CAPSTONE PROJECT-1

AIRBNB BOOKING ANALYSIS(EDA)

Team Members:

Alok Singh

Manisha Solanki

Harshit Sharma

TABLE OF CONTENTS:

INTRO-DUCTION

Airbnb, as in "Air Bed and Breakfast“ , is an online market that lets property owners rent out their spaces to travelers looking for accommodation. Travelers can rent a space for multiple people to share, a shared space with private rooms ,or the entire property for themselves.. Not only to guests but Airbnb is a portable platform to hosts also. Hosts can list their properties on Airbnb by going through certain validation process and rent their property to needy ones. NYC is the most populous city in the United States, and one of the most popular tourism and business place globally.

We are provided with the data of NYC to analyze the data and find out certain points regarding the data.

STEPS FOLLOWED IN ANALYSIS

Data collection: We collected the Airbnb data on which EDA is to be done and then get an overview of the Dataframe.

Data cleaning: we will clean the data by removing null values or replacing them. Then we will check data type and convert required data type.

DATA SUMMARY:

1.Data set name – Airbnb Booking Analysis.

2.Shape – 48895 rows * 16 columns

3.Column description – There are total 16 columns containing - ' id', 'name', 'host_id', 'host_name', 'neighbourhood_group', 'neighbourhood', 'latitude', 'longitude', 'room_type', 'price', 'minimum_nights', 'number_of_reviews', 'last_review', 'reviews_per_month', 'calculated_host_listings_count', 'availability_365′ .

Features description:

- **Column I’d and Name**: these columns are telling us the property id and property name. name can we same that’s why unique value called I’d has been given to us.
- **Host_name and host_i’d**: it is telling us property owner’s name and unique I’ds.
- **Neighbourhood groups**: giving information about the regions where properties are available.
- **Neighbourhood**: they are telling us about nearby localities of the properties.
- **Latitude and longitude:** giving the exact location of properties geographically.
- **Room_type:** telling us about the types of rooms that are provided by the various hosts
- **Price**: giving the prices of different properties with different room types

- **Minimum_nights** : giving the booking criteria provided by host as the guests have to book certain property for atleast how many number of nights.
- **Number_of_reviews** : So, the reviews that Airbnb took include a lot of variants including security, cleanliness, locality, easy to locate, guests friendly and so on. So this column can give us a full fledged idea of guests preference.
- **Last_reviews, reviews_per_month** :telling us about the last time a particular property visited by any guest. And how many visits a particular hosts are having in a month.
- **Availability_365** : telling us about the number of days a particular property is available to rent out.

PROBLEM STATEMENTS:

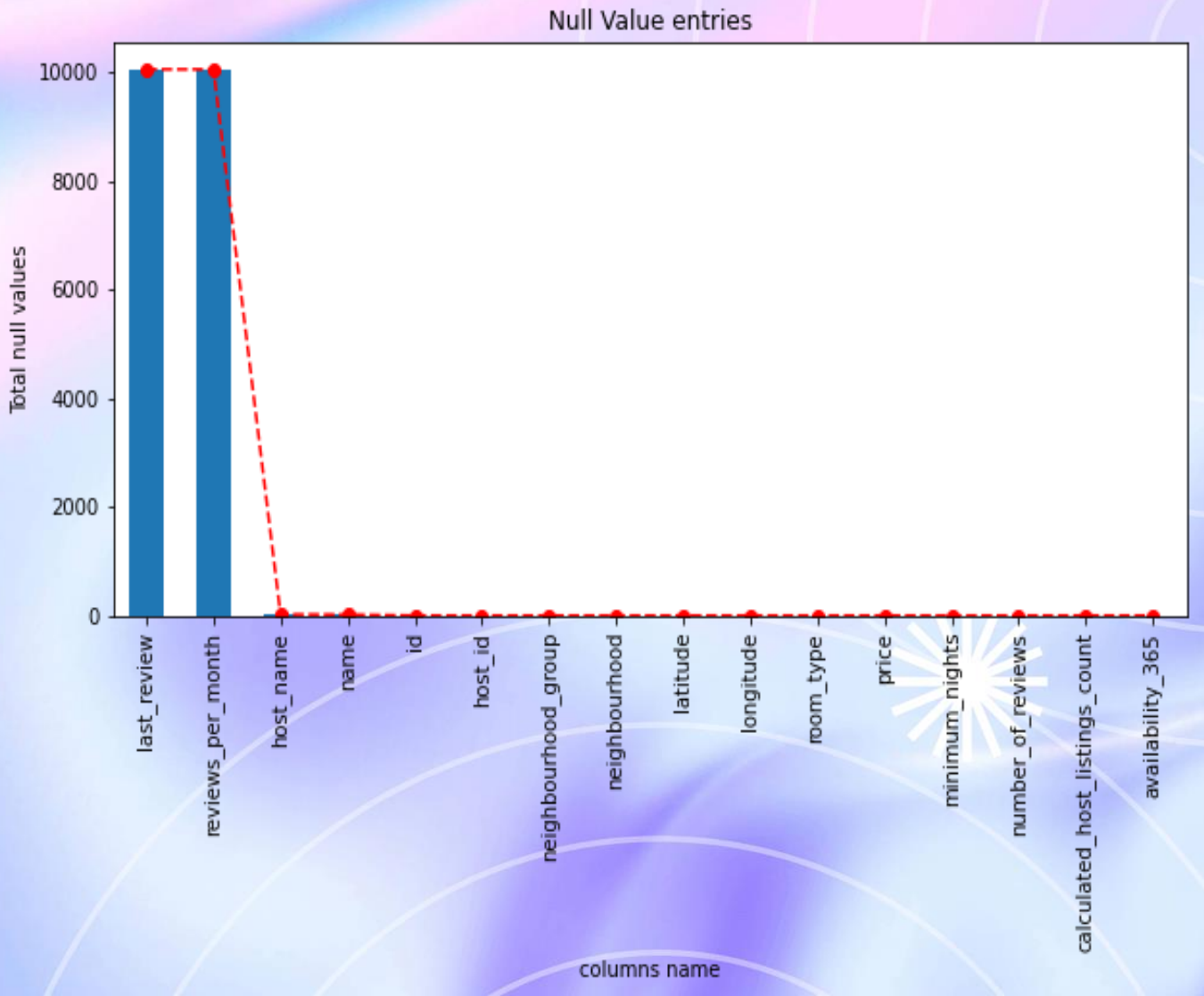
Our problem statement was to find the behaviour of guests for booking a property in New York City and to provide them with better and more options as per their demand with better services. For that we have to find:-

1. Relation between different hosts and different areas.
2. Finding the guest behaviour with prices, type of rooms and availability and locality.
3. Which locality have the expensive and cheap properties and guests preferences with them.
4. Which are the busiest hosts and why?
5. Is there any relation of the properties with the traffic in that particular locality.

CLEANING THE DATASET:

- Finding the null values in each column if any available.

- We can see only a few columns had null values.

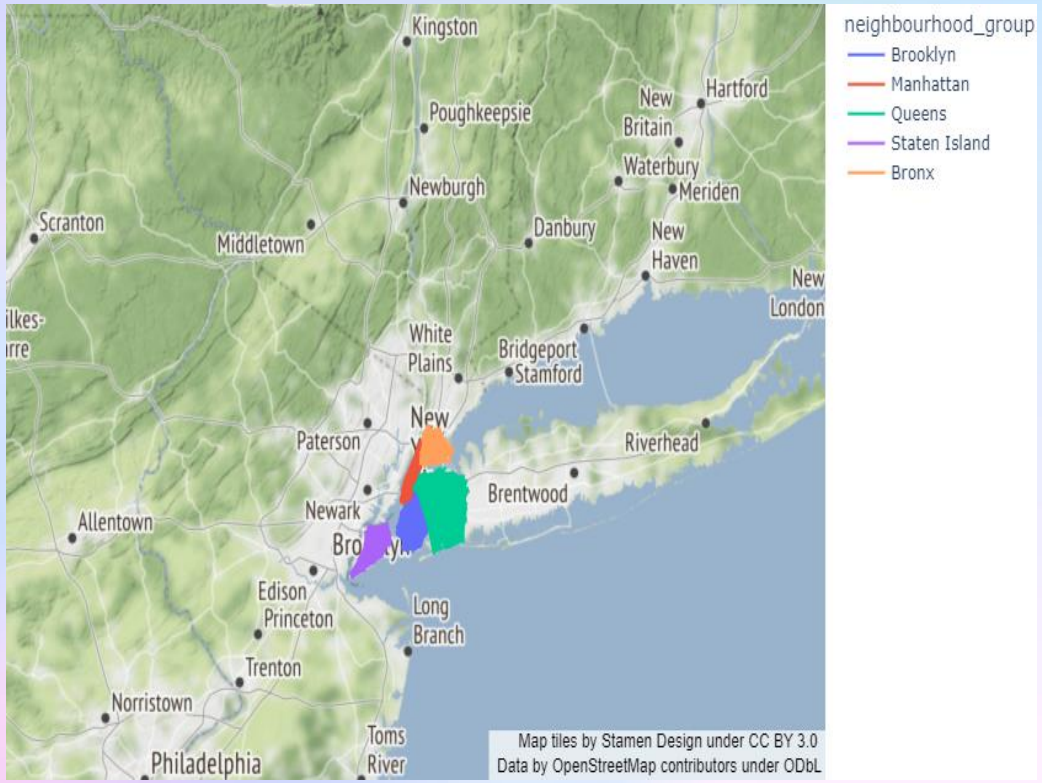


- Column ‘last_review’ and ‘reviews_per_month’ had a lot of null values and there were no other details regarding the same. So, we dropped those columns.
- Column ‘Name’, also had some null values and I’d of the properties are provided as unique value, So, we dropped the ‘Name’ column also.
- Column ‘host_name’ have 21 null values and host I’ds are also available so we replaced null values with ‘NOT AVAILABLE’.
- We can see that Price column have some 0 values but Price for a property can not be zero, that means it is a mistake. And hence we replace all the zero values with the mean.

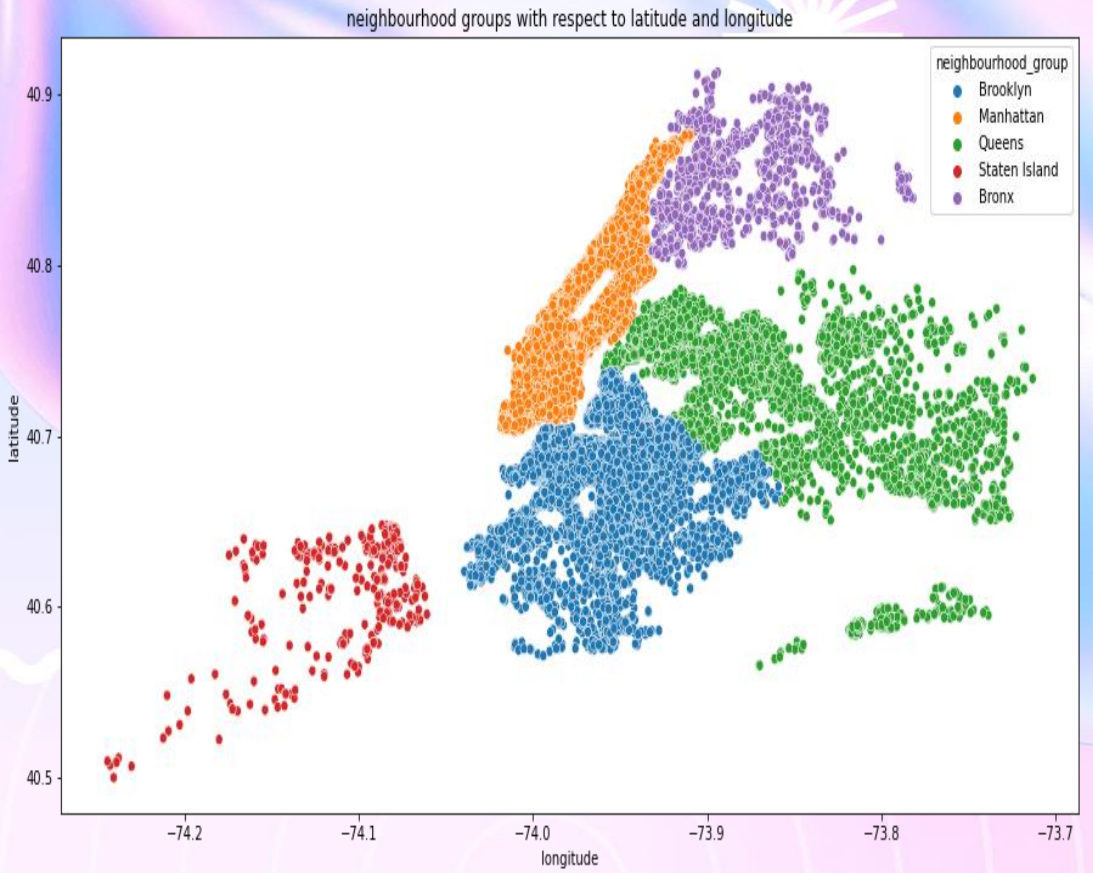
Mapping the neighbourhood groups

- From graph 1,we can see that there are five neighbourhood groups whose data we have to study.
- From graph 2,we can see the hosts available in these neighbourhood with respect to latitude and longitude.

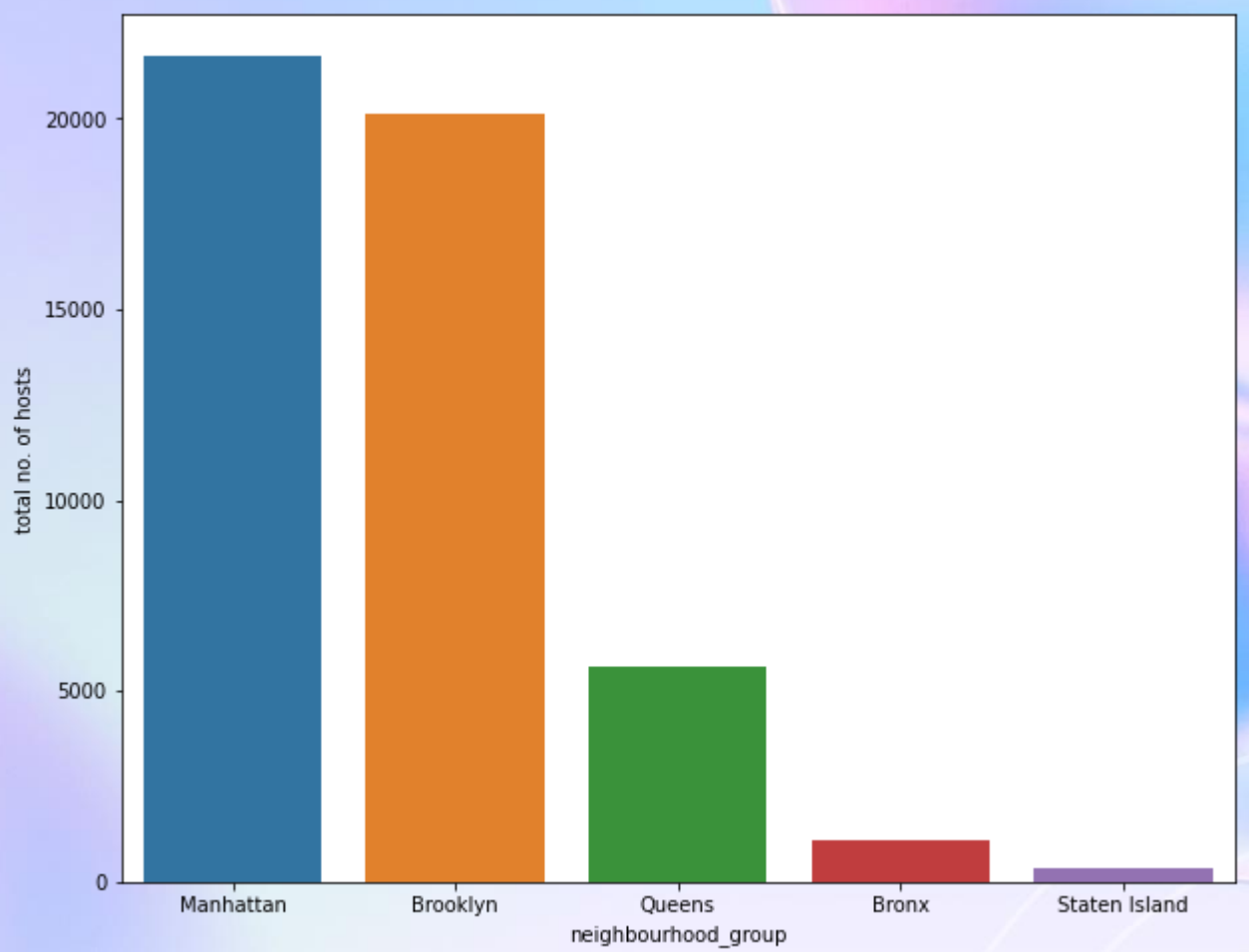
Graph 1



Graph 2



Number of hosts available in five neighbourhood groups

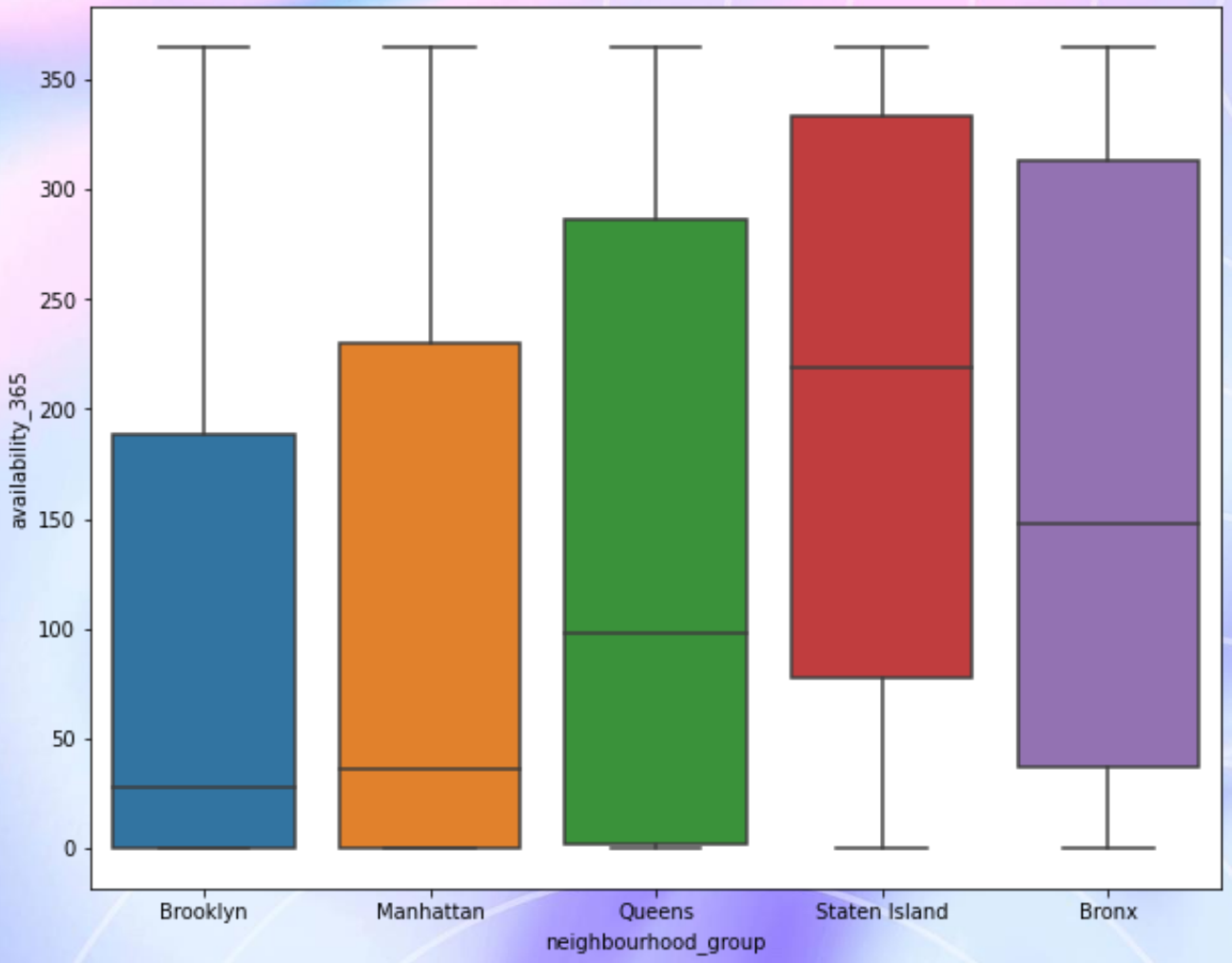


We can see that more than 22000 hosts are marked in Manhattan and around 20000 are marked in Brooklyn, while Queens have approx. 5000 hosts and Bronx and Staten island combined have total 300 hosts approx.

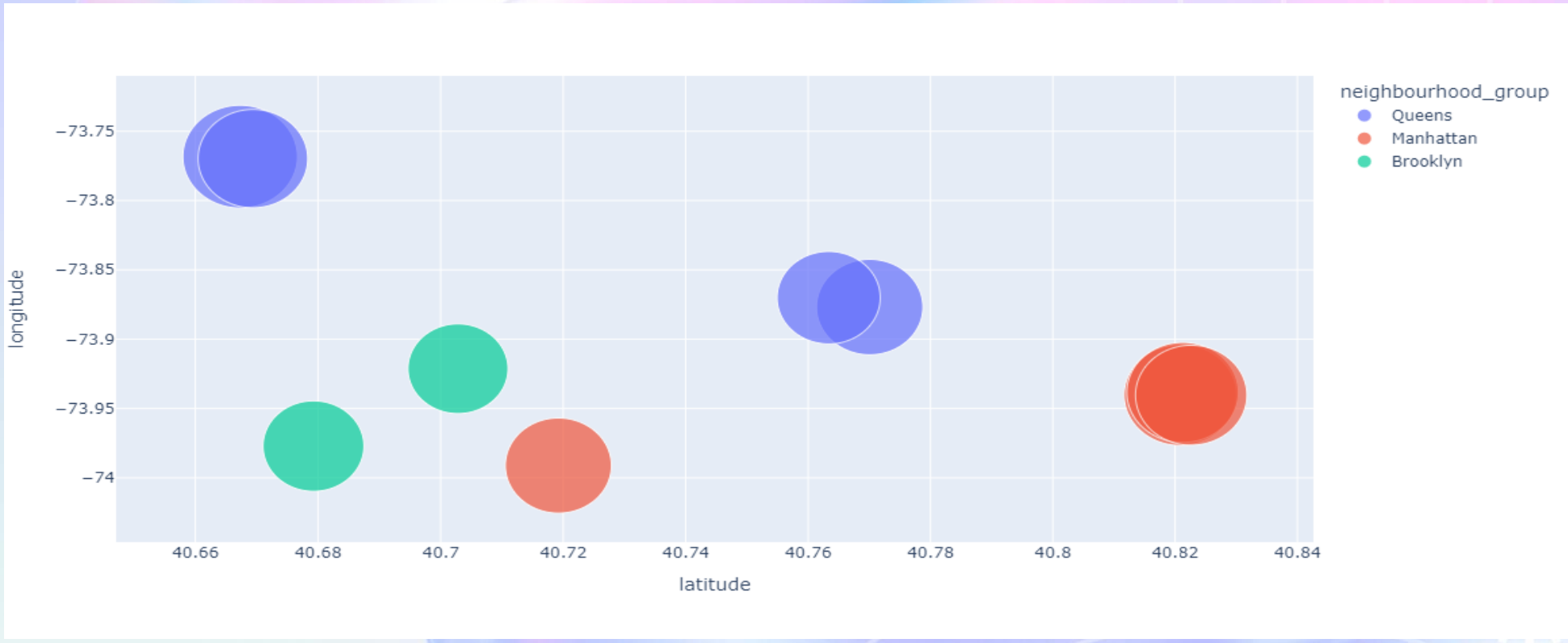
Neighbourhood groups available for how many days

Now we have created a boxplot to find out availability of different hosts in these neighbourhood groups.

- We can see that most hosts are giving availability in Staten Island and Bronx irrespective of the number of hosts, which are higher in Manhattan and Brooklyn.



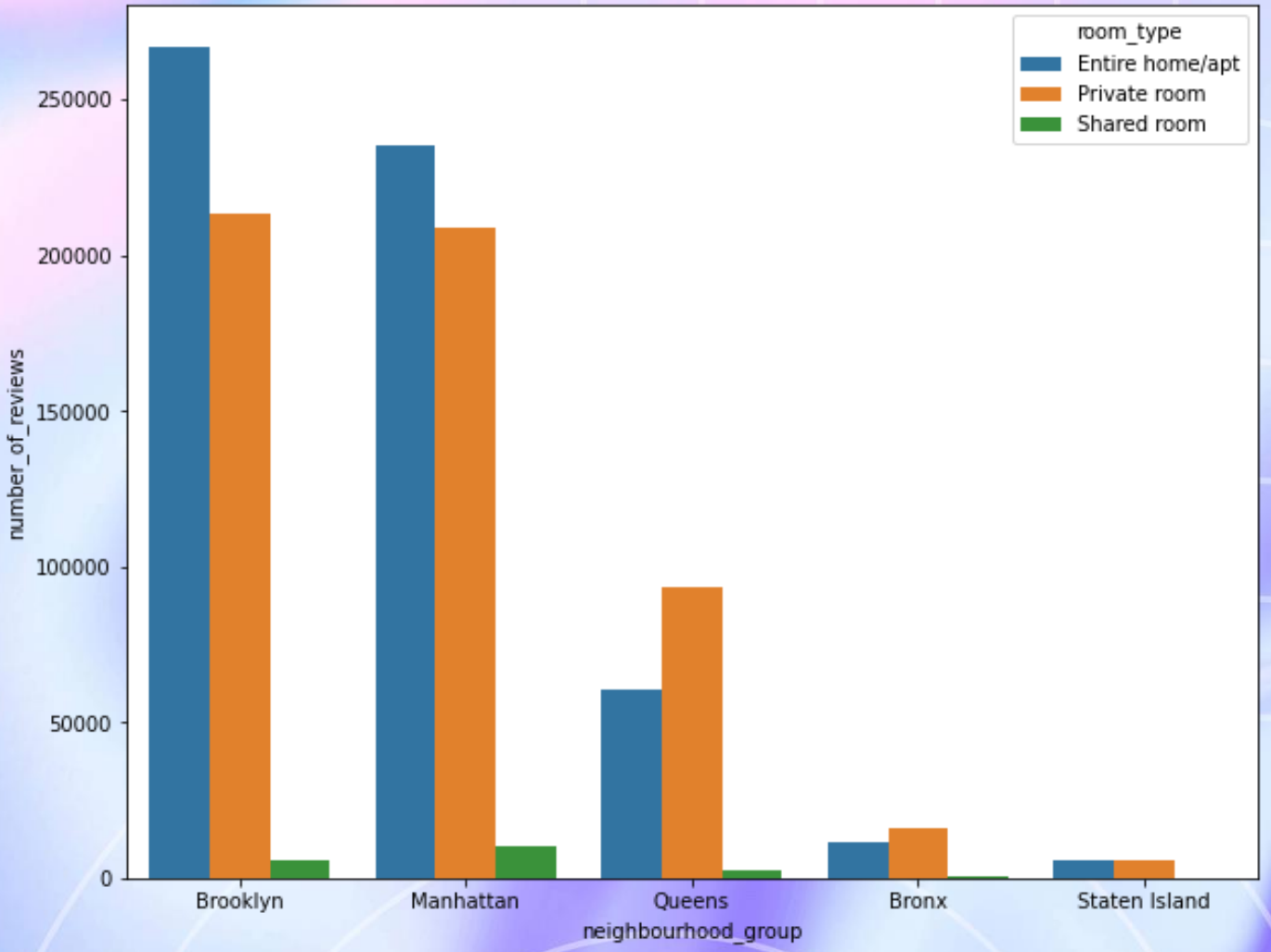
Now to check the guests preference of areas on the basis of reviews, we plotted a scatter chart of top 10 locations preferred by guests with respect to latitude and longitude.



So we can see that guests are also preferring Manhattan and Brooklyn over other neighbourhood groups

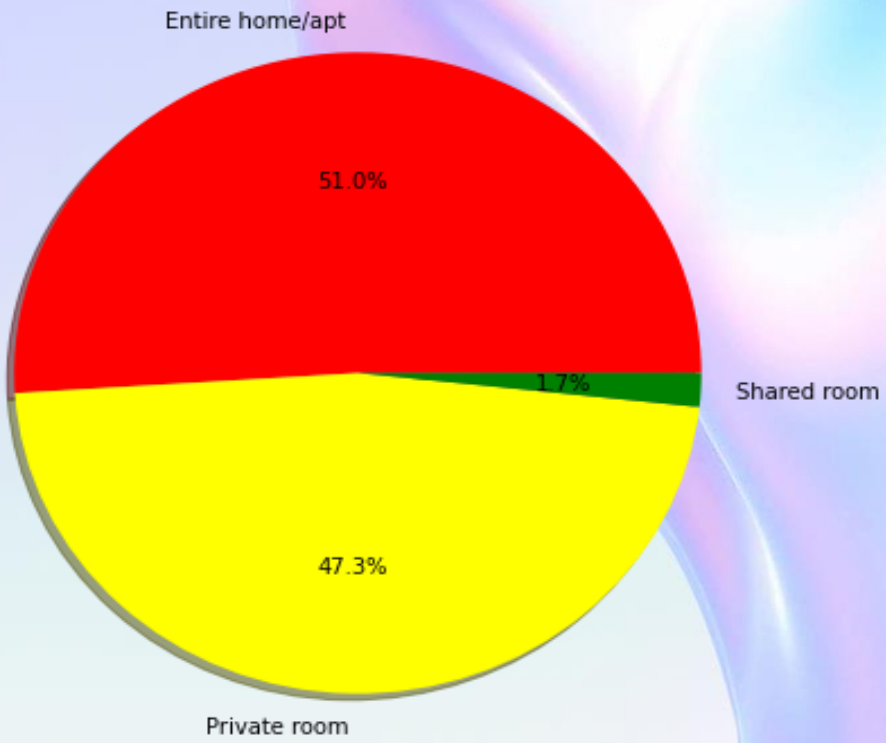
We are trying to check what type of rooms and which area is most preferrable by the guests.

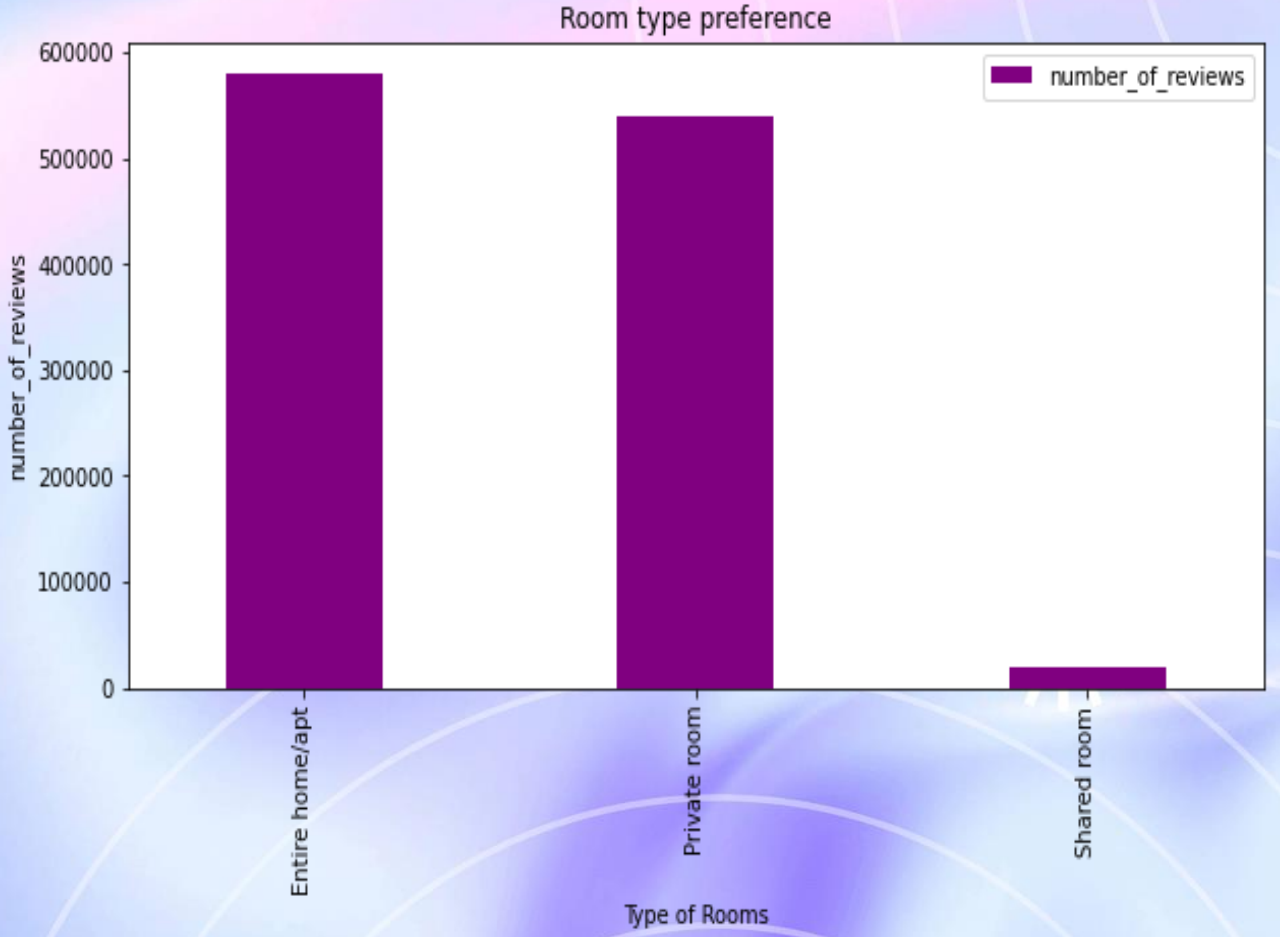
- Three types of rooms are available and as per the number of reviews, guests are preferring Entire home/apt and private rooms more.
- And guests are also preferring to stay in Manhattan and Brooklyn more.



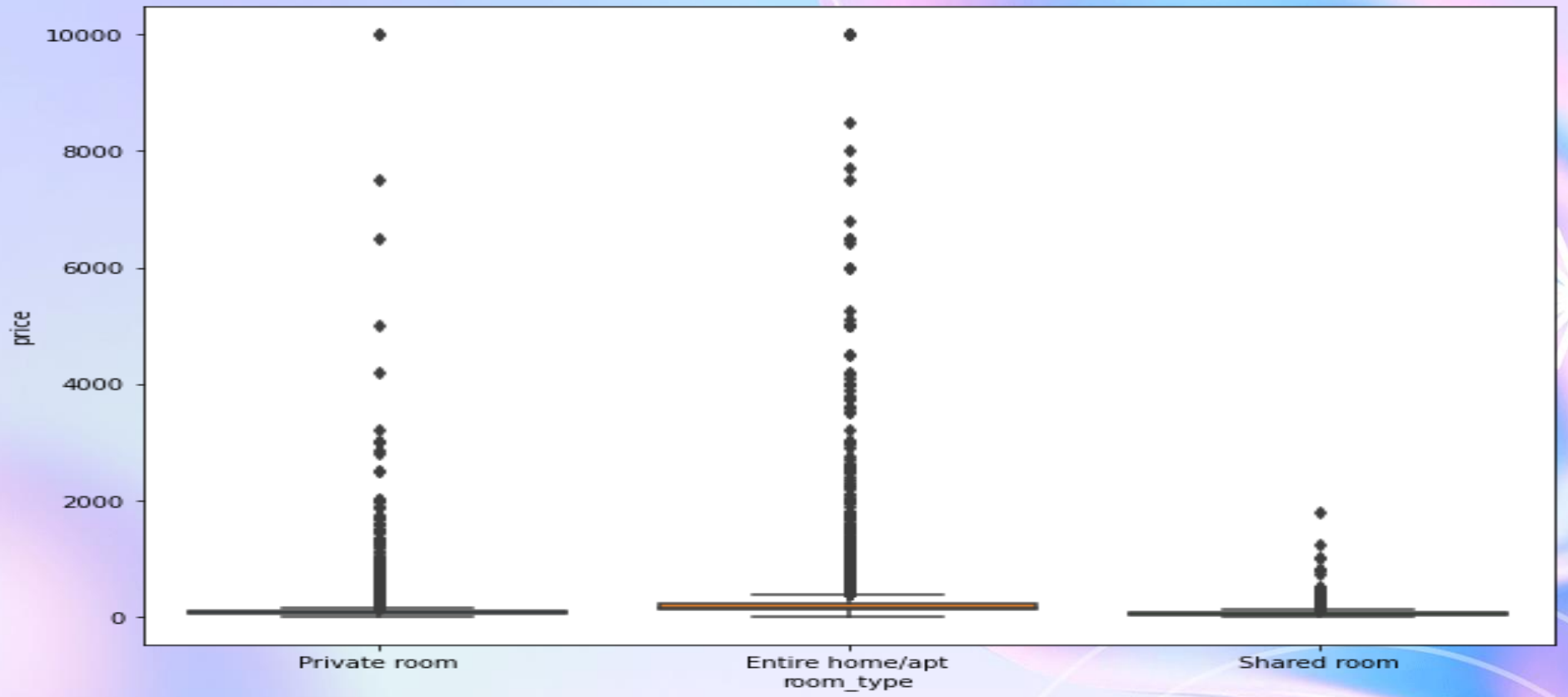
Here, we have plotted a pie chart of number of reviews for different types of rooms

- Room types reviewed by guests showing that 51% of the guests are preferring entire home and 47.3% are preferring private rooms.



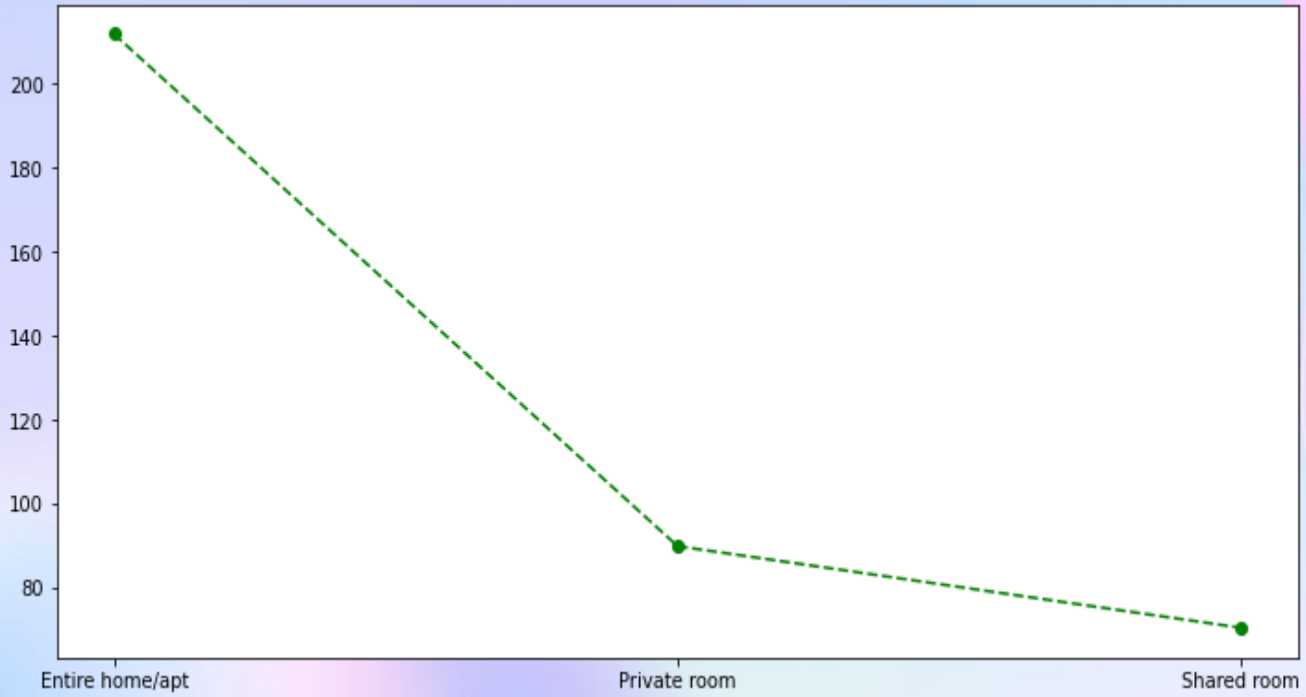


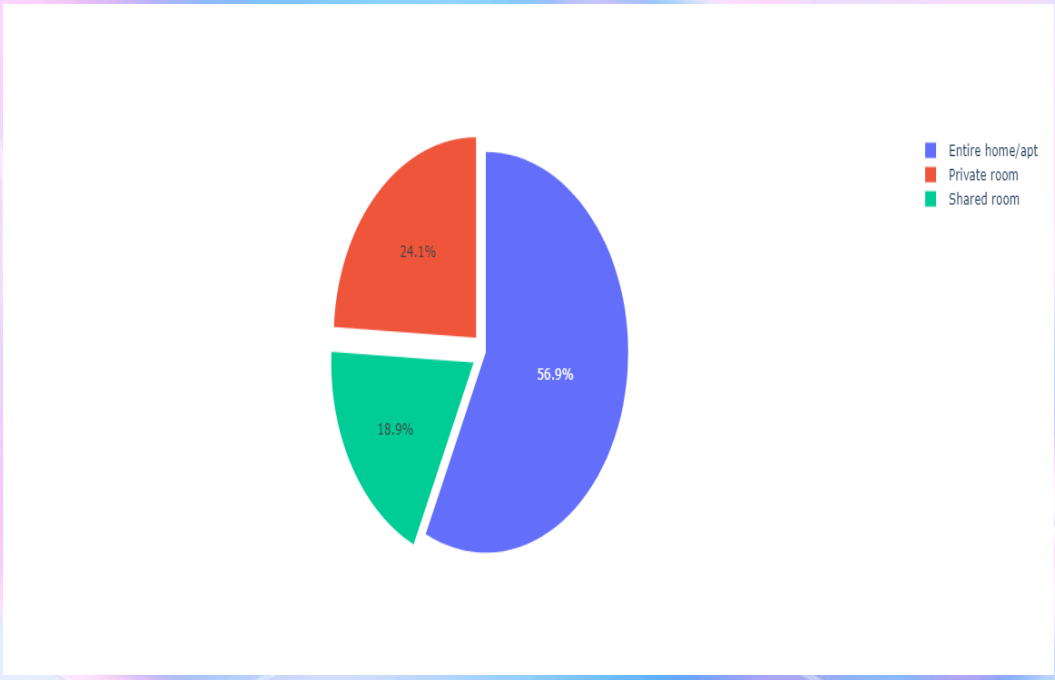
Price ranges for room types



So most of the private rooms are in the range of 2000 and entire home/apt are in the range of 3000 while shared rooms are within 500. From the above data it is clear that visitors are preferring private room and entire home more irrespective of the price range.

WE HAVE CALCULATED AVERAGE PRICE FOR DIFFERENT ROOM TYPES

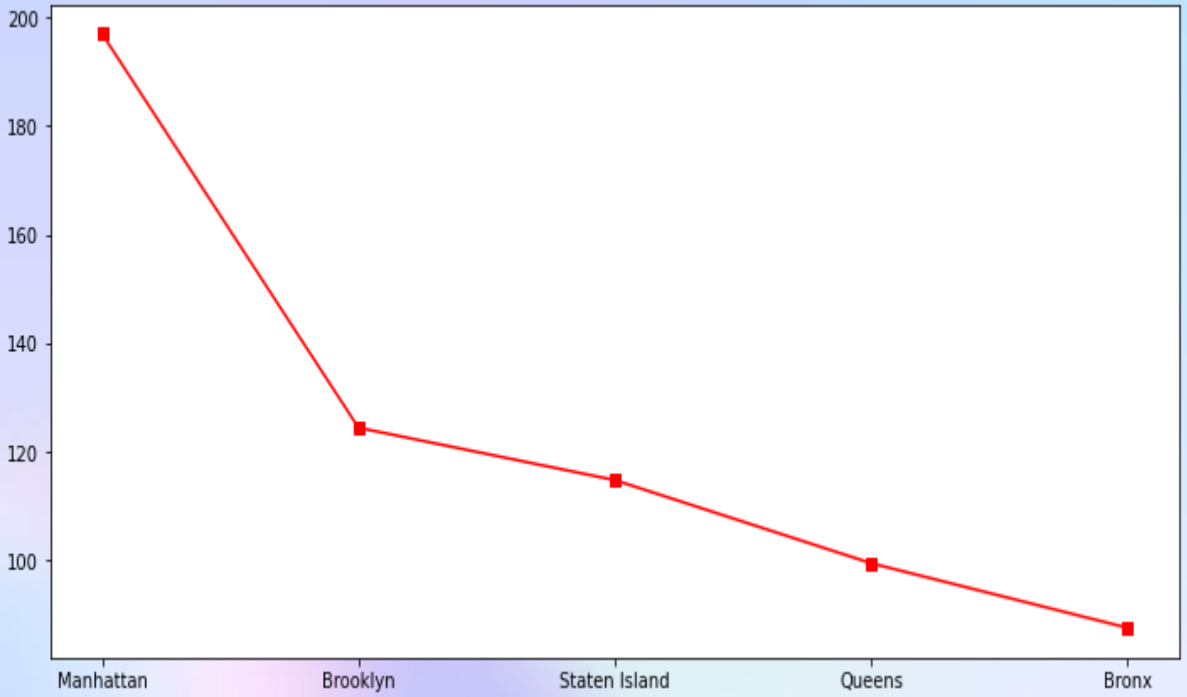


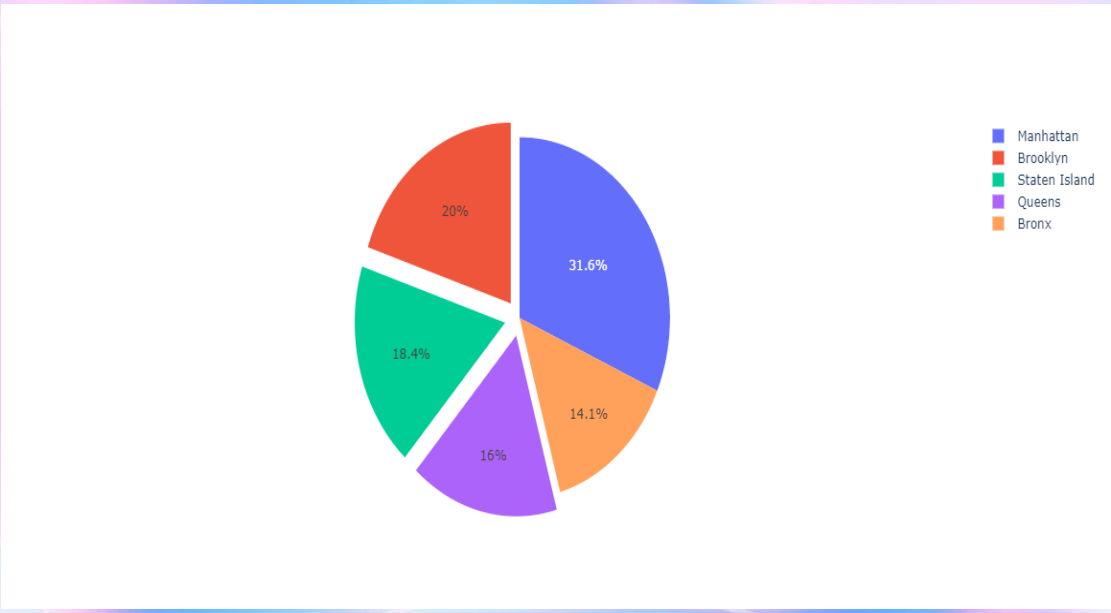


So average pricing is around 250 for entire homes and comparatively quite low for private rooms and shared rooms.

From the pie chart, we can clearly conclude that average price for entire homes is covering 51% of the total price range.

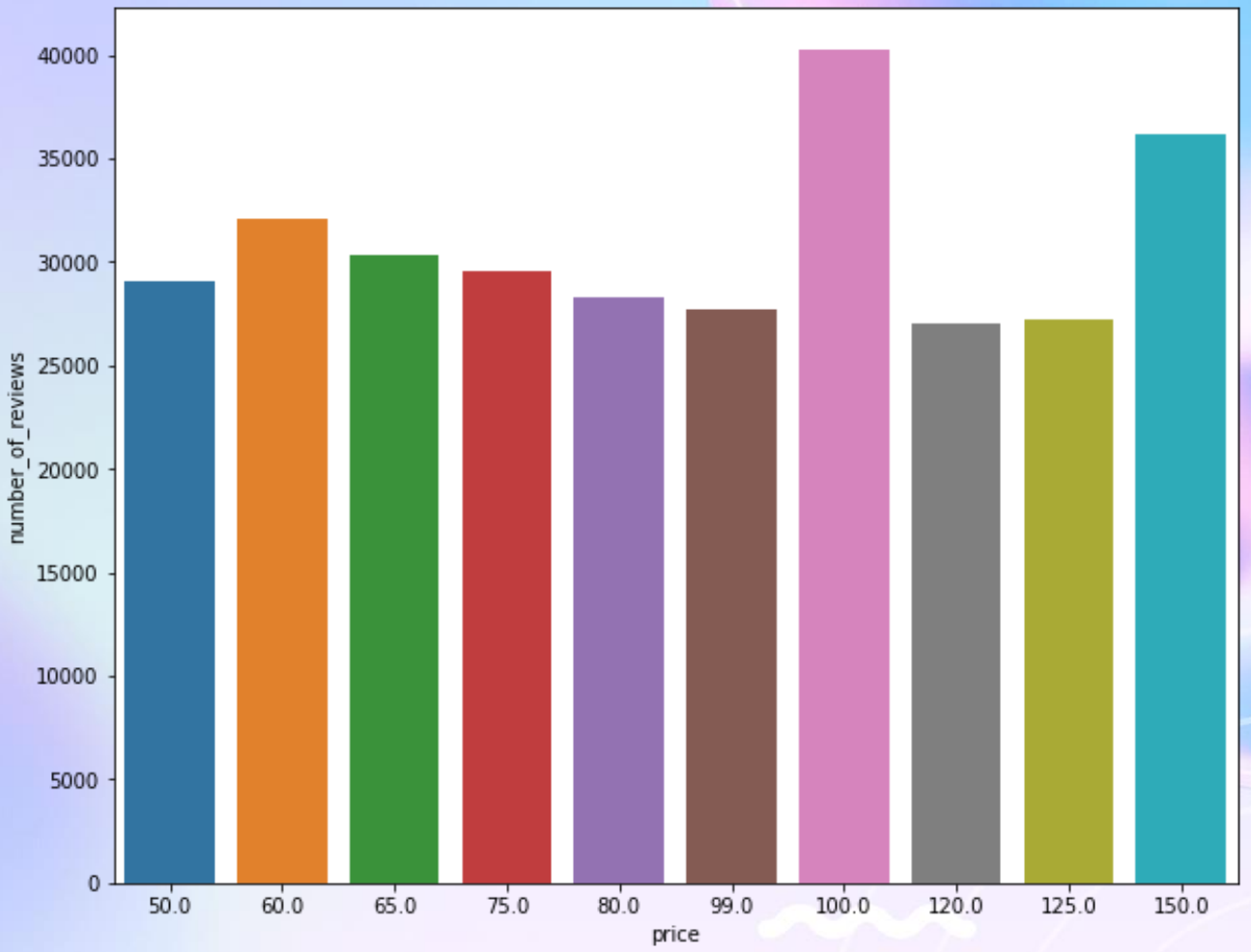
Average pricing with respect to neighbourhood groups





We can see that average price is higher for Manhattan, it is around 200 in Manhattan, that means guests can find accommodation in Manhattan for price of 200,while in Brooklyn, accommodation is available for average price of 130 and so on. The cheapest accommodation is available in Bronx.

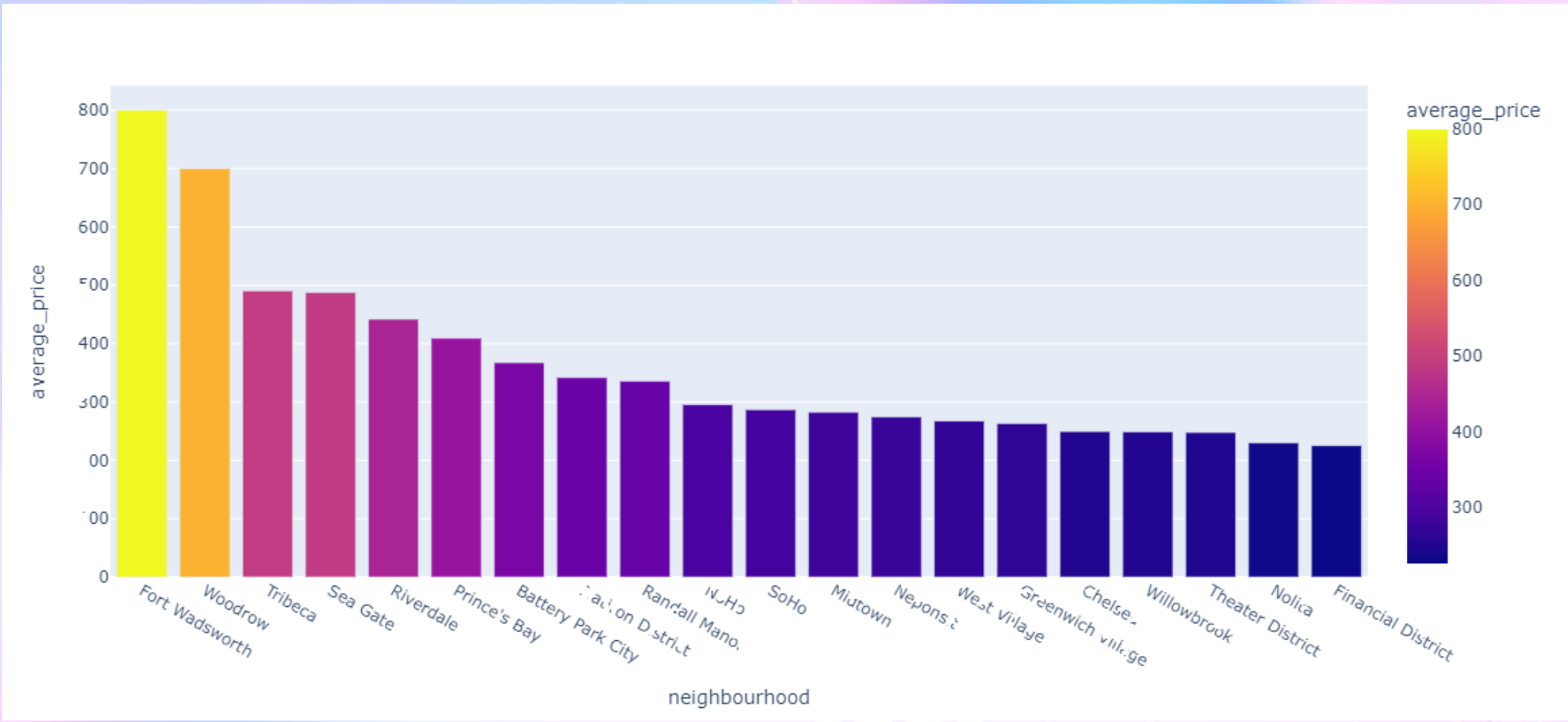
Prices preferred by guests based on the 10 highest reviewed properties:



We can see that visitors have booked more rooms for price range 50-150 .

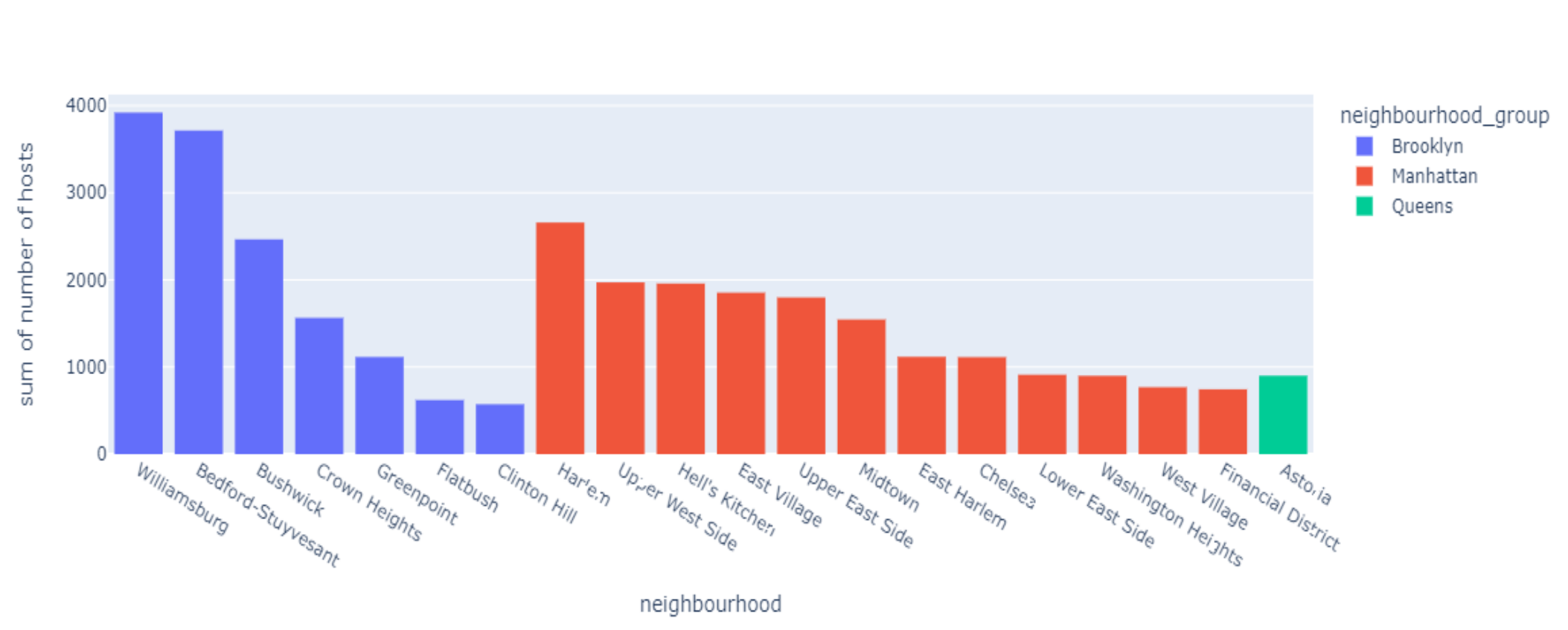
Most Expensive Neighbourhoods:

We plotted the chart of top 20 neighbourhoods with respect to the average price



Most expensive neighbourhood is Fort Wadsworth with average price of 800 and follow on.

Plotting the graph of hosts availability in different neighbourhoods

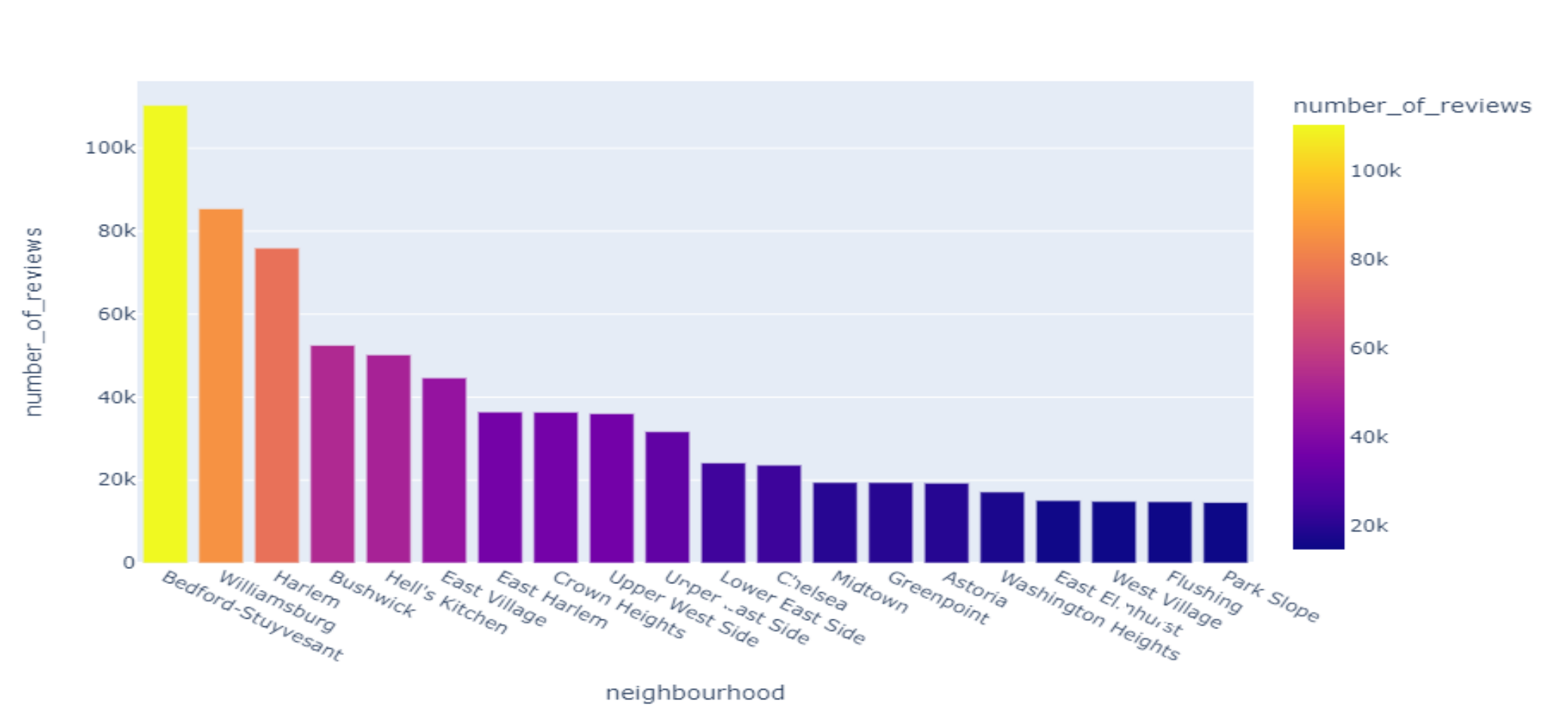


We can see that neighbourhoods are less in Brooklyn but number of hosts are maximum in Brooklyn. This makes Brooklyn the busiest area, and these localities will have the highest traffic in entire New York City.

Neighbourhood that consists least number of hosts

We can notice that the neighbourhoods with least number of reviews are the same neighbourhood that have highest average pricing. This means due to high pricing hosts are not available in those areas.

Top 20 neighbourhood with most number of reviews



Neighbourhood preferred by guests are also not from the highest average price neighbourhoods. That means guests are looking for localities with less prices.

CONCLUSION

1. Manhattan and Brooklyn are the most preferred area by both hosts and guests.

2. There are 3 types of rooms available but most of the visitors are preferring entire home/apt or private rooms. Meaning visitors are either tourists or coming for business purposes. Shared rooms are not preferred much. so we can also say that these neighbourhood groups are not student area.

3. Price ranges only matters for selection of neighbourhoods, when selecting the neighbourhood groups guests are not that concern.

4. Due to less number of neighbourhoods but highest number of hosts available in Brooklyn, we can say that Brooklyn is the busiest area and you can face traffic there.

5. Availability is higher in Staten Island and Bronx, but still not the first choice of visitors, So availability has not much role to play while selecting the room.

6.In the end we can say that there is a high probability of booking happening for Manhattan and Brooklyn in the price range of 50-150. and there is a high probability of new hosts opening properties in these areas.

SUGGESTIONS

- Our first suggestion would be to increase the number of properties in expensive neighbourhoods with lower prices.
- By doing so, we can reduce clustering of visitors in busiest neighbourhood and hence can reduce traffic in some neighbourhoods.
- Lowering the prices in expensive neighbourhoods and increasing number of properties in these neighbourhoods, guests will have more options to choose over Airbnb and hence it will improve visitors Airbnb experience.
- Another suggestion would be to change shared rooms in private rooms and offer them to visitors on the price of private room only. By doing so, their shared rooms won’t be left empty and earnings will be more.

THANK YOU