THE SECRETS OF AIRBNB IN NYC:

DATA INSIGHTS

By:
Shashwat bhansali

AGENDA

1) Objective

2) Data Life cycle

3) Analysis Method

4) Recommendations

5) Appendix:

a) Data Sources.
b) Data methodology.
c) Data model Assumptions.

OBJECTIVE

1)Conduct analysis of New York AIRBNB data set.

2)Ask effective questions that can lead to data insights.

3)Data visualization and statistical techniques.

DATA LIFE CYCLE

1)In the first phase data is captured and loaded in to the environment.

2)Once data is cleaned, EDA is done and new features are created.

3)Then meaningful insights are derived using various analytical methods.

Importing the libraries and reading the data

In [5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [6]:
```
df = pd.read_csv('AB_NYC_2019.csv')
df.head()
```

Out[6]:

	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	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

Data Wrangling

```
In [7]: df.shape
```

```
Out[7]: (48895, 16)
```

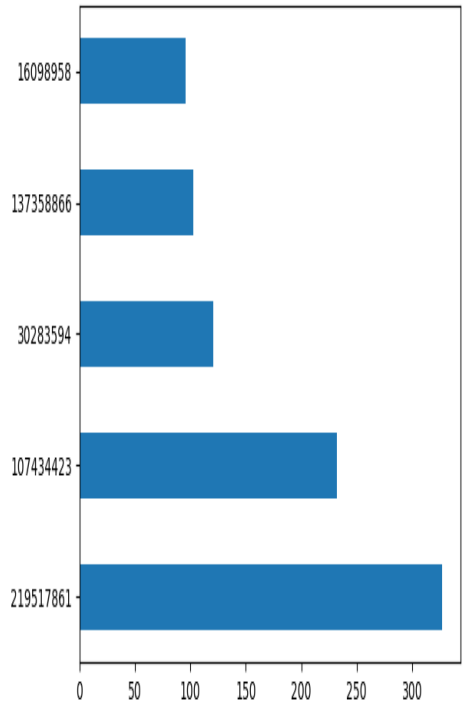
```
In [8]: df.info()
```

```
<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
```

Analysis

```
In [14]: df.host_id.value_counts().iloc[:5].plot(kind = 'barh')
```

```
Out[14]: <Axes: >
```



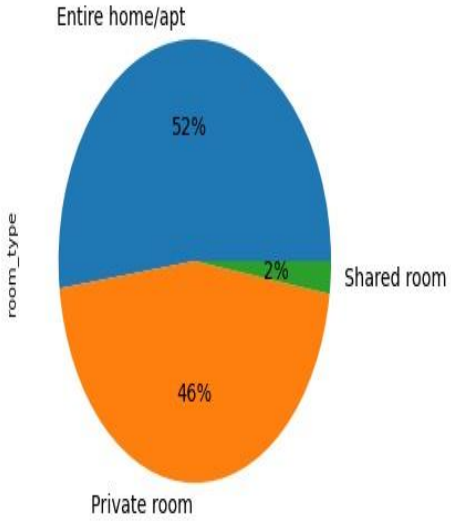
We can observe that the highest times transaction done by a customer is 327 in the year 2019.

```
In [15]: df['room_type'].value_counts()
```

```
Out[15]: Entire home/apt    25409
         Private room       22326
         Shared room         1160
         Name: room_type, dtype: int64
```

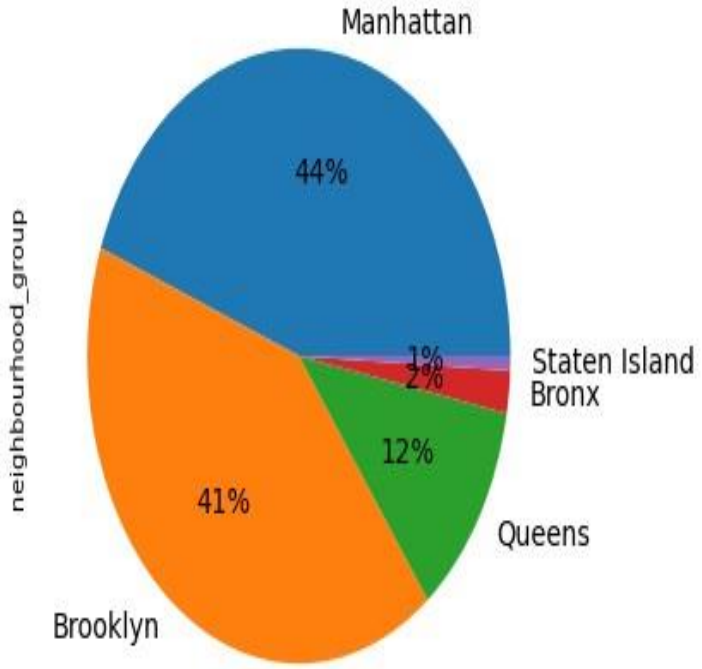
```
In [16]: fig = plt.figure(figsize=(5,5), dpi=80)
         df['room_type'].value_counts().plot(kind='pie',  autopct='%1.0f%%', startangle=360, fontsize=13)
```

```
Out[16]: <Axes: ylabel='room_type'>
```

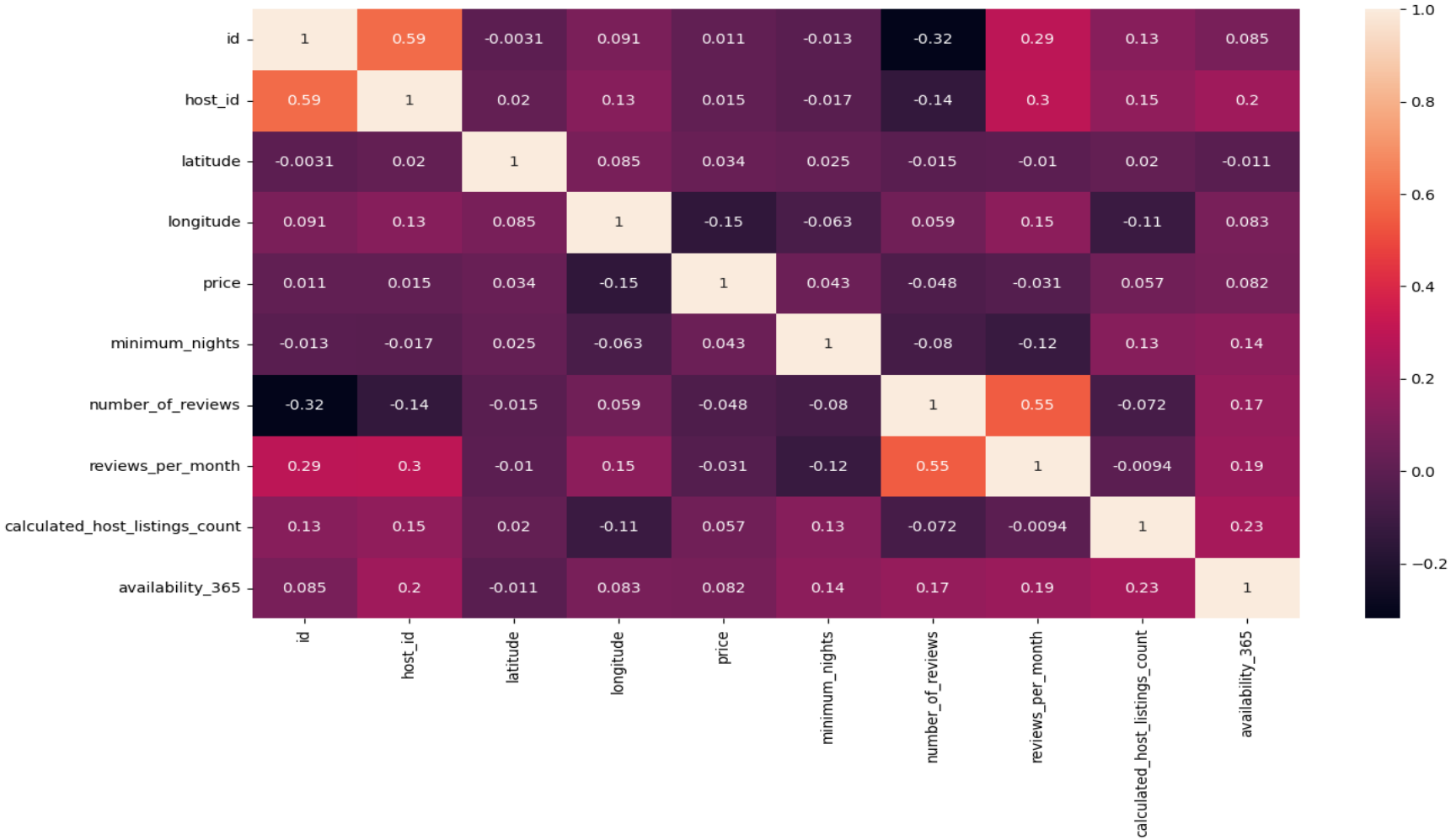


Most contributing Neighborhoods

1) Staten Island has the lower contribution.
2) 85 % of the contribution are from Manhattan and Brooklyn neighbourhood group.



Bivariate and Multi variate Analysis



Conclusion

1) Strong significant insights are derived based on various attributes on dataset.

2)Ample amount and variety of visuals can be used in the presentation in front of the stake holder.

3)Data collection team should collect data about review scores, so that it can strengthen the later analysis.

Appendix Data Sources

The columns in the dataset are self-explanatory. You can refer to 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

1)Conduct a thorough analysis of New York AIRBNB dataset.

2) Cleaned the dataset using the python.

3) Derived the necessary features.

Appendix Data Assumptions

```
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
```