Capstone Project

Airbnb Booking Analysis

Done by:

Sananda Biswas Chatterjee

&

Amit Kundu

CONTENTS

INTRODUCTION

- Airbnb is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities based in San Francisco, California.

- Airbnb doesn’t own any of the houses, instead the company acts as a broker and charges commission for each booking. It basically connecting travelers to local hosts who want to rent out their homes with people who are looking for accommodations in specific locales.

- Since its co-founders' initial idea in 2007 to let paying guests spend the night on an air mattress in their living room, the company has gone a long way. According to Airbnb's most recent data, there are more than 5.6 million listings available worldwide in more than 220 countries and more than 100,000 cities and towns.

- Airbnb is an abbreviated version of AirBedandBreakfast.com company's original name.

PROBLEM STATEMENT

- Through the Airbnb website, a customer can make a room reservation. The consumer is relocating to the neighbourhoods of Queens, the Bronx, Staten Island, Brooklyn, and Manhattan in order to explore more.
- The customer needs help finding a variety of bookings in each neighbouring nation and being guided in choosing the ideal area to live at a price they can afford, analysing the room's price so they can afford it, and choosing the right room type depending on their demands.
- These requirements will be satisfied by the volume of reviews, the frequency of reviews each month, and the room availability.

  To discover significant insights, we must study and analyse the data, such as:
  - What can we learn about different hosts and areas?
  - What can we learn from predictions? (Ex: locations, prices, reviews)
  - 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?

FEATURE INFORMATION

The dataset has 16 different features/columns with more than 49000 observations. Here are the brief discussion about these features

- **id**- Unique/distinct for each observation.
- **name**- Name of properties.
- **host_id**- There are 1270 host identifiers available in the dataset, and each id assigned to a single host.
- **host_name**- Name of host.
- **neighbourhood_group**- It represents five different locations [‘Manhattan’, ’Brooklyn’, ‘Queens’, ‘Staten Island’, ‘Bronx’].
- **neighbourhood**- It represent specific areas where the listing is located.
- **Latitude**- This column displays the property’s distance between north or south of the equator
- **Longitude** -This column displays the property's distance between the east or west prime meridian.
- **room_type**- It represent three unique type of rooms [‘Private room’, ‘Entire home/apt’, ‘Shared rooms’]
- **price**- Per night for rental property.
- **minimum_nights**- Number of nights spent by customers.
- **number_of_reviews**- Number of reviews per listing.
- **last_review**- Last date of review
- **review_per_month**- Number Of reviews per month.
- **calculated_host_llisting_count**- Total number of listings owned by the host.
- **availability_365**- The number of days in a year a specific host is available.

STEP TAKEN

Loading dataset-

We are writing our script for this project using Google Collab. We used Airbnb data that is freely available online under the Creative Commons License in order to obtain the information. We first need to import the numerous external libraries and modules required for visualization and analysis before we can load the data into our IDE.

STEP TAKEN

Data cleaning-

The next stage is to clean up the data. Frequently, the data we load has a various type of problems like missing values, incomplete data etc. The data quality can be improved through cleaning in order to be used for more in-depth analysis.

STEP TAKEN

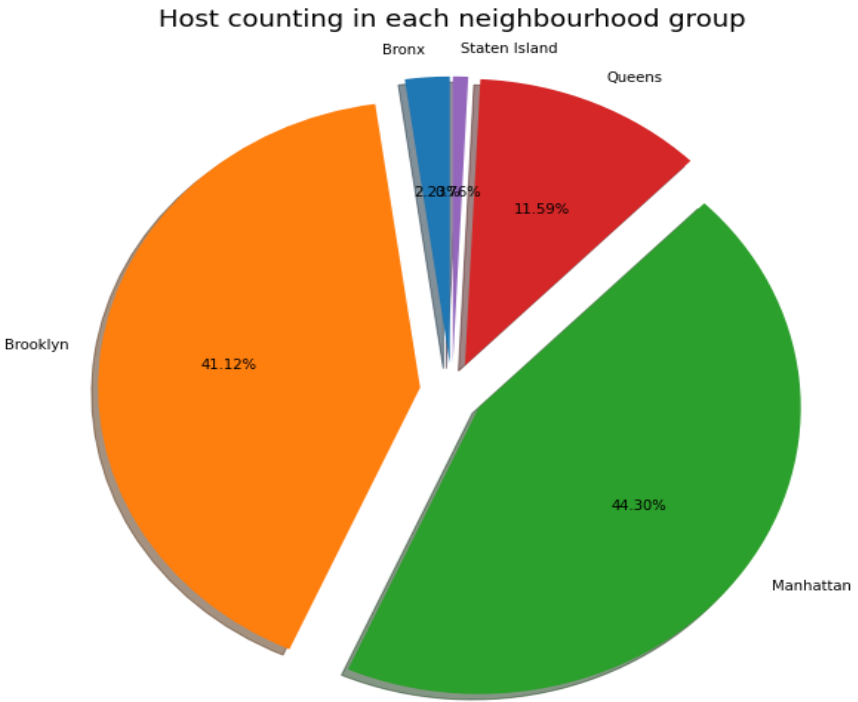
Handling missing value -

In real-world circumstances, missing data is a serious problem. NA (Not Available) values are another term for missing data in the pandas language. Isnull(), Notnull(), Dropna(), Fillna(), Replace(), and other useful functions are available in Pandas Data Frame for finding, removing, and replacing null values. We replaced all null value with a set of identified data in order to manage the missing value.

Replacing all the null value of reviews_per_months by ‘0’
Replacing all the null value of host_name by ‘No_Name’

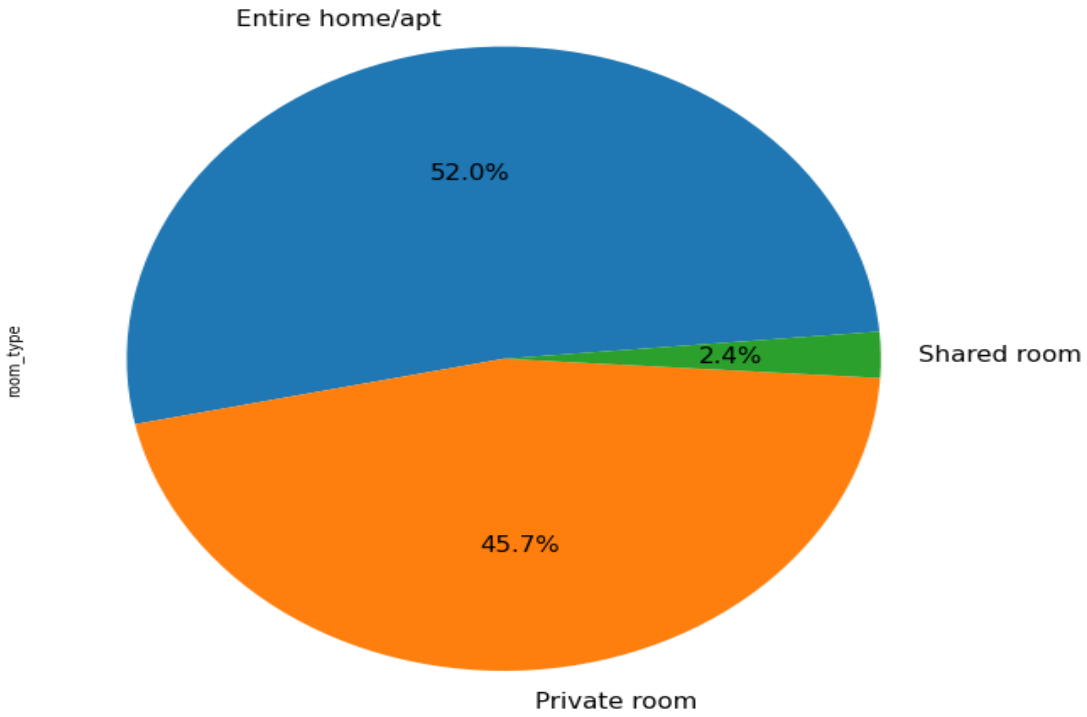
Exploratory Data Analysis

Host Counting in each neighbourhood group



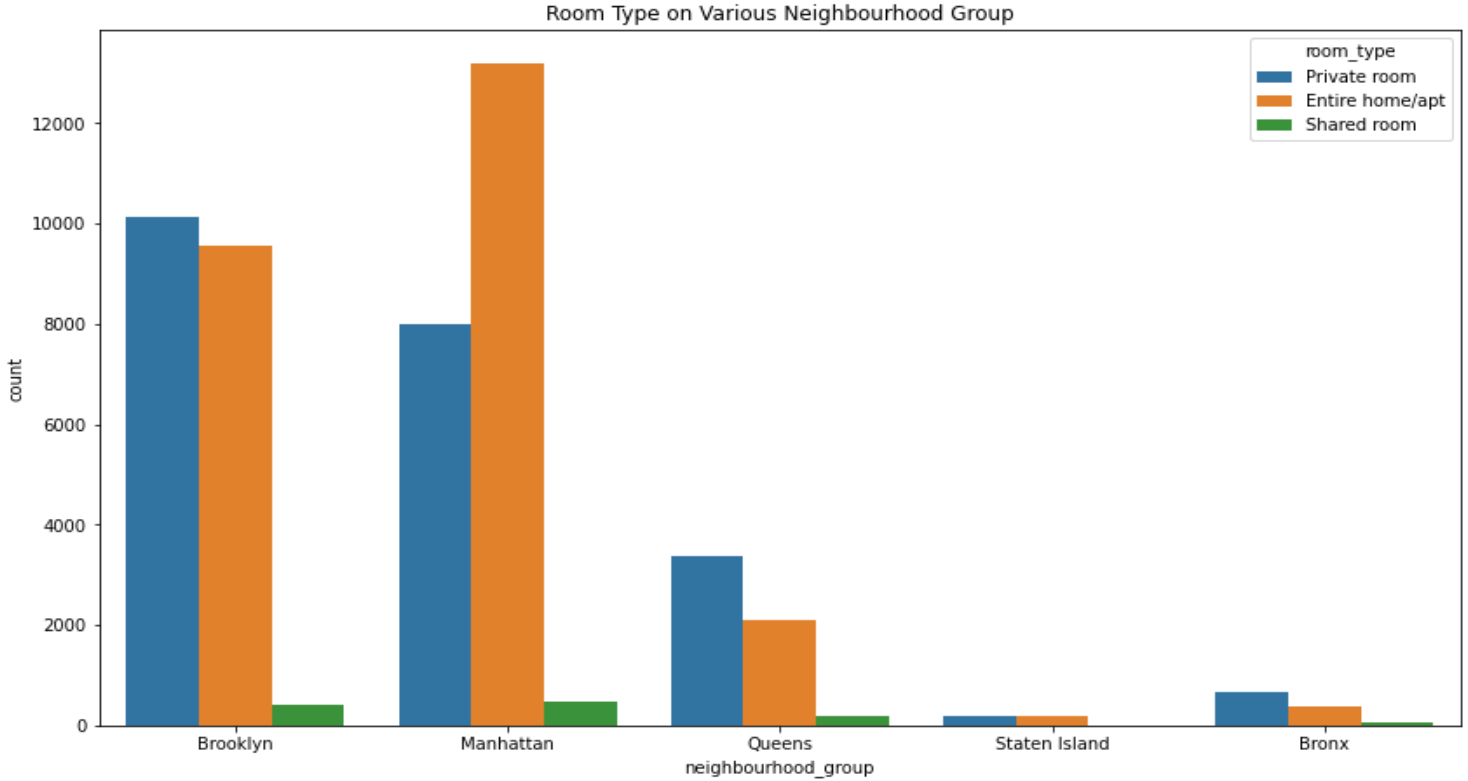
The majority of the hosts are from Manhattan and Brooklyn, making these places the most well-known.

Room type preferred by Customers



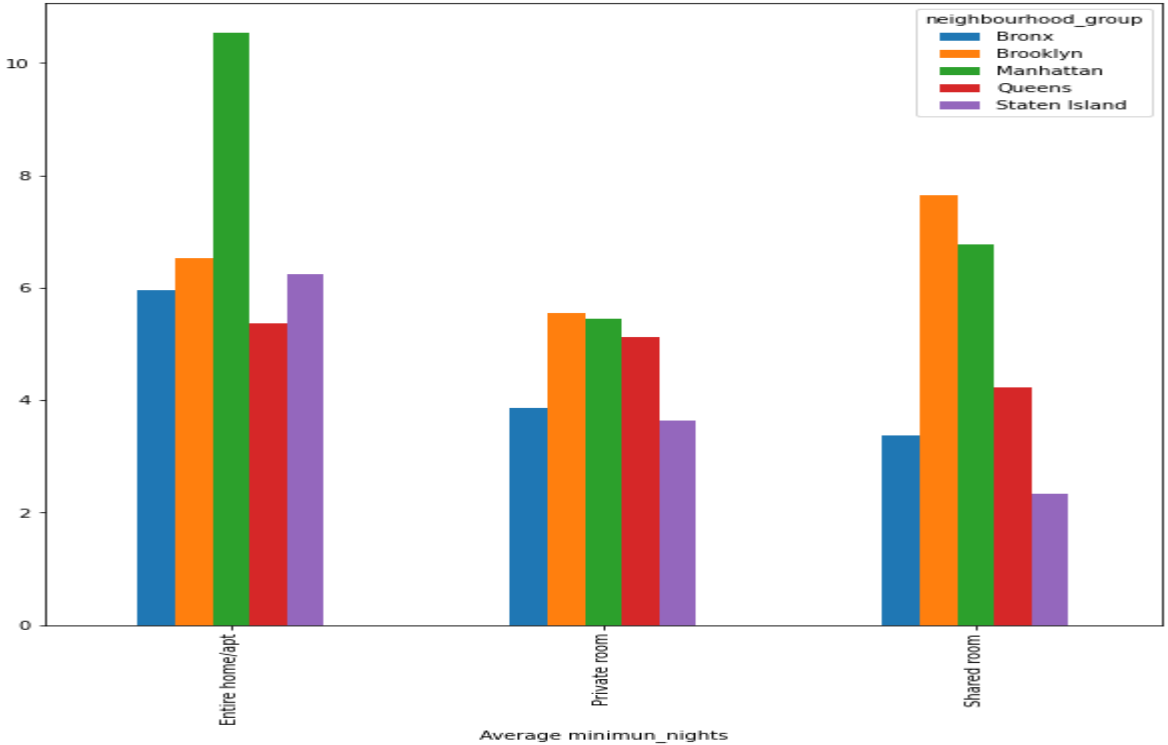
This graph describes that customers prefer living in Apartment or private rooms more than shared rooms.

Room type on various Neighbourhood Group



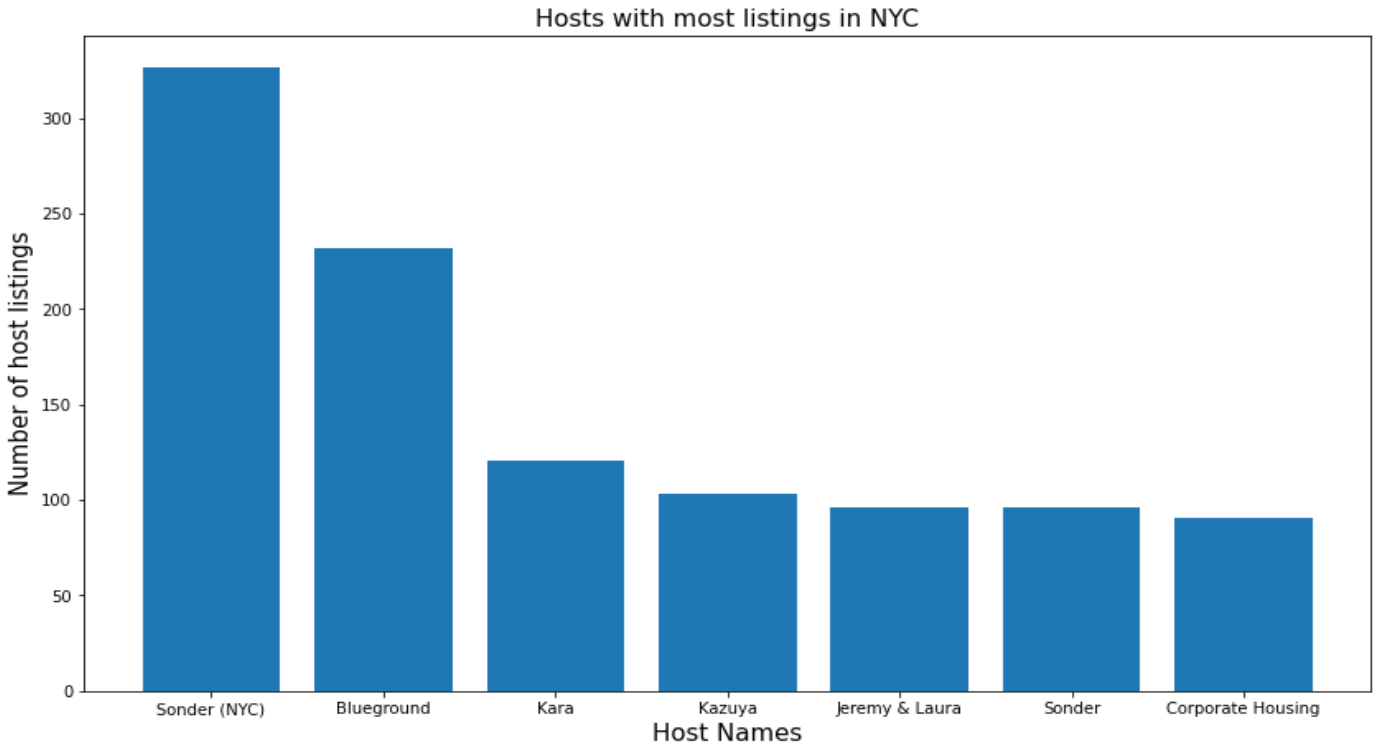
Majority hosts in the Manhattan and Brooklyn provide either a Private room or Entire home/apt.

Average Night spent in each type Room



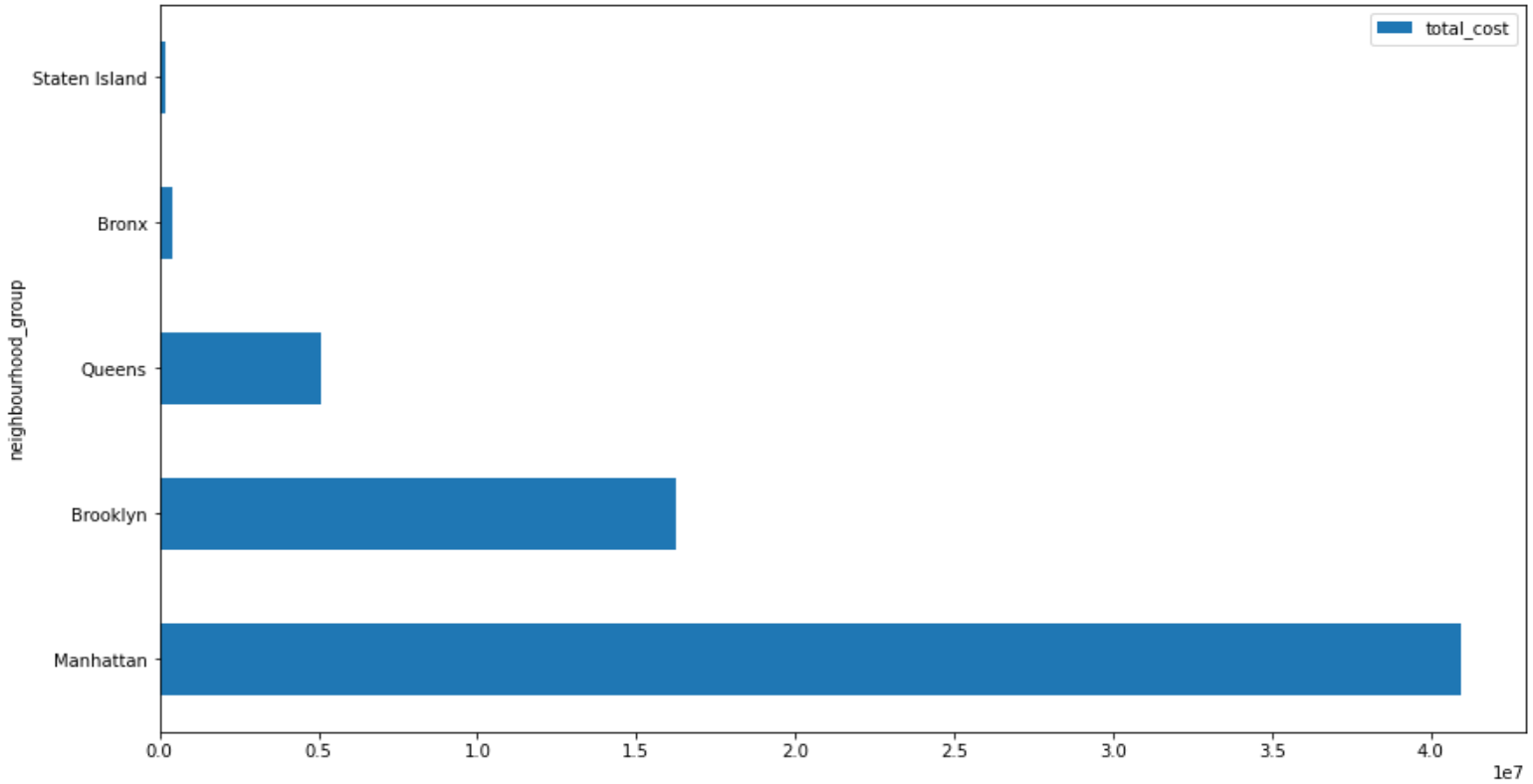
from this chart we can observed that in Manhattan, customer has spent more number of night in entire apartment compare to other room types while Queens has almost same average night spendings by customers in private and entire apartment. And Brooklyn has most number of average night spending in shared rooms.

Hosts with most listings



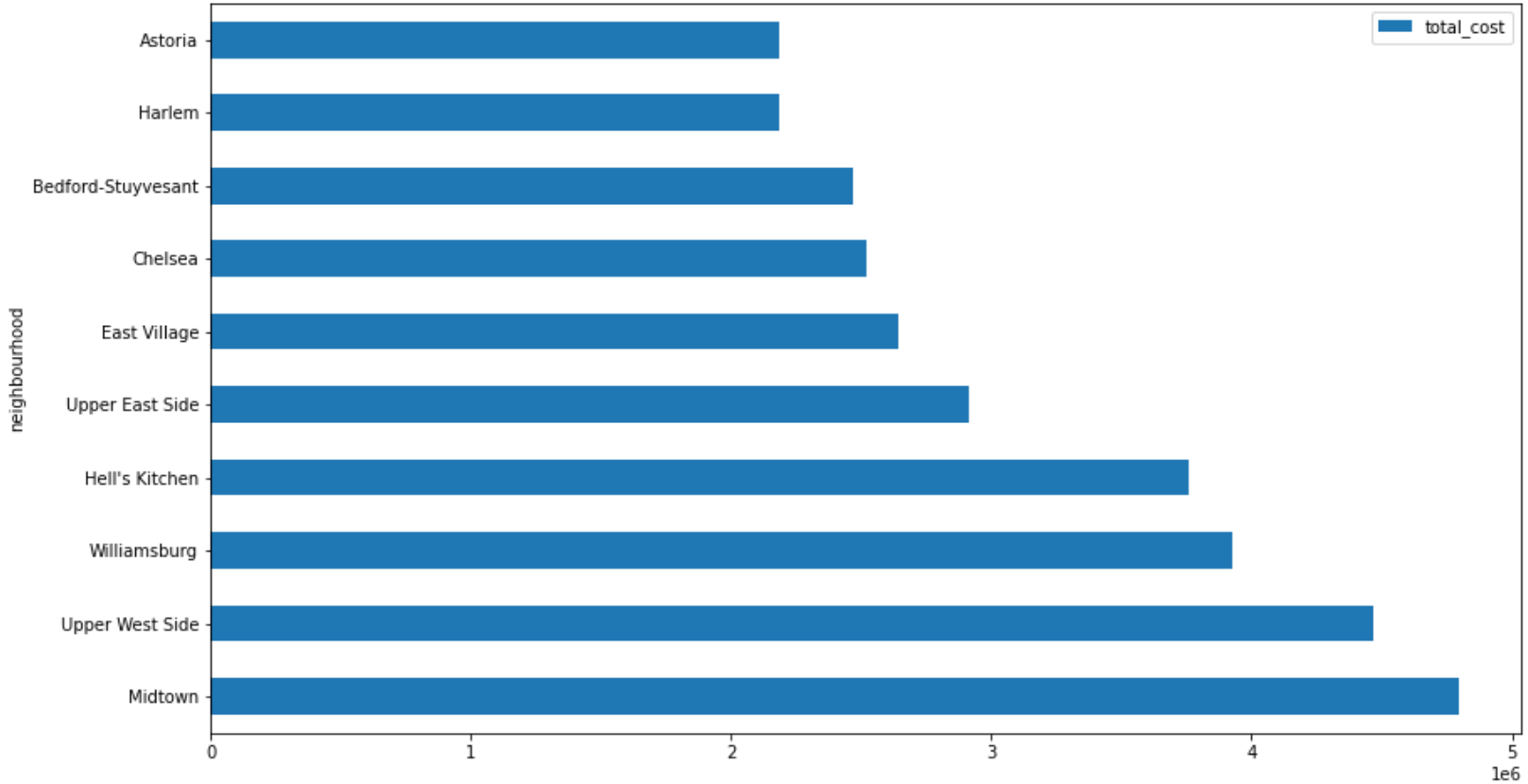
We can infer from the results above that Sonder(NYC), with 327 listings in Manhattan, is the host with the most listings.

Maximum profit among the Neighbourhood Group



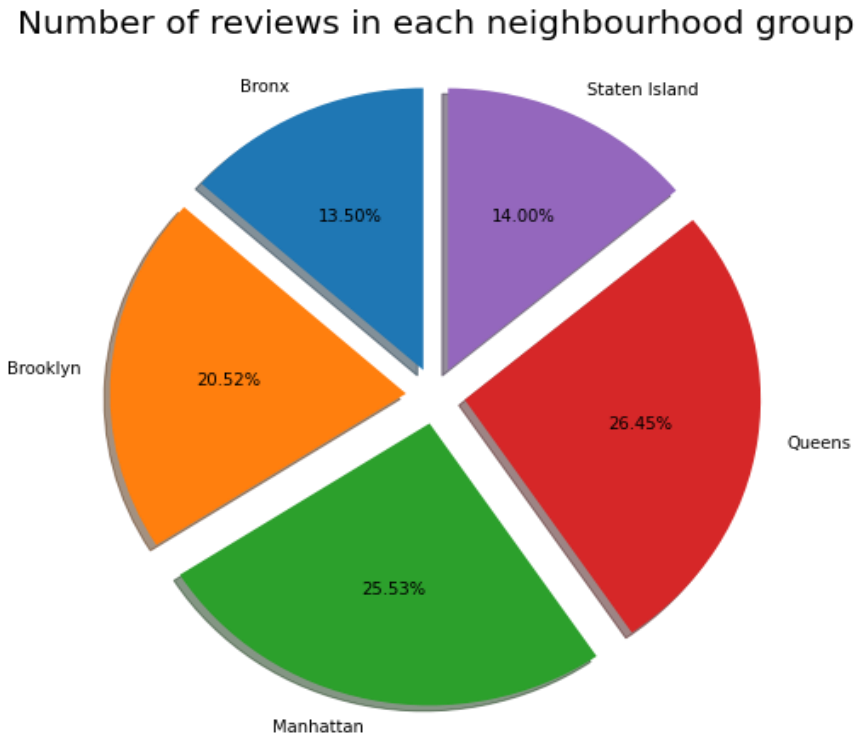
Manhattan generates the maximum revenue among the neighbourhood group.

Maximum profit among the Neighbourhood



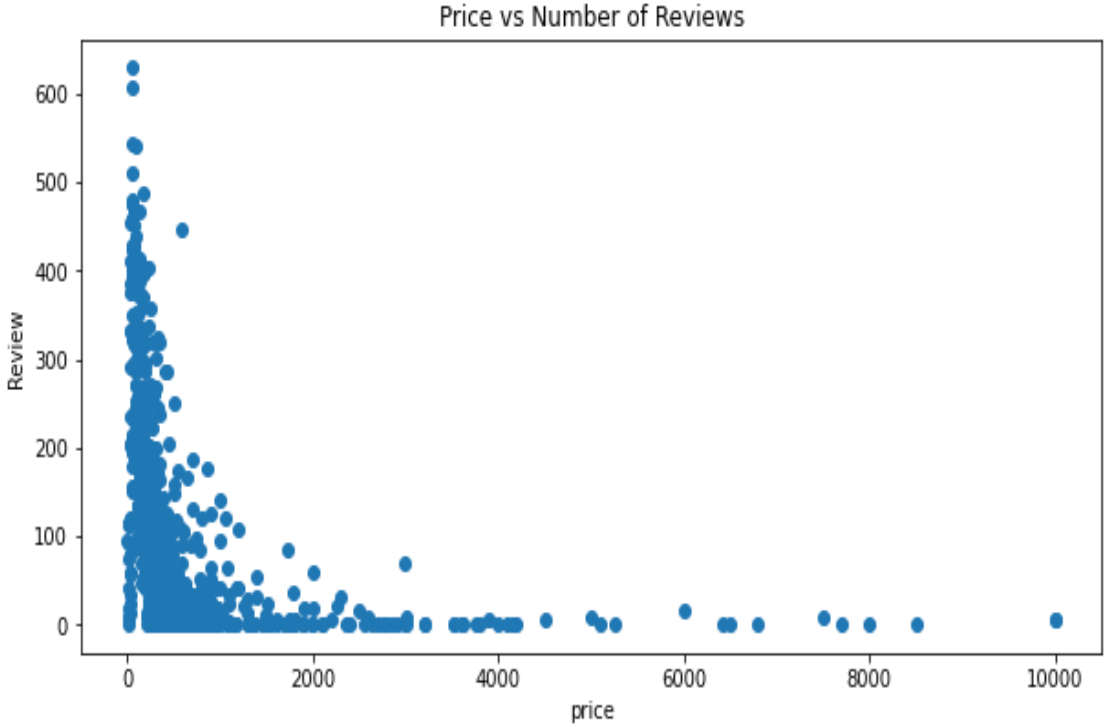
The neighbourhood that generates the maximum revenue is Midtown, which is also the most costly.

Number of reviews in each neighbourhood group



We can see that the bulk of evaluations are from Queens, Manhattan, which suggests that people like living in this area.

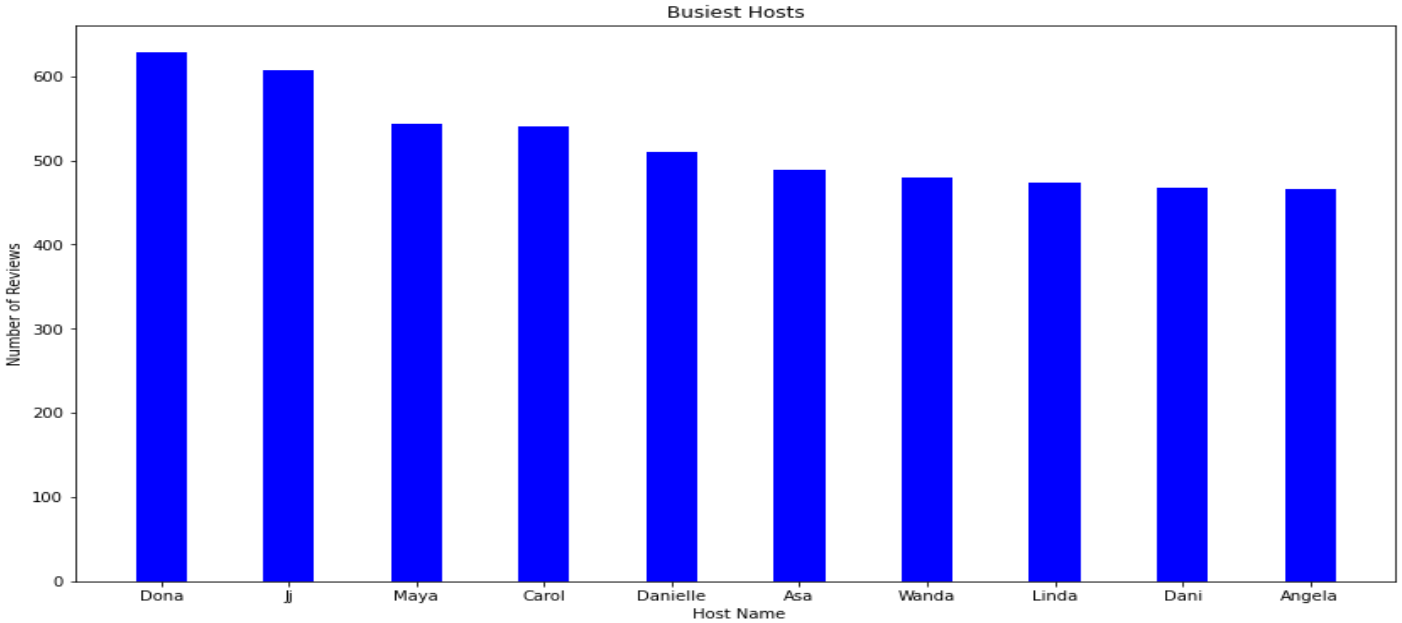
Relationship of price and number of reviews



Data analysis has revealed that consumers prefer reduced prices.

Price reduction leads to more positive reviews.

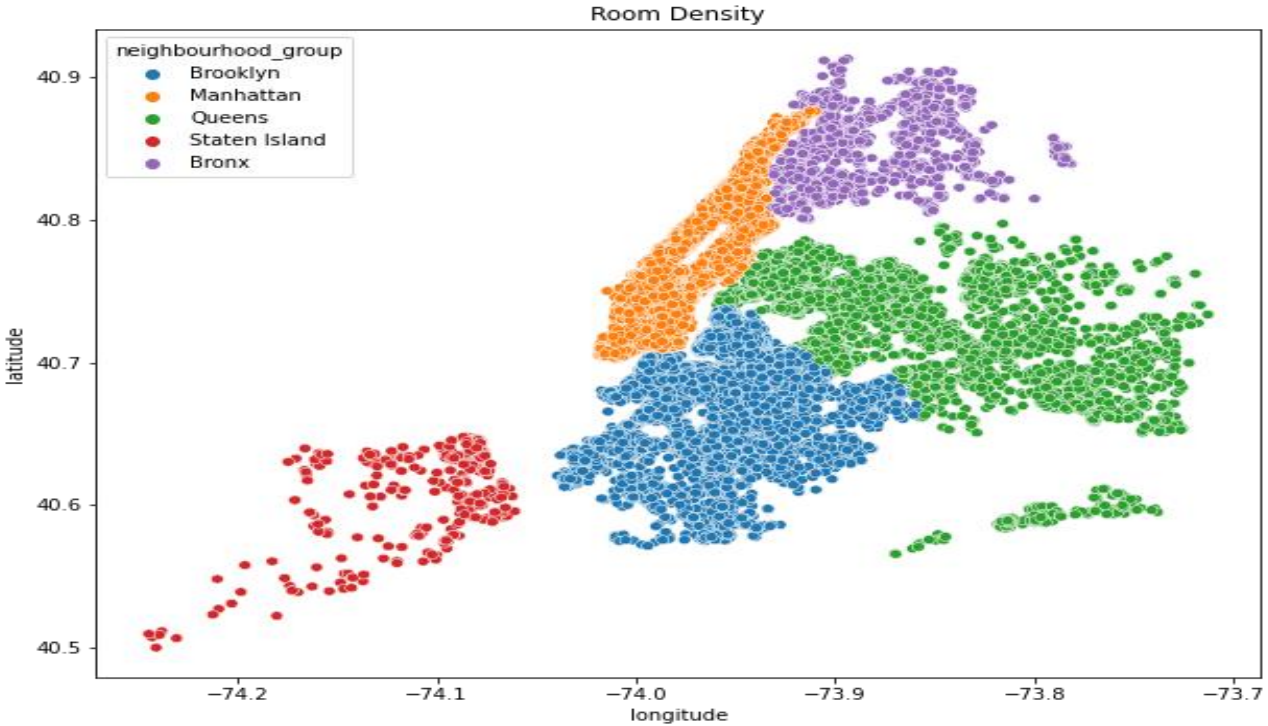
Analysis of Busiest Host



From above analysis we saw that Dona, Jj, Maya, Carol, Danielle are the busiest hosts.

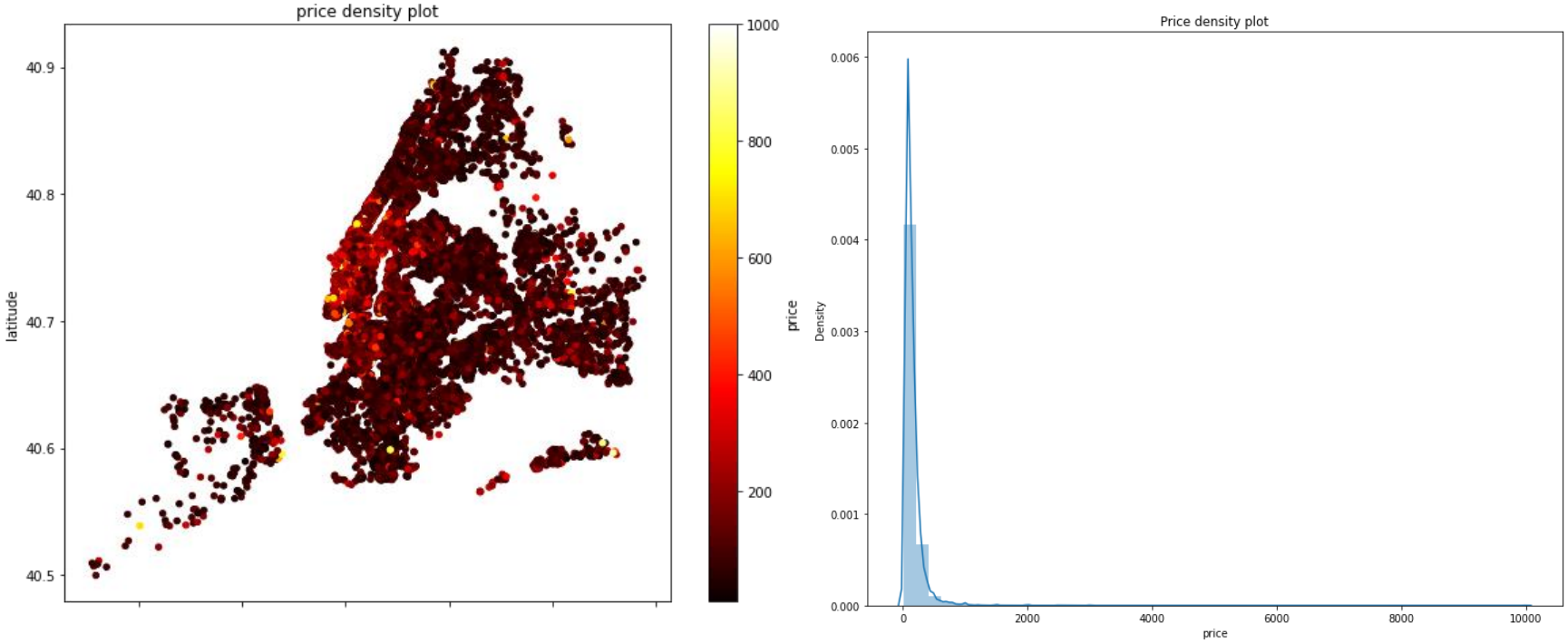
Previously, we knew that the majority of hosts in Manhattan and Brooklyn offered the most popular room types. The relationship between the number of reviews and the yearly availability of these rooms is shown in the bar plot. As a result, neighbourhood groups in Manhattan and Brooklyn get the most positive feedback, which makes their hosts the busiest of all.

Analysis of Traffic among Areas



Manhattan has a high hotel density, especially in areas around the Manhattan Financial District.

Available rooms within price range



These graphs makes it quite evident that density depends on the price rating. Density rises when the price is Minimum.

CONCLUSION

1.Manhattan and Brooklyn are the most profitable location of airbnb of New York City in compare to Staten Island, Queens and Bronx, if they increse the facilities and number of hosts in Manhattan and Brooklyn then possibly it would generate more revenue.

2.more then 80% hosts are in from Manhattan and Brooklyn.

3.Most of the customers prefer rooms with reasonable price.

4.Customers prefer to stay in Entire/apt or private rooms as compare to shared rooms so airbnb should focus to make more available entire apartment and private rooms.

5.Sonder (NYC) is the most successful host with listings of 327.

6. Density of hotels are pretty much high in Manhattan and it's nearby area.

7.Airbnb should publicity more about the famous location/sites of Bronx, Queens and Staten Island to attract the tourists and it will help them to earn more revenue from these regions as well.

Thank You

tt