Capstone Project 1:

New York City Airbnb Price Prediction

Catherine Somers

PROBLEM STATEMENT

Airbnb is an American online marketplace and hospitality service for arranging lease or rent lodging in the short-term. The company does not own any of the real estate and instead a broker receives a percentage of the services fees with each booking.

Since 2008, Airbnb has become one of the largest marketplace for hospitality services with over 6 million listings in around 190 countries around the world.

An obstacle that both Airbnb hosts and Airbnb faces is determining an optimal rent price per night. As a host on the platform, charging too much causes the renter to choose more affordable options fit to their budget. Whereas, if they charge too low, they lose out on potential revenue on the property. If Airbnb could recommend an optimal listing price to hosts, it could help maximize revenue as well as service fees.

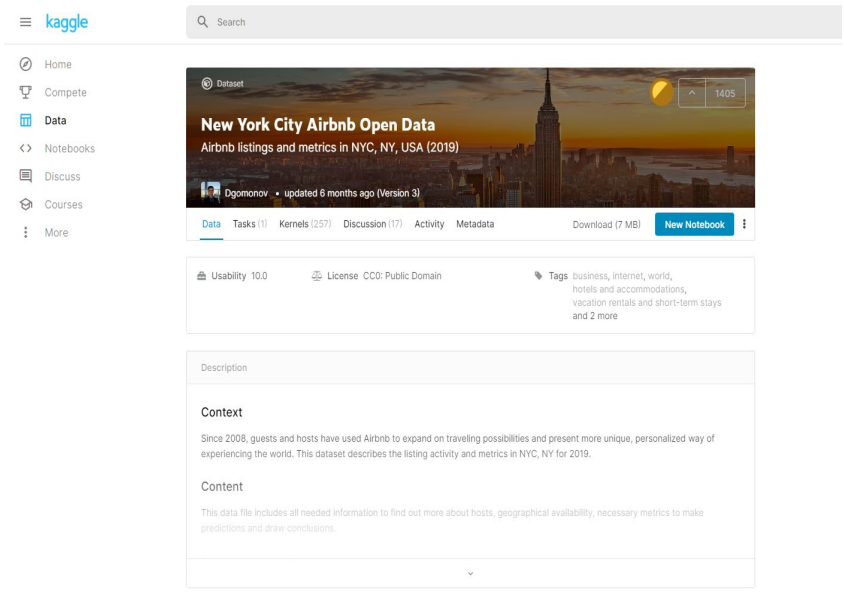
In this capstone project, we would like to be able to predict the listing price so that both the hosts and the company can optimally gain revenue.

CLIENT

The Airbnb hosts and Airbnb, the company would use price prediction for better pricing and to also maximize revenue.

DATASET

- The listing data for New York City can be found on Kaggle: New York City Open Data



The data consists of 16 different columns:

Host descriptors:

- host_id: host ID
- host_name: host name

- calculated_listings_count: amount of listing per host

Listing descriptors:

- id: ID
- name: name of listing
- room_type: living space type
- minimum_nights: amount of nights minimum
- availability_365: number of days listing is available for booking
- price: price in dollars

Review descriptors:

- number_of_reviews: number of reviews
- last_review: latest review
- reviews_per_month: number of reviews per month

Location descriptors

- neighbourhood_group: location
- neighbourhood: area
- latitude: latitude coordinates
- longitude: longitude coordinates

DATA WRANGLING

1. Data Collection and Importing
    a. Loaded in as a dataset. This dataset is publicly accessible and the .csv file contains summary information and metrics pertaining to Airbnb listing in New York City.
2. Selecting columns that may be useful in analysis
    a. Visually inspect data and select columns useful for analysis
3. Drop duplicates
    a. Removed repeat rows
4. Fix missing values
    a. Drop rows that contain null values since

5. Consider the outlier data
    a. In order for it not to skew the rest of the data we take the data between the 25th percentile and 75th percentile

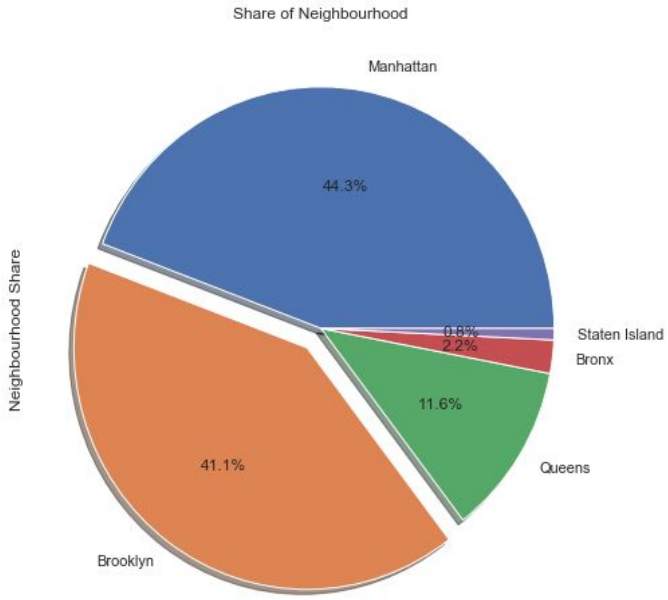
INITIAL FINDINGS

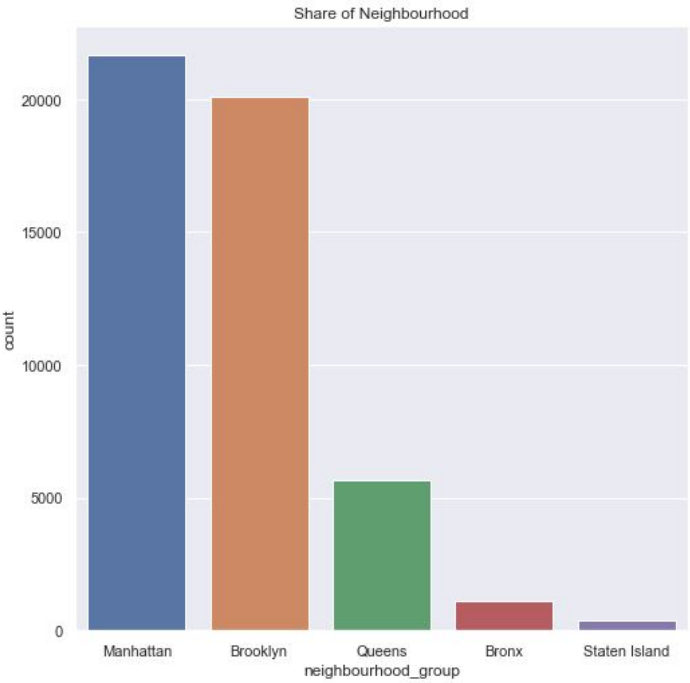
During exploratory data analysis, we asked the following questions:

1. Which neighbourhood group are these Airbnbs mostly located?
2. What is the price distribution based on the number of reviews?
3. What is the relationship between price and room type across neighbourhood and neighbourhood group?

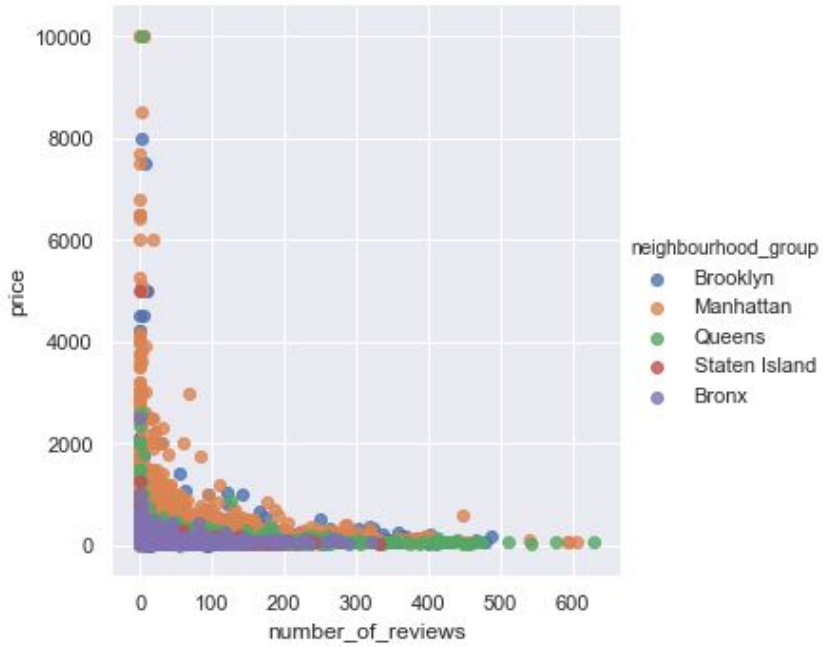
Initial findings are:

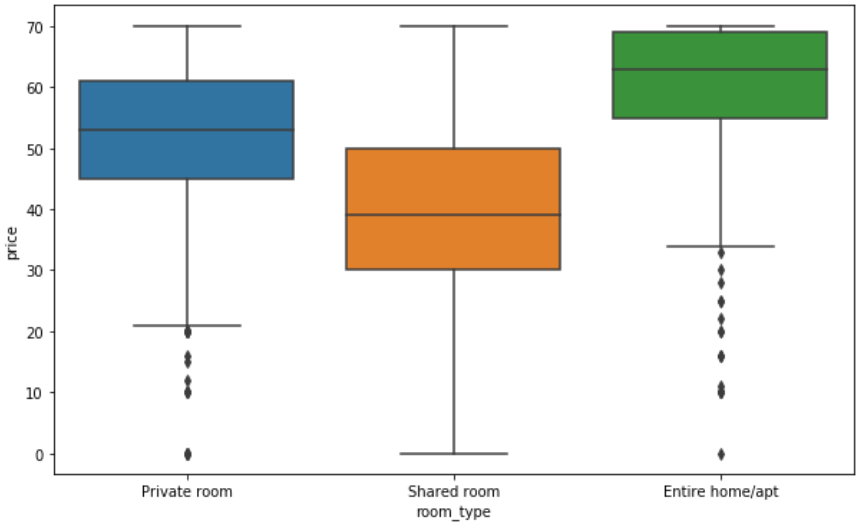
1. Most Airbnbs in New York City are located in the Manhattan and Brooklyn neighbourhood groups.
    a. 44.3% of Airbnbs located in Manhattan
    b. 41.1% of Airbnbs located in Brooklyn
    c. We can also see this evidence of both being the majority in the bar graph.
2. Among the number of reviews, price distribution of apartments are more concentrated around the Manhattan, Queens, and Bronx neighbourhood groups.
    a. Based on how we see the data spread out in the graph below
3. Price of the property is dependent on the room type across the neighbourhood and neighbourhood group.
    a. The mean price across all room types is not equal.
    b. Price is dependent on the neighbourhood group home is available in.
    c. Between the room type and neighbourhood group, the proportion of availability of a particular room type is dependent on the neighbourhood group searched for.


Share of Neighbourhood
Manhattan
44.3%
Staten Island
0.8%
Bronx
2.2%
11.6%
Queens
41.1%
Brooklyn
Neighbourhood Share

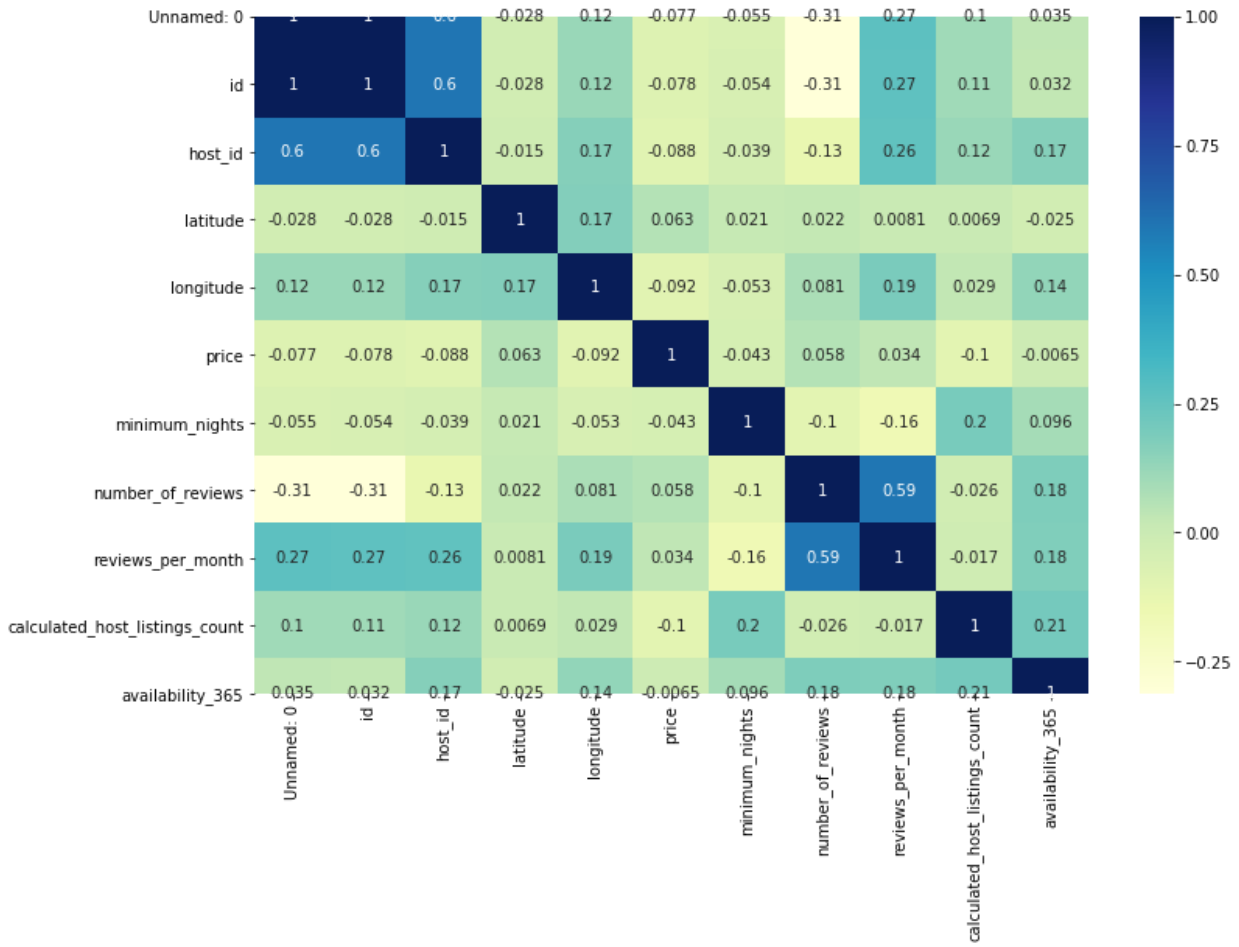

Share of Neighbourhood
20000
15000
10000
5000
0
count
Manhattan
Brooklyn
Queens
Bronx
Staten Island
neighbourhood_group


Distribution of Price of Apartments
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
Price


10000
8000
6000
4000
2000
0
price
0
100
200
300
400
500
600
number_of_reviews
neighbourhood_group
Brooklyn
Manhattan
Queens
Staten Island
Bronx


70
60
50
40
30
20
10
0
price
Private room
Shared room
Entire home/apt
room_type

In



In the heatmap seen above, the highest correlation exists between host_id and id (0.6) closely followed number_of_reviews and reviews_per_month, while the lowest correlation exist between longitude and price (-.092).