Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Ashok Kondhalkar **Email:** ashokkondhalkar98@gmail.com **Contribution:** 1.Data Wrangling 1.Hotel Dataset 2.Data Reading and inspection 3.Data Cleaning 2.Correlation with numerical data by using correlation matrix & scatterplot 3.Hotel percentage by pie chart 4.Hotel lead time and preferred stay by countplot 5.Hotel Higher bookings cancellation rate by barplot 6.Distribution channel longer average waiting time by barplot 7.Hotel having high revenue by boxplot 8.Type of meal booked by pie chart 9. Yearwise booking of hotel by plot bar **Team Member's :** **Name:** Deepanshu Kanchan **Email:** deepanshukanchan120@gmail.com **Name:** Kunal Wankhede **Email:** kunalwankhede568@gmail.com **Name:** Shajad **Email:** shehzadglocal786@gmail.com **Name:** Sushma M **Email:** sushma.gowda.m94@gmail.com
Please paste the GitHub Repo link.
Github Link:- https://github.com/ashokkondhalkar/CAPSTONE-1-Hotel-Booking-Analysis-EDA.git
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

When we decided to make Hotel Booking Analysis as an EDA project. i'm quite nervous.but,when i'm reading problem statement & csv file and imagining whole things in mind i'm feel too good.csv file contain large amount of data so, we decide to first work individually gaining some insides about EDA then we collaborate data & idea together. Thus,I'm start imaging & framing my own question and do the same random analysis to build confidence.

First I'm importing all important libraries.Mount drive and provide path for accessing csv file.Then, i'm trying to do some random analysis.

The very first problem that I faced was a csv file having duplicate rows. Also,i having columns with zero and null values.so,i try to read and inspect all data and perform data cleaning operation like *find unique values, *remove duplicate row, *finding missing values,*fillna & *changing datatype of columns, *creating new columns .Thus,I can easily use Data for Analysis.

Then I decided to do individual and team wise analysis.For which I took help of correlation matrix,pie chart,scatter plot,barplot,boxplot.etc for better visualization.

This is my best EDA in life.i'm practically understand how Data is filter used in every aspect.while filtering Data i've some question in mind where i'm try to find out answers by using Data Analysis and Visualization

Q1)What is the type of hotel in the market and percentage in each Type?
Approaches:

- Here we have data about hotel .so ,we are used .value count() for counting values and .index() for for namming purpose(i.e Pie chart)
- Also,autopct='%1.2f%%' provides percentage to Piechart

interference:

- According to Analysis city hotels are comparatively more expensive than resort hotels.
- city hotel having 66.45% and Resort hotel having 33.55% percentage in hotel market
- Resort hotel should me more expensive so,peoples are stick to city hotel

Q2)Which hotel has higher lead time and preferred stay in each hotel?
Approaches:

- Here we create a group of hotel and lead time by using .groupby and find the median of lead time.
- To find preferred stay in each
- We first find hotel having zero cancellations, we compared 'is_canceled' with zero
- Second we find stay length i.e. 'totale stay'<15 & plot the both values by using countplot

interference:

- From Analysis we found that lead time of city hotel is more than Resort Hotel
- Resort hotel having longer stay compared to city hotel.i.e.For short stay peoples are choose City Hotel

Q3) Find which hotel has a higher bookings cancellation rate?

Approaches:

- we first find hotel having no of bookings, we compared 'is_canceled' with one and making group with 'hotel and finding size of hotel by .size
- Also,We convert to DataFrame and rename column by using .rename
- Here ,we use .concat for combined two values in one table .like Total canceled bookings and Total bookings and finding its percentage
- This is represented by using barplot

interference:

- From Analysis we can say that City Hotel bookings are more canceled than Resort hotel

Q4) Find which distribution channel has a longer average waiting time?

Approaches:

Here we make group with 'distribution_channel' and 'days_in_waitng_list' and finding mean of 'days_in_waitng_list'

Also,we rename columns by using .rename and analyze whole things by using barplot.

interference:

- From Analysis we can say that TA/TO distribution channel having longer waiting time compared to Corporate and Direct

Q5)Which month hotels have high revenue?

Approaches:

- Here we are First make list of moths present in year.i.e.month_in_year
- Also,we are using .Categorical for collect all data in list month_in_year and plot figure by using box plot

interference:

- From Analysis we can say that less peoples are visited to Hotel in january month.so,revenue having huge cut off
- But,hotels have large revenue in August due to more people visiting.

Q6) Which type of meal is booked?

Approaches:

- Here we find various types of meal available in hotel so,First count values of meal by value_counts() and .index for naming purpose
- Also.to_list which converts data from array to list and explode which separate each section from one another.

interference:

- From Analysis it is clear that BB.(i.e Bed & Breakfast) is most preferable type of meal for 77.8% guest

Q7)Find from which country most guests come from?

Approaches:

- Here we apply condition to find no of bookings for 10 rows only & country values by using .value_counts()[:10]
- Then renaming columns & finding percentage to show on plotting barplot

interference:

- From Analysis it is clear that most of the peoples are come from PRT (i.e.Portugal)

Q8)Find year wise booking of hotel?

Approaches:

Here we first find unique values by .unique() then find values by using value_counts() Then this is shown in .plot bar

interference:

- From the pie chart Graph it is clear that 2016 had higher bookings compared to 2017 and 2015.
- so,according to given data there is increment of booking with alternate years

Conclusions:

- Average daily rate (adr) is directly proportional to totle_peoples.No peoples increases then revenue must be increased.
- The percentage of city hotel is 66.45%.while the percentage of resort hotel is 33.55% is use to stay.So,City hotel connect more no of peoples and having higher lead time
- For longer stay peoples are choose Resort hotel and for short stay choose city hotel
- City Hotel bookings are more canceled
- Here,TA/TO distribution channel having longer waiting time
- More people visit hotels in August and less people visit in January.
- BB.(i.e Bed & Breakfast) is most preferable type of meal for 77.8% guest
- Most of the peoples are come from PRT (i.e.Portugal)
- Year 2016 having higher bookings compared to year 2017 and 2015.

Reference:

- Almabetter
- Geeksforgeeks
- Stackoverflow
- Youtube