Capstone Project 1
Airbnb Booking Analysis

Team Members

Piyush Lanjewar
Pruthvi Raj
Yogesh Reddy

Content

- Introduction
- Problem statement
- Handling missing values
- EDA on given data
- Visualizations
- Conclusion

Introduction

Airbnb is one of the largest used companies for lodging primarily homestays for vacation rentals and tourism activities. Today Airbnb is one of the most used brands for giving a good experience to hosts and guests.

The data is used to increase the understanding of every detail to make traveling easy

and convenient. The data is utilized to show the required conclusions. The conclusions are also shown in the visualizations to make understanding easier.

Problem Statement

1) The primary factor of AIRBNB is to provide Homestays for Vacation rentals and tourism activities and none of the listed properties are owned by AIRBNB. It is just an online marketplace for all the listed properties.

2) The factor that affects this business is the Reviews. If the visited customers provide a positive review then there is a high chance that it gets booked several times based on the availability and in case of negative reviews customers are most likely don't prefer to stay there.

3) By observing the provided data around 25% of the listed properties either be Home-apt. Or private rooms or shared rooms don’t have the reviews this doesn't help to maintain a healthy relationship with the hosts of the listed properties.

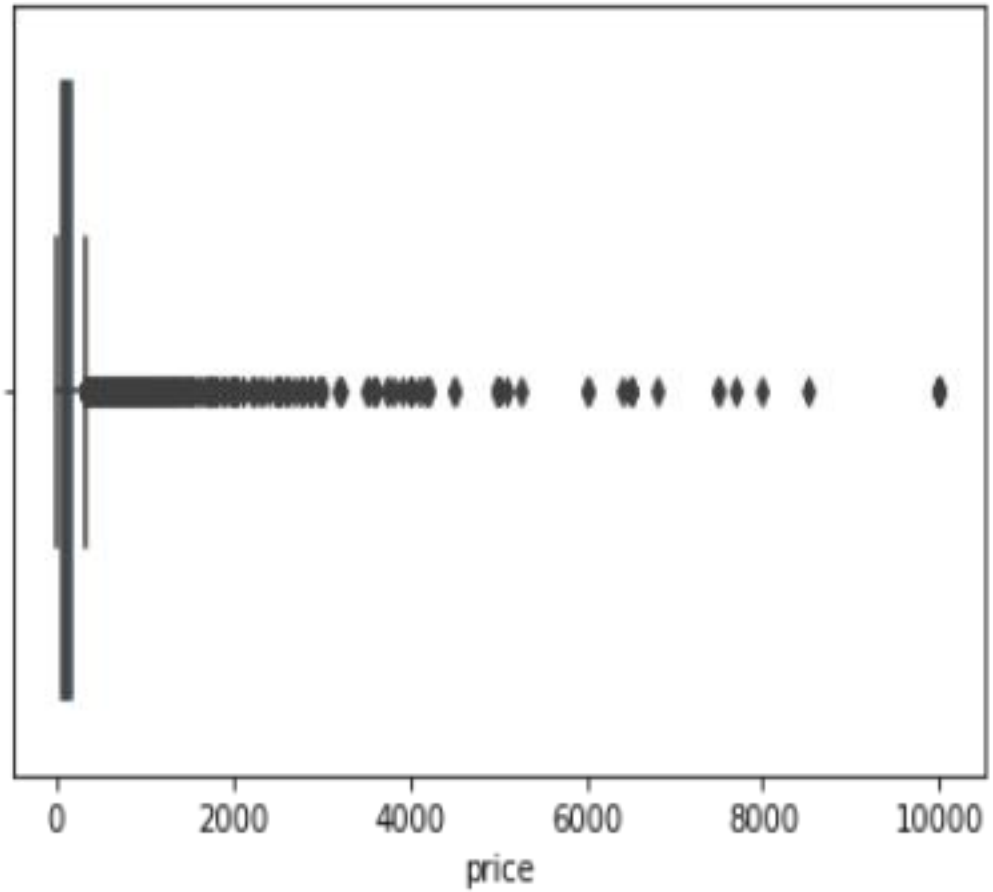
4) There is small inaccuracy in the dataset on price where around 10% of the properties are given with the price 0$ which doesn't relate at all. So to represent real data we are using median instead of mean.

Column features

- host_name: The Name of hosts who give services to guests
- Neighbourhood_group: Represents the city
- Neighborhood: Represents areas of the city
- Latitude and longitude: Represents the location of the house
- room_type: Represents the type of room(shared/private/apt)
- price: Represents price of the houses
- minimum_nights: Nights spent by customers
- number_of_reviews: Number of reviews
- last_review: Date represents the last review by customers
- reviews_per_month: Reviews per month
- calculated_host_listings_count: Host count listing
- availability_365: The availability of hosts per year

The box plot represents the Price Column showing too many outliers present in the data.

So, we will use the median in place of zero values and NA values.



	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000
mean	40.728949	-73.952170	152.755045	7.029962	23.274466	1.090910	7.143982	112.781327
std	0.054530	0.046157	240.143242	20.510550	44.550582	1.597283	32.952519	131.622289
min	40.499790	-74.244420	10.000000	1.000000	0.000000	0.000000	1.000000	0.000000
25%	40.690100	-73.983070	69.000000	1.000000	1.000000	0.040000	1.000000	0.000000
50%	40.723070	-73.955680	106.000000	3.000000	5.000000	0.370000	1.000000	45.000000
75%	40.763115	-73.936275	175.000000	5.000000	24.000000	1.580000	2.000000	227.000000
max	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

The list represents the average value of the room types corresponding to the Neighborhood group. The list can give us the estimate of price listing in different areas.

room_type	neighbourhood_group	price
Entire home/apt	Queens	139.036260
	Bronx	141.541176
	Brooklyn	202.895245
	Staten Island	266.205128
	Manhattan	291.784595
Private room	Bronx	69.025862
	Staten Island	71.394366
	Queens	72.454958
	Brooklyn	81.859242
	Manhattan	121.434183
Shared room	Staten Island	21.000000
	Bronx	46.711111
	Brooklyn	60.921212
	Queens	68.459459
	Manhattan	88.462898

EDA

What is EDA?

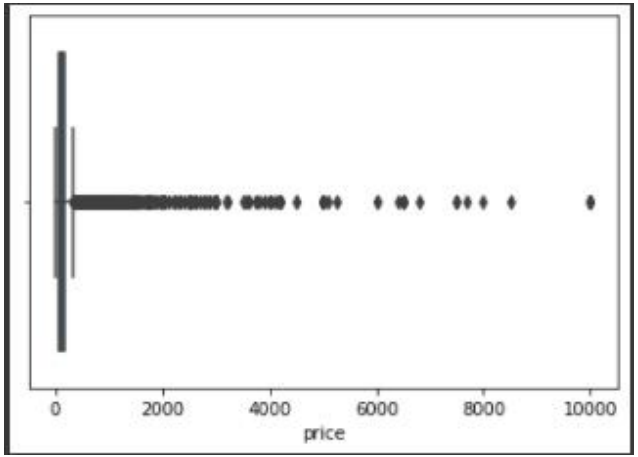
EDA is an approach to analyzing data sets using statistical graphs and many visualizations tools. EDA shows us what data can tell beyond its formal modeling or hypothesis testing task. It provides a better understanding of the data and the dependence of its features over each other. It helps us to find and handle missing values and outliers. It helps us to manipulate the data sources and get the answers we need from the data. We get a better understanding of the problem statement.

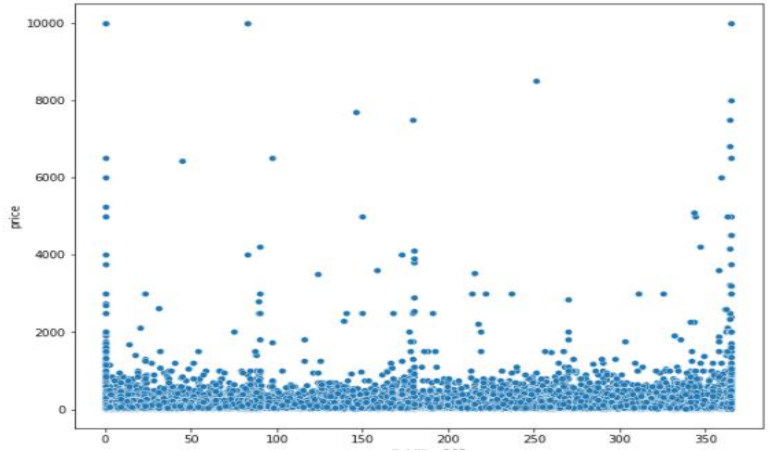
Here in our problem, we have done some visualizations to understand the data.

Outlier Data:

Outliers are the data points that affect the data at the model building which gives wrong predictions. The accuracy of the model is affected.

Here in our data, we are taking price and availability_365 for checking outliers.



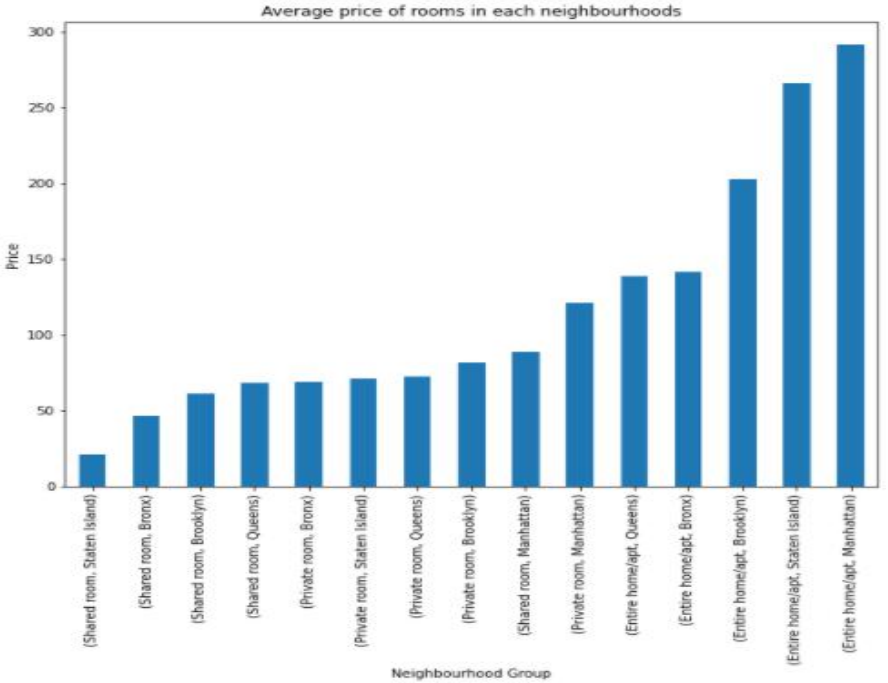


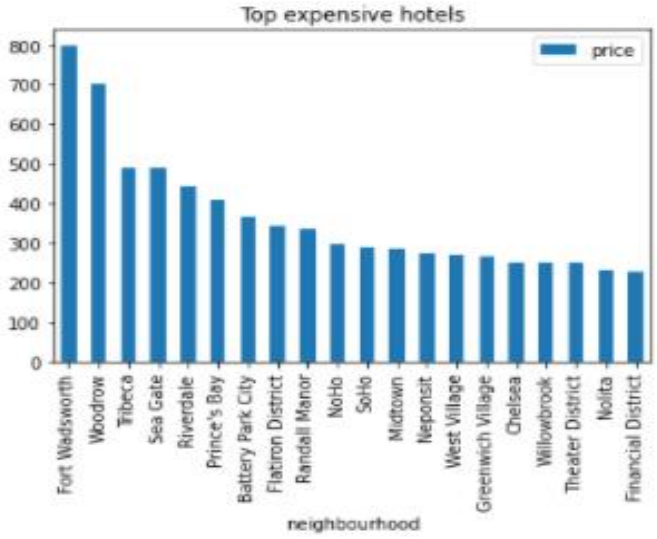
In the above two plots, we can see there are some outliers in the price feature.

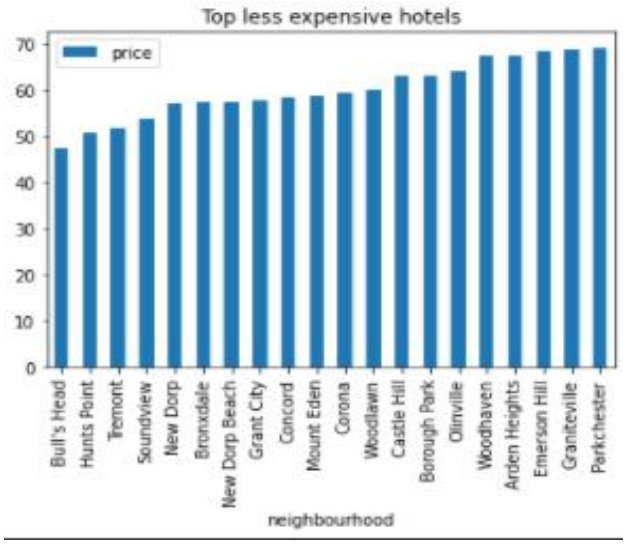
Visualizations

In this bar plot, we have visually shown the average price of the different rooms for different Neighborhood Groups.

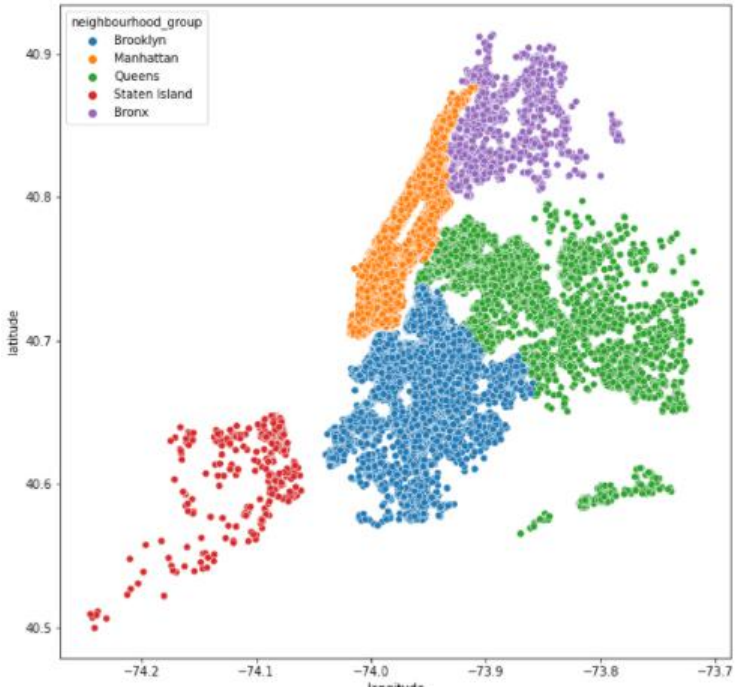
We found an Entire apt in Manhattan has the highest The average price of $291 and shared room in Staten Island has the less average price of $21.

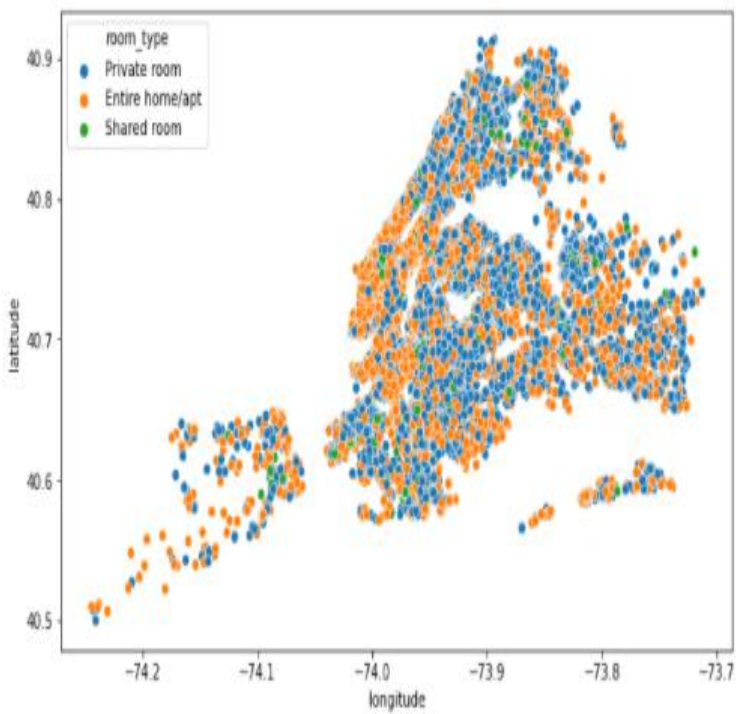




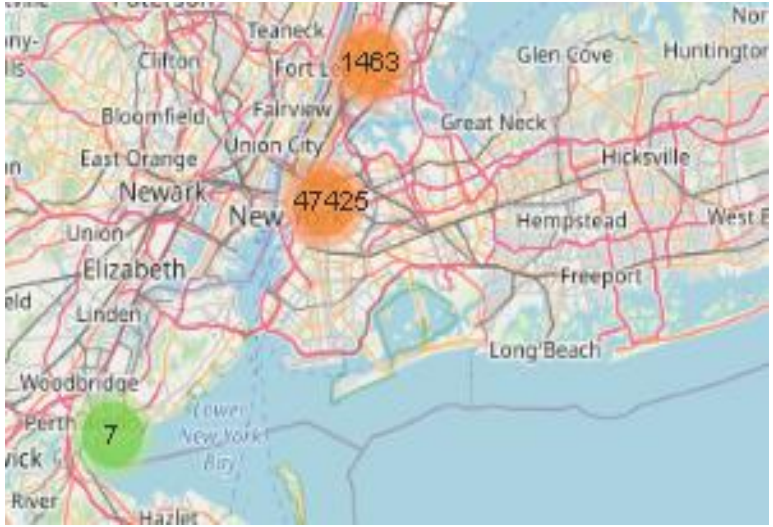


In the above bar plots, we found the top expensive hotels and less expensive hotels.
The top expensive hotel in Fort Wadsworth at $800 and the less expensive hotel is Bull’s Head at $47.

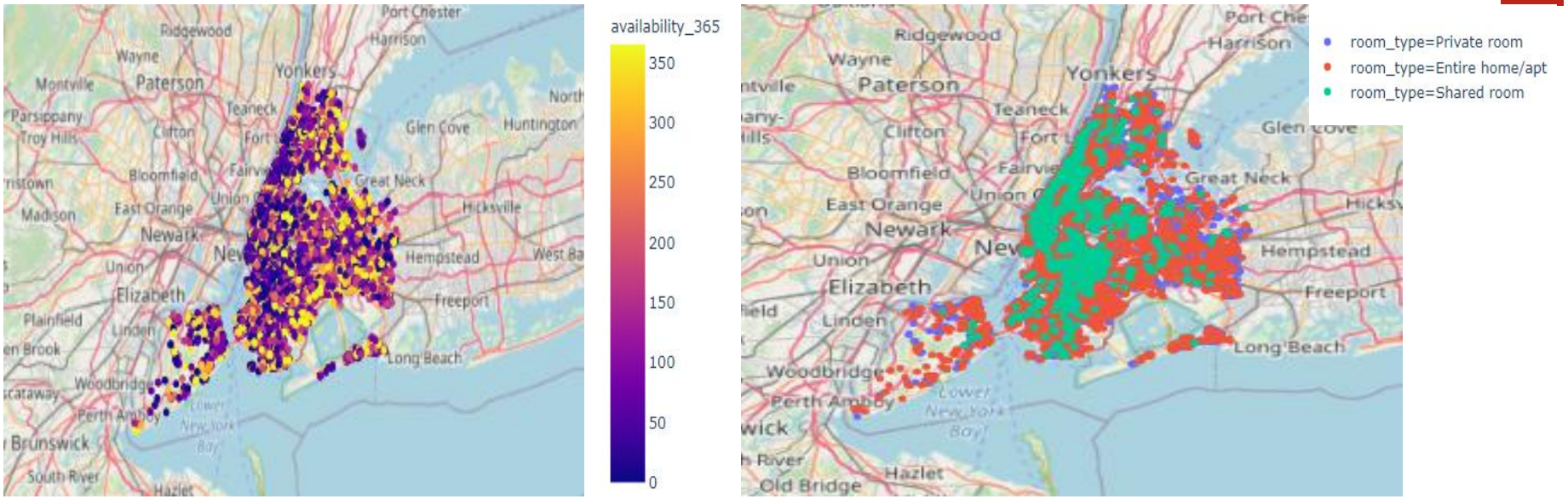




In the above two scatter plots we have used latitude and longitude data to visually see the neighborhood groups and the types of rooms shared across the New York City map.



In the above two plots, we are showing the total hotels in clusters we can see the two images one is zoomed out and the other is zoomed in to see the hotels spread across the map.



In the above two plots, we can see the hotels available for availability for 365 days and in the second plot, we can see types of rooms spread across the NYC map.

Conclusion

The data of AIRBNB since 2008 has been provided this dataset. This dataset has around 49000 observations with 16 columns and it is a mix of numerical and categorical values.

❖ The detailed Insights from the analysis:

The Neighbourhood of New York City has 5 groups:

- Brooklyn
- Manhattan
- Queens
- Staten Island
- Bronx

- Properties from Manhattan are a bit pricey followed by Brooklyn and Staten Island
1. Top 3 Hosts from the dataset are:

- Sonder(NYC)
- Blueground
- Kazuya

- Since Reviews are important here are the top 3 Hosts who hold the most reviews:

- Dona
- Asa
- Dennis & Nauko

1. There are 3 room types Entire home/apt. (~25000), Private room(~23000), Shared room(>500)
2. The visualization that is plotted on maps also shows where all the properties are located along with their price, availability, and the type of room.

Thank You