NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Harnessing Big Data to Inform Tourism Destination Management Organizations

João Pedro Martins Ribeiro da Fonseca

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2018

Harnessing Big Data to Inform Tourism Destination Management Organizations

João Pedro Martins Ribeiro da Fonseca

MGI

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

HARNESSING BIG DATA TO INFORM TOURISM DESTINATION MANAGEMENT ORGANIZATIONS

by

João Pedro Martins Ribeiro da Fonseca

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with specialization in Knowledge Management and Business Intelligence

Co Advisor: Prof. Leid Zejnilović, PhD, NOVA SBE, UNL

Co Advisor*:* Prof Miguel Neto, PhD, NOVA IMS, UNL

May 2018

ACKNOWLEDGEMENTS

I would like to thank my advisors, professors Leid Zejnilovic and Miguel Neto, for their support, advice and guidance in each stage of the development of this thesis.

Qiwei Han, for the continuous technical mentoring.

Lénia Mestrinho, for the invaluable work in the management of the project and partnerships among the involved external entities throughout the process.

Sergio Guerreiro (Turismo de Portugal) and João Ricardo Moreira (NOS) for making this project happen.

Margarida Abreu Novais, for providing the domain expert's perspectives, advice and discussions.

Iñigo Martinez and William Grimes, for the invaluable knowledge shared in our weekly data science journal papers discussions.

Data Science for Social Good Europe summer fellowship staff members and fellows, for all the knowledge shared throughout the program.

My parents and siblings, for the continuous encouragement, motivation and enormous patience.

My friends Francisco Martins, Pedro Rodrigues and Miguel Meco, for the unceasing support, thoughts and warmth.

All my friends in the music world, for the great moments that deeply supported me throughout the writing of this work.

To all others that were directly or indirectly involved in this thesis and all my friends and colleagues that accompanied me in my academic journey, my most sincere gratitude.

ABSTRACT

In the last few years, Portugal has been witnessing a rapid growth of tourism, which reflects positively in many aspects, especially in what regards economic factors. Although, it also leads to a number of challenges, all of them difficult to quantify: tourist congestions, loss of city identity, degradation of patrimony, etc. It is important to ensure that the required foundations and tools to understand and efficiently manage tourism flows exist, both in the city-level and country-level.

This thesis studies the potential of Big data to inform destination management organizations. To do so, three sources of Big data are discussed: Telecom, Social media and Airbnb data. This is done through the demonstration and analysis of a set of visualizations and tools, as well as a discussion of applications and recommendations for challenges that have been identified in the market.

The study begins with a background information section, where both global and local trends in tourism will be analyzed, as well as the factors that affect tourism and consequences of the latter. As a way to analyze the growth of tourism in Portugal and provide prototypes of important tools for the development of data driven tourism policy making, Airbnb and telecom data are analyzed using a network science approach to visualize country-wide tourist circulation and presents a model to retrieve and analyze social media. In order to compare the results from the Airbnb analysis, data regarding the Portuguese hotel industry is used as control data.

KEYWORDS

Tourism; Telecom; Social media; Airbnb; Sharing Economy; Gentrification; Tourism Flows;

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ICT	Information and Communications Technologies
INE	Instituto Nacional de Estatística
ABTA	Association of British Travel Agents
HTTP	Hypertext Transfer Protocol
API	Application Programming Interface
HTML	Hypertext Markup Language
GUI	Graphical User Interface
MCC	Mobile Country Code
MNC	Mobile Network Code
POI	Point of interest
IQR	Interquartile Range
ADR	Average Daily Rate
RevPAR	Revenue Per Available Room

1. Introduction

The ease of access to Information and Communications Technologies (ICT) have continuously improved over the last few decades, leading to a nearly effortless access to information by a growing percentage of the world's population. Learning about any place in the world is only a click away for anyone with internet access. The increasing efficiency and technology advancements have led to the continuous reduction of costs of transportation and consequently made long distance travelling unceasingly more affordable (Corwin & Pankratz, 2017). Increasing overall wealth in the world after the World War II allowed the allocation of more disposable years income for leisure activities, and the emerging image of tourism as a fashion industry with close links to the individual's status (Mason, 2016). All these factors, coupled with the reduced accommodation costs owed to emergence of sharing economies, made tourism increasingly more affordable and very popular.

Independent reports confirm such a scenario, showing that tourism industry is growing rapidly (*UNWTO Tourism Highlights: 2017 Edition*, 2017; *UNWTO Tourism Highlights: 2018 Edition*, 2018). In Portugal, tourism industry exports have grown by 33% (Pordata, 2018b) between 2016 and 2017, being distinguished with various tourism-related awards and the trust to organize major global events. Sudden growth of the tourists caught major cities in Portugal poorly prepared to accommodate such a demand in a sustainable fashion, introducing significant challenges for these cities.

The exponential growth of tourism has positive and negative impacts on society, and it is hard to strike the balance between the two. More tourists translate to more revenue, better services, job opportunities, and more attractive places for life. At the same time, with higher demand many living costs go up, and ordinary citizens may be finding it hard to make ends meet in big cities and tourism hot spots. Without facts, it is easy to jump onto conclusions. Big data and analytics may hold the potentials to inform about and optimize tourism, minimize negative and find and amplify positive impact. Providing facts and deriving insights from big data is essential for tourism management but also for healthy, substantiated, and constructive discussions and policy making. Still, there are many uncertainties about the types of data that can be useful for this purpose and how to capture the value from them. This thesis serves to explore three sources of big data and their use for tourism sector: i) social media data; ii) mobile roaming data, and iii) sharing economy (accommodation) data. In the thesis, first, for each a test of easiness to acquire data is conducted, to understand whether data are available and under which conditions. As an exercise, for the social media data, a prototype software was developed to identify viable alternatives for data collection. Second, preliminary exploratory analysis is conducted for structed data, namely AirBnB and mobile roaming data. Third, mechanisms for value capture from the data are discussed in the context of tourism management.

The motivation to focus on these three data sources is the capacity to provide information on two types of stakeholders. Mobile roaming and social media data can inform about the tourists, and provide answers to question like where do tourists go, where do they stay, what are their interest, and what was their experience like (in pictures and words). Airbnb data provides evidence on alternative accommodation supply – the service providers -, which when combined with the data published by Instituto Nacional de Estatística (INE) on the hotel accommodation supply offer a more complete view of the accommodation supply in Portugal. From the theoretical perspective, although big data sources addressed in this work are important for tourism management, they are still underutilized for tourism

research (J. Li, Xu, Tang, Wang, & Li, 2018), and particularly using a cross-method analysis of big data sources in tourism.

Part of the work done for this thesis is also presented as an interactive website on the following address: http://dssim.dssg-eu.org/reports/tdp/. Source code of the Social Media Crawler, telecom data analysis and Airbnb data analysis has been shared as an open source project on the following addresses:

1. Social Media Crawler: https://github.com/joaopfonseca/social_media_crawler
2. Telecom data analysis: https://github.com/joaopfonseca/tourism_telecom
3. Airbnb data analysis: https://github.com/joaopfonseca/airbnb_analysis

In conclusion, this dissertation focuses on shedding light over an increasingly important challenge: How can Big Data be used to inform and assist Destination Management Organizations in decision and policy making?

1.1. Research Objectives

The research objectives are specified in four main topics:

- Assess the potential of big data to inform destination management organizations
    - Study the type of information that can be collected from these sources of data
- Demonstrate this potential through the analysis of:
    - Airbnb: A market to which little is known about in Portugal
    - Telecom: Assess core information about the tourist (country of origin, destination, length of stay, etc.)
- Discuss applications of cross-method analysis and Big data in the context of tourism.
- Validate the results attained in both parts of the analysis and strengthen these findings through the usage of a cross-method analysis presented in the discussion section.

2. Background Information and Context

2.1. An Overview Of Tourism In Portugal

In 2017, for its first time, Portugal won the popular World Travel Awards title "Best European Destination" along with 30% of the awards of this renowned competition (PORTUGALPRESS, 2017). This award demonstrates Portugal's positioning as a global tourist destination. Lisbon municipality is the country's leading tourist destination. Having 504 thousand residents in 2016 (Pordata, 2018c), it welcomed over 4.5 million tourists in the same year (Pordata, 2018d).

In order to analyze the economic impact of tourism, one must understand what accounts for an export/import in the tourism industry. Eurostat defines an export of a good or service as a transaction from residents to non-residents (including smuggled goods). Therefore, export of services are those provided by residents to non-residents, regardless of where they are bought or consumed. Accordingly, Tourism exports is the sum of expenditures of non-residents in the the hosting country. In 2017, the tourism sector was responsible for €15.153,4 million in exports (Pordata, 2018a) – approximately 20% of the total exports (Pordata, 2018a), over 50% of the total exports of services and 12,1% of Portugal's GDP (tourism related activities, accommodation and catering) for the same year (Pordata, 2018e).

Similarly, this sector is also an important source of employment for the Portuguese population: According to a report issued by Turismo de Portugal, in 2016 7.4% of Portugal's working force was employed in the accommodation sector or had jobs linked with tourism, whereas another 4.7% worked in the catering sector (*População Empregada 2016*, 2017). These values have increased approximately 8% when opposed to 2015.

The tourism industry's current account has reached €10.8 billion in 2017 nationwide (Pordata, 2018b). A previous study focused on the impacts of tourism in the historical region of Lisbon (Ribeiro, 2017) demonstrate that after the 2008 crisis the investment into the rehabilitation of Lisbon's historical center arises as an opportunity to revitalize other industries, specifically the construction industry. The goal of this rehabilitation was to attract private investment and thus place Lisbon among the most sought after tourist destinations. These policies, although bringing clear benefits for Portugal, are generating social transformations on the city fabric, degradation of communities, loss of identity of historical neighborhoods, and gentrification: increased economic value of a region affects negatively the native population and lower wealth communities.

Although tourism brings economic benefits to the Portuguese economy, it is essential to also consider its volatility. A study by Banco de Portugal developed focused on the study of the volatility of the tourism sector (Daniel & Rodrigues, 2010). The study demonstrates not only the seasonality of the sector, but also that in general unanticipated shocks that affect tourism can yield a strong impact in the long run for visitors coming from the countries of origin with the highest representability: Germany, France and United Kingdom, except for domestic demand. In other words, an unanticipated event that negatively favors tourism can have a significant impact on visitors from some of the main countries of origin and is expected to endure for a considerable amount of time (Daniel & Rodrigues, 2010).

2.2. GLOBAL TRENDS IN TOURISM

According to studies focused on worldwide tourism trends conducted by the World Tourism Organization (*UNWTO Tourism Highlights: 2017 Edition*, 2017; *UNWTO Tourism Highlights: 2018 Edition*, 2018), both in 2016 and 2017 tourism represented 7% of the world's exports and 10% of the world's GDP (considering direct, indirect and induced revenue related to the sector for both years). In the table below the economic value of worldwide tourism and in Portugal is depicted:

	WORLDWIDE	PORTUGAL
TOURISM EXPORTS AS A PERCENTAGE OF GDP	10%	12,1%
TOURISM EXPORTS AS A PERCENTAGE OF TOTAL EXPORTS	7%	20%

Table 1: Comparison of Tourism's sector economic value in Portugal and worldwide (Year: 2017). Source: UNWTO and Pordata (adapted)

This comparison demonstrates the significance of the tourism sector in the Portuguese economy, given that its representability is significantly higher than in the world's economy.

2.2.1. Factors Affecting Tourism Flows

The research found regarding the general factors that affect the number of tourists visiting a country identify five broad categories that affect an economy's tourism flows (Prideaux, 2005):

- **Demand**: Lim found that the variables with the highest explanatory variables are income, relative prices and transport costs (Lim, 1999). The author includes personal preference, destination image, government regulations, personal financial capacity, political/military tensions, personal safety concerns, health epidemics and fear of crime.

- **Government Responsibilities**: Prideaux subdivides this factor into **diplomatic** (facilitate the issue of visas, recognition of passports, operation of transport modes and recognition of civil rights), **policy** (may limit the amount of tourist inflows, amount of currency taken out of the country, value and quantity of goods taken out of the country and imported by returning tourists), **marketing**, **regulatory regimes** (regulations include qualitative and quantitative controls over public and private sector organizations that operate in the tourism industry) and the **supply of goods and services** (provision of necessary infrastructures, maintenance of public health, provision of security and commercial services and education of the workforce).

- **Private Sector Factors**: Influenced by inbound and outbound operations, retail services performed by travel agents, wholesale travel services, travel insurance services and transport services.

- **Intangible Factors**: Refers to the built and natural environments, destination image, lifestyle, flows barriers and culture.

- **External Factors**: Prideaux further divides this category into 2 subcategories:

- **External Economic factors**: Includes efficiency of the hosting economy (i.e., international competitiveness – considered more efficient the lower the price level is), competition, exchange rates, national income levels and elasticity of demand.

- **External Political and Health factors**: Defined by factors beyond the ability of countries to control, such as wars, terrorism and the state of international relations.

Regardless, the author admits the existence of additional events and forces that influence tourism.

2.2.2. The Rise of the Sharing Economy

The goal of Sharing Economies is primarily to bring people with common interests together, generating niche networks with common needs and interests, which end up threatening established businesses, as peer-to-peer networks are capable of outgrowing the latter given the capability of taking benefit out of network effects (Cusumano, 2014).

The concept "Sharing Economy" derives from the term "Collaborative Consumption", coined in 1978 (Felson & Spaeth, 1978). Although, the concept only only gained popularity once internet became widespread (allowing web-based collaborative platforms to be established). In 2011 this concept was named by TIME Magazine one of the 10 ideas that will change the world (Walsh, 2011). The success of Sharing Economy services like Airbnb (founded in 2008) and Uber (founded in 2009) can be explained through a macroevironment (PEST analysis) perspective (Ertz, Durif, & Arcand, 2016):

- **Political**: Privatization trends arise, along with a de-politization of the state, giving citizens a higher sense of responsibility for their own welfare (motivated by the growth of the European Union).
- **Economic**: The rise of the concept is closely linked with the Great Recession of 2008, which led to the decline of full-time employment and purchasing power. Services operating under this concept are then seen as a source for cheaper solutions for the same needs and/or an additional source of income.
- **Societal**: Postmodernism movements brought out a more horizontal model in the consumption schemes of communities, structured as networks. The growth of the consumerism trend takes up a more central presence in the consumers' behavior.
- **Technological**: Development of the internet facilitated the effortless connection of digital communities, making the internet a source of influence in consumer behavior.

A PwC UK study identified 5 main types of Sharing Economies (Vaughan & Daverio, 2016):

- **Peer-to-Peer accommodation**: Digital platforms that enable individuals to rent unused accommodation. Some examples of companies operating in sub-sectors of this area include:
  - Rental: Airbnb
  - Home-swapping: LoveHomeSwap
  - Online-only vacation rental: HomeAway
- **Peer-to-Peer transportation**: Digital platforms that connects riders for both short or long distances. The most well-known services operating in this sector are Uber and Blablacar.
- **On-demand household services**: Crowd-based marketplaces for on-demand, crowdsourced delivery services and household task services (e.g., Uber Eats).
- **On-demand professional services**: Online platforms intended for connecting freelancers to individuals and businesses that require office tasks to be developed externally. Examples include Fiverr and Upwork.

- **Collaborative finance**: Crowdfunding and Money Lending/borrowing platforms without intermediaries such as banks, using the crowd as an investor. Two examples include Kickstarter and FundingCircle.

The rapid development of Information and communication technologies (ICT), social media, e-commerce and popularization of the Urban lifestyle, along with increasing volatility of natural resources' costs were the main driving forces for the development of Sharing Economies throughout the world and generated a whole new set of businesses.

Analyzing the case of Airbnb, some effects are linked to this platform. However, the transparency and independence of many studies is difficult to verify, as is the case of reports developed by Airbnb and Uber using internal private data (Codagnone & Martens, 2016).

Another study focused on the impacts of Airbnb on the hotel industry using data discriminated by time and space suggests that Airbnb's penetration into the market resulted in a revenue drop of 8% to 10% in the industry, being this drop non uniformly distributed. Additionally, the authors also found that hotels reduced prices, benefitting consumers (Zervas, Proserpio, & Byers, 2017).

Sharing economies emerged to reinforce ongoing trends and allowed communities to create additional sources of income by taking advantage of their unused assets, while providing users a different, more affordable solution. Although, there are few empirical studies to date on the impact of the Sharing Economies. Thus, the downfalls of said economy is yet, to some extent, unknown (Codagnone & Martens, 2016).

2.3. THE IMPACTS OF TOURISM

As observed, Portugal's economic development has been greatly assisted by the tourism industry. One must consider both the advantages and downfalls associated to the growth of the sector that is being experienced in the country.

Mason (2016) defends that tourism impacts are depend on 2 factors: When and Where. Given that tourism is (in general) a seasonal activity, in each period of a full year the impacts of tourism flows can vary in its importance/visibility. The seasonality of tourism is explained with 2 aspects: climate and holiday periods. These impacts may vary according to the region that's being analyzed, as different locations are endowed with different characteristics specific to the region, which will affect the level of sensitiveness of the region towards tourism (Mason, 2016).

Tourism impacts are distinguished in 3 types: i) Economic effects, ii) Sociocultural effects and iii) Environmental effects (Andereck, Valentine, Knopf, & Vogt, 2005; Haralambopoulos & Pizam, 1996):

2.3.1. Economic Effects

Previous studies demonstrate that tourism can yield both benefits and consequences from an economic perspective (Mason, 2016):

Positive Consequences	Negative Consequences
Contribution to foreign exchange earnings	**Inflation**: Increase in price of goods related to tourism (e.g., housing, land and food)
Contribution to government revenues	**Opportunity costs**: Cost of engaging in tourism rather than other economic activity
Generation of employment	**Over-dependence on tourism**: Making tourism the main source of a region's development
Contribution to regional development	

Table 2: Economic effects of tourism. Source: Mason, 2016 (adapted)

Mason mentions the economic effects as the most studied in tourism research literature. Although the remaining types of effects may be overlooked at first, they often represent important sources of motivation for anti-tourism movements from communities.

2.3.2. Sociocultural Effects

The term sociocultural refers to the residing population in a city (Social component), the way they interact and social relations are developed as well as material artifacts. These will set behavioral patterns and values (transmitted across generations), which will define the region's culture. This will play an important role in the promotion of tourism, due to its importance to the tourist's experience. Mason presents the following consequences of tourism in the Sociocultural panorama (Mason, 2016):

Positive Consequences	Negative Consequences
Creation of employment	**Overcrowding**: Which can cause stress for tourists and residents
Revitalization of poor or non-industrialized regions	**Loss of certain traditional activities**: When these are not perceived as attractive/interesting by tourists (e.g., farming)
Rebirth of local arts and crafts and traditional cultural activities	**Conflicting co-existence**: Tourists may have different values and engage in leisure activities, while residents are involved in working activities, which can be aggravated in the occurrence of seasonal tourism and in regions with strong religious codes
Revival of social and cultural life of the local population	**Rural-Urban Migration**: Caused by the lifestyle demonstration effect

Renewal of local architectural traditions	**Acculturation**: Deeper contact between different cultures can cause the convergence of the local culture into the second
Promotion of the need to conserve areas with aesthetic and cultural value	**Loss of culture authenticity**: Desire of tourists to experience the "real" culture can cause it to become forged

Table 3: Sociocultural effects of tourism. Source: Mason, 2016 (adapted)

One problem with the mentioned effects lie in their subjectivity. Even though they are difficult to measure, these types of consequences affect the destination's community and therefore shouldn't be disregarded (Andereck et al., 2005).

2.3.3. Environmental Effects

The natural environment refers to both natural and human features. It includes landscapes, rivers, beaches, water, climate, air etc. as well as individual buildings/structures, villages, townscapes, transport infrastructures, dams and reservoirs. In this group wildlife and the farmed environment are also included (Mason, 2016). The environmental impacts of tourism are presented as follows:

Positive Consequences	**Negative Consequences**
Stimulation of measures to protect the environment, landscape and/or wildlife	Litter generated by tourists
Promote the establishment of National Parks or Wildlife Reserves	Overcrowding of people and traffic congestion
Promote the preservation of buildings/monuments	Tourism can contribute to the pollution of water course and beaches
Money generated from tourism can be used towards the maintenance of historic buildings, heritage sites or wildlife habitats	Footpath erosion
	Can potentially generate the construction of structures (e.g., hotels) that do not fit with the present architecture
	Disturbance/damaging of wildlife habitats

Table 4: Environmental effects of tourism. Source: Mason, 2016 (adapted)

Lastly, the importance/weight of tourism impacts will differ significantly given each region of analysis.

2.4. THE TOURISM PANORAMA IN EUROPE

The number of international tourists in every continent has been continuously growing since 2009 (Statista, 2017). Out of these, over half visit Europe (excluding tourists visiting South America). This fact suggests that although the tourism sector in Portugal has had a significant growth, part of this growth is explained through the rising trend of tourism across the world.

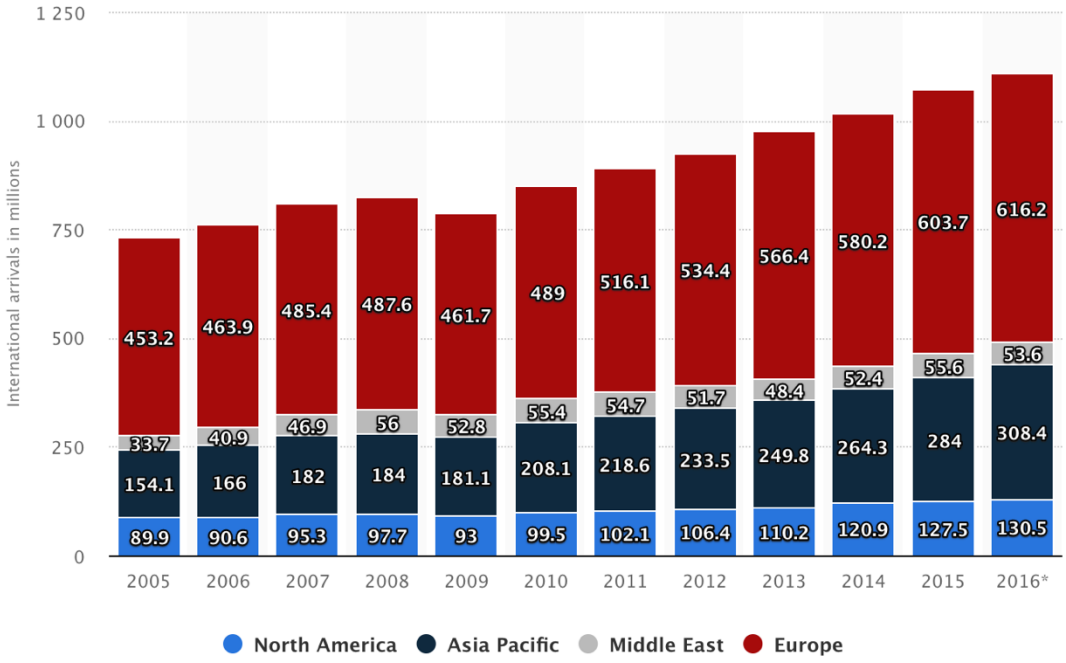


Figure 1: Number of international tourist arrivals worldwide from 2005 to 2016, by region (in millions). Source: Statista

In order to analyze some observable consequences of high tourism inflows, consider the following European cities: Berlin, Paris and Barcelona.

After Berlin's reunification (in 1990), the development of the tourism sector grew significantly over the years to follow, becoming the third largest tourist destination in Europe (Novy, 2011). Although the growth of this sector generated economic benefit for the city, it also generated many challenges (Novy, 2011). In some regions, party-tourism had a direct negative impact on residents, and both the number of complaints regarding legal or illegal conversion of rental apartments into tourist accommodations and the number of new hotels increased. These factors led to a negative impact on the city's urban fabrics, degraded neighborhoods (*Kiez*), imposition of regulation on the creation of new accommodation facilities and ultimately gentrification of the city (Novy, 2011).

After the economic crisis of 2008, Paris relied on tourism as a significant source of income to overcome it, which strengthened the ongoing adjustment of the city to mass tourism (Ribeiro, 2017). The locals resent these changes, claiming loss of identity of the city (Vallois, 2017). The cost of living in the center of this city increased, forcing many Parisians to move into the suburbs. Ever since the 2000's, the city hall implemented a set of regulations to battle the existing gentrification in the capital. Despite these efforts, between 2016 and 2017 the number of tourists in the city continued increasing (Ribeiro, 2017).

The number of tourist arrivals in the Catalan city have been continuously increasing (Statista, 2018). Accordingly, the resentment of the population against the phenomenon has been increasing over the last few years, resulting in frequent protests and other anti-tourism movements. The main motivation for this behavior is derived from the unaffordable growth of real estate in the city, leaving locals unable to pay their rents (which increased 16.5% in 2016 alone) and ultimately forced them to leave the city center (Plummer, 2017). The Association of British Travel Agents (ABTA) mentions as a part of the problem the existence of illegal (and unmanaged) tourist accommodation generated by online platforms such as Airbnb. In fact, 40% of Barcelona's tourist apartments are illegal, which undercut legal accommodation's pricing and allowed landlords to earn 4 times as much as they would by renting their apartments for long-term local tenants and made investors buy entire buildings to promote this activity (Díaz, 2017).

The city hall acknowledged the problem and is taking measures to control the phenomenon: application of taxes to tourism accommodation, stoppage of constructions of new tourist accommodation buildings since 2015 and restriction of tourist accommodation licensing depending on the area of the city (Díaz, 2017). The effectiveness of these policies are yet to be observed.

2.5. A VISION FOR SMART TOURISM IN PORTUGAL

Tourism patterns are changing. Populations can now benefit from the growth of the sharing economy (Quattrone, Proserpio, Quercia, Capra, & Musolesi, 2016) and affordable mobility options (Mason, 2016). The access to online resources allowed tourists to easily access to information, crucial to travelling aspiration and trace itineraries. These factors left cities ill-equipped to handle tourism flows that are growing to unprecedented levels. It must be every city's goal to avoid reaching a level of tourism demand that cannot be fully accommodated by its existing infrastructures.

The "Smart" tourism concept describes technological, economic and social developments that rely on Big Data, sensors, open data and new ways of connectivity and exchange of information, allowing the interconnection and synchronization of different technologies that create value and lead to innovation, entrepreneurship and competitiveness within the tourism market (Gretzel, Sigala, Xiang, & Koo, 2015).

In order to foster such concept, some efforts were already made by national authorities to better manage and promote tourism.

- A tool only existing in Portugal was developed to allow an Airbnb host to automatically register itself in Turismo de Portugal, having been implemented in December 2017, which is expected to make the detection of Local accommodation infringing the existing regulations more efficient (Machado, 2017).
- The Lisboa Card, a card that provides tourists access to 26 museums and monuments in the city, unlimited free travel by bus, metro, tram and elevators, various discounts, and free transportation to Sintra and Cascais, which facilitates mobility across Lisbon (Lisboa Card, 2018).

Sociocultural consequences of tourism are now being experienced in Portugal. According to Público (Portuguese news agency), in March 24th of 2018 a self-proclaimed Portuguese community named "Rock in Riot" has organized a parade against the excessive inflows of tourism in Lisbon (Sequeira,

2018). The motivations for this protest are the excessive tourist/citizen ratio, gentrification, as well as speculation in the real estate market.

In April 4th 2018, a new study developed by "Instituto do Planeamento e Desenvolvimento do Turismo" revealed that the Portuguese cities Lisbon and Porto have a higher tourist/citizen ratio than the ones presented by London and Barcelona (Jornal Público, 2018):

Portuguese city	**Lisbon**	**Porto**	**Albufeira**
Ratio	9	8	39
Foreign City	**London**	**Barcelona**	
Ratio	4	5	

Table 5: Tourist/Citizen Ratio in different cities. Source: Jornal Público, 2018 (adapted)

The benefits of taking a smart tourism management approach are clear for all involved parties (Neto, 2017):

Tourist	Increased support/assistance throughout its visit to the city, with instant access to crucial information that will highly impact the tourist's experience.
	Optimization of tourism flows through the recommendation itineraries, thus avoiding long waiting lines, less concentration of crowds in specific locations.
Local businesses	Personalized Marketing Campaigns (based on tourism behavior, profile and location).
	Heritage managers can clearly assess customer behavior and preferences, allowing the customization of tourism experiences and expanding demand beyond mainstream touristic locations.
	Identification of new business opportunities
Citizens	Reduced crowd cluttering across a city
	Integrated approach to planning and territorial management

Table 6: Benefits of Smart tourism. Source: Neto, 2017 (adapted)

2.6. STATE OF THE ART

In this chapter the state of the art of Big Data in tourism research will be presented, as well as tourist visitation sequences research and studies on the impact of tourism in Portugal.

Studies on the impact of tourism in Lisbon focus primarily on Lisbon's historical center. The positive and negative impacts of the city's tourism industry is analyzed, such as gentrification and territorial conflicts (Ribeiro, 2017). This study described the tourism sector and some of its impacts in the Portuguese tourism sector until 2017. Although, such analysis is difficult to replicate in posterior periods. The author finds low cost airplane passengers as one of the fastest growing forms of travelling, and hotel supply had a modest growth. Both these factors contributed to the growth of sharing economies in accommodation services.

The overall impact of the sharing economy based accommodation rental, was studied from a pricing perspective through the usage of Airbnb data, where the authors found twenty-five explanatory variables across five categories and analyzed the relationship between pricing and these determinants (Wang & Nicolau, 2017). Other studies focus on the impact of Airbnb in the hotel industry, demonstrating the negative impact on hotel revenue with the increase of supply of Airbnb listings, and offer empirical evidence that the sharing economy is successfully competing with incumbent firms (Zervas et al., 2017). Another study has been developed using survey data on the impact of Airbnb on Western Australia's Tourism Industry suggests that Airbnb users' behaviors appear to differ from the remaining tourists (Pforr, Volgger, & Coulson, 2017). Even though the economic value of these tourists is clarified, the impact of Airbnb in the overall tourism sector is not specified.

Social media data analysis has been used to study tourist behavior, specifically Miah et al. developed a big data analytics method to support decision-making in tourism destination management using geotagged photos from the website Flickr to analyze and predict tourist behavioral patterns using Melbourne, Australia, as a representative case (Miah, Vu, Gammack, & McGrath, 2017).

Studies focusing on tourism flows using mobile tracking data have varying research objectives: In one of the studies the authors propose a decomposition of tourism flows in five dimensions: i) spatial, ii) temporal, iii) compositional, iv) social and v) dynamic. and thus develop a monitoring tool that is being used by the Estonian Tourist Board (Raun, Ahas, & Tiru, 2016). Other sources of data for behavioral studies include online booking services. Batista e Silva et al. developed a Europe wide exploratory analysis of tourism with data from these sources (Batista e Silva et al., 2018). Although, both studies represent exploratory analyses for the corresponding data and little information regarding the used data was provided in all cases. No study was found complementing the analyses from all these different sources. All these studies have limitations in their data: booking services' data allows only the analysis of the location where tourists are accommodated and the granularity of the telecom and social media data used in the corresponding studies is limited, forcing these studies to develop a macro-level analysis of tourism.

Below is presented a comparative table with additional key studies on tourism flows, tourist space and tourist behavior, describing the data used, solution methods, key theories, and the purpose of the research:

Sources	Methods	Key theories	Data	Purpose
(Orellana, Bregt, Ligtenberg, & Wachowicz, 2012)	Data analytics	Computational Movement analysis was used to detect movement	GPS based tracking data	Flows of tourists in recreational destinations

(Y. Li, Xiao, Ye, Xu, & Law, 2016)	Space syntax analytics	Space syntax analysis following time series	GPS and location based sensors (high resolution video and picture)	Understanding of tourist space in China
(Zheng, Huang, & Li, 2017)	Flows' sequence analysis	Movement sequences' prediction using markov chains	GPS tracking data	Understanding tourist mobility
(Nilbe, Ahas, & Silm, 2014)	Data analytics	Analysis of travel distances between events' visitors and regular visitors	Passive mobile positioning data	Evaluate the behaviors of events visitors and regular visitors
(Shi, Serdyukov, Hanjalic, & Larson, 2013)	Predictive modelling	Usage of geotagged posts for landmark recommendation using a category-based approach	Social Media data	Personalized non-trivial landmarks recommendation
(Liu, Huang, & Fu, 2017)	Network Analysis	Visitation sequences as a measure of tourism attractions' popularity	Survey data	Exploration of underlying mechanisms of tourism attraction network
(Stienmetz & Fesenmaier, 2015)	Network Analysis	Analysis of tourists' visitation sequences	Survey data	Determine the value of the overall destination system
(Luo & Zhong, 2015)	Network Analysis	Social media networks to analyze user interactions	Survey data	Explain travel-related electronic Word-of-mouth on social networking websites
(Lee, Choi, Yoo, & Oh, 2013)	Network Analysis	Analysis of spatial centrality through constructed network structures	Roads data and village coordinates	Evaluation of integrated development strategies classified by centralities
(Sun, Wei, Tsui, & Wang, 2019)	Predictive modelling	Granger causality between tourist arrivals and search engines' index	Search engine data (Google and Baidu)	Forecasting tourist arrivals

Table 7: Comparative table on tourism flows, tourist space and tourist behavior analysis. Self-Authorship.

A recent literature review on Big Data in tourism research (J. Li et al., 2018) shows the data sources used for this type of research divided in three categories: Users (Online textual and photo data), Devices (GPS, Mobile roaming data etc.) and Operations (Web page visiting data, Online booking data etc.). The authors present an analysis on the evolution of research in this field, which has been growing since 2007, going from 3 published articles in this area for the mentioned year, up 30 articles published in 2016, dropping to 18 in the following year.

No study was found incorporating cross-method analyses. Individually, these data sources have been used in tourism research, but with common limitations: GPS data was based on a sample of tourists either carrying devices or in movement detection (thus impossible to trace visitation sequences), mobile positioning data had low granularity and often based on samples, social media data yields an error margin for location detection/sequence analysis and survey data is prone to biases and dissonance between announced and explicit behavior.

The development of data processing technologies and the generation of massive-scale data (J. Li et al., 2018), along with the development of computationally intensive sequential algorithms (such as word2vec/skip-gram models, popularization of Recurrent Neural Networks composed of Long short-term memory units), the cross analysis of different sources of data have the potential to provide novel types of analyses that up to this moment were not possible to execute. This data can be collected from existing infrastructures, meaning that limitations originated in the process of data acquisition can be avoided: instead of limited sample sizes one can use large portions of the population and eliminate sampling bias. Hence, tourist's sequence research is still in an early development stage.

Based on the analysis of the aforementioned literature, the generalist big data tourism research process has been observed to be structured as follows:

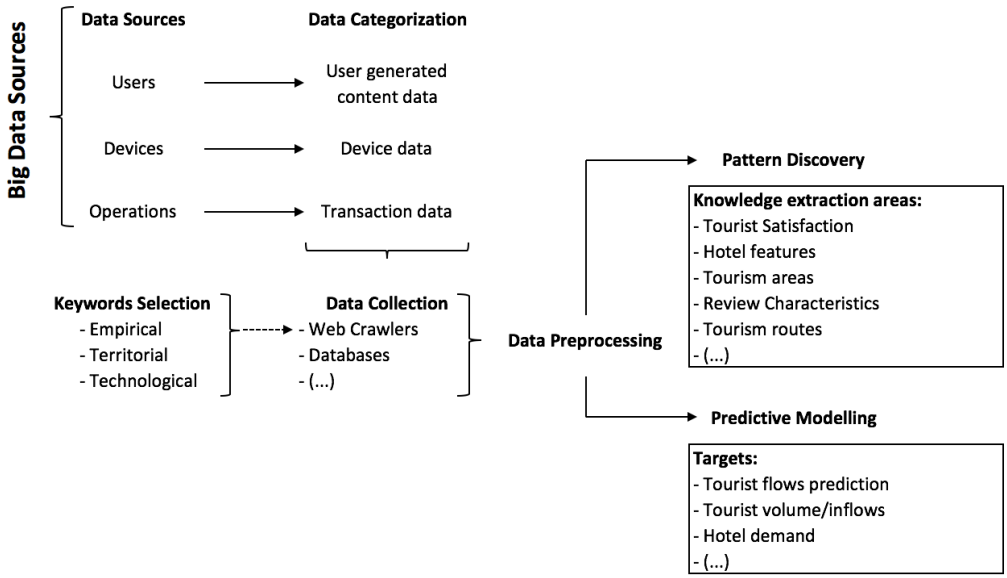


Figure 2: Big data tourism research process. Adapted from J. Li et al., 2018

As such, this thesis will follow a similar structure as the one presented in the diagram above, using the pattern discovery path through an exploratory analysis of the gathered data.

The aforementioned literature review (J. Li et al., 2018) states the importance of these big data sources for hospitality and tourism research: Tourism represents a complex system involving a series of processes such as web searching, webpage visiting, online booking/purchasing, landmark visitation, social media activity, etc., which carry effects that are often difficult to study and quantify using traditional methods (e.g., gentrification, overcrowding and coexistence conflicts). Thus, the analysis of this system as a whole using big data analytics will be capable of developing new knowledge towards a more complete understanding of the hospitality industry, tourism demand, tourist behavior and satisfaction and consequently support tourism marketing and decision making, which was not possible to achieve before the development of big data processing technologies.

2.7. Big Data Infrastructures as an Enabler of Smart Tourism

Big Data comprises three types of data: Structured, semi-structured and Unstructured data (Oracle, n.d.). Structured data consists in clearly defined types of data that are easy to query, often stored in relational databases. Although it was not until mid 2000's that unstructured forms of data such as social media data started being used. These include text, audio, video and images. Unstructured data require different approaches both in the preprocessing and analysis/modelling steps.

The voluntary sharing of personal content as publicly available data through social media platforms converted them into a valuable source of data for user behavior studies (Miah et al., 2017). The data that is extracted from these networks is usually unstructured, as it can be in the form of text, video, sound or image. Additionally, this data circulates at a high velocity, variety, volume and complexity.

Telecom data is regarded as structured data. This data can be stored in a relational database and doesn't require any special preprocessing approach. It contains additional supporting columns connected by primary keys and foreign keys, which are meant to give meaning to the codes present in the main table.

Airbnb data represents semi-structured data. Although there are some links that can be made across tables, additional preprocessing is required: Extraction of structured data (country of origin) from text data.

Although the processing of large amounts of information has been done for a long time, the concept was coined in the early 2000's by Doug Laney, when this industry analyst defined Big Data as the three V's: Volume, Velocity and Variety (Laney, 2001). Later on, new features have been added to the definition of Big Data, which vary among the entity that is defining the term. Some examples include Veracity, Value and Complexity (De Mauro, Greco, & Grimaldi, 2014).

In other words, Big Data consists in high-volume, high-velocity and high-variety of structured, semi-structured or unstructured information assets that require scalable and cost effective information processing tools that allow a new approach to data analysis and informed decision making.

A software library commonly used to handle Big Data is known as Apache Hadoop (Apache, n.d.), which is a project that develops open-source software for the "distributed processing of large datasets across clusters of computers using simple programming models", being capable of horizontal scaling (go from a single server to thousands of machines), instead of vertical scaling (increase of capacity of a single server). This software library brings several advantages to data processing: highly scalable, cost-

efficient, distributed and reliable computing power (the library is capable of detecting and handling failures in these nodes through data duplication across nodes).

Fuchs et al. show the practical usage of Big Data for strategic planning purposes in tourism destinations through Business Intelligence procedures (Fuchs, Höpken, & Lexhagen, 2014) and was adopted by a Swedish destination management organization. Another study focusing on Big Data analytics to crawl User-Generated Content to study tourist behavior in Barcelona was developed (Marine-Roig & Anton Clavé, 2015). This work proposes a method for processing data divided in four stages: Web Hosting Selection, Data Collection, Pre-Processing and Analytics. Since then more work has been developed in this area, although none of them considers the joint analysis of multiple sources of Big data.

3. METHODOLOGY

3.1. SOCIAL MEDIA CRAWLER

The aim of this project is to provide the end user all the data that is available from three social media networks (Facebook, Instagram and Twitter) regarding a specific topic of analysis (which can include multiple keywords – currently up to three), as well as an interactive dashboard containing a basic analysis of this data using traditional Business Intelligence procedures and visualizations.

With such tool, the end user is also be able to access the source data. Therefore, this tool is directed to any research group, organization or individual that wishes to take advantage from social media data analytics.

The three Social Media platforms implied different approaches given the distinct platform architecture, types of available data, and ease of access to such information.

The overall work breakdown structure of the project is as follows: Starting with the development of the tool to collect data from Facebook using a single Facebook page for experimental purposes: CMCascais. After the successful collection of this data, this method was then applied to all the pages associated with a specific keyword search. After this, the Twitter data collection tool was developed. In this situation we started by using the keyword "cascais" to fetch the most recent 2000 tweets containing such keyword. As Twitter uses their API services as a source of revenue, the data available for collection without using any paid plan is restricted. With this software it is possible to scrape 2000 tweets every 15 minutes and cannot be older than 1 week prior to the date of querying. Finally, the Instagram data crawler was included in the project. It has the capability of downloading the pictures in each post, as well as store its data, provided they contain a specified hashtag (e.g., If the keyword chosen is "cascais", the program will fetch posts containing the hashtag "#cascais").

All the data obtained with this Crawler is public and available to any user through the corresponding social media network's platform.

The web crawler starts with a list of URL's to visit (also known as seeds). So, if one requires to crawl posts from Instagram with a specific search query, e.g., "Cascais", there would be an Hypertext Transfer Protocol (HTTP) request (known as the HTTP GET method, used to retrieve data from a web server) by using the corresponding address to perform an HTTP request. The information is returned in JSON format. Application Programming Interfaces (API) are being used to get the parsed information from the pages. While a regular HTTP request returns Hypertext Markup Language (HTML) code, an API will return data in a format that is already parsed.

Finally, in order to develop a Graphical User Interface (GUI), a web framework was used. An important advantage of developing a GUI in the form of a web app is the ability to easily access and configure the software through the interface even when it is running in a remote server.

3.1.1. Problem Statement

The problem being addressed refers to the difficulty of access to structured data from these sources. Additionally, there is a lack of good, reliable and affordable solutions for this task.

3.1.2. Project Design

The program is divided in 6 parts: collection of data from the 3 social media channels, for both collection of data in a batch (when there is no stored data regarding the given keyword) and updating existing data (for all 3 channels).

Facebook Crawler:

Tools used: Facebook's Graph API (Facebook, n.d.), Facebook-SDK Python Library (Facebook & Mobolic, n.d.).

This tool was directed to fetch data regarding Facebook Pages' activities and user interactions with them. In other words, for each crawled Facebook page the program will store all the published posts (including content, creation date, ID etc.) by the page, each post's like and share number, as well as its comments, including content and creation date. It has no restriction on post creation date, and virtually no restriction on the number of requests made per period of time.

The development of the script responsible for scraping public Facebook page's posts was adapted from a GitHub project: facebook-page-post-scraper.

After performing the first batch crawl (hence storing all the historical data, tracing back to 2009), the data can then be updated regularly. The crawled posts are associated to the Facebook page ID and the comments are also associated to the corresponding post's ID.

The following diagram depicts the steps done in order to fetch data from posts in Facebook Pages and the comments linked to these posts:

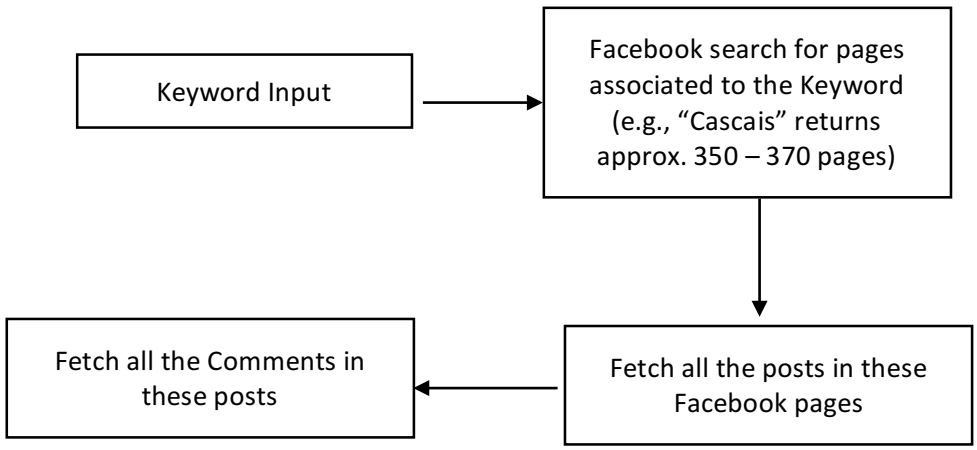


Figure 3: Facebook Web Crawling Process.

Although, given the recent Cambridge Analytica scandal that peaked in March 17th of 2018, Facebook temporarily disabled the access of its data to non-active apps (being this the case for the apps registered for the development of this project) in order to perform changes to their API's and thus avoid future exploits of the service as was the case with Aleksandr Kogan's app, which collected confidential data of over 87 million users as app users agreed to share personal information with the app, which allowed the app to collect data about the users' friends when their privacy settings were enabled (Bloomberg, 2018). Below is depicted the returned message once an HTTP request is done to Facebook's Graph API:

```
facebook.GraphAPIError: (#200) Access to this data is temporarily disabled for non-active
apps or apps that have not recently accessed this data due to changes we are making to the
 Facebook Platform. https://developers.facebook.com/status/issues/205942813488872/
```

Figure 4: Graph API Access Error Message.

Most relevant data stored (posts): Post ID, message, post type, creation date, likes count, comments count and shares count

Most relevant data stored (comments): Post ID, Comment ID, Comment message and comment creation date

Twitter Crawler:

Tools used: Twitter's Search API, Tweepy Python Library

Twitter contains 3 API products:

- **Standard Search API**: Free version of the Twitter API. Limits requests to a maximum of 100 tweets per request (although this limit is not currently being used in the project), allowing a maximum of 180 requests per each 15 minutes. Tweets crawled can only be as recent as 7 days-old.
- **Enterprise Search API**: Low-latency, full-fidelity, query-based access to the Tweet archive. Tweets crawled can as recent as 30 days-old, or alternatively all tweets from as early as 2006.
- **Premium Search API**: Similar as Enterprise Search API, for a different customer segment.

The program can crawl 2000 tweets every 15 minutes using the Standard Search API. After this, it is able to continuously update this data and append to the existing data new tweets with the same keyword search. However, it is limited regarding the storage of historical data prior to the date of crawling.

Most relevant data stored: Location Coordinates, Creation Date, Favourite Count, Tweet ID, Post Language, Location Name, Retweet Count, Tweet Message, User name, User description, User favourites count, User followers count, User friends count and User Location.

Instagram Crawler:

Tools used: Instagram-API-python Python Library (unofficial Instagram API, Python library developed by GitHub user LevPasha).

As Instagram no longer maintains their API (Instagram, n.d.), the tool that was used to crawl data from this platform was an unofficial crawler library developed in Python. The program can crawl a maximum of approximately 2 250 posts each time it runs and is able to run every 15 minutes. It does not restrict the data being fetched by its posting date. However, the API does not allow to query according by period of time. The Crawler is also capable of downloading the pictures of each post being stored through HTTP requests and are stored with the post ID as the pictures' names.

Most relevant data stored: Creation time, post ID, Message, User Name, User ID, Picture URL, like count, Video posted URL, View count, comment count, Location name, latitude and longitude.

Graphical User Interface (GUI):

Tools used: Flask Python Library and Bootstrap

The GUI allows any user to manage which keywords are being crawled and updated through a keyword submission form and select from which source the user wants to get/update the data from. Currently the software supports only a 3 keyword input, although this capacity can be expanded. The existing data for the active keywords can be analyzed through dashboards adapted for each social media dashboard, which allows the user to have a general idea of the data present in the datasets.

The GUI is composed by 3 sections:

- **Homepage**: The way the crawler works is explained.
- **Configurations**: Manage updates and set active keywords
- **Dashboards**: Basic info regarding the existing data

3.2. TELECOM

The analysis of Telecom data was done using a dataset that contained one month of telecom data (August 2017) for users with a foreign phone number, provided by the Portuguese telecommunications company NOS telecomunicações. The data consists in network events data. Part of the work in this section was inspired by a project developed in 2017's edition of Data Science for Social Good (Lozano, Flament, & Malik, 2017), out of which some of the code was used and adapted to fit the purposes of this analysis.

Network events data consists in a mixture of various protocols and network events, both active and passive, to which we will refer to as Network Events (or simply events). The provided data does not discriminate over which type of signal is being recorded and the most sensitive bits of data were either anonymized or removed. This data contains the anonymized user identifier for the customer, the nationality of the user's SIM Card, the date and time of the event, coordinates of the network tower's cell associated to the network event

Network events can originate from different network types: Public Switched Telephone Network (referred to as mobile data), text messages (SMS), phone calls, pings between the phone and the telecom tower, etc. For the majority of events there is information regarding the cell to which the tourist is connected, the tourist's nationality, the network provider and the phone's brand and model.

In addition to the network events data, telecom towers' data was provided. These cell towers are typically divided by 3 cells: responsible for controlling, formatting, receiving, routing and transmitting Network events and cover different angles out of the total 360º. In general, a device will be connected to the closest node (which might not necessarily happen, e.g., to manage overhead throughout neighboring nodes). With this data, it is possible to analyze the flows of tourists. The proposed approach uses network science principles, where nodes and edges will be generated for further analysis.

3.2.1. Data Preprocessing

The data is divided in 7 tables, containing different types of information:

Table Name	General Info	Relevant variables
continentes2	Duplicate of 'country_features' but with no economic variables	mccmnc__country (PK) continenten
country_features	Associates Continents to Countries, also contains economic variables for each country	mccmnc__country (PK) continenten gdp_nominal_per_capita longitude_avg latitude_avg
linguas_moedas	Associates Countries to currency and language	country (PK) currency language
mccmmc_optimized_new	Associates Mobile Country Code (MCC) and Mobile Network Code (MNC) to user's country of origin and network provider	mcc mnc new_country (FK) network
site_lookup_outubro	Contains data referring to the cell tower and telecom cells	ci lac protocol_ centroide_longitude centroide_latitude
tac_lookup	Associates user to corresponding phone brand and model	tac (PK) manufacturer model_name
union_all	Contains data regarding user events, associated to phone brand/model, MCC, MNC as well as the date of event	client_id enddate_ cellid lac_ protocol_ mccmnc tac

Table 8: Telecom metadata. Self-Authorship.

Out of these, only the tables “mccmnc_optimized_new”, “site_lookup_outubro” and “union_all” will be used for this analysis. As network providers will not be analyzed, the columns “mnc” and “network” will be dropped. Additionally, it was observed that the country Guam had the same MCC as the United States. Given the residual representation of Guam for the total number of tourists in Portugal, we will consider all tourists having said MCC to be tourists from the USA. Hence, we now have a primary key in the table “mccmnc_optimized_new”: MCC.

The relationships between the tables are structured as follows:

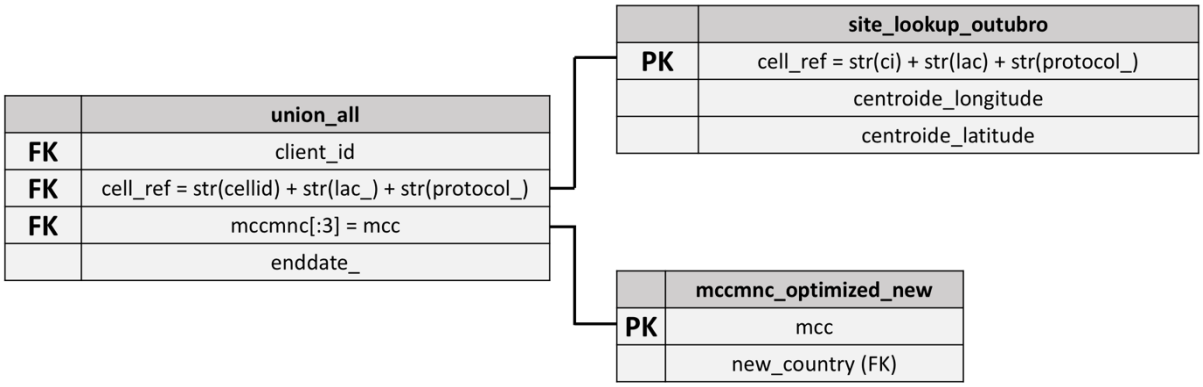


Figure 5: Telecom data - Relationships across tables.

A table containing preprocessed nodes data was generated, containing the Cell ID (= "cell_ref"), longitude and latitude of centroids and site information (district, municipality and closest POI).

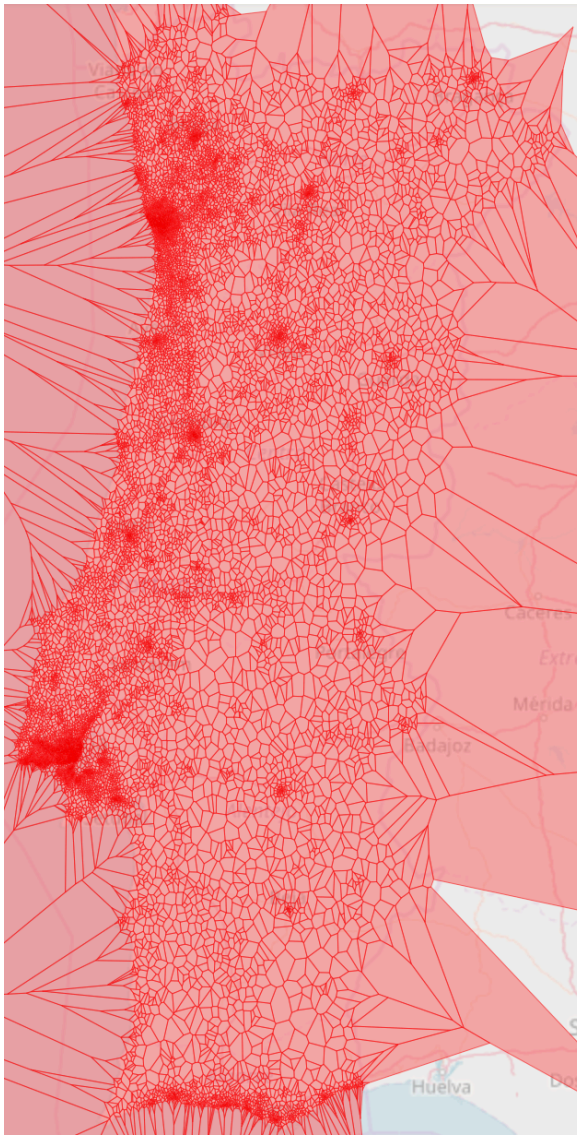
3.2.2. Analyzing telecom tower nodes

The nodes for tourism flows analysis were generated using the centroids of each node. The reasoning behind this procedure implied that whenever a user is connected to one of these nodes, it is likely that the user is located closest to the centroid of the node he/she is connected to. In other words, the user is likely located within the Voronoi Cell determined by the centroid of the Telecom tower node.

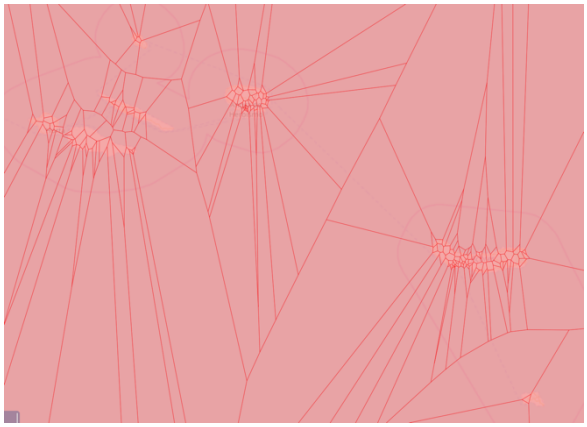
A Voronoi Cell is contained in a Voronoi Diagram, which is the partitioning of a plane in which a set of points is represented and generating a geometry (Voronoi Cell) for each represented point. These geometries are generated through the calculation of equidistant boundaries between a given point and the ones around it, i.e., for each point (seed) a geometry that will only contain the given seed will be generated, where the area of the given geometry consists of all points closer to that seed than to any other (Aurenhammer, 1991). Formally, the dominance of point p over q is defined by the problem:

$$dom(p,q) = \{x \in R^2 \mid \delta(x,p) \leq \delta(x,q)\}$$

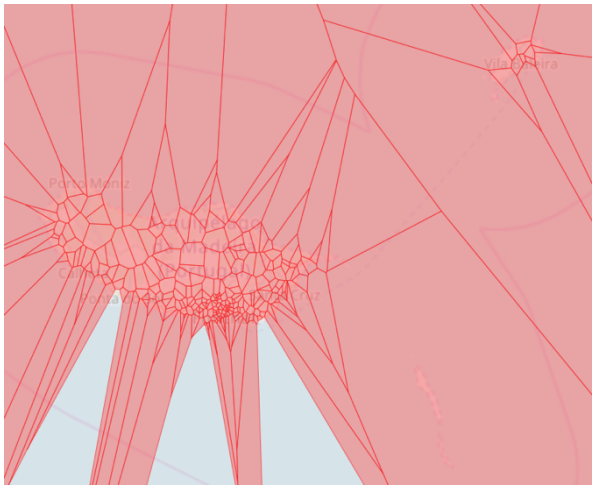
Where p, q are contained in the set of points represented in the plane and δ represents the Euclidean distance between the two points. Although other distance methods can be employed for the calculation of the Voronoi Diagram, the most popular method is the Euclidean distance, which was the one used here. The result of the Voronoi Diagram is presented below:



Continental Portugal



Azores Archipelago
(no cells in Corvo and Flores islands)



Madeira Archipelago

Figure 6: Voronoi Diagram using telecom tower's node's centroids.

The range of the nodes becomes visible, where in larger cities (such as Lisbon and Porto) the granularity of nodes is the highest. For example, when zooming in into Greater Lisbon area this is the result:

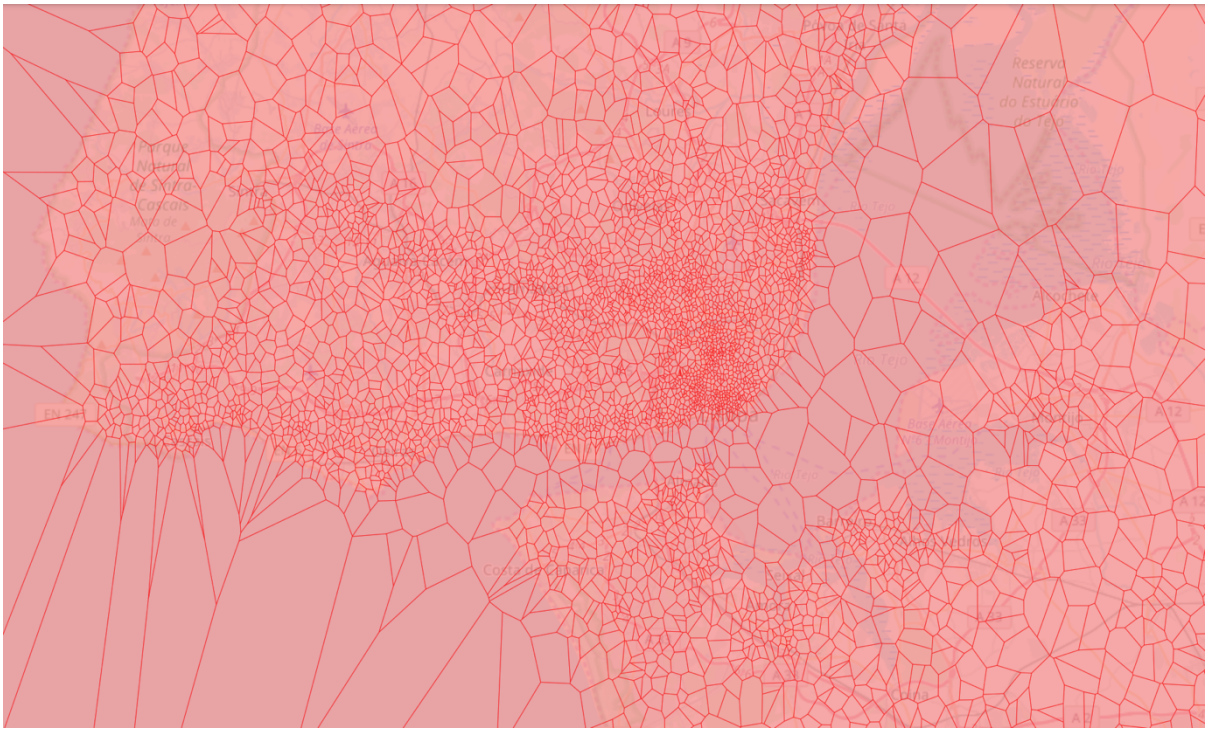


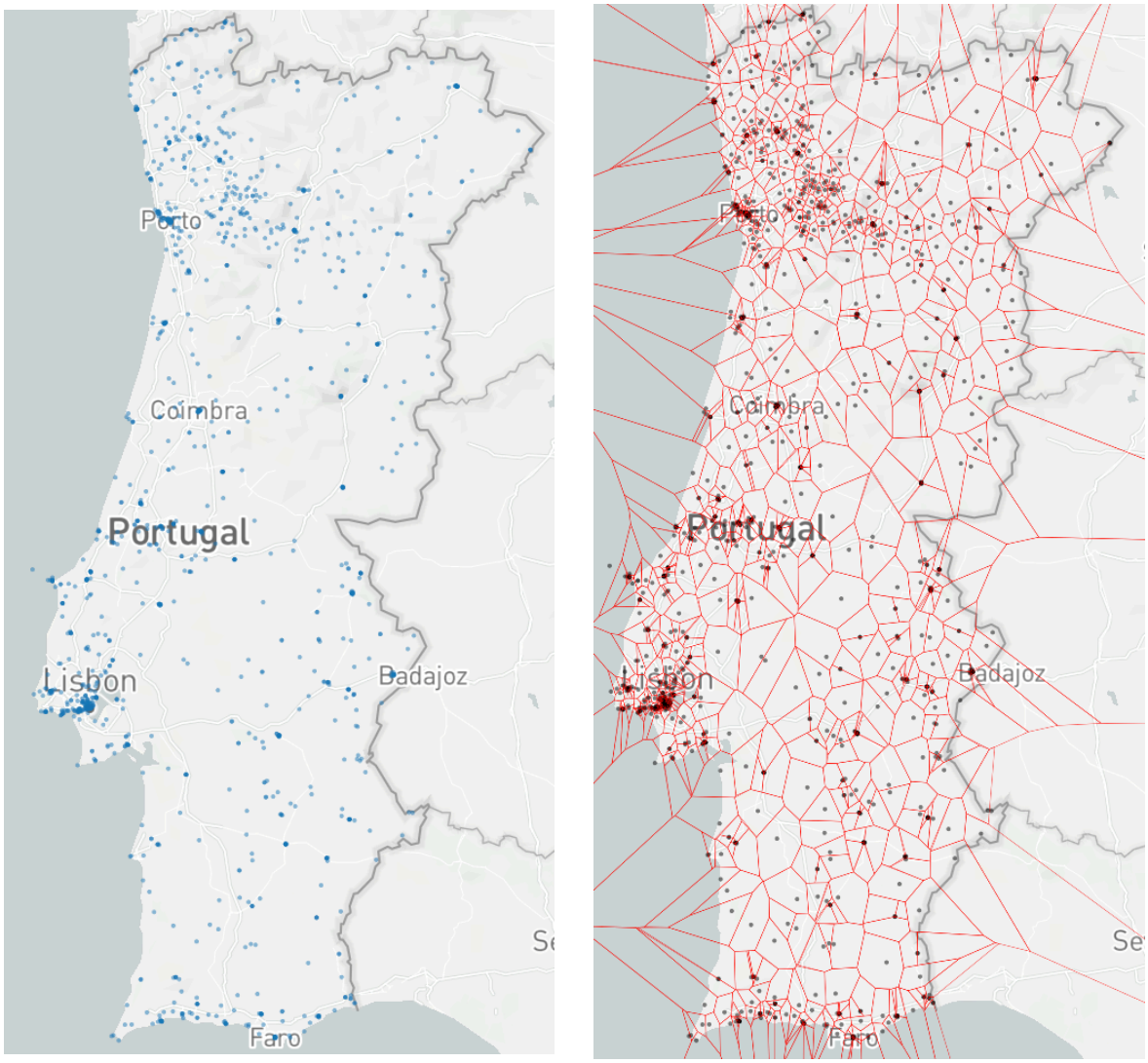
Figure 7: Voronoi Diagram using telecom tower's node's centroids in Greater Lisbon.

In these regions it is likelihood in which the user could be connected to a neighboring node from its the closest one to its location is higher, significantly increasing the error margin when estimating the user's visitation patterns. Additionally, the visualization and analysis of data using these nodes becomes difficult, due to its excessive granularity.

The way this challenge can be addressed is through the grouping of nodes. To do this, three possible methods could be used:

1. **Hierarchical clustering**: Join nodes by their relative proximity to each other.
2. **K-means clustering**: Partition nodes in K groups
3. **Manual clustering**: Set centroids' location and use the corresponding Voronoi Diagram to assign nodes to each group.

Given that the first two methods may output undesirable results, the grouping of nodes will be done using manual clustering. To do this, a dataset with touristic Points of Interest across Portugal was provided by Turismo de Portugal. This dataset contains the name, coordinates and type of attraction for a total of 1437 POIs:



Nodes' locations

Resulting Voronoi diagram

Figure 8: Voronoi Diagram using touristic Points of Interest

Using this method, the excessive granularity of the previous set of nodes was removed. To generate the network's edges for the flows visualizations, the Telecom tower nodes will be labeled using the POI's location and intra cluster flows will be filtered out.

Another table with information regarding each roamer was then developed: Day of month and Weekday of first and last contact with the network, count of days of stay, country of origin, list of locations and count of municipalities, districts and Points of Interest (POIs) visited.

3.3. AIRBNB

Four datasets were used for this analysis, containing information about all properties listed in the platform and located in Portugal, booking activities (which includes Accepted Bookings, Blocked Bookings and Cancelled/Unanswered bookings), monthly details for each property's bookings (revenue, occupancy rate and number of bookings) and Public Reviews' information associated to basic user information (review text, non-standardized customer's country of origin and job). However, it is not possible to link a public review with the corresponding booking.

In order to prepare the data for analysis, it is necessary to understand it. Below are 4 tables containing the most relevant variables in each dataset:

Listings Table	**Column**	**Description**
Most relevant variables	Property ID	ID of listing.
	Property Type	Type of property the listing is contained in.
	Listing Type	Entire home, Private room or shared room.
	Created Date	Date in which the listing was registered in Airbnb.
	City	City to which the listing belongs. Contains many missing values.
	Average Daily Rate (USD)	Daily fee charged by the landlord for the accommodation rental.
	Annual Revenue LTM (USD)	Revenue generated in the last 12 months.
	Occupancy Rate LTM	Percentage of days occupied in the last 12 months.
	Number of Bookings LTM	Number of total bookings in the last 12 months.
	Number of Reviews	Total number of reviews.
	Bedrooms	Number of bedrooms.
	Max Guests	Number of maximum guests allowed.
	Listing URL	Listing's Airbnb page.
	Latitude	GPS coordinate.
	Longitude	GPS coordinate.
	Overall Rating	Average rating attributed by users.
TOTAL VARIABLES	**52**	
TOTAL OBSERVATIONS	**112 609**	

Daily Table	**Column**	**Description**
Most relevant variables	Property ID	ID of listing.
	Date	Date of event.
	Status	R: Reserved, B: Blocked, A: Awaiting
	Price (USD)	Price to pay for the booking.
TOTAL VARIABLES	**8**	
TOTAL OBSERVATIONS	**53 405 116**	

Monthly Table	**Column**	**Description**
Most relevant variables	Property ID	ID of listing.

	Reporting Month	Month of report.
	Occupancy Rate	Percentage of days the listing was occupied for the given month.
	Revenue (USD)	Amount of revenue generated in the given month.
	ADR (USD)	Average Daily Rate in the given month.
	Number of Reservations	Total amount of reservations.
	Available Days	Sum of days without bookings.
TOTAL VARIABLES	**26**	
TOTAL OBSERVATIONS	**1 738 066**	

Reviews Table	**Column**	**Description**
Most relevant variables	Property ID	ID of listing.
	Review Date	Date the review was posted.
	Review Text	Text content of the review.
	User ID	ID of user.
	Member Since	Sign up date.
	Country	Country of origin.
	City	City of origin.
	Profile URL	User's Airbnb page
TOTAL VARIABLES	**17**	
TOTAL OBSERVATIONS	**1 192 919**	

Table 9: Airbnb metadata. Self-Authorship.

The distribution of each variable's table was plotted using Python's libraries Pandas and Matplotlib. For the sake of redundancy, such plots will not be depicted in this document.

From this initial analysis, it is possible to assess the date range of the data:

Table	**Oldest date**	**Most recent date**
Listings	December 2008	February 2018
Daily Bookings	September 2014	December 2017
Monthly Bookings	September 2014	December 2017
Reviews	October 2009	February 2018

Table 10: Date ranges for each Airbnb table

Throughout the analysis of outliers, 25% of the properties had a summed revenue throughout the previous 12 months of over $9 200, with the max value as $293 875, which although it is not possible to assess whether this value is correct, it will be considered an outlier. Other properties registered particularly high Average Daily Rates, going as high as $2 952, which was also considered as an outlier. Although, since these variables from this table will not be used, but rather from the Monthly table, these values will not be filtered out.

In the Daily Bookings table, there are also some outliers. In the variable price, some bookings can be priced as high as $540 954, which will be considered an outlier.

Regarding the Monthly Bookings table, some listings generated in revenue values over $10 000 in a single month, reaching a maximum of $80 853, and Average Daily Rates going as high as $5 775. Outliers from these variables will also be removed.

In the Reviews table no outliers were detected when analyzing interval and class type data. Although, a different problem arose: Data regarding customer profile is not standardized. In other words, the variables “City”, “Country”, “Description”, “State”, “School” and “Work” are displayed in an unstructured manner, e.g., in the variable “Country” one could see as values “I am from Portugal”, “Portuguese” or “I live in a city close to Lisbon” (illustrative examples only), even though it is clear all these three observations would have a common country of origin: Portugal. In order to parse this data from the text entries a Python script was developed using the libraries Pandas and Geograpy. Geograpy is a library developed using the NLTK library (Natural Language Processing Toolkit) with the goal of detecting locations and languages from a text and convert it to a dictionary containing the data encountered. If the algorithm detects a city in the text, it will return the city, region and the country associated to it (as a dictionary or list of dictionaries when more than one city/region/country is detected).

The outline of the processes for the parsing was developed as follows:

1. Read CSV files using Pandas;
2. Standardize text by upper casing text;
3. Create new tables with the unique values in the “City” and “Country” variables to process;
4. Use Geograpy to parse Countries each unique value;
5. Link tables with unique values to the original ones using variables “City” and “Country”, add the newly parsed countries in two different columns;
6. Create a new column using the two new variables, giving priority to the parsed countries from the variable “Country” (using parsed country from variable “City” when the parsing from the variable “Country” was unsuccessful);
7. Correct wrongly parsed countries from the final parsing;
8. Return table;

The success rate of the parsing is show after parsing:

```
Successfully parsed origins:  1181185
Unsuccessfully parsed origins:  23017
Total origins processed:  1204202
Done!
```

Figure 9: Parsing and Standardization of Airbnb users' countries of origin.

The success rate from overall process is 98%. The frequency distribution from the country of origin variable is available in the Analysis section. Although, out of this 98% of successfully parsed countries there is still a small margin of wrongly assigned countries. As the number of these occurrences is residual, so they weren’t corrected as to prevent overfitting when applying it to other datasets.

3.3.1. Outlier Removal

Given the data exploration above described, outliers will now be filtered. As Revenue is directly related to Average Daily Rate and Occupancy Rate in the monthly table, outliers in the variable “Revenue” will not be filtered out, but rather only outliers in the variable “ADR (USD)”. So, the variable “Price (USD)” from the Daily Bookings table and the variable “ADR (USD)” from the Monthly Bookings table will be filtered out.

The Interquartile Range method (IQR) will be used to remove the extreme values in the table “Daily”, column “Price (USD)”, as this method is an objective and reliable method to summarize data variability

and detect outliers in the dataset, thus removing subjectivity issues in situations where the problem of setting a threshold to detect outliers is unclear (Sullivan & LaMorte, 2016). It can be explained systematically:

1. Set IQR = Q3 – Q1
2. If a data point is below Q1 – 1.5×IQR or above Q3 + 1.5×IQR, it is viewed as being too far from the central values to be reasonable.

Additionally, in the table "Monthly" the variable "ADR (USD)" will have a maximum cutoff value of 2000$. In the "Properties" table, the variable "Annual Revenue LTM (USD)" will have a maximum cutoff value of 100000$. Additionally, the "Reviews" table had both duplicate observations and wrong datatypes (e.g., string type data in Longitude and/or Latitude variables – which should be of type float).

Below is presented the final result of the outlier removal step. The amount of outliers removed in "properties", "monthly" and "reviews" datasets is residual, in the case of the "daily" data set the percentage of removed outliers is approximately 8%.

```
Number of observations filtered:
properties :  65
daily :  4438317
monthly :  52
reviews :  97
```

Figure 10: Outlier removal results.

3.3.2. Missing Values

There are some columns with high percentage of missing values, one must consider whether to use them or not and how to fill the missing values in the variables that will be used.

There are many approaches possible for missing data imputation. Given that this analysis focuses on numerical data, using countries of origin as the only class type data (and will not require any type of imputation), the most straightforward approaches are either dropping the rows with missing values, drop the columns with missing values, use the average, median or mode, or other less intuitive processes such as K-Nearest Neighbors imputation (Gelman & Hill, 2006). In this situation, the most appropriate approaches is to use the median for columns with a low percentage of missing values and drop the ones with high percentage of missing values.

3.3.3. Correlation Analysis

It is important to avoid using correlated variables in clustering processes, as it would attribute extra weight to the attribute related with this correlation. Different methods will be used for both value and geographic clustering.

The correlation matrix is presented below:

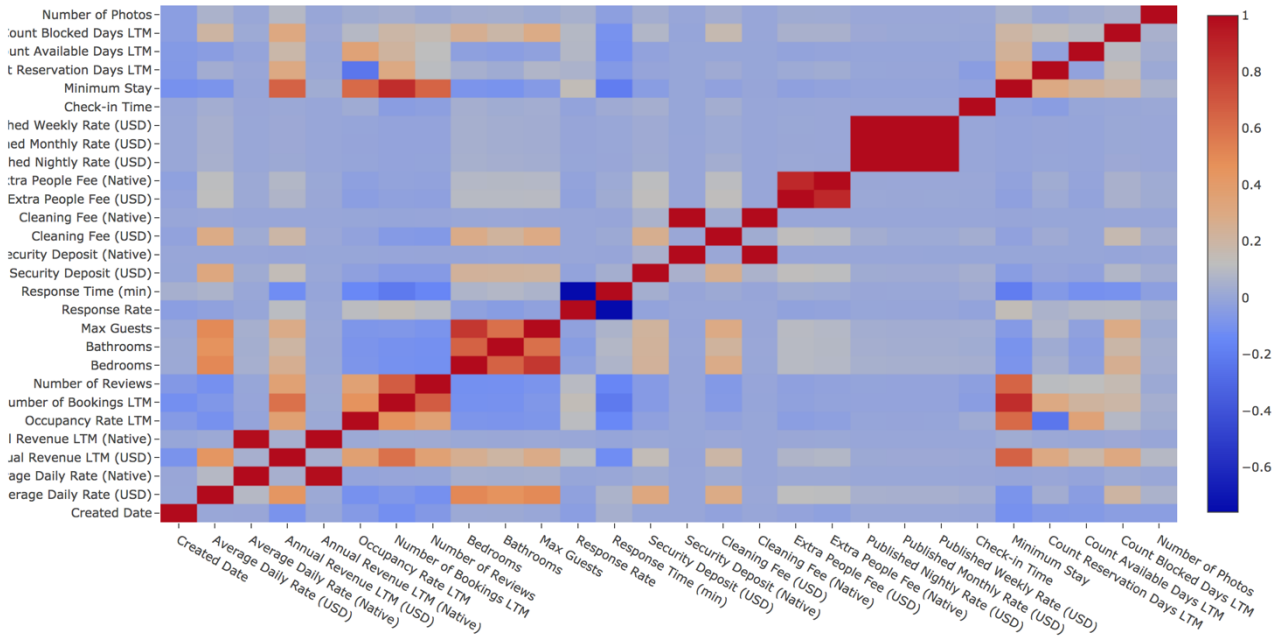
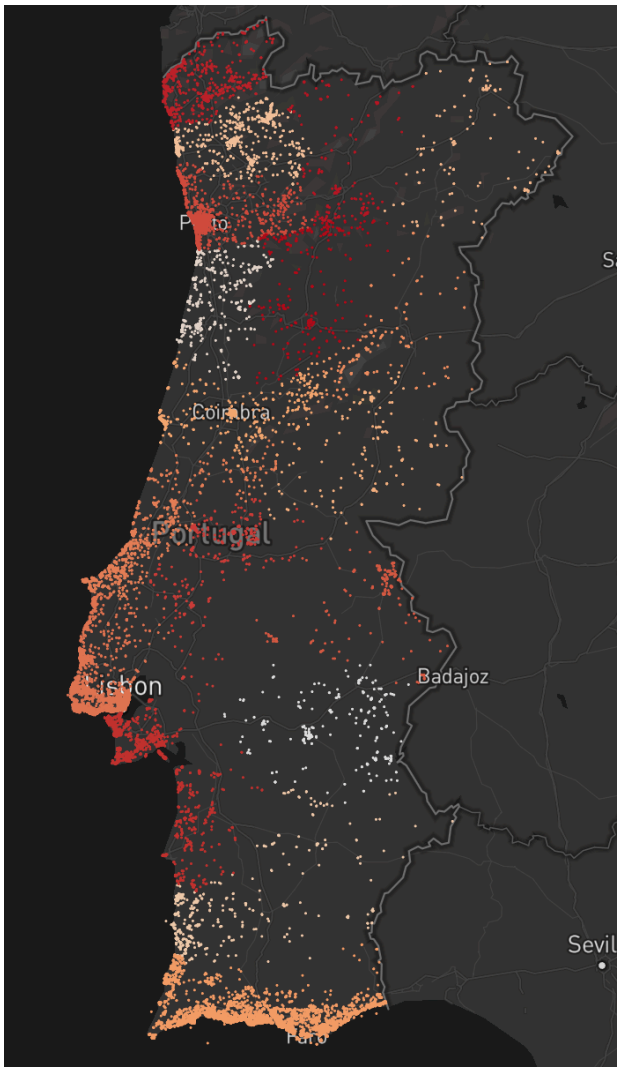


Figure 11: Pearson Correlation Matrix.

3.3.4. Geographic Clustering

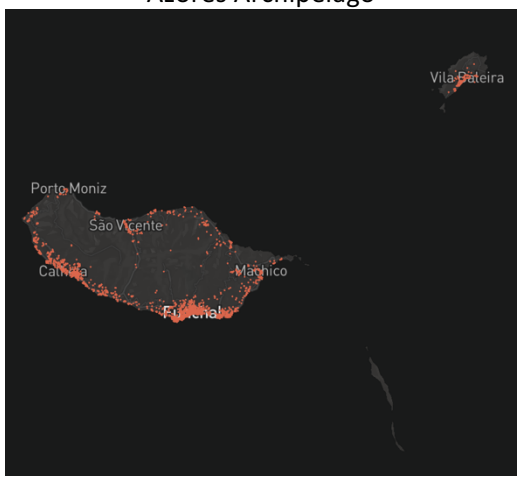
In order to have meaningful cluster boundaries, listings' clusters were made geographically using the borders of each district in Portugal. To do this, only the variables Latitude and Longitude were used. A shapefile was used (GADM, n.d.) to determine to which district each listing belonged to. Afterwards, listings with inaccurate location (i.e., some listings were located in the sea or rivers) that were not included in any district were classified using the K-nearest neighbors algorithm (with K=3). This algorithm, although simple to implement, was used because it provided the desired results.

Below is presented the characteristics of the defined regions:



Portugal (continent)

Azores Archipelago



Madeira Archipelago

Figure 12: Airbnb Listings' distribution after district labelling.

3.3.5. Value Clustering

We will use all variables that relate to the listing's value as a tourism hosting, which in this case would be the Annual Revenue, Average Daily Rate, Occupancy Rate and Number of Bookings. The reason we did not pick the Number of Reviews for a listing is because it is highly correlated with the number of bookings. Hence, in order to avoid a bias in the clustering process, this variable was discarded. Additionally, we also filtered out listings whose annual revenues were clear outliers (>$10 million) and also listings whose number of reservations was lower than 5 (as these will correspond either to overly recent listings, or inactive listings).

The elbow method will be applied to determine the number of K:

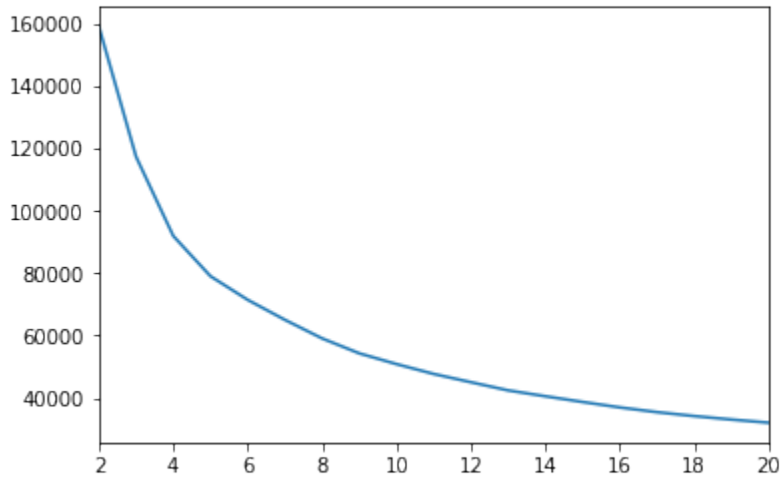


Figure 13: K-means inertia analysis for each number of clusters.

Along the increase of K, the decrease rate of inertia becomes lower when K=5. Considering the classification intuition behind value clustering (distinguish between high, medium-high, medium, medium-low and low value listings) a 5 cluster solution is preferred.

The results of the value clustering are shown in the table below with the variables used for the process and will be discussed in the analysis section.

val_cluster	frequency	Occupancy Rate	ADR (USD)	Number of Reservations	Revenue (USD)	Reservation Days	Bedrooms
0	23006	0.202	76.188	1.401	419.922	5.32	1.706
1	9343	0.655	76.409	5.716	1398.588	18.289	1.488
2	1521	0.344	266.883	2.09	3401.368	9.248	4.67
3	7345	0.287	137.885	1.618	1425.838	7.677	2.972
4	20844	0.401	70.851	2.272	713.346	10.272	1.527

Figure 14: Value clustering results

4. RESULTS AND EXPLORATORY ANALYSES

4.1. SOCIAL MEDIA CRAWLER

A diagram demonstrating the program's crawling process is available below:

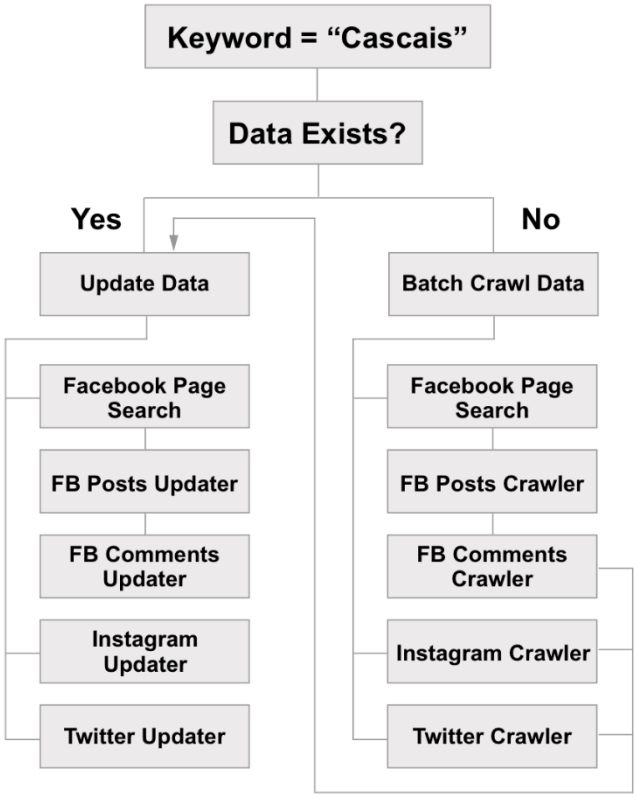


Figure 15: Diagram demonstrating the Social Media Crawling process.

The project's structure is presented as follows:

```
.
├── LICENSE
├── README.md
├── Web_Crawler
|   ├── InstagramAPI
|   ├── __init__.py
|   ├── _data
|   |   ├── carcavelos_facebook_comments.csv
|   |   ├── carcavelos_facebook_statuses.csv
|   |   ├── carcavelos_instagram_img
|   |   ├── carcavelos_instagram_posts.csv
|   |   ├── carcavelos_tweets.csv
|   |   ├── cascais_facebook_comments.csv
|   |   ├── cascais_facebook_statuses.csv
|   |   ├── cascais_instagram_img
|   |   ├── cascais_instagram_posts.csv
|   |   ├── cascais_tweets.csv
|   |   ├── estoril_facebook_comments.csv
|   |   ├── estoril_facebook_statuses.csv
|   |   ├── estoril_instagram_posts.csv
|   |   └── estoril_tweets.csv
|   ├── _master.py
|   ├── _master_batch_crawler.py
|   ├── _others
|   |   ├── data_viz_with_sample_data
|   |   └── sentiment_analysis
|   |       ├── cascais_tweets.csv
|   |       ├── sentiment_analysis_lexicon_nonNLTK
|   |       ├── sentiment_analysis_part2.py
|   |       ├── sentiment_analysis_test.csv
|   |       └── sentiment_analysis_test.py
|   ├── facebook
|   ├── facebook_access_token.py
|   ├── facebook_comments.py
|   ├── facebook_posts.py
|   ├── facebook_search_results.py
|   ├── instagram_access.py
|   ├── instagram_hashtags.py
|   ├── instagram_image_downloader.py
|   ├── twitter_tweets.py
|   ├── update_fb_comments.py
|   ├── update_fb_posts.py
|   └── update_insta_hashtags.py
├── auto_updater_3kws.py
├── gui
|   └── FlaskApp
|       ├── __init__.py
|       ├── db_facebook.py
|       ├── db_instagram.py
|       ├── db_twitter.py
|       ├── static
|       ├── support
|       ├── templates
|       └── update_manager.py
├── requirements.txt
└── social_media_crawler.py
```

Figure 16: Project Structure.

4.1.1. Implementation

The project was implemented on an Azure Server (with an Ubuntu Operating System) for short periods of time, in which the updater ran continuously. The data is then saved as .csv files. The GUI can be used in a local machine to configure the program that is running in the remote server.

4.1.2. Screenshots

Screenshots of the program's terminal view for both batch crawling and data updating is available below:

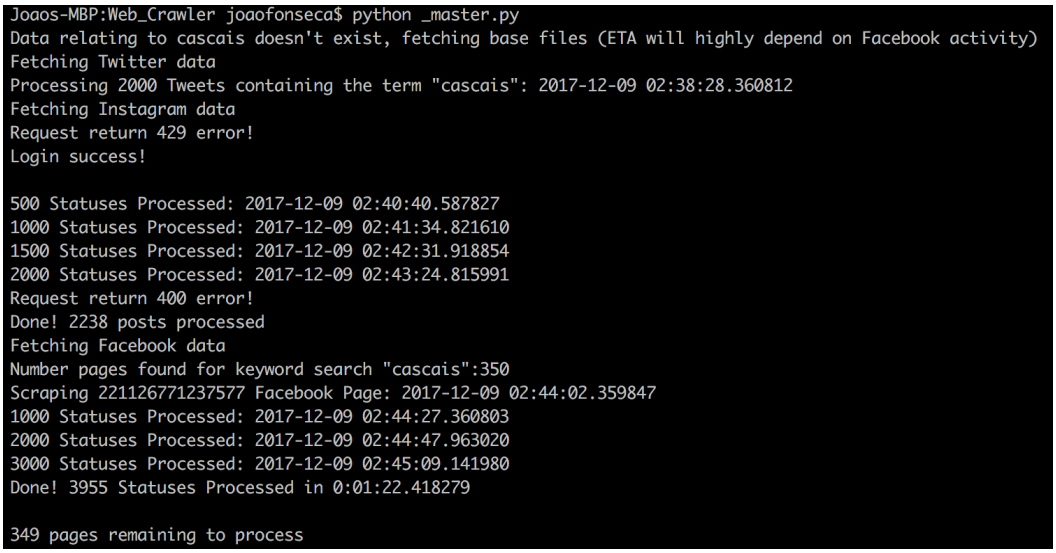


Figure 17: Social Media Crawler's Batch Crawling.

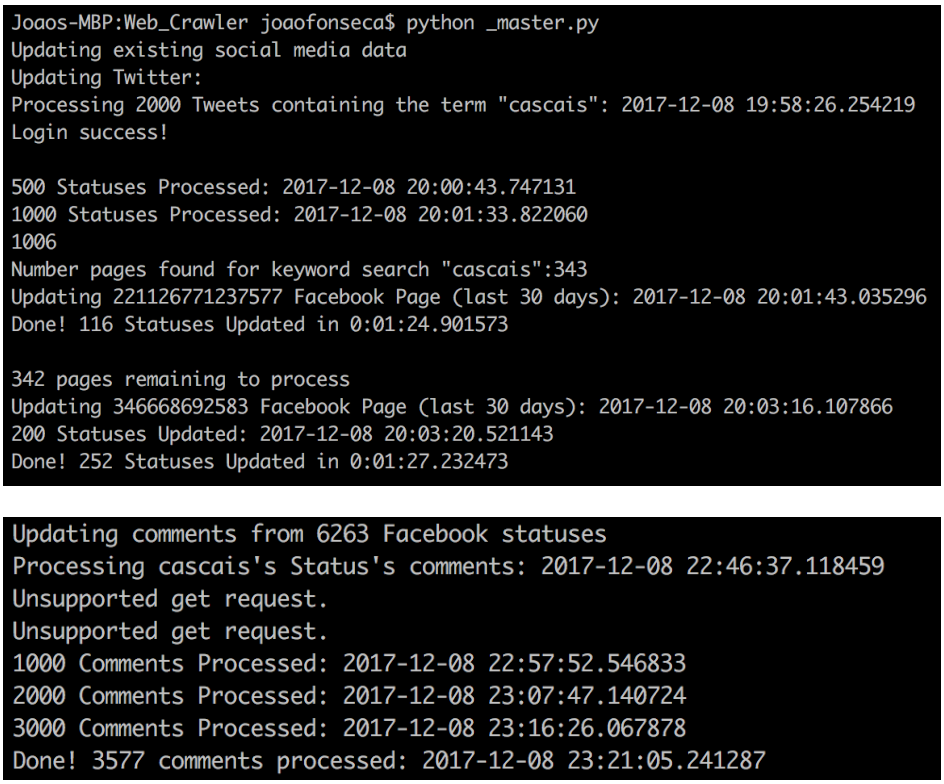


Figure 18: Updating Social Media datasets.

Given the size and complexity of the structure of the data (e.g., the commas in the posts' texts will make the table conversion in excel impossible), it cannot be directly accessible with regular spreadsheet tools such as excel, which would require previous conversion and querying (in order to reduce the amount of useless data for a given purpose passed on to the file).

Screenshots demonstrating the GUI are available below:

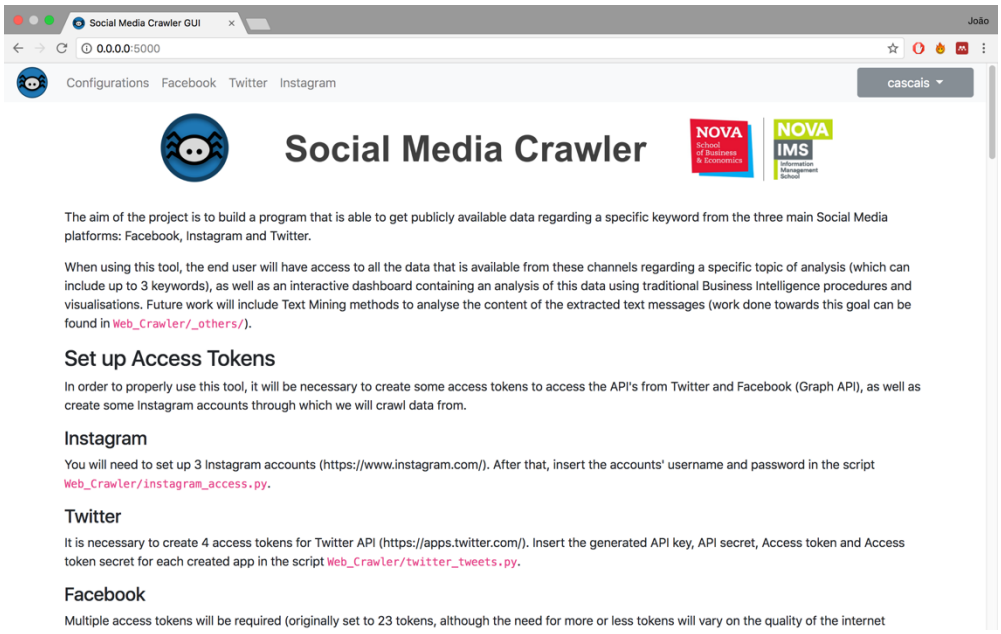


Figure 19: Social Media Crawler's Homepage.

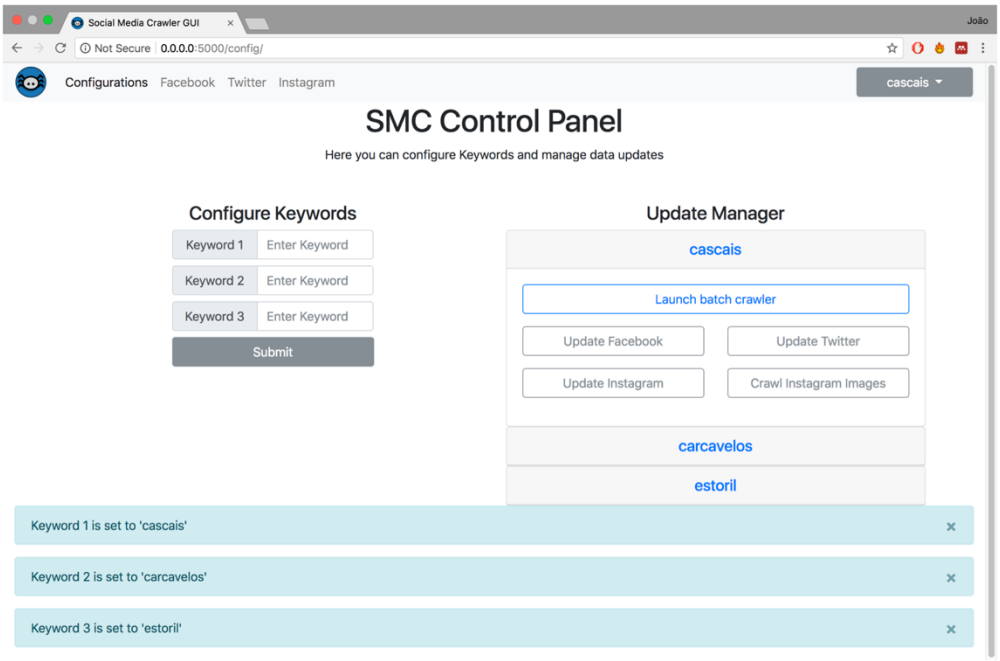


Figure 20: Social Media Crawler's Configurations panel.

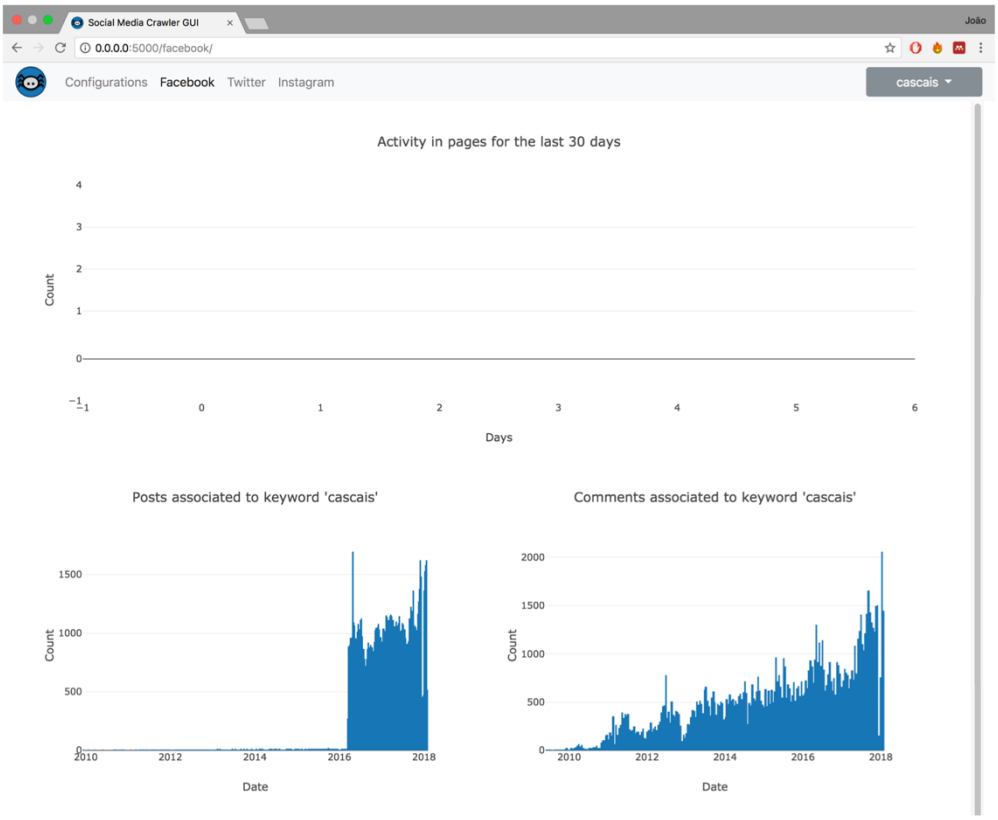


Figure 21: Facebook Dashboard. Own Authorship

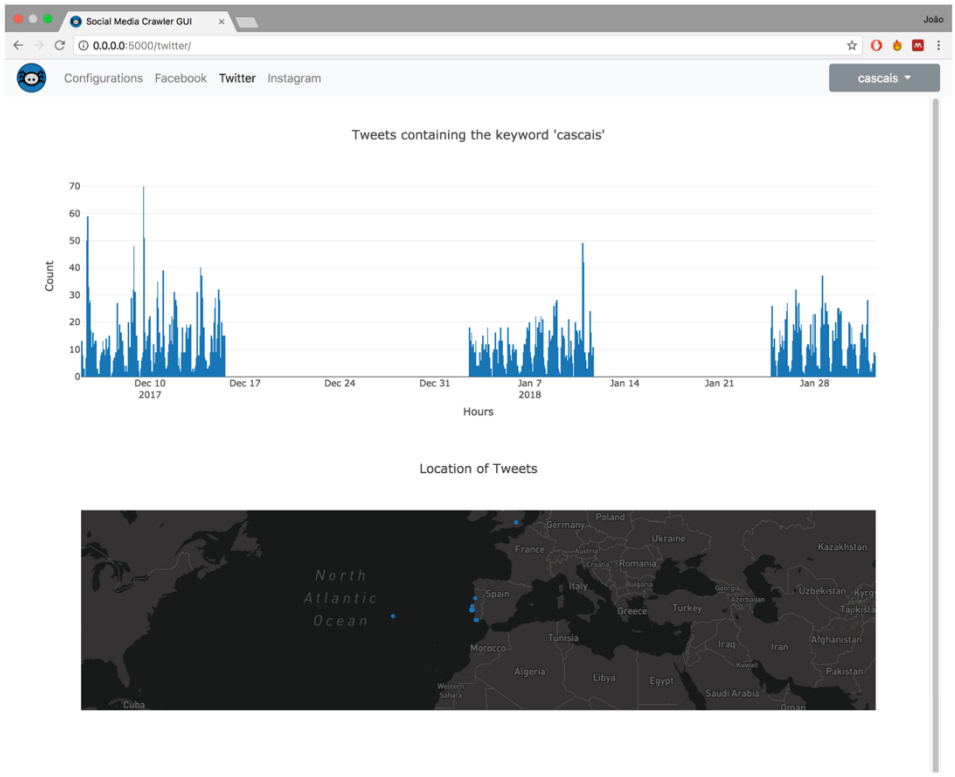


Figure 22: Twitter Dashboard.

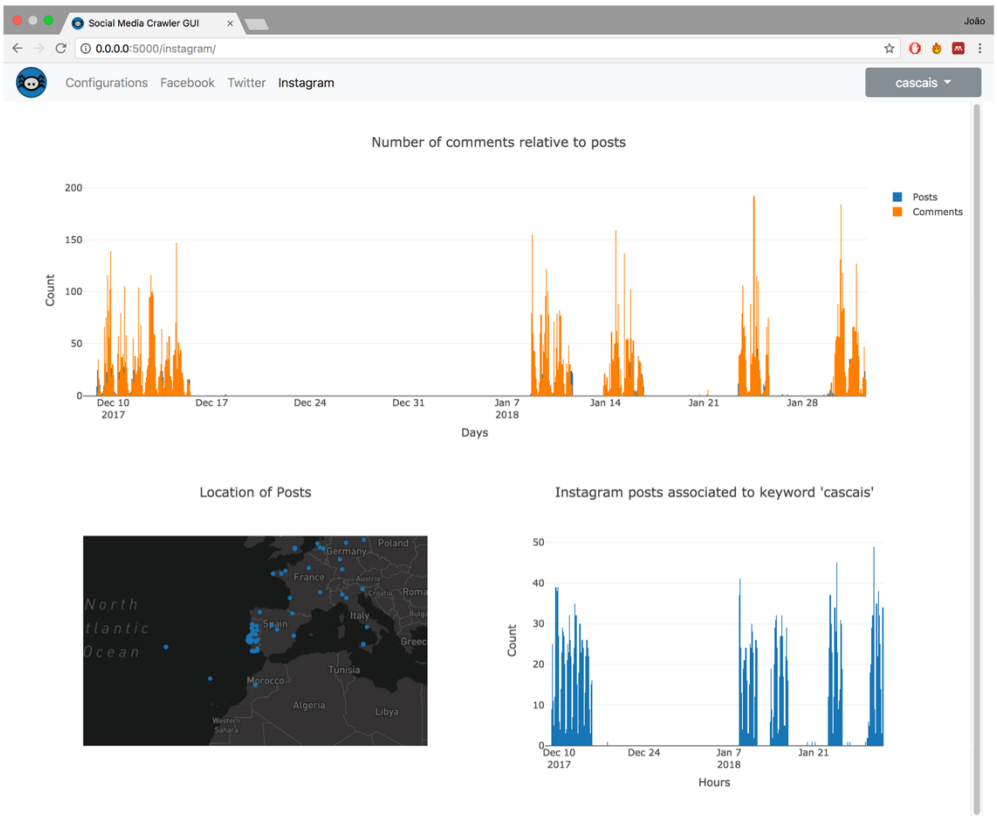


Figure 23: Instagram Dashboard.

4.2. TELECOM

4.2.1. Daily Presences of Foreign Visitors

The total presences per day for each tourist's country of origin in Portugal during August 2017 that connected to NOS' network is displayed below. Breaks are done between the highest number of visitors for each country of origin and other regions/countries of interest, with the remaining countries listed as 'others'. The majority of tourists visiting Portugal in August arrive predominantly in the second week of the month. Tourism reaches its peak until the 15th of August. After this period the number of tourists starts to fade out. As such, it is visible that the second week of August seems to be the most popular period for tourists to visit Portugal. It also clear that tourists coming using French SIM cards are among the ones presenting the highest seasonality.

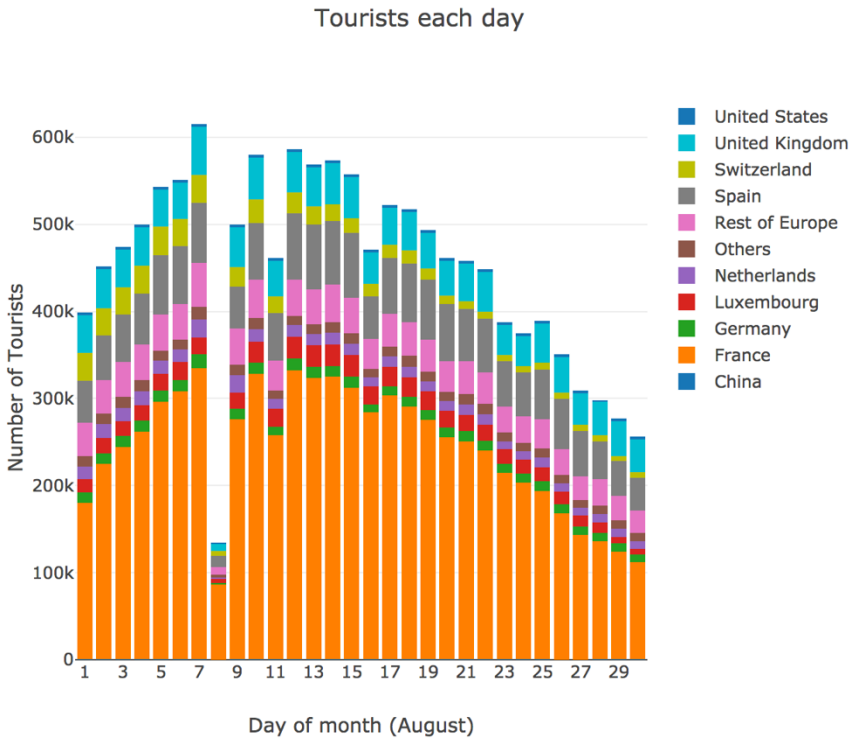


Figure 24: Tourists per day (August 2017).

The total number of tourists in August was counted for each district between August 1st 2017 and August 30th 2017. The shade of blue represents the number of visitors (darker shade = more tourists). Tourists visit mainly the Porto, Lisbon and Algarve regions. Northern districts that are normally assumed as a non-typical tourist destination are also well represented, suggesting the presence of Portuguese immigrant visitors.

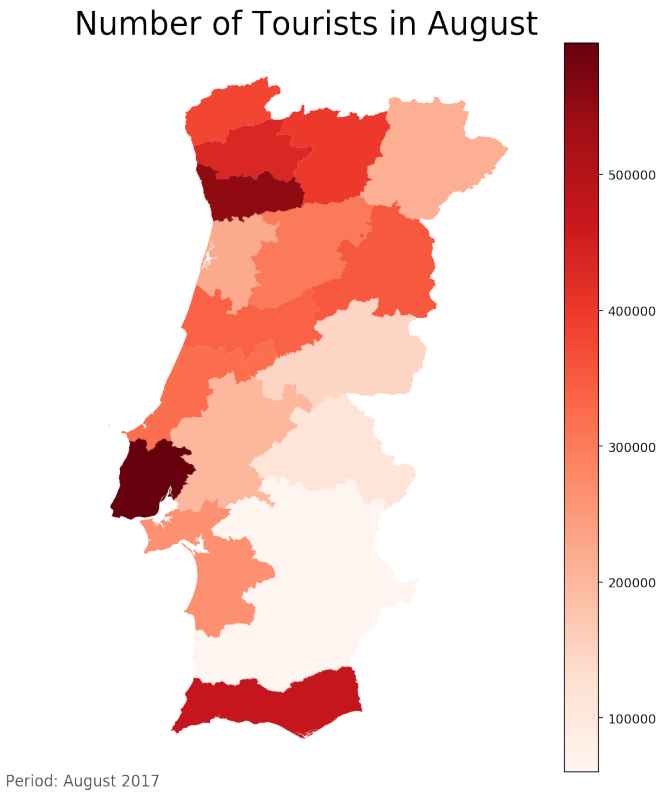


Figure 25: Tourists per district (August 2017)

The distribution of tourists' country of origin is shown below. The number of roamers coming from typical destinations for Portuguese immigrants are among the ones best represented (France, Switzerland, Luxembourg and Belgium), thus strengthening the hypothesis of the presence of Portuguese immigrants.

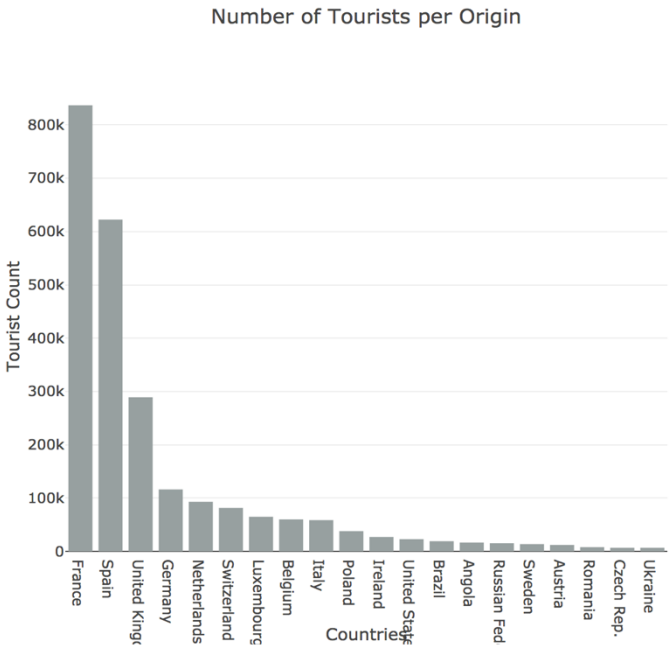


Figure 26: Tourists per Origin

The percentage of arrivals and departures per day of week was calculated. The percentage of visitors arriving on a Tuesday and leaving on a Wednesday is above the average distribution, which might be related plane tickets' prices.

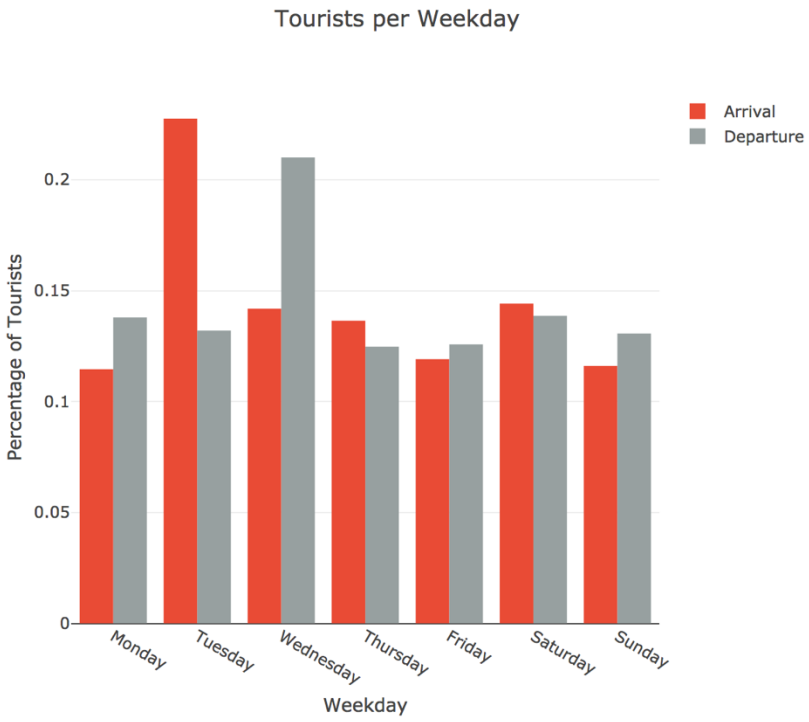


Figure 27: Tourists per Weekday

4.2.2. Duration of stay of foreign visitors

The estimates of how long foreign visitors stay in Portugal during the mentioned period, by nationality. Again, the presence of typical destinations for Portuguese immigrants are among the ones best represented. On the other hand, US visitors show one of the lowest average length of stay, followed by Italy, Spain and Germany.

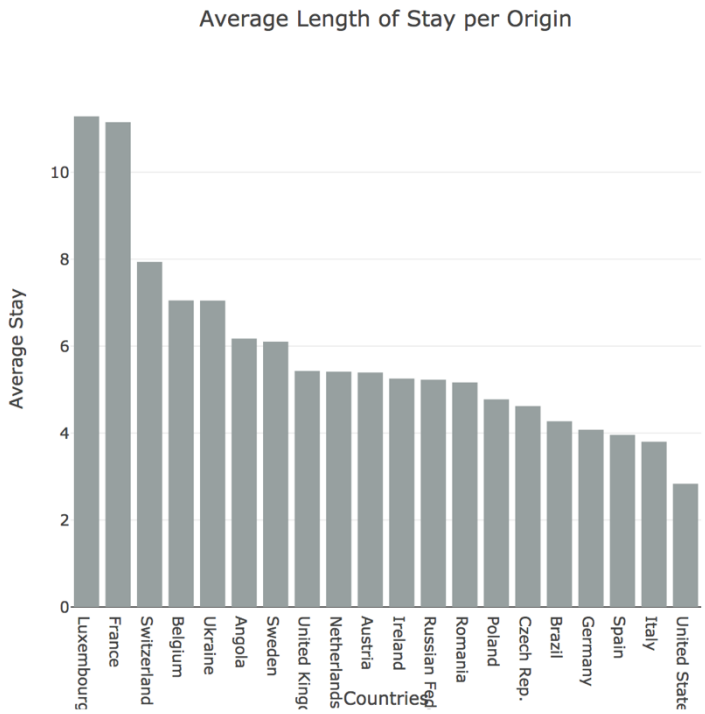


Figure 28: Average Length of Stay per Origin

The number of tourists grouped by days of stay is depicted below. The majority of tourists (52%) stay between 1 and 8 days in Portugal. A low percentage of tourists stay for longer than 2 weeks.

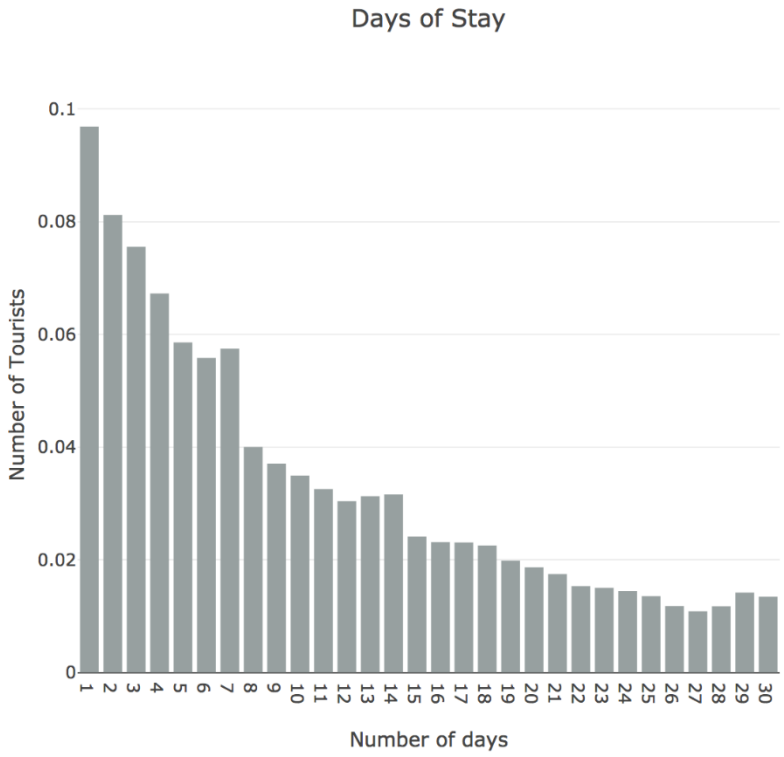
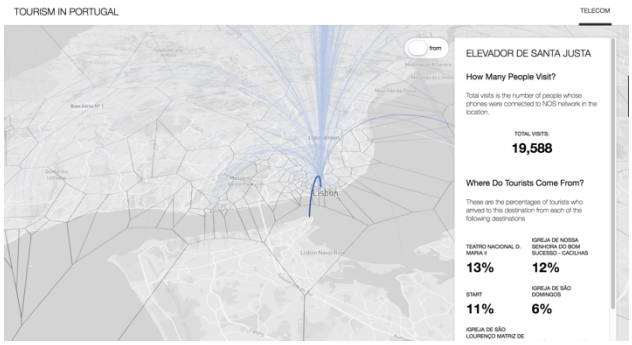
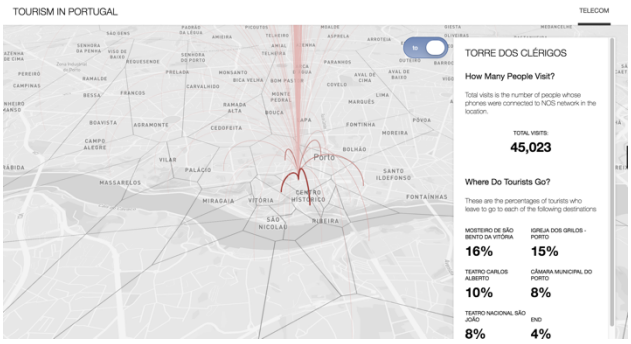


Figure 29: Percentage of Tourists per Number of Days of Stay

In order to allow the analysis of flows of roamers across Portugal (using the existing dataset), a visualization using Deck GL was developed. It allows the development of inflows and outflows visualizations for any Voronoi cell in Portugal.





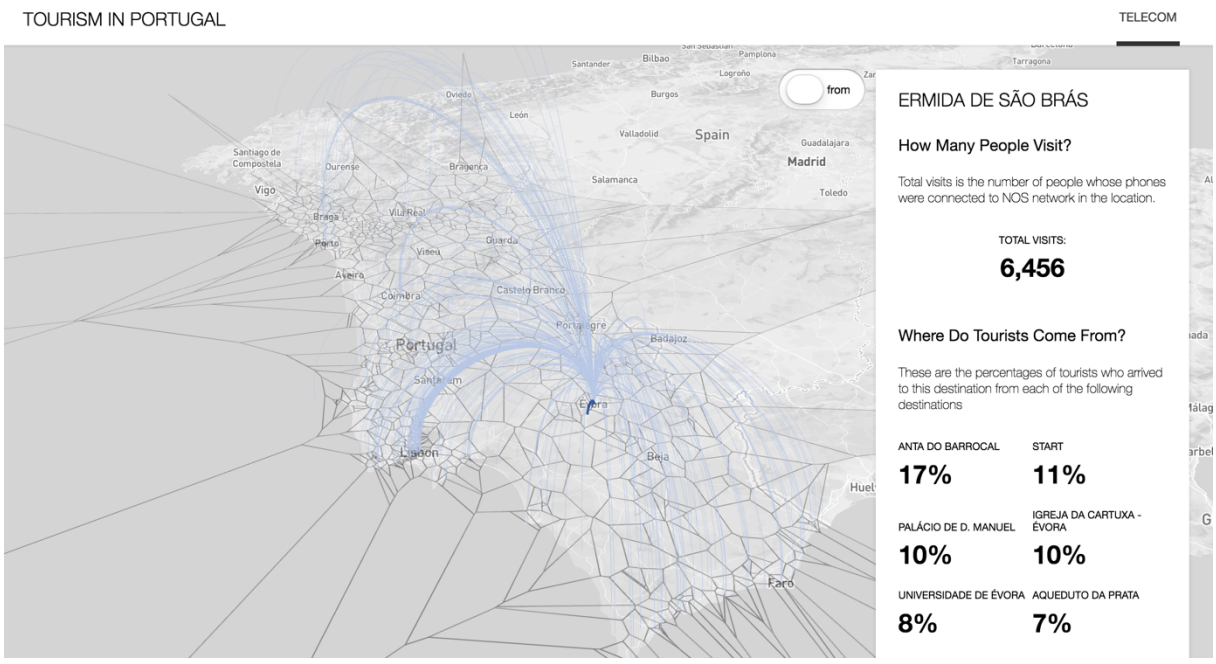
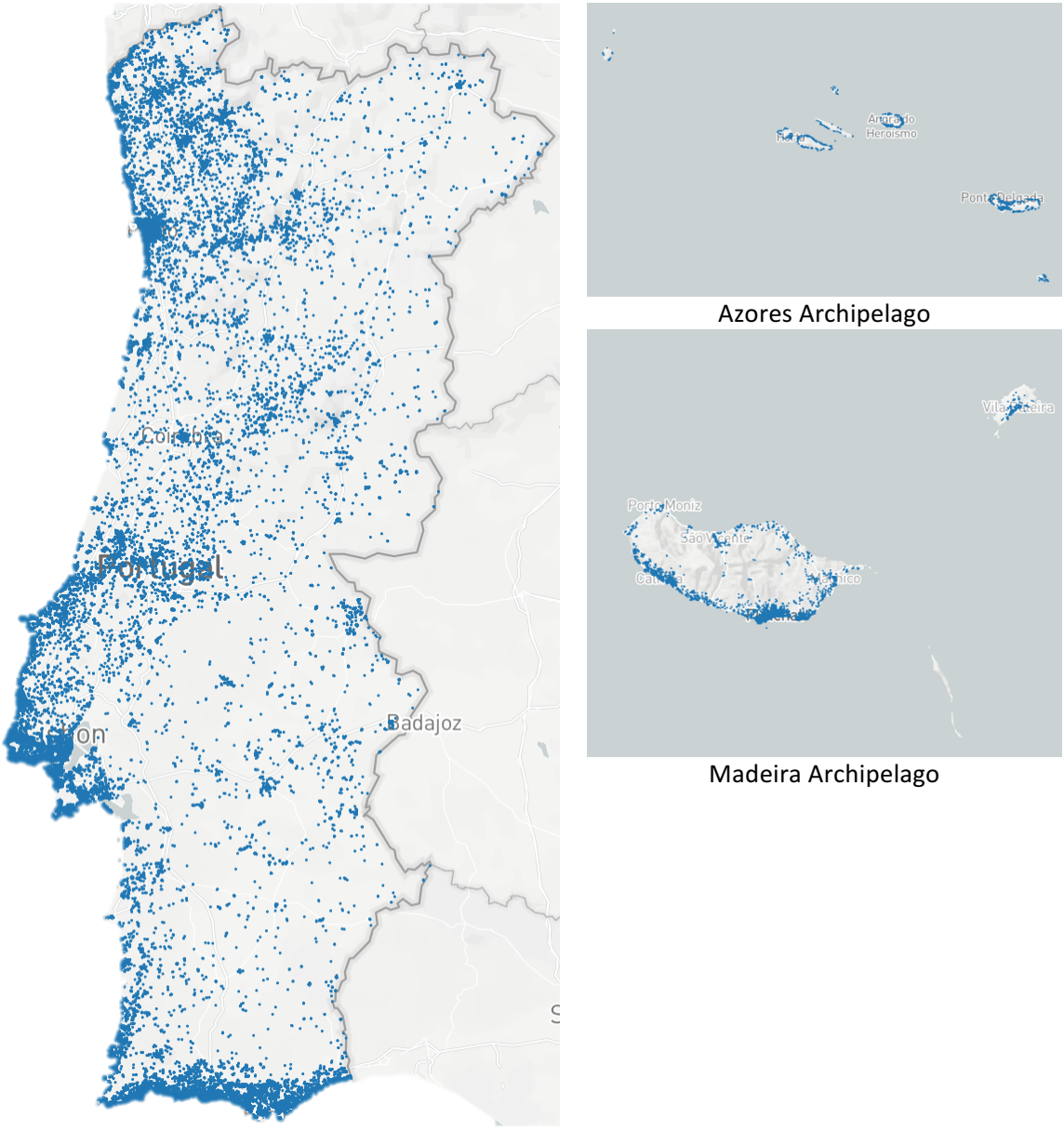


Figure 30: Screenshots of the Deck GL visualization developed

4.3. AIRBNB

4.3.1. Property Listings

Listings are mainly concentrated in in the coastal region of Portugal. The city with most property listings is Lisbon, with over 25 thousand listings. Porto comes second with about 10 thousand listings, where the remaining cities have less than half the number of listings existent in Porto. Although, it is important to mention that at this point there are many listings without a city associated to it (missing values). This challenge will be addressed in the geographic clustering process.



Azores Archipelago

Madeira Archipelago

Portugal (continent)

Figure 31: Airbnb Listings' distribution

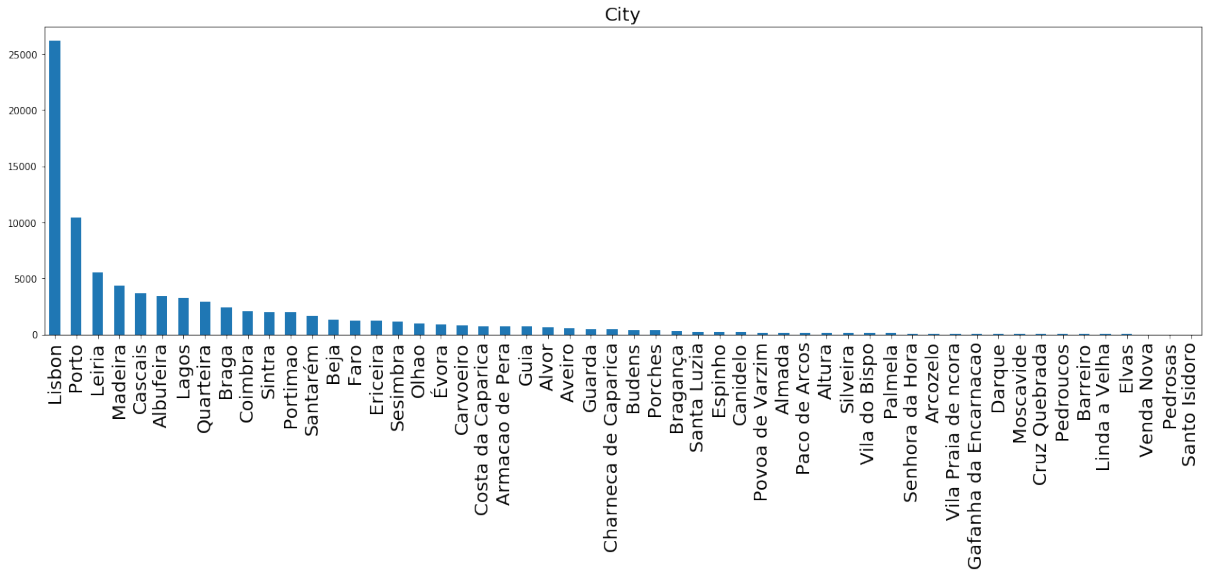


Figure 32: Number of listings per city

Most of these listings refer to entire home/apartments, as the number of private rooms represent less than half the number of the former type.

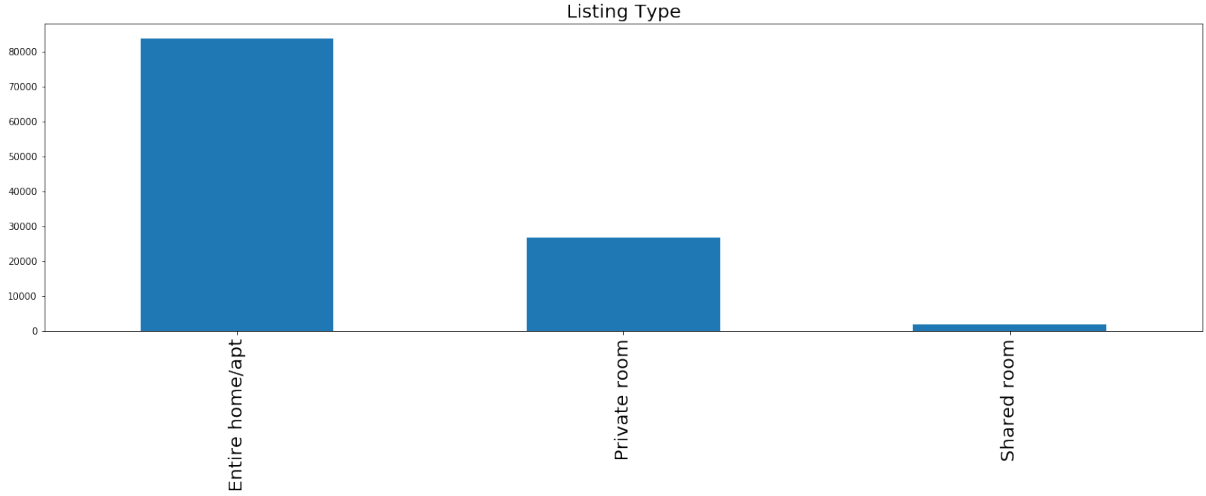


Figure 33: Type of listings

4.3.2. Daily Bookings

The daily bookings table contains information of daily booking activities for each listing. Although, aside from completed bookings, this table also includes data from blocked bookings, as well as requested bookings, which include the unanswered requests (or still awaiting an answer)..

Booking activity done in Airbnb grew exponentially over time. This might be because of two main factors: The growing popularity of Airbnb's platform, and the growing popularity of Portugal as touristic destination.

Booking activity can mean either a booking request (which means such request is waiting for approval or has been blocked), a cancelled booking, or a booking that was actually completed and went through

(which amounts to approximately 21% of the overall booking events). In the graph below is depicted the daily count of reservations completed between September 2014 and December 2017.

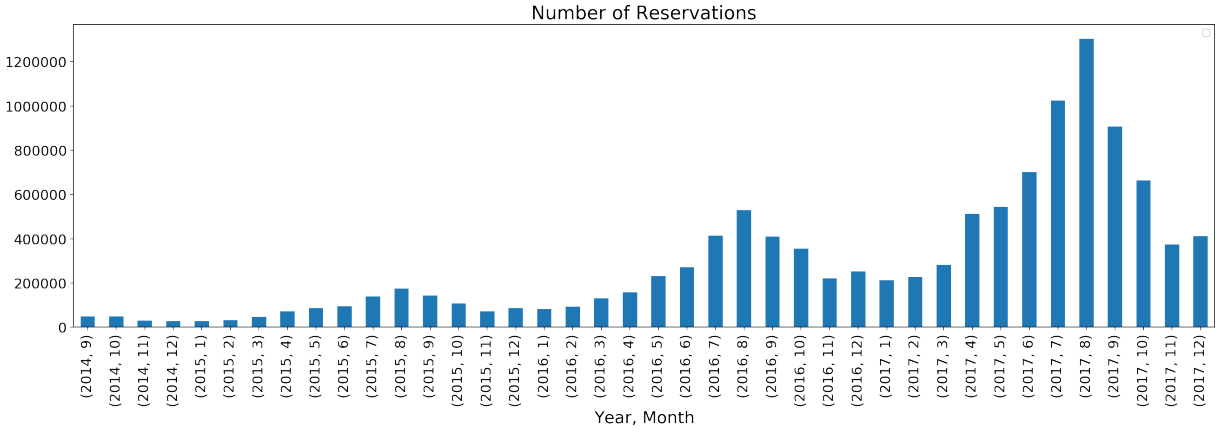


Figure 34: Reservations per month

4.3.3. Monthly Bookings

There are suggestions of inactive listings or listings with no demand, as there are many listings do not receive any reservation in several months:

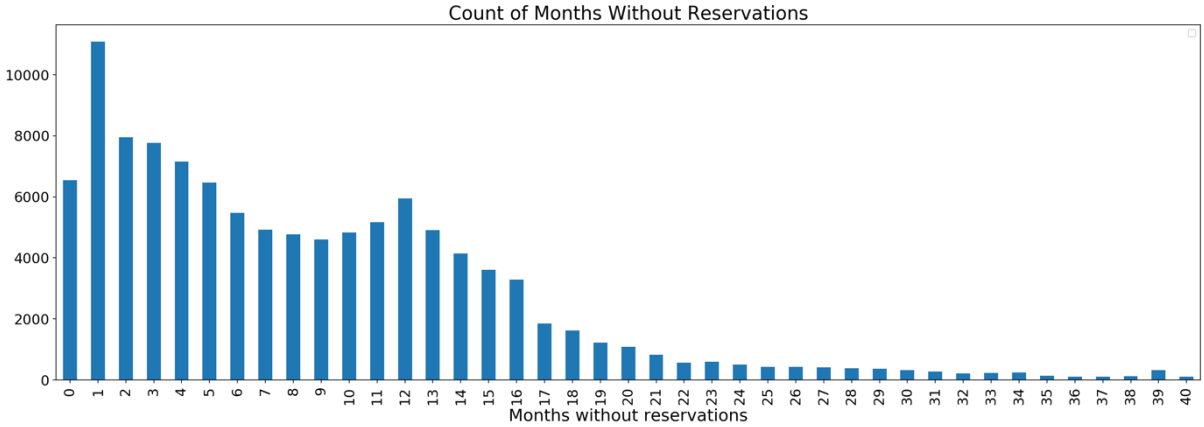


Figure 35: Number of Listings per number of months without reservations

The causes for this factor must be looked into more thoroughly, as it is important to assess whether there are specificities in these listings that cause a low demand for these offers, or on the other hand there might be too much supply in certain regions.

4.3.4. Listing Reviews

To extract information from this dataset some basic text mining techniques were required, since none of the data regarding user information is standardized. Although we have some details regarding the user's profile, namely its first name, country of origin, state (if applicable), city of origin, a brief user description, last attended teaching institution and occupation, none of it can be directly used for analysis without prior pre-processing (for the goal of this analysis, the priority was given to the parsing of the country of origin out of the non-standardized text fields: Country and City). After the parsing of the countries of origin, this is the top 50 countries of origin for Airbnb users in Portugal:

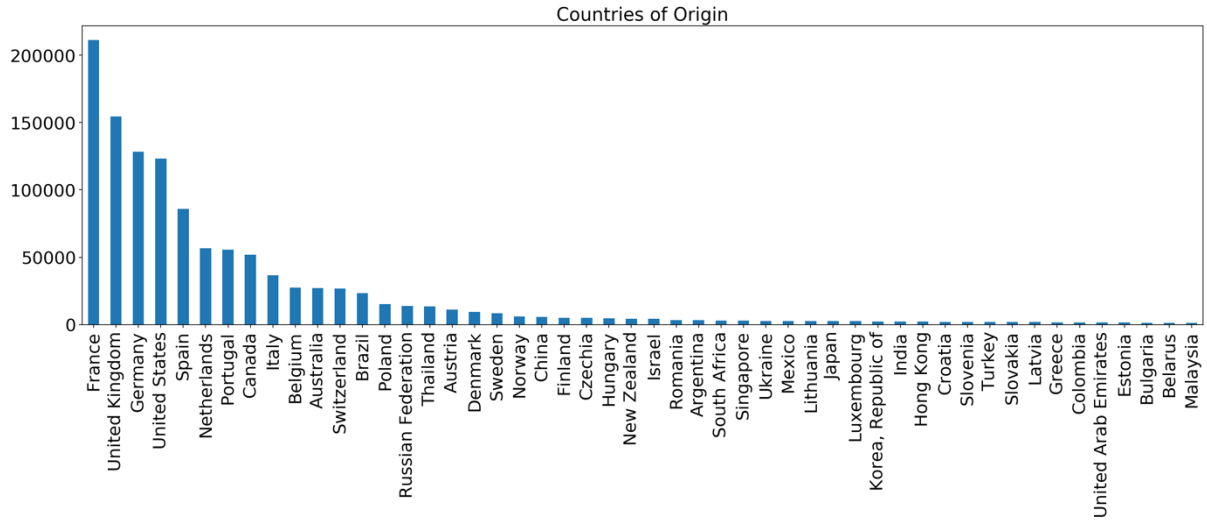


Figure 36: Countries of origin's distribution

It becomes clear that the number of French Airbnb users in Portugal is extremely high, which might be caused by two factors: 1) the seasonal immigrant flows from France into Portugal throughout the summer; 2) Airbnb's popularity in France, which according to Google trends it is in fact, the country in which Airbnb is most popular:

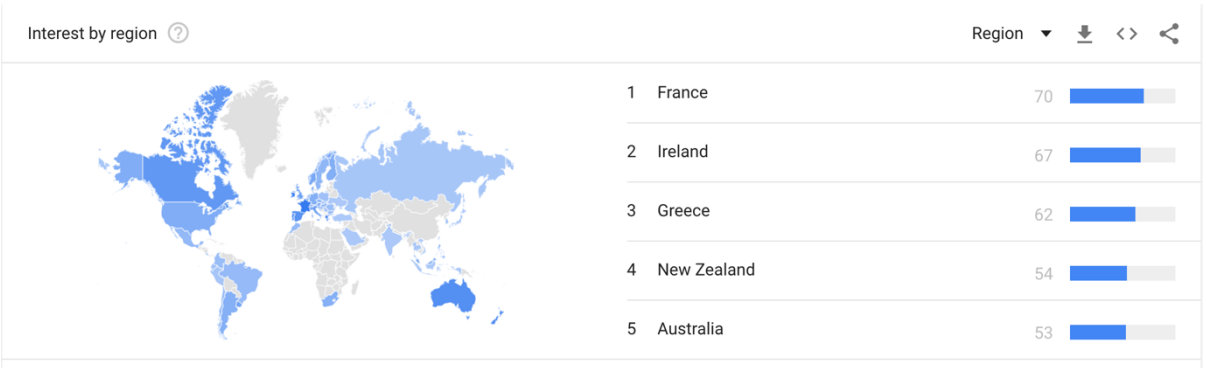


Figure 37: Google search popularity per country for the keyword "Airbnb". Source: Google trends.

4.3.5. Value Clustering Analysis

The clusters representing listings with the highest value are mostly located near the shore and large cities, whereas clusters with listings of lower value are relatively evenly distributed across the territory.

val_cluster	frequency	Occupancy Rate	ADR (USD)	Number of Reservations	Revenue (USD)	Reservation Days	Bedrooms
0	23006	0.202	76.188	1.401	419.922	5.32	1.706
1	9343	0.655	76.409	5.716	1398.588	18.289	1.488
2	1521	0.344	266.883	2.09	3401.368	9.248	4.67
3	7345	0.287	137.885	1.618	1425.838	7.677	2.972
4	20844	0.401	70.851	2.272	713.346	10.272	1.527

Figure 38: Value Clustering's results

4.3.6. Geographic Analysis

The number of reservations made each month are highly seasonal and has increased over time. However, this increase is may be be attributed to the rise in popularity of Airbnb, instead of Portugal as a touristic destination. The data analyzed refers solely to bookings in Airbnb's platform. As concluded previously, Lisbon yields a very high percentage of total bookings, whereas the second main destination corresponds to Greater Porto region and the third most booked region is Faro. Although, due to Faro's high touristic seasonality, in August 2017 Faro was the second most booked region in Portugal.

Finally, one can also conclude that the two regions with the overall lowest number of bookings are the districts in the interior of Portugal and the islands.

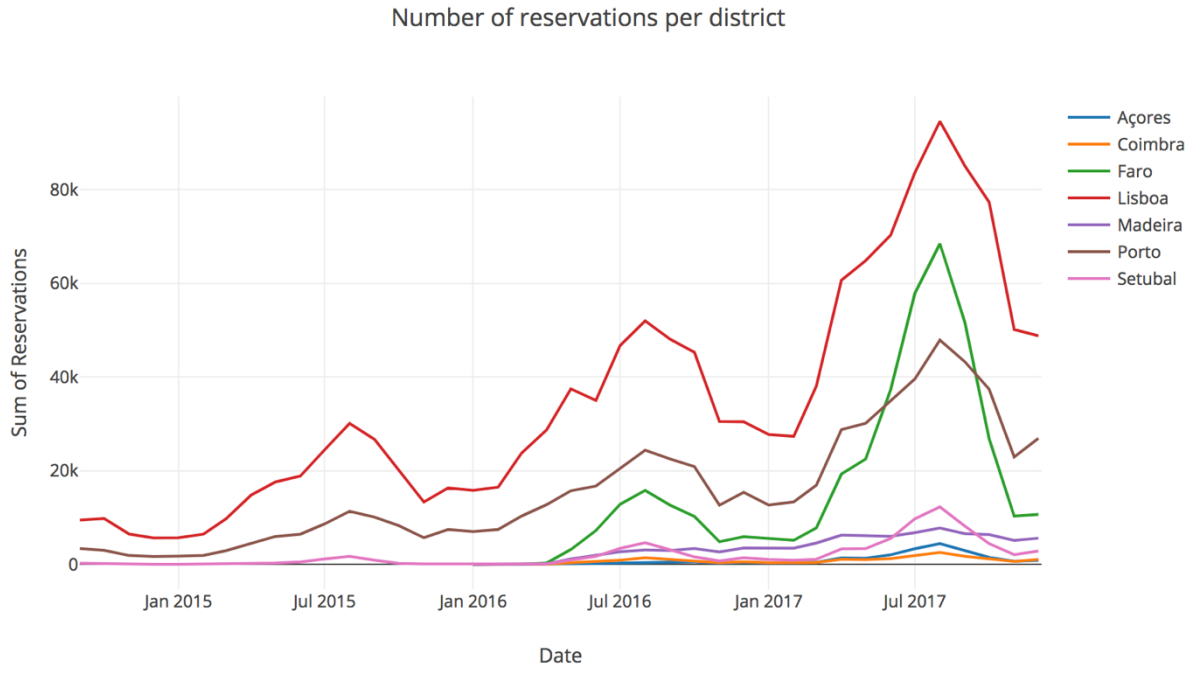


Figure 39: Number of Reservations per district

Although Porto district is well ranked in the number of bookings, it falls behind significantly on the sum of generated revenue, when considering the exponential growth of generated revenue by Lisbon and Faro. We can see that although Faro had a lower number of reservations, it is the region with most revenue generated in the high season (June to September).

This suggests that the value of listings in Porto is lower than expected. This can be caused by excessive supply and/or low demand. Although, given the above analysis, one could discard the second hypothesis. Additionally, it is also possible that bookings' Daily Rates are not yet well adjusted to the market.

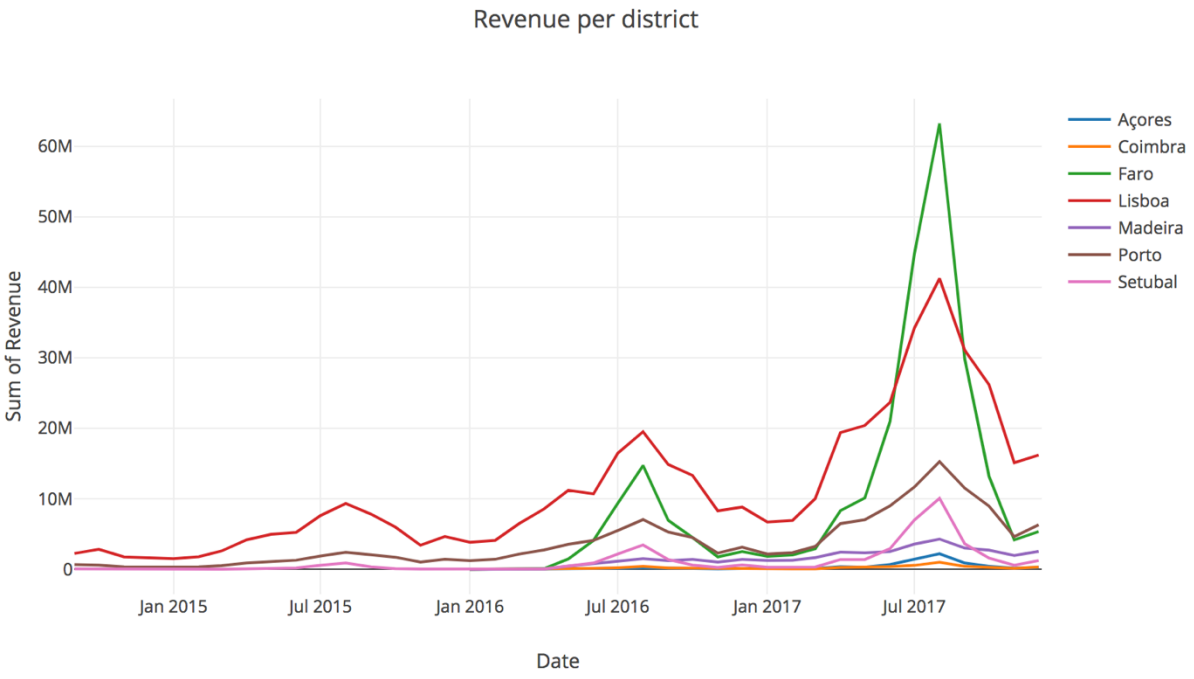


Figure 40: Revenue per district

Porto region yields one of the lowest revenue per reservation rates. On the same note, Lisbon has this same ratio as relatively average. Algarve region's previously perceived Airbnb value is confirmed, as it is the region with the highest revenue per reservation ratio since March 2016.

Being Algarve sought after as a summer touristic destination by northern European countries and English speaking countries with medium to high purchasing power, the price level in this region is expected to be higher, which will reflect on the accommodation's daily rates. Additionally, the accommodations existing in this area are typically villas, which are more expensive than entire apartments or rooms. Lastly, the existence of more villas when opposed to buildings (when compared to regions like Greater Lisbon and Greater Porto) can also lead to less accommodation supply in a region with high demand, which will drive daily rates up. Additionally, Setubal also demonstrates to be a region with one of the highest ratios, which can be related to its proximity to Lisbon and bathing regions.

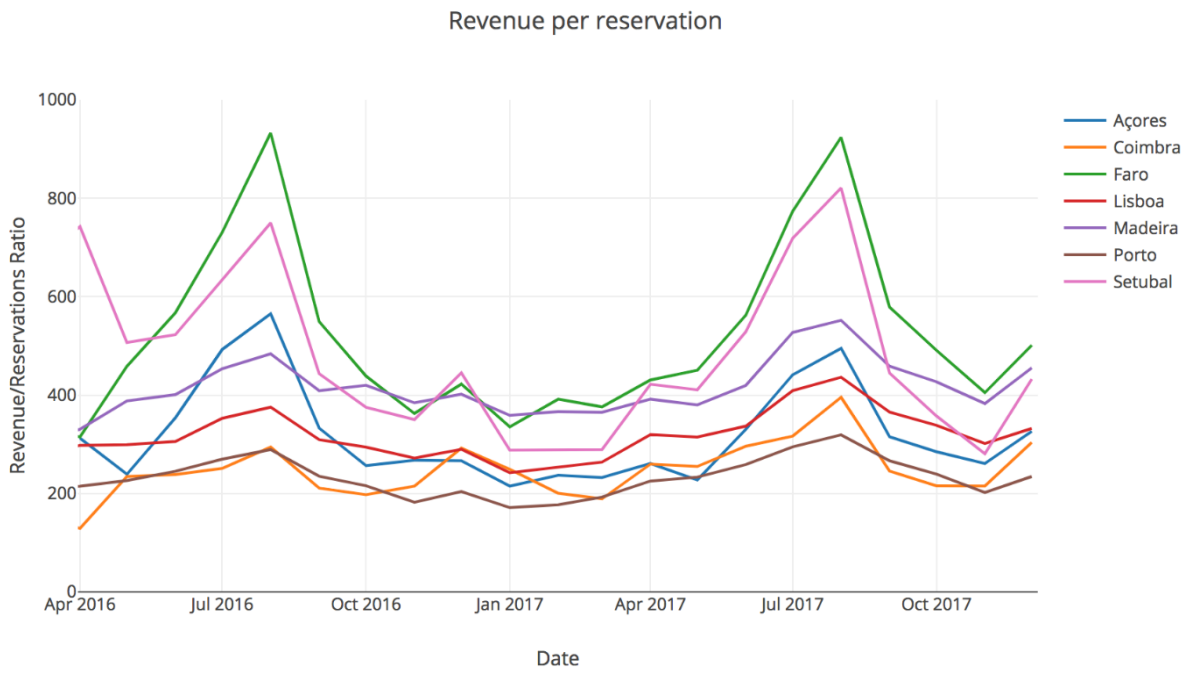


Figure 41: Revenue per reservation

Using the average occupancy rate for each district, one can now analyze the balance between Demand and supply of Airbnb bookings. The demand in Faro district in August last year represents one of the highest occupancy rates, alongside with Azores, Porto, Setubal and Lisbon.

On a different analysis, Porto and Lisbon regions did have a high amount of reservations throughout last year and are one of the regions with the highest occupancy rate throughout the year (i.e., lowest seasonality). On the other hand, Porto region's revenue per reservation and number of reservations is one of the lowest out of all districts, suggesting that either 1) this region has a higher Supply/Demand ratio when compared to the remaining regions, and/or 2) tourists to stay for shorter periods of time in this region.

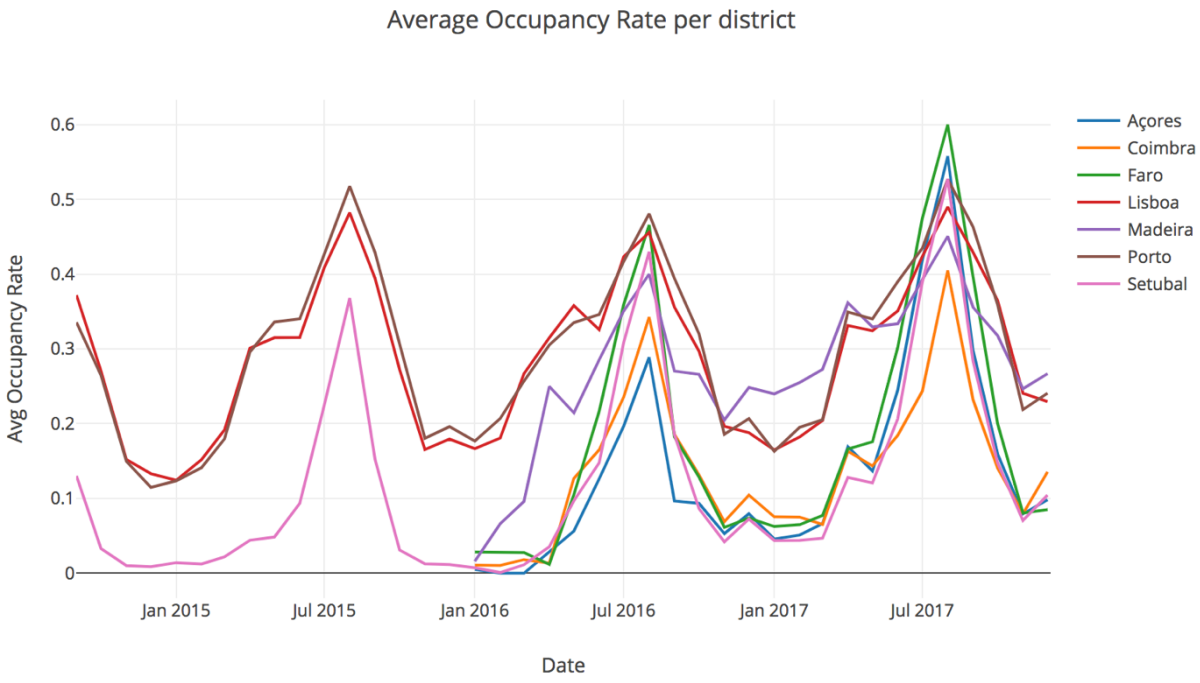


Figure 42: Average occupancy rate per district

The following graphic depicts the representation of tourists from each country of origin in each cluster. As the French are the ones using Airbnb the most in Portugal, they end up ranking either first or second in most regions. When filtering out France and "others" nationality, some differences become visible from cluster to cluster. Portuguese tourists take a significant share in the least popular regions regarding number of bookings, namely the districts of Bragança, Guarda, Portalegre, Viana do Castelo and Viseu.

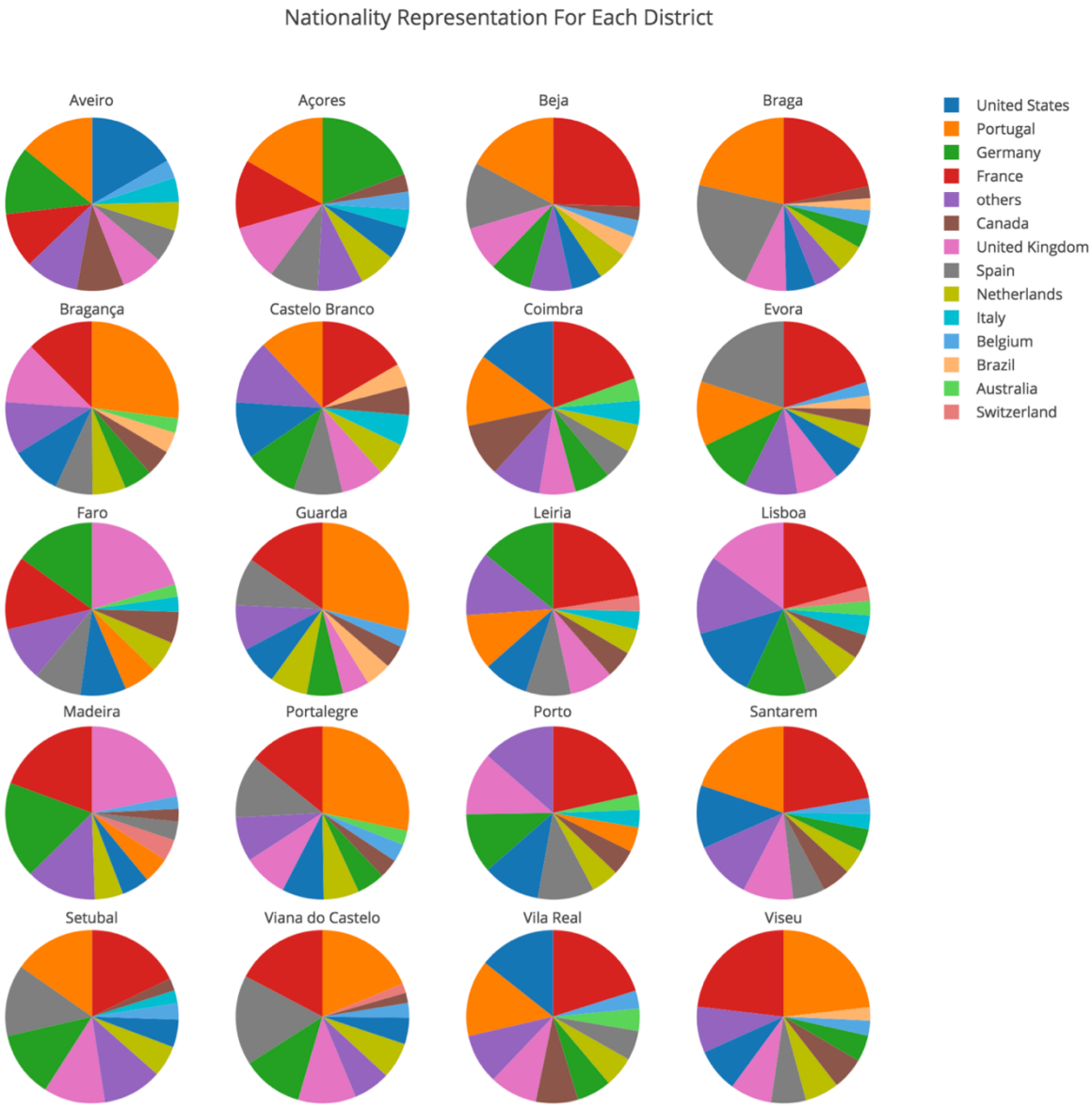


Figure 43: Countries of origin per district

Below is represented the estimated length of stay per Airbnb tourist in each district. Tourists visiting Madeira stay for longer periods of time, followed by coastal districts and Azores archipelago.

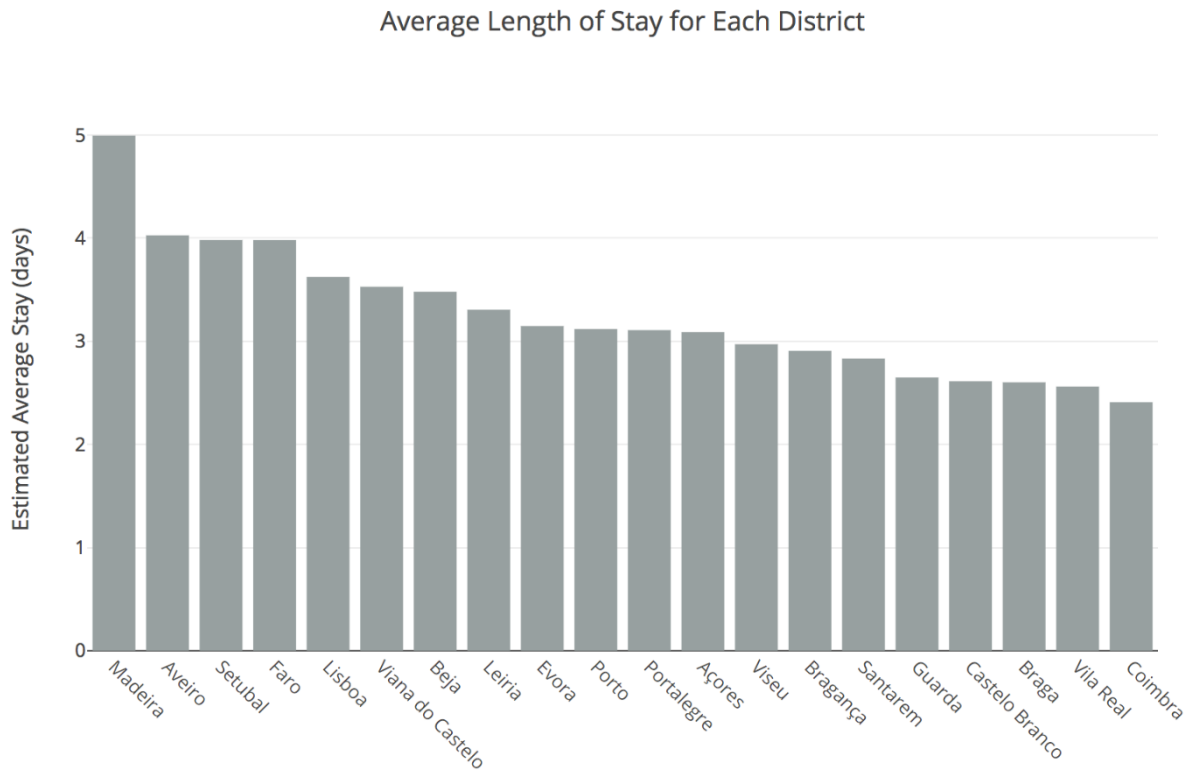


Figure 44: Average length of stay for each district

4.3.7. Country of Origin Analysis

Data regarding users' country of origin was taken from the reviews table. This table contains public reviews made by users after their stay in the listed offer. One must bear in mind that Airbnb differentiates reviews by public and private feedback, both for hosts and guests. So, in this situation we are analyzing a sample of user profiles that represent 10% of the total bookings.

The number of reviews for each month is presented below. Tourist behavior regarding month of visit is relatively equal across different nationalities, being the peak of tourists in the summer period, comprehended between June and October. Although, three different tourist patterns arise.

Tourists coming from mainly southern Europe countries such as Portugal, Spain, France and Italy are highly concentrated in the month of August, with a local maximum in the months between March and May.

Tourists from Australia, Belgium, Canada, United Kingdom and United States register similar visiting periods. These visitors, unlike the ones previously mentioned, have activity peaks in two different months, July and September.

As a third pattern, we can see Germany and Poland with their local maxima in September, with relatively high activity in the remaining months of the high season.

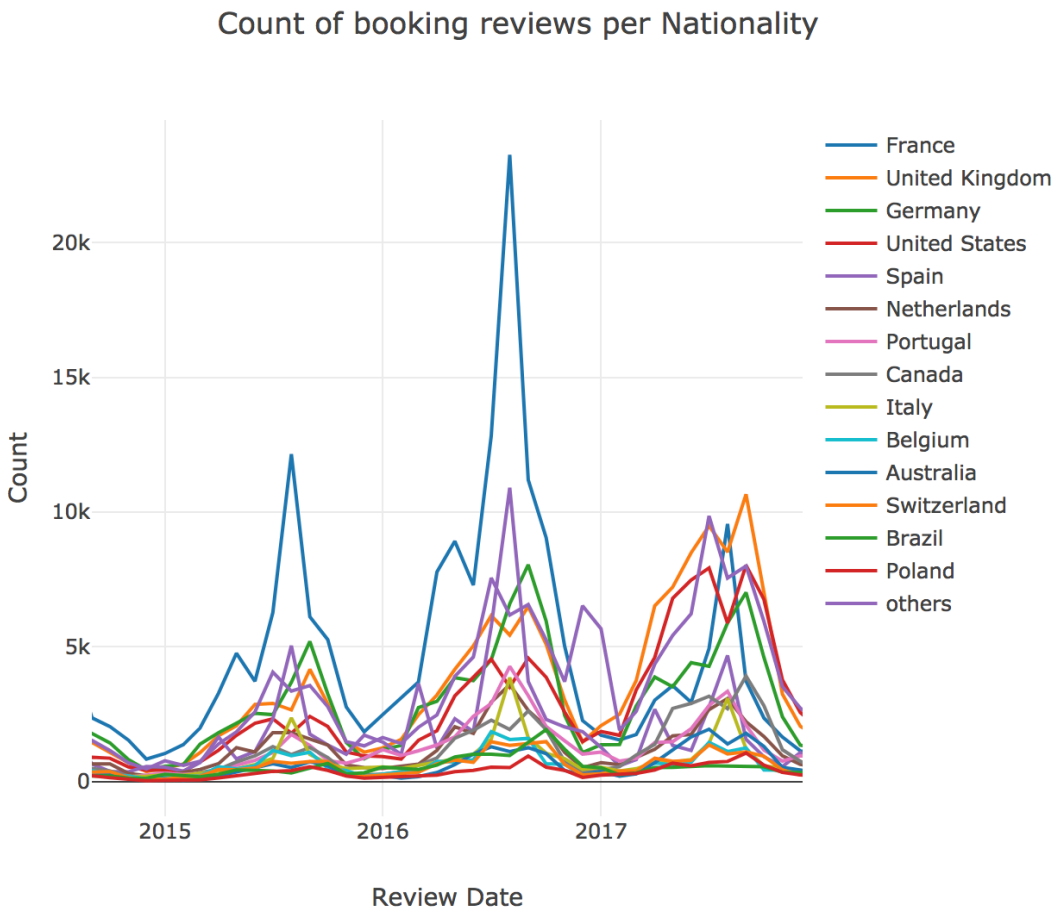


Figure 45: Booking reviews per country of origin.

Below is presented the Average Daily Rate (ADR) paid by the average tourist from each country of origin. Eastern countries represent the ones paying the lowest ADR, whereas central European tourists spend an average amount of money on Airbnb Bookings. Additionally, western European tourists from Portugal, Spain and France are spending above the average.

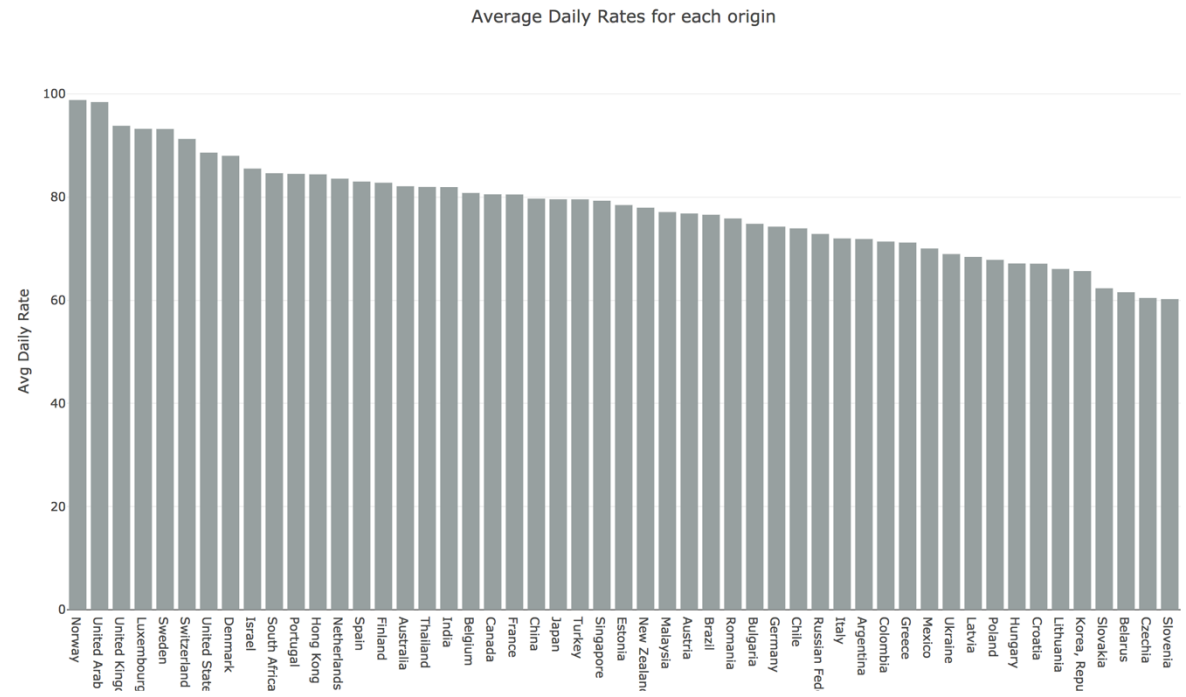


Figure 46: Average Daily Rates per country of origin.

Below is depicted the estimated average length of stay for each origin, which does not vary significantly across country of origin:

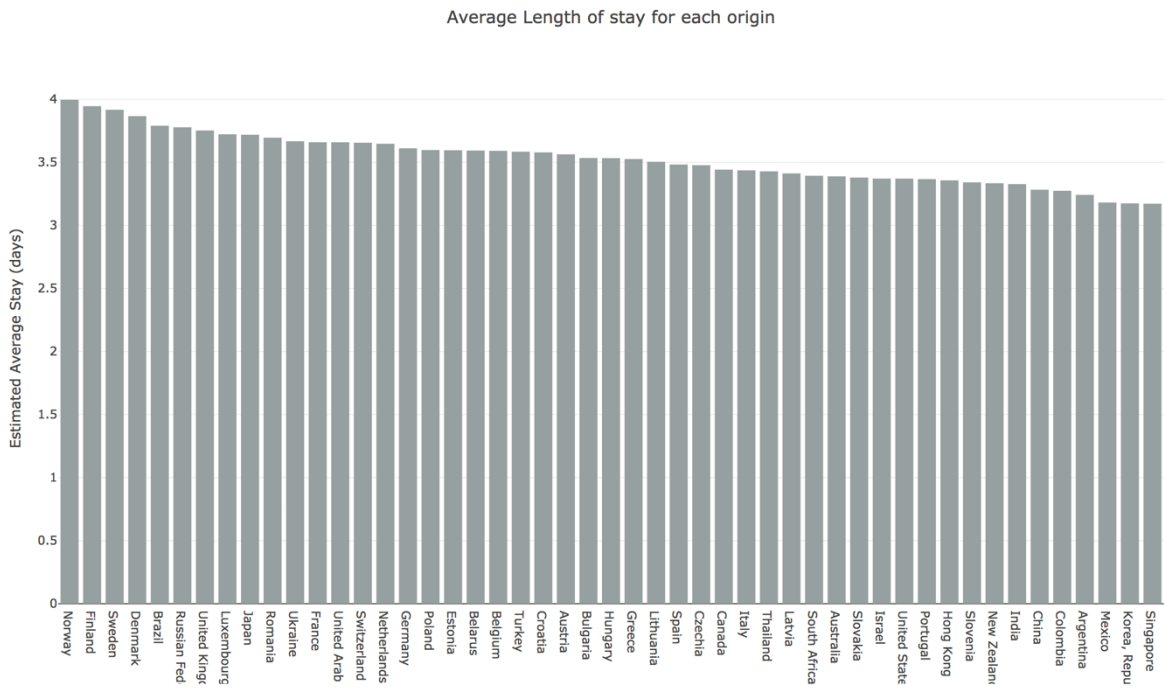


Figure 47: Average length of stay for each origin.

Although it's been concluded above that the overall main visitor is French, in 2017 France was the 4 main source of tourists using Airbnb, whereas the United Kingdom and the United States were the main sources.

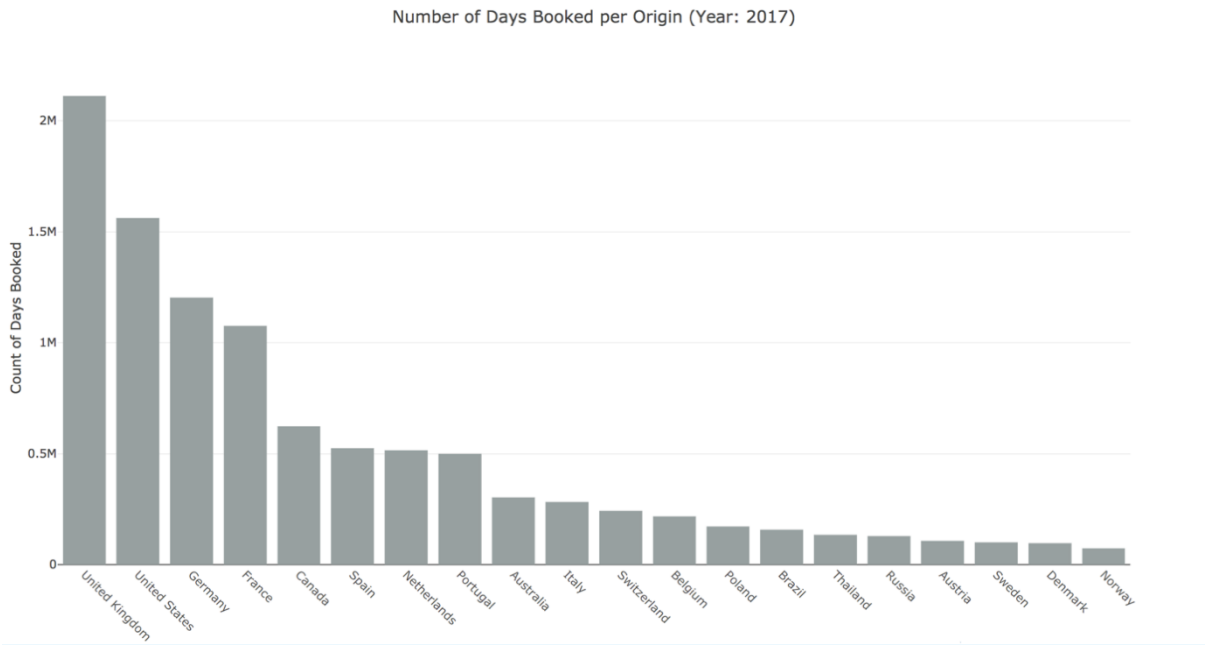


Figure 48: Number of days booked per origin (2017).

4.3.8. Value Analysis

Below is presented the weighted average daily rate for each region. The ADR was weighted using the number of reservations for each listing in each month. By doing this, we are attributing more weight to the houses with most listings and less weight to the ones with least listings and zero weight to the ones without listings. The reason to do this lies simply in the fact that listings that have no bookings don't have an impact in the market, whereas the ones with high amount of bookings have the greatest impact. Let m correspond to the month of analysis, l to each listing out of the total n listings, contained in a district, or in this case a set of listings D:

$$ADR_m = \frac{\sum_{l=1}^{n} Reservations_{l,m} * ADR_{l,m}}{\sum_{l=1}^{n} Reservations_{l,m}}, \qquad l \in D$$

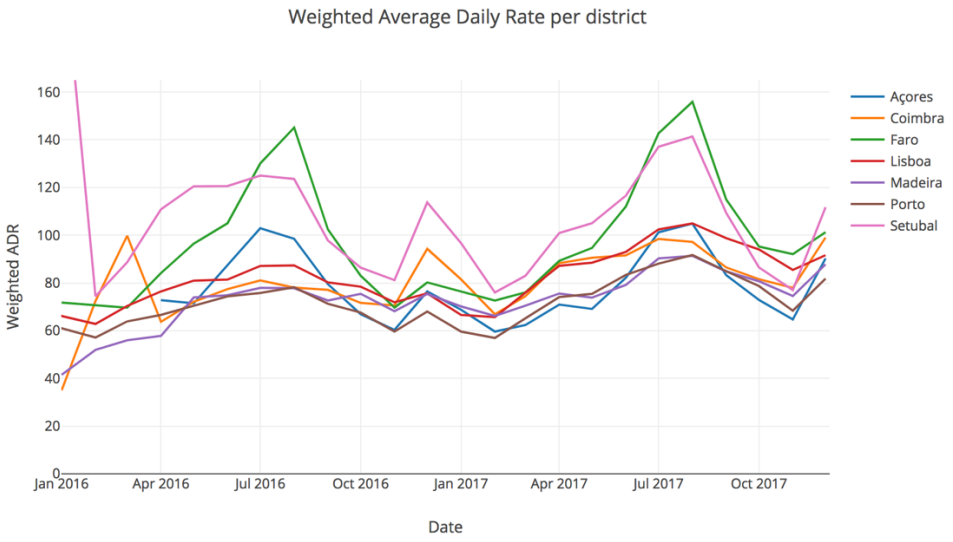


Figure 49: Weighted average daily rate per district.

Revenue Per Available Room (RevPAR) is a unit of measure that calculates the average revenue generated by a room (in this case a listing), within a set of rooms (in the case of a hotel, this would correspond to the set of rooms existing in a hotel), which will correspond to the set of listings that exist in a given district. It is once again visible the seasonal component of Airbnb's Demand. Lisbon and Porto present above average RevPAR values, with Faro being the district with the highest seasonality and highest RevPAR in the summer period of 2017.

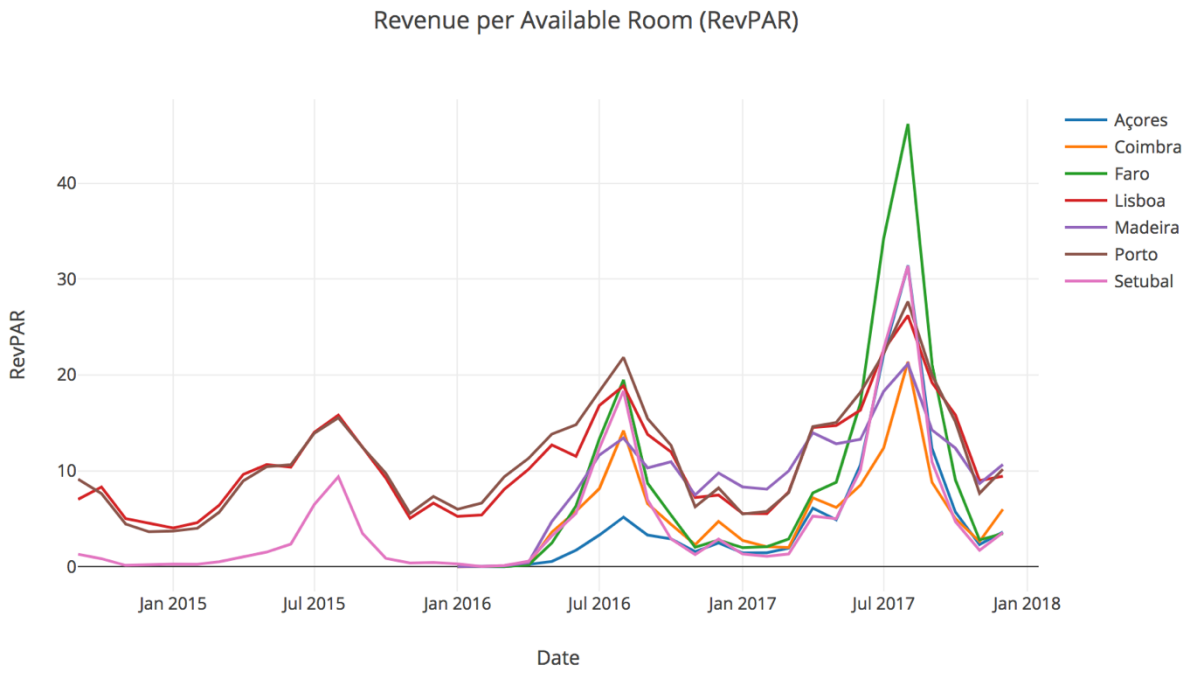


Figure 50: Revenue per Available Room.

4.4. Hotel Industry vs Airbnb Comparison

There is a clear growth of demand for Airbnb accommodation, resulting in growing revenue, reservations, number of listings and occupancy rates. In order to understand the source of the scale of this growth, it is necessary to get an overview of the Airbnb activity compared to the hotel industry. Such data is available in Turismo de Portugal's platform, TravelBI, containing information regarding Revenue Per Available Room (RevPAR), Hotel Industry Supply and Revenue (Turismo de Portugal, 2018). Additionally, a report by Instituto Nacional de Estatística on tourism statistics was also used (INE, 2017a).

4.4.1. RevPAR

From the graphics below one can confirm that in general the industry's RevPAR is significantly superior to the one from Airbnb. Although, the ranks in terms of RevPAR present one main difference: Algarve/Faro district is one of the regions with highest RevPAR in the hotel industry, while in Airbnb it only went above average in 2017, even though it grew at a significant pace from one year to the other.

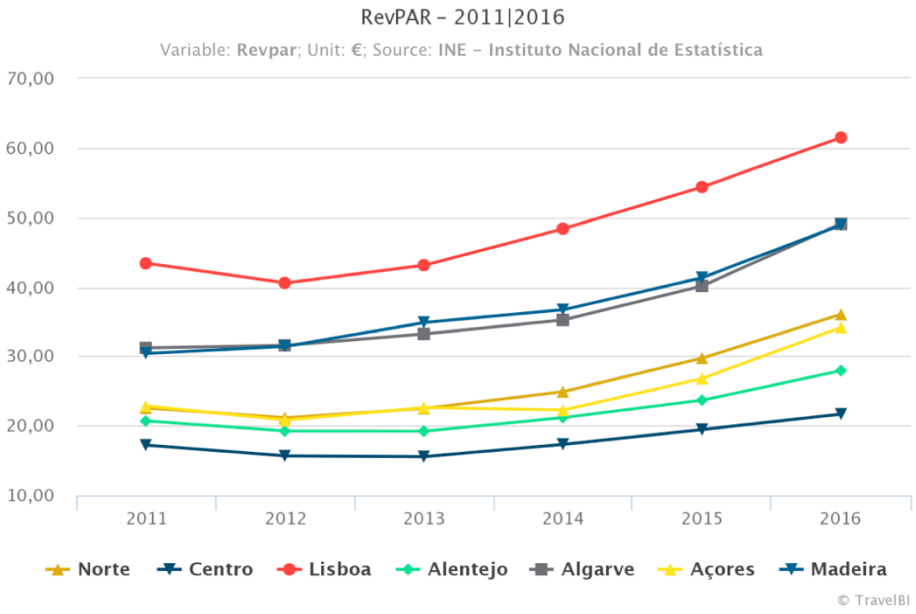


Figure 51: Hotel Industry's RevPAR. Source: TravelBI.

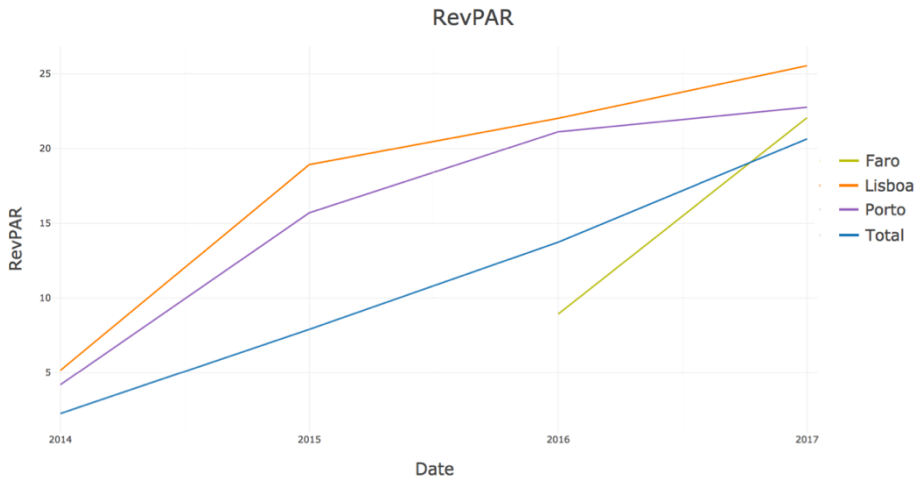


Figure 52: Airbnb's RevPAR.

Additionally, using a study from Statistics Portugal (INE, 2017a), one can now compare the RevPAR discriminating the Hotel industry by segments for the year 2016, available below. From this table it is possible to conclude that although Airbnb's RevPAR is on average less than half of the Industry's, it is approaching the RevPAR values of the average hotel with 1 or 2 stars.

NUTS	Total dos Alojamentos Turísticos	Total Hotelaria	Hotéis					Hotéis-Apartamentos				Aparta-mentos turísticos
			Total	*****	****	***	** / *	Total	*****	****	*** / **	
PORTUGAL	**40,2**	**44,6**	**47,1**	**84,7**	**47,6**	**30,3**	**25,4**	**42,9**	**56,7**	**44,4**	**33,5**	**29,2**
CONTINENTE	**40,2**	**44,7**	**47,3**	**87,4**	**49,0**	**30,5**	**25,7**	**44,1**	**61,6**	**44,1**	**38,0**	**29,3**
Norte	32,4	38,1	38,0	75,5	42,8	25,6	24,7	36,3	//	66,6	26,2	18,8
Centro	20,6	22,9	22,7	49,8	29,5	19,4	14,4	27,0	//	27,2	26,3	13,9
A.M. Lisboa	57,6	61,5	62,6	95,1	56,5	50,3	41,0	43,6	...	...	//	57,3
Alentejo	23,0	28,4	27,3	46,9	34,3	18,3	21,6	42,7	...	...	30,8	16,6
Algarve	46,1	47,6	62,5	90,1	57,5	39,1	32,1	45,7	...	...	40,7	29,8
RA AÇORES	**31,4**	**32,3**	**31,9**	**34,8**	**35,7**	**27,4**	**17,5**	**53,4**	**//**	**53,4**	**//**	**...**
RA MADEIRA	**42,5**	**48,6**	**52,3**	**76,8**	**44,1**	**28,5**	**27,5**	**38,5**	**33,7**	**45,0**	**20,0**	**...**

NUTS	Aldeamentos Turísticos	Pousadas	Quintas da Madeira	Total TER e TH	Turismo no Espaço Rural				Turismo de Habitação	Alojamento Local
					Agro-turismo	Casas de Campo	Hotéis Rurais	Outros TER		
PORTUGAL	**34,1**	**60,0**	**95,0**	**19,7**	**21,0**	**17,4**	**34,0**	**12,8**	**13,5**	**21,8**
CONTINENTE	**34,5**	**61,2**	**//**	**19,4**	**20,8**	**16,7**	**34,2**	**12,3**	**12,8**	**22,7**
Norte	...	...	//	16,9	15,0	14,1	25,5	11,0	16,9	18,2
Centro	...	...	//	19,0	28,0	15,1	38,6	7,7	12,1	12,0
A.M. Lisboa	45,6	98,0	//	37,0	...	35,6	...	//	...	35,8
Alentejo	13,7	47,7	//	18,2	21,2	16,4	38,8	11,4	5,8	11,8
Algarve	36,4	73,9	//	42,6	...	36,0	...	34,8	...	28,2
RA AÇORES	**//**	**...**	**//**	**19,7**	**...**	**19,2**	**//**	**...**	**26,1**	**x**
RA MADEIRA	**...**	**...**	**95,0**	**27,3**	**...**	**27,2**	**29,6**	**...**	**23,7**	**16,3**

Nota: RA Açores - Alojamento Local - informação não incluída por dificuldades de compatibilização

Fonte: INE – Inquérito à Permanência de Hóspedes na Hotelaria e Outros Alojamentos 2016

Table 11: Revenue per Available Room according to type of establishment, by region. Source: INE.

4.4.2. Supply

The entry in the hotel industry panorama when opposed to Airbnb is significantly different. As the maintenance costs of a hotel include significant fixed costs, as well as being in addition necessary a permit to operate legally, hotels face significant barriers of entry. To enter the Airbnb market on the other hand, it is solely necessary ownership of unused accommodation, without any permit requirements. As most landlords prior to entering Airbnb already own said unused accommodation, the most significant entry barrier for the latter are the costs associated to the ownership of the property, which most would be paying either way.

In the plots presented below is depicted the yearly evolution of bedroom supply in both Airbnb and Hotel industry. Airbnb room supply is growing exponentially and is getting close to the number of rooms in the hotel industry, in which the number of room supply is not growing as much as in Airbnb. Even though the dataset has been filtered to only include listings with 5 or more bookings in the last twelve months, there is still a possibility that some of these listings are only active for very limited periods within the year (e.g., in August, when demand is highest).

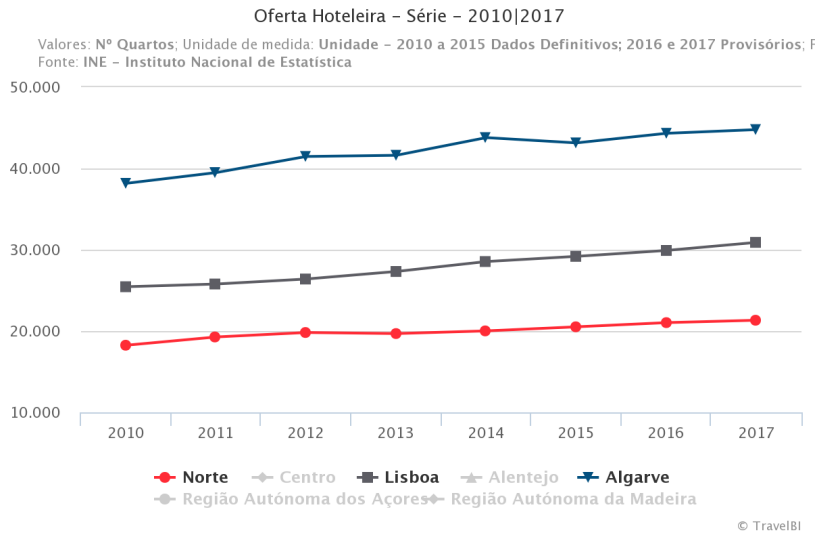


Figure 53: Bedroom Supply in the Hotel Industry. Source: TravelBI.

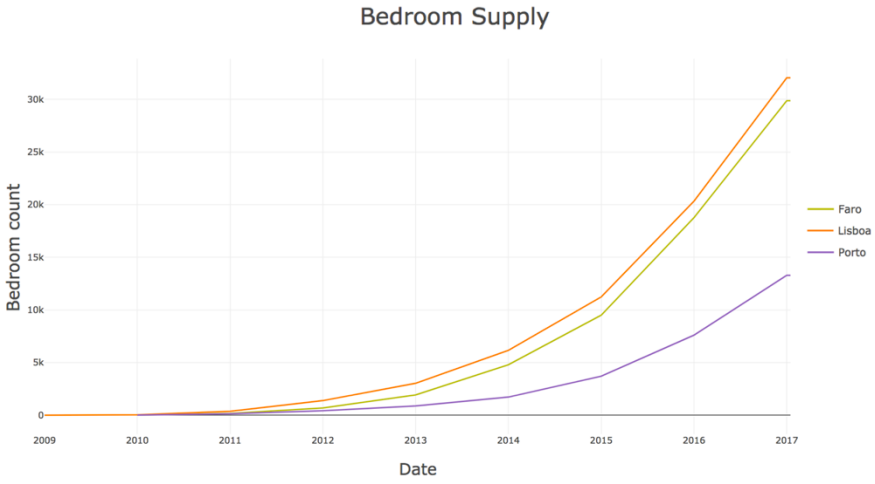


Figure 54: Bedroom Supply in Airbnb.

4.4.3. Revenue

It is now important to analyze the generated revenue by each market. This comparison clarifies significance of Airbnb in the overall tourist accommodation scenario. It's visible that in Lisbon the revenue generated by Airbnb is about one third of the revenue generated by the hotel industry in the same district. Although previously such a magnitude of significance was not clearly visible, in 2017 the existing supply in Lisbon area is nearly the same for both markets while the RevPAR is slightly above the one third ratio, thus explaining the values presented below.

Finally, the table below shows that in 2016 the revenue generated by one and two stars' hotels is the same as the generated revenue by Airbnb listings in the districts Lisbon and Porto. Although, from 2016 to 2017 the generated revenue in Airbnb listings has nearly doubled, suggesting that this economy is growing at a higher rate than the hotel industry.

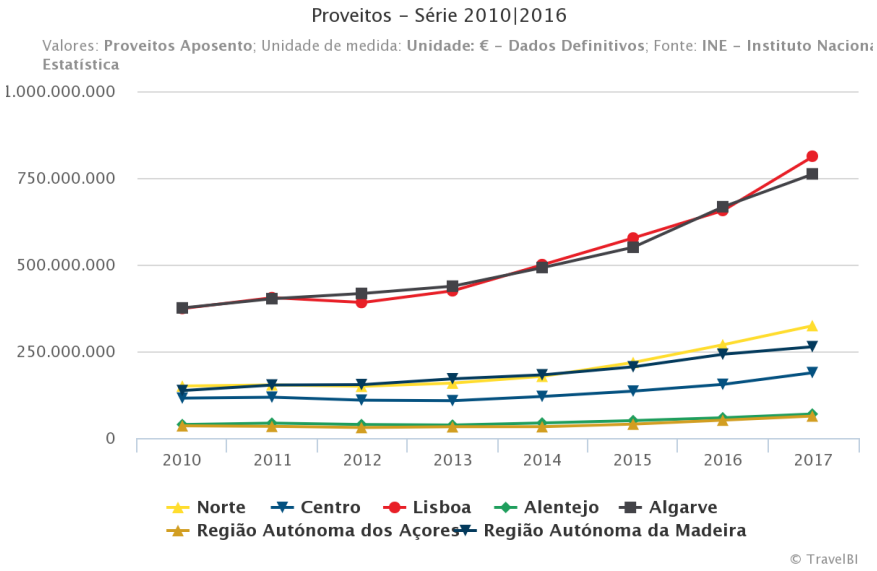


Figure 55: Hotel Industry's revenue. Source: TravelBI.

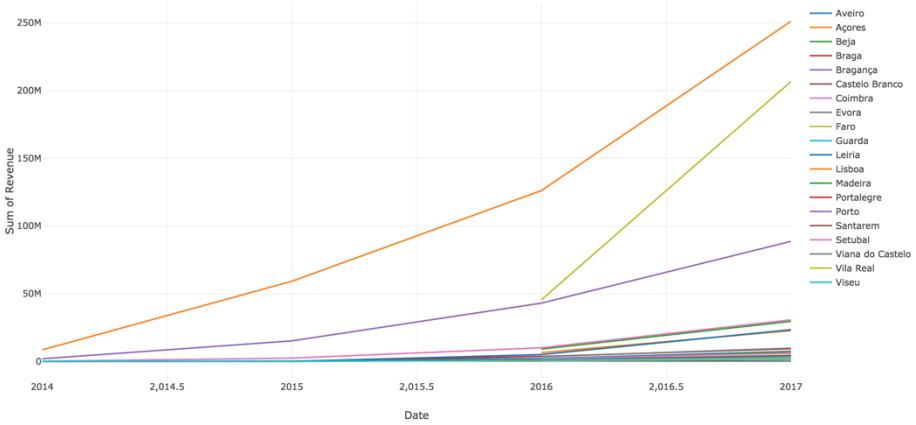


Figure 56: Airbnb revenue per district.

2016 Unidade: 10^3 euros

NUTS	Total dos Alojamentos Turísticos	Total Hotelaria	Hotéis					Hotéis-Apartamentos				Aparta-mentos turísticos
			Total	*****	****	***	** / *	Total	*****	****	*** / **	
PORTUGAL	**3103754,9**	**2823619,3**	**2190725,7**	**773949,8**	**983972,0**	**307680,1**	**125123,8**	**310840,8**	**32898,0**	**229302,1**	**48640,8**	**136856,2**
CONTINENTE	**2643565,4**	**2392238,3**	**1869196,8**	**650170,6**	**813231,7**	**286366,3**	**119428,1**	**231319,5**	**29289,8**	**162580,1**	**39449,6**	**127913,0**
Norte	403944,9	343623,0	323747,1	84743,0	146700,3	51466,2	40837,6	4 663,4	//	2489,1	2 174,3	1095,3
Centro	256513,8	212717,9	192807,5	21477,0	80236,5	70709,5	20384,5	6 311,3	//	4752,0	1 559,2	1834,4
A.M. Lisboa	931173,3	856711,3	790540,1	308729,7	337755,9	101207,4	42847,2	34102,9	...	...	//	6590,1
Alentejo	110894,5	80929,9	52702,5	9486,3	25454,8	11092,6	6668,8	15 425,6	...	...	1 918,1	3617,4
Algarve	941038,9	898256,3	509399,6	225734,8	223084,2	51890,6	8690,0	170816,4	...	...	33 798,0	114775,8
RA AÇORES	**73139,7**	**70679,0**	**63970,5**	**4 857,2**	**45588,6**	**10161,2**	**3 363,6**	**1928,3**	**//**	**1 928,3**	**//**	**...**
RA MADEIRA	**387049,8**	**360702,1**	**257558,4**	**118 922,0**	**125151,7**	**11152,6**	**2 332,1**	**77593,0**	**3608,1**	**64 793,7**	**9 191,1**	**...**

NUTS	Aldeamentos Turísticos	Pousadas	Quintas da Madeira	Total TER e TH (a)	Turismo no Espaço Rural				Turismo de Habitação	Alojamento Local
					Agro-turismo	Casas de Campo	Hotéis Rurais	Outros TER		
PORTUGAL	**117217,1**	**50195,0**	**17784,5**	**74119,0**	**11796,0**	**23811,0**	**27347,1**	**3715,0**	**7449,8**	**206016,6**
CONTINENTE	**115110,5**	**48698,4**	**//**	**67958,9**	**11319,6**	**20550,8**	**26289,7**	**3055,2**	**6743,6**	**183368,2**
Norte	...	...	//	23174,7	4113,0	5069,3	9377,2	1169,4	3445,8	37147,2
Centro	...	...	//	15748,2	2244,4	5037,1	6293,5	463,6	1709,5	28047,7
A.M. Lisboa	10890,3	14587,9	//	1609,7	...	408,5	...	//	369,7	72852,4
Alentejo	1 674,2	7 510,2	//	20565,3	3614,7	7439,1	7994,4	578,6	938,5	9399,4
Algarve	97526,8	5737,7	//	6861,1	...	2596,8	...	843,7	280,1	35921,5
RA AÇORES	**//**	**...**	**//**	**2460,7**	**...**	**1279,1**	**//**	**...**	**454,5**	**x**
RA MADEIRA	**...**	**...**	**17784,5**	**3699,4**	**...**	**1981,1**	**1 057,4**	**...**	**251,7**	**22648,4**

(a) No Continente, apenas estabelecimentos com 10 ou mais camas.

(b) RA Açores - Alojamento Local: informação não incluída por dificuldades de compatibilização

Fonte: INE – Inquérito à Permanência de Hóspedes na Hotelaria e Outros Alojamentos 2016

Table 12: Total Revenue, according to establishement type, by region. Source: INE.

5. RESULTS AND DISCUSSION

As observed in the background information section, tourism plays a fundamental role in the Portuguese economy. The benefits caused by this industry are well documented: they play an important role not only in the economy but also in sociocultural and environmental components. Although, this industry is sensitive to unforeseen shocks if such have a negative impact on tourism inflows. This impact is dependent on the nature of the shock: demand side, government, private sector, intangible factors, or external factors. As tourism does not solely bring benefits to a country, it also carries downfalls in the three components above mentioned.

The general goal of policy makers regarding the tourism industry is clear: optimize the relationship between the costs and benefits from tourism considering the three factors that are affected by tourism.

This study focuses at one stage on the extraction of information out of Airbnb and Telecom data that can ultimately prove useful to analyze and minimize one of the phenomenon mentioned: tourism over-dependence (economic factor). As a second stage, it focuses on the development of visualization and data collection tools. These tools mostly relied on social media and telecom data sources, with which the flows of tourists across Portugal can be analyzed regarding both flows' intensity and direction, as well as acquire an overview of social media activity regarding a given keyword, and the crawling of such data. All these tools were intended to be used as web applications and allow user interaction for facilitated analysis. As such, their development relied on web design programming languages and Python.

5.1. MAJOR FINDINGS

Although the analysis of Telecom data was restricted to the month of August 2017, the share of French roamers on the total roamers is the highest, followed by Spain and the United Kingdom, thus suggesting that the share of French tourists visiting Portugal is very high, same going for Spanish and UK tourists. Additionally, we can see that the majority of these roamers visit the districts of Lisbon, Porto and Faro, having also a high share of roamers visiting the districts of Braga, Viana do Castelo and Vila Real. When allied with the high share of French roamers as well as Swiss and Luxembourgish visitors along with their average length of stay (top 3), and considering the period of analysis, suggest a significant presence of visiting Portuguese emigrants. The same analysis using Airbnb data showed similar results, with less presence of Spanish tourists and higher presence of US tourists.

Tourists typically arrive on Tuesdays and depart on Wednesdays, weekdays in which plane tickets are typically cheaper, suggesting a significant presence of tourists arriving by plane at these days. Accordingly, 52% of the roamers stay in Portugal for 8 days or less.

Airbnb booking reviewers' country of origin was used as a proxy for the population of tourists in Portugal using Airbnb. Through the analysis of the number of booking reviews per nationality in each period, three different visitation patterns became clear: 1) Southern Europe visitors, such as Portugal, Spain, France and Italy have their peak activity in the platform in August (with a local maximum between March and May). 2) Anglosphere countries plus Belgium present activity peaks in two different periods, July and September. 3) German and Polish visitors have their peak activity in September, with overall high activity in the remaining months of the high season: June to August.

In general, tourists coming from eastern countries are the ones paying the lowest daily rates, whereas high GDP European countries tourists, along with United Arab Emirates and USA, are the ones paying the highest daily rates.

Airbnb listings are highly concentrated in Lisbon and Porto, being the majority of these Entire homes/apartments. There is an exponential growth of the number of Airbnb bookings in Portugal, having most of these been made in the past two years. Although, the number of inactive Airbnb bookings is also significant. Overall, the Airbnb users in Portugal have as origin France, United Kingdom, Germany and United States.

The least popular regions for Airbnb users in Portugal are the districts of Castelo Branco, Guarda, Portalegre, Viana do Castelo and Viseu. As such, these are also the regions where Portuguese tourists take the most significant share of Airbnb activity. The economic value in Airbnb of these districts is also low, being the most valuable ones Faro, Lisbon and Porto (in terms of generated revenue and RevPAR).

Although listings are mainly located in Lisbon and Porto, Faro district is the district generating the most revenue for the month of August 2017, followed by Lisbon and Porto, displaying a clear seasonality pattern. Likewise, the monthly occupancy rate per district demonstrate such seasonality, being less noticeable in Madeira. Users book Airbnb listings for a short period of time (similar to that observed in the telecom data), if excluding Luxembourg, France, Switzerland and Belgium (as many of these might be visiting Portuguese immigrants).

The contextualization of the Airbnb market using Hotel industry data revealed that Airbnb listing have under average revenue per available room (RevPAR), being comparable to the average hotel with 1 or 2 stars. Airbnb supply is growing exponentially and is close to the hotel industry's supply.

Through the analysis of the telecom data it was observed that during the month of August 2017, tourists visited Portugal mostly during the second week of the month. After that period the number of tourists begun to decay. Additionally, these tourists visited mainly the districts of Lisbon, Faro and Porto, going in accordance to the observed tourist behavior in the Airbnb platform.

As discussed above, the average length of stay of each tourist's country of origin using telecom data is biased towards the countries with the largest Portuguese communities outside of Portugal. Regardless, it also observable that the average length of stay of tourists from the USA in Portugal is approximately 3 days, significantly lower than the remainder of top visiting countries of origin. Additionally, it has been observed that 52% of visiting tourists stay in Portugal for 8 days or less.

5.2. MEANING AND IMPORTANCE OF THIS STUDY

The contribution of this work rely on three components: data used, proposed methodology, and data analysis.

This study introduces the analysis of the three sources of data altogether for the analysis of the tourism industry. Although such concept was only superficially explored, it was possible to observe that these sources allow complementary types of analysis to be developed and serve as a method to ensure the trustworthiness of the outcomes of these analyses.

The development of a Web Crawler prototype to scrape social media data demonstrates a method to easily gather data (both structured and unstructured data) in continuous basis for analysis. The main benefit of the presented methodology is its reproducibility, which allows third parties to easily adapt the presented work and adapt the crawler to their needs, as well as extend it for further functionalities and additional sources for gathering data. Such data has proven to have different goals of application: Tourism marketing (campaign monitoring, consumer profiling etc.), consumer behavior analysis, online word-of-mouth impacts, tourist experience analysis etc.

The analysis of the tourism sector using web scraped data (Airbnb) and telecommunications network events instead of the typical sources such as survey data allow the analysis of the entire population (in the case of telecom data, the population are the tourists that connected to NOS network, not the populations of tourists visiting Portugal), instead of using sample-based methods. The conclusions drawn from these analyses represent an overview of the entire population or a very large proportion of it (which is what happens in the case of the telecom data) out of reliable sources of data, making this approach a valuable source of knowledge creation. Specifically, the case of Airbnb has been observed to represent an important economic activity within the tourism sector and was previously difficult to accurately analyze (i.e., only survey data was used to analyze Airbnb in Portugal up to this point). Additional to exploratory analyses, this work demonstrates the superficial application of a multi-method analysis which, if taken to greater extent, will potentially provide more non-intuitive insights to be extracted.

Portuguese tourism promotion agencies, governmental tourism management bodies and tourism related businesses can use this work for different purposes: Set up a monitoring tool to develop analyses/dashboards with similar data, or simply benefit from the insights extracted from the data analysis in the work. The broader view of the tourism sector portrayed here is capable of being maintained in a near real-time basis, having for that reason the capacity of providing the data to develop complex analysis and dashboards to inform all the agents involved in the tourism sector, analyze the how tourists travel around Points of Interest in Portugal and thus allow the creation of policies that can alleviate the negative consequences of tourism in Portugal.

5.3. RELATION TO SIMILAR STUDIES

Part of the visualization tools presented in this work as well as its structure were inspired on key studies of the area (J. Li et al., 2018; Miah et al., 2017; Raun et al., 2016). To a large extent, the results of the analyses presented are relatively similar to other studies in Portugal using survey data (INE, 2017b), with some exceptions. This survey, conducted by Instituto Nacional de Estatística, considered only establishments listed in Airbnb located in the main land with 10 or more beds, which constitutes only part of the total listings of the platform. As such the number of listings previously estimated in the platform was much lower than the reality and consequently the estimated revenues. Regarding the proportion of countries of origin, the estimate in INE's study is close to the ones observed in this thesis, with a few variations: The representativeness of US tourists is higher than the one mention in their study, and the order of the top countries of origin are estimated differently: France takes the highest representation in this study, whereas INE's study ranks this country in third and Spain has a higher representation than estimated in INE's study, being ranked before the Netherlands. Occupancy rate was similar to the one observed in their study and the RevPAR measure is higher similar to the one observed in this study, although both measures depend on the criteria applied to filter inactive listings.

The methodologies applied in the development of the social media crawler used standard interactions with Facebook and Twitter's API, and web scraping on Instagram. For this reason, these procedures follow the methodology presented in most of the studies using social media data. Regarding telecom data, the methodology followed the regular procedures. The visualizations benefited from modern tools such as Deck GL, which allowed the visualization of flows in a manner that has been found in any other study.

Finally, although the type of analysis developed was significantly different from each other (discussed below in limitations), literature developed within this field rarely integrates more than one source of data with the goal of exploring tourism flows, tourist space or tourist behavior. The work presented has shown that the analyses developed have the capability of complementing each other and provide a broader view of the tourism sector in its whole.

5.4. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Throughout the development of the project and writing of this thesis, some limitations were detected:

5.4.1. Telecom

The data used corresponds to network events solely in NOS' network. As such, there is a clear bias in this situation: The data is not for the full population of tourists and was not randomly sampled either. For this reason, it is important for future research in this area to merge data from a plurality of network providers as a manner of approaching as much as possible the available data to the population of tourists in Portugal. Given the recent EU policy of free roaming, European citizens can now use their mobile plan as if they were in their home country, whereas non-EU citizens are still subject to roaming, which constitutes an incentive for those tourists to buy a national SIM card, which means that the proportion of non-EU citizens to EU citizen might be lower than the reality, meaning the presence of a bias in the analysis. Although there's not much one can do to eliminate this bias, it is possible to cross-validate the results with survey data, providing some understanding of the missing data and biases. Additionally, there are some inaccuracies in the data: while some users in many periods generate network events very frequently (within minutes or even seconds), others only occasionally connect to NOS' network. Future works analyze carefully these inaccuracies and check to which level of granularity is it possible to use this type of data. As this work only used one month of telecom data (the only time frame to which access was granted), it greatly limited the number of analysis that could have been performed. Consequently, as this month was August, this same analysis contained the additional bias of Portuguese immigrants visiting Portugal, which gave additional significance to French roamers (where in fact these immigrants are not tourists).

5.4.2. Social Media Crawler

As mentioned in the description of the development of the tool, the crawler (as any other web scraping tool) is very sensitive to changes in a company's policies or website structure. This means that the tool may require future updates whenever the policies of the used APIs change (as is happening with Facebook given the Cambridge Analytica Scandal), or in some situations when the structure of the website itself changes (as is the case of Instagram). As mentioned, both Instagram and Twitter have restrictions on the available time frame and number of posts crawled, which require the capability to run the social media crawler in a continuous basis, involving infrastructures/conditions that are difficult

to attain. Although it is difficult to go around this problem in the case of Instagram, Twitter sells API keys that ensure the removal of these restrictions. Future research should consider purchasing one of these keys. Alternatively, it is also important to consider acquisition of the necessary infrastructures to have the crawler run continuously.

5.4.3. Airbnb

Although Airbnb data can provide valuable information, a few limitations to this study must be pointed. It is impossible to assess whether the reviews data is accurate, given that Airbnb's accuracy in their own data is not certain. Furthermore, Airbnb's reviews can be either public or private. As we are only using publicly available data, we do not have access to user data that left a private review, or no review at all. So, we are analyzing user profiles that represent about 10% of the total bookings (≈ 1.2 million public reviews) that were actually completed. Although the user data sample extracted from the overall reviews was not randomly generated, it is highly representative. The number of total completed bookings made between September 1st 2014 and December 31st 2017 is 11.550 million Bookings which implies a minimum sample size of 16564 for a 99% confidence level and 1% margin of error. As our dataset has a depth of 1.2 million observations, it is statistically significant (although, we cannot conclude that it is an unbiased sample, as it was not randomly selected).

A second limitation would be the accuracy of data scraped by the data provider, AirDNA. In these datasets, some of the data is not very consistent. For instance, in these datasets the variables regarding monetary value (Daily rates, listing monthly incomes, etc.) are not always directly convertible between USD currency and native currency (euros), where it was found that in these situations variables with USD currency turned out to be more trustworthy than native currency. Same inconsistency applies to variables such as Booked Date, where in situations that the booking was completed this field was still left blank. Aside from these type of situations, some clear outliers were found, as is the case of some bookings that exceeded the price of $100 000 (with a maximum value of $540 954).

Moreover, given the recency of the popularity of the Airbnb platform, it is far too early to have certainty of the impact this platform can have in the overall tourism panorama. As this this new service is still going through a fast growth, markets are still accommodating to it.

Finally, it is also important to mention the impossibility to calculate the exact number of overnight stays on Airbnb. This happens because no information is provided about the number of guests staying for each booking or total guests for each month, or even maximum number of guests per listing. The only information close to this goal is the number of bedrooms, which is also not sufficient for that calculation (one bedroom can have many beds, a king sized bed, or only a single bed). So, an estimation of the number of overnight stays wouldn’t be reliable given the existing data.

5.4.4. Joint Analysis of the Data

Considering the above mentioned limitations of the available time frame for the telecom data and the social media data, in this work it was not possible to cross analyze the sources of data to a meaningful extent.

5.4.5. Applicability of Recommendations

The recommendations mentioned in the conclusions section are based on this first analysis of the tourism industry in Portugal and third party research. As such, the above mentioned policies and platform require further analysis/research in order to ensure how they could be applied and which should be applied.

6. FUTURE WORK

Research in this area, using the methods introduced, should consider extending the work developed in this thesis in a number of ways:

- Firstly, in order to attain representative data sets from social media platforms, future work should consider the adaptation and development of the necessary adjustments to overcome some of the limitations mentioned in the section above.

- Secondly, additional network science concepts should be used to study the telecom data. Through these sources of data, patterns of tourist behavior can be analyzed using sequence clustering methods and machine learning concepts such as the application of word2vec algorithms in order to understand the relationship of each point of interest regarding tourist flows.

- Thirdly, analysis of tourist feedback and trends using social media data should be executed using Text Mining techniques such as Named Entity extraction (detection of the topic in discussion) and Sentiment Analysis (association of sentiment to topic extracted from Named Entity Extraction). It might be useful to add more sources of data, both from other websites (e.g., reviews websites, hotel bookings etc.) or even better understanding local business activities and third party website activities through RSS feeds, thus allowing the user to understand what is happening in the environment in which the latter is inserted into.

Additionally, by using telecom data along with the social media data it is possible to correlate findings in each location, associating movement patterns related to a given location with the main topics discussed in social media and associated sentiment and see how social media's word-of-mouth effect impacts tourism flows. An alternative research path is to focus on the predictive analysis of tourism flows and tourist behavior, which is possible to perform through the usage of these sources of data.

7. CONCLUSION

In order to propose a tourism management approach supported by data driven methods and sources of big data, three tools have been presented in this work using different sources of information.

A preliminary study showed the current trends of tourism in Portugal, compared to the growing demand of tourism worldwide and specifically Europe. This allowed the breakdown of both positive and negative consequences of tourism growth, divided by Economic, Sociocultural and Environmental effects. Finally, it was observed that the symptoms experienced by European cities with significant anti-tourism movements and ideologies are currently being felt in Lisbon and Porto, where it was observed that such cities presented a tourist/citizen ratio above some of the most touristic cities in the world, with rising gentrification effects which is ultimately leading to general bitterness towards the growth of tourism.

The tools developed posit a proof of concept that can be used to produce production ready applications to monitor the movement tourists in Portugal. If deployed properly, these tools demonstrate the potential to become a primary source of knowledge to assist tourism management.

Telecom data was used to develop a tool that allow the analysis of tourism flows across major cities in Portugal, which can be used to design additional tools and policies to influence the spread of tourists across a city and avoid overcrowding, improve the tourists' experience in the city and minimize co-existence conflicts.

From the data analysis, significant findings include the most common weekday of arrival and departure (Tuesday and Wednesday, respectively), length of stay to be equal or inferior to 8 days for over half of the visitors and the substantial presence of visitors with France as country of origin, which hinted to Portuguese emigrants' visits and provided confirmation on the most commonly visited regions in Portugal (Lisbon, Porto and Faro).

A social media crawler was developed to fetch data from Facebook, Instagram and Twitter. Such tool crawls data not only tourists' generated content about given their experience, but also the locals' generated content about tourism. Although, given the generalist nature of this program, it can be adapted for many other applications.

Airbnb data analysis revealed tourism inflows' seasonality, exponential growth in demand and supply of accommodation units in the platform. Booking activity patterns were observed for different groups of countries of origin. The comparison of Airbnb data with Hotel data demonstrated that the average Airbnb listing is currently comparable in terms of RevPAR and generated revenue to the average low end hotel hotel service (between one and two stars) and, as previous literature suggested, the two services are perceived as direct substitutes (Zervas et al., 2017) making Airbnb a competitor of the Hotel industry.

Finally, a set of recommendations were proposed to demonstrate the potential applications of the presented tools. The approach adopted was to design an integrated set of recommendations, while still being possible to adopt each of the recommendations separately.

7.1. Recommendations and Potential Applications

Having the tools and analyses been presented in this work, the need to set up a system through which Airbnb hosting can be regulated, minimize gentrification in major cities and dilute tourists across said cities, such that it becomes possible to manage flows and accommodation in the tourism industry becomes vital in order to ensure sustainable tourism in Portugal.

The recommendation presented below represents an integrated approach towards accommodations' location management, Airbnb supply management and tourism flows management.

7.1.1. Using the Presented Tools For Tourism Analysis

The project provides three sources for data analysis that will allow authorities to make data-driven decisions. Such tools can be further improved and used continuously, potentially representing a sustainable source of knowledge.

7.1.2. Dispersing Tourist Accommodation Away From the Center of the Cities

The demand for tourist accommodation in the center of cities is justified with the proximity to its Points of Interest. Although, in the cities this phenomenon is occurring (especially Lisbon and Porto) one can assess that there is plenty of unused space in the suburbs that would benefit from such economic activity. As previously suggested, Airbnb housing is currently comparable to low tier accommodation rental, which insinuates that these users are price sensitive. As real estate value is lower in these areas, low cost accommodation could be created.

This measure generates a new challenge: ease and speed of access to the center of the city, thus minimizing the drawback created from the remoteness to said Points of Interest.

7.1.3. Optimization of Public Transportation Networks

Research towards the optimization of public transportation networks has been greatly developed. In this situation, the goal is to efficiently connect the suburbs to the corresponding major cities. To do this, two possible approaches were found:

1. The "Hub and Shuttle" approach (Mahéo, Kilby, & Hentenryck, 2017):

The model proposes a hub and shuttle model consisting of a combination of high-frequency bus routes between key hubs (corresponding to already existing bus stops), thus proposing centralized routes oriented for reduced waiting times, using secondary transportation systems (shuttles) to transport passengers between their closest bus stop to the closest hub.

2. The transit frequency optimization problem approach (Martínez, Mauttone, & Urquhart, 2014):

This model focuses on improving the efficiency of frequency in existing public transportation systems, which represents a less intrusive approach, as it simply optimizes existing transportation routes.

Next, a new challenge arises: How can we regulate the number of Airbnb properties in each location?

7.1.4. Mechanisms for Airbnb housing regulation

A study based on the analysis of benefit of the Airbnb platform in London provided recommendations based on the results obtained, which allows the constant evaluation of the impact of short-term accommodation rental in the city and thus develop a mechanism able to adjust the existing regulation at any point in time (Quattrone et al., 2016).

The idea behind these recommendations focuses on the development of a new asset: **Sharing Permits**. The asset consists of a permit to list a given accommodation in an accommodation sharing platform, as is the case of Airbnb. The implementation of this mechanism would possess the following features:

- The number of permits issued would be segmented and adjusted for each municipality, implying that in order to list an accommodation in Airbnb the owner will require a Sharing Permit for the corresponding listing and municipality. The decision process of the allocation of these permits must take into account the consequences for adoption, the impact on local economies, sustainability of tourism and avoidance of over-concentration of short-term rental supply.
- Sharing Permits would be transferable among citizens. To do this, the creation of a web platform would allow the intermediation of the process where the pricing of these permits would be based on market demand and municipality policies.
- The terms of transferable rights would change depending on whether a room or an entire apartment is rented.
- The creation of a data sharing ecosystem would allow agents to have equal access to information and thus promote fair competition among them.
- Finally, the number of permits existing for each municipality must be regularly evaluated considering the above mentioned factors and adjust the amount of circulating permits accordingly.

As demonstrated, this regulation can be monitored through the usage of updated data of Airbnb, allowing the regulatory entities to monitor which properties possess the required permit to operate in the region.

This supply regulation would additionally allow Airbnb listings' daily rates and occupancy rates to increase in city centers. This would consequently allow listings located outside of the city center to more easily appear in search queries, given that these listings' daily rates and occupancy rates are lower, raising awareness of existing supply in these areas.

Although, tourists must now be aware of the ease of access to the city center.

7.1.5. Educate tourists regarding mobility options

As a way to create awareness of the existing mobility solutions, it becomes necessary to promote and facilitate access to such knowledge. To do this, it is proposed **an extension to the above mentioned platform**, such that it would have 2 facets: **the landlord side** (for Sharing Permit trading) and the **tourist side** (for access to information regarding mobility solutions and accommodation).

The advantage of such platform is that it is not intended to compete with any other existing service, but rather promote it. This means that the information regarding transportation and route processing could be provided by services that already have sophisticated solutions for the purpose (e.g., Moovit).

Additionally, integration with accommodation providers (e.g., Airbnb and Booking) would be added as a feature.

Now it becomes necessary to develop a mechanism to be able to manage the tourism flows in the city.

7.1.6. Dilute tourist flows across the city

The tourism flows could be managed via the platform above mentioned. Using the tools presented in this project, it becomes possible to develop a recommendation system that takes into account already visited locations, current location and number of tourists in each location, either as an estimate for each hour per day per period of the year (using historical datasets) or in real time. The system would work as a tourist guide adapted for each tourist/group in such a way that their experience would be optimized while avoiding the excessive clustering of tourists in specific locations.

As the recommendation system would underlie the usage of telecom data, the success of the platform will require the mass adoption of given tool for it to have a significant impact, which is an important factor to ensure a successful implementation.

Lastly, in order to create further incentives to the adoption of the platform, other simple features could be added, such as information regarding healthcare facilities' locations (hospitals, pharmacies, etc.) and security facilities (such as police squads).

7.1.7. Creation of an Open Data Center

Tools such as the Social Media Crawler presented in this work can be further developed to develop open data centers capable of incentivizing research initiatives, assist in the development of local businesses and power events such as hackathons, which would constitute a source of continuous knowledge creation, useful for many different purposes.

8. BIBLIOGRAPHY

Andereck, K. L., Valentine, K. M., Knopf, R. C., & Vogt, C. A. (2005). Residents' perceptions of community tourism impacts. *Annals of Tourism Research*, *32*(4), 1056–1076. https://doi.org/10.1016/J.ANNALS.2005.03.001

Apache. (n.d.). Apache Hadoop. Retrieved October 14, 2018, from https://hadoop.apache.org/

Aurenhammer, F. (1991). Voronoi diagrams---a survey of a fundamental geometric data structure. *ACM Computing Surveys*, *23*(3), 345–405. https://doi.org/10.1145/116873.116880

Batista e Silva, F., Marín Herrera, M. A., Rosina, K., Ribeiro Barranco, R., Freire, S., & Schiavina, M. (2018). Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. *Tourism Management*, *68*, 101–115. https://doi.org/10.1016/J.TOURMAN.2018.02.020

Bloomberg. (2018). Facebook Cambridge Analytica Scandal: Here's What Happened. Retrieved April 29, 2018, from http://fortune.com/2018/04/10/facebook-cambridge-analytica-what-happened/

Codagnone, C., & Martens, B. (2016). *Scoping the Sharing Economy: Origins, Definitions, Impact and Regulatory Issues*. Retrieved from https://ec.europa.eu/jrc/sites/jrcsh/files/JRC100369.pdf

Corwin, S., & Pankratz, D. (2017). *Future of mobility overview*. *Deloitte*. Retrieved from https://www2.deloitte.com/content/dam/insights/us/articles/4328_Forces-of-change_FoM/DI_Forces-of-change_FoM.pdf

Cusumano, M. A. (2014). How traditional firms must compete in the sharing economy. *Communications of the ACM*, *58*(1), 32–34. https://doi.org/10.1145/2688487

Daniel, A. C. M., & Rodrigues, P. M. M. (2010). Volatility and Seasonality of Tourism Demand in Portugal. *Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies*. Retrieved from https://econpapers.repec.org/article/ptubdpart/b201003.htm

De Mauro, A., Greco, M., & Grimaldi, M. (2014). What is Big Data? A Consensual Definition and a Review of Key Research Topics. Madrid: 4th International Conference on Integrated Information. Retrieved from https://www.researchgate.net/profile/Andrea_De_Mauro/publication/265775800_What_is_Big_Data_A_Consensual_Definition_and_a_Review_of_Key_Research_Topics/links/54e61d170cf277664ff2f0b4/What-is-Big-Data-A-Consensual-Definition-and-a-Review-of-Key-Research-To

Díaz, A. (2017). Why Barcelona locals really hate tourists. Retrieved April 26, 2018, from https://www.independent.co.uk/travel/news-and-advice/barcelona-locals-hate-tourists-why-reasons-spain-protests-arran-airbnb-locals-attacks-graffiti-a7883021.html

Ertz, M., Durif, F., & Arcand, M. (2016, June 23). An Analysis of the Origins of Collaborative Consumption and Its Implications for Marketing. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2799862

Facebook. (n.d.). Graph API. Retrieved April 20, 2018, from https://developers.facebook.com/docs/graph-api

Facebook, & Mobolic. (n.d.). Facebook SDK for Python. Retrieved April 20, 2018, from

http://facebook-sdk.readthedocs.io/en/latest/api.html

Felson, M., & Spaeth, J. L. (1978). Community Structure and Collaborative Consumption: A Routine Activity Approach. *American Behavioral Scientist*, *21*(4), 614–624. https://doi.org/10.1177/000276427802100411

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations – A case from Sweden. *Journal of Destination Marketing & Management*, *3*(4), 198–209. https://doi.org/10.1016/j.jdmm.2014.08.002

GADM. (n.d.). GADM. Retrieved October 19, 2018, from https://gadm.org/

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: foundations and developments. *Electronic Markets*, *25*(3), 179–188. https://doi.org/10.1007/s12525-015-0196-8

Haralambopoulos, N., & Pizam, A. (1996). Perceived impacts of tourism: The case of samos. *Annals of Tourism Research*, *23*(3), 503–526. https://doi.org/10.1016/0160-7383(95)00075-5

INE. (2017a). *Estatísticas do Turismo - 2016*. Retrieved from https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=277048338&PUBLICACOESmodo=2

INE. (2017b). Portal do Instituto Nacional de Estatística - Airbnb. Retrieved October 13, 2018, from https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=316000944&DESTAQUESmodo=2&xlang=pt

Instagram. (n.d.). Instagram Developer Documentation. Retrieved April 20, 2018, from https://www.instagram.com/developer/

Jornal Público. (2018). Lisboa e Porto têm mais turistas por residente do que Londres e Barcelona. Retrieved April 22, 2018, from https://www.publico.pt/2018/04/04/sociedade/noticia/lisboa-e-porto-tem-mais-turistas-por-residente-que-londres-e-barcelona-1809036

Laney, D. (2001). Application Delivery Strategies. *Meta Group*, (September). https://doi.org/10.1016/j.infsof.2008.09.005

Lee, S.-H., Choi, J.-Y., Yoo, S.-H., & Oh, Y.-G. (2013). Evaluating spatial centrality for integrated tourism management in rural areas using GIS and network analysis. *Tourism Management*, *34*, 14–24. https://doi.org/10.1016/J.TOURMAN.2012.03.005

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323. https://doi.org/10.1016/J.TOURMAN.2018.03.009

Li, Y., Xiao, L., Ye, Y., Xu, W., & Law, A. (2016). Understanding tourist space at a historic site through space syntax analysis: The case of Gulangyu, China. *Tourism Management*, *52*, 30–43. https://doi.org/10.1016/J.TOURMAN.2015.06.008

Lim, C. (1999). A Meta-Analytic Review of International Tourism Demand. *Journal of Travel Research*, *37*(3), 273–284. https://doi.org/10.1177/004728759903700309

Lisboa Card. (2018). Lisboa Card. Retrieved April 22, 2018, from https://www.lisboacard.org/

Liu, B., Huang, S. (Sam), & Fu, H. (2017). An application of network analysis on tourist attractions: The

case of Xinjiang, China. *Tourism Management*, *58*, 132–141. https://doi.org/10.1016/J.TOURMAN.2016.10.009

Lozano, C., Flament, I., & Malik, M. (2017). Sustainable Tourism in Florence. Retrieved May 6, 2018, from http://dssg-eu.org/florence/index.html

Luo, Q., & Zhong, D. (2015). Using social network analysis to explain communication characteristics of travel-related electronic word-of-mouth on social networking sites. *Tourism Management*, *46*, 274–282. https://doi.org/10.1016/J.TOURMAN.2014.07.007

Machado, A. (2017). Airbnb e Turismo de Portugal fazem acordo inédito. Retrieved April 22, 2018, from https://www.jornaldenegocios.pt/empresas/turismo---lazer/detalhe/airbnb-com-ligacao-directa-ao-turismo-de-portugal

Mahéo, A., Kilby, P., & Hentenryck, P. Van. (2017). Benders Decomposition for the Design of a Hub and Shuttle Public Transit System. *Transportation Science*. Retrieved from https://arxiv.org/pdf/1601.00367.pdf

Marine-Roig, E., & Anton Clavé, S. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management*, *4*(3), 162–172. https://doi.org/10.1016/J.JDMM.2015.06.004

Martínez, H., Mauttone, A., & Urquhart, M. E. (2014). Frequency optimization in public transportation systems: Formulation and metaheuristic approach. *European Journal of Operational Research*, *236*(1), 27–36. https://doi.org/10.1016/J.EJOR.2013.11.007

Mason, P. (2016). *Tourism impacts, planning and management*. Routledge. Retrieved from https://www.routledge.com/Tourism-Impacts-Planning-and-Management-3rd-Edition/Mason/p/book/9781138016293

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. *Information and Management*, *54*(6), 771–785. https://doi.org/10.1016/j.im.2016.11.011

Neto, M. (2017). "Smart Tourism" – turismo numa cidade inteligente. Retrieved April 22, 2018, from https://observador.pt/opiniao/smart-tourism-turismo-numa-cidade-inteligente/

Nilbe, K., Ahas, R., & Silm, S. (2014). Evaluating the Travel Distances of Events Visitors and Regular Visitors Using Mobile Positioning Data: The Case of Estonia. *Journal of Urban Technology*, *21*(2), 91–107. https://doi.org/10.1080/10630732.2014.888218

Novy, J. (2011). "Berlin Does Not Love You" - Notes On Berlin's "Tourism Controversy" and its Discontents. *The Berlin Reader: A Compendium on Urban Change and Activism*, 16. Retrieved from https://rda69.files.wordpress.com/2014/11/berlin-does-not-love-you-1.pdf

Oracle. (n.d.). What is Big Data? | Oracle. Retrieved October 14, 2018, from https://www.oracle.com/big-data/guide/what-is-big-data.html

Orellana, D., Bregt, A. K., Ligtenberg, A., & Wachowicz, M. (2012). Exploring visitor movement patterns in natural recreational areas. *Tourism Management*, *33*(3), 672–682. https://doi.org/10.1016/J.TOURMAN.2011.07.010

Pforr, C., Volgger, M., & Coulson, K. (2017). *The Impact of AirBnB on WA's Tourism Industry*. Retrieved from http://bcec.edu.au/assets/The-impact-of-Airbnb-on-WAs-tourism-industry-report-web-version.pdf

Plummer, R. (2017). Are tourists still welcome after protests? Retrieved April 26, 2018, from http://www.bbc.com/news/business-40960443

População Empregada 2016. (2017). Retrieved from http://travelbi.turismodeportugal.pt/pt-pt/Documents/Análises/Alojamento/populacaoempregada2016.pdf

Pordata. (2018a). Exportações de serviços: total e por tipo. Retrieved September 14, 2018, from https://www.pordata.pt/Portugal/Exportações+de+serviços+total+e+por+tipo-2352

Pordata. (2018b). PORDATA - Balança de viagens e turismo. Retrieved April 20, 2018, from https://www.pordata.pt/Portugal/Balança+de+viagens+e+turismo-2583

Pordata. (2018c). PORDATA - Resident Population: total and by age groups. Retrieved September 14, 2018, from https://www.pordata.pt/Municipios/População+residente+total+e+por+grandes+grupos+etários-390

Pordata. (2018d). PORDATA - Tourist guests in tourist establishments. Retrieved April 20, 2018, from https://www.pordata.pt/Municipios/Hóspedes+nos+estabelecimentos+hoteleiros+total+e+por+tipo+de+estabelecimento-750

Pordata. (2018e). Produto Interno Bruto por componentes (Euro). Retrieved September 14, 2018, from https://www.pordata.pt/Europa/Produto+Interno+Bruto+por+componentes+(Euro)-2683

PORTUGALPRESS. (2017). "Best European destination": Portugal sweeps board at World Travel Awards. Retrieved April 20, 2018, from http://portugalresident.com/"best-european-destination"-portugal-sweeps-board-at-world-travel-awards

Prideaux, B. (2005). Factors affecting bilateral tourism flows. *Annals of Tourism Research*, *32*(3), 780–801. https://doi.org/10.1016/J.ANNALS.2004.04.008

Quattrone, G., Proserpio, D., Quercia, D., Capra, L., & Musolesi, M. (2016). Who Benefits from the "Sharing" Economy of Airbnb?, 1385–1393. https://doi.org/10.1145/2872427.2874815

Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management*, *57*, 202–212. https://doi.org/10.1016/j.tourman.2016.06.006

Ribeiro, M. B. (2017). *O Impacto do Turismo no Centro Histórico de Lisboa*. Universidade Nova de Lisboa. Retrieved from http://hdl.handle.net/10362/30068

Schoder, D., Gloor, P., Takis Metaxas, P., Gloor, P. A., & Metaxas, T. (2013). Social Media and Collective Intelligence: Ongoing and Future Research Streams, *27*(1). Retrieved from http://repository.wellesley.edu/scholarship

Sequeira, J. (2018). Rock in Riot – música e descontentamento contra a gentrificação de Lisboa. Retrieved April 20, 2018, from https://www.publico.pt/2018/03/23/local/noticia/rock-in-riot--musica-e-descontentamento-1807833

Shi, Y., Serdyukov, P., Hanjalic, A., & Larson, M. (2013). Nontrivial landmark recommendation using geotagged photos. *ACM Transactions on Intelligent Systems and Technology*, *4*(3), 1. https://doi.org/10.1145/2483669.2483680

Statista. (2017). International tourist arrivals worldwide by region 2016. Retrieved April 23, 2018, from https://www.statista.com/statistics/186743/international-tourist-arrivals-worldwide-by-region-since-2005/

Statista. (2018). Tourists in hotels in Barcelona 1990-2017. Retrieved October 18, 2018, from https://www.statista.com/statistics/452060/number-of-tourists-in-barcelona-spain/

Stienmetz, J. L., & Fesenmaier, D. R. (2015). Estimating value in Baltimore, Maryland: An attractions network analysis. *Tourism Management*, *50*, 238–252. https://doi.org/10.1016/J.TOURMAN.2015.01.031

Sullivan, L., & LaMorte, W. (2016). InterQuartile Range (IQR). Retrieved November 11, 2018, from http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata7.html

Sun, S., Wei, Y., Tsui, K.-L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, *70*, 1–10. https://doi.org/10.1016/J.TOURMAN.2018.07.010

Turismo de Portugal. (2018). TravelBI by Turismo de Portugal. Retrieved May 5, 2018, from http://travelbi.turismodeportugal.pt/pt-PT/Paginas/search.aspx?q=travelbicategorias:estatisticas travelbicategoriasbi:alojamento#

UNWTO Tourism Highlights: 2017 Edition. (2017). World Tourism Organization (UNWTO). https://doi.org/10.18111/9789284419029

UNWTO Tourism Highlights: 2018 Edition. (2018). World Tourism Organization (UNWTO). https://doi.org/10.18111/9789284419876

Vallois, T. (2017). OPINION: How life in Paris has gone downhill over the years. Retrieved April 25, 2018, from https://www.thelocal.fr/20171003/how-paris-has-gone-down-the-can-according-to-an-expat-whos-lived-here-for-50-years

Vaughan, R., & Daverio, R. (2016). *Assessing the size and presence of the collaborative economy in Europe*. London. Retrieved from http://ec.europa.eu/docsroom/documents/16952

Walsh, B. (2011). Today's Smart Choice: Don't Own. Share - 10 Ideas That Will Change the World. Retrieved May 1, 2018, from http://content.time.com/time/specials/packages/article/0,28804,2059521_2059717_2059710,00.html

Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, *62*, 120–131. https://doi.org/10.1016/j.ijhm.2016.12.007

Zervas, G., Proserpio, D., & Byers, J. W. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Journal of Marketing Research*, jmr.15.0204. https://doi.org/10.1509/jmr.15.0204

Zheng, W., Huang, X., & Li, Y. (2017). Understanding the tourist mobility using GPS: Where is the next place? *Tourism Management*, *59*, 267–280. https://doi.org/10.1016/J.TOURMAN.2016.08.009