# Capstone Project Report

Yuanchun Wang
3/16/2020

## Table of Contents

# I. Capstone project proposal

**New York Airbnb Data**
Background: Since 2008, more and more customers choose Airbnb to live when they are travelling. And this Airbnb data exhibits all the information about the hosts, geographical availability, necessary metrics to make prediction and draw conclusions. The data can be employed to predict the Airbnb price based on customer' s needs.

1. **What is the problem you want to solve?**
   When the customer travels to New York, they need to find a right room to rent. However, how to choose the most suitable one from so many Airbnb? Consequently, the problem for me to fix is that setting up a model for the customer to use for making better choice.

2. **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?**

   When the customer travels to New York, they need to find a right room to rent. However, how to choose the most suitable one from so many Airbnb? Consequently, the problem for me to fix is that setting up a model for the customer to use for making better choice. My model is mainly for those New York travelers who need to book the Airbnb room. I will build a website and my client just need to answer several questions, then I can recommend him/her a right Airbnb.

3. **What data are you using? How will you acquire the data?**

   I will use the New York Airbnb data download from Kaggle website
   https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

4. **Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

   a. Load the data and clean the data
   b. Visualize the data from multiple analysis
   c. Build the model for regression analysis
   d. Build a website for customer to make a better choice

5. **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

There are mainly two deliverables for this project:

1. Jupyter notebook that includes all my raw code and reasoning for the decisions I made.
2. PowerPoint presentation that summarizes the key results from the project and future directions that would be interesting to pursue.

# II. Data collection and data wrangling summary

My capstone 1 project data (New York Airbnb data) is downloaded from Kaggle website; ‘pd.read_csv’ is employed to import data to my ipynb file and named ‘data’. This data has 48895 entries and 16 columns, including ‘host_id’, ‘host_name’, ‘neighbourhood_group’, ‘neighbourhood’ and so on.

Data cleaning: Drop unnecessary and column with too many missing values, including ‘last_review’ ‘host_name’ columns. The code is ‘data.drop([‘host_name’, ‘last_review’], axis=1, inplace=True)’.

Missing values: firstly, I use ‘data.isnull().sum()’ to detect the missing value and find out the column ‘review_per_month’ has 10052 missing value; then, ‘data[data==‘ ’] =np.nan’ is used to fill ‘NaN’ in the missing value area.

Outliners: The column ‘price’ has outliners, where the price is more than 500. When I analyze the relationship between ‘price’ with ‘availability_365’, I employ new dataset data 1(‘data1=data[data.price<500]’) to do the further analysis.

Encode the data: when calculating the correlation index, the column ‘neighbourhood_group’ and ‘room_type’ are factorized to be number format, which are used for kendall correlation analysis.

# III. Data analysis

1. For the New York Airbnb data, my main goal is to find all the factors that may influence the Airbnb price. So firstly, I analyze the related factors, such as ‘neibourhood_group’ and ‘room_type’ and ‘host_id’.

   1.1 I identify that top hosts who have most listings. Interestingly, host_ID ( 219517861) host for 327 times and the host time for top 10 host range from 327 to 52. The bar plot of the top host numbers indicates the decreasing trend (Figure 1).

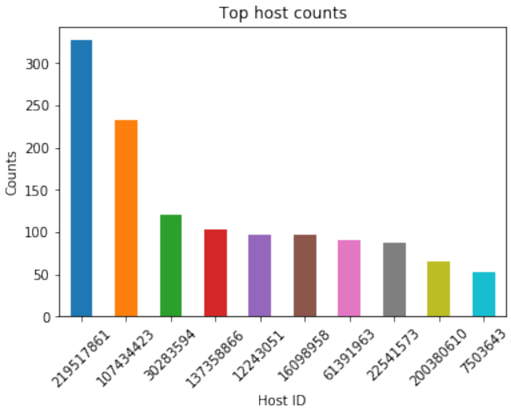

Figure 1. The bar plot of the listing numbers of the top ten host.

1.2 There are three room types in these Airbnb; more than 50% Airbnb belong to the ‘entire home/apt’ type. I build the bar plot to visualize it (Figure 2). Let’s pay attention to column ‘number_of_reviews’. Among the top ten reviewed listings, the average price per night is $65.4 and the 90% of the room type is ‘Private Room’. It is significantly higher than 45% of private room in the whole Airbnb market (Figure 1). It indicates that private room has more customer reviews, no matter it is good review or bad one.

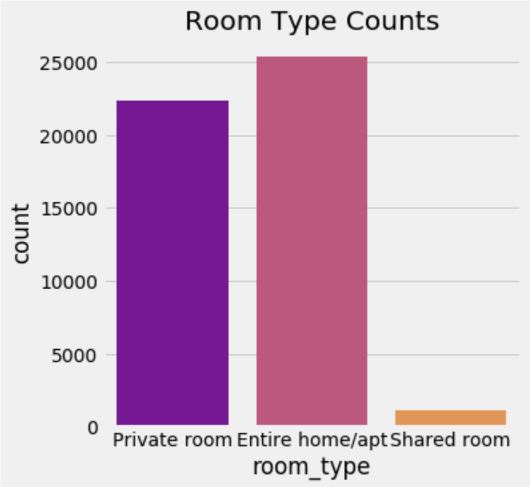

Figure 2. The bar plot of ‘room_type’ counts.

1.3 There are five ‘neigbourhood_group’ in New York and ‘Manhattan’ has the most Airbnb, followed by ‘Brooklyn’; the ‘Staten Island’ has the least Airbnb. The bar plot is employed to visualize the distribution trend in it.

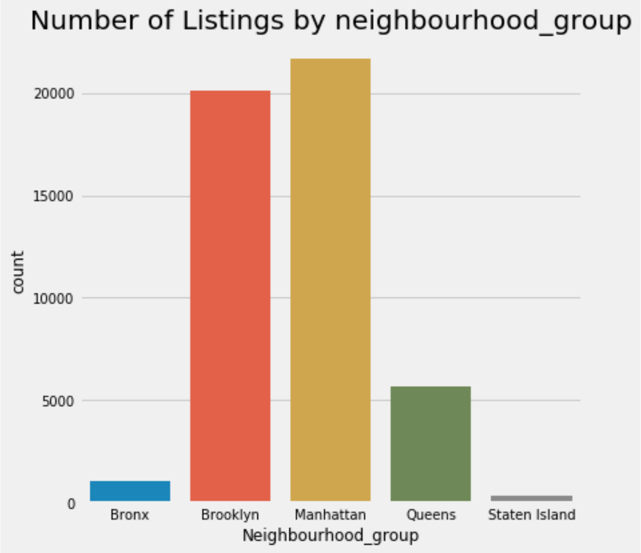

Figure 3. The bar plot of the listing numbers of the different ‘neighbourhood_group’.

2 I compare the relationship between ‘neighbourhood_group’ and ‘price’ and visualize it with violin figure(Figure 4); I find that the average price of ‘Manhattan’ has the highest price at average $150, then is ‘Brooklyn’ with $90; the ‘Bronx’ is the cheapest with $65. This indicates that ‘Manhattan’ is the most expensive neighbourhood_group and ‘Bronx’ is the cheapest one.

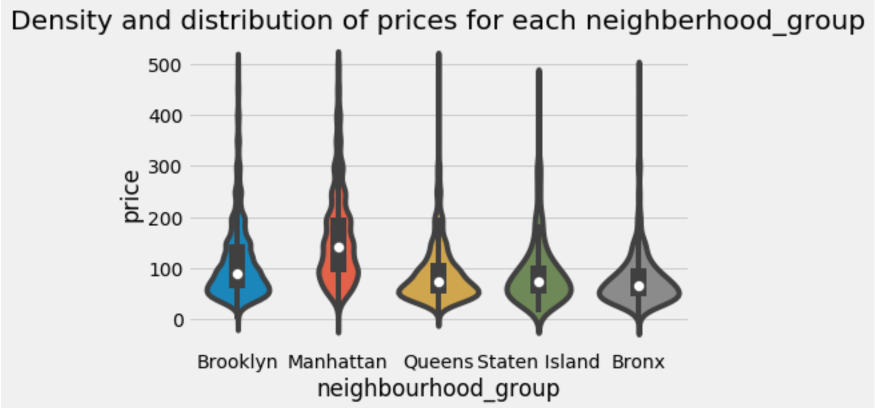

Figure 4. The violin plot of the price distribution in different ‘neighbourhood_group’.

3 In order to identify the relationship between location and room_type, a sns scatter plot has been built with the column ‘latitude’, ‘longitude’ and ‘room_type’. (Figure 5). We can predict that there is no strong distribution trend between ‘room_type’ with ‘location’.

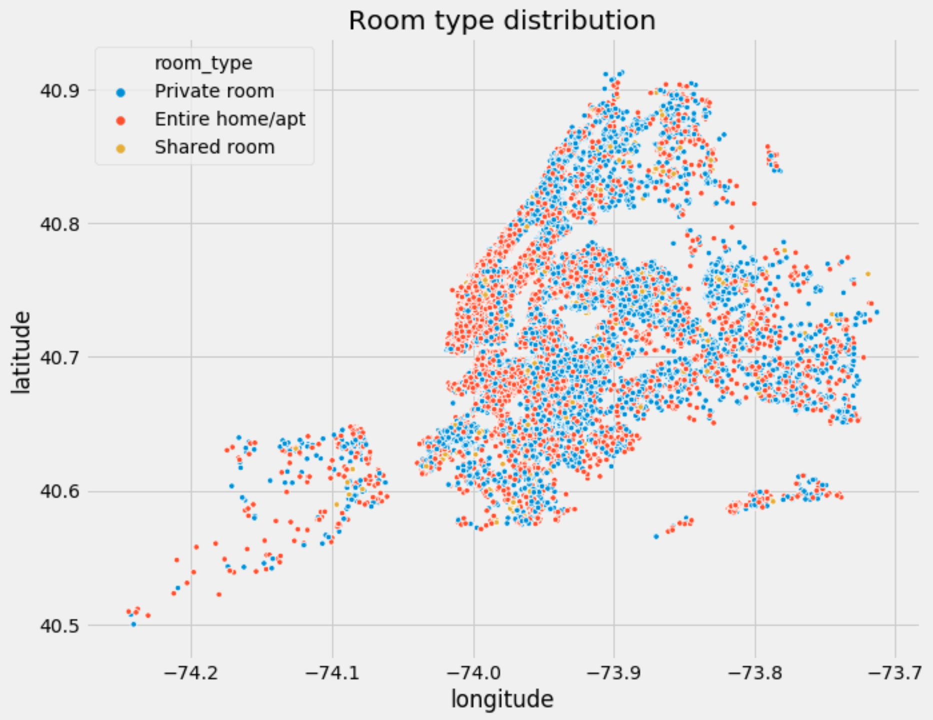

Figure 5. The scatterplot of the 'room_type' distribution among different locations of the Airbnb.

4 In order to identify the relationship between the neighbourhood and airbnb listing numbers, I count the top ten listing numbers of neighbourhood and find that the top ten numbers range from 3920 to 1545; and in the further analysis, I build the catplot of the top ten neighbourhood and the room_type and identify that "Williamsburg" has the highest number of "Entire home/apt" room type; 'Bedford-Studyvesant' has the highest number of 'Private room' room type; both of them belong to 'Brooklyn' neighbourhood_group(Figure 6). No clear trend is recorded.

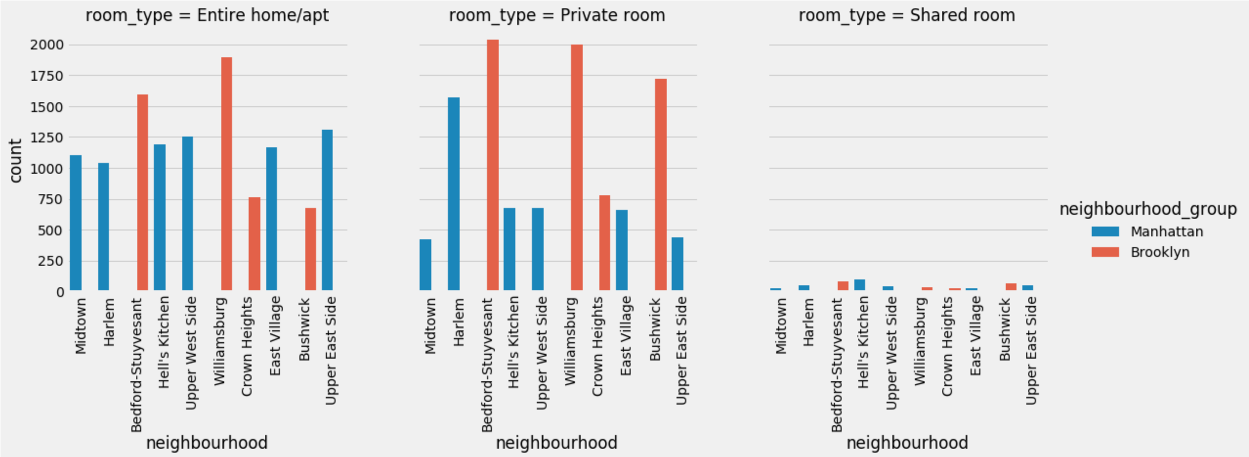

Figure 6. The catplot of top ten neighbourhood listing numbers in three different 'room_type'.

5 As we all know, neighbourhood group 'Manhattan' and 'Brooklyn' are the most travelled destination for the travelers and also they are the most Airbnb listings availability. How

about price? A color map is employed to visualize the price distribution in New York Airbnb market (Figure 7) and the color changing from blue to yellow to red to mark the corresponding price change from low to high. It is very clear that most red, yellow dots are located at ‘Manhattan’, which indicates that if you want to live in the most expensive rooms, you can find it in Manhattan. If you want to live in the cheapest rooms, you can find it in ‘Bronx’.

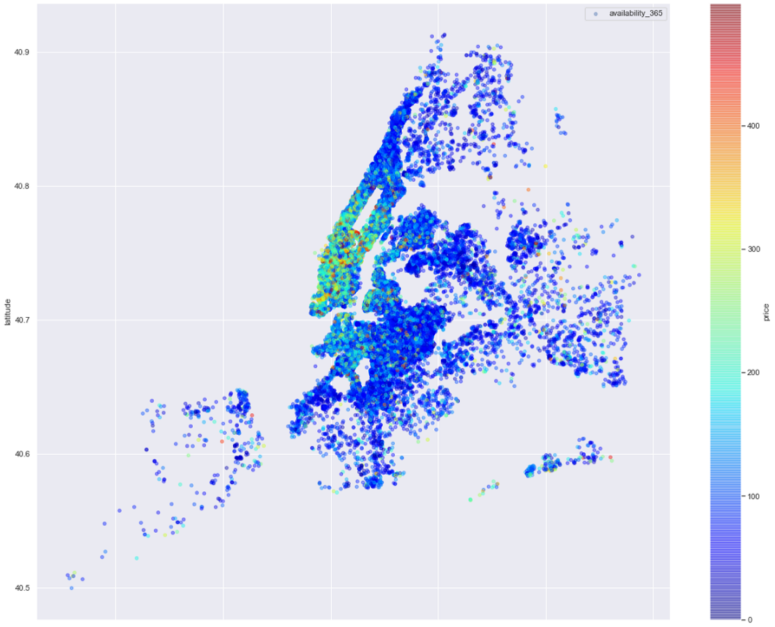

Figure 7. The color map of price distribution in the whole New York Airbnb market.

6 Among ‘neighbourhood_group’, ‘room_type’, ‘minimum_nights’, ‘caluclated_host_listings_count’,’availability_365’ these factors, which one has the strongest correlation with the Airbnb price?

I employ ‘Encode’ method to transform related factor from word to number and calculate the correlation index between them with ‘price’; The correlation index is summarized in figure 1, which is changing from 0 to 1 and the indicating color changes from dark to red to white. ‘room_type’ has strong correlation with ‘price’ (corr=0.47) (Figure 8).

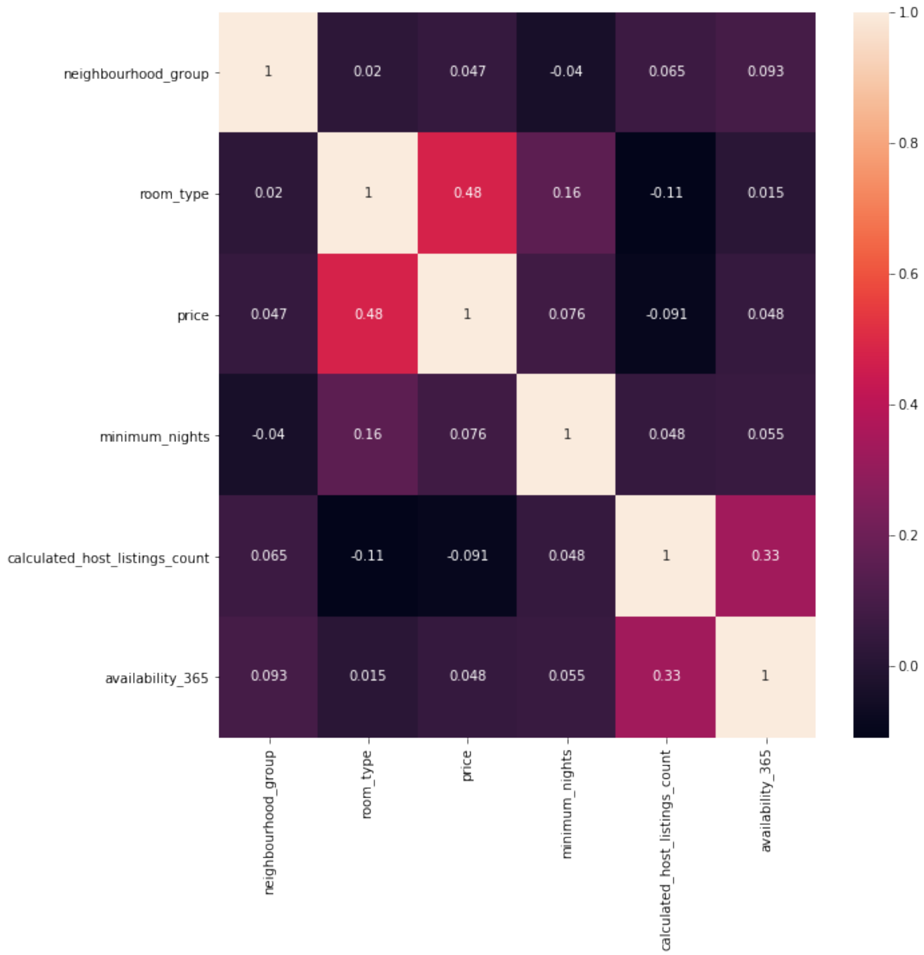

Figure 8. The Kendall correlation figure between ‘neighbourhood_group’, ‘room_type’, price, ‘minimum_nights’, ‘calculated_host_listings_count’ and ‘availability_365’.

Based on my analysis, the factor ‘room_type’ has strong correlation with Airbnb price. So my next step is to employing statistical method and machine learning to confirm this result and finish model prediction.

# IV. Statistical analysis

In my analysis, my question is how to find a right Airbnb hotel for customers based on New York Airbnb data.

1. Based on the previous analysis, the price has strong correlation (corr=0.47) with the ‘room_type’. More than 90% room are ‘private_room’ or ‘entire home/apt’. And my question is that if the price ‘private_room’ and ‘entire home/apt’ has significant difference. Thus, I calculate the p-value between ‘private_room’ price data and ‘entire home/apt’ price data.
2. Based on the bootstrap sampling method, after performing 10000 replicates immediately after setting the random seed to 47, the lower limit value for price is $150.95.

3. The average price difference between ‘entire home/apt’ and ‘private room’ is $122.01. 95% confidence interval for the difference between the entire room/apt and private room is [68,180]. Based on the p-value calculation, the p-value between subgroup ‘private_room’ and ‘entire home/apt’ at 95% confidence interval is 0.0, that is less than 0.05, which indicates the two subgroup has significant difference.

# V. Machine learning analysis

My project goal is to build a model to predict the propriate Airbnb room for the travelers. Thus, constructing a model based on machine learning is necessary. So linear regression method and Decision Tree Regressor are mainly employed to build the price prediction model to find out which one is better.
The first method is linear regression and the r2 score between ‘y_test’ and ‘y_predict’ is 0.0748 based on linear Regression model.
The second method is Decision Tree Regressor and Based on Decision Tree Regressor, r2 score between ‘y_test’ and ‘y_predict’ is 0.2534, which is higher than linear Regression model; it indicates that Decision Tree Regressor is a better fitting model.
In the unsupervised learning part, among the 'room type' factor, the first value 'Entire home/apt' is one possible value for the optimal number of dimensions (Figure 9).

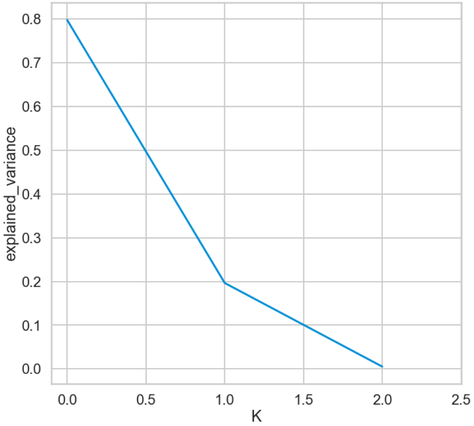

Figure 9. PCA analysis results.