METHODOLOGY

SUBMITTED BY :

ADITYA VIKRAM

SHIVAM TYAGI

SHOUVIK SHOME

PROBLEM STATEMENT

- For the past few months, Airbnb has seen a major decline in revenue
- 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

OBJECTIVE

To prepare for the next best steps that Airbnb needs to take as a business, you have been asked to analyze a dataset consisting of various Airbnb listings in New York. Based on this analysis, you need to give two presentations to the following groups.

1. Presentation - I

- **Data Analysis Managers:** These people manage the data analysts directly for processes and their technical expertise is basic.
- **Lead Data Analyst:** The lead data analyst looks after the entire team of data and business analysts and is technically sound.

2. Presentation - II

- **Head of Acquisitions and Operations, NYC:** This head looks after all the property and host acquisitions and operations. Acquisition of the best properties, price negotiation, and negotiating the services the properties offer falls under the purview of this role.
- **Head of User Experience, NYC:** The head of user experience looks after the customer preferences and also handles the properties listed on the website and the Airbnb app. Basically, the head of user experience tries to optimize the order of property listing in certain neighborhoods and cities in order to get every property the optimal amount of traction.

STEPS FOLLOWED

- **Data Understanding, Preparation, and Pre-Processing :**
  - Reading Data
  - Assigning correct datatypes
  - Treating Missing values
  - Treating outlier
- **Variable Transformation :**
  - Variable transformation and applying categorical variable transformations to turn into numerical data and numerical variable transformations to scale data
- **Exploratory Data Analysis :**
  - Univariate Analysis(Numerical and Categorical)
  - Bivariate and Multivariate Analysis

DATA ANALYSIS ALONG WITH CODE AND APPROACH

Data Understanding And Preparation

- First we imported relevant libraries
- After importing libraries we read the data and checked shape and size of the dataset
- Now, we checked datatypes of the column and converted “id” and “host id” to object datatype
- We also found that dataset contain few null values and outlier

```
 #   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)
```

```
# Analysing Numerical values
airbnb.describe()
```

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Handling Missing Values

- We identified two columns having equal percentage of missing values which were last_review and reviews_per_month of around 20.56%. Also, other two columns having minimal missing values which were host_name of 0.4% and name of the place of 0.3%.
- The values missing in “last_review” and “reviews_per_month” carrying NaN values is on purpose meaning they are not missing at random as these hosted sites/places have not receive any reviews from the customers. Hence, these places would be least preferred by the future customers and would also be facing bad business from our side
- There are in all 10052 unreviewed hosted sites on the account which is around 20% (10052/48895=20.55%)of all hosted sites

Handling Missing Values

```python
# Null Values percentage in each columns
x= (airbnb.isnull().sum()/len(airbnb)*100).sort_values(ascending=False)
x
```

```
last_review                       20.558339
reviews_per_month                 20.558339
host_name                          0.042949
name                               0.032723
id                                 0.000000
host_id                            0.000000
neighbourhood_group                0.000000
neighbourhood                      0.000000
latitude                           0.000000
longitude                          0.000000
room_type                          0.000000
price                              0.000000
minimum_nights                     0.000000
number_of_reviews                  0.000000
calculated_host_listings_count     0.000000
availability_365                   0.000000
dtype: float64
```

Imputing Missing Values

- We imputed the null values of column " reviews_per_month" with a zero
- Converted "Last_review" column to pandas dataframe and extracted year, month and date and deleted the original column and replaced NAN Values with " Not Reviewed"
- For the other 2 columns the null values were very low 0.03% of the entire data and upon checking those values , it looked like those were missed by chance and thus we imputed with mode

```
# Replacing the missing values of reviews_per_month with a zero

airbnb["reviews_per_month"] = airbnb.reviews_per_month.fillna(0)
```

```
#We will convert "Last-Review" columns to pandas dataframe and extract year ,Month and Day

airbnb['last_review_year'] = pd.DatetimeIndex(airbnb['last_review']).year
airbnb['last_review_month'] = pd.DatetimeIndex(airbnb['last_review']).month
airbnb['last_review_day'] = pd.DatetimeIndex(airbnb['last_review']).day
```

```
# Dropping the original last_review column

airbnb.drop('last_review', axis=1, inplace=True)
```

```
# Replacing the remaining missing values with a Not Reviewed option

airbnb['last_review_year'] = airbnb.last_review_year.fillna("Not Reviewed")
airbnb['last_review_month'] = airbnb.last_review_month.fillna("Not Reviewed")
airbnb['last_review_day'] = airbnb.last_review_day.fillna("Not Reviewed")
```

```
airbnb['host_name'].fillna(airbnb['host_name'].mode()[0], inplace=True)
```

```
airbnb['name'].fillna(airbnb['name'].mode()[0], inplace=True)
```

```
# recheck null values
(airbnb.isnull().sum()/len(airbnb)*100).sort_values(ascending=False)
```

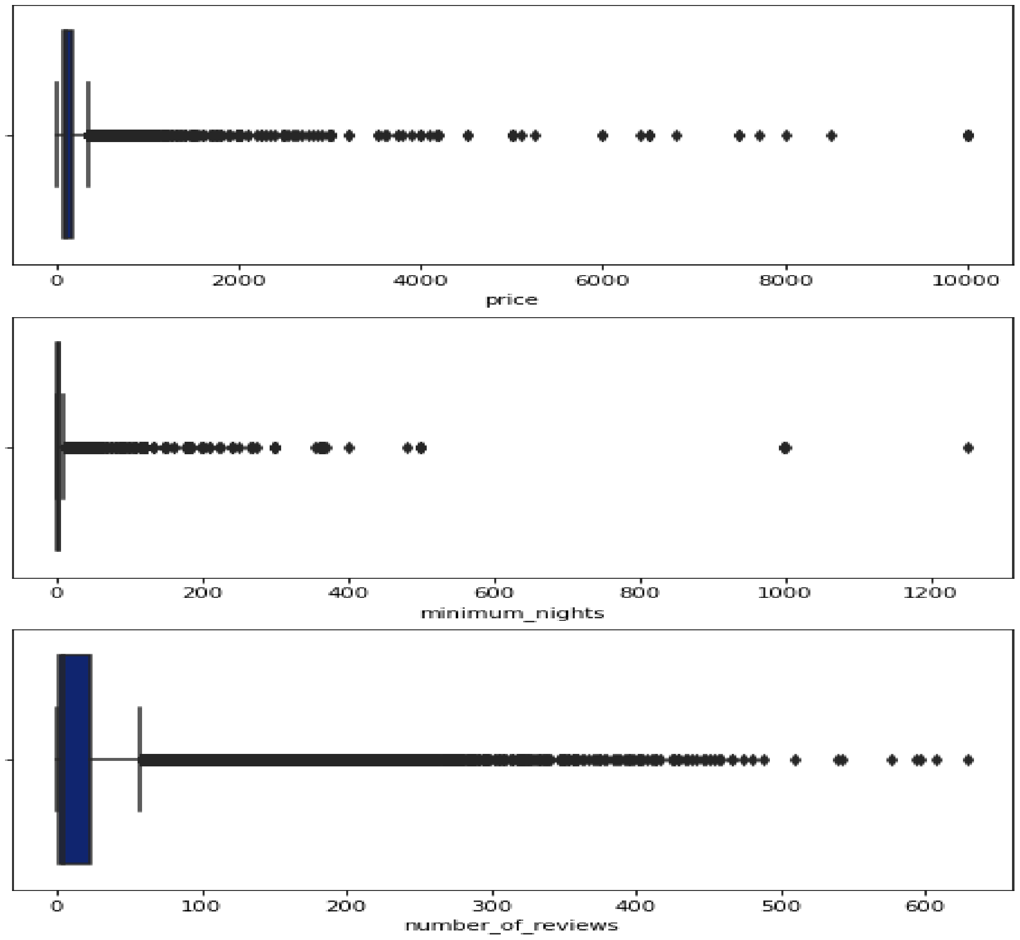
```
id                                0.0
name                              0.0
last_review_month                 0.0
last_review_year                  0.0
availability_365                  0.0
calculated_host_listings_count    0.0
reviews_per_month                 0.0
number_of_reviews             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

Handling Outliers

- We plotted box plot to check for outliers and also from describe function and there were multiple outliers
- We capped each column with outlier

```
plt.figure(figsize = (8,12))

plt.subplot(3,1,1)
sns.boxplot(airbnb['price'])

plt.subplot(3,1,2)
sns.boxplot(airbnb['minimum_nights'])

plt.subplot(3,1,3)
sns.boxplot(airbnb['number_of_reviews'])
plt.show()
```

```
plt.figure(figsize = (8,12))
plt.subplot(3,1,1)
sns.boxplot(airbnb['reviews_per_month'])

plt.subplot(3,1,2)
sns.boxplot(airbnb['calculated_host_listings_count'])

plt.subplot(3,1,3)
sns.boxplot(airbnb['availability_365'])
plt.show()
```



Outlier Treatment

```
# outlier treatment for price:
Q1 = airbnb.price.quantile(0.10)
Q3 = airbnb.price.quantile(0.90)
IQR = Q3 - Q1
airbnb = airbnb[(airbnb.price >= Q1 - 1.5*IQR) & (airbnb.price <= Q3 + 1.5*IQR)]

# outlier treatment for minimum_nights:
Q1 = airbnb.minimum_nights.quantile(0.10)
Q3 = airbnb.minimum_nights.quantile(0.90)
IQR = Q3 - Q1
airbnb = airbnb[(airbnb.minimum_nights >= Q1 - 1.5*IQR) & (airbnb.minimum_nights <= Q3 + 1.5*IQR)]

# outlier treatment for number_of_reviews:
Q1 = airbnb.number_of_reviews.quantile(0.10)
Q3 = airbnb.number_of_reviews.quantile(0.90)
IQR = Q3 - Q1
airbnb = airbnb[(airbnb.number_of_reviews >= Q1 - 1.5*IQR) & (airbnb.number_of_reviews <= Q3 + 1.5*IQR)]

# outlier treatment for reviews_per_month:
Q1 = airbnb.reviews_per_month.quantile(0.10)
Q3 = airbnb.reviews_per_month.quantile(0.90)
IQR = Q3 - Q1
airbnb = airbnb[(airbnb.reviews_per_month >= Q1 - 1.5*IQR) & (airbnb.reviews_per_month <= Q3 + 1.5*IQR)]

# outlier treatment for calculated_host_listings_count:
Q1 = airbnb.calculated_host_listings_count.quantile(0.10)
Q3 = airbnb.calculated_host_listings_count.quantile(0.90)
IQR = Q3 - Q1
airbnb = airbnb[(airbnb.calculated_host_listings_count >= Q1 - 1.5*IQR) &
                (airbnb.calculated_host_listings_count <= Q3 + 1.5*IQR)]
```

Variable Transformation

- We binned continuous numerical values columns such as "minimum_nights", "number_of_reviews","reviews_per_month","calculated_host_listings_count" and "availability_365“
- Once binning is completed we converted the datatype to object so that we can do categorical analysis and also we kept the original column so that we can do numerical analysis

```
# Creating minimum_nights into binned groups and storing it in another column
airbnb["minimum_nights_range"] = pd.cut(airbnb.minimum_nights,
                                        [0,10,20,30,40,50,60,70],
                                        labels=["<10", "10 to 20", "20 to 30", "30 to 40", "40 to 50", "50 to 60", "60+"])
airbnb["minimum_nights_range"].value_counts()
```

```
# Creating number_of_reviews into binned groups and storing it in another column

airbnb["number_of_reviews_range"] = pd.cut(airbnb.number_of_reviews,
                                           [0,50,100,150,200],
                                           labels=["<50", "50 to 100", "100 to 150", "150+"])
airbnb["number_of_reviews_range"].value_counts()
```

```
# Creating reviews_per_month into binned groups and storing it in another column

airbnb["reviews_per_month_range"] = pd.cut(airbnb.reviews_per_month,
                                           [0,2,4,6,8],
                                           labels=["<2", "2 to 4", "4 to 6", "6+"])
airbnb["reviews_per_month_range"].value_counts()
```

```
# Creating calculated_host_listings_count into binned groups and storing it in another column

airbnb["calculated_host_listings_range"] = pd.cut(airbnb.calculated_host_listings_count,
                                                  [0,3,6,9,12],
                                                  labels=["<3", "3 to 6", "6 to 9", "9+"])
airbnb["calculated_host_listings_range"].value_counts()
```

```
# Creating availability_365 into binned groups and storing it in another column
airbnb["availability_365_range"] = pd.cut(airbnb.availability_365,
                                          [0,100,200,300,400],
                                          labels=["<100", "100 to 200", "200 to 300", "300+"])
airbnb["availability_365_range"].value_counts()
```

Matrix Used For Analysis

- After we cleaned the data by handling null values , treating outlier and variable transformation, dataset was saved to do further analysis on Excel/Power BI/ Tableau
- In order to measure our analysis we created a 2x2 Matrix to provide us a direction while creating graphs using different Dimensions and Measures. This matrix involved the values needed to create the graphs with the combinations of,
- - Categorical & Numerical
- - Categorical & Categorical
- - Numerical & Numerical
- - Numerical & Categorical
- This turns out to be a road map for us, which helps in identifying which all dimensions and measures have been consolidated to get the insights from the data

Evaluation Of Methods

- The matrix which we created was evaluated at every step by creating relevant questions to see what we are trying to extract from the raw data. More importantly, to extract the relevant information that we want to recommend to our target audience. Below are the list of some questions that we curated to drive the above matrix for creating graphs.

Questions:
Which type of hosts to acquire more and where?
What are the neighbourhoods they 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.
How to get unpopular properties more traction?
What are the most popular localities and properties in New York currently?
Is there any correlation between the prices and reviews or other parameters
Which are the room types that are not performing well?
What are the price range preferred by customers?
Which properties and room types have more or less minimum nights stay?

Univariate , Bivariate and Multivariate Analysis

Univariate Analysis-Numerical Column

We used distribution and count plot from seaborn for numerical and categorical column

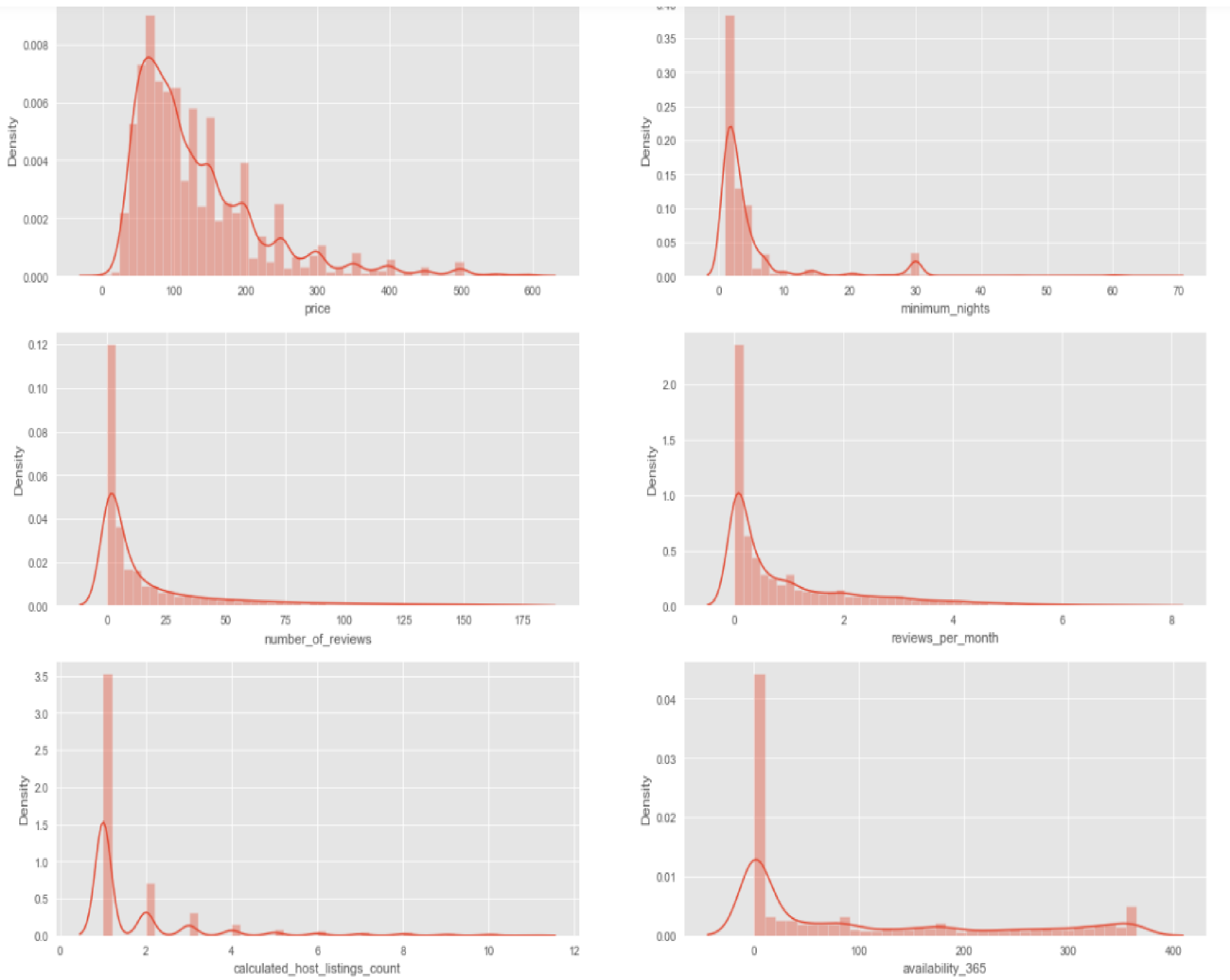
Insights which we derived from univariate analysis are ;

- Majority of Price ranges from 10$ to 100$ and it some cases it goes up to 600 $
- Minimum nights which user spend is 1-3 days
- Generally 1 review is given per month

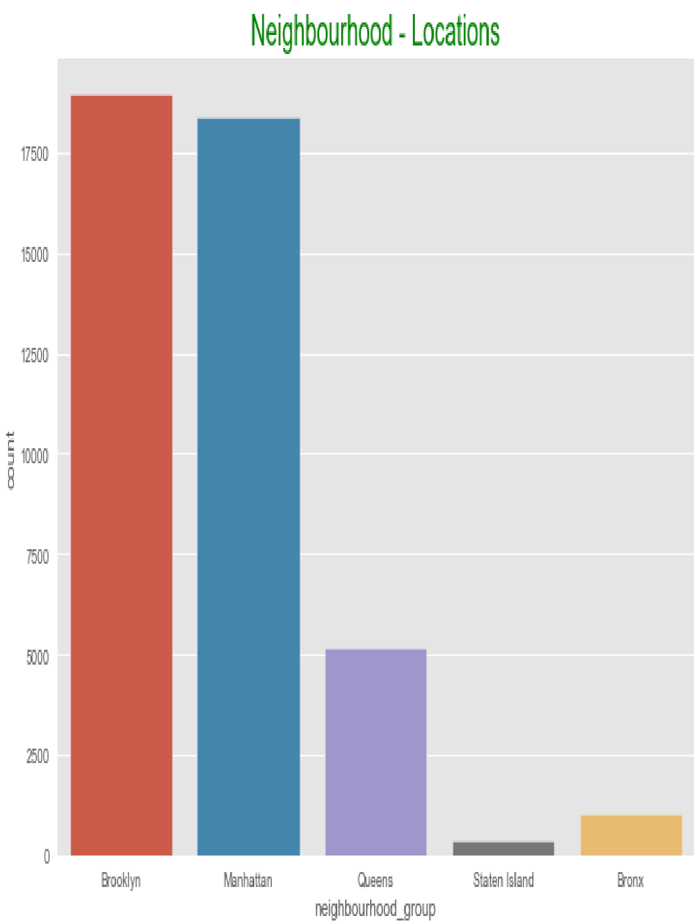
```
int_cols = airbnb.select_dtypes(include=['int64', 'float64']).columns
list(enumerate(int_cols))
```

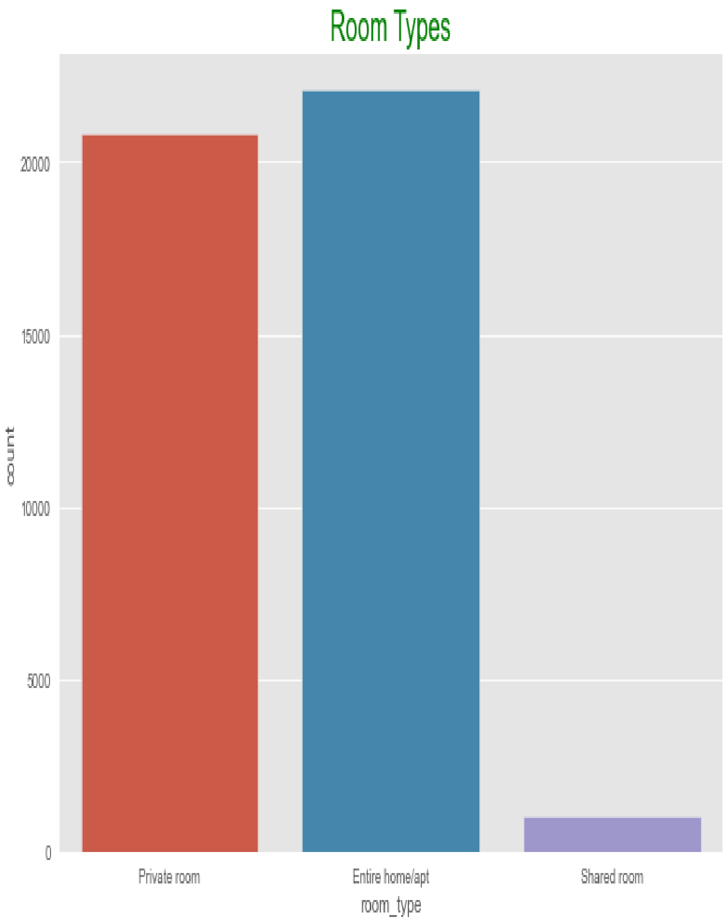
```
[(0, 'latitude'),
 (1, 'longitude'),
 (2, 'price'),
 (3, 'minimum_nights'),
 (4, 'number_of_reviews'),
 (5, 'reviews_per_month'),
 (6, 'calculated_host_listings_count'),
 (7, 'availability_365')]
```

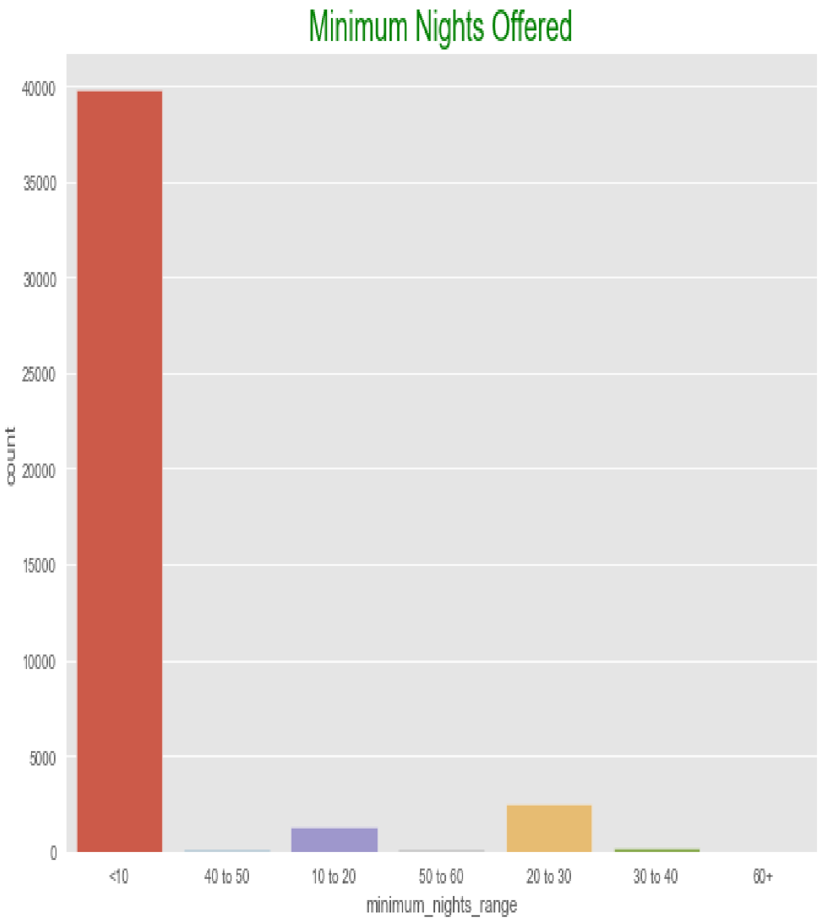
```
int_cols = airbnb.select_dtypes(include=['int64', 'float64']).columns
plt.figure(figsize=[20,18])

for n,col in enumerate(int_cols):
    plt.subplot(4,2,n+1)
    sns.distplot(airbnb[col])
```



Univariate Analysis- Categorical Column

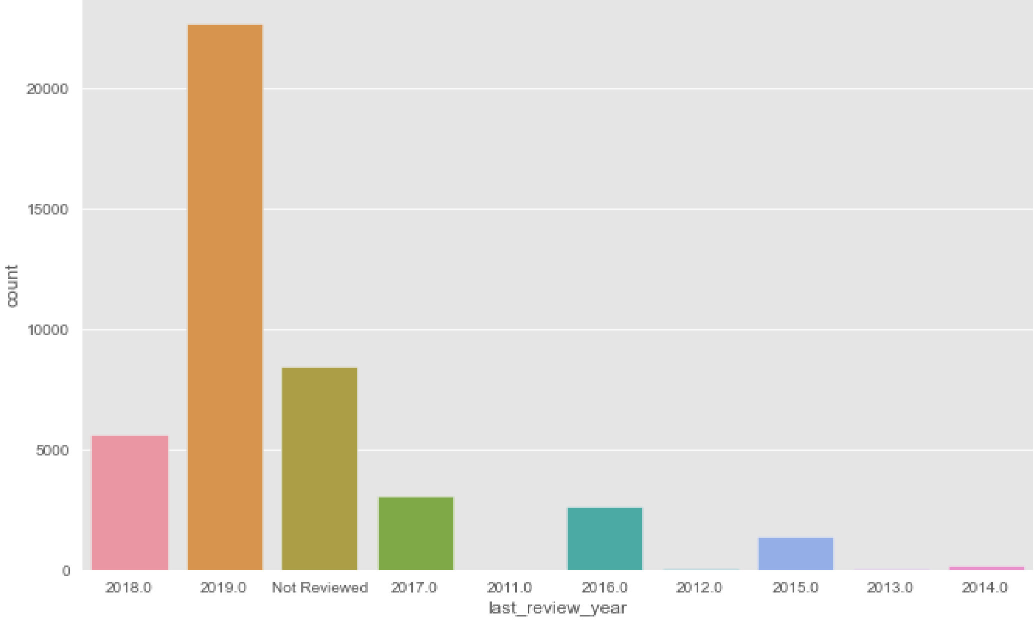




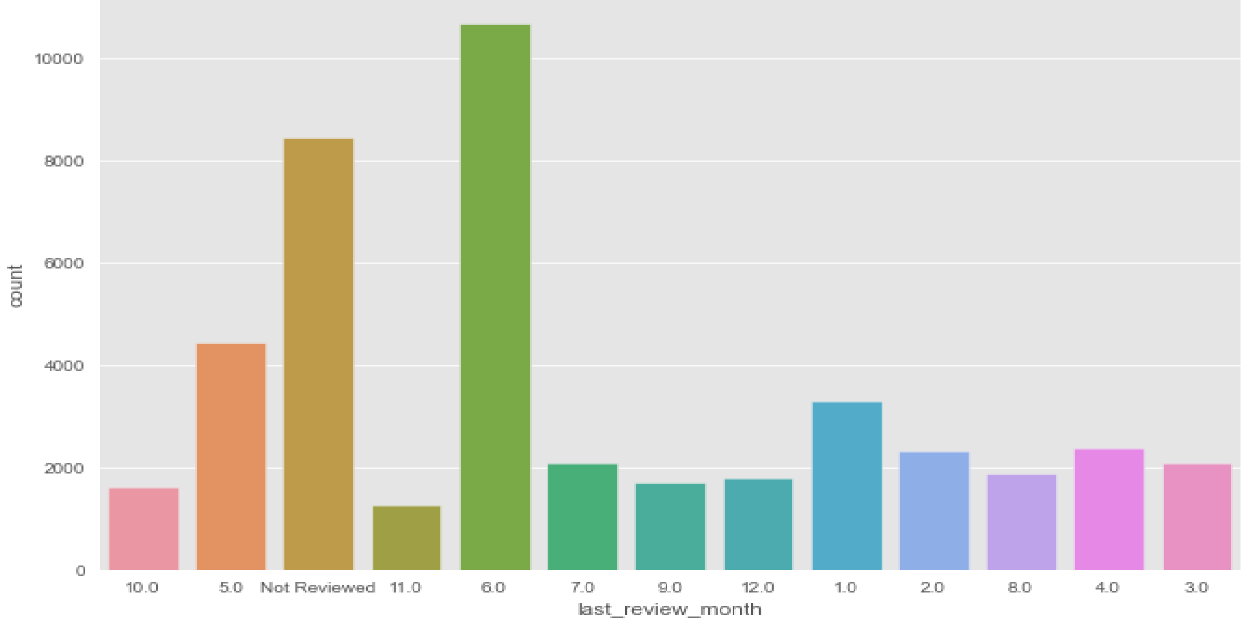


Univariate Analysis- Categorical Column

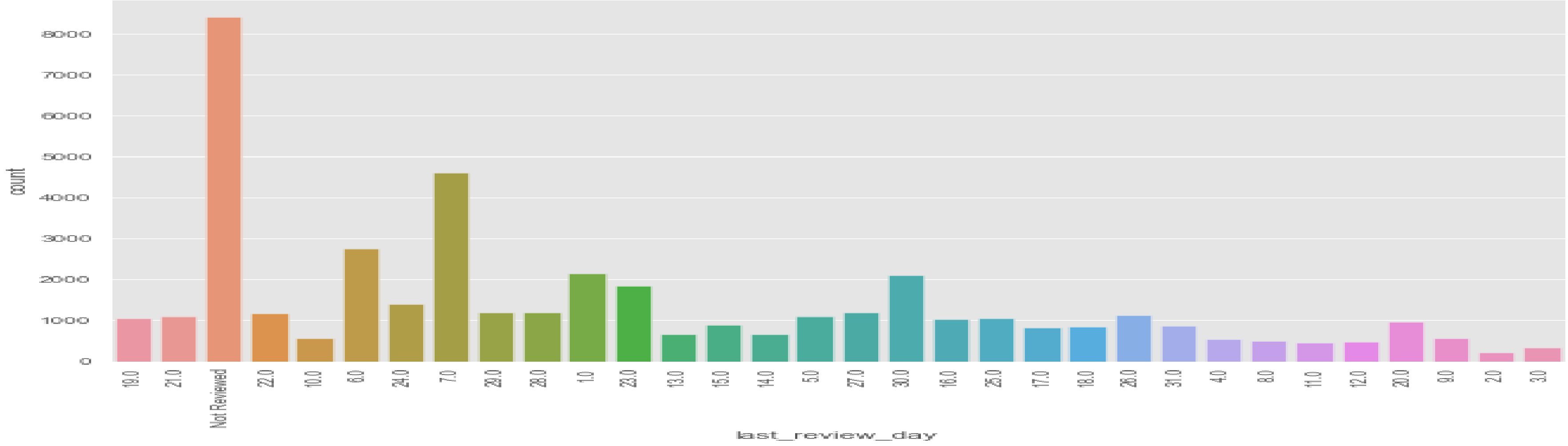
Last Review Years



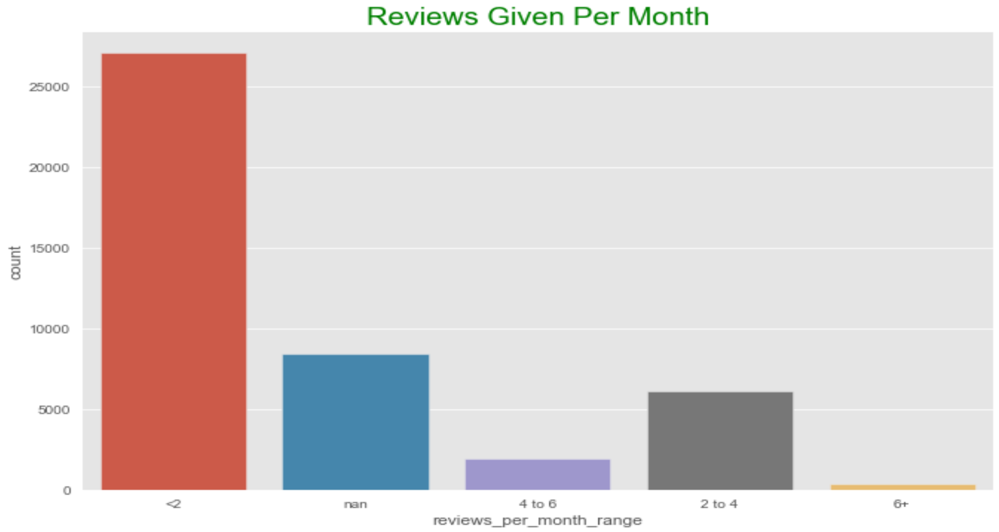
Last Review Months

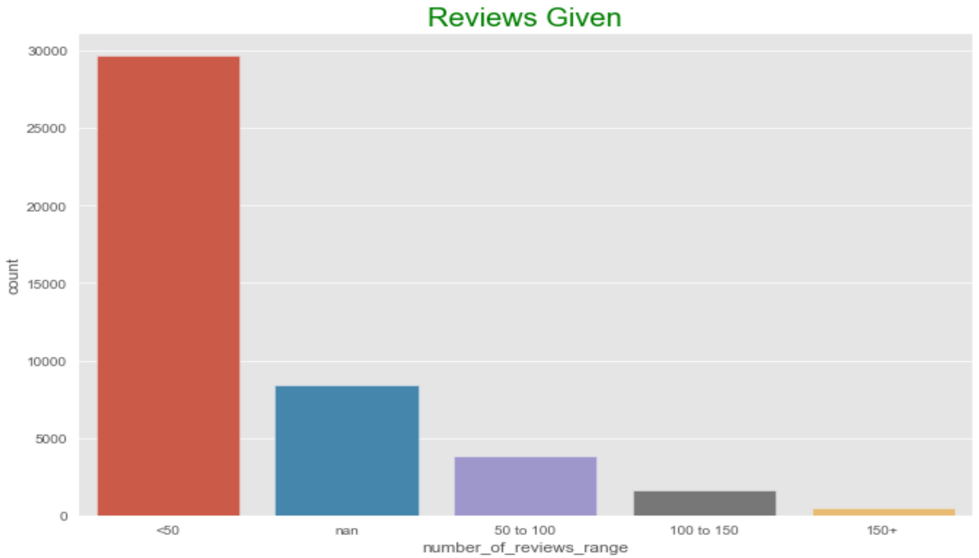


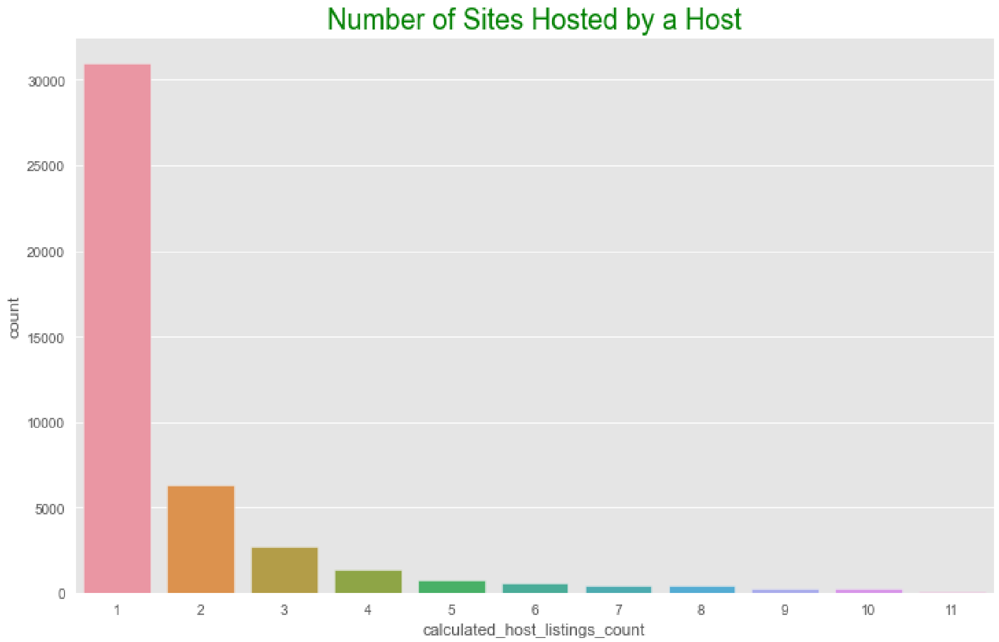
Last Review Days



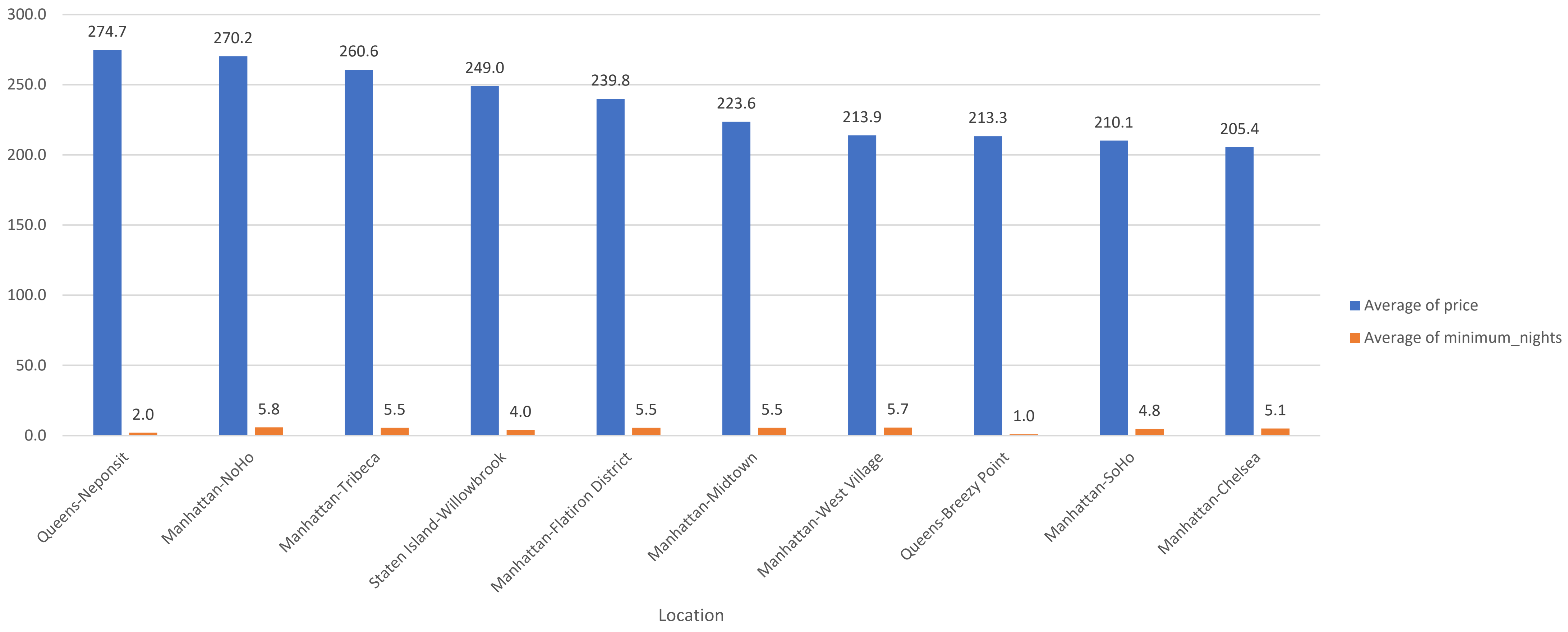
EDA- Univariate Analysis- Categorical Column



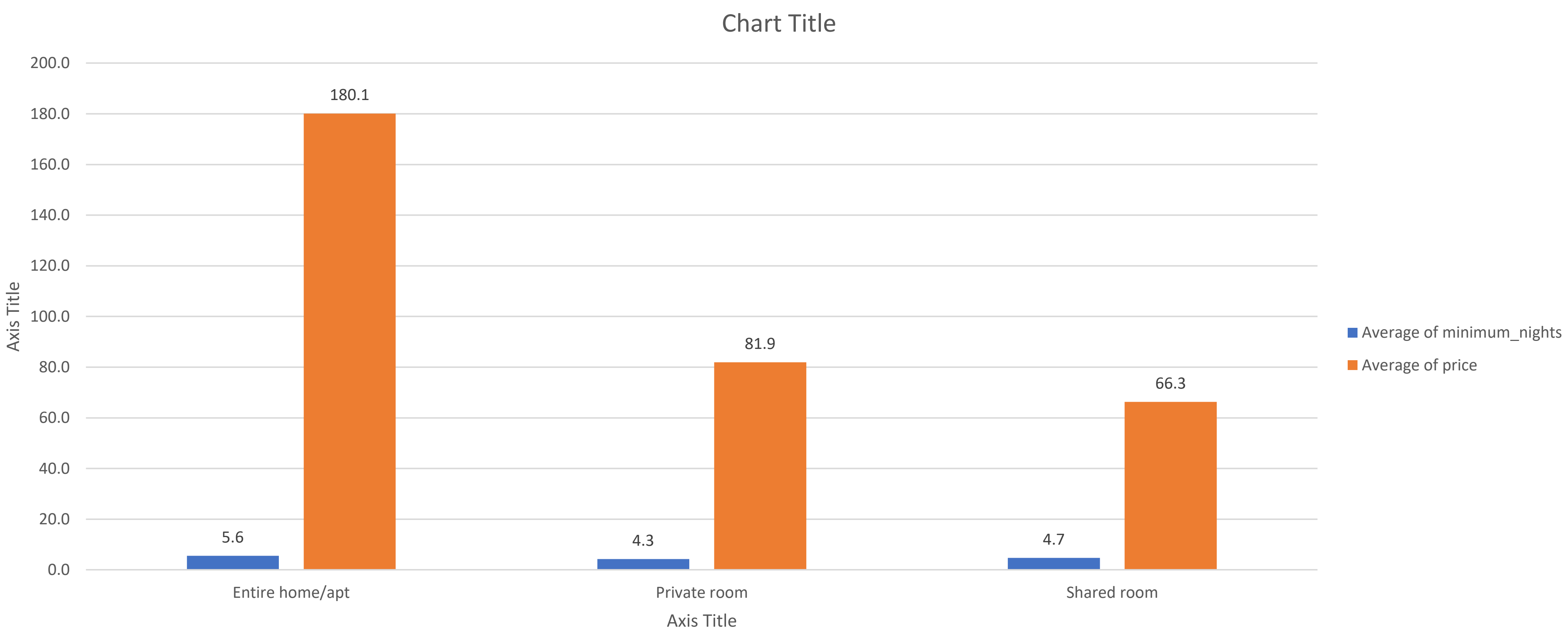




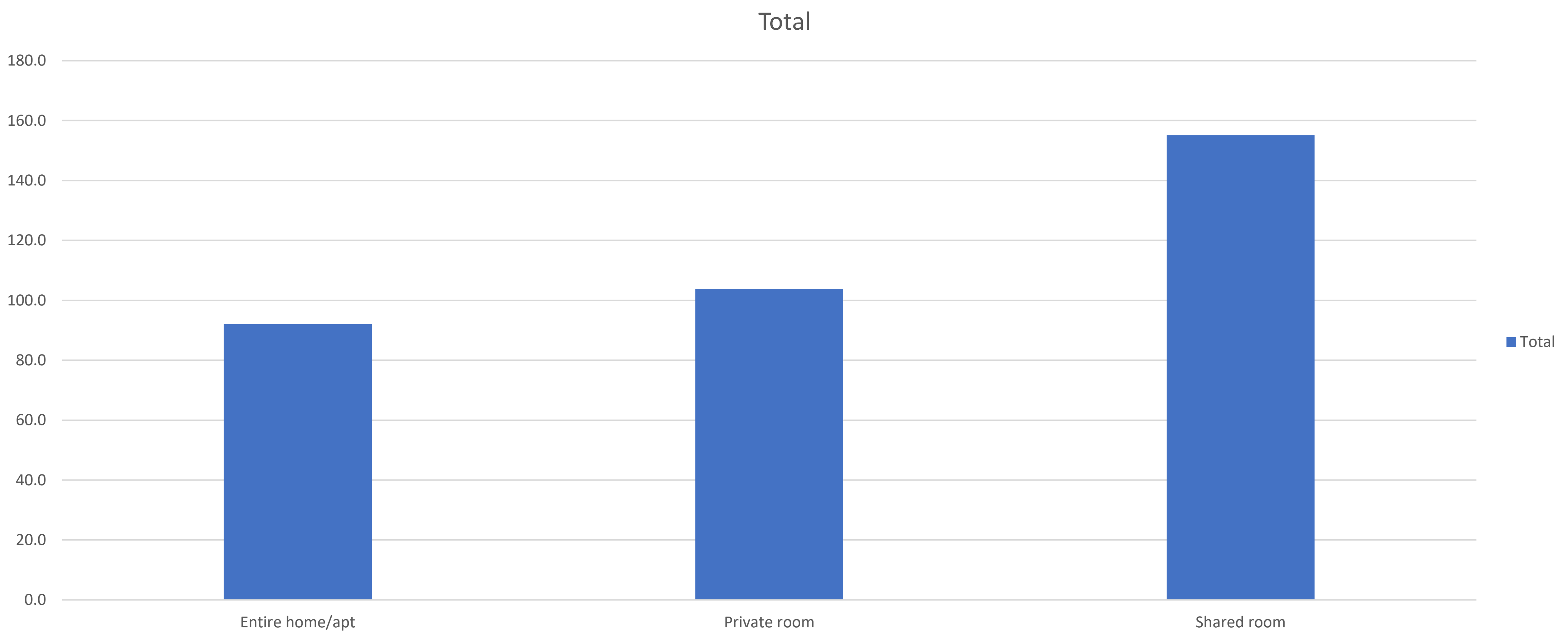
Location Wise Average Price and Average Minimum Nights



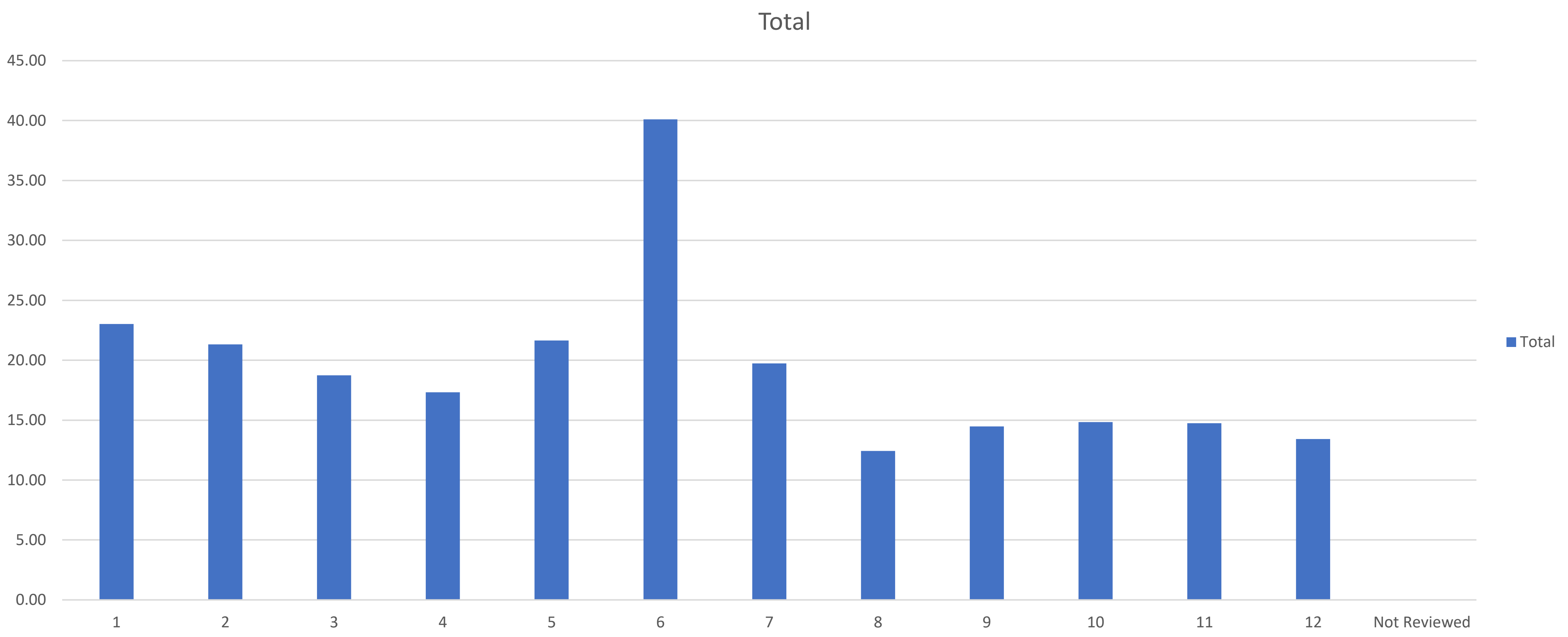
Room Type Wise Average Price and Average Minimum Nights



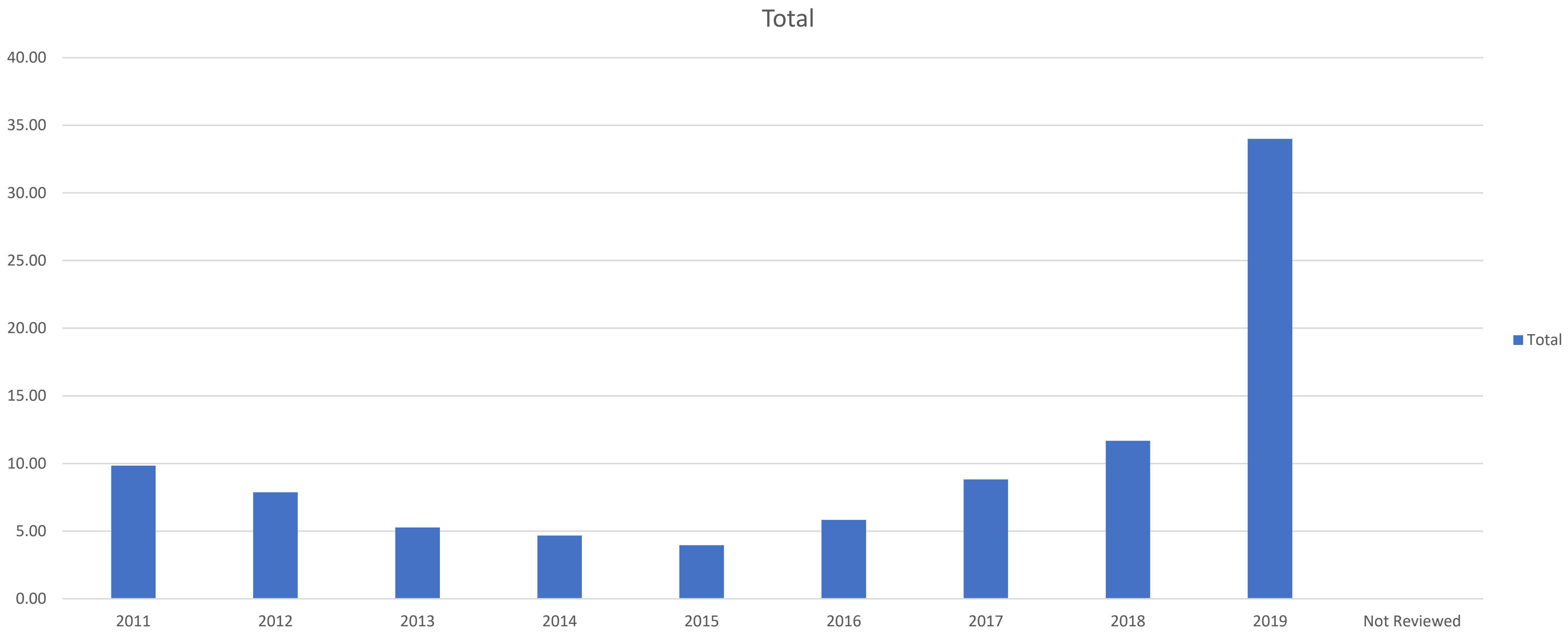
Room Type Availability



Last Month Review Vs Average Review



Last Year Review Vs Average Review



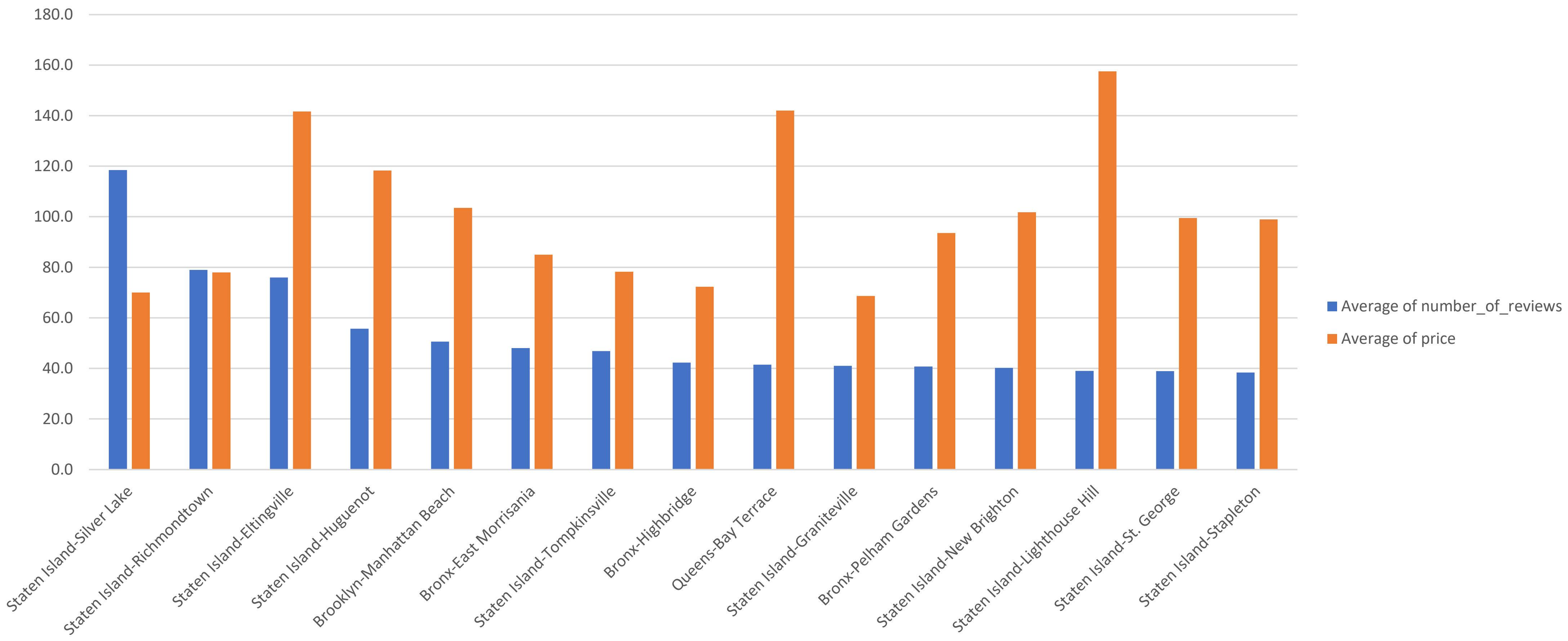
Year Wise trend of Average Review

The trend of average of Number Of Reviews for Last Review Year. Color shows average of Number Of Reviews. The marks are labeled by average of Number Of Reviews. The view is filtered on Last Review Year, which keeps non-Null values only.

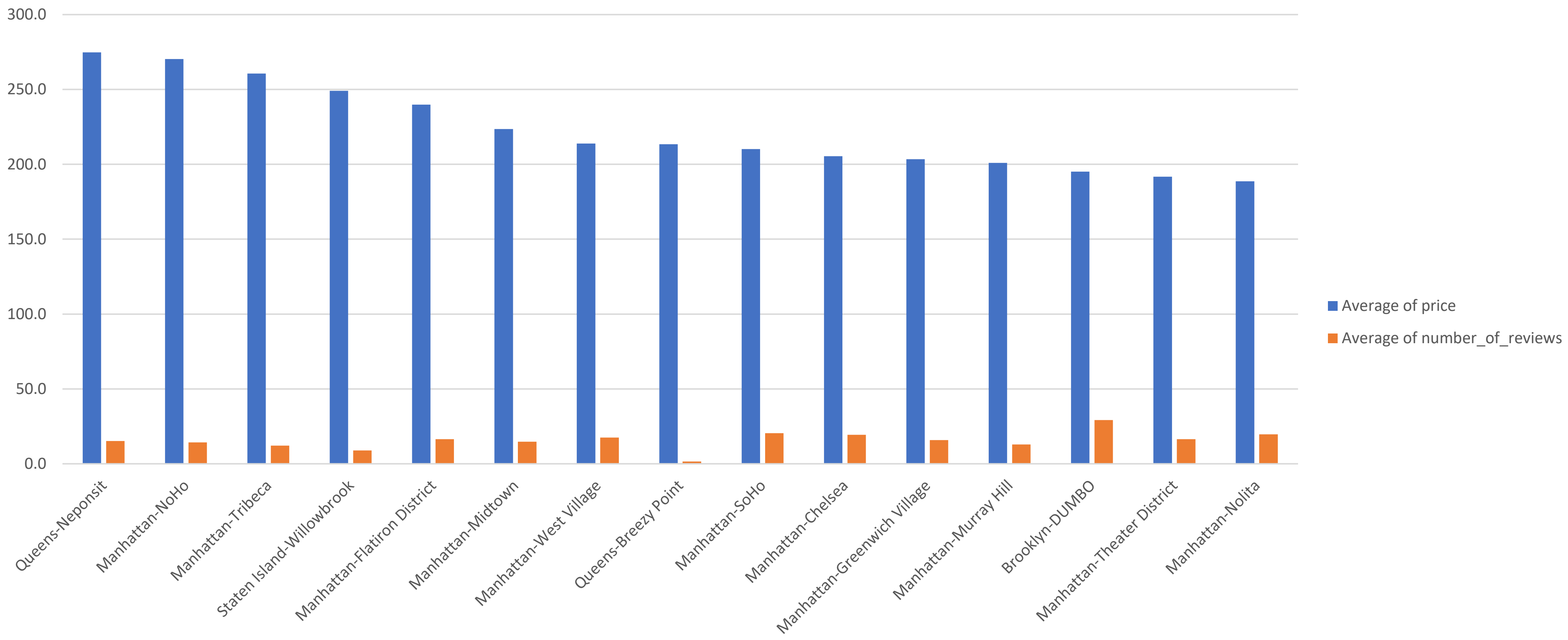
Month Wise trend of Average Review

The trend of average of Number Of Reviews for Last Review Month. Color shows average of Number Of Reviews. The marks are labeled by average of Number Of Reviews. The view is filtered on Last Review Month, which keeps non-Null values only.

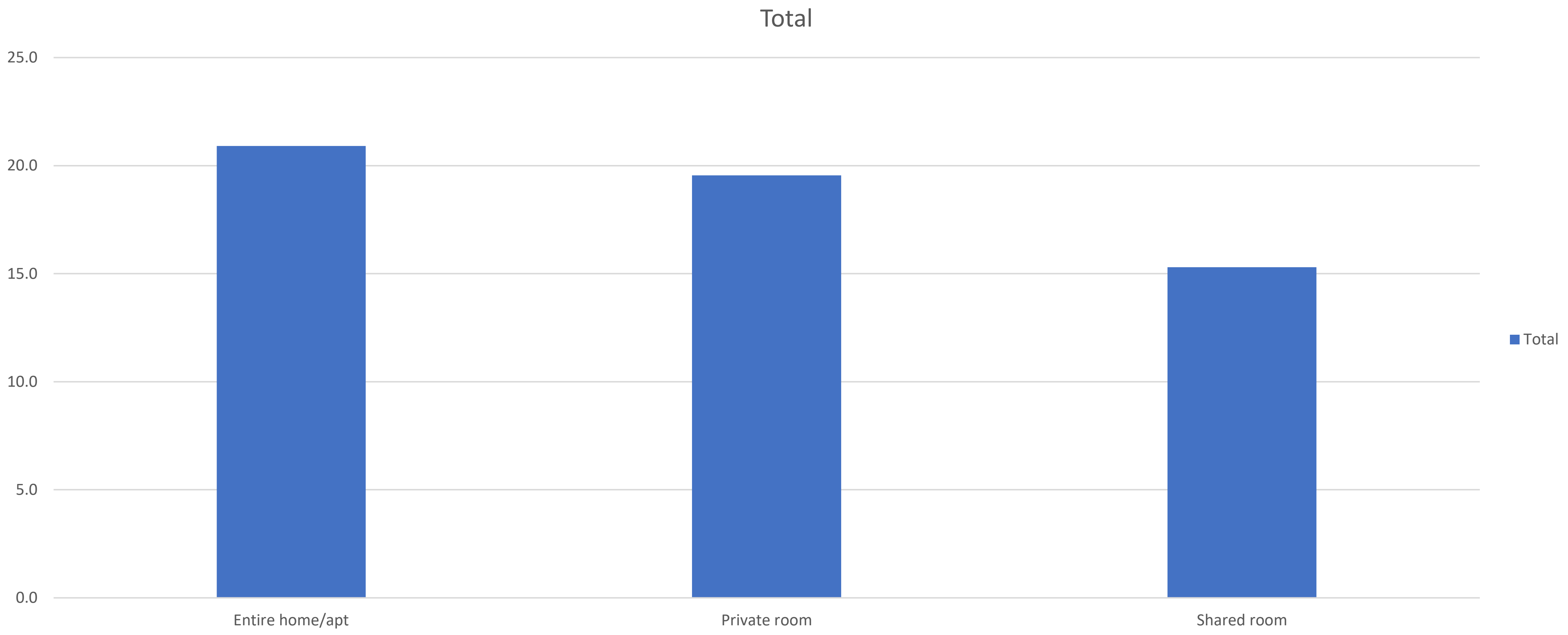
Location Vs Average Review



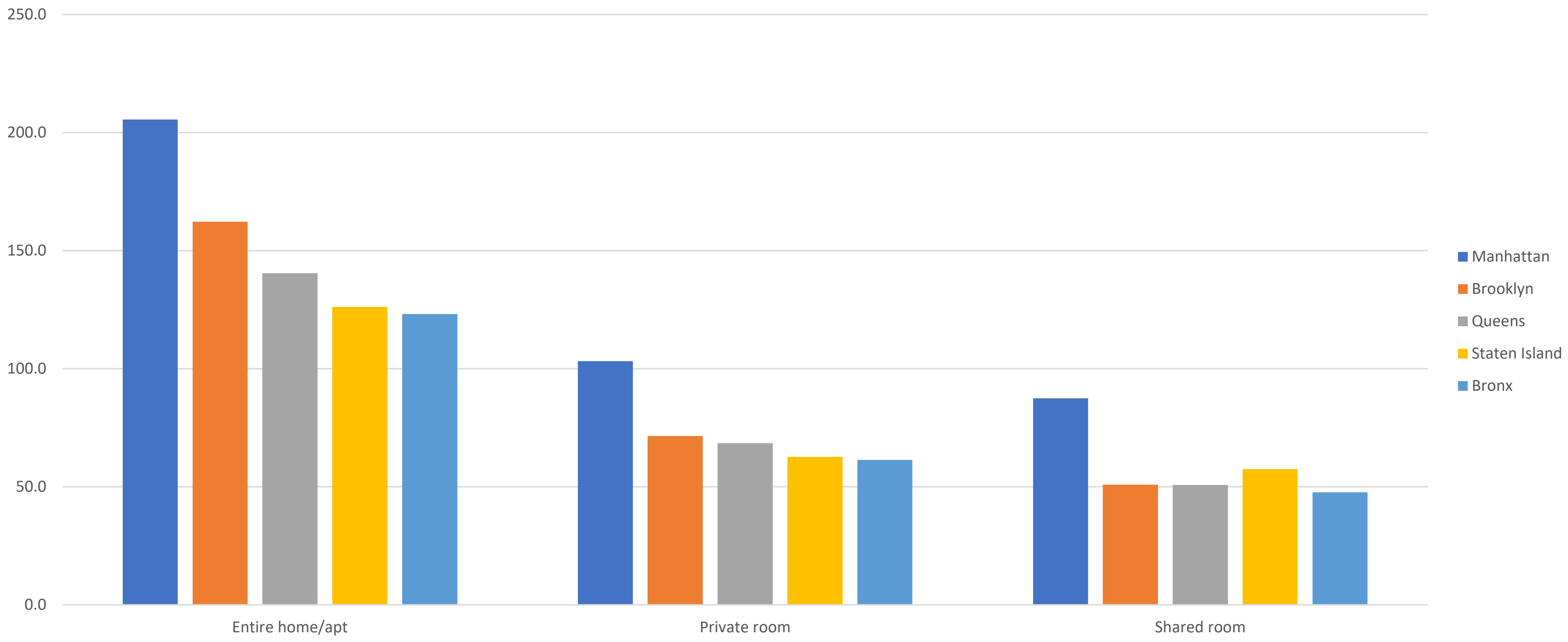
Location Vs Average Price



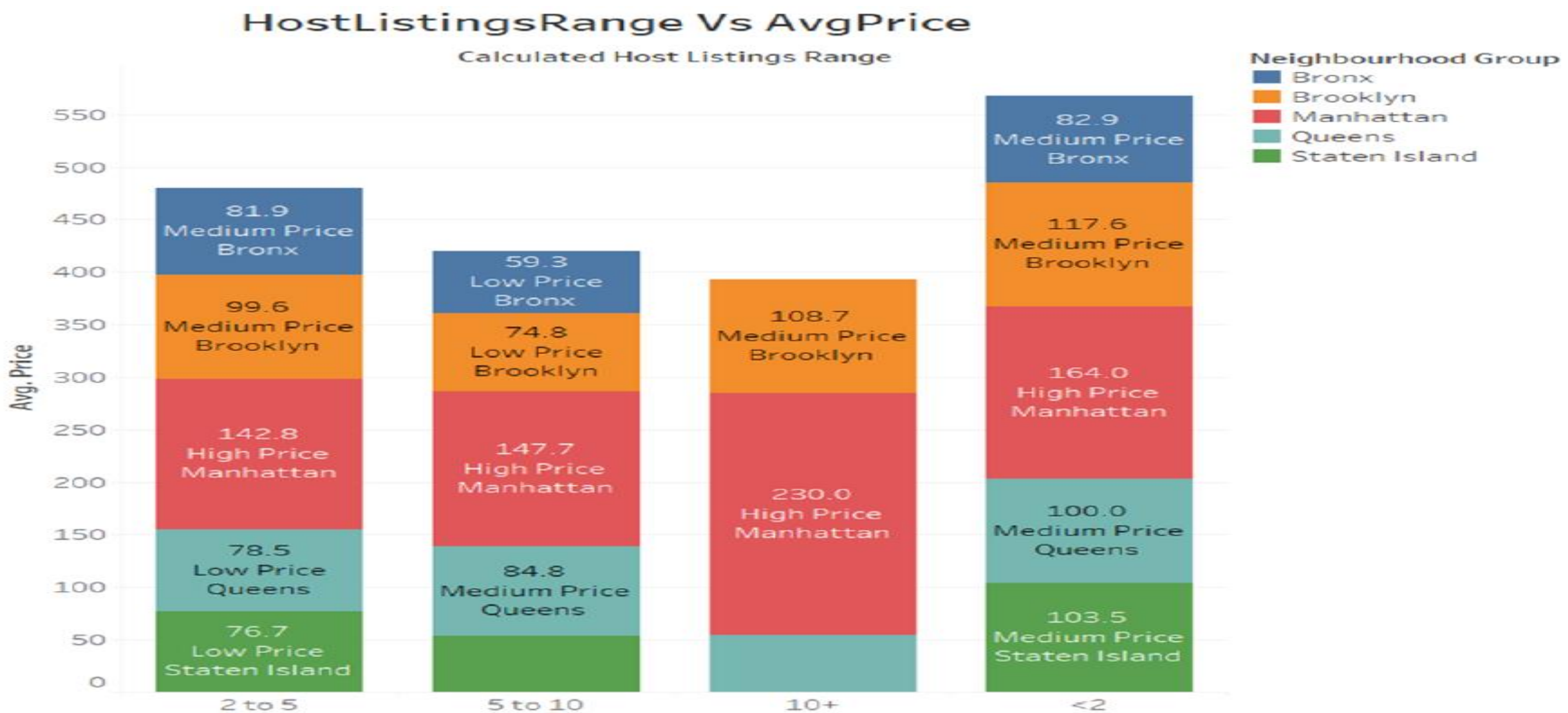
Room Type Vs Average Review



Room Type Vs Average Price

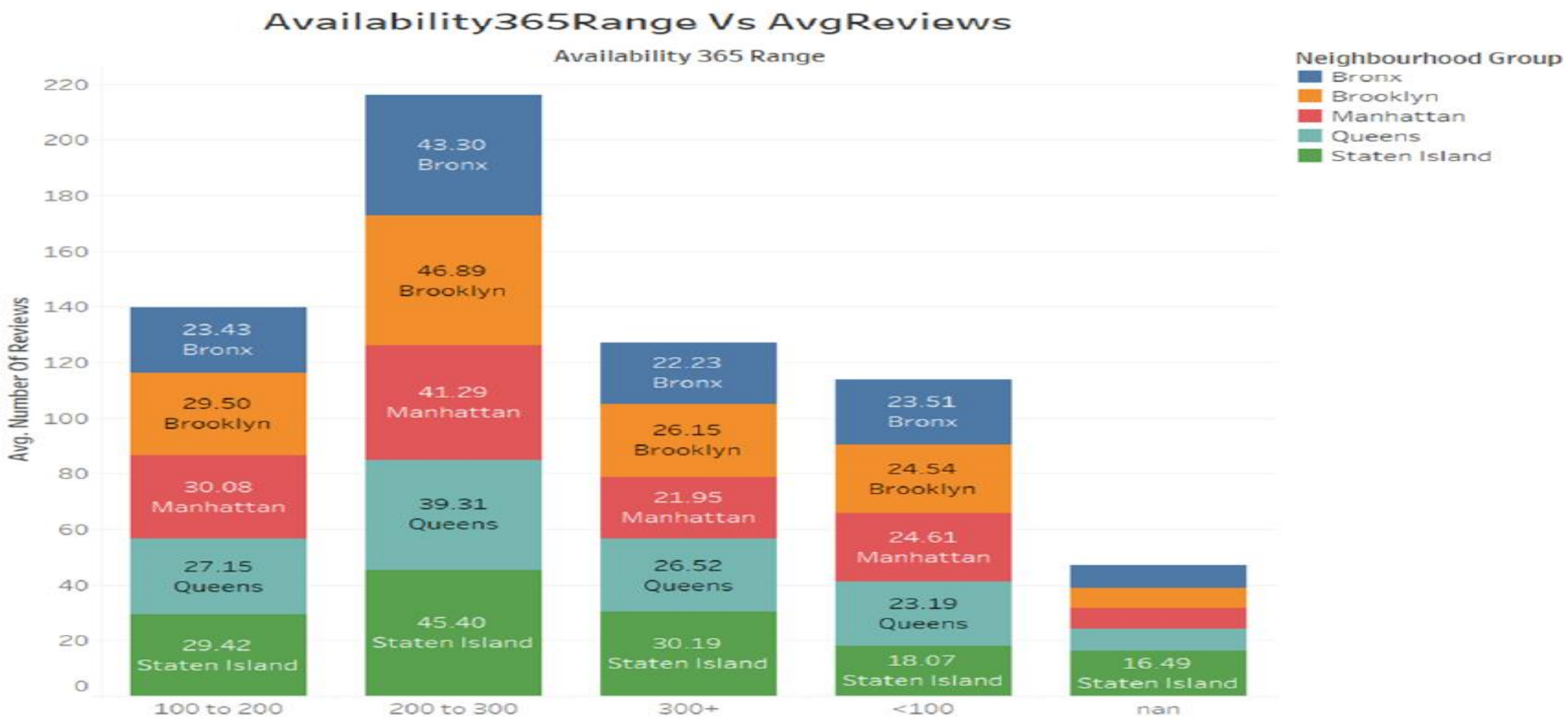


Number of Places hosted by a single host based on their Avg Price and Neighborhood



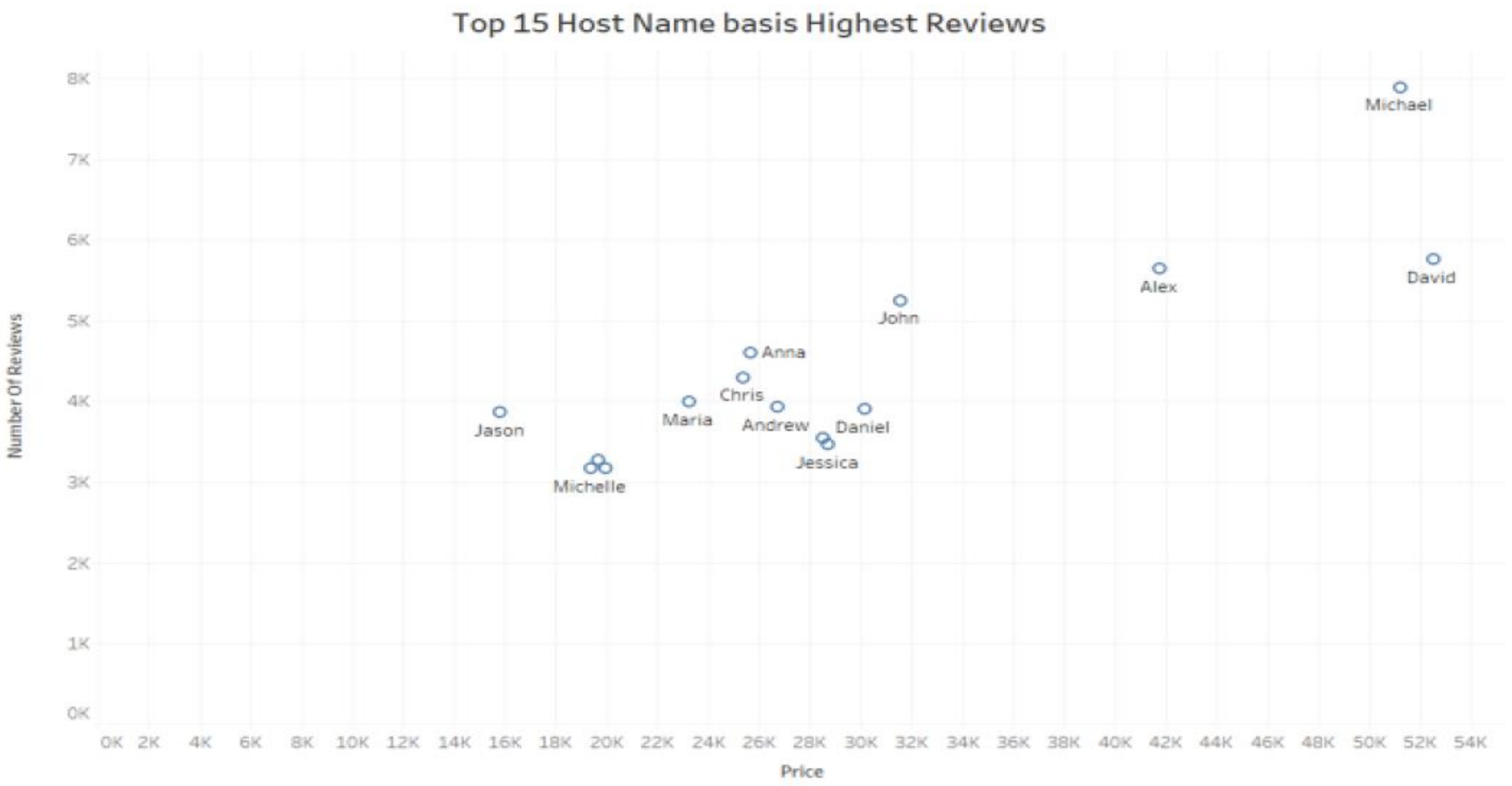
Average of Price for each Calculated Host Listings Range. Color shows details about Neighbourhood Group. The marks are labeled by average of Price, Price Range and Neighbourhood Group.

Average number of reviews given to places based on their number of days availability in a year



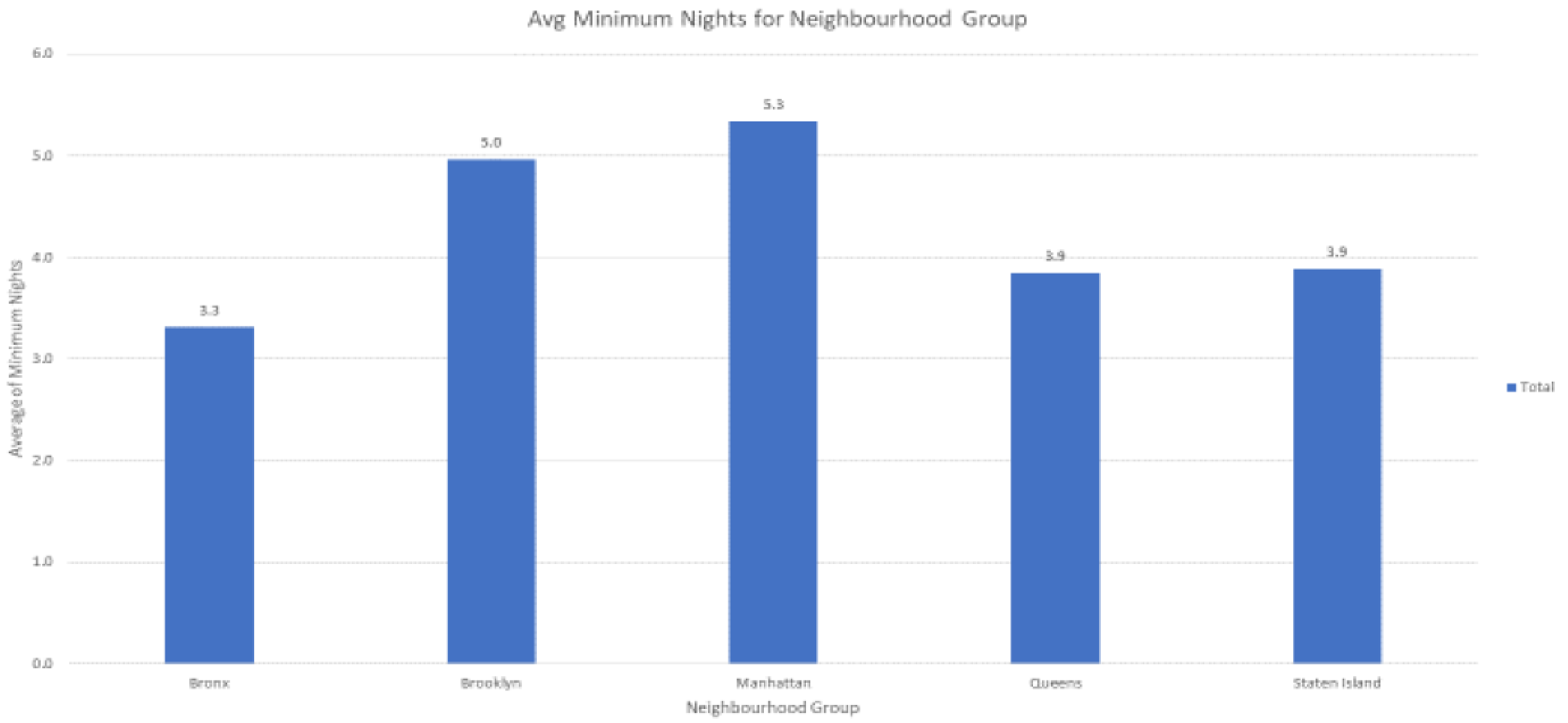
Average of Number Of Reviews for each Availability 365 Range. Color shows details about Neighbourhood Group. The marks are labeled by average of Number Of Reviews and Neighbourhood Group. The view is filtered on Neighbourhood Group, which keeps Bronx, Brooklyn, Manhattan, Queens and Staten Island.

Name of the Host who have received highest number of reviews



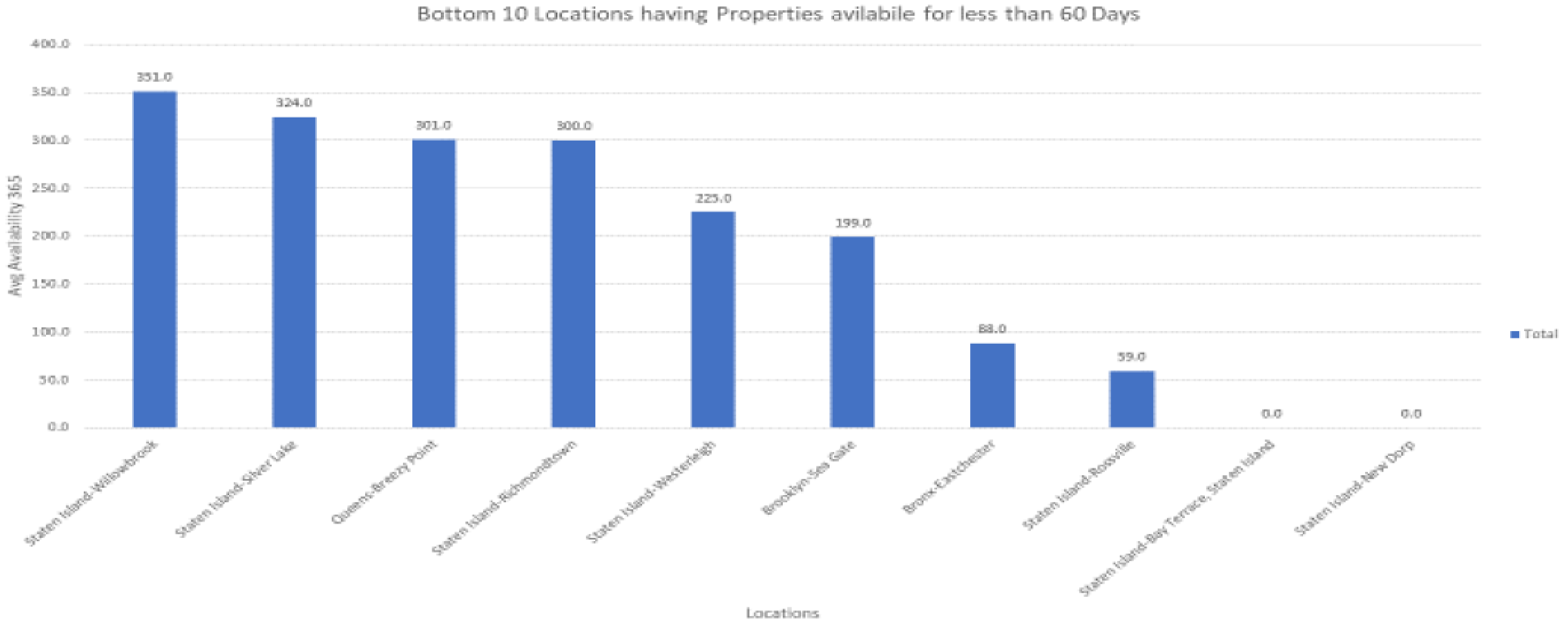
Sum of Price vs. sum of Number Of Reviews. The marks are labeled by Host Name. Details are shown for Host Name. The view is filtered on Host Name, which keeps 15 of 11,024 members.

Top Neighborhoods providing higher number of Minimum Night stay



Locations contributing more on the Platform

Locations contributing less on the Platform



Insights- I

- Brooklyn, Manhattan and Queens are dominating when it comes to listed hosting
- Majorly Private rooms or Entire apartment are provided by host
- Majority of the sites provide less than 10 nights stay at a time
- Majority of the sites have received less than 50 reviews till date
- Most the sites have received less than 2 reviews per month which indicates bad customer experience offered by majority sites
- Majority of the Host have 1 site hosted by them on the platform
- Most of the sites hosted provide 0 days availability which needs to be checked and then most of the site have less than 100 days availability compared to all 365 days
- Slowly and gradually reviews started to build up and was mostly in 2018 and 2019
- 6th month of the year i.e June seems to receive most of the last reviews in all years followed by 5th month
- Most of the times last reviews were not provided when we see Day wise. Next, majority of times it was provided on the 6th and 7th day of the month followed by 1st and last day of the month
- On an average Entire home/apt types are preferred more by the customers followed by Private rooms and then the Shared Rooms. Mostly because they are also available for a higher number of minimum nights stay window booking as compared to Private and Shared rooms

Insights- II

- Staten Island - Silver Lake, Staten Island - Richmondtown, Staten Island - Eltingville, Staten Island - Huguenot and Brooklyn - Manhattan Beach are the Top 5 locations with Low Price range that have received the highest number of reviews on average being the lowest in Price range. On the contrary, Queens - Neposit, Manhattan - NoHo, Manhattan - Tribeca, Staten Island - Willowbrook and Manhattan - Flatiron District being highest in Price range have received low number of reviews
- Michael, David, Alex, John and Anna are the Top 5 hosts that seem to have received the highest number of reviews for their listed sites and have also sites listed with High price range.
- Manhattan is the only Neighborhood in the Borough that lies in offering the Highest Price range properties on the platform followed by others with a Medium Price range on average. Prices offered above 120$ on average are High Priced, between 80$ to 120$, Medium Price range and less than 80$ to be considered Low Price range property
- Brooklyn has received the highest number of reviews based on the availability to stay open throughout the year. This is followed by Manhattan . On the other hand there are some sites in Staten Island which are not open for a single day at all and hence could be the reason they have received very low reviews from the end consumer and thus they contribute very less on the platform
- “Brooklyn-Williamsburg”, “Brooklyn-Bedford-Stuyvesant”, “Manhattan-Harlem”, “Brooklyn-Bushwick” and “Manhattan-Upper West Side” are some places providing the highest number of minimum nights window to book making Manhattan and Brooklyn the top neighborhoods in offering maximum minimum nights stay
- Majority of the customers prefer a price range of 120$ to 130$ on average for a stay. As most of them have provided a good number of reviews between this price range
- We can confirm that the greatest parameter for any customer to prefer a property and provide a review is having a maximum or minimum night stay window booking and their probability of being open for more days in a year to some extent