# **Airbnb Case Study Methodology Document**

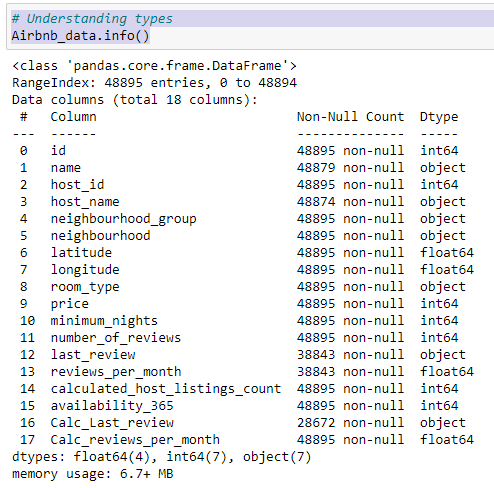
Sreehari Katageri, Vishesh Shroff, Iranna Chatti

**Data Understanding**

Below are the list of attributes and data gathered for the Analysis.

* **Id:** A unique number identifying an Airbnb listing
* **Name:** Name of the listing
* **Host\_id:** A unique number identifying an Airbnb host
* **Host\_name:** Name of the person who hosted the listing
* **Neighbourhood\_group:** Grouping of the region/ city where the listing resides
* **Neighbourhood:** It is the Region/city where the listing resides
* **Latitude:** Latitude coordinates of the listings
* **Longitude:** Longitude coordinates of the listings
* **Room\_type:** one of the “Entire home/apt”, “Private room” or “Shared room”
* **Price:** price of the listing
* **Minimum\_nights:** The minimum stay for a visit, as posted by the host.
* **Number\_of\_reviews:** The number of reviews that the listing has received from the customer. For this exercise we assume all these reviews are positive. It can also be considered as number of visits.
* **Last\_review:** latest review date
* **Reviews\_per\_month:** The number of reviews that a listing has received per month.
* **Calculated\_host\_listings\_count:** The number of listings for a particular host.
* **availability\_365:** The number of days for which a particular host is available in a year.

Understanding the Datatype

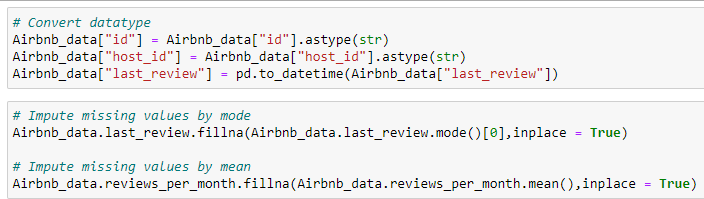


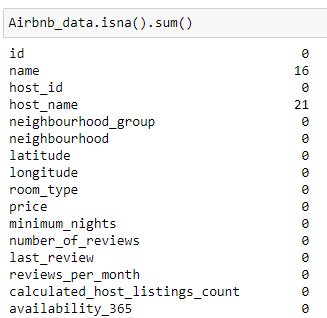
* There are **16** columns & **48,895**Rows in the Airbnb dataset.
* There are **6** Categorical columns of Object datatype
* There are **10** Numerical values of int & Float datatypes among which id, host\_id need to be converted to str datatype & last\_review to be converted to date datatype
* Though name and host\_name have missing values we will retain null values are the corresponding IDs are present
* **last\_review** & **reviews\_per\_month** have null values which we will impute by **MODE** and **MEAN**.

**Data Cleaning & Preparation:**

* There are 10 Numerical values of int & float datatypes among which id, host\_id was converted to STR datatype & last\_review was converted to date datatype.
* Though name and host\_name have missing values we have retained the null values as the corresponding IDs are present.
* last\_review & reviews\_per\_month had null values which was imputed by mode and mean.
* Tools used for Data Cleaning are **Python** & **Excel** and **Tableau** for Analysis.

**Impute Missing values & correct the datatype**

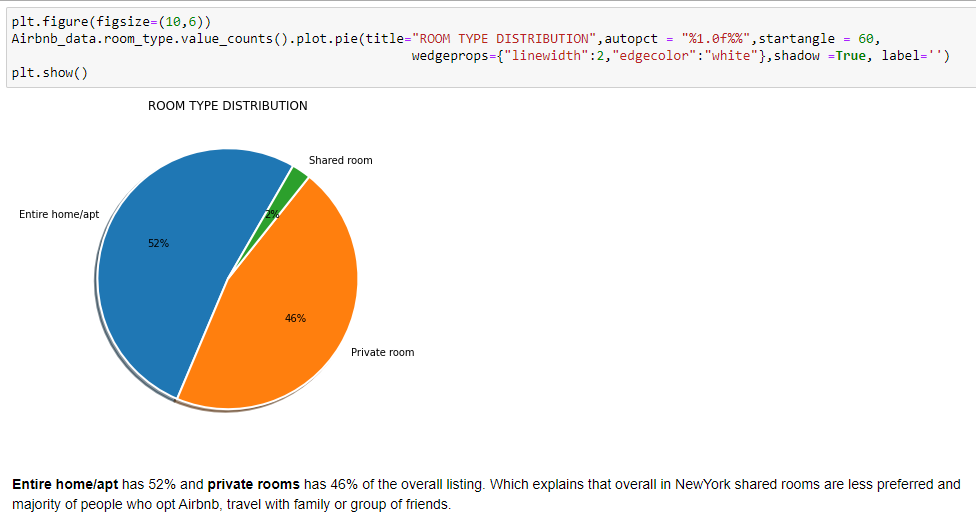




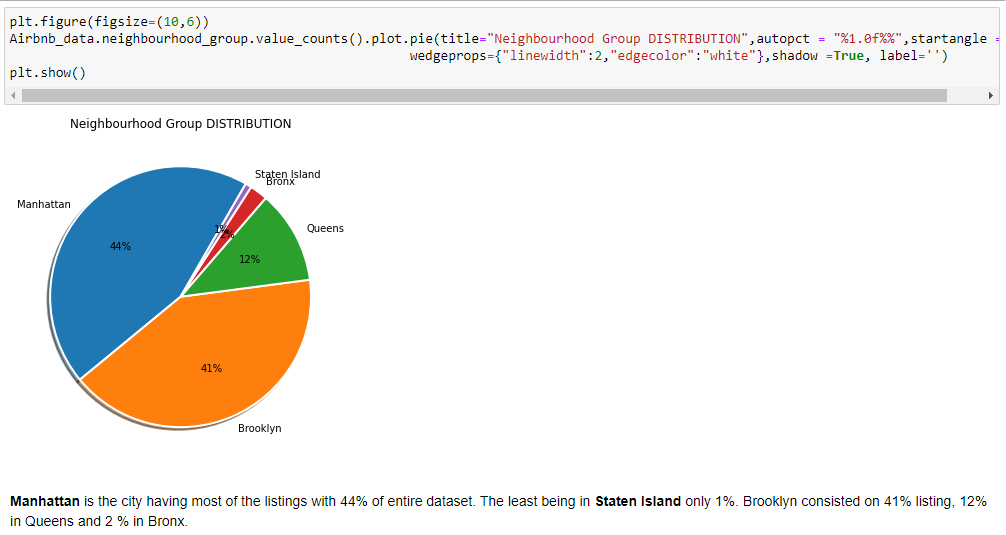
**Data Analysis:**

We have used Python and Tableau for analyzing the data through visualization. We identified the outliers in the numeric variables and accordingly excluded the extreme/ outlier values from our analysis. For Analysis we have used **bar chart**, **stacked bar chart**, **Pie Chart**, **Pareto Chart**, **Scatter Plot** and **Histogram**.

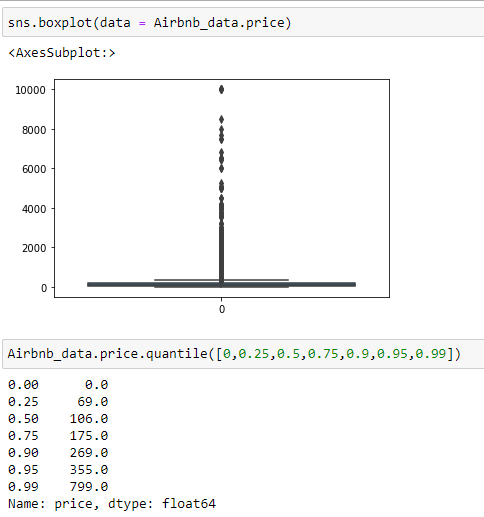
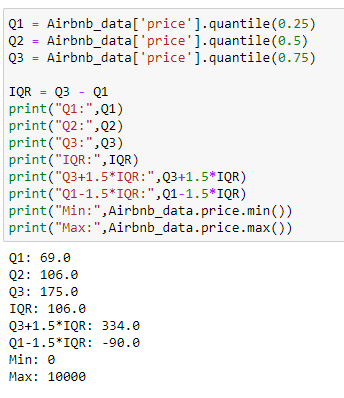
1. **Room Type Distribution**

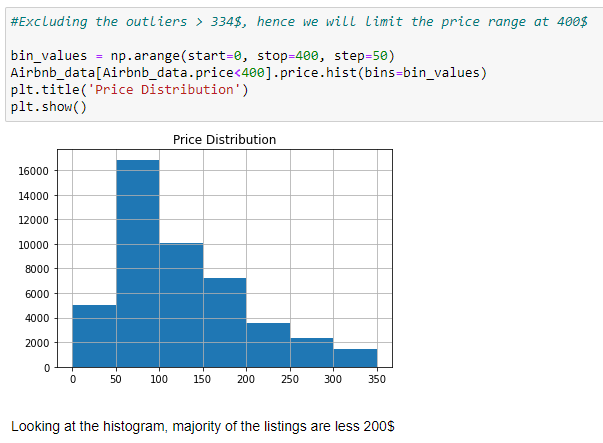


1. **Neighbourhood Group Distribution**

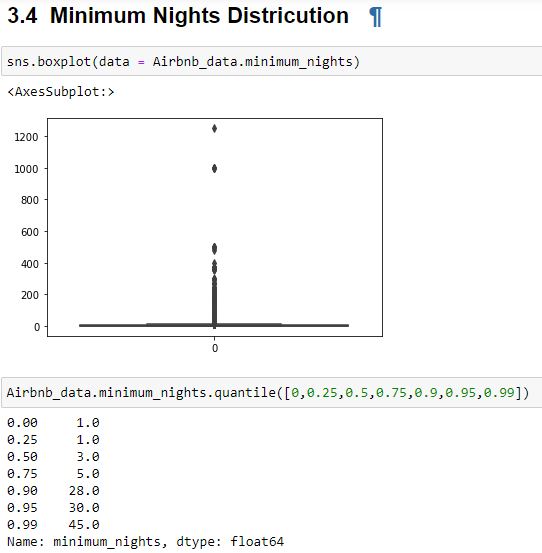
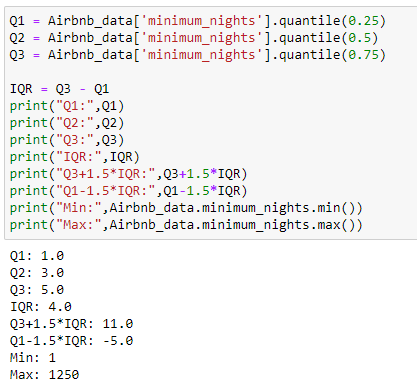


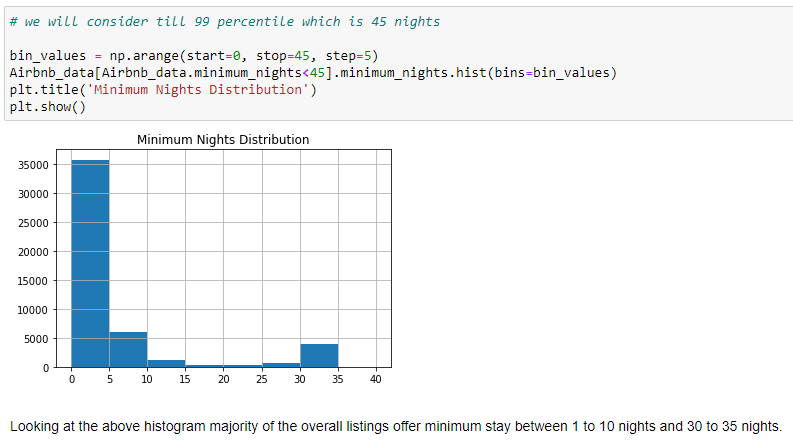
1. **Price Distribution**

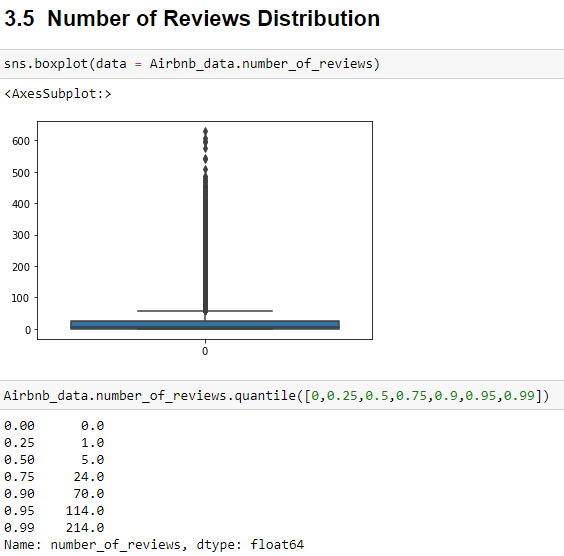
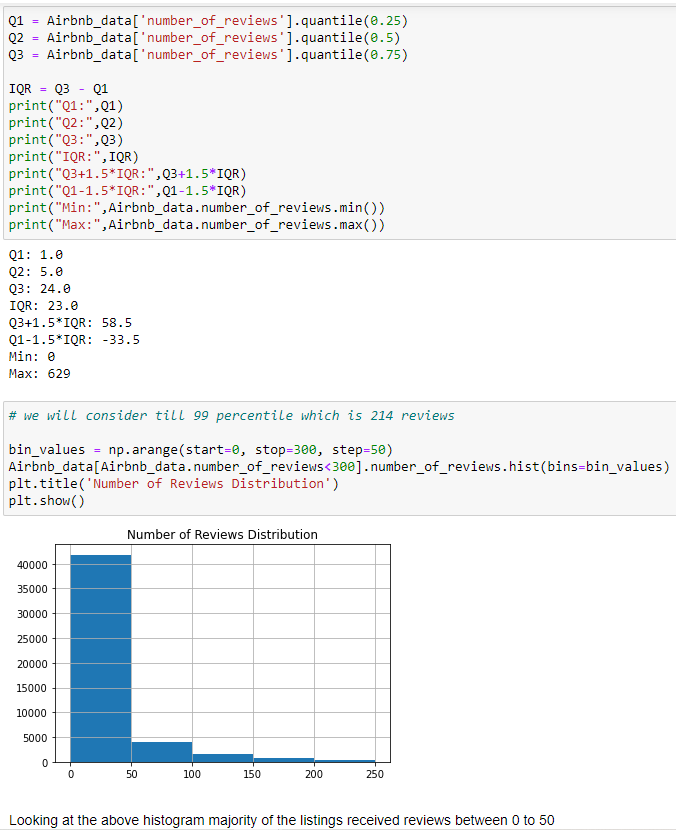
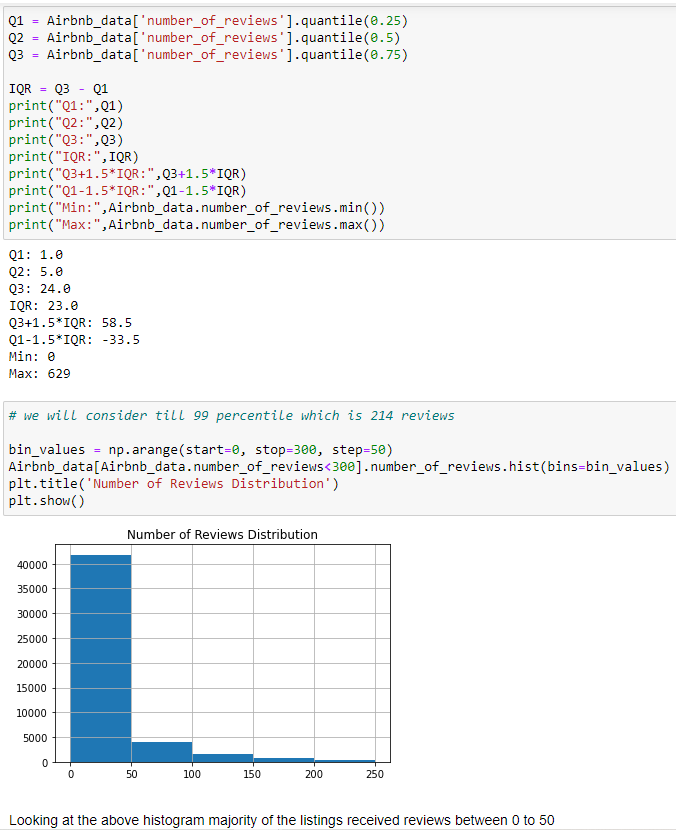


1. **Minimum Night Stay Distribution**

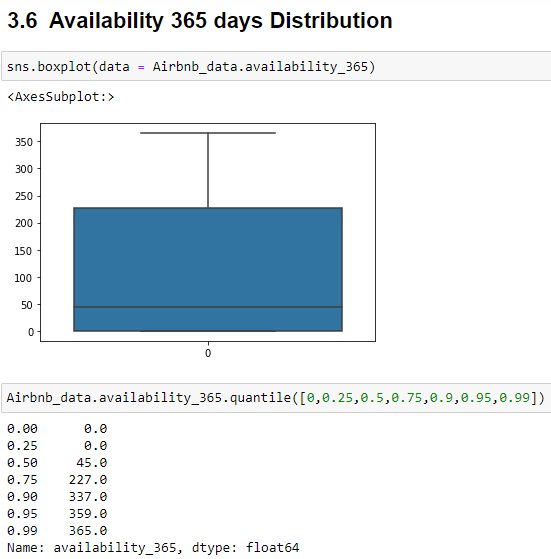
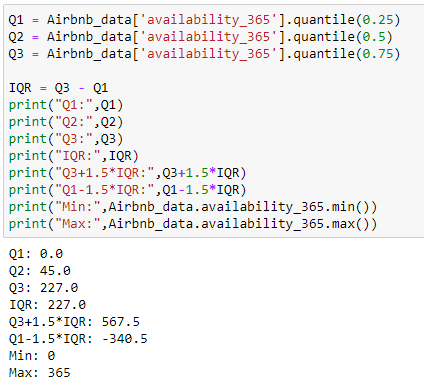
 



1. **Number of Reviews Distribution**

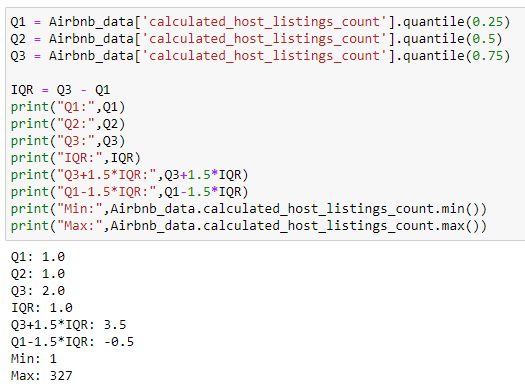
 

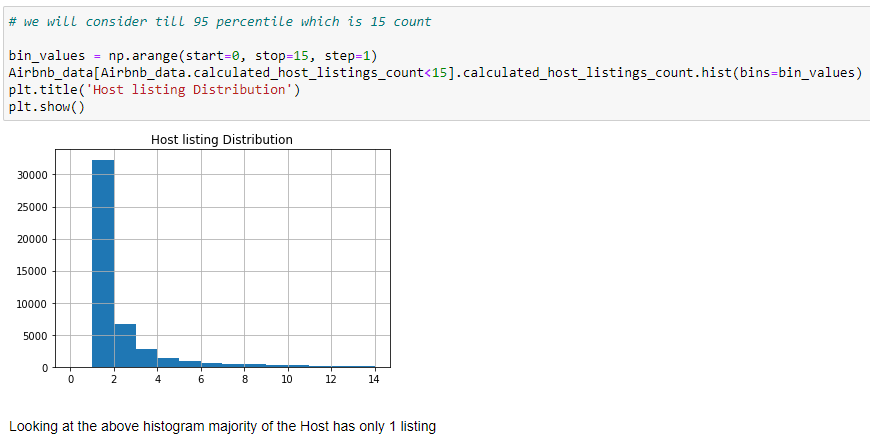
1. **Availability 365 days Distribution**

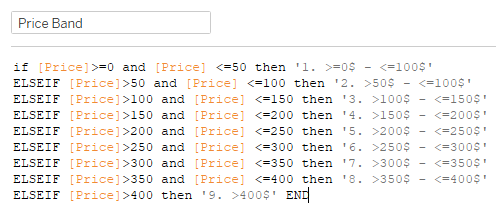
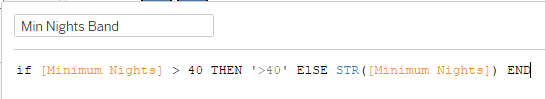
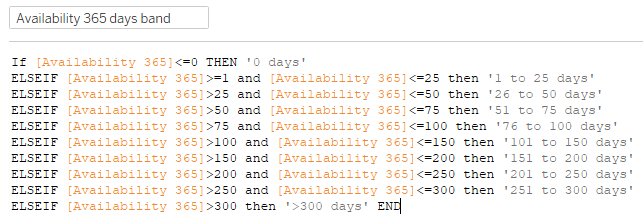


1. **Host listing Distribution**

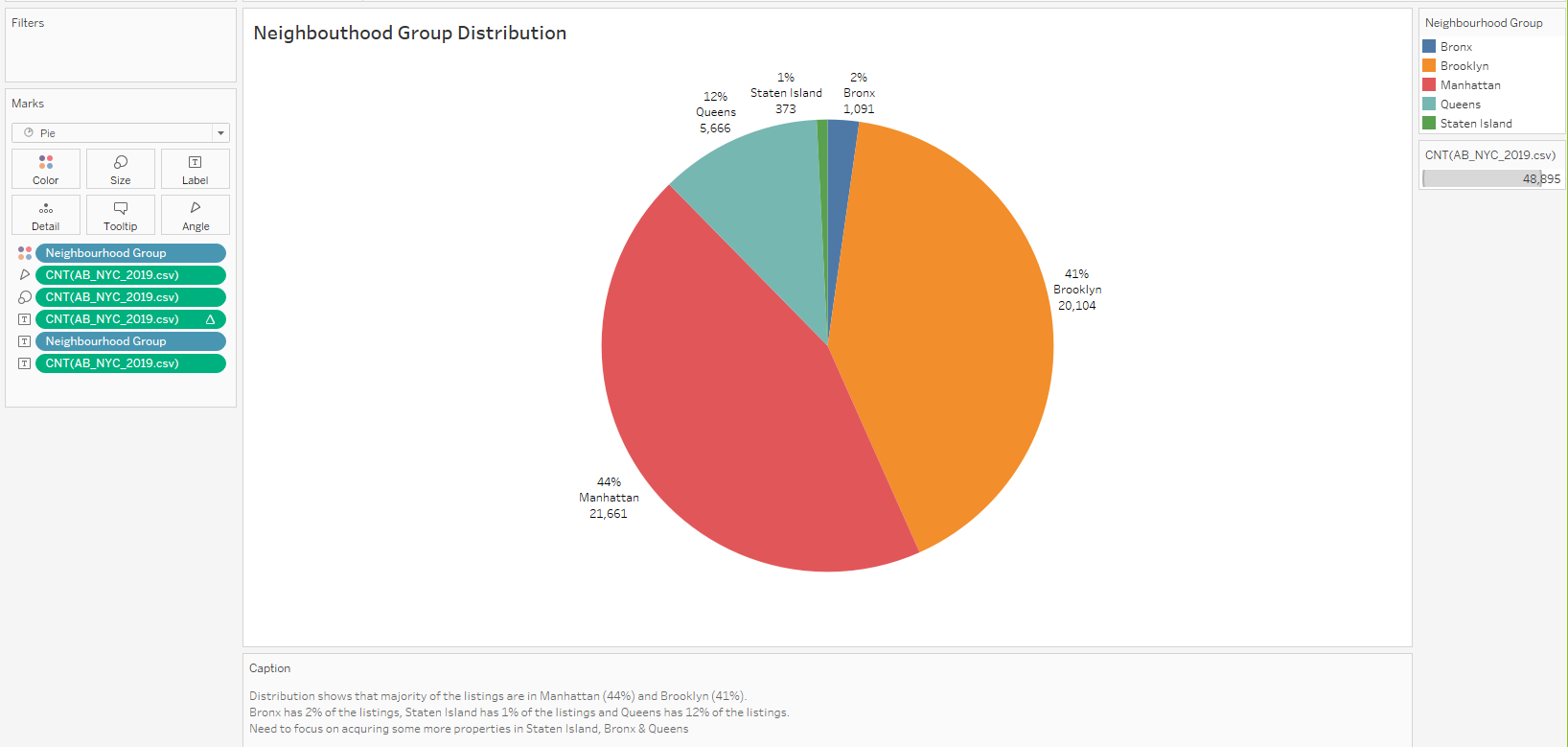


**Below are some visualizations from Tableau**

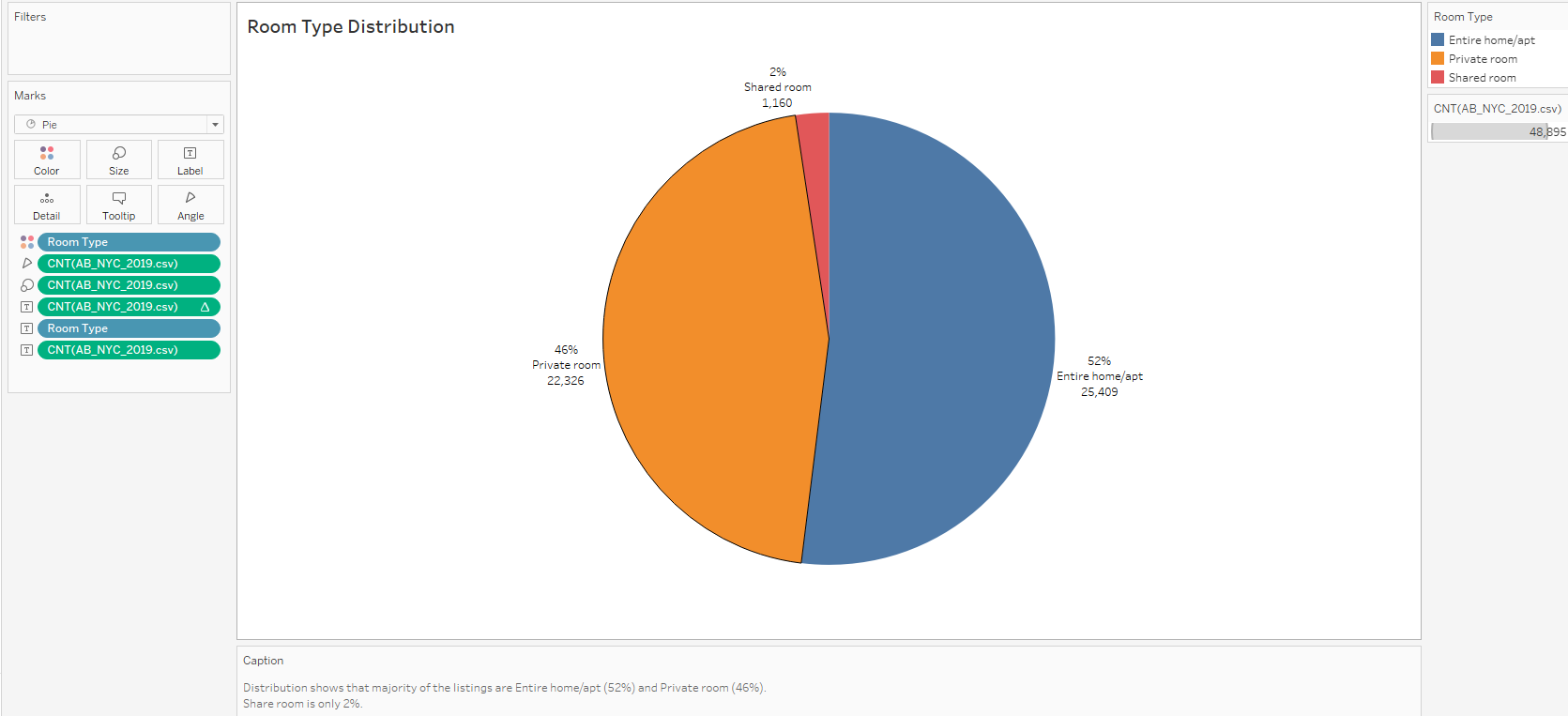
1. **Neighbourhood Group Distribution (Pie Chart)**

* Distribution shows that majority of the listings are in Manhattan (44%) and Brooklyn (41%).
* Bronx has 2% of the listings, Staten Island has 1% of the listings and Queens has 12% of the listings.
* Need to focus on acquiring some more properties in Staten Island, Bronx & Queens



1. **Room Type Distribution (Pie Chart)**

* Distribution shows that majority of the listings are Entire home/apt (52%) and Private room (46%).
* Share room is only 2%.



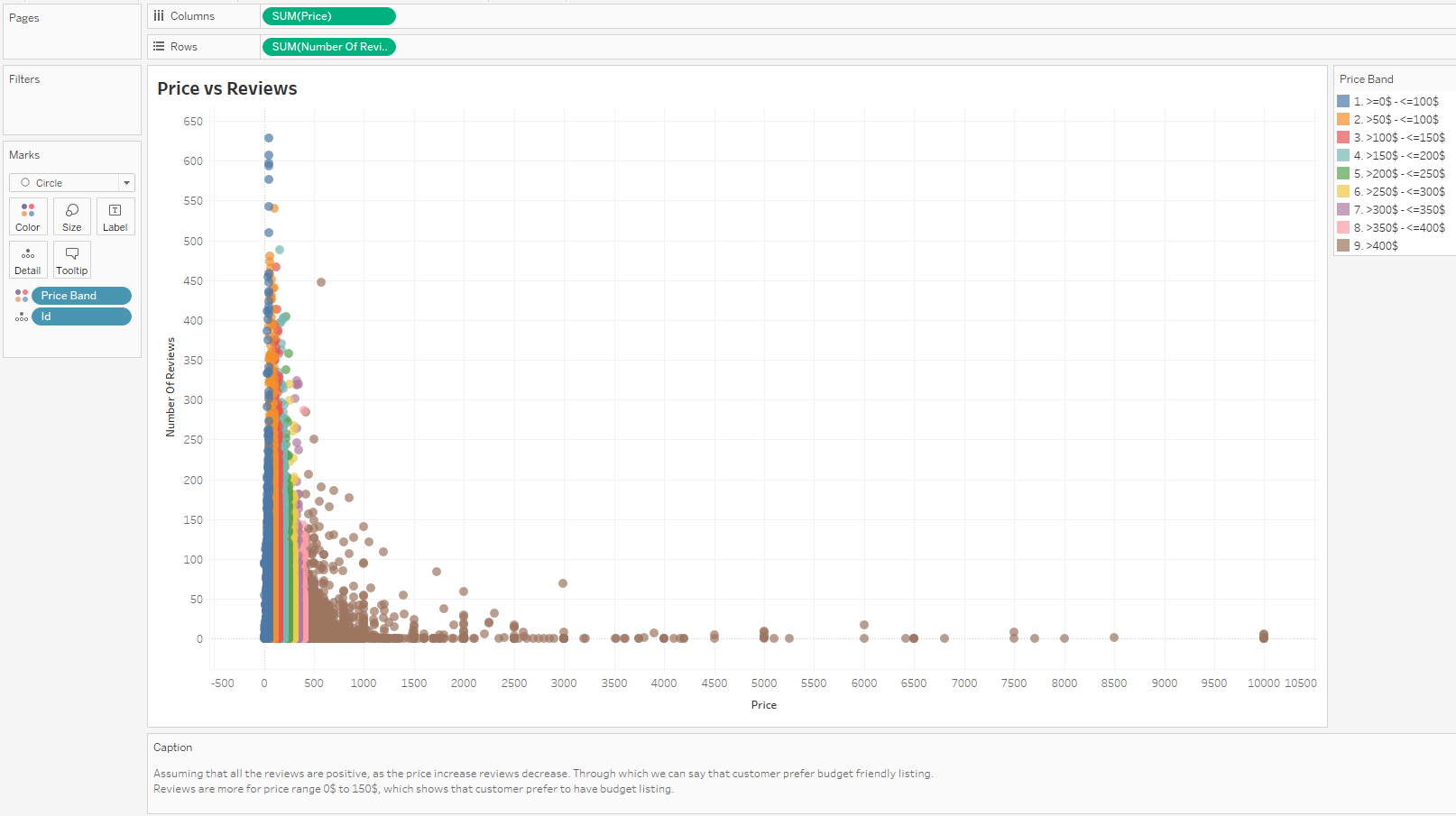
1. **Neighbourhood Group by Room Type (Stacked Bar Chart)**

* Manhattan has more of Entire home/ apt (61%) and less of Private room (37%). Hence Airbnb need to look at increasing Private room properties.
* Price at Manhattan are very high compared to other groups followed by Brooklyn. Need to focus on acquiring apartment properties with the price less than 150$



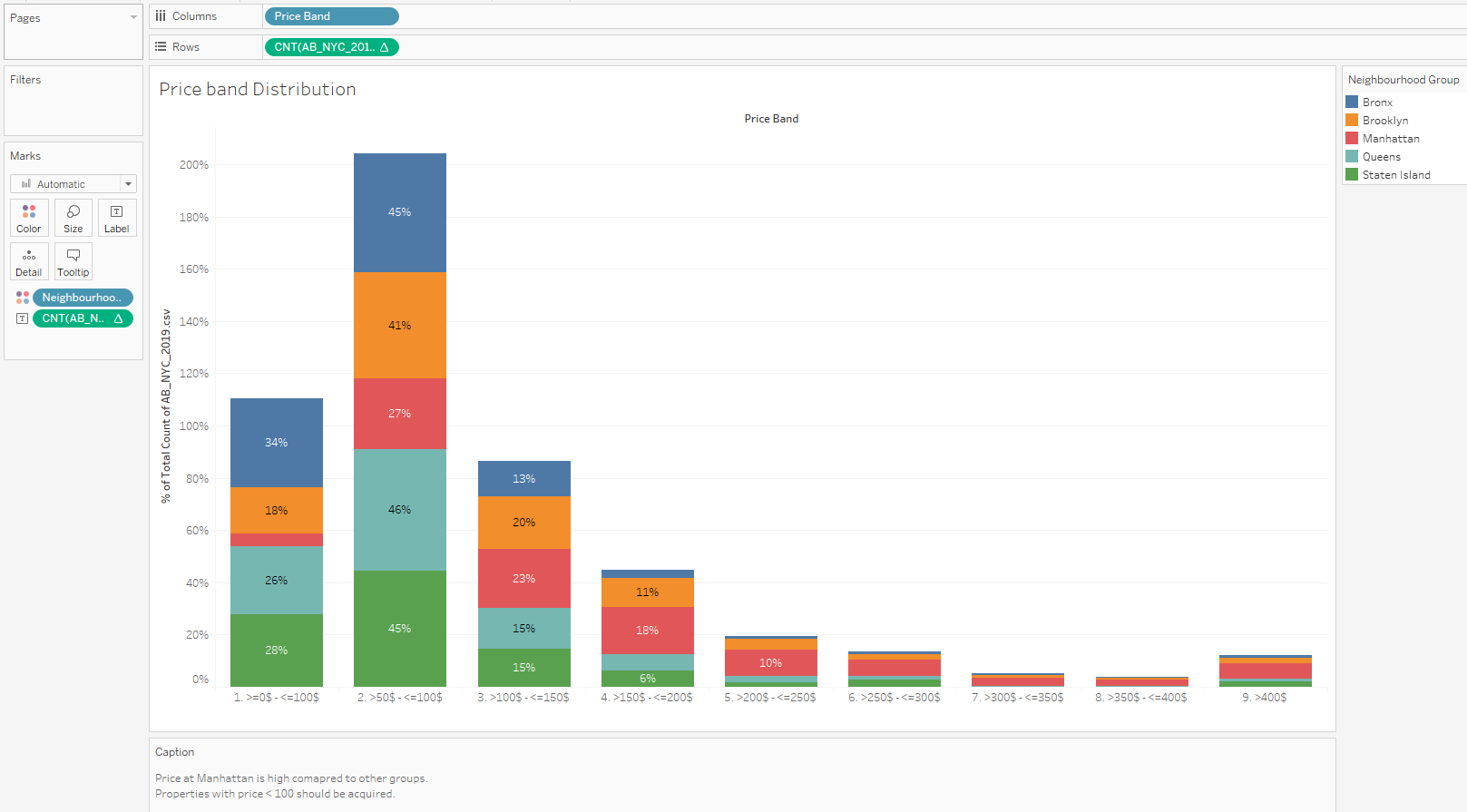
1. **Price vs Review (Scatter plot)**

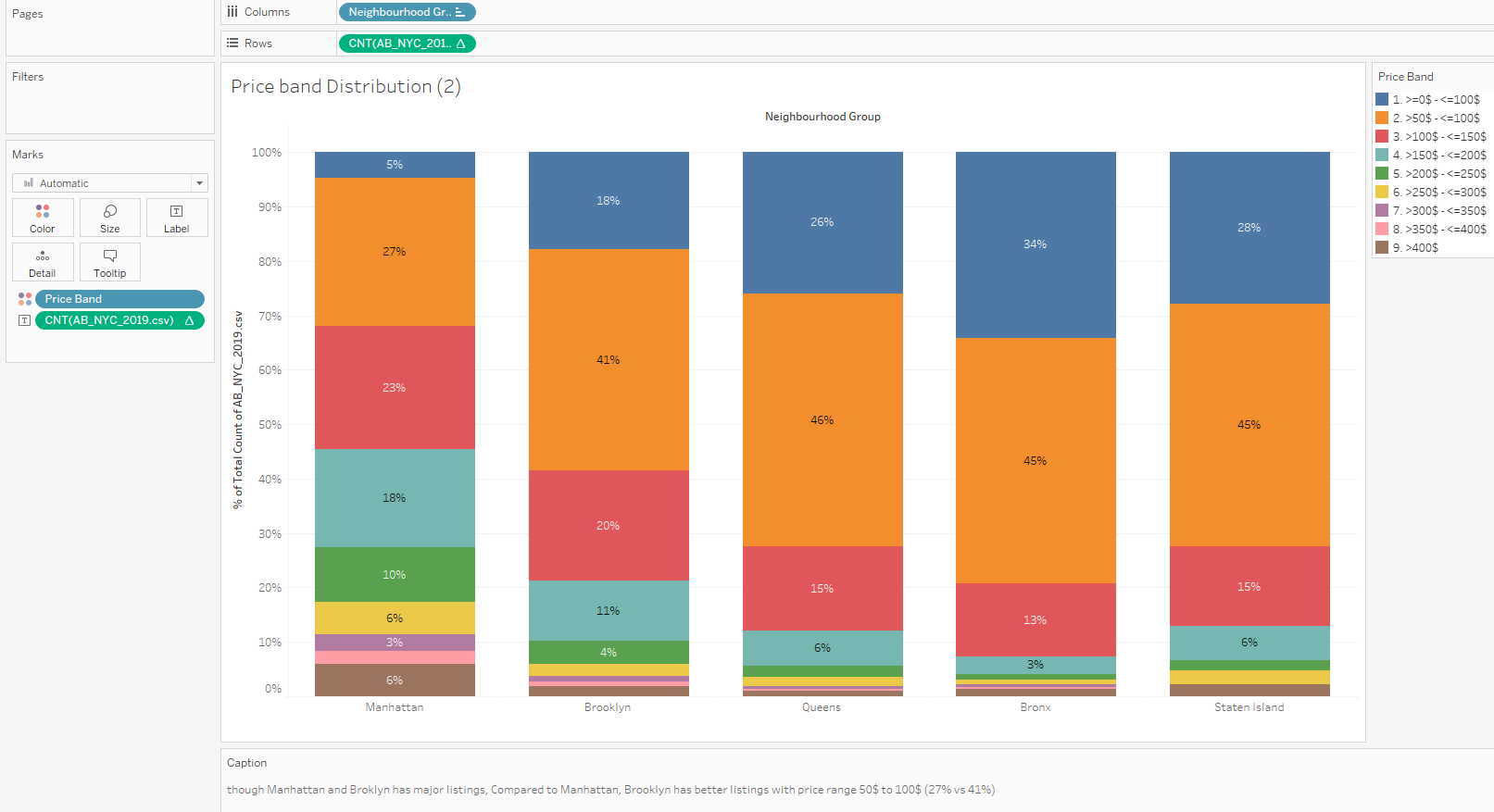
* Assuming that all the reviews are positive, as the price increase reviews decrease. Through which we can say that customer prefer budget friendly listing.
* Reviews are more for price range 0$ to 150$, which shows that customer prefer to have budget listing.



1. **Price band Distribution (Stacked Bar Chart)**

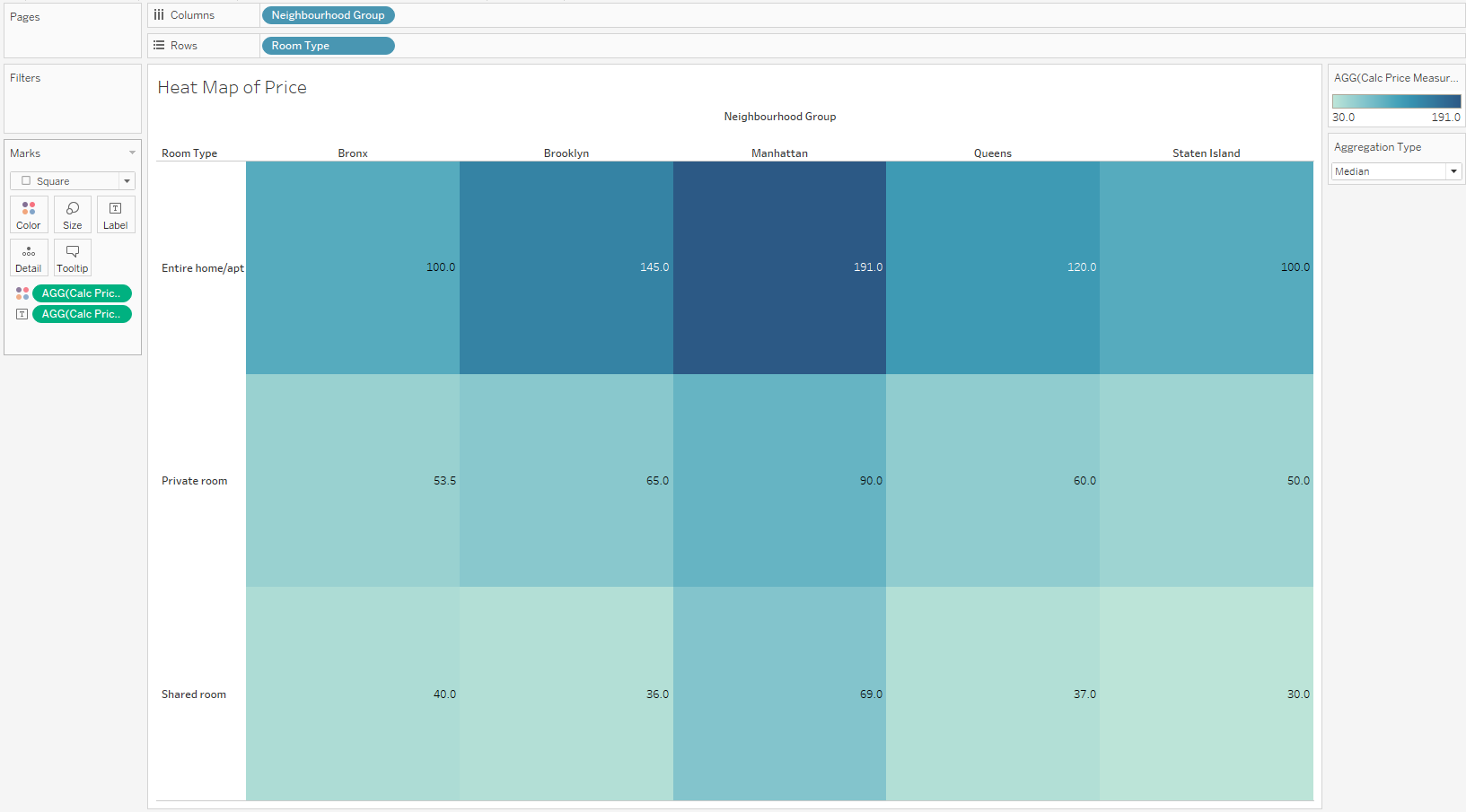
* Price at Manhattan is high compared to other groups.
* Though Manhattan and Brooklyn have major listings, compared to Manhattan, Brooklyn has better listings with price range 50$ to 100$ (27% vs 41%)





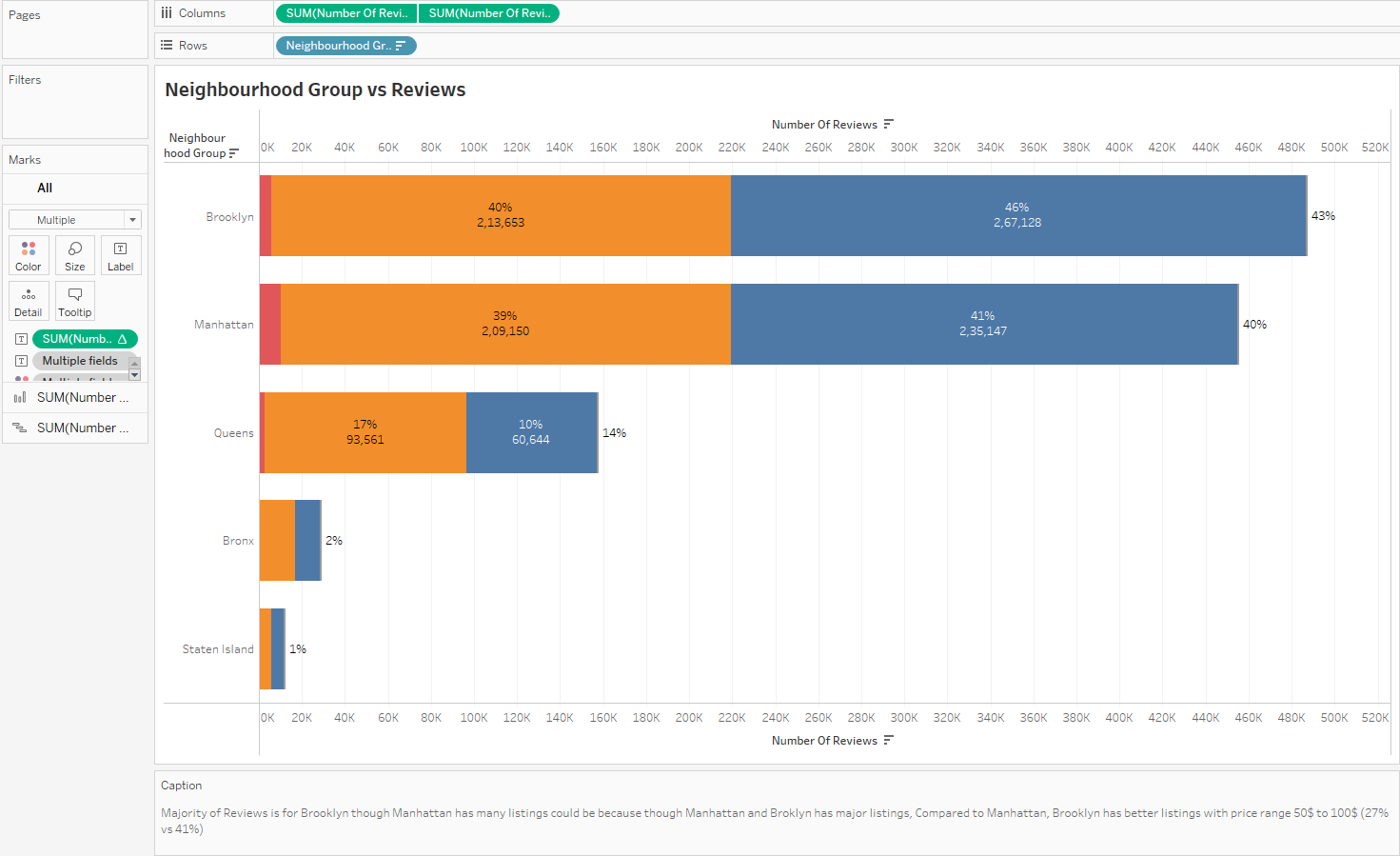
1. **Neighbourhood Group vs Room Type (Heat Map)**

* Price at Manhattan is high with Room type “Entire home/apt”.



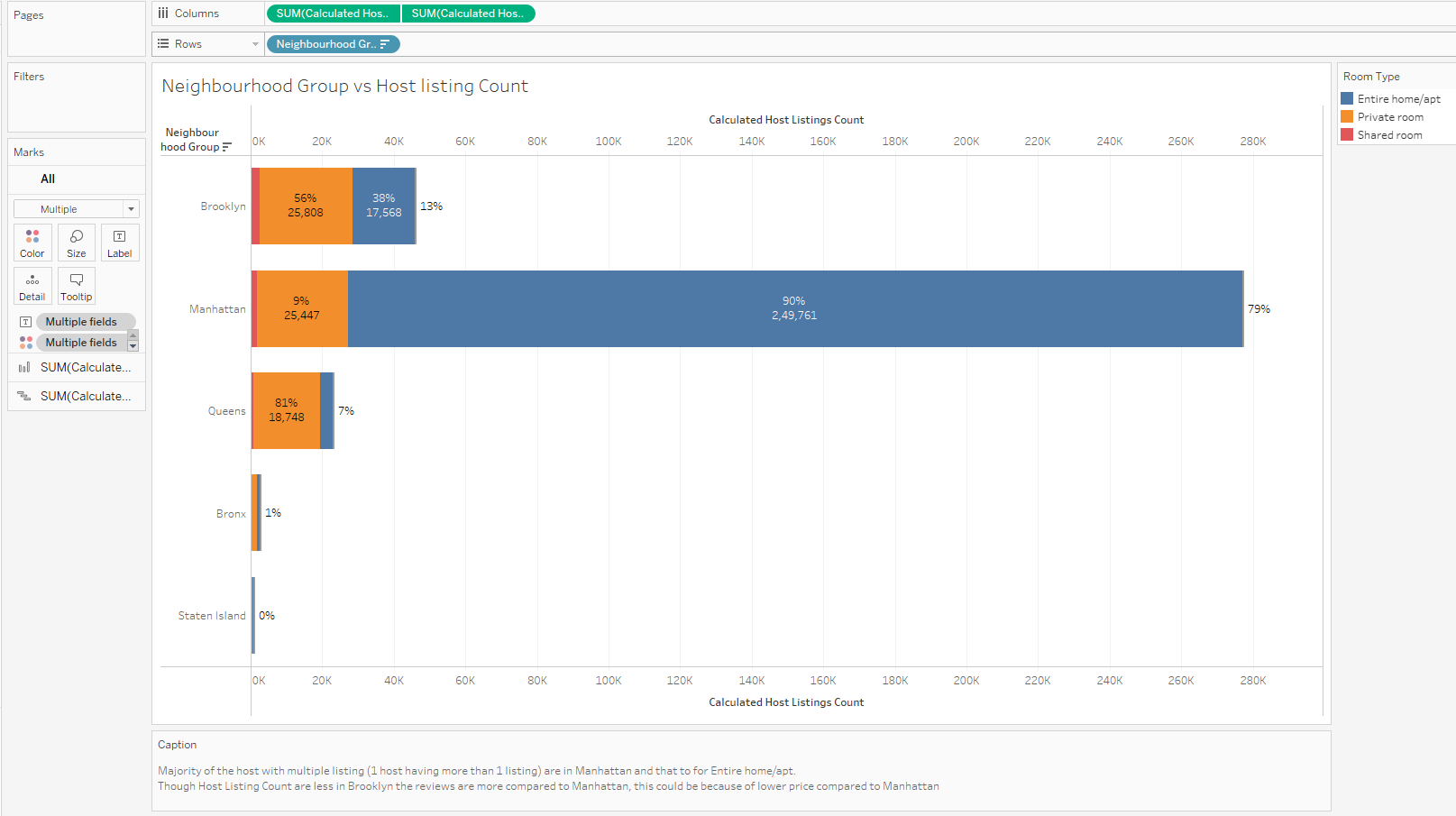
1. **Neighbourhood Group vs Review (Stacked Bar Chart)**

* Majority of Reviews is for Brooklyn though Manhattan has many listings could be because though Manhattan and Brooklyn have major listings, compared to Manhattan, Brooklyn has better listings with price range 50$ to 100$ (27% vs 41%)



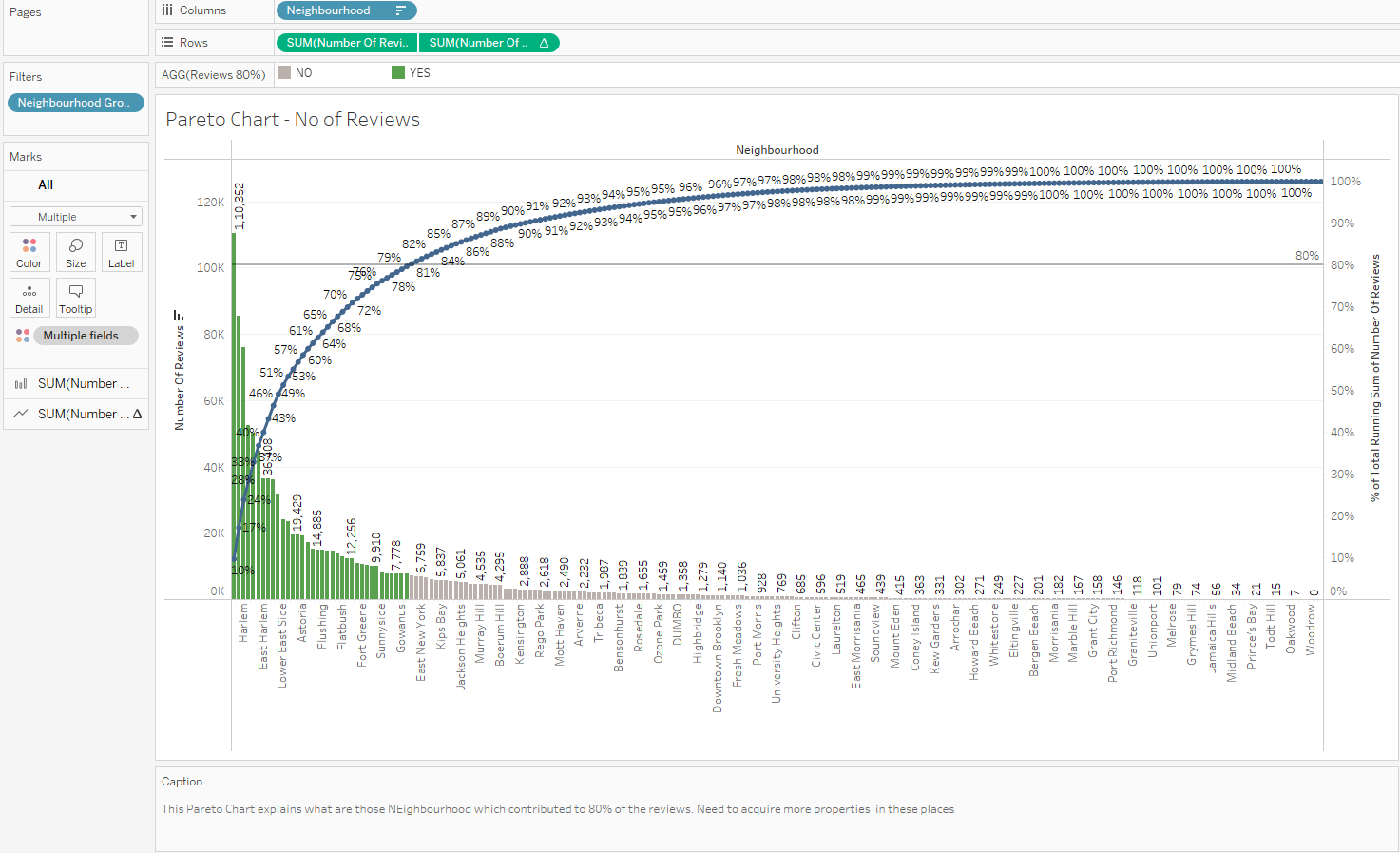
1. **Neighbourhood Group vs Host listing Count (Stacked Bar Chart)**

* Majority of the host with multiple listing (1 host having more than 1 listing) are in Manhattan and that to for Entire home/apt.
* Though Host Listing Count are less in Brooklyn the reviews are more compared to Manhattan, this could be because of lower price compared to Manhattan



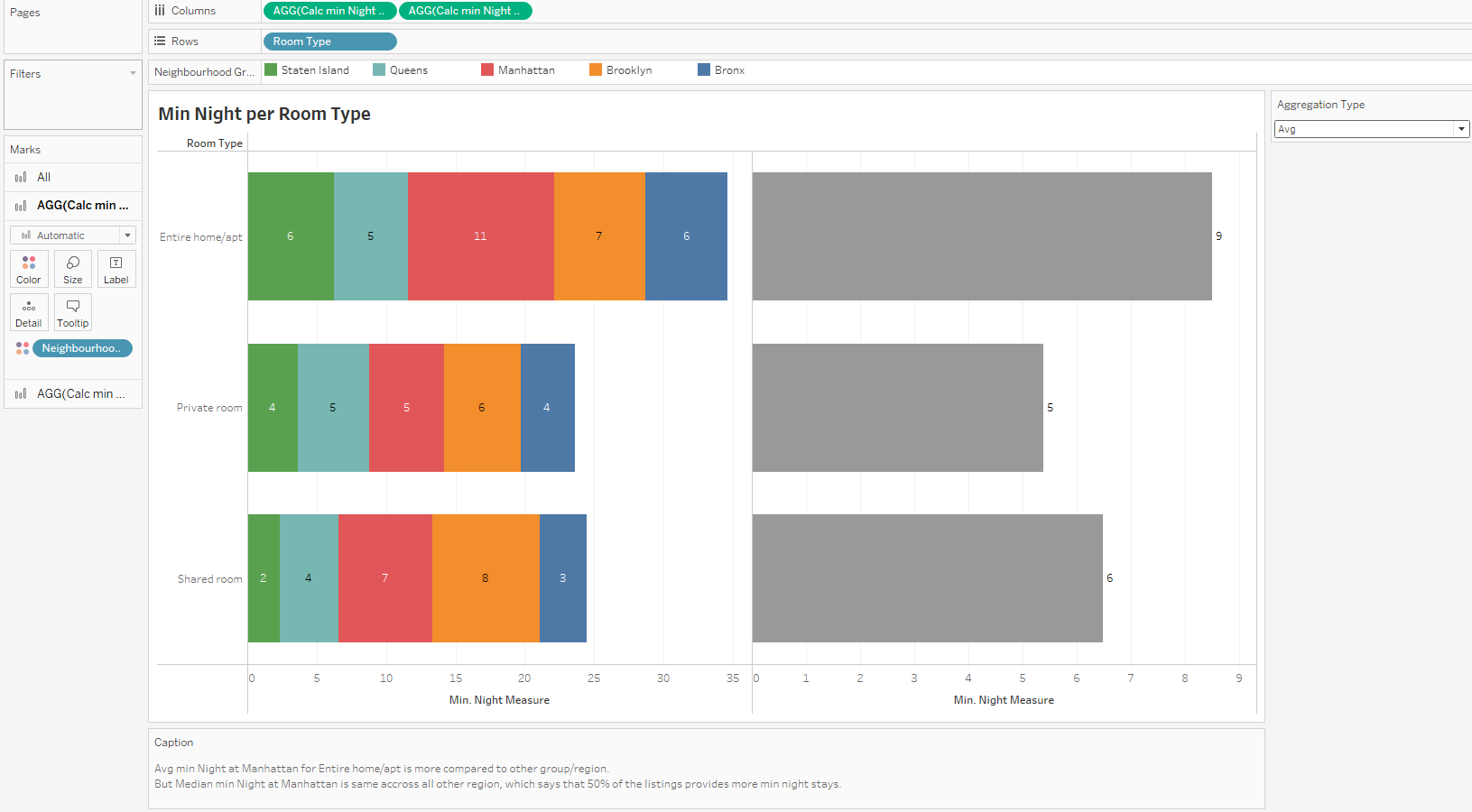
1. **Pareto Chart – No. Of Reviews**

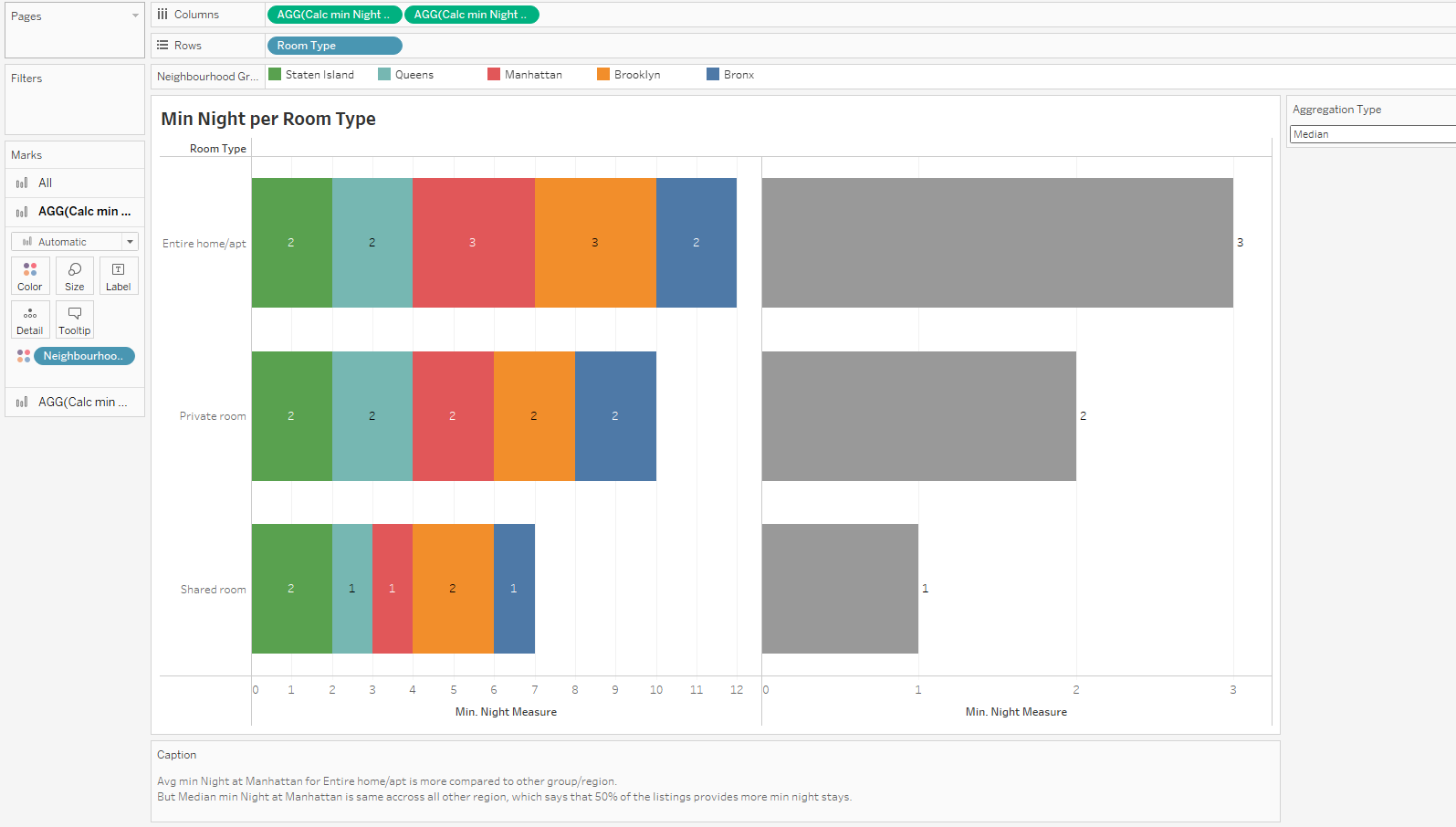
* This Pareto Chart explains what are those Neighbourhood which contributed to 80% of the reviews. Need to acquire more properties in these places.



1. **Min Night per Room Type**

* Avg. min Night at Manhattan for Entire home/apt is more compared to another group/region.
* But Median min Night at Manhattan is same across all other region, which says that 50% of the listings provides more min night stays.





1. **Min. Night stay Distribution**

* 87% of the listings offer 1-to-5-night stays.

