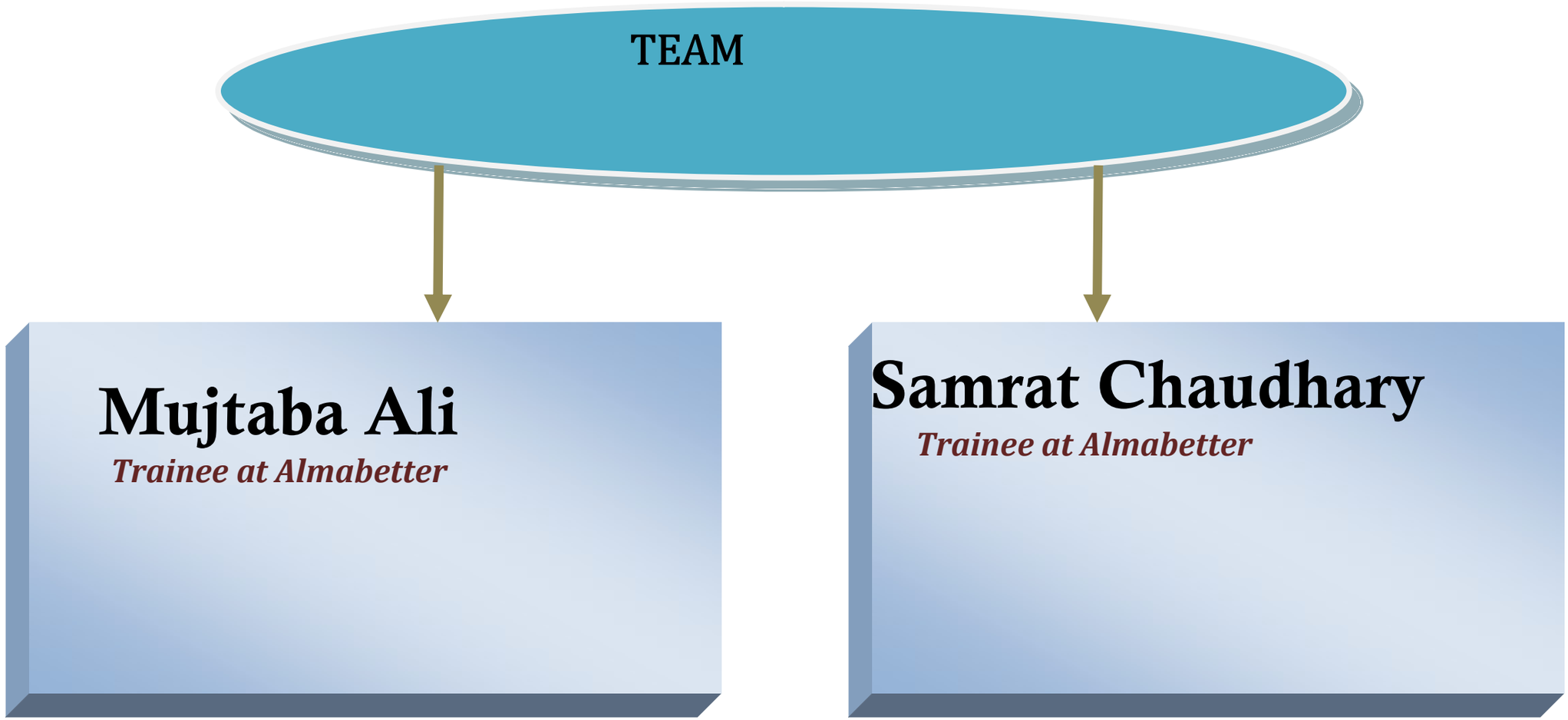
Capstone Project -1

Airbnb Bookings Analysis


TEAM
Mujtaba Ali
Trainee at Almabetter
Samrat Chaudhary
Trainee at Almabetter

❑ Introduction

- Airbnb was founded in 2008 by Brain Chesky, Nathan Blecharczyk, and Jeo Gebia.
- Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other, Airbnb enables travelers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world. Airbnb does notown any of the listed properties .
- Airbnb offers around 6 million places to stay in throughout the world. At any given time, guests book 1.9 million listings in Airbnb.

1. Here we are going to do an Exploratory Data Analysis on the data set of Airbnb NYC (2019).

- This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.
- Our main objectives of analysis will be the some of statements given to us which can be briefed as learnings from hosts, areas, price, reviews, locations etc. But we are not limited to it, we will also try to explore some more insights.

❑ Data Summary

- In this session, we will have the overview of the basic understanding of our dataset variables. What does particular features means and how it's distributed, what type of data is it. Airbnb dataset is having 16 columns in total. We can get this by basic inspection of our dataset. Some columns are not significant for our analysis which can also be kept off.

- Now let's look at some of the useful columns in our data set.

❑ Understand the variables

1. ID:

- It’s a unique id for House/apartment.

2. Name:

- Name of the listing House/apartment.

3. Host Id:

- Host Id is the government approved id for each individuals Who rent their properties on Airbnb.

4. Host Name:

- Host names are basically the name of the individual or organization Who own a room/apartment on Airbnb website.

5. Neighbourhood groups:

- Neighbourhood groups are the cluster of neighborhoods in the area.
- There are about 5 boroughs in the state.

6. Neighbourhood:

- When searching for accommodations in a city, guests are able to filter by neighbourhood attributes and explore layers of professional-quality content, including neighbourhood maps, custom local photography and localized editorial, details on public transportation and parking, and tips from Airbnb's host community.

7. Latitude:

- Latitude is the measurement of distance north or south of the Equator.

8. Longitude:

- Longitude is the measurement east or west of the prime meridian.

9. Room type:

Airbnb has 3 categories for types of space :

- Entire House/Apartment
- Private room
- Shared room

10. Price ($):

- The total price ($) of Airbnb reservation is based on the rate set by the Host, plus fee or costs determined by either the Host or Airbnb.

11. Minimum_nights:

- Minimum night is criteria for booking that guest have to pay for book that House/room or apartment.

12. Number_of_reviews:

- Number of review of each host submitted by guest.

13. Last_review:

- Latest review submitted by guest as a feedback.

14. Reviews per month:

- Number of review Host get per month.

15. calculated_host_listings_count:

- Amount of listing per host.

16. Availability 365:

- It is an indicator of the total number of days the listing is available for during the year.

❑ Steps used for EDA:

✓ ***Importing our data:*** In this section we just loaded our dataset in colab notebook and read the csv file.

✓ ***Data Cleaning and Processing:*** In this section we have tried to remove the null values and for some of the columns we have replaced the null values with the appropriate values with reasonable assumptions.

✓ ***Analysis and visualization:*** In this section we have tried to explore all variables which can play an important role for the analysis. In the next parts we have tried to explore the effect of one over the other. In the next part we tried to answers our hypothetical questions.

✓ ***Future scope:*** There are many apartments having availability as 0, which means they might stopped their business, we can find the relation of neighbourhood with these apartments if we dig deeply, various micro trends could be unearthed, which we are not able to cover during this short duration efficiently. There are various columns which can play an important role in further analysis such as number of reviews and reviews per month finding its relation with other factors or other grouped factors can play an important role.

❑ Exploratory Data Analysis on Dataset:

- ***Now we will analyze these questions that we made from dataset :***

1. Highest number of apartments owned by host.

2. Price distribution of room type.

3. Relation between neighbourhood and reviews.

4. Top 10 neighbourhood having highest number of apartments.

5. Room_type distribution over the location.

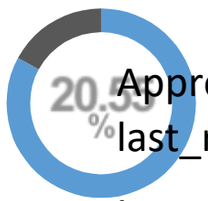
6. Learning from prediction (location).

7. Average price of each room_type in each neighbourhood_group.

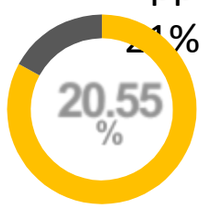
❑ EDA :

- ***Let's check the null values in data set.***

  As we can see in the data column '*last review & reviews per month'* having a large number of null values.



Approx 21% of last_reviewsarenul



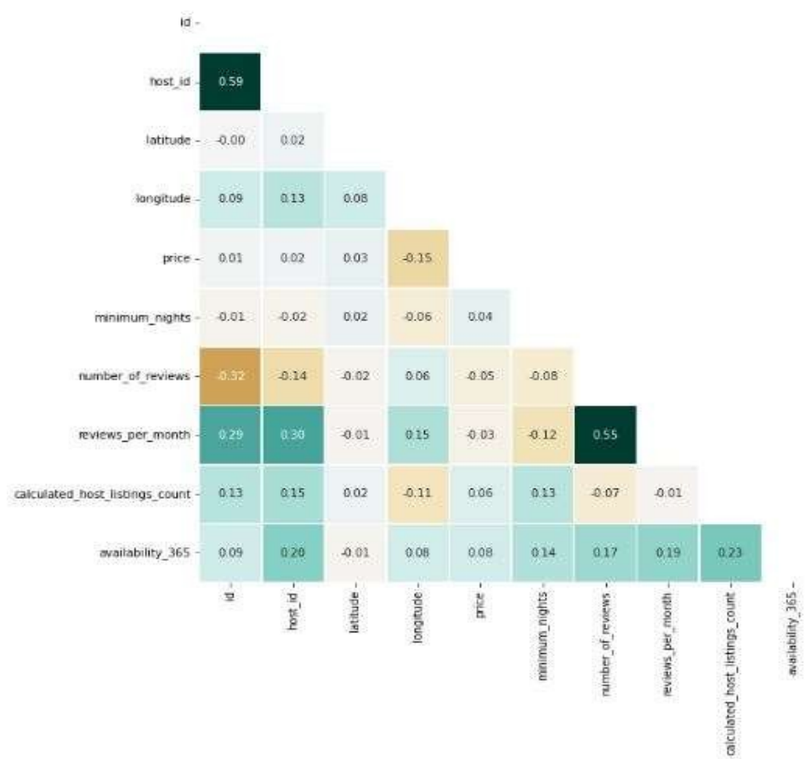
Approx
21% of reviews_per_month are null.

```
id
name
host_id
host_name
neighbourhood_group
neighbourhood
latitude
longitude
room_type
price
minimum_nights
number_of_reviews
last_review                     100
reviews_per_month               100
calculated_host_listings_count
availability_365
dtype: int64
```

- ***Let's check the correlation between columns.***

  As we can see that '*host id & id'* , *'reviews per month & number of reviews'* have good relation in dataset.
  On other side '*number of reviews &Id'* also have good relation

Feature-correlation (pearson)



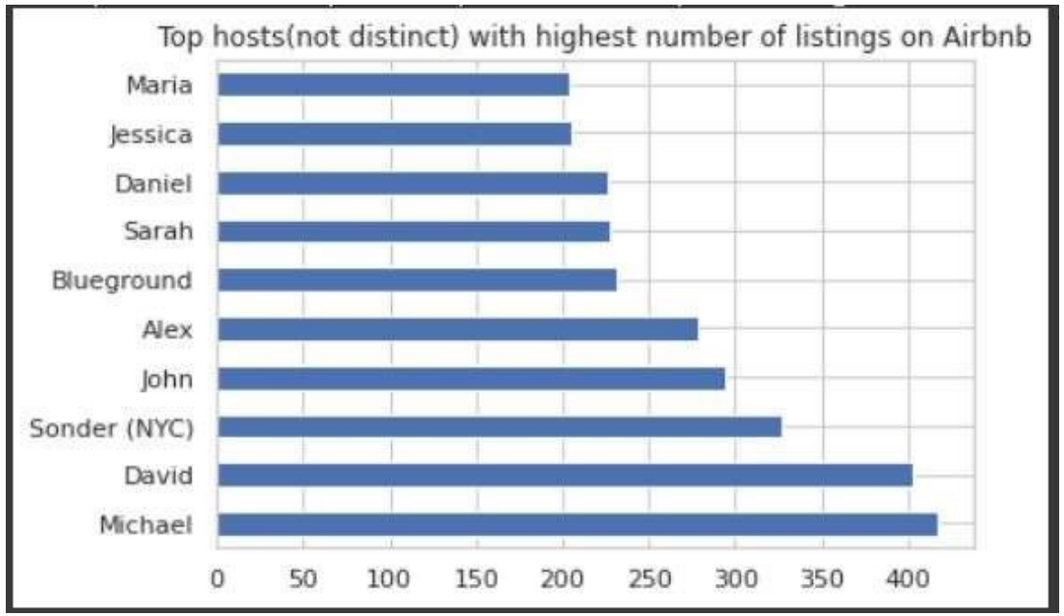
	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	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

	last_review
0	2019-01-01
1	2018-01-01
2	2019-01-02
3	2019-06-23
4	2018-01-02
5	2017-01-01
6	2019-05-27
7	2017-01-02
8	2016-01-02
9	2019-07-01
10	2016-01-03
11	2018-12-30
12	2019-01-03
13	2019-06-24
14	2018-12-31

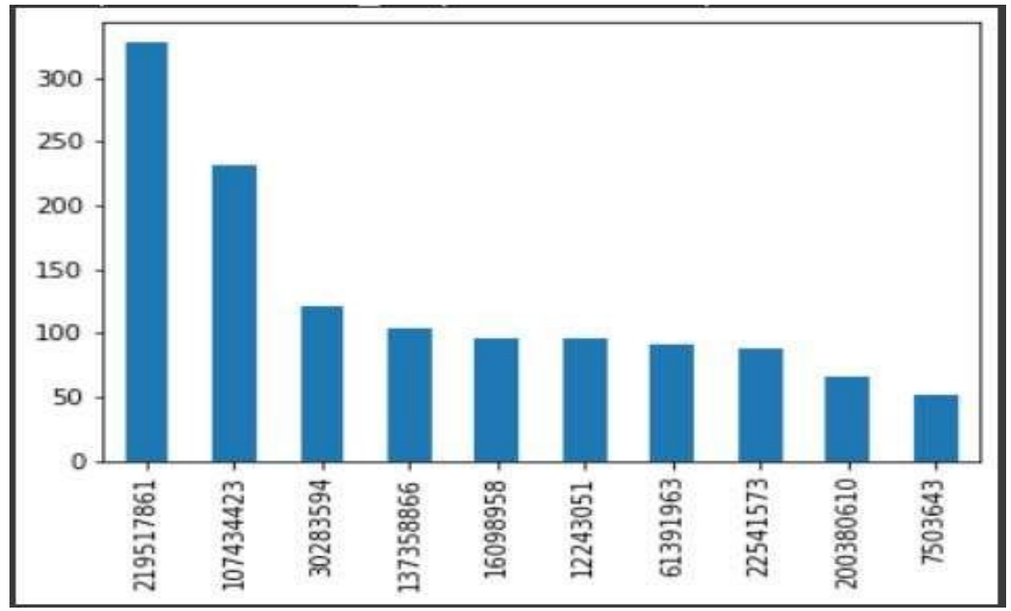
- As we can see here in “Price column” the minimum price is ‘0’ that looks strange.
- And in “*availability_365*” 25%ile of data is ‘0’ that seems awkward let’s check the Accurate data in “*availability_365*” having ‘0’ availability.
- There is approx. 36% of ‘*availability_365’* data is ‘0’ is bit shocking if you have a business providing stays on Airb availability is ‘0’ days that is an extreme case and extreme cases are shocking when it comes 36%.
- Let’s check by ‘*last_review’* either that house are still Open or Closed.
- The dataset is of the period of ‘2011-19’ and we can see in the *‘last_review*’ table there is some review which was delivered in ‘2016, 2017’ which show either that listings are already closed or not preferable.
- Now let’s fill the price table on behalf of mean price of same ‘*room_type’*

❑ Analysis:

1. Highest number of apartment owned by host.



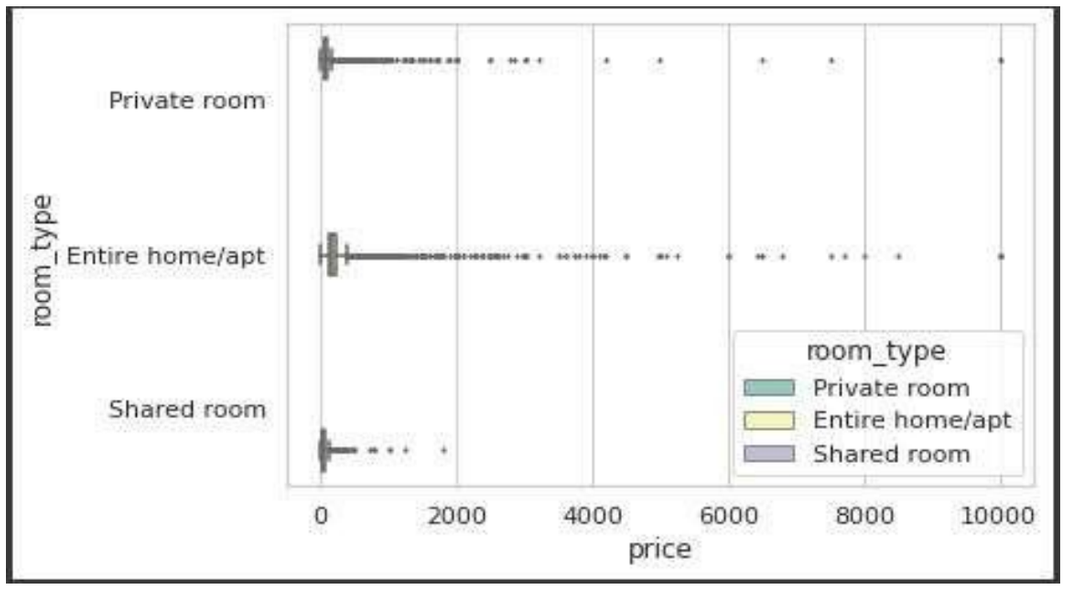
- As we can see in above fig. '*Michael*' has highest number or apartments, but as we know 'Name' is not unique here so let's check by '*Id'*.



- In fig. above we can see the '*id-219517861* which belongs to Sonder (NYC)' has highest number of apartments,
- It happened because '*Id'* is unique here but host_name '*Michael'* is more than one here.

2. *Price distribution across the room_type*

We can notice that there are many outliers for price column for each of the 'room_type' category, so let's just check why there is so high price or what else we can conclude for hosts having highest price for the rooms.



Some suggestion that can help the hosts as well as the guest.

- Kathrine and Erin have so high price and having no availability then what is the benefit of keeping too high price.
- The last review is also 2-3 years back (as the data was collected in 2019) which is also bad.
- The reviews may be low as there may be very few people who had stayed Kathrine's, Erin's and Jelena's apartment so they have very less reviews per month.
- I would have suggested to keep moderate(average).

	host_name	reviews_per_month	last_review	availability_365	price	neighbourhood_group
9151	Kathrine	0.04	2016-02-13	0	10000	Queens
17692	Erin	0.16	2017-07-27	0	10000	Brooklyn
29238	Jelena	0.00	NaN	83	10000	Manhattan

3. ***Relation between neighbourhood and reviews.***

- In the data we can clearly see that Neighbourhood '*Bedford Stuyvesant'* has highest number of *'reviews per month'* and '*number of reviews'*.

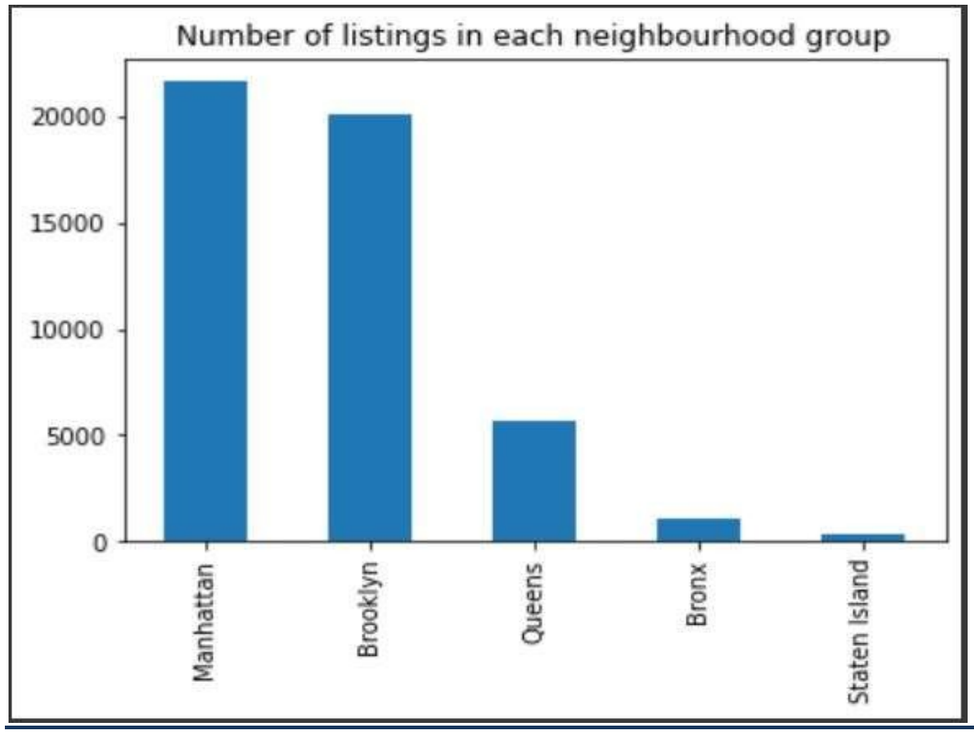
- With this data we can easily predict that as Larger number the reviews as busiest the host.

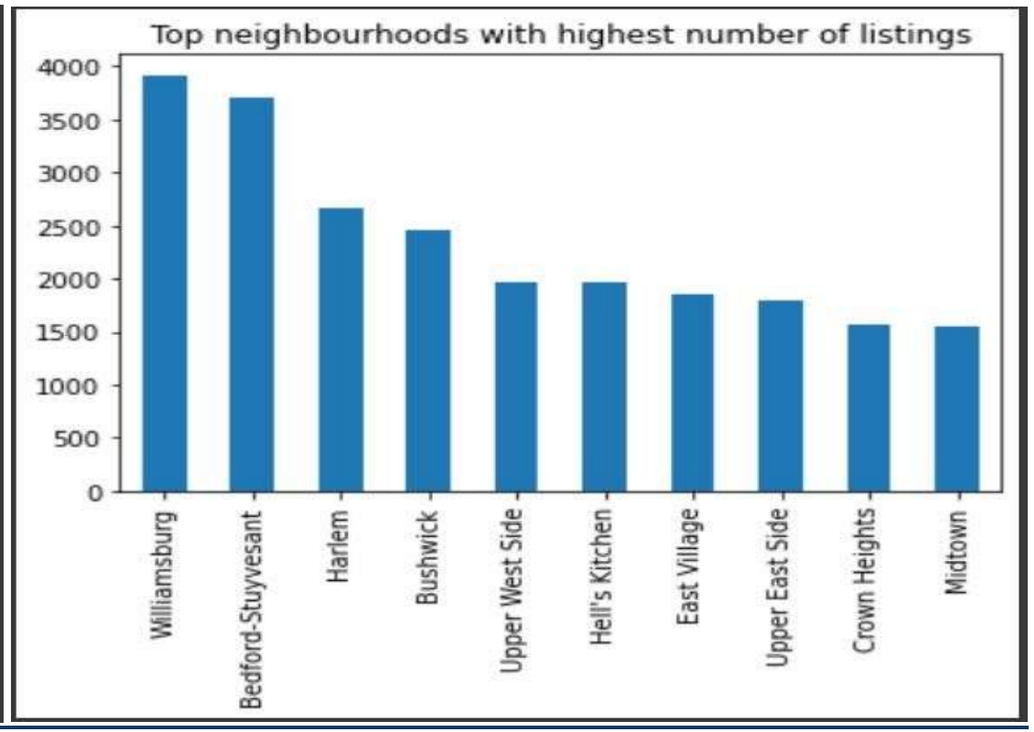
- But this can not be always true as all guest do not Submit the reviews, but this will help Us in finding Busiest host.

	neighbourhood	reviews_per_month
0	Bedford-Stuyvesant	4874.52
1	Williamsburg	3475.77
2	Harlem	2956.23
3	Hell's Kitchen	2818.79
4	Bushwick	2632.51

	neighbourhood	number_of_reviews
0	Bedford-Stuyvesant	110352
1	Williamsburg	85427
2	Harlem	75962
3	Bushwick	52514
4	Hell's Kitchen	50227

4. ***Top 10 neighbourhood having highest number of apartments.***





- We can see in fig. that *"Manhattan"* in *'neighbourhood group'* have maximum number of listing.

- We can see in fig. that *"Williamsburg"* in *'neighbourhood'* have maximum number of listing.

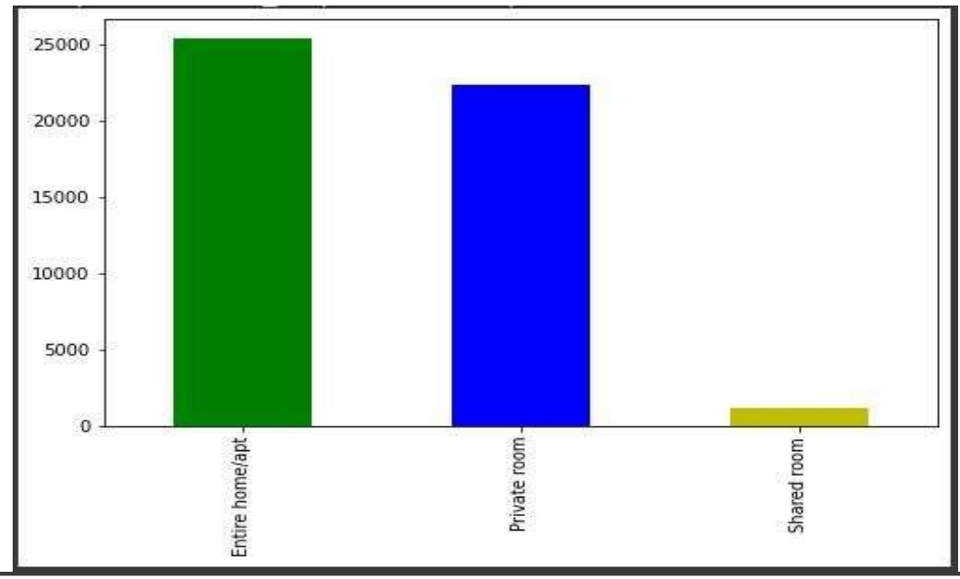
5. *Room_type distribution over the location.*

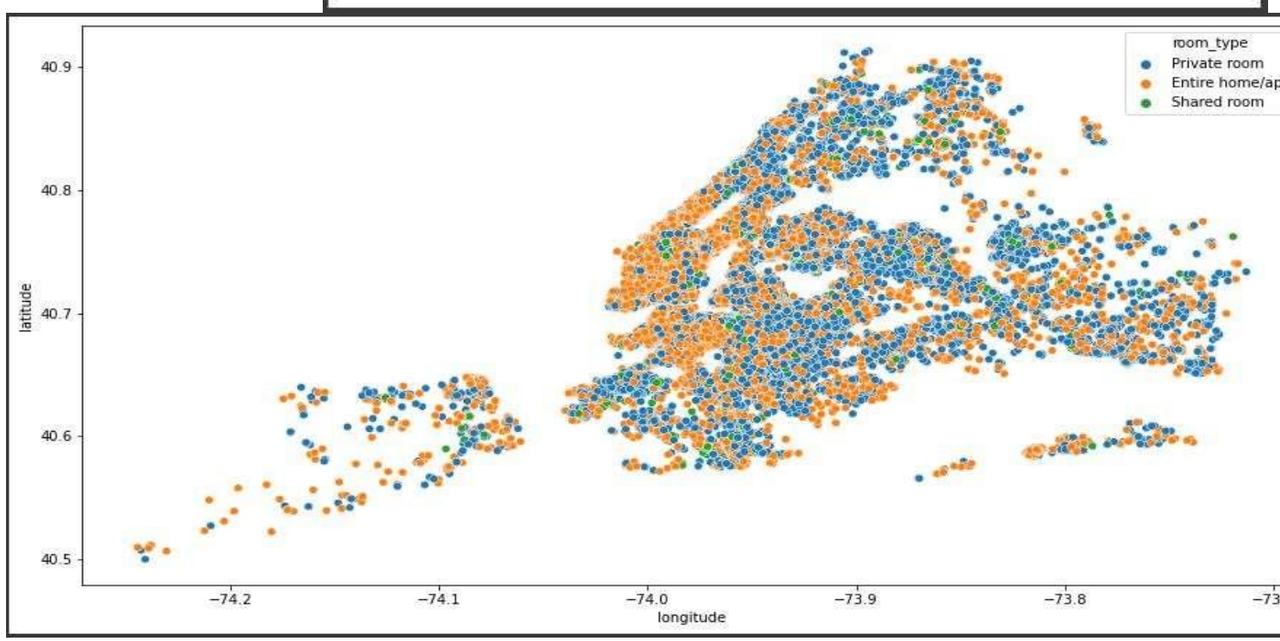
- Room type of most number of listing is ‘Entire home/apartment’, but there is not big difference in ‘private room and Entire home’.

- Scatter plot on Map show the Cluster of ‘*room_type’.*

- And we can easily predict that ‘*room_type’* is almost same in every *‘neighbourhood’*, which means the Booking of room type is almost same.

- But if we talk about ‘Shared room’ we can see a huge difference here, and on the basis of this data we can say that ‘Shared rooms’ are not preferable.



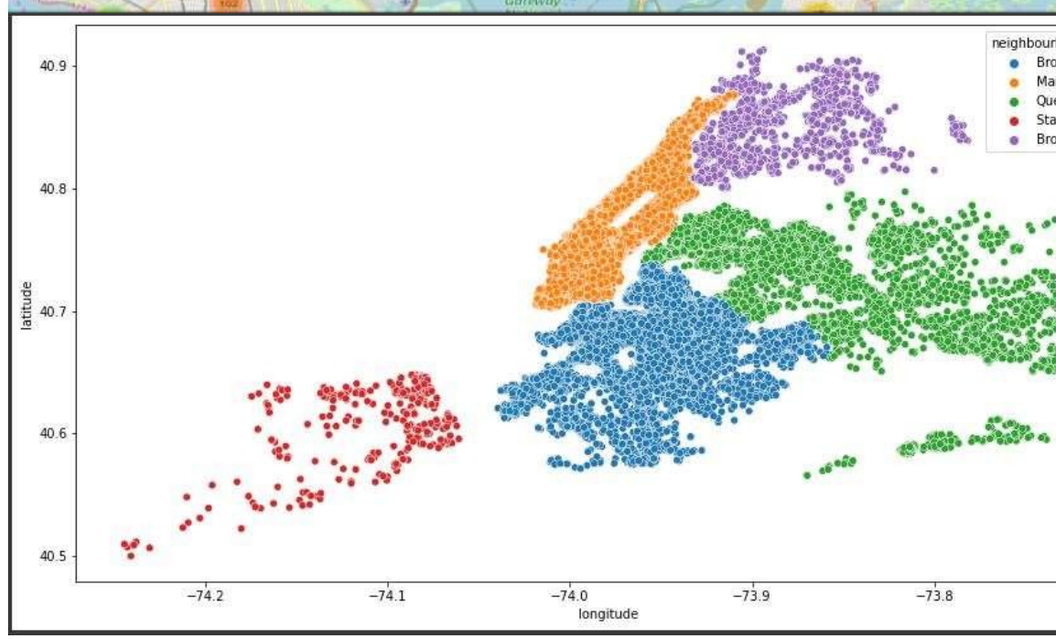


6. ***Learning from prediction (location).***

- Map shows the exact location of all the apartments
  With the help of ‘latitude’ and ‘longitude’ co-ordinates.

- Scatter plot on Map shows the cluster of *‘neighbourhood group’* with the help of *‘latitude’* and *‘Longitude’*.

- We can easily see that *‘Manhattan’* is most dense area
  Which has maximum number of listings.

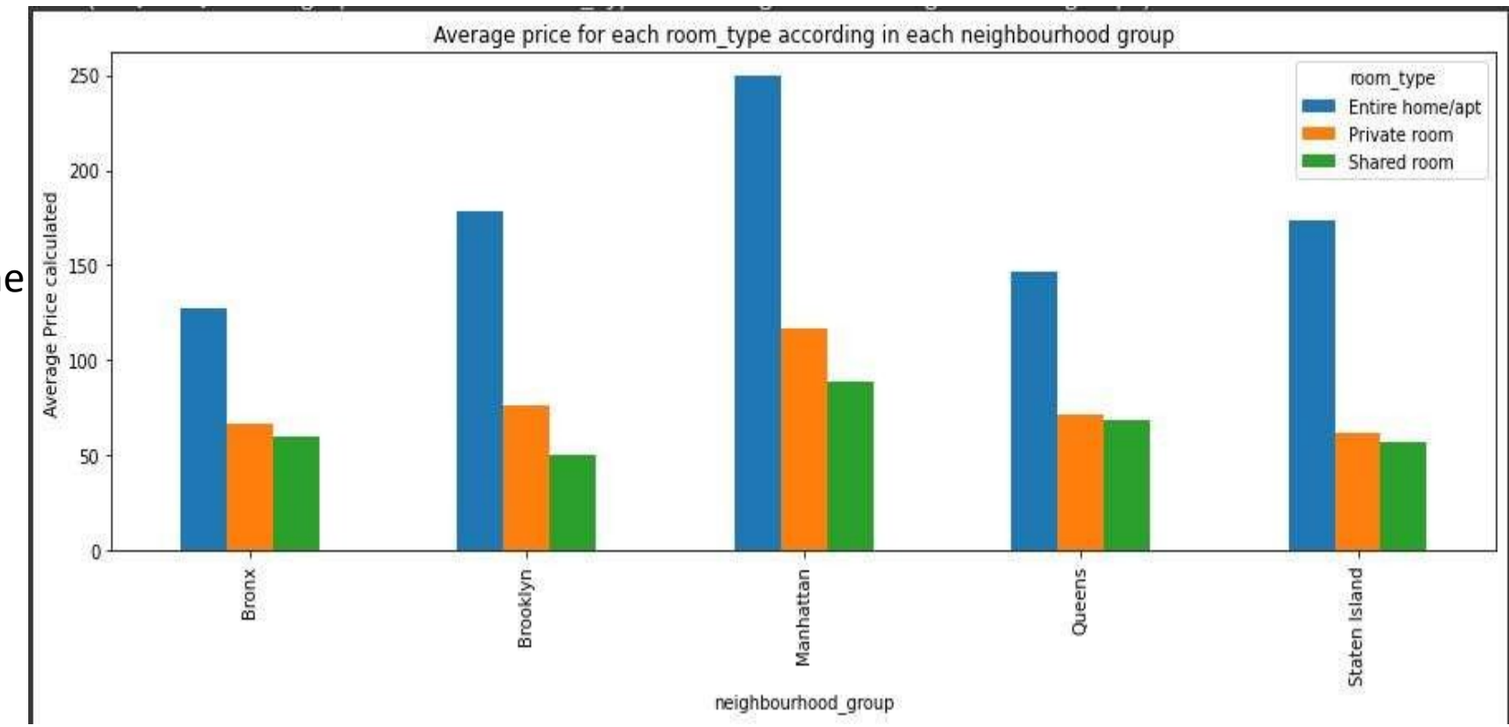


7. ***Average price of each room_type in each neighbourhood_group.***

- *Observations:* As we can see that Manhattan is most costly and Bronx is Cheap for each *room_type.*

- But I think we can make it more useful for business implementation if we do some analysis on successful hosts according to the highest no of reviews so that we can suggest that price to our hosts to get good business.

- We seen before that shared rooms are not Preferable here we can see one that.



Challenges:

- The analysis we found out that 36% of the data has 0 availability in the availability_365 column, which is an extreme case. But we didn't have other relevant required data so we couldn't alter this column.

- Then, we found out that there were many listings whose price was 0, which is not normal. So, we filled these values by the respective median price and updated the price column.

- While getting host_name with the highest listings we found out that there are many hosts whose names are the same so we went by host_id as this is unique, host_name is not unique

- Many listings whose date of the last review is very old this can mean that they must have stopped their business then those listings are of no use to us for doing analysis at present. But this assumption can also be wrong so we didn't alter this column.

- There are many outliers in the price column of some hosts which are not benefitting the host as well as the customer.

- The biggest challenge that we faced is finding the busiest hosts. If we try to find the busiest hosts by an only number of reviews then this may be not the correct metric, because we don't the current status of the host having the Highest number of reviews.

Conclusions:

-
- Sonder (NYC) host is having most number of listings on Airbnb in NYC.
- Williamsburg neighbourhood has most number of listings.
- Upper West Side, Astoria and Greenpoint neighbourhoods have costliest listing in NYC.
- Bedford-Stuyvesant neighbourhood has highest number of total reviews and Theater
- District neighbourhood has highest number of reviews_per_month.
- Most of the listings on Airbnb in NYC are either Entire Home/Apartment or Private Room. The people who prefer to stay in entire home/apartment are likely going to stay longer, whereas people who prefer to stay in private_room are likely to stay for a shorter period of time than the people who prefer to stay in entire home/apartment.
- Many rows are having values as 0 in price column, so this seems like an error which must be rectified by Airbnb.
- Maya (host) has the highest total number_of_reviews.
- Average prices of all the room_types in Manhattan are more than the average price of
- Maximum listings are listed on Manhattan and Brooklyn neighbourhood_groups. Staten
- Island and Bronx neighbourhood_group have very less numbers of listings.

THANK YOU