EDA CAPSTONE PROJECT

PROJECT 1: AIRBNB BOOKING ANALYSIS

TEAM MEMBERS:

- **RAJ SONAR** (Cohort Vindhya)
- **RAHUL JHA** (Cohort Vindhya)
- **PRAVEEN SIVAKUMAR** (Cohort Everest)

PROBLEM STATEMENT:

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. About 150 million people use Airbnb to book vacation stays or experiences and over 800 million guests have stayed at Airbnbs.

These millions of listings generate a large amount of data, which can be analysed and used to improve the business's efficiency.

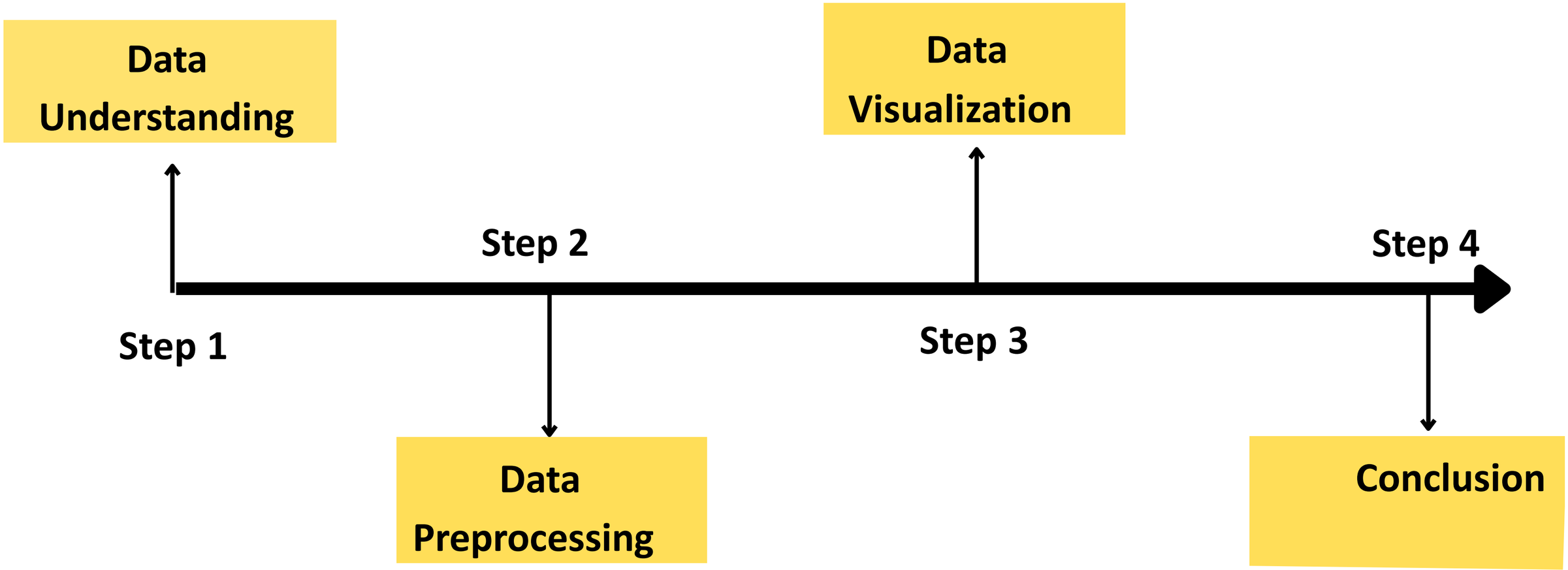
Aim of the Project:

The dataset has around 49,000 observations in it with 16 columns and it is a mix of categorical and numerical values.

- The project's goal is to evaluate the data and come up with basic questions like:

  - How geography affects hotel bookings?
  - How price is correlated with type of room booked?
  - What is the most requested room types in each neighbourhood?.....

Methodology:



DATA UNDERSTANDING:

COLUMNS:

- **ID**: It gives a unique number for each observation.
- **Name**: Basic description of the provided Airbnb.
- **Host ID**: This gives us the id of the host who owns the Airbnb.
- **Host Name**: This gives us the name of the host who owns the Airbnb.
- **Neighbourhood Group**: The 5 boroughs (a town or district which is an administrative unit) of the New York City.
- **Neighbourhood**: Towns/Cities present in the 5 boroughs.
- **Latitude**: Latitude of the Airbnb.
- **Longitude**: Longitude of the Airbnb.

Data Understanding:

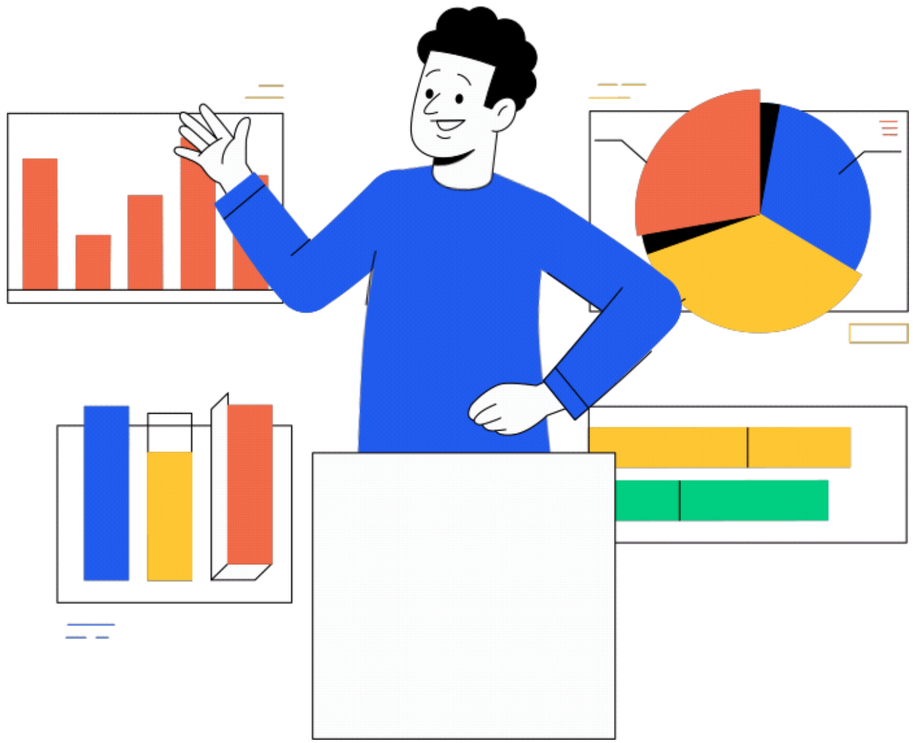
- **Room Type**: Different room types available for the Airbnb booking.
  a. Entire Home/Apartment
  b. Private Room
  c. Shared Room
- **Price**: Price of the Airbnb for one night.
- **Minimum Nights**: Number of minimun nights spent by a person in the Airbnb.
- **Number of Reviews**: Number of reviews received by the Airbnb.
- **Last Review**: Date of the last review given by the user.
- **Reviews per Month**: Mean number of reviews received by the Airbnb per month.
- **Calculated host listings count**: Count of the list of hosts.
- **Availability 365**: Availability of the Airbnb out of 365 days.

Data Preprocessing:

- No duplicate data was found.
- Dropped the columns which we did not require such as name, host name, latitude, longitude, last review and calculated host listings count.
- For missing values in review_per_month we replaced it with 0.
- Rows with price equal to 0 was replaced with its median price.
- There were some high priced Airbnbs included in the dataset which would be the lavish ones, so for price distribution we removed the outliers.

Data Visaulization:

To study and visaulize the dataset we divided the visualization on the basis of:



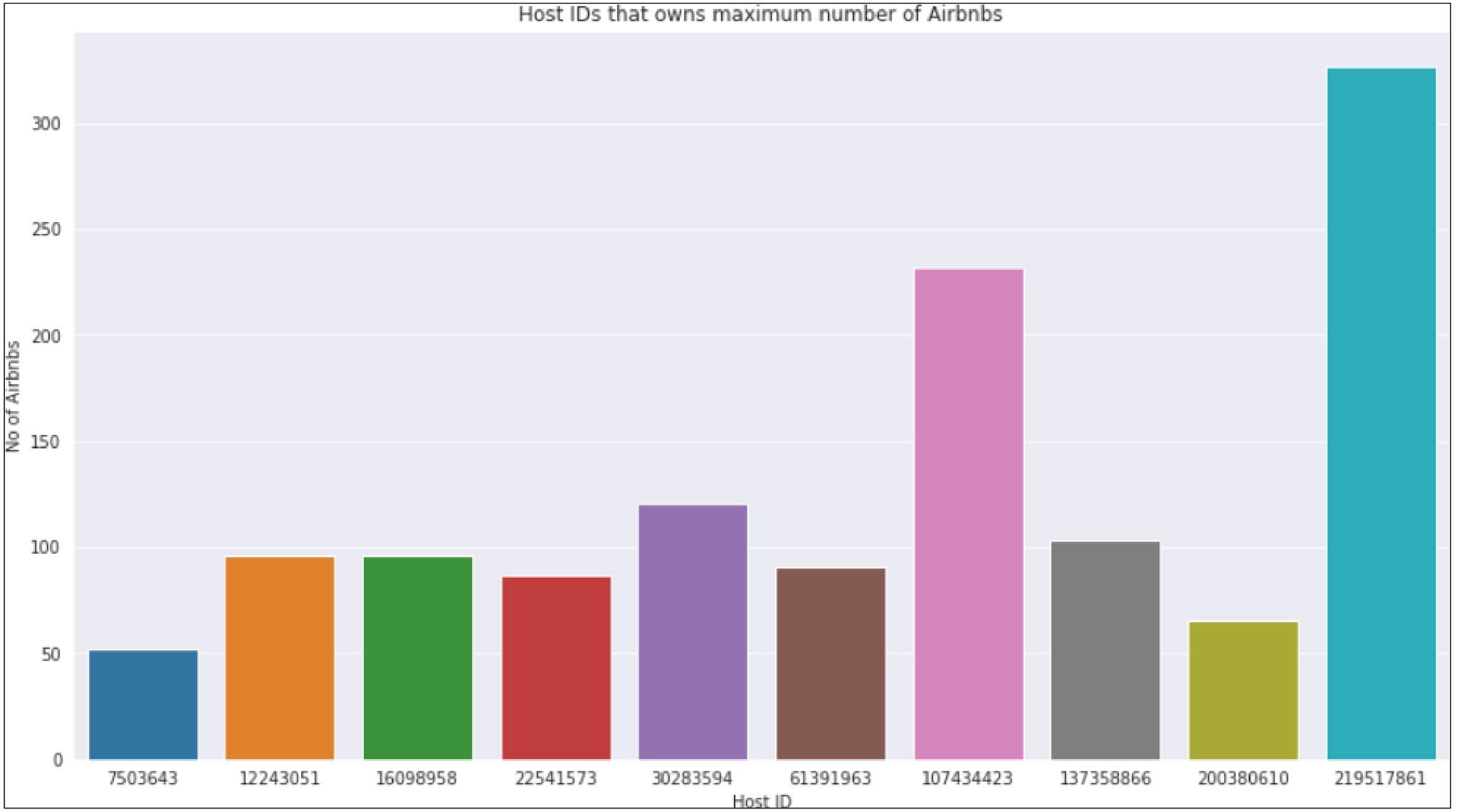
1.Univariate Analysis

It is the simplest form of data analysis where the data being analyzed contains only one variable.

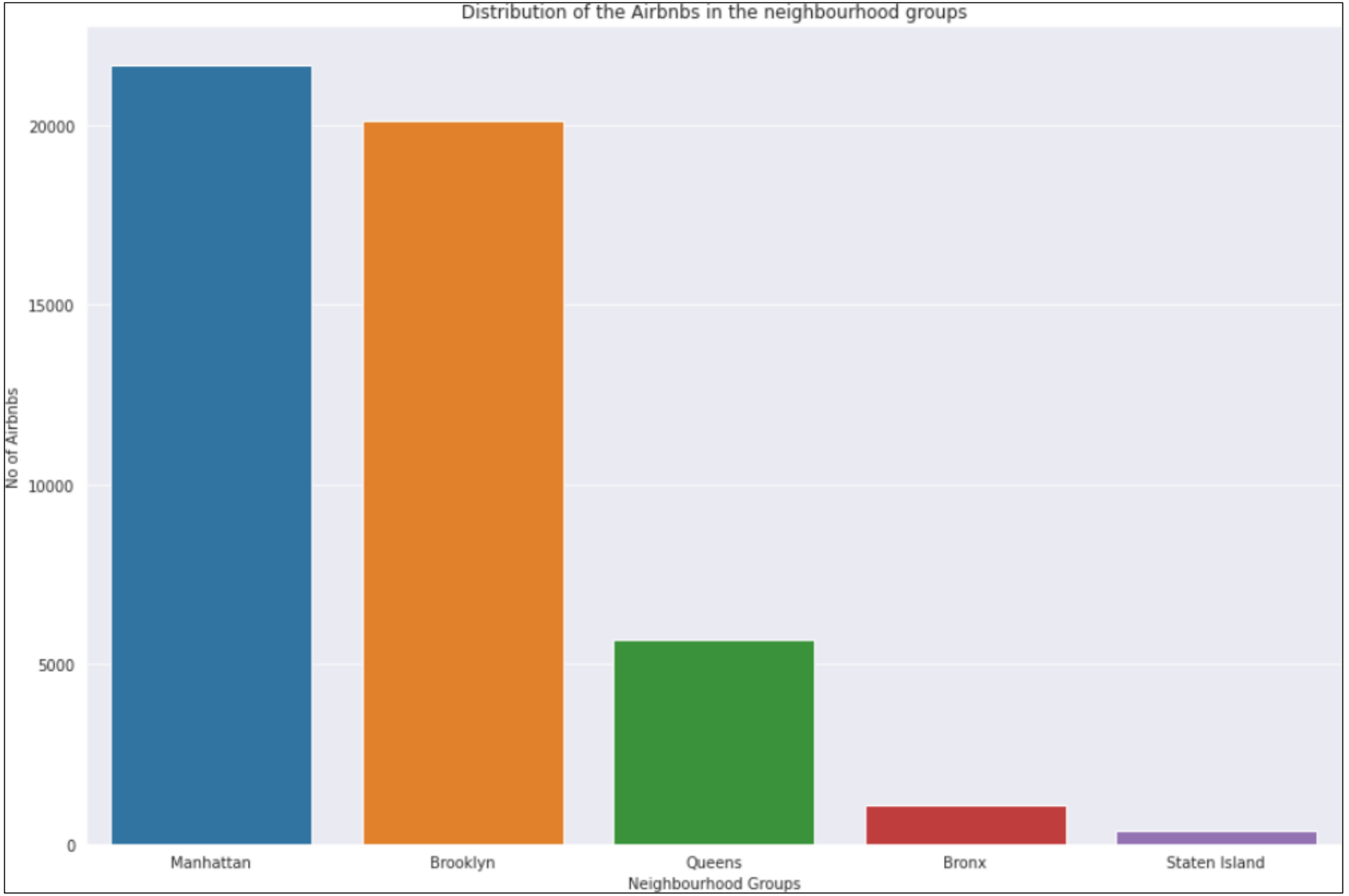
2.Multivariate Analysis

Multivariate analysis is a set of techniques used for analysis of data sets that contain more than one variable, and the techniques are especially valuable when working with correlated variables.

TOP 10 HOST IDs WITH MAXIMUM NUMBER OF AIRBNBS



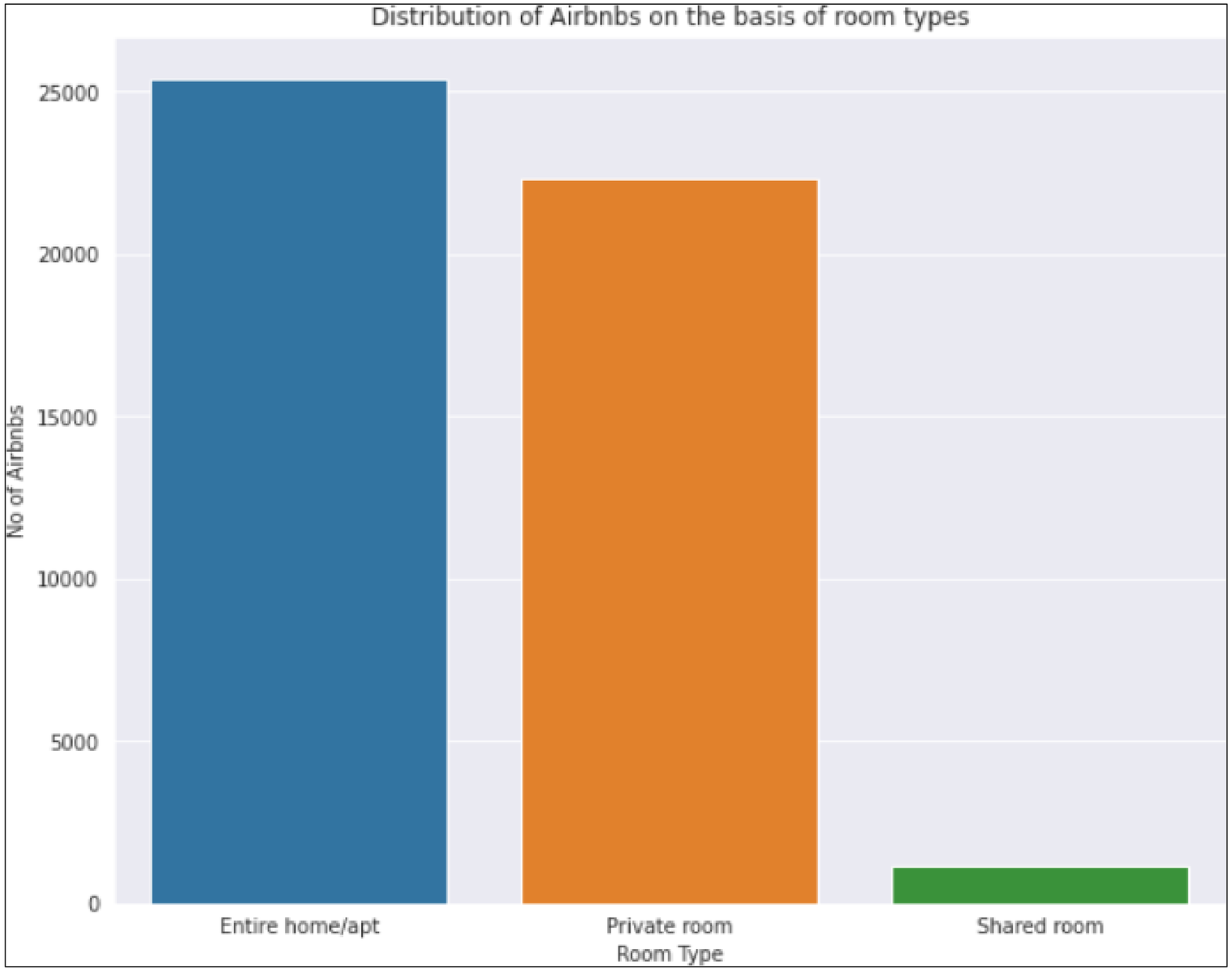
RELATION BETWEEN TOTAL NUMBER OF AIRBNBs AND NEIGHBOURHOOD GROUP



Conclusion:

- Staten Island and Bronx has the lowest numbers of Airbnbs.
- Brooklyn and Manhattan has a larger percentage of Airbnbs.

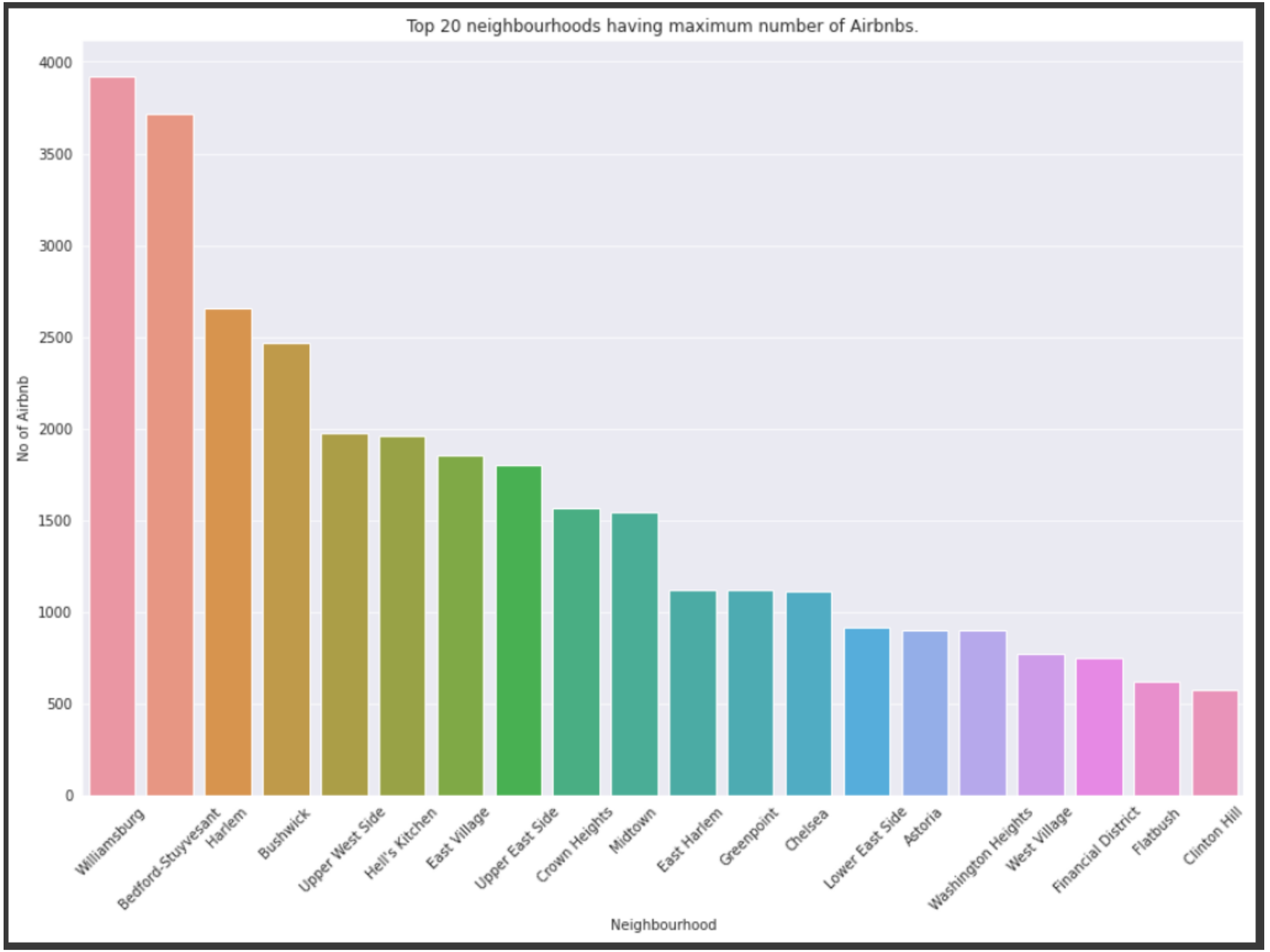
RELATION BETWEEN NUMBER OF AIRBNBs AND THEIR ROOM TYPE



Conclusion:

- **Shared room are lot less in comparison to other room types.**
- **It can be seen that count of entire home and private room is too high, so people would be booking them a lot.**

TOP 20 NEIGHBOURHOODS WITH MAXIMUM NUMBER OF AIRBNBs

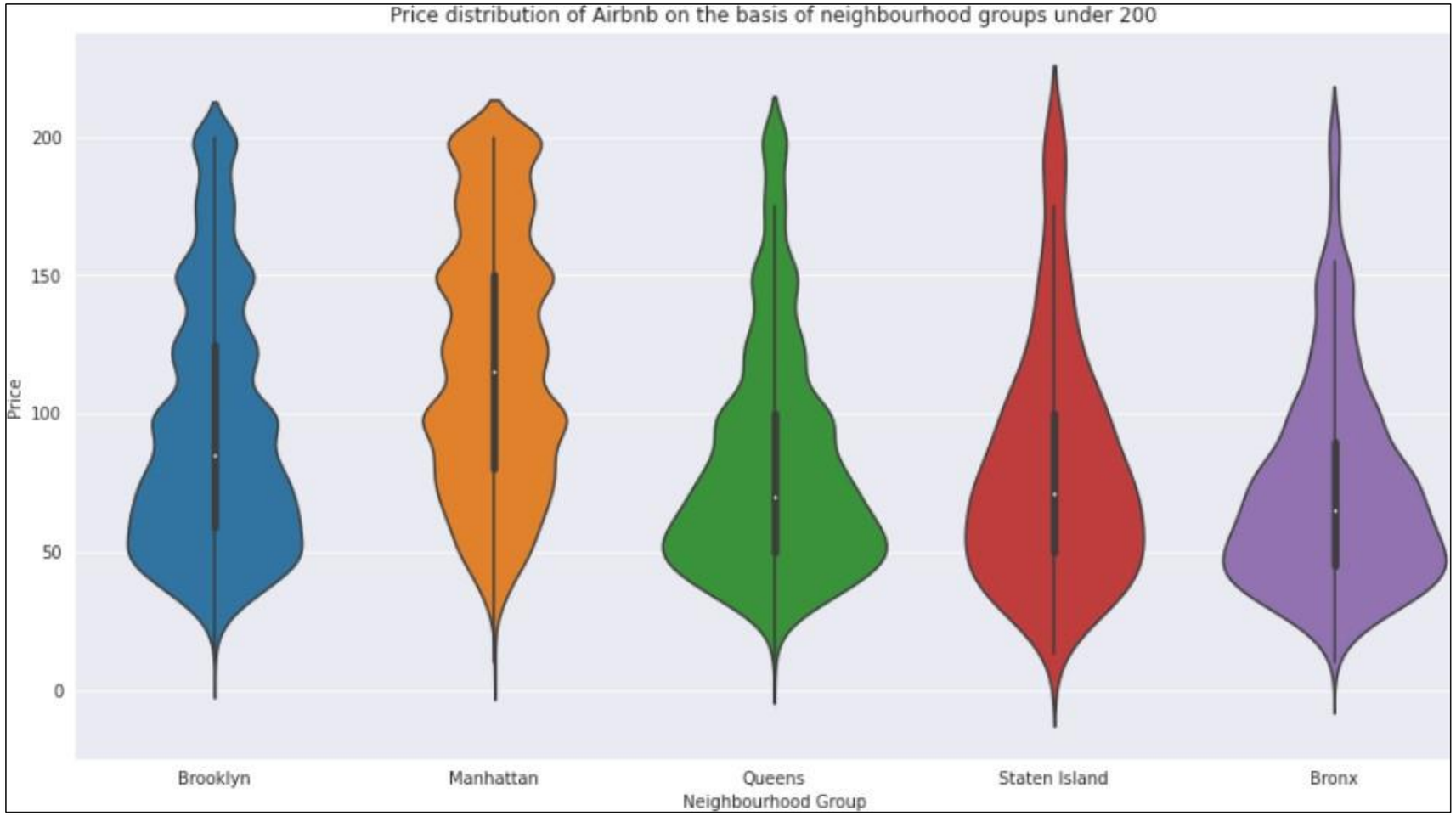


PRICE DISTRIBUTION OF AIRBNBs BELOW $200

Conclusion:

- Only a few Airbnbs had a price above $200, so to obtain a good distribution and to include maximum data we excluded the Airbnbs above the price range of $200.
- Below a price $100 we can see a larger distribution of Airbnbs.

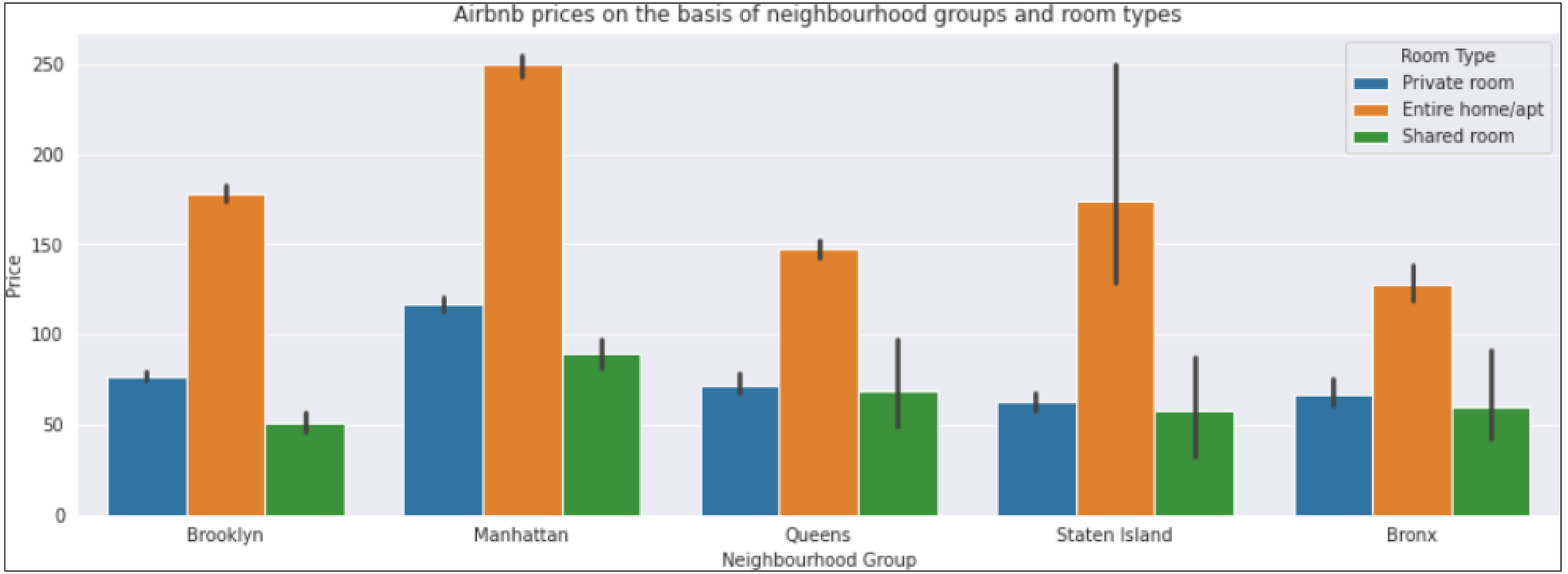
PRICE DISTRIBUTION OF AIRBNB ON THE BASIS OF NEIGHBOURHOOD GROUPS



Conclusion:

- Queens, Staten Island and Bronx have the price distribution below its mean price around 50.
- In comparison to other neighborhood groups, Brooklyn and Manhattan has a greater distribution above the price of 100 that concludes these areas have costlier hotels in the the New York City.

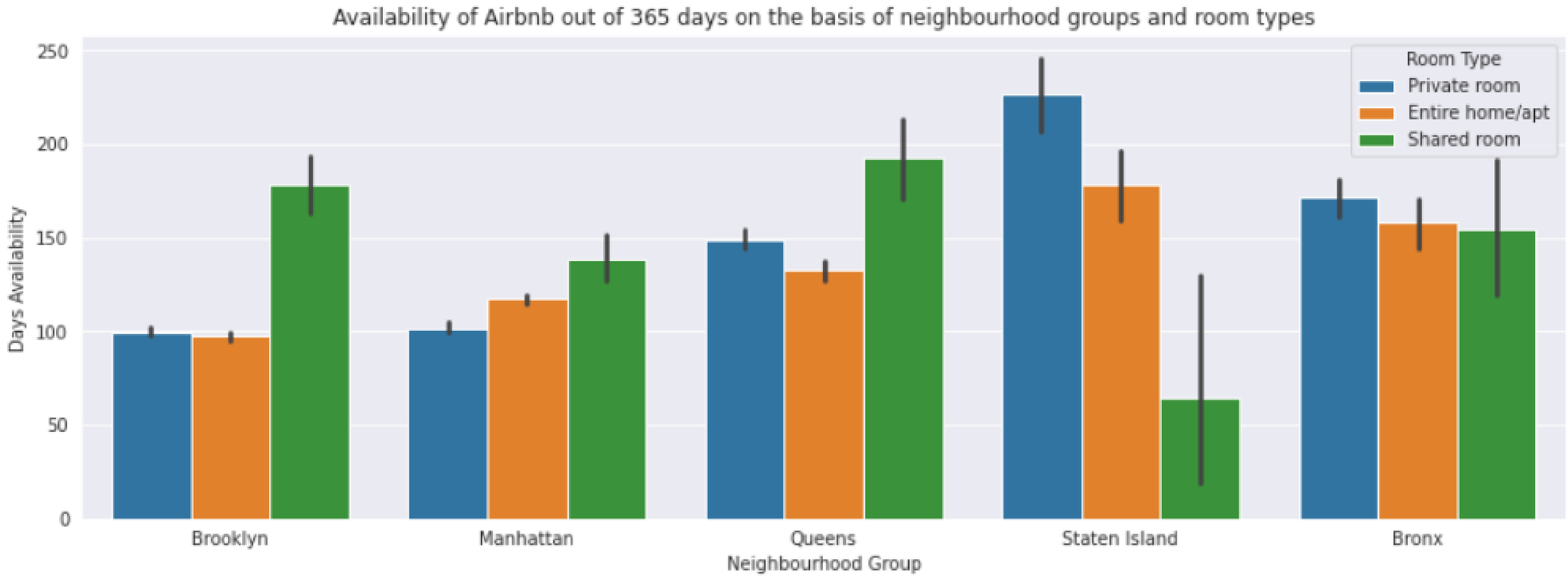
RELATION BETWEEN PRICE OF AIRBNB AND NEIGHBOURHOOD GROUP



Conclusion:

- **The price of entire home is very high in all the neighbourhood groups**
- **The price of the shared rooms are lesser throughout the neighbourhood groups**

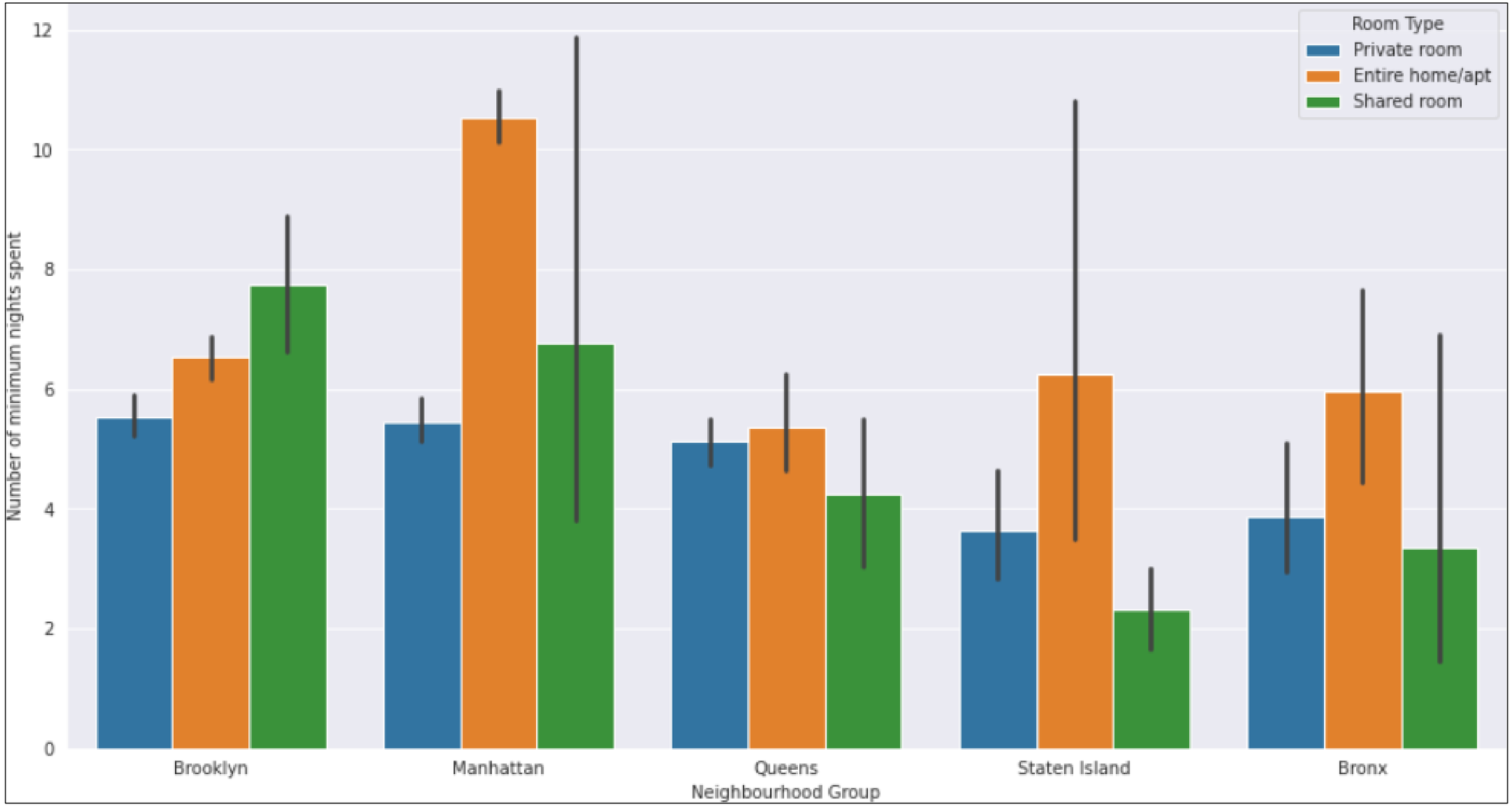
RELATION BETWEEN AVAILABILITY OF AIRBNB AND NEIGHBORHOOD GROUP



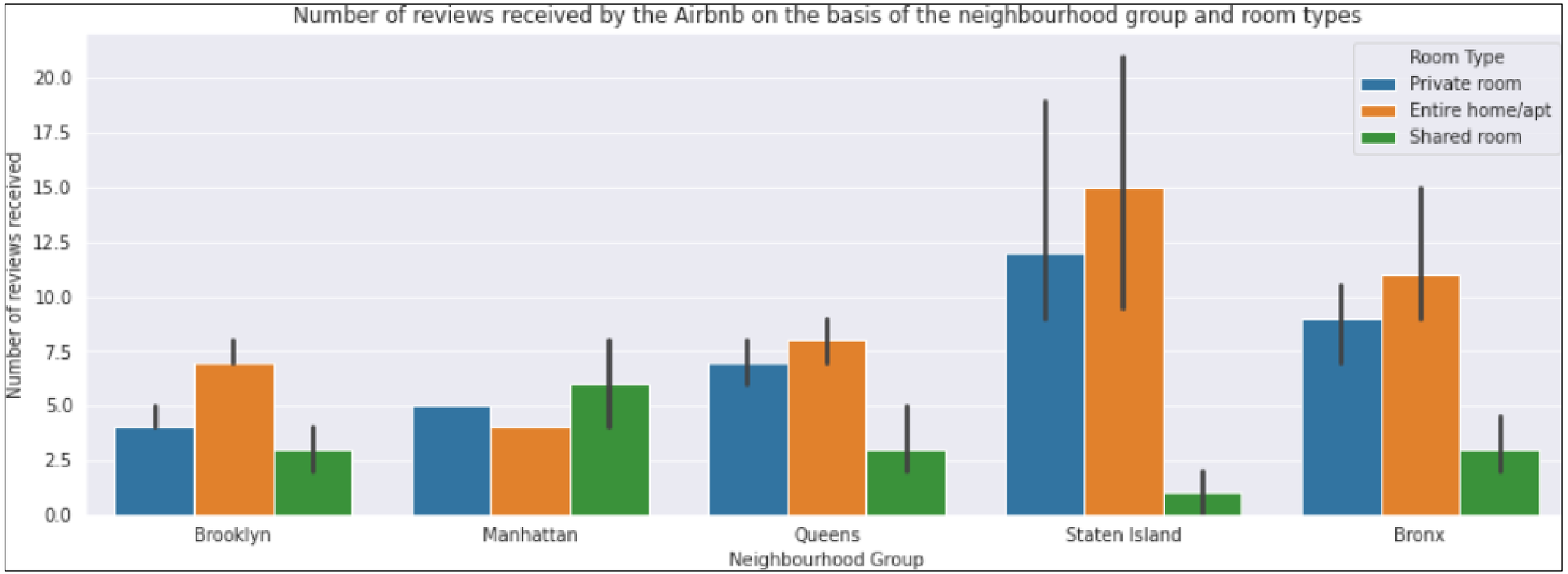
Conclusion:

- **Fewer Airbnbs are available in Brooklyn and Manhattan in comparison to others.**
- **Availability of shared rooms is quite high as compared to the other 2 room types.**

RELATION BETWEEN NUMBER OF MINIMUM NIGHTS SPENT AT EACH NEIGHBOURHOOD GROUP



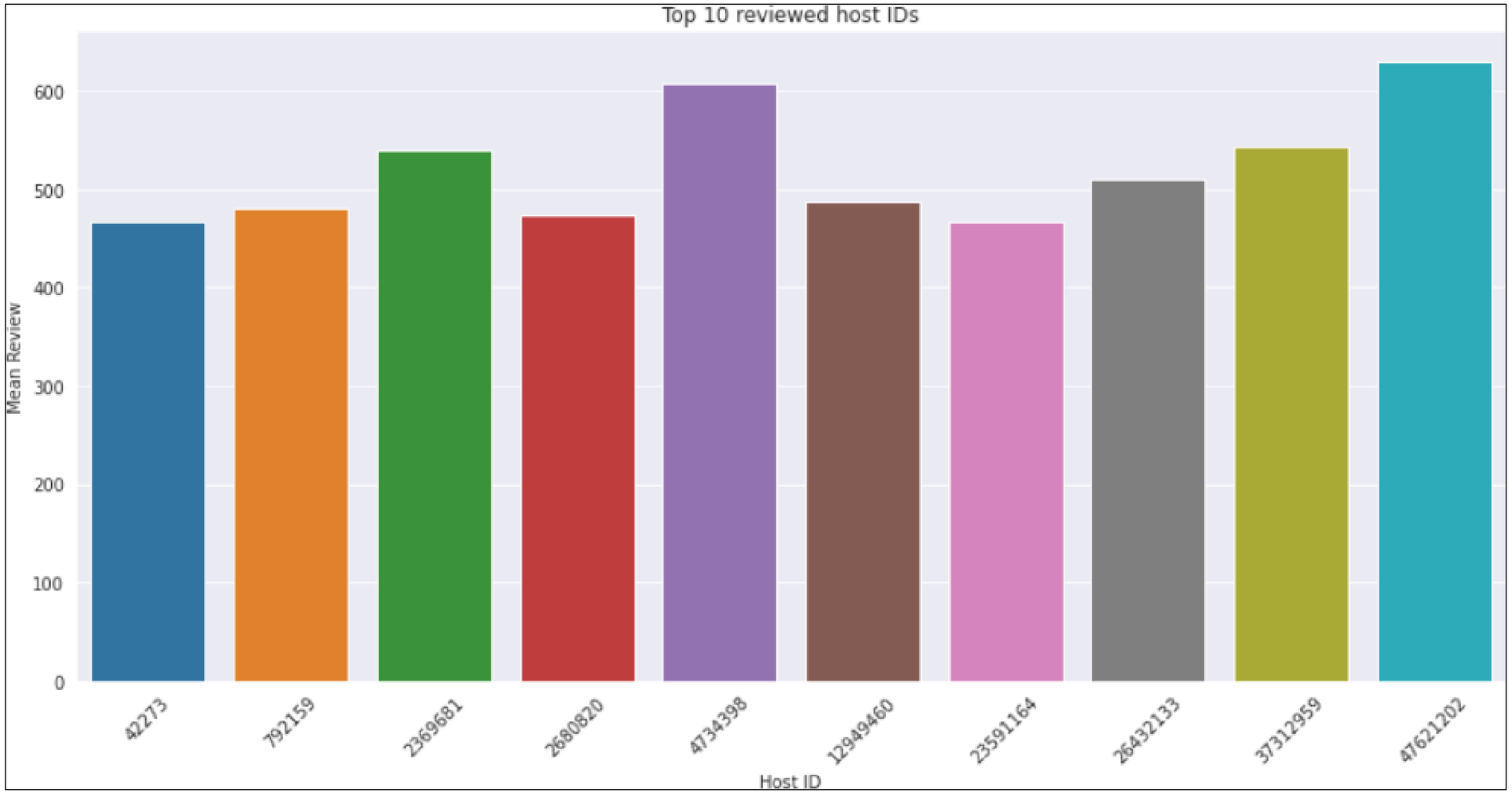
RELATION BETWEEN REVIEWS OF AIRBNB AND NEIGHBOURHOOD GROUP



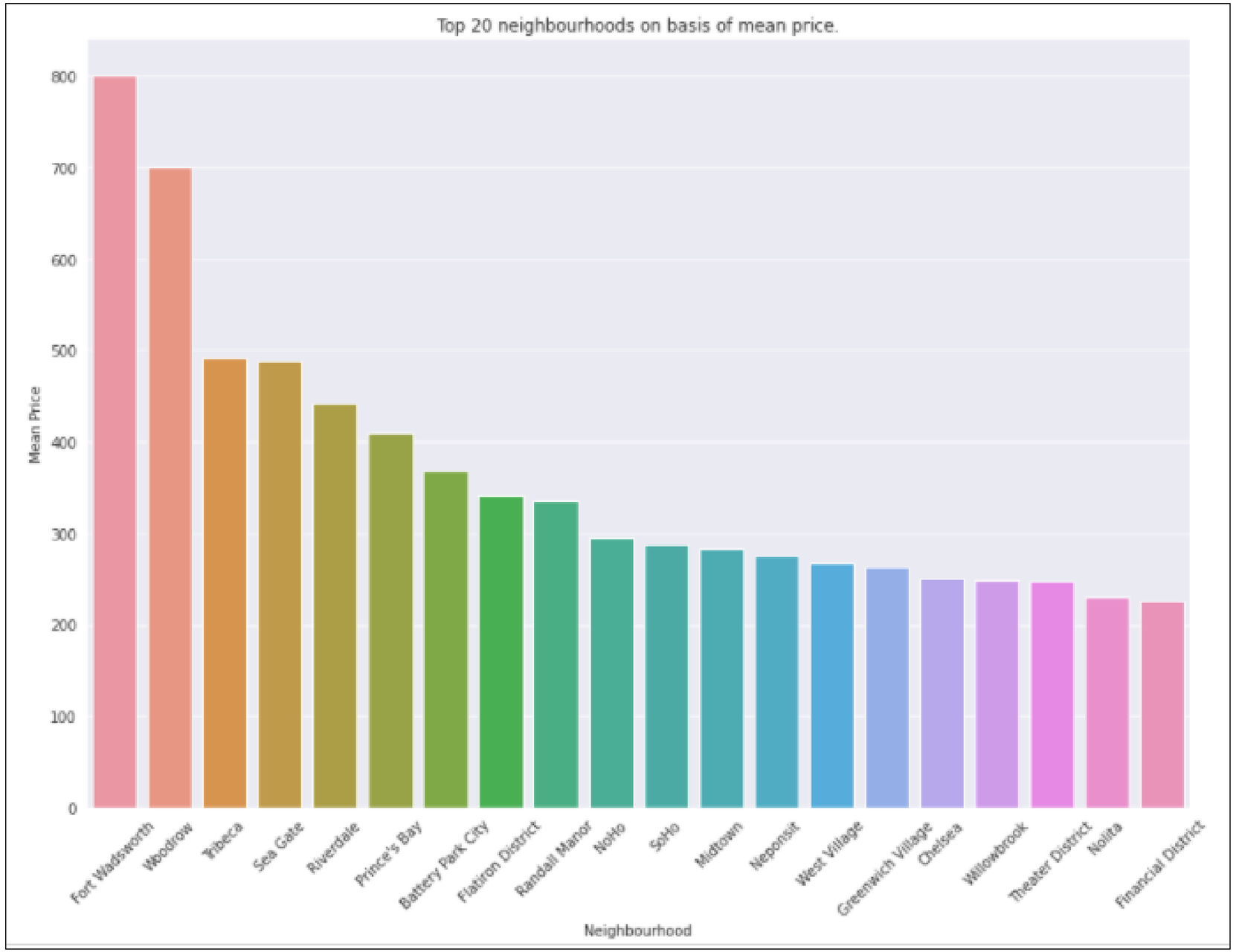
Conclusion:

- **People tend to review more in the Staten Island.**
- **People in the areas of Brooklyn and Manhattan tend to review comparatively lesser.**
- **Shared rooms are getting the least reviewed.**

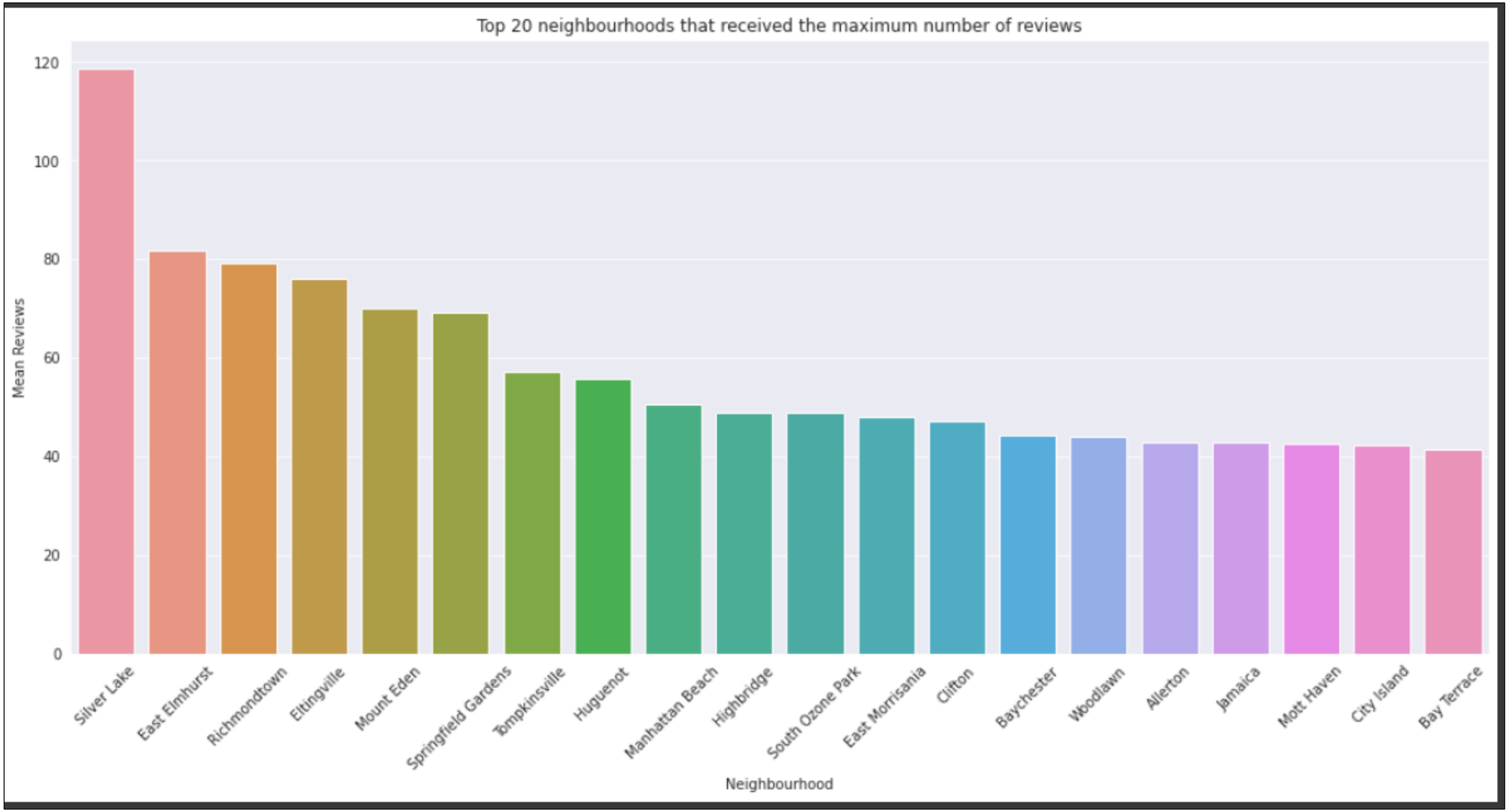
TOP 10 REVIEWED HOST IDs



TOP 20 NEIGHBOURHOODS ON THE BASIS OF MEAN PRICE



TOP 20 NEIGHBOURHOODS WITH MAXIMUM NUMBER OF REVIEWS



COMPARISON BETWEEN REVIEW AND PRICE OF AIRBNBs

Thank you