UNVEILING THE SECRETS OF AIRBNB IN NYC
DATA ANALYST POV

Dipjit Basak
Darsi Noel Jeffrey

BA || DS60
Upgrad IIIT Bangalore Executive PG Program in Data Science

Objective

- Process, Analyze & Share findings about the NYC Airbnb data, which can lead to data insights using data visualization & statistical techniques.

Data Life Cycle

Data Capture & Data Cleaning

Exploratory Data Analysis & feature Addition

Leading to Insights

Importing Library

In [1]:
```
# Pandas & Numpy
import pandas as pd
import numpy as np

# Plotting libraries - Matplotlib, Seaborn & Plotly
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.graph_objs as go
import plotly
from plotly import tools
import plotly.express as px
from wordcloud import WordCloud
# Filtering out warnings
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
# pip install wordcloud
```

In [3]:
```
inp = pd.read_csv('Downloads/AB_NYC_2019.csv')
inp.head(10)
```

Out[3]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_c
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	
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	

Data Type

- There are different data types: Categorical, Numerical etc.

4.1 Categorical

```
In [21]: inp0.columns
```

```
Out[21]: 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', 'availability_365_categories',
       'minimum_night_categories', 'number_of_reviews_categories',
       'price_categories'],
      dtype='object')
```

```
In [22]: # Categorical nominal
categorical_columns = inp0.columns[[0,1,3,4,5,8,16,17,18,19]]
categorical_columns
```

```
Out[22]: Index(['id', 'name', 'host_name', 'neighbourhood_group', 'neighbourhood',
       'room_type', 'availability_365_categories', 'minimum_night_categories',
       'number_of_reviews_categories', 'price_categories'],
      dtype='object')
```

```
In [23]: # To see the first few rows of categorical columns
inp0[categorical_columns].head()
```

Out[23]:

	id	name	host_name	neighbourhood_group	neighbourhood	room_type	availability_365_categories	minimum_night_categories	number_
0	2539	Clean & quiet apt home by the park	John	Brooklyn	Kensington	Private room	very High	very Low	
1	2595	Skylit Midtown Castle	Jennifer	Manhattan	Midtown	Entire home/apt	very High	very Low	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	Elisabeth	Manhattan	Harlem	Private room	very High	Low	
3	3831	Cozy Entire Floor of Brownstone	LisaRoxanne	Brooklyn	Clinton Hill	Entire home/apt	Medium	very Low	
4	5022	Entire Apt: Spacious Studio/Loft by central park	Laura	Manhattan	East Harlem	Entire home/apt	very Low	very High	

4.2 Numerical

```
In [24]: numerical_columns = inp0.columns[[9,10,11,13,14,15]]
numerical_columns
```

```
Out[24]: Index(['price', 'minimum_nights', 'number_of_reviews', 'reviews_per_month',
       'calculated_host_listings_count', 'availability_365'],
      dtype='object')
```

```
In [25]: inp0[numerical_columns].head()
```

Out[25]:

	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
0	149	1	9	0.21	6	365
1	225	1	45	0.38	2	355
2	150	3	0	NaN	1	365
3	89	1	270	4.64	1	194
4	80	10	9	0.10	1	0

```
In [26]: inp0[numerical_columns].describe()
```

Out[26]:

	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Analysis: Missing Value

- "last_review & reviews_per_month" have highest missing value %.

Function to check missing values

In [9]:
```
def prcnt_missing(df):
    return round(df.isnull().sum()/len(df)*100,2)

prcnt_missing(inp)
```

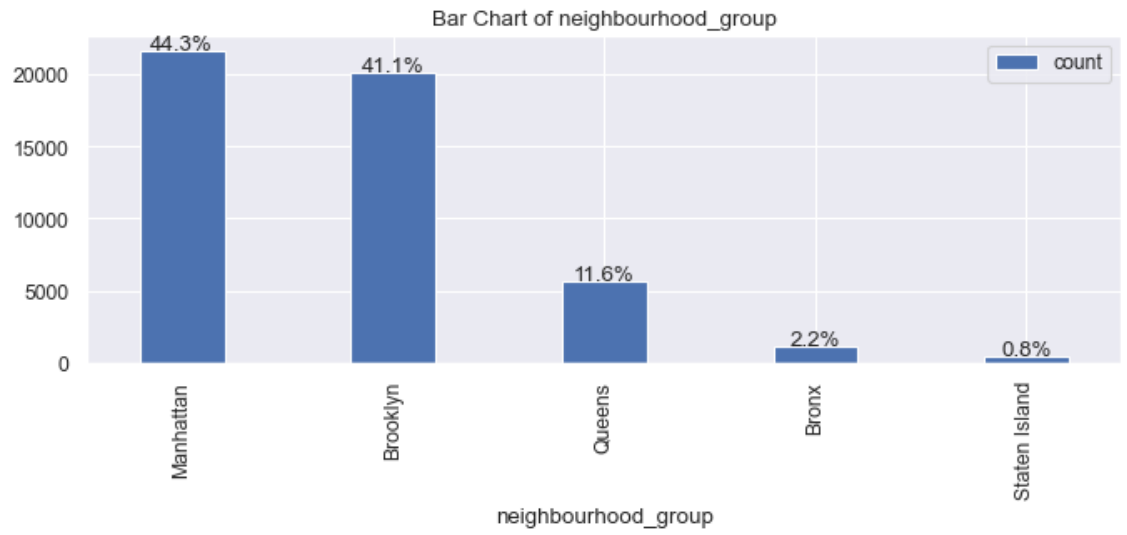
Out[9]:
```
id                                0.00
name                              0.03
host_id                           0.00
host_name                         0.04
neighbourhood_group               0.00
neighbourhood                     0.00
latitude                          0.00
longitude                         0.00
room_type                         0.00
price                             0.00
minimum_nights                    0.00
number_of_reviews                 0.00
last_review                      20.56
reviews_per_month                20.56
calculated_host_listings_count    0.00
availability_365                  0.00
dtype: float64
```

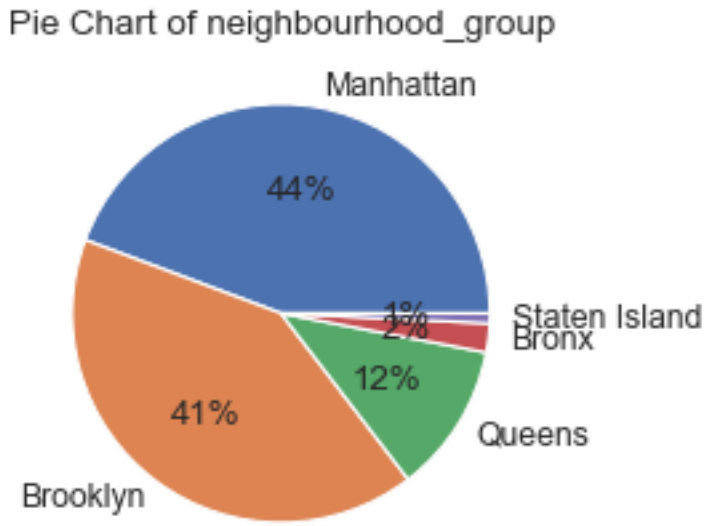
In [10]:
```
# Columns with missing values
print(prcnt_missing(inp)[prcnt_missing(inp) > 0])
```

```
name                 0.03
host_name            0.04
last_review         20.56
reviews_per_month   20.56
dtype: float64
```

Analysis: Neighborhood Group

- 85% of the listings are “Manhattan & Brooklyn” neighborhood group

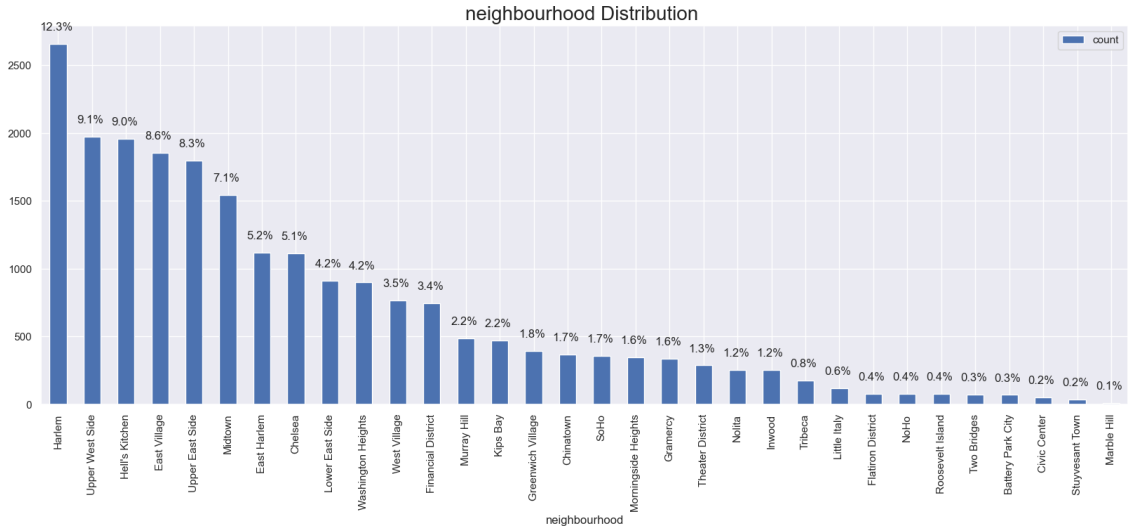


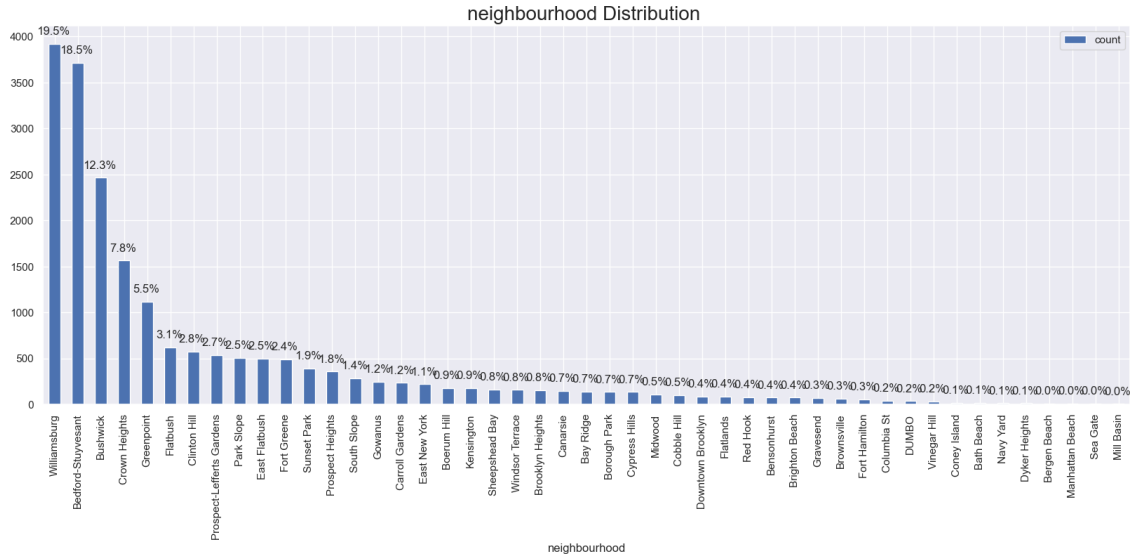


Analysis: Neighborhood Distribution

- “Manhattan

Brooklyn

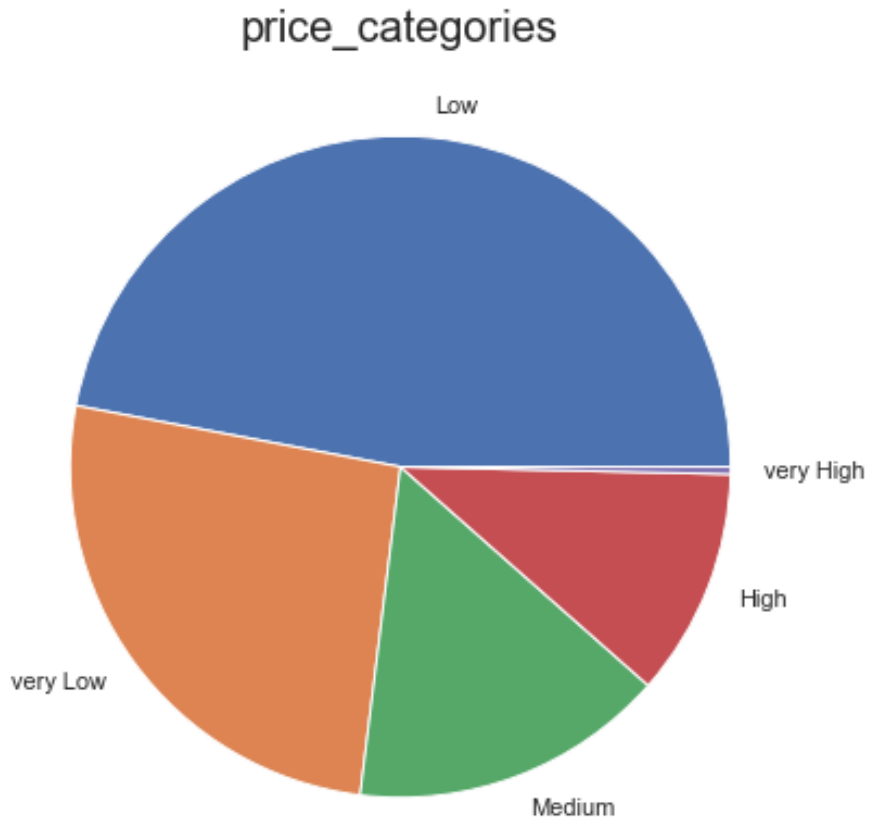


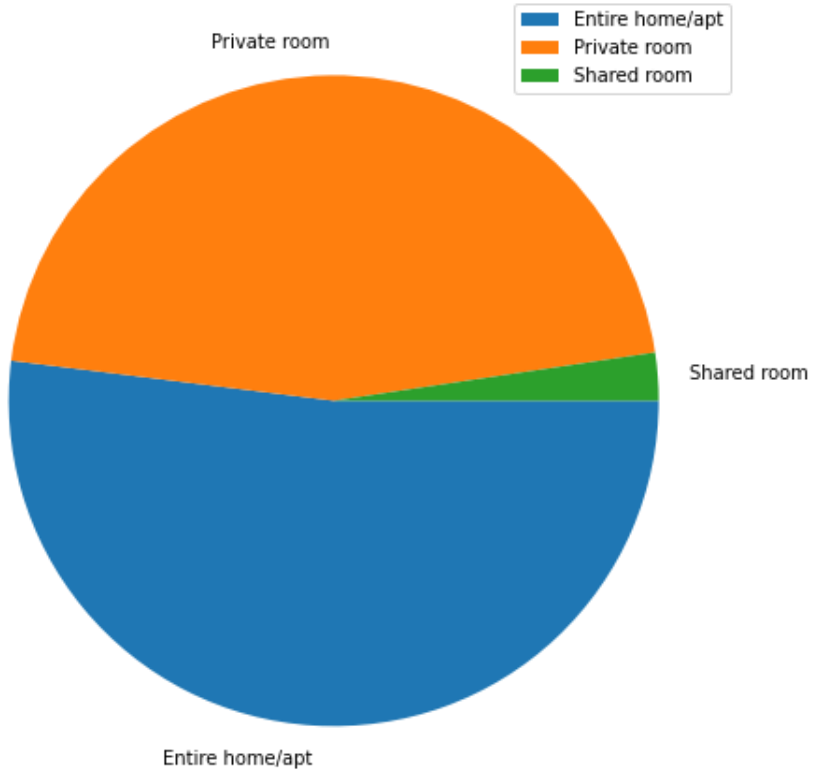


Harlem in Manhattan & Williamsburg in Brooklyn has the highest booking rate.

Analysis: Price & Room Type

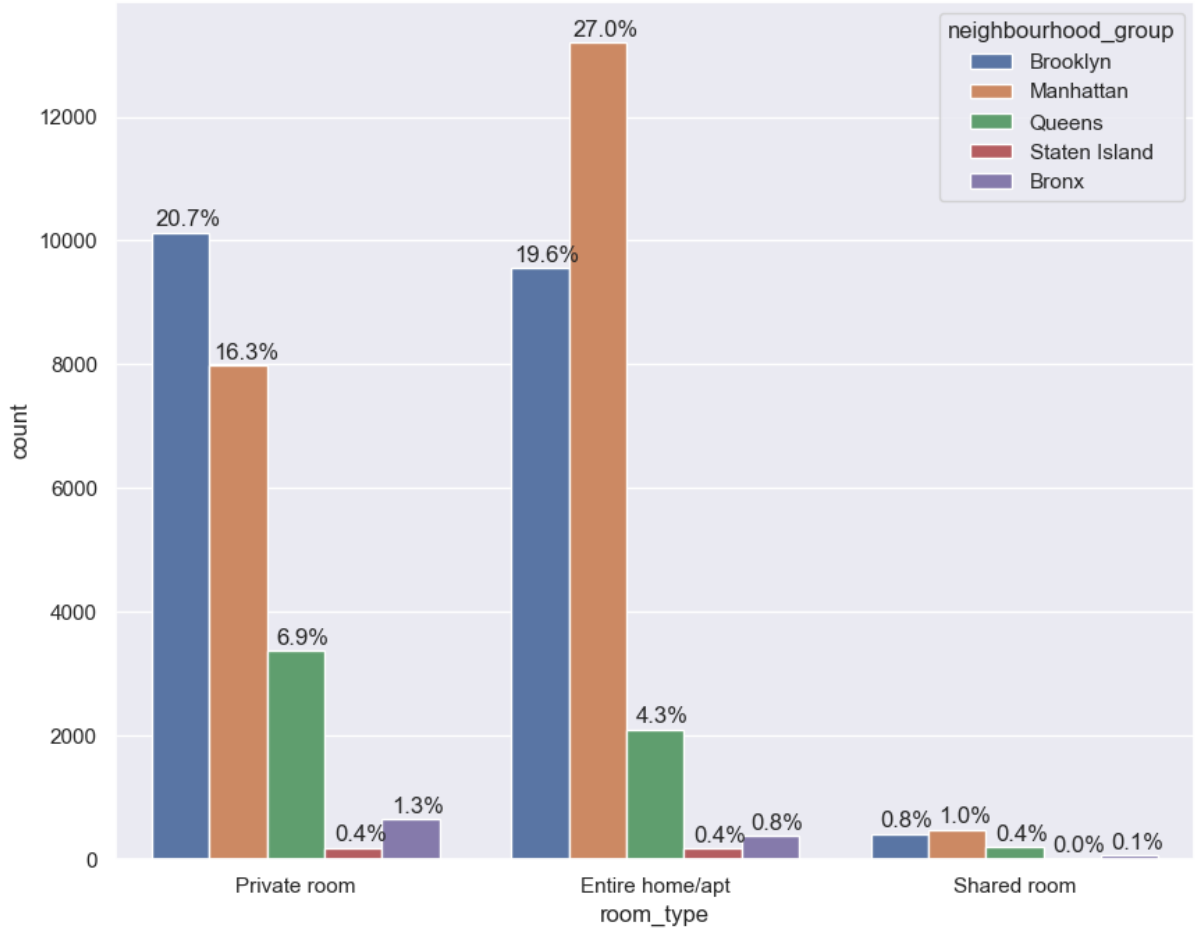
- "Low" is the most preferred "Price category".
- Least preferred "room type category" is "Shared room"



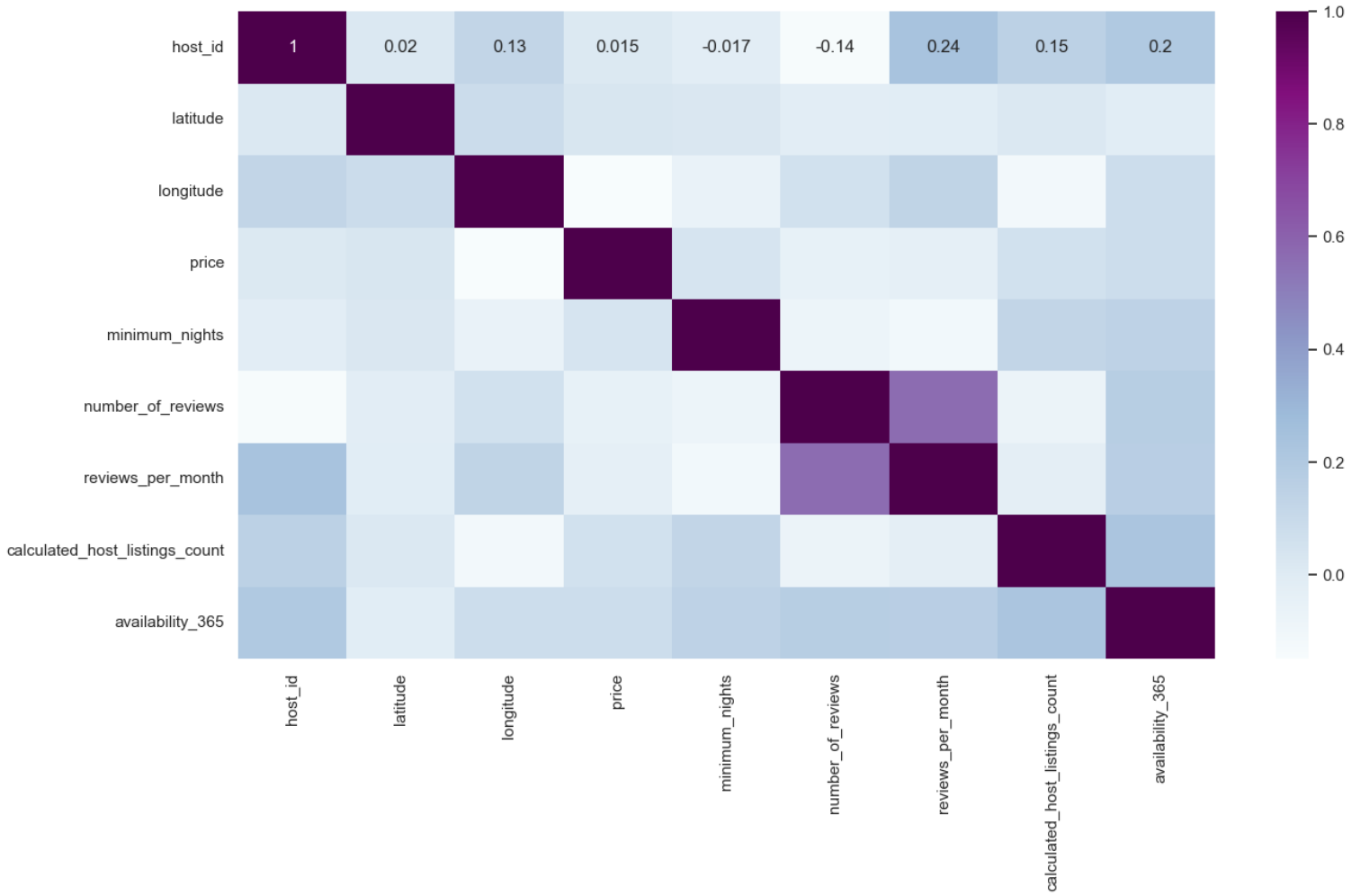


Analysis: Room Type

- In Brooklyn, Private Room booking is preferred
- In Manhattan, Entire home/apt is preferred



Uni & Bi-Variate Analysis



Conclusion

- In this Airbnb case study, 85% of the listings are “Manhattan & Brooklyn” neighborhood group
- “Low price” is the most preferred price category.
- In Brooklyn, Private Room booking is preferred where as in Manhattan, Entire home/apt is preferred.
- Least preferred room type category is “Shared room”

Thank You