

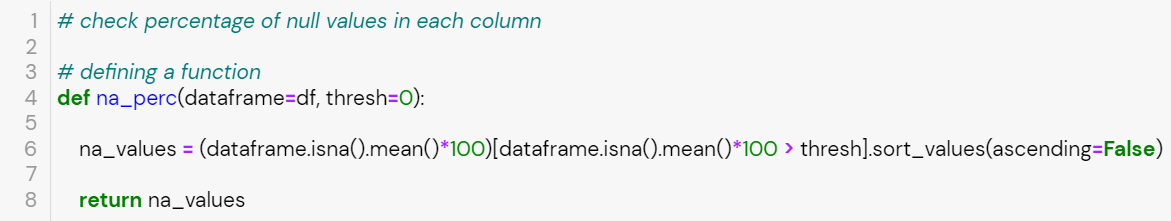
Methodology Document

# 1. Data Preparation

Imported the dataset into a Jupyter notebook to use Python to clean the dataset. This process included the following steps:

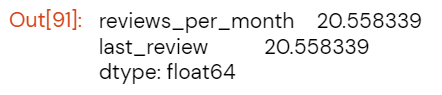
**1.1 Importing libraries and data**

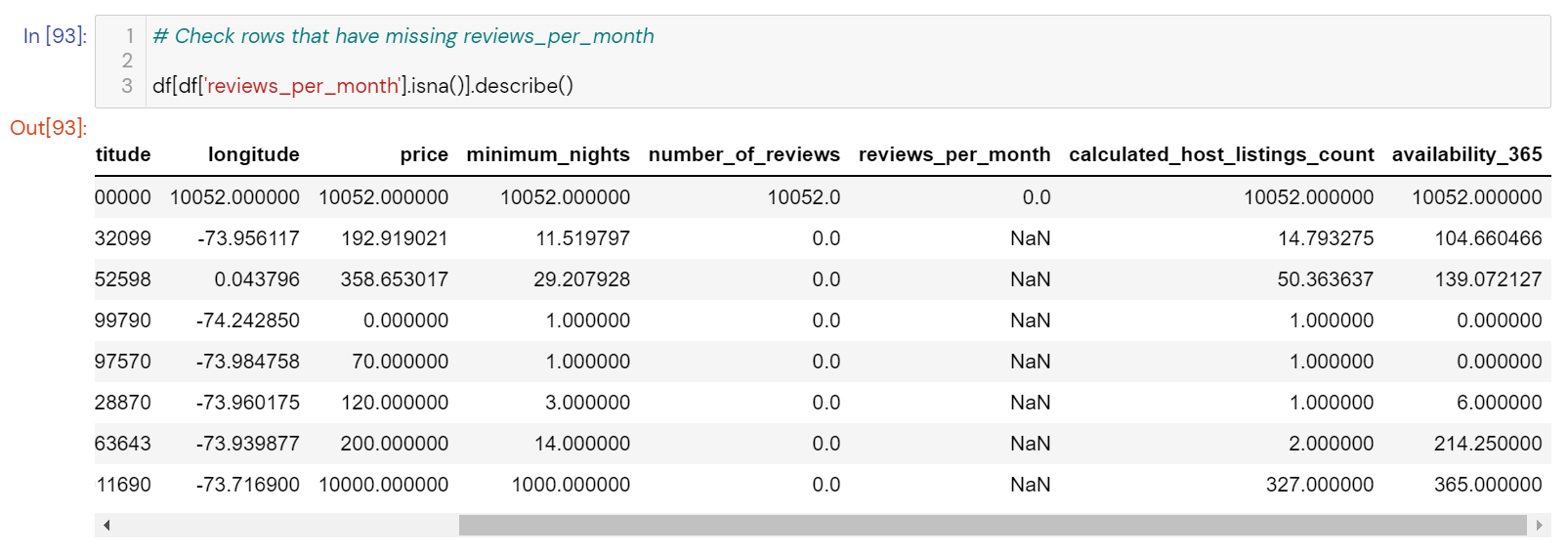
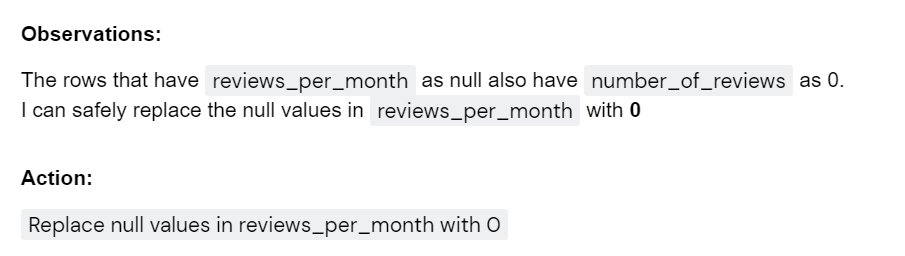
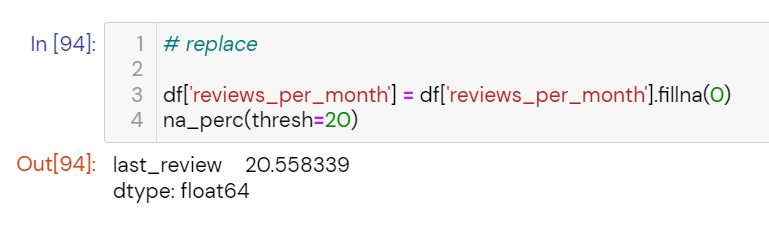
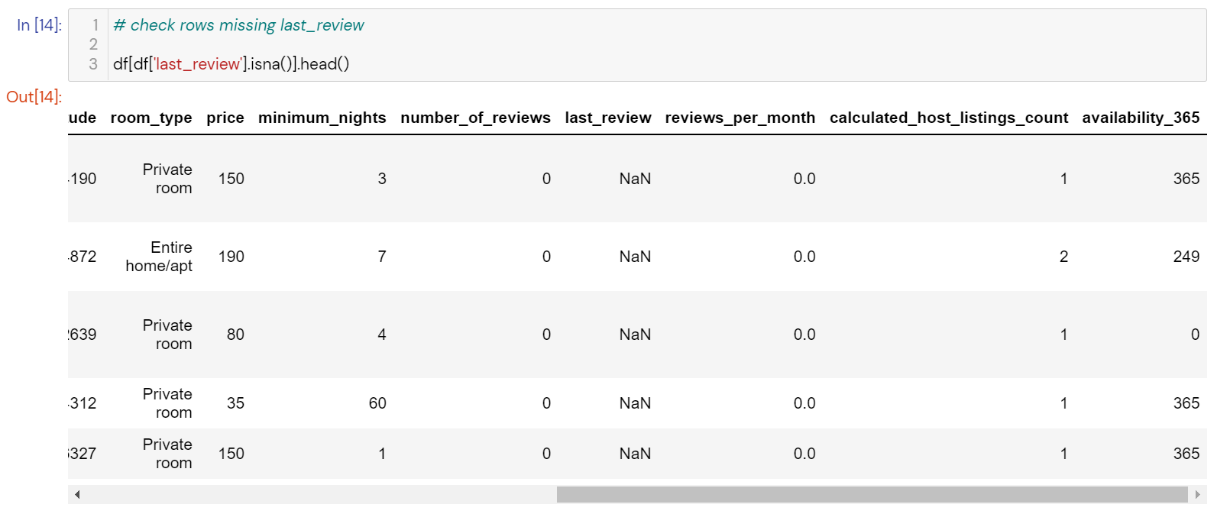
**1.2 Data Exploration**

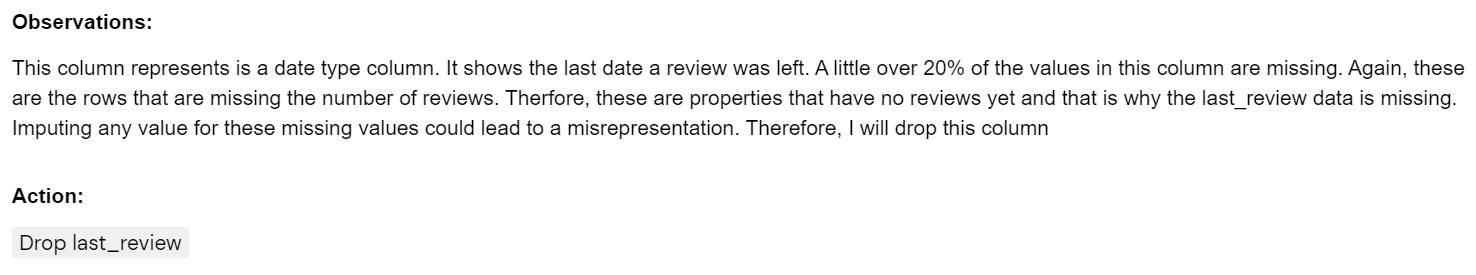
* Check the shape of dataset and the data types of the columns
* Define a function to output the percentage of null values in every column

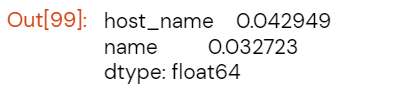
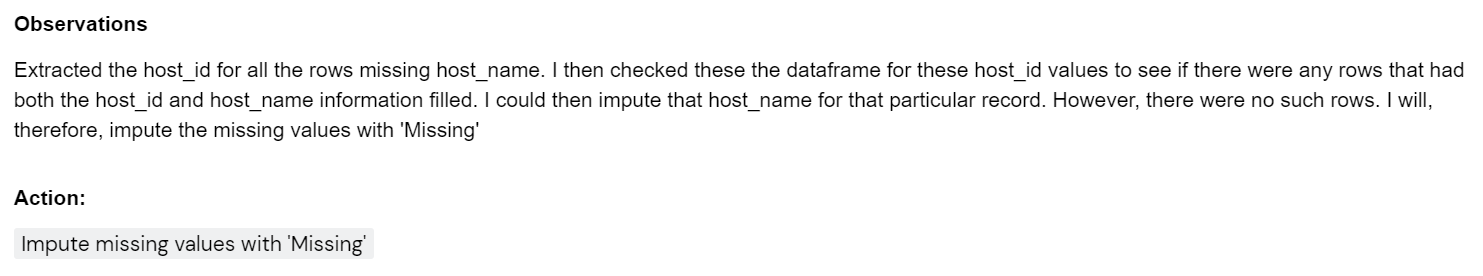
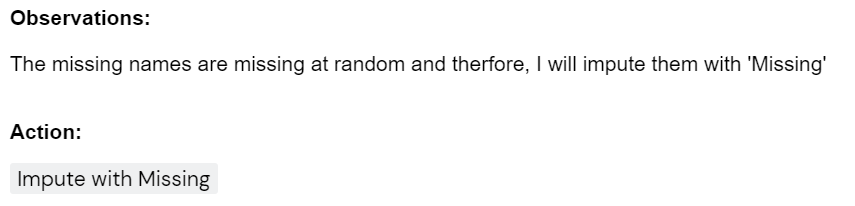
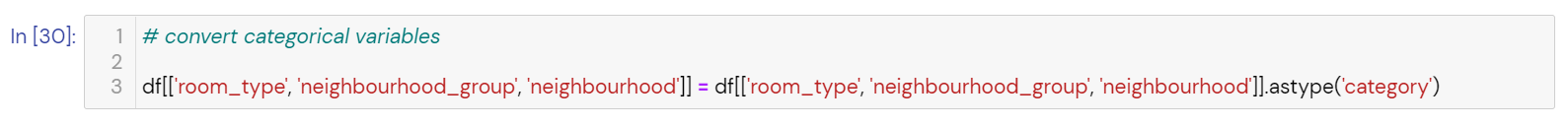
**1.3 Data Cleaning**

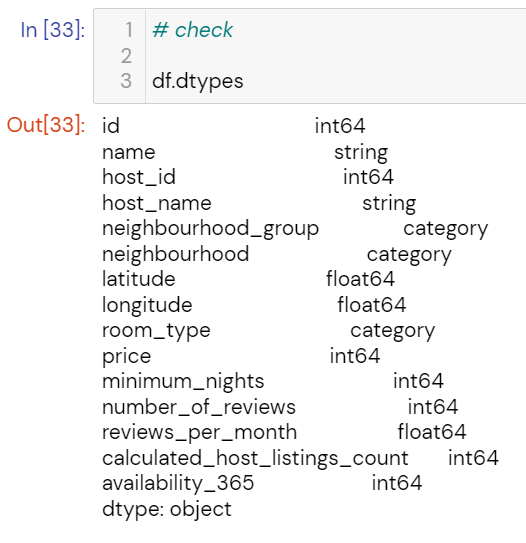
* Check columns that have more than 20% of their values as null values.



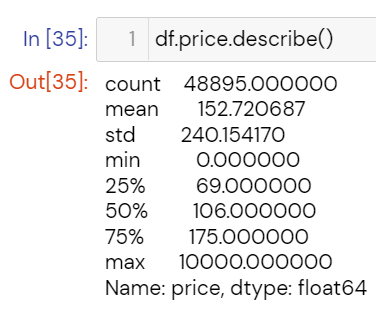
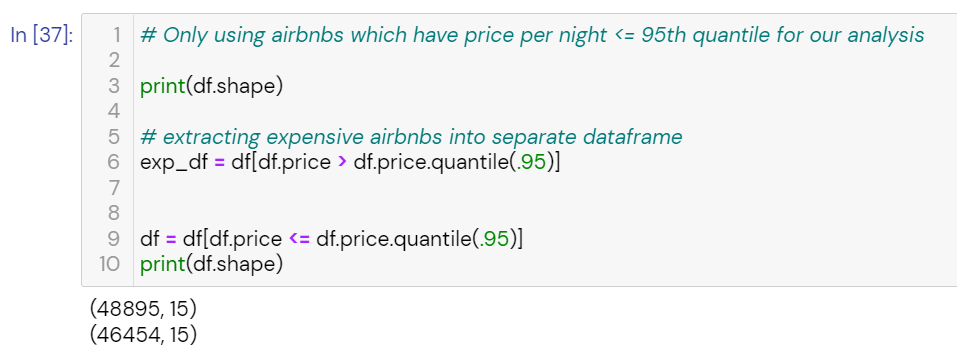
* For reviews\_per\_month:
* For last\_review:



* Check columns that still have null values
* For host\_name:
* For name:
* Converting data types of variables



**1.4 Treating for Outliers**

* For price: There is a significant difference between mean and median.

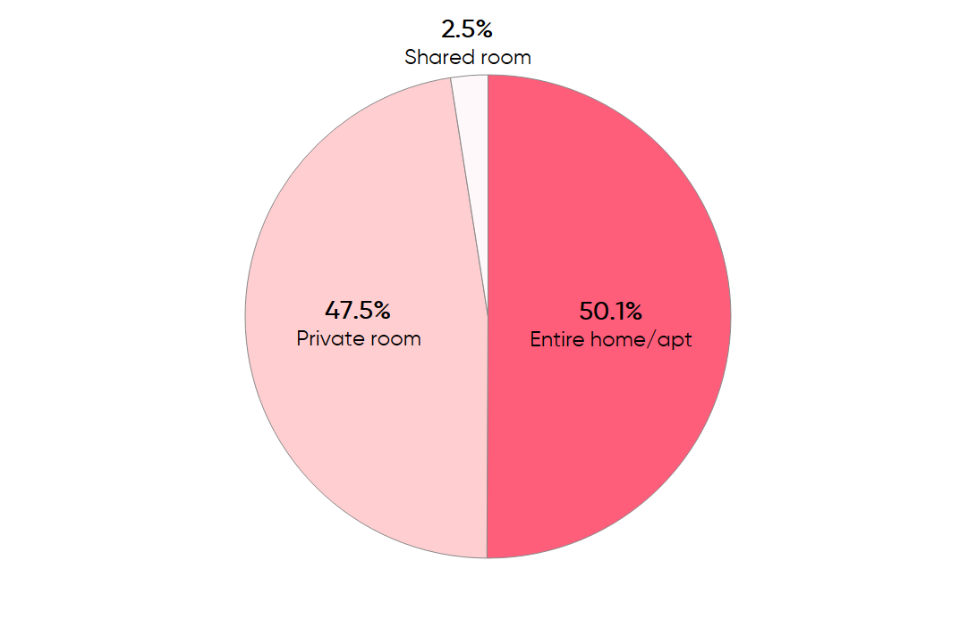
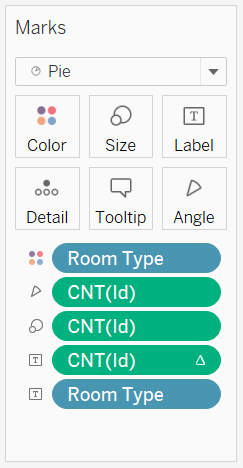
**1.5 Exporting Dataset**

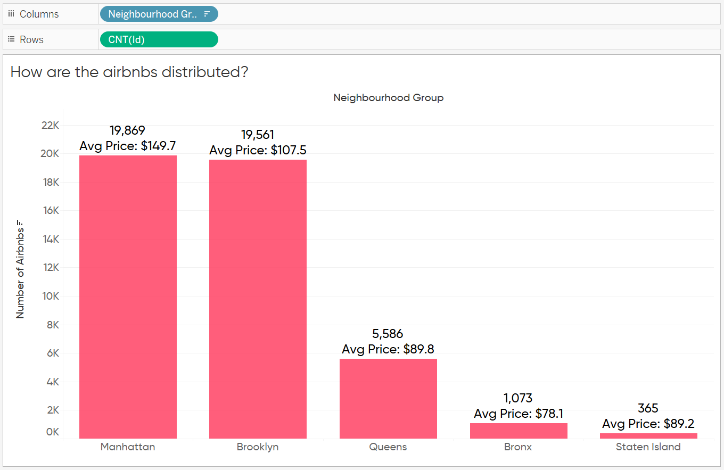
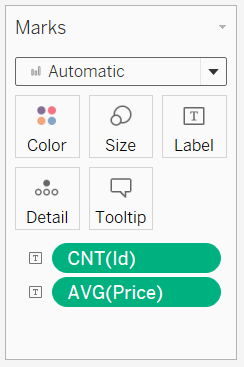
**END OF DATA PREPARATION**

# 2. Data Analysis

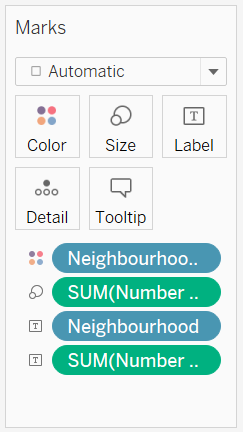
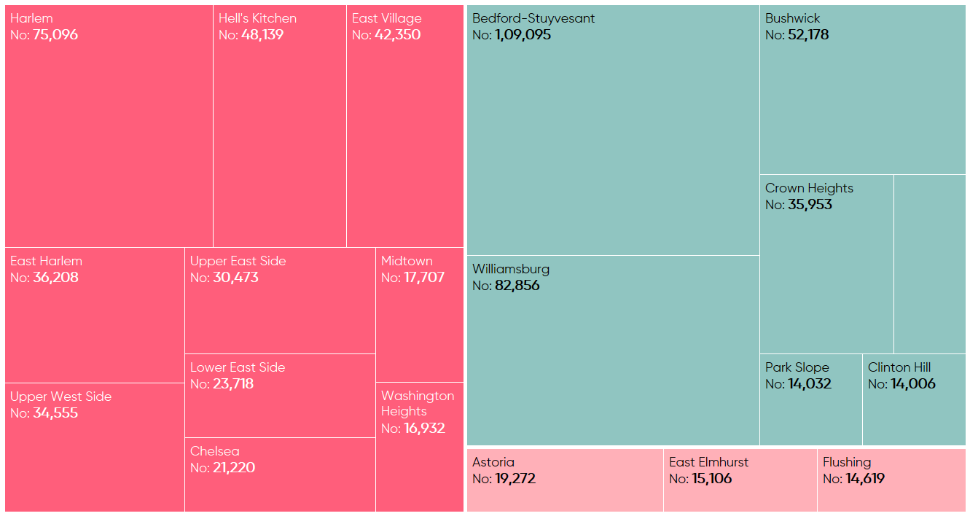
Imported the dataset into Tableau for analysis and creating visualizations. Downloaded NYC Borough Boundaries data from NYC Open Data. Joined our cleaned dataset and the borough boundaries data using a left join on neighborhood group (cleaned data) and boro name (borough boundary data)

**2.1 Host Preferences for Room Type**

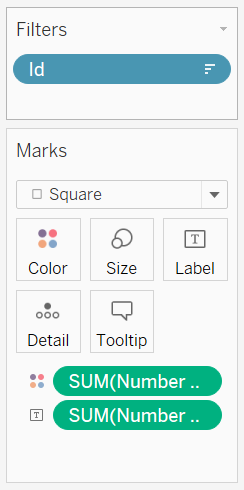
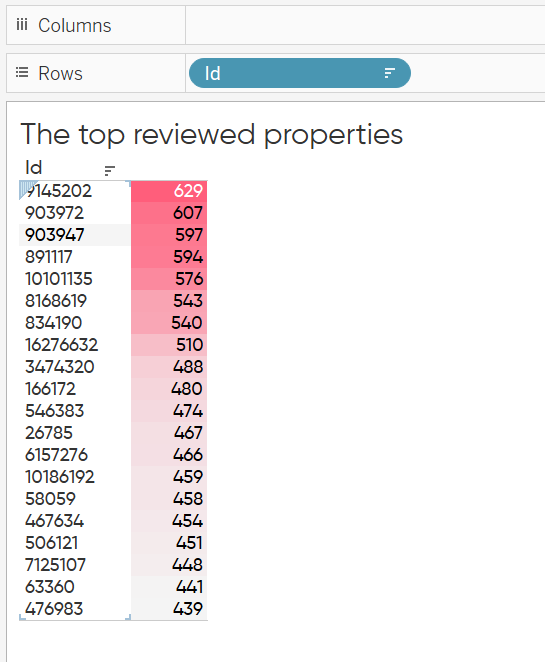
 Plotted a pie chart to see how the listings are distributed among the room types.

**2.2 Distribution of listings across boroughs -** Plotted a bar chart to see the number of listings in each borough and the average price per night

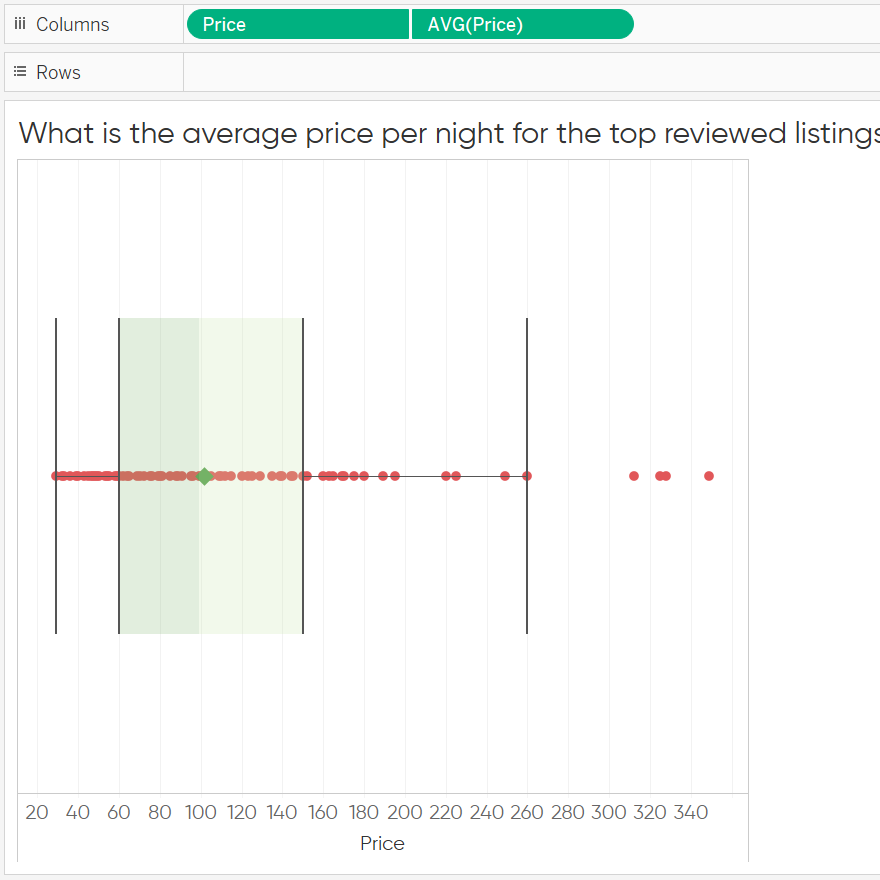
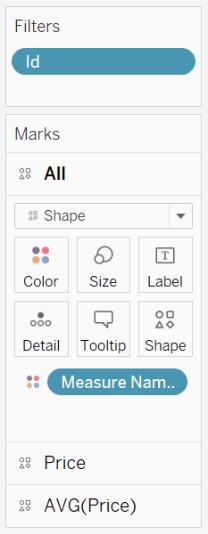
**2.3 Distribution of Reviews - Neighborhoods**

Plotted a tree map of the number of reviews for each neighborhood to identify the most popular neighborhoods.

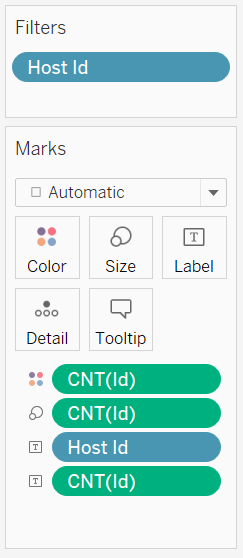
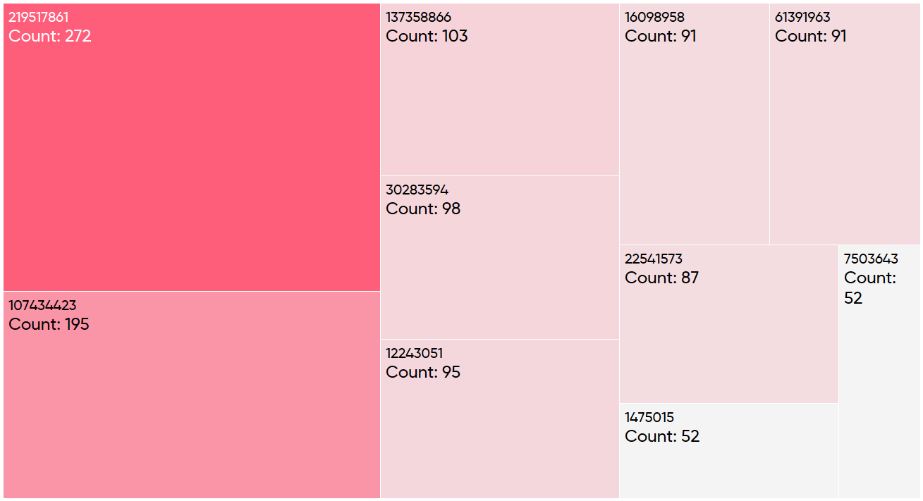
**2.4 The top reviewed properties**

Plotted a highlight table with a filter on id showing only top 20 most reviewed properties.

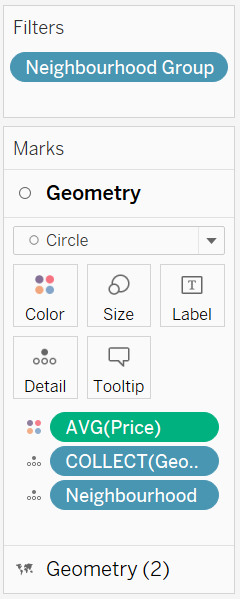
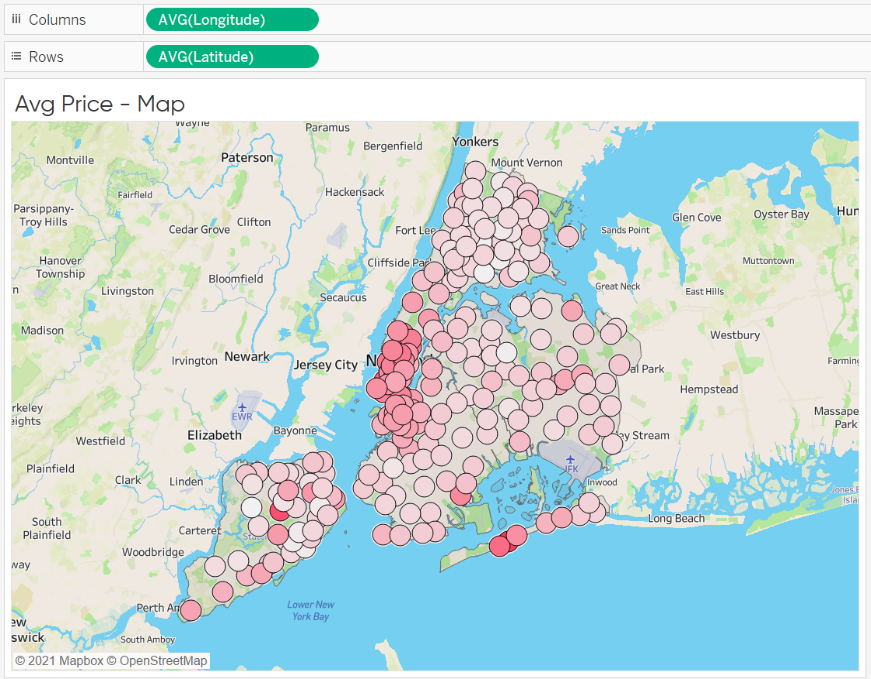
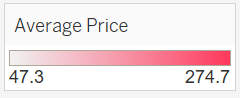
**2.5 Box Plot for prices of most reviewed listings**

Plotted a boxplot to depict the range of prices taken by the most reviewed listings. Also added a point for the average price per night for these listings using by making this a dual axis chart and synchronizing the axes. Filtering on id to get only listings having number of reviews >= 300

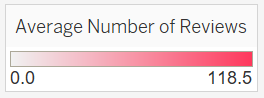
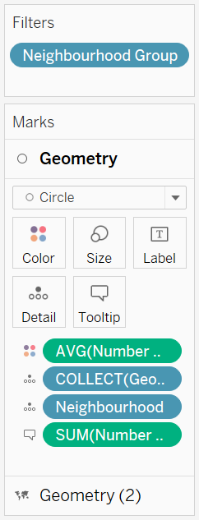
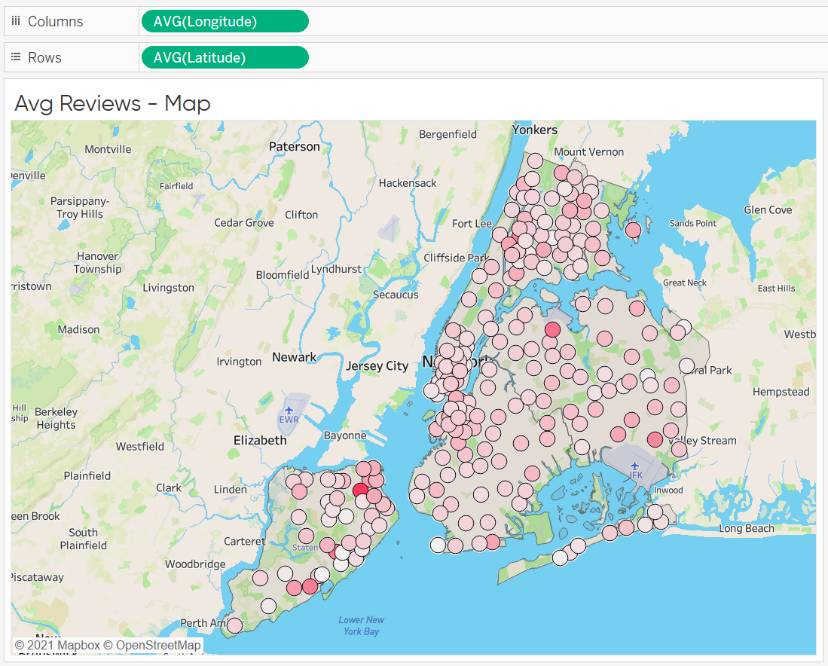
**2.6 Hosts with most listings**

Plotted a tree map to visualize the hosts with the most number of listings to their name.

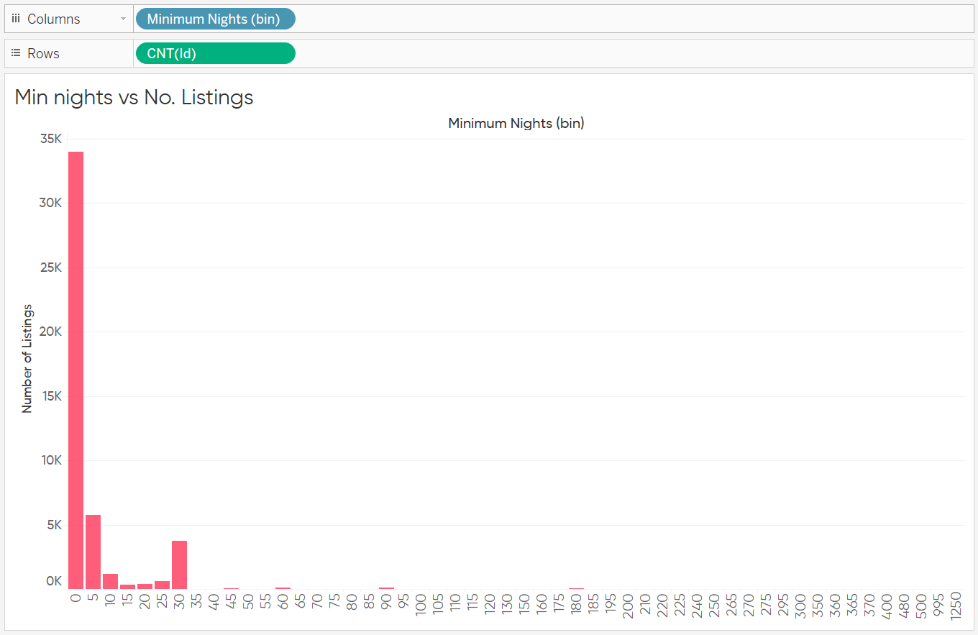
**2.7 Scatter Plot on Map to visualize average price per night for each neighborhood**

Used a marks layer where I added the map geometry and overlaid that with a scatterplot depicting the average price per night for each neighborhood.

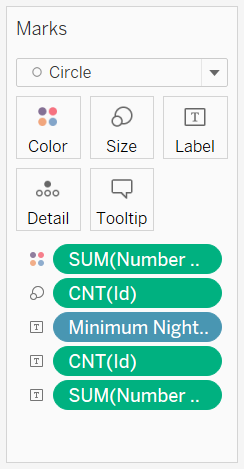
**2.8 Scatter Plot on Map to visualize average number of reviews for each neighborhood**

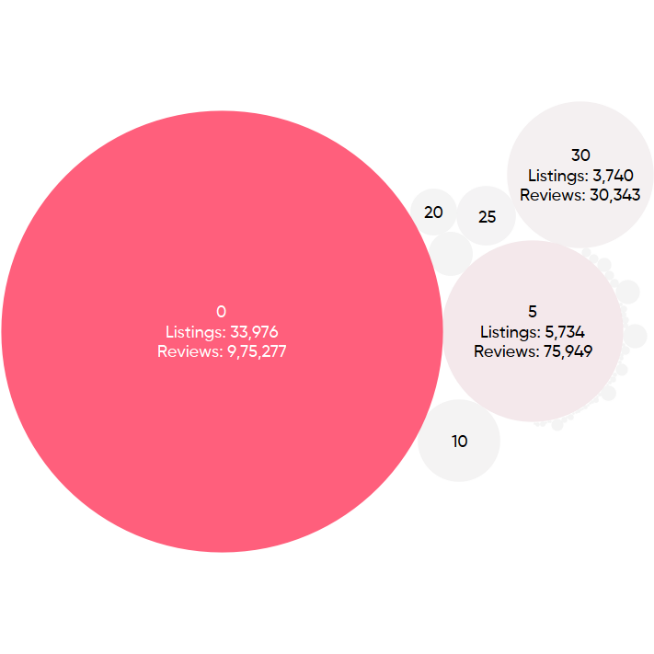
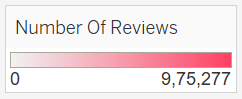
Used a marks layer where I added the map geometry and overlaid that with a scatterplot depicting the average number of reviews for each neighborhood.

**2.9 Histogram for Min Nights vs number of listings**

Created a binned variable from Minimum Nights where size of each bin was 5 nights and plotted a count of listings against it to get a visual depicting the number of listing available for each bin of minimum nights

**2.10 Bubble Plot for Minimum nights (bins) and number of reviews for each bin**

To visualize the customer preference for min nights of stay, I am visualizing it against the number of reviews which we are assuming to be a measure of popularity.



**END OF DATA ANALYSIS**