Springboard

SPRINGBOARD CAPSTONE PROJECT ONE
NEW YORK AIRBNB PRICE PREDICTION

YUANCHUN WANG

02/16/2020

Springboard

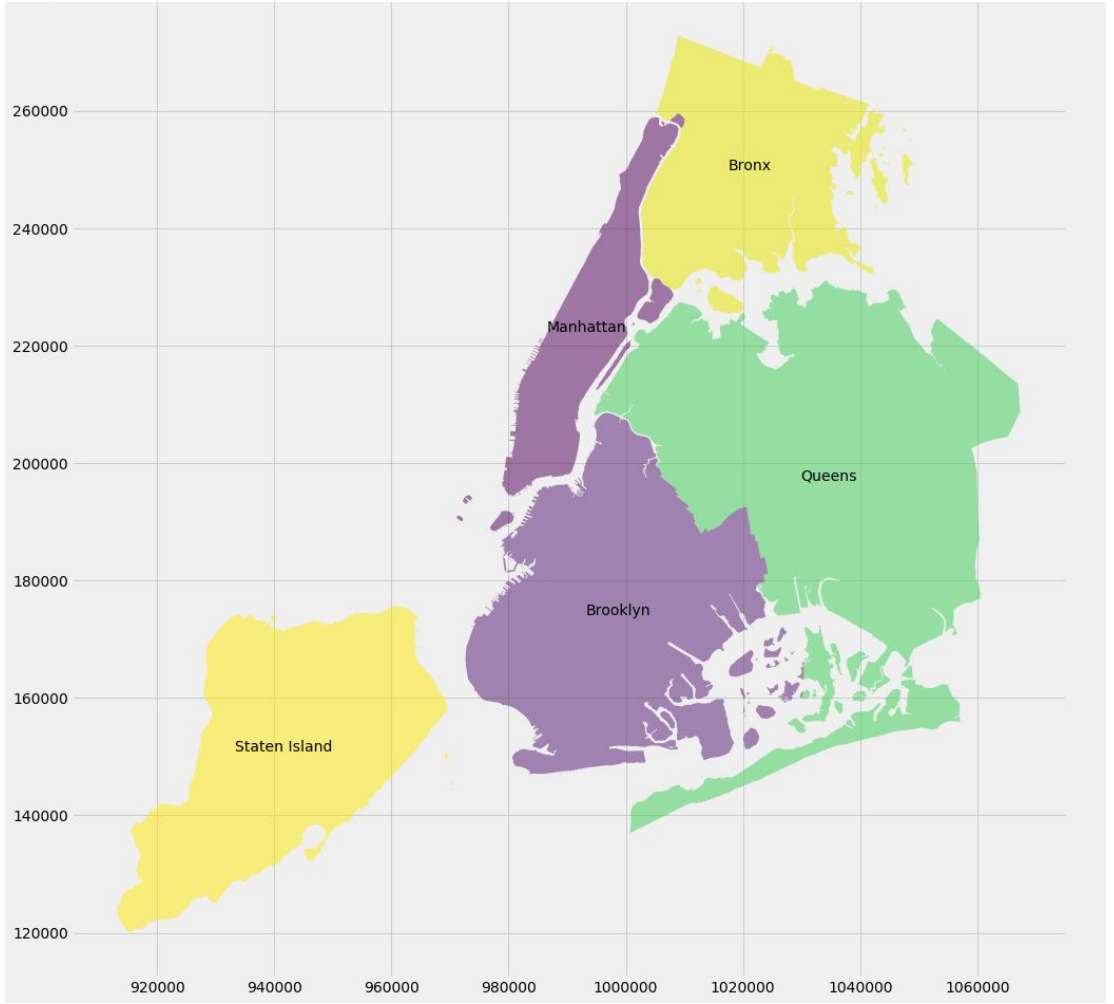
OUTLINE

- Project goal
- Data cleaning
- Data wrangling
- Data visualization
- Statistics analysis
- Machine learning prediction
- summary

Springboard

INTRODUCTION & PROBLEM STATEMENT

- Almost 5 million of travelers visit New York annually.
- New York has more than 40000 Airbnb for the travelers to choose.
- How to choose an appropriate one based on location, price and the other factors?



New York Map

Springboard

Outline

Springboard

DATA ACQUISITION & WRANGLING

- Data source:

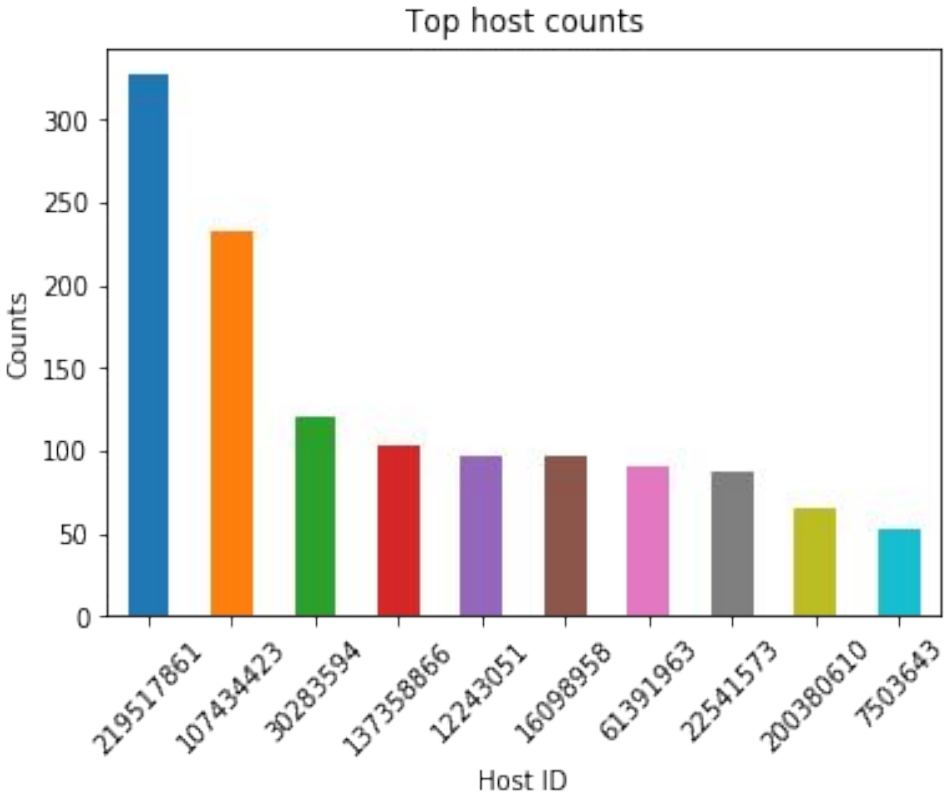
  Kaggle website: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

- Data wrangling:

1. Drop unnecessary and column with too many missing values, including ‘last_review’ ‘host_name’ columns.
2. Fill ‘NaN’ in the missing value area.
3. Remove the ‘price’ outliners.
4. Encode the data: factorize the ‘neighbourhood_group’ and ‘room_type’ to number format to finish the kendall correlation analysis.

FURTHER ANALYSIS

- 1. Identify that top hosts who have most listings



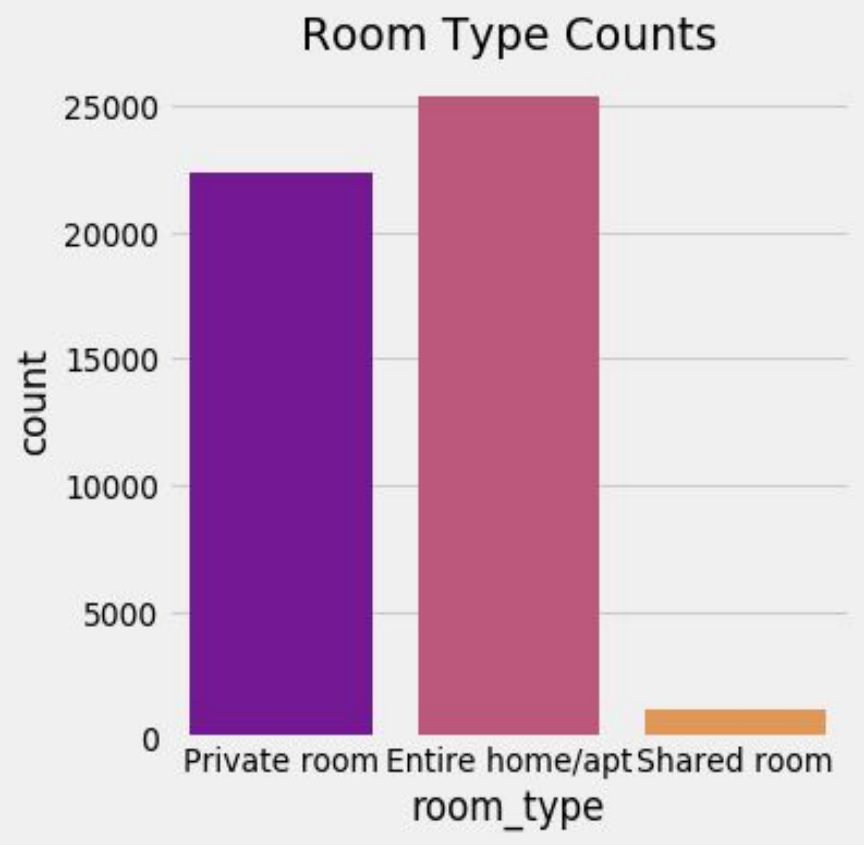
Host_ID (219517861) host for 327 times and the host time for top 10 host range from 327 to 52.

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

Springboard

FURTHER ANALYSIS

2. Identify the room type distribution trend



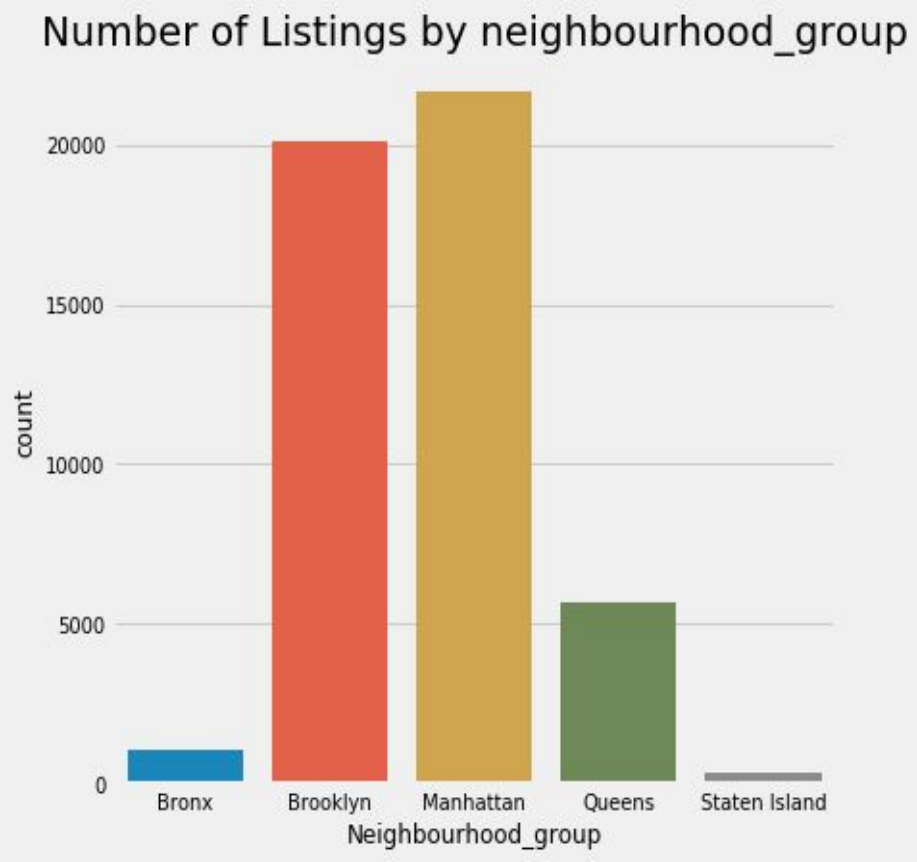
More than 50% Airbnb belong to the ‘entire home/apt’ type.

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

Springboard

FURTHER ANALYSIS

3. Summary the number of listings by neighbourhood_group.



‘Manhattan’ has the most Airbnb, followed by ‘Brooklyn’; the ‘Staten Island’ has the least Airbnb.

Springboard

FURTHER ANALYSIS

4. Neighbourhood group vs Price

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

Springboard

5. Location vs room type

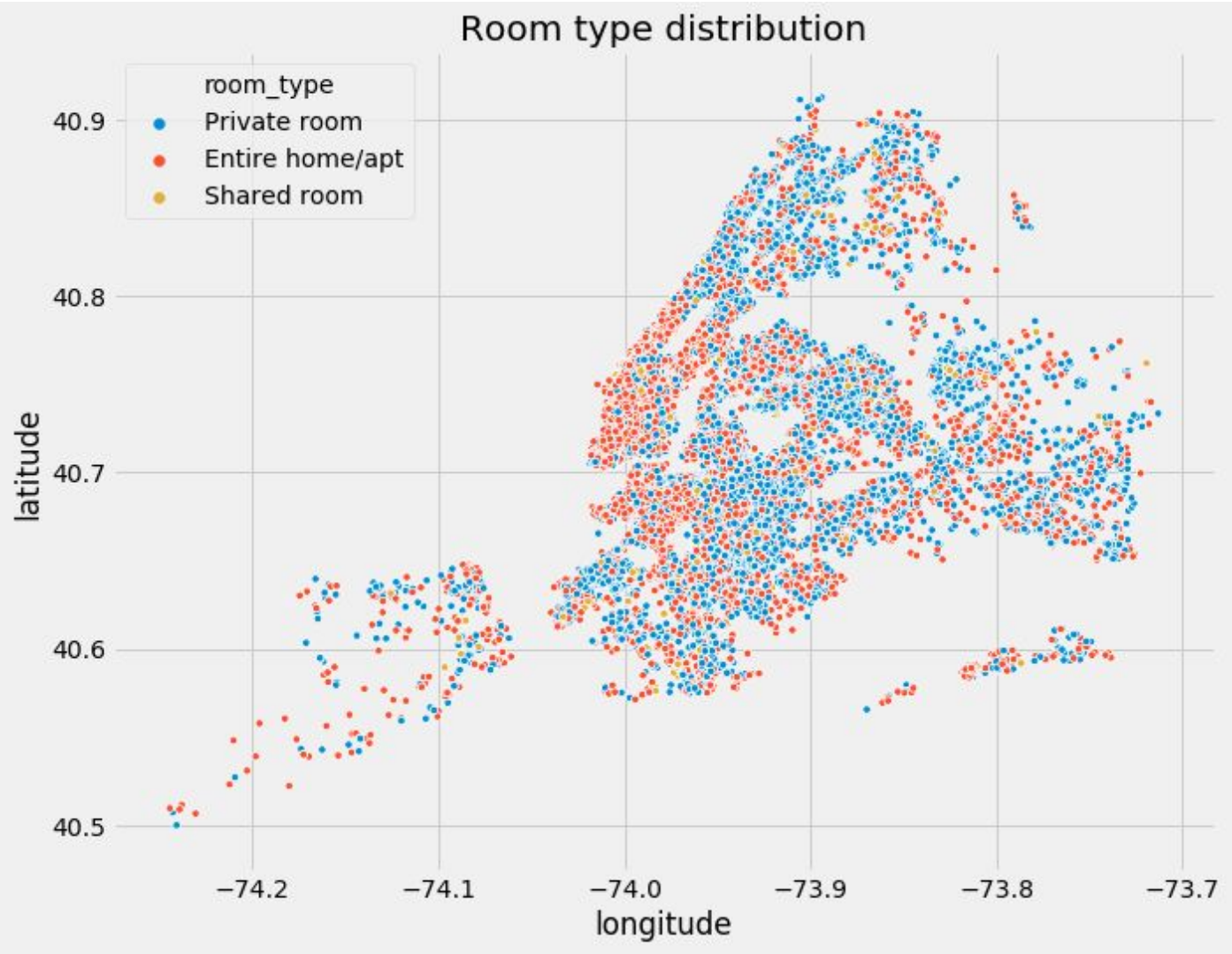


Figure 5. The scatterplot of the ‘room_type’ distribution among different locations of the Airbnb.

FURTHER ANALYSIS

6. The room type distribution of top ten neighbourhood listings

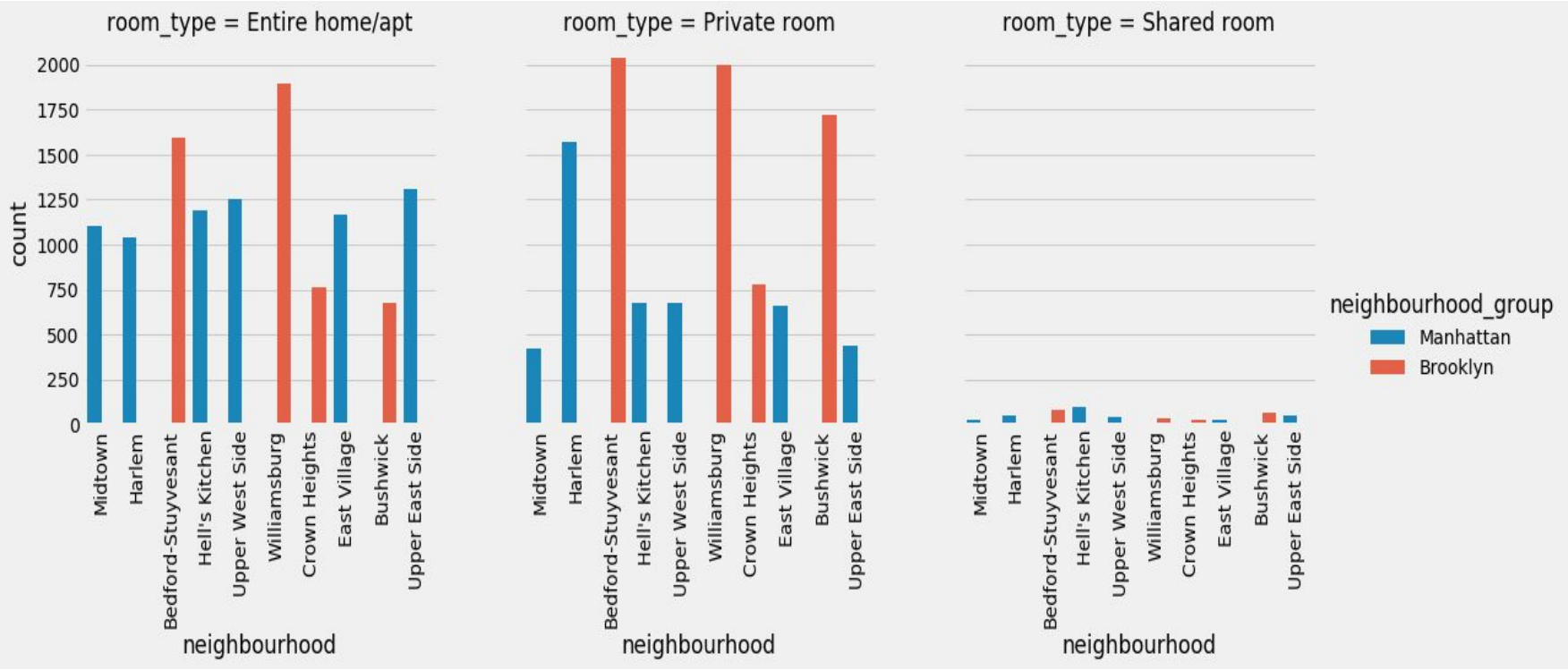


Figure 6. The catplot of top ten neighbourhood listing numbers in three different ‘room_type’.

Springboard

7. ‘Manhattan’ has more expensive Airbnb

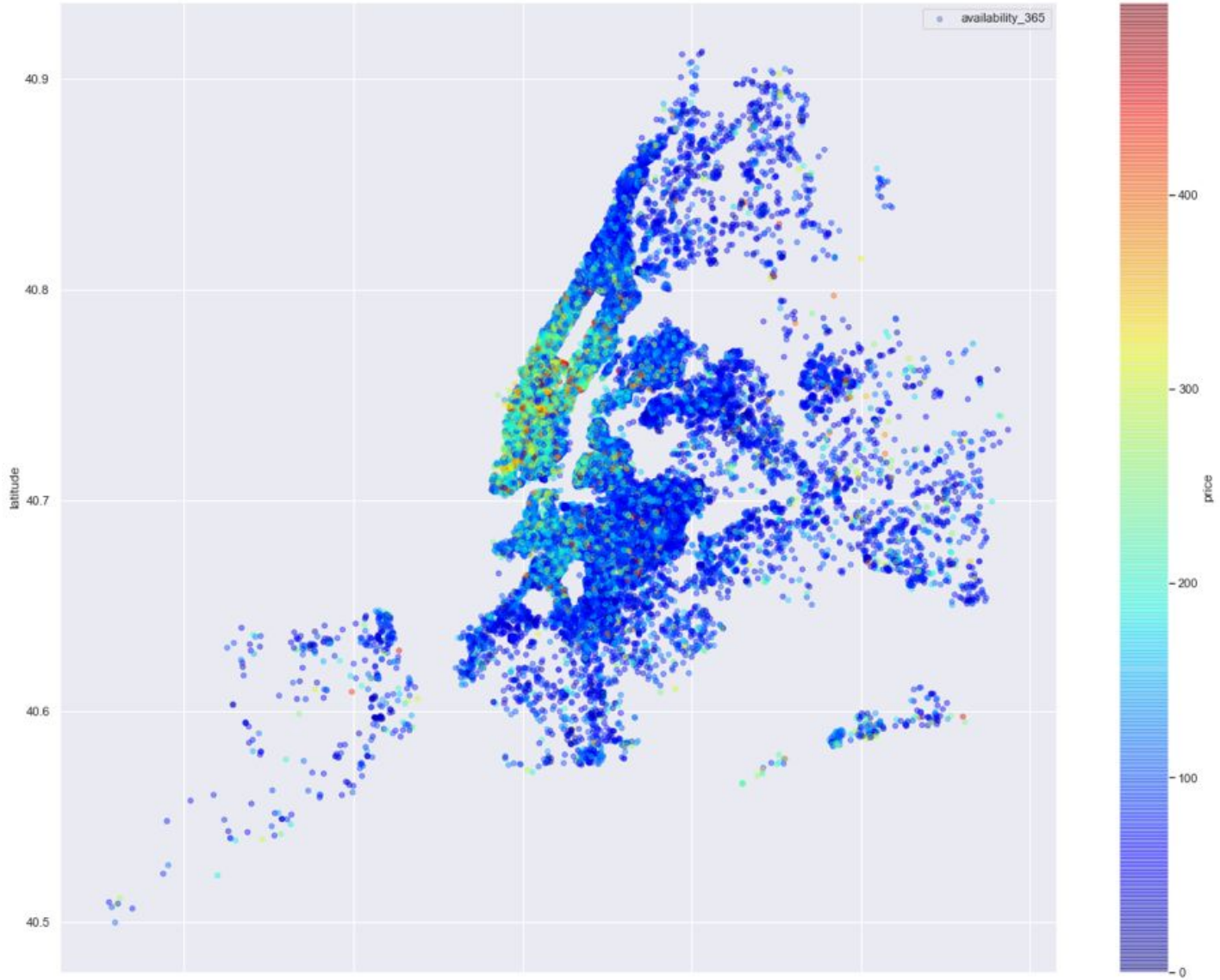


Figure 7. The color map of price distribution in the whole New York Airbnb market.

DATA SUMMARY

- 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.

Springboard

STATISTICAL ANALYSIS

- QUESTION: which factor has the strongest correlation with price?

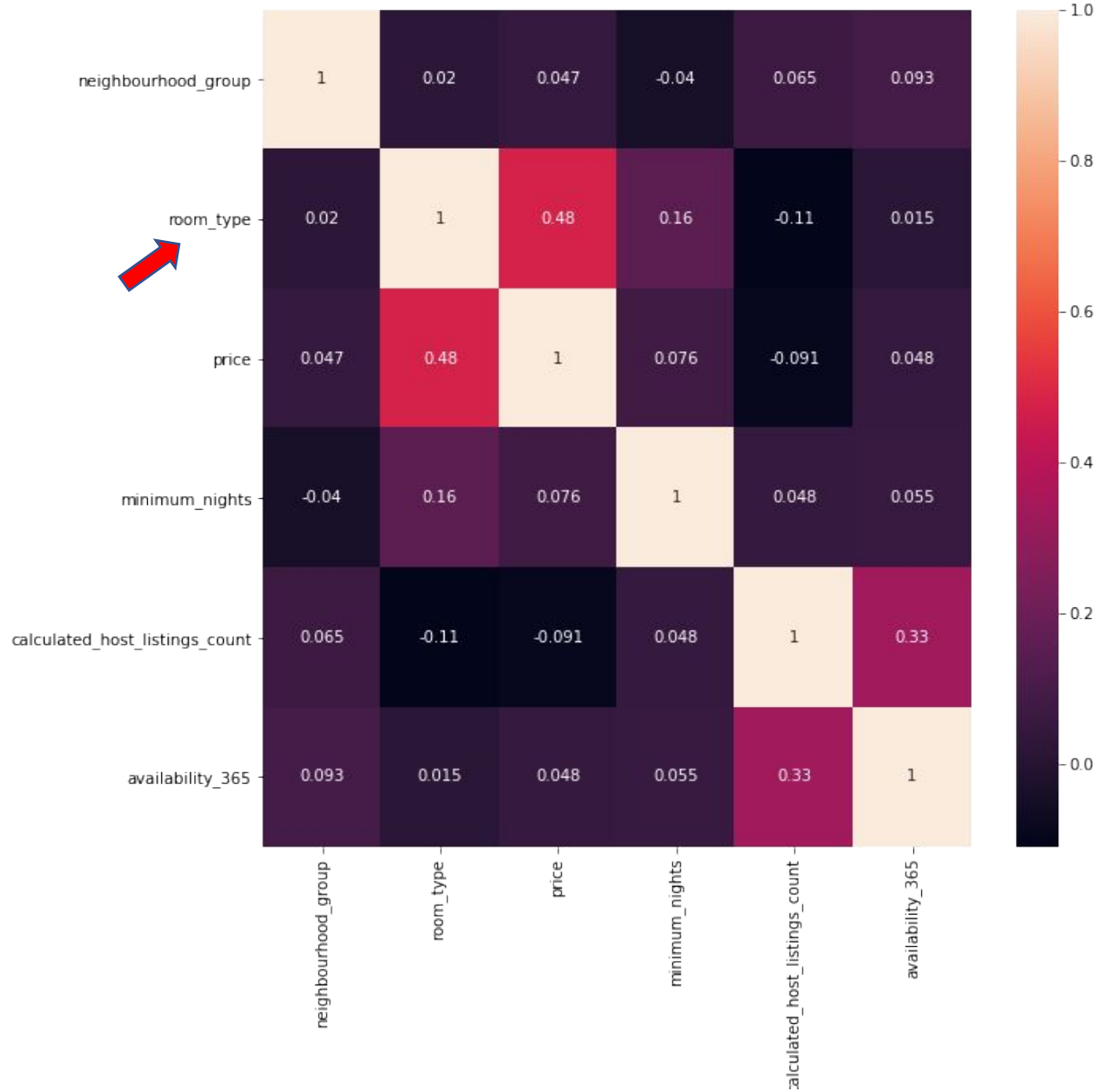


Figure 8. The Kendall correlation figure between 'neighbourhood_group', 'room_type', price, 'minimum_nights', 'calculated_host_listings_count' and 'availability_365'.

Springboard

STATISTICAL ANALYSIS

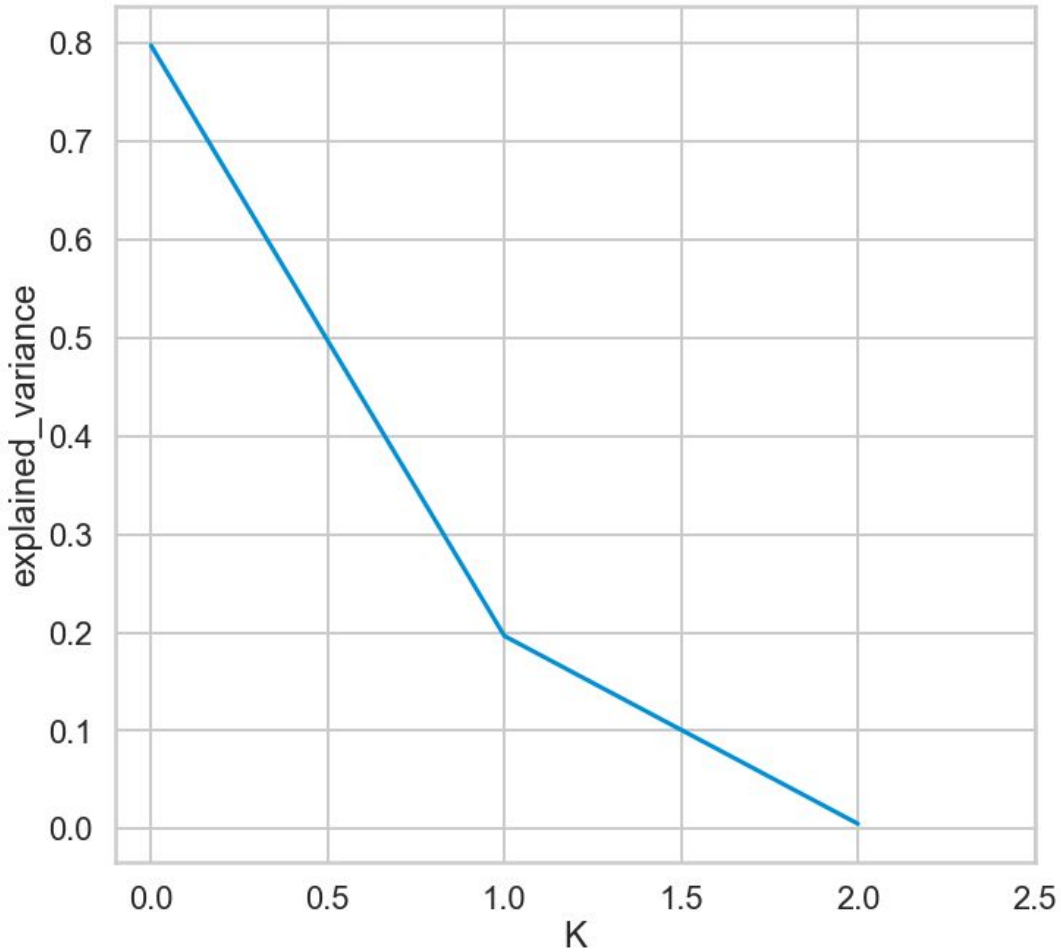
- **1. Bootstrap sampling to estimate the same 95% confidence interval lower limit is $150.**
- **2. The mean price difference between entire room and private room is $122.**
- **3. Based on the p value 0.0, we can predict the price between entire home/apt and private room has significant difference.**

Springboard

MACHINE LEARNING

- **r2 score between y_test and y_pred is 0.0748 based on linear Regression model.**
- **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.**

PCA analysis finds that among the 'room type' factor, the first value 'Entire home/apt' is one possible value for the optimal number of dimensions.¶



Springboard

FUTURE IMPROVEMENTS

- Employ more machine learning model to find the best model to fit my data, such as random forest;
- Check the overfitting for my constructing model

Springboard

THANKS FOR YOUR ATTENTION!