UNIVERSITY OF TRENTO

Department of Mathematics

Master in Data Science

Master's Degree Thesis

SHARING YOUR HOUSE, AT WHAT COST?

Analysis of a pricing model and customer segmentation for Airbnb

Thesis Supervisor:
Luigi Amedeo Bianchi

Co-Supervisor:
Diego Giuliani

Master Student:
Veronica Cipriani

Academic Year 2019/2020

UNIVERSITÀ
DI TRENTO

University of Trento
Department of Mathematics
Master in Data Science

Master Student: Veronica Cipriani

Supervisor: Luigi Amedeo Bianchi

Co-Supervisor: Diego Giuliani

I hereby declare that this dissertation is my own original work and has not been submitted before to any institution for assessment purposes. Further, I have acknowledged all sources used and have cited these in the reference section.

Signature

Acknowledgments

I would like now to thank the people who have helped me undertaking this research.

First of all, I would like to express my deepest gratitude to my supervisor, professor Luigi Amedeo Bianchi, for the continuous kind support and valuable advices. He patiently guided me during the entire process, from the choice of topic until the end of the analysis. The completion of my dissertation would not have been possible without his knowledge and nurturing.

I would like to extend my sincere thanks of my co supervisor, Diego Giuliani, and all the professors that I met during this master. A special thanks also to the University of Trento, for these incredible 5 years.

I am extremely grateful to my parents for always being a source of unlimited inspiration. They have always supported me and my choices, unconditionally, and for this I will always be incredibly thankful.

Last but not least, I cannot begin to express my gratefulness to my friends, invaluable support and inspiration. Thanks for always being there.

Master Thesis - Data Science
Academic Year 2019-2020

Contents

Introduction

It is 2007. Imagine being unemployed and almost broke, struggling with the rent. Imagine having a crazy idea and that some friends decide to give credit to that. Now, imagine that idea being worth 31 billion: this is the story of Airbnb[1].

Right in the middle of the Great Recession, many other businesses and start-up took the first steps of a great success. Uber[2], Kickstarter[3], Couchsurfing[4], BlaBlacar[5]: the lowest common denominator is to create a business model around a precise customer need.

People find answers in these new systems when sometimes they were not aware they had questions and demands in the first place. In fact, give new dignity to idle resources was not a trend until these companies started from ambitious ideas: this is the basic idea of the *sharing economy*. This disruptive movement quickly recasts the entire economic system, and new figures enter the market. Customers start to be both providers and consumers, and without huge investments are therefore able to earn some extra money. Clearly, such an innovation hides different controversy. On one side, the new collaborative system needs to fit in the existing one. On the other, the "sharing economy" wants to inspire people trust and assurance, to grow.

The thesis aims to examine this phenomenon. The Chapter 1 presents the essential concept to understand the *sharing economy*: the comprehensive literature review lays the foundation of the analysis. First of all, the concept itself of SE is defined, together with the impact of technology. An entire section focus on the protagonist of this dissertation: Airbnb and its key to success. Apart from the timeline of its journey, this part draws the relationship between this business and the existing ones - rental market and hotels. Section 1.3 focus on the price in the rental market: from the origin of the hedonic pricing model (traditionally used in the housing retail), the

[1] Airbnb, 2020. Wikipediahttps://en.wikipedia.org/wiki/Airbnb, accessed November 18th, 2020.

[2] Uber, History, Wikipedia https://it.wikipedia.org/wiki/Uber, accessed November 18th, 2020.

[3] Kickstarter, Founder Biography, Britannica https://www.britannica.com/biography/Perry-Chen, accessed November 18th, 2020.

[4] Couchsurfing, History, Couchsurfing Website https://about.couchsurfing.com/about/about-us/, accessed November 18th, 2020.

[5] Sclaunich G., "I primi dieci anni di BlaBlacar", Corriere, 2016, accessed November 18th, 2020.

discussion moves to the complex algorithm used by Airbnb (*Smart prices*). The last part of the chapter, instead, focuses on segmentation, presenting different approaches and reflections.

Chapter 2 entirely focus on the methodology. Firstly, the research question briefly introduces the framework. Overall, both the supply side of the sharing market and the demand ones are explored. The methods implemented to conduct these analyses are later presented. Then, once both the aim and techniques are defined, the data sets are described. At the end of the chapter, the data cleaning process is also precisely described. Particular attention is placed on missing values, outliers and imputation approaches.

Starting from previous studies assumptions, the first part of the analysis entirely covers Chapter 3. As mentioned, the aim is to examine the Airbnb pricing strategy in the Italian context. The data is downloaded directly from the website and completed with new attributes (2.4.1). Different machine learning techniques and statistical tools are implemented to recreate a well-fitting model, and the most interesting results are presented. Starting from the concept of hedonic pricing model, and having analysed its application in the hospitality sector, the analysis stresses the most important attributes in defining the final price. From linear regression, to LASSO an regression trees, the best fitting models are random forest and gradient boosting. For the sake of fluency, only significant results are presented, and the complete analysis is available on Github (link to the repository provided in the *Appendix*).

In the second part, the demand side is considered (Chapter 4). The aim here is to divide people into different groups to identify the typical online-booking traveller. In fact, considering socio-demographic and behavioural variable, but also purchase history, clustering can underline the most distinctive characteristic of each group. By doing so, the author wants both to highlight new potential segments and address purchase preferences. As the focus is on the Italian context, the data is downloaded from the Istat repository. Firstly, the purchase history of travellers is analysed. The *RFM* ("Recency, Frequency, Monetary") is the first applied methods. As the group are considerably unbalanced, the k-means algorithm is also performed on the same data: the results are consistent and offers interesting points. In fact, k-means recognises sort of macro-categories for the six groups identified by the RFM: globetrotters, "cheap travellers" and "luxury" ones. Lastly, the hierarchical cluster algorithm is performed to investigates socio-demographic attributes. Each section provides a summary table and plots to help understanding the results of that precise clustering methods.

Eventually, Chapter 5 concludes this dissertation. The results are fully disclosed

and commented, and the main limitations of the study are also addressed. The last section offers potential starting points for further researches.

Chapter 1

Literature Review

1.1 Sharing Economy

1.1.1 The phenomenon of sharing economy

Over the past years, the term *sharing economy* has caught increasing attention. This topic has been widely discussed in the literature (Agarwal and Steinmetz, 2019; Görög, 2018; Hossain, 2020). Despite the large number of definitions, it is often seen as a mix of practices and information (Hamari et al., 2016; Taeihagh, 2017). *Collaborative economy*, *collaborative consumption*, *access economy*, *platform economy*, *community-based economy* (Belk, 2014): each of these terms captures central aspects of this multifaceted phenomenon. Yet, one of the most complete definitions was derived by Muñoz and Cohen, 2017, who defined the sharing economy as "*a socio-economic system enabling an intermediated set of exchanges of goods and services between individuals and organizations which aim to increase efficiency and optimization of under-utilized resources in society*".

SE businesses are various and have the potential to be applied to a wide range of sectors - namely transportation, accommodation, household services, deliveries, retail commerce, consumer loans, currency exchange, project finance, computer programming (Farronato and Levin, 2015). The great variety of activities included in this model - and the difficulty to calculate their monetary volume - makes it complicated to quantify the real effect. Yet, this phenomenon is likely to become a game changer in the market and generate new ways of creating value for customers in the longer term (predictions quantify the SE as big as 335 billion U.S. dollars by 2025[1]). Therefore, even if not precisely measurable, this disruptive model clearly has a significant influence on every market it touches (Codagnone et al., 2016). In some cases, SE

[1]Statista, Value of the sharing economy worldwide in 2014 and 2025 (accessed September 15th, 2020).

is the re-framing of a business into an improved model, offering a more efficient or low-cost version of a service - or product - already present in the market. This is the case of Airbnb, competing with hotels in the hospitality sector, or Uber, in a controversial battle with taxis services around the world[2]. Other platforms, instead, aim to create services hardly available and open them to the public. In this clearly complex scenario, the common factor is *sharing excess*. Taking advantage of idle resources, SE platforms offer a huge variety of products and services with different characteristics and diffusion. In fact, the aim is often to share the excess capacity of the providers with others who need that particular good or service (Barron et al., 2020; Belk, 2014; Hossain, 2020). This dynamic system guarantees the ability to satisfy a large share of consumers (Zervas et al., 2016). In this sense, the model has the form of a peer-to-peer, collaborative economy, where the two main parties involved are a service/goods provider and a receiver/buyer. Having considered different business models and sectors, Botsman and Rogers (2010) distinguished between three types of SE, with regard to the goods for sale:

- a redistribution of used or pre-owned goods;
- a product service system where consumers pay for access to the good as a service rather than purchasing the good (access over ownership);
- the sharing of non-physical assets, based on a collaborative lifestyle.

Having depicted the different forms of this phenomenon, heterogeneity can be easily detected as one of its main features. In fact, sharing economy counts many qualities that deeply define it as a unique phenomenon, as highlighted in Eckhardt et al. (2019). First of all, the goods or service are temporarily accessed by the customer, who does not gain a permanent right on the purchased item. The second characteristic is the economical implication. In the exchange, one part pays the other a specific amount of money for the good or service, as with every transaction. As clearly specified in Belk (2014), monetary compensation is part of the definition of SE itself, as pure sharing and bartering are not considered an example of this model. Third, the peer-to-peer activity always relies on a platform that matches users and providers. For example, BlaBlaCar could not exists without the app matching occasional drivers with potential clients. Fourth, peer-to-peer companies form their supply by adding up the forces of individual consumers. As a result, the fifth and last central characteristic is that consumers have an actual role, as they are sometimes part of the suppliers as well. In this sense, consumers often switch to the role of *prosumers*[3], which usually means that users do not provide the commer-

[2]Smorto G., 2017, *Caso Uber, l'impatto su tutta la sharing economy*, Il Sole24Ore (accessed September 15th, 2020).

[3]Cambridge Dictionary, prosumer: *a customer who helps a company design and produce its products. The word is a portmanteau of the words "producer" and "consumer"* (accessed September 15th, 2020).

cial platform, but they act in this environment either as one-sided users or as both providers and consumers (Lang et al., 2020).

Another significant distinction refers to the last mentioned element. As Botsman and Rogers (2010) did with regards to the purchase side, Demaray (2015) gives a precise definition of the business model on the provider side. In fact, in the sharing economy there are mainly *peer-to-peer* (P2P) and *business-to-consumer* (B2C) models. In the first case, as already mentioned, individuals play the role of both providers and consumers. In the latter, the company that provides the platform also supply the service.

In this scenario, the impact of technology is undeniable. Another central characteristic of the sharing economy is precisely the massive use of IT. Certainly, its rise and spread have greatly boosted the growth of collaborative consumption. The emergence of the internet and digital platforms has brought up new ways of sharing and delivering both electronically transmittable services and physical ones (Belk, 2014; Codagnone et al., 2016; Hamari et al., 2016). IT-based e-commerce system has heavily reduced transaction costs, while the use of big data has made unmistakably easier the rise and improvement of peer-to-peer activities. Together with this great revolution in information technologies, the convergence of the financial crisis in 2007-2009 and the growing environmental concern gave the final push to this disruptive new model (Segato, 2016). In fact, users love this business idea, because - without making huge investments, or quitting their current job - they are able to earn some extra money (Benkler, 2004; Segato, 2016; Zervas et al., 2016). Researches have also investigated the possibility to generate a livable income from Airbnb, but geographical differences do not allow to present any generally valid conclusion (Fabo and Hudá, 2017). The great adaptability of the collaborative economy is actually one of the biggest strength of this business model[4]. But at the same time, people that share are giving a new life to idle goods, or taking advantage of the possibility to offer a service. Many studies have underlined that this approach has also an enormous resource-saving potential (Hamari et al., 2016; Mi and Coffman, 2019; Wu and Zhi, 2016). However, despite the generally positive impact of SE on the environment, one must not be caught up in the excitement, and researchers warn that the uncontrolled growth of this phenomenon could cause serious harm[5].

[4]Lees N., 2020, *What's mine is yours: The sharing economy will have to change* (accessed October 1st, 2020).

[5]Yeo S., (2018), *The SE helps fight climate change (but not as much as you think)*, The Washington Post (accessed September 15th, 2020).

1.1.2 Incentives in the sharing economy

The success of an online platform is directly connected with the ability to compete efficiently against other platforms, retailers and traditional service providers. Pricing structures and regulations need to be precisely modelled. As mentioned in the previous paragraph, the monetary transaction is an essential element to categorize a business as a part of peer-to-peer, collaborative economy. For the sharing to actually happen, each agent balance the costs and benefits of the different modalities of use and exchange of goods - or services (Benkler, 2004). Thus, since both provider and user want to get satisfaction from the trade, the two parts need to meet halfway. For this reason, the first element to be considered is the intrinsic problem of incentives: how to convince buyers and sellers to paticipate? While the consumer obtains benefits from the purchase, the provider incurs in costs related to the use of the sharing platform and the provided performance. To mitigate this highly unbalanced situation, incentives systems are implemented.

In literature, a soft scheme and a hard one are suggested (Park and van der Schaar, 2010). The first one requires a rating system, while the hard scheme uses monetary payment to reward the providers.

1.1.2.1 Hard incentives: Price

The first approach considered to encourage participation consists of modelling the hard scheme of incentives. A central element to be considered while setting a price is the relative value of the transaction. As in any other market operation, it is fundamental that the supply sets a price that is in line with the amount of money that the demand is willing to pay. An example of how this trade-off can be balanced is the mechanism of auction (Farronato and Levin, 2015). From an economical point of view, this is the most successful way to close a deal, since both parts can withdraw at any moment - if the price goes up too much (from the buyers' point of view) or if it settles to a very low level (from the seller side). The result is a perfect equilibrium between supply and demand. But in the real world, this process is far from perfect. Since auctions are highly time-consuming, nowadays many platforms have decided for the more efficient, posted price. Again, the fixed price is the result of the trade-off between efficiency and transaction cost. When pricing their products, providers want to ensure adequate profitability. Yet, as in the world of sharing economy producers may not have a real professional experience in the market they are playing in, the estimated price can mismatch the hypothetically perfect price. This gap can be due to lack of experience or inability to read information, and may require an external activation to refine the price choice. Some example of this corrective mechanism can

be found in Airbnb[6] and Uber[7,8] pricing models.

The final price offered to the consumer is also influenced by the amount of money providers pay to the sharing platform. In theory, any pricing model can be adopted. In reality, however, sharing platforms decide for simple strategies. Traditionally, two models are recognized as the most popular (Kung and Zhong, 2017). The *membership-based pricing model* charges users with a fixed fee, to access the platform itself in all its parts. The single transaction loses in value, but the amount of money paid at the beginning of each membership period counterbalances the lower price per purchase that the platform receives. The other approach requires the consumer to pay for each transaction, while no membership fee is charged (*transaction-based pricing strategy*). The combination of these two strategies introduces a third pricing model: the *cross-subsidization strategy*. In this last approach, the platform itself recompense the providers for each transaction, and therefore becomes profitable thanks to membership fees. In the same research, Kung and Zhong (2017) prove how these three approaches guarantee the same number of consumers and equivalent profit. Considering the fixed costs and desired profit, providers finally set the posted price.

While considering the aforementioned mechanisms, another central element is the dynamism of price in the sharing economy. Filippas and Gramstad (2016) precisely described the typical process for pricing, that adjusts the value required for the transaction according to popularity. Reviews, trend and consumer awareness influence the positive and negative variations, which can be quite significant in an online environment. The traditional markets work in a similar way, responding to the supply-demand mechanisms, but given the more limited capacity to adjust to the market, their "tuning" process is slower. Studies prove that SE prices are highly competitive in the market - see the example of *Airbnb*, a colossus in the SE presented in the next paragraph of this chapter. This is one of the greater advantages of this system, that is often able to offer major benefits in comparison to traditional markets: lower prices and broader options (Roma et al., 2019; Sundararajan, 2016).

With regard to the traditional market, we have already mentioned the disruptive effect caused by the sharing economy. This new agent affects also "traditional" prices: finding themselves competing with online opponents, markets experience different effects according to their characteristics (Roma et al., 2019): the price can be positively adjusted, while in other cases it drops because of competition. However, the strategic benefits that sharing platform can capitalize has some limits. The

[6] Airbnb Answers: pricing suggestions (accessed September 17th, 2020).

[7] Uber, Surge pricing: What's happening when prices surge? (accessed September 17th, 2020).

[8] Gurley B., 2014, *A Deeper Look at Uber's Dynamic Pricing Model. Above The Crowd* – online edition(accessed September 17th, 2020).

pervasive power of internet retailing has been less performing than expected, despite pricing advantages (Grewal et al., 2004). In fact, the analysis on the *willingness to pay* of consumers by Chatterjee and Kumar (2017) detects how differences in situations and type of purchases affect the maximum amount of money a potential buyer has decided to spend, and where they decide to spend it. Online markets certainly have the possibility to create a rich and multifaceted environment, with many different options that can be associated with the idea of perfect competition. But at the same time, goods that need to be touched or tried are more challenging for online stores, and retail stores can try to fill the gap. Personal attention and suggestions and post-purchase services are also excellent opportunities for a physical shop, especially for expensive and high-risk purchases. In fact, the perceived risk connected with online shop persists as one of the main motivators in the definition of the willingness to pay.

1.1.2.2 Soft incentives: Reviews

The second approach has its roots in the general concept of trust. At the very bottom, the key point in the relationship between providers and consumers is that the first needs to inspire reliability on the latter. The soft scheme, as presented in Park and van der Schaar (2010), is the second, fundamental ingredient for a successful sharing platform. The revolution of the peer-to-peer economy has its root in this concept: this phenomenon has in fact legitimized trusting stranger[9] by creating a safe space where this is actually possible. But at the same time, for the peer-to-peer economy to exists, people need to take part in it. In literature, different motivators are investigated, and the relative importance of economic, social and environmental factors are measured (Böcker and Meelen, 2017). The platform is the one and only actor that governs the exchange and distribution of information between the two sides. In the first place, it needs to offer a convenient condition to providers, whether they are meant to supply a long-run arrangement or just occasional deals (Farronato and Levin, 2015). The higher the number of providers - and therefore, of variety and quantity of products - the more the possibility that consumers join the platform. As a direct consequence, the increasing number of consumers attracts more providers on the platforms, as there are more possibilities to close deals. This circular phenomenon is defined *cross-side network externality*[10] (Kung and Zhong, 2017).

The essential starting point for a sharing business is therefore attractiveness, and the platform providers need to prove its consistency to providers in the first place,

[9] Botsman R, 2016, *We've stopped trusting institutions and started trusting strangers*, Ted Talk, (accessed September 6th, 2020).

[10] https://guides.co/g/the-network-effects-bible/121735 (accessed September 17th, 2020).

but also inspire buyers with trust. Once users get to the platform, engagement is the next vital step. Many online cases prove gamification to be a successful incentive, for example with badges and quality guarantee (Li et al., 2012). At the same time, providers find themselves in the position when they hardly control awareness - since little personal advertising can be done (Filippas and Gramstad, 2016).

In this context, the rating system provided in the soft scheme creates a sort of history that helps the consumer understanding the level of reliability of the shopper, therefore enabling them to trust - *and buy from* - strangers. Many studies have proven how customer reviews are trusted, therefore boosting sales (N. Hu et al., 2008; J. B. Kim, 2014; Lawani et al., 2019). What is absolutely counter-intuitive is that this positive effect on purchase comes both from positive *and* negative reviews. In fact, whether a review negatively or positive outline a product or service, what really makes the difference is the level of details. The quality of a review - rather than the rating or the total number of comments - is the turning point. The more customers perceive a subjective point of view, the more they think they can make a conscious choice (Berger et al., 2010; Ghose and Ipeirotis, 2011). For this reason, a successful review system often corresponds to a trustworthy sharing platform.

1.1.3 The tourism sector in the sharing economy

As mentioned, the disruptive power of sharing economy is hardly measurable, despite the large number of studies focusing on the shock caused in the market by this new player. Having mentioned all the different sectors that collaborative economy entered, this section will focus on the effects on the hospitality industry, which is by far the most affected - and the one that has mostly benefited - from this innovation.

In 2017, tourism gross revenue reached $1.6 trillion, based on bookings (Hong, 2018). The following year, travel sales worldwide were worth more than $ 560 billion[11]. The positive trend of this sector is confirmed by the World Travel & Tourism 3 Council in their last Economic Impact Report, as "WTTC's latest annual research, in conjunction with Oxford Economics, shows the Travel & Tourism sector experienced 3.5% growth in 2019"[12]. 2019 really was a success for the tourism sector, and this great growth is confirmed also by other sources - UWTO[13]. The same organization also forecast a +3,+4% in 2020, confirming the positive trend - before the drop due to Covid-19 global pandemic.

Traditionally, all the main services connected with this sector have been provided by tour operators, by hotels and B&B with regards to lodging, by airlines and rail

[11] Statista, Online travel market. 2019 report (accessed September 20th, 2020).
[12] WTTC, Economic impact report 2019. (accessed September 20th, 2020).
[13] World Tourism Organization, Tourism Barometer, January 2020 (accessed September 20th, 2020).

when dealing with transportation.

Nowadays, the scenario is completely different. Despite the large growth, the role of many of these subjects has been downsized and they are not a key player in this market anymore. So, as the sharing economy and the hospitality industry are growing together, the revolution of P2P is also changing and shaping the way of doing tourism[14]. In general, many of the services that can be offered between peers are seen as direct alternatives to those that are provided by professionals in the hospitality industry[15]. Namely, home-sharing, home exchange and connected service - like dining services and cleaning services - are the main categories of sharing economy contribution to tourism. At the same time, a multitude of other services and segments are directly connected with the hospitality industry. For example, transportation is considerably influenced in all its forms - air travel, rail, cruise, car rental - but also with new, innovative competitors from the sharing economy - such as car sharing.

On the other hand, the relationship with the digital world represents another important element in the success of tourism. As mentioned, digitization has marked a massive change in different aspects of everyday life in the last decades. Online platforms enable many to easily access the increasing volume of information available on the net. At the same time, social media enriches this scenario even more, offering a variety of tools that opens to everyone the real possibility to access numerous new options for travelling. The massive volume of information offers the potential tourist a wider choice, both in terms of destinations and awareness. However, the planning process becomes incredibly complex precisely due to this overload of information (Nikoli and Lazakidou, 2019; Sciarelli et al., 2018). The search for information has become much simpler, but balancing all the factor in the perfect, final choice is challenging. In this scenario, as digitization creates a problem, it also provides a solution. While travel organizations rely on digital tools and take advantage of this new possibility to attract new tourist, the turning point in this net of information is being able to evaluate who and what to trust. Social networks, travel communities, blogs and other platforms are trying to answer this new need. Again, in the era of sharing economy, trusting strangers is fundamental[16].

Tourists constantly ask for personal advice or search for online reviews before booking. As selecting reliable information from the internet can be problematical, the exchange of information and personal experience between consumers become a priceless supports in planning trips and holidays (Pan et al., 2007; Schmallegger and Carson, 2008). The rise of *user-generated content* (UGC) has an incredibly important

[14] Grüman R., 2018, *Sharing Economy is Changing the Tourism Industry* (accessed September 20th, 2020).
[15] Ferrer R., 2018, *The sharing economy and tourism* (accessed September 1st, 2020).
[16] See 8.

role in the online world, and particularly in the hospitality sector. Social media and blogs have the intriguing ability to move customers' preferences and - therefore - enter their decision process. As aforementioned, advises comes no longer just from family and friends, but also strangers: trusting people online can make it incredibly easier to understand what you are purchasing[17]. Many pieces of research show an increasing number of people consulting these online reviews (Andersson, 2010), and highlight the role of word of mouth in the dynamics of purchases and the definition of the final price as well (Babić Rosario et al., 2020). The proliferation of *UGC*, in fact, has created a whole new phenomenon, where consumers share their experience and opinions, therefore spreading *electronic word-of-mouth*. The positive attitude of a potential customer depends on many factors, but the quality of online information and the perceived reliability of reviews are game-changers. Having depicted the essential role of user-generated content and electronic word-of-mouth, trustworthiness and reliability will be further analysed in subsection 1.3.3.1, as well as influence on the pricing model.

As clearly demonstrated, ICT plays a critical role in *e-tourism* - that is "the digitisation of all the processes and value chains in the tourism, travel, hospitality and catering industries that enable organisations to maximise their efficiency and effectiveness" (Wahab, 2007). In addition to the phenomena previously illustrated, studies identify other forms of innovations connected to digitization. First of all, the web presence of tourism destination strategically plays an important role (Mich, 2013). Internet is an invaluable tool, yet it can be the source of hidden danger: successfully branding the tourism product and diversifying from competitors decide the fortune of the entire business. Diversifying is, in fact, the second great revolution of digital tourism. Personal marketing and personalized proposal offer potential consumers the best fitting product, increasing customer satisfaction (Mitrović et al., 2019). Segmentation is a very valuable and interesting technique to guarantee a great understanding of different groups of customers. While customizing a product, defining different prices according to the willingness to pay offers strategic advantages as well. Tourism products are often intangible, experience-based goods that are difficult to evaluate before purchase (Casaló et al., 2015), thus the effective use of price lever assures the highest possible profit, as the model adapts to price elasticity (Nieto-García et al., 2017). Pricing is probably one of the most sensitive issues for customers, especially in the multifaceted framework just depicted, therefore the quality/price relationships require special attention.

To conclude, another essential element has to be analysed: in the mixture of *tourism* ingredients, transportation is paramount. Together with the digital revolution, an-

[17] Gans J., 2011, *The Rise of Content Platforms*, Harvard Business Review.

other aspect to consider with regards to the great success of the hospitality industry is in fact the level of transport facilities. The great improvement of transport services, combined with competitive prices, attracts every day an increasing number of tourists. In fact, transportation is one of the most important contributors to tourism success: the speed of domestic trips, as well as the ease and relatively low prices for international mobility, are essential ingredients in the mix (Gierczak, 2011). The need to ensure accessibility depends upon multiple factors: not only the ability to reach places all around the globe but also a decent level of transport services once in the destination. The close relationship between tourism and transport facilities is evident: the success of a destination is partly ensured by efficient transportation (Celata, 2007; Curry and Falconer, 2014). As aforementioned, the sharing model of peer-to-peer economy can be potentially applied also to transportation. The literature mainly focuses on ride-sharing, its limit and the regulation (Standing et al., 2019). However, while the impact on the hospitality industry is wider and covers many aspects of this sector, sharing service transportation is usually limited to short distances. Therefore, the impact of sharing economy on transportation is still quite unclear. What is sure is that future trends in mobility will definitely affect tourism as well (Franckx and Mayeres, 2016).

1.2 A successful example: Airbnb

The second section of the review focuses on a successful example in the sharing economy: Airbnb. This company is unmistakably one of the most positive examples of understanding and adapting to changes. Being a distinctive mark of the collaborative economy, *sharing* - and sharing in the lodging sector - predate the internet (Belk, 2014). One of the greatest achievement of Airbnb is that its platform perfectly adapts to these new needs of customers. As the customers want to be the protagonist of both demand-side and supply process (Prahalad and Ramaswamy, 2004), they decided to invest their client with both roles.

1.2.1 What was the key to success?

Despite the great growth, Airbnb had a troubled beginning. A brief overview of the history can help to understand the turning point of this business[18]. Back in October 2007, the Airbnb idea first came to light[19,20]. The business started as an air-mattress renting-service, as a way to make some extra bucks for three roommates struggling

[18] Airbnb, *Abount us* (accessed October 2nd, 2020).
[19] Wikipedia, *Timeline of Airbnb* (accessed October 1st, 2020).
[20] Wikipedia, *Airbnb* (accessed October 1st, 2020).

with San Francisco rent. After the first *home made* experience, Brian Chesky, Joe Gebbia and Nathan Blecharczyk decided to give this idea a chance: they put up and launched a website, that was online on August 11th, 2008. The first customers arrived, yet the beginning was hard. The breakthrough was at the beginning of 2009: Y Combinator noticed the project and decided to invite Air Bed&Breakfast - as known at the time - into the prestigious startup accelerator. Then, the turning point: after some month, West Coast investors noticed the company[21]. From that moment on the company started to grow vertically.

Being one of the most outstanding companies in the collaborative economy, thousands of tourists decided to take a leap of faith and trust the business. However, the bigger the success, the more complex the problems. Airbnb was experiencing great appreciation but started also to draw attention from local authorities. Policymakers were not expecting this disruptive power to enter the market, and the answer from local authorities has taken different approaches (D. Guttentag, 2017). In response to this unstoppable phenomenon, city laws sometimes made it illegal to rent out your unit without being present for less than thirty days. Every night, ordinary people were renting their spare room, unaware that a large number of these accommodations were illegal. As the regulatory conflict became an ordinary issue in the early years of the company, so did the legal trouble[22]. To please cities, they also started to collect hotel taxes and promised to give insight into their data.

Despite the legal issues, the ground rule held: to keep the *belong anywhere* promise[23]. The numbers, in fact, shows a worldwide success: nowadays there are more than 7 million listings, more than 100 thousands cities spread all over the world, covering more than 220 country and regions[24]. On the tourists' side, Airbnb counts an overall number of more than 750 million guests, with an average of 2 million per night. As already mentioned, the great appreciation from customers directly originates from the great capacity of the company to adapt to their specific needs. Indeed, one fortune of this business is probably great timing. While the concept of value is mutating, Airbnb understood that ownership was not essential anymore. In the battle between ownership and access, the rise of sharing economy clearly states that the latter is overcoming the first[25,26]. Nowadays, people value more experience, and access to service trumps possession (Oskam and Boswijk, 2016).

[21] Aydin R, 2019, *How 3 guys turned renting air mattresses in their apartment into a $31 billion company, Airbnb* (accessed October 1st, 2020).

[22] Coldweel W., 2014, *Airbnb's legal troubles: what are the issues?* The Guardian (accessed October 1st, 2020).

[23] Chesky B., 2014, *Belong Anywhere* (accessed October 1st, 2020).

[24] See 18.

[25] Colao J. J., 2012, Welcome To The New Millennial Economy: Goodbye Ownership, Hello Access (accessed October 2nd, 2020).

[26] Gidopoulos Y., 2019, Access vs Ownership: Really a Revolution? (accessed October 2nd, 2020).

Luck and timing apart, another essential ingredient in the mix is trust. "A well-designed reputation system is key to get it right" says one of the founder, Joe Gebbia[27]. In his speech, he precisely illustrates the design of trust in the company and underlines that the platform highly encourages hosts and guests to leave reviews. Lately, they also modified the review procedures: Zervas et al. (2015) argue that Airbnb ratings are above the average, compared to other platforms. Therefore, they encourage guest to be honest, for new potential clients to feel they are approaching a transparent service. Moreover, they considerably enhance security, through identity verification systems. Technology has clearly created new schemata in which people need to fit, yet the mechanisms that support customers make it easier to trust the online world than governments and institutions. As mentioned while presenting soft incentives in 1.1.2, a well-designed information system can significantly reduce the information asymmetries due to not knowing anything about the host. Reviews, ratings and comments have therefore a great impact on purchase and booking (Lawani et al., 2019) - see 1.3.3.1 for further details.

To conclude, according to Pralahad and Ramaswamy (2004), in the era of *prosumers*, a collaborative platform needs to apply the so-called *DART model*: dialogue, access, risk management and transparency. These four blocks ensure a great result. At first, *dialogue* and interactivity maintains a loyal community and gives customers exactly what they want. Then, *access* to information on the company is a priority, as it allows *risk management*. The more a business gives insight not only on data but also on the methodology to assess personal risk associated with products and services, the more people can trust the company. Therefore, *transparency* is the last turning point in the process. In fact, by reducing information asymmetry - which companies traditionally benefit from - the business can gain in trust and loyalty. This is the reason why, as previously underlined, Airbnb decided to design for trust and transparency.

1.2.2 Airbnb, hotels and private sector

The disruptive power of the sharing economy has been widely discussed. In the housing industry, the effect is similar, as key players like Airbnb has a significant role in shaping the supply side of the market: the rise of house sharing has determined a turning point. In fact, this out of many companies has become a threatening competitor for the traditional hospitality industry (D. A. Guttentag and Smith, 2017; Nowak et al., 2015; Oskam and Boswijk, 2016). Therefore, many researchers spend their time studying the reason why tourists choose not to stay in traditional accommodation for their trips.

[27] Gebbia J., 2016, How Airbnb designs for trust (accessed September 6th, 2020).

As already mentioned, Airbnb business is not always well-welcomed, to such an extent that some cities decided for strict regulation on short-term rentals. This is often due to the fact that, as some private individuals switch to home-sharing, the market changes and there is less offer for long-term renting. As a consequence, the house renting experiences an increase in price for residents (Barron et al., 2018), which can lead to serious consequences for city administration. What city administration should consider, however, is the propensity of owners to rent their properties for short-term rather than long-term, when they are given the chance to choose. The overall effect of Airbnb and, more in general, of home-sharing, has its roots in the choice made by hosts: in the end, this is mainly a matter of costs and benefits. After the financial crisis, the increased taxation on properties really put pressure on homeowners. Thus, the possibility to generate extra income through the sharing economy sound very appealing (Roma et al., 2019).

While Airbnb settles in the rental market, the hotel industry is affected as well. In the past years, the traditional businesses in the hospitality industry have grown, and hotels have certainty benefited from the trend. Initially, house-sharing platform mainly addressed long-term tourist and did not touch the share of hotels and B&B. However, as these platforms started to grow and attract both more hosts and more guests, the increasing number of listings is convincing these companies to expand as much as possible (Saussier, 2015). There is still little evidence that house-sharing will disrupt the hospitality industry, and the debate counts divergent positions. One of the CEO of the company firmly denied the competition, as he suggested that Airbnb attracts travellers on the longer-term[28]. On the other hand, media[29,30,31] and scholars share the same opinion: Airbnb - and any sharing platform, weighting for their power - endangers hotels revenues (D. A. Guttentag and Smith, 2017; Nowak et al., 2015; Oskam and Boswijk, 2016; Zervas et al., 2016). Delving into this hypothesis, Roma et al. (2019) analyses the Italian market and find evidence that the impact on the hospitality industry is not uniformly distributed. One of the main advantages of Airbnb is certainly the competitive price: the smaller number of fixed costs for hosts (that barely has barriers to enter the platform) and nature itself of the listing (an otherwise idle resource) control the final rate. On the other hand, traditional businesses incur in higher cost, therefore requiring higher fares to ensure profitability. Precisely for this reason, hotels incumbent are not equally affected, as lower-end hotels suffer more sharing platform competition (Roma et al., 2019;

[28] Business Insider Intelligence (2017), Airbnb CEO speaks on disrupting hotel industry (accessed October 2nd, 2020).

[29] Harvard Real Estate Review, 2019, A New Era of Lodging: Airbnb's Impact on Hotels, Travelers, and Cities, Medium (accessed October 2nd, 2020).

[30] HBS Working Knowledge, 2018, The Airbnb Effect: Cheaper Rooms For Travelers, Less Revenue For Hotels, Forbes (accessed October 2nd, 2020).

[31] Griswold A., 2016, It's time for hotels to really, truly worry about Airbnb (accessed October 2nd, 2020).

Zervas et al., 2016). In the end, the type of consumers looking for accommodation decide the final effect, and they do not see Airbnb as a competitor when looking for high-end hotels. Despite the common belief that radical technological innovation would disrupt the market, each case has its own specific peculiarity. As Xie and Kwok (2017) stated in their work, being able to adapt is a key factor for incumbents to save performance and profitability.

Having depicted the panorama of leisure tourism, business trips are also an important share in the market. Airbnb guests are tourists, visitors and - in small part - business travellers. Despite the limited numbers of the latter, one of the innovations introduced by the company in the years is *Airbnb for work*[32]. From short trips to a longer stay, this service covers any needs of employees: not only classic listings but also work-friendly homes for team works and boutique hotel. In his literature review, Guttentag (2019) also presents the *"For work trips"* filter that points out the most appropriate listings for business. As mentioned, the overall number of business travellers is not very significant: back to 2015 and 2016, in a span of seven months, a total of 50 thousands employees of approximately 5 thousand companies. However, the company report a growing level of satisfaction between customers[33]. Certainly, the company has a positive margin for potential growth.

1.2.3 The case study: Airbnb in Italy

The previous paragraphs give an overview of the model and core business of Airbnb. However, the company has inherent quarrel that partially derives from the incredibly quick growth, year after year. While becoming successful, more and more problem appeared, and they were not always ready to properly face the problem. Starting from security, both from the host and the guest side, Airbnb has to face and manage daily challenges to its model. In this sense, reviewing and rating process on one side, 24/7 support lines and the *Host Guarantee*[34] tries to solve security issues.

Legal controversies are another burning issue for the platform. Zooming on the Italian case, tax fraud and unfair competition are serious problems threatening the business itself (Rolando, 2018). In 2018, Federalberghi published a report denouncing the number of the so-called *Shadow Economy*. Even though Airbnb is presented as one of the biggest examples in the sharing economy, the organization claims that some pillars of the business are in fact big lies. Data at hand, Airbnb seem to be the one and only source of income for a lot of hosts, some of whom count up to 4000 listings. Moreover, usually, owners do not share the experience with guests, as most

[32] Airbnb, Airbnb for Work (accessed October 3rd, 2020).

[33] Airbnb, 2019, Airbnb for Work Uncovers Surprising Trends in Needs of Modern Workers (accessed October 3rd, 2020).

[34] Airbnb, Airbnb's Host Guarantee, (accessed October 24th, 2020).

ads on the website are for entire apartment[35], and they rent out their property for more than just occasional period, as claimed by the company. According to Italian law, short leases apply the *cedolare secca*[36], preferential taxation of 21%. However, tax fraud is estimated for a total amount of 200 million euro in the first year only, pointing out a critical issue.

Legislation on the phenomenon is neither uniform nor clear. Local regulation, in fact, assumes different position towards Airbnb: what is allowed in some cities is forbidden in others[37]. The novelty of this business model certainly pose various dilemmas to the authority, but there is an urgent need for more clarity and regulations.

1.3 Pricing the lodging sector

As widely discussed in the previous sections, one of the main strategic advantages of Airbnb over traditional businesses in the hospitality industry is the price. The platform has fine-tuned a complex pricing mechanism over the year, providing hosts with a variety of supportive tools to derive the optimal final rate. Having disclosed the intricate relationship with both hotels and rental market, this section discusses the traditional pricing model for the rental market. The Airbnb model is analysed, to compare the two strategies and highlight possible similarities. From the traditional hedonic model, this section marks the way towards dynamic pricing.

1.3.1 Price in the rental market

The debate over the importance of real estate as a central sector in the economy has always attracted many scholars. Trends, characteristics and cycles have been widely observed in the last decades, and the resulting studies have often undercovered a positive relationship between housing and economic development (Harris and Arku, 2006; Pu and Zhao, 2018). In particular, prices appear to be the real fuel to economic growth - or, on the other side, the cause of the decline. In their research, Miller et al. (2009) reveal the close relationship between house pricing and economy. Setting the right price in the lodging sector is therefore not only convenient from an individualistic, short-term point of view, but also a broader one. For this reason, the aim of this section is to investigate the pricing model in the rental market, to understand the most important feature to be considered when setting the final price. The findings laid the foundations for the comparison with the sharing economy model.

[35] 76.68% in 2018, as reported in Federalberghi, 2018.
[36] DL n. 50/2017. Containing the so-called "regime fiscale delle Locazioni Brevi Turistiche".
[37] *See* Federalberghi, "COSA ACCADE ALL'ESTERO".

In general, pricing the housing market has always been very challenging, as it needs to combine a variety of heterogeneous characteristics. The final value is influenced by a multitude of features of the building, and - as already mentioned while describing the relationship between price and economic growth - neighbourhoods can play an essential role too. Eventually, time and general economic situation need to be considered as essential ingredients in the final price, as well. Moreover, the actual value perceived by customers is only really captured once the house is sold, as underlined in the work of Nagaraja et al. (2011). For this reason, a common method to underpin pricing trends, and therefore predict them, employ historical data of houses sold multiple times (Bailey et al., 1963; Nagaraja et al., 2011.

Clearly, the debate on the phenomena and structural characteristics that drive price fluctuation is rich and varied (Bragoudakis et al., 2016; Knoll et al., 2014). Traditionally, the *hedonic pricing method* has been a widely used tool in the econometric studies of the urban housing market (Epple, 1987; Rosen, 1974). Other methodologies have been successfully implemented, and in some cases, predictions are even more accurate than standard hedonic regression (Horowitz, 1992). Broadly speaking, the housing price level formation needs to consider both microeconomic and macroeconomic elements (Gaspareniene et al., 2014). Yet, the *HPM* is firmly established in the literature and is undoubtedly one of the most popular approaches. Researchers have employed many different functions through the years (parametric, non-parametric and semi-parametric models) (Owusu-Ansah, 2013), and although the hedonic pricing method does not have a pre-defined functional form, Rosen (1974) suggests a non-linear relationship between price and house as the most suitable. This approach consists of examining each dwelling as a commodity built up by different attributes. The idea behind the hedonic regression is in fact that commodities are basically the sum-up of their characteristics, hence their value can be computed adding the estimated values of these features (Herath and Maier, 2010). Therefore, any location is described as a vector $z = (z_1, z_2, ..., z_n)$, z_i being the level of the $_i th$ characteristic on a total of n features. The price function $P = p(z)$ is simply described as set out below:

$$p(z) = p(z_1, z_2, ..., z_n).$$

and the final value P depends on the form of p. The hedonic price function, therefore, establishes a relationship between the expenditure and all the different characteristics of the house itself. In fact, the final value of a house is derived as a function of its attributes, and therefore the main assumption of this model is that properties could be divided into different characteristics: structural feature, location and infrastructures, social and natural environments. Some recurrent features are listed in

Malpezzi (2003): number of rooms, square footage, category of the house (single or multifamily, number of floor, etc.), heating or cooling systems, age of the building, structural feature and quality of finish. As clearly explained, this function takes into consideration two different sets of attributes: characteristics of the residential structure and location elements (Can, 1992). The importance of position is justified by the accessibility: usually, the closer to the city centre - distance from the *Central Business District* - the easier to access services and places (Herath and Maier, 2011). Due to its architecture, one of the intrinsic characteristics of HPM is also its biggest flaw, as the model requires large data set with very detailed info to be precise (Hamish, 2018). Therefore, all the structural characteristics add up to the final price. Namely, two houses with the same characteristics - and also, similar location - usually have more or less the same price. However, literature shows how similar houses with the same attributes and located close to each other can be valued differently (Maury and Tripier, 2010). This leads to another key factor in deriving the final price, that is the market. Buyers and sellers are, in fact, the agents that can move up and down the final selling price or the rent of a house.

To sum up, there are some characteristics that (Hamish, 2018; Iacobini and Lisi, 2013; Rosato et al., 2017):

- Location characteristics - which include environmental and urban quality, location, distance to amenities and from the centre;
- Technical characteristics - which comprehend all the structural properties of the dwelling (such as dimension, architecture typology, finishing, maintenance, floor area, number of rooms, ...);
- Economic characteristics.

Clearly, the picture is complex and pricing the lodging sector is a delicate issue that requires a considerable amount of information. The market of hotel rooms and is similar to the renting, but more variables are to be considered (Andersson, 2010; Gibbs, Guttentag, Gretzel, Morton, et al., 2018).

1.3.2 Airbnb prices

Scholars have investigated the effect of different features of a listing as pricing determinants in online accommodation platforms - such as Airbnb. In particular, different studies (Cai et al., 2019; Choudhary et al., 2018; Dogru and Pekin, 2017; Olimov, n.d.; Perez-Sanchez et al., 2018) test for the effect of reviews, ratings, host photos, location, number of rooms and bedrooms, but also cancellation policies, bathrooms, type of room and type of listing, distance from the centre. As mentioned in various sections, price is one of the main agents of Airbnb ability to compete in the mar-

ket[38]. Airbnb is competing with different players within the hospitality industry: getting the price right is therefore essential, as it basically determines the success of the short-rent posted offer. The listing options on the platform have undeniable cost advantages over hotels, as rates have been reported as about 30-60% cheaper than the ones offers by traditional hospitality businesses, all around the world (Hong, 2018).

As mentioned, the hedonic model has been both in real estate and in tourism, given the large of analogies between the two sectors. Unmistakably, the number of differences is also significant, therefore necessitating other explanation in the pricing process. In fact, while the hedonic model applied to the traditional hospitality sector is similar to the one suggested by the real estate market, in the tourism industry it considers also a brand-new element. In particular, variable such as services and reviews plays a very important role, and are usually added to the mix (Andersson, 2010; Gibbs, Guttentag, Gretzel, Morton, et al., 2018).

In literature, a considerable number of researches and studies deepens the analysis of Airbnb price determinants and the underlying model. The first, innovative revolution in the lodging, peer-to-peer business is the central role of the host. Unlike the great majority of hotels, the host can set the price of his/her own properties (Voltes-Dorta and Sánchez-Medina, 2020; K. L. Xie and Kwok, 2017). While the traditional business adapts the final fare according to revenue-management practices, at the beginning Airbnb gave complete freedom to host. As mentioned while describing incentives in subsection 1.1.2, this freedom can cause serious drawbacks: the often complete lack of professional experience might complicate the pricing process for the host. Precisely for this reason, the platform offers two different choices to the host: manually set prices or the support of *Smart Prices*. In November 2015, the company launched this tool as the result of listening to a very urgent need[39]. Apparently, challenging as it is, many hosts were facing serious problems in setting the price (Hill, 2015). As the task can be both complicated and time-consuming, the purpose of this tool is to offer hosts what they needed the most: personalization and control, as well as support. The complex algorithm does the majority of the work, as it aims to maximize the number of booking by adjustments on the price. This tool has in fact some essential information. When a host that set a price, the *spillover effect*[40] suggests that the final value is influenced by similar places in the surrounding area (Chica-Olmo et al., 2020). The owner, however, does not know if nearby prices are effectively leading to bookings. Smart Prices suggests, therefore, the most appealing price, and hosts have the possibility to set a minimum and

[38] Airbnb vs Hotels: A Price Comparison (accessed October 4th, 2020).

[39] Bray, J., The Price is Right - How we used host feedback to build personalized pricing tools Airbnb Design (accessed October 6th, 2020).

[40] Wikipedia, Spillover (accessed October 6th, 2020).

maximum price to avoid excessively large fluctuations[41].

The complex model includes more than 70 influencing factors. Intuitively, it counts variables as *position*, *the ratio between supply and demand* - as mentioned -, *the number of people looking at the ad*, *the time spent, on average, on the ad*, *seasonality*, *type of listing and its characteristics*, *services offered by the host*. The price of the lodging can also be influenced by other actors, such as ratings, but also accessibility and nearby services (Chica-Olmo et al., 2020; Sánchez-Ollero et al., 2014). Voltes-Dorta and Sánchez-Medina (2020) compiled a precise review on the topic, presenting many recent studies, each one covering and inspecting a specific factor.

However, as shown by Gibbs et al. (2018), only a small fraction of hosts take advantage of this tool, and when they do it is usually because they have acquired more experience and they manage more listings. Possibly leading to great loss from an economical point of view, the underuse of dynamic strategy needs to be contrasted by company policies: Airbnb needs to encourage hosts to optimize prices.

1.3.3 Price prediction

To create a credible, efficient predicting model for listing prices, many different variables need to be combined. As already mentioned, even though the price can be arbitrarily chosen by the host, there are some factors that inevitably affect it. The literature cites countless researches on Airbnb pricing model, but also on the individual effect of specific characteristics (Chica-Olmo et al., 2020; Falk et al., 2019; Lawani et al., 2019; Lorde et al., 2019; McNeil, 2020; Voltes-Dorta and Sánchez-Medina, 2020; Wang and Nicolau, 2017). To shape a predictive algorithm, different statistical models are implemented (Ordinary Least Squares, quantile regression, random effect model, geographically-weighted regression). The hedonic price model is usually recognized as the most suitable and fitting tool.

Accommodation prices are essentially driven by three considerations: the physical characteristics, the factors which impact user perception, and the position. Moreover, Falk et al. (2019) proved that these components have different roles in shaping the final cost of the stay across location and price range. The following review presents those factors that are universally listed as the main influences on the price.

1.3.3.1 Reviews and host reputation

In an online platform, where people do business with strangers, reviews are a powerful tool. The information asymmetries cause by not knowing the host, on one side, and the potential guest, on the other, can be reduced with reviews, comments, and

[41] Airbnb, Prezzi Personalizzati (accessed October 6th, 2020).

ratings. For this reason, the platform highly encourages users to rate their stay. As stated in a TEDTalk by one founder, Joe Gebbia[42], "*a well-designed reputation system is key to get it right*".

The phenomenon of review is complex, and in online platforms, it is often seen as the digitization of the traditional *word-of-mouth* (WOM) communication. Based on a common idea, these two phenomena have however significant differences (Huete-Alcocer, 2017). Traditionally, word-of-mouth includes all forms of information exchange, and the electronic version of it basically consists in *"statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet"* (Hennig-Thurau et al., 2004). Reviews, which are part of the eWOM phenomenon, significantly and positively impact the consumers' behaviour (Goh, 2015; Lawani et al., 2019; Liang et al., 2017). Although some studies state that different readers or hotels category are differently influenced by review (Anderson, 2011; Gretzel and Yoo, 2008), online reputation is proven to generally influence the tourism and hospitality sectors, as mentioned (Litvin et al., 2008; Pan et al., 2007; Ye et al., 2011). In the era of trusting strangers, similarity plays a central role: similar age, location and geography lead to more openness (Abrahao et al., 2017). However, the power of close ones can still outweigh the web. Gellerstedt and Arvemo (2019) study the effect of a good friend positive review against a negative online majority, proving the greater effectiveness of the first. As mentioned in 1.1.2.2 - trust is one of the milestones of the success of Airbnb. Boosted by the era of social media, reputation systems can assist users in new transactions since usually, these take place between parties that have never interacted before (Jøsang et al., 2007). One controversial point is an incredibly large scale of information and comments available on the web. This might lead to an overload of data to be processed for consumers, and will most probably create new dynamics in the market (Dellarocas, 2003).

In literature, scholars have studied both the *quantitative and qualitative* impact of reviews (Goh, 2015). The effects can be split up into volume (total number of reviews about a stay/listing) and valence (the ratings, both average and variance). Focusing on the former, many studies show how the number of reviews has a positive impact on online sales (Y. Chen et al., 2003; P.-Y. Chen et al., 2004). The effect of rating, instead, is more disputed. Though some researchers show that the nature of reviews - positive reviews - can also play a central role, what is proven to influence more the number of booking is polarity. Simply put, the smaller the variance, the more positive the trend in bookings (Ye et al., 2009). However, more recent studies show how the effect is little to none (W. G. Kim et al., 2015). And when the effect

[42] See 27

exists, it is different for lower-tier hotels rather than upper-tier. Luxury properties, for example, may benefit more from very high scores, rather than a high number of reviews. While the first inspire quality, the latter can be associated with the idea of exclusivity (Blal and Sturman, 2014).

Another interesting property of ratings is the skewed shape of distribution: as the most satisfied and the most discontented consumers are usually the ones posting, the average might be misleading (Koh et al., 2010). Related to that, a good platform also needs to take care of complaints. Being able to detect in the first place, and eventually correct failures, is crucial, in order not to lose potential new clients (Floyd et al., 2014).

Having discussed the effect of *eWOM* on occupancy rate (Viglia et al., 2016), online reputation greatly affects also the price. Usually, the higher the rating, the higher the price (Y. Chen and Xie, 2017; Lawani et al., 2019). Deepening this relationship, studies have depicted a multifaceted scenario. Lorde et al. (2019), for example, show that the number of reviews has no real effect on the price, while Zhang et al. (2017) also proved the rating score to be negatively connected with the price. Moreover, as mentioned, lower-priced listings are affected differently than high-end ones (Öğüt and Onur Taş, 2012).

Intimately connected with the previous section, host reputation can play an incredibly important role in consumer behaviour and price definition. Hosts behaviour, therefore, might also charge a premium in the final price. For example, Chen and Xie (2017) reported a positive effect of immediate responsiveness on the nightly fee, while hosts picture increase trusts (Ert et al., 2016). To help consumers measuring hosts trustfulness, Airbnb launched in 2009 the *Superhost program*[43], which has the double goals of rewarding particularly successful and experienced hosts and at the same time trigger others to improve and do better. This program is open to every host in the platform, and only require them to meet four criteria[44]: very high ratings and responding rate, a little number of cancelled reservations and a minimum numbers of stay. Guests pay more attention to Superhosts' ads, therefore leading to a probable higher number of booking and review. The badge is perceived as a guarantee of quality, thus increasing the perceived value of the product offered by the host. This quality assurance is proven to be connected with higher prices: consumers are willing to pay a premium for this kind of stays (Ert and Fleischer, 2019; Liang et al., 2017).

[43] https://blog.atairbnb.com/superhost/ (accessed October 13th, 2020).

[44] Airbnb Superhost: il meglio dell'ospitalità (accessed October 13th, 2020). Hosts that want to take part in this program need to meet the following criteria: an average rating of 4.8 in the previous year, more than 10 stays, less than 1% cancellation rate in the last year - which means a total of zero for hosts with less than 100 bookings - and 90% of response rate within the first 24 hours.

1.3.3.2 Location

One of the foundations of house pricing lays in the attribute *position*. In real estate, the motto "location, location, location" describes the reasons why people are willing to pay a premium, under the same structural condition (Can, 1998). There are two levels of influence: the first one captures the surroundings, and comprehends all the adjacencies, while the other refers to neighbourhood characteristics (Kauko, 2006; Schirmer et al., 2014). Simply put, the location has a variety of ways to influence the final price, that range from neighbourhood to distance from the city centre, from nearby services to quality of surroundings, from the physical environment to green areas, and so on (Archer et al., 1996).

However, the importance of these variables is weighed according to the context: while many studies proved the importance of the *distance from the city centre* (Dogru and Pekin, 2017; Gibbs, Guttentag, Gretzel, Yao, et al., 2018; Wang and Nicolau, 2017), this significance is deeply connected on the general market in the scope. For instance, Kauko (2006) discloses how the location is essential in an urban context, while it is less interesting on a smaller scale - a village. Similarly, being close to the city centre might be not essential in maritime cities.

Moreover, whiting the urban context, different area have different influence on the price. Perez-Sanchez et al. (2018) model a multi-variable analysis that also investigates this dimension. The results show - for the 4 Spanish cities under the scope - a 16% increase in price for sightseeing area and a 5% increase for eating and shopping sites. To conclude, other elements that are intimately connected with location and therefore affect the price are points of recreation and sports, transportation facilities, service, retail.

1.3.3.3 Size and rooms

As aforementioned, structural characteristics of the listing are crucial factors in influencing the price. This dimension may be captured in different variables in an ad. In fact, the host can post the number of bedrooms and bed, the number of bathrooms and total rooms, but also the guests capacity. When booking a stay, filtering for the number of guests is actually the very first step. Then, the users can decide which type of room to choose, whether that is an entire apartment, a private or shared room or a hotel suite. In their study, Gibbs et al. (2018) model a hedonic regression that results in proving as a very significant factor for the final price both the type of booking - the entire apartment is, on average, the most expensive option - and the capacity and size of the dwelling. This result is confirmed in other researches

as well: *site attributes*[45] are in fact proven to be statistically significant (Dogru and Pekin, 2017; Lorde et al., 2019; McNeil, 2020; Perez-Sanchez et al., 2018). Therefore, on one side the foot square of the apartment is important, on the other number of rooms, bathrooms and accommodations play a central role as well (Y. Chen and Xie, 2017). These findings come from studies that focus on various cities and can be therefore considered universally valid - even though some differences might persist.

1.3.3.4 Services, amenities and policies

The previous section widely discussed the hedonic pricing model and its implementation in the rental market. Again, the HPM lays its basis on a simple concept: the price is the cumulative value of parts and attributes. Airbnb case perfectly fits in this description, as literature proves that not only size and number of rooms, but also many attributes and services are significantly affecting the final fare. Having demonstrated the importance of structural characteristics, services and amenities are presented as follows. In their review, Cai et al. (2019) illustrate precisely the complex framework. Findings are sometimes contradictory, and some variables have a mixed effect. However, some patterns can be recognized. Among amenities and services, some are considered nowadays essential: TV, Wifi, air conditioning, heating are just some example of variables that are almost in every rental. At the same time, analysis usually does not consider those comforts that are listed only in a limited number of houses. Dogru and Peking (2017) analyse the market in Boston, revealing a positive influence of appliances in general, but also a negative effect of the kitchen. The same study also proved a strong significance of offered breakfast, while Wang and Nicolau (2017) show the opposite results - in a more general context, considering the listing offered on the platform in 33 cities all around the world. As the number of bedrooms is proved to be important - see 1.3.3.3 -, so does the type of bed - real bed increase the perceived value of the listing (Wang and Nicolau, 2017). The general status of the property is of course important: user-generated reviews point out that cleanness is a positive influence on the final price (Y. Chen and Xie, 2017). Moreover, policies can also play an important role, and while instant booking is connected with the idea of cheap and has therefore a negative effect (Gibbs, Guttentag, Gretzel, Morton, et al., 2018), relaxed cancellation policies can increase the price (Y. Chen and Xie, 2017; Perez-Sanchez et al., 2018). Another interesting amenity is parking. In literature, this variable is often investigated and usually turns out as significant (Gibbs, Guttentag, Gretzel, Morton, et al., 2018; Thrane, 2007; Wang and Nicolau, 2017).

To conclude, the number of amenities that can affect the price is enormous, as each

[45] As defined in Lorde et al., 2019.

host can independently choose to add new services to his/her listing. For simplicity, this research only investigates the most common, to test their significance.

1.4 Segmentation

The last section of this *Literature review* illustrates an essential technique for market analysis: segmentation. Customer segmentation is a *consumer-oriented process* (Camilleri, 2018) that simply consists of dividing customers into groups or clusters, based on their characteristics, behaviour, purchase habits, or a mix of these variables. Generally, there are two macro-categories of variables: *descriptive* and *behavioural* ones, together with product-related factors (Camilleri, 2018; Ma, 2015). Based on these assets, the different approaches can be used to define groups (Camilleri, 2018; Sari et al., 2016) are listed as follows:

- *past purchases*: analysis on past purchase characteristics, such as price, shipping method, type of product, overall satisfaction;
- *profit potential*: next step to the previous method, it allows to categorize customers according to their *profitability*. Variable such as transaction frequency, last purchase, customer lifetime value and average expense are combined to identify most loyal customers, big spenders, those who are almost lost and those who need attention, potential and promising customers;
- *demographic*: relatively easy to perform, this segmentation simply identifies groups in terms of socio-demographic characteristics;
- *psychographic*: based on values, personality traits and interest, this kind of segmentation analysis can provide interesting insight when matched with the purchasing behaviour of the subjects;
- *behavioural*: groups according to the purchase behaviour.

The methodology is varied: clustering is surely the most comprehensible one, but decision trees, neural network, generalized association rule models do also their job (Ma, 2015; Tsiptsis and Chorianopoulos, 2010). Grouping customers and predicting their purchase habits is certainly a concerning problem for the market. However, the most challenging task is not to find these segments, but to actually obtain an actionable business strategy in line with those (Chapman and Feit, 2015). What is valuable is the underling information: whether there are any differences between groups, if these differences are relevant, or they are telling a story. This tool must provide companies and sellers with all the information to properly communicate with potential consumers, and identify those segments that can easily be targeted. Both the resulting targeting - more efficient - and the improved communication

both enhance profit and improve customer satisfaction. According to its resources, a company can decide whether to address a multitude of segments or just a limited number.

The phenomenal growth of sharing economy and the huge revolution that came along has surely change tourists' behaviour. On the other side, the digital era affected also suppliers. Personalized marketing and offers are some of the main innovations in tourism research, and guarantee tourists with the most suiting option. Within this sector, segmentation focuses on the choice of accommodation and destination, motivators to travel, expense and budget, use of technology while booking (D. Guttentag et al., 2017). Researches show that participation on sharing platforms is strictly connected with some feature, and therefore users and non-users can be generally identified upon the presence of some specific realizations. For example, the *Flash Eurobarometer 438* recognizes age and education as influencing variables, while gender seems not to be very relevant. Lutz and Newlands (2018) also highlight the importance of further analysis: once users are identified, it is significant to discriminate according to different groups, as this is the turning point for personalized contents. This step consists in check whether socio-demographic variables and specific guests profiles affect the segmentation. Again, they prove that the choice of the type of accommodation, for example, is significantly connected with age, income and gender. Comparing the characteristics of those who stay in an entire apartment and the ones that, instead, stay in a shared room, differences emerge. This second type of rooms usually targets lower-income guests, and age is also a recurrent discriminant. Gender and social status also play an important role: women show the tendency not to stay in hostels and dorms, as well as couples. A broad range of different motivators move guests in their choices (D. Guttentag et al., 2017), and being able to identify them can provide useful insight on the platform side.

Chapter 2

Data and Method

2.1 Research Question

The first chapter explores the main theoretical aspects that are related to the broad concept of sharing economy. Many different elements are presented, from the rise of the phenomenon to the case study of Airbnb. This review addresses all the concepts that are necessary to understand the following chapters. Thus, the first step is to understand the aim of the research. The goal is to outline the relationship between tourism and the sharing economy, with a particular focus on the Italian case. Customer preferences are also identified and examined, as this study wants to investigate both the demand and the supply side. The idea is to delineate a clear and complete picture of the house-sharing market.

The large contribution to the price prediction issue is thoroughly discussed in Chapter 1. Many methodologies are explored in a variety of contexts, and a large number of attributes are proved to be significant in shaping the price in the house-sharing market. Therefore, the first step is to check whether there is consistency between attributes usually proved significant in Airbnb models and the particular case of Milan. At the same time, the tendencies of Italian tourists are explored to group people and be able to identify different targets in the market. Simply stated, the author wants to prove the effect of socio-demographic attributes to define those who participate in the house-sharing. In the end, the aim of the research is to understand on one side the most significant house attributes in defining the price, on the other the tourists' preferences. The findings of the quantitative research are eventually combined with the results of the clustering process. The second part, in fact, focuses on those people that are willing to take part as customers into the house-sharing business, as they are an essential part of shaping the price as well.

Once the aim is clarified, the reasons for this research are also reviewed.

Despite the high number of paper debating on the most important attributes while deriving the price of a listing in Airbnb, only a small fraction of these study make predictions. Moreover, the Italian market is never analysed in its own peculiarities, and it seldom appears in comprehensive analysis - one example is in Wang and Nicolau, 2017. As a direct consequence, a complete lack of such predictive studies is registered for the Italian case, therefore making this final project one of the first exploratory studies on the price model of Airbnb in this framework. The research wants to investigate the price, but also significance of listings' attributes with respect four macro-categories: *reviews and host reputation* (referred as *trust* in the analysis), *location*, *structural attributes* and *services*, as presented in Chapter 1. Overall, Chapter 3 wants to answer the following questions: what are the most significant characteristics? Which categories do they belong to? What base-price would have a listing with given features?

The same is true for the demand side. Although examples of segmentation can be found in the Italian touristic framework, the particular focus on the house-sharing market is a peculiarity of this research. Understanding customer preferences is another essential step in shaping the price: what are the main characteristics of Italian travellers? Why do they choose Airbnb? What price are they willing/able to pay for their vacation?

The *raison d'être* of this thesis is to answer all these questions. On one side, the justification is to add details to the complex picture of the house-sharing market with an Italian case study. On the other, the comprehensive approach of studying the phenomenon from both the supply and the demand side applied in the project, as aforementioned, is seldom encountered and represents a new application.

2.2 Methodology

The research, as already explained, is mainly composed of two different parts: the first, that uses supervised learning to obtain an approximation of the price, and the second, where customers are grouped according to their attributes and purchase habits.

The first set of methods focuses on price prediction. Due to computational reasons, each model is first tested on a 10 per cent sample of the complete data set, to perform the first round of parameter tuning. According to the results, the final fine-tuning is performed on the complete training data set, to obtain the best performing parameters. Moreover, thanks to the high dimensionality of the data, it is possible to divide the final set into training and test of equal dimension, applying a 50-50 rule.

The different methods are briefly presented as follow. Given the idea of *hedonic pricing model* and its function, presented in 1.3, the price prediction starts with multiple linear regression - the most common form of linear regression - encoded as follows:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon$$

with n = number of predictors.

However, this model is highly influenced by outliers, and also requires a high number of assumptions (James et al., 2013), that can be checked via diagnostic plots in R[1]:

- linear relationship between predictors and predictors. To detect violations, the residuals plot must be checked;
- errors must be normally distributed: $\epsilon_i \sim \mathcal{N}(\mu, \sigma^2)$. The normal Q-Q plot shows that this assumption is violated;
- homoscedasticity, as errors must have the same variance: $Var(\epsilon_i) = \sigma^2$, for $i : 1, ..., n$. To detect the violation, one needs to refer to the Scale-Location plot;
- absence of multicollinearity, when two or more predictors are highly correlated. This assumption can be tested through a correlation matrix;
- absence of auto-correlation: $\epsilon, \epsilon_1, \epsilon_2, ..., \epsilon_n$ must be uncorrelated. .

However, many of these are violated. The log-transformation of the dependent variable solve some problems, but the errors' distribution is still non-symmetric. LASSO ("least absolute shrinkage and selection operator") is then implemented, as this method loosens the aforementioned assumptions (of all, the normality of the errors). This non-linear shrinkage method reduces dimensionality by selecting a smaller set of predictors. The performance, however, does not significantly improve, and therefore other models are tested.

In particular, tree-based models are great alternatives to sidestep all the a priori conditions of traditional regressions. As stated in Phelps and Merkle (2008), trees, forests and boosting are non-parametric alternatives to regressions that can handle a wider variety of data. Regression trees give a little contribution to the analysis, and more complex algorithms are preferred - at the expense of clearly displayed results. The analysis moves therefore to the next models. Random forests are a collection of de-correlated trees built on a bootstrap sample of the original data, by selecting only a n fraction of variables - usually $n = p/3$ for regression, with p = number of predictors (Battiti and Brunato, 2014; Breiman, 2001; Hastie et al., 2017). The last method presents a similar procedure, but as the output of Random Forest is the

[1] Bommae K., Understanding Diagnostic Plots for Linear Regression Analysis, 2015 (accessed December 4th, 2020).

ensemble of independent trees, in Gradient Boosting trees (or stumps) are grown sequentially, based on residuals (Hastie et al., 2017; Kuhn and Johnson, 2013).

The accuracy of each model is calculated with the RMSE ("root mean square error"), to facilitate comparison: the lower the value, the better the fit. The following table sums up the train and test errors, and schematically presents the models.

Model	Train error	Test Error
Linear Regression	-	72.71
Lasso	-	70.58
Regression Tree	70.25	70.46
Random Forest	39.02	38.96
Gradient Boosting	19.08	35.60

The detailed description of the analysis is presented in Chapter 3, while limitations and results can be found in Chapter 5.

The second part illustrates the segmentation. As these operations are carried out on a smaller set of information, no particular computational problems are encountered.

First of all, the purchase preferences of each tourist are examined: as clearly underlined in 2.1, customers are essential players in the pricing mechanism. For this reason, RFM ("Recency, Frequency, Monetary") profiles are identified, and people are divided according to these attributes. By doing so, the author identifies the average expense of those who actually book accommodation on Airbnb or similar sharing platform, therefore having a first glint of the typical customer.

Then, these three main indicators - days since the last trip, trips per year and total expenses - are added to the data set, and used for unsupervised clustering. The aim here is to check whether socio-demographic variables can predict the RFM profiles, therefore proving that booking and travelling habits are correlated with different attributes. In particular, two different methods are explored: k-means and hierarchical cluster. While the former, by structure, can only handle continuous numeric variables, the latter can process any type of data by using the Gower distance - see 4.4 for further details.

2.3 Data Description

As aforementioned, the research studies the two different side of the lodging industry in the sharing economy, with a specific focus on the case of Airbnb. Therefore, there are different data sources to be investigated.

To create a predictive model, information on the listings in Milan are retrieved from

Airbnb online archives. The data is downloaded directly from the websites[2], and can be used under a Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication" license. Having decided to focus on the Italian market, choosing which city to analyze is completely arbitrary, as the aim is to create a model that can also be reshaped on other cities for further studies. Given the significant impact of the current pandemic on almost every aspect of everyday life, the analysis focuses on 2019. In the following section, the data sets are described. The company revises monthly both listings and bookings for each city, therefore providing a precious source of up-to-date data for researchers. The *calendar* data sets contain information about *bookings*. The *listing* data set, instead, contains all the essential attributes that shape the price per night.

As widely presented in Chapter 1, the final price is a mixture of a variety of characteristics. The main groups of variables are described below.

- listing identification and basic information provided on the website: `id`, `name`, `summary`, `space`, `description`;
- `host_id`, `host_name`, `hosts_since`, `host_location`, `host_about`: basic information on the host;
- `host_response_time`, `host_response_rate`, `host_acceptance_rate`;
- `host_verifications` and `host_identity_verified`: variables that describe the reliability of the house owner;
- `host_has_profile_pic` and `host_is_superhost`;
- `number_of_reviews`, `last_review`, `reviews_per_month`, but also the ratings for different aspects of the stay (`review_scores` variables);
- `note`, `transit`, `access`, `interaction`: basic info for the guest to reach the house;
- `house_rules`, `cancellation_policy` and `instant_bookable`: policies of the listing;
- `neighbourhood`, `latitude` and `longitude` (WGS84 EPSG:4326 coordinate system);
- `room_type` (type of listing, recorded as *Private room*, *Entire home/apt*, *Shared room* and *Hotel room*[3]) and `property_type`;
- `bathrooms`, `bedrooms`, `beds`, `amenities`, `square_feet`: main structural feature;
- `price`: integer, price of the listing;
- `security_deposit`, `cleaning_fee`, `extra_people` as added costs to the price.

[2] Inside Airbnb (accessed October 9th, 2020).

[3] Airbnb offers independent hotel the possibility to serve ads on the websites. As revealed in their *Newsroom*, Hotel Boutique is one of the new types of property available on the platform. Airbnb Newsroom Airbnb Unveils Roadmap to Bring Magical Travel to Everyone https://news.airbnb.com/airbnb-unveils-roadmap-to-bring-magical-travel-to-everyone/ (accessed October 9th, 2020).

To perform segmentation on (potential) customers, information is retrieved from ISTAT, the Italian Institute of Statistics[4]. Again, the data can be used under a licence Creative Commons – Attribution – 3.0 version. The data set is provided together with a .R file with preliminary code, to rename variables and recode them as factors when necessary. Overall, the data set contains 51 different information, as grouped below:

- `mese` and `annrif`: month and year data refers to;
- `sesso`, `eta10`, `staciv4`, `pnasc`, `reg`, `rip`: socio-demographic info on the individual;
- `istr4`, `condogg`, `posiz4`, `ateco3`, `cond4`, `ORARIO`, `RAPP`, `CODPROF7`: education level and occupancy status, together with working sector, position and work schedule;
- `progvia` registers the progressive number of trips for the single, analysed individual;
- `TIPOVGG`: reason for the trip, together with month and year of the trip;
- `DEST_PR`, `DEST_RE`, `DEST_IE` give precise info on the destination, from detail level - Italian province - to general direction - Italian region or foreign country - to macro level - Italy/abroad binary info;
- period (`TRIM`, `GGINIZ`, `MMINIZ`, `AAINIZ`) and total duration (`DURATA`) of the trip;
- main reason for the trip (`MOTVAC`, `MOTLAV` and `tipo`), means of transport (`MEZZO`) and accommodation (`ALLOG`); `ALPART` and `npart`, to describe whether other members of the family participated;
- `ORGALL` and `ORGTRA`) describes the type of organization - individual, tour operator, none - and `IORGALL` and `IORGTRA` register whether the individual used internet while booking transportation and/or accommodation;
- `TIPOMARE`, `TIPOCROC`, `TIPOMONT`, `TIPOCITTA`, `TIPOCAMP`, `TIPOALTRO`: type of destination;
- coefficients to calculate the number of trips on the population level (`W` and `COEV`);
- average expense overall (`ESPE_CO`) and per day (`ESPE_GIO`) for single travellers;
- `piattall1`: binary variable, to describe whether or not the booking was through online travel agencies (such as Bokking.com, TripAdvisor, and so on);
- `piattall2`: binary variable, to describe whether or not the booking was through an online sharing platform (Airbnb, for example);
- `piattall3`: binary variable, to describe whether or not the booking was on the online website of the travel agencies or travel operator.

[4]Istat, micro dati accessed September 4th, 2020).

The sample design and data collection process are precisely described in the *Nota Metodologica* available online[5]. The reference population is composed of Italian families, divided into five geographical subdivision: north-east, north-west, centre, south, islands. Moreover, to capture seasonality, the research is continuous as the sample is divided into 12 sub-sample groups. Each sample participates during a specific month of the year, with a three-step survey. The first and the last of the three meetings cover the topic of trips and travel. It is important to notice that the data set also contains the survey's weights, that allow changing sample estimates to the entire population. The methodological note precisely describes the computational steps to obtain this *calibration estimators*.

2.4 Data Cleaning and Feature Engineering

The first step before performing the actual analysis is data cleaning. To obtain the best fitting results, this action is undertaken together with feature engineering. As some information is dropped and the data is prepared for the analysis, some variables are coded as the result of combination and manipulation of other information.

2.4.1 Airbnb data set

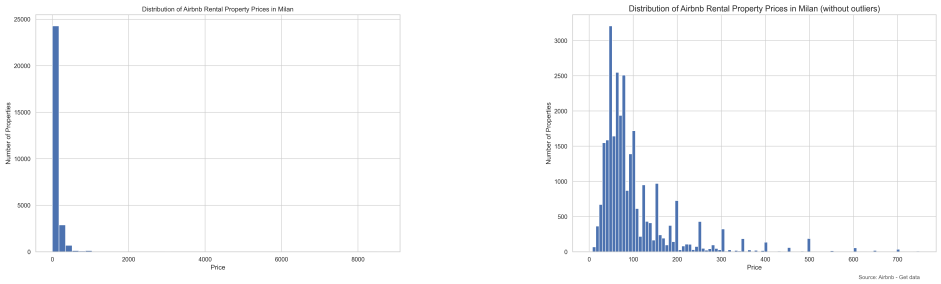
First of all, the *calendar* data is considered. The monthly update by Airbnb collects information that goes from the day of the update till the following year, covering a span of 365 days. As mentioned, the aim is to train the price model on 2019 data, to appreciate differences due to seasonality and correct therefore predictions. However, bookings from following updates shows that travellers regularly cancel, change or make a new reservation a few weeks before the stay. This tendency suggests that data from January 2019 might not be optimal to investigate the entire following year. For this reason, all the twelve updates from 2019 are downloaded: from January to December, the final data set is put together as the union of observation from each month. Basically, each download is divided into days before the next updates, and days after; only the observations from the first group are part of the final data set. The operation is computed for each month, and each subset is used to calculate the booking rate for each listing in the span of one month. The resulting value is compared to the average occupancy rate per month, creating `diff_from_avg`: this variable describes the availability of each listing compared to the average. The idea behind this information is the simple rule of supply-demand: if the demand is higher, the price goes up. A positive rate suggests that the listing is experiencing a higher than average number of bookings. On the other hand, a negative value describes the

[5]ISTAT, Nota Metodologica, accessed October 24th, 2020.

situation in which the listing, for that month, registers - on average - less booking that the other listings in the city for the same period of time.

The remaining part of data cleaning is performed on the *listings* data sets. First of all, the variables are properly coded. Numeric variables originally saved as strings (such as `price` and `host_response_rate`) are changed to integer. Dates are converted to a date object, to numerically encode *first* and *last_review*. The resulting `days_from_last_review` and `days_from_first_review` variables facilitates and accelerates the understanding the information. It is also easier to add them to a predictive model - as it avoids the use of dates.

Then, the binary variables are changed into dummies, as well as the variable recording descriptions of the listings and basic info for the guest - these are coded as 1 if present, 0 otherwise. The variables `host_response_time`, `cancellation_policy` and `bed_type` are converted to factors and numeric variables are saved *as.numeric*. `Host_verifications` is changed to a simple count of tools, and the list is removed for simplicity. Similarly, `number_amenities` registers the amenities and services. However, some particular services are isolated to check for their specific presence/absence with ad hoc dummies. Having highlighted the importance of free parking near the house in literature, the amenities *Free street parking* is saved. *Kitchen* and *smoke policies* often appear in literature as well and are therefore added. Given the importance of connection, especially while travelling, the presence/absence of *Wifi* is also registered. Arbitrarily, this study also investigates the effect of *Host greets you*, *Bed linens*, *Pets allowed* and *Hairdryer*. The former wants to investigate the eventual importance or influence of having human contact with the host; on the other side, considering the amount of space that bed linens and hairdryer require in the luggage, the other two explores whether people travelling appreciate this kind of extra or if they prefer their own. Once this analysis is completed, the twelve months are bind together.



(a) Price distribution of the whole dataset (b) Price distribution after removing outliers

Figure 2.1: Histograms with price distribution.

Other essential steps for a nice fitting model consists of dropping outliers, as they

might cause the predictions to be highly unstable, and dealing with NAs value. First of all, by displaying the price distribution (*Figure* 2.1), it is easy to foresee the strong effect of outliers on the main statistics. Even though the price distribution presents a long right tail, the complete data set (a) is highly left-skewed. Therefore, observations that are extremely positioned are deleted from the set[6]. The variables `accommodates`, `guests_included`, `bedrooms`, `bathrooms` and `bed` also present some outliers, meaning that some listings register some information with respect to a single booking, while some other refers to the entire structure[7]. These are simply coded to the most likely value: guests cannot be more than the total capacity of the listing, and in this case, `accommodates` is used instead; *private rooms*' and *shared rooms*' `bathrooms` and `bedrooms` are changed back to 1. A little number of cases (8) are dropped, as there are inconsistencies in the data.

Then, missing information needs to be changed or dropped, upon the possibility to find a suitable replacement. Apart from NA, missing values are also found in the form of strings ("N/A", ""). With a very small coverage[8], `square_foot` is dropped, as the available information is not enough for substitution and other attributes can describe the dimension of the dwelling. Missing `neighbourhood` are dropped[9]. On the contrary, `bedrooms`, `bathrooms`, `beds` and `host_total_listings_count` are numeric and NAs are therefore changed with the average. `host_response_time`, `host_response` and `host_acceptance_rate` that corresponds to missing information are saved as "No info". In fact, these missing realizations can be due to different reasons. On one side, some NAs might be because the host has never received a message/booking. On the other side, the cause might simply be that they leave the message to be read, therefore providing a bad service. In both cases, as it is not possible to make distinctions, the new factor level is saved as the most penalizing, as 1.1.2.2 discusses the importance of reviews and trust to take part in the sharing. Lastly, a significant problem is how to deal with missing values in reviews. First of all, the variable `reviews_per_month` registers the absence of values exactly when the listing has 0 total review. These NAs are simply saved as 0. The `ratings` attributes, instead, presents a more complex situation, as the platforms allow guests to have a certain level of freedom when assigning scores: someone might decide to review the location, someone else the communication, others might leave a score for each item. Literature discourage the imputation of the average value, as it might be highly misleading (Koh et al., 2010). However, a great amount of these absent

[6]The listing with a price higher than 750 are dropped. The value is chosen as $mean(price) + 3 \times sd(price)$.

[7]In fact, many *Private room* counts more than 10 bedrooms and bathrooms. This goes against the definition of the private room itself by Airbnb, that describes one-room ads. By cross-checking with the *property_type*, many of these correspond to facilities such as B&B and hostel. The hypothesis is that these facilities upload the total capacity of the property, even though each booking is separated for each guest.

[8]This attributes never counts more than hundreds realization out of thousands of entries per month.

[9]No package in R allows to combine the info of a district with coordinates.

values corresponds to zero reviews, and for this reason are coded as 0 as well. In the pre-processing step, briefly presented in Chapter 3, this information is however corrected by creating new variables that describe the relationship between reviews' score and the number of reviews, given the high correlation of the two.

The second part of feature engineering focuses on positional attributes[10]. For simplicity, the computation described below are performed on a smaller data set, obtained as the combination of all the uniquely registered listing throughout the year, to have a single entry per listing. A central element in deriving the final price of a listing is proven to be its location. The complete analysis presented in Chapter 1 shows how many attributes of location influence the rate per night of a stay: neighbourhood, distance from the central district, nearby services, and so on (Archer et al., 1996). For this reason, the distance for the city center is calculated in kilometers

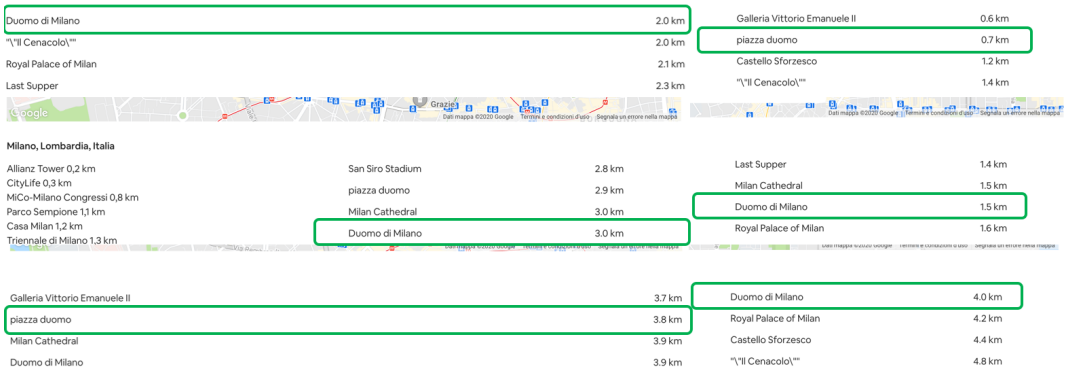


Figure 2.2: Screens from Airbnb website.

from the Duomo di Milano[11]. This particular spot is chosen arbitrarily both because many ads in Airbnb includes the distance from this building in the description - see *Figure* 2.2[12] - and because the cathedral is the heart of the city.

Together with the distance from the centre, the location of *points of interests* plays also an interesting role in shaping people choice. As presented in Giglio et al. (2019) and Huang et al. (2018), human mobility is deeply connected with the location of hot spots in the city. When planning a trip, people often decide a priori which attraction to visit based on personal interests. Technology is, again, great support for this complex tasks: an increasing number of online trip planning systems automatically generates a routing plan for visiting specific Points of Interest, pre-selected by the user (Sylejmani and Dika, 2011). To catch this dimension, the study checks also

[10] For simplicity, *Geopy* is preferred to the package *geosphere* in R, as the Nominatim function (Geopy, Geopy's documentation accessed October 18th, 2020) gives the same output of distHaversine (geosphere package, distHaversine, accessed October 18th, 2020).

[11] Address: Piazza del Duomo, 20122 Milano MI, Google Maps https://www.google.it/maps (accessed October 10th, 2020).

[12] Screen from Airbnb website: Airbnb Milano (accessed October 10th, 2020).

for the number of tourist attractions located nearby each listing. The variables `count_iso`, `number_pois`, and `avg_distance` represent exactly this measure. The aim is to represent the location of a house differently, to test which one is more important while defining the price.

Name	Category
Leonardo3	Museum
Torre Branca	Historical Landmark
Pinacoteca Ambrosiana	Art Gallery
Il Duomo	Building, Cathedral, Church
Castello Sforzesco	Historical Landmark
Galleria Vittorio Emanuele II	Shopping Mall
Teatro alla Scala	Opera House
Santa Maria delle Grazie	Church
Navigli Neighbourhood	Historical Landmark
Museo del Novecento	Art Gallery, Museum
Pinacoteca di Brera	Art Gallery, Architectural Landmark
Parco Sempione	Park
San Maurizio al Monastero Maggiore	Church
Monumental Cemetery	Cemetery
Fondazione Prada	Art Gallery, Cinema
Torre Velasca	Building
Corso Como	Historical Landmark

Table 2.1: Table with the selected POIs and their categories.

`count_iso` and `number_pois` share a similar idea: while the first variable counts the number of tourist attractions within a defined number of minutes walking distance, the latter simply takes the distance and considers the POIs in an n-kilometre radio. The initial idea was to calculate `count_iso`. Different approaches are tested for the calculations, but no significant results are produced. In particular, this study explores two different ways to create isochrones - that are lines *drawn on a map connecting points at which something occurs or arrives at the same time*[13]. The first approach consists in creating them around each listing with OpenRouteService, resulting in geolocated polygons containing all the places within n - with n arbitrarily chosen - minutes walk. For computational reason[14], this operation can not be run on the entire set of entries, and an alternative solution had to be implemented. The second approach consists in using the package *osmnx* in Python, which refers to OpenStreetMap. In this case, the provider is overloaded and makes it difficult to perform the large numbers of requests[15]. Therefore, due in general to computational

[13]Isochrone map, Wikipedia (accessed October 10th, 2020).

[14]The API key from OpenRouteService used to create isochrones has only a Free plan with the following restrictions: *Isochrones (500 requests per day @20 requests per minute)*, and computing the operation on the 28255 listings would take months. More information available at accessed October 10th, 2020).

[15]In many trials, while a small subset of entries give the expected result and create the polygon of the isochrone, the following error is produced on the entire data set: *HTTPConnectionPool(host=', port=80): Read timed out. (read timeout=10)*. See the issue at the following link (https://forums.x-

reasons, the variable *count_iso* can not be computed on the whole data set, and *number_pois* is used instead.

In the web, different websites of suggestions are searched and a final list of tourist attractions are selected - a certain level of arbitrariness is inevitable[16] - as in *Table* 2.1.

Having decided for a radio of 1 kilometre around each listing, the number of points within that distance is recorded in the data set. As clearly shown in the map of *Figure* 2.3, the great majority of locations are condensed in the central districts[17]. However, this tendency for centrality causes the majority of location (17974) to register no tourist attractions in the span of one kilometre. To correct this measure and add details and complete the data set, the *avg_distance* of each listing from the 17 spots is computed. As for the distance from the city centre, this variable is coded in kilometres, as a float with a single decimal - as in Airbnb[18] website ads. The additional information - `CBD`, `number_pois` and `avg_distance` - is added to the final dataset.

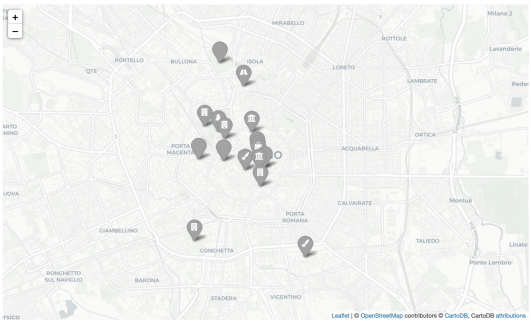


Figure 2.3: Location of tourist attractions in the city.

To conclude, the importance of these positional attributes is widely discussed in the literature (Archer et al., 1996), but this is also confirmed as we can expect comparable listings that are near one another to be priced similarly. This phenomenon, mentioned in Chapter 1, is the *spillover effect*: the externality suggests that the final price per-night is influenced by the similar places in the surrounding area (Chica-Olmo et al., 2020). Simply stated, the price might be affected by spatial autocorrelation (Haining, 2001). Further researches might investigate this topic more carefully.

plane.org/index.php?/forums/topic/115382-btileopenstreetmaporg-timed-out-error/) for further details (accessed October 11th, 2020).

[16]The list of 17 tourist attractions is chosen between the ones presented in *Culture trip* (https://theculturetrip.com/europe/italy/articles/20-must-visit-attractions-in-milan/) (accessed October 11th, 2020). No restaurants, railway stations or bus stops, metro or similar services are included, to keep a limited amount of points.

[17]This finding is confirmed by looking at the heat maps from Instasights, that displays the most visited places in the city per category. This tool gathers data from millions of users, according to four different categories (sightseeing, eating, shopping, nightlife), and make them available for consulting. However, the data from the map cannot be downloaded to be coded as variables. InstaSights, https://www.instasights.com (accessed October 11th, 2020).

[18]See 12.

2.4.2 ISTAT data set

For the second part of the analysis, the data from ISTAT previously presented is analysed. To classify and group individuals, the central variables in the data set are `IORGALL`, which describes whether individuals book accommodation via web and `piattall2`, recording if any platform such as Airbnb is used in the process.

Overall, the amount of information is such that no more variables need to be added to perform the analysis. Therefore, data preparation only consists of cleaning and recoding of pre-existing variables. First of all, redundant information is dropped, as well as non-necessary details. For example, the aim of the study is to focus on the booking process and purchase behaviour of customers, rather than precisely point out their favourite destinations. Then, categorical variables are mapped into factors and dummies, when needed (`age`, `education`, `working status`, but also `type of destination`, and so on).

Chapter 3

Analysis: supply side

The following section illustrates the different steps undertaken to obtain a prediction model for prices in Airbnb listings. In particular, the first step consists of *Exploratory Data Analysis*. Then, the best fitting models tested to predict the final price are presented.

3.1 Exploratory Data Analysis

Together with data cleaning and data engineering, exploratory data analysis is performed for general insight on the data. Thanks to the *neighbourhood.geojson* file provided directly by Airbnb, the listings are plotted on a map. In Figure 3.1 , the

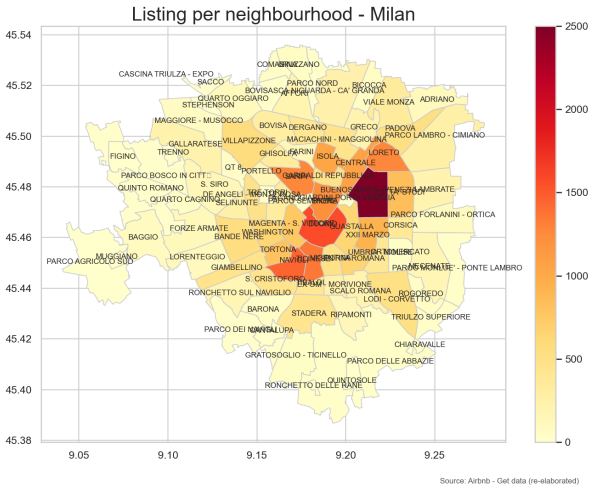
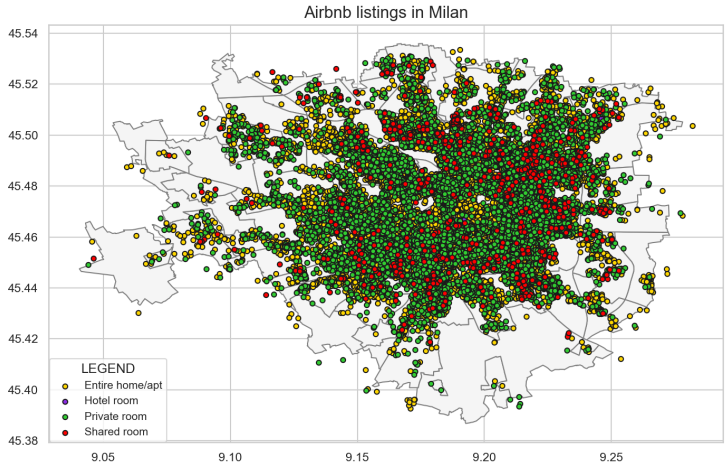


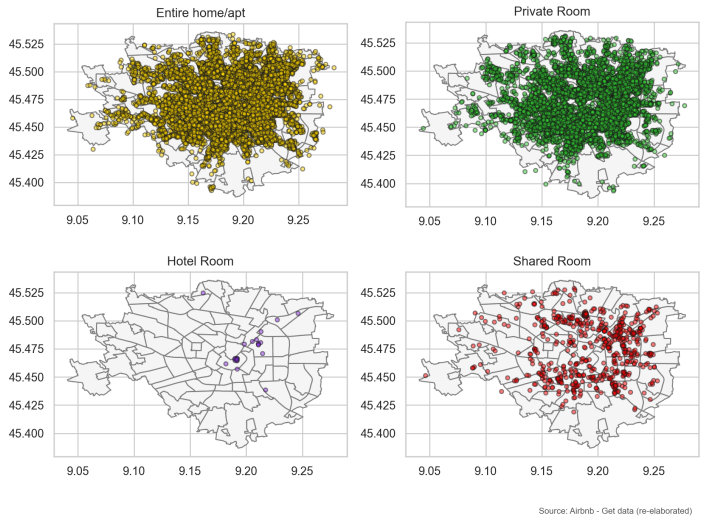
Figure 3.1: Heat map for the most "crowded" neighbourhoods.

number of listing per neighbourhood is represented through a red scale choropleth

map, while Figure 3.2 display the points in a scatter. In particular, (a) presents the location of houses, (b) shows the different location according to the type of room. The *Entire home/apt* is clearly the most frequent option (72.96%) followed by *Private room* (24.76%), *Shared room* (2.2%) and *Hotel room* (0.08%) - see Table 3.1. Apart from the distribution of room type, it is also interesting to investigate the average price for each group.



(a) Scatter plot with every type of listing.



(b) Scatter plots per listing type.

Figure 3.2: Scatter plots

The box plots in Figure 3.3 show a similar price for the first three categories, while hotels have generally a higher price - but also a huge dispersion around the sample mean. Similarly to what one would expect, the average daily rate decreases from an entire apartment to a private room and again shifting from this to a shared room.

	Room type			
	hotel room	entire apartment/house	private room	shared room
perc.	72.96	24.76	2.2	0.08

Table 3.1: Room type distribution for Airbnb in Milan.

However, this dimension is highly influenced by the size of each class: in fact, the distribution of room type is skewed (see Table 3.1). Given the large number of neighbourhoods, box plots are not very informative to explore the price per each district. For this reason, the choropleth map is displayed instead.

The first gap shows a decrease of almost 80 euros in average, and the most expensive district - *Giardini Porta Venezia* - is nearly three times more expensive than the average prices (after removing the outliers, the average price is 127.6 euros). The price quickly decreases towards the average level, as only the most expensive neighbourhoods register a significantly higher price than the overall mean. *Figure* 3.4 clearly show this trend, as most of the district are in green and blue shade of tone. Moreover, this map highlight another intuitive trend of pricing strategies in general: the closer to the city centre, the higher the price.

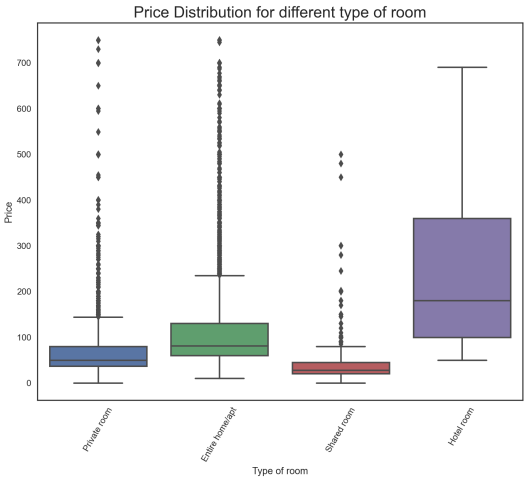


Figure 3.3: Box plot with the average price per room type.

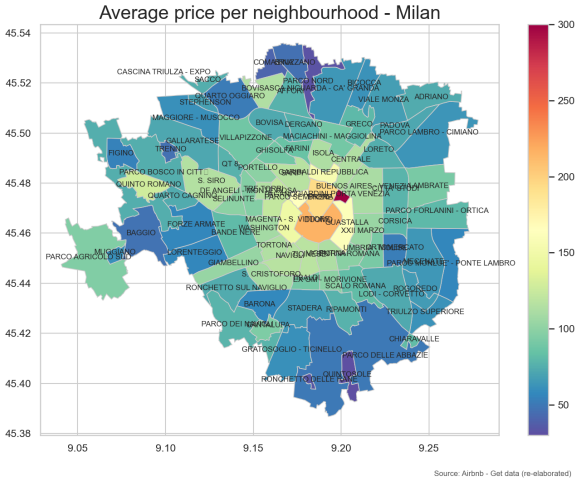


Figure 3.4: Choropleth map in Spectral palette.

Another aspect that, usually, influences the price is *seasonality*. As explained in the "Price"section in Chapter 1, month and day of the week are both considered in the *Smart price* algorithm. The data is already corrected for outliers, and the results are presented below. The fluctuation does not seem very relevant on the `month` variable, as both average and standard deviation present a stable tendency. This result seems to be in contrast with the literature, as seasonality is usually proven to be significant in the hospitality sector. However, the daily update of prices in the *calendar* data set might provide more precise information on this measure[1].

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Avg	121	122	125	130	129	128	127	127	129	130	130	131
Sd	87.3	88.6	92.8	94.2	92.8	92.1	90.9	90.6	91.6	93.3	92.9	93.7

The last variable inspected is `reviews`, that cover a range between 0 and 718. However, the number of listings that has no review on the platform is incredibly high: 11286 out a total of 28255. The variable has therefore a highly skewed distribution. To measure customer satisfaction, ratings might be a more solid option. The average score, however, is incredibly high, as addressed in Zervas et al. (2015) (see section 1.2). For example, isolating for the listing in January, the statistics are presented below[2]:

	Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
January	20.00	90.00	96.00	93.17	100.00	100.00	4059

As mentioned while dealing with NAs values, for this reason using the average can corrupt the analysis and it is highly discouraged in literature (Koh et al., 2010).

3.2 Analysis: Price Prediction

The analysis is completely implemented within the Rstudio environment in R. The dependent variable, `price` is estimated and then predicted with different models. As presented in the research question, the final aim is to discover the most significant variables and offer a good prediction of the price, despite the several limitations of the data set - that will be covered in Chapter 5. As presented in the Methodology in Chapter 2, there are different regression model that can be implemented to predict

[1] In this dissertation, due to the high volume of information contained in this data set - millions of entries per month - the price is shape on the monthly price of each listing.

[2] During the year, the average ratings are as follows: Jan 93.2, Feb 93.2, Mar 93.3,Apr 93.4, May 93.5, Jun 93.4, Jul 93.4, Aug 93.3, Sep 93.4, Oct 93.3, Nov 93.3, Dec 93.3. However, it is important to remember that the majority of listings is recorded every month, so apart from those listing that get some new rating from one month to the other, the expected value tends to be approximately identical throughout the year. Nevertheless, the value is noticeable high.

prices (Madhuri et al., 2019; Truong et al., 2020; Xin and Khalid, 2018). The following sections present step by step the model selection.

3.2.1 Linear Regression and LASSO

The first methods explored, as briefly mentioned in 2.2, are linear regression and LASSO.

To use the function `lm`, the author supposes that all the assumptions presented in Section 2.2 (linearity, normally distributed and independent error, homoscedasticity and absence of multicollinearity between predictors) hold. However, after performing the complete multivariate linear regression, the diagnostic plots detect the violations of different assumptions (see Figure 3.5).

The log-transformation is applied, and the most important variables are selected backwards (*backward selection*), deleting one by one all the non-significant predictors. The final model counts 25 variables, and the RMSE is 72.71. However, although this transformation solves the heteroscedasticity and flattens the linear relationship, distribution of the errors presents heavy tails - see Normal Q-Q plot in 3.5.

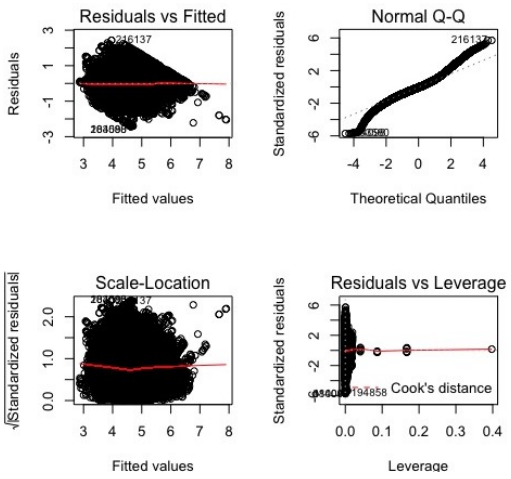


Figure 3.5: Diagnostic plots for the multiple linear regression, log(price).

As it suffers less high variability (Hastie et al., 2017), LASSO is also performed for variables selection and predictions. The assumptions require numeric, independent and standardized variables (Tibshirani, 1996). Therefore, after the preprocessing and the creation of dummies from factor, the function `glmnet` is implemented. The accuracy is slightly better - 70.58 euros of error - but as the error is still large, other assumptions-free models are explored: trees, forests and gradient boosting.

3.2.2 Tree and Random Forest

First of all, regression trees are implemented with `rpart`. The first trial uses default parameters (`minsplit` equal to 20, 30 for `maxdepth` and `cp` of 0.01) and the call is

as follows:

$$rpart(formula = price\ ., data = trainset, method = "anova")$$

and the resulting model includes a variety of outputs. The `rpart.plot` is displayed in Figure 3.6. First of all, the *cp table* displays the number of splits and the corresponding error. The object `variable_importance`, that collects all the most significant predictors, is also an interesting element to inspect. Comparing the output with the previous parametric models, some results are confirmed, as in Table 3.2. However, a single tree with default parameter gives worse results compared to linear regression and LASSO: the RMSE is 78.70 in the training set, the test error is not that different (78.04).

	Macro-categories				
	Structural	Location	Trust	Services	Other
Log-lin. Regr.	5	2	6	9	5
Regr. Tree	5	3	5	2	1

Table 3.2: Variable importance.

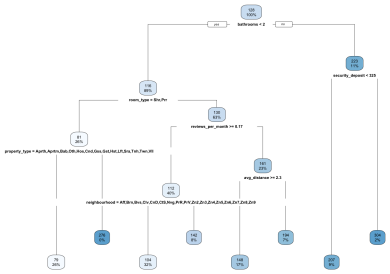


Figure 3.6: Single tree with default parameters.

One of the main disadvantages of trees is, in fact, the high variance: this cause regression trees to have little accuracy, especially on extreme values (Kuhn and Johnson, 2013). Therefore, parameters tuning can be implemented to improve accuracy and prediction performance. First of all, higher complexity is forced (0.001), at the expense of clarity and easy visualizations. The pruned tree obtained with the best complexity parameter - the one that minimizes the error - have an error of 70.31 (on the test set). However, this improvement is associated with much less ease of representation. Thus, the next step is to search a grid of parameters to find the best performing `cp` - that controls complexity - `max_depth` - that can influence overfitting - and `min_split`, to regulate the minimum number of observations per node. The aim is to obtain the best performing model[3]. The improvement in the accuracy, however, is not particularly significant.

As the performance of a single tree is not impressive, the following step of the analysis is to sacrifice visualization for the sake of prediction accuracy. Thus, random forests are implemented. As mentioned in 2.2, the high dimension of the data forces

[3] rpart.control Rdocumentation.

to optimize the process and find a shorter and less-consuming solution. For this reason, this method is implemented through the `ranger` package, that contains a fast implementation for high dimensional data[4].

As for the single tree, the set of parameters in the model is tuned to obtain the highest accuracy. First, a 10 per cent sample of the data set is tested on a larger grid, then the search is refined over a smaller grid, using the full testing dataset. Compared to the previous models, the accuracy greatly improves: the training error is 39.02, 38.96 on the test set. Once the model is proved to fit the data, the variable importance is explored.

The plot 3.7 (a) describing random forest considers the 16 most important variables. At the end of the literature review - precisely, in section 1.3.3 - the attributes that affect the price are divided into 4 macro-categories: *reviews and host reputation*, *location*, *structural attributes* and *services* provided by the listing. By comparing these groups with Figure 3.7 (a), we can identify the most influential categories. The biggest group of attributes - `reviews_per_month`, `review_score_rating`, `days_last_review` for the reviews side, while for the host side `day_host_since`, `host_total_listings_count`, `host_verifications` - refers to trust variables. Location ones - `avg_distance`, `neighbourhood`, `number_pois` - and structural ones - `bathrooms`, `bedrooms`, `room_type` - both accounts for three out of sixteen attributes. The last variables identifiable in one of the four macro categories is `number_amenities` - services offered by the host. Apart from these 4 macro categories, `security_deposit`, `maximum_nights`, `availability_30` are also identified as important. The table below presents the complete summary:

	Macro-categories				
	Structural	Location	Trust	Services	Other
RandonForest	3	3	6	1	3

To account also for the relative importance, the list is narrowed to `bathrooms`, `avg_distance`, `bedrooms`, `room_type`, `reviews_per_month`, the top five variables. Therefore, we can conclude that some basic structural characteristics, the location and the reputation of the hosts might be enough to define a base-price.

3.2.3 Gradient Boosting

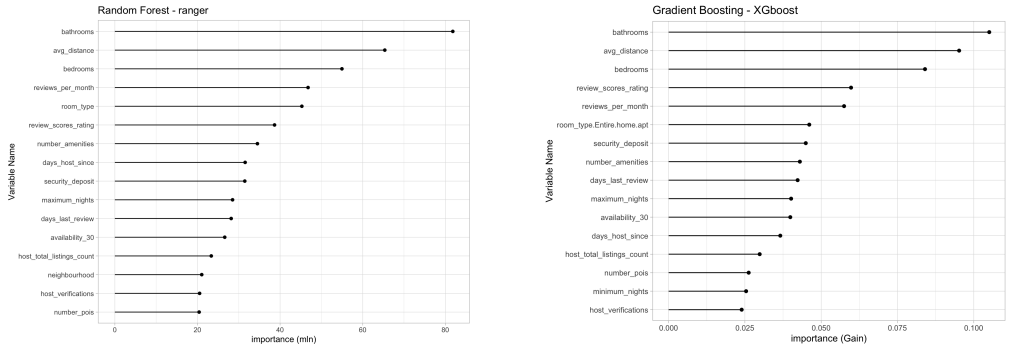
Gradient Boosting concludes the first part of the analysis. Given the large success of this method in a variety of contexts, this research also tests its performance. By sequentially fitting weak learners on the pseudo-residuals, this method constructs an

[4]The function `ranger` is 6 times faster than RandomForest. See documentation: Rdocumentation.

additive regression model (Friedman, 2002). In particular, the following implementation considers only a subset of the training data, therefore falling into the so-called "Stochastic gradient boosting".

As mentioned in 2.2 and later in 3.2.2, the computational cost of implementing this algorithm on the whole dataset is high. Therefore, the analysis in R adopts the package `xgboost`, a faster implementation[5] of gradient boosting with respect to the common `gbm` presented in James et al. (2013). The algorithm uses numerical matrixes. Thus, the first step is to encode factors into dummies[6], and the entire dataset is then converted into a matrix[7]. Then, the analysis follows the same procedure of the random forests. First, a smaller sample of data is used to perform parameters tuning with the `xgb.cv` function and the best parameters are finally included in the final model - implemented with the function `xgboost`. The resulting set of best parameters is as follow:

- `eta` (corresponding to shrinkage in gbm), which describe the learning rate: 0.1;
- `max_depth`, value that controls the depth of each tree (base learner), and therefore complexity: 7;
- `min_child_weight`, the minimum number of entries per final node: 5;
- `subsample`, that in the context of Stochastic GB defines the portion of train observation that model each trees: 0.8;
- `colsample_bytree`, the portion of feature supplied to a tree in the training process: 0.9.



(a) Random forest (b) Gradient boosting

Figure 3.7: Comparison of variables importance.

The errors suggest that the final model suffers from overfitting: the training error (19.08) is certainly much smaller than the test one (35.60). This is probably due

[5] Xgboost presentation accessed December 4th, 2020).

[6] The function `dummyVars` from `caret` is used.

[7] The `xgboost` package provide the `xgb.DMatrix` function.

to the nature of the algorithm that learns from "mistakes" (*pseudo-residuals*), and therefore gives more weight to special cases. Again, the variable importance is checked and compared to the previous list.

Figure 3.7 displays the importance of variables in both random forest and gradient boosting: the similarities are impressive, and we can therefore infer that these attributes play a central role in shaping the price. As plotted in 3.7 (b), the variables are divided among the macro-cateogories as follows:

	Macro-categories				
	Structural	Location	Trust	Services	Other
XGBoost	3	2	6	1	4

Chapter 4

Analysis: demand side

After exploring the pricing mechanisms of Airbnb, this Chapter illustrates the demand side. The author here analyses which elements can identify different types of customers. Different data mining techniques are explored and customers are grouped according to profit potential, past purchases and socio-demographic variables. Simply stated, the aim is to uncover patterns and useful information from a large data set. In particular, segmentation and different clustering algorithms are implemented, to analyze data and group observations independently from any known class labels. For the sake of clarity, the following section only highlights the most interesting results.

4.1 Exploratory Data Analysis

The second part of the analysis wants to identify segments of travellers with similar habits and/or similar characteristics, and capture their opinion towards sharing platforms. First of all, a quick overview on the data forces to reconsider a more general approach: in the data set, only 2282 entries out of 4393 registers the information of booking accommodation via the web, and - out of this subgroup - we have only partial information on the use of `piattall2`. This small number (98 out of 1415 observations) suggests to investigate people that use the internet for booking, and then capture potential sub-segments, in order to have a larger sample of 2282 observations.

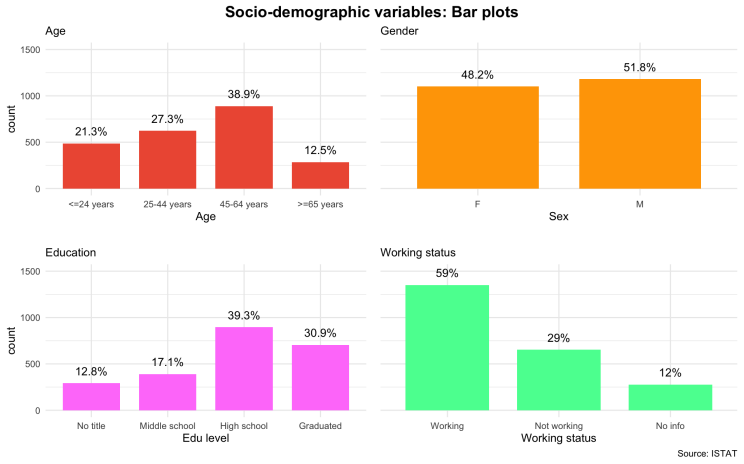


Figure 4.1: Overview on socio-demograpihc variables.

First of all, *Figure* 4.1 presents the distribution of individuals. The presented variables are then investigated for segmentation, as well as booking habits and trip characteristics. As already demonstrated in literature - see Section 1.4 - age is recognized as a central factor in influencing the booking process, together with the level of education. On the other side, gender is not that significant in the booking process, while it is more interesting for the choice of accommodation.

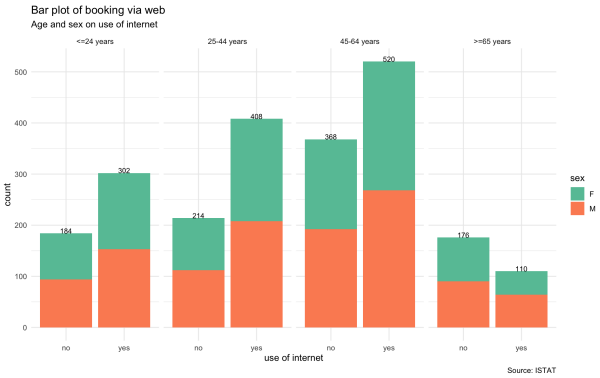


Figure 4.2: Booking accommodation through the internet.

Figure 4.2 gives a complete picture of people using internet for accommodation. Clearly, younger people have a higher tendency to book online accommodation, with a peak in the age range 25-44, as more than 65% entries (408 observations) refers to bookings via web. Individuals under 24 years of age use internet in 62% of occasions (total of 302), followed by the

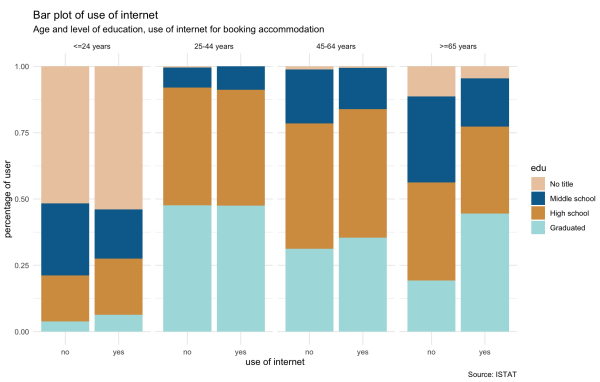


Figure 4.3: Booking accommodation on internet.

age group of 45 to 64 (a total of 520 bookings via web, 58 percent). The elderly rarely use internet, for a total of 110 bookings out of 286 (38%). The relative distribution of gender in each bin - the figure clearly show that the green and orange colour occupies more or less the same amount of space - confirms that this attribute does not play a central role. The effect of level of education, instead, seems mixed (*Figure* 4.3). The proportions remain the same for individuals younger ages, while for 45-64, and especially people older than 65, the difference is tangible. Moreover, it must be considered that percentage in the first age group are highly influenced by the fact that, usually, people get a diploma around 19 years old, and a degree around 22 or more. In the end, we can conclude that age is the main actor, while education has a significant impact on older people.

The same attributes are analysed with respect to use of sharing platforms. As already mentioned, the amount of missing information in `piattall2`, together with the great difference between people using and not using these online tools, result in the complex plot in *Figure* 4.4. Again, gender is confirmed to be a non interesting attribute in exploring booking tendency. Age and level of education are explored in *Figures* 4.4 and 4.5.

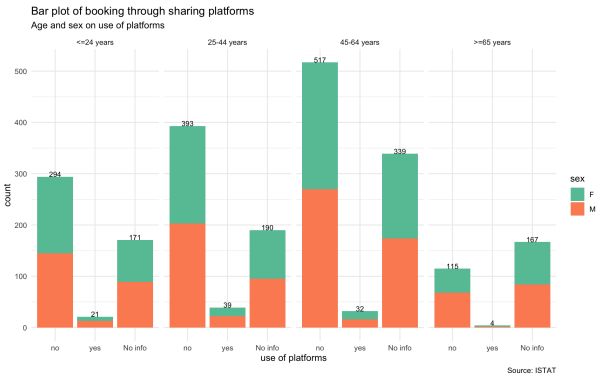


Figure 4.4: Booking accommodation through the internet.

With relatively smaller numbers, booking via online platforms copies the previous trends: with a 9% of online booking, people between 25 and 44 represent the peak of the graph, followed by young people (6.6%), the age group of 45 to 64 (5.8%) and, in the end, elderly people (with a small 3.3%) - see 4.4, the percentage from the absolute numbers presented.

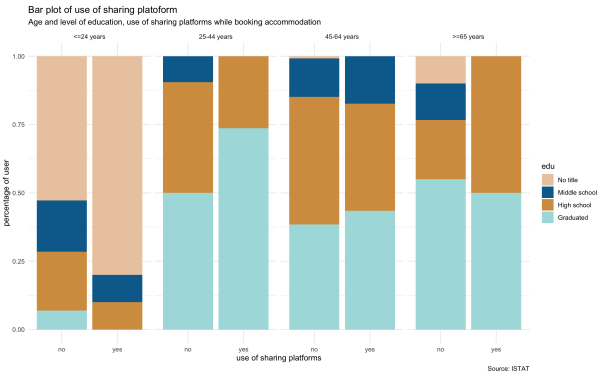


Figure 4.5: Booking accommodation on sharing platforms.

To conclude, *figure* 4.5 represents booking habits per age group according to a different level of education. As before, the plot suggests that for older people the

positive effect of education is stronger. However, absolute numbers highly influence the distribution - only 4 people over 65 used sharing platforms - and therefore conclusions need to be more prudent.

4.2 RFM analysis

The first adopted model is RFM (Recency, Frequency and Monetary), often used in marketing to quantify customer behaviour (Wei et al., n.d.). As the frequency variable is recorded as the number of trips in only one year, the ISTAT data set certainly has some limitations. Simply stated, `progvia` covers the range 1 to 4, with a highly skewed distribution[1]. However, the analysis is carried out and individuals are clustered based on their profile and then assigned to a group, according to the three-digit rfm score[2].

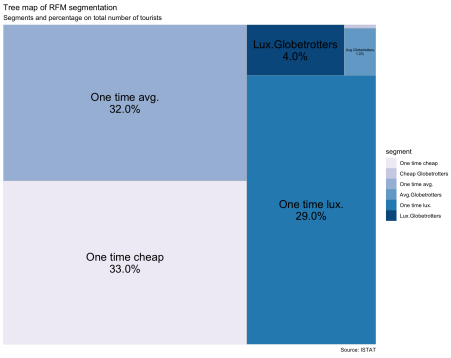


Figure 4.6: RFM segments.

First of all, the three measures are computed on the subset: recency is coded as days from the last trips (considering December 31st, 2019 as reference), frequency as the number of distinct trips, monetary as the overall expense for every individual during the last year. Results are computed on a three-bins scale[3], resulting in 27 possible combinations. Starting from these combinations, segments can be created in a number that usually varies between 3 and 11-12. For this study, a total of 6 different groups are identified - see *Figure* 4.6 -, according to the number of trips and the average amount of money spent:

- *one time luxury travellers*: individuals that travelled only once during the year, spending a great amount of money (591-4970 euros/person) on that single trip; higher ratio expenses/number of trips;
- *luxury globetrotters*: people that travelled 2 to 4 times in the analysed period, spending a considerable total amount of money;
- *one time average-spending travellers*: individuals that travelled only once during the year, spending between 251 and 590 euro per person;

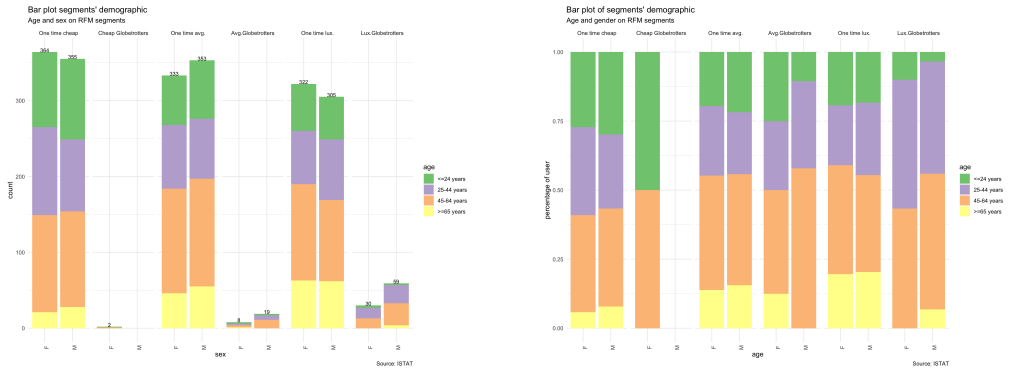
[1] 2118 out of 2282 entries in the subset refers to individuals that travelled just once during the last year, 145 travelled twice, only 17 three times and just a couple of entries register 4 `progvia`.

[2] As presented in Wei et al. (2010), another approach requires to give each score the same weight and calculate a composite score as total. The choice here is dictated by the decision to classify customers according to different values of each score, as in accessed November 2nd, 2020).

[3] In literature the default number is five. However, since the frequency range is uncommonly small, the parameter in the R function is manually changed.

- *average globetrotters*: travellers that made 2 or 3 trips in the year, spending an average amount of money;
- *one time cheap travellers*: people that booked a single trip in 2019, spending between 20 and 250 euro per person;
- *cheap globetrotters*: people that travelled between 2 and 4 time in 2019, smaller ratio expenses/number of trips;

The six groups are highly unbalanced, as the general tendency shows that people prefer to travel just once per year[4]. However, socio-demographic characteristics are investigated to highlight any differences.



(a) Absolute distribution

(b) Percentage distribution

Figure 4.7: Socio-demographics attributes in RFM segments.

The two plots in 4.7 show, on the left side the absolute distribution, while the (b) plot display percentages. Differences do not appear to be very significant, but we also need to consider the great imbalance of absolute numbers in the plot (a). The only interesting fact is that the percentage of younger people travelling decreases while the average cost of the trip increase. On the other side, the elderly tend to spend more on their trips. However, this correlation might be mostly due to the fact that individuals younger than 24 years old are rarely working, and therefore has less money to spend on trips.

To conclude, *Table* 4.1 completes the analysis. By dividing people according to the number of trips, the number of days since the last one and the total amount of money spent travelling some elements clearly emerge. After reporting the use of the internet and sharing platforms[5], conclusions are drawn. First of all, travelling more than once per year seems strictly related to business trips. Also, age appears to be connected with the available budget, while gender does not play a role: the major-

[4] A total of 2032 individuals travelled just once, as aforementioned.

[5] In order to obtain a numeric summary, RFM information is merged to the numeric data set used for k-means clustering. Therefore, each dummy is changed into a binary 1-0 Variables are coded as follow: *Gender Female* is recorded as 1, as well as *Work Yes*, *Trip Holiday* and *Destination Italy*.

ity of younger travellers belongs to *cheap globetrotters*, decreasing in each "upper" group. Focusing on the booking process, the percentage of people using the internet for accommodation changes between 50 and 70 per cent with no clear pattern (the only exception are average globetrotters). One explanation might be, again, the age distribution. The house-sharing phenomenon registers unexpected results: despite what the literature review points out, there is no correlation between "cheap travellers" the use of platforms such as Airbnb. The group *Cheap Globetrotters* registers an incredibly high percentage of use of sharing platform, but the RFM segments are clearly unbalanced - this segment only counts two individuals.

	Size	Gender	Work status	Trip Type	Dest.	Days	Monetary	Avg /trip	Use Internet	Use shar. platf.
OT cheap	719	0.506	0.580	0.896	0.900	2.63	163	163	0.598	0.0292
CheapGlobe	2	1	0	0.75	1	1.75	178	88.8	0.25	0.5
OT avg.	686	0.485	0.548	0.911	0.754	4.45	391	391	0.631	0.0612
AvgGlobe	27	0.296	0.741	0.463	0.889	1.76	426	207	0.370	0.0185
OT lux.	627	0.514	0.576	0.917	0.510	8.91	1084	1084	0.523	0.0399
Lux.Globe	89	0.337	0.809	0.502	0.660	3.49	1134	543	0.676	0.0337

Table 4.1: Summary table for RFM segments.

4.3 K-means

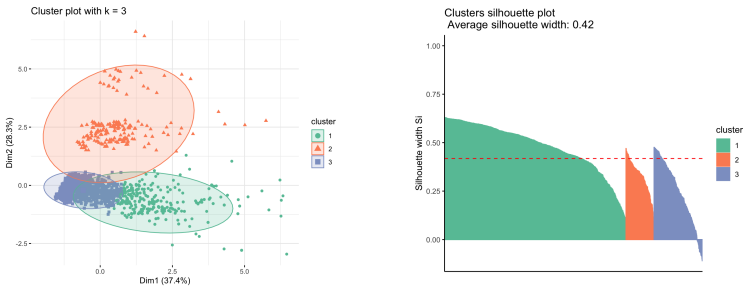
Having explored the profitable side of each individual by identifying the *rfm parameters*, the k-means algorithm is performed, on the same set of variable, to confirm the consistency of the chosen categories and add some further considerations. Then, the next section discusses categorical variables and the hierarchical clustering method.

As mentioned, the *k-means* is performed on a small subset of the numeric version of the data. *Rfm parameters* are simply included in the clustering data[6], together with the average number of days of vacations per individuals, and the algorithm is implemented (Khajvand, 2011).

One last element to keep in mind is that, for the nature of the k-means algorithm, the partitioning is more affected by attributes with a larger variation, then a narrow range. For this reason, standardization is here applied.

This approach for partitioning allows obtaining k non-overlapping cluster (James et al., 2013). Since the algorithm requires a priori the k parameter, the first step is to find the best number of partitions. The measures of silhouette and between and within-cluster sum of squares suggest 3 as the best number of cluster. *Figure* 4.8 (b) show that the partitioning has some limitations and that some observations are misclassified, although the average silhouette is 0.41, sufficiently good (this value

[6] A first trial also comprehends age and level of education, but since these attributes are not continuous numeric variables - age is in range and education is a categorical ordinal variable - this was incorrect and it has therefore been deleted.



(a) Cluster plot. (b) Silhouette.

Figure 4.8: Evaluation of k-means, $k = 3$.

range from 1 when the observation is well clustered, -1 otherwise). However, the 3 resulting groups have some interesting implications.

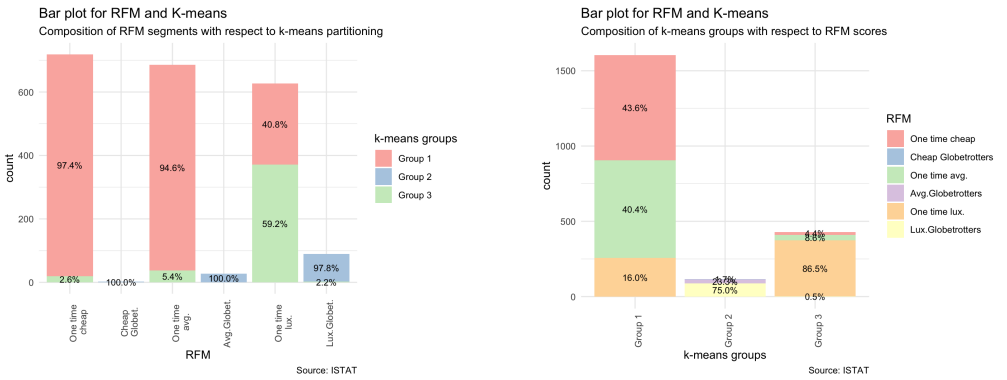
First of all, *group 3* identifies those individuals that spend the greatest amount of money on trips, and that - on average - travel for a higher number of days - see *Table* 4.2 - for interpretation, see 4.2, that gives the reference for binary values. The peculiarity of people in *group 2*, instead, is the number of trips per year. *Group 3*, instead, collects the one-time travellers that over the year spend little money on vacations. Information on use of internet and sharing platforms is also added to the table, in order to check if segments have any differences.

	Size	Gender	Work. stat.	Trip Type	Dest.	Days	Recen.	Freq.	Monetary	Use internet	Use shar. plat.
K1	1605	0.491	0.573	0.899	0.778	3.43	179	1	344	0.612	0.0467
K2	248[a]	0.336	0.784	0.336	0.707	3.16	157	2.12	950	0.612	0.0345
K3	429	0.541	0.548	0.939	0.555	11.8	167	1.00	1202	0.610	0.0303

[a] The number of individuals (Figure 4.9 (b)) is way smaller, as here the algorithm counts every single trip.

Table 4.2: Summary table for k-means partitions.

The first important element to highlight is the - at least apparent - correlation between total expenses and use of sharing platform. Moreover, by clustering on the same data as RFM, k-means secures the consistency of the groups by identifying three macro clusters: "cheap travellers", globetrotters, and "luxury travelers". The same conclusion is confirmed by the following plots - 4.9 (b). In fact, by plotting RFM information on the partitions resulting from k-means, one recognizes correspondence between RFM segments and the three aforementioned groups.



(a) RMF on k-means groups.

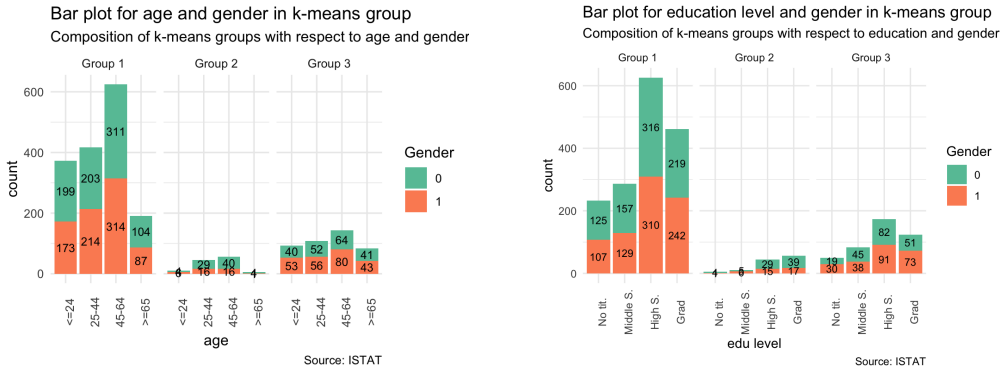
(b) k-means on RFM segments.

Figure 4.9: Different grouping.

By looking at the (b) subplot in *Figure* 4.9 and checking for the silhouette output[7], the first important thing to notice is that group 2, a total of 248 individuals, counts almost exclusively *globetrotter* - namely, individuals that travelled more than once in 2019. Also, it is easily noticeable that the majority of entries in group 3 refers to "One-time luxury" travellers, while only a small percentage (around 13%) are one-time cheap and average, as the majority of entries in groups 1. The first hypothesis was that this fraction actually belongs to this group, as *Figure* 4.8 (b) show that a small portion of group 3 entries are misclassified. However, further analysis shows that, although the misclassified entries of groups 3 actually belong to 1, this small number of entries refers to "One-time luxury" travellers. The reason is that, although the RFM profile is different, these travellers do not fit perfectly in any group. Simply stated, these individuals spend an amount of money that is more similar to group 2 (around 900 euros, on average). The average duration is 7 days, perfectly in between 3 (group 1) and 11 (group 3). The number of trips is almost identical - 1 for misclassified entries and 1.004662 for group 3 - but since the variables are in a 1 to 4 range, the profiles better fit the group with identical total count: group 1.

Overall, k-means provides some interesting results. As already mentioned, being able to define three specific groups with different travelling habits is fundamental while setting listings' price. To conclude, k-means' groups are better defined with respect to socio-demographic variables and travelling preferences. While education does not seem to be relevant, Figure 4.10 (b) shows that age has different distribution across groups. The majority of young travellers, in fact, belongs to group 1, while older people are more evenly distributed.

[7]The silhouette function in the cluster package gives a three-column data frame output: *cluster* - assigned group - *neighbor* - closer cluster - and *sil_width* - silhouette value. For further information, see accessed November 5th, 2020).

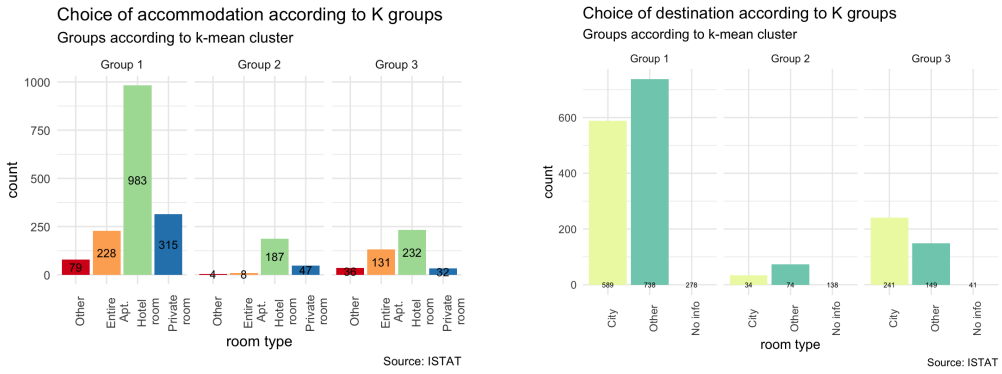


(a) Age slots per k-mean group

(b) Education levels per k-mean group

Figure 4.10: Socio-demographic attributes for k-means.

By looking at the following plots - in particular, Figure (a) 4.11 - we notice another interesting element: being the group that uses sharing platform the most, "cheap travellers" includes the majority of people that book private room, typical of the house-sharing context - in Milan, this `room_type` account for the 25% of the totality of Airbnb listings. However, entire apartments are also largely available - more than 70% of ads in the city of Milan: as a good percentage of group 3 chose either a room or an apartment, this segment might hide some potential customers for sharing platforms. Moreover, focusing in particular on the case study, the majority of these people preferred to visit a city.



(a) Type of accommodation booked per group

(b) Type of destination per k-mean group

Figure 4.11: Travelling preferences for k-means.

4.4 Hierarchical cluster

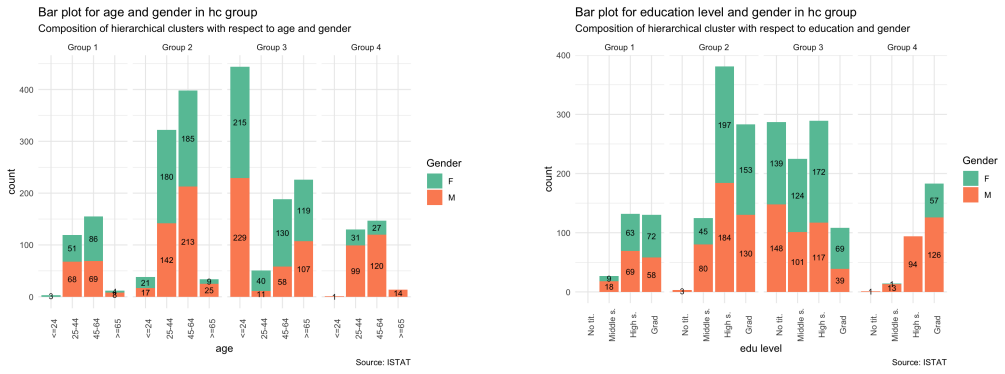
This section focuses on behavioural characteristics and socio-demographic attributes. The previous section briefly mentioned one of the limits of k-means: because of its

distance computation, it is not possible to add categorical variables to the model. Therefore, to explore socio-demographical attributes such as age, gender or education, hierarchical cluster with *Gower Distance* is implemented[8]. Together with recency, frequency and monetary, the distance matrix comprehends gender, age, level of education, working status, type of trip (holiday/business trip) and number of days. The variables use of the internet (`IORGALL`) and/or use of sharing platform (`piattall2`) to book accommodation are explored once the clusters are defined. The tuning process suggests using the *Ward method* to compute the distance between clusters, suggests a partition either in three or four clusters. However, since the 4-groups segmentation results in a better partitioning, this choice is preferred. The schematic summary, where categorical data are changed into numbers for the sake of briefness, presents the main results.

	Size	Gender	Work. stat.	Trip Type	Dest.	Days	Recen.	Freq.	Monetary	Use internet	Use shar. plat.
HC1	289	0.498	0.962	0.965	0.0104	6.60	172	1.07	926	0.668	0.0657
HC2	792	0.499	0.997	0.963	0.955	4.43	175	1.05	444	0.643	0.0556
HC3	909	0.554	0	0.982	0.746	5.54	175	1.06	528	0.535	0.0319
HC4	292	0.199	0.969	0.0993	0.781	2.87	178	1.64	704	0.521	0.0137

Table 4.3: Summary table for hierarchical cluster partitions, $k = 4$.

Table 4.3 does not provide any further insights, apart from the consideration that business trips - where the almost entirety of bookings involves hotel rooms - corresponds to the lowest percentage of use of sharing platform. In Milan, only 0.08% of total listing are hotel room, therefore proving that sharing platforms are not fitting very well this type of preferences.

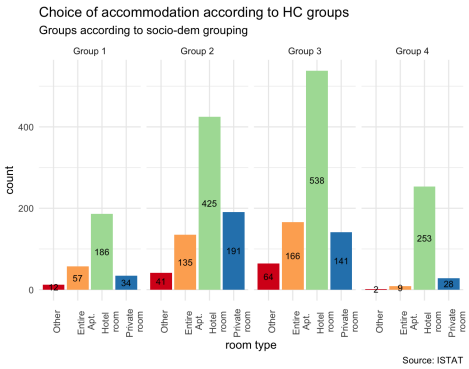


(a) Age slots per hierarchical cluster

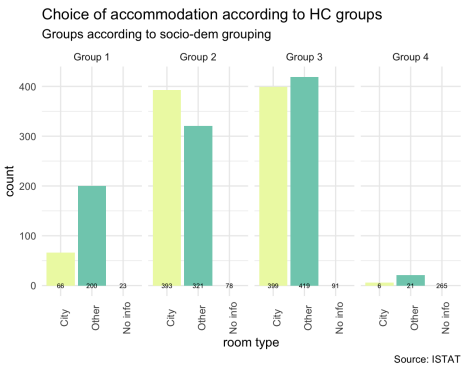
(b) Education levels per hierarchical cluster

Figure 4.12: Socio-demographic attributes for hc groups.

[8]To model categorical variables, the algorithm is applied to the distance matrix created with function *daisy* and the *Gower* metric, which treats each variable differently according to its type - numeric, ordinal or categorical (https://dpmartin42.github.io/posts/r/cluster-mixed-types, accessed October 31st, 2020).



(a) Type of accommodation booked per hierarchical cluster



(b) Type of destination per hierarchical cluster

Figure 4.13: Travelling preferences for hc groups.

Chapter 5

Conclusions

This chapter concludes the research, presenting results, discussing limitations and suggesting further research directions. However, before going into details, a brief answer to the research questions 2.1 is provided. First of all, despite important computational limits, the analysis of Chapter 3 offers a clear description of significant attributes and produces a nice prediction model. Chapter 4, instead, delineates different customer profiles. First, RFM and k-means focus on purchase habits. Then, the hierarchical cluster adapts 4 profiles on the socio-demographic attributes.

5.1 Results

In this section, final results are presented. Focusing on the content of Chapter 3, the price analysis confirms the findings presented in Chapter 1. In fact, the majority of the most significant predictors falls into one of the four macro categories discussed: *reviews and host reputation*, *location*, *structural attributes* and *services*. The table below allows comparing the distribution of significant variables in the different models analysed, counting them according to their macro-category:

	Macro-categories				
	Structural	Location	Trust	Services	Other
Log-lin. Regr.	5	2	6	9	5
LASSO	5	3	10	11	6
Regr. Tree	5	3	5	2	1
RandonForest	3	3	6	1	3
XGBoost	3	2	6	1	4

Concerning the price, another result deserves attention. In fact, as widely proven in the analysis, basic models - such as Linear Regression - fail at obtaining a good prediction. This is mainly due to the high number of outliers and skewed distribution

of the `price` variable: these problems are largely addressed in Chapter 2. However, despite the high prediction error (RMSE), these models offer a good starting point to customize the final price (as the majority of the errors are underestimating the real price). Moreover, being more robust, the non-parametric tree models significantly improve the performance in terms of predictive accuracy, at the cost of increased computational costs. Since the data includes outliers and errors are not normally distributed, we can conclude that tree models have in general a better and adequate performance. As in the table presented in Chapter 2, the errors are displayed below:

Model	Train error	Test Error
Linear Regression	-	72.71
Lasso	-	70.58
Regression Tree	70.25	70.46
Random Forest	39.02	38.96
Gradient Boosting	19.08	35.60

In general, the prediction more than satisfies the research question, as the aim is to understand the pricing mechanism and to offer a good base-prediction.

On the other side, customer segmentation also provides some interesting results. Although it is not possible to fully explore preferences - Istat data only provides some variables, with little or no information on accommodations choices (see next section) - the analysis portrays different profiles. The first approach consists of dividing people according to purchases' records: RFM segmentation results in six groups. The k-means, performed on the same data, confirms the consistency of the groups by identifying three macro-clusters: "globetrotters", "cheap travellers" and "luxury travellers", providing interesting insights into the booking process. In fact, as in table 4.2, as the expenses decrease, the percentage of people using sharing platform for booking accommodation grows (while `IORGALL` - use of the internet to book accommodation - stays the same). This result offers an important element to be kept in mind while defining prices on a sharing platform. Moreover, although most of the people booking hotels belong to the "cheap travellers", individuals with this profile are also the most likely to book a private room. 25% of Airbnb listing are, in Milan, private room: therefore, "cheap travellers" might actually find on Airbnb what they are looking for, both in terms of price and type of accommodation. Therefore, the analysis of purchase history provides interesting points for discussion.

To conclude, the hierarchical cluster, instead, focuses on grouping people according to socio-demographic variables and adds some further insights. In particular, working trips are connected with a smaller percentage of sharing platform use, as most of the bookings involve hotel room. Also, trips abroad are usually more expensive and are concentrated in the first group. It is also interesting to notice that the smaller

monetary value do not correspond to HC3 (not-working people), but to the second one. However, the age plot clarifies the result, as group 3 mainly includes people younger than 24 - who are often studying and seldom working - and retired ones. Group 2, instead, puts together people between 24 and 65, and we can suppose that as they have little time to travel (confirmed by the short average duration of trips), the total expenses are also lower.

5.2 Limitations

As often mentioned during the exposition of the analysis, limitations mostly affect the first part of the analysis - see Section 3.2. The main problem here involves computational costs, as some machine learning algorithm (random forest and especially gradient boosting) require time-consuming parameters' tuning. For this reason, the tuning process includes two steps: a first search is performed on a large grid using only a subset of the training set. Based on these preliminary results, the grid of parameters is reduced. The final tuning is implemented on the entire training set, resulting in the best set of values.

Connected with the dimension of the data set, another great problem is how to deal with missing values. Although the Data Cleaning section widely discusses the topic, further investigation might point out more efficient ways to impute NAs. This is particularly relevant for variables such as reviews, which present many missing values but at the same time are fundamental in assessing trustworthiness, extremely important for any house-sharing platform. New ways of combining existing information might be the key: one example is suggested in Kim and Im (2018).

Another issue is how to identify and treat outliers. In the analysis, we presented some inconsistency in the data and, in the case of Airbnb, these are mainly due to the business itself. In a few words, a business that started as a "simple" house-sharing business, nowadays opens its door also to Bed&Breakfasts, Hostels, and Hotels. This is the most probable explanation for the incongruity between `guests_included` and `accommodates`, but also `room_type` and `bedrooms` available. This is particularly evident in the last-mentioned case: some private or shared rooms count as many as twenty bedrooms.

More in general, the main limitation when trying to predict Airbnb prices is the amount of information that the platform itself lacks (or is not willing to share), such as the current occupancy rate of listings in the neighbourhood, the interaction of single visitors with the websites, the number of times that a single listing has been viewed, and so on. For this reason, the results presented in the previous section can be considered more than satisfactory. In fact, the selected model provides a great

starting point to anyone who wants an estimated base-price.

The main limitation of the second part of the analysis lies again in the data. Chapter 4 wants to investigate booking habits starting from the only freely available data online: this certainly avoids extra costs - connected for example to *ad hoc* surveys - but sets precise limits to the result itself. In particular, given the small total sample available, the author decided to review the use both of internet and of sharing platforms in the booking process. Otherwise, the risk was to obtain completely inconsistent results - as only 98 entries include meaningful values for the variable of interest `piattall2` (use of a sharing platform such as Airbnb to book accommodation).

To conclude, the analysis certainly presents some flaws. A clear limit is the lack of harmony between the data sets from Airbnb and Istat, as essential information on customer's preferences are not registered in the second set. However, the general approach represents an interesting starting point for further studies.

5.3 Further research directions

As presented in the research question in Chapter 2, this dissertation wants to be an exploratory study of the house-sharing market. Therefore, many opportunities are disclosed and still open for further researches.

The previous section offers suggestions to deepen this topic. First of all, as already mentioned, this research aims to be a comprehensive study and covers both the demand and the supply side. However, for such researches, a full comparison of costumers preferences with the goods or services offered on the sharing platform is fundamental. Therefore, as Airbnb data was retrieved directly from their website, the easiest way to match these two sides is with *ad hoc surveys*. In particular, apart from accommodation preferences, budget and booking process, it would be necessary to explore more in depth travellers' profiles: what services are considered crucial in a listing, what type of structure do people usually prefer to stay in, and so on.

Another idea for future researchers is the use of daily prices for prediction. As mentioned in the Exploratory Data Analysis of Chapter 3, the computational limits of the available machines confined the research to a single price per-night, updated on a monthly basis. Hypothetically, the prediction would benefit from a more fine-grained analysis. Moreover, as mentioned in Chapter 2, house and hotel prices can suffer the price of competitors in the surrounding area (Chica-Olmo et al., 2020, causing *spatial autocorrelation* between prices. To control for this externality, positional attributes are here added to the set of available information. However, complex econometrics models might control this phenomenon in a better way, and

also identify and use spatial spillover to improve predictions.

Considering the data available, this last section leaves to further research also the possibility for better data cleaning and imputation. In particular, as widely discussed in Chapter 2, effectively treating and identifying outliers might be the changing point for accuracy. In fact, also in this brief exploration, the presence of outliers not only in the price but also in the other variables greatly affects the performance of the models. Chapter 2 provides a glimpse of the issue and also presents the different approaches to deal with them. Missing values are a great concern as well: data pre-processing and feature engineering can certainly help. As suggested in the limitations, combining different information and techniques might be the game-changer.

Overall, the exploratory nature of this thesis lays the foundations for many further pieces of research.

Appendix

The complete code is available at the following link Github:
https://github.com/vero203602/Airbnb_price_and_segmentation.

Data cleaning

NA detection and imputation

```
gg_miss_var(l2019) + theme_bw()

miss_variables = gg_miss_var(l2019)$data$n_miss
col_names = gg_miss_var(l2019)$data$variable

do.call(rbind, Map(data.frame, names = col_names,
                   miss_var = miss_variables))%>%
  filter(miss_var > 0) %>%
  arrange(desc(miss_var))

l2019$bathrooms <- round(l2019$bathrooms)

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

l2019 %>%
  group_by(bedrooms) %>%
  summarise(
    mode_bath = Mode(bathrooms)
```

```
  )

l2019$bathrooms[is.na(l2019$bathrooms)] <-
                Mode(l2019[l2019$bedrooms ==
                (l2019$bedrooms),]$bathrooms)

l2019 %>%
  group_by(bathrooms) %>%
  summarise(
    mode_bed = Mode(bedrooms)
  )
l2019$bedrooms[is.na(l2019$bedrooms)] <-
                Mode(l2019[l2019$bedrooms ==
                (l2019$bathrooms),]$bedrooms)

l2019 %>%
  group_by(bedrooms) %>%
  summarise(
    mode_b = Mode(beds)
  )
l2019$beds[is.na(l2019$beds)] <-
            Mode(l2019[l2019$bedrooms ==
            (l2019$bedrooms),]$beds)

l2019$host_total_listings_count[is.na
            (l2019$host_total_listings_count)] <-
            round(mean(l2019$host_total_listings_count,
                       na.rm = TRUE)
                  )

l2019[, c(37, 55:61)]
     [is.na(l2019[, c(37, 55:61)])] <- 0
l2019 <- subset(l2019,
                select = -c(square_foot))


l2019 <- l2019 %>%
  mutate_at(vars(host_response_time,
```

```
                host_response_rate,
                host_acceptance_rate),
            na_if, 'N/A') %>%
  mutate_at(vars(host_response_time,
                host_response_rate,
                host_acceptance_rate),
            na_if, '')

l2019 <- l2019[!(l2019$neighbourhood == ""), ]
```

Outlier detection and imputation

```
df_month %>%
  dplyr::select(id, guests_included, accommodates,
                bathrooms, bedrooms, beds,
                price, property_type) %>%
  dplyr::filter(guests_included > accommodates)

df_month <- mutate(df_month,
                   guests_included = case_when(
                     (guests_included > accommodates) ~
                     as.numeric(accommodates),
                     TRUE ~ as.numeric(guests_included)))
df_month %>%
  dplyr::group_by(accommodates) %>%
  dplyr::summarise(guest_included =
                    mean(guests_included))

df_month %>%
  dplyr::group_by(bedrooms) %>%
  dplyr::summarise(guest_included =
                    mean(guests_included),
                   bed_included   =
                    mean(beds))

df_month %>%
  dplyr::select(id, guests_included, accommodates,
                bathrooms, bedrooms, beds,
                price, property_type, room_type) %>%
  dplyr::filter(bedrooms > accommodates) %>%
```

```
  dplyr::filter(room_type == 'Entire home/apt'
                | room_type == 'Hotel room')

df_month %>%
  dplyr::select(id, guests_included, accommodates,
                bathrooms, bedrooms, beds,
                price, property_type, room_type) %>%
  dplyr::filter(bedrooms > accommodates) %>%
  dplyr::filter(room_type != 'Entire home/apt'
                | room_type != 'Hotel room')

df_month <- mutate(df_month,
                   bedrooms = case_when(
                     (room_type == 'Private room') ~ 1,
                     (room_type == 'Shared room')  ~ 1,
                     TRUE ~ as.numeric(bedrooms)))
df_month %>%
  dplyr::group_by(bedrooms) %>%
  dplyr::summarise(guest_included =
                    mean(guests_included),
                   bed_included   =
                   mean(beds),
                   accomm         =
                   mean(accommodates))

df_month[df_month$bedrooms == 8
        | df_month$bedroom == 13 ,]
df_month      <- df_month[! df_month$bedrooms == 8
                          | df_month$bedroom == 13 ,]
df_month %>%
  dplyr::group_by(guests_included) %>%
  dplyr::summarise(bedrooms     = mean(bedrooms),
                   bed_included = mean(beds),
                   accomm       = mean(accommodates))
df_month[df_month$guests_included == 16 ,]
df_month <- df_month[!df_month$guests_included == 16 ,]

df_month %>%
  dplyr::group_by(bathrooms) %>%
```

```
  dplyr::summarise(guest_included =
                     mean(guests_included),
                   bed_included   =
                     mean(beds))

subset_bat <- df_month %>%
  dplyr::select(id, guests_included, accommodates,
                bathrooms, bedrooms, beds,
                price, property_type, room_type) %>%
  dplyr::filter(bathrooms >= 7)

df_month <- mutate(df_month,
                   bathrooms = case_when(
                     (room_type == 'Private room') ~ 1,
                     (room_type == 'Shared room')  ~ 1,
                     TRUE ~ as.numeric(bathrooms)))
df_month %>%
  dplyr::group_by(bathrooms) %>%
  dplyr::summarise(guest_included =
                     mean(guests_included),
                   bed_included   =
                     mean(beds))

df_month %>%
  dplyr::select(guests_included, beds,
                bathrooms, bedrooms) %>%
  dplyr::group_by(guests_included) %>%
  dplyr::summarise(bathrooms = mean(bathrooms),
                   bedrooms  = mean(bedrooms),
                   beds      = mean(beds))
```

Analysis: price prediction

Ranger parameter tuning: Random Forest

```
ranger_grid2 <- expand.grid(
  mtry       = seq(12, 18, by = 1),
  node_size  = c(15, 20),
  sampe_size = .8,
```

```
  OOB         = 0,
  RMSE        = 0
)

for(i in 1:nrow(ranger_grid2)) {

  set.seed(11)
  model <- ranger(
    formula         = price ~ .,
    data            = trainset,
    num.trees       = 500,
    mtry            = ranger_grid2$mtry[i],
    min.node.size   = ranger_grid2$node_size[i],
    sample.fraction = ranger_grid2$sampe_size[i],
  )
  predRanger  <- predict(model, data = testset)
  predR       <- predRanger$predictions
  RMSEranger1 <- round(RMSE(predR, testset$price), 3)

  ranger_grid2$OOB[i]  <- sqrt(model$prediction.error)
  ranger_grid2$RMSE[i] <- RMSEranger1

}

ranger_grid2 %>%
  dplyr::arrange(RMSE) %>%
  head(10)
```

XGBoost parameter tuning: Gradient Boosting

```
xgb_tune <- expand.grid(
  eta              = c(.05, .1),
  max_depth        = 7,
  subsample        = c(.65, .8),
  min_child_weight = c(5, 10),
  colsample_bytree = c(.75, .9),
  trees_best       = 0,
  RMSE             = 0
)
```

```
for(i in 1:nrow(xgb_tune)) {

  set.seed(11)
  xgbtuning <- xgb.cv(
    params              = list(
      eta               = xgb_tune$eta[i],
      max_depth         = xgb_tune$max_depth[i],
      min_child_weight  = xgb_tune$min_child_weight[i],
      subsample         = xgb_tune$subsample[i],
      colsample_bytree  = xgb_tune$colsample_bytree[i]
    ),
    data                    = train_matrix,
    nrounds                 = 1500,
    nfold                   = 5,
    objective               = "reg:squarederror",
    verbose                 = 0,
    early_stopping_rounds   = 50
  )

  xgb_tune$trees_best[i] <- which.min(xgbtuning$
                              evaluation_log$test_rmse_mean)
  xgb_tune$RMSE[i]       <- min(xgbtuning$
                              evaluation_log$test_rmse_mean)
}

xgb_tune %>%
  dplyr::arrange(RMSE)%>%
  head(5)

best_parameters <- xgb_tune %>%
  summarise(
    min_RMSE          = min(RMSE),
    eta               = xgb_tune$eta
                         [which.min(RMSE)],
    max_depth         = xgb_tune$max_depth
                         [which.min(RMSE)],
    subsample         = xgb_tune$subsample
                         [which.min(RMSE)],
    min_child_weight  = xgb_tune$min_child_weight
```

```
                        [which.min(RMSE)],
    colsample_bytree = xgb_tune$colsample_bytree
                        [which.min(RMSE)]
  )
```

Analysis: Segmentation

RFM

```
rfm_data <- iorgall %>%
  group_by(progind) %>%
  dplyr::summarise(revenue= sum(ESPE_CO),
                   most_recent_visit = min(date),
                   number_of_orders =n_distinct(progvia),
                   recency_days = min(recency))

colnames(rfm_data) <- c('customer_id',
                        names(rfm_data)[2:5])
date_rfm <- lubridate::as_date("2019-12-31")
rfm_data$most_recent_visit <-
                as.Date(rfm_data$most_recent_visit)

rfm_df_results <- rfm_table_customer(rfm_data, customer_id,
                                     number_of_orders,
                                     recency_days,
                                     revenue, date_rfm,
                                     recency_bins = 3,
                                     frequency_bins = 3,
                                     monetary_bins = 3)

most_prof   <- c(123, 133, 223, 233, 323, 333)
onetime_lux <- c(113, 213, 313)
avg_globe   <- c(132, 222, 232, 323, 332)
avg         <- c(112, 212, 312)
cheap_often <- c(131, 231, 331)
less_p      <- c(111, 121, 211, 221, 311, 321)

rfm_segment <- as.vector(rfm_df_results$rfm$rfm_score)
rfm_segment[which(rfm_segment %in% onetime_lux)] <-
```

```
                                    "One time lux."
rfm_segment[which(rfm_segment %in% most_prof)]   <-
                                  "Lux.Globetrotters"
rfm_segment[which(rfm_segment %in% avg)]         <-
                                    "One time avg."
rfm_segment[which(rfm_segment %in% avg_globe)]   <-
                                  "Avg.Globetrotters"
rfm_segment[which(rfm_segment %in% less_p)]      <-
                                    "One time cheap"
rfm_segment[which(rfm_segment %in% cheap_often)] <-
                                 "Cheap Globetrotters"
```

Bibliography

Abrahao, B., Parigi, P., Gupta, A., & Cook, K. S. (2017). Reputation offsets trust judgments based on social biases among Airbnb users. *Proceedings of the National Academy of Sciences*, *114*(37), 9848–9853. https://doi.org/10.1073/pnas.1604234114

Agarwal, N., & Steinmetz, R. (2019). Sharing Economy: A Systematic Literature Review. *International Journal of Innovation and Technology Management*, *16*(06), 1930002. https://doi.org/10.1142/S0219877019300027

Anderson, C. (2011). The Impact of Social Media on Lodging Performance. *Cornell Hospitality Report*, *12*(15), 6–11. https://scholarship.sha.cornell.edu/cgi/viewcontent.cgi?article=1004&context=chrpubs

Andersson, D. E. (2010). Hotel attributes and hedonic prices: An analysis of internet-based transactions in Singapore's market for hotel rooms. *The Annals of Regional Science*, *44*(2), 229–240. https://doi.org/10.1007/s00168-008-0265-4

Archer, W. R., Gatzlaff, D. H., & Ling, D. C. (1996). Measuring the Importance of Location in House Price Appreciation. *Journal of Urban Economics*, *40*(3), 334–353. https://doi.org/10.1006/juec.1996.0036

Arimond, G., & Elfessi, A. (n.d.). A Clustering Method for Categorical Data in Tourism Market Segmentation Research, 7.

Babić Rosario, A., de Valck, K., & Sotgiu, F. (2020). Conceptualizing the electronic word-of-mouth process: What we know and need to know about eWOM creation, exposure, and evaluation. *Journal of the Academy of Marketing Science*, *48*(3), 422–448. https://doi.org/10.1007/s11747-019-00706-1

Bailey, M. J., Muth, R. F., & Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, *58*(304), 933–942. http://www.jstor.org/stable/2283324

Barron, K., Kung, E., & Proserpio, D. (2018). The Sharing Economy and Housing Affordability: Evidence from Airbnb. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3006832

Barron, K., Kung, E., & Proserpio, D. (2020). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *Marketing Science*, mksc.2020.1227. https://doi.org/10.1287/mksc.2020.1227

Battiti, R., & Brunato, M. (2014). *The LION Way: Machine Learning plus Intelligent Optimization*.

Bayler, M. J., Muth, R., & Nourse, H. (1963). A regression method for real estate price index construction. *American Statistical Association Journal*, *58*, 933–942. https://doi.org/http://dx.doi.org/10.1080/01621459.1963.10480679

Belk, R. (2014). You are what you can access: Sharing and collaborative consumption online. *Journal of Business Research*, *67*(8), 1595–1600. https://doi.org/10.1016/j.jbusres.2013.10.001

Benkler, Y. (2004). Sharing Nicely: On Shareable Goods and the Emergence of Sharing as a Modality of Economic Production. *The Yale Law Journal*, *114*(2), jstor 10.2307/4135731, 273. https://doi.org/10.2307/4135731

Berger, J., Sorensen, A. T., & Rasmussen, S. J. (2010). Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Marketing Science*, *29*(5), 815–827. https://doi.org/10.1287/mksc.1090.0557

Blal, I., & Sturman, M. C. (2014). The Differential Effects of the Quality and Quantity of Online Reviews on Hotel Room Sales. *Cornell Hospitality Quarterly*, *55*(4), 365–375. https://doi.org/10.1177/1938965514533419

Böcker, L., & Meelen, T. (2017). Sharing for people, planet or profit? Analysing motivations for intended sharing economy participation. *Environmental Innovation and Societal Transitions*, *23*, 28–39. https://doi.org/10.1016/j.eist.2016.09.004

Botsman, R., & Rogers, R. (2010). *What's Mine is Yours*. Collins.

Bragoudakis, Z., Emiris, M., & Constantinescu, M. (2016). Selected Review of the Empirical Literature on House Price Modelling and Forecasting: What Does the Literature Say? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3160945

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Burgess, S., Sellitto, C., Cox, C., & Buultjens, J. (2009). User-Generated Content (UGC) in Touris: benefit and concerns of online consumers. *17th European Conference on Information Systems*.

Cai, Y., Zhou, Y., (Jenny) MA, J., & Scott, N. (2019). Price Determinants of Airbnb Listings: Evidence from Hong Kong. *Tourism Analysis*, *24*(2), 227–242. https://doi.org/10.3727/108354219X15525055915554

Camilleri, M. A. (2018). Market Segmentation, Targeting and Positioning. *Travel Marketing, Tourism Economics and the Airline Product*, 17. https://doi.org/10.1007/978-3-319-49849-2_4

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, *22*(3), 453–474. https://doi.org/10.1016/0166-0462(92)90039-4

Can, A. (1998). GIS and Spatial Analysis of Housing and Mortgage Markets. *Journal of housing research*, *Vol. 9*(1), 61–86. https://doi.org/10.2307/24833659

Casaló, L. V., Flavián, C., Guinalíu, M., & Ekinci, Y. (2015). Avoiding the dark side of positive online consumer reviews: Enhancing reviews' usefulness for high risk-averse travelers. *Journal of Business Research*, *68*(9), 1829–1835.

Celata, F. (2007). Geographic marginality, Transport accessibility and Tourism development, 10.

Chapman, C., & Feit, E. M. (2015). *R for Marketing Research and Analytics*. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-14436-8

Chatterjee, P., & Kumar, A. (2017). Consumer willingness to pay across retail channels. *Journal of Retailing and Consumer Services*, *34*, 264–270. https://doi.org/10.1016/j.jretconser.2016.01.008

Chattopadhyay, M., & Mitra, S. K. (2019). Do airbnb host listing attributes influence room pricing homogenously? *International Journal of Hospitality Management*, *81*, 54–64. https://doi.org/10.1016/j.ijhm.2019.03.008

Chen, P.-Y., Wu, S.-y., & Yoon, J. (2004). The Impact of Online Recommendations and Consumer Feedback on Sales, 711–724. https://doi.org/10.1145/1134707.1134743

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications*, *8*(5), 241–251. https://doi.org/10.1016/j.elerap.2009.03.002

Chen, Y., & Xie, K. (2017). Consumer valuation of Airbnb listings: A hedonic pricing approach. *International Journal of Contemporary Hospitality Management*, *29*(9), 2405–2424. https://doi.org/10.1108/IJCHM-10-2016-0606

Chen, Y., Fay, S., & Wang, Q. (2003). Marketing Implications of Online Consumer Product Reviews, 36.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58–70. https://doi.org/10.1016/j.ijhm.2018.04.004

Chica-Olmo, J., González-Morales, J. G., & Zafra-Gómez, J. L. (2020). Effects of location on Airbnb apartment pricing in Málaga. *Tourism Management*, *77*, 103981. https://doi.org/10.1016/j.tourman.2019.103981

Choudhary, P., Jain, A., & Baijal, R. (2018). Unravelling Airbnb Predicting Price for New Listing. *General Finance*, 10. https://arxiv.org/abs/1805.12101

Codagnone, C., Abadie, F., & Biagi, F. (2016). *The future of work in the 'sharing economy'.* LU, Publications Office. Retrieved September 12, 2020, from https://data.europa.eu/doi/10.2791/431485

Conyette, M. (2011). Demographics for Segmentation in Online Travel, 6.

De Maeyer, P., & Estelami, H. (2013). Applying the peak-end rule to reference prices. *Journal of Product & Brand Management*, *22*(3), 260–265. https://doi.org/10.1108/JPBM-04-2013-0290

de Castro Cardoso, D. M. (2018). What reasons lead guests to choose Airbnb and Booking.com to book peer to peer rentals? A comparison between the two major booking platforms. *Final Dissertation - Master in Management.* https://repositorio-aberto.up.pt/bitstream/10216/117167/2/300875.pdf

Deboosere, R., Kerrigan, D. J., Wachsmuth, D., & El-Geneidy, A. (2019). Location, location and professionalization: A multilevel hedonic analysis of Airbnb listing prices and revenue. *Regional Studies, Regional Science*, *6*(1), 143–156. https://doi.org/10.1080/21681376.2019.1592699

Dellarocas, C. (2003). The Digitization of word-of-mouth: promise and challenges of online reputation mechanisms, 38. http://ccs.mit.edu/dell/digitization%20of%20word-of-mouth.pdf

Demary, V. (2015). Competition in the Sharing Economy. *IW Institut der deutschen Wirtschaft: Policy Paper*, *19*, 28.

Dogru, T., & Pekin, O. (2017). What do guests value most in Airbnb accommodations? An application of the hedonic pricing approach, *5*(2), 15.

Dolnicar, S. (2008). Market segmentation in tourism. (A. G. Woodside & D. Martin, Eds.). In A. G. Woodside & D. Martin (Eds.), *Tourism management: Analysis, behaviour and strategy.* Wallingford, CABI. https://doi.org/10.1079/9781845933234.0129

Dolničar, S. (2004). Beyond "Commonsense Segmentation": A Systematics of Segmentation Approaches in Tourism. *Journal of Travel Research*, *42*(3), 244–250. https://doi.org/10.1177/0047287503258830

Dorner, V., Giamattei, M., & Greiff, M. (2020). The Market for Reviews: Strategic Behavior of Online Product Reviewers with Monetary Incentives. *Schmalenbach Business Review*, *72*(3), 397–435. https://doi.org/10.1007/s41464-020-00094-y

Eckhardt, G. M., Houston, M. B., Jiang, B., Lamberton, C., Rindfleisch, A., & Zervas, G. (2019). Marketing in the Sharing Economy. *Journal of Marketing*, *83*(5), 5–27. https://doi.org/10.1177/0022242919861929

Epple, D. (1987). Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products. *The Journal of Political Economy*, *95*(1), jstor 1831299, 59–80.

Ert, E., & Fleischer, A. (2019). The evolution of trust in Airbnb: A case of home rental. *Annals of Tourism Research*, *75*, 279–287. https://doi.org/10.1016/j.annals.2019.01.004

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*, *55*, 62–73. https://doi.org/10.1016/j.tourman.2016.01.013

European Commission. (2016). The use of collaborative platforms. *Flash Eurobarometer*, *438*.

European Commission. (2018). *Environmental potential of the collaborative economy: Final report and annexes.* LU, Publications Office. Retrieved September 17, 2020, from https://data.europa.eu/doi/10.2779/518554

Fabo, B., & Hudá, S. (2017). Can Airbnb provide livable incomes to property owners?, *AIAS Working Paper*(168), 30.

Falk, M., Larpin, B., & Scaglione, M. (2019). The role of specific attributes in determining prices of Airbnb listings in rural and urban locations. *International Journal of Hospitality Management*, *83*, 132–140. https://doi.org/10.1016/j.ijhm.2019.04.023

Fang, Z., Huang, L., & Wierman, A. (2017, April 3). Prices and Subsidies in the Sharing Economy, In *Proceedings of the 26th International Conference on World Wide Web*. WWW '17: 26th International World Wide Web Conference, Perth Australia, International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3038912.3052564

Farronato, C., & Levin, J. (2015). The rise of peer-to-peer businesses. Global Investor 2.15, Investment Strategy & Research. https://www.oxfordmartin.ox.ac.uk/downloads/GI_215_e_GesamtPDF_01_high.pdf

Federalberghi. (2018). Turismo e shadow economy - Tutela del consumatore, concorrenza leale ed equità fiscale al tempo del turismo 4.0. *Edizioni ISTA*, *Istituto Internazionale di Studi e Documentazione Turistico Alberghiera*.

Feubli, P., & Horlacher, J. (2015). What's the value added of the sharing economy? Global Investor 2.15, Investment Strategy & Research. https://www.oxfordmartin.ox.ac.uk/downloads/GI_215_e_GesamtPDF_01_high.pdf

Filippas, A., & Gramstad, A. R. (2016). A Model of Pricing in the Sharing Economy: Pricing Dynamics with Awareness Generating Adoptions, 17.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How Online Product Reviews Affect Retail Sales: A Meta-analysis. *Journal of Retailing*, *90*(2), 217–232. https://doi.org/10.1016/j.jretai.2014.04.004

Franckx, L., & Mayeres, I. (2016). Future trends in mobility: The rise of the sharing economy and automated transport. *MIND-sets*, 113.

Fränti, P., & Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, *93*, 95–112. https://doi.org/10.1016/j.patcog.2019.04.014

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Garcia-López, M.-À., Jofre-Monseny, J., Mazza, R. M., & Segú, M. (n.d.). Do short-term rent platforms affect housing markets? Evidence from Airbnb in Barcelona. *IEB Working Paper 2019/05*, 38. https://doi.org/doi.org/10.1016/j.jue.2020.103278

Gaspareniene, L., Venclauskiene, D., & Remeikiene, R. (2014). Critical Review of Selected Housing Market Models Concerning the Factors that Make Influence on Housing Price Level Formation in the Countries with Transition Economy. *Procedia - Social and Behavioral Sciences*, *110*, 419–427. https://doi.org/10.1016/j.sbspro.2013.12.886

Gellerstedt, M., & Arvemo, T. (2019). The impact of word of mouth when booking a hotel: Could a good friend's opinion outweigh the online majority? *Information Technology & Tourism*, *21*(3), 289–311. https://doi.org/10.1007/s40558-019-00143-4

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, *35*(1), 46–56. https://doi.org/10.1080/10548408.2017.1308292

Gibbs, C., Guttentag, D., Gretzel, U., Yao, L., & Morton, J. (2018). Use of dynamic pricing strategies by Airbnb hosts. *International Journal of Contemporary Hospitality Management*, *30*(1), 2–20. https://doi.org/10.1108/IJCHM-09-2016-0540

Gierczak, B. (2011). The History of Tourist Transport After the Modern Industrial Revolution. *Polish Journal of Sport and Tourism*, *18*(4), 275–281. https://doi.org/10.2478/v10197-011-0022-6

Giglio, S., Bertacchini, F., Bilotta, E., & Pantano, P. (2020). Machine learning and points of interest: Typical tourist Italian cities. *Current Issues in Tourism*, *23*(13), 1646–1658. https://doi.org/10.1080/13683500.2019.1637827

Goh, S.-K. (2015). The Effect of Electronic Word of Mouth on Intention to Book Accommodation via Online Peer-to-Peer Platform: Investigation of Theory of Planned Behaviour. *The Journal of Internet Banking and Commerce*, *01*(s2). https://doi.org/10.4172/1204-5357.S2-005

Görög, G. (2018). The Definitions of Sharing Economy: A Systematic Literature Review. *Management*, 175–189. https://doi.org/10.26493/1854-4231.13.175-189

Gretzel, U., & Yoo, K. H. (2008). Use and Impact of Online Travel Reviews (P. O'Connor, W. Höpken, & U. Gretzel, Eds.). In P. O'Connor, W. Höpken, & U. Gretzel (Eds.), *Information and Communication Technologies in Tourism 2008*. Vienna, Springer Vienna. https://doi.org/10.1007/978-3-211-77280-5_4

Grewal, D., Gopalkrishnan R Iyer, & Levy, M. (2004). Internet retailing: Enablers, limiters and market consequences. *Journal of Business Research*, *57*(7), 703–713. https://doi.org/10.1016/S0148-2963(02)00348-X

Guttentag, D. (2017). Regulating Innovation in the Collaborative Economy: An Examination of Airbnb's Early Legal Issues (D. Dredge & S. Gyimóthy, Eds.). In D. Dredge & S. Gyimóthy (Eds.), *Collaborative Economy and Tourism*. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-51799-5_7

Guttentag, D. (2019). Progress on Airbnb: A literature review. *Journal of Hospitality and Tourism Technology*, *10*(4), 814–844. https://doi.org/10.1108/JHTT-08-2018-0075

Guttentag, D. A., & Smith, S. L. (2017). Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. *International Journal of Hospitality Management*, *64*, 1–10. https://doi.org/10.1016/j.ijhm.2017.02.003

Guttentag, D., Smith, S., Potwarka, L., & Havitz, M. (2017). Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study. *Journal of Travel Research*, *57*(3), 342–359. https://doi.org/10.1177/0047287517696980

Haining, R. (2001). Spatial autocorrelation (N. J. Smelser & P. B. Baltes, Eds.). In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences*. Oxford, Pergamon. https://doi.org/https://doi.org/10.1016/B0-08-043076-7/02511-0

Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for*

Information Science and Technology, *67*(9), 2047–2059. https://doi.org/10.1002/asi.23552

Hamish, A. (2018). Value of nature implicit in property prices – Hedonic Pricing Method (HPM) methodology note. *Office for National Statistics*, 10. https://www.ons.gov.uk/economy/environmentalaccounts/methodologies/valueofnatureimplicitinpropertypriceshedonicpricingmethodhpmmethodologynote

Harris, R., & Arku, G. (2006). Housing and economic development: The evolution of an idea since 1945. *Habitat International*, *30*(4), 1007–1017. https://doi.org/10.1016/j.habitatint.2005.10.003

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The Element of Statistical Learning* (Vol. Second Edition).

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52. https://doi.org/10.1002/dir.10073

Herath, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research: A review of the literature. *Institute for Regional Development and Environment*, 1–21. https://ro.uow.edu.au/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1977&context=buspapers

Herath, S., & Maier, G. (2011). Hedonic house prices in the presence of spatial and temporal dynamics.

Hill, D. (2015). How much is your spare room worth? *IEEE*, *52*(9), 32–58. https://doi.org/https://ieeexplore.ieee.org/document/7226609

Hong, J. (2018). Rise of the Sharing Economy and the Future of Travel and Tourism Industry. *Journal of Business and Hotel Management*, *07*(02). https://doi.org/10.4172/2169-0286.1000180

Horowitz, J. L. (1992). The role of the list price in housing markets: Theory and an econometric model. *Journal of Applied Econometrics*, *7*(2), 115–129. https://doi.org/10.1002/jae.3950070202

Hossain, M. (2020). Sharing economy: A comprehensive literature review. *International Journal of Hospitality Management*, *87*, 102470. https://doi.org/10.1016/j.ijhm.2020.102470

Hu, M. (Ed.). (2019). *Sharing Economy: Making Supply Meet Demand* (Vol. 6, chapter 8). Cham, Springer International Publishing. https://doi.org/10.1007/978-3-030-01863-4

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology*

and Management, *9*(3), 201–214. https://doi.org/10.1007/s10799-008-0041-2

Huang, L., Yang, Y., Zhao, X., Gao, H., & Yu, L. (2018). Mining the Relationship between Spatial Mobility Patterns and POIs. *Wireless Communications and Mobile Computing*, *2018*, 1–10. https://doi.org/10.1155/2018/4392524

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 22.

Huete-Alcocer, N. (2017). A Literature Review of Word of Mouth and Electronic Word of Mouth: Implications for Consumer Behavior. *Frontiers in Psychology*, *8*, 1256. https://doi.org/10.3389/fpsyg.2017.01256

Iacobini, M., & Lisi, G. (2013). Estimation of a Hedonic House Price Model with Bargaining: Evidence from the Italian Housing Market. *XLI Incontro di Studio del Ce.S.E.T.*, 41–54. https://doi.org/10.13128/Aestimum-13123

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY, Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Johns, N., & Gyimóthy, S. (2002). Market Segmentation and the Prediction of Tourist Behavior: The Case of Bornholm, Denmark. *Journal of Travel Research*, *40*(3), 316–327. https://doi.org/10.1177/0047287502040003009

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, *43*(2), 618–644. https://doi.org/10.1016/j.dss.2005.05.019

Kauko, T. (2006). What makes a location attractive for the housing consumer? Preliminary findings from metropolitan Helsinki and Randstad Holland using the analytical hierarchy process. *Journal of Housing and the Built Environment*, *21*(2), 159–176. https://doi.org/10.1007/s10901-006-9040-y

Khajvand, M. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 7.

Kim, J. B. (2014). Impact of Online Customer Reviews and Incentives on the Product Sales at the Online Retail Store: An Empirical Study on Video Game Titles at Amazon.com, 11.

Kim, J., & Im, J. (2018). Proposing a missing data method for hospitality research on online customer reviews. *International Journal of Contemporary Hospitality Management*, *30*(11), 3250–3267. https://doi.org/10.1108/IJCHM-10-2017-0708

Kim, W. G., Lim, H., & Brymer, R. A. (2015). The effectiveness of managing social media on hotel performance. *International Journal of Hospitality Management*, *44*, 165–171. https://doi.org/10.1016/j.ijhm.2014.10.014

Knoll, K., Schularick, M., & Steger, T. M. (2014). No Price Like Home: Global House Prices, 1870-2012. *CESifo Working Paper Series*, *5006*. https://ssrn.com/abstract=2512724

Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, *9*(5), 374–385. https://doi.org/10.1016/j.elerap.2010.04.001

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY, Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Kung, L.-C., & Zhong, G.-Y. (2017). The optimal pricing strategy for two-sided platform delivery in the sharing economy. *Transportation Research Part E: Logistics and Transportation Review*, *101*, 1–12. https://doi.org/10.1016/j.tre.2017.02.003

Kwok, L., & Xie, K. L. (2019). Pricing strategies on Airbnb: Are multi-unit hosts revenue pros? *International Journal of Hospitality Management*, *82*, 252–259. https://doi.org/10.1016/j.ijhm.2018.09.013

Lang, B., Botha, E., Robertson, J., Kemper, J. A., Dolan, R., & Kietzmann, J. (2020). How to grow the sharing economy? Create Prosumers! *Australasian Marketing Journal (AMJ)*, *28*(3), 58–66. https://doi.org/10.1016/j.ausmj.2020.06.012

Lawani, A., Reed, M. R., Mark, T., & Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. *Regional Science and Urban Economics*, *75*, 22–34. https://doi.org/10.1016/j.regsciurbeco.2018.11.003

Li, Z., Huang, K.-w., & Cavusoglu, H. (2012). Quantifying the Impact of Badges on User Engagement in Online Q&A Communities.

Liang, S., Schuckert, M., Law, R., & Chen, C.-C. (2017). Be a "Superhost": The importance of badge systems for peer-to-peer rental accommodations. *Tourism Management*, *60*, 454–465. https://doi.org/10.1016/j.tourman.2017.01.007

Lisi, G. (2011). Price Dispersion in the Housing Market: The Role of Bargaining and Search Costs, 12.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, *29*(3), 458–468. https://doi.org/10.1016/j.tourman.2007.05.011

Lorde, T., Jacob, J., & Weekes, Q. (2019). Price-setting behavior in a tourism sharing economy accommodation market: A hedonic price analysis of AirBnB hosts in the caribbean. *Tourism Management Perspectives*, *30*, 251–261. https://doi.org/10.1016/j.tmp.2019.03.006

Lu, S., Li, Z., Qin, Z., Yang, X., & Goh, R. S. M. (2017, December). A hybrid regression technique for house prices prediction, In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, IEEE. https://doi.org/10.1109/IEEM.2017.8289904

Lumsden, S.-A., Beldona, S., & Morrison, A. M. (2008). Customer Value in an All-Inclusive Travel Vacation Club: An Application of the RFM Framework. *Journal of Hospitality & Leisure Marketing*, *16*(3), 270–285. https://doi.org/10.1080/10507050801946858

Lutz, C., & Newlands, G. (2018). Consumer segmentation within the sharing economy: The case of Airbnb. *Journal of Business Research*, *88*, 187–196. https://doi.org/10.1016/j.jbusres.2018.03.019

Ma, H. (2015). A Study on Customer Segmentation for E-Commerce Using the Generalized Association Rules and Decision Tree. *American Journal of Industrial and Business Management*, *05*(12), 813–818. https://doi.org/10.4236/ajibm.2015.512078

Mack, C., Su, Z., & Westreich, D. (2018). Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide. *Agency for Healthcare Research and Quality (US)*. https://www.ncbi.nlm.nih.gov/books/NBK493614/

Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House Price Prediction Using Regression Techniques: A Comparative Study, In *2019 International Conference on Smart Structures and Systems (ICSSS)*. 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, IEEE. https://doi.org/10.1109/ICSSS.2019.8882834

Maury, T.-P., & Tripier, F. (2010). Strategies for search on the housing market and their implications for price dispersion. *Working Papers hal-00480484, HAL*, 59. https://hal.archives-ouvertes.fr/hal-00480484/document

McNeil, B. (2020). Price Prediction in the Sharing Economy: A Case Study with Airbnb data. *Honors Theses and Capstones*, *504*, 16. https://scholars.unh.edu/honors/504

Melián-González, S., Bulchand-Gidumal, J., & González López-Valcárcel, B. (2013). Online Customer Reviews of Hotels: As Participation Increases, Better Evaluation Is Obtained. *Cornell Hospitality Quarterly*, *54*(3), 274–283. https://doi.org/10.1177/1938965513481498

Mi, Z., & Coffman, D. (2019). The sharing economy promotes sustainable societies. *Nature Communications*, *10*(1), 1214. https://doi.org/10.1038/s41467-019-09260-4

Mich, L. (2013). La web presence delle destinazioni turistiche, 15.

Miller, N., Peng, L., & Sklarz, M. (2009). House Prices and Economic Growth. *The Journal of Real Estate Finance and Economics*, *42*(4), 522–541. https://doi.org/10.1007/s11146-009-9197-8

Mitrović, D. M., Simović, O., & Raičević, M. (2019). Personalized Marketing in the Function of the Tourist Destination Improvement. *Economics*, *7*(1), 127–137. https://doi.org/10.2478/eoik-2019-0011

Mohamad, I. B., & Usman, D. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303. https://doi.org/10.19026/rjaset.6.3638

Monterrubio, C., Andriotis, K., & Rodríguez-Muñoz, G. (2020). Residents' perceptions of airport construction impacts: A negativity bias approach. *Tourism Management*, *77*, 103983. https://doi.org/10.1016/j.tourman.2019.103983

Moon, H., Miao, L., Hanks, L., & Line, N. D. (2019). Peer-to-peer interactions: Perspectives of Airbnb guests and hosts. *International Journal of Hospitality Management*, *77*, 405–414. https://doi.org/10.1016/j.ijhm.2018.08.004

Muñoz, P., & Cohen, B. (2017). Mapping out the sharing economy: A configurational approach to sharing business modeling. *Technological Forecasting and Social Change*, *125*, 21–37. https://doi.org/10.1016/j.techfore.2017.03.035

Nagaraja, C. H., Brown, L. D., & Zhao, L. H. (2011). An autoregressive approach to house price modeling. *The Annals of Applied Statistics*, *5*(1), arxiv 1104.2719, 124–149. https://doi.org/10.1214/10-AOAS380

Newlands, G., Lutz, C., & Fieseler, C. (2018). Navigating Peer-to-Peer Pricing in the Sharing Economy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3116954

Nieto-García, M., Muñoz-Gallego, P. A., & González-Benito, Ó. (2017). Tourists' willingness to pay for an accommodation: The effect of eWOM and internal reference price. *International Journal of Hospitality Management*, *62*, 67–77. https://doi.org/10.1016/j.ijhm.2016.12.006

Nikoli, G., & Lazakidou, A. (2019). The Impact of Information and Communication Technology on the Tourism Sector. *Almatourism*, *19*. https://doi.org/10.6092/issn.2036-5195/8553

Nowak, B., Allen, T., Rollo, J., Lewis, V., He, L., Chen, A., Wilson, W. N., Costantini, M., Hyde, O., Liu, K., Savino, M., Chaudhry, B. A., Grube, A. M., & Young, E. (2015). Global Insight: Who Will Airbnb Hurt More—Hotels or OTAs. *Morgan Stanley Research*. https://docplayer.net/28102842-Global-insight-who-will-airbnb-hurt-more-hotels-or-otas.html

OECD. (2016, March 9). *OECD Tourism Trends and Policies 2016*. OECD. https://doi.org/10.1787/tour-2016-en

OECD. (2020, March 4). *OECD Tourism Trends and Policies 2020*. OECD. https://doi.org/10.1787/6b47b985-en

Öğüt, H., & Onur Taş, B. K. (2012). The influence of internet customer reviews on the online sales and prices in hotel industry. *The Service Industries Journal*, *32*(2), 197–214. https://doi.org/10.1080/02642069.2010.529436

Oh, H., Parks, S. C., & DeMicco, F. J. (n.d.). Age- and Gender-Based Market Segmentation: A Structural Understanding, 22.

Olimov, F. (n.d.). Dynamic Pricing in a Decentralized Platform: The Case of Airbnb, 46. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=ESWM2019&paper_id=287

Oskam, J., & Boswijk, A. (2016). Airbnb: The future of networked hospitality businesses. *Journal of Tourism Futures*, *2*(1), 22–42. https://doi.org/10.1108/JTF-11-2015-0048

Owusu-Ansah, A. (2013). A review of Hedonic Pricing Models in housing research. *Journal of International Real Estate and Construction Studies*, *1*(1), 19–38.

Pan, B., MacLaurin, T., & Crotts, J. C. (2007). Travel Blogs and the Implications for Destination Marketing. *Journal of Travel Research*, *46*(1), 35–45. https://doi.org/10.1177/0047287507302378

Park, J., & van der Schaar, M. (2010, March). Pricing and Incentives in Peer-to-Peer Networks, In *2010 Proceedings IEEE INFOCOM*. IEEE INFOCOM 2010 - IEEE Conference on Computer Communications, San Diego, CA, USA, IEEE. https://doi.org/10.1109/INFCOM.2010.5461952

Perez-Sanchez, V., Serrano-Estrada, L., Marti, P., & Mora-Garcia, R.-T. (2018). The What, Where, and Why of Airbnb Price Determinants. *Sustainability*, *10*(12), 4596. https://doi.org/10.3390/su10124596

Phelps, M. C., & Merkle, E. C. (n.d.). Classification and Regression Trees as Alternatives to Regression, 2.

Prahalad, C., & Ramaswamy, V. (2004). Co-creating unique value with customers. *Strategy & Leadership*, *32*(3), 4–9.

Qian, F. (2008). A Study on CRM and Its Customer Segmentation Outsourcing Approach for Small and Medium Businesses (L. D. Xu, A. M. Tjoa, & S. S. Chaudhry, Eds.). In L. D. Xu, A. M. Tjoa, & S. S. Chaudhry (Eds.), *Research and Practical Issues of Enterprise Information Systems II*. Boston, MA, Springer US. https://doi.org/10.1007/978-0-387-76312-5_67

Raykov, Y. P., Boukouvalas, A., Baig, F., & Little, M. A. (2016). What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm (B.-J. Yoon, Ed.). *Plos One*, *11*(9), e0162259. https://doi.org/10.1371/journal.pone.0162259

Rimer, R. S. (2017). Why do people choose to stay with Airbnb? *Master Thesis - Master of Science in International Tourism Management*, 66. https://www.semanticscholar.org/paper/Why-do-people-choose-to-stay-with-Airbnb-%C3%96nder-Rimer/074dcb356dfc8dbf7886439e04d71e4907f5c6ef?p2df

Rolando, L. (2018). *Il fenomeno Airbnb e l'abitare contemporaneo*. Politecnico di Torino. https://webthesis.biblio.polito.it/9969/1/tesi.pdf

Roma, P., Panniello, U., & Lo Nigro, G. (2019). Sharing economy and incumbents' pricing strategy: The impact of Airbnb on the hospitality industry. *International Journal of Production Economics*, *214*, 17–29. https://doi.org/10.1016/j.ijpe.2019.03.023

Rosato, P., Breil, M., Giupponi, C., & Berto, R. (2017). Assessing the Impact of Urban Improvement on Housing Values: A Hedonic Pricing and Multi-Attribute Analysis Model for the Historic Centre of Venice. *Buildings*, *7*(4), 112. https://doi.org/10.3390/buildings7040112

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55. https://doi.org/10.1086/260169

Sánchez-Ollero, J. L., García-Pozo, A., & Marchante-Mera, A. (2014). How Does Respect for the Environment Affect Final Prices in the Hospitality Sector? A Hedonic Pricing Approach. *Cornell Hospitality Quarterly*, *55*(1), 31–39. https://doi.org/10.1177/1938965513500709

Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on Customer Segmentation Technique on Ecommerce. *Advanced Science Letters*, *22*(10), 3018–3022. https://doi.org/10.1166/asl.2016.7985

Saussier, J. (2015). Sharing as a disruptive force. Global Investor 2.15, Investment Strategy & Research. https://www.oxfordmartin.ox.ac.uk/downloads/GI_215_e_GesamtPDF_01_high.pdf

Schirmer, P. M., Van Eggermond, M. A., & Axhausen, K. W. (2014). The role of location in residential location choice models: A review of literature. *Journal of Transport and Land Use*, *7*(2), 3. https://doi.org/10.5198/jtlu.v7i2.740

Schmallegger, D., & Carson, D. (2008). Blogs in tourism: Changing approaches to information exchange. *Journal of Vacation Marketing*, *14*(2), 99–110. https://doi.org/10.1177/1356766707087519

Sciarelli, F., Della Corte, V., & Del Gaudio, G. (2018). The evolution of tourism in the digital era: The case of a tourism destination. *Sinergie, Italian journal of management*, *36*(105). https://doi.org/10.7433/s105.2018.09

Segato, G. (2016). *The Sharing Economy*. Università degli studi di Padova. http://tesi.cab.unipd.it/51835/1/Segato_Gianluca.pdf

Sharing or paring? Growth of the sharing economy. (n.d.), 32.

Soler, I. P., & Gemar, G. (2018). Hedonic price models with geographically weighted regression: An application to hospitality. *Journal of Destination Marketing & Management*, *9*, 126–137. https://doi.org/10.1016/j.jdmm.2017.12.001

Standing, C., Standing, S., & Biermann, S. (2019). The implications of the sharing economy for transport. *Transport Reviews*, *39*(2), 226–242. https://doi.org/10.1080/01441647.2018.1450307

Sundararajan, A. (2016). *The sharing economy: The end of employment and the rise of crowd- based capitalism.* Cambridge, MIT Press. http://pinguet.free.fr/sundararajan.pdf

Taeihagh, A. (2017). Crowdsourcing, Sharing Economies and Development. *Journal of Developing Societies*, *33*(2), 191–222. https://doi.org/10.1177/0169796X17710072

Thrane, C. (2007). Examining the determinants of room rates for hotels in capital cities: The Oslo experience. *Journal of Revenue and Pricing Management*, *5*(4), 315–323. https://doi.org/10.1057/palgrave.rpm.5160055

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Torres, E. N., Singh, D., & Robertson-Ring, A. (2015). Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry. *International Journal of Hospitality Management*, *50*, 77–83. https://doi.org/10.1016/j.ijhm.2015.07.012

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, *174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Tsiptsis, K., & Chorianopoulos, A. (2010). *Data Mining Techniques in CRM: Inside Customer Segmentation.* Wiley and Sons, Hoboken. http://dx.doi.org/10.1002/9780470685815

Viglia, G., Minazzi, R., & Buhalis, D. (2016). The influence of e-word-of-mouth on hotel occupancy rate. *International Journal of Contemporary Hospitality Management*, *28*(9), 2035–2051. https://doi.org/10.1108/IJCHM-05-2015-0238

Voltes-Dorta, A., & Sánchez-Medina, A. (2020). Drivers of Airbnb prices according to property/room type, season and location: A regression approach. *Journal of Hospitality and Tourism Management*, *45*, 266–275. https://doi.org/10.1016/j.jhtm.2020.08.015

von Hippel, P. T. (2004). Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician*, *58*(2), 160–164. https://doi.org/10.1198/0003130043204

Wahab, I. N. (2007). Role of Information Technology in Tourism Industry: Impact and Growth. *International Journal of Innovative Research in Computer and Communication Engineering*, (2), 4.

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

Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (n.d.). A review of the application of RFM model. *Afr. J. Bus. Manage.*, 8.

Wu, X., & Zhi, Q. (2016). Impact of Shared Economy on Urban Sustainability: From the Perspective of Social, Economic, and Environmental Sustainability. *Energy Procedia*, *104*, 191–196. https://doi.org/10.1016/j.egypro.2016.12.033

Xie, K. L., & Kwok, L. (2017). The effects of Airbnb's price positioning on hotel performance. *International Journal of Hospitality Management*, *67*, 174–184. https://doi.org/10.1016/j.ijhm.2017.08.011

Xie, K. L., So, K. K. F., & Wang, W. (2017). Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach. *International Journal of Hospitality Management*, *62*, 101–110. https://doi.org/10.1016/j.ijhm.2016.12.004

Xin, S. J., & Khalid, K. (2018). Modelling House Price Using Ridge Regression and Lasso Regression. *International Journal of Engineering & Technology*, *7*(4.30), 498. https://doi.org/10.14419/ijet.v7i4.30.22378

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*(1), 180–182. https://doi.org/10.1016/j.ijhm.2008.06.011

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, *27*(2), 634–639. https://doi.org/10.1016/j.chb.2010.04.014

Zervas, G., Proserpio, D., & Byers, J. (2015). A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2554500

Zervas, G., Proserpio, D., & Byers, J. (2016). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Boston U. School of Management Research Paper*. https://doi.org/https://dx.doi.org/10.2139/ssrn.2366898

Zhang, T. C., Jahromi, M. F., & Kizildag, M. (2018). Value co-creation in a sharing economy: The end of price wars? *International Journal of Hospitality Management*, *71*, 51–58. https://doi.org/10.1016/j.ijhm.2017.11.010