DATA PIRATES

BIG DATA PROJECT DOCUMENTATION

Airbnb Prices in European Cities

PREPARED FOR:

Dr. Kris Manohar.

PREPARED BY:

Jade Ganga
Jessica Mohammed
Paul Chanka
Avin Heeralal

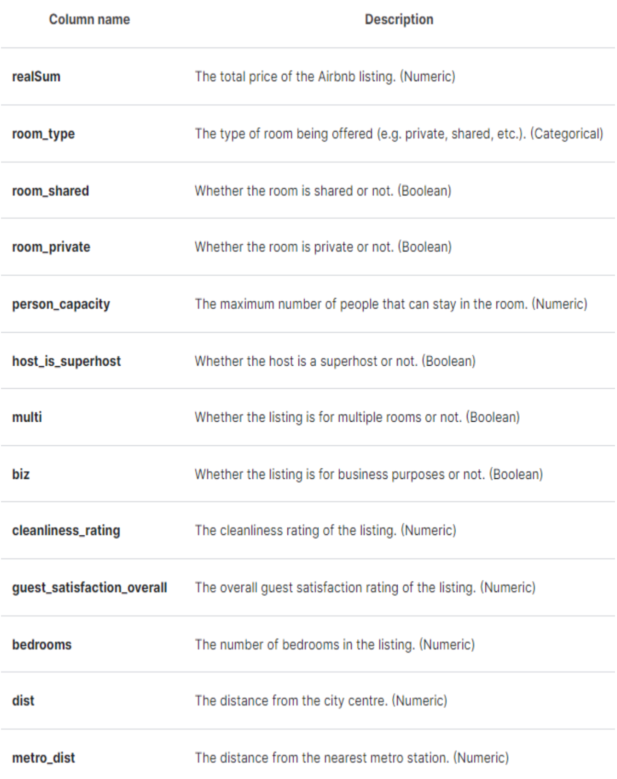
Abstract:* *The following document would delve into the processes of how the "Airbnb prices in European Cities" dataset was ingested and investigated, to provide predictive outcomes based on the features within the dataset using Jupyter Notebook and various python technologies. The predictions made by the investigation would aid new Airbnb hosts when adding new listings on the Airbnb site as well as clients looking for a satisfactory location to stay.

Introduction:

The ever increasing age of digitalization has no limits to the types of services which can be found online. With this vision a platform known as Airbnb was created in 2008 by Joe Gebbia, Nathan Blecharczyk and Brian Chesky.[1] The platform allowed regular individuals to provide temporary rentals of their personal property such as a room, apartment or entire house for travelers or locals in need of a place to stay. Once a free account is created on Airbnb.com individuals would have access to advertising their available place for stay, together with their rates and relative information about the property such as room type, contact information, pricing etc. From this same site tenants can make their bookings and get into contact with a host of their choosing.

Over the years Airbnb has seen tremendous growth, some recent reports from Airbnb claims that there exists over six million listings on their site spanning across two hundred and twenty countries around the world. It is also noted that approximately 95% of Airbnb properties boast of an average 4.5 star rating.[2] However, most other hotels' ratings are much lower. According to reports, an average of 2 million people rest their heads in an Airbnb property each night with this number growing exponentially.[3] The cities which showed the largest number of listings on the platform were London, Paris and New York.

Although there has been a great demand for Airbnb service it would be beneficial to know what makes these listings successful and satisfactory to the clients. Firstly to do this a relative dataset containing the necessary information to be investigated must be supplied. From Kaggle, an online resource, a dataset was chosen "Airbnb Prices in European Cities". The following is a brief overview of the columns found within the Airbnb dataset.[4]



Column name	Description
realSum	The total price of the Airbnb listing. (Numeric)
room_type	The type of room being offered (e.g. private, shared, etc.). (Categorical)
room_shared	Whether the room is shared or not. (Boolean)
room_private	Whether the room is private or not. (Boolean)
person_capacity	The maximum number of people that can stay in the room. (Numeric)
host_is_superhost	Whether the host is a superhost or not. (Boolean)
multi	Whether the listing is for multiple rooms or not. (Boolean)
biz	Whether the listing is for business purposes or not. (Boolean)
cleanliness_rating	The cleanliness rating of the listing. (Numeric)
guest_satisfaction_overall	The overall guest satisfaction rating of the listing. (Numeric)
bedrooms	The number of bedrooms in the listing. (Numeric)
dist	The distance from the city centre. (Numeric)
metro_dist	The distance from the nearest metro station. (Numeric)

Figure 1.0 - Features of the Dataset

Using the given columns above it can be assumed that the 'guest_satisfaction_overall', may be directly linked to the other factors such as 'cleanliness_rating' or 'dist'. From this assumption a predictive analysis can be completed on the dataset.

This analysis would be most useful since it can help airbnb hosts determine which factors actually influence the guest satisfaction of the listing.

Related Works:

From the online resource Kaggle it was an ease to find similar work done on airbnb data. The following are some works related to airbnb datasets:

- Airbnb Analysis, Visualization and Prediction[5] done by Chirag Samal based on New York City Airbnb Open Data. This notebook followed basic data analytic steps of data cleaning, data visualization and Regression analysis where the models used were Linear Regression, Decision Tree Regression and Random Forest Regression.
- Airbnb Prices in European Cities by Istiyaque Ahmed this was done using the same dataset chosen for the project. The varying csv files were merged into one and various investigations and visualizations were done.[6]
- Airbnb Prices in European Cities by Abdullah Babar, this notebook went into detail of the experiments at each step from data cleaning to the regression models used.[7] This investigation used the same csv and went into feature selection of the Airbnb listings to generate a predictive model based on prices to derive a link between features and pricing of Airbnb Models.

Another similar project was found on a github repository by Deepak Karkala which included an interface being built for inputting features such as to predict the target price. [8]

Since there was a range of related work available online various assumptions could have been made. This streamlined the process of deriving a solution and applying workable models.

Proposed Solution:

The proposed solution is to determine relationships between features by calculating the correlation score between the various features of the cleaned dataset. The features with the strongest relationship would then be used to build a predictive model. The model would be built such that it uses the features with the strongest correlation to "guest_satisfaction_overall" to make guest satisfaction rating predictions. A correlation matrix was used to determine the relationships.

Based on the dataset, it was decided that linear regression and random forest regression algorithms would be used to perform predictions of guest satisfaction rating from the cleanliness rating.

The correlation matrix was chosen because it is a powerful tool that determines the correlation coefficient between features of a dataset. These correlation coefficient values can then be used to determine which features of the dataset have the strongest relationship. Correlation coefficients closer to 1 indicate strong relationships and those that are 0 or less tend to be weaker relationships.

Linear regression is an analysis tool used to predict values of a feature based on another feature.[10] The linear regression model would build upon the results of the correlation matrix, using the features with the strongest relationships. In this project, the dependent variable was "guest_satisfaction_overall" and the independent variable was "cleanliness rating."

Random forest is a powerful classification algorithm that can be used for predictions.[11] It uses a number of decision trees to obtain more accurate predictions.

The performance of these algorithms was monitored by calculating and comparing the models' f1 score, accuracy, precision, recall and mean squared error values.

Discussion:

The "Airbnb Prices in European Cities" Dataset from Kaggle.com consisted of two datasets for 10 different cities in Europe. One dataset was for weekends while the other contained data from weekdays. All 20 datasets were   combined[9] and then cleaned.

The cleaning process consisted mainly of checking for missing values and dropping columns that were not needed, such as longitude and latitude. Graphs were plotted for cleanliness rating against guest satisfaction as well as guest satisfaction against price (realSum)

The scatter plot for realSum vs guest satisfaction shows that listings with a high guest satisfaction rating tend to have high realSum values. The graph illustrates a relationship between these features whereby the higher a listing's guest satisfaction rating, the higher its realSum value. Later on in the investigations, it was seen that the relationship between realSum and guest satisfaction in the correlation matrix was on the lower side of the scale with -0.0019.

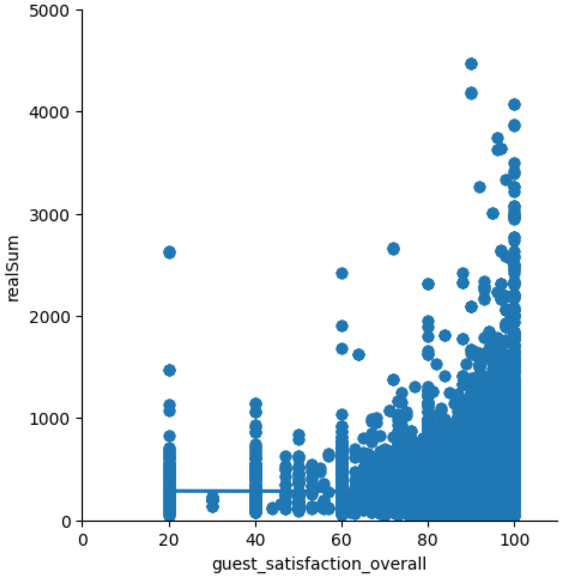


Figure 2.0 - Scatter Plot of "guest_satisfaction" vs "realSum".

The cleanliness rating versus guest satisfaction scatter plot reflects a linear relationship between the features. It shows that as cleanliness rating increases, the guest satisfaction rating also increases. The plot shows that higher cleanliness ratings tend to relate to higher guest satisfaction ratings. Upon further investigation of the relationship between these features, using the correlation matrix, it was proven that the highest correlation existed between ‘guest_satisfaction_overall’ and ‘cleanliness_rating’ with a value of 0.71.

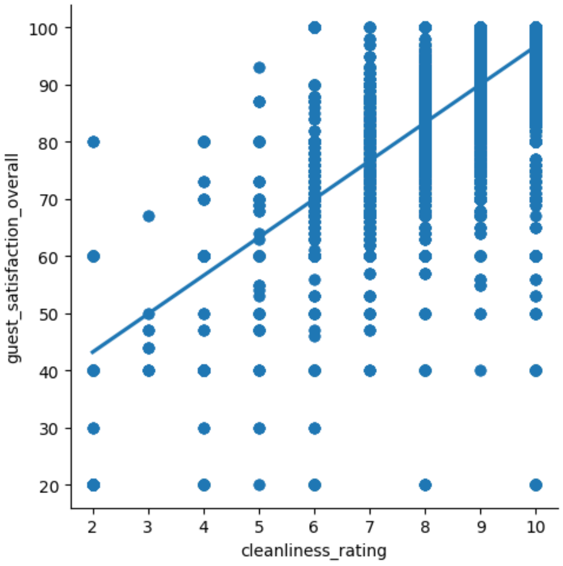


Figure 3.0 - Scatter Plot of “guest_satisfaction” vs “cleanliness_rate”.

The correlation matrix results indicate that the working dataset has very few correlation. There were 2 strong correlations including the relation between guest satisfaction rating overall and cleanliness rating as well as the relationship between metro distance and distance from a city center. There is one more correlation although it is not as strong as the other two mention it is strong on its own. This correlation is the bedrooms and persons capacity.

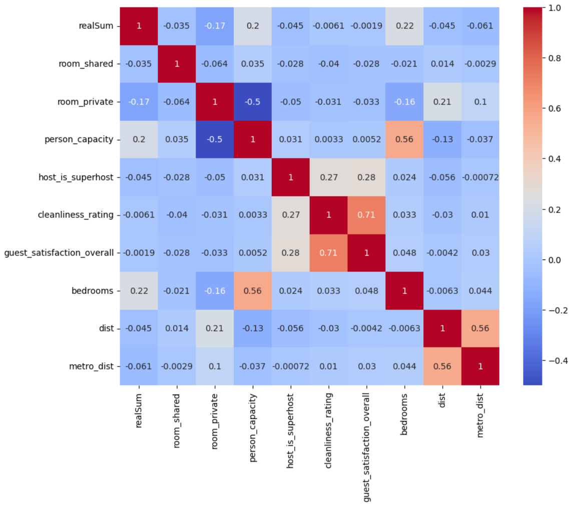


Figure 4.0 - Correlation Matrix of Dataset Features

The linear regression model was given a cleanliness rating of 5 and predicted a guest satisfaction rating of 63. The mean squared error, f1 score, accuracy, precision and recall values were calculated as metrics to analyze the performance of the model. The model had a mean squared error of 38.947.

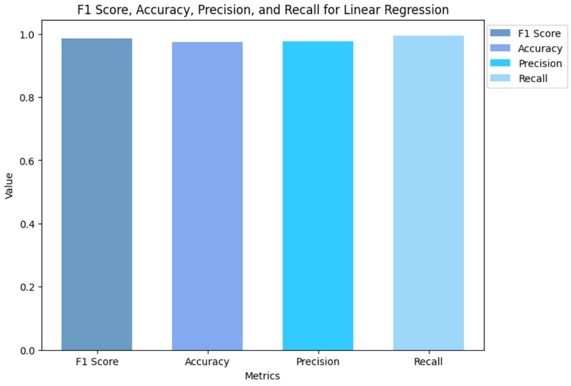


Figure 5.0 - Bar Chart Showing Performance Metrics for Linear Regression Model

The results of the linear regression model were used to plot a graph of predicted values versus actual values from the dataset. This graph showed that the predicted values plotted a straight line reflecting the model’s high accuracy, f1 score, recall and precision values.

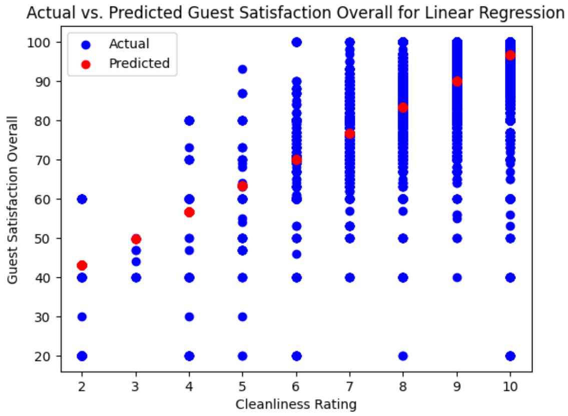


Figure 5.1 - Scatter Plot of Actual vs. Predicted "guest_satisfaction_overall" for Linear Regression Model

In order to validate the linear regression model a sensitivity analysis was conducted on the model. This was used to stress test the model by implementing a variety of values into the model to monitor how the system's output changes in response to the changes in the cleanliness rating. The perturbations of the sensitivity analysis varied within -0.1 and +0.1 with the increment of each perturbation the output of the sensitivity was displayed. The sensitivity analysis proved that the predictive model was strong and held a linear relationship since the sensitivity at perturbation -0.1 was -0.05291119 and for perturbation +0.1 was +0.05291119.

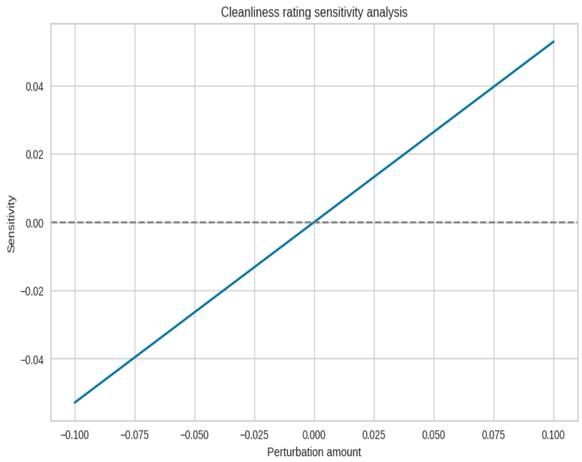


Figure 5.2 - Graph Showing Sensitivity Analysis for "cleanliness_rating" for Linear Regression Model

Also giving the random forest regressor model a value of 5, returned a value of 61. This model produced a mean squared error of 38.004. The metrics that were calculated for this model were also f1 score, accuracy, precision and recall.

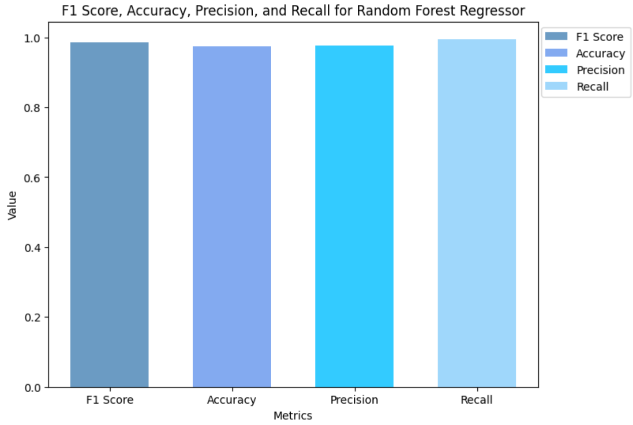


Figure 6.0 - Bar Chart Showing Performance Metrics for Random Forest Regressor Model

The random forest regressor had a smaller mean squared error than the linear regression model which means that it made more accurate predictions. This means that the difference between the random forest model's predicted values and actual values is smaller than those of the linear regression model.

The results of the random forest regressor were used to plot a graph of actual versus predicted values. This resulted in an almost straight line of predicted values which reflects the performance of the model.

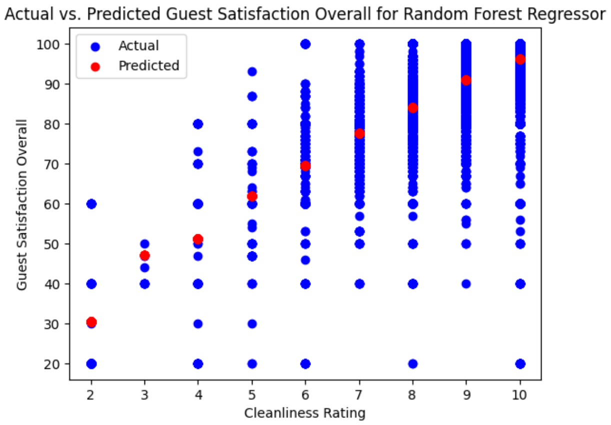


Figure 6.1 - Scatter Plot of Actual vs. Predicted "guest_satisfaction_overall" for Random Forest Regressor Model

Both models had the same accuracy and f1 scores which means that they were making the same number of correct and incorrect predictions.

On the other hand the models had slightly different mean squared error values which indicates that their magnitude of error varied.

Overall both models performed significantly well and had almost the same performance metrics.

Future Works:

To better improve on the models done within this investigation certain criteria when combining the various datasets could be implemented such as adding a "City" column to keep track of various Airbnb features throughout the listings within cities in Europe. By adding these steps the models can be made more specific and prevent overfitting of the predictive models.

K-fold cross validation can also be implemented in the case that there is overfitting. It can be used to estimate the skill and ability of the model using new data as well. Also introducing external data sources that are separate from the data in the Airbnb dataset (such as events in the area and economic indicators, property amenities etc.) can provide additional insight into factors that influence guest satisfaction.

Introducing data from other regions where there may be differences in pricing and guest satisfaction, can also impact these predictions that were made.

Additionally, other models such as Support Vector Machines or Neural Networks could be explored to find more accurate predictions.

Conclusion:

To conclude, the investigations conducted were able to accurately predict the "guest_satisfaction_overall" based on the "cleanliness_rating". Since they were able to prove an exponential growth of guest satisfaction values when the cleanliness rating was increased this showed that there existed a linear relationship between both features.

Unfortunately, no other features were worth investigating for predictions. Since the correlation metric between other features were proven to be significantly below a suitable value.

The predicted values of guest satisfaction based on cleanliness rating can help hosts to better understand what might make their listings more satisfactory to clients. It can also help to ensure that their listings keep a good rating on Airbnb.com.

Reference List

[1]"What is Airbnb and how does it work? - Airbnb Help Center," *Airbnb*. https://www.airbnb.com/help/article/2503

[2]G. Zervas, D. Proserpio, and J. W. Byers, "A first look at online reputation on Airbnb, where every stay is above average," *Marketing Letters*, vol. 32, no. 1–16, Nov. 2020, doi: https://doi.org/10.1007/s11002-020-09546-4.

[3]H. Sherwood, "How Airbnb took over the world," *the Guardian*, May 05, 2019. https://www.theguardian.com/technology/2019/may/05/airbnb-homelessness-renting-housing-accommodation-social-policy-cities-travel-leisure

[4]"Airbnb Prices in European Cities," *www.kaggle.com*. https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities?resource=download.

[5]"Airbnb Analysis, Visualization and Prediction," *kaggle.com*. https://www.kaggle.com/code/chirag9073/airbnb-analysis-visualization-and-prediction

[6]"Airbnb Prices in European Cities," *kaggle.com*. https://www.kaggle.com/code/istiyaque6ty3/airbnb-prices-in-european-cities.

[7]"Airbnb Prices in European Cities," *kaggle.com*. https://www.kaggle.com/code/ibabarx/airbnb-prices-in-european-cities.

[8]D. Karkala, "Airbnb-Data-Science," *GitHub*, Dec. 06, 2022. https://github.com/deepak-karkala/airbnb-data-science.

[9]T. Bugdani, "How to combine CSV files using Python? - AskPython," Dec. 29, 2022. https://www.askpython.com/python-modules/pandas/combine-csv-files-using-python

[10]"sklearn.linear_model.LinearRegression — scikit-learn 0.22 documentation," *Scikit-learn.org*, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[11]scikit-learn, "3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.20.3 documentation," *Scikit-learn.org*, 2018. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html