Capstone Project

on

Airbnb Booking Analysis

Team Members:

Shreyash Movale
Faissal Shah
Eshaan Sosa
Ajinkya Jumde
Neha Gupta

What is 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.

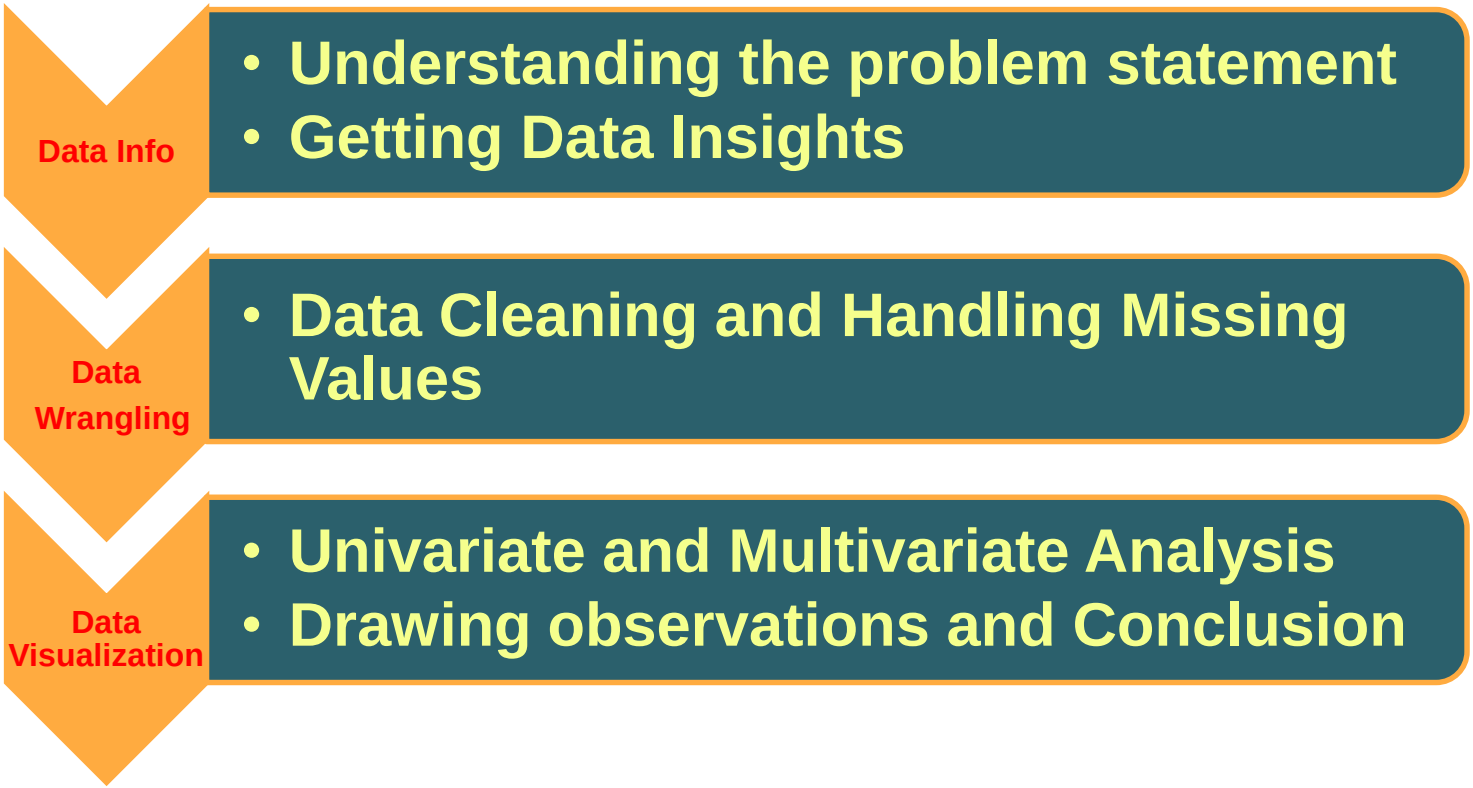
Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Wikipedia

What can we analyze? (most frequent questions)

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

- **Which neighborhood area has most of the properties?**
- **Which hosts are busiest and why?**
- **Which type of rooms are preferred by customers and where?**
- **What can we learn from predictions?**
- **Which area shows most of the bookings?**
- **Is there any noticeable difference in bookings among different areas?**

Roadmap of Analysis



Data Information

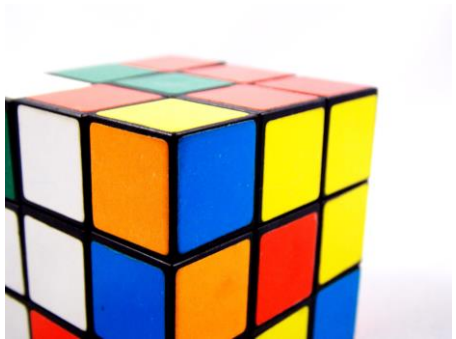
Categorical

- **Room Type**
- **Areas**
  - Neighbourhood
  - Neighbourhood group
  - Latitude & Longitude
- **Host**
  - Host id
  - Host name

Numerical

- **Host Listings Count**
  - Availability
- **Reviews**
  - Reviews Per Month
  - Last Review
  - Number of reviews
- **Minimum Nights**
- **Price**

Data



A Quick Look of Data

```
<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
```

- The output of data info extracted from pandas library of python.
- It shows the columns in tables, with non-null values counts and data type of each column
- Also, it shows Class and memory usage

Data Insights

- **Neighbourhood Groups:** One of the important data categories on the basis of which we're dealing with various other dependent parameters such as price, total properties located in the particular area, and key factors in determining the area of interest for the investors. Neighbourhood groups consist of child areas which belong to that particular neighbourhood group.

- **Latitude/longitude**: These parameters give the location of neighbourhood groups on the world map which will be helpful for us in studying variations according to the locations of that area. A graphical representation can be done on a location basis to understand better distributions of properties in NYC.

Data Insights Continued…

- **Price Category:** We have divided the price into three categories (Expensive, affordable, cheap) to get our data predictions in a more concise manner.

- **Reviews**: A combination of the Number of reviews a particular host and property got, last review, and reviews per month. The review-based analysis is done by taking into the loop the neighbourhood and room type.

- **Room type**: The Airbnb dataset has three categories of rooms. They are a Private room, an entire home/apartment, a Shared room, etc.

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.
- From a data analysis point of view name and hostname will not be that important as it's a categorical feature and will have lots of categories and this will not contribute to exploring the data and also from a security point of view of the host.
- Data cleaning is an important part while performing data operations to maintain the flow of the program codes without interruption

Data Analysis and Visualizations

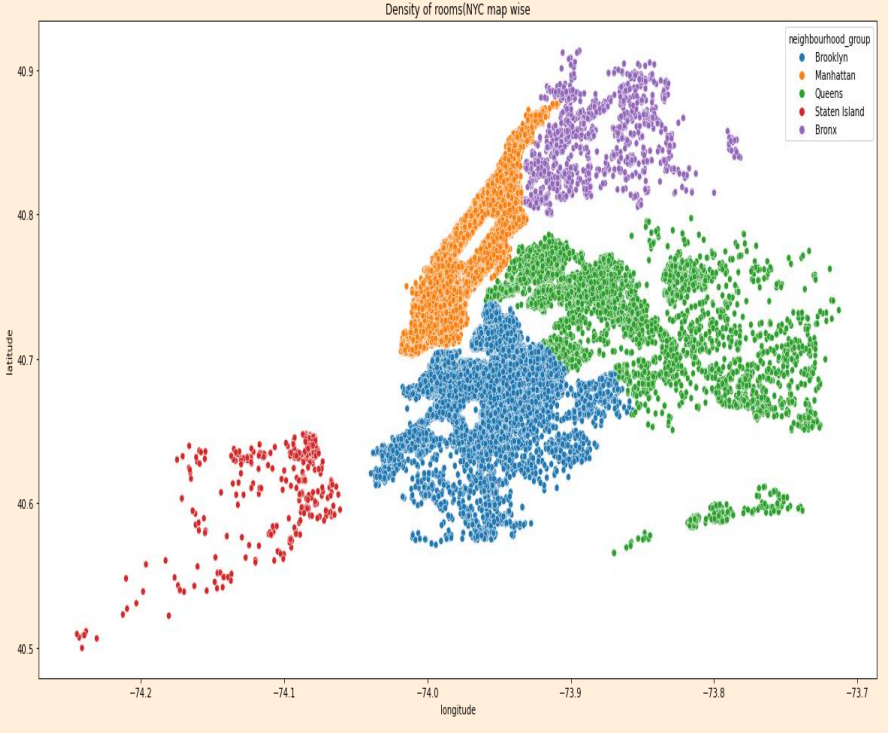
Which neighbourhood area has most of the properties?

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.

What can we predict from this?

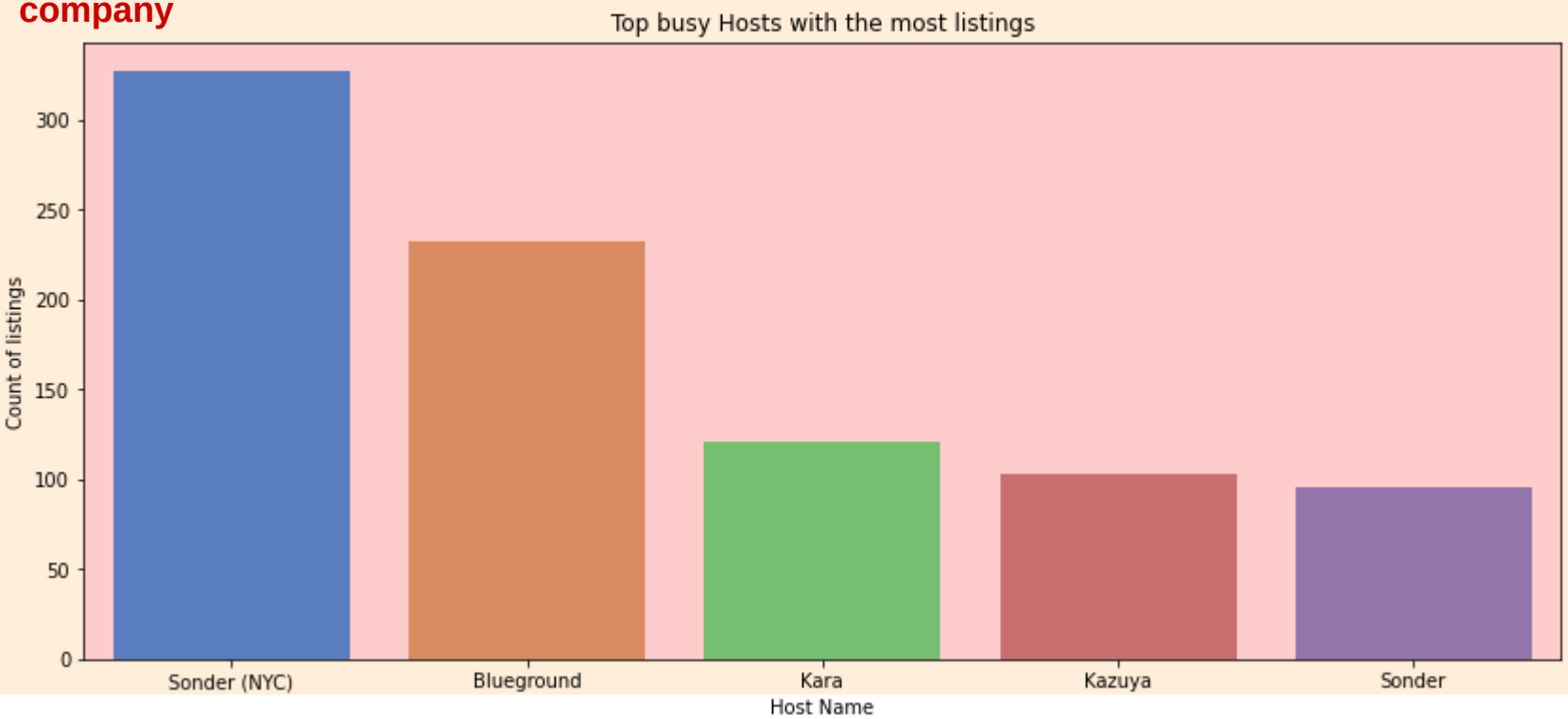
From the above scatter plot map, we can conclude that manhattan (orange) is the most preferred area. Most of the investors are interested in manhattan followed by Brooklyn and Queens.

Brooklyn gains a little less attraction while Staten Island has the least density of properties



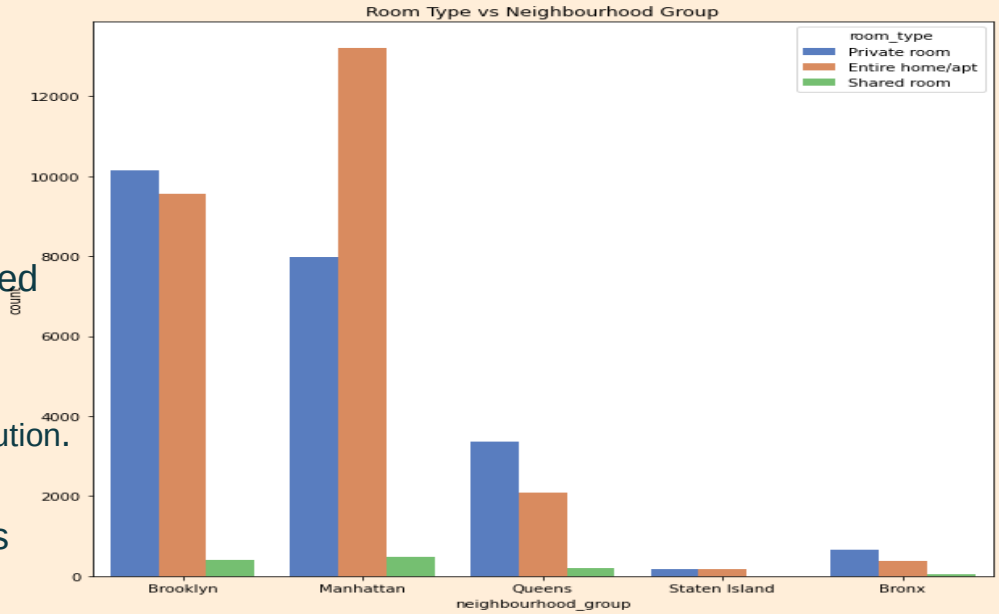
Which hosts are busiest and why?

This graph helps us to list some top hosts which contribute a large share towards business of company



Which type of rooms are preferred by customers and where?

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

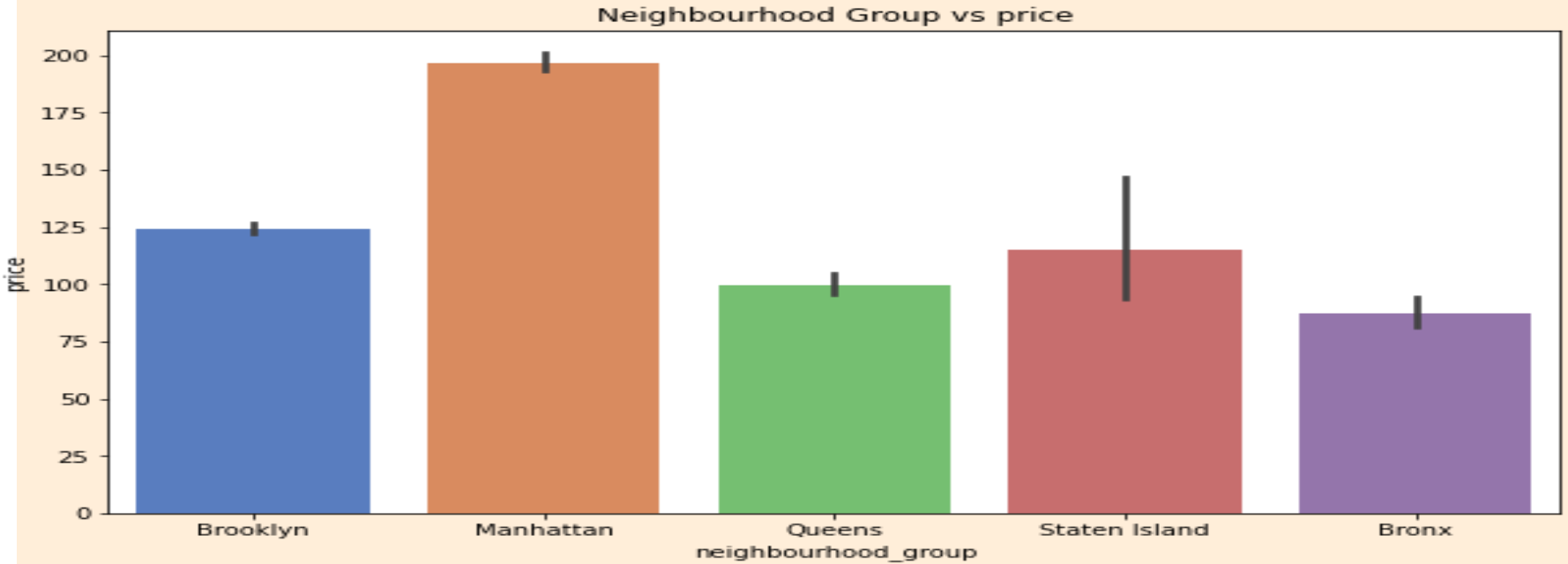


What do we learn from this?

- The above graph shows that most of the booking, s 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 Island

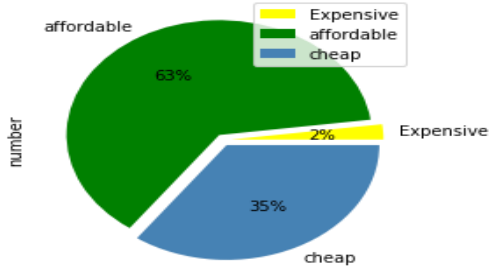


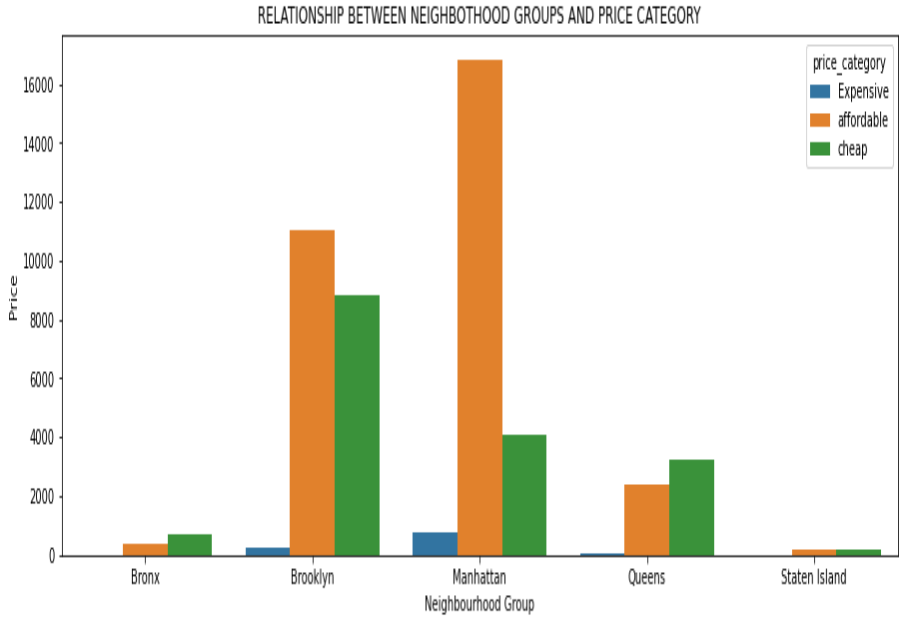
Analysis of price distribution using price categories

***Here we consider the price less than or equal to 80 as cheap, more than 80 but less than 500 as affordable, and more than 500 as expensive.*

The above bar graph shows the relationship between the neighborhood groups and the price category.

The least people prefer the expensive category, instead, the maximum people prefer the affordable category followed by the cheap category in all the neighborhood groups except in the case of Bronx and Queens where the relationship is reversed.

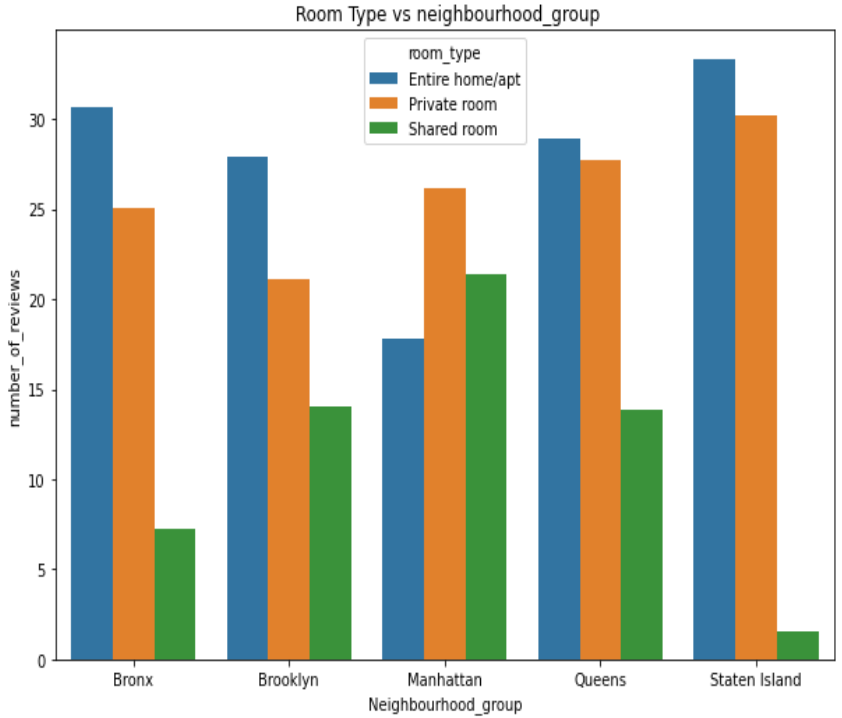




What can we learn from reviews predictions?

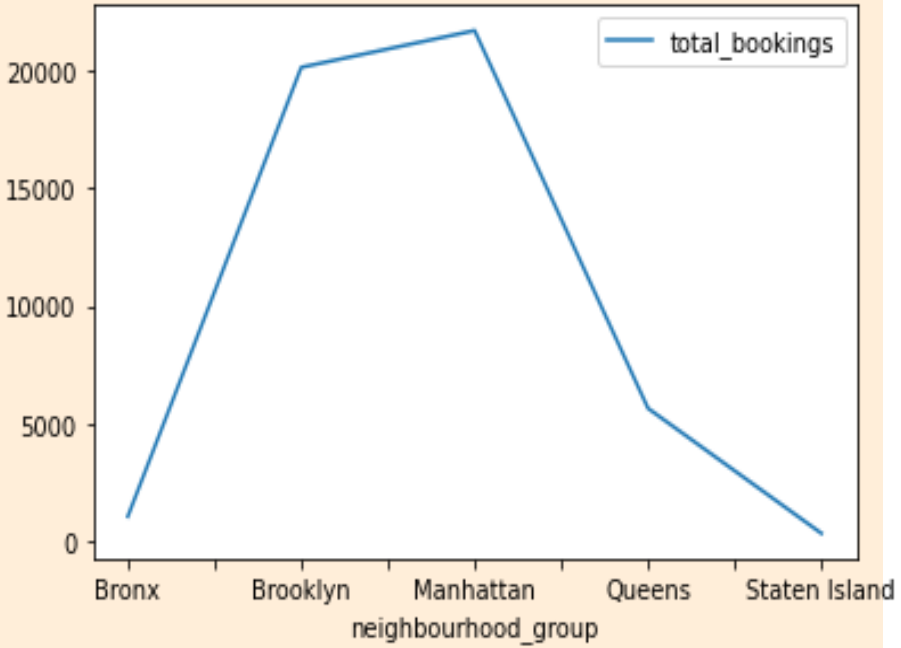
- Number of reviews has an equal average distribution over the neighborhoods groups

- Shared rooms show got lesser reviews as people mostly do not prefer shared rooms



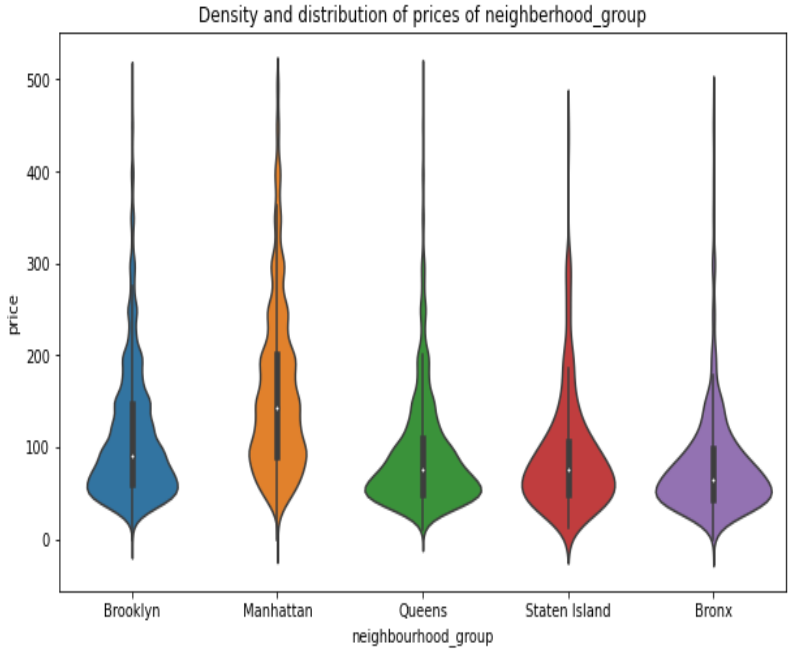
Which area shows most of the bookings?

- The neighborhood areas with most bookings have been plotted with the help of a line graph.
- The line showing bookings is at peak in Brooklyn and Manhatten which makes it crucial area for business purposes



Is there any noticeable difference in bookings among different areas?

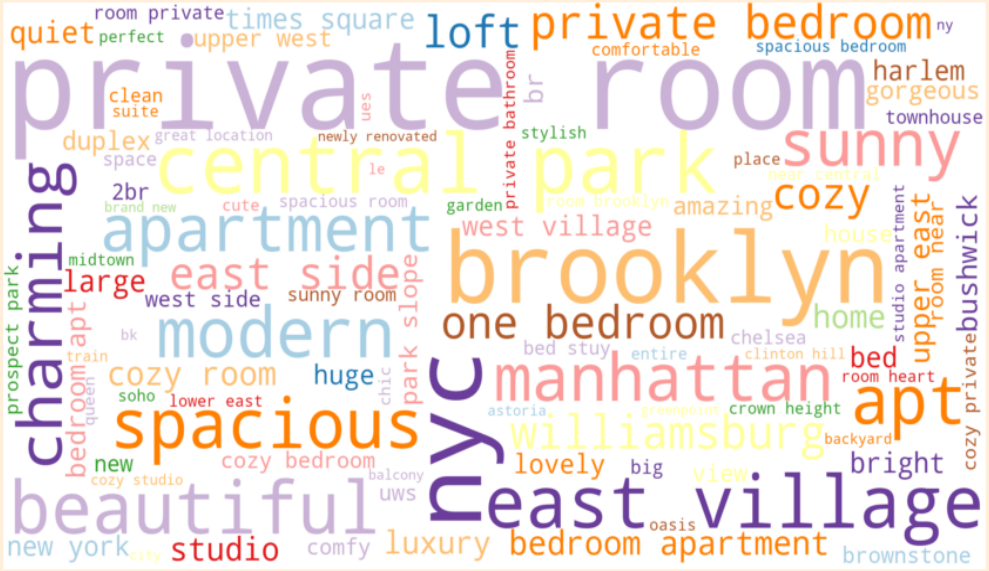
- With a violin plot we can definitely observe a couple of things about distribution and density of prices for Airbnb in NYC Groups.
- First, we can state that Manhattan has the highest range of prices for the listings with $150 price as average observation, followed by Brooklyn with $90 per night.
- Queens and Staten Island appear to have very similar distributions, Bronx is the cheapest of them all.



Frequently searched words

These are some frequently Searched words by the Customers.

So, we can highlight this Words on website or portfolio to get customer attraction

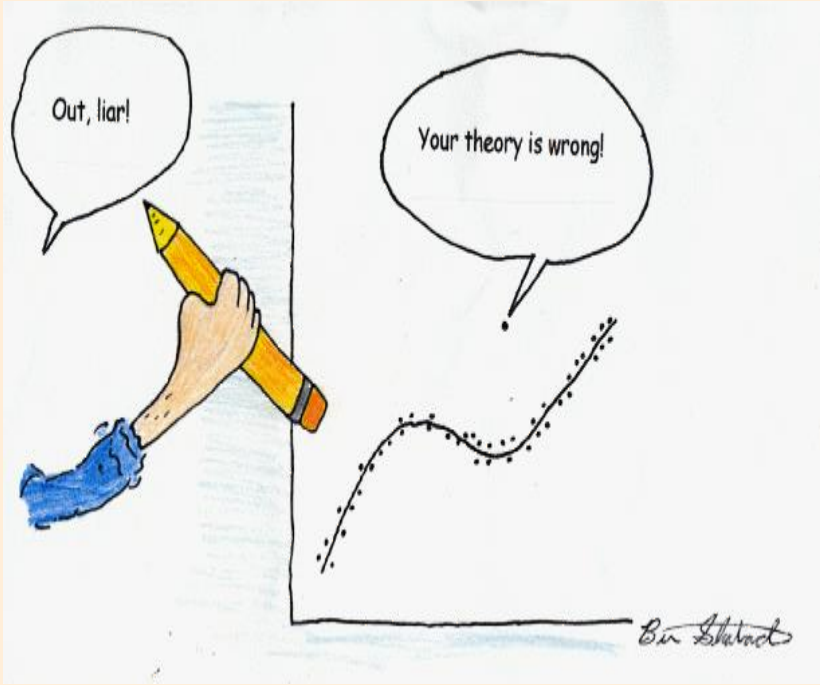


Outliers

While dealing with the data we come to know that there are many outliers throughout the data set

These outliers deviated the graphs towards themselves which we dealt with taking mean values wherever possible.

Most of the outliers belong to Manhatten among neighborhood groups



Conclusion

- We have done various analyses 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 flat/apt than shared room making or business concern towards the private room or flat/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**