AlmaBetter EDA Capstone Project 1 Airbnb

This project is a part of the AlmaBetter Premium Program , Banglore/Bengaluru ,Karnataka , India

The data we are going to analyse is the data of Airbnb NYC (2019). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas, price, reviews, locations etc. but not limited to.we will also try to explore some more insights.

Project Status: [Completed]

About the dataset:

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

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.

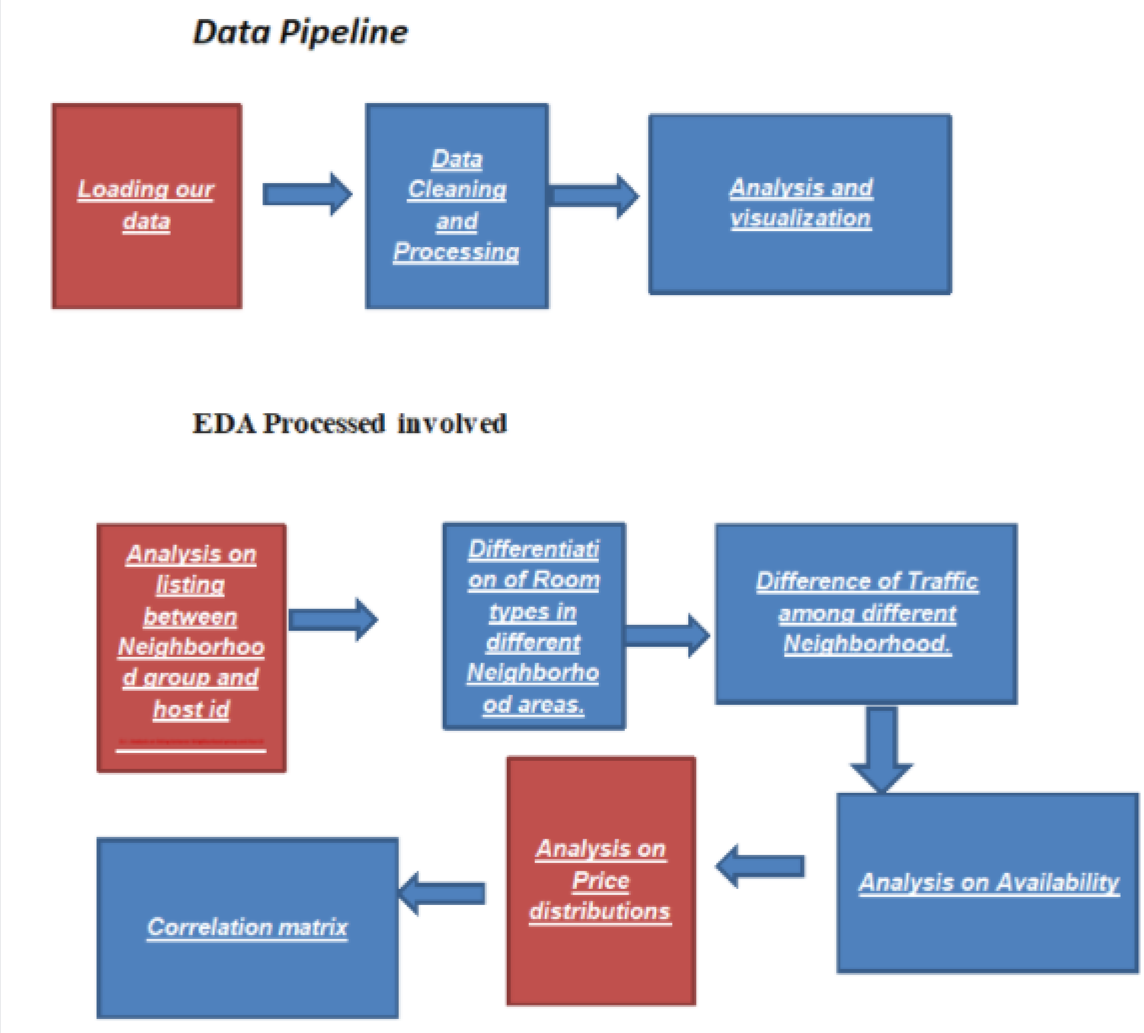
The data we are going to analyse is the data of Airbnb NYC (2019). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas, price, reviews, locations etc. but not limited to.we will also try to explore some more insights.

Approach used:

the approach we have used in this project is can be defined in the given format-

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 : EDA on hosts, neighborhood, neighborhood groups and room types. 3.2 : EDA on price distributions 3.3 : EDA on availability, reviews and correlation matrix


Data Pipeline
Loading our data
Data Cleaning and Processing
Analysis and visualization
EDA Processed involved
Analysis on listing between Neighborhood group and host id
Differentiation of Room types in different Neighborhood areas.
Difference of Traffic among different Neighborhood.
Analysis on Availability
Analysis on Price distributions
Correlation matrix

Future scope of work:

Working out on Top neighbourhoods and group and compare their positions based on different indicators Considering the amount of indicators in the data, if we dig deep enough, various micro trends can be unearthed, which we were not able to extensively cover during this short duration. This data can also help to understand varioues properties and their demands. The review column can be helpful to get insights from different neighbourhoods.

Data Visualization Methods Used

- Pie Chart Plots
- Count Plot
- Bar Plot
- Scatter Plot
- Heatmap

Python Libraries used

For Graphing :

- Matplotib
- Seaborn
- Numpy
- Pandas

What is EDA?

“Exploratory Data Analysis is a detective work. EDA can never be the whole story, but nothing else can serve as the foundation stone — as the first step”. This is a quote by the renowned Data Scientist John W. Tukey in 1970. In simple language, exploratory data analysis (EDA), also known as Data Exploration is a step in the Data Analysis process, where several techniques are used to better understand the dataset being used. The key steps involved in EDA are

- Acquire and loading data
- Understanding the variables
- Cleaning dataset
- Exploring and Visualizing Data
- Analyzing relationships between variables

Understanding the Variables

In this session, we do the basic inspection of our data set and familiarise the columns. In the Airbnb data set of NYC, we have 16 columns in total, which is a combination of numerical and categorical variables. By doing the basic inspection, we can easily figure out which ones are categorical variables and which ones are numerical variables. Some columns are not significant for our analysis which can also be kept off. Now let’s look at some of the useful columns in our data set.

Host Id

· Host Id is the government approved id for each individuals who rents their properties on Airbnb.

· This is one of the numerical variables associated with each host.

· There are about 37457 unique values in the data set.

· There exist multiple listings corresponding to a particular host id.

Host Name

· Host names are basically the names of the individuals or organisations who rent a rooms/apartment in Airbnb website.

· In order to protect the privacy of both hosts and guests, they don't share last names until after a booking is confirmed.

· Also there are three types of hosts.

o Listing owner: The person who lists the space in their Airbnb account. This is usually the person who owns the property or living in the property.

o Co-Host: Someone, usually a friend or family member, who helps the Host manage their place. The listing owner decides whether the co-Host appears as the primary Host on the listing or not.

o Hosting team: A hosting team is a business or team of people that manages places to stay on behalf of the listing owner.

· There are about 11453 unique values out of 48895 observations.

· This variable is categorised as a categorical variable since a particular individual or organisation can own multiple types of rooms.

Neighbourhood

· When searching for accommodations in a city, guests are able to filter by neighbourhood attributes and explore layers of professional-quality content, including neighbourhood maps, custom local photography and localized editorial, details on public transportation and parking, and tips from Airbnb’s host community.

· By looking at the neighbourhoods, the guest can match neighbourhood’s personality with their own.

· In Airbnb dataset, neighbourhood is a categorical variable

Neighbourhood groups

· Neighbourhood groups are the clusters of neighbourhoods in the area.

· In NYC the neighbourhood groups are generally the boroughs. There are about 5 boroughs in the state.

· Since there are many neighbourhoods in each borough, it is a categorical variable.

Room type

Airbnb has 3 categories for types of spaces:

· Entire house/apartment

· Private room,

· Shared room.

Entire house or apartment generally means a full unit with bedroom, bath and kitchen. Private room generally means you get your own private bedroom where no one has access to (i.e. you are not sharing it with the host or another guest). Shared room means you are sharing a room with someone who is another guest or your host. You are also supposed to stay on the couch in the host's living room, so it is not really private, it's a living room during the day, and you sleep there at night.

Price

The total price of your Airbnb reservation is based on the rate set by the Host, plus fees or costs determined by either the Host or Airbnb.

Types of fees

· Airbnb service fee: Guest service fee charged by Airbnb, this provides 24 hours community support and helps everything run and ensure the system runs smoothly.

· Cleaning fee: Charged by some Hosts to cover the cost of cleaning their space (applicable to all countries except China).

· Extra guest fee: Charged by some Hosts for each additional guest beyond a set number.

· Security deposit: Some reservations may require a security deposit requested by the Host or Airbnb—find out more about security deposits.

· Value Added Tax (VAT, JCT, and GST): Charged to guests who live in certain countries—find out more about VAT.

· Local taxes: Charged based on the location of the Host's place—find out more about local taxes.

Other relevant variables

· Reviews per month: insights into frequency of visits of the listing

· Minimum nights: indicator of minimum stay length, to be used with the number of monthly reviews

· Availability 365: It is an indicator of the total number of days the listing is available for during the year.

Data Pipeline

EDA Processed involved

4. Null Value treatment

Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

5. Data Visualisation

Data visualization is the graphical representation of information using visual elements like charts, graphs, and maps. Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions. Here we make use of the tools matplotlib and seaborn to visualize data. The distributed plot, bar plot and scatter plot are the few graphical representa

Final Summary

This Airbnb-NYC(2019) dataset is very informative dataset having 48895 rows and 16 columns. we have found that any host can have many id's and host_id's based on their properties(2.2). we have found hosts that take good advantage of the Airbnb platform and provide the most listings; we found that our top host has 327 listings.Then we have seen that Manhattan has highest number of listings followed by Brooklyn and so on. After that, we proceeded with analyzing neighbourhood groups ,neighborhoods and found that in top 10 neighbourhoods only Manhattan and Brooklyn groups take part . Then we moved to analyze price and observed that Price is more distributed across the dataset in a specified range (20,300). also there are some outliers.Another very interesting insights 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 . In final section we observed that we have noticed that the minimum avg. availability is in Brooklyn followed by Manhattan and the maximum availability is in Staten. then we have found top 10 busiest hosts across NYC and noticed that Micheal is the busiest host followed by David and Sonder(NYC). This order is the as we had seen in section 2.1 for most listings. This can be due to high number of reviews in the neighbourhood. Overall, we discovered a very good number of interesting relationships between features and explained each step of the process. This analysis can help leaders in high level business decisions, control over the platform, marketing initiatives, implementation of new features and much more.

Contributing Team Members:

Name	Email
ABDUL QUADIR	abdulec1002@gmail.com
Mohammad Atique Najmuddin Shaikh	Atiqueshaikh0141@gmail.com