Exploratory Data Analysis of Airbnb NYC Dataset

Pushkar Srivastava,Rahul Pandey
Data science trainees,
AlmaBetter, Bangalore.

Abstract:

This study aims to understand the customer's and host's behaviour and performance on the platform of Airbnb through an exploratory study that accounts for around 49,000 observations in New York, the United States of America, in neighbourhoods with a high number of accommodations listed on the Airbnb platform.

Our analysis can help in the understanding of different hosts and areas, which hosts are the busiest, which type of property listing is most preferred by customers, and which neighbourhoods have the highest traffic.

Keywords: Exploratory Data Analysis, Airbnb NYC dataset.

Problem Statement:

Since 2008, guests and hosts have used Airbnb to expand on travelling 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) behaviour 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 of categorical and numeric values.

Explore and analyze the data to discover key understandings (not limited to these) such as :

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

Introduction:

Exploratory data analysis (EDA) is an approach using descriptive statistics and graphical tools to better understand data. It is used mainly to maximize insight into a

dataset, detect outliers and anomalies, and test underlying assumptions. It is a robust first step before the application of other statistical methods.
The NYC Airbnb dataset contains 49000 observations with 16 columns which describes information regarding the Airbnb property listings, host, location, property type, price, minimum nights, number of reviews, and availability.
Our goal here is to perform an exploratory data analysis on the Airbnb NYC dataset, which could help in understanding the story the dataset entails.

Steps Involved:

- **Ask**
  In this phase of the analysis, we asked ourselves what is the main problem statement we are trying to answer.

- **Prepare**
  In this phase, we made sure that the dataset we were given was good enough to answer our queries or that some additional data is required for our analysis.

- **Process**
  In this phase, the dataset cleaning process was started.

  Duplicate values removal-Firstly we checked for any duplicated observation through spreadsheet's remove duplicated functionality which was not found in our dataset.

  Null values removal-Secondly, we checked for null values in the dataset and found that our dataset contained a large number of null values therefore for accurate results we dropped the columns containing nulls. In the review_per_month column, we assumed that NAN would imply zero reviews that month so we replaced it with zero.

  Data validation-Thirdly, we validated our cleaned data based on data type, range, constraint, structure, and consistency.

- **Analyze**
  This phase was all about making sense of data by identifying trends and relationships within the data. This involved four steps

  Organize data-Locate or access data for the problem

  Format and adjust data-Filtering and sorting data

  Get input from teammates-Getting valuable insights from teammates on possible trends.

  Transform data-Identifying trends and relationships in data.

- **Share**
  In this phase, we used visualizations to explain our findings mainly through matplotlib and seaborn libraries in python.

- **Act**
  Using our solutions from above to solve the problem statement.

Observation

We dealt with missing data and outliers. That's a lot of work that Python helped us make easier. This analysis gave us great insight into which neighbourhood group has the highest and lowest average listing which can be correlated with traffic density in respective neighbourhood groups. Also, we learned which type of listing among the three(Apt./entire home, private room and shared room) are most prefered in each neighbourhood group.

Conclusion

That's it! We reached the end of our study. Throughout the analysis, our goal was to investigate each variable and uncover as many hidden facts about the data as possible. Though it is also true that we spent the most time on the price variable and its relationship with other variables since most of the important information regarding traffic density and customer preference is correlated with price.

References

GeeksforGeeks
Researchgate.net