Statistical Analyses of Registered AirBnB Properties in New York City using Artificial Intelligence Models

Introduction

For this report we are analyzing a dataset containing information about all properties in New York City (NYC) that are registered with 'AirBnB', a business that acts as an online marketplace for individuals to both offer and book out lodging. We obtained this dataset from the online community of data scientists 'Kaggle', but the original source is from 'Inside AirBnB', an independent and non-commercial website that collates publicly available data about AirBnB listings.

The dataset contains 16 columns, please see the table below for a list of each column and the information it provides:

Column Number	Column Name	Column Description	Column Datatype
1	id	Unique identifier for each listing.	Discrete Int
2	name	Name of the listing.	String
3	host_id	Unique identifier for the host of the listing.	Discrete Int
4	host_name	Name of the host.	String
5	neighbourhood_group	NYC borough the listing is located in.	String
6	neighbourhood	Local neighbourhood the listing is located in.	String
7	latitude	The listing's latitude coordinate.	Continuous Float
8	longitude	The listing's longitude coordinate.	Continuous Float
9	room_type	Type of space the listing is.	String
10	price	Price of renting the space for one night in US dollars.	Discrete Int
11	minimum_nights	Minimum number of nights that must be booked.	Discrete Int
12	number_of_reviews	Number of reviews the listing has received.	Discrete Int
13	last_review	Date of the last review the listing received.	Date
14	reviews_per_month	Average number of reviews listing receives per month.	Continuous Float
15	calculated_host_listings_count	Number of listings host owns.	Discrete Int
16	availability_365	Number of days in the year listing is available for booking.	Discrete Int

As you can see, this is a comprehensive dataset which provides us with a wealth of information about every listing. This dataset contains 48,895 rows which means, at the time this dataset was compiled, there were roughly 48,895 listings registered with AirBnB in the NYC area.

We will be statistically analyzing this dataset using Artificial Intelligence (AI) methodologies in order to tackle various problem domains we have put forward.

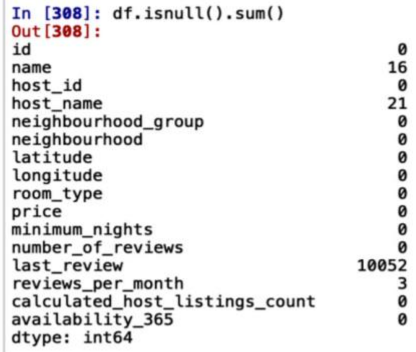
Section 2: Regression Models of Pricing

Methodology

Using a regression model that predicts the price for homes and rooms in New York, we want to see what the relation between the type of listing and if there is any difference in prices between the different areas of New York.

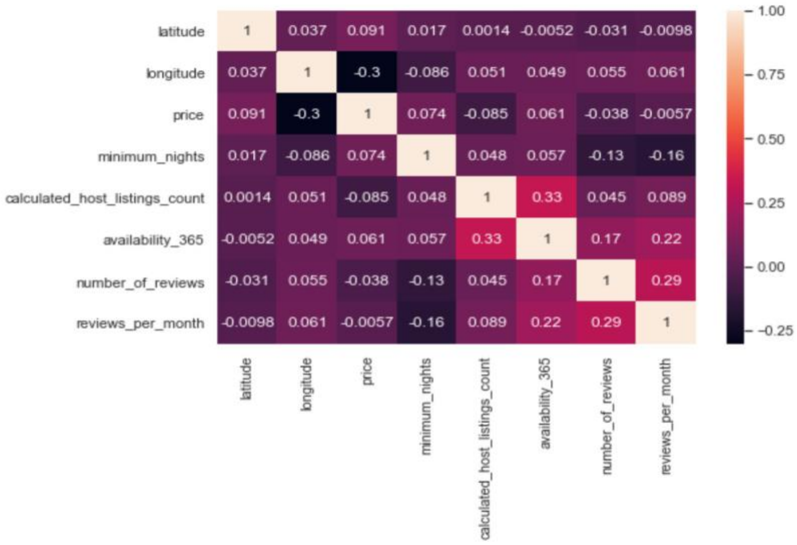
To produce a good model, we have to clear our data and prepare it for fitting.

Looking at the shape of our dataset we can see that we have 16 columns and 48895 rows. We can then check the types of the columns using “df.dtypes” to better understand our dataset. We then use the Pandas' method “.isnull().sum()” to check if there is any missing data. The results show that the columns: 'name', 'host_name', 'last_review' and 'reviews_per_month' have missing values. To cope with these missing values we use Pandas' “.filna” method and we change them into zeros. Moreover, we have to check if there is any categorical data, which has to be changed to dummy variables.



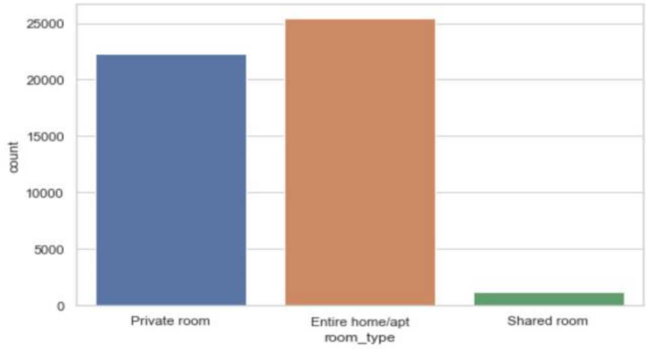
```
In [308]: df.isnull().sum()
Out[308]:
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                     3
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

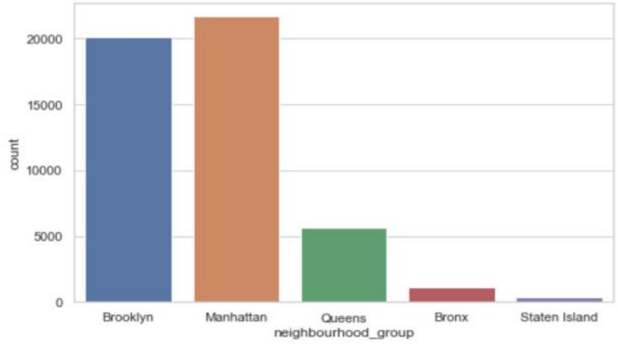
For a better understanding of the data, we decided to separate the columns with categorical and numerical data into two different lists, followed by dropping the columns that are not needed for the model. Since the column "last_review" has more than ten thousand missing values, we have dropped this column. Other columns that we have removed are: 'name', 'host_id', 'host_name', 'neighborhood' , 'id' since they all contain values of type "string”, which unlike numerical data cannot be converted to dummy variables due to the uniqueness of every value. After making these changes our dataset has 10 columns and 48,895 rows.



We then applied some plots to better understand our data. The first plot is a heatmap that shows the correlation between all the columns in the dataset. In this heatmap, we used the default colours to visualize the coefficients. We have dark purple to black for negative values and then purple to pale yellow for the positive ones. By this heatmap, we can see that our target "price" is positively related to the latitude but negatively related to the longitude. A positive correlation can be seen between price and the columns: "minimum_nights" and "availability_365", indicating that the price might get higher for a listing with more available nights. On the other hand, the correlation with the columns "calculated_host_listings_count", "number_of_reviews" and "reviews_per_month" is negative, which means that the price might go down depending on the reviews for the apartment or the room.

The count plot for the columns: "room_type" and "neighbourhood_group" shows us that both have a categorical value that can be converted to dummy variables. The "room_type" column will be replaced by three additional columns, each of which will indicate one category out of the



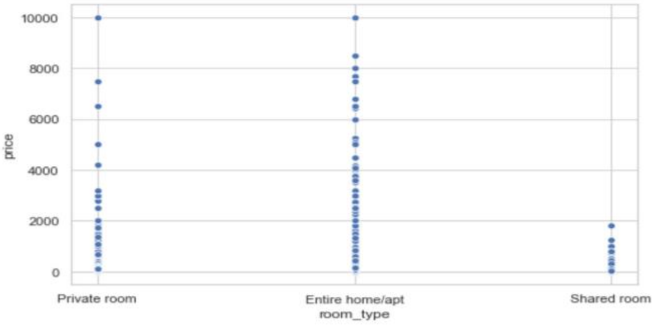


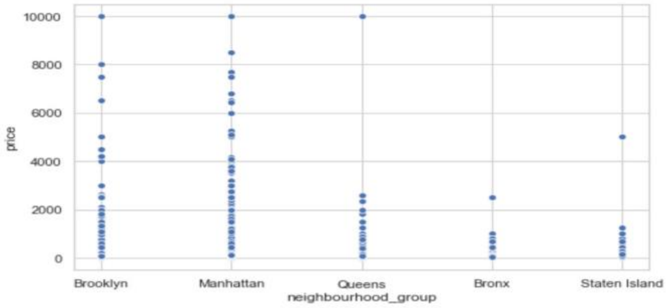
three the initial column contains: Private room, Entire home/apartment and Shared room. From this plot, we can observe that the entire homes/apartments and private rooms are dominating over the shared rooms in New York.

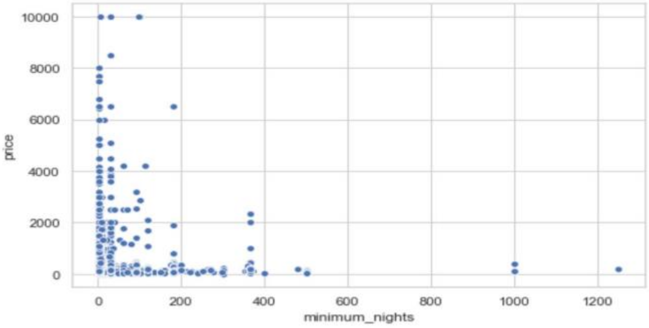
The new columns which are: room_type_Private_room, room_type_Entire_home_apr, room_type_Shared_room will have the values 0 and 1 – 0 indicating that the value is not this room type and 1 – indicating that the value is the type of room which the column specifies.

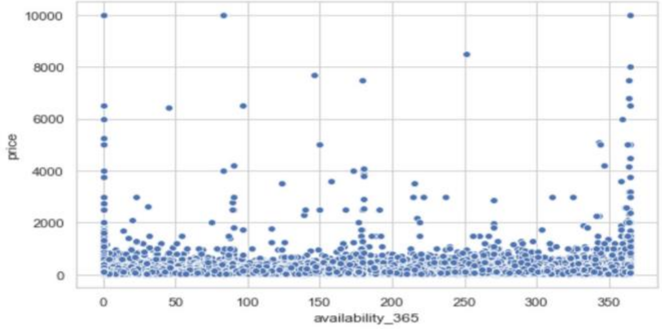
Similarly, the "neighbourhood_group" will be replaced by: "neighbourhood_group_Bronx", "neighbourhood_group_Brooklyn", "neighbourhood_group_Manhattan", "neighbourhood_group_Queens", "neighbourhood_group_Staten Island."

Next plots show the relation between price and number of different columns:





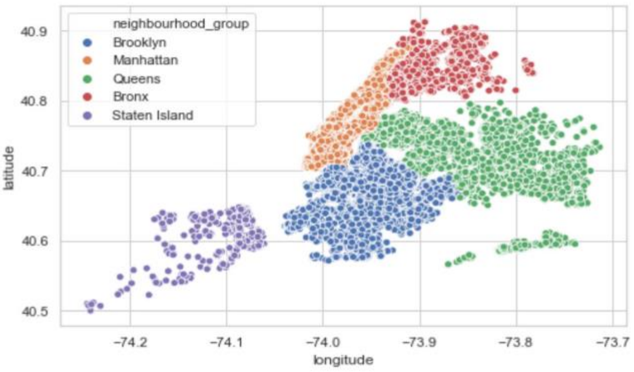


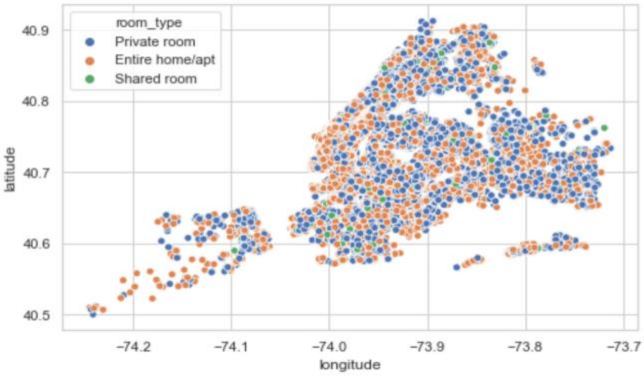


Observations that can be made following these plots are that the prices for entire homes and apartments is usually between “$0” and “$8000”, while for a shared room the maximum price is “$2000” and usually between “$0” and less than “$4000” for a private room. Higher prices can be expected in Manhattan and Brooklyn, while for Queens the range is “$0-2500” with one “listing costing “$10000”; and less than “$2000” for Bronx and Staten Islands.

We’ve produced some more interesting plots with the scatter graphs below, that let us analyze which type of room is more abundant in each NYC borough.

We can see that in Brooklyn we have a majority of shared rooms, in Manhattan between the 40.7 and 40.8 latitudes and around -74.0 longitude, we have more entire homes/apartments than private rooms and almost no shared rooms. The Bronx looks like it has similar numbers of private rooms and entire homes, while Queens tends to have a number of entire homes located at -73.8 longitude and 40.7-40.8 latitude and private rooms around this area. For Staten Island, the private homes and apartments are located more between -74.1 and -74.3 longitude and 40.5 – 40.6 latitude, while the private rooms are over 40.6 latitude.





The two models that we chose for this dataset are “Linear Regression” and the “Decision Tree”, the reasons are as follows:

Regarding “Linear Regression”, we have two variables – price and a set of predictor variables necessary for the model like the location of the Airbnbs. For these two variables, we can apply a linear equation to find a relation between the price and the other information. From the heatmap, we saw that there is a positive relation between price and other columns like latitude, minimum nights and availability_365, which is needed for determination before trying to fit the linear regression model. This model is good in predicting future events depending on trends, which in our case – we have to forecast the prices in particular areas, where the homes have rates decided by some tendency. Linear Regression is very good with linearly separable datasets, it is decent for sales, pricing, determining marketing effectiveness, etc. Our predicted outcome before carrying out the analysis, the “Linear Regression” model will have some good points which are near the line. However, one of the disadvantages of Liner Regression is that it would not be able to catch the points that are further away – it is sensitive to outliers. Another disadvantage is that the data should be independent, however, as seen in one of the plots, we have clustering of the entire homes/apartments in the center of Queens. In my opinion, the more luxurious homes and rooms will be predicted at a lower price than their actual price.

Decision Tree is the model that can capture better than Linear Regression if the model has not well-defined linear relation. In our case, although there is some linear relation, we have variables with negative correlation with the price. From the plots, we can see that our model cannot be separated precisely with a line. Decision Tree is good for handling datasets with many missing values or errors, which is also a good idea for our model. One of the major advantages for this model is that it does not require scaling or normalization of the data. On the other hand, the disadvantage that it can be insufficient in regression sometimes may cause a problem for us.

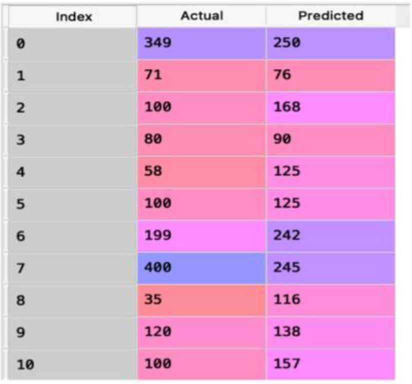
The expected outcome for Decision Tree model is to perform better analysis than the linear regression model for the reason which we previously mentioned – not perfect linear relation between the data in the set.

Results (Regression model)

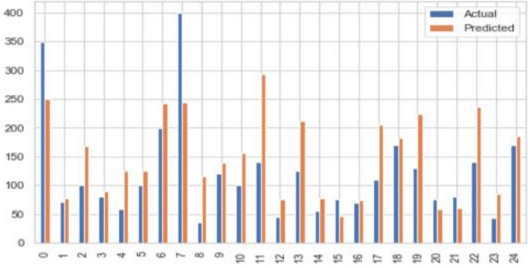
After splitting the dataset to 25% test set and 75% training set and fitting it with the Linear Regression model, we have the following results:

Mean of Linear Regressor: 153.71662303664922

Root Mean Squared Error of Linear Regressor: 231.96310275628431



Index	Actual	Predicted
0	349	250
1	71	76
2	100	168
3	80	90
4	58	125
5	100	125
6	199	242
7	400	245
8	35	116
9	120	138
10	100	157

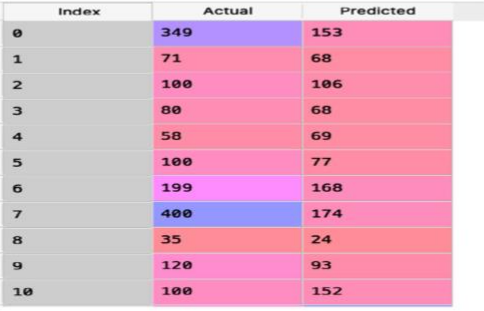


As can be seen, there are some good predictions but much noise as well.

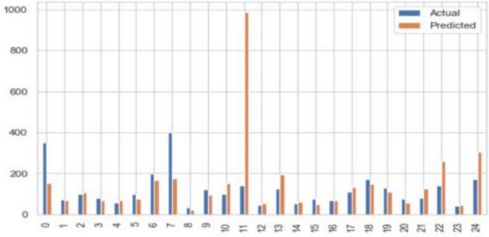
For the Decision Tree Regressor we have the following outcome:

Mean of Decision Tree Regressor: 153.71662303664922

Root Mean Squared Error of Decision Tree Regressor: 246.57515345624944



Index	Actual	Predicted
0	349	153
1	71	68
2	100	106
3	80	68
4	58	69
5	100	77
6	199	168
7	400	174
8	35	24
9	120	93
10	100	152



Evaluation

The problems we had with the dataset and the regression models are: the relationship between the variables was not that precise which made our model not good enough for accurate predictions; the missing data in the "last review" - about one-third of the values in this column, had to be removed otherwise the outcome may have been disturbed by the zeros which replaced the missing data.

Conclusion

To summarize, The Linear Regression has lower Root Mean Squared Error, which means it has better results than the Decision Tree Regressor, which is not what we were expecting. Although it is the better model, the dataset is not good enough for exact predictions.