**Question 1**

a :- Load the dataset into Python and identify the type of the dataset features and report them.

I have first loaded the dataset on the jupyter notebook but since the file is to big than I uploaded in colab and start working on it. With very first I call the library below library by uploading the data on gooledrive than I Than I call the library pandas and used it to read the file after giving the path the file is in csv format and below is the code which I used.mount it. Below are the codes which I used.

from google.colab import drive

drive.mount('/content/drive')

import pandas as pd

path = "/content/drive/MyDrive/train\_small.csv"

df = pd.read\_csv(path)

The type of dataset feature are as follow.

b :- Perform summary statistics and explain what issues these statistics reveal.

Summarizing Data

The **describe()** function computes a summary of statistics pertaining to the DataFrame columns.

Code :-

df.describe()

In the code section you will see the output.

C :-Perform exploratory analysis to identify any collinearities and explain which issues collinearity causes.

Following are the code which is used.

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

import pandas.util.testing as tm

X = df

# VIF dataframe

vif\_data = pd.DataFrame()

vif\_data['feature'] = X.columns

# calculating VIF for each feature

vif\_data['VIF'] = [variance\_inflation\_factor(X.values, i) for i in range(len(X.columns))]

print(vif\_data)

issues collinearity causes. TypeError: ufunc 'isfinite' not supported for the input types, and the inputs could not be safely coerced to any supported types according to the casting rule ''safe''

d :- Identify the issues that are encountered in the dataset.

Since the data is to long and it takes lot of time also to get execute it has approximately 2380557 rows and 54 columns I have run the all code in colab but on running it colab also I faced a lots of issue.Earlier I was running with hardware acclerator with none which is in automatic setting but when i run the code with the GPU by changing the setting same issue I faced it takes lot of time to get execute.

e :- Propose and implement the solutions for the issues you have found. The issues listed here are not exhaustive. If you encounter any other issues, please propose and implement solutions for those as well.

Since the data have many outlier which is very difficult for me to handle for this I have used the 2 methods which are IQR and the ZSCORE but I have not able to implement it since the data was to large and when I was running it it was taking almost a 4 hrs since I have spend a 2 days regulary to execute it but I was not able to execute it since in the GPU I face the issue of

**Warning: you are connected to a GPU runtime, but not utilizing the GPU.**[**Change to a standard runtime**](https://colab.research.google.com/drive/1DV7McWx4DzO-MBhWarnZyjkjxVh_78Fw)

I dont know i am getting this issue and from the starting I have to run the cells again for me its was really a big challenge to understand the data and execute it and find the fruitful information from it.

Your session has been crashed and restart again I faced this issue more then 50 time when I try to run the code.

f :- Are there any data privacy and security issues in this dataset? If so, what are they? How would you solve these problems?

No , My reason for No is that their is no personal information of the customer or any information which customer wanted to hide.

**Question 2**

1. Calculate Click-Through Rates and Conversion Rates.

Click-through Rate (CTR) is the percentage of people that clicked on your hotel pahe to visyour landing page or website. You can calculate it with the following formula:

Clicks ÷ Impressions = CTR

For example, if your ad received 50 clicks and 1000 impressions, then the CTR is 5%.

This metric shows you the percentage of site visitors that completed the desired action. You can calculate it with the following formula:

conversions  ÷ clicks on hotel page or webside

For example, if you received 50 conversions from 1,000 clicks, your conversion rate would be 5%.

b:- How do property review score and property star rating affect click-through and conversion rates? Employ statistical techniques to formally compare the effects of different levels of these variables.

Property reviews score will give the positive or negative and it provide the information to the customer and property star rating also.

**Question 3**

1. This dataset has 3 possible outcome variables: clicked, booking\_value and booked. Select one of those variables to model and train a machine learning model of your choice.

From the 3 outcome variable I have chossed the booked feature if the listing is booked 1, else 0.

1. You are required to explain the following details:
2. Which outcome variable did you choose and why?

I have choosed the Booked feature Since from this I will get the correct customer who have booked the hotels and and done the payments.

1. Which variables did you include in the initial model?

Booked

1. Which transformations/pre-processes did you apply? Why did you apply those transformations? What would have happened if you didn't apply those techniques?

Since their is so many null values in the dataset which I replace it with the mean.

Since it is very important for me to remove the null value and replace it than only I can further work on the project

1. Which machine learning technique did you choose? Why?

Since the problem is binary classification Because I have choosed the booked in which 1 and 0 are their result

1. Which evaluation metric did you use to optimize the machine learning model?

Since by apply the algorithms

KNN=KNeighborsClassifier(n\_neighbors=6)

LR=LogisticRegression()

DT=DecisionTreeClassifier(random\_state=6)

GNB=GaussianNB()

classification\_report,confusion\_matrix,accuracy\_score,roc\_curve,auc are the evaluation matric

I have choosed.

1. Why did you choose it? What are the implications of the final metric value?

Since my code was not run in the colab I not able to see the result of metric value due to always I get issue when I am try to run the code In colab

1. How would you interpret this model? If the model is too complex to be directly interpreted, how would you approach this problem?

The model was very complex since because of the data is to large I have try to understand and give the timeI apply my knowledge and the so many approach but since I faced a lot of issue while working on colab I wound not directly interpreted because I have not even see by code to get execute in colab since I always get the message you memory get crashed it start again.

|  | **search\_id** | **site\_id** | **user\_country\_id** | **user\_hist\_stars** | **user\_hist\_paid** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **competitor8\_price\_percent\_diff** | **clicked** | **booking\_value** | **booked** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |