EXPLORATORY DATA ANALYSIS ON AIRBNB BOOKINGS

By

Shreyash Movale, Neha Gupta, Ajinkya Jumde, Eshaan Sosa, Shahfaissal I Dharwad

ABSTRACT

This study examined the relationship between various parameters of the AIRBNB dataset such as host id, hostname, neighbourhood group, neighbourhood, room type, price number of reviews, availability. An exploratory data analysis using field data points collected from the Airbnb listings in the metropolitan area of New York city reveals intriguing findings. The analysis helps us in understanding the most preferred hosts and neighbourhood groups by guests, the density of properties across the various neighbourhood, the number of room types belonging to each neighbourhood group, expensive neighbourhood groups, busiest hosts, 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, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via the website and mobile app. Source; Wikipedia

Airbnb mostly has a business in properties on rentals. Airbnb does not own any property instead works as an intermediator between property owners and customers looking for properties for rent.

This booking of room process can be done from the website or app of Airbnb. Airbnb receives bookings from all over the world so they have to work on building a relationship with the property owners which they wish to put on rent for the visitors.

What we have done is that we had analysed some major areas based on the given dataset for NYC

To help Airbnb with the decision to an expansion of business in particular areas.

There are a 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 the host which is termed as a 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. HOSTNAME**:**

Hosts who have listed their properties on Airbnb have a name which is termed the Hostname in the data frame

e. NEIGHBOURHOOD GROUP**:**

The name of groups of different hosts who have listed their property on Airbnb is termed a neighbourhood group.

f. NEIGHBOURHOOD**:**

Different localities of New York City are known as a neighbourhood in the data frame.

g. LATITUDE & LONGITUDE:

Latitude and longitude can be utilized to identify specific locations, which can also help identify landmarks.

h. ROOM TYPE:

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

i. PRICE**:**

Every property listed on Airbnb has a rental price for owing over some time.

j. MINIMUM NIGHTS**:**

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

k. NUMBER OF REVIEWS**:**

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

l. REVIEWS PER MONTH**:**

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

m. CALCULATED HOST LISTING COUNT:

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

n. AVAILABILITY 365:

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

2. PROBLEM STATEMENT

Airbnb is 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 about the world. Today, Airbnb has become one of the most widely used and recognized worldwide services. Data analysis of 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 a paragraph and numerical values.

This dataset has a 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 neighbourhood?
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: panda is an open-source library that provides high-performance data manipulation in Python.

Mat plot: 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 Data frame:

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

c. Recognize the Data Frame:

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 Null Values:

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 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 finding 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 neighbourhood, Booking in the city, Host in the city, Different room types, Top host, Average Nights in the room, Price distribution

finally, from all the results after performing exploratory data analysis meaningful conclusions were drawn which are included 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, word cloud, line chart, few of the important analysis is shown below.

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

b. Room type within different neighbourhoods:

Datasets of different room types and neighbourhood groups are utilized for visualization which is done by grouping the data. Bar plot is taken into account for visualization.

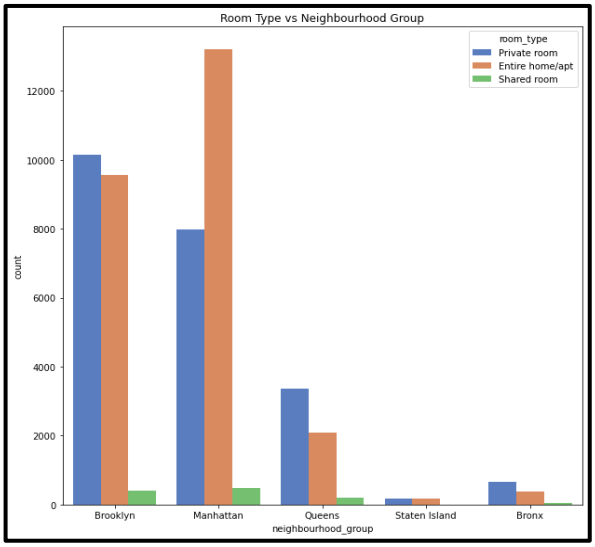

Figure 1.1 Room types within different neighborhoods

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

c. **Price over neighbourhood groups**
Datasets of different room types and neighbourhood groups are utilized for visualization which is done by grouping

the data. Bar plot is taken into account for visualization

It shows that Manhattan is quite an 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, group by 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, the maximum people prefer the affordable category followed by the cheap category in all the neighbourhood groups except in the case of Bronx and Queens where the relationship is reversed.

5. CONCLUSION:

That's all from our side for now. Although we know that we have not analyzed many of the aspects which might have been missed in our analysis.

We have done room type distribution analysis throughout the map and concluded that there are more demands for private rooms and flats/apartments. So, the company must focus on merging more property owners with these types of rooms.

We have studied that there is a huge price gap between Manhattan and Bronx as well as system island neighbourhood groups making Manhattan an expensive area for living. So the company can provide various offers to visitors taking into consideration, the visitor's demand in these areas.

We have also analyzed prices and reviews showing the major contribution of customers for the affordable category.

Based on the Collab study we concluded that the Manhattan and Bronx can be major business centers with Airbnb with private rooms and the price category must be medium/affordable.

So, the company must show interest in acquiring more business with such types of property owners.

References:

Wikipedia

https://www.geeksforgeeks.org/

https://stackoverflow.com/

https://towardsdatascience.com/exploratory-data-analysis