The Data Analyse of The New York City Airbnb Open Data

Yilin Wang

Introduction:

This dataset is a collection of listing activity and metrics of the Airbnb in NYC, NY in 2019. With the development of economics, more and more people choose to use Airbnb as a traveling experience. Personally, Airbnb is one of the main aspects of my traveling because it is the best experienced from near the local culture, and it brings us sweet feeling like home. It al provide a more convenient, comfortable and economical method of travel. The data file from public Kaggle Datasets shows all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

Overall goal is to answer the following questions

Note: since the neighborhood contains too much information and it is easy to confuse readers so that I decide to use neighbourhood_group which is the exact location for analysis.

1. Which are the top 10 neighbourhood_groups and neighbourhoods contians most airbnb?

2. What is the relationship between the location/area (neighbourhood_group/neighbourhood) and the Pageviews (overall reviews/ monthly reviews)?

3. What is the relationship between the room style and price base on each location (neighbourhood_group)?

Since Airbnb is one of the main aspects of my traveling, the price and location are the most important elements that I considered as I search the Airbnb so that I am very interested in using my analyzing to answer this question based on the topic.

Data Description:

- id: listing ID
- name: name of the Aribnb listing
- host_id: ID of the host
- host_name: name of the host
- neighbourhood_group: location
- neighbourhood: area
- latitude: latitude coordinates
- longitude: longitude coordinates
- room_type: listing space type of the aribnb
- price: price in dollars
- minimum_nights: amount of nights minimum
- number_of_reviews: how many reviews of the airbnb
- 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

Note that:

Since this dataset "AB_NYC_2019" is a summary information and metrics for listings in New York City in 2019 so that it contain with the limitations of integrality because it only contain all information that post online during that time period (August, 12th, 2019).

In [6]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sbn

sbn.set(font_scale = 1.5)
sbn.set_style('white')

%matplotlib inline

from scipy.stats import linregress, ttest_ind
```

In [7]:
```
data = pd.read_csv('AB_NYC_2019.csv')
data.head()
```

Out[7]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851

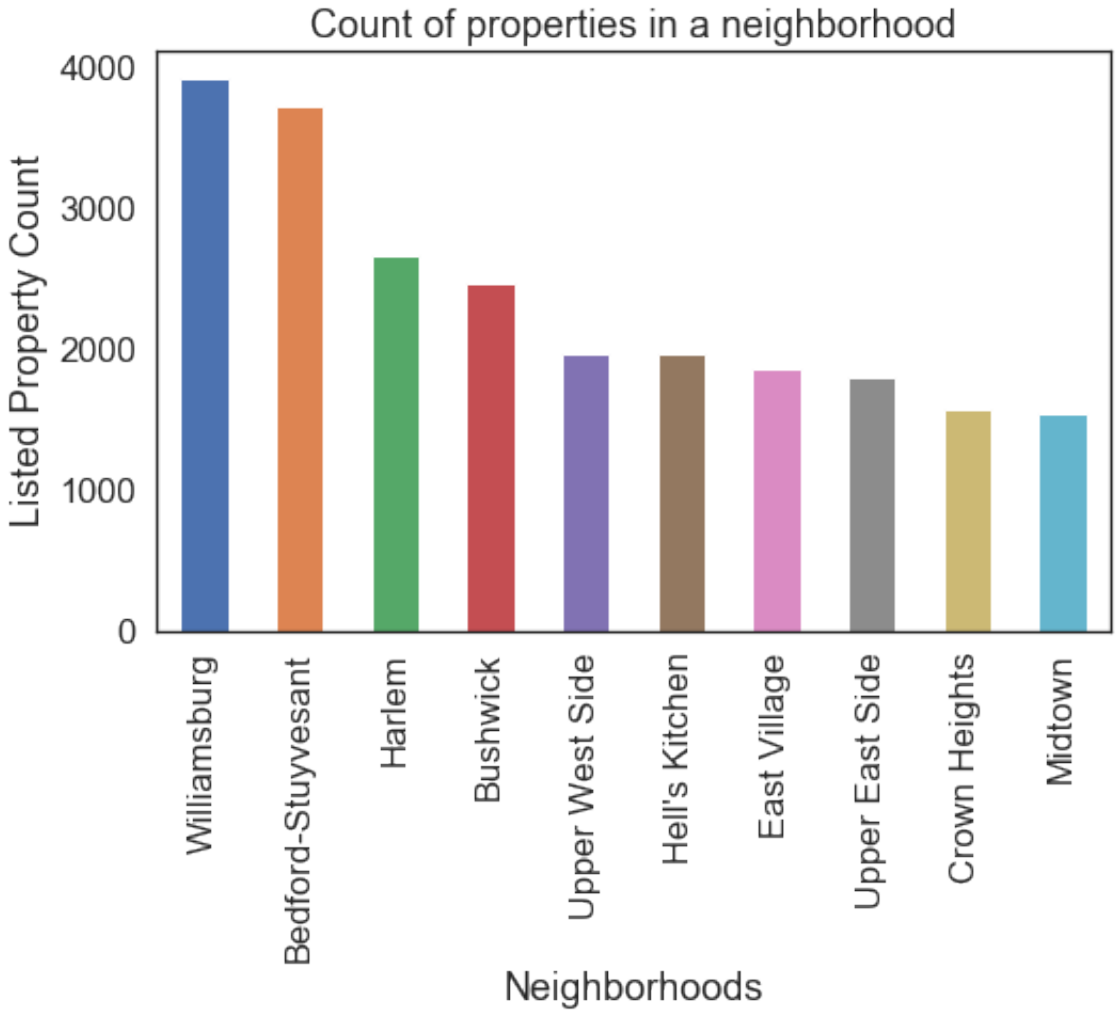
In [82]:
```
# count the properties value of neighborhood on the list
list_neighborhood = data.neighbourhood.value_counts()
print(num_neighborhood)

# obtain the top ten neighborhood
top_10_neighborhoods = data.neighbourhood.value_counts().head(10)
print(top_10_neighborhoods)
#plotting the top ten neighborhoods
plt.figure(figsize=(8,5))
top_10_neighborhoods.plot.bar()
plt.xlabel('Neighborhoods')
plt.ylabel('Listed Property Count')
plt.title('Count of properties in a neighborhood')
plt.show() #optional
```

```
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
East Harlem                     1117
```

```
Greenpoint                   1115
Chelsea                      1113
Lower East Side               911
Astoria                       900
Washington Heights            899
West Village                  768
Financial District            744
Flatbush                      621
Clinton Hill                  572
Long Island City              537
Prospect-Lefferts Gardens     535
Park Slope                    506
East Flatbush                 500
Fort Greene                   489
Murray Hill                   485
Kips Bay                      470
Flushing                      426
Ridgewood                     423
Greenwich Village             392
                             ...
Oakwood                         5
Little Neck                     5
New Brighton                    5
Emerson Hill                    5
Holliswood                      4
Todt Hill                       4
Olinville                       4
Mill Basin                      4
Castleton Corners               4
Spuyten Duyvil                  4
Prince's Bay                    4
Arden Heights                   4
Graniteville                    3
Eltingville                     3
Neponsit                        3
Breezy Point                    3
Huguenot                        3
Lighthouse Hill                 2
Silver Lake                     2
West Farms                      2
Westerleigh                     2
Bay Terrace, Staten Island      2
Howland Hook                    2
Co-op City                      2
Willowbrook                     1
Rossville                       1
Richmondtown                    1
Woodrow                         1
New Dorp                        1
Fort Wadsworth                  1
```

```
Name: neighbourhood, Length: 221, dtype: int64
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
```



What is the most popular Neighborhoods(contains more airbnb)?

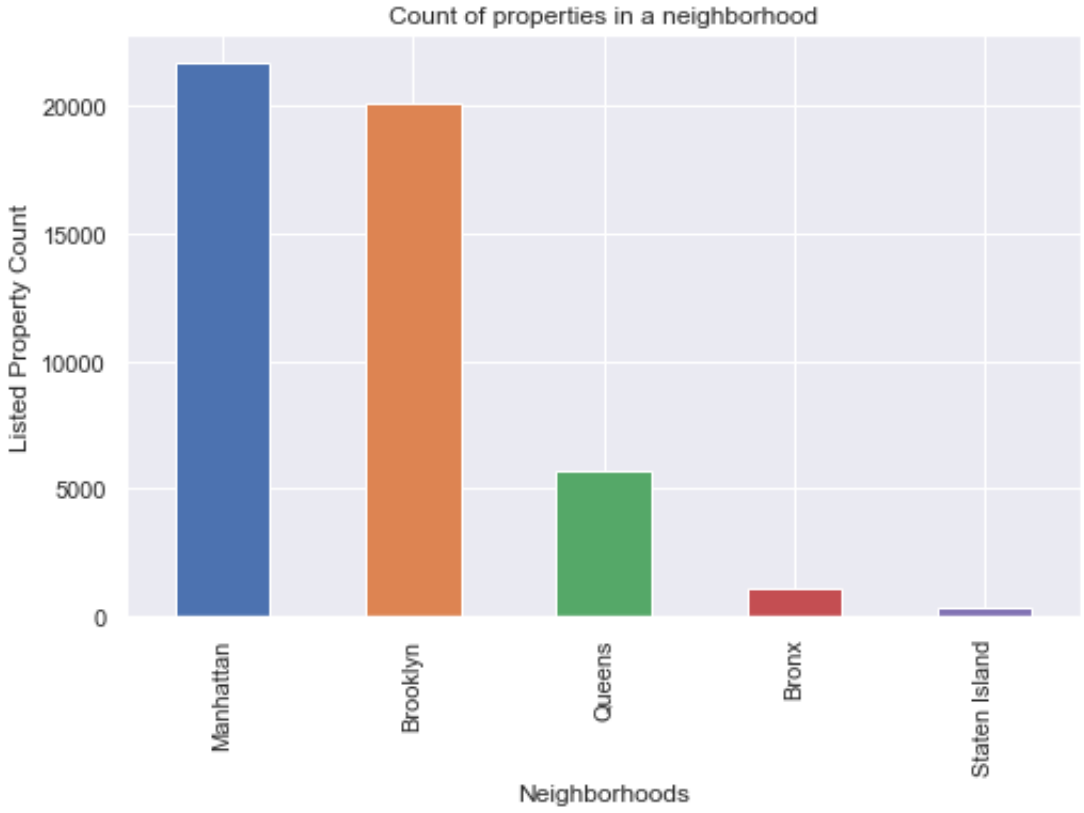
Figure 1

Williamsburg is the most popular Neighborhood. The pgraph shows the pupulation of airbnb in the listing of the Neighborhoods.

In [136]:
```
# count the properties value of neighbourhood_groups on the list
list_neighbourhood_group = data.neighbourhood_group.value_counts()
print(list_neighbourhood_group)

# since there are only five neighbourhood_groups on the list, just pri
nt all groups
#plotting the neighbourhood_groups
plt.figure(figsize=(8,5))
list_neighbourhood_group.plot.bar()
plt.xlabel('Neighborhoods')
plt.ylabel('Listed Property Count')
plt.title('Count of properties in a neighborhood')
plt.show() #optional
```

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



What is the most popular Neighborhoods_group/area(contains more airbnb)?

Figure 2

Manhattan is the most popular Neighborhood group. The pgraph shows the pupulation of airbnb in the listing of the Neighborhoods group.

In [42]:
```
# obtain the overall information of the reviews_per_month
data.loc[:,'reviews_per_month'].describe()
```

Out[42]:
```
count    38843.000000
mean         1.373221
std          1.680442
min          0.010000
25%          0.190000
50%          0.720000
75%          2.020000
max         58.500000
Name: reviews_per_month, dtype: float64
```

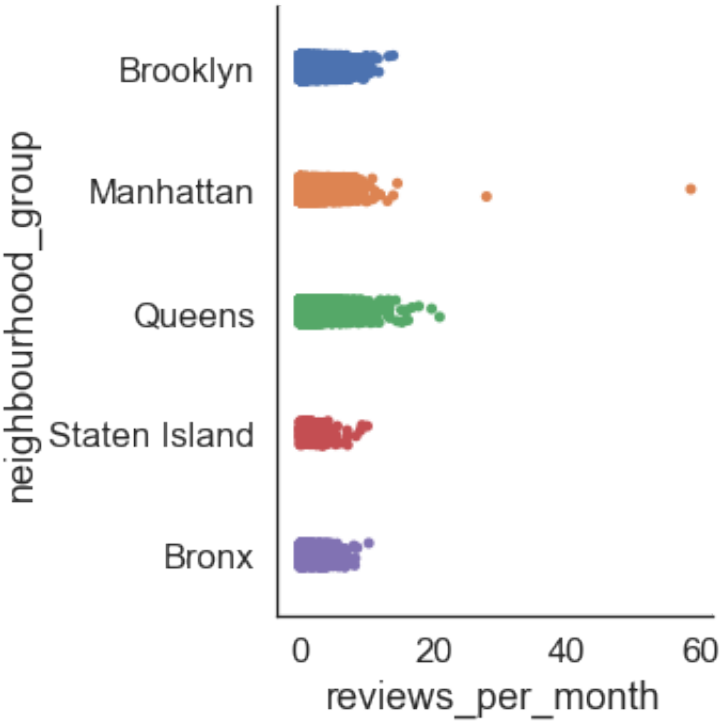
In [84]:
```
#
pd.pivot_table(data,
               index = 'neighbourhood_group',
               values = 'reviews_per_month',
               aggfunc = ['mean', 'std'])
```

Out[84]:

	mean	std
	reviews_per_month	reviews_per_month
neighbourhood_group		
Bronx	1.837831	1.673284
Brooklyn	1.283212	1.516259
Manhattan	1.272131	1.628252
Queens	1.941200	2.213108
Staten Island	1.872580	1.685495

In [119]:
```
sbn.catplot(data = data,
            x = 'reviews_per_month',
            y = 'neighbourhood_group',
            kind = 'strip')
```

Out[119]: <seaborn.axisgrid.FacetGrid at 0x1a222d7208>



Which Neighborhood group contains more reviews per month?

Figure 3

Queens contains more reviews per month. The pgraph shows reviews in each month in the listing of the Neighborhoods. Based on the figure, it shows Queens conatins the most reviews per month.

In [ ]:
```
# obtain the table about the relationship between neighbourhood and nu
mber_of_reviews
pd.pivot_table(data,
               index = 'neighbourhood',
               values = 'number_of_reviews',
               aggfunc = ['mean', 'std'])
```

In [105]:
```
# obtain the overall information of the number_of_reviews
data.loc[:,'number_of_reviews'].describe()
```

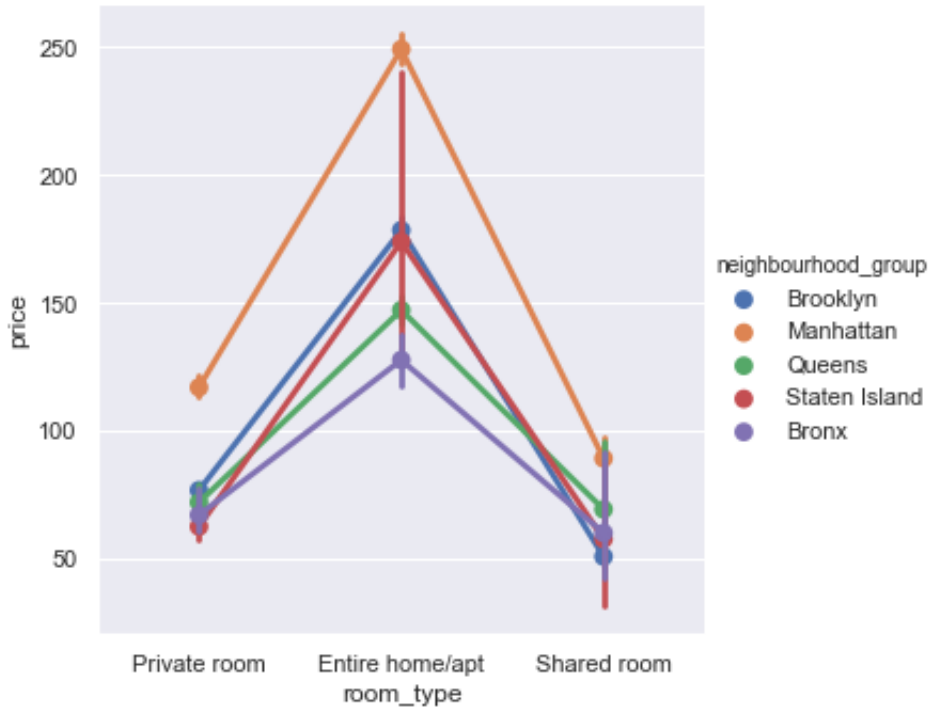
Out[105]:
```
count    48895.000000
mean        23.274466
std         44.550582
min          0.000000
25%          1.000000
50%          5.000000
75%         24.000000
max        629.000000
Name: number_of_reviews, dtype: float64
```

In [139]:
```
# obtain the max of reviews, and find its correspondent neighbourhood
and neighbourhood group
maxnumber_of_reviews = data['number_of_reviews'] == 629
area = data.loc[maxnumber_of_reviews, 'neighbourhood']
location = data.loc[maxnumber_of_reviews, 'neighbourhood_group']
print ('The neighbourhood contains most of reviews is {}.'. format(are
a))
print ('The neighbourhood group contains most of reviews per month is
{}.'. format(location))
```

```
The neighbourhood contains most of reviews is 11759    Jamaica
Name: neighbourhood, dtype: object.
The neighbourhood group contains most of reviews per month is 11759
Queens
Name: neighbourhood_group, dtype: object.
```

In [135]:
```
# creat a catplot that shows the relationship between the price and ro
om style base on
#---each neighbourhood_group
sbn.catplot(data = data,
            hue = 'neighbourhood_group',
            y = 'price',
            x = 'room_type',
            kind = 'point')
```

Out[135]: <seaborn.axisgrid.FacetGrid at 0x1a2dad1940>



Which room_style is more expensive, and where it locate (Neighborhood group)?

Figure 4

The entire home is the most expensive room style with the location in Manhattan The plot shows relationship between the the price and room style base on each neighbourhood_group. Based on the figure, it shows the entire home which is located in Manhattan is most expensive which is make sence because it is the most dwelling location in NY.

In [133]:
```
# creat a catplot that shows the relationship between each elemnt in t
he list base on
#---each neighbourhood_group
sbn.pairplot(data, hue = 'neighbourhood_group')
```

```
/Users/Genie/anaconda3/lib/python3.7/site-packages/statsmodels/nonpa
rametric/kde.py:448: RuntimeWarning: invalid value encountered in gr
eater
  X = X[np.logical_and(X > clip[0], X < clip[1])] # won't work for t
wo columns.
/Users/Genie/anaconda3/lib/python3.7/site-packages/statsmodels/nonpa
rametric/kde.py:448: RuntimeWarning: invalid value encountered in le
ss
  X = X[np.logical_and(X > clip[0], X < clip[1])] # won't work for t
wo columns.
```

Out[133]: <seaborn.axisgrid.PairGrid at 0x1a24eaaeb8>

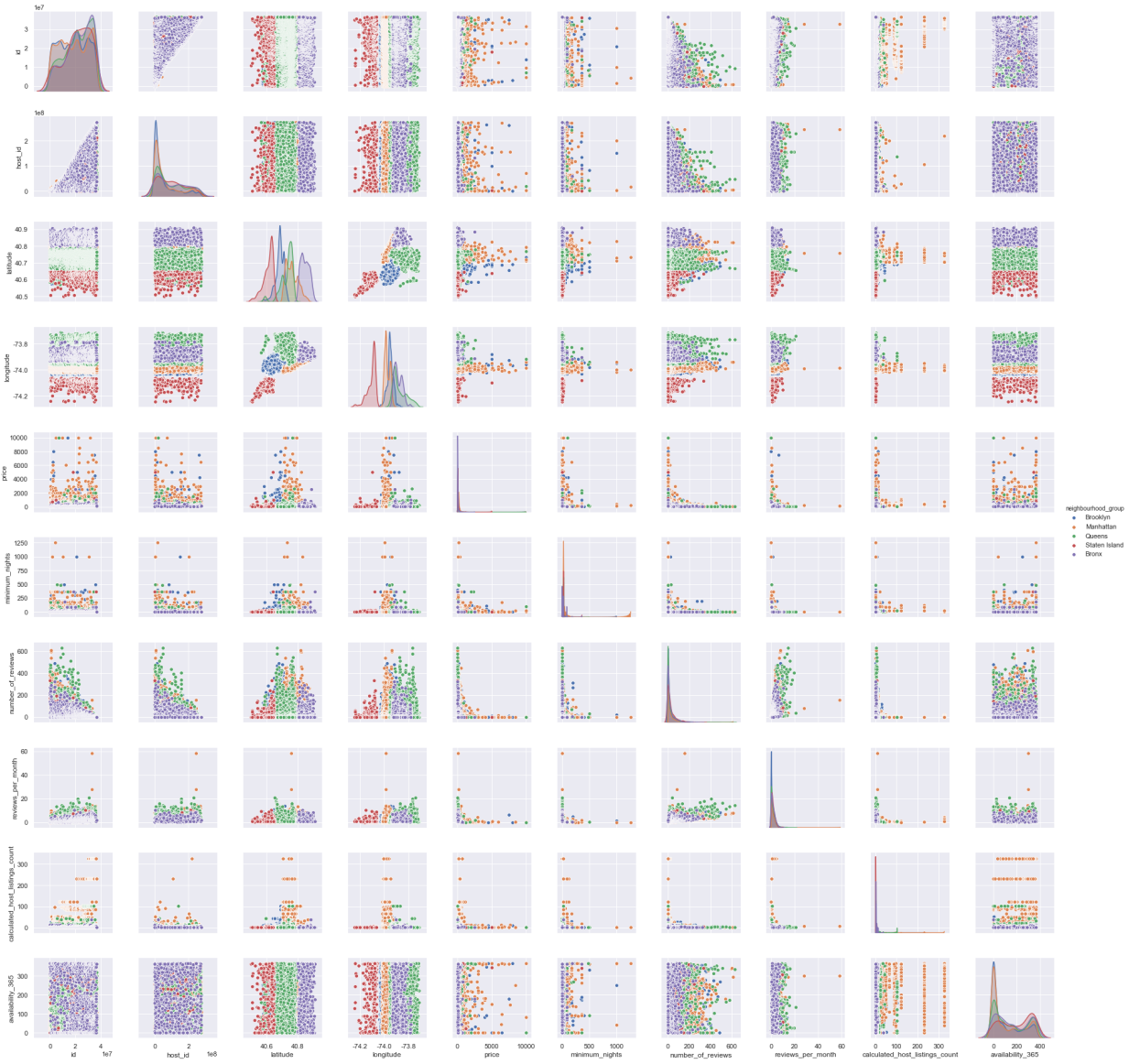


Figure 5

The plot shows the relations between each element based on neighbourhood_group so that the user can base on their unique requirement to find the figure for searching the information.

In [ ]: