Capstone Project

Airbnb Bookings Analysis

By- Vaibhav Satish Taware

Points for Discussion

- Data summary
- Handling missing values

Research Questions

- Which place is good to go in terms of price and availability?
- Which room type is preferred by the host?
- Which month has received highest last reviews?
- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc.)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?
- Which area is most expensive ? And why?
- Who are the top 10 hosts who have listed their property maximum number of times? In which room type they have invested in? And in which area their property belongs to?

Data Summary

Airbnb NYC 2019 : This dataset has around 48,895 observations in it with 16 columns and it is a mix between categorical and numeric values.

Let's understand the each column in detail:

- id- This is the identity number of the property listed by a particular host.
- Name- Description of the property.
- host_id- This is the identity number of the host who have registered on Airbnb website.
- neighbourhood_group- This is the name of the neighbourhood group.
- neighbourhood- This is the name of the neighbourhood present in the neighbourhood group.

- latitude- This represents the coordinate of latitude of the property listed.
- longitude- This represents the coordinate of latitude of the property listed.
- room_type- This represents types of room listed by the host.
- price- This is rent of listed property in USD.
- minimum_nights- This is the minimum number nights to be spend by the guest in that property to rent a property.
- number_of_reviews- This represents the number of reviews the particular property received.
- last_review- The date at which property was last reviewed.
- reviews_per_month- Reviews per month property received.
- calculated_host_listings_count- Number of listing done by a particular host on Airbnb.
- availability_365- This represents the number of days property available in 365 days.

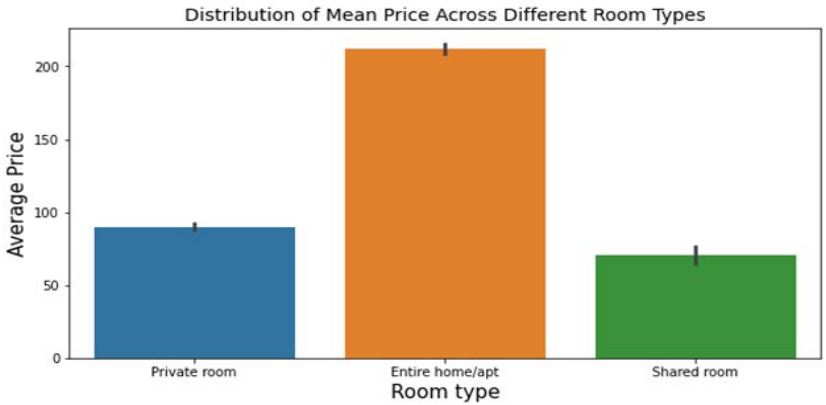
Handling missing values

The columns which contains missing values are name,host_name,price, last_review and reviews_per_month. We drop the missing value rows which has missing value in name and host_name columns and then the number of rows in a dataframe reduced to 48858.

In barplot we can see that price is depend upon room type so we replace missing values for price column category wise i.e using room type.

Replaced missing values of column reviews per month by mean of the reviews_per_month.

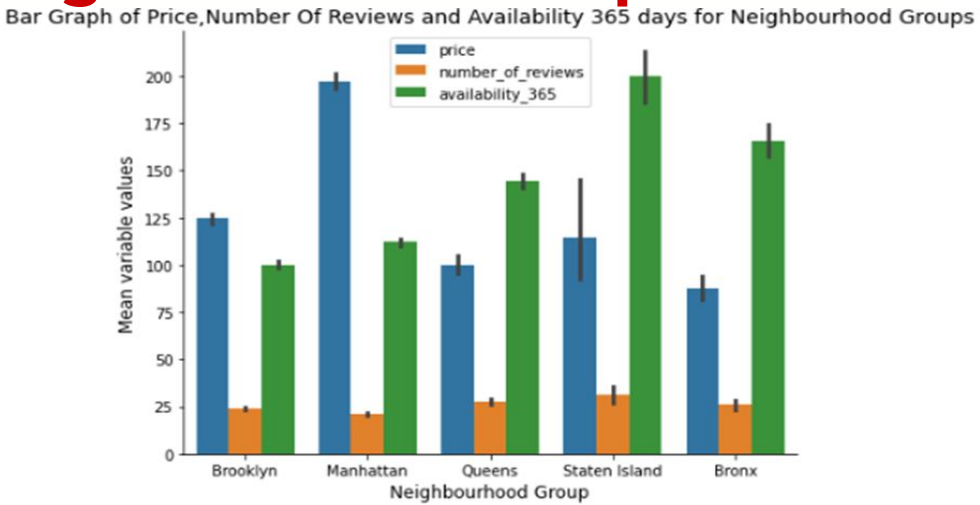
except last_review column any other column does not have missing values.



Which place is good to go in terms of price and availability ?

- From bar graph we could see Bronx is good neighbourhood group to go since this neighbourhood has minimum mean price and second highest availability 365 days.

- But we narrow it down to neighbourhood Bull's Head is most affordable place to go which belongs to Stater Island.



Neighbourhood Group	Neighbourhood	Mean Price
Staten Island	Bull's Head	47.333333
Bronx	Hunts Point	50.500000
Bronx	Tremont	51.545455
Bronx	Soundview	53.466667
Staten Island	New Dorp	57.000000

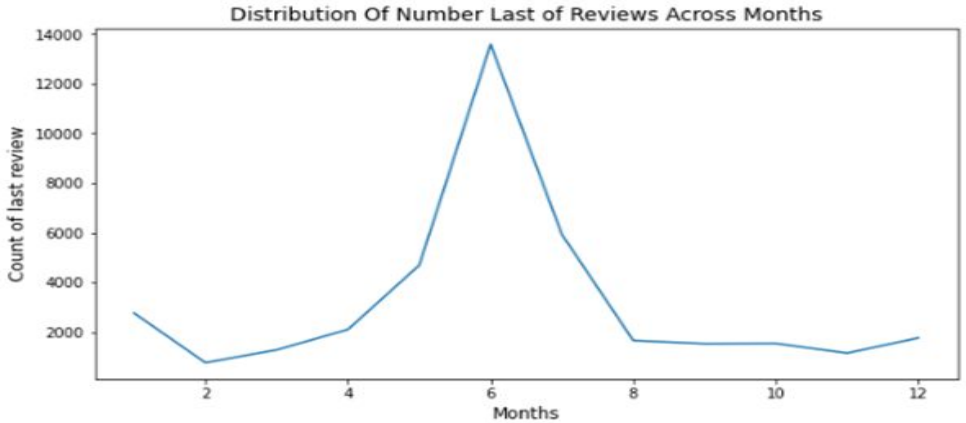
Which room type is preferred by the host?

- Most of the hosts prefer to rent the Entire home/apt then followed by Private room and then Shared room.

Room Type	Count of Rooms	Percentage of a Rooms
Entire home/apt	25393	51.973065
Private room	22306	45.654755
Shared room	1159	2.372181

Which month has received highest last reviews ?

- The last review is mostly given by customers in the month of June (6).

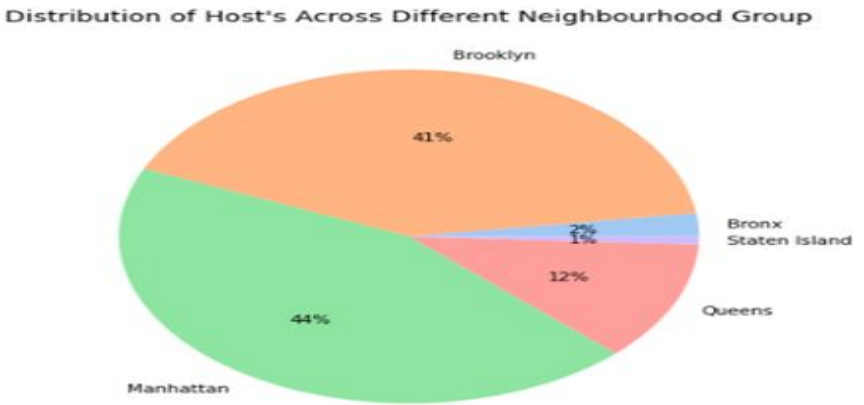


- The year 2019 has the highest percentage of last review in the month June.

Last Review Year	Count of Last Review in June	Percentage of count of last review in June
2019.0	12599	92.735169
2018.0	382	2.811718
2016.0	247	1.818048
2017.0	247	1.818048
2015.0	89	0.655086
2014.0	16	0.117768
2013.0	4	0.029442
2012.0	2	0.014721

What can we learn about different hosts and areas?

- Manhattan neighbourhood group has high percentage of hosts followed by Brooklyn and so on.



- Top 5 neighbourhood who has high count of hosts.

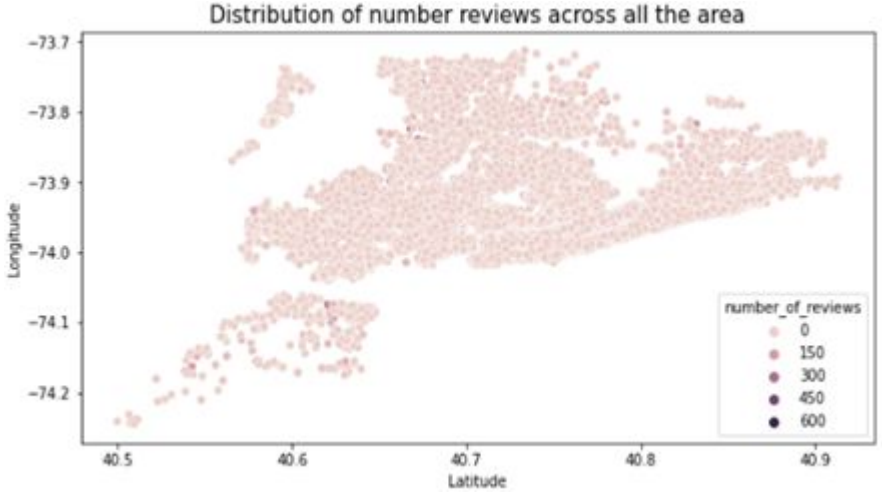
Neighbourhood group	Neighbourhood	Count of host Id
Brooklyn	Williamsburg	3917
Brooklyn	Bedford-Stuyvesant	3713
Manhattan	Harlem	2655
Brooklyn	Bushwick	2462
Manhattan	Upper West Side	1969

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

- People usually go for Entire home/apt or Private room so after removing the shared room type we obtain the table which shows that the Graniteville neighbourhood has the mean minimum price among all neighbourhood's.

Neighbourhood group	Neighbourhood	Room Type	Mean Price
Staten Island	Graniteville	Private room	20.000000
Staten Island	Grant City	Private room	29.500000
Bronx	Van Nest	Private room	36.666667
Bronx	Castle Hill	Private room	38.600000
Staten Island	New Dorp Beach	Private room	38.666667

- From scatter plot we could see that the number of reviews across all the locations are mostly lies between 0 to 150.
- Lets a consider outlier data in terms of number of reviews to see in which area those outliers belong to ?
- Outliers are calculated using IQR we considered those review whose number of review are more than upper whisker in boxplot i.e more than Q3+1.5 * (I.Q.R)
- Brooklyn has the highest number of outlier reviews.



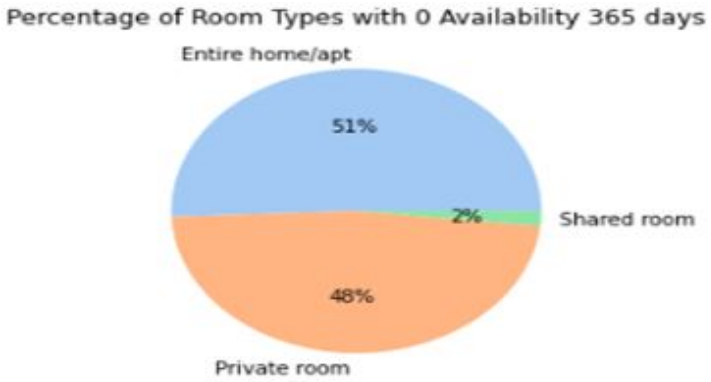
Neighbourhood Group/Area	Count of outlier reviews
Brooklyn	2611
Manhattan	2347
Queens	832
Bronx	157
Staten Island	68

Which hosts are the busiest and why?

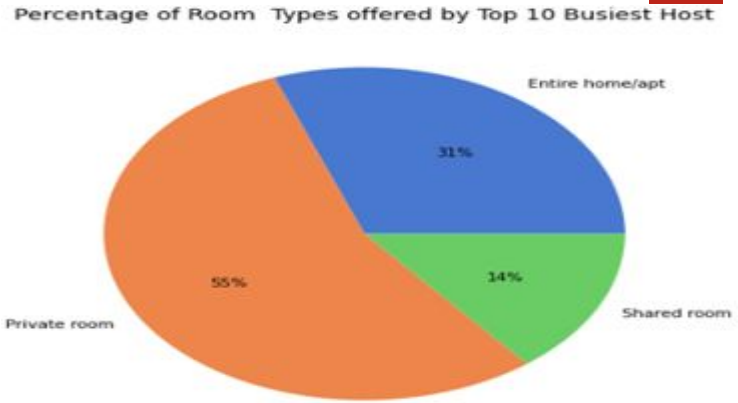
- To see the busiest host's we consider data with 0 availability 365 days.
- These are the top 10 hosts those are most busiest among all 16410 hosts who have 0 availability 365 days and here Count of 0 availability 365 days column shows that number of time that host has 0 availability.

Host ID	Host Name	Count of 0 availability 365 days
19303369	Hiroki	16
100238132	Michael	12
51596474	Antony	12
137358866	Kazuya	11
204852306	Dee	11
193502084	Linda	8
24831061	Hosteeva	8
51913826	The Bowery House	8
732460	Nôm	6
187487947	Diego	6

- Lets see percentage of room types with 0 availability 365 days.
- From pie chart we can see that shared rooms have only 2% 0 availability 365 days it means shared rooms does have not have a high traffic whereas Entire home/apt and Private rooms are high in demand.



- Lets see percentage of room types offered by top 10 busiest host's
- Most busiest top 10 host's offers Entire home/apt and Private room in higher percentage so this could be the one reason of being most busiest hosts.



Is there any noticeable difference of traffic among different areas and what could be the reason for it?

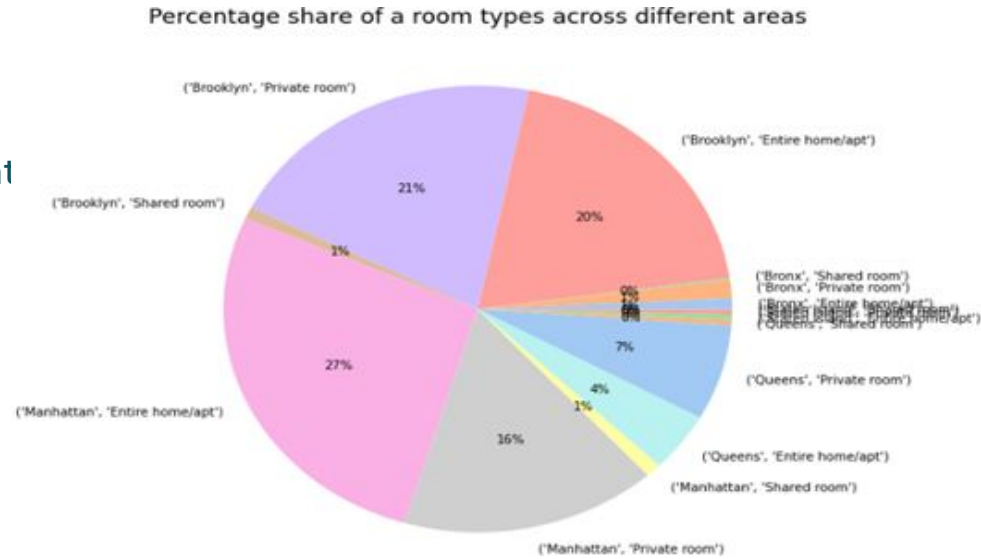
- The higher is the total number of reviews this implies more is the number of the customers have visited to that area and higher is the traffic in that area. Table shows the top 3 high traffic areas.

Neighbourhood Group/Area	Total Number of reviews
Brooklyn	486174
Manhattan	454126
Queens	156902

- Lets see the reason of being in a top 3
- The mean minimum nights to be spend in top 3 high traffic areas is also high this means that customer has to spend more nights in these areas than any other areas so this could be the one reason of high traffic.

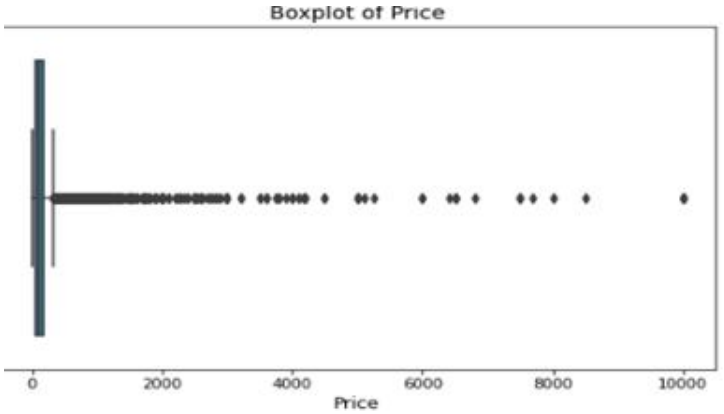
Neighbourhood Group/Area	Mean Minimum Nights
Manhattan	8.538188
Brooklyn	6.057693
Queens	5.182910
Staten Island	4.831099
Bronx	4.564738

- In pie chart percentage share of room types across different area we can see that the high traffic areas such as Manhattan, Brooklyn and Queens have higher percentage of both Entire home/apt or Private room types so this could be the second reason for high traffic in these areas



Which area is most expensive ? and Why ?

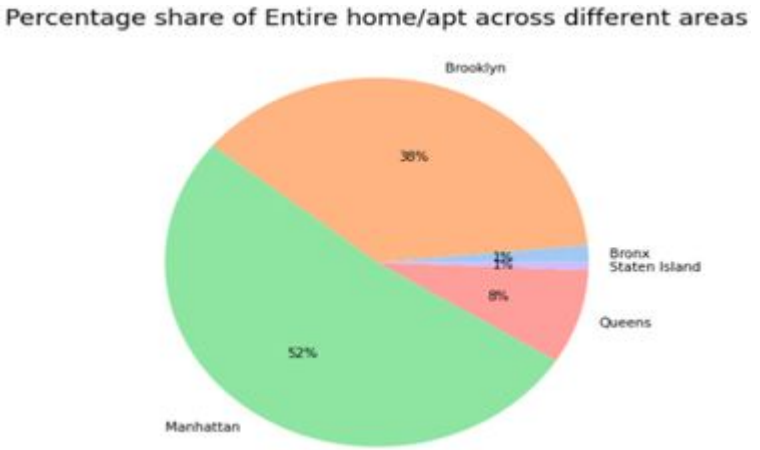
- After seeing the boxplot of price we can see that there are so many outliers present above the upper whisker.



- After considering the outlier data in terms of price along with their room type reason to consider room type since price is dependent on the room type
- In table we can see that most of outliers are from Manhattan area and their room type is Entire home/apt which accounts to 64.1 % of the data and then followed by Brooklyn area with same room type i.e Entire home/apt with 20.8 % and so on.

Neighbourhood Group	Room Type	Count of Room types in Percent
Manhattan	Entire home/apt	64.086166
Brooklyn	Entire home/apt	20.767418
Manhattan	Private room	7.909795
Queens	Entire home/apt	2.490744
Brooklyn	Private room	2.356109
Queens	Private room	0.706833
Bronx	Entire home/apt	0.538539
Manhattan	Shared room	0.504881
Staten Island	Entire home/apt	0.269270
Bronx	Private room	0.134635
Queens	Shared room	0.134635
Brooklyn	Shared room	0.067317
Bronx	Shared room	0.033659

- Lets see why Manhattan area has highest outliers ?
- In pie chart we can see that Manhattan area has the highest Entire home/apt so it has higher number of outliers.



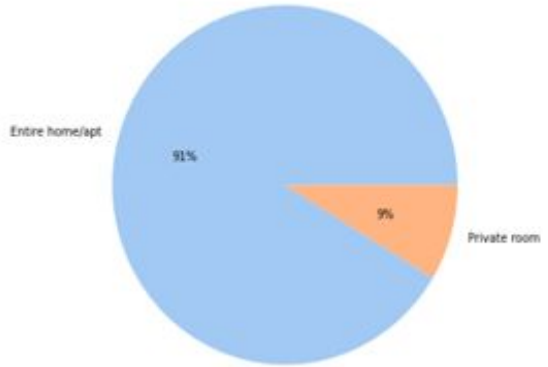
Who are the top 10 hosts who have listed their property maximum number of times ? In which room type they have invested in ? and in which area their property belongs to ?

- These are the top 10 host's with highest number of property listing count.

Host ID	Listing Count of host id
219517861	327
107434423	232
30283594	121
137358866	103
12243051	96
16098958	96
61391963	91
22541573	87
200380610	65
1475015	52

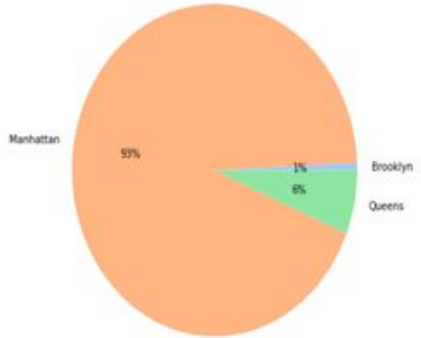
- Top 10 hosts only prefers to rent the Entire home/apt or Private room and not the shared room the reason behind it could be the price.

The percentage of room types listed by top 10 host's who have listed their property highest number of times



- Top 10 host's only prefers to rent property in Manhattan,Queens and Brooklyn area and not in Staten Island and Bronx area. The reason for it could be the minimum nights to be spent in these areas is higher than other two area's which is shown in below table.

The percentage of neighbourhood group (Area) listed by top 10 host's who have listed their property highest number of times



Neighbourhood Group (Area)	Mean Minimum Nights
Manhattan	8.538188
Brooklyn	6.057693
Queens	5.182910
Staten Island	4.831099
Bronx	4.564738

Conclusions:

- Bull's Head is most affordable and good availability place to go which belongs to Staten Island.
- Most of the hosts prefer to rent the Entire home/apt then followed by Private room and then Shared room.
- The year 2019 has the highest percentage of last review in the month June.
- Manhattan neighbourhood group has high percentage of hosts.
- People usually go for Entire home/apt or Private room and Graniteville neighbourhood has the mean minimum price among all neighbourhood's.
- Shared rooms does have not have a high traffic whereas Entire home/apt and Private rooms are high in demand. Most busiest top 10 host's offers Entire home/apt and Private room in higher percentage so this could be the one reason of being most busiest hosts.
- Brooklyn has high traffic among all neighbourhood group.
- Manhattan area is most expensive.

Thank you