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.

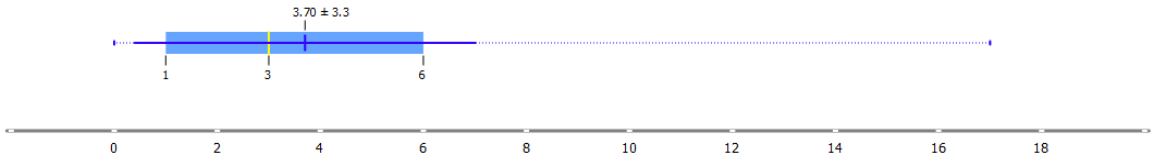
Age: Age (years).

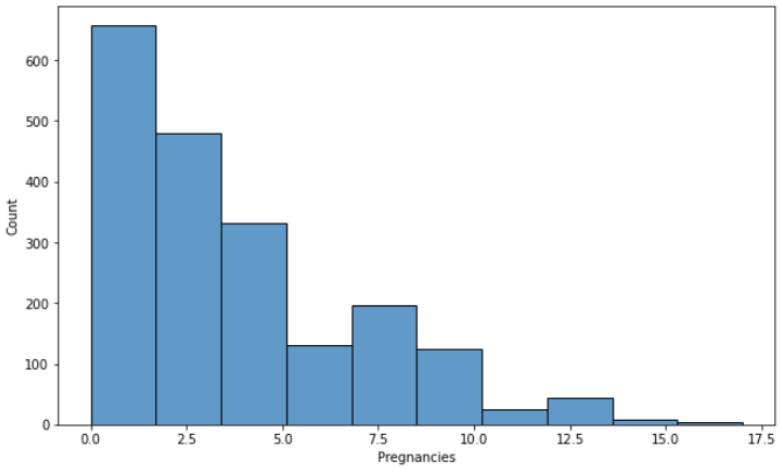
Outcome: 0 = Diabetic, 1 = Not Diabetic

2) Appy basic statictical descriptions for the dataset

2.1. Pregnancies

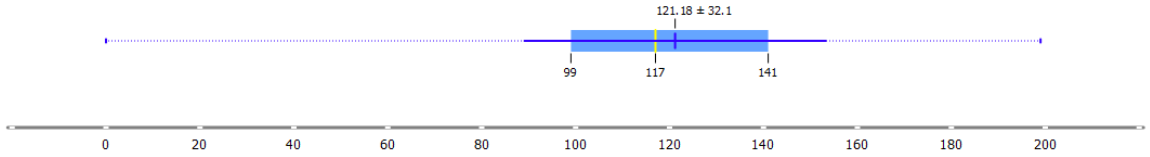
	count	mean	std	min	25%	50%	75%	max
Pregnancies	2000.0	3.7035	3.306063	0.0	1.0	3.0	6.0	17.0

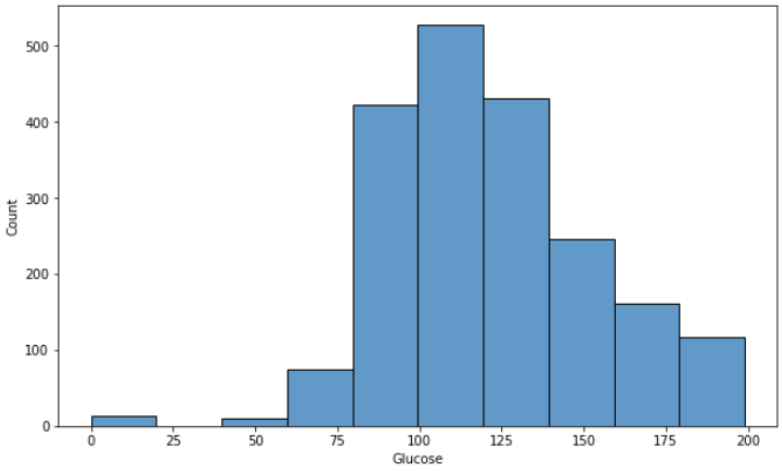




2.2. Glucose

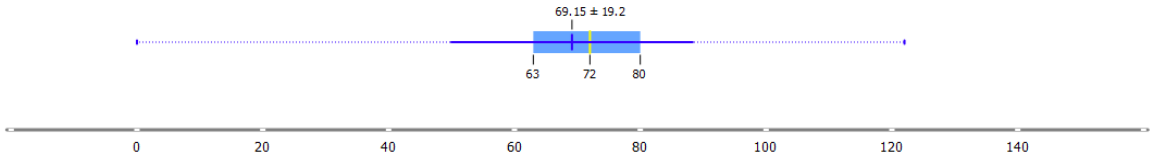
	count	mean	std	min	25%	50%	75%	max
Glucose	2000.0	121.1825	32.068636	0.0	99.0	117.0	141.0	199.0

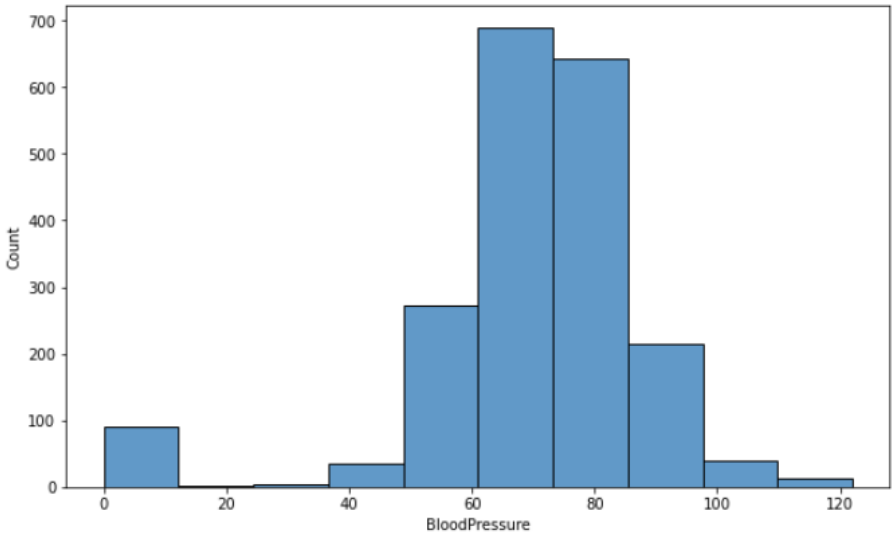




2.3. BloodPressure

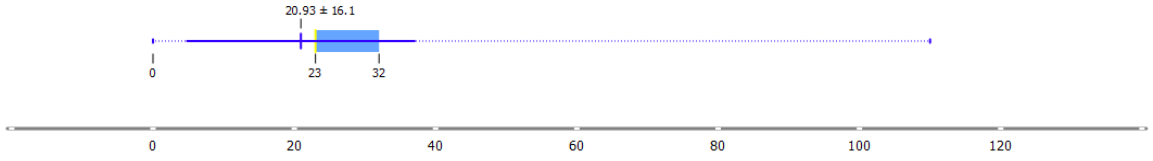
	count	mean	std	min	25%	50%	75%	max
BloodPressure	2000.0	69.1455	19.188315	0.0	63.5	72.0	80.0	122.0

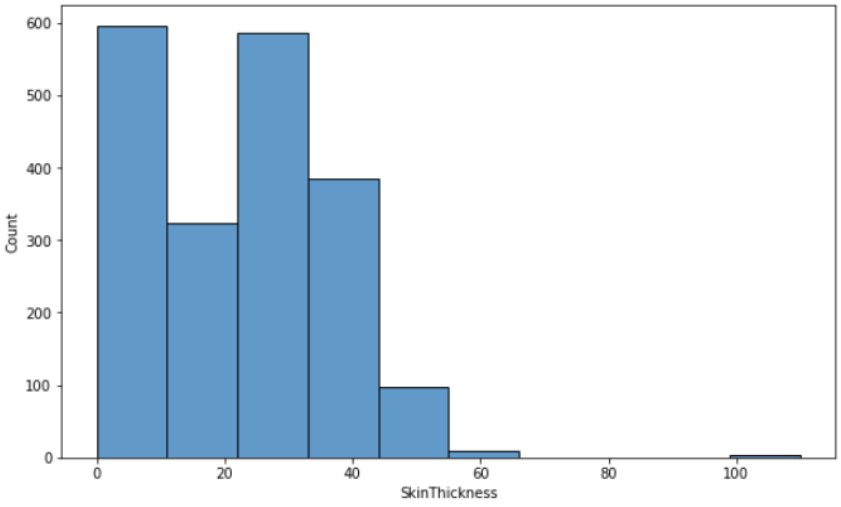




2.4. SkinThickness

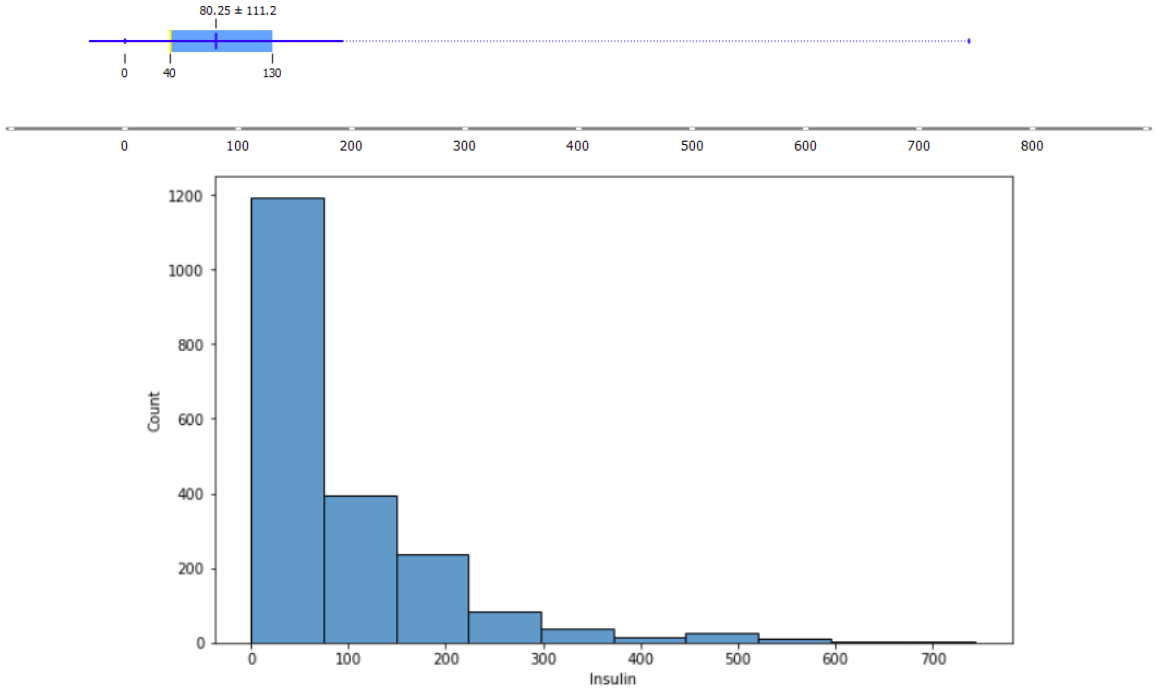
	count	mean	std	min	25%	50%	75%	max
SkinThickness	2000.0	20.935	16.103243	0.0	0.0	23.0	32.0	110.0





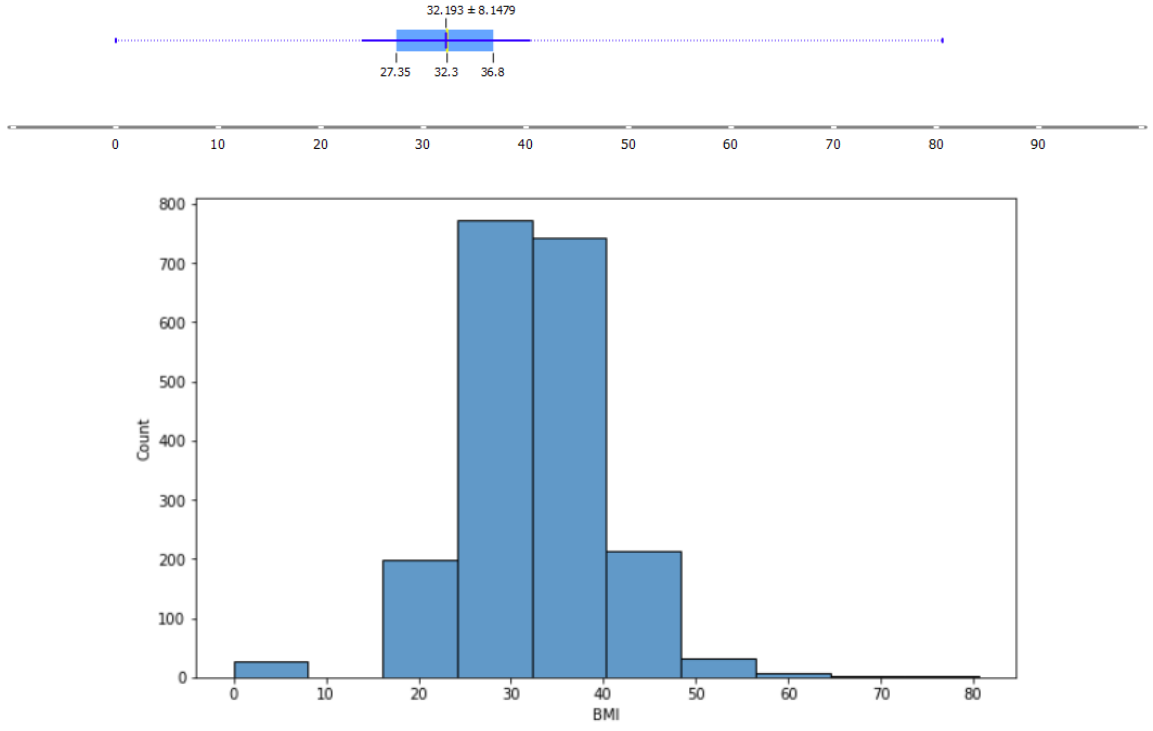
2.5. Insulin

	count	mean	std	min	25%	50%	75%	max
Insulin	2000.0	80.254	111.180534	0.0	0.0	40.0	130.0	744.0



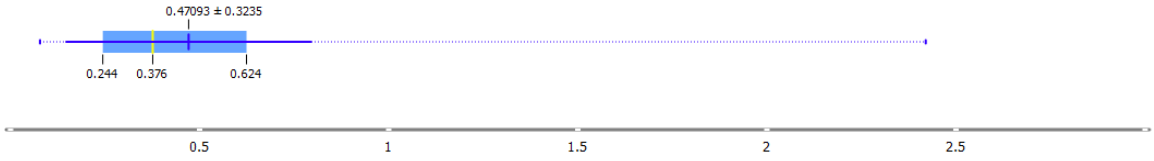
2.6. BMI

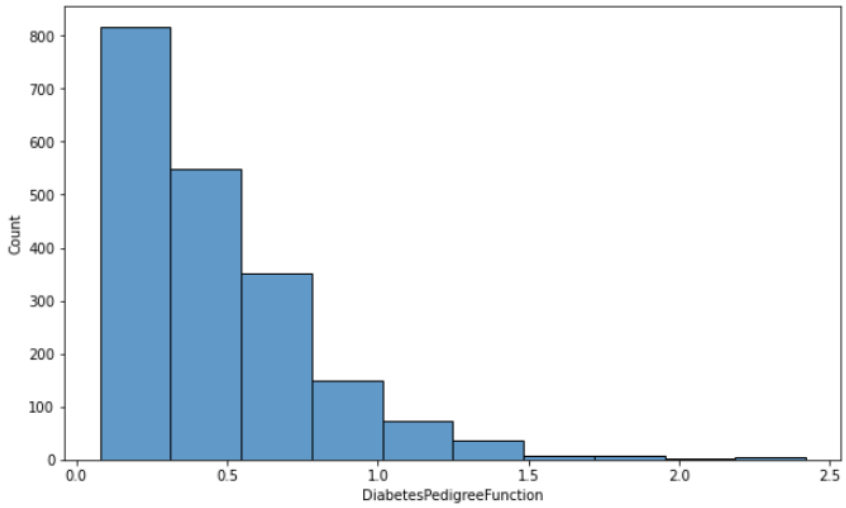
	count	mean	std	min	25%	50%	75%	max
BMI	2000.0	32.193	8.149901	0.0	27.375	32.3	36.8	80.6



2.7. DiabetesPedigreeFunction

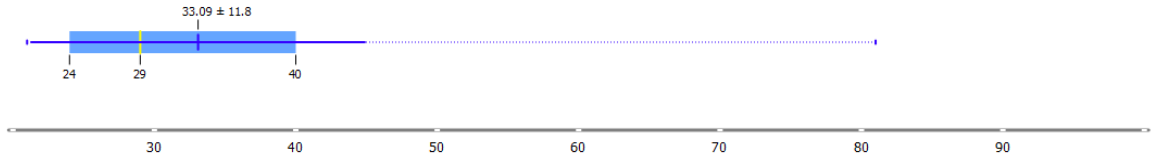
	count	mean	std	min	25%	50%	75%	max
DiabetesPedigreeFunction	2000.0	0.47093	0.323553	0.078	0.244	0.376	0.624	2.42

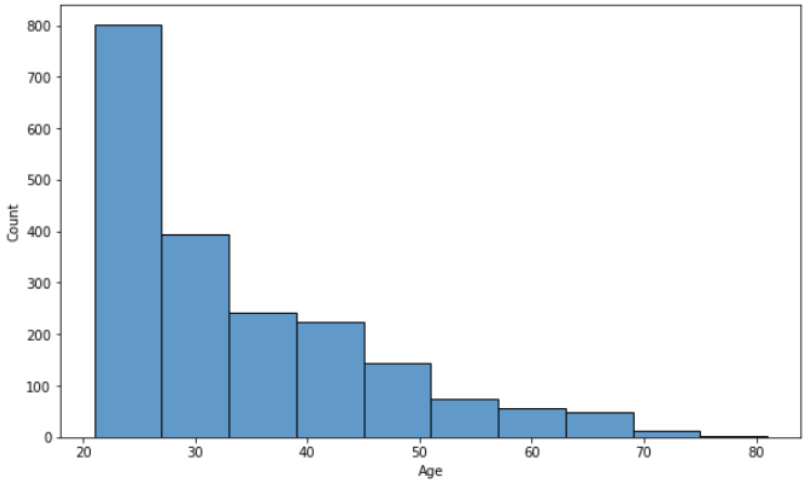




2.8. Age

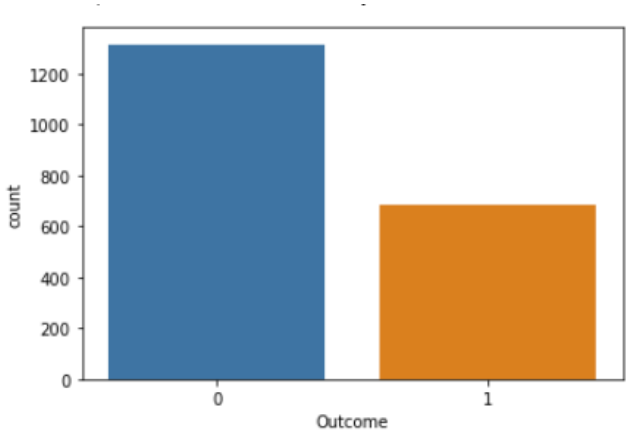
	count	mean	std	min	25%	50%	75%	max
Age	2000.0	33.0905	11.786423	21.0	24.0	29.0	40.0	81.0



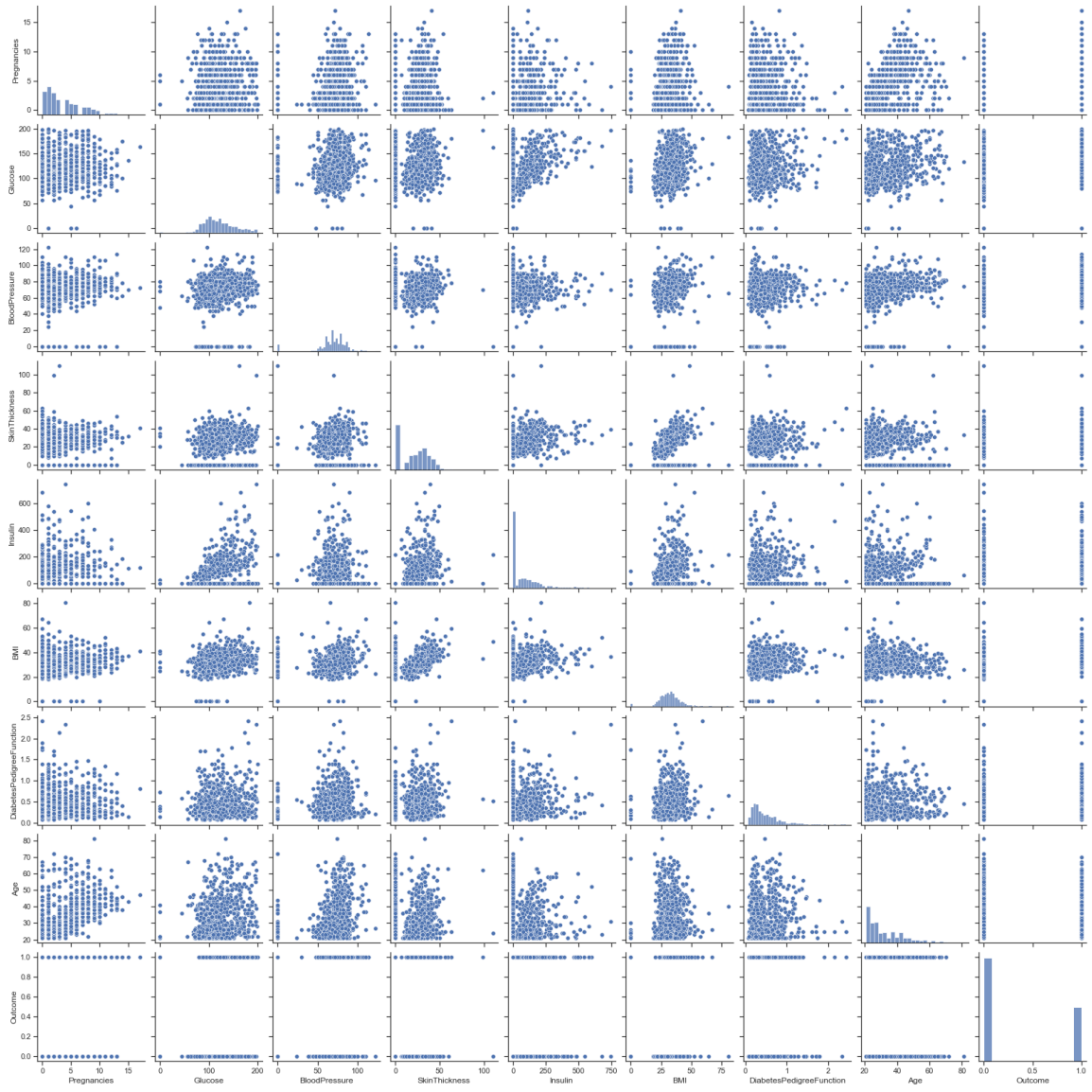


2.9. Outcome

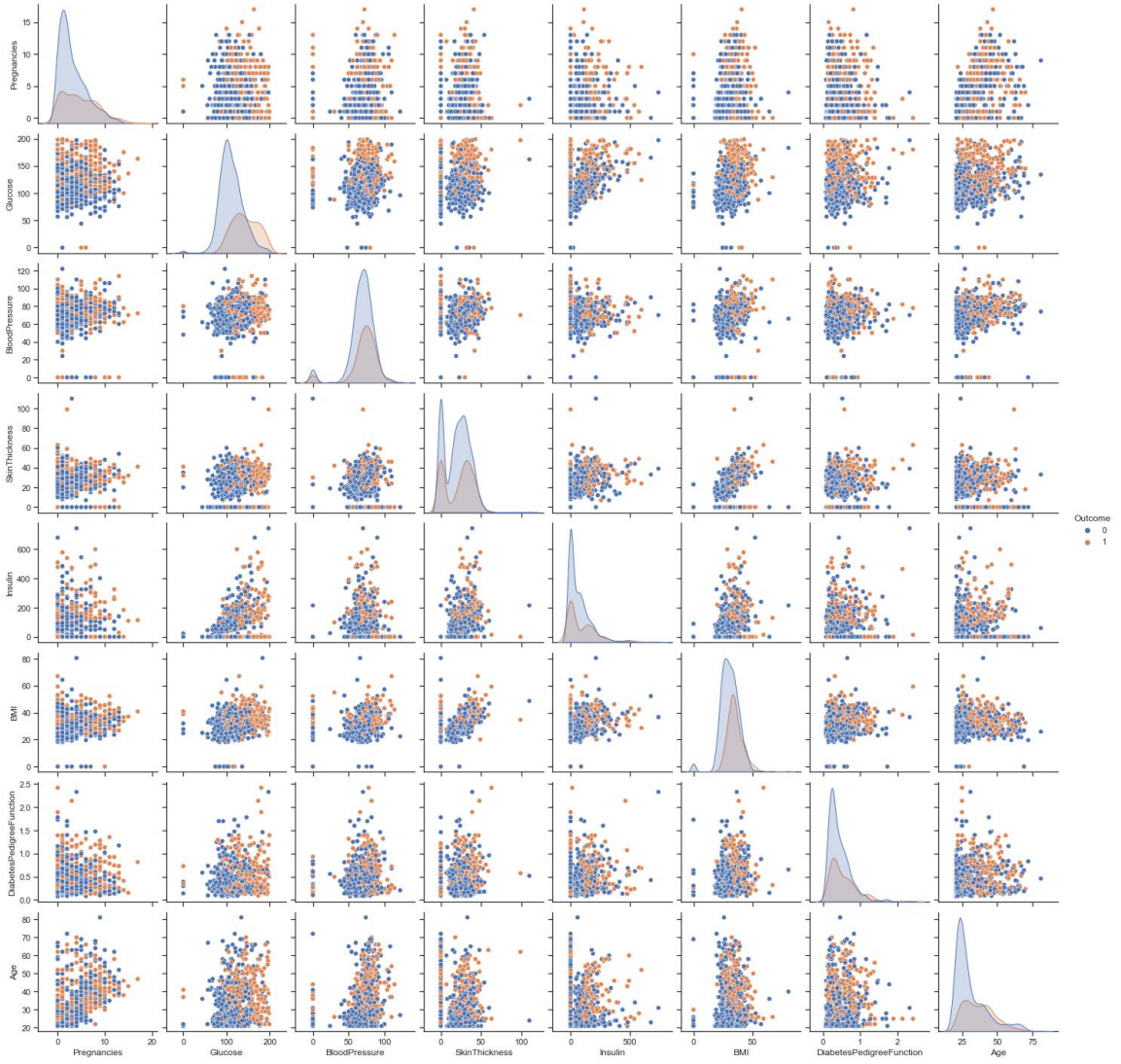
Outcome	
0	1316
1	684



3) Visualize this dataset by using scatterplot matrix.
Scatter Plot without distinct Outcome result.

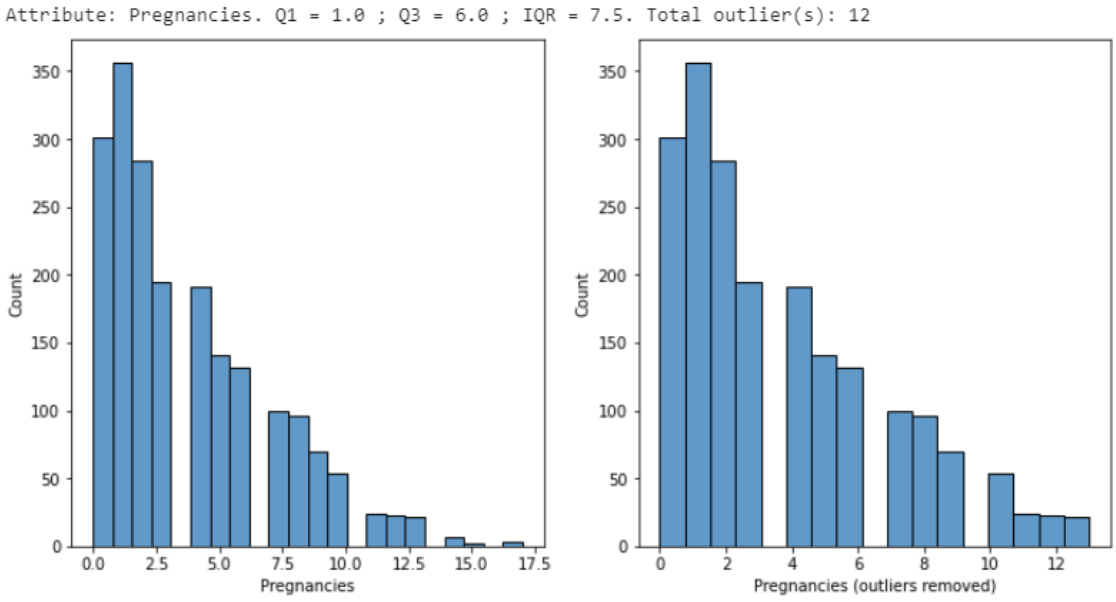


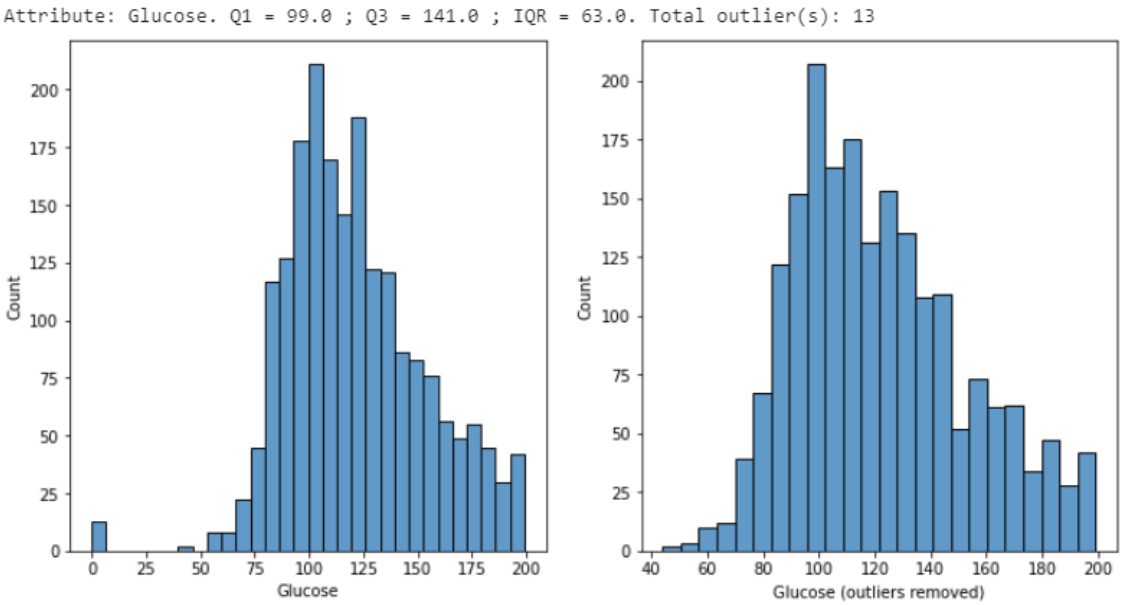
Scatter Plot with distinct Outcome result.



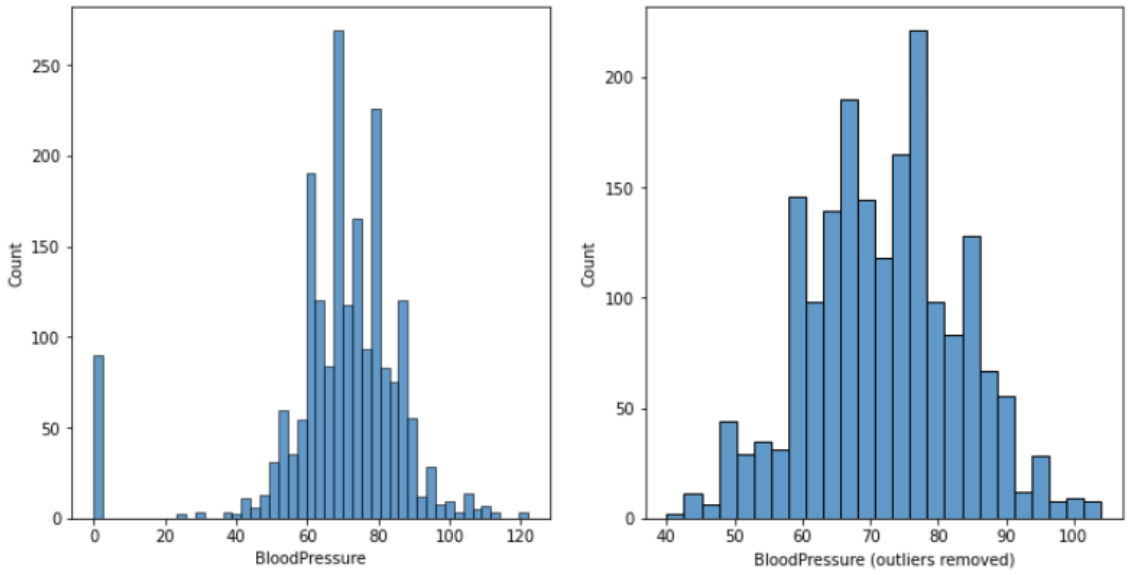
4) Do we have outliers in this dataset? For each attribute, list them out.

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.

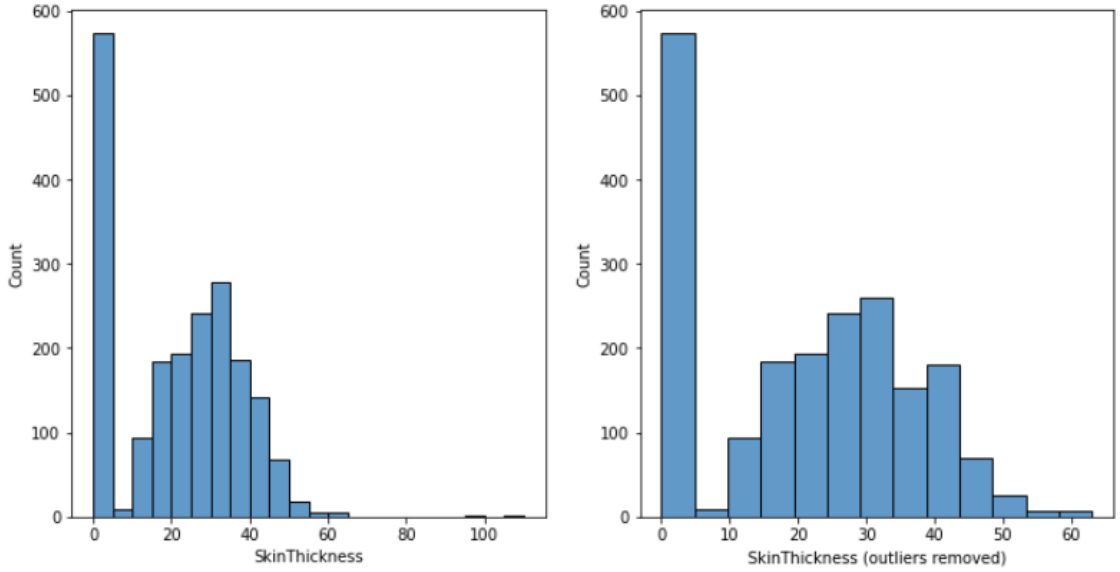




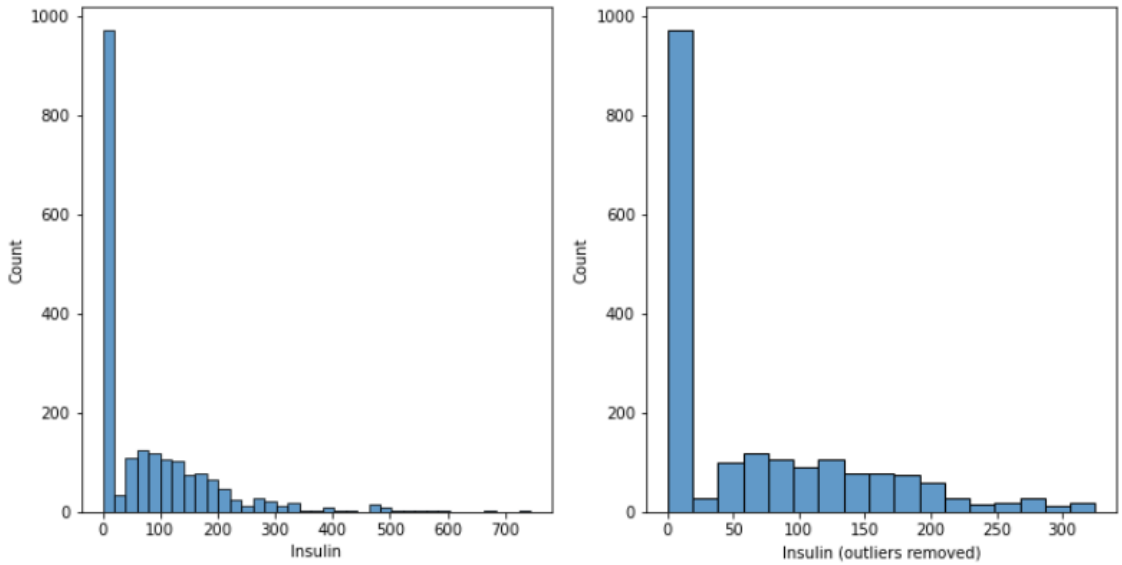
```
Attribute: BloodPressure. Q1 = 63.5 ; Q3 = 80.0 ; IQR = 24.75. Total outlier(s): 125
```



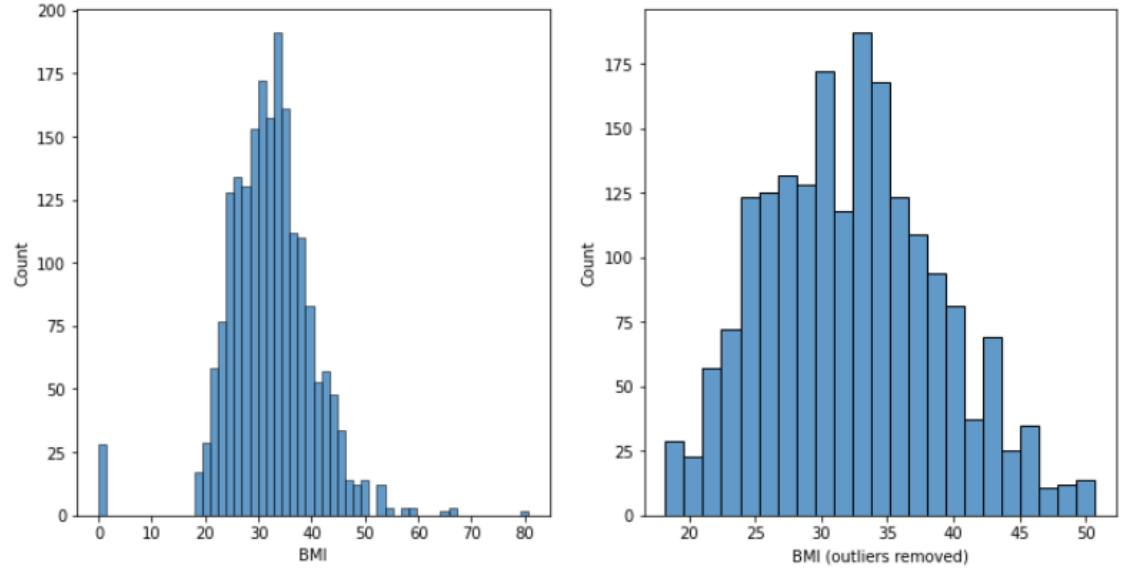

```
Attribute: SkinThickness. Q1 = 0.0 ; Q3 = 32.0 ; IQR = 48.0. Total outlier(s): 4
```



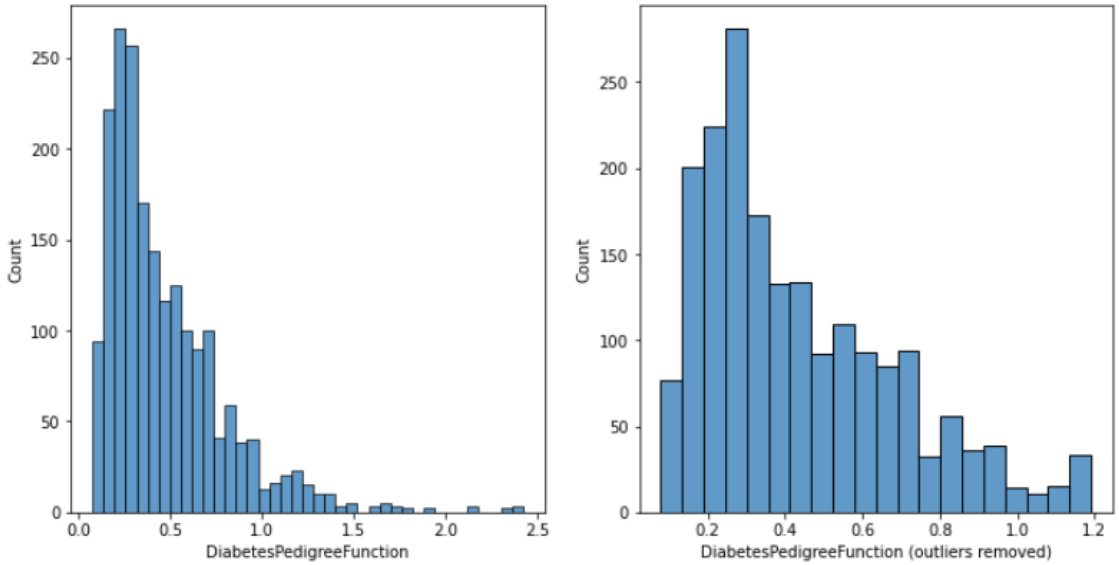
Attribute: Insulin. Q1 = 0.0 ; Q3 = 130.0 ; IQR = 195.0. Total outlier(s): 73



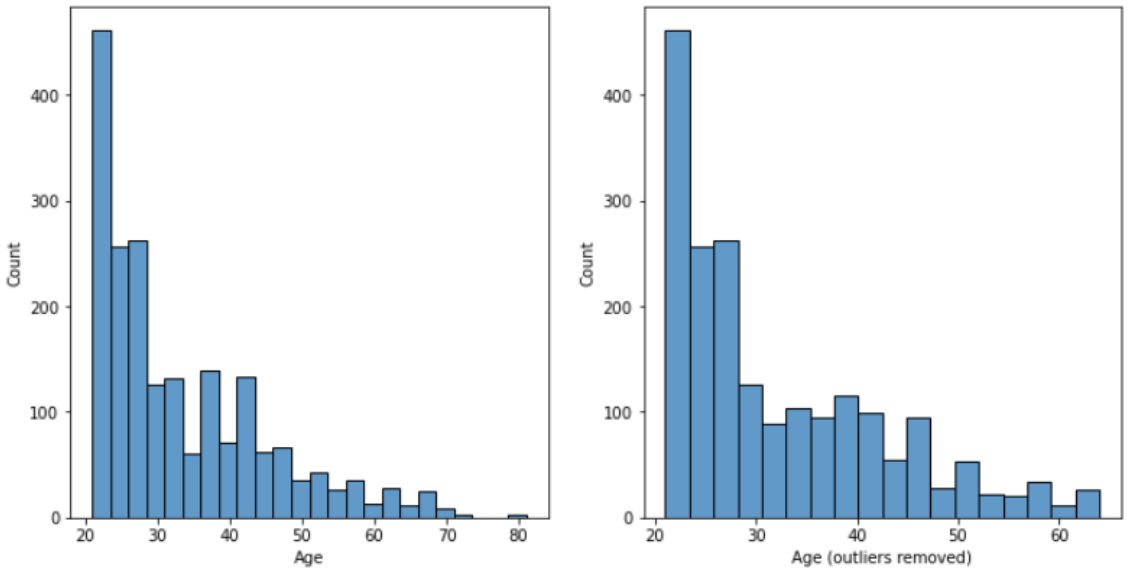
Attribute: BMI. Q1 = 27.375 ; Q3 = 36.8 ; IQR = 14.137499999999996. Total outlier(s): 56



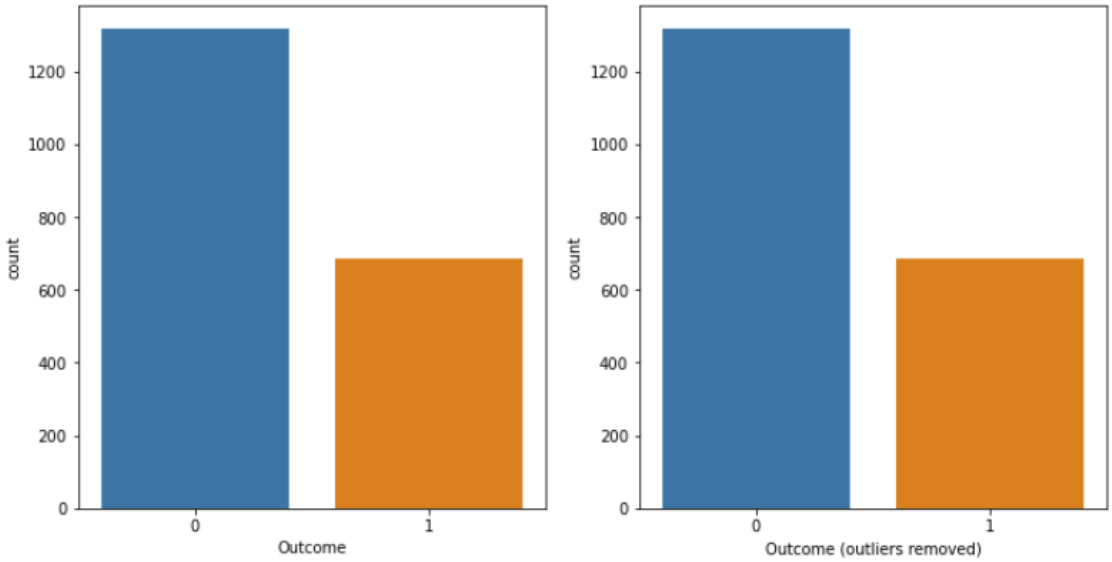
```
Attribute: DiabetesPedigreeFunction. Q1 = 0.244 ; Q3 = 0.624 ; IQR = 0.5700000000000001. Total outl
ier(s): 68
```




```
Attribute: Age. Q1 = 24.0 ; Q3 = 40.0 ; IQR = 24.0. Total outlier(s): 48
```



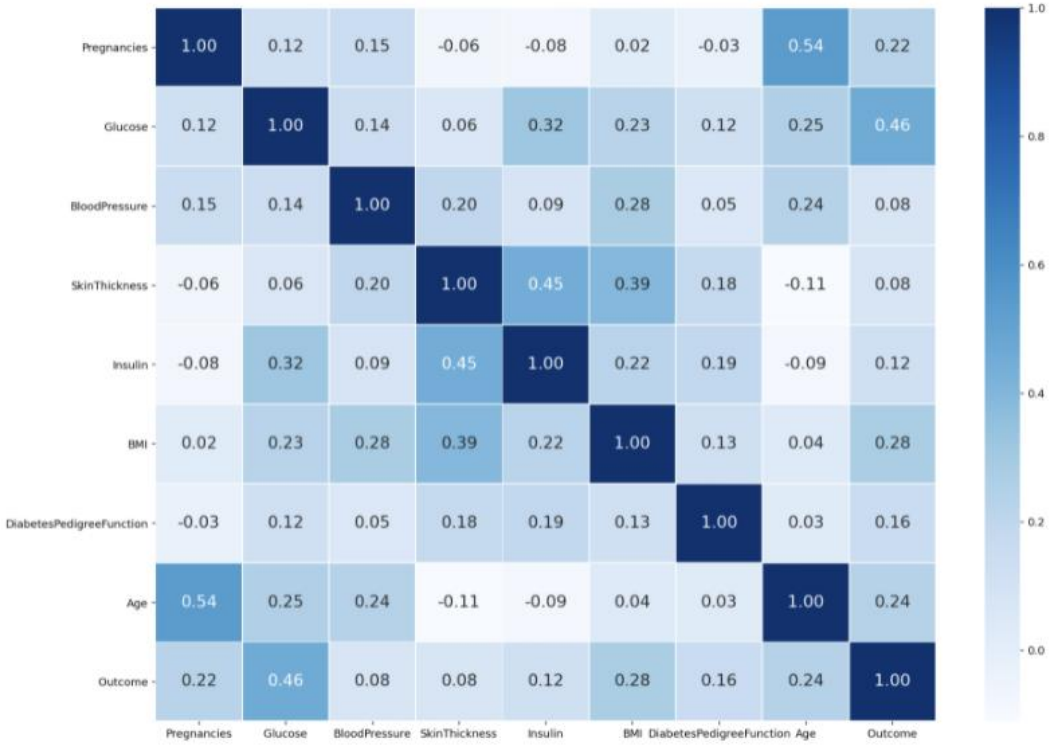
```
Attribute: Outcome. Q1 = 0.0 ; Q3 = 1.0 ; IQR = 1.5. Total outlier(s): 0
```



5, 6) Check the relationship of the two attributes

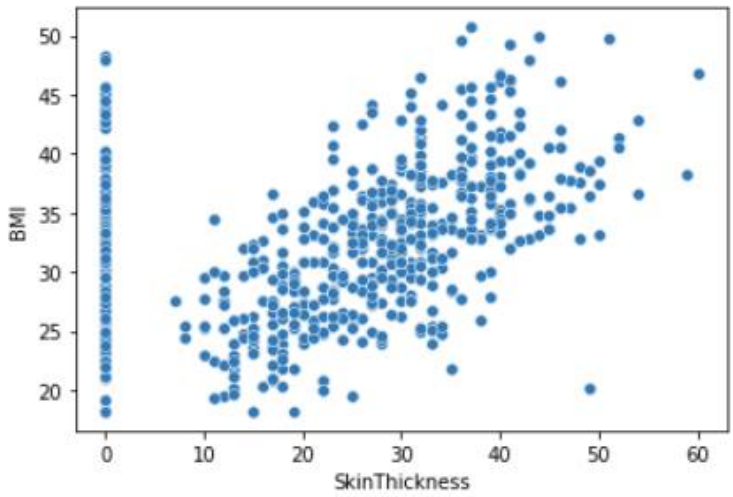
The correlation of dataset is show as below.

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



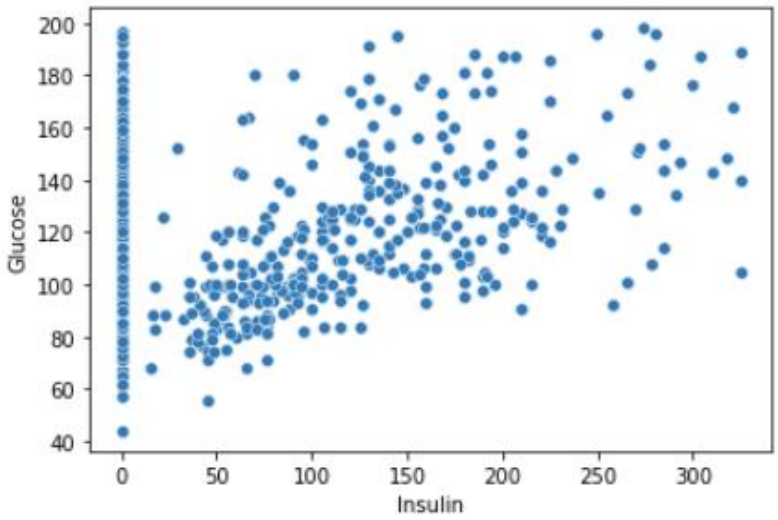
*5) **SkinThickness** and **BMI**.*

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:

```
df = data.drop(['Outcome'], axis=1)
```

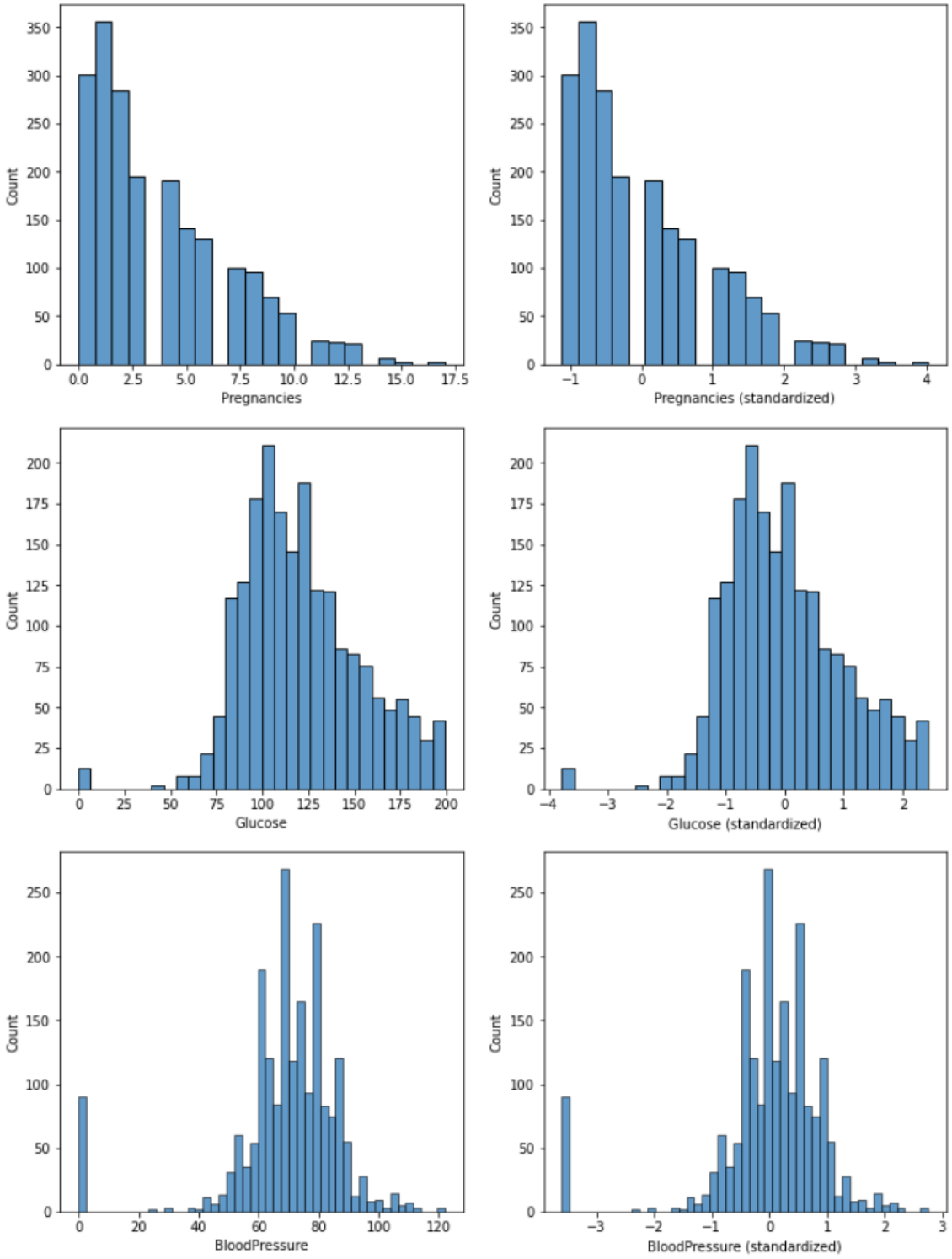
```
df.mean()
```

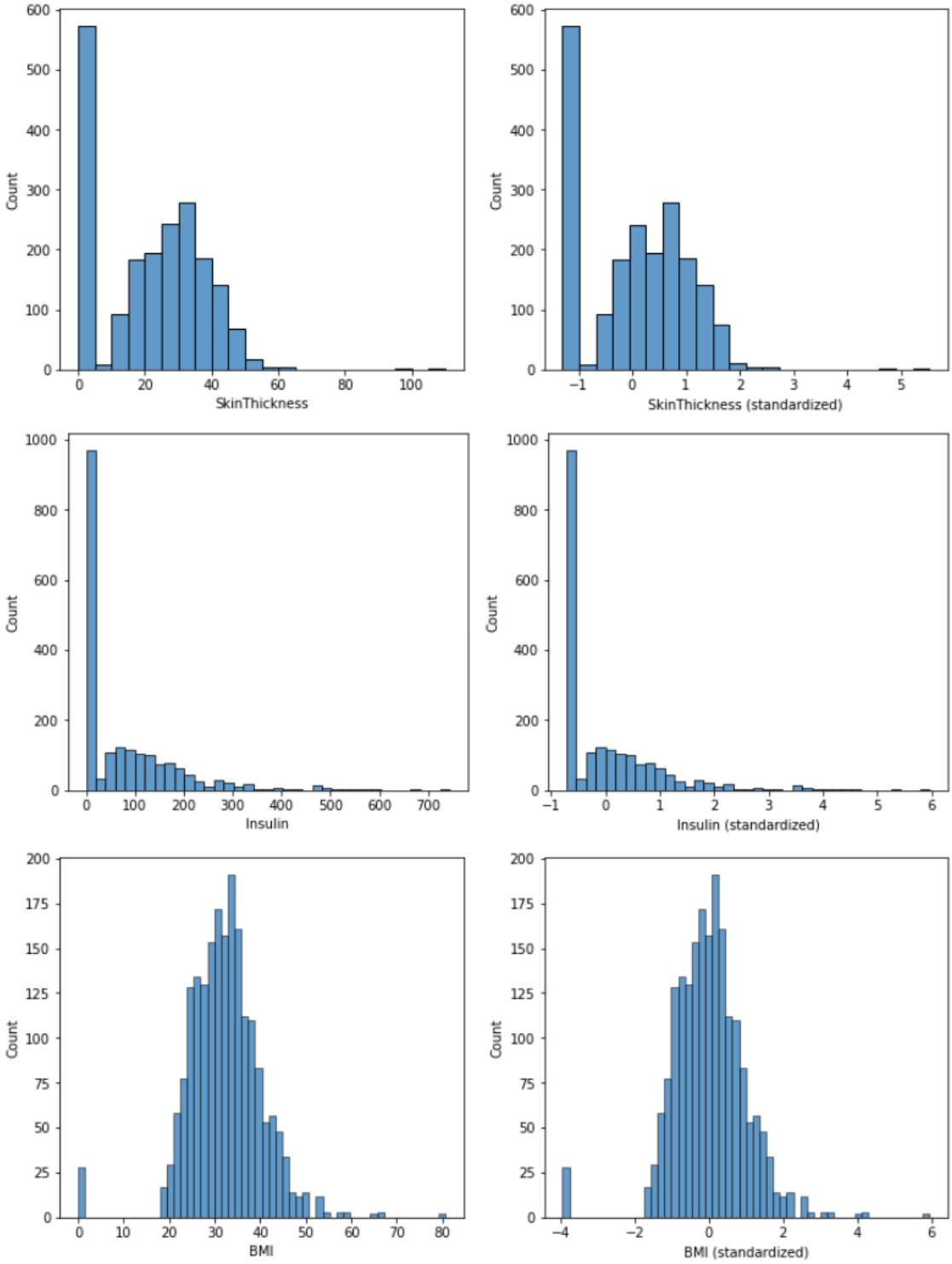
```
Pregnancies                  3.70350
Glucose                    121.18250
BloodPressure               69.14550
SkinThickness               20.93500
Insulin                     80.25400
BMI                         32.19300
DiabetesPedigreeFunction     0.47093
Age                         33.09050
dtype: float64
```

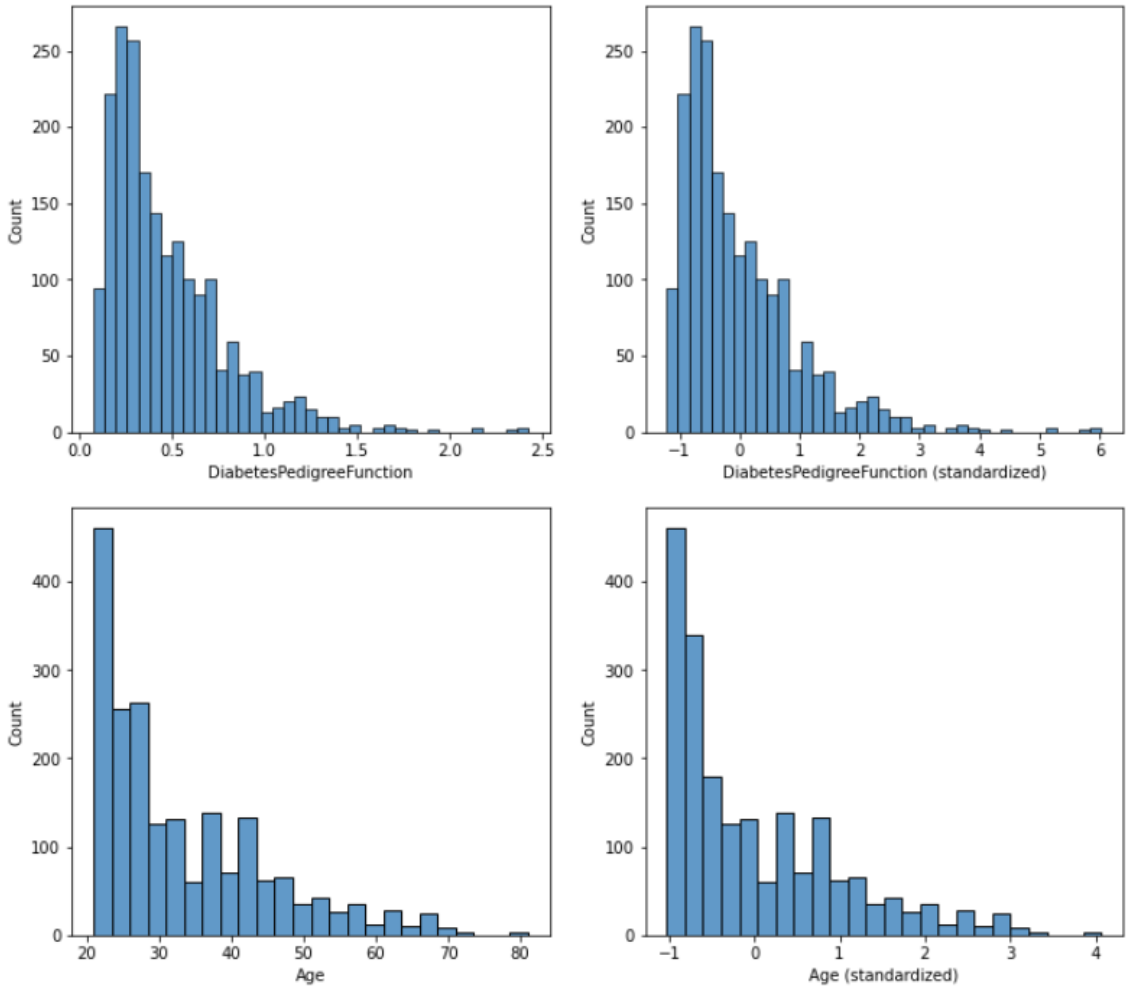
```
df.std()
```

```
Pregnancies                  3.306063
Glucose                     32.068636
BloodPressure               19.188315
SkinThickness               16.103243
Insulin                    111.180534
BMI                          8.149901
DiabetesPedigreeFunction     0.323553
Age                         11.786423
dtype: float64
```

Compare histogram result between and after standardized:




Count
SkinThickness
SkinThickness (standardized)
Insulin
Insulin (standardized)
BMI
BMI (standardized)


Count
DiabetesPedigreeFunction
DiabetesPedigreeFunction (standardized)
Age
Age (standardized)