UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

TRABAJO DE FIN DE GRADO

TÍTULO: Analysis of the behavior of Airbnb in four different Spanish areas.

AUTOR: Fátima Domínguez Plata

TUTOR: Jesús Barreal Pernas

CURSO ACADÉMICO: 2019/2020

CONVOCATORIA: Junio

ABSTRACT

The appearance and rise of the Airbnb business has created a huge amount of debate and a new way (and a useful tool) to understand tourism. Being Spain highly dependent on this sector, by using the data provided by the website InsideAirbnb this essay tries to analyze the behavior of the business (and consequently, of tourism) in four Spanish areas that have different characteristics: Euskadi, Madrid, Málaga and Mallorca. Results suggest that the areas differ in the type of properties they offer, have similar peaks on tourism during certain months (except for Madrid), and that the prices depend on the size of the accommodation as well as on its location. It also shows that what customers value most is not the price, but the location, cleanliness and the value perceived.

Keywords: tourism, Spain, Airbnb, hedonic pricing

INDEX

CHAPTER 1. INTRODUCTION 2
CHAPTER 2. AIRBNB 3
2.1. Overview of the company 3
2.2. Airbnb and data analysis with R 4
CHAPTER 3. DATA 5
CHAPTER 4. EXPLORATORY ANALYSIS 7
4.1. Type of properties 7
4.2. Demand analysis 9
4.3. Price analysis 11
4.4. Text analysis 14
4.5. Spatial analysis 20
CHAPTER 5. REGRESSION ANALYSIS 24
5.2. Madrid 30
5.1. Málaga 31
5.4. Mallorca 32
5.5. Interpretation of results 33
CHAPTER 6. LIMITATIONS AND FURTHER ANALYSIS 36
CHAPTER 7. CONCLUSION 38
APPENDIX 40
BIBLIOGRAPHY 45

CHAPTER 1. INTRODUCTION

Being tourism one of the main economic engines in Spain, the topic chosen comes from a need to understand a new type of emerging business that has quickly grown and is starting to even change some tourists actions. Furthermore, with this growth has also come a huge debate on its impact on other markets and communities, and it is therefore important to understand as much as one can from it.

This paper is going to analyze the characteristics of Airbnb in four different Spanish areas: Euskadi, Madrid, Málaga and Mallorca. These areas are popular touristic destinations and they have different attributes that could make them subject to different types of tourism.

The main objective is to see how touristic accommodations behave in the different areas and understand the tourism that takes place in them. Understanding this would allow people to take advantage of the situation by knowing what a good accommodation offers, what customers value most, or how to adjust the price of the accommodation depending on different factors.

The structure of this work will be the following: First the company Airbnb will be described and the potential of data analysis to the business will be shown. Afterwards the data used will be described. The next chapter will be the exploratory analysis, which comprises an analysis of the types of accommodations offered, an analysis of the demand and its growth, an analysis of the price and its variability and seasonality, a text analysis of the reviews previous customers wrote, and a visualizations with maps (geographical and density map) of the listings locations. Chapter 4 is the regression analysis, where the methodology for the regression will be explained and the results will be shown and interpreted. The quality of the regression will be also checked with different tests such as the Jarque-Bera test and the Breusch-Pagan test. The next chapter will contain the many limitations of the paper and suggestions for further analysis. Finally, the last chapter will present the conclusion with the main ideas obtained from the analysis. In addition, at the end of the paper there is an annex with more detailed graphs that are not included in the chapters to maintain a more formal style.

CHAPTER 2. AIRBNB

2.1. Overview of the company

Airbnb was founded in 2008 in San Francisco and since then it has grown quickly and expanded across the globe. It is currently in 191 countries and 65.000 cities (*Fast Facts - Airbnb Newsroom*, n.d.), and it has challenged the hotel industry with its model of collaborative economy placing itself at an estimated value of 38$ billion (*As A Rare Profitable Unicorn, Airbnb Appears To Be Worth At Least $38 Billion*, n.d.).

Its exponential growth can be explained by several reasons (Oskam & Boswijk, 2016): its value proposition of offering a "local experience"; the engagement of its community (customers are part of the business and there exists a network rather than a hierarchy); being a pioneer of its field; innovation; a good digital interface; and more importantly, the rise of technology and digital communication, which allows it to reach new markets fast and cheaply.

Its business model consists on allowing people to rent their houses for a short period of time through a website. Airbnb makes guests and hosts come together and charges them both: for guests it charges a percentage under 14.2 of the reservation subtotal and for hosts a 3% of the service fee. Its objective is therefore to rent the properties offered in its portal as often as possible. Something really important in this type of business is consequently what customers think and how they value their experience because it will have an important impact on future potential customers.

The impact of Airbnb has not only been on the hospitality industry but also on the general way people approach tourism: it is believed to have a negative impact on hotels and a positive impact on tourism. Its impact has also been studied on the rental housing market, and on its legality and regulations -or lack of them-. Some cities, such as Barcelona, Amsterdam or Berlin are undergoing discussions and have had protest movements because many citizens believe that some areas are becoming just touristic renting areas and causing a bad impact on the community. Some governments have answered to this by restraining Airbnb activities and controlling them more thoroughly, but the market is so new that there is still a great debate around it.

2.2. Airbnb and data analysis with R

R is a programming language used mainly for statistical purposes and Airbnb has been known for using this tool to draw insights from their business data. They actually acknowledge part of their business success is due to their data science and analytics team (Bion et al., 2018). They make use of the data generated by their website portal to assist in decisions and get a better knowledge of what customers think and want. This work of their data science team can be divided in three main areas: Product Insights, Experimentation, and Predictive Modeling. (Bion et al., 2018).

Product Insights tries to use data to make the product Airbnb offers -touristic lodging- better for customers. They analyze the type of customers per region, their behavior and preferences; what are the most and the least booked hosts and why; the supply characteristics of each city… Based on all this information Airbnb is able to form hypothesis, test them and create new opportunities for its business. Here they use R for exploratory data analysis, data visualization and writing their research.

The *Experimentation* area does A/B testing with the hypotheses formed in the Product Insights area. When a hypothesis is tested and accepted, that might be reflected in an update on the website or the App. R is used here to run some of the more complex experiments that are not yet automated.

Finally, the *Predictive Modeling* area uses Machine Learning to, for example, assign prices to lodgings based on their characteristics -Smart Pricing-. When a host registers an apartment or house, Airbnb website can provide him with a price based on their predictive analysis algorithm -obviously, the host can always choose to ignore this suggestion and set his own price-. This algorithm takes into account the market popularity in the area of the apartment, seasonality, lead time, popularity of the apartment or the host, past reviews and previous bookings (*What's Smart about Smart Pricing? – The Airbnb Blog – Belong Anywhere*, n.d.). R has been used to, for example, build a marginal returns model, to optimize the guests service fee, or for regression trees to analyze how users are affected by different changes in features (Bion et al., 2018).

Airbnb has also encouraged people to use R (and other languages such as python) by organizing a recruiting competition in Kaggle where they made public a dataset with information about fictional users and asked people to predict in what country would those users make their first Airbnb booking (*Airbnb New User Bookings | Kaggle*, n.d.). People that did the best were offered a job opportunity at the Data Science and Analytics team in the company.

CHAPTER 3. DATA

The data used was obtained from the InsideAirbnb website (*Inside Airbnb. Adding Data to the Debate.*, n.d.), created by Murray Cox. The website is not associated with Airbnb and the data is gathered from public information of the company using web scraping. This compilation is available to make use of it for discussions and public analysis.

The data used for this essay is from four different areas with significant tourism in Spain: Euskadi, Madrid, Málaga and Mallorca. For each city I have used several datasets:

- **Listings:** 16 variables for each listing. The ones used for this study were: *id* (the identification number for each of the listings on the website), *latitude*, *longitude*, *room_type*, and *price*.
- **Detailed_listings:** 106 variables. The ones used for the regression analysis were: *id*, *latitude*, *longitude*, *price*, *bedrooms* (number of bedrooms in the apartment), *bathrooms* (number of bathrooms), *review_scores_rating (*rating that reviews gave to the apartment and its host), *number_of_reviews*, *cleaning_fee*.
- **Reviews:** 2 variables for each listing. *Date* and *id*. It was used to measure an approximation to the demand of Airbnb.
- **Reviews_detailed:** 6 variables with information about the reviews. The variable *comments* (the body text of each review) was used for the text analysis.
- **Calendar:** 4 variables: *id*, *date*, *availability*, and *price*. This dataset was used for the price analysis.

One of the big parts of the work was to clean and sort the data in a way that could be used for the different types of analysis and for its visualization. Some points to note are the following:

i) Data from *Shared room* and *Hotel room* were dropped, as well as null values in any of the variables, except for *reviews_per_month* (the null values of this variable were set to 0 in order to keep those listings).

ii) The prices in the datasets are in USD, therefore I converted them into euros using the USD/EUR exchange rates from the European Central Bank (*ECB Statistical Data Warehouse,* n.d.).

The prices were also deflated with the Spanish Hotel Price Index (Índice de Precios Hoteleros) obtained from the INE (*Índices de Precios Hoteleros*, *Índices y Tasas de Variación Interanual Por Comunidades Autónomas(12156)*, n.d.).

iii) All the coordinates used for the city center, beach and airport locations were obtained manually by me from Google Maps and inputted into R.

iv) The areas to which the data corresponds have some differences: for Euskadi and Málaga the data corresponds to the whole autonomous community (comunidad autónoma), for Madrid it corresponds to only the municipality (municipio), and for Mallorca it corresponds to the whole island, which is part of the Islas Baleares.

The R packages used were: dplyr for data manipulation; ggplot2 for data visualization; lubridate to work with dates; leaflet, sp, sf, raster, adehabitatHR, rgdal, rgeos, tmap, for the spatial visualization; tm and tidy text for the text analysis; igraph for the word network visualization; lmtest and olsrr for the regression analysis; grid, scales, and wesanderson for different aesthetic visualization features.

All the output, graphs and images were made by me. This was my first time using R and therefore there is a big amount of work and time behind each output.

CHAPTER 4. EXPLORATORY ANALYSIS

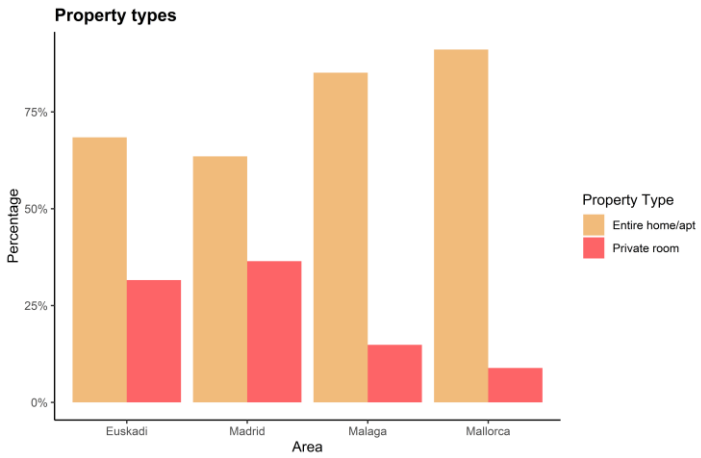
Exploratory analysis tries to summarize datasets by reducing very complex data into values or figures. For the purpose of this essay I will extract as much information as possible from the data scraped from the Airbnb website and will focus on: the type of properties offered, the demand, the prices, the reviews and the location.

4.1. Type of properties

The number of observations varies in relation to the dataset and what it is used for. In the *listings dataset* we are considering Euskadi (5,142 apartments/private rooms), Madrid (19,895); Málaga (6,060), and Mallorca (15,982). This already gives us an idea of how the distribution of listings across the different areas is, meaning that Airbnb is more used in Madrid and Mallorca than in Euskadi and Málaga.

Airbnb offers four types of properties: *Entire home/apartment* (it usually includes a bedroom, a bathroom and a kitchen), *Private room* (in which the guest rents only one room and shares the rest of the facilities with other people), *Shared room* (where different guests sleep in the same room), and *Hotel room* (some hotels are allowed to announce their rooms in the website). Since the *Shared room* and *Hotel room* are not significantly relevant (the numbers are very low), I decided to drop them for the whole analysis. I used therefore the count of entire homes and private rooms and divided them by the total number of listings offered in each of the areas.

Figure 1. Types of Airbnb properties.



Source: Compiled by the author.

In general, there is a higher offer of entire home/apartment (35,758) than private rooms (11,321). This is intuitive, since it is what Airbnb is mainly known of, and in addition people wanting to rent a private room, might as well choose to do so in a hotel.
In the chart we can already see a difference in the four areas of study. Those areas that are characterized by a "beach tourism" have a higher number of entire homes/apartments and not as many private rooms. This might be due to the fact that people usually travel to these places with their family, and a house will give them more freedom, apart from resulting cheaper than a hotel. They also tend to stay for longer while doing this type of tourism. A relevant fact to take into account is that there is a higher ownership of second homes in this type of areas from people that live in other cities and rent their houses while they are not there, being Spain the country with the highest amount of second home ownership in Europe (Guisan & Aguayo, 2010), both from Spaniards and foreigners.

Even though in Madrid and Euskadi, the offer of entire home/apartment is also higher than private rooms, the difference is not as big. This is due to people traveling to these areas for fewer days and with different purposes, such as business.
The distribution of the types of properties across the areas reflects the heterogeneity of the different types of tourism that take place in Spain: the areas that have been historically focused on a sun-and-beach tourism have a higher amount of second dwelling ownership and offer those accommodations to tourists.

As for the number of hosts, Euskadi has 2,806 unique hosts, Madrid 9,167 hosts, Málaga 2,372 and Mallorca 4,206. Calculating how many listings (accommodations) each host has would be interesting since that is one of the things that is getting a big amount of debate, and this is shown in the table below.

Table 1. Number of listings per host.

	1 listing	**2 listings**	**3 listings**	**> 3 listings**
Euskadi	77.15%	11.63%	4.15%	7.07%
Madrid	74.06%	12.75%	4.58%	8.61%
Málaga	66.70%	13.97%	5.73%	13.6%
Mallorca	69.95%	11.87%	4.22%	13.96%

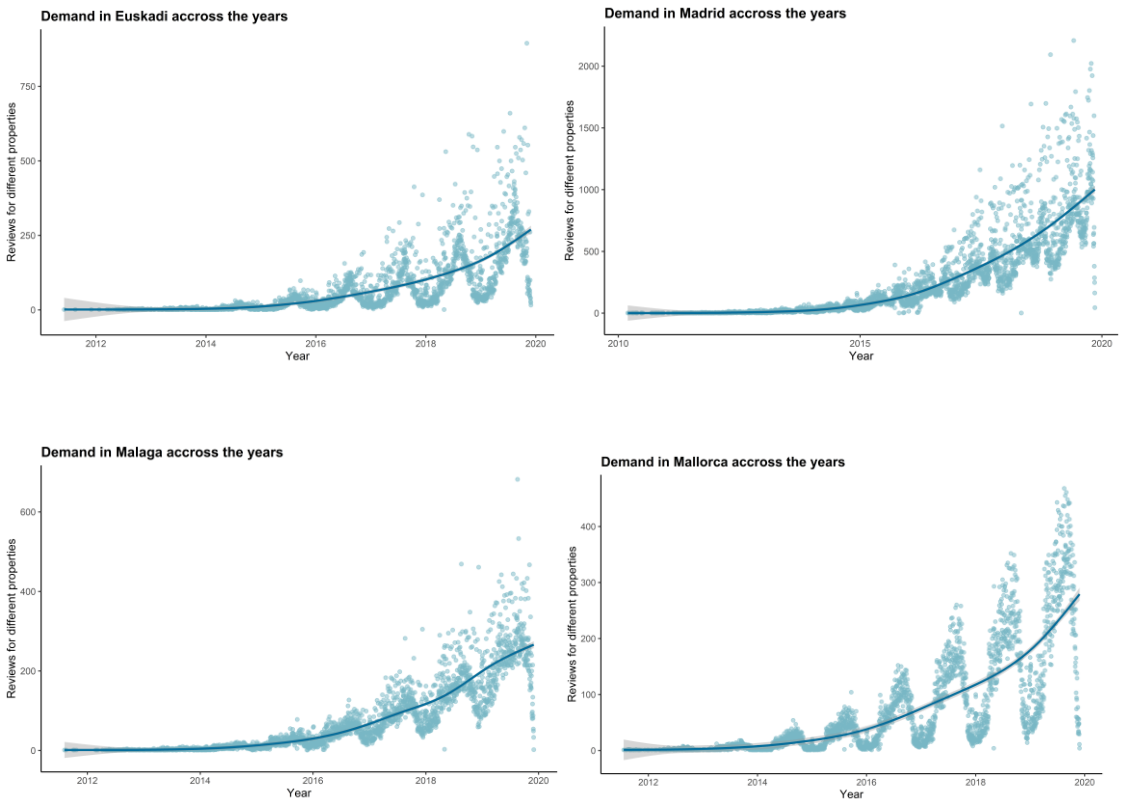
Source: Compiled by the author.

Most hosts have only one listing, but it is also quite common to have two or even more than three listings. This can be due to them making a mistake and posting their accommodation more than once, to those multiple listings being private rooms in a same house, or to them actually having more than one accommodation on renting, which is not something strange since many people treat Airbnb as a job and have many apartments that they constantly rent to tourists, and this happens mostly in Mallorca and Málaga.

4.2. Demand analysis

We are not able to access the information about how many times a property has been rented in Airbnb, therefore, to measure the demand I will use the number of reviews. Reviews can only be written by people after using the accommodation and even though not everyone writes a review it is a very common practice in Airbnb, but we can only assume that the demand is a bit higher than the one reflected in the study. Therefore, the count number of reviews at each different day for each area was considered to make the figures.

Figure 2. Demand of Airbnb across years.



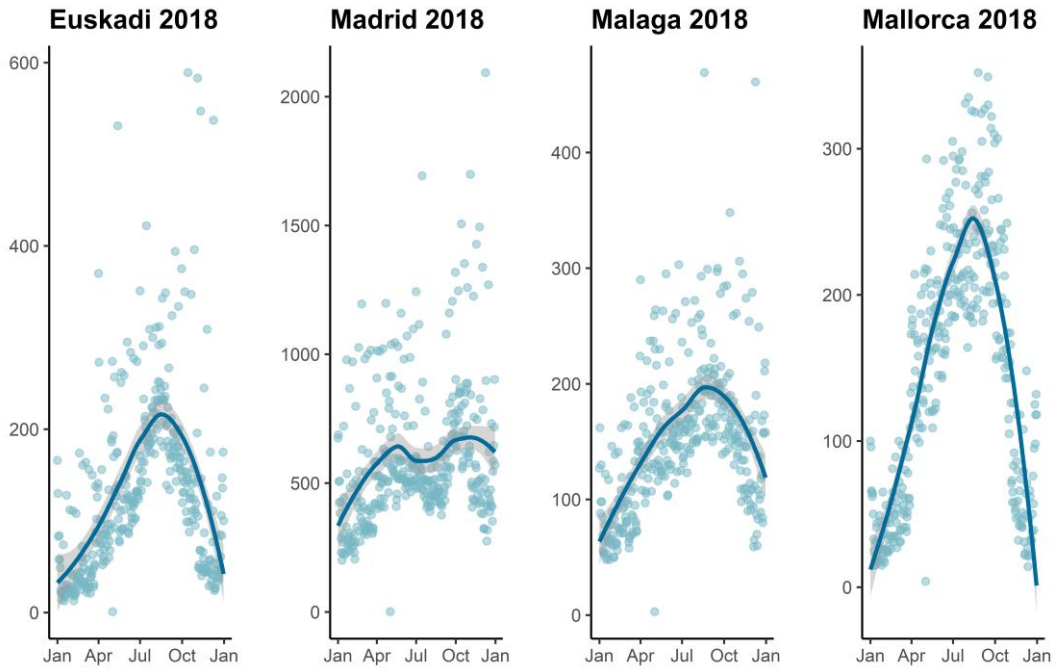
Source: Compiled by the author.

The data goes from 2010-2012 (it was in 2012 when Airbnb opened offices abroad, and in between those, one in Barcelona.) to the end of 2019. We can observe how across the years, the demand of Airbnb increased greatly in all the areas and it looks like it will increase even more in the future due to the rapid growth we talked about previously.

The seasonality can also be observed, as most tourism destinations experience it. Seasonality here refers to the fluctuations of tourism due to factors such as holidays or weather conditions. There are two types of main causes of seasonality in tourism (BarOn, 1972): natural seasonality, which refers to the weather (not only the sun but the snow as well) and institutional seasonality (religious events, and school and public holidays). Some other causes (Chung, 2009) could be sporting events, social pressure and inertia and tradition travelers.

In Figure 3 we have a more detailed graph of the year 2018 that reflects how some cities present a huge seasonality, and others not as much. This seasonality has great impacts since it can be reflected in an increase of tourists' expenditure, tourism employment and transportation.

Figure 3. Demand of Airbnb across the months.



Source: Compiled by the author.

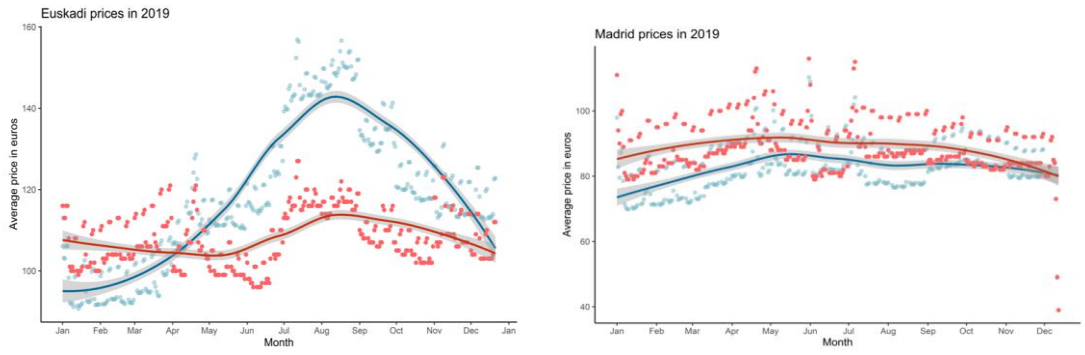
Since the scale is different (some cities, such as Madrid, are bigger and have a higher demand than others) I decided to keep each graph at their scale for easier comparison. As we can see, Mallorca stands out for its huge seasonality. As expected, its demand increases during summer, being at a peak in July and August. It occurs similarly in Euskadi and Malaga. And Madrid is the city that presents least seasonality, which can be because its type of tourism does not depend on the good weather and the beach and consequently it is more spread across the different months of the year, but also because people do not rent a house in Madrid for weeks for vacations since the tourism in the capital is shorter in length and of a different nature.

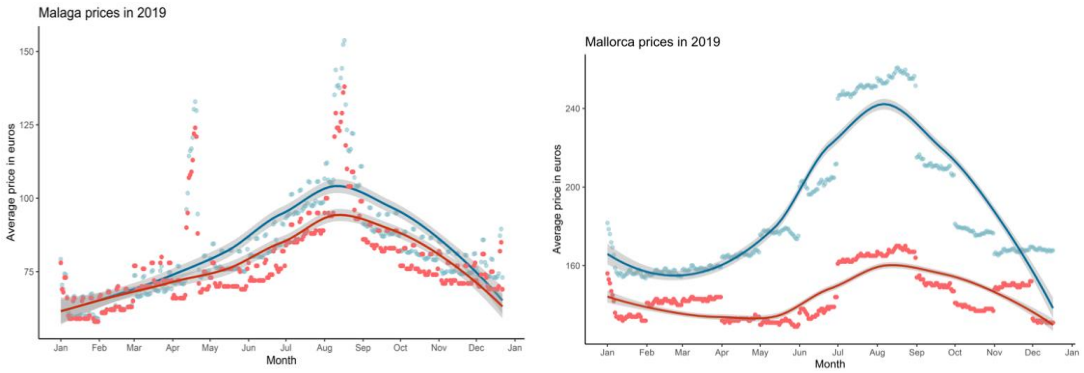
Sutcliffe and Sinclair made a study in 1980 where they analyzed the seasonality in Spanish tourism using tourists arrivals in the years 1952-1975 and the graph (after the Spanish Stabilization Plan in 1958) shows a very similar pattern to those of Airbnb (except Madrid's graph), decades after. They said this high seasonality should be slowed down by implementing counteractive policies such as taxes on tourism during the peak season, encourage domestic tourism during off-season, and developing other forms of tourism.

4.3. Price analysis

The prices were originally in USD and I converted them to euros using the exchange rate from the European Central Bank. The rates for the weekends and for holidays were not available, therefore I used the rate of the previous Friday for each of them. In the following graphs, the deflated prices ($\frac{Price}{HPI\ for\ the\ area}$) are shown in red.

Figure 4. Evolution of prices in 2019.





Source: Compiled by the author.

The prices' evolution across the year is quite similar to the demand evolution. We have peaks during the summer, except in Madrid where we have more stability. This increase in prices is especially noticeable in Mallorca, which has a huge peak in August. We can also see how there is basically no demand -and therefore very low prices- during some months (during the winter).

Málaga has also an increase in prices during the Easter holidays due to the Semana Santa celebrations. As for Madrid, the prices are stable during the whole year with some increases in July or June from people that might want to visit the city for a few days.

In table 2 we can see a detail of the prices: as for average prices, Mallorca is the most expensive, which is reasonable since if we recall it from before, most of their listings are from entire homes/apartments. In addition, Málaga has the cheapest average price, but not because its prices are always lower than those of the other areas, but because during the off-season its prices get very low.

Table 2. Detailed prices.

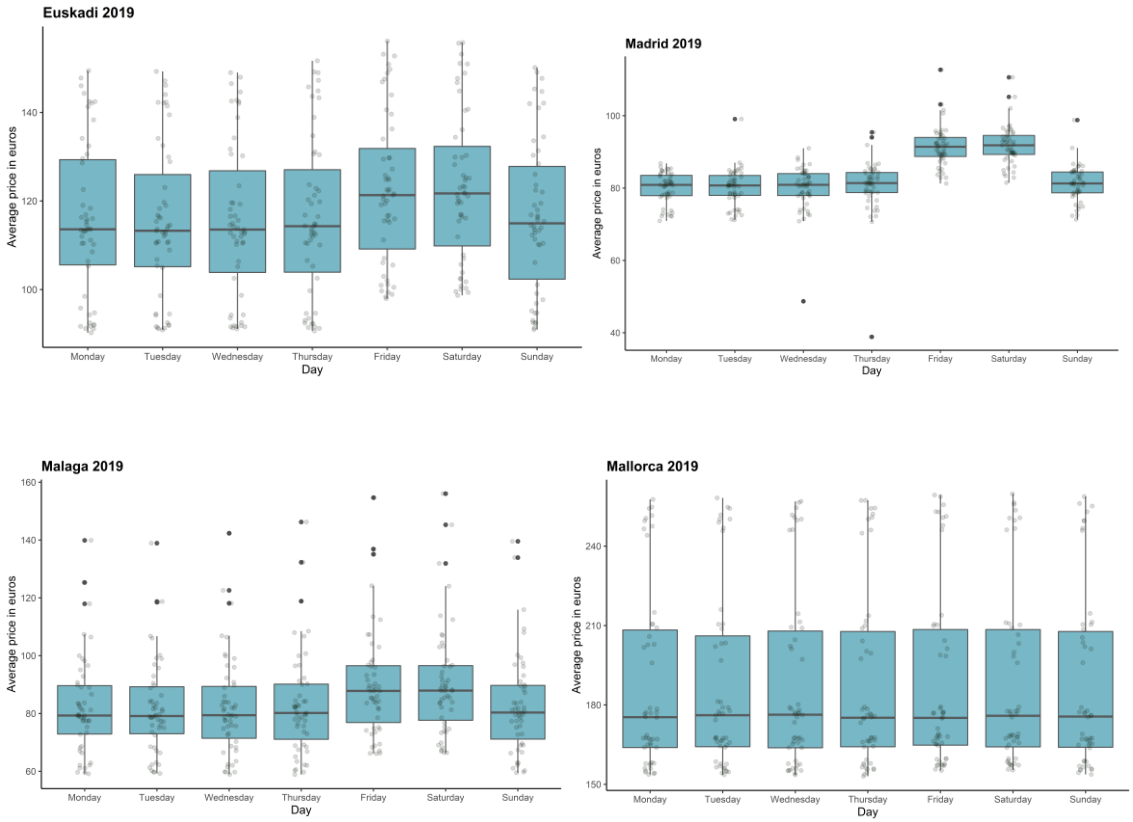
	Average price	**Maximum price**	**Minimum price**
Euskadi	108, 01€	127€ (12th July)	96€ (6th June)
Madrid	88,72€	116€ (31 May)	39€ (12th December)
Málaga	78,04€	138€ (17th August)	58€ (28th January)
Mallorca	144,11€	170€ (16th August)	129€ (28th May)

Source: Compiled by the author.

Prices also vary during the days of the week, which can be very important in determining the characteristics of tourism. For the boxplots in Figure 5 the average prices were grouped into days regardless of the month.

As we can see, the weekends (Fridays and Saturdays) have higher prices. This is due to the fact that most people travel on the weekends. Interestingly this is not the case in Mallorca, where prices are stable throughout the week. So here we can see another relevant difference in how these different areas behave in term of tourism: Mallorca is a special area characterized by a type of tourism that lasts more than one week and even months, therefore the price per day does not vary as much as in the other areas.

Figure 5. Average prices during the week.



Source: Compiled by the author.

To check the relationship between the price and some other variables (the review rating, the number of bedrooms and the number of bathrooms) I used the Pearson correlation coefficient, which is a very common way of identifying relationships between two variables.

$$r = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y} \quad (1)$$

A coefficient of 1 means that there is a perfect positive relationship. A coefficient of 0 means that there is not a linear correlation between the variables. And a coefficient of -1 means that there is a perfect negative relationship.

Table 3. Pearson correlation.

	Price – Nº bedrooms	Price – Nº bathrooms
Euskadi	r = 0.5513 p-value = 2,2e-16 t = 43.525	r = 0.4984 p-value = 2,2e-16 t = 37.864
Madrid	r = 0.1186 p-value = 2,2e-16 t = 15.022	r = 0.2349 p-value = 2,2e-16 t = 30.388
Málaga	r = 0.4416 p-value=2,2e-16 t = 35.202	r = 0.4060 p-value = 2,2e-16 t = 31.778
Mallorca	r = 0.5995 p-value = 2,2e-16 t = 75.78	r = 0.6311 p-value = 2,2e-16 t = 82.309

Source: Compiled by the author.

The p-value shown is very low, which is common in large datasets, and that is why R only shows p-values above the 2.2e-16 level. Therefore, it does not give us much relevant information about the significance of the correlation. Since the p-value < α (0,05) we reject the null hypothesis that r = 0, that is, the hypothesis that there is absolutely no relation between the variables. So we can say that there is a moderate positive relationship between price and number of bedrooms (especially for Mallorca -as people rent the whole apartment-) and Euskadi. There is also a moderate positive relationship between price and number of bathrooms (again, in Mallorca and Euskadi this is more noticeable).

4.4. Text analysis

Text analysis and text mining consist on the processing of different texts (in this case, online reviews from previous customers) to find unknown information or relationships that are not easily seen at a first glance or from a huge amount of data. With text analysis we can see different patterns and find very useful insights. Analysis in reviews is

important because negative comments, even when they touch a smaller number of topics than positive comments and are fewer in number, have a stronger impact (Humphreys, 2019).

Due to hardware limitations I was not able to do the text analysis on the whole dataset. Therefore, the analysis was done on a random sample that constituted a 40% of the whole dataset. First, the variable "comments" of the dataset (the one with the different reviews) is transformed into a vector source, and afterwards into a corpus (a collection of text). After that, we remove the punctuation and put everything into lowercase. Most of the reviews are in English, Spanish and German. Stop words -those that are very common and with little meaning for the task intended, such as determinant or pronouns- from the three languages were also removed. Other words I removed because of their irrelevance were the following -very repeated- terms:

Removed words	Apartment, madrid, stay, apartamento, está, piso, casa, also, est, flat, can, one, lugar, trés, malaga, city, nous, haus, palma, bilbao, town, city, pour, estancia, Sebastian…

After removing this we were left with a data set of Euskadi (101,598 words), Málaga (136,116 words), Madrid (116,786 words), Mallorca (136,116 words).

Figures 17-20 in the Appendix show the 20 most repeated words in the reviews for the different areas. There is one limitation for this type of analysis, and that is that since the reviews are done in many different languages, what is showed is just the count for the specific word in an specific language, but if we put together all the words from the different languages that have the same meaning, the count for the concept would be higher.

As we can see, the four areas have very positive words as the most common, with "great" leading the ranking in all of them. This gives us the impression that the customers have had a good experience in their rented apartments. From the most common words we can also infer what consumers find important, since that is what they would talk about, and those things are: location, cleanness, a welcoming host, and that the flat is within a walking distance of different places -implied in the words "walk" and "ubicación"(location)-. One of the things that can be surprising is that "price" is not one

of the most used words, meaning that it is not as important for customers as the other factors mentioned.

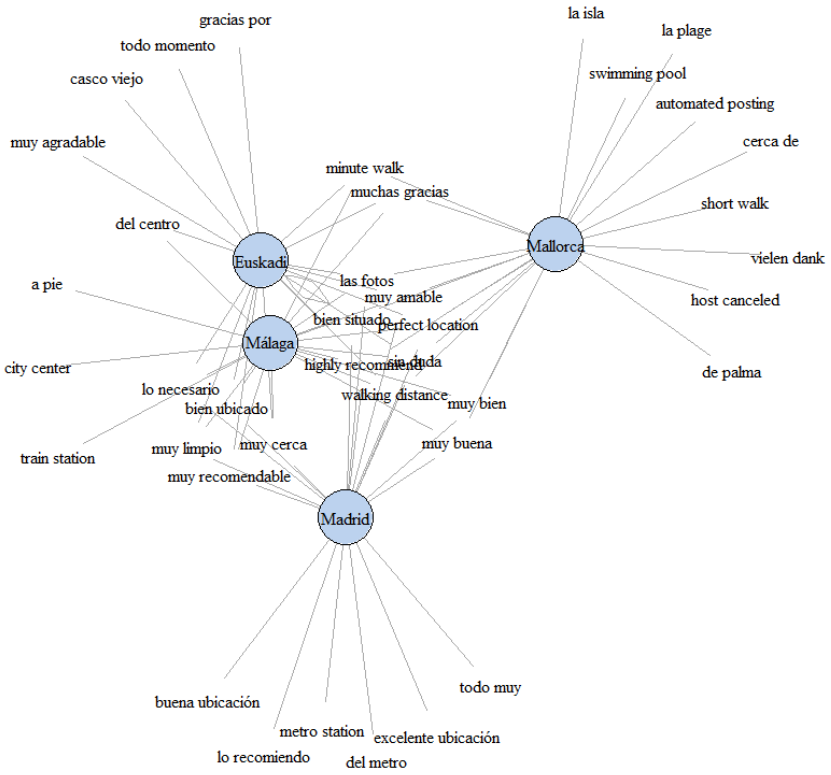
In Madrid it is also important that the flat is close to the metro, which is expected since it is a more urban area so public transportation has a greater weight. In Málaga and Mallorca the relevance is placed on the beach and summer related activities, but there is a bit of a difference since Málaga reviews focus more on the city center (the old town is a famous touristic attraction) while in Mallorca there is a bit of more focus in the "pool" (that is because the typical houses rented there are individual chalets with swimming pools, while the typical house rented in Málaga are apartments close to the beach).

It was also interesting to see how the top words for Mallorca (before stop words were removed) were German stop words, which just comes to show how Mallorca's main tourism comes from German speaking countries.

As for the word "recommend", which is a very relevant one in this topic, its frequency was: Málaga (5,86%), Madrid (5,69%), Mallorca (6,29%), Euskadi (5,15%). So customers that went to Málaga had a better experience than those in Euskadi, but overall the use of the word has a very similar frequency.

Perhaps a more revealing metric that could give us more information on customers wants would be not to see only individual words, but bigrams -"bags" of pair of words-. Figure 6 shows a word vector for the top 20 bigrams in the four different areas and table 4 documents the 10 most common ones with their frequency count.

Figure 6. Bigrams.



Source: Compiled by the author.

Table 4. The 10 most frequent bigrams.

Euskadi		Madrid		Málaga		Mallorca	
Bigram	Count	Bigram	Count	Bigram	Count	Bigram	Count
highly recommend	5.615	muy bien	58.216	city center	11.337	highly recommend	11.238
sin duda	5.482	walking distance	30.573	muy bien	9.677	walking distance	4.343
muy recomendable	5.040	muy buena	29.707	highly recommend	8.960	muy bien	4.110
muy amable	5.023	highly recommend	24.760	del centro	8.569	bien situado	3.157
muy limpio	4.847	bien ubicado	22.678	walking distance	7.275	vielen dank	2.625
las fotos	4.290	muy limpio	22.338	minute walk	6.643	perfect location	2.545
walking distance	4.102	las fotos	22.187	muy limpio	4.503	de palma	2.496
bien situado	3.934	muy amable	20.849	bien situado	4.364	minute walk	4.622
casco viejo	3.721	sin duda	19.871	muy buena	4.362	automated posting	2.440
lo necesario	3.548	lo necesario	15.770	las fotos	3.949	host canceled	2.323

Source: Compiled by the author.

As we can see, Málaga and Euskadi are the closest in similarity when it comes to reviews. It can draw our attention the presence in Mallorca's bigrams of "automated posting" and "host canceled", this is because when someone books an apartment and afterwards the host cancels it, an automatic review in Airbnb is posted with an "automated posting" message to warn potential customers. By the frequency of it we can assume that hosts in Mallorca are to be a bit less trusted or that the administration of the different rentings has some problems, maybe because of the huge demand of it during the summer.
The presence of the word "fotos" (pictures) shows how important it is for a listing to post pictures of the apartment since customers will definitely want to see it beforehand and will afterwards check whether the pictures were coherent with reality. Apart from this, we can assume what we already inferred before: having a clean apartment is important for customers, the closeness to different places of relevance such as the beach or the city center, and the closeness to different means of transportation (train station, metro station) are also important factors. In this regard, it is interesting how the airport does not appear as one of the relevant words. As for the host, it is also an important factor: the host's niceness and that the accommodation offered all what was needed is something that customers will also value.

Another way in which these bigrams could be useful is by searching different terms and checking what words come more often with them. When searching for "dirty" we get more than 200 bigrams referring to it, such as: "bit dirty", "dirty dishes", "dirty towels", "extremely dirty", "pretty dirty", "dirty bathroom", "dirty bed".
When searching for "uncomfortable" we also get a frequent "uncomfortable mattress". Another of the most frequent negative bigrams is the "mucho ruido" (lot of noise) one, with more than 500 counts. We can assume therefore, that in order to have a good touristic apartment, the host should have it clean and provide a comfortable bed and a quiet location.

Further research on this topic could be done with a sentiment analysis, which tries to analyze the "emotional intent" of the words in the reviews and classify them in "negative words" or "positive words", as well as seeing what emotions do they reflect most: joy, sadness, surprise, fear… This type of analysis is becoming increasingly important in today's world since industries have now a huge (and increasing) number of reviews in their platforms. Companies can make use of it to get a better customer understanding and

to improve the features of their products and services to, consequently, improve profits. In the Airbnb case, by doing a review analysis, hosts would be able to see what they lack and improve their competitiveness.

As for the ratings, users can rate their experience in the reviews in five different areas: *overall experience* (the variable in the dataset is called "review scores rating", and it is the general rating of the stay), *cleanliness* (how tidy and clean the accommodation was), *accuracy* (whether the pictures posted and the description of the accommodation were accurate), *value* (whether the relation price-accommodation was good or not), *communication* (how easy it was to communicate with the host, if he responded quickly and answered to all the questions…), *check-in* (how easy the check-in was), and *location* (rating of the neighborhood, closeness to relevant sites, amount of noise, safety…).

To check the correlation between all these ratings and the price I used the Spearman's correlation since the ratings variable is a rank variable and there might not be a linear relationship. The Spearman's correlation is calculated as follows:

$$\rho = 1 - \frac{(6 \cdot \sum di^2)}{n(n^2 - 1)} \qquad (2)$$

Where *di* is the difference between the ranks of i^{th} pairs of two variables, and *n* is the number of pairs of observations.

Figure 18 in the appendix shows the correlation matrix of the different ratings and prices for each area. The p-value for all of them, as with the Pearson correlation, is very low, therefore we just get a p-value of 2,2e-16. As we can see the price is positively correlated with the different scores, being the "location" the one with highest correlation in Madrid and Euskadi (ρ =0.2). Even though in Málaga and Mallorca there is also a ρ =0.2 of correlation between price and the general rating of the accommodation. But in any case, it is a low correlation.

The highest pair of correlation in all the areas (ρ = 0.7 and 0.6) is that between the general rating and the "value rating", which means that to have a higher rating in your accommodation the most important thing is that people perceive that the price paid is justified with the value they assign to the accommodation.

The correlation between general score rating and the accuracy of pictures and description is also a positive one. The same happens to score rating and cleanliness.
But overall there is not a significant difference between the areas (the only differences are of a few decimals), all of them have similar correlations. We could conclude with what we already stated: location, cleanliness and the value people perceive they got by the price they paid are the most important factors for customers. In addition, the price is not relevant when setting a rating score for the accommodation since there is barely any correlation between the price and the different ratings.

Table 5. Mean values of ratings.

	Euskadi	**Madrid**	**Málaga**	**Mallorca**
Overall experience	93,95	92,45	92,14	93,28
Cleanliness	9,52	9,38	9,40	9,41
Accuracy	9,67	9,53	9,53	9,53
Value	9,26	9,19	9,18	9,17
Communication	9,78	9,65	9,65	9,63
Checkin	9,73	9,63	9,63	9,68
Location	9,57	9,66	9,54	9,46

Source: Compiled by the author.

In table 5, the overall experience rating goes from 0 to 100 and the other ratings go from 0 to 10. As we can see, Euskadi is the area with best ratings except in the variable of location, where Madrid has a score a bit higher. The areas with worse ratings are Madrid and Málaga and their lowest ratings are in value (probably the accommodation was too expensive for what it offered).

4.5. Spatial analysis

In the spatial analysis I will try to display the distribution of the different Airbnb accommodations. The first type of map shows the listings over a geographical map (in the appendix it can be seen with bigger detail) and the second type of maps are density maps.

In order to avoid the limitations from a zone-based density map, that highly depends on how the different zones (neighborhoods) are delineated, and can change greatly from one zone to another producing a drastic un-continuous data visualization at the boundaries of each neighborhood (since it computes the average of each neighborhood), the maps were done using the kernel density estimation (Anselin et al., 2016). Kernel density estimation (KDE) is a technique often used in exploratory point data analysis to estimate density of that point-based data (the coordinates of the flats in this case). So kernel distribution does not assume an equal distribution of points within a zone/neighborhood as the zone-based density does. The default Kernel in R is the bivariate normal kernel (Gaussian kernel), which gives the probability density that an Airbnb listing, according to the geographical coordinates, is at a point. Therefore, in the maps, the most yellow parts are those in which there is a higher concentration of touristic flats. In this case, the estimation is done by calculating "the fraction of an observation point at location x based on the kernel function used and the distance between location x and the observation point" (Wilson & University Consortium for Geographic Information Science., n.d.). The kernel is calculated as (*How to Calculate Home Ranges in R: Kernels*, n.d.):

$$K\ (x) = \frac{1}{2\pi} e^{(-\frac{1}{2}x^i x)} \quad (3)$$

$$f\ (x) = \frac{1}{nh^2} \sum_{i=1}^{n} K(\frac{1}{h}(x - X_i)) \quad (4)$$

With *h* being the smoothing parameter. The higher the smoothing parameter, the larger the size of the range will be. The default smoothing parameter in the R function used was:

$$h = \left(0.5 * \left(sd_x + sd_y\right)\right) * n^{-\frac{1}{6}} \quad (5)$$

Some of the patterns we can infer from the maps are that most of the listings are situated on the coastline, which is one of the main touristic attractions of Spain. These accommodations are also situated in areas that already have an important tourism sector, therefore they can be a big competitor of hotels, or maybe there could be enough demand for both of them to work without affecting each other.

In Euskadi the offer of Airbnb flats is around the major cities (Bilbao, San Sebastián and

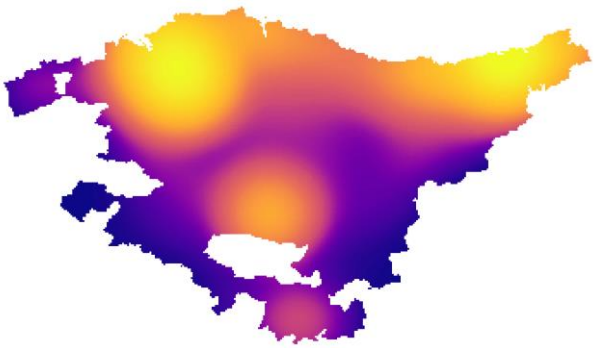
Vitoria). In Madrid, it is very concentrated in the city center. In Málaga the highest concentration is on the coastline but there are also many accommodations across the area because Málaga is small and all of the locations are therefore still close to the beach in any way. In Mallorca, the Airbnb accommodations are across the whole island and it is greater in the area of Pollença and Alcùdia than in Palma.

Due to several constraints I was not able to spatially analyze the dataset further, but it could be useful to calculate the spatial correlation with a Moran I test to check how the distance influences different variables. Research has been done on this on the study *"Airbnb Offer in Spain—Spatial Analysis of the Pattern and Determinants of Its Distribution"* (Adamiak et al., 2019), which shows how the location of Airbnb listings is affected in a positive way by the coast, a high number of secondary dwellings and interestingly, the hotel accommodation supply.

Figure 7. Geographical map Euskadi.

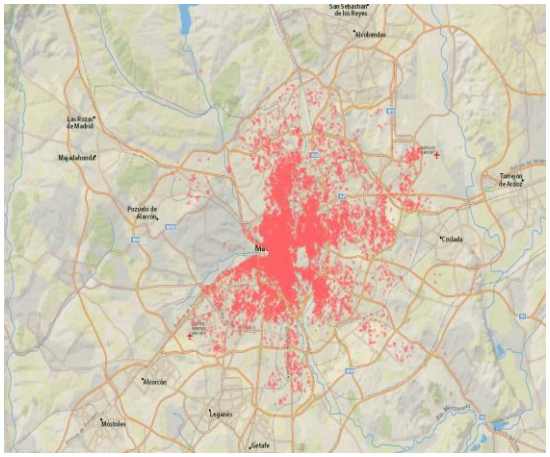
Source: Compiled by the author.

Figure 8. Density map Euskadi.



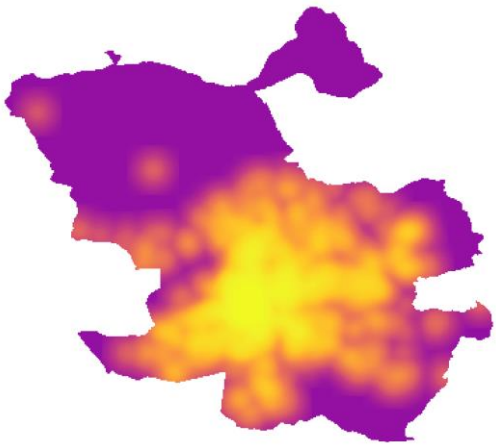
Source: Compiled by the author.

Figure 9. Geographical map Madrid.



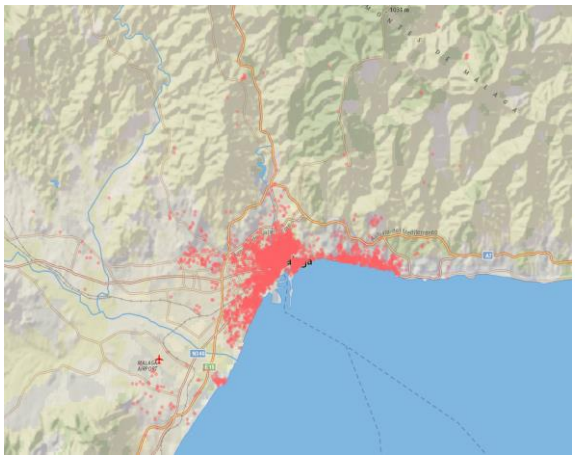
Source: Compiled by the author.

Figure 10. Density map Madrid.



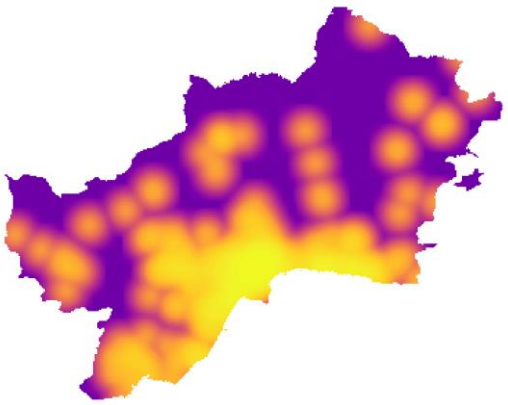
Source: Compiled by the author.

Figure 11. Geographical map Málaga.



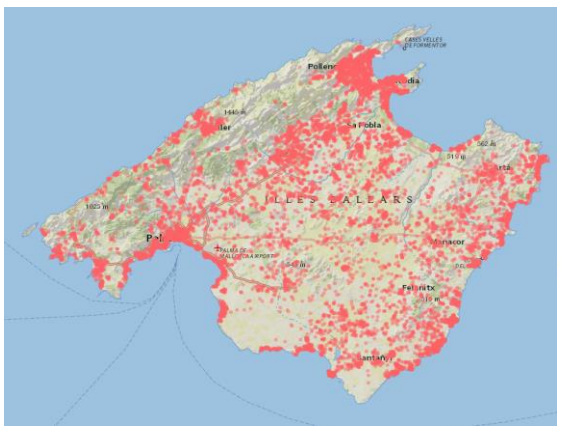
Source: Compiled by the author.

Figure 12. Density map Málaga.



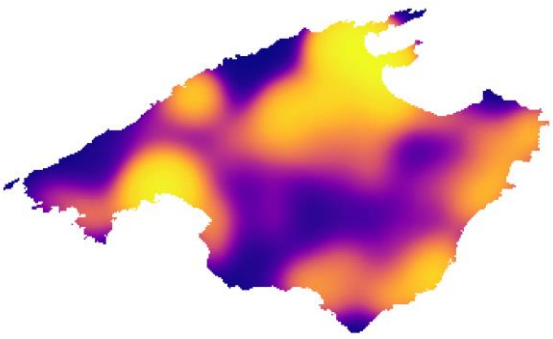
Source: Compiled by the author.

Figure 13. Geographical map Mallorca.



Source: Compiled by the author.

Figure 14. Density map Mallorca.



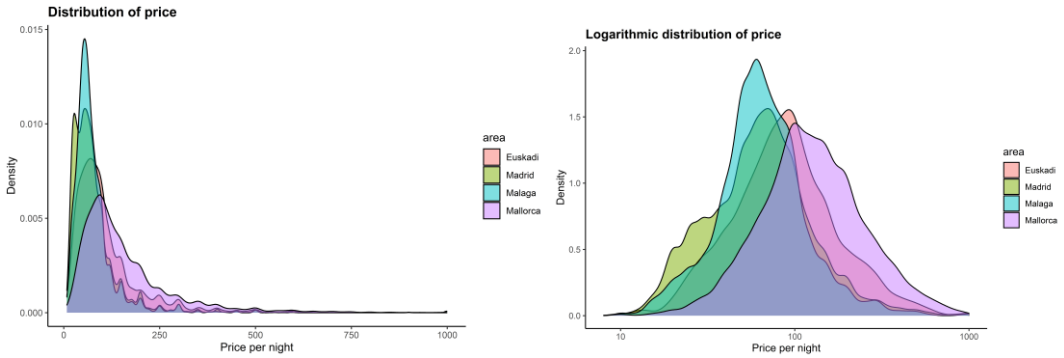
Source: Compiled by the author.

CHAPTER 5. REGRESSION ANALYSIS

In the regression analysis I will try to see the effect of different variables on the price in 2019 of the Airbnb flats. Due to the fact that the data is from public information scraped from the Airbnb website there are serious limitations for this analysis. First, we only know the price at which the flat was published when the data was scraped. Therefore, it is not possible to keep track of the evolution of the prices of each individual listing. It is only possible to see the evolution of the average price of all the listings across time. Being the time variable a hugely important one when it comes to the rental of tourism apartments, the regression model is going to obviously lack a lot of relevant information, but we can still see the role other variables play in determining the price of the accommodation.

The distribution of the dependent variable (Price per night) is very skewed to the right, so we will first do a logarithmic transformation to make the distribution less skewed and better for the regression.

Figure 15. Price distribution



Source: Compiled by the author.

The methodology followed is the spatial hedonic model with an OLS method. This model is very common and useful in economics, and tries to evaluate the contribution of different attributes to the final price of a good (Kuethe & Foster, 2008). Because of this, it has been widely used when analyzing housing prices. The basic equation is as follows:

$$P = f(S, E, L) \qquad (6)$$

Where P is a vector of prices, S has the characteristics of the house -the size, for example-, E has different environmental and socioeconomic variables, and L refers to the location. Here, the S will comprise the number of bedrooms, number of bathrooms and the cleaning fee. The E will comprise the number of reviews the accommodation has and the score it achieved, and L will comprise the distance to the city center and the distance to the beach. The model will therefore be the following:

$$\text{Ln}\,(Price_{area}) = \mu + \beta x_i + \varepsilon_i \quad (7)$$

Table 6. Regression variables.

Variable	**Description**
PRICE	Price per night when the data was scraped, log transformed
Accommodation characteristics	
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
cleaning_fee	Fee charged by hosts for the cleaning. Additional cost
"Socio economic"	
number_of_reviews	Number of reviews the listing has
reviews_score_rating	The general rating the listing got in the reviews
Location	
dist_beach	Distance to the beach (in km)
dist_center	Distance to the city center (in km)
dist_airport	Distance to the airport (in km). Only for Mallorca

Source: Compiled by the author.

Model 1:

$$\text{Ln}(Price_i) = \beta_0 + \beta_1 Nbedrooms_i + \beta_2 Nbathrooms_i + \beta_3 DistBeach_i + \beta_4 DistCenter_i + u_i \quad (8)$$

Model 2:

$$\text{Ln}(Price_i) = \beta_0 + \beta_1 Nbedrooms_i + \beta_2 Nbathrooms_i + Ln(\beta_3 DistBeach_i) + Ln(\beta_4 DistCenter_i) + u_i \quad (9)$$

Model 3:

$$\text{Ln}(Price_i) = \beta_0 + \beta_1 Nbedrooms_i + \beta_2 Nbathrooms_i + Ln(\beta_3 DistBeach_i) + Ln(\beta_4 DistCenter)_i + \beta_5 Cleaning\_fee_i + \beta_6 Number\_of\_reviews_i + \beta_7 Review\_scores\_rating_i + u_i \quad (10)$$

For the distance, some points to note are:

i) The beach chosen for Málaga was the Playa de la Malagueta, and the city center was the centro histórico (historic center).

ii) Madrid does not have a beach, so its models do not have that variable. As for the city center, the coordinates from the Plaza Mayor were chosen.

iii) In Euskadi, to choose the city center coordinates I grouped the accommodations into three areas: Guipúzcoa, Vizcaya, and Álava. And for each of them I chose a different city center and a different beach, as it is shown in the table below.

Table 7. Coordinates chosen for Euskadi.

	City center	**Beach**
Guipúzcoa	San Sebastián (Plaza de la Constitución)	Playa de la Concha
Vizcaya	Bilbao (Plaza Nueva in the Casco viejo)	Plentzia beach (at 25 km from Bilbao)
Álava	Vitoria (Catedral de Santa María in the Casco antiguo)	Playas de Garaio (inland beaches)

iv) The same was done with Mallorca, for better analysis I grouped the cities according to their proximity to eight mayor cities.

Table 8. Grouping and coordinates chosen for Mallorca.

	Cities	City center	Beach
Palma	Estellenchs, Andrach, Marrachí, Calviá, Bañalbufar, Santa María del Camino, Santa Eugenia, Palma, Valldemosa, Puigpuñent, Esporlas	Catedral-Basílica de Santa María de Mallorca	Playa de Can Pere Antoni
Pollença	Pollença, Escorca	Pollença Plaza mayor	Playa de Port de Pollensa
Alcùdia	La Puebla, Alcùdia	Church St.Jaume	Playa d'Alcudia
Manacor	Manacor, San Lorenzo de Cardessar, Porto cristo, Felanich, Villafranca de Bonany, Ariany, Petra, San Juan	Torre del Palau	Cala Magraner
Artà	Artá, Son Servera, Capdepera	Historic Center	Playa Arenal De Sa Canova
Lluchmayor	Lluchmayor, Montuiri, Campos, Porreres, Santañy, Las Salinas, Algaida	Plaza de España	Cala Blava
Sóller	Fornalutx, Buñola, Deyá, Sóller	Plaza de la constitución	Playa de So Calobra
Inca	Inca, Sancellas, Campanet, Consell, Binisalem, Sineu, Mura, Alaró, Búger, Costitx, Mancor del Valle, Lloseta, Lloret de Vista Alegre, María de la salud, Santa *Margarita*, Llubí	Historic center	Playa de Muro

In addition, I added the variable of distance to the airport since I think it is important considering that Mallorca is an island.

The distance calculated was the haversine (half-versed-sine) distance. It calculates the distance between two different points of latitude and longitude on earth, on the surface of the sphere. When the points are as closed together as here (in the same city) it really does not make that much of a difference to calculate the haversine distance or the euclidean distance since they will return a very similar value. The reason to choose the haversine one was merely because of the easiness when coding it into the regression. In addition, the output in R was expressed in meters, but I converted it to kilometers before inputting it into the regressions.

Given point A and point B, with their respective longitude and latitude, the haversine distance is calculated as follows (*James Inman*):

$$d = 2r \arcsin\left(\sqrt{sin^2\left(\frac{\Delta latitudeDifference}{2}\right) + \cos(lat1)\, cos(lat2) sin^2\left(\frac{\Delta longitudeDifference}{2}\right)} \right) \quad (11)$$

Where:

ΔlatitudeDifference = lat2 – lat1; ΔlongitudeDifference = lon2 – lon1; and r = radius of earth (the default in R is 6378137 meters).

5.1. Euskadi

Table 9. Euskadi regressions.

	Model 1	**Model 2**	**Model 3**
(Intercept)	1.690***	1.735***	1.458***
	(0.009)	(0.009)	(0.046)
bedrooms	0.138***	0.139***	0.132***
	(0.004)	(0.004)	(0.004)
bathrooms	0.032***	0.028***	0.024***
	(0.006)	(0.006)	(0.006)
dist_center	0.003***		
	(0.001)		
dist_beach	-0.008***		
	(0.000)		
log(dist_center)		0.003	-0.002
		(0.003)	(0.003)
log(dist_beach)		-0.062***	-0.059***
		(0.003)	(0.003)
cleaning_fee			0.001***
			(0.000)
number_of_reviews			-0.001***
			(0.000)
review_scores_rating			0.003***
			(0.000)
R^2	0.377	0.419	0.437
Adj. R^2	0.377	0.418	0.436
Num. obs.	4339	4339	4339

***p < 0.001; **p < 0.01; *p < 0.05

Source: Compiled by the author.

5.2. Madrid

Table 10. Madrid regressions.

	Model 1	**Model 2**	**Model 3**
(Intercept)	1.761***	1.715***	1.548***
	(0.006)	(0.005)	(0.022)
bedrooms	-0.003*	-0.003	-0.000
	(0.001)	(0.001)	(0.001)
bathrooms	0.092***	0.090***	0.042***
	(0.005)	(0.004)	(0.004)
dist_center	-0.033***		
	(0.001)		
log(dist_center)		-0.081***	-0.067***
		(0.002)	(0.002)
cleaning_fee			0.004***
			(0.000)
number_of_reviews			-0.000***
			(0.000)
review_scores_rating			0.001***
			(0.000)
R^2	0.109	0.117	0.269
Adj. R^2	0.109	0.116	0.269
Num. obs.	15803	15803	15803

***p < 0.001; **p < 0.01; *p < 0.05

Source: Compiled by the author.

5.1. Málaga

Table 11. Málaga regressions.

	Model 1	**Model 2**	**Model 3**
(Intercept)	1.620***	1.573***	1.408***
	(0.008)	(0.008)	(0.030)
bedrooms	0.104***	0.108***	0.094***
	(0.003)	(0.003)	(0.003)
bathrooms	0.053***	0.051***	0.036***
	(0.007)	(0.006)	(0.006)
dist_center	0.034***		
	(0.005)		
dist_beach	-0.055***		
	(0.006)		
log(dist_center)		-0.026***	-0.030***
		(0.004)	(0.004)
log(dist_beach)		-0.039***	-0.037***
		(0.006)	(0.006)
cleaning_fee			0.002***
			(0.000)
number_of_reviews			-0.001***
			(0.000)
review_scores_rating			0.002***
			(0.000)
R^2	0.284	0.300	0.363
Adj. R^2	0.283	0.300	0.363
Num. obs.	5116	5116	5116

***p < 0.001; **p < 0.01; *p < 0.05

Source: Compiled by the author.

5.4. Mallorca

Table 12. Mallorca regressions.

	Model 1	**Model 2**	**Model 3**
(Intercept)	1.653***	1.592***	1.480***
	(0.007)	(0.011)	(0.024)
bedrooms	0.087***	0.087***	0.078***
	(0.002)	(0.002)	(0.002)
bathrooms	0.075***	0.074***	0.069***
	(0.003)	(0.003)	(0.003)
dist_center	-0.002***		
	(0.000)		
dist_beach	0.003***		
	(0.000)		
dist_airport	0.001***		
	(0.000)		
log(dist_center)		-0.009***	-0.014***
		(0.002)	(0.002)
log(dist_beach)		0.028***	0.026***
		(0.002)	(0.002)
log(dist_airport)		0.020***	0.010**
		(0.003)	(0.003)
cleaning_fee			0.000***
			(0.000)
number_of_reviews			-0.001***
			(0.000)
review_scores_rating			0.002***
			(0.000)
R^2	0.476	0.478	0.501
Adj. R^2	0.476	0.477	0.501
Num. obs.	10152	10152	10152

***p < 0.001; **p < 0.01; *p < 0.05

Source: Compiled by the author.

5.5. Interpretation of results

I will focus on Model 3 since it is the most complete one. To interpret the coefficients, we must notice that the dependent variable (Price) was transformed to logarithms, so the coefficients (β_i) will be interpreted as percentages.

The first coefficient to interpret is that of *bedrooms*, which shows how much the price of the accommodation increases when there is one bedroom more. When there is one bedroom more, the price increases in $\beta_1 \times 100$. It affects the most to the prices in Euskadi, where one room increases the price by a 13.2%, followed by Mallorca (+9.4%) and Málaga (+7.8%), and there is not a relevant increase in the price of Madrid. This aligns with what we inferred before from the descriptive analysis, that Madrid does not have a "family"-type of tourism, therefore the amount of rooms in the apartment won't be as relevant.

The second coefficient, *bathrooms*, indicates how much the price increases when there is one bathroom more. The interpretation is done in the same way as with the bedrooms coefficient. As we can see from the regression tables, bathrooms increase the price less than bedrooms. One bathroom has the most impact in Mallorca (+6.9%), interestingly this is followed by Madrid (+4.2%), Málaga (3.6%) and Euskadi (+2.4%). While the number of bedrooms does not seem to have a big impact in the price in Madrid, the number of bathrooms does. It is interesting to see that the opposite happens with Euskadi, the number of bedrooms has a higher impact in price than in the other areas whereas the number of bathrooms has a lower impact.

Afterwards come the distance variables. Since they are also in logarithms, the interpretation of the coefficient is that if the distance to the center/beach increases a 1%, the price of the accommodation increases a $\beta_3\%$ and $\beta_4\%$ respectively.

If the distance of the apartment to the center increases, that is, that the apartment is more far away from the center, the price of the accommodation decreases, and this happens in all the areas: Madrid (-0.067%), Málaga (-0.030%), Mallorca (-0.014%) and Euskadi (-0.02%).

As for the beach, the more far away the apartment is, the lower is the price, except in Mallorca. In Euskadi the price decreases a 0.059%, and in Málaga it decreases a 0.037%. Interestingly, in the model it shows that in Mallorca, the price actually increases the more

far away the apartment is from the beach (+0.026%), this is maybe because the houses close to the beach are smaller in size (probably accommodations in apartments communities with shared swimming pools) and have less luxury, while the more inland it gets, the bigger the houses are and they have gardens and private swimming pools or different services. It happens the same with the airport, the more far away the apartment is from the airport, the higher the price (it increases a 0.010% if the distance increases a 1%).

The next coefficient is the *cleaning fee* that the accommodation may or may not have. This cleaning fee is something the host chooses to charge the guests due to the cleaning that takes place after they leave. This coefficient is interpreted the same way as the first two coefficients, $\beta_i \times 100$. Based on the results we can say that increases in the cleaning fee rise the price of the apartment.

The next one, *number of reviews* checks how much the price increases if the apartment gets reviews (independently of if they are positive or negative). In all the areas (except in Madrid where it has no impact) one additional review decreases the price by -0.1%. Even though this probably comes to say that those apartments that have lower prices have more reviews since they are rented by more people.
Lastly, the *review scores rating* coefficient, measures how much the price increases or decreases when the total rating of the apartment increases by one unit. The results show that when the score increases, the price increases too. It has the most effect in Euskadi (the price increases +0.3%), Málaga (+0.2%), Mallorca (+0.2%), and finally Madrid (+0.1%).

All the coefficients are significant, as shown by the p-value (it is less than 0.001). As for the R^2 coefficient, that shows the goodness of fit is of: Euskadi (43.7%), Madrid (26.9%), Málaga (36.3%) and Mallorca (50.1%). That means that the variables used explain the prices of 43.7% of the Airbnb accommodations in Euskadi, and the same interpretation can be done for the other areas. The remaining percentage is probably due to the date at which the accommodations are being rented, as we saw in the descriptive analysis, where price fluctuated greatly depending on the month, being very high in the summer. Some of that percentage could also be due to spatial dependence.

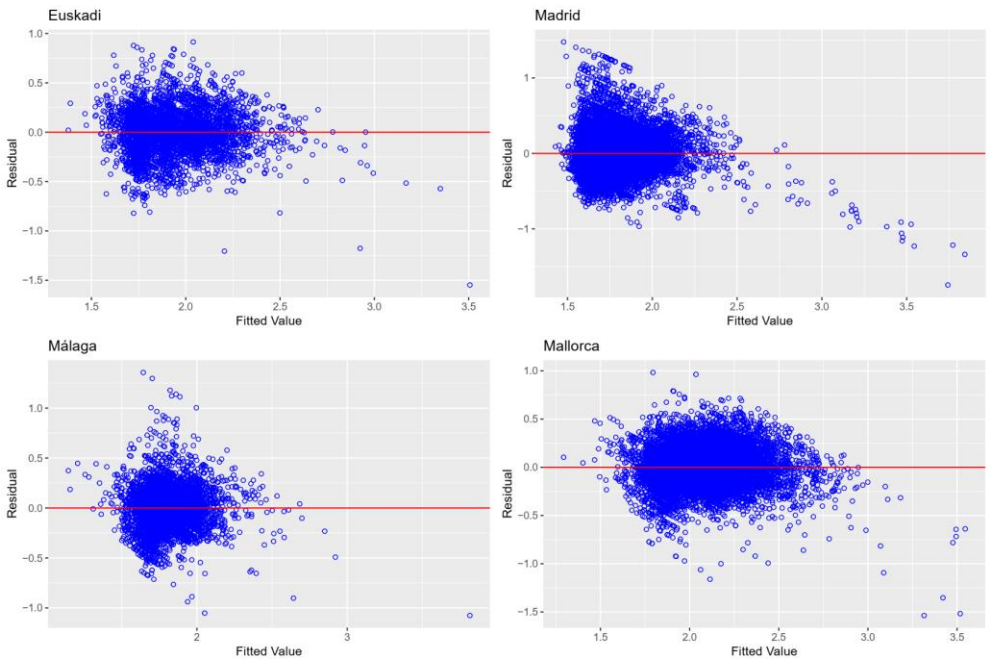
To check the existence of an approximate multicollinearity I calculated the VIF (Variance Inflation Factors):

$$VIF = \frac{1}{1 - R_j^2} \qquad (12)$$

The VIF for all the variables in the four different areas is lower than 4, therefore we do not have multicollinearity in the model. Even if there was, our coefficients would still be efficient since the approximate multicollinearity does not affect the basic LS properties, but we would not have precision in the confidence intervals.

Furthermore, in an OLS model the error terms are assumed to be: i) normally distributed, ii) independent, iii) and homoscedastic. To check if this is true in the model, it is useful to plot the residuals versus the fitted values.

Figure 16. Residuals vs fitted values.



Source: Compiled by the author.

At a first glance the variability does not seem to be more or less equal, the variance of the residuals seems to be decreasing, which means that the variance is not constant. In order to see this better, the Breusch-Pagan test, which is used to determine the existence of heteroskedasticity (when the variance of the errors is not constant), is calculated. The null hypothesis is that there exists homoscedasticity, that is, that the variance of the errors is

constant. This null hypothesis is rejected for the four regressions, so there is heteroskedasticity in the model. Similarly, the White test also rejected the null hypothesis of homoscedasticity for the model.

In addition, in order to check normality, I calculated the Jarque-Bera test in the error terms of the regressions. The null hypothesis in this test is that the residuals follow a normal distribution, but the Jarque-Bera test rejects this null hypothesis in the four regressions. It is common that the residuals don't follow a normal distribution and this will not affect the unbiasedness, efficiency and consistency of the model, which will still hold, but the hypothesis tests and confidence intervals won't be valid.

CHAPTER 6. LIMITATIONS AND FURTHER ANALYSIS

There are many limitations in this analysis. One of them is that the data is scraped from the Airbnb website, it is therefore public data and only shows what was posted by hosts. There is no way to tell if those characteristics (the number of bedrooms, bathrooms, location…) were real, nor if the final price at which the accommodation was rented is the same one as the price that was posted on the website.
We also do not have information from the demand, as that is private information, and have inferred it from the number of reviews (but, as said earlier, the demand must be higher than the number of reviews written).
As for the text analysis, as we noticed before, since there are different languages used, the count of "meanings" is not exactly accurate. We are only counting words, and if they are in different languages they will be counted separately even if they mean the same.

In regard to the models used, the R^2 is relatively low, which means that not all the factors that influence the price have been analyzed. One of the greatest limitations is the fact that we cannot use the variable of time, which is one of the most important factors when it comes to tourism and prices, since renting an apartment by the beach during the summer is not the same as renting it during the winter. The prices used for the regression are those that were posted in the moment of the web scraping and they are highly dependent on the dates. More factors that influence the price of the accommodations and that were not being considered are the proximity to other Airbnb similar accommodations and how influenced they get by their neighbors' prices, the proximity to public transportation

(especially in Madrid) or to supermarkets and restaurants. And finally, the subjective opinion of the hosts, how much do they think their accommodation is worth per night, their volatility, and how well-equipped the accommodation is (the amount of furniture, for example) or if it has nice views.

We must also be wary about the distance to the city center and to the beach, since for example only one beach was chosen for Málaga. For Euskadi and Mallorca, more beaches where chosen depending on location, but there might obviously be other beaches closer to certain apartments than the ones I chose. As for the city center, it is quite wide in Madrid, so instead of a specific point (the Plaza Mayor in this case) maybe it would be more accurate to choose a radius and calculate the distance of the apartments from it. In Mallorca, even though the cities were separated into eight different groups and each of them had a different city center point (the one in the biggest city) those other cities obviously have a city center and the price of the apartments in them will be also influenced by the proximity to it, not only by their proximity to the city center in the big neighboring city.

Lastly, the model could have also been better if another methodology was chosen instead of the OLS, such as the Geographically Weighted Regression (GWR), that helps explore the varying spatial relationships.

Further analysis could be done by doing a thorough sentiment analysis on the reviews in all the languages in which they appear and including those sentiments into the regression. Another suggestion is to include socioeconomical variables such as the amount of immigration or the average income of the neighborhood. Other interesting suggestions would be to include the street-level noise (since we saw that it is important for customers) or the closeness of the accommodation to hotels.
In addition, for a more complete analysis, the variable of time should be added in order to see the fluctuations of the price. With this we could also be able, not only to analyze how public and school holidays affect the price, but also how much different events (religious, sport…) affect tourism. Lastly, it would also be interesting in the future to see how the COVID19 affected the Airbnb business and tourism in Spain, and how long it takes to go back to the previous pace.

CHAPTER 7. CONCLUSION

The paper used the data provided by InsideAirbnb (*Inside Airbnb. Adding Data to the Debate.*, n.d.) to analyze the characteristics of these touristic accommodations and to try to see a bit of how tourism behaves in four different areas in Spain. Even though there are serious limitations for this analysis the insights we can get from it are still useful and relevant.

From the exploratory analysis we gained insights on the growth of this business, what types of properties are being offered the most in the areas studied, when do we have peaks of tourism in each of the areas (considering the months and also the days of the week), how prices fluctuate according to this demand, what customers value the most according to the reviews they wrote, and where are the listings located.

Airbnb has grown greatly and has an increasing tendency (even though this conclusion would have to be further checked in the future because of the COVID19). The accommodations offered are mainly entire houses (especially in Málaga and Mallorca) and in a lower amount, private rooms. The demand is at its highest during the summer in all the areas except in Madrid, where it behaves in a flatter way. Also, during the weekends, the demand increases. As for the prices, they behave in a similar way to the demand, being the highest during the summer.

In regard to what customers value, we can state that they give the most importance to: location, cleanliness and the value perceived. Interestingly, price is not a very common topic in the reviews.

Lastly, in the maps we saw how the accommodations are very concentrated in the city center and in the coastline, and that Madrid seems to have the most concentration, whereas Mallorcas' listings seem to be more spread across the whole island.

For the regressions the dependent variable, price, was transformed to logarithms for better analysis. The independent variables used were: number of bedrooms, number of bathrooms, distance to the city center, distance to the beach, distance to the airport, cleaning fee, number of reviews and the score of the reviews.

The regression analysis confirmed what we inferred before, that the size of the accommodation increases the price. In addition, the closeness to the city center and to the

beach increases the price as well (except in Mallorca, where the closeness to the beach appears to decrease the price, probably because the houses will be smaller). In addition, as could be expected, the score obtained in the reviews increases the price of the accommodations as well.

In conclusion, thanks to the Airbnb public information we can understand a bit better tourism and how it is different in each Spanish area, which can help to take advantage of it by adjusting the offer better to customers, knowing what to expect from tourists in order to be prepared for it, and predict future prices.

APPENDIX

Table 17. 10 most common words in Euskadi reviews.

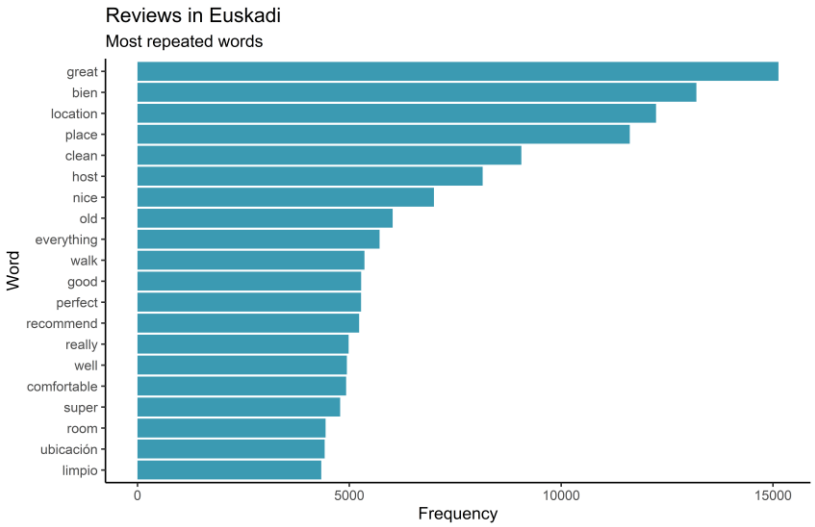


Table 18. 10 most common words in Madrid reviews.

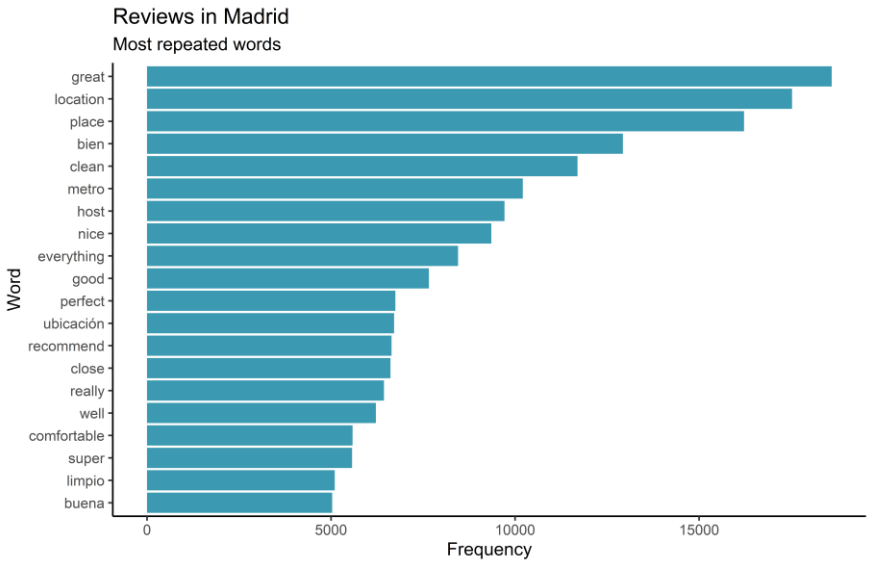


Table 19. 10 most common words in Málaga reviews.

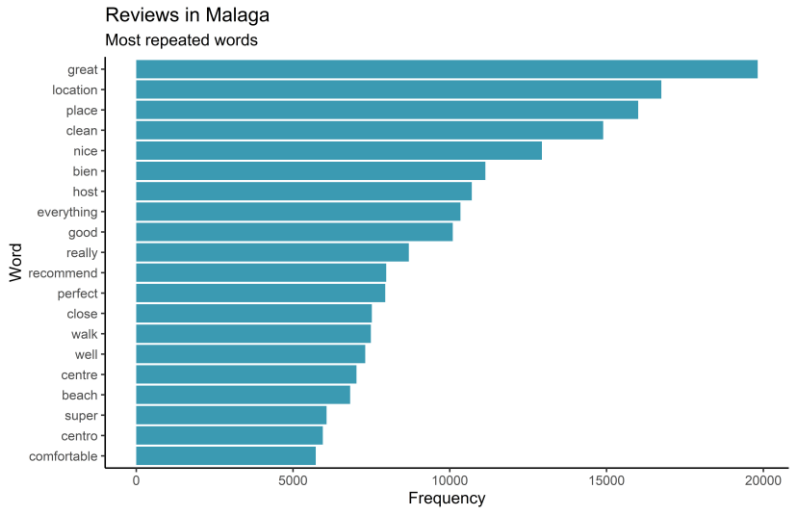


Table 20. 10 most common words in Mallorca reviews.

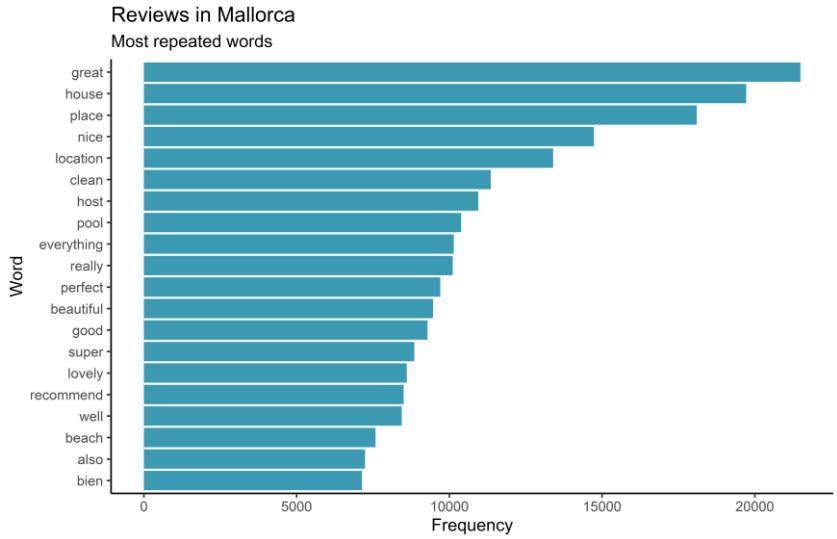


Table 21. Scores correlation matrix.

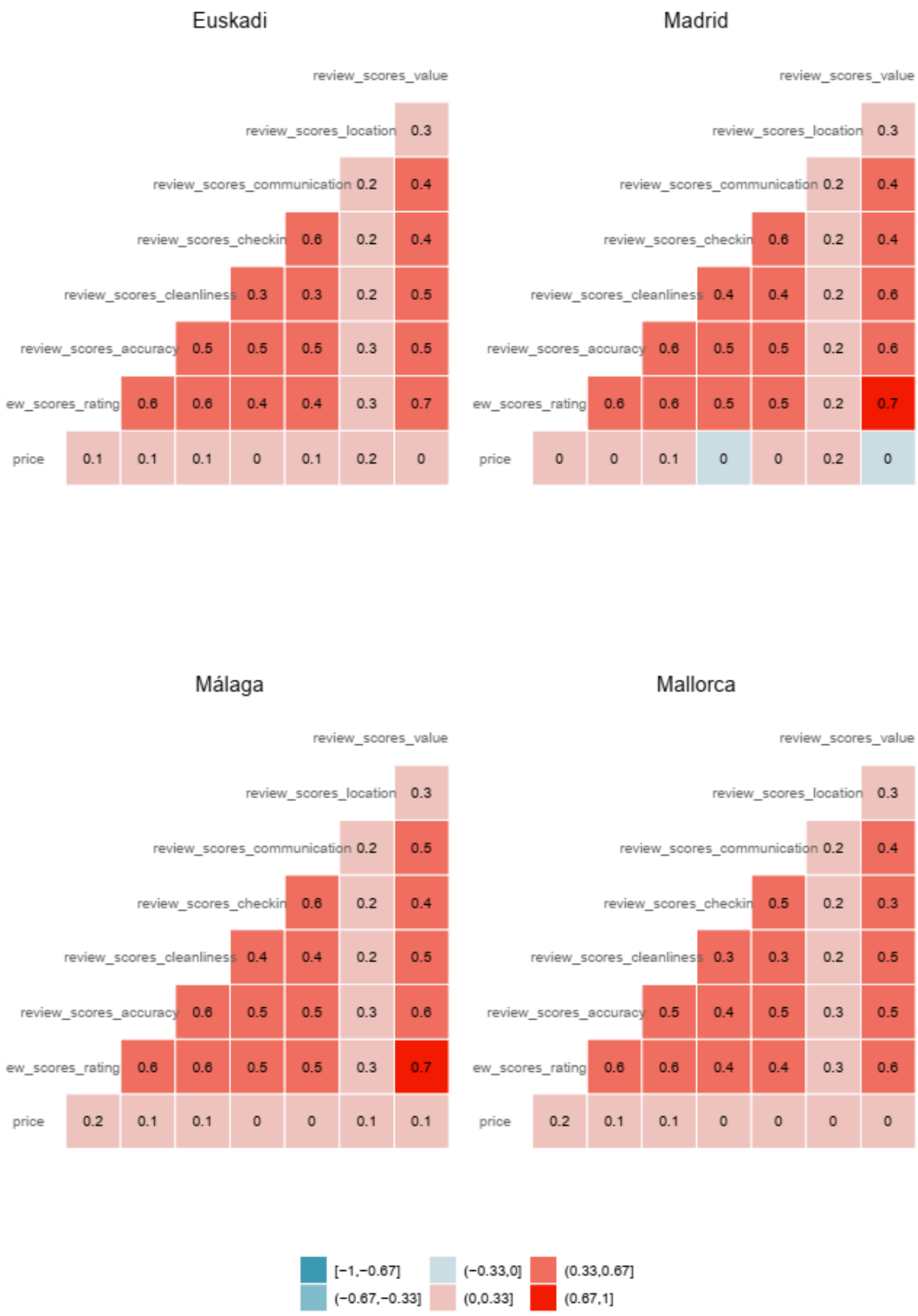


Table 22. Euskadi Geographical map.

Table 23. Madrid Geographical map.

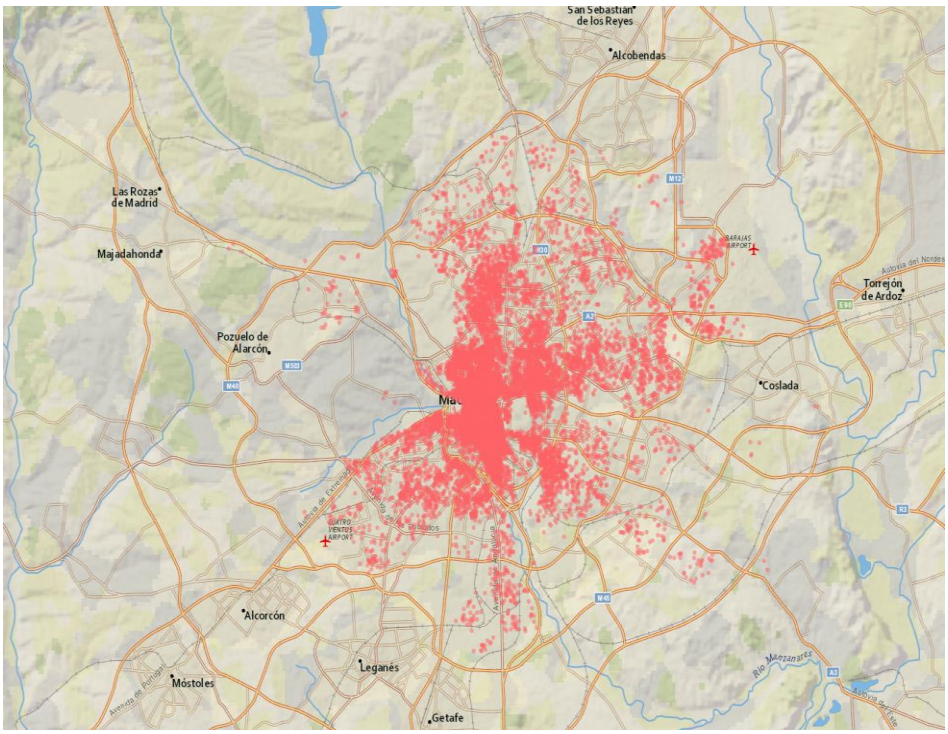


Table 24. Málaga geographical map.

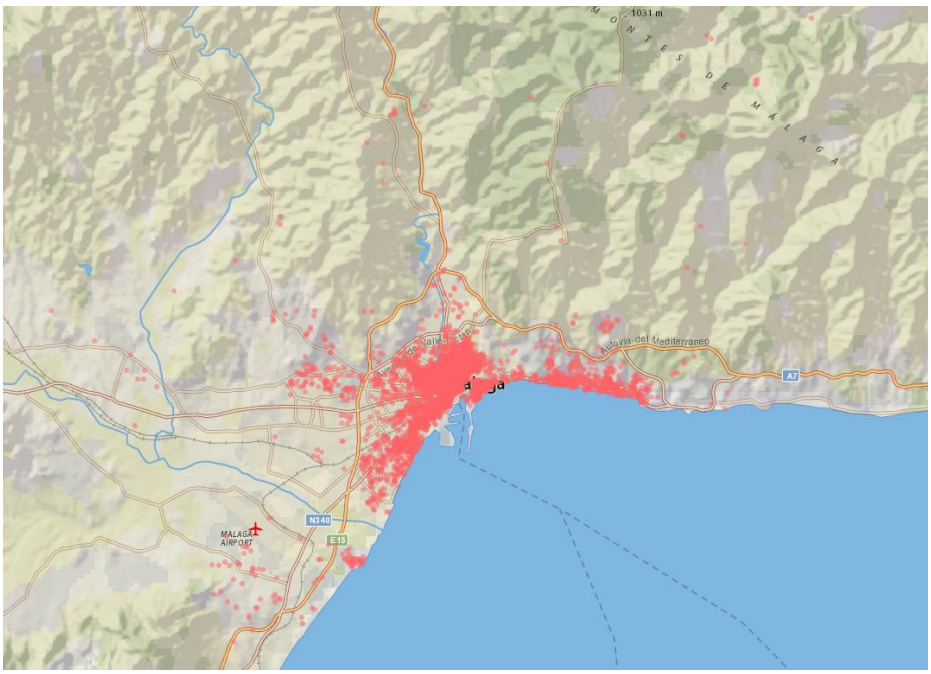
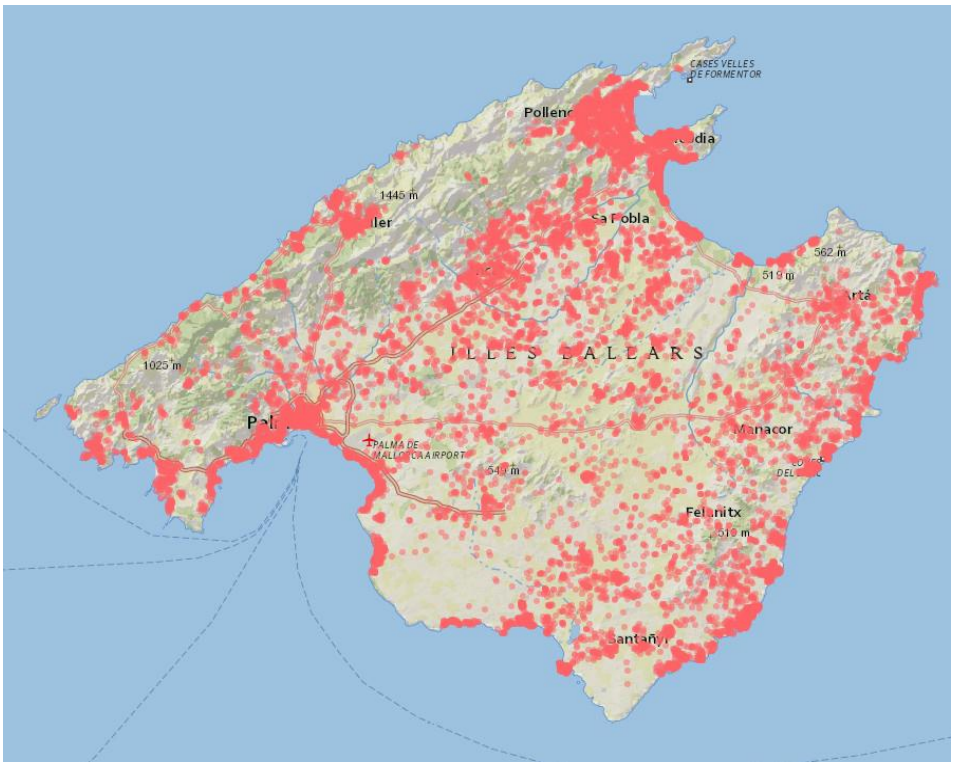


Table 25. Mallorca geographical map.



BIBLIOGRAPHY

Adamiak, C., Szyda, B., Dubownik, A., & García-Álvarez, D. (2019). Airbnb offer in Spain-Spatial analysis of the pattern and determinants of its distribution. *ISPRS International Journal of Geo-Information*, *8*(3). https://doi.org/10.3390/ijgi8030155

Airbnb New User Bookings | Kaggle. (n.d.). Retrieved May 16, 2020, from https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings

Anselin, L., Syabri, I., & Kho, Y. (2016). An Introduction to Spatial Data Analysis and Visualisation in R. *Geographical Analysis*, *38*, 5–22. https://doi.org/10.1111/j.0016-7363.2005.00671.x

As A Rare Profitable Unicorn, Airbnb Appears To Be Worth At Least $38 Billion. (n.d.). Retrieved May 15, 2020, from https://www.forbes.com/sites/greatspeculations/2018/05/11/as-a-rare-profitable-unicorn-airbnb-appears-to-be-worth-at-least-38-billion/#15d937c82741

Bion, R., Chang, R., & Goodman, J. (2018). How R Helps Airbnb Make the Most of its Data. *American Statistician*, *72*(1), 46–52. https://doi.org/10.1080/00031305.2017.1392362

BarOn, R. R. V. (1972), Seasonality in tourism – part I, International Tourism Quarterly, Vol 4, 40-64

ECB Statistical Data Warehouse. (n.d.). Retrieved May 16, 2020, from https://sdw.ecb.europa.eu/browseSelection.do?node=qview&SERIES_KEY=120.EXR.D.USD.EUR.SP00.A

Fast Facts - Airbnb Newsroom. (n.d.). Retrieved May 16, 2020, from https://news.airbnb.com/fast-facts/

Guisan, M. C., & Aguayo, E. (2010). Second homes in the Spanish regions: Evolution in 2001-2007 and impact on tourism, GDP and employment. *Regional and Sectoral Economic Studies*, *10*(2), 83–104.

How to calculate home ranges in R: Kernels. (n.d.). Retrieved May 16, 2020, from https://jamesepaterson.github.io/jamespatersonblog/04_trackingworkshop_kernels

Humphreys, A. (2019). Automated Text Analysis. *Handbook of Market Research*, *May*, 1–32. https://doi.org/10.1007/978-3-319-05542-8_26-1

Índices de precios hoteleros, índices y tasas de variación interanual por comunidades

autónomas(12156). (n.d.). Retrieved May 16, 2020, from https://www.ine.es/jaxiT3/Tabla.htm?t=12156&L=0

Inside Airbnb. Adding Data to the Debate. (n.d.). Retrieved May 16, 2020, from http://insideairbnb.com/

James Inman - Wikipedia. (n.d.). Retrieved June 25, 2020, from https://en.wikipedia.org/wiki/James_Inman

Kuethe, T., & Foster, K. (2008). A Spatial Hedonic Model with Time-Varying Parameters: A New Method Using Flexible Least Squares. *American Agricultural Economics Association Annual Meeting*, 1–22. http://ageconsearch.umn.edu/bitstream/6306/2/467673.pdf

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

What's smart about Smart Pricing? – The Airbnb Blog – Belong Anywhere. (n.d.). Retrieved May 16, 2020, from https://blog.atairbnb.com/smart-pricing/

Wilson, J. P. (John P., & University Consortium for Geographic Information Science. (n.d.). *Geographic information science & technology body of knowledge.*