Capstone Project

Airbnb EDA project

Team members

Prince Nayak

Gokul Jagannadh

Mohd Sharik

Introduction to Airbnb :

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.

How Airbnb works

- ***A very brief Airbnb business model may be defined as follows-***

- ***Hosts list out their property details on Airbnb along with other factors like pricing, amenities provided etc.***
- ***Airbnb sends a professional photographer to the property location in order to take high quality photographs.***
- ***Travelers search for a property in the city where they wish to stay and browse available options according to price, amenities etc.***
- ***Booking is made through Airbnb where traveler pays the amount mentioned by host and some additional money as transaction charges.***
- ***Host approves the booking. Traveler stays there and finally Airbnb pays the amount to the host after deducting their commission.***

Problem Statements:

Since our data has many categorical and numerical columns hence in this EDA project our main aim will be to explore the dataset with respect to columns as much we can. Some of our problem statements are given as. Again we are not limited to these problems :

1. What we can learn about hosts and areas ?

2. Difference of Traffic among different Neighbourhood.

3. Prediction on different columns of dataset (exp:-price, availability etc)

4. Which hosts are the busiest and why ?

Data Pipeline

1. Loading our data : **In this section we simply loaded our data in google colab to further EDA.**

2. ***Data Cleaning and Processing :*** **In this section we have removed unnecessary features and then we have cleaned out data by filling null values based on certain reasonable assumptions.**

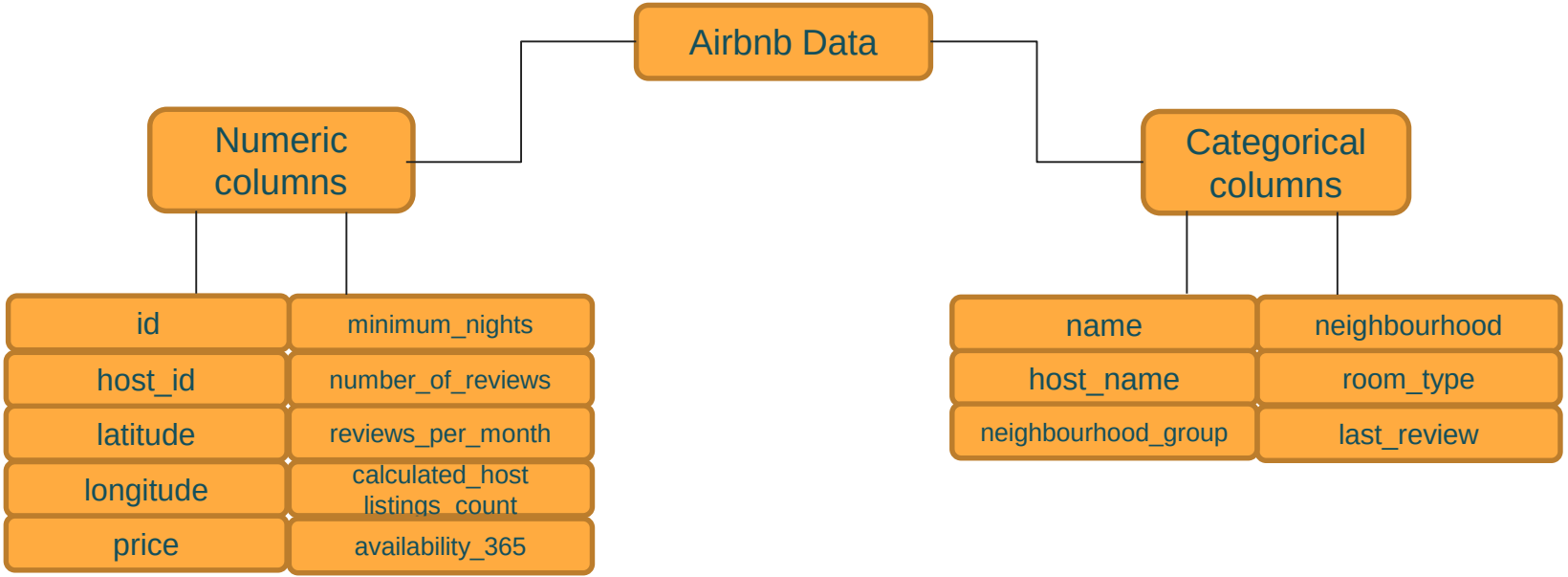
3. Analysis and visualization : **This section is divided into three parts.**

3.1 : Analysis on hosts, Neighbourhood, Neighbourhood groups and room types.

3.2 : Analysis on price distributions

3.3 : Analysis on availability, reviews and correlation matrix

Data summary :

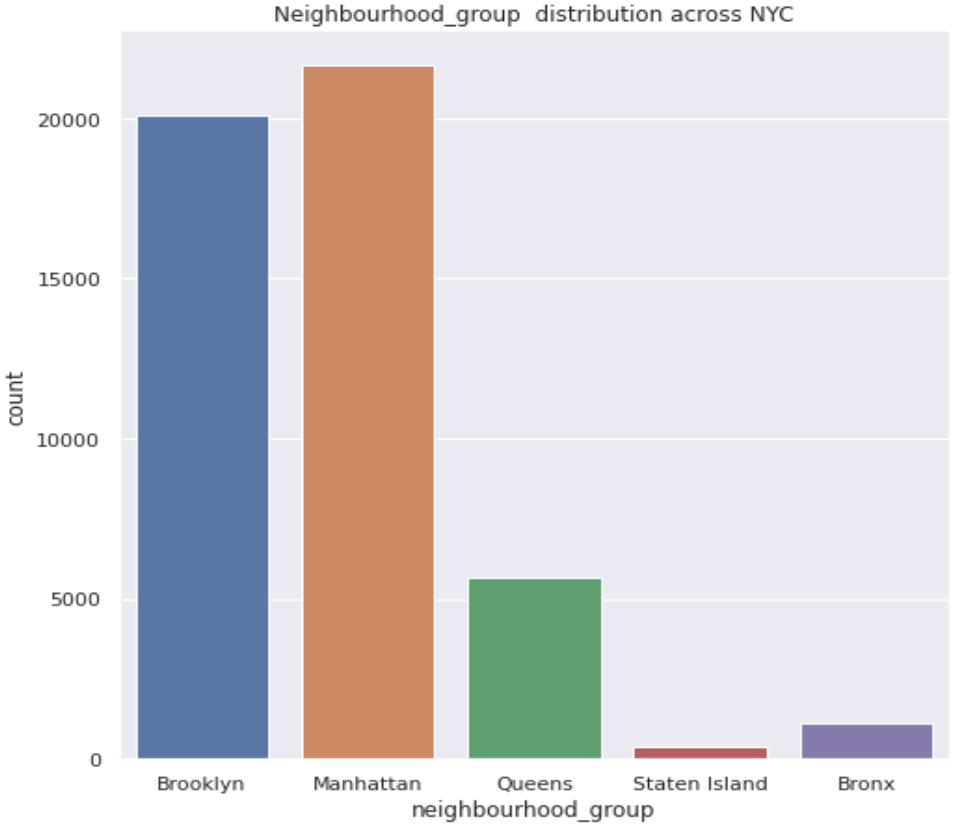


2. Data Cleaning and Processing

- **In this section first we have dropped 3 columns and these columns are-**
  - *1. id (Since all the values in id column are unique hence this column has no use in analysis).*
  - *2. name (Since we have 47905 unique values and some missing values hence this column also has considered of no use in analysis)*
  - *3. last_review (it's simply the date of review and since we already have num of reviews and review per month so we have drop this one too)*
- *Then we also get rid of missing values.*

3.1 : Analysis on hosts, Neighbourhood, Neighbourhood groups and room types.

Total number of listings in a Neighbourhood group

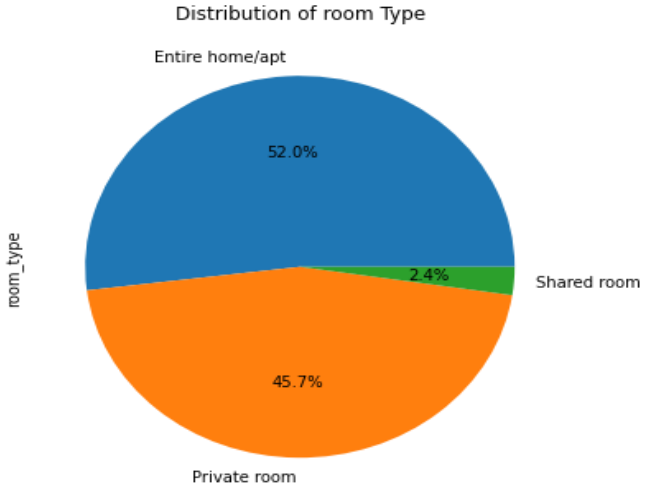


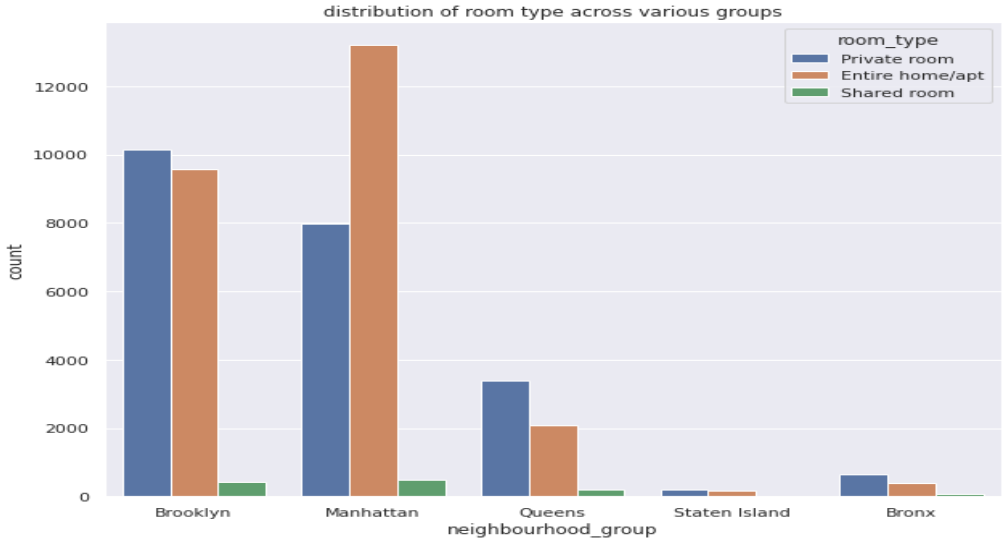
First we plotted a count plot between Neighbourhood groups and total listing counts. we observed that Manhattan has highest number of listings followed by Brooklyn. And minimum listings are listed in Staten island

3.1 : Analysis on hosts, Neighbourhood, Neighbourhood groups and room types.

Differentiation of Room types in different Neighbourhood areas.

Then we observed that although Manhattan has highest number of listing but in private rooms Brooklyn is leading and Manhattan Is leading because of higher number of Entire room\apartment.





We also have seen distribution of different Room types and found that people are More interested in entire room\aprt. and Only few people are interested in shared rooms.

3.1 : Analysis on hosts, Neighbourhood, Neighbourhood groups and room types.

Top 10 Neighbourhoods having most number of listings.

neighbourhood	neighbourhood_group	total_listings
Williamsburg	Brooklyn	3920
Bedford-Stuyvesant	Brooklyn	3714
Harlem	Manhattan	2658
Bushwick	Brooklyn	2465
Upper West Side	Manhattan	1971
Hell's Kitchen	Manhattan	1958
East Village	Manhattan	1853
Upper East Side	Manhattan	1798
Crown Heights	Brooklyn	1564
Midtown	Manhattan	1545

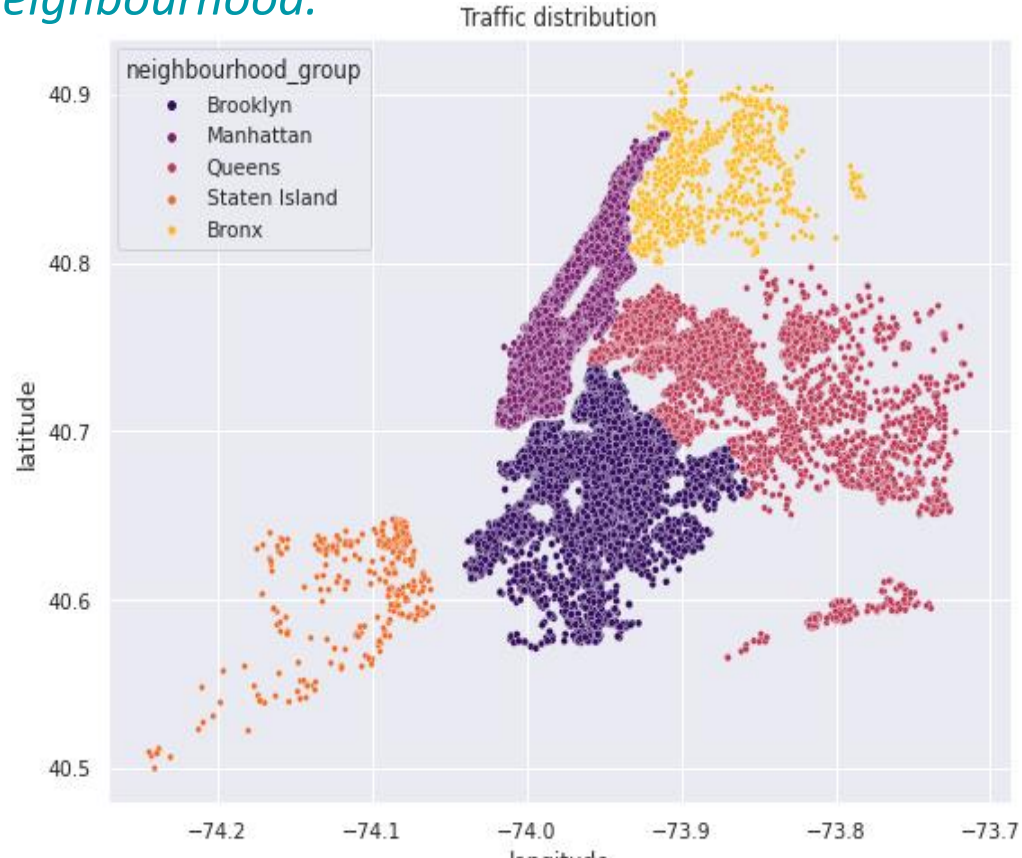
This table is giving us the top 10 neighbourhood having most number of listings. We can see that out of these 10 neighbourhood 6 are from Manhattan and 4 are from Brooklyn.

3.1 : Analysis on hosts, Neighbourhood, Neighbourhood groups and room types.

Difference of Traffic among different Neighbourhood.

To understand this we have plotted a scatter plot between Neighbourhood groups and total number of listings in the neighborhood. With this plot we find that Manhattan has highest number of listings followed by Brooklyn and Staten island has minimum number of listing followed by Bronx and Queens.

neighbourhood_group	listing_counts
Manhattan	21661
Brooklyn	20104
Queens	5666
Bronx	1091
Staten Island	373



Reason for traffic :

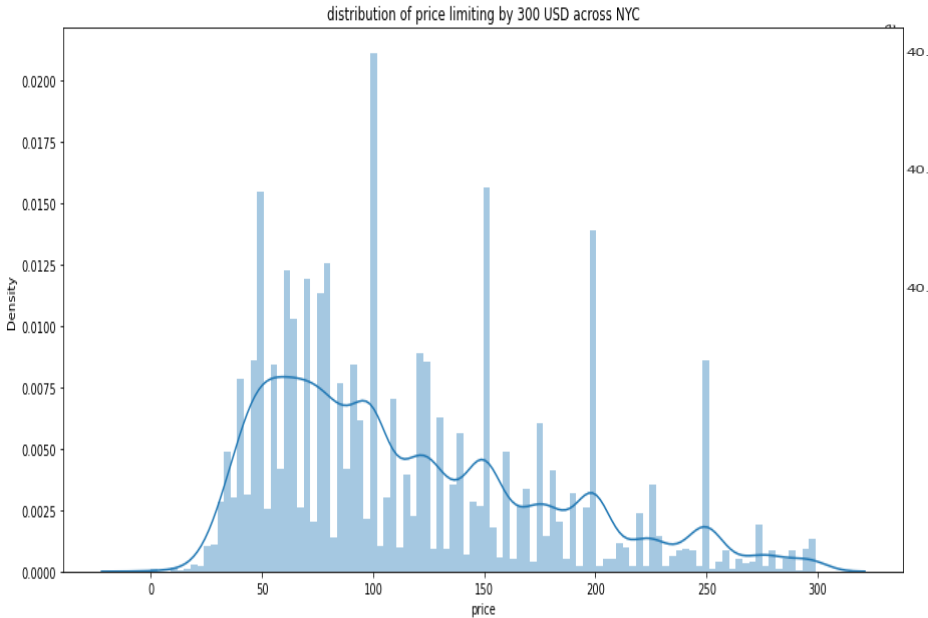
Airbnb users rate their stay on the basis of location, cleanliness and bunch of other parameters and to solve this we have to consider number of reviews column in our dataset. We have noticed that Brooklyn and Manhattan has highest number or reviews. We conclude that people who are searching rooms online go through the reviews and select their rental properties.

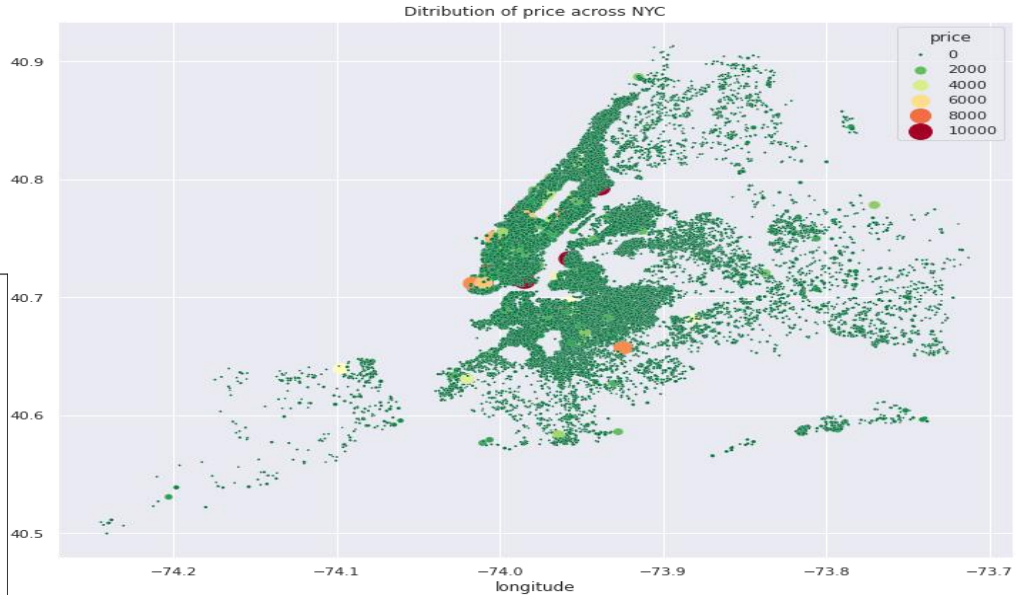
Neighbourhood group	Reviews count
Manhattan	454569
Brooklyn	486574
Queens	156950
Bronx	28371
Staten Island	11541

3.2 : Analysis on Price distributions

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

Using scatterplot we have seen that we have some outliers in our data(red, yellow and orange points)





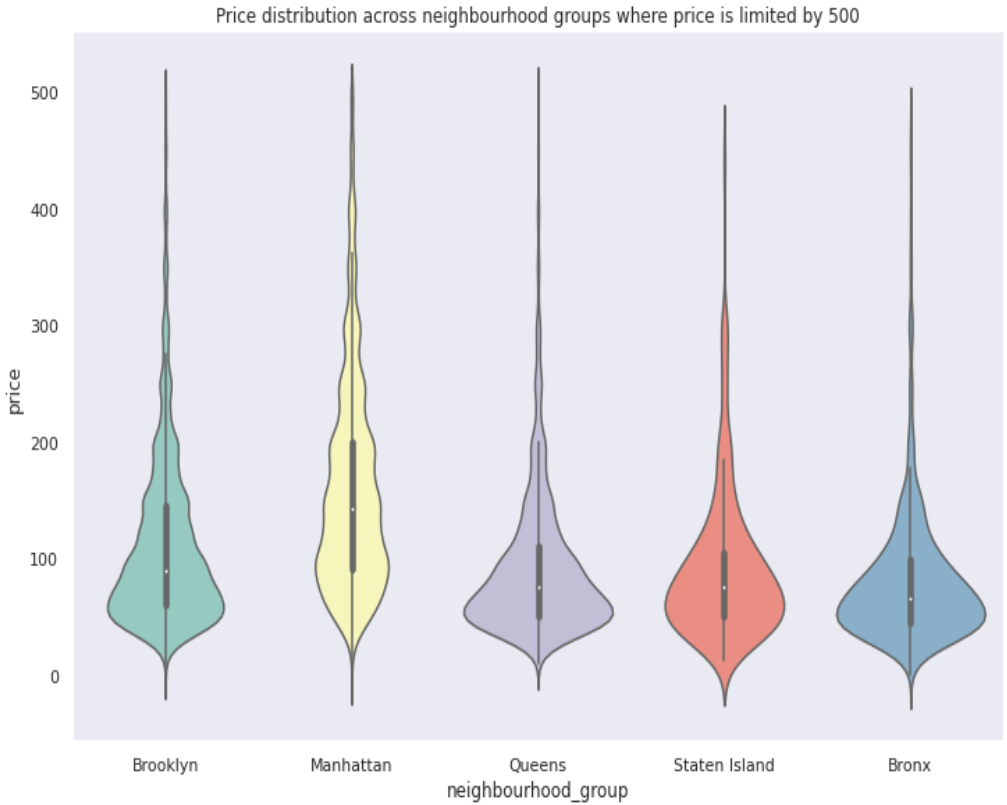
We also have seen that price density is Very high in range(40,200)

3.2 : Analysis on price distributions

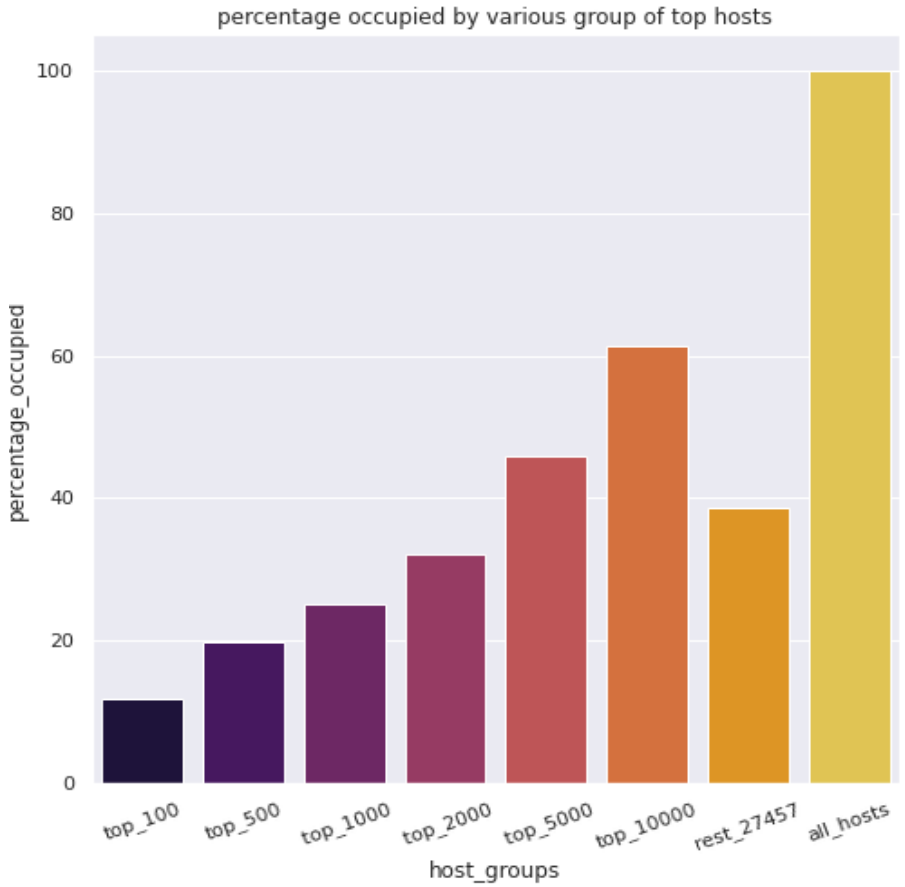
What we can learn from predictions (ex: locations, price, reviews) ?

To understand the distribution of price in Neighbourhood groups we used a violin plot And found that Manhattan has highest Mean price and Bronx has minimum. We can also summarize that Manhattan has highest noise in price followed by Brooklvn and so on.

neighbourhood_group	median	mean	min	Q75	Q90	Q97	max
Bronx	65.0	87.496792	0	99.0	145.0	282.4	2500
Brooklyn	90.0	124.383207	0	150.0	210.0	350.0	10000
Manhattan	150.0	196.875814	0	220.0	333.0	599.0	10000
Queens	75.0	99.517649	10	110.0	175.0	275.0	10000
Staten Island	75.0	114.812332	13	110.0	184.0	299.0	5000



Revenue analysis



Another very interesting insight we found by analyzing revenue made by top hosts(top revenue makers) that top1000 hosts occupy approx. 25% of all revenue and top 10000 occupy approx. 61% of all revenue.

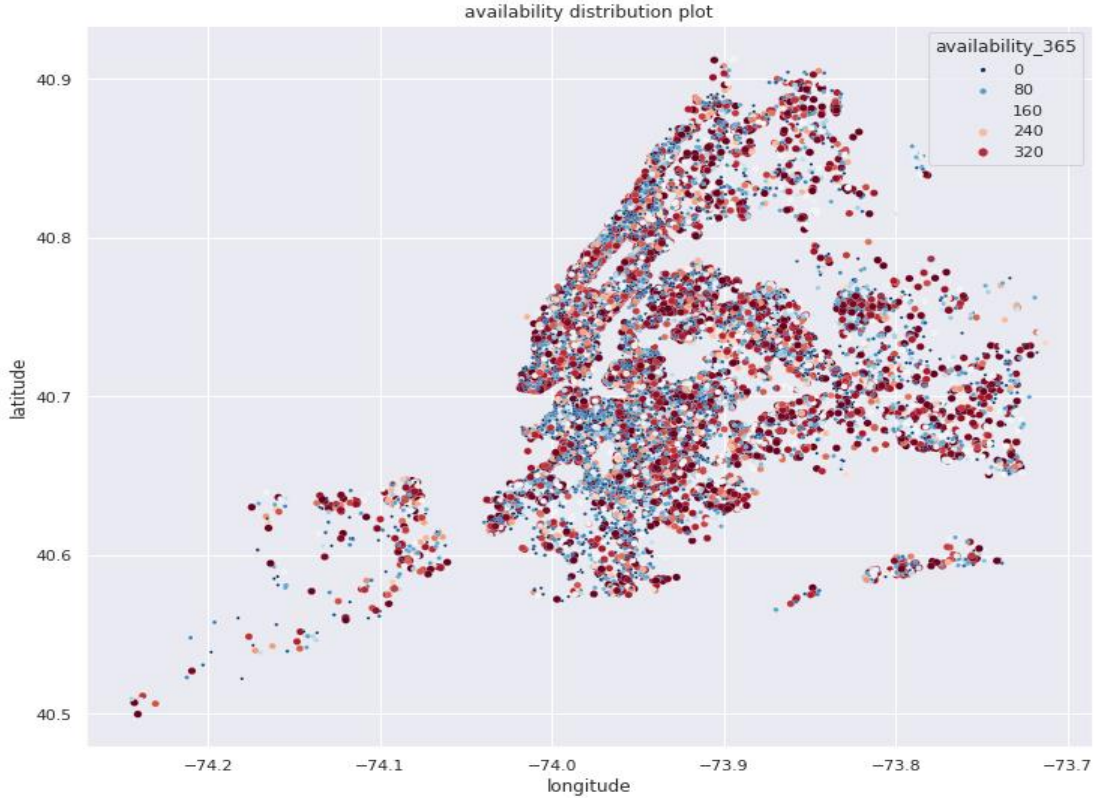
	host_groups	revenue_of_group	percentage_occupied
0	top_100	880814	11.80
1	top_500	1485833	19.90
2	top_1000	1874550	25.10
3	top_2000	2398463	32.12
4	top_5000	3431809	45.96
5	top_10000	4582349	61.37
6	rest_27457	2884929	38.63
7	all_hosts	7467278	100.00

3.3 : Analysis on availability, reviews and correlation matrix

Analysis on Availability

To understand the availability we Plotted a scatterplot based on Availability with longitude and latitude of the city to observe the Graph distribution and the index shows different availability Across the city

neighbourhood_group	mean
Bronx	165.758937
Brooklyn	100.232292
Manhattan	111.979410
Queens	144.451818
Staten Island	199.678284



3.3 : Analysis on availability, reviews and correlation matrix

Which hosts are the busiest and why ?

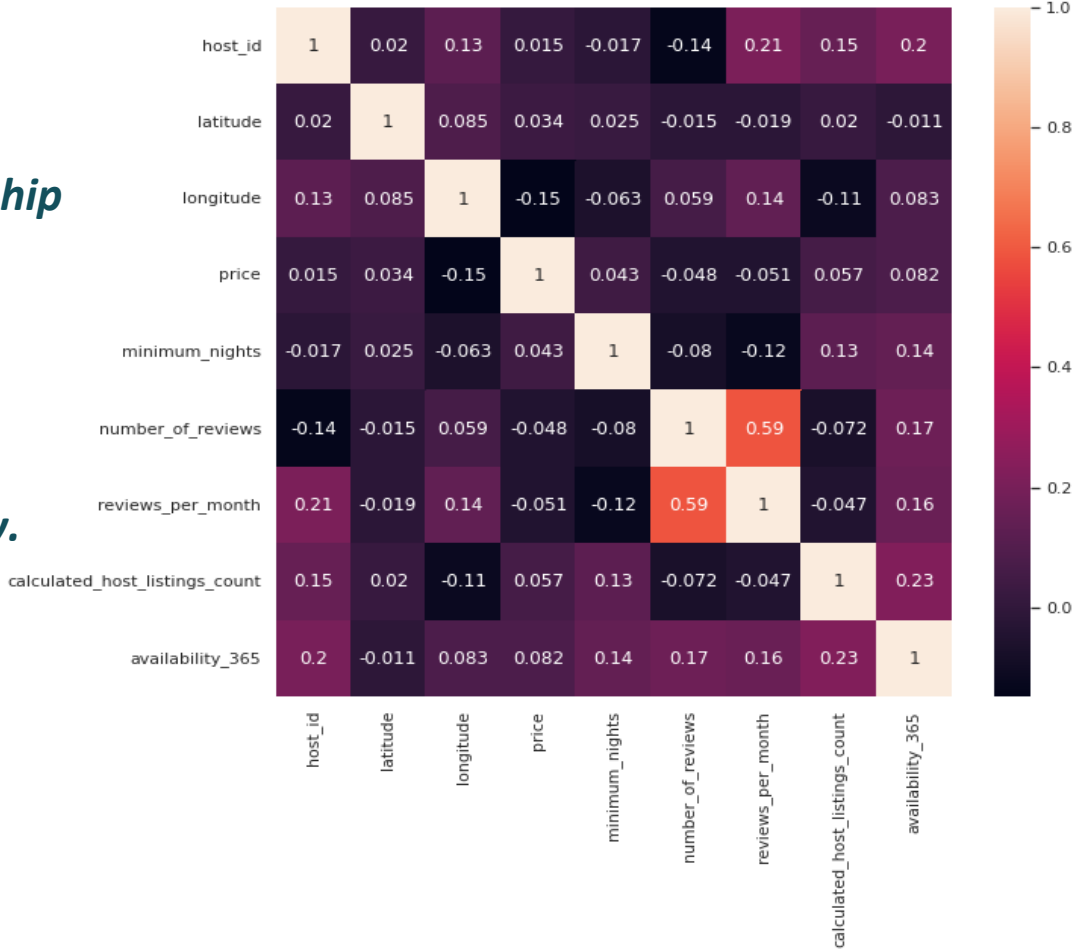
host_id	host_name	revenue	total_listings	avg_availability
219517861	Sonder (NYC)	82795	327	301.49
107434423	Blueground	70331	232	253.81
30283594	Kara	33581	121	313.42
137358866	Kazuya	4514	103	196.48
12243051	Sonder	20451	96	276.61
16098958	Jeremy & Laura	20060	96	292.32
61391963	Corporate Housing	13308	91	241.92
22541573	Ken	18743	87	312.17
200380610	Pranjal	18865	65	338.03
7503643	Vida	7758	52	297.71

To solve the above problem we have to consolidate three column as seen in the data frame next to it, so we calculated the top 10 busiest hosts based on total number of listings and avg. availability. This table has shown that the busiest host is Sonder (NYC) followed by Blueground and Sonder Kara.

These hosts are busiest because they have higher number of listing and availability is less.

3.3 : Correlation matrix

Finally, we plotted a correlation Matrix this matrix shows relationship between all column in the dataset Headed by a threshold between -1 to 1 and we have found that there is a positive corr. between total listing counts and availability. and price is negatively corr. With longitude(this may be because Manhattan has highest price in data set and hence creating this anomaly).



Conclusions :

Through this exploratory data analysis and visualization project, we gained several interesting insights into the Airbnb rental market.

- ***1. Learning about hosts and areas:-Manhattan has highest number of listing but in private rooms Brooklyn is leading and Manhattan Is leading because of higher number of Entire room\apartment.***

- ***2. Traffic among different Neighbourhood. Manhattan has highest number of listings followed by Brooklyn and Staten island has the minimum number of listings we have seen it after plotting Neighbourhood with total number of listings in the Neighbourhood.***

Conclusions :

- *We also have seen top 10 Neighbourhoods in the NYC having maximum listings.*
- *3. Prediction on different columns of dataset (exp :-price, availability etc.):- Manhattan has highest mean price and Bronx is minimum.*
- *We can also summarize that Manhattan has highest noise in price followed by Brooklyn and so on.*
- *4. Which are the busiest host and why :- the busiest host is Sonder(NYC) followed by Blueground and Kara. These hosts are busiest because they have higher number of listing and availability is less.*

Challenges :

- *We have faced challenges in selecting the columns for the analysis.*

- *Although there were only few columns which had missing values but it was a bit difficulty in treating these missing values.*

- *We also have face some challenges in choosing the right plot as visualization is the most important part of the project hence selecting the appropriate features was very important.*

Thank you

Regards:

Mohd Sharik

Prince Nayak

Gokul Jagannadh