Capstone Project
Hotel Booking Analysis (EDA)
(Snehendu Jana)

Data science student
Alma Bette

Insights of the project:

We are provided with a records sheet which includes all the applicable data required to examine the booking system of the hotel industry. Utilizing the principles of EDA (Exploratory Data Analysis), Data Wrangling, and Data Visualization we will analyze the elements affecting hotel bookings. These elements may be used for reporting the developments, causes and the trends in the bookings.

Algorithm followed:

- Importing necessary packages and libraries
- Mount the drive in colab and read the .csv [Comma Separated file].
- Analysing the data sheet.
- Removing null/NAN/duplicate rows.
- Fixing the outliers.
- Dropping certain columns/combining certain columns to make our dataset free of any irrelevant data.
- Applying the concept of Data Wrangling and Data Visualization so we can analyse the dataset and retrieve necessary information.

Step 1:

- importing the package:
- NumPy[Numerical Python] the most significantly used package in python. Efficiently used for solving array computation.
- Pandas library is good for analyzing tabular data. You can use it for exploratory data analysis [data cleaning , data wrangling ,data manipulation].
- Matplotlib is a library using which we provide a better visuals to our tabular data in the form of [Pie Chart, Bar Graph, Line Graph ,Histogram etc]which help us to make appropriate analysis.

Step 2:

- Analyzing the information sheet and informing the NULL/NAN/Missing values if any. we want to get rid of these values as they influence the accuracy in results. we tend to treat the outliers too, essentially they are data points that are far away from different data points and that they will distort visual results

Step 3:

- Now our dataset is free from all the ambiguity we will proceed with the

Capstone Project
Hotel Booking Analysis (EDA)
(Snehendu Jana)

Data science student
Alma Bette

process of EDA (Exploratory Data analysis] referring to the process of performing analysis on data so as to get the visual results.

Step 4:

- Presenting the data to visualization, Matplotlib works efficiently with data frames and arrays. It treats figures and axes as objects. It contains various stateful APIs for plotting.

Frequently used terminologies:

Data Wrangling

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. Transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
- Functionalities performed under it: -

Data exploration, dealing with missing /NAN values, Duplicate values, filtering the data.

Data Visualization

It's a process of transforming the tabular data into graphic visuals, to make it easy and quick for user to understand.

- As the vision to a graphical data is much more effective in results than visualizing in tabular data visualization is used. Visualization closely integrates statistical and verbal descriptions of data set. In this way, areas that need improvement are identified and addressed.

Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Some important functions used:

1. df.head( ): By default, it returns the first 5 rows of the Dataframe. To change the default, we may insert a value between the parenthesis to change the number of rows returned.

Capstone Project
Hotel Booking Analysis (EDA)
(Snehendu Jana)

Data science student
Alma Bette

2. df.tail( ): By default, it returns the last 5 rows of the Dataframe. This function is used based on position.

3. df.describe( ): Return a statistical summary for numerical columns present in the dataset. This method calculates some statistical measures like percentile, mean and standard deviation of the numerical values of the Series or DataFrame.

4. df.isnull( ).sum( ): Return the number of missing values in each column.

5. df.shape: It shows the number of dimensions as well as the size in each dimension. Since data frames are two-dimensional, what shape returns is the number of rows and columns.

6. df.size: Return an int representing the number of elements in this object. Return the number of rows if Series, otherwise returns the number of rows times the number of columns if DataFrame.

7. df.info( ): It helps in getting a quick overview of the dataset. This function is used to get a brief summary of the dataframe. This method prints information about a DataFrame including the index dtype and column dtypes, non-null values, and memory usage.

Tools using which we create graphical representation/Pictorial data

Line chart :

- are used to represent the relation between two data X and Y on a different axis.
- Syntax:- plt.plot(x, y, color ="red")

Bar Chart :

- A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.
- Syntax:- plt.bar(x, height, width, bottom, align)

Capstone Project
Hotel Booking Analysis (EDA)
(Snehendu Jana)

Data science student
Alma Bette

Box Plot:

- A box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. The most important visual it provides is of an 'outlier'.

This is what each boxes describe in box plot.

Heat maps:

- A **heatmap** (or **heat map**) is a graphical representation of data where values are depicted by color. A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale. It is very easy to understand the correlation using heatmaps it tells the correlation of one feature(variable) to every other feature(variable).

Scatter Plot:

- Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

Alterations made in data-set

1) Created a copy of original data sheet named it as Hotel_df.
2) Renaming the column 'adr' by Average_Daily_Rate.
3) Creating new columns to data set: -
   - I. Total_member-created by adding the values from Children, Babies, Adults columns.
   - II. Full_Stay – created by adding the values from stays_in_weekend_night and stays_in_week_nights column.
   - III. kids-created by adding the values from Children, Babies column.
   - IV. Converting the data type of columns Country, Company, agent to int.
   - V. Created a subset from main data set as City_df , Resort_df.

Key indicators:

Key performance indicators (KPIs) are targets that help you measure progress against your most strategic objectives. While organizations can have many types of metrics, KPIs are targets that are "key" to the success of our business.

Capstone Project
Hotel Booking Analysis (EDA)
(Snehendu Jana)

Data science student
Alma Bette

Some key indicators are:

Bookings:

- Here in the data sheet we observe the bookings for two hotel [City, Resort] Based on this we figured out the rate of visit of repeated guests is very low on which we need to work as it would impact the overall rating of hotel. If the hotel is not visited by repeated guests that means that are not happy with the services and switch for other options.

- Secondly, we observed despite of having no deposit for advance bookings we could see rate of cancellation is higher. So, hotel should charge some amount in advance for booking this would help in decreasing the cancellation rate.

Average Daily Rate:

- The Average Daily Range is a technical indicator used to measure volatility in an asset. Typically, this indicator is used to signal a significant change in price action over the short term. The Average Daily Range (ADR) is like a moving average in that it reflects the average of previous values. It is measured as the total revenues generated by all the occupied rooms in a hotel divided by total number of occupied rooms over given time period.

Challenges faced and overcome:

- A lot of resources were referenced by each one of us so it took us much time to finalize the key indicators.
- While plotting the char/ string values for country column first we need to covert it in list then we can plot.
- After analysing the data set a lot of observations were noted by each one of us so in what flow we should proceed in project was a challenge! So we distributed it accordingly like whoever has suggested the observation will write a code by their own and then we would debug[if required] and collaborate it at the end.
- Choosing the appropriate visualization techniques to use was difficult.

Conclusion

From the above-mentioned procedure, we are able to analyze the data set to its fullest and have made the decisions prioritizing certain factors which govern the booking, cancellation, average daily rate followed by many such factors.

References:

1. www.geeksforgeeks.org
2. www.tutorialspoint.com

Capstone Project
Hotel Booking Analysis (EDA)
(Snehendu Jana)

Data science student
Alma Bette

3. https://pandas.pydata.org/pandas-docs/stable/index.html

Team Members

I. Akshay Fasale
II. Kanika Kakra
III. Rishikesh Damale
IV. Shubham Joshi