Airbnb NYC- Data-Driven Analysis Methodology

By:

Sagar Barge, Vunna Praveen Kumar, Satbir Kaur

Methodology

1. **Understanding problem & objective:**
   Airbnb has experienced a significant decline in revenue over recent months, likely due to reduced travel demand. So, Airbnb wants to find solutions for this with help of listing data.
2. **Understanding the Data-**
   This data includes all Airbnb listings in New York City. Contains details such as host information, property type, price, location, availability, and customer reviews.

Data Dictionary

Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

3. **Discussing about Tools required:**
   - After discussion with team members we decided to different software to achieve this objective.
   - Tools: Python, Excel, PowerBI, Tableau, QGIS.

4. **Import required libraries-**

Import Liabraries

In [1]:
```
# Import Liabraries
import numpy as np
import pandas as pd
import matplotlib.pylab as plt
import seaborn as sns
import geopandas as gpd
from shapely.geometry import Point
import warnings
warnings.filterwarnings ('ignore')
```

5. **Checking data type, and basic statistics to get an overview of data:**
   - *.shape, df.info(), df.describe()*

6. **Checking duplicate values:**

```
# Find duplicates
df.duplicated().sum()
```

```
0
```

7. **Checking null values :**

```
# Checking null values

df.isnull().sum()
```

```
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
```

8. **Filled Null values :**
   - Filled 'name', 'host_name' as ‘Unknown’. This column is useful for finding top listing names.
   - Filled Null values as 0 >> Because “number_of_reviews” column having 0 value, so “reviews_per_month” will be 0.
   - Changing data format to date

```
# Filled Null values as Unknown
df[['name', 'host_name']]= df[['name', 'host_name']].fillna('Unknown')
```

```
# Filled Null values as 0 >> Because number_of_reviews column having 0 value, so reviews_per_month will be 0.
df['reviews_per_month']= df['reviews_per_month'].fillna(0)
```

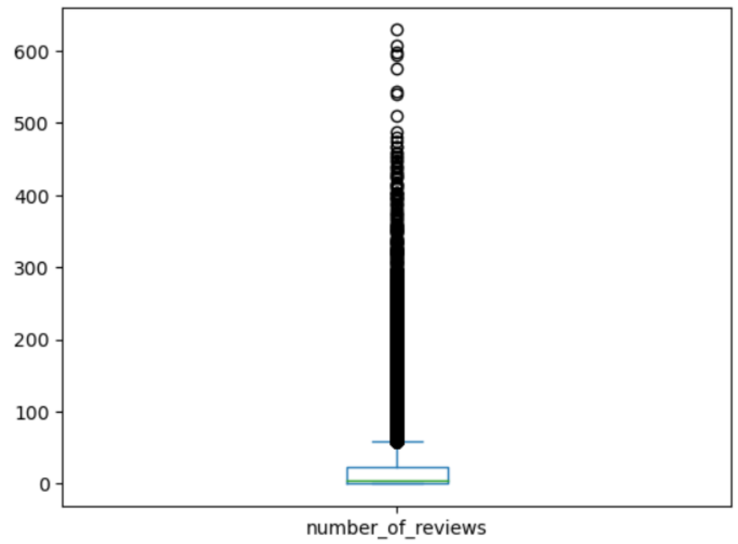
```
# Changing data format to date
df['last_review'] = pd.to_datetime(df['last_review'])

```

9. Checking outlier:

- For checking outlier used box plot & value count
  Example-

```
df.number_of_reviews.plot.box()
```

```
<Axes: >
```



- "minimum_nights" capped at 500, because after 500, there were only a few values.

```
# Capping at 500

df['minimum_nights']= df['minimum_nights'].where(df['minimum_nights']<500,500)
df.minimum_nights.plot.box()
```

10. Dropping Unnecessary columns

```
# Dropping Unnecessary columns
df = df.drop (['last_review', 'id' ], axis=1)
```

11. Creating new Column with Low review (0), to understand unpopular listings

```
# Adding Column with Low review (0)
df['Low_review'] = df['number_of_reviews']==0
df['Low_review_01'] = (df['number_of_reviews']==0).astype(int)
```

12. Export final dataframe to .csv for more analysis work >>>>>>>>>>>

```
# Export dataframe to .csv

df.to_csv('Airbnb_NYC_final.csv')
```

13. Correlation Matrix to understand relation between different variables

14. Hotspot map created using geopandas

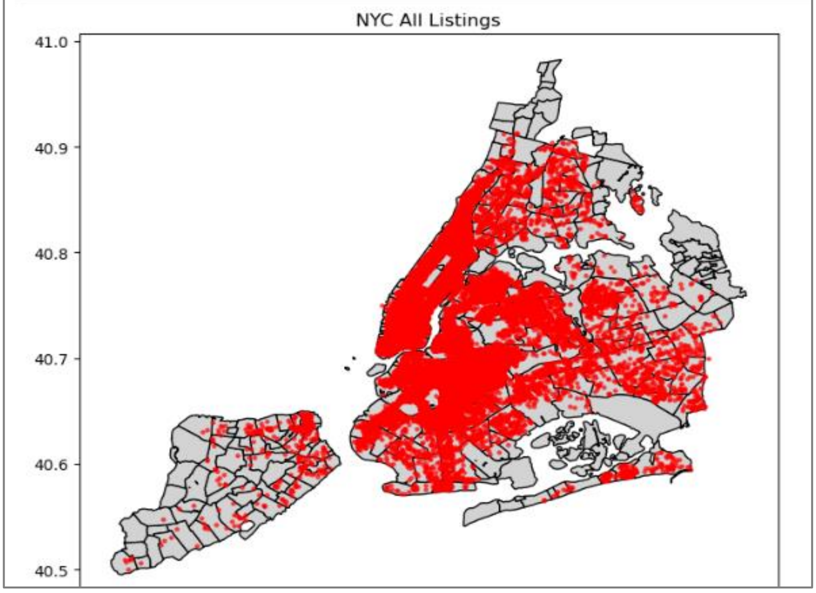
- Created a GeoDataFrame from .csv (lat long to point)

```python
# Create a GeoDataFrame from csv (lat long to point)
nyc_zip = [Point(xy) for xy in zip(df['longitude'], df['latitude'])]
nyc_point = gpd.GeoDataFrame(df, geometry=nyc_zip)
nyc_point.plot()
```

```
<Axes: >
```

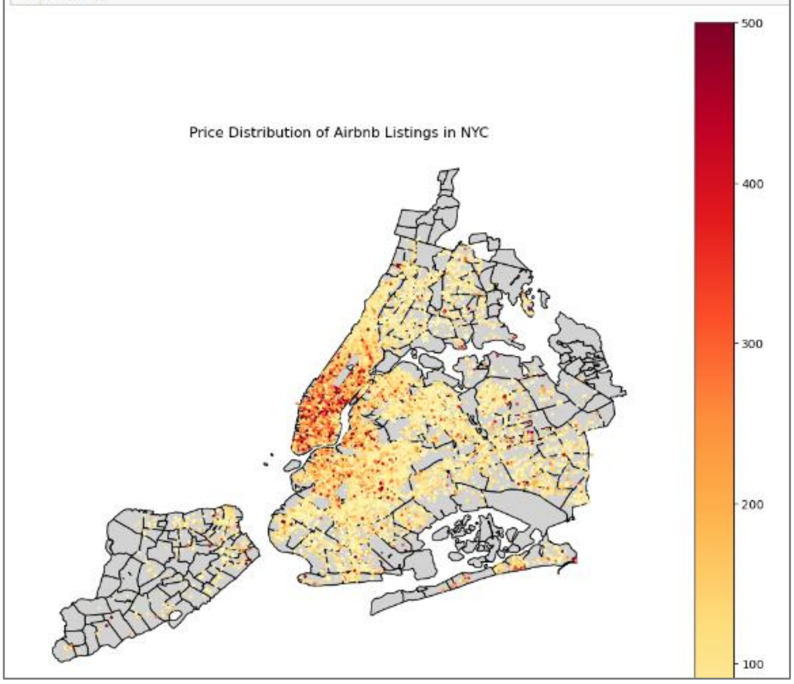
- Loaded NYC map shapefile over listing point

```python
# Loaded a NYC map shapefile
nyc_area = gpd.read_file('NYC_Boundary.shp')

ax = nyc_area.plot(figsize=(8, 7), color='lightgrey', edgecolor='black')

# Overlay points on top of the polygons
nyc_point.plot(ax=ax, color='red', markersize=5, alpha=0.6)
plt.title('NYC All Listings')
plt.show()
```



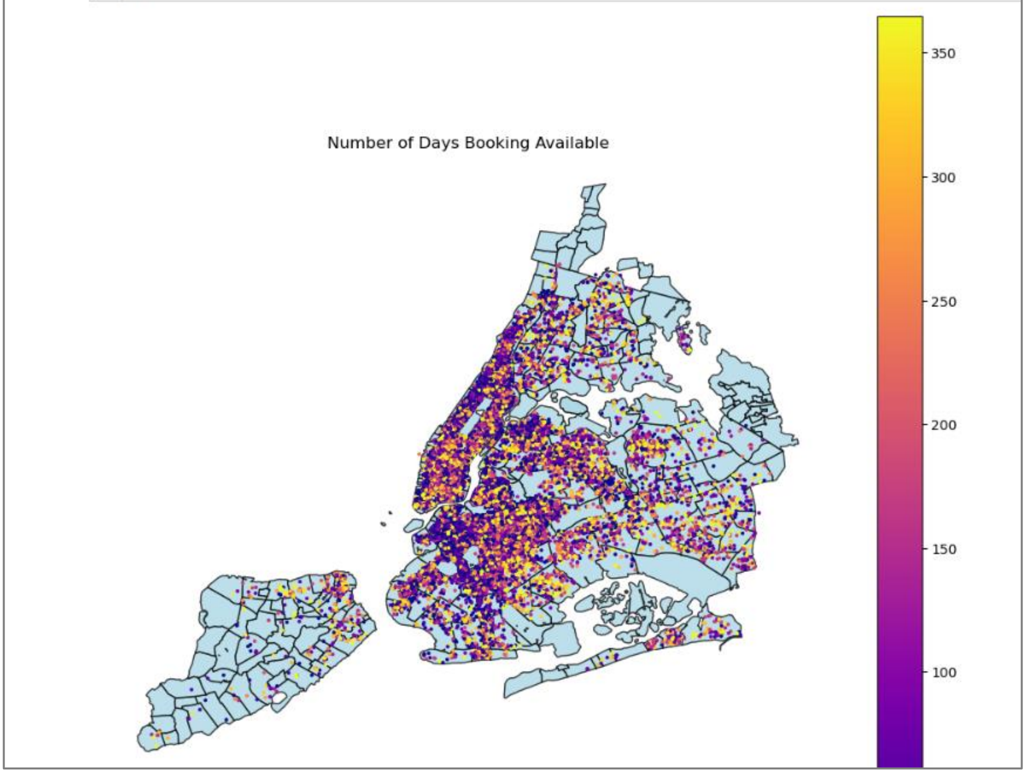
- Price Distribution

```python
fig, ax = plt.subplots(figsize=(12, 12))

# Plot the NYC area polygons with a base color
nyc_area.plot(ax=ax, color='lightgrey', edgecolor='black')  # Base color for NYC area

# Plot the points, adjusting the colormap and setting custom color scale limits
# Set vmin and vmax to control the color mapping and emphasize high-price areas
nyc_point.plot(column='price', ax=ax, legend=True, cmap='YlOrRd', markersize=3, vmin=10, vmax=500)

# Title and other plot settings
plt.title("Price Distribution of Airbnb Listings in NYC")
plt.axis('off')
plt.show()
```



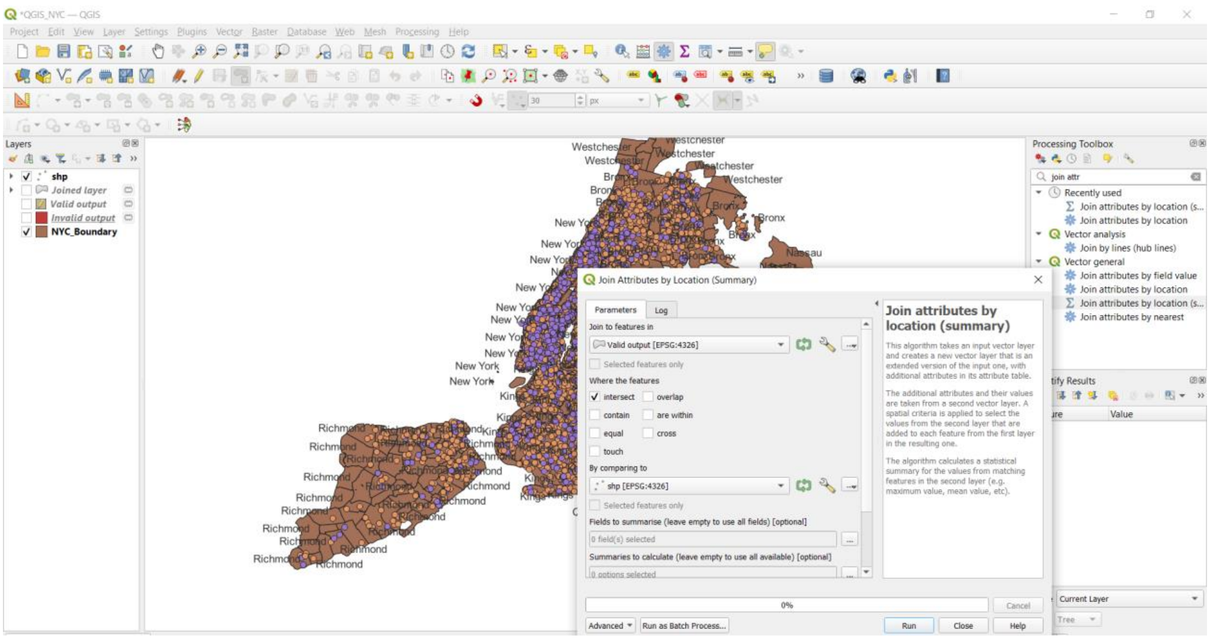
- Number of Days Booking Available

```
In [37]: fig,ax = plt.subplots(figsize=(12,12))
         nyc_area.plot(ax=ax,alpha=0.8,edgecolor='black',color='lightblue')
         nyc_point.plot(column='availability_365',ax=ax,legend=True,cmap='plasma',markersize=3)

         plt.title("Number of Days Booking Available")
         plt.axis('off')
         plt.show()
```



- Join attribute in QGIS (point and neighborhood boundary layer)

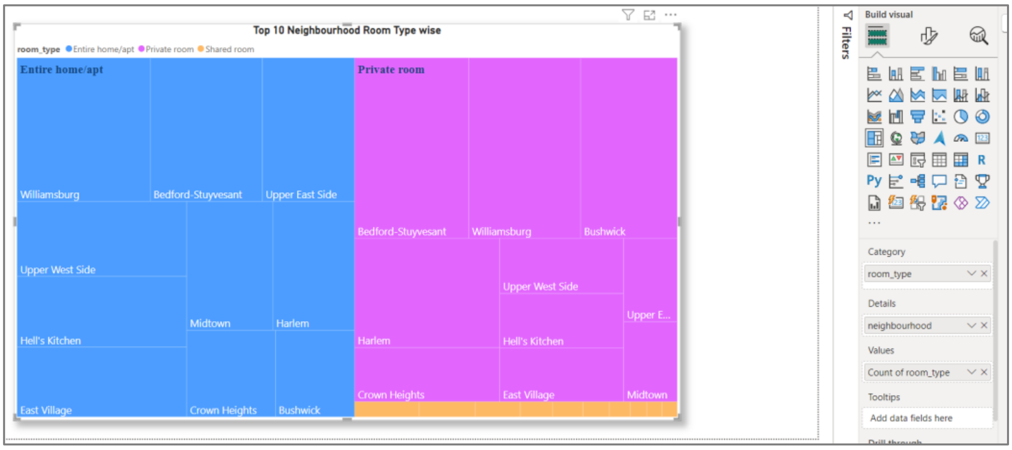
To identify Unpopular Properties, we created new columns as “Low_review” where 1= 0 review & 0 = other reviews. we sorted out properties with No Reviews and showed them on a QGIS map.



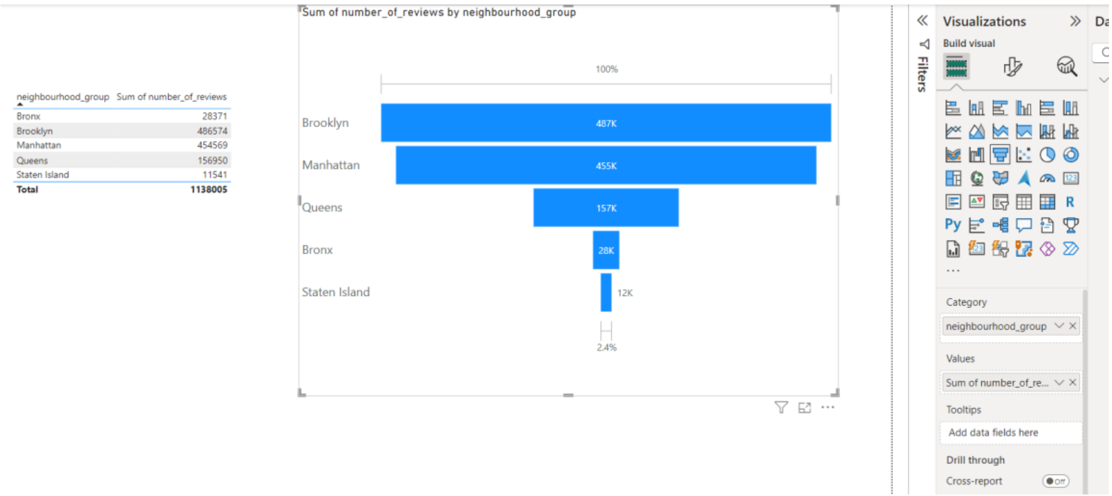
Power BI, Tableau

15. Importing data in Power BI, Tableau & Analysis

- To analyze which type of host to acquire more and where, we analyzed the Top 10 neighborhoods with the type of rooms to count of listings. the following chart shows which listings are available and where they are.

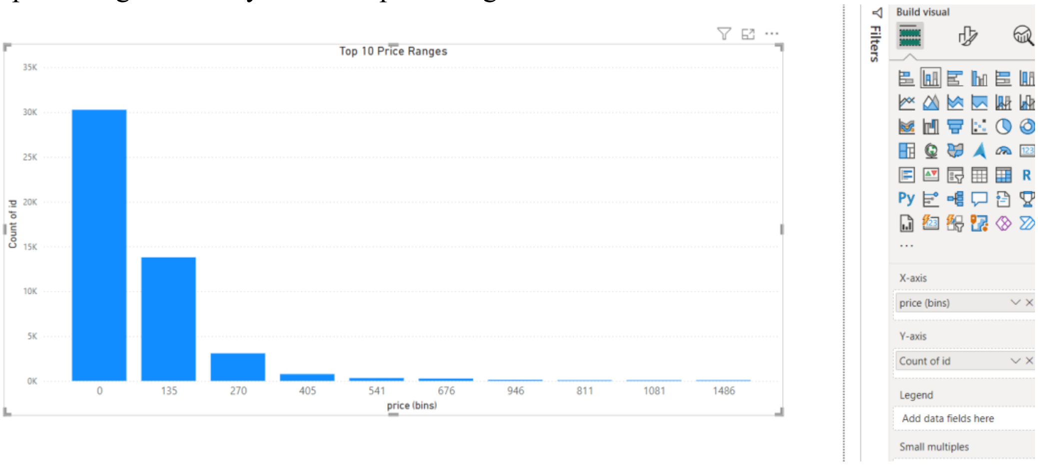


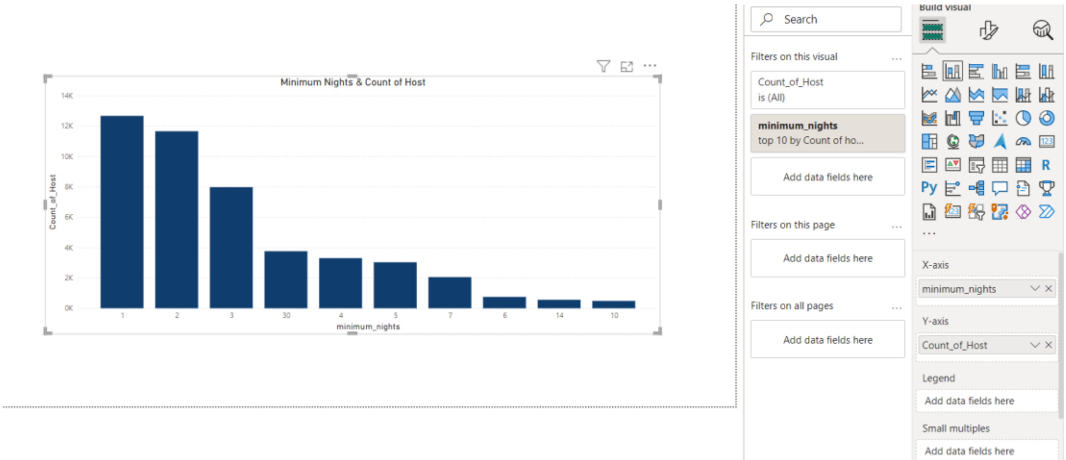
- Neighborhoods group by count of listings, Here we understand listing count with their size of shape.



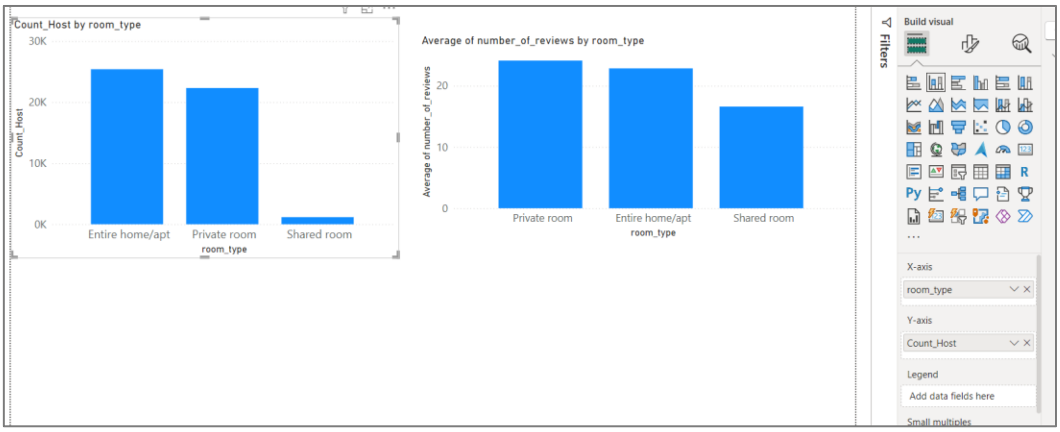
Top 10 Neighborhoods by **Sum of No. of reviews**

- To know the categorization of customers based on their preferences we created bins of price ranges & to know preferred stay for days. There was an outlier in the data so filtered as top 10 price ranges to easily visualize price ranges.

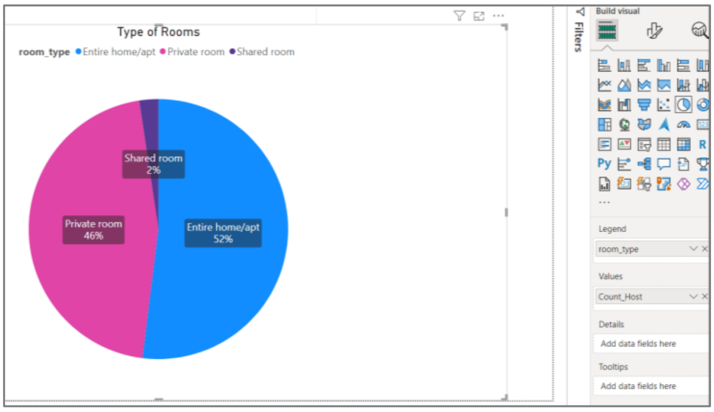


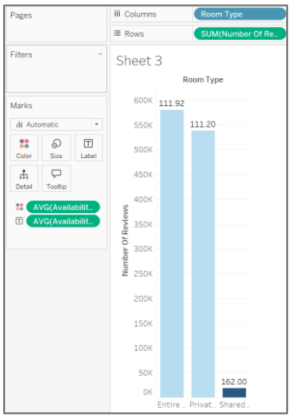


- To know The various kinds of properties that exist w.r.t. customer preferences, we calculated count of host by room type & Average review room type

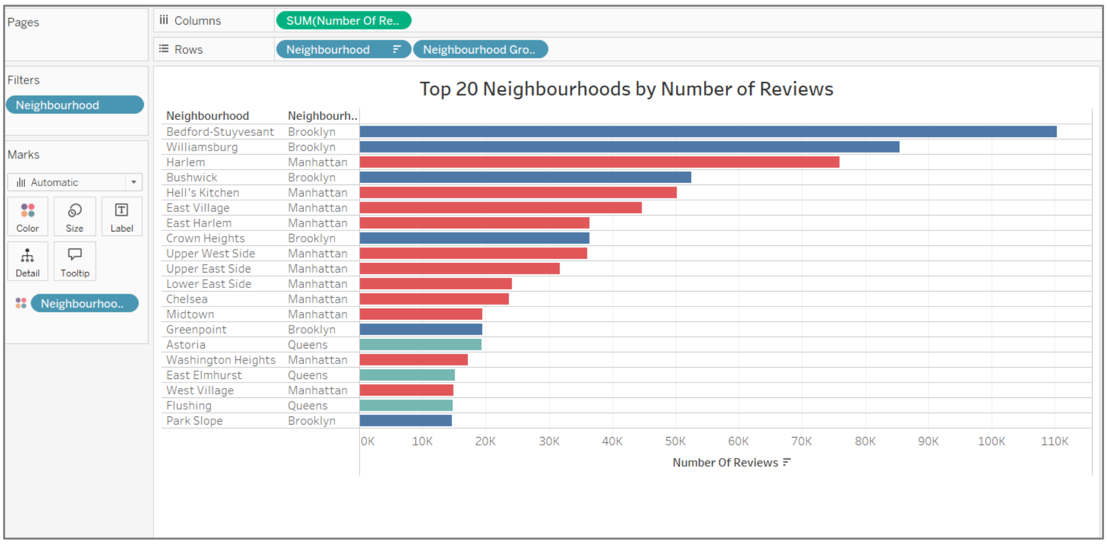


- To know adjustments in the existing properties to make it more customer-oriented in the first pie-chart share of different room type presented. In second plot bar chart created with room type in x axis, sum of number of reviews on y axis, the average availability given to text labels to bars & color shades according to average availability.

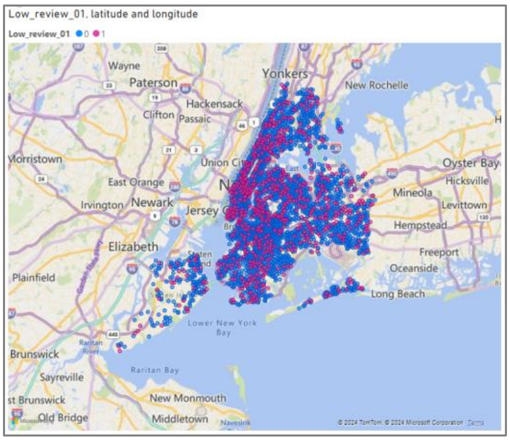




- When calculating the most popular localities in New York, we used a bar chart here, with the sum of the number of reviews in columns and the neighborhood and neighborhood group in rows. We then filtered the top 20 Properties by the number of reviews.



- To identify Unpopular Properties, we created new columns as "Low_review" where 1= 0 review & 0 = other reviews. we sorted out properties with No Reviews and showed them on a location map. The map highlights unpopular properties. with this we presented Unpopular Neighbourhood Group, top 20 Unpopular Properties by



neighbourhood	neighbourhood_group	Sum of Low_review_01
Williamsburg	Brooklyn	757
Bedford-Stuyvesant	Brooklyn	573
Midtown	Manhattan	559
Bushwick	Brooklyn	521
Upper West Side	Manhattan	489
Harlem	Manhattan	452
Hell's Kitchen	Manhattan	426
Upper East Side	Manhattan	393
East Village	Manhattan	363
Crown Heights	Brooklyn	299
Chelsea	Manhattan	285
Greenpoint	Brooklyn	247
Financial District	Manhattan	235
Astoria	Queens	191
Washington Heights	Manhattan	178
Murray Hill	Manhattan	174
East Harlem	Manhattan	173
Lower East Side	Manhattan	173
West Village	Manhattan	158
Theater District	Manhattan	140
Kips Bay	Manhattan	133

- **Assumptions:**
  - Airbnb assumes that after covid-19 pandemic travel activity will increase.
  - Identified customer preferences using the number of reviews given by customers