# Storytelling Case Study: Analysis of New York Airbnb Dataset

BY: PRIYANSHI SRIVASTAVA, ATHIRA ANIL & KOWSHIK SHARMA

# Content

- BACKGROUND
- OBJECTIVE
- DATA ANALYSIS
- RECOMMENDATIONS
- APPENDIX
  - DATA SOURCES
  - DATA METHODOLOGY
  - DATA MODEL ASSUMPTION

# 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.

The different leaders at Airbnb want to understand some important insights based on various attributes in the dataset as to increase the revenue such as -

1. Which type of hosts to acquire more and where?
2. The categorization of customers based on their preferences.
    a. What are the neighborhoods they need to target?
    b. What is the pricing ranges preferred by customers?
    c. The various kinds of properties that exist w.r.t. customer preferences.
    d. Adjustments in the existing properties to make it more customer-oriented.
3. What are the most popular localities and properties in New York currently?
4. How to get unpopular properties more traction?

# Objective

To conduct analysis on dataset consisting of various Airbnb listings in New York.

Gain answers for important insights based on various attributes in the dataset so as to increase the revenue.

To process, analyze and share findings by data visualization and statistical techniques.

# Data Cleaning

- Eliminating NULL values and dropping certain columns that are not efficient to the dataset.

```
# Calculating the missing values in the dataset
airbnb.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
```

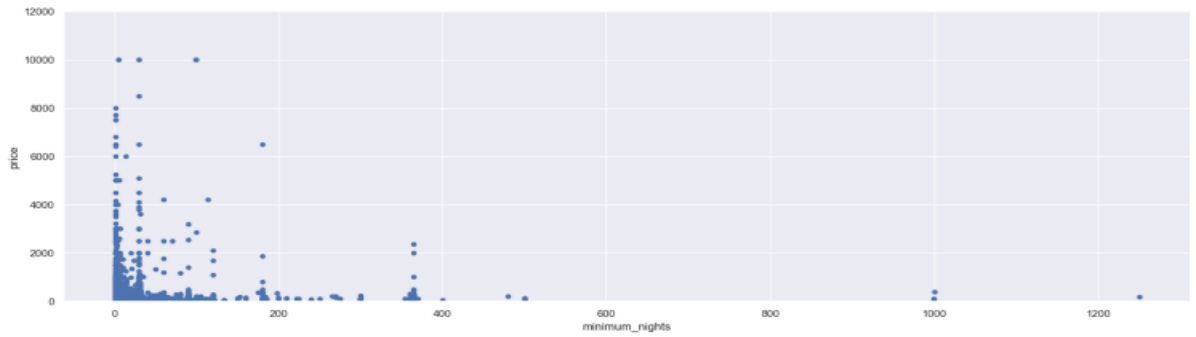
```
# Now we have the missing values, there are certain columns that are not efficient to the dataset
airbnb.drop(['id','name','last_review'], axis = 1, inplace = True)
```

# Analysis

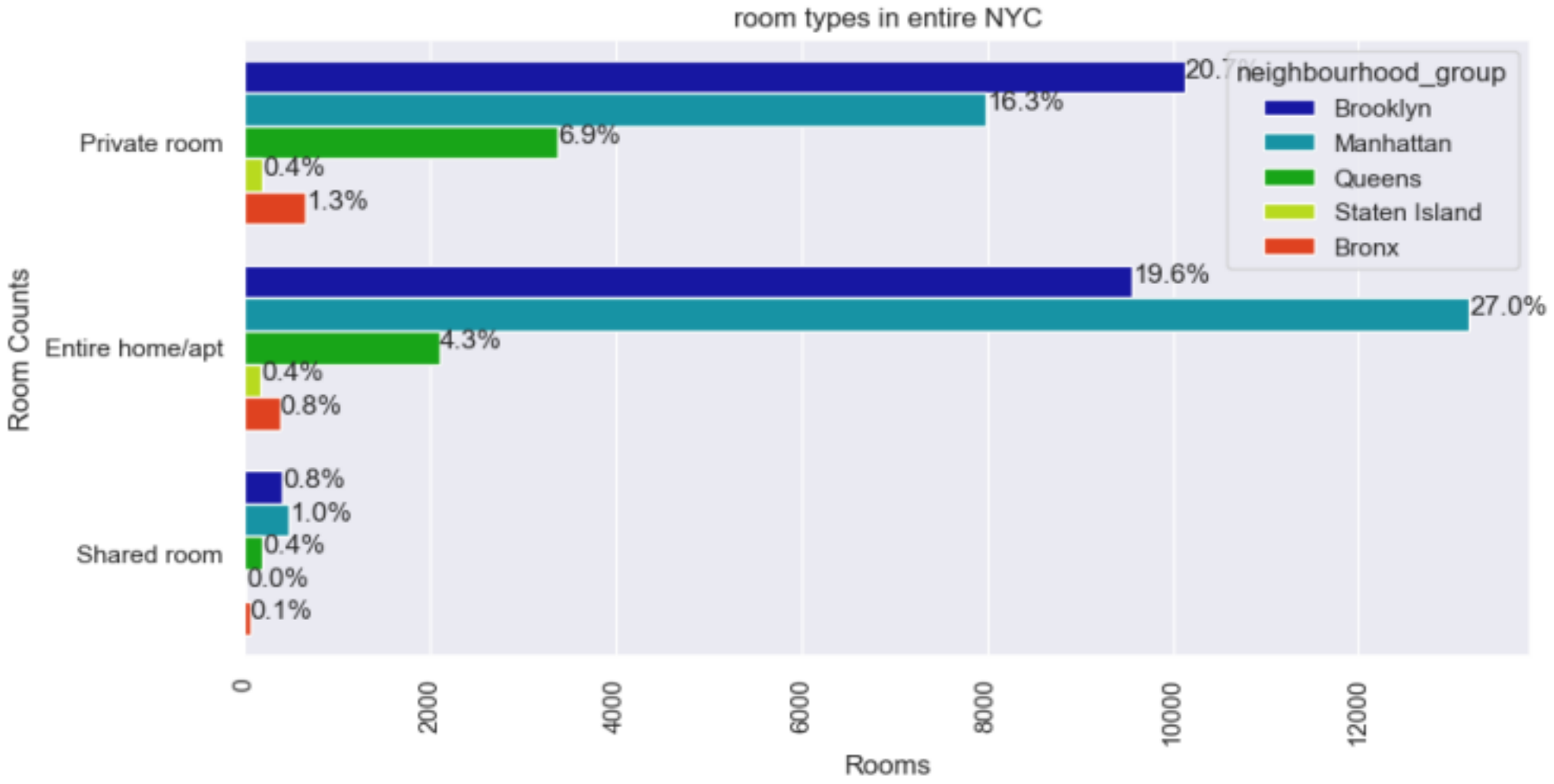
In [41]:
```
var='minimum_nights'

data=pd.concat([abnb_sur['price'],abnb_sur[var]],axis=1)
data.plot.scatter(x=var,y='price',ylim=(0,12000))
```

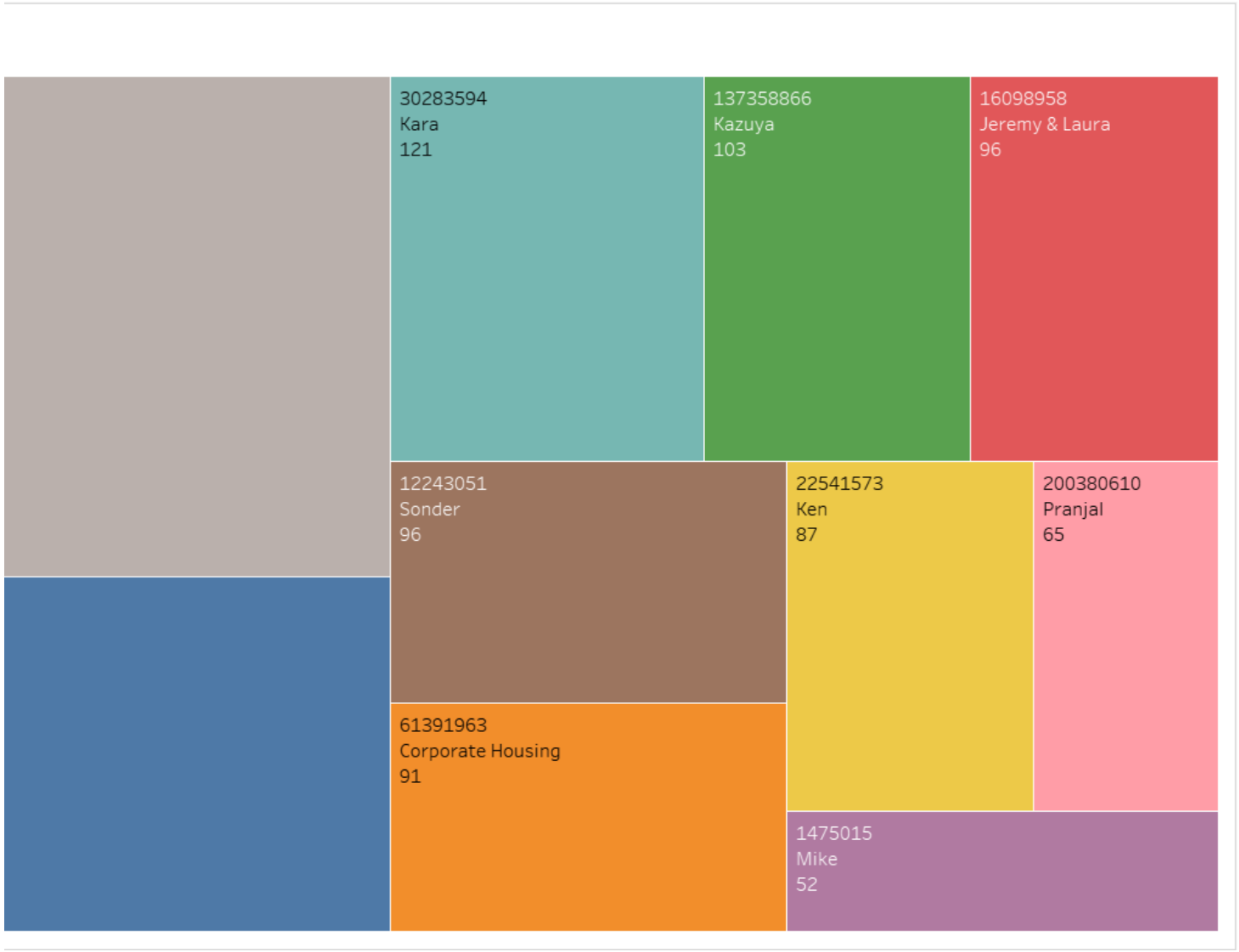
```
*c* argument looks like a single numeric RGB or RGBA sequence, which should be avo
ided as value-mapping will have precedence in case its length matches with *x* & *
y*.  Please use the *color* keyword-argument or provide a 2D array with a single r
ow if you intend to specify the same RGB or RGBA value for all points.
```

Out[41]:
```
<AxesSubplot:xlabel='minimum_nights', ylabel='price'>
```



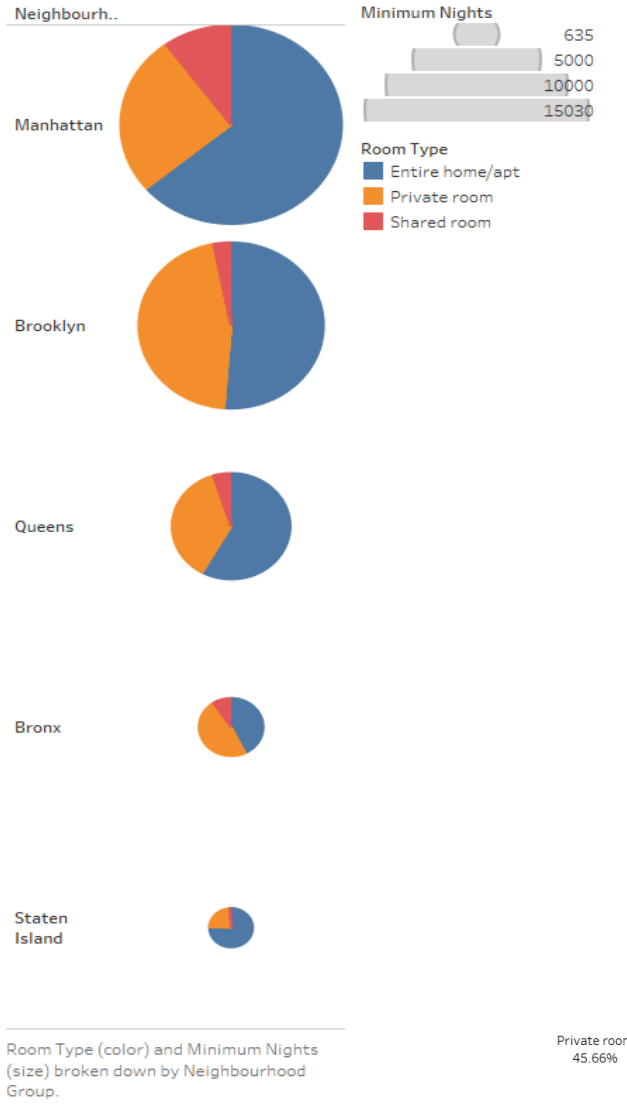

In [45]:
```
abnb_sur.boxplot(column=['price'])
plt.show()
```

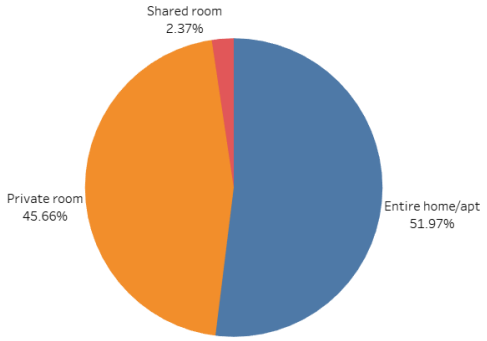

room types in entire NYC
neighbourhood_group
Brooklyn
Manhattan
Queens
Staten Island
Bronx
Private room
Entire home/apt
Shared room
Room Counts
20.7%
16.3%
6.9%
0.4%
1.3%
19.6%
27.0%
4.3%
0.4%
0.8%
0.8%
1.0%
0.4%
0.0%
0.1%
0
2000
4000
6000
8000
10000
12000
Rooms




# Which type of hosts to acquire more and where?

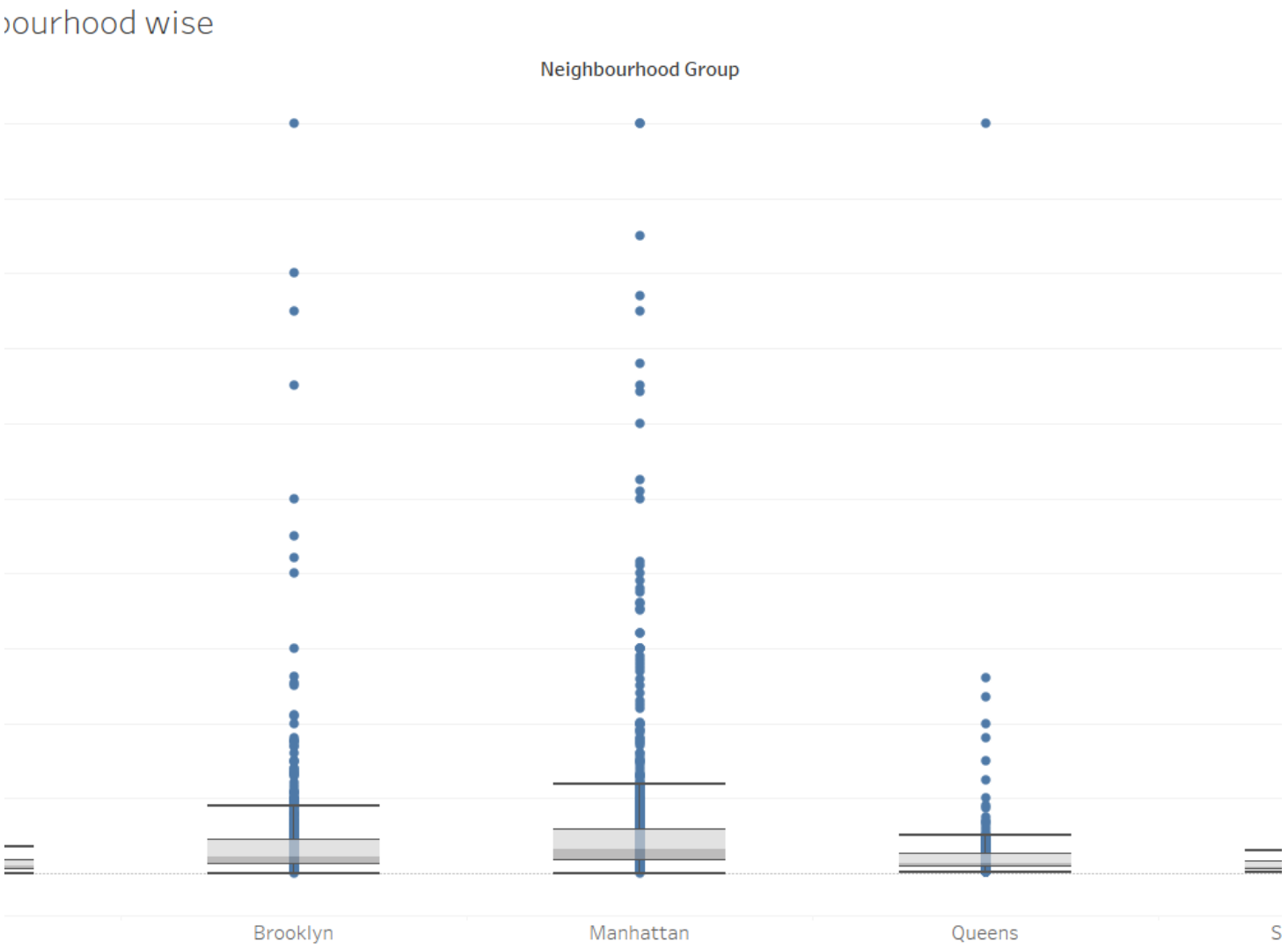
- Host Sonder (id 219517861), has been booked most number of times i.e. 327.
- Host Blue ground is the second popular host.
- Then there are other hosts like Kara, Ken, Pranjal, Jeremy and Mike that fall under top 10 hosts.



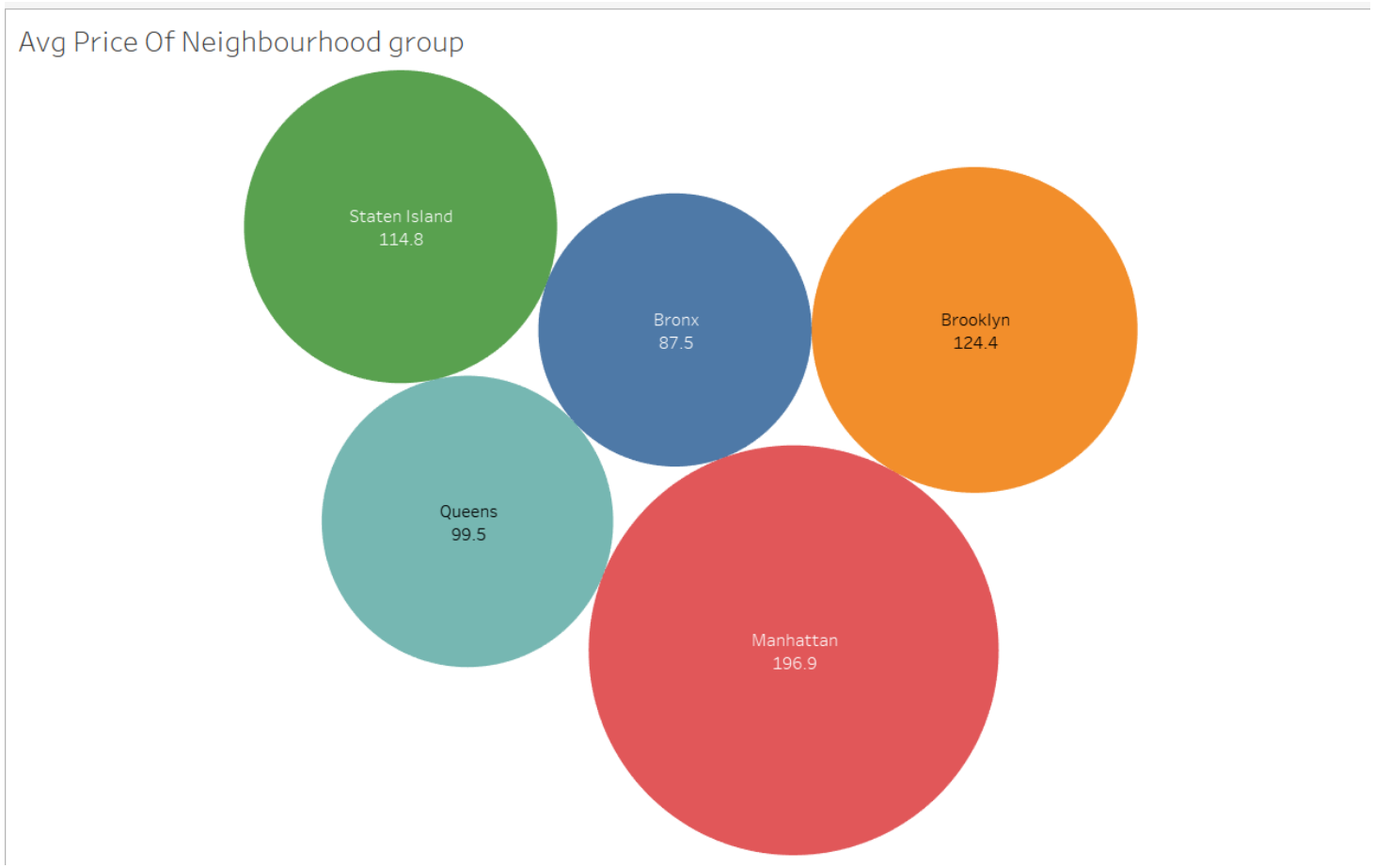

# Room type with respect to Neighbourhood group

- There are three types of rooms - **Entire home/Apartment, Private room & shared room.**
- Overall, customers appear to prefer **private rooms (45%) or entire homes (52%) in comparison to shared rooms (2.4%).**
- Airbnb can concentrate on promoting shared rooms with discounts to increase bookings and also acquire more private listings.
- Queens & Bronx contribute 60% each to private rooms, more than the combined ratio of 45% Whereas, Manhattan has a higher contribution in entire home (61%), compared to the combined ratio of 52%.



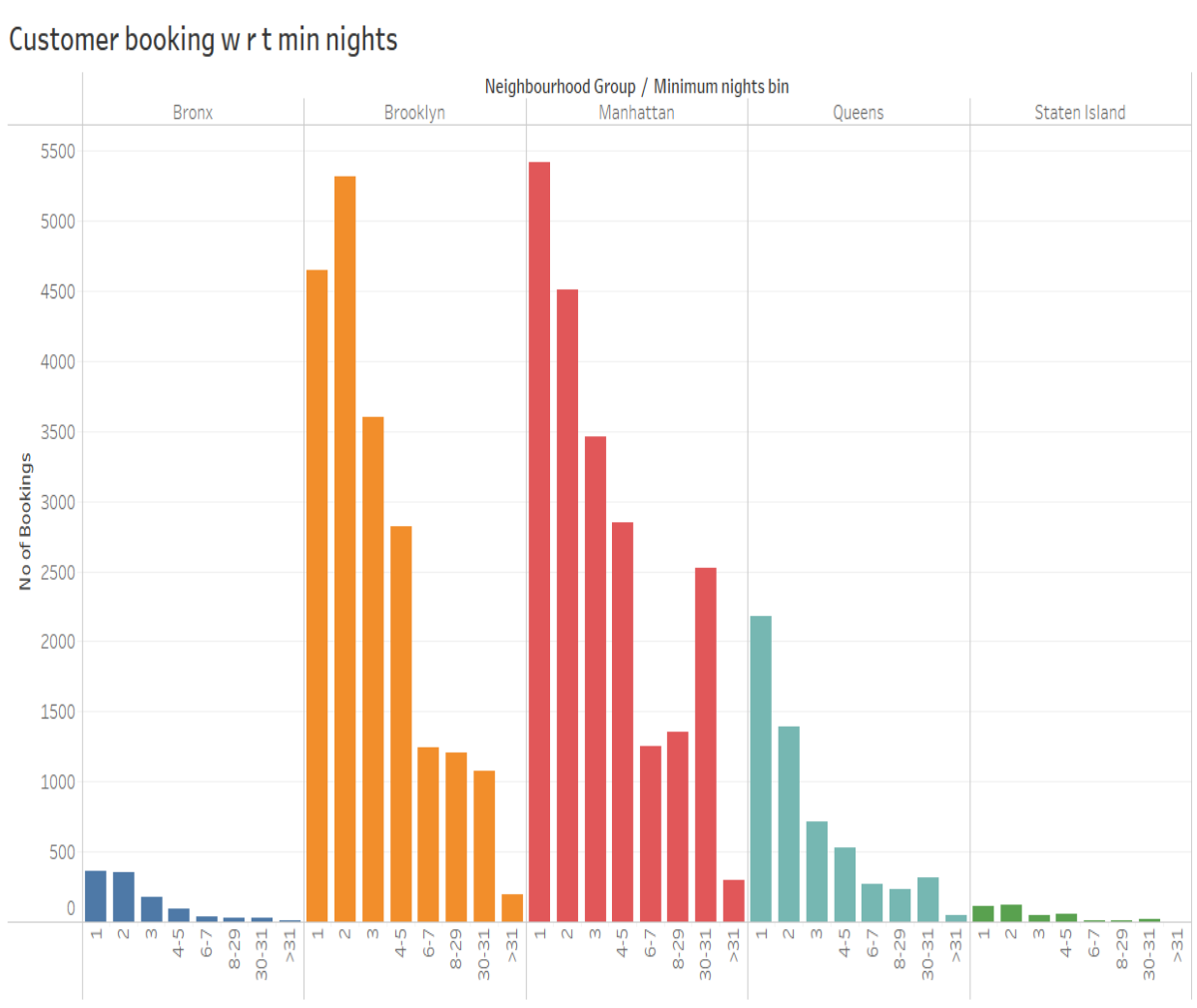

# Price Analysis Neighbourhood wise

- Most of the outliers in Price column are for Brooklyn and Manhattan.
- Also, Manhattan has the highest range of prices for the listings.
- Bronx is the cheapest of them all.
- We can see the median price of all neighbourhood groups lying between $ 80 to $ 300.
- Price was highly positively skewed so median was very close the lower quartile with some outliers as seen in the boxplot below.




# Average price of Neighbourhood groups

- The average price of listed properties in Manhattan is around 196.9, which is highest among all neighbourhoods.
- Average price for Brooklyn is second highest i.e. 124.4.
- Bronx appears to be an affordable neighbourhood as the average price is almost half than Manhattan's average price.




# Customer Booking with respect to minimum nights

The listings with Minimum nights 1-5 have the most number of bookings. We can see a prominent spike in 30 days, this would be because customers would rent out on a monthly basis.

After 30 days, we can also see small spikes, this can also be explained by the monthly rent taking trend.

**Manhattan & Queens** have higher number of 30 day bookings compared to the others. The reason could be either tourists booking long stays or mid-level employees who opt for budget bookings due company visits

# Recommendations

- Airbnb can concentrate on promoting shared rooms with discounts to increase bookings and also acquire more private listings.
- Ample amount and variety of visuals have been used in the presentations for the stake-holders.
- Data collection team should collect data about review scores so that it can strengthen the later analysis.
- A clustering machine learning model to identify groups of similar objects in datasets with two or more variable quantities can be made.

# Appendix-Data Sources

The columns in the dataset are self-explanatory. These can be referred in the diagram given below to get a better idea of what each column signifies

| 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 |

# Appendix: Data Methodology

**Data Cleaning:**

1- Cleaned data to remove any missing values and duplicates.

2- Used python to handle NULL values, eliminated columns that are not efficient to dataset.

**Exploratory Data Analysis:** Once data is cleaned, EDA is done.

Used group aggregation, pivot table and other statistical methods.

**Gathering Insights:** After EDA, meaningful insights are drawn.

Created charts and visualizations to visualize the data for showcasing the insights in a better way.

# Appendix-Data Assumptions

Classified the variables into different types – categorical, numeric, location and time.

```
Categorical Variables:
    - room_type
    - neighbourhood_group
    - neighbourhood

Continous Variables(Numerical):
    - Price
    - minimum_nights
    - number_of_reviews
    - reviews_per_month
    - calculated_host_listings_count
    - availability_365
- Continous Variables could be binned in to groups too

Location Varibles:
    - latitude
    - longitude

Time Varibale:
    - last_review
```