EDS 6397 – INFORMATION VISUALIZATION

Semester Project

Group –8

Unveiling New York Airbnb Trends

Aiswarya Lalichetti	2252761
Kowsalya Chowdary Nannuri	2251454
Manogna Sai Talluri	2166179
Rohith Bejjam	2138034
Sai Kiran Kandulapati	2156539
Vaishnavi Kulkarni	2251980
Viniktha Gadde	2242893

Objective: We are examining Airbnb listings in New York to understand the factors driving their popularity. and to help users make smarter Airbnb choices.

Dataset Overview:

The dataset contains 102600 rows and 26 columns, some of which directly correspond to the analytical directions provided. For the analysis, we retained only the relevant columns that correspond to potential analytical directions and removed the remaining columns.

Cleaning the dataset:

The data cleaning steps we undertook were critical in refining the dataset to a state that ensures reliability and validity in our subsequent analytical procedures. The steps we performed to clean our dataset are listed below:

Step1: Loading the Dataset:

The dataset was loaded into a pandas Data Frame to facilitate cleaning and manipulation.

Step 2: Initial Review and Column Selection:

The dataset was initially reviewed to understand the columns and types of data present. Columns were selected based on their relevance to the specified analytical directions.

Step 3: Columns Removed:

country, country code columns are Removed due to redundancy, as the dataset pertains to New York and does not require country-level differentiation. 'license' Column is also removed as there is no data mentioned in that column at all and it is irrelevance to the analysis, because it did not contribute to the specified analytical goals.

Step 4: Handling Missing Values:

Rows with missing values in neighbourhood, neighbourhood group, latitude, and longitude columns were removed. The percentage of missing values in these columns are exceptionally low, which is under 0.03%. In this case, dropping these rows could be the best approach as it maintains the integrity of the geographical analysis without significantly reducing the dataset size.

Step 5: Data Type Conversion:

The missing values in "last review "column were replaced with a placeholder date (1900-01-01) to allow for consistent data formatting and to maintain the integrity of the dataset for any future time-based analysis that might be conducted and also to avoid NaN values that could disrupt time-based calculations or visualizations. Commonly used placeholder dates might be 1900-01-01, 1970-01-01, the Unix epoch start date.

Step 6: Cleaning Text and Numerical Data:

The price and service fee columns were cleaned to remove non-numeric characters (like $ and commas) for accurate numerical analysis.

The missing values in the column "NAME" was replaced with place holder text "Listing name not provided" and similarly missing "host name" column entries were replaced with "Host Name Not Provided" and another column "host_identity_verified" is replaced with "unconfirmed" because If we take a conservative stance, we could assume that unless a host is explicitly verified, their status should be

'unconfirmed'. The misspelled neighborhood names 'Manhatan' and 'Brookyln' were corrected to 'Manhattan' and 'Brooklyn.

Step 7: Imputing Missing Numerical Data:

Missing values in other numerical columns were imputed with the median of the respective column to preserve the central tendency of the dataset without the influence of outliers.

Step 8: Categorical Data Consistency:

Categorical columns like instant_bookable, cancellation_policy, and room type were checked for consistency. Missing values were imputed with the mode of the column.

Step 9: Removing Duplicate Entries:

The dataset was checked for duplicate entries using unique identifiers like id and host id, and any duplicates found were removed to ensure the uniqueness of each listing.

Step 10: Final Data Type and Integrity Check:

A final review of the data types for each column was conducted to ensure they were appropriate for the type of analysis to be performed.

Step 11: Saving the Cleaned Dataset: The cleaned dataset was saved to a new CSV file, preserving the cleaning and formatting that had been applied.

Potential Analytical Directions

1.Geographical Analysis:

Explore the distribution of Airbnb listings across different neighborhoods in New York. Identify popular areas for Airbnb accommodations and their characteristics.

No.	Columns	Rows	Filters	Parameter
1.	Longitude	Latitude	Room Type, Neighborhood Group, Is in Selected Range	Price Range

Details of Calculated Field:

Is in Selected Range | airbnb_data_with_binary_house_rules

```
CASE [Price Range]
  WHEN 1 THEN [Price] >= 50 AND [Price] <= 280
  WHEN 2 THEN [Price] >= 281 AND [Price] <= 510
  WHEN 3 THEN [Price] >= 511 AND [Price] <= 740
  WHEN 4 THEN [Price] >= 741 AND [Price] <= 970
  WHEN 5 THEN [Price] >= 971 AND [Price] <= 1200
  WHEN 6 THEN [Price] >= 50 AND [Price] <= 1200
  ELSE FALSE
END
```

Using the details above, we created a geographical analysis dashboard with each chart conveying the following information.

Chart1: Map

Title: "New York's Airbnb Landscape: A Neighborhood Overview."

This chart displays the distribution of Airbnb listings across various neighborhood groups, categorized by different price ranges as depicted in the calculated field.

Chart 2: Bar Chart

Title:" Comparing prices of Airbnb listings across different Neighborhood Groups."

This chart highlights the quantity of Airbnb listings categorized by neighborhood group and distributed across different price ranges.

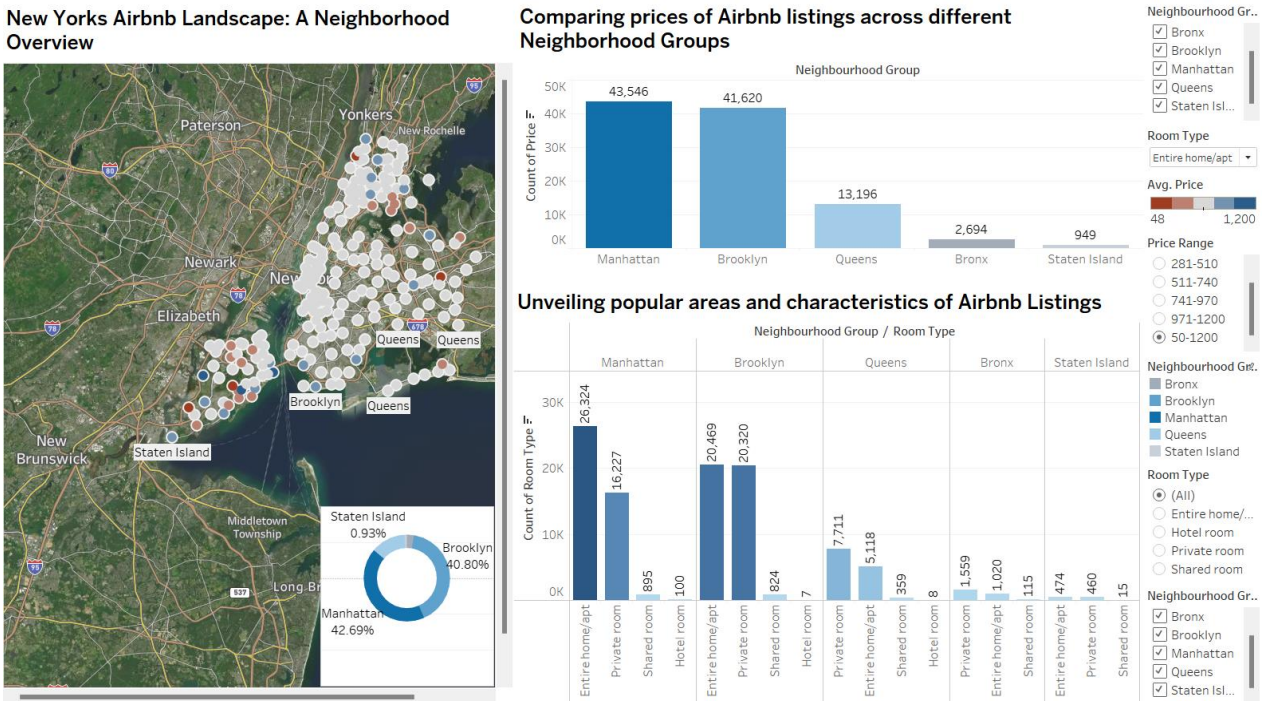
Analysis: The Manhattan neighborhood group exhibits a higher number of Airbnb listings compared to Brooklyn, Queens, Bronx, and Staten Island. This indicates a discernible popularity pattern among the five regions in New York.

Chart 3: Bar Chart

Title: "Unveiling popular areas and characteristics of Airbnb Listings."

This chart illustrates how room types are distributed within each neighborhood group.

Analysis: Within each neighborhood group, a distinct pattern is observed in the distribution of room types, with the common trend being: Entire home/apt > Private room > Shared Room > Hotel Room.



Plot1: Exploring Geographical Patterns and its characteristics

2. Listing Review Rates and Popularity:

Analyze the relationship between review rates, the number of reviews, and the overall popularity of listings. Do high review rates significantly impact listing popularity?

No.	Columns	Rows	Filters
1.	Review Rate Number	Popularity, Number of Reviews	Popularity, Number of Reviews

Title: "Unveiling the Influence of Review Rates on Listing Popularity."

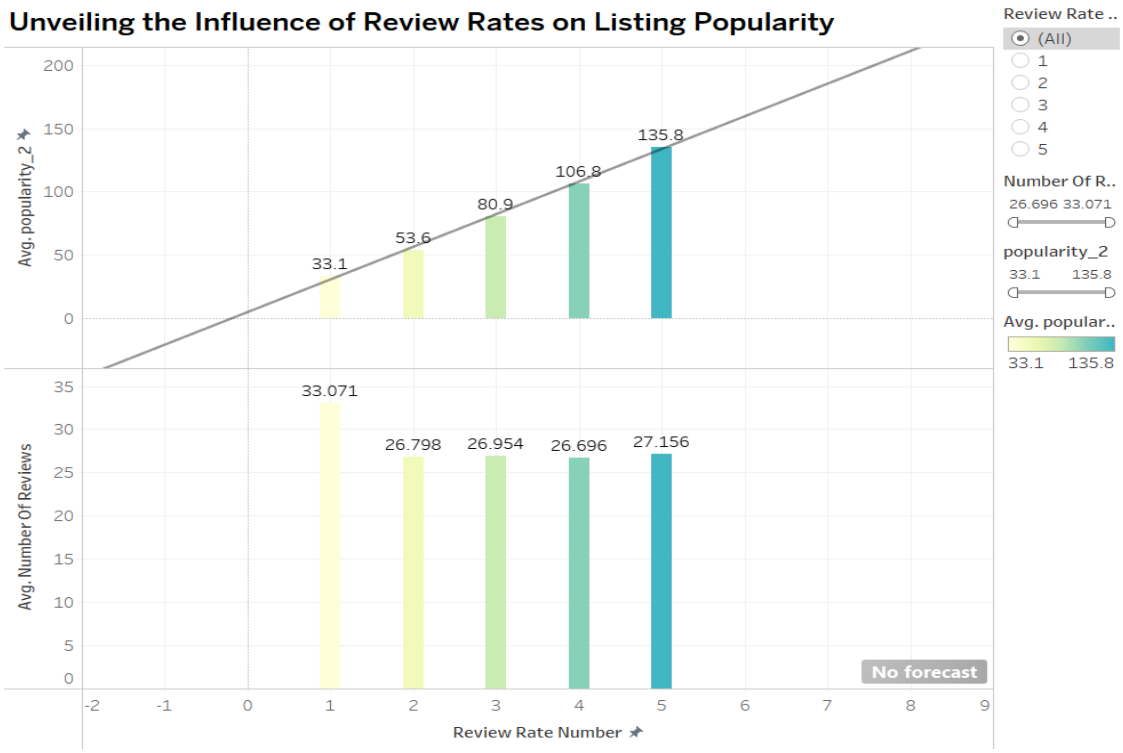
Chart 1: Bar Chart

We formulated a calculated field for popularity by multiplying the review rate number and the number of reviews.

Analysis: The bar plot depicts the relationship between average popularity and the number of review rates, and it suggests a direct proportional connection between popularity and review rate number. The trend line also depicts the same proportionality relationship.

Chart 2: Plot between average Number of Reviews and Review Rate Number.

Analysis: No correlation can be identified, as all five review rates fall within a similar range.



Plot2: Evaluating Popularity Through Review Rate Number and Number of Reviews

3. Pricing and Minimum Nights:

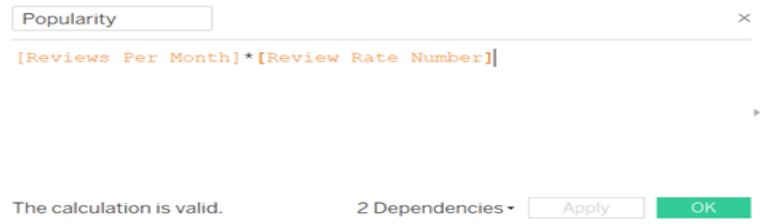
No	Columns	Rows	Filters	Detail	Color
1.	Price	Availability 365 and Minimum Nights	Popularity	Name	Availability 365(Red) Minimum Nights (Light Teal)

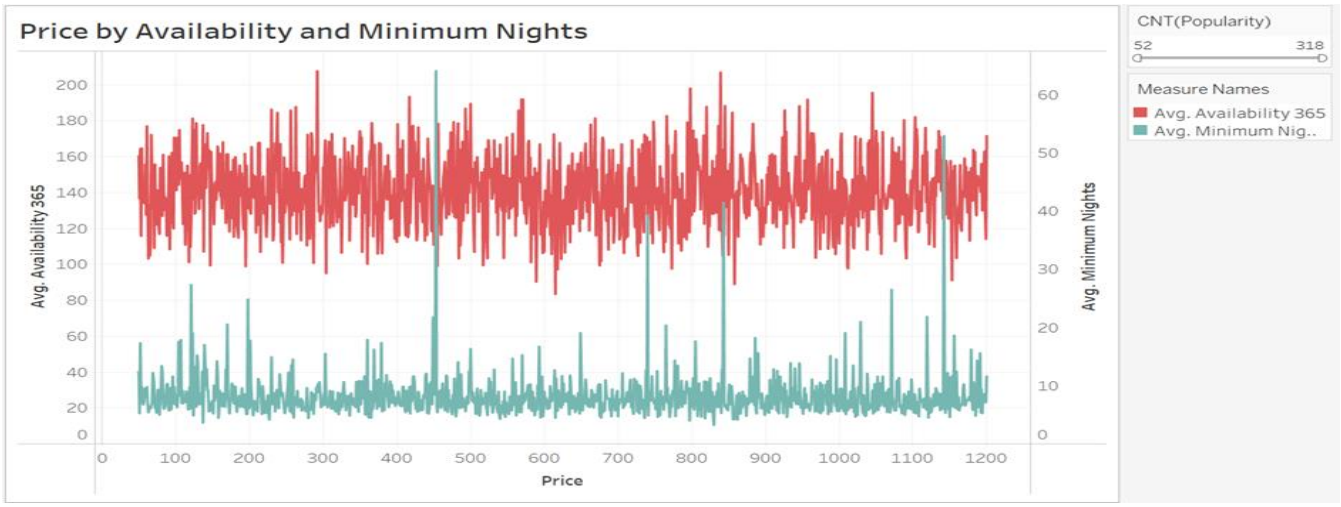
Title: " Analyzing the Price by Availability and Minimum Nights "

Chart: Dual Lines Chart

Analysis: It employs a Dual Lines Chart to visually represent the relationship between pricing and key metrics. Popularity is defined as the product of reviews per month and review rate number. Using this popularity metric as a filter, the resulting dual lines chart illustrates the interplay between Availability and Minimum Nights, offering a comprehensive view of how these factors influence pricing dynamics.

Details of calculated field (Popularity):





Plot3: Price by Availability and Minimum Nights

4. Host Performance Metrics:

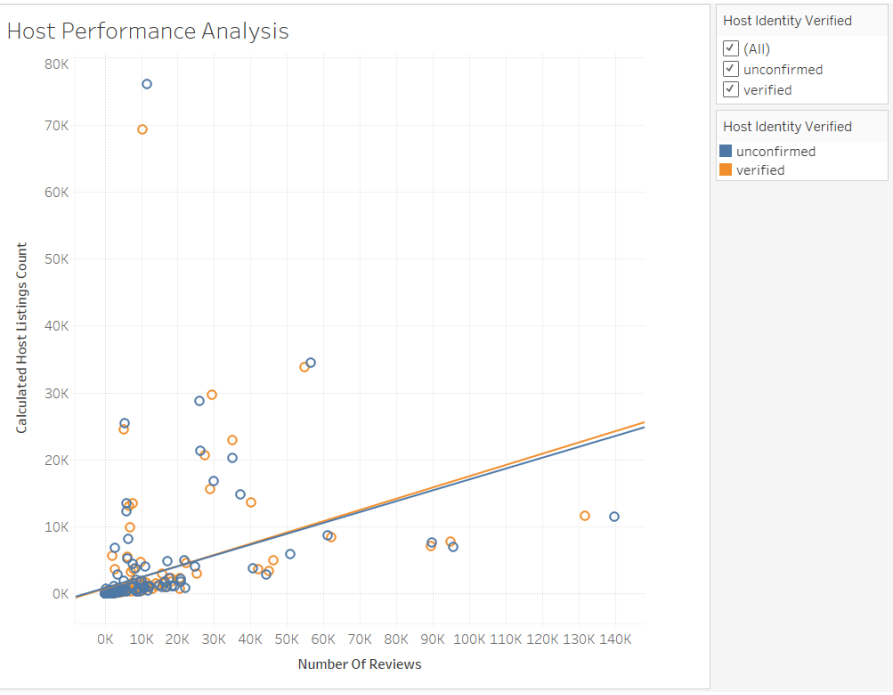
Analyze metrics related to hosts such as calculated host listings count, host identity verification, and the average number of listings. How do these variables correlate with listing success?

No	Columns	Rows	Filters	Color	Details
1.	Number of Reviews	Calculated Host Listing Count	Host Identity Verified	Host Identity Verified	Neighborhood, Neighborhood group

Chart: Scatter Plot

Title: "Unveiling the Correlation Between Host Characteristics and Airbnb Listing Effectiveness"

Analysis: The below is the scatter plot showing the relationship between the calculated host listings count and the listing success rate for Airbnb listings in New York City. The scatter plot shows that there is a negative correlation between the calculated host listings count and the listing success rate. This means that as the number of listings a host has increases, the likelihood of a listing being successful decreases. Overall, the scatter plot suggests that there is a relationship between the calculated host listings count and the listing success rate for Airbnb listings in New York City.



Plot 4: Impact of Host Listing Quantity on Listing Success

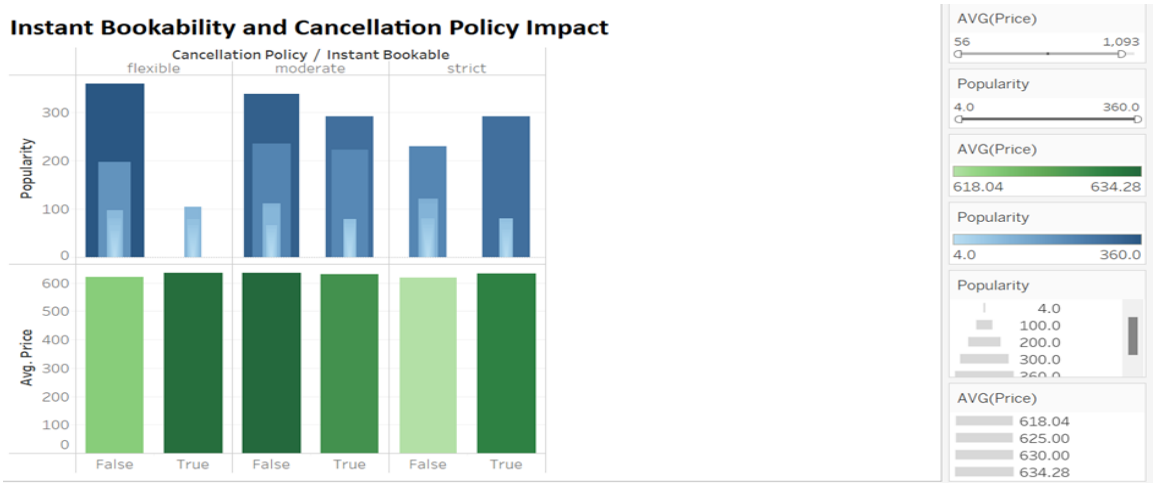
5. Instant Bookability and Cancellation Policy Impact:

No	Columns	Rows	Filters	Size	Color	Tooltip
1.	Cancellation Policy and Instant Bookable	Popularity and Price	Popularity and Price	Popularity and Price	Popularity and Price	Name

Title: "Unraveling the Interplay of Instant Bookability and Cancellation Policies on Popularity and Price Dynamics"

Chart: Bar in Bar Chart

Analysis: This Chart is employed to elucidate the relationships at play. Popularity, determined by the product of reviews per month and review rate number, serves as a filter, facilitating customer decision-making. The chart visually represents how Instant Bookability and Cancellation Policies factor into this decision process, aiding users in making informed choices based on their preferences and priorities.



Plot 5: Price by Instant Bookability and Cancellation Policy

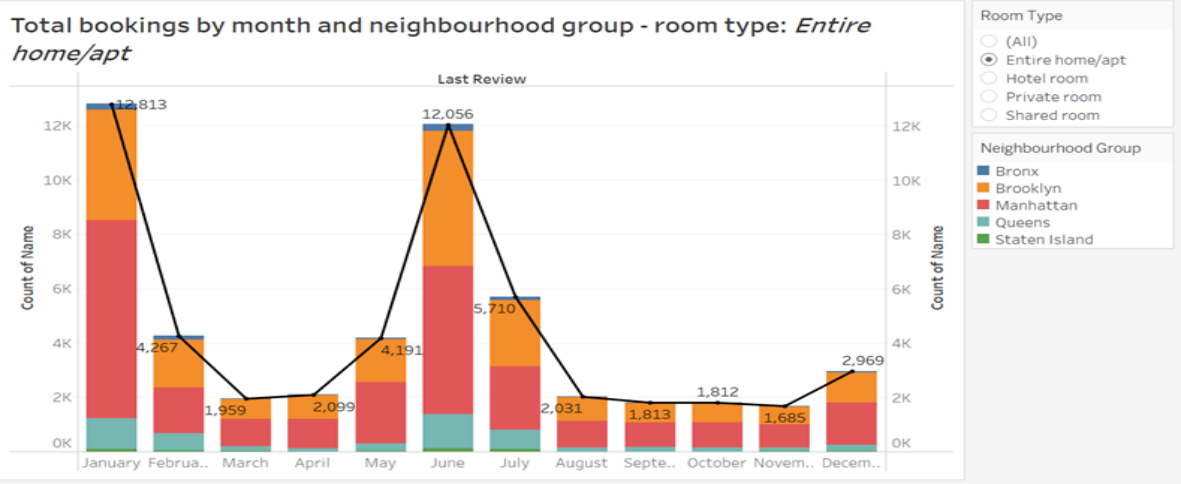
6. Total bookings by month and neighborhood group

No	Columns	Rows	Filters	Color
1.	Month	Count of name Neighborhood Group	Room Type	Neighbourhood Group

Title: "Exploring Monthly Trends and Neighborhood Dynamics**"**

Chart: Bar and Line Chart

Analysis: This analysis presents the total bookings categorized by both month and neighborhood group. Furthermore, it delves into the variance in pricing across different room types within each neighborhood group, providing insights into the relationship between booking patterns, months, neighborhood groups, and room type pricing.



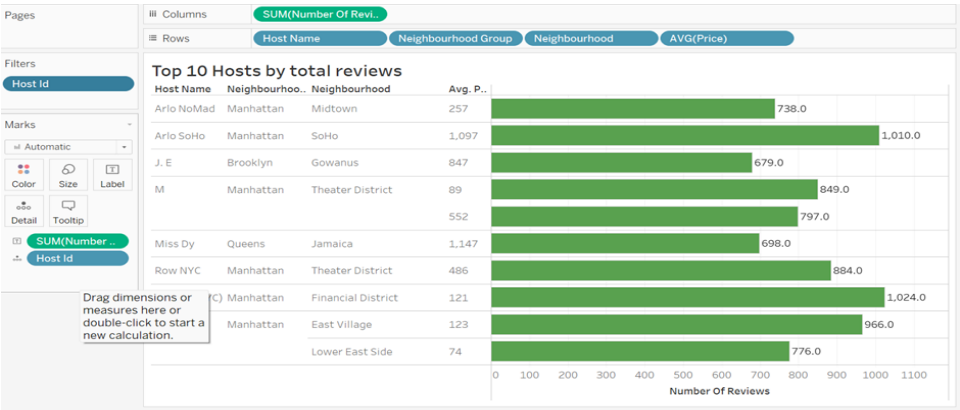
Plot 6: **Total bookings by month and neighborhood group – Entire home/ apt**

7. Top 10 hosts by total reviews.

No	Columns	Rows	Filters	Size	Detail	Label
1.	Number Of Reviews	Host Name Neighborhood Group Neighborhood Price	Host ID	Popularity and Price	Host ID	Number Of Reviews

Title: "A Comprehensive Analysis of Top 10 Hosts Based on Total Reviews and Multidimensional Factors"

Analysis: This structured approach provides a comprehensive evaluation of hosts' performance, visually comparing their total reviews. The horizontal bar chart enhances clarity, offering insights into the influence of factors such as popularity and pricing on host performance.



Plot 7: **Top 10 Hosts by total reviews**

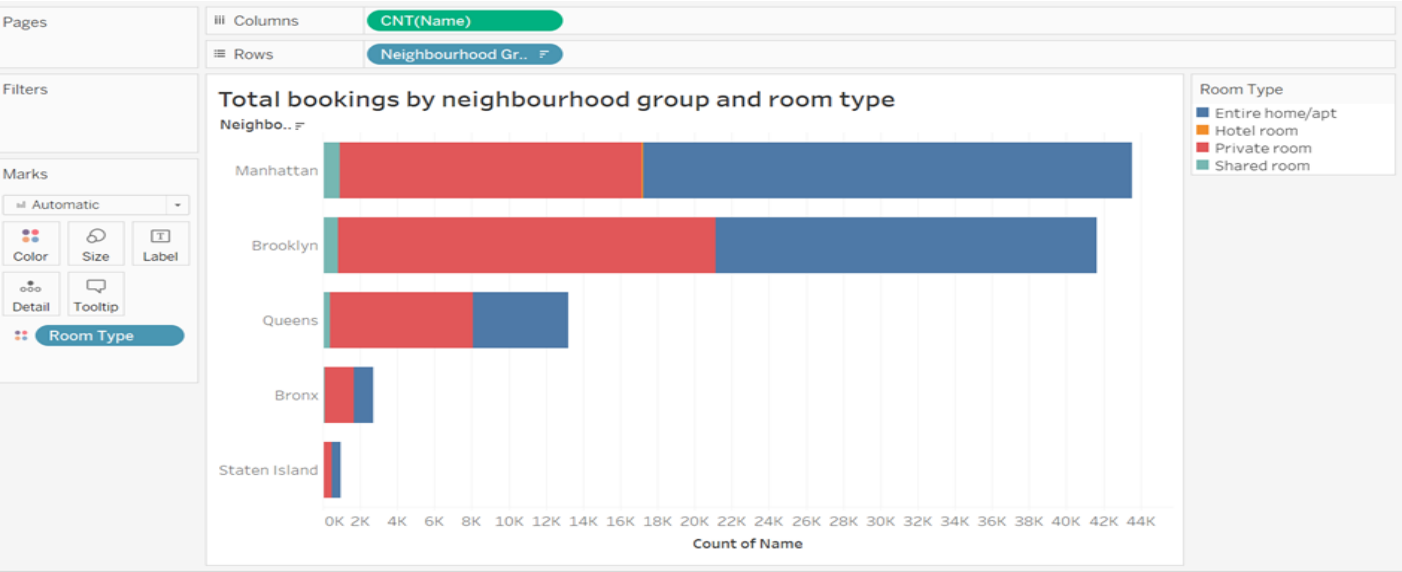
8. Total bookings by a neighbourhood group and room type.

No	Columns	Rows	Color
1.	Name	Neighbourhood Group	Room Type

Title: "Exploring Neighborhood Group and Room Type Interplay**"**

Chart: Horizontal bar

Analysis: Designed a horizontal bar chart to analyze total bookings by combining neighborhood groups and room types. The bars represent total bookings for each combination, and color coding enhances differentiation between neighborhood groups or room types. This visualization offers a concise overview of booking distribution, revealing insights into popularity and demand patterns across diverse neighborhood groups and room types.



Plot 8: Total bookings by neighborhood group and room type

Conclusion: The analysis highlights critical factors influencing Airbnb listing success in New York, such as location, pricing, and host attributes. Findings from this project provide valuable guidance for hosts to enhance their offerings and for guests to make better-informed booking decisions, fostering a more effective Airbnb market in the city.

Dashboard Link:
https://public.tableau.com/views/IV_ProjectGroup8_AirbnbNYC/Overview?:language=en-US&:display_count=n&:origin=viz_share_link