Capstone Project for Health Care

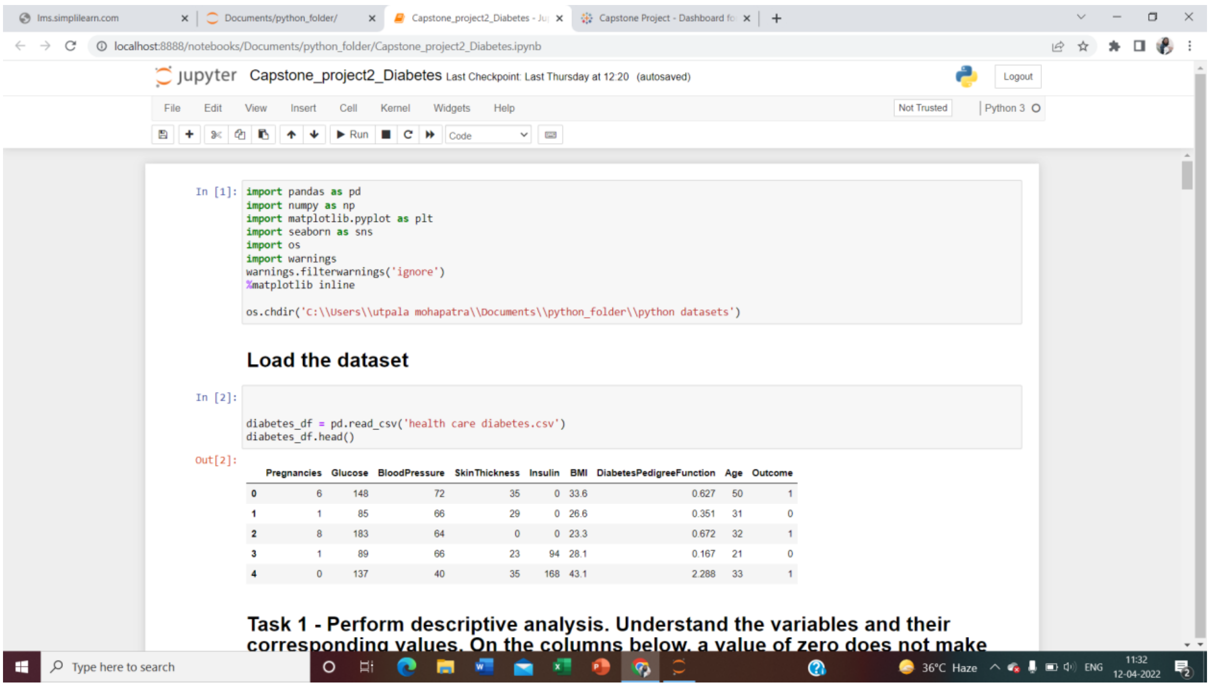
Description- NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.

- The dataset used in this project is originally from NIDDK. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
- Build a model to accurately predict whether the patients in the dataset have diabetes or not.

Data Exploration-

Task 1 - Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:

Load Dataset


jupyter Capstone_project2_Diabetes Last Checkpoint: Last Thursday at 12:20 (autosaved)

In [1]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

os.chdir('C:\\Users\\utpala mohapatra\\Documents\\python_folder\\python datasets')
```

Load the dataset

In [2]:
```python
diabetes_df = pd.read_csv('health care diabetes.csv')
diabetes_df.head()
```

Out[2]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Task 1 - Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make

Descriptive Analysis of the data.

Capstone_project2_Diabetes Last Checkpoint: Last Thursday at 12:20 (autosaved)

```
Out[6]: Pregnancies                 0
        Glucose                     0
        BloodPressure               0
        SkinThickness               0
        Insulin                     0
        BMI                         0
        DiabetesPedigreeFunction    0
        Age                         0
        Outcome                     0
        dtype: int64
```

No null values present in the database.

Descriptive Analysis

```
In [7]: diabetes_df.describe()
```

Out[7]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Task 2 - Visually explore these variables using histograms. Treat the missing

Check for missing values.

Capstone_project2_Diabetes Last Checkpoint: Last Thursday at 12:20 (unsaved changes)

Find out which columns have 0 values as 0 values means missing values and then Impute the missing Values.

```
In [5]: (diabetes_df==0).sum()
```

```
Out[5]: Pregnancies                 111
        Glucose                       5
        BloodPressure                35
        SkinThickness               227
        Insulin                     374
        BMI                          11
        DiabetesPedigreeFunction      0
        Age                           0
        Outcome                     500
        dtype: int64
```

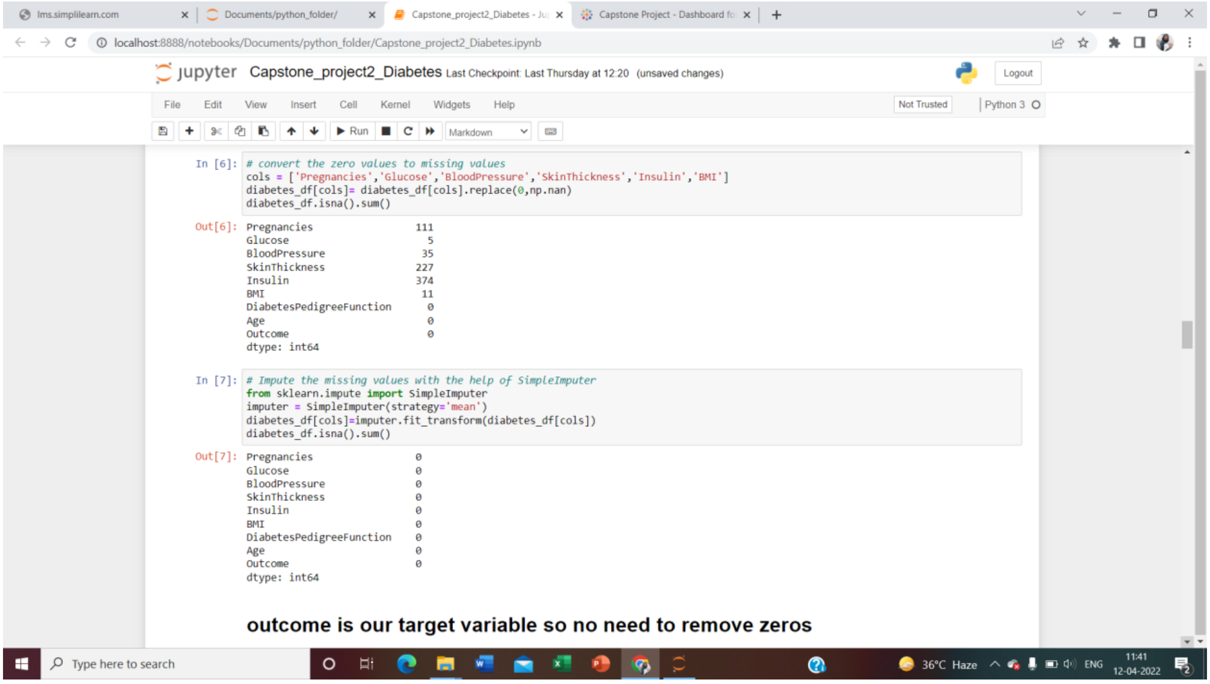
```
In [6]: # convert the zero values to missing values
        cols = ['Pregnancies','Glucose','BloodPressure','SkinThickness','Insulin','BMI']
        diabetes_df[cols]= diabetes_df[cols].replace(0,np.nan)
        diabetes_df.isna().sum()
```

```
Out[6]: Pregnancies                 111
        Glucose                       5
        BloodPressure                35
        SkinThickness               227
        Insulin                     374
        BMI                          11
        DiabetesPedigreeFunction      0
        Age                           0
        Outcome                       0
        dtype: int64
```

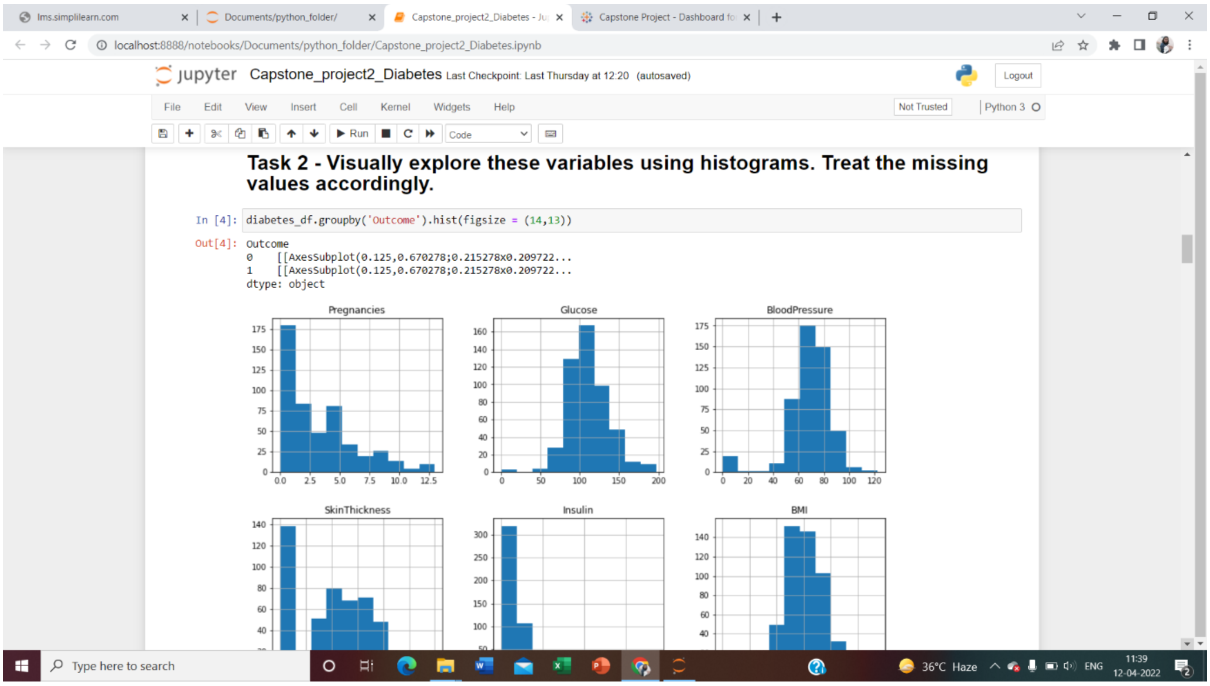
```
In [7]: # Impute the missing values with the help of SimpleImputer
        from sklearn.impute import SimpleImputer
```

Task 2 - Visually explore these variables using histograms. Treat the missing values accordingly.

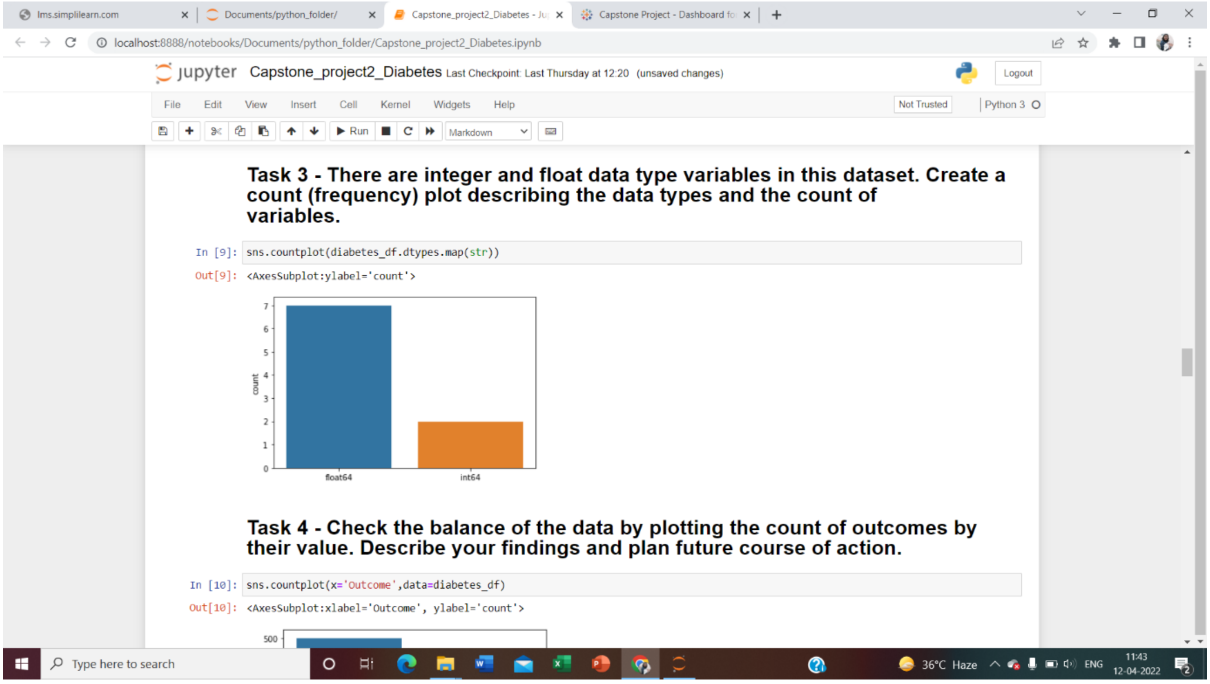
Impute missing values



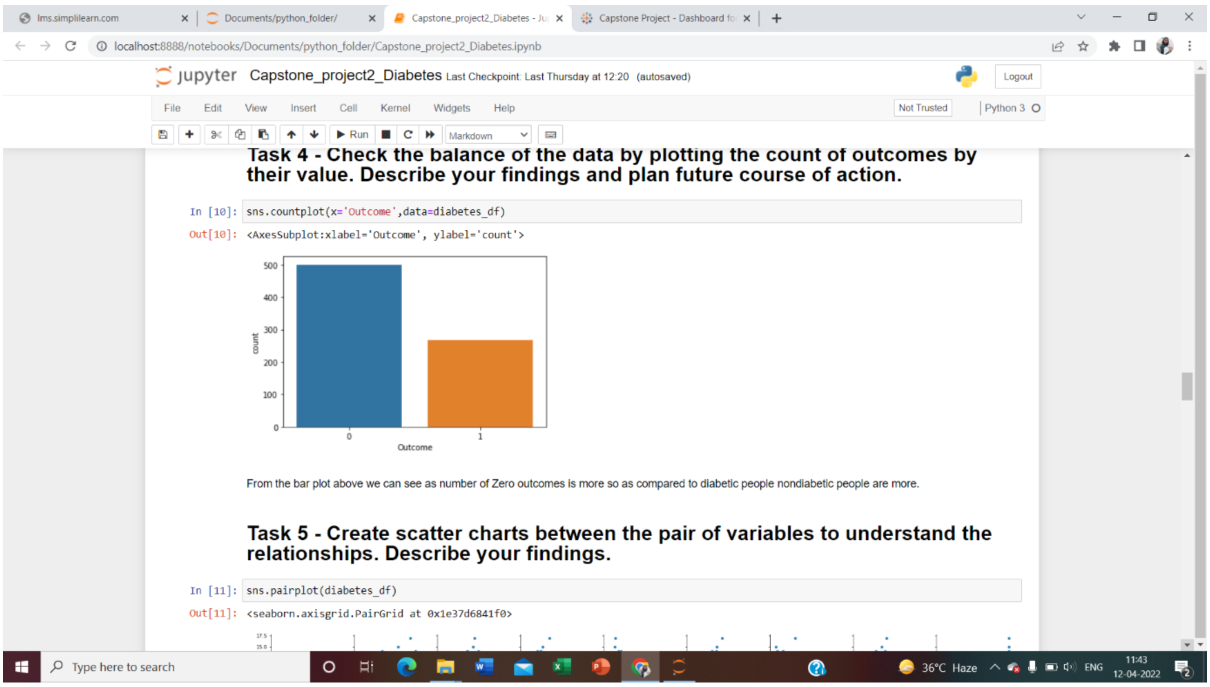
Visual analysis by Histograms



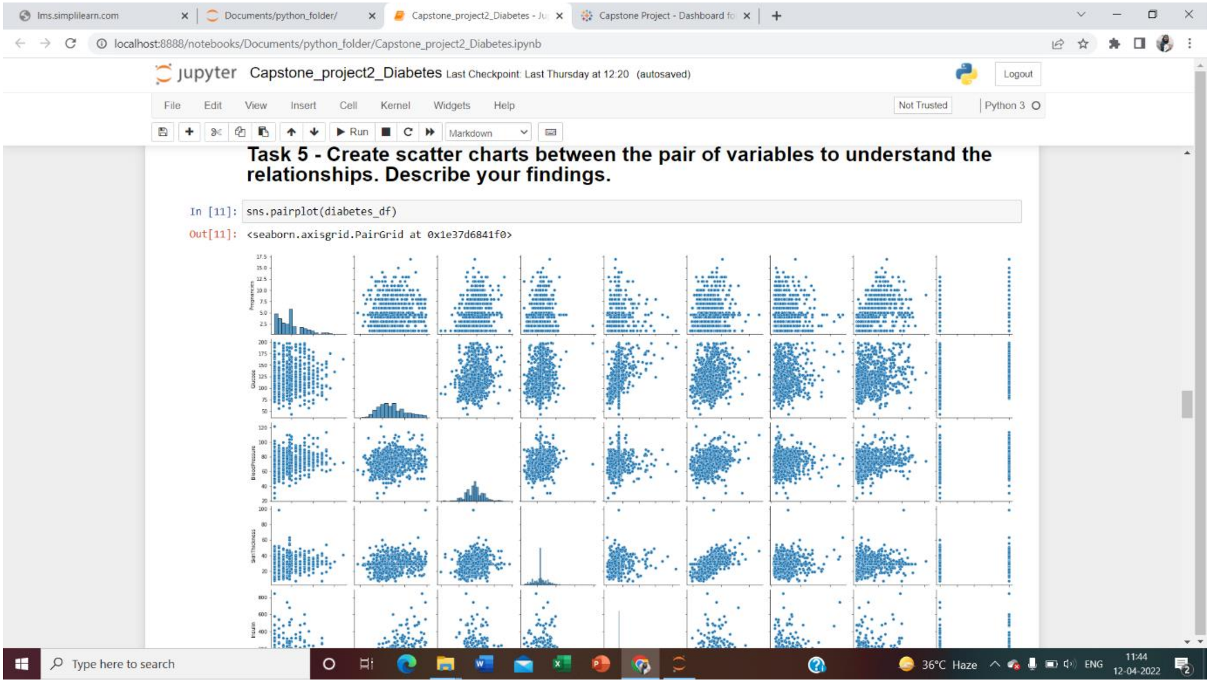
Task 3 – There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.



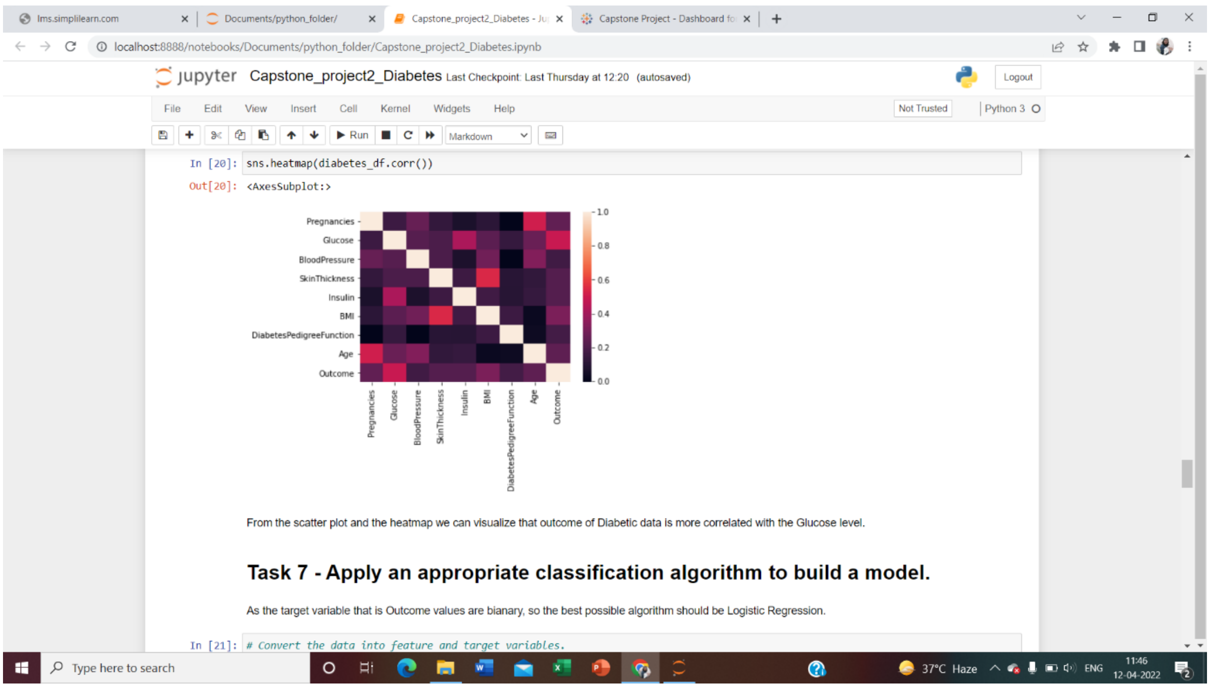
Task 4 – Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.



Task 5 – Create scatter charts between the pair of variables to understand the relationships. Describe your findings

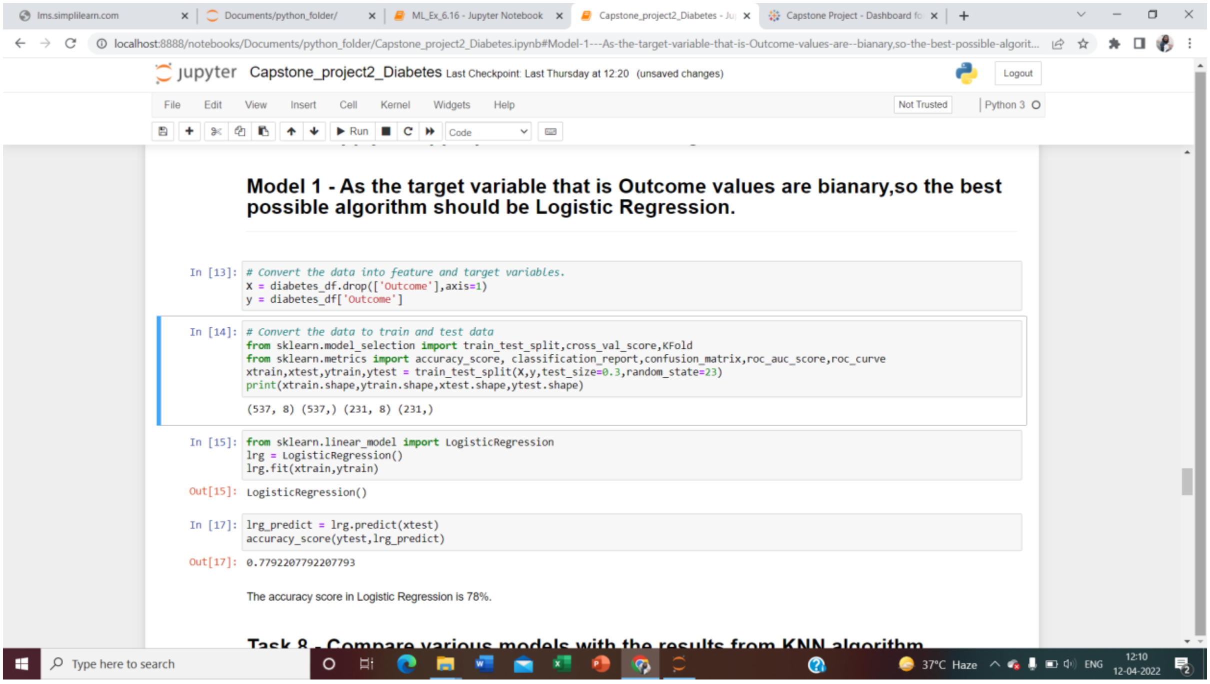


Task 6 – Perform correlation analysis. Visually explore it using a heat map.



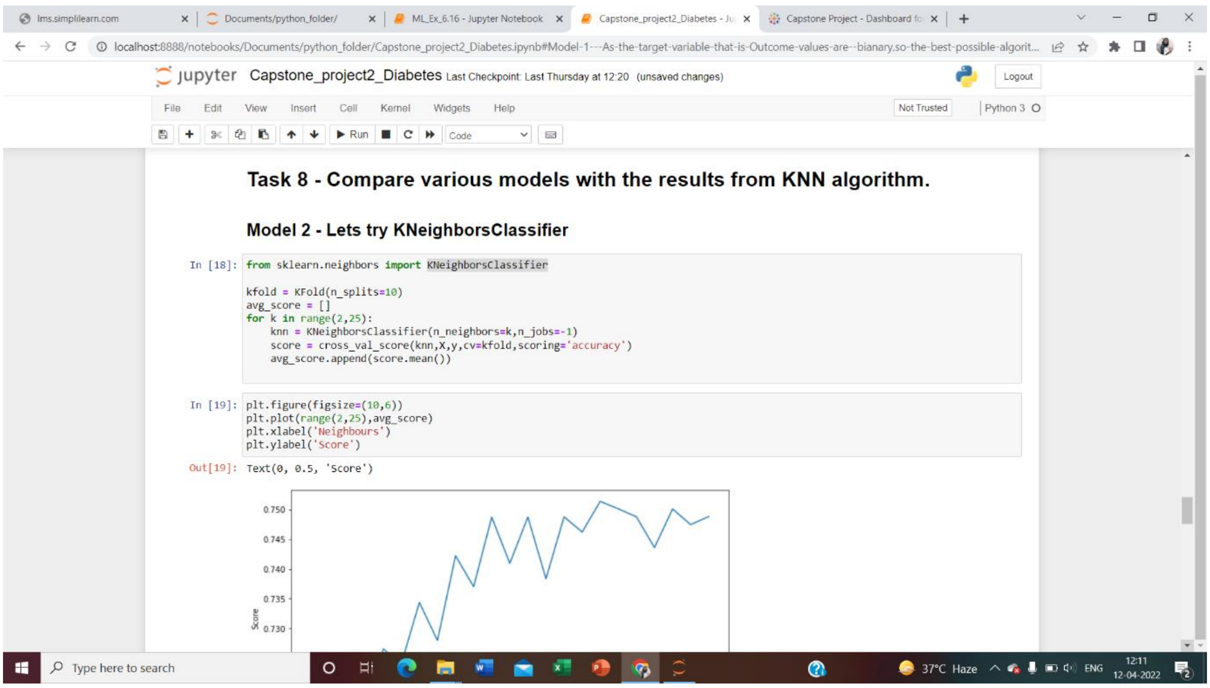
Task 7 – Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.

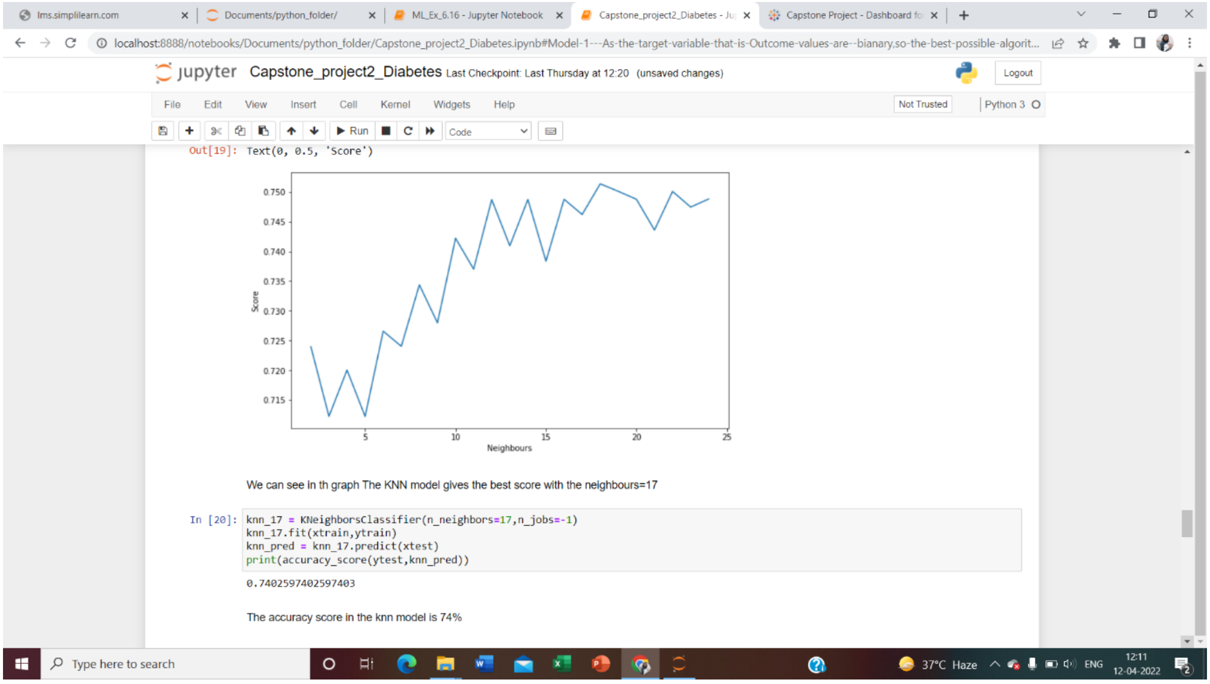
ANS - As the target variable is binary so Logistic regression should be the suitable model for this data to predict best results.



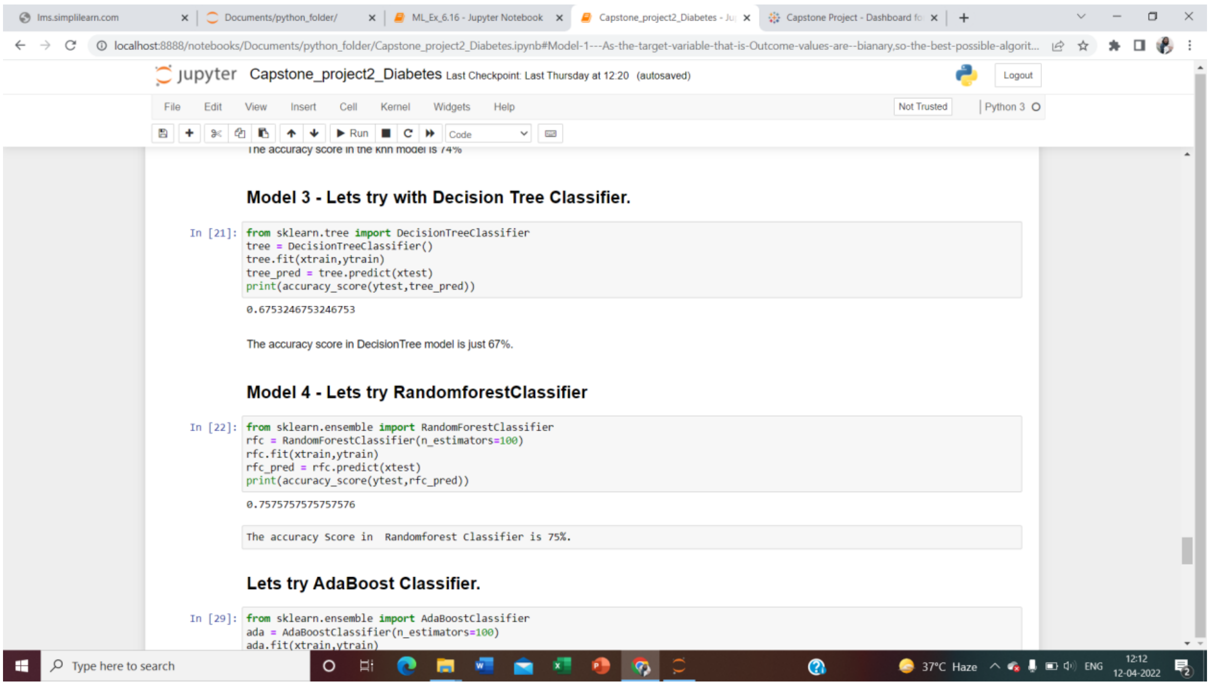
Task 7 – Compare various models with the results from KNN algorithm.

KNN model

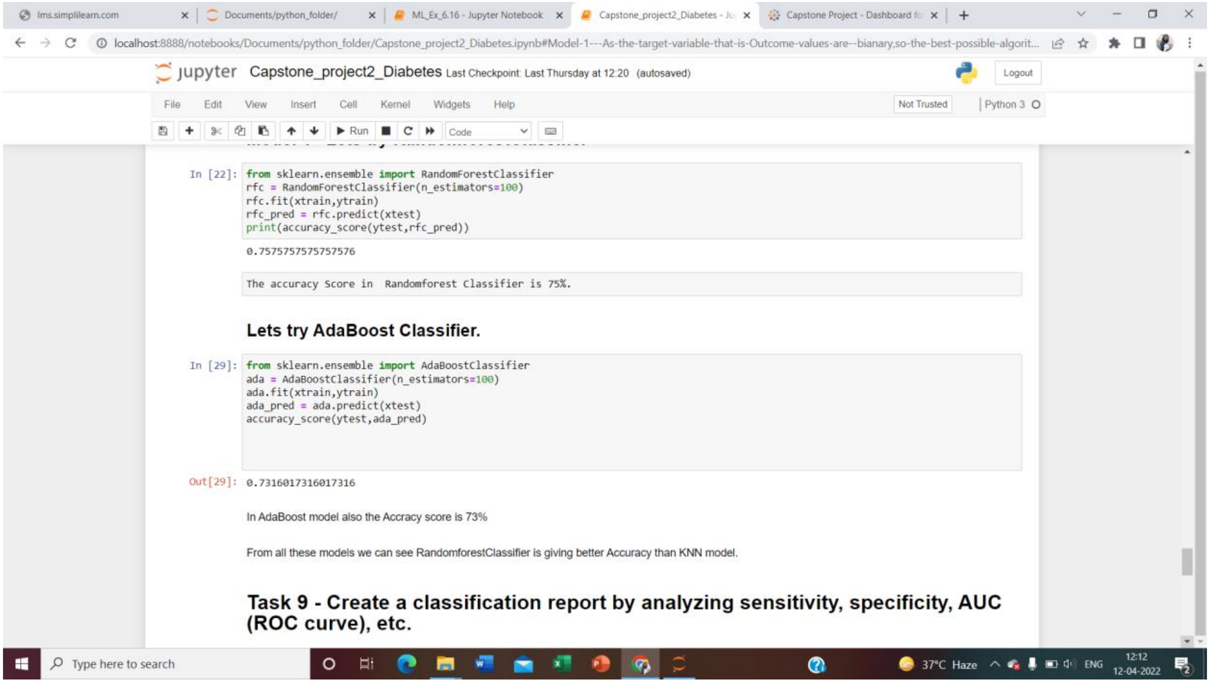




DecisionTree and RandomForest Classifier

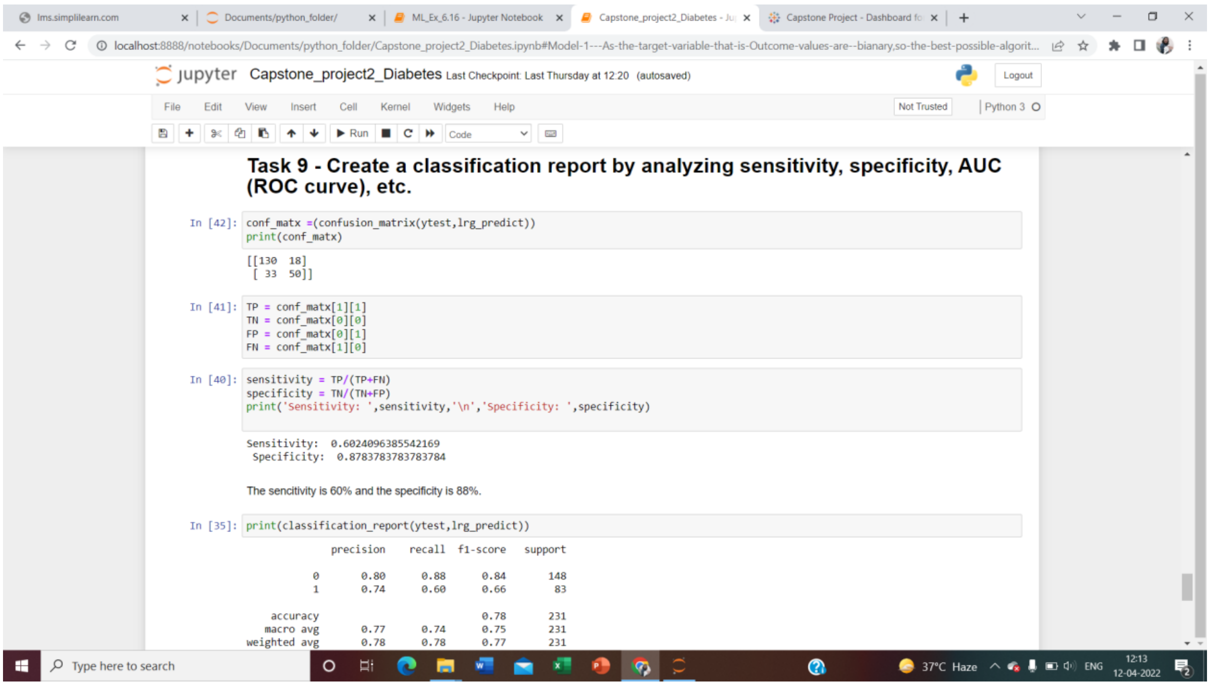


Adaboost Classifier

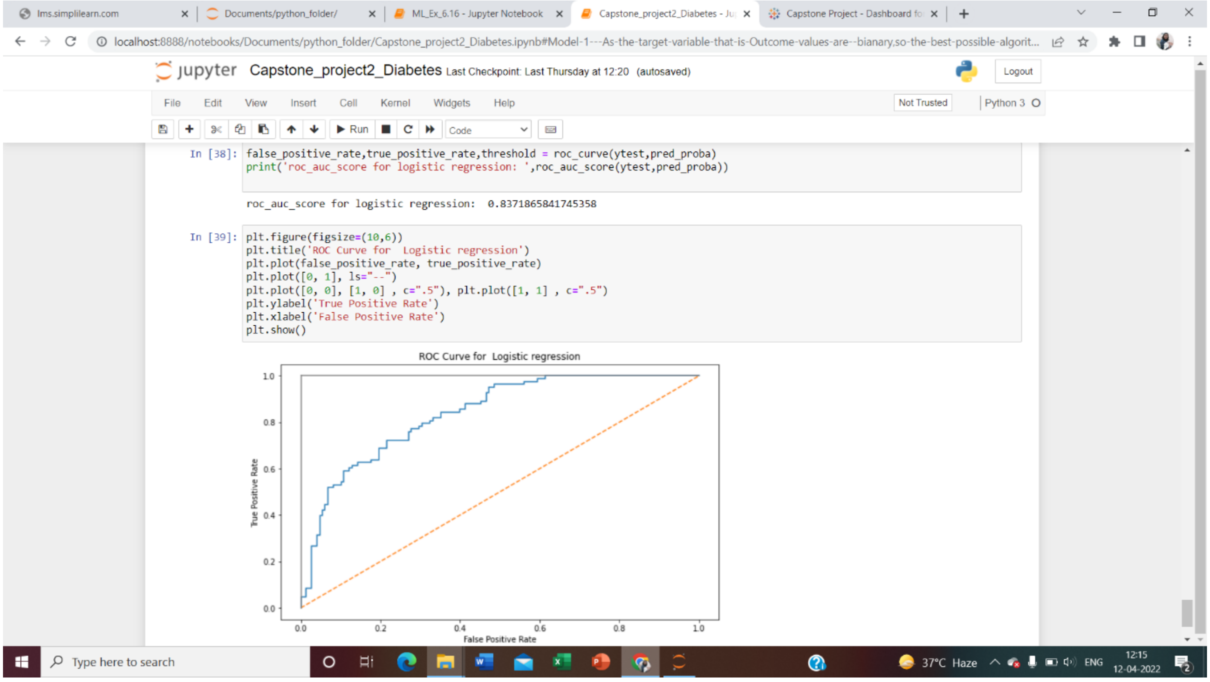


Task 8 – Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc.

Sensitivity, Specificity and Classification Report



ROC curve



Task 5 - Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

- Pie chart to describe the diabetic or non-diabetic population
- Scatter charts between relevant variables to analyze the relationships
- Histogram or frequency charts to analyze the distribution of the data
- Heatmap of correlation analysis among the relevant variables