**Course: Data Mining**

**Homework 2**

**Student ID: 20C14001**

**Student Name: Le Duong Tuan Anh**

**Dataset: “diabetes.csv”**

*1) Desbribe the dataset*

Dataset type: Record.

This dataset has 2,000 rows and 9 columns. The last column, “Outcome”, shows the result that the person has a diabetes or not.

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Type | Datatype | Has missing value? |
| Pregnancies | Interval-scaled | Int64 | No |
| Glucose | Interval-scaled | Int64 | No |
| BloodPressure | Interval-scaled | Int64 | No |
| SkinThickness | Interval-scaled | Int64 | No |
| Insulin | Interval-scaled | Int64 | No |
| BMI | Interval-scaled | Float | No |
| DiabetesPedigreeFunction | Interval-scaled | Float | No |
| Age | Interval-scaled | Int64 | No |
| Outcome | Categorical Data Binary (0/1) | Int64 | No |

**Pregnancies**: Number of times pregnant

**Glucose**: Plasma Glucose Concentration.

**BloodPressure**: Diastolic Blood Pressure.

**SkinThickness**: Estimate body fat.

**Insulin**: 2-Hour Serum Insulin.

**BMI**: Body Mass Index.

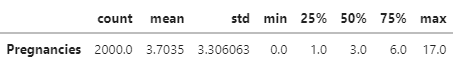
**DiabetesPedigreeFunction**: Iinformation about diabetes history in relatives and genetics.

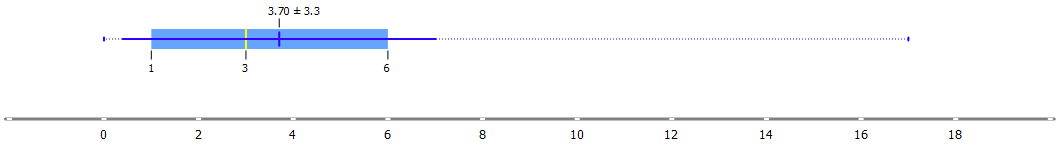
**Age**: Age (years).

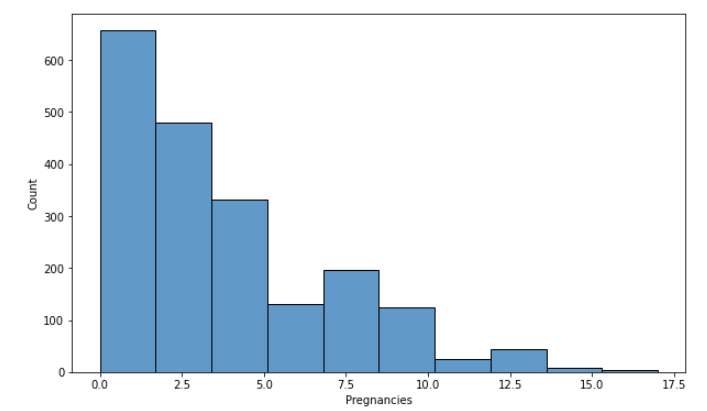
**Outcome**: 0 = Diabetic, 1 = Not Diabetic

*2) Appy basic statictical descriptions for the dataset*

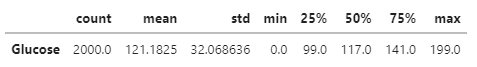
**2.1. Pregnancies**

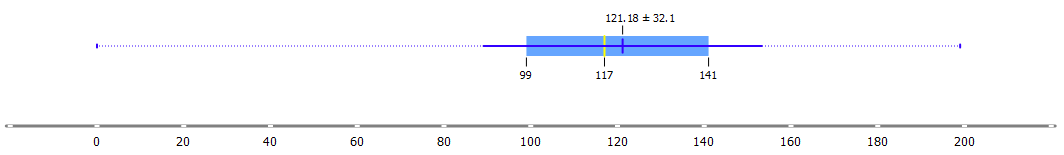
****

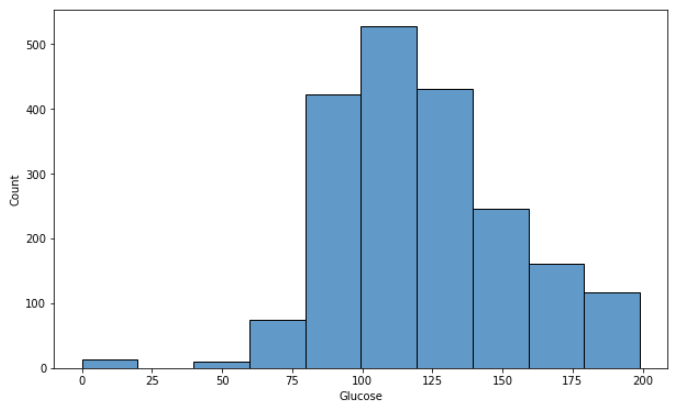
****

****

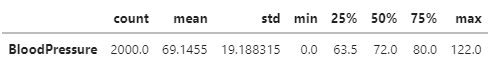
**2.2. Glucose**

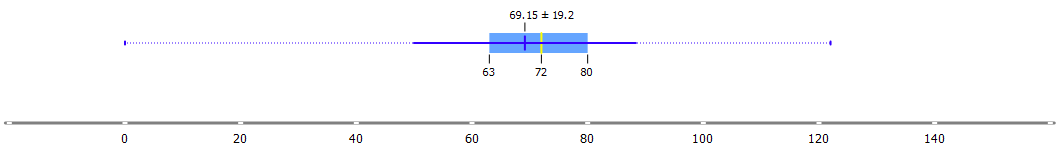
****

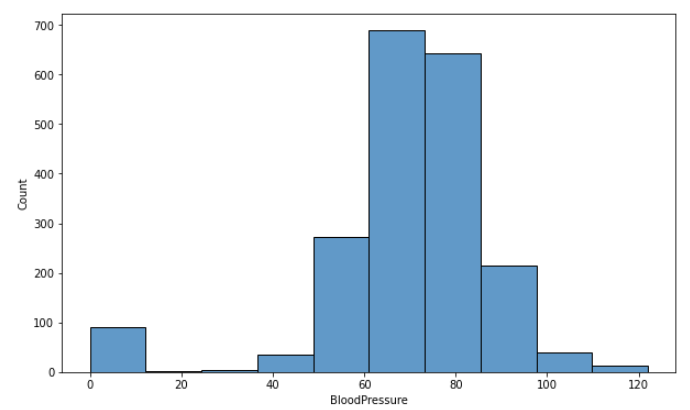
****

****

**2.3. BloodPressure**

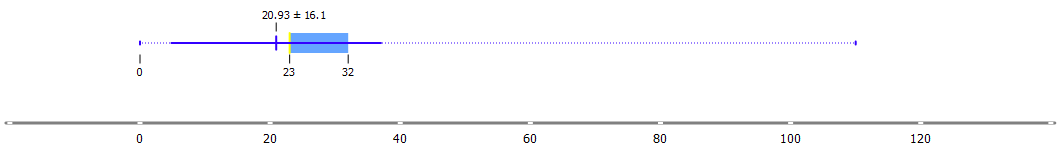
****

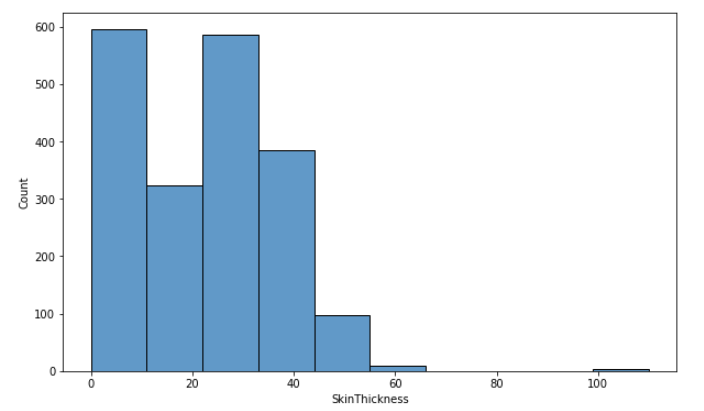
****

****

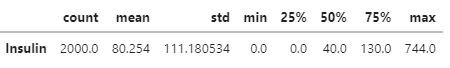
**2.4. SkinThickness**

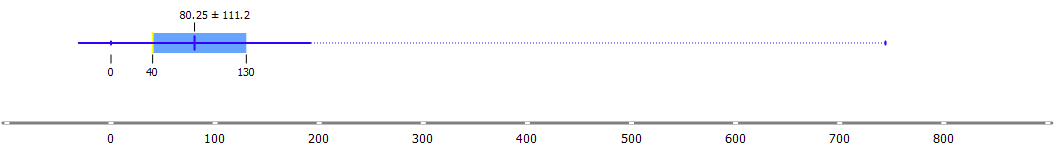
****

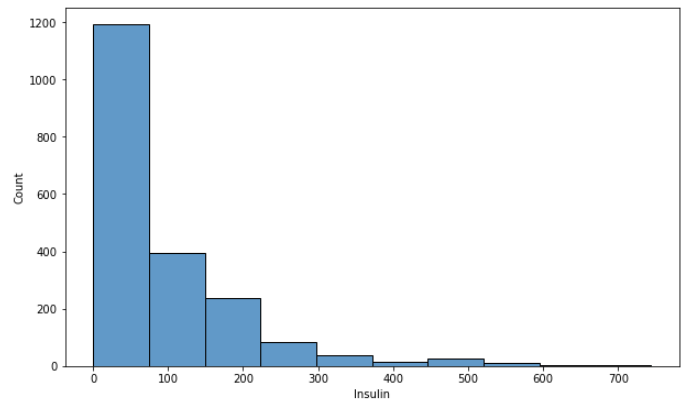
****

****

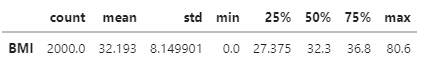
**2.5. Insulin**

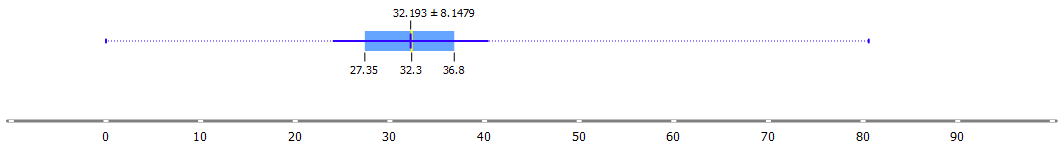
****

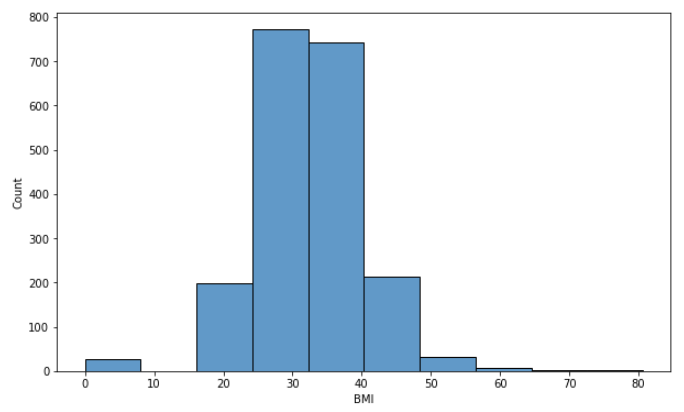
****



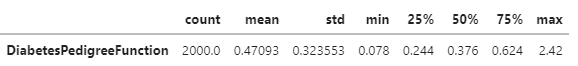
**2.6. BMI**

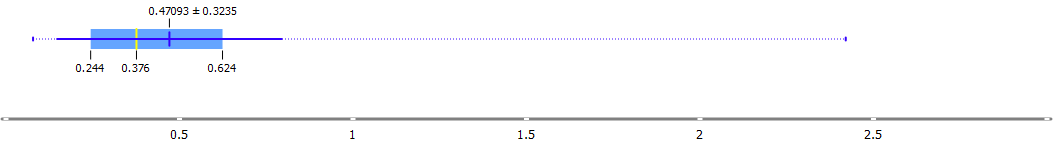
****

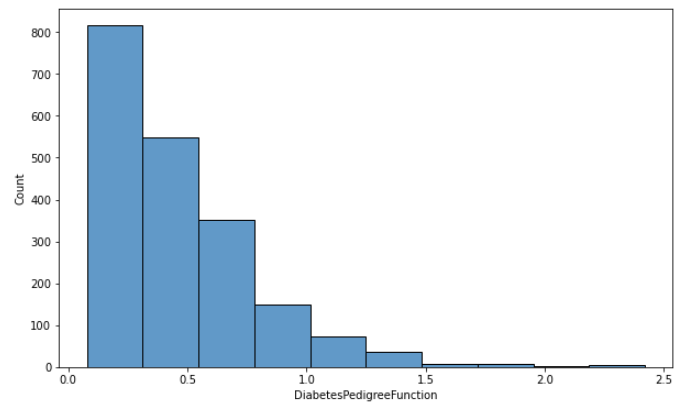
****

****

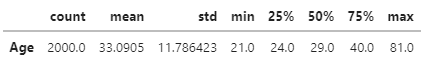
**2.7. DiabetesPedigreeFunction**

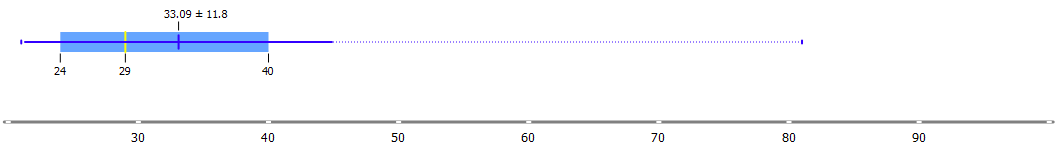
****

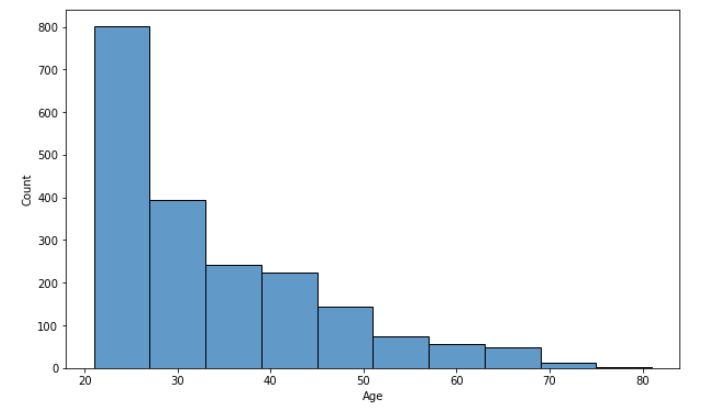
****

****

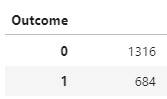
**2.8. Age**

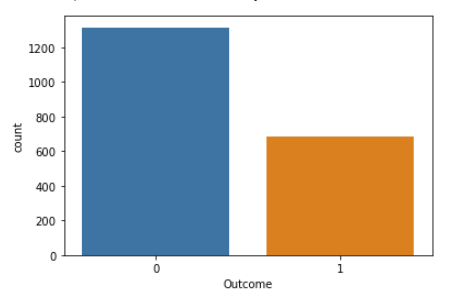
****

****

****

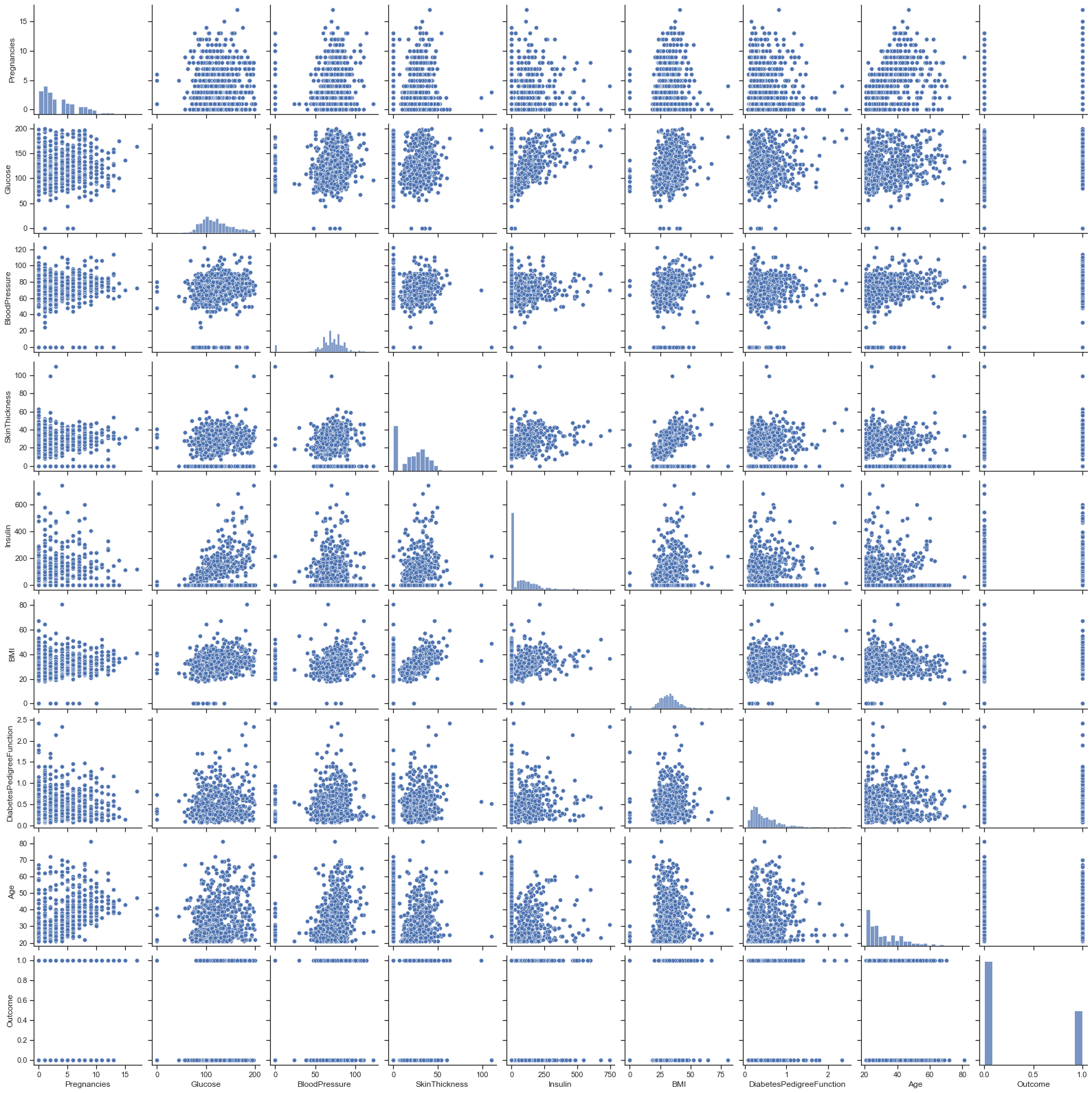
**2.9. Outcome**

****

****

*3) Visualize this dataset by using scatterplot matrix.*

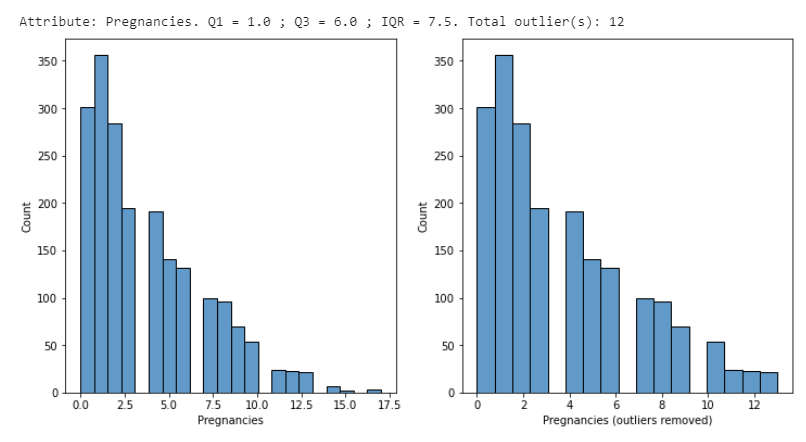
**Scatter Plot without distinct Outcome result.**

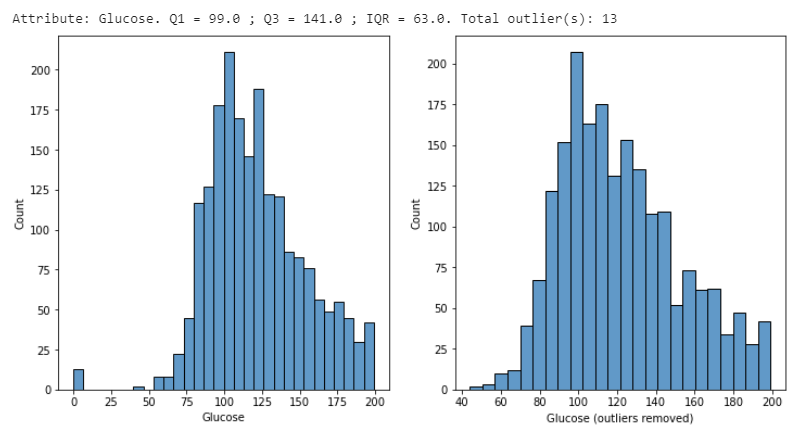


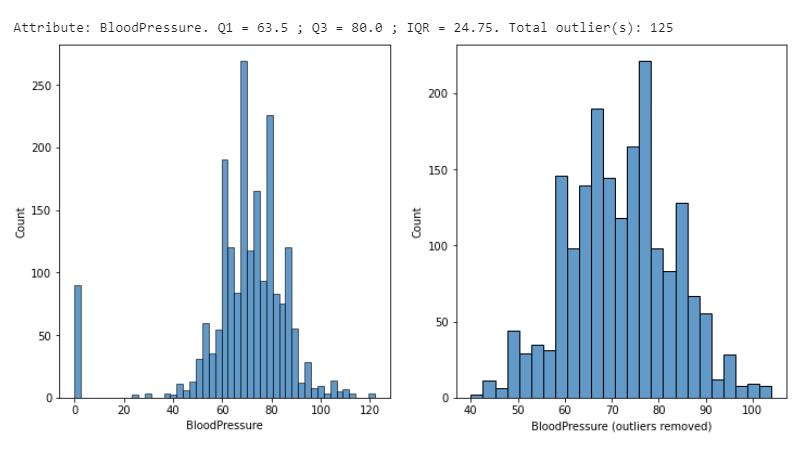
**Scatter Plot with distinct Outcome result.**

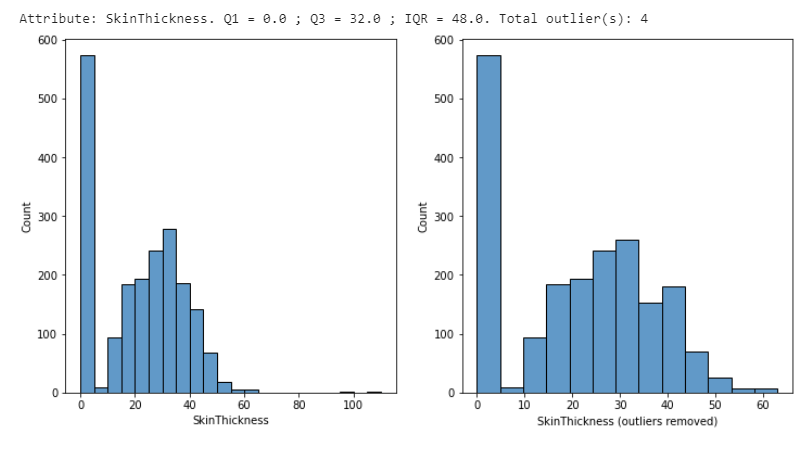


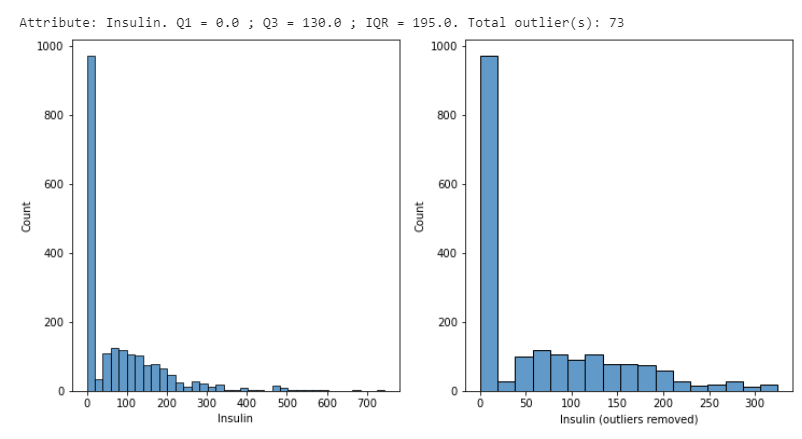
*4) Do we have outliers in this dataset? For each attribute, list them out.*

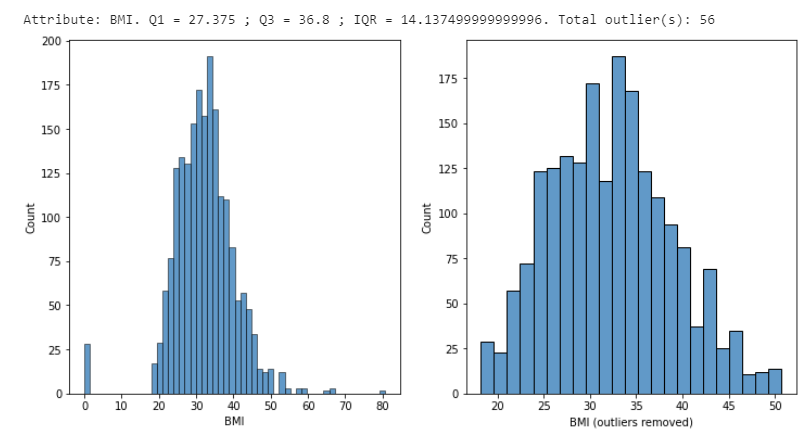
We would consider outliers by using Inter-quartile Range method. In case the data point is out of range [ Q1 – 1.5\*IQR, Q3 + 1.5\*IQR ], it might be outliers. 

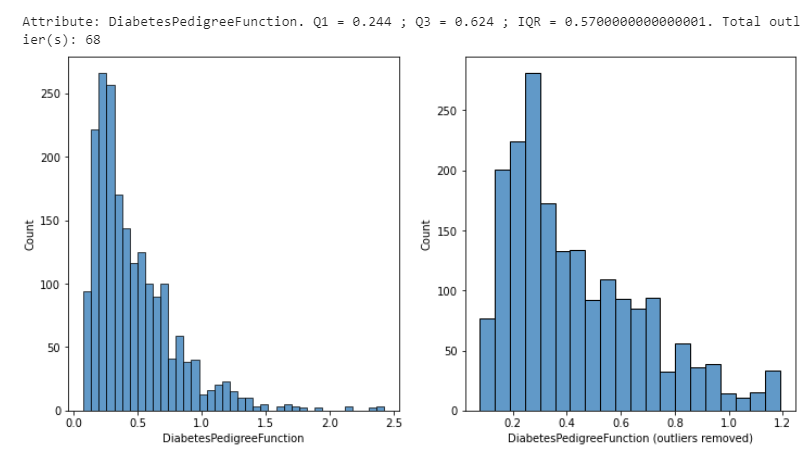


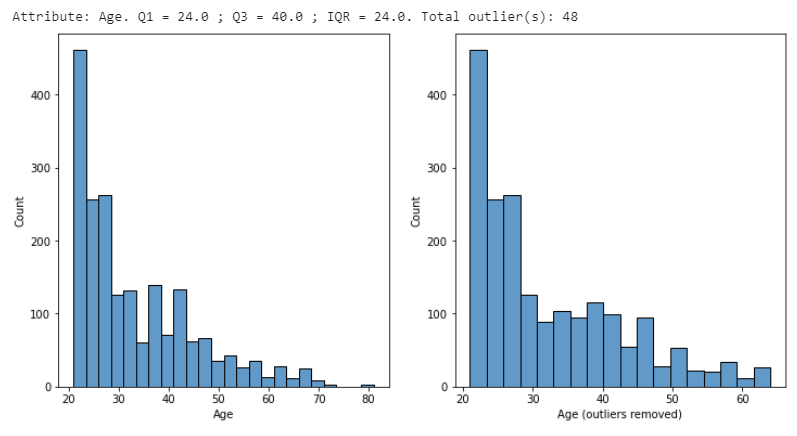


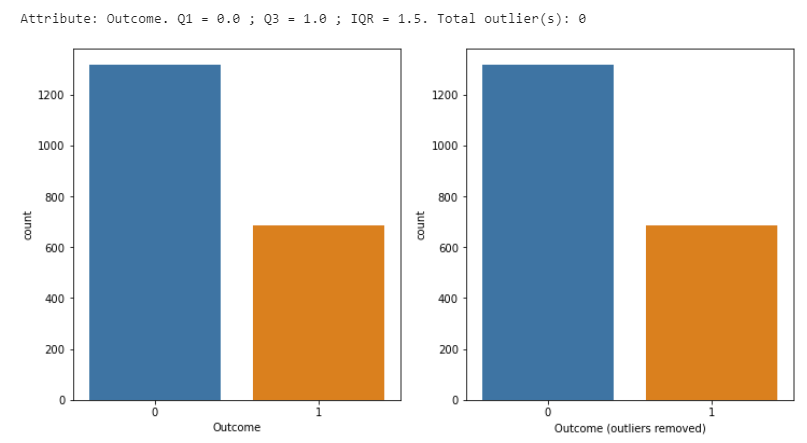








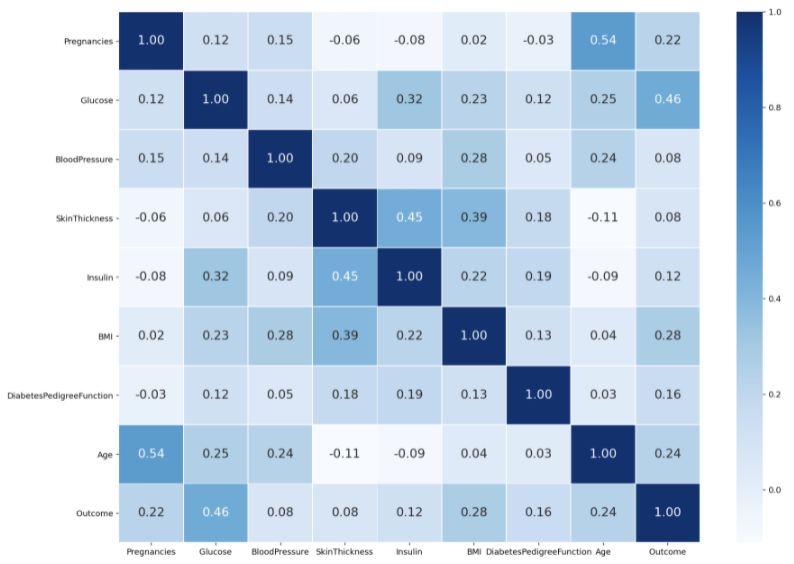




*5, 6) Check the relationship of the two attributes*

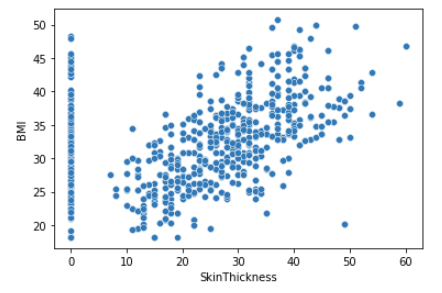
The correlation of dataset is show as below.

(Standard correlation co-efficient, **Pearson**)



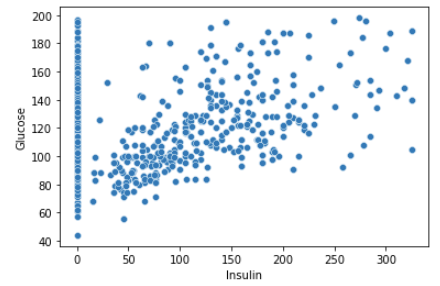
*5)* ***SkinThickness*** *and* ***BMI****.*

The correlation between SkinThickness and BMI is quite strong, 0.45.

**

*Note that, the SkinThickness should not be 0 (as the thickness of skin is quite not reasonable when equal to 0.*

*6)* ***Insulin*** *và* ***Glucose****.*

**

*Note that, the Insulin should not be 0.*

*7) Standardize this dataset such that all attributes have the same data unit.*

We standardize features by removing the mean, then scaling dataset to unit variance. Note that we will not standardize the **Outcome** column, as it is a **categorical column**.

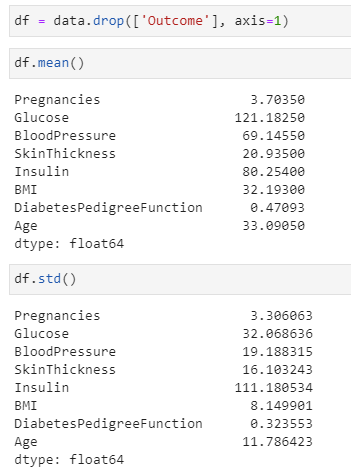
The standardized score is calculated as below:

Z = (X – U) / S

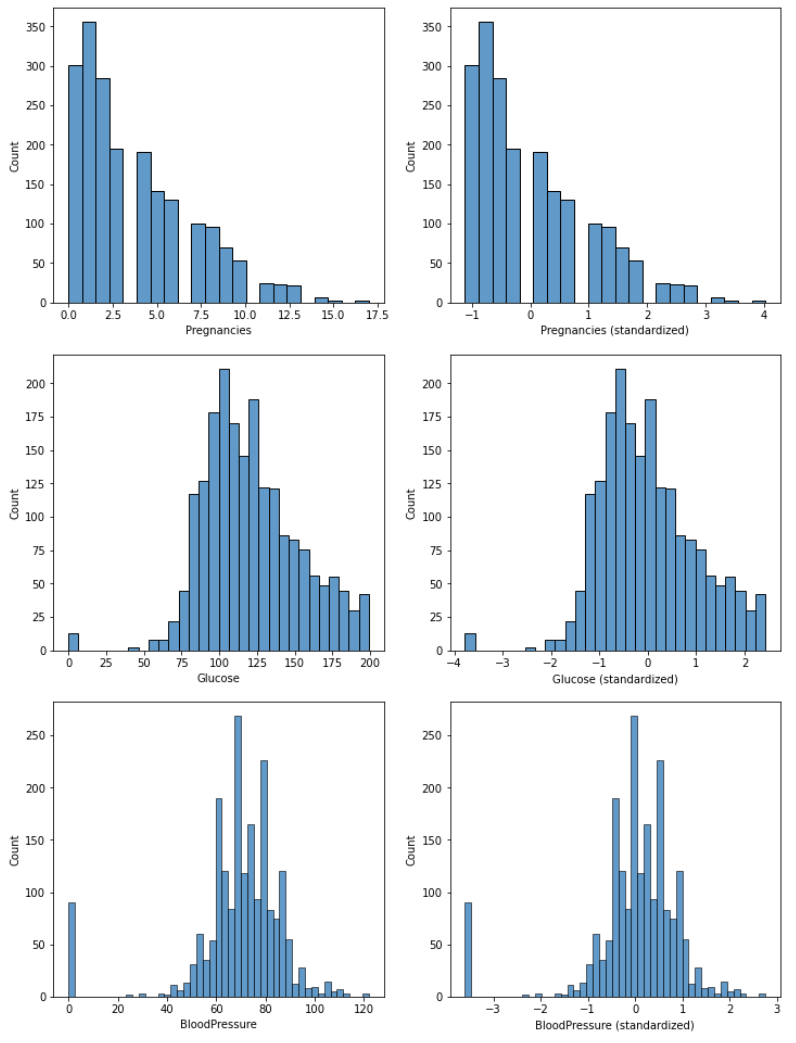
**Where:**

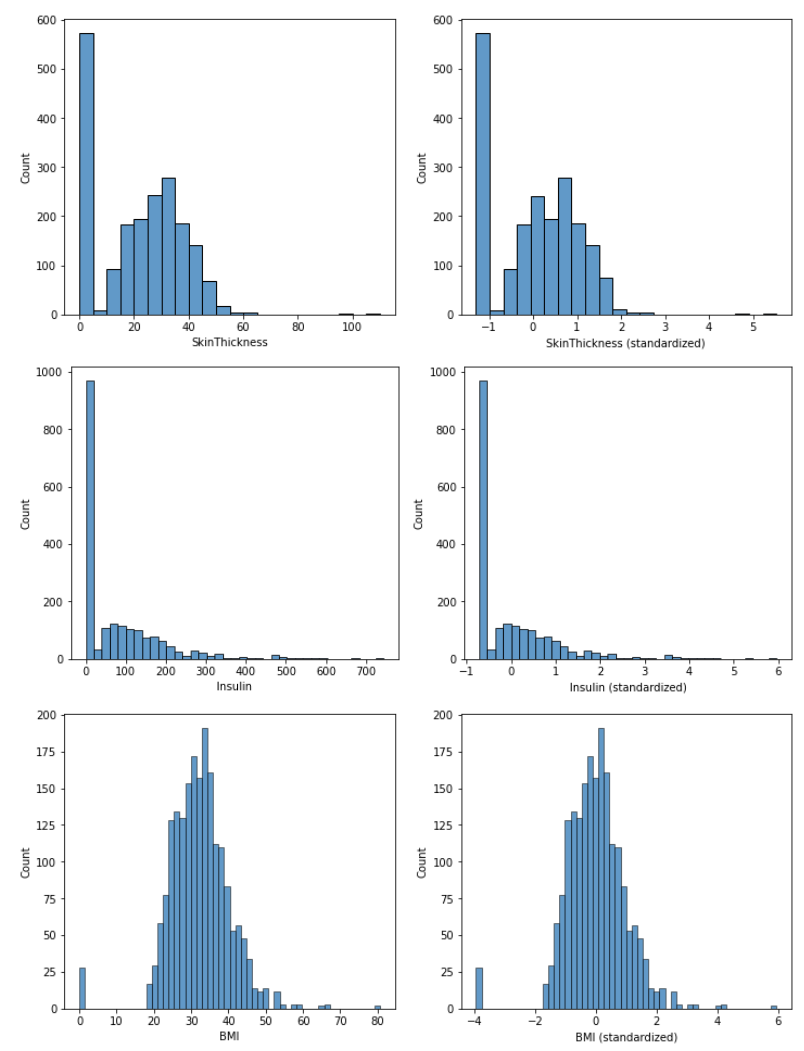
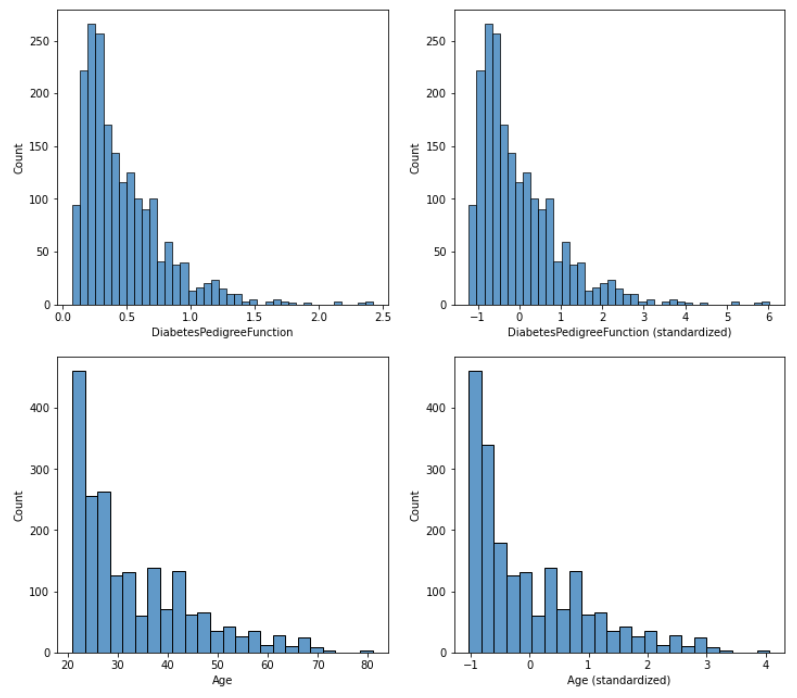
* X: The value of sample in dataset.
* U: The mean of dataset.
* S: The standard deviation of dataset.

**Code:**



**Compare histogram result between and after standardized:**

****

**** ****