Airbnb Booking Analysis

Names

Data Science Trainees

Alma Better

**Abstract:**

Our team has extensive experience with Airbnb Booking Analysis. We intend to concentrate solely on New York City. The investigation is mostly focused on the boroughs of Brooklyn, Manhattan, Queens, Staten Island, and the Bronx. This information includes information about the host, lodging prices, and reviews, among other things. With the use of exploratory data analysis, data wrangling, visualization, and other tools, we can extract a lot of information from the analysis.

**1.Problem Statement:**

The customer can book a room through the Airbnb website. To explore more, the customer is moving to the Queens, Bronx, Staten Island, Brooklyn, and Manhattan neighborhoods. The customer requires assistance in identifying a number of bookings in each borough country and guiding that customer in selecting the best place to live at an affordable cost, analyzing the room's price so that the customer can afford it, and selecting the appropriate room type based on their needs. The quantity of reviews, the number of reviews per month, and the availability of the rooms will all meet these criteria. All of these insights will be obtained through EDA and a thorough comprehension of the data.

* **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:**

Airbnb is a database of accommodation bookings. This data is only available for New York City from 2011 to 2019. We can grasp all of the features by exploring the data. The objective is to look into the data and then evaluate it with all of the necessary information. We must extract crucial insights from the data through data exploration, data cleansing, and analysis. So that the insights can be leveraged to guide marketing and gain a better knowledge of customers' and hosts' behaviour.

**Steps Involved:**

**Exploratory Data Analysis:**

Using the pandas package, we loaded our data. We printed the top 5 rows and did different operations, such as casting the columns into their appropriate data types, to better comprehend the data. Statistical data, univariate analysis, and the accuracy of data in each column were all investigated. Understanding the relationship of all numerical and categorical columns. In addition, I performed Bivariate Analysis on both numerical and categorical data. We also plot a numerical column distribution to see if it follows a Gaussian distribution. We used a variety of plots to gain a better understanding of the data and presented it in an attractive way.

**Visual Analysis:**

Visualization consolidates thousands of crucial data points into a single graph. To acquire a better grasp of the data, we employed a variety of charts and presented it in an enjoyable manner. Seaborn, plotly, and matplotlib are used to visualise the data. We used a variety of plots, including bar charts, pie charts, map charts, box plots, violin plots, waterfall plots, treemaps, and paretto charts, among others. After cleaning the data, it is necessary to conduct a data analysis. This makes the data more interactive, and it allows you to receive immediate insights from the graph to help you make better decisions and meet your business objectives.

Checked correlation and done univariate analysis

**Treating Null Values:**

The provided data contains null values. These null values appear as a result of some sensitive data that the user did not want to specify and neglected to enter. A heat map plot can be used to visualise it. The null value handling is essential for gaining deeper insights and useful data. This null value handling can be inferred in a variety of ways.

**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:**

An outlier is a data point that deviates from the normal distribution of data. It also cause the data's original distribution. The Outlier is produced by experimental errors, incorrect data entry, and other factors. Box plots and violin plots can be used to visualise the outlier. The technique employed here is Inter Quantile Range (IQR) to remove outliers (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 is a tool that allows you to make quantitative conclusions about a process or set of activities. Our goal is to do a statistical test to see if the prices of rooms in different neighbourhoods are similar or not. To do so, we built a group of neighbourhoods and took the mean of the prices, then chose two neighbourhood groups with prices that are closer to each other, then we used Hypothesis Testing to get what we wanted. When the prices of rooms in neighbouring groups are equal, the Null Hypothesis is accepted, and when they are not, the Alternate Hypothesis is accepted.

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.

**Transformation:**

In EDA, transformation is important. Because the Normal Distribution is used in many Machine Learning models. With the help of the Central Limit Theorem, we can convert non-Gaussian numerical columns into Normal Distribution. The central limit theorem's goal is to convert any distribution into a normal one. It requires n samples from the population and number samples to draw from the population. Once n number of samples drawn it calculates the mean of each drawn samples. The distribution will then approach the normal distribution once plotted.

**Conclusion**