San José State University

Department of Applied Data Science

DATA 230

Data Visualization

Fall 2023

Section 27

Instructor: Andrew H. Bond

**Term Project: Price Predictions and Regional Characteristics in Airbnb Data.**

**Dinesh Reddy Kothur**

**017475367**

**Abstract:**

Since its establishment in 2008, Airbnb completely changes the travel and hospitality sector by offering guests a unique and personalized alternative for conventional lodging. This project takes a holistic approach to a thorough analysis of Airbnb data.

In this exploration of Airbnb data, I have used python for data manipulation and analysis. Libraries like pandas and NumPy played a crucial role in conducting Exploratory Data Analysis(EDA). I have used pandas for efficient data manipulation ,cleaning and structuring of the data.

With the help of Tableau visualization tool, complex patterns and insights within Airbnb dataset were bought to life. Tableau served as a catalyst for a thorough representation of pricing dynamics, neighborhood diversity, host contributions and temporal trends through interactive dashboards and visually appealing graphics. Theses graphs and charts played a crucial role in better understating of minute details of Airbnb’s ecosystem, leading to a greater comprehension of user preferences and accommodation choices.

In this exploration of Airbnb data, the journey starts with careful examination of average cost by the type of accommodation in various regions. This visualization reveals complex patterns in pricing dynamics and reveals the preferences of Airbnb users in different places. This comprehensive analysis aims to shed light on pricing and regional variations in addition to adding to the current conversation about how to improve the Airbnb experience.

**TABLE OF CONTENTS**

ABSTRACT……………………………………………………………………………………1

DATA DESCRIPTION………………………………………………………………………3

DATA CLEANING……………………………………………………………………………4

DATA ANALYSIS AND VISUALIZATION…………………………………………….8

1. MULTILINE CHARTS…………………………………………………….8
2. DONUT CHARTS…………………….……………………………………9
3. HORIZONTAL BAR CHART……………………………………………10
4. BUBBLE CHART………………………….……………………………….11
5. LOLLIPOP CHART………………..………………………………………12
6. HORIZONTAL BAR CHART……………………………………………13
7. TREE MAP…………………………………………………………………..14
8. STACKED BAR CHART..…………………………………………………15
9. AREA MAP…………………………………………..……………………..16
10. DASHBOARD……………………………………………………………….17
11. USE CASES…………………………………………………………………...19
12. CONCLUSION……………………………………………………………….20

Dataset Description:

|  |  |
| --- | --- |
| **Variables** | **Description** |
| id | Identifier for the listing. |
| Name | Name of the listing |
| host\_id | Identifier for the host. |
| hostname | Name of the host. |
| neighbourhood\_group | The place in which the property is located |
| neighbourhood | The neighborhood in which the property is located |
| latitude | Latitude coordinate of the property. |
| longitude | Longitude coordinate of the property. |
| room\_type | Type of room in the listing (e.g., Entire home/apt, Private room, Shared room). |
| price | Price of the listing per night. |
| minimum\_nights | Minimum no of nights required to book. |
| number\_of\_reviews | No of reviews for the listing. |
| last\_review | Date of the last review |
| reviews\_per\_month | Average no of reviews received per month |
| calculated\_host\_listings\_count | Count of listings for the host. |
| availability\_365 | No of days the property is available for booking in a year |

**Exploratory Data Analysis:** Downloaded the dataset from Kaggle. Using Python libraries like pandas and numpy for EDA.

Loading the data:

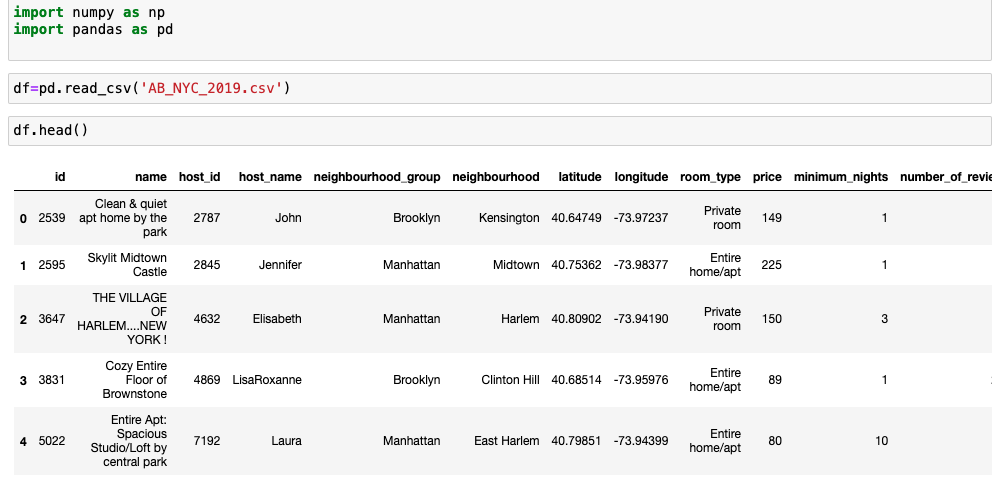
****

Figure 1:Loading data

The above pic shows loading of the data and displaying the top 5 rows in the data with df.head()

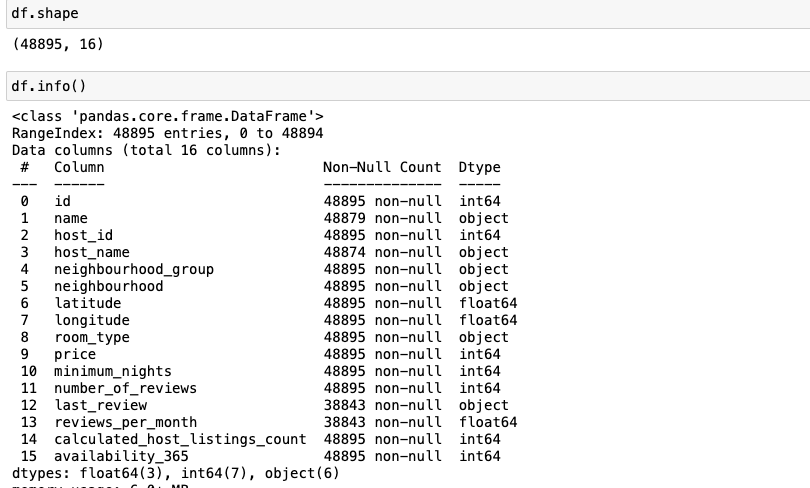


Figure 2: Data Information

Finding shape of the dataset using df.shape() , our dataset has 48895 rows and 16 columns.

df.info(): it gives information about our columns and their data types.

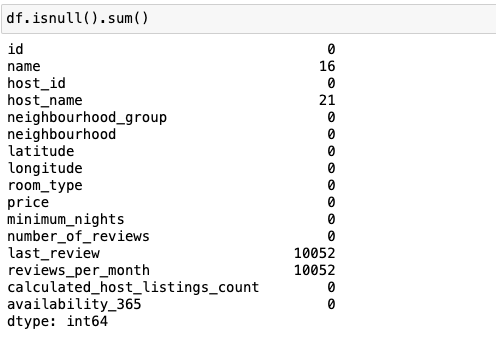


Figure 3: Columns

*df.isnull().sum()* function is used to find the total no of null values present in our dataset.

From the above figure we can see that there are 16 null values in our ‘name’ column and 21 null values in ‘host\_name’ column.  
we also have null values in last review and reviews per month which is around 25% of our data.

**Handling the Missing data:** We can handle these missing values with various techniques like replacing with mean ,median etc. But in our case the missing values are present in our name and host\_name section. These values can not be replaced so it is better to drop these columns .

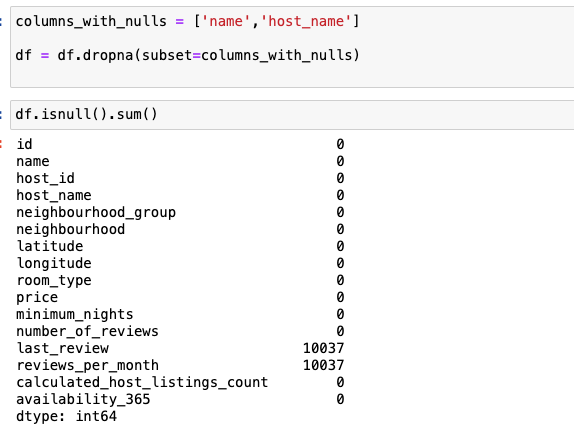


Figure 4: Finding nulls

Now we have dropped columns ‘host\_name’ and ‘name’ using df.dropna method .

Now let us check if there are any abnormal values in our data like zeros.

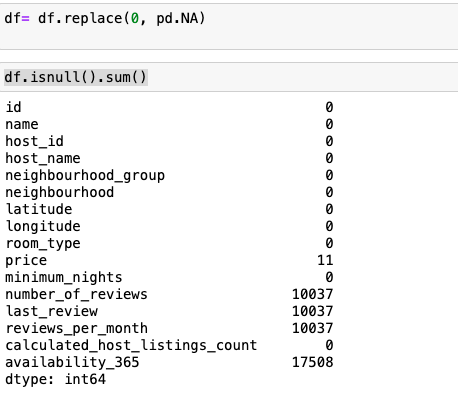


Figure 5: cleaning null values

Here we are 1st replacing our zeros with NA values and counting the no of null present in it using *df.isnull().sum()*

We can see that there are around 17508 zeros in our availability\_365 column. This contributes around 35% of the data. Using this column causes inaccuracy in predictions. So, it will be better to drop this column.  
We can do that by using *df.drop(‘availability’,axis=1).*

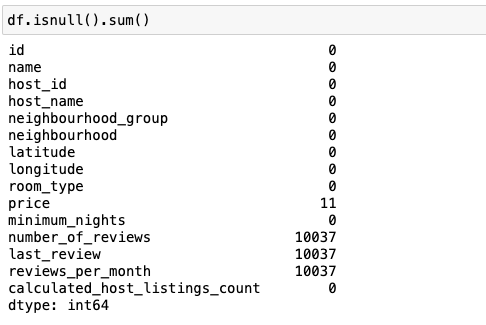
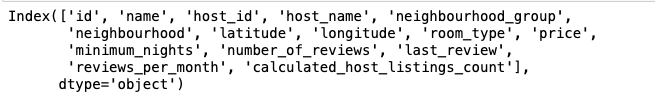


Figure 6 :dropping columns.

From the above picture we can see that the column has been dropped .

We also observe that there are few more null values in our data In ‘price’ as well as ‘reviews columns’ .  
Price is assumed to be zero in those few instances and those null values In reviews columns are replaced with zeros .



The above picture shows the all the columns present in our data.

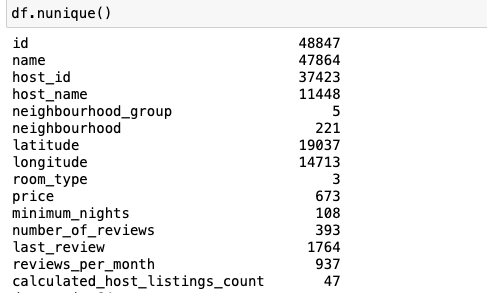


Figure 7:unique values

Unique values: using *df.unique()* we can find the total no of unique values present in our data.

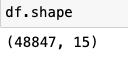


Figure 8: dataframe shape.

Now after all cleaning the shape of our dataset is (48847,15).

Discriptive analysis of our dataset.

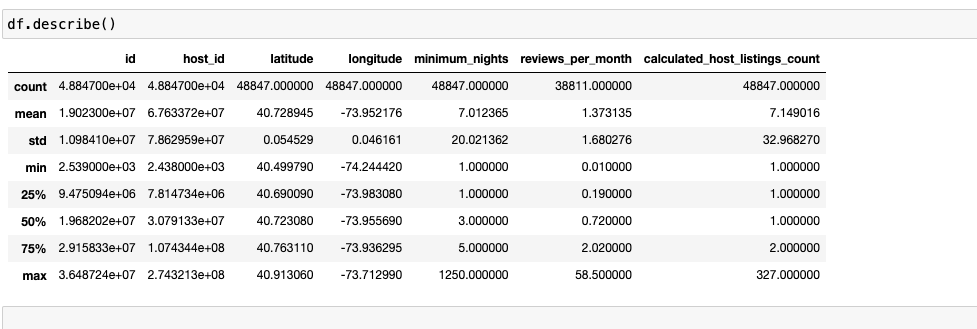
****

Figure 9:Discriptive analysis

Now we have performed all the data cleaning ,we need to download this data .

We can do it using pandas. df.to\_csv( ) function downloads the newly created data set in csv format. We also can specify the path to download.

****

Figure 10:Downloading the cleaned data.

**Data Visualization with Tableau:**

**Worksheet 1: Number of reviews per month**

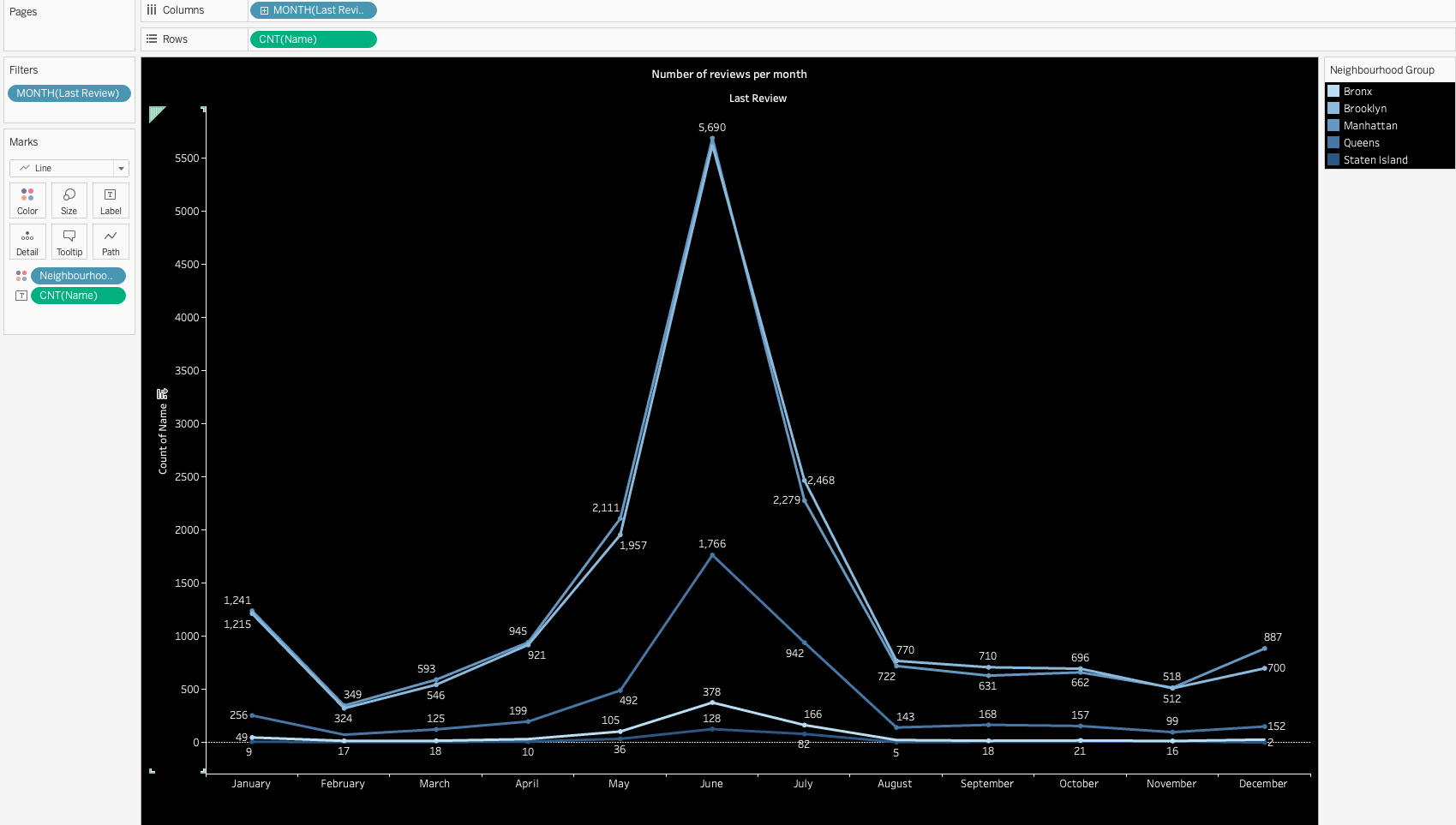


Figure 11: Multiline chart reviews per month per neighbourhood group

* The line chart shows monthly reviews of 5 different neighbourhood groups namely Bronx, Brooklyn, Manhattan, Queens, State island .
* For the above graph we have Months on the columns and count of names in the Rows.
* We also added a filter by month which gives us the monthly distribution of reviews.
* To differentiate the neighborhood groups in the line chart we added neighbourhood groups to color section .
* To show the monthly reviews we added count of names to the label section.
* We have set the title of the worksheet as ”Number of Reviews per Month” .
* From the graph we ca deduce that out of all the months June has seen the peak no of reviews in all the neighbourhood groups.
* We also can say that Brooklyn and Manhattan dominated with highest reviews in the month of June with 5690 and 5626 reviews. Bronx and staten island have lowest no of revies overall
* Overall the graph shows the distribution of reviews over time on location basis. We can also observe that the peak in may, June is due to the holiday season.

**Worksheet 2: Total Neighbourhoods by neighbourhood groups**

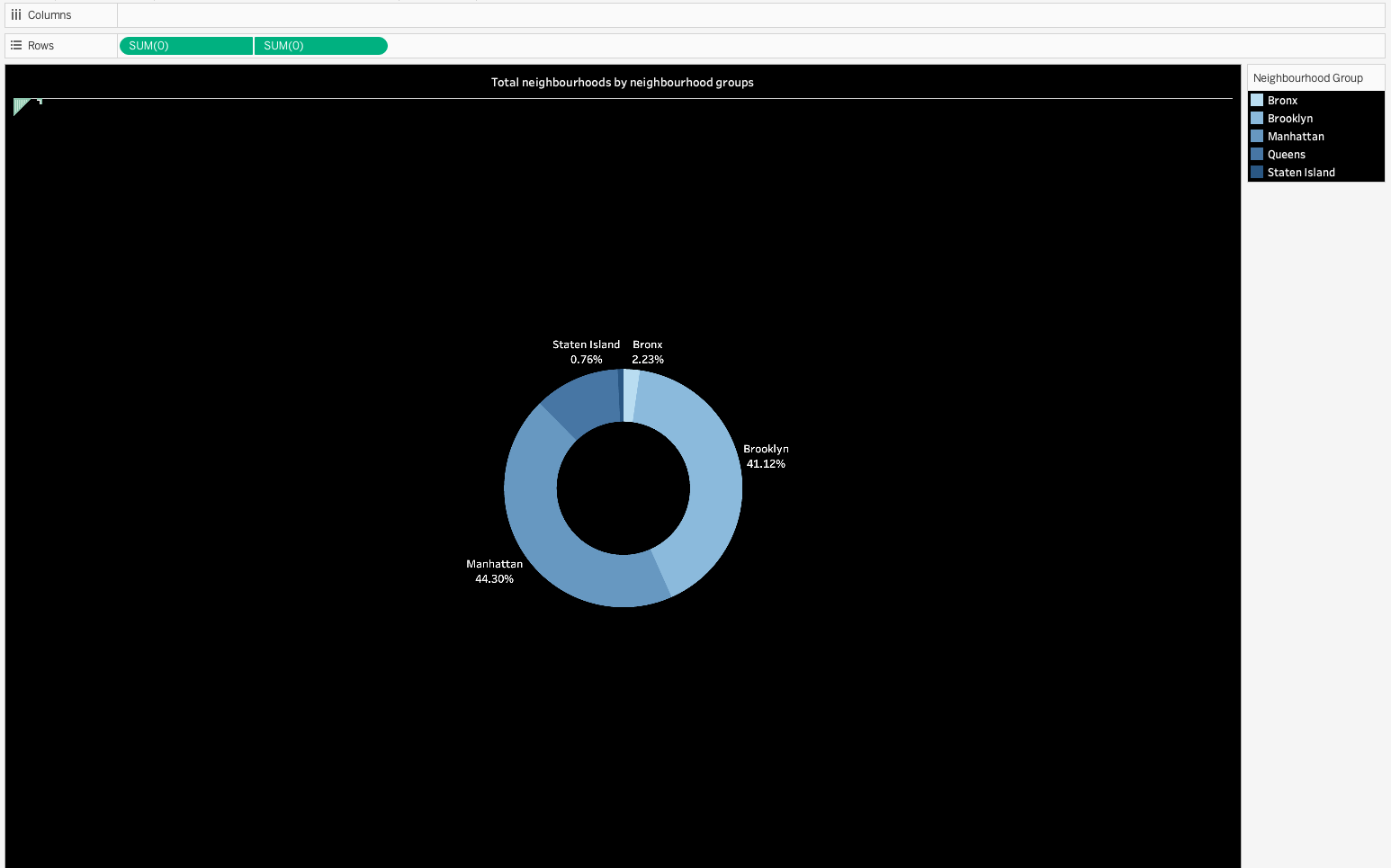
****

Figure 12:Donut Chart showing neighbourhoods by neighourhood groups

* The donut chart is divide into segments representing different New York city areas, with each areas percentage share.
* We have chosen color palate to be blue.
* From the above pie chart we can deduce that Manhattan and Brooklyn contribute to 44.25% i.e it has total of 21,661 neighborhoods and 41.12% i.e 20,104 neighborhood’s.
* Staten Island has the lowest number of neighborhoods with just 373 and it contributes to 0.76% of the total share.
* Bronx is the second lowest neighbourhood group with 1091 neighborhoods and it contributes 2.23% of the total share.

**Worksheet 3: Top hosts by total reviews.**

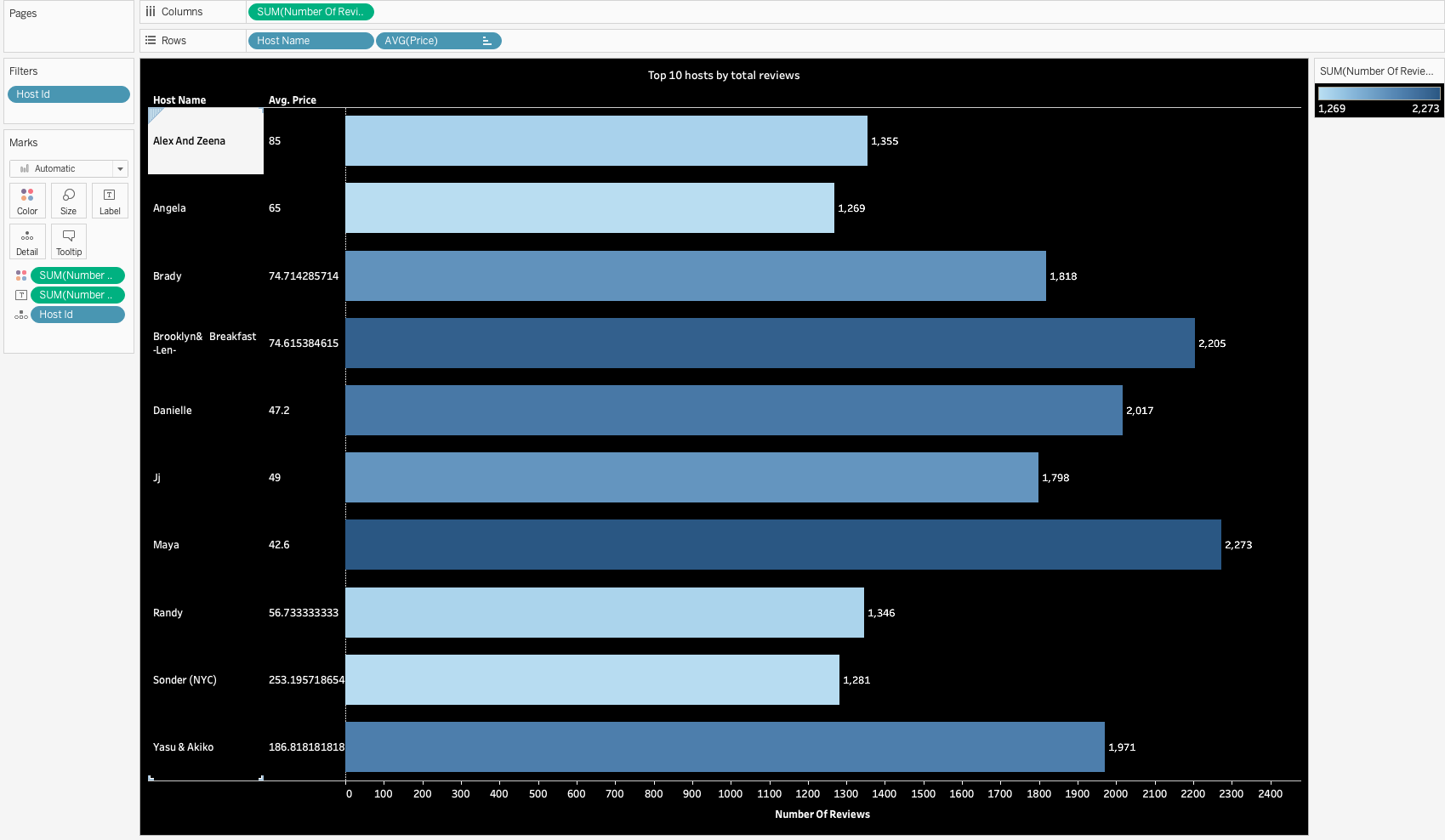
****

Figure 13: Horizontal bar chart of top hosts by reviews

* The above horizontal bar graph shows over all top hosts by reviews in a particular year.
* The chart lists host name in the Y-axis and along with hosts we also added Average price in the rows which shows the average price of that particular host.
* In the column section we have used Sum of Total no of reviews.
* We have added a filter with hosted .
* In order to show the no of reviews a host received In the graph we have used label for sum of no of reviews.
* In the we have hosts on the left and their respective average price in middle and a horizontal bar showing the sum of reviews that user got.
* From the above graph we can see that Sonder is the star host with most no of reviews with 2269 and Maya is the last host in the top 10 hosts with 973 reviews.
* The average price next to host name shows the correlation between price and popularity
* We also can observe that out of top 10 hosts 9 of them have average prices under 100.

**Worksheet 4: Reviews by year**

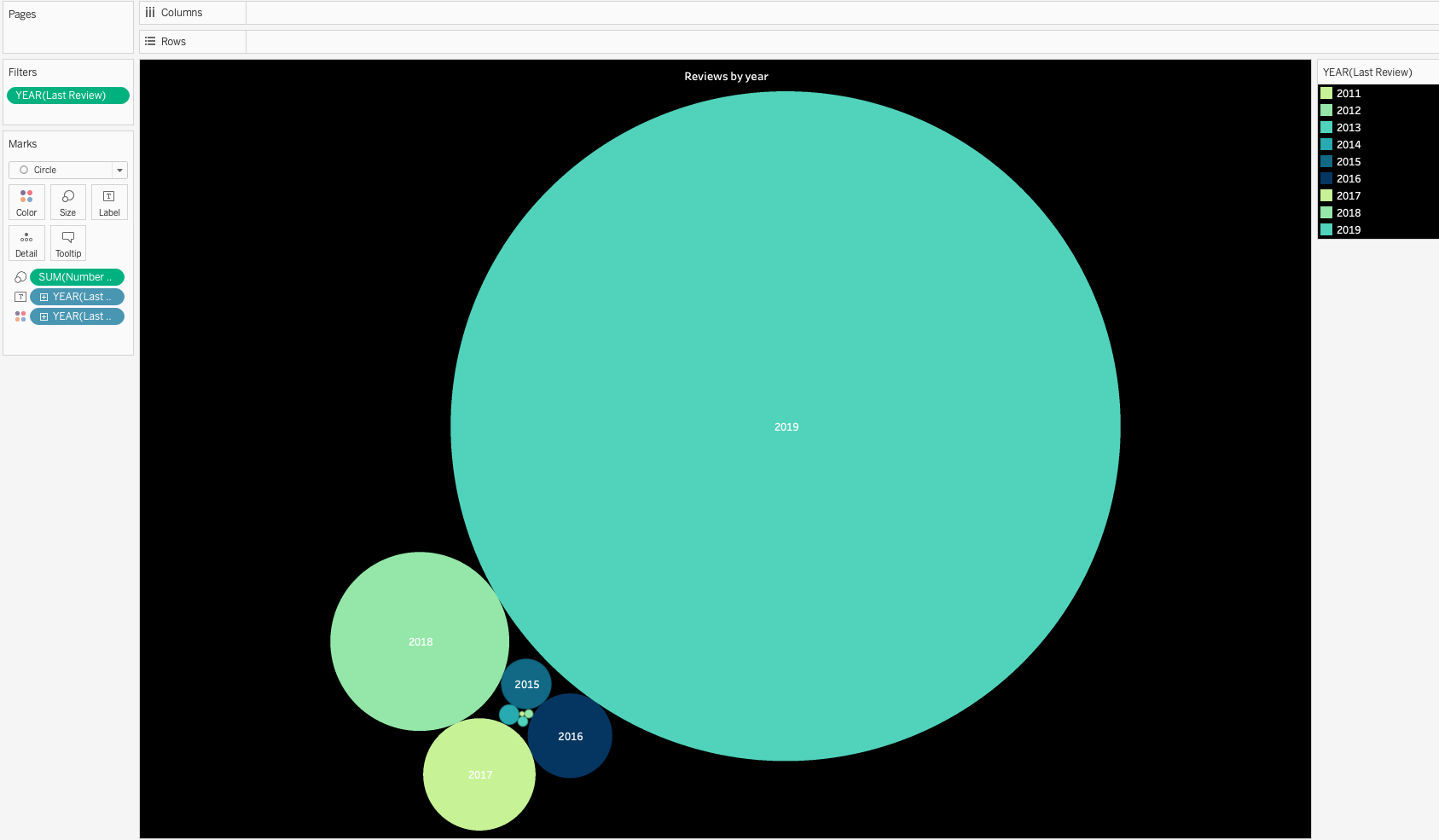
****

Figure 14: Bubble chart showing yearly reviews

* The above screenshot shows the bubble chart of reviews by year
* Here we have total data of reviews from the year 2011 to 2019.
* We have set a legend to differentiate between reviews of each year.
* The bubbles in the bubble chart indicate the sum of total no of reviews in a particular year.
* The size of Bubble in the bubble chart is proportional to the no of reviews.
* From the chart we can observe that 2019 has the most no of reviews with 1,013,755
* The year 2011 has the lowest no of reviews ,it has a total of 61 reviews.
* Airbnb was started in 2008 and the users were less initially so less reviews in initial years.
* From the bubble chart we can deduce that with the increase years the sum of reviews has increased steadily with which we can say that the business is steady and gaining popularity over the years.

**Worksheet 5: Average price by neighbourhood group and room type.**

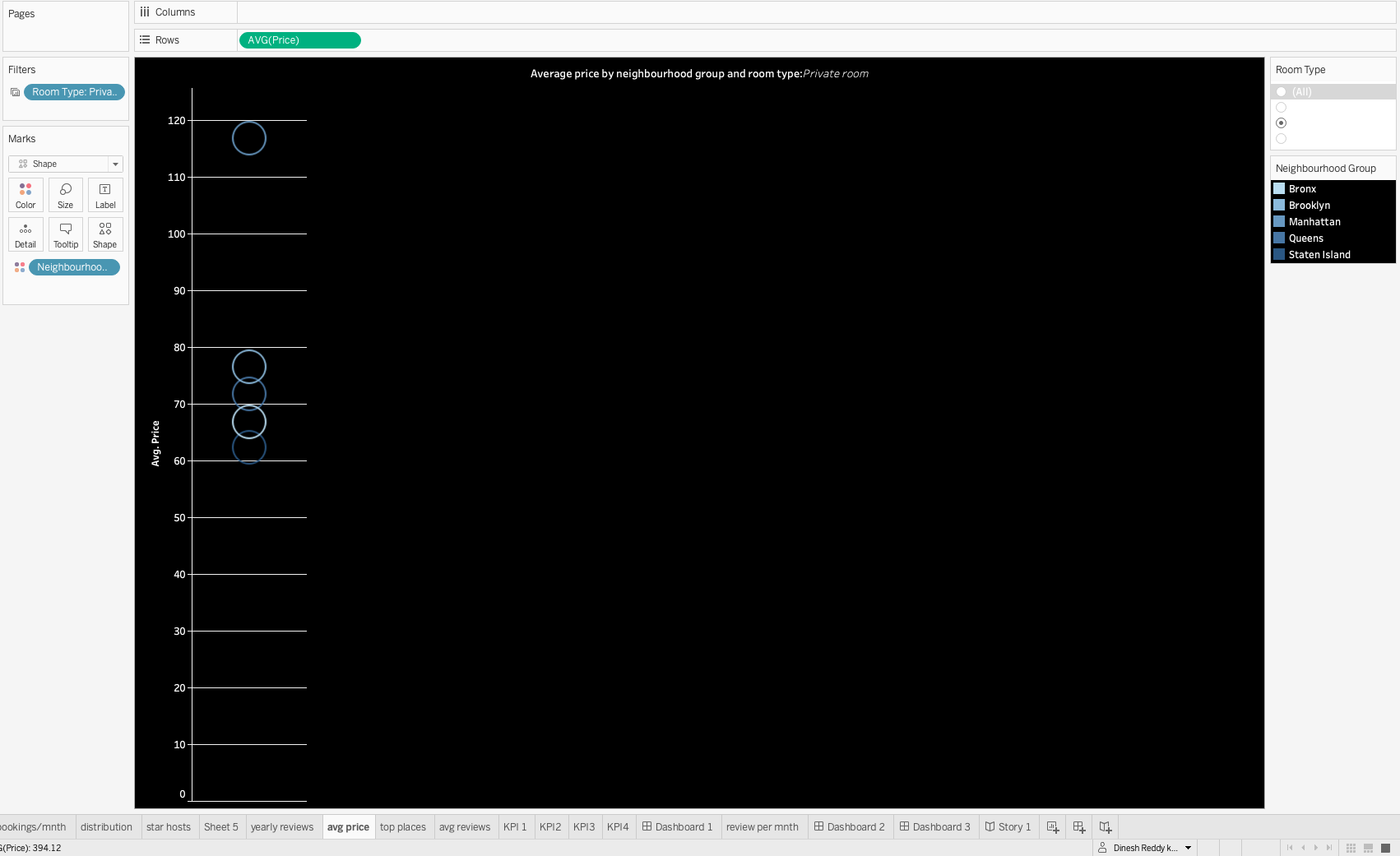
****

Figure 15: lollipop chart of average price by neighbourhood group and room type.

* The above chart shows a lollipop chart for Average price by neighbourhood groups and room type.
* This chart combines bar charts and scatter plots to show datapoints along a line with a circle representing the data point value.
* Here we inserted Average price in the Rows shelf which range from 0 to 120.
* The circles are plotted over the x axis.
* We have added a legend with neighbourhood groups to differentiate .
* Added filter ‘room type ’ to check the average prices of each room type namely apartment, private room and shared room in a neighbourhood.
* Overall Manhattan has the highest average price with 196.88 dollars and Bronx has lowest average price with 87 dollars.
* We can change the filters and check the average prices of particular room type over the neighborhoods. From the graph we can observe that Manhattan has the Highest average price in all room types.

**Worksheet 6: Top Neighbourhoods with highest Average price**

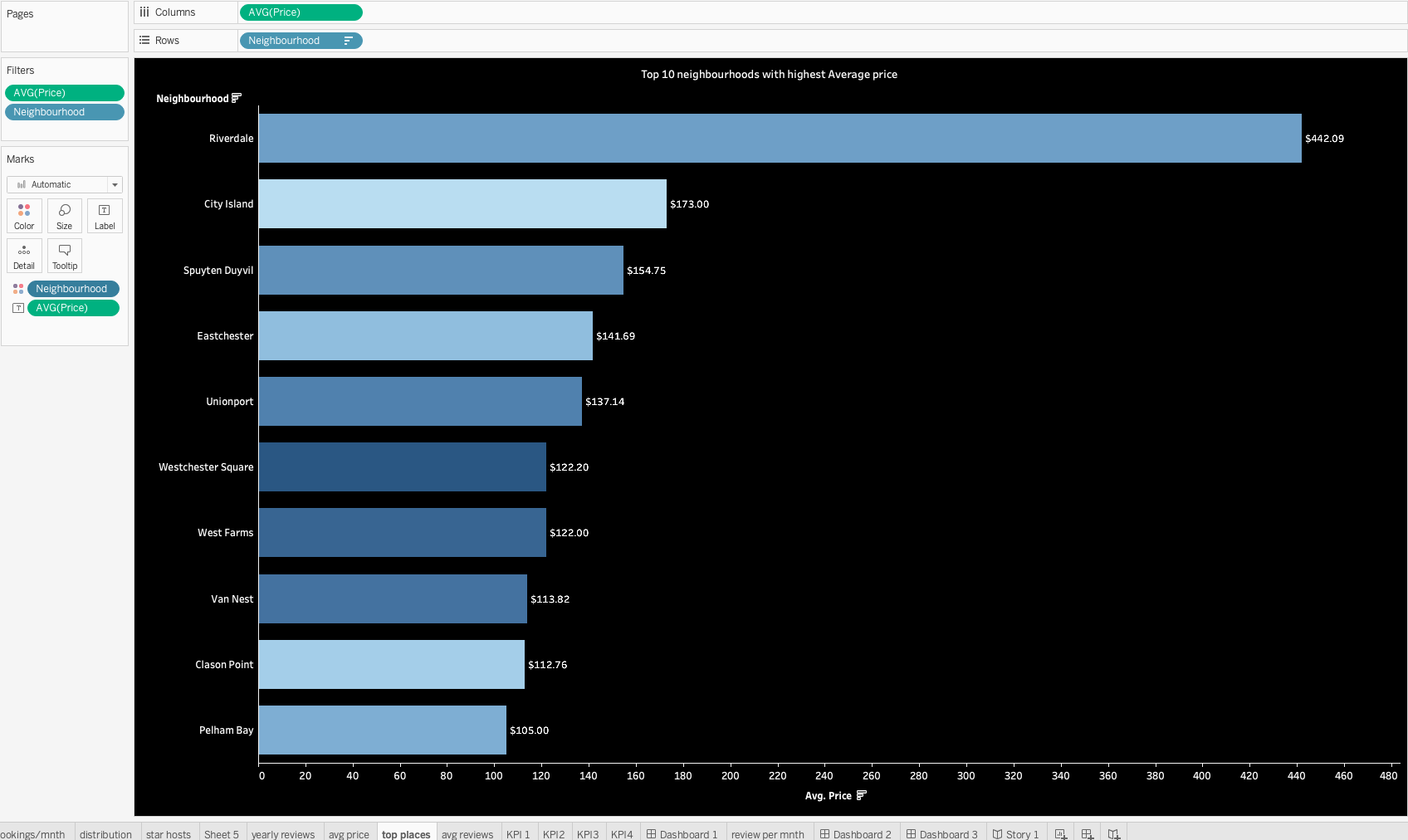
****

Figure 16: Bar chart of top 10 Neighborhoods with highest average prices.

* The above chart is a horizontal bar chart for the top 10 neighbourhoods with highest average price.
* We have added neighbourhood In the rows shelf and Average price in the column shelf
* We have added two filters Average price and Neighbourhood.
* For labelling the average price we have added average price label ,also we have sorted over graphs in descending order of the average prices.
* In the graph we can observe that Riverdale has the highest average price with 442.09 dollars and city Island follows with 173 dollars.
* We can observe a huge gap between the top two neighborhoods, which can tell us that Riverdale might have all luxurious rooms.
* There is no big difference between average prices among the rest top 9.

**Worksheet 7: Average reviews by room type and neighbourhood group.**

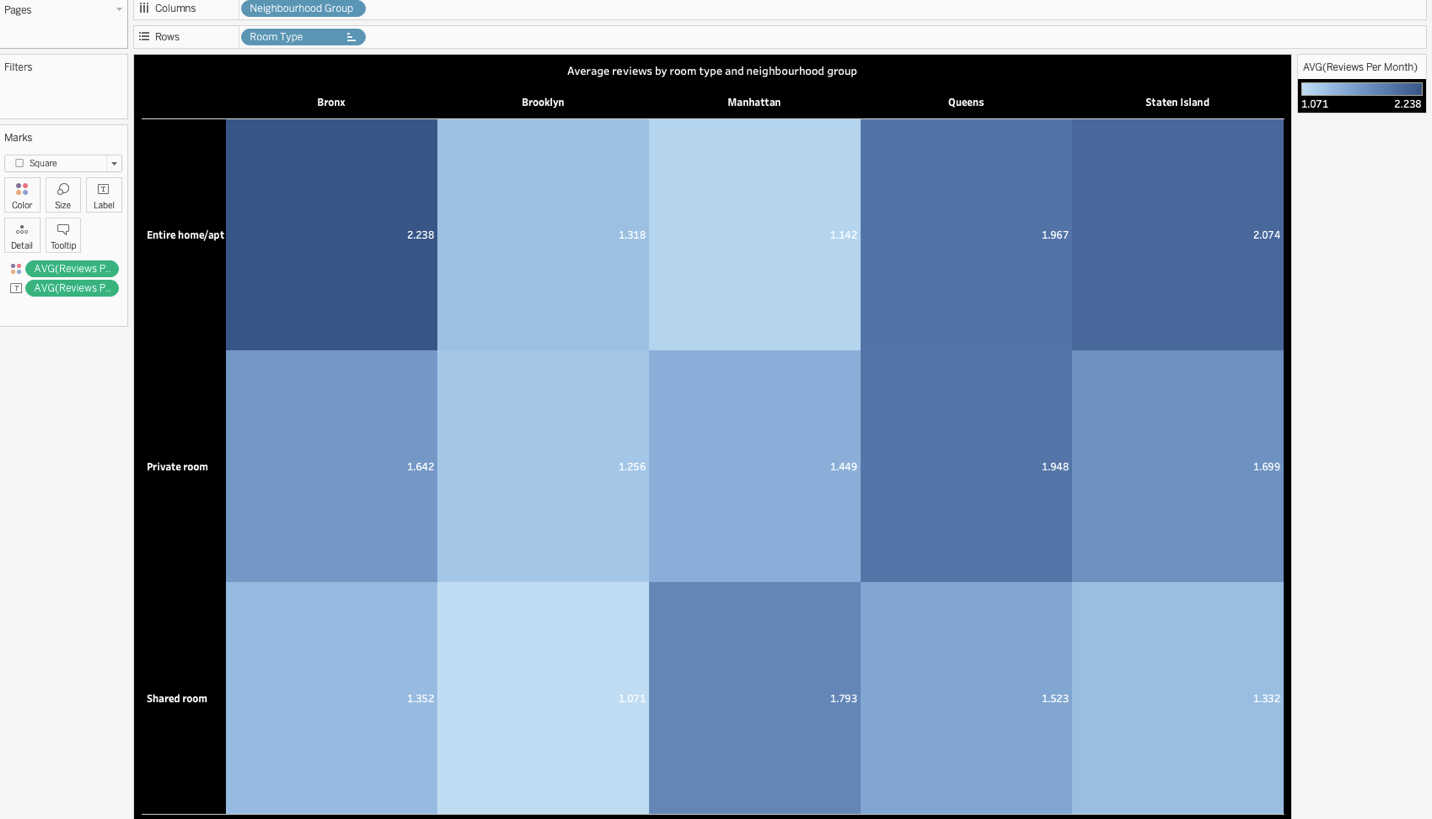
****

Figure 17: Tree map of avg reviews by room type and neighbourhood group

* The picture shows the tree map of Avg reviews by room type and neighbourhood group.
* To generate this tree map we have added Neighbourhood Group to the columns shelf and Room type in Rows shelf.
* For the purpose of showing the label we have added average reviews to the label section, also to add legend we have added avg reviews to color section.
* In the chart in the left side we have Apartment types namely Entire apt, Private room, shared room .
* On the x axis we have neighbourhood groups .
* From the graph we can observe that for Entire Apartment ,Bronx neighbourhood has the highest average review and Brooklyn has the lowest reviews.
* For Private rooms Queens has the highest average review and Brooklyn has the lowest average review
* For Shared room Manhattan has the highest average review and Brooklyn has the lowest average review.

**Worksheet 8: Bookings by neighbourhood groups and room type.**

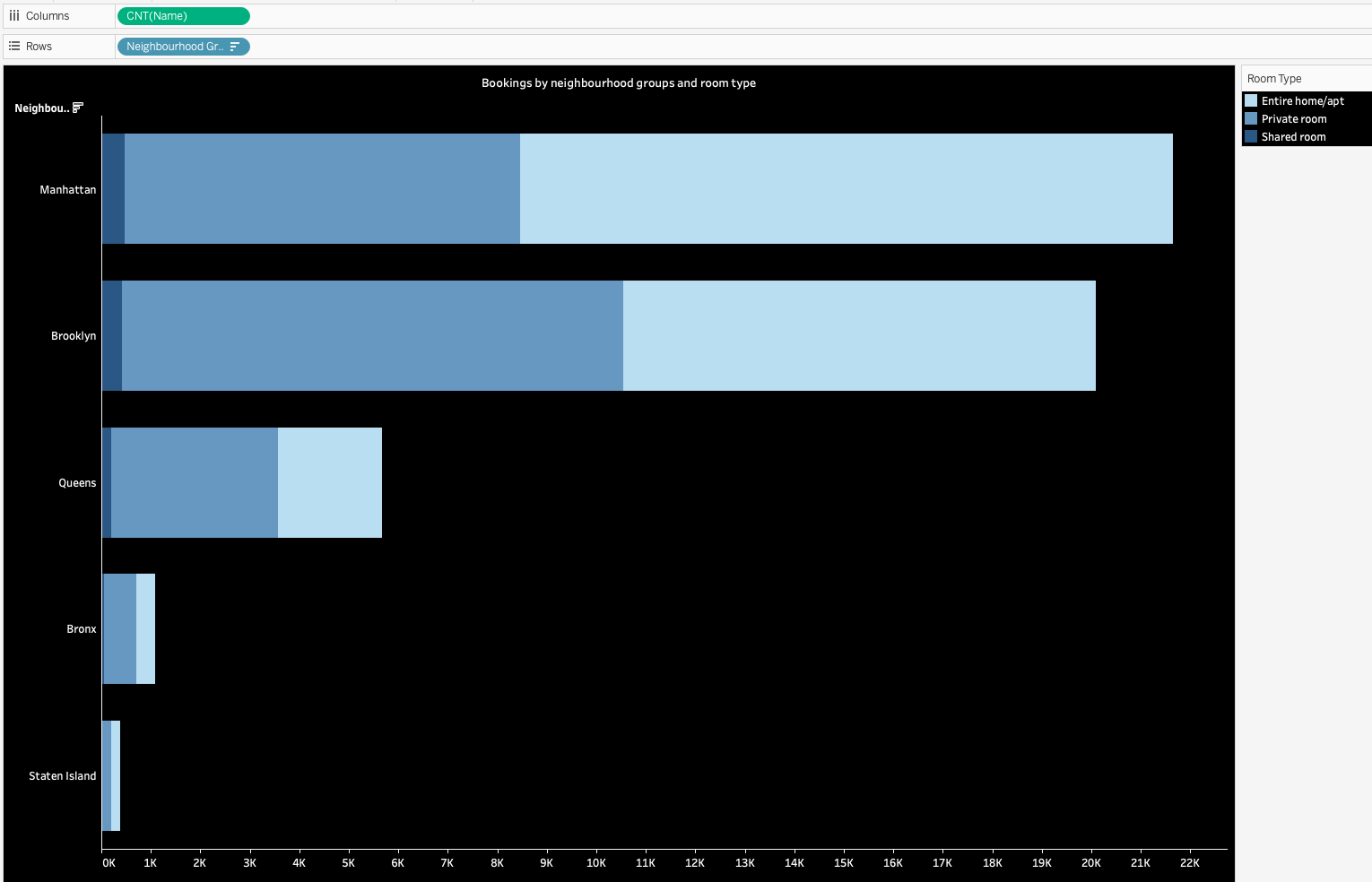


Figure 18: Stacked Bar chart of bookings by neighbourhood groups and room type.

* The above chart shows stacked bar chart of bookings by neighbourhood groups and room type.
* To generate this stacked bar chart we have put Neighbourhood groups in Rows shelf and count of name in column shelf.
* we have neighbourhood groups on the y axis and no of bookings in the x-axis.
* We have selected blue legend to differentiate between neighbourhood groups.
* From the graph we can observe that Manhattan has the highest no of bookings with 22000. Most people preferred entire apartments in Manhattan .
* Brooklyn has the second highest booking among the neighbourhood and In Brooklyn majority of people chose private rooms for the stay
* Staten Island has seen the lowest no of total bookings which is less than 500 .
* Considering business ,it is better to setup business in Manhattan or Brooklyn.

**Worksheet 9: Average price of neighborhoods and room type.**

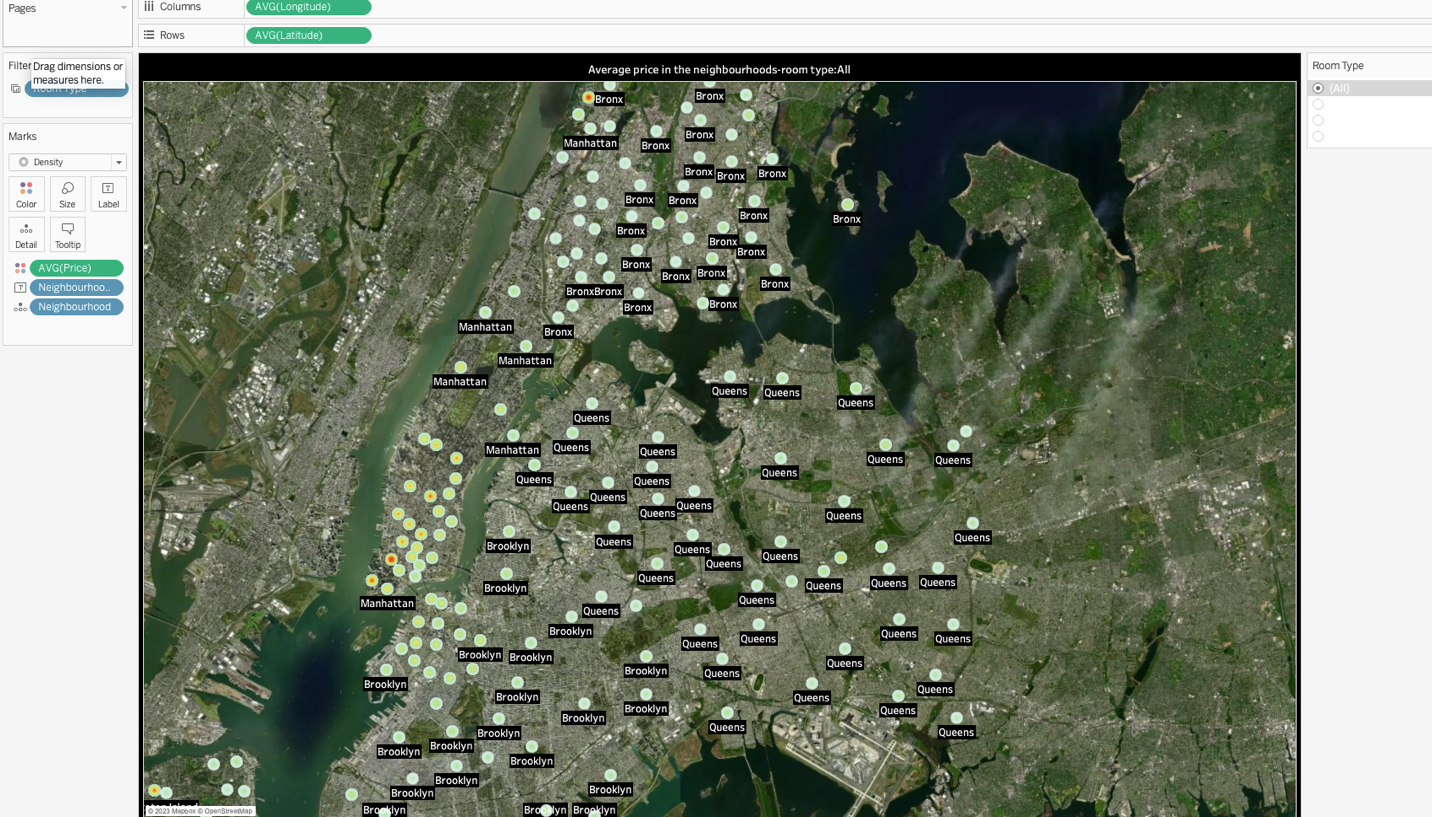
****

Figure 19: map chart of average price of neighbourhoods and room type.

* The above picture shows the map chart of the neighbourhood .
* The picture shows the distribution of different types of rooms over the New york.
* From the above map chart we used density mapping based on the average prices.
* The lightest color shows lowest average price and the most dense and bright color dot shows the costliest room in the area.
* From the graph we can observe that Manhattan has the brighter spots meaning the average price of room in Manhattan is high
* Also Bronx has the most lightest spots meaning ,less average prices in the area.
* Overall from the above graph we can say that Manhattan has the highest average prices and Bronx has lowest average prices making it most affordable to stay.

**Dashboard: Price Predictions and Regional Characteristics in Airbnb Data**

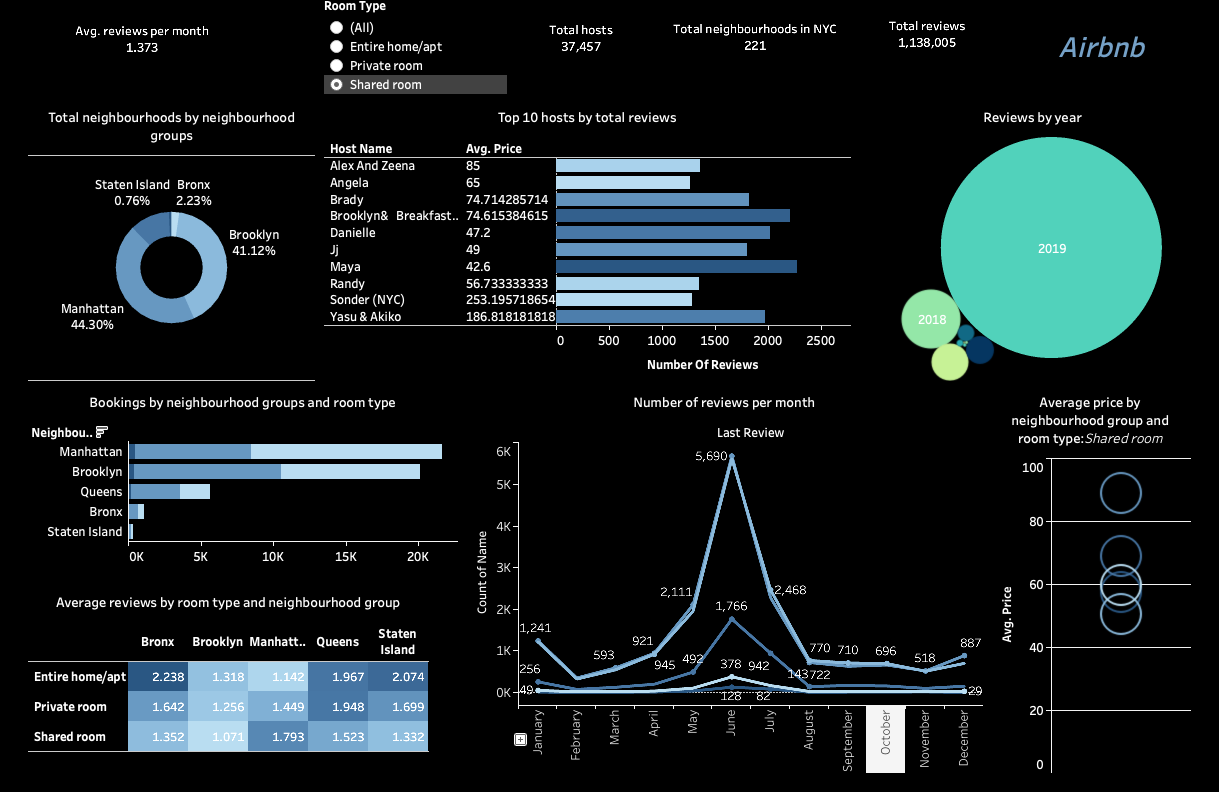


Figure 20: Dashboard

* The above picture shows the dashboard of our dataset, which includes all the graphs and charts like pie chart, Bubble chart, tree map, line chart, stacked bar chart and etc that we created above.
* In the above dash board we have included 4 key performance indicators(KPI).
* First one is Average reviews per month :1,373.

This is the average number of reviews that Airbnb rooms got everymonth.

* Total hosts: 37,457.

This is the total no of hosts present in our dataset.

* Total Neighborhoods in NYC:221.

It is the total no of Neighborhoods present in our dataset.

* We could have added one more KPI i.e Total Neighbourhood groups which will be 5.
* We also have one filter present in our dashboard through which we can change the room types. When changes are made in the filter respective charts and graphs with that particular data also change.

***From the dashboard the Quick Insights that we get are:***

* From Donut chart we can observe ,Manhattan has the most number of neighborhoods and Staten island has the lowest
* From stacked bar chart we can observe, Manhattan has the highest total number of bookings and Staten island has the least total number of bookings.
* Sonder is the best host in the neighbourhood and Maya is the second best host. Also the average prices of the top 10 hosts is less than 90 dollars , from which we can observe that , people are more satisfied with affordable prices.
* From multiline chart we can observe that June month has seen the peak in reviews , this might be because of holiday season.
* From lollipop chart we can observe that , Manhattan has the overall highest average prices .
* From bubble chart we can observe that 2019 have the highest number of reviews and the no of reviews is increasing steadily over years ,from which we can know that the business is gaining popularity and growing steadily.

**Use Cases:**

**1)Host Pricing Strategies:** Hosts can leverage the this analysis and gain average pricing insights in the neighbourhood, for a particular room type and list their prices accordingly to improve their business.

**2)Estate Investment Decisions:** Investors those who want to setup new business can gain insights from this analysis and setup business in the most happening places. So it can help in decision making.

**3)Market Expansion Analysis for Airbnb:** Airbnb’s corporate team can identify potential markets for business expansion by analyzing the neighborhoods popularity and understanding where users are preferring to stay.

**4)Tourism Industry Competitiveness:** Hotels and other accommodation providers can compare Airbnb's pricing and occupancy trends with their own to stay competitive. This could influence promotional offers, pricing during peak seasons, and service offerings.

**5)Consumer Travel Planning:** Travel agencies and trip planning websites can use these insights to help clients make informed accommodation choices.

**6)Business Expansion for Auxiliary Services:** Small business like cleaning services and property management services can use these insights to setup these their business in the most happening places.

**7)Customer Experience Enhancement:** Airbnb itself can use this analysis to enhance user experience by offering personalized recommendations based on popular neighborhoods and price points preferred by similar user profiles.

**Conclusion:**

The Airbnb platform has seen significant growth and preference among users for affordable accommodation options, with unique patterns of popularity and pricing that vary by neighborhood and time of year. Affordable pricing, particularly by top-rated hosts, drives higher satisfaction and repeat bookings, while Manhattan emerges as a premium market with higher average prices. The consistent rise in reviews over time, particularly in busy months like June, suggests that people are becoming more and more confident in Airbnb as a service. These insights can be a strong starting point for a number of strategic initiatives, such as market expansion, improved customer experiences, and host pricing adjustments and real estate investments.

**Github Repository:** https://github.com/KothurDineshReddy/Airbnb\_Data\_visulization

**REFERENCES:**

1.Kaggle Dataset : <https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data/data>

2.Lecture Notes from DATA 230.

3.Medium articles:

<https://medium.com/@blog.linhcao/a-data-driven-look-at-airbnb-in-nyc-market-trends-insights-and-best-practices-b672d16ec0c9>