**Methodology Document Of AirBnB Data Story Telling Case Study**

Submitted by,

Hetal Khanapure

**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: Data Analysis Managers &Lead Analysts & Head of Acquisitions and Operations, NYC.

This document is comprised of various steps involved to come up with **the insights and Recommendations** to help Business to take important decisions for the successful running of the business, starting from **Data cleaning, Exploratory Data Analysis and Data Visualisations.**

**Problem Statement:**

Suppose that you are working as a data analyst at Airbnb. 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.

**Tools Used:**

We used tools such as Python Jupiter Notebook,Tableau and Microsoft Excel mainly to perform various Data Transformations and Data Analysis to come up with insights and Recommendations.

**Assumptions:**

* As the dataset does not talk about the Number of Booking made per property, we assume that number of reviews column as Number of Booking made.
* Estimated Revenue is calculated based on Number of reviews and price column as per the assumption made in the point 1.
* We assumed the data prior to the COVID – 19 period was achieving the desired revenue.
* The company’s strategies are decided considering the travel will be increased in the post COVID period.
* We assume that if availability 365 is 0 , means that property is fully booked or delisted

**Derived columns:**

* We have created calculated fields in Tableau to find Total Revenue using price and Number of Reviews (assumed Number of Reviews = Number of bookings
* Price Range: Created bin for price column for the better categorization.
* Created calculated column for Availibility\_365 = 0 to understand how many properties are delisted.

**Steps:**

The various steps involved in the Data Analysis are

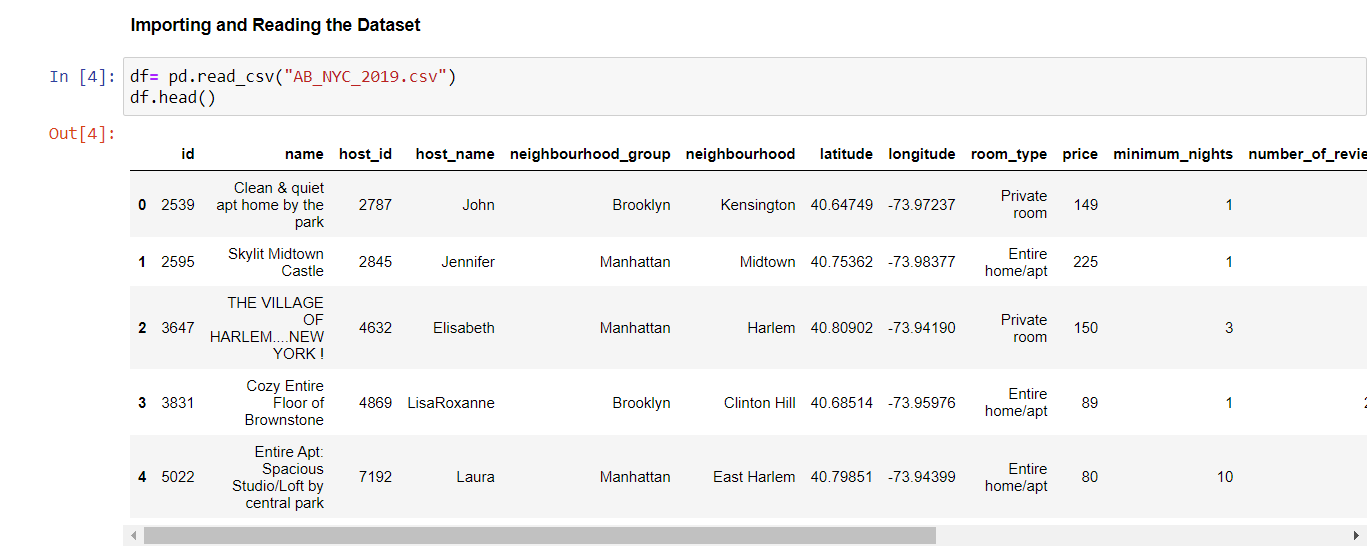
1. Importing the data into Python dataframe using Pandas Library.
2. Inspecting and Cleaning the data: Missing Value treatment, Outlier Treatment.
3. Exploratory Data Analysis
4. Data Visualization
5. Insights and Recommendations

Let’s see everything in detail with screen shots:

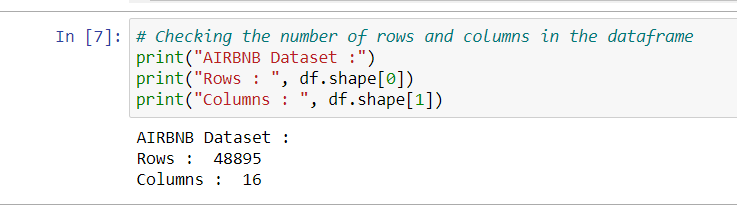
1. Importing the data into Pandas Dataframe:



1. Importing and Reading the Dataset:

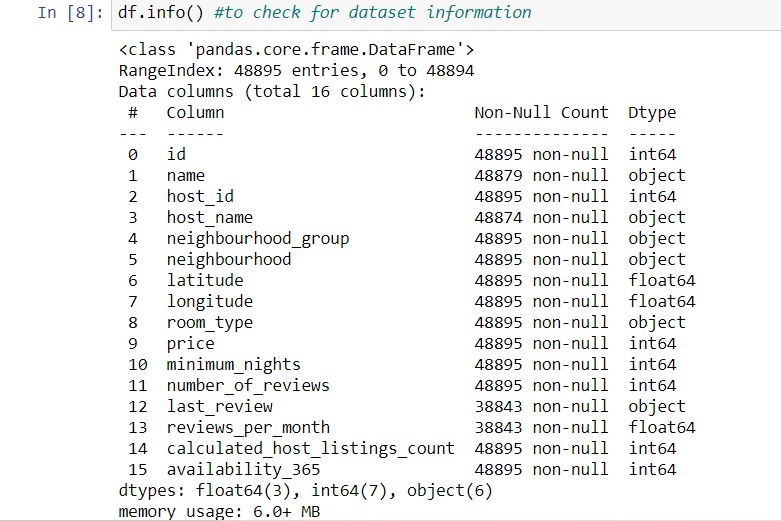


Verifying the number of rows and columns present in the data.

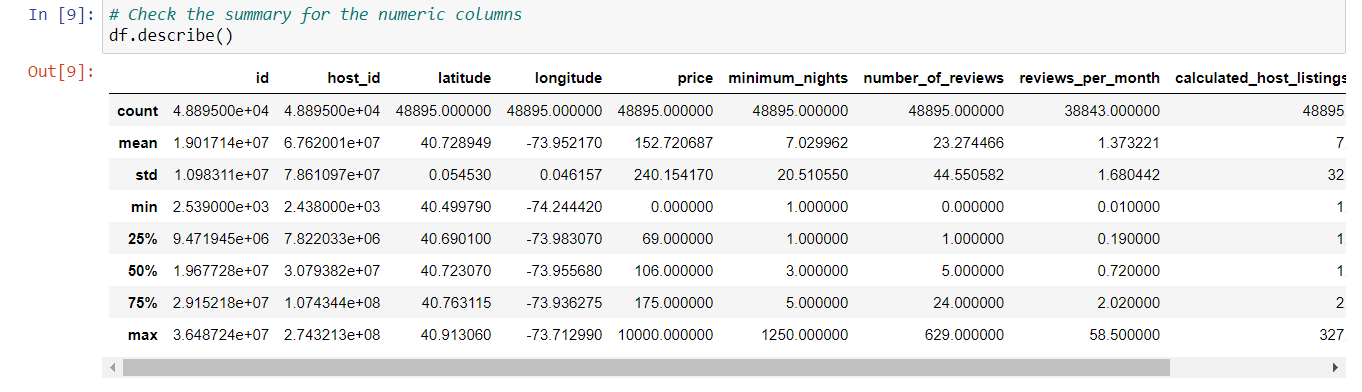


We can see there are 48895 rows and 16 columns present in the dataset.

Now checking datatypes and columns names using info()



Next checking numerical column details using describe ()



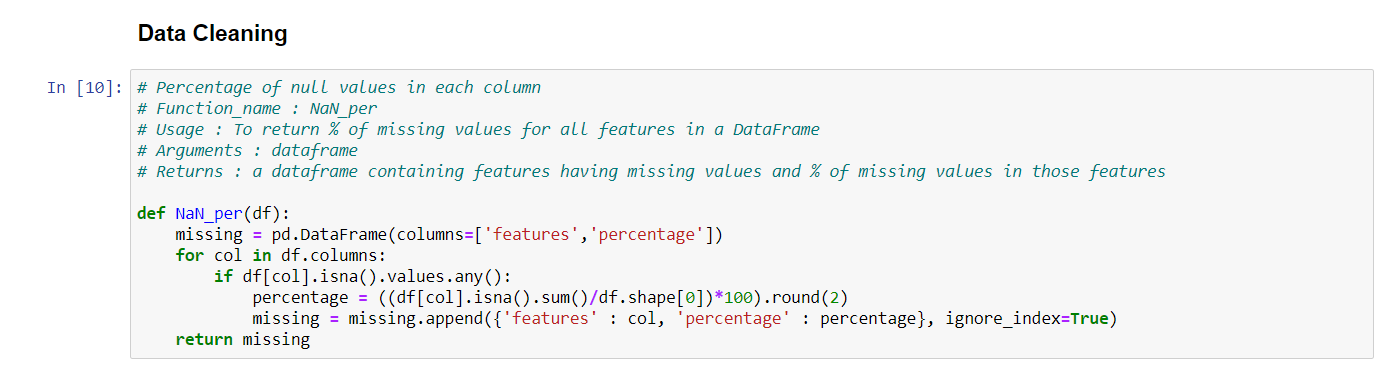
Using describe (), we can know the min, max , count,mean,id and different quantiles.

**Note**: ID, Host\_id columns are required for Data Analysis purpose. As we are not doing any Machine Learning Techniques here, we keep these columns.

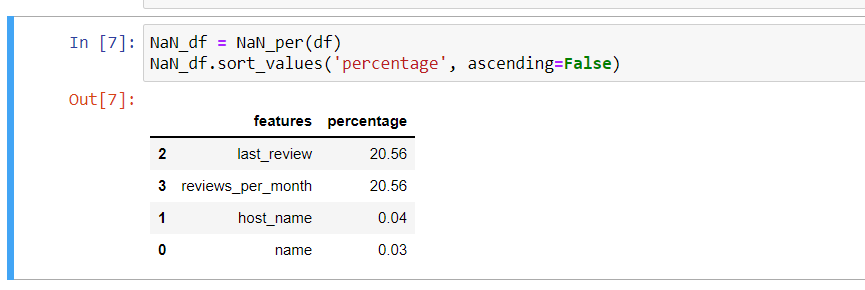
1. Data Cleaning:

* Missing values

Below code snippet is a function defined for retrieving columns having missing values and its percentage.



Below are the results:



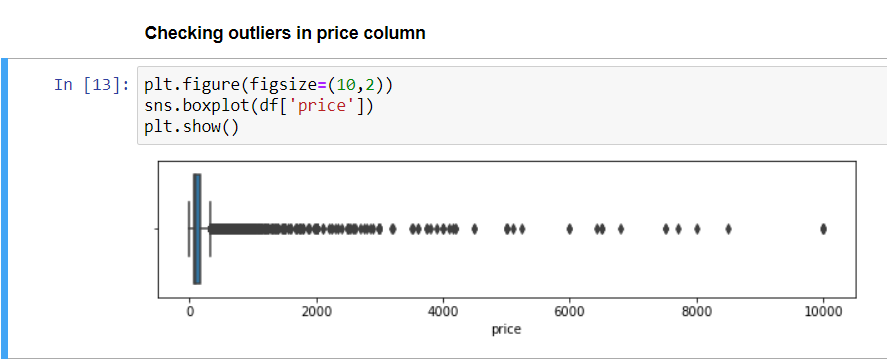
**Observation**: As we can see in the results, last\_review and reviews\_per\_month columns are having 20.56% of missing values with respect to total data present.

**Note**: Missing value Imputations are not being performed here as we are only performing Data Analysis and getting insights. So missing values will not affect our objective and we will keep them as it is.

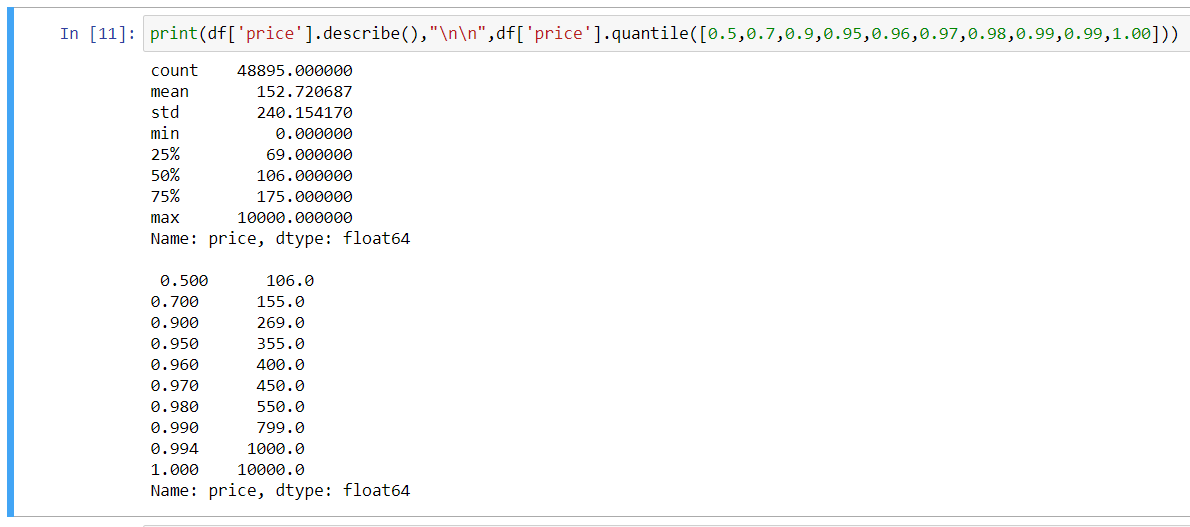
* Outlier Analysis:

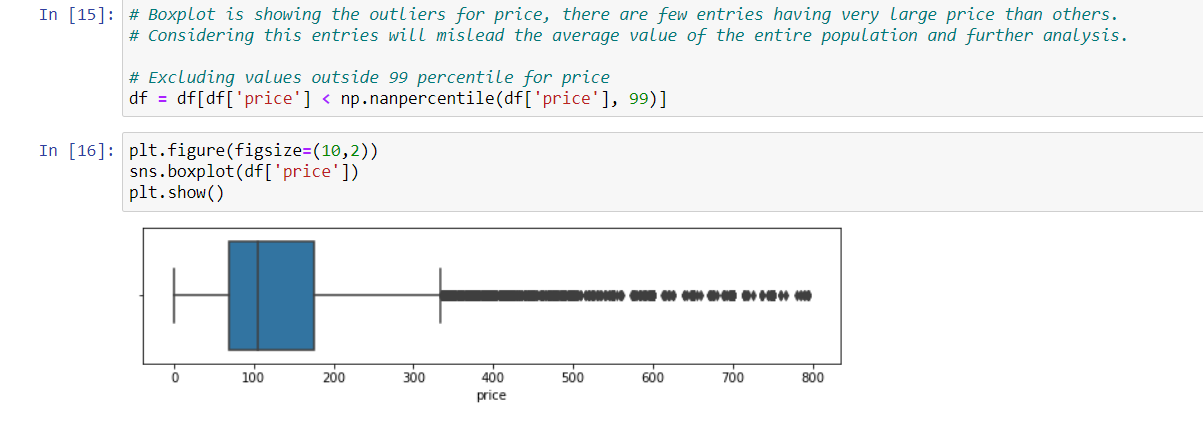
We perform Outlier Analysis for all the numerical columns and check if required to exclude any outliers.

**price** column:



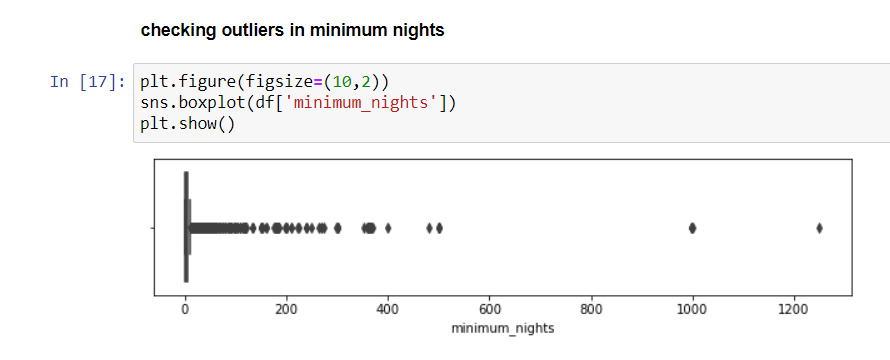
Checking various quantiles:

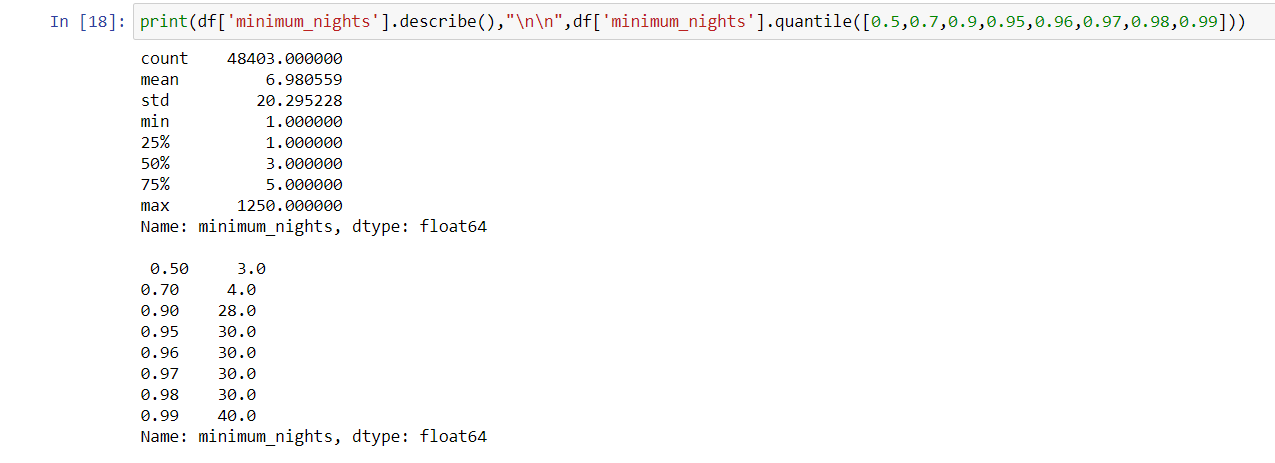


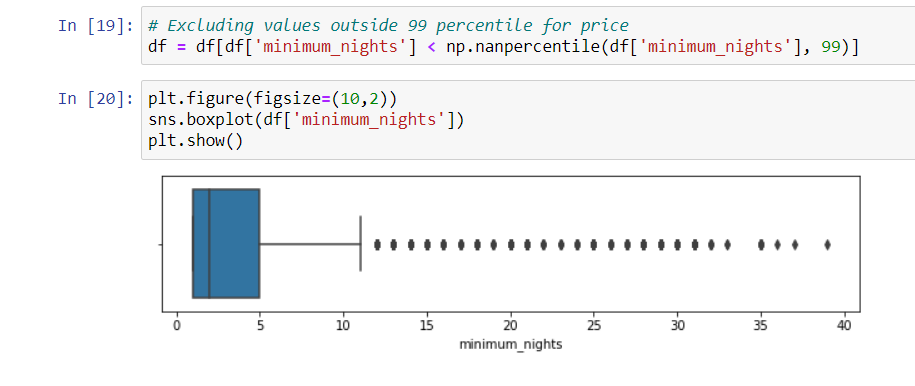


**Observation**: We can see there are outliers in price column. We are considering taking the entries till 0.99 quantiles and the values are continuous and it may affect the analysis when taking the average price. Again, as the data Analysis is concerned, we may need to analyze outlier cases also and study how it affects the business. We plan to exclude the outlier for price column at certain places and perform analysis with Outlier too.

**minimum\_nights** column:

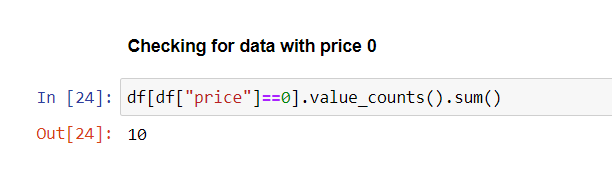
****

****

****

**Observation**: We can see there are outliers in minimum\_nights columns. We are considering taking the entries till 0.99 quantiles and the values are continuous and it may affect the analysis when taking the average minimum nights. Again, as the data Analysis is concerned, we may need to analyze outlier cases also and study how it affects the business. We plan to exclude the outlier for minimum\_nights column at certain places and perform analysis with Outlier too.

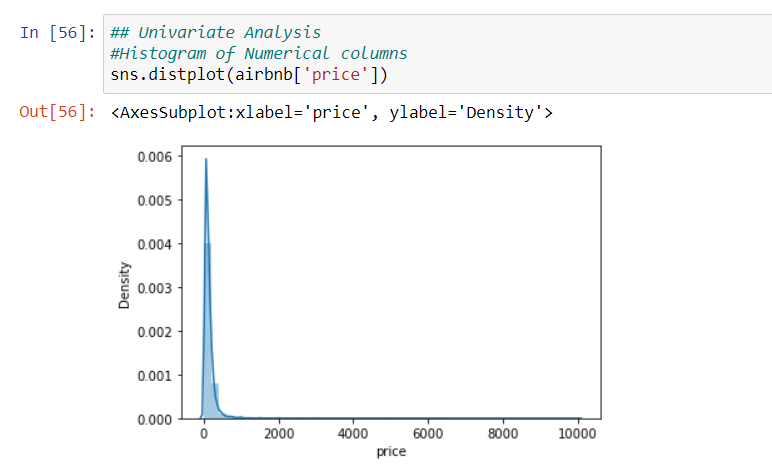
Price column having 0 is field values:



Observation: There are 10 entries having price column as 0.

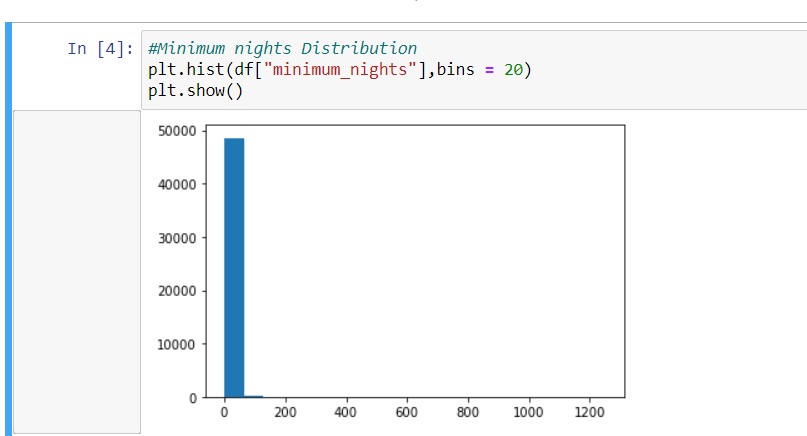
1. **Exploratory Data Analysis:**

* **Univariate Analysis:**

****

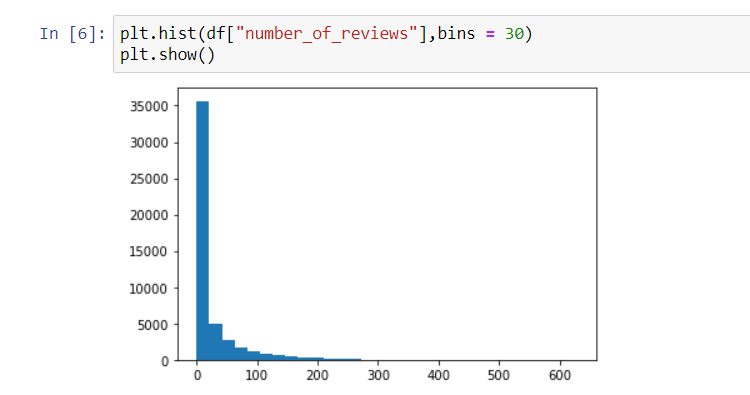
***Observation:*** *The count of price varies from 0 to 10000 being from 1500 to 10000 as outliers. We can see there the price ranges from 0 to 1000 got the highest number / density and there are not much counts for outlier and very minimal. So we may want exclude the outlier and perform the analysis.*

* minimum\_nights Distribution:

**

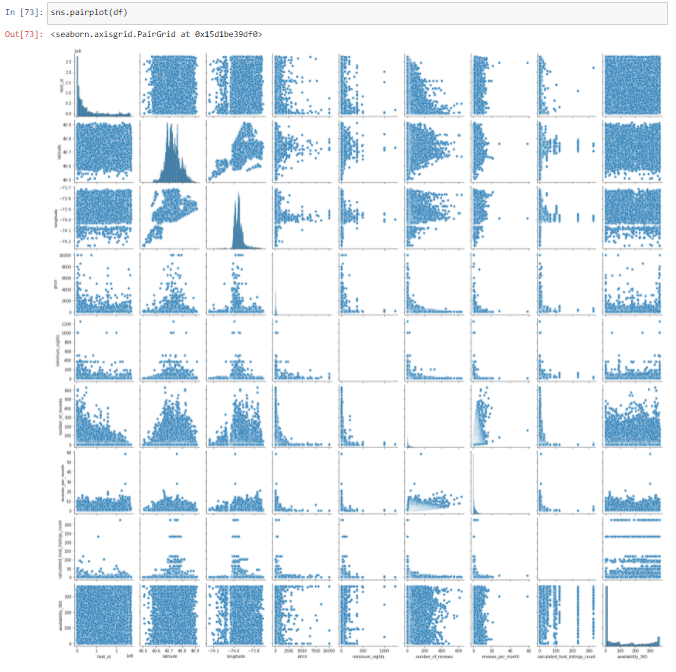
***Observation****: Majority of the data points fall under 0 to 5 minimum\_nigh stay.*

* Number of reviews distribution:

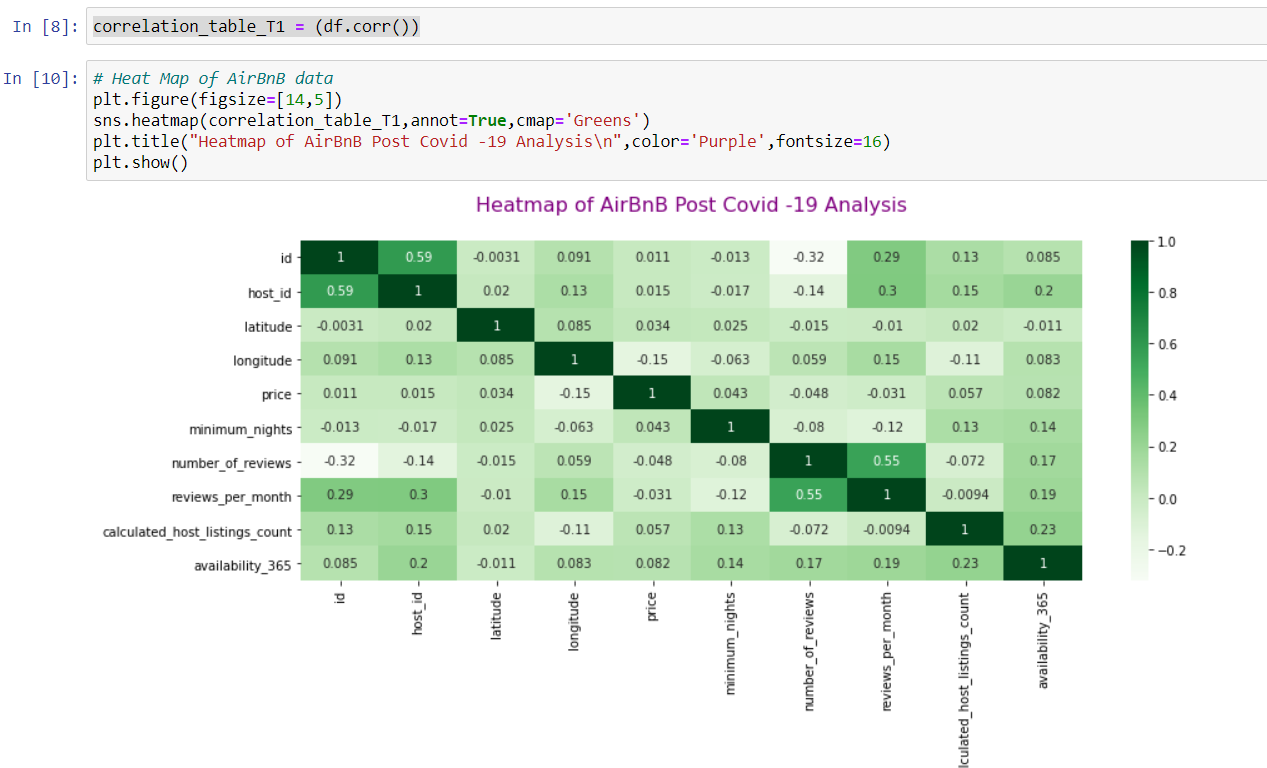


***Observation****: The data points fall under 0 to 300 and data points fall under 0 – 20 have the highest number of reviews.*

**Bivariate Analysis:**

* Pair plot of the data frame to understand the correlation between the numerical columns.
* 

Heatmap of AirBnB data:

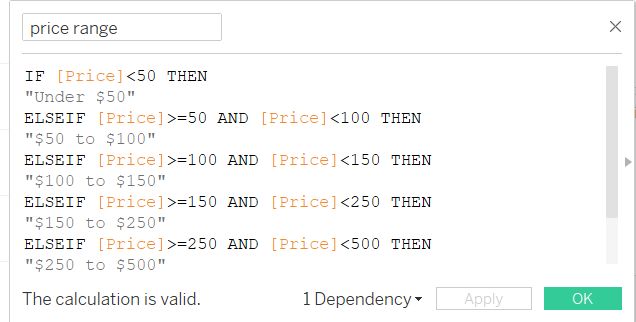


***Observations****: From the pair plot and correlation matrix we can know there is a negative correlation between price, minimum nights and Number of reviews. We can see there is a positive correlation between Calculated\_host\_listings\_count and minimum\_nights & availability \_365 columns.*

1. **Data Visualization and Analysis using Tableau:**

* Calculated Fields/Derived columns created in Tableau:

Price range : Code snippet as follows:



* Availability\_zero:



* Estimated Revenue:

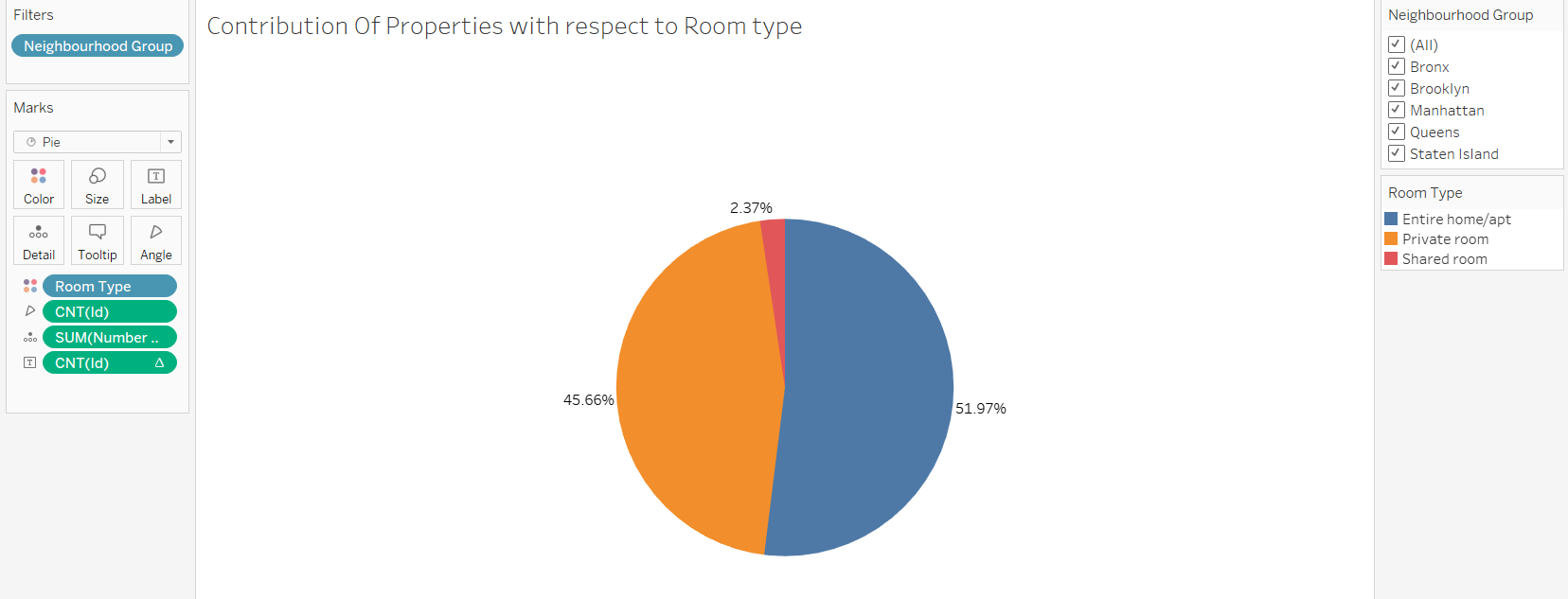


* Minimum nights excluded outlier:



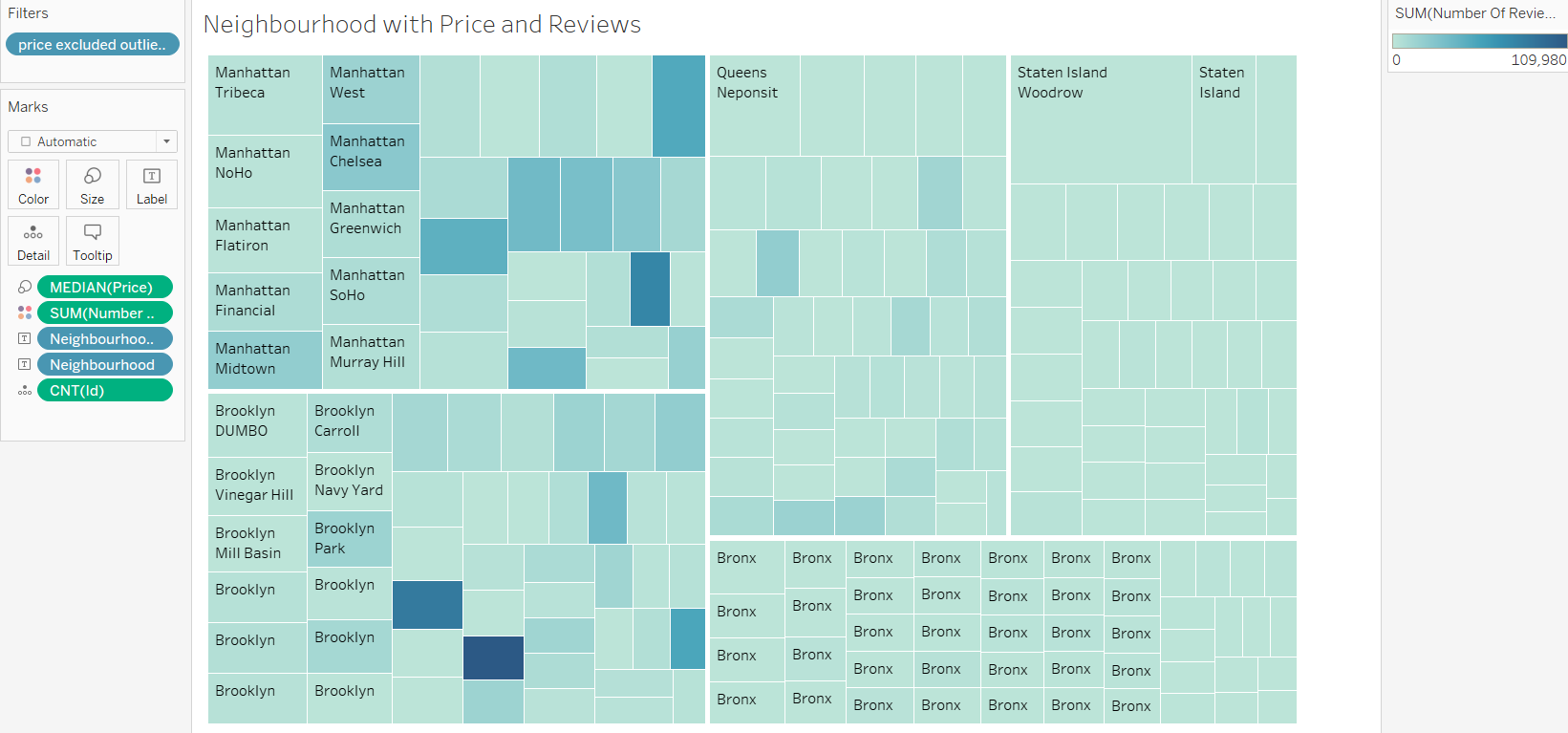
We will use Tableau for Data Visualization and Analysis to come up with Insights and observations .Recommendations are made from the insights and observations drawn from the Analysis.

* Below chart shows contribution of Properties based on Neighbourhood Group and Room Type.



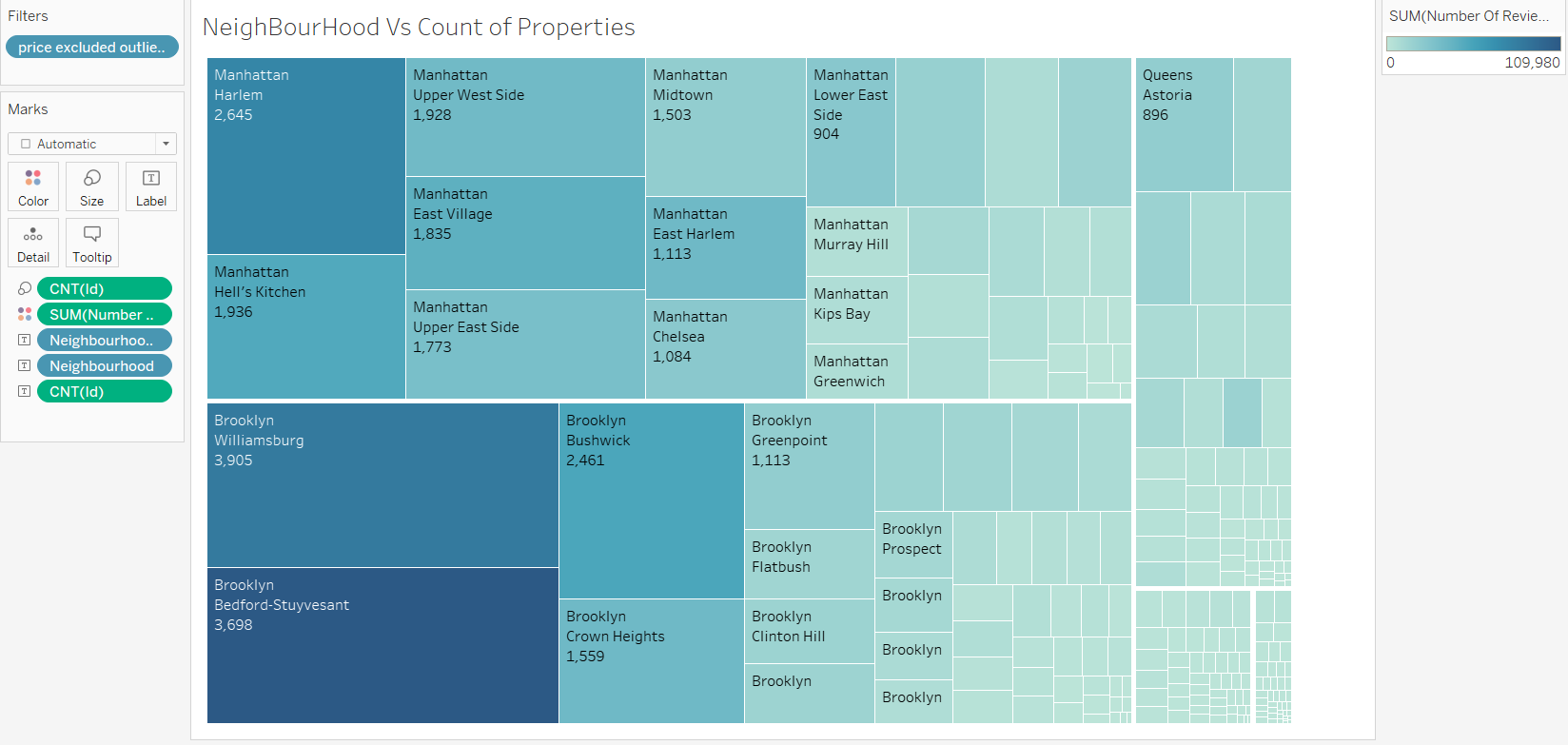
***Observation/Insights****:* *About 96% of the properties are based on Entire Home/apt. Very less people use Shared room across all the Neighbourhood. Another interesting insight is all the Neighbourhood except Manhattan has high number of Private rooms whereas in Manhattan majority of rooms are of Type Entire home/apartment.*

* Neighbourhood with Median Price and Number of Reviews



***Observation/Insights****: The general trend is booking is majorly focused on Price. Here number of reviews (booking) is more for lower priced properties. We can see a contrary trend also for some properties with high price got more bookings too. So some people do not mind price if good facilities may be offered.*

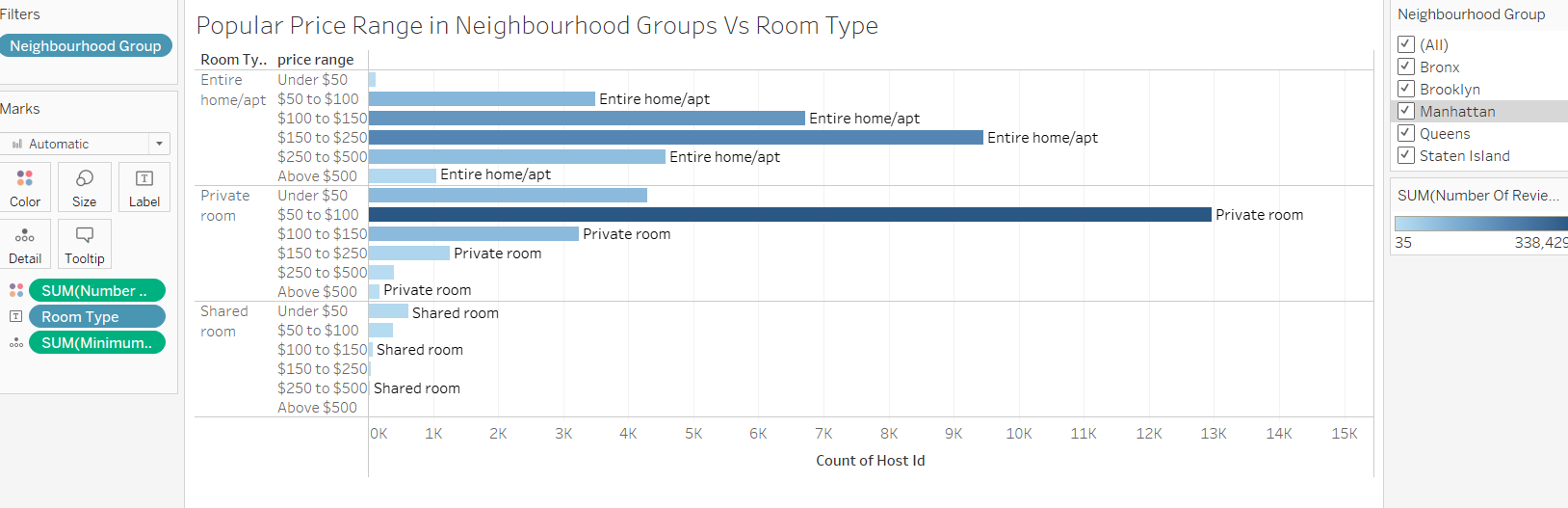
* Neighbourhood vs Count of Properties

**

***Observation /Insights:*** *We can see a relation between Number of Reviews(booking) and number of properties. Manhattan and Brooklyn has got the highest number of bookings where there are highest properties lie. Queens also have a reasonable number of properties and bookings whereas Bronx and Staten lsland has got less number of bookings where there are very less number of properties.*

*The* ***insight*** *is when properties are increased, bookings also increase. Bronx and Island are the locations where business can leverage more on adding the properties to have a higher revenue.*

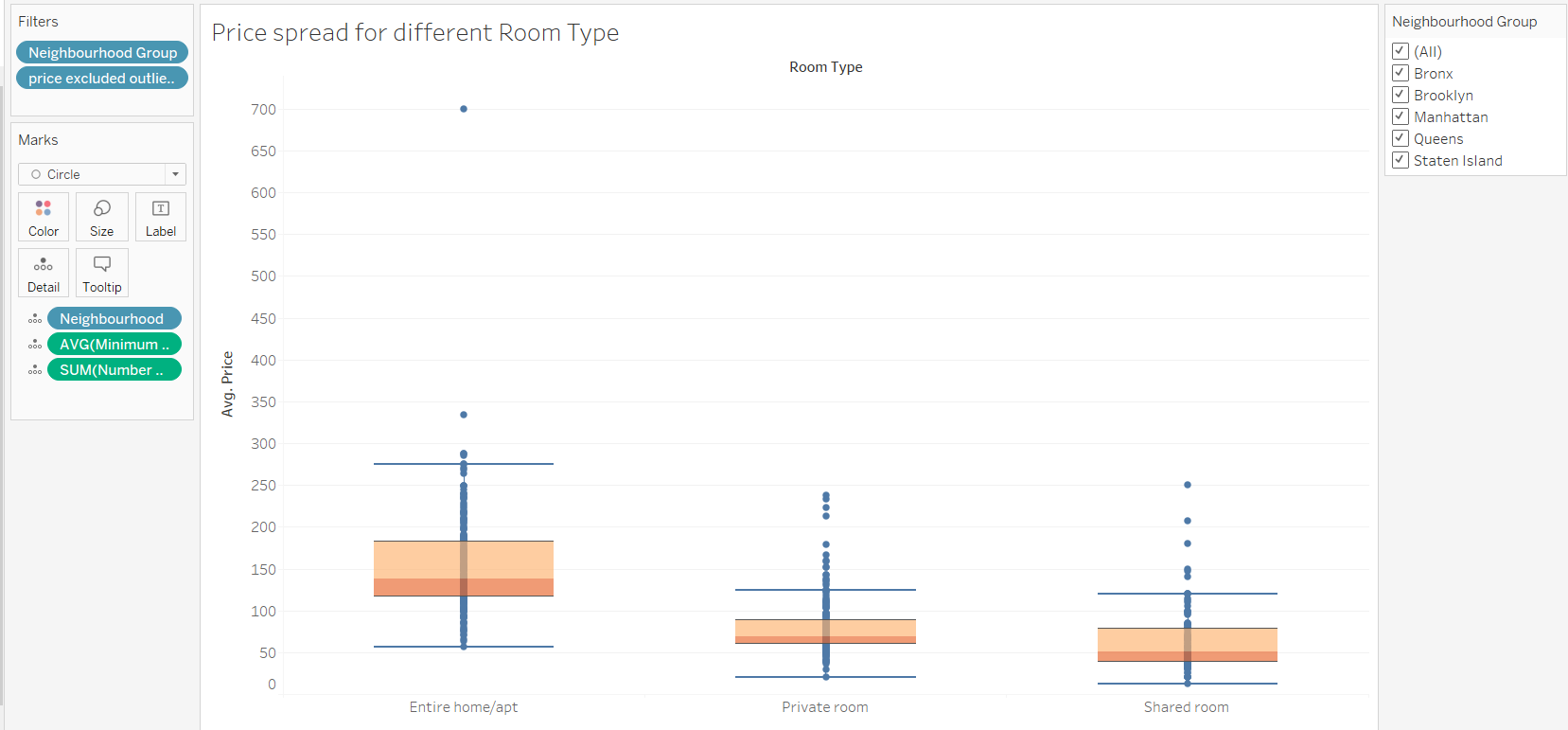
* Popular Price Range in Neighbourhood Groups Vs Room Type

****

***Observations/Insights: The popular price ranges are from 50 to 250 range. Above 250 , not many booking are seen.***

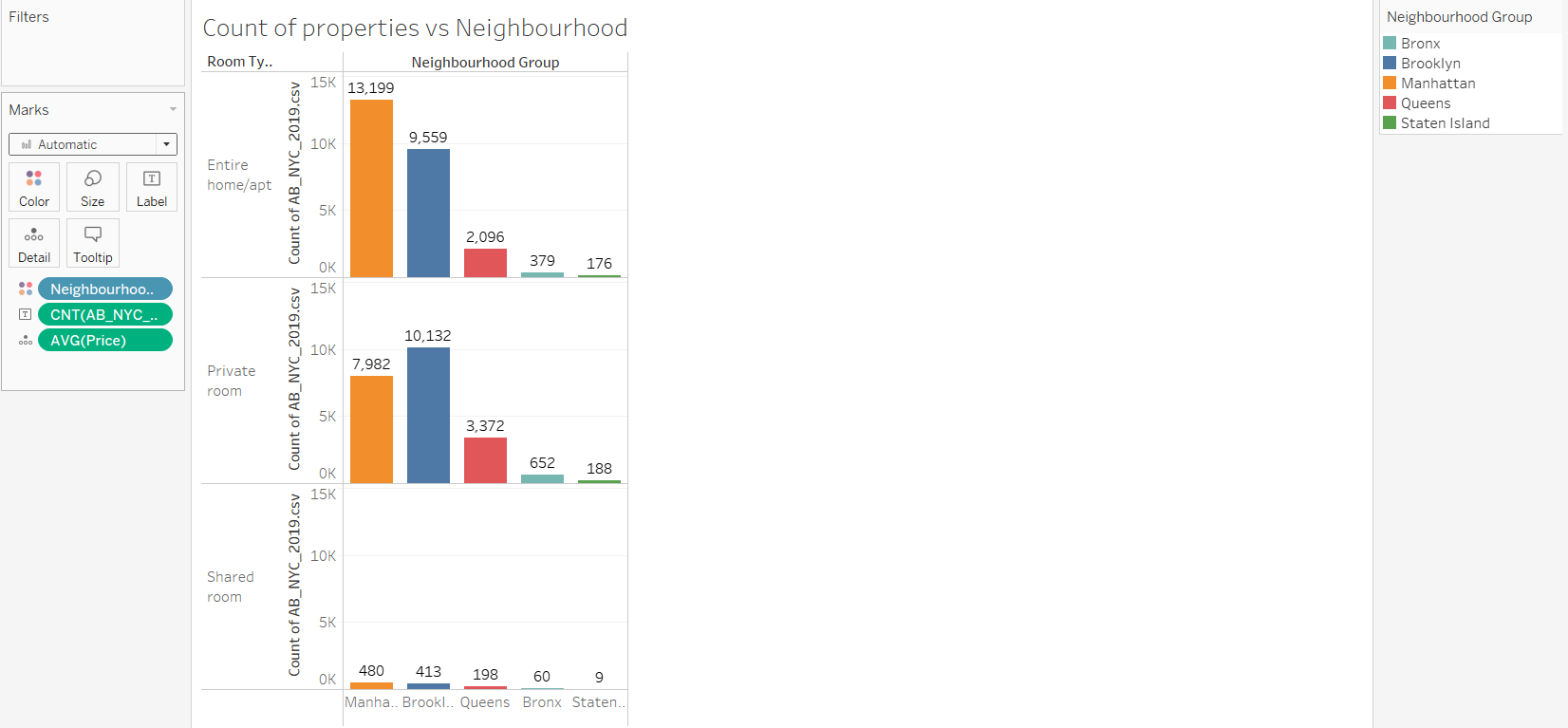
***Insight:*** *More booking are towards less priced properties and popular Room Type is Entire Home/Apt and Private Rooms. We can see there are less number of shared rooms only. But people are booking even in shared rooms in relatively lower price. So leveraging Shared rooms also may contribute to higher revenue.*

* Price spread for different Room Type

******

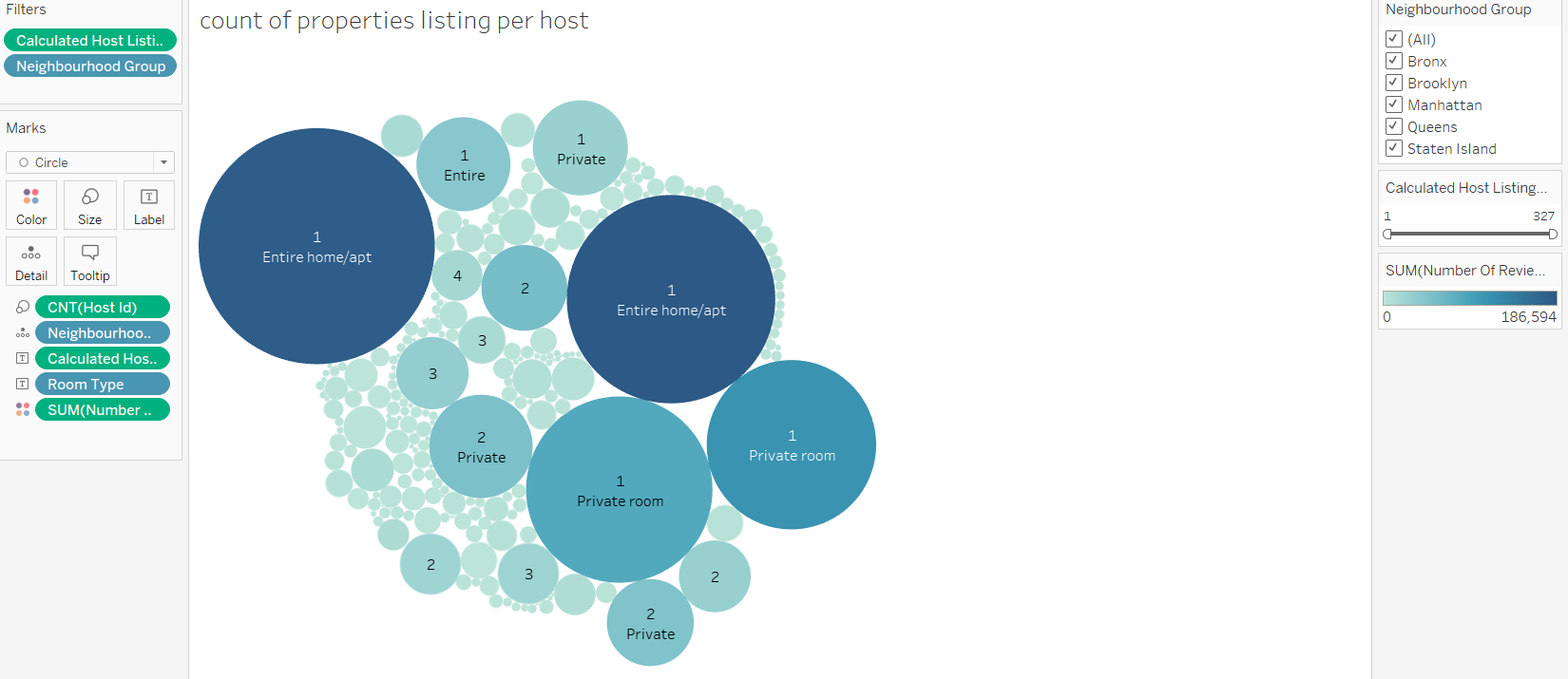
**Observation/insights:** *Entire Home/Apt is costly compared to Private rooms and shared rooms .People prefer mostly Entire Home or Private rooms based on the availability*.

* Number of Properties in each Neighbourhood Group



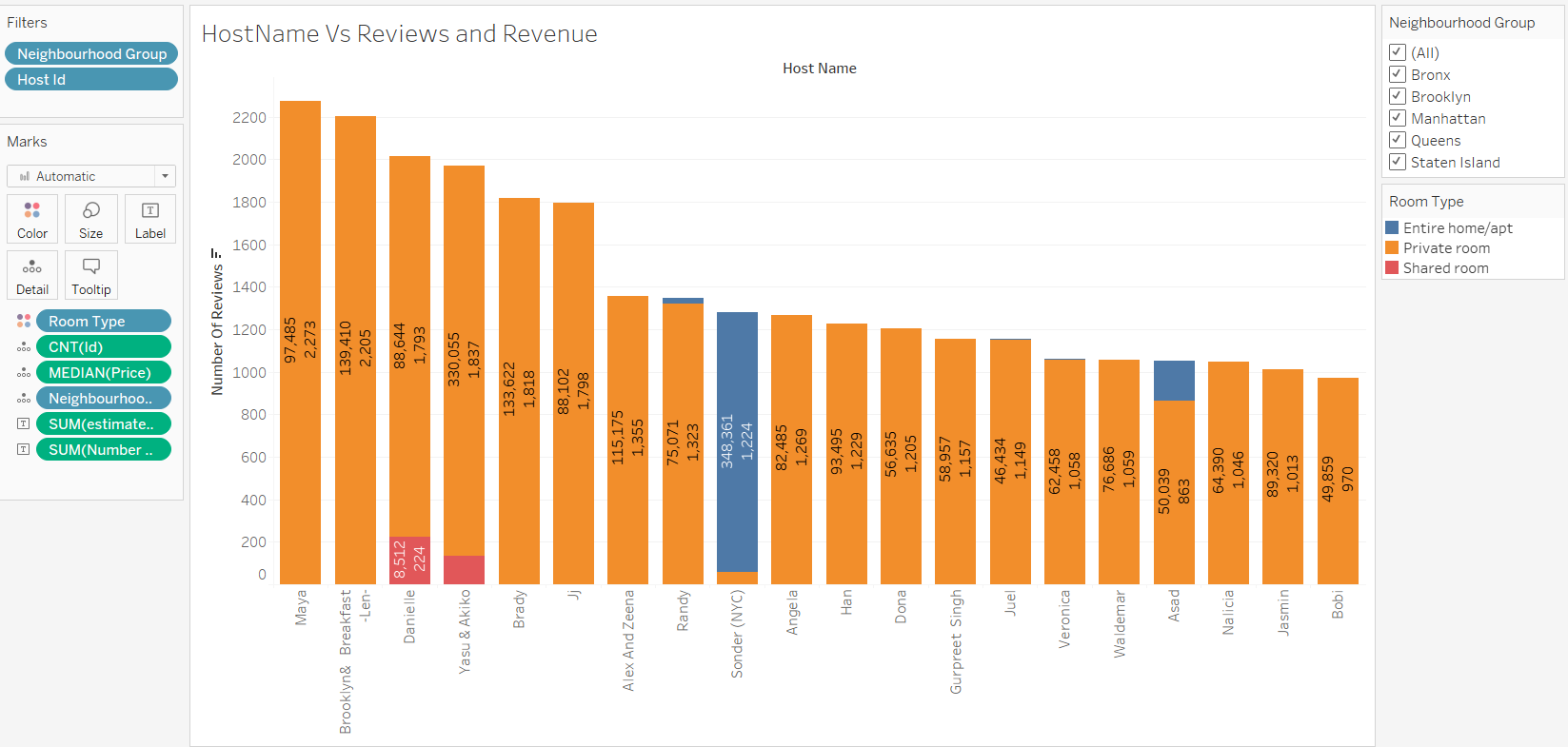
***Observation/Insights:****Majority of the properties are in Manhattan and Brooklyn.Naturally bookings also will be higher in these areas. Entire home is more than private rooms in Manhattan whereas all other regions private rooms are higher. It is clear that Families are more interested in Entire Home and they are interested in Manhattan region as there are many choice in terms of selecting the Rooms.Business can add more properties in other regions such as Bronx,Queens and Staten Island especially Entire Home and private rooms. It is also a good option to increase the shared rooms also in all the regions as there it is lower priced.*

* Property Listing Hostwise Vs Bookings

******

***Observation/Insight****: Some people are very picky , they might think the host having very less number of listings of Homes/Rooms would be good to choose and maintained well.They might focus on the hosts having less listing and book. Business can think about adding Host having less than 5 listings.*

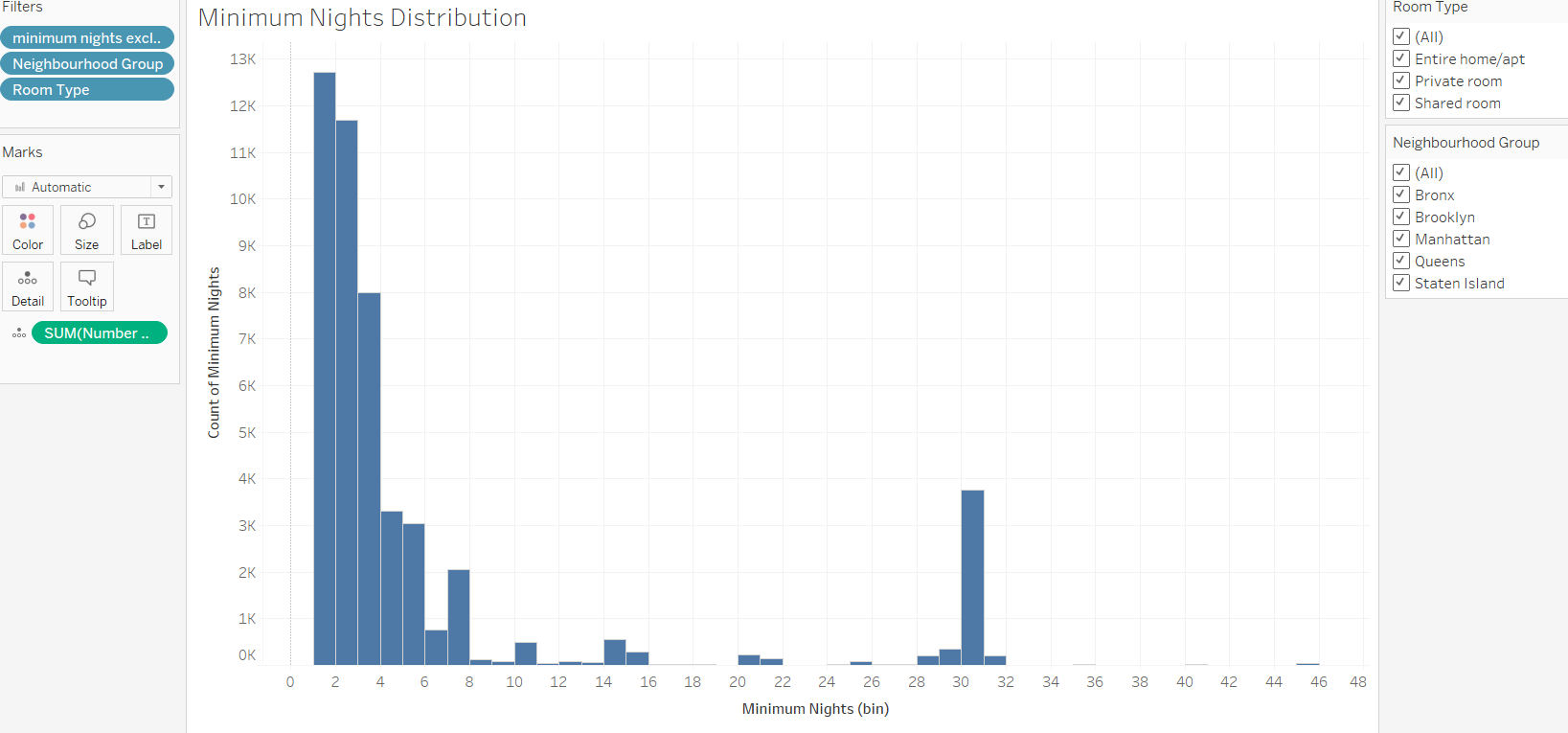
* HostName Vs Reviews and Revenue

**

*Observations/Insights: Above chart shows Top 20 Highest reviewed people or Hosts who hosts the properties in the website.There are many people who book the properties based on the reviews.They just go to the site and check who hosted and the reviews about them. Once booked, when they plan to book again, if they have good experience earlier , then they will definitely book with the same host.*

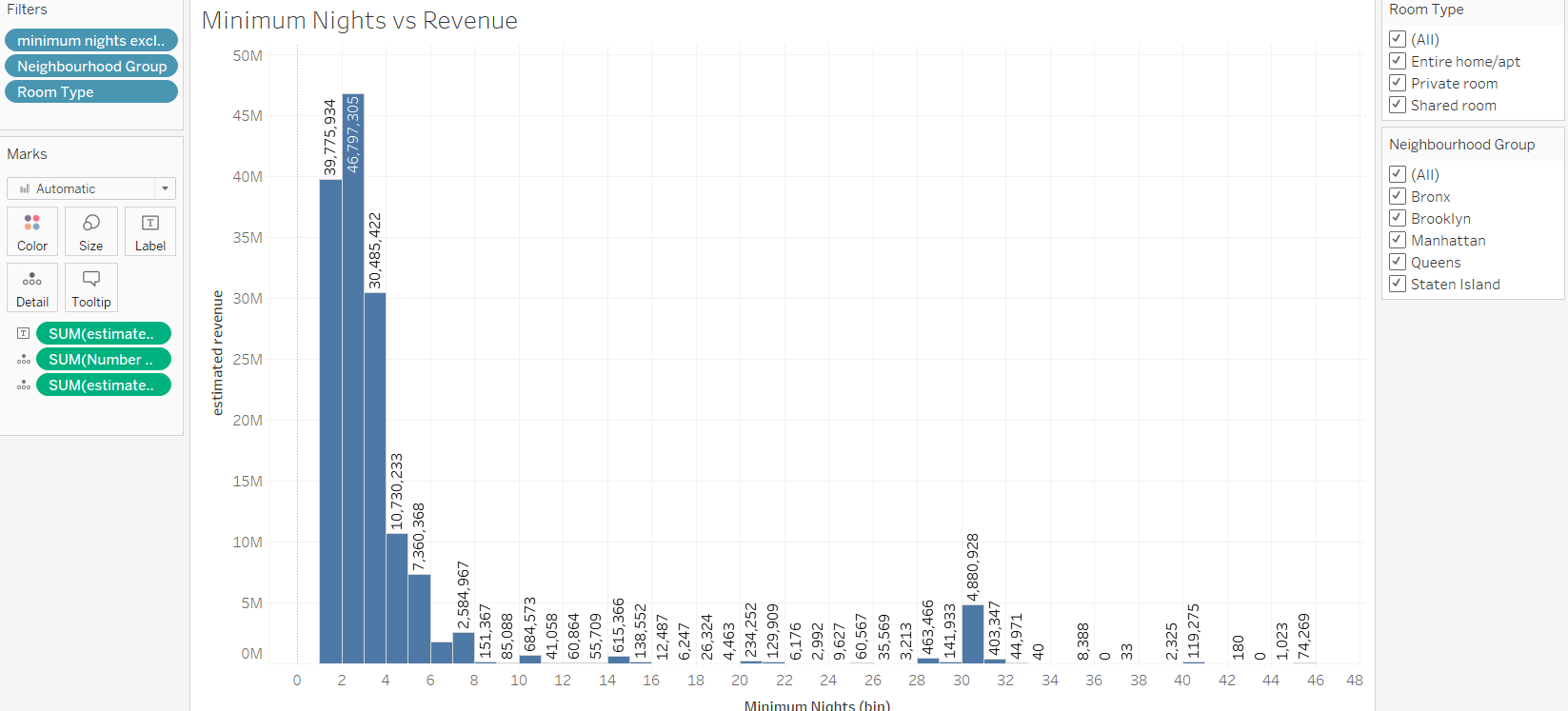
*Business can arrange properties with the Host having good amount of reviews and revenue.Why revenue is that, reviews can be bad also.When we consider both high Reviews and high Revenue, we can arrange the properties under the same host which would be useful in increasing the Business revenue.*

* Minimum Nights Distribution



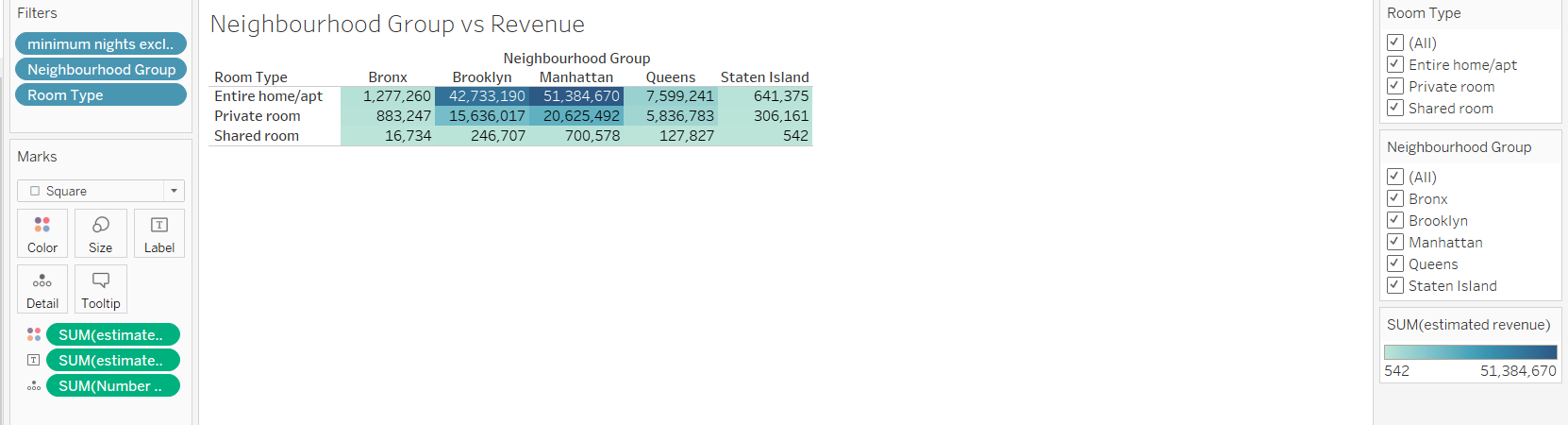
*Observation/Insights:Maximum Properties are for Minimum night stay 1,2 and 3 days and 30 days for all the neighbourhood groups and Room Type.*

* Minimum Nights vs Revenue.

**

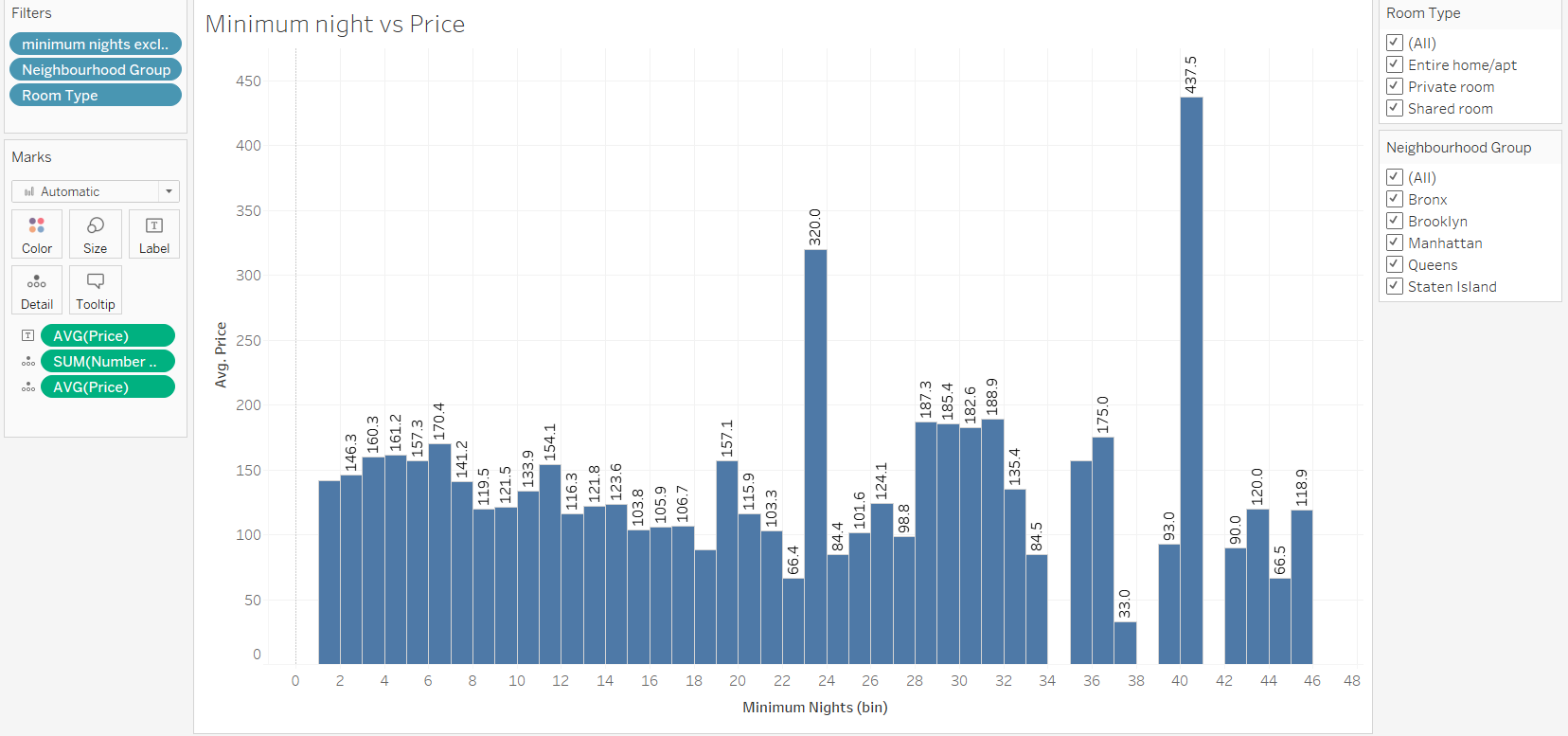
*Observation/insights: The hike in the revenue is for Minimum stays 1 to 5 days , mainly 1, 2 and 3 night stay and 30 days. Business can think about the properties having min night stays in these categories across all locations.*

* Neighbourhoodgroup Vs Revenue



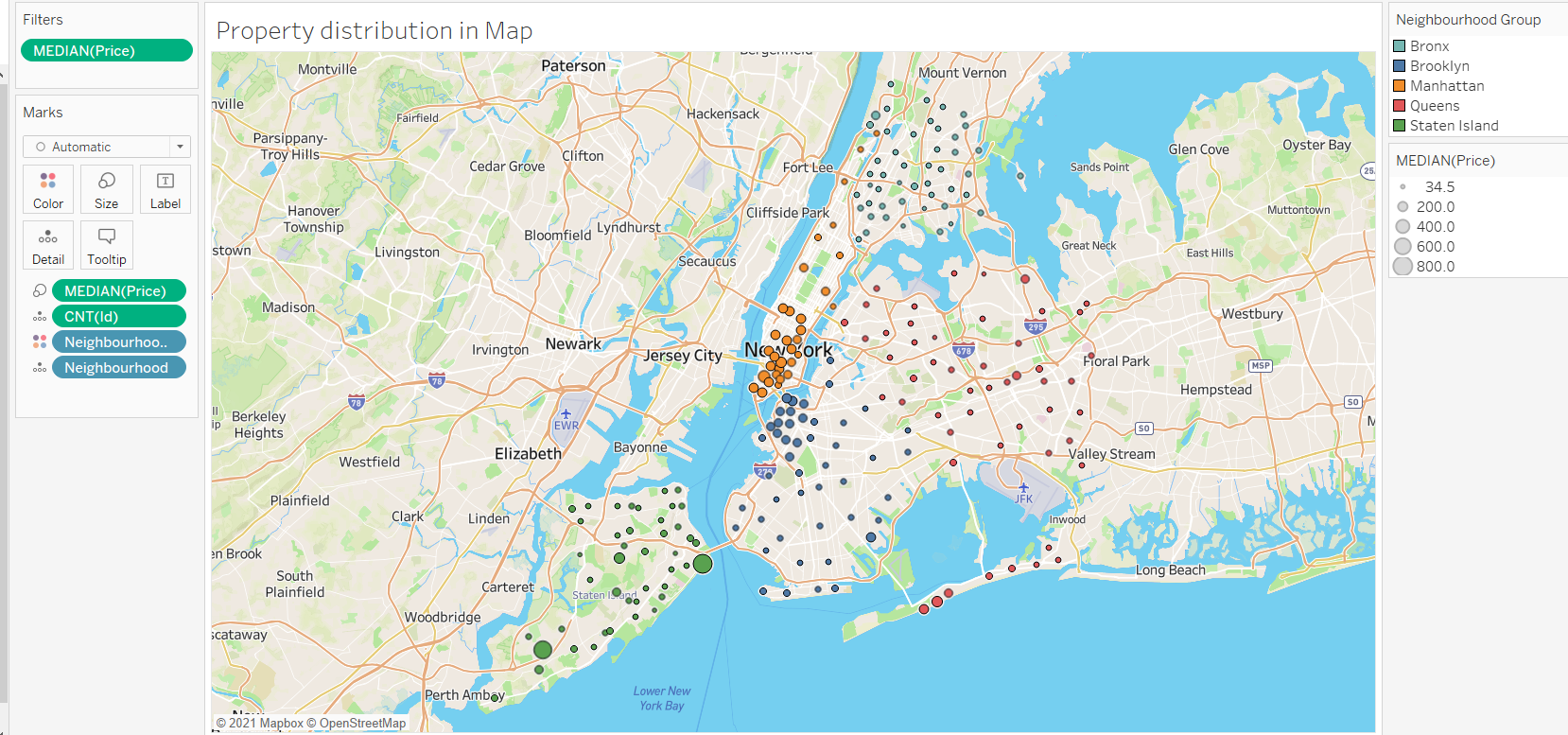
***Observations/Insights:*** *Overall Manhattan has got the highest revenue across all the Room Types followed by Brooklyn and Queens. Staten Island and Bronx are on the lower side where Business can put efforts to bring the revenue higher.*

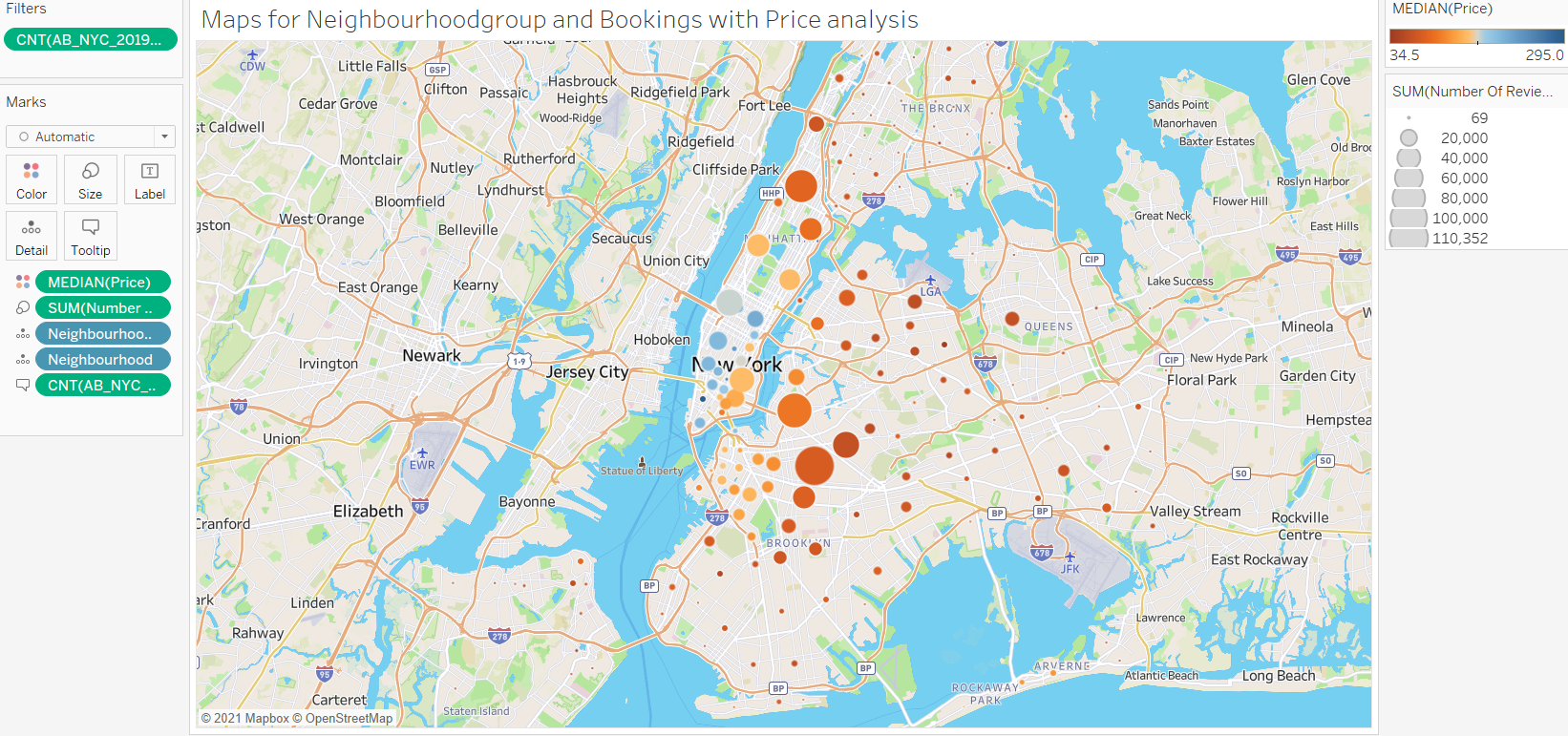
* Average price Vs Minimum Night stay

**

*Observation/Insights: The average price for Minimum night stay is around 150$ and can see there is a hike in 23 days and 41 days.*

* Property distribution in Map

**

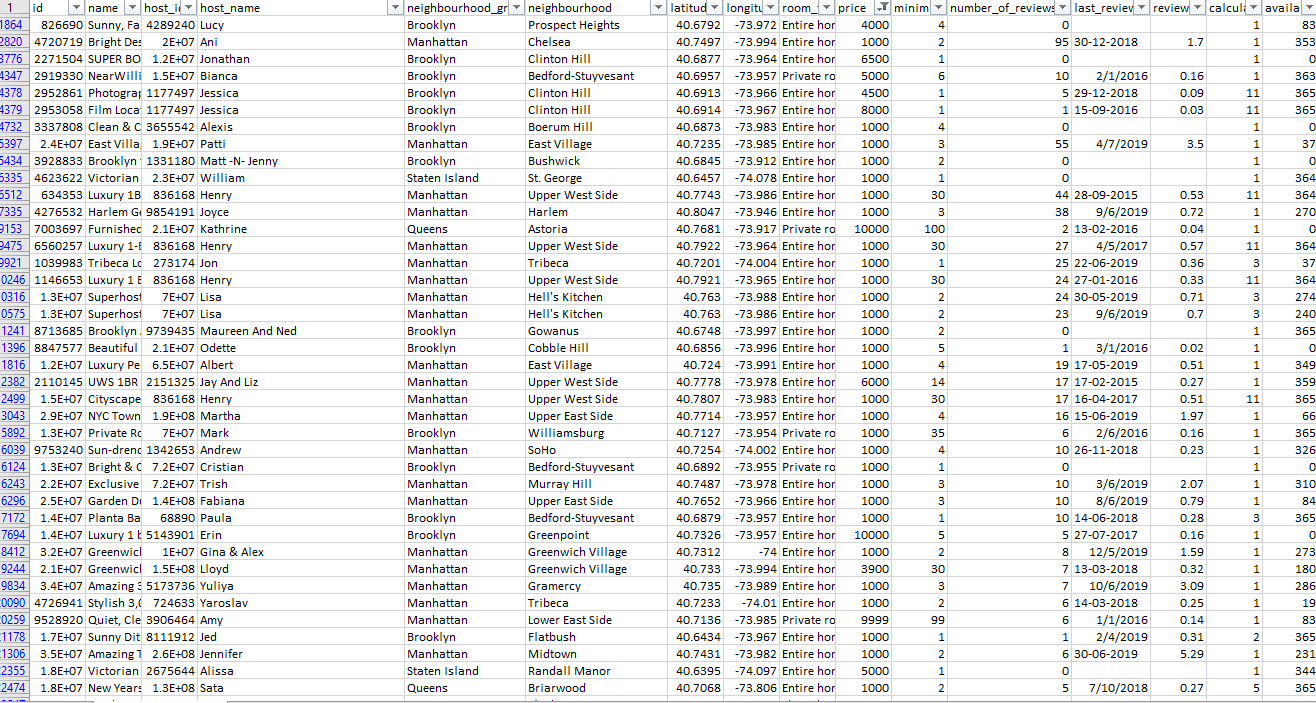
**

*Observation/Insights:*

*Map1: It is clearly seen that the median price for properties are affordable except few properties.Manhattan and Brooklyn at an average, price is medium to high ranges compared to Bronx and Queens. In Staten Island we see the median price is higher in some areas like Woodrow and FortWadsworth than other regions like Manhattan and Brooklyn.*

*Map2: It is visible that number of reviews are higher in lower median price regions.*

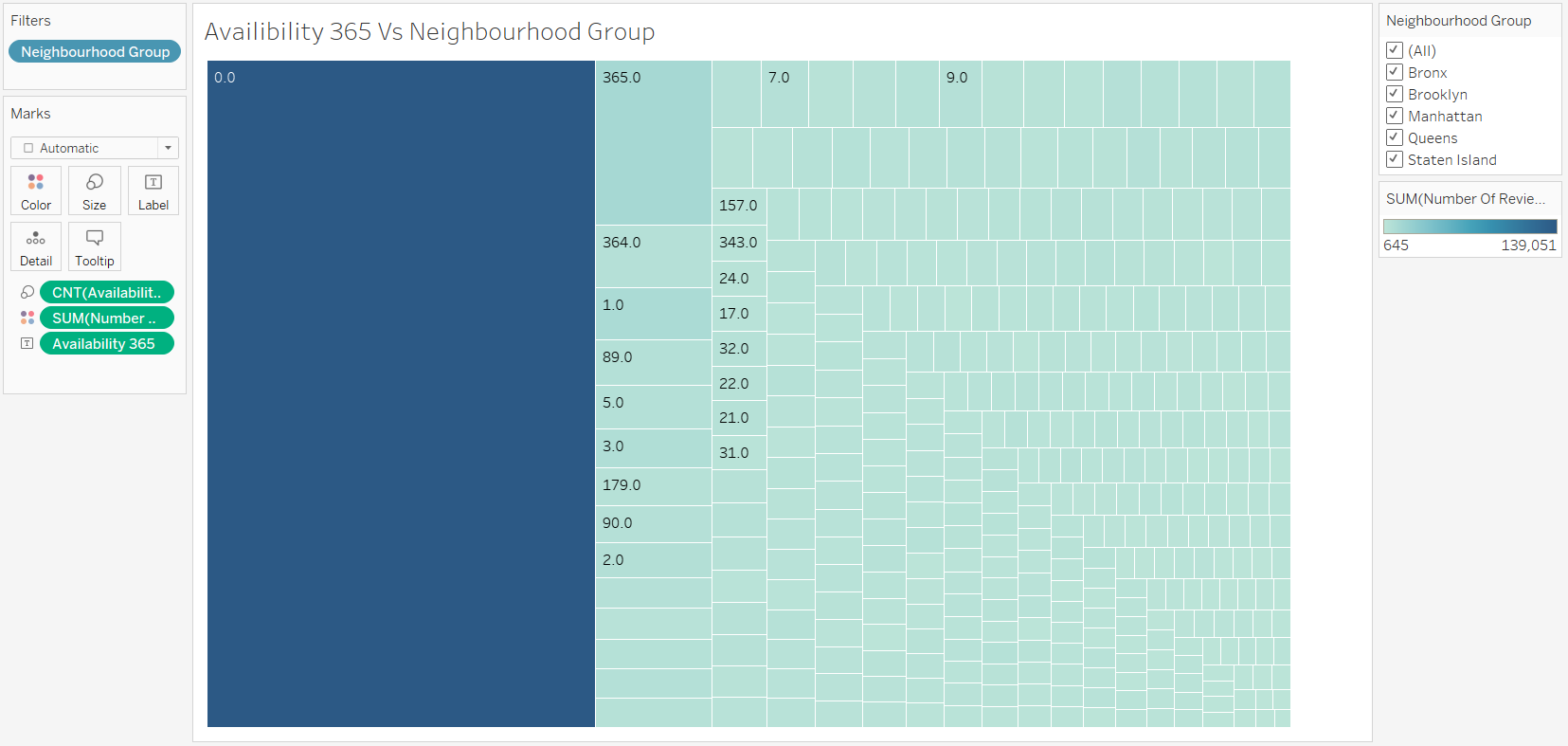
*Outliers in Price:*

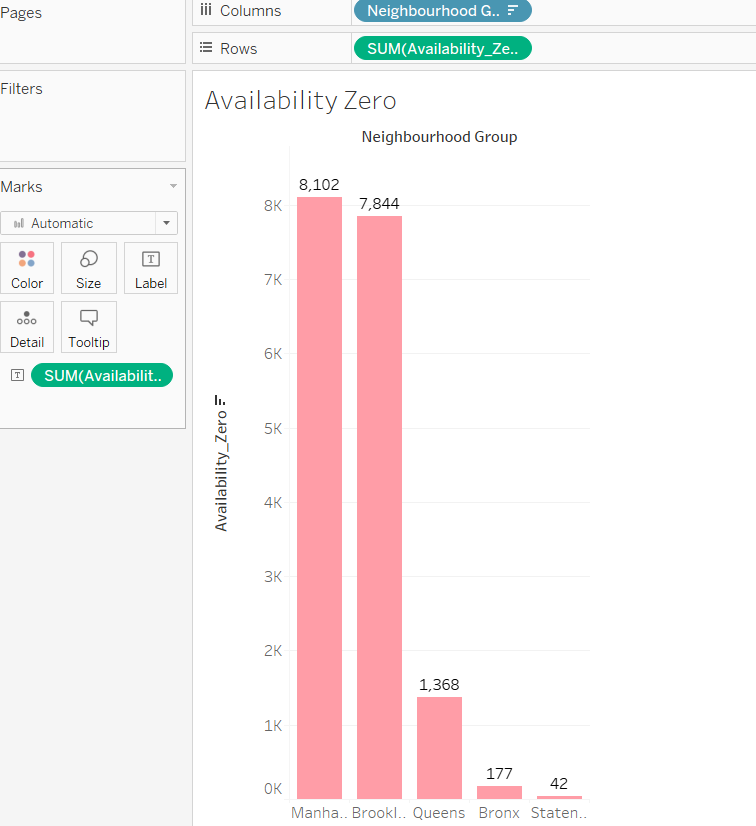
**

*Above excel screenshot shows the outliers in the Price column which received very less number of reviews / bookings and most of the locations are from Manhattan and Brooklyn.*

***Insight****: As we have seen, people are not likely to stay in Luxury villas for min nights stay 1/2/3 and their preference is low priced and convenient place to stay.Business may consider reducing the cost of the properties and increase the bookings and to increase revenue in staten Island where properties are less and high priced , may think about adding the properties which are less priced.*

* Availibility\_365 distribution with reviews



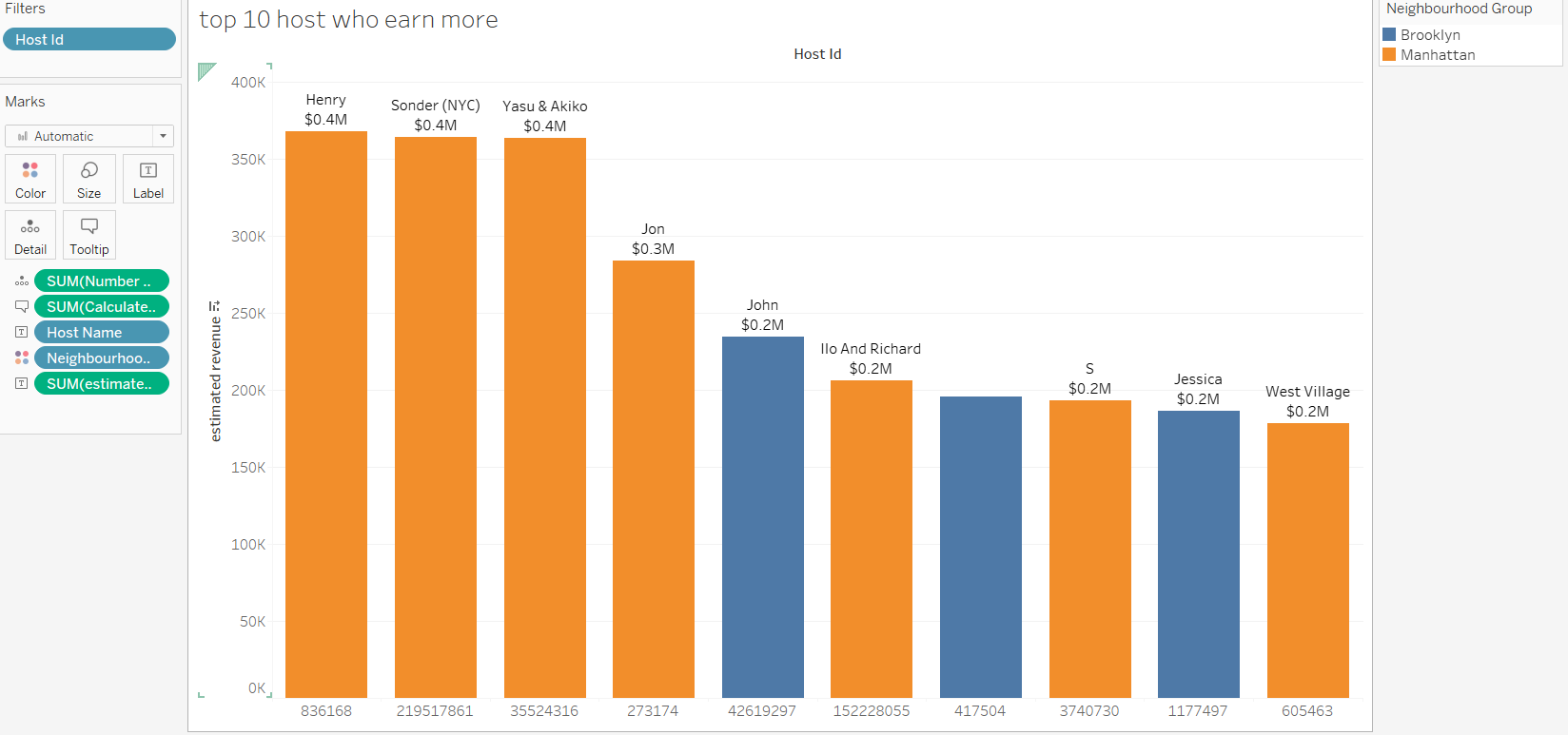
**

***Observation****: The figure 1 represents the highest review received is for Availability as “0” compared to other Availability\_365 entries.*

***Insight****: It is possible that many reviewers or people are interested in these properties and like to visit. Assuming that, we can restart the service inorder to increase the revenue.*

*Fig 2 represents the number of properties not available for listing*

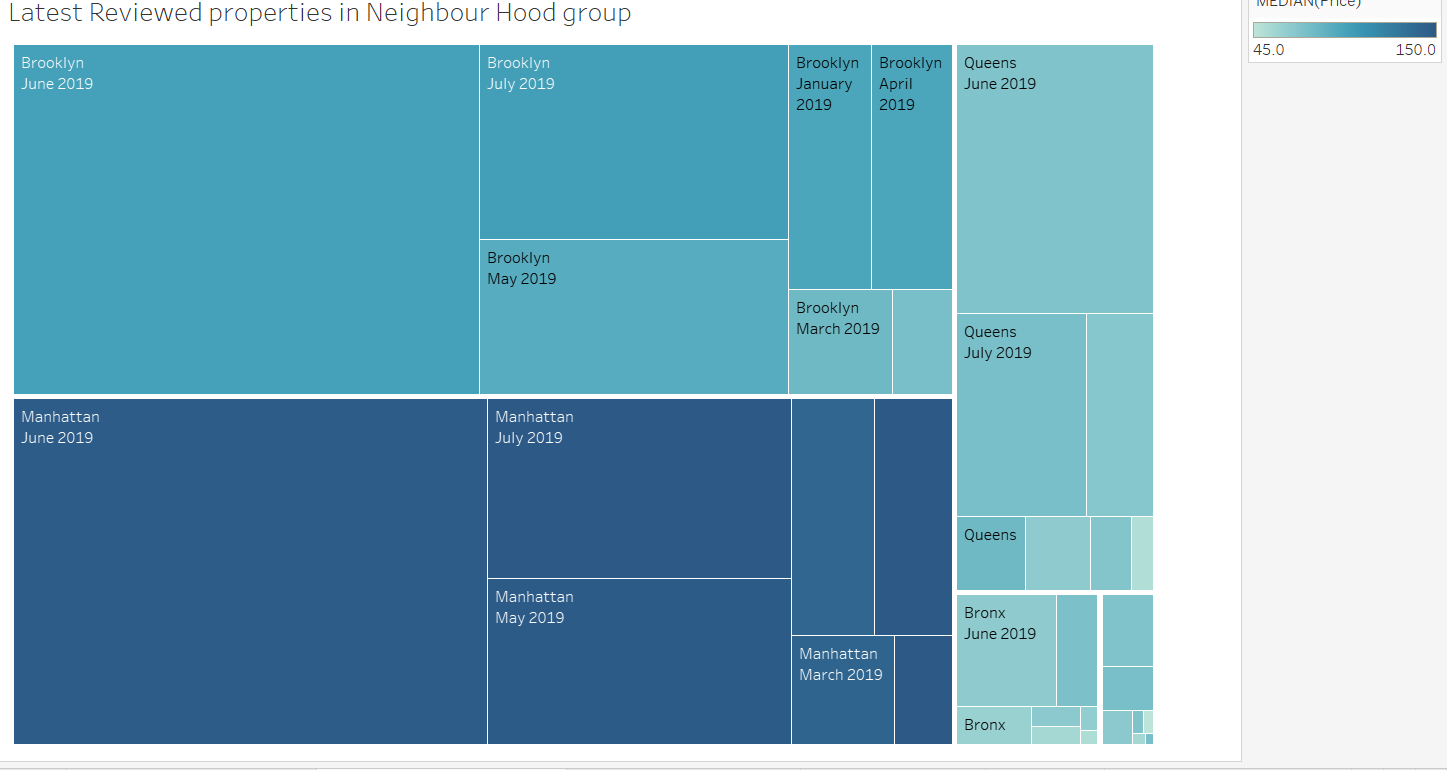
* Top Earning Hosts

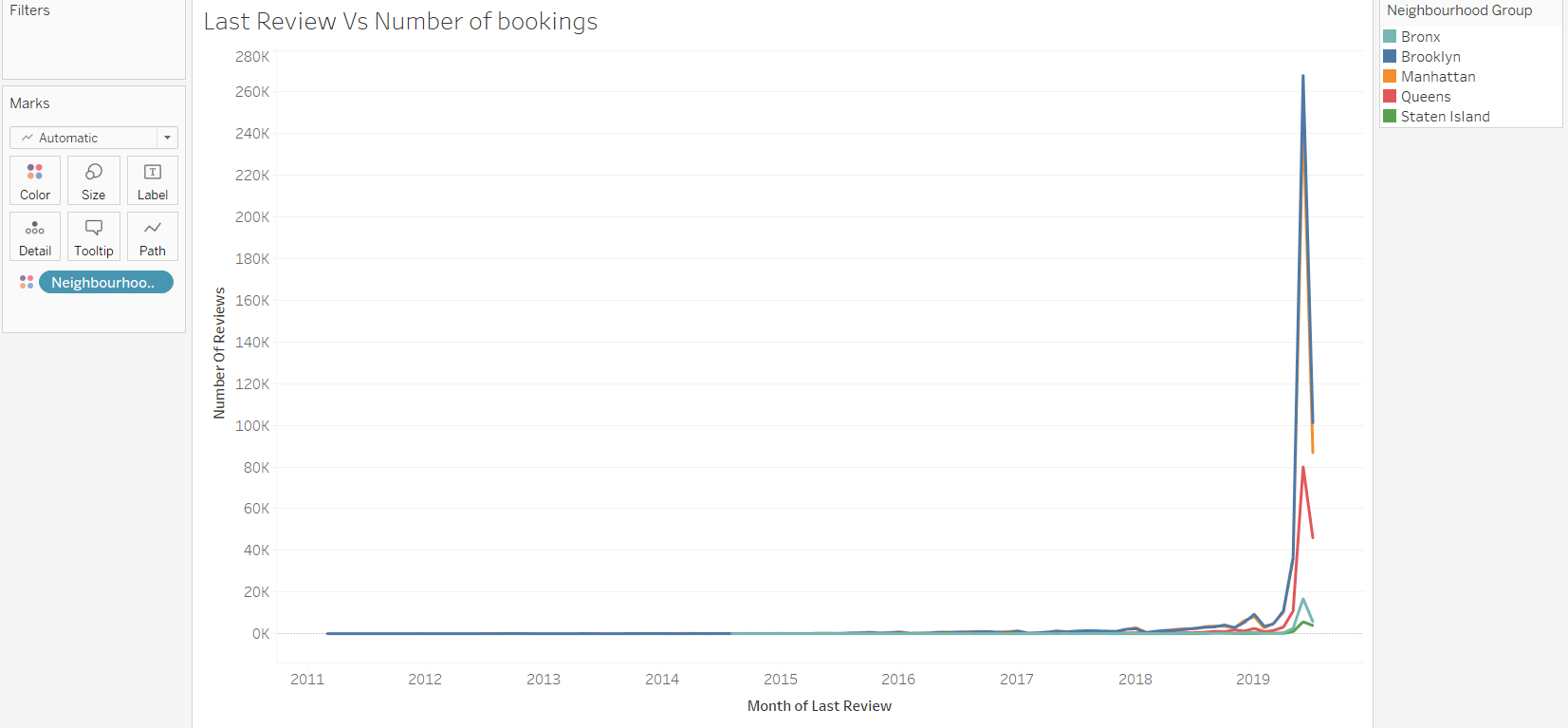


*Observations: Above chart shows the top 10 host who earn highest revenues.*

*Insight: Business can take these Host help and increase the revenue by contacting them to add property listings under them.*

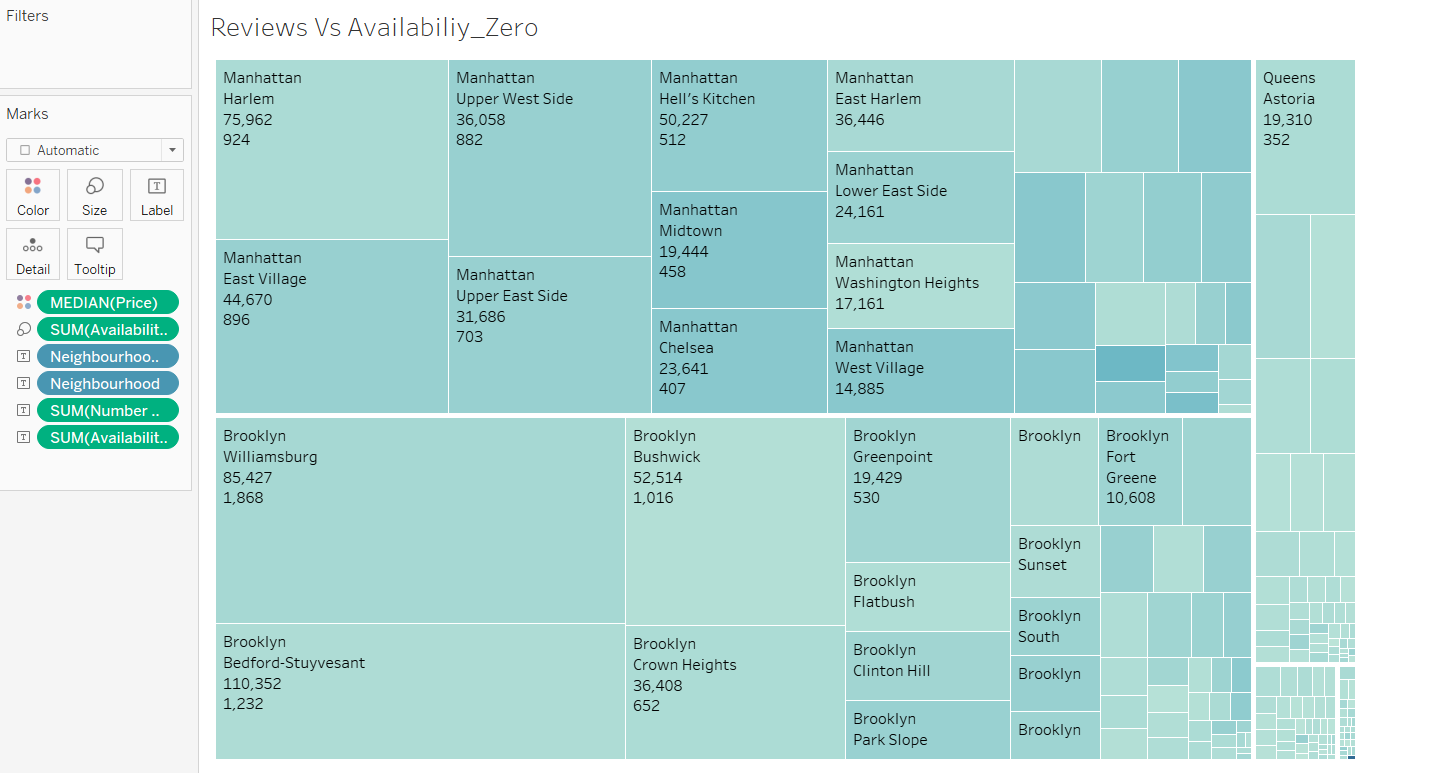
* Latest Review Vs Neighbourhood group

**

**

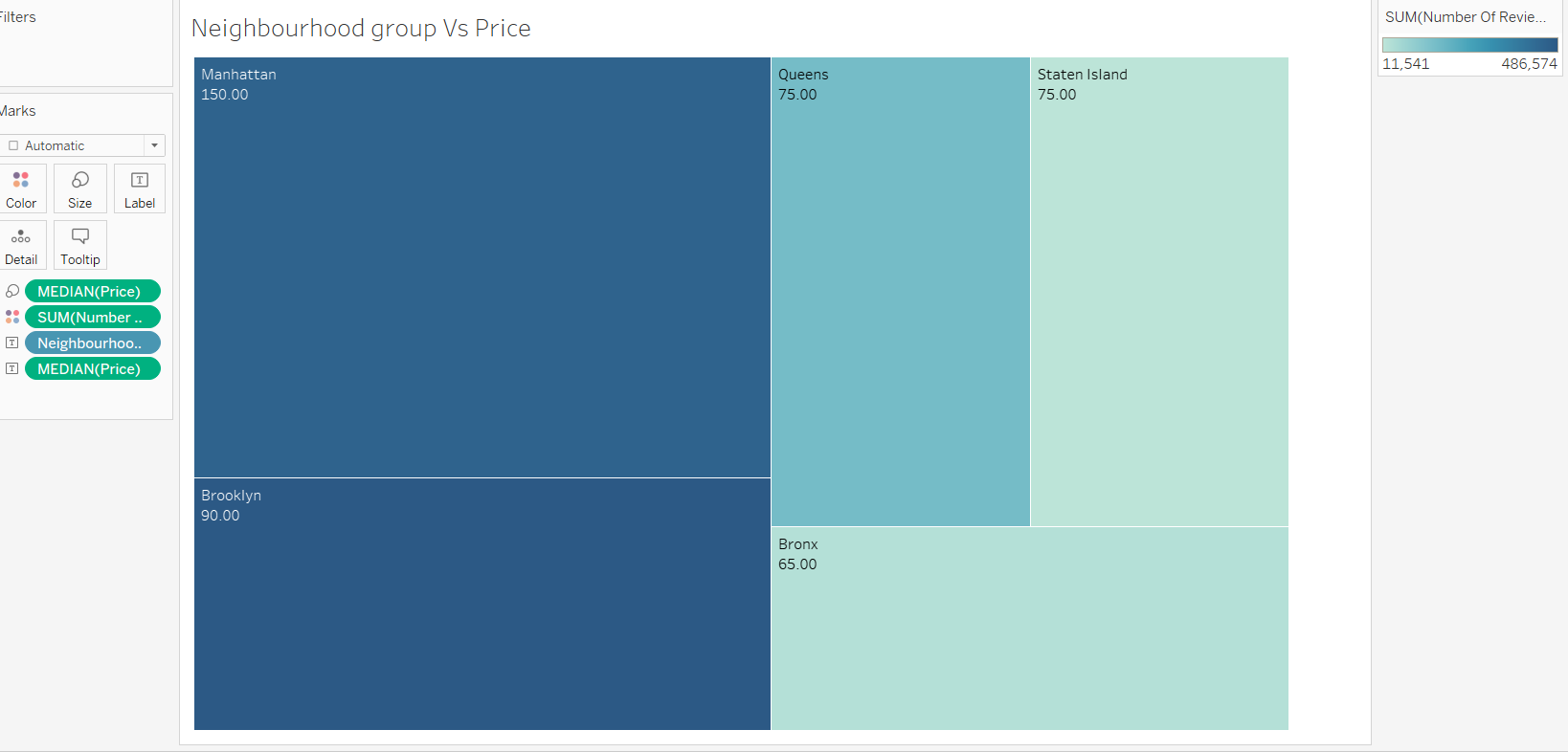
***Observation/Insights****: We can clearly see Manhattan and Brooklyn reviews play a major role in the business which means the bookings are mainly from Manhattan and Brooklyn. There are contributions from other places in terms of reviews which means business looked like active till this period.*

* Review Vs Availability zero

**

*Observations: Above chart shows the locations /neighbourhood areas where AirBnB received good reviews for the listing having availability\_365 = 0 (Not available in the website).If business could re-think of starting them again, it would be helpful to increase the revenue.*

* NeighbourHood group Vs Price



***Observations***: *We can see the median price of all the neighbourhood group and it is obvious that Manhattan is the economic among all others.*

***Insight****: The rate of properties in Manhattan is higher compared to other groups and people can get properties in reasonable price if add more properties to the listing in other locations other than Manhattan. There are possibilities that people may end up taking in Manhattan when they have no options in terms of listings of their choice.*