Lisbon School of Economics & Management
Universidade de Lisboa

Programming and Data Science

1st Semester, S14

Students:

Alice Castro – 58563

Daniela Gonçalves – 57261

Filipa Machado – 58512

Professor: Carlos J. Costa

Abstract

Airbnb is an innovative platform in terms of how renters and owners interact with each other, and it is worth noting that it has boosted the rental of properties for a short period. The characteristics of this business domain mean that the rental price of each of the properties is defined by several factors. In this paper, an evaluation was made of which factors can influence the price of each property as well as the best models to predict their value.

Keywords: Machine Learning, Regression, Seaborn Correlation Heatmap, Histogram, Box Plot, Scatter Plot, Airbnb.

Table of Contents

Table of Figures

1. Introduction

This work is presented in the scope of the Programming and Data Science curricular unit, part of the master's degree in Management Information Systems taught at the Lisbon School of Economics and Management by Professor Carlos J. Costa.

Airbnb is a platform created in 2008 that serves as a marketplace for advertising and renting properties which can take multiple topologies, such as single rooms or full houses, allowing the interaction between hosts and guests and the exchange of payments. Although it was created in the USA, it is currently present in several parts of the world. For this paper, Amsterdam was the city selected as the focus of the study. This paper is developed through a data set from 2019, which allows the analysis of the several variables provided by this popular hosting application.

The purpose of the work performed in this paper is to analyse the 2019 Airbnb data set in Amsterdam, understand what the relationships between the variables are and evaluate which factors influence the property renting price.

The variables that were considered most interesting to analyse are those that result from Airbnb user reviews and those that correspond to the price of each Airbnb per night. Through these variables, an attempt was made to answer the following questions: Is there a direct relationship between the price and the quality of the property? Does the number of reviews impact the review value? Does the location of the city influence the price of the property? For this purpose, the CRISP-DM data science life cycle was followed.

2. Literature Review

The Airbnb platform provides quarterly data sets of the various parts of the world where they operate. This provision allows several students and researchers to work on real data, improving their knowledge of others. In addition, it also allows Airbnb to have several studies and articles related to their business that can be used in managing it.

This platform presented something different in 2008, as it was not the typical hotel platform, but a business domain where the way of assigning prices is independent from the typical hotel standards, where there is a marketplace where the host is entitled to set the specific price for its own property.

Being marked by the difference it represents, Airbnb was the target of study by several data scientists, to assess the best way to treat this data and draw conclusions from it.

Considering the price of the property, several studies depict the diversity of factors that can influence it. (Dhillon, et al., 2021) considers that the availability of properties in the area, the number of people looking for a property, seasonality or the day of the week are decisive factors in defining the price of each property.

Other important factors that are evaluated in the article (Islam, et al., 2022) are geography and the host's description of the property. Geography was considered an important factor as the location of the property can also influence the price set, as this represents proximity to places of interest increasing, or not, the visitor's interest in the property.

Several articles have tried to find the best machine-learning model to make price predictions for properties on the Airbnb platform. The authors come to different conclusions depending on how the data is handled and the information that the data sets have. Examples of different perspectives are (Dhillon, et al., 2021) who considered that the best model to make predictions would be Random Forest and (Li, Zhu, & Xie, 2020) who considered the Bagging model.

3. Empirical Work

3.1 Data Context

Background Information

The data set used in this project is a sample taken from the Airbnb platform housing registry in Amsterdam, Netherlands.

Airbnb is a property-renting business where a user can rent properties for a certain period and price. Each owner can define their price and how many people they can accommodate. They can also define how many extra guests are accepted, the value for each, and the minimum number of nights accepted.

Airbnb's main objective is to ensure that all the guests are well treated and that there are few complaints about the service. For that reason, the clients' reviews in the comments are one of the best ways to evaluate the service in a determined location.

Airbnb also attributes a Super host badge to the best-rated users so that they have benefits like appearing first when another user searches for a place near their property.

Assessing the situation

With the evaluation provided in this document, the goal is to analyse the number of well-rated properties available and understand if the overall quality of the service is guaranteed. In addition, Airbnb must control customers' reviews so they can remove scams and other problematic properties from the platform to guarantee its quality.

Considering the main objectives of the Airbnb platform, it is assumed that in the end, most of the properties have a good review since the bad ones should already have been removed from the platform or are in the process of being. However, there is a risk of the number of bad reviews being higher than expected, which could be interpreted as inadequate platform management.

Goals

The main objective of this work is to understand if there is a direct relationship between the price and property quality. For this, the value of the evaluations and the price will be compared.

Another objective is to understand if there is a control of the properties available in the platform through the evaluation of the number of bad reviews and to understand if the number of evaluations impacts the review value.

Additionally, the impact of the property’s type and location on pricing will be evaluated.

In the end, there is a modelling evaluation, comparing different models to achieve the one that fits better, in this data set case, to be used in predictions of the property's price.

Project Plan

For this project, a fixed data set from 2019 will be used through Python Notebook to transform and evaluate the data. The data set has a total of 7833 rows and 33 columns, representing all the properties available in Amsterdam in 2019.

The first step of data processing consisted of removing the NaN values to evaluate the data set better. For this, it was necessary to start by assessing the amount of those NaN values in each one of the variables.

The variables that produced the most significant number of NaN values were those representing review-related fields ('review_scores_rating',' 'review_scores_accuracy',' 'review_scores_cleanliness',' 'review_scores_checkin',' 'review_scores_communication',' 'review_scores_location',' 'review_scores_value'). To avoid losing data consistency, these attributes were replaced by the average of the remaining values in each column. The same operation was performed for the variables 'host_response_rate', bathrooms, bedrooms and beds.

Evaluating the correlation between the data set's variables was the first step to better understanding the data set using a correlation heatmap.

Then, the data was represented in the form of histograms. Through them, it was possible to draw some conclusions regarding the operation of the Airbnb business and confirm some of the assumptions made at the beginning. The analysis of some attributes using box plot graphs allowed a more detailed evaluation of them.

To evaluate the impact that the property location might have on price, the properties were represented in a scatter plot which had the information of Amsterdam city and city centre borders. For this, the geopandas library was used.

Lastly, the data was represented in pie charts to get more information about the impact of the property type on price.

In the end, a Python library (Pycaret) was used to compare several models and, through several metrics, conclude which models are best used in this data set.

3.2 Data Collection

The Google Dataset Search platform was used to search for a data set for the development of the present work. The website that contains the selected data set is data.world, and the data set id is Airbnb Raw Data.

The data set has 7833 rows and 33 columns that include the following data:

- Host id (integer);
- Hostname (string);
- The year the user became a host (year);
- The day and month that the user became a host (string);
- Id of the property (integer);
- Neighbourhood of the property (string);
- City of the property (string);
- State of the property (string);
- Zip Code of the property (string);
- Country of the property (string);
- Latitude of the property (integer);
- Longitude of the property (integer);
- Location of the property (geopoint);
- Property type (string; possible values: apartment, house, bed & breakfast, boat, loft, cabin, camper/RV, villa, dorm, yurt, chalet, earth house, hut, treehouse, other);
- Room type (string; possible values: entire home/apt, private room, shared room);
- Number of people that can be accommodated (integer);
- Number of bathrooms (integer);
- Number of bedrooms (integer);
- Number of beds (integer);
- Type of bed (string);
- Price (integer);
- Number of guests that are included in the price (integer);
- Extra price for each extra person (integer);
- Minimum nights that the client must spend to book the property (integer)
- Host response time (string, possible values: within a few hours, within an hour, within a day, a few days or more, N/A);
- Host response rate (integer);
- Number of reviews (integer);
- Rating review value (integer);
- Accuracy review value (integer);
- Cleanliness review value (integer);
- Check-in review value (integer);
- Communication review value (integer);
- Location review value (integer);
- Primary review value (integer).

This data is sufficient to achieve and analyse the proposed objectives.

3.3 Exploring Data

The data set is composed only of one table with one primary key: the property id.

All the different review values are a part of an attribute population since they represent the evaluation of the property.

To analyse the data, it was necessary to treat it, such as replacing the NaN values with the median values of the column. Next, various ways of visualising it were implemented to examine the data provided and turn it into knowledge to extract information as described below.

Seaborn Correlation Heatmap

The Seaborn Correlation Heatmap concept can be separated into two parts:

Firstly, the Heatmap is a type of data visualisation where darker colours generally represent higher activity while lighter colours represent the opposite.

Secondly, correlation is a statistical measure that relates two variables whose values cover the range [-1, +1]. Thus, when the result is -1, it is considered a weak relationship because as one variable increases, the other decreases. In opposition, when the result is +1, both variables increase, which shows a strong relationship between the two variables. When the correlation is 0, there is no linear trend or relationship between the two variables analysed.

In short, the correlation heatmap represents a 2D correlation matrix. This matrix represents the values of the first dimension in the rows and the second dimension in the columns. Furthermore, as the colour of the cells is displayed on a monochromatic scale, it becomes simpler to analyse the data obtained.

Taking this into consideration, a Seaborn Correlation Heatmap of the variables available in the data set was appropriate to detect the existing correlation between the various variables, as can be seen in the following figure:

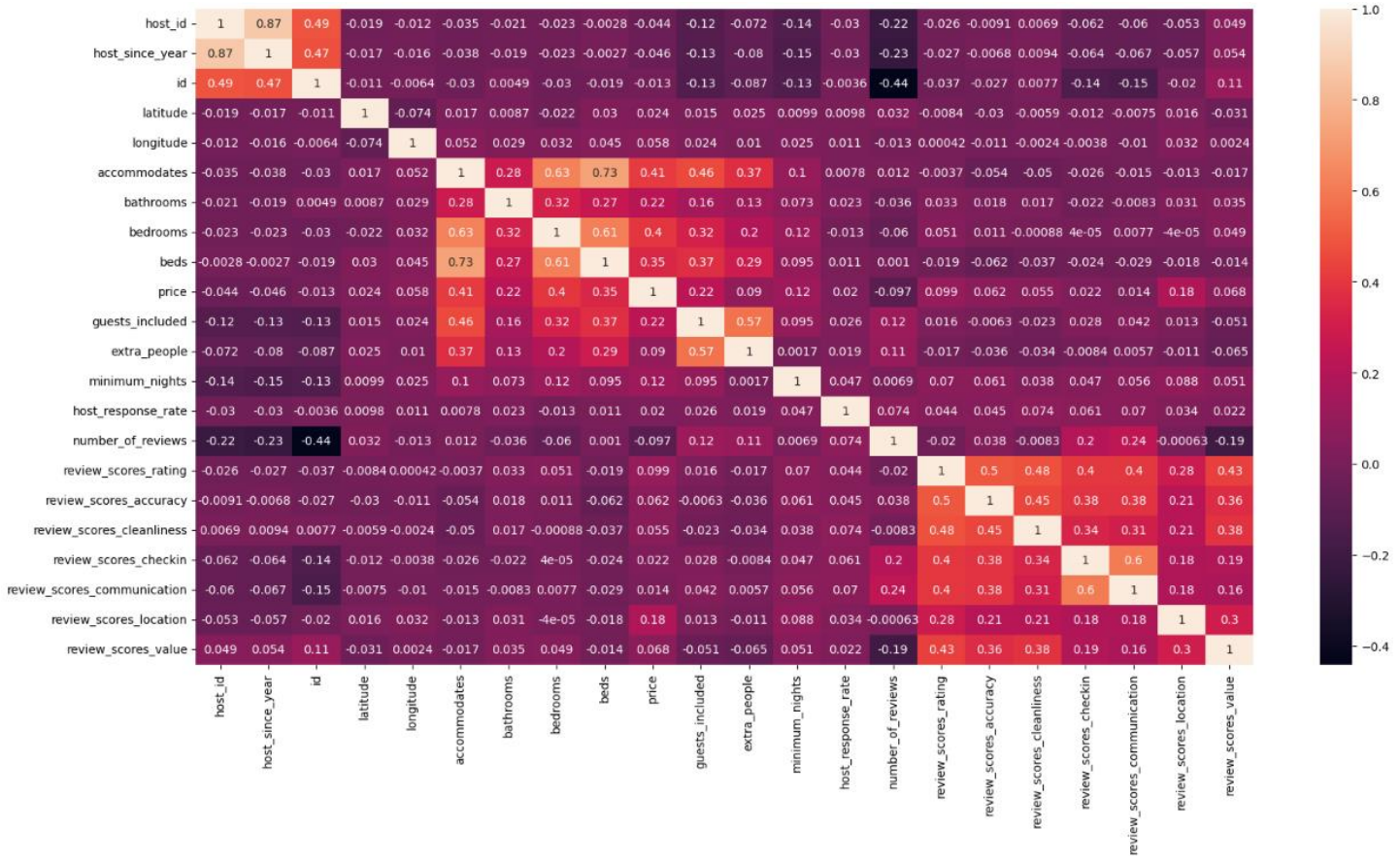

Figure 1 - Seaborn Correlation Heatmap for Airbnb in Amsterdam.

In this way, the diagonal of this graph assumes the value one because it is correlating the variable with itself, translating to a perfect correlation. In this case, data represented in purple represents a lower correlation than those represented with lighter colours. It can also be stated that the heatmap is symmetric since the same two variables are correlated in 2D.

It should be noted that the variables with the lowest correlation are the number of reviews and the id of the apartment, with a correlation of -0.44. Considering the application of this value in the Airbnb business, it is possible to confirm its validity since the number of reviews has no direct relationship with the property id. Instead, the number of reviews depends on the number of people who have visited the property and chosen to write a review about it.

On the other hand, the variables that show the best relationship, not considering the correlation of the variable with itself, are the ones with a correlation of 0.87 resulting from the relationship between 'host_id' and 'host_since_year,' implying that for each host id there is also stored information about the year of adherence of the host to the application.

It is also worth noting the relationship between the variables 'bedrooms' and 'accommodates', which correlates to 0.63; the relationship between the variables 'beds' and 'accommodates', assuming a value of 0.74 and finally, the relationship of the variables 'beds' with 'bedrooms' with a correlation of 0.61. These values fit the business goal since the more spaces and beds a property has, the more people it can accommodate.

Histogram

A histogram is a tool to observe the distribution of data on a given scale. Thus, it is possible to assess the frequency of a given value in a given attribute.

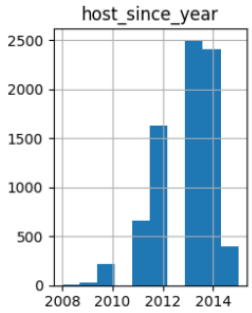

Figure 2 - The year the host joined Airbnb.

As can be observed in the histogram of Fig. 2, there has been an evident growth over the years in the number of people who have become hosts. This variation in values may be due to the growth of the platform and its dissemination. The platform created in 2008 started having a small number of hosts until 2010, but the trust in the platform and its good performance positively impacted the number of hosts over time.

In Fig. 3, it is possible to observe the number of people that properties can accommodate and the number of bathrooms, bedrooms, and beds. Through the evaluation of these histograms, the direct relationship between the number of beds, bedrooms, and people that the property accommodates, as described in the previous sub-chapter, is clear.

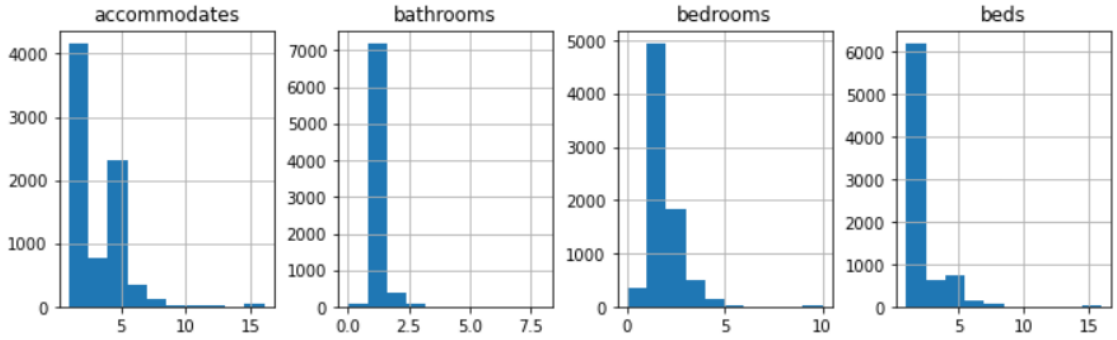

Figure 3 - Characteristics of Airbnb's properties - accommodation capacity, number of bathrooms, bedrooms, and beds.

Next, Fig. 4 shows the number of guests included by each Airbnb and the fixed price for each extra person.

Notably, most of the prices to be paid are below 50€ but, from a total of 7833 properties, 4.316 do not accept extra people beyond the initially established value.

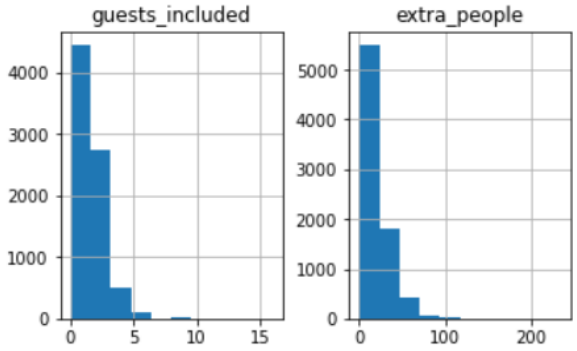

Figure 4 - Number of guests and price.

The following Fig. 5 shows the host response rate, and it is possible to see through the analysis that the vast majority of hosts respond to messages from interested parties, even if they take some time.

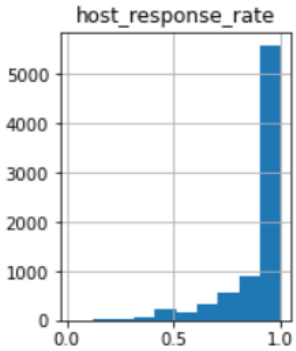

Figure 5 - Host response rate on Airbnb.

Fig. 6 shows the number of reviews, the review scores rating, the review scores accuracy, the review scores cleanliness, the review scores check-in, the review scores communication, the review scores location, and the review scores value.

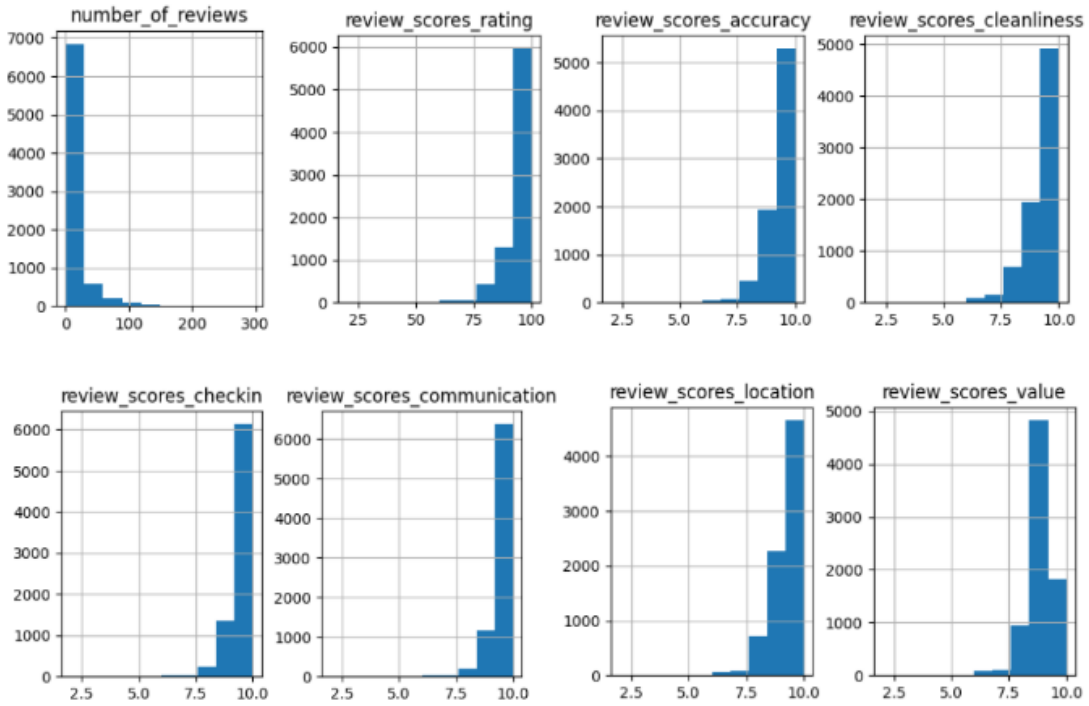

Figure 6 - Airbnb reviews in Amsterdam - number of reviews, review scores rating, review scores accuracy, review scores cleanliness, review scores check-in, review scores communication, review scores location, and review scores value.

Starting the analysis with the number of reviews, almost 7,000 Airbnb's have approximately between 0 and 25 reviews.

Next, in the graph relating to the review scores rating, it is possible to see that the most significant influx of rating scores is from values higher than 75 to 100, which is a good indication that Airbnb users are satisfied with the services provided.

Subsequently, in the analysis of the review scores accuracy graph, similar to the previous graph, the highest influx of accuracy scores are valued higher than 7.5 up to 10, which means that Airbnb's accuracy is true to what the host described, which contributes to a better alignment of expectations between customers and hosts.

In the next step, the review scores cleanliness chart analysis shows that users are satisfied with the cleaning services provided, as the values were primarily above 7.5 and only a few users gave a low value.

The graphs that express the review scores check-in and the review scores communication also admit a similar trend to those previously analysed, where the values are significantly favourable because the most significant proportion of values are above 8, demonstrating that users are pleased with the check-in service and with the host's communication towards them.

Next, in the analysis of the review scores location graph, it is also possible to observe that a large part of the scores is located in the range of [7.5;10], which means that guests are pleased with the location of the properties.

Finally, in the analysis of the review scores value graph, the scores overall have values higher than 8, which indicates that guests are satisfied with Airbnb's property renting price.

In summary, it is possible to conclude that the graphs presented above all have the same trend of results, and it is possible to establish several relationships between the histograms.

Starting by relating the graph of the host response rate with the review score communication, it is possible to state that if the customer does not have a good experience with, for example, the way or the delay of the host response, it will undoubtedly leave a less satisfactory review, in the host response rate, in the review score communication, which will cause adverse effects on the review score rating.

It is also possible to relate the review score accuracy graph to the review score rating because if the host has not made a realistic description of Airbnb, the review accuracy will be low, and consequently, so will the review score rating because the customer may feel deceived by the host, to the extent that the description provided does not correspond to reality.

The review score cleanliness graph and the check-in are also related to the review score rating graph since if all these experiences are not pleasant for the user, certainly the review score rating will be influenced, which also causes a feeling to the user of more or less cost/benefit about the price he paid and the conditions he had.

Box Plot

A box plot is a graphic tool whose purpose is to present data distribution and outliers through quartiles, thus enabling the comparison of the medians of data dispersion and outliers between box plots.

The following charts were created to analyse the data regarding the following variables: 'review_scores_rating', 'price', 'review_scores_cleanliness', and 'review_scores_location'.

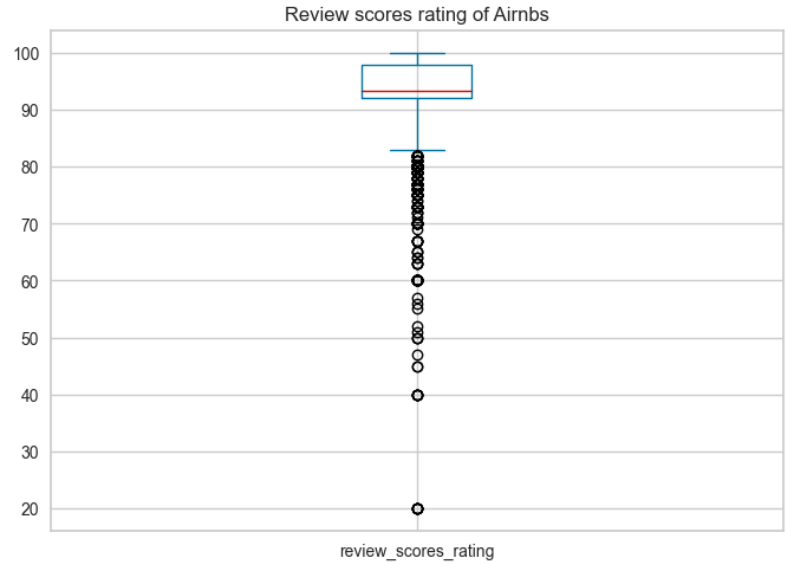

Figure 7 - Review scores rating of Airbnb's in Amsterdam.

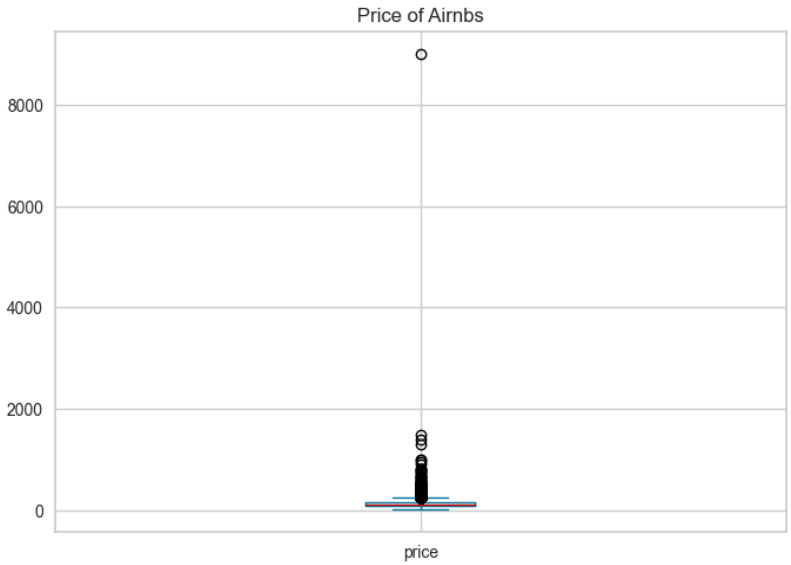

Figure 8 - Price of Airbnb's in Amsterdam.

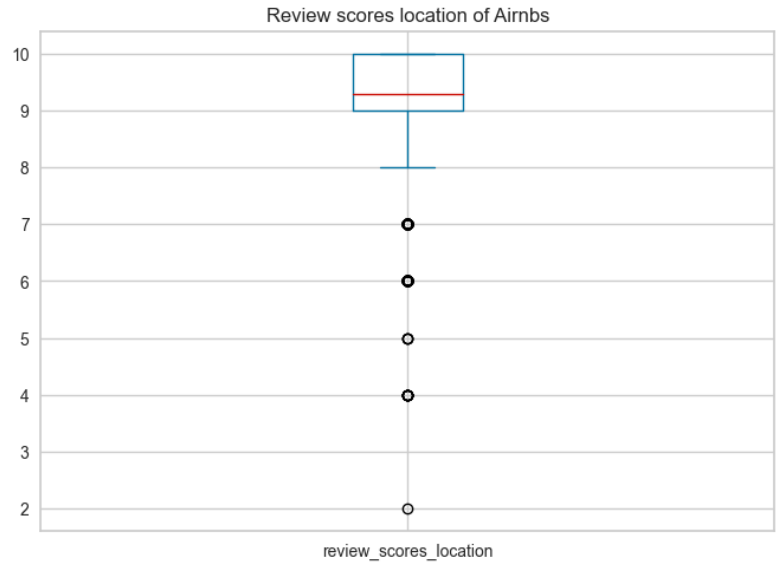

Figure 9 - Review scores location of Airbnb's in Amsterdam.

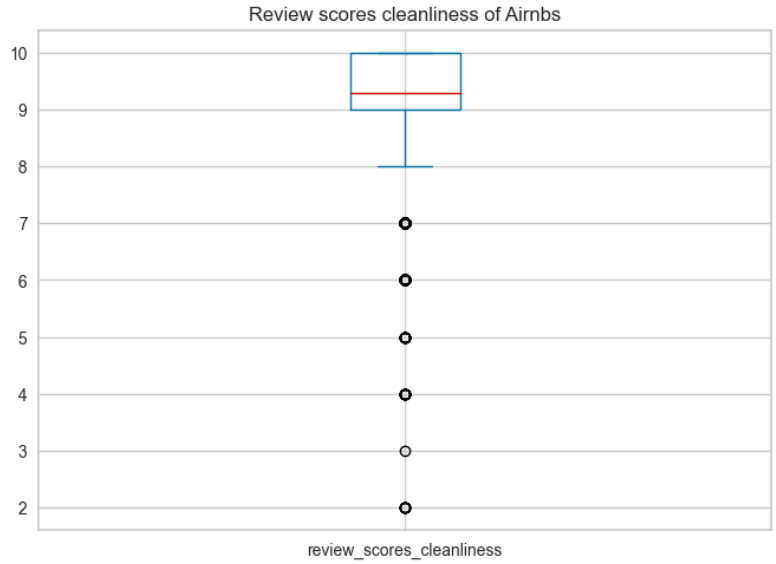

Figure 10 - Review scores cleanliness of Airbnb's in Amsterdam.

In Fig. 7, the box plot of the review scores rating of Airbnb in Amsterdam shows that the maximum score is 100, the median, that is, the second quartile, is close to 93.34 and the minimum score is close to 83. The third and the first quartile are between [92, 98], which shows that, in general, users are quite satisfied with the service provided. Nevertheless, there are several outliers, the most extreme takes the value 20, and even so, there is an intense concentration between the values [83, 60], which translates the existence of several reviews that attribute to Airbnb the ratings that diverge from the average, revealing that the hosts can still significantly improve Airbnb's and their conditions in general.

Fig. 8 shows the graph relating to the price of Airbnb in Amsterdam, where due to the outlier that assumes a value greater than 8.000€ and the dispersion of the remaining values, the analysis of the respective box plot was hampered. But through auxiliary calculations, using the panda's library and the quantile() method, it was possible to accurately determine the median, where it assumes a value of 109.0€ and where the third and first quartiles are between [85.0; 150.0]. It is possible to deduce that generally for

each Airbnb, the price varies between €85 and €150 per night, with the most common value being €109. However, all the outliers recorded are above the third quartile, i.e., there are prices of up to 150 per night.

Fig. 9 illustrates the graph concerning the variable the review scores location of Airbnb in Amsterdam, where the maximum limit is equal to the maximum value of the third quartile, with the box being between [9,10], the minimum limit assumes the value of 8 and the median of 9.29 scores, it should be noted that there are some outliers in the values 7, 6, 5, 4 and 2. This means that users are satisfied overall and on average they evaluate with 9.29 scores the location of the chosen Airbnb.

Finally, Fig. 10 shows the graph relating to the variable the review of cleanliness of Airbnb in Amsterdam, and it is possible to see that this box plot is similar to Fig. 9 but has added more outliers, namely in score 3, and presents a median of 9.28 scores, which translates that users think that the Airbnb they rented is in great conditions regarding the hygiene of the place.

In general, terms, as shown in Figs. 7, 9 and 10, there are similarities between the box plots, namely in the positive asymmetry, as the median is closer to the first quartile than to the third. Similarly, the data dispersion between the third and first quartiles is also closer to the maximum value limit. It can be concluded that users are generally very satisfied with Airbnb for the topics previously analysed.

Scatter Plot

A scatter plot is a diagram in which each pair of values in the data set is represented by a point. It is also possible to add the representation of more variables, by changing the colour or size of the points, for example.

To facilitate the interpretation of each of the graphs, the first step was to represent the borders of the city of Amsterdam, in black, and the borders of the city centre, in red.

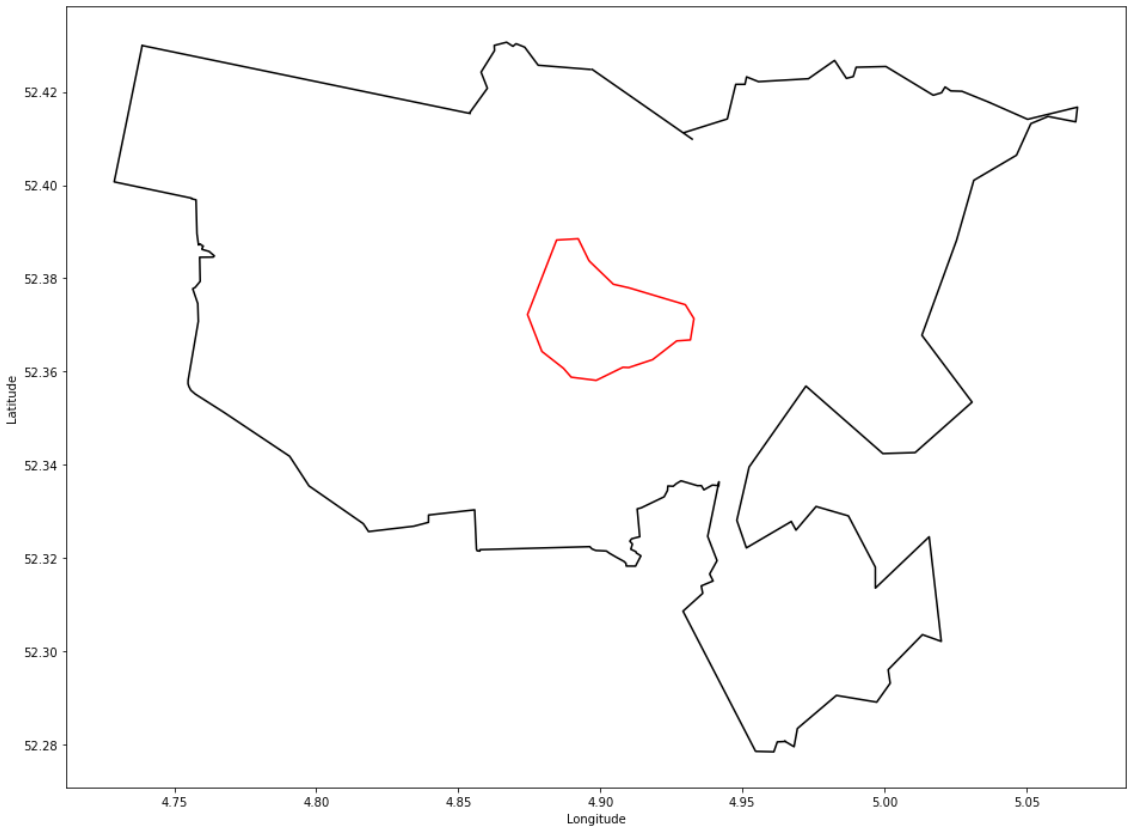

Figure 11 - Amsterdam city and city centre borders.

In Fig. 12 it is possible to observe only the coordinates of the houses whose value is less than 1000€ because the houses with a higher value were disregarded since there are only four of them and their inclusion would make the diagram difficult to observe since it would imply that the colour scale represents a greater number of values, with a smaller gradient difference between values with significant differences.

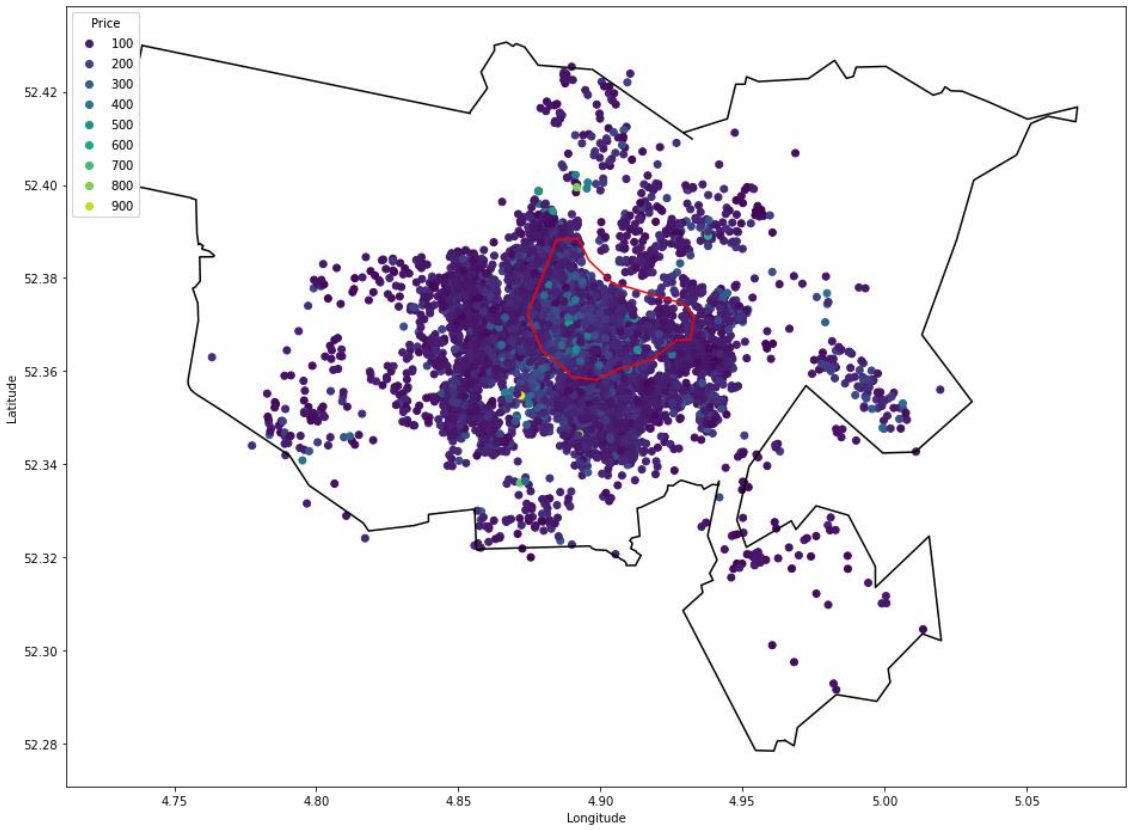

Figure 12 - Properties with price less than 1000.

Contrary to the expected, the highest values are not mostly found in the city centre. However, this may be due to several factors, namely the number of people each property accommodates also influences the rental price of the property, i.e. the more people a property accommodates, the higher its price can be.

Fig. 13 represents only the properties that accommodate 1 person. This way it is possible to make a more accurate comparison.

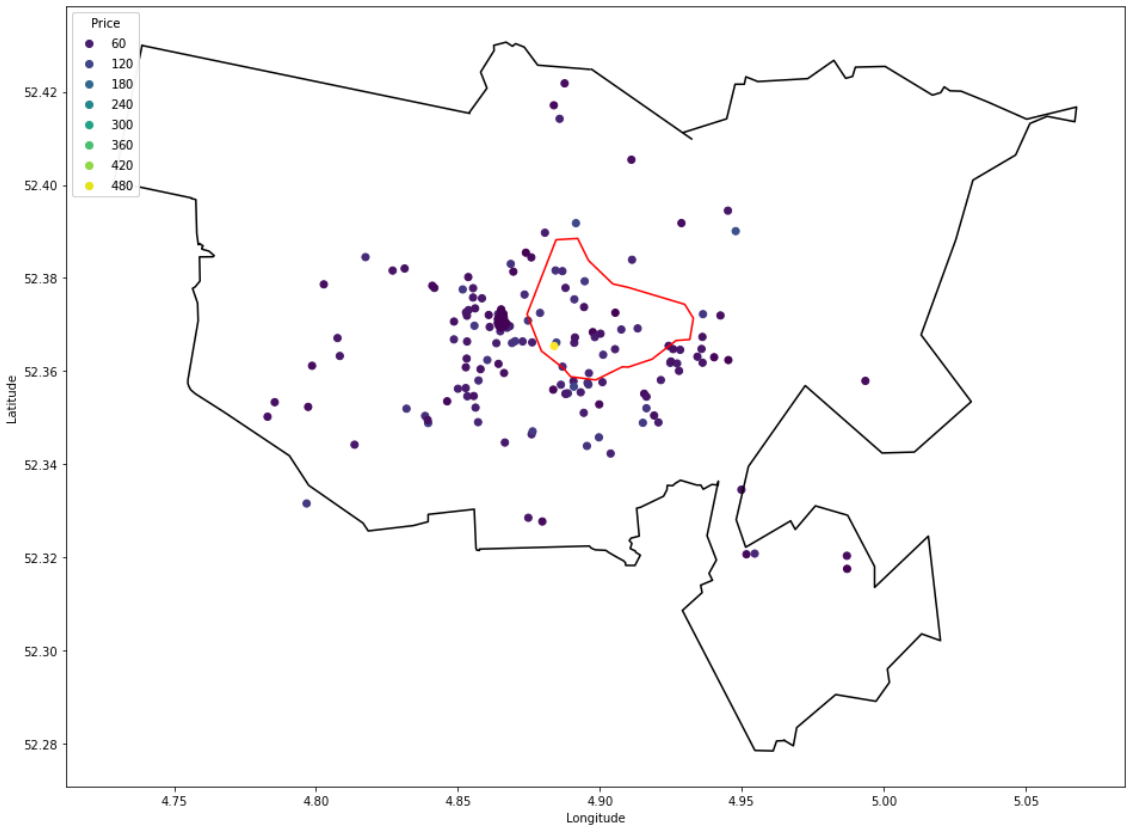

Figure 13 - Properties that accommodate only one person.

The study (Gyódi & Nawaro, 2021) found that, in 10 of Europe's largest cities, rents in city centres were higher than in suburban areas. This figure shows that the most expensive property, represented in yellow, is exactly in the city centre.

However, due to the size of the scale used, there is no great notion of the difference between the values of the remaining properties. In this way it is possible to observe in Fig. 14, houses that accommodate only one person and whose value is lower than 470€, to remove the outlier and make the visualisation and interpretation of the remaining data more detailed.

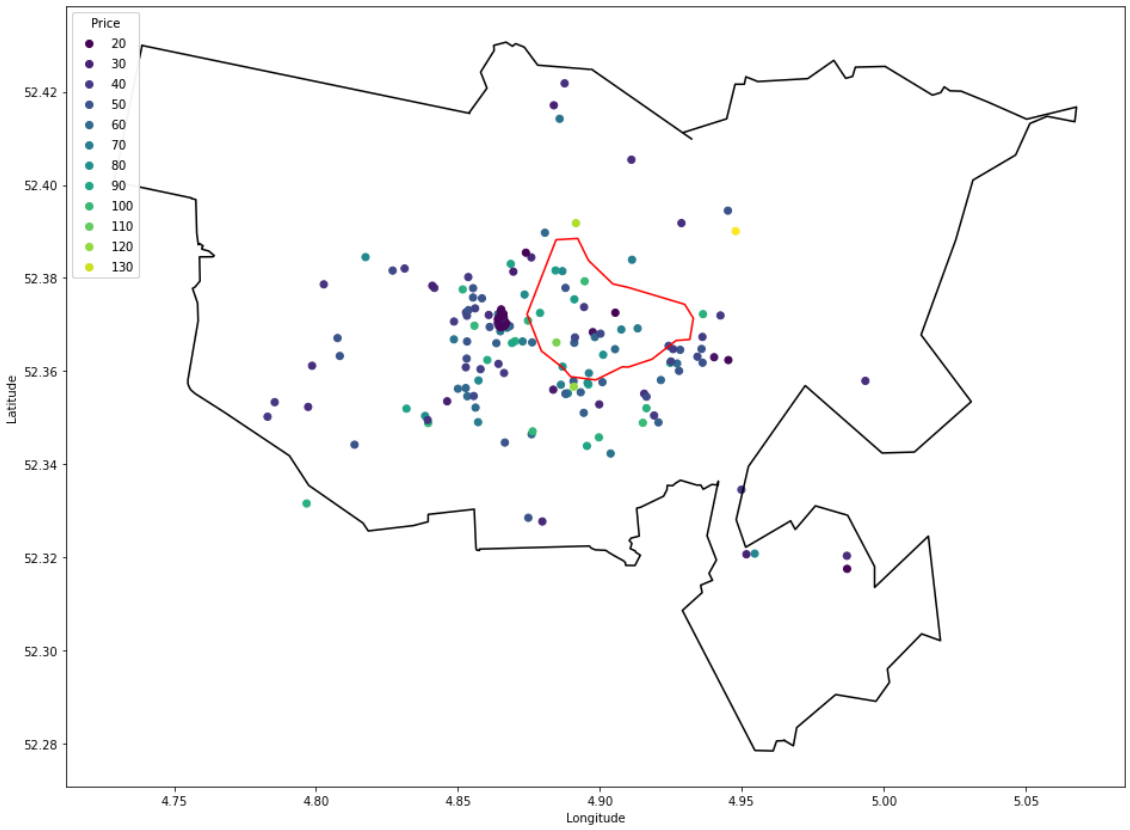

Figure 14 - Properties that accommodate only one person, and the price is less than 470.

Looking at the same case in more detail, it is apparent that not only is it the most expensive property outside the centre, but there are properties of various values distributed across the city.

This observation raises the following questions: What influences the price difference? Is it the valuations of each property? Is it the extra conditions it provides?

To be able to answer the questions raised above and consider the limitations of the data set, restrictions were made regarding the property ratings. In Fig 15 it is possible to see the properties that accommodate only one person, with prices lower than 470€ and with a review score rating higher than 95, with a maximum of 100 scores.

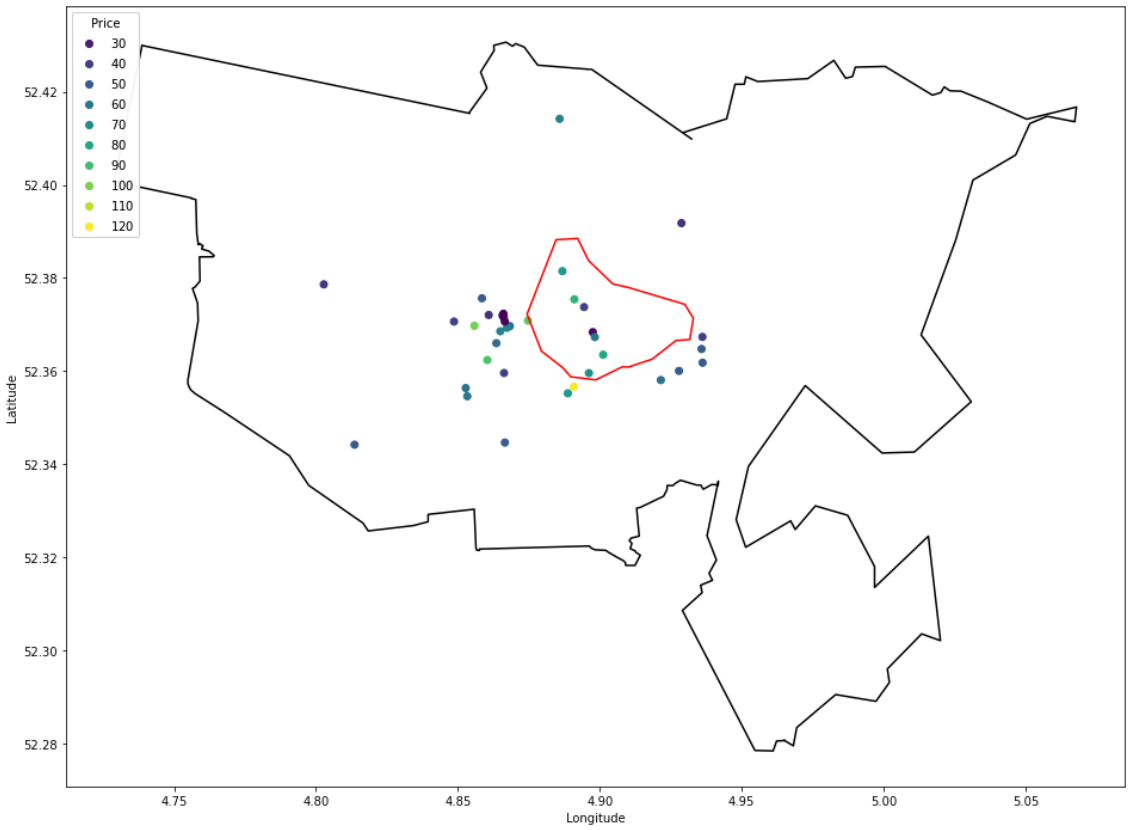

Figure 15 - Properties that accommodate only one person, the price is less than 470 and the rating is above 95.

It can therefore be seen that the number of properties has decreased significantly. However, the disparity in values remains, which reflects the fact that other factors influence the price of properties.

Some of the factors that can influence the price are the physical attributes of the property; neighbourhoods; time of the week; the month or year and location of the property **(Islam, et al., 2022).**

Another evaluation carried out was the relationship that exists between the number of evaluations and their value, represented in Fig 16.

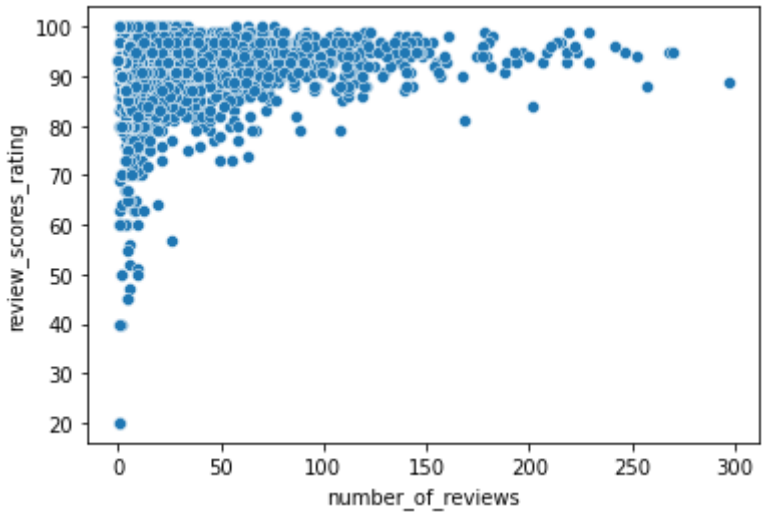

Figure 16 - Relationship between the number of evaluations and their value.

As we can see, the properties that have a good rating are comprised between few or many ratings. However, only properties with few ratings have a low rating, this can represent the control done by the Airbnb platform. Therefore, properties that have a low evaluation, represent a bad service.

Pie Chart

Another factor that can influence is the type of accommodation the property has. Thus, an assessment was made of the price related to the type of property represented in figures 17, 18 and 19.

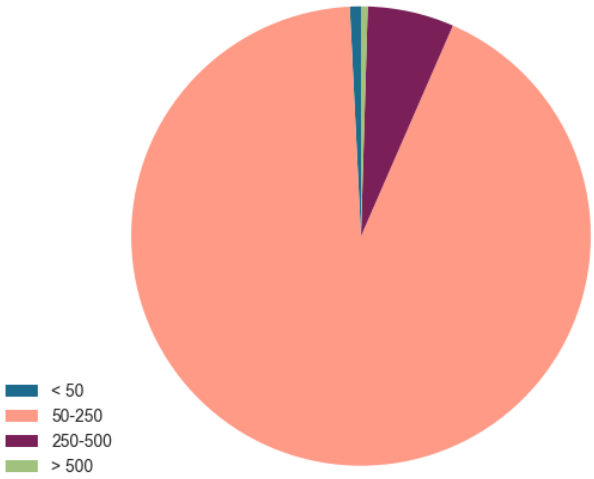

Figure 17 - Price distribution for Entire Home/Apt.

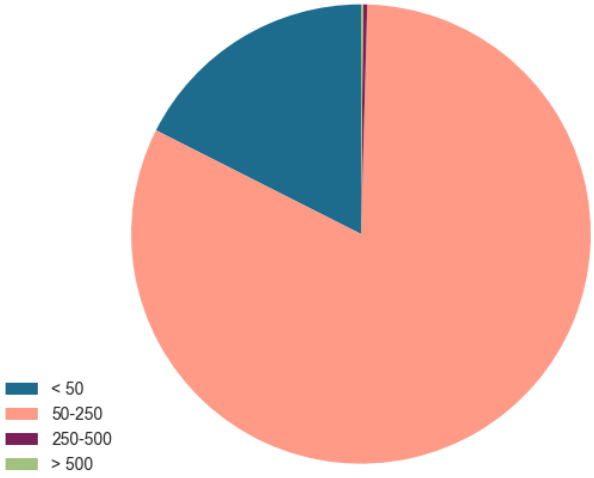

Figure 18 - Price distribution for Private Room.

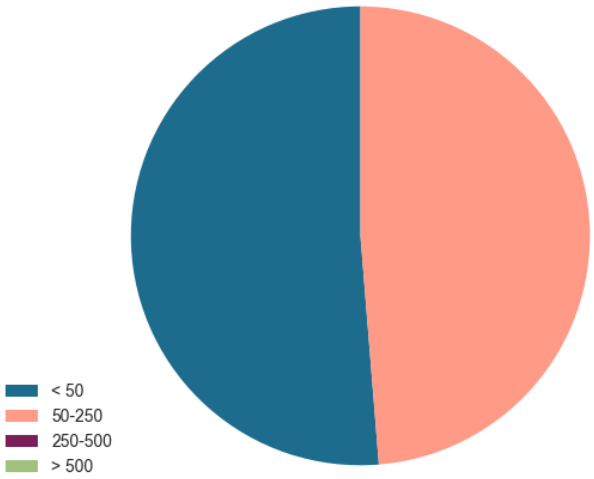

Figure 19 - Price distribution for Shared Room.

There is indeed an influence of the type of property on its price.

The type of property Entire Home/Apt shows that over 75% of users are willing to pay between 50€ and 250€, while in Shared Room, there is a less significant margin of difference between users willing to pay 50€ and users willing to pay between 50€ and 250€. It is therefore possible to conclude that users are more willing to pay higher values for a better type of property.

3.4 Data Modelling

Since the data set has labelled data, supervised machine learning techniques should be used, aiming to discover relationships between dependent and independent variables. They can be further classified into two main categories: classification and regression. As the target variable, price, is a continuous variable, regression models were used for the forecasts.

The best models to use were selected with the help of PyCaret (Python Library). This library was chosen because of the ease with which it is used, replacing several lines of code for something much more straightforward.

This library allows the data to be evaluated with different models, making it possible to compare them and select the best models to apply. Notably, the proportion of the data set used for training and validation used by this library is by default 0.7, and numerical missing values are transformed and calculated from the average value of the training data set to feature.

The models used to test the data set were:

- Bayesian Ridge (**br**);
- Elastic Net (**en**);
- Lasso Regression (**lasso**);
- Light Gradient Boosting Machine (**lightgbm**);
- Linear Regression (**lr**);
- Random Forest Regressor (**rf**);
- Extra Trees Regressor (**et**);
- Decision Tree Regressor (**dt**);
- Dummy Regressor (**dummy**);
- Lasso Least Angle Regression (**llar**);
- Gradient Boosting Regressor (**gbr**);
- Ridge Regression (**ridge**);
- K Neighbors Regressor (**knn**);
- Huber Regressor (**huber**);
- AdaBoost Regressor (**ada**);
- Orthogonal Matching Pursuit (**omp**);
- Passive Aggressive Regressor (**par**);
- Least Angle Regression (**lar**).

With the model comparison provided by PyCaret, it was possible to compare the metrics relating to each model and understand which one was best to use.

Note that the Least Angular Regression and the Passive Aggressive Regressor were discarded from the following analyses because, in this case, they admit very high values, being very bad models to use, as it prevents an efficient and detailed analysis of the graphs of the remaining models.

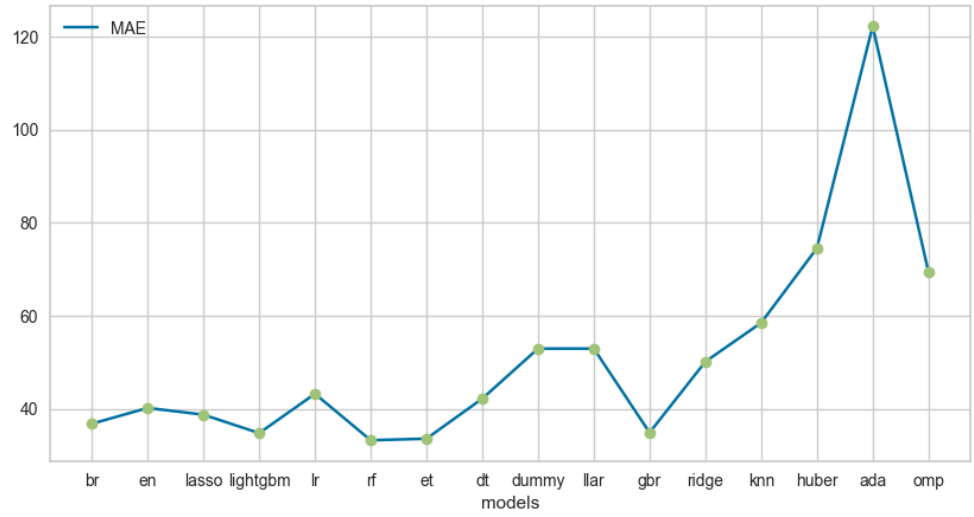

Figure 20 - Mean Absolute Error of the models.

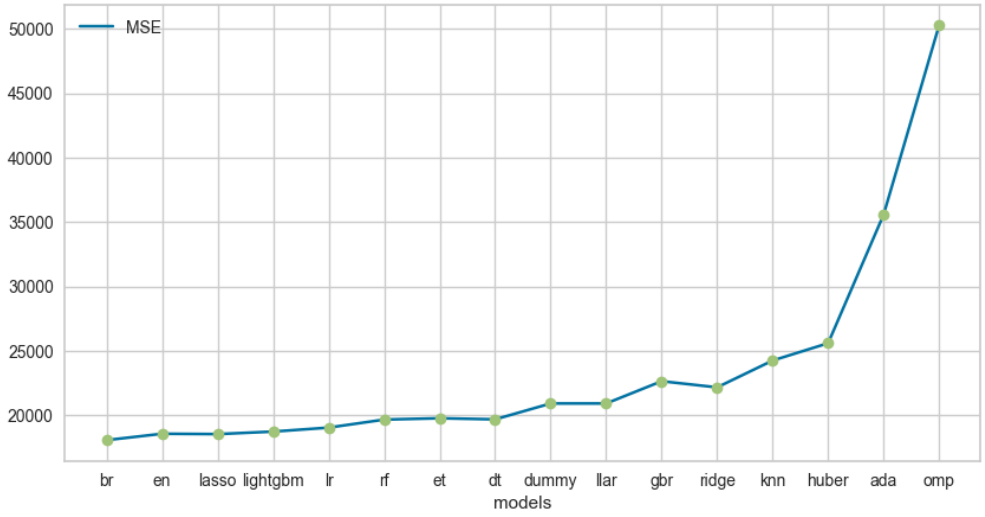

Figure 21 - Mean Squared Error of the models.

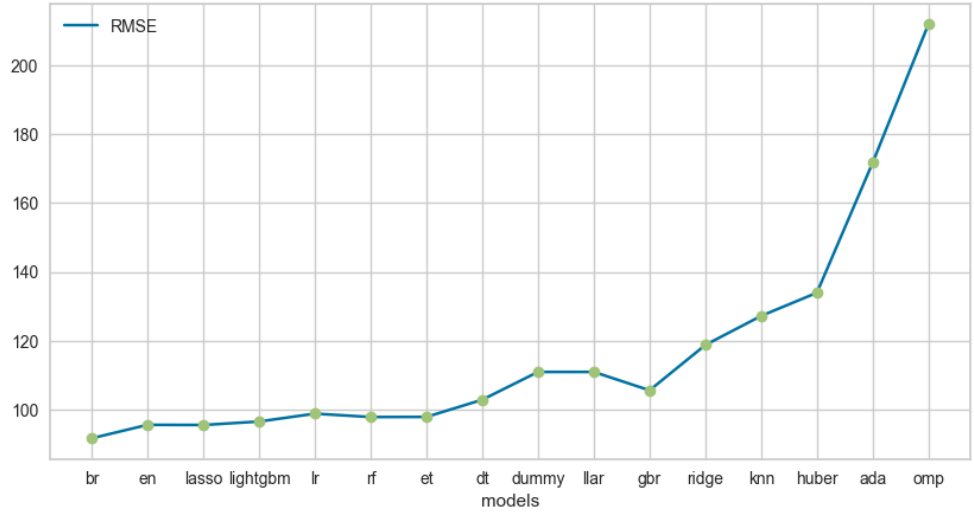

Figure 22 - Root Mean Square Error of the models.

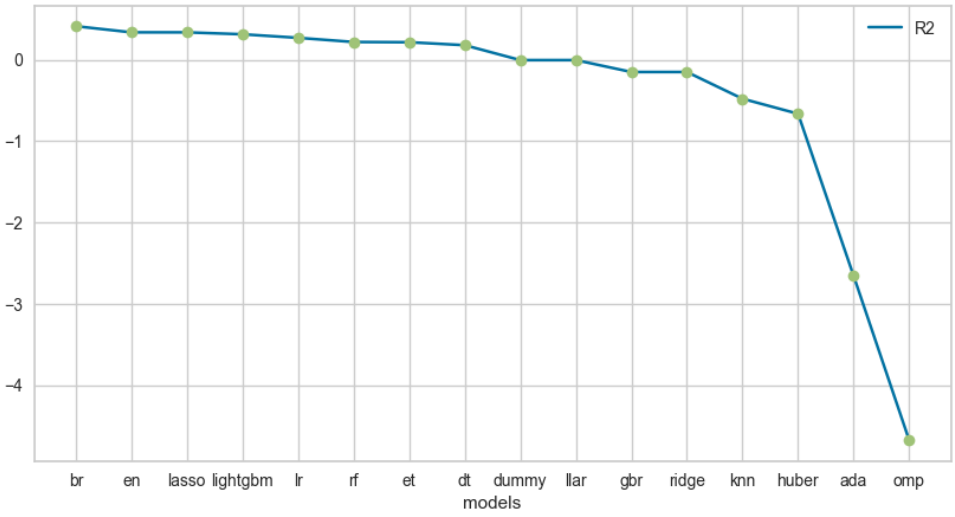

Figure 23 - R-squared of the models.

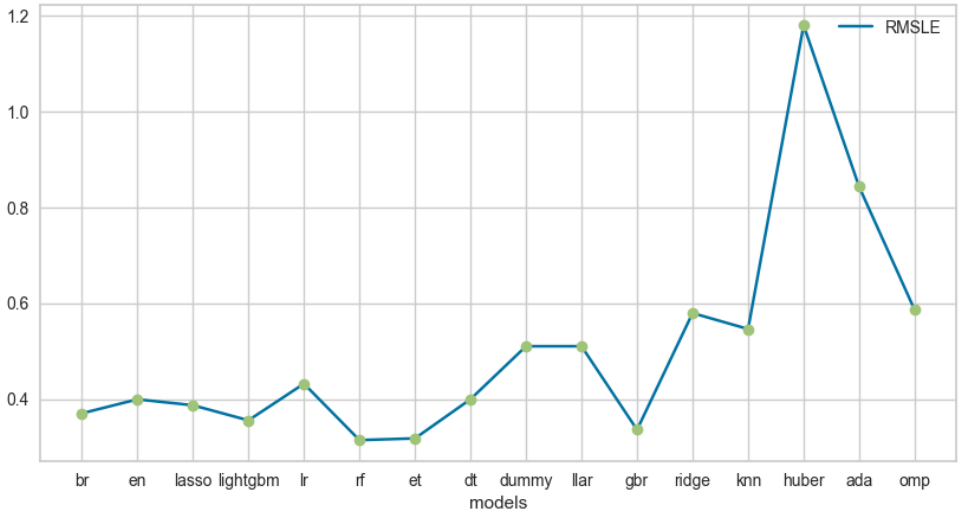

Figure 24 - Root Mean Logarithmic Error of the Models.

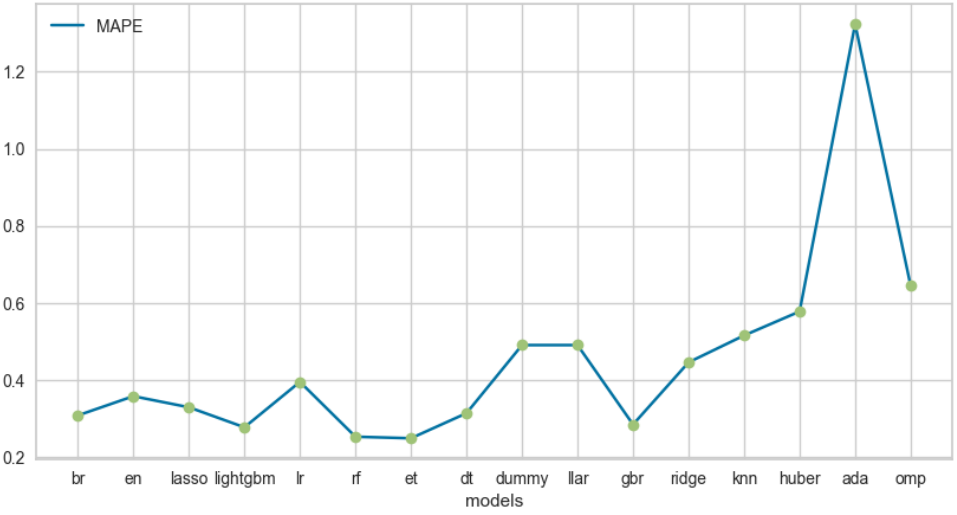

Figure 25 - Mean Absolute Percentage Error of the models.

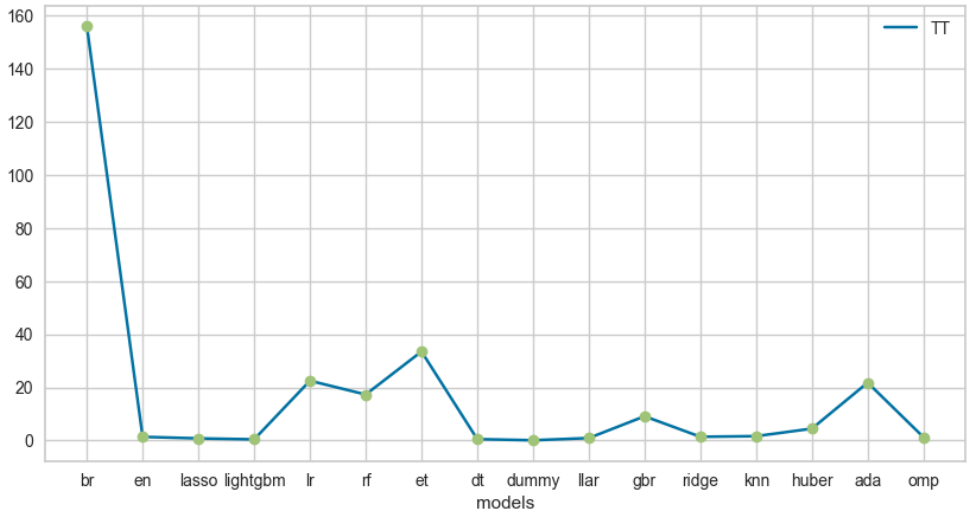

Figure 26 - Execution Time of each model.

3.5 Evaluation

Based on the graphs shown in the previous section, Fig. 20 illustrates the models considering only the MAE metric (Mean Absolute Error), and it is possible to conclude that the best models representing this metric are: the Random Forest Regressor with a value of 33.2209; followed by the Extra Trees Classifier with 33.5666; and finally, the Gradient Boosting Machine with 34.7624, which means that the Random Forest Regressor for being the model with the lowest absolute value of errors is therefore the best to be used according to this metric that in a way measures accuracy.

Moving on to Fig. 21, the MSE (Mean Squared Error) measures the mean squared difference between the estimated values and the actual value, in this case the TOP 3 of the models analysed: Bayesian Ridge with 18072.9586, Lasso Regression with 18540.2089 and the Elastic Net model with 18567.5508.

Next, Fig. 22 is considered the RMSE (Root Mean Square Error) that evaluates the average distance between the predicted values of the model and the real ones, briefly the model shows how concentrated the data are around the line of best fit, and the models that best represent this measure are the Random Forest Regressor with 91.7316, followed by the Lasso Regression with 95.5863.

Moving on to Fig. 23, R2 (R-squared) is the metric that predicts how well the model predicts the outcome. When the result is 0 the model does not predict the outcome, when the result is 1 the model predicts the outcome perfectly and when the value is between [0,1] then the model partially predicts the outcome. Therefore, the two models which are closest to the value 1 are: Bayesian Ridge with 0.4134 and Elastic Net with 0.3382.

Fig. 24 represents RMSLE (Root Mean Squared Logarithmic Error), which despite being like an RMSE metric, it uses logarithms. The best models are the Random Forest Regressor with 0.3145 and the Extra Trees Regressor with 0.3181 for this metric which only considers the relative error between the Predicted and the actual value.

Next, Fig. 25 considers the MAPE metric (Mean Absolute Percentage Error), which evaluates the average difference between the predicted and the actual value. The models that best represent this metric are the Extra Trees Regressor with 24.82%, followed by the Random Forest Regressor with 25.22%.

Finally, Fig. 26 represents the execution times (TT) of each model, and the model that stands out the most is the Bayesian Ridge because it takes the longest execution time, followed by the et, lr, ada, rf, gbr and huber models. Apart from these models mentioned above, the others do not present great differences regarding execution times, presenting good times.

In conclusion, the Bayesian Ridge model and the Random Forest Regressor model are the most appropriate models to be used, since they present the best results in the metrics, and it should be noted that their applicability varies depending on what we are evaluating.

4. Results & Discussion

Through the developed analysis, it is possible to understand that there is a strong correlation between bedrooms and beds with accommodation, and so there is also a relationship between beds and bedrooms, and it is obvious that the more beds and bedrooms there are, the more guests Airbnb will receive.

Regarding the histograms, the biggest conclusion is that the better each review is, for example regarding cleanliness and accuracy, the better Airbnb's score rating and score value are. It is also worth noting that in general, all guest reviews are higher than 7.5 on a scale of 1 to 10, although there is always room for improvement.

In addition, through the box plot, the analysed reviews have medians higher than 9 points on a scale of 1 to 10, supporting the results obtained in the histograms.

The impact that the location of the property can have on the property price was also assessed. However, it was concluded that this factor alone is not enough, although it does have a significant impact.

Property type was also a factor assessed, with the conclusions being much more direct as there is a significant relationship of importance between this factor and property types.

Finally, the PyCaret library was used and through the results obtained it was possible to conclude that the two best regression models to be applied to the data set are the Bayesian Ridge and the Random Forest Regressor.

5. Limitations & Conclusions

The main difficulty found during the data analysis was the amount of data and the treatment of the null values. The limitation that most restricted the analysis was the outlier higher than 8,000€ of the price variable which hampered the entire analysis of this variable because it influenced the values obtained and was considered an extreme outlier which is an exception to the other moderate outliers found

Thus, it is believed that it would have been possible to obtain a better accuracy value, i.e., less MAE if the data set had more complete data because then it would not be necessary to base the project on assumptions and correlations, since the null values and outliers make the results have less accuracy and are biased, even replacing these values by means or medians, for example.

For future research, it would be interesting to group the prices by holidays and analyse the price evolution over several years to observe if it would be possible to find any trend.

Another interesting method to continue this project and improve the accuracy of the models would be to use ensemble methods, i.e., bagging and boosting, because, although more complex than traditional models, they combine the results of multiple weak models and produce better results.

6. References

Awan, A., & Soofi, A. (2017). Classification Techniques in Machine Learning: Applications and. *Journal of Basic & Applied Sciences*.

Dhillon, J., Priyanka Eluri, N., Kaur, D., Chhipa, A., Gadupudi, A., Cherupulli Eravi, R., & Pirouz, M. (2021). Analysis of Airbnb Prices using Machine Learning.

Gyódi, K., & Nawaro, Ł. (2021). Determinants of Airbnb prices in European cities: A spatial econometrics approach. *Tourism Management*.

Islam, M., Li, B., Saiful Islam, K., Ahasan, R., Mia, M., & Haque, M. (2022). Airbnb rental price modeling based on Latent Dirichlet Allocation and.

Li, R., Zhu, A., & Xie, Z. (2020). Machine Learning Prediction of New York Airbnb. *Third International Conference on Artificial Intelligence for Industries*.

pycaret. (2022, 11 25). Retrieved from Regression: https://pycaret.readthedocs.io/en/stable/api/regression.html

Used Data Set. (2022, 11 10). Retrieved from https://data.world/aewart/airbnb-raw-data