Airbnb Booking Analysis

Names

Data Science Trainees

Alma Better

**Abstract:**

Our team has a deep knowledge about Airbnb Booking Analysis. We aim to focus only on New York city. The analysis is majorly done on borough country they are Brooklyn. Manhattan, Queen, Staten Island, Bronx. This data describes all host details, price of rooms, reviews etc. From analysis we get huge information with help of Exploratory Data Analysis, Data Wrangling, Visualization etc.

**1.Problem Statement:**

The Customer is able to book a room in Airbnb. Customer is moving to Queens, Bronx, Staten Island, Brooklyn and Manhattan neighbourhoods to explore it.

The customer is in need to identify a number of bookings in each borough country & guide that customer to choose a best place to live at affordable cost, analyse the price of the room so that customer can afford it, type of room according to their needs. All of these requirements will be satisfied by number of reviews, reviews per month and availability of the rooms. We will get all of these insights by performing EDA and understanding the data itself.

* **Id -**Id of Each row
* **Name -**Name of Each room given by host
* **Host\_id –**Id of each Host
* **Host\_name –**Name of each Host
* **Neighbourhood group –** Borough country Name
* **Neighbourhood –**All cities of Each Borough Country
* **Latitude and Longitude –**Geographical Data of Borough Country
* **Room type** –Name of each room type
* **Price –**price of each rooms
* **Minimum Nights-** price for Minimum nights to stay
* **Number of review-** Total reviews given by customer
* **Last reviews-** Date of last reviews given by customer
* **Reviews per month –**Average review rate per month
* **Calculated host listing count** –Count of Host listing in Airbnb
* **Availabilty\_365 –** Number of days availability

**Introduction:**

The Airbnb is a room bookings data. This data is given only for New York City from the year of 2011 to 2019.By exploring the data we are able to understand all the features. The objective is to explore the data and then analyse the data with all the required information. With the help of data exploration, data cleaning and analysis we have to discover key understandings from it. So that the understandings can be used for guiding marketing, understanding of customers’ and hosts’ behaviours.

**Steps Involved:**

**Exploratory Data Analysis:**

Once our data loaded using pandas library. To understand the data, we printed top 5 rows and performed various operation like casting the columns into their respective data types. Creating new column **price\_per\_night** for fast analysis. Understanding the relationship of all numerical and categorical columns. Also done the Bivariate Analysis with respect to numerical data and categorical data. We also plot a distribution of numerical column to verify whether it follows a Gaussian Distribution.

**Visual Analysis:**

Visualization tells thousands of meaningful data in to single plot. The visualization is done with help of seaborn, plotly, matplotlib etc. We performed various plot like bar chart, pie chart map chart, box plot, violin plot and waterfall chart etc. Data Analysis is performed before cleaning the data.

**Treating Null Values:**

Null values are present in the given data. These null values occur due to some sensitive information which user may not like to specify those data. It can be visualized using heat map plot.

**Imputation techniques for each column**

1. Removing **hostname** column datas from Dataframe
2. Replacing **name** column null values by **'Missing'** Labels
3. **Median Imputation**Techniques used for **Last review** column
4. **Mode Imputation** Techniques used for **reviews\_per\_month** column

**Performing Outliers Removal:**

Outlier is a data which differ from actual distribution of data. It also causes the Original distribution of data. The Outlier can be visualized using box plot and violin plot. The technique used here is to perform removal of outlier using Inter Quantile Range (IQR).

Q1=25th percentile

Q3=75th percentile

IQR=Q3-Q1

Lower\_bound= Q1 - (1.5 \* IQR)

Upper\_bound= Q3 + (1.5 \* IQR)

**Statistical Tests**

A statistical test provides a mechanism for making quantitative decisions about a process or processes. Our intent is to perform a statistical test to find out *whether the price of rooms across all neighbourhood groups are similar or not*. For that we have created a group of neighbourhoods and taken the mean of the price & then picked two neighbourhood group’s prices where the values are nearer to each other and to get what we want we are using Hypothesis Testing**.**

Where the Null Hypothesis is accepted when the price of rooms of neighbourhood groups are equal and alternate Hypothesis when they are not equal.

Result of the Hypothesis were:

price of rooms in Bronx and Queens are not equal

price of rooms in Brooklyn and Staten Island are not equal

price of rooms in Queens and Staten island are equal.

price of rooms in Manhattan are not equal to any other.

Another Statistical test we have done to find out whether the room type are similar to neighbourhood groups so to find out we have used Chi Square test.

* HO: Not Related (no association)
  + alpha > 0.05
* HA: Related (association)
  + alpha < 0.05

## **Chi-square = ∑(fe−fo)2fe**

Our goal was to find out the value of alpha and alpha value was less than 0.05 so we can say

# neighbourhood groups and room types are related

**Conclusion**