Capstone Project
Title: Airbnb Booking analysis

Presented by: Ghadage Manisha

Index

- Problem Statement
- Data Summary
- Data Pipeline
- Data Handling
- Exploratory Data Analysis(EDA)
- Conclusion

Problem Statement

Since 2008,guest and host have used Airbnb to expand on travelling possibilities and present amore unique, personalised way of experiencing the world. Today, Airbnb become one of kind service that is used and recognised by the whole world. Data analysis on millions of listings provided through Airbnb is crucial factor for the company. These millions of listings generate a lot of data that can be analysed and used for security, business decision, understanding of customers and providers(hosts) behaviour 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 numerical values.

Explore and analyse data to discover key understandings (not limited to these) such as

- What can we learn about different hosts and areas?
- What can we learn from predictions?(ex.location, price, 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?

Data Summary

The dataset has 16 different features with more than 49,000 observations. The most important features of given dataset are

Host id: It is id given to specific host and there are 1270 host id available in given dataset.

Neighbourhood group: It represent location, given data set contain 5 different locations

Neighbourhood: It represent specific areas where the listing is located.

Room type: It represent category of room type being listed.

Minimum nights: It represents number of nights spend by customer in given listing.

Number of reviews: It represents the number of reviews for listings.

Availability 365: It represents number of days in year for which given property is available for rent.

Price: It represent rate for given room type in given location for one night.

Data Pipeline:

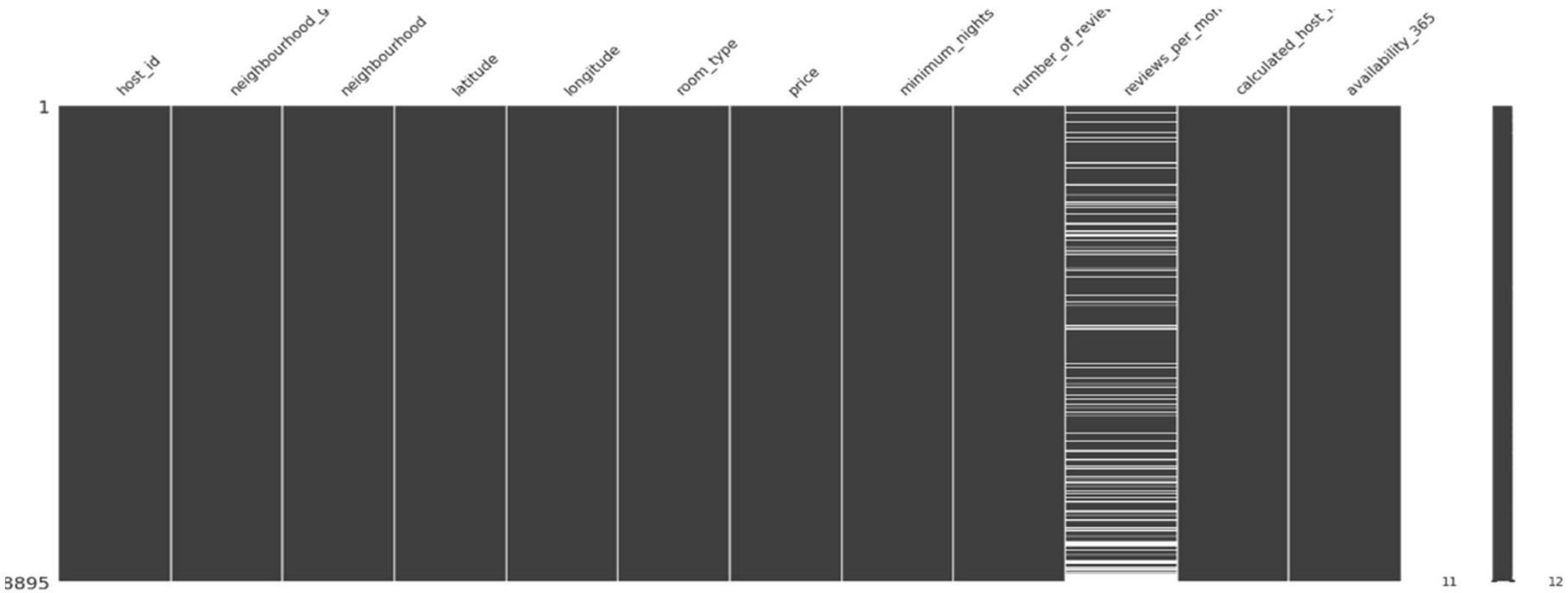
Data Processing: In this part I have checked the data and its all features.

Data Handling: In this part I have checked for Nan values, missing values and duplicate observations and done the refinement dataset.

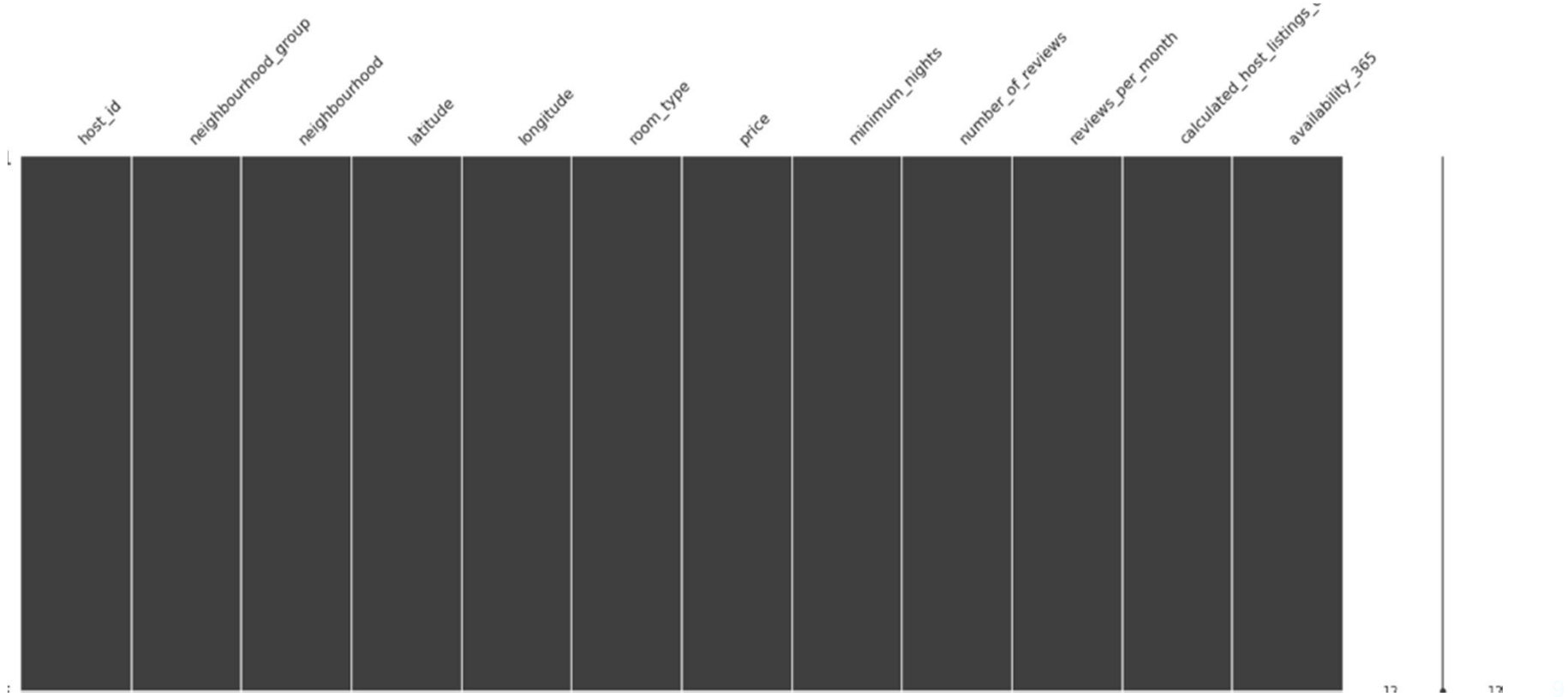
Exploratory Data Analysis: In this part I do some exploratory data analysis on selected important features.

Data Visualization: Finally, I visualize the data using distinct plots for distinct features of the data, try to analyse the relationship between the features of data with the help of different plots. I also cleared all the point that's why given thing happens.

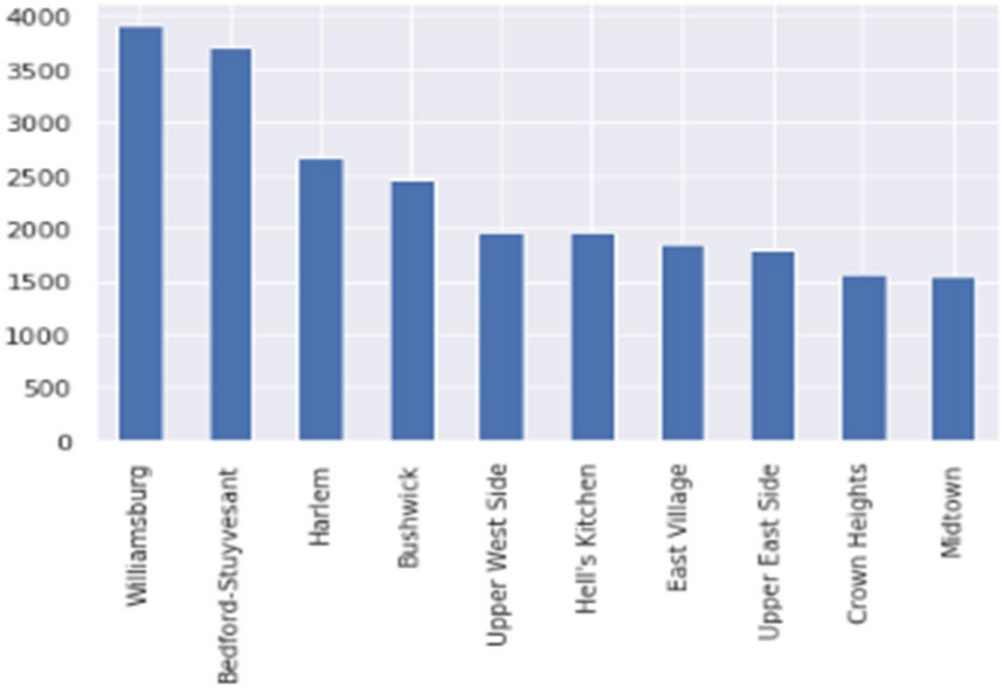
Data Handling (Missing, Nan values)



Data Handling (Missing, Nan Values)



EDA(Top 10 Listing Areas)

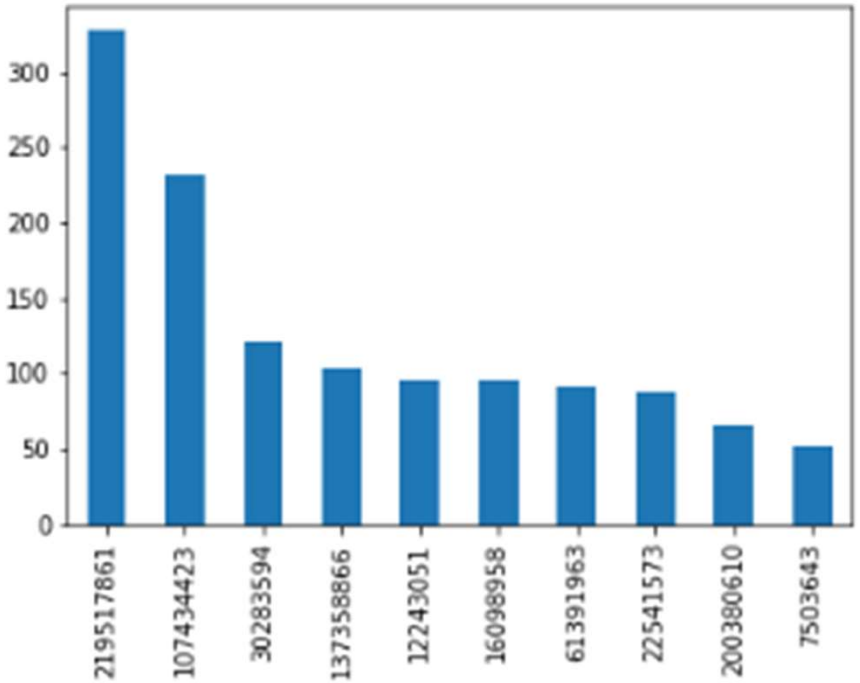


Observations:

Data set contains total 221 different areas .

Williamsburg area has maximum number of listings followed by Bedford-Stuyvesant.

EDA(Top 10 Hosts)

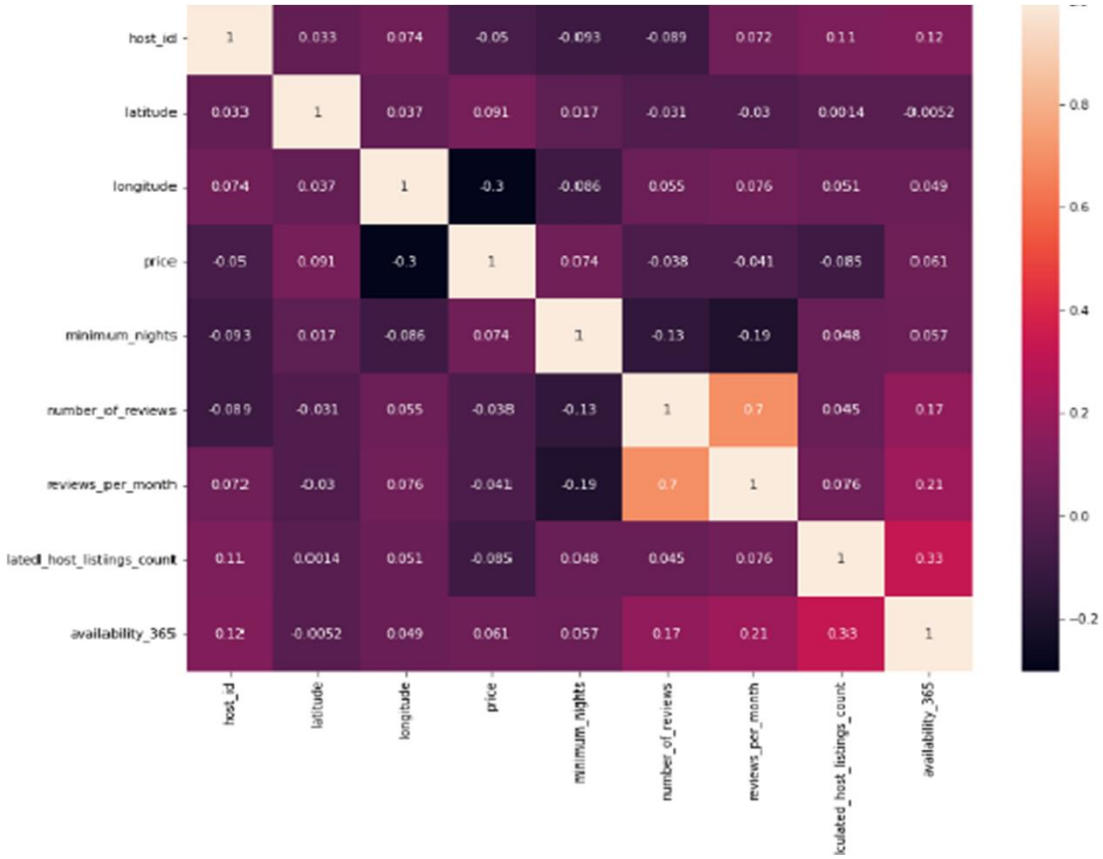


Observations:

In given dataset total 37457 hosts are available.

From bar plot it is observed that host with host-id 219517861 is top most host with 327 number of listings

EDA(Correlation)

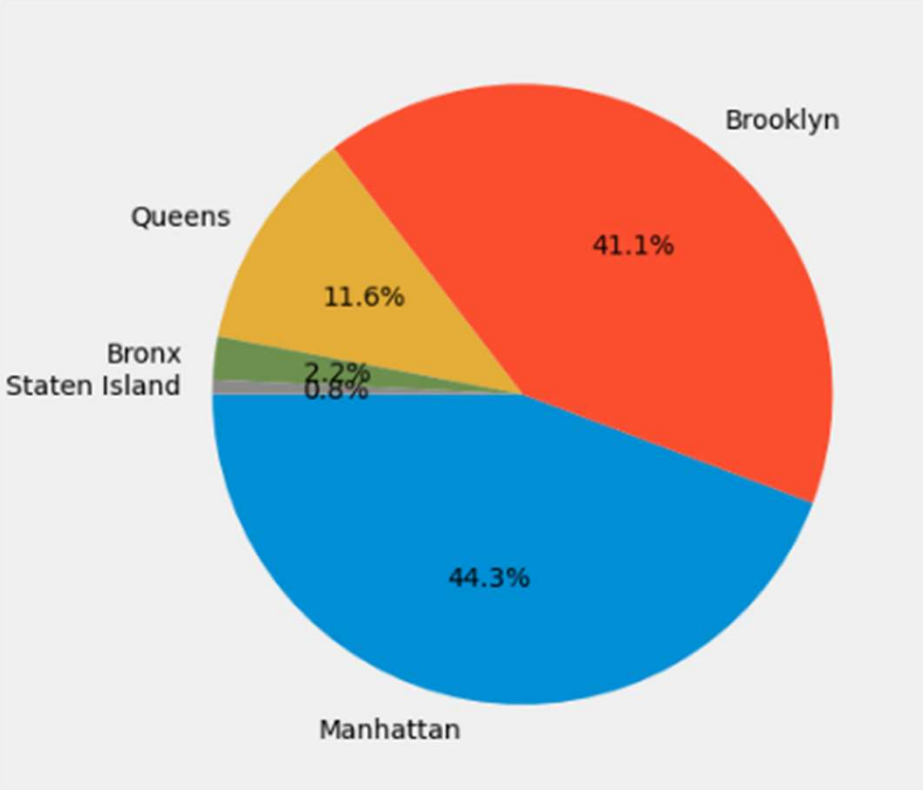


Observation:

No strong correlation is observed between given different features of dataset.

Number_of_reviews and availability-365 ,calculated_host_listings and availability_365 are weakly correlated

EDA(Top most location with maximum number of listings)

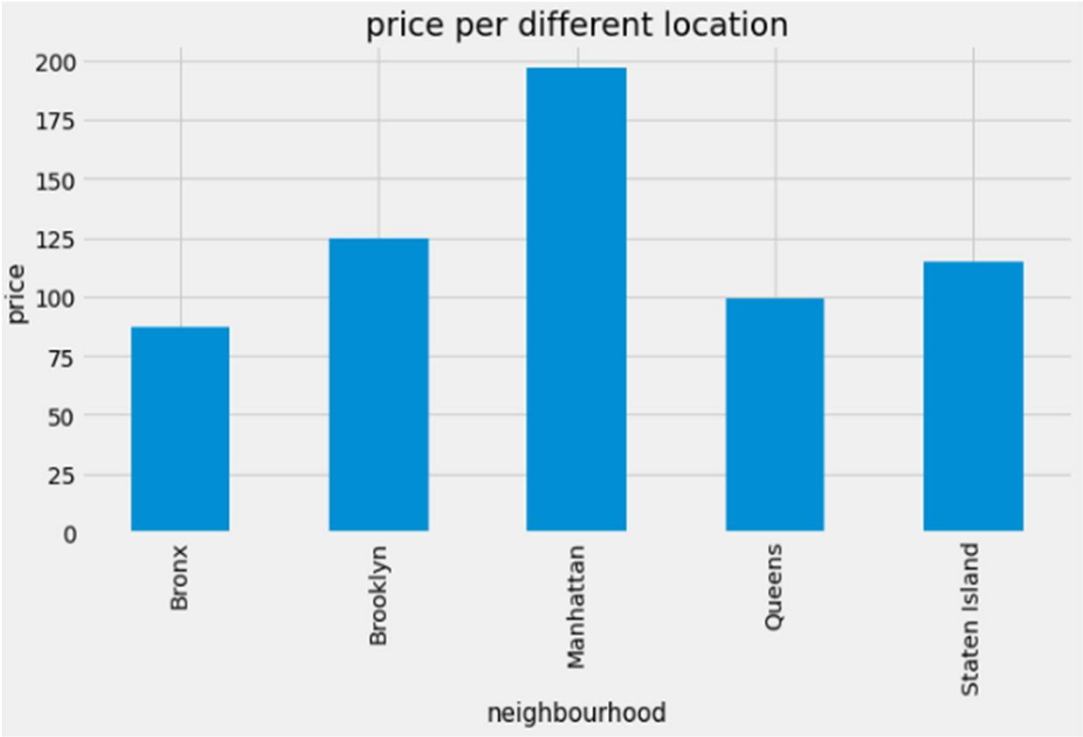


Observation:

Dataset contains 5 different location in new york city.

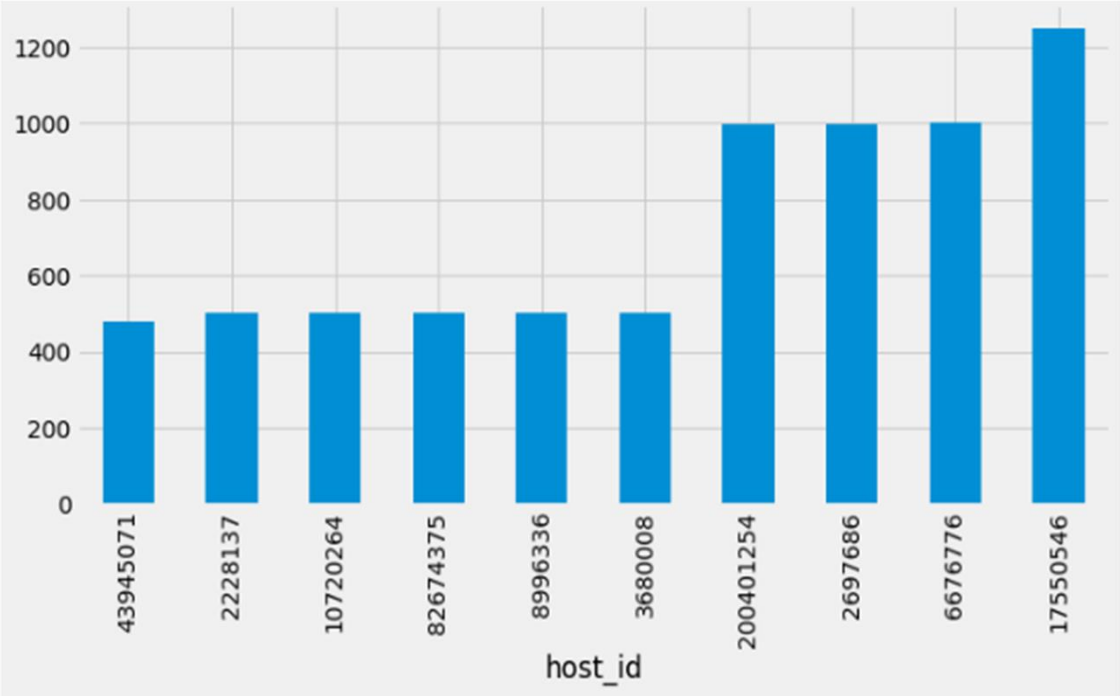
From pie chart it is observed that maximum number of listings are found in Manhattan(44.3%)

EDA(price per location)



Observation:
From bar chat it is observed that Manhattan is most expensive location followed by Brooklyn

EDA(Top 10 busy hosts)



Observation:

Host with host id 17550546 is busiest host among top 10 busy hosts.

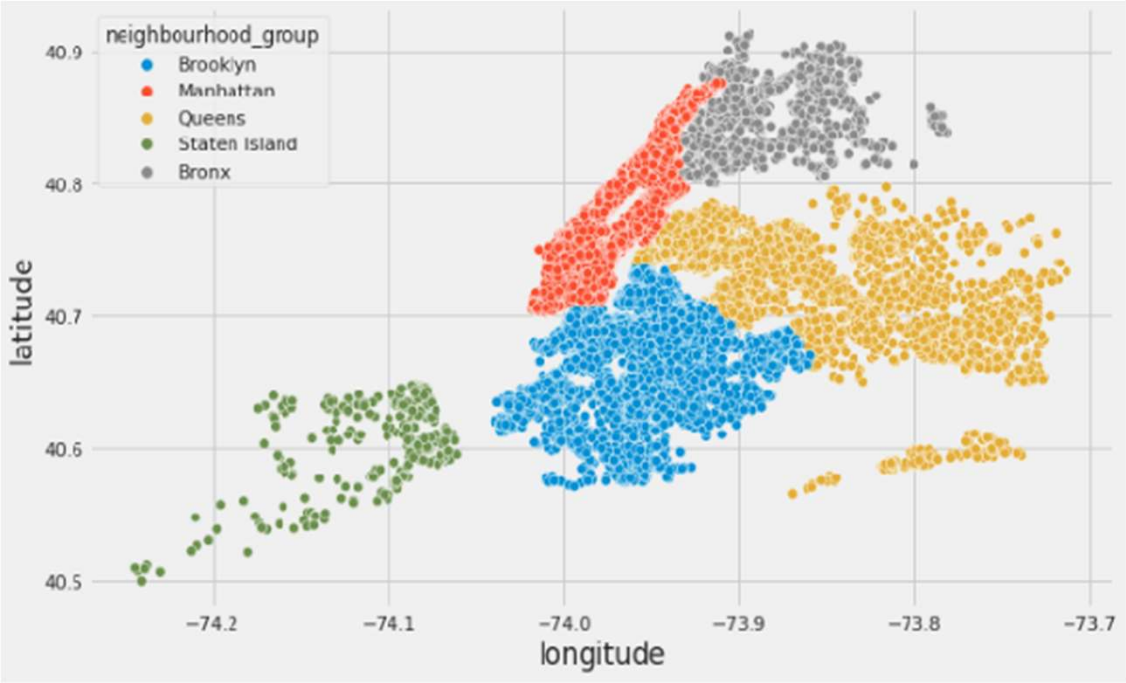
This is because number of minimum nights spend at listing belongs to host id 17550546.

EDA(Top Reviewed Listings)

neighbourhood_group	number_of_reviews
Queens	629
Manhattan	607
Manhattan	597
Manhattan	594
Queens	576
Queens	543
Manhattan	540
Queens	510
Brooklyn	488
Brooklyn	480

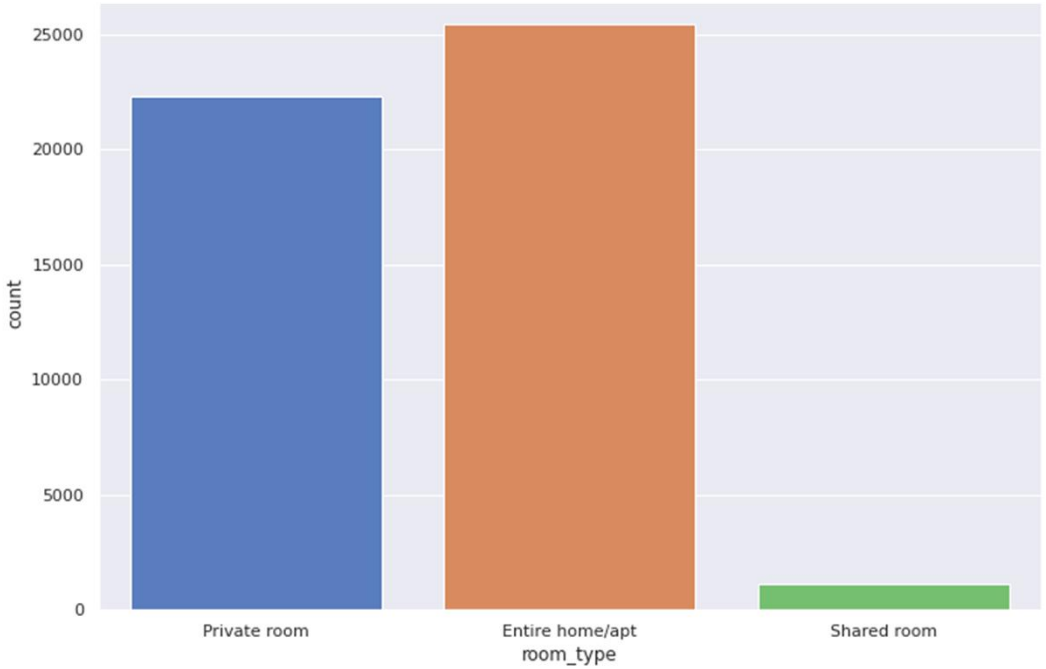
Observation:
From table it is observed that number of reviews are maximum for listings in Manhattan.

EDA(Noticeable difference in traffic among different areas)



Observation:
Majority of traffic(number of listings) are observed in Queens and Brooklyn followed by Manhattan.

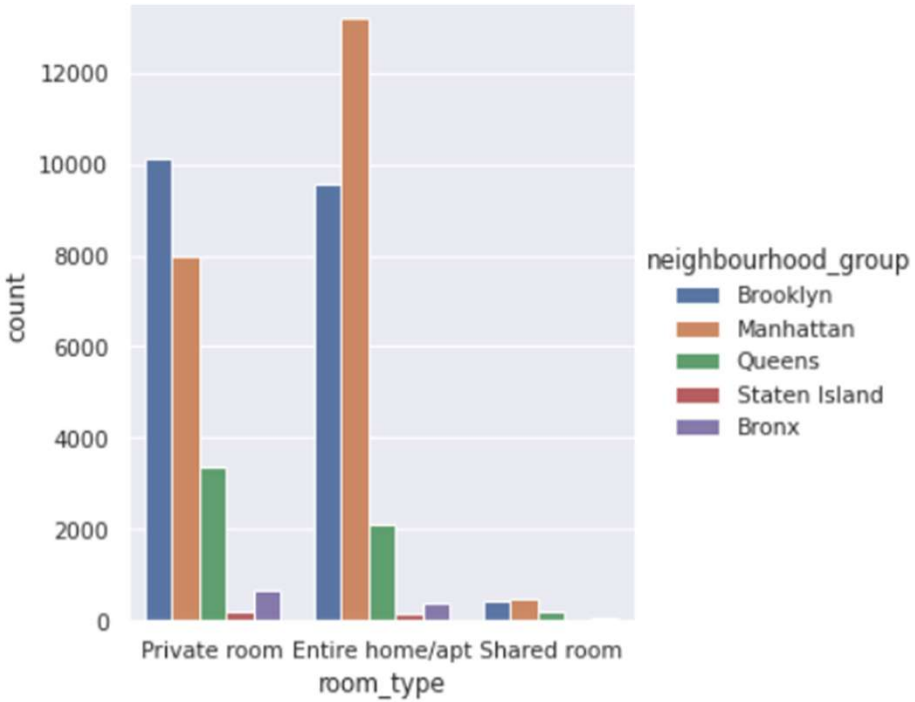
EDA(Room types available)



Observation:

From plot we observe that entire home / apartment has highest share, followed by private rooms, and least preferred is shared rooms

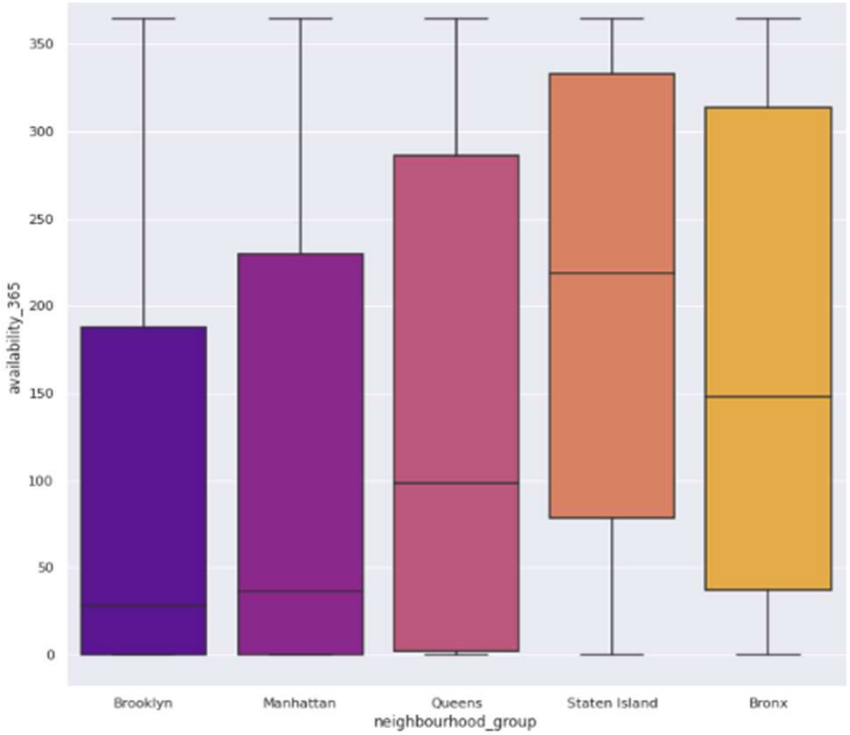
EDA(Room type available per location)



Observation:

Three different room types are available in given dataset. From plot it is observed that majority of entire home type is located in Manhattan and most of the private rooms are concentrated at Brooklyn

EDA(Availability_365 vs. neighbourhood group)



Observation:

The graph shows relationship between availability room and neighbourhood group

Conclusion:

The given dataset appear to be very rich dataset with a variety of columns that allowed us to do deep exploration on each significant column presented.
For given data set I found that there are total 221 different areas out of which ‘Williamsburg’ has maximum number of listings.
There are total 37457 hosts and host with host_id 219517861 is top host with 327 listings.
No strong correlation observed among price, reviews and location.
Out of 5 different locations in dataset Manhattan is most crowded location with 44.3%listings.
Host with host_id 17550546 is busiest host as listings where minimum nights spend are more belongs to him.

Major traffic of listing observed in ‘Queens, Brooklyn and Manhattan’ This is because availability of listing for year and number of reviews.
Data set contains thee different room types, it is observed that ..
Majority of entire homes are located in Manhattan
Majority of private rooms located in Brooklyn.

Future Scope:

This data analytics will be very useful for higher level on Airbnb Data/Machine Learning team for better business decision, control over the platform, Marketing initiatives, implementation of new feature and much more.

QUESTIONS
Q&A
ANSWERS

Thank You!!!