UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Machine Learning

# Project
# Airbnb Price Forecasting

*by*

Laura Isabella Forero Camacho
Liliia Aliakberova

*supervised by*

Marta Arias Vicente
Bernat Coma Puig

**Table of Contents**

## Introduction

This comprehensive report presents the in-depth analysis and findings of our research conducted as part of the Machine Learning course. Our investigation focused on the dataset titled "Airbnb Price Determinants in Europe," which we obtained from Kaggle [1]. Our primary objective was to develop a robust predictive model for forecasting Airbnb prices, a task of utmost importance in the dynamic and competitive landscape of the short-term rental market. By accurately estimating prices, hosts can strategically set competitive rates, attract a wider customer base, and optimize their rental income. Equally significant, this forecasting information empowers guests to make informed decisions and find the best value accommodation options that align with their preferences and budget constraints.

To accomplish our goal, we embarked on a thorough exploration of the determinants that influence Airbnb prices across various cities in Europe. By analyzing this rich dataset, we aimed to unravel the underlying factors driving price variations and identify the most influential features contributing to price differentials. Armed with this knowledge, we sought to develop a robust and reliable model capable of accurately predicting future Airbnb prices. Finally, we discussed the implications of our findings and highlighted potential areas for future research.

## 1. Data Exploration

### 1.1 General descriptive statistics

We conducted our analysis using the Airbnb Price Determinants in Europe dataset, which includes data from 10 cities: London, Amsterdam, Athens, Barcelona, Berlin, Budapest, Lisbon, Paris, Rome, and Vienna. The dataset comprises information collected for both weekdays and weekends. In total, it contains 51,707 rows when combining the data for weekdays and weekends, with 21 columns. Among these columns, 8 are categorical variables, and 13 are numeric variables. Notably, this dataset is complete, with no missing values and duplicates. The analysis and diagrams presented in this Data Exploration section are based on the ExploratoryAnalysis.ipynb notebook, which contains the full code and visualisations.

On the base of the describe() function for numerical data the dataset contains information on various numeric variables related to the rental properties.

| | realSum | person_capacity | cleanliness_rating | guest_satisfaction_overall | bedrooms | dist |
|---|---|---|---|---|---|---|
| count | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 | 51707.00000 | 51707.000000 |
| mean | 279.879591 | 3.161661 | 9.390624 | 92.628232 | 1.15876 | 3.191285 |
| std | 327.948386 | 1.298545 | 0.954868 | 8.945531 | 0.62741 | 2.393803 |
| min | 34.779339 | 2.000000 | 2.000000 | 20.000000 | 0.00000 | 0.015045 |
| 25% | 148.752174 | 2.000000 | 9.000000 | 90.000000 | 1.00000 | 1.453142 |
| 50% | 211.343089 | 3.000000 | 10.000000 | 95.000000 | 1.00000 | 2.613538 |
| 75% | 319.694287 | 4.000000 | 10.000000 | 99.000000 | 1.00000 | 4.263077 |
| max | 18545.450280 | 6.000000 | 10.000000 | 100.000000 | 10.00000 | 25.284557 |

| | metro_dist | attr_index | attr_index_norm | rest_index | rest_index_norm | lng | lat |
|---|---|---|---|---|---|---|---|
| count | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 | 51707.000000 |
| mean | 0.681540 | 294.204105 | 13.423792 | 626.856696 | 22.786177 | 7.426068 | 45.671128 |
| std | 0.858023 | 224.754123 | 9.807985 | 497.920226 | 17.804096 | 9.799725 | 5.249263 |
| min | 0.002301 | 15.152201 | 0.926301 | 19.576924 | 0.592757 | -9.226340 | 37.953000 |
| 25% | 0.248480 | 136.797385 | 6.380926 | 250.854114 | 8.751480 | -0.072500 | 41.399510 |
| 50% | 0.413269 | 234.331748 | 11.468305 | 522.052783 | 17.542238 | 4.873000 | 47.506690 |
| 75% | 0.737840 | 385.756381 | 17.415082 | 832.628988 | 32.964603 | 13.518825 | 51.471885 |
| max | 14.273577 | 4513.563486 | 100.000000 | 6696.156772 | 100.000000 | 23.786020 | 52.641410 |

*Figure 1 - Output of the "describe" function for numerical values*

Here are the key findings:

- The count column indicates that there are 51,707 data points available for each variable.
- "realSum": The average cost of the rental properties is approximately 279.88 units, with a minimum value of 34.78 and a maximum value of 18,545.45.

- "person_capacity": The average capacity of the rental properties is around 3.16 people, ranging from a minimum of 2 people to a maximum of 6 people.
- "cleanliness_rating": The cleanliness rating of the properties has an average value of 9.39, with a standard deviation of 0.95. The ratings range from a minimum of 2 to a maximum of 10.
- "guest_satisfaction_overall": The overall guest satisfaction rating is high, with an average value of 92.63 out of 100. The ratings vary from a minimum of 20 to a maximum of 100.
- "bedrooms": On average, the rental properties have approximately 1 bedroom, with a minimum of 0 and a maximum of 10 bedrooms.
- "dist and metro_dist": These variables represent the distances to certain locations. Their values range from a minimum of 0.015 to a maximum of 25.28.
- "attr_index" and "attr_index_norm": These variables are related to an attraction index. The average values are 294.20 and 13.42, respectively, with a range from a minimum to a maximum value.
- "rest_index" and "rest_index_norm": These variables are related to a restaurant index. The average values are 626.86 and 22.79, respectively, with a range from a minimum to a maximum value.
- "lng" and "lat": These variables represent the longitude and latitude coordinates of the rental properties.

The next step is the summary of the categorical variables in the dataset based on the output of the describe() function.

| | room_type | room_shared | room_private | host_is_superhost | multi | biz | city | type |
|---|---|---|---|---|---|---|---|---|
| count | 51707 | 51707 | 51707 | 51707 | 51707 | 51707 | 51707 | 51707 |
| unique | 3 | 2 | 2 | 2 | 2 | 2 | 10 | 2 |
| top | Entire home/apt | False | False | False | 0 | 0 | London | Weekends |
| freq | 32648 | 51341 | 33014 | 38475 | 36642 | 33599 | 9993 | 26207 |

*Figure 2 - Output of the "describe" function for categorical values*

Here are the key findings:

- "room_type": The dataset includes three unique types of rooms. The most common room type is "Entire home/apt", which appears 32,648 times in the dataset.
- "room_shared": This variable indicates whether a room is shared or not. The dataset contains two unique values, with "False" being the most frequent value, appearing 51,341 times.
- "room_private": This variable represents whether a room is private or not. It has two unique values, with "False" being the most common value, appearing 33,014 times.
- "host_is_superhost": This variable indicates whether a host is a superhost or not. It has two unique values, with "False" being the most frequent value, appearing 38,475 times.
- "multi": The variable denotes whether a property accommodates multiple guests or not. It has two unique values, with "0" being the most common value, appearing 36,642 times.
- "biz": This variable indicates whether a property is a business property or not. It has two unique values, with "0" being the most frequent value, appearing 33,599 times.
- "city": The dataset includes data for 10 unique cities. The city with the highest frequency is "London", which appears 9,993 times.
- "type": This variable represents the type of data (either weekdays or weekends). It has two unique values, with "Weekends" being the most common value, appearing 26,207 times.

In addition to the summary of categorical variables, it is important to note that the dataset does not contain any duplicates or null values. This means that each row in the dataset is unique, and there are no missing values in any of the categorical columns. The absence of duplicates and null values

ensures the integrity and completeness of the dataset, allowing for reliable analysis and conclusions based on the available data.

### 1.2 Distribution

In this section, we delve into the exploration of the "Airbnb Price Determinants in Europe" dataset, focusing on the distribution of numerical and categorical values. Analysing the distribution of numerical variables provides insights into central tendency, spread, and potential outliers. Visualising numerical distributions through histograms helps identify patterns and anomalies. Similarly, exploring the distribution of categorical variables reveals the prevalence and relationships between different categories. By examining these distributions, we aim to uncover valuable insights that may contribute to understanding the determinants of Airbnb prices in Europe. The results of this exploratory analysis serve as a foundation for subsequent modelling and further analysis in our endeavour to predict Airbnb prices accurately.

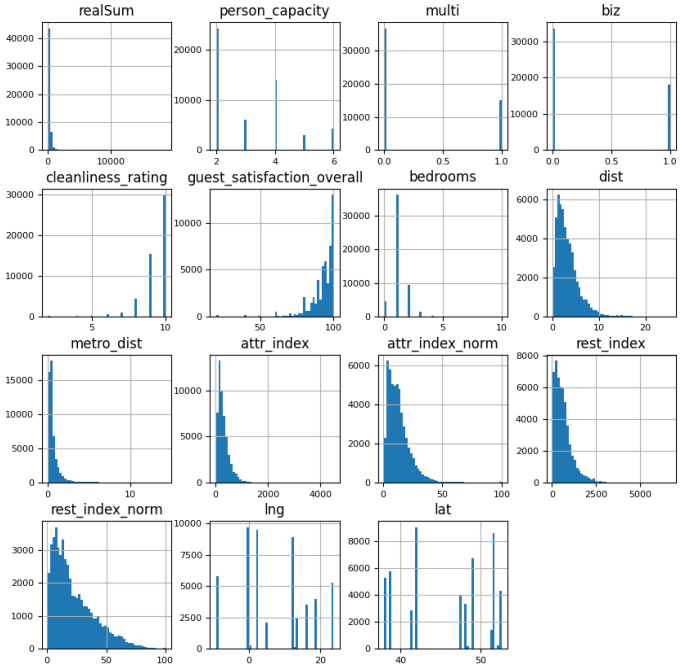


*Figure 3 - Distribution of numerical values*

In our analysis, we examined the histograms and statistical measures of several numerical variables in the dataset. Upon analyzing the histograms and statistical measures of the numerical variables in our dataset, we observed some interesting patterns and skewness in the distributions. "realSum" variable is skewed to the right. The tail of the distribution extends towards higher prices, indicating the presence of a few expensive listings. "multi" or "biz" both variables have a significant skewness towards the left, as indicated by their means being lower than their medians. This suggests that the majority of listings do not have the "multi" or "biz" feature. "Person_capacity", "cleanliness_rating", and "guest_satisfaction_overall" exhibit relatively symmetrical distributions. These variables have means and medians that are close to each other, indicating a balanced distribution of values. "Bedrooms", "dist", and "metro_dist show" minimal skewness, as their means and medians are quite similar. This suggests a relatively even distribution of values for these variables.

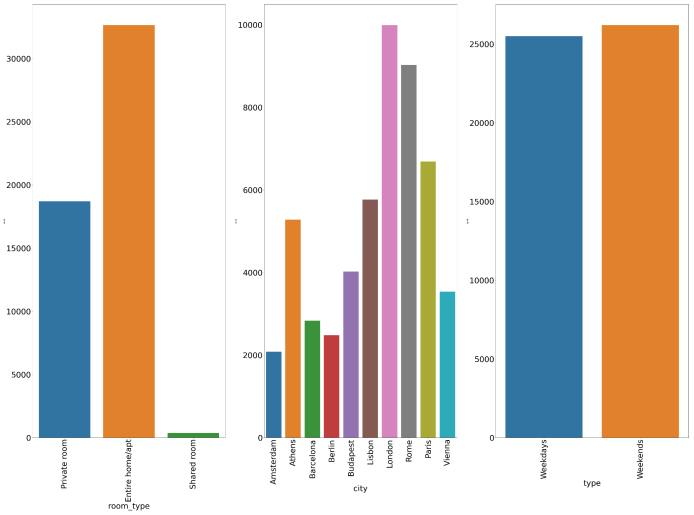


*Figure 4 - Distribution of categorical values*

The histograms offer a comprehensive view of the data distribution for the variables "room_type," "city", and "type." The most prevalent category among listings is "Entire home/apt," indicating a preference for full accommodations. London emerges as the dominant city, boasting the highest number of listings, followed by Rome, Paris, Lisbon, Athens, Budapest, Vienna, Barcelona, Berlin, and Amsterdam. In terms of the type of stay, the dataset exhibits a relatively balanced distribution between "Weekends" and "Weekdays." These histograms provide valuable insights into the composition of the dataset, enabling us to discern prominent categories and patterns within it.

### 1.3 Outliers

In this section, we focus on analysing the presence of outliers within the numeric variables of the dataset. Outliers can have a significant impact on analysis. To gain insights into their occurrence, we present boxplots for each numeric variable. These boxplots provide a visual representation of the data's distribution, including measures of central tendency and dispersion.

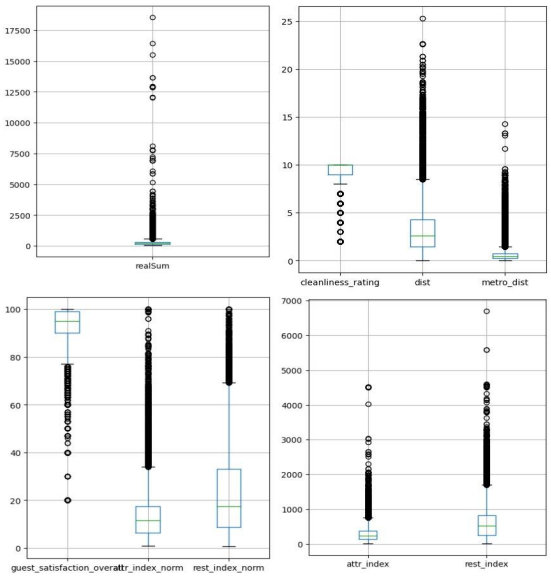


*Figure 5 - Boxplots of numerical values*

As we can observe, the box plots illustrate the presence of outliers in several attributes of the Airbnb Price Determinants in Europe dataset. These outliers represent data points that deviate significantly from the overall distribution of the data.

For the attribute "realSum," the box plot shows that there are some properties with extremely high or low total prices. These outliers may indicate unique or exceptional properties that stand out from the rest in terms of pricing.

In the attribute "guest_satisfaction_overall," outliers suggest properties with exceptionally high or low guest satisfaction ratings. These outliers could represent properties that have received overwhelmingly positive or negative feedback from guests.

The attributes "attr_index_norm" and "rest_index_norm" represent normalized attractiveness and restaurant indices, respectively. Outliers in these attributes indicate properties located in areas with extremely high or low levels of attractiveness or restaurant availability. These outliers may represent properties in highly popular or less-visited areas.

Similarly, the attributes "attr_index" and "rest_index" reflect the non-normalized attractiveness and restaurant indices. Outliers in these attributes provide insights into properties situated in areas with exceptional attractiveness or limited restaurant options.

### 1.4 Correlation

The correlation section of this report aims to analyze the relationships among the numerical variables in the dataset. Correlation measures the strength and direction of the linear relationship between two variables. By examining the correlation matrix, we can gain valuable insights into how different variables are associated with one another. This information can help us understand the dependencies and patterns within the data, which can be valuable for further analysis and decision-making. In this section, we will explore the correlations among the numerical variables and highlight key findings.

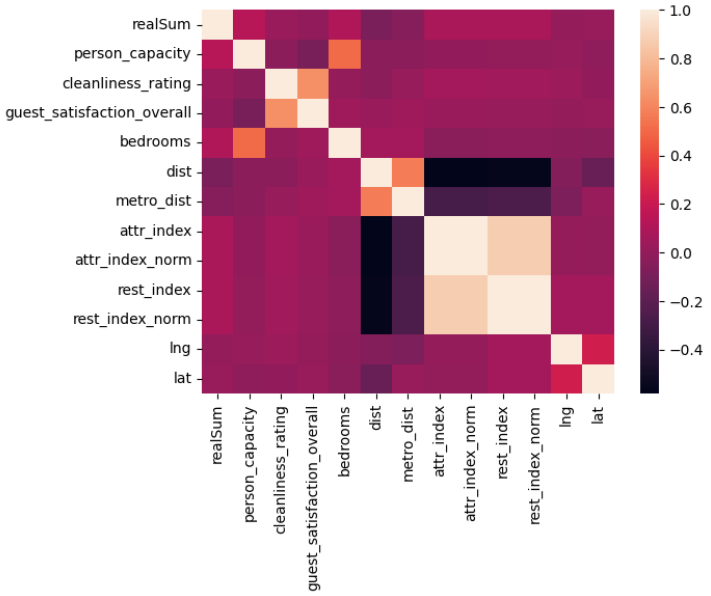


*Figure 6 - Visualisation for correlation matrix for numerical values*

The correlation matrix reveals several key points:

- "realSum" (total prices) has a positive correlation with "person_capacity" (accommodation capacity) and "bedrooms" (number of bedrooms), indicating that higher prices are associated with larger and more spacious listings.
- There is a positive correlation between "cleanliness_rating" and "guest_satisfaction_overall," suggesting that higher cleanliness ratings are associated with higher overall guest satisfaction.
- "dist" (distance from attractions) and "metro_dist" (distance from the nearest metro station) have negative correlations with "bedrooms," indicating that listings with more bedrooms tend to be located closer to attractions and metro stations.
- "attr_index" and "rest_index" (indices related to attractions and restaurants) have positive correlations with each other, suggesting that areas with more attractions tend to have more restaurants.
- The variables "attr_index_norm," "rest_index_norm," "attr_index," and "rest_index" have a strong positive correlation with each other, indicating that the normalized and non-normalized indices have similar relationships with other variables.
- There is no strong correlation between "realSum" and other variables like "cleanliness_rating," "guest_satisfaction_overall," "dist," and "metro_dist," suggesting that price is not directly influenced by these factors.

In conclusion, the correlation analysis reveals several key findings. There is a positive correlation between the total prices of Airbnb listings and their capacity. Higher cleanliness ratings are associated with higher overall guest satisfaction. The number of bedrooms shows a positive correlation with the accommodation's capacity. Additionally, accommodations that are closer to attractions tend to have higher attraction indices. These correlations provide valuable insights into the relationships among the numerical variables in the dataset.

### 1.5 General price overview

Current section provides a comprehensive analysis of the pricing patterns and trends in the Airbnb listings dataset. By delving into the pricing dynamics, we can gain valuable insights into the market and understand the factors that contribute to the variations in prices. This analysis will assist in making informed decisions related to pricing strategies and understanding the competitiveness of the listings in the dataset.

Firstly, we explored the total prices of accommodations across different cities and day types. A diagram illustrating the comparison of total prices is presented below to provide a visual representation of the data.

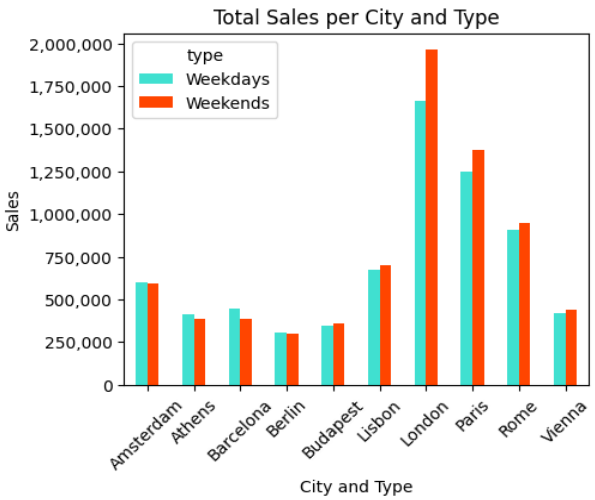


*Figure 7 - Total sales per city and day type*

When comparing the total sales for accommodations across different cities and day types, the following observations can be made:

- London has the highest total sales, with 1,662,102.82 on weekdays and 1,960,052.45 on weekends. It is followed by Paris with 1,248,202.30 on weekdays and 1,377,047.72 on weekends.
- Amsterdam, Barcelona, and Rome also show significant total sales, with Amsterdam having slightly higher sales compared to Barcelona and Rome.
- Lisbon and Vienna have relatively lower total sales compared to the other cities, but they still show a considerable amount of revenue.
- Athens, Berlin, and Budapest fall in the lower range of total sales among the cities mentioned.

In summary, London stands out as the city with the highest total sales for accommodations, followed by Paris. Amsterdam, Barcelona, Rome, Lisbon, and Vienna also demonstrate substantial sales. Athens, Berlin, and Budapest have comparatively lower total sales.
Secondly, we explored the average prices of accommodations across different cities and day types. A diagram illustrating the comparison of average prices is presented below to provide a visual representation of the data

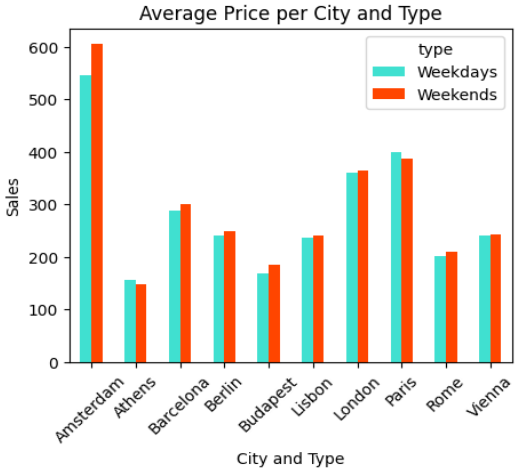


*Figure 8 - Average price of accommodation per city and day type*

The average prices of Airbnb listings vary across different cities and day types. Here is a comparison of the prices:

- Amsterdam has the highest average prices, with 545.02 on weekdays and 604.83 on weekends. It is followed closely by Paris with 398.79 on weekdays and 387.03 on weekends.
- Athens has the lowest average prices among the cities mentioned, with 155.87 on weekdays and 147.58 on weekends.
- London, Barcelona, and Berlin fall in the middle range of average prices, with London having slightly higher prices compared to Barcelona and Berlin.
- Budapest and Rome have relatively lower prices compared to the other cities, but they still show an increase during weekends.
- Lisbon and Vienna have similar average prices, with Lisbon having slightly lower prices on both weekdays and weekends.

In summary, Amsterdam stands out as the city with the highest average prices, while Athens has the lowest. The other cities fall in between, with London, Paris, Barcelona, and Berlin showing slightly higher prices, and Budapest, Rome, Lisbon, and Vienna having relatively lower prices.

In addition, we conducted a visualisation of a heatmap to analyse the density of housing points for all 10 cities based on their respective coordinates. As an example the Amsterdam map is presented below.

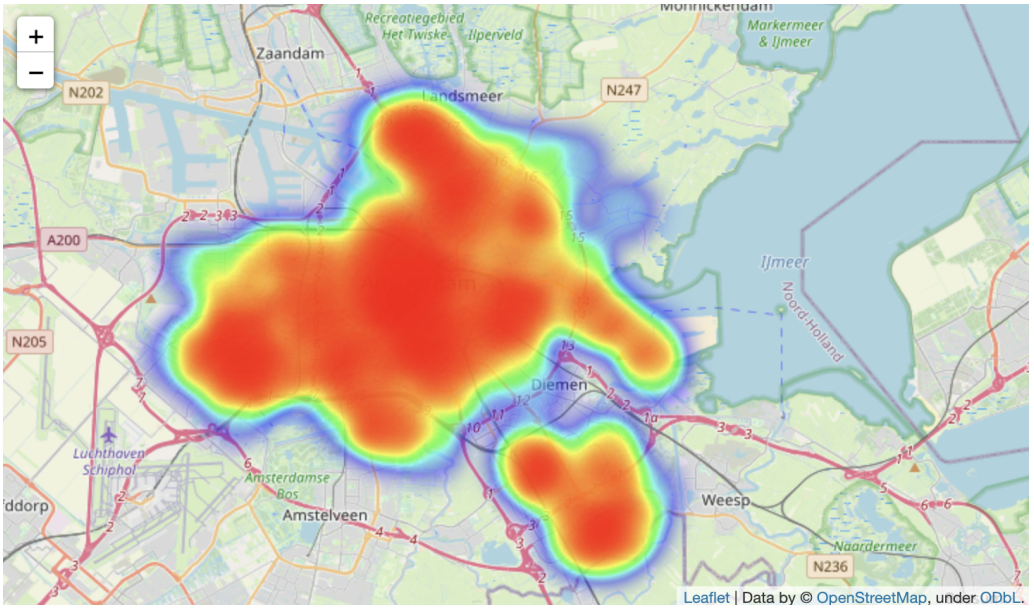


*Figure 9 - Density heatmap for Amsterdam*

This heatmap provides valuable insights into the distribution and concentration of accommodations in different areas. We observed that the city centres tend to have the highest density of housing options, while certain areas exhibit a lack of available accommodations. Further investigation and additional information are necessary to uncover any hidden patterns or factors influencing the housing locations across the cities. The heatmap visualisation for Amsterdam, along with other cities, can be found in the ExploratoryAnalysis.ipynb notebook.

Furthermore, we conducted a visualisation of housing points analysed by their prices using colour-coded scales for all 10 cities. This visualisation provides a clear picture of the price distribution and identifies any outliers within the dataset. The map for Budapest is presented below.

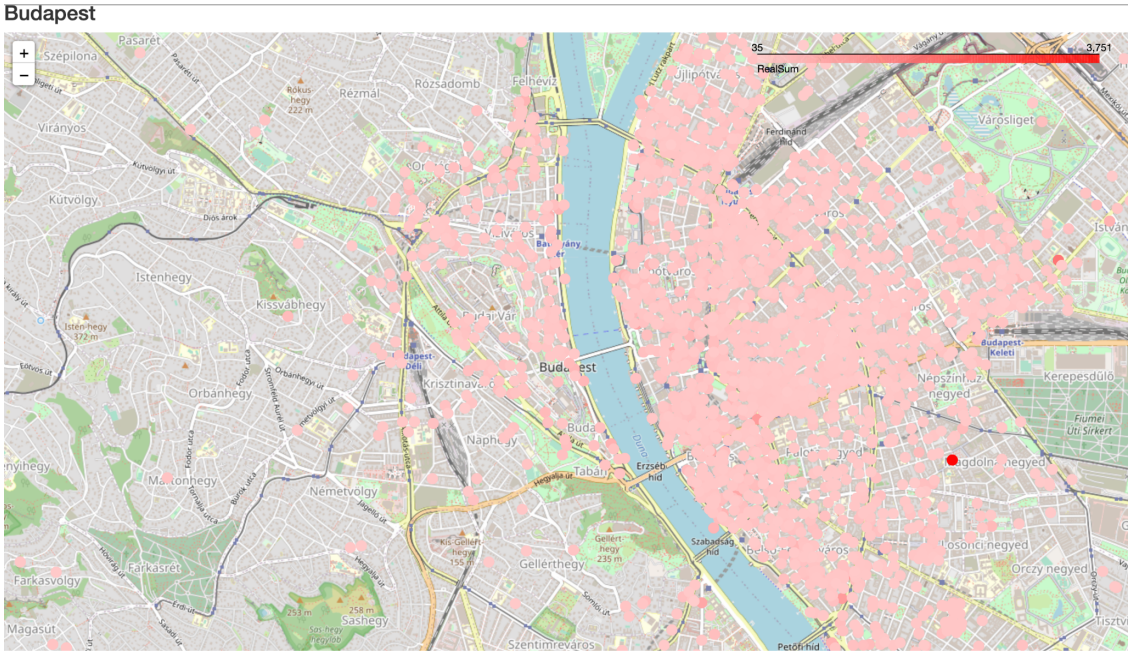


*Figure 10 - Map of Budapest with housing points analysed by their prices*

This visualization will be further explored in a separate section dedicated to outlier detection and analysis.

## 2. Data Preprocessing and Cleaning

### 2.1 Outlier analysis

In order to identify outliers in the dataset, we employed the Z-score model for each city, focusing on the "realSum" variable, which is known to contribute to the majority of outliers. The Z-score calculates the deviation of each data point from the mean in terms of standard deviations. We set the threshold at 3 standard deviations.

After applying the Z-score model, we successfully detected outliers in the dataset. Specifically, a total of 421 outliers were identified and subsequently removed from the analysis. This step helps to ensure that the remaining data accurately represents the majority of observations and reduces the impact of extreme values on our analysis.

In addition to the Z-score model, we also performed outlier analysis using the Isolation Forest algorithm on other columns. We utilized the IsolationForest model with a contamination rate of 0.1 and a random state of 42. By applying this model, we were able to identify and isolate outliers in the dataset. As a result, we obtained a set of remaining predictors without multicollinearity, which includes variables such as 'room_shared', 'room_private', 'person_capacity', 'host_is_superhost', 'multi', 'biz', 'cleanliness_rating', 'guest_satisfaction_overall', 'bedrooms', 'dist', 'metro_dist', and 'lng_London'. These predictors were deemed to be reliable and independent for further analysis. The detailed results and analysis of the outlier detection process can be found in the FinalModel_General.ipynb notebook.

The removal of outliers has had a significant impact on improving the final model results. By eliminating these extreme values, we have reduced the potential influence of outliers on our analysis and modeling process. This step ensures that our final model is more robust and provides more accurate insights into the underlying patterns and relationships within the data.

### 2.2 Data format

In this stage, we standardise the format of categorical variables, ensuring consistency and enhancing their usefulness in subsequent analysis. For instance, categorical binary variables like room_shared, room_private, and host_is_superhost are transformed into numerical equivalents of 1 and 0. Furthermore, we employ one-hot encoding for city and type variables, allowing us to effectively incorporate these variables into the regression model.

In addition, we have implemented a data pre-processing step that involves integrating location information from Airbnb listings. Recognizing the significant impact of geographic coordinates on prices, we have added latitude and longitude variables for each of the 10 cities in our dataset (Amsterdam, Athens, Barcelona, Berlin, Budapest, Lisbon, London, Paris, Rome, and Vienna). This inclusion ensures that our model captures the specific location details associated with each listing, allowing it to account for spatial variations in pricing trends. Adding this information allows us to discover patterns and correlations specific to different areas and neighborhoods within each city, which improves the accuracy of our price predictions.

## 3. Feature selection

In the variable selection process, we utilized three techniques: correlation-based selection, KBest, and cross-validation feature selection. The first model involved conducting a correlation analysis to determine the relationships between the features and the target variable. Based on the absolute correlation values, we selected the most influential features. In the second model, SelectKBest with the f_regression scoring function was employed to identify the best important features using logistic regression. Lastly, the third model incorporated Cross Validation and utilized GradientBoostingRegressor as the estimator. This model calculated the mean absolute error and performed cross-validation to identify the optimal combination of variables based on this metric. It is important to note that the metrics were calculated using the logarithmic transformation of the target variable (price), as detailed in the next section.

After implementing and comparing these methods, we concluded that the best model for this case was achieved through cross-validation feature selection with GradientBoostingRegressor as the estimator. This approach allowed us to select the most significant variables that greatly impact the prediction of Airbnb prices. The results can be seen in the following image:

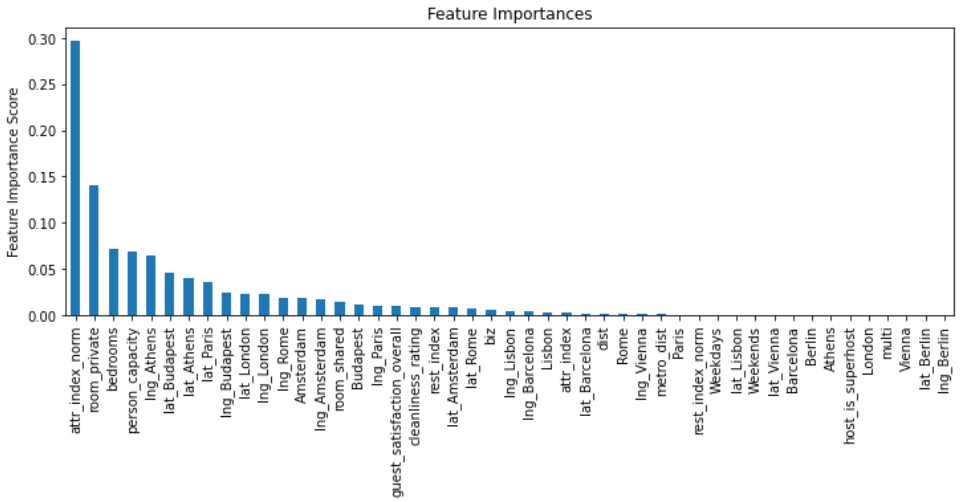


*Figure 11 - Feature selection scores using GradientBoostingRegressor estimator*

The feature selection process using the GradientBoostingRegressor model reveals the importance of several variables in predicting Airbnb prices. In this case, the mean absolute error was 0.2209. The importances of the features were found in the range of 0 to 0.297103. In particular, variables like "attr_index_norm", "room_private", "bedrooms" and "person_capacity" showed relatively high importance. These variables capture important attributes related to property attractiveness, room type, number of rooms, and capacity.

In addition, location-related variables such as "lng_Athens", "lat_Budapest", "lat_Athens" and "lat_Paris" also demonstrate significant contributions. These variables reflect the geographic coordinates of specific cities, indicating that location plays a role in pricing decisions. Additionally, features such as guest satisfaction ratings, cleanliness ratings, and distance-related metrics show moderate importance.

It is important to note that variables with significance greater than 0.0005 are considered significant contributors to the model's predictions. These importances represent the relative importance of each variable within the GradientBoostingRegressor model, with higher scores indicating greater predictive power.

Overall, the feature selection process allows us to identify the most influential variables in predicting Airbnb prices. With this we can build a robust predictive model that focuses on the essential factors that affect pricing decisions. Ensuring that the model avoids incorporating noisy or irrelevant features, leading to a more accurate and reliable prediction of Airbnb prices in the market.

**4. Data Transformations**

In order to improve the performance of the model, a logarithmic transformation was applied to the target variable (realSum), using the expression np.log(y_n + 0.000001). When performing this transformation, we use the logarithmic scale of the target variable during training and the calculation of the metric.

The application of a logarithmic transformation to the target variable, in this case the price, aims to improve the performance of the model. This transformation uses logarithmic scales during training and metric calculation, which offers several benefits to improve the effectiveness of the model. First, it helps mitigate the impact of heteroscedasticity, which refers to the unequal variance of errors in regression models. By reducing the influence of heteroskedasticity, the transformation promotes a more consistent and stable relationship between the predictor variables and the target variable. Additionally, the log transformation addresses skewness in the price distribution, resulting in a more symmetric distribution suitable for the regression model. Additionally, it can help linearize the relationship between the predictors and the target variable, allowing the model to capture underlying patterns more accurately.

To address potential skewness in the predictor variables, we applied a series of transformations. We conducted a skewness analysis to identify variables exhibiting skewness, which indicates departures from symmetry in the distribution of data. Positive skewness suggests a longer tail on the right side, while negative skewness indicates a longer tail on the left side. For variables with left skewness (negative skewness), indicating a longer tail on the left side, a power transformation was applied. The expression np.power(X[left_columns], 3) was used, raising the variable to the power of 3. This choice of power transformation with an exponent of 3 was made to further emphasize the correction of the left skewness and shift the distribution towards a more symmetrical form.

On the other hand, for variables displaying right skewness (positive skewness), indicating a longer tail on the right side, a logarithm transformation was used. The expression np.log(X[right_columns] + 0.000001) was employed, with the addition of a small constant (0.000001) for numerical stability and to handle cases with zero or near-zero values. The logarithm transformation compresses larger values and expands smaller values, effectively reducing the right skewness and achieving a more symmetric distribution. These transformations help ensure that the assumptions underlying the regression models, such as linearity and homoscedasticity, are more likely to hold.

**5. Model Selection and Training**

The metric chosen to evaluate the performance of the model was the mean absolute error (MAE), this offers interpretability, robustness against outliers and a clear understanding of the average magnitude of the prediction errors. MAE measures the average absolute difference between the predicted and actual values, providing a direct and intuitive measure of model accuracy. It is less sensitive to outliers compared to other metrics such as root mean square error, which makes it suitable for this case, where outlier issues were identified. In addition, MAE preserves the original unit of the target variable, allowing for meaningful interpretation of the metric in real-world applications.

For model selection, four regression models were trained and compared: GradientBoostingRegressor, LinearRegression, RidgeCV, and LassoCV. The results of each model are summarized in the table below:

| **Model Name** | **Train MSA (log y)** | **Test MSA (log y)** | **Test MSA (y)** | **R2 (y)** |
|---|---|---|---|---|
| Gradient Boosting Regressor | 0.223 | 0.213 | 54.838 | 0.661 |

| | | | | |
|---|---|---|---|---|
| Linear Regression | 0.243 | 0.268 | 68.253 | 0.592 |
| Ridge | 0.413 | 0.832 | 229.225 | -3.074 |
| Lasso | 0.379 | 0.420 | 103.430 | 0.208 |

Based on the chosen evaluation metric, mean squared error (MSA), the Gradient Boosting Regressor model was selected as the best performer. The train MSA and test MSA (log y) metrics represent the mean squared error calculated based on the transformation of the target variable. On the other hand, the test MSA (y) and R2 (y) metrics were obtained by applying the inverse transformation to obtain the original variable and facilitate result interpretation. The Gradient Boosting Regressor exhibited the lowest MSA values, indicating better performance in predicting the target variable compared to the other models.

Moreover, the selected model, Gradient Boosting Regressor, was further tuned by applying tuning parameters to optimize its performance. The tuning process involved using the scikit-learn GridSearchCV function and evaluating different combinations of hyperparameters. The parameter grid used for the grid lookup included three different loss functions (huber, ls, and lad), three values for alpha (0.9, 0.95, and 1.0), and three options for the number of estimators (100, 200, and 300). Grid search was performed using 10-fold cross-validation and negative mean absolute error as the scoring metric.

The best model was obtained along with its corresponding optimal parameters. The best parameters for the Gradient Boosting Regressor model were found to be: alpha = 0.9, loss = huber, and n_estimators = 300. These parameters were selected based on their ability to minimize negative mean absolute error during cross-validation.

Then, the best model on the full training set was trained using the optimal parameters. It was found that the selected parameters, obtained through the grid search, gave the best results in terms of mean absolute error and R2 score in the test set. These were 54.838 EUR( y) and the R2 score was 0.661.

**6. Model Evaluation and results**

For our predictive model, we chose the Gradient Boosting Regressor. Using this model, we observed that the predicted prices differed from the baseline value of 54.838. Since the currency information was not provided in the Kaggle dataset, we relied on the conditional monetary units implied by the dataset creator, which could be either dollars or euros.

To analyze the error difference between the predicted price and the average price for each city and day type (weekday or weekend), we calculated the percentage deviation. We divided 54.838 by the average housing price and multiplied it by 100%. We also created visualisations to provide an overview of the results and the performance of our model that is provided below.

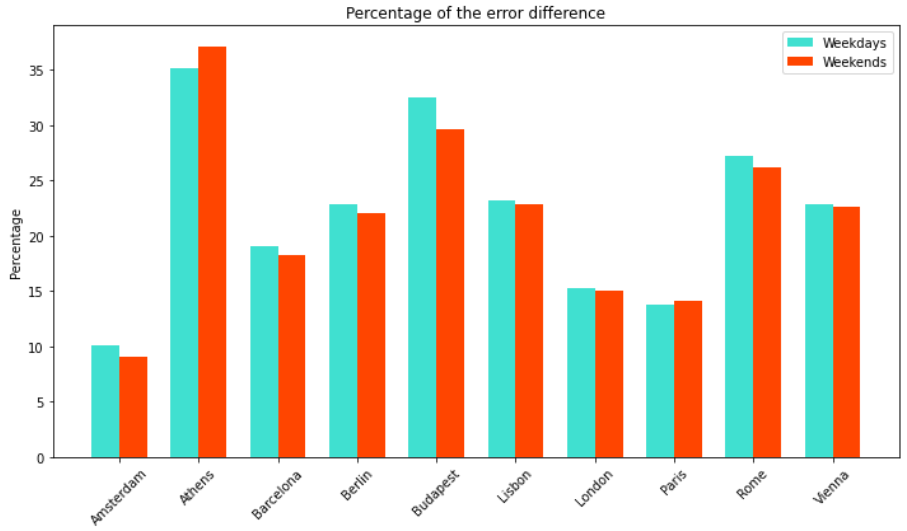


*Figure 12 - The percentage deviations of predicted housing prices to average prices*

In comparing the percentage deviations of predicted housing prices to average prices across the cities, we can observe the following trends:

- Amsterdam stands out with the lowest deviation of 9.91% for both weekdays and weekends, indicating a relatively accurate prediction of housing prices. This suggests that the model performs well in capturing the price dynamics specific to Amsterdam.

- Athens exhibits the highest deviation among all cities, with a considerable 34.64% difference between predicted and average prices for both weekdays and weekends. This indicates a larger discrepancy in the model's ability to accurately predict housing prices in Athens. Factors unique to the Athens market might contribute to this higher deviation.

- Barcelona, Berlin, Budapest, Lisbon, London, Paris, Rome, and Vienna display relatively similar levels of deviation, ranging from 13.54% to 32.06% across both weekdays and weekends. These cities demonstrate moderate variations between the predicted prices and the average prices. Further analysis is necessary to understand the factors influencing these discrepancies.

Overall, the comparison highlights variations in the model's predictive performance across different cities. It suggests that the model performs relatively well in accurately predicting prices in Amsterdam, while Athens poses a greater challenge due to larger deviations. The other cities show moderate deviations, indicating potential room for improvement in capturing the price dynamics accurately.

**Conclusion**

In conclusion, this report analysed Airbnb housing prices in 10 European cities using the Airbnb Price Determinants in Europe dataset. The analysis encompassed data for weekdays and weekends, examining various factors influencing pricing trends.

The chosen model, Gradient Boosting Regressor, was employed to predict housing prices, resulting in a deviation of 54.838 from the actual prices. Further evaluation of the deviation was performed by comparing it to the average prices for each city and day type. The percentage deviations revealed variations across the cities, with Amsterdam displaying the lowest deviation and Athens exhibiting the highest.

In light of the findings, several proposals for further analysis can be suggested. Firstly, conducting a deeper exploration of the factors influencing housing prices in Athens would help address the observed higher deviation. Additionally, investigating the underlying patterns of housing locations in the cities with the highest deviations could provide valuable insights into the factors contributing to these discrepancies.

Furthermore, incorporating additional datasets, such as socio-economic indicators or tourism statistics, could enhance the predictive model's accuracy and capture more comprehensive insights into pricing determinants. As another approach there can be done separate ML analysis for each city separately in order to improve the accuracy of the predicted prices and help customers and Airbnb companies.

Overall, this report presents a comprehensive analysis of Airbnb housing prices in European cities, shedding light on key trends and patterns. The findings and proposals for further analysis provide a solid foundation for future research and decision-making in the domain of short-term accommodation rentals.

**References**

1. Airbnb Price Determinants in Europe. (2023, February 13). Kaggle. Retrieved from https://www.kaggle.com/datasets/thedevastator/airbnb-price-determinants-in-europe
2. Documentation scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation. (n.d.). https://scikit-learn.org/0.21/documentation.html