**Technical Documentation**

**CAPSTONE EDA PROJECT**

**ON**

**HOTEL BOOKING ANALYSIS**.

**Insights from Problem:**

**When is the best time for booking a hotel?**

The best days to book a hotel room can be major holiday when management is not working.

* Weekdays
* Weekends
* Month by month
* Year by year

The best days of the week to book a hotel room can be Friday, Saturday, Sunday.

**When is the peak seasons to hotel industry ?**

* The months of November through March are often off-peak for hoteliers. The attendance is at its peak, runs from June through the end of August.

**who are the target audience who book a hotel?**

* Target clients for a hotel could include business travellers, tour groups, small conference groups, and individual tourists, vacationers, workers, and event planners are some examples of a hotel.

**How hotels generate revenue?**

* Hotels generate income by charging guests to stay. Other services like room rents, food and beverage sales etc generate revenue for them.

**Algorithm followed:**

* importing the required libraries and packages
* Using Colab, mount the device and open the.csv [Comma Separated file].
* examining the data table.
* deleting duplicate, NAN, and null rows.
* removing the anomalies.
* Applying the concept of Data Wrangling and Data Visualization so we can analyse the dataset and retrieve necessary information.

***Step 1:***

* importing the required libraries:
* The most widely used Python package is NumPy [Numerical Python]. efficiently utilised for array calculation problems.
* Analysing tabular data is an excellent use of the Pandas library. It can be used for exploratory data analysis, including cleaning, wrangling, and manipulating data.
* With the use of the Matplotlib package, we can better visualise our tabular data in the form of [Pie Chart, Bar Graph, Line Graph, Histogram, etc.] which aids in doing the necessary analysis.

***Step 2:***

* Analysing the information sheet to identify any NULL/NAN/Missing values. As these numbers affect the correctness of the results, we wish to get rid of them. We frequently treat outliers as well; in essence, these are data points that are greatly dispersed from other data points and that would skew visual outcomes.

***Step 3:***

* Now that our dataset is clear of all ambiguity, we will go forward with the process of exploratory data analysis (EDA), which refers to the act of analysing data in order to produce visual results.

***Step 4:***

* Matplotlib effectively uses data frames and arrays to present the data for visualisation. It views axes and figures as objects. There are several stateful plotting APIs in it.

**Terminology that is frequently used:**

***Data Wrangling***

* Data wrangling is the process of organising and cleaning up large and disorganised data sets so they are simpler to access and analyse. the process of transforming and mapping data from one "raw" data type into another with the goal of improving its suitability and value for a range of downstream uses, including analytics.
* Functionalities performed under it:

Data exploration, handling NAN and missing values, removing duplicates, and data filtering.

**Data Visualization:**

It involves converting tabular data into graphic images so that users may quickly and easily understand it.

* It is employed because visualising graphical data yields considerably better outcomes than doing so for tabular data. The verbal and statistical descriptions of the data set are closely integrated during visualisation. This identifies problem areas and directs attention to them.

**Exploratory Data Analysis**

* Exploratory data analysis is a crucial procedure that entails performing early investigations on data in order to find patterns, identify anomalies, test hypotheses, and validate assumptions with the aid of summary statistics and graphical representations.

Some important functions used:

1. df.head( ):The Dataframe's top five rows are what are returned by default. The number of rows returned can be changed by inserting a value between the parentheses to override the default.
2. df.tail( ): The latest five rows of the Dataframe are returned by default. Position determines how you utilise this function.
3. df.describe( ):For each of the dataset's numerical columns, provide a statistical summary. This method computes some statistical measurements for the numerical values of the Series or DataFrame, such as percentile, mean, and standard deviation.
4. df.isnull( ).sum( ): Provide the total number of values missing from each column.
5. df.shape:It displays both the total number of dimensions and their respective sizes. The shape that is returned is the number of rows and columns because data frames are two-dimensional.
6. df.info( ): It helps in quickly gaining an overview of the dataset. To obtain a concise summary of the dataframe, use this function. This method prints details about a DataFrame, such as the non-null values, column and index dtypes, and memory utilisation.

**Tools using which we create graphical representation/Pictorial data**

***Lineplot:***

* on a different axis to depict the relationship between two data points, X and Y.

***Barplot :***

* A bar chart, often known as a bar graph, is a diagram that displays categorical data as rectangular bars with heights or lengths proportional to the values they stand for. You can plot the bars either vertically or horizontally. A vertical bar chart may also be referred to as a column chart.

***Boxplot:***

* A box plot is a visual representation of the location, dispersion, and skewness groups of quartiles of numerical data. The image of an "outlier" is the most significant one it offers.

***Stripplot:***

* An independent stripplot is created. When all data and a representation of the underlying distribution are provided, it is a useful addition to a boxplot or violinplot. On the basis of the category, a scatter plot is created.

***Pieplot:***

* A circular statistical graphic called a pie chart can only show one series of data at a time. The overall percentage of the provided data is represented by the chart's area. The percentage of the data pieces is represented by the area of the pie slices. Pie charts are frequently used in corporate presentations since they offer a quick summary of information like sales, operations, survey results, resources, etc.

***Countplot:***

* Utilizing bars, a method is utilised to display the numbers of observations in each category bin.

***Heatmaps:***

* A heatmap is a graphical display of data in which values are represented by colours. A heatmap that displays a 2D correlation matrix between two discrete dimensions and uses coloured cells to represent data from typically a monochromatic scale is called a correlation heatmap. Heatmaps make the association between one feature (variable) and every other feature incredibly simple to understand (variable).

**Key indicators:**

* Key performance indicators (KPIs) are benchmarks that allow you to track your progress towards your most important strategic goals. Despite the fact that businesses can use a wide range of indicators, KPIs are objectives that are "essential" to the development of our company.

**Some key indicators are:**

***Bookings:*** We can see the reservations for two hotels [City, Resort] in this dataset.

Based on this, we deduced that the percentage of returning visitors is quite low, an area where we need to focus. If a hotel does not receive repeat business, this indicates that the customers are dissatisfied with the services

***Average Daily Rate:*** The Average Daily Range (ADR) reflects the average of earlier values, much like a moving average. It is calculated as the sum of all hotel room revenues for a specified time period divided by the total number of booked rooms.

***Conclusion***

With the help of the above mentioned approach, we were able to fully evaluate the data set and prioritise key parameters that influence booking, cancellation, average daily rate, and many other factors.

***Project by***

Ravi Kumar