Capstone Project 1

Airbnb Prediction

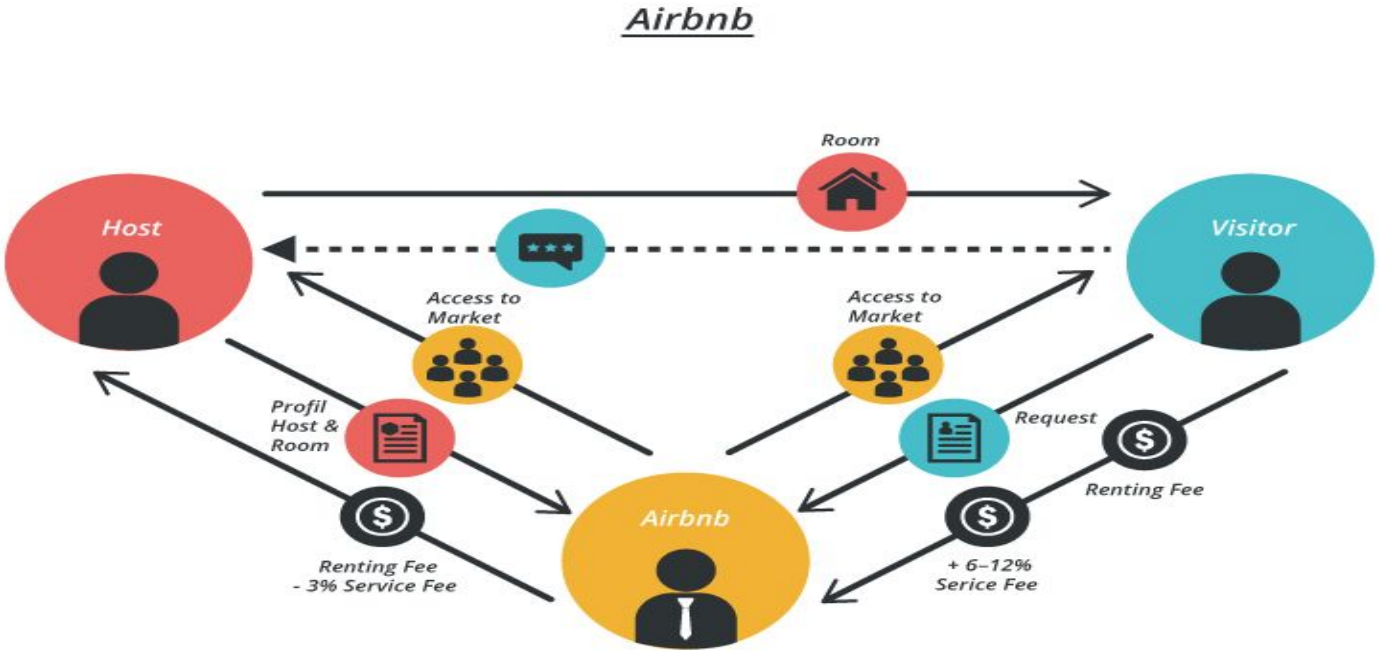
Team Members

Shanawaz Anwar
Varshith Kola
Mrinmoy Kar

Finding The best places to Stay

1. Understanding The Problem Statement
2. Analyzing The Data
3. Cleaning and Preprocessing The Data
4. Visualizing The Data using ML Libraries
5. Finding Inferences from the Data

Airbnb Business Model



Data Pipeline

- Data Preprocessing 1 : In the first phase, we get rid of any unwanted columns from the data
- Data Preprocessing 2 : In this phase, We take care of missing values and do the type casting
- Visualizing Data(EDA) : In the final phase,We analyze the data and plot different different plots like bar chart, pie chart, scatter plot etc

Challenges

- There were 4 columns with null values
- We performed different kinds of imputation to impute the null values
- Or if it was required we dropped rows with null values

Null Values

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

Data Summary

- **id : It is an unique ID which represents a particular row**
- **name : This column represents the name of the Airbnb listing**

Data Summary

- **host_id : Unique identification of hosts when they are listing on Airbnb**
- **host_name :Name of the host**

Data Summary

- **neighbourhood_group**: It consists of several neighbourhoods categorized under 1 major group, there are such groups in this column
- **neighbourhood** : It has names of all the neighbourhoods

Data Summary

latitude : Describes the horizontal coordinates of the listing

longitude :Describes the vertical coordinates of the listing

Data Summary

room_type : Describes the types of rooms listed in listings by the hosts. There are three category of listings in his column

price: It has the price of the different type of rooms keeping in mind different neighbourhood groups

Data Summary

minimum_nights : The minimum number of nights that a tourist has to book

number_of_reviews : Number of reviews that a particular place or a listing got. With this we can determine how much popular that place is amnog the tourists

Data Summary

last_review : When was the last time that place was reviewed, this demonstrates how frequently, the host lists his place

reviews_per_month : Number of reviews a place is getting per month on average

Data Summary

Calculated_host_listings_count : How many times on average does the host lists his property

Data Summary

availability_365 : availability of the places listed by the host throughout the year

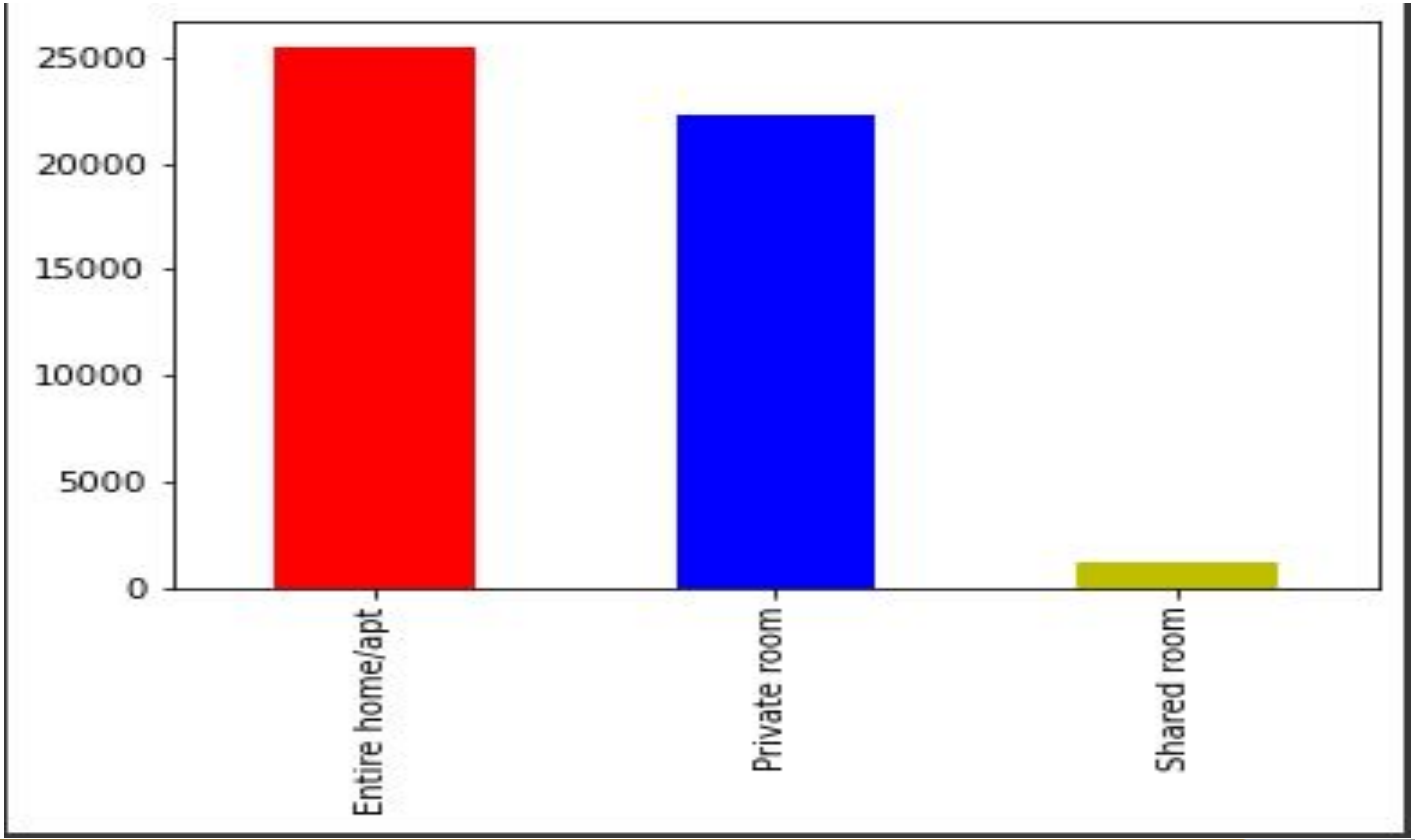
Description of The Data

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000

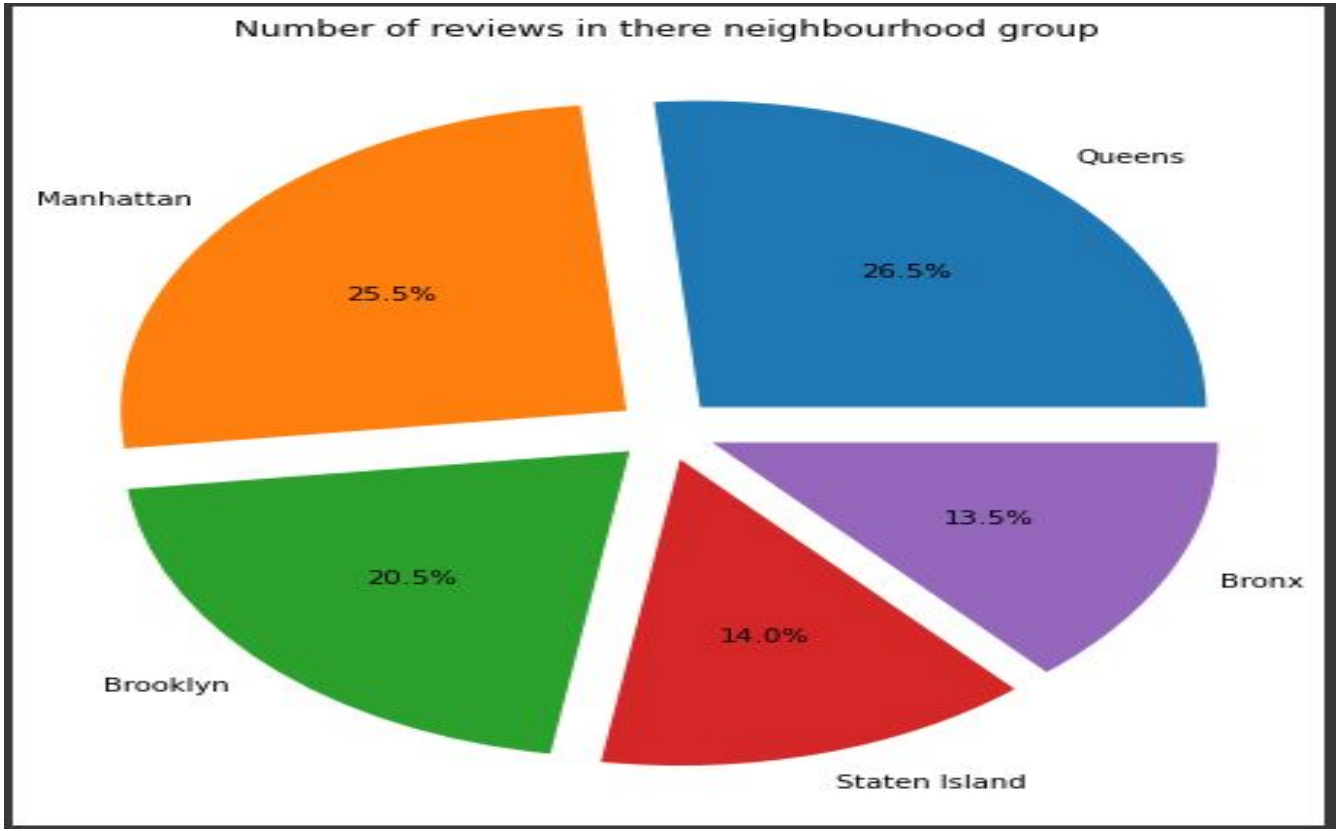
Count of Listings According to Neighbourhood

	host_name	neighbourhood_group	calculated_host_listings_count
13221	Sonder (NYC)	Manhattan	327
1833	Blueground	Brooklyn	232
1834	Blueground	Manhattan	232
7275	Kara	Manhattan	121
7478	Kazuya	Brooklyn	103
7479	Kazuya	Manhattan	103
7480	Kazuya	Queens	103
6540	Jeremy & Laura	Manhattan	96
13220	Sonder	Manhattan	96
2901	Corporate Housing	Manhattan	91

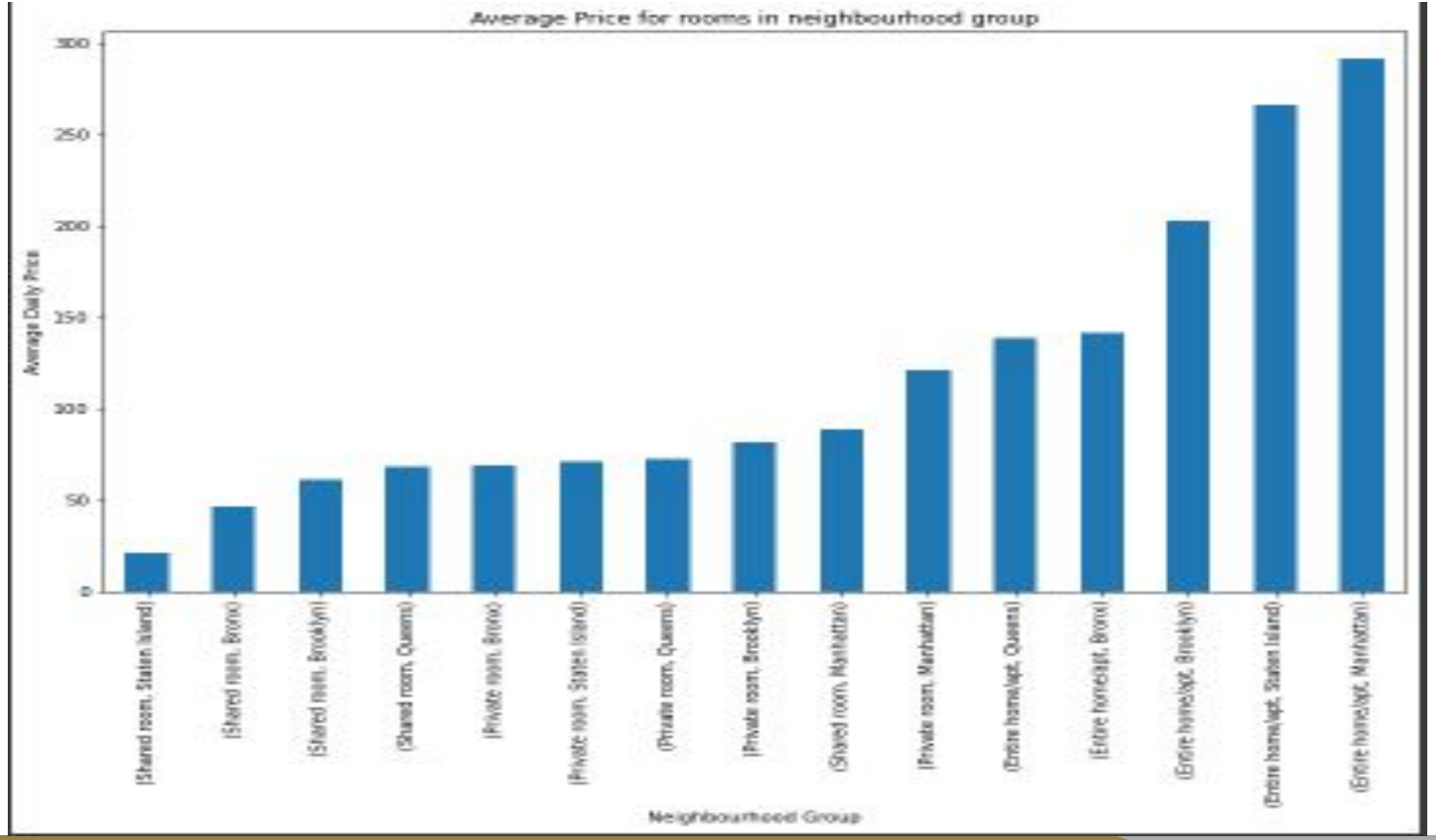
1)Most Preferred Room Types



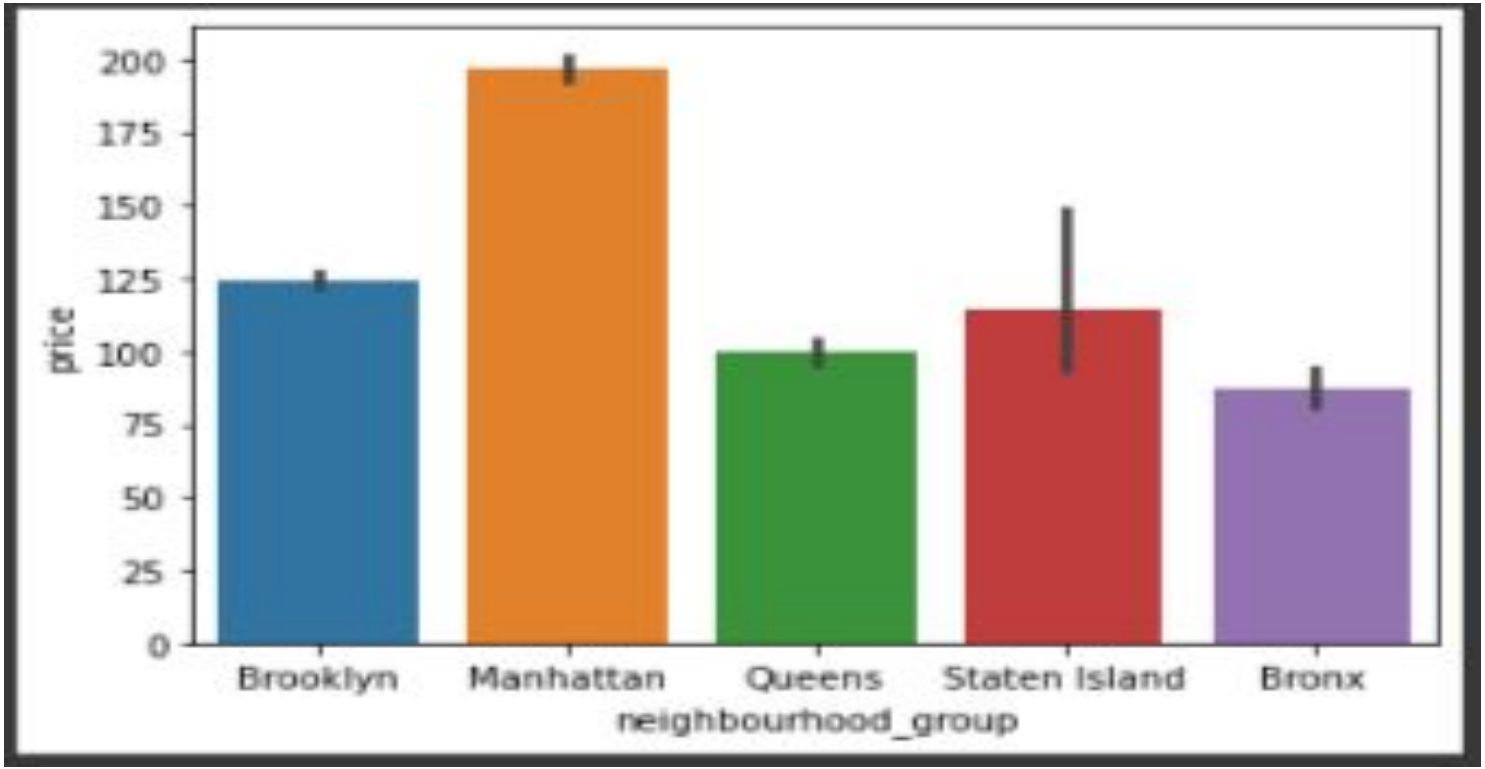
2) Most Reviewed Neighbourhood Group



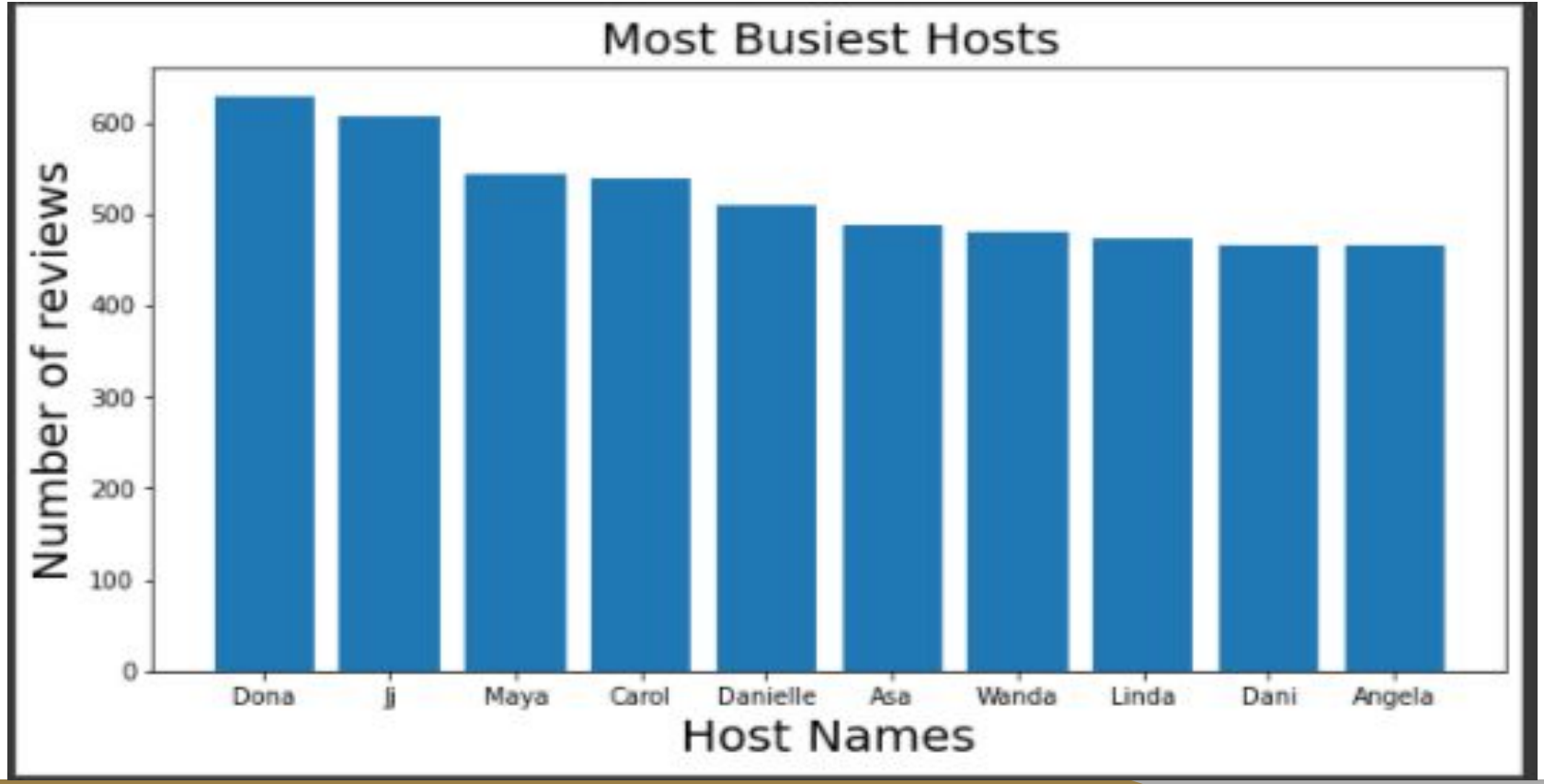
3) Avg Price of Rooms



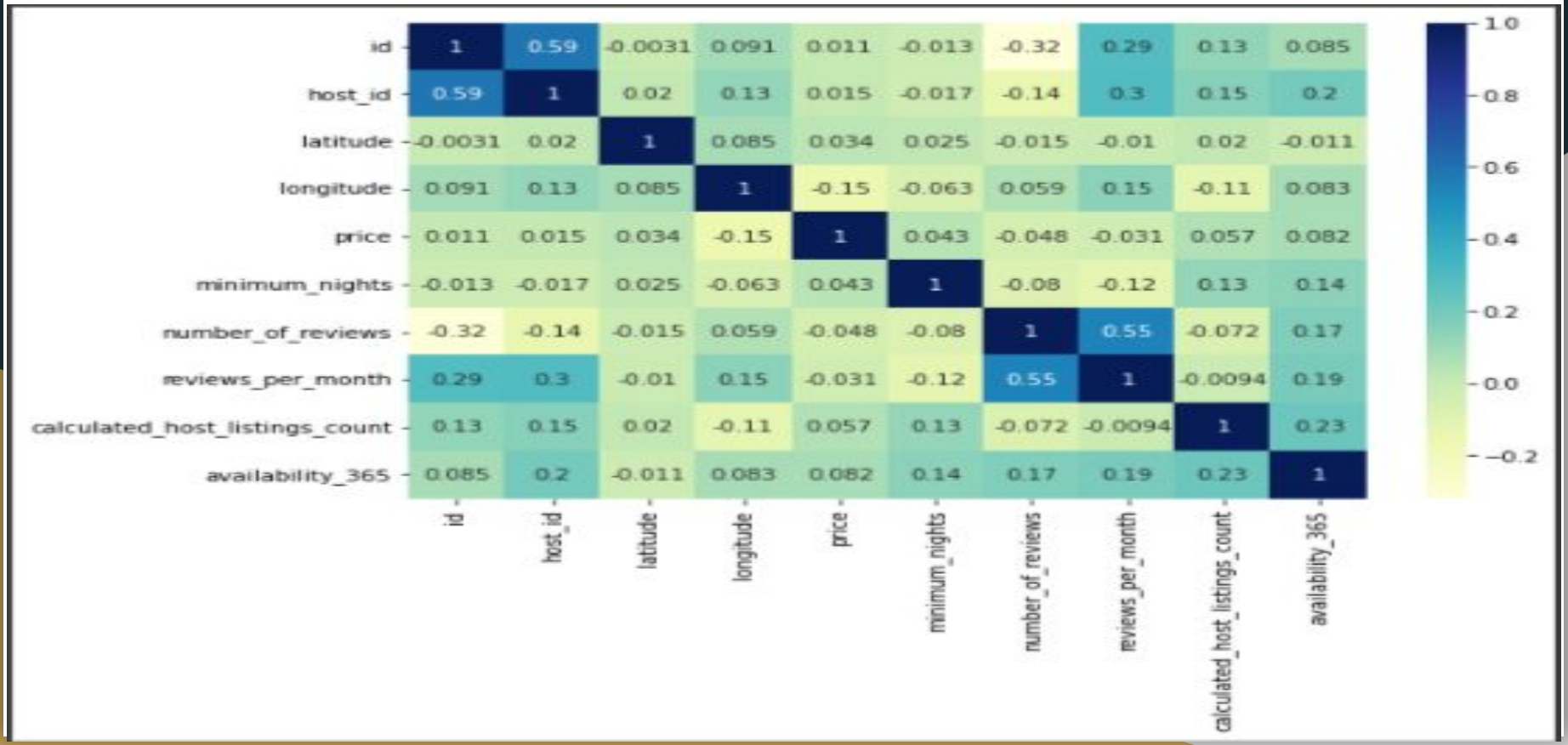
4) Price Comparison of Different Neighbourhood Groups



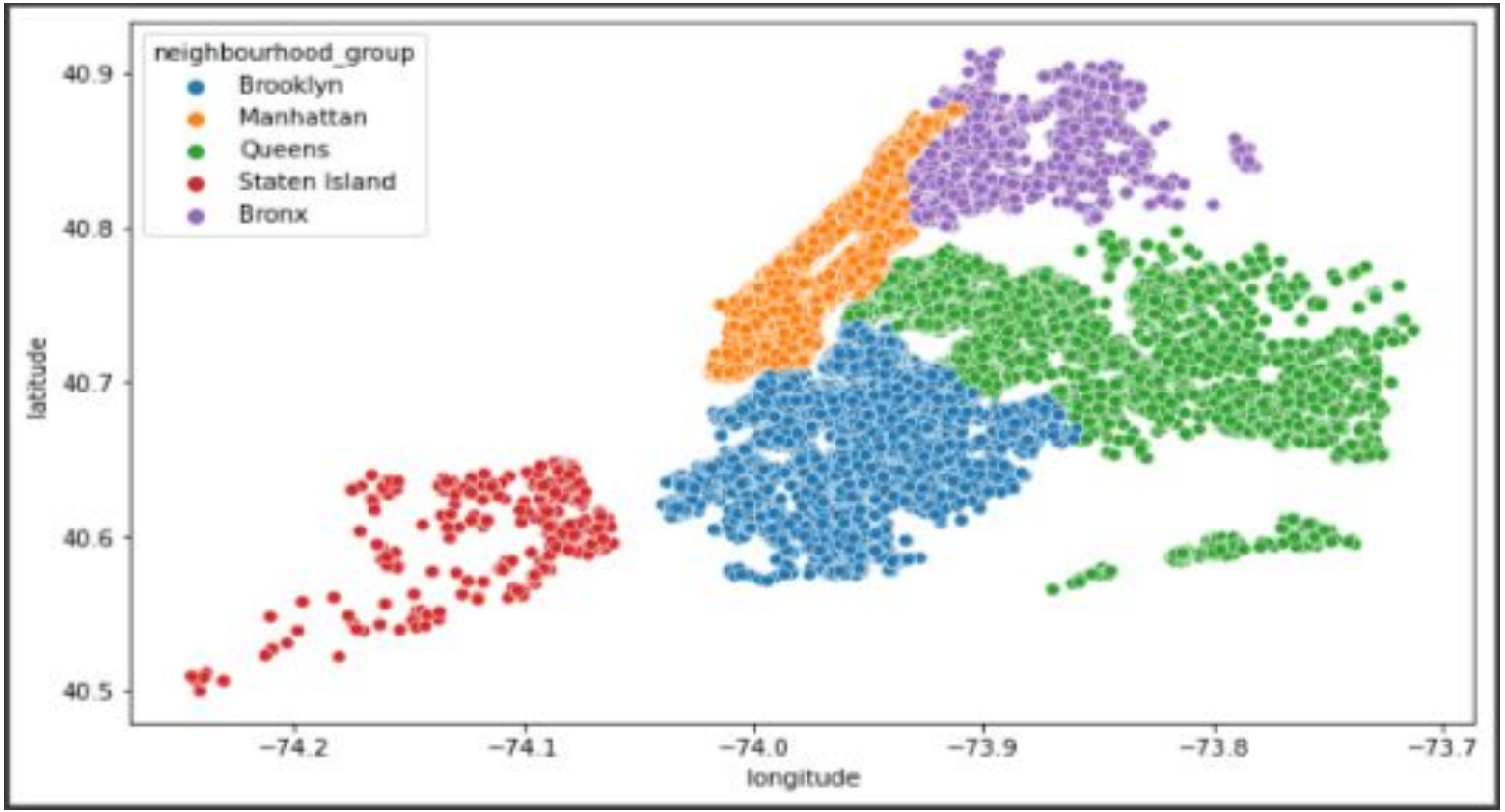
5) Most Busiest Hosts



6) Correlation Between Different Var



7) Listings in Different Geographic Regions



Libraries Used

- Pandas
- Matplotlib
- Seaborn
- Numpy

Conclusion

- Sonder has the highest number of listings and neighbourhood is Manhattan
- Most Preferred room type is ‘Entire home /apt’
- ‘Queens’ is the most reviewed neighbourhood group

Conclusion

- Manhattan has the highest ‘avg daily price’ for the room type ‘Entire home /apt’
- Manhattan also has the highest avg price for any room in the 5 neighbourhood groups

How can this analysis help ?

- From the analysis we can help customers to choose the best places to stay
- This analysis also helps the hosts,through which they can better their customer service
- This analysis helps customers save heir time and money

THANK YOU

Q&A