Methodology Document

Case Study Purpose: **Methodological Document for Data Analysis on Airbnb**

Methodology Approach:

Tools Used: Python**,** Tableau, Excel

1. Problem Background:

For the past few months, Airbnb has noticed a major decline in revenue due to the lockdown imposedduring the pandemic.

Currently that the limitations 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 a websiteand mobile app.

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

3. Type of Data required to Analyze:

Fall in the revenue could be for two major causes, either the sites hosted on the platform are not able to deliver a better user experience or there could be a competitor in the market catching the market share. Keeping the above in mind, we first try to work on the first reason as that is something internal to the companyand can have the data in hand to identify the reasons behind the plummeting 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 host's details, prices of the hosted sites and reviews received by the end consumer

4. How was the Data obtained?

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 propertiesoffer falls under the purview of this role.

Head of User Experience, NYC: The head of a user experience looks after the customer preferences and also handles the properties listed on the website and the Airbnb app. The head of a user experience tries to optimize the order of property listing in certain neighborhoods and cities 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 theirchallenges for being fully functional for the 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 communityfor better improvement and value-generating ideas.

Provide discounted commission rates to property owners on keeping the minimum night stay booking windowfor more than 10 days and property functional for the maximum number of days in a year.

Method Analysis Code with Python:

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,

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.pyplot as plt, seaborn as sns
import plotly.express as px
%matplotlib inline

pd.set_option('display.max_columns',200)
pd.set_option('display.max_rows',500)
```

Ignore the warnings we might get,

```python
# Supress Warnings

import warnings
warnings.filterwarnings('ignore')
```

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

```python
airbnb = pd.read_csv("AB_NYC_2019.csv")
airbnb.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

```python
airbnb.shape
```

```
(48895, 16)
```

2. Columns With DATA-TYPES:

```
airbnb.info()

<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
```

Data Description:

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:

Numerical Data

```python
airbnb[['price', 'minimum_nights', 'number_of_reviews',
        'reviews_per_month', 'calculated_host_listings_count',
        'availability_365']]
```

```python
# Analyzing categorical values

airbnb.select_dtypes(include=['object']).describe()
```

3. Handling Missing Values and Outliers:

Using Isolation Forest:

Using Isolation Forest, we can not only detect anomalies faster but we also require less memory compared to other algorithms. Isolation Forest isolates anomalies in the datapoints instead of profiling normal data points.

```python
# Importing Isolation Forest Library
from sklearn.ensemble import IsolationForest
```

```python
# Setting Model with data set
model = IsolationForest(n_estimators=100, max_samples=50000, contamination=0.02)
model.fit(airbnb[['price']])
```
0

Handling outliers

```python
# Price field seems to have outliers, hence needed to be handled
airbnb['price_anomaly'] = model.predict(airbnb[['price']])

# anamoly scores
airbnb['price_scores'] = model.decision_function(airbnb[['price']])
```

```python
# price outliers percentage

price_anomaly = len(airbnb.loc[airbnb['price_anomaly'] == -1])
print('anomaly detection of price: ', round((price_anomaly/len(airbnb))*100,4), '%')
```
anomaly detection of price: 1.9859 %

```python
data_clean = airbnb.loc[airbnb['price_anomaly'] == 1]
```

Handing outliers in "Minimum_nights" column

```python
# Minimum night column also has outliers which can be handled

anomalies = len(data_clean.loc[(data_clean['minimum_nights'] > 60)])
print('anamaly in minimum nights values:', round((anomalies/len(data_clean))*100,4),'%')
```
anamaly in minimum nights values: 0.6448 %

```python
data_clean = data_clean.loc[airbnb['minimum_nights'] <= 60]
data_clean.info()
```

Outliers- availability_365

```python
anomaly = len(airbnb.loc[(airbnb['availability_365'] == 0)])
print(anomaly)
print('anomaly in availability_365: ', round((anomaly/len(airbnb))*100,4), '%')
```

```
17533
anomaly in availability_365:  35.8585 %
```

Since the amount of data is representative in this case and it may not represent an outliers, so its not a good idea to remove them

```python
# null values in last_review are filled with 'Not recieved'

airbnb.last_review.fillna('Not Recieved', inplace= True)
```

```python
# null values in host_name is filled with others

airbnb.host_name.fillna('others', inplace=True)
```
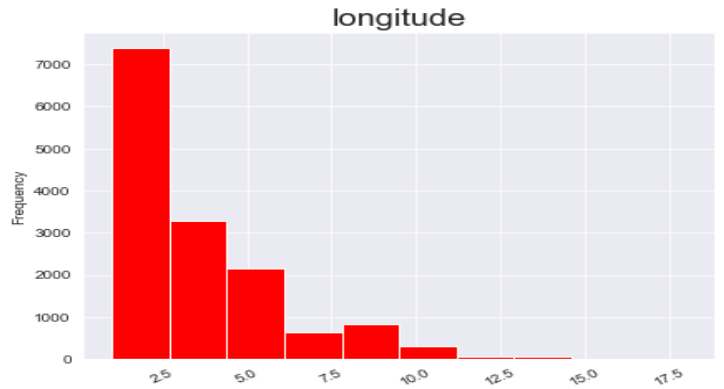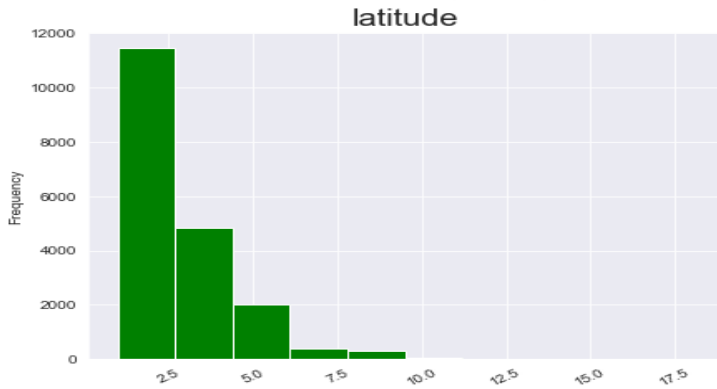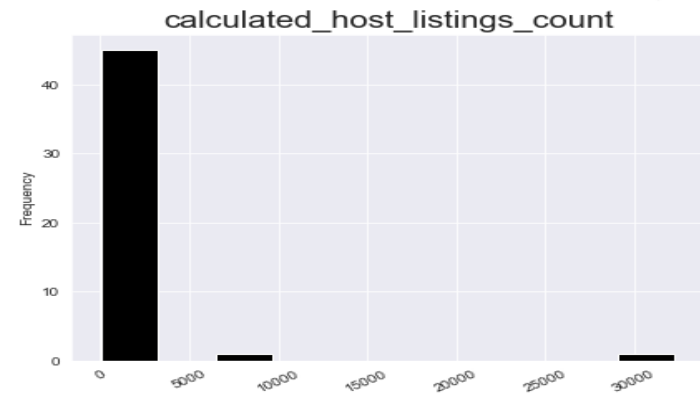
```python
# And remaining null values are dropped
airbnb.dropna(inplace=True)
```
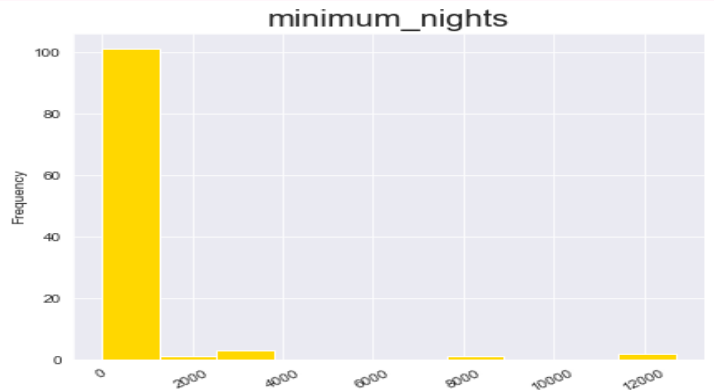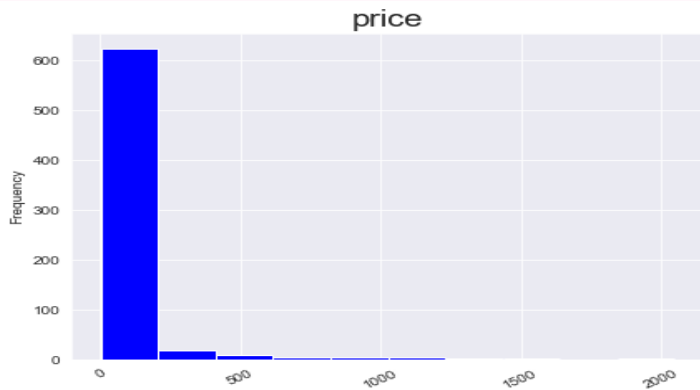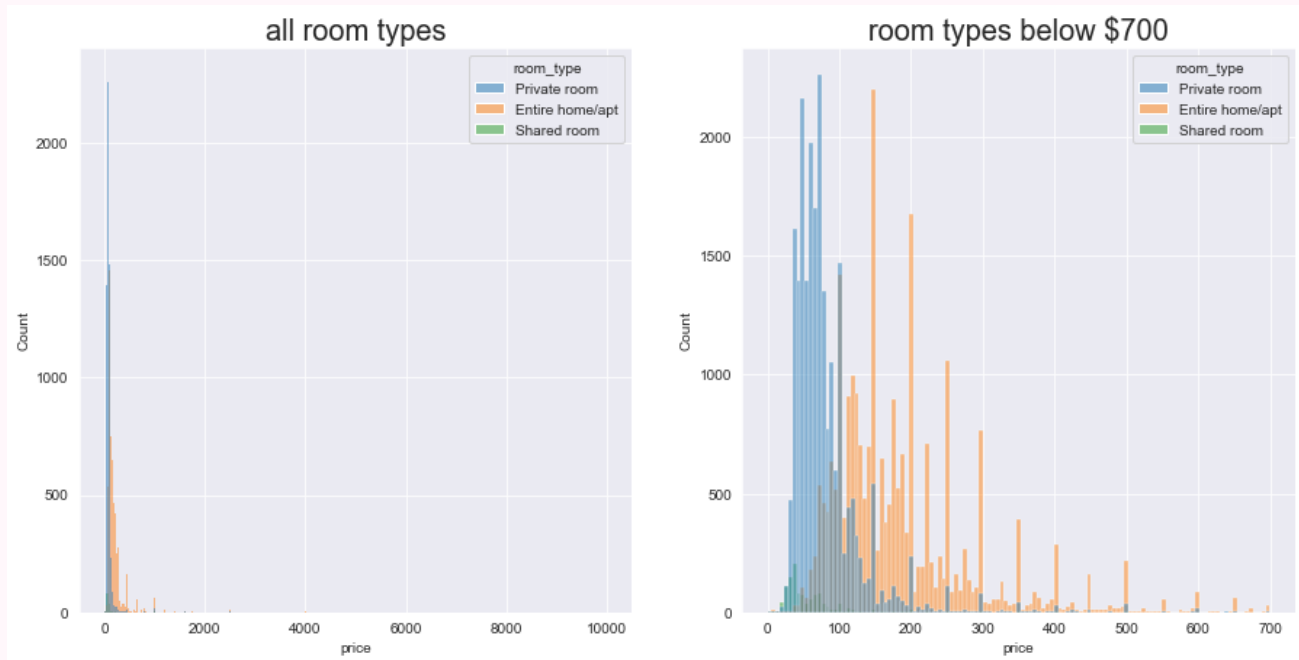
4. Analyzing Methods and Data Distribution :
a. Univariate Analysis:
General Univariate Analysis on Numeri columns. For numeric columns usedcount hist plots

b. Bi-Multivariate Analysis:

Here we first plotted a pair plot of all the numeric columns using seaborn library inPython itself. Below is the code for the same.
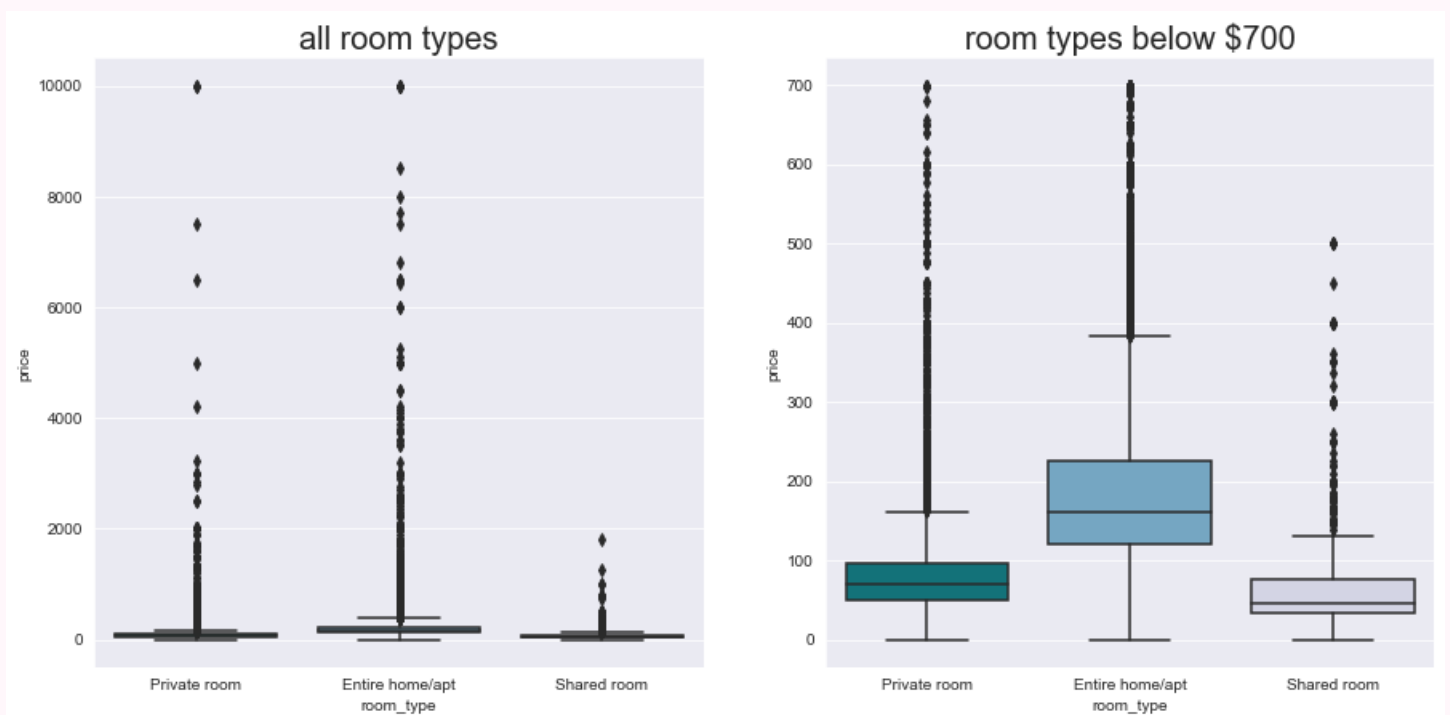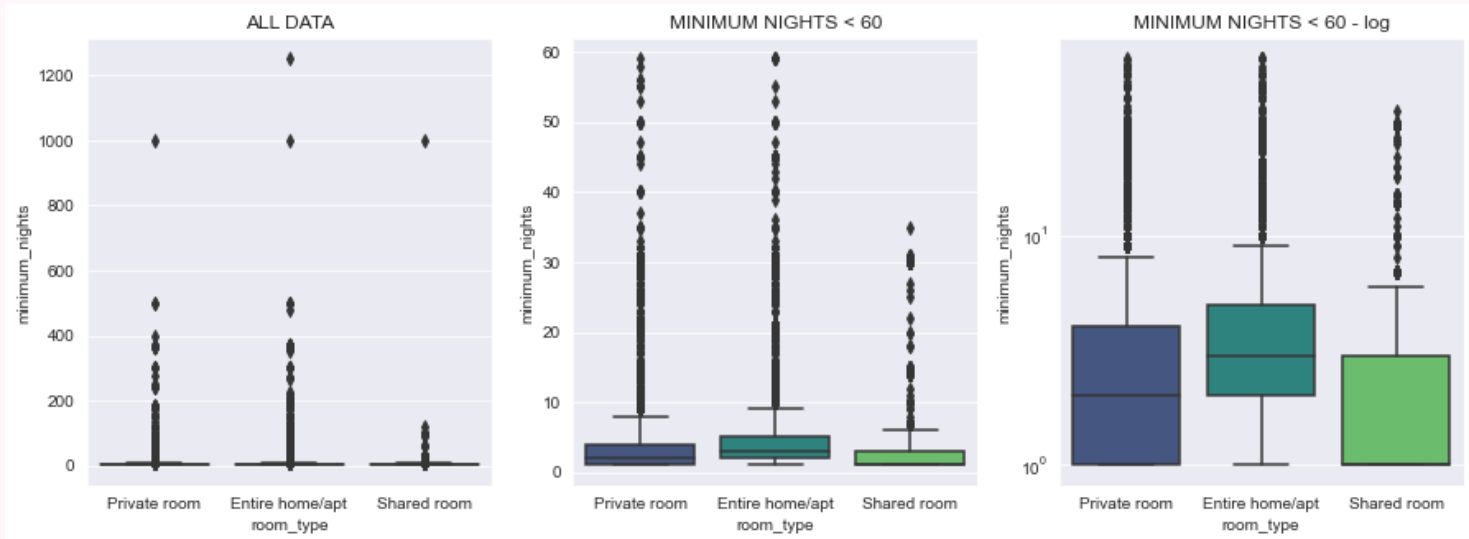
5. Matrix used for Analysis:

In order to measure our analysis, we created a categorical and Numerical Matrix to provide us a direction while creating graphs using different Dimensions and Measures. This matrix involved the values needed to create the graphs. which helps in identifying which all dimensions and measures have been consolidated to get the insights from the data. Below isthe Matrix

1. Categorical & Numerical
2. Categorical & Categorical
3. Numerical & Numerical
4. Numerical & Categorical

6. Evaluation of Methods:

conducted an empirical evaluation of whether doing good or doing in the right way to check at every step by creating relevant questions to see what we are trying toextract from the raw data. its significantly, to extract the relevant information that we want to recommend to our target audience. Above matrix taken as a resource for creating graphs and charts taking insights from them to Answer the following questions.

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

Findings & Insights:

7. 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

data frame.There seems to be many columns with missing values.

8. Visualization Analysis:

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

Caption

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 average of Price and Neighbourhood Group.

2. The trend of count of Number Of Reviews for Last Review Month

Number Of Average Reviews monthwise
Last Review

3. The trend of count of Number of Reviews for Last Review year

Number Of Average Reviews monthwise
Last Review

Caption

The trend of count of Number Of Reviews for Last Review Year. Size shows sum of Number Of Reviews. The marks are labeled by average of Number Of Reviews. The data is filtered on Last Review Month, which excludes Null.
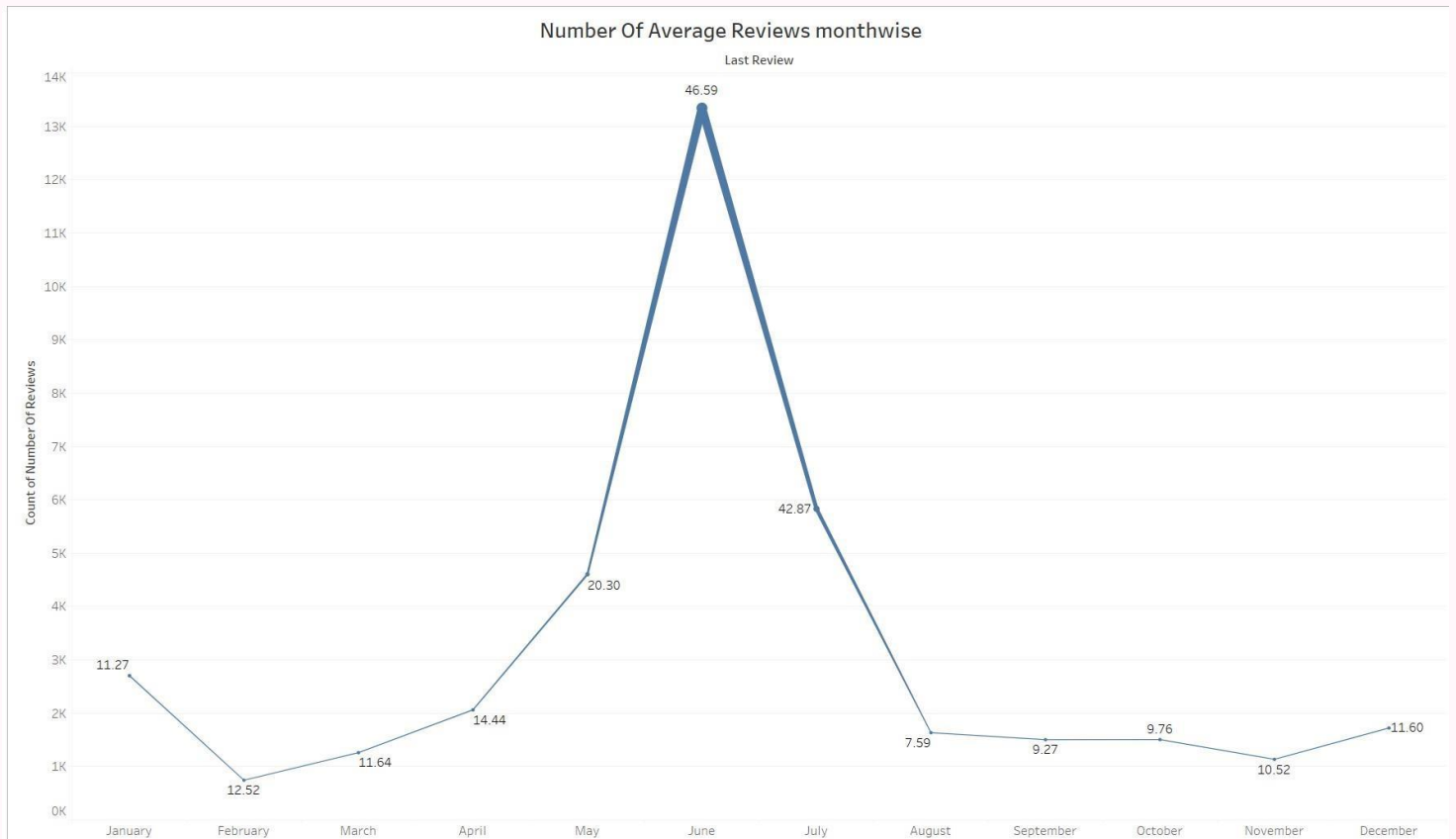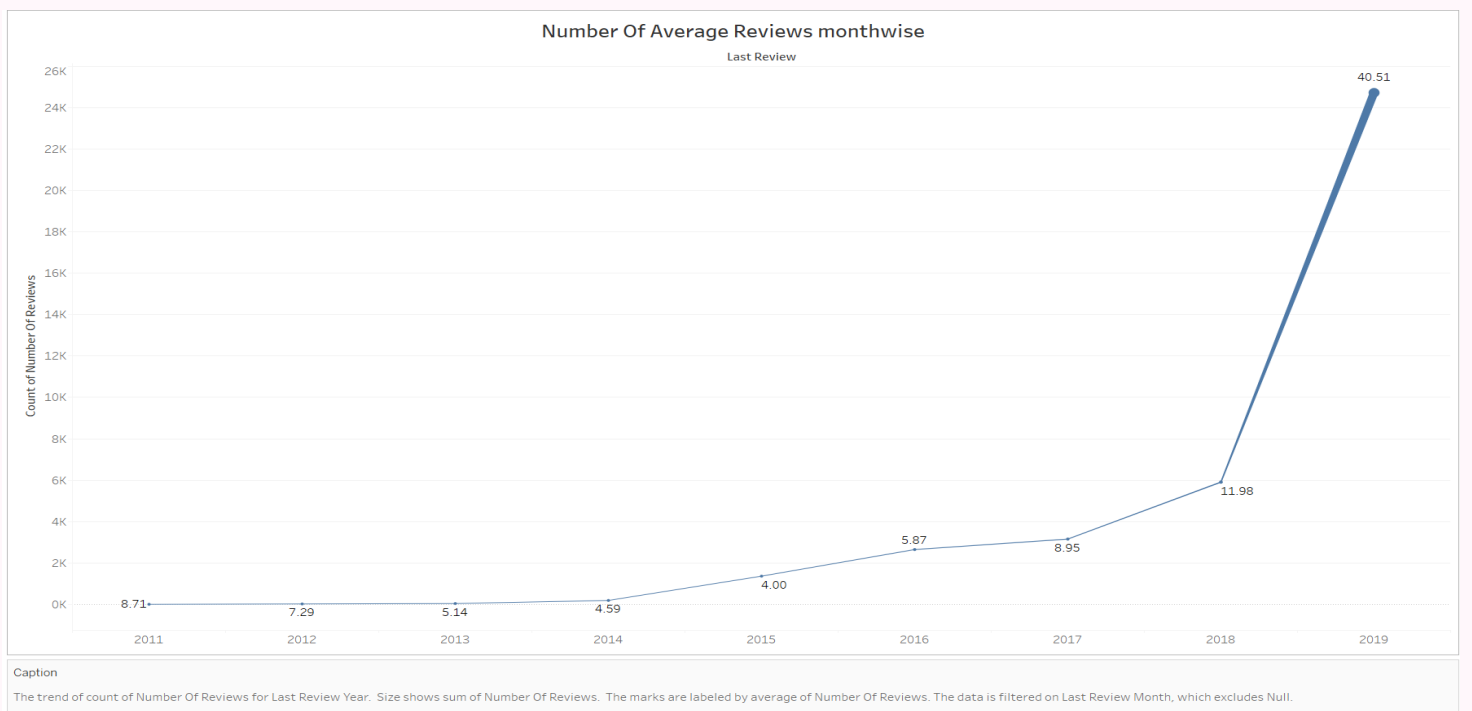
4. Overview on Room Availability on with Room types, colour code shows insights forAverage Price Range

Average price Vs Room type

Room Type

AVG(Price)	
69.3	186.6

Summary	
Count:	3
AVG(Availability 365)	
Sum:	382.07
Average:	127.36
Minimum:	108.94
Maximum:	162.58
Median:	110.55
AVG(Price)	
Sum:	339.1
Average:	113.0
Minimum:	69.3
Maximum:	186.6
Median:	83.3

Caption

Average of Availability 365 for each Room Type. Color shows average of Price.
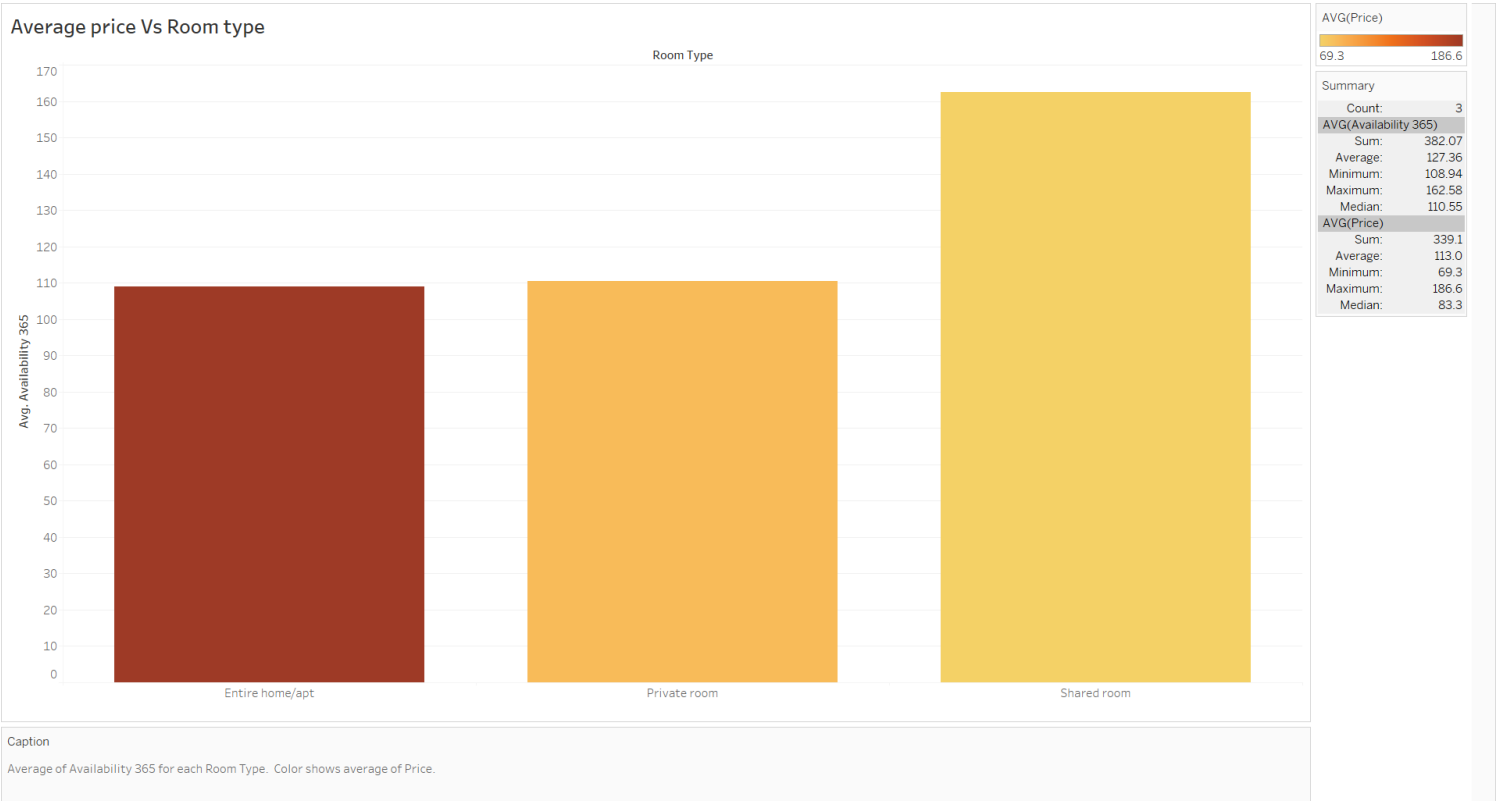
5. Overview on Hosts vs average Price, color code shows insights for Average Price Rangeon different Neighborhood groups

Average Price Vs Hostlisting range

Host lisiting scale

AVG(Price)	
51.7	174.5

Caption

Average of Price for each Host listing scale. Color shows average of Price. The marks are labeled by Neighbourhood Group and average of Price.
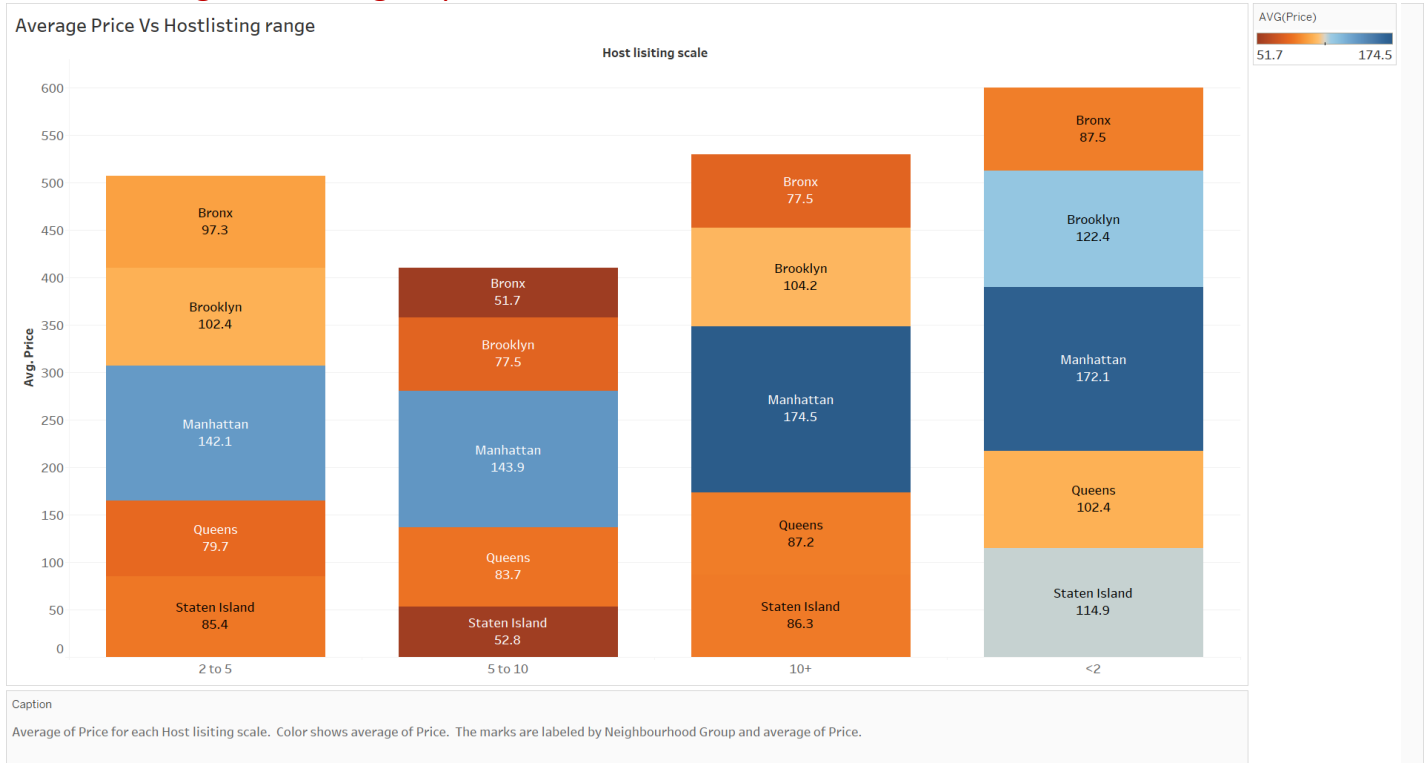
6. Overview on Availability vs average Price, color code shows insights for Average Price Range on different Neighborhood groups

Availability365 Vs Avg.Price

Availability 365 (bin)

Caption

Average of Price for each Availability 365 (bin). Color shows details about Neighbourhood Group. The marks are labeled by Neighbourhood Group and average of Price. The view is filtered on Availability 365 (bin), which keeps 0, 100, 200 and 300.
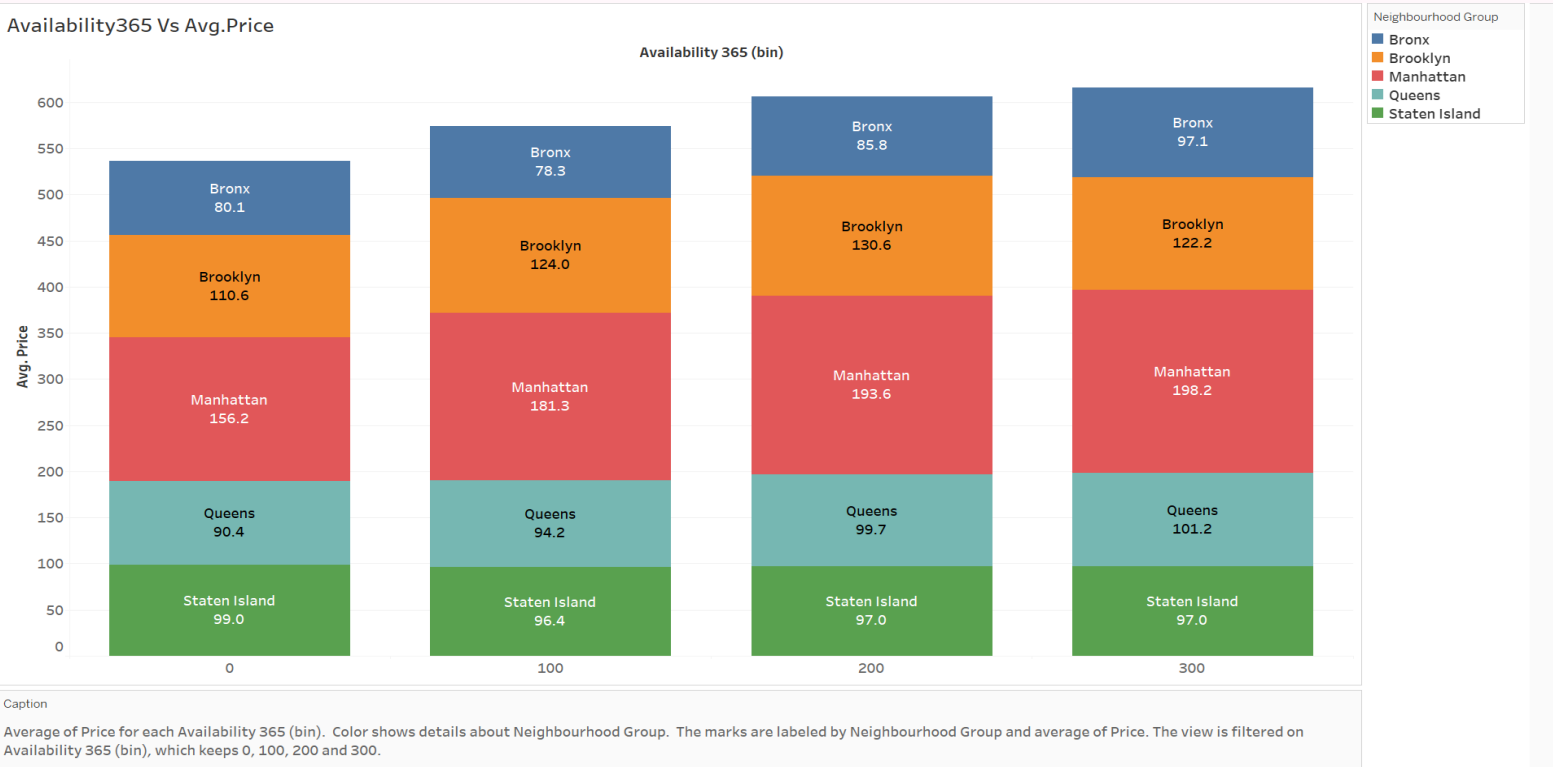
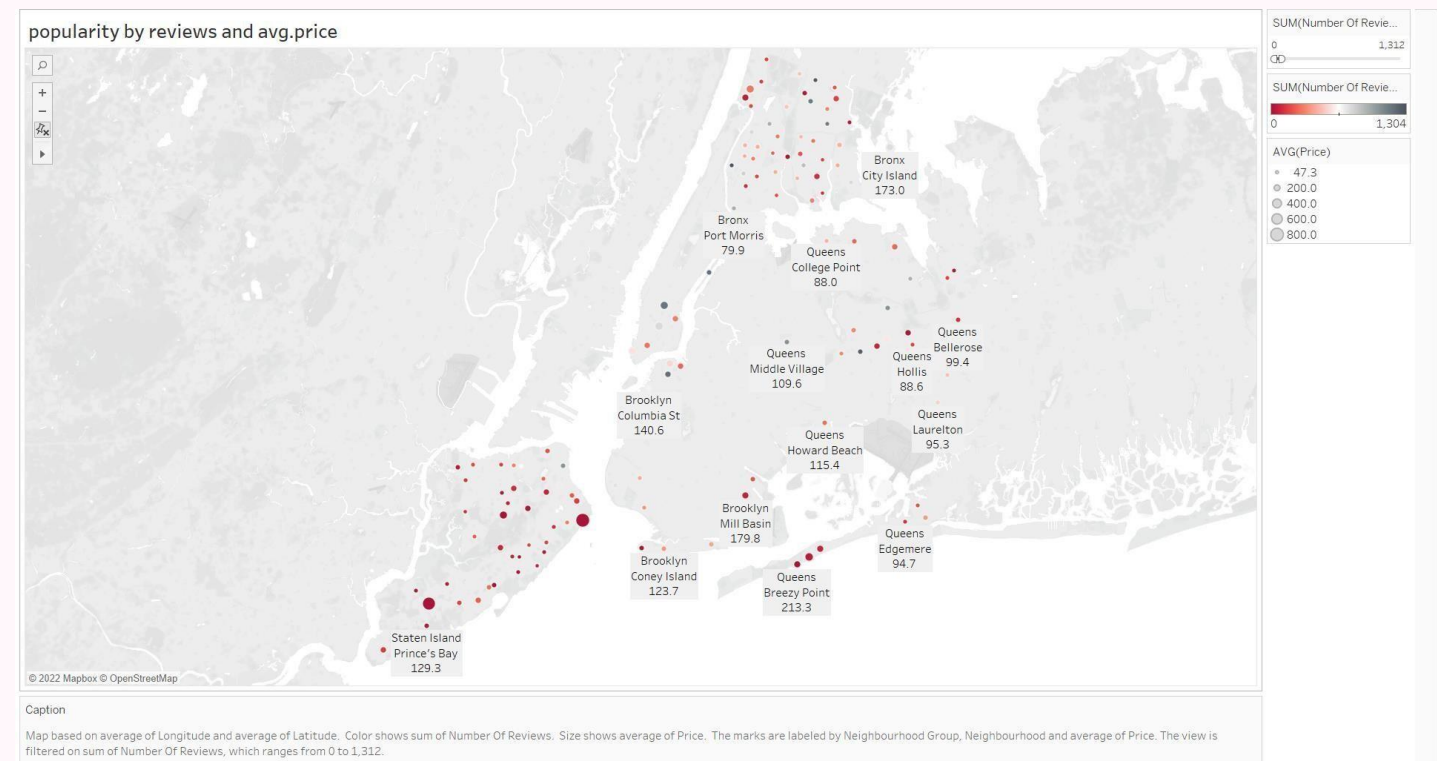## 7. Overview on Number of Reviews vs Availability, colour code shows insights for count of different Neighbourhood groups with average price

Availability Vs Avg Reviews

Availability 365 (bin)

Caption

Average of Number Of Reviews for each Availability 365 (bin). Color shows details about Neighbourhood Group. The marks are labeled by count of Neighbourhood Group, Neighbourhood Group and average of Price.
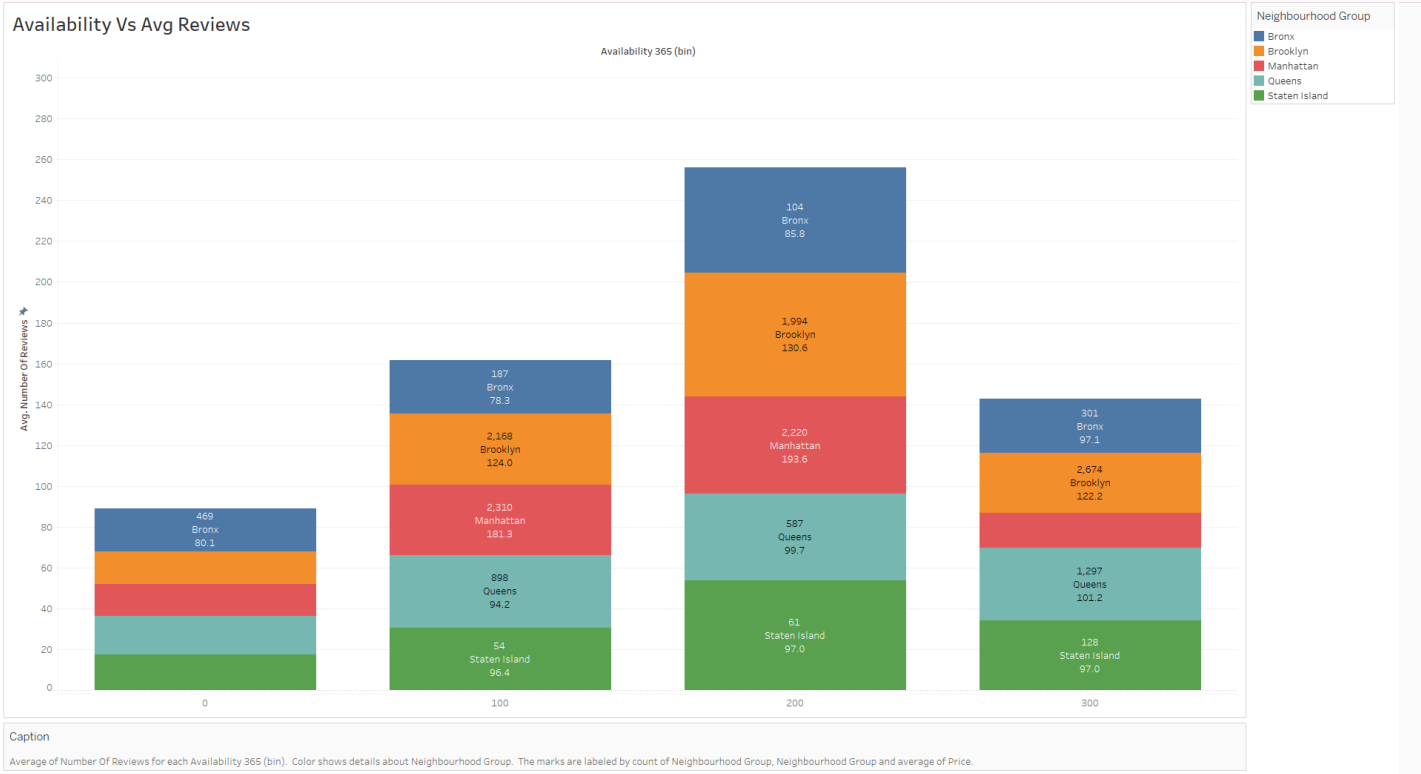
8. Overview on Top 20 Hosts by Reviews, colour code shows insights for sum of reviews

Top 20 Host by Reviews

Host Name	

Limit
Top 20 by SUM([Number Of Reviews])

SUM(Number Of Revie...

3,781   10,634

Caption

Host Name and sum of Number Of Reviews. Color shows sum of Number Of Reviews. Size shows sum of Number Of Reviews. The marks are labeled by Host Name and sum of Number Of Reviews. The view is filtered on Host Name, which keeps 20 of 11,249 members.

Number of reviews Vs Average p...   Neighbourhood Avg Price   Average price Vs Room type   Average Price Vs Hostlisting ran...   Availability Vs Avg.Price   Availability Vs Avg Reviews   **Top 20 Host by Reviews**   popularity by reviews and avg.pr...   avg
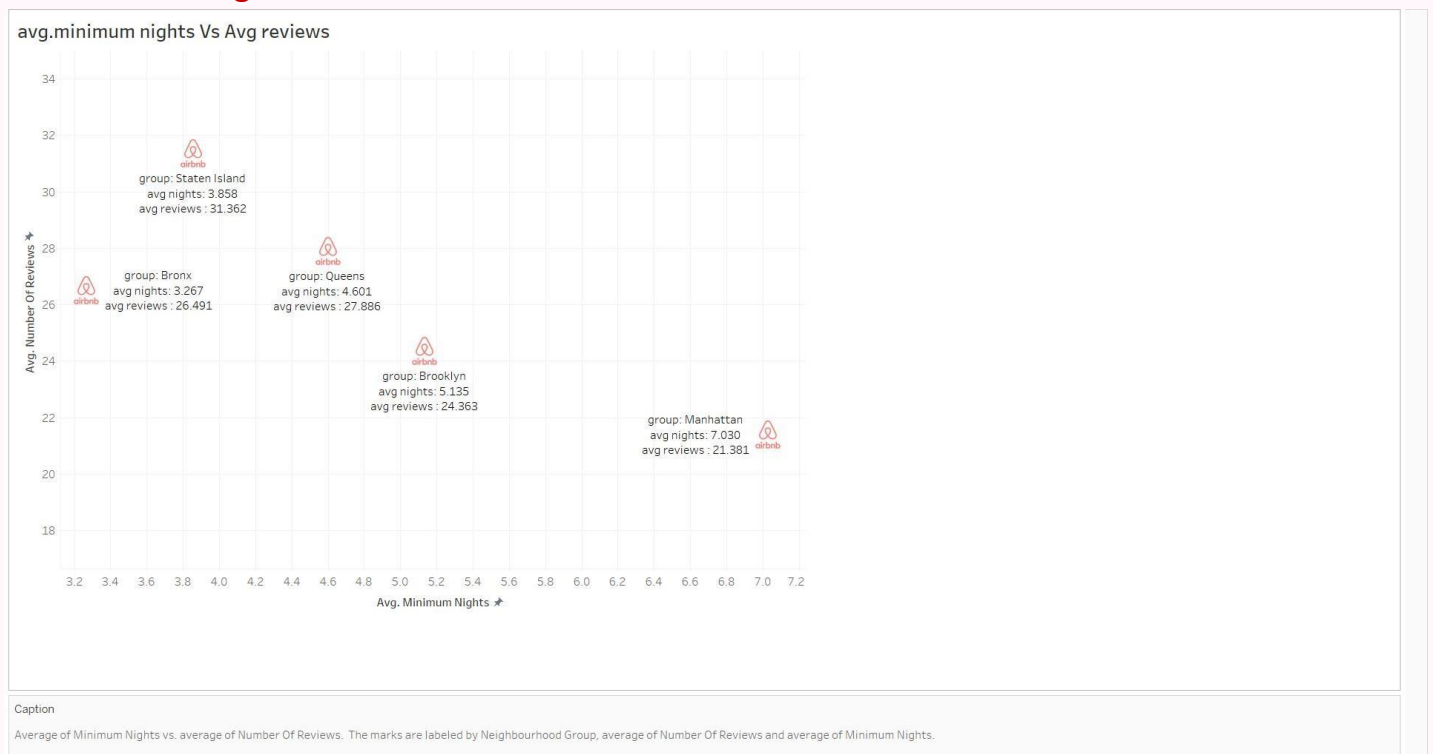
9. Map Showing the Location most reviewed neighbourhood groups and size showing the estimation of average price range

popularity by reviews and avg.price

SUM(Number Of Revie...
0   1,312

SUM(Number Of Revie...
0   1,304

AVG(Price)
· 47.3
○ 200.0
○ 400.0
○ 600.0
○ 800.0

© 2022 Mapbox © OpenStreetMap

Caption

Map based on average of Longitude and average of Latitude. Color shows sum of Number Of Reviews. Size shows average of Price. The marks are labeled by Neighbourhood Group, Neighbourhood and average of Price. The view is filtered on sum of Number Of Reviews, which ranges from 0 to 1,312.
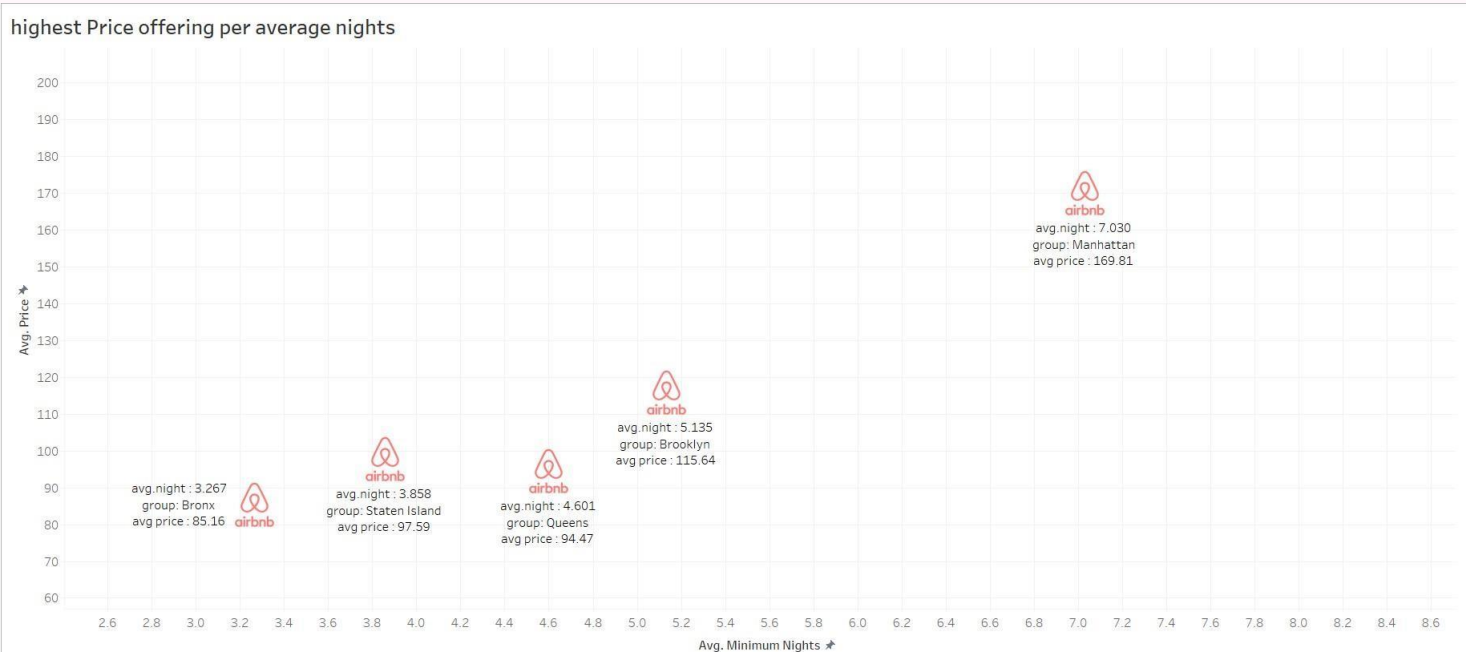
10.Number of Reviews Vs Minimum Nights Graph Showing the Number Review takenwith minimum night stay
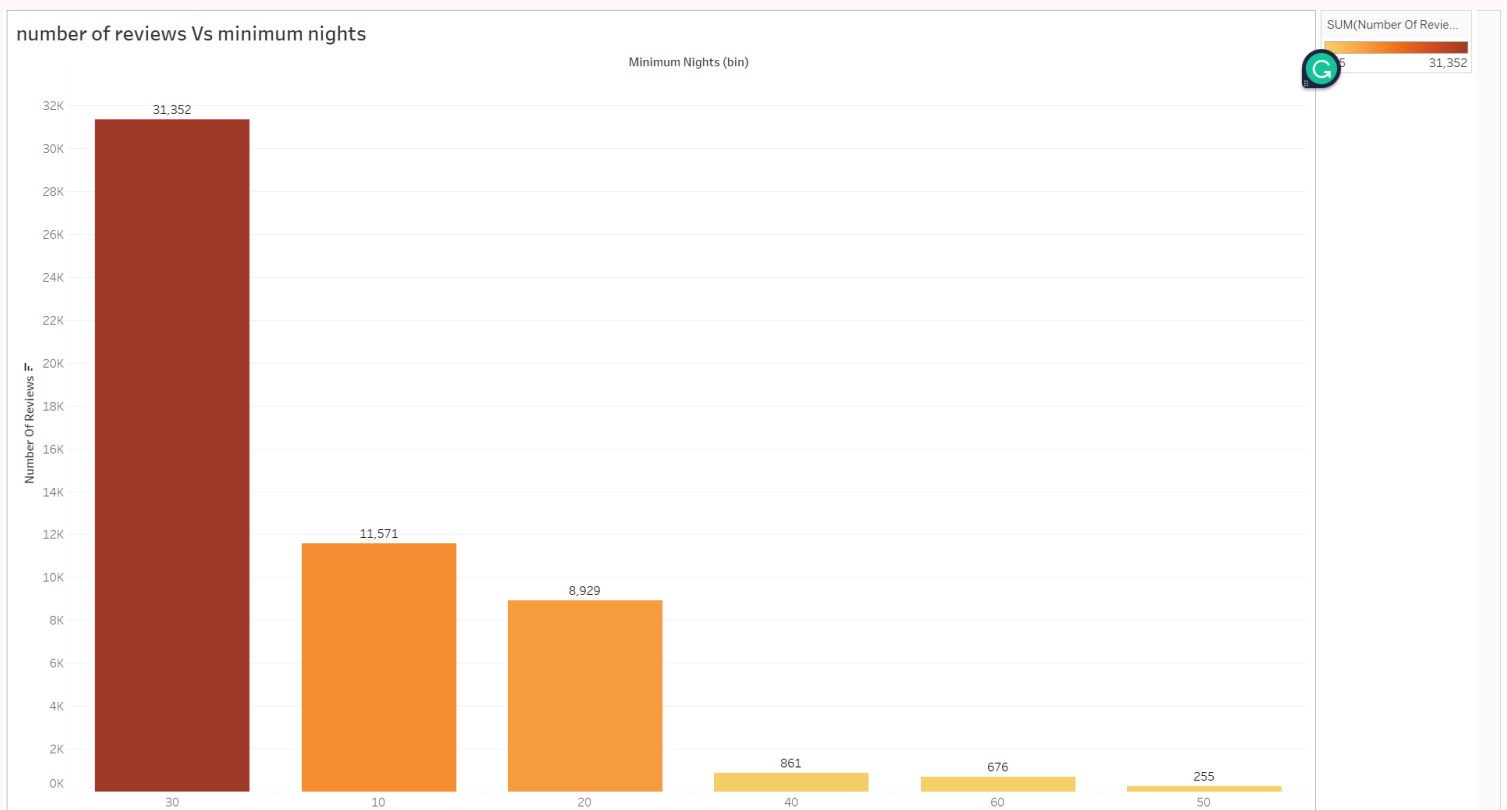
11.Number Reviews taken by the Neighbourhood with average minimum Stay and averageReviews

Caption

Average of Minimum Nights vs. average of Number Of Reviews. The marks are labeled by Neighbourhood Group, average of Number Of Reviews and average of Minimum Nights.

12. Average Price charges by the Neighborhood with minimum nights

highest Price offering per average nights

Avg. Price

avg.night : 3.267
group: Bronx
avg price : 85.16

avg.night : 3.858
group: Staten Island
avg price : 97.59

avg.night : 4.601
group: Queens
avg price : 94.47

avg.night : 5.135
group: Brooklyn
avg price : 115.64

avg.night : 7.030
group: Manhattan
avg price : 169.81

Avg. Minimum Nights

Caption

Average of Minimum Nights vs. average of Price. The marks are labeled by average of Minimum Nights, average of Price and Neighbourhood Group. Details are shown for Neighbourhood Group.

9. Key Findings from the Above Graphs:

- The marks are labelled by average of Price and Neighbourhood Group, Manhattan is having a highest Average Price of $169.81 , followed by the Broklyn and Manhattan are dominating when it comes to listed hostings followed by Queens.
- Most of the Reviews are received In the month of June with average 46.69, from themonth May to July most of the Reviews are received.
- Initial years which were 2011 till 2014, last reviews were negligible. After that it is slowlygoing up and most of the last reviews are received in the recent years of the data that is2019 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 bythe 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 badcustomer experience offered by majority sites.
- Also, Majority of the sites have received less than 50 reviews till date which is kind of lessas 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 lessShared rooms.

Conclusion:

Important Insights from the Data:

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

2) Highest Price range Manhattan is the only Neighbourhood in the Borough that lies in offering the Highest Price range of properties on the platform followed by others with a Medium Price range on average. Prices offered above 120$ on average are considered to bea High Price, between 80$ and 120$, Medium Price range and less than 80$ to be considered Low Price range properties.

3) Having a high price range, Entire home/apt types rooms are available for less than 100 dayson 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 hostwith an average price of 230$ followed by Brooklyn with an average price of 108$. On theother hand, all the hosts have less than 2 properties listed in either of the Borough on anaverage price range between 80$ and 170 $.

5) Brooklyn has received the highest number of reviews based on the availability to stay openfor 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 atall 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 mostof them have provided a good number of reviews within 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 a high pricerange.

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 ranges that have received the highest number of reviews on average being the lowest in the Price range. On the contrary, Queens - Neposit, Manhattan - NoHo, Manhattan - Tribeca, Staten Island - Willowbrook and Manhattan - Flatiron District is thehighest in the Price range and have received a low number of reviews.

9) "WELCOME TO BROOKLYN" PARKSIDE VIEW STUDIO APT", "Oasis on the Park", "HELLO BROOKLYN" PARKSIDE VIEW NEWLY RENO APT", "Comfy Home Away From Home/Multiple rooms", "LOVE BROOKLYN" Newly Renovated Studio APT" and

"CozyRetreat" in North Crown Heights" are the Top 6 listed places that have received the highest number of reviews.

10) On average Entire home/apt types are preferred more by the customers followed by Private rooms and then Shared Rooms. Mostly because they are also available for a highernumber 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 bookmaking Manhattan and Brooklyn thetop neighborhoods in offering maximum minimum night's stay.

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

14) 5766 properties are open for more than 300 days a year. Around 2286 of them are from Brooklyn followed by Manhattan 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.