



STORY TELLING CASE STUDY : AIRBNB

Presented By :
SONIA DHANDA
SHANTAM GARG
SUDHANSHU
MISHRA



EXPLORATORY DATA ANALYSIS ON AIRBNB DATA SET

**UNDERSTANDING DATA & PROVINDING
INFORMATION TO : DATA ANALYSIS MANEGER & LEAD
DATA ANALYST**



AGENDA

Key Findings :

- To provide information – Understanding Data
- To explain the process – Exploratory Data Analysis
- To share challenges & insights

Appendix :

- Data Assumptions
- Data Source
- Data Dictionary
- Data Methodology
- Data Visualizations



OBJECTIVE :

- Here we are going to do an Exploratory Data Analysis on the data set of Airbnb NYC (2019).
- Our main objective is to understand critical pre-COVID period data & make insights from the Airbnb NYC business.
- Make recommendations to concerned department on how to prepare for post pandemic changes.
- Improve our shared understanding about our customers.
- Provide early recommendations to our marketing and operations teams.



BACKGROUND :

What is Airbnb ?

Airbnb is a marketplace where travelers get to stay at the property listed on the site by the local hosts. Airbnb is a community built on sharing. It is based on an aggregator business model where the company collects the data about a specific good/service providers, make the providers their partners, and sell their services under its brand.

Problem Background :

For the past few months, Airbnb has seen a major decline in revenue due to COVID situation. Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.

So, analysis has been done on a dataset consisting of various Airbnb listings in New York.

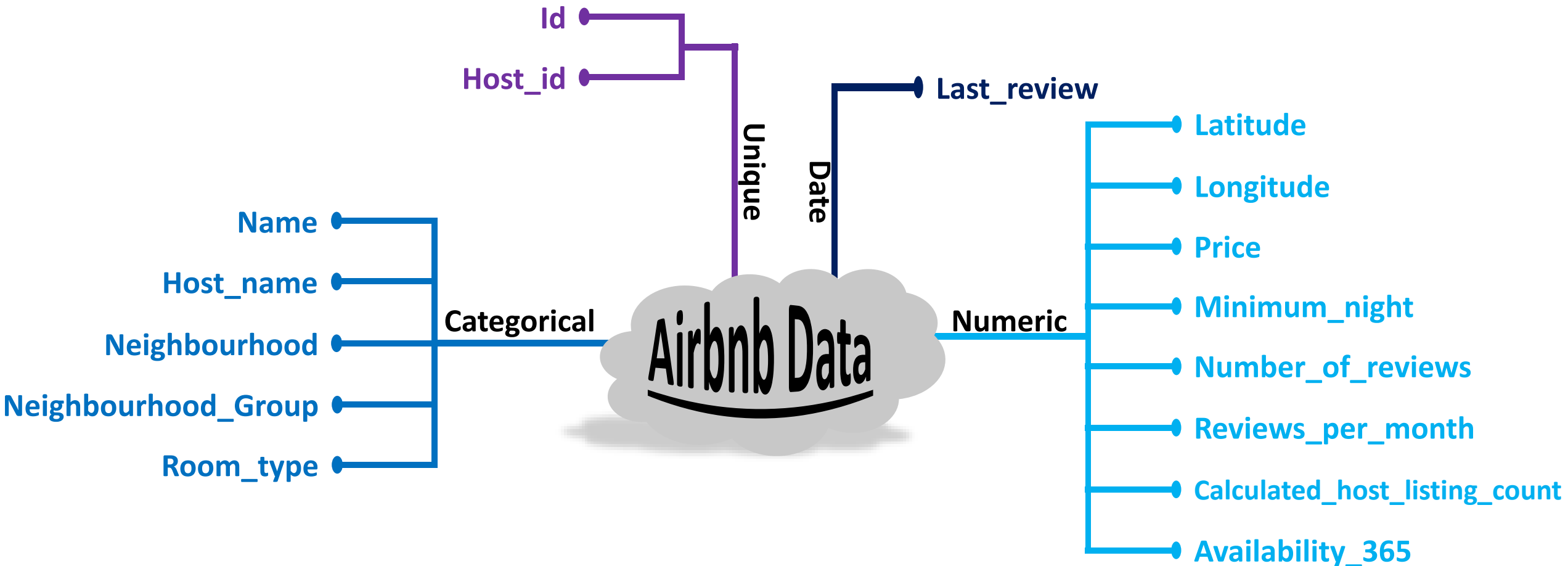


ABOUT THE DATASET :

- Lets have the overview of the basic understanding of our dataset variables.
- This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.
- Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other hand, Airbnb enables travelers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world.
- The data we are going to analyze is the data of Airbnb NYC (2019). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas, price, reviews, locations etc. but not limited to, we will also try to explore some more insights.



Understanding Data :





Understanding Data Variables :

- ❑ **ID:**
 - It's a unique id for House/apartment.
- ❑ **Name:**
 - Name of the listing House/apartment.
- ❑ **Host Id:**
 - Host Id is the government approved id for each individuals Who rent their properties on Airbnb.
- ❑ **Host Name:**
 - Host names are basically the name of the individual or organization Who own a room/apartment on Airbnb website.
- ❑ **Neighbourhood groups:**
 - Neighbourhood groups are the cluster of neighborhoods in the area.
 - There are about 5 boroughs in the state.
- ❑ **Neighbourhood:**
 - When searching for accommodations in a city, guests are able to filter by neighbourhood attributes and explore layers of professional-quality content, including neighbourhood maps, custom local photography and localized editorial, details on public transportation and parking, and tips from Airbnb's host community.
- ❑ **Latitude:**
 - Latitude is the measurement of distance north or south of the Equator.
- ❑ **Longitude:**
 - Longitude is the measurement east or west of the prime meridian.



❑ Room type:

- Airbnb has 3 categories for types of space :
 - **Entire House/Apartment**
 - **Private room**
 - **Shared room**

❑ Price (\$):

- The total price (\$) of Airbnb reservation is based on the rate set by the Host, plus fee or costs determined by either the Host or Airbnb.

❑ Minimum_nights:

- Minimum night is criteria for booking that guest have to pay for book that House/room or apartment.

❑ Number_of_reviews:

- Number of review of each host submitted by guest.

❑ Last_review:

- Latest review submitted by guest as a feedback.

❑ Reviews per month:

- Number of review Host get per month.

❑ calculated_host_listings_count:

- Amount of listing per host.

❑ Availability 365:

- It is an indicator of the total number of days the listing is available for during the year.



Exploratory Data Analysis :

What is EDA ?

- In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

The following steps are involved in the process of EDA:

- Acquire and loading data
- Understanding the variables
- Cleaning dataset
- Exploring and Visualizing Data
- Analyzing relationship between variables



Approach used for EDA:

The approach we have used in this project is defined in the given format.

1. **Loading our data:-** In this section we just loaded our dataset in python notebook and read the csv file using pandas library.
2. **Data Cleaning and Processing:-** In this section we have tried to check for the null values and for some of the columns we have replaced the null values with the appropriate values with reasonable assumptions.
3. **Analysis and visualization:-** In this section we have tried to explore all variables which can play an important role for the analysis using Tableau. In the next parts we have tried to explore the effect of one over the other. In the next part we tried to answers our hypothetical questions.
4. **Future scope of Further Analysis:-** There are many apartments having availability as **0**, which means they might stopped their business, we can find the relation of neighbourhood with these apartments if we dig deeply, various micro trends could be unearthed, which we are not able to cover during this short duration efficiently. There are various columns which can play an important role in further analysis such as number of reviews and reviews per month finding its relation with other factors or other grouped factors can play an important role.



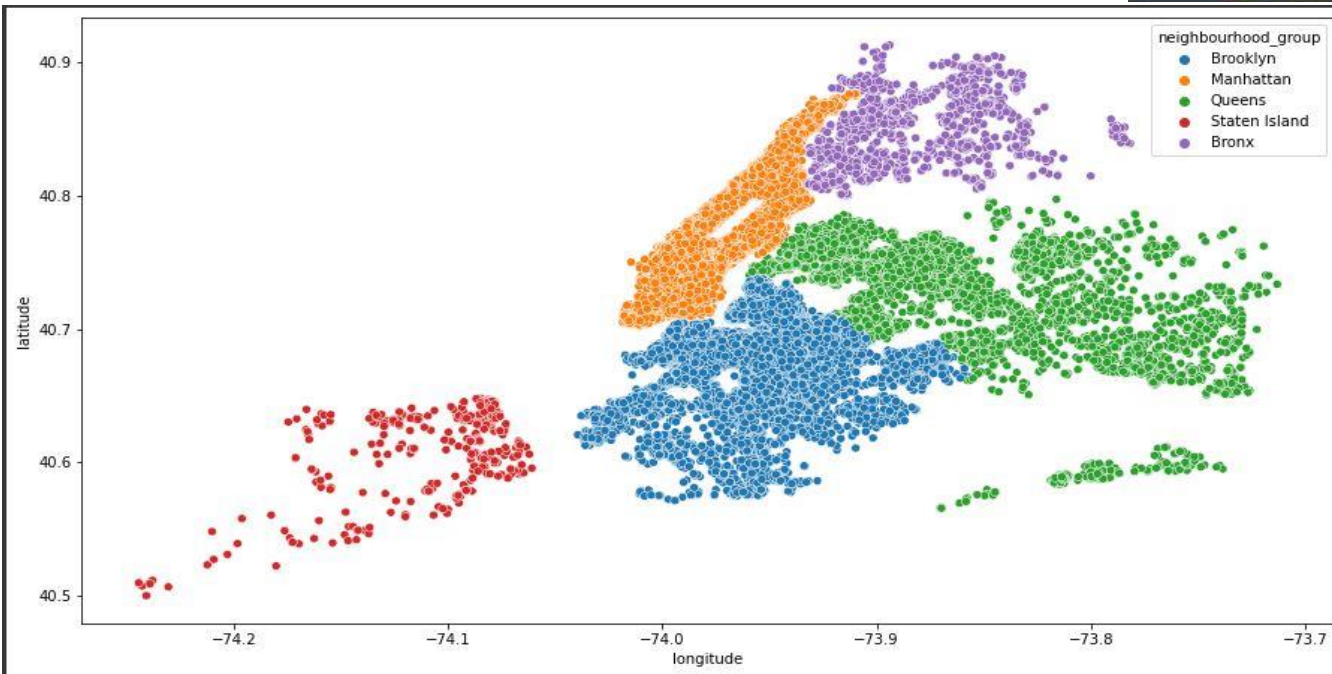
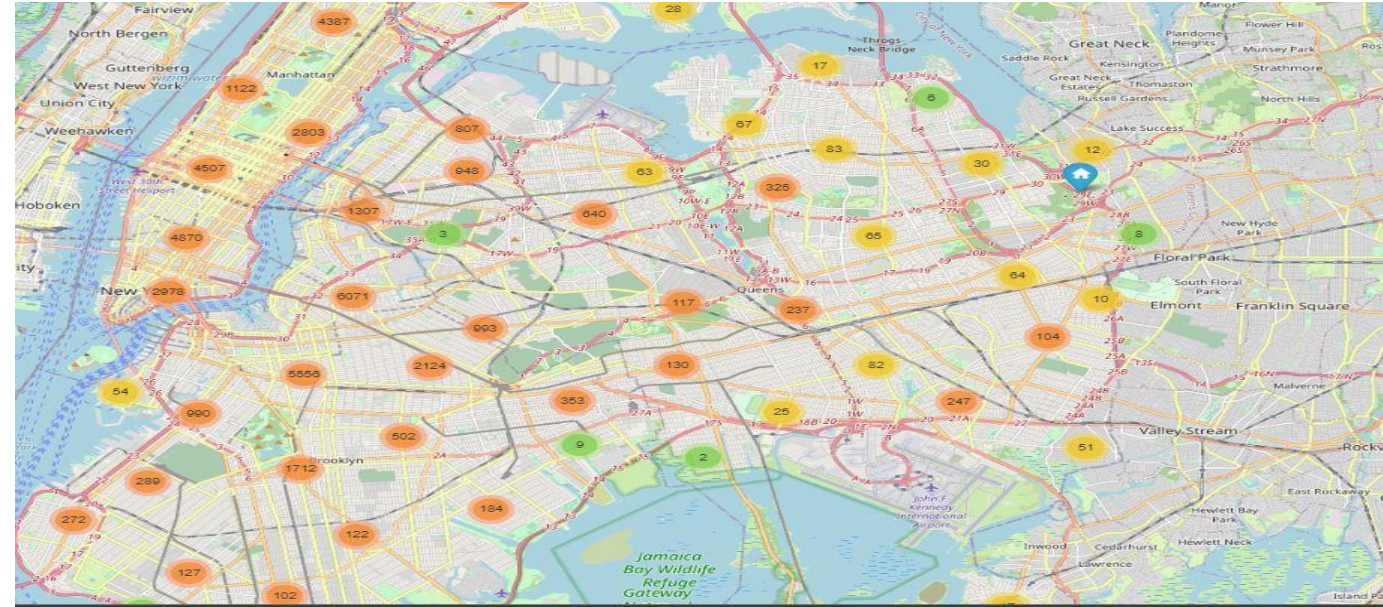
Now we will analyze the Data and will get the answer of following question :

- Which type of hosts to acquire more and where?
- What are the neighborhoods we need to target?
- What is the pricing ranges preferred by customers?
- The various kinds of properties that exist w.r.t. customer preferences.
- Adjustments in the existing properties to make it more customer-oriented.
- What are the most popular localities and properties in New York currently?
- What is the average preferred price of neighborhood group to the location?
- Distribution of room type across neighborhood groups.
- Most Preferred room type.
- Top 10 hosts with most listings & as per most reviews



EDA :

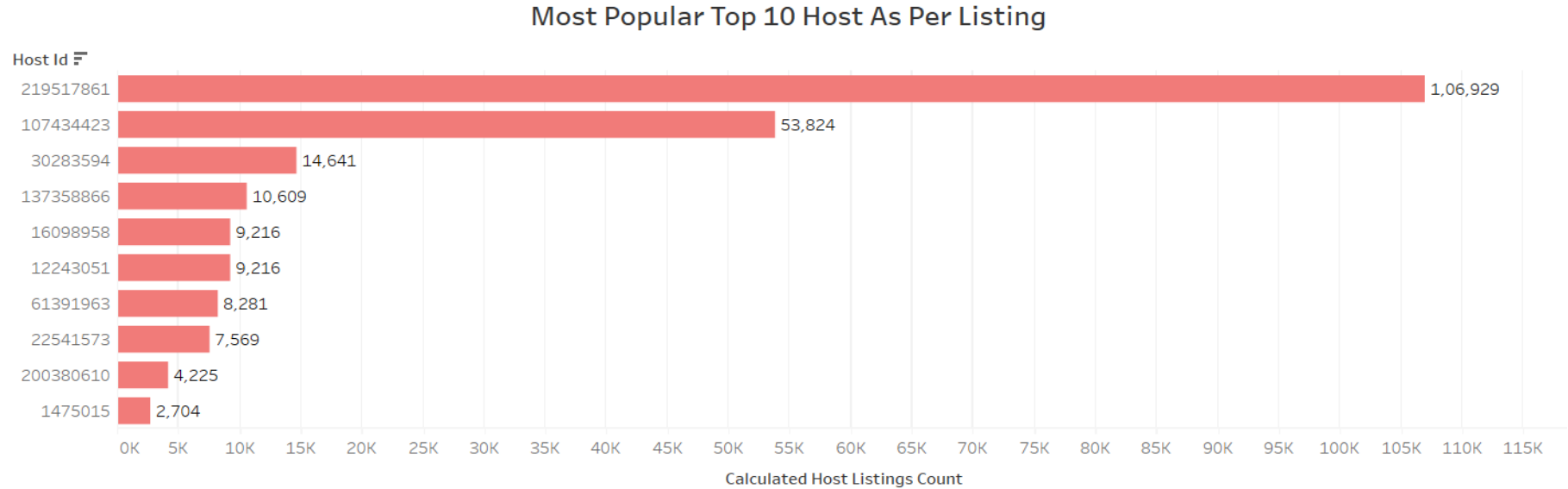
- Map shows the exact location of all the apartments with the help of 'latitude' and 'longitude' co-ordinates.



- Scatterplot on Map shows the cluster of 'neighbourhood_group' with the help of 'latitude' and 'longitude'.
- We can easily see that 'Manhattan' is most dense area which has maximum number of listings.



Most Popular Hosts As Per Listing



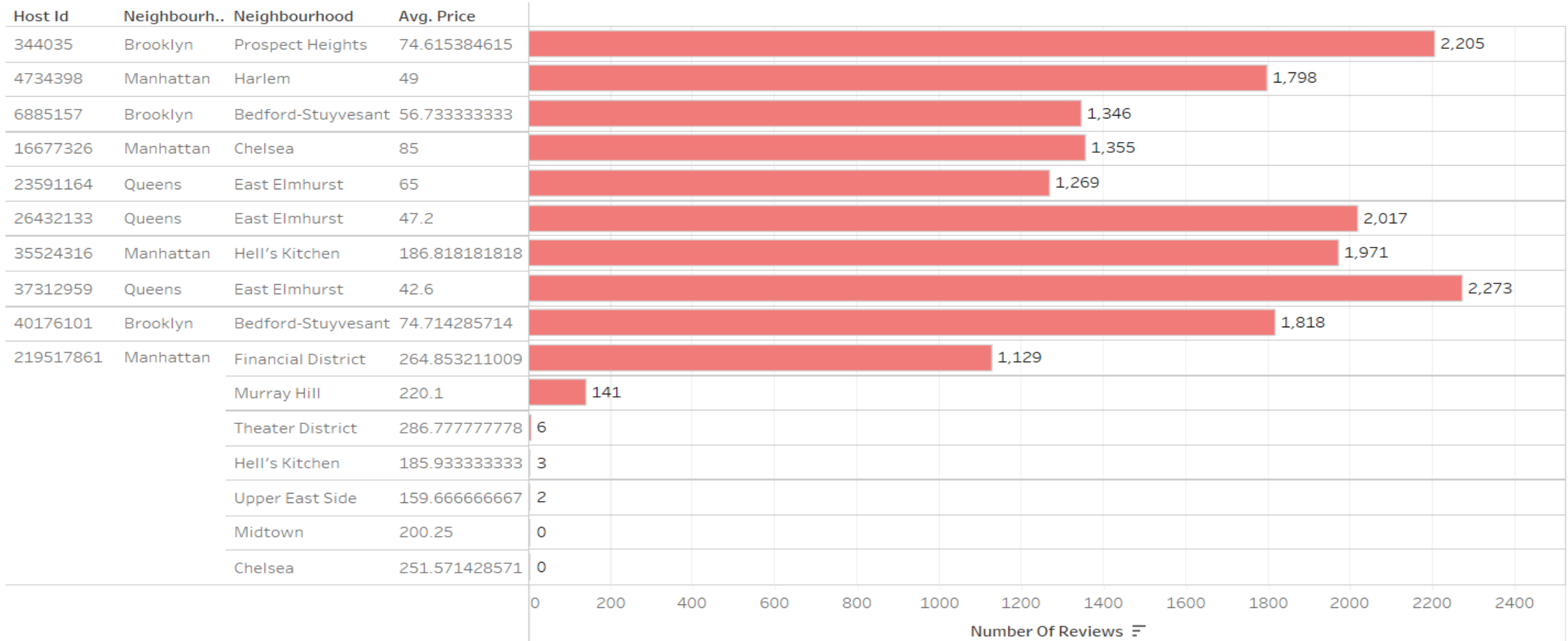
Total Hosts – 37,457

Host with id **219517861** is the most popular as per total sum of listing count & host with id 107434423 is second most popular.



Most Popular Hosts As Per Total Reviews

Top 10 Hosts As Per Total Reviews Across Neighbourhood





Top 10 Hosts As Per Total Reviews Across Neighbourhood

Host Id	Neighbourh..	Neighbourhood	Avg. Price	
344035	Brooklyn	Prospect Heights	74.615384615	2,205
4734398	Manhattan	Harlem	49	1,798
6885157	Brooklyn	Bedford-Stuyvesant	56.733333333	1,346
16677326	Manhattan	Chelsea	85	1,355
23591164	Queens	East Elmhurst	65	1,269
26432133	Queens	East Elmhurst	47.2	2,017
35534316	Manhattan	Upper West Side	100.010101010	1,071

Host with id **344035** is the **most popular** across neighbourhood as per total reviews received as 2,205 & belongs to Prospect Heights area in Brooklyn location.

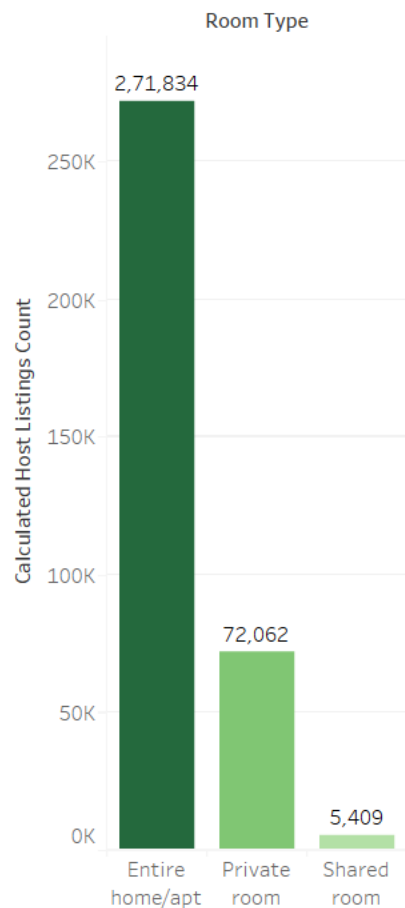
Average Price offered by the most popular host with id 344035 is approx. **75\$ per night** which is quiet reasonable.

Host with id 4734398 is the second most popular as per total reviews as 1,798 & belongs to Harlem area in Manhattan location having average price per night as 49\$.



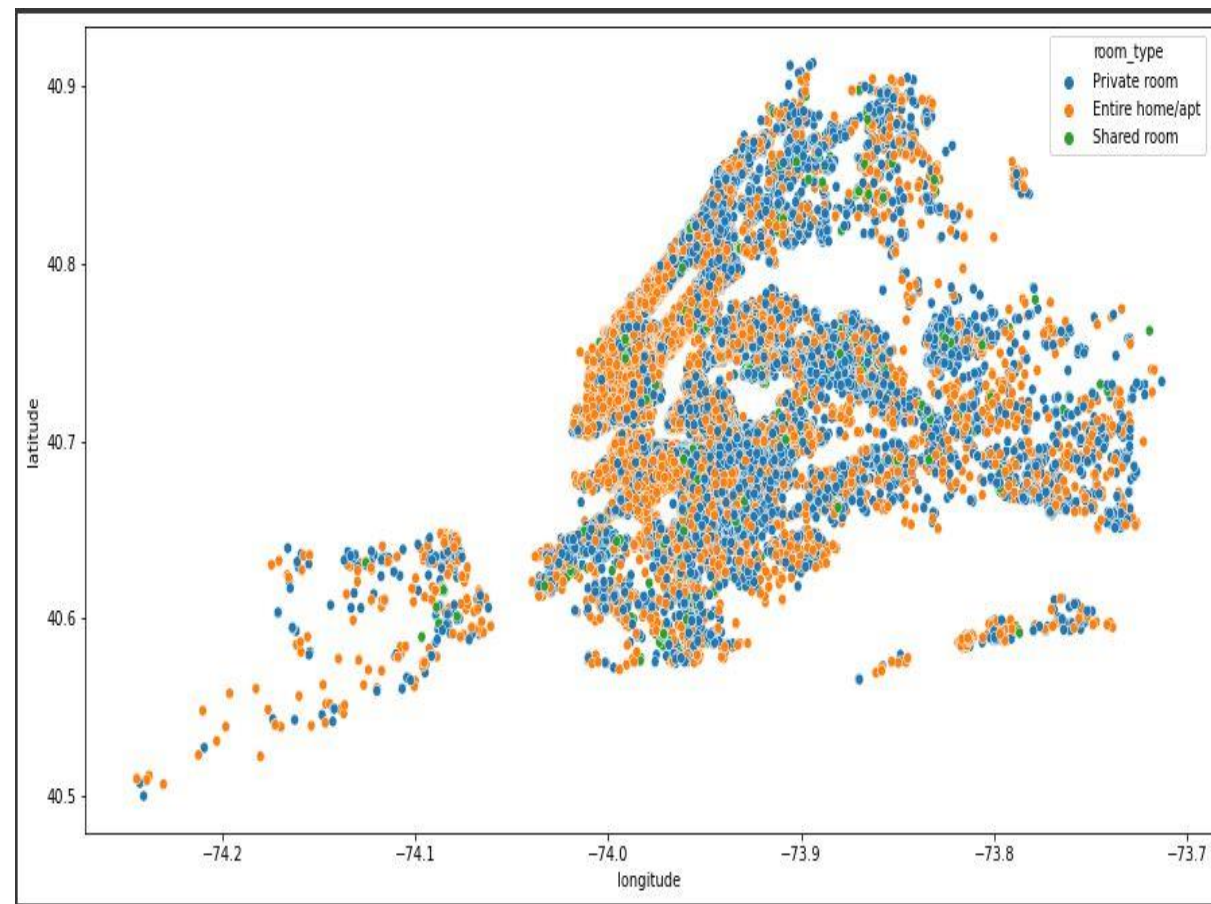
Room type distribution over the location :

Host Listings Per Room Type



Room type of most number of listing is **‘Entire home/apartment’**.

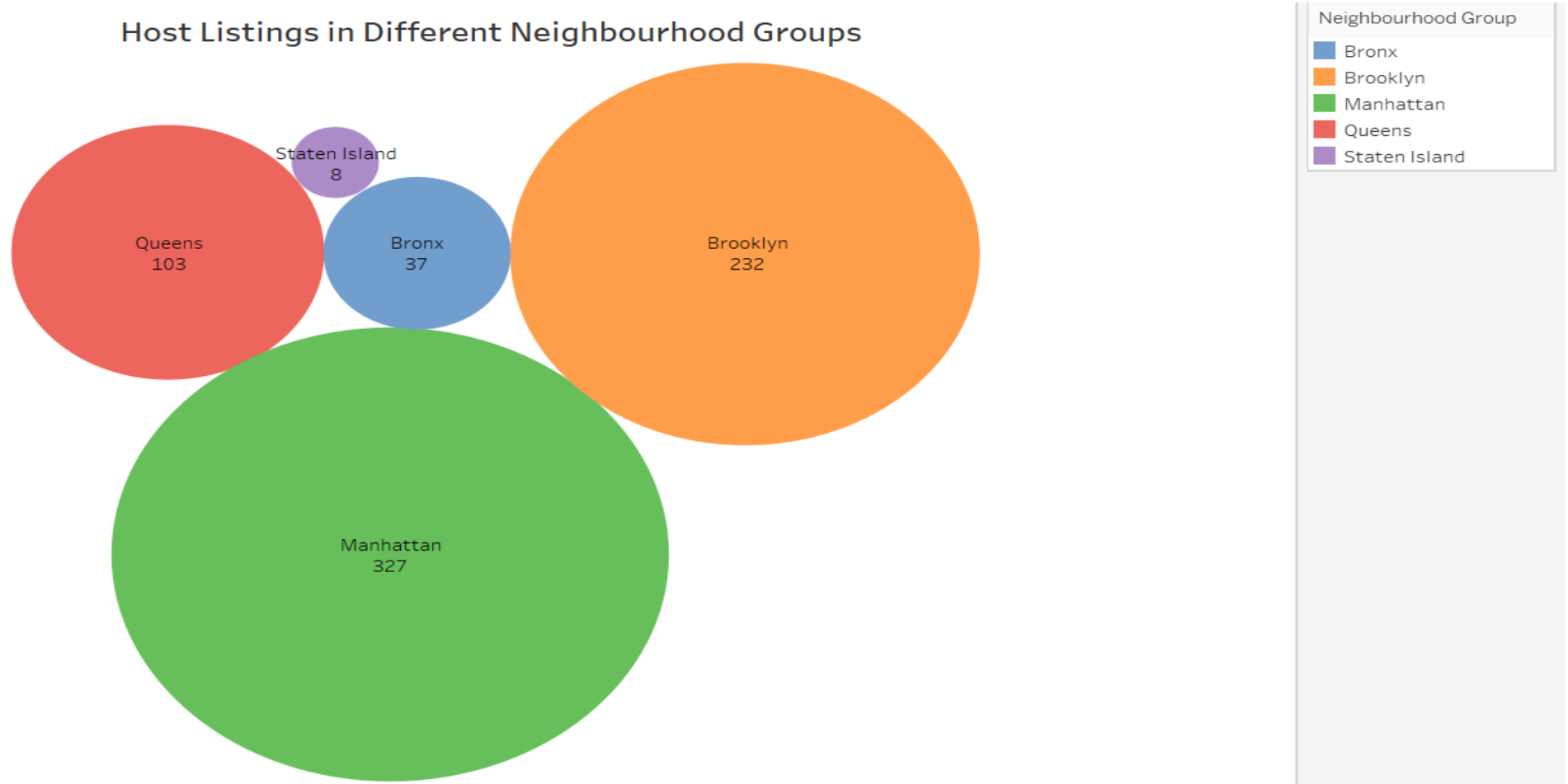
- Scatterplot on Map show the cluster of **‘room_type’**.
- And we can easily predict that for **‘room_type’ private room & entire home**, distribution is almost same in every **‘neighbourhood’**, which means the booking of room type is almost same in every **‘neighbourhood’**.



Scatterplot on Map show the cluster of **‘room_type’**



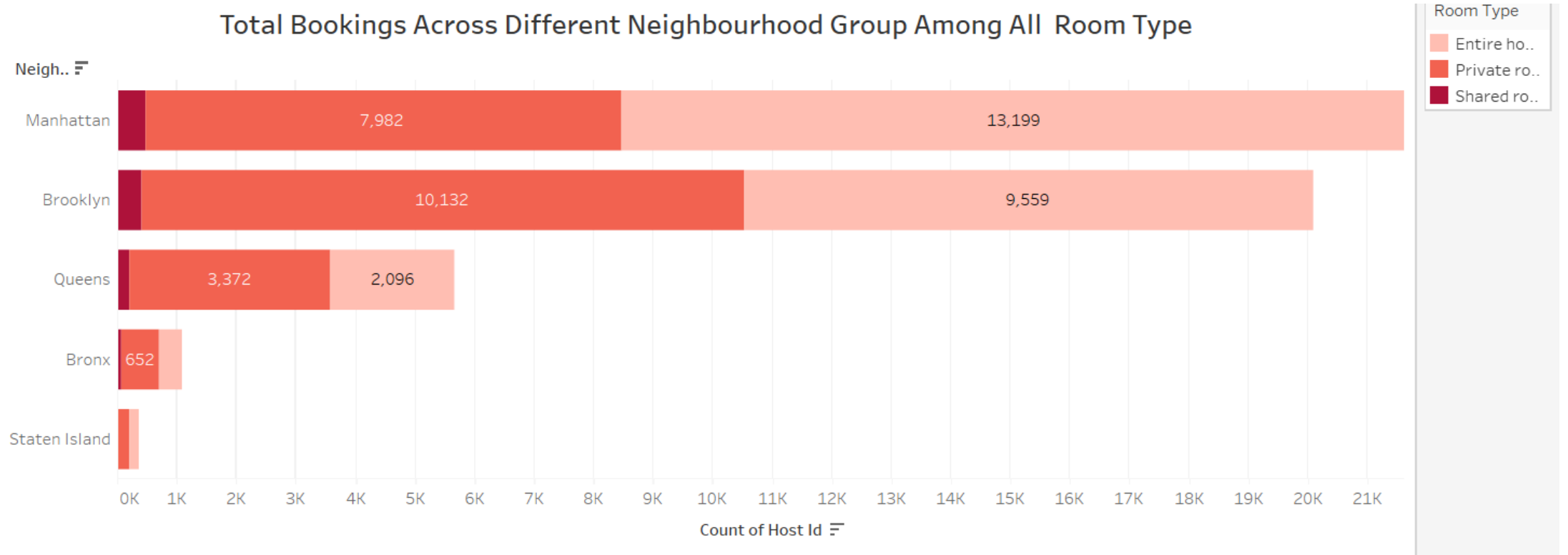
Maximum Host Listings In Different Neighbourhood Groups



Maximum Host Listings is present for **Manhattan** location



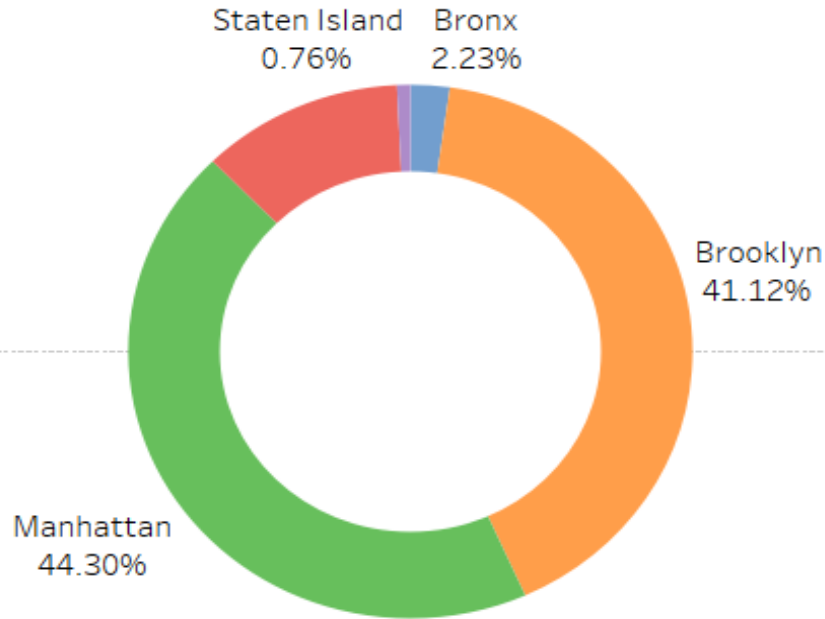
Total Bookings



Highest bookings are done for entire home/apt room type for the location Manhattan followed by Brooklyn

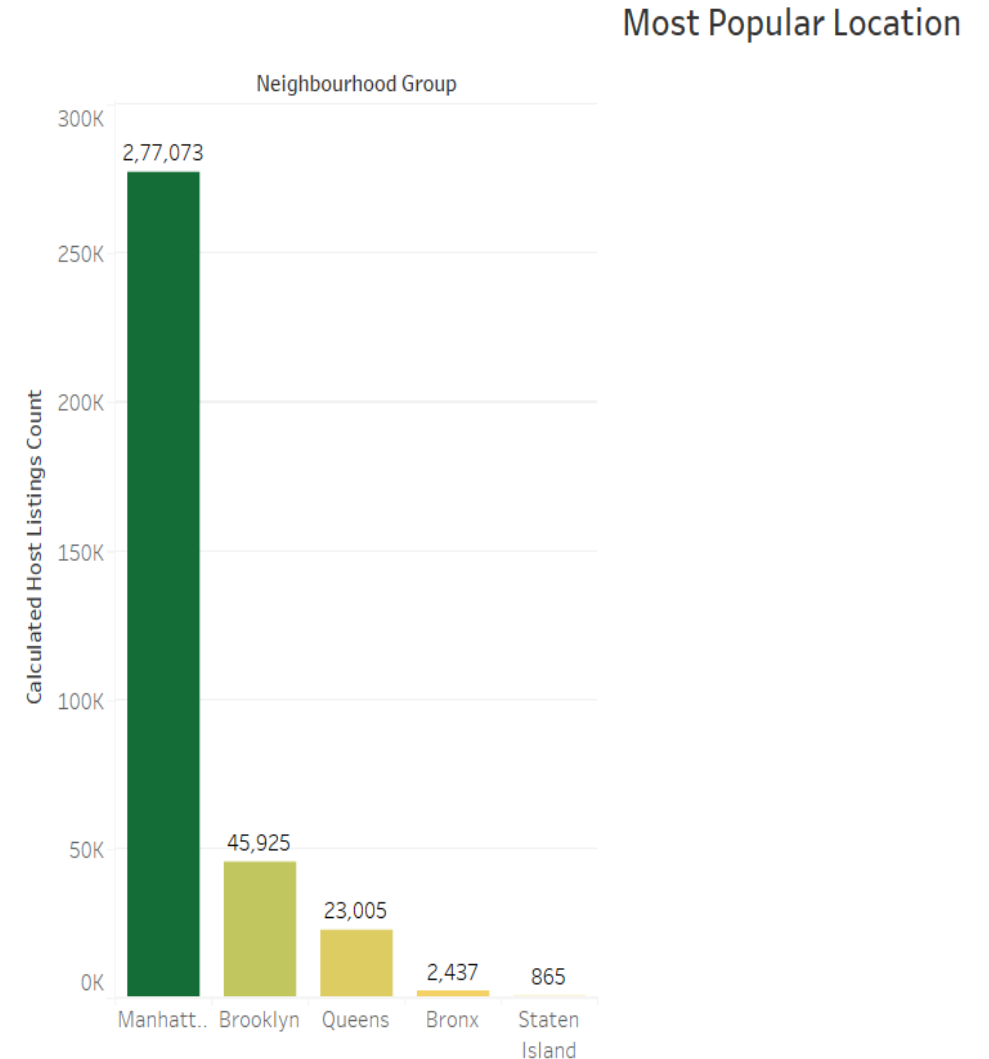


Total Neighbourhood By Neighbourhood Groups



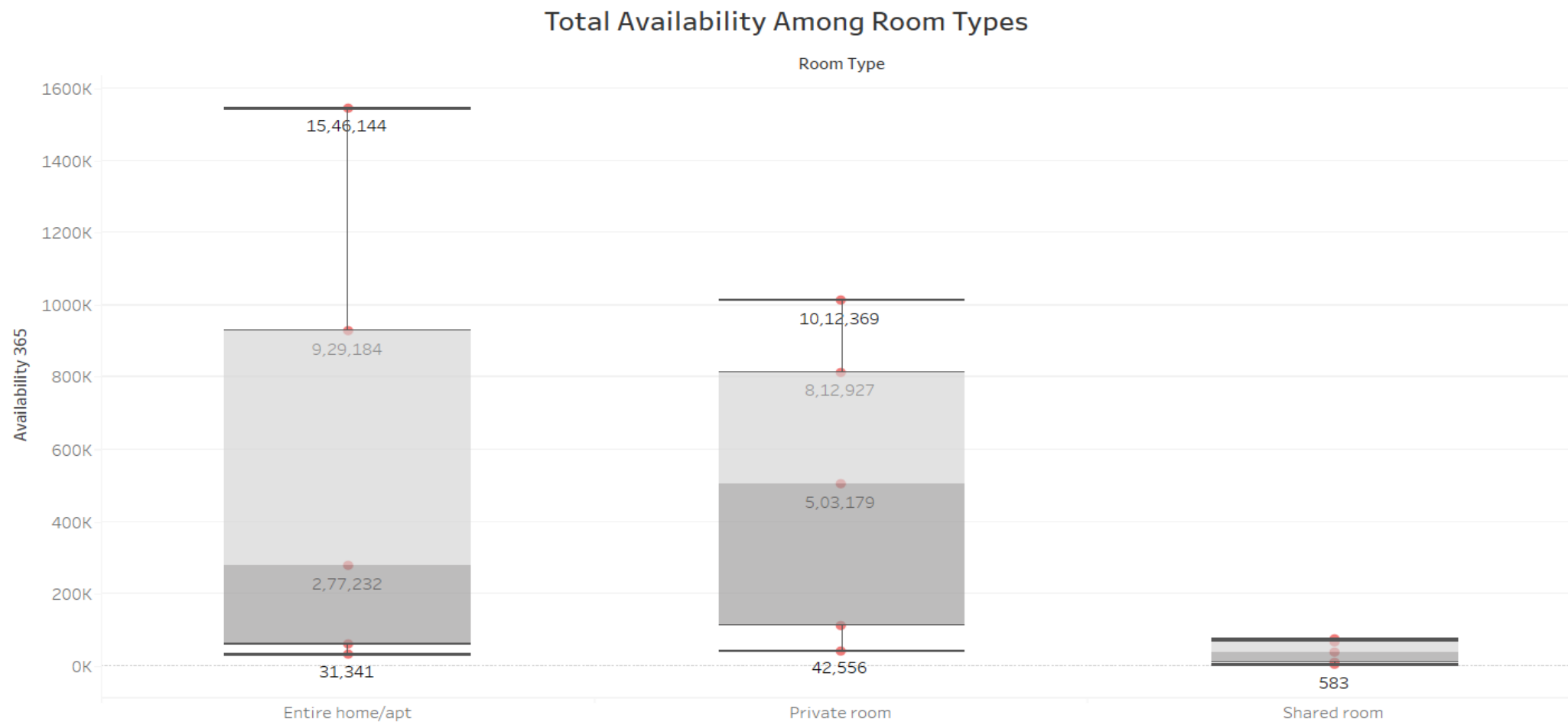
Total Neighbourhood
in Neighbourhood
Groups = 221

Manhattan & Brooklyn
constitute maximum
percentage of
neighbourhoods ~85%





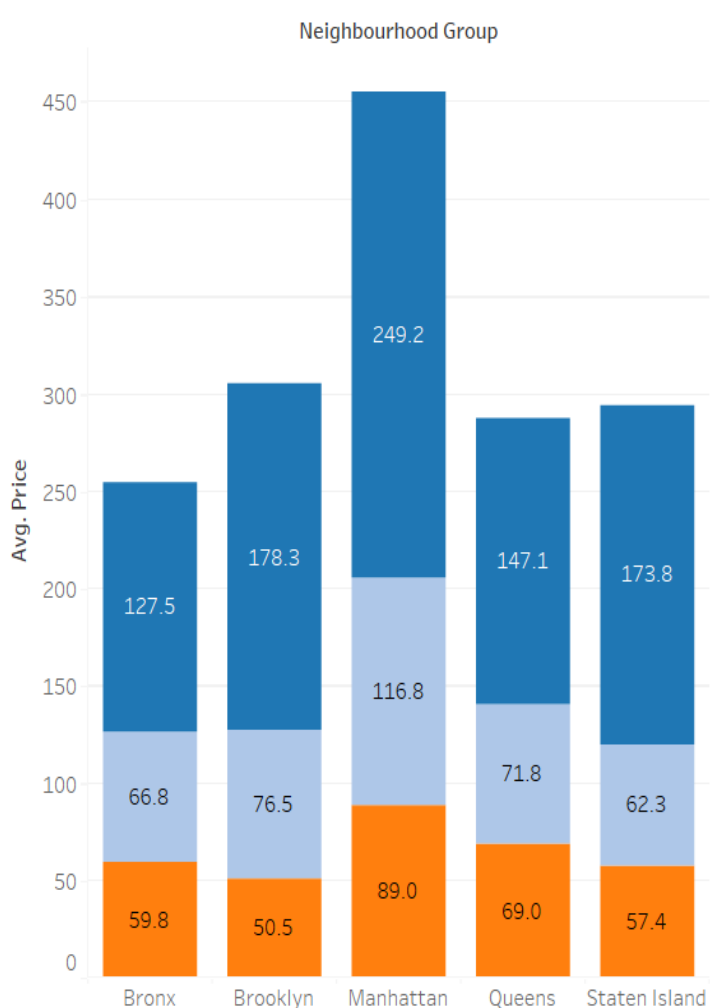
Availability Among Different Room Types





Avg Price Across Neighbourhood Group Among Different Room Types

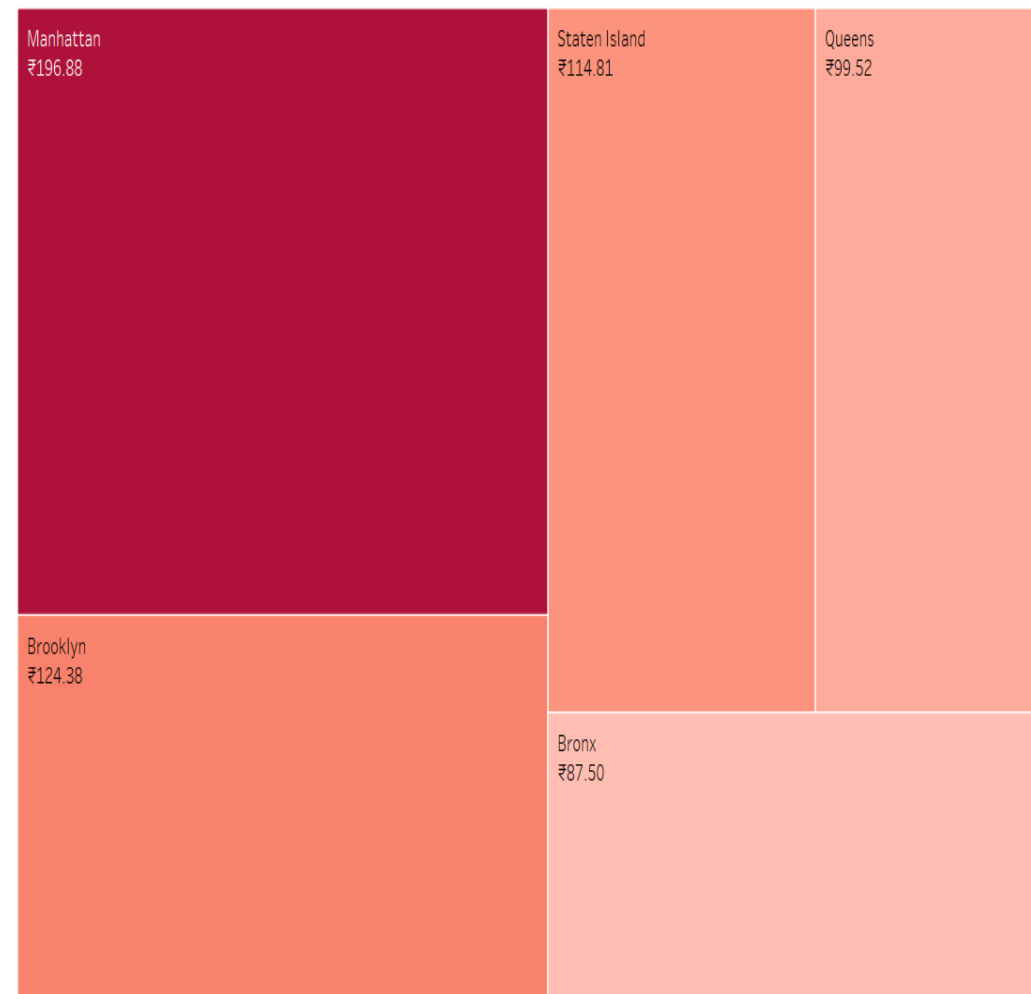
Avg Price Across Neighbourhood Group Among Different Room Types



Average price is highest for Entire home/apartment & maximum for the location **Manhattan**.

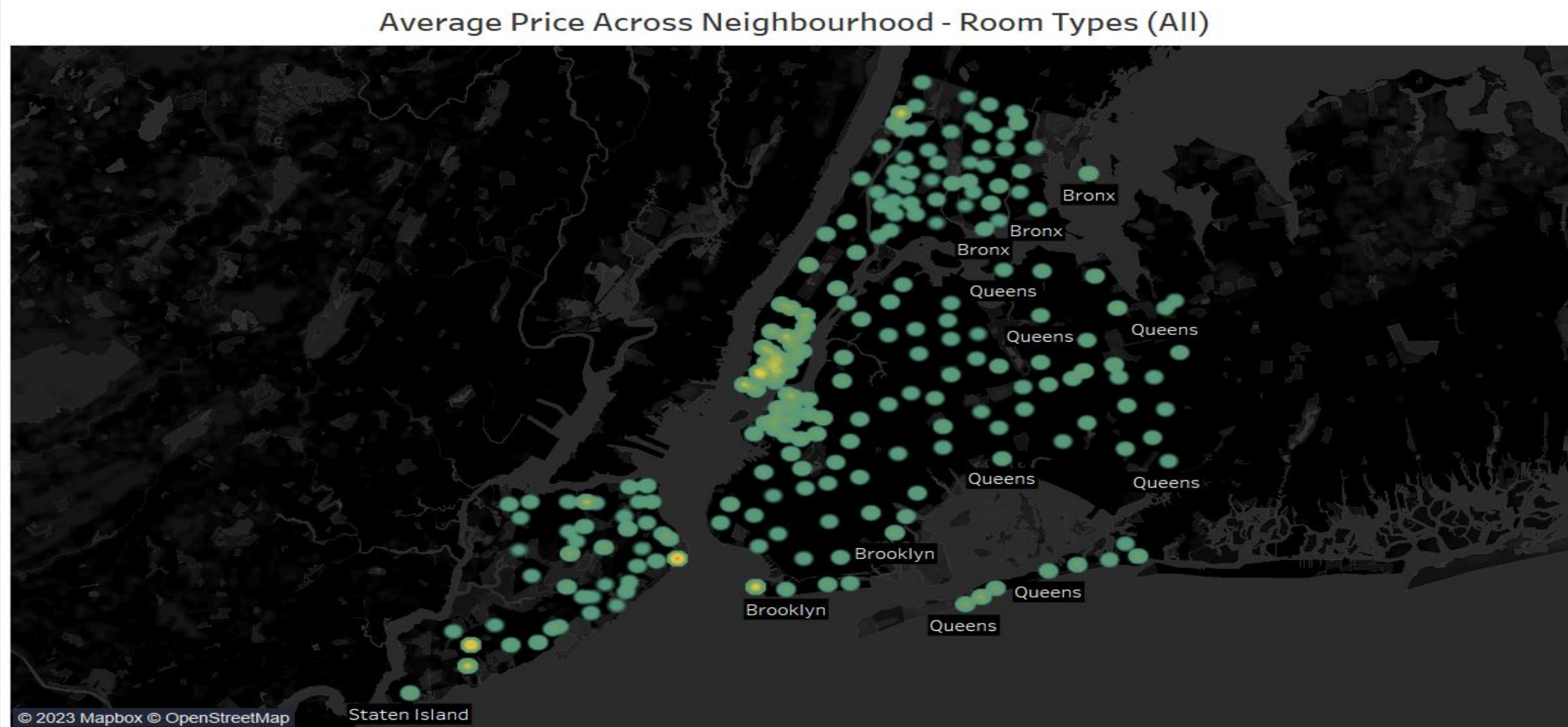
Least average price is for Shared rooms.

Average Price Across Neighbourhood Groups





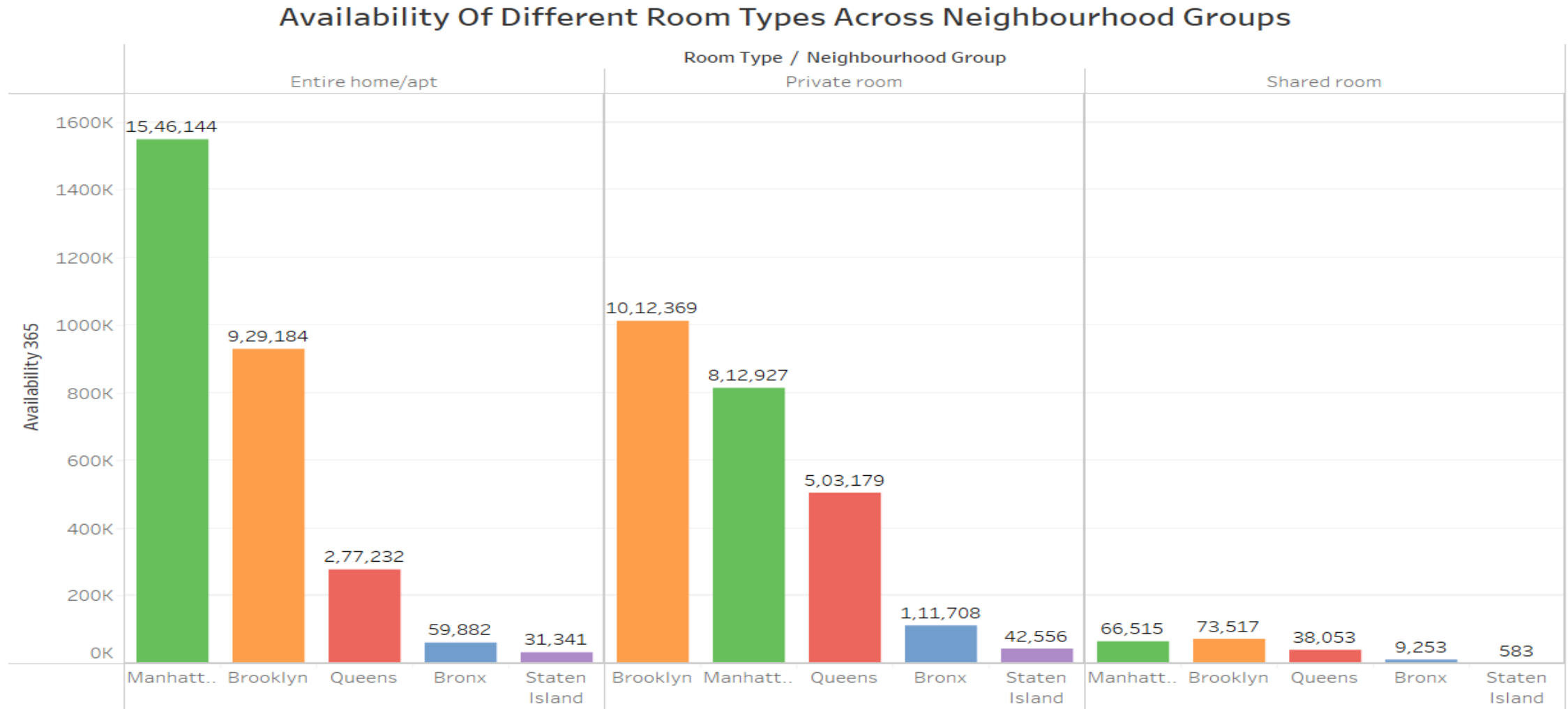
Distribution Of Avg Price Across Neighbourhood



Average price is maximum for the rooms in areas that are near sea soar / coastal area



Availability Of Room Types Across Neighbourhood Group

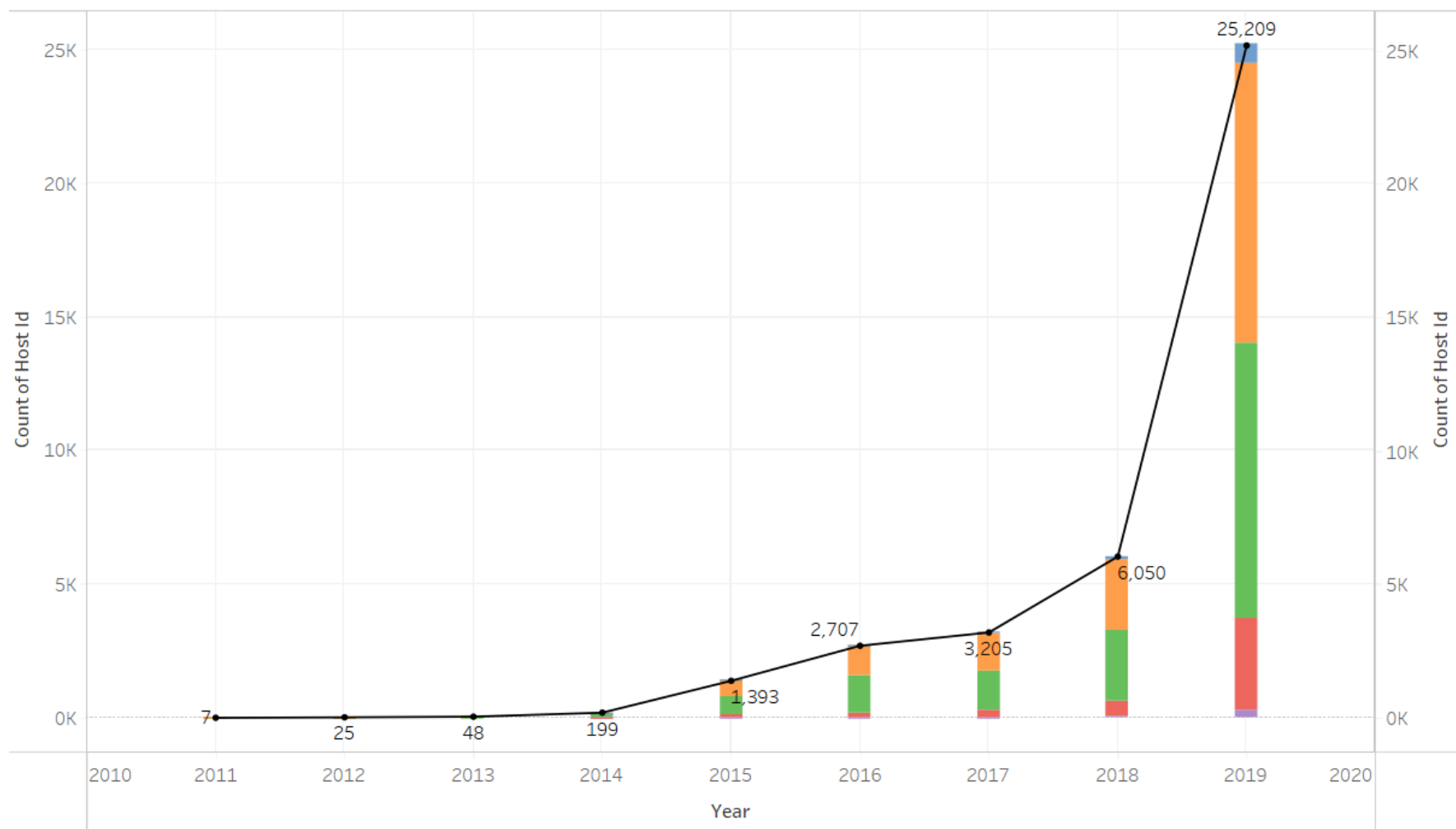


Maximum room type available is entire home/apt & for the location Manhattan followed by Brooklyn.



Total Bookings Over Years

Total Bookings By Year Across Neighbourhood Group For All Room Types



Neighbourhood Group

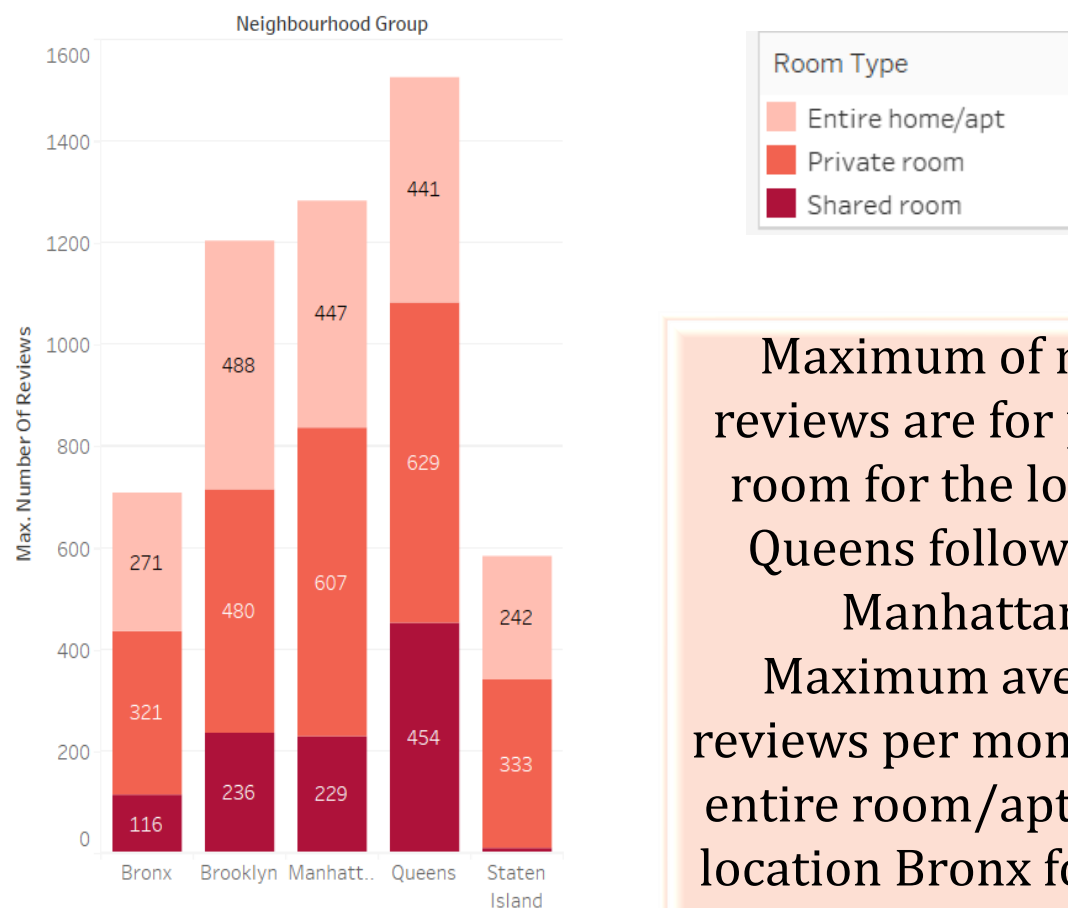
- Bronx
- Brooklyn
- Manhattan
- Queens
- Staten Island

Total bookings have significantly increased over years for all room types & most preferred locations are Manhattan & Brooklyn.



Reviews Per Month By Room Type & Neighbourhood Group

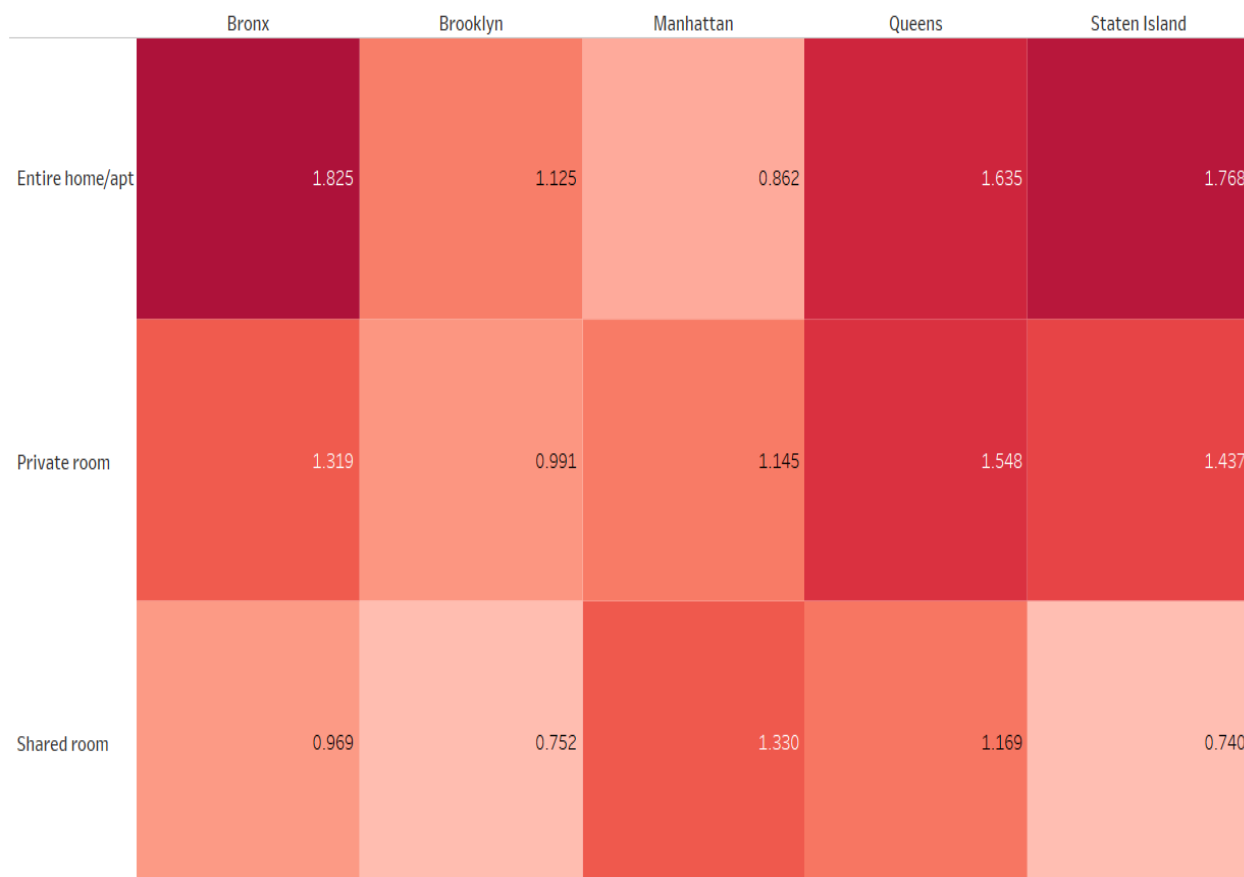
Maximun No. Of Reviews Across Neighbourhood Group - All Room Type



Maximum of no. of reviews are for private room for the location Queens followed by Manhattan.

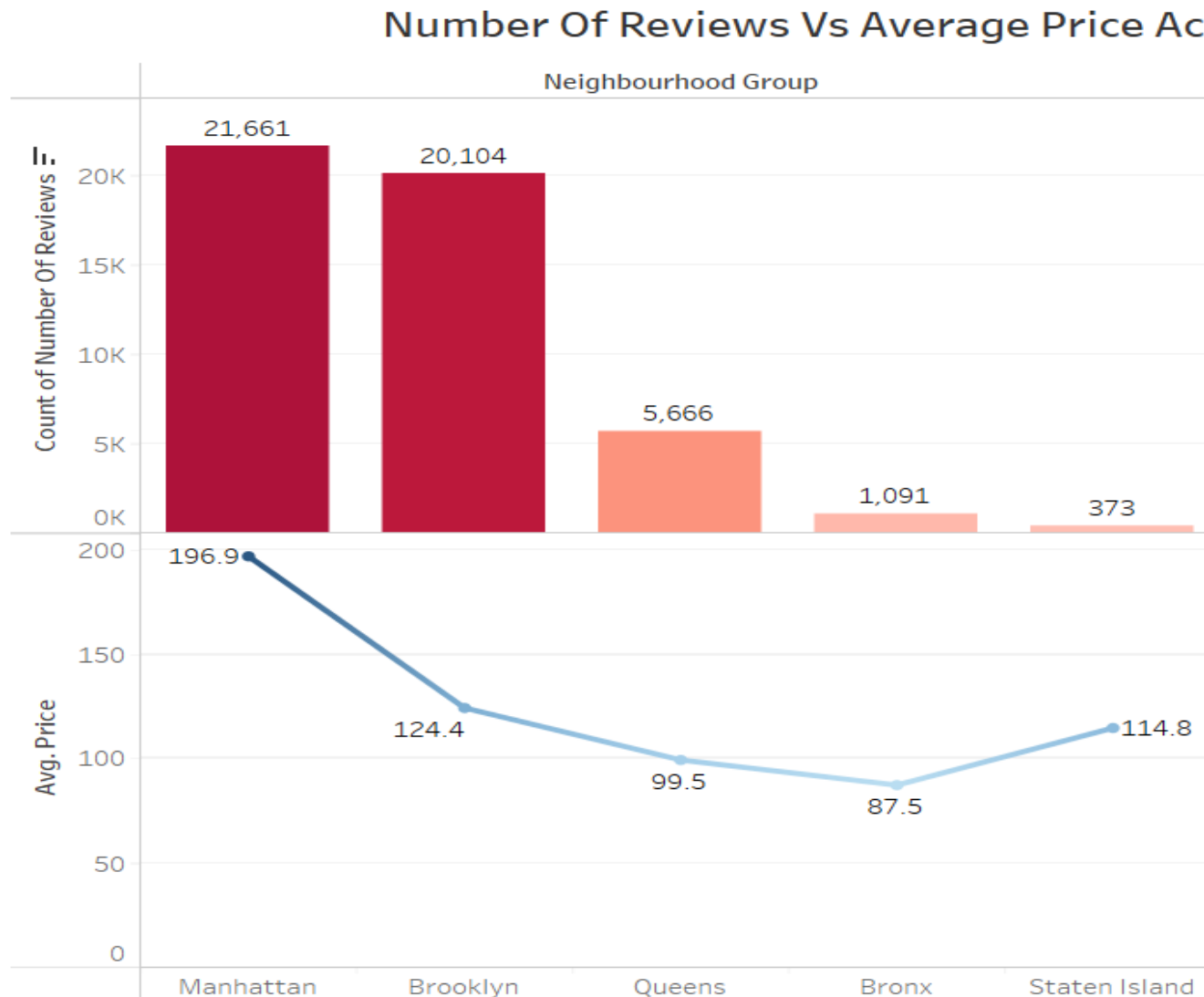
Maximum average reviews per month is for entire room/apt for the location Bronx followed by States Island

Average Reviews Per Month By Room Type & Neighbourhood Group





Comparing Avg Price & No. Of Reviews Across Neighbourhood Group

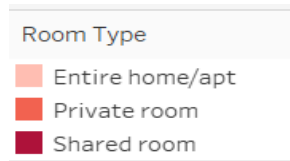
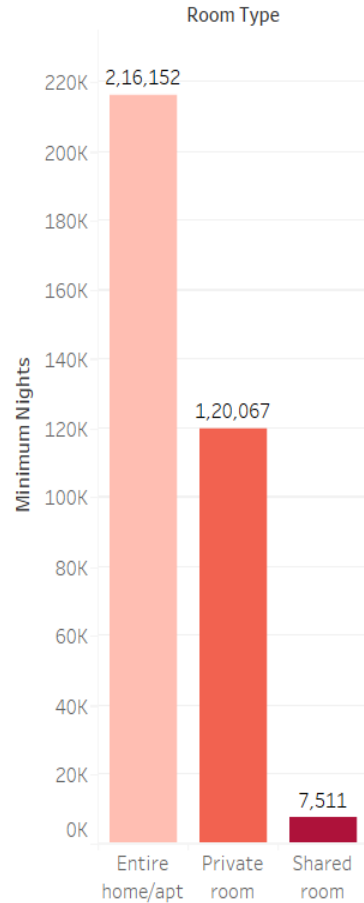


As we can observe that average price is maximum for Manhattan location followed by Brooklyn. Also the no. of reviews is also maximum for Manhattan neighbourhood group.



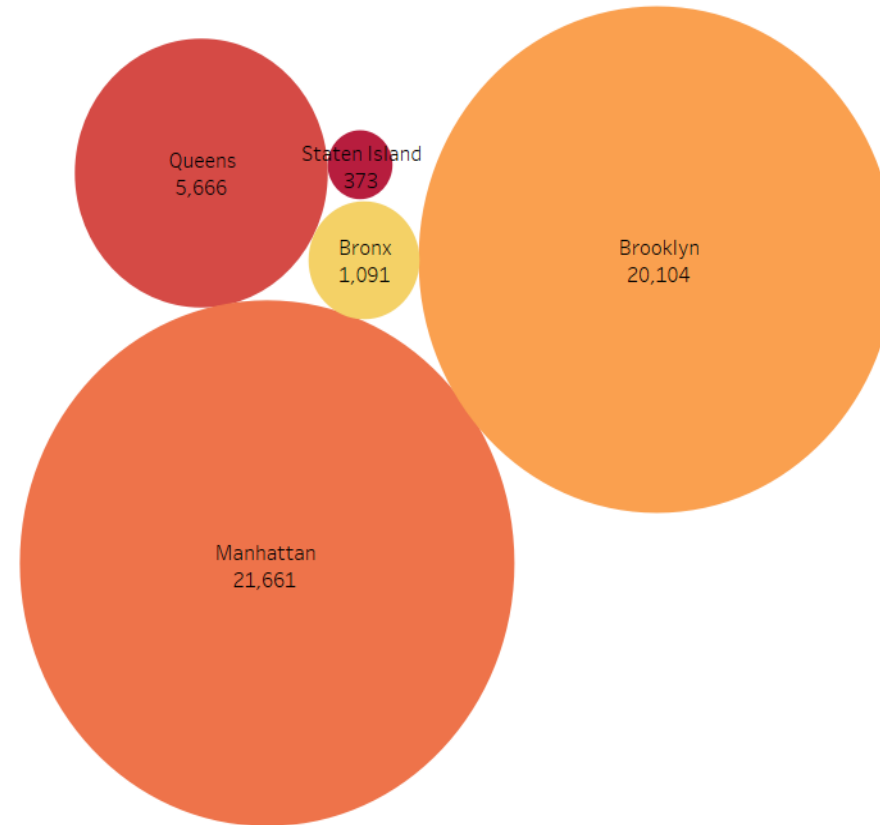
Minimum Night Stay For All Room Types Across Neighbourhood Groups

Minimum Night Stay Vs Room Type



Most min. night stayed room type is entire room & most preferred location for the same is Manhattan followed by Brooklyn.

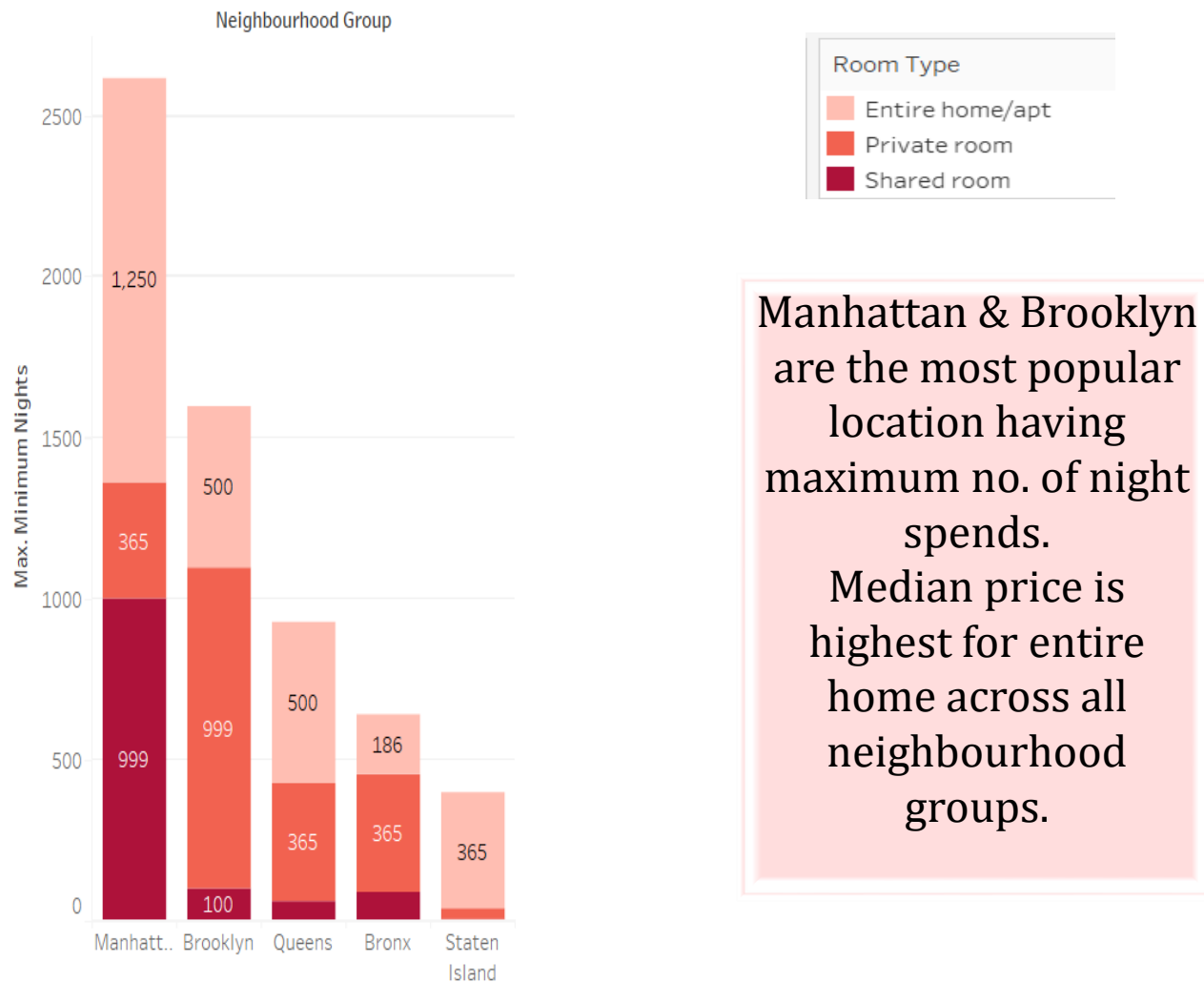
Minimum Night Stay Across Different Neighbourhood Group



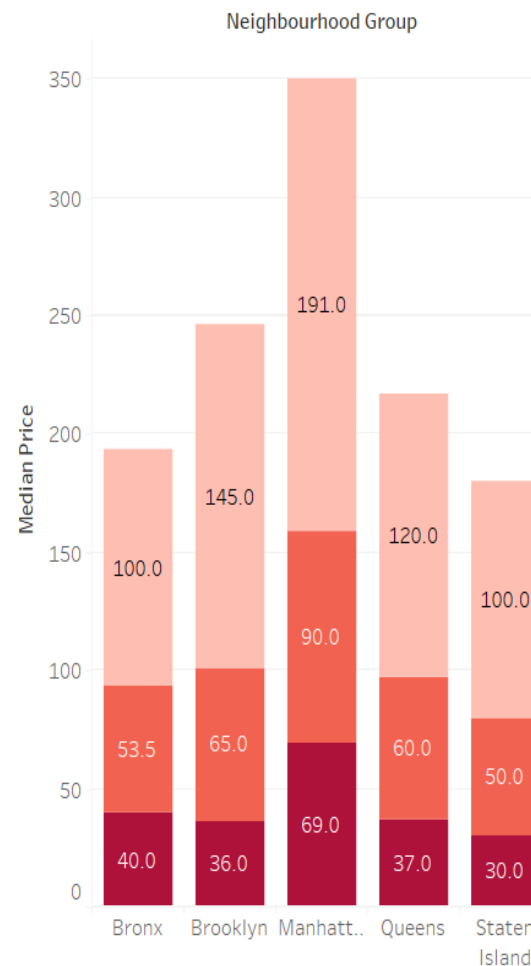


Max. No. Of Nights Spends VS Price Bracket Across Neighbourhood Groups

Maximum Number Of Night Spends In Neighbourhood Group



Price Bracket In Neighbourhood Groups





Key Recommendations & Action Items :

- **Manhattan** is the most popular location in New York for hosts to do Airbnb business followed by Brooklyn : As people preferred Manhattan to stay when in need for maximum no. of night stay in Entire apartments room type. Although the average prices are on higher side for this location still it's the most popular among travellers. Maximum Host listings is also for Manhattan only.
- Maximum room type available is **entire home/apt** & for the location Manhattan followed by Brooklyn.
- **Manhattan & Brooklyn** constitute maximum percentage of neighbourhoods ~85% .
- **Average price is highest for Entire home/apartment** & maximum for the location **Manhattan**.
- **Shared rooms** have fewer listings but are more affordable and have more availability, so they can be maximised.
- **Average price is maximum for the rooms in areas that are near sea / coastal area.**
- Host with id **219517861** is the most popular as per total sum of listing count & host with id 107434423 is second most popular.
- Host with id **344035** is the most popular across neighbourhood as per total reviews received as 2,205 & belongs to Prospect Heights area in Brooklyn location.



- **Average Price** offered by the most popular host with id 344035 is approx. **75\$ per night** which is quite reasonable & competitive.
- The no. of reviews is higher at lower priced properties as people are less likely to book expensive rooms.
- Most of the listed properties **are private rooms & complete homes/apt**, which also account for majority of the total price share.
- Expensive prime locations like Manhattan & Brooklyn can be targeted for non premium properties & location like Queens & Bronx for premium properties.
- Acquire Private rooms & entire home/apt since they are more popular room type having more no. of reviews per listings.



Challenges Faced :

- Reading the data set & understanding the columns.
- For answering some of the questions we had to understand the business model of Airbnb that how they work.
- Handling NAN values & imputing missing data.
- The most common challenge in the data analysis phase involves choosing how to present the data, that is, how to use charts to represent them.
- Designing multiple visualizations to summarise the information in the dataset & successfully communicating the results & trends to the reader.
- The interpretation of results and identification of patterns.



APPENDIX :

Data Assumptions :

- Assumed that pre-pandemic data was generating desired revenue.
- Assumed that the company does not wish to expand into new markets in New York.
- To learn about customer preferences, used the no. of reviews per listings as a popularity metric.
- Assumed no. of reviews provided to be use as a base measure to find customer preferences.
- Null values are assumed to have no effect on the analysis.



Data Source & Data Dictionary :

- New York Airbnb Dataset provided – **AB_NYC_2019.csv**
- Data Dictionary attached

Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

Dataset Description



Data Methodology & Visualisations :

- Data understanding & pre processing is done in python using pandas library.
- Basic data cleaning & handling such as missing value imputation is also done with the help of pandas.
- After cleaning the data set its imported as excel file to do further analysis.
- Data analysis & visualisation is done using Tableau.

Methodology document attached for further reference: [Methodology.docx](#)



THANK YOU