Project Report

GitHub URL

https://github.com/ConorSaund/UCDPA_conorsaunders

Abstract

In this project, the analysis was of pricing of Airbnb listings in ten European capital cities. The dataset consisted of twenty files, including weekday and weekend data for each city. The aim was 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. We all know the reasons that causes rental spaces to increase their pricing, it is interesting to try to prove it with certain factors.

Dataset

The dataset used was of Airbnb pricing in ten European capital cities. The attributes used in the data sets were room types, cleanliness, distance from the city centre to name a few. The data sets consist of weekday data and weekend data for the ten cities, leading to twenty files to work with. 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 some of the most popular 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 was using their online notebook for coding. It is the default code used to pull datasets straight from Kaggle. The datasets used throughout the assignment were then input using an integration where it 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, we need 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 quick 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 "roo_type", sorting of the room types by count in a descending order, and formatting of the graph. Capitalisation of the city names, along with adding axis labels, and a city by count legend on the right handside of the graph.

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 Suerhosts and not. 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. Based on this we went further and removed rows that contained missing cells and did not include them in any calculations. 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.

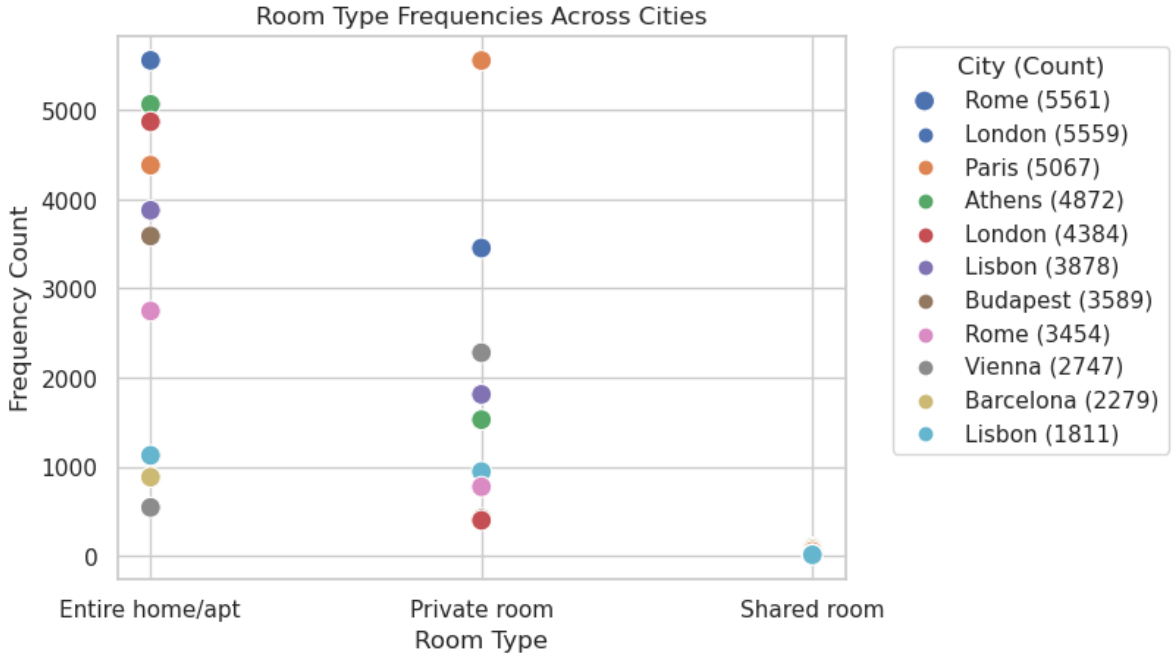
Another linear regression analysis was carried out on the distance from metros, cleanliness rating, number of rooms and the attraction index. The code was to first

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, alinear 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 locations of each Airbnb location as was given in the dataset by longitude and latitude of each. 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 would lead to a higher price on average. 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.

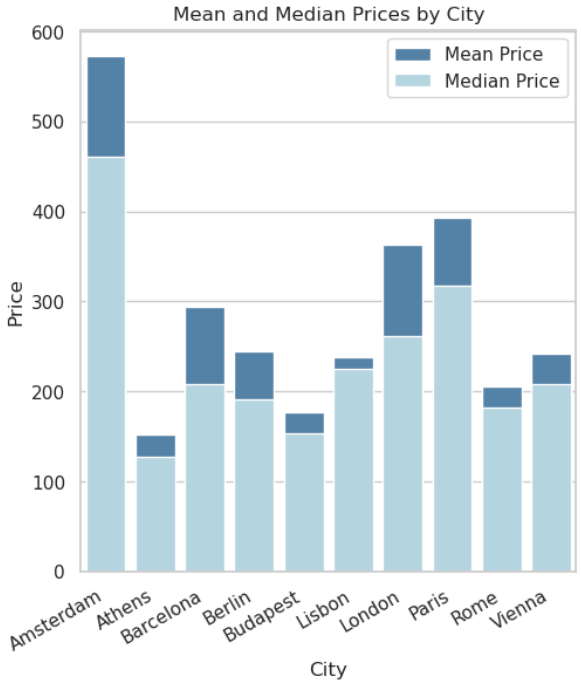
Results

(Include the charts and describe them)



(figure 1)

Figure 1 shows the counts of room types amongst the ten European cities. It is clear that the most common room type was an "Entire home/ apt", with " Private room" slightly less and finally "Shared room" with the lowest count.



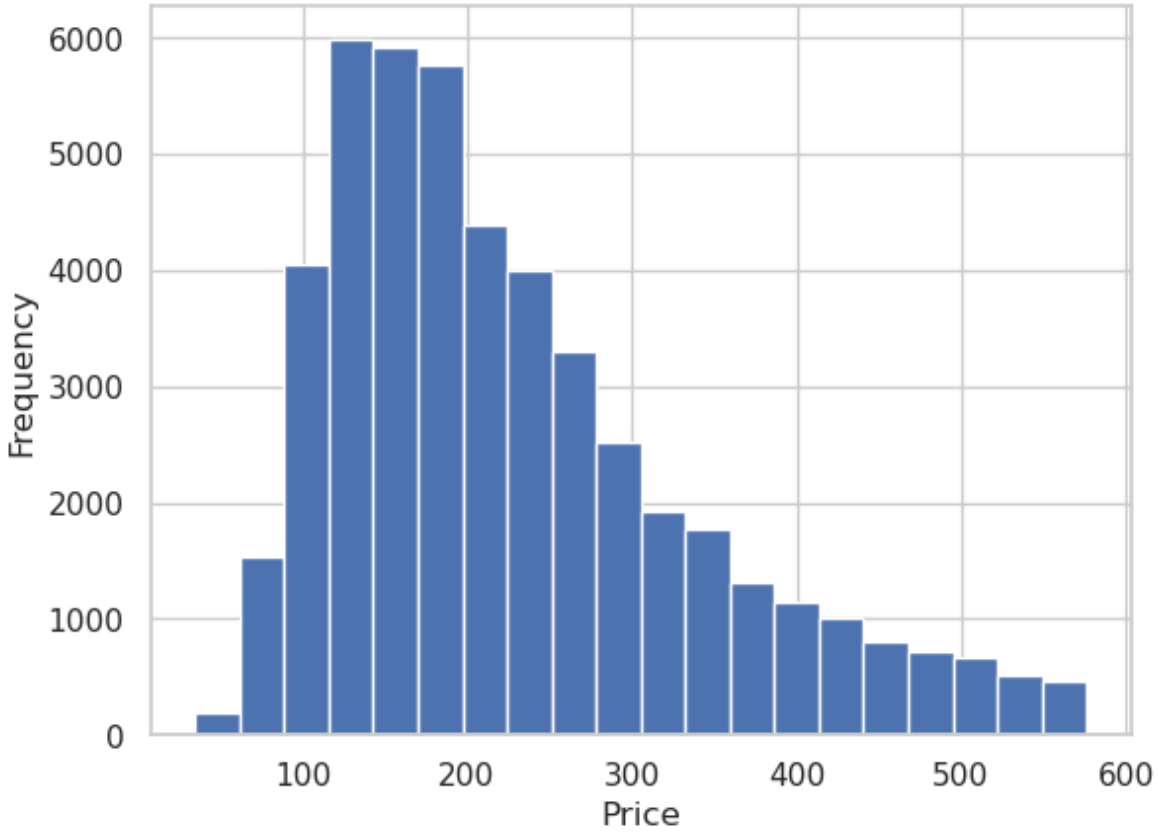
City	Mean Price	Median Price
Amsterdam	573.11	460.24
Athens	151.74	127.72
Barcelona	293.75	208.3
Berlin	244.58	191.18
Budapest	176.51	152.98
Lisbon	238.21	225.38
London	362.47	261.29
Paris	392.53	317.6
Rome	205.39	182.59
Vienna	241.58	208.49

(figure 2)

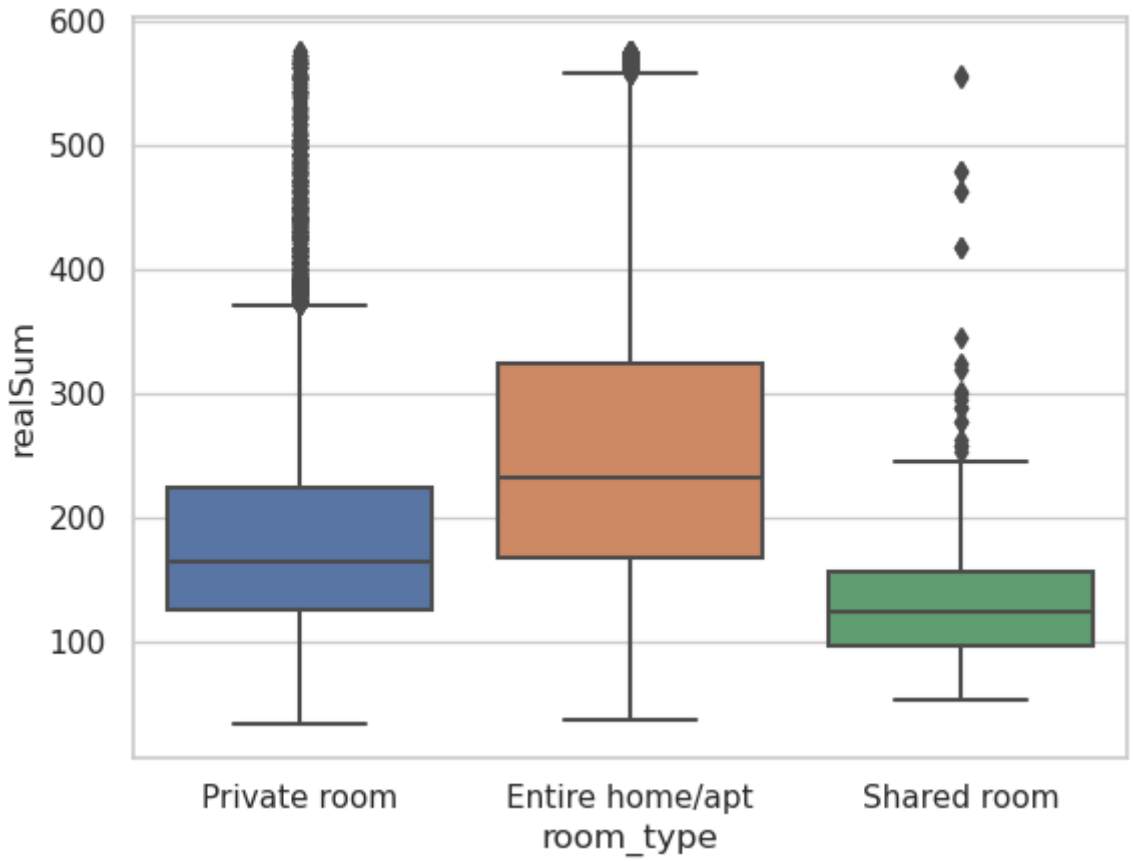
Figure 2 shows the mean and median prices of the European cities. From the graph and table we can see that Paris has the highest mean price at 392.53, which is higher than most of the other European cities. Athens has the lowest mean price at 151.74, which is significantly lower than the other cities. This suggests that Athens could be a relatively affordable city to visit or rent in. The median rice for all cities is greater than the mean 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 is suggesting 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 accommodation which is 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)


6000
5000
4000
3000
2000
1000
0
Frequency
100
200
300
400
500
600
Price

(figure 4)

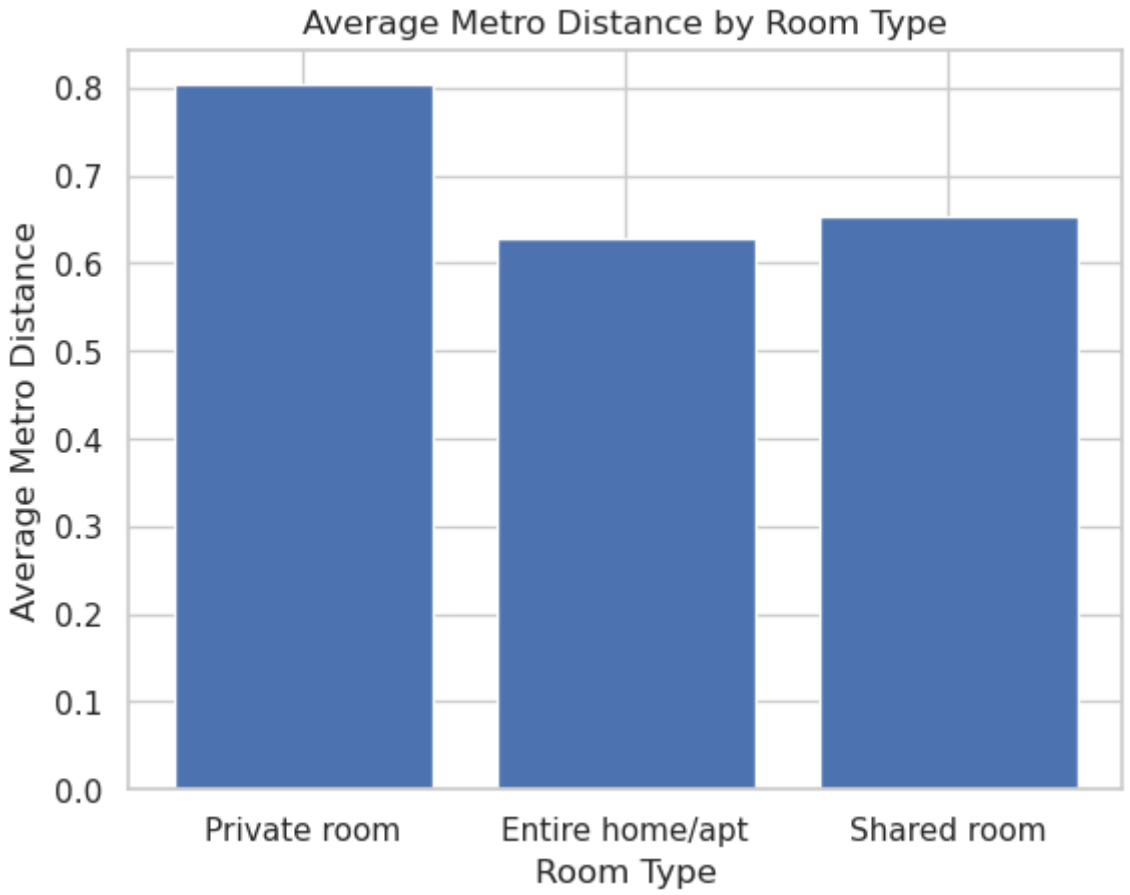


(figure 5)

```
                            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)


Average Metro Distance by Room Type
Average Metro Distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Private room
Entire home/apt
Shared room
Room Type

(figure 7)

```
                            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 8)

Insights

(Point out at least 5 insights in bullet points)

References

(Include any references if required)