Capstone Project- 1

Airbnb Bookings Analysis

by-

Team Datavengers

Akshay Auti

Kunal Mahadik

Mehul Bansal

Ramesh Manglav

Points of Discussion

- About the Dataset
- Problem Statements
- Features description
- Data Exploration
- Data Cleaning
- Hosts and neighbourhood groups
- Price distribution across neighbourhoods
- Popular neighbourhood by reviews
- Preferred room type
- Conclusion

About the dataset – Airbnb NYC 2019

- Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily home stays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app.
- The dataset that we would be analyzing consists the booking information on Airbnb from 2008 till 2019.

Problem Statements:

With the help of exploratory data analysis techniques', we will try to answer the following 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?

Features description:

The features in the dataset can be described as follows:

1. id - This is the identity number of the property listed by a particular host.
2. name - It stands for the name of the property listed by the host.
3. host_id - It is the identity number of the hosts who have registered on Airbnb website.
4. host_name - These are the names of the hosts who have listed their properties.
5. neighbourhood_group - These are the names of the neighbourhood groups present in the NYC.
6. neighbourhood - These are the names of the neighbourhood present in the neighbourhood groups in NYC.

7. latitude - These represent the coordinates of latitude of the property listed.
8. longitude - These represent the coordinates of longitude of the property listed.
9. room type - This represent the various types of room listed by host.
10. price - This is the rent of the property listed in USD.
11. minimum nights - This represent the minimum number of nights customer rented the property.
12. Number_of_reviews - This represent the number of customers reviewed the property.
13. last_review - This represent the date when the property was last reviewed.
14. reviews_per_month - It is the count of reviews per month which the property received.
15. calculated_host_listings_count - It is the number of listings done by a particular host.
16. Availability_365 - This represent the number of days the property is available among 365 days.

Data Exploration

Checking the first 5 rows of the dataset.

```
# Importing dataset
airbnb_data = pd.read_csv('/content/drive/MyDrive/Datasets/Airbnb NYC 2019.csv')
airbnb_data.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10

The dataset consist of 48895 observations (rows) and 16 features (columns).

```
[5] #Understanding the data
    # Checking the shape of datset
    print(f'The shape of Airbnb Dataset is {airbnb_data.shape}')

    The shape of Airbnb Dataset is (48895, 16)
```

Checking out the 16 features:

```
# Checking the feature names
print(f' The names of the features present in the dataset are: ')
list(airbnb_data.columns)
```

```
 The names of the features present in the dataset are:
['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']
```

Checking for the categorical columns in the dataset:

```
#Checking for categorical columns
cat_cols = airbnb_data.select_dtypes(include = 'object').columns
print(f' The following are the categorical features in the dataset: {list(cat_cols)}')
```

```
 The following are the categorical features in the dataset: ['name', 'host_name', 'neighbourhood_group', 'neighbourhood', 'room_type', 'last_review']
```

Checking for the non categorical columns in the dataset:

```
62] # Checking for numeric/ non categorical columns
    num_cols = airbnb_data.select_dtypes(exclude = 'object').columns
    print(' The following are the non categorical features in the dataset:')
    list(num_cols)
```

```
 The following are the non categorical features in the dataset:
['id',
 'host_id',
 'latitude',
 'longitude',
 'price',
 'minimum_nights',
 'number_of_reviews',
 'reviews_per_month',
 'calculated_host_listings_count',
 'availability_365']
```

Checking for null values:

The columns like number_of_reviews and reviews_per_month have largest number of null values.

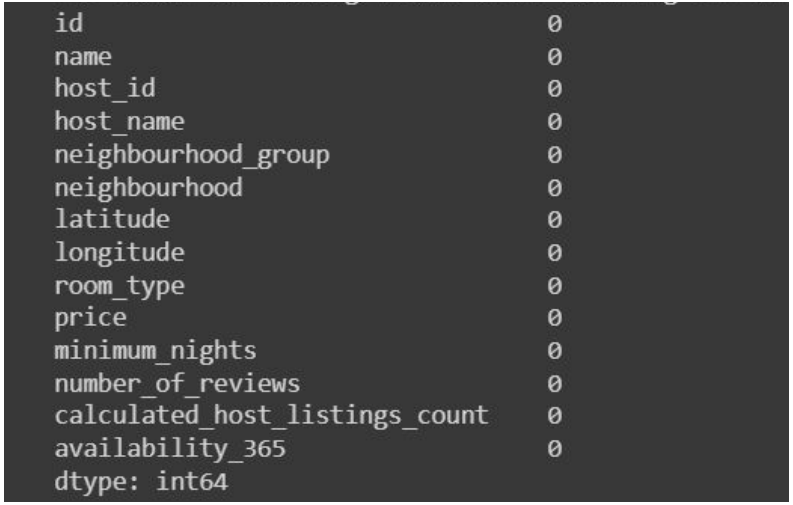
The columns like name and host_name contain fewer number of null values.

```
The missing values before cleaning the data are:
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
```

Data Cleaning

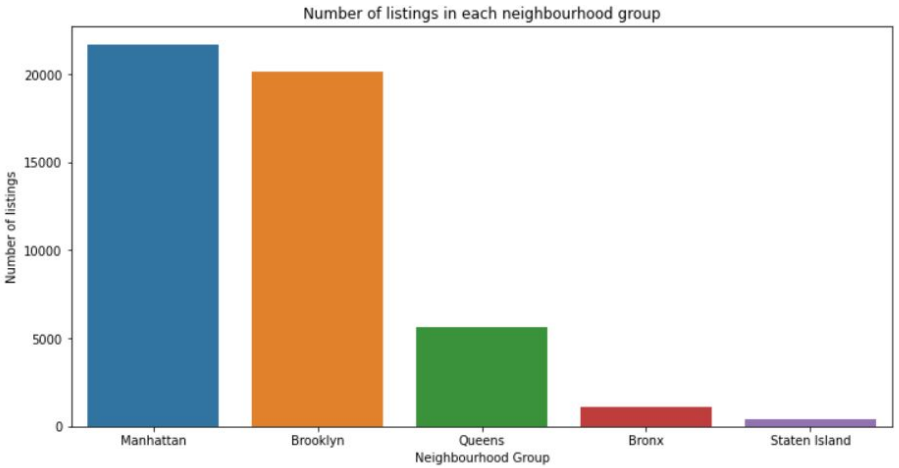
Fixing the null values:
We have dropped the unnecessary columns like number_of_reviews and reviews_per_ month.

Null values after the cleaning the data are as shown.



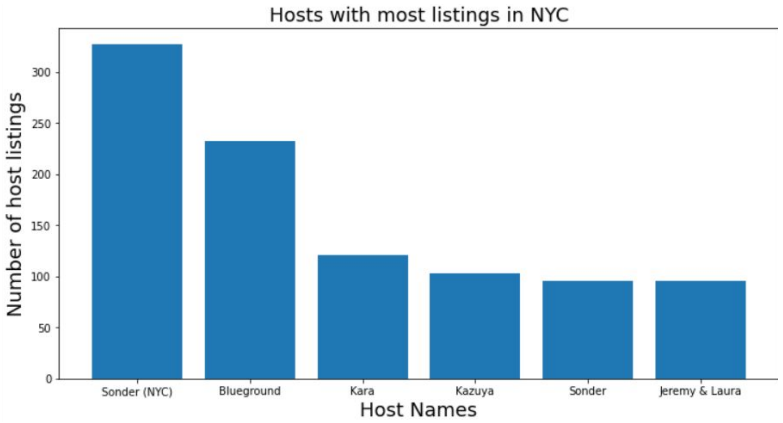
Number of listings across neighbourhood groups.

- It is observed that the Manhattan has highest number of listings of 21661 which is 44.3% of total listings done on Airbnb.
- Brooklyn has 2nd highest number of listings of 20104 which is 41.13% of the total listings.
- Queens comes in 3rd place with 5666 listings whereas Bronx and Staten Island have least number of listings.



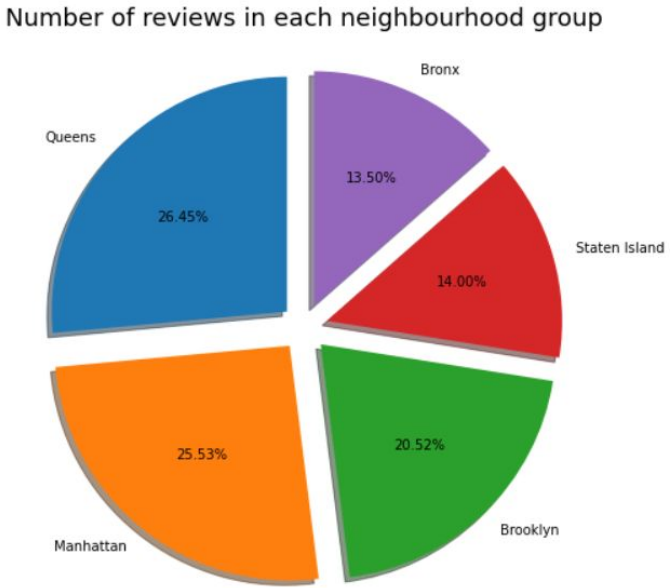
Hosts with maximum listings

- As shown in the adjacent bar chart, we can see there is a good distribution among the top 6 hosts.
- The host named Sonder(NYC) has highest number of listings of 327 in Manhattan neighbourhood group.
- The host named Bluegorund has 2nd highest listings of 232 in Manhattan Neighbourhood group.
- The host Blueground also has 232 listings in Brooklyn.



Areas with maximum reviews

- The number of reviews feature in the dataset represent the customers who have given the reviews to a particular property they have stayed in.
- Looking at the pie chart, Queens has 26.45% of total reviews which is a maximum share.
- Manhattan has 2nd highest number of reviews constituting 25.53%.
- Bronx has 13.50% of total reviews.



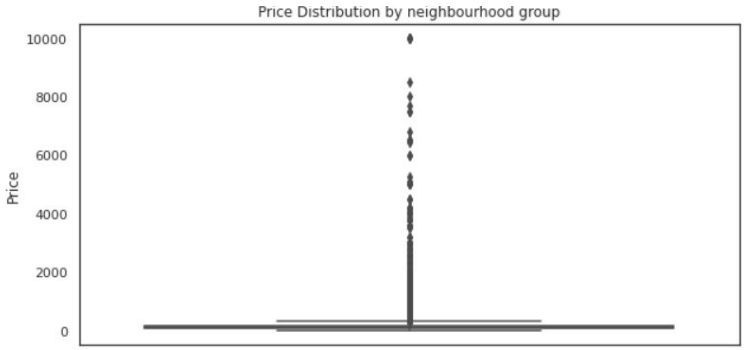
Analysis of rental prices

- Looking at the statistics parameters of the features in the dataset. Average price is 152.72$ and it varies from 0$ to 10000$. 10000$ seems to be very high price which can be a luxurious apartment.

```
# Checking the descriptive statistics
airbnb_data_new.describe()
```

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

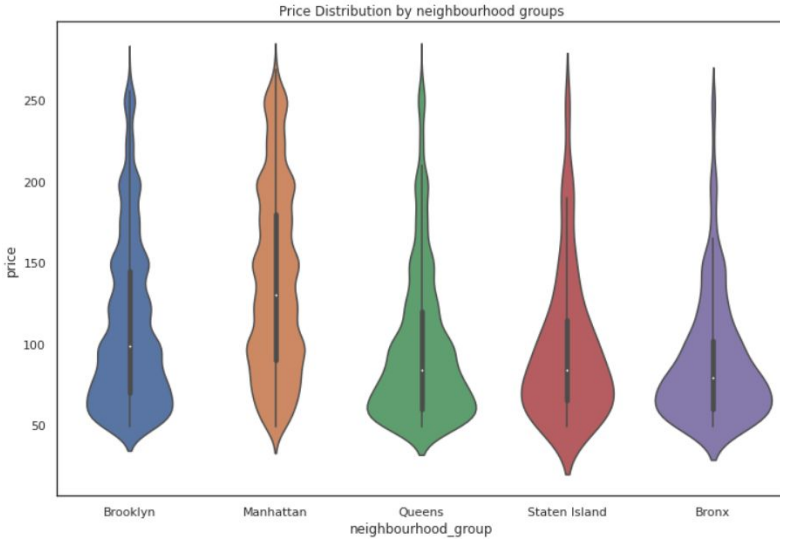
- We can see from the adjacent box plot that price data is skewed. There are outliers present in the data.
- We have eliminated the outliers based on quantile flooring and capping.
- We have considered price range which falls under the 10th percentile and 90th percentile range.
- The range is from 49$ to 269$



- After capping and flooring the price data, we get the below shown boxplot which is much cleaner.

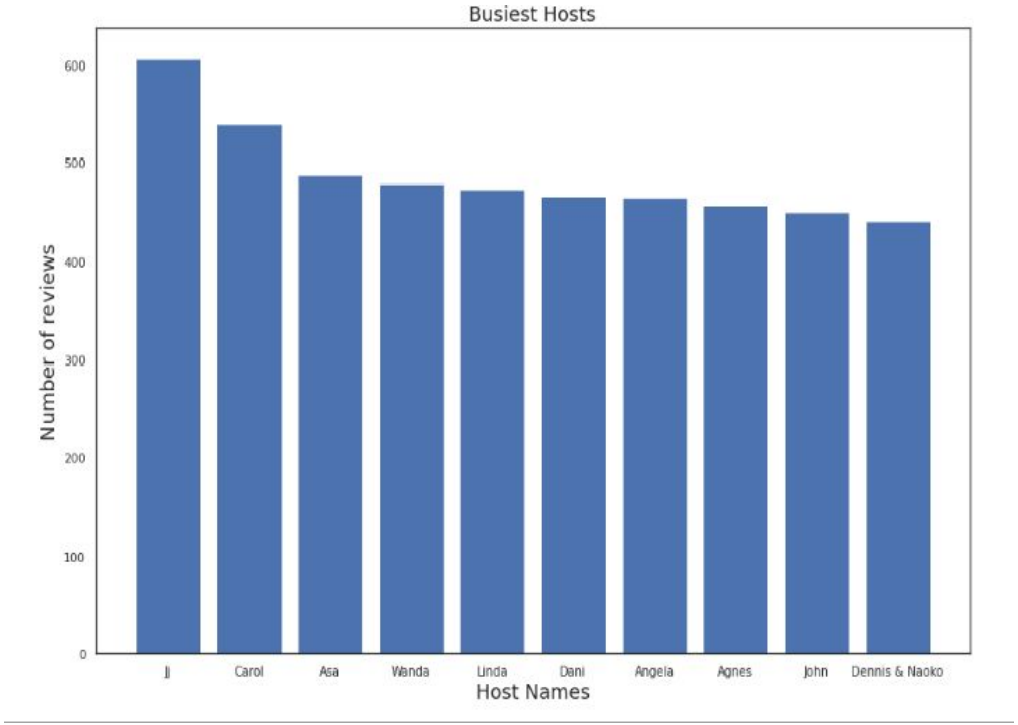
Price distribution across neighbourhood groups

- The adjacent violin plot represents the price distribution across the neighbourhood groups.
- It is observed that, Manhattan has highest price range and is the most expensive one.
- Brooklyn comes in second which has higher range of price but is cheaper with respect to Manhattan.
- Queens, Bronx and Staten Island have narrower price range and have cheaper options.



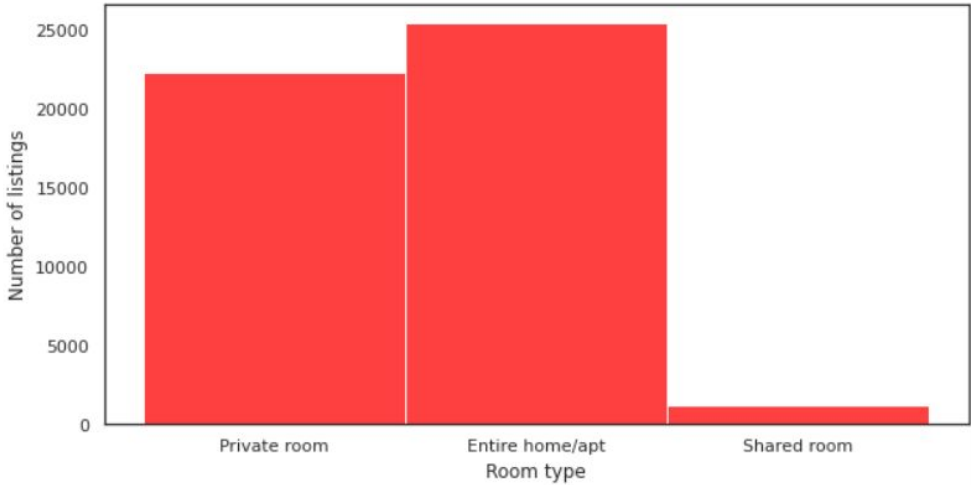
Busiest Hosts

- The adjacent bar plot shows the top 10 hosts with respect to number of reviews.
- Among them Jj has highest numbers of reviews and we can assume that Jj is the busiest host.
- The top hosts have listed private room, entire home/apt.



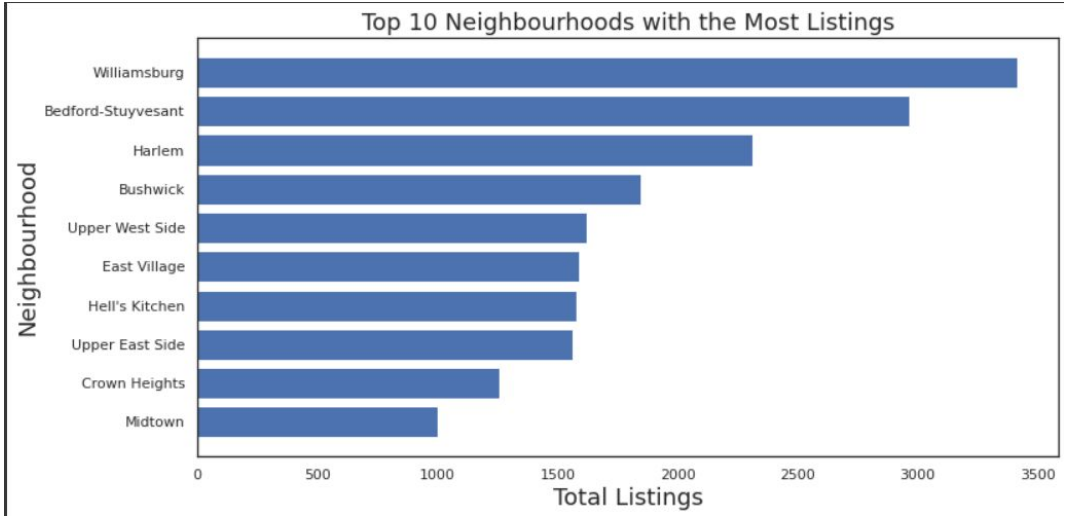
Most Preferred Room Types

- Looking at the adjacent histogram, we can say that there are 3 room types listed in the entire dataset namely Private room, Entire home/apt, Shared room.
- Among this types the most preferred room type is Entire home/apt as well as private room.
- Shared room is least preferred by people.



TOP 10 Neighbourhoods with the Most Listings

- From the various Neighbourhoods top 10 Neighbourhoods are listed in the chart
- Williamsburg has the highest numbers of listing which is around 3500 and it is located in Manhattan.



Conclusion:

- Manhattan has most number of listings, followed by Brooklyn and Queens. Staten Island has least number of listings.
- Manhattan and Brooklyn make up for 87% of listings available in NYC.
- Brooklyn and Manhattan are most liked neighbourhood groups by people.
- Queens has significantly less host listings than Manhattan. So, we should take enough steps to encourage host listings in Queens as there is decent demand in the neighbourhoods of Queens.
- The maximum demand is for private rooms and entire home/apartment. People are more interested in cheaper rentals.
- The top 10 neighborhoods with the most listings are located either in Manhattan or Brooklyn, with Harlem and Williamsburg presenting leading numbers in each borough, respectively.

- Manhattan is the top neighbourhood group in terms of listings as well as highest price range. It was assumed that Brooklyn might have most number of listings as it is a quite popular place.
- Given that Manhattan is world-famous for its museums, stores, parks, and theaters, also its substantial number of tourists throughout the year, it makes perfect sense that prices are much higher in this neighbourhood group.
- Brooklyn comes in second with significant number of listings and cheaper prices as compared to the Manhattan. With most listings located in Williamsburg and Bedford Stuyvesant — two neighborhoods strategically close to Manhattan — tourists get the chance to enjoy both boroughs equally while spending less.