Capstone Project

Exploratory Data Analysis

Team Members

Meghna Goud
Neha Pasi

Introduction To Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing intial investigation on the data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is good practice to understand the data first and try to gather as many insights from it.

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, the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking.

The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk and joe Gebbia. Airbnb is shortened version of its original name, AirBedandBreakfast.com

Imported libraries

Steps Involved :

*Data Understanding :

*Data Cleaning:

1. Import important libraries.
2. Observe and understand each variable in columns.
3.checking for null, missing, dublicate values.
4. Change the data type.
5. Check value counts for a specific column.
6. Check correlation between variables.
7.checking for outliers.
8.Dropping unnecessary columns.

*Discover the patterns by visualising data.

Data Summary

Numeric – 1. price
2. number_of_reviews
3. minimum_nights
4. reviews_per_month
5. calculated_host_listings
6. avaliability_365
7. lattitude
8. longitude

Categorical – 1. room_type
2. neighbourhood_group
3. neighbourhood

Unique - 1. id
3. host id

String – 1.name
2.hostname

Data Summary

This dataset has 48895 observations in it with 16 columns and it is a mix between categorical and numeric values.

1. id : Column id is a unique column in the dataset
2. name : This column contains the name of the listing.
3. host_id : This column contains the host IDs of the various hosts. Each host has a unique host ID.
4. host_name : This column contains the name of the hosts for a listing.
5. neighbourhood_group : It is an categorical column containg different neighbourhood groups.
6. neighbourhood : It is an categorical column containg the various neighbourhoods of a listing.
7. lattitude : It is an numerical column containg the latitude of the geographical location of the listing.
8. longitude : It is an numerical column containg the longitude of the geographical location of the listing.
9. room_type : It is an categorical column containg different room types.

Data Summary

10. price : This column contains the price of the listings.
11. minimum_nights : It contains the minimum number of nights spend by tourists in a listing.
12. number_of_reviews : This column shows how many reviews are there for a particular listings.
13. last_review : This column contains the last date when the listing was reviewed.
14. reviews_per_month : This column contains the number of reviews for a particular listing in a month.
15. calculated_host_listings_count : This column shows number of listings of a particular host.
16. avalaibility_365 : This column shows the avalaibilty of a listing on yearly basis.

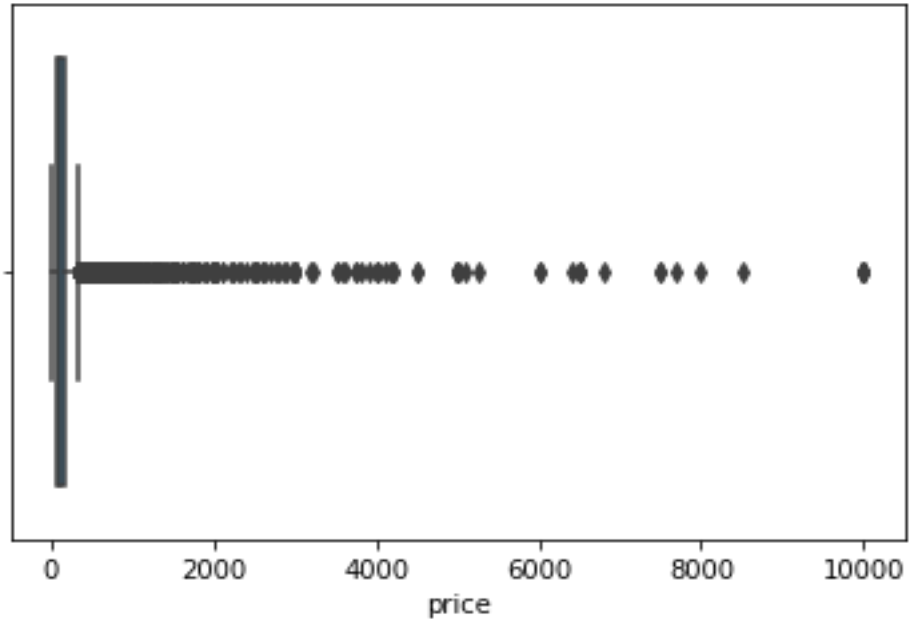
Futuristic Features

count_host_id – Top 10 host count according to their host id.
count_neighbourhood – Count of highest listing in neighbourhood.
average_price_df – Average price based on location and room type.
min_and_max_df - Minimum and maximum price in different neighbourhood according to room types.
count_room_type – Total count of each room type according to neighbourhood group.
areas_reviews - listing count of neighbourhood group according to number of reviews.
avalaibilty_type - Distribution of neighbourhood group, room types listings with avalaibilty of rooms.
avalaibilty_df - relationship between neighbourhood group and avalaibilty of rooms.

EDA:

Box Plot of “Price” column:

we can observe that the minimum price is zero that is unjustified.



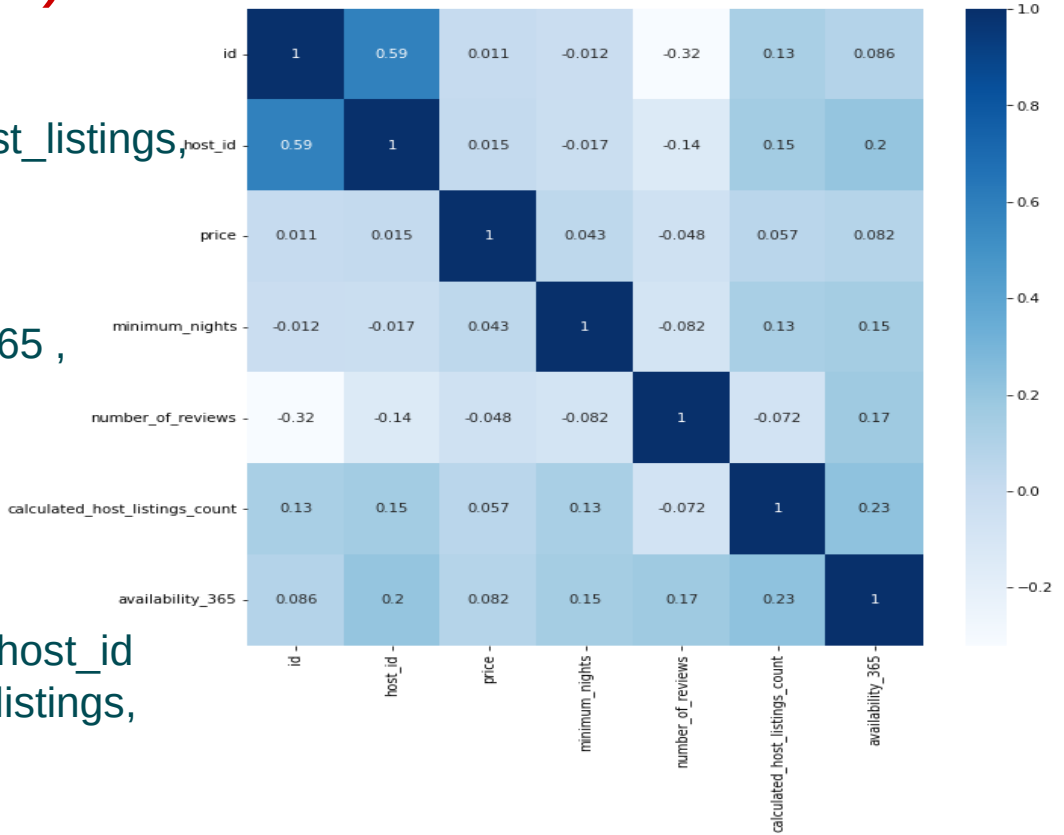
EDA : (correlation matrix)

Positive Correlation

1.Price - avaliability_365 , calculated_host_listings, minimum_nights , host_id
minimum_nights - avaliability_365 , calculated_host_listings , price
calculated_host_listings - avaliability_365 , minimum_nights , price , host_id

Negative Correlation

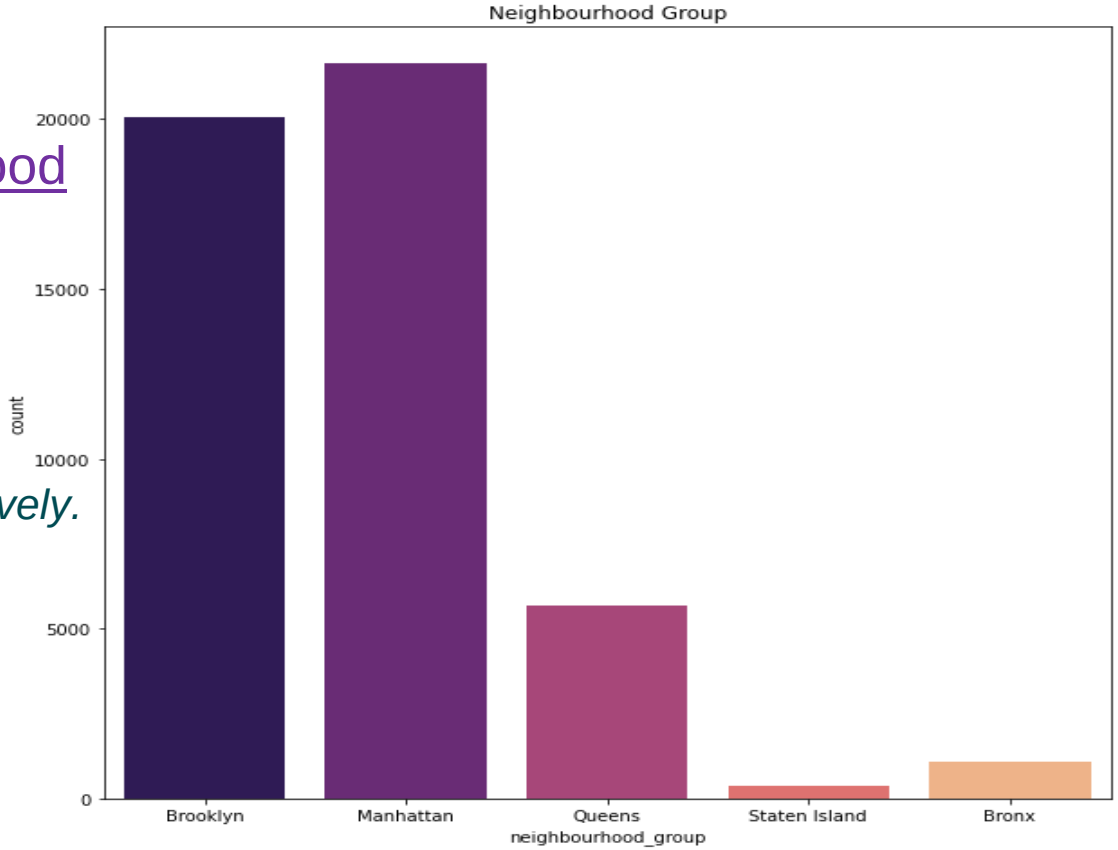
1.price - number_of_reviews
2.minimum_nights - number_of_reviews,host_id
3.number_of_reviews - calculated_host_listings, minimum_nights , host_id



EDA:

Count Plot of Neighbourhood group:

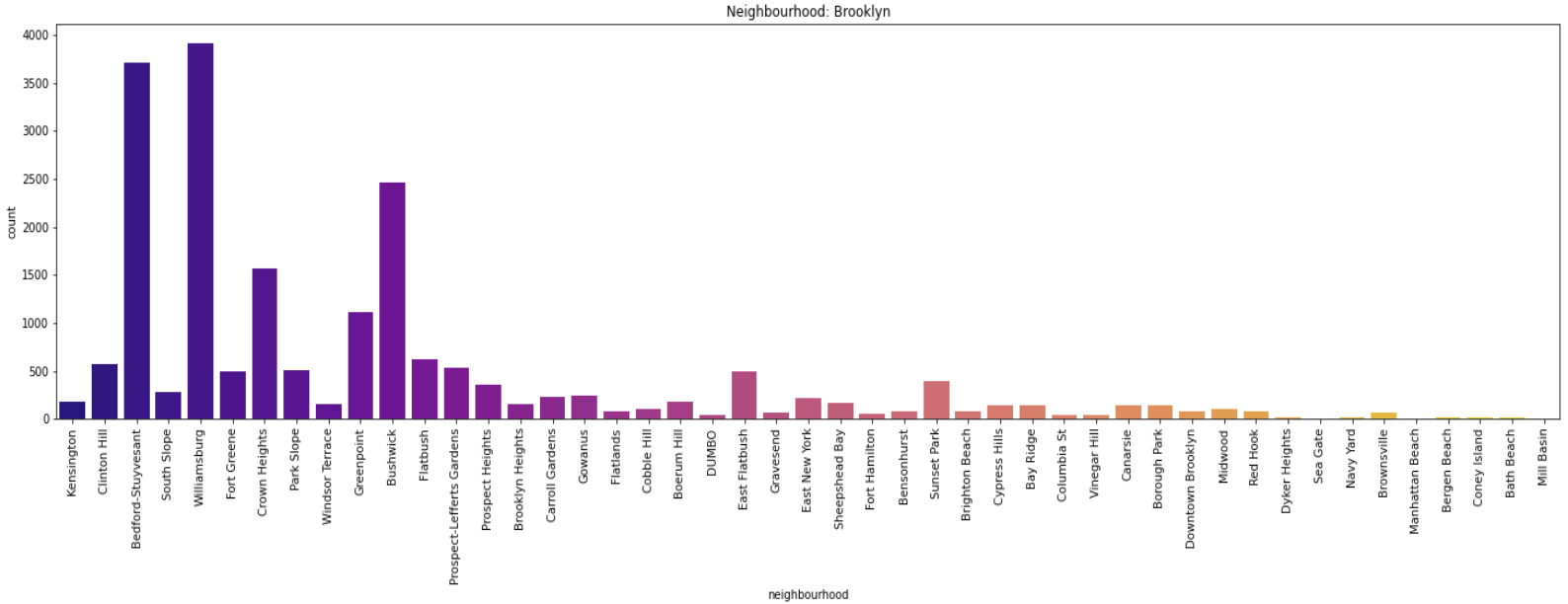
most of the properties are located in Manhattan and Brooklyn followed by Queens, Bronx and Staten Island respectively.



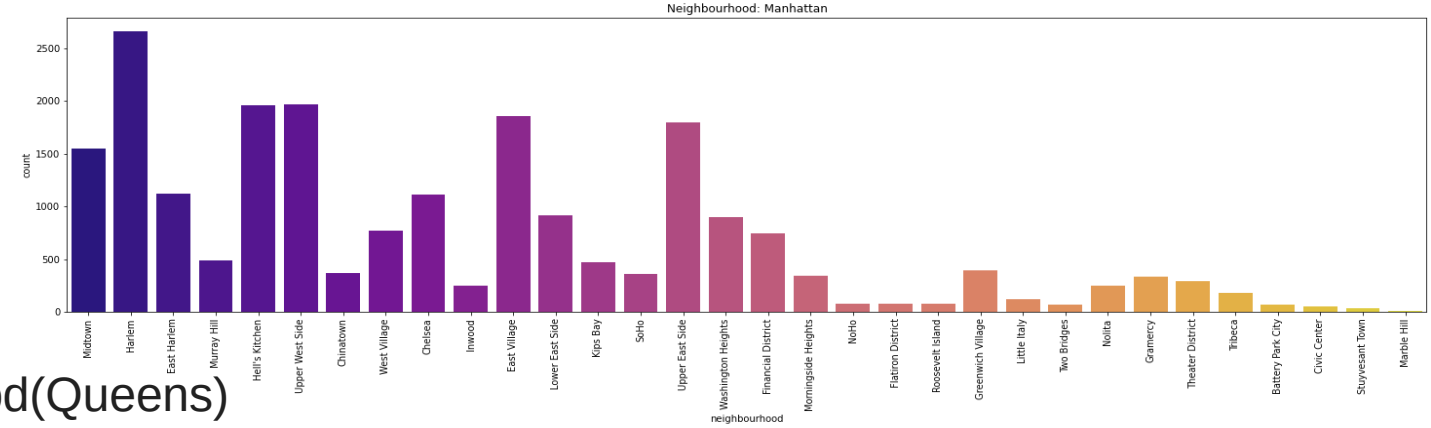
EDA:

Neighbourhood count plots ,each plot was created based on their respective neighbourhood_group category

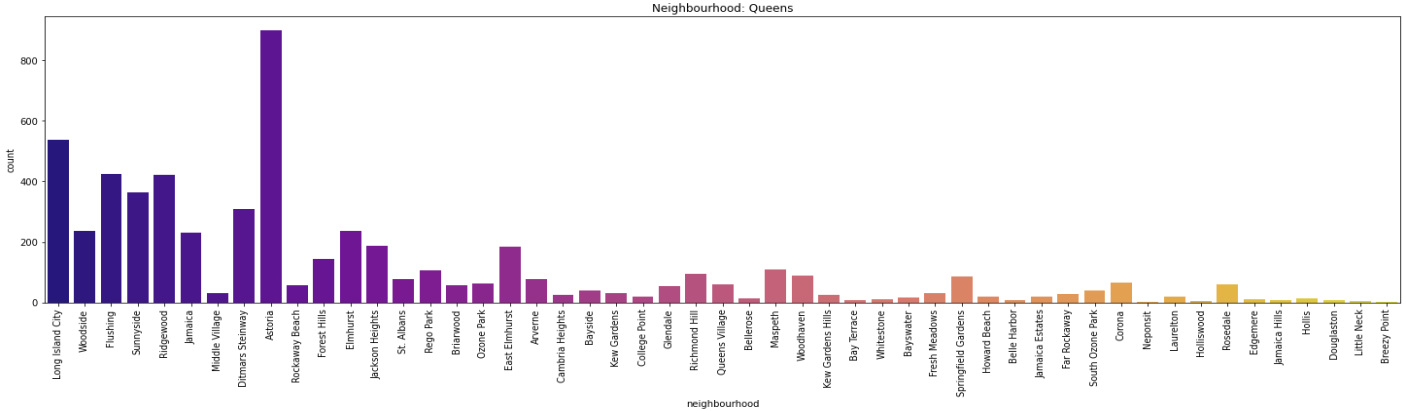
1. Neighbourhood(Brooklyn)



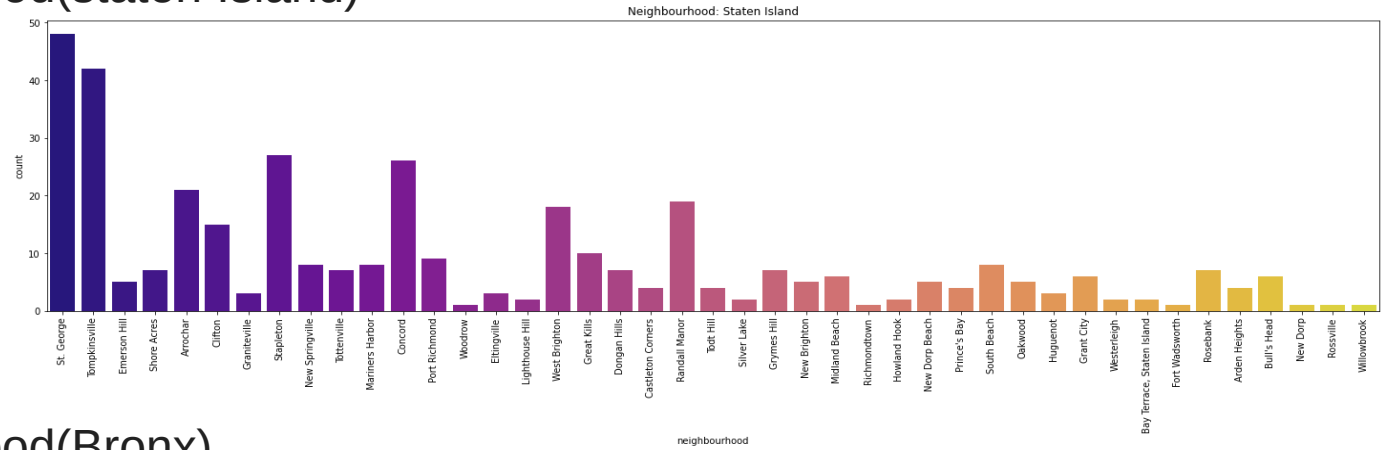
2. Neighbourhood(Manhattan)



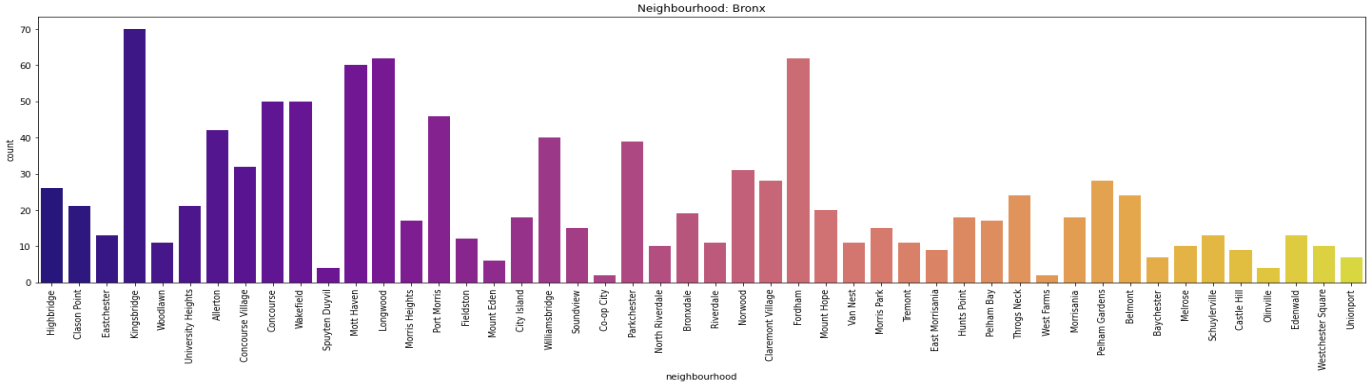
3.Neighbourhood(Queens)



4.Neighbourhood(staten Island)



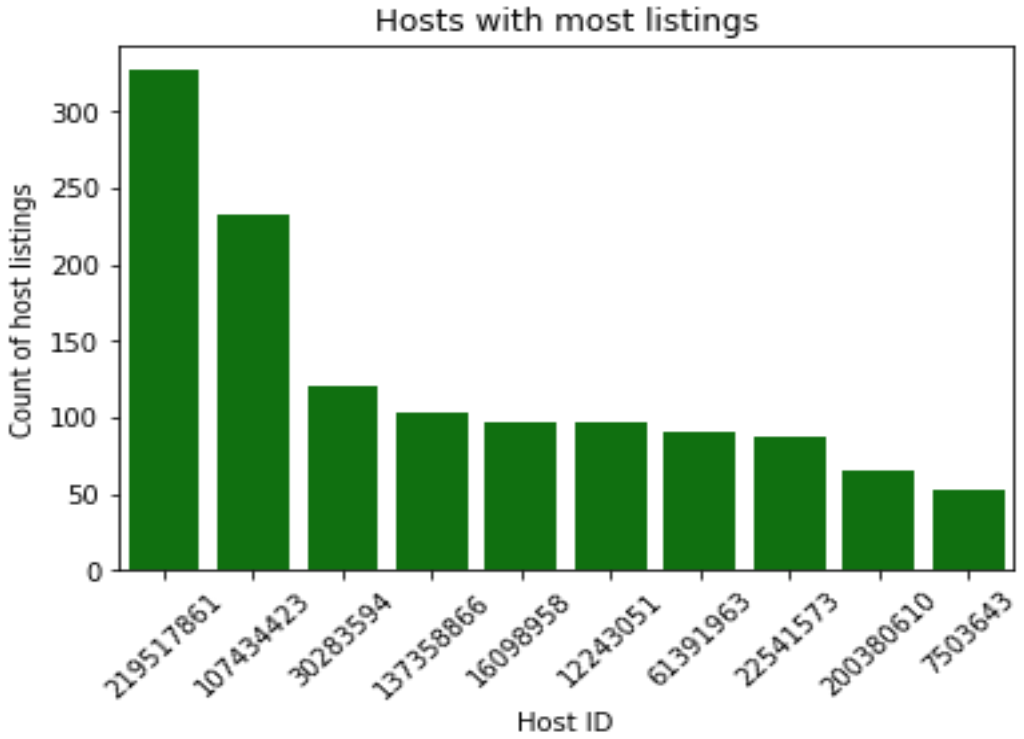
5.Neighbourhood(Bronx)



EDA:

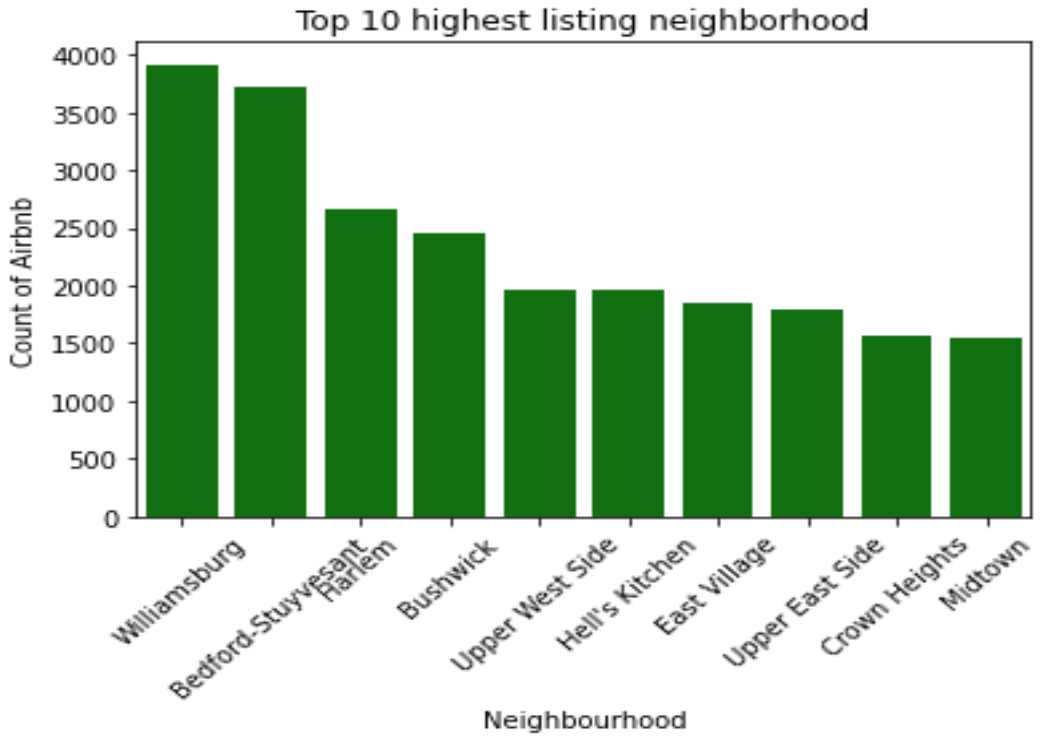
Count plot of Top 10 host :

Host with host id 219517861 has maximum number of listings with 327 listings



EDA:

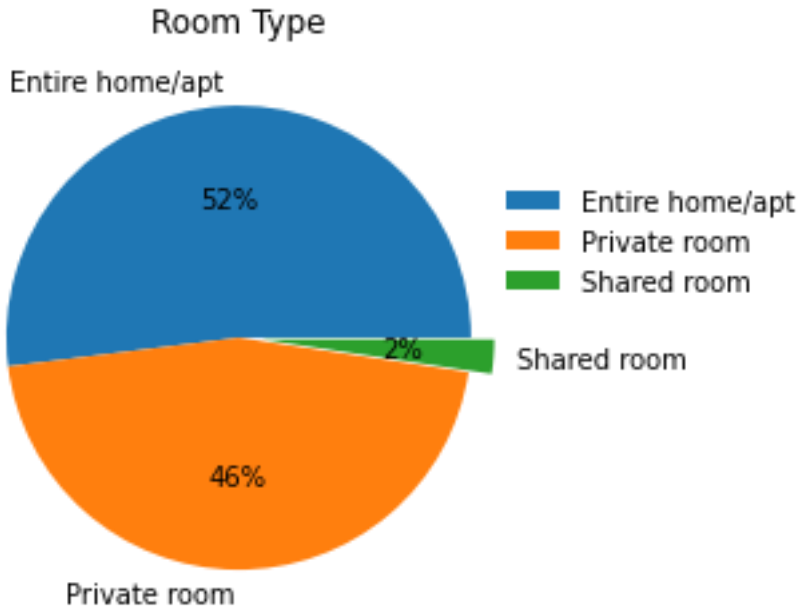
Count Plot of Top 10 highest listing neighbourhood



EDA:

Pie chart for room type:

Entire home/apt has highest count
followed by private room.
Shared room has least number of counts.

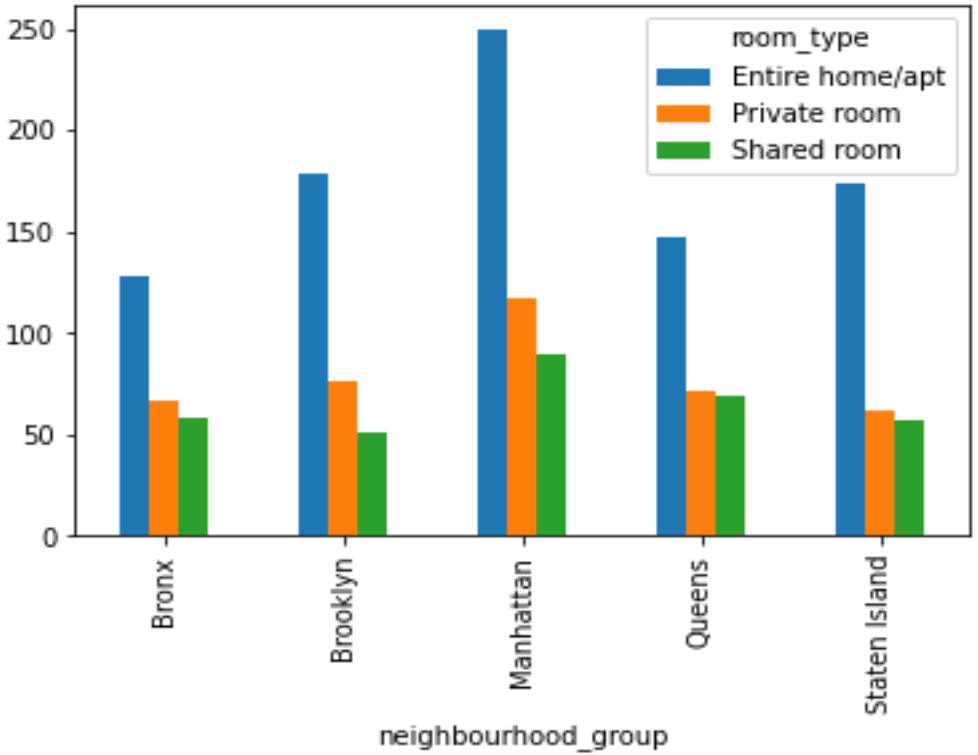


EDA:

Bar Plot of price of room type for neighbourhood group:

From above plot we can observe that average price for entire home/apt is highest compared to private and shared rooms.

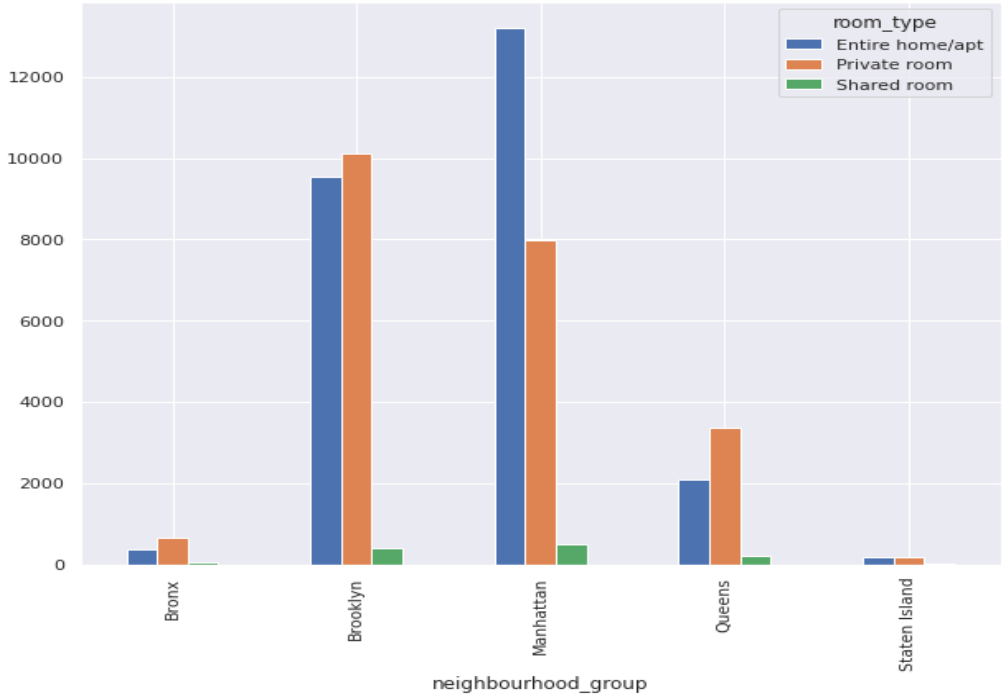
However, there is not much difference in price of shared and private rooms.



EDA:

Count plot of room type for neighbourhood group:

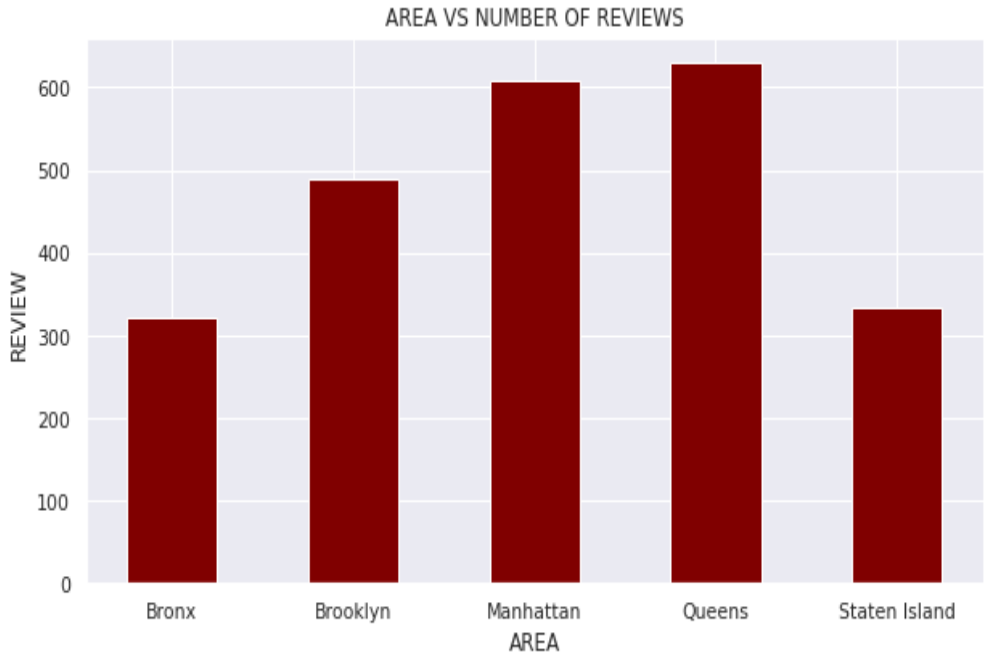
Manhattan has the maximum number of Entire home / apt as room_type and Brooklyn has the maximum number of Private room



EDA:

Bar plot of Number of reviews for each neighbourhood group

Here, Queens and Manhattan has large number of reviews and Bronx and Staten Island has less number of reviews as compared to others



EDA:

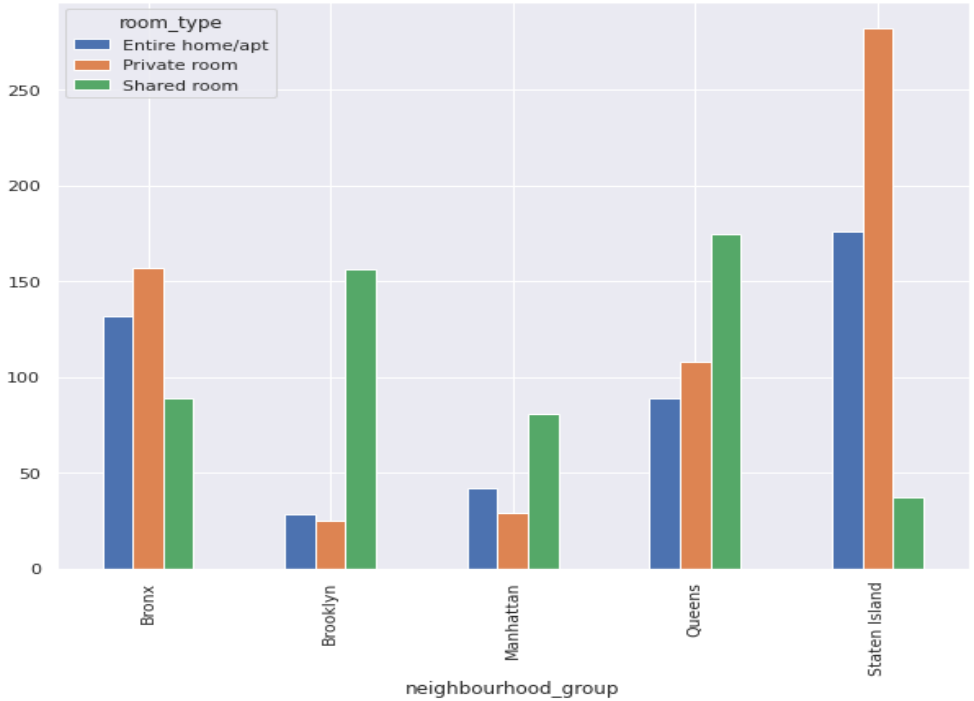
Scatter plot of price v/s Number of reviews:

From the above analysis we can say that most people prefer to stay in place where price is less

EDA:

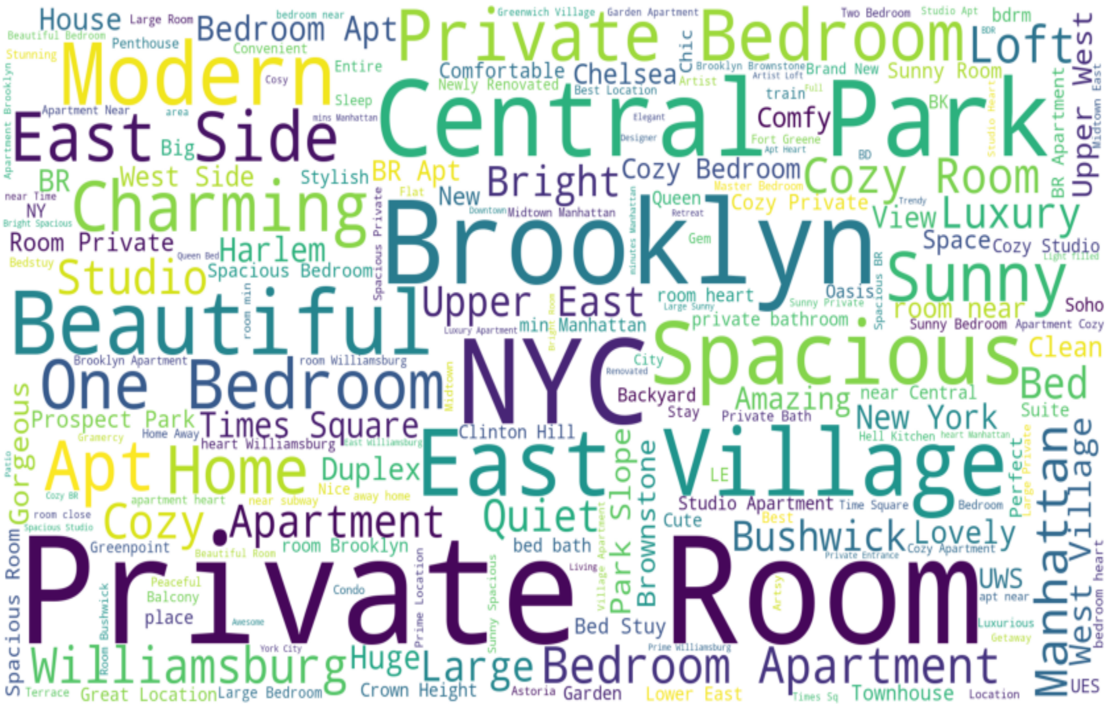
Bar plot of count of Availability of each Room type for each Neighbour hood group.

The avalaibilty of all room types is relatively low in Manhattan hence we can say that it is the most occupied neighbourhood of NYC.



EDA:

Wordcloud : To find the most used words in name of listings.



Conclusion:

1. Manhattan is the most spread area in New York for hosting.

2. The most popular room type is 'Entire home/apt,' with 52 percent of listings, and the least popular is 'Shared Room,' with only 2.4 percent of listings.

3. The average price for an entire home/apt is the highest, but there is no notable difference between the pricing of a Shared room and a Private Room in most of the neighbourhood groups.

4. In every neighbourhood, people stay in the entire home / apt room type for longer periods of time.

5. Since the avalaibilty of Entire home/apt and Private room is relatively low in Manhattan new hosts should invest in these room type.

6. The properties are usually described in the listing's name.