Capstone Project -1

Airbnb Bookings Analysis (EDA)

Vaibhavkumar Gupta

Priyanka Pal

Bhavik Verma

Shayan Somanna

Dilkhush Sharma

❖ Table of Content

❖ Introduction

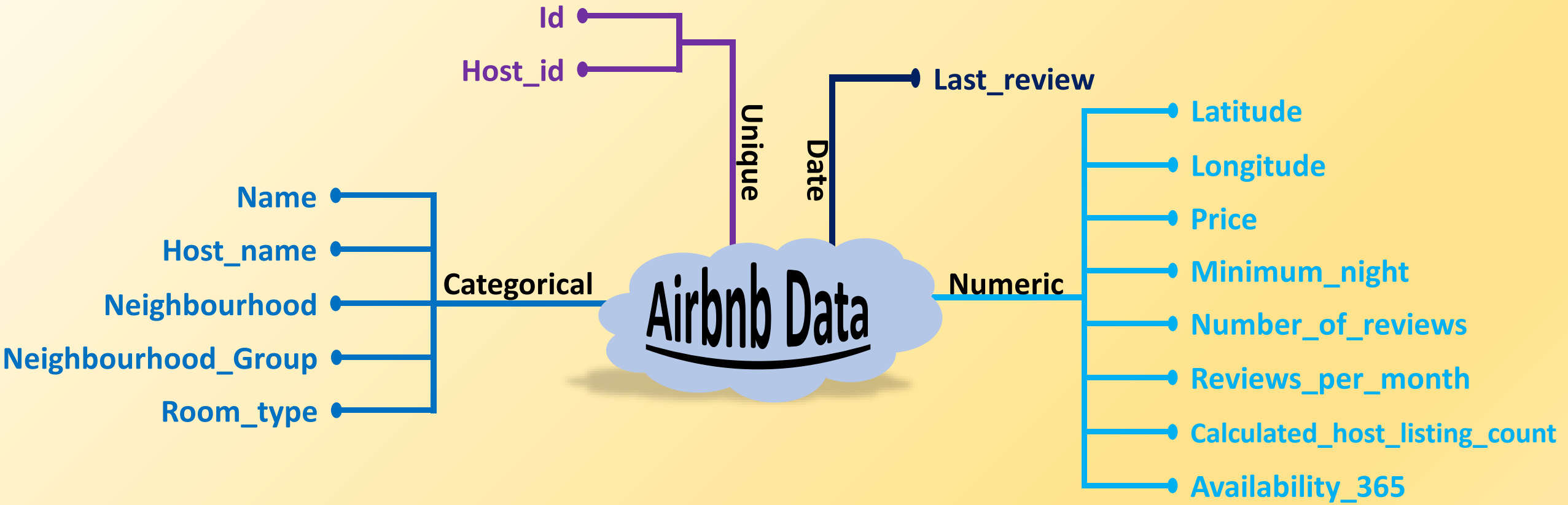
- Airbnb was founded in 2008 by Brain Chesky , Nathan Blecharczyk and Jeo Gebia.
- Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other, Airbnb enables travelers 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 not own 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.

- 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 its 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.

❖ Data Summary continued….



❖ Understand the variables

- **ID:**
  - It's a unique id for House/apartment.
- **Name:**
  - Name of the listing House/apartment.
- **Host Id:**
  - Host Id is the government approved id for each individuals Who rent their properties on Airbnb.
- **Host Name:**
  - Host names are basically the name of the individual or organization Who own a room/apartment on Airbnb website.
- **Neighbourhood groups:**
  - Neighbourhood groups are the cluster of neighborhoods in the area.
  - There are about 5 boroughs in the state.
- **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.
- **Latitude:**
  - Latitude is the measurement of distance north or south of the Equator.
- **Longitude:**
  - Longitude is the measurement east or west of the prime meridian.

- **Room type:**
  - Airbnb has 3 categories for types of space :
    - Entire House/Apartment
    - Private room
    - Shared room
- **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.
- **Minimum_nights:**
  - Minimum night is criteria for booking that guest have to pay for book that House/room or apartment.
- **Number_of_reviews:**
  - Number of review of each host submitted by guest.
- **Last_review:**
  - Latest review submitted by guest as a feedback.
- **Reviews per month:**
  - Number of review Host get per month.
- **calculated_host_listings_count:**
  - Amount of listing per host.
- **Availability 365:**
  - It is an indicator of the total number of days the listing is available for during the year.

❖ Exploratory Data Analysis:

AI

❑ What is EDA ?

“Exploratory Data Analysis “ is very important in machine learning. Whenever we start our work on any project we must analyze the factors deeply. Hypothetical questions and that hypothetical questions lead to some hidden facts. This collaborative work is simply known as EDA.

- **The following steps are involved in the process of EDA:**

- **Acquire and loading data**
- **Understanding the variables**
- **Cleaning dataset**
- **Exploring and Visualizing Data**
- **Analyzing relationship between variables**

❖ Approach used for EDA:

The approach we have used in this project is defined in the given format.

1. **Loading our data:-** In this section we just loaded our dataset in colab notebook and read the csv file.
2. **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.
3. **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.
4. **Future scope of Further Analysis:-** 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.

❖ Graphs used for EDA:

❑ Types of Graphs we have used for Data Visualization:

- ➢ Count Plot
- ➢ Bar Plot
- ➢ Scatter Plot
- ➢ Heatmap
- ➢ Box plot

❖ Libraries used for EDA:

❑ Python Libraries we used for graphs:

- ➢ Matplotlib
- ➢ Seaborn
- ➢ Klib
- ➢ Numpy
- ➢ Pandas
- ➢ Folium

❖ Exploratory Data Analysis on Dataset:

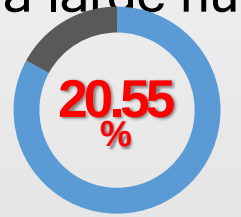
- **Now we will analyze the Data and will get the answer of following question :**

1. Highest number of apartments owned by host.
2. Top 10 neighbourhood having highest number of apartments.
3. Top 3 neighbourhood having highest price in each neighbourhood_group.
4. Relation between neighbourhood and reviews.
5. Learning from location.
6. Room_type distribution over the location.
7. Price distribution of room type.
8. Top 5 host which received highest number of reviews.
9. Average price preferred by customers.
10. Average price preferred to get good number of reviews.
11. Top 10 busiest Host.

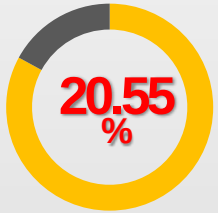
❖ EDA :

➢ 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_reviews are null.

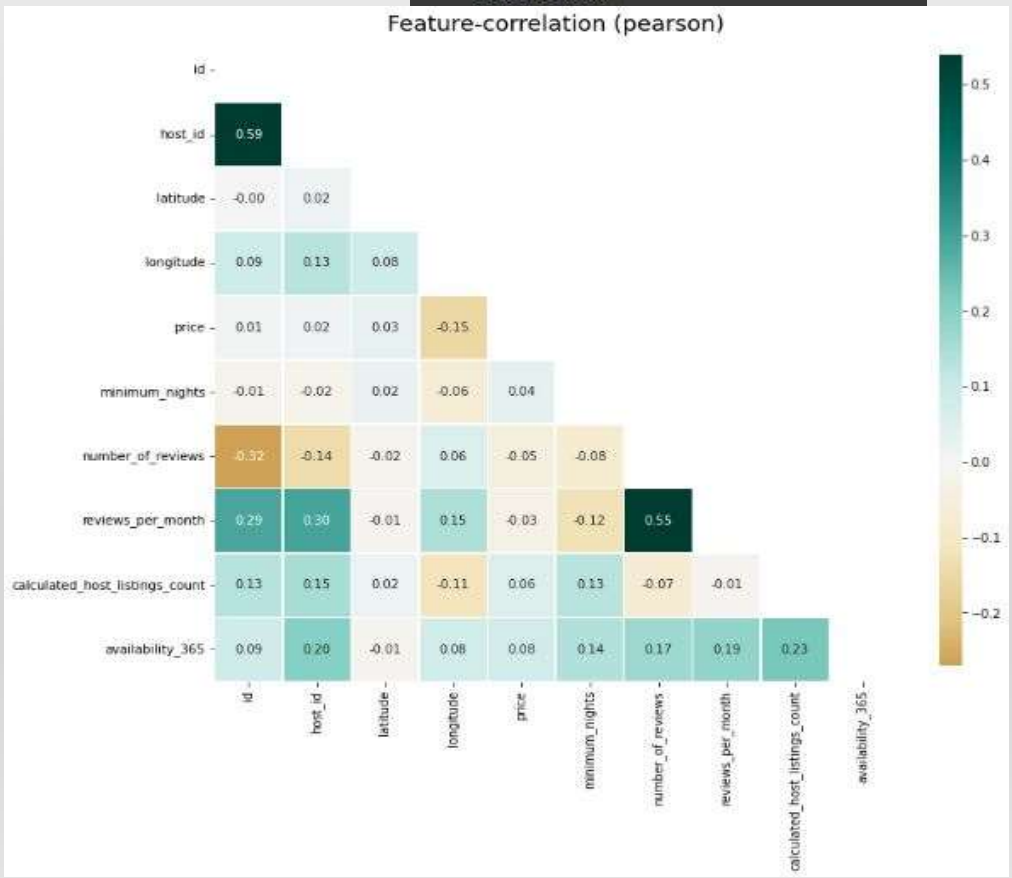


Approx 21% of reviews_per_month are null.

```
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
```

➢ 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.



❖ EDA :

	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
0	2019-01-01	194
1	2018-01-01	142
2	2019-01-02	129
3	2019-06-23	90
4	2018-01-02	86
5	2017-01-01	85
6	2019-05-27	75
7	2017-01-02	73
8	2016-01-02	67
9	2019-07-01	63
10	2016-01-03	61
11	2018-12-30	61
12	2019-01-03	60
13	2019-06-24	59
14	2018-12-31	57

- 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 Airbnb and the availability is ‘**0**’ days that is an extreme case and extreme cases are shocking when it comes 36% .

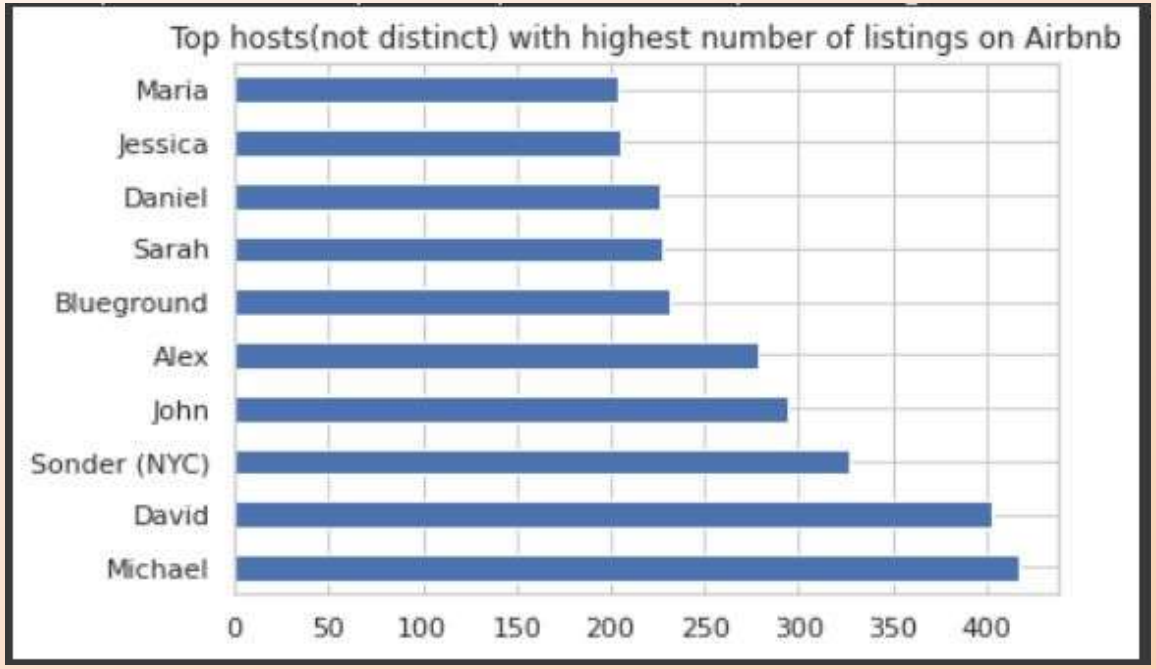
➢ Let’s check by **‘last_review’** either that house are still Open or Closed.

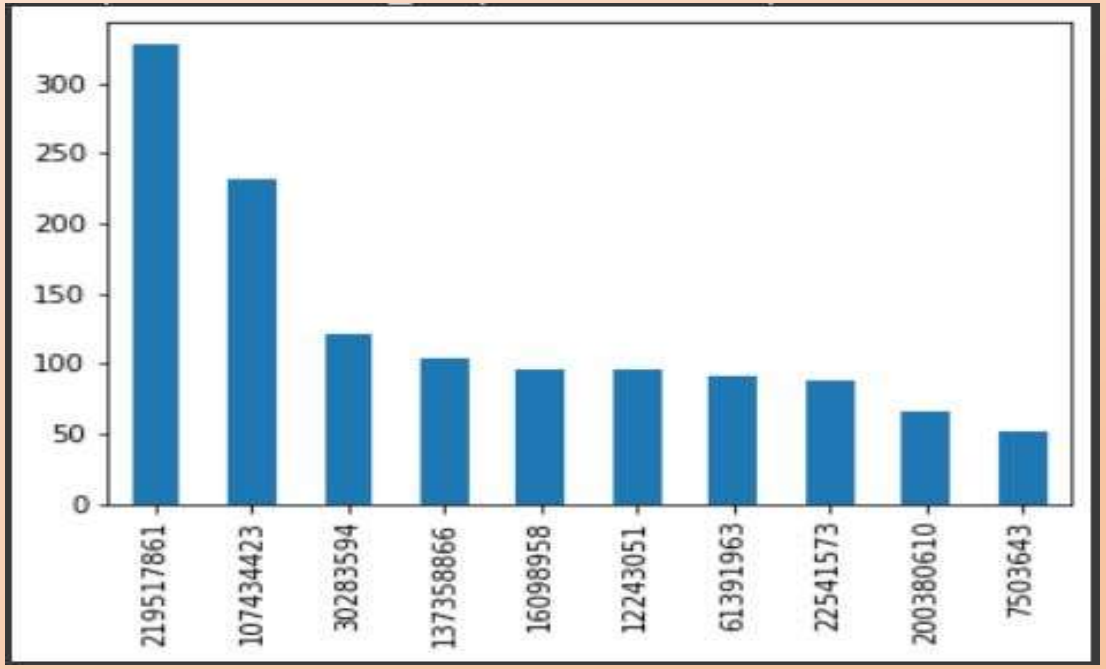
✓ The dataset is of the period of **‘2011-19’** and we can see in the **‘last_review’** table there is some review which was delivered in **‘2016, 2017’** which show either that listings are already closed or not preferable.

❑ Now let’s fill the price table on behalf of mean price of same **‘room_type’**

❖ EDA :

1. Highest number of apartment owned by host.



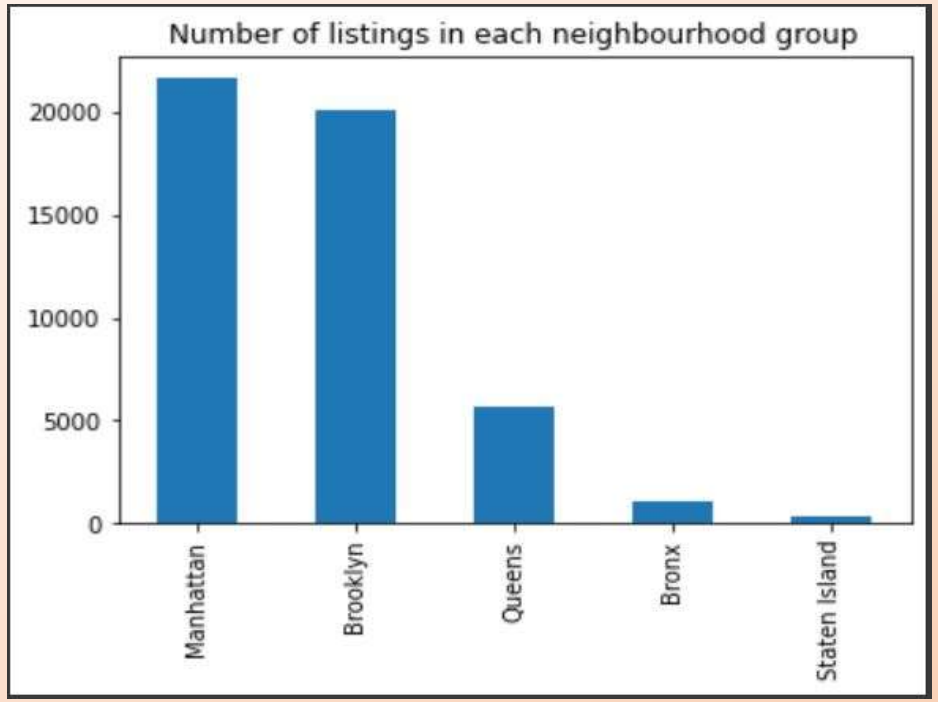


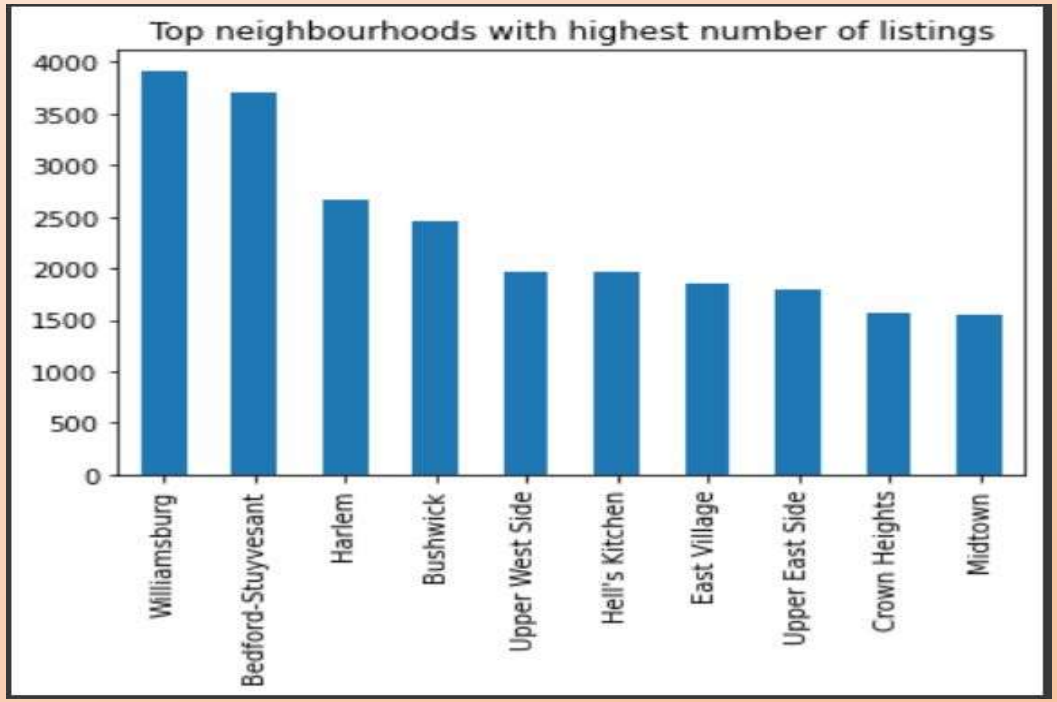
- As we can see in above fig. **'Michael'** has highest number or apartments, but as we know **'Name'** is not unique here so let's check by **'Id'**.

- In fig. above we can se 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 .

❖ EDA :

2. Top 10 neighbourhood having highest number of apartments.





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

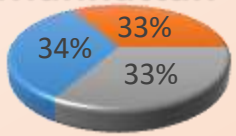
- We can see in fig. that "**Williamsburg**" in '**neighbourhood**' have maximum number of listing.

❖ EDA :

3. Top 3 neighbourhood having maximum price in each neighbourhood_group.

	neighbourhood	price
0	Upper West Side	10000
1	East Harlem	9999
2	Lower East Side	9999

Manhattan

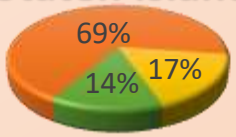


Upper West Side
East Harlem
Lowe East Side

- As we can see in neighbourhood_ group **'Manhattan'** there is approx. no different in

	neighbourhood	price
0	Randall Manor	5000
1	Prince's Bay	1250
2	St. George	1000

Staten Island

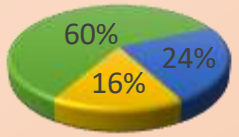


Randall Manor
Prince's Bay
St. George

- As we can see in neighbourhood_ group **'Staten Island'** there is much different in price of first

	neighbourhood	price
0	Riverdale	2500
1	City Island	1000
2	Longwood	680

Bronx

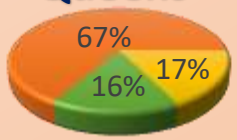


Riverdale
City Island
Longwood

see in neighbourhood_ group **'Bronx'** there is much different in price of all top three **'neighbourho**

	neighbourhood	price
0	Astoria	10000
1	Bayside	2600
2	Forest Hills	2350

Queens

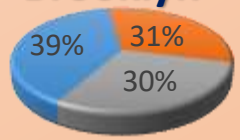


Astoria
Bayside
Forest Hills

group **'Staten Island'** it's much larger different in price of first and second **'neighbourho od'**.

	neighbourhood	price
0	Greenpoint	10000
1	Clinton Hill	8000
2	East Flatbush	7500

Brooklyn



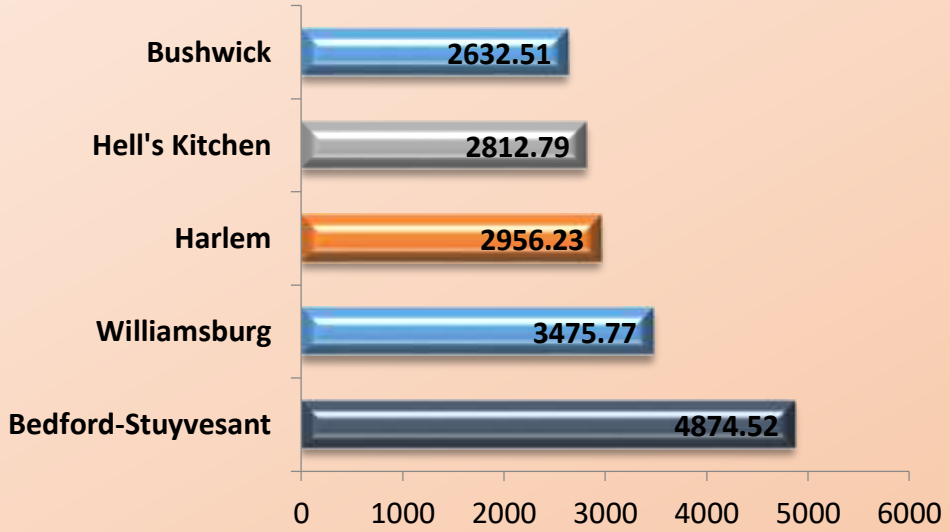
Greenpoint
Clinton Hill
East Flatbush

'Manhattan' there is not much different in price of top 3 three **'neighbourho od'**.

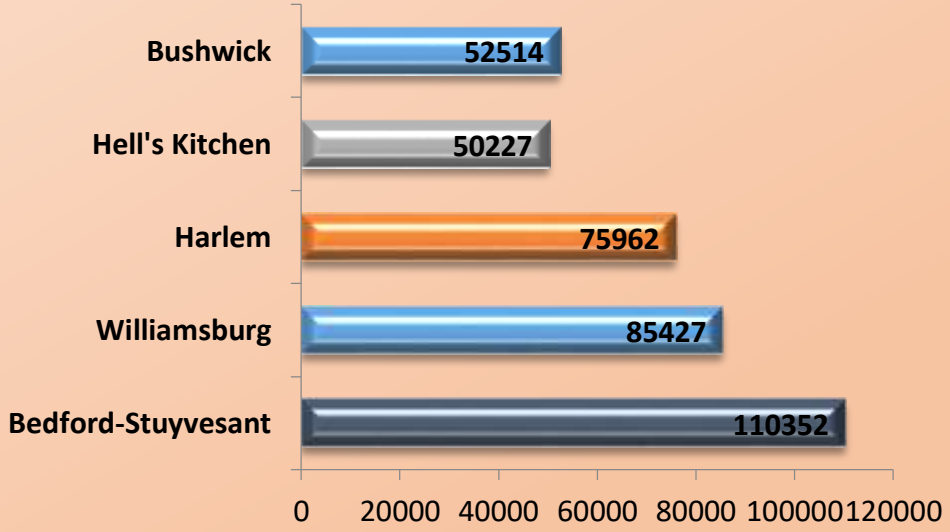
❖ EDA :

4. Relation between neighbourhood and reviews.

- In the data we can clearly see that the 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



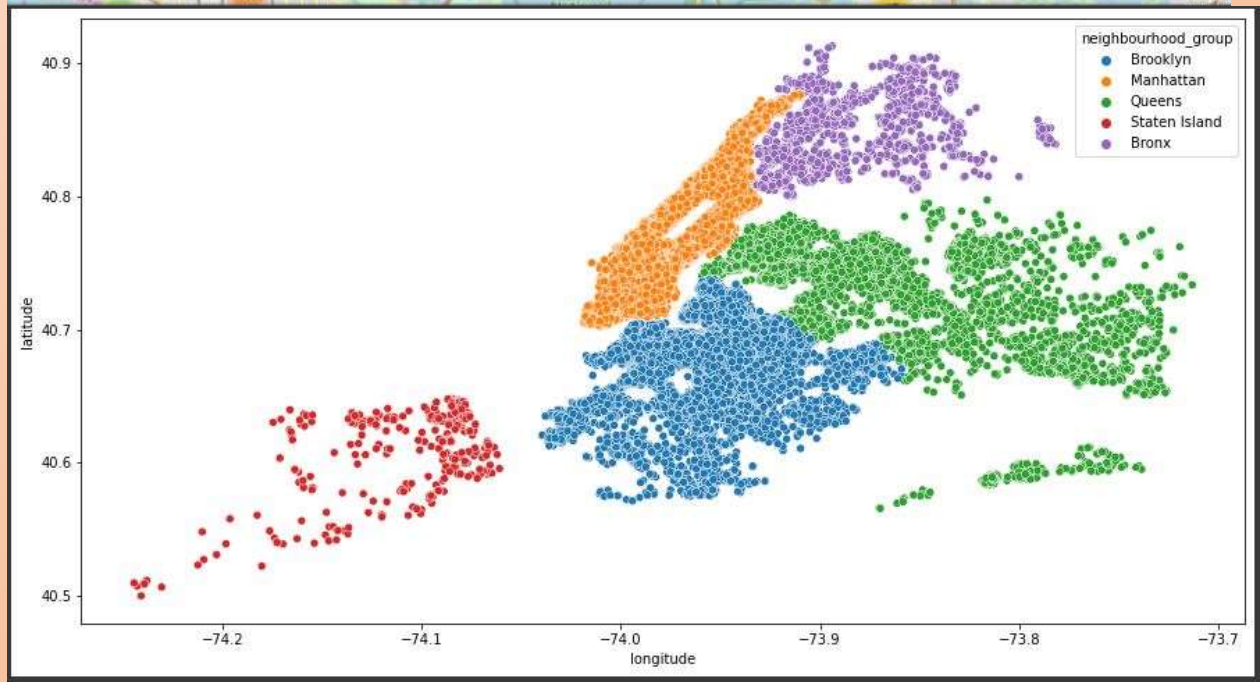
	neighbourhood	number_of_reviews
0	Bedford-Stuyvesant	110352
1	Williamsburg	85427
2	Harlem	75962
3	Bushwick	52514
4	Hell's Kitchen	50227

❖ EDA :

5. Learning from prediction(location).

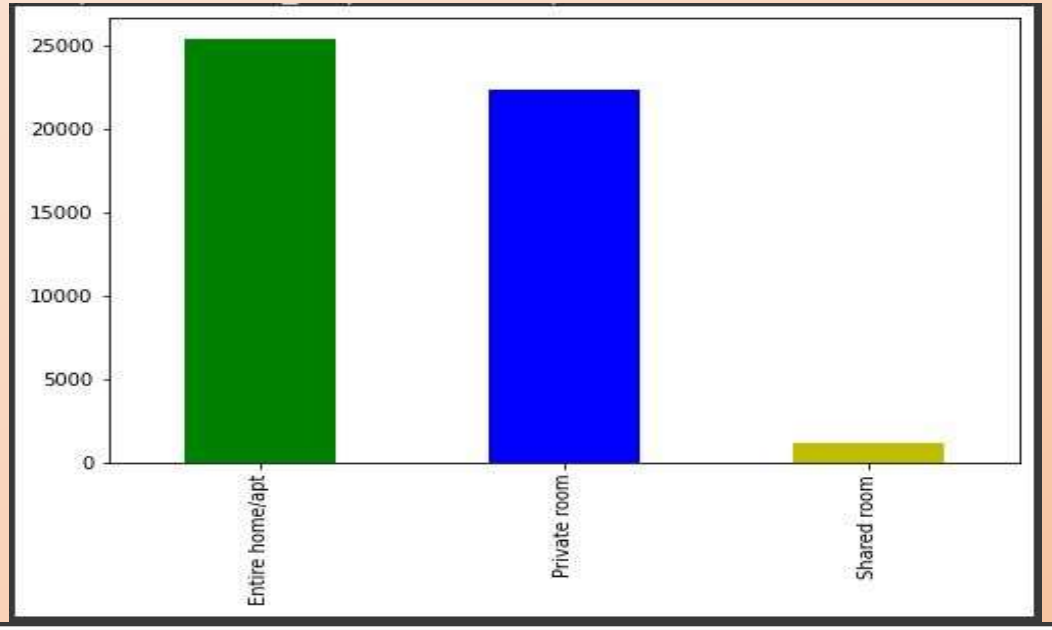
- Map shows the exact location of all the apartments with the help of **'latitude'** and **'longitude'** co-ordinates.

- Scatterplot 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.



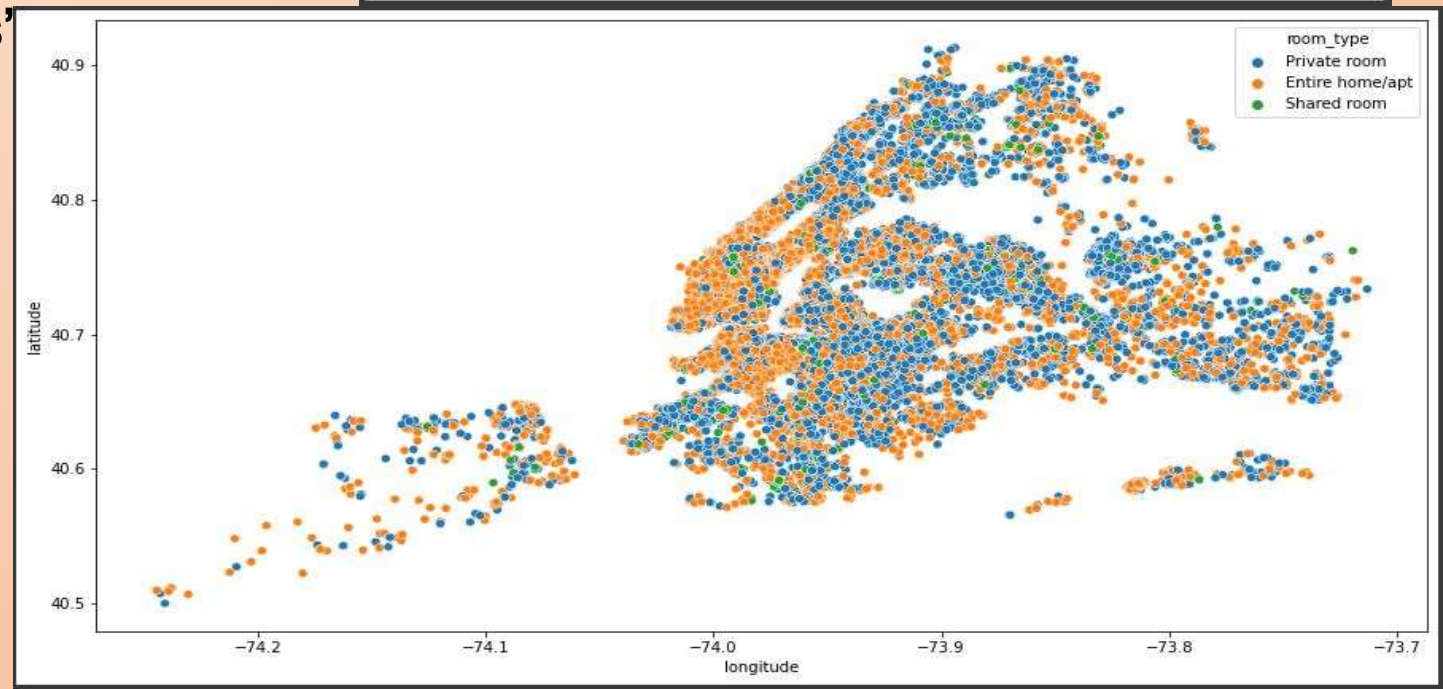
❖ EDA :

6. Room_type distribution over the location.



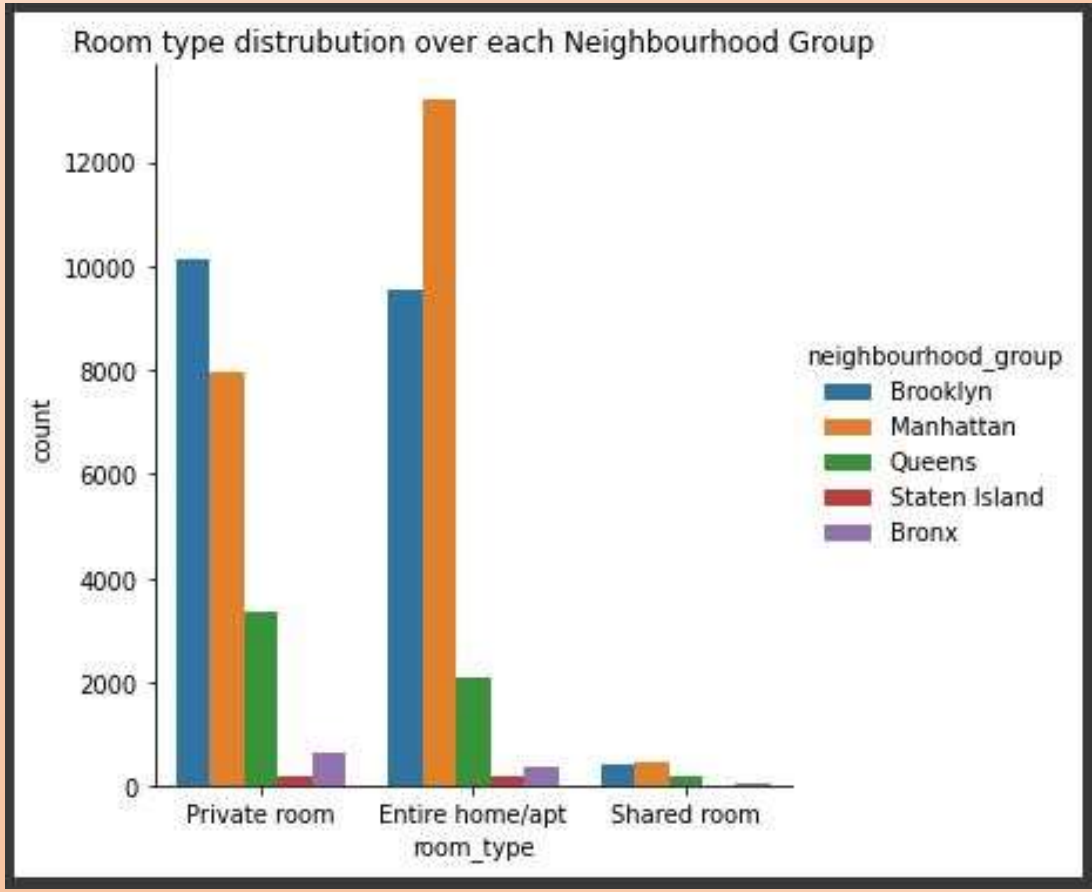
- Room type of most number of listing is **'Entire home/apartment',** but there is not big difference in '**private room and Entire home'.**
- But if we talk about **'Shared room'** we can see a huge difference here, and on the basis of this data we can say that **'Shared rooms'** preferable .

- Scatterplot 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 in every '**neighbourhood**'.



❖ EDA :

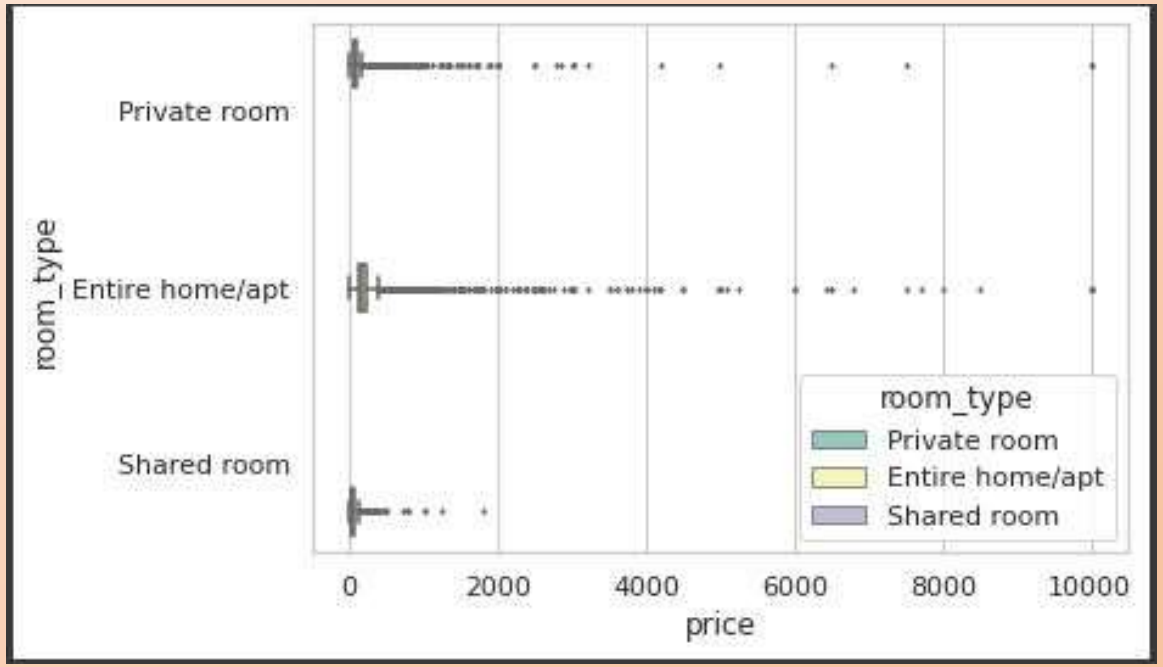
- With the help of this data we can understand the distribution of different **‘room_type’** in different **‘neighbourhood_group’**.
- According to data **‘private rooms’** are most located in **‘Brooklyn’** and **‘Entire home/apartment’** are most located in **‘Manhattan’** and **‘Shared rooms’** are most located in **‘Manhattan’**.
- 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 almost not preferable .
- **‘Brooklyn’** and **‘Manhattan’** are almost similar to each other but if we see other **neighbourhood_group** there is a huge different in



❖ EDA :

7. 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 lets 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.

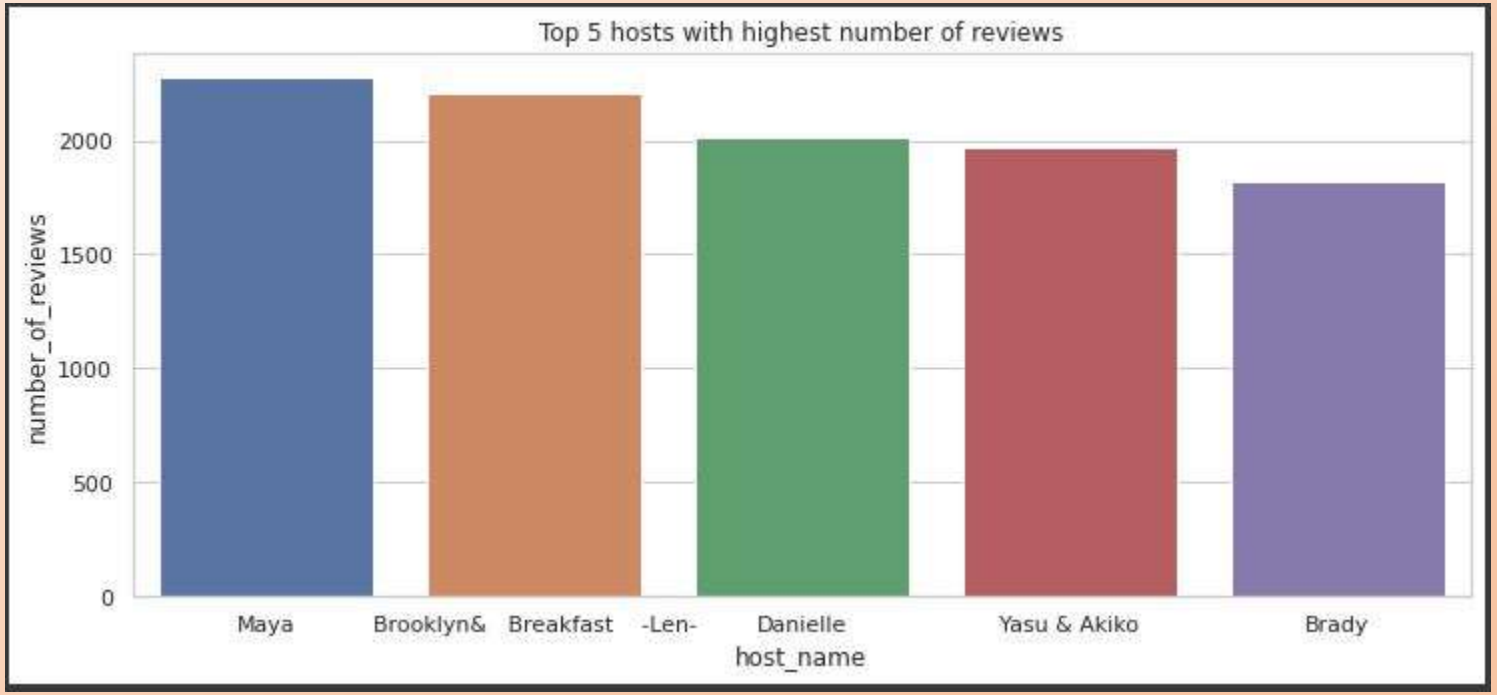
- Kathrine and Erin have so high price and having no availability then what is the benefit of keeping too high price.
- The last review is also 2-3 years back (as the data was collected in 2019) which is also bad
- The reviews may be low as there may be very few people who had stayed Kathrine's, Erin's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

	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

❖ EDA :

8. Top 5 host who received highest number of reviews.

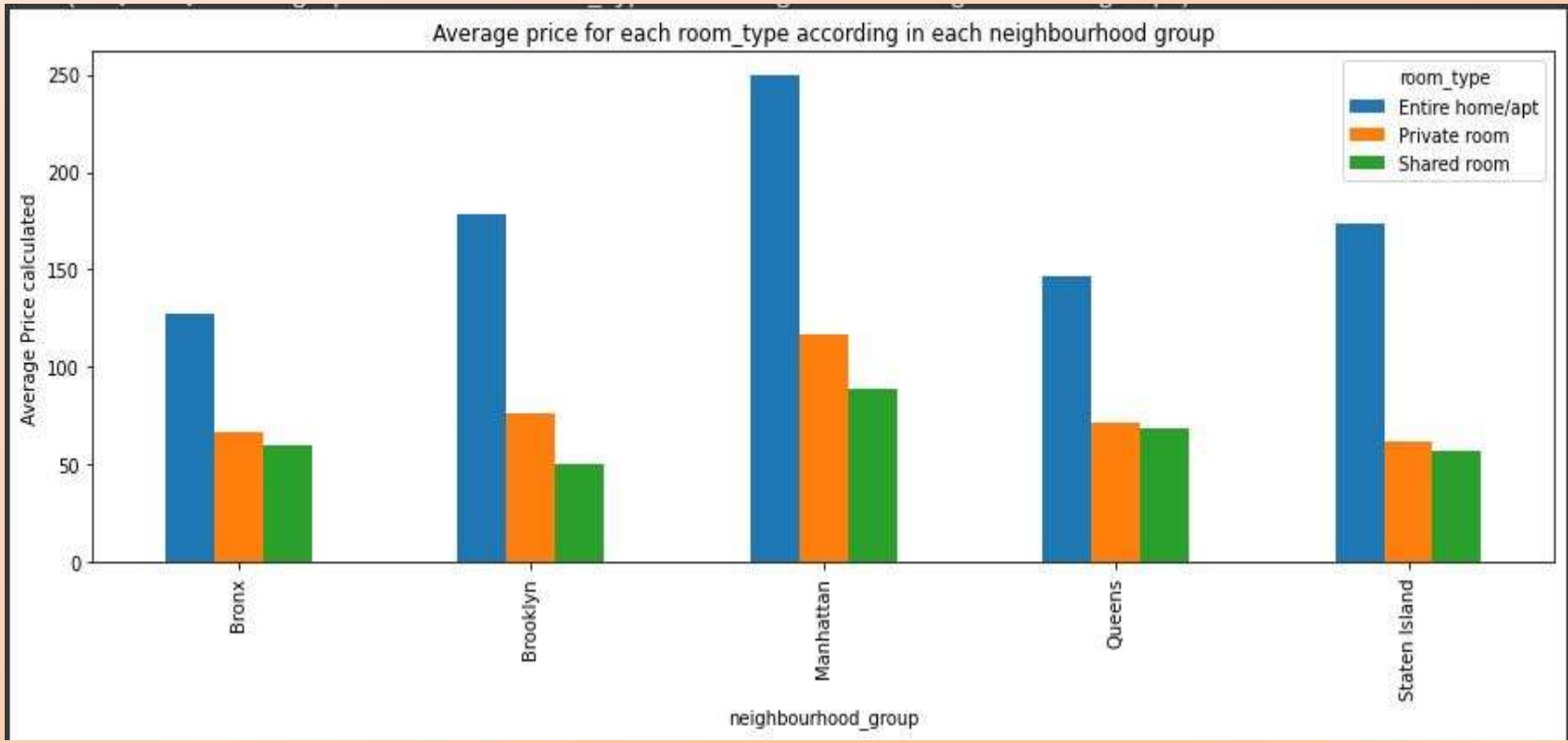
- We can see in all top 5 hosts there is not a big difference.
- We can assume with this data that all 5 host may be the busiest host according to review, but as we have already seen that some of the reviews are from past years that may infer that the homes are already closed or not preferable.
- But if we could filter the data then reviews can be a source to find busiest host.



❖ EDA :

9. Average price of each room_type in each neighbourhood_group.

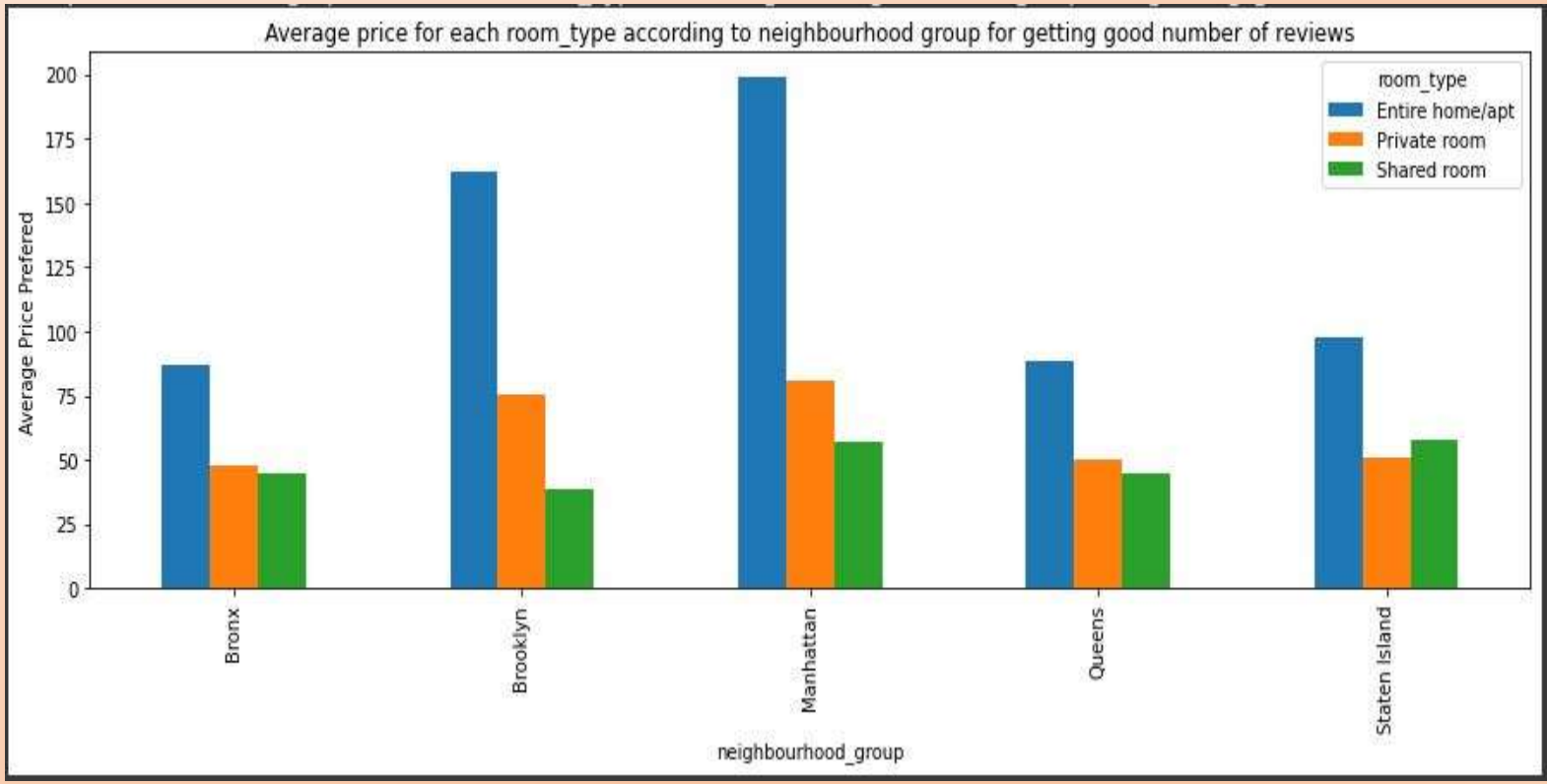
- **Observations:** As we can see that **Manhattan** is most costly and **bronx** is cheap for each **room_type**.
- But I think we can make it more useful for business implementation if we do some analysis on successful hosts according to the highest no of reviews so that we can suggest that price to our hosts to get good business.
- We seen before that shared rooms are not preferable here we can see one that



❖ EDA :

10. Average price preferred to get good number of reviews.

- **Observations:**
  - Clearly if we compare the results with previous result **(i.e when we calculated average price preferred by people in each neighbourhood_group with different room_types)** we can see that this result is bit different and more useful.
  - As an analyst we would suggest to keep price in this range to get more number of reviews in specific room type and at a particular place.



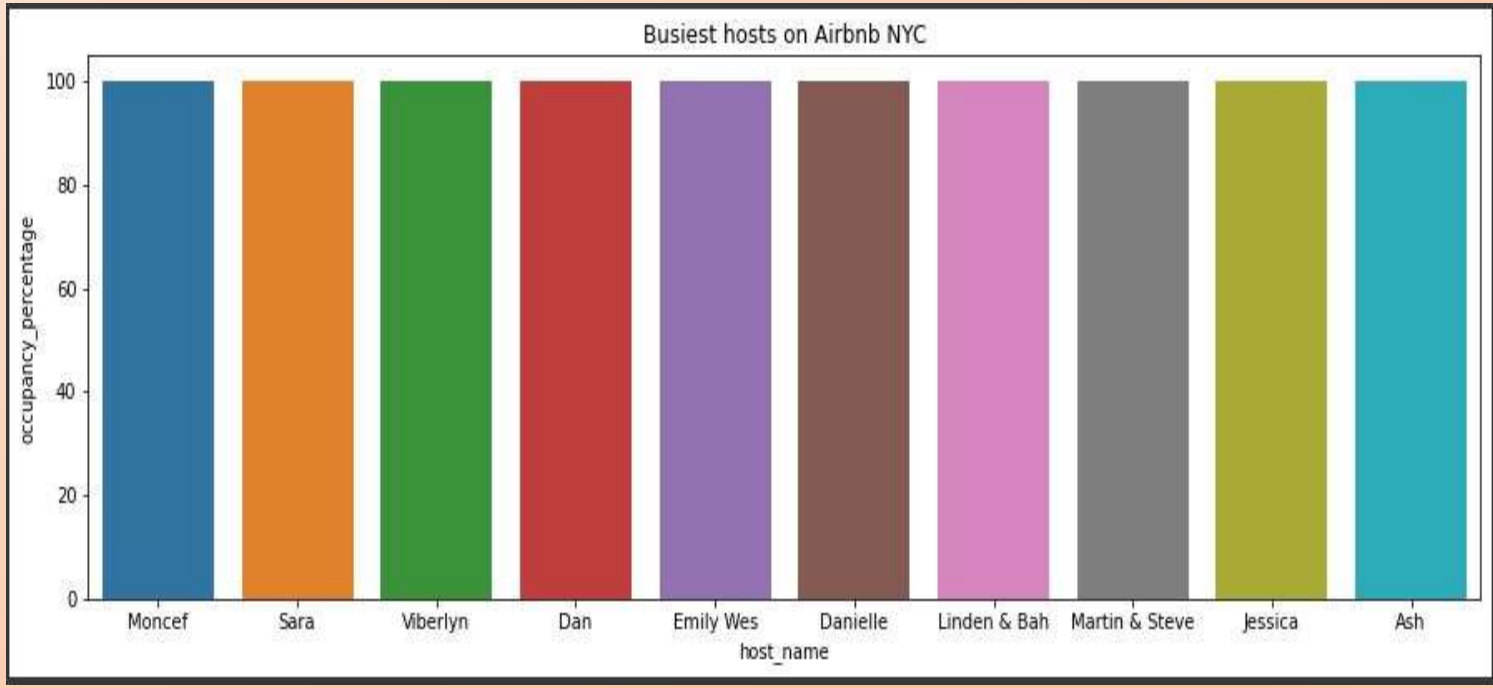
❖ EDA :

11. Top 10 busiest host.

Observations:

The `host_name` column as discussed bef ore contains
repeating names, the `calculated_host_listi ngs_count`
gives us a fair idea of how many properties in total a host
owns. The use of `host_id` as primary key for the groupby
function is the only way. Followed by `host_ name` for
groupby since the names of the hosts is im portant in the
analysis.

The mean is used as the aggregating functi on for the
`occupancy_percentage` and price column



❖ EDA :

Formula use to find Top 10 busiest host.

- A metric is a system of measurement in this case **'busiest'** which gives a relative comparison between the hosts. The metric defined below is a proxy to estimate the busyness of a host. The metric here using the fore-mentioned columns estimates the percentage of occupancy the property has seen in one period of business. The metric mean across various properties for a host gives the average occupancy rate/percentage the host. The higher the percentage, the busier a host is said to be.
- The 3 columns afore-mentioned are taken into consideration for calculating this metric.
- Firstly, the metric needs the available months (one period of business) the host is open for business/accepting bookings :-
  - "available months = available days / (365/12)"
- For the given months the property is open for business, next is to estimate the maximum possible bookings a property can have through the available days, here the assumption is that, every customer stays exactly equal to minimum nights required by the listing :-
  - "total possible bookings = available days / minimum nights"
- The next step is to estimate the actual number of bookings that occurred in the year. The assumption made here is that the number of reviews received per month is analogous to that many customers on average booked/stayed in this property. Hence we will estimate bookings as :-
  - "estimated bookings = reviews per month x available months"
- Using all the above calculations, the percentage of occupancy throughout the year is gives as :-
  - "Occupancy % = estimated bookings / total possible bookings x 100"
- **POINT TO NOTE:-** According to the assumptions and calculations done above to calculate the metric, a property with 1 customer over the entire period of business as the property's total possible booking records a 100% when the estimated bookings is also 1. In simpler terms, if the expected booking count is calculated to be 1 and the property hosts 1 customer, then the property is said to be 100% busy.

❑ "The **host_name** column as discussed before contains repeating names, the **calculated_host_listings_count** gives us a fair idea

❖ Challenges faced :

- ✓ While doing 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.
- ✓ Further, 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.
- ✓ There are many listings whose date of the last review is very old this can mean that they must've 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 were many outliers in the price column of some hosts which weren't 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. For example, if we check that one host has x number of reviews which is highest but when we check the date of last_review and find out that the reviews are very old than the current date, then we can infer that business is currently shutdown so how can we take such hosts into consideration for knowing the busiest hosts. Ideally, the busiest host should be that one whose occupancy is almost full or full..

❖ 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.
- ✓ Maximum listings are listed on Manhattan and Brooklyn neighbourhood_groups. Staten Island and Bronx neighbourhood_group have very less numbers of listings.
- ✓ 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.
- ✓ Keeping high price of the listing and have 0 availability isn't benefitting the host as the consumer is ready to pay the price but even after that there are no available rooms then what's the benefit of paying such a premium.
- ✓ Maya (host) has the heighest total number_of_reviews.
- ✓ Average prices of all the room types in Manhattan are more than the average price of

Thank you

"There is no great genius without some touch of madness." **– Seneca**