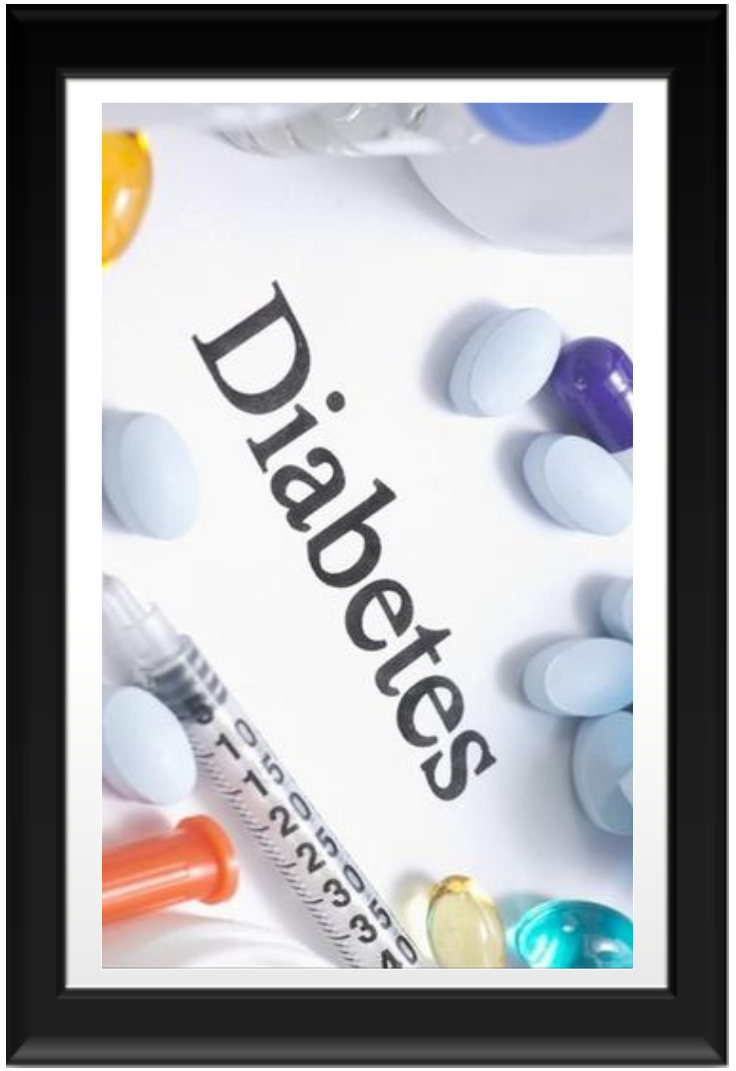
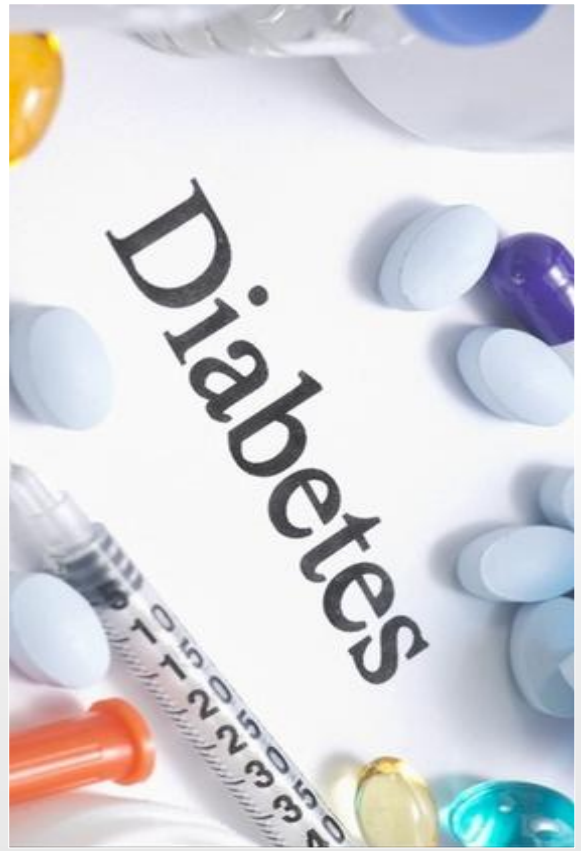
Indian Diabetes Prediction



Prof. MRUNALINI K.
(Statistics / Mathematics / Python /R /Julia / Data Science / ML / DL Trainer) , Mumbai.
G – mail : mrunalini0107@gmail.com

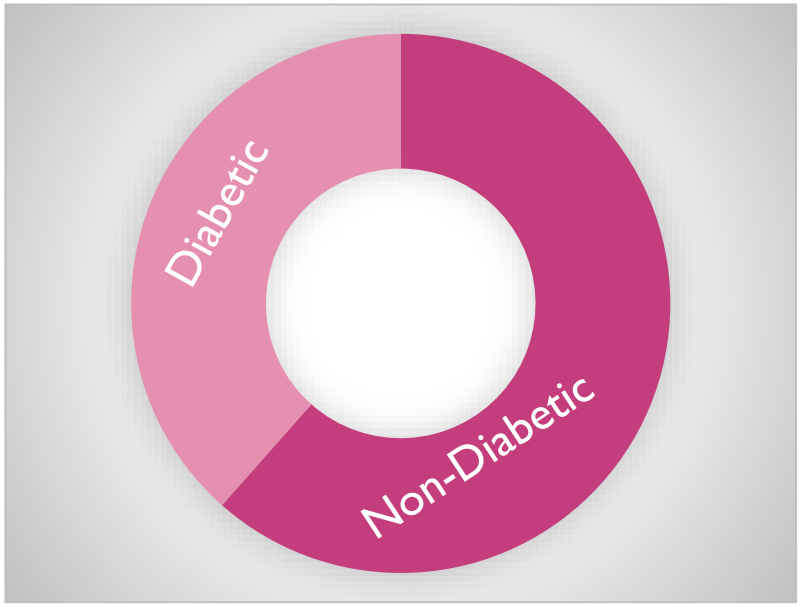
Our Activity

- Introduction
- Problem statement
- Dataset & Analyzing the Data
- Exploratory Data Analysis (EDA)
- Hypothesis Testing
- Analysis of Variance (ANOVA)
- Questionnaire
- Splitting the dataset into Training and Test data
- Conclusion



Introduction

INDIAN DIABETES PREDICTION

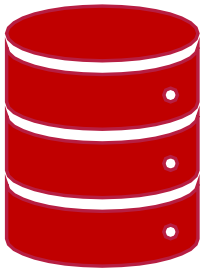


Diabetes is a chronic (long-lasting) health condition that affects how our body turns food into energy. Diabetes is a metabolic disorder in which the body has high sugar levels for prolonged periods of time. The lack of insulin causes a form of diabetes.

- **Type-I Diabetes:** It is a medical condition that is caused due to insufficient production and secretion of insulin from the pancreas. Type 1 diabetes is thought to be caused by an autoimmune reaction (the body attacks itself by mistake). This reaction stops your body from making insulin. Approximately 5-10% of the people who have diabetes have type 1
- **Type-2 diabetes**: With type 2 diabetes, your body doesn't use insulin well and can't keep blood sugar at normal levels. About 90-95% of people with diabetes have type 2.

Problem Statement

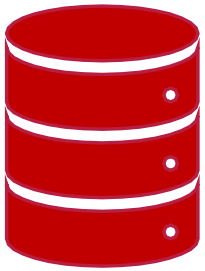
1. To predict if a particular observation is at a risk of developing diabetes, given the independent factors.
2. Is Glucose the most important factor in determining the onset of diabetes followed by BMI and Age?
3. What are the chances of risk of diabetics for persons aged more than 37 years?
4. What are the chances that persons have a glucose level more that 140, are risk to be diabetic?
5. Which approach is the best to classify a person to be Diabetic or not ?(Based on Accuracy)



Dataset

The **Indian Diabetics Prediction** dataset, also known as the **Pima Indians Diabetes Dataset (PIDD)**, is a collection of medical records of female patients aged 21 and above. It contains 768 instances, each with nine attributes.

- **Pregnancies** : Number of times the patient has been pregnant.
- **Glucose** : Result of the Oral Glucose Tolerance Test. This test checks how the body moves sugar from the blood into tissues like muscle and fat.
- **Blood Pressure** : Diastolic blood pressure values in mm Hg. This is the pressure in the arteries when the heart rests between beats.
- **Skin Thickness** : Triceps skin fold thickness in mm. Its thickness gives information about the fat reserves of the body.



Dataset

- **Insulin** : Insulin levels in the blood.
- **BMI** : Body Mass Index, a measure of body fat based on height and weight.
- **Diabetes Pedigree Function**: A function that scores likelihood of diabetes based on family history.
- **Age** : Age of the patient.
- **Outcome** : Whether the patient has diabetes or not. It's a binary value

  where 1 indicates positive for diabetes and

  0 indicates negative.

Indian Diabetics Prediction dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and can be used to predict whether a patient has diabetes based on certain diagnostic factors. Starting off, We use Python 3.3 to implement the model.

Introduction to Data

Attributes	Data Type	Remarks
Pregnancies	int64	The state of carrying a developing embryo or fetus within the female body.
Glucose	int64	A simple sugar which is an important energy source in living organisms and is a component of many carbohydrates.
Blood Pressure	int64	The pressure of the blood in the circulatory system, often measured for diagnosis since it is closely related to the force and rate of the heartbeat and the diameter and elasticity of the arterial walls.
Skin Thickness	int64	Diabetes, you're more likely to have thin skin. High blood sugar (glucose) can cause this.
Insulin	int64	Insulin is a hormone that lowers the level of glucose (a type of sugar) in the blood.
BMI	float64	Body Mass Index (BMI) is a person's weight in kilograms divided by the square of height in meters.
Diabetes Pedigree Function	float64	Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
Age	int64	The length of time that a person has lived.
Outcome	int64	The way a thing turns out

Introduction to Data

Understanding Parameters :

- **Pregnancies :** No. of times pregnant
- **Glucose :** Result of the Oral Glucose Tolerance Test.
- **Blood Pressure** : Diastolic blood pressure (mm Hg).
- **Skin Thickness** : Triceps skin fold thickness (mm)
- **Insulin** : Insulin levels in the blood.
- **BMI** : Body mass index (kg/m^2)
- **Diabetes Pedigree Function :**
- **Age :** Age of the patient in years
- **Outcome :** The target column which we are interested in finding out. $1-diabetic$, $0-nondiabetic$

Range of BMI:

$BMI<18.5$	$-underweight$
$18.5 < BMI < 24.9$	$-idealweight$
$25 < BMI < 29.9$	$-overweight$
$29.9 < BMI$	- obese

Analyzing the Data

Examining all variables in the data. Data understanding is to gain general insights about the data, which covers :

- The number of rows and columns
- Values in the data
- datatypes
- Missing values in the dataset.
- shape – shape will display the number of observations(rows) and features(columns) in the dataset
- There are 768 observations and 9 variables in our dataset

Display the top 5 observations of the dataset

	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

EDA : Exploratory Data Analysis

Exploratory Data Analysis refers to the crucial process of performing initial investigations on data to discover patterns to check assumptions with the help of summary statistics and graphical representations.

- EDA can be leveraged to check for outliers, patterns, and trends in the given data.
- EDA helps to find meaningful patterns in data.
- EDA provides in-depth insights into the data sets to solve our business problems.
- EDA gives a clue to impute missing values in the dataset

Statistics Summary

- The information gives a quick and simple description of the data.
- Can include Count, Mean, Standard Deviation, median, mode, minimum value, maximum value, range, standard deviation, etc.
- Statistics summary gives a high-level idea to identify whether the data has any outliers, data entry error, distribution of data such as the data is normally distributed or left/right skewed
- In python, this can be achieved using describe()
- describe() function gives all statistics summary of data

Statistics Summary

It is used to get the basic information about variables in a dataset and to find the potential relationships between variable.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Statistics Summary

- Based on the understanding of the parameters, it seems highly unlikely that glucose, blood pressure, skin thickness, insulin and BMI levels are 0.
- What will do is to default the 0 values to the mean of each parameter.
- Not amend the pregnancies column as it is possible that the woman in the dataset has never been pregnant.
- This should impact the distribution of the data, and reduce variance in it.

Statistics Summary

Create a copy of the original dataset and replace the 0 values of the impacted columns with the mean values

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.000000	3.000000	6.00000	17.00
Glucose	768.0	121.681605	30.436016	44.000	99.750000	117.000000	140.25000	199.00
BloodPressure	768.0	72.254807	12.115932	24.000	64.000000	72.000000	80.00000	122.00
SkinThickness	768.0	26.606479	9.631241	7.000	20.536458	23.000000	32.00000	99.00
Insulin	768.0	118.660163	93.080358	14.000	79.799479	79.799479	127.25000	846.00
BMI	768.0	32.450805	6.875374	18.200	27.500000	32.000000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.243750	0.372500	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.000000	29.000000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.000000	0.000000	1.00000	1.00

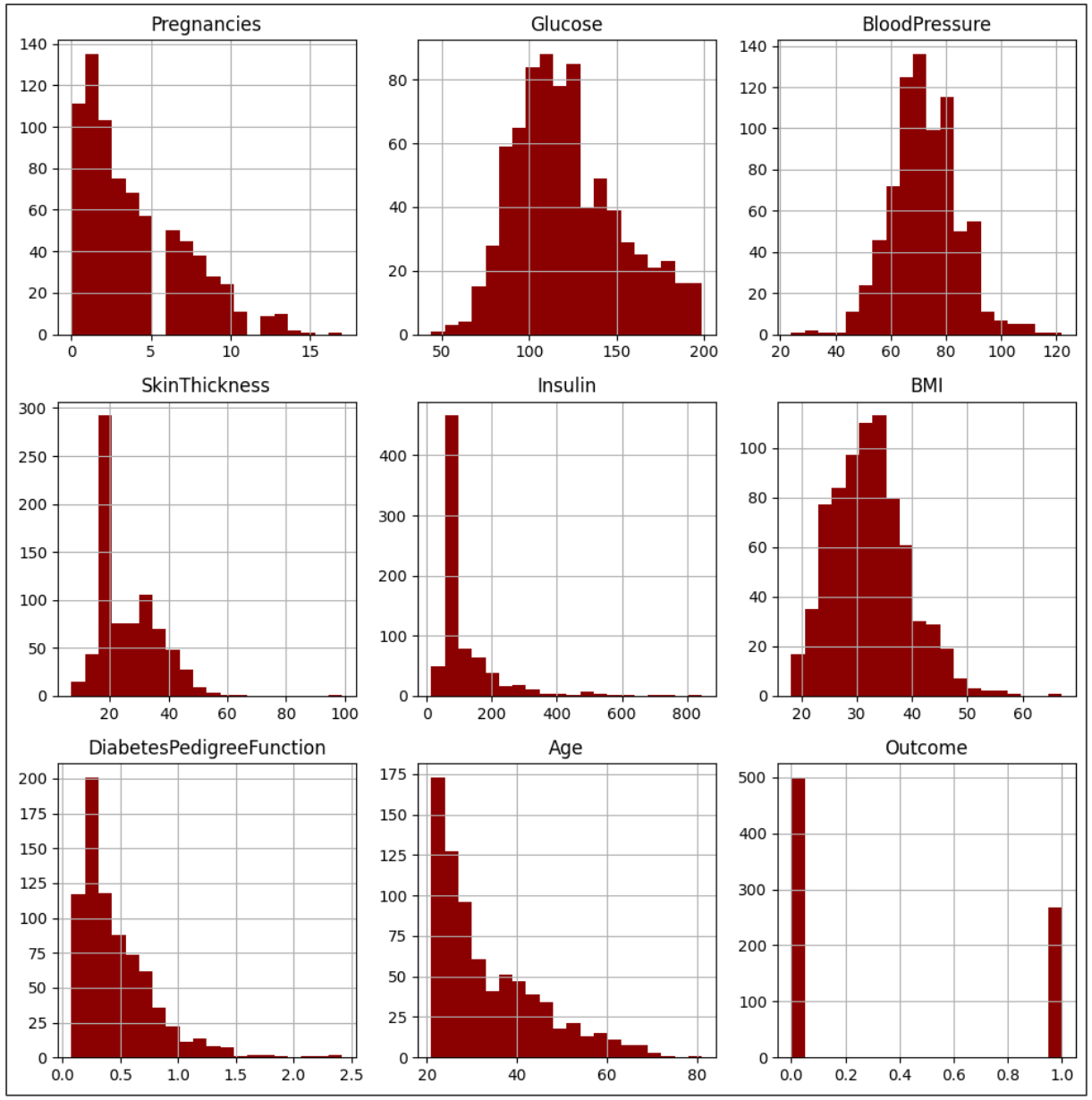
Now that the 0 values are accounted for, The rest of the Exploratory Data Analysis.

EDA Univariate Analysis

Analyzing/visualizing the dataset by taking one variable at a time:

- Data visualization is essential; we must decide what charts to plot to better understand the data. Visualize our data using Matplotlib and Seaborn libraries.
- **Matplotlib** is a Python 2D plotting library used to draw basic charts we use Matplotlib.
- **Seaborn** is also a python library built on top of Matplotlib that uses short lines of code to create and style statistical plots from Pandas and Numpy
- Univariate analysis can be done for both Categorical and Numerical variables.
- **Categorical variables** can be visualized using a Count plot, Bar Chart, Pie Plot, etc.
- **Numerical Variables** can be visualized using Histogram, Box Plot, Density Plot, etc.
- **In our example, we have done a Univariate analysis using Histogram and Box Plot for continuous Variables.**

EDA Univariate Analysis



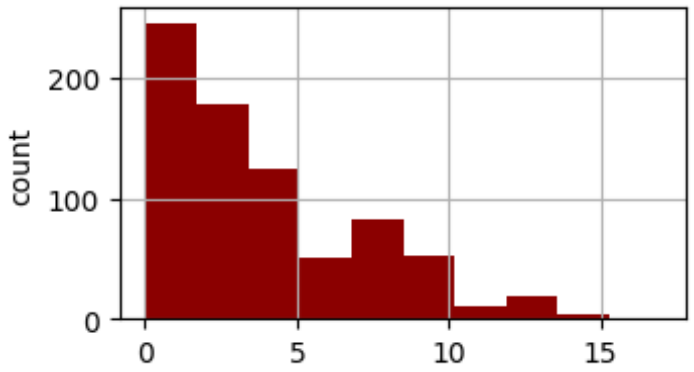
Histogram Plot of data

- The above histogram plots give a high-level view of the bucket distribution of the dataset parameters.
- At first glance, most of them appear to be positively skewed, with Glucose and Blood Pressure with the closest distribution to a normal distribution. Outcome is a bimodal distribution which is to be expected. There appear to have very few outliers for each of the parameters.
- From the histogram plot above also, there doesn't seem to have that many outliers.
- From the dataset description, only Insulin is the only parameter which has a very huge outlier but it should be fine to leave it in the dataset as it is.

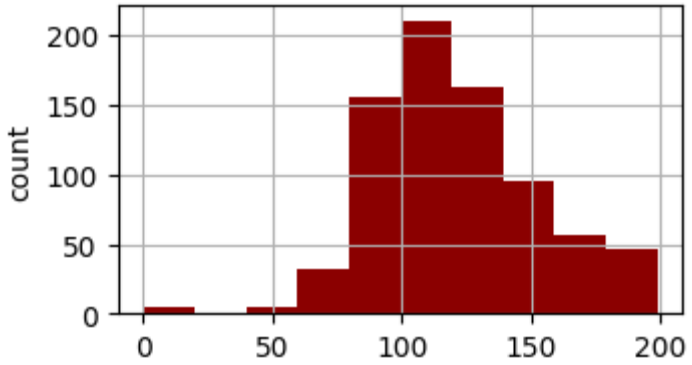
EDA Univariate Analysis

Histogram Plot of data

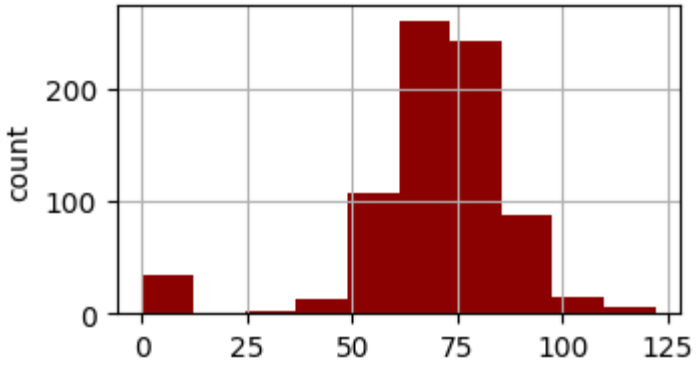
Pregnancies
Skew : 0.9



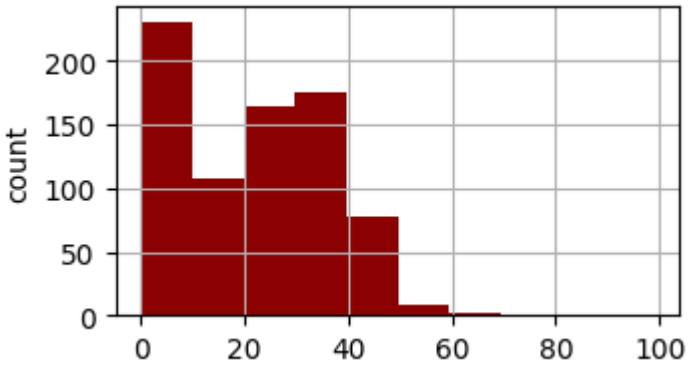
Glucose
Skew : 0.17



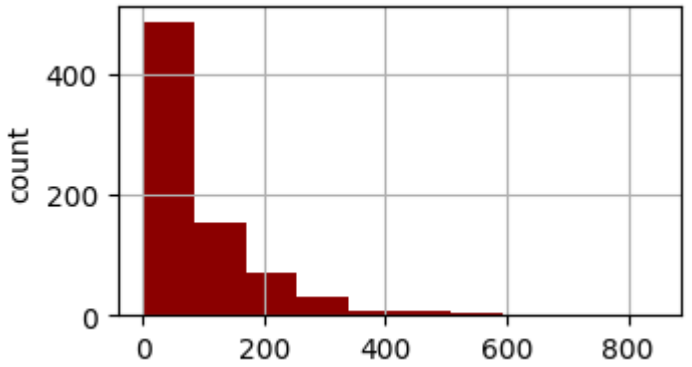
BloodPressure
Skew : -1.84



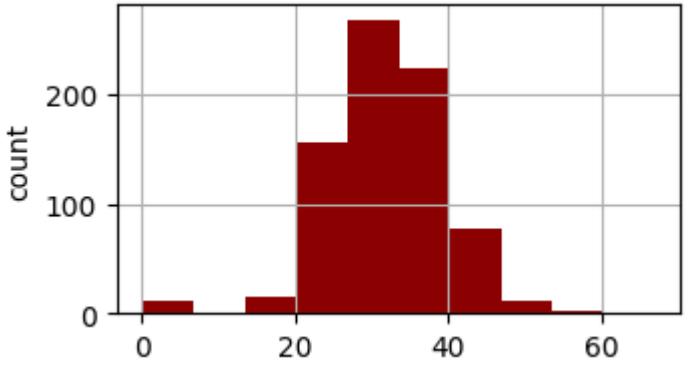
SkinThickness
Skew : 0.11



Insulin
Skew : 2.27



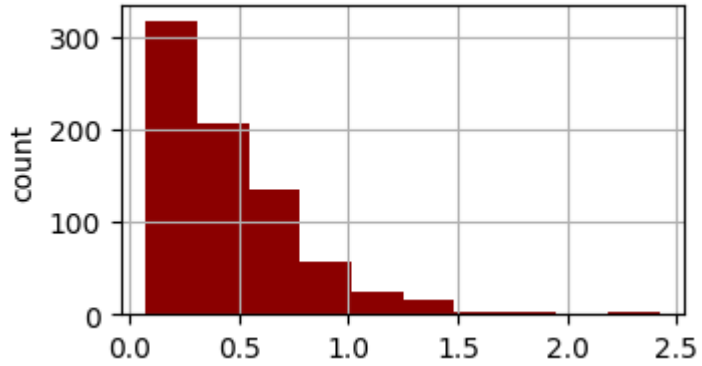
BMI
Skew : -0.43



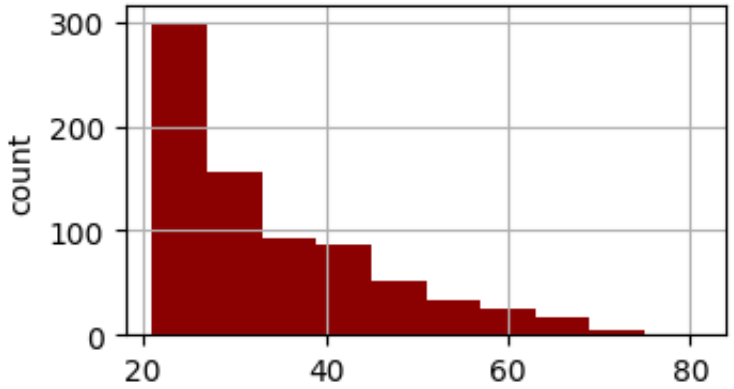
EDA Univariate Analysis

Histogram Plot of data

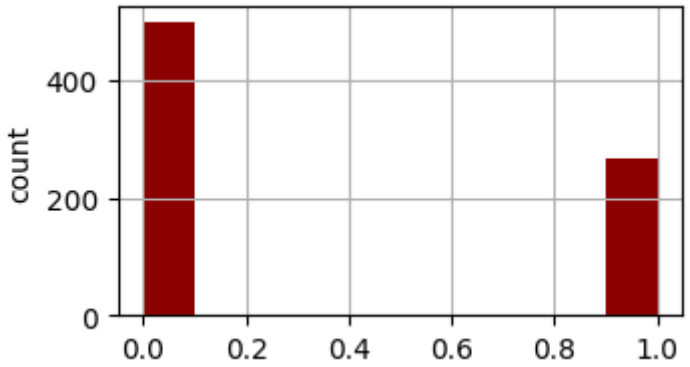
DiabetesPedigreeFunction
Skew : 1.92



Age
Skew : 1.13



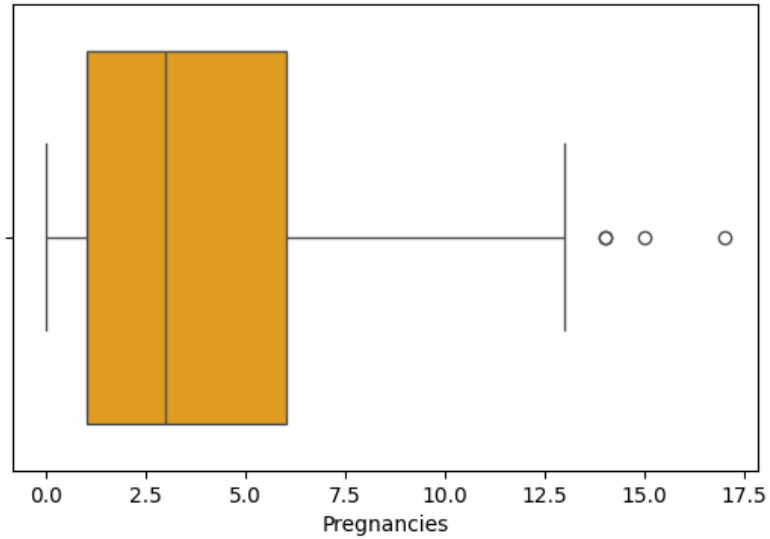
Outcome
Skew : 0.64



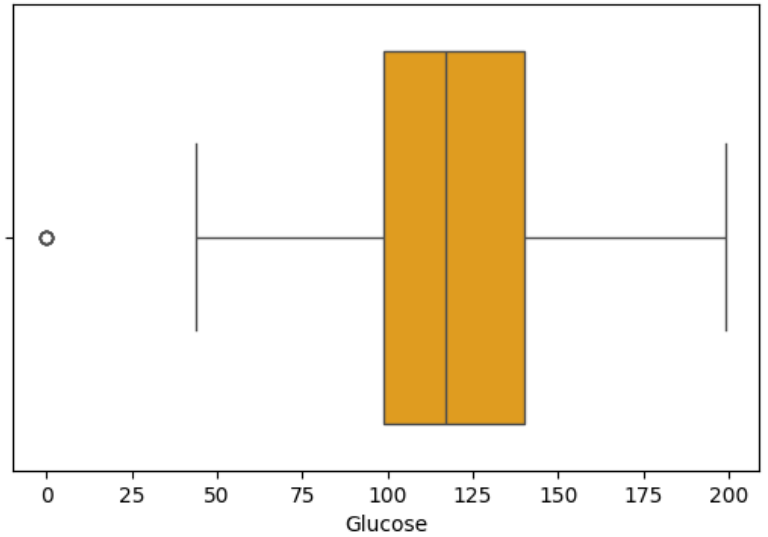
EDA Univariate Analysis

Box Plot of data

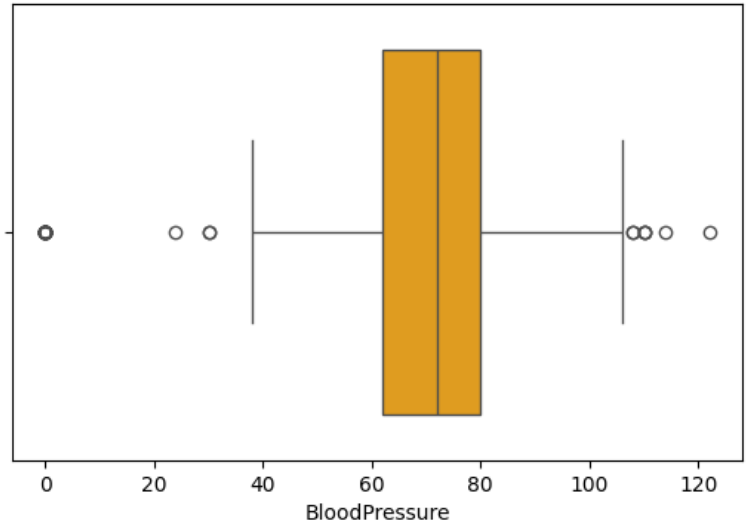
Pregnancies
Skew : 0.9



Glucose
Skew : 0.17



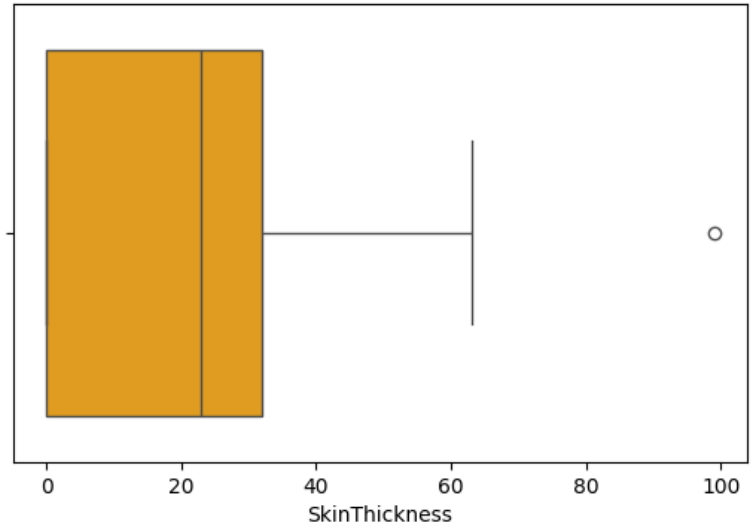
BloodPressure
Skew : -1.84



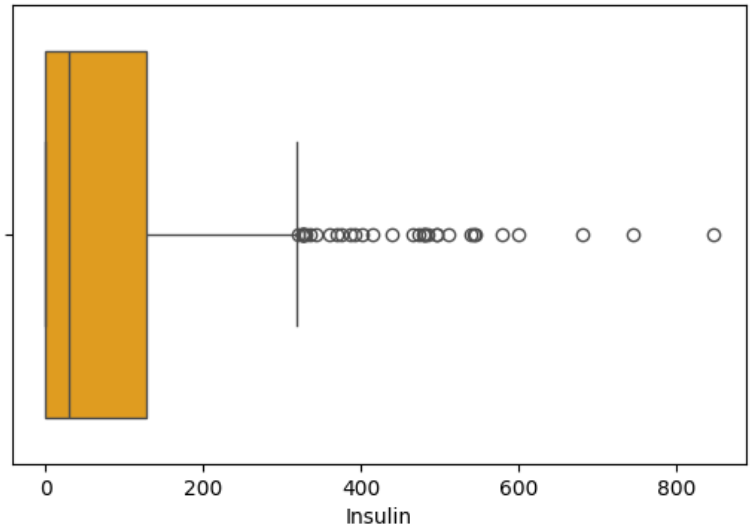
EDA Univariate Analysis

Box Plot of data

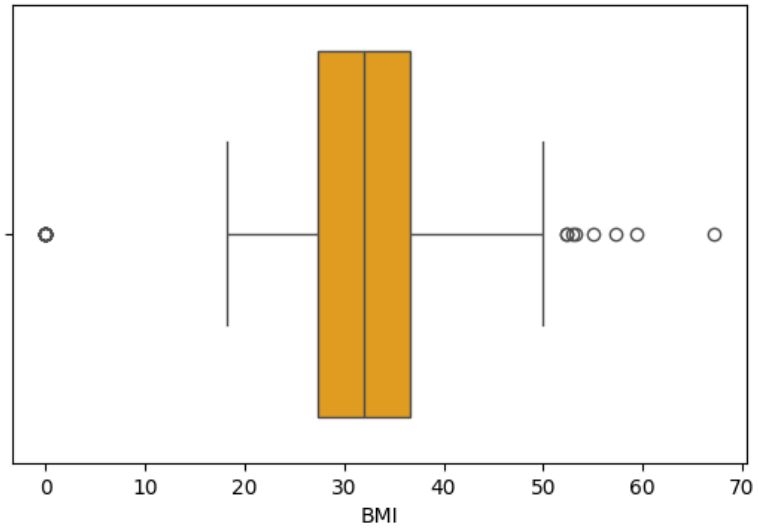
SkinThickness
Skew : 0.11



Insulin
Skew : 2.27



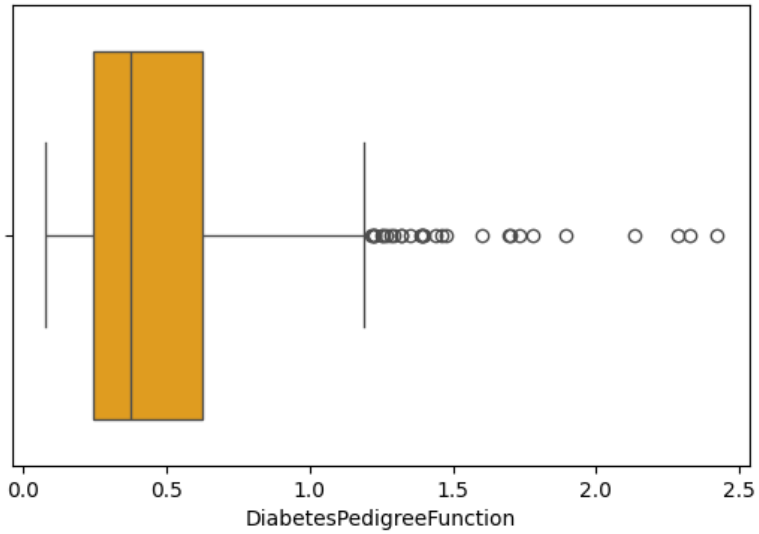
BMI
Skew : -0.43



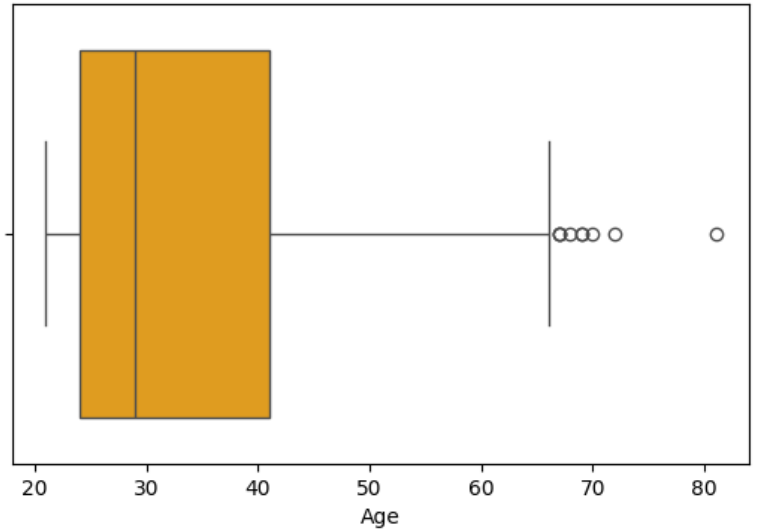
EDA Univariate Analysis

Box Plot of data

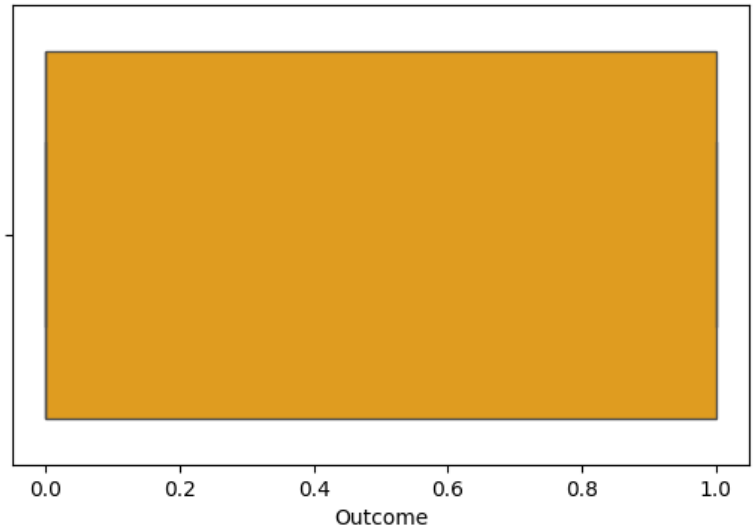
DiabetesPedigreeFunction
Skew : 1.92



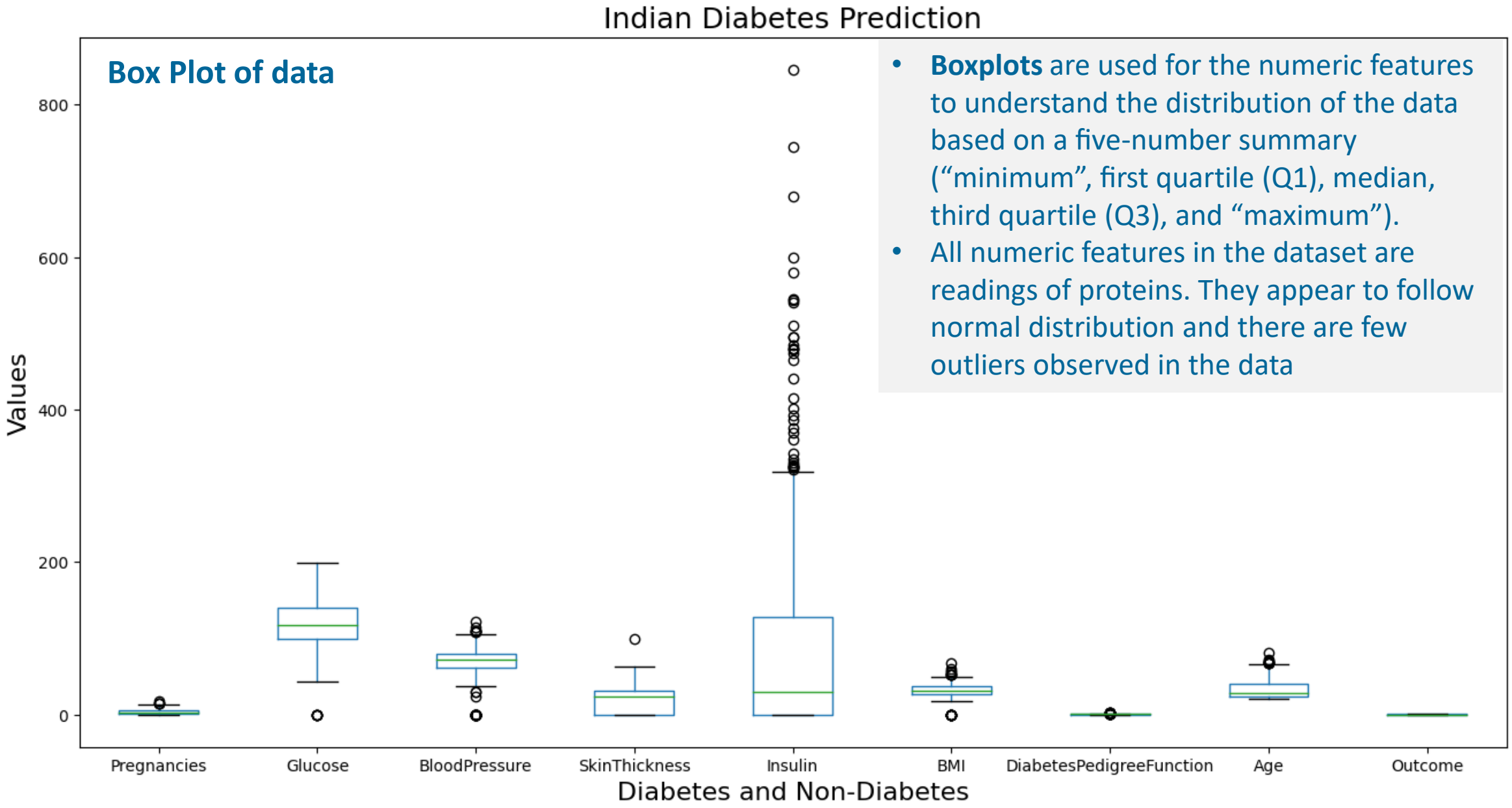
Age
Skew : 1.13



Outcome
Skew : 0.64



Box Plot of data



- **Boxplots** are used for the numeric features to understand the distribution of the data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").
- All numeric features in the dataset are readings of proteins. They appear to follow normal distribution and there are few outliers observed in the data

EDA Bivariate Analysis

- `Bivariate Analysis` helps to understand how variables are related to each other and the relationship between dependent and independent variables present in the dataset.
- For `Numerical variables`, Pair plots and Scatter plots are widely been used to do Bivariate Analysis
- A Stacked bar chart can be used for `categorical variables` if the output variable is a classifier. Bar plots can be used if the output variable is continuous

Correlation of data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.127964	0.208984	0.013376	-0.018082	0.021546	-0.033523	0.544341	0.221898
Glucose	0.127964	1.000000	0.219666	0.160766	0.396597	0.231478	0.137106	0.266600	0.492908
BloodPressure	0.208984	0.219666	1.000000	0.134155	0.010926	0.281231	0.000371	0.326740	0.162986
SkinThickness	0.013376	0.160766	0.134155	1.000000	0.240361	0.535703	0.154961	0.026423	0.175026
Insulin	-0.018082	0.396597	0.010926	0.240361	1.000000	0.189856	0.157806	0.038652	0.179185
BMI	0.021546	0.231478	0.281231	0.535703	0.189856	1.000000	0.153508	0.025748	0.312254
DiabetesPedigreeFunction	-0.033523	0.137106	0.000371	0.154961	0.157806	0.153508	1.000000	0.033561	0.173844
Age	0.544341	0.266600	0.326740	0.026423	0.038652	0.025748	0.033561	1.000000	0.238356
Outcome	0.221898	0.492908	0.162986	0.175026	0.179185	0.312254	0.173844	0.238356	1.000000

- From the above correlation matrix, we can see that there doesn't appear to have any parameters which have very strong correlation to each other.
- The parameter with the highest positive correlation to each other is BMI and SkinThicknes
- Glucose and BloodPressure are the only parameters which most resemble a normal distribution.

EDA Multivariate Analysis

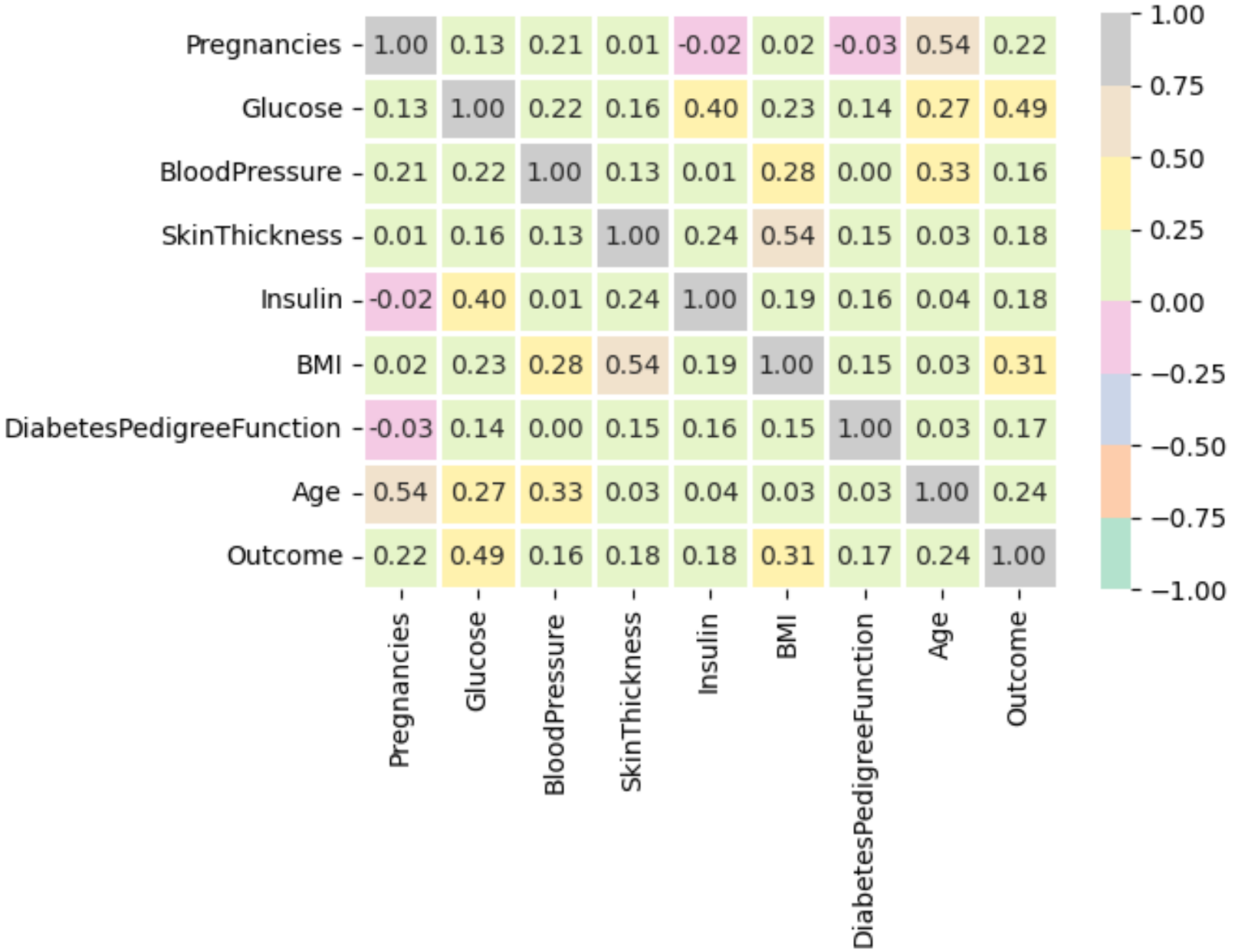
Multivariate analysis looks at more than two variables. Multivariate analysis is one of the most useful methods to determine relationships and analyze patterns for any dataset.

A heat map is widely been used for Multivariate Analysis

Heat Map gives the correlation between the variables, whether it has a positive or negative correlation.

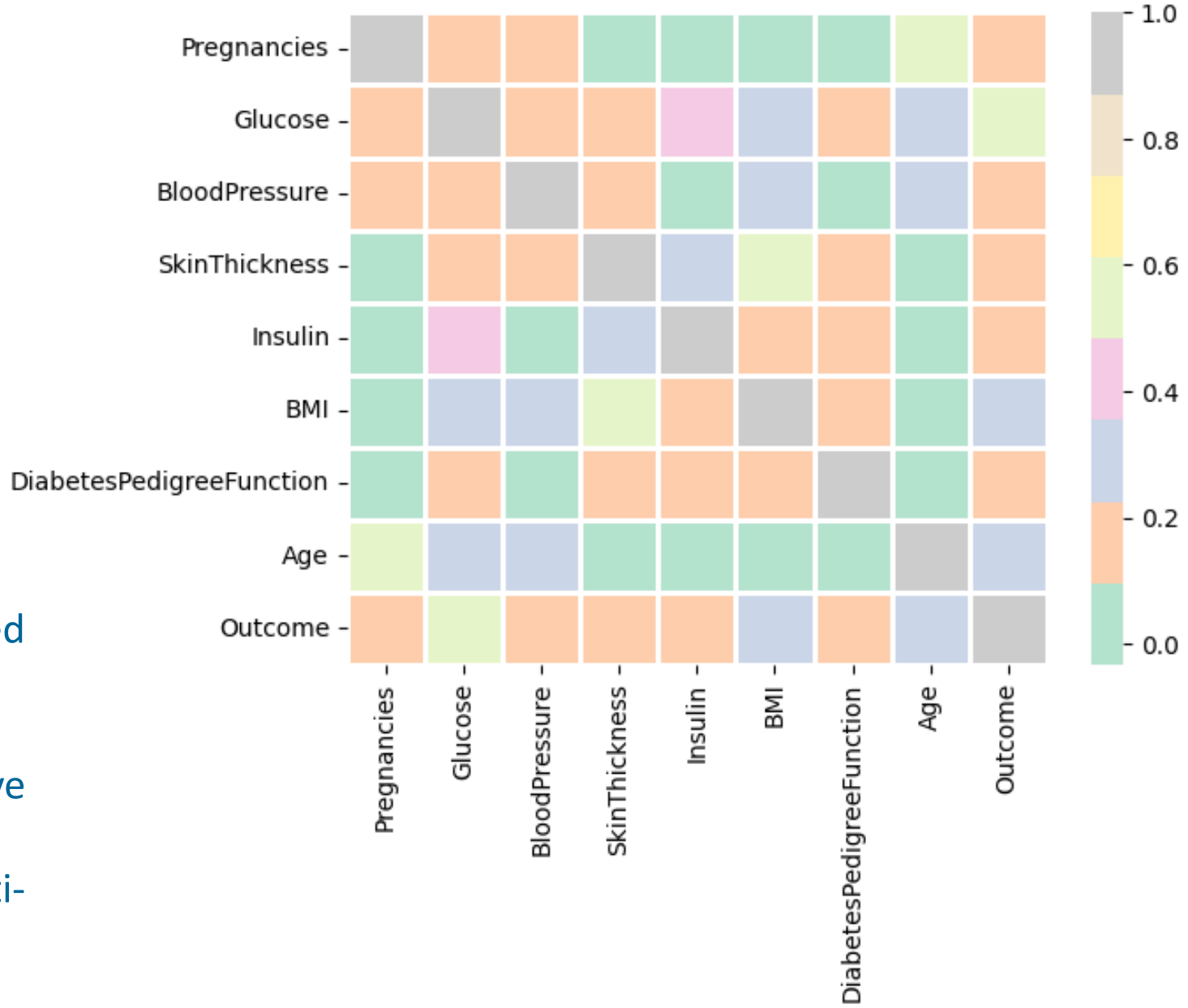
In our example heat map shows the correlation between the variables.

Heatmap


Pregnancies 1.00 0.13 0.21 0.01 -0.02 0.02 -0.03 0.54 0.22
Glucose 0.13 1.00 0.22 0.16 0.40 0.23 0.14 0.27 0.49
BloodPressure 0.21 0.22 1.00 0.13 0.01 0.28 0.00 0.33 0.16
SkinThickness 0.01 0.16 0.13 1.00 0.24 0.54 0.15 0.03 0.18
Insulin -0.02 0.40 0.01 0.24 1.00 0.19 0.16 0.04 0.18
BMI 0.02 0.23 0.28 0.54 0.19 1.00 0.15 0.03 0.31
DiabetesPedigreeFunction -0.03 0.14 0.00 0.15 0.16 0.15 1.00 0.03 0.17
Age 0.54 0.27 0.33 0.03 0.04 0.03 0.03 1.00 0.24
Outcome 0.22 0.49 0.16 0.18 0.18 0.31 0.17 0.24 1.00
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Correlation Matrix

```
Pearsons correlation:
PearsonRResult(statistic=0.544
341284023389,
pvalue=1.862812832863466e-60)
```

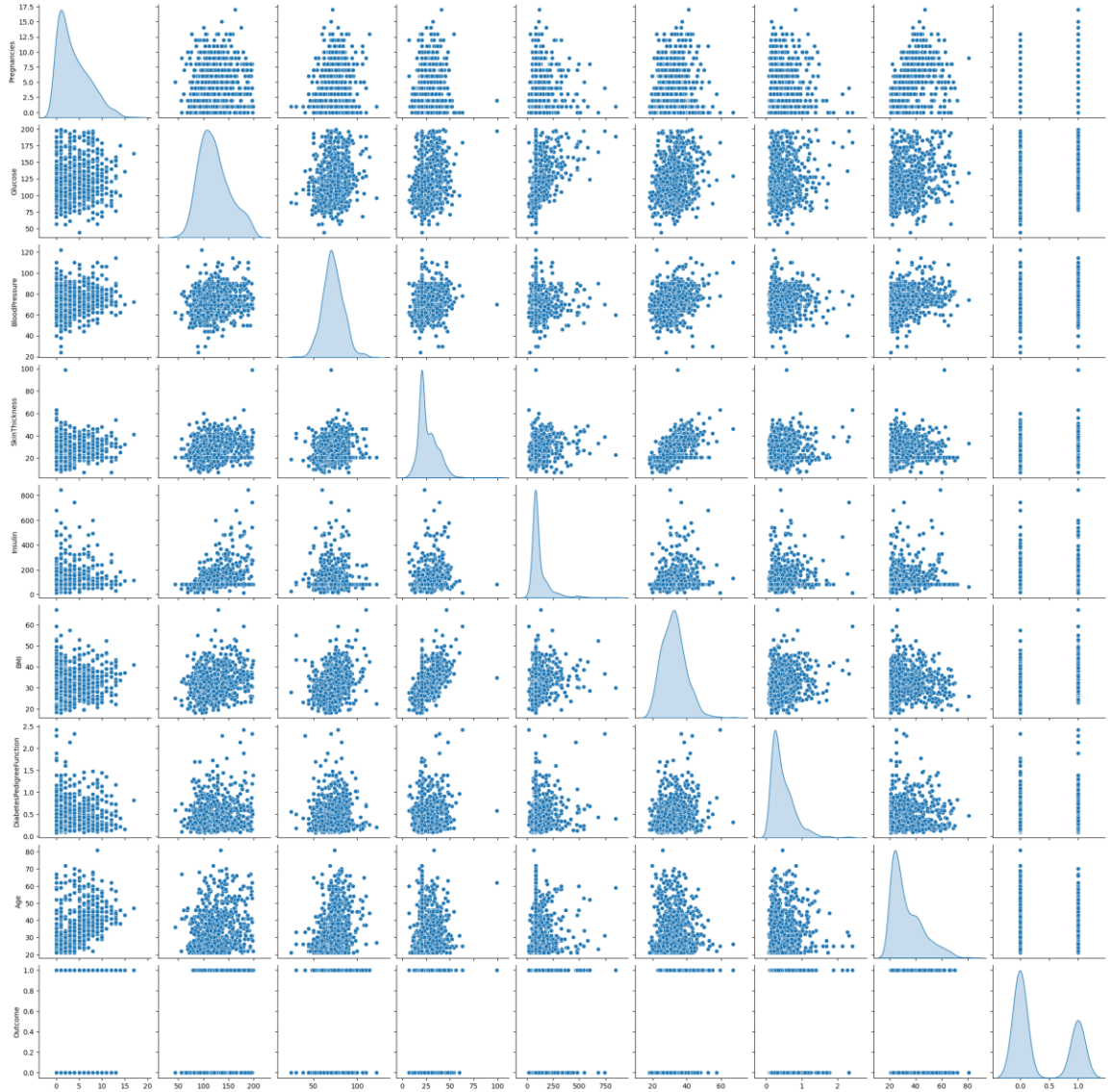


From the figure,

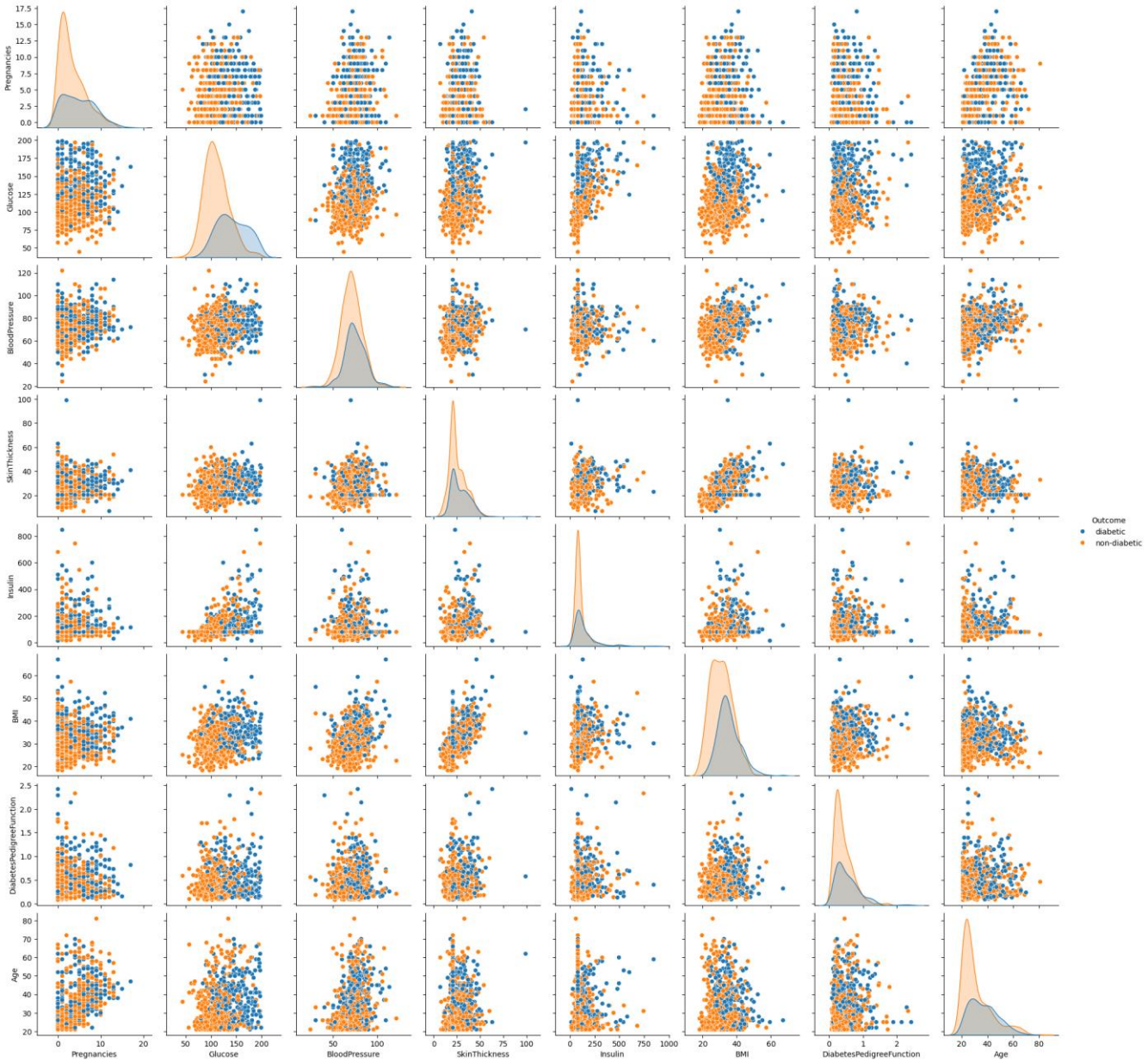
- A significant correlation can be observed between Pregnancies and Age.
- The correlation coefficient (r) is 0.544.
- By a rule of thumb, in case of an r above 0.70, multi-collinearity is expected.
- Hence, No significant case of multi-collinearity is observed.

pairplot

Glucose and **BloodPressure** are the only parameters which most resemble a normal distribution.



pairplot



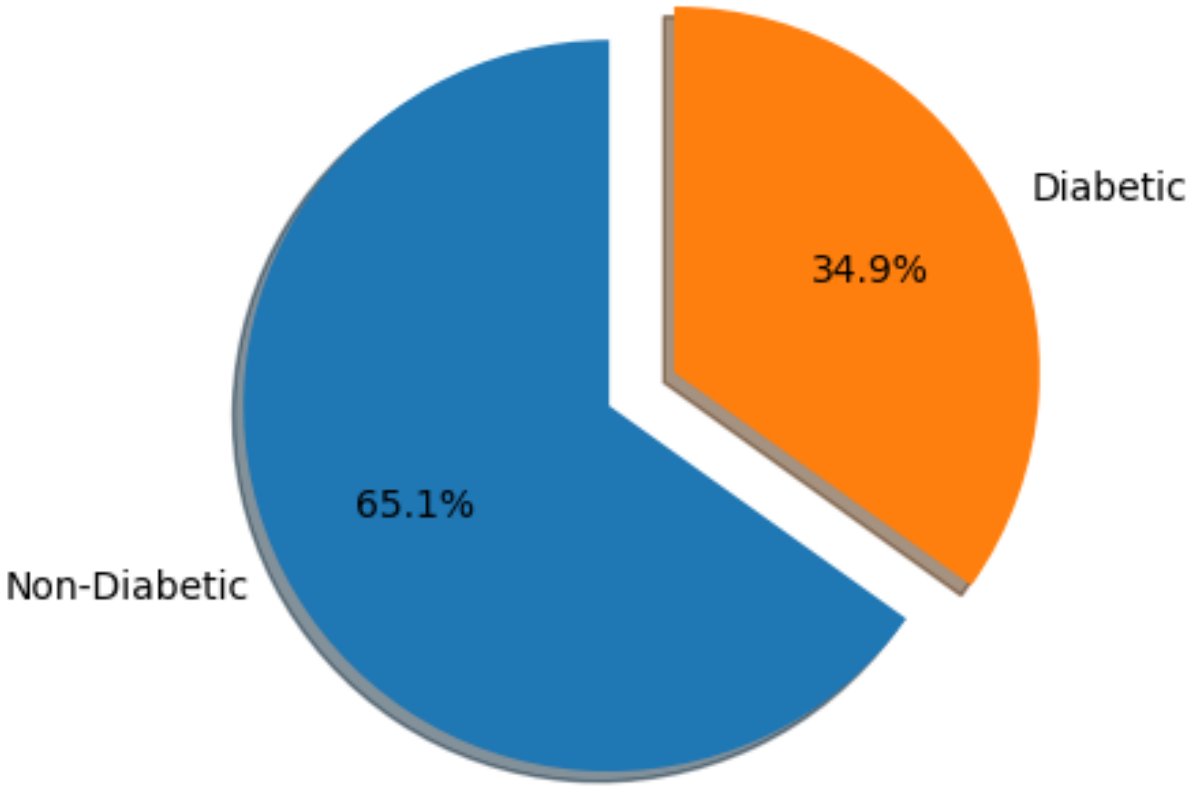
To plot a pairplot to see which parameters might have a stronger correlation with either outcomes of diabetic patient and non-diabetic patient.

- we cannot really see a strong correlation between the parameters and the outcome, with the exception of Glucose, where the higher the glucose level, the greater the chance of a diabetic patient.

EDA for Categorical Data

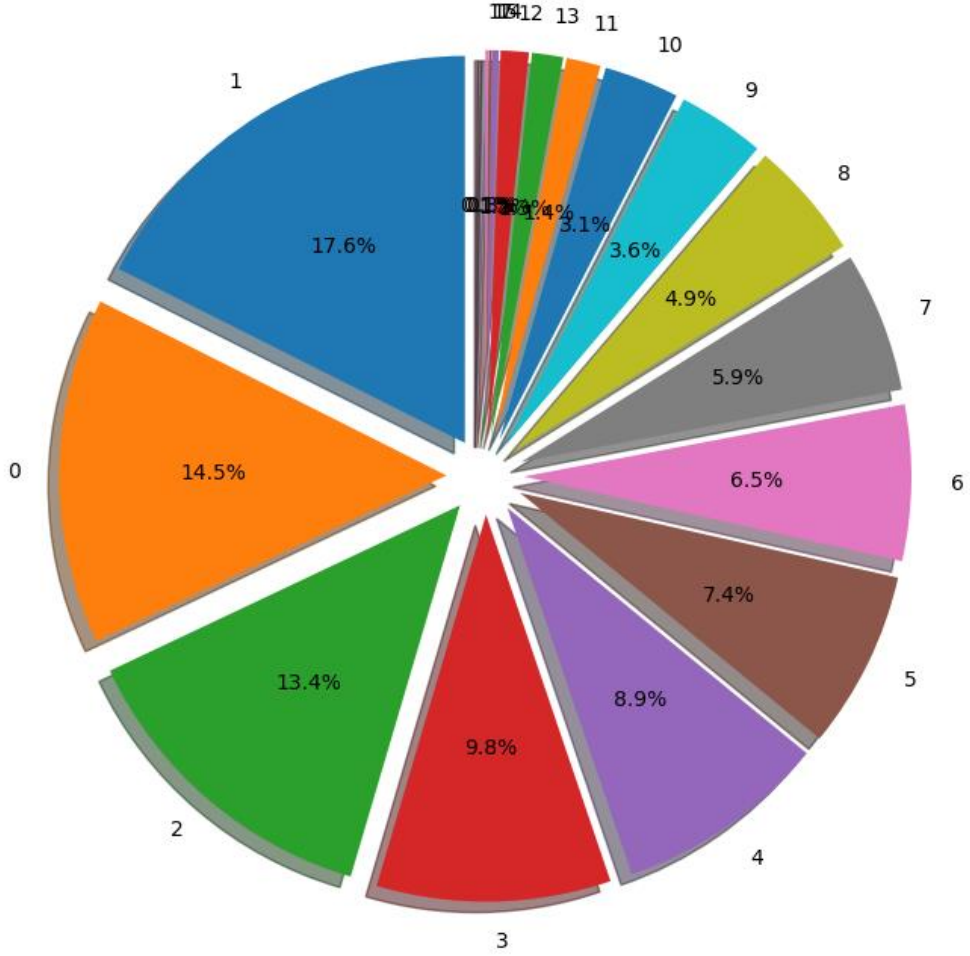
Pie chart :

Proportion of people having diabetes displayed in pie chart.

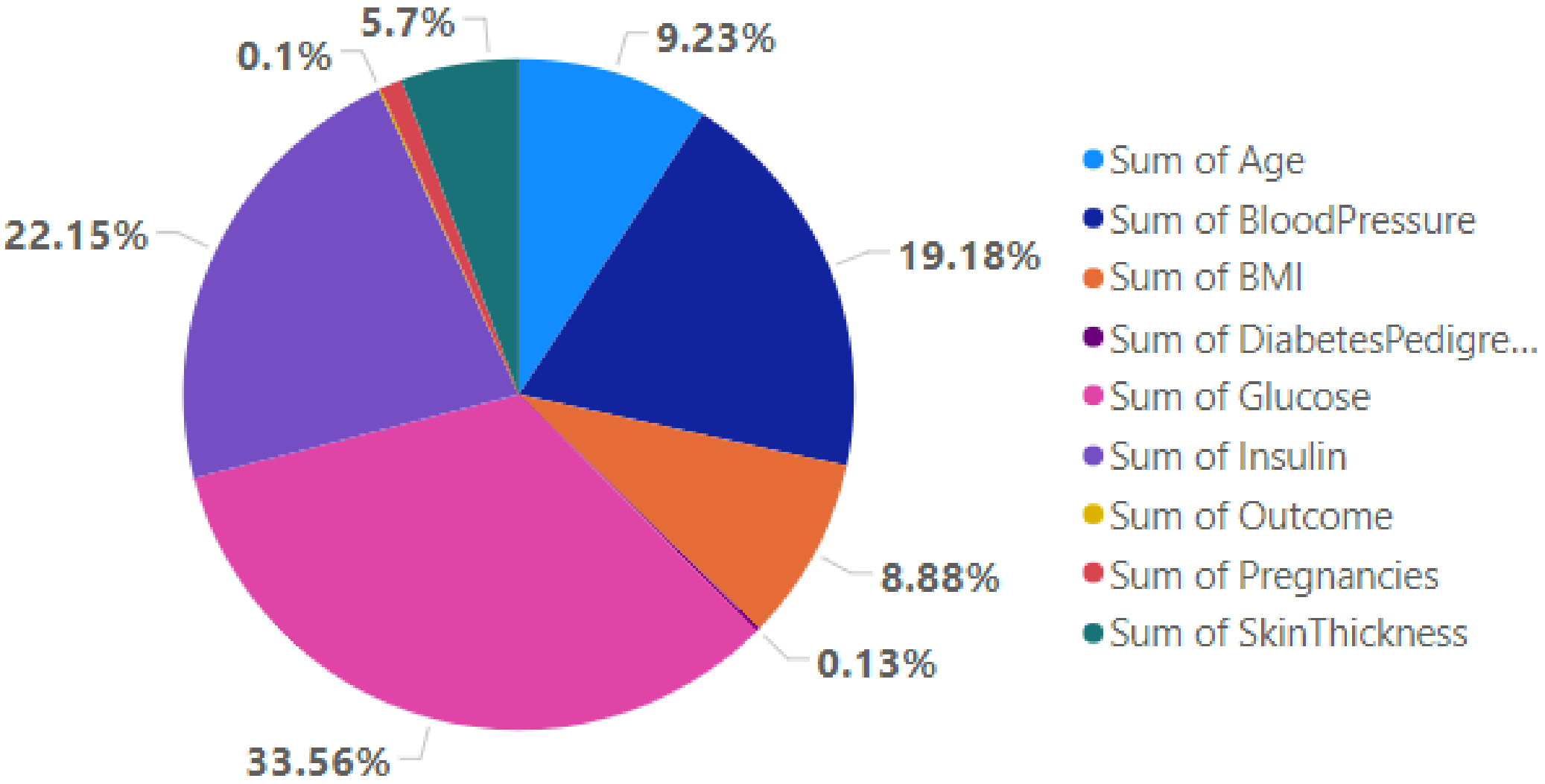


Pie Chart of Pregnancy :

Proportion of people having diabetes displayed in pie chart



EDA for Categorical Data



EDA for Numerical Data

Group by the data to look for the `average values` of the Diabetic and Non Diabetic Patients

	Non-diabetic	Diabetic
Pregnancies	3.298000	4.865672
Glucose	109.980000	141.257463
BloodPressure	68.184000	70.824627
SkinThickness	19.664000	22.164179
Insulin	68.792000	100.335821
BMI	30.304200	35.142537
DiabetesPedigreeFunction	0.429734	0.550500
Age	31.190000	37.067164

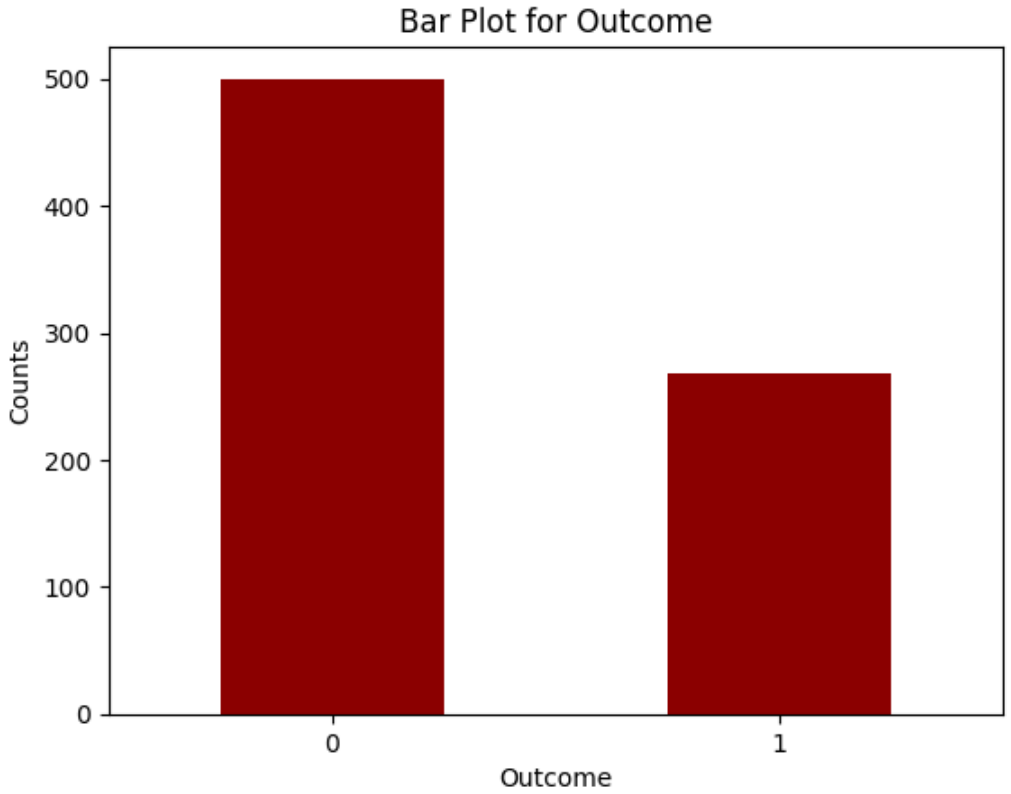
EDA for Numerical Data

Group by the data to look for the `Standard Deviation` of the Diabetic and Non Diabetic Patients

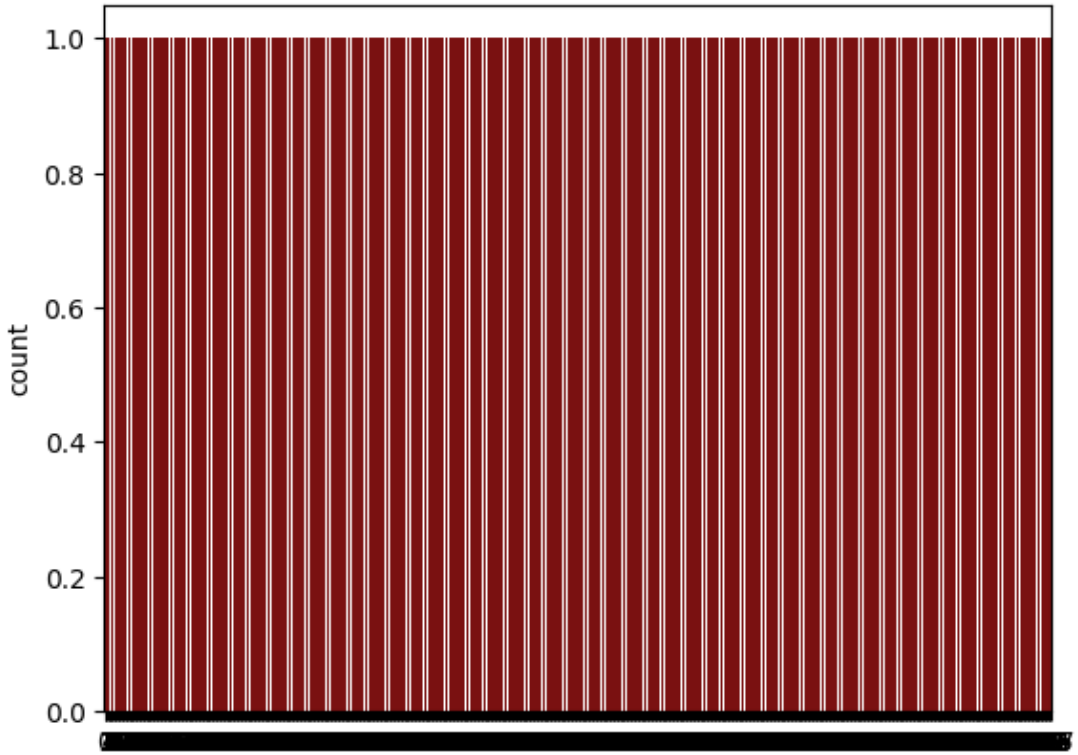
	Non-diabetic	Diabetic
Pregnancies	3.017185	3.741239
Glucose	26.141200	31.939622
BloodPressure	18.063075	21.491812
SkinThickness	14.889947	17.679711
Insulin	98.865289	138.689125
BMI	7.689855	7.262967
DiabetesPedigreeFunction	0.299085	0.372354
Age	11.667655	10.968254

EDA for Numerical Data

Bar Plot of Outcome

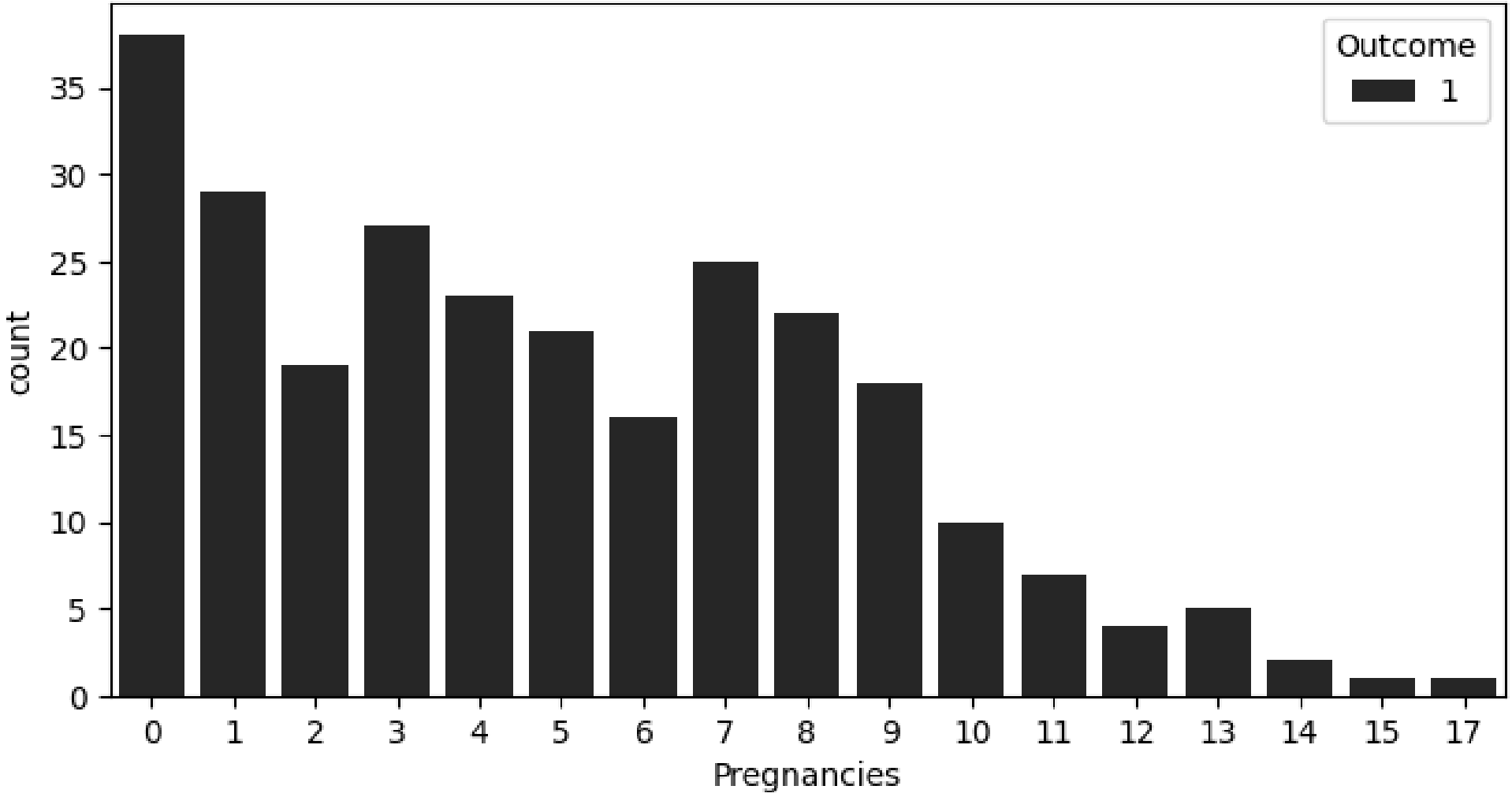


Bar Plot of Pregnancies

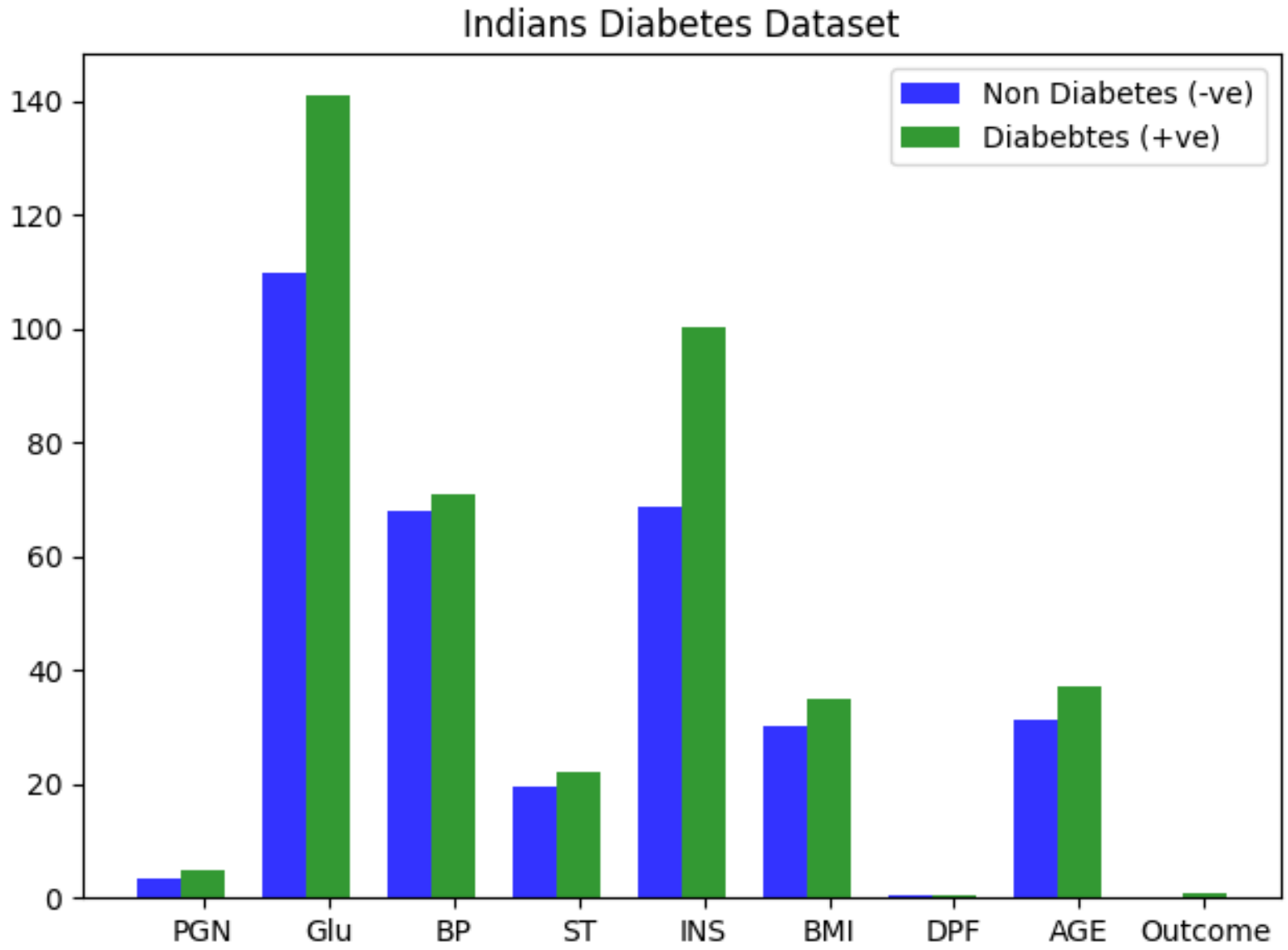


EDA for Numerical Data

The distribution of Diabetic Patients according to Pregnancy



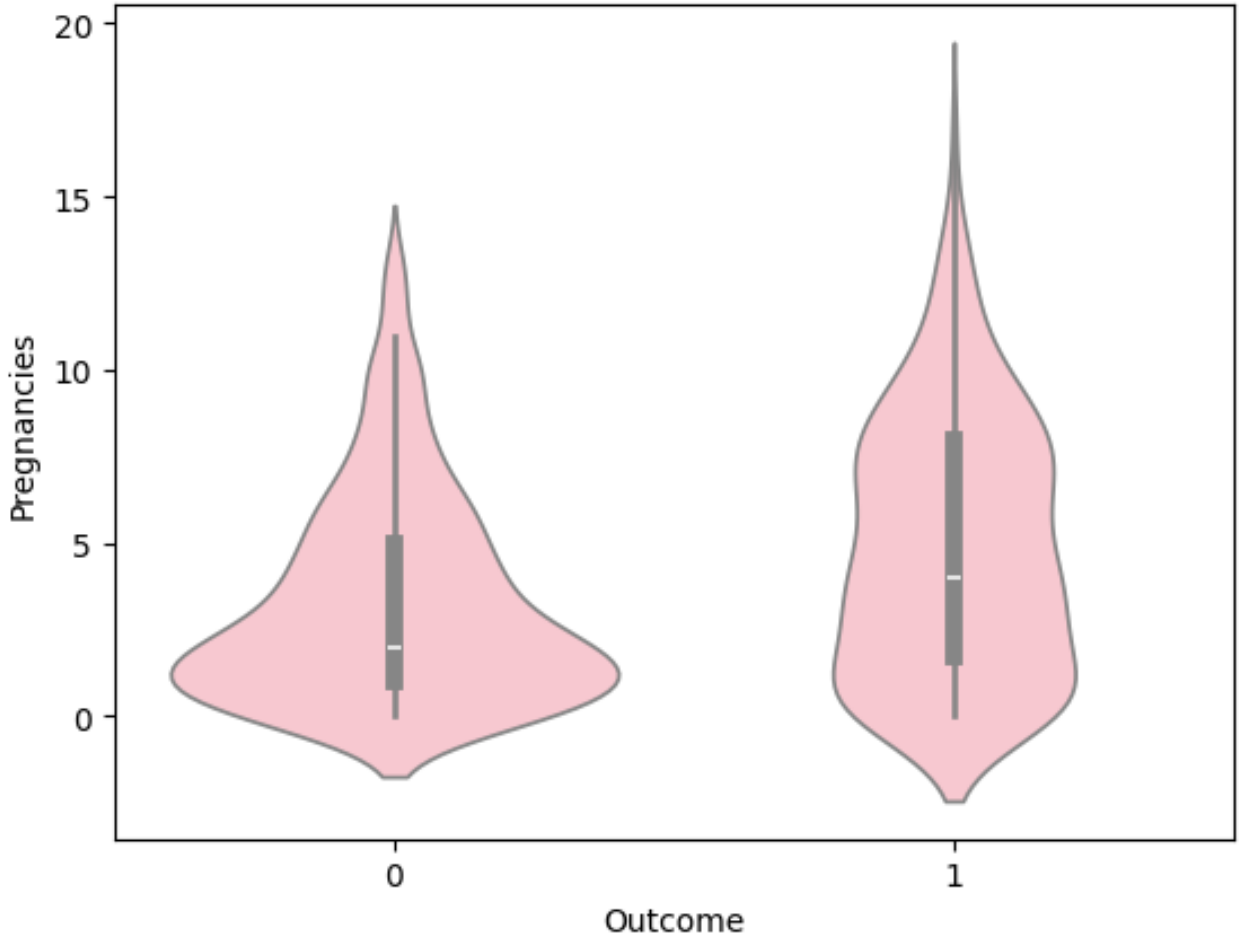
A Multiple Bar Plot can be used to **show the relationship between Categorical variables and continuous variables**



EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

Classifying the `Pregnancies based on Outcome`

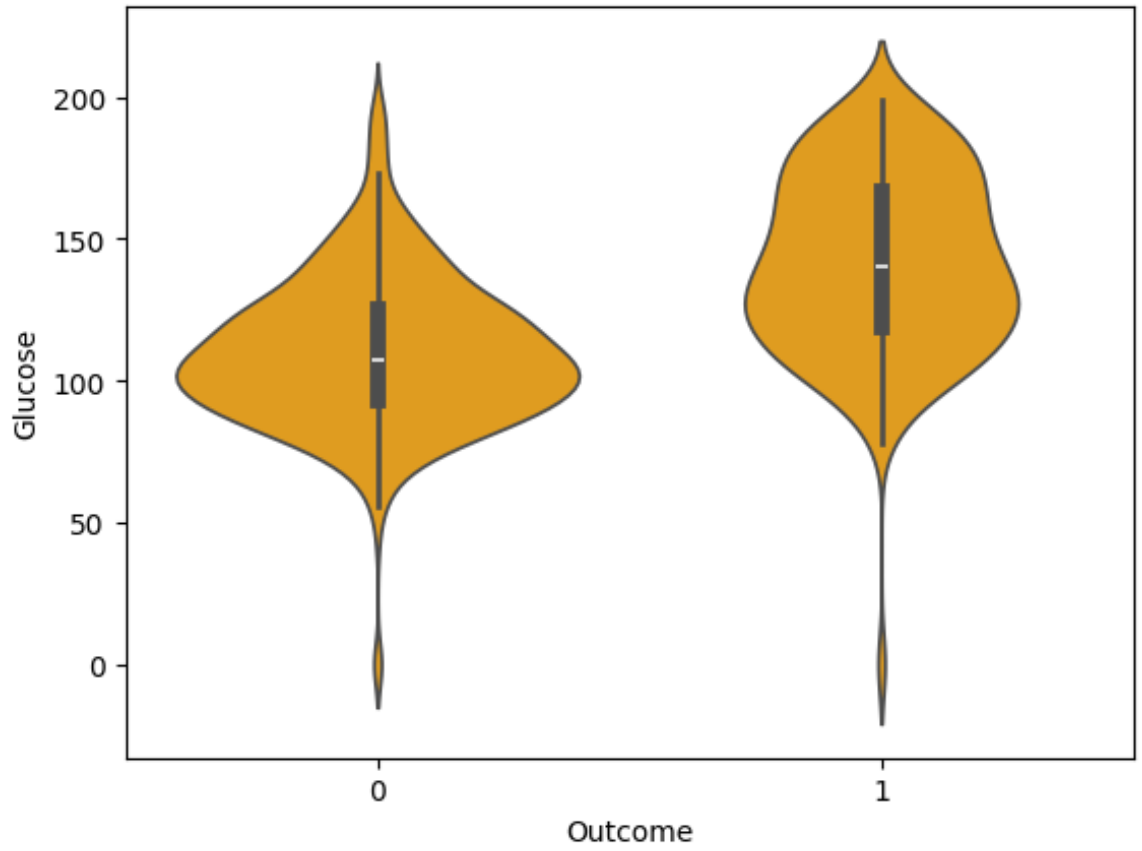


we observe diabetic women had more pregnancies than non-diabetic

EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

Classifying the `Glucose based on Outcome`

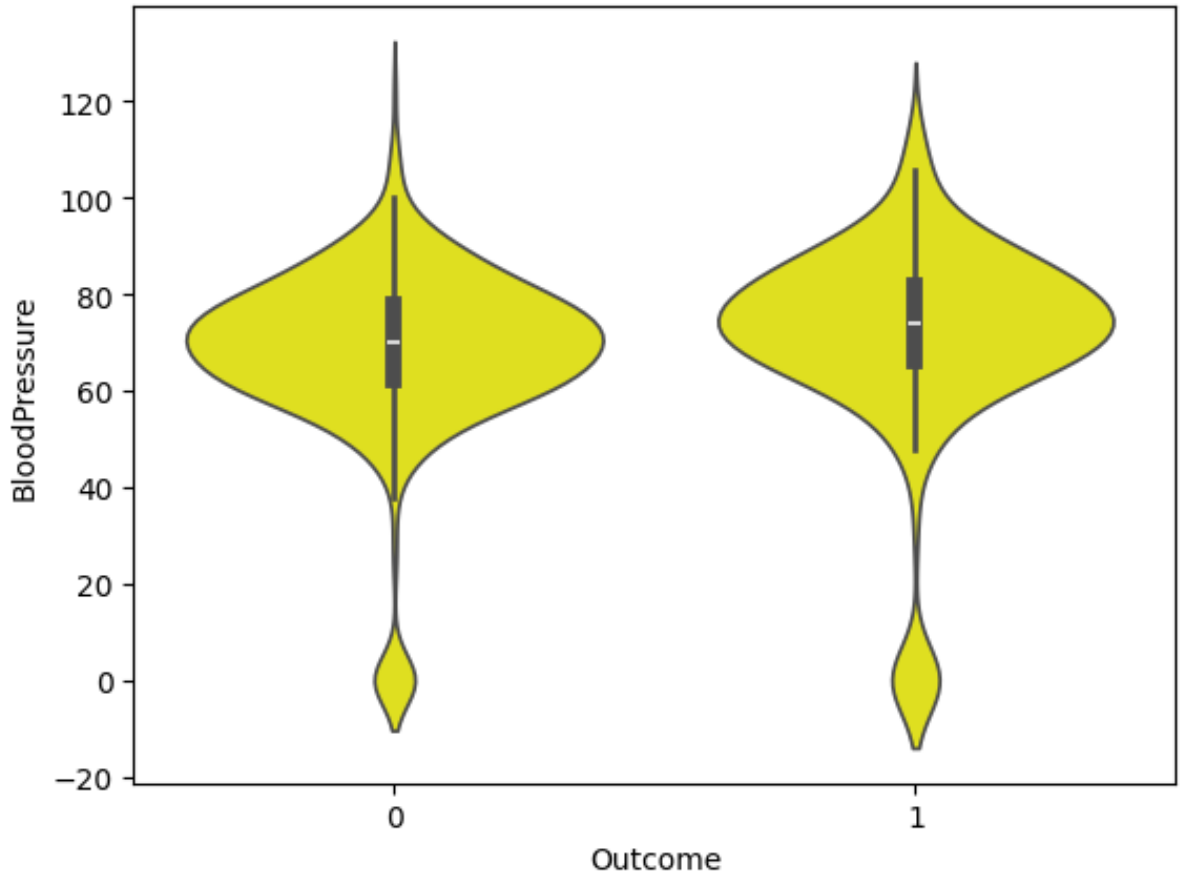


- We see a massive vertical distance between the box-plot for Diabetics and Non-Diabetics.
- This indicates that Glucose can be a very important variable in model-building

EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

Classifying the `Blood Pressure based on Outcome`



- The bottom tail of the violins indicates the zero values we need to replace.
- We will replace the zeros for 1 with median of 1 and same for 0.

EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

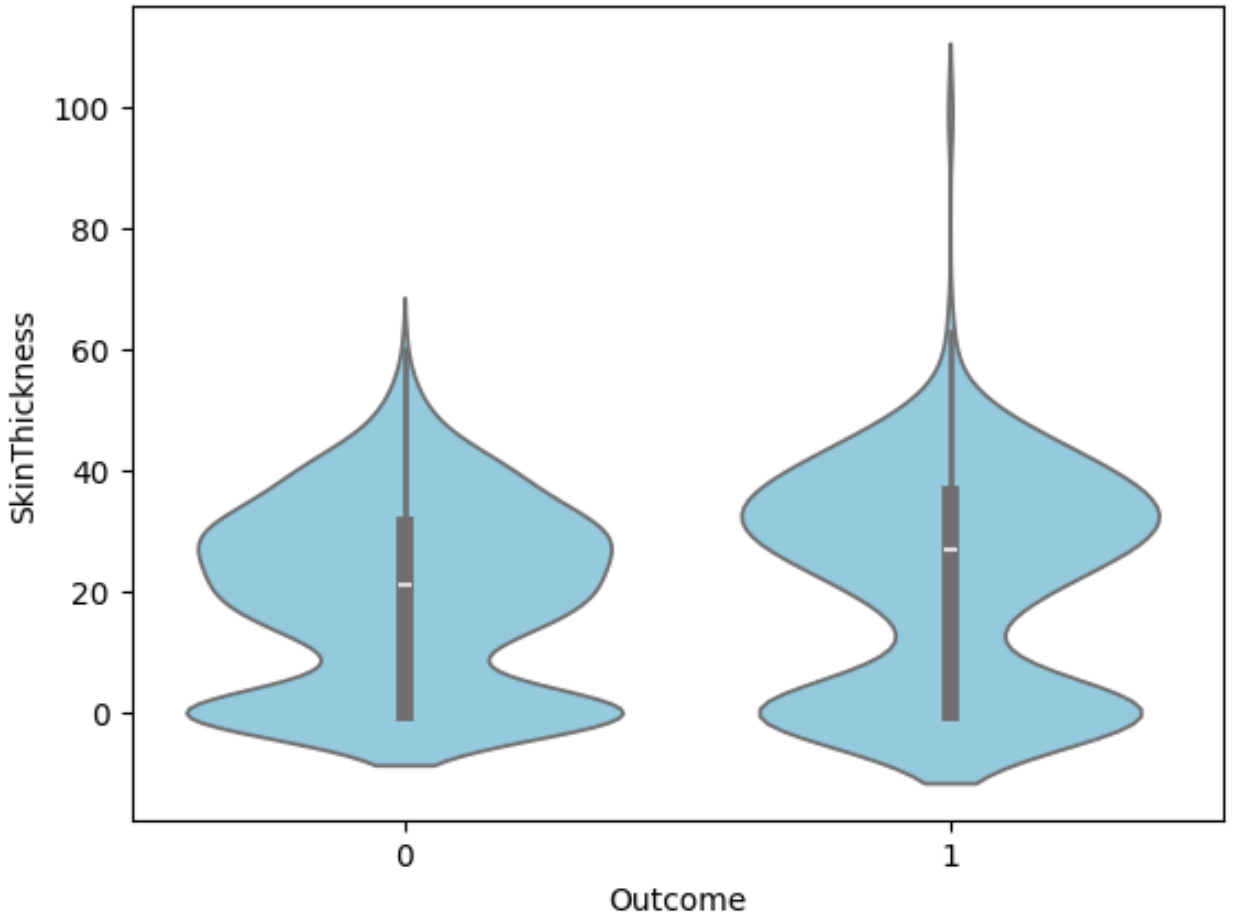
Replacing the `zero-values for Blood Pressure`

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
2	8	183	64	0	0	23.3	0.672	32	1
4	0	137	40	35	168	43.1	2.288	33	1
6	3	78	50	32	88	31.0	0.248	26	1
8	2	197	70	45	543	30.5	0.158	53	1

EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

Classifying the `Skin Thickness based on Outcome`

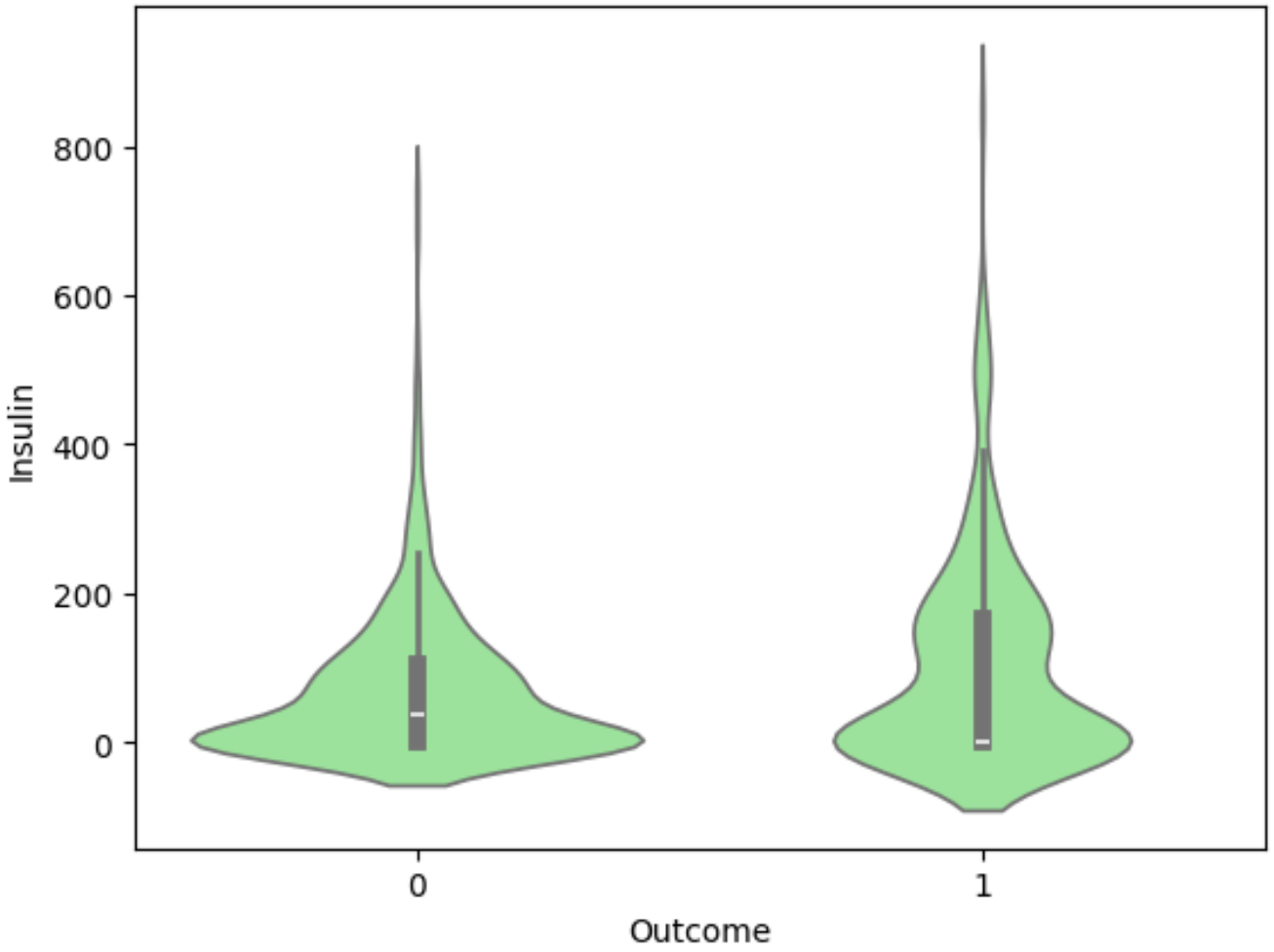


Skin Thickness for Diabetics is more than that of Non-Diabetics.

EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

Classifying the `Insulin based on Outcome`

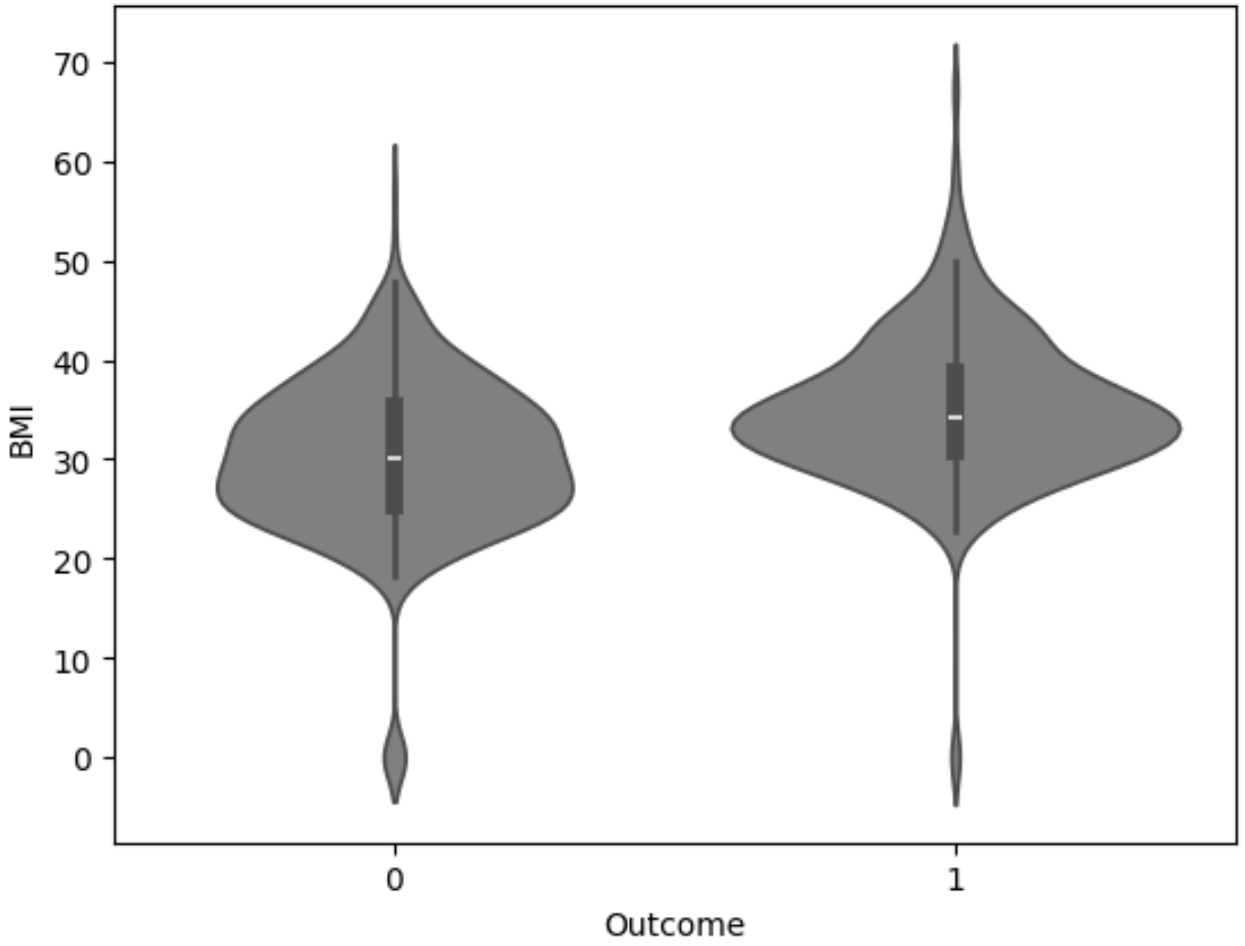


Insulin is a little higher. It can be roughly hypothesized that Insulin for Diabetics is lower than Non-Diabetics

EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

Classifying the `BMI based on Outcome`

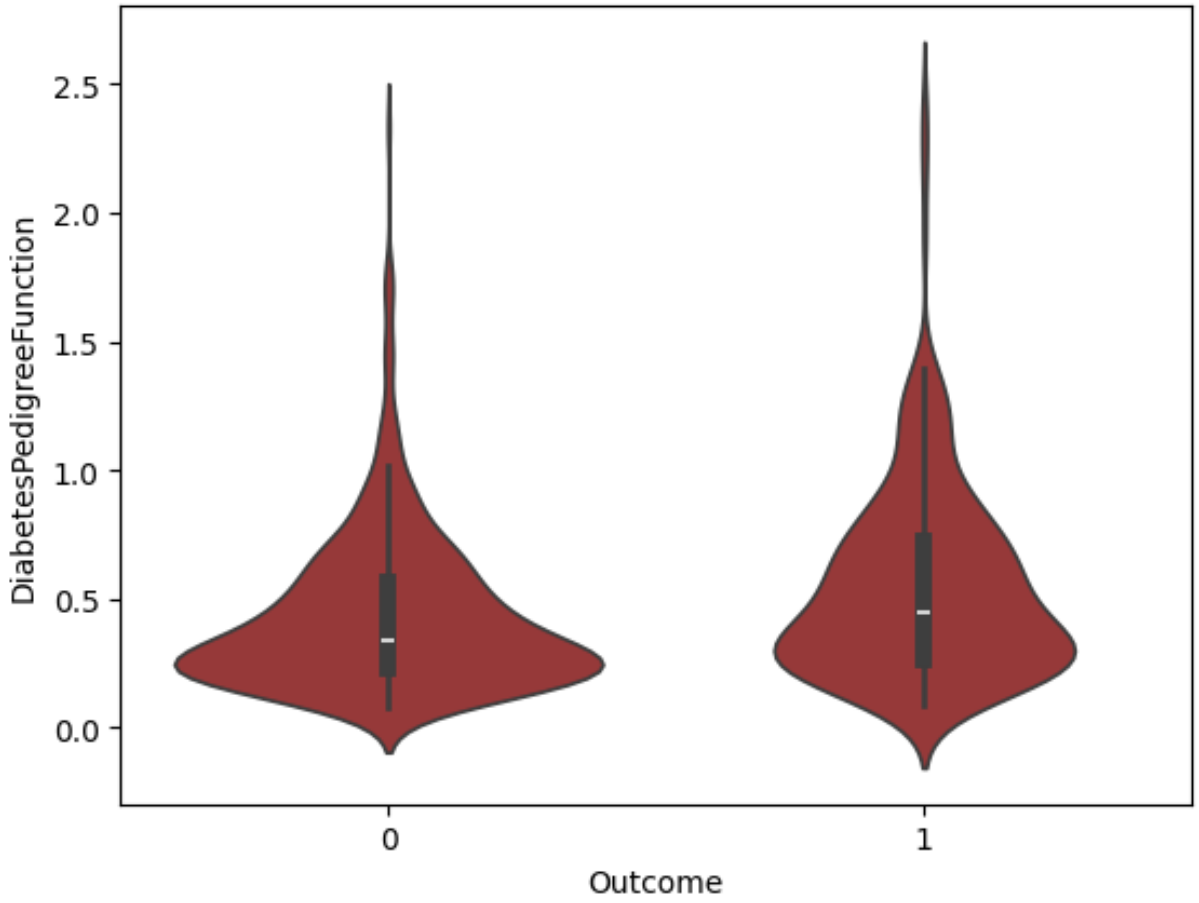


Observed using box plot , BMI for diabetics is more than BMI for non-diabetics .

EDA for Draw a Violin Graph

The outliers or anomalies in the data based Diabetic and Non Diabetic patients.

Classifying the `Diabetes Pedigree Function` based on Outcome



Diabetics seem to have a higher pedigree function that the non-diabetics.

Hypothesis Testing

First hypothesis test : `if Glucose data has a normal distribution.`

Next hypothesis test would be that based on the above hypothesis test

To test the correlation between Glucose and the target outcome.

First Hypothesis Test:

`To Test :`

`Null Hypothesis:` The sample comes from a normal distribution.

`Alternative Hypothesis:` The sample does not come from a normal distribution

```
Statistics = 35.373,
p_value =0.000
As p-value is 0.0 , which is lower than the significance level,

we reject the null hypothesis.
```

From the above test result, **we reject Glucose having a normal distribution.**

Hypothesis Testing

Second Hypothesis Test :

To perform a Pearson correlation test using stats.pearsonr

`Null Hypothesis:` Both sets of data are uncorrelated.

`Alternative Hypothesis:` Both sets of data are some what correlated

```
Statistics = 0.493,
p_value =0.000
As p-value is 0.0 , which is lower than the significance level,
we reject the null hypothesis.

The correlation coefficient between Glucose and the Target Variable is: 0.4929
```

Analysis of Variance (ANOVA)

Test if the treatments are significantly different and print the ANOVA table.

To Test:

Null Hypothesis: H0: There is no statistically significant for each experimental(Treatment) group.

Alternative Hypothesis: H1: There is a statistically significant for each experimental (Treatment) group.

```
F-Test Statistic: 793.3060406325525
p-Value of F: 0.0
Critical Value of F-test : 1.9397493579111722
```

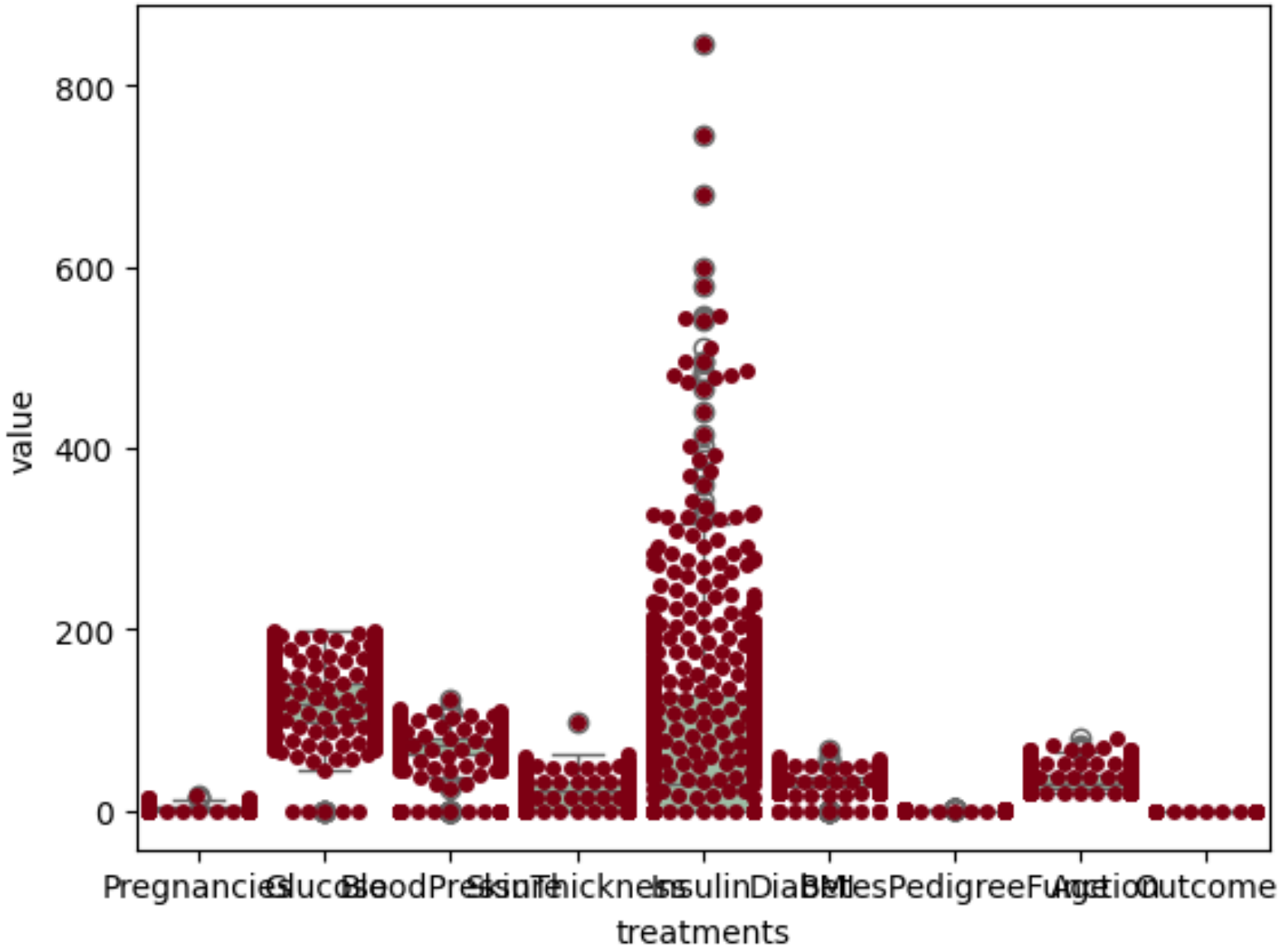
Analysis of Variance Table is:

	sum_sq	df	F	PR(>F)
C(treatments)	1.067945e+07	8.0	793.306041	0.0
Residual	1.161599e+07	6903.0	NaN	NaN

Analysis of Variance (ANOVA)

Interpretation:

- The $p-value$ obtained from ANOVA analysis is significant (p<0.05), and
- Therefore, we conclude that there are significant differences among treatments.



Questionnaire

1. If a person is above 37 years old what are the chances that he will be having diabeties?

```
fav_out = data[(data['Age'] > 37) &(data['Outcome'] == 1)]['Outcome'].count()
tot_out = len(data[data['Age'] > 37]['Outcome'])
probability = fav_out/tot_out * 100
print(f'\nThe probability would be:', round(probability,2),"%")
```

```
The probability would be: 51.49 %
```

Questionnaire

2. What are the chances that if the person has a glucose level more that 140 he has diabetes?

```
fav_out = data[(data['Glucose'] > 140) &(data['Outcome'] == 1)]['Outcome'].count()
tot_out = len(data[data['Glucose'] > 140]['Outcome'])
probability = fav_out/tot_out * 100
print(f'\nThe probability would be:',round(probability,2),"%")
```

```
The probability would be: 68.75 %
```

Questionnaire

3. Which approach is the best to classify a person to be Diabetic or not ? `(Based on Accuracy)`

- Total 768 patients record
- Using 650 data for training
- Using 100 data for testing
- Using 18 data for checking

Questionnaire

```
# Use the our training data to create a bayesian classifier.
diabetesCheck = SVC()
diabetesCheck.fit(trainData, trainLabel)
```

```
▾ SVC
SVC()
```

```
# After we train our bayesian classifier , we test how well it works using our test data.
accuracy = diabetesCheck.score(testData,testLabel)
print("Accuracy = ",accuracy * 100,"%")
```

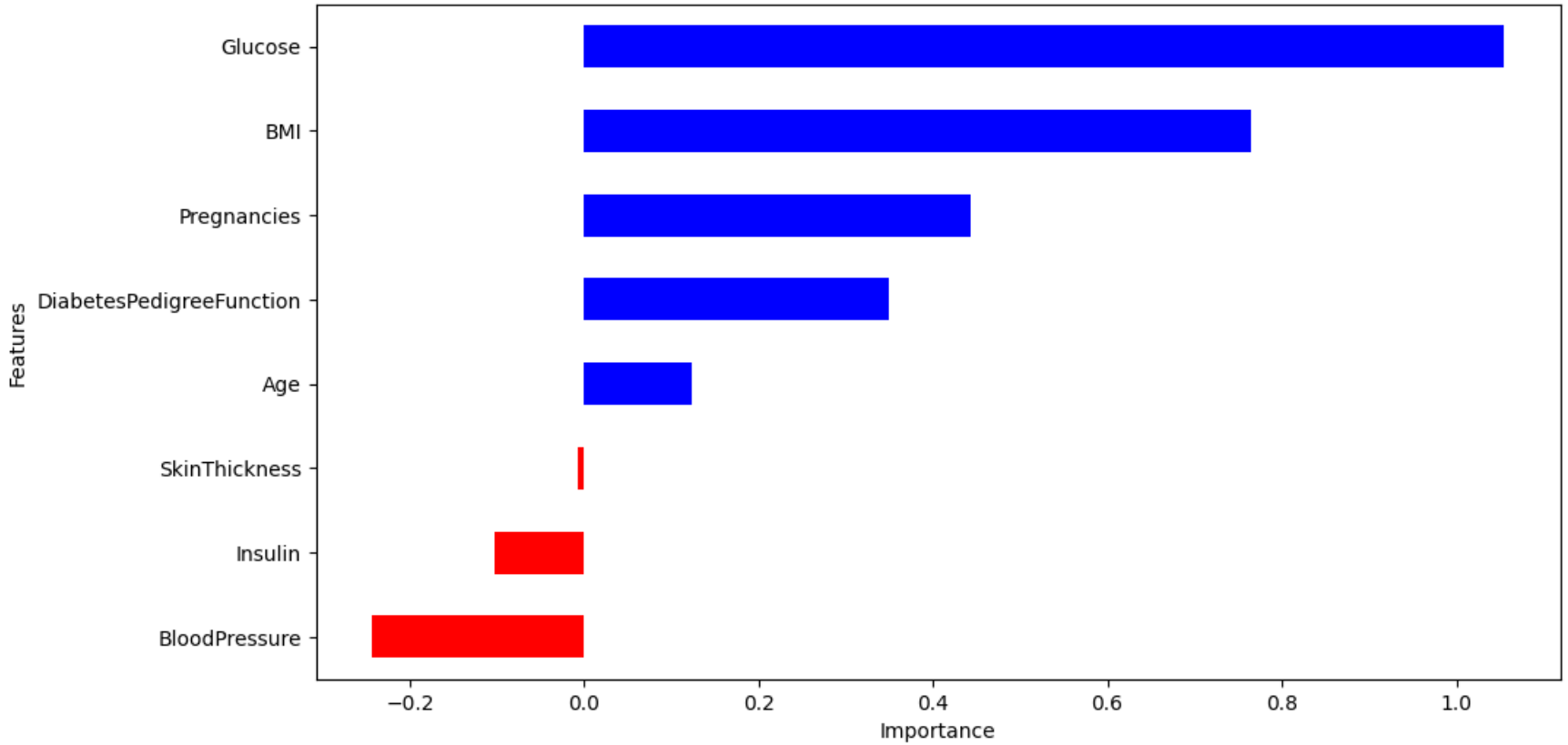
```
Accuracy =  75.0 %
```

```
diabetesCheck = LogisticRegression()
diabetesCheck.fit(trainData,trainLabel)
accuracy = diabetesCheck.score(testData,testLabel)
print("Accuracy = ",accuracy * 100,"%")
```

```
Accuracy =  78.0 %
```

Questionnaire

4. Is Glucose the most important factor in determining the onset of diabetes followed by BMI and Age?



- Glucose is the most important factor in determining the onset of diabetes followed by BMI and Age.
- Other factors such as Diabetes Pedigree Function, Pregnancies, Blood Pressure, Skin Thickness and Insulin also contributes to the prediction.

Splitting the dataset into Training and Test data

Train and Test Split on all features

```
Ratio of Diabetes to Non-Diabetic Labels in training dataset is: 0.66
Ratio of Diabetes to Non-Diabetic Labels in testing dataset is: 0.63
```

Seems to be fairly evenly distributed between the training and testing dataset.

Feature Scaling

using a StandardScaler to perform feature scaling. This will retain the mean and the standard deviation of the sample distribution of the data set, and reuse it to transform the X_train and X_test subsequently. I try to reuse the mean and standard deviation obtained from the training set and apply it to the testing set as well. Standardizing data after data splitting is to prevent data leakage from test dataset into train dataset.

Feature Scaling

```
In [84]: data_0['Outcome'].value_counts()

Out[84]: Outcome
         0    500
         1    268
         Name: count, dtype: int64
```

From the above, it would seem that there is imbalanced data, number of samples of patients without diabetes is approximately twice the number of samples of patients with diabetes. Some up sampling may be required to adjust the number of samples. It would seem that it would be better if up sampling is done after feature scaling.

```
Before Upsampling, no. of samples in the training dataset: 537
Before UpSampling, counts of label '1': 183
Before UpSampling, counts of label '0': 354
After Upsampling, no. of samples in the training dataset: 708
After UpSampling, counts of label '1': 354
After UpSampling, counts of label '0': 354
```

Retrospective

- **Unbalanced Dataset**
  – Same number of Diabetic [34.9%] & Non-Diabetic [65.1%] patient's data not present.
- **Hard to train the model with such unbalanced data set.**
- **Using non-supervise learning approach for the same was not giving satisfactory outcome.**
- **Given dataset is more accurate[83%] for Support Vector Machine[SVM]**
- **Random forester classifier is providing accuracy of 79%**

Conclusion:

`Support Vector Machine (SVM)` Classifier is the right model due to high accuracy, precision and recall score. One reason why `SVM Classifier` compared to `Random Forest Classifier` showed an improved performance was because of the presence of outliers. Here is the accuracy for the different approaches

- Support Vector Machine – 83%
- Random Forest Classifier – 79%
- KMean – 67%

As mentioned before, since Random Forest is not a distance-based algorithm, it is not much affected by outliers, whereas distance-based algorithm such as KMean showed a lower performance.

Based on the feature importance:

`Glucose` is the most important factor in determining the onset of diabetes followed by BMI and Age. Other factors such as Diabetes Pedigree Function, Pregnancies, Blood Pressure, Skin Thickness and Insulin also contributes to the prediction.

- As we can see, the results derived from Feature Importance makes sense as one of the first things that actually is monitored in high-risk patients is the Glucose level.
- An increased BMI might also indicate a risk of developing Type II Diabetes. Normally, especially in case of Type II Diabetes, there is a high risk of developing as the age of a person increases (given other factors).

THANK YOU!

Prof. MRUNALINI (Data Science Trainer) , Mumbai

G – mail : mrunalini0107@gmail.com