CASE STUDY REPORT

Detailed Summary of Milestone 1-3

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset used for this project describes the listing activity and metrics in NYC, NY for 2019. This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions such as it has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Business Problem/Data

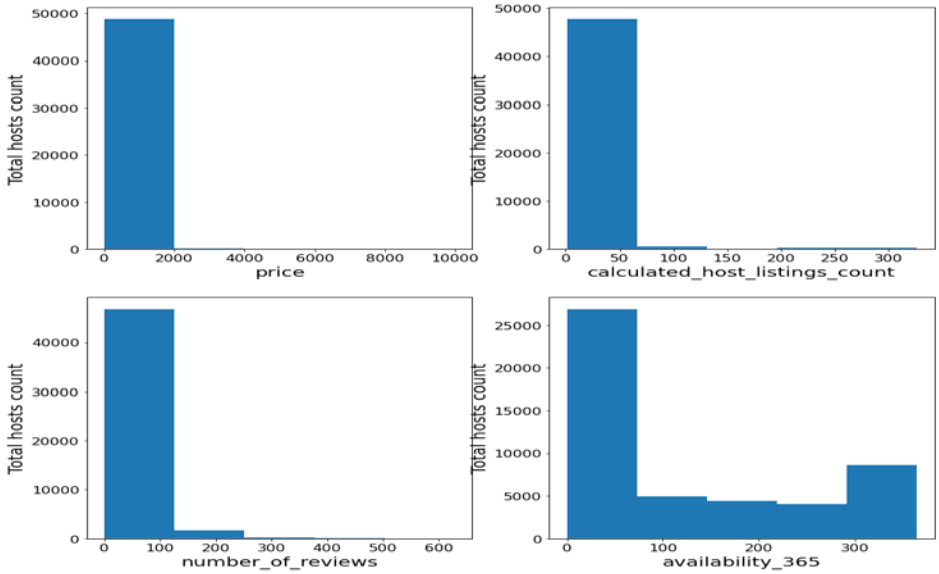
For milestone 1, I used Kaggle for the data which was from Airbnb 2019 data. From the dataset I wanted to learn about different hosts and areas, which hosts was the busiest and what type of room different neighborhood had customers really wanted to take on usual basis. For the graphical analysis I used neighborhood type and room type to understand which borough had high popularity and which room type had high demand respectively. Below we have the data frame showcasing features (I took 2 snapshot since it couldn't fit in 1) which were later on manipulated and also the graphs showing different story from our dataset. In the following graphs we can see that the prices for the Airbnb ranged between less than 2000, number of reviews were mostly 0-100 reviews for the hosts, highest availability fell between 0- 90 days whereas for the neighborhood group which had highest demand/occupancy were in the order of Manhattan, Brooklyn, Queens followed by Bronx and little on Staten Island. Additionally, the room type which had highest demand/occupancy

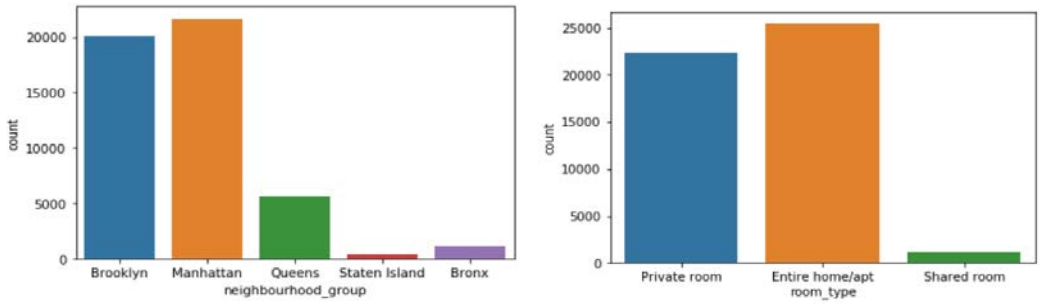
were Entire home/apartment followed by private room and lastly shared room which had less than 2000 occupancy out of 48,000 observations.

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	l
2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	

last_review	reviews_per_month	calculated_host_listings_count	availability_365
2018-10-19	0.21	6	365
2019-05-21	0.38	2	355

Graphical Analysis





For milestone 2, I dropped ‘id’, ‘name’, ‘host id’, ‘host name’ columns which were irrelevant to the model building. The reason being that later on I wanted to learn more on the neighborhood, room type which didn’t need above features. After the comments provided in Milestone 2, I decided to drop other features which were irrelevant for my model building such as availability 365 column which showcased number of availability of room in 1 year.

	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	calculated_host_listings_count	Brooklyn	Manhattan	Queens	Staten Island
0	40.64749	-73.97237	Private room	149	1	9	6	1	0	0	0
1	40.75362	-73.98377	Entire home/apt	225	1	45	2	0	1	0	0
2	40.80902	-73.94190	Private room	150	3	0	1	0	1	0	0
3	40.68514	-73.95976	Entire home/apt	89	1	270	1	1	0	0	0

Dimensionality & Feature Reduction and Feature Engineering

In Milestone 1, I learned about the features with stories told through visuals/graphs. Well, in Milestone 2, I found other analysis through numbers using functions such as nunique(), value_counts(), groupby()[ ].count(), .sum(). An observation I learned was that "219517861" host_id is the busiest one with 327 times he attended a guest. His name is "Sonder (NYC)" in neighbourhood_group "Manhattan". Further I learned by using the function df[‘neighbourhood_group’].nunique() that Manhattan borough had 21,661 total guests using Airbnb followed by Brooklyn with 20,104 guest, Queens with 5,666 guest, Bronx with 1091 guest and Staten Island with 373 guest. Another observation I found was that Entire home/apartment room type had 25,409 guest whereas Private room had 22,326 guest and Shared room had 1160 guest. In addition, I had other observation made as well such as groupby(‘room_type)[‘Price’], groupby(‘neighborhood_group’)[‘room_type’] to name a few.

To make changes from Milestone 1, I used seaborn which is a Python data visualization based on matplotlib. I wanted to see how different it would look using different package. I found it colorful and more attractive for display which showcased frequency of guest on type of room they chose. Another graph I updated using seaborn was using neighbourhood_group to see which Borough had highest to lowest guest comparatively. I would like to admit that I liked seaborn package with countplot() function which can be thought of as a histogram across a categorical instead of a quantitative variable. It was colorful, engaging and to the point graph. I created dummy variables for neighborhood group feature since they were in categorical variables such as Manhattan, Queens and so on. After creating the dummy variables, I dropped the neighborhood_group column which had categorical variable.

Model Selection & Evaluation

For milestone 3, I used logistic regression classification for machine learning model building. I used this model because it provided us with solution on predicting what type of rooms customers usually wants to take. However, the result was not fruitful because of which I didn't get adequate result. Hence, I changed my direction towards decision tree classification where based on the decision tree classification model I got the following result: Entire home/apt had 84% precision, 86% recall and 85% f1 score. Private room had 81% precision, 81% recall and 81% f1 score. Shared room had 65% precision, 34% recall and 45% f1 score. Accuracy is 82% which means that the model has 82% correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Based on above we have higher precision on entire home and private room which means that we have low positive false rates. Recall is the ratio of correctly predicted positive observations to the all observations in actual

class - yes. F1 Score is the weighted average of Precision and Recall. Accuracy is 82% which means that 82% of ratio is correctly predicted observation to the total observations.

Conclusion

Based on using two different model (logistic regression classification and decision tree classification) it was evident that decision tree classification was the right choice. Other analysis I could make based on above analysis from all the milestone is that Airbnb is highly used or we can assume had high number of guest in Manhattan followed by other Boroughs such as Brooklyn, Queens, Bronx and Staten Island respectively. Another stunning observation I found was how low guest were in Shared room in comparison to Entire room/apt and Private room. Recommendation based on above finding I would like to make to Airbnb is to find more on why there has been low frequency of guests in Borough such as Queens, Staten Island and Bronx. Compared to other Manhattan and Brooklyn had similar frequency count. Similarly, another recommendation would be to increase the listings of Shared rooms. Compared to Entire home/apt and Private room the listing for Shared room was highly low. This could be an area Airbnb can focus on because there might be many tourists who would like to visit New York city but are unable to afford the cost of staying in Entire home/apt or Private room in the city which could be costly at times. Shared room are economical in nature and it can attract more tourist. The challenge for the project was that the dataset I retrieved was only from 2019 and not few years which could have brought our more samples and observations to manipulate on.