



CAPSTONPROJECT -1

AIRBNB BOOKING ANALYSIS
(EDA)

TABLE OF CONTENT

:

- ▶ Introduction to 'Airbnb'
- ▶ Data description and value
- ▶ Exploratory data analysis
- ▶ Difficulty faced and their solution
- ▶ Conclusion





INTRODUCTION :

- Airbnb was founded in 2008 by Brian Chesky , Nathan Blecharczyk , Joe Gebbia.
- Airbnb based in San Francisco, California, operates an online marketplace focused on short-term homestays and experiences. The company acts as a broker and charges a commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com. The company has been the subject of criticism for lack of regulations and enabling increases in home rents. Airbnb based in San Francisco, California, operates an online marketplace focused on short-term homestays and experiences. The company acts as a broker and charges a commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com. The company has been the subject of criticism for lack of regulations and enabling increases in home rents.

- Airbnb features a review system in which guests and hosts can rate and review each other after a stay. Hosts and guests are unable to see reviews until both have submitted a review or until the time period to review has closed, a system that aims to improve accuracy and objectivity by removing fears that users will receive a negative review in retaliation if they write one. However, the truthfulness and impartiality of reviews may be adversely affected by concerns of future stays because prospective hosts may refuse to host a user who generally leaves negative reviews
- Here we are going to do an exploratory data analysis on the data set of guest and host (2019)
- This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.
- The main objective of analysis will be some of statement given to us which can be briefed as learnings from host, areas, price, reviews, locations and other learning objective and also try to explore some more insights.

Data description :

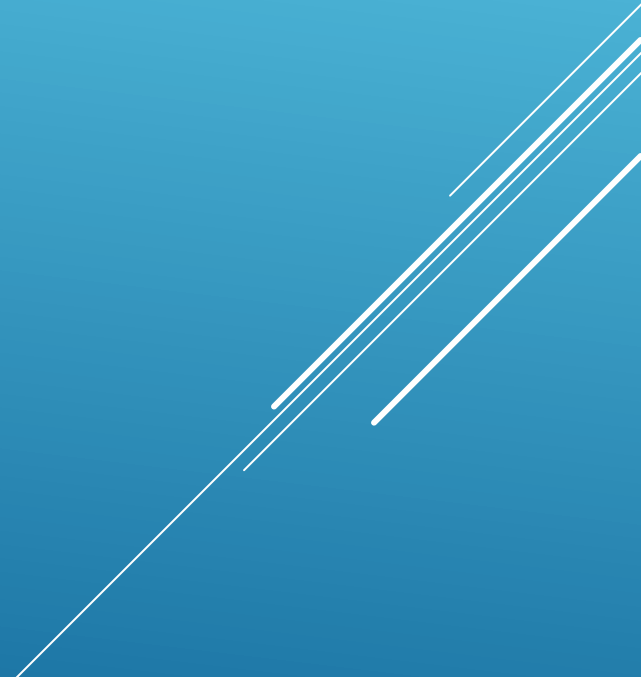


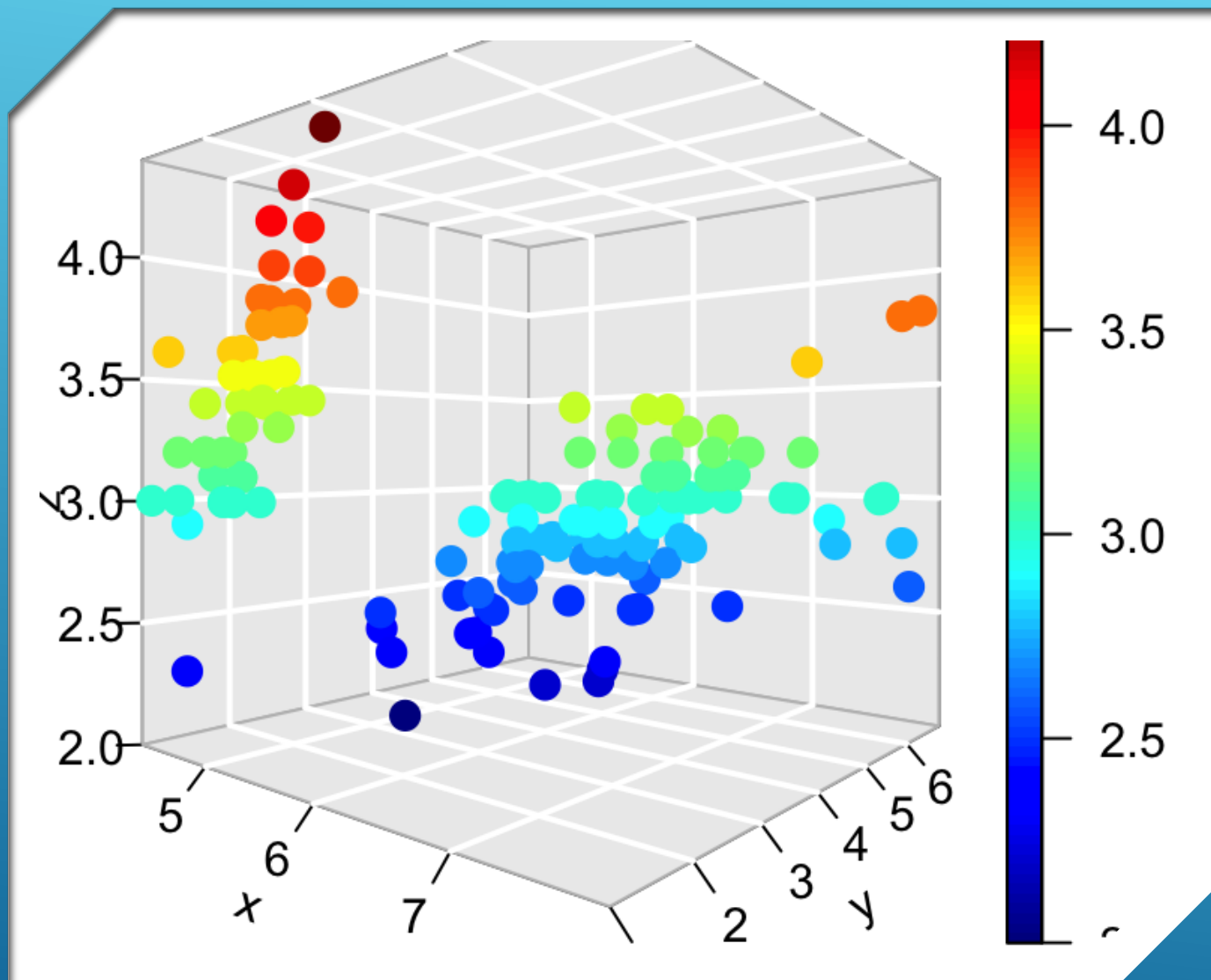
VALUES INFORMATION (VARIABLES USE IN AIRBNB DATASET) :

- Id : it's unique id for house/apartment.
- Name: name of listing house/apartment
- Host id: host id is government approved id for each individuals who rent their property on Airbnb
- Host name: host name is the name of the individual or organization who own room/apartment on Airbnb
- Neighborhood groups: Neighborhood groups are the cluster of neighborhoods in the area. There are about 5 brought in the state
- Neighborhood: when searching for accommodation in a city, guests are able to filter by neighborhood attributes and explore layers of professional quality content, including neighborhood 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 spaces are entire house/apartment, private room shared room.
 - Price(\$): the total Price of Airbnb reservation is based on the rate set by the host, plus fees 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
 - Review per month: number of review host get per month.
 - Calculated host listing count: amount of listing per host.
 - Availability 365: it is indicator of days the listing is available for during the year.
- 

EXPLORATORY DATA ANALYSIS(EDA) :

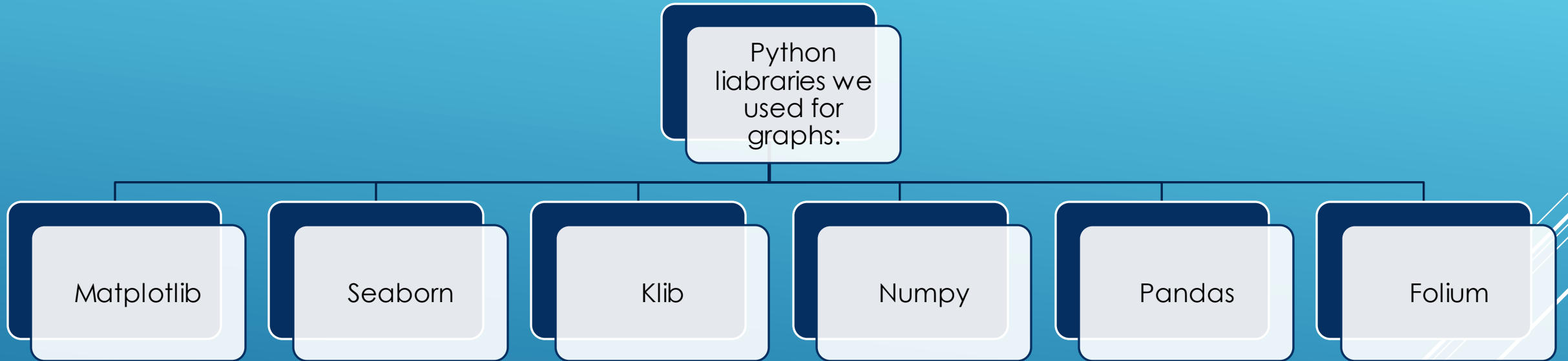
- WHAT IS EDA?
 - "Exploratory data analysis" is very important in machine learning. Whenever we start our work on any project we must analyze the factors deeply. Hypothetical questions and that hypothetical questions lead to some hidden facts. This collaborative work is simply known as EDA
 - The following steps are involved in the process of EDA:
 - i. Acquire and Loading data
 - ii. Understanding the variable
 - iii. Cleaning dataset
 - iv. Exploring and visualizing the data
 - v. Analyzing relationship between variables
- 
- A series of white diagonal lines of varying lengths and thicknesses are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.



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 :



Exploratory Data Analysis on Dataset :

Now we will analyze the data from relevant dataset question and their answer, the question as follow:

1. Which hosts are having highest number of apartments ?
2. Which are the top 15 neighborhood which are having maximum number of apartments for Airbnb in the respective neighborhood ?
3. top 5 neighborhood in each group which are having maximum prices in their respective neighborhood group ?
4. How neighborhood is related with reviews ?
5. What can we learn from predictions? (ex: locations, prices, reviews, etc.) location learning
6. What is the distribution of the room type and its distribution over the location ?
7. How does the Room type is distributed over neighborhood Group are the ratios of respective room types more or less same over each neighborhood group ?
8. How the price column is distributed over room type and are there any Surprising items in price column ?
9. Which are the top 10 hosts that have obtained highest no. of reviews ?
10. What is the average preferred price by customers according to the neighborhood group for each category of Room type?
11. What is the average price preferred for Keeping good number of reviews according to neighborhood group ?
12. Which neighborhood have most of the host focused to own property ?
13. Overall price analysis?
14. Overall room description?

EDA :

1. First we view the all null values in dataset.
 - As we can see the data column 'last review' and 'review per month' having most number of null values.
 - Approx 21% last review and review per month are null in dataset
2. Lets check the relation between data column.
 - 'host id and id' , 'review per month and number of reviews' they relate to each other in dataset.
 - 'Number of review and id' also relate to each other.

```
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
```

EDA :

id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
le+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
le+07	6.762001e+07	40.728949	-73.952170	152.739094	7.029962	23.274466	1.373221	7.143982	112.781327
e+07	7.861097e+07	0.054530	0.046157	240.146276	20.510550	44.550582	1.680442	32.952519	131.622289
le+03	2.438000e+03	40.499790	-74.244420	10.000000	1.000000	0.000000	0.010000	1.000000	0.000000
ie+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
le+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
le+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
le+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

- There is approx. 35% of 'availability 365' data is '0' is practically closed room /apartment if you have a business providing stay on Airbnb the availability is '0' day that is an extreme case
- Let's check the last review either that house is still open or closed.
- The dataset is period of '2011-19' and we can see in the last review there is some review which was delivered in 2016 , 2017 that means listing is already closed or not preferred.
- Filling the price table on behalf of mean price of same 'room type' holding more price than '0'.

```
last_review
2019-01-01    194
2018-01-01    142
2019-01-02    129
2019-06-23     90
2018-01-02     86
2017-01-01     85
2019-05-27     75
2017-01-02     73
2016-01-02     67
2019-07-01     63
dtype: int64
```


EDA :

1. Which host having highest number of apartment ?

- From table we can see that host name Michael its appearing 417 times in the host_name column , so this might imply that Michael is having highest number of rooms , but from the host_id column its showing highest appearance of any host_id is 327 , so this clearly implies that there can be multiple person may have same name that's why we are getting different highest appearance in host_name as compared to host_id
- As per host_id Sonder (NYC) is having maximum numbers of rooms for the guest, For Airbnb he might be very important person then.
- By help of python we identify host that having highest number of apartment and his detail which will be helpful for Airbnb to provide proper facility to most important host and encourage him for more business.

Michael	417	219517861	327
David	403	107434423	232
Sonder (NYC)	327	30283594	121
John	294	137358866	103
Alex	279	16098958	96
...		...	
Rhonycs	1	23727216	1
Brandy-Courtney	1	89211125	1
Shanthony	1	19928013	1
Aurore And Jamila	1	1017772	1
Ilgar & Aysel	1	68119814	1
Name: host_name, Length: 11452, dtype: int64		Name: host_id, Length: 37457, dtype: int64	

As per host_name

As per host_id

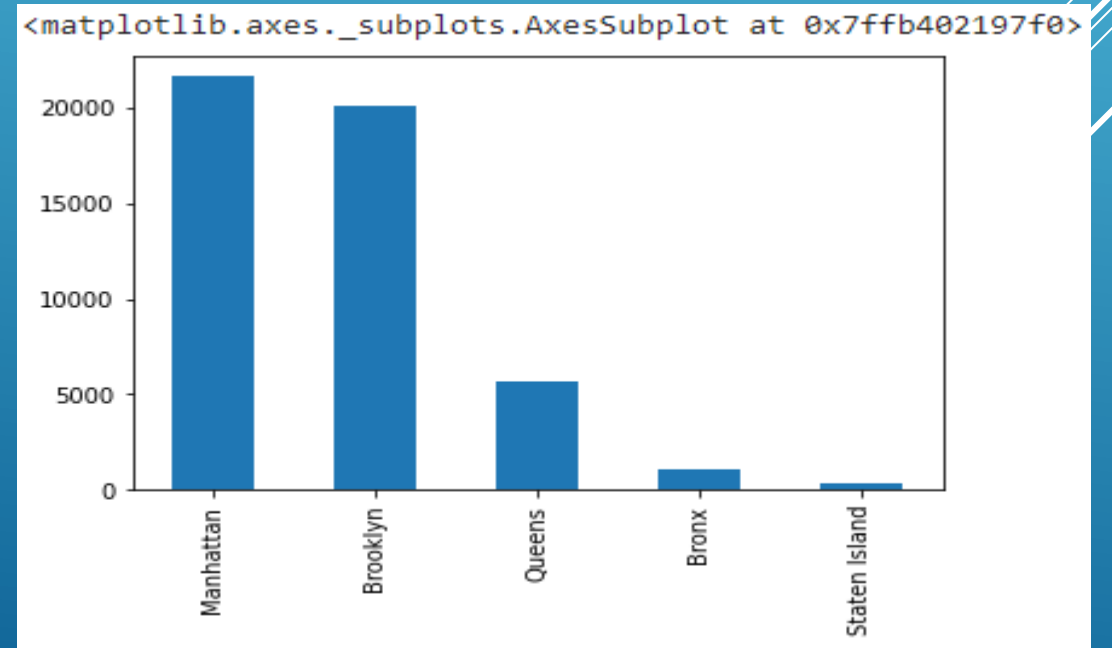
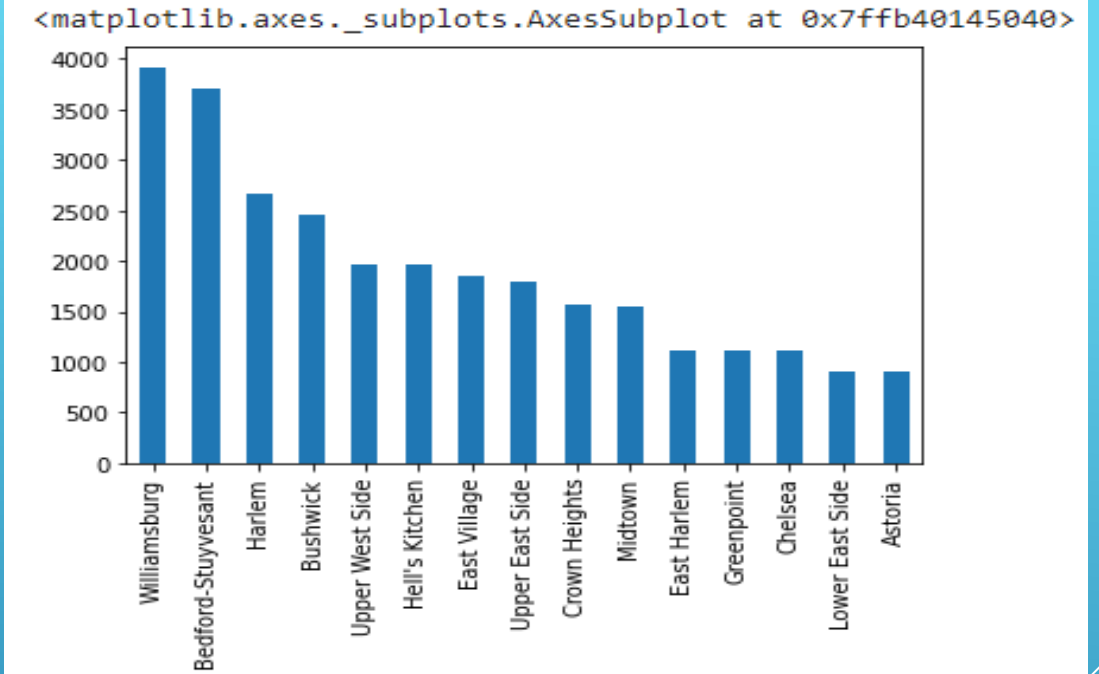
	host_name	neighbourhood_group	neighbourhood	latitude	longitude
38293	Sonder (NYC)	Manhattan	Financial District	40.70637	-74.00645
38294	Sonder (NYC)	Manhattan	Financial District	40.70771	-74.00641
38588	Sonder (NYC)	Manhattan	Financial District	40.70743	-74.00443
39769	Sonder (NYC)	Manhattan	Murray Hill	40.74792	-73.97614
39770	Sonder (NYC)	Manhattan	Murray Hill	40.74771	-73.97528
39771	Sonder (NYC)	Manhattan	Murray Hill	40.74845	-73.97446
39772	Sonder (NYC)	Manhattan	Hell's Kitchen	40.76188	-73.99616
39773	Sonder (NYC)	Manhattan	Hell's Kitchen	40.76037	-73.99744
39774	Sonder (NYC)	Manhattan	Murray Hill	40.74884	-73.97589
39775	Sonder (NYC)	Manhattan	Hell's Kitchen	40.76079	-73.99807

Host having highest number of apartment

EDA :

2. Which are the top 15 neighborhood which are having maximum number of apartments for Airbnb ?

- As per data Williamsburg, Bedford Stuyvesant, Harlem, Bushwick, upper west side, crown heights, midtown, east Harlem, Greenpoint, Chelsea, lower east side, astoria this are top 15 neighborhood which having highest number of apartment.
- As per data Manhattan has highest neighborhood then as follow Brooklyn, queens, Bronx, Staten island.
- As per data Williamsburg having highest neighborhood and Manhattan having highest number of neighborhood group which will helpful information for Airbnb to give more workforce to these places for more crowd handling



EDA :

3. top 5 neighborhood in each group which are having maximum prices in their respective neighborhood group ?

neighbourhood	price	neighbourhood	price	neighbourhood	price	neighbourhood	price	neighbourhood	price
0 Upper West Side	10000	0 Randall Manor	5000	0 Riverdale	2500	0 Astoria	10000	0 Greenpoint	10000
1 East Harlem	9999	1 Prince's Bay	1250	1 City Island	1000	1 Bayside	2600	1 Clinton Hill	8000
2 Lower East Side	9999	2 St. George	1000	2 Longwood	680	2 Forest Hills	2350	2 East Flatbush	7500
3 Tribeca	8500	3 Fort Wadsworth	800	3 Westchester Square	670	3 Long Island City	2000	3 Bedford-Stuyvesant	5000
4 Upper East Side	7703	4 Woodrow	700	4 Mott Haven	500	4 Arverne	1500	4 Cypress Hills	5000
In Manhattan upper west side has highest price then east Harlem, lower east side, Tribeca, Upper east side are follow.		In Staten island Randall mirror has highest price then prince's Bay, St. George, Fort Wadsworth, Woodrow are follow.		In Bronx Riverdale has highest price then City island, longwood, Westchester square, mott haven are follow.		In queens Astoria has highest price then Bayside, Forest Hills, long island city, Arvene are follow		In Brooklyn Greenpoint has highest price then Clinton hill, East Flatbush, Bedford-Stuyvesant, Cypress hills are follow.	

EDA :

4. How neighborhood is related to reviews ?

- As per review per month theater district has highest number of reviews receive in month as per this information we can say theater district has highest number of host.
- As per number of review Bedford-Stuyvesant has highest number of reviews receive overall as per the overall review we also say that Bedford- Stuyvesant has highest number of host
- This will not be accurate busiest host but we consider this when there is no alternate option to finding busiest host.

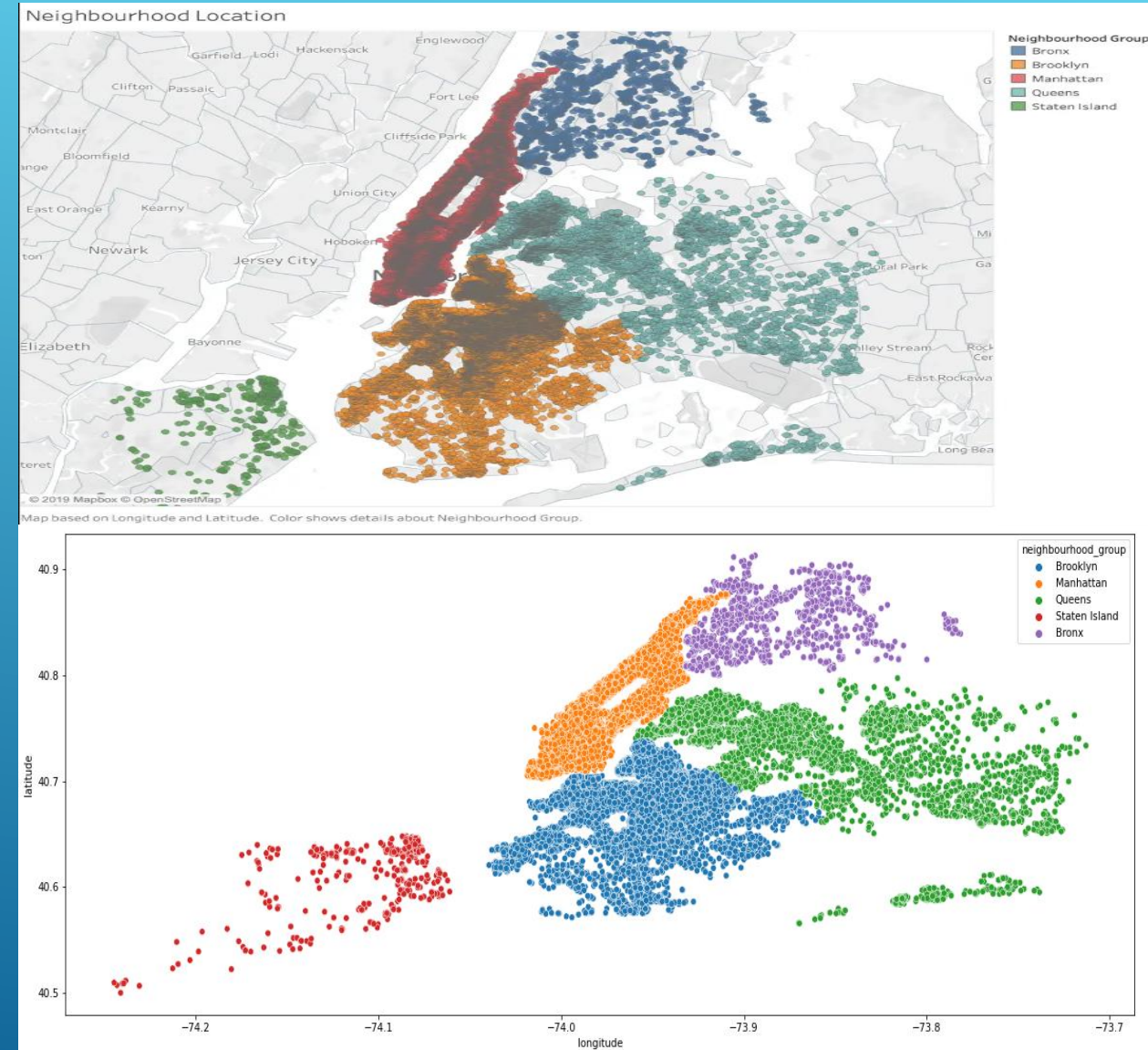
	neighbourhood	reviews_per_month
0	Theater District	58.50
1	Rosedale	20.94
2	Springfield Gardens	19.75
3	East Elmhurst	16.22
4	Jamaica	15.32
5	Williamsburg	14.00
6	Hell's Kitchen	14.00

	neighbourhood	number_of_reviews
0	Bedford-Stuyvesant	110352
1	Williamsburg	85427
2	Harlem	75962
3	Bushwick	52514
4	Hell's Kitchen	50227
5	East Village	44670
6	East Harlem	36446

EDA :

5. What can we learn from predictions? (ex: locations, prices, reviews, etc.) location learning

- Map shows exact location of apartments based on the longitude and latitude co-ordinates.
- We use the scatterplot on map to find out cluster of neighborhood group with the help of latitude and longitude.
- As per scatterplot we can easily say that 'Manhattan' is most dense area which has maximum number of listing.
- 'Staten island' is least dense area which has least number of listing.

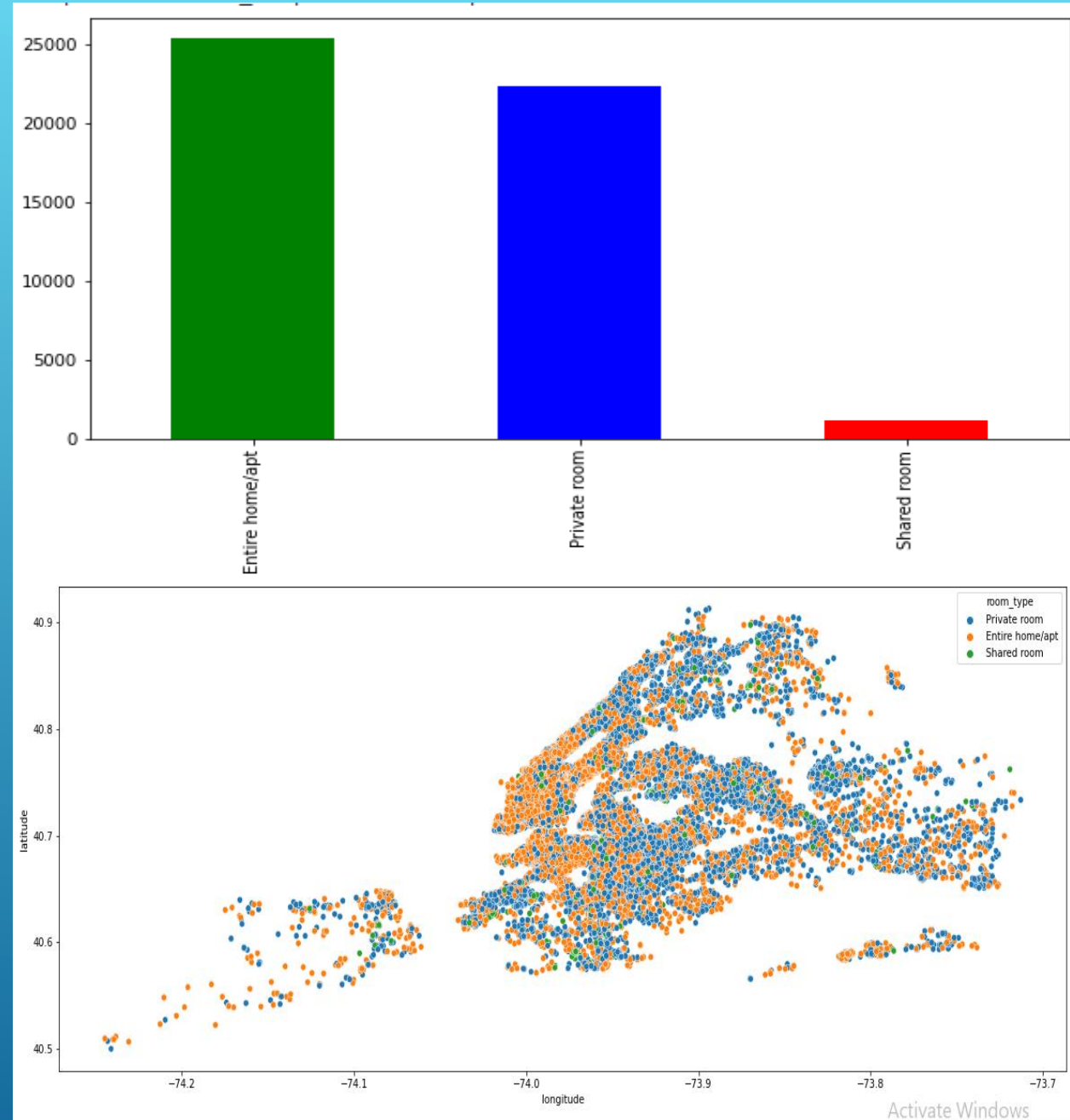


EDA :

6. What is the distribution of the room type and its distribution over the location ?

So we can notice the following

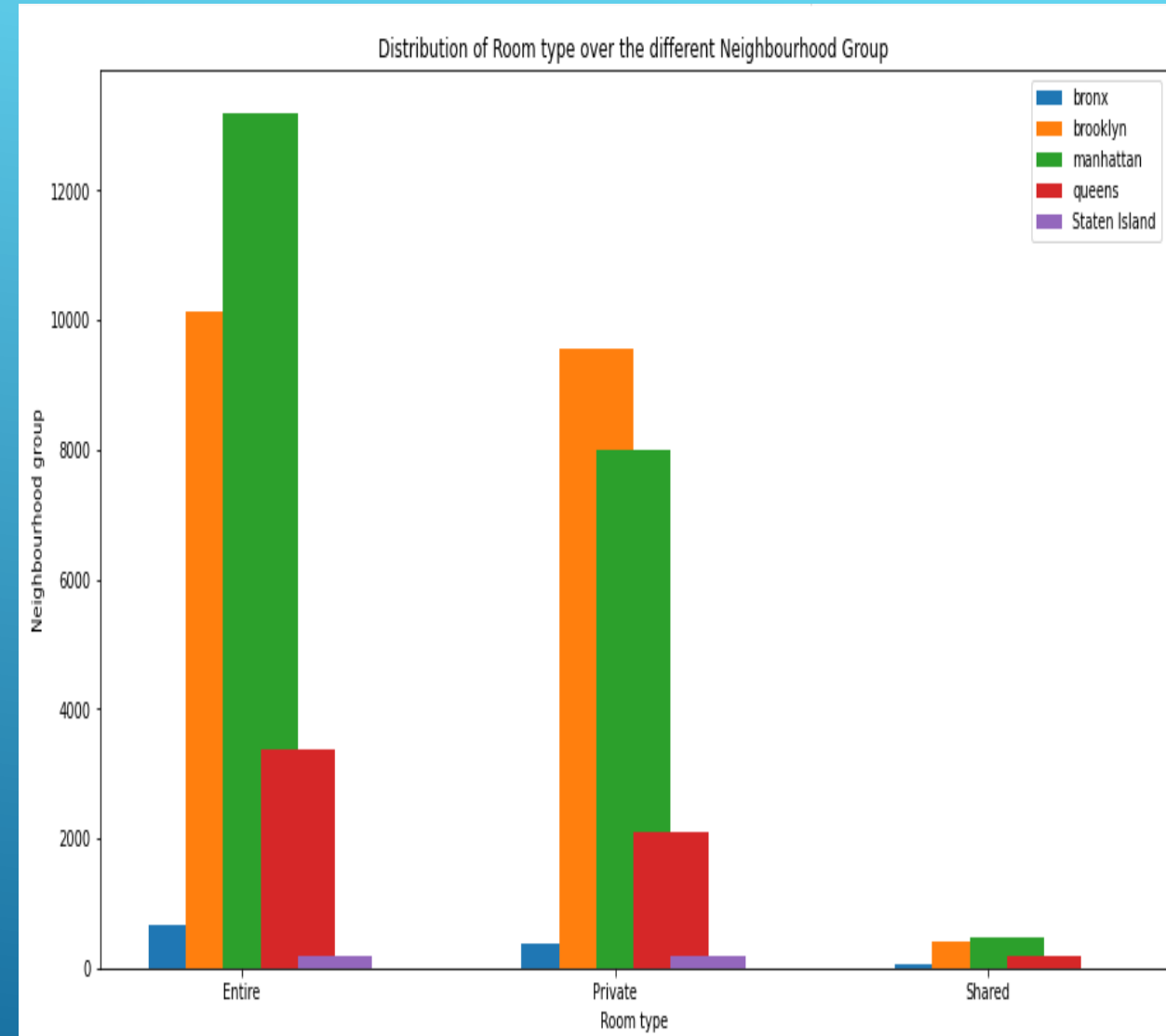
- that maximum numbers of room are Entire home/Apartment and Private room there are only few shared rooms.
- So mostly host prefer to give Entire home/Apartment or Private Rooms rather than Shared rooms.
- As we can see scatterplot on map show the cluster of room type.
- We can also say that room type is almost same in every neighborhood, which mean the booking of room type is almost same in every neighborhood.



EDA :

7. How does the Room type is distributed over neighborhood Group are the ratios of respective room types more or less same over each neighborhood group ?

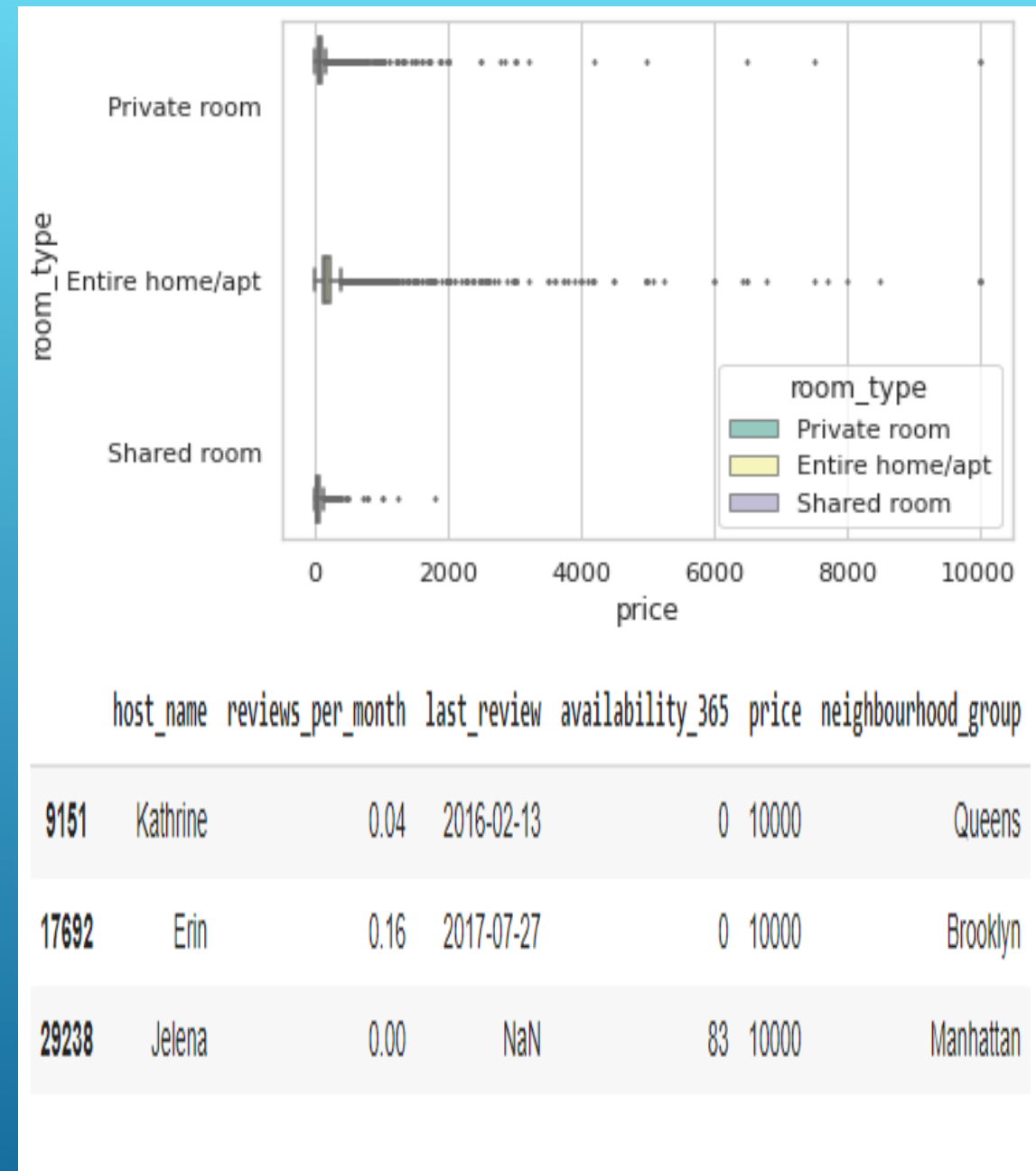
- As we can see that ratio of entire, private and share room is same in every neighborhood group.
- Where entire room has maximum demand and then private room and followed by shared room.
- And bar also show the density of room type in every neighborhood group, where in entire room Manhattan is leading and then Brooklyn.
- In private room type Brooklyn have highest density then followed by Manhattan.
- In shared room type Manhattan is leading and followed by Brooklyn, by this we can conclude that shared room have high demand in dense area.



EDA :

8. How the price column is distributed over room type and are there any Surprising items in price column ?

- we can notice that there are many outliers for price in each of the room type category, where we see that entire room has highest price range then followed by private and shared rooms.
- I would have suggested the following to Airbnb
- Kathrine and Erin have price so high 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 review may be low as there may be very few people who is staying in Kathrine, Erin and Jelena apartment so might have less reviews per month
- I would have suggested to keep moderate (average) price so that more people would visit and stay in her apartment, it would also increase her reviews per month



EDA :

9. Which are the top 10 hosts that have obtained highest no. of reviews ?

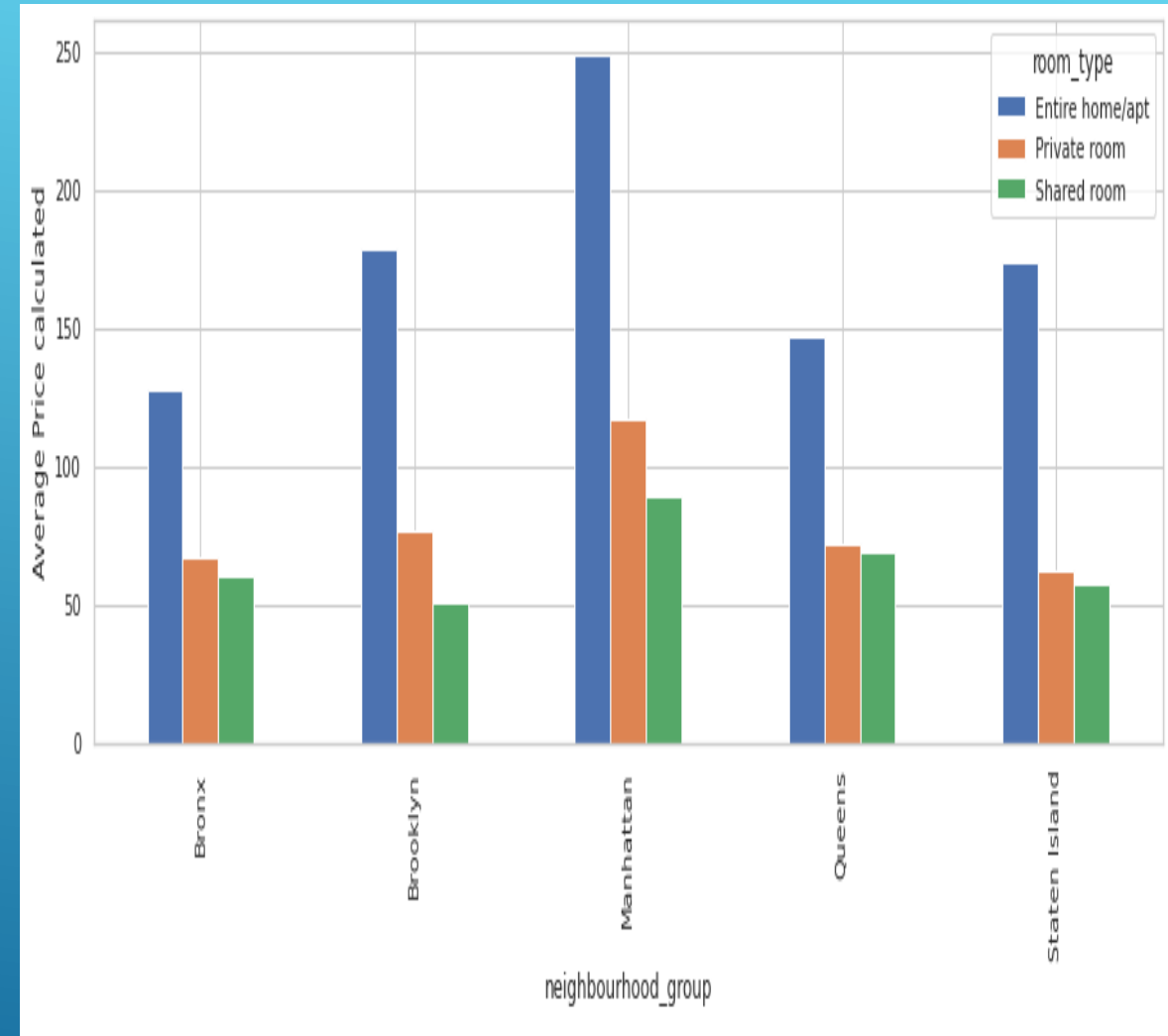
- Top 10 host that have highest number of reviews as follow Maya, Brooklyn& Breakfast Len, Danielle, Yasu & Akiko Brady, ji, Alex And Zeena, Randy, sonder(NYC), Angela.
- We can assume that all 10 host may be the busiest host according to review, but as we have already seen that some of host having multiple room which can affect the number of review and also the reviews are from past year that may infer that the homes are already closed or not preferable.
- If we could filter the data then reviews can source to find the busiest host.

	host_id	host_name	number_of_reviews
21304	37312959	Maya	2273
1052	344035	Brooklyn& Breakfast -Len-	2205
18626	26432133	Danielle	2017
20872	35524316	Yasu & Akiko	1971
21921	40176101	Brady	1818
7362	4734398	Jj	1798
14707	16677326	Alex And Zeena	1355
9201	6885157	Randy	1346
34629	219517861	Sonder (NYC)	1281
17517	23591164	Angela	1269

EDA :

10. What is the average preferred price by customers according to the neighborhood group for each category of Room type?

- Observations: As we can see that Manhattan is most costly and Bronx is cheap for each room type
- As 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 this price to our host for good business.
- We see before that shared room is not preferable, we can see one of the reason is price of room, where the difference between private and shared room price is almost same in most of neighborhood groups that's why most of customer preferred private room over shared room.

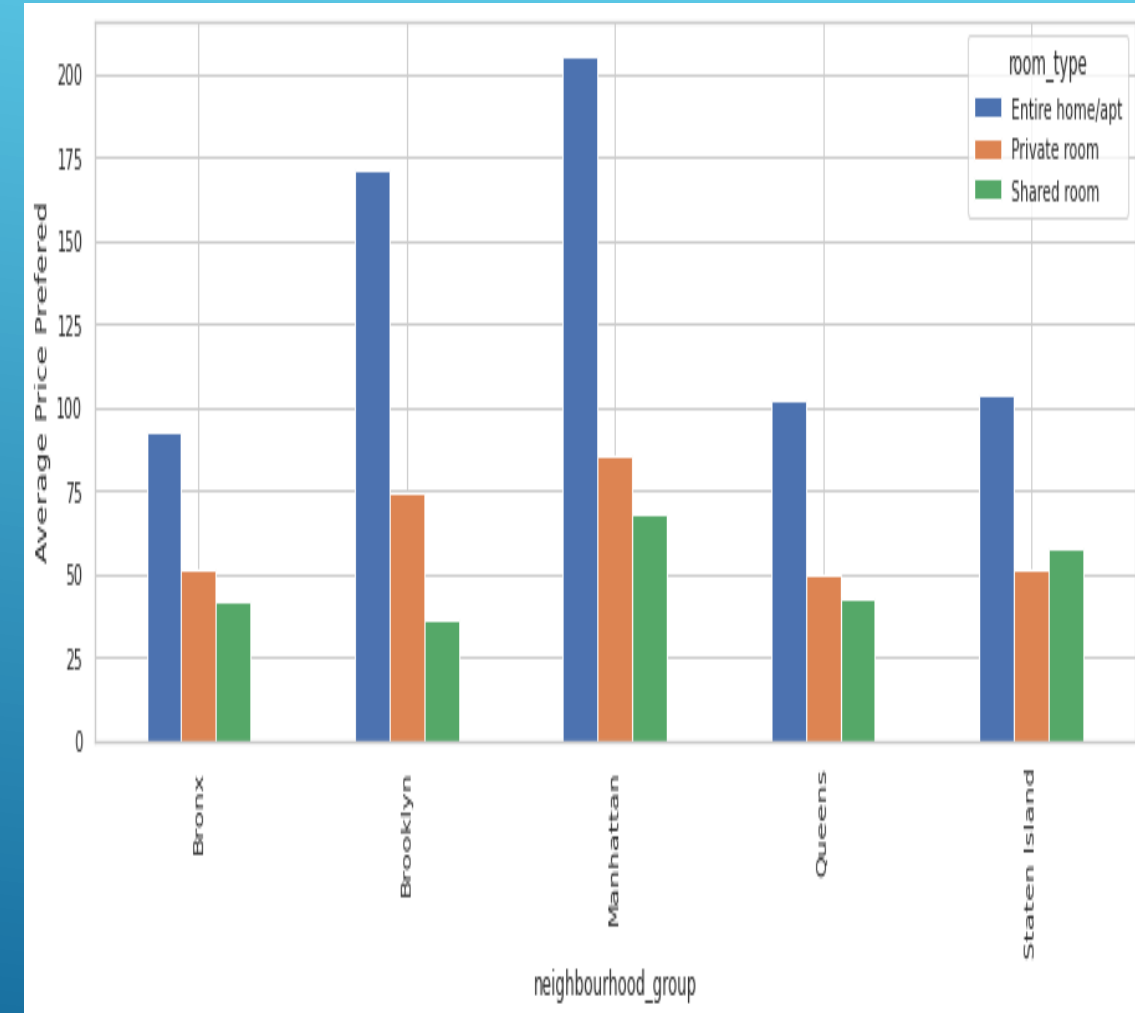


EDA :

11. What is the average price preferred for Keeping good number of reviews according to neighborhood group ?

OBSERVATIONS

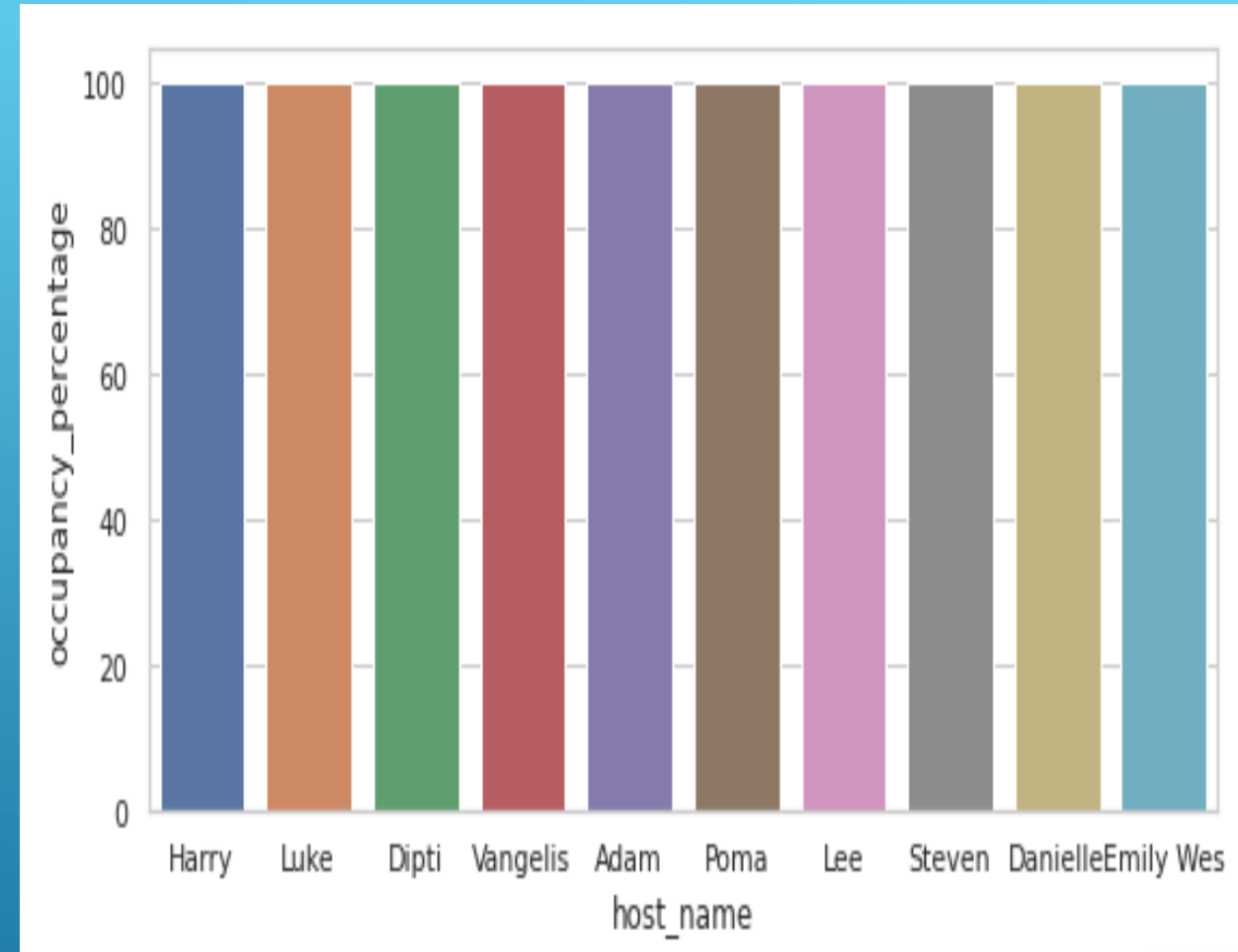
- clearly if we compare the results with previous result (i.e. when we calculated average preferred price by people in each neighborhood group with different room types) we can see that this result is bit different and more useful.
- As an analyst I would suggest to keep price in this range to get more number of reviews in specific room type and at particular place.
- And also suggest for work on price of shared room and also to check why customer prefer very low for private room in Staten island.



EDA :

12. Which neighborhood have most of the host focused to own property ?

- As we can see that harry, Luke, Dipti, Vangelis, Adam, Poma, Lee, Steven, Danielle, Emily are top 10 host with 100 percent occupancy percentage.
- As per data we can say they are busiest host so we can suggest that Airbnb should survey them or monitor them.
- As per survey or monitor trained another host or suggest them how to increase occupancy percentage.
- Which will helpful for both of them Airbnb and host also.



EDA : FORMULA TO FIND 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 aforementioned are taken into consideration for calculating this metric. the metric needs the available months (one period of business) the host is open for business/accepting bookings :-

$$\text{available months} = \text{available days} / (365/12)$$

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 :-

$$\text{total possible bookings} = \text{available days} / \text{minimum nights}$$

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 :-

$$\text{estimated bookings} = \text{reviews per month} \times \text{available months}$$

Using all the above calculations, the percentage of occupancy throughout the year is given as :-

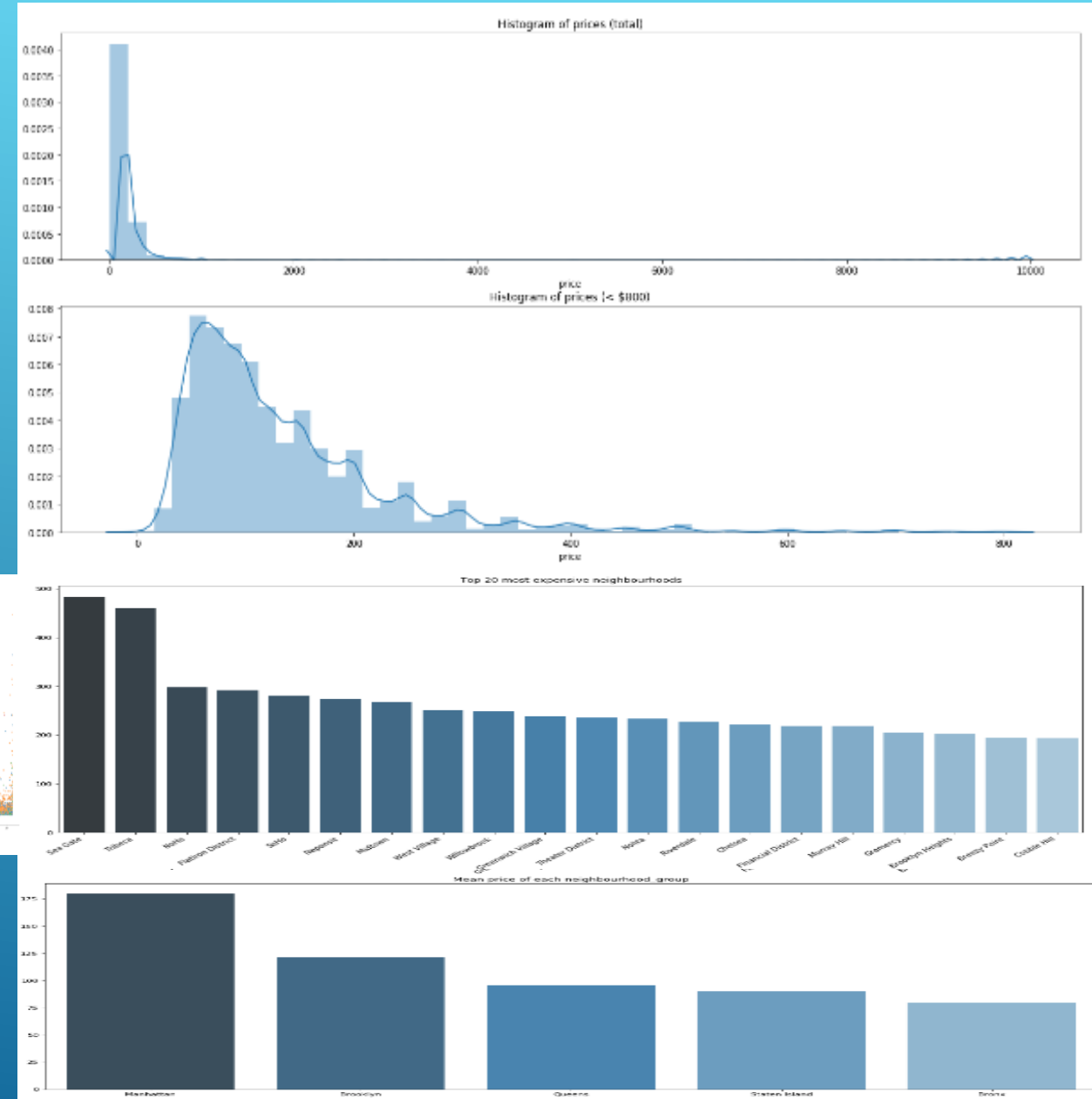
$$\text{occupancy \%} = \text{estimated bookings} / \text{total possible bookings} \times 100$$

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.

EDA :

12. Overall price analysis ?

- The first graph shows the total prices histogram of prices(total).
- The second graph shows the prices above \$800 histogram of prices(<\$800).
- Third graph shows the top 20 most expensive neighborhood.
- Fourth graph shows the mean prices for each neighborhood group.



- By the help of seaborn we plot graph over price to latitude, longitude, number of review, review per month, calculated host listing count, availability 365.

EDA :

conclusions

We can see several things:

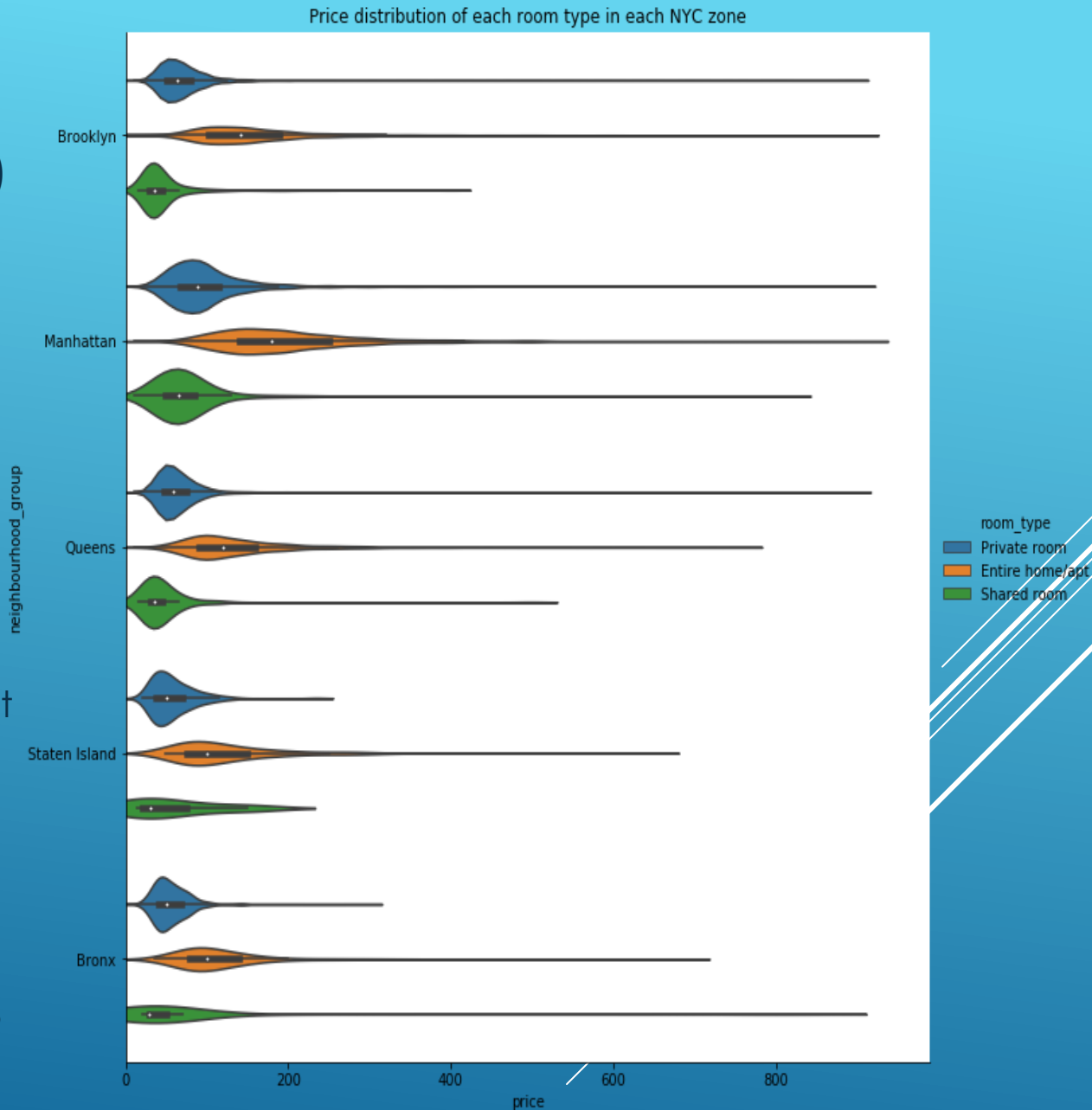
1. Entire apartments are the most expensive ones (obviously)
2. Shared rooms tend to be more in the city center.
3. The price distributions of each room type in the 5 different NY zones look similar, the most notable difference is on their means.

4. The price distributions of the room types point that, in general, shared and private rooms have similar prices (less deviation). On the other hand, entire apartments have more variability (and are more expensive, obviously). We don't have data about the properties, but we can guess that other variables like square ft. or being near a metro station affect the price.

5. When looking at the Popularity/Price plot, we can see that rooms with more reviews tend to be cheaper (although the Entire apartment class is more noisy than the others). This can be explained as: the more expensive is an apartment, the less people stay there and therefore, the less reviews. This would be interesting to use in a model to predict prices or the popularity of a room.

6. It's generally cheaper to stay in rooms between 14 and 28 nights.

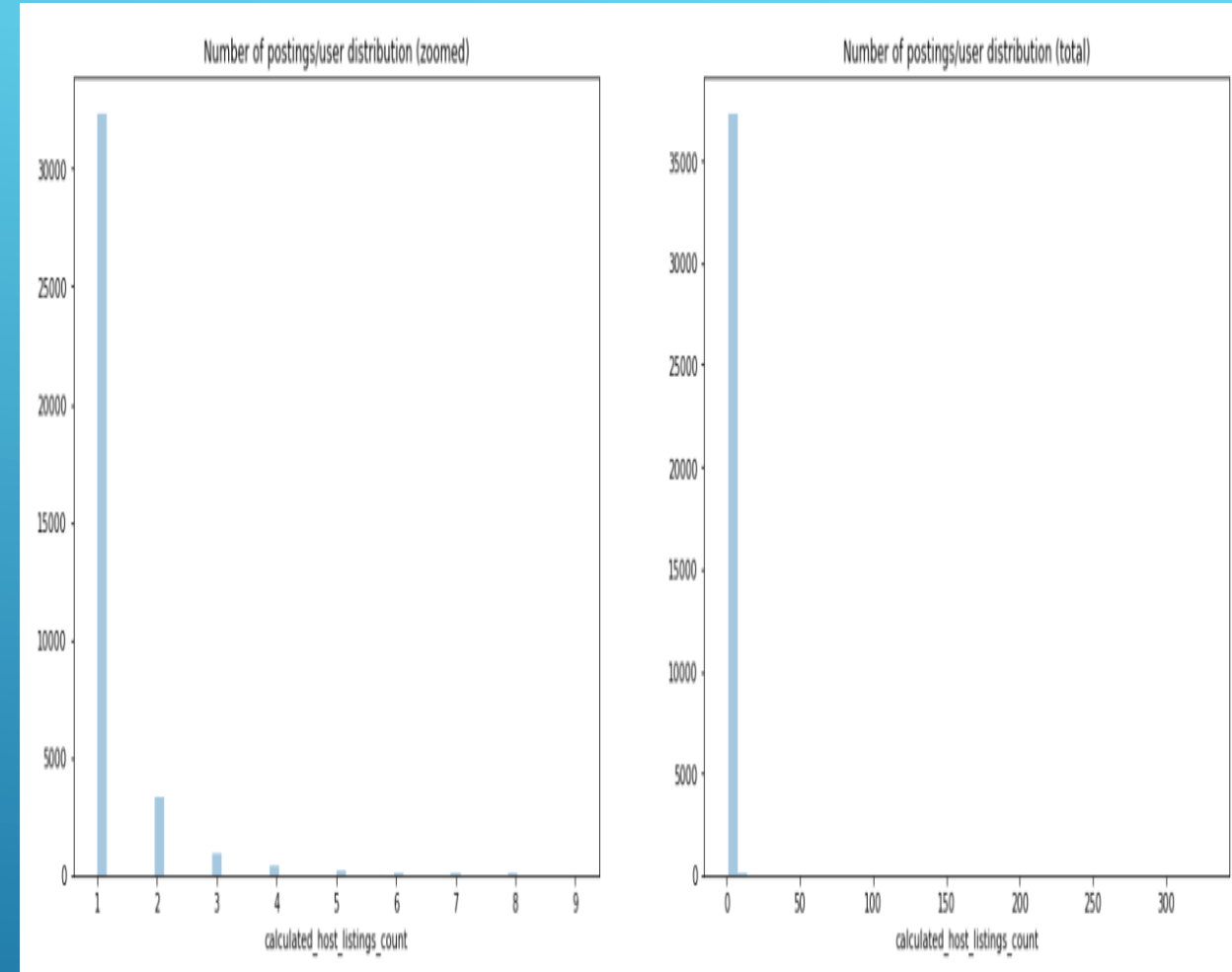
7. Usually, the minimum required nights to stay in a room is around 2.



EDA:

13. Overall posting analysis ?

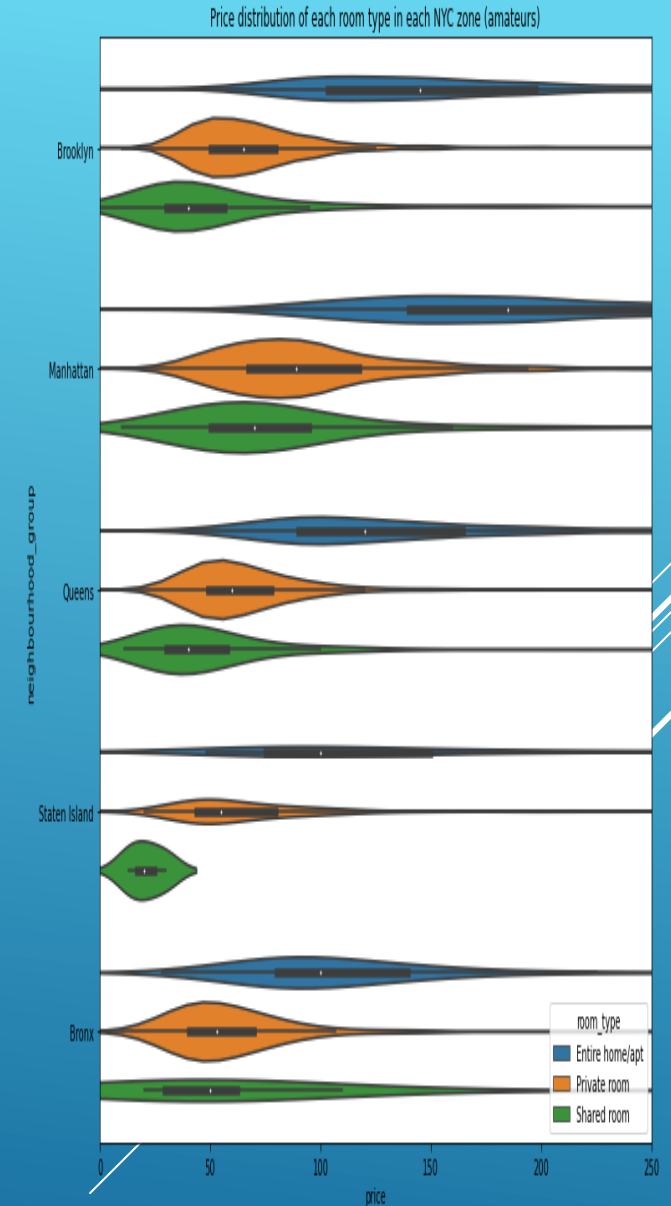
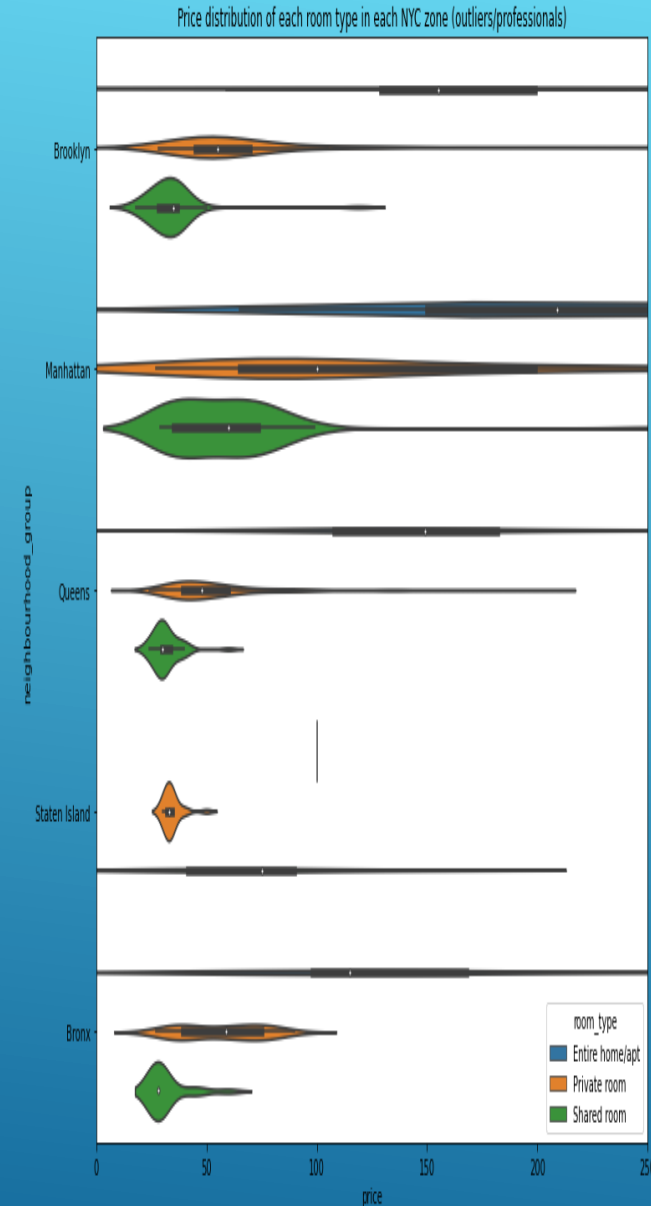
- In this section the postings themselves are studied to infer information about the users who post rooms
- The distribution plot says us that most users have around one listing on the platform, however, it strikes me that there are some outliers that have >300 . Maybe they are professional users/agencies. Let's take a look. We can define the outliers as the ones who have p postings as $p \geq \mu + 2\sigma$.
- The first graph is number of posting/user distribution(zoomed) at calculated host listing count.
- The first graph is number of posting/user distribution(total) at calculated host listing



EDA :

Section Conclusions:

1. There are two type of hosts: professional and amateurs.
2. The more rooms a user has, the more they tend to have clustered properties, maybe because they buy or rent the apartments/rooms in bulk.
3. The professional users post rooms with slightly higher price than the amateur ones.
4. The professional users tend to prefer centric zones




EDA : Summary

The NYC Airbnb dataset it's a contains a very well curated list of room postings from the New York city. Although it would be nice to have few more features about the rooms, users, comments, etc., I think a great number of task can be accomplished even with the limited number of entries in the dataset, at least, we all can learn something new about the vacation room rental in NYC.

This analysis is light in terms of the things tested as I really hadn't any particular goal to accomplish (curiosity driven)

I've structured this analysis in three main areas (maybe more focused on a business analytics part): Prices, Postings(users) and a light textual analysis of the room titles.

Several white lines of varying lengths and thicknesses are drawn diagonally across the bottom right corner of the slide, extending from the right edge towards the bottom left.

EDA : What have we found ?

We have tested several hypothesis against the data, and the conclusions obtained are listed at the end of each section, however, here are the most interesting ones (IMO).

- NYC shared rooms tend to be grouped in the city center, maybe because there are thought for travelers who want to visit the most iconic city places.
- Relating the price/popularity variables suggest that people who travel and use Airbnb tend to prefer the posts which are cheaper
- There are two types of user posting rooms: Professionals, which are outliers, each one holding a high number of rooms; and Amateurs, who usually have only a few. Although amateurs can be making money as a business to, their volume is clearly inferior to the professional ones.
- The professional posts are located in the city center.
- The way rooms are announced is different between professionals and amateurs. The first use more objective terms to describe the room whereas the second use more subjective.
- Having a room "near to" things affect to popularity (maybe it's a good idea to include this words in the title of the room).

CHALLENGES FACED :

- Reading the dataset and understanding the column.
- For answering some of the questions we had to understand the business model of Airbnb that how they work.
- Handling Nan values, null values and duplicates
- Designing multiple visualizations to summarize the information in the dataset and successfully communicate the result and trends to the reader.
- Removing the outliers for some data set. Finding and sorting of few complicated dataset.
- The biggest challenge that we faced is finding the busiest hosts. If we try to find the busiest host by an only number of reviews then this may be not the correct metric. We don't use the current status of the host having the highest number of reviews. If we check the date of last review and find out the review are very old than current date, then we can infer that business currently shutdown hence the busiest host should be that one whose occupancy is almost full.

CONCLUSION :

This Airbnb ('AB_NYC_2019') dataset for the 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented. First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings; we found that our top host has 327 listings. After that, we proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another. Next, we put good use of our latitude and longitude columns and used to create a geographical heatmap color-coded by the price of listings. Further, we came back to the first column with name strings and had to do a bit more coding to parse each title and analyze existing trends on how listings are named as well as what was the count for the most used words by hosts. Lastly, we found the most reviewed listings and analyzed some additional attributes. For our data exploration purposes, it also would be nice to have couple additional features, such as positive and negative numeric (0-5 stars) reviews or 0-5 star average review for each listing; addition of these features would help to determine the best-reviewed hosts for NYC along with 'number of review' column that is provided. Overall, we discovered a very good number of interesting relationships between features and explained each step of the process. This data analytics is very much mimicked on a higher level on Airbnb Data/Machine Learning team for better business decisions, control over the platform, marketing initiatives, implementation of new features and much more. Therefore, I hope this kernel helps everyone.

THANK YOU

