswered. Heuristic evaluation was in that sense used to confirm the conclusions deriving from web indicators, specifying usability issues in each page. Still, web indicators can represent an advantage from a business perspective, providing information on financial data and rates of goal conversion. In this study, six characteristics were evaluated: navigation, internal search, site architecture, content and design, customer service, and the purchasing process. Heuristic methodologies work in this sense as a complementary procedure for obtaining specific information on the usability issues identified by web analytics.

4 Google Analytics interface

The Google Analytics interface is the first environment to explore after setting up the tracking code (GATC) in our web pages. It provides the user with hundreds of default reports and segments, which can be used to explore the main trends on our website, as well as share information with stakeholders in the company. Access to information is subdivided into multiple levels, including the administrator and users' profiles. Using the same login, we can create unlimited accounts as well as properties, each corresponding to a domain or subdomain. Each property can also be associated with up to 50 profiles, with different levels of authority and access to the information. In this way, to different stakeholders in the company we can provide access to the data and the opportunity of visualizing, adding filters, editing or creating new conversion goals, segments, alerts, schedule e-mails, create shortcuts or annotations. The administrator also manages users' access, account settings and the integration of information from other sources, such as AdSense or AdWords. This is particularly important for the configuration of the GA interface and the realization of tests such as A/B tests, which depends on the integration with Google Webmaster Tools.

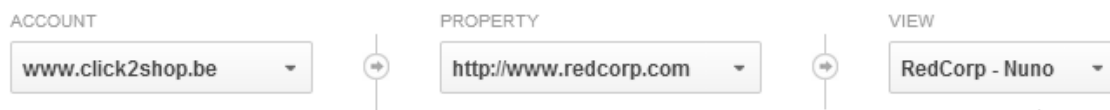


FIGURE 9 - LEVELS OF ACCESS IN GA

In this case, we are using a profile with access customization options, allowing us to create reports, segments, dashboards, goals, filters, annotations or alerts. For this particular website we had already defined a set of goals which aim to reflect the engagement of users with the website, as well as the acquisition of new prospects:

Goal 8 – Engaged users – per visit, for sessions with more than 10 page views;

Goal 7 – Engaged users – duration, for sessions lasting longer than 5 minutes;

Goal 6 – Newsletter subscriptions;

Goal 5 – Order process flow, consisting of the five steps taken to place an order, including the cart, login, shipping, payment and summary pages;

In order to have a consistent analysis we also need to have a significant amount of data. However, when granted access to a view, data is only available starting from that same day. Because of that, a period of time is needed in order for us to have enough observations that allow us to start exploring some relations. For this section, we are going to be using data collected from Monday, the 13th of January of 2014 to Sunday, the 30th of March of 2014, an eleven week period in which we can clearly begin to identify patterns in the behavior of visitors, the importance of different channels and the interaction of visitors with some of the main pages and site features. Some of the aggregate values for that time period show us that we had almost 100.000 visits from about 62.000 unique visitors, with clearly higher traffic during weekdays (expected in a B2B environment).

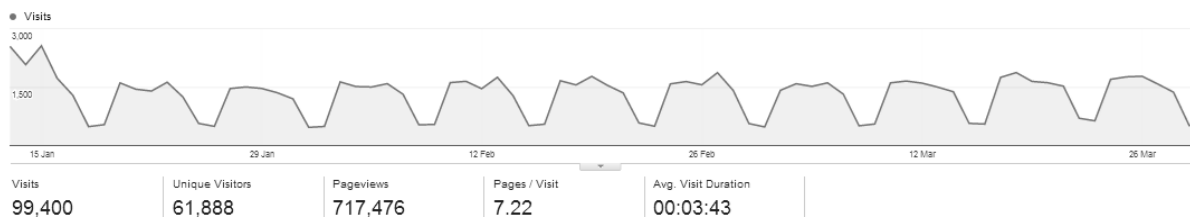


FIGURE 10 - SESSIONS PER DAY AND BASIC INDICATORS

4.1 Intelligence Events

4.1.1 Definition

The intelligence section is intended to help users identify significant variations in the metrics and can be subdivided into automatic and custom alerts. Automatic alerts are in this sense calculated by GA, regardless of the indicator, automatically capturing any significant variations. This is done based on each metric's past performance for comparable periods of time, calculating its average values and standard deviation according to the principles of normal distribution. The sensitivity of an alert can thus be triggered to oscillate between 1 to 7 times the standard deviation, from the highest to the lowest level of sensitivity. In this sense, at highest sensitivity, if a metric suffers a deviation of only 1 standard deviation, an automatic alert will warn us about a possible behavioral change. As we know from the three sigma rule in

statistics, nearly all values are contained within three standard deviations from the mean, with a sequence that goes from 68.27%, 95.45% to 99.73% as we get further away from the mean value. This may be a very useful feature, since GA does this for all data in the profile. Even if it is not a relevant metric for our digital strategy, significant oscillations are still going to be communicated (Google Inc., 2014). Due to the high number of metrics, we are often unable to individually monitor each one. However, through this we have the possibility to passively monitor major changes, so we can then decide whether or not those are relevant oscillations.

Metric	Segment	Period	Date	Change	Importance ↓
Bounce Rate	Medium: email	Weekly	9 Feb 2014 - 15 Feb 2014	>500%	
Revenue	Country/Territory: France	Weekly	9 Feb 2014 - 15 Feb 2014	>500%	
Bounce Rate	Exit Page: /Default.aspx	Daily	23 Feb 2014	>500%	
Revenue	Country/Territory: Belgium, City: Liege	Weekly	2 Mar 2014 - 8 Mar 2014	>500%	
Visits	Source: newsletter	Daily	27 Feb 2014	>500%	

TABLE 3 – AUTOMATIC INTELLIGENCE ALERTS

Furthermore we can also choose to customize our own alerts, for receiving information about variations regarding specific metrics of particular relevance. These might apply to different periods in time, for all traffic or only certain segments defined by dimensions. As an example, we can segment our alerts according to the type of visitors, traffic channels, behaviors, users’ devices or ecommerce objectives. On the other hand, metrics can also be related to site usage, goal completion, ecommerce, specific content or clicks on campaigns. The personalization of intelligence alerts in this context focus either on the percentage variation for each metric or on the definition of absolute threshold values.

Alert Conditions

This applies to

Alert me when

Condition

Value

Compared to

FIGURE 11 – CUSTOMIZED ALERTS

4.1.2 Analysis

For the month of March, the intelligence alert tells us there has been an increase in the number of total visits, with this month registering 10% more visits than the last comparable period. For this, was especially important the growth in the number of New Visitors, particularly in the case of visits generated by Google searches. This is in this sense positive information for our site, which can reflect the result of digital marketing strategies. If for example we were investing in SEO this could be an indication of success, especially in the case of the acquisition of new visitors. However this issue must be interpreted in terms of relative evolution and not the absolute values. The identification visitors using *page tagging* combined with the impossibility of consulting the queries which generated this increase in traffic (because of (not set) keywords constituting about 94% of total organic traffic) makes it impossible to identify the exact number of returning visits. It is thus more relevant to comprehend periodic variations, consulting multiple reports and indicators.



FIGURE 12 – ALERT FOR AN INCREASE IN TRAFFIC WITH VISITOR TYPE AND SOURCE

4.2 Audience

4.2.1 Definition

According to Google Inc. (2014) the intent of audience reports is to provide insights into the demographic variables which compose our audience, technologies used to reach our site and assess some aspects on loyalty and engagement of our public. One of the most useful sections here is the geographical reports, which comprehend the language and location reports, also graphically representing the origin of visits from around the world. This is made by an approximation of the area for our visitors by using the IP address to estimate their location. Because of that, it is not a 100% accurate tool, demarcating users by region and service provider (ISP), rather

than exact locations. One of the best uses for this feature is to count the visits originated from a certain region, also retrieving the approximated latitude and longitude using the API and integrating multiple layers of information.

The processing of the demographics and interest reports is on the other hand made by calculation of user categories and per website affinity based on the users' searches and website visualizations. The way this segmentation is done is by monitoring the data from previous visited sites by each user, for determining their interest and age groups. However, this can only be done by approximation, with each unique visitor associated with one or more devices (Google Login) and vice-versa. In this way one device might be used by multiple users, while the reverse is also true. Since it demands for modifications to the GATC as well as contact with Google administrators, this feature goes beyond the scope of this work. Still, it is here worth mentioning as an additional feature. In this way, we will mainly be considering geographical, technological and (in-page) behavioral aspects for our audience.

4.2.2 Analysis

4.2.2.1 Location

For our case, 54.4% percent of all visits come from Belgium, while in the second position we have the USA with only 3.4%, followed by the UK with 3.2%. However, 90.8% of the revenue provides from Belgium, with Germany and France accounting for just about 2.6% and 2.4% respectively. In spite of the same relative weight in terms of revenue, Germany had almost double the number of transactions.




	Sessions ?	% New Sessions ?	Pages / Session ?	Avg. Session Duration ?	Transactions ?	Revenue ? ↓	Ecommerce Conversion Rate ?
 Belgium	54,056 (54.38%)	41.73%	10.02	00:05:24	2,540 (87.74%)	€1,208,186.32 (90.79%)	4.70%
 Germany	3,334 (3.35%)	62.69%	6.58	00:03:45	110 (3.80%)	€35,064.89 (2.63%)	3.30%
 France	3,148 (3.17%)	60.20%	5.70	00:03:13	60 (2.07%)	€31,986.72 (2.40%)	1.91%

TABLE 4 - INDICATORS FOR THE 3 MAIN REVENUE-GENERATING COUNTRIES

This might lead us at first to think that average orders from France were more valuable, which could be an inaccurate statement. In order to investigate the

distribution of order value, we use the API to extract more insight on the characteristics of the orders made in each of these three countries. In this way we can observe not only the mean values for the orders, but also the interquartile differences and the type of distribution followed in each case.

As we can see below, these are all highly skewed distributions, with strong influence of extreme values in the average values of aggregate data. Because of that, average order value tells us also about the nature of our business, since much of the revenue is provided by few transactions. As stated by Provost & Fawcett (2013) this type of distribution is a very common characteristic in web data, with the behavior of users fluctuating widely according to different metrics. This is due to the ease of access and lack of costs for additional visits, as well as the absence of the conditionings presented by traditional physical environments. Clicks take very little effort and users also feel protected by the anonymity of the web. Also, as we have stated, B2B environments involve much longer decision processes, with multiple levels of hierarchy, influencers and decision makers. In this sense, all investment need to be duly justified for their functionality, with the disregarding of emotional factors (Leek & Christodoulides, 2012). In this sense, brands can be important mostly from a relational perspective, inducing a sense of trust and reducing the perceptual risk. Interpersonal relationships are often very important, with sales representatives being the face and synonym of trust in the brand.

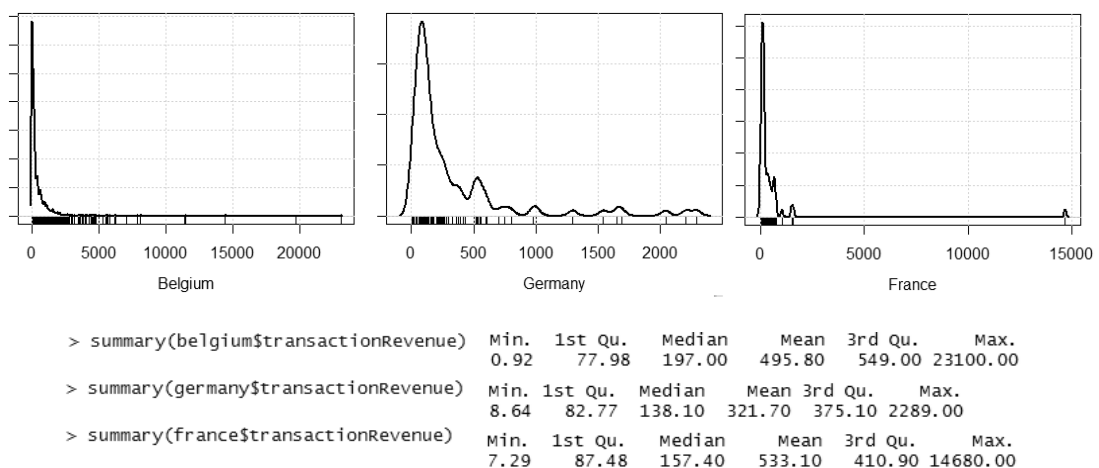


FIGURE 13 – DISTRIBUTION AND INTERQUARTILE RANGE OF TRANSACTIONS BY COUNTRY (USING R AND THE API)

As we can see, the distribution of order value highly skewed in every case, with the distribution of visits and revenue in Belgium concentrated in Brussels, with almost 24% of the visits and 29% of the revenue. Antwerp follows in terms of relative visits with 6.5% and 5.5% of revenue, and Louvain-la-Neuve in spite of only contributing with 2% of total visits accounting for 5.55% of revenue, followed closely by Liege. This confirms the importance of Belgium and some of its major cities, particularly Brussels, for the business volume of this company, with the majority of traffic and revenue concentrated in a limited geographic region. It seems therefore there is an unexplored window of opportunity for other countries, especially inside the European area, where transactional costs and cultural barriers are reduced.

We have already mentioned France and Germany, but ecommerce conversion rate is also high for Switzerland (11.3%), Denmark, (3.3%), Sweden (2%), Luxembourg (1.7%) or Norway (1.4%), where few visits convert more rapidly. Still, the contribution in percentage for total revenue is very limited and most of these transactions are originated from a limited number of territories, which may translate into just a few returning customers. These tendencies may be observed in the following table, where we explore the behavior of the 10 most revenue generating cities outside of Belgium. As we can see, engagement is also high in most cases when compared to the site average (7.2 pages and 3:43 length per session), with a general lower rate of new sessions (57.8% site average):

		Sessions ?	Pages / Session ?	Avg. Session Duration ?	Revenue ? ↓	Ecommerce Conversion Rate ?	Transactions	% New Sessions
Le Cres	▣▣	17 (0.04%)	10.82	00:08:29	€18,734.61 (15.02%)	5.88%	1 (0.27%)	39.34%
Kastrup	▣▣▣	60 (0.14%)	15.78	00:12:31	€7,992.24 (6.41%)	23.33%	14 (3.78%)	32.34%
Pulheim	▣▣▣	83 (0.19%)	13.75	00:07:45	€7,721.01 (6.19%)	20.48%	17 (4.59%)	43.31%
Idstein	▣▣▣	80 (0.18%)	18.52	00:14:54	€7,351.92 (5.90%)	18.75%	15 (4.05%)	40.41%
Luxemburg City	▣▣▣	578 (1.31%)	9.80	00:05:20	€6,660.60 (5.34%)	2.42%	14 (3.78%)	41.11%
Stockholm	▣▣▣	211 (0.48%)	4.40	00:01:10	€5,628.92 (4.51%)	4.74%	10 (2.70%)	30.81%
(not set)	▣▣▣	447 (1.01%)	9.15	00:05:08	€5,453.58 (4.37%)	5.37%	24 (6.49%)	51.64%
Paris	▣▣▣	507 (1.15%)	6.60	00:03:55	€4,521.42 (3.63%)	3.75%	19 (5.14%)	35.06%
Milan	▣▣▣	241 (0.55%)	4.41	00:01:01	€4,156.59 (3.33%)	5.81%	14 (3.78%)	38.12%
Grenoble	▣▣▣	86 (0.19%)	10.78	00:05:02	€3,238.30 (2.60%)	5.81%	5 (1.35%)	52.77%

TABLE 5 – TOP TEN CITIES OUTSIDE BELGIUM

4.2.2.2 User type

Besides geographical dimensions, the Audience report also allows for an analysis of the public according to their behavior, pages' stickiness and the motivation generated on users to return. It therefore provides information on the type of visitors (new vs. returning), recency and frequency of visits. In this case, we can see that new versus returning traffic is for the most part hand in hand, with a rate of new visits of 57% and 43% for returning visits. It is important however not stick only to these numbers, looking at other metrics in order to better comprehend the impact of each type of visitors for our business. In this sense, we can see that in spite of in average over half the sessions are coming from new arrivals to the site, returning customers engage much more with the content of the pages and have greater visit duration, when compared to new visits.

According to our conversion goals, about 27% returning visits last longer than 5 minutes and look into more than 10 pages, while only about 9% of sessions from new visitors do so. Additionally, over 86% of the revenue was unquestionably generated by returning customers. Again, this may in reality be even a greater number, due once again to the inaccuracies of page tagging methodologies.

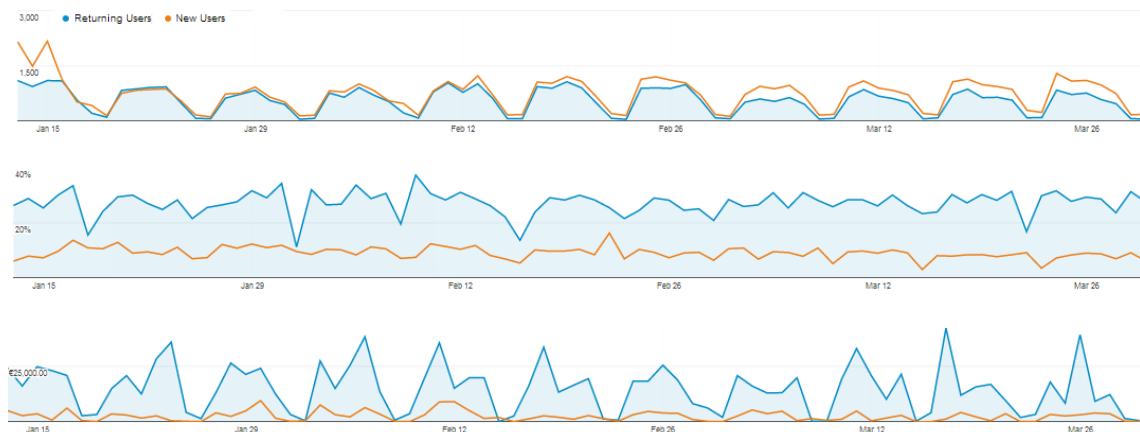


FIGURE 14 – RETURNING (BLUE) AND NEW (ORANGE) USERS PER NUMBER OF SESSIONS; % OF ENGAGED VISITORS (PAGE VIEWS >10); AND DAILY REVENUE

Another unmistakable fact is the relative greater importance of returning visits in terms of both value and engagement when compared to new visits. In this sense, running a correlation matrix in R between these variables and using a binary for identifying returning and new visitors, we can further explore the relation between

visitor type and its relation to conversions and value. In spite of a negative correlation between returning visits and total traffic (there are generally more new than returning visits), there is a conversely clear positive effect of returning visits on both revenue and goal 8 conversion (session page views > 10).

	isreturning
visits	-0.3805966
goal8ConversionRate	0.9391798
transactionRevenue	0.6325175

TABLE 6 – CORRELATION MATRIX FOR THE EFFECT OF RETURNING VISITS ON THE NR OF VISITS, GOAL 8 (PAGE VIEWS >10 PER SESSION) CONVERSION AND REVENUE

In the case of the frequency and recency reports these also correspond to highly skewed distributions, in relation to both visit count and the number of days since the users' last visit. In this case, only a few devices will truly preserve this information, which will lead to the reporting of progressively few extreme values. The solution provided by GA is to bin the distributions, increasingly widening the limits of each group, according to the number of occurrences. This might however induce error in the treatment of data, since we will be considering different intervals for each bin, while treating them as equal. In other words, we would be taking users (devices) which had a visit count of for example 15-25 and compare them to the ones which registered 101-200 sessions. The same happens with the Engagement report where GA uses bins according to the page depth (pages per session) and duration of sessions. This may in this sense be used as a mere indicator, however not consistent from our point of view. Still, we can retrieve the exact count for each value by using the Reporting API, avoiding GA's default values and further exploring the relations established recurring to the use of our statistical software.

4.2.2.3 Technology

Still considered within the audience tab, we can also explore the technologies used by visitors to access the website. In this view, the mobile report provides information about the devices used to consult the site, which in this case corresponded to about 9.6% of total visits for the given time period. However, the generated revenue was only at about 0.16%, with also seemingly lower engagement

than desktop devices. Page views per visit and average visit duration for desktop devices averaged in this sense at about 7.6 pages and almost 4 minutes, while mobile and tablet averaged at respectively 3.27 pages and 1 minute and 4.27 and just under 2 minutes session duration. This thus seems to be a relatively inexpressive technology, which can still be used for occasional visits or to view specific products. As we can see in the Visitors Flow report, from all mobile visitors who do not drop off in the starting pages (only about 15%), at least 31.6% uses the internal search feature on a first interaction. This indicates these are users looking for specific categories of products.

4.2.2.4 Visitors flow

The Users Flow report in this sense provides an interactive visualization of the main pages consulted by users according to the funnels of traffic on different pages. Furthermore, we can also use segmentation features in order to explore either the main landing pages of the public, problematic drop-off zones in the conversion funnel or the effectiveness of campaigns. One of the campaigns identified with an abnormally high number of new visits during the first week of data (figure 14), which can be traced back to the Google/CPC medium. In this sense, if we isolate this time period and the medium in order to explore user interactions at each level, we can see that this was a campaign generating mostly unqualified traffic, with 98.3% drop-offs in the landing page. This is thus an ineffective campaign, with users quickly leaving.

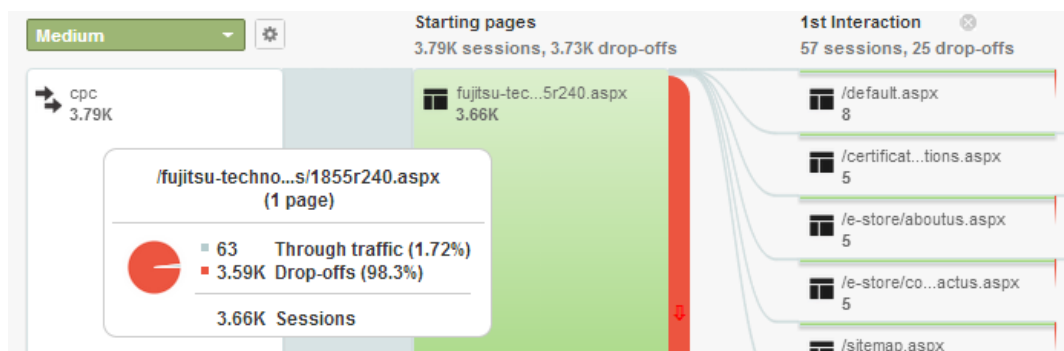


FIGURE 15 –POOR PERFORMING CPC CAMPAIGN: 98.5% DROP OFFS BEFORE THE FIRST INTERACTION

On the other hand, we can also see that most users are dropping off at the very first page they visit, with about 39.1 thousand visits for the month of March and 24.3

thousand drop offs in the landing page, which represents 62% of total traffic. This difference is even greater when we consider new visitors, with a drop off rate of 77% in the starting page, compared to 39% from returning users. In the Flow Visualization report, we can further explore these relations with organic traffic corresponding to 58% of total traffic, facing only 22.5% of direct visits. However just under 25% of these visits accounted for through traffic, in which most were landing on the home page (default.aspx) and almost half (about 10% of total) continue searching the site using the internal search.



FIGURE 16 – PATH FROM ORGANIC TO INTERNAL SEARCH ON THE 1ST INTERACTION

One relevant factor to take in consideration is that for every landing page seen by new visitors, about two-thirds drop off before having any interaction. On the other hand, returning visitors reflect a much higher percentage of through traffic from the landing page, with 39% drop-offs, with a home page at 90% through traffic, search page at 46% and product page at 25%. Hewlett-Packard pages are also mentioned among the main landing pages for both returning and new users, with respectively 20% and 3% through traffic, as well as Samsung in the case of new traffic. Most of these are generated from organic visits (80%).

4.2.2.4 User Networks

Lastly in this section, it is also worth mentioning the network dimension, which indicates the users' service provider. This is a geographical attribute which can be combined with other dimensions, including spatial data. This is mostly useful for understanding the distribution of users as well as the diversity of connections used (Google Inc., 2014). One of the main features of this dimension is that we can sometimes approach the user at a device level, singling out in some cases specific

organizations. In this case, we can identify a few specific networks, such as those used in universities. The latitude and longitude dimensions, exported to R using the API can further provide better comprehension on the location of visitors. To explore the interactions between dimensions, we thus combine multiple criteria, in order to isolate a specific organization’s network. In the case of “Université Catholique de Louvain” network for example, we can see that in spite of 13 different Operating Systems having had consulted the website, windows 7 and XP devices accounted with over 91% of value from all the sales. Furthermore, we also see that the ecommerce conversion rate for XP for this network is over 91%, having registered high levels of engagement per session. Almost every session from XP in this network thus ends up in a sale. This might be truly important information, since this is the number 3 network in terms of overall revenue, allowing us to adapt strategies and target customers. Due to the great number of networks, we are only going to be looking into the most important identifiable source of revenue as an example of what can be done in this sense:

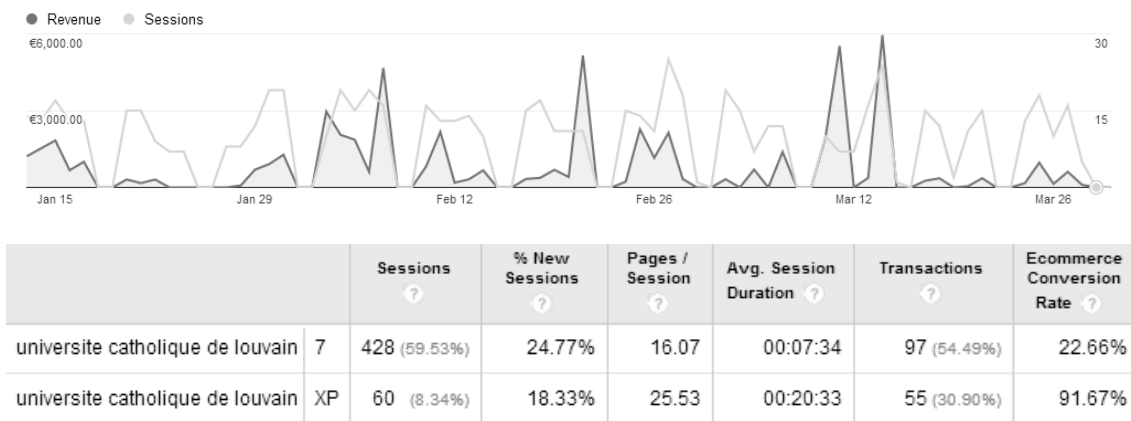


TABLE 7 – REVENUE AND SESSIONS FOR THE “UNIVERSITE CATHOLIQUE DE LOUVAIN” FOR THE TWO MAIN OPERATING SYSTEMS

4.2.3 Summary

To wrap up this section, some of the main insights here were:

Geographically, 90.9% of the revenue comes from Belgium, while only 56% of the visits also do so. Germany and France follow as the most representative countries, with 2.6% and 2.4% of revenue and 3.4%, and 3.2% of visits respectively.

The distribution of order values is strongly skewed with great contribution, in most cases, of extreme values for the overall performance of each territory. This is particularly well illustrated in the case of Germany, which in spite of having almost twice the transactions of France with an ecommerce conversion rate of 3.3% over 1.9%, both countries account for almost the same revenue, due to extreme high values on the side of France. The maximum value in Germany was in this case 6.4 times lower than France's maximum.

Some cities outside Belgium might indicate important clients, exhibiting lower than average percentage of new visits, little absolute number of sessions but high engagement and conversion rates. We in this case highlight for example the cities of Kastrup, Denmark (23.3% conversion rate), Pulheim (20.5%) and Idstein (18.8%), both in Germany, besides the number one region outside Belgium, Le Cres (with only one transaction, raising the question of a fortuitous event).

The importance of returning versus new visitors is manifest both in terms of sales (86% of revenue – which is obvious since users have to login to place an order), but more importantly in terms of engagement with 27% over 9% from new visits., with differences varying with the provenience of users.

Similarly, 90.4% of visits come from desktop users, corresponding to 99.8% of revenue.

For both returning and new visits, the most used navigation feature is the web shop internal search for a first interaction, with differences in the drop-off rate of users. While 76.6% of new visitors drop-off on the landing page, only 38.7% returning do so. In the first and second interactions 84.2% and 89.9% of the original traffic volume drops off for new traffic, while only 56.1% and 67.6% do so for returning visits.

4.2.4 Period Comparison

- Belgium cities contribute with great majority of revenue and visits (90.9% and 56%), while Germany and France follow;

	Sessions ?	% New Sessions ?	Transactions ?	Revenue ?	↓	Ecommerce Conversion Rate ?
Belgium						
13, 2014 - Mar 30, 2014	54,056 (54.38%)	41.73%	2,540 (87.74%)	€1,208,186.32 (90.79%)		4.70%
31, 2014 - Jun 28, 2014	55,607 (47.52%)	42.65%	2,577 (86.27%)	€1,208,748.81 (88.25%)		4.63%
Germany						
13, 2014 - Mar 30, 2014	3,334 (3.35%)	62.69%	110 (3.80%)	€35,064.89 (2.63%)		3.30%
31, 2014 - Jun 28, 2014	3,909 (3.34%)	65.64%	130 (4.35%)	€38,516.57 (2.81%)		3.33%
France						
13, 2014 - Mar 30, 2014	3,148 (3.17%)	60.20%	60 (2.07%)	€31,986.72 (2.40%)		1.91%
31, 2014 - Jun 28, 2014	4,012 (3.43%)	67.92%	77 (2.58%)	€37,495.76 (2.74%)		1.92%

TABLE 8 - TOP 3 COUNTRIES BETWEEN THE TWO PERIODS

As we can see from the previous table, there does not seem to be a significant difference between the two periods we have been collecting data, with only small changes to each indicator. The more significant change in percent terms is an increase on the percentage of new sessions for France, which went up 7.72%. In order to test the statistical significance of these differences we can in this way recur to our software, running a t-test for comparing the proportion for each period. This will help us compare the two periods, as we would for control and treatment groups, or the evaluation of periods pre and post promotional campaigns (Kaushik, 2006). Due to ease of translation, we are going to use Excel for running automated proportions test at **95% confidence** (With t-statistic of -1.96 and 1.96 for 2-tailed tests):

$$Z \cong \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

FORMULA 1 – STATISTICAL TEST FOR COMPARING PROPORTIONS

In this way, we can see that the difference in the proportion of transaction conversions for Belgium was not statistically significant, with a test value of 0.506. The

same happened in the case of Germany ($Z= -0.062$), and France ($Z= -0.04$) for the number of transactions. On the other hand, in spite of the increase in the absolute number of visits for the second period, the proportion of visits from Belgium was higher during the first period ($z= 31.81$), which means that other countries gained relative importance during this period. According to this, while there was no statistically significant change for Germany, the proportion for France was lower during the first period at 95% confidence ($z= -3.39$). Likewise, the proportion of sessions for “Other” countries also increased for the second period ($z= 30.99$). The best example of this was the US, which went from 4.91% of sessions to 6.52% (4877 to 7631).

Generally speaking, all top 20 countries increased their absolute number of sessions, with a statistically significant increase. However, that tendency was not accompanied by the number of transactions, which showed no significant increase at 95% ($Z= 1.668$). This is because most of the increase on the number of visits came mostly from new users, reducing the overall ecommerce conversion rate. The percentage of new sessions, in this way went up from 57.8% to 61.7%, a statistically significant change at 95% confidence ($Z=18.64$). While there was also a statistically significant change for Belgium in the percentage of new sessions ($Z=3.08$), it was only a difference under 1%, while for Germany and France these were differences of almost 3% ($Z=2.61$) and 7.7% ($Z=6.78$). In this way, in absolute terms Belgium had about 1159 more new sessions in relation to the first period, while Germany and France had 476 and 830 more new visits.

	Sessions	% New	New Sessions	# Diff.
Belgium	54056	41.73%	22558	
	55607	42.65%	23716	1159
Germany	3334	62.69%	2090	
	3909	65.64%	2566	476
France	3148	60.20%	1895	
	4012	67.92%	2725	830

TABLE 9 – SESSIONS FOR BOTH PERIODS IN THE TOP 3 COUNTRIES

- Returning customers generate more revenue (86%) and engagement (27% vs 9%), while the traffic volume is not significantly different;

For the second period there was a higher percentage of new visitors (61.77% over 57.78%), which is a statistically significant difference at 95% confidence ($Z=18.86$). The ecommerce conversion rate in this sense also seems to have retracted, from 0.8% to 0.68% for new visitors, as well as 5.8% to 5.56% for returning users. At the same level of confidence, the difference in conversion rates for new visitors are in this way also statistically significant ($Z=2.52$), while for returning customers there were no statistically significant oscillations in terms of sales ($Z=1.53$). One interpretation for this is that while we were able to attain more new visitors, especially due to the organic channel, these users manifest lower conversion rates as expected, while sales for existing customers remained relatively stable. Were this “false” new visitors (*e.g.* blocking cookies), theory would suggest that conversion rates would remain unaltered.

The proportion of transactions from returning customers is on the other hand of 84.1% during the first period and 83.5% for the second. This is not a statistically significant difference ($Z=0.65$), which also indicates that the proportion of buying customers we can identify as returning also remains stable.

On the other hand, the conversion of engagement objectives also decreased for the second period, from 9% to 7.2% for new visitors and 27.84% to 26.03% for returning users in relation to pages seen per visit (goal 8). This was for both cases a statistically significant decrease at 95%, ($Z=11.89$ and $Z=6.01$), particularly relevant in the 2% drop in the case of newcomers.

4.3 Acquisition

4.3.1 Definition

The acquisition reports predominantly refer to our main traffic channels, analyzing the precedence of visitors and assessing the performance of campaigns. In this section, we can also explore the most used keywords that lead to our site with

regard to organic search engine visits, and the way social media contributes with traffic and sales. For GA to identify the provenience of users, it is nevertheless necessary to reference the campaigns, linking the precedence of a user to his behavior during a session and registering interactions with other channels over time. This is a simple procedure, which involves the customization of URLs through the introduction of a set of parameters which allow GA to automatically identify each campaign. To help us in that task, Google also provides a free URL builder, which automatically assigns a new URL containing the required information (Google Inc., 2014). In order to set up a custom campaign, the parameters must thus be added to the end of each URL, using proper syntax and respecting the defined structure. We should also be aware that GA is case sensitive, so that “google” is different from “Google”. However, setting up campaigns does not require any modification to the GATC.

Google Inc. (2014) thus identifies a total of five parameters to keep track of referrals or to provide campaign information. The following list contains the three main parameters, used to identify traffic sources:

- **utm_source**: used to identify the website or the advertiser;
- **utm_medium**: used to identify the marketing strategy;
- **utm_campaign**: used to identify the campaign name.

There is thus a distinction between marketing source, medium and campaign, constituting each of these attributes a different dimension possible of being evaluated independently. The Google source might for example contain multiple mediums, such as organic or CPC, but certain mediums might be also be presented in various sources, such as Organic Google or Yahoo. Other parameters include **utm_term** and **utm_content**, which may be used to provide additional information, regarding paid keywords in the first case or to differentiate between similar contents or links using the same ad. The parameters correspond to a specific structure and should be separated from the URL using a question mark and from each other using an ampersand. The following is an example of a custom campaign for the web source, using the banner medium for the apple store campaign:

**redcorp.com/WebShop/AppleStore/Home.aspx?utm_source=web&utm_medium=ban
ner&utm_campaign=AppleStore_Banner_10_10_2012**

Some channels however, do not require to be tagged, since GA automatically references these sources. For example, for active AdWords campaigns auto-tagging can be enabled in order for information to be automatically available in GA. Furthermore, incoming traffic from organic searches and referral websites is also automatically identified, with no need for any modification. Some of the best practices nonetheless include the consistent use of parameters, in order to guarantee the regularity of information. In this sense, fragmentation can make it difficult for us to identify the precedence of visits or get lost in the amounts of data. Lastly, Google also aims to preserve the privacy of users, so that no personally identifiable information (PII) should be collected using any of these tools. Campaign referencing may however sometimes be used to work around these rules, through the personalization of tags, for example in e-mail marketing. This is as we have said, against Google policies and may result in the closing of an account.

Some traffic sources are in this context common to all accounts, while campaigns differ depending on the strategy for the website. The main traffic sources are however *Direct traffic*, from users who enter the website using an URL or a bookmark, *Referrals*, which consist of links from other websites to our pages, *Search engines*, including organic and paid traffic and allowing us to analyze some of the most used queries to reach our site, and *Other* campaigns which we have configured as described supra (Waisberg & Kaushik, 2009).

The acquisition report in this sense has strict relation with the `_utmz` cookie, which is the file responsible for storing the information about each visitor's provenience. It has a default expiration period of 6 months and is updated every time data is exchanged between GA and the user. This may however pose the question of multi-touch conversions and how can we attribute credit to other channels which also contributed to the conversion. This is one of the topics covered in the next sections, in the multi-channel funnels analysis, with the comparison of different attribution models considering multiple interactions through the customer's lifecycle. These dimensions

are closely related to the `_utma` cookie, which registers each unique visitor's ID with a default expiration period of 2 years. We can nevertheless edit these values in the GATC, by adding a snippet which allows us to define the expiration value, as such:

```
_gaq.push(['_setVisitorCookieTimeout', value]) (Sharma, 2012b).
```

4.3.2 Analysis

4.3.2.1 Traffic Channels

For our case study while the *default channel grouping* identifies only seven main traffic channels, the overview report on the source and medium tab tells us there are 431 source/medium combinations, deriving mainly from the great amount of referrals and customized campaigns. On the other hand, the *default channel grouping* lists traffic by organic search, direct, (other), referral, display, e-mail and social channels (by volume of traffic respectively).

For the Redcorp website, the main acquisition channels (source and medium) are the organic search from Google, the direct channel, web campaigns and the e-mail (newsletters), in terms of volume of traffic. From an ecommerce conversion rate perspective, some of the most successful mediums are however the direct channel and e-mail with respectively 21.6% and 2.6% of all traffic to the site, an engagement of over 2 page views more and about 2 minutes over the average length of visit to the site, generating 30.5% and 8.1% of all sales during the time period. Direct is in this way the most revenue-generating channel. On the contrary, in spite of the great amount of generated visits by the Organic channel (53.4% of all visits to the site), the number of page views per visit is 2 pages below the site average, while visits last 1 minute less when compared to the site's average. The amount of generated revenue accounts in this sense for 23.6% of the site's total, making it the number 3 source of revenue. Still, when compared to the visits to value ratio, this is still a channel primarily dedicated to acquire traffic, since 69.8% of all visits who come from this channel are new. This is about 12% over the website's average, only matched by referral traffic (about 71% new visits), largely surpassed however in terms of traffic volume.

4.3.2.2 An Organic Issue

The organic channel is one of the main sources for driving traffic volume, with one of the highest relative percentages of newcomers. In this way, the Keywords section, primarily designed to help us assess the most common phrases used to reach our site, could be very useful in terms of exploring the users' interests, but also for optimization and SEO purposes. However, since late 2011 to every logged-in user Google started encrypting sessions via SSL (Secure Sockets Layer). This means they were switched to navigate using https and Google's secure search, tendency being followed by the major search engines (Kaushik, 2013b).

In practical terms, the aim of using secure search (https) is to protect the privacy of users, resulting however in (not provided) search terms for web analytics platforms. This hinders our chance of getting closer insights into the users' interests, exploring the way they reached our site. Optimization procedures (such as SEO) thus face new challenges, with professionals looking for alternatives to keyword analysis. This has in fact impact regardless of the methodology used, whether we are talking about *Page Tagging* or *Logs*. Clifton (2013) evidences that even browsers are now incorporating this feature, which means that using Chrome (which has already surpassed Internet Explorer with over 42% market share) or Firefox, we are already encrypting searches, having a huge impact on keywords.

What is however questionable is the distinction Google makes between Organic and AdWords traffic. While all the information concerning organic searches is gone, AdWords keywords suffered no impact by these measures (Clifton, 2013). Because of that, this is a questionable approach to privacy, where only non-paid services are affected.

This is particularly relevant in the case of the Redcorp website, where (not provided) keywords represent 94.9% of all the organic searches. Furthermore, from all the revenue-generating keywords, we can only identify one which does not contain the term "redcorp". So from all of these which generated income, we can only identify one that might have been originated from non-returning customers. If we exclude all searches containing (not provided) or the term "redcorp", as well as the terms where the percentage of new visits is greater than zero, the total amount of visits which we

can maybe relate to new visits is only of 2% of total traffic and 0.01% of revenue. In this way, the analysis of keywords becomes irrelevant, with no particularly actionable information deriving from it.

4.3.2.3 Keyword Alternatives

Some of the solutions for this, proposed by Kaushik (2013), are to make use of tools which provide analogous information in order to overcome the limitations of “not provided”. In the ambit of SEO for example Google Webmaster Tools, Google Keyword Planner or Google Trends are already options used by professionals, with a slight different approach from web analytics. Trends for example, gives us an analysis of the most common search terms, according to four different parameters: type of search (web, images, YouTube...), geographical location, look back period and category of terms. However, this tool only gives us a normalized variation on the relative level of interest of a search phrase (Price, 2013). Because of that, the AdWords Keyword planner might be a great addition for paid advertising, giving us ideas for keywords, inherent cost and assessing their performance (Alpar, 2013). Lastly, the Webmaster Tools are an essential component of SEO, allowing us to better comprehend the website from Google’s perspective. In this way, we gain insight of what pages might have been indexed, what links are referring to our pages and the most common keywords used to reach our site. The Webmaster Tools actually provide a more holistic approach to keywords, giving us a role of indicators to assess search performance: the **number of queries** which returned pages from our site, the specific **query** which we are ranked for, the number of **impressions** from searches in which our pages appeared as a result, the number of **clicks** on our site’s listing, the **CTR** (click-through rate), which is the number of impressions that actually generated a visit to our site, and the **average position** in the SERP (search engine results page). These indicators allow us to either evaluate over time trends, drill down into specific keywords and refine each of our pages’ strategy (DeMers, 2013).

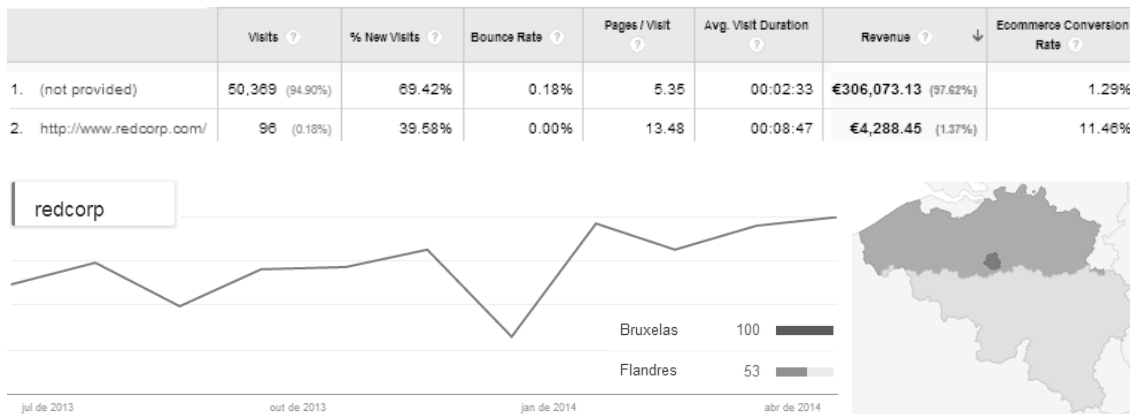


FIGURE 17 - ANALYTICS KEYWORDS REPORT AND GOOGLE TRENDS FOR THE TERM “REDCORP” (12 MONTHS)

4.3.2.4 Traffic Sources and Mediums

The *All Traffic* reports on the other hand allow us to identify what other channels are contributing for the acquisition of incoming traffic, demonstrating in a similar structure the main indicators of performance and engagement, with the possibility of drilling and combining additional dimensions. These sections, as highlighted by Google Inc. (2014), focus mainly on the users’ ABC cycle – Acquisition, Behavior and Conversion.

In the Redcorp website we specifically see that in absolute terms, both direct and organic mediums account for the highest number of both visits and value. However, there is a much different visitor behavior in relation to each channel, with the organic medium exhibiting the lowest average values for session engagement as well as the lowest rate of visits to value. In this sense, in spite of generating over 53% of all traffic to the website, it generates only about 23.5% of revenue. This may be problematic because while we can infer by mere logic that direct and e-mail are for the great majority returning users, we can only see the variation of the percentage on new visitors for the organic and referral channels, in order to comprehend what are the major trends in our website. Had we access to this information, we would be able to explain the low conversion rates for organic and referral traffic, when compared to the direct and email channels. The latter are obviously connoted with returning traffic, while the former originate traffic from external sources, hence the high percentage of new visits (website average at 58%). Still, the major problem here is our access to the users’ cookies, which makes it more appropriate to do an internal comparative

analysis, rather than a consideration of the absolute data (Clifton, 2012a; Kaushik, 2010b).

	Visits ?	% New Visits ?	Bounce Rate ?	Pages / Visit ?	Avg. Visit Duration ?	Transactions ?
2. (none)	21,427 (21.56%)	56.45%	0.46%	9.54	00:05:05	989 (33.47%)
4. email	2,569 (2.58%)	17.98%	1.36%	9.22	00:05:34	171 (5.91%)
11. referral	6,625 (6.66%)	71.47%	0.42%	6.53	00:03:02	122 (4.21%)
13. organic	53,077 (53.40%)	69.79%	0.18%	5.30	00:02:31	673 (23.25%)

TABLE 10 -TRAFFIC SOURCES PER NUMBER OF PAGES PER SESSION

The importance of analyzing the precedence of visitors primarily has to do with the task of determining which channels might be generating visits and revenue, so we can see which areas to invest, analyzing the effectiveness of our campaigns. In this case, we can see there is a clear behavioral difference between new and returning customers, with the latter exhibiting a much more similar behavior to visitors from other sources. On the contrary, new visits (which constitute almost 69% of visits from the organic channel), are less involved with contents, which results in fewer conversions. Even so, organic and referral sites are the only sources which might be truly acquiring new traffic, with the organic channel registering an increase in the percentage of new visits, as we saw from the intelligence reports, which contributed with an increase of 10% in the number of total traffic for the month of March. This channel has in fact registered a steady weekly increase in the percentage of newcomers, from 66% to 82% over 11 weeks. Referral and organic when compared to other channels, have respectively 25% and 21% more new visits in general terms, thus constituting the main channels for obtaining new prospects.



		Visits ? ↓	Pages / Visit ?	Avg. Visit Duration ?	Transactions ?	Ecommerce Conversion Rate ?
google / organic	New Visitor	35,397 (68.74%)	3.67	00:01:13	83 (11.74%)	0.23%
google / organic	Returning Visitor	16,095 (31.26%)	9.16	00:05:31	624 (88.26%)	3.88%

FIGURE 18 - WEEKLY % OF NEW VISITS, FROM 66% TO 82%; AND INDICATORS FOR ORGANIC NEW AND RETURNING VISITORS

During the first week of data there was a Google/CPC campaign running, which brought almost 3 thousand visits to the site, with 83% of new visits. This was however a poor performing campaign with 96% of sessions with about 2 pages seen and an average session length of only 9 seconds. This was a campaign targeted at a specific product, with a product landing page for the Fujitsu Lifebook e753. However, the goal conversion rate was only equivalent to 1%, and specifically for engagement goals. Furthermore, almost half of these conversions occurred from returning visits, which means only a small fraction of the visitors acquired by this campaign returned to the website (15% of returning visits). Still, no ecommerce conversions were so far attributed to this campaign. Hence no monetary value was assigned to it.

Besides this, no other major advertising had been running during the time period, apart from social media advertising. These consisted primarily on sponsored posts and display ads (on Facebook), which aimed to engage with customers, generate more “Likes” to the page and to drive new visits to the website via posts on relevant themes for our target audience. In this sense, most of these sponsored posts focused primarily on ICT, targeted per language, adult male users living in Belgium, interested in the theme of technology. Below, we can see the indicators given by Facebook for the period, which are for the most part analogous to those used by GA and the Webmaster tools, containing the number of impressions (number of visualizations of each ad), clicks and actions. This last indicator includes likes to the page and the installation of apps, without users necessarily clicking the ad. Other difference is CPM, which refers to the cost per 1,000 impressions, beyond CPC (cost per click).



FIGURE 19 -FACEBOOK AD MANAGER METRICS

In general terms however, GA tells us that social media accounts for only 0.35% (Facebook and LinkedIn) of the traffic acquisitions with a value of only 0.2% of all total revenue (from LinkedIn). In May, Redcorp’s Facebook page had 653 likes, while LinkedIn had only 123 followers. These are relatively inexpressive values, which can however present a window of opportunity from unexplored marketing channels and the acquisition of new traffic. LinkedIn seems in this sense to be the more effective social platform, which in spite of having an average of about 19% session engagement, has an ecommerce conversion rate of 4.55%. This represents over a half more than the average of other sources. However, 58% of the visits for the whole time period happened between the 3rd and the 5th of February, with 72% of the revenue generated on two transactions (from a total of four) on the 5th of February. These are in this sense inexpressive numbers, which might still indicate an area to gather further prospects.

Visits ?	% New Visits ?	Pages / Visit ?	Avg. Visit Duration ?	Transactions ?	Revenue ?	Ecommerce Conversion Rate ?
88 % of Total: 0.09% (99,400)	51.14% Site Avg: 57.75% (-11.46%)	5.89 Site Avg: 7.22 (-18.45%)	00:03:00 Site Avg: 00:03:43 (-19.30%)	4 % of Total: 0.14% (2,895)	€2,514.01 % of Total: 0.19% (€1,330,766.32)	4.55% Site Avg: 2.91% (56.07%)

TABLE 11 - INDICATORS FOR THE LINKEDIN.COM SOURCE

4.3.2.5 The Web Source

Another issue with the configuration of campaigns for this site is related to the “web” source and corresponding traffic mediums (bort and banner). This is one of the main sources of acquisition, generating high user engagement and also exhibiting high conversion rates. This seems therefore to be a very effective channel, with interested visitors that not only browse through multiple pages but also buy products. However, this is a misleading assessment, since all mediums contained in the web source correspond to internal campaigns. In this case, *bort* medium corresponds to the

“related products” section, while *banner* correspond to the tagged banners on the website. So this is a configuration issue which might induct in error, since this is not an external channel, posing attribution problems.

That said, campaigns are normally assigned to a conversion using the traditional last click interaction model. In this sense, the last campaign or source consulted by a user before a conversion will be assigned the value of that visitor’s conversions. This is the simplest model of attribution, considering only the last interaction as the determinant channel leading to conversion (we should here recall that information regarding campaign acquisition is stored for a default 6 months on the user’s `_utmz` cookie). There is however an exception to this rule, which has to do with direct accesses to the site. In this case, GA does not overwrite campaign information if the last interaction is originated from a direct visit. The rationale behind this is that if it weren’t for the previous campaign, the user could not have reached the site, so it makes sense to attribute the acquisition to that channel (Reynolds, 2010).

In this case, what is happening is that users might be coming to the site from different channels, for example via social referrals, but information is overwritten by contents consulted by users’ naturally navigating the site. If for example a user coming from Facebook makes a purchase, but only uses the internal search bar, the products section on the main page or the navigation tabs, the social media referral will be correctly assigned the value originated from that visit. However, if on the other hand a user recurs to any of the web shop banners or the related products section on the product pages, the conversion will wrongly be assigned to one of the internal campaigns (`web/bort` or `web/banner`) and not the original source of precedence.

This methodology, while in fact helpful to determine which the most used sections of the website, leads to information loss concerning the original acquisition channels for our visitors. This also explains the low percentage of new visits assigned to these (internal) campaigns, since acquisitions are made when visitors are already in the site. A customer might in this sense arrive to the website as a new visitor, but because session information is preserved, the moment he clicks the banner or related products section he is already considered a returning customer. Because of that,

indicators for this source are well above average, with this being the most revenue-generating channel, with a very low percentage of new visits.

	Sessions ?	% New Sessions ?	Pages / Session ?	Avg. Session Duration ?	Transactions ?	Ecommerce Conversion Rate ?
1. web	12,541 (12.62%)	4.85%	12.42	00:07:11	941 (32.50%)	7.50%
2. (direct)	21,427 (21.56%)	56.45%	9.54	00:05:05	969 (33.47%)	4.52%
3. google	55,265 (55.60%)	70.41%	5.14	00:02:24	664 (22.94%)	1.20%
4. newsletter	1,494 (1.50%)	15.73%	9.25	00:05:41	103 (3.56%)	6.89%

TABLE 12 - ACQUISITION SOURCE PER GENERATED REVENUE

In conclusion, it is safe to assume that most of the business value relies on returning users, constituting a great amount of traffic and ecommerce transactions. In this sense, the major channels for incoming visits are those which indicate previous contact between the organization and the user, such as direct and e-mail. This also emphasizes the importance of establishing relational connections with customers in B2B. As proposed by Miletsky (2010), since in this environment sales-cycles can be fairly long, web resources should therefore focus on reinforcing the brand name. Leek & Christodoulides (2012) also highlight the importance of trust in B2B relations, reducing the perceptual risk and proving the company can deliver efficient practical solutions. Having available sales representatives can in this way help motivating potential clients to take action, something which is an important part of this company's business model.

B2B websites, should thus be organized according to this perspective in a formal manner, providing brochures and informative videos, authoring papers on industry topics, offering password-protected client areas and presenting the contact information, as well as a description about the evolution of the company (Miletsky, 2010). These are all areas already explored by the company, with regular uploads made to the YouTube channel, as well as a newsletter and a news section in the website, shared through regular social media updates. There is also an available about page containing the company information, as well as the contacts for all representatives from the company. One of the main problems with B2B, especially in

the web, is however adjusting our filters to get the attention from business owners and managers, selecting the right communication channels according to the target audience.

4.3.3 Summary

To sum up this section the main bullet points are:

The acquisition section focus on the ABC cycle, aiming to identify the source of acquisition of conversions and exploring the typical behavior of users associated with each traffic channel.

The main traffic mediums for the Redcorp website are by far organic (53.4%) and direct (21.6%), followed by referral (6.7%) and email (2.6%) traffic. CPC in spite of being associated with 3% of the traffic volume was a one-time poorly conducted campaign, generating no ecommerce conversions.

The main revenue generating source/mediums are on the other hand direct (30.5%), web/bort (24.4%) and google/organic (23.3%), followed by web/banner (8.5%) and the newsletters (3.4%). Some of the highest ecommerce conversion rates are therefore registered for the internal web sources (bort and banner mediums, with respectively 7.3% and 8.9% conversion rate). Direct and email also exhibit expressive numbers at 4.3% and 6.9% conversions. On the other end of the spectrum, Google/organic only exhibits a 1.3% rate.

There is a clear difference between organic returning and newcomers, with 68.9% of new visitors from organic at only 0.2% conversions. Contrariwise, 3.9% of returning organic visitors make a purchase, exhibiting high engagement levels (9.16 pages and 5min30secs compared to 3.67 pages and 1min13secs per session) constituting 89% of organic revenue.

Organic and referral are the channels generating a significative amount of traffic from newcomers, with respectively a 69.8% and 71.5% percentage of new visits for each of these mediums.

On the other hand, social media is still an inexpressive channel, generating few visits and conversions in spite of the campaigns developed and regular YouTube activity. Special attention to the LinkedIn should be paid though, which might reveal important in the future, especially in a B2B environment.

The keywords report in this section can tell us little about our users' interests, since 94.9% of keywords are not provided. Still, Google Webmaster Tools and Google Trends can be used to overcome these impediments and exploring the most searched queries that lead to our site or to improve search engine marketing.

Lastly, some configuration issues emerge in this case, since the web source contains the referentiation of several internal mediums. This makes it difficult to trace back the original acquisition source of users (particularly new), since campaign information is being overwritten and traced back to these channels. Also, the newsletter source must be consistent over time, associated with the e-mail medium. Contrary to what has been happening, since for each month a new newsletter is being referenced as a different source for the email medium. This could instead be information contained on the campaign parameter, rather than the source parameter (which results in the fragmentation of information).

4.3.4 Period Comparison

- The majority of visits (53.4%) is originally acquired by the organic channel, followed by direct traffic (21.6%), Referrals (6.7%) and others. However, only 23.3% of revenue is attributed to organic, while direct (30.5%) and web/bort (24.4%) collect the most value.

During the first period the channel driving most traffic was undoubtedly the organic, having had the highest amount of acquisitions in both absolute terms of visits as well as percentage of new sessions, second only to referral sources. In terms of relative importance, while referral maintained its importance of about 6% for both periods ($Z=0.38$), organic had a statistically significant increase to 57.9% of total traffic ($Z=20.81$). This reinforced the channel's position as number one driver of sessions,

especially due to the increase on the percentage of new sessions. While during the first period these were about 69.8% of traffic for this channel, the numbers increased to 73.2%, a statistically significant difference ($Z=13.1$). Consequently, returning organic decreased in proportion in relation to total website sessions, from 16.1% to 15.5% at a statistically significant difference at 95% confidence ($Z=3.97$). On general terms, the percentage of new sessions also increased for all major channels, from 57.8% to 61.8% of total traffic ($Z=18.9$).

In terms of revenue however, direct was again the channel with the highest value, from 30.5% to 33.2% of turnover. The total number of transactions however had only a 0.1% oscillation, with no evidence of difference between the two periods ($Z=0.07$). The same happened for the second channel, concerning the web source, including the “related products” (bort) and banner campaigns, which went from 23.14% to 21.76% of operations ($Z=1.27$), and the organic medium, from 23.25% to 22.59% ($Z=0.6$). These were thus differences without statistical significance at 95% confidence. However, in relation to the web source there are indeed clear differences between mediums, with bort collecting over 91% of transactions. The conversion rate between the two periods also went down for this and the banner campaign, from a significant 9.04% to 7.76% ($Z=3.03$) for bort, as well as from 5.51% to 2.86% for banner ($Z=1.59$), a value with no statistical significance due to the small number of hits.

On the contrary, one of the channels with the biggest improvements was email, almost doubling the number of total visits between the two periods, from 2.58% to 3.71% of total traffic volume, a significant increase at $Z=14.84$. Likewise, the number of total transactions also followed this trend, from 5.91% to 9.73% of operations ($Z=5.45$). In spite of this increase, the revenue value corresponding to the operations had a shier evolution, from only 8.1% to 9.8%, with the channel maintaining exactly the same conversion rate, at 6.7% transaction conversion. What this reflects is that the behavior of newsletter subscribers remained very stable, with sales driven by an increase in the absolute number of visits.

As for the direct channel, ecommerce conversion rate decreased slightly from 4.52% to 3.97% ($Z=4.98$), while overall values for organic also decreased from 1.27% to 0.99% rate ($Z=4.61$). This difference is mainly due to the high number of new sessions

for the second period, which reflects in the aggregate values of performance for the channel. Drilling down into the data we can thus see that Returning Google organic only oscillated from 3.93% to 3.88% conversion rate, a difference that exhibits no statistical significance ($Z=0.24$). Organic New on the other hand had only a 0.2% conversion rate, with the increase in their absolute numbers having an impact on the aggregate values for the whole channel.

Lastly, again failing to impress is the social source, with only 0.35% and 0.19% of total sessions, as well as 0.14% and 0.1% of transactions. In this way, while the decrease in visits was statistically significant ($Z=7.25$), the number of transactions was not ($Z=0.44$) at 95% confidence. The proportion of new sessions is also not significantly different from the website's average ($Z=0.8$), which means that there is no evidence of the channel being particularly effective in generating new prospects. In its constitution, the social channel includes Facebook as the main driver of traffic, with a significant increase from 56.32% to 65.77% of all social traffic ($Z=2.25$), LinkedIn which went from 25.29% to only 5.41% ($Z=-6.08$), and YouTube with 8.91% to 22.52% ($Z=4.66$) of social. Transactions however are very sporadic, at only 4 operations for LinkedIn during the first period and 3 for Facebook during the second.

- The main channels generating new prospects are organic (69.8%) and referral traffic (71.5%), representing a significant difference in relation to the website's average.

One of the issues we have mentioned throughout this work with the page tagging methodology is the fact that users can restrict access to cookies, making it inaccurate to interpret literally the data for new sessions. On the contrary, Kaushik (2010) evidences the importance of understanding the overtime evolution of the percentage of new visits, interpreting website trends at the light of ongoing campaigns and actions, rather than considering the exact values for this particular metric. One example of that is the direct channel, which we would intuitively guess it would have a low percentage of new visitors, given that users must already know the URL before they come in the site. In that way, only visits from returning customers on new devices

or offline campaigns (e.g. flyers or business cards) could bring new visitors to the site through this source.

For the first period however there were 56.45% of new sessions for this source, just under the 57.78% website average, still registering a statistically significant difference at 95% ($Z=3.57$). Contrary to that, during the second period this percentage increased to 61.47% for direct, in line with the 61.77% for the website's average, exhibiting no significant differences ($Z=0.89$).

On the other hand, generating not only the highest number of visits, but also new visits is the organic channel, which went from having 69.79% new sessions ($Z=46.01$ when compared to the website's average), to a significant increase to 73.2% during the second period ($Z=12.09$). That fact reflected an overall increase in traffic volume, mainly due to these new visits. In this way, the website had during the first period 57.78% new visits, while during the second 61.77% of sessions were new ($Z=18.9$). On the same page, the most consistent channel in bringing new prospects to the site are referral sources, with 72.14% and 74.49% for the two periods, a slight but significant increase ($Z=3.05$), in a medium that has only just over 6% of total traffic for both periods.

At the other end of the spectrum, email is non-surprisingly the external source with the highest proportion of returning customers. This is obviously due to the fact that users must sign up to the newsletter in order to receive emails, driving them to the site. Still, from 18%, the number went up to 27% of new visits in the second period, corresponding to 2.58% and 3.71% of total traffic for the periods, a small but significant increase ($Z=15.88$).

4.4 Behavior

4.4.1 Definition

The behavior section contains reports which help us comprehend the performance of the elements and pages on the site and the way users interact with it (Google Inc., 2014). Because of that, this report's dimensions focus primarily on the site's features, adopting metrics which are not only indicators of behavior, but which

provide technical information for the improvement of the website's functionality. As a result, the overview section begins by presenting some general data about each page's performance (content section), focusing on the page (url extensions) and page title categories. The main indicators are in this case the number of page views, time on page, bounce and exit rates. Internal search terms are also dissected in this report, with the indication of keywords and number of occurrences per search. Lastly, we look into the events triggered, which are defined by the user by calling the `_trackEvent ()` method in the source code of the web page.

In this case, we will mainly mention the uses of this feature, since for our case we are only tracking two events – the utilization of the search bar, for internal search phrases, and the clearance of shopping carts (revealing users which drop off in the middle of an ecommerce conversion). However, the default site search report already covers the first event more effectively, while the second has a relatively unexpressive number of occurrences, with only 77 unique events for the first period. Clifton (2012) in this sense refers to the usage of event tracking mainly for in-page elements which do not generate a page view. Because of that, events are independently reported, especially useful in the case of dynamic content such as embedded Ajax or Flash elements, downloads or outbound links. This could for example be an appropriate substitute to our poorly configured internal Web campaigns.

Tagging events also allows us to distinguish between bounced visits and exits, due to the fact that bounces correspond to visits which only generate one page view. These are commonly associated with bad user experiences, interpreted as lack of interest in the part of the visitor. However, we argue that single-page sessions can also be related to effectiveness. In single-page websites or a campaign with a properly defined landing page, visitors might still have meaningful experiences while going through only one page. A way to solve this issue is therefore resorting to event tracking, due to the fact that when an event occurs, single-page exits will no longer be considered as bounces. GA in this way considers an interaction with page elements, reflecting clear interest on the user's part.

As for the Content reports, GA primarily focus on the performance of pages, responses of visitors, the pages' value (given by (Transaction revenue + Goal value) /

Unique page views), as well as the pages' loading times. This last is a particularly important report in order to assess the more technical aspects of our website, its development and the way pages respond to different devices. As we have seen, website loading times are one of the important aspects of customer loyalty, contributing for online surfing experience and customer retention. According to this perspective, a study by Akamai Inc. (2009) considers quick page loadings essential for a satisfactory ecommerce experience, with the online environment also influencing traditional physical environments.

According to this, two seconds is the acceptable threshold value for 47% users, while about 40% abandon the web shop after a period of 3 seconds waiting. This also affects sales in the short term, with up to 79% of users who go through a bad online experience affirming they will not return to the same website. 27% of times this will also affect the perception associated with physical stores, and consequently their sales. One of the main problems with poorly-developed web pages is not only short-term losses, but especially the effect on the long-term. When waiting for a page to load, visitors become distracted, leaving the site or start looking for other options. Because of this, speed is determinant for user engagement and customer retention.

This is however a study conducted in 2009, introducing the evolution of customers from 2006 to 2009. Visitors' expectations are in this way continuously increasing, with the development of new technologies and the higher speeds of internet connection available in the market. One of the emerging channels is in this line of thought mobile shopping, with the proliferation of smartphones and tablets in the communications industry. This introduces a new field of research in the disciplines of web design and development, which is reflected in the concept of responsive web design (EDIT, 2014). This is a concept tied to the optimization of websites to multiple platforms, meeting the demands of both regular and mobile users.

Due to the large number of devices and different configuration of screens with internet access, websites are now challenged to respond to context, using fluid grids and flexible images. This is particularly relevant for mobile users, due to the great variety of screen sizes and generally slower speed of internet connection. Responsive

design thus takes in account the diversity of platforms for the development of websites, having in mind the multiplicity of devices that can access our website.

4.4.2 Analysis

4.4.2.1 Site Speed

In our case study, we have already seen the greater importance of regular desktop traffic over mobile connections, not only in terms of value but also the total amount of visits, page views as well as average session engagement. We can also see in the Page Timing report that most pages load in up to 3 seconds (73%), while 93% do so in the first 7 seconds, for all sessions. There is however a clear difference between mobile and non-mobile traffic, since while 74% of non-mobile pages take up to 3 seconds to load (32% only take up to 1 second), 61% of mobile take between 1 and 7 seconds, where most (41%) take 3 to 7 seconds. Furthermore, about 26% of all pages for mobile take 7 to 13 seconds to load, which is a considerable amount of time and sessions, while only 4.7% of pages for non-mobile took that long to load. There are therefore clear discrepancies in terms of technology and the necessity of adapting contents and objects to different devices.

It is however important to explore the relative importance of mobile traffic during the decision cycle, which in this case might be relatively inexpressive. Since we are dealing with B2B and mostly high involvement purchases, these involve multiple interactions, mostly in an office environment. Even so, the inexistence of a mobile version also hinders the possibility of decision processes going through these platforms. During this period, only about 4% of all sessions came from mobile.

Still, one of the features which might help identifying problems and optimizing user experience is the speed suggestions tab, which presents web developers with automatic recommendations for the technical development of each page. This analysis is subdivided into mobile and non-mobile, evaluating both user experience (legibility and interaction with elements), as well as speed, considering back-end code, images and further recommendations. The homepage (default.aspx), which is also our main landing page, in this sense has a classification of 80/100 for desktop, while mobile

experience is evaluated at only 59/100. Poor mobile evaluation is in this case mostly due to with legibility and dimensioning issues, as well as the lack of a viewport for adapting the visualization of the website to these devices. Currently, all pages are being processed the same way both for mobile and non-mobile traffic, resulting in poor legibility and functionalities for the mobile audience.

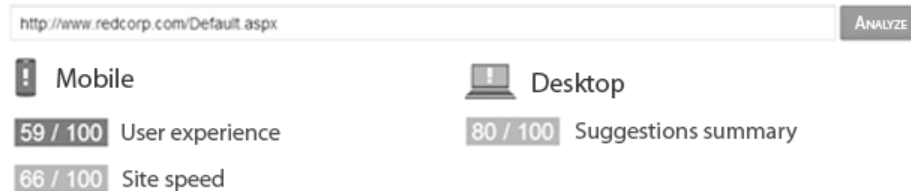


FIGURE 20 - PAGE SPEED SUGGESTIONS FOR THE DEFAULT.ASPX PAGE

Each page might in this sense be individually subjected to this evaluation and compared to the site's average using the page timings report. In this view, we have access to each URL, having indication on the number of page views, which indicates the most consulted pages, as well as the percent variation for each page in relation to the average for all sessions or for certain segments. Here we can see that the home page has a slightly better performance than most pages, with an average waiting time of 2.71 for all sessions. In the case of non-mobile this value is slightly reduced for 2.68 seconds, while mobile has an average value over 121% more than the site average, at 7.13 seconds.

Still, while non-mobile homepage views account for 14.7% of all visualizations, mobile homepage represent only 0.2% of this number. Again, the Fujitsu notebook (associated with CPC campaign) is mentioned as a poor performing landing page, one of the slowest to load due to mobile traffic. This has been the most consulted page for this segment, with a very poor performance skewing the values for other observations. The average loading time was for this page almost 25 seconds, while the second most visited homepage had an average loading time of 7 seconds. In the case of non-mobile users, the most popular pages are also the best performing, with loads averaging at 3.21 seconds.

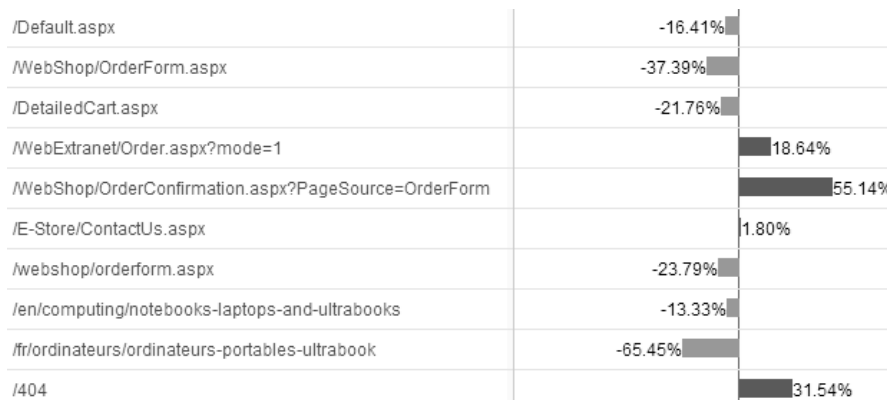


FIGURE 21 - PAGE LOADING TIME FOR NON-MOBILE (COMPARED TO 3.21 AVERAGE)

However, page loading times are not only associated with the devices but also the technology used to access our site. Because of that, it is important to have in mind our target audience and the resources at their disposal. That is why strategic planning and website development is crucial for performance, considering the type and number of elements to integrate on each page. As stated, it is not only important to adapt contents, but also promote legibility or reduce heavy content for our target audience. This has a clear relation with mobile traffic, but also the geographical distribution of our target audience.

For our case, this is not however a critical issue, given that most of the visits (54.5%) and revenue (91%) provide from urban areas from Belgium, followed by France (3.3% and 2.8%) and Germany (3.4% and 2.5%). The speed of connection for these countries is because of that very high, with few exceptions in France. Belgium therefore has an average of 2.23 seconds, 27% less than the site's average. In this case, only 17% of all page loadings for non-mobile take more than 3 seconds, while only less than 4% take more than 7 seconds to load. On the other hand, for Germany only under 32% of pages take more than 3 seconds, while less than 5% take more than 7. In France however, 77.5% take between 1 and 7 seconds, while 9.3% of loadings take between 7 to 13 seconds. This is therefore a significant difference, especially in the Ile-de-France region, with an average time of 5.57 seconds.







Country / Territory ?	Mobile	Avg. Page Load Time	Pageviews ? ↓	Page Value ?
 Belgium	No	2.23	525,857 (72.58%)	€86.00(130.02%)
 Belgium	Yes	10.83	22,577 (3.12%)	€0.80 (1.21%)
 Germany	No	2.99	21,310 (2.94%)	€46.18 (69.81%)
 France	No	4.93	18,231 (2.52%)	€74.11(112.04%)
 United Kingdom	No	4.21	15,061 (2.08%)	€13.62 (20.59%)
 United States	No	6.15	14,345 (1.98%)	€10.12 (15.30%)
 Netherlands	No	2.36	13,939 (1.92%)	€9.65 (14.59%)
 Luxembourg	No	2.54	8,546 (1.18%)	€35.03 (52.95%)

TABLE 13 - PAGE VIEWS AND AVERAGE LOAD TIMES BY COUNTRY AND DEVICE

4.4.2.2 Internal Search Usage

Still in the behavior section, the search report usage is dedicated to exploring the use of the internal search engine. For websites with great diversity of products (such as Redcorp’s), this may be a critical feature for user experience and ease of navigation. Moreover, having an internal search system also provides additional sources of information for research, since it is an opportunity to assess the phrases and topics users are interested in. As highlighted by Clifton (2012), this information may in some cases not only be used by marketers, in order to improve campaigns, but also content creators, product managers or other functional areas of the company. Additionally, it is also possible to follow the users’ behavior in the product research process, the number of interactions, search refinements or if this feature is helping improve user experience and generating conversions.

We might however argue that referring to unique search terms in order to extract actionable insights may be a frustrating effort, due to the high number of different phrases searched by users. In this sense, the number of terms for the given period was 43 437, from 65 953 total searches. This means that most terms are only used once, while only two terms correspond each to about 1% of researches – the phrases “Toshiba Z series” and “netgear”, with the third most popular phrase is “Toshiba” at only 0.2%. The high number of different terms thus hinders the possibility of drawing meaningful conclusions from these reports.

Still, we can recur to the usage report, which simplifies our approach dividing users by interaction, comparing those who use the search feature versus those who don't. This is a simpler, more useful procedure which reflects that internal search is an important feature in the website, generating great interaction with content. For the Redcorp website we can see that almost 28% of all visits use the internal search feature, having in average a much higher engagement than visits which do not. The number of pages per session is in average much higher (4.07 versus 15.42), as is the average duration of visits (1 min 41 secs compared to over 9 mins). Moreover, from the total number of transactions, almost 78% make use of the internal website searches, which is reflected on an ecommerce conversion rate of 8.18% (compared to 0.89% from non-search). These are meaningful numbers which emphasize the importance of internal search and tell us about the way users navigate the site. Search users are not only more engaged, but more likely to return (at least 68% returning versus only 33% from non-search) and generate higher revenue.

Because of this, we argue that buying sessions are originated from educated visitors, proactively assuming the direction of their sessions. Furthermore, it also reflects that the end of an ecommerce conversion cycle often ends with a session which uses internal searches, with less resort to other campaigns. Because of that, Internal campaigns such as *bort* or *banner* are in this case less important, playing an important role especially in the decision process. Buying sessions however, are more direct, driven by the user. Still, the web source (internal campaigns) was the last consulted channel for 20.3% of traffic, generating about 33.6% of revenue, showing that roughly a third of people (31% of unique transactions) use the banners or the related products sections as the last influencer (campaign) on their search process.

From all the traffic, about a quarter of transactions come from sessions in which both internal search and internal campaigns are used, while another quarter are direct visitors who also use the search feature. We should also notice that the traffic sources which generate the majority of visits and the highest percentage of newcomers (Google organic followed by Direct), have generally lower engagement and do not use site search. These are thus users browsing through the site, getting to know the products or still in the decision process.

		Sessions ?	% New Sessions ?	Pages / Session ?	Avg. Session Duration ?	Transactions ?	Ecommerce Conversion Rate ?
Visits With Site Search	web	5,718 (5.75%)	0.31%	19.41	00:12:08	734 (24.49%)	12.84%
Visits With Site Search	(direct)	9,920 (9.98%)	46.68%	14.50	00:08:02	858 (28.63%)	8.65%
Visits With Site Search	google	8,943 (9.00%)	29.68%	15.39	00:09:04	541 (18.05%)	6.05%
Visits Without Site Search	web	6,938 (6.98%)	8.23%	6.67	00:03:22	211 (7.04%)	3.04%
Visits Without Site Search	(direct)	11,825 (11.90%)	62.85%	5.65	00:02:46	188 (6.27%)	1.59%
Visits Without Site Search	google	46,154 (46.44%)	77.57%	3.20	00:01:07	166 (5.54%)	0.36%

TABLE 14 - SITE SEARCH USAGE PER SOURCE

4.4.2.3 In-Page Heat Map

Lastly, it is also worth mentioning the in-page analytics feature, which provides a visual representation of the clickstream and conversion rates associated with each web section. In this way, each page may be displayed in our browser, as it would for a regular customer visiting the site. The main difference is the chromatic hierarchy established with each link and image, hierarchizing hotter sections, associated with goal and segments. This is a simplified version approaching other types of usability tests, which aim to identify areas of denser activity and value. At the present date, Google has launched a new Chrome extension which allows us to navigate the website and select our metrics as we go, comparing different periods and pages. Below, there is an example of this application:

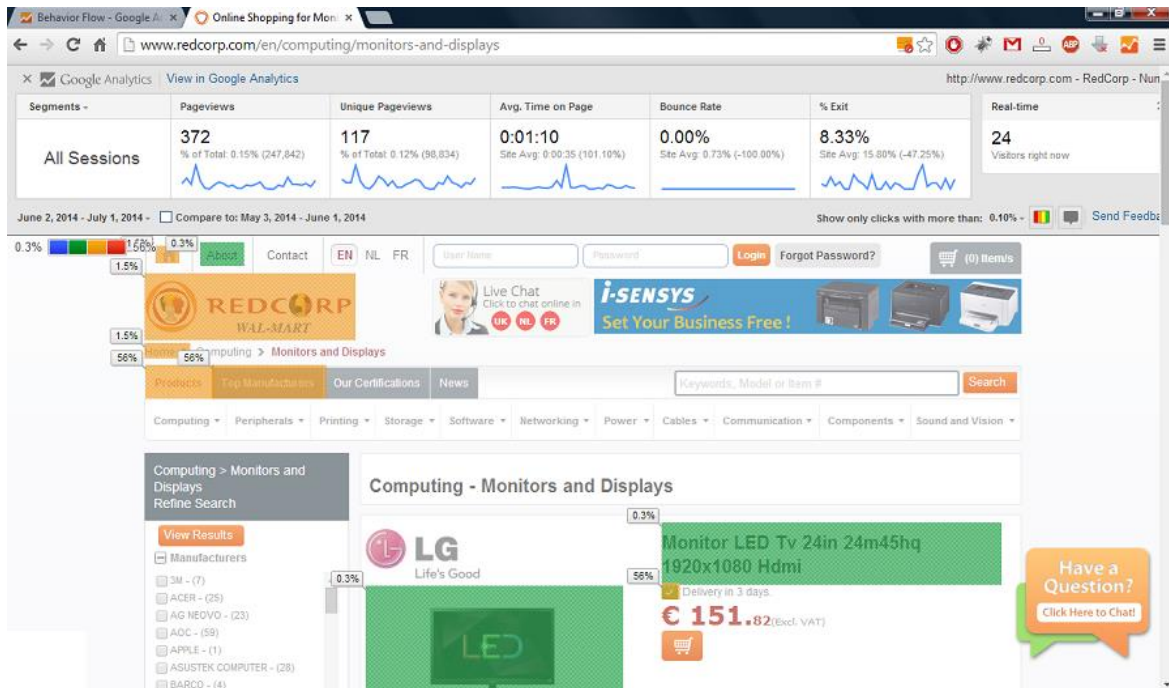


FIGURE 22 – PAGE ANALYTICS EXTENSION – CLICK RATE FOR THE “MONITORS AND DISPLAYS” SECTION

4.4.3 Summary

The behavior section provides technical information about our site’s architecture and the utilization of its features by users. In this way:

Page loading times were in our case influenced by the devices used to access the web site, with 71% of non-mobile users taking up to 3 seconds per page, while mobile varied with 41% between 3 to 7 seconds and 26% with 7 up to 13. These are differences that strongly affect user experience. Yet, only a little over 4% of visits provide from mobile users, which may be pose a chicken-and-the-egg problem.

In the suggestions page for the main landing page (default.aspx), the main issues had to do with legibility and dimensioning, as well as the lack of a viewport for mobile users, resulting in poor user experience for this segment.

At a geographical level, the Benelux region for non-mobile users thus respects the 3 second threshold value for page loading times, which also includes Germany, another important country for this firms business. On the other hand, these values are almost double for the UK, when compared to Belgium, and are much higher for France and the USA (almost triple the value for Belgium). Still, drilling down into different

regions and cities certain differences emerge, specific to the devices and each user. Nonetheless, there seems to be a geographical effect on load time due to distance.

Probably the most interesting insight drawn from this section was the behavioral differences of internal search users and non-users. Search users are about 27.7% of sessions, but have a higher return rate (32.85% new sessions versus 67.34% from non-search), with much higher conversion rates (8.18% versus 0.89% for ecommerce and 44.13% versus 6.55% for engaged users per page views). The combination of this with the source dimension also reveals the influence of marketing channels and its effect on conversion rates – particularly web.

4.4.4 Period Comparison

- Site search users constitute about 27.7% of traffic, with a third of new sessions, while about two thirds of non-search users are newcomers, with great impact on engagement and transactions conversion.

Indeed, the great majority of sessions visitors do not use the internal search feature, as we can see by the Search Usage report. The percentage of search users has in fact slightly decreased from 27.71% to 27.29% in the second period. A slight but significant difference at 95% confidence ($Z=2.18$). Again, both these segments also suffered a significant increase in the proportion of new visits, especially in the case of search users. In this way, from 67.34% the number of without search sessions rose to 70.21% ($Z=12.25$), while search sessions went from 32.85% to 39.27% ($Z=16.25$). This also seems to have had an impact on bounce rate, which went up from 0.12% to 0.36% ($Z=5.86$), as well as ecommerce conversion rate, down from 8.18% to 7.31% ($Z=-3.97$). Still, it was a much better rate of conversion than for visits without search, down from 0.89% to 0.76% ($Z=-2.85$), a statistically significant difference at 95% confidence.

Likewise, the engagement of visitors also went down for the aggregate value for the dimensions, from 6.55% to 5.29% of non-search visits with over 10 page views per session, as well as 44.13% to 38.65% of search users ($Z= 10.59$ e $Z=13.56$). However different types of users clearly have a distinct behavior with returning search having

45.51% engagement over 29.75% for new search users ($Z=-27.62$) and 10.95% over 2.86% for returning and new non-search users ($Z=-47.19$), for the second period. Also, depending on the user type, search users have always statistically significant higher conversion rates, even when comparing new search to returning non-search users (3.45% vs 2.8% at $Z=3.4$).

4.5 Conversions

4.5.1 Definition

The conversions report section is designed to help us assess the achievement of our goals in the website. In this sense, Google Inc. (2014) describes a conversion as the completion of an important activity to the success of the business by our users. Because of that, conversions refer not only to purchases, but also to key actions connoted with positive feedback from our visitors. The configuration of each goal is in this sense a manual task, which derives from our online strategy and sets benchmarks for the desired actions we wish our users will take. In the GA administrator screen we can therefore access the tab to define the conditions of new goals. Goal type might in this sense refer to a specific URL destination, the duration of a visit, the pages seen per session or the completion of an event (such as a visualization of a video). This last requires a set-up of an event, which might be different depending on the GA version we are using (Universal *analytics.js* or Classic *ga.js*).

As we have seen earlier, events refer to user interactions which can be tracked independently from a page load. These are frequently interactive contents, downloads, gadgets, flash elements, videos or other embedded elements. Using the `_trackEvent()` method we include the parameters to get information concerning the category of the objects tracked, the action made by the user, and three optional fields including label (string of additional dimensions), value (integer of numerical data) and non-interaction (Boolean)(Sharma, 2010). The following is an example of usage of this method for classic analytics:

```
<a href="#" onClick="_gaq.push(['_trackEvent', 'Videos', 'Play', 'Video name']);">Play video </a>
```

In addition to this, we might also want to track third-party outbound links, taking part for example in the conversion funnel. In this case, through event traffic, GA gives us the option to track outbound traffic by setting up an event for specific URLs. This uses the same configuration as event tracking and can help us determine some of the exit paths taken by our visitors, or assess the effectivity of a campaign or a call to action. Following is an example of usage for *analytics.js* given by Google Inc. (2014):

```
<a href="http://www.example.com" onclick="trackOutboundLink  
(‘http://www.example.com’); return false;"> Check out this excellent example. </a>
```

Additionally, goals can also be monetized according to the approximate value calculated for a conversion, deriving for example from past sales. This may in this sense be helpful for determining the ROI of campaigns, to assess the most effective channels by attributing an approximate value to each acquisition. One example is if we have an ongoing campaign where research tells us 10% of acquisitions end up making a purchase, we can take the average transaction (or customer) revenue in order to calculate the value of new acquisitions. If for example, the average order value is 100€, for a 10% commerce conversion rate each acquisition might be assigned a value of 10€ (Google Inc., 2014). In this sense, we can better determine how much campaigns are worth and the levels of investment that should be made on each channel (Clifton, 2012a). This methodology however requires constant monitoring and the interpretation of data, due to the dynamism and constant changes in values.

One pertinent issue to account for are the characteristics of customers, the nature of our products, as well as rate of conversions. As we have seen, some major differences occur between the B2B and B2C environments, with very different buying behavior and motivations. The B2B decision process is as we’ve seen generally much more complex, involving multiple levels of influence and hierarchy. Also the distribution of session value is much more skewed and unpredictable, so assigning value to each session might be much more difficult (or even undesirable). In order to attribute a value to non-transactional goals, we must first explore the impact of

conversions on our sales, on-line or not, trying to assess the correlation between engagement and transactional conversions for each segment and channel.

Tanner & Raymond (2012) also refer to the level of involvement required by products, which reflect the psychological relationship the consumer establishes with the product and the level of information he needs to make a decision. In this sense, this is a continuum between fairly routine decisions, which do not require a great deal of monetary or psychological investment, and heavy purchases with extensive consideration. The authors in this context distinguish between three levels of involvement, from low, to limited and high involvement. This perception of involvement is also something that depends on the personal characteristics of consumers, with some products obviously more commonly associated with higher levels than others.

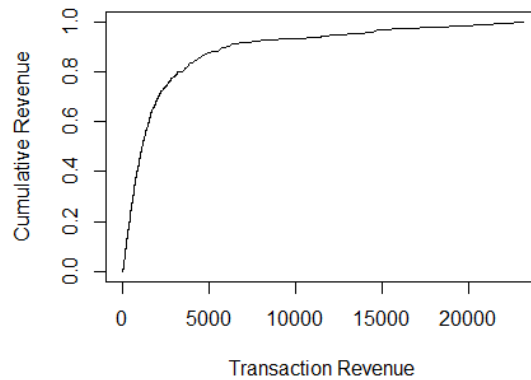
To different categories are also associated typical response behaviors, particularly in the case of low involvement purchases. Routine response behaviors are in this sense almost automatic decisions consumers make on a regular basis, centered on information gathered in the past. Impulse buying is an example of low involvement behavior, which while not necessarily a repeated action, reflects a low perceived risk. On the other end, high involvement carries high risks for the consumer, referring to more sporadic purchases, bearing great significance to the buyer. Because of that, it takes an extended problem solving process, where pre and post purchase assistance might be necessary for providing information and reducing anxiety.

4.5.2 Analysis

4.5.2.1 Types of Conversion

As we can see from the distribution of transaction revenue to our website, the attribution of value to non-monetary conversions must be carefully considered, especially for this type of market. This is because the distribution of visitor and transaction value follows a highly skewed distribution, with extreme values contributing for a great part of the business. Average session value must thus be taken in careful consideration along with the distribution of order revenue for both

transactions and customers. In this sense, while the great majority of orders is under the value of 506€ (75% of orders which account for only about 25% of the revenue), just a small minority of transactions accounts for most of the revenue.



```
> trans[2084,] transactionRevenue  cumperc
                    506.4  0.2502458
```

FIGURE 23 - PERCENTAGE OF CUMULATIVE REVENUE BY THE VALUE OF EACH TRANSACTION, WITH OBSERVATION 2084 (OF 2789 – 3RD QUARTILE) AT ONLY 25% CUMULATIVE VALUE (DATA FROM THE API)

Because of that, merely taking average session value for monetizing conversions might in this case be inadequate, since we would be characterizing very different sessions according to an one-dimensional behavioral category, knowing at start that there are many variables influencing session value. We therefore argue that this methodology fails to capture the customer’s lifetime experience, looking at visitors from an overly simplified session perspective.

In the case of Redcorp, the goals defined for the website primarily concern the engagement of users and the subscription of the newsletter. These are in that sense indicators which relate to session duration (Goal 7), the number of pages per session (Goal 8) and lastly the visualization of a the “Thank you” page after users have signed up for the newsletter (Goal 6). Goal 1 and 5 are also configured as destination pages which will however be disregarded in this analysis because these are related to the completion of transactions, which are already tracked by the ecommerce report included in this section. One of the factors we again verify is that conversion rates seem to be higher especially during weekdays, with lower rates for weekend sessions.

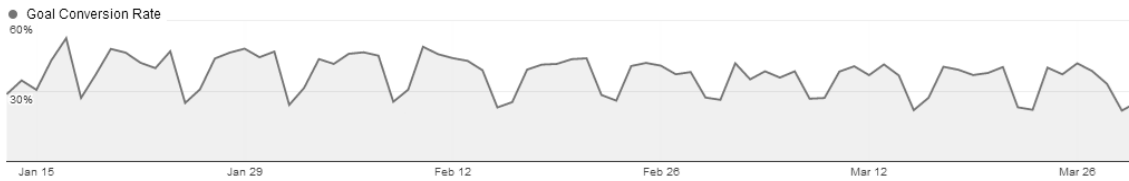


FIGURE 24 - GOAL CONVERSION RATE FOR ALL GOALS

4.5.2.2 Funnel Analysis

The *Reverse Path* report allows us to track the pages seen by users prior to the conversion occurred. This might be especially important in relation to destination goals, in order to understand the main sections of the website that are leading to conversion. The feature contains a look back window of up to 3 steps, for visualizing the most common paths prior. However, the amount of possible pages to have been seen is in some cases so high, that the number of combinations results in unactionable information. For this example, engagement goals each have over 15000 combinations of different steps that lead to a conversion for the time period, with no particular insights possibly extracted from here.

In other cases, we have access only to obvious relations, of which we have the example of Goal 5. This corresponds to the necessary flow of an order, in which at least 92% of the conversions refer only to the pages required for the order form. This has to do with configuration issue, as well as the possible look back window. In this case, it would probably be more useful to resort to other tools, such as event tracking. There are in this way some configuration and interpretation issues, which might result in unintelligible or irrelevant data.

Goal Completion Location ?	Goal Previous Step - 1 ?	Goal Previous Step - 2 ?	Goal Previous Step - 3 ?	Engaged Users - per visit (Goal & Completions) ↓
1. /WebShop/OrderForm.aspx	/WebShop/OrderForm.aspx	/WebShop/OrderForm.aspx	/DetailedCart.aspx	132 (0.78%)
2. /WebShop/OrderForm.aspx	/funnel_G1/OrderForm_Billing.html	/WebShop/OrderForm.aspx	/DetailedCart.aspx	110 (0.65%)
3. /WebShop/OrderForm.aspx	/funnel_G1/OrderForm_Login.html	/WebShop/OrderForm.aspx	/DetailedCart.aspx	67 (0.40%)

TABLE 15 – REVERSE PATH FOR ORDER PLACEMENTS

The Funnel Visualization report is in this sense a much better way of understanding the conversion funnel, as well as the steps at which traffic might be

diverging to other pages. In this example, Goal 5 corresponds to the placement of an order by users where the conversion funnel is defined by a series of destination pages, required in the payment process for inserting the shipment information. Through funnel analysis, we can see the amount of people who initiates the process, in which stage of the funnel they do so, from which pages these visits are originated and, if they leave the process flow, where are they leaving to.

In this case, we can see that over half the visitors who initiate the process from the Detailed Cart Page ends up making the purchase. On the other hand, users who abandon the funnel do so almost only on the first and second steps of the process. These are the steps containing information about the order and the user (Cart page and Order login), which is the entry door for the purchase to happen. The image below also shows that very few of these diversions immediately exit the site, continuing to browse through other pages. This is thus a positive factor since visitors don't completely cut contact, but continue looking for other contents. Knowing the shape of the conversion funnel is an important resource, since we can here identify and interpret major resistance points.

In the example, the first step (cart page), can be seen by any user regardless of their membership. The information of products is preserved by using session cookies, which makes this an available feature even for non-members. On the other hand, logging in requires the user to sign up, giving us personal and company information. Moreover, as this is a B2B website, only professionals can make a purchase on this site. Because of that, these are the main steps for the diversion of users. After that, drop offs become much more uncommon with almost all users (99.1%) who go through to the third step converting into a sale.

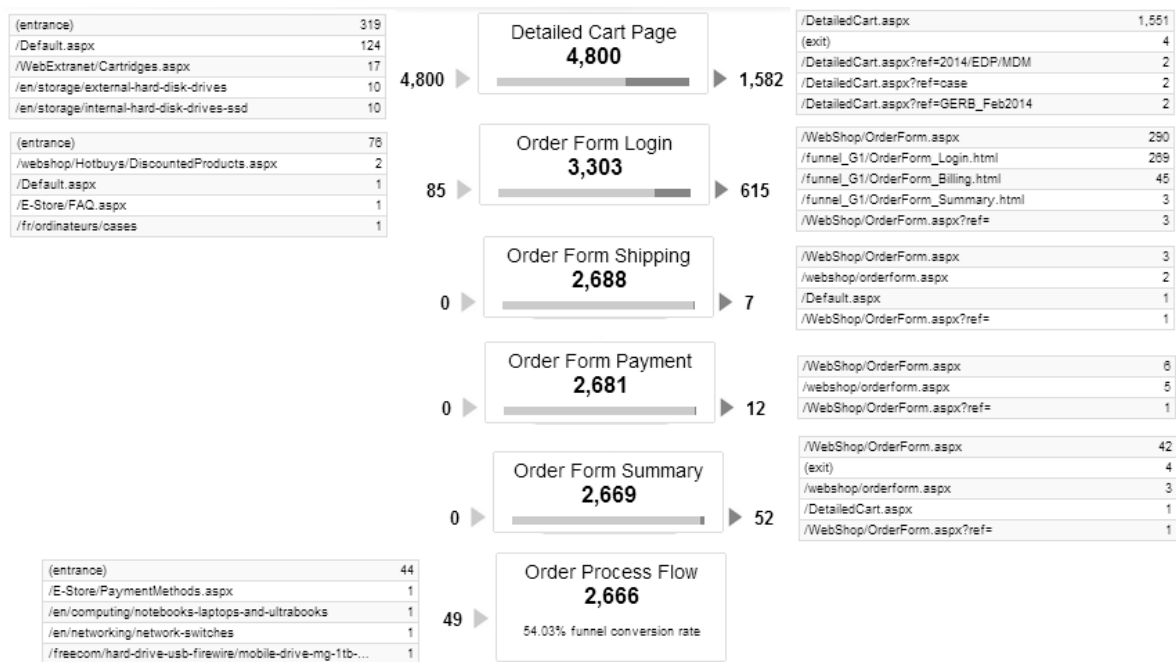


FIGURE 25 - CONVERSION FUNNEL FOR THE ORDER PROCESS FLOW

4.5.2.3 Attribution Models

The attribution of conversions to each channel is also an additional concern we would like to address, since the traditional attribution model uses only last click interaction in order to assign campaigns with a determined value. This is as we have seen an over-simplification of reality, especially in the case of ecommerce conversions. Users often use various sources during the decision process, each contributing to the process of decision. In this sense, multiple campaigns may be seen consulted during different periods in time, so attributing one with the whole value of a transaction is an inaccurate assessment of the contribution of each channel. Because of that, GA enables us to explore the influence of various channels in visits prior to conversion, through the Multi-Channel Funnel (MCF) analysis. This is a set of reports constituted by the Assisted conversions report, the Top conversion paths and the Model comparison tool as the main features to explore the effect of multiple interactions.

In this context, Kaushik (2013a) explores the differences between each attribution model, as well as the most appropriate tools to use when weighing the influence of different channels along the customer lifetime cycle. We thus begin by exploring the Assisted conversions tool, where we start by selecting the conversion

goal and the look back number of days before the conversion happened. The maximum number here is 90 days (roughly three months), so if we had a campaign which ended 91 days before the conversion, conversions credit would not be assigned to it on this report. However, that information might still be displayed in other sections, such as the Acquisition reports, since this information depends on the `_utmz` cookie which preserves acquisition information for a default period of 6 months.

Furthermore, the assisted conversion analysis also makes the distinction between last click and assisted conversions, concerning the times a channel was used as part of the funnel or as a last click interaction. In this way, the following table gives us, on the right-hand column, the relative importance of each channel in relation to its position on the conversion funnel. Higher values therefore stand for a greater importance of assisted over direct conversions, while infinite stands for the inexistence of last click conversions. On the other hand, values closer to 0 indicate that these are channels often used as a last influencer. As depicted below, we can see that the most valuable source is direct, with an expectedly high ratio of last click conversions. All other channels have greater importance of assisted over last click interactions, meaning users are prone to consider them as part of the decision process, but don't usually consider them decisive for the purchase. Assisted to last click ratio does not however equal channel value, with (not set) being the second more valuable for assisted conversions, while organic is second for last click. (not set) values in this case stand for the web source, while social network had only one LinkedIn assisted ecommerce conversion for the period.

Conversion: 1 Conversion Type Selected ▾ Type: All AdWords Lookback Window: Set 90 days prior to conversion

MCF Channel Grouping ?	Assisted Conversions	Last Click or Direct Conversions	Assisted / Last Click or Direct Conversions ↑
1. Direct	2,148 (56.99%)	2,436 (84.15%)	0.88
2. Referral	68 (1.80%)	49 (1.69%)	1.39
3. Organic Search	582 (15.44%)	238 (8.22%)	2.45
4. (not set)	786 (20.85%)	160 (5.53%)	4.91
5. Email	184 (4.88%)	12 (0.41%)	15.33
6. Social Network	1 (0.03%)	0 (0.00%)	∞

TABLE 16 - ASSISTED CONVERSIONS REPORT FOR ECOMMERCE TRANSACTIONS

However, this last report only gives us information about the overall performance of channels and not their evolution. Because of that, we have no indication about its position on the conversion funnel, but only if it made or not part of it as a last or assistant interaction. Information on channels position in the funnel might however be important in order to explore the channels which introduce the brand, the ones stimulating an ongoing relationship or which are decisive for the buying decision. Kaushik (2013a) in this context refers to channel attribution models, which provide different levels of information in this ambit. The evaluation of exact *top conversion paths* is from his perspective a relatively vague exercise, due once again to the incredibly high number of different combinations it might take for users to reach a conversion. There are simply too many possible combinations of channels to consider, so trying to control the exact path followed by users is a fruitless action. Each conversion is relative to each context, with different points of interaction in time.

In relation to our website, we again face the problem of proper identification of devices as well as missing identification of returning costumers. Therefore, there is a high amount of direct conversions generating the great majority of conversions. Intercalated with these, we occasionally have a few conversion paths using more than one or two sources, as depicted below. The total number of conversion paths is in this case 3.514, which reflects an inoperable number in practical terms for their individual analysis. Below, we have an example of two generic conversion paths (using google,

web and direct sources), compared to a more specific one, which naturally generated fewer conversions.

		Conversions	Conversion Value
google	web (direct)	90 (0.33%)	(0.50%)
google	(direct) × 12		
google	(direct) × 3		
14-JAN-18 - Flash - Bundle / HP Pro 3500	(direct) × 2	5 (0.02%)	(0.49%)
web			
(direct) × 2	web	208 (0.76%)	(0.43%)

FIGURE 26 - CONVERSIONS AND % VALUE FOR TOP CONVERSION PATHS

Probably the more complete channel attribution tool is in this sense the *Model comparison* tool, which allows us to simultaneously compare channels using up to three models of value attribution. This report uses the same look back window as the assisted conversions report, with the difference of weighing each channel's importance according to its position in the conversion funnel. The selection of attribution models allow us to compare different perspectives defined by the each model's rules, determining how credit is assigned to each channel in each transaction. With this we evaluate channel performance pondering the number of conversions initiated, assisted or concluded for each source, medium or channel group.

Google Inc. (2014) and Sharma (2012a) in this context classify attribution models into two different categories including baseline (default) or custom attribution models. Among baseline attribution models we find the *last click* interaction and the *first click* interaction models, assigning 100% credit to the last and first interactions, the *last non-direct click*, ignoring direct clicks and attributing 100% credit to the last non-direct channel, the *last AdWords click* for AdWords campaigns, the *linear attribution* model, assigning equal credit to all interactions in a conversion path, the *position based model*, an hybrid between last click, first click and the linear models which splits the credit in a 40-20-40 ratio for first, in-between and last interactions, and the *time decay* model, attributing more importance to interactions closer to the moment of conversion. This last one, works on the basis of an exponential decay of the value of a conversion, with a half-life decay of 7 days. This means that with each week passed, the credit assigned to each channel will be cut in half. So an interaction

happened 14 days ago will be weighed at about a quarter the importance of the last interaction.

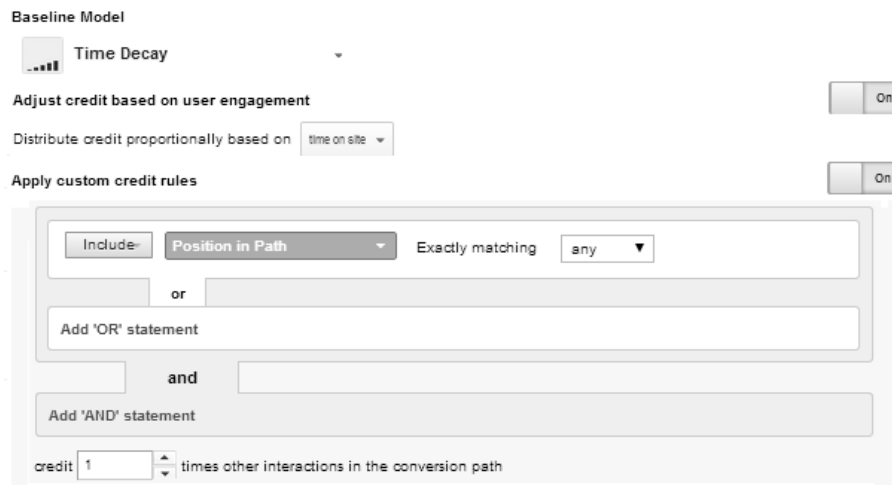


FIGURE 27 – CONFIGURATION OF CUSTOM ATTRIBUTION MODEL

On the other hand, we can also choose to customize our own attribution models, based on the default baseline given by GA. Supra, we have depicted an example of time decay applied to a custom model of attribution. In this example, we adjust the attribution of credit by user engagement based on the time on site metric, applying the credit rules to any of the channels in the conversion path. We can also attribute different weigh to different channels, if for example we consider certain campaigns are more influential than others. In order to do so, we therefore include custom credit rules in which we define the set of criteria to match the desired weighing. Chau (2013) for example refers to the fact that direct interactions should not actually be considered marketing channels, since they reflect actions from the user and not really a response to any kind of content or campaign.

For our case, we set a look back window of 90 days prior to conversion so we can use the model comparison tool to explore the different position of channels and the importance of each in the conversion funnels. In this sense, we will be using the medium dimension to compare the evolution of marketing channels and their relative contribution at each point in time. The benchmark model for this will be the *time decay* model, for which we will be considering the first six higher revenue-generating

mediums. We then selected the *linear*, *position based*, our *custom 1* as well as the *first* and *last* interaction models in order to compare the value of each to the benchmark model. The interpretation of this allows us to make assumptions on the importance of mediums and the influence they have on the user’s lifecycle. Below, we have the percent value for both the number and the value of conversions according to the benchmark model, as well as the variation according to our model comparison tool:

Time Decay		% change in Conversion Value (from Time Decay)					
Conversions	Conversion Value	Linear	Position Based	Custom 1	First Interaction	Last Interaction	Last Non-Direct Click
(none) (68.92%)	(81.35%)	-0.91%	0.16%	-1.37%	-0.62%	1.82%	-41.72%
organic (21.09%)	(10.96%)	4.12%	9.60%	-2.40%	49.47%	-24.13%	77.99%
bort (3.63%)	(3.73%)	3.66%	-15.02%	28.04%	-95.77%	47.25%	445.05%
referral (4.05%)	(1.62%)	-0.10%	-13.04%	1.74%	-7.35%	-26.34%	-3.38%
email (0.85%)	(1.15%)	3.04%	-7.69%	12.37%	-3.29%	-28.08%	315.85%
banner (1.16%)	(1.06%)	8.90%	-24.67%	12.41%	-100.00%	17.94%	431.59%

TABLE 17 - MODEL COMPARISON TOOL BY MEDIUMS

According to the different perspectives, we can see that many variations can occur, with different approaches influencing our response. In this sense, we can see that for all cases the direct medium (none) is the most important channel. However, when compared to the linear model, only the direct and referral channels lose importance. All other channels benefit from being attributed the same weighing regardless of their position, meaning they usually play an assistant role, rather than being last interactions in the decision cycle. Furthermore, our custom model also reflects the channels with higher engagement, using the same model of our benchmark, but making use of time on site to attribute higher credit to more engaging channels. Most mediums in this case benefit from the feature, with the exception of direct and organic, which as we saw (especially in the case of the latter), have the more heterogeneous public. Because of that, average values will result in a penalization of overall channel performance, since we are only taking in consideration aggregate behaviors. Because of that, e-mail and internal campaigns, such as bort and banner, are non-surprisingly the mediums most benefiting from user engagement. In

this case, e-mails are associated with returning customers, as well as internal campaigns, which are not the causes for but the result of user engagement. That said, these mediums also rank high on last click interactions, while every other fail to impress in this area.

In addition, the only medium that actually improves when considering first click interactions is the organic channel, which reflects the inability of all others to generate new acquisitions and leads. What it tells us is that even if any other campaign is generating new visits, the only channel generating conversions are organic searches. Once again, it seems that the only channel possible of generating significant new prospects is engine-searches. This is also the most valuable non-direct (and non-internal) medium, with 19.5% of all non-direct last click value conversions, followed only by e-mail at 4.8% and referral traffic at 1.6% of total value. Lastly, the position based model also reflects the relative higher importance of direct and organic traffic, in the first case due to last interaction importance, in the second due to first click.

4.5.2.4 Transactions

Lastly, we explore the ecommerce section, which gives us information about the products that were bought, quantity, revenue, shipping costs and tax information, as well as the performance of sales and the distribution of days from first visit to purchase. In order to track ecommerce transactions, three methods are however required our software, using the source code of our web pages. In this sense, we first have to create a transaction object, by using the `_addTrans()` method in our webpage. Secondly, we need to be able to track the items associated with a transaction by calling the `_addItem()` method, specifying each product's price, category and quantity. Lastly we need to submit this information to GA by using the `_trackTrans()` method.

A major setback of this feature is however not having the products tied to a specific page for getting engagement data, which could be an interesting resource for exploring pre and post purchase behavior. If for example we want to see how much time was spent by users on a specific product page, this has to be done manually, by identifying the pages and make them correspond to each referenced product. This has however to be done using another interface or programming language, since the

default GA environment does not do this by itself. In this case, there are 3.223 products referenced for the site at the moment. Because of this, we can only use the default interface to gather descriptive facts about quantity or value of transactions and products, being often impossible to retrieve values associated with the user, such as page views per visit (even using the API), being only possible to recover values associated with the product.

Another relevant matter is the distribution of order value, versus the number of unique orders and average value. In this sense, some of the most profitable products can be associated with a number of unique orders, without however taking in account the distribution of amounts per order. GA again takes only absolute and average values, which is a limited approach, especially in a B2B environment. To illustrate this we can take for example the product with the identification 'M852R237' (Thinkpad Edge E530 computer), having sold 35 units in 6 unique orders. The average quantity per order is therefore 5.83, which makes it the 4th most generating revenue product for the time period. The fact however is that from the 6 transactions, 5 only sold one unit, while one transaction sold the remaining 30 units. This is therefore an interpretation issue, because even though the product had visibility enough to sell six times, its popularity clearly wasn't consistent over time.

Because of this, average values are often not a good indicator of performance due to the fact that one product which might be underperforming one day, the other it can be among the top rated articles, given the characteristics of B2B. These variations thus have to be inspected manually, by accompanying the evolution of each product.

In that sense, sales performance can additionally be tracked either by accompanying the revenue or the number of transactions generated per day, with reference to each unique transaction. In this way, to each order placement there is an associated revenue (as well as tax and shipping information) and the quantity of items purchased. Drilling down into each unique order's ID, we also access information about the order's items and the generated revenue. In the next example, we also used the segmentation feature in order to make the distinction between returning and new users, examining this period's biggest order (89080.2120140124) and the products selected. Here we access information about quantity and price, with reference to the

date in the last segment of the order number (20140124) and the interface timeline. Information that we cannot access however, is that of the clients who placed the order. GA's interface does not directly communicate such information and in order to do so we would have to integrate this with other applications.

Product ?	Product Revenue ?	Quantity ?
Returning Users	€23,097.00 % of Total: 1.74% (€1,330,786.32)	360
New Users	€0.00 % of Total: 0.00% (€1,330,786.32)	0
1. Monitor LED Backlit	€15,888.00 (68.79%)	60 (16.67%)
2. Notebook Case 15.4in Corporate	€3,072.60 (13.30%)	60 (16.67%)
3. combination Portable Laptop Lock	€1,292.40 (5.60%)	60 (16.67%)
4. Adapter Mini Displayport To DVI	€1,282.20 (5.55%)	60 (16.67%)
5. Keyboard Preferred Pro USB Be/uk	€1,204.80 (5.22%)	60 (16.67%)
6. Oem Rx250 Optical Mouse Black	€357.00 (1.55%)	60 (16.67%)

TABLE 18 - UNIQUE TRANSACTION REVENUE AND QUANTITY PER ITEM

4.5.3 Summary

The conversions section contains as we have seen some of the most important reports to help us trace back the effectiveness of channels and assess the overall performance of goals. Some of the insights from this section were in this way the following:

The monetization of conversion goals may in some cases be a tool for calculating the approximated ROI of campaigns, with the possibility of integrating online and offline marketing. However, attention should be paid to the distribution of order value and to other metrics, such as time and visits to conversion and the number of channels used.

Most of the transactions have little contribution to the bulk of total revenue. Because of that, the distribution of order value is highly skewed, with extreme outlying values contributing decisively for business performance. In this way, roughly 75% of transactions contribute to only 25% of revenue. The remaining quarter is thus extremely important for us, in just a fraction of the users who visited the site.

Managing the customer lifecycle is in this way much more appropriate than targeting users at session level.

Goal conversions happen primarily during weekdays, not only at an absolute level but also at a percentual level. This is true not only for transactional conversions but also for engagement goals (visit duration and page views goal 7 and 8) and the subscription of the newsletter (in spite of the low number).

Defining a funnel for goal conversions might help us identify the main points of diversion. In our case, for the order process flow we were able to identify the main source of diversion as being (1) the Detailed Cart page at about 32.3% drop offs and the (2) Login Order Form at 18.6%. After that these numbers are greatly reduced with over 99% of users finishing the purchasing process.

The Multi-Channel Funnel Analysis consists of several reports, such as the Assisted Conversions and the Model Comparison reports in which we explore the users ABC (Acquisition-Behavior-Conversion) cycle according to each channel's position on the conversion funnel. Having a 90 days look back window and the time decay as our benchmark model, we were able to define that:

The direct channel is for all cases the most important channel;

The direct and referral channels lose importance when compared to the linear model, which indicates that these are the channels closer to the conversion while every other has more of an assisting role;

Our custom model, pondering time on site as a metric for favoring channels generating engagement, only detracted direct and organic. These are however the channels with the largest amount of visits and more heterogeneous public;

The only channel benefiting from First Click interaction is Organic, which again supports our belief that this is the only channel introducing our site to new prospects;

The Position Based model favors mostly the organic and direct channels, seemingly the first and last channels to be consulted before conversions;

4.5.6 Period Comparison

- Transactions have a highly skewed distribution in terms of turnover, with 75% of revenue coming from 25% of transactions, and vice-versa.

As we saw throughout this work outliers and extreme values have a very important contribution for this website and business in general. One of the most evident variables in relation to that is the revenue per transaction, in which we saw a great deal of the company's business is constituted either by very large orders when compared against the value for the first three quartiles of the distribution. Following, there is a summary and the histogram for the distribution of values for the first and second periods, in which we can see that this is a consistent tendency over time.

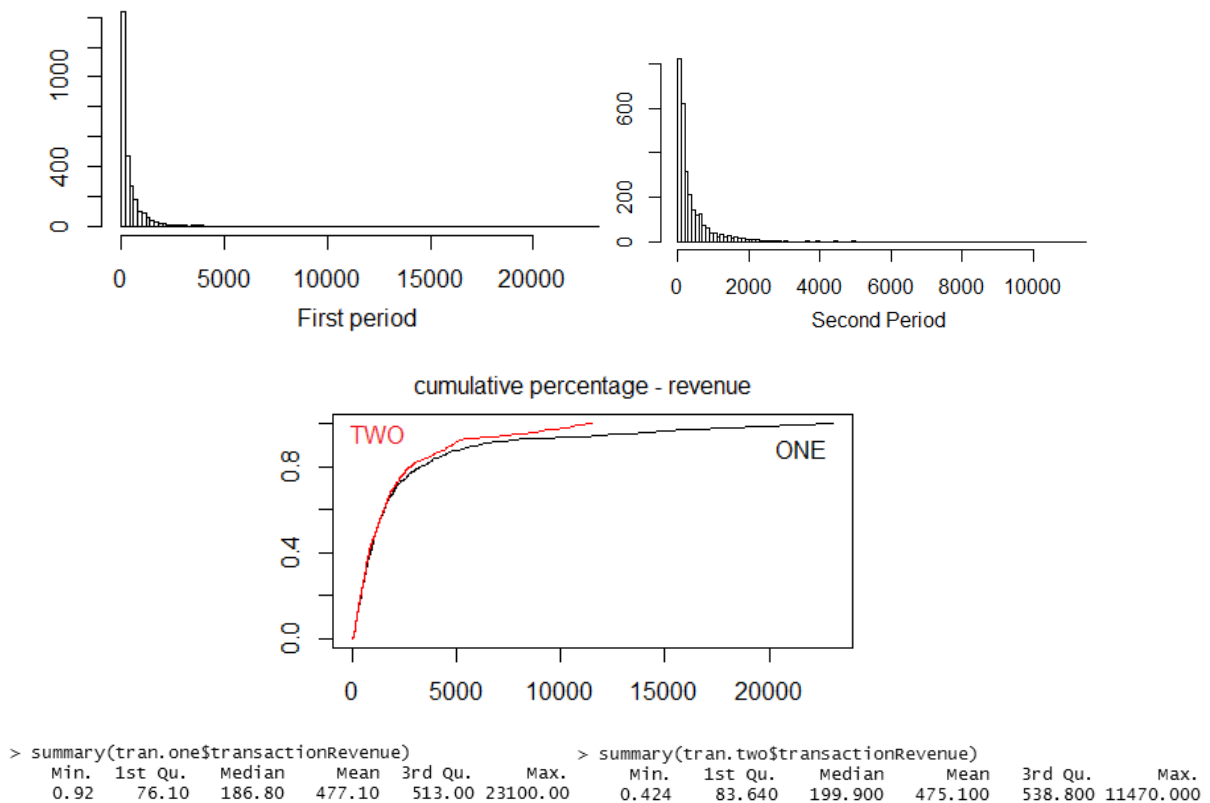
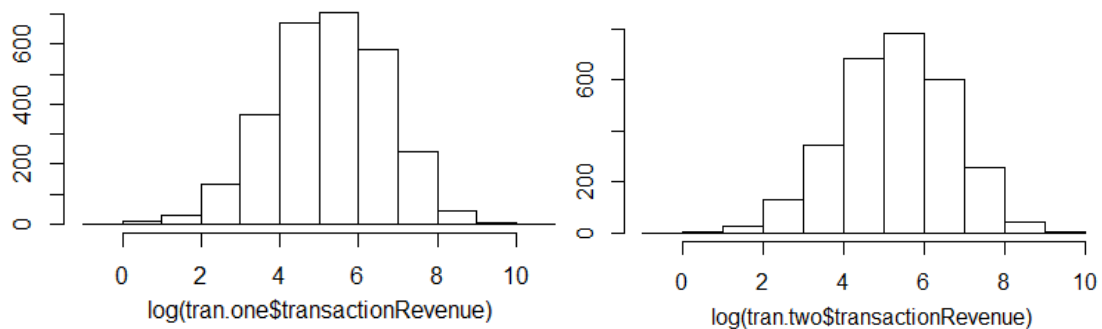


FIGURE 28 - DISTRIBUTION OF TRANSACTION VALUE FOR THE 1ST AND 2ND PERIODS (ONE AND TWO)

For the second period however, we see that the higher numbers are about half of those during the first period, while interquartile range however remains to be roughly the same. In this case, we have access to the value information of each

transaction and consequently the values for the mean and standard deviation in the distributions. Because of that, we can compare the two by running a t-test, which can tell us if there is a significant difference between the average values in the observations. Due to the fact that the original values are not normally distributed however, we chose to transform the variables, taking the logarithm of each value, as such:



Two sample t-test

```
data: log(tran.one$transactionRevenue) and log(tran.two$transactionRevenue)
t = -1.7367, df = 5646.877, p-value = 0.08249
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.13796302  0.00834561
sample estimates:
mean of x mean of y
5.227094  5.291903
```

FIGURE 29 - T-TEST FOR LOG TRANSFORMED TRANSACTION VALUE FOR THE TWO PERIODS

As we can see, at 95% confidence the t-test does not reject the null hypothesis of the logarithm for the transactions' revenue having the same average value, which means that there is no evidence of difference in the distribution of transactions value for the two periods. Consequently, we argue that there is no evidence of denial to our "75/25" assumption, meaning that the great majority of transactions accounts for little revenue, with great importance of sporadic extreme values to this business.

- The main diversion point of the conversion funnel for transactions is the Shopping Cart page (32.3%), followed by login (18.6%). After that, nearly all users conclude the transaction.

In relation to the conversion funnel, we can see again that the main step of diversion for users who initiate the order process flow is the first stage, which is the detailed cart page. In this way, only 67.04% of users went through this stage during the first period, while 67.87% did so in the second. Running a t-test on the difference in these proportions we can see that this is not a statistically significant difference at 95% confidence ($Z=0.88$). However, in relation to the second stage, a more significant percentage of people diverged from the Login page during the second period, from 81.38% through traffic to only 79.46% during the second period. This is a statistically significant decrease of almost 2% ($Z=1.99$), which might also be a consequence of the increase in the percentage of new users. That is because only enterprises can buy from this website, while sales to regular consumers are not permitted. From this point, 99.2% of users in the first period went through the remaining steps, including shipping, payment and summary information, while 100% of users in the second period did so. This is also a small but significant difference ($Z=4.71$), which reflects the very high probability of users not diverging once they logged in.

5 Statistical Procedures

Up until this point we have been discussing the utilization of Google Analytics as a reporting tool using mainly its interface, through the utilization of the browser or the GA application. While this is as we have seen a very complete, comprehensive tool giving us access to a great deal of indicators, dimensions and reports, it provides us only with descriptive aggregate values for the behavior of visitors. In this way, one of our major limitations in terms of exploring the data is we do not get access to data for singled-out behaviors, but only aggregate values for dimensions such as time (*e.g.* date), webpages or marketing channels. In this sense, only the Premium version gives us access to un-sampled data to retrieve using Google Big Query.

While the architecture of data works very well for the GA environment, including the segmentation feature, when transported into other environments, we thus have to be careful with the dimensions we choose to combine in order to guarantee the intelligibility of the data.

Throughout this work, we used some of the functionalities in R, particularly to illustrate the distribution of values or to explore the correlation between variables. However, it would also be relevant to explore the extent to which we can apply other statistical procedures to GA data. Having in mind the particular data architecture, we however know at start that much of the information we will be using is aggregate according to the different dimension, lacking information about the distribution of values and individual cases for our users. Because of that, our analysis was up until now based on rates which reveal the tendencies of each dimension (segment), such as goal conversion rates or the percentage of new sessions. In statistical terms, we therefore based the analysis in proportions testing, of observed occurrences over the total number of observations. In the following section, we are also going to be using from the 13th of January to the 28th of June (24 weeks), in order to explore some techniques we can use to explore the metrics relation in accordance to the possible segmentation according to the available dimensions.

5.1 Modeling with R

So far we have been exploring the available metrics especially using rates of conversion and proportions, because of the fact that metrics are structured to return especially aggregate values for different segments. This limits our access to the user dimension, leaving us only with more general approaches to our units of analysis. In this way, we already talked about most of the dimensions that help us segment our audience, as well as the metrics that with those can be combined. Because this is a structured environment designed for a specific application, many combinations do not work, so selecting appropriate indicators is important for the adequateness of analysis.

Throughout this work and having in mind the perspectives we could adopt, some of the most interesting dimensions available are session-related, reflecting visitors' engagement at the level of each visit, as well as the dimensions having to do with time and our marketing channels. Some perspectives (Correia, 2010b; Kosny, 2014; Simpson, 2014) also looked to work around the default dimensions, using customization to create a unique user ID dimension, either using PII and non-PII. This however would require the customization or an update of the GATC, as well as an additional period of data collection.

5.1.1 Session Dimensions

In this example, using RGoogleAnalytics we extract Visit Length and Page Depth dimensions from GA, combining the two for characterizing individual sessions with engagement values for both duration and number of pages seen. To these, we also add the date dimension for having a time reference, resulting in a total of over 58 thousand combinations of observations.

	duration	depth	revenue
duration	1.00000000	0.5348295	0.09092837
depth	0.53482946	1.00000000	0.23151378
revenue	0.09092837	0.2315138	1.00000000

TABLE 19 – CORRELATION TABLE BETWEEN VALUE AND ENGAGEMENT VARIABLES

As we can see from the previous table, using the Pearson's correlation coefficient, duration and depth exhibit a relatively strong correlation to each other,

while relating poorly with our transactional indicator. What this means is that there is a weak linear correlation between engagement metrics of a session and its value, suggesting that sessions associated with higher engagement levels do not necessarily relate to sessions with higher values. Even so, considering these might be insufficient variables to introduce in a model, we include additional variables, concerning the precedence of users, the use of the internal search feature, the day of the week, as well as the number of previous visits. In this way, we are not limited to an approach based on merely engagement indicators, extending our analysis into the utilization of different website sections, external referrers, as well as time indicators.

In order to explore these relations we will use a linear regression as proposed by Polancic (2007) for the realization of a test on the effect of these variables on session revenue. However, resorting to this type of methodology imposes the need for verifying the appropriateness of data and the type of distributions we find. In this particular case, by plotting the data and summarizing the descriptive statistics for each we acknowledge that these are highly skewed distributions. In order to satisfy all the assumptions of OLS, because these resemble Poisson distributions, the estimators will therefore have to be transformed. There are in the beginning no identifiable linear relationships between Y and X (Root, 2010), particularly in relation to the engagement and visitors' number of previous sessions.

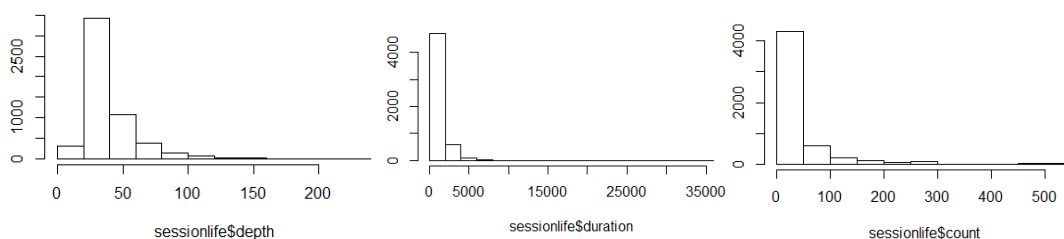


FIGURE 30 – DISTRIBUTION OF ENGAGEMENT AND VALUE VARIABLES

As we can see by the previous histograms, there are again problems with the distribution of variable values, with Provost & Fawcett (2013) arguing that these are common issues in complex scenarios, with similar occurrence frequent with online data. Assuming normality for this kind of distributions is therefore often not correct,

and we should focus first on data appropriateness and its necessary transformations, in order for it to be operable and transmit the right information. On the same page, Root (2010) also points out the inadequacy of OLS in relation to skewed distributions. This type of data always has a high proportion of number zero outcomes, with nonlinear relations between the explanatory and the response variables, exhibiting heteroskedastic errors. For us to deal with this problem, the authors then suggest that we transform Poisson-distributed variables, taking its logarithmic form for the representation of the same reality. This should result in a Gaussian distribution, maintaining the data values we are interested in, only with a different interpretation of results. Satisfying the assumptions of OLS estimators, as well as correcting the inconsistencies in data is in this way one of the major challenges with this type of datasets, which we will then be looking to correct.

For exploring the relations in the regression model in, we therefore chose to do a manual logarithmic transformation of all the variables, resorting to the $\log()$ function in R, transforming each observation to its natural logarithm. New variables were thus created, with an approximately normal distribution and sessions with zero duration value also excluded. Additionally, the categorical variables were introduced as dummy variables, in order to identify visitors' channels, weekends or sessions using the internal search. Using these, a linear model was introduced in order to explore the relations of the explanatory variables with session value:

```
Call:
lm(formula = revenue ~ log(duration) + log(depth) + log(count) +
    search + direct + organic + internal + email + weekend, data = sessionlife)

Residuals:
    Min       1Q   Median       3Q      Max
-816.9  -392.4  -250.8    55.7  22581.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -525.679    123.231  -4.266  2.03e-05 ***
log(duration)   34.400     15.917   2.161  0.03073 *
log(depth)    207.627     41.107   5.051  4.54e-07 ***
log(count)      7.826      9.540   0.820  0.41208
search        -78.462     33.670  -2.330  0.01982 *
direct         99.425     68.999   1.441  0.14965
organic       106.456     69.104   1.541  0.12349
internal      137.072     68.224   2.009  0.04457 *
email         211.450     80.877   2.614  0.00896 **
weekend        13.835     76.352   0.181  0.85622
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 970 on 5418 degrees of freedom
Multiple R-squared:  0.01722, Adjusted R-squared:  0.01559
F-statistic: 10.55 on 9 and 5418 DF, p-value: 2.355e-16
```

MODEL 1 - COEFFICIENTS FOR LINEAR REGRESSION ON SESSION VALUE

This is as expected a poor performing model, given the fact that only under 2% of the variation in the response variable can be explained by the variables contained in the model (R-squared). Moreover, only the constant and five of the nine variables exhibit statistical significance for explaining the variations in session revenue. Still, as affirmed by Frost (2013), low R-squared values are often common with variables reflecting human behavior, and in most cases we can still draw insights from the significance of variables. In this way, the constant variable, logdepth and email (dummy), are statistically significant variables at 99% confidence, while logduration search and internal search (dummies) are at 95% confidence.

However, we are again having issues with our values' distribution, with non-normal (skewed) errors, as indicated by the diagnostics plot, as well as the residuals distribution (Faraway, 2005 *cit in* CrossValidated, 2014). One reason for this is as we have seen the highly skewed distribution of session revenue. In this case, we are challenged with the facts that most sessions result in no conversions and the huge difference in their value. Because of that, even if we take the square root of revenue as our response variable, the problem, while reduced, will still be an issue. With this regression however, logduration and logdepth, search and internal become statistically significant at 99% confidence, while internal campaigns remain so at 95%. The goodness of fit of the model also increased, with an R-squared of 3.5%.

```
Call:
lm(formula = sqrt(revenue) ~ log(duration) + log(depth) + log(count) +
    search + direct + organic + internal + email + weekend, data = sessionlife)

Residuals:
    Min       1Q   Median       3Q      Max
-20.342  -8.763  -3.313   5.372  133.750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.53069    1.65088  -0.321  0.747877
log(duration)  0.82244    0.21324   3.857  0.000116 ***
log(depth)    3.68546    0.55070   6.692  2.42e-11 ***
log(count)    0.01467    0.12780   0.115  0.908591
search       -1.77738    0.45106  -3.940  8.24e-05 ***
direct        1.31551    0.92435   1.423  0.154742
organic       1.40417    0.92576   1.517  0.129384
internal      2.03212    0.91398   2.223  0.026231 *
email        3.41147    1.08348   3.149  0.001649 **
weekend      1.17407    1.02285   1.148  0.251087
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.99 on 5418 degrees of freedom
Multiple R-squared:  0.03456, Adjusted R-squared:  0.03296
F-statistic: 21.55 on 9 and 5418 DF, p-value: < 2.2e-16
```

MODEL 2 - LINEAR MODEL FOR SESSION VALUE WITH TRANSFORMED RESPONSE VARIABLE

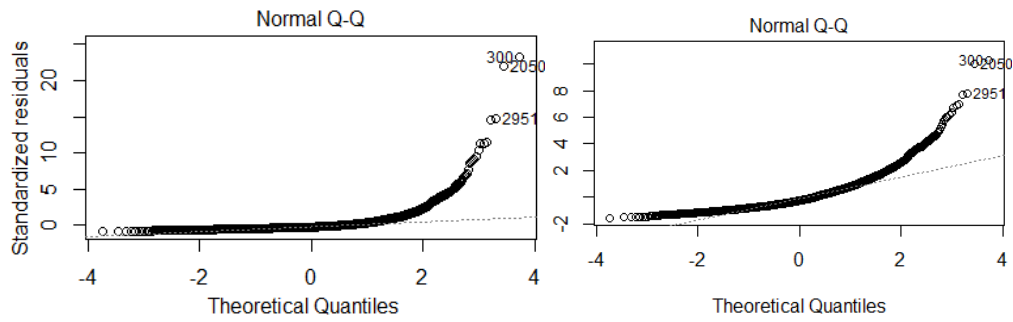


FIGURE 31 - Q-Q PLOT FOR THE RESIDUALS OF MODEL 1 AND 2

Because of the high number of zero values in session revenue however, solving the problems for the assumptions for our model by transforming the variables has in this way proven to be an unproductive effort. Because of that, instead of assumptions on session revenue, Araripe, Gondaliya, & Shah (2013) resort to logistic regression trying to predict the probability of a specific outcome for one user. Nevertheless, in order to explore the effect of our variables in buying user, we disregard zero-value sessions, using the same indicators to see the effect of our explanatory variables on session revenue of *buying* users. In order to do this we used the same data base, excluding rows associated with no revenue (n=5427).

As we can see in the following histograms, with the variables transformation we can assume that they approximately follow a normal distribution, with the exception of session count. This is explained by the great number of single-session users, which we already mentioned to be one of the biggest problems of the online world and the available data collection methodologies. Therefore, this is not going to be a statistically significant variable for our model and we can thus exclude it from our analysis. The assumption of normality is also corroborated by our regression's diagnostic plots, with our Q-Q plot displaying an approximately straight line.

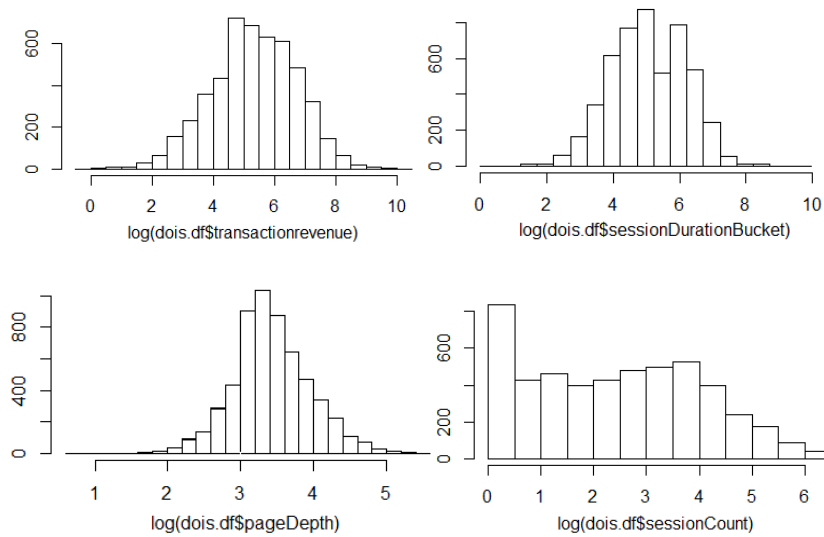


FIGURE 32 – LOGARITHMIC VARIABLES DISTRIBUTION

```
> cor(sessionl[1:4])
      duration    depth    count    revenue
duration 1.0000000 0.53993458 0.107508284 0.081500963
depth    0.53993458 1.00000000 0.055905588 0.121320972
count    0.10750828 0.05590559 1.000000000 0.001594892
revenue  0.08150096 0.12132097 0.001594892 1.000000000
```

TABLE 20 - CORRELATION OF VARIABLES FOR USERS' BUYING SESSIONS

```
Call:
lm(formula = log(revenue) ~ log(duration) + log(depth) + log(count) +
    search + direct + organic + internal + email + weekend, data = sessionl)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-5.5995 -0.8909  0.0288  0.9646  4.7042
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.34379    0.17377   19.243 < 2e-16 ***
log(duration)  0.10431    0.02244    4.647 3.44e-06 ***
log(depth)    0.39693    0.05797    6.848 8.34e-12 ***
log(count)   -0.01453    0.01345   -1.080 0.28012
search       -0.22311    0.04748   -4.699 2.68e-06 ***
direct        0.05450    0.09729    0.560 0.57537
organic       0.09910    0.09745    1.017 0.30923
internal      0.16185    0.09620    1.682 0.09255 .
email         0.32782    0.11404    2.875 0.00406 **
weekend       0.18107    0.10766    1.682 0.09266 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.368 on 5417 degrees of freedom
Multiple R-squared:  0.04116, Adjusted R-squared:  0.03957
F-statistic: 25.84 on 9 and 5417 DF, p-value: < 2.2e-16
```

MODEL 3 - LINEAR REGRESSION AND DIAGNOSTIC PLOTS FOR USERS' BUYING SESSIONS

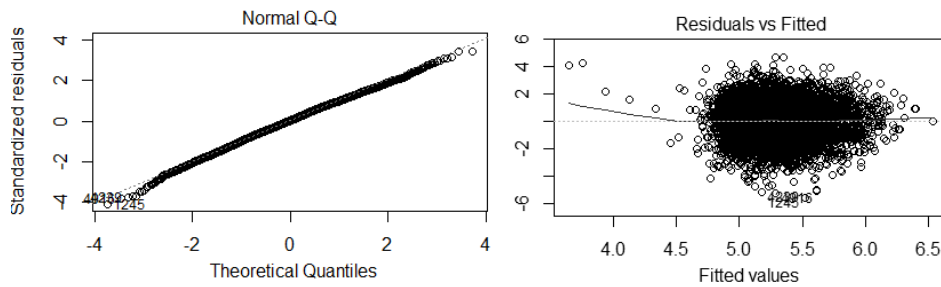


FIGURE 33 - MODEL 4 DIAGNOSTIC PLOTS

One important factor with this is that by transforming the variables we guaranteed the approximate normal distribution of the variables, maintaining however so heteroskedasticity of the error term. However, there are still a few outliers detected in our charts. Again, the logarithm of duration, depth, and the search and email dummies are significant at 99% confidence, while internal campaigns and weekend days are so at 90%. Another factor worth noticing is still the negative effect of search usage on revenue, which our previous research would suggest otherwise. A 1% increase in duration and depth will in this sense result in respectively 0.1% and 0.4% increase in revenue value, *ceteris paribus*, while the binary variable with the strongest effect is email, which results in a 0.33% increase in the dependent variable, while internal campaigns result in a 0.16%. Contrary to what we would expect, transactions during weekends also seem to have slightly higher value (0.18%) and the search feature surprisingly exhibits a negative effect (-0.22%).

However, previous research focused primarily on the *occurrence* of transactions, rather than value. In this case we are moreover looking at each individual sessions, and not aggregate values, in different units of analysis. The usefulness of this is nonetheless again merely descriptive, intending to explore the extent to which session variables can contribute to explain session value, in this particular case. Again the model had only and R-squared of 4.1%, which represents the percentage variation in the response variable that the explanatory variables can capture. It therefore seems a poor performing model, with many off-line variables seeming to be missing.

As we acknowledged, little variations on session revenue can be captured used this model and the dimensions associated to individual sessions, which again reminds us of the importance of offline interaction and multiple layers of decision in B2B

organizational purchases. For that matter, it is thus pertinent to find ways of exploring a broader relation between indicators and the time variation for each user, rather than focusing on visits. In terms of user value, it thus seems to make much more sense to comprehend the entirety of the lifetime cycle than to restrict him to a particular period in time. In this way we will be looking in the next sections to explore the role of other dimensions, reflecting aggregate metrics for a certain period of time.

5.1.2 Channel Dimensions

In this example we are going to use the medium dimension associated with our marketing channels in order to segment our traffic, aggregate engagement values as well as temporal references such as date and day of the week in order to explore and try to anticipate the variables effect on channel value. Therefore, we first select the correspondent dimensions and metrics to our model, introducing the dimensions to segment our variables, and the metrics for the desired values, as following:

```
dimensions = "ga:date, ga:medium, ga:dayofweek",  
metrics = "ga:transactionrevenue, ga:visits, ga:pageviews, ga:timeonsite",
```

Because the metrics' values are returned in their raw state, and due to data inconsistencies, we then have to transform most of our variables. In this way, we started by filtering meaningless mediums to the business, which revealed to have had no generated turnover. Following that, categorical variables were transformed into dummies, indicating the marketing medium and weekend days. Numeric variables also suffered a logarithmic transformation, in order to meet the assumptions of OLS estimators and normal distributions. Lastly, the data was randomly divided into the Train and Test subsets (80%-20% - n=778 and n=195), in order for us to employ the supervised learning method for predicting the value of the test dataset and evaluating model performance. After several tests and variable transformations, the selected model is thus as follows:

```

Call:
lm(formula = lrevenue ~ lvisits + (lpviews:ltimesite) + web +
    direct + organic + referral + email + weekend, data = train.df)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8181 -0.6210  0.0961  0.7067  2.7968

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.748744    0.324924   8.460 < 2e-16 ***
lvisits        -0.817621    0.090703  -9.014 < 2e-16 ***
web            0.882511    0.345325   2.556 0.010792 *
direct         0.970374    0.371092   2.615 0.009100 **
organic        0.768468    0.400746   1.918 0.055533 .
referral       0.721882    0.357036   2.022 0.043535 *
email          1.255493    0.341900   3.672 0.000257 ***
weekend       -0.297854    0.207625  -1.435 0.151814
lpviews:ltimesite 0.103128    0.006143  16.789 < 2e-16 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.106 on 769 degrees of freedom
Multiple R-squared:  0.4569, Adjusted R-squared:  0.4512
F-statistic: 80.86 on 8 and 769 DF,  p-value: < 2.2e-16

```

MODEL 4 – LINEAR MODEL FOR CHANNEL REVENUE USING THE TRAIN SUBSET

In this regression, most coefficients exhibit statistical significance at 99% confidence. However, “web” is only statistical significant at 95%, while “organic” is at 90% confidence. The “weekend” variable exhibits no statistical significance, in spite of the expected negative effect on transactions. In this way, *ceteris paribus*, the channel which generates higher revenue is email, with a 1.26% increase in the response variable. The direct channel follows, with an increase of 0.97% and internal web campaigns, with 0.88%. Organic and referral respectively reflect a 0.77% and 0.72% increase, *ceteris paribus*. Other minor mediums, such as social were not included in any separate category for their relevance.

The variable associated with traffic, lvisits, on the other hand, contrary to what we would maybe expect, generates a negative impact of 0.82% for each 1% increase in its value. This is probably because these are the channels in our dataset that have the most appearances in our dataset, used more often by our visitors, and accumulate the most page views and time on site, which as we’ll see have a positive effect on turnover. However, some channels with few visits but high engagement (such as email and web), are more effective in converting (high conversion rates), as opposed to high traffic channels (direct and organic), in which the number of poorly qualified visits dilute the ability to generate revenue. Conversely, we also introduced an interaction

variable between accumulated time on site and page views per medium, which are two highly correlated metrics, with this being a significant variable at 99%. In this case, a percent increase in (pages*time) will result in a 0.1% increase in channel turnover.

The following plots confirm the assumption of normality for our variables (q-q plot) only one outlier in the residuals plot, revealing however some heteroskedasticity. In alternate versions of this model (see appendix), this was not such of a problem, but revealed to be less accurate in the following test, so that these were the selected variables.

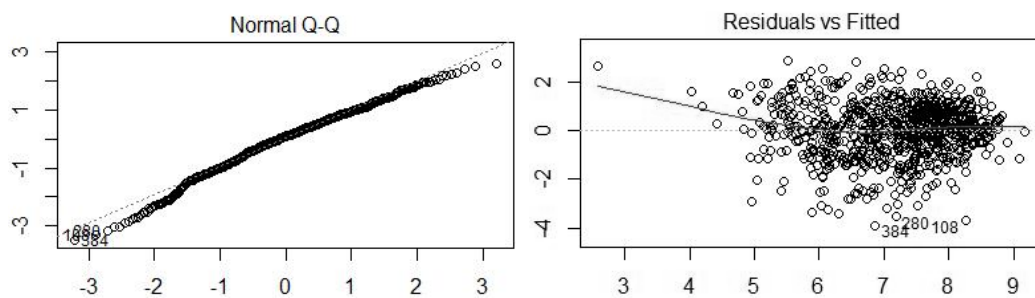


FIGURE 34 - DIAGNOSTIC PLOTS FOR MODEL 4

According to the coefficient of determination, about 46% of the variations in the response variable can be explained by the variables included in the model, which exhibits global significance at 99% confidence. However, we will also try to predict the value on the test subset, in order to understand to what extent could the model contribute to the prediction of channel value, based on the given metrics.

In that sense, we run the regression using the predict function, which based on the training set calculates the medium value per date, then comparing the differences between expected and real revenue, as well as the paired percent difference between the prediction and the actual value. In overall terms, the model **underspecified** the value of channels, attributing **69.2% of the actual value** to the bulk of transactions. The distribution of paired differences is as such:

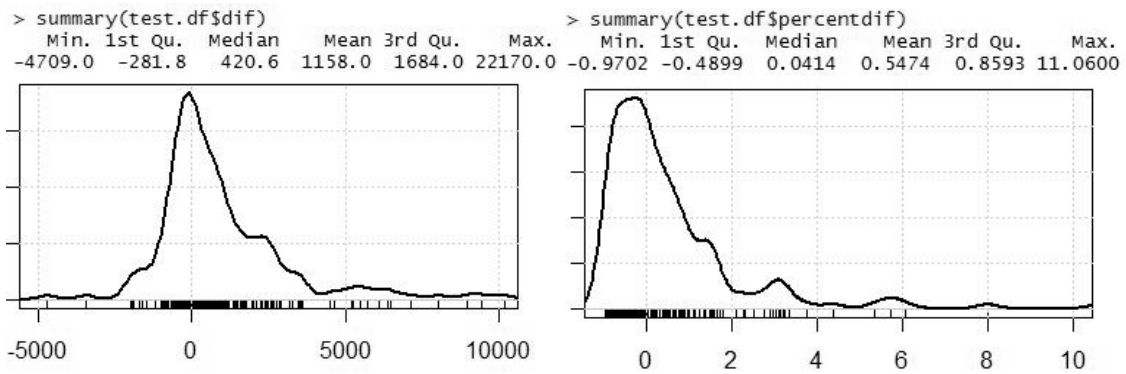


FIGURE 35 - DISTRIBUTION OF DIFFERENCES BETWEEN PREDICTED AND ACTUAL VALUE IN ABSOLUTE AND % DIFFERENCE

However, depending on the source similar tests reveal different capacity in determining channel value. As an illustrative example, we performed a test for predicting the aggregate value of each channel, which resulted in a predicted evaluation of **69.2%** of actual **overall** value, while mediums were attributed 72% of the value for referral traffic, 78.2% for direct, 64% for organic, 40.9% for referral and 69.7% for web. The higher the number of observations, the better the accuracy of the model.

5.1.2.1 Future Applications

These models rely on different variables to try to predict an expected outcome, given the historical weighing of each regressor. In this sense, if we have access to the values for each observation, the model relies on past indicators for trying to anticipate future or ongoing trends. The problem in this case is that when we get access to the indicators that allow us to infer on the value of each channel, in fact, we already know its value, since it is automatically given by GA. This model might however be used to compare time variations and the expected effect of the increment of new visits, through new campaigns and investments. Furthermore, the model gives us, according to the variations on the indicators, the expected performance according to past business conditions. In this sense, comparing the predicted versus the real value of observations can give us an indication of variation in both the behavior of customers and its effect on revenue for the company.

In this sense, with the current regressors we were only able to evaluate the value for the major sources of revenue, comparing for example expected trends versus current performance (*e.g.* if a model for the first half of the year has an accuracy of 90% and evaluates observations for the next month in only 70% of its value). The training process was in this case based on 778 observations, 80% of the data from this period. The higher the number of observations for each channel, the higher the precision of the attributed predicted value.

Other applications might also relate to the use of different dimensions, such as unique and anonym user ID (Kosny, 2014; Simpson, 2014), or types of regression, such as generalized linear models for predicting binomial distributions (logit), as in the example given by Araripe et al. (2013). Siegel (2013) also demonstrates the more advanced uses and future tendencies of these methodologies, from the use of client data for assessing a client's level of risk, from applications to the management of elections and other campaigns for selecting our target audience.

6 Concluding Remarks

Throughout this work we sought to explore the multiple dimensions of web analytics, starting by contextualizing the ambit of application and the main technologies available for the collection of online user data. In this way, this tool is perceived primarily as a monitoring tool, which helps us constantly monitor the most important trends and indicators in our site. Because of that, after an introductory section, we conducted a thorough analysis of the available reports, combining metrics and segments across the multiple reports and available dimensions. This led us to various interpretations on the website's traffic and the company's business, summarized for each report in a *Summary* section, included in our analysis.

In a second examination, which aimed to corroborate or disprove some of the main remarks made by the first set of reports, we monitored a second period of observations, conducting a series of tests (mostly proportions tests) exploring the changes in behavior of users or modifications in the nature of our business between periods. These are techniques also often employed after the realization of a marketing campaign, through which we evaluate the response of users to such actions, or if there are any evidences of significant changes over time. Working mostly with our spreadsheet, we were in this case able to compare in an agile manner different rates and proportions, revealing the significance of changes over time, or between different segments.

Lastly, we also conduct a series of analysis to some of the most relevant dimensions, adopting the possible segmentations to evaluate the extent to which the available metrics on user behavior can explain the variations on session, channel and total turnover. In this case we used linear regression to explore the effect of session engagement metrics on value, which as expected resulted in a poor performing model of only about 4.1% R-squared. This means that, in spite of the statistical significance and positive effect exhibited by engagement variables, as well as our channel and weekend variables significance, the model lacks much of the information that helps explaining the variation in the response variable. Because of that, looking at value from a session point of view is a highly limited perspective, particularly obvious in our B2B, high-involvement sales environment.

Following this, the utilization of the medium dimension also allowed us to segment traffic by their entrance channel, as well as the typical behaviors, type of customers and involvement generated by each channel. In this case, only the weekend variable failed to exhibit statistical significance, in a model that revealed to have a significantly better goodness of fit than in the first case, with an R-squared of 45.7%. In this way, it seems that when taking in account aggregate session values, the capacity of the model of explaining variations in turnover increasing, and with it its predictive capabilities. In this way, we used the supervised learning procedure, with a train and a test set to assess model performance, trying to predict the value of each channel on a certain date and comparing it with the actual values. The utility of having a fine-tuned model is in this way of establishing a benchmark of the expected performance, comparing it to actual business results. Wide variations would of course reflect major changes in business conditions.

Furthermore, this type of procedure is already used, with other applications and dimensions, to try to predict outcomes and the probability of events, for example at the user level, having in mind the metrics and statistical procedures at our disposal. One possible application of this for future research would be to use the anonym User ID (Universal Analytics version) dimension to employ relational techniques, in real-time, to follow the user lifetime cycle and explore the extent to what Unique User IDs can tell us something about user value or the probability for certain actions.

References

3 Scale Networks. (2011). What is an API? Your guide to the internet business (R)evolution. Retrieved May 1, 2014 from <http://www.3scale.net/wp-content/uploads/2012/06/What-is-an-API-1.0.pdf>

Akamai Inc. (2009). Akamai Reveals 2 Seconds as the New Threshold of Acceptability for eCommerce Web Page Response Times. Retrieved April 21, 2014, from http://www.akamai.com/html/about/press/releases/2009/press_091409.html

Alpar, A. (2013). Google AdWords Keyword Planner vs. Keyword Tool: SEO & PPC Feature Comparison. Retrieved May 10, 2014, from <http://searchenginewatch.com/article/2289304/Google-AdWords-Keyword-Planner-vs.-Keyword-Tool-SEO-PPC-Feature-Comparison>

Araripe, C., Gondaliya, A., & Shah, K. (2013). How to perform predictive analysis on your web analytics tool data. Retrieved June 20, 2014, from <https://www.youtube.com/watch?v=4zexsGKdlgw>

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., & Stoica, I. (2010). A view of cloud computing. *Communications of the ACM* (vol. 53) , 50-58 Retrieved December 10, 2013 from <http://dl.acm.org/citation.cfm?id=1721672>

Balamurugan, S., Vasuki, M., Angayarkanni, A., & Aurchana, P. (2013). Extend the Efficiency of a Website Using Web Analytics, *IJCTT Journal* (vol. 4 issue 6), 1693–1697.

Bloquiaux, L., & DeVuyst, P. (2013). Belgian e-commerce ready for the future. Retrieved May 10, 2014, from <http://www.insites-consulting.com/belgian-e-commerce-is-ready-for-the-future/>

Boslaugh, S., & Watters, P. (2008). *Statistics in a Nutshell* (1st edition.). O'Reilly Media.

Burby, J., & Atchison, S. (2007). *Actionable Web Analytics*. Wiley Publishing.

Calabrese, F. (2013). Un réseau d'autoibus redessiné grâce au téléphone mobile. *La Recherche* (nr. 482). 32-36.

Chau, R. (2013). Google Analytics attribution model comparison tool. Retrieved April 12, 2014, from <http://www.whymeasurethat.com/2013/06/26/google-analytics-attribution-model-comparison-tool/>

Clifton, B. (2012). *Advanced Web Metrics with Google Analytics* (3rd ed.). Indianapolis: Wiley Publishing.

Clifton, B. (2013). The rise and rise of “not provided” keywords. Retrieved May 10, 2014, from <http://www.advanced-web-metrics.com/blog/2013/02/01/the-rise-and-rise-of-not-provided-keywords/>

Coon, T. (1992). GNU General Public License - Terms and Conditions for Copying, Distribution and Modification. Retrieved February 1, 2014, from <http://www.r-project.org/COPYING>

Correia, J. (2010). Google Analytics PHP cookie parser. Retrieved July 22, 2014, from <http://joaocorreia.pt/google-analytics-scripts/google-analytics-php-cookie-parser/>

CrossValidated. (2014). Interpreting the residuals vs. fitted values plot for verifying the assumptions of a linear model. Retrieved July 22, 2014, from <http://stats.stackexchange.com/questions/76226/interpreting-the-residuals-vs-fitted-values-plot-for-verifying-the-assumptions>

Decuyper, A., & Blondel, V. (2013). Une vie privée est-elle encore possible? *La Recherche* (nr. 482), 38–42.

Delen, D., & Demirkan, H. (2013). Data , information and analytics as services. *Decision Support Systems* (vol. 55), 359–363. doi:10.1016/j.dss.2012.05.044

DeMers, J. (2013). How to Use Google Webmaster Tools to Maximize Your SEO Campaign. Retrieved May 10, 2014, from <http://searchenginewatch.com/article/2273660/How-to-Use-Google-Webmaster-Tools-to-Maximize-Your-SEO-Campaign>

Deprest, J. (2012). Belgium and its ICT industry. *Information Technology in Government Forum*. Retrieved July 1, 2014, from http://pt.slideshare.net/E-Gov_Center_Moldova/belgium-and-its-ict-industry

EDIT. (2014). Industry Sessions - Responsive Design. Retrieved February 20, 2014, from <http://vimeo.com/84622243>

Elisa DBI. (2013). *Google Analytics Case Study: Improving donations and email registrations for Merlin.org.uk*. Retrieved April, 12, 2014, from http://www.elisa-dbi.co.uk/wp-content/uploads/2013/02/GA_Case_Study_Merlin.pdf

Enge, E., Spencer, S., Stricchiola, J., & Fishkin, R. (2012). *The art of SEO* (2nd Edition). O'Reilly Media.

Fagan, J. C. (2013). The Suitability of Web Analytics Key Performance Indicators in the Academic Library Environment. *The Journal of Academic Librarianship*. doi:10.1016/j.acalib.2013.06.005

Fang, W. (2007). Using Google Analytics for Improving Library Website Content and Design : A Case Study. *Library Philosophy and Practice*. Retrieved February 12, 2014, from <http://digitalcommons.unl.edu/libphilprac/121>

Frost, J. (2013). Regression Analysis: How to Interpret R-squared and Assess the Goodness-of-Fit. Retrieved July 22, 2014, from <http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>

Google. (2013). What Is The Core Reporting API - Overview. Retrieved January 30, 2014, from <https://developers.google.com/analytics/devguides/reporting/core/v3/>

Google Inc. (2013). Google Analytics Cookie Usage on Websites. Retrieved March 10, 2014, from <https://developers.google.com/analytics/devguides/collection/analyticsjs/cookie-usage?hl=pt-PT>

Google Inc. (2014). Google Analytics Help Center. Retrieved March 27, 2014, from <https://support.google.com/analytics/>

Gupta, R., Mehta, K., Bhavsar, K., & Joshi, H. (2013). Mobile Web Analytics, *IJARCSSE 2* (3), 288–292.

Hasan, L., Morris, A., & Proberts, S. (2009). Using Google Analytics to Evaluate the Usability of E-commerce Sites. *Loughborough University Repository*.

Henoch. (2014). Ecommerce Belgium grows 26% to €1.91bn. Retrieved May 10, 2014, from <http://ecommercenews.eu/ecommerce-belgium-grows-26-to-e1-91bn/>

Hines, K. (2013). How to Use the New Google Analytics Advanced Segments. Retrieved March 5, 2014, from <http://blog.kissmetrics.com/new-google-analytics-advanced-segments/>

ICO (2011). The EU cookie law (e-Privacy Directive). Retrieved December 4, 2013, from http://www.ico.org.uk/for_organisations/privacy_and_electronic_communications/the_guide/cookies

James, J. (2012). Are regression models useful? Retrieved March 13, 2014, from <http://getdelve.com/2012/04/are-regression-models-useful/>

Kaushik, A. (2006). Excellent Analytics Tip#1: Compute Statistical Significance. Retrieved June 25, 2014, from <http://www.kaushik.net/avinash/excellent-analytics-tip1-statistical-significance/>

Kaushik, A. (2007). Data Mining And Predictive Analytics On Web Data Works? Nyet! Retrieved March 22, 2014, from <http://www.kaushik.net/avinash/data-mining-and-predictive-analytics-on-web-data-works-nyet/>

Kaushik, A. (2009). Manifesto for Web Marketers and Analysts. Retrieved November 26, 2013, from <http://www.kaushik.net/avinash/manifesto-web-marketers-analysts/>

Kaushik, A. (2010a). *Web Analytics 2.0*. Indianapolis: Wiley Publishing.

Kaushik, A. (2010b). *Web analytics 2.0: The Art of Online Accountability & Science of Customer Centricity*. Wiley Publishing.

Kaushik, A. (2011). The Difference Between Web Reporting And Web Analysis. Retrieved January 11, 2014, from <http://www.kaushik.net/avinash/difference-web-reporting-web-analysis/>

Kaushik, A. (2013a). Multi-Channel Attribution Modeling: The Good, Bad and Ugly Models. Retrieved March 15, 2014, from <http://www.kaushik.net/avinash/multi-channel-attribution-modeling-good-bad-ugly-models/>

Kaushik, A. (2013b). Search: Not Provided: What Remains, Keyword Data Options, the Future. Retrieved February 25, 2014, from <http://www.kaushik.net/avinash/secure-search-not-provided-keyword-analysis-data-sources/>

Kent, M. L., Carr, B. J., Husted, R. A., & Pop, R. A. (2011). Learning web analytics: A tool for strategic communication. *Public Relations Review*, 37(5), 536–543. doi:10.1016/j.pubrev.2011.09.011

Kosny, C. (2014). Custom dimensions and metrics in Universal Analytics. Retrieved July 25, 2014, from <http://www.knowledge.com/en/blog/2014/02/custom-dimensions-metrics-universal-analytics/>

Kutuçku, S. (2010). Using Google Analytics and Think-Aloud study for improving the information architecture of metu informatics institute website: a case study. Middle East Technical University. Retrieved January 30 2014 from <http://etd.lib.metu.edu.tr/upload/12612584/index.pdf>

Lee, H. J. (2011). Google Analytics for Digital Library Evaluation. *Tallinna Ulikool, Hogskolen i Oslo, Universita Degli Studi di Parma*. Retrieved March 21, 2014, from <http://hdl.handle.net/10642/987>.

Leek, S., & Christodoulides, G. (2012). A framework of brand value in B2B markets: The contributing role of functional and emotional components. *Industrial Marketing Management*, 41(1), 106–114. doi:10.1016/j.indmarman.2011.11.009

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing — The business perspective. *Decision Support Systems*, 51(1), 176–189. doi:10.1016/j.dss.2010.12.006

Miletsky, A. (2010). *Principles of Internet Marketing*. Course Technology.

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big Data Imperatives*. Apress.

Pakkala, H., Presser, K., & Christensen, T. (2012). Using Google Analytics to measure visitor statistics: The case of food composition websites. *International Journal of Information Management* 32, 504–512. Retrieved January 10, 2014, from <http://www.sciencedirect.com/science/article/pii/S026840121200062X>

Plaza, B. (2011). Google Analytics for measuring website performance. *Tourism Management*, 32(3), 477–481. doi:10.1016/j.tourman.2010.03.015

Polancic, G. (2007). Empirical Research Methods Poster. Retrieved February 2, 2014, from <http://www.itposter.net/itPosters/researchmethods/researchmethods.htm>

Price, C. (2013). How to Use Google Trends for SEO. Retrieved May 10, 2014, from <http://searchenginewatch.com/article/2292198/How-to-Use-Google-Trends-for-SEO>

Provost, F., & Fawcett, T. (2013). *Data science for Business: What you need to know about Data Mining and Data-Analytic thinking*. O'Reilly Media.

Reynolds, W. (2010). Important Exception to Google Analytics Last Click Attribution. Retrieved March 23, 2014, from <http://www.seerinteractive.com/blog/important-exception-to-google-analytics-last-click-attribution>

Root, E. (2010). Poisson regression. Retrieved April 20, 2014, from http://www.colorado.edu/geography/class_homepages/geog_4023_s11/Lecture07b_PoissReg.pdf

Sharma, H. (2010). Event Tracking Google Analytics & Universal Analytics. Retrieved May 10, 2014, from <http://www.optimizesmart.com/event-tracking-guide-google-analytics-simplified-version/#comments>

Sharma, H. (2012a). Advanced Attribution Modelling in Google Analytics. Retrieved April 5, 2014, from <http://www.seotakeaways.com/advanced-attribution-modelling-google-analytics/>

Sharma, H. (2012b). Google Analytics Cookies Explained in Great Detail. Retrieved March 10, 2013, from <http://www.seotakeaways.com/google-analytics-cookies-ultimate-guide/>

Siegel, E. (2013). *Predictive Analytics - Power to predict who will click, buy, lie, or die*. Wiley Publishing.

Simpson, D. (2014). How to send user IDs to Google Analytics. Retrieved July 22, 2014, from <http://davidsimpson.me/2014/04/20/tutorial-send-user-ids-google-analytics/>

Sultan, N. (2013). Knowledge management in the age of cloud computing and Web 2.0: Experiencing the power of disruptive innovations. *International Journal of Information Management*, 33(1), 160–165. doi:10.1016/j.ijinfomgt.2012.08.006

Tanner, J., & Raymond, M. (2012). *Marketing Principles*. Creative Commons. Retrieved February 12, 2014, from <http://2012books.lardbucket.org/books/marketing-principles-v2.0/index.html>

Villegas, D., Bobroff, N., Rodero, I., Delgado, J., Liu, Y., Devarakonda, A., Parashar, M. (2012). Cloud federation in a layered service model. *Journal of Computer and System Sciences*, 78(5), 1330–1344. doi:10.1016/j.jcss.2011.12.017

W3Techs Inc. (2013). Usage of traffic analysis tools for websites. Retrieved March 12, 2014, from http://w3techs.com/technologies/overview/traffic_analysis/all

Waisberg, D., & Kaushik, A. (2009a). Web Analytics 2.0: Empowering Customer Centricity. *SEMJ*, 2(1).

Wheble, D. (2013). How To Forecast Traffic Using Regression Analysis. Retrieved March 24, 2014, from <http://website-analytics.com.au/how-to-forecast-traffic-using-regression-analysis/>

Zhao, Y. (2013). R Reference Card for Data Mining. Retrieved April 10, 2014, from <http://www.rdatamining.com/>

Appendix

Channel Dimensions – Models and Diagnostics

Baseline Model

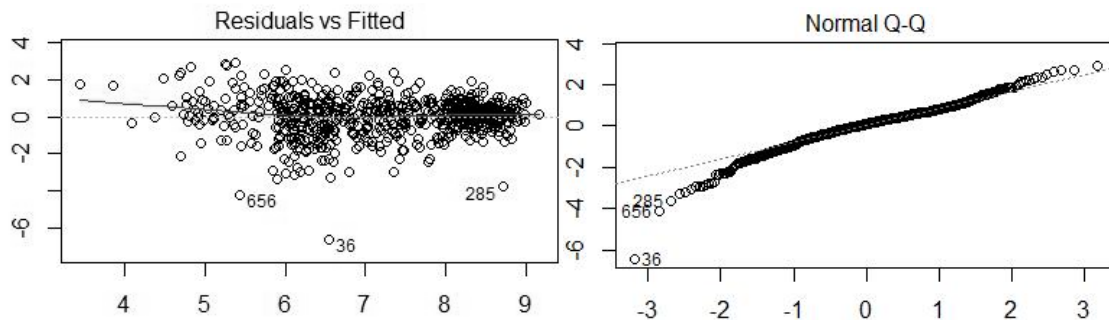
```
call:
lm(formula = lrevenue ~ lvisits + lpviews + ltimesite + direct +
    organic + referral + email + weekend, data = train.df)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-6.6315 -0.5268  0.0950  0.6131  2.9369
```

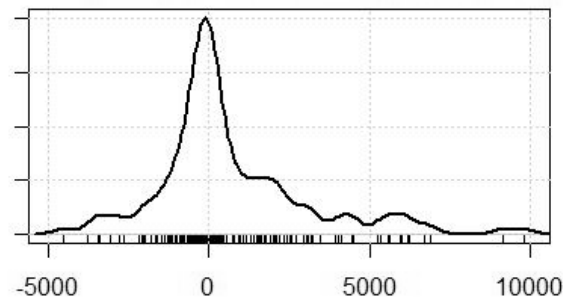
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8374     0.7076  -1.183  0.23706
lvisits       -0.7871     0.1562  -5.039 6.01e-07 ***
lpviews        2.2975     0.2655   8.652 < 2e-16 ***
ltimesite     -0.3636     0.1550  -2.346  0.01926 *
direct        -0.3741     0.1498  -2.497  0.01278 *
organic       -0.6297     0.2051  -3.070  0.00223 **
referral      -0.6751     0.1687  -4.001 7.00e-05 ***
email         0.2415     0.1309   1.845  0.06554 .
weekend      -0.4338     0.1389  -3.123  0.00186 **
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 675 degrees of freedom
Multiple R-squared: 0.5555, Adjusted R-squared: 0.5502
F-statistic: 105.4 on 8 and 675 DF, p-value: < 2.2e-16



```
> summary(test.df$dif)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4552.0 -526.0   66.3   780.5  1577.0 24740.0
```



Extended Model

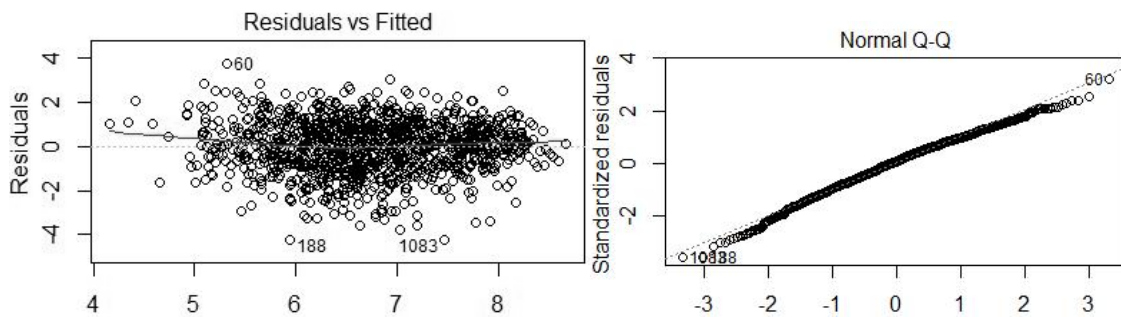
```
call:
lm(formula = lrevenue ~ lvisits + lpviews + ltimesite + search +
  direct + organic + referral + email + weekend + returning +
  mobile, data = train.df)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-4.2325 -0.7733  0.0731  0.8393  3.7734
```

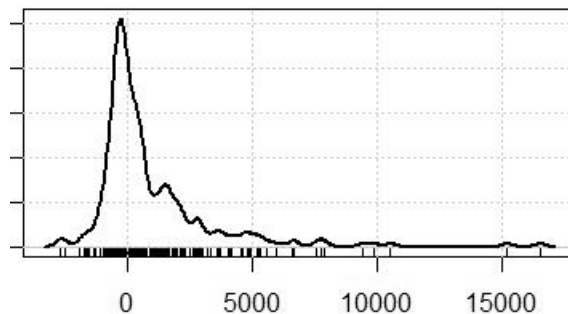
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01687   0.45006   0.037  0.9701
lvisits     -0.56088   0.10983  -5.107 3.84e-07 ***
lpviews      1.50169   0.17401   8.630 < 2e-16 ***
ltimesite   -0.09300   0.08530  -1.090  0.2758
search      -0.11603   0.13255  -0.875  0.3816
direct       0.01155   0.10719   0.108  0.9142
organic     -0.29519   0.11677  -2.528  0.0116 *
referral    -0.01556   0.15172  -0.103  0.9183
email       0.33152   0.13509   2.454  0.0143 *
weekend     0.10580   0.19367   0.546  0.5850
returning   0.60424   0.09772   6.184 8.74e-10 ***
mobile      0.33629   0.32060   1.049  0.2944
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.183 on 1131 degrees of freedom
Multiple R-squared:  0.3237, Adjusted R-squared:  0.3171
F-statistic: 49.22 on 11 and 1131 DF, p-value: < 2.2e-16
```



```
> summary(test.df$dif)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2672.0 -303.9   202.0   925.6 1507.0 16520.0
```



Selected Model

Call:
`lm(formula = lrevenue ~ lvisits + (lpviews:ltimesite) + web + direct + organic + referral + email + weekend, data = train.df)`

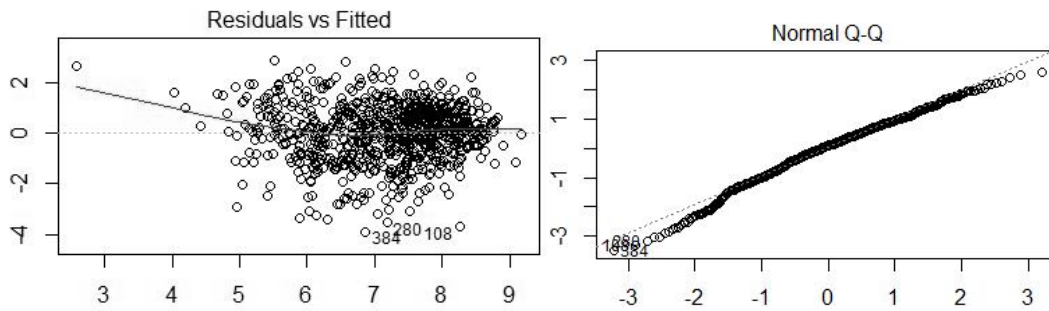
Residuals:
 Min 1Q Median 3Q Max
 -3.8181 -0.6210 0.0961 0.7067 2.7968

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.748744	0.324924	8.460	< 2e-16 ***
lvisits	-0.817621	0.090703	-9.014	< 2e-16 ***
web	0.882511	0.345325	2.556	0.010792 *
direct	0.970374	0.371092	2.615	0.009100 **
organic	0.768468	0.400746	1.918	0.055533 .
referral	0.721882	0.357036	2.022	0.043535 *
email	1.255493	0.341900	3.672	0.000257 ***
weekend	-0.297854	0.207625	-1.435	0.151814
lpviews:ltimesite	0.103128	0.006143	16.789	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.106 on 769 degrees of freedom
 Multiple R-squared: 0.4569, Adjusted R-squared: 0.4512
 F-statistic: 80.86 on 8 and 769 DF, p-value: < 2.2e-16



```
> summary(test.df$dif)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4709.0 -281.8  420.6  1158.0 1684.0 22170.0
```

