EDA on Airbnb NYC Dataset

Vinayak Marathe, Riya Patel,

Muskan Kasere

Abstract - The purpose of the project of Airbnb bookings is to do multiple manipulations on Airbnb dataset and find some meaningful insight from it. The research problem is to find the best listing in Airbnb NYC based on their price, location, and customer reviews. We have used many python library like numpy for scientific computing, pandas for data manipulation, matplotlib for data visualization and graphical plotting, seaborn for statistical graphics plotting, wordcloud to find the most frequently used keyword & folium for visualizing the map of New York City. Then, we performed some data analysis and found different business insights.

Keywords: numpy, pandas, matplotlib, seaborn, folium, word cloud

I. INTRODUCTION

Airbnb is identified for "Air Bed and Breakfast", is an American online hospitality company focused on short time length homestays or rental platform. On Airbnb, users can book a nearby place to stay according to their convenience in more than 34000+ cities in some 200+ countries. Airbnb started in 2008 by Brian Chesky and Joe Gebbia, specifically based in San Francisco California. Airbnb platform is reachable via mobile app and website. The company does not own any lodging, even though it is a type of broker which receives percentage services fees from both the party (guest & host) with every booking. As we know, Airbnb is an online platform started in 2008 offering homestays but does not own any property because it works as a broker b/w customers and hosts and earn value by commission from each booking.

After 2009, many property owners opened their doors to Airbnb. It was a great success but the hosts always have a problem with "What price they need to put on their property?" The consequence of the problem was if the price set too high, the number of customer lending properties will be less and if set too low, the host will be at the loss.

In order to explain the nature of our data, the data is collected and initialized in the first stage then we have checked our data with the shape, size and type. Then in the second stage, we are focusing on our variable and understanding the behaviour of our variables. At the third stage, we have analysed our data with different Exploratory Data Analysis (EDA) techniques, accompanied with the aid of the software of some standardization to correct the information in the event of some empty data cell errors. Again, the data is analysed in order to study various kinds and samples of information below the EDA techniques. In order to check all the information in the data of Airbnb, we have raised some questions and tried to find some practical business insights. And then at the last stage, we have visualized our data with the different visualization techniques using the python libraries. In the last, we have come up with the achieved business insights and conclusion of our project.

II. Problem Statement

The objective of the project is to perform an exploratory data analysis, data pre-processing, data cleaning & imputation and at the end, apply different Data Visualization techniques to get the meaningful insight from the given data. This project aims to apply some amazing Python Libraries such as Folium and Word Cloud which will give a boost to our visual understanding of the data. These thousands and lots of records generate a lot of statistics/data and this data can be analysed, used for security, grasp of customers/providers behaviour, business preferences & performance on the platform, implementation of modern additional service and for a lot of things. We are analysing the various aspects with different use cases which covers many aspects of Airbnb.

III. METHODOLOGY

The proposed methodology's implementation begins with downloading the dataset. Then data cleaning and normalization is executed as a step of pre-processing of data. After cleansing, the data is analysed. And then after analysing the data, data visualization is done. The file's visualization helps to have a show up at the traits and the relationship of attributes in the dataset. At last, all the business insights carried out in this project.

A. Datasets

The on hand dataset is used in analysis. It has information about New York City. In this project, the dataset of NYC AirBnb Booking Analysis is used. With 48895 instances and 16 properties, it is a multivariate dataset. The dataset consists of every expertise that is helpful and attributes that are no longer useful. So, in pre-processing the beneficial statistics is chosen and statistics cleansing is performed to get rid of the null values. Let's understands our variables:

- id - unique listing id
- Name- Represents Accommodations
- Host id - Unique id for hosts
- Host Name - Registered name for hosts
- Neighbourhood Group - Group of area/Locations
- Neighbourhood - Area under neighbourhood group
- Latitude - location of listing
- Longitude - location of listing
- Room Type - unique types of each room\Price - price of properties in dollar
- Minimum Nights - minimum night stay required for single visit
- Number Of Reviews - total rating
- Last review - latest review given
- Reviews Per Months - ratings received per month
- Calculated Host Listings Count - total number of properties registered under hosts
- Avaliability_365 - number of days for which host is available in a year for bookings

B. Pre-Processing

As at this preprocessing level, this is the necessary step; significant data is derived from the dataset of NYC Airbnb Booking Analysis. This section is compulsory because the raw data is now not reliable and unfinished, so pre-processing is carried out for greater steps to render geared up raw data. In this technique at some stage in pre-processing, 16 attributes are used to apprehend the area of New York City. These 16 attributes encompass host id, host name, listings name, room types, neighbourhood group, availability 365, number of reviews and many more. The attribute's values are normalized and descriptive statistical analysis is done.

C. Data Cleaning & Visualization

The quality of data performs a fundamental role, and the most cautiously depicted trouble to be. For this research, data cleansing has extended the quality of our dataset. Data cleansing is essential as it gets rid of inappropriate attributes of data from the dataset. This step of the model will make the dataset elevated particular and exact. In this

part, the null (NaN) values are eliminated from the dataset to make it greater and some null values are handled by filling it with its appropriate value. Next, we have visualized our facts graphically (as seen in the fig.1 below) as the dataset is in tabular shape and it is hard to appear at and understand the data in this or any other way. Data visualization helps in grasping the style of the data. Data visualization in this method is a graphical illustration of the data. In this analysis, the utilization of bar charts, box plot, dist. plots, scatter plots, point plots, pie charts, histograms and sub plots, the cleaned records obtained thru pre-processing is visualized. It makes it easy to maintain the attribute's tough relationship with the beneficial resource of graphical representation.

As mentioned above, this visualization performs an essential attribute in data exploration. A variety of parameters of the dataset plotted based on the available attributes as viewed in figure below.

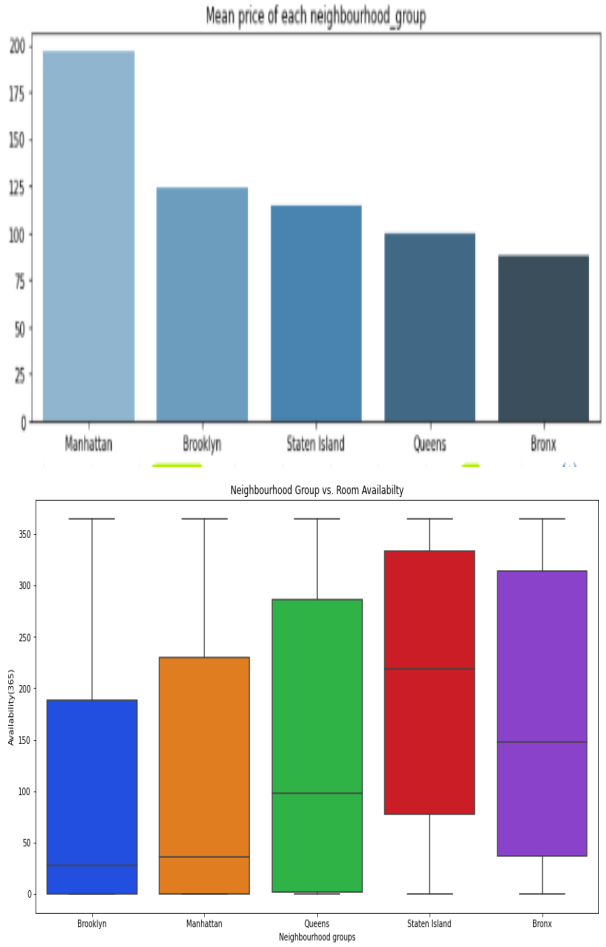

fig.1 EDA based on attributes

D. Exploratory Data Analysis (EDA)

After loading the dataset, we look at data shape, there are 48895 rows and 16 columns. There are 10 numerical variables and 6 categorical variables. In EDA, we have analysed different questions and tried to find the necessary solutions. Some questions we have analysed in our EDA are as follows:

1. Which room type and locality are mostly in demand in NYC?
2. Top 10 Host with most properties/listing on Airbnb website?
3. Top 10 Host with maximum number of reviews on the basis of reviews per month?
4. Top 10 most reviewed properties or listings on Airbnb?
5. Top 3 Most Expensive Listings on Airbnb?
6. Which Property type and room type is cheapest and more available on Airbnb?
7. What is the average price/day of Airbnb listings with respect to neighbourhoods in NYC? (Top neighbourhoods in NYC)
8. Properties with less number of reviews.
9. Any Particular Neighbourhood or Location with Maximum no. of Bookings and revenue from room type?

IV. COMPETITIVE ADVANTAGES & FUTURE WORK

The main advantages of Airbnb are as follows:

- Ease of Use – Search by price, locations and check-in/check-out dates
- Home Incentive – They can make money by sitting at home
- List Once – Hosts post one time only
- First to Market – For transaction-based temporary housing site
- Profiles and Bookings – Browse profiles of the hosts and book in just few steps

And some of the future work are as following:

- Scrape additional fields such as amenities and compare which amenities in both kinds of businesses, comments which can help people to understand the proper insights of that place.

- The major problem faced by guests or property owners using Airbnb is the trust factor. Giving your space to any stranger as a host and staying with strangers at their place as a guest might not be easy. So one can work on it also.

V. CONCLUSION

In the end, the useful technology used by us is python and its library. In this analysis project, about 10 different use cases were analysed on the given dataset to make better business decisions and help analyse customer trends and satisfaction, which can lead to new and better products and services. It has been found that Most of the Bookings took place for the "Williamsburg" of around "27%" followed by "Bedford-Stuyvesant", "Harlem" which has "25%" & "18%" respectively. Additionally, we also find-out the Top Earners (Host), relationship between neighbourhood group and Prices, Price comparison in terms of Room Type, Preference of Guests with respect to Room Type. Furthermore, we have also analysed Maximum Number of Bookings, Customer Reviews and many more.