Project Name -

#AIRBNB BOOKING ANALYSES Project by :- Shaurya Pradhan

Project Summary -

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific area.Airbnb offers people an easy, relatively stress-free way to earn some income from their property.

Since 2008, guests and hosts have used Airbnb to travel in a more unique, personalized way. This dataset describes all the listing activity of homestays in New York City.NYC is not only the most famous city in the world but also top global destination for visitors drawn to its museums, entertainment, restaurants and commerce.

Dataset has 48895 observations across 16 attributes, ranges from host information, geographical information, booking logistics, house/room information to reviews and availability.

Our motivation to explore this dataset is because we want to provide guidance for travelers to New York City before they make any decisions in terms of Airbnbs. Specifically, we want the travelers to have a general understanding of where would be the cheapest and best reviewed homestays.

Different variables: price, name, host id, host name, Neighborhood-group, neighborhood, latitude, longitude, room-type, minimum-nights, number of reviews, last review, review per month, calculated host listings, and availability 365 days.

Problem Statement

Write Problem Statement Here.

- One of the biggest challenges for companies is to maintain positive customer experience along with having a financially profitable business model for property owners. How factors are affecting the price for the Airbnb listing in NYC? What is the overall location distribution of Airbnb NYC? Which neighborhood has a better average price for the Airbnb listing?

- We will be doing analysis on every Airbnb listing based on their location, including their price range, room type, listing name, and other related factors.

- Our objective would be to find out the key metrics that influence the listing of properties on the platform.

Define Your Business Objective?

Trying to answering following question for airbnb :

Q1. In New York where are the highest share of airbnb hotels ?

Q2. Which room type has the highest and the lowest number of booking ?

Q3. According to the neighborhood group which room type has the most number of booking ?

Q4. What is the average price per night for different room type based on neighborhood group ?

Q5. What are the Top ten neighborhoods with the most expensive prices ?

Q6. What are the Top ten neighborhoods with the cheapest prices ?

Q7. What are the top ten neighborhoods with most number of booking ?

Q8. What are the top ten neighborhoods with the least number of booking ?

Q9. Which neighborhood group have most number of reviews ?

Q10. Which neighborhood group has the most booking show using scatter plot ?

Q11. Show a relation between neighborhood group and price using box plot ?

Q13. Who are the top earning host ?

Q14. Which room type has been occupied for the most number of nights ?

Let's Begin !

1. Know Your Data

```
# Importing Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import missingno as msno
```

```
 # Loading Dataset
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly
remount, call drive.mount("/content/drive", force_remount=True).
```

```
airbnb = pd.read_csv('/content/drive/MyDrive/Airbnb NYC 2019.csv')
```

```
# Dataset First Look
airbnb
```

```
             id                                              name
host_id  \
0          2539                Clean & quiet apt home by the park
2787
1          2595                             Skylit Midtown Castle
2845
2          3647               THE VILLAGE OF HARLEM....NEW YORK !
4632
3          3831                   Cozy Entire Floor of Brownstone
4869
4          5022  Entire Apt: Spacious Studio/Loft by central park
7192
...         ...                                               ...
...
48890  36484665   Charming one bedroom - newly renovated rowhouse
8232441
48891  36485057     Affordable room in Bushwick/East Williamsburg
6570630
48892  36485431           Sunny Studio at Historical Neighborhood
23492952
48893  36485609              43rd St. Time Square-cozy single bed
30985759
48894  36487245  Trendy duplex in the very heart of Hell's Kitchen
68119814

           host_name neighbourhood_group       neighbourhood  latitude
\
0               John            Brooklyn          Kensington  40.64749

1           Jennifer           Manhattan             Midtown  40.75362

2          Elisabeth           Manhattan              Harlem  40.80902

3        LisaRoxanne            Brooklyn        Clinton Hill  40.68514

4              Laura           Manhattan         East Harlem  40.79851

...              ...                 ...                 ...       ...

48890        Sabrina            Brooklyn  Bedford-Stuyvesant  40.67853

48891        Marisol            Brooklyn            Bushwick  40.70184

48892  Ilgar & Aysel           Manhattan              Harlem  40.81475

48893            Taz           Manhattan      Hell's Kitchen  40.75751
```

```
48894     Christophe           Manhattan      Hell's Kitchen  40.76404


       longitude        room_type  price  minimum_nights
number_of_reviews  \
0      -73.97237     Private room    149               1
9
1      -73.98377  Entire home/apt    225               1
45
2      -73.94190     Private room    150               3
0
3      -73.95976  Entire home/apt     89               1
270
4      -73.94399  Entire home/apt     80              10
9
...          ...              ...    ...             ...
...
48890  -73.94995     Private room     70               2
0
48891  -73.93317     Private room     40               4
0
48892  -73.94867  Entire home/apt    115              10
0
48893  -73.99112      Shared room     55               1
0
48894  -73.98933     Private room     90               7
0

      last_review  reviews_per_month
calculated_host_listings_count  \
0      2018-10-19               0.21                               6

1      2019-05-21               0.38                               2

2             NaN                NaN                               1

3      2019-07-05               4.64                               1

4      2018-11-19               0.10                               1

...           ...                ...                             ...

48890         NaN                NaN                               2

48891         NaN                NaN                               2

48892         NaN                NaN                               1
```

```
48893         NaN                NaN                              6

48894         NaN                NaN                              1


       availability_365
0                   365
1                   355
2                   365
3                   194
4                     0
...                 ...
48890                 9
48891                36
48892                27
48893                 2
48894                23

[48895 rows x 16 columns]
```

```
# Dataset Rows & Columns count
airbnb.shape
```

```
(48895, 16)
```

Our dataframe have 48895 rows and 16 columns.

```
# Dataset Info
airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
```

```
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

Removing the Duplicates

```
# Dataset Duplicate Value Count
airbnb.duplicated().sum()
airbnb.drop_duplicates(inplace = True)
```

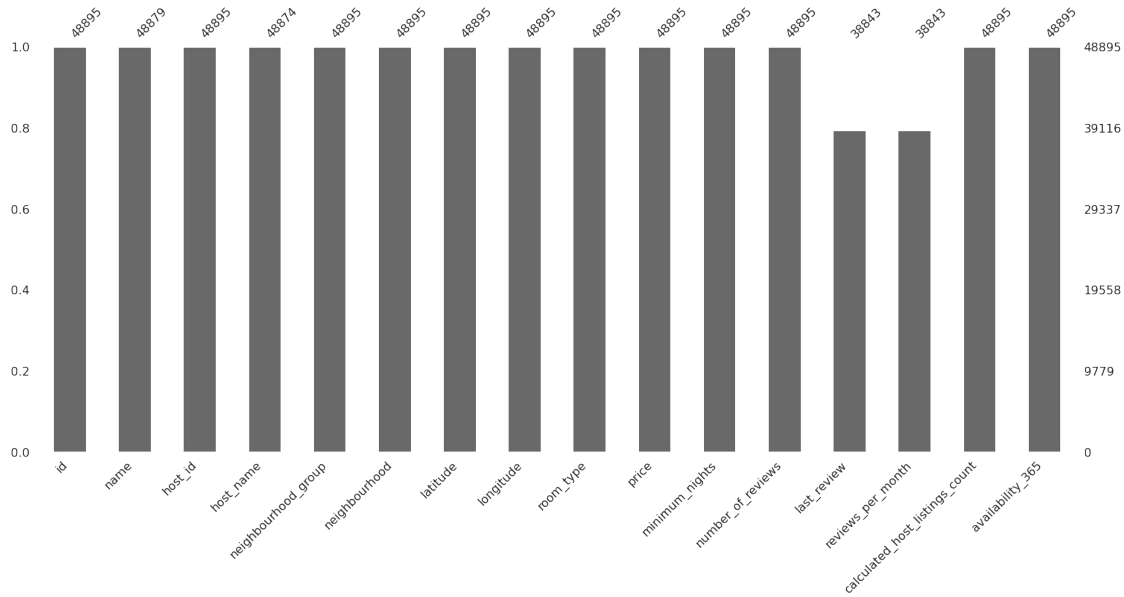
####Checking Missing Values

```
# Missing Values
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
```

```
# Visualizing the missing values
plt.rcParams['figure.figsize'] = (10,5)
msno.bar(airbnb)
```

```
<AxesSubplot:>
```



What did you know about your dataset?

As of now we know our dataset have 48895 rows and 16 columns and out of them last_review and reviews_per_month columns are the one which have largest number of missing values. Hence we should remove them

2. Understanding Your Variables

```
# Dataset Columns
airbnb.columns

Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')

# Dropping columns that are unnecessary for our analysis
airbnb.drop(['id','name','last_review'], axis = 'columns', inplace =
True)

# Replacing all NaN value in reviews_per_months with 0.
airbnb.reviews_per_month.fillna(0,inplace = True)
airbnb.host_name.fillna(0,inplace = True)

# Write your code to make your dataset analysis ready.
airbnb.isnull().any()

host_id                           False
host_name                         False
neighbourhood_group               False
neighbourhood                     False
latitude                          False
```

```
longitude                          False
room_type                          False
price                              False
minimum_nights                     False
number_of_reviews                  False
reviews_per_month                  False
calculated_host_listings_count     False
availability_365                   False
dtype: bool
```

```
# Descriptive statistics for numerical values
airbnb.describe()
```

```
            host_id      latitude     longitude         price
minimum_nights  \
count  4.889500e+04  48895.000000  48895.000000  48895.000000
48895.000000
mean   6.762001e+07     40.728949    -73.952170    152.720687
7.029962
std    7.861097e+07      0.054530      0.046157    240.154170
20.510550
min    2.438000e+03     40.499790    -74.244420      0.000000
1.000000
25%    7.822033e+06     40.690100    -73.983070     69.000000
1.000000
50%    3.079382e+07     40.723070    -73.955680    106.000000
3.000000
75%    1.074344e+08     40.763115    -73.936275    175.000000
5.000000
max    2.743213e+08     40.913060    -73.712990  10000.000000
1250.000000

       number_of_reviews  reviews_per_month
calculated_host_listings_count  \
count       48895.000000       48895.000000
48895.000000
mean           23.274466           1.090910
7.143982
std            44.550582           1.597283
32.952519
min             0.000000           0.000000
1.000000
25%             1.000000           0.040000
1.000000
50%             5.000000           0.370000
1.000000
75%            24.000000           1.580000
2.000000
max           629.000000          58.500000
327.000000
```

```
       availability_365
count      48895.000000
mean         112.781327
std          131.622289
min            0.000000
25%            0.000000
50%           45.000000
75%          227.000000
max          365.000000
```

What all manipulations have you done and insights you found?

First we have checked for duplicate values and then droped them secondly we have cheched for null values in our data set and found there are two variables with highest number of null values to solve this we have droped last_review and filled review_per_months with zero. Thirdly we have droped all the variables which is not useful for our analysis.

```
airbnb.head(5)
```

```
   host_id neighbourhood_group neighbourhood  latitude  longitude  \
0     2787            Brooklyn    Kensington  40.64749  -73.97237
1     2845           Manhattan       Midtown  40.75362  -73.98377
2     4632           Manhattan        Harlem  40.80902  -73.94190
3     4869            Brooklyn  Clinton Hill  40.68514  -73.95976
4     7192           Manhattan   East Harlem  40.79851  -73.94399

         room_type  price  minimum_nights  number_of_reviews  \
0     Private room    149               1                  9
1  Entire home/apt    225               1                 45
2     Private room    150               3                  0
3  Entire home/apt     89               1                270
4  Entire home/apt     80              10                  9

   reviews_per_month  calculated_host_listings_count  availability_365

0               0.21                               6               365

1               0.38                               2               355

2               0.00                               1               365

3               4.64                               1               194

4               0.10                               1                 0
```

4. Data Vizualization, Storytelling & Experimenting with charts : Understand the relationships between variables

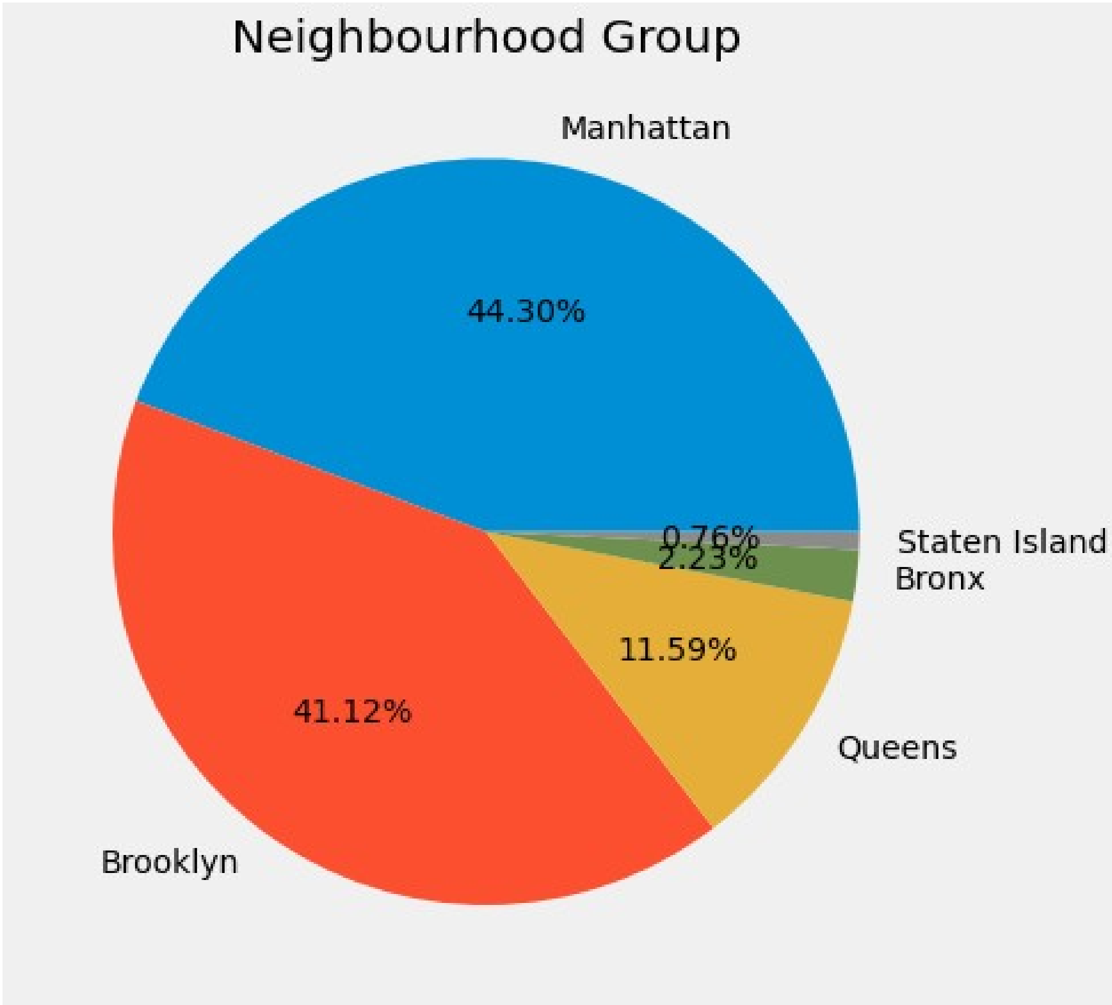
```
airbnb.head(5)
```

```
   host_id    host_name neighbourhood_group neighbourhood  latitude  \
0     2787         John            Brooklyn    Kensington  40.64749
1     2845     Jennifer           Manhattan       Midtown  40.75362
2     4632    Elisabeth           Manhattan        Harlem  40.80902
3     4869  LisaRoxanne            Brooklyn  Clinton Hill  40.68514
4     7192        Laura           Manhattan   East Harlem  40.79851

   longitude        room_type  price  minimum_nights
number_of_reviews  \
0  -73.97237     Private room    149               1
9
1  -73.98377  Entire home/apt    225               1
45
2  -73.94190     Private room    150               3
0
3  -73.95976  Entire home/apt     89               1
270
4  -73.94399  Entire home/apt     80              10
9

   reviews_per_month  calculated_host_listings_count  availability_365

0               0.21                               6               365

1               0.38                               2               355

2               0.00                               1               365

3               4.64                               1               194

4               0.10                               1                 0
```

Neighborhood Group

Q1. In New York where are the highest share of airbnb hotels ?

```
# Chart - 1 visualization
plt.style.use('fivethirtyeight')
plt.figure(figsize=(10,7))
plt.title('Neighbourhood Group')
plt.pie(airbnb.neighbourhood_group.value_counts(), labels
=airbnb.neighbourhood_group.value_counts().index,autopct='%.2f%%' )
plt.show()
```



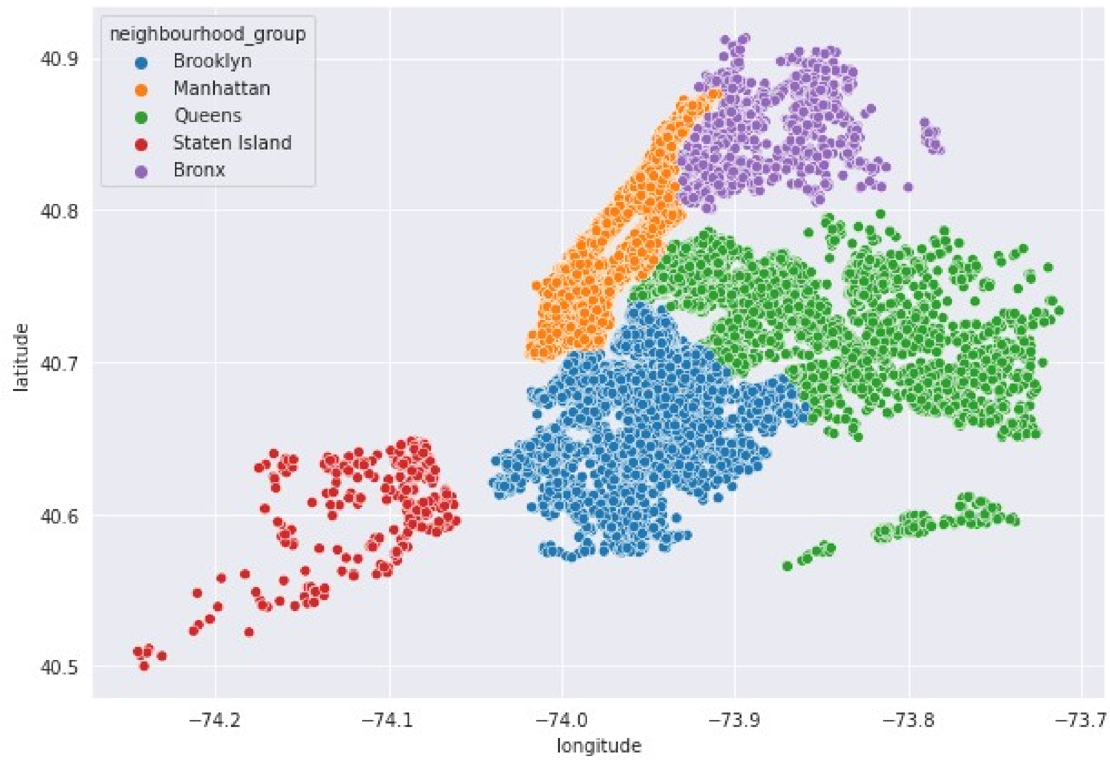
#**Observation**

The above pie chart shows that the highest number of airbnb listing is from Manhattan and Brooklyn Boroughs of New York city.

And Manhattan boroughs have most active number of host.

The same can be drived from the map of neighborhood group.

```
plt.figure(figsize = (10,7))
sns.scatterplot(x = airbnb.longitude, y = airbnb.latitude, hue =
airbnb.neighbourhood_group)
plt.show()
```

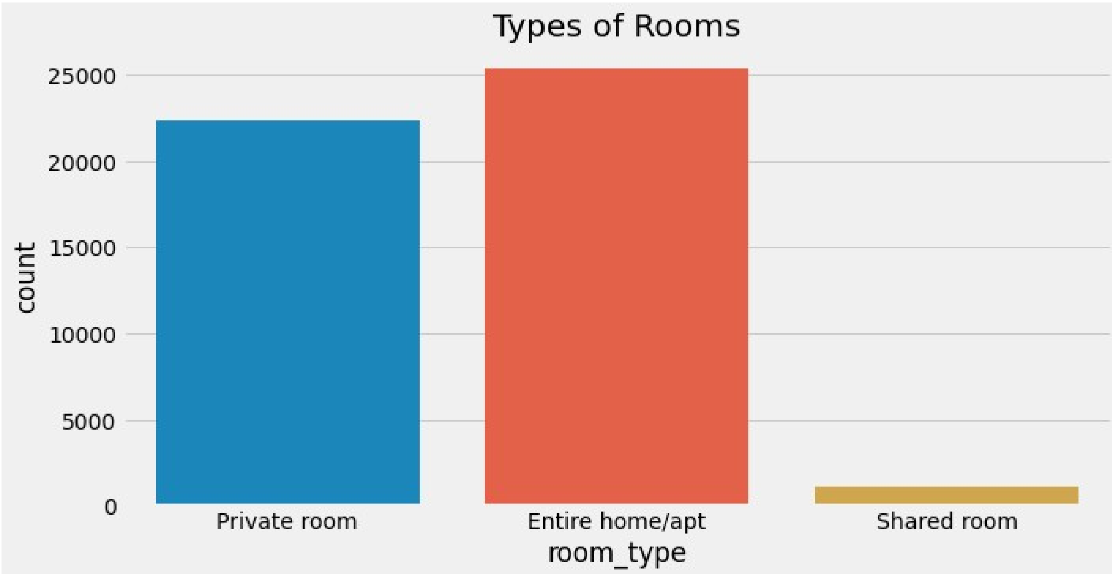


Room details

Q2. Which room type has the highest and the lowest number of booking ?

```
# Chart - 2 visualization code
plt.figure(figsize = (10,5))
plt.title('Types of Rooms')
sns.countplot(airbnb.room_type)
fig = plt.gcf()
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From
version 0.12, the only valid positional argument will be `data`, and
passing other arguments without an explicit keyword will result in an
error or misinterpretation.
  warnings.warn(
```



Insight found from the chart?

We can see that entire home/apt have the highest booking followed by private room and the least booked is shared room.

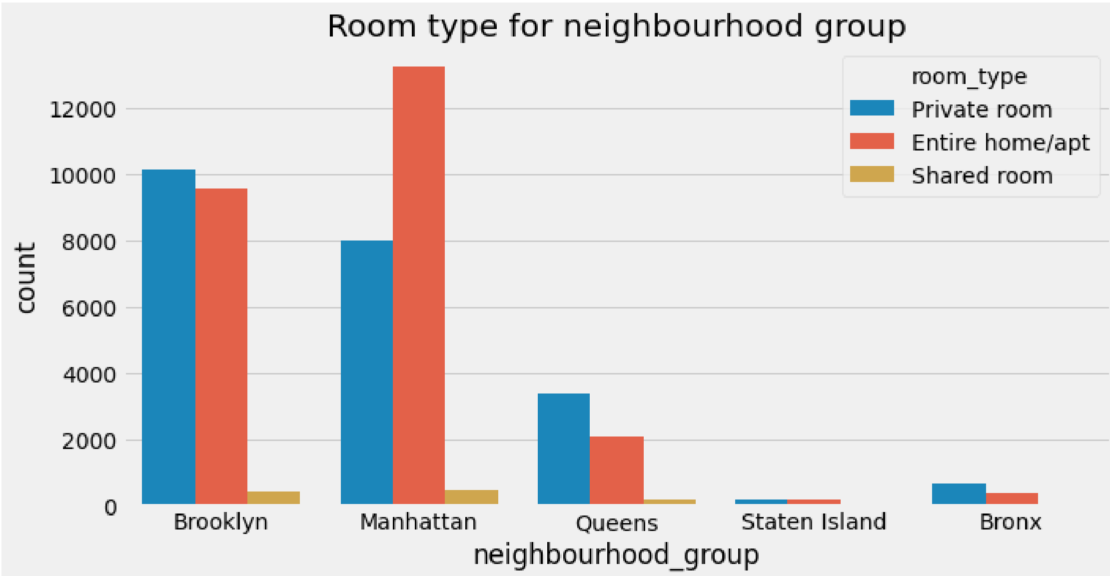
Chart - 3

Room type for neighborhood group

Q3. According to the neighborhood group which room type has the most number of booking ?

```
# Chart - 3 visualization code
plt.figure(figsize = (10,5))
plt.title('Room type for neighbourhood group')
sns.countplot(airbnb.neighbourhood_group, hue = airbnb.room_type)
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From
version 0.12, the only valid positional argument will be `data`, and
passing other arguments without an explicit keyword will result in an
error or misinterpretation.
  warnings.warn(
```



Insight found from the chart?

It clearly shows in Manhattan entire home/apt is the most booked while in Brooklyn private room is the most booked which is also same in queens and bronx but in staten island entire home/apt is slightly higher.

Price Exploration

```
#checking null values
airbnb['price'].isna().sum()
```

```
0
```

```
airbnb['price'].head(10)
```

```
0    149
1    225
2    150
3     89
4     80
5    200
6     60
7     79
8     79
9    150
Name: price, dtype: int64
```

```
airbnb['price'].describe()
```

```
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
```

Observation

From statistics summary the price range is from 0-10000. But the maximum price is of 10000 which can be due to location,room type, neighbourhood , season etc. we also have minimum price of 0 which can be due to dynamic pricing or the willingness of not to share the price with the Airbnb or may be there was no booking at all.

```
airbnb['price'][airbnb['price'] == 0].value_counts()
```

```
0    11
Name: price, dtype: int64
```

```
# Replacig a 0 with mean in price column
airbnb['price'].replace(to_replace = 0,value = airbnb['price'].mean(), 
inplace = True)
airbnb['price'].describe()
```

```
count    48895.000000
mean       152.755045
std        240.143242
min         10.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```

#**Highest Prices in 5 boroughs of New York city**

```
#high prices
high_price = airbnb.groupby('neighbourhood_group',as_index = False)
['price'].max()
high_price
```

```
  neighbourhood_group    price
0               Bronx   2500.0
1            Brooklyn  10000.0
2           Manhattan  10000.0
3              Queens  10000.0
4       Staten Island   5000.0
```

#**Observation** From above table we can see Brooklyn, Manhattan and Queens have prices of 10,000.

```
#low prices
low_price = airbnb.groupby('neighbourhood_group', as_index = False)
['price'].min()
low_price
```

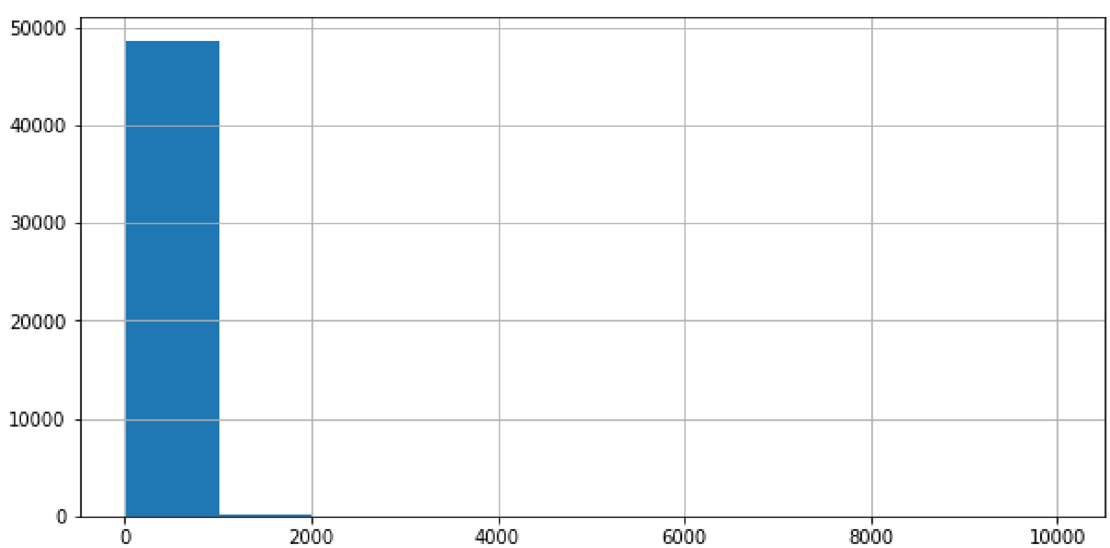
```
  neighbourhood_group  price
0               Bronx   10.0
1            Brooklyn   10.0
2           Manhattan   10.0
3              Queens   10.0
4       Staten Island   13.0
```

#**Observation** Lowest prices are at places Bronx, Brooklyn, Manhattan and queens.

```
airbnb['price'].describe()
```

```
count    48895.000000
mean       152.755045
std        240.143242
min         10.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```

```
hist_price = airbnb['price'].hist()
```

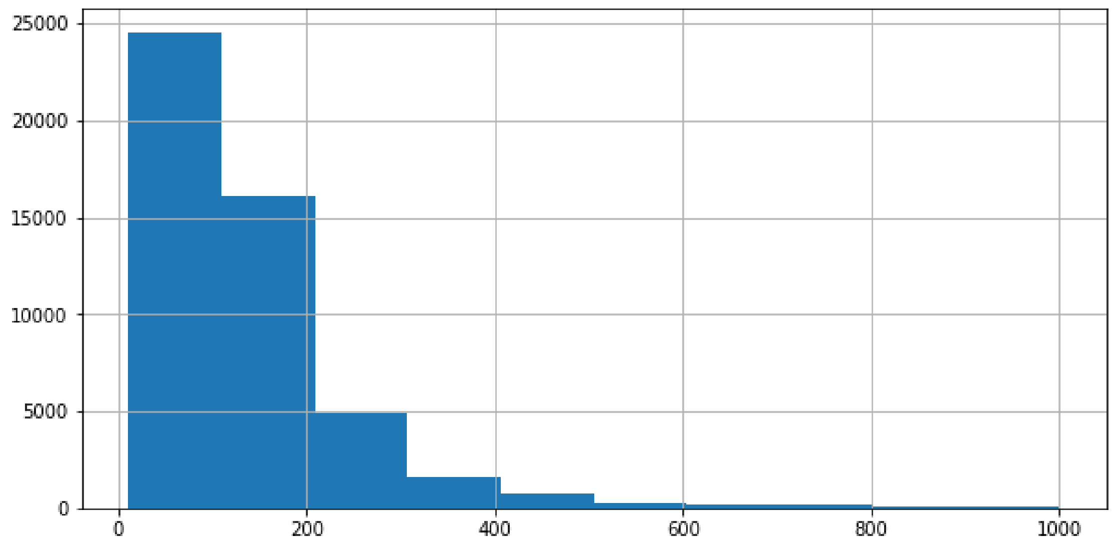


Observation :

Most value is less than 1000. So the most dominanting price range is under 1000 which hold nearly 48,000 properties.

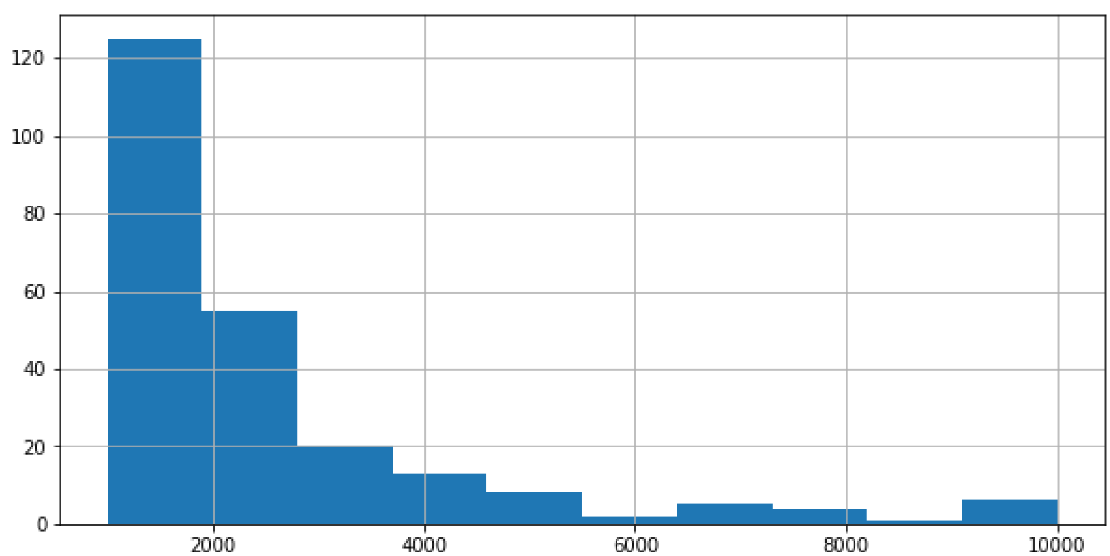
So there are 48,000 properties with price less than 1000.

```
# Focusing on rental rate which are less than 1000
hist_price = airbnb['price'][airbnb['price']<=1000].hist()
```



#**Observation** Price range is more dominating under 200. which means even if we know that most properties are under price range 1000 even inside it the most dominating one is with 200.

```
# Rantel rate greater than 1000
hist_price = airbnb["price"][airbnb['price'] > 1000].hist()
```



#**Observation** It is safe to say that most of the properties fall under the price range of 2000 in it. And most dominanting range is from 1000 to 4000.

```
# Counting price greater than 1000
price_greater_than_2000 = airbnb[airbnb['price'] >
1000].value_counts()
print(price_greater_than_2000)
```

```
host_id    host_name    neighbourhood_group  neighbourhood    latitude
longitude  room_type        price   minimum_nights  number_of_reviews
reviews_per_month  calculated_host_listings_count  availability_365
8730       Allison      Manhattan            Chelsea          40.73692
-73.99219  Entire home/apt  1495.0  1                 11
0.22               1                               0
1
75110137   Christina    Brooklyn             Gowanus          40.68494
-73.98850  Private room     1333.0  50                 0
0.00               1                               365
1
60535711   Bruce        Manhattan            Midtown          40.76040
-73.97410  Entire home/apt  1100.0  2                 20
0.53               2                               268
1
63492343   Lenore       Manhattan            Chelsea          40.73971
-73.99611  Entire home/apt  1050.0  2                  0
0.00               1                               365
1
65562107   Gina         Manhattan            Tribeca          40.71868
-74.00765  Entire home/apt  1200.0  5                 13
0.98               1                               347
1


..
11460768   Brian        Manhattan            Upper West Side  40.80020
-73.96045  Entire home/apt  1500.0  1                  0
0.00               1                               0
1
11461854   Lauren       Manhattan            West Village     40.73295
-74.00755  Entire home/apt  1500.0  1                  0
0.00               1                               0
1
11490872   Nick         Manhattan            Kips Bay         40.74422
-73.97822  Entire home/apt  1550.0  2                  0
0.00               1                               0
1
11492501   Victoria     Manhattan            Stuyvesant Town  40.73205
-73.98094  Entire home/apt  1500.0  1                  0
0.00               1                               0
1
272166348  Mary Rotsen  Manhattan            Upper East Side  40.78132
```

```
-73.95262  Entire home/apt  1999.0  30              0
0.00               1                               270
1
Length: 239, dtype: int64
```

Observation

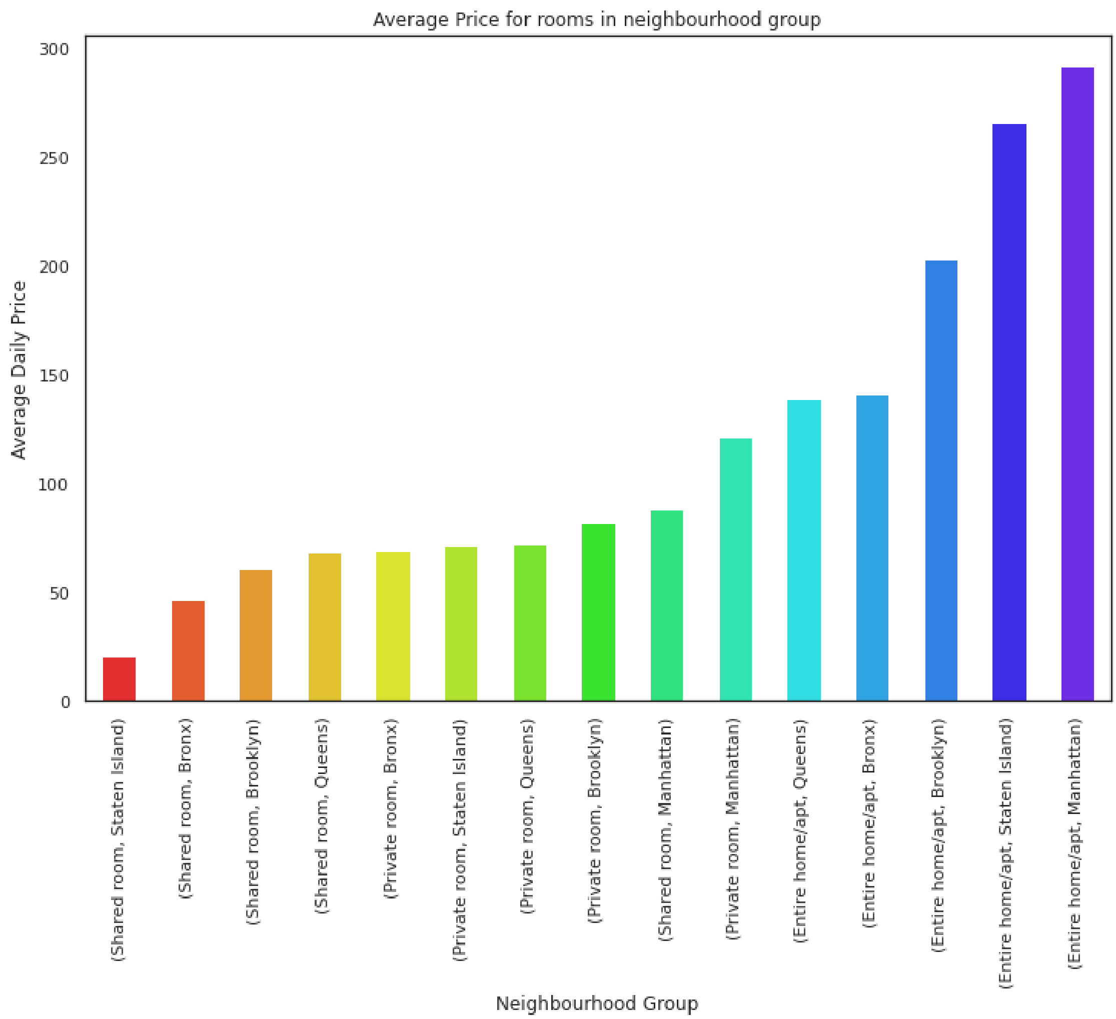
We can easily see their are 239 listing whose prices are greater than 1000.

So they might be luxurious property or it can be a error during input.

Average room price for locality

Q4. What is the average price per night for different room type based on neighborhood group ?

```
# Average room price per night
plt.figure(figsize=(12,8))
df = airbnb[airbnb['minimum_nights']==1]
df1 = df.groupby(['room_type','neighbourhood_group'])
['price'].mean().sort_values(ascending=True)
df1.plot(kind='bar', color =('#e33030',
'#e35d30','#e39830','#e3c230','#dae330','#b0e330','#7be330','#39e330',
'#30e37e','#30e3b0','#30dde3','#30a4e3','#3081e3','#3d2ee6','#6e2ee6')
)
plt.title('Average Price for rooms in neighbourhood group')
plt.ylabel('Average Daily Price')
plt.xlabel('Neighbourhood Group')
plt.show()
print('List of Average Price per night based on the neighbourhood
group')
pd.DataFrame(df1).sort_values(by='room_type')
```



List of Average Price per night based on the neighbourhood group

room_type	neighbourhood_group	price
Entire home/apt	Queens	139.036260
	Bronx	141.541176
	Brooklyn	202.895245
	Staten Island	266.205128
	Manhattan	291.784595
Private room	Bronx	69.025862
	Staten Island	71.394366
	Queens	72.454958
	Brooklyn	81.713284
	Manhattan	121.434183
Shared room	Staten Island	21.000000
	Bronx	46.711111
	Brooklyn	60.921212
	Queens	68.459459
	Manhattan	88.462898

Observation

From looking at plot few things are clear

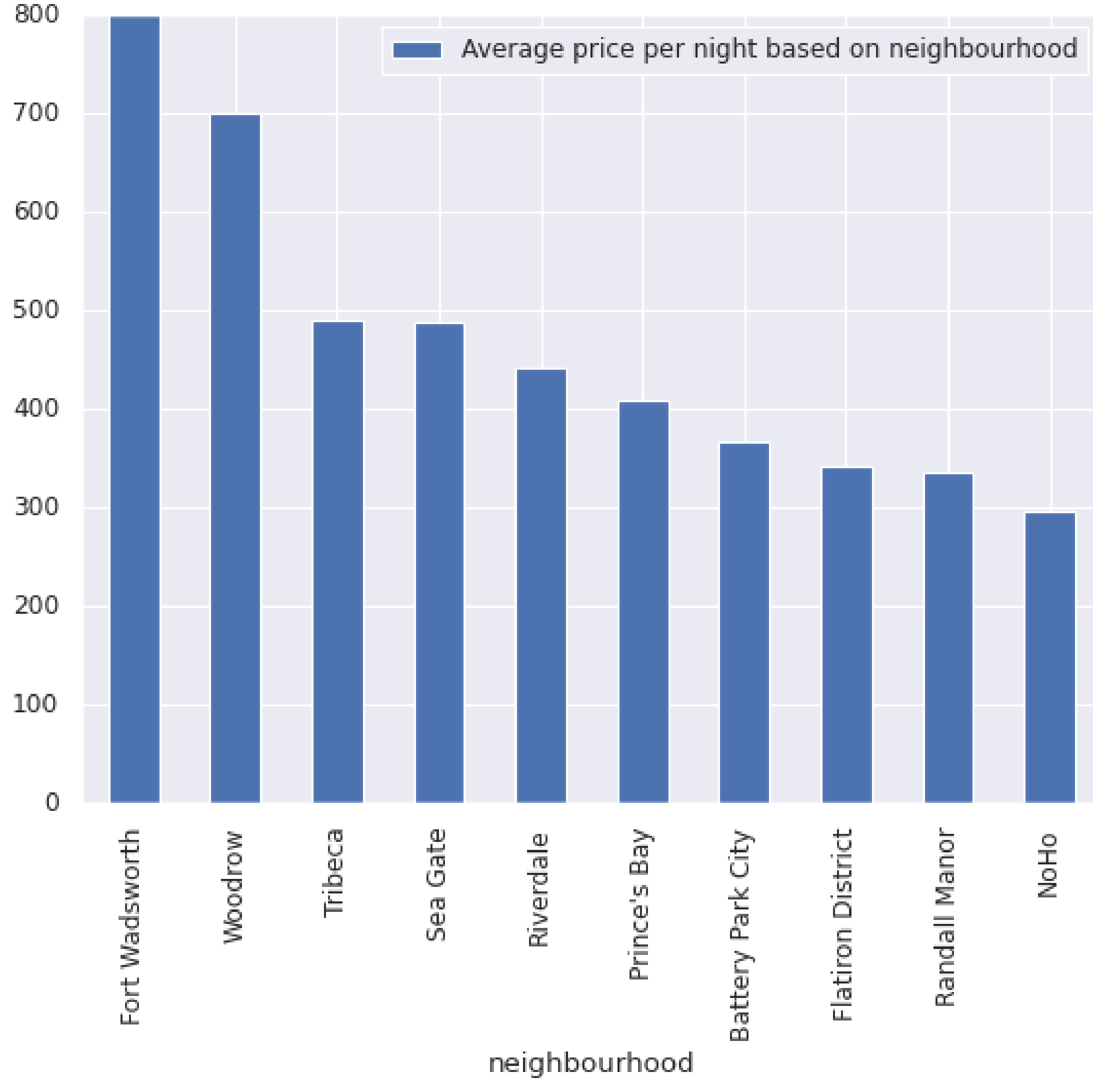
1. Shared room at staten Island is the most cheapest stay per night whereas Renting a Entire apartment/Home at Manhattan per night is the most expensive.
2. Average price for Private room is also considerably expensive at manhattan so is the shared room at Manhattan is expensive than other private rooms of the neighbourhood. This clearly states that Manhattan is the expensive stay than any other locality.
3. Bronx is the most cheapest stay in terms of neighbourhood group comparison in respect to room type.
4. Though Shared room at Staten Island is the cheapest whereas Apartment renting is not cheapest at Staten Island

Costly Neighborhood

Q5. Top ten neighborhoods with the most expensive prices ?

```
# Now we will be checking average prices across each neighborhood
# Top 10 neighborhood
print('Top 10 most expensive locality in Airbnb listing are :')
df4 = airbnb.dropna(subset=["price"]).groupby("neighbourhood")
[["neighbourhood", "price"]].agg("mean").sort_values(by="price",
                         ascending=False).rename(index=str,
columns={"price": "Average price per night based on
neighbourhood"}).head(10)

df4.plot(kind='bar', color = (''))
plt.show()
pd.DataFrame(df4)
```

```
Top 10 most expensive locality in Airbnb listing are :
```



neighbourhood	Average price per night based on neighbourhood
Fort Wadsworth	800.000000
Woodrow	700.000000
Tribeca	490.638418
Sea Gate	487.857143
Riverdale	442.090909
Prince's Bay	409.500000
Battery Park City	367.557143
Flatiron District	341.925000
Randall Manor	336.000000
NoHo	295.717949

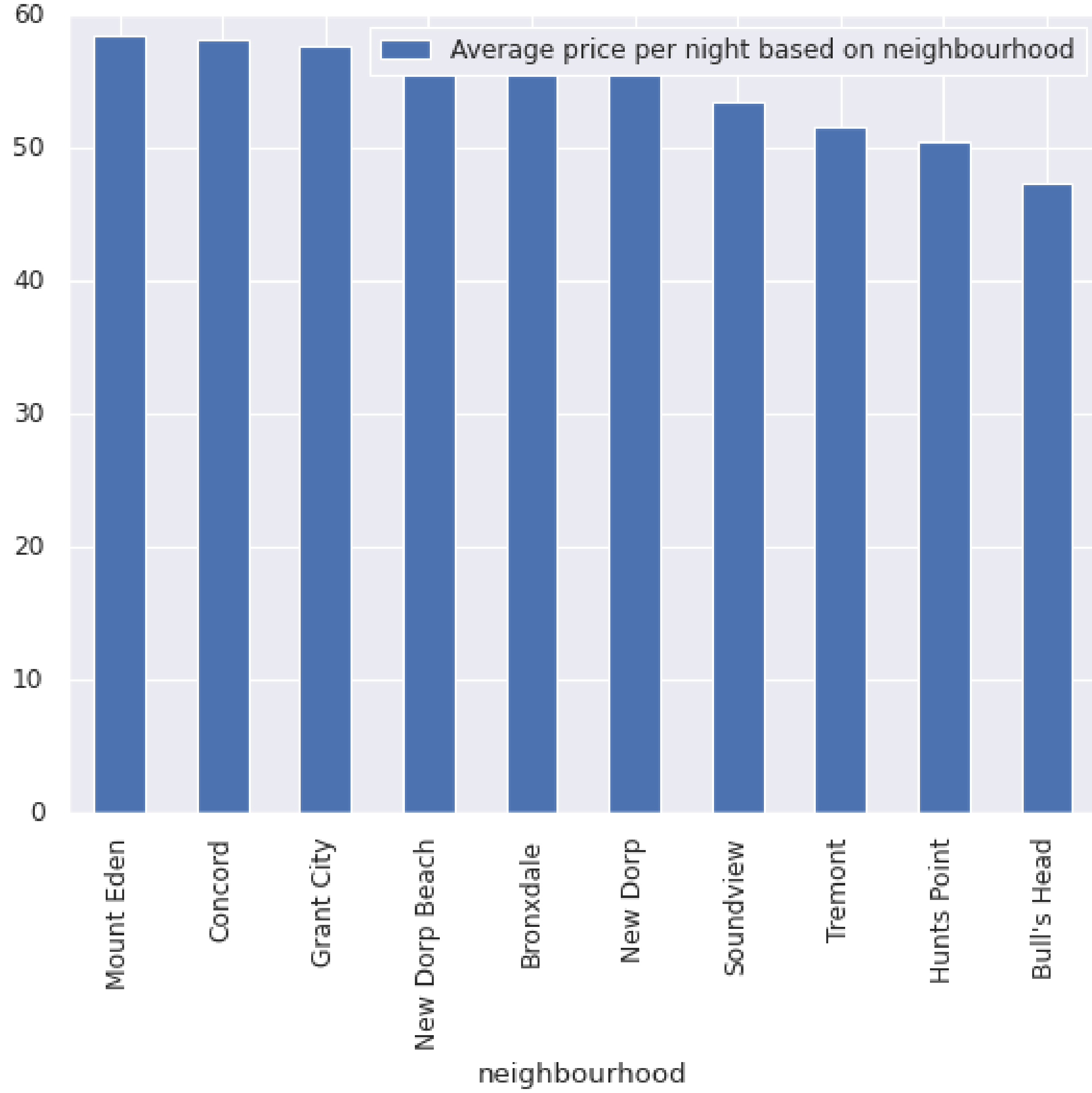
Observation

According to the graph and data Fort Wadsworth is most costly neighborhood.

Q6. Top ten neighborhoods with the cheapest prices ?

```
# top ten least costly neighborhood
print('Least expensive neighbourhood according to Airbnb listing are')
df4 = airbnb.dropna(subset=["price"]).groupby("neighbourhood")
[["neighbourhood", "price"]].agg("mean").sort_values(by="price",
       ascending=False).rename(index=str, columns={"price": "Average
price per night based on neighbourhood"}).tail(10)

df4.plot(kind='bar')
plt.show()
pd.DataFrame(df4)
```

```
Least expensive neighbourhood according to Airbnb listing are
```



neighbourhood	Average price per night based on neighbourhood
Mount Eden	58.500000
Concord	58.192308
Grant City	57.666667
New Dorp Beach	57.400000
Bronxdale	57.105263
New Dorp	57.000000
Soundview	53.466667
Tremont	51.545455
Hunts Point	50.500000
Bull's Head	47.333333

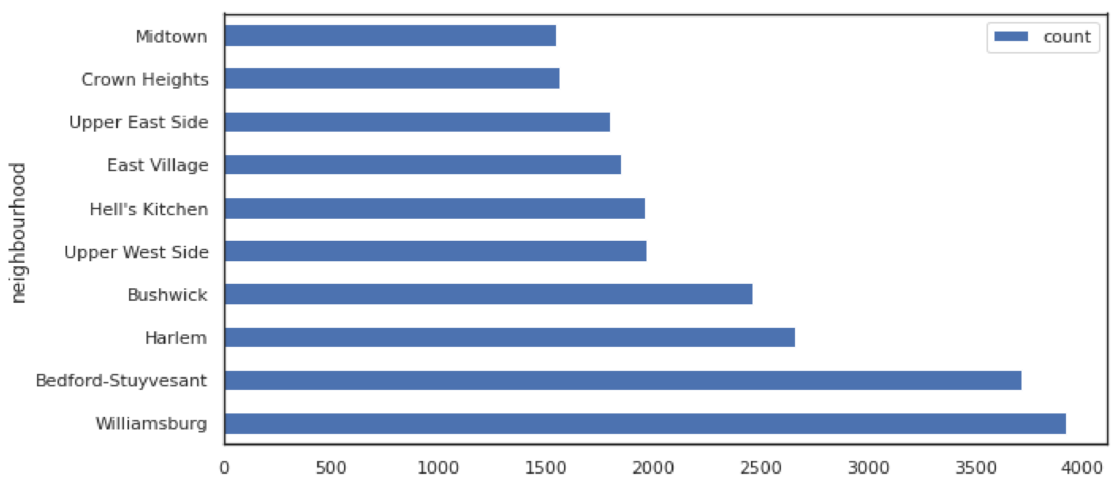
Observation

Least costly or the cheapest neighborhood to stay is Bull's Head.

Neighborhood which have most number of booking

Q7. What are the top ten neighborhoods with most number of booking ?

```
# Top 10 listing based on neighborhood
df5 = airbnb.groupby('neighbourhood')
[['neighbourhood','host_name']].agg(['count']
              )
['host_name'].sort_values(by='count',ascending=False).rename(index=str
,columns={'Count':'Listing Count'})

df5.head(10).plot(kind='barh')
plt.show()
pd.DataFrame(df5.head(10))
```




```
                    count
neighbourhood
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
```

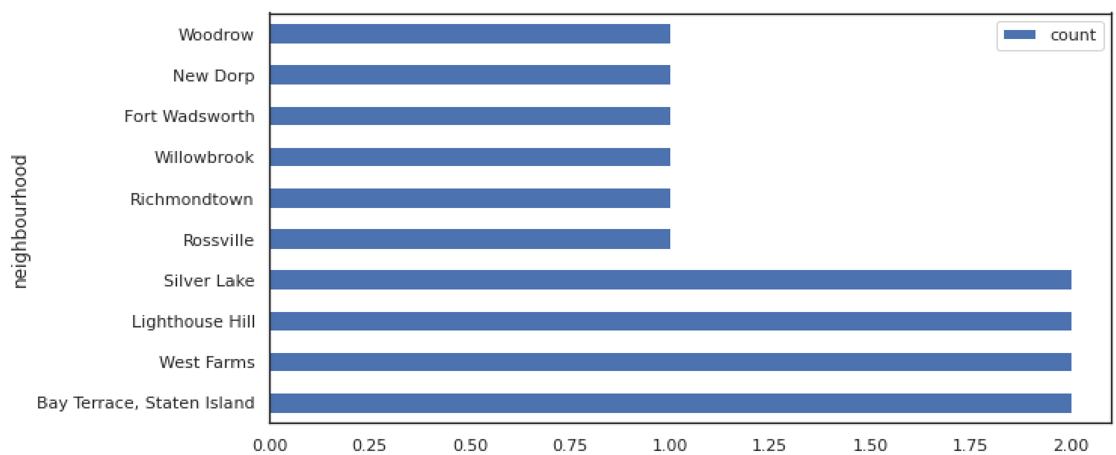
Observation

We can see Williamsburg has most number of listing count.

Q8. What are the top ten neighborhoods with the least number of booking ?

```
# Bottom 10 listing based on neighborhood

print('Least Listing number of count')
df5 = airbnb.groupby('neighbourhood')
[['neighbourhood','host_name']].agg(['count']
       )
['host_name'].sort_values(by='count',ascending=False).rename(index=str
,columns={'Count':'Listing Count'})

df5.tail(10).plot(kind='barh')
plt.show()
pd.DataFrame(df5.tail(10))
```

```
Least Listing number of count
```




```
                            count
neighbourhood
Bay Terrace, Staten Island      2
West Farms                      2
Lighthouse Hill                 2
Silver Lake                     2
Rossville                       1
Richmondtown                    1
Willowbrook                     1
Fort Wadsworth                  1
New Dorp                        1
Woodrow                         1
```

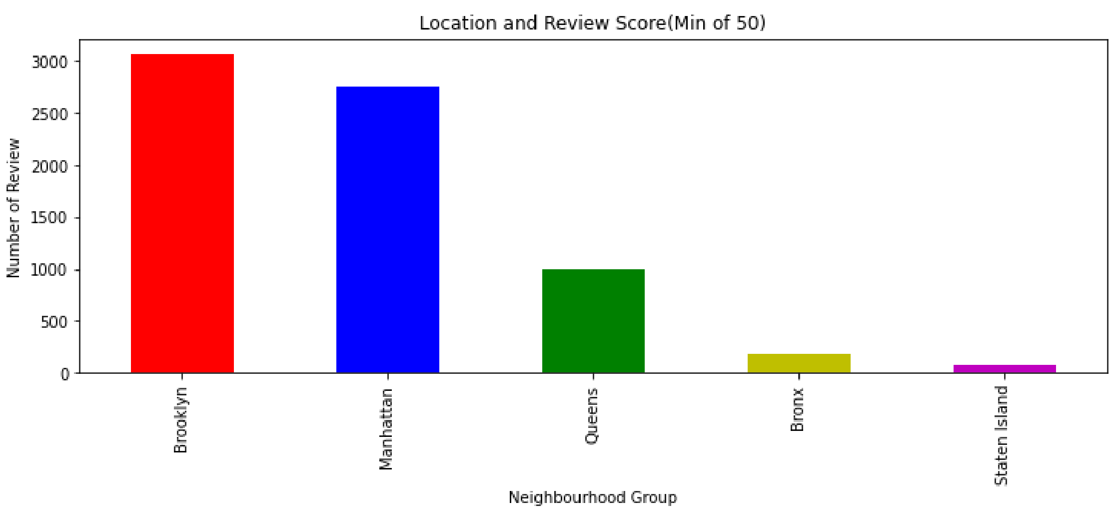
Observation

So Rossville, Richmondtown, Willowbrook, Fort Wadsworth, New Dorp and Woodrow have least number of booking.

Analysis on number of reviews neighborhood groups have got

Q9. Which neighborhood group have most number of reviews ?

```
fig = plt.figure(figsize=(12,4))
review_50 = airbnb[airbnb['number_of_reviews']>=50]
df2 = review_50['neighbourhood_group'].value_counts()
df2.plot(kind='bar',color=['r','b','g','y','m'])
plt.title('Location and Review Score(Min of 50)')
plt.ylabel('Number of Review')
plt.xlabel('Neighbourhood Group')
plt.show()
print(' Count of Review v/s neighbourhood group')
pd.DataFrame(df2)
```




```
 Count of Review v/s neighbourhood group

               neighbourhood_group
Brooklyn                      3065
Manhattan                     2751
Queens                         997
Bronx                          187
Staten Island                   81
```

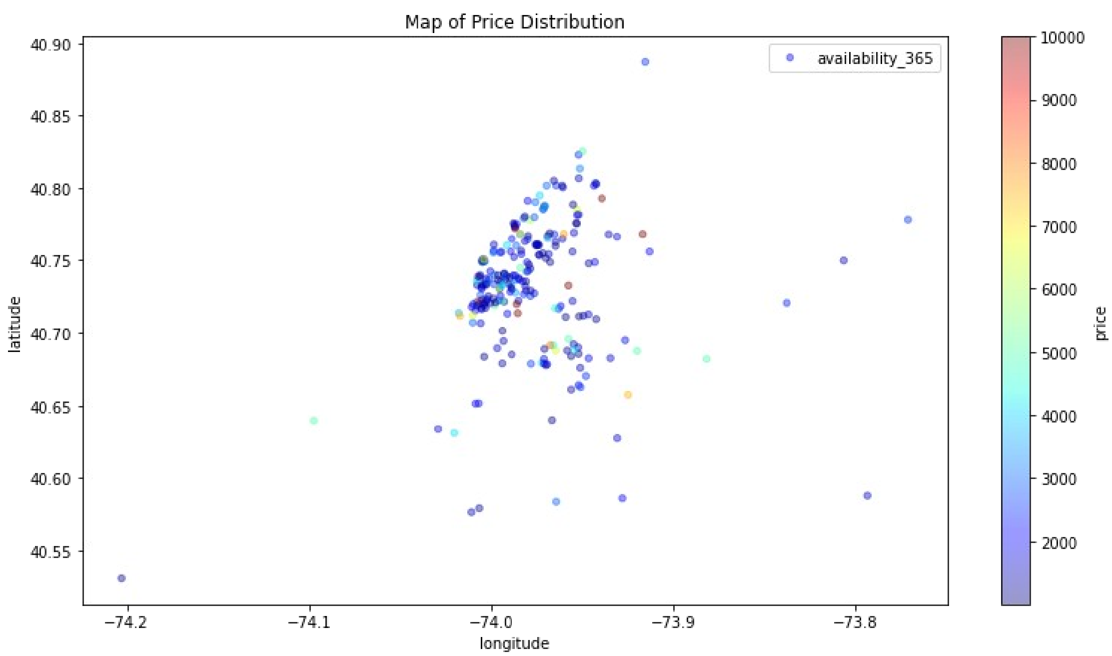
Observation

So brooklyn have most number of reviews where as staten island have least number of reviews.

Neighbourhood Group Price Distribution

Q10. Which neighborhood group has the most booking show using scatter plot ?

```
plt.figure(figsize=(13,7))
plt.title("Map of Price Distribution")
ax=plt.gca()
ax=airbnb[airbnb.price>1000].plot(kind='scatter',
x='longitude',y='latitude',label='availability_365',c='price',ax = ax,
cmap=plt.get_cmap('jet'),colorbar=True,alpha=0.4)
ax.legend()
plt.ioff()
plt.show()
```

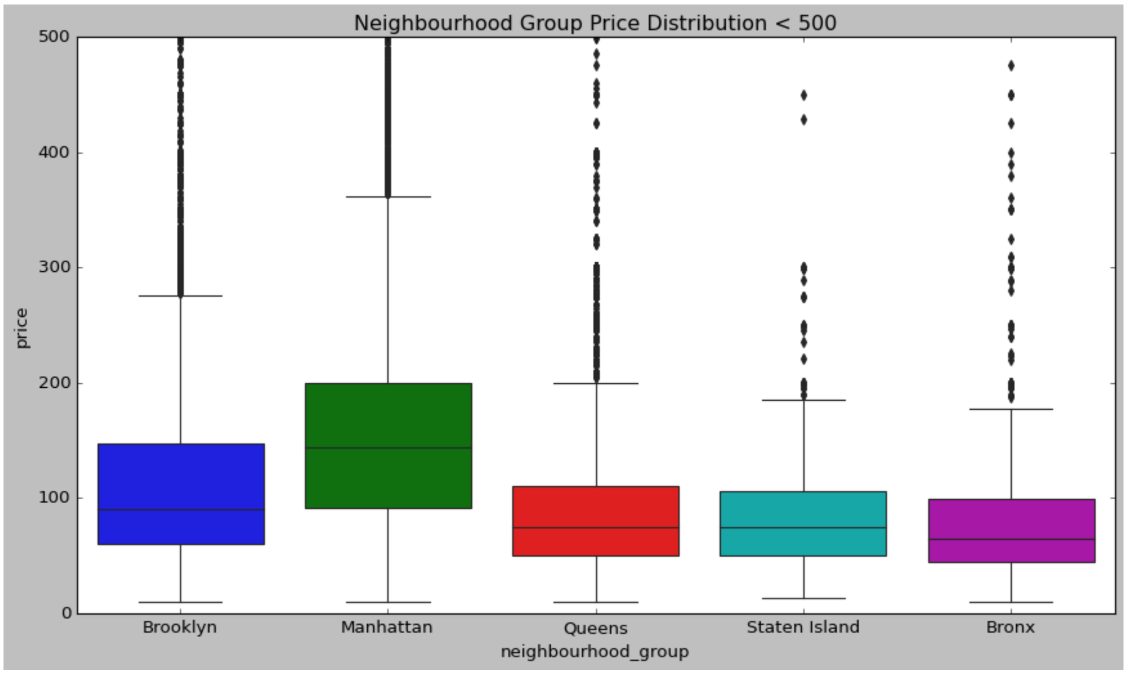


Observation

Most listing with prices greater than 1000 are from Manhattan which means Manhattan has the highest prices.

Q11. Show a relation between neighborhood group and prices using box plot ?

```
plt.style.use('classic')
plt.figure(figsize=(13,7))
plt.title("Neighbourhood Group Price Distribution < 500")
sns.boxplot(y="price",x ='neighbourhood_group' ,data =
airbnb[airbnb.price<500])
plt.show()
```



Observation

Boxplot above, we can observe a couple of things about the distribution of prices for Airbnb in NYC.

1. We observe that Manhattan has the highest prices for the listing with average being 140 which is followed by Brooklyn with 90.
2. Queens and Staten Island seem to have a very similar distribution.
3. Cheapest listing are from Bronx.

Host Name

Q13. Who are the top earning host ?

```
airbnb['host_name'][airbnb['host_name'] == 0].value_counts()
```

```
0    21
Name: host_name, dtype: int64
```

```
top_host=airbnb.groupby(['host_name','host_id'])
['price'].sum().reset_index()
top_host.rename(columns={'price':'total_price'},inplace=True)
top_host.head()
```

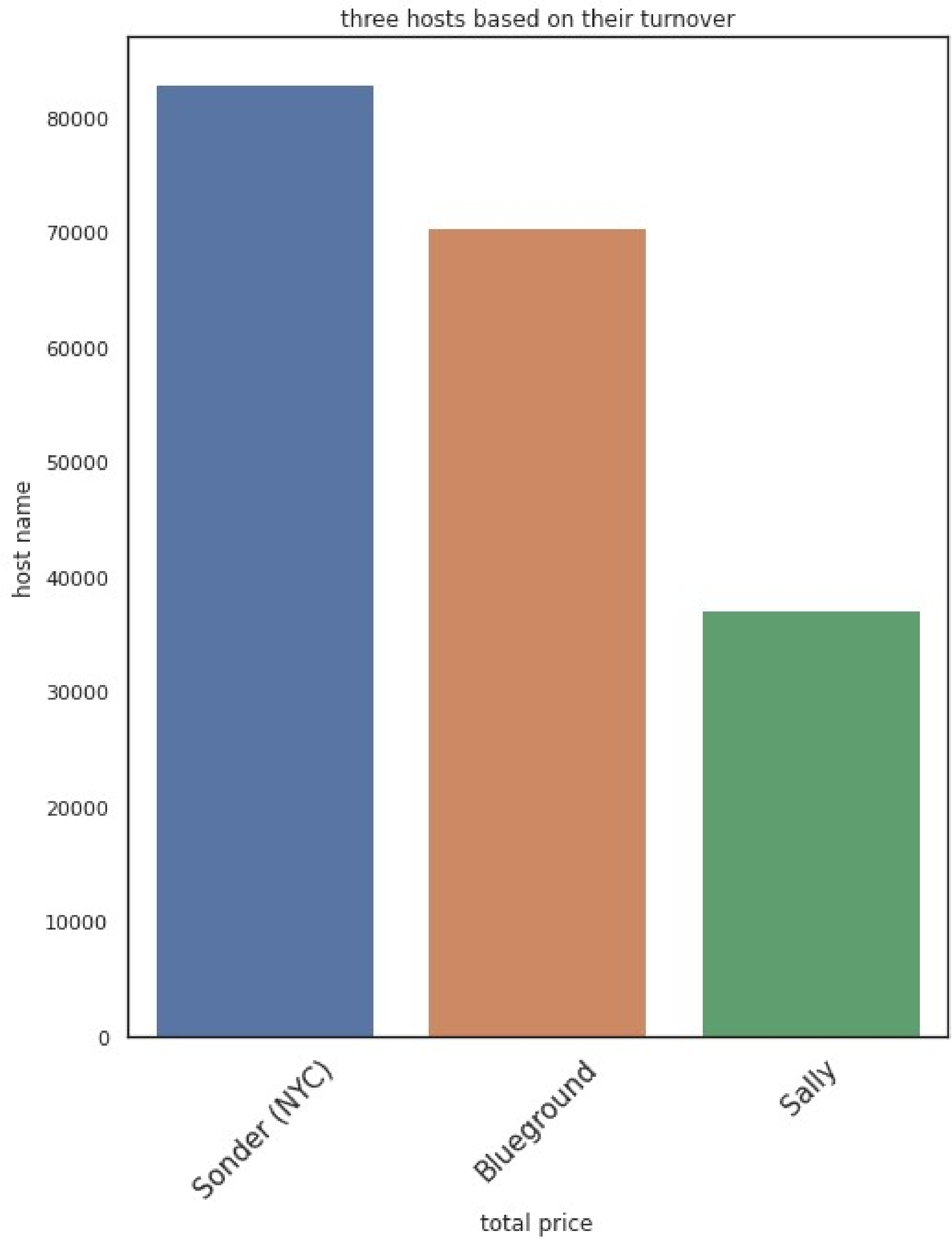
```
  host_name  host_id  total_price
0         0   415290          325
1         0   526653           50
2         0   919218           86
3         0  5162530          145
4         0  5300585          220
```

```
top_3=top_host.sort_values('total_price',ascending=False).iloc[:3,:3]
top_3
```

```
           host_name    host_id  total_price
33240  Sonder (NYC)  219517861        82795
4876     Blueground  107434423        70331
31247         Sally  156158778        37097
```

```
sns.set(rc={'figure.figsize':(8,10)})
sns.set_style('white')
abc= sns.barplot(x='host_name',y='total_price',data = top_3)
abc.set_title('three hosts based on their turnover')
abc.set_ylabel('host name')
abc.set_xlabel('total price')

abc.set_xticklabels(abc.get_xticklabels(),rotation = 45,size='15')
```

```
[Text(0, 0, 'Sonder (NYC)'), Text(1, 0, 'Blueground'), Text(2, 0,
'Sally')]
```



#**Observation** So sonder, blueground, sally are the top hosts.

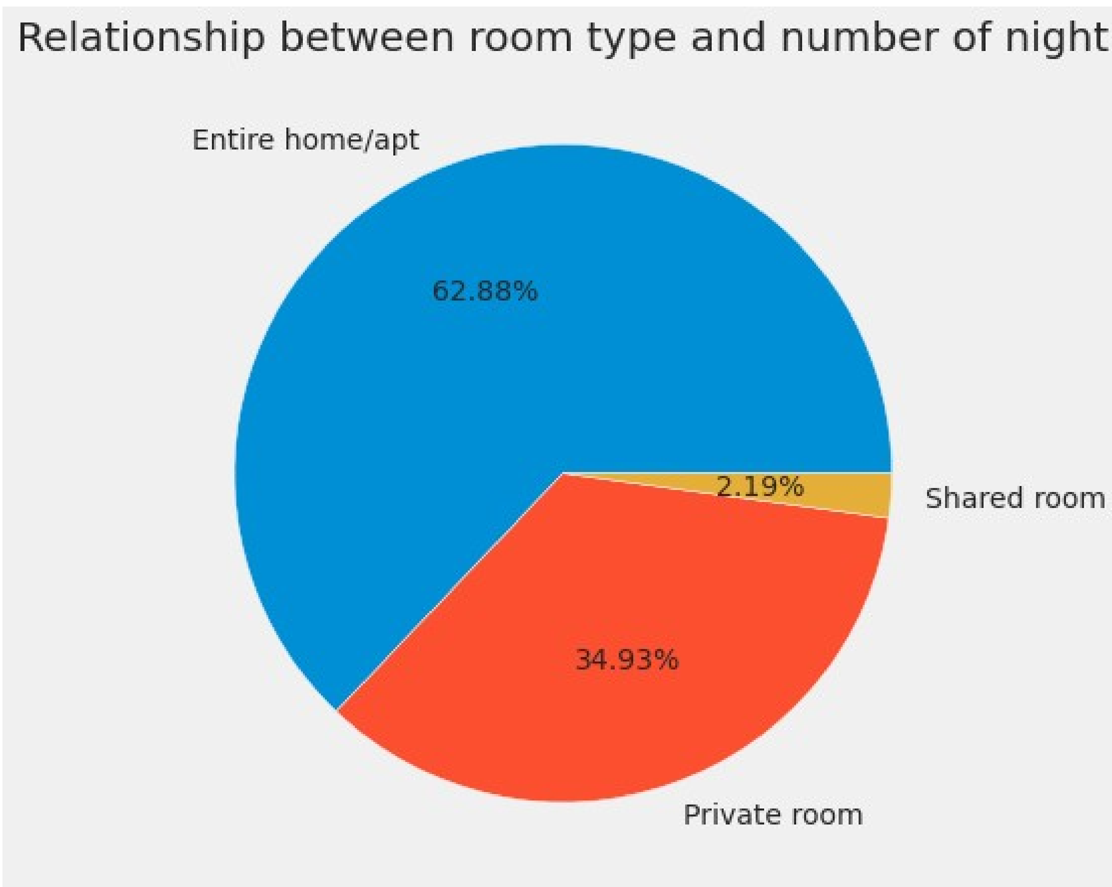
Q14. Which room type has been occupied for the most number of nights ?

```
# Finding the sum of minimum_nights
sum_room = airbnb.groupby('room_type')
```

```
['minimum_nights'].sum().reset_index()
sum_room
```

```
         room_type  minimum_nights
0  Entire home/apt          216152
1     Private room          120067
2      Shared room            7511
```

```
# Ploting the chart
plt.style.use('fivethirtyeight')
plt.figure(figsize=(10,7))
plt.title('Relationship between room type and number of night')
plt.pie(sum_room['minimum_nights'],labels
=sum_room['room_type'],autopct='%.2f%%' )
plt.show()
```



#**Observation**

- From pie chart we can determine that 63.2% customers spend night in entire home/apt.
- Only 1.6% customers spend night in shared room.

#**Correlational matrix**

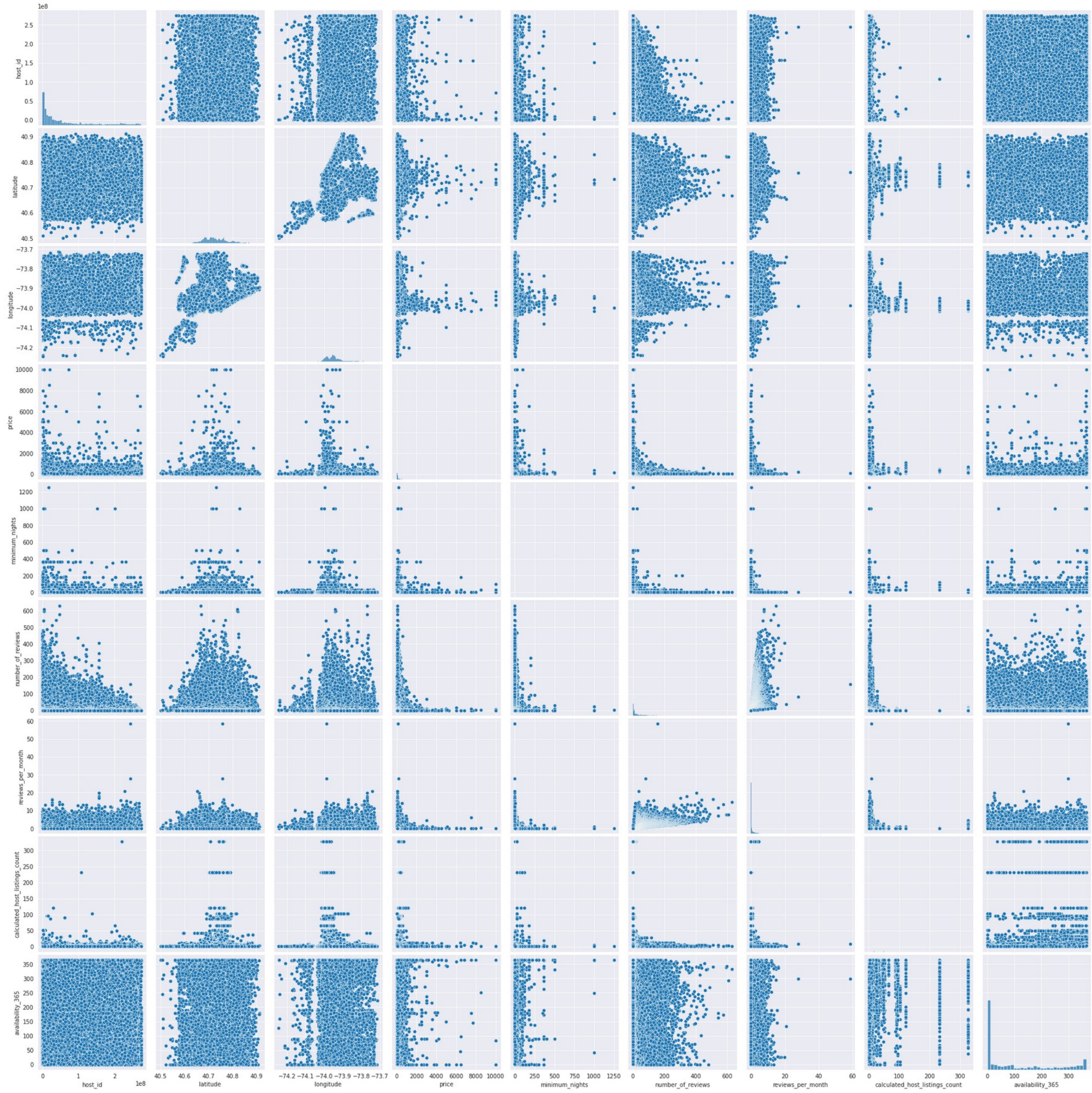
```
# Corelation between different variables
corr = airbnb.corr(method='kendall')
plt.figure(figsize=(13,10))
plt.title("Correlation Between Different Variables\n")
sns.heatmap(corr, annot=True)
plt.show()
```

#**Observation** positive correlation with minimum number of nights, availability of 365 days. Calculated host listings have negative correlation with price and the above graph also shows least correlation with number of reviews.

#**Pair Plot**

```
# Visualizating our data set through pairplot
plt.figure(figsize=(30, 30))
sns.pairplot(airbnb, height=3, diag_kind="hist")
plt.show()
```

```
<Figure size 2160x2160 with 0 Axes>
```



Observation

- latitude and longitude have a normal distribution, most of the hosts are concetrated in specific area.
- reviews_per_month has a lot of outlayers, because of the missing values filled by average.
- availability_365 the most of the hosts are not available all the year.
- price most the host has a price under $1000

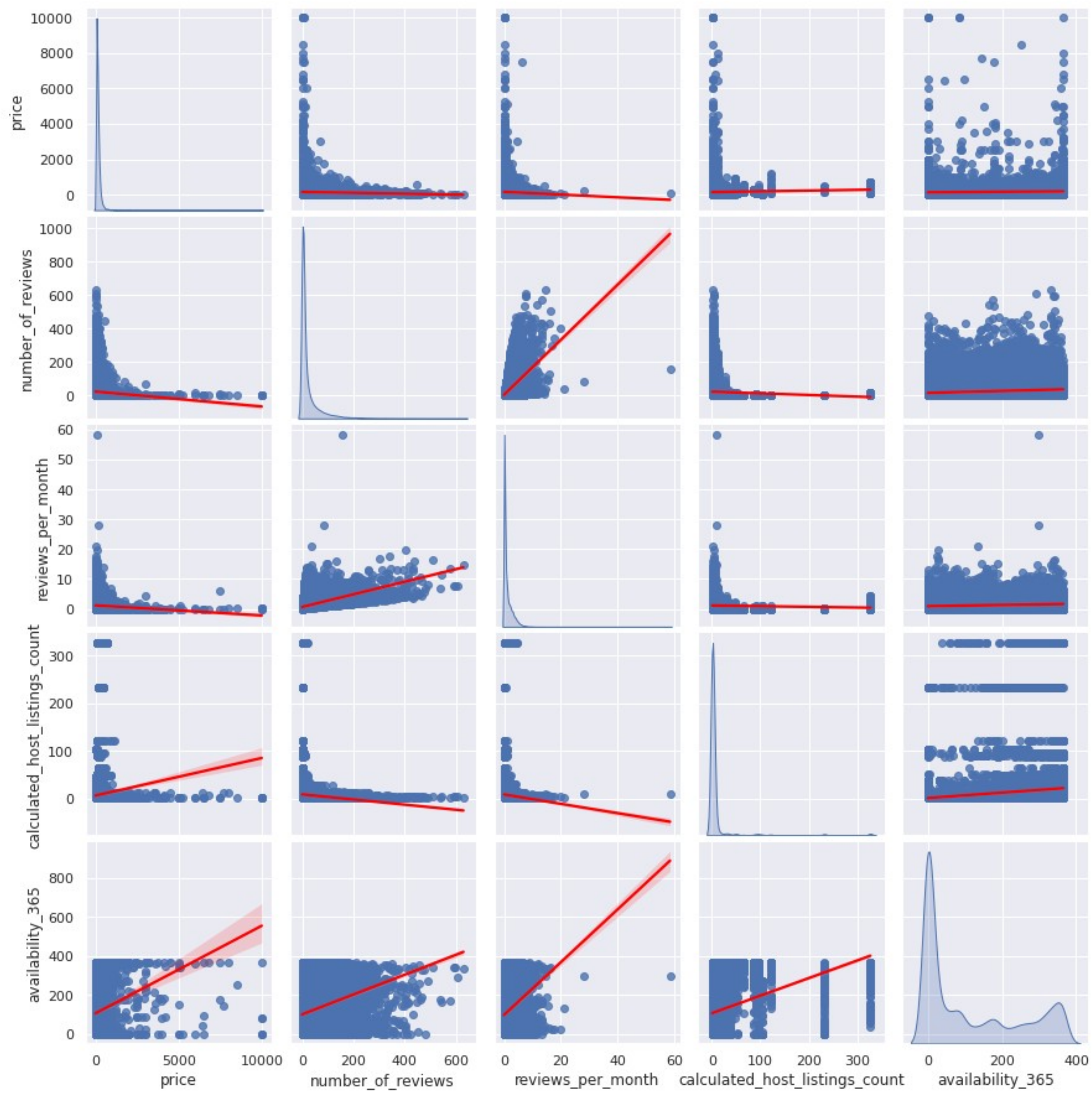
```
airbnb.head(5)
```

```
   host_id    host_name neighbourhood_group neighbourhood  latitude  \
0     2787         John            Brooklyn    Kensington  40.64749
```

```
1     2845     Jennifer          Manhattan       Midtown  40.75362
2     4632    Elisabeth          Manhattan        Harlem  40.80902
3     4869  LisaRoxanne           Brooklyn  Clinton Hill  40.68514
4     7192        Laura          Manhattan   East Harlem  40.79851

   longitude        room_type  price  minimum_nights
number_of_reviews  \
0  -73.97237     Private room    149               1
9
1  -73.98377  Entire home/apt    225               1
45
2  -73.94190     Private room    150               3
0
3  -73.95976  Entire home/apt     89               1
270
4  -73.94399  Entire home/apt     80              10
9

   reviews_per_month  calculated_host_listings_count  availability_365

0               0.21                               6               365

1               0.38                               2               355

2               0.00                               1               365

3               4.64                               1               194

4               0.10                               1                 0
```

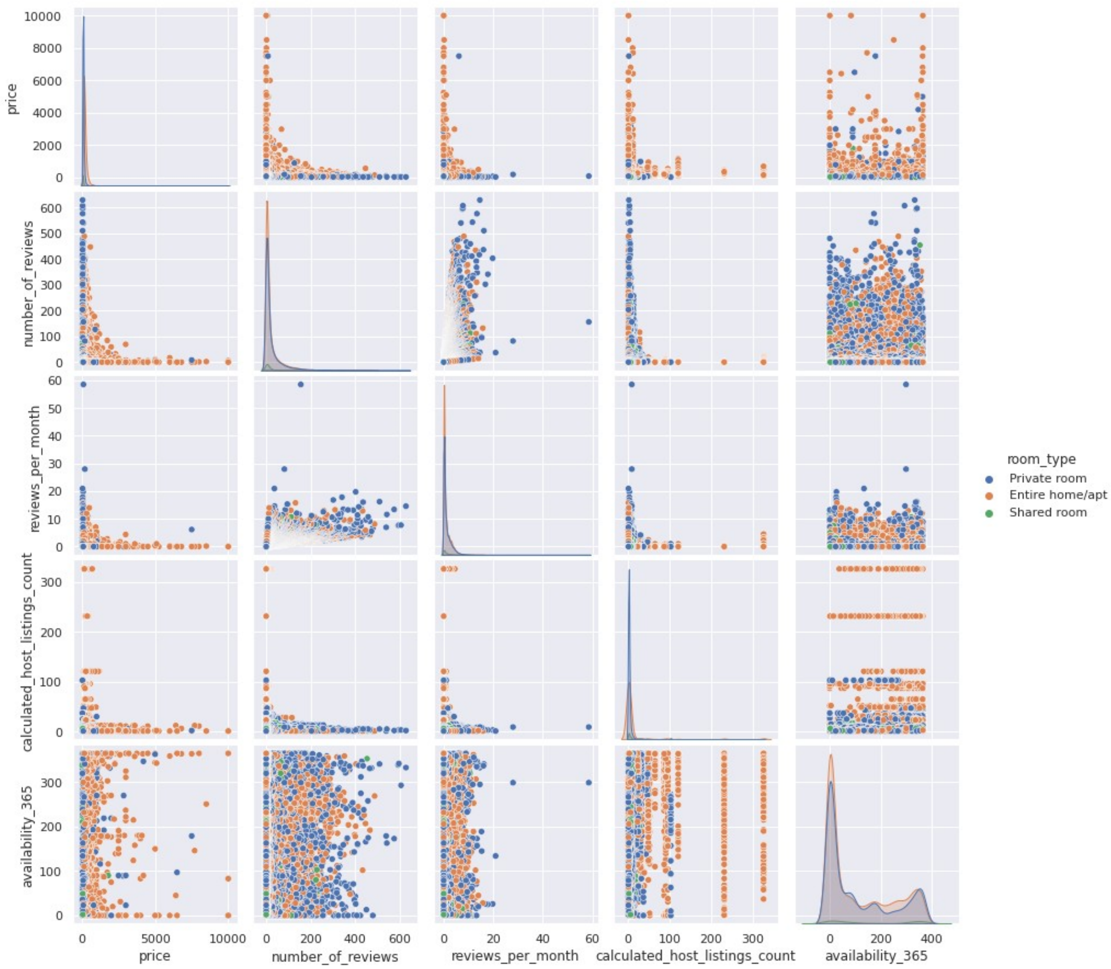
```python
# Selecting required columns
col_to_plot =
['price','number_of_reviews','reviews_per_month','calculated_host_list
ings_count','availability_365','room_type']
# Ploting pairplot
sns.set_style('darkgrid')
sns.pairplot(airbnb[col_to_plot], kind='reg', diag_kind='kde',
             plot_kws={'line_kws':{'color':'red'}})
```

```
<seaborn.axisgrid.PairGrid at 0x7f291521d790>
```



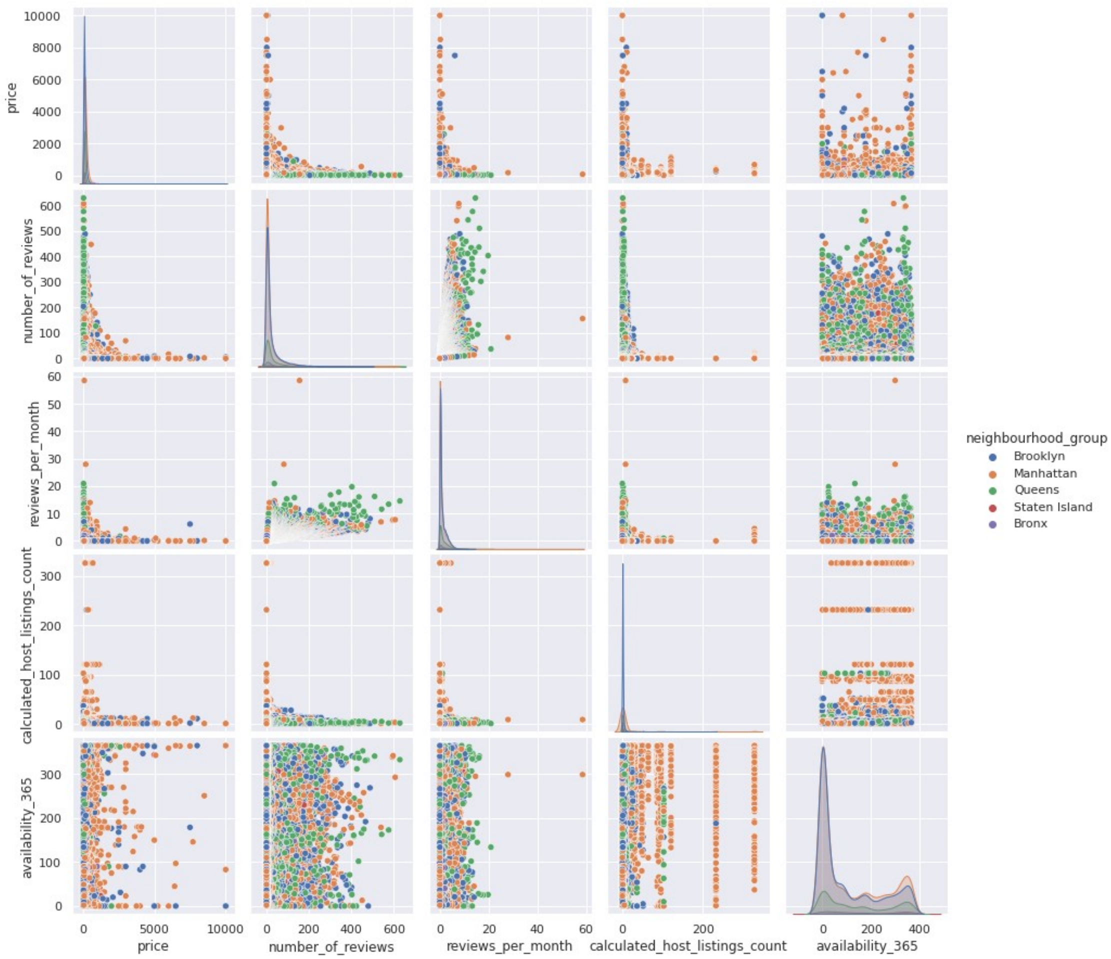

```
sns.pairplot(airbnb[col_to_plot], hue ='room_type')
```

```
<seaborn.axisgrid.PairGrid at 0x7f294e572670>
```




```
# Selecting required columns
col_to_plot =
['price','number_of_reviews','reviews_per_month','calculated_host_list
ings_count','availability_365','neighbourhood_group']
# Ploting pairplot
sns.set_style('darkgrid')
sns.pairplot(airbnb[col_to_plot], hue = 'neighbourhood_group')
```

```
<seaborn.axisgrid.PairGrid at 0x7f290bec3e20>
```



Conclusion

- Manhattan is the cream of the croop for airbnb as it has most number of customer as compared to other Boroughs of new york city.
- Customer pays highest amount of 10,000 for airbnb booking in Manhattan, Brooklyn and queens and the lowest amount is of 10.
- As we know that the entire home/apt has been occupied for the most number of night which is exactly 2,16,152 and in percentile it will 62.88% of all room type.
- So the average price for entire home/apt per night in different neighborhoods groups are as follows:

Queens 139.036260

#####Bronx 141.541176

#####Brooklyn 202.895245

#####Staten Island 266.205128

#####Manhattan 291.784595

- Most expensive neighborhood in airbnb listing are Fort Wadsworth with the average price of 800.000000 followed by Woodrow with 700.000000 and then Tribeca with 490.638418
- And the neighborhood with the cheapest price are Mount Eden wih the average price of 58.500000 followed by Concord 58.192308 and Grant City with 57.666667.
- Top three neighborhoods with the most number of booking are Williamsburg with 3920 followed by Bedford-Stuyvesant with 3714 and then Harlem with 2658.
- Neighborhood with the least number of booking are Rossville Richmondtown Willowbrook Fort Wadsworth New Dorp Woodrow all which have only one booking.
- Neighborhood groups according to there reviews Brooklyn 3065 Manhattan 2751 Queens 997 Bronx 187 Staten Island 81.
- Top earning hosts are sonder with amount of about 82,795.0, blueground with 70331.0 and sally with 37097.0 .