# Airbnb Case Study

PPT-I

By:
Kunal Sadana

# Agenda

Objective

Background

Key Insights

Recommendations

Appendix

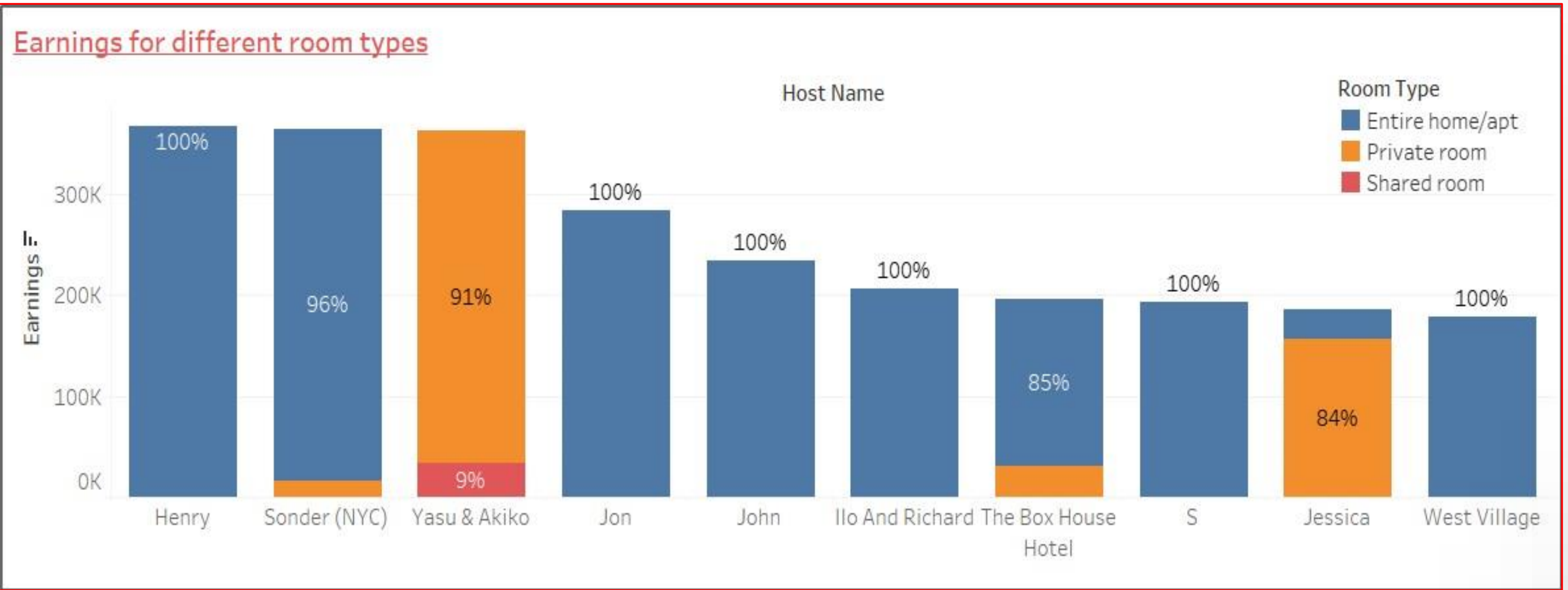
# Objective

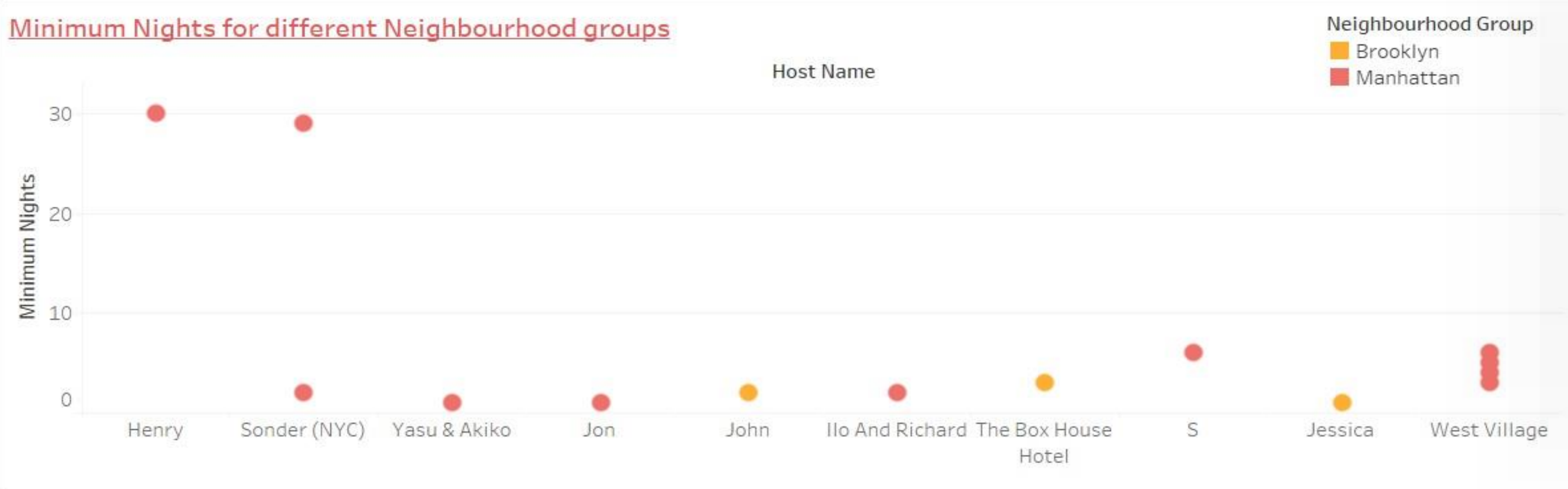
- Analyse the dataset consisting of various Airbnb listings in New York.
- Understand some important insights based on various attributes in the dataset so as to increase the revenue.

# Background

- For the past few months, Airbnb has seen a major decline in revenue.
- Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.

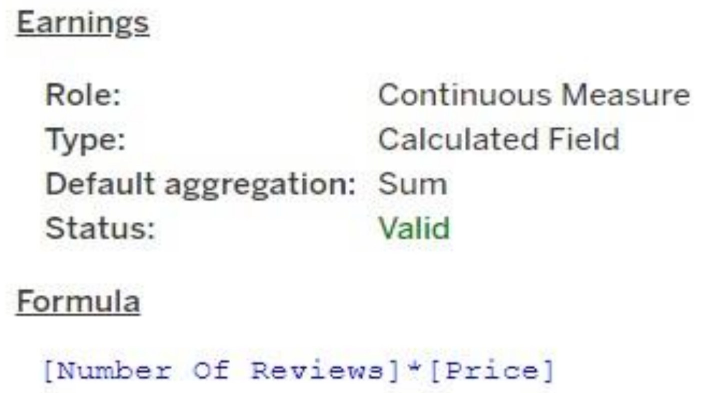
# Top Hosts



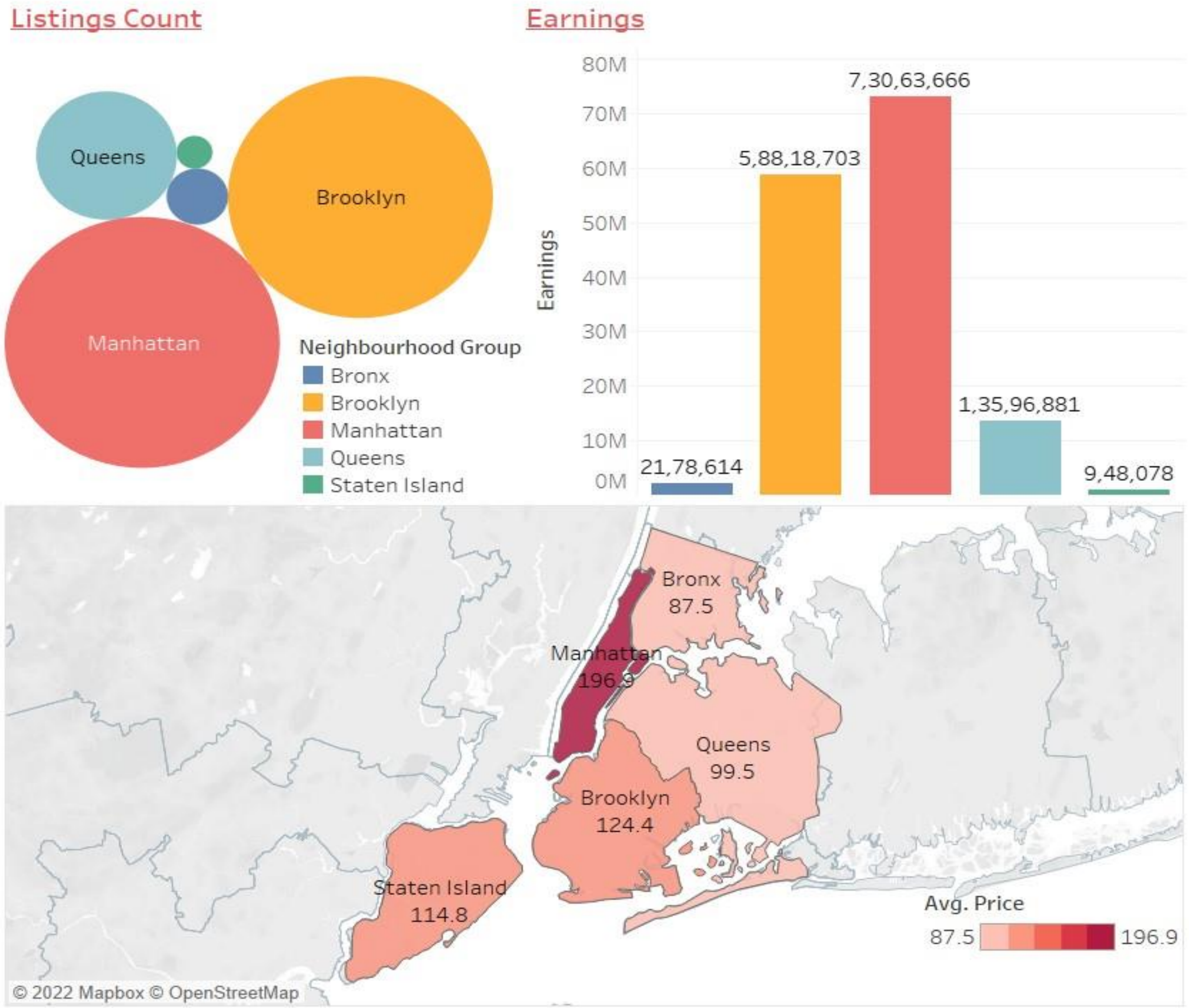




- Top 10 hosts by earnings.
- Most of them have Entire Home/apartment as room type.
- All have listings in Manhattan and Brooklyn area.

NOTE: "Earnings" is a calculated feature derived by given formula.



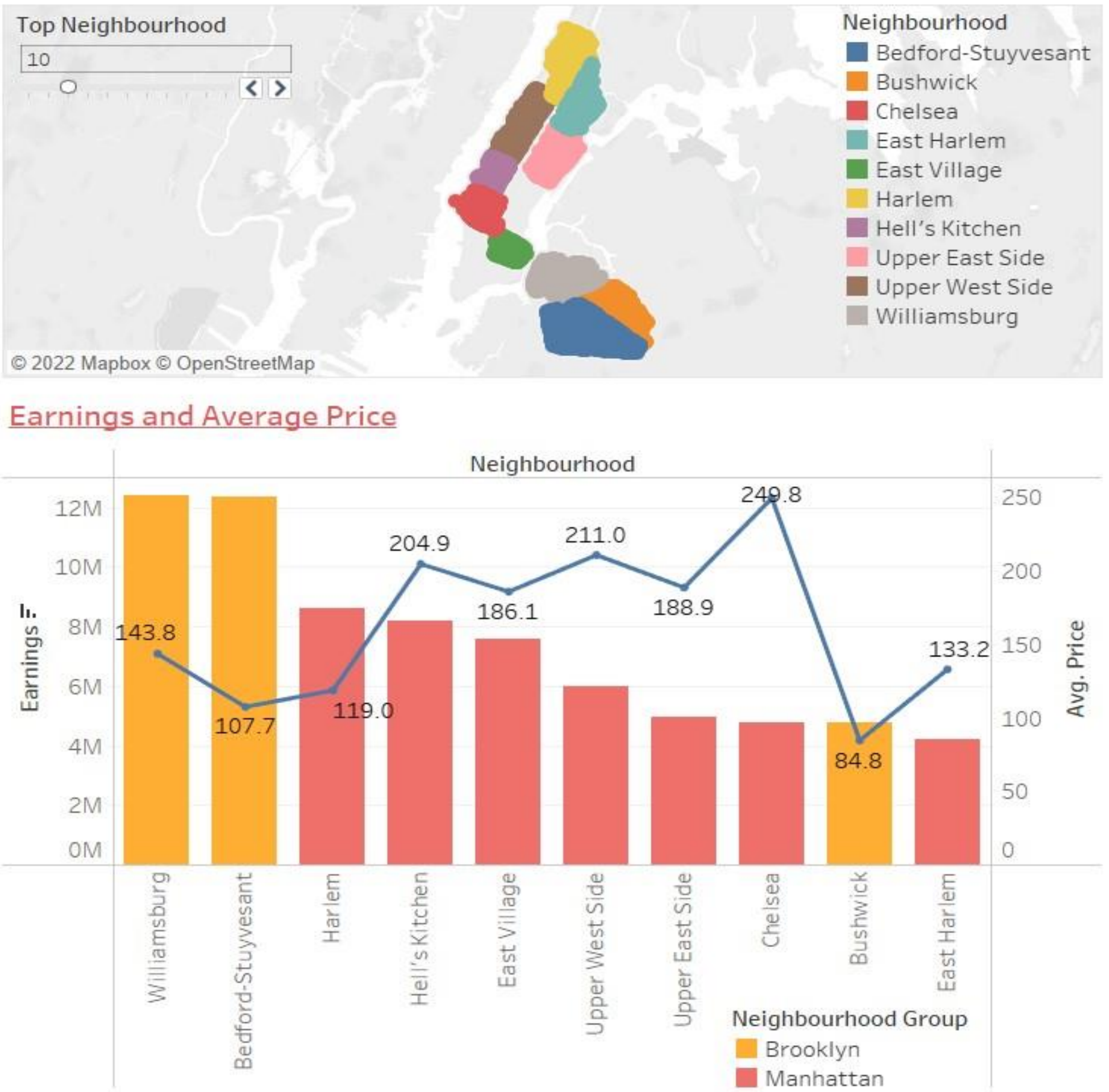
# Neighbourhood Groups




- The top New York City boroughs are Manhattan, Brooklyn, and Queens in this order.
- Manhattan has the most volume, highest Average Price and is the most popular destination for Airbnb.

As one of the quietest boroughs, Staten Island is ideal for a peaceful stay from the hustle and bustle of the city and can be promoted more.

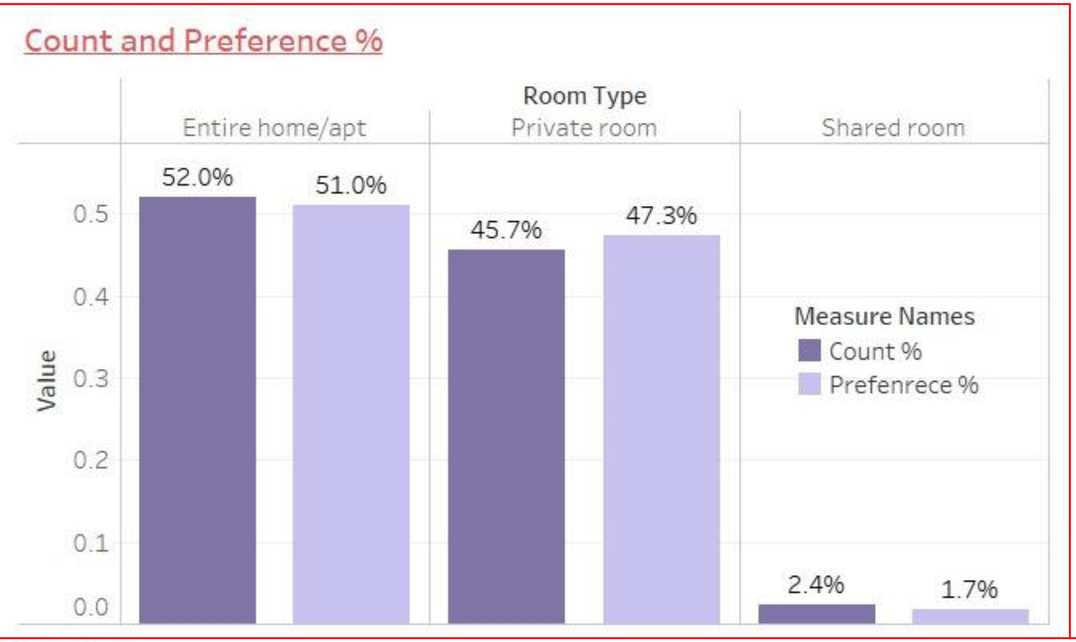
# Top Neighbourhoods

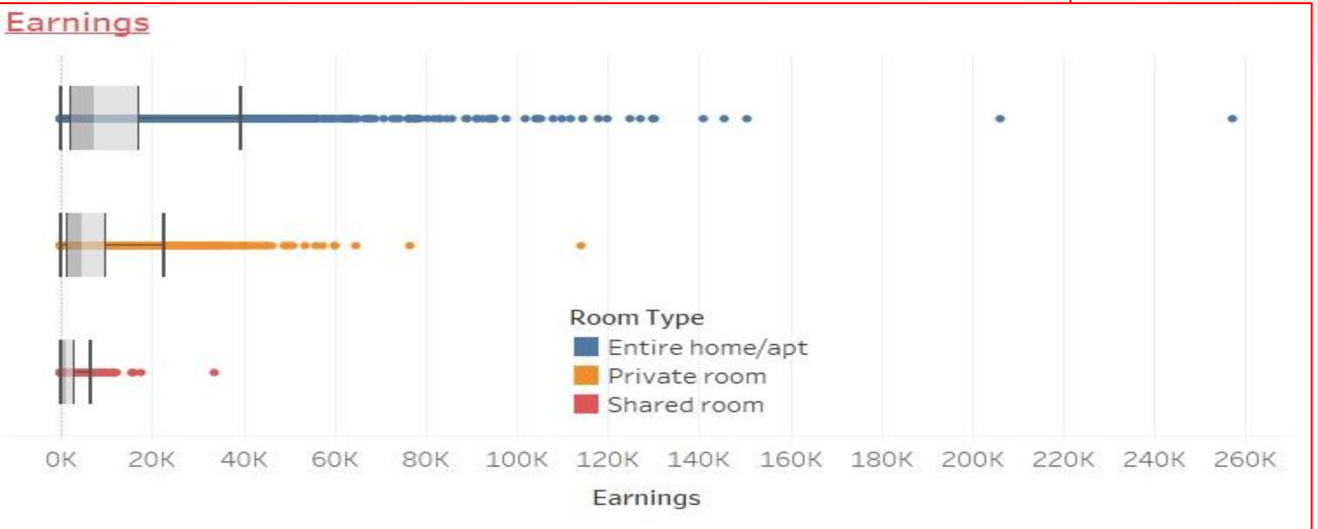



- Top 10 Neighbourhoods by earnings.
- Most of them lie in Manhattan followed by Brooklyn.
- Bushwik, Bedford-Stuyvesant and Harlem offer lowest Pricing.

Proximity to city attractions, the subway, and hip places are all factors that contribute to making these neighborhood famous and ideal for more investment.

# Property Types



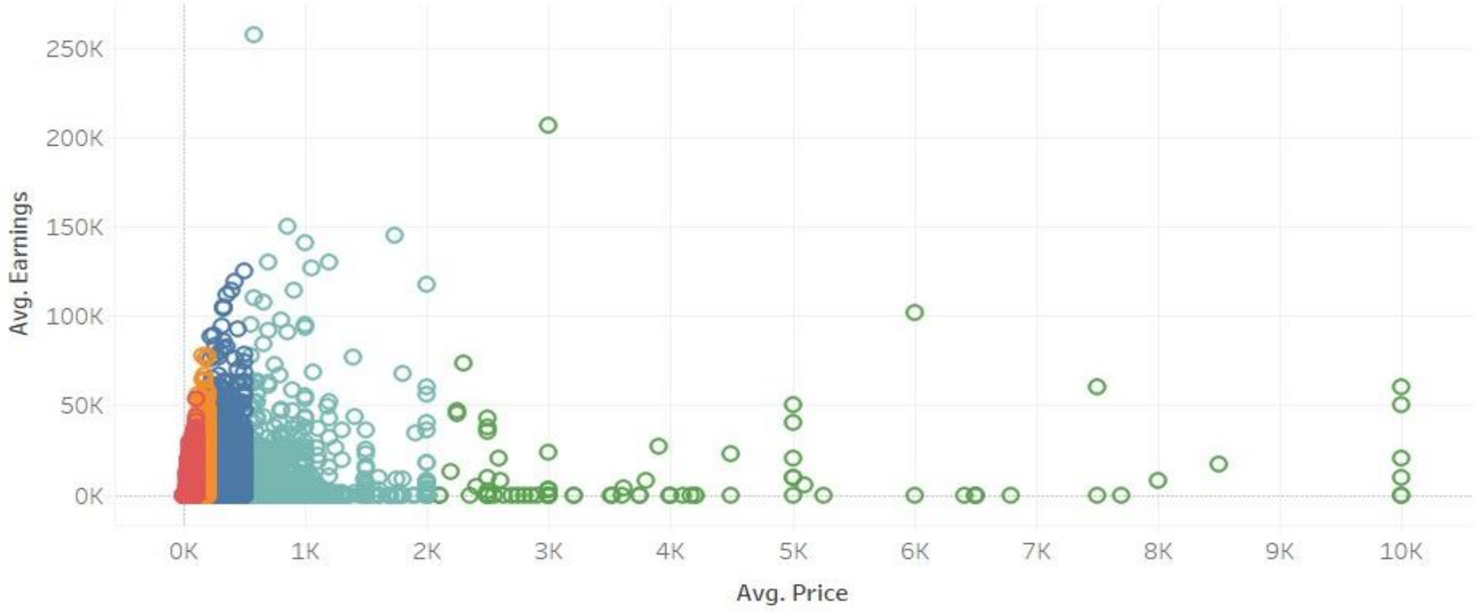




- Entire home/apartments have highest share of listing counts followed closely by Private rooms.
- Most of the customers prefer Entire home/apartments and Private rooms with very few preferring Shared rooms.
- Earnings via Entire home/apt is quite high making them ideal for more investments.

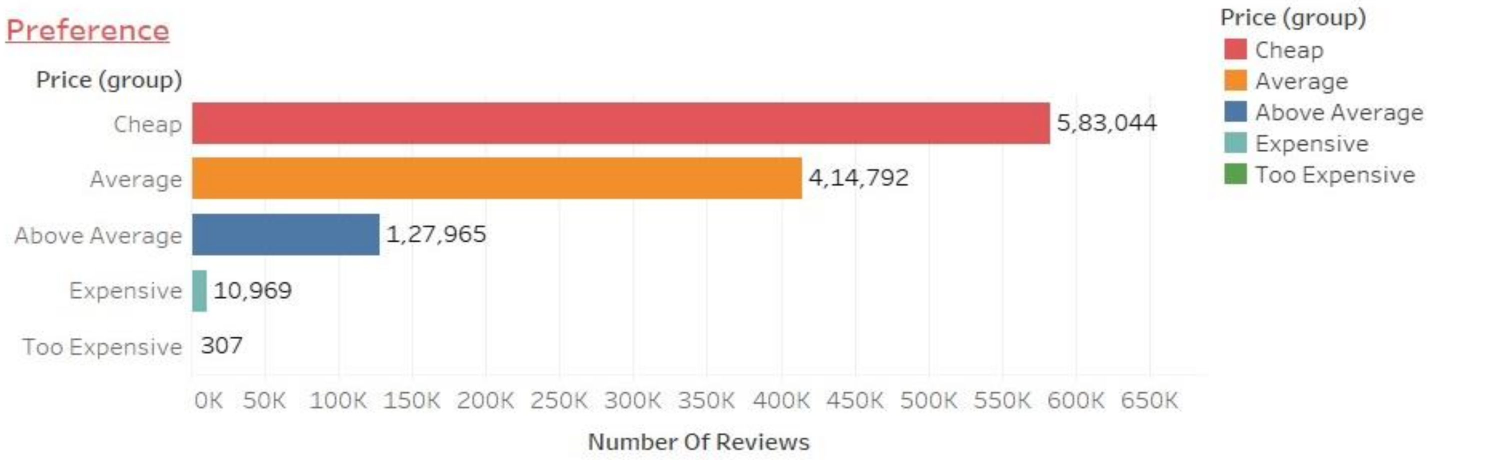
<u>NOTE</u>: "Preference" is calculated based on "SUM(Number of Reviews)"

# Pricing Ranges

Average Price vs Average Earnings




Preference




- Customers preference is inversely proportional to Pricing.
- Average and Above Average Pricing categories result in most Earnings.

<u>NOTE</u>: “Price (groups)” is a calculated feature where Price has been binned into following categories:

| Group | Range |
|---|---|
| Cheap | < 100$ |
| Average | 100$-200$ |
| Above Average | 200$-500$ |
| Expensive | 500$-2000$ |
| Too Expensive | > 2000$ |

# Availability for Booking



- A lot of properties have been listed as “Not Available”. Low Earning being one of the reason.
- People have preferred properties which are available for booking for more days in an year.
- Properties available for 6-9 months top in Average Earnings.

NOTE: “Availability 365 (groups)” is a calculated feature where Availability has been binned into following categories:

| Group | Range |
|---|---|
| Not Available | 0 |
| 1 month | Up to 30 days |
| 1-3 months | 30-90 days |
| 3-6 months | 90-180 days |
| 6-12 months | Above 180 days |

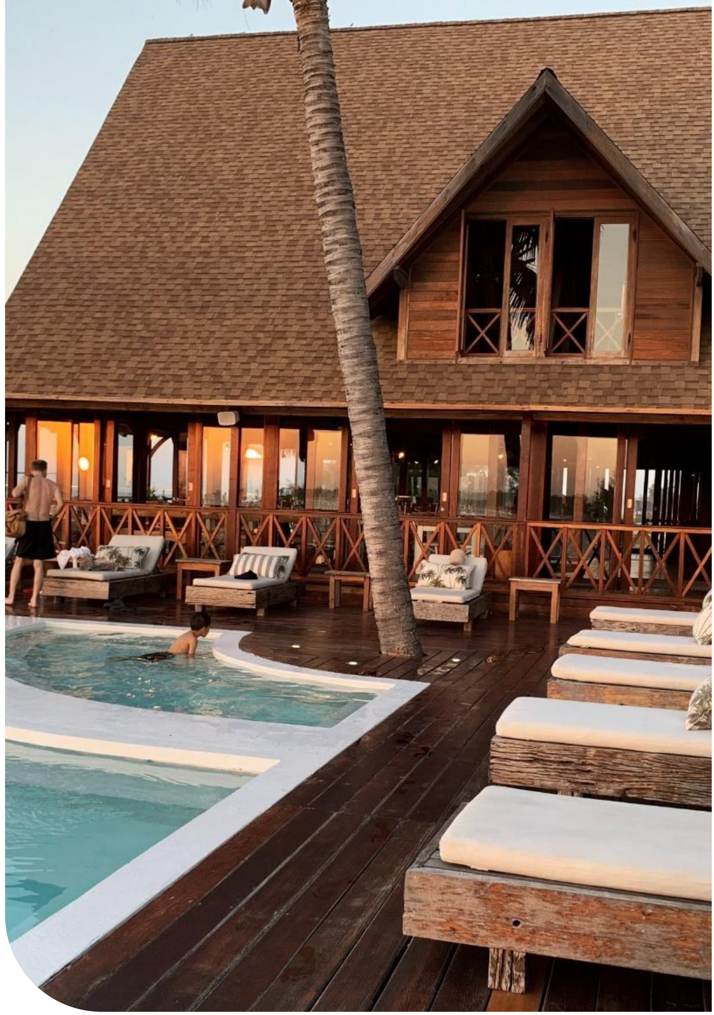
# Recommendations

## Following insights should be focused on for further development or investment:

- Hosts who have **less number of properties** listed and have **Entire home/apartments** as property type.
- **Manhattan** and **Brooklyn** are the top New York city boroughs.
- **Harlem** in Manhattan and **Bedford-Stuyvesant** in Brooklyn offer unique investment opportunity with less Average Price and high Earnings.
- Listings having **Average or Above Average Price** range.
- Listings available for booking for **6-9 months**.

# Appendix

- A thorough analysis of the given dataset was conducted.
- Data Methodology document has been attached with the file.



# Appendix

## Methodology

**Data**

- PYTHON (JUPYTER) used for initial Data understanding and processing.
- Here is a snapshot of 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 |

**Data Pre-processing**

1. NULL values treatment

```
# percentage of NULL values
airbnb.isnull().sum()*100/len(airbnb)

id                                0.000000
name                              0.032723
host_id                           0.000000
host_name                         0.042949
neighbourhood_group               0.000000
neighbourhood                     0.000000
latitude                          0.000000
longitude                         0.000000
room_type                         0.000000
price                             0.000000
minimum_nights                    0.000000
number_of_reviews                 0.000000
last_review                      20.558339
reviews_per_month                20.558339
calculated_host_listings_count    0.000000
availability_365                  0.000000
dtype: float64
```

- Deleting all rows where either "name" or "host_name" is missing

```
# Deleting all rows where either `name` or `host_name` is NULL
airbnb.dropna(subset=['name','host_name'], inplace=True)
```

- Filling NULL values with 0 for "reviews_per_month" since the same rows which had NULL values for this column had 0 for "number_of_reviews".

```
# Filling NULL values with 0 for `reviews_per_month`
airbnb.reviews_per_month.fillna(0, inplace=True)
```

2. Conversion of datatype
   - Converting "last_review" to datetime

```
# Converting `last_review` to datetype
airbnb.last_review = pd.to_datetime(airbnb.last_review)
```

3. Saving the file
   - Saving the pre-processed file to be used in TABLEAU for visualisation

```
# creating a new file to be used for visualisation in TABLEAU
airbnb.to_csv('airbnb_treated.csv', index=False)
```

**Data Visualisation**

TABLEAU used for the Data Visualisation purposes

- Creation of new features/variables to help in better analysis via visualisation:
  a. Earnings: "Earnings" is a calculated feature derived by given formula.

Earnings

Role: Continuous Measure
Type: Calculated Field
Default aggregation: Sum
Status: Valid

Formula

[Number Of Reviews]*[Price]

  b. Availability 365 (groups): "Availability 365 (groups)" is a calculated feature where Availability has been binned into following categories:

Availability 365 (group)

Role: Discrete Dimension
Type: Ad-hoc group
Remote column: [airbnb_treated.csv].[availability_365]
Remote type: Eight-byte, signed integer
Contains NULL: Unknown
Locale:
Sort flags: Case-sensitive
Status: Valid

Domain (6 members)

1 Month
1-3 Months
3-6 Months
6-9 Months
9-12 Months
Not available

| Group | Range |
|---|---|
| Not Available | 0 |
| 1 month | Up to 30 days |
| 1-3 months | 30-90 days |
| 3-6 months | 90-180 days |
| 6-12 months | Above 180 days |

  c. Price (groups): "Price (groups)" is a calculated feature where Price has been binned into following categories:

Price (group)

Role: Discrete Dimension
Type: Ad-hoc group
Remote column: [airbnb_treated.csv].[price]
Remote type: Eight-byte, signed integer
Contains NULL: Unknown
Locale:
Sort flags: Case-sensitive
Status: Valid

Domain (5 members)

Above Average
Average
Cheap
Expensive
Too Expensive

| Group | Range |
|---|---|
| Cheap | < 100$ |
| Average | 100$-200$ |
| Above Average | 200$-500$ |
| Expensive | 500$-2000$ |
| Too Expensive | > 2000$ |

- "Preference" of customers is based on "SUM (number_of_reviews)".
- Various types of plots used:
  - Bar graphs
  - Stacked Bar charts
  - Scatter plots
  - Bubble charts
  - Maps
  - Box plots
  - Pie Charts