Science and Information Technology Department

Rate management for home holiday rentals: a data analytics approach

João Martins Rebelo Fontoura

Dissertation submitted as partial fulfillment of the requirement for the degree of

Master in Computer Science and Business Management

Supervisor:

Prof. Raul M. S. Laureano Assistant Professor, ISCTE Business School, Department of Quantitative Methods for Management and Economics

October 2019

(This page was intentionally left blank)

Acknowledgments

I would like to thank everyone that believed in me and my effort towards a successful ending to this research work.

First, my sincere gratitude to my Supervisor Raul Laureano, for all the support, availability, guidance, patience, and knowledge provided. I would like to thank him for never letting me give up, this would have not been possible without him.

To Marta, a big thank you for all the support given through the good and bad phases of this dissertation, and for understanding my absence in the last few weeks.

To my friend Duarte, for accompanying me through this stressful final stage of my academic cycle.

To all my family and friends, for the support, concern, and motivation given in carrying out this dissertation.

(This page was intentionally left blank)

Abstract

Home holiday rentals is a growing industry in the city of Lisbon, following the increase of the tourism volume, yet very little explored and with a great lack of studies and information about it. This leads to a need for research, in order to find new patterns in the guest's and host's behavior that allow the property's owners to maximize their profits, by increasing their monthly occupancy rates. This study was made using several Data Mining techniques.

This dissertation used the data from the property management company FeelsLikeHome (FLH), namely the properties information and reservations' historical data from January 2017 to May 2019.

The relationship between the monthly average rate per night and the monthly occupancy rate was studied, to understand if they affect or explain one another. After this, there was a need to understand which are the variables that better explain and predict the occupancy rate. Finally, with this information, a set of matrices were built based on the most important predictors, displaying the corresponding occupancy rate, with the objective of proposing changes to the rates per nigh currently implemented.

A predictive model was obtained for the occupancy rate, through the interpretation of patterns in the properties' occupancy. With this, properties' profiles with high and low occupancy were identified and coefficients of rate change were proposed. These models offered useful knowledge for FLH and for the industry professionals, since it allowed them to develop marketing strategies to improve profits.

Keywords: Data Mining, FeelsLikeHome, Occupancy rate, Rate per night, CRISP-DM, Predictors.

(This page was intentionally left blank)

Resumo

O aluguer de casas de férias é uma indústria em crescimento na cidade de Lisboa, que tem acompanhado o aumento do volume do turismo, no entanto, é ainda muito pouco explorada, verificando-se uma grande falta de estudos e informação sobre esta. Isto leva a uma necessidade de investigação, a fim de encontrar padrões no comportamento dos hóspedes e dos anfitriões que permitam ao dono da propriedade maximizar os seus lucros, aumentando a taxa de ocupação mensal da propriedade. Este estudo foi feito com recurso a diversas técnicas de Data Mining.

Esta dissertação utilizou dados da empresa de gestão de propriedades FeelsLikeHome (FLH), nomeadamente informação das propriedades e dados históricos das reservas de Janeiro de 2017 a Maio de 2019.

A relação entre o preço por noite médio mensal e a ocupação média mensal foi estudada, a fim de entender se eles se afetam ou explicam mutuamente. Depois disto, houve uma necessidade de entender quais as variáveis que melhor explicam e preveem a taxa de ocupação. Finalmente, com esta informação, um conjunto de matrizes foi construído com base nos preditores mais importantes, exibindo a taxa de ocupação correspondente, com o objetivo de propor alterações aos preços por noite praticados atualmente.

Foi obtido um modelo preditivo para a taxa de ocupação, através da interpretação de padrões na taxa de ocupação das propriedades. Com isto, foram identificados perfis de propriedades com predições de taxas de ocupação altas e baixas e foram propostos coeficientes de alteração do preço. Estes modelos oferecem conhecimento útil para a FLH e para os profissionais da indústria, uma vez que lhes permite desenvolver estratégias de marketing para aumentar os seus lucros.

Keywords: Data Mining, FeelsLikeHome, Taxa de ocupação, Preço por noite, CRISP-DM, Preditores.

(This page was intentionally left blank)

Index

Index of Tables

Index of Figures

List of abreviations

UNTWO	World Tourism Organization
CRISP-DM	Cross Industry Standard Process for Data Mining
IBM	International Business Machines
RM	Revenue Management
ADR	Average Daily Rate
ORS	Online Reservation System
INE	Instituto Nacional de Estatística
OLS	Ordinary Least Squares
NCR	National Cash Register
DT	Decision Tree
CART	Classification and Regression Tree
ANN	Artificial Neural Network
MAE	Mean Absolute Error
RMSE	Root-Mean-Square Error
MLR	Multiple Linear Regression
R^2	Coefficient of Determination
FLH	FeelsLikeHome

(This page was intentionally left blank)

1. Introduction

1.1. Theoretical background and motivation

In recent years, a big growth in the tourism industry, economically and socially, has been verified in all the major cities around the world. This industry is one of the industries with bigger economical potential, and it has opened several doors regarding new business opportunities (Khan, 2014; UNWTO, 2016).

With more and more tourists arriving in their cities, local communities and second homes' owners looked at this as an investment opportunity and realized that by renting a room in their houses for a short-term period, or even renting their second homes, they could make an extra fee (Mohlmann, 2015). Tourists responded very positively to this new emerging market segment, since this would offer them a wider variety of prices available, and it would provide them a real local experience, living among local communities (Guttentag, 2013; Saló & Garriga, 2011). Since tourism is a rising industry, the holiday accommodation rentals sector still has much to grow, and it is predicted to achieve numbers that could compete with bigger accommodation segments, like hotels (Guttentag, 2013).

At first, one of the biggest difficulties that accommodation's owners had, was in making their renting offers known by potentially interested tourists, due to lack of distribution channels for this kind of service. This led to the emergence of a series of platforms, as a way of resource distribution and communication (Mohlmann, 2015; Wang & Nicolau, 2017). Nowadays, the biggest online platform for people to list and book unique accommodations around the world is Airbnb, operating in 192 countries and with more than 500,000 listings by 2017 (Wang & Nicolau, 2017).

The growth of Airbnb facilitated the listing of rental opportunities for hosts, although, it did not manage those listings. The revenue management process of each listing still had to be defined by its rental manager (the host itself or an intermediary company managing the accommodations), more especially, the rate management component, regarding the delineation of the most suitable rental price. At first, rental managers looked at hotel's pricing systems to find the best rate for their rentals and increase their profits to the fullest, but fast it was realized that the indicators that affect price in the hospitality industry are unfit for the holiday accommodation rentals industry, and so, the price was not optimized, and rentals managers

were not taking the most advantage possible of their rentals (Guttentag, 2013; Wang & Nicolau, 2017).

Since there aren't many studies or models for rate optimization processes in holiday accommodation rentals, it is imperative that price determinants with the most relevance are identified in accommodation rentals' data, in order to create an efficient model, capable of, basing on those determinants, predicting the best price to apply to each listing.

This study is motivated by the need that hosts of holiday accommodations and accommodation's management companies, in this case, specifically FeelsLikeHome, have in optimizing their rate management processes, like defining the best rates for each rental listing, with the main objective of making the most profit possible of any rental opportunity.

FeelsLikeHome is a property management company, founded in 2012, who offers full-service management focused on the tourism market, more specifically, the short-term renting segment. The company provides its customers with investment opportunities, by fully managing the rental of a room or a full house/apartment, without the owners having to be concerned with their property management (FeelsLikeHome, 2018).

In this context, with large samples of accommodation rentals' data, data mining techniques are recommended, in order to discover valuable knowledge, as patterns concerning possible correlations between price and a list of explanatory variables (Rygielski, Wang, & Yen, 2002).

1.2. Objectives

The main objective of this study is to identify which are the indicators that have a bigger impact on short term rental's rates, in order to build a set of matrices that can suggest changes to the currently implement rates, in order to improve them and make them more profitable. The list of possible explanatory variables is built based on a literature review of rate management processes in other industries, and in several studies regarding the price determinants in holiday accommodation rentals. More detailed, the objectives are:

1. Characterize the rental rate per night of short term second home rentals, the occupancy rate, and study their relation and how they affect one another. This will help with finding out, at first, if and how the occupancy rate affects and is explanatory of the rate established by rental managers in their listings. We consider that there is a relation if the correlation coefficient if over 0.3.

2. Identify the explanatory variables for the occupancy rate, *i.e.*, estimate the occupancy rate regarding all the possible predictors, in order to identify the ones that have a bigger impact on the occupancy rate. This will tell us which characteristics of the house, location, external environment and reservations that have a bigger influence on the property's occupancy rate. Doing this, it is possible to identify properties´ profiles, that have a higher occupancy rate or a lower one. This will help the company make a decision about in which houses to raise or lower the rate per night because the property is very crowded or with few guests.
3. Create a set of matrices to propose changes to the currently implemented rate, based on the most important predictors for the occupancy rate, displaying the corresponding occupancy rate.

Following these objectives, we will be capable of finding out if there is any relation between the occupancy rate of holiday accommodation for rental and its rental price. Then, from the list of possible explanatory variables created based in the literature review, we will identify the factors or characteristics that have a bigger impact on accommodation's occupancy rate, reflecting what customers value or depreciate the most. Finally, with all this information, we will be able to build several profiles using the most relevant variables, which will be linked with higher or lower occupancy rates. Finally, some measures for those profiles will be proposed, in order to maximize occupancy rates and, in the last instance, profits.

1.3. Methodology

This study regards the case study of the FeelsLikeHome company, and it treats and analyzes historical booking data and price registers of about 400 accommodations, during a period between 2017 and 2019, in the FeelsLikeHome platform. These data, provided by the property management company, do not consider canceled and non-canceled booking data, but on the other hand includes information about the accommodations characteristics, arrival, and departure date, as well as other reservation' information, guests nationality and the rates per night implemented during the period in question.

In order to deal with big amounts of data, this study used CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which outlined a process model that allows users to perform data mining projects, with an appropriated framework to perform those projects. This

methodology is composed of six phases, however, in this work only the first five are taken into account: business understanding, data understanding, data preparation, modeling and evaluation (IBM, 2018).

1.4. Document structure

This work is divided into five chapters. In chapter two the literature review is presented, namely general concepts of Revenue and Rate management, as its growing importance in several different industries. The evolution of rate management in the hospitality and airline industries are explored, since there are several similarities between these industries and the short-term home holiday rental industry, regarding the pricing system of its products or services. Still in the literature review, the growing of the short-term rental industry is an important subchapter.

Chapter three describes the methodology used, in this case, the CRISP-DM methodology, as its 5 phases. In chapter four there is the demonstration of the results for the defined objectives, as its discussion in the business context. Finally, in chapter five there is the conclusion, contributions, and limitations of the work.

2. Literature review

This chapter exposes and summarizes all the relevant information, theory, and studies, about the rate management, a small, yet important component of a company´s revenue management, and the rising of the short-term home rentals industry in Lisbon. First, the concept and importance of the rate management process are analyzed and described, then his application and different models in the airline industry and hospitality industry are presented since those are the two industries where the rate management has evolved more markedly. The market of the short-term home holiday rentals is construed and his impact on local neighborhoods is analyzed, such as the emergence of the Airbnb platform. Finally, a relation is established between an optimized and improved rate management process and this growing market, and how the second can benefit from the first.
In order to perform an organized and accurate research, an ad-hoc protocol was used to select the best articles to be analyzed.

Ad-hoc protocol

List of selected sources of review (platforms): IEEE Xplore, ACM Digital Library, Science Direct, IBM research, Google Scholar.

Keywords used: [(“Rate Management” OR “Revenue Management” OR “Yeld Management” OR Pricing OR rates OR fares)] AND [(“Second home rental” OR “Short-term rental” OR “Local accommodation” OR “Hospitality industry” OR “Airline industry” OR “tourism”) AND NOT (“Food industry” OR Stores)] AND [(“Data Mining” OR “Statistical analysis”)] IN (Title) OR (Abstract).

Review procedures: Inclusions criteria (IC): Studies that include large samples; Computational based rate management models; Studies that can be related to the second home rental market segment.
Exclusion criteria (EC): Studies that are not in English language; Poster or tutorials; Studies in areas not related to rate management or pricing; Studies that are not fully available or unfinished.
Period: 2008- 2018

2.1. From Revenue to Rate management: concept and importance

Revenue Management (RM) represents the practice of gaining the highest possible revenue in the selling of a company´s service product, as this management technique is more usual in service firms (Ng, 2007), it is also recognized as a management principle that fits into multiple business areas including marketing, strategy and consumer behavior (Cross, Higbie, & Cross, 2009; Guillet & Mohammed, 2015; Ivanov, 2014). Ng (2007) adds that practitioners of this discipline use different tools, as targeted pricing, market segmentation, and demand forecasting, in order to sell the limited capacity of the firm at the highest rate possible.

It is not too much to say that over the past decades, well-implemented pricing and revenue management techniques have added tens of billions of dollars to the net profits of hundreds of firms (Cross, Higbie, & Cross, 2011) , and so, it is easier to understand why it gained so great importance in companies´ way of operating, and why is considered by many firms as an indispensable part of their marketing and operating strategies (Cross et al., 2011). Other evidence of the impact and the growing importance of revenue management was the launch of two academic journals on the subject: *Journal of Revenue and Pricing Management* and the *International Journal of Revenue Management*.

Companies´ revenue management strategies started in the airline industry between 1970 and 1980, as Yield Management, where the core ideologies were aimed at maximizing yield per available seat, combining the plane's capacity with the best-fixed seat fare possible. In the late 1980s, the hospitality sector adapted these strategies from the airline industry, in order to be able to offer different prices and time-sensitive products to their customers, diversifying market segments, and thereby increase hotel profits (Cross et al., 2009). As these strategies were adapted into the hospitality industry and other service industries, it became more appropriate to redefine yield as revenue, reflecting a more strategic concept (Cross et al., 2011; Ivanov, 2014).

In the early 2000's, the discipline faced a few challenges, not being able to prevent the demand decrease, and some changes had to be made. With access to extensive databases and the ability to analyze customer behavior, firms recognized that they could do far more than simply manage inventory (Cross et al., 2009), as they did in the beginning, when the scope of revenue management was limited to capacity planning and allocation for a given set of fixed prices (Ng,

2007). Instead of inventory-focused, a new role was emerging that encompassed marketing, sales and channel strategy, customer-focused, as the scope of revenue management started to include pricing and demand behavior in the optimization process (Ng, 2007). This kind of thinking leads firms to the fundamental issues of pricing and customer value (Cross et al., 2009). This phenomenon happened mostly in the hospitality and tourism sectors, as a customer-centered approach is more and more usual, with prospects of integrating customer relationship management into the subject (Guillet & Mohammed, 2015).

We need to bear in mind that even though there are many general revenue management principals applied across different industries, each industry has specific characteristics that need specific principals applied to the companies that work in it (Ivanov, 2014). Rate management systems are more and more a common practice among several industries, with certain specifications, strategies, and objectives according to the market and business requirements.

The concept of rate management is directly related to revenue management, being the component of the RM, which defines pricing strategies for the products or services, in order to maximize revenues. By analyzing previous hotel RM studies, we can conclude that more and more, pricing analysis and optimized techniques have been recognized as the main strategic tools that hotel organizations use to achieve their objectives (Guillet & Mohammed, 2015), and in his studies, Cross et al (2009) state that the emergence of rate management process started when the revenue managers asked themselves if the fixed rates they were stablishing for their products/services, were the right ones to begin with. As revenue management began focusing on strategy and pricing, rate management systems became more popular, and leading firms realized that new pricing analytics were required to support this decision process.

Automated rate management systems are a solution to many companies, saving them a lot of time analyzing customer data. These emerging systems are able to set rates for each property (in the hospitality sector) or arrival date (airline sector) based on several factors, like demand forecast, the elasticity of demand in the correspondent market segmentation and competitive rates. The system processes these data and recommends optimal rates, to maximize revenue (Cross et al., 2009). However, these systems are still in a very preliminary state, since they can´t consider or predict many issues that can alter demand or any market-related aspect.

Over the years, there has been much skepticism among company executives regarding rate management practices, however, there is less and less of it as more companies successfully

increase their profits by adopting systematic and scientific approaches in this area (Lieberman, 2010). Pricing models determined by the companies variate according to the industry in which they are included. The integration of pricing and revenue management with the company´s supply chain presents a great opportunity to expand this science (Cross et al., 2011).

Pricing strategies are defined as the decisions that a company has to make on how much to charge for a product/service, and its study involves, among other factors, the value these products have for the buyers, the quantity the company wants to sell, organizational implications, the competitors´ prices and how to achieves the objectives of the firm (Ng, 2008). According to Cross et al. (2009), while the concept of price optimization is still in his embryonic stage, many believe it will be the next big step in revenue management, since it brings together an understanding of customer buying habits and market dynamics to predict what each customer segment is willing to pay for a certain product/service in a wide variety of circumstances

Those pricing strategies need to be very well planned, with the ability to adapt according to several factors, related to the target customer profile, market-related changes or any contextual inconvenience, as it has been recognized that overcapacity will lead to lower prices and uncontrolled discounting, while raising prices in an environment of constrained capacity may be harmful to the company, and encourage competition (Cross et al., 2011), and so, pricing is a very delicate process, which can determine a company´s success. Customer´s behavior, preferences, valuations and competitor actions are some factors that can change over time, and so, a well-designed pricing system should be able to instantly react to this, allowing the firm to be prepared and to adapt to these changes (Lieberman, 2010).

Cross et al. (2011) affirm that the core of this discipline is in the process of understanding and analyzing the value that customer gives to each product, and align product prices, placement and availability with each customer segment. Besides the great number of studies on pricing strategies and methods, many focused on other aspects of pricing, like identifying which ones are the characteristics of product/service attributes that consumers are willing to pay extra for, and which have no effect on price (Guillet & Mohammed, 2015).

Nowadays, there are several different pricing tools, used in different context, by different companies, like Ivanov (2014) presents in his work: price discrimination, which means different values are charged to the customers for the same product/service, according to their price sensitivity; lower price guarantee is also a common tool, used mostly by the hospitality

sector, and it means that if the customer finds a lower price for the same or similar hotel, they match that price; price framing refers to the way a price is presented to the customers, for example, a low price can be presented directly as a low price, or as a discount from a higher price, forming a different expectation about the value of the product/service; dynamic pricing, where the price reflects the current level of demand for the product/service, and adjust it according to changes in demand rate, or is set according to the date when the purchase is made.

In the service industries, the pricing strategies are mostly dynamic, segmenting their own market in different parts, by the length of the lead time to the end of the service product´s lifetime. For example, when a reservation is made very close to the target date, for a hotel room, a flying seat or other service product, the rates will be higher than for reservations made a long time before the target date. Because of this rate policy, most service products are sold in advance, incurring in an opportunity loss for the service providers, because the profit margin is higher if the purchasing date is closer to the date (Guo, Ling, Yang, Li, & Liang, 2013), as the firms realize this, they are starting to set prices based on forecasted patterns of demand so that the capacity sold to in advance would not deny the firm of obtaining more profits from customers arriving late (Ng, 2007).

So, although this is a very common strategy, the date when the reservation is made shouldn't be the only factor to consider if we want an optimized rate management system. According to Cross et al. (2009), demand is also important, when it is low, rate management systems (or yield management systems) recommend opening the lowest rate programs, and although many in the industry claim that there are times when low rates neither stimulate nor capture additional demand, no analytics existed to support those theories.

Although the main objective of a company´s rate management system is to define the best pricing strategy, in order to increase their profits, this is not consensual, and a lot of other objectives of pricing are proposed, varying according to the company´s objectives (in long or short-term), philosophy and strategy. Those objectives can be market share maximization, price differentiation, service quality leadership, creation of prestige image for the company, among many others (Ng, 2008).

2.1.1. Rate management in the hospitality industry

Revenue management in the hospitality industry was adapted from the airline industry (Cross et al., 2009), and in the past several decades it began to focus much more on a strategic view

on revenue management systems, moving away from a capacity centered management view, to a profit maximization based approach (Altin & Schwartz, 2017). Because of the industry's high competition, some hotel firms realized that, if they wanted to gain some advantage, offering their customers the best price possible was a good and achievable solution (Bojanic, 1996). Ivanov (2014) proposed revenue management in the hospitality industry as a system, as we can see illustrated in Figure 1, defined as a series of structural, procedural and human resource elements, committed to achieving the hotel firm's objectives. The system consists of four structural elements (Data and information, Hotel revenue centers, RM software, and RM tools), RM process and RM team. The component regarding rate management and pricing is included in the RM tools, once the pricing systems can be seen as a tool to be used by hotel firms to increase revenues.

Marriot International, one of the pioneers in hotel revenue management, added between 150 and 200 million dollars to its top line after the adoption of these techniques. This was seen as an amazing discovery, and by the late eighties, a lot of other hotels followed the same path. This new way of management allowed them to offer a broader range of fares to their guests, and to allocate how many rooms should be priced at different fares (Cross et al., 2009).

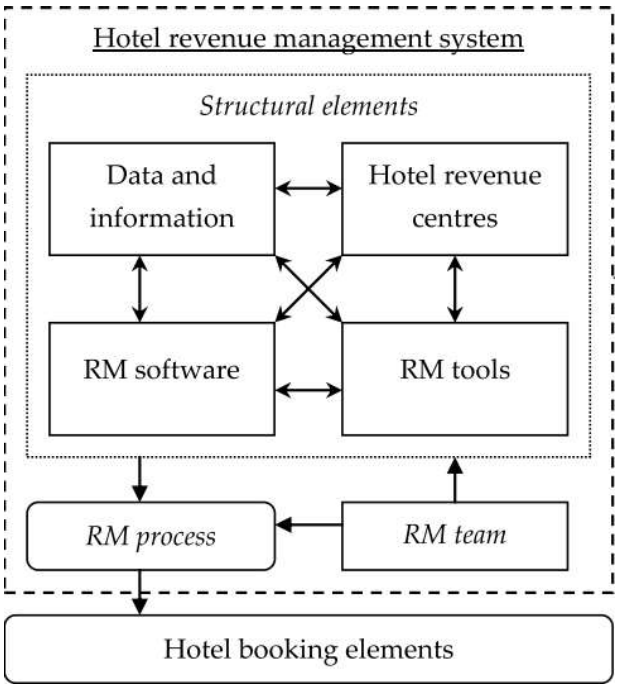


Figure 1: Hotel revenue management system

Source: Ivanov & Zhechev (2012)

Ivanov (2014) adopts a seven-stage revenue management process, as illustrated in Figure 2. As we can observe, the pricing systems and pricing decisions are included in the "decision" stage. This stage oversees the development of models to recommend optimal levels of pricing, rate structures, overbookings and help to make proper decisions. These decisions are ultimately made by the revenue managers and could relate to opening or closing particular price levels to all or specific market segments, opening or closing dates for sale to all or specific market segments, revising price levels, among others. Modica, Landis, and Pavan (2009) complement that historical demand data, as historical booking records, help on predicting future demand, and setting the right price to the right room, encouraging or discouraging demand.

Although it focuses much on generating revenues by reducing prices, revenue management is not simply a discounting method to achieve more profits, but a scientific and well-thought definition of price discrimination for the proper customer segments. Some look at historical demand data in order to predict future demand and customer's arrival and departure, and with this information, set the right rates for each room (Modica et al., 2009).

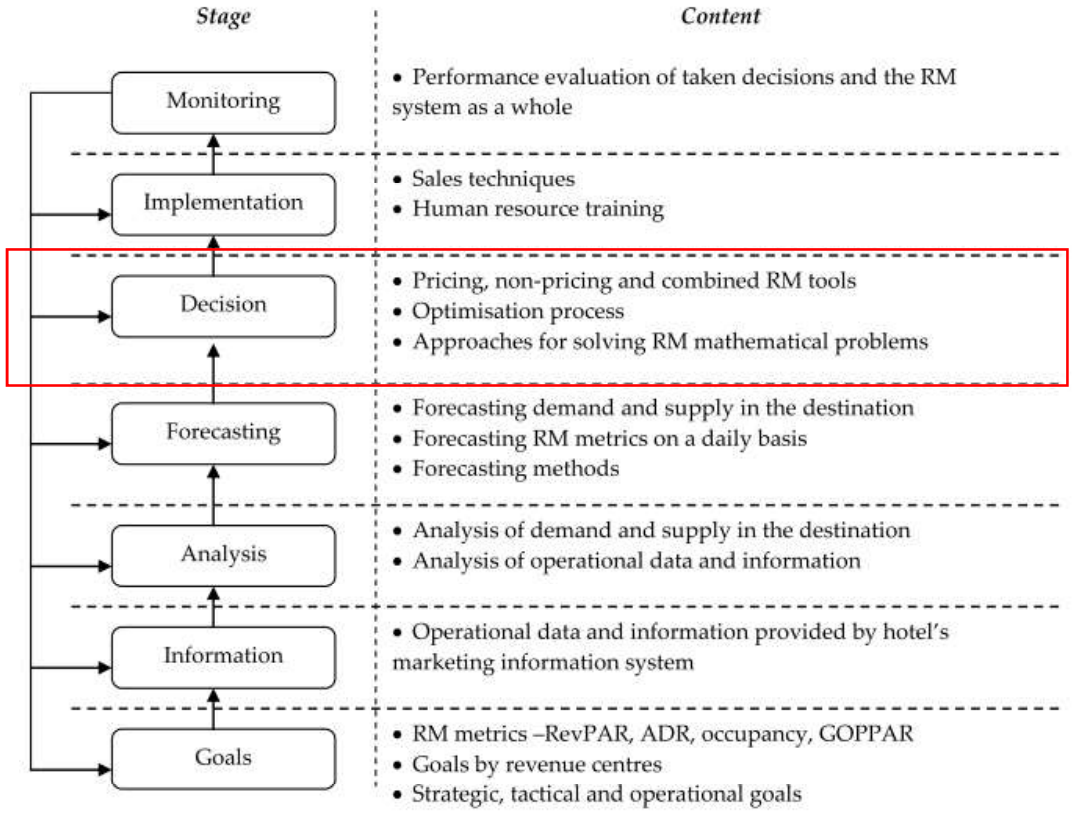


Figure 2: Hotel revenue management process

Source: adapted from Ivanov & Zhechev (2012)

Price is one of the most important instruments of rate management, since it originates profit for the hotel firm. It can be changed over time, through different customer segments, according to specific conditions and factors, being this called dynamic pricing. This may attract the right customers and generate higher revenues. Pricing tools include a group of techniques that influence hotel's prices, including price discrimination, dynamic pricing, lowest price guarantee, among other techniques (Ivanov, 2014).

Pricing strategy, or pricing system, and price optimization are components of a typical hotel's rate management system, included in the revenue management system. The first one implicates setting up price rates in order to determine the minimum price to assign for different market segments and variate these prices over time to maximize revenues. The second one is directly related to inventory, as it includes the sequence of decisions to ensure that the right inventory is sold at the right price, at the right time, to the right market segment based on consumers' willingness to pay (Guillet & Mohammed, 2015). Through databases available, hotel firms focused on analyzing and even manipulating customer behavior, and so, with well-established pricing strategies, revenue management started to create demand, instead of only managing it (Cross et al., 2009).

Ivanov (2014) describes the average daily rate (ADR) as one of the metrics that show the efficiency and performance of the hotel's revenue management, as it relates to the average price charged by the hotel for one night. It reflects the hotel's ability to generate profit from higher prices instead of more overnights. According to the author, this average price can be influenced by the day of the week when the reservation is made, the period of the year (tourism season and lower season), market segments, special events (for example web summit), type of rooms booked by customer and contract conditions with distribution channels. Webb & Schwartz (2017) reinforce the importance of this metric, adding that its principal objective, is to measure the efficiency of the short-term decisions, in this case, decisions regarding room prices.

Initially, hotels' pricing strategies were based on analyzing a series of factor, like historical and future reservations, and booking patterns by market segment (Cross et al., 2009), and more recently pricing techniques changed from demand-driven to value-based, capitalizing the opportunities offered by data analysis (Altin & Schwartz, 2017). On the other hand, Modica et al. (2009) give great importance to demand-based pricing, where different prices are charged to different customers depending on the current level of demand. Cross et al. (2009) add that

the importance assigned to price results in studying the things that customers value the most and segmenting the market according to those things. Ivanov (2014) quoting El Haddad, Roper & Jones (2008), states that dividing customers into different segments is essential for matching supply and demand, and therefore maximize profits.

The pricing problem is very complex, and each decision involving price must consider a lot of factors, like occupancy rates, customer demand, competition´s price positioning and cost of structure. Plus, each of these may alter by many other conditions, like the day of the week, season, weather and big events nearby (Cross et al., 2009; Modica et al., 2009). In some hotel firms, the price attributed to each room may change according to several factors, as type of room, food board, room view, period of stay, date of booking, booking terms (cancellations, payment terms), distribution channel, guest's loyalty, group size, etc., all kinds of criteria that justify the application of different prices (Ivanov, 2014), as evidenced in the Table 1.

Among the large amount of literature in hotel's price levels and pricing techniques, it was concluded that there are several popular factors normally considered when making decisions regarding price definition, being those factors: hotel's star category, online rating, hotel location, distance to the beach and city center, booking date, the characteristics of the reservation, the customer's profile, season of the year, room facilities, type of hotel, competition nearby, temperature and economical and environmental context (Vives, Jacob, & Payeras, 2018). On the other hand, Wang & Nicolau (2017) divide price determinants into five categories: location characteristics, hotel's star category, hotel's services, and comforts and external competition factors.

The customer's willingness to pay, or perceived price, must be studied in order to find the right rate for the right room, at the right moment. Customers take into account things like time, search costs and convenience in their determinations of the perceived price (Bojanic, 1996). Customer demand does not have a uniform price elasticity, or willingness to pay, it changes by segment and market conditions, and so, hotel firms charge different market segment different prices, so that customers with higher willingness to pay, don't buy at lower rates (Ivanov, 2014).

Table 1: Factor influencing pricing decisions in the hospitality industry

Factor	Impact on pricing
Category	Higher prices for higher category properties
Quality / value	Higher prices for hotels delivering higher value to their customers
Image	Positive image leads to higher prices than competing hotels
Product lifecycle stage	Lower prices during introduction and decline stages, higher during maturity
Additional services included in the price	More included services lead to price escalation
Location	Hotels closer to tourist resources boast higher prices
Competition	Serves as a benchmark
Sales volume	Lower prices (discounts) for guests (distributors) booking more rooms
Demand	Higher prices during periods of high demand (e.g. special events)
Demand elasticity	Lower prices for price sensitive customers, higher prices for less price sensitive customers
Affiliation to a hotel chain	Payment of franchise/management fees increases costs and, thus, prices
Bargaining power of distributors	High distributor bargaining power leads to lower prices
Company's marketing strategy and goals	Focused differentiation strategy is related to high prices, while market penetration to low prices
Organisational structure	Determines who has the responsibility for pricing within the company
Taxation	Positive relationship with prices
Government regulations	Setting price ceilings (maximum prices) or price floors (minimum prices)
Costs	From an accounting perspective costs are a pricing factor. From a marketing perspective customers are interested in their costs (price, time, social costs, etc.) and the value they receive, not the costs of the company.

Source: adapted from Ivanov (2014)

Many theories define the relationship between price and demand as linear, if one decreases, the other one increases, and vice versa (Modica et al., 2009). Variations in tourism demand influence hotel firms to look for solutions to manage demand, and attract more customers during slow periods (Ivanov, 2014). With low customer demand, hotel firms tend to leave all discount rates available, hoping for an increase in demand and revenue (Bojanic, 1996; Cross et al., 2009; Modica et al., 2009). This cause some concern among the industry, because this practice usually leads to a competition for the lowest price, and sometimes, these low rates don't even have any effect on demand. This is considered an unconscious use of pricing systems (Cross et al., 2009).

Another technique used by hotel firms to attract customers is allowing them to book in advance, weeks or months before the arrival date. This allows customers to be guaranteed that they will have their room booked a certain time in the future, while until their arrival date hotel firms can profit from their room. In some hotels, there is the lowest price guarantee offered by the

hotel, and if a customer finds a lower price for the same hotel or similar, that price is matched, to safeguard the customer's stay (Ivanov, 2014).

Dynamic pricing strategies in the hospitality industry has been increasingly adopted by many hotel firms, as they typically segment their market of service products by how much time before are the reservations made. A higher rate for customers booking their rooms one day before the arrival date, and a lower rate for those who book in advance (Guo et al., 2013; Modica et al., 2009). It allows a hotel to increase their profits by offering a price reflecting the current level of demand and occupancy, and adapting it according to changes in any of these factors (Ivanov, 2014; Modica et al., 2009). Customer´s willingness to pay varies as the arrival date approaches, as normally customers booking one or two days before the arrival date are less price-sensitive, and so, are willing to pay a higher price (Cross et al., 2009; Modica et al., 2009).

In most recent times, with the advance of technologies, many hotels integrated informatics systems into their managing processes in order to optimize and simplify making decisions. Information and Communication Technologies were a big improvement in price optimization processes, since, through the analysis of big volumes of customer's data, the customer-oriented pricing processes were facilitated and encouraged (Vives et al., 2018). Forecasting prediction systems are gaining great importance too and are increasingly being developed with the prospect of helping optimize operations. Integrating Big Data analysis with this forecasting systems, Bing Pan and Yang Yang (2017) propose a forecasting model of weekly hotel occupancy for a destination, using several tourism data sources, including the hotel's website traffic and weekly weather traffic. This can be very helpful, since it facilitates the demand-based pricing decision. Big data from Customer Relationship Management systems have also been very helpful, since it allows the hotel firms to outline their customer's profile and create new market segments according to the different types of customers that stay in the hotel, first-timers or repeaters. This is still in an early stage, but more and more this systems have been developed, to facilitate pricing decisions to different market segments (Talón-Ballestero, González-Serrano, Soguero-Ruiz, Muñoz-Romero, & Rojo-Álvarez, 2018).

Cross et al. (2009) describe the traditional optimization system for hospitality, based on allocating a certain number of rooms to be sold at each rate (from a predefined and optimal set of rates) in order to maximize revenue. If the rates at which the rooms are sold match the customers' perception of value for the rooms, the profits will be optimized. Emerging price optimization systems, while still in its embryonic stage, are able to find the best rates for each

room and arrival date based on demand forecast, its elasticity in the market segment, and competitive rates, predicting what the customer is willing to pay in a diversity of situations (Cross et al., 2009).

In their work, Vives et al. (2018) designed a price optimization scheme for the hospitality sector, relying this process in two imperative sub-activities, price differentiation, and customer segmentation. The scheme, which we can see in Figure 3, is built demonstrating the sources of price variability and demand segmentation in the hospitality sector.

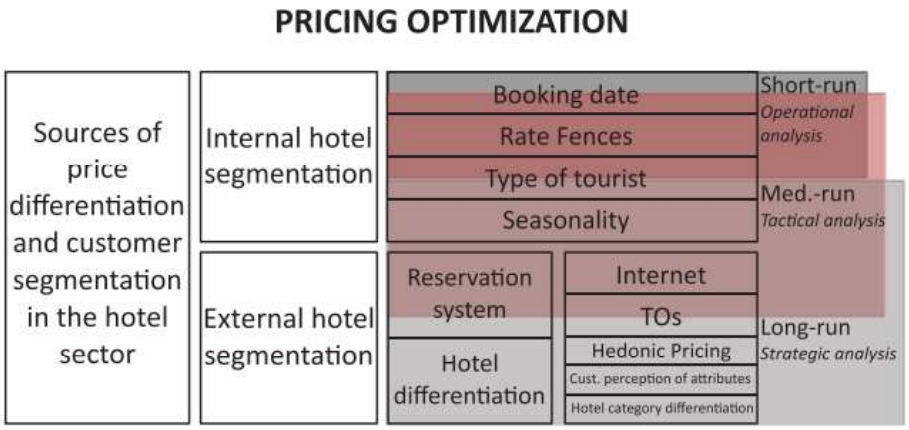


Figure 3. Sources of price variability and demand segmentation in the hospitality industry.

The Online Reservation System (ORS) is widely used in the hospitality sector, as it allows customers to book hotel rooms at any time. This informatic tool also offers different segments, with different prices for the rooms, according to the date of booking, evidencing a dynamic pricing strategy (Guo et al., 2013; Ivanov, 2014). Decisions of revenue management also can be automated, when a well-developed RM software is integrated in real-time with the OSM, it is possible for the software to set different rates to different customers during different search times, depending on the reservations already made for that date and how often the accommodation for that date has been searched via website. However, the lack of the human factor could make the software set a higher price than necessary, and dissuade customers from booking the room (Ivanov, 2014).

In their work, Guo et al. (2013) propose a pricing model in using of OSM, considering different demand seasons (regarding weather, global events, and different cultures) and different hotel capacities, and based on market segmentation, with a static demand function, as well as a finite

capacity of hotel rooms. The applicability of the model is verified by a series of algorithms, through different demand values, using the hotel´s segmentation and corresponding rate for each segment. The results suggest that market segmentation is beneficial for the profits of the hotel firms.

2.1.2. Lessons learned from other industries: Airline industry

The airline industry was the first one to have revenue management systems operated by their companies, and consequently rate management, through which they defined their pricing strategies. Nowadays, this industry is considered to have one of the most complex pricing systems in the world (Escobari & Gan, 2007). It all started in the seventies, as a desperate strategy for airlines in economic difficulties (Cross et al., 2009, 2011), when the airline industry was decontrolled and the interest in the topic began to increase. Many airlines reported increases in revenue of 5% or more after starting revenue management systems (Ng, 2007).

Revenue management in the airline industry has been founded on revenues of the airline seats, grouped into different booking classed, each with a specific fare and restrictions (Alderighi, Nicolini, & Piga, 2015), and many researchers assume the time of purchase as the main aspect for discriminating revenues, whether through prices or capacity allocation (Alderighi et al., 2015; Eneckere & Peck, 2012; Ng, 2007), evidencing what was done by the airline industry in the early days of revenue management applications, when they offered discounted seats, with advanced booking requirements, and with a limited number of seats that could be sold in discount, limiting their availability to those who could plan in advance (Cross et al., 2009).

Cross et al (2011) describe the first real application on rate management with the objective of incrementing profits. The American Airlines introduced the “Super Saver Fares”, very low prices for their flight seats, restricted to capacity and advanced purchase. Rapidly they realized that they needed to consider a lot more external factors if they wanted to have an optimized system.

As in most industries, revenue management can be decomposed in several components: demand forecasting, overbooking, capacity allocation, and finally pricing (Marcotte & Savard, 2003). This last one is involved in the rate management system. Pricing policies can be used to produce a wide variety of fares according to different booking classes, among other flight perks, on the same flight (Alderighi et al., 2015), allowing the airline company to optimize their profits. They need to solve two issues in order to do that, first, decide how many seats to

make available for the demand season, then, decide how to price them in order to adjust fares over time as a response to strong demand (Eneckere & Peck, 2012).

It is recognized that in this industry, pricing has received low attention, comparing to the other components of revenue management, mostly because the several difficulties that companies find in the implementation of a practical and optimized decision support system (Marcotte & Savard, 2003), that´s why there is not much information available on the topic.

In her work, Ng (2007), quoting Cary (2004), evidence that in the US airline industry, pricing strategies and revenue management function very differently, being separated components, as opposed to what Marcotte & Savard (2003) said, about one being another's component, both of them in need of great adjustment. Pricing is almost entirely motivated by the actions and reactions of the competitors, and revenue management is motivated by the patterns of historical demand data (Ng, 2007).

Marcotte & Savard (2003) affirm that the application of a well-established pricing system in real-life situations, requires the analysis of a great amount of data, like the fares practiced by the competitors, demand forecast, historical sales pattern, among other factors. Cross et al (2009), based on historically facts, and airline companies' practices, point customer demand patterns tracking as the main statistic to consider on this matter, and quoting Talluri and van Ryzin (2004), conclude that the essence of airline revenue management reside in forecasting demand at the fare class level and then opening and closing, increasing and decreasing, until the airplane is filled with the most profitable composition of passengers.

Alderighi et al. (2015), through the studies of Dana (1999a) and Deneckere & Peck (2012), and the analysis of data from the airline company RyanAir, show some evidence of capacity-based theories of pricing, and reinforced an existing relationship between a flight´s occupancy rate and the evolution of fares, or price dispersion. Escobari & Gan (2007) also verified that price fares are on average higher in fully occupied flights, emphasizing these theories. On the other hand, there are some time-based theories on pricing, where the airlines use price discrimination depending on the purchase date, to exploit customer´s variation on terms of willingness to pay and uncertainty about departure time, with prices typically increasing over time (Möller & Watanabe, 2010).

In their work, Marcotte & Savard (2003) propose a solution methodology for the pricing optimization issue. They developed a bi-level mathematical model driven by decision variables that fall into two different categories. The first category consists of quantities describing the

purchase behavior of each client group, considering criteria like the nominal value of an airline ticket, flight duration, quality of service, customer fidelity to a given airline. The second category is associated with commercial targets set by the airline, such as revenue or market targets. This methodology was tested in an American airline, although was well planned, it still has some downsides.

Alderighi et al. (2015), quoting Dana (1999a), describe a basic model for fare optimization, where the correlation between fares and seat availability is evidenced, when the airplane´s capacity doesn´t change, by presuming that fares are set before demand is predicted. Price distribution result from the uncertainty of demand, and from the low probability of selling an extra seat when the flight occupancy increases. Lowest fares are attributed when there is a high probability of seats being sold, while the highest fares are attributed when there is a low probability of selling an extra seat (Escobari & Gan, 2007). However, a solution where prices can not be adjusted over time is ineffective, as the airline cannot amend their pricing decisions as it learns new information about demand (Eneckere & Peck, 2012).

Eneckere & Peck (2012) propose a pricing model where current prices are based on historical data about the evolution of demand in previous periods, and so, price variation is allowed. With this practice, customers may decide whether they want to purchase their tickets in advance or not, basing their decisions on their own expectations about future prices and historical data on demand.

In conclusion, in order find the best rates to apply to their flight seats, many airlines appeal to revenue and rate management process, as this is, more and more, becoming a usual thing among the airline industry (Cross et al., 2009, 2011; Ng, 2007). Although this is a relatively recent matter, many theories emerge, considering a different factor in the pricing decision process (Marcotte & Savard, 2003), as still does not exist a universal optimized process.

2.2. Short-term home holiday rentals: A growing market

Tourism is more and more becoming a major subject around the world, mostly because of its unmeasured growth, but also because of the business opportunities it provides.

In 2017, there was an increase of 84 million tourist arrivals around the world, when compared to 2016, making a total of 1,323 million tourists in the world in 2017. When focusing on the tourism accommodation activity (including hotels, tourism in rural areas, lodging tourism and

local accommodation), in July 2017 were registered 5840 establishments in operation, with 2663 of those being local accommodations. Finally, were registered around 3.4 million guests in second home rentals and 8 million overnights (INE, 2017), being this evidence of the fast growth of this segment in the tourist accommodation market.

In recent years, a new business model known as "sharing economy" has gained great importance in the tourism accommodation sector, regarding the share of assets between private individuals, for a pre-determined fee (Mohlmann, 2015), originating the second home rental segment. This phenomenon needs to be considered as an example of tourism gentrification (Gant, 2016). Wang & Nicolau (2017), quoting Zervas et al. (2016), describe the use of this business model as the connection of people who own a second home accommodation, with people who need temporary accommodation, usually, these transactions are made through a digital marketplace. This business model has grown mainly due to increased tourist demand (Guttentag, 2013; Karlsson & Dolnicar, 2016).

Second-home rentals are one of the biggest segments in the tourist accommodation market, and it increases more and more with the growing number of tourists visiting the big cities. Second homes link tourism and migration, and its importance in housing markets and tourism is being increasingly acknowledged (Müller, 2014). This phenomenon can be explained because it offers an alternative way to get accommodation for these tourists, sometimes a cheaper way. Despite the increasing importance gained by this market segment, there is not much information about management procedures of second homes, since this is one of the least studied matters in the rental market (Saló & Garriga, 2011). The growth of this industry can also be explained by the bigger variety of prices offered, so as a differentiated experience, when comparing to other accommodation options (Guttentag, 2013).

The growth of second home rentals is also related to the recent innovation in sharing economy business, since it provides a way for resource distribution and consumption (Lamberton & Rose, 2011). When applied to the second home rental sector, this business model encouraged the emergence of a series of platforms, making it possible for homeowners to rent their properties, and profit with it (Mohlmann, 2015; Wang & Nicolau, 2017).

To provide a real case of second home rentals for holiday integration in a region, Gant (2016) describes the case of Barcelona. Since the late nineties, tourism has become one of the main industries for this city, which lead investors and hotel companies to buy entire apartment buildings and transformed them into holiday renting accommodations, also lifestyle migrants

bought second homes in Barcelona so they could rent them to short term visitors while they were away, so, holiday rentals became a business opportunity for many.

This rented holiday home market has been grown specifically in Spain, in the past few years, as homeowners are availing the arrival of tourist to grab the business opportunity. Nowadays, sixteen percent of all residences are officially considered second homes in that country, and in Costa Brava, a popular tourism destination in Spain, second home rentals represent more than fifty percent of total accommodation options, followed by hotels and "bed and breakfast" (Saló & Garriga, 2011). Also, in Switzerland, the investment in second homes was significant during the nineties, revolutionizing the accommodation structure near ski resorts. Demand for private apartments to overnight has increased in the last years, while overnights in hotels have decreased. Although the market of second home rentals is huge in Switzerland, many second homeowners, instead of renting, have their families, friends or relatives living there (Bieger, Beritelli, & Weinert, 2007).

From the second house owners, we can distinguish three types of hosts. First, those who continue living in their houses, together with renters, secondly, those who are only temporarily absent from their homes, and during that time, the house is being rented, and third, owners who run permanently rental business on their second homes (Guttentag, 2013). On the other hand, there are second homes which are not being rented and are considered a market opportunity loss, they can be considered very harmful for the region when they are empty, since they affect the appearance of a touristic destination. Many owners claim that they could be motivated to rent, with the help of information, communication platforms, administrative support, and service like maintenance and cleaning (Bieger et al., 2007).

There are some factors that could discourage second homeowners to rent their properties, divided into three categories. Economical- lack of precise fee of return, because of the high taxes on returns from rent; transaction costs; no interest in return fees due to the owner's wealth. Social- fear of losing the local community's respect and friendship, due to the host's bad behavior; reluctance to hosts with different habits (smoking, house pet, etc.). Psychological- Loss of the opportunity to use the second home at any time; the fear of losing control and intimacy (Bieger et al., 2007). Local governments and residents in touristic areas report that a lot of problems may arise from short-term holiday accommodations penetrating residential areas and express concerns about the negative impact that these accommodations could have

in the neighborhood (Gant, 2016), like noise, inconvenience, traffic, parking, waste issues, or even drunken behavior when talking about big groups (Gurran & Phibbs, 2017).

On the other hand, Karlsson & Dolnicar (2016) show some reasons that could lead second homeowners to rent their homes, and divide them into three categories, as we can see in Figure 4.: Income, focusing on the rental's revenue. Social interaction, reflecting the will of the host to meet and connect to new people. Sharing, related to the unused space, and the waste of not being in the house most of the time.

The owners of second homes do not always live near the house they are renting, and even in those cases, they are not only financially, but also socially attached to the destination (i.e location of the house), interacting with the local community and contributing to the local culture. In resume, they establish a good relationship with the local community, so it is easier for their business to grow (Bieger et al., 2007), since many local residents are against holiday renting homes in their living areas, because it could lead to a substitution of residential life by tourism, and their neighborhood's identity loss (Gant, 2016).

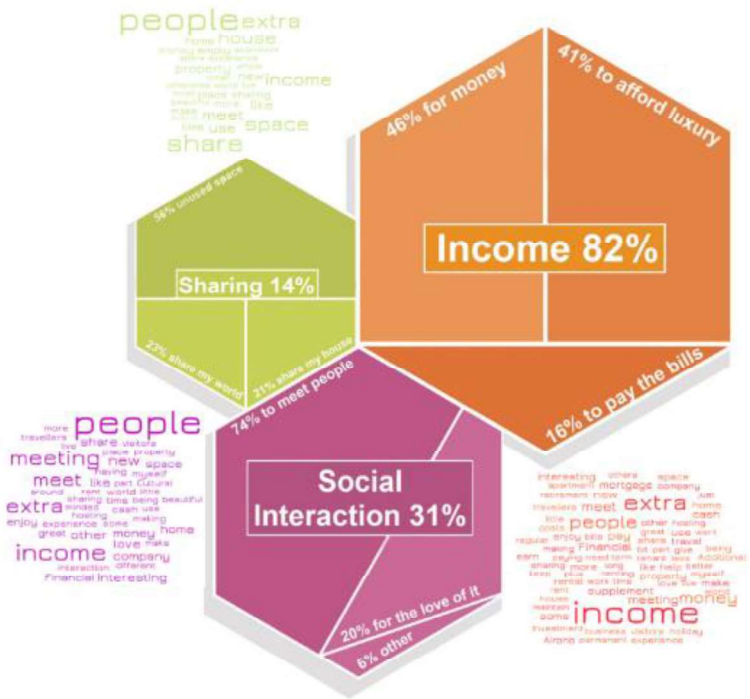


Figure 4. Host reasons for offering accommodations

Source: adapted from Karlsson & Dolnicar (2016)

Looking at the demand side, tourists point out many benefits in second home rentals, being the lower price the main advantage, but also the opportunities to cultural mix and interaction with their hosts and local communities (Balck & Cracau, 2015; Guttentag, 2013).

Hotels and second homes have their own characteristics, and so, many people prefer one or the other. Second homes accommodations offer a more varied set of prices than hotels, each one related to a different experience in different accommodation, with low prices for those who are more price-sensitive, and the opposite (Guttentag, 2013). Hotels have room service for their customers, many different indoor and outdoor sports facilities, planned entertainment activities and whether is a hotel or an aparthotel. On the other hand, second homes normally have a greater number of rooms, the surface area, the terrace area, the type of home (apartment, single-family home) and the possibility of sea view. When comparing second home rentals to hotel firms, one difference that comes up is regarding price seasonality. After analyzing a series of price records in hotels and second homes, in the region of Costa Brava, Spain, it was concluded that hotels are more seasonal than second homes, in terms of pricing, this means that exists a clear price differentiation as the season changes (Saló, Garriga, Rigall-I-Torrent, Vila, & Sayeras, 2012).

In conclusion, the second home rental segment has gained great importance in the holiday accommodation, since it is seen as a response to the increasing volume of tourists arriving every day in touristic destinations (Guttentag, 2013; Karlsson & Dolnicar, 2016; Müller, 2014). It's a business opportunity to gain revenues with non-used second homes (Mohlmann, 2015).

2.2.1. The emergence of Airbnb

Airbnb is a customer-to-customer accommodation sharing service, which is the largest operating online community marketplace for accommodations in the world (Mohlmann, 2015). Zervas, Proserpio, and Byers (2017) describe Airbnb as a provider of holiday accommodations and a pioneer of the sharing economy. Alternatively, Airbnb describes itself as a "trusted community marketplace for people to list, discover and book unique accommodations around the world", and also as a customer-to-customer marketplace in the sharing economy (Zervas et al., 2017).

The increasing number of holiday renting accommodations contributed to the creations of Airbnb in 2008 by Brian Chesky and Joe Gebbia (Gant, 2016; Wang & Nicolau, 2017), and since its launch, it has grown very rapidly, with more fifty million guests using the platform by

2015. This kind of customer-to-customer sharing platforms transformed the accommodation sector in a very drastic way (Guttentag, 2013; Zervas et al., 2017). Airbnb offers over 500,000 listings in 33,000 cities and 192 countries (Wang & Nicolau, 2017). It started in San Francisco and New York, accomplishing one million bookings by 2011, and spreading to the rest of the US by 2012.

Airbnb has shaken the traditional market for tourism accommodations, like hotels, since it provides an online platform, working as a marketplace to rent spaces (like apartments, rooms, houses) from one ordinary person to another (Guttentag, 2013). Airbnb insists that it's not competing with the hospitality industry, but instead is expanding the accommodation market (Gurran & Phibbs, 2017), while on the other hand, independent studies suggest that this platform will have negative impacts on local hotel revenues (Oskam & Boswijk, 2016). These online platforms like Airbnb tend to increase the tourism volume in residential areas, which is the case of Sydney, Australia, where, since its spreading to that area in 2011, Airbnb listings more than double each year (Gurran & Phibbs, 2017).

It gives the option for the user to be a host, to pay a fee per month and share their renting options through this platform, or be a guest, and focus on finding the best deal on holiday home rentals (About us – Airbnb, 2018). Hosts list their second homes or spare rooms on the platform and establish their own nightly, weekly or monthly prices (Zervas et al., 2017). It is permitted to list all kinds of accommodations, and Airbnb claims that 57% of its spaces are entire apartments or homes, 41% are private rooms and 2% are shared rooms (Airbnb, 2018). Traditional "Bed & Breakfast" can list themselves on the platform.

Airbnb can be considered as a novel accommodation business model, which is built on modern internet technologies, overcoming some difficulties that owners of second homes had in making their accommodations known to potential guests. Hosts can generate and publish content regarding their available accommodations (Guttentag, 2013).

On this platform, hosts need to attend to four basic requirements: general evaluation, high frequency of response, reservation cancellations and reservations acceptance. (About us – Airbnb, 2018). After this, they offer their guests a local living experience, inferring that guests inflict no additional problems to the neighborhood and to the community, instead, bringing returns to local hosts and businesses (Oskam & Boswijk, 2016; Khadem, 2016). Although, there are many media reports on conflicts between Airbnb guests and local residents, because

of a series of reasons, such as noise, bad habits, garbage on the street, among others (Khadem, 2016).

The creation of Airbnb just expanded the situation that existed before, with more business opportunities for investors, tourism firms and property owners, and gave more visibility for those who wanted to rent rooms in their homes (Gant, 2016), while it could serve as a substitute for hotel stays and thus, affecting hotel revenues (Zervas et al., 2017). Through its invasive marketing, Airbnb expands the potential of the holiday rental homes sector by far (Gurran & Phibbs, 2017). Although Airbnb is a great business opportunity, many rentals are actually illegal, due to short-term rentals regulations in some countries (Guttentag, 2013).

Although the many economic benefits it brings to some regions, platforms like Airbnb bring up a series of questions about the effectiveness of the existing urban policies and planning on tourist and residential accommodations, and how can this business model affects the local housing market (Gurran & Phibbs, 2017).

Airbnb has developed an online reputation system, which allows and encourages guests to evaluate and review their stay, and hosts to evaluate their guests, creating a reputation for one and the other. Using star ratings, they can evaluate different features of their stay, like cleanliness, noise, location, etc., and post their written reviews online in the platform (Oskam & Boswijk, 2016; Zervas et al., 2017). Studies have shown that hosts usually charge higher prices if their accommodations are well-rated in the Airbnb platform (Gutt & Herrmann, 2015).

To use the platform, one searches by destination, dates, group size, and the website returns a list of accommodations available regarding those parameters, and then individual adverts can be selected for more detail or reservation, if you have an Airbnb profile (Guttentag, 2013).

Guttentag (2013), through the studies of Dolnicar & Otter (2003), points that, despite its big growth since it was created, the demand for this kind of services is not given, since it lacks a lot of areas that tourists and customers in general value very much, like service quality, staff friendliness, and security. This can lead to some limitations regarding the evolution of Airbnb, since it only appeals to a certain niche market, and it will never match the traditional accommodations in terms of general quality. Airbnb states that its service complements hotels, since it doesn’t target the same type of customers (Lawler, 2012).

In conclusion, Airbnb has grown exponentially, with some advantages and disadvantages, but one thing is right, that this platform has been a game-changer in the tourism accommodation

rentals industry, as it allowed second homeowners to share their rental opportunities (Gurran & Phibbs, 2017; Guttentag, 2013; Mohlmann, 2015; Zervas et al., 2017).

2.3. Rate optimization on short-term holiday rentals: A conceptual process

There aren't many studies regarding the best rate optimization process to apply a second home rental business, although, a series of factors have been pointed out as being crucial to consider when defining a pricing system for this segment.

With the fast-growing of shared economies for holiday accommodation rentals, examinations of pricing to apply to this accommodations offers important knowledge to stakeholders, since they can improve their profits by optimizing pricing processes (Wang & Nicolau, 2017).

The hospitality industry considers many price indicators in their price optimizations processes that are unfit to the accommodation offers in sharing economies, such as star ratings and corporate affiliations, while in this industry, the majority of price determinants are related to existing assets used for residential purposes (Guttentag, 2013). On the other hand, in the accommodation rental sector, demand volume changes depending on the season of the year, big events, holidays, stay length, far or close to the arrival date, and even days of the week (Figure 5.). Prices should be reduced for low demand days and increased during demand peaks, following a dynamic pricing strategy (PriceLabs, 2018).

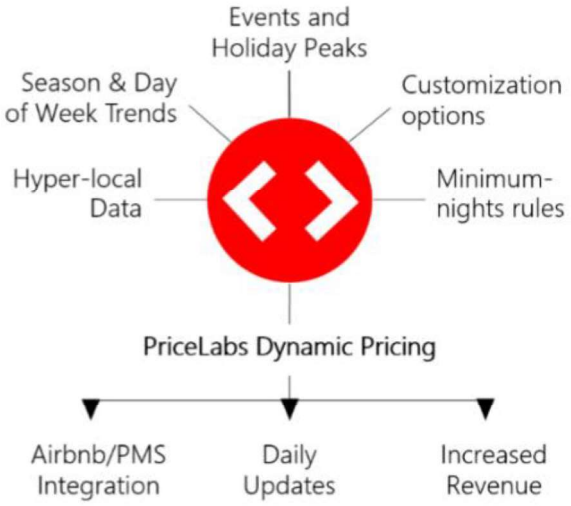


Figure 5. Dynamic Pricing with PriceLabs

Source: PriceLabs (2018)

In their work, Wang & Nicolau (2017) studied and identified the price determinants of shared economy's holiday accommodation rentals in a sample of around 180, 000 accommodations provided by the Airbnb platform. Through ordinary squares and quantile regression analysis, and among twenty-five different explanatory variables, the relationship between price and the other variables was studied (Table 2). OLS (Ordinary Least Squares) results revealed that twenty-four of the twenty-five variables under study are good price determinants, while QR analysis indicated that all of the twenty-five variables have a significant effect on price. These variables are divided into five subcategories: host attributes, site and property attributes, features and services, rental rules and online review ratings.

Table 2: Sources of price variability and demand segmentation in the hospitality industry

Variables	OLS	Quantiles				
		0.1	0.25	0.5	0.75	0.9
Constant	2.4083[a]	1.8154[a]	2.0438[a]	2.3263[a]	2.7156[a]	3.0724[a]
	(0.0296)	(0.0443)	(0.0426)	(0.0451)	(0.0484)	(0.0447)
Host Attributes						
Superhost	0.0837[a]	0.1206[a]	0.1102[a]	0.0807[a]	0.0534[a]	0.0422[a]
	(0.0039)	(0.0061)	(0.0048)	(0.0041)	(0.0046)	(0.0064)
Host listings count	0.0006[a]	0.0007[a]	0.0006[a]	0.0008[a]	0.0008[a]	0.0005[a]
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Host's profile picture	−0.1154[a]	−0.0856[a]	−0.0994[a]	−0.1142[a]	−0.1313[a]	−0.1473[a]
	(0.0241)	(0.0267)	(0.0330)	(0.0392)	(0.0414)	(0.0357)
Host identity verified	0.0856[a]	0.0679[a]	0.0843[a]	0.0948[a]	0.0861[a]	0.0767[a]
	(0.0025)	(0.0039)	(0.0031)	(0.0029)	(0.0033)	(0.0042)
Site & Property Attributes						
Distance (km)	−0.0059[a]	−0.0053[a]	−0.0054[a]	−0.0058[a]	−0.0062[a]	−0.0063[a]
	(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0003)	(0.0003)
Accommodation type 1	−0.0827[a]	−0.1390[a]	−0.1073[a]	−0.0619[a]	−0.0422[a]	−0.0329[a]
	(0.0033)	(0.0056)	(0.0043)	(0.0038)	(0.0042)	(0.0058)
Accommodation type 2	−0.0890[a]	−0.2483[a]	−0.1175[a]	−0.0303[b]	0.0345[a]	0.0556[a]
	(0.0088)	(0.0193)	(0.0147)	(0.0119)	(0.0113)	(0.0162)
Entire home/apartment	0.8948[a]	1.0171[a]	0.9555[a]	0.8419[a]	0.7535[a]	0.6667[a]
	(0.0090)	(0.0151)	(0.0127)	(0.0128)	(0.0171)	(0.0187)
Private room	0.3419[a]	0.4097[a]	0.3727[a]	0.3162[a]	0.2582[a]	0.1735[a]
	(0.0090)	(0.0148)	(0.0128)	(0.0128)	(0.0169)	(0.0183)
Accommodates	0.0616[a]	0.0554[a]	0.0589[a]	0.0660[a]	0.0711[a]	0.0780[a]
	(0.0010)	(0.0015)	(0.0012)	(0.0014)	(0.0016)	(0.0021)
Bathrooms	0.1085[a]	0.0515[a]	0.0798[a]	0.1237[a]	0.1705[a]	0.2092[a]
	(0.0027)	(0.0039)	(0.0040)	(0.0042)	(0.0055)	(0.0067)
Bedrooms	0.1248[a]	0.0976[a]	0.1111[a]	0.1216[a]	0.1318[a]	0.1275[a]
	(0.0021)	(0.0031)	(0.0027)	(0.0029)	(0.0034)	(0.0043)
Amenities & Services						
Real bed	0.1555[a]	0.0819[a]	0.1244[a]	0.1831[a]	0.2106[a]	0.2336[a]
	(0.0055)	(0.0070)	(0.0060)	(0.0055)	(0.0083)	(0.0085)
Wireless Internet	0.0951[a]	0.1331[a]	0.1262[a]	0.0886[a]	0.0646[a]	0.0751[a]
	(0.0052)	(0.0089)	(0.0073)	(0.0068)	(0.0076)	(0.0099)
Breakfast	−0.0106[a]	0.0108	−0.0031	−0.0103[b]	−0.0253[a]	−0.0210[a]
	(0.0042)	(0.0068)	(0.0054)	(0.0048)	(0.0055)	(0.0069)
Free parking	0.0811[a]	0.1184[a]	0.1103[a]	0.0891[a]	0.0433[a]	0.0084
	(0.0029)	(0.0049)	(0.0037)	(0.0033)	(0.0036)	(0.0049)
Instant bookable	−0.0665[a]	−0.0606[a]	−0.0614[a]	−0.0607[a]	−0.0680[a]	−0.0756[a]
	(0.0031)	(0.0048)	(0.0040)	(0.0036)	(0.0041)	(0.0051)
Rental Rules						
Cancellation policy (Moderate plus strict)	0.0448[a]	0.0446[a]	0.0431[a]	0.0478[a]	0.0490[a]	0.0406[a]
	(0.0028)	(0.0043)	(0.0035)	(0.0034)	(0.0039)	(0.0049)
Smoking allowed	−0.2654[a]	−0.2253[a]	−0.2536[a]	−0.2804[a]	−0.2876[a]	−0.2588[a]
	(0.0035)	(0.0053)	(0.0042)	(0.0040)	(0.0050)	(0.0063)
Required guest's profile picture	0.0102	0.0096	−0.0008	0.0178	0.0096	0.0197
	(0.0082)	(0.0126)	(0.0101)	(0.0103)	(0.0095)	(0.0129)
Required guest's phone verification	0.0220[a]	0.0217	0.0303[a]	0.0172	0.0272	0.0141[a]
	(0.0071)	(0.0115)	(0.0085)	(0.0090)	(0.0088)	(0.0105)
Online Reviews: Number & Ratings						
Reviews per year	−0.0010[a]	0.00004	−0.0002[b]	−0.0007[a]	−0.0015[a]	−0.0022[a]
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Review scores for rating	0.0087[a]	0.0091[a]	0.0091[a]	0.0091[a]	0.0086[a]	0.0082[a]
	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0002)

Notes: [a] prob <1%; [b] prob <5%.
Standard errors in parenthesis.

Source: Wang & Nicolau (2017)

Albert Saló & Anna Garriga (2011) also studied the attributes that can explain the overall price of a second home rentals' sample of Costa Brava, Spain. This sample contains over one thousand observations, and a semilogarithmic specification has been used to make it easier the interpretation of coefficients. The OLS regression model has been used in this process. Through studied literature, the authors considered about twelve relevant attributes (listed in Figure 6.) divided into three categories: Inside and outside housing characteristics, both tourist and housing characteristics and tourist characteristics. Results of this studies indicate that among the twelve attributes considered, home size, the existence of a swimming pool, beach distance and seasonality are the attributes which affect the most accommodation's price, although the number of rooms, the housing type and the physical location have also a contribution to accommodation's final price.

Variable	Coded name
Rooms	Rooms
Home size	*size home: less 50 m²*
	size home: 50–100 m²
	size home: 101 m² and more
Terrace/garden size	*terrace/garden: less 10 m²*
	terrace/garden: 10–30 m²
	terrace/garden: 31 m² and more
Common garden	common garden
Housing type	*terraced house*
	detached house
	apartment
Car park	car park
Swimming pool	swimming pool
Beach distance	beach distance: 0–30 m
	beach distance: 30–100 m
	beach distance: 101–300 m
	beach distance: 301–1,000 m
	beach distance: 1,001–3,000 m
	beach distance: 3,001–5,000 m
	beach distance: 5,001 m and more
Sea views	sea views
Physical location (municipality)	Castelló d'Empúries
	Begur
	l'Escala
	l'Estartit
	Llançà
	Lloret
	Castell-Platja d'Aro
	Roses
	Calonge
	St Feliu de Guixols
Commercialization channel (intermediaries)	wholesaler
	Internet
Time period	*low season*
	medium season
	medium-high season
	high season
	upper-high season
	extreme-high season

Figure 6. Variables used in the hedonic regression

Source: Saló et al. (2012)

Although there aren't many studies about rate optimization in the shared economy holiday accommodation rentals, there are already some revenue management solution models developed to help firms optimize their pricing approaches. PriceLabs is an RM solution for short term rentals, who focuses on finding the best rate possible for every night in the holiday accommodation rental market, using hyper-local supply, demand trends and user preferences to do that (PriceLabs, 2018).

In order to recommend the best price, PriceLabs consider a lot of factors that can change through time: Base Price- a base price is estimated for the listing, using listing characteristics and performance. The indicators of the listing performance will help the user making a decision regarding price reduction or increase of a specific listing. Historic Trends- historical data for vacation rentals and hotels around the year has been gathered to better understand seasonal, weekly and holiday/event trend. The PriceLabs algorithm is constantly learning from historical trends to also understand booking behaviors around the world. Future Forecasts- Information about future-looking data and booking trends is predicted, and hotel prices and availability are also analyzed to understand the upcoming demand. The algorithm knows that if, for example, listings in the user's area are booked to seventy percent, they can increase their listing price. Customizations- The user is able to set some customized rules and restrictions about price and minimum stay, which is a benefit for users running their operations differently. Data Specific Overrides- The user can apply specific data about the location or another variable that the algorithm doesn't know, and so, the recommendation doesn't get compromised (PriceLabs, 2018).

(This page was intentionally left blank)

3. Methodology

In this work, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology will be the one chosen, in order to better deal with the big amount of data gathered. It provides a structured approach to planning a data mining project. CRISP – DM was developed in 1996 by three analysts who represented NCR Systems Engineering Copenhagen, DaimlerChrysler AG, and SPSS Inc.

This methodology is business-oriented and so, it is more appropriate for this study. CRISP-DM breaks the process of data mining into six different phases (Figure 7): business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al., 2000). In this work, however, only the first five phases will be taken into account, since the results will be presented in a further chapter, as well as its discussion.

The first phase, or business understanding, focus on learning the project´s main objectives and collecting the more appropriate requirements to achieve the proposed goals. This knowledge is to be converted into a data mining problem and a plan is designed to achieve the main objectives.

In the second phase, or data understanding, a big amount of data is collected, and many tasks are put in motion in order to discover data problems, evaluate the data quality and even to find patterns previously unknown that can be the first step to a new hypothesis.

The third phase, or data preparation, complements the previous phase. The data is arranged in order to go to data mining tools and be treated. A process of selecting tables and attributes, as well as a cleaning and transformation process from the initial raw data is done until all the data is ready and well-suited for the business objectives.

The fourth phase, or the modeling phase, is in charge of choosing and selecting the best and more adequate modeling techniques available and apply them, to achieve the best values possible. Since some techniques have specific requirements regarding the input data, stepping back to the data preparation phase is usually necessary.

The fifth phase, the evaluation phase, reviews the model, step by step, and makes sure that the model is correctly built to achieve the business objectives. It is important to find out if any important business issue has not been considered.

Finally, the last phase, or deployment, is in charge of organizing and presenting the information in a way that is convenient to the customer, plus the plan maintenance, and the production of the final report of the project.

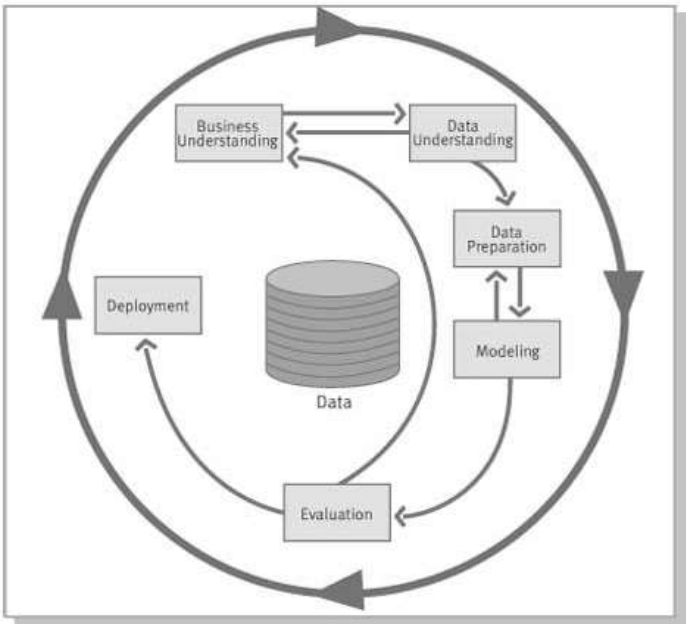


Figure 7. CRISP-DM different phases

Source: Chapman et al. (2000)

3.1. Business understanding

In this first phase, the objective is to understand the business context, and its requirements in order to achieve its main goals, that is, to identify the important factors that can influence the outcome of the project. It is very important to get to the customer´s main objectives and preferences, otherwise, there will be right answers for the wrong questions (Chapman et al., 2000).

All the resources, constraints, assumptions and other facts should be considered in determining the requirements and data analysis, and there will have to be more detailed fact-finding about these (Chapman et al., 2000).

In this study, it was pointed that the arrival of a big amount of tourists every day in Lisbon, can be a huge business opportunity for short-term rentals, and the companies responsible for the properties' management, namely FeelsLikeHome (the study-case company for this work), can

maximize their profits with an optimized rating process that guarantees the maximum occupancy rate for every house, every month.

There are not many studies about optimized processes to maximize profits in the short-term rental industry, and so, studying these processes in the airline industry and the hospitality industry, allows us to understand that, to maximize the revenues, rates per night should be flexible, and change according to different indicators.

To achieve this optimized model, the rating process should incorporate historical data that allow the user to collect information about which are the indicators that costumers value the most when choosing and deciding to rent a house for their vacations, and increasing or decreasing the rate per night according to the indicators that have a bigger impact on the renting price.

Through the analysis of FeelsLikeHome´s historical data records, regarding the properties available for renting in Lisbon, and every reservation´s details in the period between January 2017 and May 2019, the main business objective is to predict the profiles of the houses in which the rate per night can be raised or diminished, for having a very high or a very low occupancy rate, respectively. In analytic terms, the first objective is to study if there is any relation between the occupancy rate and the rate per night implemented by FeelsLikeHome in their properties, and if there is, how strong is it. Then, to study which are the variables that have a bigger influence in the occupancy rate and build properties´ profiles, gathering several variables, in which the occupancy rate is higher and profiles in which the occupancy rate is lower, so a decision about the increasing or decreasing of the rate per night can be made. Finally, the set of matrices to propose changes to the currently implemented rates, based on the most important predictors for the occupancy rate, through the corresponding occupancy rate, shall present coefficients meaning an increase (a number above 100) or decrease (a number below 100) of the currently implemented rate per night (number 100).

3.2. Data understanding

The data understanding phase includes four sub-phases in order to successfully have a proper understanding of the data collected. First, the collection of data that are listed in the project resources, and a consequent listing of all dataset acquired. Then, a meticulous description of that data, including its properties and quantity, in order to report if it is everything according to plan. The third thing to do is an exploration of the data, or analyzing the data through, the

relations between a small number of attributes, and to improve the data description and quality reports. Finally, verification of data quality is imperative, in order to make sure that the data is complete, to identify the error and how frequent are they and to find solutions for these errors (Chapman et al., 2000).

The data used in this study was provided by FeelsLikeHome, and included information about 480 houses located in the urban area of Lisbon, with 10 variables (Table 3), some describing features of the house, and others pointing relevant information about the history of the property in the company. From the initial sample, several properties were excluded, due to criteria defined previously and discussed with the company FeelsLikeHome. Only the properties that have been active anytime between January 2017 and May 2019 were selected, with these two months as lower and upper boundaries. The properties that did not have any information about the neighborhood in which they were located or about the typology of the house were excluded (very relevant information for the study in progress), so as the properties that did not have any reservations made in the established period, and thus, no relevant information about the occupancy rate. From the initial sample of 480 properties, 333 were left to be analyzed.

Regarding the reservations, a table with 12 variables (Table 4), the same period was defined for this domain, between January 2017 and May 2019, and every reservation outside this interval was excluded, because a relatively recent sample will better predict the preferences of the guests arriving in Lisbon. Canceled reservations were also excluded, since the company does not register cancelation rates neither uses this information to set their rates. Two variables were created for this table, “Year” and “Month”, to be easier to filter and find reservations in certain periods.

The sample provided also counted on a table with information about the countries´ presence on the reservation history, in order to investigate which are the nationalities that contribute the most to this business and adjust the rates to their economic possibilities. In this matter, for every month between January 2017 and May 2019, every nationality that has ever booked anything in FeelsLikeHome´s properties are listed, and there are 9 variables (Table 5), related with the reservations made by each and every country, every month.

Finally, two auxiliary tables from FeelsLikeHome´s database were used, one with more detailed information about the properties´ monthly occupancy rates, including 12 variables (Table 6), and other related to every property´s statistical data, in terms of reservations made,

using 12 variables (Table 7) for the purpose. A main Excel table was created with all the relevant information gathered, divided by Client ID and month, between the period from January 2017 and May 2019. This was reflected in 5999 lines of Excel.

Table 3: Information about Table "Properties"

Variable ID	Variable name	Description	Nature	Domain	Decision
ClientID	Client ID	The Id which identifies each property	Ordinal	Num	Property Ident.
Neighborhood	House´s Neighborhood	The neighborhood where the property is located	Nominal	String	Data preparation /Modelling
City	City	The city where the property is located	Nominal	String	Excluded
Typology	House´s typology	The number of rooms in the property	Ordinal	String	Modeling
FloorNumber	House´s Floor number	The floor where the property is located	Ordinal	Num	Modeling
Elevator	Elevator	Existence of an elevator	Flag	String	Modelling
Parking	Parking	Existence of car parking	Flag	String	Modeling
MaxOccupancy	Maximum occupancy	Maximum number of guests that the property can shelter	Nominal	Num	Data preparation /Modelling
Inactive	Inactive	Activeness or inactiveness of the property	Flag	String	Excluded (filter only)
StartDate	Active start date	The date when the property became available for renting	Ordinal	Date	Excluded (filter only)
InactiveDate	Inactive date	The date when the property became inactive	Ordinal	Date	Excluded (filter only)

Table 4: Information about Table “Reservations”

Variable ID	Variable name	Description	Nature	Domain	Decision
AccessID	Access ID	The Id that identifies the reservation	Ordinal	Num	Reservation ident.
Site	Booking Site	The website from where the reservation was made	Nominal	String	Data preparation
ClientID	Client ID	The Id which identifies each property	Ordinal	Num	Excluded (filter only)
Year	Year	The year of the registration	Ordinal	Num	Modeling (filter)
Month_name	Month name	The month of the registration	Nominal	String	Modeling (filter)
CheckInDate	Check-in date	The date when the check-in was made	Ordinal	Date	Data preparation
CheckOutDate	Check-out date	The date when the check-out was made	Ordinal	Date	Data preparation
GuestCountry	Guest’s country	Guest’s nationality	Nominal	String	Data preparation
NumofGuests	Number of Guests	Number of guests included in the reservation	Nominal	Num	Data preparation
Canceled	Canceled	Cancelation or non-cancelation of the reservation	Flag	String	Excluded
ReservationDateCreated	Reservation´s creation date	The date when the reservation was created	Ordinal	Date	Data preparation
ReservationSoldValue	Reservation sold value	The value that the costumers paid for all the reservation	Ordinal	Date	Excluded

Table 5: Information about Table “Nationalities”

Variable ID	Variable name	Description	Nature	Domain	Decision
Year	Year	The year of the registration	Ordinal	Num	Registration ident.
Month	Month	The month of the registration	Nominal	String	Registration ident.
CountryCode	Country´s code	Letter code of the country	Nominal	String	Excluded
PercentageOfTotal	Percentage of total	Percentage of total reservations made by each country	Nominal	Num	Data preparation
Country	Country	Country´s name	Nominal	String	Data preparation/Registration ident.
NumReservations	Number of reservations	Number of reservations made by each country	Nominal	Num	Excluded
TotalValue	Total value	Total value of revenue provided by each country	Nominal	Num	Excluded
TotalDays	Total days	Total days that guest from the country have stayed in FeelsLikeHome´s properties	Nominal	Num	Excluded
AvgRates	Average rates	Average rates per night paid by each country´s guests	Nominal	Num	Excluded

Table 6: Information about Table “Occupation rate by month”.

Variable ID	Variable name	Description	Nature	Domain	Decision
ClientID	Client ID	The Id which identifies each property	Nominal	Num	Registration ident.
Year	Year	The year concerning the occupancy´s information	Ordinal	Num	Registration ident.
Month	Month	The month concerning the occupancy´s information	Nominal	String	Registration ident.
ClientName	Client name	The name of the property´s owner	Nominal	String	Excluded
Typology	Typology	The number of rooms in the property	Ordinal	String	Data preparation/Modelling
NumReservations	Number of reservations	Number of reservations made for that property	Nominal	Num	Modeling
Num_days_occup	Total Days booked	Total number of days the property was booked	Nominal	Num	Data preparation
TotalExtraDays	Total Extra Days	Total number of days the property was occupied by its owner	Nominal	Num	Excluded
Available_days _for_rent	Days available for renting	Total number of days, in the concerning month and year, the property was available for rent	Nominal	Num	Data preparation
CalendarDays	Calendar days of the month	Number of days in the concerning month	Nominal	Num	Excluded

Table 7: Information about Table “AVG reservations”.

Variable ID	Variable name	Description	Nature	Domain	Decision
ClientID	Client ID	The Id which identifies each property	Nominal	Num	Registration indent.
Year	Year	The year concerning the property´s reservations’ data	Ordinal	Num	Registration ident.
Month_name	Month name	The month concerning the property´s reservations’ data	Nominal	String	Registration ident.
ClientName	Client name	The name of the property´s owner	Nominal	String	Excluded
Typology	Typology	The number of rooms in the property	Ordinal	String	Modeling
MaxGuests	Maximum Guests	Maximum Guests allowed in the property	Nominal	Num	Data preparation/Modelling
NumReservations	Number of Reservations	Number of reservations made for each property	Nominal	Num	Excluded
TotalSellValue	Total Sell Value	Income of each property	Nominal	Num	Excluded
TotalDays	Total of Days	Total of days that each house was occupied per month	Nominal	Num	Excluded
TotalGuests	Total of Guests	Total number of guests each house had per month	Nominal	Num	Excluded
AvgDuration	Average Duration	Average duration of the stay in each property	Nominal	Num	Excluded
Avg_rate_night	Average Rate	Average rate per night implemented in each property	Nominal	Num	Modeling

Among the 5 tables, some variables were excluded due to the lack of relevance that they had for the study, or simply because they had many blank answers, which prevented any conclusions drawn to be accurate and realistic. Those variables are:

From Table 3 “Properties”:

- City (All properties are located in Lisbon)
- Inactive (used as filter only)
- StartDate (used as filter only)
- InactiveDate (used as filter only)

From Table 4 “Reservations”:

- ClientID (used as filter only, and to identify the property booked)
- Year (used as a filter only, and to identify the year of the reservation)
- Month_name (used as a filter only, and to identify the month of the reservation)
- Canceled (Irrelevant for the case, since all canceled reservations are excluded)
- ReservationSoldValue (Irrelevant for the case)

From Table 5 “Nationalities”:

- CountryCode (Countries are identified by its name instead of a country code)
- NumReservations (Irrelevant for the case, since it doesn´t discriminate de property)
- TotalValue (Irrelevant for the case, since it doesn´t discriminate de property)
- TotalDays (Irrelevant for the case, since it doesn´t discriminate de property)
- AvgRates (Irrelevant for the case, since it doesn´t discriminate de property)

From Table 6 “Occupation rate by month”:

- ClientName (Name of the property´s owner is irrelevant)

From Table 7 “Avg Reservations”:

- ClientName (Name of the property´s owner is irrelevant)
- NumReservations (This information had already been taken from Table 4)
- TotalSellValue (Irrelevant, since it doesn´t give the average rate per night)
- TotalDays (This information had already been taken from Table 4)
- TotalGuests (Irrelevant, since there is information about the average number of guests)
- AvgDuration (Irrelevant for the case)

In the following image (Figure 8) is presented every table provided by the company (FeelsLikeHome) and their relations, plus which attribute is the identifier of each table. The

table “Nationalities” doesn´t have any relations with the other tables, since it is independent and presents only the general information about which countries have the biggest amount of reservations made with the company.

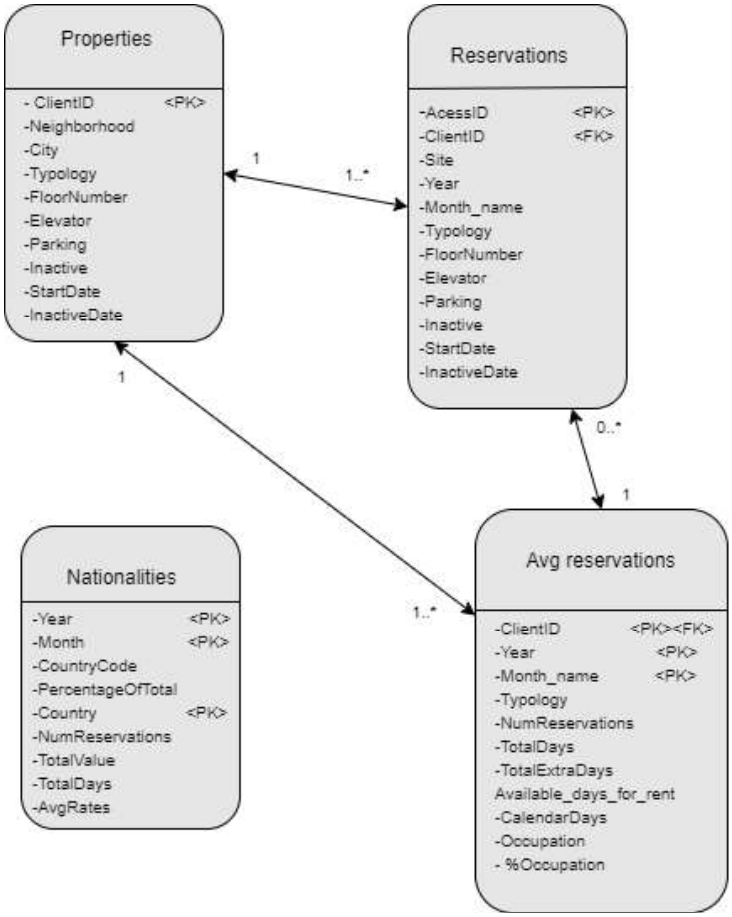


Figure 8. Tables provided by FeelsLikeHome and their relations.

3.3. Data preparation

In this phase, the final dataset is produced, in order to go to the modeling phase without any errors and with every variable correctly implemented. A selection of the relevant data is made, considering many criteria important for the data mining goals. Next, cleaning, constructing and integrating data is very important, through different analysis techniques, subsets of data may be excluded if not suitable for the purpose, and the construction of derived attributes may be needed, by transforming old values into new ones for existing attributes, or creating new variables by means of old ones. Finally, to assure that the modeling process runs as expected, is imperative that the transformed data is formatted, to prevent the loss of meaning by some variables (Chapman et al., 2000).

Regarding the Table “Properties”, two new variables were created with old variables from this table (Table 8). The variable Neighborhood was used to create a new variable Historical_area (Whether the property is located in a historical area of the city or not) in order to divide the many neighborhoods into two different groups, “Yes” or “No”. Also, the new variable Touristic_area was created based on the variable Neighborhood (Whether the property is located in a touristic area or not), also to reduce the number of categories.

The division of neighborhood into historical and non-historical was made using the following characterization (Câmara Municipal de Lisboa, n.d):

- Historical neighborhoods (Yes): Alfama, Almirante Reis, Areeiro, Av. Da Liberdade, Av. Novas, Bairro Alto, Baixa, Bica, Cais do Sodré, Campo de Ourique, Castelo, Chiado, Estefânia, Estrela, Graça, Intendente, Lapa, Madragoa, Marquês de Pombal, Martim Moniz, Mercês, Mouraria, Príncipe Real, Rato, Restauradores, S.Bento, S.José, Santa Catarina, Santos-o-velho, Sé.

- Non-historical neighborhoods (No): Ajuda, Alcantâra, Beato, Belém, Benfica, Campolide, Expo, Laranjeiras, Olivais, Restelo, Saldanha, Santa Marta.

The division of Lisbon´s neighborhood into touristic and non-touristic is a little bit different, focusing on shopping spots and nice views, besides the previously referred, center of the city. So, this segmentation is made using the following characterization (Lisboando, n.d):

- Touristic neighborhoods (Yes): Baixa, Chiado, Bairro Alto, Príncipe Real, Alfama, Graça, Belém, Av. Da Liberdade.
- Non-touristic neighborhoods (No): Ajuda, Alcantâra, Alfama, Almirante Reis, Areeiro, Av. Novas, Beato, Benfica, Bica, Cais do Sodré, Campo de Ourique, Campolide, Castelo, Estefânia, Estrela, Expo, Intendente, Lapa, Laranjeiras, Madragoa, Marquês de Pombal, Martim Moniz, Mercês, Mouraria, Olivais, Rato, Restauradores, Restelo, S.Bento, S.José, Saldanha, Santa Catarina, Santa Marta, Santos-o-velho, Sé.

The variable For_Two is created through the existing variable MaxOccupancy, and it tells if a property is prepared to shelter a family (more than two people) or if it is a small house, only for one or two people, being the categories “Yes” and “No”.

Table 8: New variables created based Table “Properties”.

Variable ID	Variable Name	Description	Nature	Domain	Origin variable
Historical_area	Historical area	Indicator if the property is located in a historical area	Flag	String	Neighborhood
Touristic_area	Touristic area	Indicator if the property is located in a touristic area	Flag	String	Neighborhood
For_two	House for two	Indicator if the property is appropriate only for two people or less	Flag	String	MaxOccupancy

Based on Table 4 (Table “Reservations”), three new variables were created to support this study (Table 9). First of all, the variable “Days booked in advance” (id: DaysBookedInAdvance) was created, so it is possible to get some information about the guest´s tendencies concerning the reservation making process, whether they book their houses in advance, or if they prefer to do that more in the nick of time. This new variable was created using two old variables, “Check in Date” (id: CheckInDate) and “Reservation´s Creation Date” (id: ReservationDateCreated), being the difference between these two dates, the value of this new variable. However, the number of days in advance with which the guests usually book their properties doesn´t contribute for further analysis, since it gives a single value for each reservation made. So, through a pivot table created in Excel, another new variable was created, “Average booking advance” (id: Avg_booking_advance), so it was possible to associate an average number of days in advance with which the guests made their reservations, to every property, per month.

The variable “Booking site” (id: Site) was used to create the variable “Favorite distribution channel” (id: fav_dist_channel), which indicates the most commonly used platform for booking FeelsLikeHome´s properties, per month. The categories for this variable boil down to “Airbnb”, “Booking”, “FeelsLikeHome” and any others besides those three is defined as “Other”. If a month has more than one most common preference, regarding the distribution channel, it is labeled as “Don´t have”.

Table 9: New variables created based on Table “Reservations”.

Variable ID	Variable name	Description	Nature	Domain	Origin variable
DaysBookedInAdvance	Days booked in advance	Number of days in advance with which the guests made their reservations	Nominal	Num	CheckInDate/ReservationCreatedDate
Avg_booking_advance	Average booking advance	Average number of days in advance with which the guests made their reservations	Nominal	Num	DaysBookedInAdvance
Fav_dist_channel	Favorite distribution channel	Most commonly used platform for booking properties	Nominal	String	Site

In order to better understand the impact of the different countries in this business area, two variables were created (Table 10), to identify the nationality that influences the most the success of the company´s business core, and how strong is that influence. First, the variable “Percentage of predominant country” (id: perc_predominant_country) reflects the biggest percentage of reservations made by the same nationality, and it was created using the variable “Percentage of total” (id: PercentageOfTotal). Secondly, the variable “Predominant Country” (id: pred_country_month) was created with the previously created “Percentage of predominant country”, by identifying which is the country with the biggest percentage of rentals.

Table 10: New variables created based Table “Nationalities”.

Variable ID	Variable name	Description	Nature	Domain	Origin variable
Perc_predominant_country	Percentage of predominant country	Biggest percentage of rentals by one nationality	Nominal	Num	PercentageOfTotal
Pred_country_month	Predominant country	Nationality with the biggest volume of reservations made	Nominal	String	Perc_predominant_country/Country

Information about the properties´occupancy rates is also very relevant, and thus, through the variables in Table 6 (Table “Occupation rate by month”), adjustments were made, and three

new variables were created (Table 11). First, a variable named “Occupation” was created, and it reflected the coefficient of occupancy of the property, in a certain month. This variable is calculated by dividing the number of days in which the property was in fact occupied, by the number of days the house was available for renting, using, for this, the variables “Total days booked” and “Days available for renting”. After this, the “Property´s occupation rate” variable was calculated by assigning a percentage to the “Property´s occupancy” variable, and finally, a new variable was created to indicate the occupancy rate for the previous month .

Table 11: New variables created based on Table “Occupation rate by month”

Variable ID	Variable name	Description	Nature	Domain	Origin variable
Occupation	Property´s occupancy	The property´s occupancy, in decimal values	Nominal	Num	Num_days_occup/Available-days_for_rent
Occupation_rate	Property´s occupancy rate	The property´s occupancy, in percentage values	Nominal	Num	Occupation
Occ_rate_month_before	Occupancy rate of the previous month	Indicates the value of the occupancy rate for the previous month, for the house in question.	Nominal	Num	Occupation_rate

Regardless the data collected, two more variables were created (Table 12), one to verify the influence of the month in the guest´s decision to rent a house or not, and other to verify if the occurrence of a big event in the city affects directly the willingness of the guests to rent more houses. First, the variable “Season” was created, and it indicates if the month is of high volume of tourist or low volume of tourists, accordingly to the following criteria, based on the main holidays:

- High Season: April, June, July, August, December.
- Low Season: January, February, March, May, September, October, November.

The variable “Big Event” was created to indicate in which months big events are happening in the city. Those big events can include music or business events and religious holidays and were defined through the following criteria:

- Big Events (Yes):

- Carnival (February)
- Easter (April)
- "Santos Populares" (Lisbon´s typical holiday) (June)
- Music Festivals in Lisbon – NOS Alive (July)
- Web Summit (November 2017 and November 2018)
- Christmas (December)

Table 12: New variables created without any original data.

Variable name	Variable ID	Description	Nature	Domain	Original variable
Season	Season	Indicates whether the month in question is classified as high season or low season	Flag	String	None
Big Event	Big_Event	Indicates whether the month has any big event happening on it	Flag	String	None

At the end of the data preparation phase, the most relevant variables for the problem were selected, new ones created and prepared to be inserted into the models chosen (Table 13). In Tables 14 and 15 there are some descriptive statistics to these variables, as information about the missing values of the variables.

Table 13: Final variables to use as input in the modeling phase.

Variable ID	Variable name	Description
Month_number	Month number	The correspondent number of the month
Occupation_rate	Property´s occupancy rate	The property´s occupancy rate in the corresponding year and month
Occ_rate_month_before	Occupancy rate for the previous month	Indicates the value of the occupancy rate for the previous month, for the house in question.
NumReservations	Number of reservations	Number of reservations made for each property, in the corresponding year and month
Max_occupancy	Maximum occupancy	Maximum number of guests that each property can shelter
Avg_rate_night	Average rate	Average rate per night implemented in each property, in the corresponding year and month
Season	Season	Indicates whether the corresponding month is classified as high season or low season
Big_event	Big Event	Indicates whether the corresponding month has any big event happening on it
Typology	Property´s typology	The number of rooms in the property
For_two	Property for two	Indicator if the property is appropriate only for two people or less
FloorNumber	Property´s floor number	The floor where the property is located
Fav_dist_channel	Favorite distribution channel	Most commonly used platforms for booking properties
Neighborhood	Property´s neighborhood	The neighborhood where the property is located
Historical_area	Historical area	Indicator if the property is located in a historical area
Touristic_area	Touristic area	Indicator if the property is located in a touristic area
Pred_country_month	Predominant country	Nationality with the biggest volume of reservations made in the corresponding year and month
Percentage_pred_country	Percentage of predominant country	Biggest percentage of rentals by one single nationality
Avg_booking_advance	Average booking advance	Average number of days in advance with which the guests made their reservations, in the corresponding year and month
Parking	Parking	Existence or non-existence of car parking spot
Elevator	Elevator	Existence or non-existence of elevator

Table 14: Descriptive Statistics of the final variables to use as inputs (I).

Variable ID	Descriptive statistics			Missing Values
Month_number	Count		5999	
	Highest Frequency	5	632	
	Lowest Frequency	6	401	
Occupation_rate	Count		5999	
	Mean		54.47	
	Std.Deviation		31.67	
	Mode		0	
NumReservations	Count		5999	
	Highest Frequency	7	832	
	Lowest Frequency	14	1	
	No reservations	0	935	
Max_occupancy	Count		5999	
	Highest Frequency	4	3040	
	Lowest Frequency	10	4	
Avg_rate_night	Count		5049	950 (15.84%)
	Mean		99.41	
	Std.Deviation		56.99	
	Mode		52	
Season	Count		5999	
	Highest Frequency	Low	3719	
	Lowest Frequency	High	2280	
Big_event	Count		5999	
	Highest Frequency	No	3110	
	Lowest Frequency	Yes	2889	
Typology	Count		5999	
	Highest Frequency	T2	2605	
	Lowest Frequency	T6	13	
For_two	Count		5999	
	Highest Frequency	No	5223	
	Lowest Frequency	Yes	776	

Table 15: Descriptive Statistics of the final variables to use as inputs (II).

Variable ID	Descriptive statistics			Missing Values
FloorNumber	Count		5721	278 (4.63%)
	Highest Frequency	2	1366	
	Lowest Frequency	13	7	
Fav_dist_channel	Count		5999	
	Highest Frequency	Airbnb	2331	
	Lowest Frequency	Booking	150	
	Frequency	Don´t have	2661	
Neighborhood	Count		5999	
	Highest Frequency	Baixa	587	
	Lowest Frequency	Benfica	6	
Historical_data	Count		5999	
	Highest Frequency	Yes	5448	
	Lowest Frequency	No	551	
Touristic_area	Count		5999	
	Highest Frequency	No	3026	
	Lowest Frequency	Yes	2973	
Pred_country_month	Count		5999	
	Highest Frequency	Spain	3143	
	Lowest Frequency	Portugal	193	
Percentage_pred_country	Count		5999	
	Mean		12,9	
	Std.Deviation		3.72	
	Mode		12,79	
Avg_booking_advance	Count		5999	
	Mean		38.36	
	Std.Deviation		41,6	
	Mode		0	
Parking	Count		5999	
	Highest Frequency	No	5452	
	Lowest Frequency	Yes	547	
Elevator	Count		5999	
	Highest Frequency	No	4187	
	Lowest Frequency	Yes	1812	

3.4. Data analysis techniques and modeling

This phase focuses primarily on selecting the most adequate modeling techniques to apply to the data and calibrating their parameters to optimal values. A suitable model is built, and validity tests are applied to verify its quality (Chapman et al., 2000). In this study, an outlier analysis was made, and the predictive approach was used.

3.4.1. Descriptive techniques: Univariate and bivariate

Descriptive statistics will be used to accomplish the objective to study and describe the "rate per night" and the "occupation rate" variables in separate, and their relationship between one another. These techniques provide some basic information about the sample (one or more variables) and about observations made. That information may be quantitative, like calculating the *mean* or the *std.deviation* for example, or it can be visual, in the form of graphics. These summaries withdrawn from the description process of a variable, being so simple and limited, may only be the initial description as part of extensive analysis, or they may be sufficient for the purpose of the study (Hand, Mannila, & Smyth, 2001).

Within the descriptive statistics, we have two ways of describing variables. First, the univariate analysis, which involves describing variables individually, with no relation to other variables. In this analysis there are three characteristics of a variable that we need to look at, the distribution, a summary of the frequency or the percentage of individual values for a variable; the central tendency, estimated by the mean, median and mode; and finally its dispersion, referring to the spread of values around the central tendency, being the range and the standard deviation the most common measures for this part (Trochim, 2006).

The bivariate analysis and the multivariate analysis are suited for situations where the sample includes more than one variable, and it focuses not only on describing the variables individually, but also to describe the relationship between pairs of variables, through quantitative measures of dependence, as the correlation coefficient, which indicates if there is any relationship between the variables, positive or negative, depending on whether the coefficient is positive or negative, and how strong that relationship is, and as covariance, a numerical measure of linear association between two variables. A positive value of covariance indicates a positive linear relationship between the variables, while a negative value indicates a negative linear relationship. A covariance of 0 indicates no linear relationship. Cross-

tabulations and scatter pots are often used to capture something about the relationship between two or more variables. The first method can be used regardless of whether the variables are quantitative or qualitative and it is basically a presentation of two or more variables in a tabular form, to help to identify a relationship between the variables. A scatter plot, on the other hand, is a graphical presentation of the association between two or more quantitative variables, displayed on a different axis. The relationship between the variables is reflected by the pattern of the plotted points in the graphic. Finally, the description of conditional probabilities also helps in finding out and describing the relationship between two variables, since it is a probability distribution that, given two events A and B, can predict the probability of A given B, when B is known to exist (Stinerock, 2018).

3.4.2. Predictive modeling

Predictive modeling is a group of statistical techniques used to predict certain events based on current and historical data. These models exploit patterns in historical data to identify risks and opportunities, through the identification of relationships among many factors and determination of potential associated with a set of conditions (Turban, 2011). The functional and expectable effect of these techniques is that the conclusions withdrawn will have an impact or influence organizational processes that include a large number of individuals. In this study, some predictive techniques will be used, such as linear regression, decision tree, and artificial neural network.

3.4.2.1. Linear regression

The regression analysis is a frequently used methodology for predictions, being the base of predictive analytics, and will be used for both objectives 1 and 2, in order to search for relationships between different variables and sets of variables, by exploiting how can one variable predict the other, or, in statistical terms, how can the explanatory variables, or independent variables, influence and predict the value of the dependent value, which is always only one. The objective is to find a mathematical equation that serves as a model to represent the interactions between two or several variables used as input. There are a variety of regression techniques to be applied to different situations, in this study, only the simple linear regression technique and multiple regression will be used (Weiss & Davidson, 2010; Finlay, 2014).

These techniques can be used to demonstrate a mathematical relationship to prove, or not, the cause-effect between the dependent variable and the explanatory ones, or to obtain a relationship that can predict the value of the dependent variable through future observations of the explanatory variables (Larose, 2005).

Regarding the assumptions related to these techniques, the sample must be representative of the population for the inference prediction, the explanatory variables must be linearly independent and measured with no error, the error needs to have a mean of zero on the explanatory variables and its variance must be constant across different observations (Pestana & Gageiro, 2005).

For the multiple regression case, to estimate the unknown coefficients of a linear regression model, normally the ordinary least squares (OLS) model is used. This technique seeks to find the best adjustments for a set of data by minimizing the sum of the residuals (the difference between an observed value and the value provided by a model) made in the results of every single equation. The OLS estimator is consistent when the explanatory variables are exogenous, *i.e.*, when those variables are fixed in repeated sampling.

Other techniques can be used instead of the OLS, like least absolute deviations, which minimizes the absolute sum of the residual, and the Theil-Sen estimator, which chooses a line whose slope is the median of the slopes determined by pairs of sample points.

The linear regression formula is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Where Y is the dependent variable, x_j are the j explanatory variables, α and β_j are regression coefficients to be estimated by the OLS method. Moreover, the ε is the error, j represents each variable (from 1 to k) and i represents each observation. The estimate regression formula is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$$

Simple linear regression

This technique is used to provide an equation that describes the relationship of one variable Y to another variable X (Stinerock, 2018). The simple linear regression is characterized by having only one explanatory variable, being the other variable the dependent variable, and it finds a

linear function that can predict the dependent variable value as a function of the explanatory variable (Lane, 2013).

The goal is to find the most appropriate fit for the data points. “Best” fit should be understood as in the OLS method approach, i.e, a line that minimizes the sum of squared residuals.

Multiple regression

Contrary to what happens in simple linear regression, the multiple regression technique searches for a relationship between one dependent variable and more than one explanatory variable (Pestana & Gageiro, 2005). While in the simple linear regression the data are modeled in order to build an approximately straight line, in this technique, the goal is to build a hyperplane, since there are more than two variables used as input.

The degree to which two or more predictors (explanatory variables) are related to the dependent variable is expressed in the correlation coefficient R, which can assume values between 0 and 1. To interpret the direction of the relationship between variables, the sign of the regression coefficients (the β values in the equation) is crucial. If the coefficient is positive, then the relationship of this variable with the dependent one is positive, as in, if one grows, the other grows too, and vice versa.

Although the multiple regression technique is very simple do use, since it works through a mathematical equation that expresses the relationship between the dependent and the explanatory variables, it has some disadvantages, such as not being able to be used with nonlinear data, contrary to decision trees and artificial neural networks, analyzed in the chapters below.

3.4.2.2. Decision trees

The decision tree is one of the most used techniques for regression problems, since it is an easy model to understand and it is capable of identifying and classifying complex structures. This technique is suitable for problems with many variables because it is a nonlinear predictive model and offers a clear representation and visualization of the information. The model is compound by tree-shaped structures that represent sets of decisions.

A decision tree consists of branches and nodes (Figure 9), and there are three types of nodes, the decision nodes, the chance nodes and the leaf nodes (nodes without any “child nodes”).

Each branch represents the outcome of a test on an attribute to classify a pattern, and each node represents a decision taken after computing all attributes.

The decision tree is built recursively, and the division of data by the branches depends on the type of the attribute used in such division (Turban, 2011). Usually, the users of these decision trees prefer less complex trees, and this complexity is measured by one of the following indicators: total number of nodes, total number of leaf nodes, depth of the tree or number of attributes used. In short, the building process of a decision tree is based on the following steps:

1. Creating a root node and assign it to the totality of data.
2. Selecting the best division attribute.
3. Add a branch to the root node for each division value. Divide the data into exclusive subsets along the specific division lines.
4. Repeat steps 2 and 3 for every node until reach the objective.

It is necessary to take into account some criteria when building a decision tree, like deciding which is the best node splitting rule based on variables´ values, in order to differentiate observations based on the dependent variable. Once this rule is selected, it is applied to every child node, over and over again, until the process detects no further gain in splitting the child nodes, or some pre-set stopping rule is met. These stopping rules should be defined when building the model, to set a stopping point when the child nodes become leaf nodes (Gama, Carvalho, Faceli, Lorena, & Oliveira, 2012).

When the DT is finished, it is normally large, with many ramifications and in need of improvement due to a big amount of different variables and rules, and so, using the prune technique, some branches are excluded, in order to build a smaller tree, yet more precise when applied to unknown cases (Han & Kamber, 2001). The prune technique can be applied during the construction of the tree, if any of the stopping criteria have been satisfied, or after it is done, when has been already analyzed and learned, which is what happens in most of the algorithms.

The use of decision trees brings some advantages that support the fact that this is one of the most used prediction techniques (Bose & Mahapatra, 2001):

- Flexibility – Decision trees can work with different types of variables, whether they are continuous, nominals and ordinals.
- Variables selection – The building process of a decision tree chooses the most adequate variables to include in the model, preventing the problem of redundant variables.

- Interpretation – The decision nodes are based on the values of the variables used in the problem´s description, and they turn complex problems into simpler problems in each node division.
- Robustness – This is a method that can easily adapt itself to missing data, plus, it is unaffected by variables transformation.

Although there are many advantages supporting the use of this technique, there are also some disadvantages (Bose & Mahapatra, 2001):

- Replication – The algorithm duplicates the sequence of rules and tests in different branches of the same tree, leading to a redundant model.
- Missing values – Without the use of an appropriate algorithm, variables with unknown values can cause problems in subsequent branches.

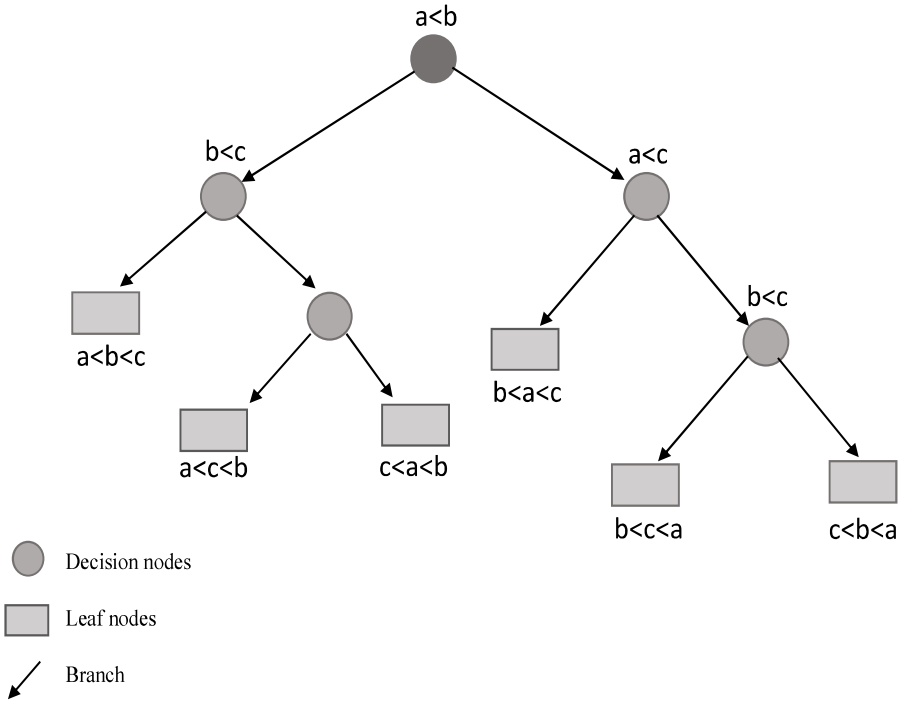


Figure 9. Representation of a decision tree.

There are many different algorithms that can be used when building a decision tree. In this work, the CART – Classification and Regression Tree, will be used. Classification and Regression Trees are a non-parametric decision tree learning technique that builds

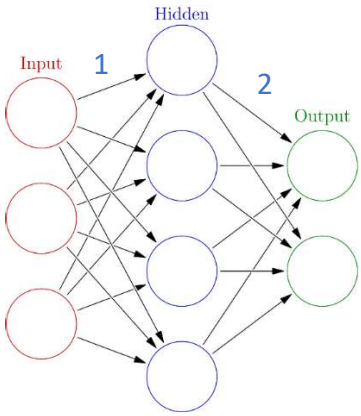
classification or regression trees, according to whether the variables are categorical or numerical, respectively. Presented by Breiman, Friedman, Olshen, and Stone (1984), it gives us easily interpretable binary Trees, in which each node has exactly two child nodes, and assigning to each variable its relative importance. A child node is always more homogeneous than his parent and the division from every node into two child nodes is repeated until it is not possible to divide more and while this division can produce a great accuracy improvement (Li, Sun, & Wu, 2010).

The CART algorithm handles cases of missing values by using surrogate splitting, in order to try to make a prediction even when the split variable is missing and identifying variable importance (Delen, Kuzey, & Uyar, 2013).

3.4.2.3. Artificial neural network

Artificial neural networks (ANN) were created at the beginning of the 50´s with the studies of McCulloch & Pitts (1943) and were originally inspired in trying to simulate some functions of the human brain. They can be used as a classification method or a regression method. In this work, ANN will be evaluated to see if this is a good prediction technique for the occupancy rate. ANN are compound by layers of nodes with different functions in the network: there are an input layer, intermediate layers (or hidden layers) and an output layer, and the connections between layers have different “costs” associated (Figure 10).

The ANN bases its structure and functioning on human neurons, trying to simulate the capacity of the human brain in obtaining knowledge and trying to reproduce a nonlinear type of learning as the human neural networks have (Gama et al., 2012). Currently, ANN have a huge importance in processing data, since it allows the understanding and reliability of the data, and enables the generalization through approximations with high trust, even when the data are totally different from the ones used in the training phase.



1 – Weight between the input layer and the hidden layer

2 – Weight between the hidden layer and the output layer

Figure 10. Representation of an Artificial Neural Network.

This model is composed of artificial neurons, which receive input data and combine this data with their internal state, being this process called the activation. The nodes have an activation function, which is applied to the input values' sum with the cost of the connection, or, in this work´s case, the variables. The result of this sum is passed to the next node layer and this process is repeated for every layer, until it reaches the output layer. The ANN learn incrementally by modifying the costs of the connections between layers, so that in the training phase, the results predicted for the output value will be closer to those observed in this phase, obtaining the less error possible.

The artificial neural networks are very suitable for regression tasks, since the numeric numbers can be passed through the layers of nodes and finally passed to the output layer.

In 1958, Rosenblatt created an algorithm, called the perceptron, which consists of each artificial neuron, or node, calculates the sum of the multiplication between the signals used as input and the costs associated with the connection between the nodes. This weighting is used as input in the activation function, and then the result is passed to the next node, which may or may not process it again (Rosenblatt, 1958). The perceptron networks have the particularity of not having intermediate layers, only the input layer and output layer, which allows an easier understanding of the data and the outputs, although with some limitations regarding linear problems.

The most used neural network is the Multilayer Perceptron (Papert & Minsky, 1969), and it is compound by three essential layers: the input layer, including the selected data to be analyzed, the intermediate layer, or hidden layer, which can be one or more than one layers, where the data processing is made, and finally the output layer, where we can find the results obtained.

Regarding the advantages that this technique offers, learning and generalization are two of those, since the technique provides a description of the whole from small samples. A fast and parallel processing allows complex tasks to be completed in a short period of time, and the adaptability inherent to the ANN allows the model to adapt to environmental changes. The robustness offers a capacity to process incomplete data efficiently and still maintain a high performance when some nodes are disabled (Cortez, 2002; Gama et al., 2012).

Contrary to the decision tree technique, which presents clear graphics, easy to understand and to decode, since it creates decision borders that can be easily translated into decision rules, the ANN show a big lack explanations in its models, regarding the number of layers of the model, which are the activation functions and its relation with the data provided (Graupe, 2007). The use of a big number of input variables may lead to a big delay when processing the data.

The ANN are not very efficient when analyzing unknown data, due to the difficulties in understanding the model operation. One of the solutions for this problem is dividing the sample in training and testing set, so the first one can be used in conceptualizing the model and the second one in detecting the prediction´s efficiency (Gama et al., 2012).

3.5. Evaluation

3.5.1. Quality metrics

In order to choose the statistical method that obtains better results, is necessary to evaluate the prediction capacity of those methods trough some evaluation metrics, depending on the problem. In regression models, the most common metrics are:

- Coefficient of determination (R^2): In statistics, the coefficient of determination is the proportion of the variance in the dependent variable that is predicted from the independent variables. It tells us how well the data fit the model. This technique is mostly used in statistic models, to predict future results or as a hypothesis test. The R^2 varies from 0 to 1, being 1 the perfect adjustment between variables. Although the R^2

varies from 0 to 1, there may be times where this value is negative, this may occur when the data adjustment functions are not linear.

- Mean absolute error (MAE): The mean absolute error is typically used to measure how close the observed results are to the predicted results. This metric is an average of the absolute errors, and so, the lower this metric is, the better for the model (Hyndman & Koehler, 2006). The mean absolute error is given by the following formula:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

  Considering a scatter plot of *n* points, where point *i* has de coordinates (xi, yi), the mean absolute error is the average vertical distance between each point and the identity line.

- Root-mean-square error (RMSE): Is a frequently used metric to measure the differences between values predicted by a model and the values observed through a standard deviation. Those differences are called residuals when the calculations are made on the sample used for the estimate or forecast errors when calculated out of the sample. It only compared the forecast errors of different models for one single variable.

3.5.2. Validation methods

When analyzing a certain dataset, it is essential that the higher accuracy possible is guaranteed, and so, the partition of the original data sample into two different sets is very common, a training set and a testing set. The way the model treats the data is evaluated in the two different samples, to find out about its efficiency and accuracy (Li et al., 2010). The most used techniques to partition the sample are the holdout technique and cross-validation.

In the holdout method, data points from the whole data set are assigned to two subsets, usually called the training set and the testing set. For this work, 70% of the data set were used for training, while the other 30% was used for testing. The random partition can be problematic, since it does not guarantee that the subgroups are all representative. For them to be representative, the data is divided ensuring that every class is represented in every subset (Witten & Frank, 2005). The main characteristic of the holdout method is swapping the training and the testing subsets, *i.e.*, using the test subset for training the model, and the training subset to test the model. The average of the two results reduces the problem of the data not being representative. Although being very useful, the holdout method proposes that the data sample

is divided into two equal parts, and, on the other hand, the model is more efficient if the subset training is bigger than the testing subset (Witten & Frank, 2005).

The process of training and testing can be repeated with different subsets from the data sample. Each time this happens, the data is selected randomly for the subsets. In this procedure, the error rate is the average of the errors from each time this process occurs. This is a branch of the holdout method called the repeated holdout (Witten & Frank, 2005).

In cross-validation, the original data sample is partitioned into *k* equal-sized subsets. From all the subsets, one is selected as the testing subset, and the rest as training subsets. The cross-validation process is repeated *k* times, since every subset will serve as testing exactly once. The *k* results can be averaged to produce a single estimation. The 10-fold cross-validation is the most commonly used, since is said to be the one which obtains the most correct error rate (Witten & Frank, 2005).

With the use of different techniques and a big set of data to make predictions, it is possible the occurrence of overfitting and underfitting.

Overfitting occurs when the model produces an analysis that corresponds too closely to the training set, and, therefore, fails to fit the additional data and to find a predictive rule that can be generalized. This will increase the error of the testing set and decrease the error of the training error, making the distance between the two values quite large (Dietterich, 1995).

On the other hand, Underfitting occurs when some parameters or variables that would appear in a specified model are missing, making the model unable to adequately analyze the structure of the data. It would occur, for example, when trying to fit non-linear data into a linear model. A model with underfitting has a poor predictive performance.

3.5.3. Ensembles

In statistics, ensemble methods are used to obtain a better predictive performance in multiple learning algorithms and techniques. It consists of a finite set of alternative models, offering flexible alternatives within the different algorithms and techniques used to make predictions about a data set (Opitz & Maclin, 1999).

Usually, bagging and boosting are the statistical ensemble methods applied to the training data to vary the models produced by the statistical techniques. These methods are used in order to

make better and more accurate predictions, by having a larger number of predictive models (Quinlan, 2006).

The bagging method creates multiple models and uses a randomly drawn subset of the data sample to train each one, to promote model variance. It is important that the training subsets are different and able to produce different classifiers.

On the other hand, boosting uses all the training set at each repetition of model training, so, this method trains each new model instance to emphasize the instances that the previous model failed to classify or analyze. Bosting offers a great increase in the model´s accuracy.

In this work, the boosting and bagging methods were used in the CART algorithm and in the Artificial Neural Network algorithm, since they were available to be applied to these algorithms in SPSS Modeler software.

3.5.4. Parameterization

Multiple regression

With the main objective of understanding which are the factors or variables that contribute the most to explain the dependent variable, the multiple linear regression (MLR) technique was used. The study of the MLR was performed using the IBM SPSS Modeler. The program verifies and validates the assumptions for this technique, allowing the data entered to be analyzed. In this study, the stepwise regression method was used to build the regression model, since this method selects the most significant variables and create a model excluding the non-significant ones.

In the context of this technique, was introduced a level of significance, related to the probability of rejection of the null hypothesis when the null hypothesis is true.

In the Stepwise method, the choice of the predictive variables is done by an automatic procedure. In each step, a random variable is considered for addition or subtraction from the set of explanatory variables, based on pre-established criteria. The last model built by the stepwise method is usually the one chosen to be analyzed.

The multiple regression model does not accept string variables as input, and so, for the string variables in the explanatory variable set, new dummy variables were created to be used as

input. Dummy variables are numeric, quantitative variables that represent categorical data and can take only two quantitative values, usually used as flags for a certain condition.

Decision tree – CART

Decision trees, using the CART algorithm, were performed in order to understand if any of the variables gathered and prepared in the data preparation phase can explain the occupancy rate of each property, per month, and how strong is the relationship between those variables and the dependent variable occupation_rate. The models were created in IBM SPSS Modeler.

From the initial set of variables gathered and presented in the modeling phase, not all the variables were included in the decision tree model, since this method is used to search for the variables that better predict the occupancy rate of each house. Based on the literature review, some of the variables were not considered as possible predictors, for being out of context in the process of obtaining the occupancy rate and not having any theoretical relationship with the dependent variable.

As said in the previous chapter, the sample was divided into a training set and testing set, in this case, 70% of the sample was assigned to the training subset, while 30% was assigned to the testing subset. The CART algorithm allows to variate the decision tree´s parameters, and so, several decision trees were created, diversifying the parameters between models. This parameterization will prevent the overfitting and underfitting, through a prune technique. The most relevant models are shown in Table 14, as the values of each one´s parameters.

The parameterization of the models was performed by manipulating the following parameters: the minimum number of cases in the parent node; the minimum number of child nodes; the maximum tree depth; the inclusion of any improvement method and what was it (boosting or bagging); the application or nonapplication of the prune model.

Wang and Nicolau (2017) state that price determinants should be divided into categories, according to which segment of the business do they belong to. For this process of parameterization to be more extensive and efficient, the set of variables was punctually divided into two categories, and parametrization was applied to both the resulting subsets of variables, and the whole set without any division, all of this in order to achieve the best model. So, the variables used in these decision trees models were:

Total Set (T): Month_number; Max_Occupancy; Season; Big_event; Typology; For_two; FloorNumber; Neighborhood; Historical_area; Touristic_area; Parking; Elevator; Pred_country_month; Fav_dist_channel; Avg_booking_advance.

Property´s Characteristics set (PC): Max_occupancy; Typology; For_two; FloorNumber; Neighborhood; Month_number; Historical_area; Touristic_area; Parking; Elevator.

Reservation´s Characteristics set (RC): Big_Event; Season; Pred_country_month; Fav_dist_channel; Avg_booking_advance; Month_number

The NumReservations variable was excluded from the data set for this phase, since it was decided that this should not be considered as a predictor for the occupancy rate. Although the results of the model quality were much better when the variable was included, the number of reservations could not be controlled by FLH, and there was nothing that the company could do to improve their occupancy rates based on the current number of reservations per month. It would not be ethic if the company limited the maximum number of nights in order to get a higher number of reservations.

Table 16: Decision trees Models parameterization

		Model A	Model B	Model C	Model D	Model E
Data set used		T	T	RC	PC	PC
Parameters	Min. cases in Parent Node	2	2	2	2	2
	Min. cases in Child Node	1	1	1	1	1
	Depth	10	10	5	10	5
	Improvement method	Boosting	Bagging	-	Bagging	-
	Prune	No	No	Yes	No	No

T: Total set; RC: Reservation´s characteristics; PC: Property´s characteristics

Artificial Neural Networks

As in the parameterization of decision trees, also in the Artificial Neural Networks (ANN), different models were tested, with different parameters, to find out which is the combination of parameters that produces the model with the most significant results. As had already happened with the decision trees, also in this technique the original sample was partitioned into 70% for training subset and 30% for testing subset.

The software used to build these models was the IBM SPSS Modeler, and the Multilayer Perceptron was the model of artificial networks chosen, since it allows for more complex relationships, and for this work, this is the most suitable ANN model. Following the procedures in the parameterization of the DT, also for the ANN the sample was tested as a whole and divided into two subsets, the Reservation's characteristics set and the Property's characteristics set, both these subsets are described in the decision tree's parameterization's subchapter. Regarding the parameters manipulated, the number of hidden layers in the network was tested as one or two, and also the number of neurons in the hidden layers was changed until the best model is found, without any limitation. Finally, ensemble methods were applied to this algorithm, to see if through a better performance, the results were also better. ANN models with boosting (enhanced accuracy) and bagging (enhanced stability) were tested.

After several models with different parameterizations were built, it was concluded that the number of hidden layers and the number of neurons presented better results when set automatically by the Modeler software, regardless of the other parameters.

So, for the whole sample and for each subset, three models were built, one without any ensemble method, one using the boosting method and one using the bagging method. The most relevant models for each set of variables were the ones using the bagging method. In resume, for each set of variables (Total set; Property´s characteristics set; Reservation´s characteristics set), the most relevant models had parameters presented in Table 17.

Table 17: Artificial Neural Network: Parameters for the most relevant models

Parameters	ANN model	Multilayer perceptron
	Number of hidden layers	Automatically computed
	Number of neurons in hidden layer	Automatically computed
	Improvement method	Bagging

4. Results and discussion

4.1. Relation between occupancy rate and average rate per night

To understand if there is any relationship between the value per night charged by the company FeelsLikeHome in their houses (dependent variable), and the occupancy rate of those houses (independent variable), and, if there is, how strong is that relationship, a simple linear regression technique was used. This study was made using IBM SPSS Statistics software. The software verifies and validates the assumptions of this technique.

Two separate analyses were made, one trying to investigate if the occupancy rate of the properties can explain the average rate per night of those properties in the same month, and other analyses to investigate if the occupancy rate of the month before can explain the average rate per night. Basically, this chapter focuses on finding out if FeelsLikeHome has taken any action regarding the value per night charged in their properties, according to the occupancy rate values from the same month or the month before. For this purpose, the dependent variable is avg_rate_night, while the independent variables, separately, are the occupation_rate, and the occ_rate_month_before.

In Table 18 are presented the descriptive statistics of the average rate per night and the occupancy rate. Moreover, in Table 19 are presented the descriptive statistics of the average rate per night and the occupancy rate referring to the previous month.

Table 18: Descriptive statistics of the average rate per night and the occupancy rate

Variables	Mean	Standard Deviation	Minimum	Median	Maximum
Avg_rate_night	99.41	56.99	5.44	89.08	1059.00
Occupation_rate	64.56	23.18	3.23	70.00	100.00

Valid N: 5049 observations

Table 19: Descriptive statistics of the average rate per night and the occ. rate of the prev. month

Variables	Mean	Standard Deviation	Minimum	Median	Maximum
Avg_rate_night	99.91	57.64	5.44	89.38	1059.00
Occ_rate_month_before	61.86	26.03	0.00	67.86	100.00

Valid N: 4876 observations

When a simple linear regression model is made, to see if whether the model is a good predictor, we first take a look at the *p-value* and the *t*. Usually, for $\alpha = 0.05$, this means that if $p < 0.05$, then the correlation is considered to be significant, and the hypothesis null is rejected, and when $p > 0.05$, the null hypothesis is not rejected. The ANOVA test presented as a null hypothesis that the independent variable occupation_rate, does not present any predictive power towards the dependent variable avg_rate_night, and thus, that the correlation between the two is nonexistent. In this case, the *p-value* is 0.099>0.05, therefore, the correlation is not significant, and the relationship between the variables is not confirmed. It is concluded that there is no relationship between the two variables.

When the same test is applied to the occupancy rate of the previous month, the results are a bit different, yet very close to one another. In this case, the *p-value* for the regression between the variables avg_rate_night (dependent variable) and occ_rate_month_before (independent variable) is 0.115>0.05, and so, as it happened in the previous test, the correlation is not significant and is legit to conclude that there is no relation between the occupancy rate of a property in the previous month and the average rate per night implemented in a month.

In order to better understand the possibility of a relationship between the two variables, a Spearman ordinal correlation was also used, with the variables avg_rate_night and occupation_rate as input. Looking at the results (Spearman (5049) = 0.087; $p<0.001$), we can see that the positive correlation is significant, although very weak (about 8.7%). When the same test is applied to the pair of variables avg_rate_night and occ_rate_month_before, the results are similar (Spearman (4876) = 0.085; $p<0.001$). As it happens in the previous model, although the correlation is significant, is very weak, almost nonexistent (about 8.5%).

So, through the compound of results given by these two methods, we verify that there is no association between the avg_rate_night and the occupation_rate variables, and that the occupation_rate does not explain the avg_rate_night.

As we can see in Figure 11, the average rate per night and the occupancy rate have, in fact, a similar evolution over the few months, if we do not take into account the corresponding months. Although this evolution is quite similar, there are some irregular variations that do not allow conclusions to be drawn regarding any possible linear relationship between the two variables. Figure 12 shows the relation between the average rate per night and the occupancy rate of the previous month over the months, and as we can see, there is no linear relation.

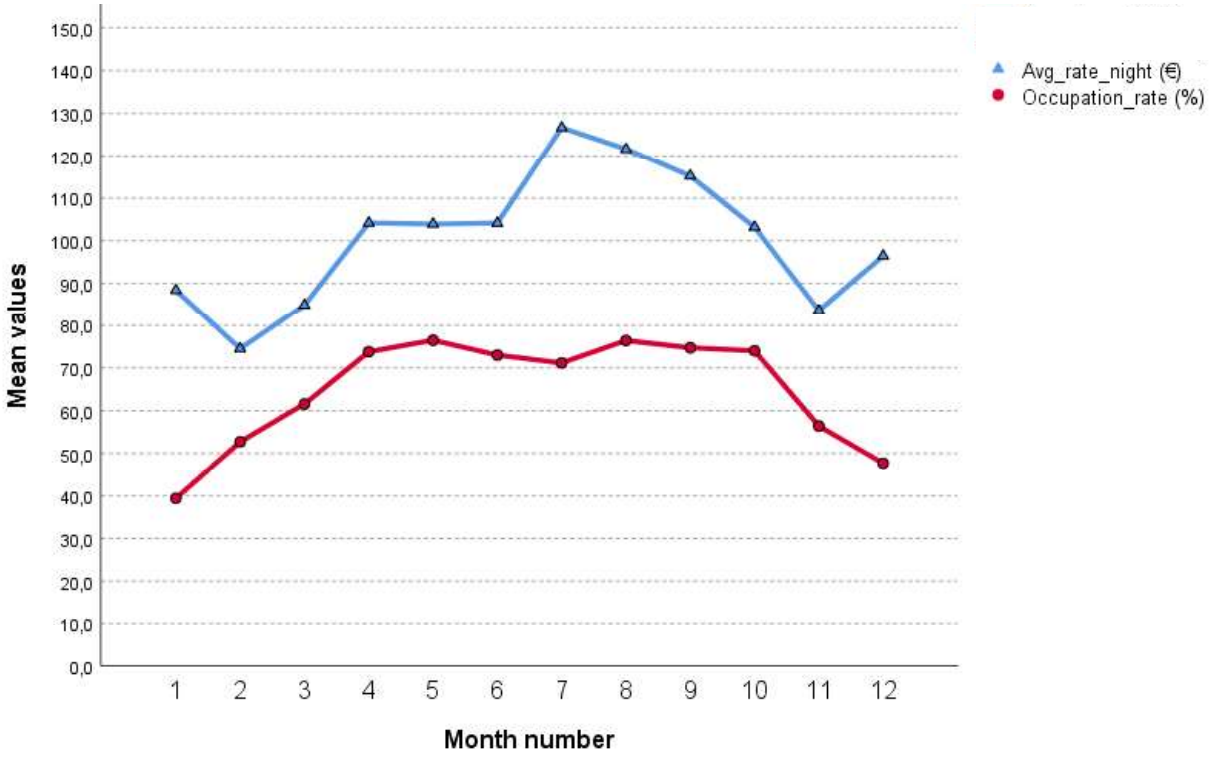


Figure 11. Relation between the average rate per night and the occupancy rate

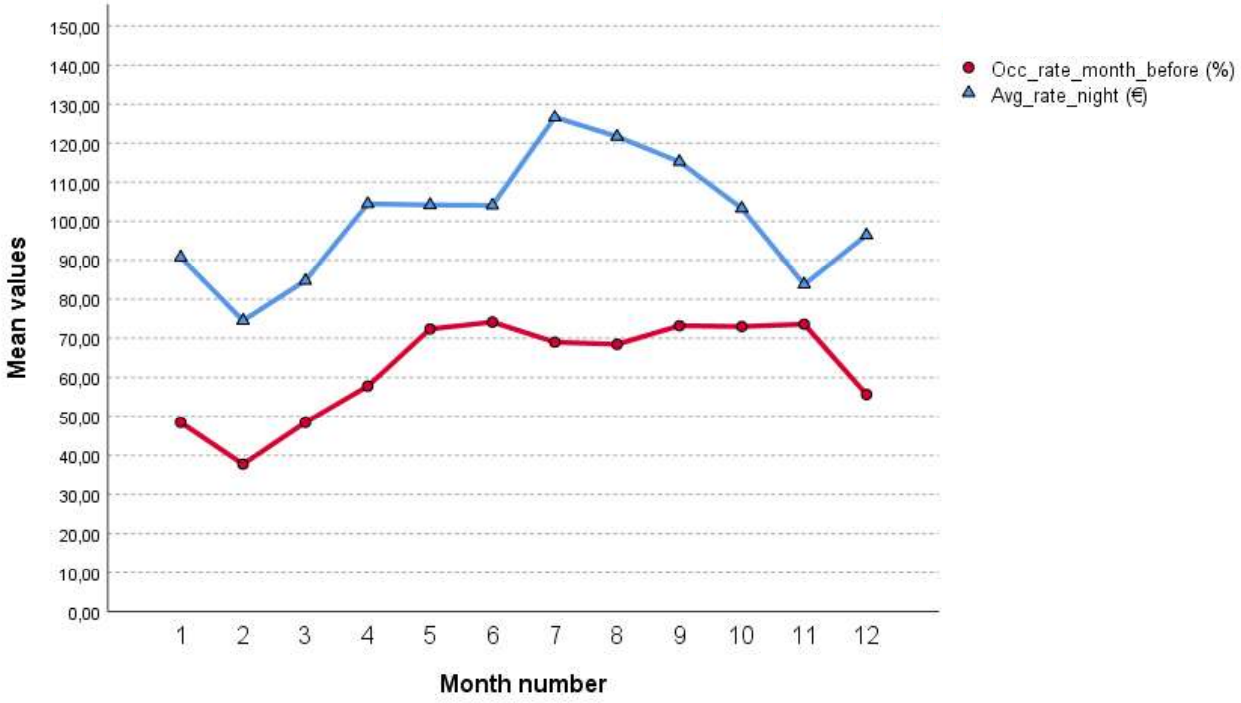


Figure 12. Relation between the average rate per night and the occ. rate prev. month

Alternative analysis

After concluding that the average value per night assigned to each house per month cannot be explained by either the occupancy rate in the same month or the occupancy rate in the previous month, an alternative scenario was taken in account, and more statistical tests were made to understand if the same variables of occupancy rate, considered as a whole product, discarding the individuality of each property, can explain the average rate per night, also without the context of property, through a mean value, just taking in account each month and each year. In the Table 20, we can see the results for the linear regression for the avg_rate_night and occupation_rate, while in Table 21 we can see the results for the same technique but for the pair of variables avg_rate_night and occ_rate_month_before.

Table 20: Relation between the average per night and the occupancy rate in alternative scenario

Dependent variable	Independent variable	Coefficient	Std. Deviation	t Test	Sig	Model Quality
Avg_rate_night	Occupation_rate	0.79	0.18	4.39	0.000	F Test = 19.24; p<0.001 R^2 adjusted = 39.5% Std. Error of Estimate = 0.18

Table 21: Relation between the average per night and the occ. rate in the prev. month in an alternative scenario.

Dependent variable	Independent variable	Coefficient	Std. Deviation	t Test	Sig	Model Quality
Avg_rate_night	Occ_rate_month_before	0.74	0.20	3.77	0.000	F Test = 14.23; p<0.001 R^2 adjusted = 32.9% Std. Error of Estimate. = 0.20

Looking at the results, we can see that the correlation is significant, since in both cases the p-*value* < 0.01. In this case, both the variables related to the occupancy rate and the occupancy rate for the month before can explain the average rate per night in about 39.5% and 32.0%, respectively, which is a better result than we have seen in the previous scenario.

We can conclude that for the overall sample, considering all the statistical tests done, both the variables in question cannot explain the average rate per night. Although, when we consider the occupancy rate as a separate variable, excluding the fact that it belongs to individual properties, there is, in fact, a relation to be pointed. Since this work focuses on obtaining the best rate per night to each property, the results from the alternative scenario are not enough to conclude that the relation is valid.

4.1.1. Discussion

Through the observation of results, some conclusions can be drawn. Literature Review points out that the pricing optimization processes should take into account customer buying habits and market dynamics (Cross et al., 2009), and, in the current work, this is reflected in the occupancy rate of each property. Modica et al. (2009) reinforce the idea that demand-based pricing is very important, where prices are set depending on the current level of demand, and points a linear relationship between the two variables, if one decreases, the other increases. Pricing systems should react instantly to every factor that changes over time (Lieberman, 2010), and in this case, as pointed by Cross et al. (2009; 2011), when demand is low, it is recommended that lowest rates programs are opened, and, consequently, when demand is high (corresponding to a high occupancy rate), the highest rates programs should be opened.

In the first instance, it could be concluded that FeelsLikeHome does not follow a process of rate management based on the occupancy rate of each property, *i.e.*, does not set their rates based on the information provided by the occupancy rate, contrary to what Ng (2007) pointed out (targeted pricing and demand should be related), and instead, focuses their rate management

process in other explanatory variables. Another possible explanation for the results is that the company does not set their rates only based on the observation and analysis of the occupancy rate and that the rates are not abruptly and unconsciously changed. Other factors are probably analyzed with the intent of finding out if the change in the rate per night is profitable.

These results were discussed with an FLH specialist, who tried to explain some of the results obtained.

At first, a hypothesis arose for non-relation of the occupancy rate to the average rate per night implemented. That hypothesis was based on the possibility that the values of the rate per night and the occupancy rate could be defined in different moments, being impossible to set a relationship between the two variables, by observing the historical data provided. However, that hypothesis was discarded, since the values of the rates per night used to calculate the average rate per night, were the values paid at the moment when each reservation was made for a certain month, which corresponded to a momentary occupancy rate for that same month. The value of the rate per night changed with the approximation to the beginning of the month, as did the occupancy rate, so, this relation could be observed even if before the month began, if the price paid per night had accompanied the changes or lack of changes of the occupancy rate.

It was pointed, by the specialist, that FeelsLikeHome changed their rate manager twice through the established data time between January 2017 and May 2019. In the first year (2017), the rate manager in charge was not a specialist in the area, so the process was not consistent, and, in some periods, there was not a systematic verification of the suitability of the rates per night in every property. In the next year, for a short period of time, a new rate manager took place, and so, there were some changes in the process, until finally the current rate manager took action and started applying some profitable processes.

All these events could explain the existent relation between the occupancy rate and the average rate per night when considering the occupancy rate as a whole variable, independent from the property it belongs to, contrary to what happens when checking the data using a property point of view, since this existent relation can point out that in the majority of the houses, in several months, there was, in fact, a concern when defining the rates per night, despite not being a consistent and constantly verified process.

In conclusion, the results contradict the literature review regarding the necessity of taking into account the occupancy rate when defining a profitable pricing system (Ng, 2007). However,

this can be explained by the instability of rating processes lived by the company, and the lack of commitment in one single rating process.

4.2. Predictors of occupancy rate

To achieve the objective of understanding which are the variables that can better explain a property´s occupancy rate, a set of independent variables was gathered and used as input in several statistical techniques and models. Those variables were retrieved from historical data provided by FeelsLikeHome and prepared for the modeling process. In order to describe them with a correct context, the descriptive statistics of the independent variables were separated into three different Tables (Appendix A, B, and C).

For this purpose, three different algorithms were chosen (MLR, DT, and ANN) and applied to the three different variable subsets defined in the parameterization chapter (Total set, Reservations' characteristics set, and Properties' characteristics set) (Stream from IBM SPSS Modeler in Appendix D). Several models were built and tested for each algorithm, and the parameters for the most relevant models in each algorithm are presented in the parameterization chapter (Table 16 and Table 17).

In conjunction with the FLH specialist, it was discussed which was the most appropriate data subset to use to find the predictors of the occupancy rate, from a business perspective. At first, it was pointed out that the division of the initial data set including all variables into two was very convenient, since it allowed a deeper study of two different market segments of the company. If only the variables regarding the reservations were used, that could define some customer profiles that tend to book more houses or even pay higher prices, and some marketing decisions could be made to attract those profiles of customers in the right conditions. On the other hand, if the properties' characteristics variables were chosen, the houses that guaranteed a higher occupancy rate could be more advertised. However, the data set chosen was the total set, with all the variables together, since it offers an overall view of the important variables for the business, and it makes it easier for the company to make important decisions of what to invest more in. In conclusion, both the subsets for the reservations' characteristics and properties' characteristics were considered, although it was concluded that the best one for the context was the total data set, which corresponds to model B in the decision trees parameterization (Table 16) and to the artificial neural network that uses the total set. As

previously pointed in the validation methods chapter, the original sample was divided into two subsets, 70% for training and 30% for testing.

The evaluation of the models was made using quality metrics described in the modeling chapter, namely the R^2, the MAE and the RMSE (Witten & Frank, 2005). For each algorithm, the best model for the total data set was evaluated and the results were compared. However, the choice for the best model took into account not only the values provided by this quality metrics, but also the convenience of the chosen model for the business, in this case, the company FeelsLikeHome. In Table 22 are presented the results of the evaluation for each algorithm.

Table 22: Evaluation of each algorithm used

Algorithm	R^2		MAE		RMSE	
	Training	Testing	Training	Testing	Training	Testing
MLR	0.591	0.609	21.166	19.907	25.794	24.552
DT	0.702	0.695	17.297	17.471	22.799	22.570
ANN	0.830	0.701	14.144	16.956	18.743	21.999

The R^2 is probably the best indicator for comparing the models of regression, this coefficient shows the distribution of the predicted values and observed through a line or a curve. Its value oscillates between $-\infty$ and 1, and the closer to 1 the value is, the smaller is the difference between the predicted and observed values. The mean absolute error is an average of the individual errors, with no regard to the signal of the values, the smaller the MAE, the better the model. Finally, the RMSE is a value related to the error in the model, and so, the smaller this value, the better.

By looking at Table 22, some conclusions can be taken regarding the most adequate model to use. Statistically, the R^2 value for the testing subset is higher for the artificial neural networks, however, the decision trees present the closest results between training and testing subsets, and still with values around 0.70 , so the difference between the coefficient of determination for the decision trees and artificial neural networks does not compensate the usage of the ANN algorithm. In addition, it is more advantageous for the business that the model chosen is the decision tree, since it facilitates the definition of profiles based on the explanatory variables that present a higher or lower occupancy rate.

At first, a single model of decision tree was generated, without the addition of a bagging ensemble, in order to be possible, the visual observation of the DT and its nodes. The quality of this model in the test sample is approximately equal to the same model with bagging

incorporated (R^2 = 0.668). The decision tree including all the variables as input allows withdrawing some profiles with corresponding occupancy rates. The occupancy rate was segmented into three subgroups according to the FLH specialist's suggestion, to its value: Low occupancy rate (less than 30%); Medium occupancy rate (between 30% and 65%) and High occupancy rate (greater than 65%). Below, some relevant nodes were described:

1. Profiles for low occupancy rates:
    - Node 8: if the favorite distribution channel in which the reservation is made is indifferent (can be any channel) and the booking advance is less or equal than 5.125 days and the property's maximum occupancy is more than 5.5 people, then the predicted occupancy rate is 10.194% (n=118).
    - Node 31: if the favorite distribution channel in which the reservation is made is indifferent and the booking advance is less or equal than 5.125 days and the property's maximum occupancy is less or equal than 5.5 people and the customer's nationality is Brazil, France, Portugal or Spain and the property's typology is Studio, T0, T1, T1 Duplex or T2 Duplex, then the predicted occupancy rate is 16.194% (n=223).
    - Node 61: if the favorite distribution channel in which the reservation is made is indifferent and the booking advance is less or equal than 5.125 days and the property's maximum occupancy is less or equal than 5.5 people and the customer's nationality is Brazil or Portugal and the property's typology is T2 or T3, then the predicted occupancy rate is 11.371% (n=37).

2. Profile for medium occupancy rate:
    - Node 24: if the favorite distribution channel in which the reservation is made is Airbnb, Booking, FeelsLikeHome or any other in which the company list their houses in, and the Month number is less or equal than 3.5 and the booking advance is less or equal than 26.850 days and the month number is bigger than 1.5, then the predicted occupancy rate is 52.922% (n=163).

3. Profiles for high occupancy rate:
    - Node 26: if the favorite distribution channel in which the reservation is made is Airbnb, Booking, FeelsLikeHome or any other in which the company list their houses in, and the month number is less or equal than 3.5 and the booking advance is more than 26.85 days and the month number is bigger than 2.5, the predicted occupancy rate is 69.694% (n=76).
    - Node 28: if the favorite distribution channel in which the reservation is made is Airbnb, Booking, FeelsLikeHome or any other in which the company list their houses in, and the month number is less or equal than 3.5 and the month number is less than 10.5 and the booking advance is more than 15.2 days, the predicted occupancy rate is 77.629% (n=657).
    - Node 81: if the favorite distribution channel in which the reservation is made is indifferent and the booking advance is more than 5.125 days and the month number is bigger than 2.5 and the month number is less or equal than 10.5 and the typology is Studio, T0, T1, T1 Duplex, T2, T3 Duplex or T4, and the booking advance is more than 24.367 days and the month number is bigger than 3,5 and the typology is T0 or T1 Duplex, the predicted occupation rate is 81.702% (n=34).

After the DT without the bagging ensemble was analyzed, for robustness purposes, it was generated a tree where the target was the predicted occupation rate by the decision tree model with the bagging incorporated (robustness model). This tree helps to understand how the bagging model predicts the occupation rate. The tree structure is similar to the single tree built and used to identify the profiles, and so, the model is not presented. However, the quality of this tree is very good as it predicts the predicted occupancy rate with very few errors (MAE in test sample = 1.488; RMSE in test sample = 2.514).

Regarding the identification of the most important explanatory variables in predicting the occupancy rate, Figure 13 shows the relative importance of those variables, in the bagging model (best predictive model). This result is confluent with the ones obtained from the other two trees built (simple tree model and robustness model), as shown in Figures 14 and 15.

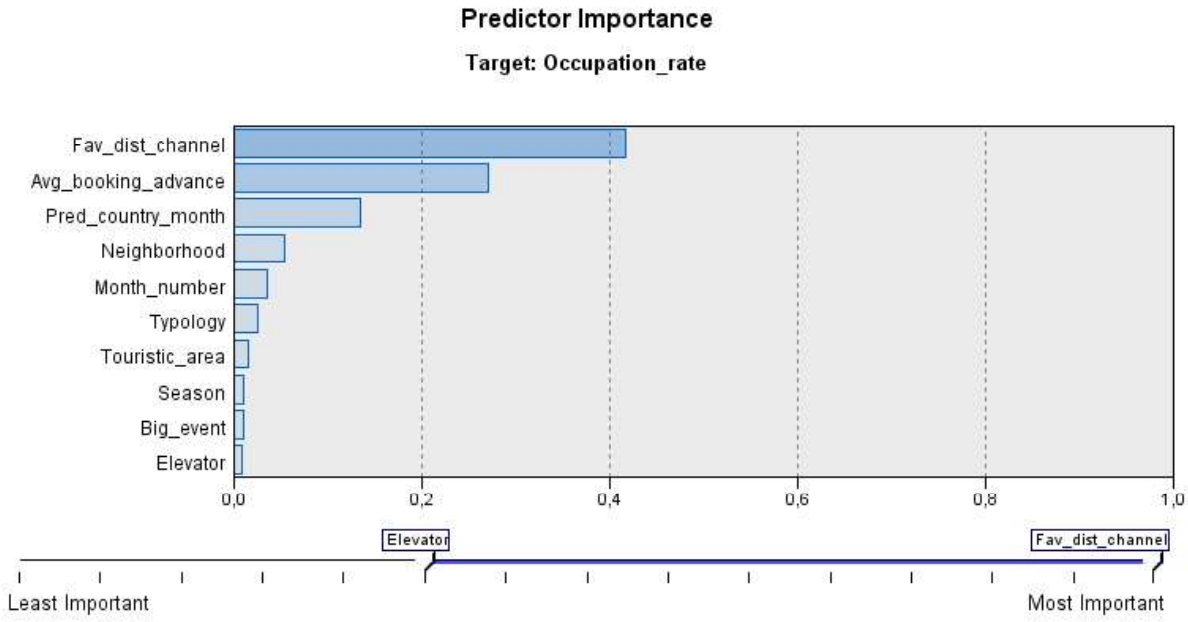


Figure 13. Predictor importance for the decision tree model with bagging

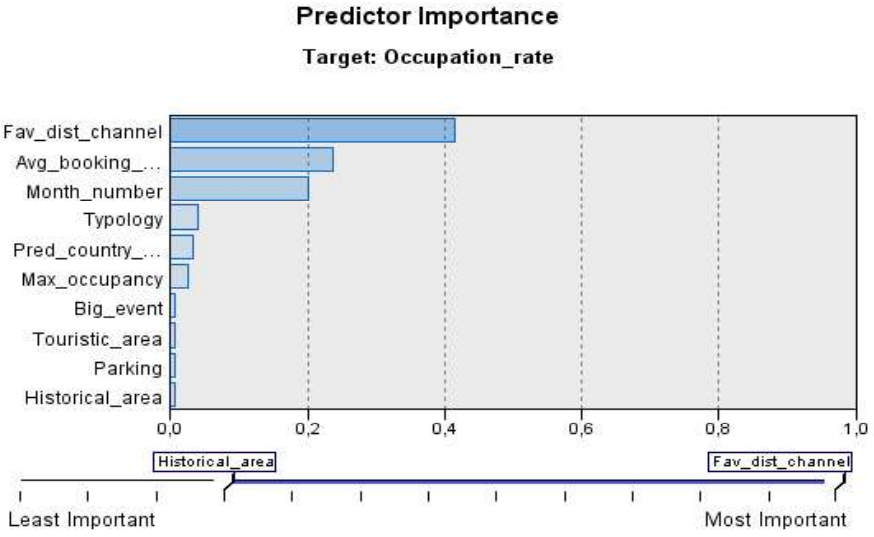


Figure 14. Predictor importance for the simple decision tree built

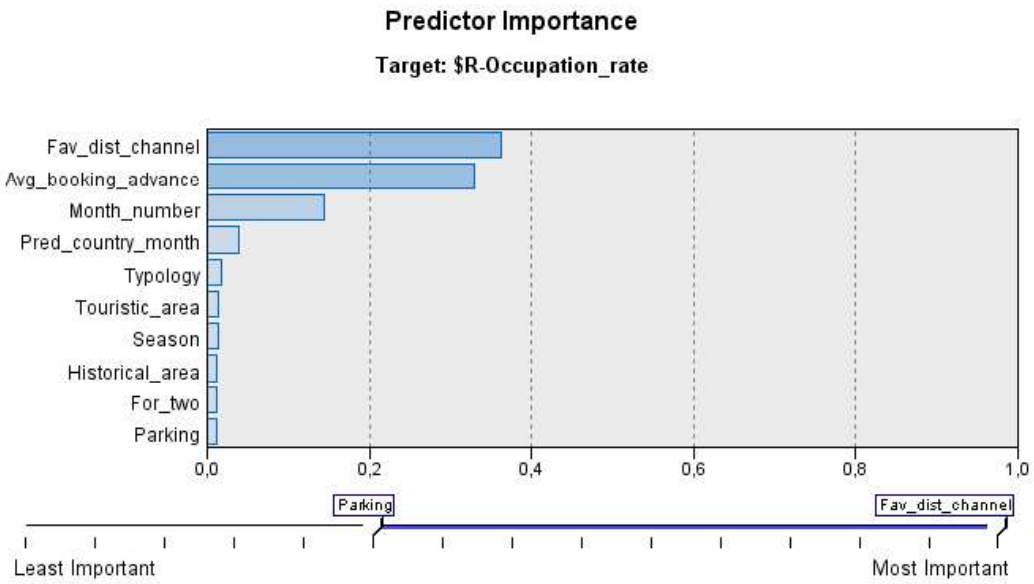


Figure 15. Predictor importance for the robustness tree model

4.2.1. Discussion

Through the observation of Figures 13 and 14, it can be concluded that the distribution channel where the reservation is made is the best predictor of the occupation rate, however, the decision tree shows that the main division to this variable is between being indifferent which is the distribution channel used, or the favorite being one of the big distribution channels where FeelsLikeHome list their houses. Also, the booking advance is a good predictor, and in general bigger booking advances tendentially lead to a higher occupancy rate (Guo et al., 2013), while when the booking is made very close to the check-in date, the occupancy rates tend to be lower. The Nationality of the customers and the month of the check-in are also important predictors.

Several studies support the idea that the main scope of revenue management or rate management processes should be the customer. Through the analysis of the customer's behavior, the rate management processes can be optimized and profitable (Ng, 2009), and Cross et al. (2011) add that the main point of those processes is analyzing and understanding the value that customers give to each.

By observing and analyzing the results provided by the decision tree, they coincide in some points with the information gathered and presented in the Literature review chapter, namely regarding the importance of the distribution channel used to make the reservation, and the booking advance time (Ivanov, 2014).

Doing a parallelism with the hospitality industry, Modica et al. (2009) reinforces the idea that historical data can be very useful to predict future demand, and several other variables are said to influence this demand, like the day of the week when the reservation is made, the season (whether is high or low season), the weather and if there are any big events nearby (Cross et al., 2009; Modica et al., 2009; Guo et al., 2013). Still in the hospitality industry, it is said that by giving a book in advance option with a big-time range, hotels tend to have higher occupancy rates when the booking advance is big. Besides this, the most used indicators to manage the profit are the type of room, duration of the stay, the booking advance, distribution channel and group size (Ivanov, 2014). All these indicators can be contextualized to the short-term rentals industry. So, by looking at the overall results, we see that some of the most important predictors are coherent with the theory, like the distribution channel or the average booking advance, while others do not show up in the model as important, even though the literature review points them as important predictors, like the existence of a big event or the season (if it is high or low). Theoretically speaking, the neighborhood, or whether the property is located in a historical or touristic are, would be an important predictor, since the tourists tend to prefer staying in historical or touristic areas when they travel. However, the city of Lisbon has been the victim of a huge increase in the tourism industry, which makes the tourists be indifferent to where they stay, since they have a place to stay.

4.3. Matrices to propose changes to the currently implemented rates

Based on the results provided by the model where the bagging method was incorporated, several matrices were built, in order to propose changes in the rate per night. For each occupancy rate level, a coefficient is suggested to apply to the currently implemented rate as follows:

- 0.8 of the current rate if the predicted occupancy rate is less than 30% (Low occupancy)
- 1.0, *i.e.*, the same rate, if the predicted occupancy rate is between 30% and 65% (Medium occupancy)
- 1.2, *i.e.*, a 20% increase in the current rate if the predicted occupancy is greater than 65% (High occupancy)

Several matrices were built, each one combining two of the most important predictors, and showing the predicted occupancy rate. Figure 16 shows the occupancy rate by typology and month number, while Figure 17 shows the same matrix, but for each distribution channel

category. Finally, Figure 18 shows the occupancy rate by typology and month number, considering different levels of booking advance. Three levels were created in agreement with the FLH specialist: Small advance (less or equal than 7 days); Medium advance (between 8 and 30 days); Long advance (more than 30 days). It is noteworthy that the best predictors chosen, allow to estimate the occupancy rate with RMSE around 11%.

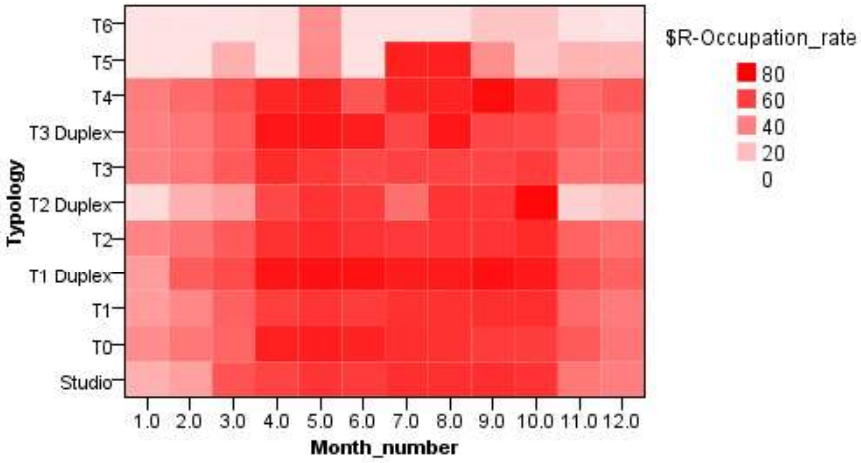


Figure 16. Matrix for the occupancy rate per typology and month number

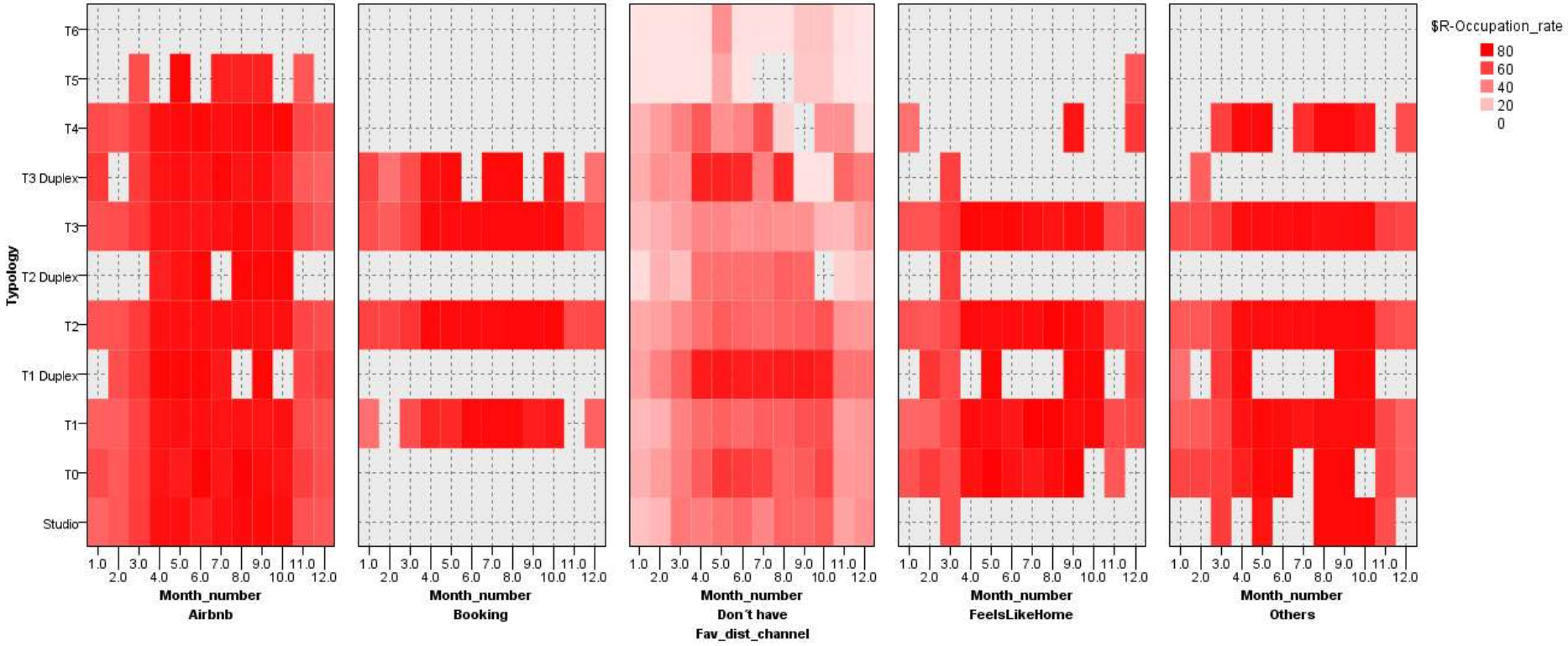


Figure 17. Matrix for the occupancy rate per typology and month number, for each dist. channel

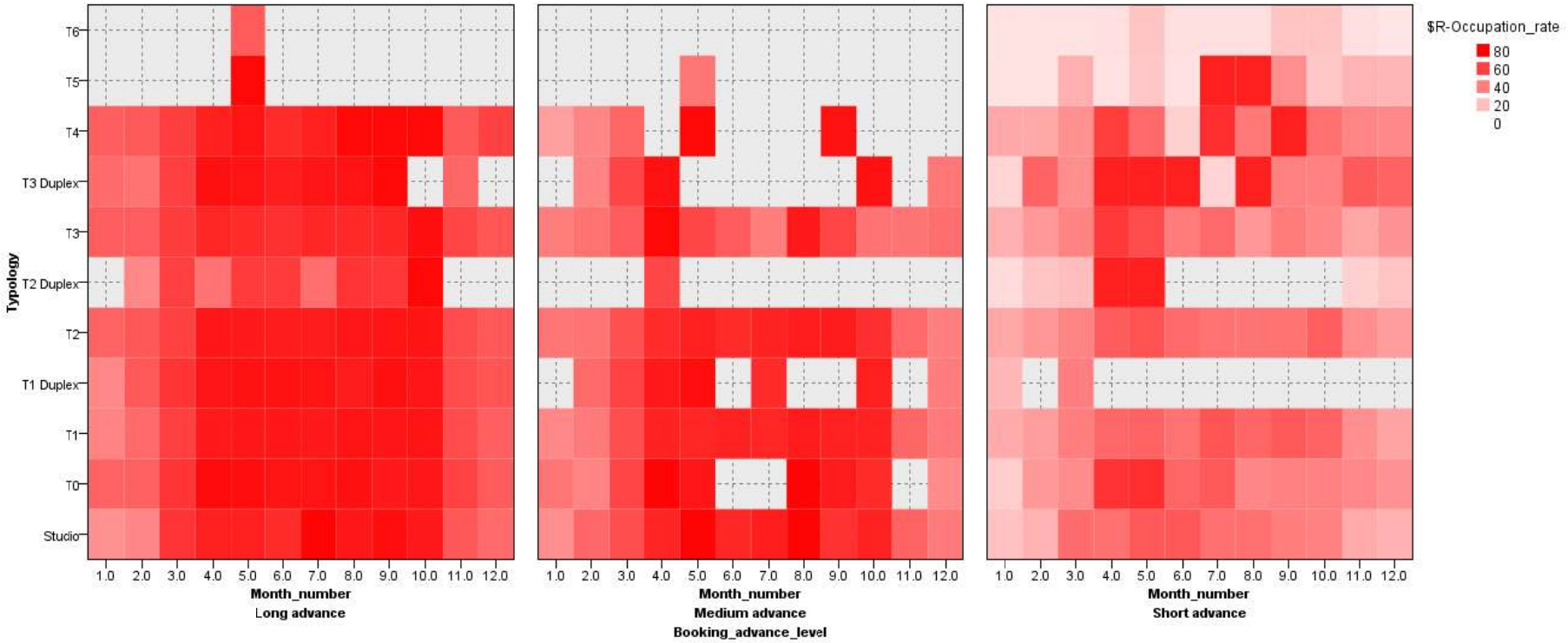


Figure 18. Matrix for the occupancy rate per typology and month number, for each dist. booking advance catg.

4.3.1. Discussion

The matrices built, give a notion to whether and where it is recommended to increase or decrease the rates, based on the more important variables, according to the predictive models. The coefficients proposed are not accurate for the best occupancy rate possible and can be interpreted and changed by the company accordingly to their wills and perspectives.

For the values of low occupancy, the coefficient of 0.8 to apply to the current rates implemented can be even lower and can come down to 0.5 if the occupancy rate is close to 0%. Since the occupancy rate is divided by levels, and the interval of values for low occupancy can go from 0% to 30%, also the coefficient of rate change can variate. So, if the occupation is still very low with the decrease of 0.8, this value can be altered for less.

In the case of medium occupancy, the value of the coefficient recommended was 1.0, *i.e.*, without any change to the current rate implemented. However, this value can vary between 0.9 and 1.1, to test how the occupancy rate follows these changes. If the occupancy rate responds positively, this coefficient can be implemented definitely.

Finally, for the high occupancy rate cases, it is risky to change the current rates implemented, because they can affect negatively the occupancy rate, and so, a small value of increase was suggested (about 20% above the current rate implemented). This value can be higher if the occupancy rate stays higher and does not respond negatively to this change. Some tests should be done before making a final decision on how much to raise the current implemented rate.

Although the theoretical objective is to raise the occupancy rates, the company's objective is to increase profits, and that is why this model with general coefficient recommendations was built, so it is easy for the company to interpret, and change the coefficients at their own will.

(This page was intentionally left blank)

5. Conclusion

5.1. Summary

This work was made as a study case of FeelsLikeHome, a property management company focused on the tourism market, more specifically, the short-term renting segment. The main objectives defined were: 1. To study the relationship between the rate per night of short term second home rentals and the corresponding occupancy rate, and whether the occupancy rate is a predictor of the rate per night; 2. Identify the explanatory variables for the occupancy rate and identify property's profiles with a higher or lower occupancy rate; 3. Create a set of matrices to propose changes to the currently implemented rates.

These objectives were looked at with a data mining perspective, more specifically, as a regression problem. CRISP-DM was the methodology used, since it is business-oriented.

There was a gathering and preparation of data from the historical records provided by FLH, where some variables were excluded, and several others were created, in order to help the purpose. Then, for the first objective to be accomplished, a simple linear regression was performed, to understand if the occupancy rate was an explanatory variable of the rate per night. To better understand the possibility of a correlation between the two variables, a spearman ordinal correlation was also performed. The results showed that for the occupancy rate did not explain the rate per night, *i.e.*, it was not a good predictor, contrary to what the literature review said. Although the results were not as expected, the first objective was accomplished, since through a conversation with an FLH specialist, reasons were found f

After that, based on the literature review, several variables were selected from the total sample of variables provided by FLH, including property's characteristics and reservation's characteristics, and used as input to understand the most important explanatory variables for the occupancy rate, *i.e.*, the best predictors. A predictive model was created, and three different techniques were performed and evaluated, in order to select the one which offered better results regarding the most important explanatory variables. Those techniques were: multiple linear regression, CART decision tree, and artificial neural network. Through the quality metrics chosen (R^2, MAE, and RMSE), it was concluded that the CART decision tree was the best technique. After the most efficient CART model was found, it was performed, and by the

observation of results, it was possible to identify property's profiles with a higher or lower occupancy rate, as well as the most important predictors of the occupancy rate.

Finally, several heatmaps showing the occupancy rate were built using the most explanatory variables, and three levels of occupation were defined: lower occupancy (below 30%), medium occupancy (from 30% to 65%) and high occupancy (above 65%). To each one of those levels of occupancy, a coefficient to be applied over the currently implemented rate was proposed. That coefficient can be altered depending on the company's intentions. In this way, all the objectives were accomplished.

5.2. Recommendations for FLH

Through the observation of the identified properties' profiles and the matrices built, some recommendations can be made to FLH in order to possibly increase their profits:

- Lower the rates prices for properties with capacity for six people, when the booking advance is a week or less, since in these cases, the occupancy rates tend to be very low.
- T5 and T6 properties have a very low predicted occupancy rate, or even null in some cases. It is proposed the creation of promotions for these properties, in the winter months (January, February, November, and December), when the occupancy rate is very low.
- Between the months of May and October, the T1 and T2 properties tend to be almost fully booked, since its summer and Lisbon has become a main attraction in the past few years. An increase in the rates per night could be a good way to make the most advantage of this, since there will always be people willing to pay to stay in Lisbon.
- Airbnb offers higher and most consistent occupancy rates. The investment in this growing platform could be a good way to gain some visibility for the properties and for the company. Also, some agreements with the platform regarding the rates, as advertisement could be good ways to increase profit.

More generally, it is recommended that FeelsLikeHome adopt a consistent and profitable rate management system, preferably basing their rates per night in the historical occupancy rates. All records of historical reservations' and properties' data are imperial to find the best predictors of the occupancy rates.

5.3. Contributions

The short-term holiday rentals is a growing industry, with great potential, but with a big lack of studies regarding the optimization of its processes. With a well-designed rate management system, this industry could be much more profitable than it is. There is a need to study the contextualization and behavior of every "actor" of this industry, guests, and hosts, in order to gather enough information to build a useful rate management system.

By doing an extensive literature review, the importance of the rate management system was evidenced, as its application in the short-term holiday rental industry. The hospitality and the airline industry served as comparison bases, and the characteristics that influenced the rate per night and the occupancy rate were pointed out in these two industries. This information provided some notions about how the second homeowners should look at the business if they want to rent their houses in a short-term condition.

This work intended to take the first steps in understanding the dynamics and possibilities of this industry. By finding out which are the most important predictors of the occupancy rate, as well as identify some properties' profiles with a predicted high or low occupancy rate, important information was provided to the investigators and professionals.

Through many different techniques of Data Mining, patterns were found in historical data. For the professionals of the area, this shows the utility that DM techniques can have in finding new information about the business that the eye cannot see, and improve and change some habits and processes for better economical results.

For FeelsLikeHome, the results achieved can be a starting point to a well-designed and profitable rate management system, and some further analysis can be made using the same techniques as presented in this wok. The same analysis procedures used in this work can be used by the company to analyze another dataset and discover some new useful conclusions.

5.4. Limitations

In this work, some limitations were found, in the process and in the results:

- The data provided by FLH had some inconsistencies in the values, as some variables that could have been important but had a large amount of missing data.

- This work was made only for properties located in Lisbon, and so, the results cannot be representative of other cities.
- Although the sample gathered consisted of 333 properties, and a time interval of two and a half years, this is still a small sample for the results to be generalized
- Several possible predictors of the occupancy rate suggested in the literature review were not used, since the data gathered did not have information about them.

5.5. Further research

Although this work provided some useful results, there is still a long way ahead regarding the understanding of rate management and possible improvements. Future research may be:

- An improvement in the procedures, with the inclusion of new DM predictive techniques and other parameters to enrich the process.
- Perform the same analysis in a bigger sample, with more variables, as the weather, the guests' age and the online rating of each property, to build a model that can be generalized.
- With more information about the business model of the company, propose a matrix with specific rates per night that guarantee a maximum occupancy rate, and thus, the best profit.

References

Airbnb (2018). Airbnb fact sheet. Retrieved from http://assets.airbnb. com/press/press-releases/Airbnb%20Fact%20Sheet_en.pdf

Airbnb (2018). About us section. Retrieved from https://www.airbnb.pt/b/setup

Alderighi, M., Nicolini, M., & Piga, C. A. (2015). Combined effects of a capacity and time on fares : Insights from the yield management of a low-cost airlines, *Review of Economics and Statistics, 97*(4), 900–915. doi: 10.1162/REST

Altin, M., & Schwartz, Z. (2017). "Where you do it " matters : The impact of hotels ' revenue-management implementation strategies on performance. International Journal of Hospitality Management, *67*, 46–52. doi: 10.1016/j.ijhm.2017.08.001

Balck, B., & Cracau, D. (2015). Empirical analysis of customer motives in the shareconomy: A cross-sectoral comparison. Working Paper.

Bieger, T., Beritelli, P., & Weinert, R. (2007). Understanding second home owners who do not rent: Insights on the proprietors of self-catered accommodation. *International Journal of Hospitality Management*, *26*(2), 263–276. doi: 10.1016/j.ijhm.2006.10.011

Bojanic, D. C. (1996). Consumer perceptions of price, value and satisfaction in the hotel industry: An exploratory study. *Journal of Hospitality and Leisure Marketing, 4* (1), 5–22.

Bose, I., & Mahapatra, R. K. (2001). Business data mining — A machine learning perspective. *Information & Management, 39*(3): 211–225. doi: 10.1016/S0378-7206(01)00091-X

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Câmara Municipal de Lisboa (n.d.). Zona centro histórico. Retrieved from http://www.cm-lisboa.pt/visitar/lazer-entretenimento/frente-ribeirinha/depois/zona-centro-historico

Chapman, P. & Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide*. USA: SPSS Inc.

Cortez, P. (2002). *Previsão de Séries Temporais*. Braga: Universidade do Minho - Escola de Engenharia.

Cross, R. G., Higbie, J. A., & Cross, D. Q. (2009). Revenue management's renaissance: A rebirth of the art and science of profitable revenue generation. *Cornell Hospitality Quarterly*, *50*(1), 56–81. doi: 10.1177/1938965508328716

Cross, R. G., Higbie, J. A., & Cross, Z. N. (2011). Milestones in the application of analytical pricing and revenue management. *Journal of Revenue and Pricing Management*, *10*(1), 8–18. doi: 10.1057/rpm.2010.39

Delen, D., Kuzey, C. & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications 40* (10), 3970-3983. doi: 10.1016/j.eswa.2013.01.012

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27* (3), 326-327. doi: 10.1145/212094.212114

Eneckere, R., & Peck, J. (2012). Dynamic competition with random demand and costless search: A theory of price posting. *Econometrica*, *80*(3), 1185–1247. doi: 10.3982/ECTA8806

Escobari, D., & Gan, L. (2007). Price dispersion under costly capacity and demand uncertainty. *National Bureau of Economic Research Working Paper Series*, *13075*, 1–38. doi: 10.3386/w13075

FeelsLikeHome(2018). About us. Retrieved from http://rentals.feelslikehome.pt/about-us

Finlay, S. (2014). *Predictive analytics, data mining and big data: Myths, misconceptions and methods*. Basingstoke: Palgrave Macmillan.

Gama, J., Carvalho, A., Faceli, K., Lorena, A., & Oliveira, M. (2012). *Extracção de conhecimento de dados - Data mining*. Lisboa: E. Sílabo.

Gant, A. C. (2016). Holiday rentals: The new gentrification battlefront. *Sociological Research Online*, *21*(3), 1–9. doi: 10.5153/sro.4071

Graupe, D. (2007). *Principal of artificial neural netwroks*. Chicago: Word Scientific.

Guillet, B. D., & Mohammed, I. (2015). Revenue management research in hospitality and tourism. *International Journal of Contemporary Hospitality Management*, *27*(4), 526–560. doi: 10.1108/IJCHM-06-2014-0295

Guo, X., Ling, L., Yang, C., Li, Z., & Liang, L. (2013). Optimal pricing strategy based on

market segmentation for service products using online reservation systems : An application to hotel rooms. *International Journal of Hospitality Management*, *35*, 274–281. doi: 10.1016/j.ijhm.2013.07.001

Gurran, N., & Phibbs, P. (2017). When tourists move in: How should urban planners respond to Airbnb? *Journal of the American Planning Association*, *83*(1), 80–92. doi: 10.1080/01944363.2016.1249011

Gutt, D. & Herrmann, P. (2015). Sharing means caring? Hosts' price reaction to rating visibility. *ECIS Research-in-Progress Papers, 54* 1-13.

Guttentag, D. (2013). Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, *18*(12), 1192–1217. doi: 10.1080/13683500.2013.827159

Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco: M. K. Publishers.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, Mass.: MIT Press.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688. doi: 10.1016/j.ijforecast.2006.03.001

IBM (2018). CRISP-DM Help Overview. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm

INE (2018). Tourism activity. *Destaque Press Release*, 1-8.

Ivanov, S. (2014). *Hotel revenue management: From theory to practice*. Varna: Zangador.

Ivanov, S., & Zhechev, V. (2012). Hotel revenue management – a critical literature review-working paper. *Turizam: Znanstveno-Stručni Časopis*, *60*(1989), 1–35. doi: 10.2139/ssrn.1977467

Karlsson, L., & Dolnicar, S. (2016). Someone's been sleeping in my bed. *Annals of Tourism Research*, *58*, 159–162. doi: 10.1016/j.annals.2016.02.006

Khadem, N. (2016). Noise, nudity, foul language: Airbnb hosts should be fined, says report.

Sydney Morning Herald. Retrieved from http:// www.smh.com.au/business/the-economy/noise-nudity-foul-language- airbnb-hosts-should-be-fined-says-report-20160412-go4em1.html

Khan, M. A. (2014). A broad view of prospects of tourism industry with reference to India. *The Journal of Management Awareness, 17* (2), 56-63. doi: 10.5958/0974-0945.2014.00005.3.

Lamberton, C. P., & Rose, R. L. (2011). When is ours better than mine? A framework for understanding and altering participation in commercial sharing systems. *Journal of Marketing*, *76*(4), 109–125. doi: 10.2139/ssrn.1939289

Lane, D. M. (2013). *Introduction to statistics*: *An interactive eBook*. University of Houston.

Larose, D. T. (2005). *Discovering knowledge in data: An introdution to data mining*. New Jersey: John Wiley & Sons, Inc.Wiley.

Lawler, R. (2012). Airbnb: Our guests stay longer and spend more than hotel guests, contributing $56m to the San Francisco economy. *TechCrunch*. Retrieved from http:// techcrunch.com/2012/11/09/airbnb-research-data-dump/

Li, H., Sun, J. & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications, 37*, 5895-5904. doi: 10.1016/j.eswa.2010.02.016

Lieberman, W. (2010). From yield management to price optimization: Lessons learned. *Journal of Revenue and Pricing Management*, *10*(1), 40–43. doi: 10.1057/rpm.2010.44

Lisboando (n.d.). Bairros de Lisboa. Retrived from https://lisboando.pt/bairros/

Marcotte, P., & Savard, G. (2003). A bilevel modelling approach to pricing and fare optimisation in the airline industry. *Journal of Revenue and Pricing Management, 2*(1), 23–36. doi: 10.1057/palgrave.rpm.5170046

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4): 115–133.

Modica, P., Landis, C., & Pavan, A. (2009). Yeld management and coastal hospitality industry demand. *TILTAI, 3*, 53–67.

Mohlmann, M. (2015). Collaborative consumption: Determinants of satisfaction and the likelihood of using a sharing economy option again. *Journal of Consumer Behaviour*, 14(3), 193-207. doi: 10.1002/cb.1512

Möller, M., & Watanabe, M. (2010). Advance purchase discounts versus clearance sales. *Economic Journal*, *120*(547), 1125–1148. doi: 10.1111/j.1468-0297.2009.02324.x

Müller, D. K. (2014). Progress in second-home tourism research. In A. A. Lew, C. M. Hall, & A.M. Williams (Eds.), *The wiley blackwell companion to tourism* (pp. 389–400) West Sussex: John Wiley & Sons, Ltd. doi: 10.1002/9781118474648.ch31

Ng, I. C. L. (2007). Advance demand and a critical analysis of revenue management. *Service Industries Journal*, *27*(5), 525–548. doi: 10.1080/02642060701411682

Ng, I. C. L. (2008). *The pricing and revenue management of services: A strategic approach.* New York: Routledge.

Opitz, D. & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169–198. doi: 10.1613/jair.614

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

Pan, B., & Yang Y. (2017). Forecasting destination weekly hotel occupancy with Big Data. *Journal of Travel Research 56*(7), 1-14. doi: 10.1177/0047287516669050

Papert, S. & Minsky, M. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, Mass.: The MIT Press.

Pestana, M. & Gageiro, J. (2005). *Análise de dados para ciências sociais: A complementaridade do SPSS*. Lisboa: E. Sílabo.

PriceLabs (2018). Revenue management for vacation rentals. Retrieved from https://pricelabs.co

Quinlan, J. R. (2006). *Bagging, Boosting, and C4.5*. In *Proceedings of the thirteenth national conference on artificial intelligence* (Vol. 1, pp. 725-730). AAAI Press.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408.

Rygielski, C., Wang, J., & Yen, D. C. (2002). Data mining techniques for customer relationship

management. *Technology in Society, 24*, 483–502. doi: 10.1016/S0160-791X(02)00038-6

Saló, A., & Garriga, A. (2011). The second-home rental market: A hedonic analysis of the effect of different characteristics and a high-market-share intermediary on price. *Tourism Economics*, *17*(5), 1017–1033. doi: 10.5367/te.2011.0074

Saló, A., Garriga, A., Rigall-I-Torrent, R., Vila, M., & Sayeras, J. M. (2012). Differences in seasonal price patterns among second home rentals and hotels: Empirical evidence and practical implications. *Tourism Economics*, *18*(4), 731–747. doi: 10.5367/te.2012.0141

Stinerock, R. (2018). *Statistics with R - A beginner´s guide*. London: SAGE Publications Ltd.

Talón-Ballestero, P., González-Serrano, L., Soguero-Ruiz, C., Muñoz-Romero, S., & Rojo-Álvarez, J. L. (2018). Using big data from customer relationship management information systems to determine the client profile in the hotel sector. *Tourism Management*, *68*, 187–197. doi: 10.1016/j.tourman.2018.03.017

Trochim, W. M. K. (2006). Descriptive statistics. Research Methods Knowledge Base

Turban, E. (2011). *Business intelligence: A managerial approach*. Boston: Prentice Hall.

UNWTO (2016). Tourism Highlights - 2016 Edition. Retrieved from http://www.e-unwto.org/doi/pdf/10.18111/9789284418145

Vives, A., Jacob, M., & Payeras, M. (2018). Revenue management and price optimization techniques in the hotel sector : A critical literature review. *Tourism Economics, 20*(10) 1–33. doi: 10.1177/1354816618777590

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. doi: 10.1016/j.ijhm.2016.12.007

Webb, T., & Schwartz, Z. (2017). Revenue management analysis with competitive sets: Vulnerability and a challenge to strategic co-opetition among hotels. *Tourism Economics, 23* (6), 1206-1219. doi: 10.1177/1354816616671473

Weiss, G.M., & Davidson, B. (2010). *Data mining.* In H. Bidgoli (Ed.) *The handbook of technology management* (pp. 2-17). New Jersey: John Wiley and Sons.

Witten, I.H. & Frank, E. (2005). *Data mining: Pratical machine learning tools and techniques*.

San Francisco: Morgan Kaufmann Publishers.

Zervas, G., Proserpio, D., & Byers, J. (2017). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Journal of Marketing Research*, *54*(5), 687–705. doi: 10.2139/ssrn.2366898

(This page was intentionally left blank)

Appendix

A: **Property's characteristics**

Characteristics	Count	%	Mean	Std.Dev	Minimum	Median	Maximum
Typology							
Studio	306	5,1					
T0	216	3,6					
T1	1783	29,7					
T1 Duplex	87	1,5					
T2	2605	43,4					
T2 Duplex	27	0,5					
T3	749	12,5					
T3 Duplex	54	0,9					
T4	120	2					
T5	39	0,7					
T6	13	0,2					
Total	**5999**	**100**					
Max_occupancy			4,2	1,4	1	4	10
Total	**5999**						
For_two	5223	87,1					
No	776	12,9					
Yes							
Total	**5999**	**100**					
FloorNumber			2,2	1,9	-1	2	13
Total	**5721**						
Parking							
No	5452	90,9					
Yes	547	9,1					
Total	**5999**	**100**					
Elevator							
No	4187	69,8					
Yes	1812	30,2					
Total	**5999**	**100**					

B: **Property's location characteristics (Neighborhood)**

Neighborhood	Count	%
Ajuda	81	1,4
Alcântara	11	0,2
Alfama	436	7,3
Almirante Reis	66	1,1
Areeiro	16	0,3
Av. da Liberdade	298	5,0
Av. Novas	204	3,4
Bairro Alto	539	9,0
Baixa	587	9,8
Beato	12	0,2
Belém	59	1,0
Benfica	6	0,1
Bica	23	0,4
Cais do Sodré	54	0,9
Campo de Ourique	156	2,6
Campolide	13	0,2
Castelo	20	0,3
Chiado	256	4,3
Estefânia	13	0,2
Estrela	121	2,0
Expo	143	2,4
Graça	348	5,8
Intendente	146	2,4
Lapa	78	1,3
Laranjeiras	7	0,1
Madragoa	277	4,6
Marquês de Pombal	256	4,3
Martim Moniz	273	4,6
Mercês	29	0,5
Mouraria	25	0,4
Olivais	67	1,1
Príncipe Real	450	7,5
Rato	23	0,4
Restauradores	121	2,0
Restelo	57	1,0
S. Bento	72	1,2
S. José	37	0,6
Saldanha	61	1,0
Santa Catarina	402	6,7
Santa Marta	34	0,6
Santos-o-Velho	22	0,4
Sé	100	1,7
Total	**5999**	**100,0**

C: **Reservation's Characteristics**

Variables	Count	%	Mean	Std. Dev.	Minimum	Median	Maximum
Avg_booking_advance			38,36	41,6	0	30	358
Total	**5999**						
NumReservations			5	3	0	5	14
Total	**5999**						
Month_number							
1	589	9,8					
2	602	10,0					
3	619	10,3					
4	620	10,3					
5	632	10,5					
6	401	6,7					
7	403	6,7					
8	418	7,0					
9	425	7,1					
10	427	7,1					
11	425	7,1					
12	438	7,3					
Total	**5999**	**100,0**					
Season							
High	1660	27,7					
Low	4339	72,3					
Total	**5999**	**100,0**					
Big_event							
No	3110	51,8					
Yes	2889	48,2					
Total	**5999**	**100,0**					
Fav_dist_channel							
Airbnb	2331	38,9					
Booking	150	2,5					
Don´t have	2661	44,4					
FeelsLikeHome	417	7,0					
Others	440	7,3					
Total	**5999**	**100,0**					
Pred_country_month							
Brasil	589	9,8					
France	1385	23,1					
Germany	689	11,5					
Portugal	193	3,2					
Spain	3143	52,4					
Total	**5999**	**100,0**					

D: Output from IBM SPSS Modeler