Driving Factors Behind Airbnb Pricing -
A Multilinear Regression Analysis

Jonathan Herrgård, Johan Flöjs

Maj 2023
Abstract

With a high increase of users in the world’s ever expanding sharing economy, Airbnb has become a customary solution in short term rentals of accommodations. In this market, it is the host’s job to choose a pricing which sufficiently corresponds to what tenants are willing to pay. There can be multiple methods of choosing the price but this study aims to determine and evaluate which factors have a significant impact on short term rental pricing of housing and to what degree. By modelling this issue, the reader can make a better understanding of what to pay or charge for an accommodation. This study also serves as ground work for further investigations exploring nested and multi-leveled factors.

The study is limited to the Spanish short term rental market, taking a more in-depth look at the cities of Barcelona, Madrid and Palma. Moreover, listings between 2015 and 2017 are considered in the study. In the end, factors identified as significant on accommodation pricing were *Entire Home, Accommodates, Bathrooms, Review Scores Rating* etc. Some of the factors are interchangeable as they have a miniscule effect on the accommodation pricing. Conversely, *Entire Home and Accommodates* is seen as absolute necessities for the model as they, together, explain two-thirds of the variations in price.
Sammanfattning


Contents

1 Introduction 6
   1.1 Background . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
   1.2 Previous Studies . . . . . . . . . . . . . . . . . . . . . . . . . . 7
   1.3 Purpose . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
   1.4 Scope and Limitations . . . . . . . . . . . . . . . . . . . . . . . 8
   1.5 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . 8

2 Theory 9
   2.1 Microeconomical Theory of Supply & Demand . . . . . . . . . . 9
   2.2 Hedonic Pricing Model . . . . . . . . . . . . . . . . . . . . . . . 9

3 Method 10
   3.1 Data Set . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
      3.1.1 Categorical Grouping . . . . . . . . . . . . . . . . . . . . . 10
      3.1.2 Variable Creation . . . . . . . . . . . . . . . . . . . . . . . 10
   3.2 Initial Selection of Variables . . . . . . . . . . . . . . . . . . . . 10
      3.2.1 Initial Data Cleaning . . . . . . . . . . . . . . . . . . . . . 13
   3.3 Regression Model . . . . . . . . . . . . . . . . . . . . . . . . . . 13
   3.4 The 5 Assumptions in Linear Regression . . . . . . . . . . . . . . 13

4 Results 16
   4.1 Initial Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
   4.2 Linearity . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
   4.3 Residual Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . 18
   4.4 Transformation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
   4.5 Dependencies between variables . . . . . . . . . . . . . . . . . . 21
      4.5.1 Correlation . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
      4.5.2 Multicollinearity . . . . . . . . . . . . . . . . . . . . . . . . 23
      4.5.3 Handling of variables based on dependencies . . . . . . . . 23
   4.6 Analysis of outliers . . . . . . . . . . . . . . . . . . . . . . . . . 24
      4.6.1 Influence and leverage . . . . . . . . . . . . . . . . . . . . . 24
   4.7 Outliers . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
   4.8 Variable Selection . . . . . . . . . . . . . . . . . . . . . . . . . . 25
      4.8.1 Forward Selection and Backward Elimination . . . . . . . 26
      4.8.2 Revised Forward Selection and Backwards Elimination . . 27
   4.9 Final Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28
   4.10 Validation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
      4.10.1 Validation based on $R^2$, $p$-value and $t$-statistic . . . 29
      4.10.2 Cross Validation . . . . . . . . . . . . . . . . . . . . . . . . 30
   4.11 Marginal Effect of Pricing . . . . . . . . . . . . . . . . . . . . . 31
## 5 Discussion

### 5.1 Discussion of Model

- 5.1.1 Positively Influential Factors
- 5.1.2 Negatively Influential Factors

### 5.2 Model Problems and Future Studies

## 6 Conclusion

## 7 List of References
1 Introduction

1.1 Background

With the help of globalization and digitization, the travel industry has changed drastically during the beginning of the 21st century. With the world becoming smarter and more integrated, a great number of alternatives to the traditional economy instruments have arouse, changing the way we look at the economy. One of these systems is the sharing economy which facilitates peer-to-peer transactions between two parties in which services or goods is exchanged (Corporate Finance Institute, 2022).

In the sharing economy, individuals may share assets such as homes or cars. These transactions often occur through online marketplaces or platforms that connect providers with users and facilitate payment and exchange. This creates an economic environment in which individuals and businesses can share their underutilized assets or resources. The sharing economy has had a significant impact on the travel business, especially in the areas of transport and accommodation. A number of different sharing-platforms has arouse based on this system, disrupting the traditional hotel industry. One of these new challengers is Airbnb, an online marketplace for B2B (Business-to-business, B2C (Business-to-consumer or C2C (Consumer-to-consumer), enabling people or companies to rent out their homes, apartments, spare rooms or offices to people looking for a place to stay and or work.

As of 2022, Airbnb captures around 10-12 percent of the travel demand in cities such as New York, Paris and London (HotelTechReport, 2022). The open access that Airbnb provides for its users and customers leads to accommodations in a variety of different sizes, styles and locations being published on the platform. Today, the platform presents 6.6 million active listings worldwide in over 100 thousand cities and towns (Airbnb, 2022). Studies from Goldman Sachs show that the popularity of Airbnb and other peer-to-peer-platforms seems to increase, over a period of 5 years, the amount of people preferring traditional accommodation over peer-to-peer lodging nearly halved (Business Insider, 2016). Taking the sheer number of active listings into account, getting the pricing of the listing right is of utter importance for attracting customers. This especially holds true in major cities with a lot of competition.

Under the assumption of an efficient market, the pricing of these rentals can be assumed to be affected by the consumer value each accommodation offers the customer. Therefore this study will analyze so called internal factors which can be linked to consumer values. These factors include parameters such as number of bathrooms, number of beds and amenities available for use during the stay. Moreover, the profile and characteristics of the host will be taken into consideration. In this way, the study will take a consumer based approach and will be performed under micro-economic assumptions. Based on these as-
sumptions this study aim to find the key factors influencing Airbnb pricing in cities in Spain and estimate a reasonable listing price per night based on the characteristics of the accommodation.

1.2 Previous Studies

There have been several previous studies which have similar investigations on the matter. While the conclusions, method and choice of regression model varies between reports, Tong & Gunter (2020) discovered that the main positive factors in pricing are overall rating and factors indicating size whereas the main negative ones are number of reviews and distance to city center. Similarly Wang & Nicolau (2017) found that size factors such as how many people can be accommodated and number of bathrooms clearly affected the price positively. The report also concluded that review scores are a positive pricing factor meanwhile number of reviews decrease the price. Just as Wang & Nicolau (2017), a study made by Dogru & (2017) found factors affecting the experience having a positive influence on pricing. However, their study also takes cleanliness into consideration, showing guests tend to pay more for accommodations applying a mandatory cleaning fee.

In a report by Gibbs et. al (2017), they found similar positive variables as above but were surprised in their findings on the variables superhost, number of reviews and instant booking. They found that the variable superhost in many cases was insignificant and that number of reviews and instant booking had a negative impact on pricing. Further, a report by Chen et. al (2017) investigating eventual correlations between host behaviours and pricing found significant positive effects on the price when hosts respond fast and apply a moderate cancellation policy.

These previous findings will act as the basis for research in this study.

1.3 Purpose

The aim of this study is to develop and evaluate a multilinear regression model for estimating Airbnb rental prices, with the purpose of providing a comprehensive understanding of the underlying factors that drive price variations in the short-term rental market. By analyzing a data set containing a wide range of property features and local factors, the study seeks to identify the key internal factors that significantly influence Airbnb rental prices, such as property size, location, amenities and host profile. This information will not only contribute to the existing body of research on short-term rental pricing dynamics but also offer valuable insights for hosts looking to optimize their rental income and guests seeking to make informed decisions when booking accommodations through Airbnb.

Furthermore, the thesis will delve into the methodology of linear regression,
discussing its fundamental principles and assumptions, as well as the process of model selection, feature engineering, and parameter estimation. By examining the strengths and limitations of linear regression in the context of Airbnb rental price prediction, the study aims to determine the suitability and accuracy of this approach for the problem at hand.

1.4 Scope and Limitations

The study will be made under the assumption that the market for the listings included in the study is efficient. Efficient markets leads to a supply matching the demand and thus listings are assumed to be made in a fashion that caters to this assumption. Hence, if an accommodation has an abundance of demand the host of said accommodation will raise the rates of its customers. Therefore we can also assume that attractive internal factors of a property and its host will drive up the price. Actions will be taken in order to assure that only listings with at least one review is part of the study, the minimum of one review will act as insurance for an actual demand of the property in question.

Geographically, the study will be limited to major cities in Spain. This will ensure that there is enough data points in each location for the model to make adequate estimations based on the availability of data. Making the geographical limitation more generous would cause the study to become more extensive than needed. Instead, the main focus will be directed towards the characteristics of the listings allowing an effort to pin-point factors driving the pricing and then in turn be able to base estimations on these factors.

The data used in this study will contain observations in a time period from 2015 to 2017. This was a time period of normal travelling behavior and a stable economy, allowing the model to bring out the important internal factors instead of macro-economic ones such as inflation. By using a stable period the model will in turn be more generalized for use in other similar periods of normal economic activity. Note that the model will not be able to give a real time inflation adjusted price, but can still offer a valuable foundation for understanding Airbnb pricing.

1.5 Problem Statement

Initially, the problem statement of this study can be divided into two parts:

- Can a hedonic pricing model be made on Airbnb listings?
- Which are the main driving factors behind Airbnb pricing in Spain?
- Does price differ between different cities in Spain?

These statements together with previous studies will form the basis for the research done within this thesis.
2 Theory

2.1 Microeconomical Theory of Supply & Demand

A key idea in microeconomics is the supply and demand curve, which shows how the quantity of an item or service supplied and the quantity desired at various prices relate to one another. According to the law of supply, the quantity of an item or service supplied will grow as its price rises, if all other variables are equal. Contrarily, according to the rule of demand, an item or service’s quantity will decline in demand as its price rises, if all other variables are equal (Investopedia, 2023). The intersection of the supply and demand curve determines the equilibrium and a change in one or the other curves will shift this intersection.

2.2 Hedonic Pricing Model

A hedonic pricing model is a pricing model based on linear regression. The method is used to estimate the price of a product based on its underlying characteristics and attributes (Investopedia, 2021). Most commonly, the model is applied within the real-estate market to determine the underlying factors behind the price of an asset using the price as the response variable. The regression model ought to assign values and weights to every constituent element or contributing factor to calculate the overall value of the composite product. Hedonic approaches come in various forms, such as linear, non-linear, variable interaction, and other valuation scenarios with varying levels of complexity (Corporate Finance Institute, 2023), although the linear approach is applied in this study.
3 Method

3.1 Data Set

The data set used in this study is taken from Opendatasoft (Opendatasoft, 2020), which is a comprehensive open hub for sharing data. The website is known to have reliable data which is sourced from a vast amount of trusted providers. Examples of large companies using Opendatasoft are among Schneider Electric and SNCF a government based railroad company in France (Opendatasoft, 2023).

More specifically, the origin used for our data set is Airbnb’s listings page, where the data has been webscraped. For this report the observations in the data is presumed to be reliable, independent and accurate but an acknowledgment has to be made whether all data points are legitimate as the market in question can involve scams and non-serious hosts. Several measures presented below has been taken to further clean the data set. This will rid of unnecessary, incomplete and false data which makes the final model more precise.

3.1.1 Categorical Grouping

The data set includes variables that are both qualitative and quantitative. While the quantitative variables could be left in their initial form, the qualitative ones had to be sectioned into categorical groups and then represented in numerical form. After grouping together different aspects in a certain variable we represented these groups using dummy variables in order to separate the significance of the different groupings. An example of this type of grouping is the type of bed available. We narrowed several different bed types into the groups Real Bed and Other Bed. This process was done for all qualitative variables where the amount of internal categories were too many relative their number of observations.

3.1.2 Variable Creation

Two of the variables included in the initial data set, ”Features” and ”Amenities”, contained small informational texts about the features and amenities of the listing. While features covers information about the host and their Airbnb profile, amenities explains what appliances, services and facilities that are included in the stay. To represent these variables as their own we scraped these texts and created dummy variables for those that were of particular interest. Examples of variables found here are TV, Washer, AC, Superhost and more.

3.2 Initial Selection of Variables

As mentioned earlier, the data set used for this study is vast and detailed. Each data point is described by 89 separate variables, forcing us to select a subset of the data to continue working with. As the study mainly focuses on consumer values and micro-economic factors, variables full-filling these criteria will
be taken into consideration whilst creating the subset. Variables were also selected, or removed, based on domain knowledge and/or whether they were able to be quantified or not.

The initial variable selection was based on an iterative process where the selection was made in three steps. First off, variables describing the main physical factors of the accommodation were selected. Variables were then selected based on the amenities and general properties of the listing and at last, focus was directed towards the host and their profile characteristics.

After analyzing the data set at hand, the variables in the tables below were chosen as a base for the study:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Origin</th>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodates</td>
<td>-</td>
<td>1-16 people</td>
<td>Number of people the accommodation can house during the stay</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>-</td>
<td>0-8 bathrooms</td>
<td>Number of bathrooms</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>-</td>
<td>0-10 bedrooms</td>
<td>Number of bedrooms</td>
</tr>
<tr>
<td>Beds</td>
<td>-</td>
<td>0-16</td>
<td>Number of beds</td>
</tr>
<tr>
<td>Cleaning Fee</td>
<td>-</td>
<td>0-$100</td>
<td>Obligatory fee needed to be paid by guests for cleaning</td>
</tr>
<tr>
<td>Days as Host</td>
<td>-</td>
<td>11-2942 days</td>
<td>Number of days since the host published its first listing on the platform</td>
</tr>
<tr>
<td>Distance</td>
<td>-</td>
<td>0-95000m</td>
<td>Distance to city center</td>
</tr>
<tr>
<td>Host Response Rate</td>
<td>-</td>
<td>0-100%</td>
<td>Describes how many booking requests the host responds to</td>
</tr>
<tr>
<td>Minimum Nights</td>
<td>-</td>
<td>1-200 days</td>
<td>Minimum number of nights allowed per booking</td>
</tr>
<tr>
<td>Number of reviews</td>
<td>-</td>
<td>1-600 reviews</td>
<td>Number of reviews on the listing</td>
</tr>
<tr>
<td>Review Scores Rating</td>
<td>-</td>
<td>0-100</td>
<td>Average review score of the listing</td>
</tr>
<tr>
<td>Reviews Per Month</td>
<td>-</td>
<td>0.02-30 reviews/month</td>
<td>The number of reviews left on the listing per month</td>
</tr>
<tr>
<td>Security Deposit</td>
<td>-</td>
<td>0-$950</td>
<td>The deposit needed for allowance to stay on the property</td>
</tr>
</tbody>
</table>

Figure 1: Showing the quantitative variables included in the study.
<table>
<thead>
<tr>
<th>Qualitative Variables</th>
<th>Origin</th>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Amenities</td>
<td>0,1</td>
<td>AC available</td>
</tr>
<tr>
<td>Elevator</td>
<td>Amenities</td>
<td>0,1</td>
<td>Elevator installed in the building</td>
</tr>
<tr>
<td>Free Parking</td>
<td>Amenities</td>
<td>0,1</td>
<td>Free Parking on, or just outside, the property</td>
</tr>
<tr>
<td>Gym</td>
<td>Amenities</td>
<td>0,1</td>
<td>Gym available for free</td>
</tr>
<tr>
<td>Pool</td>
<td>Amenities</td>
<td>0,1</td>
<td>Pool available for free</td>
</tr>
<tr>
<td>TV</td>
<td>Amenities</td>
<td>0,1</td>
<td>TV available</td>
</tr>
<tr>
<td>Washer</td>
<td>Amenities</td>
<td>0,1</td>
<td>Washer available for use</td>
</tr>
<tr>
<td>Other Bed</td>
<td>Bed Type</td>
<td>0,1</td>
<td>Whether the bed offered is a couch, futon or pull-out sofa or not.</td>
</tr>
<tr>
<td>Real Bed</td>
<td>Bed Type</td>
<td>0,1</td>
<td>Whether a real bed is available or not.</td>
</tr>
<tr>
<td>Flexible</td>
<td>Booking Policy</td>
<td>0,1</td>
<td>Whether a flexible booking policy is used or not. Flexible policy means that guests need to cancel their booking at least 48 hours prior to check-in days to get a full refund</td>
</tr>
<tr>
<td>Moderate</td>
<td>Booking Policy</td>
<td>0,1</td>
<td>Whether a moderate booking policy is used or not. Moderate policy means that guests need to cancel their booking at latest 30 days before check-in to get a full refund</td>
</tr>
<tr>
<td>Strict</td>
<td>Booking Policy</td>
<td>0,1</td>
<td>Whether a strict cancellation policy is used or not. Strict policy means that guests need to cancel their booking within 48 hours after it is made to get a full refund</td>
</tr>
<tr>
<td>Barcelona</td>
<td>City</td>
<td>0,1</td>
<td>Whether the property is located in Barcelona or not.</td>
</tr>
<tr>
<td>Madrid</td>
<td>City</td>
<td>0,1</td>
<td>Whether the property is located in Madrid or not</td>
</tr>
<tr>
<td>Palma</td>
<td>City</td>
<td>0,1</td>
<td>Whether the property is located in Palma or not</td>
</tr>
<tr>
<td>Checkin 24</td>
<td>Features</td>
<td>0,1</td>
<td>Whether it is possible to check in 24/7 or not</td>
</tr>
<tr>
<td>Exact Location</td>
<td>Features</td>
<td>0,1</td>
<td>Whether the exact location of the property is revealed before booking or not</td>
</tr>
<tr>
<td>Instant Booking</td>
<td>Features</td>
<td>0,1</td>
<td>Whether it is possible to book without conversation with host before confirmation.</td>
</tr>
<tr>
<td>Superhost</td>
<td>Features</td>
<td>0,1</td>
<td>Whether the host is a superhost or not</td>
</tr>
<tr>
<td>Apartment</td>
<td>Property Type</td>
<td>0,1</td>
<td>Whether it is an apartment or not</td>
</tr>
<tr>
<td>House</td>
<td>Property Type</td>
<td>0,1</td>
<td>Whether it is a house or not</td>
</tr>
<tr>
<td>Entire Home</td>
<td>Room Type</td>
<td>0,1</td>
<td>Whether the accommodation is a entire home or not</td>
</tr>
<tr>
<td>Private Room</td>
<td>Room Type</td>
<td>0,1</td>
<td>Whether the accommodation is just a private room</td>
</tr>
<tr>
<td>Shared Room</td>
<td>Room Type</td>
<td>0,1</td>
<td>Whether the accommodation is a shared room or not</td>
</tr>
</tbody>
</table>

Figure 2: Showing the qualitative variables included in the study.
3.2.1 Initial Data Cleaning

After removing problematic variables, the vast majority of observations had a complete data set, and only approximately 1% were missing data. Because of this minuscule discrepancy we either removed those observations or changed the data to something more appropriate. Some data points were also removed on the basis of the scope of the study.

Initially, all data points with a missing price was removed from the data set as the study aims to predict the price of listings, hence making the price required for each data point. Listings with prices below $10 were also removed as these were seen as non-serious and or too specific in their nature. Example of housing types in this price range were caves, tents and kayaks which are not of interest to this study. Therefore any data point with a property type that could not adequately fit into the groups "House" and "Apartment" were also removed. Listings that required a minimum stay of more than a year were also removed as we limit this study to more short term housing.

Furthermore, we excluded all observations which did not have recorded data on number of accommodates and bathrooms as these were often missing other data as well. To further make sure that the listings were truthful and ready to be rented out, not just a pricing test, we removed the observations which had missing data on the host’s response rate and response time. Lastly we decided to change the missing data on security deposit and cleaning fee to 0. This was a reasonable measure as there were no observations which already had 0 under these variables, suggesting that much of the missing data was actually 0.

3.3 Regression Model

In this study, we will use a multiple linear regression model which utilizes a least squares approach. The method has been prevalent in previous works on the matter (Wang & Nicolau, 2017) & (Gibbs et. al, 2017), and fits well under the hedonic pricing assumption. As the aim of the study is to provide a pricing model, the response variable will be price per night.

3.4 The 5 Assumptions in Linear Regression

There are several assumptions that have to be made when applying a linear regression model to a problem in order to obtain feasible predictions and these will have to be validated. The assumptions are (Statisticssolutions, 2023):

- **Linearity**: The assumption that there are linear relationships between the dependent variable and independent variables.
- **Independence**: The assumption that the observations in the data set are independent from each other.
• **Homoscedascity**: The assumption that there is a constant variance of the errors across all levels of the independent variables.

• **Normality**: The assumption that the errors follow a normal distribution with mean zero.

• **No Multicollinearity**: The assumption that the independent variables do not have a high correlation with each other.
4 Results

4.1 Initial Model

Figure 3: Illustrating all included variables, their estimated coefficients, errors, t-values and p-values as well as model R-square and F-statistics in the initial model.
The initial model contains 33 different quantitative and qualitative independent variables. Not all variables are of significance, hence further analysis will be performed to remove some or all of them. At this point the model reaches an Adjusted R-squared of 0.5593 which indicates that the variance in price is decently explained by the independent variables (Montgomery, 2012, p.87). Although this value of Adjusted R-squared can be improved by additional measures after further variable selection and analysis.

By looking at the coefficients in Figure 3, we can evaluate how reasonable they are using domain knowledge. For example, Accomodate’s coefficient of 6.15 tells us that there is about a $6 increase for every extra person the housing accommodates. This is reasonable enough to not be suspicious of the estimates. The only significant variable where the coefficient seems peculiar is Washer which is estimated to decrease the price by $3.5. This variable could still be useful in combination with other variables and is therefore kept at this stage.

One might note that the variable Palma is missing in Figure 3. This is due to Palma being the baseline for the model, meaning that the variables Barcelona and Madrid explain the difference in price compared to Palma.

## 4.2 Linearity

To confirm the assumption of linearity we have plotted our quantitative variables one by one against the actual pricing (Figure 4). This will give a visual of how linear their relationships are and is expected to show a band moving in a common direction. Qualitative variables on the other hand can not be checked using this method as the variables only take on True or False values of 0 and 1. These do not need to be checked as the categorical variables are always linearly related to the response (Bookdown, 2023).
Figure 4: Illustrating variables plotted against Price.

Figure 4 shows that some of the variables have a linear relationship whereas others seem more random, for example Days as Host or Host Response Rate which are both insignificant. At this point we evaluate these relationships to be linear enough to continue using a linear model approach.

4.3 Residual Analysis

Figure 5: Illustrating studentized residuals on the left and studentized residuals plotted against fitted values on the right (After transformation).

In Figure 5 we have the residuals plotted against index and the fitted values from the model. The left hand side shows that the residuals, or errors, are centered
around zero. The right hand side shows that the residuals are greater when the predicted price is increased. This means that the errors are not constant over all levels, indicating heteroscedasticity.

Heteroscedasticity directly oppose one of our stated assumptions to use the model and is a cause for concern. This can be corrected by transforming the data to make the errors more normally distributed. By observing the Normal QQ-plot in Figure 6 we can clearly see how the normality assumption of errors are not fulfilled and a transformation will be necessary.

![Normal QQ-plot of the initial model](image)

**Figure 6: Illustrating the Normal QQ-plot of the initial model.**

### 4.4 Transformation

Due to the data not full filling all the said requirements for linear regression, a transformation was made. The box-cox transformation resulted in a $\lambda = 0.010101010$. As this value is approximately 0, a log-transformation of the response variable was used instead for better understanding (Statistics How To, 2022). The log-transformation led to the following equation:

$$\log(Price) = \beta X$$
After transforming the response variable in the data set, we can see that the new residuals are more centered around zero and that they are more evenly spread across all levels of fitted values (Figure 8).

The new Normal QQ-plot (Figure 7) is also improved, showing a betterment in the normality of the errors. Hence, we have a model which better suits the homoscedasticity and normality in the linear regression assumptions presented earlier.
In the case of linearity, the transformation makes the linear dependencies more clear on some of the variables, whereas a few are still showing random patterns seen in Figure 9.

### 4.5 Dependencies between variables

#### 4.5.1 Correlation

Correlation between individual regressors might make it troublesome to distinguish on which level each predictor contributes to the response variable. Thus removing any regressors with high correlation is of importance. As we can see in the figure presented (Figure 8), some of the variables included in the model highly correlates with each other. According to theory, an absolute value of the correlation coefficient between two regressors exceeding 0.5 is of large scale and thus, one of the coefficients should be removed (Andrews University, 2005).
As stated in figure 10, all three of Accommodates, Beds and Bedrooms seems to be highly correlated to each other. Examining their respective correlation with the response variable, Price, shows that Beds and Bedrooms possibly could be removed from the model due to Accommodates having a greater level of correlation with Price than the pair of them.

Further examination of the correlation matrix shows a high correlation between Number of Reviews and Reviews per Month. Removing one of these in favor of the other might prove beneficial for the model.

High levels of correlation can also be spotted whilst analyzing the variables describing the cities included in the study. As the study takes three different cities into consideration, two of them is needed as regressors in the model, whilst the third act as baseline. Given the binary values and the relatively low amount.
of observations for the baseline city Palma, it is not unexpected to see high correlation between Barcelona and Madrid. This is because 89% of the listings not located in Barcelona is located in Madrid. For the purpose of this study, both Madrid and Barcelona will be a part of the final model despite their correlation level.

4.5.2 Multicollinearity

To check for Multicollinearity, the variance influence factor (VIF) is used. VIF is a measurement used to quantify the level of multicollinearity in a regression model. As a measure, VIF calculate how much the variance of a certain regressor increases under influence from correlations between the regressors of the model (Montgomery, 2022, p296). Usually, values over 10 are seen as high and thus 10 marks a possible cut-off for VIF-values. Looking at figure 11, both Private Room and Entire Home exceeds this limit. However, as both these variables originates from the same categorical variable, the removal of Private Room is enough to lower the value of Entire Home to a more acceptable value of 1.427347 and therefore Private Room could be removed from the model. Accommodates, Bed and Bedrooms are other variables showcasing VIF-values which might be higher than desirable, however further examination is required.

![Figure 11: Illustrating the VIF-values of the different variables in the initial model.](image)

4.5.3 Handling of variables based on dependencies

Based on the examination of both correlation and multicollinearity, it is evident that removal of some variables is beneficial for the model. Due to a combination of high correlation with other variables and VIF-values that in some form separated them from the rest, the following variables were removed:

- Beds
- Bedrooms
- Number of Reviews
- Private Room
4.6 Analysis of outliers

4.6.1 Influence and leverage

Analysis of the influence and leverage was performed using *Cook’s Distance* and *hat-values*. The results for both of these measures is presented in the figures down below. In short, *Cook’s Distance* measures the effect of deleting a single observation on the estimated coefficients of the regression model (Montgomery, 2012, p.220) whilst the the diagonal values of the hat matrix represent the leverage of each observation in the data set. Leverage refers to the degree of which an observation affects the regression model’s fitted values.

![Hat Values vs. Price](#)  ![Cook's Distance vs. Price](#)  ![Hat Values vs. Cook's Distance](#)

**Figure 12:** Illustrating the hat values to the left, cook’s distance in the center and hat values plotted against cook’s distance on the right (initial transformed model).

After examination, a number of influential points were detected. Despite there being a certain amount of points with values separating them from the rest, none of the data points exceeded the direct *Cook’s Distance* cut-off-limit of 1 (Pennsylvania State University, 2018). Further investigation was made and the *Cook’s Distance* for each point was plotted against the hat value for each point to identify problematic data points. Points with a hat value larger than 0.1 or a *Cook’s Distance* larger than 0.1 was then removed from the data as they were seen as harmful for the fit of the model (Figure 12). When checking the data points which were removed, it could be noted that some of their values were extreme in contrast to the others. For example some points had a distance from the city center of 80 km, indicating that the housing in reality is in another city than stated. Similarly some listings had unrealistically low prices compared to the amount of people it accommodated. These data points were therefore removed on a domain knowledge basis.
4.7 Outliers

*COVRATIO* is a statistical measure that compares the covariance matrices of two sets of variables and can be used for detection of outliers in the dataset. Individual points with a *COVRATIO* less than 1 are seen as harmful for the model (Montgomery et. al, 2012, p.219). The analysis of the *COVRATIO* was based on the scatter plot down below.

![Covariance Ratio Plot](image)

Figure 13: *Illustrating the covratios for each observation (initial transformed model).*

Analyzing the plot, a number of different points has a value lower than one and are thus damaging the precision of the model. However, to not exclude too many points, a limit was set at 0.8. Any point with a value exceeding that limit was then removed.

4.8 Variable Selection

Variables for the final model was selected based on two different methods, backwards elimination and forward selection. Both methods were evaluated based on three different parameters, adjusted R-squared, BIC and Mallow’s CP.
4.8.1 Forward Selection and Backward Elimination

Figure 14: Illustrating adjusted R-squared, BIC and Mallow’s CP per variable added in the forward selection process.

When maximizing the adjusted R-squared, Forward Selection included all variables in the model with a maximum adjusted R-squared value of 0.7184. In terms of BIC, forward selection only included 21 variables in the model with a minimum value of -26000. Mallow’s CP was minimized with a model containing 26 variables giving a value of 22. However, there is a clear pattern showing in all three of the graphs above, the improvements seem to stagnate after the 11th variable is selected. This pattern is also eminent whilst analyzing the figure down below (Figure 15).

Figure 15: Illustrating the best combinations of variables depending on adjusted R-squared, BIC and Mallows’s CP from the forward selection process.

Clearly, a number of different combinations result in the same values regarding adjusted R-squared, BIC and Mallow’s CP. This is indicating that the model could be describing the response variable just as well with fewer variables. Further investigation was performed with the help of backwards elimination.
Observing figure number 16 above, it can be seen that the function describing each of the measurements start to plateau after the addition of the 11th variable. This is indicating that the predictive ability of the model is increasing steadily up until 11 regressors are selected for the model. When adding one more regressor to the model, neither the adjusted R-squared, BIC or Mallows’ CP improves in large scale.

4.8.2 Revised Forward Selection and Backwards Elimination

Based on the results from Forward Selection and Backwards Elimination, a revised version was made to implement 11 as the maximum amount of variables. In Figure 18, the new forward selection and backward elimination tables show the 11 most important variable choices based on adjusted R-squared, BIC and Mallows’ CP.
As seen in Figure 18, both selection processes gave the same result with an adjusted R-squared of 0.7189, BIC of -26000 and Mallow’s CP of 330. Hence our final model includes the given variables in addition to Madrid, which is kept to be able to analyse the effects on price of the city.

Table 1. Illustrating the best combinations of variables and their addition to adjusted R-squared, BIC and Mallows’s CP.

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable</th>
<th>$R^2_{Adj}$</th>
<th>$C_p$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Entire Home</td>
<td>0.5021</td>
<td>15214</td>
<td>-14987</td>
</tr>
<tr>
<td>2</td>
<td>Accommodates</td>
<td>0.6152</td>
<td>6838.9</td>
<td>-20547</td>
</tr>
<tr>
<td>3</td>
<td>Barcelona</td>
<td>0.6463</td>
<td>4582.0</td>
<td>-22324</td>
</tr>
<tr>
<td>4</td>
<td>Bathrooms</td>
<td>0.6614</td>
<td>3474.6</td>
<td>-23249</td>
</tr>
<tr>
<td>5</td>
<td>Reviews Scores Rating</td>
<td>0.6707</td>
<td>2787.7</td>
<td>-23840</td>
</tr>
<tr>
<td>6</td>
<td>Reviews per Month</td>
<td>0.6884</td>
<td>2055.9</td>
<td>-24490</td>
</tr>
<tr>
<td>7</td>
<td>Cleaning Fee</td>
<td>0.6884</td>
<td>1486.1</td>
<td>-25009</td>
</tr>
<tr>
<td>8</td>
<td>AC</td>
<td>0.6947</td>
<td>1020.7</td>
<td>-25441</td>
</tr>
<tr>
<td>9</td>
<td>Distance</td>
<td>0.7007</td>
<td>587.64</td>
<td>-25850</td>
</tr>
<tr>
<td>10</td>
<td>Pool</td>
<td>0.7027</td>
<td>427.09</td>
<td>-25999</td>
</tr>
<tr>
<td>11</td>
<td>Security Deposit</td>
<td>0.7040</td>
<td>332.56</td>
<td>-26084</td>
</tr>
</tbody>
</table>

4.9 Final Model

Our final transformed model is presented below. The beta-coefficients can be found in Figure 17 under estimates. The X-values are the observation values for each corresponding variable.

$$Log(Price) = \beta_0 + \beta_{Accomodates} \cdot x_1 + \beta_{Bathrooms} \cdot x_2 + \beta_{CleaningFee} \cdot x_3$$
\[ + \beta_{\text{SecurityDeposit}} \cdot x_4 + \beta_{\text{MinimumNights}} \cdot x_5 + \beta_{\text{ReviewScoresRating}} \cdot x_6 \\
+ \beta_{\text{ReviewsPerMonth}} \cdot x_7 + \beta_{\text{AC}} \cdot x_8 + \beta_{\text{Pool}} \cdot x_9 + \beta_{\text{EntireHome}} \cdot x_{10} \\
+ \beta_{\text{Distance}} \cdot x_{11} + \beta_{\text{Madrid}} \cdot x_{12} + \beta_{\text{Barcelona}} \cdot x_{13} \]

Figure 19: Illustrating all included variables, their estimated coefficients, errors, t-values and p-values as well as model R-square and F-statistics for the final model.

4.10 Validation

4.10.1 Validation based on $R^2$, $p$-value and t-statistic

The adjusted $R^2$ for the final model was 0.7052, compared to the initial model with a adjusted $R^2$ of 0.5593. This shows that the final model can explain 71% of the variation in the pricing.

Looking at the t-values for each regressor in the final model, they all sit outside the interval $(-2, 2)$, which indicates that the estimated coefficient is unlikely to have occurred by chance alone, and is therefore considered statistically significant. Due to the fact that each absolute value of the t-statistic is greater than
the critical value, we can reject the null hypothesis that the coefficient is equal to zero and conclude that there is evidence of a significant relationship between the regressor and the dependent variable. The only exception is for the variable Madrid.

Testing for p-value, results show that there is strong a relationship between the model and the response variable. The measured $p$-value for the model was $2.2 \cdot 10^{-16}$. Evaluation of each individual regressor yields the same result, meaning that the inclusion of each variable in the model has a significant impact on the outcome variable and that we can be confident that this relationship is not due to chance. Once again, the only exception is for the variable Madrid.

4.10.2 Cross Validation

A k-fold cross validation was made in order to test prediction errors on 10 different folds, or subsets, of data. This method allows the testing of several models and data by leaving out one fold at a time as testing data. The cross validation gives a good measure of how under or over-fitted the model is at various number of variables by measuring the errors of predictions.

As seen in Figure 20, once again we find diminishing returns around 11 variables but that the errors keep going down when including up to 24 variables. If the max amount of variables is set to 11 we can see in Figure 21 that the best model generated includes the same variables as in our final model, except for our inclusion of Madrid for city comparison.

If all variables should be included is up for discussion. The question is if the extra complexity of another 13 variables is worth the decrease of $\pm 1\%$ in the mean prediction errors.
4.11 Marginal Effect of Pricing

Calculation of the marginal effect each variable had on the pricing was made by raising $e$ to the power of the corresponding regressor, multiplied with the value each variable can take. To exemplify, calculation of how much a review score of 90 would add to price was done in the following fashion:

$$\text{Marginal Increase} = e^{\beta_{\text{Review Scores Rating}^{90}}} \approx 1.96,$$

meaning that a review score of 90 would increase the price with 96% compared to a accommodation of exactly the same type but with a review score of 0. Figure 20 down below shows all these marginal increases per quantitative variable.
Figure 22: Illustrating marginal increase in price per added unit for each quantitative variable.

Using the graphs above, the average marginal increase per quantitative variable were then calculated. This was done by taking $e$ raised to the power of beta for the corresponding regressor, multiplied with the average value of the corresponding variable. To exemplify, the average marginal increase for *accommodates* was calculated accordingly:
Average Marginal Increase

\[ e^{\beta_{\text{Accommodates}}(\text{Average number of accommodates})} \approx 1.242, \]

meaning that, all other things equal, a accommodation housing 3.53 people (The average number of accommodates) tends to be priced 24.2% higher than a accommodation only housing 1 person. The average for all quantitative variables is marked in the figure above where the vertical, and horizontal, red line cross each other. Table number 4, down below, show these averages for each quantitative variable.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Average Marginal Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodates</td>
<td>24.2%</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>18.6%</td>
</tr>
<tr>
<td>Cleaning Fee</td>
<td>5.49%</td>
</tr>
<tr>
<td>Security Deposit</td>
<td>2.56%</td>
</tr>
<tr>
<td>Minimum Nights</td>
<td>-1.35%</td>
</tr>
<tr>
<td>Review Scores Rating</td>
<td>94.3%</td>
</tr>
<tr>
<td>Reviews Per Month</td>
<td>-7.15%</td>
</tr>
<tr>
<td>Distance</td>
<td>-7.29%</td>
</tr>
</tbody>
</table>

Table 2. Showing the average marginal increase for quantitative variables included in the final model

The marginal increases for each qualitative variable was calculated using the same method as for the quantitative ones. However, the qualitative variables can only take the values 1 or 0. Analyzing table 5 down below, we can see that entire home has the biggest impact on the price, meaning that if the accommodation is an entire home, and not shared with someone else, the model estimates the price 94.5% higher, all other things equal.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Marginal Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>12.9%</td>
</tr>
<tr>
<td>Pool</td>
<td>15.8%</td>
</tr>
<tr>
<td>Entire Home</td>
<td>94.5%</td>
</tr>
<tr>
<td>Madrid</td>
<td>-1.13%</td>
</tr>
<tr>
<td>Barcelona</td>
<td>25.5%</td>
</tr>
</tbody>
</table>

Table 3. Showing the marginal increase for qualitative variables included in the final model
5 Discussion

5.1 Discussion of Model

Our final choice of model includes 12 independent variables of which 11 of them are statistically significant with a p-value of $\approx 0$. The 12th variable Madrid was kept to be able to distinguish the difference between the three cities used in the data and what interaction effects are present based on city. Upon further investigation of this insignificance, it could be determined that there is no statistical significant difference in price between Palma and Madrid, hence making their general price-effect almost the same.

5.1.1 Positively Influential Factors

We can see that Barcelona increases the general price by $e^{0.23} - 1 \approx 26\%$ but that the variations in prices are large and overlap between the cities, indicating that a hedonic model is more suitable than a set price level for each city when explaining the pricing. Similar work (Tong, Gunter, 2020) came to the same conclusion, stating that no equilibrium could be found for prices in Barcelona, Madrid and Seville. They argue that different listings are "highly but not perfectly substitutable", which fits well with the hedonic pricing model where a host can add factors to increase pricing. Accommodates and Bathrooms are both highly positively influential on the price which comes to no surprise. These variables explain the size of the accommodation well and size is known to increase prices in more conventional markets. Previous studies (Tong, Gunter, 2020; Wang & Nicolau, 2017) has seen comparable results of size factors being among the most positively influential variables. In our study we found that accommodates and bathrooms on average increase the price by 24% and 19% respectively, all other factors equal. These were therefore among the top four biggest contributors to the average price. It is also shown in Figure 20 that accommodates and bathrooms had the biggest maximum increases in price of 300% and 190% among all the variables. Although the biggest average increases were given by the review scores rating and if the entire home was rented.

Studies (Tong & Gunter, 2020; Wang & Nicolau, 2017) show that review scores are among the most influential aspects of a listing which rings true to this study. The average increase in price by review scores were 94%, all other factors equal. This is because the average listing has an exceptionally high score of 90% meanwhile a score of 0% gives the baseline. The review scores can be interpreted as an insurance that the host delivers what is promised, hence a sort of insurance for customer satisfaction which ought to be paid for. The second biggest average increase came from the variable "Entire Home" which states if the entire accommodation is rented instead of just a shared or private room. It is reasonable to assume that a tenant is ready to pay the average 95% marginal increase our model came up with in order to not live together with the host.
For amenities such as AC and pool the average price increase was 13% and 16% respectively, all other factors equal. As expected people are willing to pay for extra amenities, especially those that fit the travel profile well. Higher Cleaning fees and security deposits also had a small positive influence of 5% and 3% respectively, all other factors equal. This could be explained by the fact that a more expensive housing, either because it is bigger or more luxurious, will take on a more tedious cleaning job and higher risk. The variables can still help to provide accuracy for the model if a price is to be predicted, even though the relationship is non-casual. Another study (Dogru & Pekin, 2017) argues that the relationship between cleaning fees and price might follow from the tenant’s desire for the housing to be properly cleaned, possibly because a higher fee indicates that a cleaning firm is hired. This idea explains a more causal relationship between the two.

5.1.2 Negatively Influential Factors
The two biggest average marginal decreases in price came from the variables Reviews Per Month and Distance, which had a decrease of 7% each, all other factors equal. As distance explains how remote the accommodation is, it is reasonable that it has a more negative influence the greater the distance to the city center, as in most traditional housing markets. Reviews per month being a negative influence on the other hand is more surprising but is in part confirmed by several other studies (Tong & Gunter, 2020; Wang & Nicolau, 2017; Gibbs et. al, 2017) which looked at the negative impact by number of reviews. These two variables have a high correlation and can be interpreted similarly. The negative impact likely stems from the fact that cheaper accommodations are rented out more often, sequentially leading to more reviews. Further, hosts who use Airbnb as a main source of income might lower prices if there is a risk of not finding a tenant, therefore pushing the price down on accommodations which are rented more often. This suggests a reversed causal relationship where the price is affecting the amount of reviews per month and not the other way around. The variable Minimum Nights also had a negative influence on price which is reasonable in normal economic environments. This is because a long term contract lessens the risk for the host and in turn the host is willing to accept a lower return. On average Minimum Nights decreases the price by 1%, which seems rather low. This is because the vast majority of Airbnb listings are for short term rentals, many times just a few nights. Figure 20 shows that listings that require around 180 nights are estimated to decrease the price by around 50%.

5.2 Model Problems and Future Studies
Fully explaining the variations in price using a model based around linear regression and quantitative variables is most likely impossible. If a model were to be perfect, it would need to include parameters which are hard to put a measurable value on. Trying to build a model taking such information into
consideration would be complex due to both data availability and evaluation of emotional connections and with certain aspects. Therefore, capturing the influence of such external factors is difficult using “easy to measure” variables alone. It is not unreasonable to believe that these subjective and emotional factors such as styling, atmosphere, mood and spirit can explain the remaining variance in the price. To further investigate this matter, data would first have to be collected on how tenants subjectively perceived the housing. This might, to an extent, already be covered by the rating variable but there is possibly more detailed data which could impact the model positively. However, as rating is a very subjective measurement, and might depend on the anticipation before hand, opinions could change after a second visit, purely based on the customers prior knowledge.

When it comes to more quantitative measurements, one that is missing is the area of the accommodation. The data set used for the study included a measurement in square feet, but due to 85% of the listings missing info in this regard it was simply not considered when building the model. Instead, the size of the accommodation is purely represented by the number of accommodates and bathrooms which can skew the fit of the model and its predictability. Leaving out such variable creates problems as the sizing tends to be highly correlated to price when evaluating the housing market, hosts with bigger apartments could thus be more likely to increase the rent.

Although distance to the city center is taken into consideration, proximity to public transport and tourist attractions is excluded from the model. A study made by Kholodilin and Maksimova (2019), suggests that public transport has a significant impact on rental pricing on the long-term market, leading one to believe that the same conclusions possibly could be drawn even within the short-term rental market. Further, when Kim et. al (2020) researched the Chicago hotel market, they found that short distances to tourist attractions had a significant positive impact on pricing. From the direct competition between the hotel market and companies such as Airbnb, it is likely that Airbnb customers value the same parameters when booking. Especially, this might hold true in tourist oriented cities like Barcelona where, according to Murrillo et al. (2013), the accommodation sector receives 93% of their annual revenue from visitors.

Future studies on the subject could benefit from taking seasonal changes into consideration. The data used in this study only has the price for one, specific date, for each listing. This means that seasonal changes in pricing are not considered in the model. During peak travel seasons, such as summer or major holidays, the demand for Airbnb listings might increase, leading to higher prices than in general. Conversely, during periods with less traveling and tourism, prices might tend to fall. Additionally, Spain is a country filled with both cultural celebrations and different sporting events which may add temporarily price fluctuations under a selected few days. For example, during peak travel seasons, such as summer or major holidays, demand for Airbnb listings may be
high, resulting in increased prices. Conversely, during low season, demand may be lower, leading to lower prices. Additionally, certain regions may have specific peak seasons based on local events, such as music festivals, sporting events, or cultural celebrations, which can impact pricing.

More in-depth comparisons and direct separation of cities is factors that could increase the predictability of the model and should be considered in future studies. As earlier discussed, our model shows that if the accommodation is located in Barcelona the prices increase compared to if the location is in Madrid or Palma. However, building three separate models for each city would most likely increase the ability of explaining variations in price as guests might have different priorities when renting in different cities and thus value factors such as amenities differently. Furthermore, external factors, such as local economy, could impact the pricing, causing the model to be skewed if cities of various types and economies is bunched together into one model.
6 Conclusion

This study explores how different pricing factors affect the price of Airbnb accommodations using a data set of 88 independent variables. After reaching a final model with 12 independent quantitative and qualitative variables, we gained more insight in how pricing is determined and its complexities. Using this model, a price can either be predicted or estimated for commercial or private use in order to determine price of an accommodation in Barcelona, Madrid or Palma which is to be rented out.

It was clear that a hedonic pricing model can be made as the final model included 11 variables of physical and non-physical attributes which significantly affected the price. Further we could also make a distinction between cities, confirming our beliefs that different cities have different price levels. The main driving factors in pricing, not including city, were size factors such as number of accommodates, bathrooms and if the housing was private as well as review ratings and if the accommodation had a pool or not. In spite of the fact that this study is limited, it still answers our problem statements adequately. Although, further research can be made to explore nuances in how different factors explain price at various price levels and how different factors affect price in contrasting cities, which will give a much deeper understanding of pricing in short term rental markets.
7 List of References


Andrews University, Correlation Coefficients (no date). Available at: https://www.andrews.edu/calkins/math/edrm611/edrm05.htm


OpenDataSoft (2023) *Customers - Opendatasoft*. Available at: https://www.opendatasoft.com/en/customers/

PennState (2018). *Identifying Influential Data Points*. Available at: https://online.stat.psu.edu/stat462/node/173/: :text=Using%20Cook's%20distance%20measures.&text=Here%20are%20the%20guidelines%20commonly,quite%20likely%20to%20be%20influential.


