Airbnb Booking Analysis

Sanjay Yadav
Data science trainee
AlmaBetter, Bangalore

Abstract:

Purpose - The purpose of this project is to review and analyze the bookings on Airbnb –one of the most significant innovations in the tourism sector.

Methodology/Approach - Exploratory Data Analysis of dataset containing around 49000 records and 16 attributes .

Findings - Finding relationship between Airbnb hosts , Location, Prices and other features provided in the dataset.

Practical implications - Better understanding of large dataset and how to analyze to get meaningful sense out of the provided data.

Keywords – EDA , Airbnb, NumPy, Pandas, Matplotlib, Seaborn

1. Problem Statement

Airbnb is an online platform through which individuals can rent out their spaces as tourist accommodation. These spaces typically entail either an “entire place”(house, condominium, etc. or a “private room " in a residence where the host is also present. Airbnb’s diverse inventory additionally includes some fairly exotic accommodations (castles, igloos, treehouses, etc. Airbnb listings range from quite modest to extremely luxurious.

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.

Explore and analyze the data to discover key understandings.

2. MAIN OBJECTIVES :

• Explore and clean the data for any missing values and duplicate data.

• Perform analysis on the price distribution in various major cities of California.

• Check if the price has a relation with the minimum stays as well as property type along with location.

• Perform geographical plotting of the price and location, and to see of if I can find any interesting aspects near the Silicon Valley and the tourist locations in LA.

• Analyze the behavior of hosts with multiple listings.

• Understand the relation between reviews count and neighborhood.

2. Introduction

In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA) which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Our goal here is to analyze the dataset provided and find correlation between various attributes through plotting using matplotlib and seaborn libraries.

3. Libraries Used

1 . NumPy

- NumPy is a Python library used for working with arrays.
- It also has functions for working in domain of linear algebra, Fourier transform, and matrices.
- NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.
- NumPy stands for Numerical Python.

2. Pandas

- Pandas is a Python library used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Why Use Pandas?

- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant.
- Relevant data is very important in data science.

3.Matplotlib

- Matplotlib is a low level graph plotting library in python that serves as a visualization utility.
- Matplotlib was created by John D. Hunter.
- Matplotlib is open source and we can use it freely.
- Matplotlib is mostly written in python, a few segments are written in C, Objective-C and JavaScript for Platform compatibility.

4.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.
- Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

5.Plotly

- Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python.

4. Steps involved:

❑ **Data Cleaning**

Data cleaning means fixing bad data in your data set.

Bad data could be:

- Null values treatment
- Data in wrong format
- Duplicate data
- Outliers

❑ **Data Categories**

To analyze data, we also need to know the types of data we are dealing with.

Data can be split into three main categories:

- Numerical - Contains numerical values. Type – int64 , float64
- Categorical - Contains values that cannot be measured up against each other. Type - object

By knowing the type of your data, we will be able to know what technique to use when analyzing them.

❑ **Finding Relationships**

A great aspect of the Pandas module is the corr() method.

The corr() method calculates the relationship between each column in our data set.

The similar can be achieved through seaborn correlation heatmap plot.

❑ **Analyze the Data**

When we have cleaned the data set, we can start analyzing the data using imported libraries.

- **Plotting**

We can find and establish relationships between various attributes such as neighborhood groups ,price, reviews etc. through various statistical graphs provided by matplotlib and seaborn.

5. Conclusion:

Starting with loading the data so far we have done data cleaning including null values treatment, data in wrong format, empty cells , analyzing and finding relationships between various attributes and plotting .

Through our analysis we find that there is great correlation between rent price and traffic since larger reviews are available for the places with low rent.

More listings are rented for shorter period of time .

Also people prefer to rent entire apartments rather than sharing which can be seen by the high traffic difference .

References-

1. GeeksforGeeks
2. W3Schools
3. Stack Overflow