Exploratory Data Analysis on Airbnb

By: Ajay Negi & Mantresh Kumar

Cohort Rimo, Almabetter

Abstract:

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.

EDA of Airbnb will help us to investigate the data to discover the patterns, spot anomalies, test hypotheses, and check the assumptions with the help of the summary statistics and graphical representations.

Keywords: availability, neighbourhood, neighbourhood groups, boroughs

1. Problem Statement

The customers need a platform that allows them to navigate and negotiate details involved with renting and booking event spaces easily. Costumers need to be able to sort through venue options for the ideal space while letting the hosts monetize their property and feel comfortable renting their property to new clients. As price and availability is an important concern for customers booking travel online, hotels leave you disconnected from the city and its culture. It is the best way that exists to book a room with a locale or become a host.

2. Introduction

Airbnb dataset has information about hosts, costumer's reviews, availability of rooms, neighbourhood groups, neighbourhoods, prices of room types and many more. So, with Exploratory Data Analysis we are analysing the patterns, changes, fluctuations in the prices and availability, the behaviour of hosts, expenditure and priority and relevancy of the people, and many things can be taken out as a conclusion from this dataset which can be used for the future development of the company.

Our goal here is to provide future-oriented conclusions for Airbnb. So that they can take decisions based on that conclusions and earn fruitful results.

3. Challenges Faced

- Reading the dataset and understanding the meaning of some columns.
- For answering some of the questions we had to understand the business model of Airbnb and how they word.
- Handling NaN values, null values, and duplicates.
- Fixing the invalid values
- Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.

4. Neighbourhood Groups in Dataset

The sub-division of New York City is known as **“Neighbourhood Groups”**. There are total 5 Neighbourhood Groups:-

- Manhattan
- Brooklyn
- The Bronx
- Queens
- Staten Island

5. Types of Rooms in Dataset

Among these 5 neighbourhood groups, there are 3 types of rooms:-

- Entire Home/Apartment: High Price and High Demand
- Private Rooms: Mid Price and Mid Demand
- Shared Rooms: Low Price and Low Demand

6. Libraries Used in EDA

- **Pandas:** Pandas is one of the most popular libraries of Python that helps to present the data in a way that is suitable for analysis via its Series and DataFrame data structures. It provides various functions and methods to both simplify as well as expedite the data analysis process.

- **NumPy:** NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

- **Matplotlib:** Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits.

- **Seaborn:** Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with Pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on DataFrames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

- **Folium:** Folium is a Python library used for visualizing geospatial data. It is easy to use and yet a powerful library. Folium is a Python wrapper for Leaflet.js which is a leading open-source JavaScript library for plotting interactive maps.

7. EDA Procedure

- Importing Libraries
- Basic Data Analysing
    - Segregating Numerical and Categorical Columns
- Data Wrangling
    - Removing Null Values
    - String to Date Conversion
    - Invalid Data Fixing
    - Duplicate Removal
    - Outliers Handling
- Plotting Graphs by using Matplotlib and Seaborn
- Observation of graphs and plots
- Taking out Conclusions

8. Conclusion and Final Outcomes

- Prices are higher in Manhattan as people have a higher opportunity for growth and availability of facilities on their doorsteps.

- Shared rooms were not introduced till the year 2013, in 2014 the shared rooms were introduced.

- Most people prefer Entire Rooms/Apartments and then Private Rooms because privacy and personal space are more important to people.

- Airbnb got the most number of reviews in 2019 because the people stayed for most no. of nights in the year of 2019.

- The per capita income of people in New York is enough that they can afford private rooms and entire home apt.