Project Report

GitHub URL

https://github.com/ConorSaund/UCDPA_conorsaunders

Abstract

In this project, the analysis was of the pricing of Airbnb listings in ten European capital cities; Amsterdam, Athens, Barcelona, Berlina, Budapest, Roma, Vienna, Lisbon. The dataset consisted of twenty files, including weekday and weekend data for each city. The aim is to gain insights into the factors affecting the pricing of Airbnb listings in these cities.

Introduction

The various attributes in the dataset, such as pricing, satisfaction levels, and location to a metro, provide a rich source of information that can be used to extract insights. Working with this dataset requires creative thinking and the application of various analytical techniques to derive meaningful results.

Dataset

The dataset used was Airbnb pricing in ten European capital cities. The attributes analysed in the data sets were;

- realSum
- room_type
- room_private
- host_is_superhost
- cleanliness_rating
- guest_satisfaction_overall
- bedrooms
- dist
- metro_dist
- lng
- lat

The data sets consist of weekday data and weekend data for the ten cities, resulting in twenty files with which to work. There were approximately 51’700 observations of non-null entries.

The dataset was chosen as it provides an interesting opportunity to analyse the Airbnb market in a handful of European tourist destinations. The dataset was also accredited by the National Science Centre in Poland under project 2017/27/N/HS4/00951. By examining the pricing trends and characteristics of Airbnb listings in these cities we can gain insights into the broader tourism sector, and with further analysis it could be used to identify future areas of growth.

Implementation Process

The initial code is filled by Kaggle, as I used their online notebook for coding. It is the default code used to incorporate datasets straight from Kaggle. The datasets used throughout the assignment were imported using an integration that took all CSV files available from the domain in Kaggle. Following this, an initialisation of the main packages that were to be used throughout the code was needed.

A load of all of the data that was to be used was then carried out. This consisted of twenty CSV files, of ten European cities of weekday data and weekend. A list was created of each of the links for the CSV files. This was called 'dataframe' for ease of use as it was not going to be used past this point. Before we could concatenate all the files and condense them into one file, it is important to ensure that all files have the same number of columns. This was carried out with a 'for' loop to iterate through the files. 'df.shape' was used to output the number of rows and columns in each file. 'df.columns' was also used to output the headers of each file. A manual check was then carried out to ensure the number of columns were the same along with the headers.

A function was later defined to concatenate two CSV files into one file. This was used for the weekday and weekend pricings of each city to be merged into one file for each city and any errors or blanks removed.

The next block of code for "1. Comparison in room types", a comparison was carried out of room types from across all of the cities in the datasets. A grouping was done of each "city" and "room_type", sorting of the room types by count in a descending order, and formatting of the graph and capitalisation of the x-axis labels was carried out.

A second graph was created to show the mean and median pricing of the house lets. Similar to the formatting of the previous graph, Capitalisation was carried out for the x-axis labels, and a table formed for an easier overview of the pricings.

The second insight used for the datasets of Airbnb lettings amongst ten European cities was Superhost proportions in the lets. The code is to iterate through the cities and find the counts of superhosts. A calculation is made showing the proportion of Superhosts per city. A basic table is printed showing the mean satisfaction level of each "room_type".

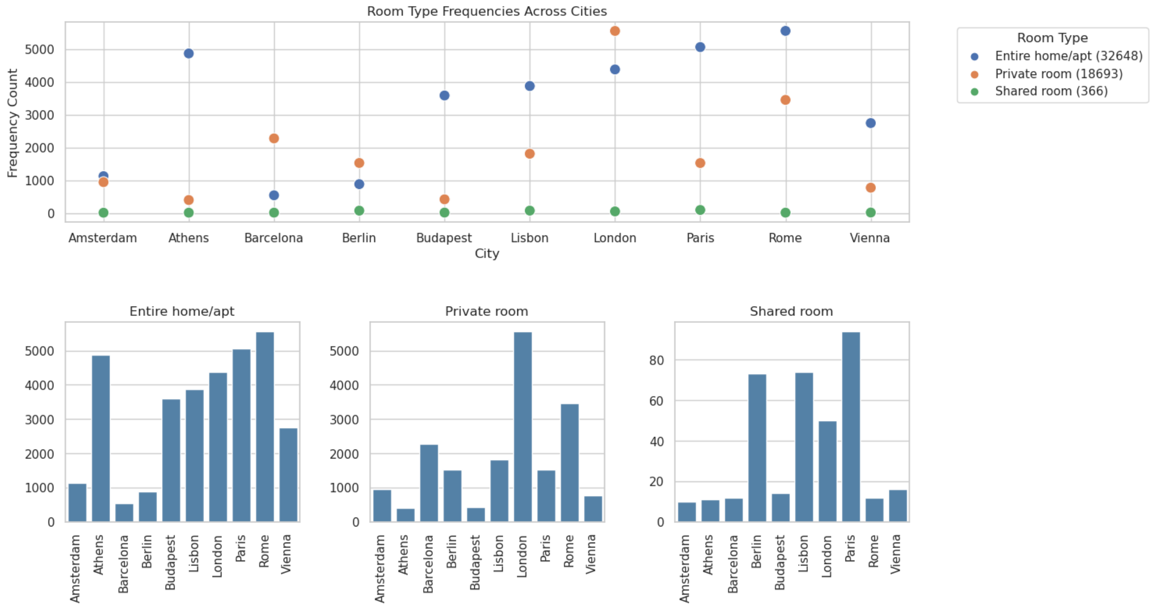
A multiple linear regression was carried out on the independent variables of "cleanliness_rating" and "metro_dist" along with a dependent variable of "realSum", to explore the relationship between them. The first step was to check for any missing or error values in the dataset using ".isnull()". This returned a boolean for each cell indication whether it contained a missing value or not, any empty cells were removed. A check for incorrect values was carried out using ".describe()", and further went on to remove any outliers that would skew the data. Calculations of all necessary data were carried out and a regression model used, which produces an output to interpret.

Another linear regression analysis was carried out on the distance from metros, cleanliness rating, number of rooms and the attraction index. The code was to first calculate the average distance to metros for each room type. A bar plot is then formed for each room type. After the plot was created, a linear regression model was used, OLS - Ordinary Least Squares method. The regression output shows various statistics including the R-squared value which indicates the proportion of the variation in the dependent variable by the independent variable.

The final block was divided into two parts, a heat map and cluster map showing the longitudinal and latitudinal locations of each Airbnb location as was given in the dataset. These maps show the distribution of the locations around each city, for which we can determine if locations are in tourist areas, and if this influences a higher average price. The first part imports the relevant packages and creates a centred map of the cities. It creates a MarkerCluster overlay on the map and allows for interactive uses on the map. The second map shows a similar result but leaning on a heatmap overlay.

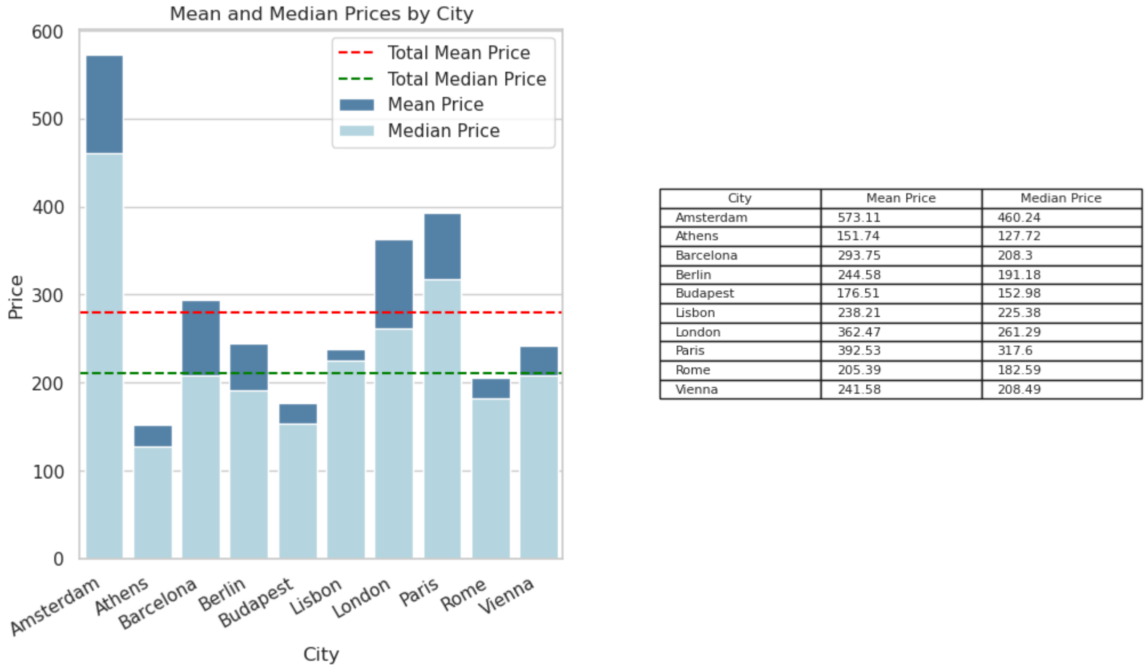
There are many ways that machine learning can be used. For example we can use machine learning and deep learning to analyse data on customer behaviour and interactions within a company. A model can then be trained to predict if a customer is at risk of leaving the company in the future. Whether to use classification or regression models depends on the type of prediction the desired would be. Classification models are used on categorical variables whilst regression is used when the output is continuously variable, similar to the prediction of a price at a certain time, such as Figure 6. For example, if a customer is to leave the company in the future, classification methods would be ideal as we would be categorising if a customer is to leave or not. It would be one or the other and could not be in the middle. The output could be boolean of True or False.

Results



(figure 1)

Figure 1 shows the number of room types amongst the ten European cities. It is clear that the most common property type was an "Entire home/ apt", followed by " Private room" and lastly "Shared room" with the lowest amount.



City	Mean Price	Median Price
Amsterdam	573.11	460.24
Athens	151.74	127.72
Barcelona	293.75	208.3
Berlin	244.58	191.18
Budapest	176.51	152.98
Lisbon	238.21	225.38
London	362.47	261.29
Paris	392.53	317.6
Rome	205.39	182.59
Vienna	241.58	208.49

(figure 2)

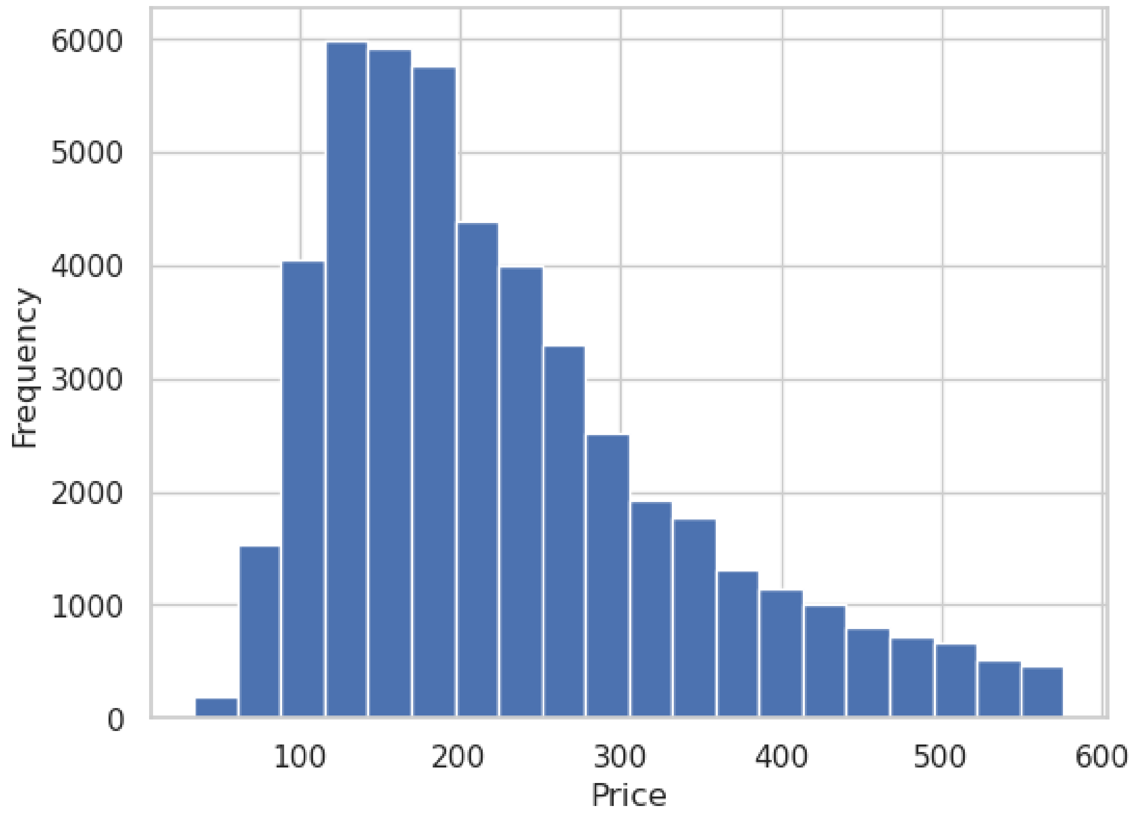
Figure 2 shows the mean and median prices of the European cities. From the graph and table we can see that Paris has the highest mean price at 392.53, which is higher than most of the other European cities. Athens has the lowest mean price at 151.74, which is significantly lower than the other cities. This suggests that Athens could be a relatively affordable city to visit or rent in. The mean price for all cities is greater than the median signifying a skew towards the higher end. Amsterdam has a significantly higher pricing compared to all of the other cities, showing that accommodation can be quite expensive and driving up the pricing. Lisbon, Barcelona and Vienna all have relatively similar mean and median pricing, which suggests the distribution of prices are relatively even in the aforementioned cities. London has a relatively high mean price but lower median. This shows that London has some Airbnb listings which are driving up the average pricing in the area.

```
Proportion of superhosts: False    0.744097
True     0.255903
Name: host_is_superhost, dtype: float64
Mean guest satisfaction by room type:
room_type
Entire home/apt    92.888691
Private room       92.231156
Shared room        89.674863
Name: guest_satisfaction_overall, dtype: float64
```

(figure 3)

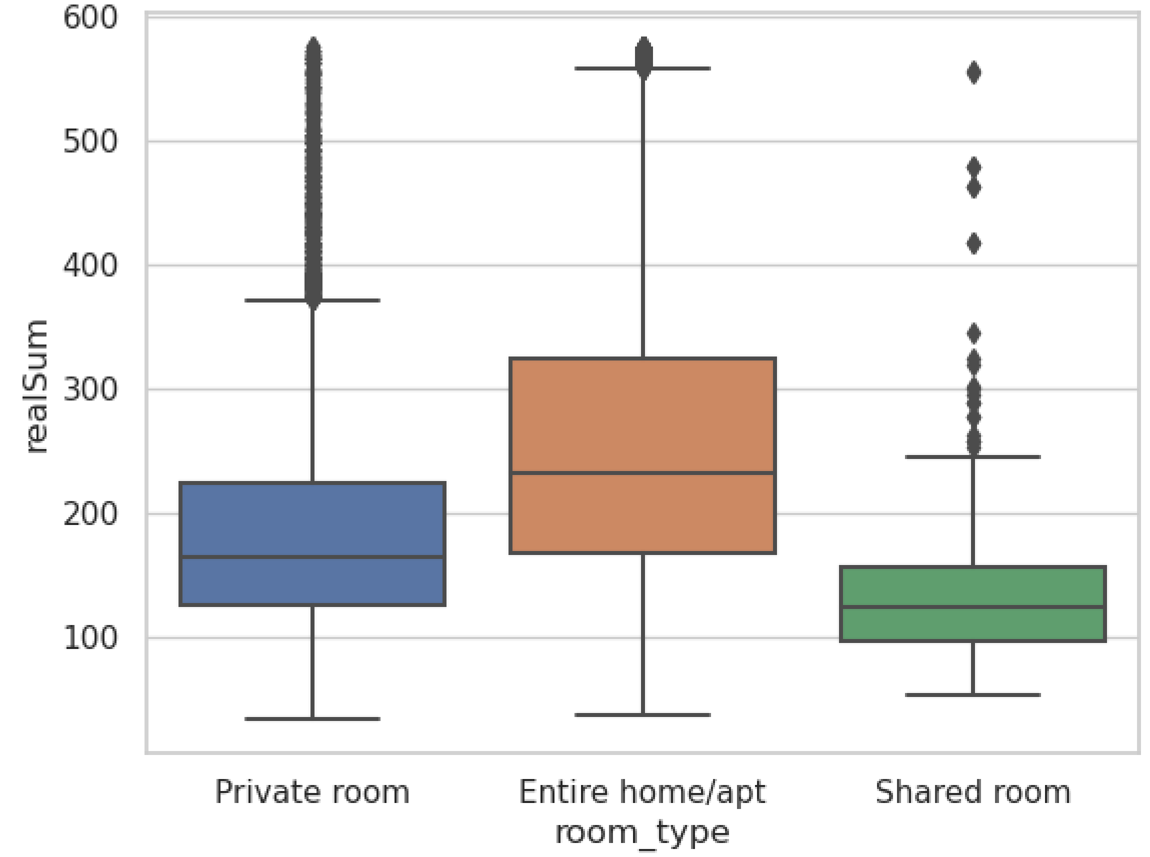
Figure 3 shows the proportion of Superhosts amongst the datasets. Working through the output it can be noted that approximately 25% of hosts were superhosts, and therefore majority of the stays were with non-superhosts.We can also see from the output that lets with an "Entire home/ apt" had higher satisfaction ratings than "Private room" or "Shared room".

More research would be required to investigate the disparity in superhost use. It is interesting that more customers did not choose lets with Superhosts. Typically a let with a superhost would generally increase the price for the extra care and attention that the customer would receive. It would be an interesting collection of data to see the reasoning behind not selecting a superhost.



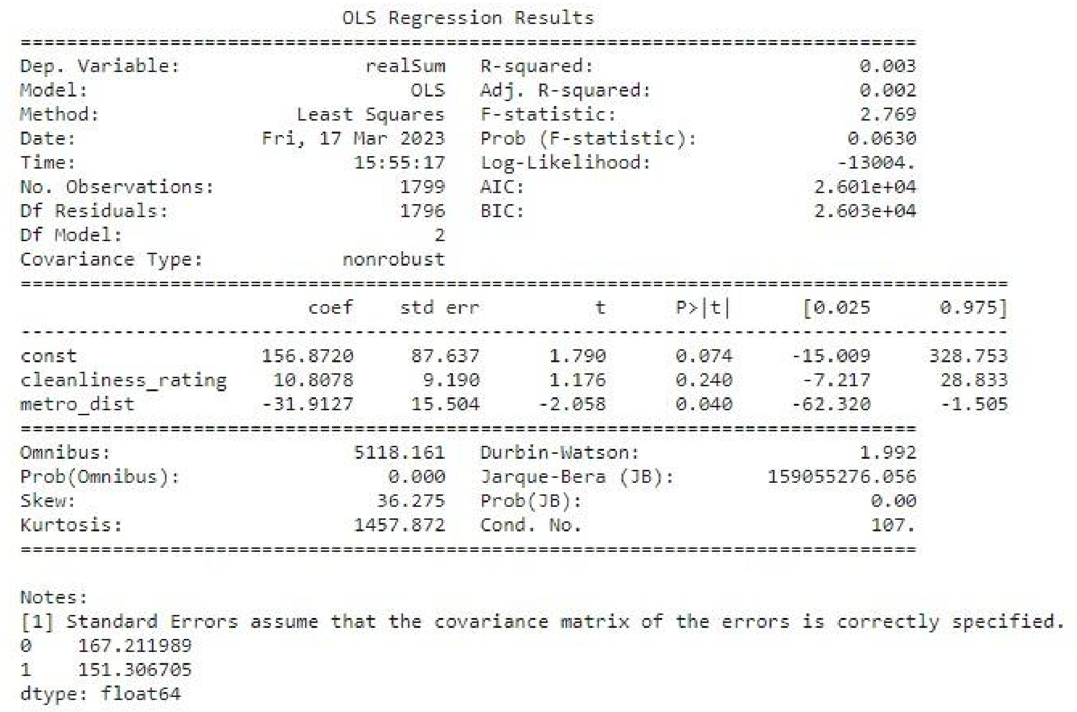
(figure 4)

Figure 4 shows a bell-shaped histogram of the frequency of "realSum", or the pricing of the letting. Combining this figure with Figure 1, we note that the graph is skewed due to the number of units let for "Entire home/apt" compared to the cheaper comparison of "Shared room". The mode of the graph was approximately 120.00, mean roughly 280.00 and median 220.00.



(figure 5)

Figure 5 shows a boxplot of the room types against the pricing. A box plot is a standardised way of displaying the distribution of data based on Q1, Q3, minimum, maximum and median. We can immediately see the number of outliers for all of the room types. All of the boxplots' minimum was roughly the same, with "Shared room"'s maximum being significantly less than the other room types. "Entire home/ apt" maximum was substantially greater. All of the boxplots above are all skewed towards the upper end of the pricing, signifying there are more properties at a higher value than lower.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                realSum   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     2.769
Date:                Fri, 17 Mar 2023   Prob (F-statistic):             0.0630
Time:                        15:55:17   Log-Likelihood:                -13004.
No. Observations:                1799   AIC:                         2.601e+04
Df Residuals:                    1796   BIC:                         2.603e+04
Df Model:                           2
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                156.8720     87.637      1.790      0.074     -15.009     328.753
cleanliness_rating    10.8078      9.190      1.176      0.240      -7.217      28.833
metro_dist           -31.9127     15.504     -2.058      0.040     -62.320      -1.505
==============================================================================
Omnibus:                     5118.161   Durbin-Watson:                   1.992
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        159055276.056
Skew:                          36.275   Prob(JB):                         0.00
Kurtosis:                    1457.872   Cond. No.                         107.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
0    167.211989
1    151.306705
dtype: float64
```

(figure 6)

Figure 6 shows the summary output of an OLS (ordinary least squares) linear regression model, which is used to predict rental prices, the dependent variable, based on two independent variables, cleanliness rating and distance to the nearest metro station.

The R-squared value represents the proportion of variance in the rental prices that is explained by the independent variables in the model. In this case, the R-squared is very low at 0.003, indicating that the model is not very effective at explaining the variation in rental prices.

The coefficients represent the estimated effect of each independent variable on rental prices. In this case, the cleanliness rating has a positive coefficient of 10.81, indicating that as cleanliness rating increases, rental prices tend to increase as well.

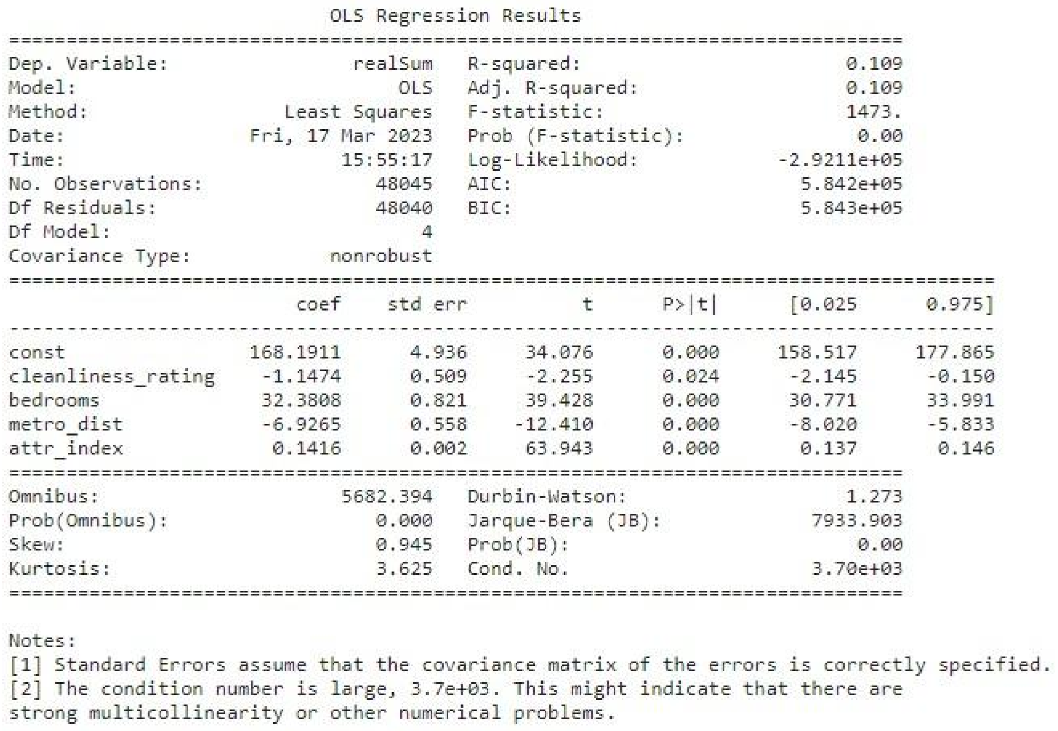
The P-value indicates the significance of each coefficient, or the likelihood that the observed effect of each independent variable on rental prices is due to chance. In this case, the p-value for cleanliness rating is 0.24, which is not statistically significant at the conventional level of 0.05. The p-value for distance to the metro station is 0.04, which is statistically significant at the 0.05 level.

The distance to the nearest metro station has a negative coefficient of -31.91, indicating that as distance to the metro station increases, rental prices tend to decrease.

The confidence interval values indicate the range of values within which the true coefficients are likely to fall with a certain level of confidence. In this case, we can be 95%

confident that the true value of the cleanliness rating coefficient falls between -7.22 and 28.83, and that the true value of the distance to the metro station coefficient falls between -62.32 and -1.51.

Finally, the output includes the predicted rental prices for two new observations with cleanliness ratings of 4.5 and 4.8, and metro distances of 1.2 and 1.8. The predicted prices are 167.21 and 151.31, respectively.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                realSum   R-squared:                       0.109
Model:                            OLS   Adj. R-squared:                  0.109
Method:                 Least Squares   F-statistic:                     1473.
Date:                Fri, 17 Mar 2023   Prob (F-statistic):               0.00
Time:                        15:55:17   Log-Likelihood:            -2.9211e+05
No. Observations:               48045   AIC:                         5.842e+05
Df Residuals:                   48040   BIC:                         5.843e+05
Df Model:                           4
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                168.1911      4.936     34.076      0.000     158.517     177.865
cleanliness_rating    -1.1474      0.509     -2.255      0.024      -2.145      -0.150
bedrooms              32.3808      0.821     39.428      0.000      30.771      33.991
metro_dist            -6.9265      0.558    -12.410      0.000      -8.020      -5.833
attr_index             0.1416      0.002     63.943      0.000       0.137       0.146
==============================================================================
Omnibus:                     5682.394   Durbin-Watson:                   1.273
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7933.903
Skew:                           0.945   Prob(JB):                         0.00
Kurtosis:                       3.625   Cond. No.                     3.70e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.7e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(figure 7)

The output above is the result of running an OLS (Ordinary Least Squares) regression analysis on the provided data. The model aims to predict the dependent variable "realSum" (i.e., the actual price of the accommodation) based on the independent variables "cleanliness_rating," "bedrooms," "metro_dist," and "attr_index."

The R-squared value of the model is 0.102, which indicates that only 10.2% of the variation in the "realSum" variable is explained by the independent variables included in the model. The adjusted R-squared value is the same, which suggests that adding or removing independent variables from the model does not significantly affect the explanatory power of the model.

The coefficients of the independent variables indicate the direction and magnitude of their effect on the dependent variable. The intercept (or constant) term is 163.9941, which represents the expected value of "realSum" when all independent variables are equal to zero.

The coefficient of "cleanliness_rating" is -0.3427, which means that a one-unit increase in the cleanliness rating is associated with a decrease of 0.3427 units in "realSum," holding all other variables constant. However, this coefficient is not statistically significant (i.e., P>|t| > 0.05), indicating that cleanliness_rating does not have a significant effect on the actual price of the accommodation.

The coefficient of "bedrooms" is 24.4770, which means that a one-unit increase in the number of bedrooms is associated with an increase of 24.4770 units in "realSum," holding all other variables constant. This coefficient is statistically significant (i.e., P>|t| < 0.05), indicating that the number of bedrooms has a significant effect on the actual price of the accommodation.

The coefficient of "metro_dist" is -5.9582, which means that a one-unit increase in the distance to the nearest metro station is associated with a decrease of 5.9582 units in "realSum," holding all other variables constant. This coefficient is statistically significant (i.e., P>|t| < 0.05), indicating that the distance to the nearest metro station has a significant effect on the actual price of the accommodation.

The coefficient of "attr_index" is 0.1269, which means that a one-unit increase in the attraction index is associated with an increase of 0.1269 units in "realSum," holding all other variables constant. This coefficient is statistically significant (i.e., P>|t| < 0.05), indicating that the attraction index has a significant effect on the actual price of the accommodation.

Overall, the results of the regression analysis suggest that the number of bedrooms, the distance to the nearest metro station, and the attraction index are significant predictors of the actual price of the accommodation, while cleanliness rating does not have a significant effect.

Insights

1. By sorting and grouping the room types by city we can see the distribution of the types of room in each location. This can provide an insight into the most popular types of accommodation in each city and how they differ amongst each location.
2. A basic table output shows the proportion of superhosts in each letting. This can help to identify if there is an influence in using lets from superhosts or if it is an important factor in guest satisfaction and also how it can affect price.
3. By carrying out linear regressions we can explore the relationship between multiple variables and see how they affect the pricing of the lets. The output from the model can be used to make data-driven decisions on price.
4. A plot and regression was used to visualise the relationship between the metro distance and price. This can be useful in identifying which room types are most popular in areas with better access to public transport.
5. The two maps show the distribution of the lettings and can help to identify how location affects pricing and whether it is an important factor for guests when choosing their rental.

References

"The Devastator", *Airbnb prices in European cities*, *Kaggle*. https://zenodo.org/record/4446043#.ZBW_n3bP1D9. Available at: https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities (Accessed: March 2023).