Capstone Project

on

Airbnb Booking Analysis

By

L . Srinadh

Introduction to 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 a shortened version of its original name, AirBedandBreakfast.com.

The company is regulated by many jurisdictions, including the European Union and cities such as San Francisco and New York City.

Introduction to Exploratory Data Analysis

Exploratory Data Analysis is a process of examining or understanding the data and extracting insights or main characteristics of the data. EDA is generally classified into two methods, i.e. **graphical** analysis and **non-graphical** analysis.

EDA is very essential because it is a good practice to first understand the problem statement and the various relationships between the data features before getting your hands dirty.

Technically, The primary motive of EDA is to

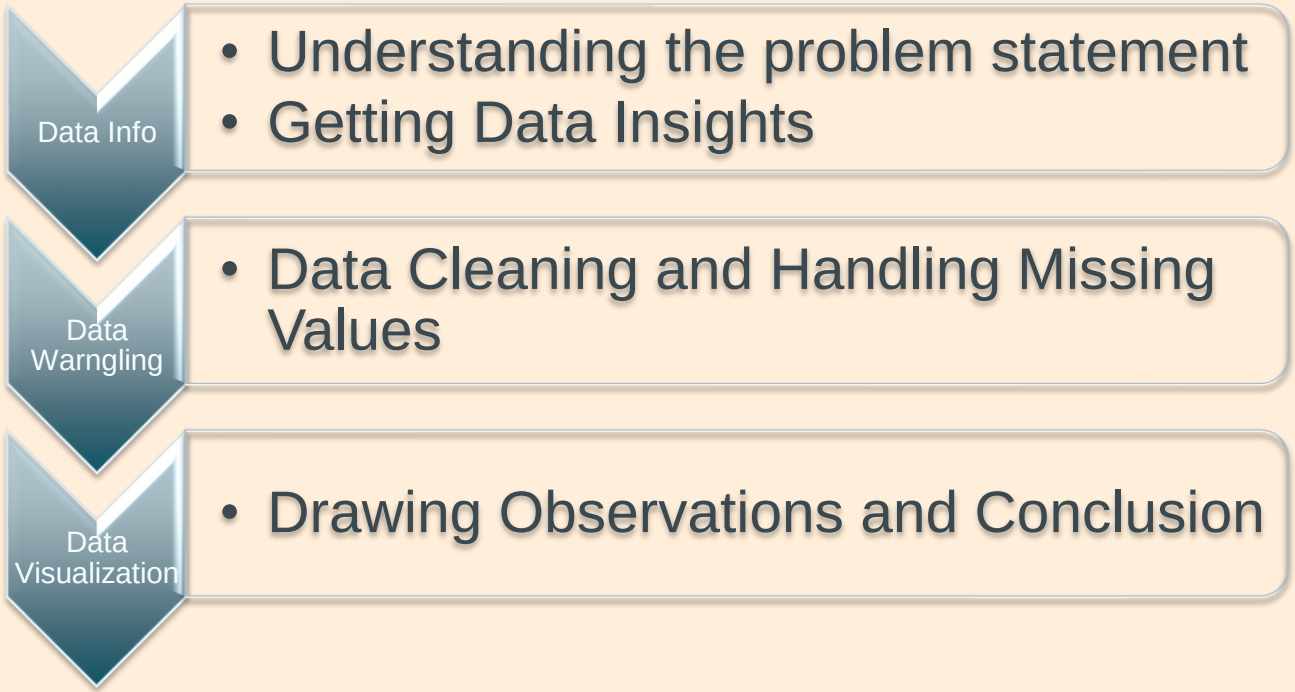
- Examine the data distribution
- Handling missing values of the dataset
- Handling the outliers
- Removing duplicate data
- Encoding the categorical variables
- Normalizing and Scaling

What we can analyze from this data?

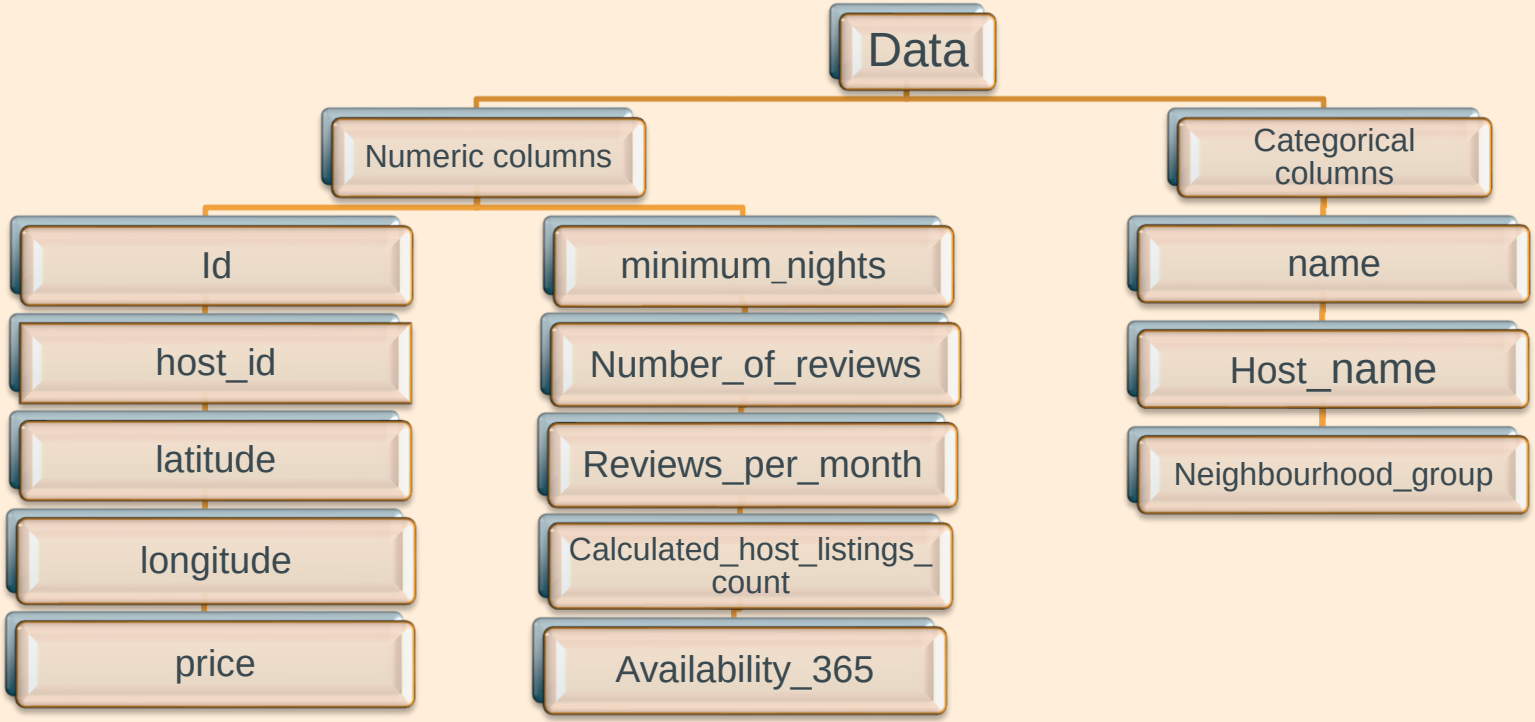
A lot of questions come to mind while thinking about analyzing the whole data set. So I picked the frequent and important ones.

- What are the top hosts in NYC according to neighbourhood and Airbnb listings?
- Which location got most of the properties?
- Analysis on room type on basis of area
- What can we learn from price predictions?
- What can we learn from review predictions?
- Which hosts are busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?
- Finally analyze on correlation matrix

Detailed analysis plan



Data Information



Lets take a quick look on data set

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

- This output data is extracted with the help of pandas.
- The output shows the columns which are non-null values counts of the data.
- It also shows the Dtype of all the columns.
- And it gives the shape of the data set i.e it shows the rows(48895) and columns(16) of the data.
- Finally it shows the memory usage.

Data Cleaning and Handling Missing Values

A dataset may contain lots of data as null values. These null values may cause an error while executing any code or while plotting graphs. So, these null values must be checked before operating on data. So I handle these null values by replacing it with zero or I can aslo remove the column if it is not necessary.

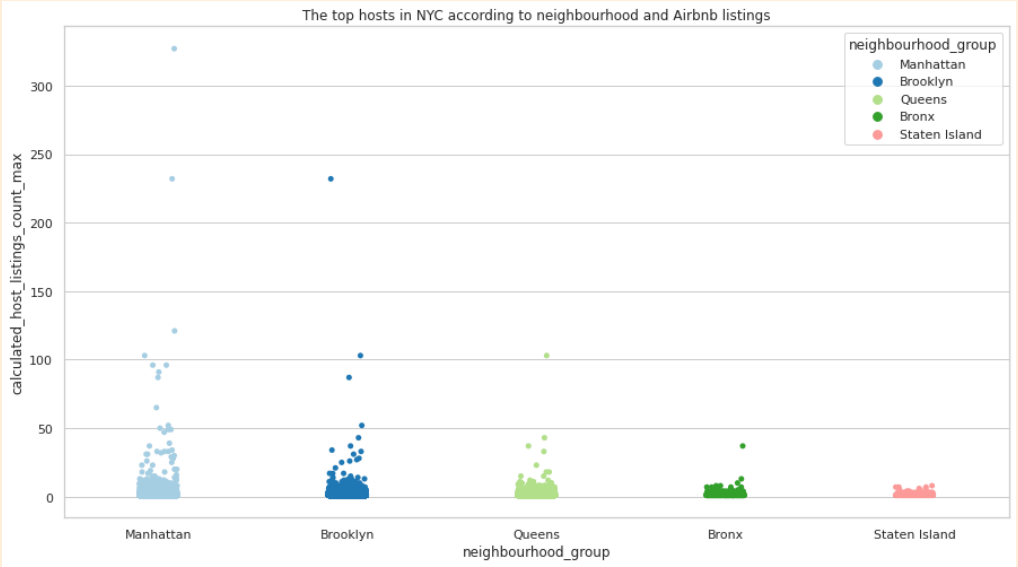
Data cleaning is an important part while performing data operations to maintain the flow of the program codes without interruption.

Data Analysis and Visualizations

What are the top hosts in NYC according to neighbourhood and Airbnb listings?

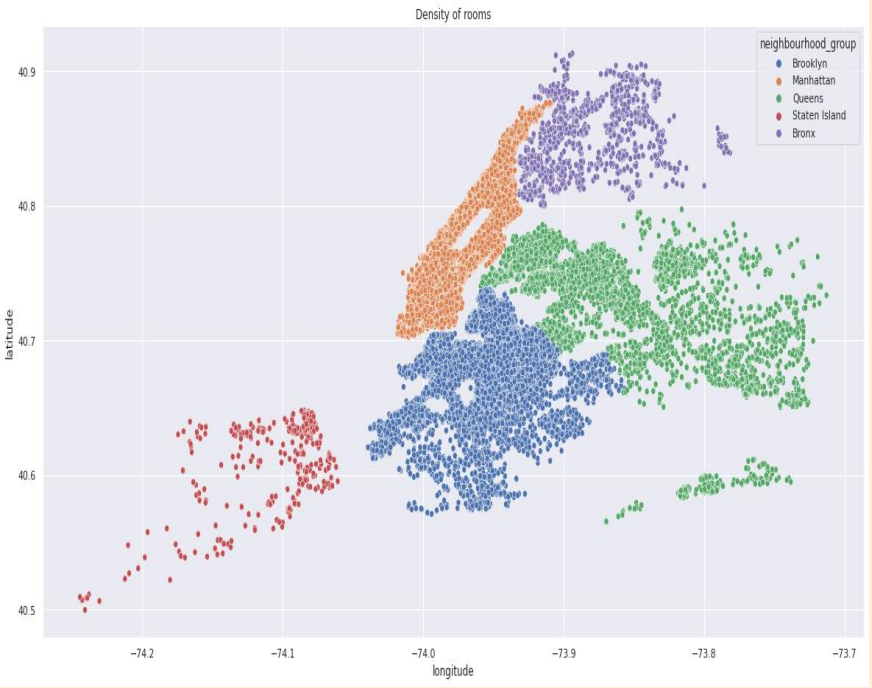
From this data we can see that most of the hosts are from manhattan.

So we can say that most of the traffic of hosts is from manhattan area.



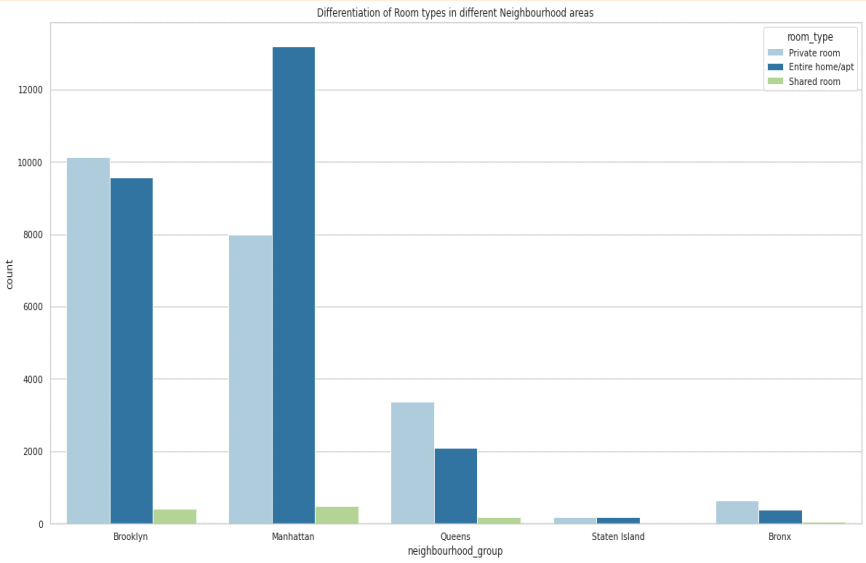
Which location got most of the properties?

- This analysis has been done on the basis of neighbourhood groups grouping it with host listings across NYC and plotted a scatter representation of it using latitude & longitude.
- The scatter plot map shows that manhattan is the most preferred area and most of the investors are interested in manhattan followed by Brooklyn and Queens. Staten island have least density of properties



Analysis on room type on basis of area

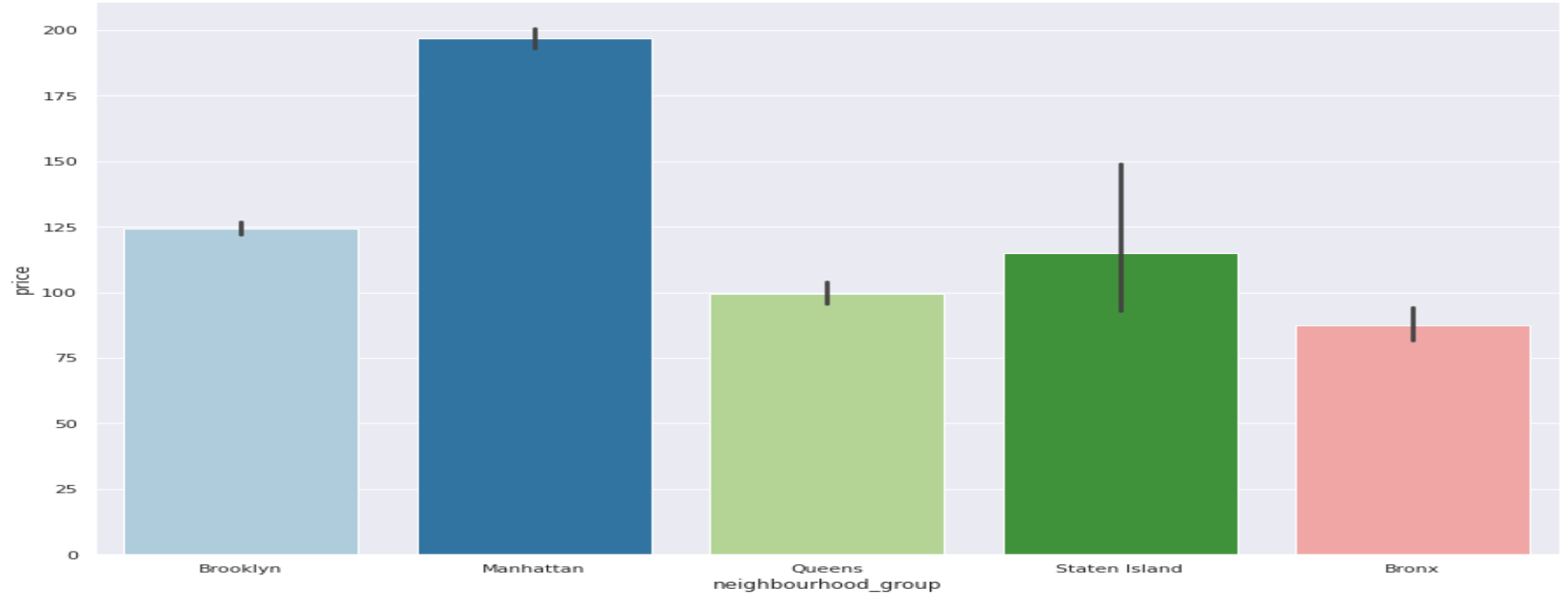
This Graph shows the relationship Between room types with respect to Neighbourhood group



- The graph shows that most of the bookings are for Home/apt followed private room.
- Shared rooms have the very least contribution.
- Most of the preferred room type bookings come from Manhatten and Brooklyn

What can we learn from price predictions

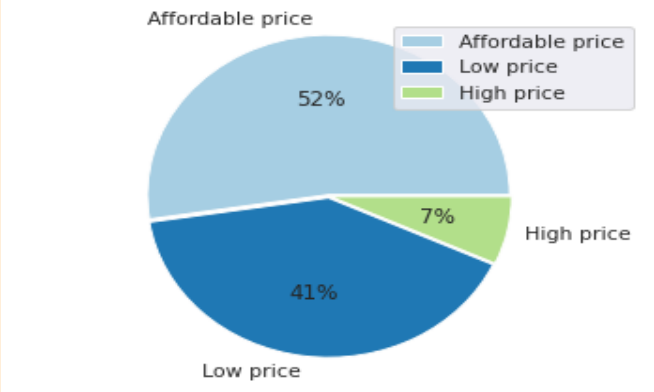
Manhatten is the most expensive neighborhood followed by Brooklyn and Staten

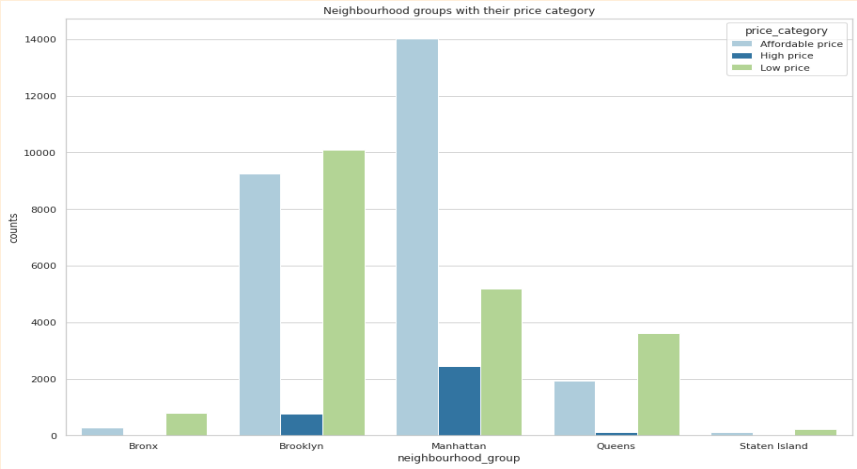


Analysis on price distribution using price categories

Here I consider the prices less than or equal to 90 as low prices and more than 90 but less 300 as affordable prices and more than 300 as high prices.

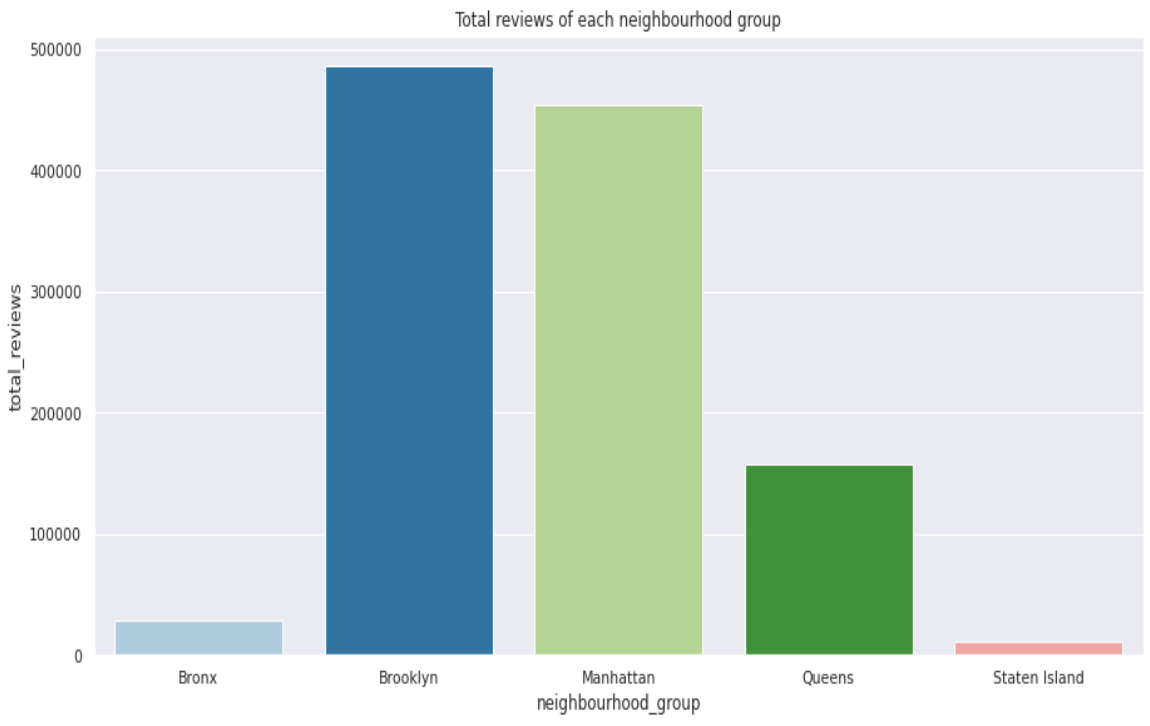
The pie chart shows that most of the people are prefer to the Affordable price category and least people prefer to the high price category.





The bar plot shows that the neighbourhood groups manhattan has more affordable prices and also high prices and brooklyn has more low prices comparing to all the neighbourhood groups.

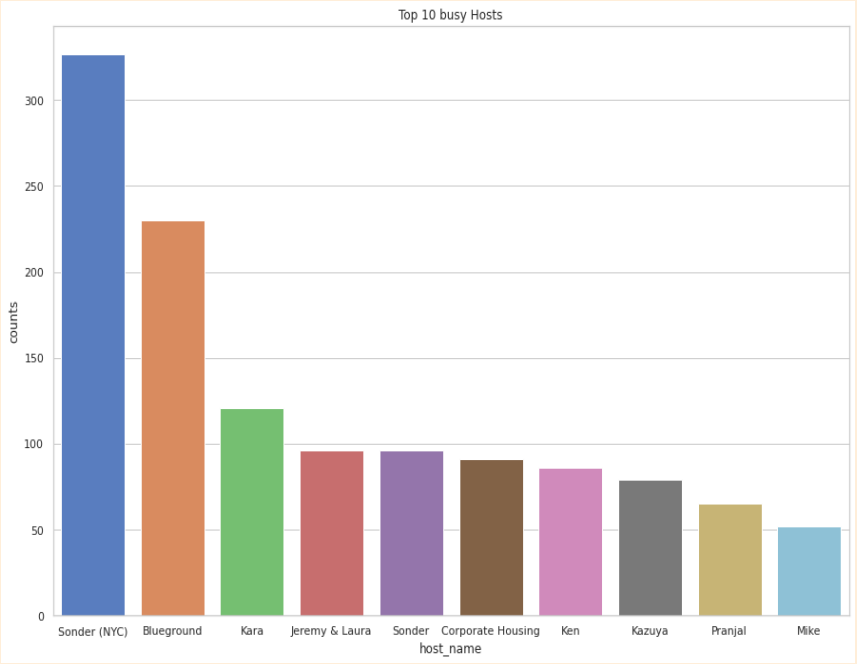
What can we learn from review predictions



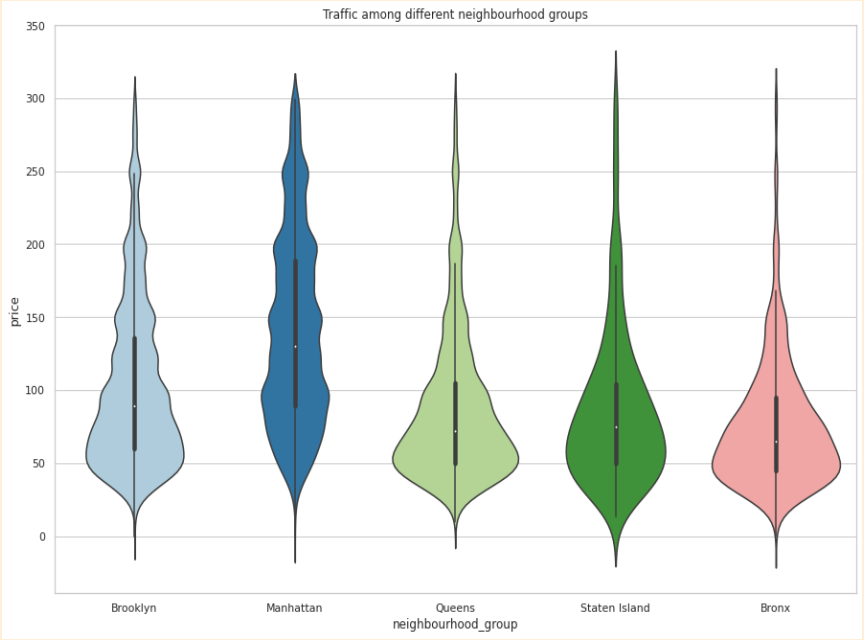
From the data and visualization. I can predict that most of the people who visiting the neighbourhood group brooklyn are giving the reviews.

Which hosts are the busiest and why

We can see the top 10 busy hosts on the plot that shows sondar(NYC) is the most busy host because these all hosts are from the neighbourhood group manhattan which have the most affordable prices and most visitings are intrested by the people.



Is there any noticeable difference of traffic among different areas and what could be the reason for it



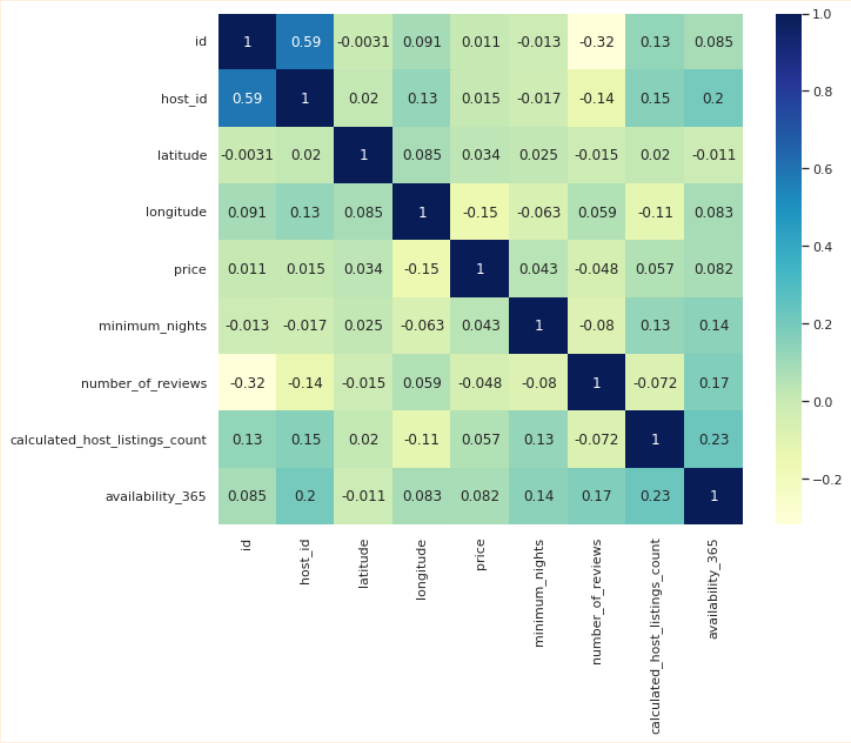
From the violinplot. I can observe that the most traffic in manhanttan area because that area have all facilities as for high class people they providing expensive rooms and for all the people can stay with affordable prices.

Finally analyze on correlation matrix

As calculated host listings count increases, the number of rooms available would also increase which accounts for the positive correlation among the availability 365 and calculated host listings counts.

Also if the availability is higher more people will choose that host and give reviews which accounts for the positive correlation between availability and reviews per month.

As we go to right of the map, the longitude value increases and the id counts also increases which in turn accounts for the increased number of reviews per month.



Difficulties I faced during my analysis

- I faced challenges in selecting the columns for the analysis.
- And also I have face some challenges in choosing the right plot as visualization is the most important part of the project hence selecting the appropriate features was very important.

Conclusion

I have done various analysis which directed us to the below conclusions.

- The scatter plot of neighborhood groups with property density shows that Manhattan and Brooklyn are the most favorable area for investors as well as customers. So more property owners can be considered for business in this area.
- Analysis on room type showed that more people are interested in renting a private room or entire home/apt than shared room making or business concern towards the private room or entire home/apt category
- Price predictions show that Manhattan is the most expensive area followed by Brooklyn while the Bronx is the cheapest although price category analysis showed that most visitors come from the affordable category so concentrating our business towards affordable properties.
- Analysis for most bookings again showed that Manhattan and Brooklyn are best for investment.
- Violin type graphical analysis showed that the Bronx and Staten island show that property rates are a little lower there as compared to others so it will be better to invest there for future business.


Thank You