Airbnb NYC EDA

Piyush Lanjewar
Pruthvi Raj
Yogesh Reddy

Abstract:

Airbnb is the largest platform to find and rent homestays and apartments for vacations and holidays. We are provided with data from the year 2008, to perform EDA. EDA helps us to identify important features for the better improvement of the service.
It helps us to understand the data in different aspects.
Keywords: EDA, plotly, plotly express, folium, plotting.

1. Problem Statement

The primary factor of AIRBNB is to provide Homestays for Vacation rentals and tourism activities and none of the listed properties are owned by AIRBNB. It is just an online marketplace for all the listed properties.

2. Introduction

Airbnb is one of the largest used companies for lodging primarily homestays for vacation rentals and tourism activities. Today Airbnb is one of the most used brands to give hosts and guests a good experience. The data is used to increase the understanding of every detail to make traveling easy and convenient. The data is utilized to show the required conclusions. The conclusions are also shown in the visualizations to make understanding easier.

The factor that affects this business is the Reviews. If the visited customers provide a positive review, then there is a high chance that it gets booked several times based on the availability. In case of negative reviews, customers most likely don't prefer to stay there. By observing the provided data around 25% of the listed properties either be Home-apt. Or private rooms or shared rooms don't have the reviews this doesn't help to maintain a healthy relationship with the hosts of the listed properties.

There is small inaccuracy in the dataset on price where around 10% of the properties are given with the price 0$ which doesn't relate at all. So to represent real data instead of using the mean to fill the faulty prices we applied Median.

3. Types of Houses and Different Neighbourhood Groups

There are three different types of Houses listed in Airbnb.

- Entire room/Apt
- Shared Room
- Private Room

There are five neighborhood groups.

- Brooklyn

- Manhattan
- Queens
- Staten Island
- Bronx

4. About the Dataset

The dataset has approximately 49000 rows and 16 columns.

- host_name: The Name of hosts who give services to guests
- Neighbourhood_group: Represents the city
- Neighborhood: Represents areas of the city
- Latitude and longitude: Represents the location of the house
- room_type: Represents the type of room(shared/private/apt)
- price: Represents price of the houses
- minimum_nights: Nights spent by customers
- number_of_reviews: Number of reviews
- last_review: Date represents the last review by customers
- reviews_per_month: Reviews per month
- calculated_host_listings_count: Host count listing.
- availability_365: The availability of hosts per year.

5. Observations

- **Price Variation:**

The prices are different for each house type, on average the prices of an Entire home/apt are more expensive and a shared room is less expensive.
When we talk about the neighborhood-wise prices Manhattan has expensive houses of all types.

- **Distribution of Houses across the map**

There are around 49000 houses listed. In which Manhattan has more listing of houses of 21661 and Staten Island has less listing of houses of 373.

- **Availability of Hosts**

The dataset availability_365 column tells about the host availability for 365 days, if it is 0 then the host is busy.

6. Steps involved:

- **Exploratory Data Analysis**
  After loading the dataset we have performed EDA using various plotting techniques. We have visually plotted several graphs on features of the dataset. We found several insights from the dataset.
- **Null values Treatment**
  Our dataset contains a large number of null values which might affect our analysis. For the reviews column, we have replaced null values with 0's. For the price column, we have replaced null values with a median. Unnecessary columns for the analysis were removed.
- **Visualizing data on a Leaflet map**
  We were provided with the latitude and longitude data of each house.

We have used folium and plotly libraries to visually represent the data on a map. We visually represented neighborhood groups and each house using scatter plots and cluster markers.

7. Conclusion:

- The data of AIRBNB since 2008 has been provided in this dataset. This dataset has around 49000 observations with 16 columns and it is a mix of numerical and categorical values.
- The detailed Insights from the analysis:
    - The Neighbourhood of the Newyork city has 5 groups:
        - Brooklyn
        - Manhattan
        - Queens
        - Staten Island
        - The Bronx
    - Properties from Manhattan are a bit pricey followed by Brooklyn and Staten Island
    - Top 3 Hosts from the dataset are:
        - Sonder(NYC)
        - Blueground
        - Kazuya
    - Since, Reviews are important here are the Top 3 Hosts who hold the most reviews:
        - Dona
        - Asa
        - Dennis & Nauko
    - There are 3 room types Entire home/apt. (~25000), Private room(~23000), Shared room(>500)
    - The visualization that plotted on maps also shows where all the properties are located along with their price, availability, and the type of the room

References-

1. Medium
2. GeeksforGeeks
3. Analytics Vidhya