Methodology Document

Storytelling Case Study: Airbnb, NYC

Submitted By:

NAVEEN

NARGIS

MUKESH

1. Research Problem

- For the past few months, Airbnb business has seen a significant decline in revenue due to travel restrictions because of the Covid-19 pandemic.
- The revenue took the largest hit in NYC in the Q2 of 2020.
- Now that the restrictions have started lifting and people have started traveling, Airbnb wants to make sure it is fully prepared for the change.

2. Objectives

- Improve our strategies to revive the impact of Covid-19 on the economic and market conditions of Airbnb, NYC.
- Understand the customer preference and user experience trends for Airbnb, NYC.
- Provide recommendations for new acquisitions and improve customer experience

3. Data Assumptions

- Assumed that the data prior to the Covid-19 period was achieving the desired goals.
- Airbnb wants to continue its business in NYC and has no plans of expanding to other territories.
- The strategies decided were considered keeping in mind that there will be no further travel restrictions.

4. Data Methodology

- Tools used – Python for analysis, Tableau for visualization

Data Understanding and Preparation:

- The following relevant libraries were imported.

```
#importing the required libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings(action='ignore')
```

- The dataset was loaded, datatypes of variables were checked and along with that the dimensions and size of the data frame was checked.

In [2]:
```
#using pandas library and 'read_csv' function to read csv file
dataf=pd.read_csv("AB_NYC_2019.csv")
#examine head
dataf.head(5)
```

Out[2]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	2
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

- # Handling Missing Values and Outliers:
  - The missing values and outliers were checked in the data frame.
  - The following columns had missing values – last review, reviews per month, host name, and name.
  - These columns had NaN values – last review and reviews per month indicating some listed properties didn’t receive reviews.
  - Missing values are imputed accordingly with median and mode

```
#looking to find out first what columns have null values
#using 'isnull' function will show us how many nulls are found in each column in dataset
print((100*dataf.isnull().mean()).sort_values().to_string())
```

```
id                                  0.000000
host_id                             0.000000
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
calculated_host_listings_count      0.000000
availability_365                    0.000000
name                                0.032723
host_name                           0.042949
last_review                        20.558339
reviews_per_month                  20.558339
```

- Dropping columns that are not significant for our future data predictions.

```
#dropping columns that are not significant or could be unethical to use for our future data exploration and predictions
dataf.drop(['id','last_review'], axis=1, inplace=True)
#examine the changes
dataf.head(10)
```

	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
0	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9
1	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45
2	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0
3	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270
4	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9

- The following columns had outliers - price, minimum nights, number_of_reviews, reviews_per_month, and calculated_host_listings_count and it was treated using capping.

```
#imputing missing values...
dataf['reviews_per_month'].median()
dataf['reviews_per_month'] = dataf['reviews_per_month'].fillna(0.72)
```

```
dataf.dtypes
```

```
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
reviews_per_month                  float64
calculated_host_listings_count       int64
availability_365                     int64
dtype: object
```

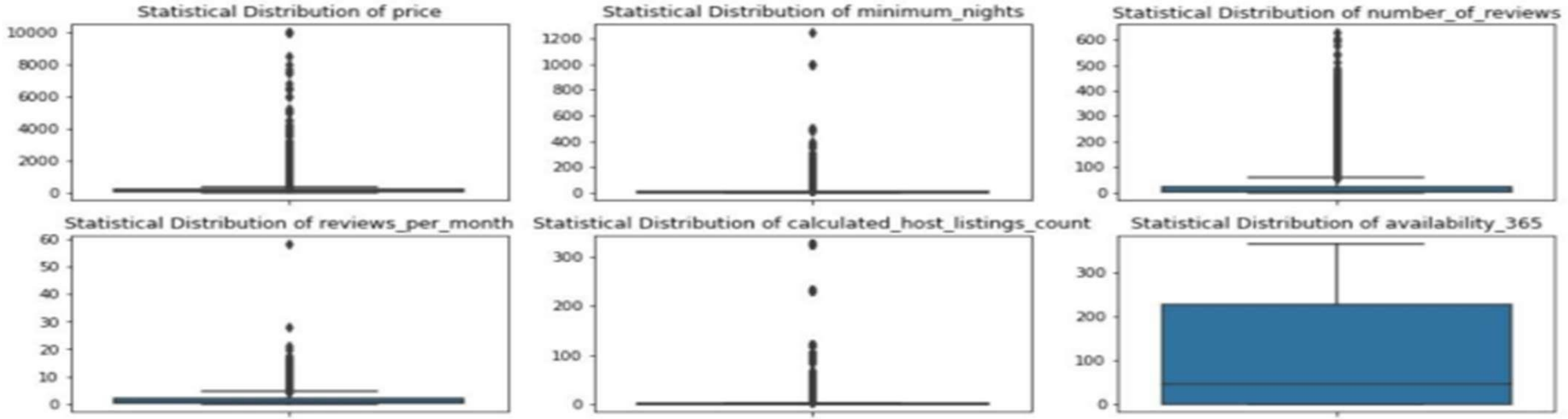
```
cat_cols = dataf.select_dtypes(include = ['object']).columns
cat_cols
```

```
Index(['name', 'host_name', 'neighbourhood_group', 'neighbourhood',
       'room_type'],
      dtype='object')
```

```
cont_cols = dataf.select_dtypes(include =['float','int']).columns
cont_cols
```

```
Index(['host_id', 'latitude', 'longitude', 'price', 'minimum_nights',
       'number_of_reviews', 'reviews_per_month',
       'calculated_host_listings_count', 'availability_365'],
      dtype='object')
```

```
# box plot for checking outliers
plt.figure(figsize=(16,6))
for i in enumerate(cnt):
    plt.subplot(2,3,i[0]+1)
    sns.boxplot(y=df[i[1]])
    plt.title("Statistical Distribution of "+i[1])
    plt.ylabel("")
```




```
# treating outliers with capping method
for i in cnt:
    q1=df1[i].describe()["25%"]
    q3=df1[i].describe()["75%"]
    iqr=q3-q1
    ub=q3+1.5*iqr
    df1[i]=np.where(df1[i]>ub,ub,df1[i])
```

5. Data Analysis And Visualization:

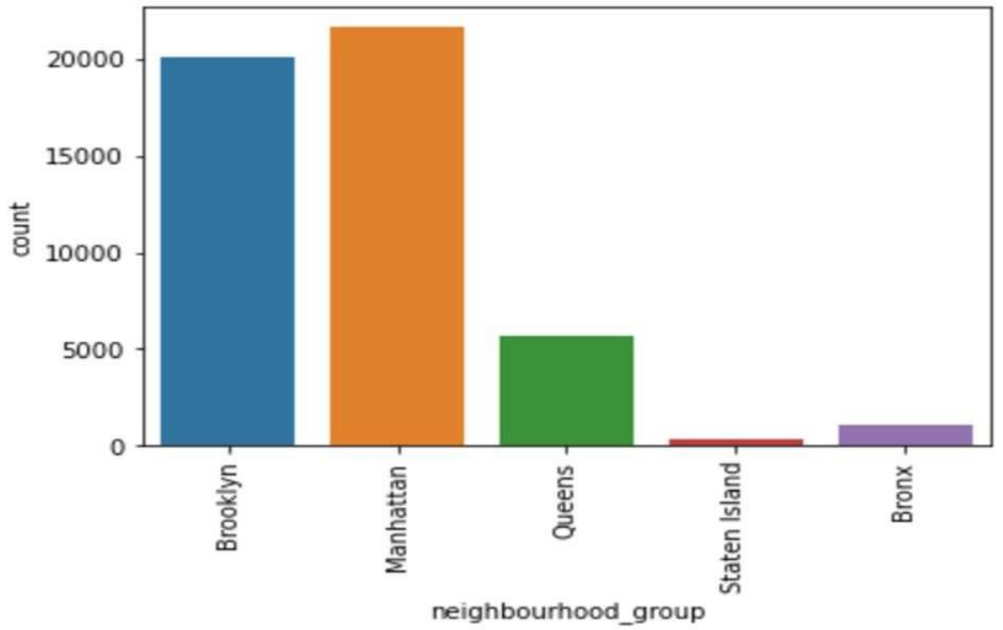
- Loading the data to csv file fo further visualization in tableau.

```
#Loading clean and balanced data to csv
dataf.to_csv('air_bnb_.csv', index=False)
```

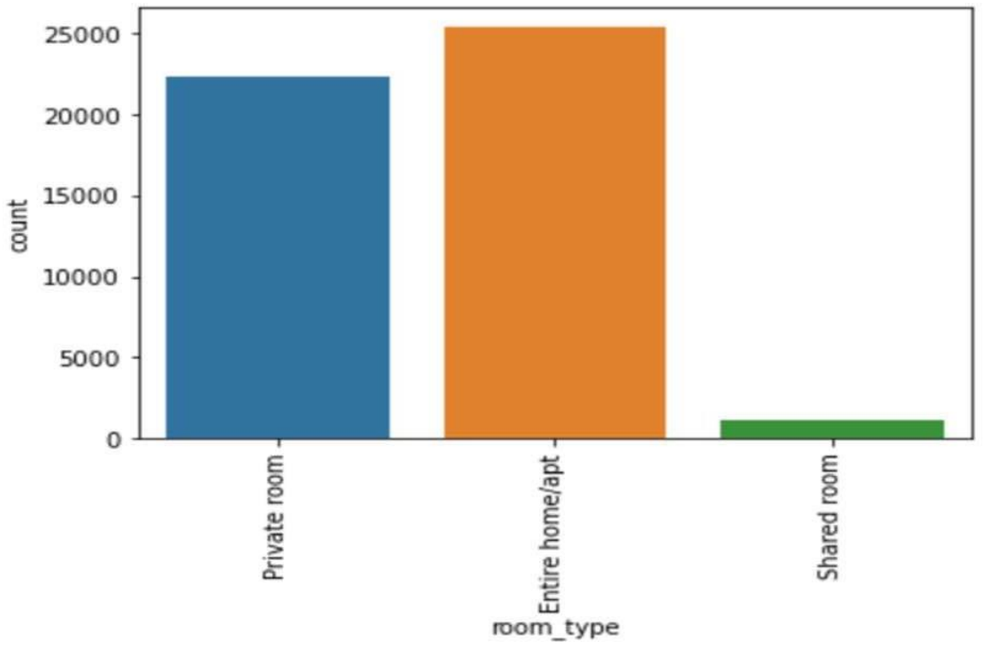
- Checking for data imbalance in dataset

```
In [8]: for i in cat_cols_airbnb:
            print(i)
            sns.countplot(airbnb_data[i])
            plt.xticks(rotation=90)
            plt.show()
```

neighbourhood_group



room_type



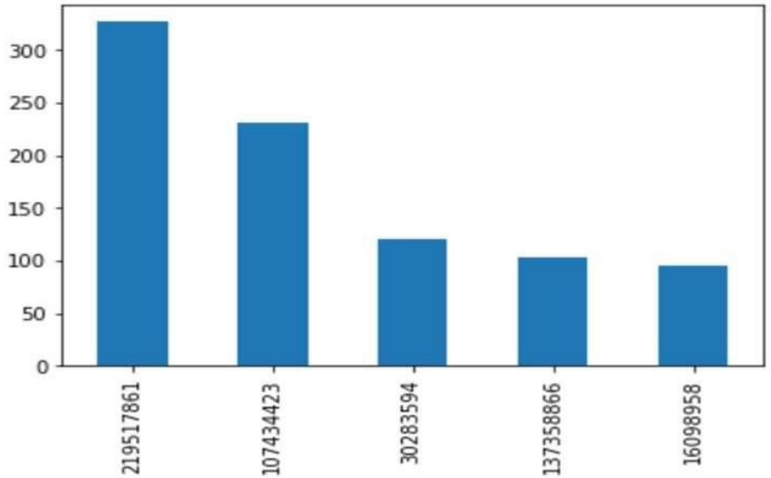
- Analysis of columns present in Airbnb dataset.

Analysis host_id

In [10]:
```
airbnb_data.host_id.value_counts().iloc[:10]
```

Out[10]:
```
219517861    327
107434423    232
30283594     121
137358866    103
16098958      96
12243051      96
61391963      91
22541573      87
200380610     65
7503643       52
Name: host_id, dtype: int64
```

In [50]:
```
airbnb_data.host_id.value_counts().iloc[:5].plot(kind = 'bar')
plt.show()
```



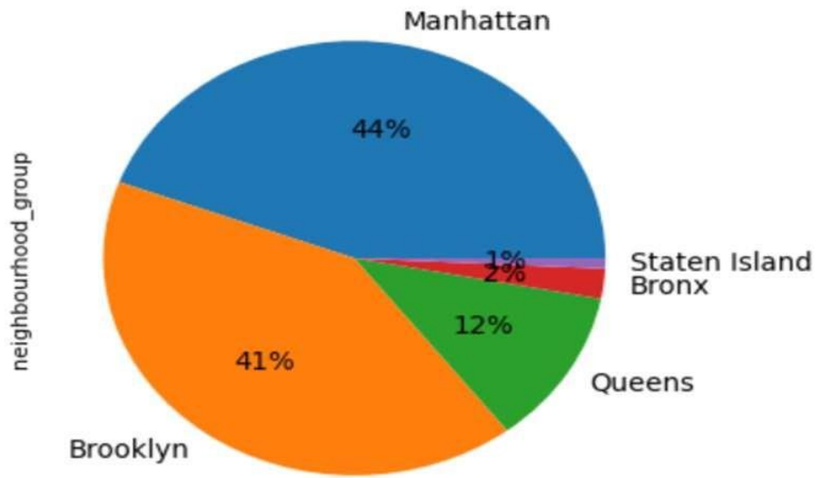
Here we notice that highest number of stays by a host is 327 out of 365 days.

Analysis neighbourhood_group

In [12]:
```
airbnb_data['neighbourhood_group'].value_counts()
```

Out[12]:
```
Manhattan        21661
Brooklyn         20104
Queens            5666
Bronx             1091
Staten Island      373
Name: neighbourhood_group, dtype: int64
```

In [49]:
```
fig = plt.figure(figsize=(5,5), dpi=80)
airbnb_data['neighbourhood_group'].value_counts().plot(kind='pie', autopct='%1.0f%%', startangle=360, fontsize=12)
plt.show()
```



In Manhattan(44%) and Brooklyn(41%) cities most Airbnb transactions happens.

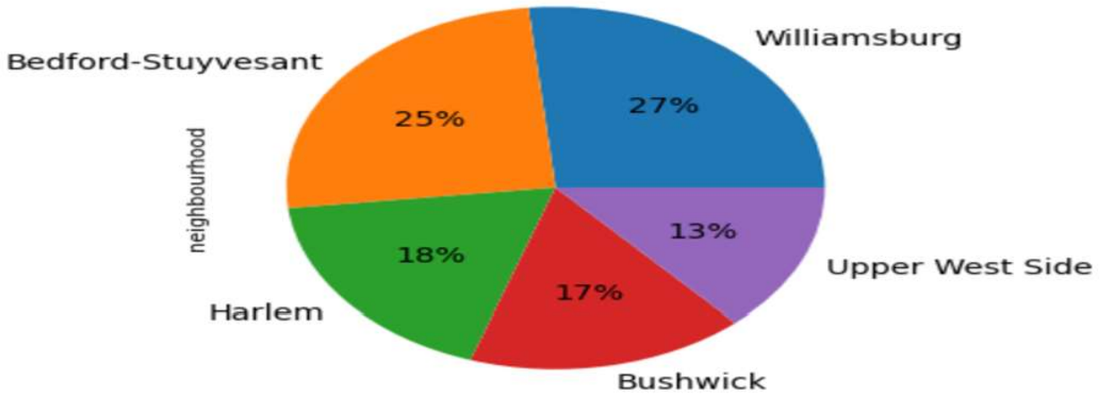
In Staten Island city(1%) least Airbnb transactions happens.

Analysis neighbourhood

In [6]:
```
airbnb_data['neighbourhood'].value_counts().iloc[:10]
```

Out[6]:
```
Williamsburg            3920
Bedford-Stuyvesant      3714
Harlem                  2658
Bushwick                2465
Upper West Side         1971
Hell's Kitchen          1958
East Village            1853
Upper East Side         1798
Crown Heights           1564
Midtown                 1545
Name: neighbourhood, dtype: int64
```

In [10]:
```
fig = plt.figure(figsize=(5,5), dpi=80)
airbnb_data['neighbourhood'].value_counts().iloc[:5].plot(kind='pie',  autopct='%1.0f%%', startangle=360, fontsize=13)
plt.show()
```



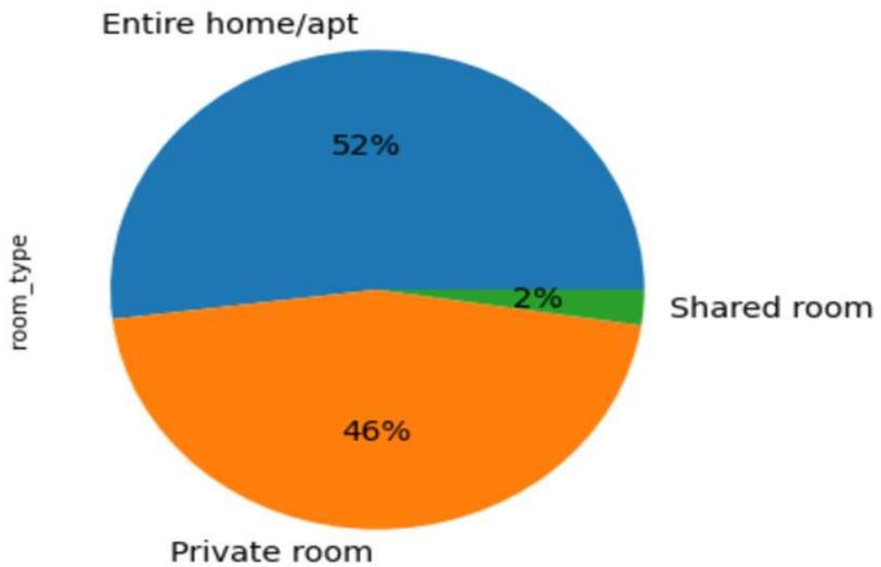
We can see that Williamsburg is the area where high number of transaction happens.

Analysis of room_type

In [11]:
```
airbnb_data['room_type'].value_counts()
```

Out[11]:
```
Entire home/apt    25409
Private room       22326
Shared room         1160
Name: room_type, dtype: int64
```

In [12]:
```
fig = plt.figure(figsize=(5,5), dpi=80)
airbnb_data['room_type'].value_counts().plot(kind='pie',  autopct='%1.0f%%', startangle=360, fontsize=13)
plt.show()
```



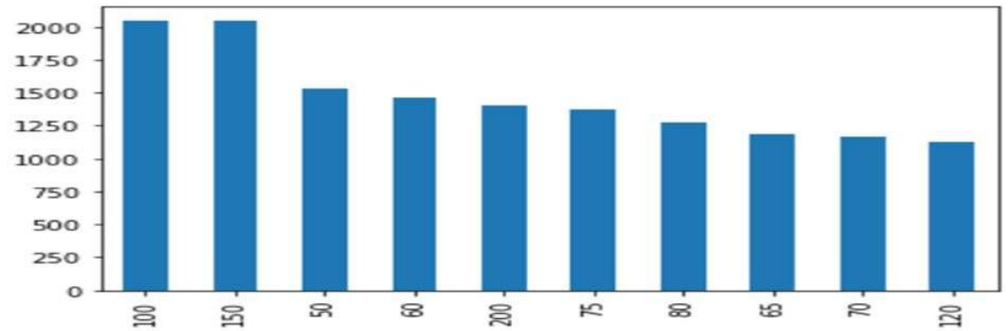
Around 25k people (52%) choose to use a house/apt while 22k(46%) for a private room. Only 1k(2%) people choose a shared room. This could mean more people who use airbnb, use it with family maybe for tours,visits,etc.

Analysis of price

In [16]:
```
airbnb_data.price.value_counts().iloc[:10]
```

Out[16]:
```
100     2051
150     2047
50      1534
60      1458
200     1401
75      1370
80      1272
65      1190
70      1170
120     1130
Name: price, dtype: int64
```

In [15]:
```
airbnb_data.price.value_counts().iloc[:10].plot(kind = 'bar')
plt.show()
```



In [14]:
```
airbnb_data.price.describe()
```

Out[14]:
```
count     48895.000000
mean        152.720687
std         240.154170
min           0.000000
25%          69.000000
50%         106.000000
75%         175.000000
max       10000.000000
Name: price, dtype: float64
```

The average pricing is around 152 dollars.

50% of data has price greater than 106 dollars.

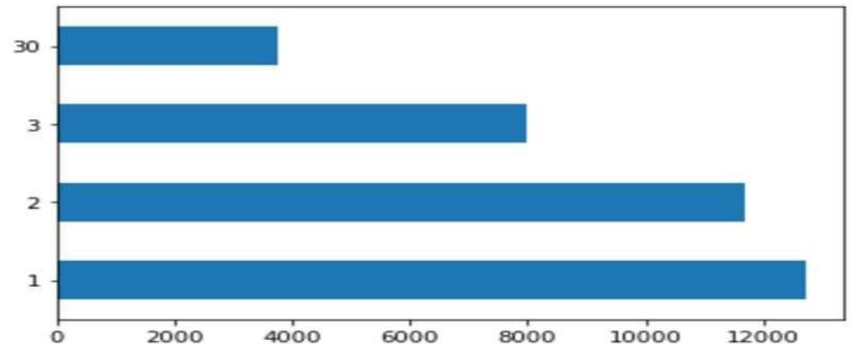
The costliest airbnb has around 10k dollars as price.

Analysis of minimum_nights

In [17]:
```
airbnb_data['minimum_nights'].value_counts()
```

Out[17]:
```
1        12720
2        11696
3         7999
30        3760
4         3303
5         3034
7         2058
6          752
14         562
10         483
29         340
15         279
20         223
28         203
31         201
21         135
8          130
60         106
90         104
```

In [18]:
```
airbnb_data['minimum_nights'].value_counts().iloc[:4].plot(kind = 'barh')
plt.show()
```



We can observe that most of almost 12k people used 1 night stay in airbnb.

11k people choose 2 night stay while 7k choose 3 night stay.

Almost 3.7k stayed upto a month.

Analysis of availability_365

```
In [19]: airbnb_data['availability_365'].value_counts()
```

```
Out[19]: 0       17533
         365      1295
         364       491
         1         408
         89        361
         5         340
         3         306
         179       301
         90        290
         2         270
         6         245
         363       239
         8         233
         4         233
         342       230
         188       225
         7         219
         88        200
         311       199
```

```
In [20]: airbnb_data[airbnb_data['availability_365'] == 365].describe()
```

Out[20]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_cc
count	1.295000e+03	1.295000e+03	1295.000000	1295.000000	1295.000000	1295.00000	1295.000000	1295.000000	1295.000
mean	1.940195e+07	8.554698e+07	40.729014	-73.943275	250.769884	19.60000	10.220849	0.793089	13.158
std	1.197265e+07	8.786960e+07	0.057781	0.059799	550.497373	65.05093	22.095983	0.897942	36.224
min	2.539000e+03	2.787000e+03	40.507080	-74.242850	20.000000	1.00000	0.000000	0.010000	1.000
25%	8.725256e+06	8.931349e+06	40.687990	-73.983210	72.000000	1.00000	0.000000	0.240000	1.000
50%	2.065068e+07	4.634351e+07	40.730990	-73.954270	125.000000	3.00000	2.000000	0.720000	2.000
75%	3.027040e+07	1.565055e+08	40.762095	-73.921715	225.000000	30.00000	10.000000	0.720000	7.000
max	3.648315e+07	2.733930e+08	40.893740	-73.721730	9999.000000	1250.00000	183.000000	8.940000	327.000

Costliest airbnb with 365 days availablity costs around 10k dollars with average of 250 dollars.

Analysis of reviews_per_month

In [22]:
```
airbnb_data['reviews_per_month'].max()
```

Out[22]:
```
58.5
```

In [23]:
```
airbnb_data[airbnb_data['reviews_per_month'] == 58.5]
```

Out[23]:

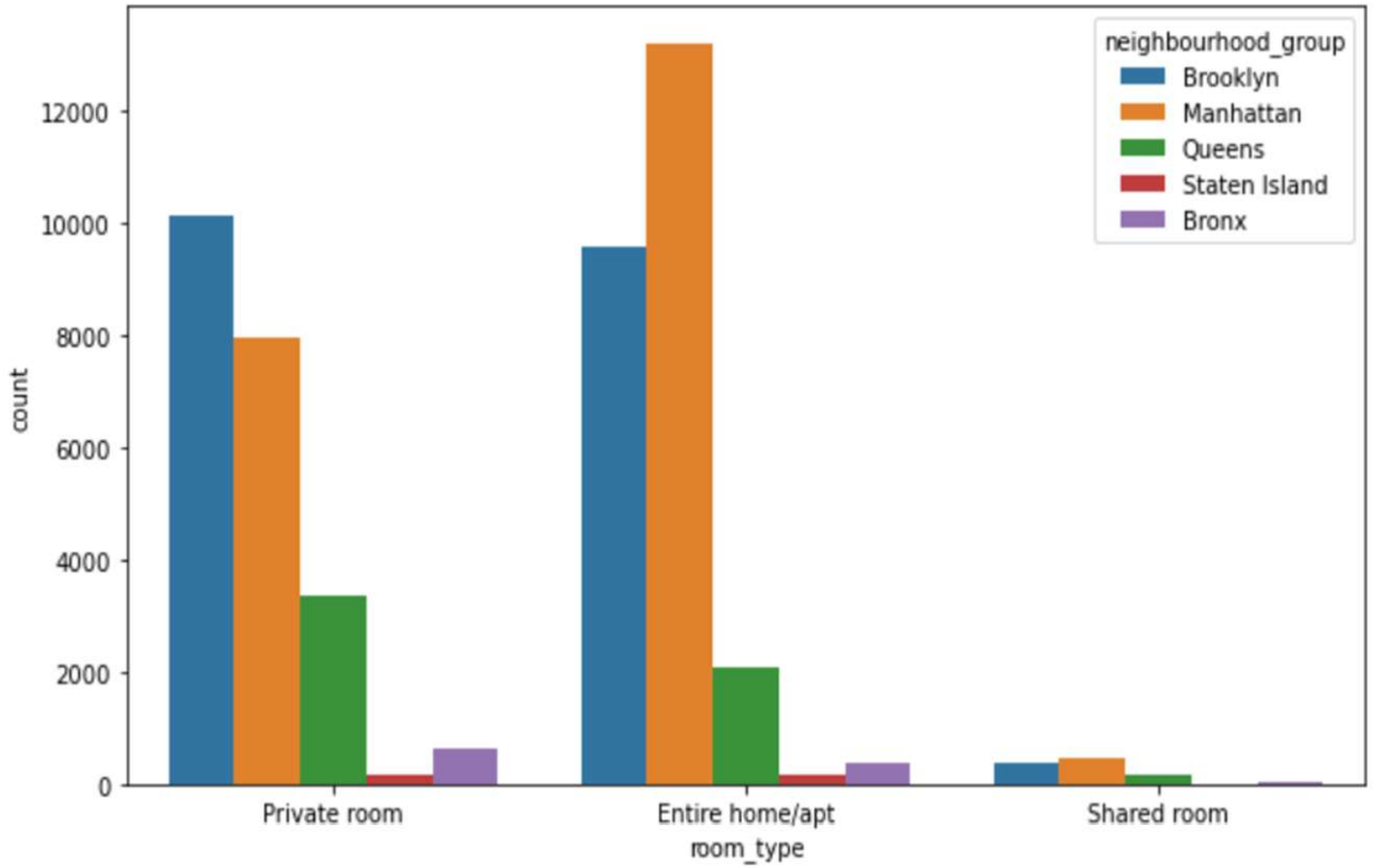
	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_rev
42075	32678719	Enjoy great views of the City in our Deluxe Room!	244361589	Row NYC	Manhattan	Theater District	40.75918	-73.98801	Private room	100	1	

Enjoy great views in Manhattan has the highest reviews per month. They offer Private room and is worth 100 dollars a night.

- Bivariate Analysis on python:

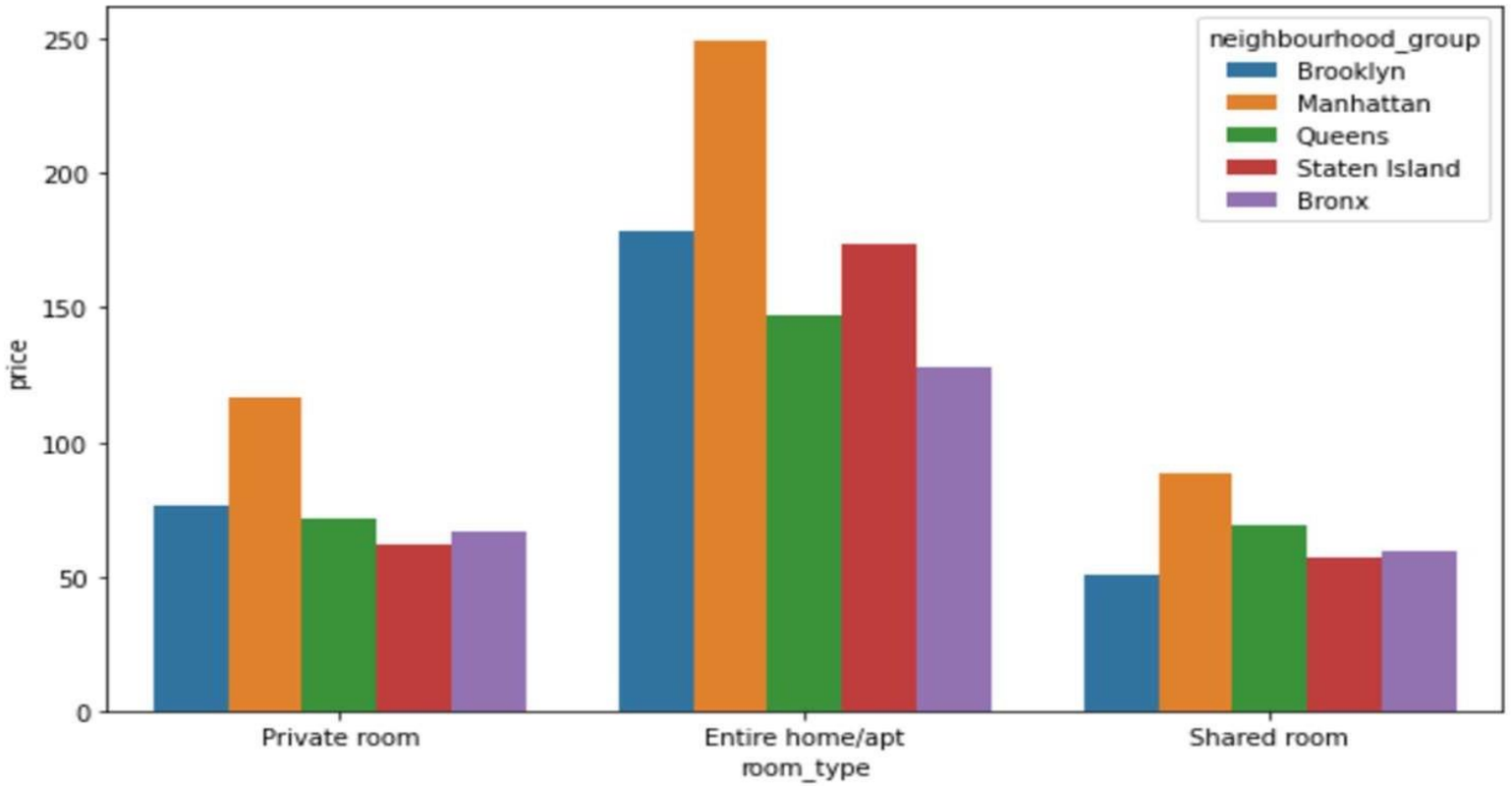
```
In [25]: corr = airbnb_data.corr()
         plt.figure(figsize=(15,8))
         sns.heatmap(corr, annot=True)
         plt.show()
```

In [8]:
```
plt.figure(figsize=(10,6))
sns.countplot(data = airbnb_data, x = 'room_type', hue = 'neighbourhood_group')
plt.show()
```



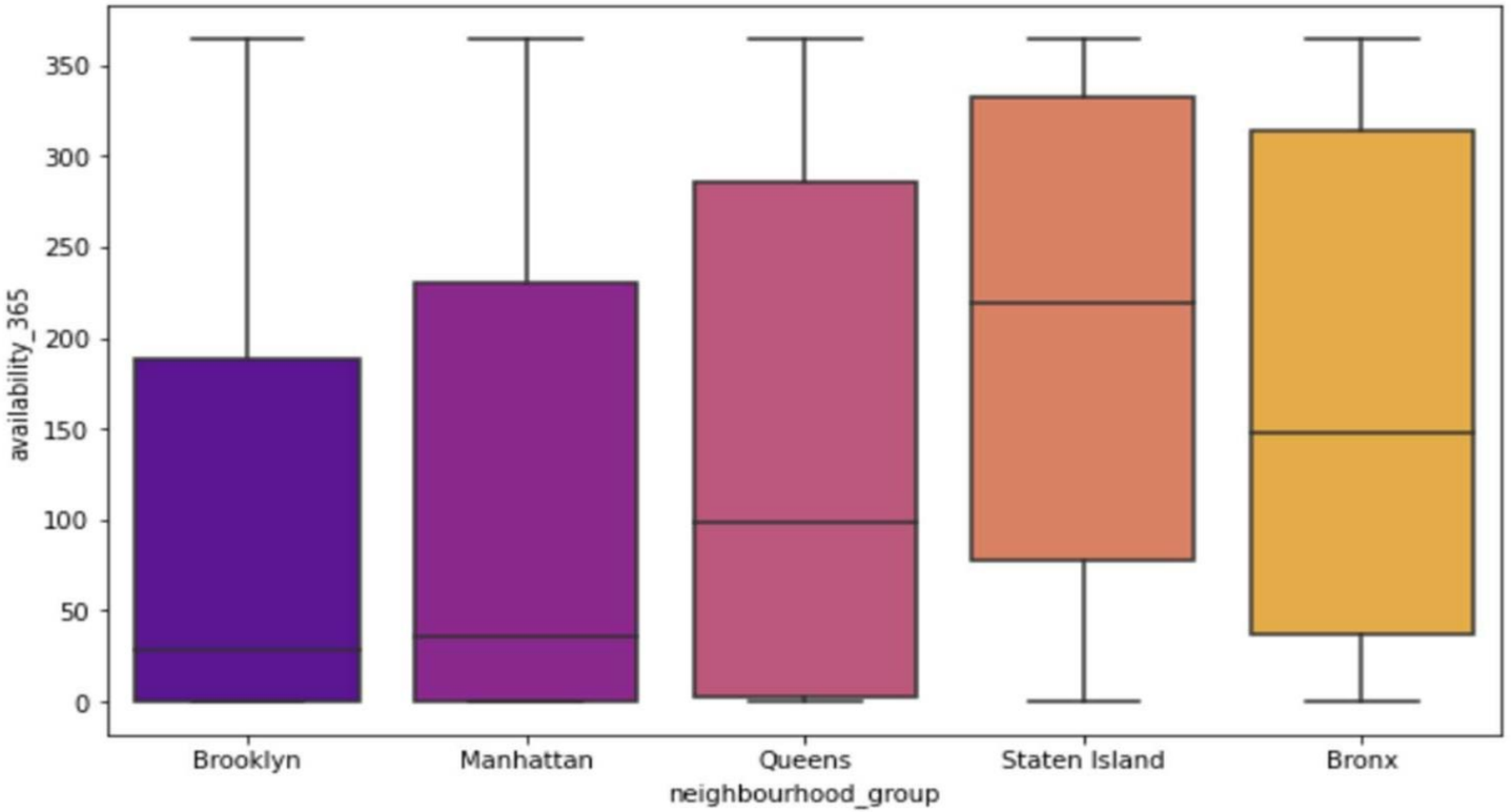
Home service seems to be most used by people and the highest in Manhattan. This is also the highest service used across New York City.

In [11]:
```
plt.figure(figsize=(10,6))
sns.barplot(x = 'room_type',
            y = 'price',
            hue = 'neighbourhood_group',
            data = airbnb_data, ci=0)
plt.show()
```



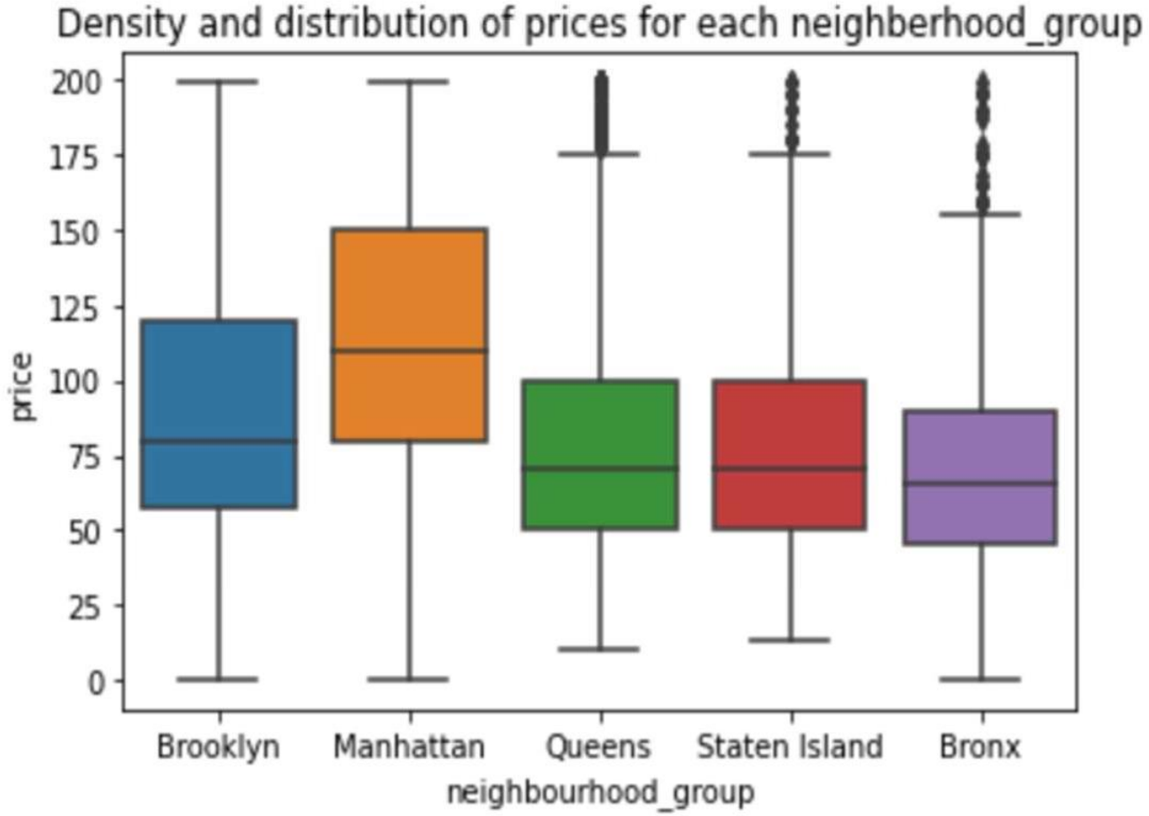
Manhattan has the most expensive rental properties, while Bronx has the least expensive.

In [29]:
```
plt.figure(figsize=(10,6))
ax = sns.boxplot(data=airbnb_data, x='neighbourhood_group',y='availability_365',palette='plasma')
plt.show()
```



Staten Island has th highest average airbnb availablity.

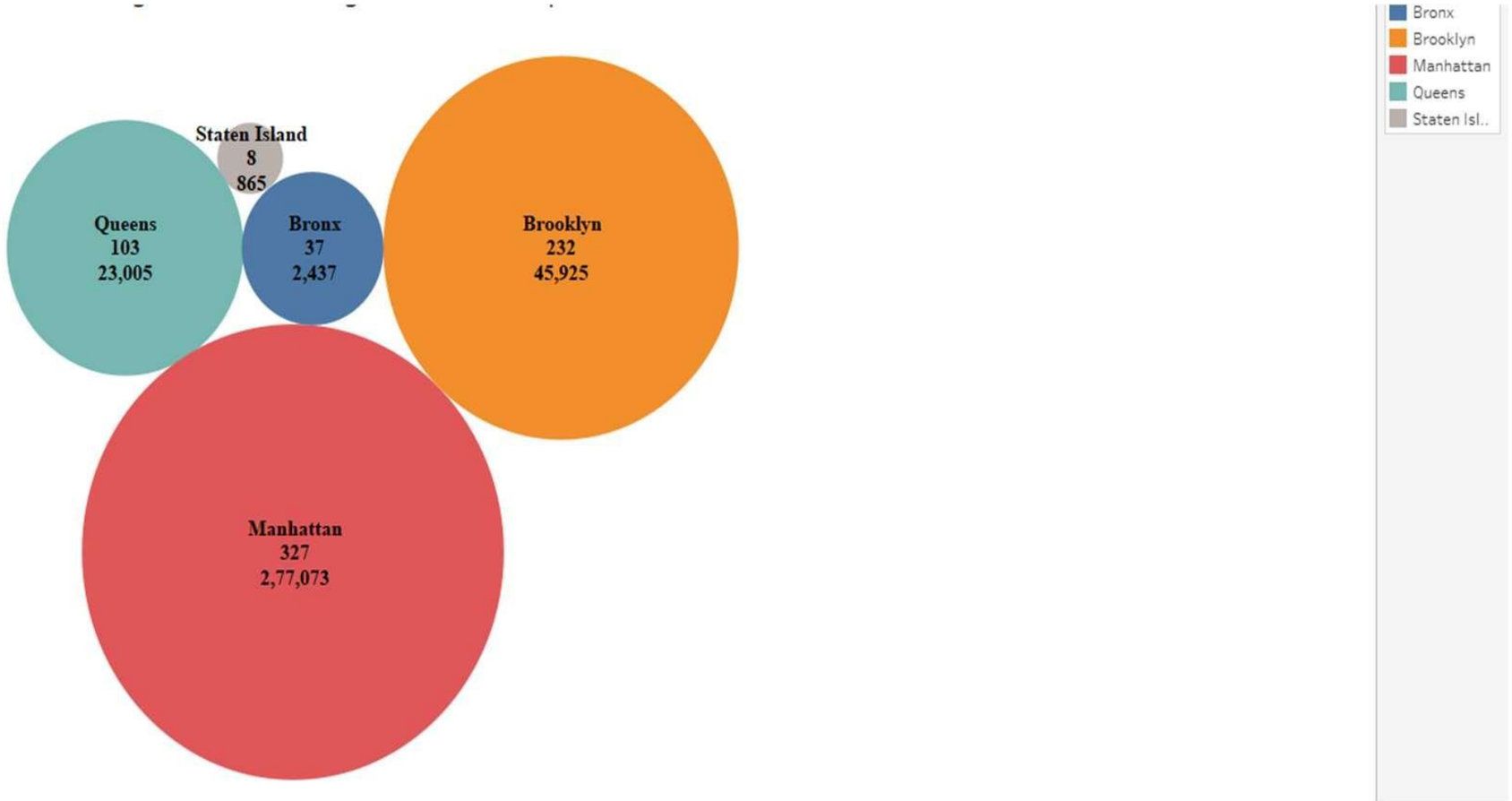
In [30]:
```
v2=sns.boxplot(data=airbnb_data[airbnb_data.price < 200], x='neighbourhood_group', y='price')
v2.set_title('Density and distribution of prices for each neighberhood_group')
plt.show()
```



Manhattan airbnb's has the highest average price.

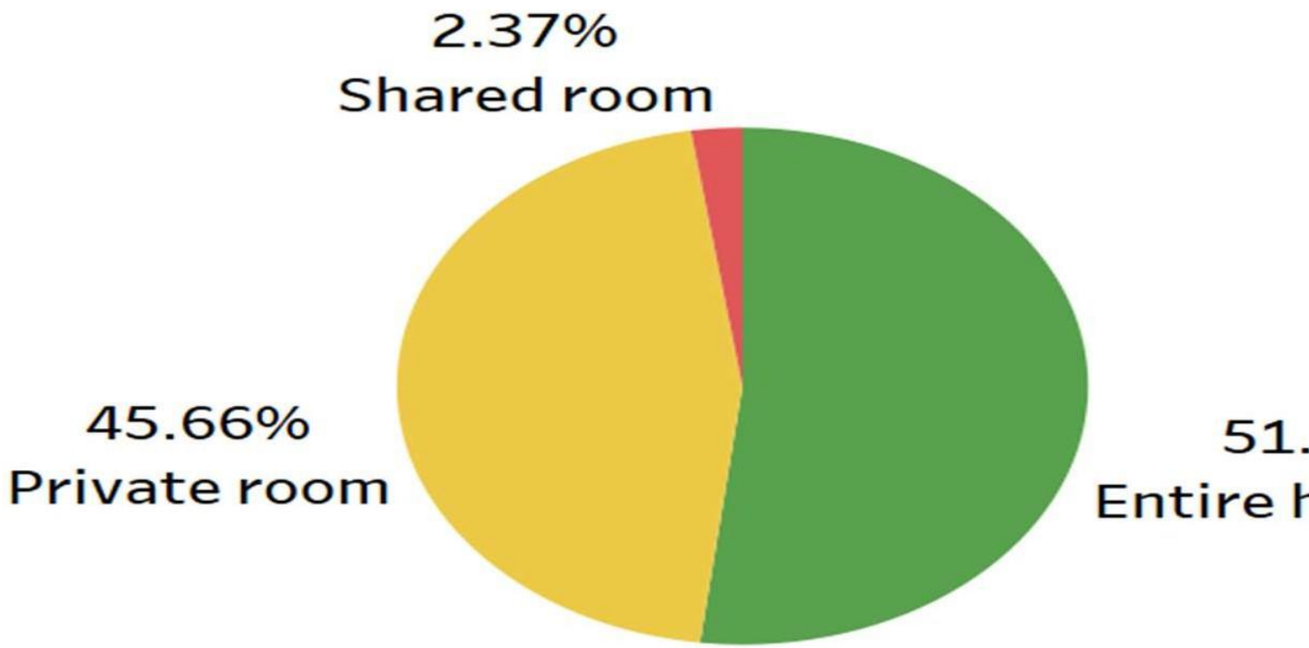
Visualization on Tableau

- Neighborhoods with Most Listings



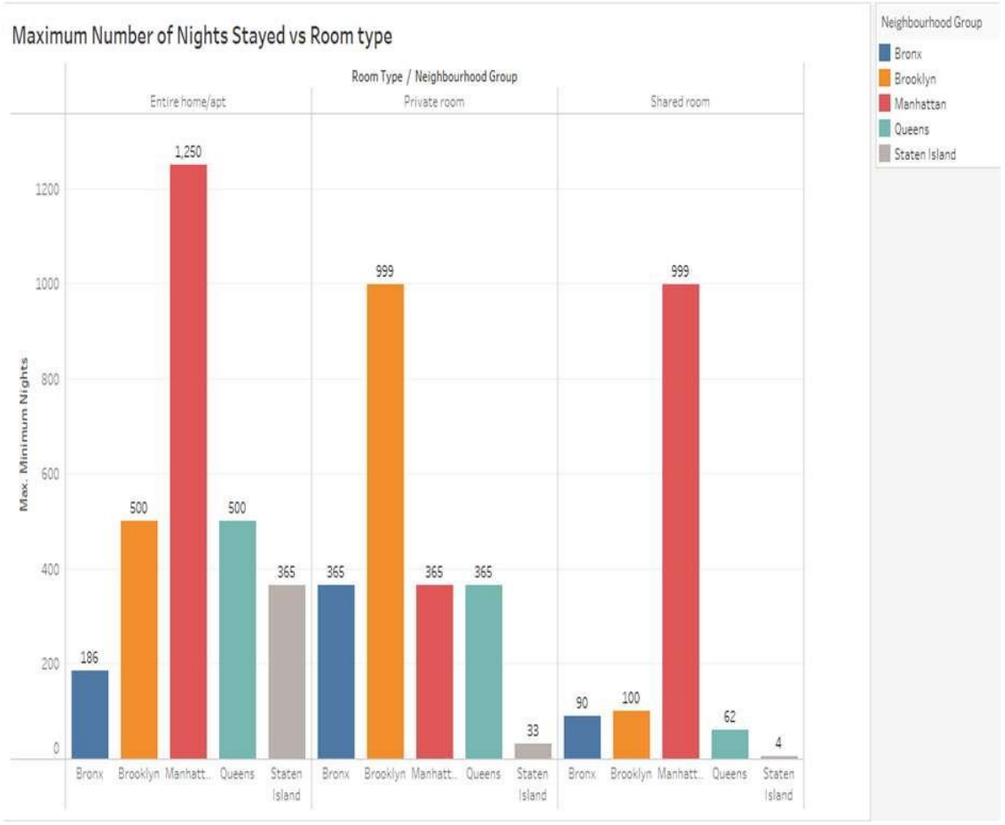
We observe most listings are in Manhattan and Brooklyn

- Room Preference

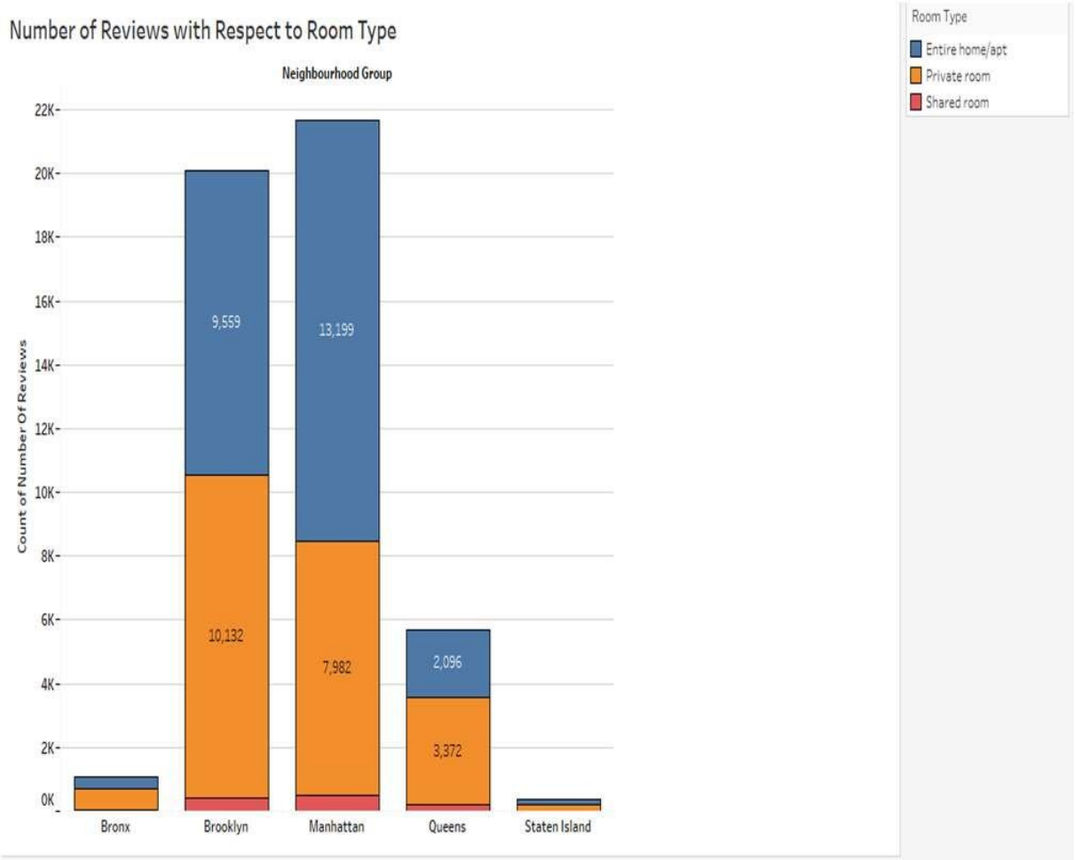


People prefer entire homes and apartments followed by private roomsand shared room are the least preferred

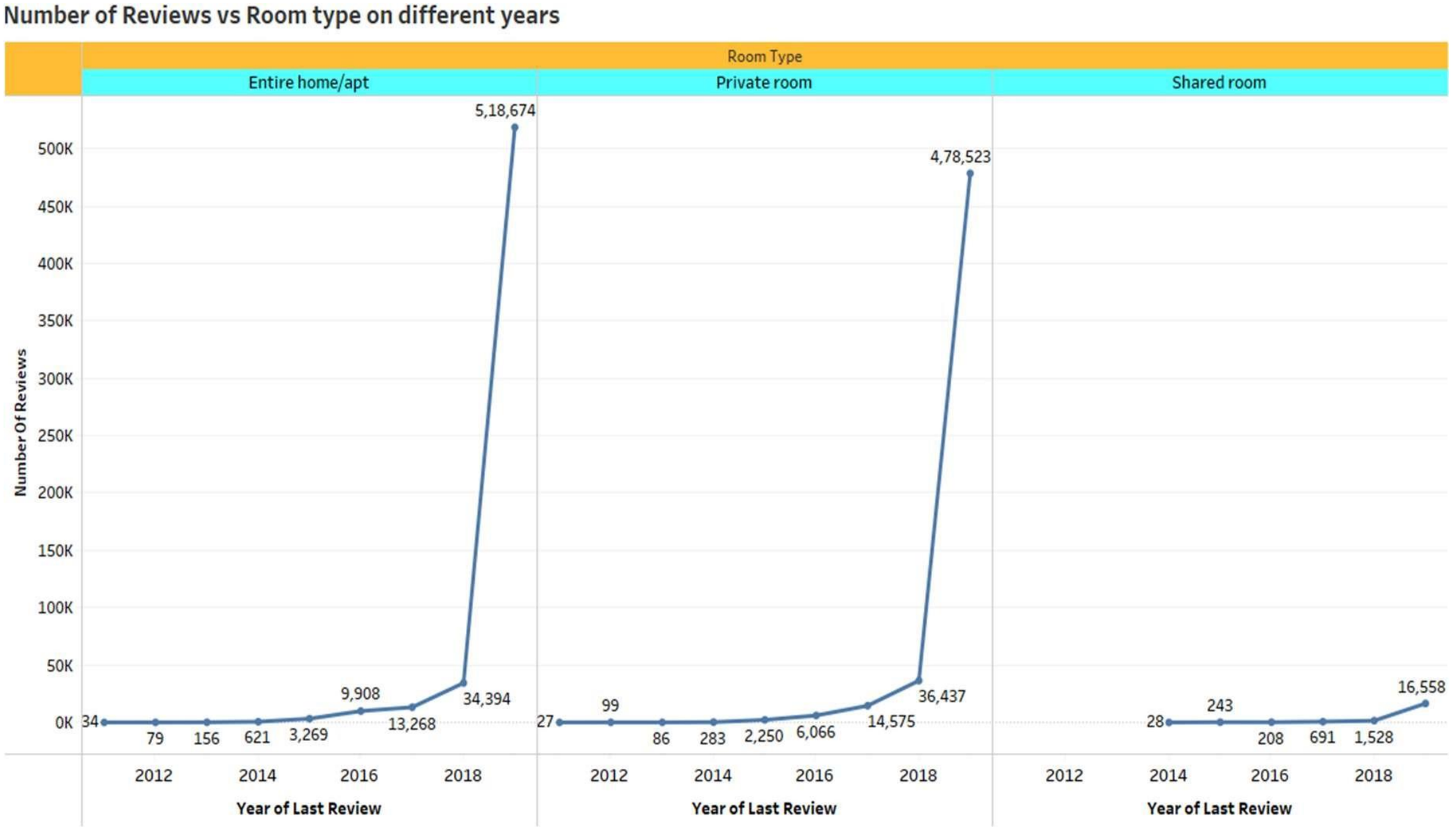
- Maximum number of nights stayed in different room types



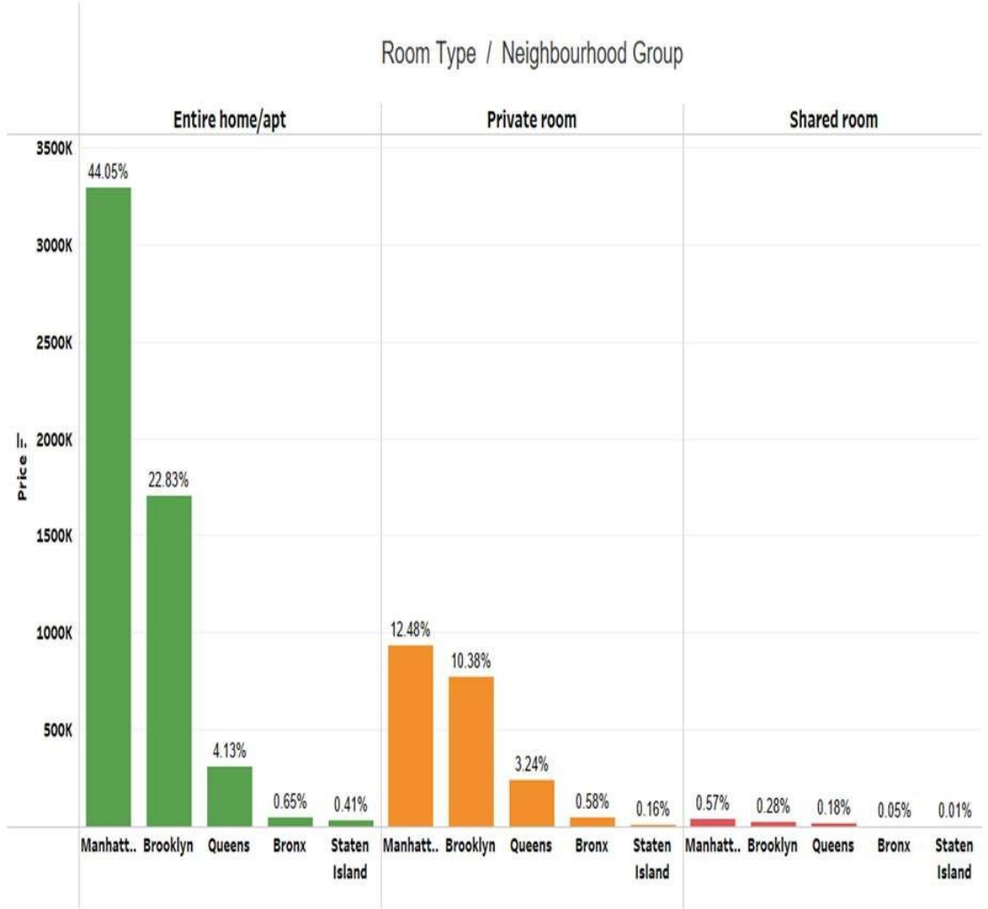
- Number of reviews with room type



- Number of Reviews vs Room type on different years (2012-2019)

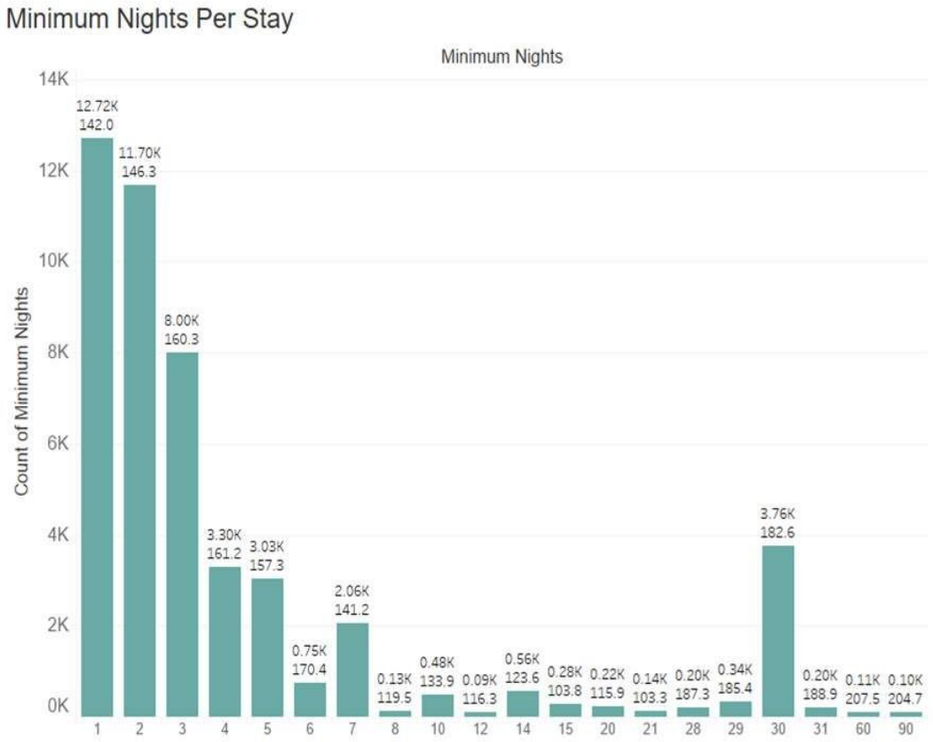


- Prices of Properties and Room Type in Neighbourhood Group

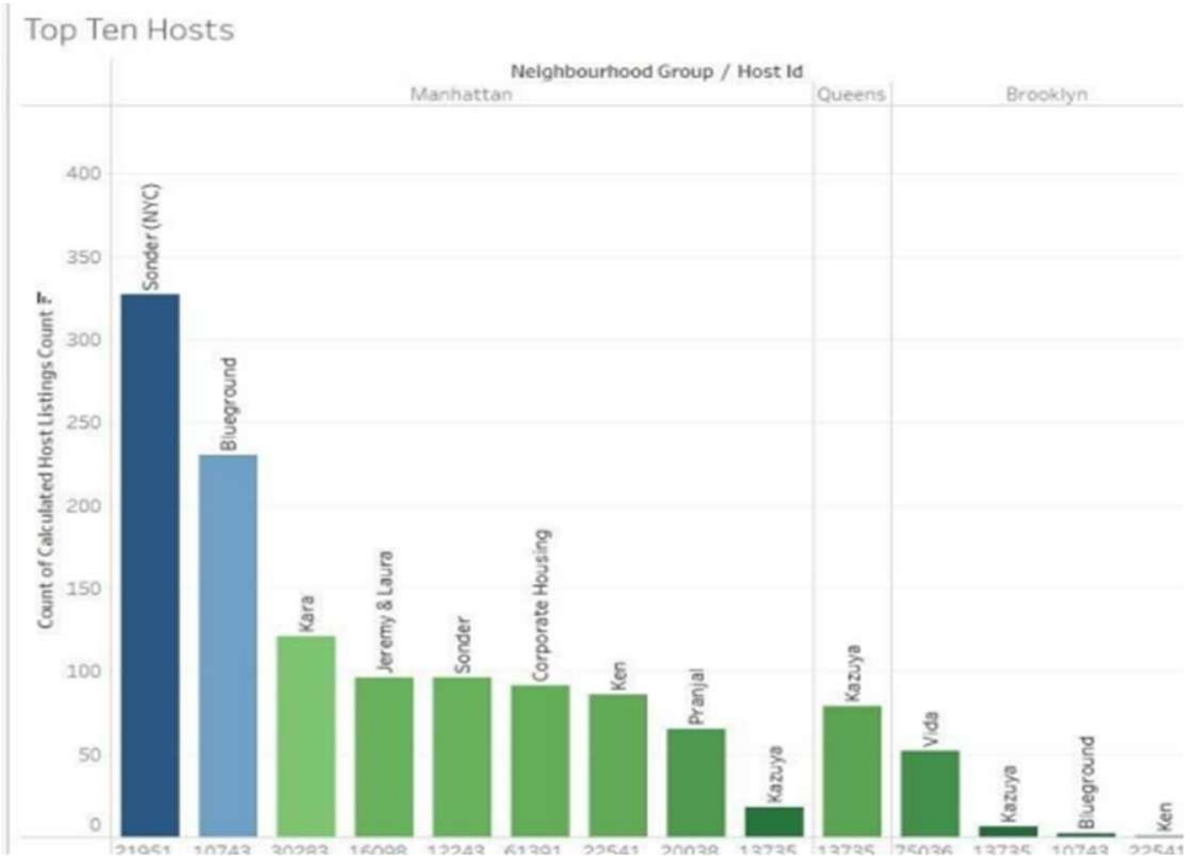


- Average Prices of Properties and Room Type in Neighbourhood Group

- Customer Preferences of Minimum Nights Per Stay



- Top Ten Hosts



Most host with highest listings is in Manhattan

6. Key Insights and Recommendations:

- Renters in New York City who use Airbnb are privileged to entirehouses or apartments, plus private rooms above shared rooms.
- Manhattan has the most expensive rental properties followed by Brooklyn, while the Bronx and Staten Island have the least expensive.
- People show interest in the host Sonder and spend most nights here.
- Pay attention to popular areas like Manhattan and Brooklyn where more people are interested.
- Since there is a lower likelihood that people will choose a highpriced room, there are more evaluations at lower prices than at higher prices.
- People show interest in the host Sounder and spend more nights there also Michael is most reviewed host among all.
- Majority of the people like to spend one day followed by two days.

Thank you