Insights from Airbnb Technical analysis in the Post Covid period

By

Harsha Wardhan

&

Samanyu Ghose

Problem Statement

- Airbnb has seen a major decline in revenue due to pandemic.

- As the effect of pandemic has started to decline and the restrictions have started to lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for the change

- The different leaders in Airbnb wants to understand some important insights based on various attributes in the dataset so as to increase the revenue

Outline

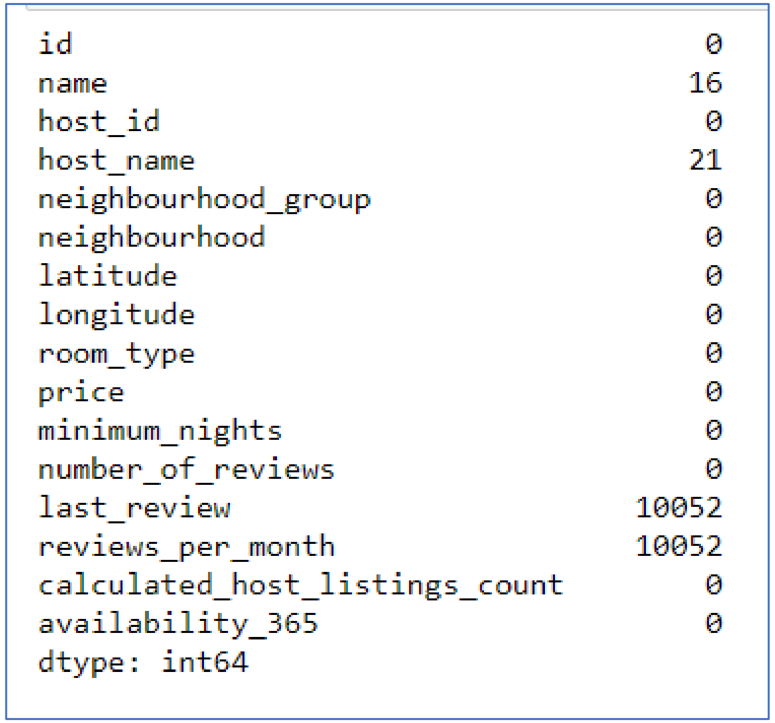
- ❑ Reading and understanding the data
  - Inspecting dataset
  - Missing value
  - Outlier

- ❑ EDA

- ❑ Analysis

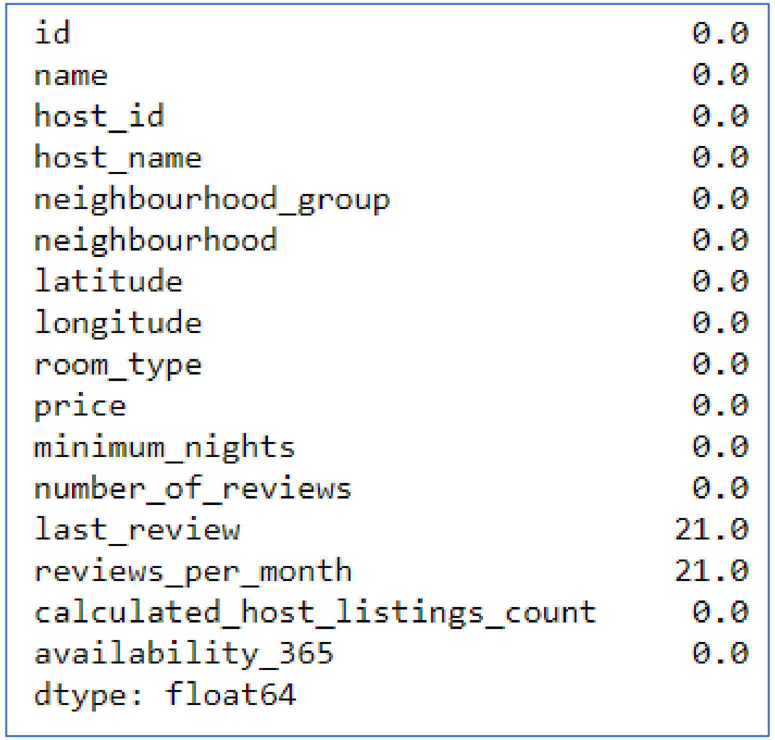
Handling missing values

Missing values by numbers



```
id                                 0
name                              16
host_id                            0
host_name                         21
neighbourhood_group                0
neighbourhood                      0
latitude                           0
longitude                          0
room_type                          0
price                              0
minimum_nights                     0
number_of_reviews                  0
last_review                    10052
reviews_per_month              10052
calculated_host_listings_count     0
availability_365                   0
dtype: int64
```

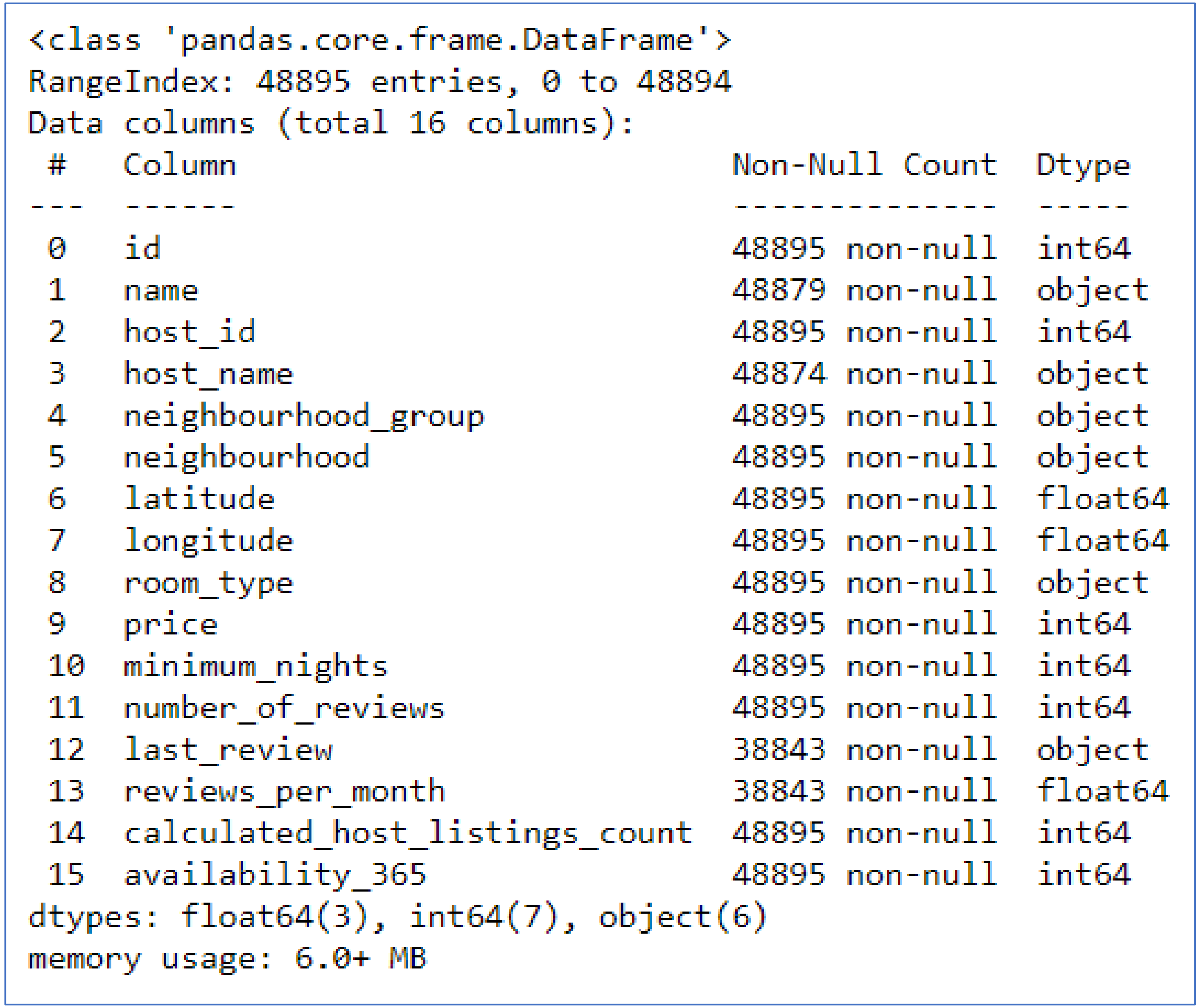
Missing values by percentage



```
id                                 0.0
name                               0.0
host_id                            0.0
host_name                          0.0
neighbourhood_group                0.0
neighbourhood                      0.0
latitude                           0.0
longitude                          0.0
room_type                          0.0
price                              0.0
minimum_nights                     0.0
number_of_reviews                  0.0
last_review                       21.0
reviews_per_month                 21.0
calculated_host_listings_count     0.0
availability_365                   0.0
dtype: float64
```

- There are 16 missing value in name, 21 in host_name, 10052 in last_review and 10052 reviews_per_month columns
- As percentage of missing values is not much high, so we have decided not to do any imputation or in other words, we have kept it as it is

Dataset



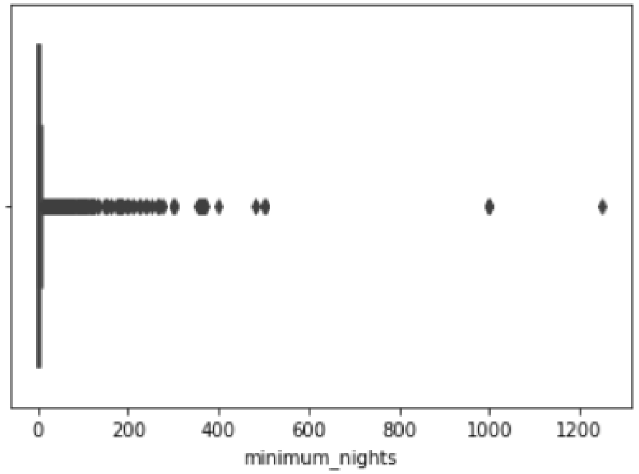
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   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
```

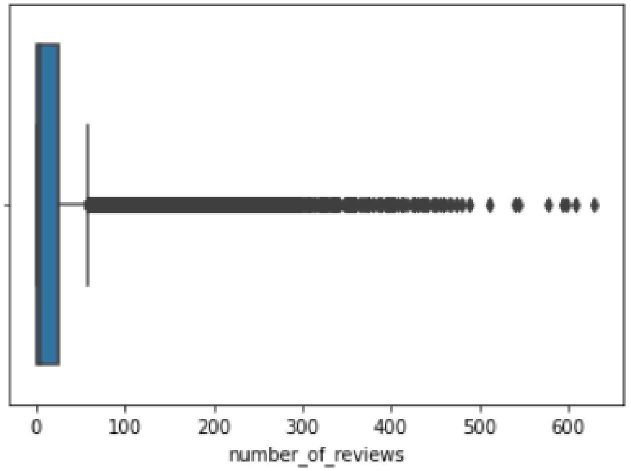
- Dataset has 48895 rows and 16 columns

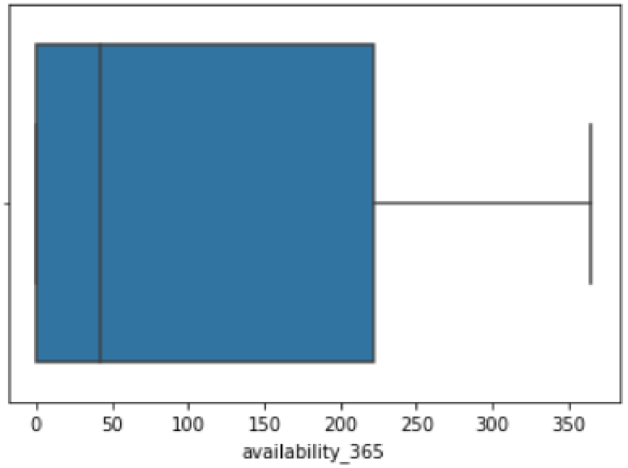
- There are missing values in ‘name’, ‘host_name’, ‘last_review’,’reviews_per_month’ columns which we have ignored

Handling Outliers

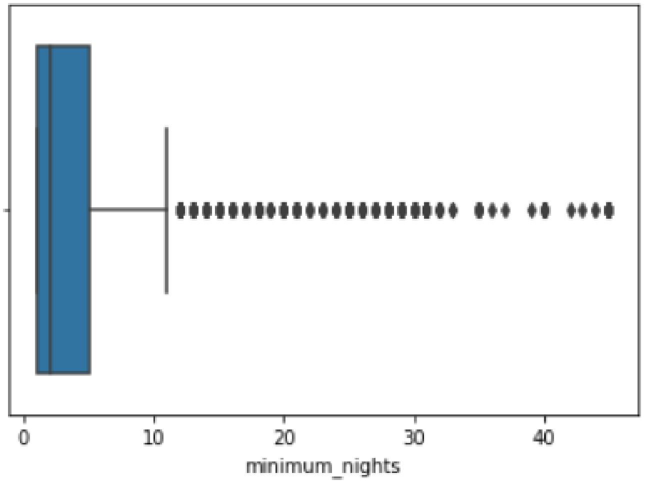
Before handling outlier

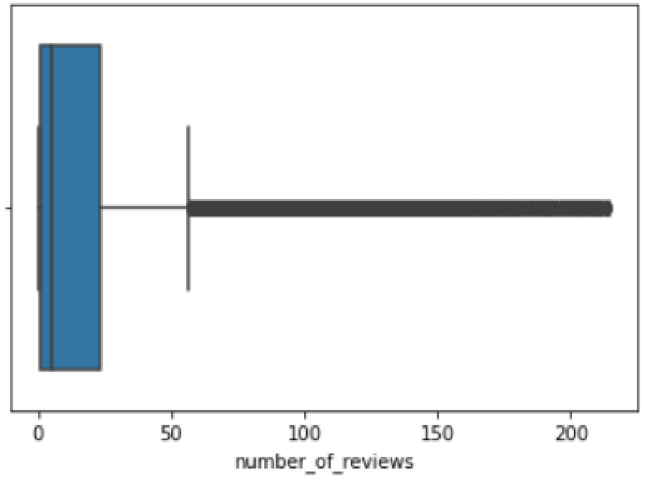






After handling outlier

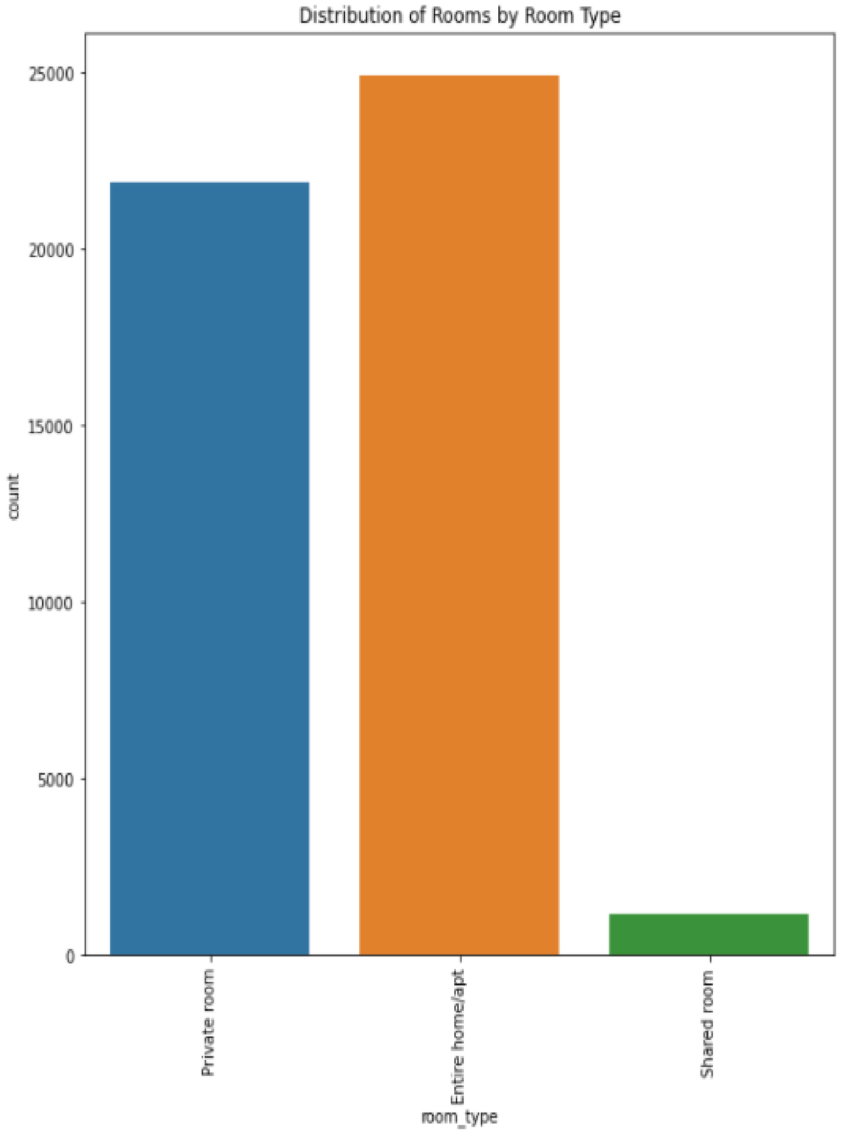


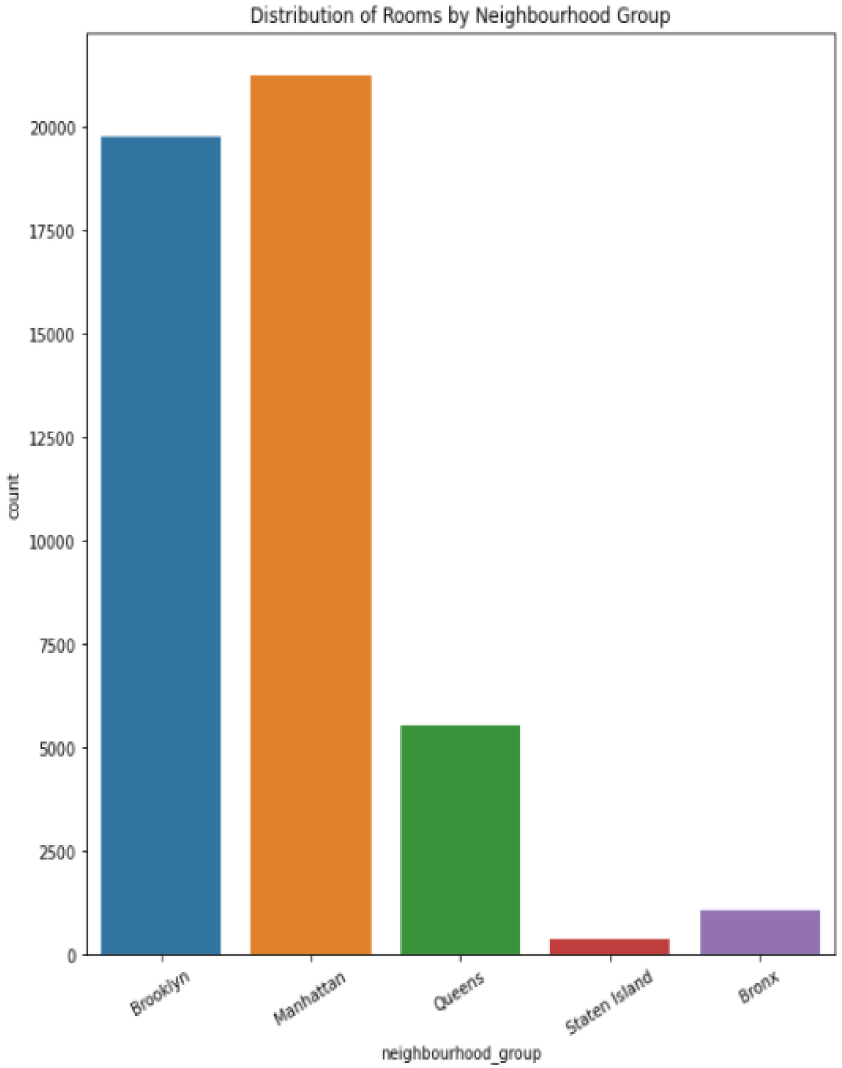


- We have handled outlier in ‘minimum_nights’ and ‘number_of_reviews by taking till 0.99 quantile and ignoring the rest

- Since there is no outlier in availability_365, so we have kept as it is

EDA : Univariate Analysis on Categorical variables

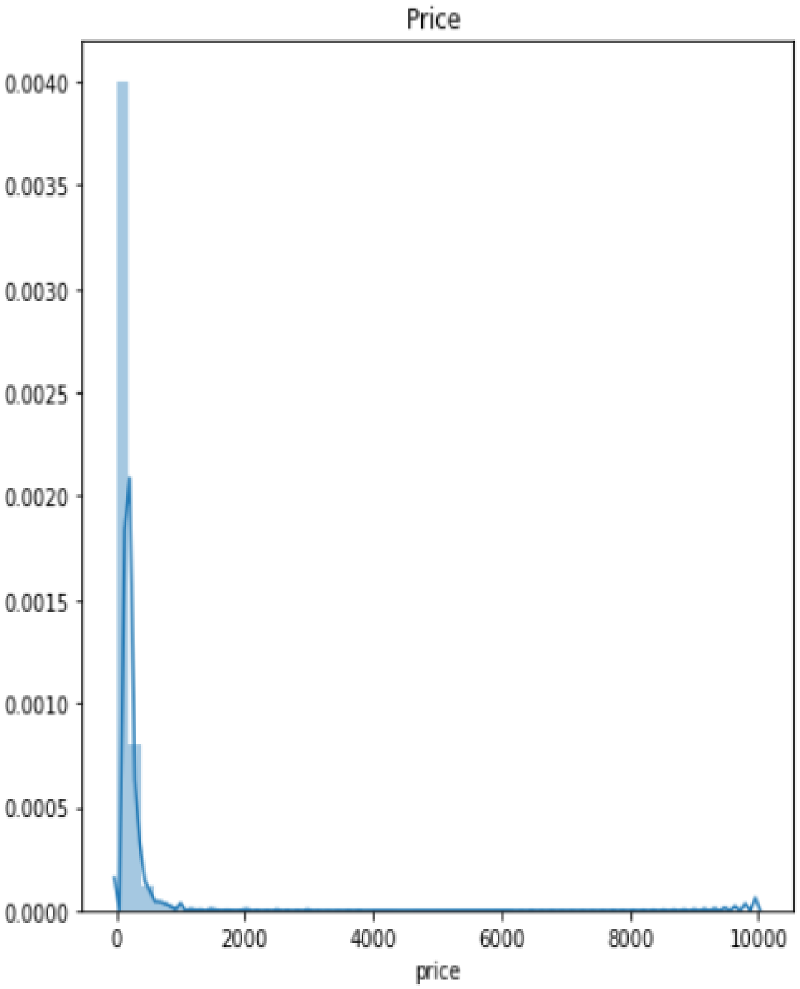


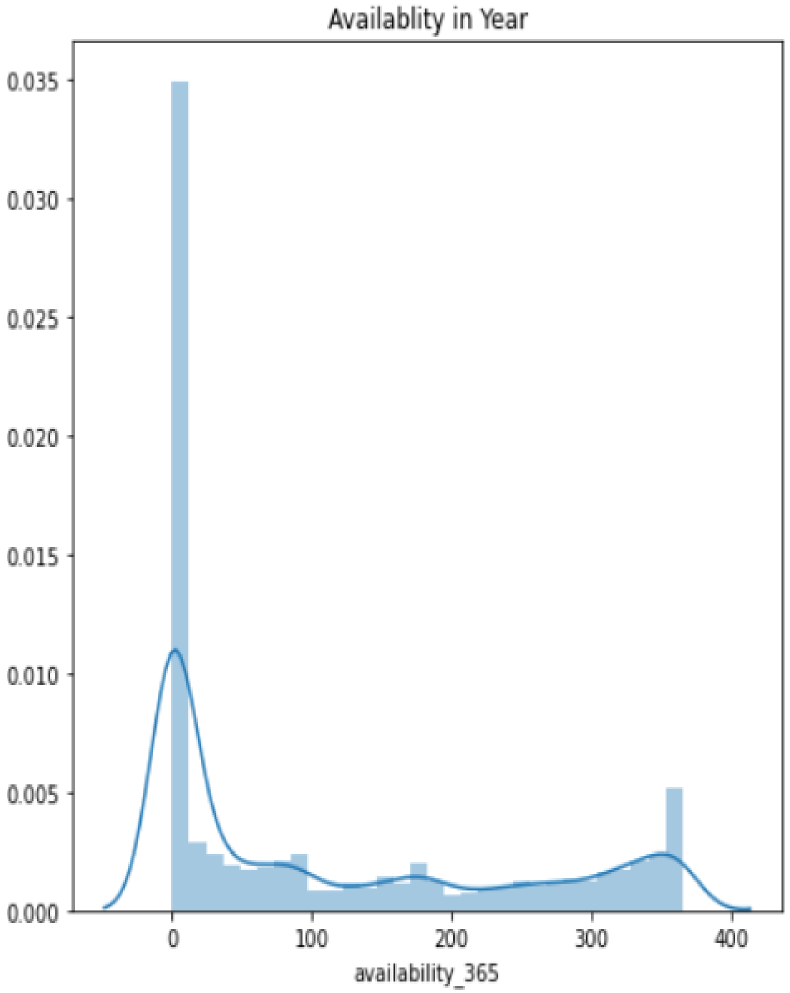


- Here we can see that 'Entire home/apt' is more than other types in general

- There are more rooms in 'Manhattan' than the other places

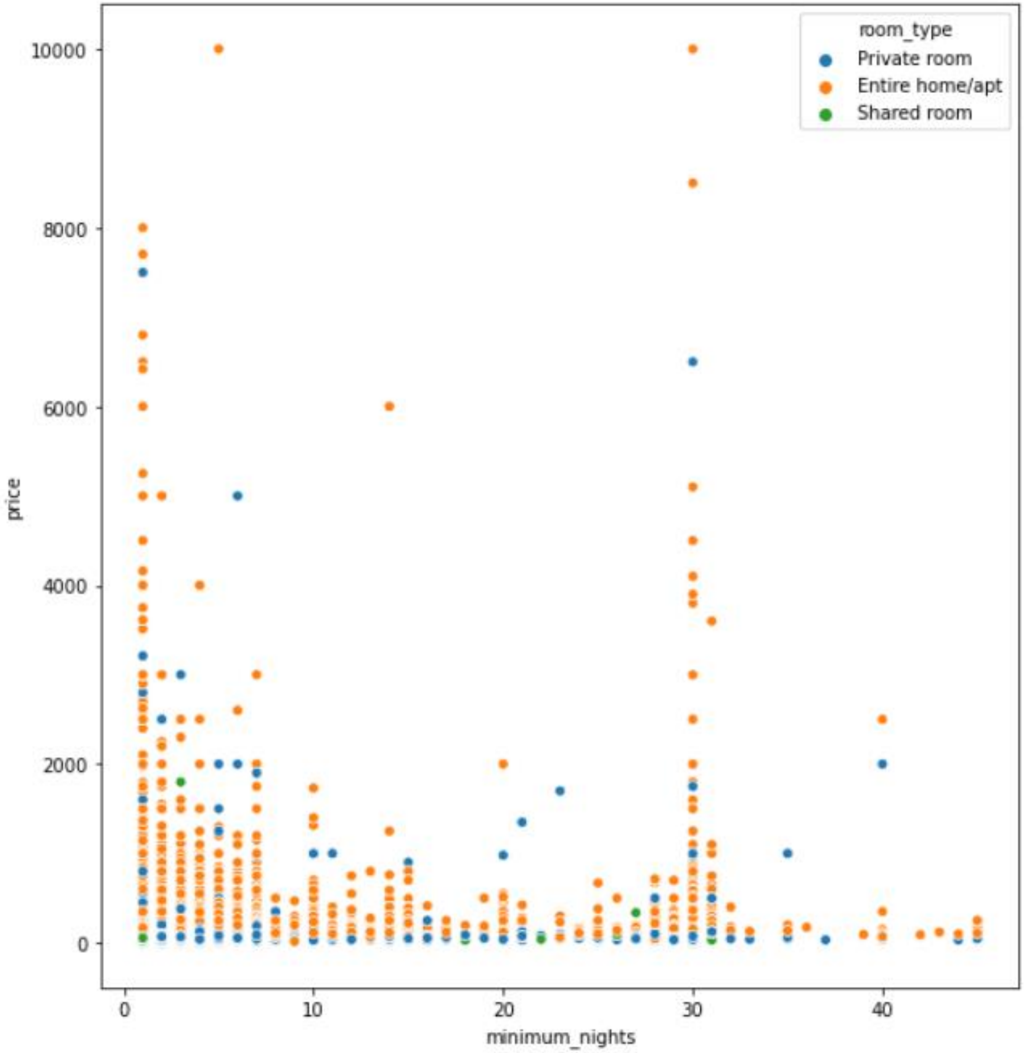
Univariate Analysis on Numerical Variables

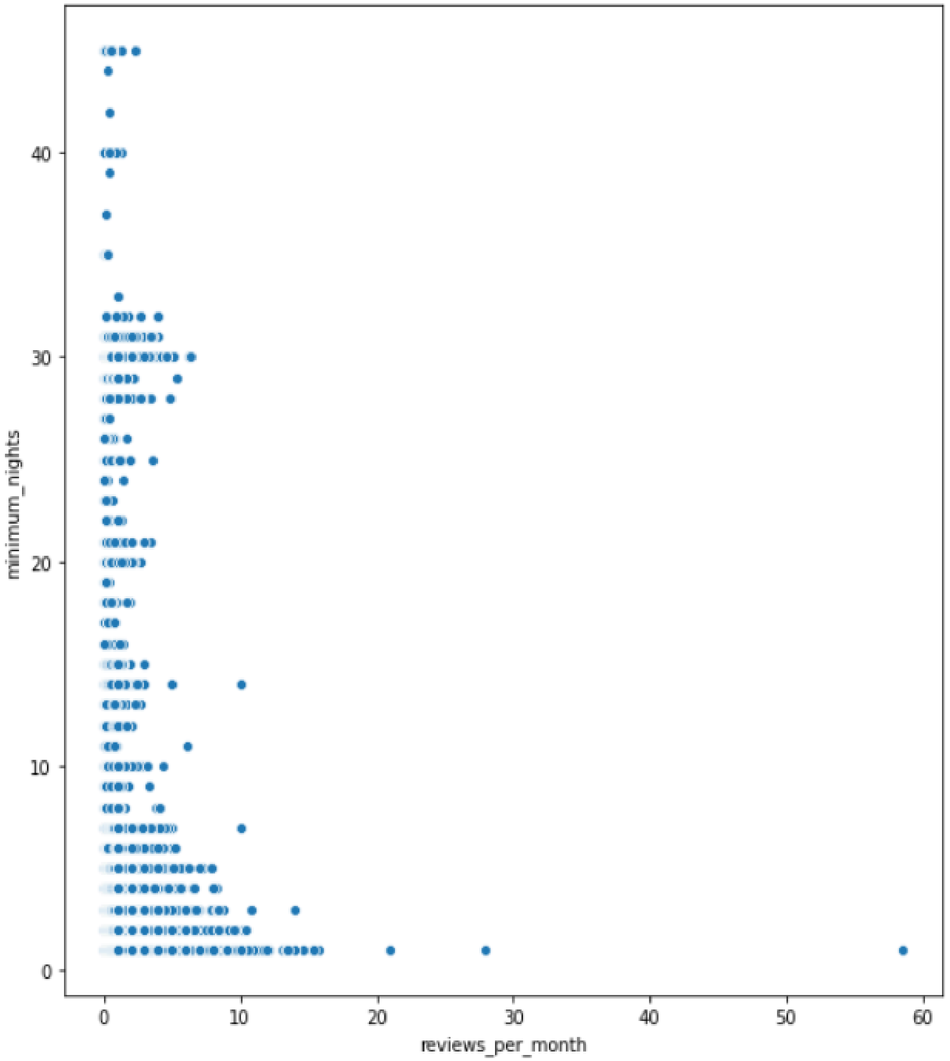




- Most prices are less than 2000 but some are even close to 10000 showing some luxury places
- Availability of room are mostly for 1 nights and some are little bit more than usual for 350 days

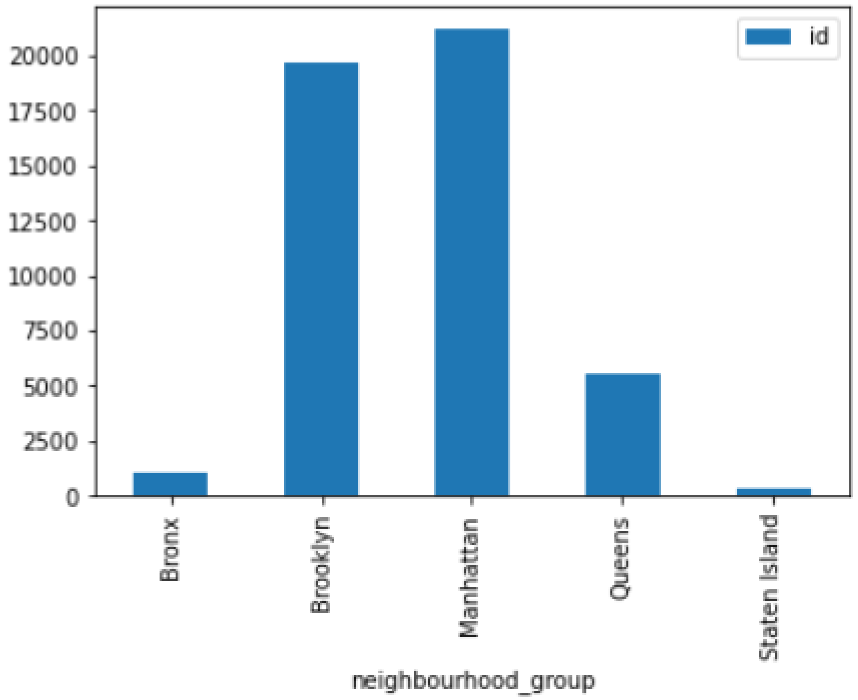
Bivariate Analysis (Price vs Minimum Nights & Reviews per month vs Minimum Nights)

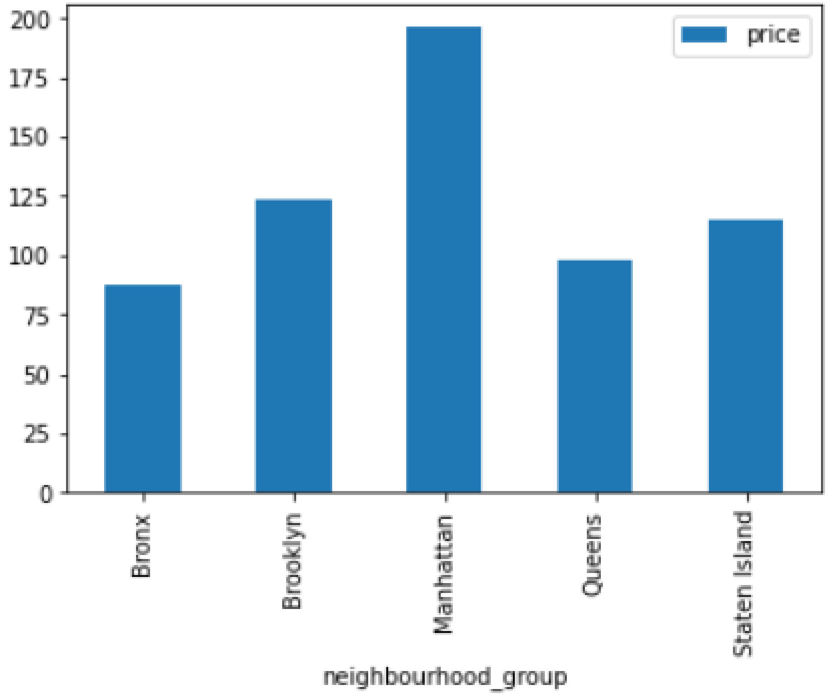




- In the first image,price for minimum_nights is some cases are very high for 1,5,30 nights
- In the second image, the number of review_per_month is very high for 1 nights as compared to more than 1 nights

Bivariate Analysis (Id vs neighborhood_group & Price vs neighborhood_group)

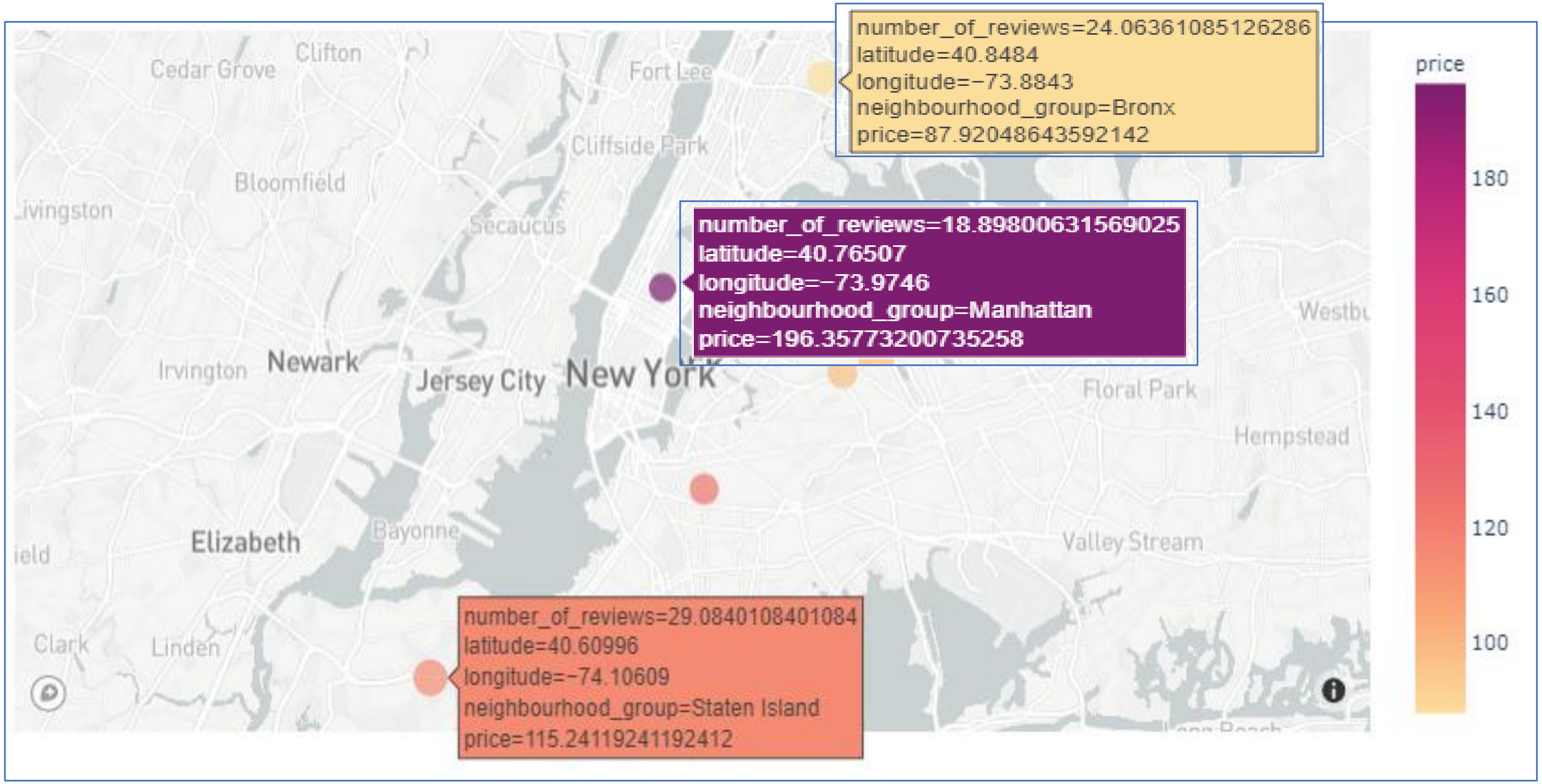




- Brooklyn and Manhattan have the highest number of listings
- Staten Island and Bronx have the least number of

- Manhattan has the highest average price
- Brooklyn is the next highest price
- Bronx and Staten Island is little cheaper compared to the other cities

Distribution of price with reviews through bubble map



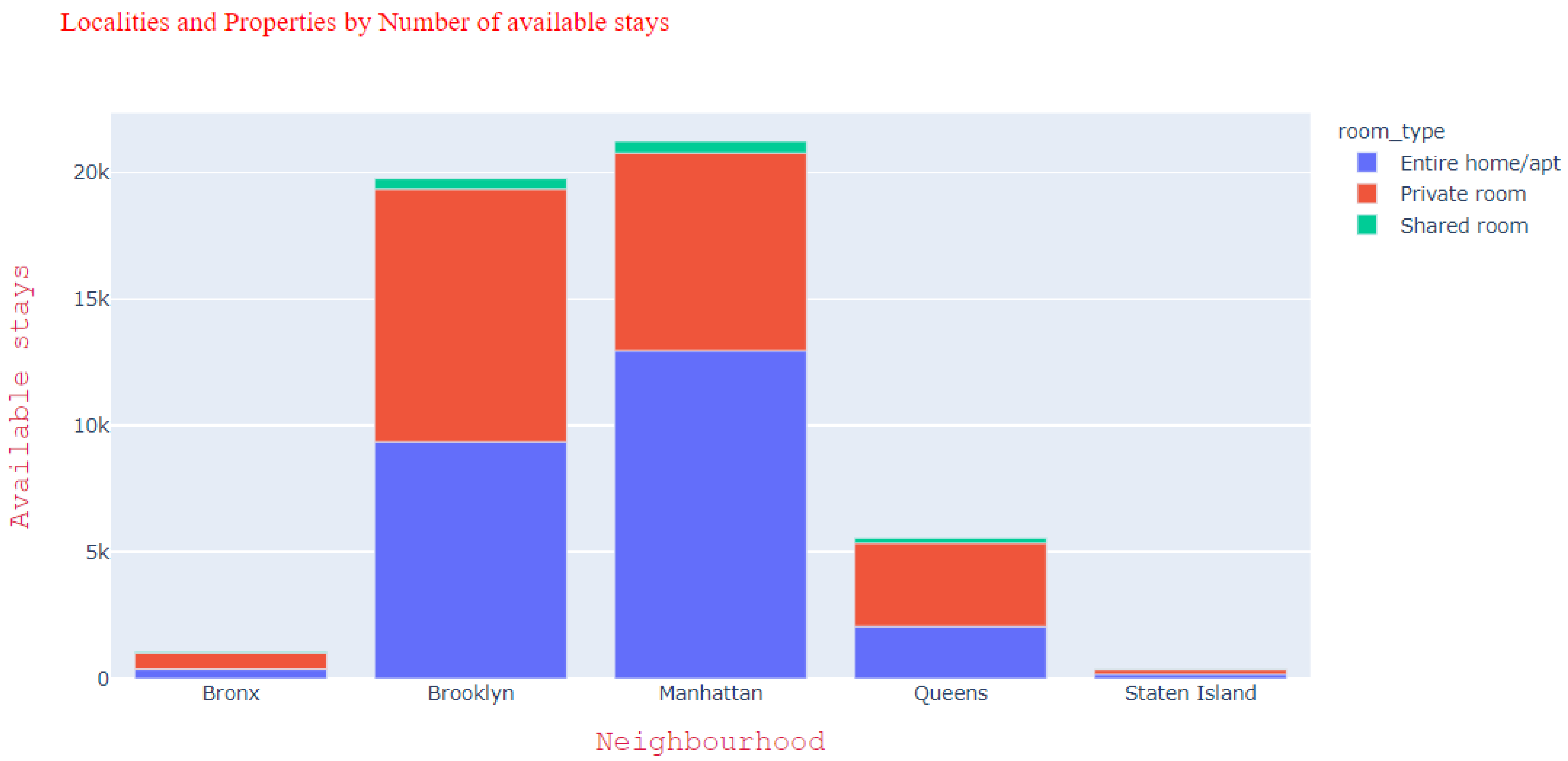
number_of_reviews=24.06361085126286
latitude=40.8484
longitude=−73.8843
neighbourhood_group=Bronx
price=87.92048643592142

number_of_reviews=18.89800631569025
latitude=40.76507
longitude=−73.9746
neighbourhood_group=Manhattan
price=196.35773200735258

number_of_reviews=29.0840108401084
latitude=40.60996
longitude=−74.10609
neighbourhood_group=Staten Island
price=115.24119241192412

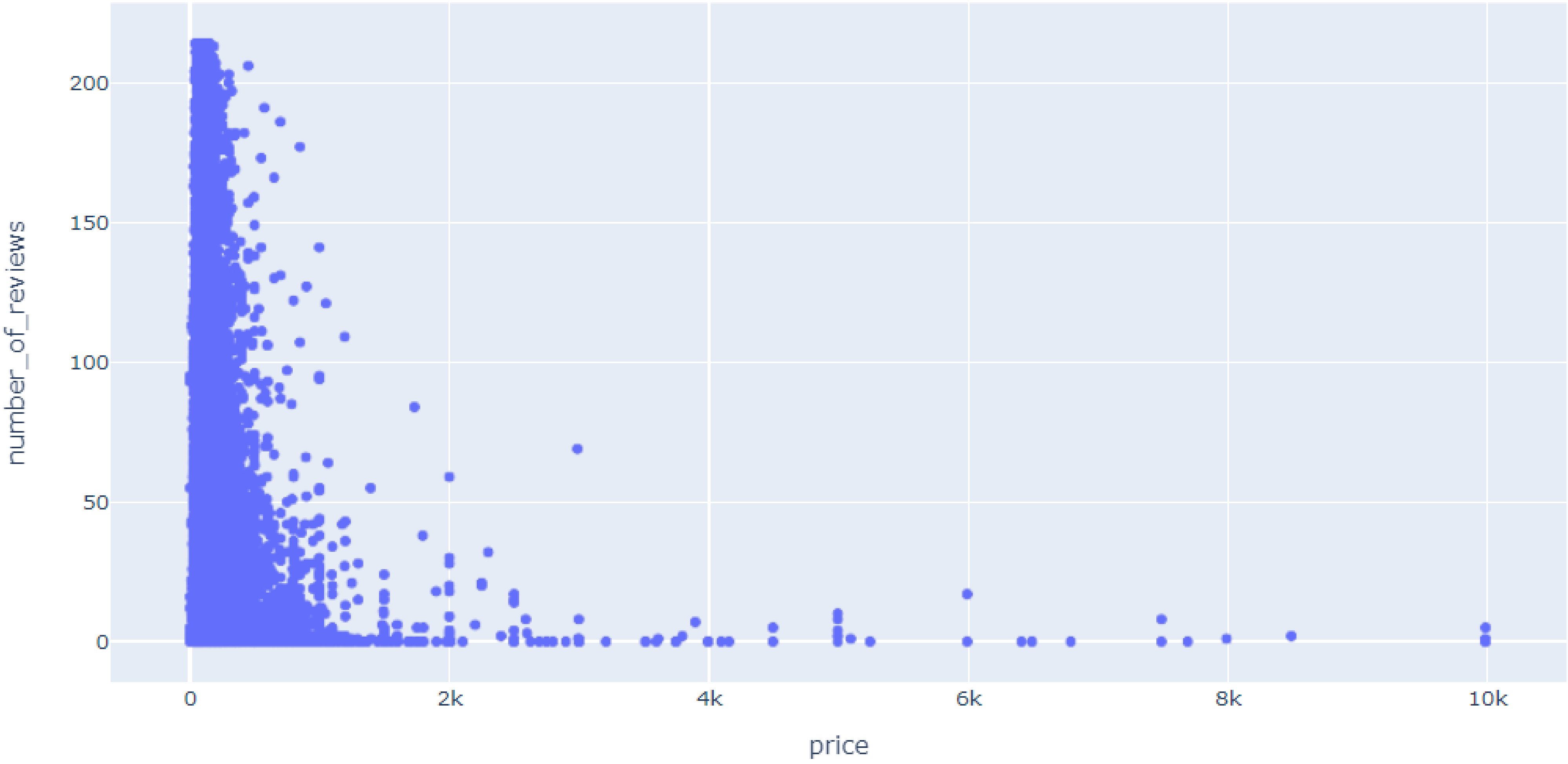
- Manhattan is the borrow with highest average price
- Brooklyn and Staten Island is the borrows with next highest price
- Queens and Bronx is little cheaper compared to the other borrows

Neighbourhood group: by available stays



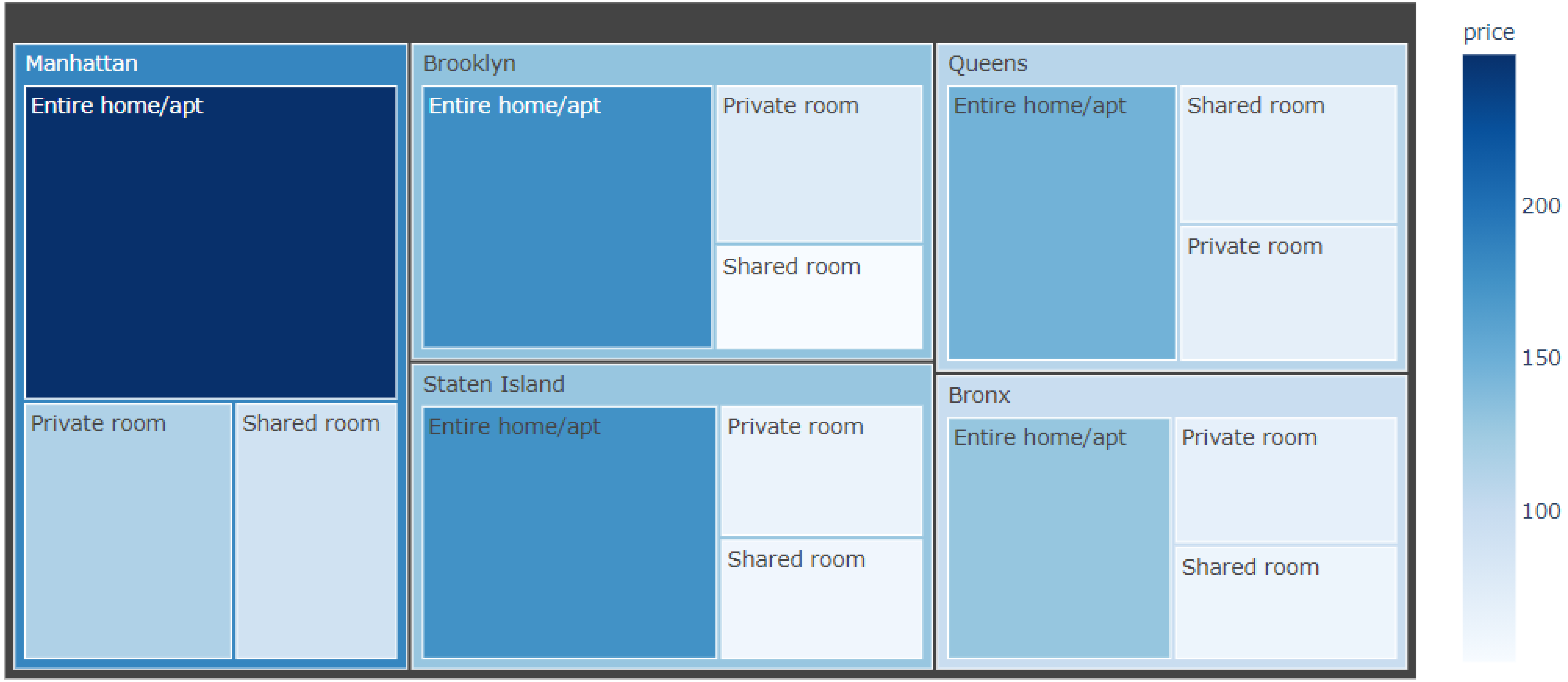
- In Manhattan, more number of homes and apartments are available for stays
- In Brooklyn, more number of private rooms are available for stays
- Bronx and Staten Island have very less number of listings
- Shared rooms are very less in number compared to the other two types of listings
- It is better to convert all shared rooms to private rooms

Price Range- Customer Preferences



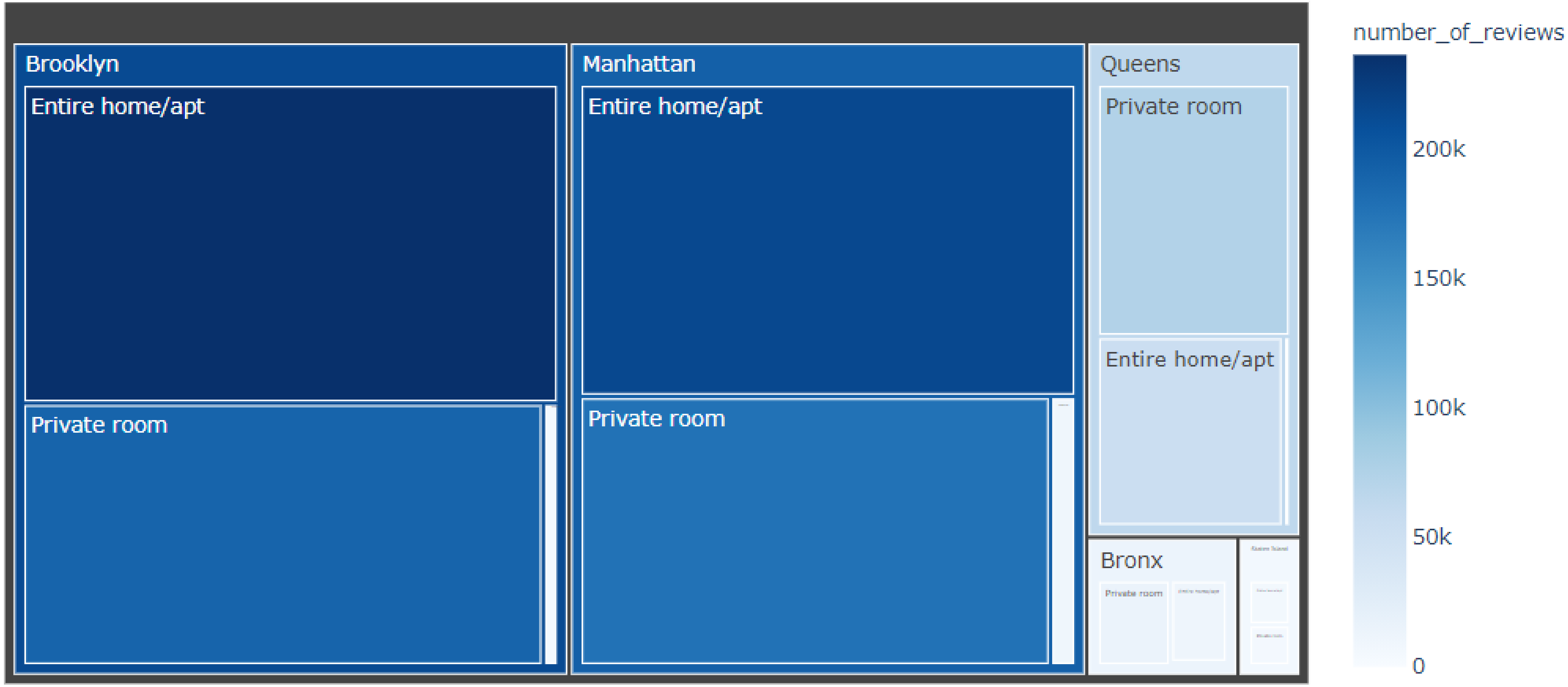
- Most of the people prefer the listings that cost less than $ 2000 per night

Average price of listing in neighborhood group



- In Manhattan, the average price of entire home/apt is higher than any other type of listings
- Shared rooms and private rooms are costing lesser than renting an entire home/apt
- In Queens, Staten island and Bronx, even the shared rooms are costing equal to private rooms

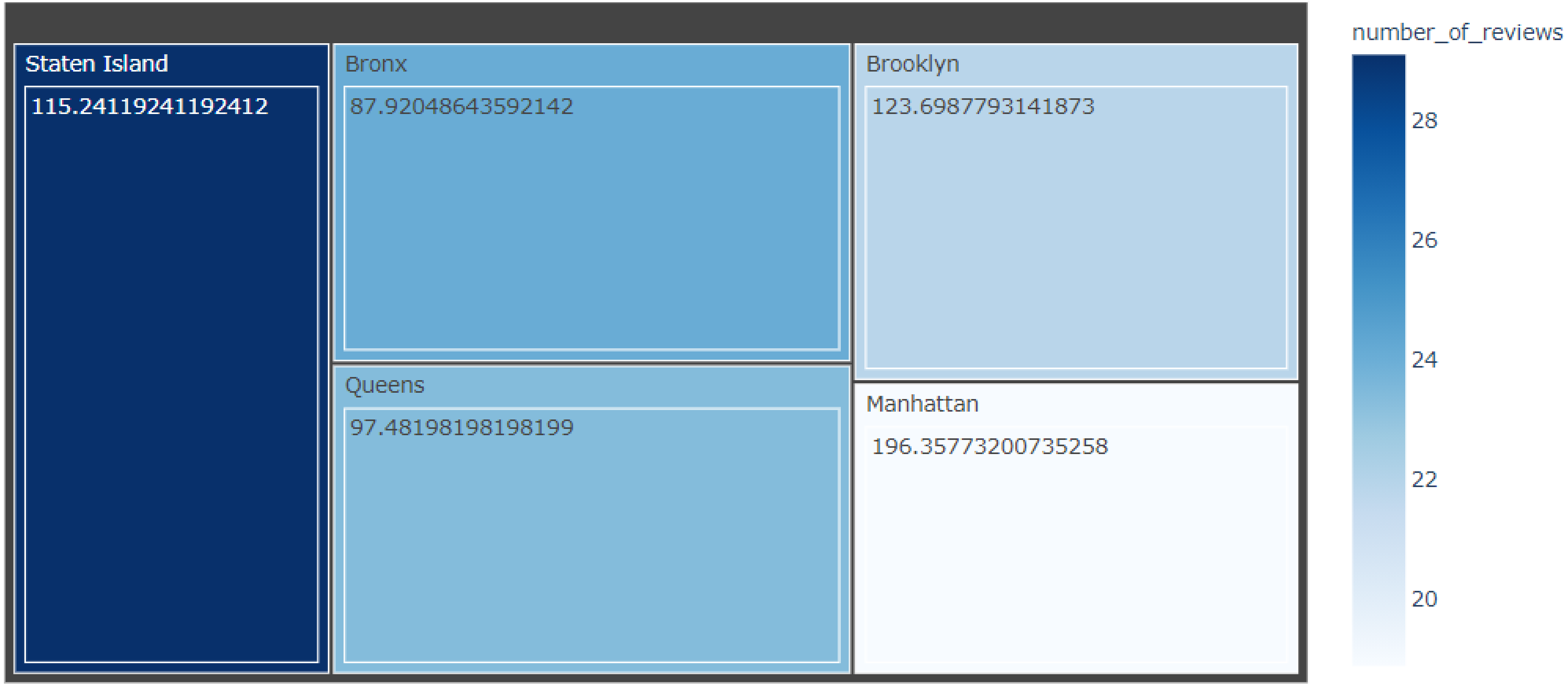
Most preferred areas and types of rooms



PREFERRED

- In Manhattan and Brooklyn, home stays are being preferred
- Private rooms are also being equally preferred as homes
- In Queens, Bronx and Staten island, Private rooms are being preferred than homes/apts

Preferred area based on average of price and reviews

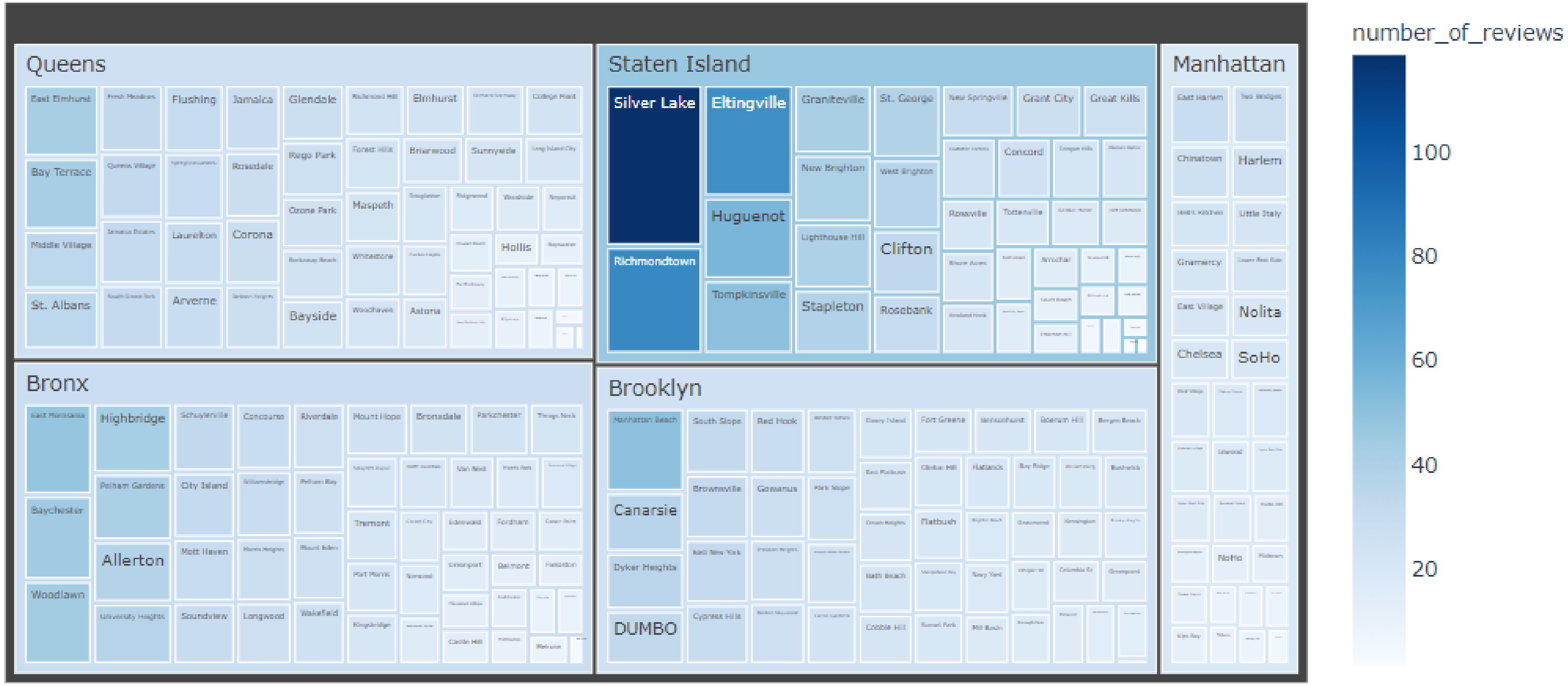


PREFERRED

- On the basis of average price and reviews, Manhattan is having the highest average followed by Brooklyn, Staten Island, Queens and Bronx

Preferred neighborhoods in each borrow

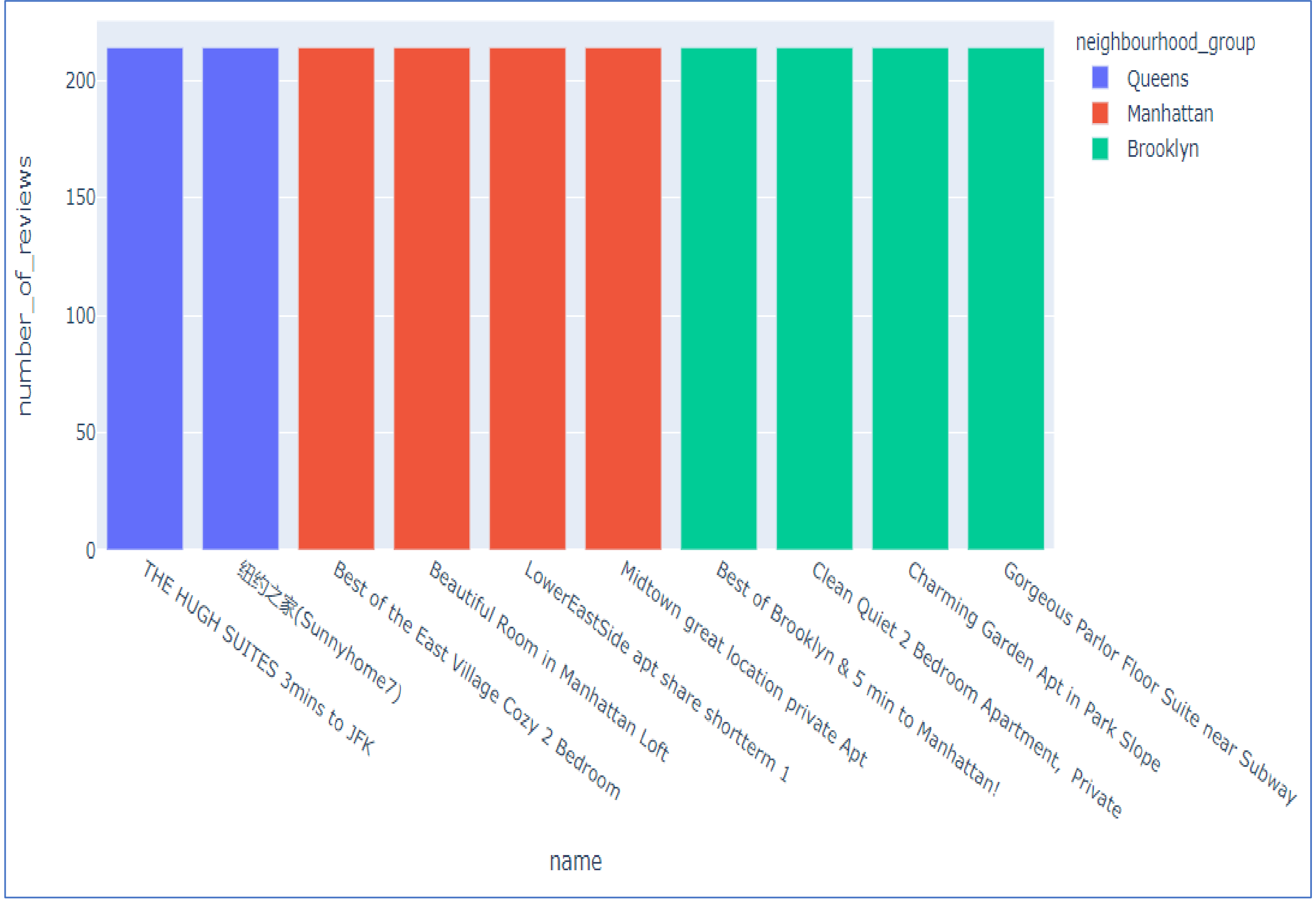
PREFERRED



The most preferred neighbourhoods in the entire Newyork city are

- Tribeca (Manhattan)
- Jamaica (Queens)
- South Slope (Brooklyn)
- Richmond Hill (Queens)
- East Elmhurst (Bronx)

Top 10 preferred stays in Newyork city



10 Most preferred stays are :-

Manhattan :-
1.Mistown great location private apt
2.LowerEastside apt share shortterm1
3.Beautiful Room in Manhattan Loft
4.Best of the East Village Cozy 2 Bedroom

Brooklyn :-
5.Gorgeous Parlor Floor Suite near Subway
6.Charming Garden Apt in Park Slope
7.Clean Quiet 2 Bedroom
Apartment, Private
8.Best of Brooklyn & 5 min to Manhattan

Queens :-
9. Sunnyhome7
10. The Huge Suites 3mins to JFK

Recommendation I

- Since Manhattan is most preferred city on the basis of reviews as well as price and Manhattan has most number of homes and apartments for stay, so it becomes a crucial business destination for the company

- During this pandemic period where people have less in their pockets to spend, prices can be slashed and kept around $ 2000 per night to give momentum to the business and these prices are generally preferred by the customers

- It’s better to convert the shared rooms to private rooms in Brooklyn and Manhattan as private rooms are most preferred

Recommendation II

- As Queens and Bronx have rooms with cheaper price, we can consider this as business opportunity to acquire more rooms and run the marketing campaign to increase the visibility of these two borrows

- Increase number of rooms in
  - Silver Lake (Staten Island)
  - Richmondtown (Staten Island)
  - Elting ville (Staten Island)
  - East Elmhurst (Queens)
  - Manhattan Beach (Brooklyn)

Appendix

Data Sources :-

- Here is a snapshot of our Data Dictionary
  - Host details such as host id, host name along with name of the stay and the listing ID based on type of rooms, location and area
  - Rooms type such as Private rooms, Entire Home/Apt, Shared rooms with prices of each kind of rooms along with minimum nights stayed by the customer and number of reviews provided by each customer along with last review date, reviews per month, availability of each type of room in neighborhood
  - Various areas comes under have been categorized into 5 neighborhood groups named Bronx, Brooklyn, Manhattan, Staten Island and Queens
  - Exact location of each neighborhood through latitude and longitude coordinates

- We have used the following data source
  - AB_NYC_2019

Continued…….

Appendix

Data Methodology :-

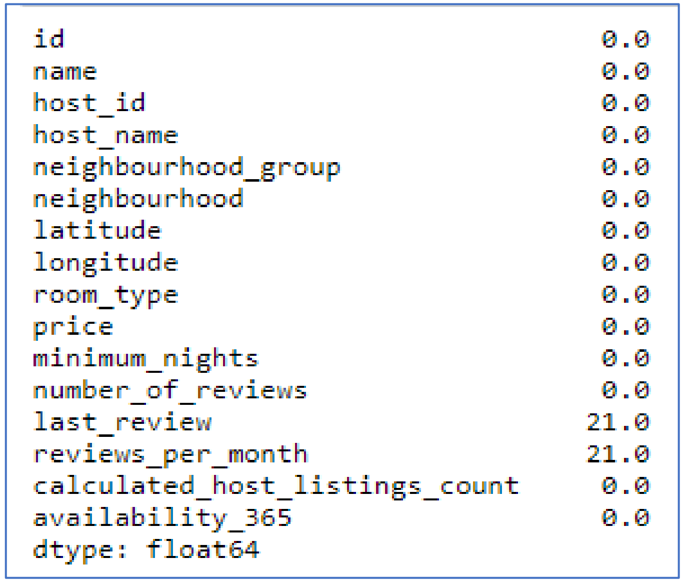
- We conducted a thorough analysis of AirBnb dataset. The process included :-
  - Cleaning the data set using visualization technique like boxplot, distplot to remove outliers
  - We did EDA through Univariate and Bi-variate analysis using countplot, distplot, scatterplot
  - We created various visualization to derive insights for further recommendation

Continued…….

Appendix

Data Assumption :-

- It is assumed that the company was achieving the desired revenue before Covid19 period
- Companies strategy will be based on the assumption that the travel will increase once things are in place post covid
- We also assumed that company has no plans to expand its business in new territories
- We have done analysis on missing value percentage and there were 21% of missing values in reviews_per_month and last_review but as the percentage was not much, we have not done any imputation



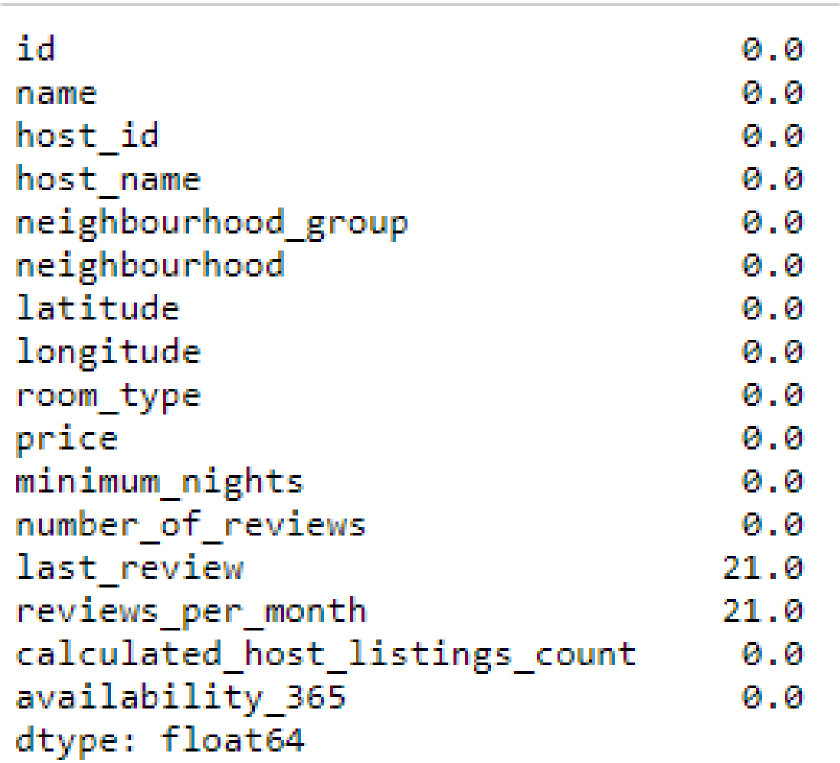
```
id                                 0.0
name                               0.0
host_id                            0.0
host_name                          0.0
neighbourhood_group                0.0
neighbourhood                      0.0
latitude                           0.0
longitude                          0.0
room_type                          0.0
price                              0.0
minimum_nights                     0.0
number_of_reviews                  0.0
last_review                       21.0
reviews_per_month                 21.0
calculated_host_listings_count     0.0
availability_365                   0.0
dtype: float64
```

Continued…….

Appendix

Data Assumption :-

- We have done analysis on missing value percentage and there were 21% of missing values in reviews_per_month and last_review but as the percentage was not much, we have not done any imputation



```
id                                 0.0
name                               0.0
host_id                            0.0
host_name                          0.0
neighbourhood_group                0.0
neighbourhood                      0.0
latitude                           0.0
longitude                          0.0
room_type                          0.0
price                              0.0
minimum_nights                     0.0
number_of_reviews                  0.0
last_review                       21.0
reviews_per_month                 21.0
calculated_host_listings_count     0.0
availability_365                   0.0
dtype: float64
```