New York Airbnb Case Study

Data Understanding:

There are 16 Variables in the Data of which 10 are Numeric Columns, 5 are Categorical and 1 is a Date column.

```
id                                 int64
name                              object
host_id                            int64
host_name                         object
neighbourhood_group               object
neighbourhood                     object
latitude                         float64
longitude                        float64
room_type                         object
price                              int64
minimum_nights                     int64
number_of_reviews                  int64
last_review                       object
reviews_per_month                float64
calculated_host_listings_count     int64
availability_365                   int64
dtype: object
```

```
df.count()
```

```
id                                48895
name                              48879
host_id                           48895
host_name                         48874
neighbourhood_group               48895
neighbourhood                     48895
latitude                          48895
longitude                         48895
room_type                         48895
price                             48895
minimum_nights                    48895
number_of_reviews                 48895
last_review                       38843
reviews_per_month                 38843
calculated_host_listings_count    48895
availability_365                  48895
dtype: int64
```

Handling Missing Values:

```
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
last_review                   10052
reviews_per_month             10052
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

last review and **reviews per month** have 20% missing values, we can drop them so that it won't Impact the Analysis.

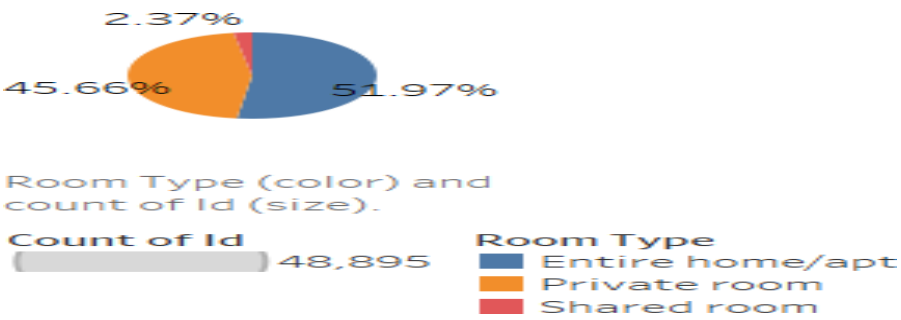
Cleaning of Data

```
df.dropna(subset=['last_review','reviews_per_month'],axis=0,inplace=True)
```

Data Analysis and Visualization:

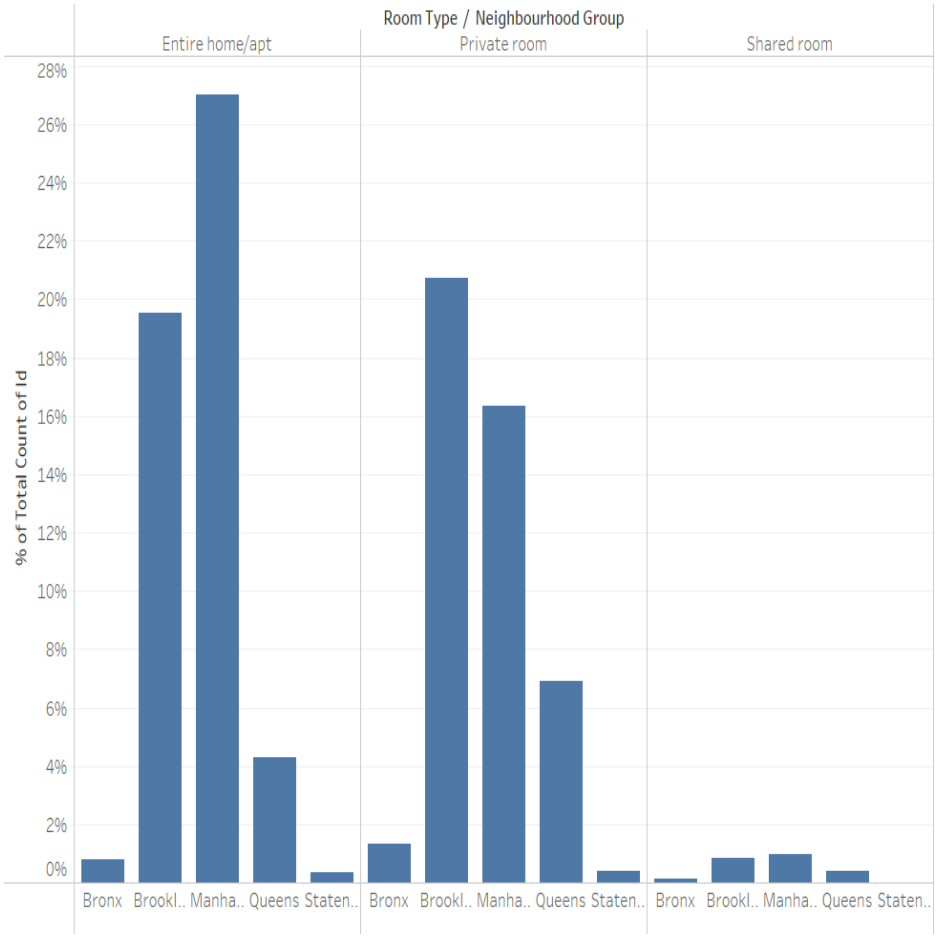
Analysis is done based on Customer Preferences i.e., Room Type, Prices and Regions to get Insights which will be useful for the Business in Post Covid Period.

Customer Preferences of Room Types:



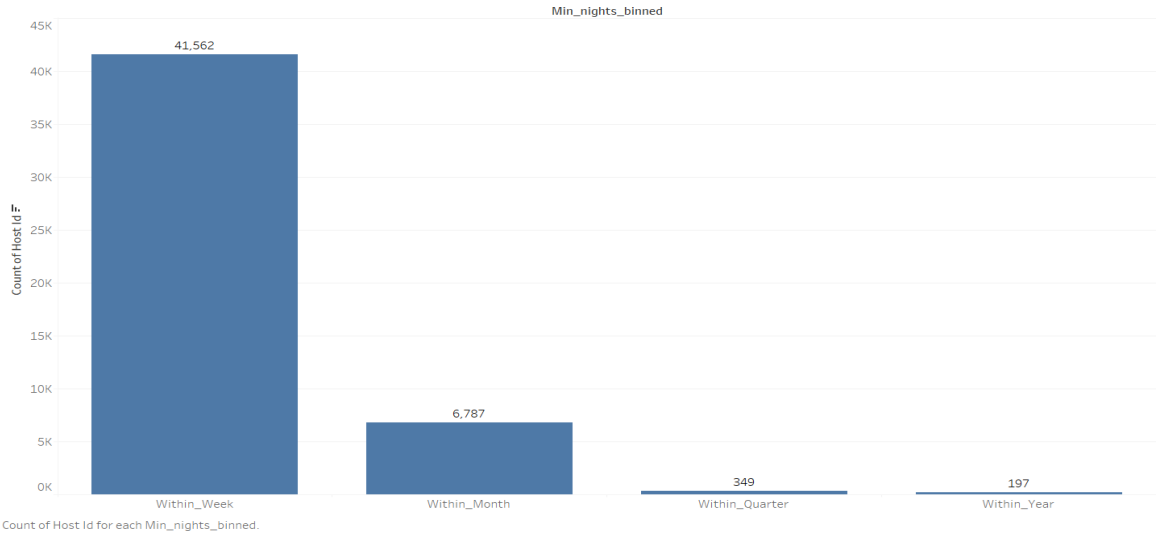
Customers Prefer Properties with Lower Prices across all regions

Customer Preferences Based on Region and Room Type:



% of Total Count of Id for each Neighbourhood Group broken down by Room Type.

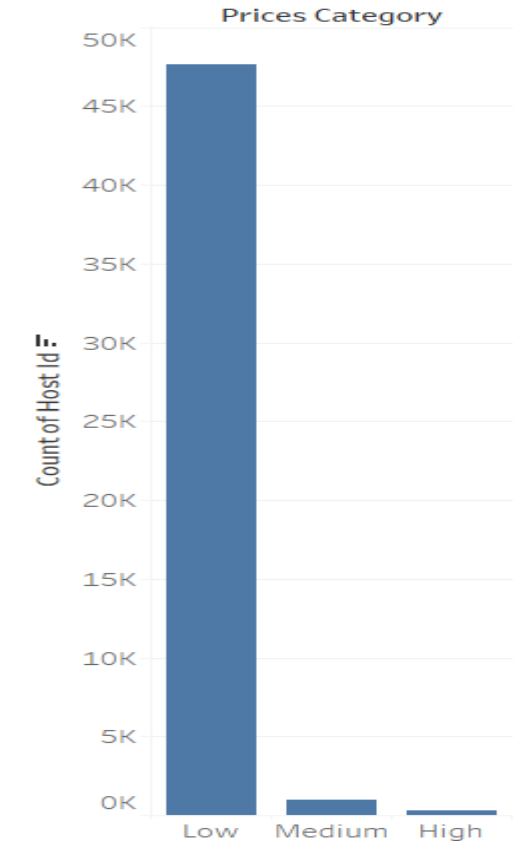
Preference based on Min no of Nights:



Data Methodology:

Binning is done on Prices and Min no of Nights Variable to get the segments which have high/low Priority as per Customer Preference.

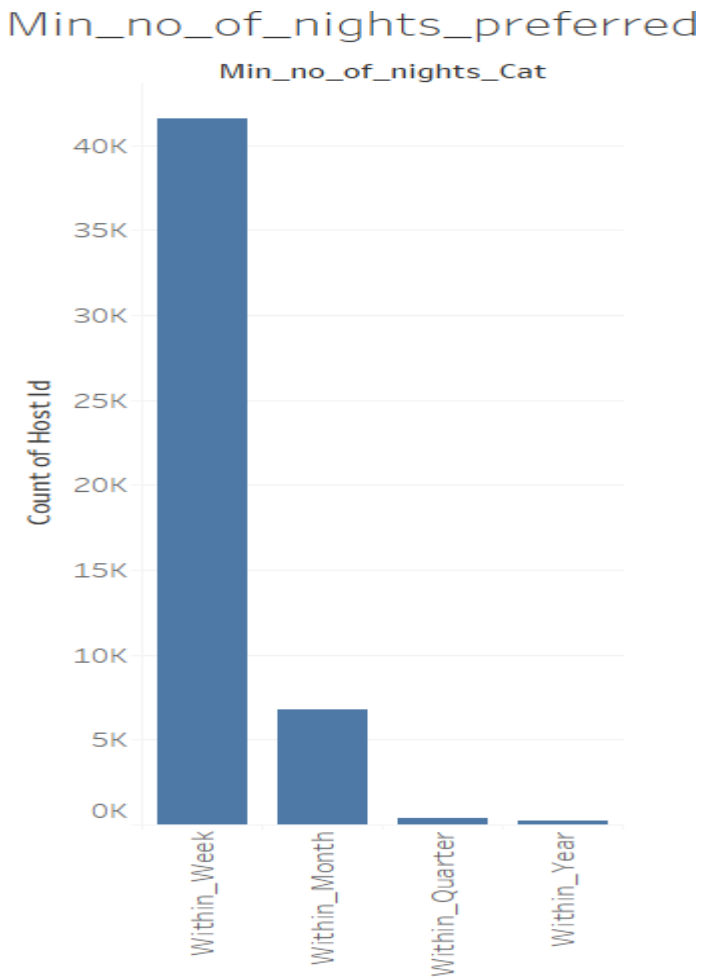
Price_bins



Minimum number of nights Binning:

Min_no_of_nights_Cat

```
if [Minimum Nights]<8 then "Within_Week"
ELSEIF [Minimum Nights]<32 then "Within_Month"
ELSEIF [Minimum Nights]<91 then "Within_Quarter"
else "Within_Year"
END
```



Business Understanding:

Airbnb has seen a major decline in revenue due to COVID-19 pandemic, now since people have started to travel more Airbnb wants to make sure that it is fully prepared for this change.

This Airbnb ('AB_NYC_2019') dataset for the 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented.

First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings; we found that our top host has 327 listings.

After that, we proceeded with analysing boroughs and neighbourhood listing densities and what areas were more popular than another. Next, we put good use of our latitude and longitude columns and used to create a geographical heat map color-coded by the price of listings.

Further, we came back to the first column with name strings and had to do a bit more coding to parse each title and analyse existing trends on how listings are named as well as what was the count for the most used words by hosts. Lastly.

Tool used for Data wrangling: Python

Tool used for Binning and Visualization: Tableau 2021.3

Prepared By:

Korakolla Umesh