Airbnb Price Prediction
New York City 2019
MBA 540 Data Mining
Final Report

Professor: Dr. Keli Xiao
Group Member: Yifan Dai, Zhe Zhou, Loreto VillalbaRubio

Airbnb Price Prediction

•Abstract:

The major goal of this project is to provide an analysis of Airbnb, specifically at New York City Metro area in 2019. We want to mine the data and uncover interesting observation about the different hosts and areas. Examine the data, analyze outliers using data exploration techniques; boxplot, scatter plot, pie chart, classification and regression. In return, we want to discover the price variation based on different factors. As a result, Manhattan is the most expensive booking, Bedford-Stuyvesant is the busiest area and Brooklyn has the highest number of reviews.

•Introduction: Discuss the following items in the introduction:

○The report explores the New York City Airbnb Open Data and uncover the true data within the datasets. Given the dataset Airbnb NYC 2019, our assignment or goal is to predict the price variations by create multiple analysis using different techniques learned throughout the semester. The importance of our analysis in this report are as follows:

- Examine and clean the dataset to find out any anomalies in the data.
- Examine how the prices of the Airbnb changes within different neighborhood in the New York City.

- Take top 5 and bottom 5 of the price of the Airbnb, analyze and plot the trend.

- Correlation analysis using attributes provided in the dataset and find the most positive and most negative correlations.

- Analysis longitude and Latitude using scatter plot.

- Find if there's relations between busiest area V.S. most expensive area.

- Classification of Airbnb price prediction and Regression of Airbnb price prediction.

•Methodology:

During this report, we have used data exploration by have tested out the "price" and "availability_365" attributes in the dataset, we removed the data that had either of these as part of cleaning the data. First, we used boxplot, scatter plot to find the price variation among the neighborhood in NYC. Then during the group meeting, we received feedback from the professor that we need to use multiple techniques to determine the data accuracy. Therefore we added classification and regression of the Airbnb price prediction using python.

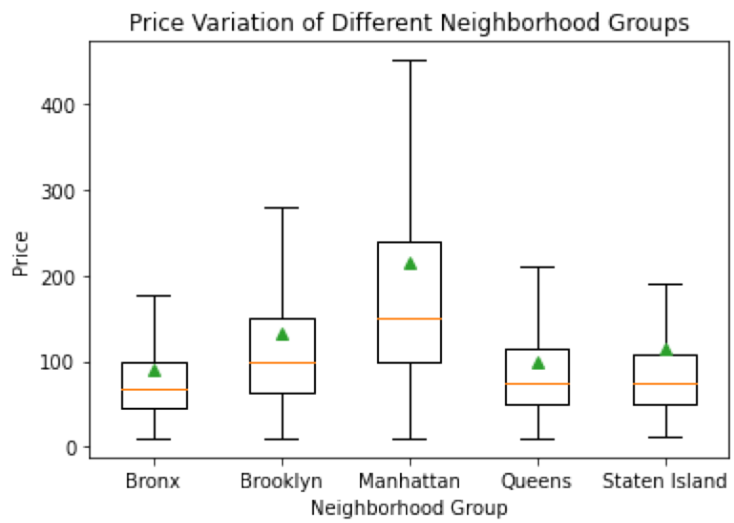
We also did a correlation analysis to determine the most positive and most negative correlations between Airbnb bookings. And we did a scatter plot to analyze the longitude and latitude of various Airbnb location. We excluded any price that is greater or equal to $1000 to avoid luxury booking.

Data and Experimental settings:

Data Exploration: We have tested out the “price” and “availability_365” attributes in the dataset, we removed the data that had either of these as part of cleaning the data. We grouped the dataset by “Neighbourhood” and removed any group that size is less than 5 then filtered down to median price in each group. In this part, we used boxplot method and we eliminated any luxury Airbnb booking data as they may affect the outcome as outliners.

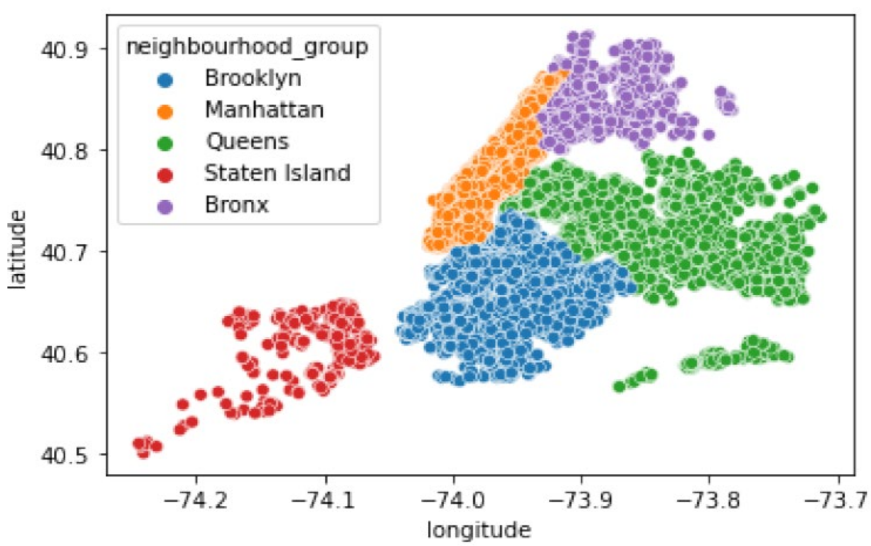
neighbourhood	price
Concord	34.5
Castle Hill	39.0
Hunts Point	40.0
Corona	40.0
Tremont	41.0

neighbourhood	price
Tribeca	309.0
Flatiron District	299.0
NoHo	250.0
Midtown	225.0
West Village	218.0



Correlation Analysis: To determine the correlation we created a table using "Price", "Minimum_nights", "Number of reviews", "Reviews per month", "Host listings count" and "Availability_365". To find the most positive and most negative correlations.

Longitude & Latitude Analysis (Scatter Plot): We made X-axis represents longitude and Y-axis represents Latitude and using a scatter plot. First



we grouped by neighbourhood, then we removed any price that is larger or equal to $1000 as they represent luxury booking. At the end, the second scatter plot will show where the most expensive booking among all neighborhood groups. We also want to consider whether the busiest area will have impact or relation on neighborhood pricing.

neighbourhood	count	price	minimum_nights	number_of_reviews	availability_365
Bedford-Stuyvesant	2478	115.354722	6.350686	39.598870	174.545198
Williamsburg	2051	161.171136	7.753291	35.203315	142.779132
Harlem	1734	129.643022	7.662053	38.734141	163.369666
Bushwick	1447	91.409122	6.852799	31.941949	162.447132
Hell's Kitchen	1446	213.183264	9.569848	31.527663	188.009682

Classification: In this part, we decide to build a random forest regression model to do the classification of Airbnb Price Prediction

Package and Function used in Python:

- "DecisionTreeClassifier( )" in **sklearn.tree**

Variables used to fit the regression model (selected from previous analysis):

- "price"
- "latitude"
- "longitude"
- "room_type"
- "reviews_per_month"
- "calculated_host_listings_count"
- "availability_365"

Since some variables are categorical, we need to convert them into numeric values before fitting the random forest regression model.

For "price", according to mean value in "price" variable, we decide to separate it into 3 types:

- low (0): less than $100
- medium (1): between $100 and $200
- high (2): above $200

Airbnb Price Prediction

In the random forest model, we don't need to consider if a categorical variable is nominal or ordinal. However, the function require numeric input, so we need to use numeric values to represent each type in categorical variables. For "room_type", use 0 to represent shared room, 1 to represent private room, and 2 to represent entire home/apt.

Regression: Package and Function used in Python:

- "LinearRegression( )" in sklearn.linear_model

Variables used to fit the linear regression model (selected from previous analysis):

- "price"
- "neighbourhood_group"
- "room_type"
- "minimum_nights"
- "number_of_reviews"
- "reviews_per_month"
- "calculated_host_listings_count"
- "availability_365"

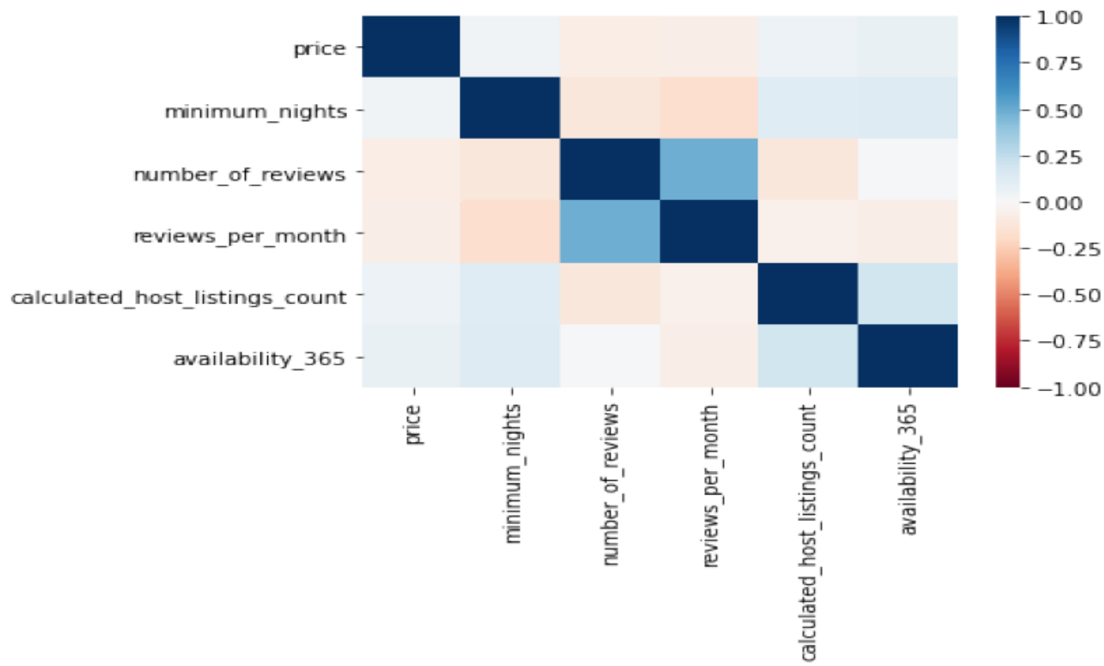
Since some variables are categorical, we need to convert them into numeric values before fitting the linear regression model.

For "neighbourhood_group", since it's a nominal variable, so it cannot be converted to numeric values directly. We replace it by four variables "Brooklyn", "Manhattan", "Queens", and "Bronx", and let 1 represents "Yes", 0 represents "No". Note that, all 0 represents "Staten Island".

For "room_type", we consider it as a ordinal variable, so we can convert it to numeric values directly. We use 0 to represent shared room, 1 to represent private room, and 2 to represent entire home/apt.

•**Data and Experimental settings:**

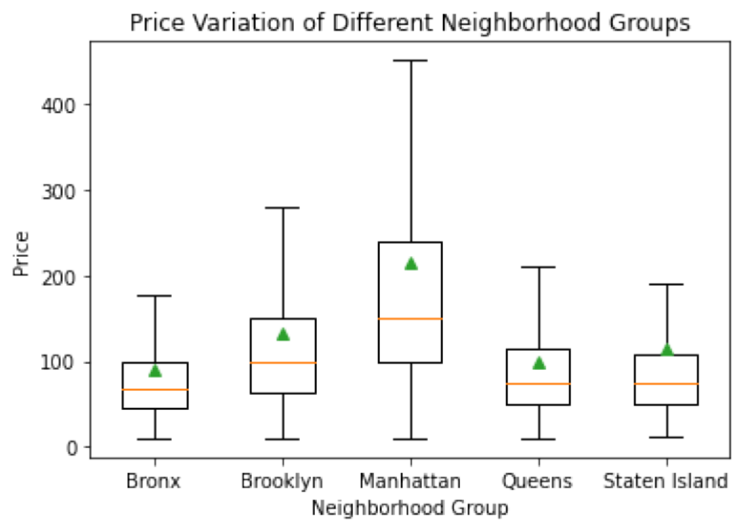
•**Results and analysis:**



Darker color implies the absolute value of the correlation coefficient is closer to 1. Therefore, according to the heatmap, the "number of reviews" and "reviews per month" have the most positive correlations and "reviews per month" and "minimum nights" have the most negative correlations." We only need to focus on the result of the first row since we only want to figure out the relationship between the count and four variables. From the heat map, we find that the relationship between the number of listings and price, minimum nights, and the number of reviews is positive, but the relationship between the number of listings and the number of days when the listing is available for booking is negative.

If the price of an area is higher, more hosts want to make listings here. Also, more reviews imply a higher number of customers, and hosts are willing to make listings in a place that has a lot of customers. And the minimum number of nights can demonstrate the requirement to Airbnb in the area. Hence, if the number of minimum nights increases, hosts may think there is a chance to earn money in the area. Similarly, the number of days when a listing is available for booking also represents the requirement to Airbnb, on the contrary, the smaller it is the more people live in this area. This is the reason why the coefficient is negative between the number of listings and the number of days when the listing is available for booking.

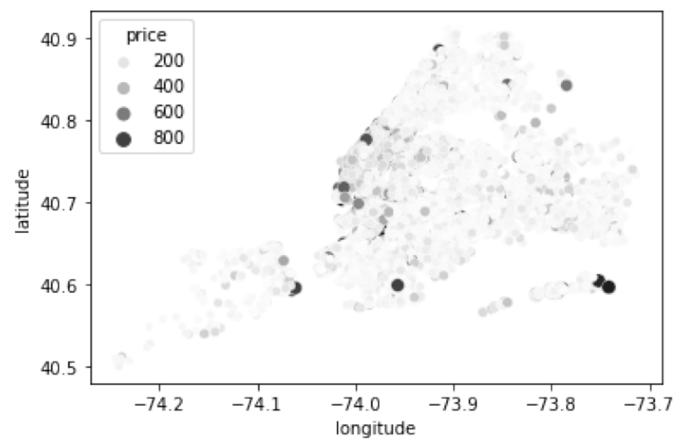
Airbnb Price Prediction

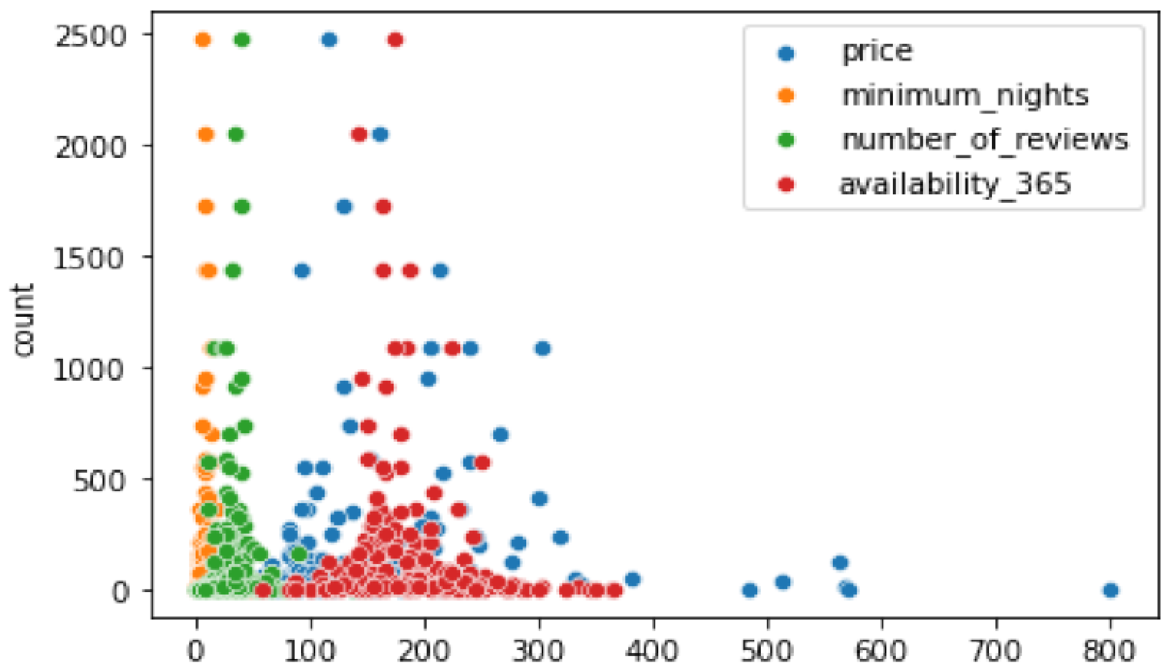


According to the chart, Manhattan has the most expensive Airbnb among all neighborhood groups. (also highest median and mean price.) The Airbnb price and price variation will increase in Manhattan.

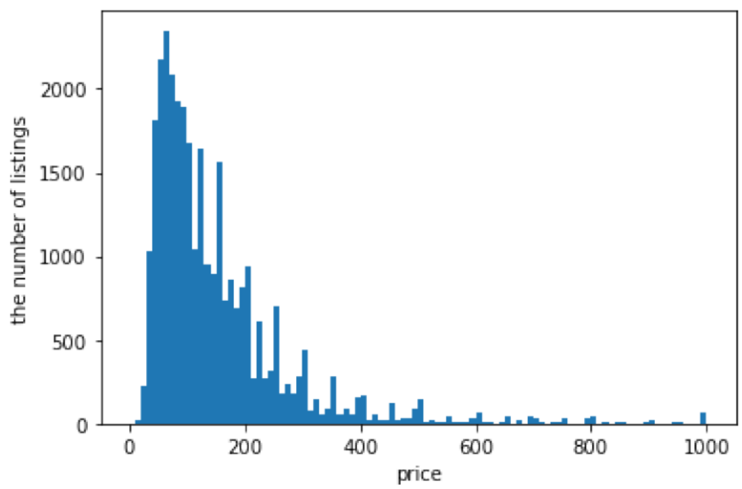
The higher the price, darker the point will be.

And as per the scatter plot, Manhattan is once again have the most expensive booking among all neighborhood groups.

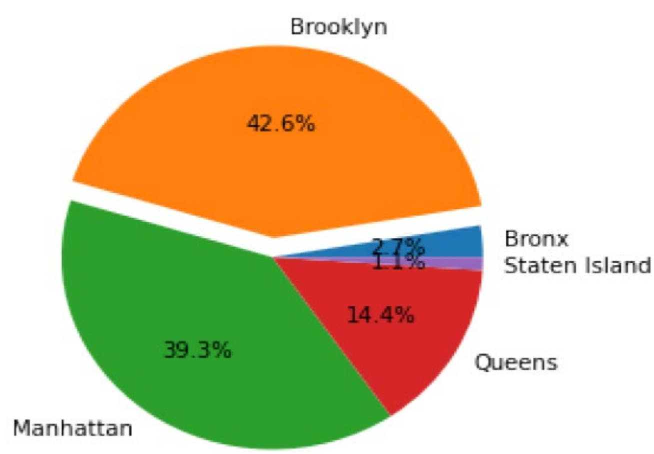




We found that there's no linear relationship between price, min nights, number of reviews and availability using the scatter plot. If the price of an area is higher, more hosts want to make listings there. Also, more reviews imply a higher number of customers. From the data, Bedford-Stuyvesant is the busiest area cause it has the highest number of listings. With Average price of $115.35 and mean # of reviews of 40.



According to the data, we can see that prices of most listings are between 0 and $200, with a very small proportion of all listings in the dataset will have their price of more than $200.



According to the data (the number of reviews), Brooklyn has the highest number of reviews, with Manhattan right behind it at 39.3% of the market share in NYC.

```python
# Find optimal depth for random forest regression model
depth = 0
accuracy_score = 0
for i in range(2, 10):
    CLF = DecisionTreeClassifier(criterion="entropy", max_depth=i)
    CLF = CLF.fit(X_train, y_train)
    y_pred = CLF.predict(X_test)
    if metrics.accuracy_score(y_test, y_pred) - accuracy_score > 0.01:
        depth = i
        accuracy_score = metrics.accuracy_score(y_test, y_pred)

# Model evaluation
CLF = DecisionTreeClassifier(criterion="entropy", max_depth=depth)
CLF = CLF.fit(X_train, y_train)
y_pred = CLF.predict(X_test)
print("Depth:", depth)
print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
```

```
Depth: 6
Accuracy: 0.589242053789731
```

For the random forest regression model, we need to decide the value of depth. The goal is to maximum the accuracy and make depth as small as possible. The algorithm above fit the model by different values of depth from 2 to 10, and pick the depth that has a significant effect on the accuracy of the regression model. So, if the accuracy of the model increases 0.01, we will consider this increase as significant. In the end, the algorithm tells the best depth is 6, and the accuracy of the model is 58.92%

•Discussion and Conclusion:

This project investigated several techniques learned throughout the semester and challenged every member in term of data mining. Two members never touched coding before, ever. We performed a series of prediction analysis, extract the outliners and compare between different prediction to find any connection and true data. RMSE of the linear regression model is around 250, so errors will be large when using this model to predict the price. Not perform detailed variable analysis before fitting the model, so some variables may not very significant to price. Not consider the interaction between different variables. For classification, classifying randomly will make a 33% accurate prediction, so the accuracy of the random forest regression model may not be acceptable. Only do classification and regression on price, clustering may be a better approach method in the future.