EXPLORATORY DATA ANALYSIS ON AIRBNB BOOKINGS

By

Shahfaissal I Dharwad,Shreyash movale,Neha gupta, Ajinkya jumde, Eshaan sosa,

ABSTRACT

This study examined the relationship between various parameters of the AIRBNB dataset such as host id,host name , neighbourhood group, neighbourhood,room type,price number of reviews,availbility. An exploratory data analysis using field data points collected from theAirbnb listings in the metropolitan area of New York city reveals intriguing findings.The analysis helps us in understanding most preferred hosts and neighbourhood groups by guests, density of properties across various neighbourhood, number of room types belonging to each neighbourhood groups, expensive neighbourhood groups, busiest host's, preference of room types by guests, price of various room types.

This analysis helps draw insights from the data and can be utilised for security, business decisions, understanding of customers and providers, behaviour and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

Keywords: *Airbnb,Price,Neighbourhood group,Hosts,Room type,Number of reviews,Apartment,Reviews per month*

1. INTRODUCTION

Airbnb is an online marketplace that connects people looking to rent homes and people looking for accommodation in certain areas. Airbnb, Inc. is an American company with an internet market for sleeping, primarily rented accommodation, and tourism. jobs. A Airbnb does not have any properties. It provides a platform where people can rent their own spaces or spare rooms for guests. Prices are set by property owners and revenue is collected through the Airbnb app. The idea for Airbnb is simple: Find a way for local people to make more money by renting their backup houses or a place for tourists. There are many different types of Airbnb. You can rent a room at someone's house or on the whole island and everything in between.

Airbnb was founded in august 2008 by Joe Gebbia,Brian Chesky and Nathan with their headquarters in San Francisco. The idea came to Chesky again. Joe Gebbia in 2007 when they could not afford to rent their apartment. They transformed their living room into a bedroom to 'share' their home with three guests and provide them with a homemade breakfast. This was the beginning of the Airbnb concept. Travelers who use this platform get advertising for their employment to millions of people around the world, with the assurance that a large company will handle payments and provide support where needed.

There are few attributes of the AIRBNB bookings given below:

a. ID**:** There is a unique ID number for every entry in the dataset, with this unique ID information from the data can be easily extracted and identified

b. NAME**:** Every neighbourhood group has different hotels or renting rooms owned by host which is termed as name in the data frame.

c. HOST ID**:**

Same Hosts may have properties in different neighbourhood groups so a unique ID for the host is given as Host ID.

d. HOST NAME**:**

Hosts who have listed their properties on Airbnb has name as which is termed as Host name in the data frame

e. NEIGHBOURHOOD GROUP**:**

Name of groups of different hosts who have listed their property on Airbnb is termed as neighbourhood group.

f. NEIGHBOURHOOD**:**

Different localities of New York City are known as neighbourhood in the dataframe.

g. LATITUDE & LONGITUDE:

Latitude and longitude can be utilized to identify specific locations, which can also be helpful in identifying landmarks.

h. ROOM TYPE:

Different types of rooms are available which are categorized as private room, entire home/apartment, shared room

i. PRICE**:**

Every property listed on the Airbnb has rental price for owing over a period of time.

j. MINIMUM NIGHTS**:**

This data gives us the information about the time period of stay by guests in hotel or renting houses

k. NUMBER OF REVIEWS**:**

Contains information of the Count of reviews given by particular guest staying at rooms

l. REVIEWS PER MONTH**:**

Count of reviews per month by every guest is stored in this column.

m. CALCULATED HOST LISTING COUNT:

Every hosts owns different properties across different neighbourhood groups and count of this properties of every host is listed .

n. AVAILIBILTY 365:

This data helps us in knowing Number of days the hotel or renting place is available in financial year.

2. PROBLEM STATEMENT

Airbnb an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals and tourism activities based in New York city. Since 2008, tourists and hosts have used Airbnb to expand their travel opportunities and introduce a unique, personal way of feeling the world. Today, Airbnb has become one of the most widely used and recognized worldwide service. Data analysis in the millions of listings provided by Airbnb is an important aspect of the company.

This database has approximately 49,000 views in it and 16 columns and is a mixture between paragraph and numerical values.

This dataset has few problems in it such as

a.) What can we learn from the various tourists and places?

b.) What can we learn from the prophecies? (e.g., locations, prices, reviews, etc.)

c.) Which host places are busy and why?

d.) Is there a noticeable difference in traffic between different areas and could be the reason for that?

e.) How do prices of listings vary by location?

f.) How does the demand for Airbnb rentals fluctuate across the year and over years?

g.) Are the demand and prices of the rentals correlated?

h.) What are the different types of properties in NYC? Do they vary by neighborhood?

i.) What localities in NYC are rated highly by guests?

j.) Do regular hosts and super hosts have different cancellation and booking policies.

3. STEPS INVOLVED

a. Python Library:

NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Pandas: pandas is an open-source library that provides high-performance data manipulation in Python.

Matplot: Matplotlib is a python library used to create 2D graphs and plots.

Seaborn: Seaborn is a library for making statistical graphics in Python.

Word Cloud: Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.

b. To Import Dataframe:

Data frame has been imported from google drive and read data frame applying read_csv.

c. Recognize The DataFrame:

Dealing with a huge data set is a time-consuming part. To minimize the workload and efforts we must have to distribute data and analyze the contents first.

d. Dealing With NullValues:

As we have seen, our data frame contains a large number of null values so we need to deal with null values at the beginning of our project to discard null values from the data frame to improve our accuracy.firsts we have calculated the null value in each column such as **name,hostname,last review and reviews per month** have **16 21, 10052, 10052**" respectively.

e. Deal with Data:

Univariate Analysis: - ***Univariate Analysis*** is the key to understanding each and every variable in the data. Learn how to visualize and interpret ***univariate*** data.

Multivariate Analysis: ***multivariate*** data is to make a matrix scatterplot, showing each pair of variables plotted against each other.

f. Function & Method applied for Data frame:

Group By function: ***groupby***() function is used to split the data into ***groups*** based on some criteria.
Statistical method: To find some statistical summary like mean, max, min,count, standard deviation etc
Using statistical data, we have represented the various types of graphs.

g. Creating heat map and find co-relation between different columns with each other:

We have created a heat map between columns to find the co-relationship between all the columns with the help of correlation of statistical method

h. Performing Analysis:

for finding out the most availability_365, top neighborhood, Booking in city, Host in city, Different room types, Top host, Average Nights in room, Price distribution

finally,from all the results after performing exploratory data analysis meaningful conclusions were drawn which is include at the end of the document

4. Data Analysis:

Eda is performed with the data frame on various variables which are dependent on each other and visualization of the result is done using various plots such as scatter plot, boxplot, bar plot, violin plot, histogram, heatmap, wordcloud, line chart, few of the important analysis is shown below.

a. **Density of neighborhood across the different locations:** Latitude and longitude data is used to known the density of neighbourhood groups across the location. The data is visualized with the help of scatter plot. Latitude and longitude forms a grid system which helps to identify the exact or absolute, locations on the surface of earth. Latitude and longitude can be utilized to identify specific locations, which can also be helpful in identifying landmarks.

b. Room type within different neighbourhoods :

Dataset of different roomtypes and neighbourhood groups are utilized for visualization which is done by grouping the data. Barplot is taken into account for visualization.

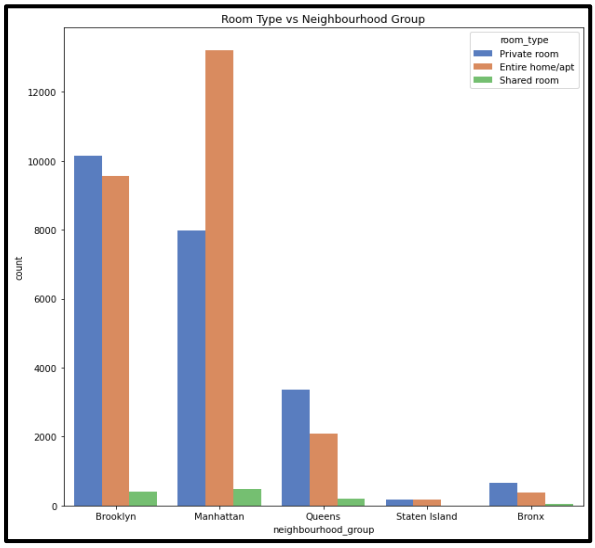

Figure 1.1 Room types within different neighborhoods

The above results showed that customers are more interested in booking Entire home/apt followed by private rooms.

c. **Price over neighbourhood groups**

Dataset of different roomtypes and neighbourhood groups are utilized for visualization which is done by grouping the data. Barplot is taken into account for visualization

It shows that Manhattan is quite expensive neighbourhood group compared to others.

d. Unique price category counts:

Price is classified into 3 groups i.e. price below 80 is classified as cheap, price between 80 and 150 is classified as affordable and price above 150 is classified as expensive, groupby function on neighbourhood group along with price category is applied to get the count of neighbourhood groups based on price category .bar plot is taken into account for visualization .It is observed that the least people prefer the expensive category, instead maximum people prefer affordable category followed by cheap category in all the neighbourhood groups except in case of Bronx and Queens where the relationship is reverse.

5. CONCLUSION:

We have studies here all the different aspects of data analysis which will be beneficiary for the business development of the company. We have checked the dependency of some key parameters of company with others like pricing and reviews analysis. We get some information about hosts and neighbourhoods,etc. Although there will always be some queries which we have to resolve apart from the above analysis. This analysis will help company to take decisions like which areas they are leading and lagging, where they need improviisation and what necessary changes they need to take for better functionality of their buisness.