Exploratory Data Analysis of Airbnb Booking Dataset

Shubham Patil, Swapnil Zambare, Bhushan Patil
Data science trainees, Almabetter, Bangalore

Abstract

Airbnb is an online marketplace that connects people who want to rent out their homes with people looking for accommodations in that locale. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world.

Nowadays, Airbnb has become a kind of service that is used by the whole world. Data analysts become a crucial factor for the company that provided millions of listings through Airbnb. These listings generate a lot of data that can be analyzed and used for security, business decisions, understanding of customers' and providers' behavior on the platform, implementing innovative additional services, guiding marketing initiatives, and much more.

1. Introduction

We present here our exploratory data analysis, visualizations, interactive plots, animations and lots of other interesting insights into the Airbnb data. We focus on New York City's data, for the very reason that we live in New York and second, we wish to perform an in-depth analysis on one of the most densely populated cities in the world.

1.1. Following are a few questions that we aim to answer through our analysis:

- What can we learn about different hosts and areas?
- What can we learn from predictions? (Ex: locations, prices, reviews, etc.)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

2. Data Explanation

Libraries Used:

We have used several packages during the analysis of the historical data of Airbnb in NYC in order to make data manipulation and visualization.

The list of libraries used in this project can be seen below:

- NumPy: For mathematical operations involving arrays.
- Pandas: For data manipulation.
- Matplotlib: For plotting.
- Seaborn : High level Matplotlib wrapper library

3. Understanding data

We found out that our dataset has 48895 listings in 16 categorical data. The features include listing name, host id, location information, location coordinate, room type, the price per night, and so on.

To understand the data we should know the different Airbnb room types. Based on the information on the Airbnb website, the definition of each room type are:

- **Private room** ✈ Guests have exclusive access to the bedroom/sleeping area of the listing. Other parts such as the living room, kitchen, and bathroom are likely open either to the host or even to other guests. Entire home/apt
- **Entire home/apt** ✈ Guests have the whole place for themselves. It usually includes a bedroom, bathroom, and kitchen.
- **Shared Room** ✈ Guest sleep in a bedroom or a common area that could be shared with others.

4. Data cleaning:

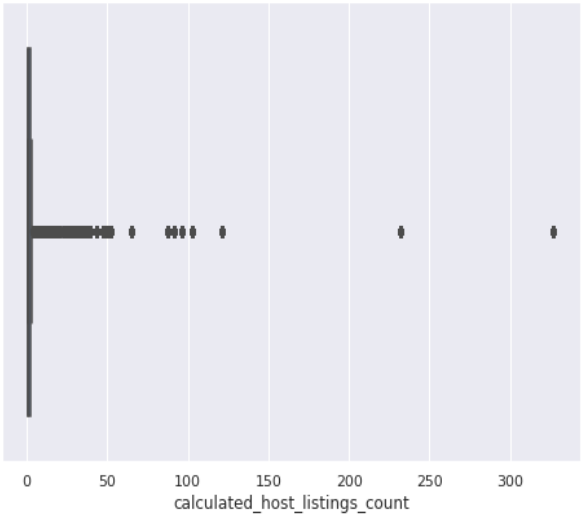
I. **We find out null and missing values:** In our case, the missing values that are observed do not need too much treatment.

II. **Removing redundant variables:** Looking into our dataset, we can state columns ‘id' and ‘host_name’, ‘last_review’ are irrelevant and unethical for further data exploration analysis. Therefore, we can get rid of those columns.

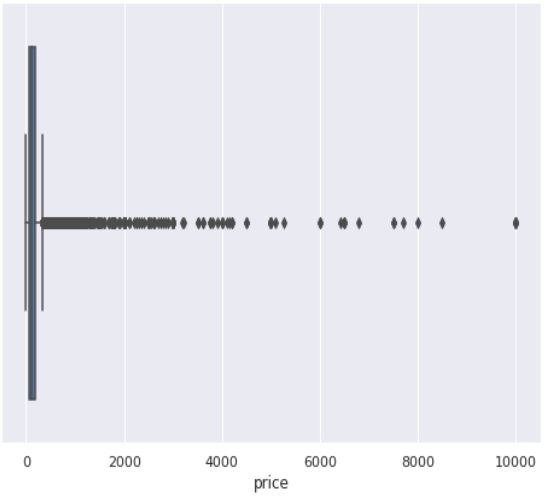
III. **Replacing the missing values:** Next, we need to replace all the missing values in the ‘review_per_month’ and 'price' column with 0 (zero) to make sure the missing values do not interfere with our analysis.

5. Finding Outlier:

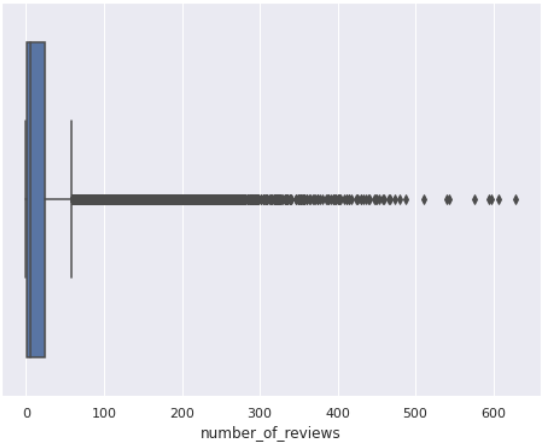
5.1. Host listing count:



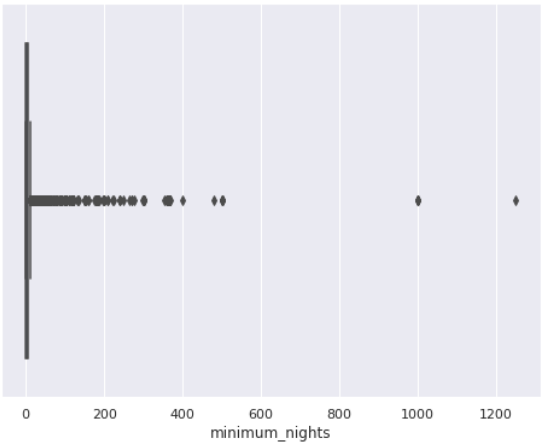
5.2.Price:



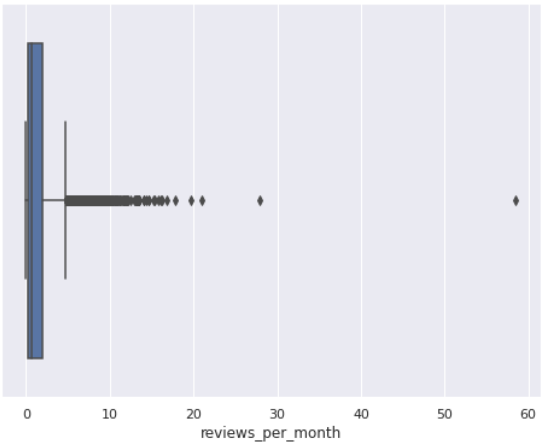
5.3. Number of reviews:



5.4. Minimum nights:

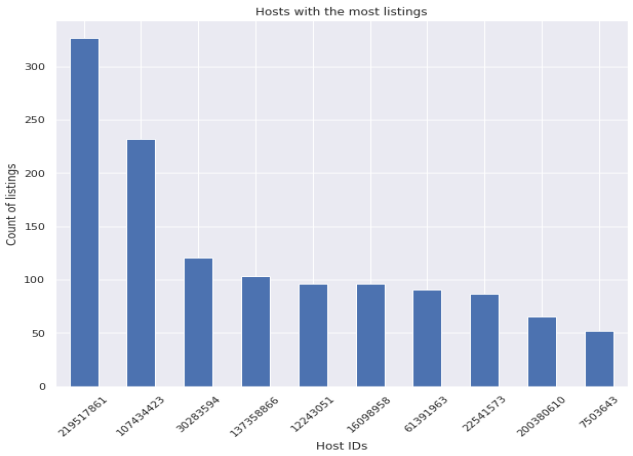


5.5.Reviews per month:



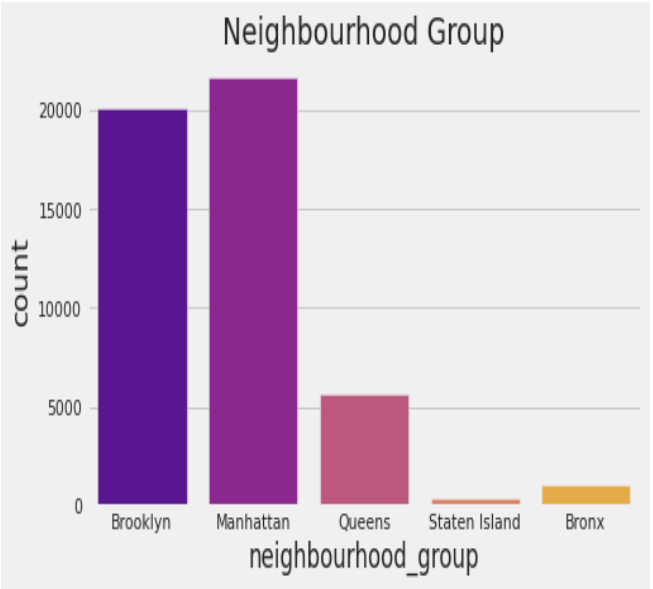
6. Data Visualisation:

6.1. Top listing count:



From the chart above, we can see the total of top 10 hosts is almost 2 %(1270 listings) of the whole dataset (48895 listings). Even one of the hosts has more than 325 listings**.**

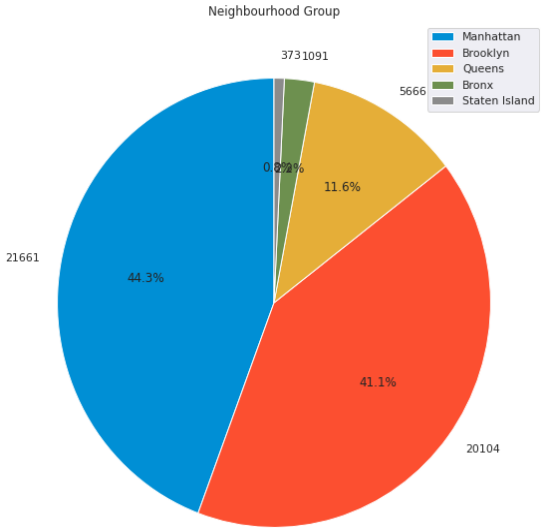
6.2. Counting Neighbourhood Groups:



As we can see from the graph, the maximum listing is from the Manhattan

(21661 listing) neighbourhood group while Staten Island has very few listings, only 373 listings.

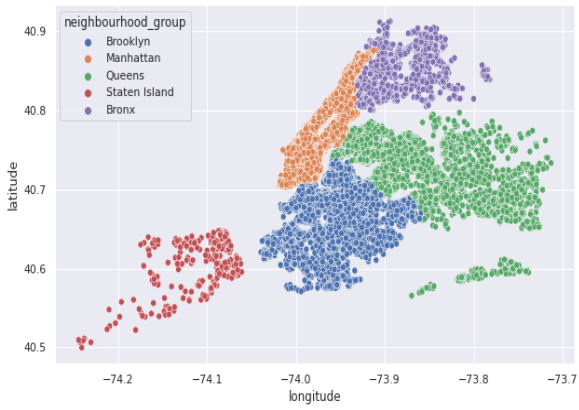
6.3. Top region area



From the chart above, we see the Manhatta n has the most listings with almost 21661 l istings number, covering more than 44.3% of the total listings.which was followed by Brooklyn with 20104 listing covering abou t 41% of total listing.

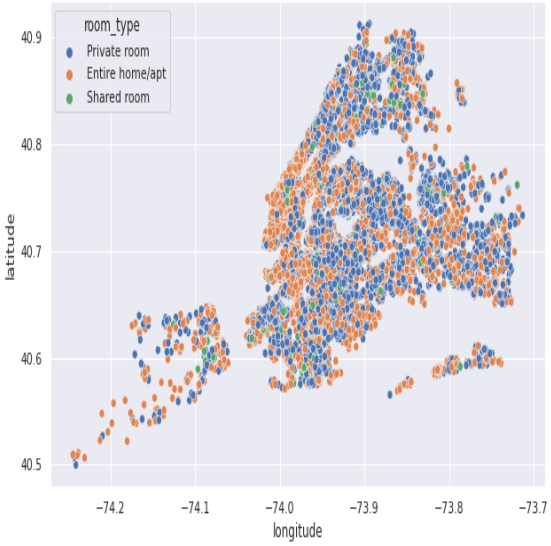
6.4. Visualisation map of every listing and Neighbourhood group

I. Density of listing:



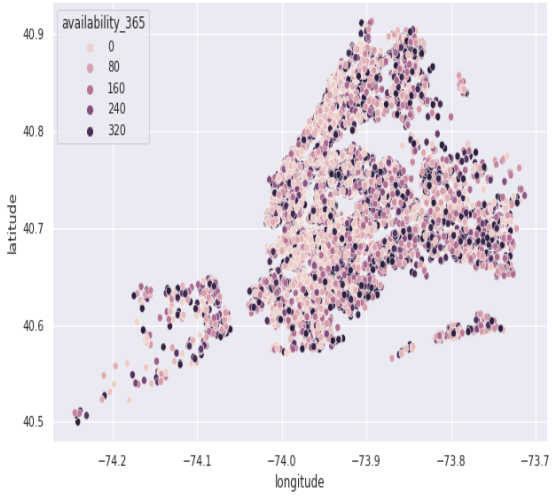
From the map above, we can see clearly where the densest listing is located, shown by the orange colour which is the area of Manhattan region is the densest while the area shown in red scatter plot is of Staten Island has very less density of listing.

II. Types of room preferred:



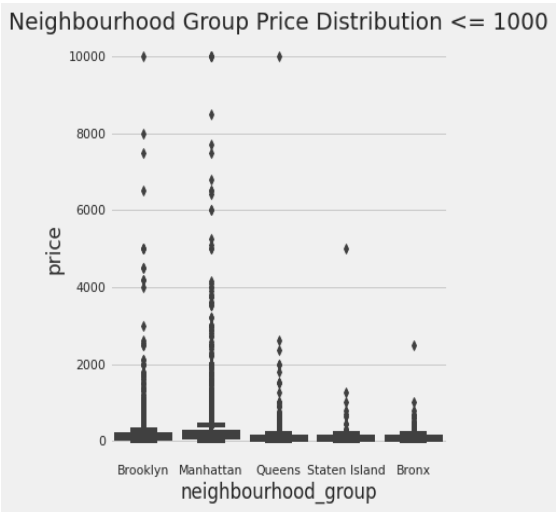
From the above scatter plot it has seen that people mostly prefer Entire Home/apt (25409 listings) which gives them more privacy and more roaming area than a shared room. Also we can see that all types of rooms are available throughout the neighbourhood group.

III. Room availability during 365 days:

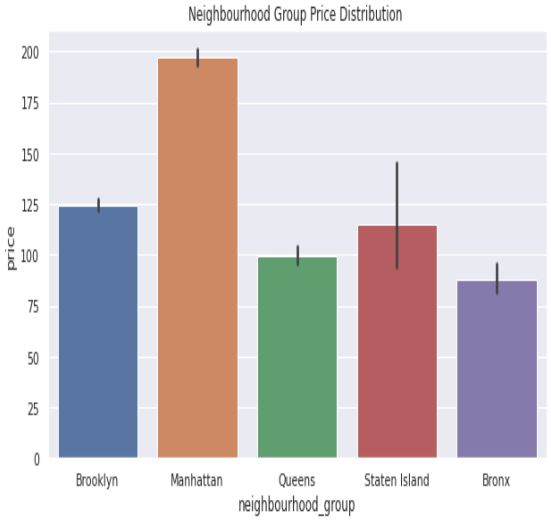


From the above scatter plot we can see that the faintly plotted area which is mostly in Manhattan has less room availability during 365 days. As people mostly prefer these areas for bookings.

6.5. Price Group Analyses of Neighbourhood Group

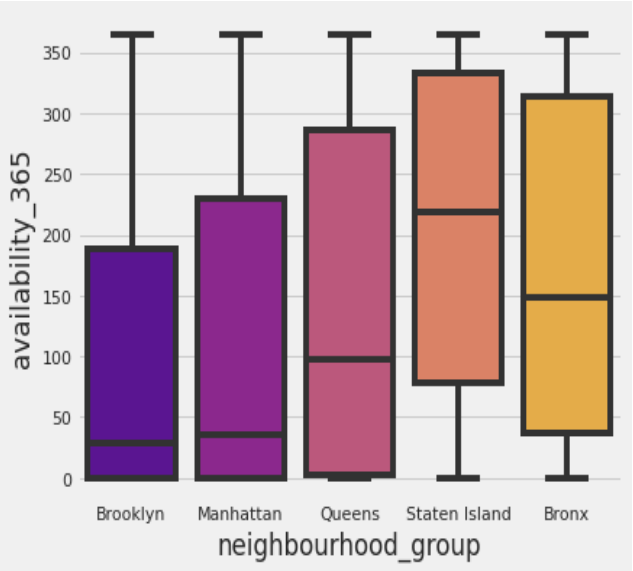


From the graph above at some of the places there is a very high price per night about $ 10000 as seen in the Manhattan and Brooklyn region. We also consider it as an outlier.



From the data above, we see the Manhattan Region has the most expensive price per night with a median $ 195.

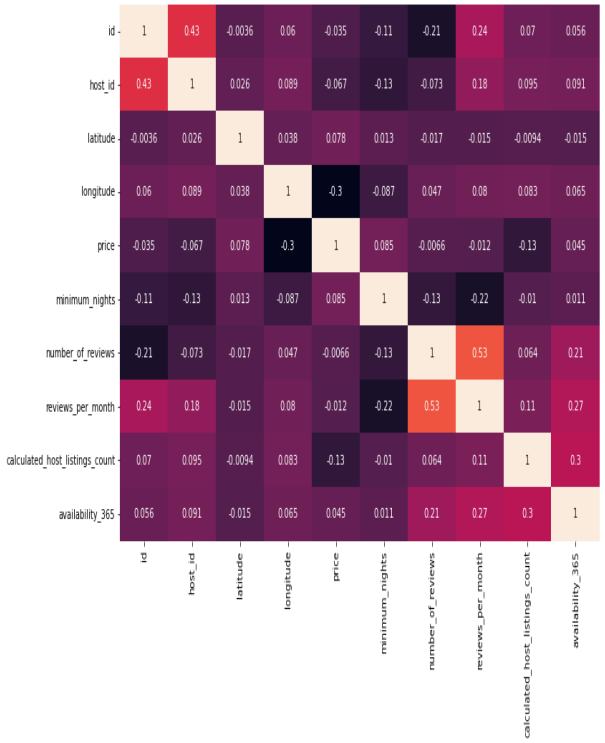
6.6. Availability of Room



From the above box plot Brooklyn and Manhattan region is the most busy region as there is very less availability of room while in Staten Island there is more than average 200 days rooms are available as

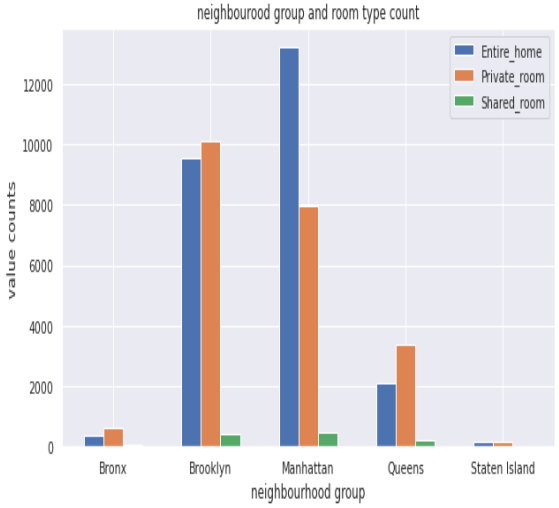
people are less preferred than other regions.

6.7. Correlational matrix:



Correlational matrix shows the correlation coefficients for different variables of the dataset.

6.8.Neighbourhood group and room type count



From the above bar graph we can say that the maximum room type count is of Entire Home or apartment. Also maximum availability is in Manhattan while very less in the Staten Island region.

7.Conclusion:

Simply by performing EDA on the dataset,we've identified various new in sight on how the Airbnb listings distribute d on New York, we know where are the list ings located, found out the Manhattan is dominating the listing number and have the highest price range, how the listing pri ce might be related to the number of booki ngs and room availability.

Also we have found that people mostly choose the room where they are available at cheap rate and giving them more privacy and roaming area as they prefer Entire home or apartment for rental purpose than shared ones.

1. The people who prefer to stay in Entire home or Apartment they are going to stay bit longer in that particular Neighbourhood only.

2. The people who prefer to stay in Private room they won't stay longer as compared to Home or Apartment.

3. Most people prefer to pay less price.

4. If there are more number of Reviews for particular Neighbourhood group that mea ns that place is a tourist place.

8. References

To prepare this report we use some directive notes, reports, and web pages that are listed below:

- Lecture Notes
- Kaggle Data Set

The Extra Notebooks in Kaggle

- Exploratory Data Analysis(EDA) of NYC Airb
- NYC Airbnb EDA
- Analytics Edge