A. Methodology Approach

1. Research Problem

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

2. Business Understanding

Airbnb is an American company based in San Francisco, California. It operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. The platform is accessible via website and mobile app.

Afterall, being an online marketplace for hosting personal home stays and private apartments in majority, the company had two types of customers. One who hosts their place and the another who books the place for a particular time that is the end consumer utilizing the hosted place. Airbnb earns commission from both ends and hence have to make sure both of its customers are able to generate value from their business. They also have to make the hosted place offered on their platform provide the best services at reasonable prices and look out for the best technology to ease out the booking process for the end consumer without hassle.

3. Type of Data required to Analyze

Decline in the revenue could be for two major reasons, either the sites hosted on the platform are not able to provide better user experience or there could be a competitor in the market capturing the market share. Keeping the above in mind, we first try to work on the first reason as that is something internal to the company and can have the data in hand to identify the reasons behind the plummeting of the revenue.

Hence, we use the information of the hosted places on the platform to see where and what can be done to improve the end consumer experience. The data would majorly include the location and region of the hosted places, in our case we are targeting Borough (New York City) — the Bronx, Brooklyn, Manhattan, Queens and Staten Island, followed by their hosts details, prices of the hosted sites and reviews received by the end consumer.

4. How was the Data was acquired? (Assumption)

- The provided data is captured from the CRM tool used by Airbnb to manage their customers that are hosting sites on their platform.
- The reviews provided in the data frame are assumed to be positive as it is not mentioned whether they are negative or positive reviews.

5. Whom are we presenting?

- **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.
- **Head of Acquisitions and Operations, NYC:** This head looks after all the property and hosts 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.

6. Recommendations

- One to one interaction with some property owners in Staten Island, Queens and Bronx to identify their challenges for being fully functional for maximum number of days in a year and allow a booking of more than 10 days of minimum night stay.
- Create some sort of interaction between the Top 5 hosts to share their experience with the rest of the community for better improvement and value generating ideas.
- Provide discounted commission rates to property owners on keeping the minimum night stay booking window for more than 10 days and property functional for maximum number of days in a year.

B. Method of Analysis along with code:

1. *Data Understanding and Preparation*

Before we start the basic understanding of the data in hand, we imported relevant libraries available in Python. Below are the libraries that we imported,

```
# Importing Libraries

import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('seaborn-dark-palette')
from scipy import stats
import datetime as dt
import plotly
import plotly.express as px

# Ignore warnings

import warnings
warnings.filterwarnings("ignore")
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

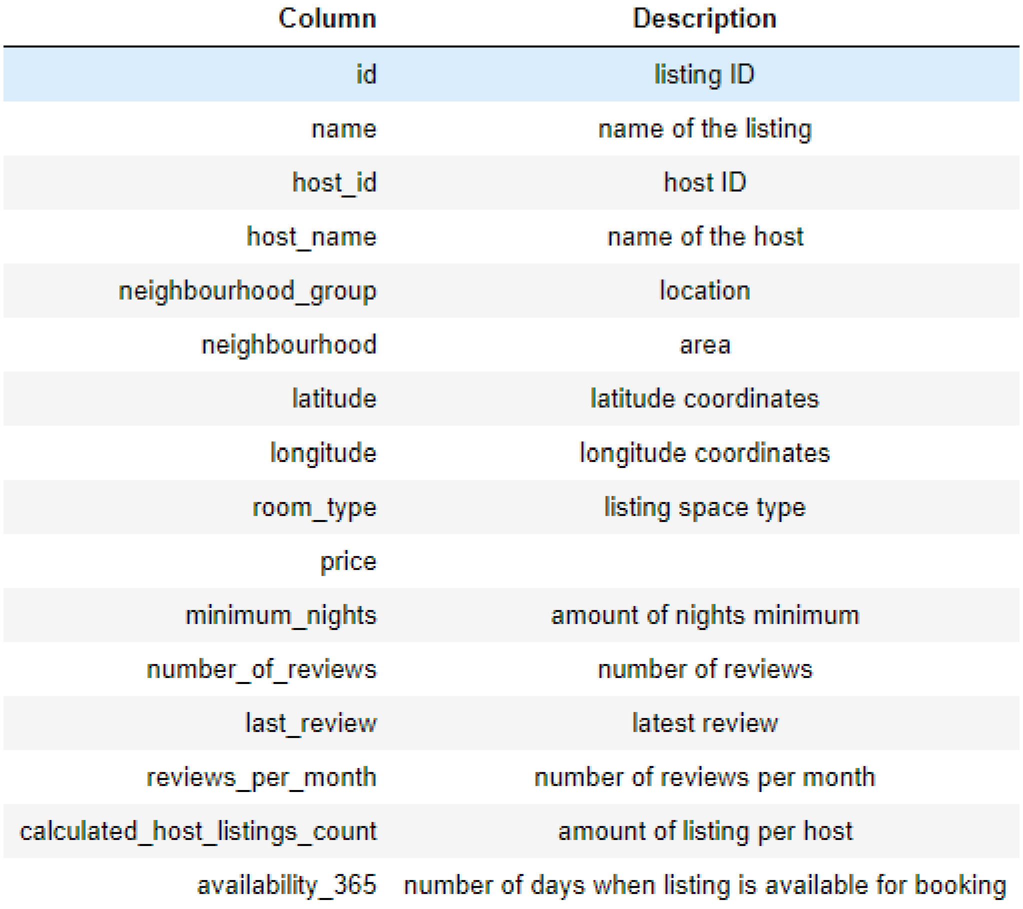
We started with **Understanding the Data** in hand provided by running basic functions to load and interpret the variables, data types of the variables, dimensions and size of the dataframe. Below is the code used for the same.

```
# load the dataset
airbnb = pd.read_csv("AB_NYC_2019.csv")
airbnb.head()

# Dimensions
Airbnb.shape

# Data-Types
airbnb.info
```

2. *Variables in the dataframe:*



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

The above understandings lead us to perform basic **Numeric and Categorical analysis** in depth by using the following function along with some basic wear and tear,

```
# Numeric Analysis
airbnb.describe()

# Analyzing categorical values
airbnb.select_dtypes(include=['object']).describe()
```

3. *Handling Missing Values and Outliers:*

- Then we moved to handle missing values and outliers in the dataframe. Starting with the missing values, we identified two columns having equal percentage of missing values which were last_review and reviews_per_month of around 20.56%. And also, other two columns having quiet minimal missing values which were host_name of 0.4% and name of the place of 0.3%.
- Then we analyzed the values missing in last_review and reviews_per_month carrying NaN values on purpose, meaning they are not missing at random as these hosted sites/places have not received 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.

- Then we identified that we have 16 places and 21 host names that are missing and then we cross check them with the help of their id's to verify whether they are missing at random or by chance.
- After analyzing, it seems like these values (host name and their place name) are missing by chance hence we need to collect this information from the Host acquisition and operations team. But for now we left these rows blank.
- Finally, we just imputed the missing values of reviews_per_month with a 0.

Below is the code we used to identify missing values. We also imported missingno library to do the same.

```
# Checking missing values columns
import missingno as msno
msno.bar(airbnb)
```

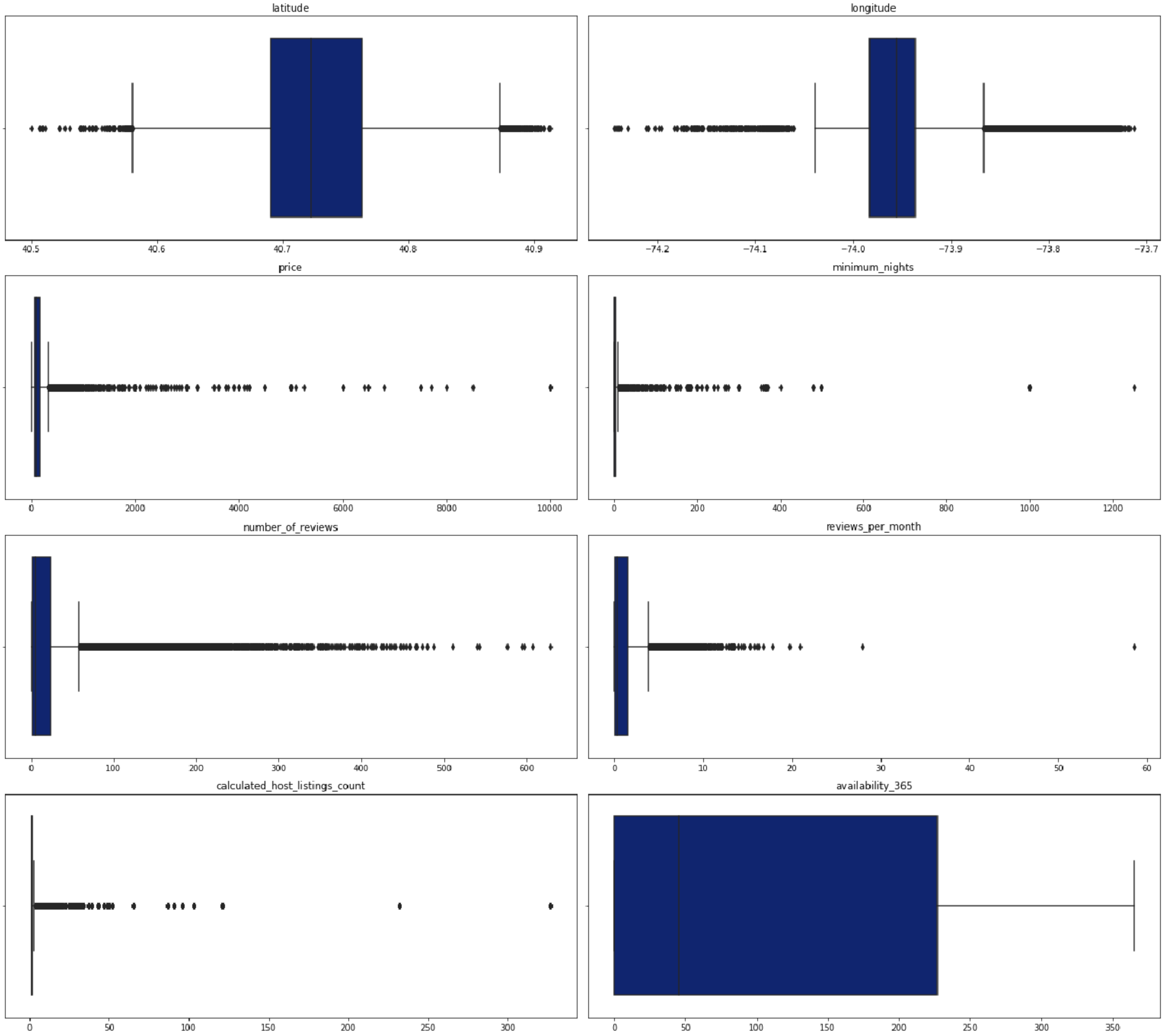
```
# Checking missing values percentages
def null_values(airbnb):
   return round((airbnb.isnull().sum()/len(airbnb)*100).sort_values(ascending = False),2)
null_values(airbnb)
```

Post analyzing and treating the missing values accordingly we treated the spread in the data frame i.e. outliers. Below is the code we used to identify the spread of the outliers.

```
# Extracting Numeric columns:
int_cols = airbnb.select_dtypes(include=['int64', 'float64']).columns
```

```
# Tagging them:
list(enumerate(int_cols))
```

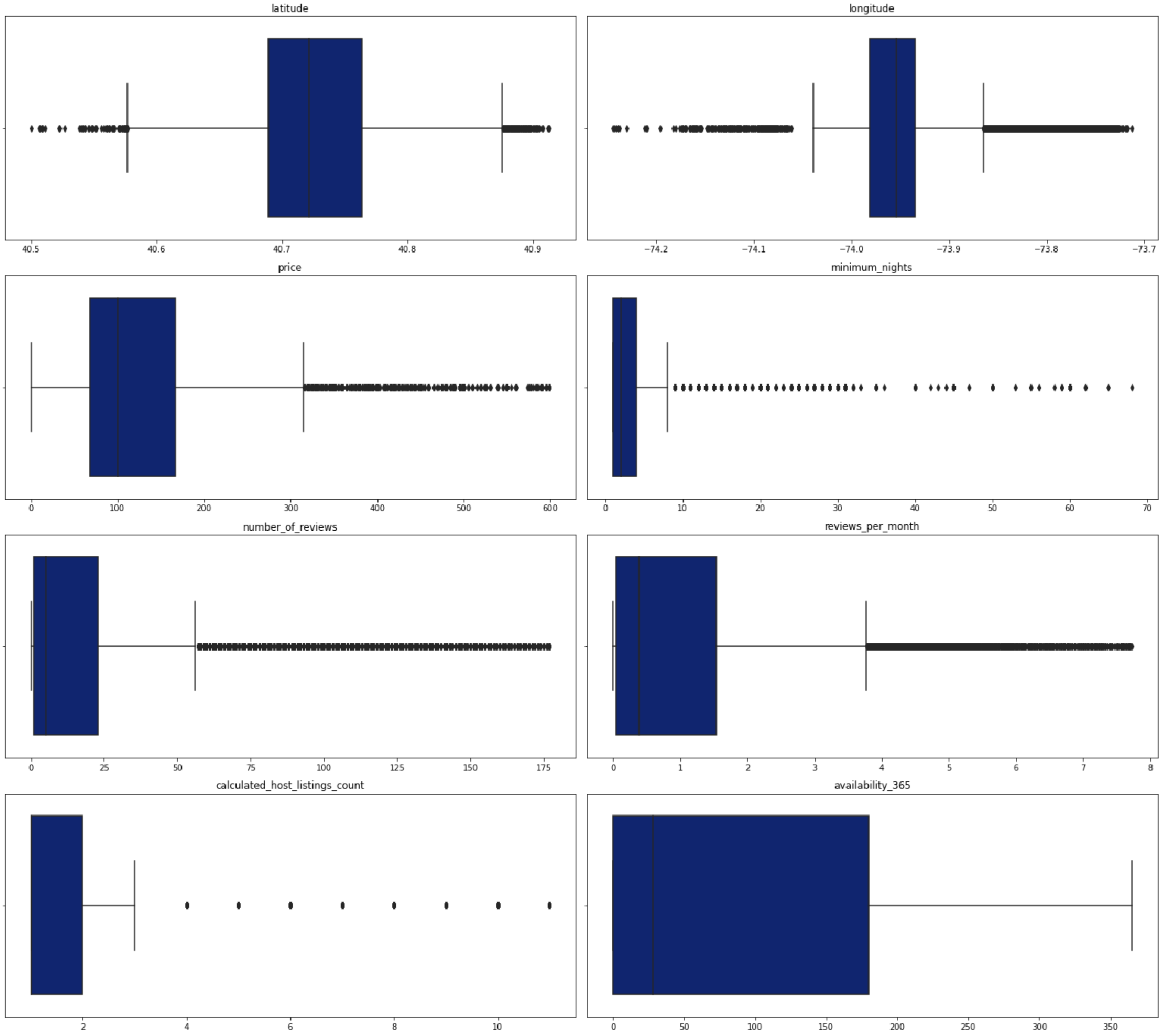
```
# Plotting the spread of outliers:
plt.figure(figsize=([20,22]))
for n,col in enumerate(int_cols):
   plt.subplot(5,2,n+1)
   sns.boxplot(airbnb[col], orient = "h")
   plt.xlabel("")
   plt.ylabel("")
   plt.title(col)
   plt.tight_layout()
```



The method we used to treat them was by capping them by 10% Below is the code for the same.

```
# Capping (statistical) outliers
# 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 minimum_nights:
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)]
```



Looks like we were able to manage the outliers, enough to analyze the information in EDA.

4. *Feature Selection / Engineering*

The most important step of our Data Preprocessing was to convert some of the numeric features into categorical variables by creating bins of them. Yet post conversion, we kept the numeric one’s handy tool for analyzing. Below is the table of all the variables that were engineered for our further analyses.

Dimensions	Measures
name	price
host_name	minimum_nights
Location (neighbourhood_group + neighborhood)	number_of_reviews
room_type	reviews_per_month
minimum_nights_range (<10, 10-20, 20-30, 30-40, 40-50, 50-60, 60+)	calculated_host_listings_count
number_of_reviews_range (<50, 50-100, 100-150, 150+, Nan)	availability_365
reviews_per_month_range (<2, 2-4, 4-6, 6+, Nan)	
calculated_host_listings_range (<2, 2-5, 5-10, 10+)	
availability_365_range (<100, 100-200, 200-300, 300+, Nan)	
last_review_year (2011 to 2019 & Not Received)	
last_review_month (1 to 12 & Not Received)	
last_review_day (1 to 31 & Not Received)	

5. *Analyzing Methods:*

a. Univariate Analysis:

We started our general Univariate Analysis of Numeric and Categorical columns. For numeric columns we used a Distribution plot from seaborn and for categorical columns we used a Countplot from the same library seaborn. Below are the codes for the same.

```
# Extracting and Tagging the Numeric Columns:
int_cols = airbnb.select_dtypes(include=['int64', 'float64']).columns
list(enumerate(int_cols))

# Plotting the Numeric Variables Distribution:
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])
```

```
# Checking the count of Neighborhood Groups
plt.figure(figsize=[12,7])
sns.countplot(airbnb.neighbourhood_group)
plt.title('Neighborhood - Locations', fontdict={'fontsize': 20, 'fontweight': 5, 'color': 'Green'})
plt.show()
```

Similarly we used the above countplot code for the rest of the categorical plots created.

b. Bi-Multivariate Analysis:

Here we first plotted a pairplot of all the numeric columns using seaborn library in Python itself. Below is the code for the same.

```
# Plotting the pairplot
sns.pairplot(airbnb)
plt.show()
```

Parallel to we created all other Bivariate and Multivariate plots using **Excel and Tableau**.

6. Matrix used for Analysis

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. Below is the Matrix.

Matrix	Numerical	Categorical
Categorical	HostedPlace (Name) Vs Average Price HostedPlace (Name) Vs Average Reviews Room Type Availability basis Price Difference HostListingsRange Vs Average Price basis Neighbourhood Group Availability 365 Range Vs AvgReviews basis Neighbourhood Group Last Review Year Vs Number of Reviews Last Review Month Vs Number of Reviews Last Review Day Vs Number of Reviews Average Minimum Nights & Price basis Location Average Minimum Nights & Price basis Room Type Avg Reviews basis Minimum Nights Stay Hosting Sites Range basis their avg. availability for 365 days	Map Showing Neighborhood Groups Basis Price Range
Numerical	Top 15 Host Name basis Highest Reviews Top Locations basis AvgPrice Vs AvgReviews Top 20 Hosted Place Name basis Price Range	Average Price Vs Review Range Location Vs Average Price Location Vs Average Reviews RoomType Vs Average Price basis Neighbourhood Group Room Type Vs Average Reviews Avg Minimum Nights for Neighborhood Group Avg Minimum Nights for Location Top Properties Available for more than 365 Days Top 10 Locations having Properties Available for 365 days Bottom 10 Locations having Properties available for less than 60 Days

7. *Evaluation of Methods*

The above matrix 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.

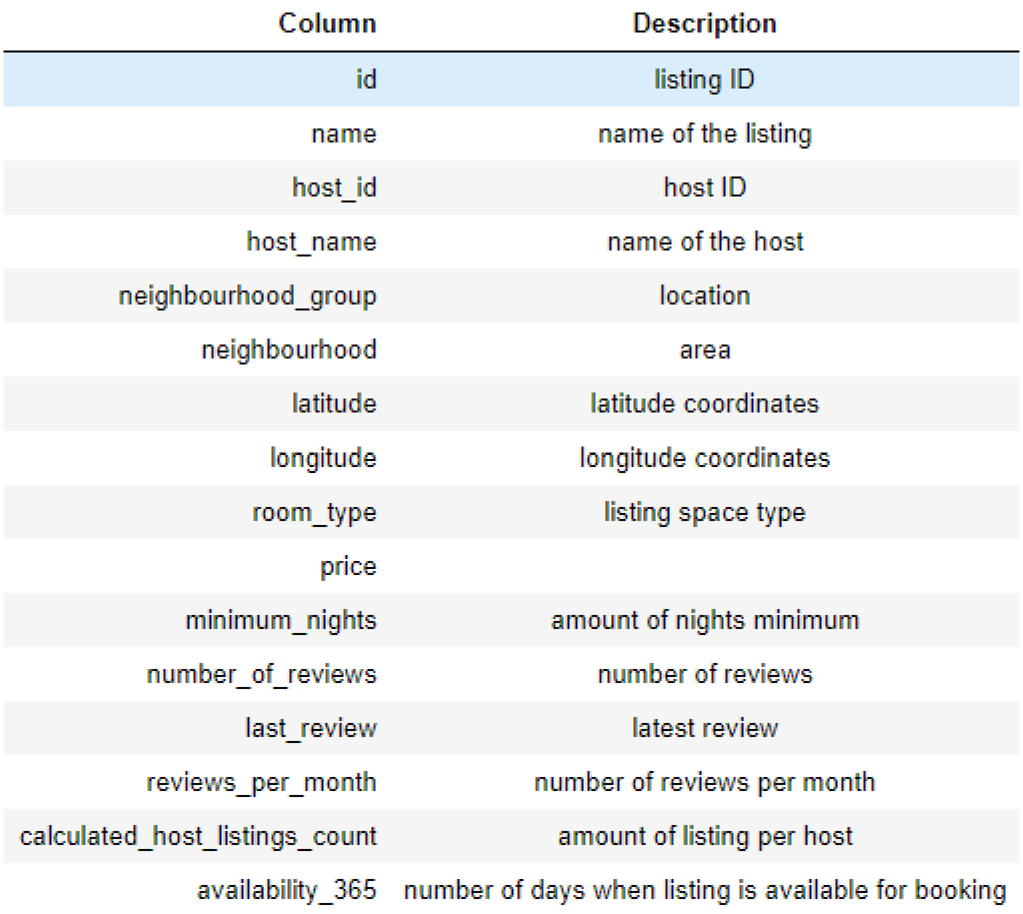
Evaluating Questions
Which locations are getting more traction?
Which locations are price and review sensitive?
What are the pricing ranges preferred by end customers?
What type of properties are preferred by the customers?
Which properties are available for more days in a year and in which location?
In what time period the properties have received more or less number or reviews?
What are the most popular localities and properties in New York currently?
Which properties and room types have more or less minimum night stay?
How many sites are hosted by a single host and what are its success metrics?
Which hosts have received better reviews?
Which are the locations that are not performing well based on reviews and other parameters?
Which are the room types that are not performing well?
Which parameter makes the customer prefer the property and provide a review?
Is there any correlation between the prices and reviews or other parameters?
Which location has properties functioning for more than 300 days in a year or less than 50 days?

C. Findings & Insights

8. *Basic Data Interpretation:*

- There are 16 columns and 48895 rows in the dataframe.
- There are 3 floats, 7 integers and 6 objects data type values in the dataframe.
- There seems to be many columns with missing values.
- Need to check the reason behind the missing values and some feature engineering needed too.

9. *Variables in the dataframe:*

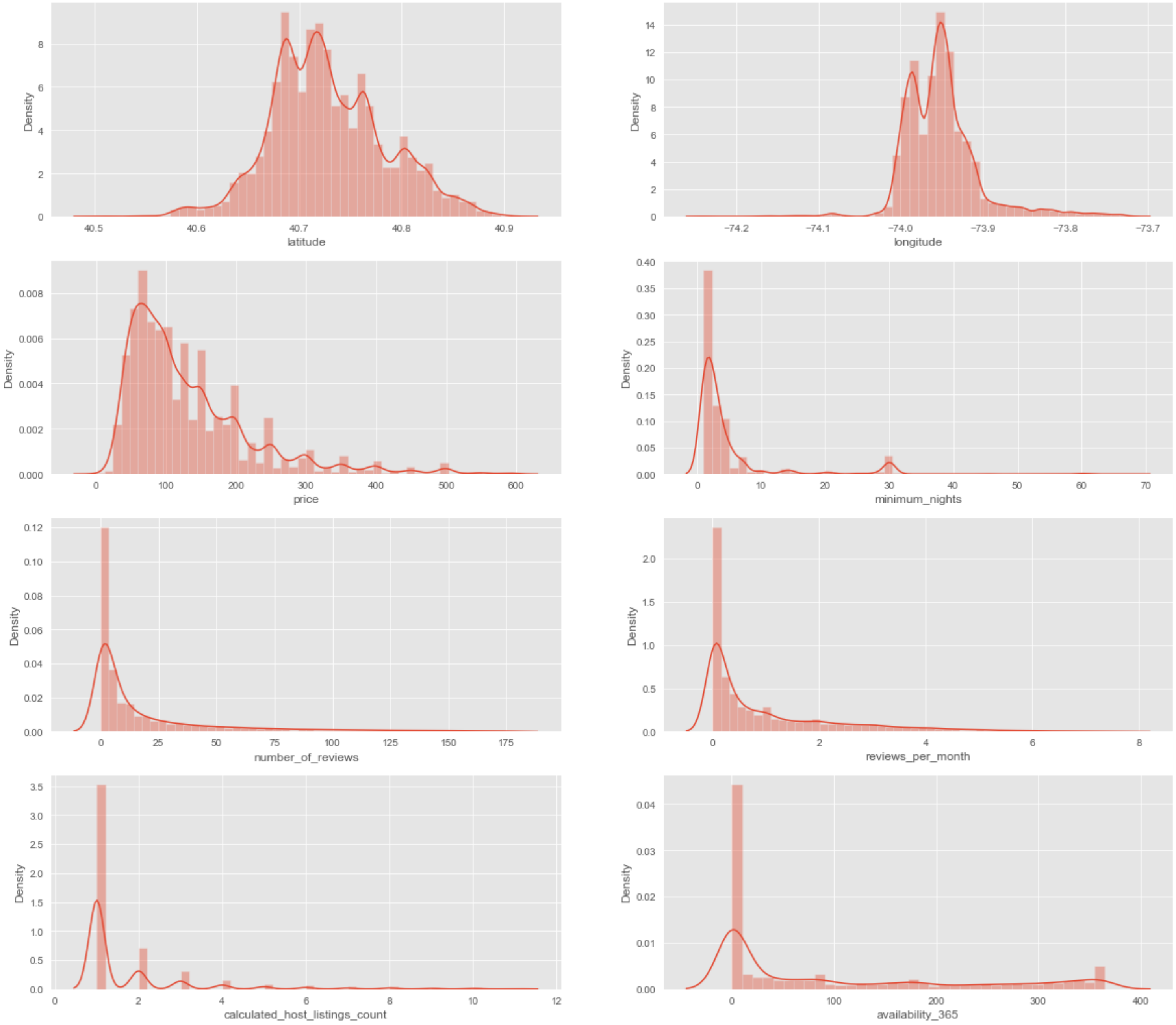


Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	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

10. *Numeric and Categorical Analysis:*

- Latitude and Longitude obviously belong to New York city as we have the data from the same.
- We can see prices starting from 0 dollars going upto 10 grand in dollars for hosting a place. Question: why is there a 0 dollar price for a hosted place on Airbnb and for which place?
- There seems to be a huge variance in minimum_nights, number_of_reviews, reviews_per_month and calculated_host_listings_count columns. Some things need to be checked in detail too under these columns.
- Manhattan seems to be the most friendly neighborhood group and Williamsburg the most known location underneath that.
- Hillside Hotel is the top hosted place which is the Entire home/apt room type.
- This place is hosted by Michael.

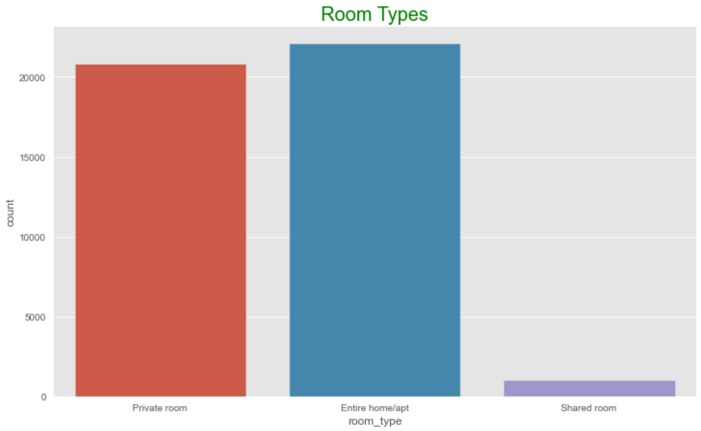
11. Numerical Univariate Analysis:

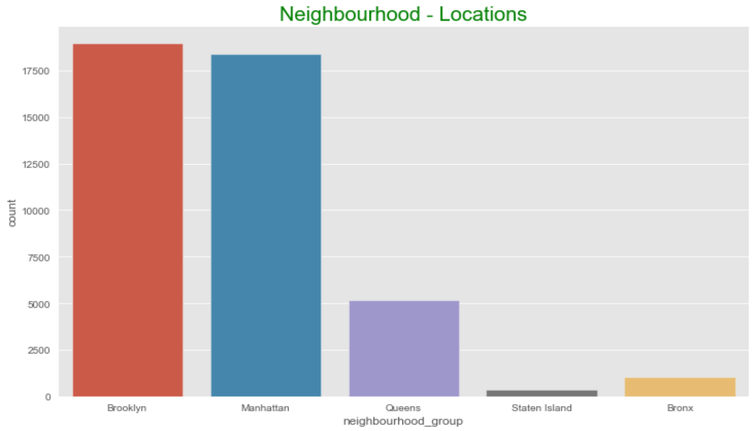


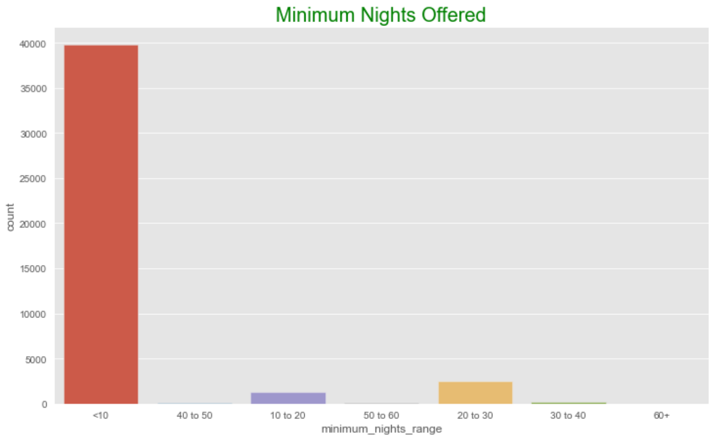
Findings:

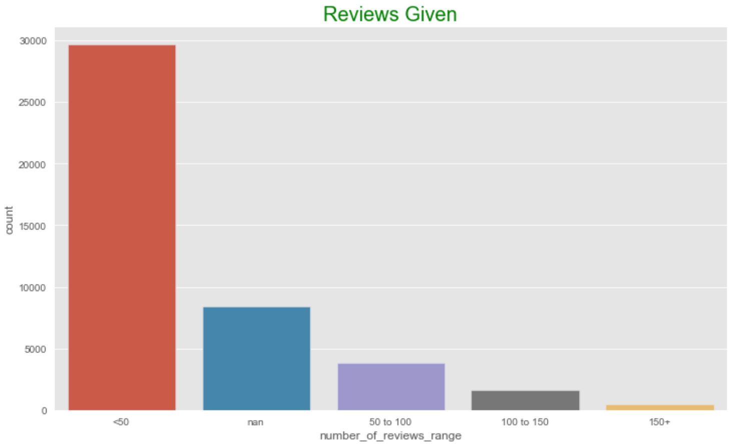
- The Highest price range seems to be between 30 dollars to 150 dollars per day stay for most of the sites hosted.
- Still we can see there are many sites which cost more than 200 dollars per day and can even go upto 500 dollars.

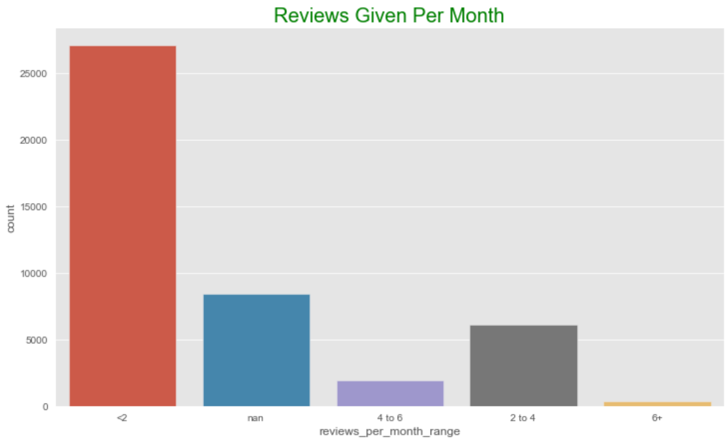
12. Categorical Univariate Analysis:

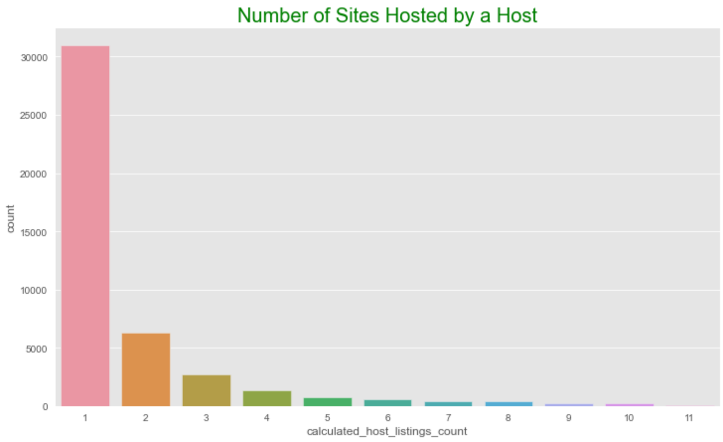

Room Types

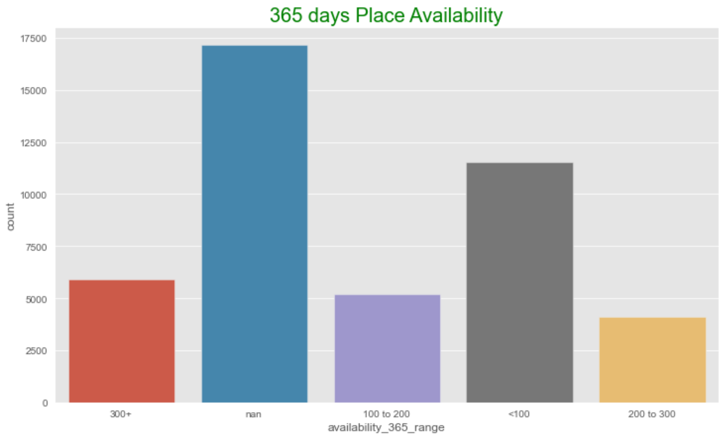

Neighbourhood - Locations

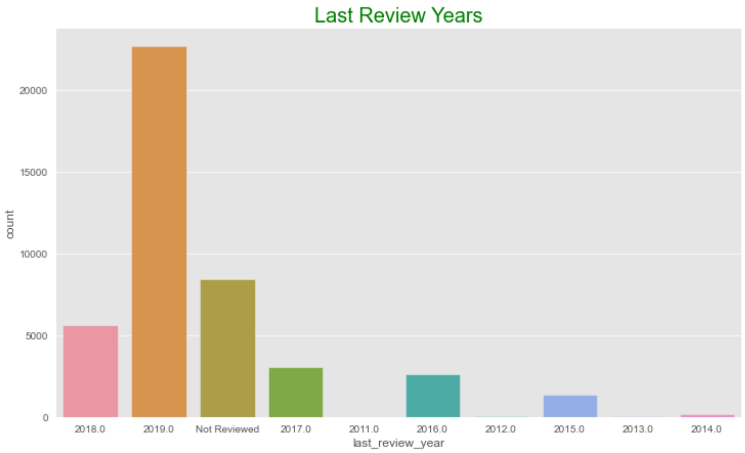

Minimum Nights Offered

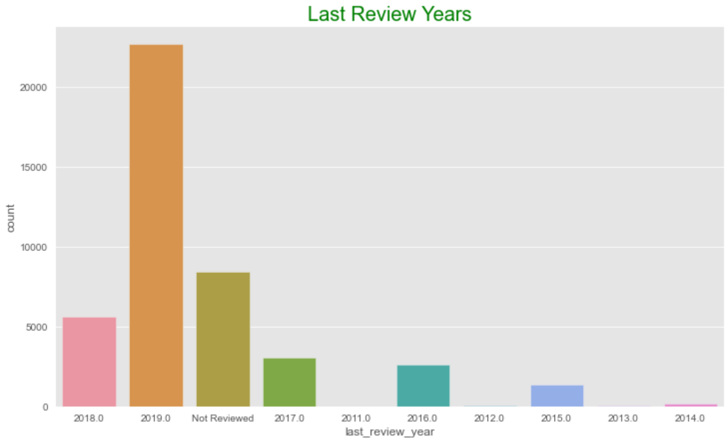

Reviews Given

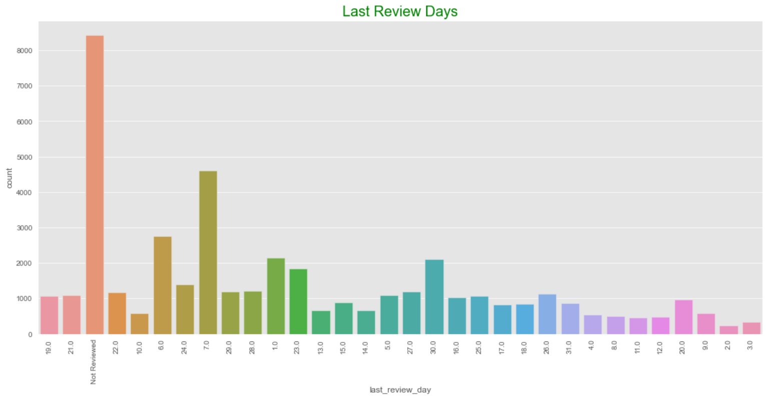

Reviews Given Per Month


Number of Sites Hosted by a Host







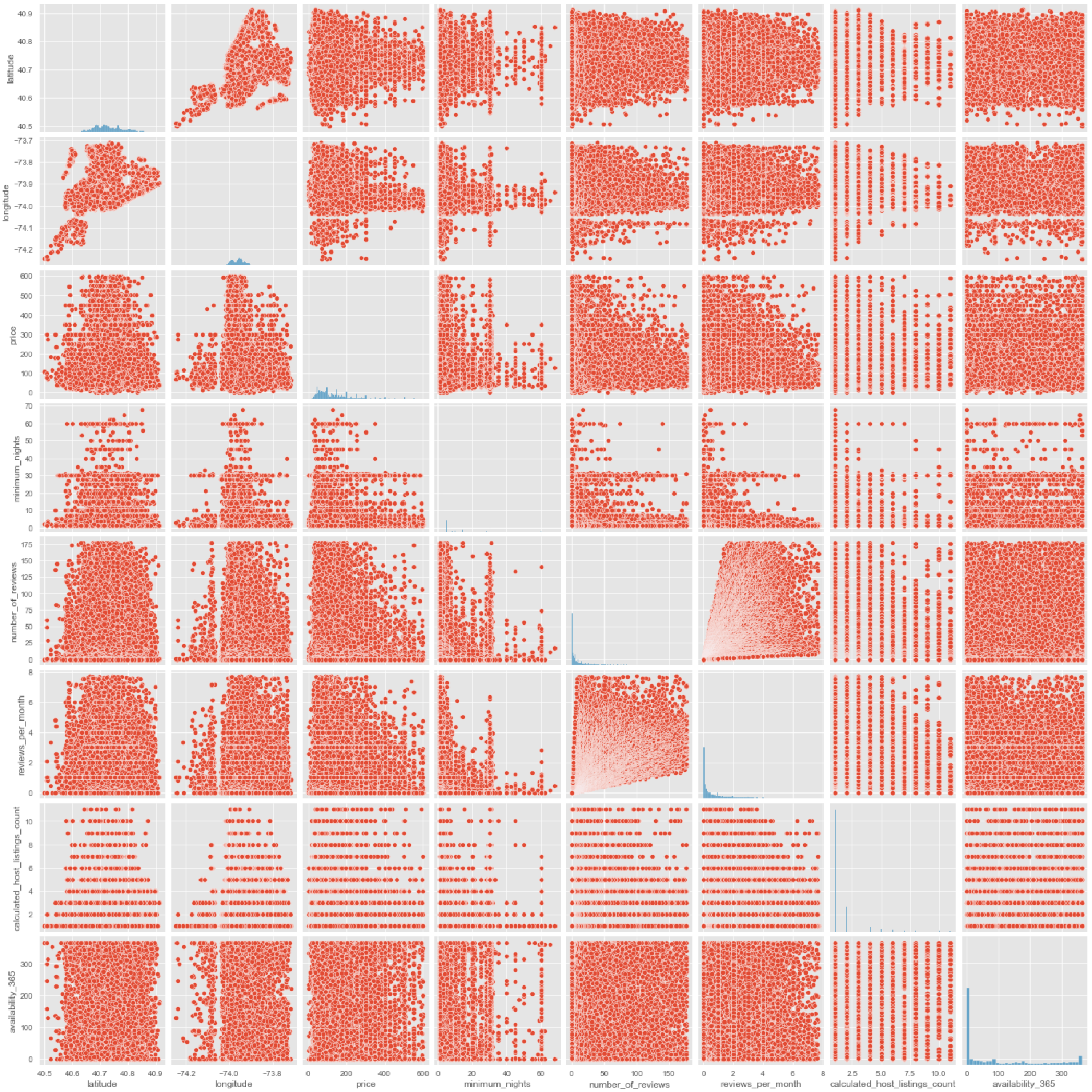


Findings:

- Most of the time last reviews were not provided when we see Day wise. Next, the majority of times it was provided on the 6th and 7th day of the month followed by the 1st and last day of the month.
- The 6th month of the year i.e June seems to receive most of the last reviews in all years followed by 5th month.
- Initial years which were 2011 till 2014, last reviews were negligible. After that it is slowly going up and most of the last reviews are received in the recent years of the data that is 2019 and 2018.
- Most of the sites hosted have less than 100 days availability in comparison to all 365 days. Also, majority of them have provided 0 days availability which has to be cross-checked by the Hosting Acquisition and Operations teams to know the reason.
- Majority of the hosts have less than 2 sites hosted by them on the platform.
- Most of the sites have received less than 2 reviews per month which indicates bad customer experience offered by majority sites.
- Also, Majority of the sites have received less than 50 reviews till date which is kind of less as per social norms.
- Majority of the sites provide less than 10 nights stay at a time.
- Majority of the sites hosted are either Private rooms or Entire apartments but very less Shared rooms.
- Broklyn and Manhattan are dominating when it comes to listed hostings followed by Queens.

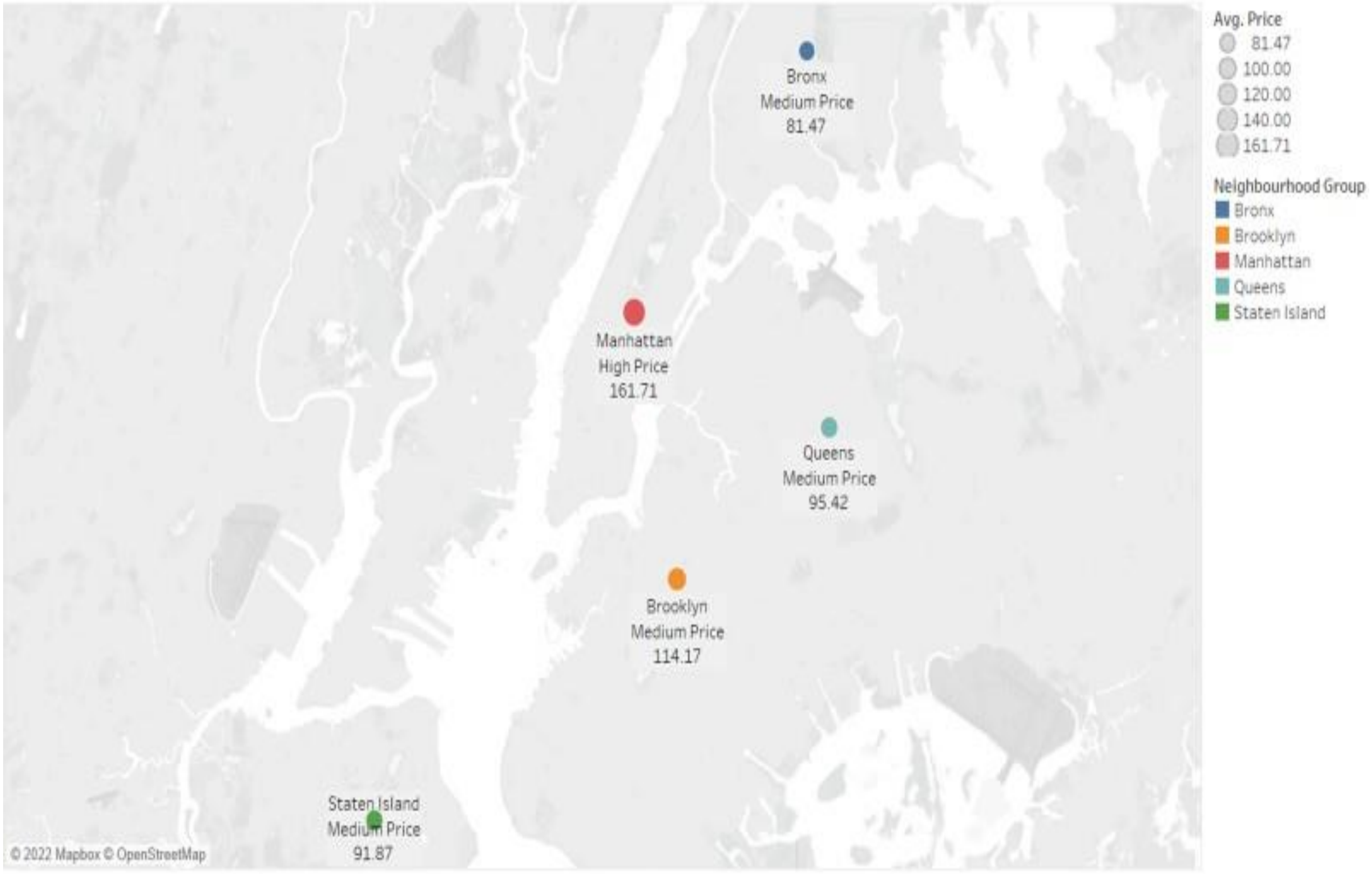
13. *Multivariate Analysis:*

# Numeric Correlation Analysis



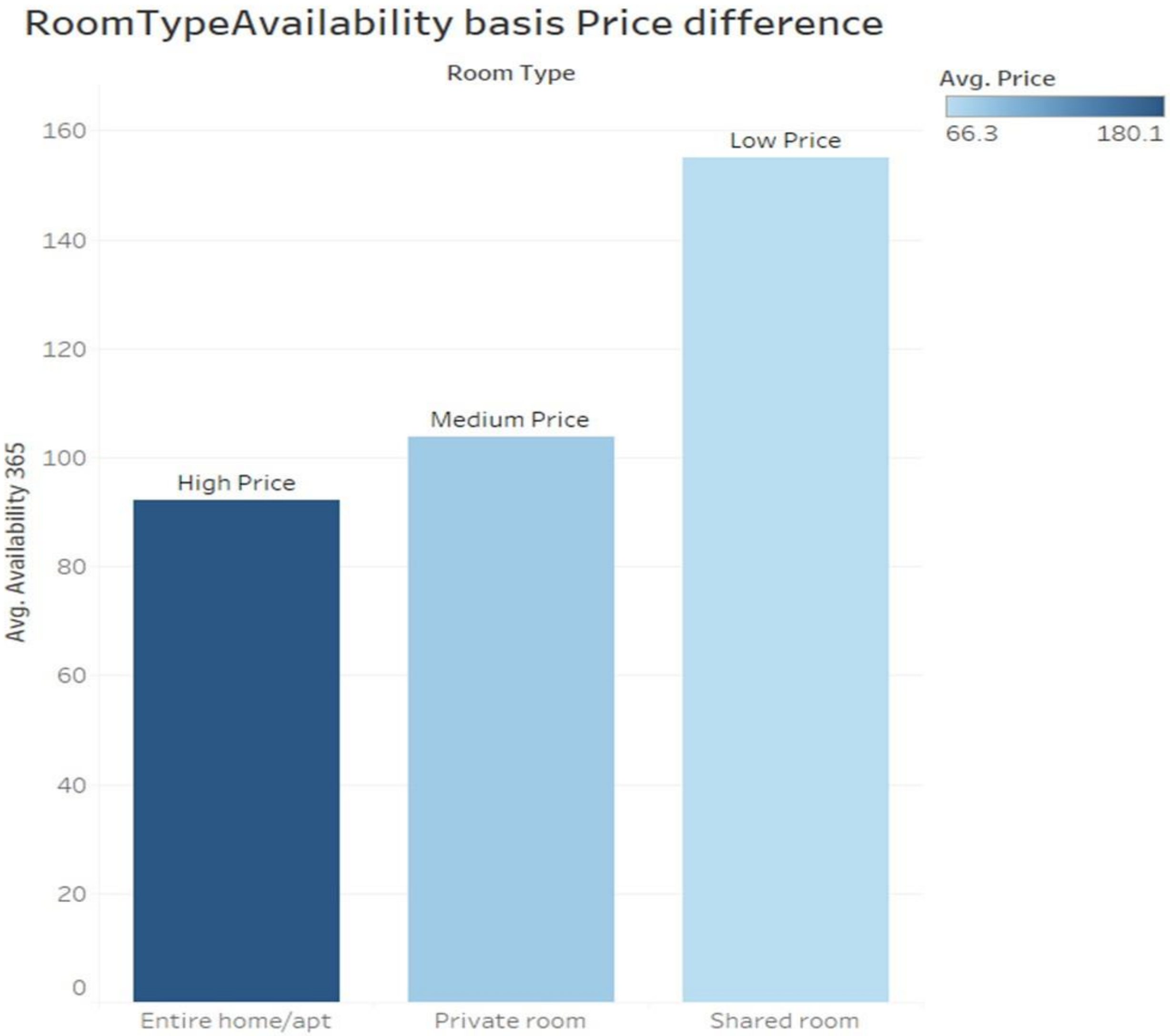
Overview of the Major Boroughs (New York City) based on Price Range

Neighbourhood Groups Basis Price Range



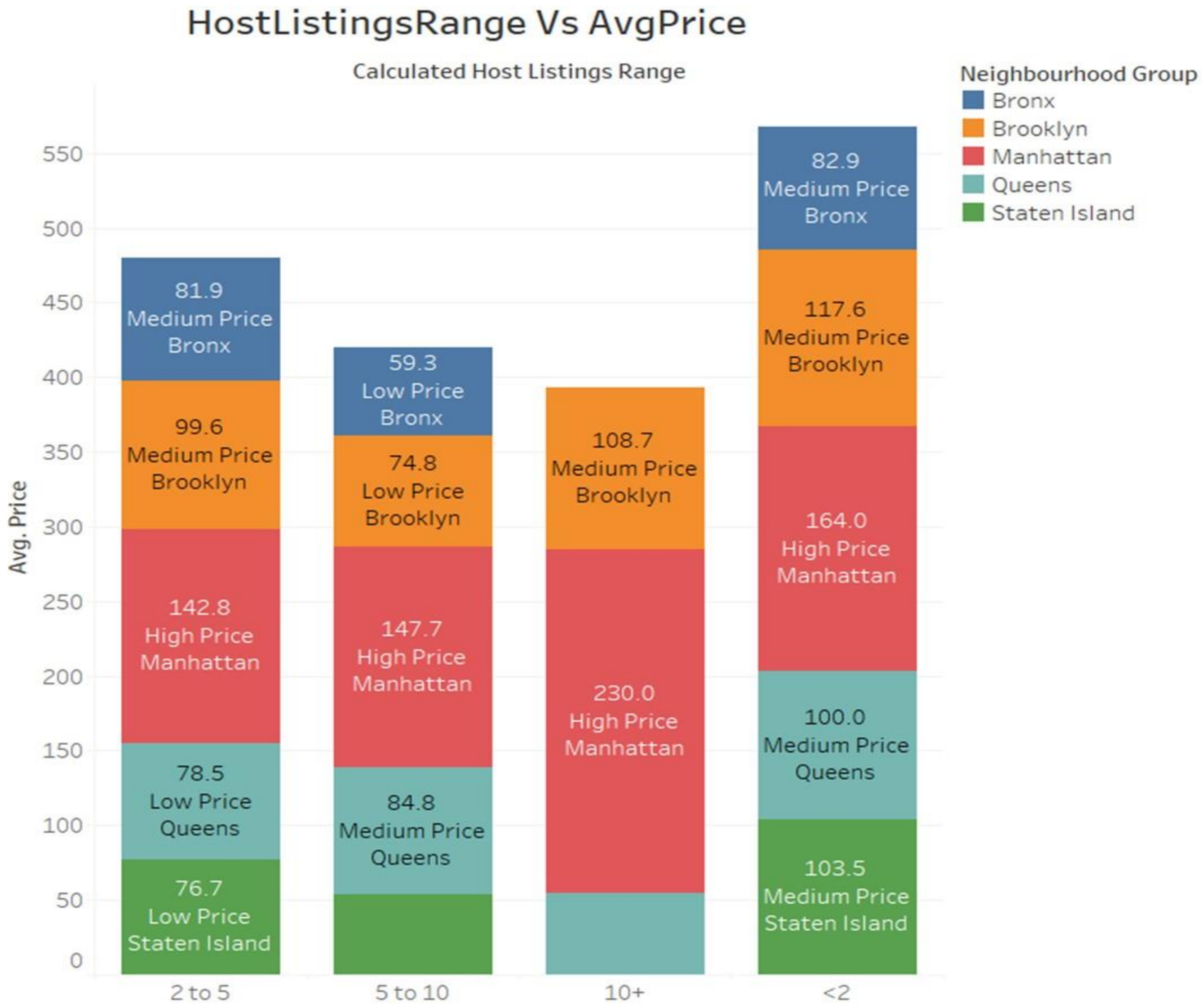
Map based on average of Longitude and average of Latitude. Color shows details about Neighbourhood Group. Size shows average of Price. The marks are labeled by

Types of Rooms available for N number of days in a year based on their price range.



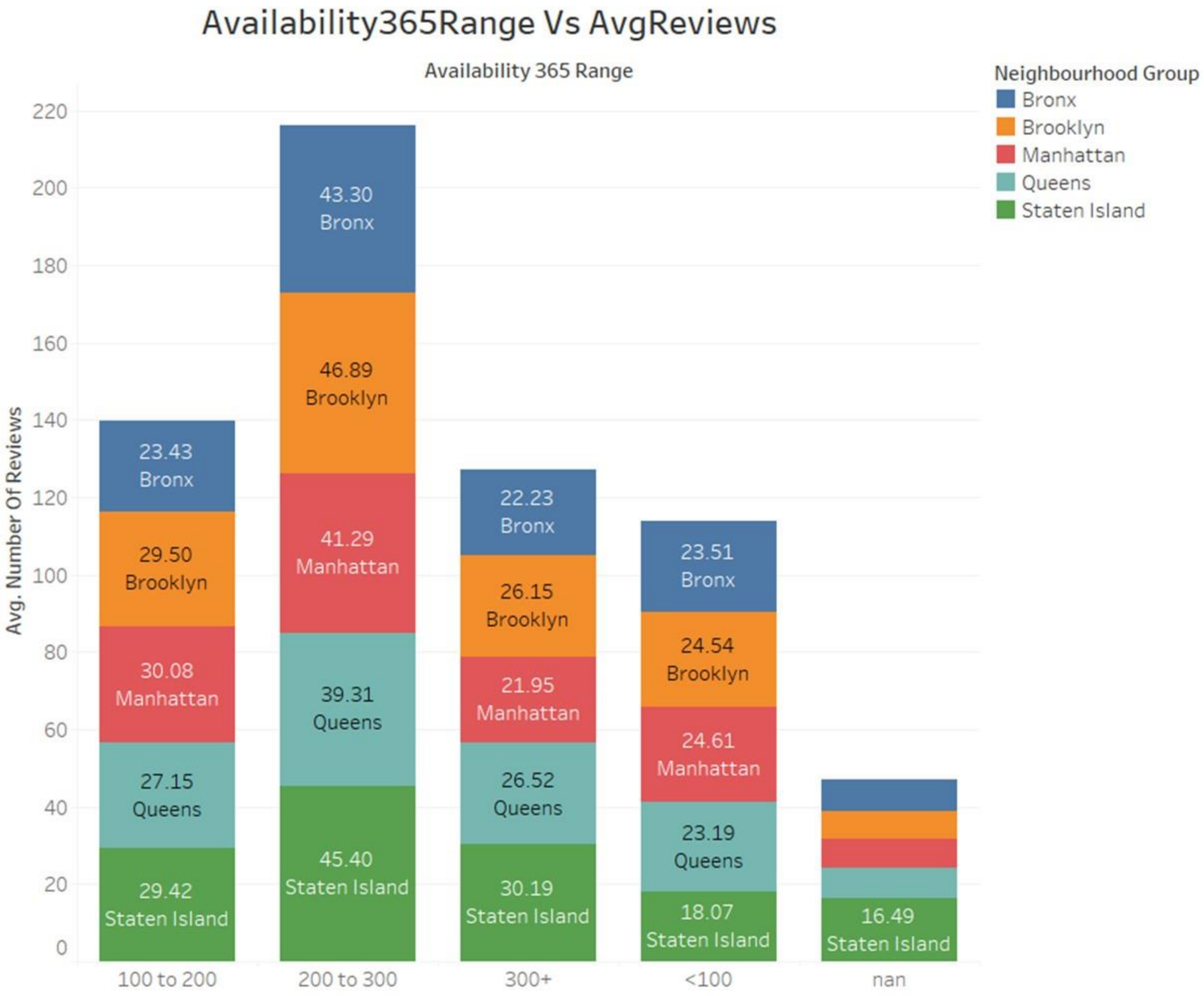
Average of Availability 365 for each Room Type. Color shows average of Price. The marks are labeled by Price Range.

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



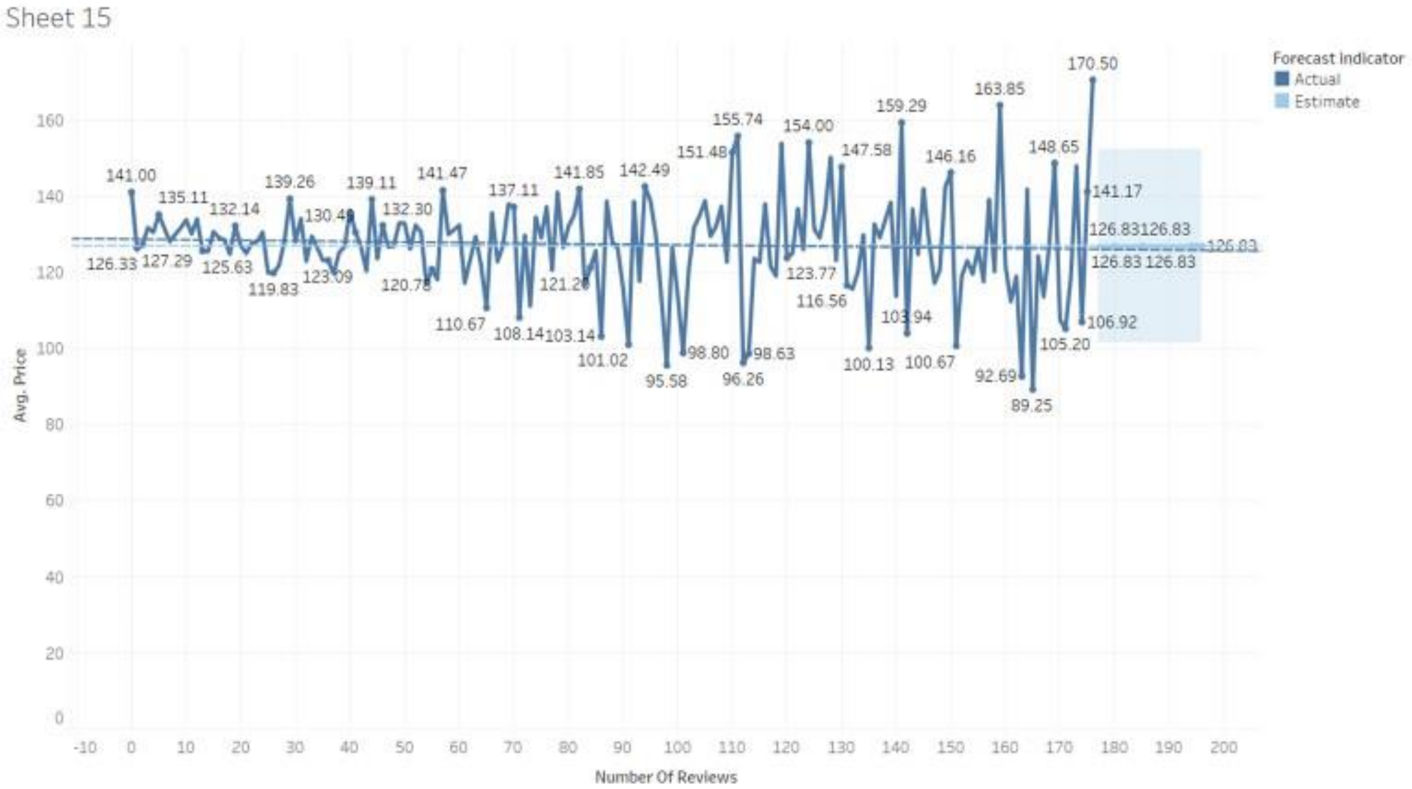
Average of Price for each Calculated Host Listings Range. Color shows details about Neighbourhood Group. The marks are labeled by average of Price, Price Range and Neighbourhood Group.

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



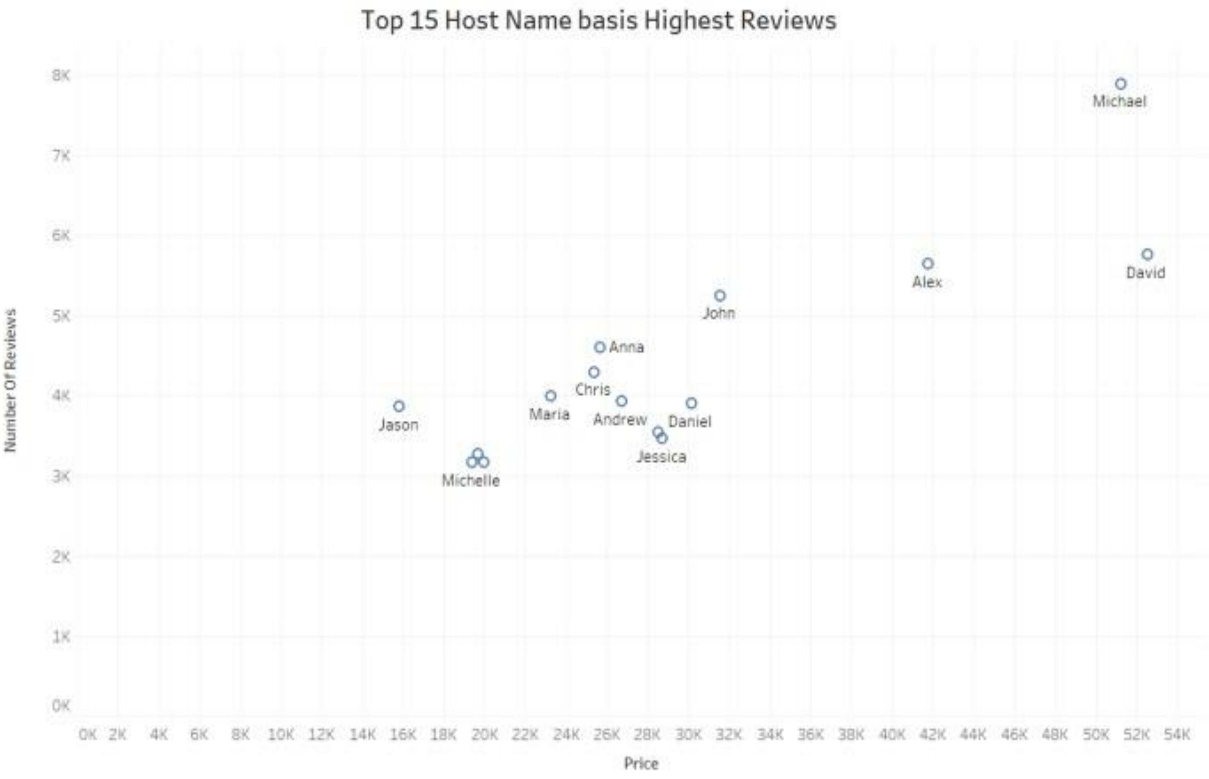
Average of Number Of Reviews for each Availability 365 Range. Color shows details about Neighbourhood Group. The marks are labeled by average of Number Of Reviews and Neighbourhood Group. The view is filtered on Neighbourhood Group, which keeps Bronx, Brooklyn, Manhattan, Queens and Staten Island.

# Price range preferred by the customers



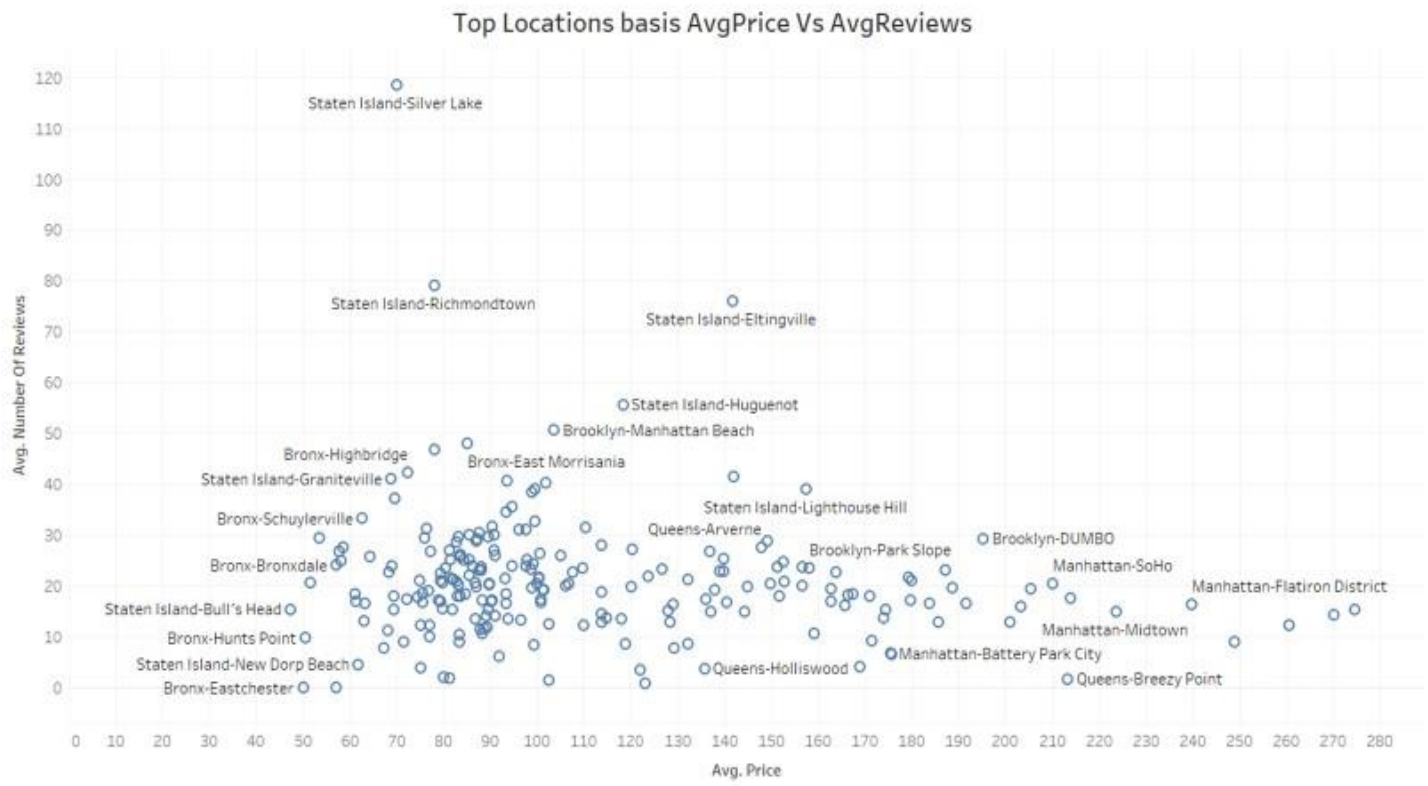
The trend of average of Price (actual & forecast) for Number Of Reviews. Color shows details about Forecast indicator. The marks are labeled by average of Price (actual & forecast).

# Name of the Host who have received highest number of reviews



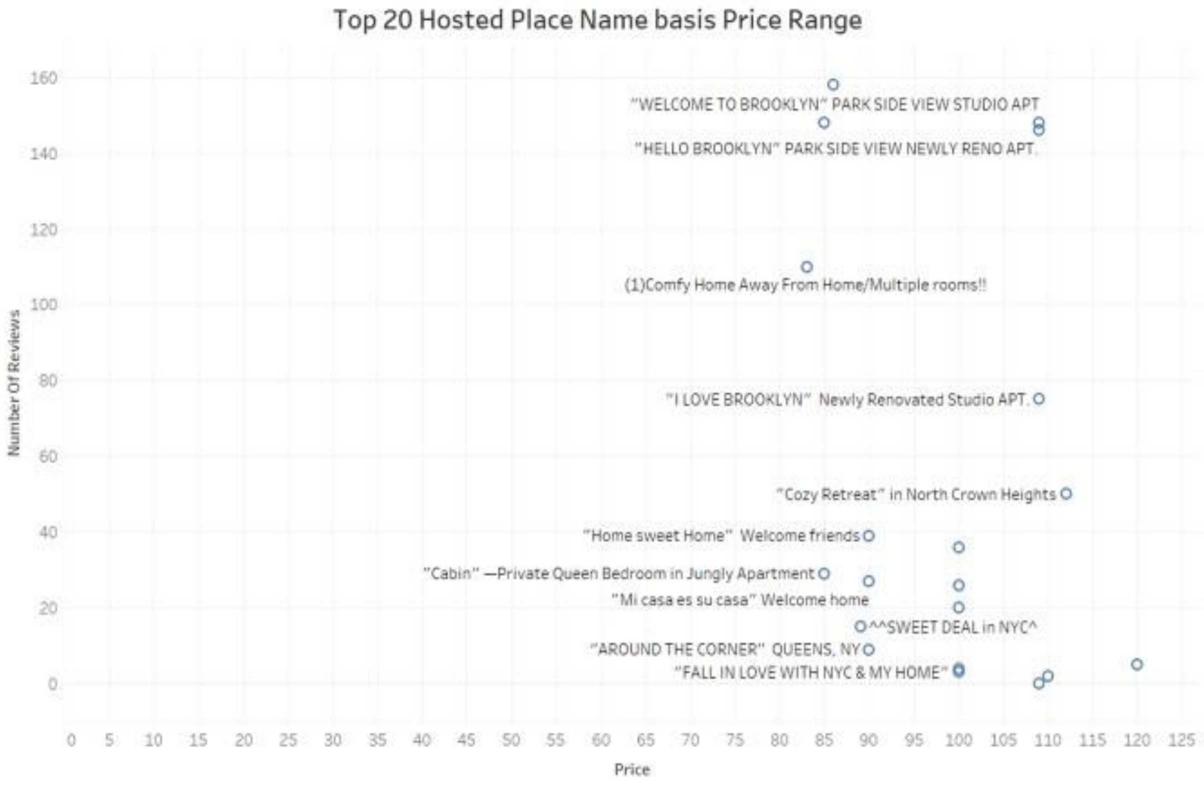
Sum of Price vs. sum of Number Of Reviews. The marks are labeled by Host Name. Details are shown for Host Name. The view is filtered on Host Name, which keeps 15 of 11,024 members.

# Name of the Locations that have received highest number of the Reviews



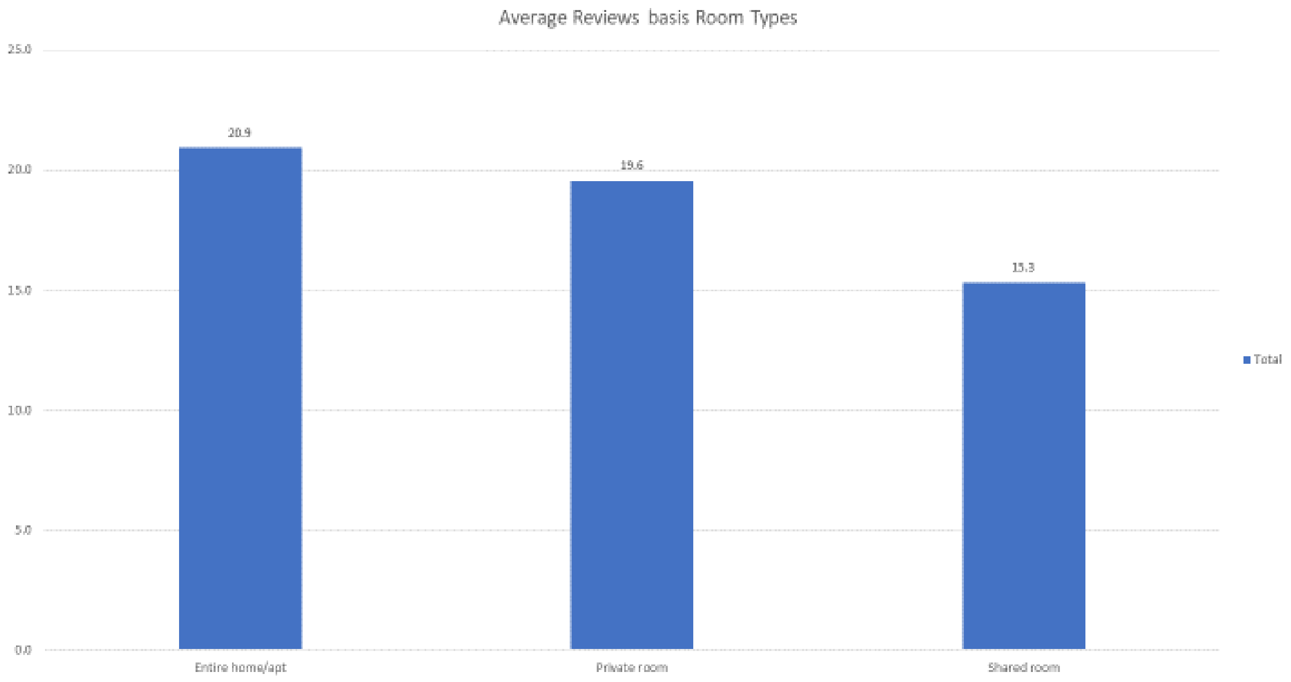
Average of Price vs. average of Number Of Reviews. The marks are labeled by Location. Details are shown for Location.

# Name of the Hosted Places that have received highest number of reviews and lies in High Price range

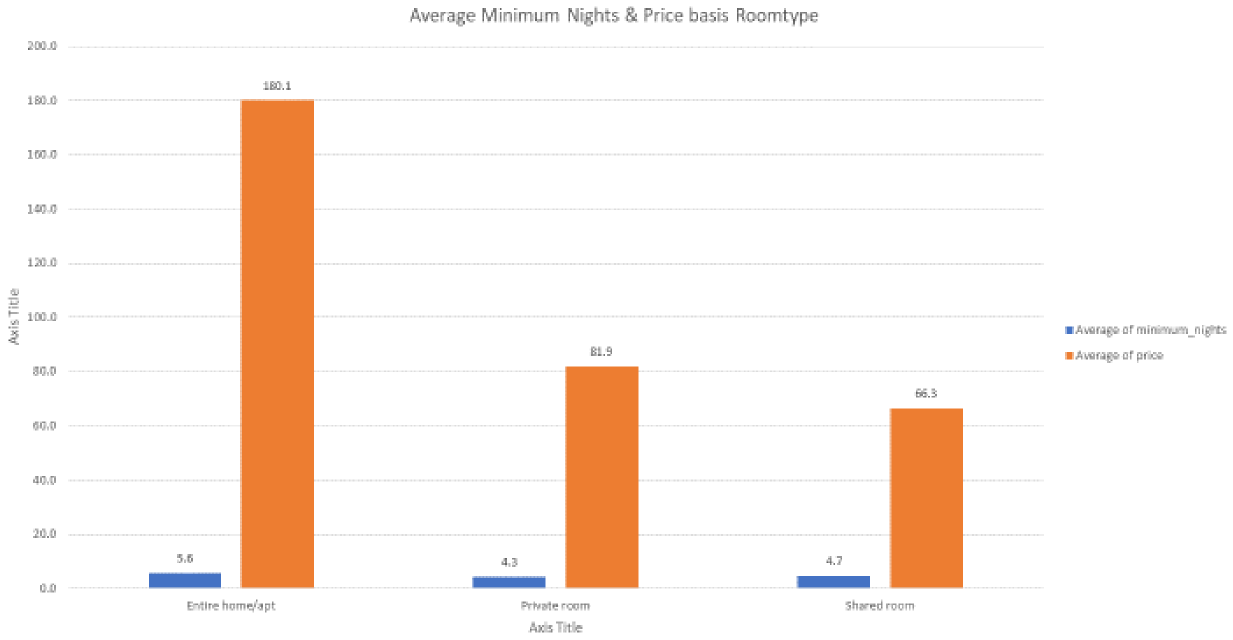


Sum of Price vs. sum of Number Of Reviews. The marks are labeled by Name. Details are shown for Name. The view is filtered on Name, which keeps 20 of 42,696 members.

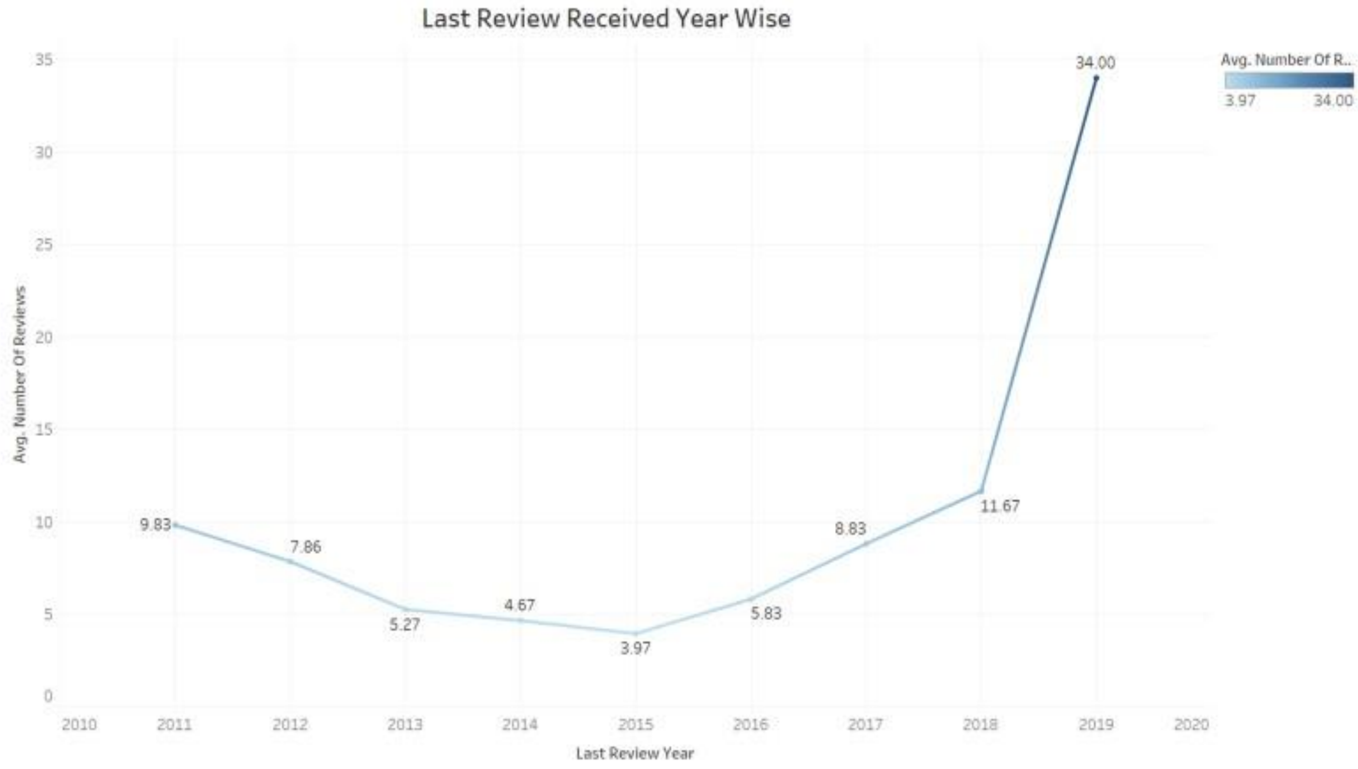
# Room Types preferred more by Customers



# Room Types available for more minimum night stay contradicting their price range

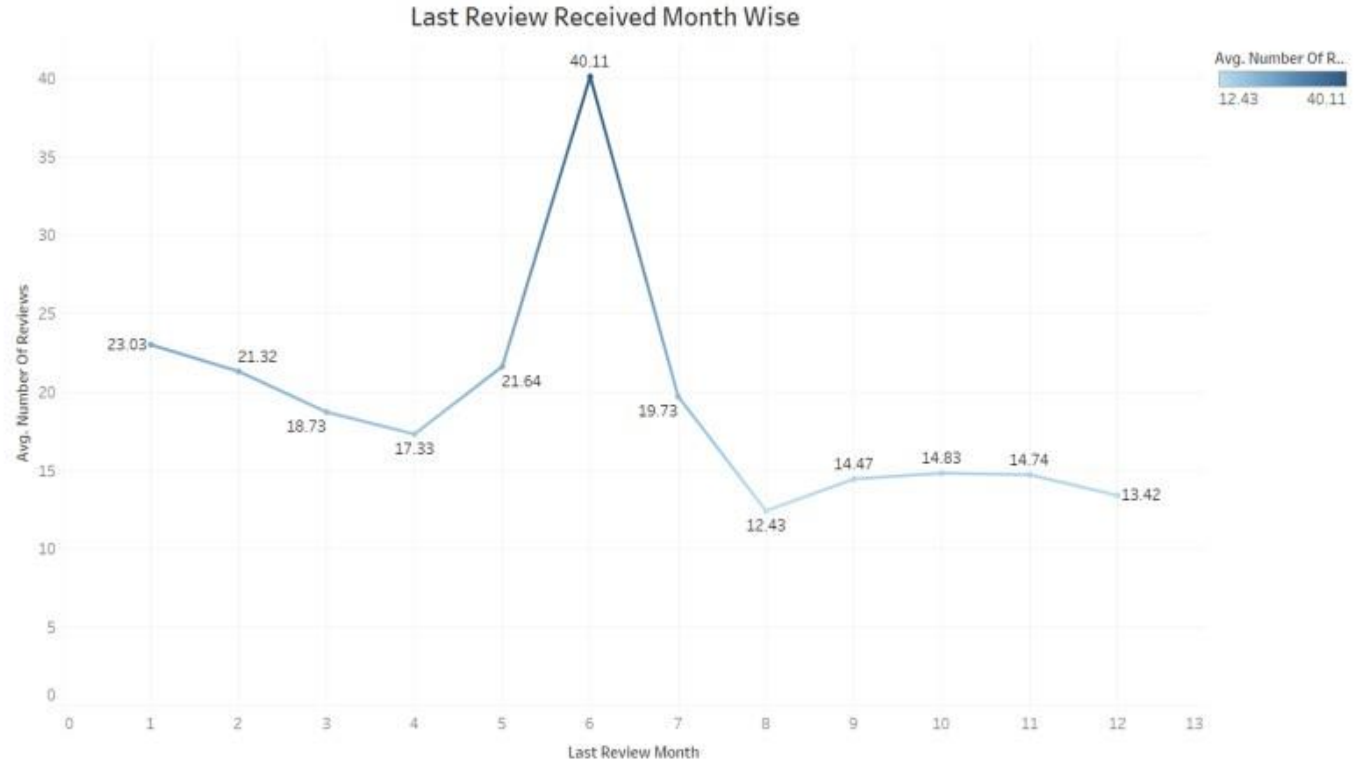


# Years which have impacted more in customer reviews



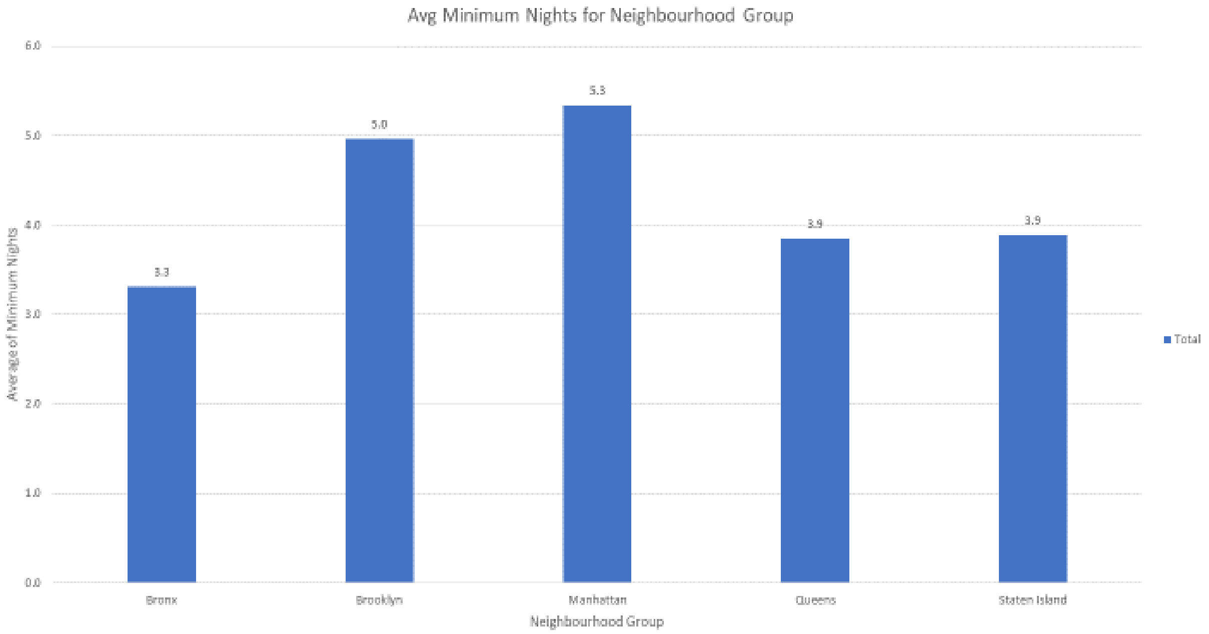
The trend of average of Number Of Reviews for Last Review Year. Color shows average of Number Of Reviews. The marks are labeled by average of Number Of Reviews. The view is filtered on Last Review Year, which keeps non-Null values only.

# Months in which the end customer are active more in providing reviews

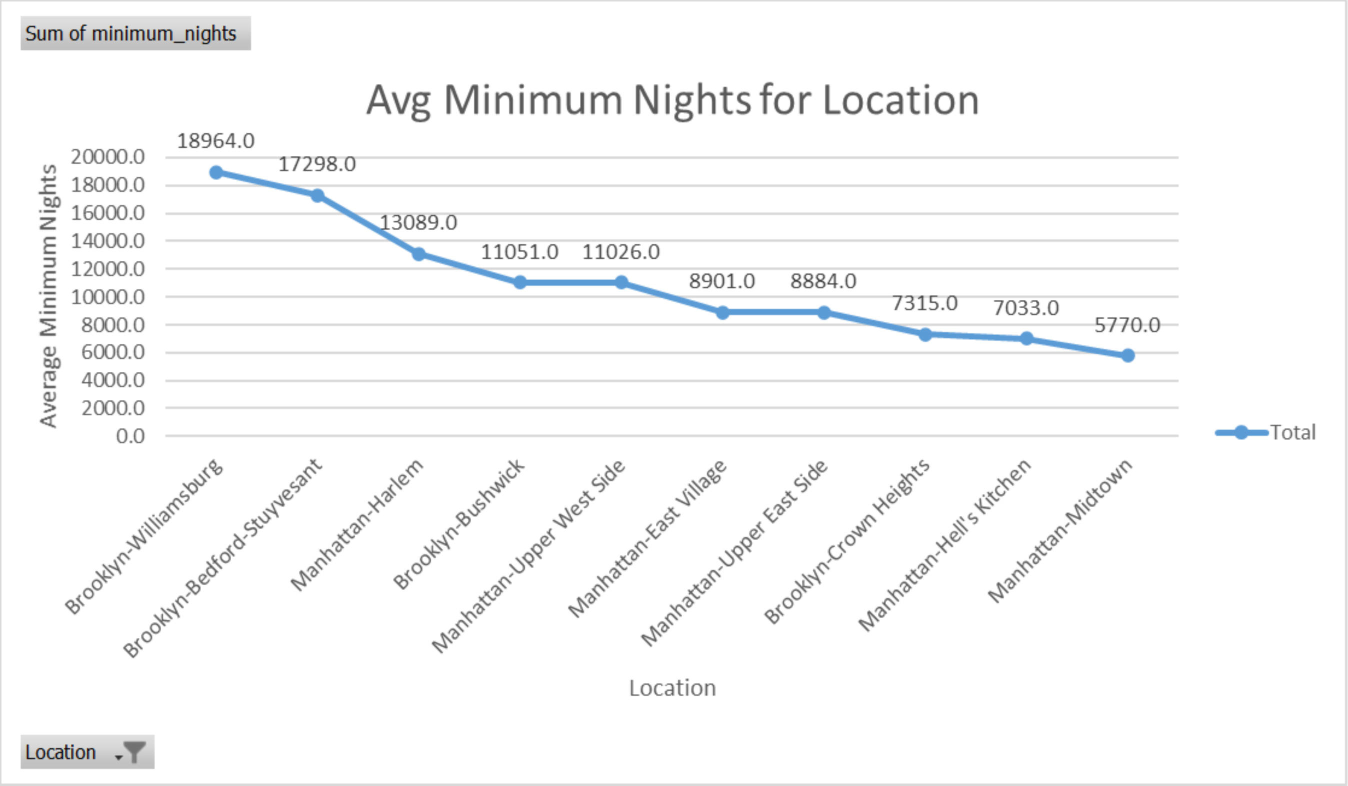


The trend of average of Number Of Reviews for Last Review Month. Color shows average of Number Of Reviews. The marks are labeled by average of Number Of Reviews. The view is filtered on Last Review Month, which keeps non-Null values only.

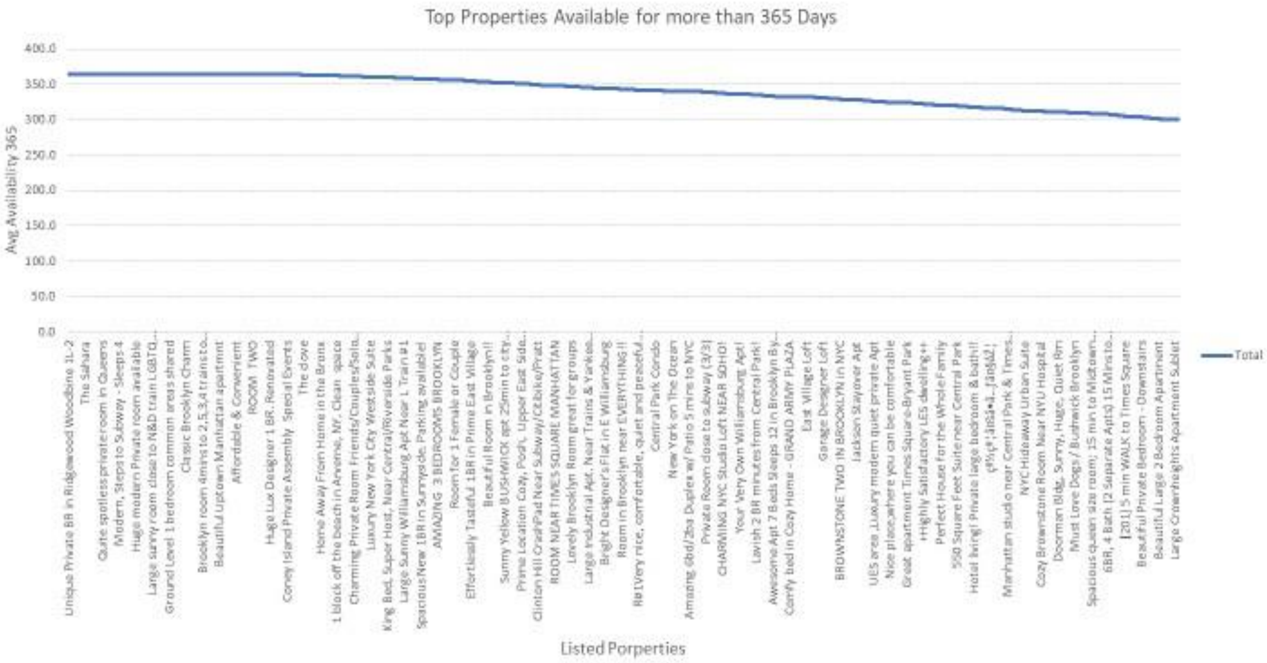
# Top Neighborhoods providing higher number of Minimum Night stay



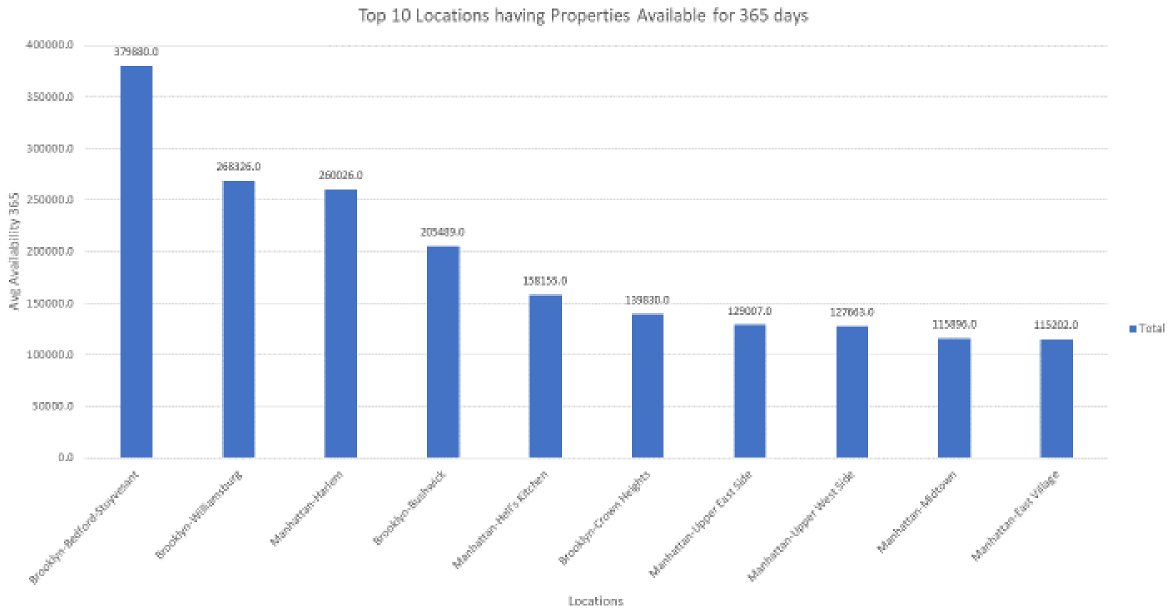
# Top Locations providing higher number of Minimum Night stay



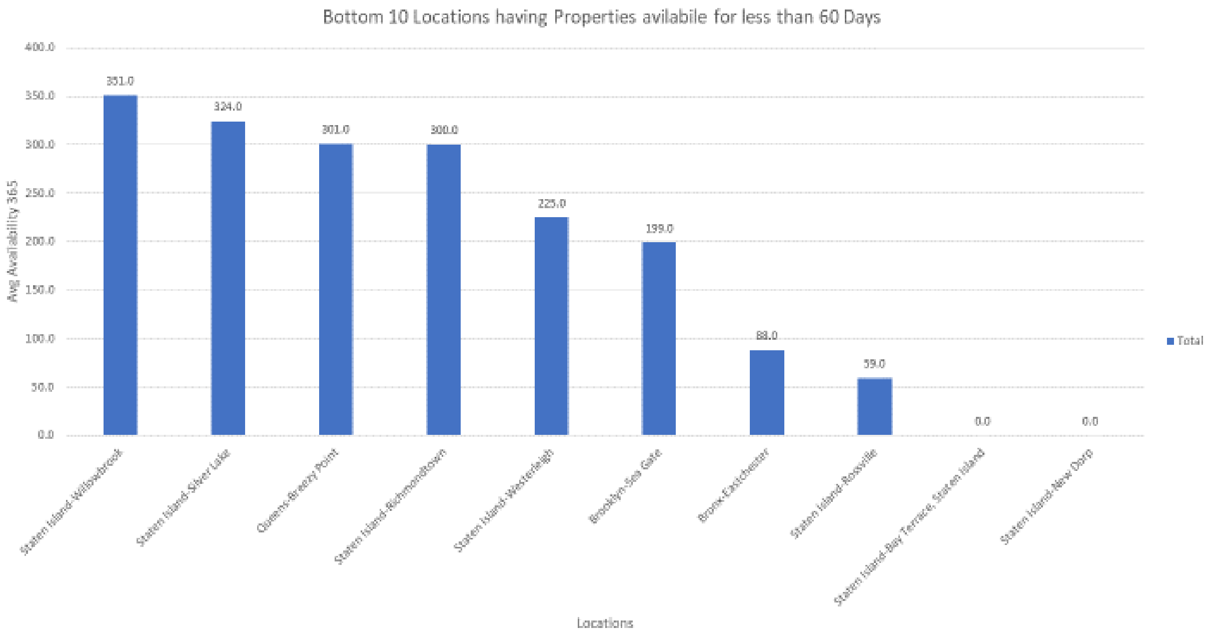
# Properties contribution more on the Platform



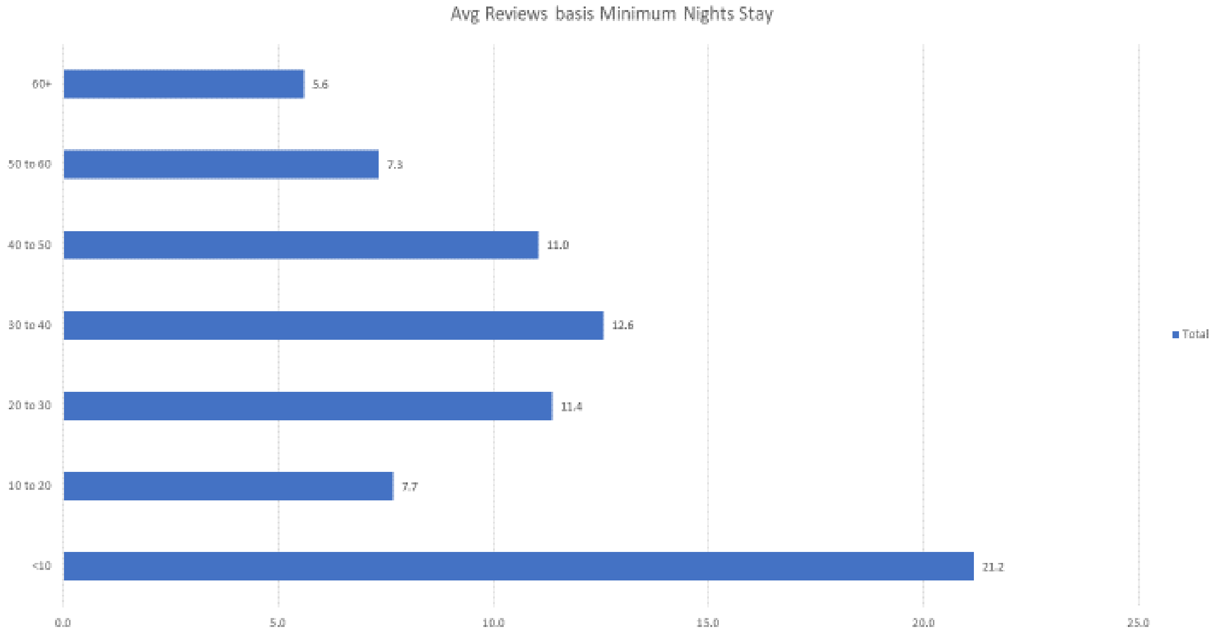
# Locations contributing more on the Platform



# Locations contributing less on the Platform



# Does higher number of Minimum night stay receive the highest number of reviews



Important Findings:

1. There seems to be no positive or any type of correlation between the numerical variables.

2. 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 considered to be a High Price, between 80$ to 120$, Medium Price range and less than 80$ to be considered Low Price range property.

3. Having a high price range, Entire home/apt types rooms are available for less than 100 days on average followed by Private rooms on an average of 105 days and Shared rooms around 155 days on average being the lowest in price.

4. Manhattan has the highest number of places listed around more than 10 by a single host with an average price of 230 $ followed by Brooklyn with an average price of 108$. On the other hand, all the hosts have less than 2 properties listed in either of the Borough on an average price range between 80 $ to 170 $.

5. Brooklyn has received the highest number of reviews based on the availability to stay open for more than 200 days in a year. This is followed by Staten Island and then the Bronx. 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. **We need to check which are these places and what issues are they facing?**

6. 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.

7. **Michael, David, Alex, John and Daniel 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.
8. 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.

9. **““WELCOME TO BROOKLYN"" PARK SIDE VIEW STUDIO APT"” , ““Oasis on the Park””, ““HELLO BROOKLYN"" PARK SIDE VIEW NEWLY RENO APT””, ““Comfy Home Away From Home/Multiple rooms”, “”LOVE BROOKLYN"" Newly Renovated Studio APT”” and ““Cozy Retreat"" in North Crown Heights”**” are the Top 6 listed places that have received highest number of reviews.

10. 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.

11. **“Modern Duplex - Central Chelsea!!!” in Manhattan-Chelsea, “Spacious & Bright 3BRs Near Subways, Parks, Shops” in Brooklyn-Cobble Hill, “NYC LUXURY3 BEDROOMS IN MIDTOWN EAST & GYM& BALCONY” in Manhattan-Murray Hill, “An Artist's Inspiration: Sun-Soaked Chelsea Loft” in Manhattan-Chelsea and “Upper West Side elegance. Riverside” in Manhattan-Upper West Side** are the Top 5 hosted places with highest price offerings.

12. **“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.

13. The average number of reviews started increasing exponentially after 2015-2016. And majority of the customers provide higher number reviews either between the months of May till July or in the starting of the year which shows the higher booking window in a year.

14. There are 5766 properties that are open for more than 300 days in a year. Around 2286 of them are from Brooklyn followed by Manhattan of around 1947 properties. And on the other hand, the properties that stay open for less than 50 days a year belong to Queens or Staten Island.

15. 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.

D. Tools Used

- ***Python*** *used for Data Understanding, Pre-processing and general Univariate and Multivariate Analysis.*
- ***Tableau & Excel*** *used for in-depth Bi-Multivariate Analysis.*

THANK YOU