Methodology for New York Airbnb's Dataset

By:
Vrushali Rane

Data Cleaning and Preparation

• **Handling Missing Values** :

There are two columns **name** and **host_name** has missing values **16** and **21,** respectively. These missing values represent a small proportion of the total dataset. We have addressed these missing values with imputation method as '**Unknown**'.

Column Name	Count
id	0
name	16
host_id	0
host_name	21
neighbourhood_group	0
neighbourhood	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
calculated_host_listings_count	0
availability_365	0
location	0

Data Cleaning and Preparation

• **Summarizing Numeric Columns**:

Each column gives statistical information about different aspects of the data, including measures of central tendency (mean, median, mode), dispersion (standard deviation, variance, range), and shape of the distribution (kurtosis, skewness). This data is helpful for understanding the distributions and characteristics of the variables in our dataset.

Summary	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
Mean	152.7206872	7.029962164	23.27446569	7.143982002	112.7813273
Standard Error	1.086070222	0.092756653	0.201474998	0.14902406	0.595246998
Median	106	3	5	1	45
Mode	100	1	0	1	0
Standard Deviation	240.1541697	20.51054953	44.55058227	32.95251885	131.6222889
Sample Variance	57674.02525	420.6826422	1984.75438	1085.868499	17324.42692
Kurtosis	585.6728789	854.0716624	19.52978807	67.5508883	-0.997534045
Skewness	19.118939	21.82727453	3.690634572	7.9331739	0.763407577
Range	10000	1249	629	326	365
Minimum	0	1	0	1	0
Maximum	10000	1250	629	327	365
Sum	7467278	343730	1138005	349305	5514443
Count	48895	48895	48895	48895	48895

Data Cleaning and Preparation

- **Correlation between Numeric Columns:**

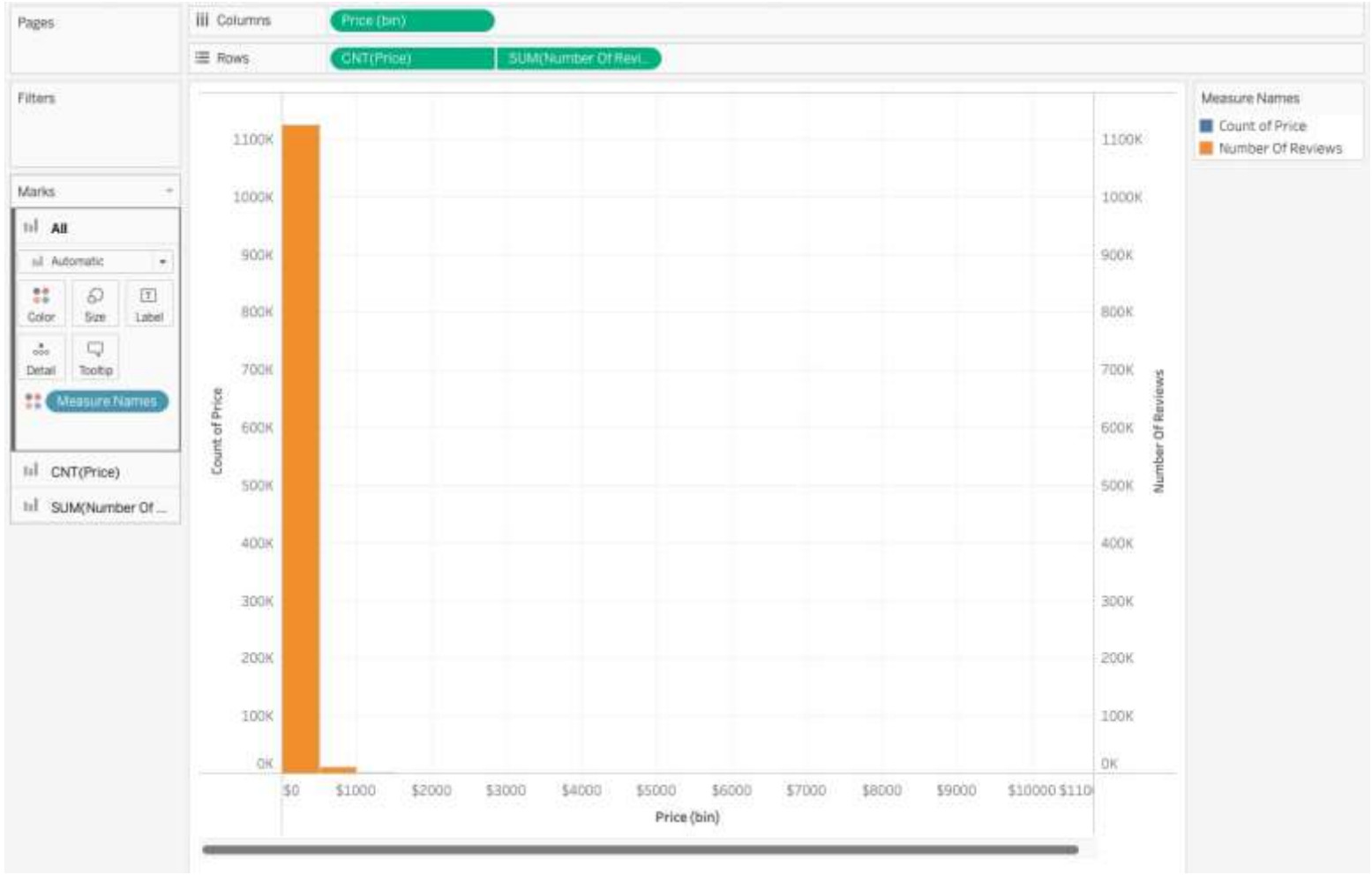
None of the correlations are particularly high, indicating that the relationships between these variables are not strongly linear. However, there are some mild tendencies observed where certain variables show slightly stronger correlations with each other compared to others.

	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
price	1				
minimum_nights	0.043	1			
number_of_reviews	-0.048	-0.080	1		
calculated_host_listings_count	0.057	0.128	-0.072	1	
availability_365	0.082	0.144	0.172	0.226	1

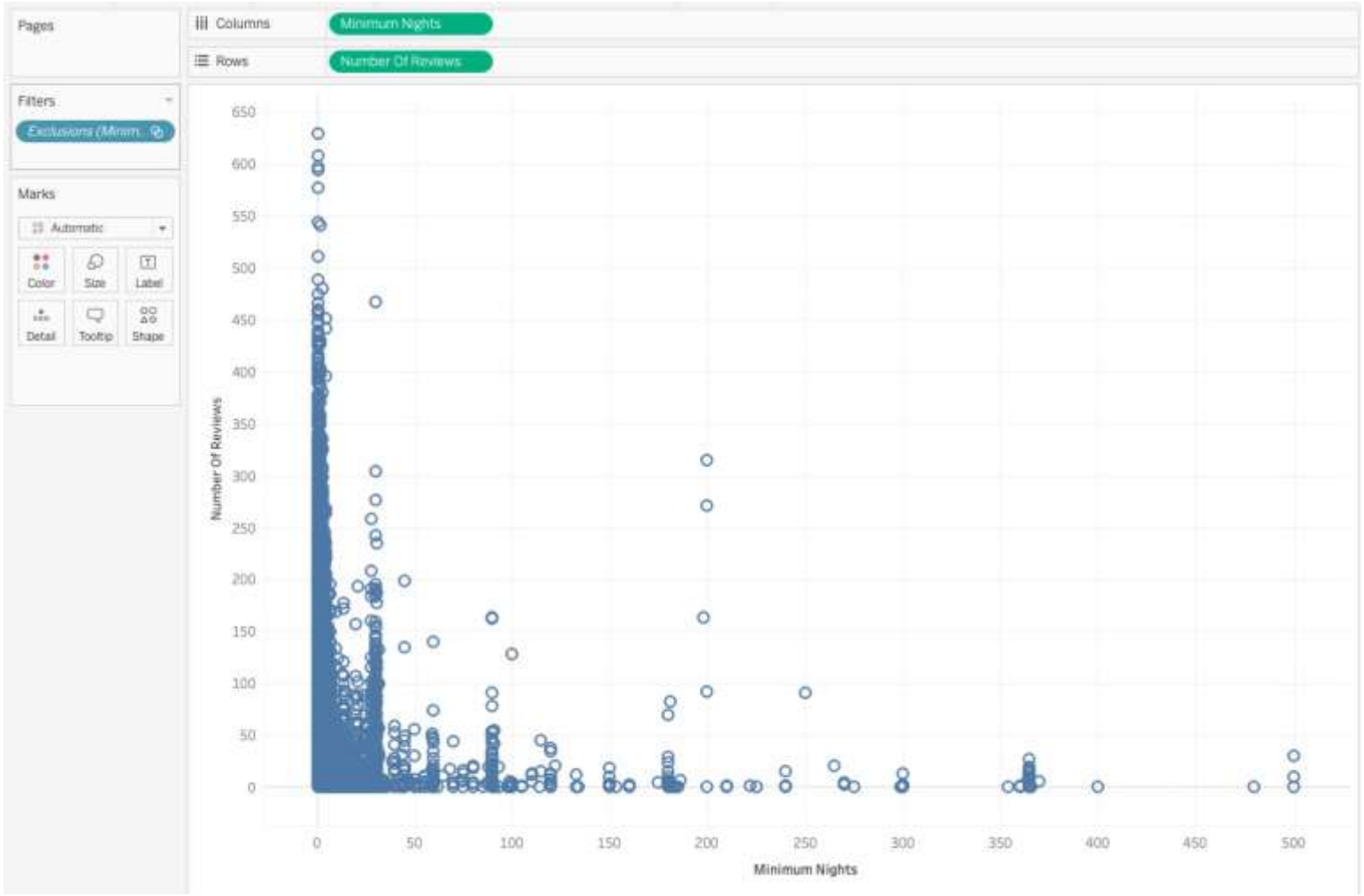
To analyze which type of hosts to acquire more in specific neighborhoods, we created **Treemaps**, utilizing variables such as **Neighborhood**, **Neighborhood Group**, and **Count of Room Types.** We filtered **Top 5 Neighborhoods** based on **Count of Room Type.** It will help in understanding of where and what types of hosts might be more beneficial to acquire.

To analyze which neighborhood to be targeted, we created **Packed Bubbles**, utilizing variables such as **Neighborhood**, **Neighborhood Group**, and **Sum of Number of Reviews.** We filtered **Top 5 Neighborhoods** based on **Sum of Number of Reviews**. It will help in understanding of neighborhood based on the customer's demand and preferences within those areas.

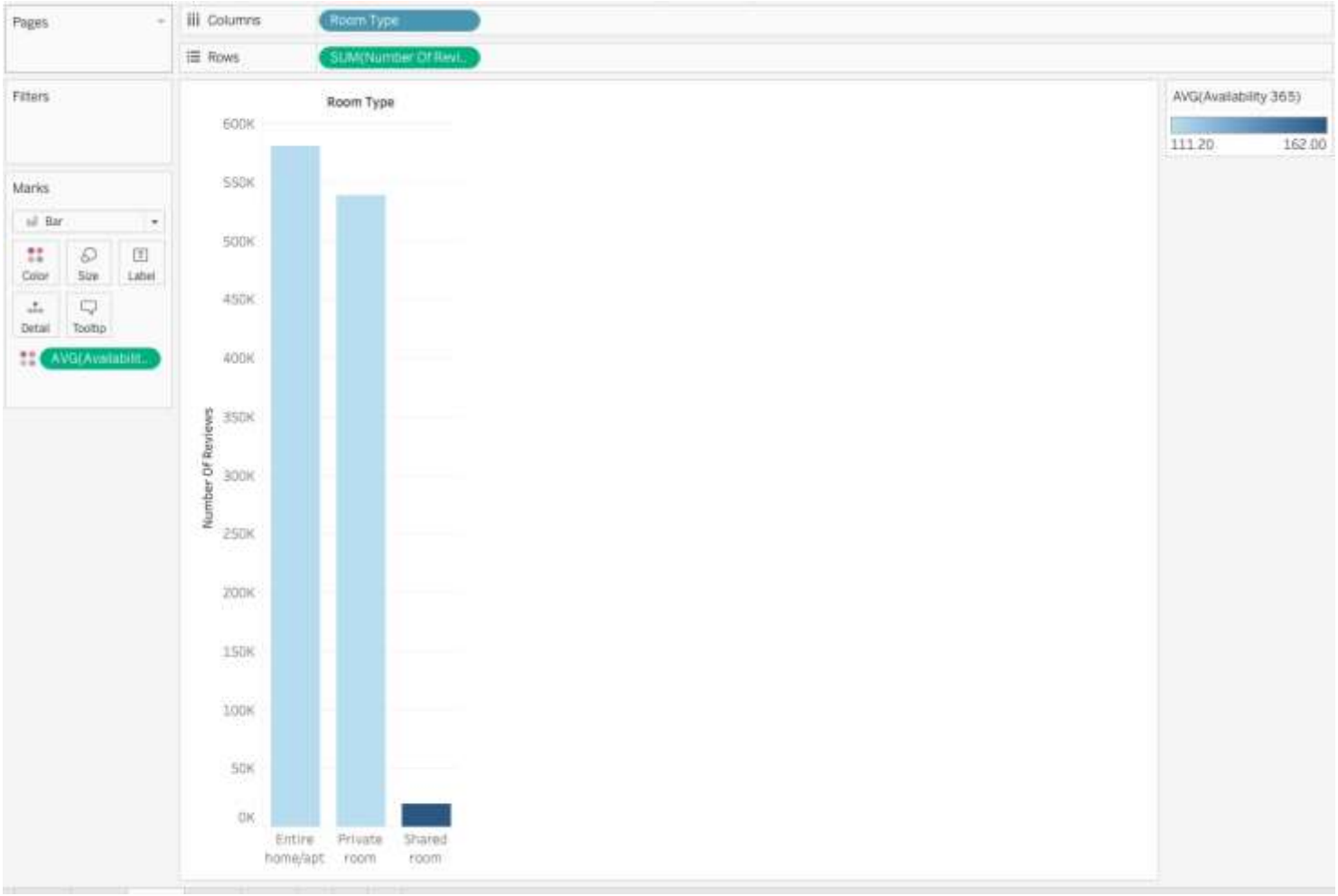
To uncover the preferred pricing ranges by customers, we created a **Dual-Axis Chart** utilizing variables such as **Price(bins), Count of Price, Number of Reviews.** It will help understanding of pricing preferences, budget, aiding in making informed decisions for pricing and service improvements.



To understand customer's stay preferences, we created a **Scatter Plot** utilizing variables such as **Number of Reviews** and **Minimum Nights**. It will help understanding of customer behaviors concerning stay durations and satisfaction, aiding in making informed decisions for pricing and service improvements.



To identify adjustments for existing properties in New York City to make them more customer-oriented, we created **Bar Chart** using variables such as **Room Type**, **Sum of Number of Reviews**, and **Average Availability_365**. It will help in aiding in making targeted improvements and strategic decisions to cater better to customer preferences in New York City.



To identify most popular localities and properties in New York City, we created **Bar Chart** using variables such as **Neighborhood**, **Neighborhood Group, Name**, and **Sum of Number of Reviews**. We filtered **Top Neighborhoods, Name** based on **Sum of Number of Reviews.** It will help in aiding in making targeted improvements and strategic decisions to cater better to customer preferences in New York City.