Technical Document

EDA on Airnb Booking Analysis

Contribution- Individual

Abdul Rehman Ansari

Abstract:

Airbnb is an online marketplace connecting travelers with local hosts. The platform enables people to list their available space and earn extra income in the form of rent. Founded in San Francisco in 2008 as a start-up, the company has become a worldwide booking platform. Airbnb is present in over 190 countries across the world.

We are provided with dataset which has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Our experiment can help Customers in the sense, which hotels to book in different Neighborhoods by looking at our analysis on reviews of different Neighborhoods , what Room Type they should get according to their needs and it will also help them to decide at what price they should pay and many more. It will also help the Local Hosts to analysis the Customer behavior regarding their preference in different areas.

Introduction:

We are provided with dataset which has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values. We have many columns but we shortlisted some important columns such as “host_id , host_name , neighbourhood_group , neighbourhood , room_type , price , minimum_nights and number_of_reviews”. This will help us to analysis the guests and their preference different Neighbourhood , Prices of listings and lot more.

Our goal here is to present our exploratory data analysis, visualizations, interactive plots and lots of other interesting insights into the Airbnb data. Which could help Guests in choosing the perfect Place to stay according to their preference and provide information to different Hosts so that they can attract more customers.

Problem Statement:

Data on 49000 observation is provided by Airbnb. Our goal is to analysis the data so that this can be used for security, business decisions, understanding of customers' and hosts behaviour and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more. We have been provided with a data which lots of Null values. We saw that the price column has some irregularity as the minimum price of the apartment is 0 i.e. free stay. So we have deal with Null values and make some assumptions for the price column.

We have Hosts and Neighborhood group column, we have to do some analysis based on these column to get the data regarding the behavior of hosts. Then we have location, price, and reviews column we will analyze this column to get the data about price and reviews in different locations.

We have to analyze which hosts are the busiest and why with the help of data regarding their minimum stays. Lastly we focused on any noticeable difference of traffic among different areas and what could be the reason for it. With the help of Neighborhood group and minimum nights column.

Approach:

1. First we deal with null values, last_review and reviews_per_month both had 10052 null values so we dropped both these columns, we also dropped the rows which had nullvalues. Then we drop the unnecessary columns such as longitude and latitude. Then we saw that the price column has some irregularity as the minimum price of the apartment is 0 i.e. free stay so we took an assumption and replace all the ‘zeros’ with median price of the column.

2- In first analysis we analysed the Neighbourhood group column to find out the Neighbourhood group which has most no of listing properties.

3-We analysed hosts column with calculated_host_listing_count tofind out the Top hosts by listings.

4- We analysed price, room type, neighbourhood group, neighbourhood and reviews column to find out which neighborhood group has highest no of reviews. What preference of customer in respect to price and at what range of price most properties are listed. Room type share and which type of room guest usually prefer. Which are the most expensive Neighborhoods in our data set etc.

5- We tried to find Top 10 busiest host by counting their maximum reviews. Taking top 100 busiest as sample we tried to find their choices regarding room type by comparing busiest host with room type.

6- Lastly we analysed the noticeable difference of traffic among different areas and what could be the reason for it by comparing Neighborhood group and minimum night's column and looked into some other factors such as places of tourist attraction in the Neighborhood.

Conclusion:

That's it! We reached the end of our exercise. Starting with loading the data so far we have done EDA, null values treatment, took assumption for price and analyzed each columns.

Represented our observation with the help of different type of charts and visualization.

Our analysis Airbnb data can be used for business decisions, understanding of customers' and hosts behaviour and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.