Exploratory Data Analysis

CAPSTONE PROJECT

Airbnb Booking Analysis

Prepared by:

PRAVEEN SIVAKUMAR

Data Science Trainees,
Almabetter

Abstract:

Today's Airbnb services provide people an easy,relatively stress-free way to earn some income for the hosts from their property as well as for the customers such as home benefits,personalized services,authenticity and social connection around the world.Thus it has curated a large dataset having both dependent and Independent variables.

This dataset describes the listing activity and metrics in NYC,NY for 2019. Our Exploratory Data Analysis can help understand what could be the reason for classification of prices and distribution of which room type in various neighbourhood groups and their respective neighbourhood.

1. Problem Statement:

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.The dataset contains the following columns for our interpretation,

- **Id:** Id for each airbnb
- **Name:** name of each airbnb
- **Host_id:** Id for each host
- **Host_name:** name of the hosts who owns the airbnb
- **Neighbourhood_group**:States in which the neighborhoods belongs
- **Neighbourhood:**Cities in which airbnb's are distributed
- **Latitude:** represent the latitude of each airbnb
- **Longitude**: represent the longitude of each airbnb
- **Room_type:** Indicates the room type
- **Price:** Cost per day in Dollars
- **Minimum_nights:** how many days has it been used constantly
- **Number_of_reviews:** Indicates the count of reviews received
- **Last_review**:last updated review date

- **Reviews_per_month:** monthly average review count
- **Calculated_host_listing_counts:** host listing count
- **Availability_365:** availability of rooms in a calendar year

2. Introduction:

We are a team of three had present our exploratory data analysis,visualization,Interactive plots and lot of other interesting insights into the airbnb data.We focus on New york City's data that it would give us a great demonstration to the exploratory data analysis techniques and also wish to perform an in-depth analysis on one of the most densely populated cities in the world.

3.Objective:

The main objective is to do a exploratory data analysis which could help to come to conclusion for the few questions that we aim to answer through our analysis.

The following are the few questions that we aim to answer:

1. What can we learn about different hosts and areas?
2. What can we learn from predictions? (ex: locations, prices, reviews, etc)
3. Which hosts are the busiest and why?
4. Is there any noticeable difference of traffic among different areas and what could be the reason for it?
5. Top neighbourhoods in NYC with respect to average price /day of airbnb listings?
6. Room types vs price on different neighbourhood groups?
7. On an average for how many nights people stayed in each room types?
8. How monthly reviews varies with room types in each neighbourhood groups?
9. Top10 reviewed hosts on the basis of reviews/months?
10. Room types and their relation with availability in different neighbourhood groups?

Imported Modules :

Numpy:

adding support for large multi-dimensional mathematical functions

Pandas:

Data analysis and manipulation tool for the dataset

Matplotlib:

For creating static,animated and interactive visualization representations.

Seaborn:

High level interface for informative statistical graphics.

Loading the dataset:

Mounted using the google drive to load the dataset in the form of CSV(comma separated values)

Dataset Attributes:

The dataset exactly contains 48895 rows and 16 Columns with around 20000 missing values which should be cleaned before processing. There are 5 **Categorical columns** namely name,host_name,neighbourhood_group,neighbourhood and room_type And 11 **numerical columns** namely Id,host_id,latitude,longitude,price,minimum_nights,number_of_reviews, last_review,reviews_per_month,calculated_host_listing_count and availability_365

Data cleaning:

Dropped columns:

We have dropped the last_review column as there is no derivable information with that.

Handling null values:

Since the columns last_review and reviews_per_month has most of the null values,we have to either drop or replace the values.As the last_review column has been dropped,the null values in the review_per_month denotes that the respective airbnb have not received the reviews, thus the null values can be replace with 0.

And also some of the airbnb's has price 0.Since the price cannot be equal to zero,so replacing the 0 priced airbnbs with mean of the priced column.

Challenges Faced:

1. In the reviews_per _month column,20% of the data are null values which made us to difficult to understand whether the particular airbnb didn't receive any reviews or the values just went missing.

- We progressed the analysis by providing 0 to all the null values which is a huge risk.since it would affect the impressions of other columns.

2. The price column has a lot of outliers which doesn't make sense as it ranges from 0 to 10000 USD.

- We tackled the 0 price by taking mean on the room_type and neighbourhood.

Exploratory Data Analysis:

Univariate Analysis:

It involves with analysis of a single feature or variable.Such analysis that we have undergone are,

1. Top 10 host IDs that owns the maximum number of airbnbs in the New york City.

It has been found that:

- Hosts id's from Manhattan owns most number of airbnbs in the New york city.

2. Distribution of the airbnbs in the neighbourhood groups

It has been found that:

- Queens, Staten Island, and the Bronx have main distribution below its mean price around below 60.
- In comparison to another neighborhood group in Brooklyn and Manhattan, we can see a greater distribution above price 100 that concludes there are costly hotels in urban area.

3. Distribution of Airbnbs on the basis of room_types

It has been found that using a count plot

- Shared room are lot less in comparison to other room type.
- As count of entire home and single room are more,it concludes that more people book them.

4. Price distribution of the airbnbs across the new York City

It has been found that:

- It can be observed that 75% of the Airbnbs has a price below 200.
- To accommodate maximum data in the histogram and to include maximum number of Airbnbs in the graph, we would be plotting a graph having Airbnb prices equal to and lesser than 200.

5. Top 20 neighbourhoods having maximum number of airbnbs

It has been found that:

- Williamsburg has most number of airbnbs in all of New york city followed by Bedford-Stuyvesant in which both belongs to Brooklyn neighbourhood.

Multivariate Analysis:

It involves the analysis of multiple variables.The multivariate analysis we have undergone are,

1. Comparison of neighbourhood_group in terms of price.

It has been found that:

- Manhattan being the most costliest place to live in, have price more than 140 USD followed by Brooklyn with around 80 USD on an average for the listings.
- Queens, Staten Island are on the same page with price on listings

2. Comparison of neighbourhood_group in terms of room_type

It has been found that:

- Manhattan has more listed properties with Entire home/apt around 27% of total listed properties followed by Brooklyn with around 19.6%.
- Private rooms are more in Brooklyn as in 20.7% of the total listed properties followed by Manhattan with 16.3% of them. While 6.9% of private rooms are from Queens

3.Average availability of Airbnbs on the basis of neighbourhood groups and room types

It has been found that:

- Both Private room and Entire home/apt from Staten Island has the most average availability.
- On the other hand, the shared rooms from Brooklyn has found that most available when compared to other neighbourhood groups.

3. Comparison of reviews in terms of room_type in each neighbourhood_group

It has been found that:

- People tend to review more in Staten Island of Entire home and single room.
- People of urban towns are less to leave a review for Airbnb.
- Shared room of small town has low percentage of review.

4. Average number of nights spent in the Airbnbs on the basis of neighbourhood groups and room types

It has been found that:

- The average number of nights spent is high in Entire room/apt in all of the neighbourhood groups.

5. Top 20 neighbourhoods who received the maximum number of reviews

It has been found that:

- Silver lake from Manhattan neighbourhood has the maximum number of reviews followed by East elmhurst.

6.Top 20 neighbourhoods on the basis of mean price

It has been found that:

- Fort wadsworth from Manhattan has the most mean price followed by woodrow from Staten Island.
- Most of the neighbourhoods in the top 20 belongs to Manhattan Neighbourhood

7.Top 10 host_id on basis of mean review

It has been found that:

- The Hosts Row NYC has the most reviews per month with 43 reviews per month followed by Louann with 21.

8.Correlation of every features.

It has been found that:

- We have used the Pierson correlation
- Review per month is very strongly correlated with number of review.

Conclusion:

We have reached the end of our project!

Starting with loading the data so far we have done the EDA by importing the necessary modules,going through the dataset attributes,data cleaning,univariate and multivariate analysis with their respective visualization.

In perspective of all the visualization we have able to answer some critical questions.

This would in turn help the customers in matching the right airbnb's with the right customers efficiently as well as quickly.

Reference:

1. Geek for geeks
2. W3 schools
3. Wikipedia