EDA – Capstone Project

On

Airbnb Booking Analysis

Created By-

Raja Chowdhury

Airbnb

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities, based in San Francisco, California, New York City and the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties, instead it earn profits by receiving commission from each booking. In general, Airbnb is cheaper than hotels because Airbnb does not have to pay for the overhead costs of a hotel or the general management of such a large operation.

Points of Discussion

1. Problem Statements
2. Understanding Dataset
3. Data Preparation
4. Exploratory Data Analysis
5. Key Findings & Conclusions

Problem Defining Statement

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.
This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Explore and analyze the data to discover key understandings (not limited to these) such as :

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

Understanding The Dataset

As the objective is clear the data needs to be analyzed and this process starts with understanding our datas et of Airbnb NYC(New York City) and we come to understand the following things:

Our dataset contains 48895 rows and 16 variables as follows:

1. **id**- refers to the identity number of the property listed by a particular host.

2. **name**- refers to name of the listed property in Airbnb.

3. **host_id**- refers to the identity number of the host in Airbnb.

4. **host_name**- refers to name of the hosts, who listed their properties.

5. **room_type**- represents the various types of room in the listed property.
   - Entire home/ap
   - Private room
   - Shared room

6. **price**- refers to the cost of the room per night in USD.

7. **number_of_reviews**- refers to the number of customers reviewed the property.

8. **last_review**- refers to the date when the listed property was last reviewed.

9. **reviews_per month**- refers o the count of reviews per month the property received.

10. **longitude**- these represents the longitude coordinates of the listed property.

11. **latitude**- these represents the latitude coordinates of the listed property.

12. **neighbourhood_group**- refers to the names of the neighborhood groups present in the NYC.
    1. Manhattan
    2. Brooklyn
    3. Queens
    4. Bronx
    5. Staten Island

13. **neighbourhood**- refers to the neighborhoods present in Airbnb NYC.

14. **availability_365-** refers to the availability of the listed property out of the total 365 days in a year.

15. **minimum_nights**- refers to the minimum number of nights stayed by the customer.

16. **calculated_host_listings_count-** refers to the number of properties listed under a particular host.

Data Preparation

Data processing requires cleaning of data and preparing it for further analysis. Our cleaning process involved the following parts:

1. We observed that there were some null values present in few of our columns:

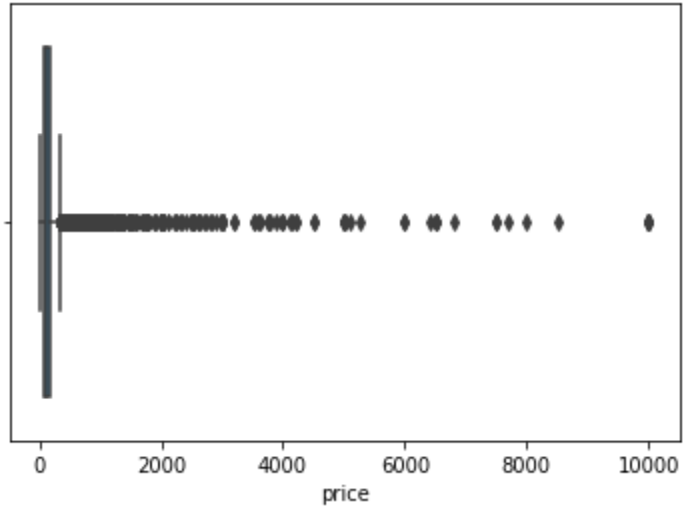
- **last_review** : last_review column has more than 20% of the null values and this is quite irrelevant column for our analysis so we have simply dropped this column.

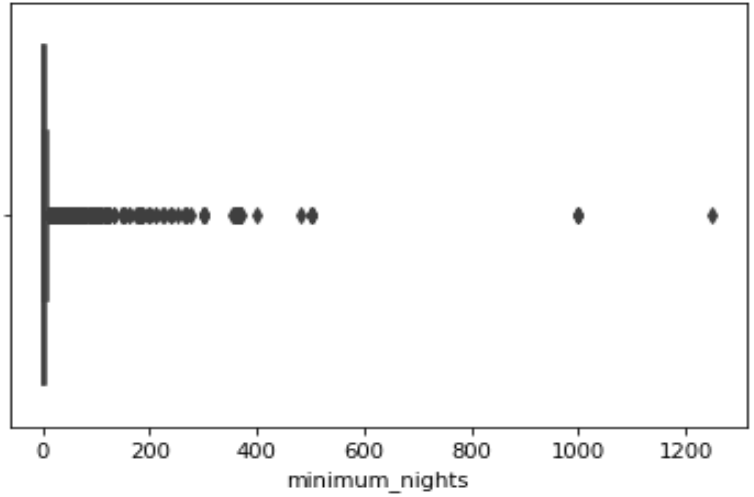
- **host_name** & **name** : Missing values are 21 & 16 in host_name and name columns respectively and both are less than 0.5% so we have simply dropped the rows corresponding to the missing values as they are very less in numbers and that won't affect our visualization.

- **reviews_per_month** : We also found some null values in reviews_per_month column and from analysis point of view it is an important column, so we have replaced all the null values of review_per_month column with zero ('0').

2. We have also observed that our dataset have 11 entries with zero price values so we have simply dropped those rows for better analytical result.

3.Two columns, price & minimum_nights are having outliers confirmed using Boxplot chart so we added two more column for z-score “z_price & z_min_nights” then we filter our data set by putting condition and removed the outliers.





Exploratory Data Analysis

Basically we have two categorical columns in our Airbnb dataset and on the basis of which we are going to do the further visualizations. So lets start our visualization with those data.

Room Type

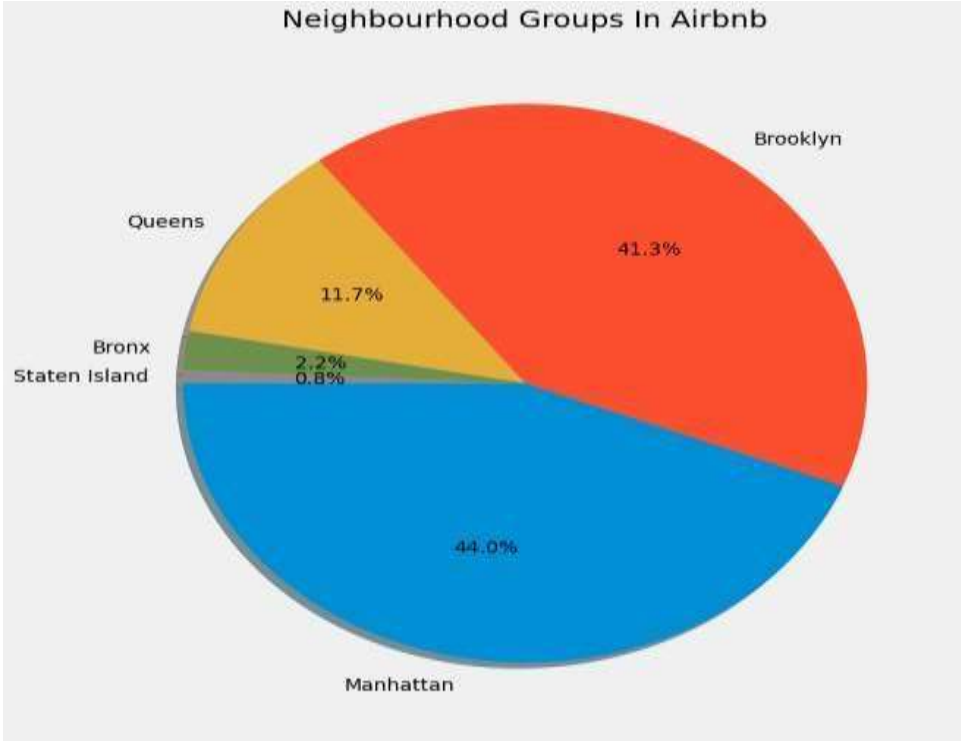
- Entire home/apt
- Private room
- Shared room

Neighbourhood Group

- Manhattan
- Brooklyn
- Queens
- Bronx
- Staten Island

Neighbourhood Groups

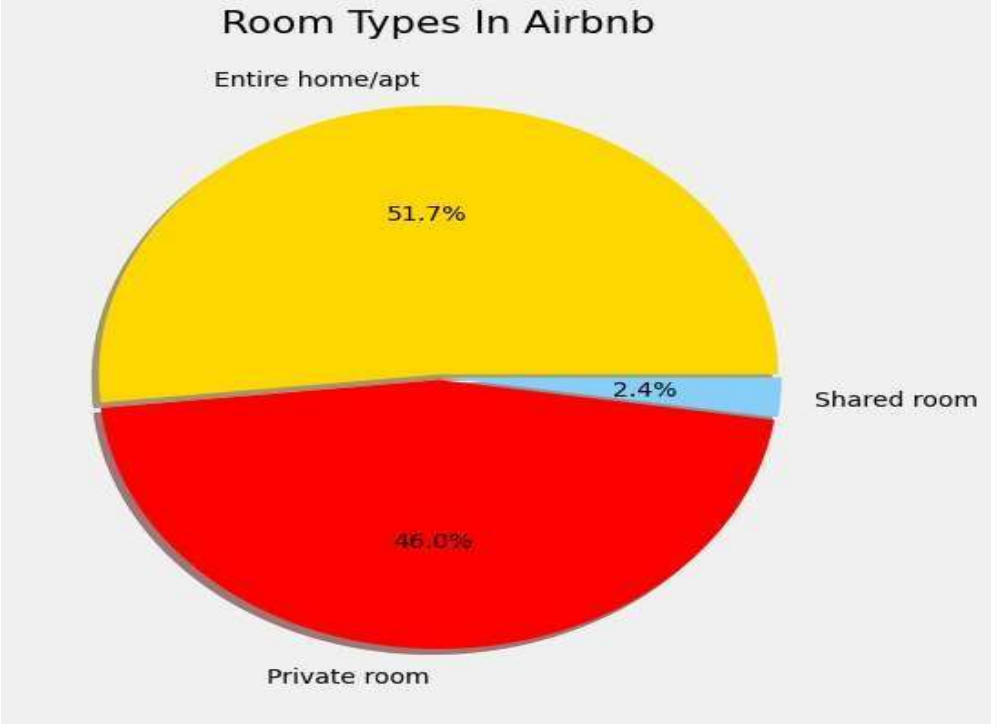
Here Manhattan has highest number of booking followed by Brooklyn and they together holds more than 85% of the total booking while Staten Island holds the least percentage of booking in the neighborhood group.



EDA

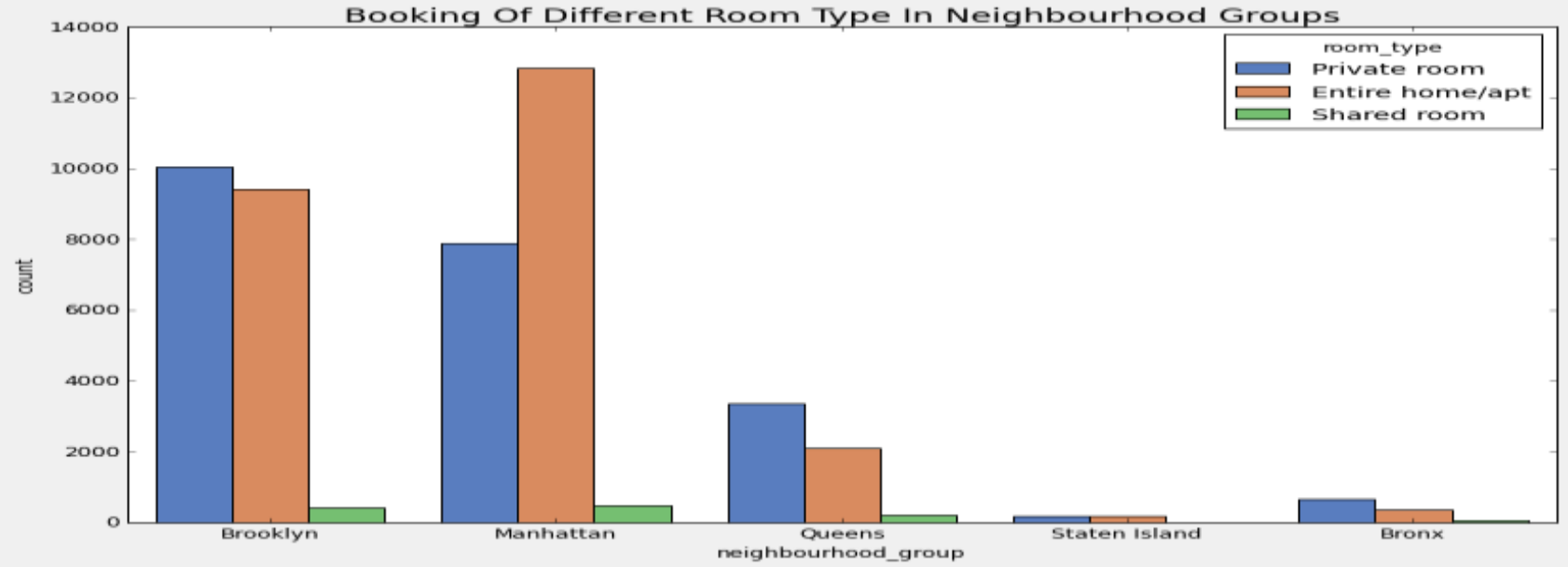
Room Types

Here Entire home/apt holds the highest number of booking followed by Private room and they together holds around 97% of the total booking while Shared room holds the least percentage of booking.



EDA

Room Type Demand In Neighbourhood Groups

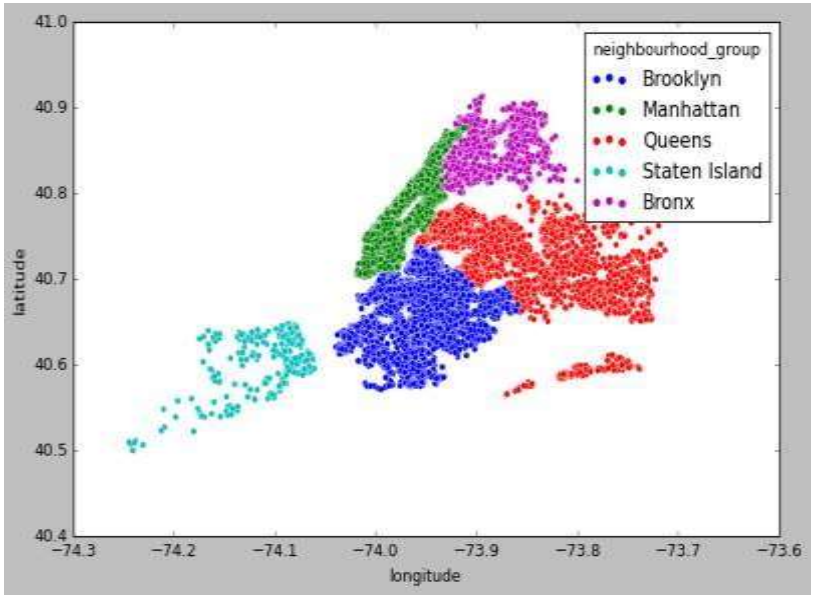


- In almost all neighborhood groups, private room has been more preferred followed by Entire Home/apt.
- Except in Manhattan, where Entire Home/Apt is more preferred than Private rooms.

EDA

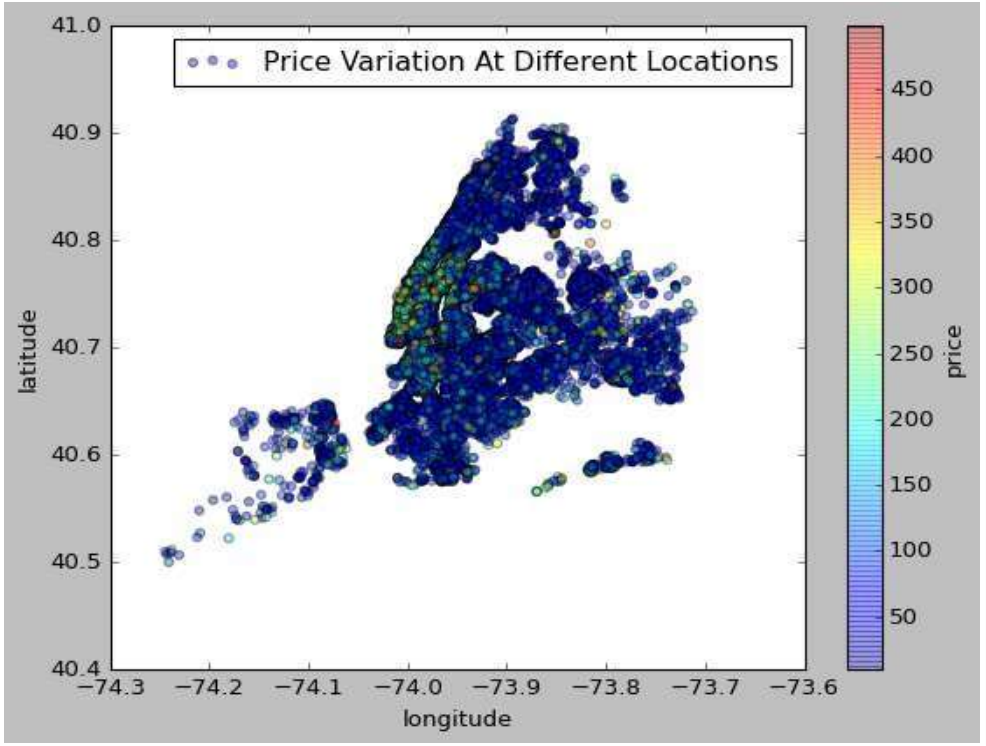
Hosts and Areas

Here we can conclude that Manhattan is the most densely populated neighborhood group by hosts and the reason for it can be the presence of Manhattan Island, bounded by the Hudson, East and Harlem rivers, and other tourist places and therefore it has highest number of host within lesser area space than any other neighborhood group.



Location and Price

Here we can see that area around Manhattan has higher prices and the reason for it can be the presence of Manhattan Island, bounded by the Hudson, East and Harlem rivers, and other tourist places as we have seen in the previous graphs.



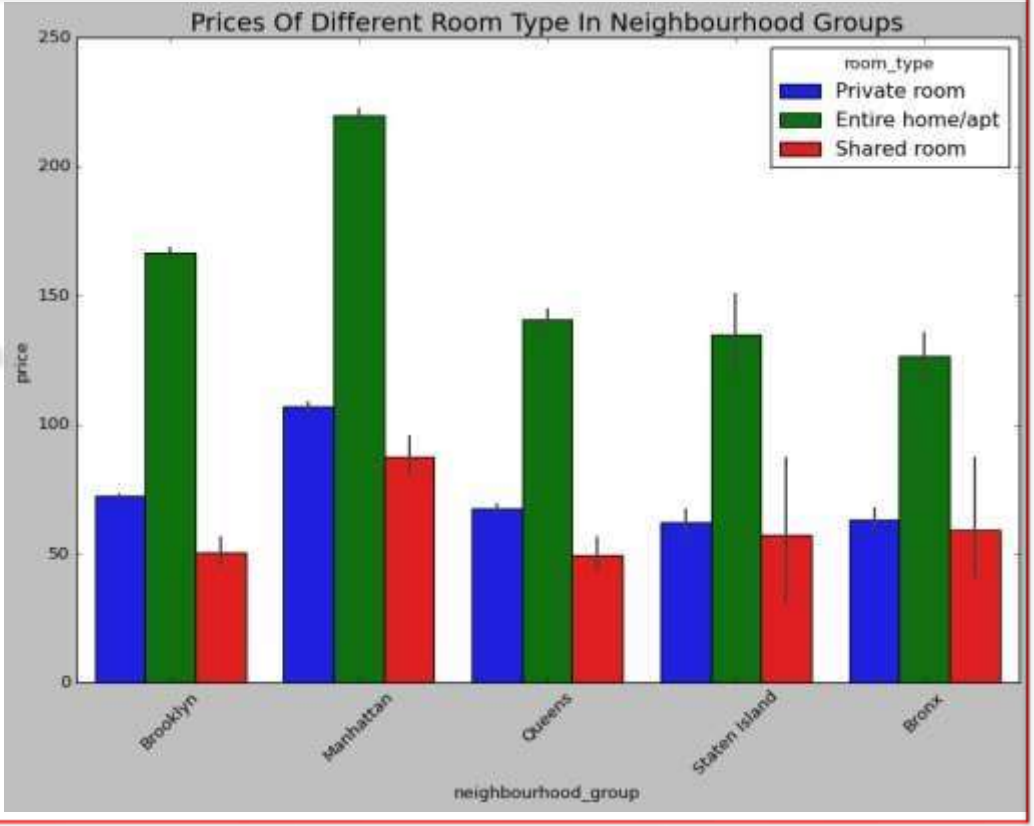
EDA

Prices In Neighbourhood Groups

- Firstly, price of the Entire Home/Apt is very high in all Neighborhood Group followed by Private Room.

- Also, the price in Manhattan is relatively high in all its room type followed by Brooklyn and almost similar for Queens, Staten Island and Bronx.

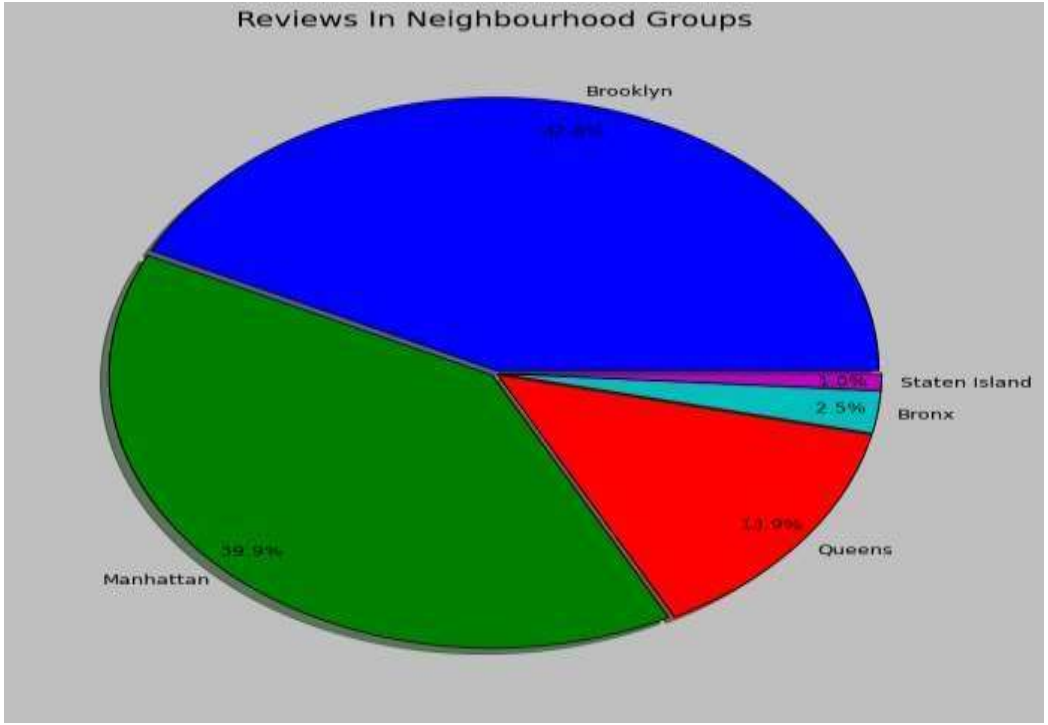
- We can think of as demand for Entire home /Apt is higher specially in Manhattan and Brooklyn thus push the price at higher side.



EDA

Reviews in Neighborhood Groups

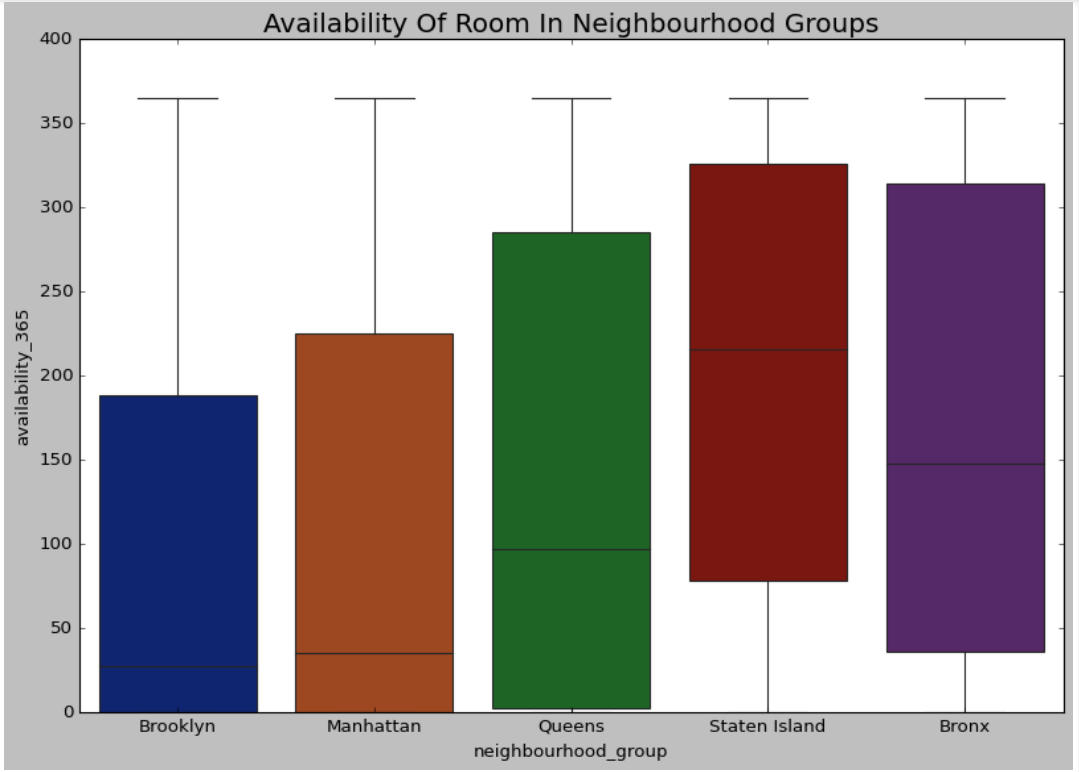
Here Brooklyn got the maximum percentage of reviews followed by Manhattan despite of Manhattan having the highest number of booking and they together holds the 82% of the reviews. While Staten Island has the least percentage of total reviews.



EDA

Availability Of Room In Neighborhood Groups

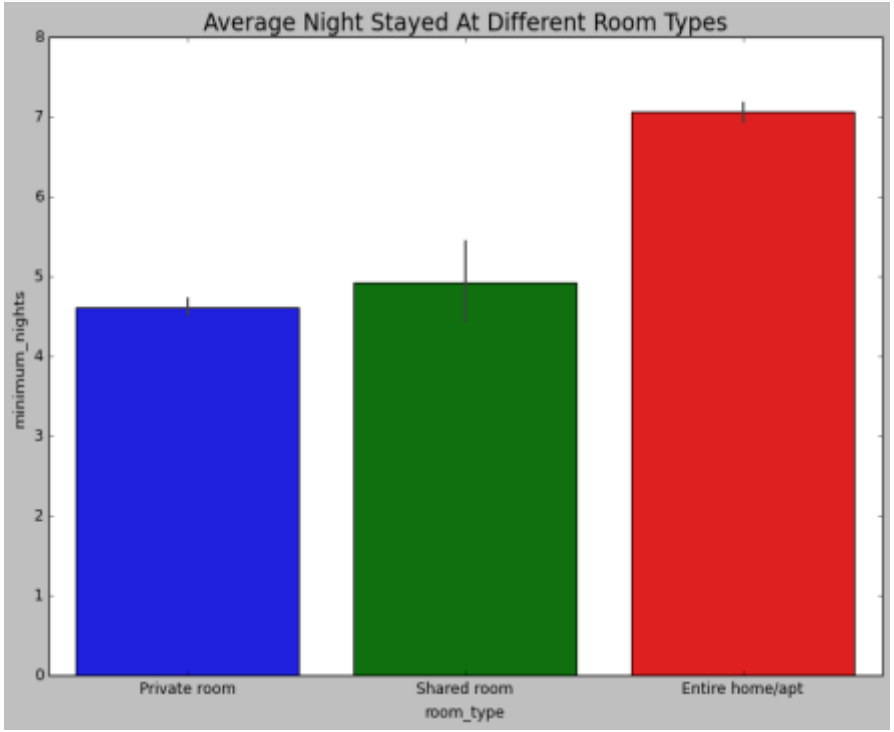
We have already seen that Brooklyn and Manhattan have higher number of booking compared to the other neighborhood groups therefore they both have less availability of rooms. Similarly Staten Island and Bronx have least number of booking therefore they have more availability of rooms compare to others.



EDA

Average Night Stay In Room Types

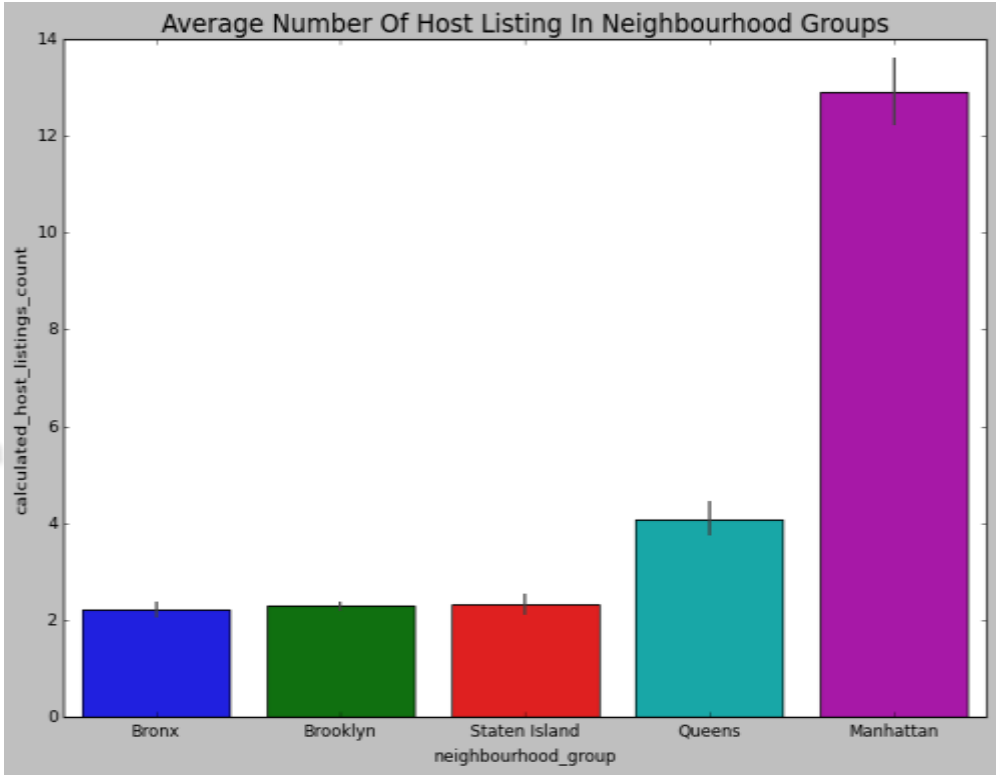
We can see that Entire Home/Apartment has higher average night stay and reason for it, we can assume is that mostly Entire Home/Apt are booked for family vacations, group tours or for events so they tends to stay longer. While Private room and Shared room are most likely to be booked for official purpose or for limited time stay.



EDA

Host Listing In Neighborhood Groups

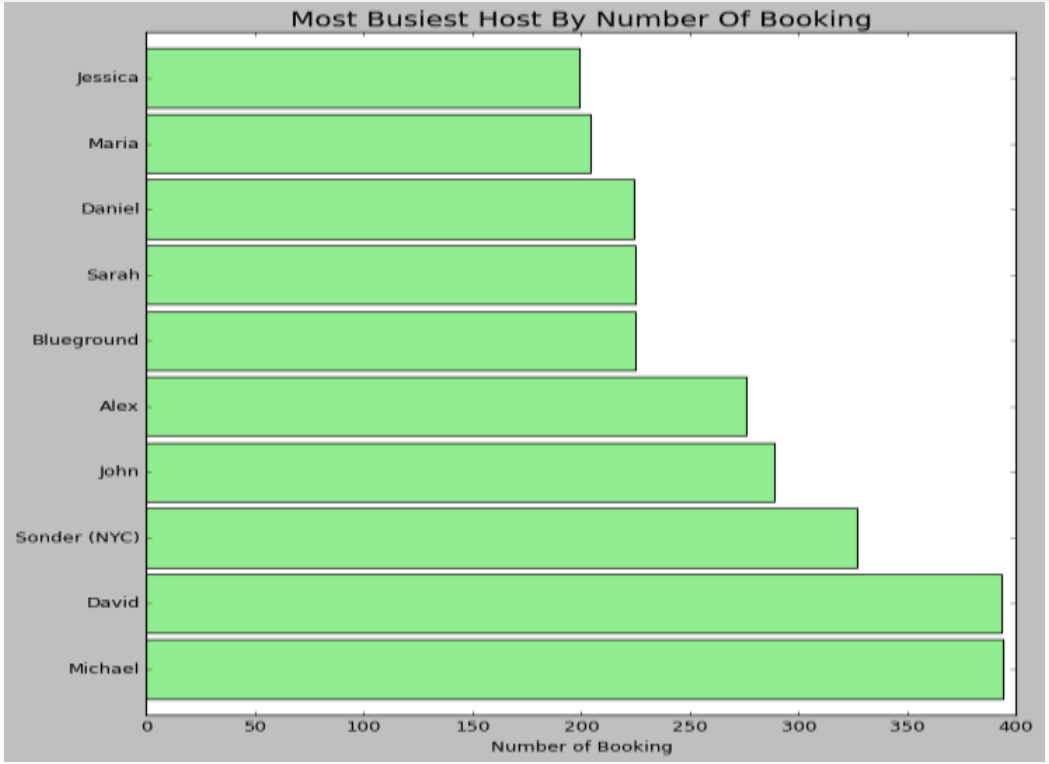
Here we can conclude that the hosts of Manhattan neighborhood group have high number of properties listed in Airbnb and this could be because Manhattan has higher demand as well as higher price compared to any other neighborhood groups location which encourage hosts to list more of his properties.



EDA

Most Busiest Host

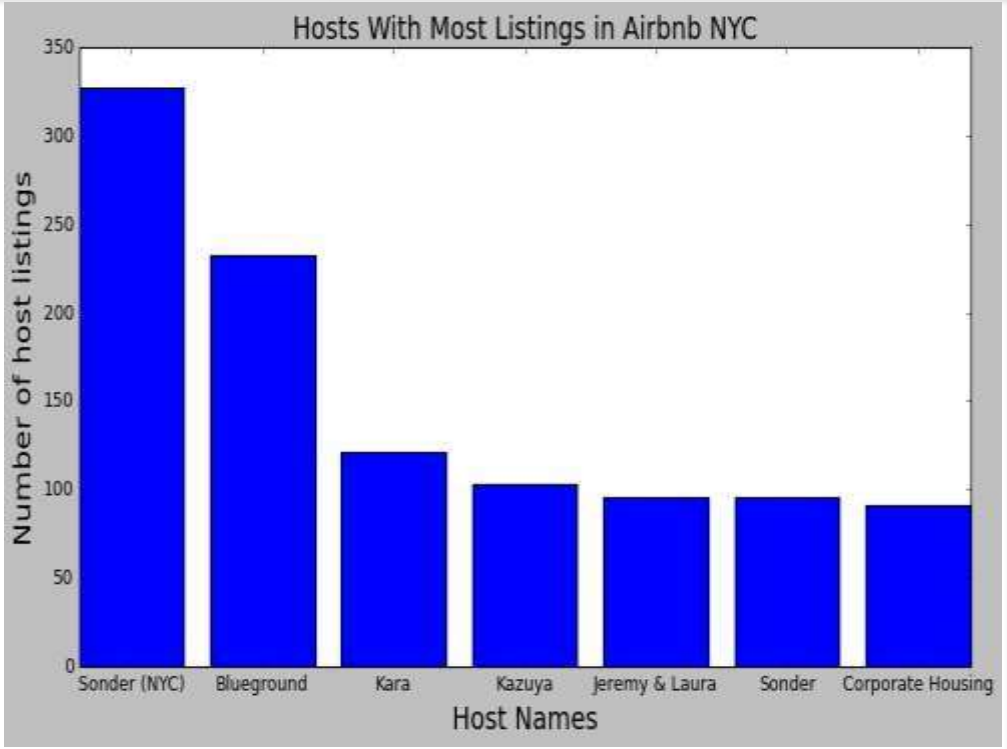
Here we can see that **Michael** is the most busiest host in the Airbnb NYC dataset as it has most number of bookings. The reason for it can be its location as it is situated in Manhattan neighborhood group and has Entire Home/apartment where demand are at peak with higher price that we have already seen in the previous visualizations.



EDA

Host With Most Listing

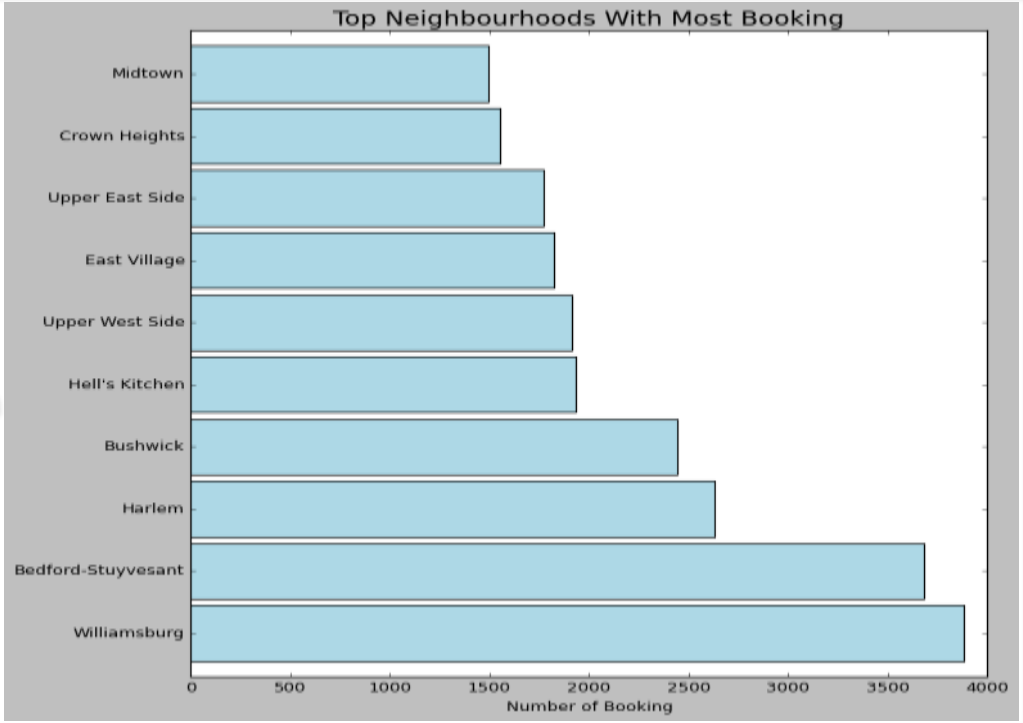
Here we can see that **Sonder (NYC)** is the host with most number of listed property in Airbnb NYC. Sonder (NYC) belongs from Manhattan neighborhood group and most of its listing have Entire home/apartment.



EDA

Most Demanded Neighborhoods

Here we can see some top demanded neighbourhoods, where **Williamsburg** is the neighbourhood with most number of bookings in our Airbnb dataset and seven out of ten top neighbourhoods belongs from Manhattan neighbourhood group with Entire home/apartment room type.



Key Findings

- ✓ Manhattan and Brooklyn neighborhood groups hold around 85% of the total booking while Staten Island has the least number of bookings.
- ✓ Manhattan neighborhood hosts have higher number of properties listed in Airbnb.
- ✓ Entire home/apt is the most preferred room type followed by private room and they hold around 97% of the total booking.
- ✓ Queens has significantly less host listings than Manhattan. So, we should take enough steps to encourage host listings in Queens.
- ✓ Seven out of ten top neighborhood belongs from Manhattan neighborhood group with Entire home/apartment room type.

Conclusions

- We can conclude from the analysis that Manhattan is the top neighborhood group with the highest number of bookings and host listings. Seven out of ten top hosts are from Manhattan. One of the probable reasons for being the most preferred Neighborhood Group is that Manhattan is world-famous for its museums, stores, parks, music and cultures. It has high number of tourist places where Entire Home/Apartment are more preferred as stay options and also for longer periods. These factors increased the demand and led to higher prices compared to any other neighborhood groups.

- Brooklyn has significant number of bookings because Brooklyn also has some famous bridges, parks, museums, islands and other tourist places but with more affordable prices as compared to Manhattan. It also received the maximum number of reviews followed by Manhattan.

Conclusions (cont.)

- Rest 3 neighborhood groups namely Queens, Bronx and Staten Island are observing very less number of bookings and hosts, especially on Staten Island. Considering that these are less popular areas, it is possible that some guests choose these locations to save up money or for official purposes who want to stay for limited time periods. And these neighborhood groups have higher room availability than Brooklyn and Manhattan.

AI

THANK YOU