Capstone Project- 1

Airbnb Bookings Analysis

by-

Team Data Avengers

Rahul Chavan

Sagar Sanap

Rutuja Ahire

Rishu Gupta

Points of Discussion

- Dataset Overview
- Problem Statements
- Features in Dataset
- Importing Libraries and Data Cleaning
- Data Exploration
- Hosts and Areas
- Neighbourhood Group Price Distribution
- Busiest Hosts by Reviews
- Room Type Preferred
- Neighbourhood Popularity
- Conclusion

Dataset Overview

- Airbnb, as in “Air Bed and Breakfast”, is a service that lets property owners rent out their spaces to travelers looking for a place to stay. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves.

- The dataset Airbnb NYC 2019 that we are analyzing consists of the booking data on Airbnb from 2008 to 2019.

Problem Statements

- What can we learn about different hosts and areas?
- What can we learn from predictions(prices, reviews,etc.)
- Which hosts are busiest and why?
- Which room type is preferred in most popular neighbourhood?
- Is there any noticeable difference of traffic among different areas, what could be the reason for it?

Features in dataset:

The features in the dataset can be described as follows:

- id - It is the identity number of the property listed by a particular host.
- name - It is the name of the property listed by the host.
- host_id - It is the identity number of the hosts
- host_name – It is the name of the host
- neighbourhood_group - It is the name of the neighbourhood groups in NYC
- neighbourhood – It is the name of the neighbourhood in NYC.
- latitude – It is the coordinates of latitude of the property listed.
- longitude – It is the coordinates of longitude of the property listed.

- room type – It is the type of room listed by host.
- price - It is the rent of the property listed in US Dollars.
- minimum nights – It is minimum nights customer rented the property.
- Number_of_reviews – It is the number of customers reviews on the property.
- last_review – It is the date when the property was last reviewed.
- reviews_per_month - It is the count of reviews per month on the property
- calculated_host_listings_count - It is the number of listings done by a host
- Availability_365 – It is the number of days the property is available in a year

Importing Libraries and Data Cleaning

Importing the required libraries.

```
[1]  import pandas as pd
     import numpy as np
     import matplotlib
     import matplotlib.pyplot as plt
     import seaborn as sns
```

The dataset has of 48895 rows and 16 features.

```
[4]  airbnb_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
```

Dropping non-relevant features from the dataset.

```
[5]  airbnb_df.drop(['name','last_review'],axis=1,inplace=True)
     airbnb_df.head()
```

Replacing NaN values from features.

```
[6]  airbnb_df.host_name.fillna('Unavailable',inplace=True)
     airbnb_df.reviews_per_month.fillna(0,inplace=True)
     airbnb_df.isnull().sum()
```

Updated dataset has no NaN values.

```
id                                0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

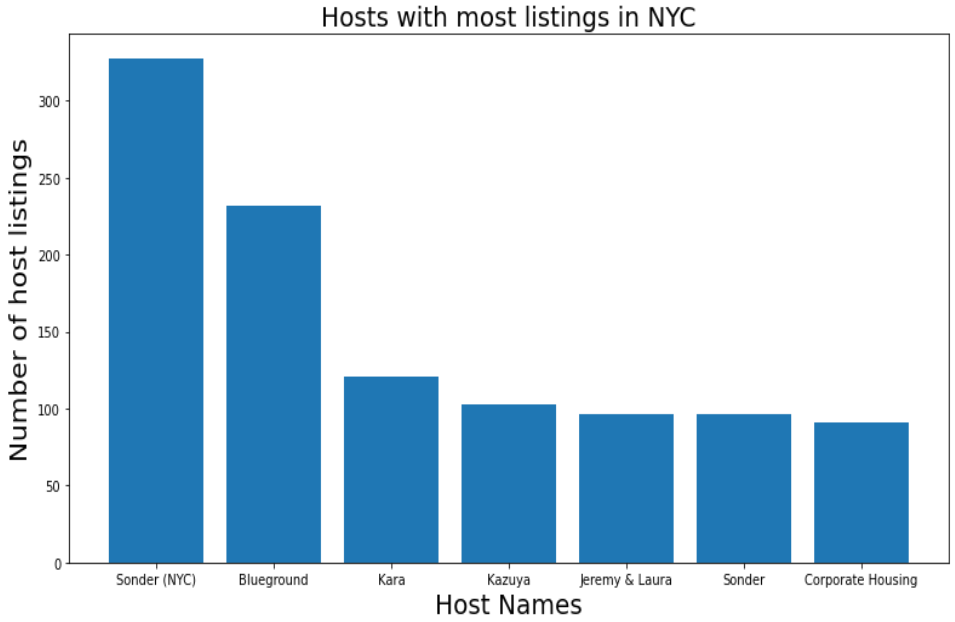
Hosts and Areas

Top 10 hosts across different neighbourhood groups

	host_name	neighbourhood_group	calculated_host_listings_count
13217	Sonder (NYC)	Manhattan	327
1833	Blueground	Brooklyn	232
1834	Blueground	Manhattan	232
7275	Kara	Manhattan	121
7480	Kazuya	Queens	103
7479	Kazuya	Manhattan	103
7478	Kazuya	Brooklyn	103
6540	Jeremy & Laura	Manhattan	96
13216	Sonder	Manhattan	96
2901	Corporate Housing	Manhattan	91

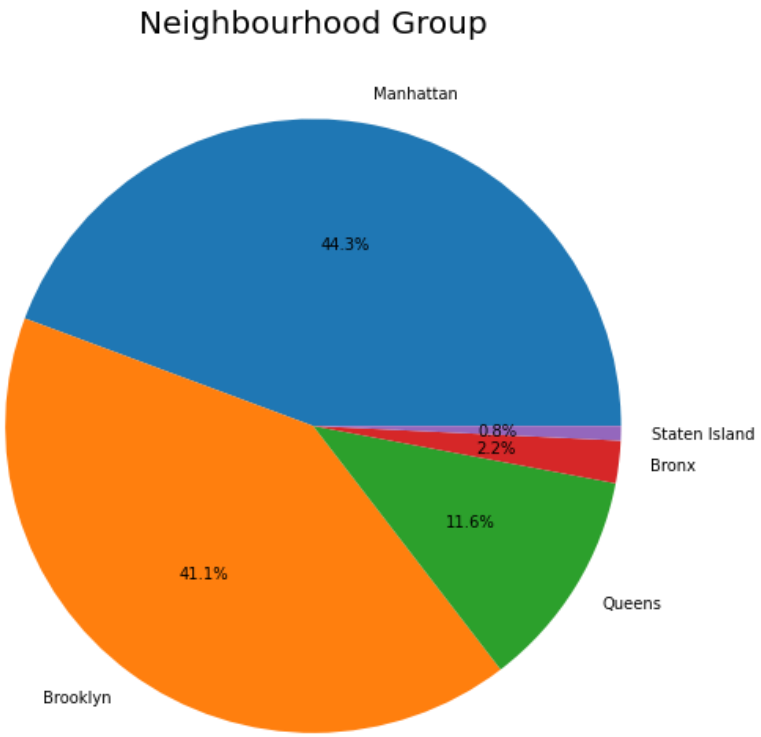
Hosts with most listings

- The host named Sonder(NYC) has highest number of listings of 327 in Manhattan neighbourhood group.
- The host named Blueground has 2nd highest listings of 232 in Manhattan Neighbourhood group.
- The host Blueground also has 232 listings in Brooklyn.



Neighbourhood Group

- The pie chart shows Manhattan has 44.3% of the total listings which is the highest share.
- Its followed by Brookyln with a share of 41.1% of total listings.
- This means just Manhattan and Brooklyn have about 85% of the total NYC listings which is a lot.
- Staten Island has the lowest number of listings in NYC with a share of only 0.8%



Dataset Statistics

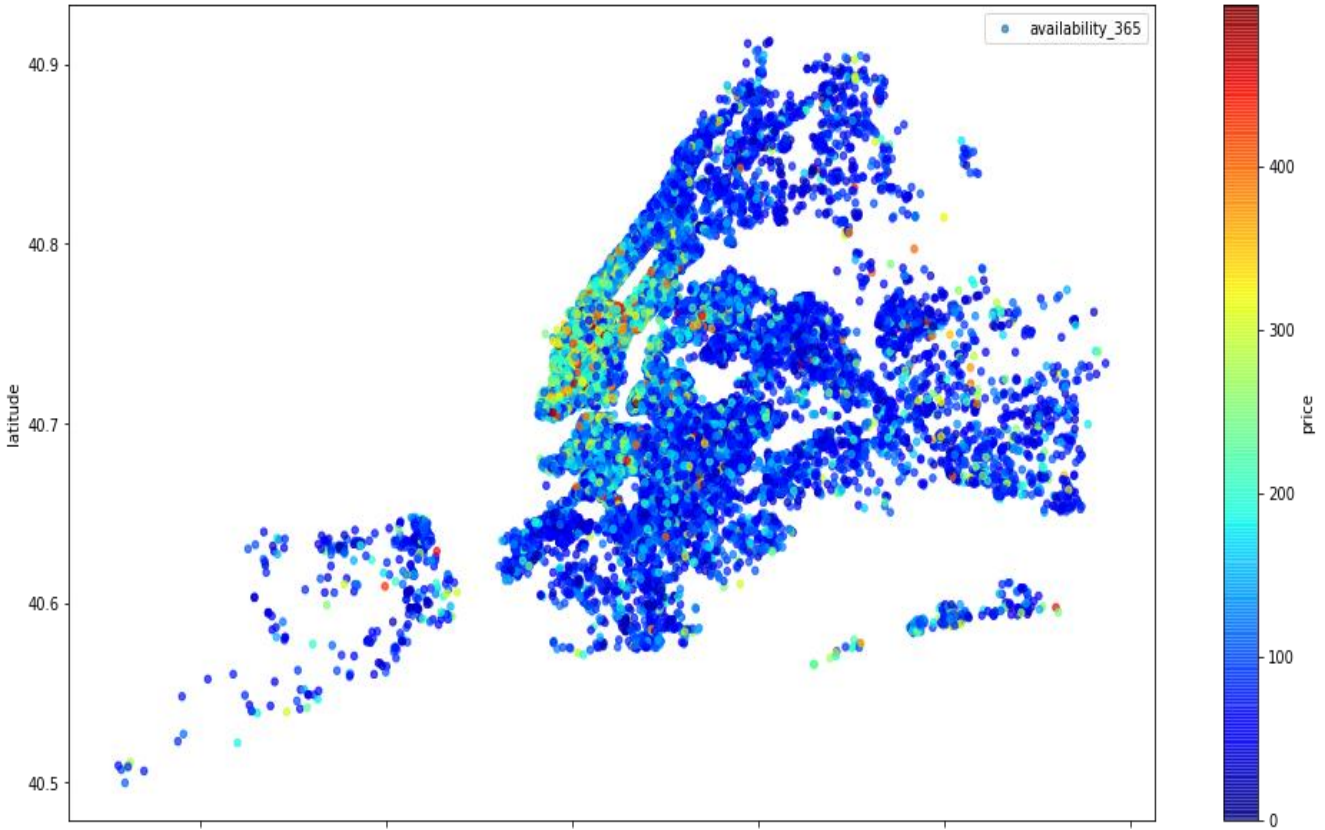
From the statistics we can see that maximum rental price is $10000 which is absurdly high and most probably skewed. Hence, we have to limit the price range to [price<500] in the plots to get an accurate display of the price ranges in different neighbourhoods.

```
[10] airbnb_df.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.090910	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.597283	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.000000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.040000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.370000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	1.580000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

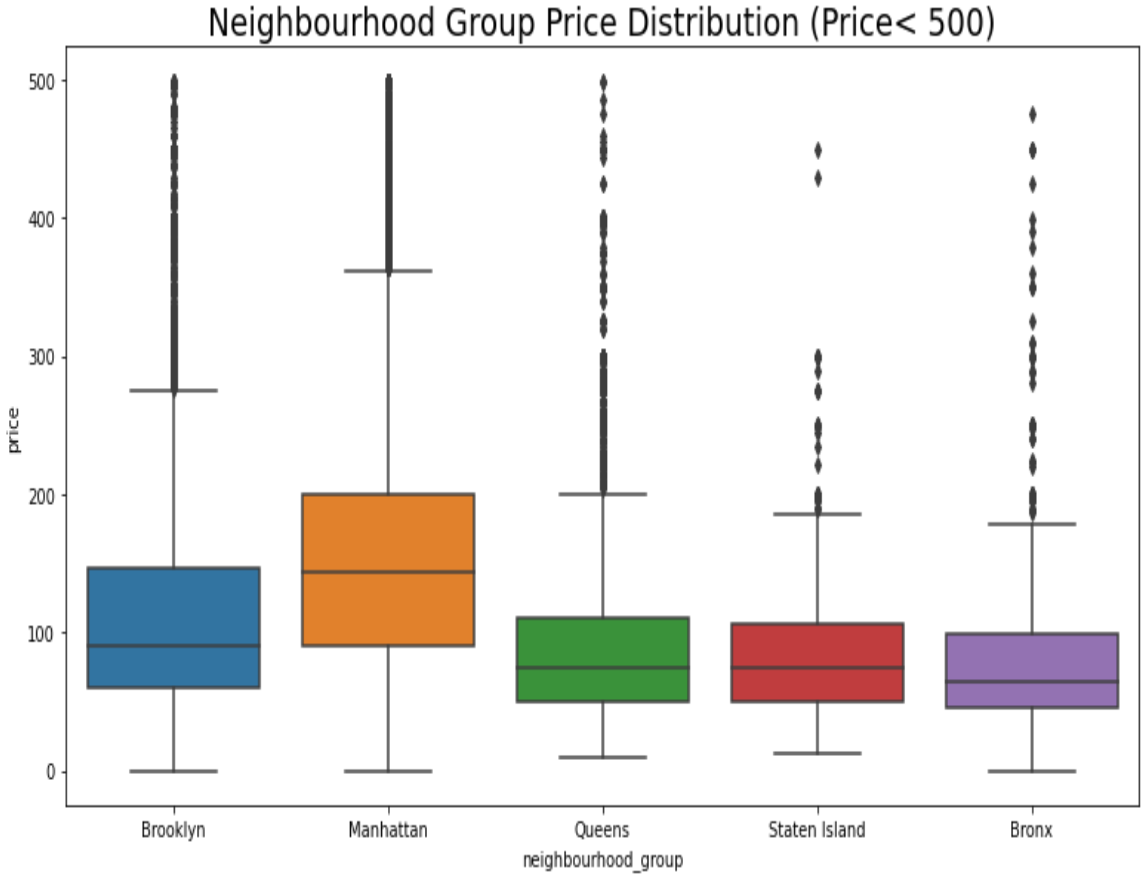
Neighbourhood Group Price Distribution

- Red color dots are the rooms with a higher price and blue color dots are the rooms with a lower price.
- Manhattan has the largest variation in price range.
- Manhattan has the most number of green, yellow and red dots which means it has more expensive rooms than other neighbourhood groups



Price distribution further illustrated.

- Manhattan has the highest range price for the listings with about 140 USD as an average price, followed by Brooklyn with 90 USD.
- Queens and Staten Island have a very similar range price.
- Bronx has the lowest average price.

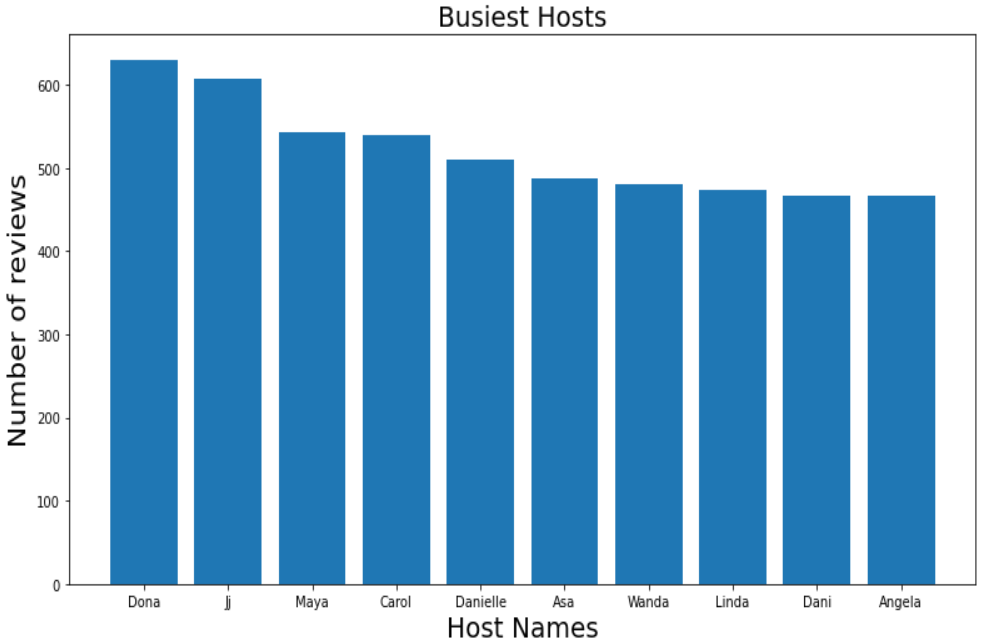


Top 10 busiest hosts by reviews

	host_name	host_id	room_type	neighbourhood_group	number_of_reviews
10310	Dona	47621202	Private room	Queens	629
17755	Jj	4734398	Private room	Manhattan	607
25626	Maya	37312959	Private room	Queens	543
6259	Carol	2369681	Private room	Manhattan	540
8973	Danielle	26432133	Private room	Queens	510
3966	Asa	12949460	Entire home/apt	Brooklyn	488
37848	Wanda	792159	Private room	Brooklyn	480
22556	Linda	2680820	Private room	Queens	474
8651	Dani	42273	Entire home/apt	Brooklyn	467
2953	Angela	23591164	Private room	Queens	466

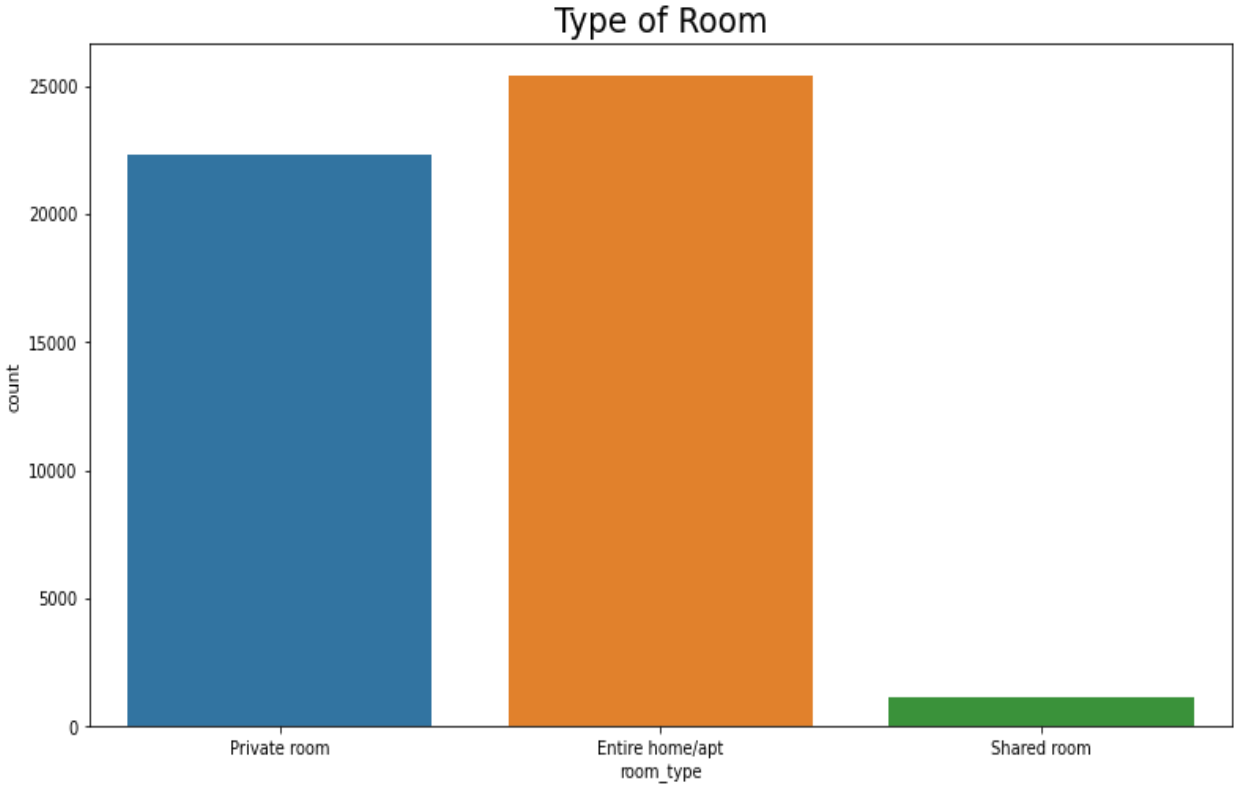
Busiest Hosts by Reviews

- Dona has highest numbers of reviews and we can assume that Dona is the busiest host followed by Jj and Maya
- The chart also shows that the busiest neighbourhoods groups are Manhattan, Queens and Brooklyn.



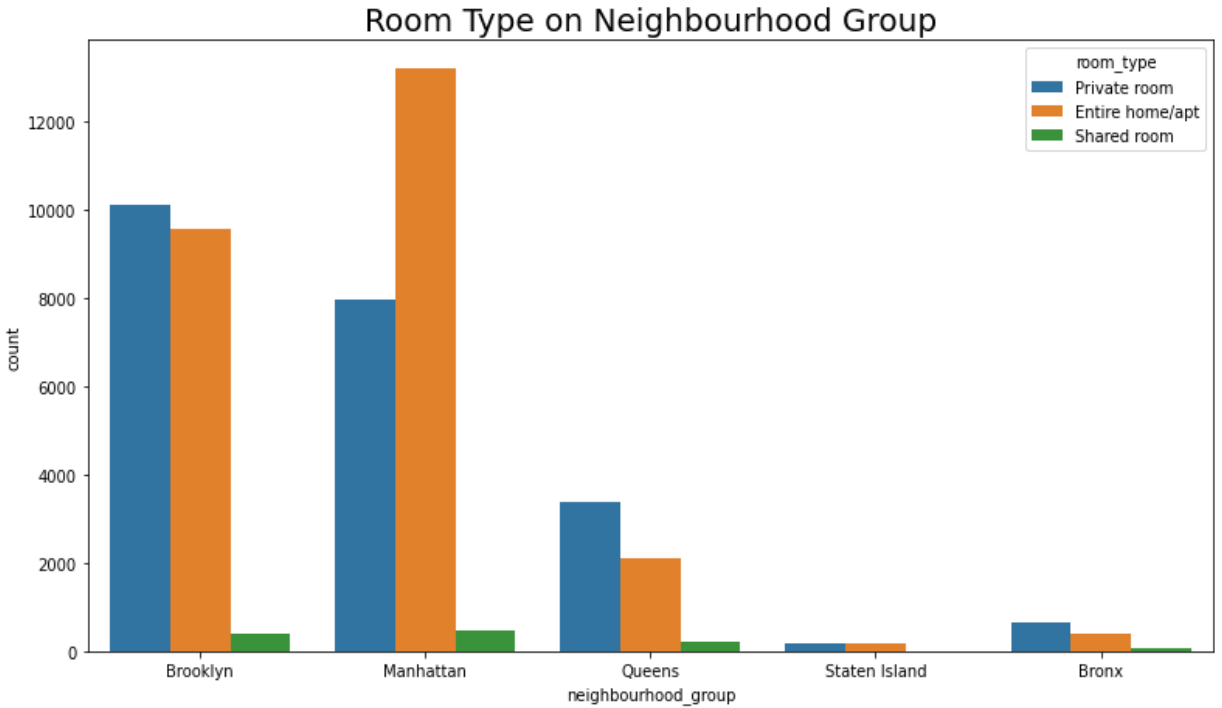
Room Type Preferred

- The most preferred room type is Entire home/apt as well as private room.
- Shared room is least preferred by people.



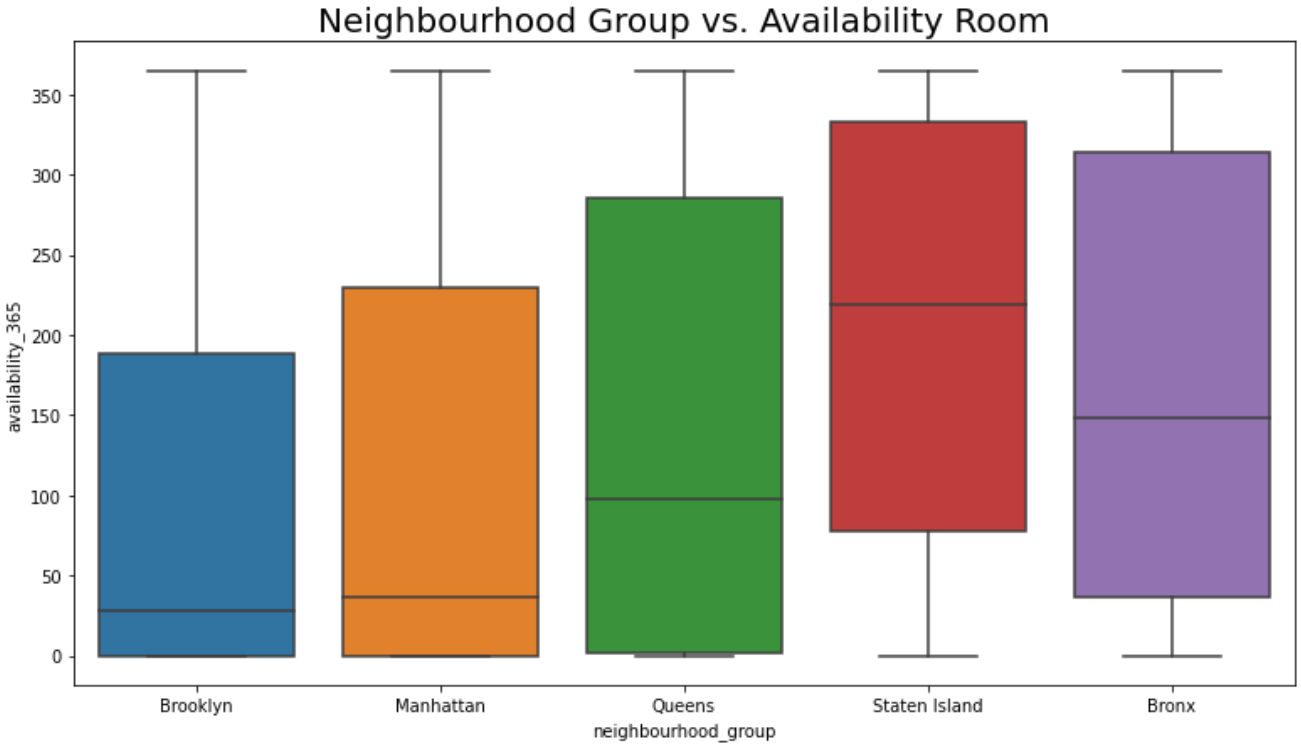
Room Type on Neighbourhood Group

- Brooklyn has almost the same number of private rooms and entire home/apartments
- Manhattan has significantly more entire home/apartments than private rooms
- Queens has somewhat more private rooms than entire home/apartments



Neighbourhood Popularity

Neighbourhood group room availability

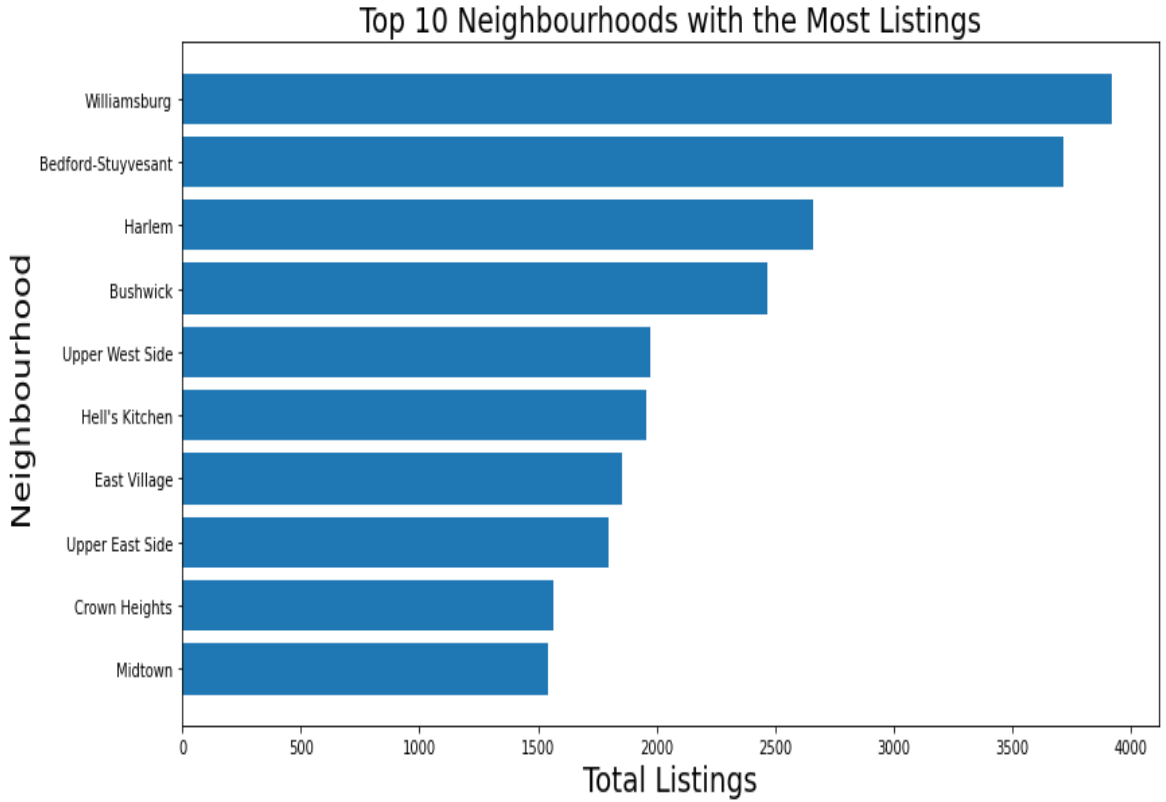


Top 10 neighbourhoods with most listings

```
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
Name: id, dtype: int64
```

Neighbourhoods with the Most Listings

- Williamsburg In Mahhattan has the highest numbers of listings which is 3920.
- It is followed by Bedford-Stuyvesant in Brooklyn with 3714 listings.
- Harlem in Manhattan comes third with 2658 listings



Top 10 most popular neighbourhoods by reviews

	neighbourhood_group	neighbourhood	room_type	number_of_reviews
398	Queens	Jamaica	Private room	629
273	Manhattan	Harlem	Private room	607
369	Queens	East Elmhurst	Private room	543
288	Manhattan	Lower East Side	Private room	540
214	Brooklyn	Park Slope	Entire home/apt	488
146	Brooklyn	Bushwick	Private room	480
379	Queens	Flushing	Private room	474
230	Brooklyn	South Slope	Entire home/apt	467
399	Queens	Jamaica	Shared room	454
259	Manhattan	East Village	Private room	451

Conclusion

- Hosts with most listings are from Manhattan and Brooklyn clearly stating that hosts prefer these two neighbourhood groups.
- Manhattan has the most number of listings out of all neighbourhood groups. Manhattan and Brooklyn have about 85% of the total NYC listings.
- Queens has significantly lower host listings than Manhattan. Staten Island has the lowest number of listings in NYC.
- Manhattan is the most expensive followed by Brooklyn. Bronx has the cheapest listings in NYC.
- The busiest hosts are from Manhattan, Queens and Brooklyn which is normal considering these are the most urban areas of NYC.
- Private room and entire home/apartment are the most in-demand room types.

- Manhattan has skyscrapers, world-famous museums, central parks, fast-paced lifestyle and a high standard of living. Brooklyn is home to beautiful parks, dining experiences and culture. This is why these neighbourhood groups are most preferred by hosts.
- Queens is geographically the largest of the five neighbourhood groups. Queens neighbourhoods have almost all the big-city perks. It has incredible cuisine and dining and has a fabulous culture and connectivity. This can be a strong enough motivation for a new hosts.
- Staten Island has the lowest number of listings in NYC which is based because Staten Island is the most suburban neighbourhood group in NYC. Public transit is very lackluster, no subway connections to other neighbourhood groups and the standard of living is relatively low.
- The neighbourhoods with the most listings are either in Manhattan or Brooklyn. Harlem and Williamsburg are the two neighbourhoods with the most listings in NYC