SWEET OR SOUR

Diabetes Prediction in women

Pearploy Chaicharoensin
6381278

TABLE OF CONTENTS

01 INTRODUCTION

02 GET TO KNOW OUR DATA

03 DATA CLEANING

04 EXPLORATORY DATA ANALYSIS

05 PREDICTION MODELS

06 DISCUSSION

01

INTRODUCTION

INTRODUCTION

DISCLAIMER

Initially, my research proposal focused on predicting early symptoms of diabetes. However, further investigation led me to realize that my original topic was too broad, prompting me to narrow it down. Therefore, I decided to shift the focus of my study to predicting diabetes specifically in females.

PROBLEM STATEMENT

Diabetes is one of the most common human diseases and has become a significant public health concern around the world. The research from Centers for Disease Control and Prevention (2022) found that diabetes increases the risk of heart disease (the most common diabetes complication) by about four times in women but only about two times in men.

INTRODUCTION

CONTEXT

This project will drive deep in Diabetes Prediction in Females based on certain diagnostic. That is because when detected early, may prevent the progression of the disease and avoid other complications.

This project uses the combination of data science and machine learning, where I will analyze data and test in two of prediction models, which are logistic regression and random forest. At the end, I will discuss about what features can be used in the prediction and what kind of prediction model is better.

02

GET TO KNOW OUR DATA

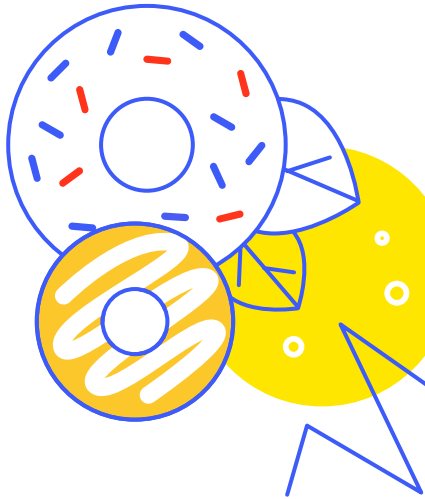
DOWNLOAD DATASET

from the National Institute of Diabetes and Digestive and Kidney Diseases

```
file = '/Users/mudmi/Desktop/Data Sci/datasci project/diabetes.csv'
df = pd.read_csv(file)
df.head()
```

	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

THE PROVIDED FEATURES



Pregnancies Number of times pregnant

Glucose Plasma glucose concentration a 2 hours in an oral glucose tolerance test

Blood Pressure Diastolic blood pressure (mm Hg)

Skin Thickness Triceps skin fold thickness (mm)

Insulin 2-Hour serum insulin (mu U/ml)

BMI Body mass index (weight in kg/(height in m)^2)

Diabetes Pedigree Function A mathematical formula used to calculate the risk of diabetes in individuals based on their family history

Age Age (years)

Outcome Class variable (0 or 1) whether the patient is diabetic or not

03

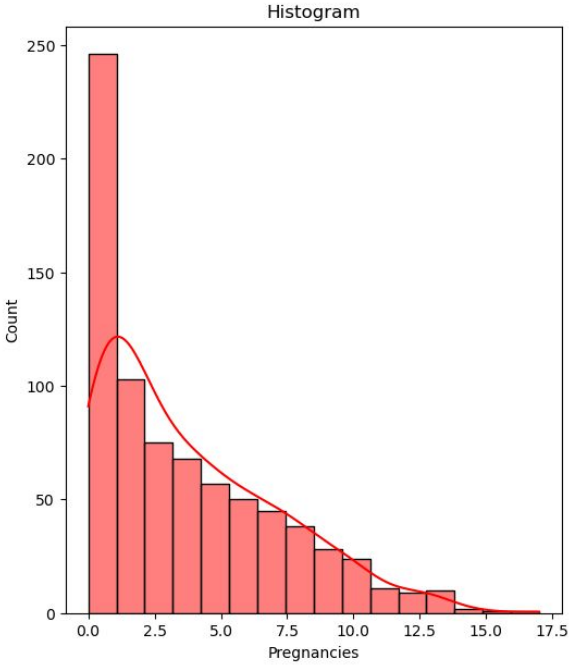
DATA CLEANING

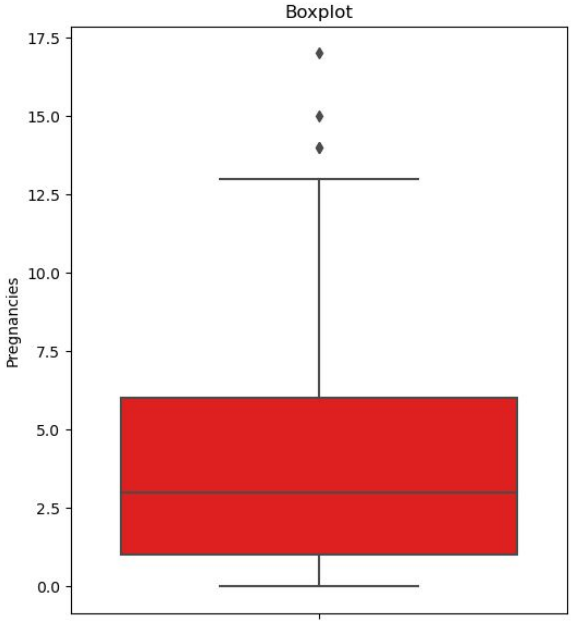
Drop duplicated and check null values

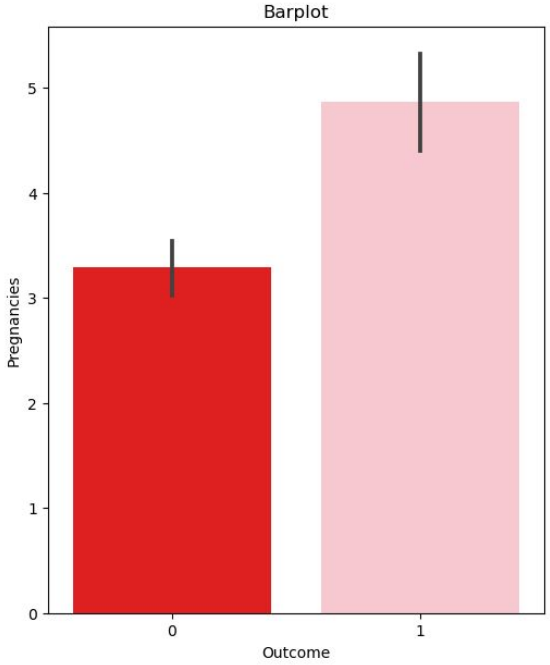
```
df = df.drop_duplicates()
df.isnull().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Pregnancies

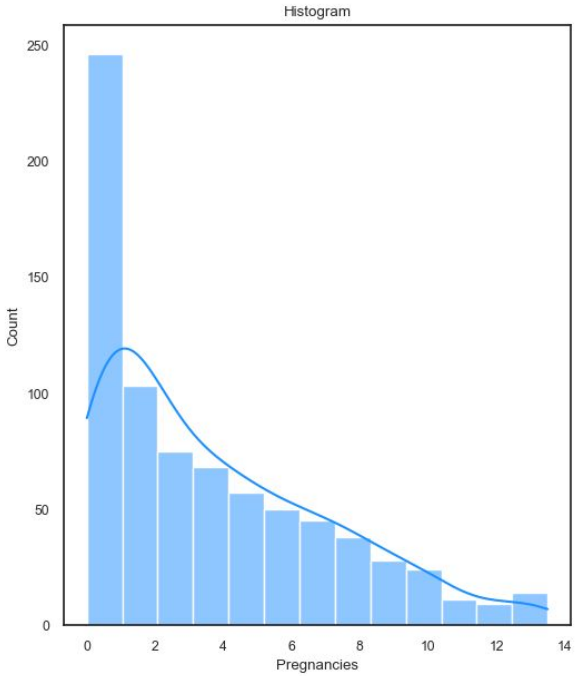


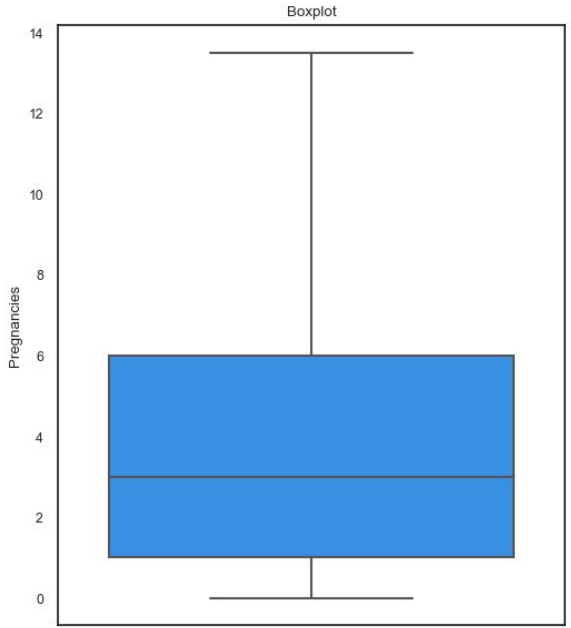


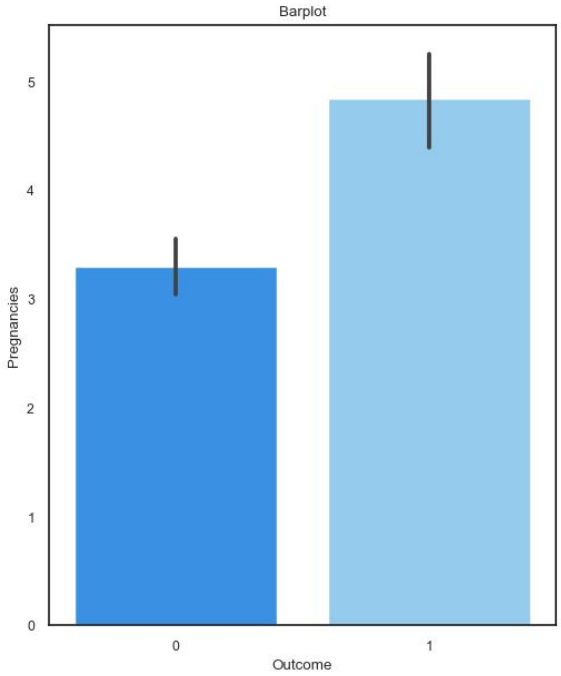


Outlier Boxplot Discussion: it's possible that a woman can give birth up to 17 children but pretty rare

Pregnancies

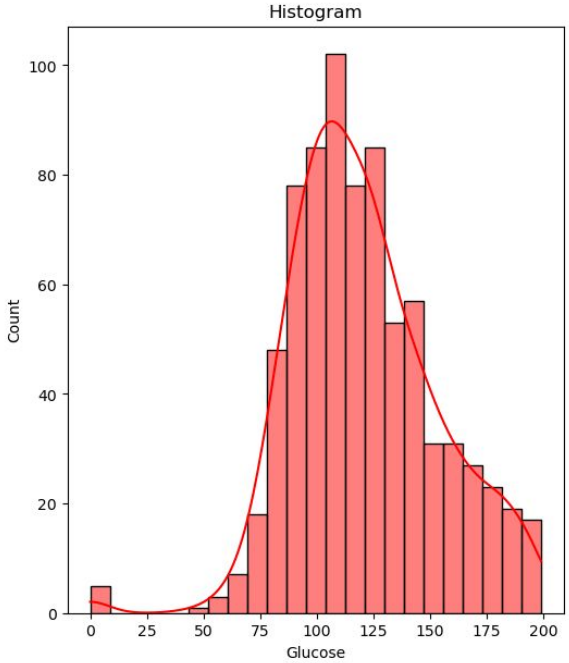


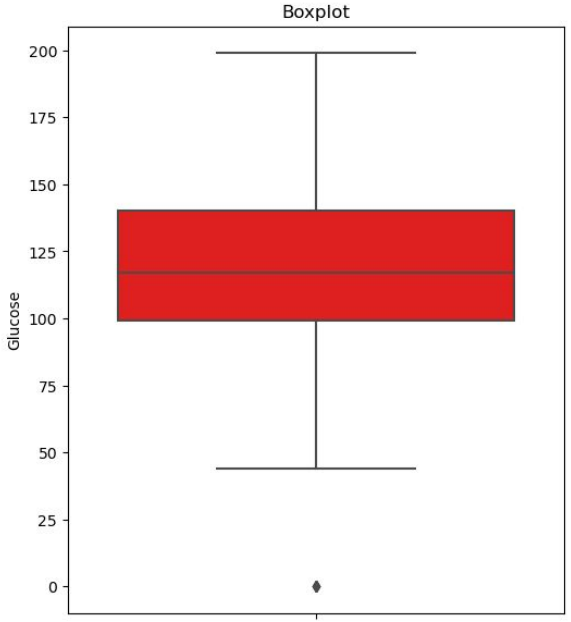


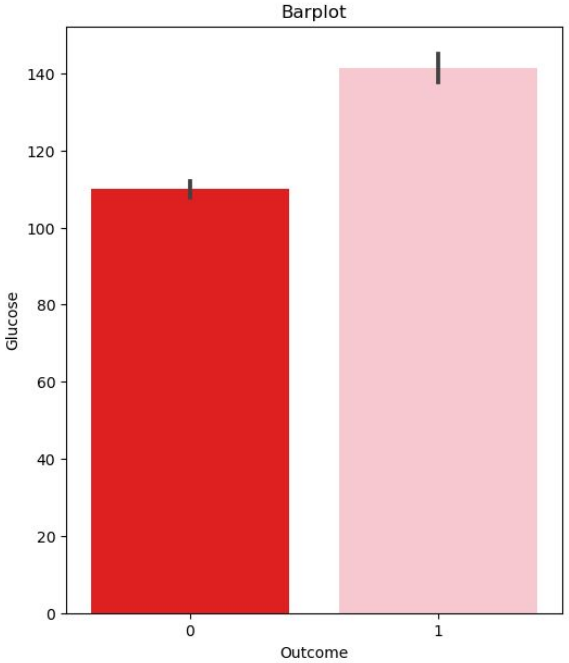


What has changed?: outlier removed

Glucose

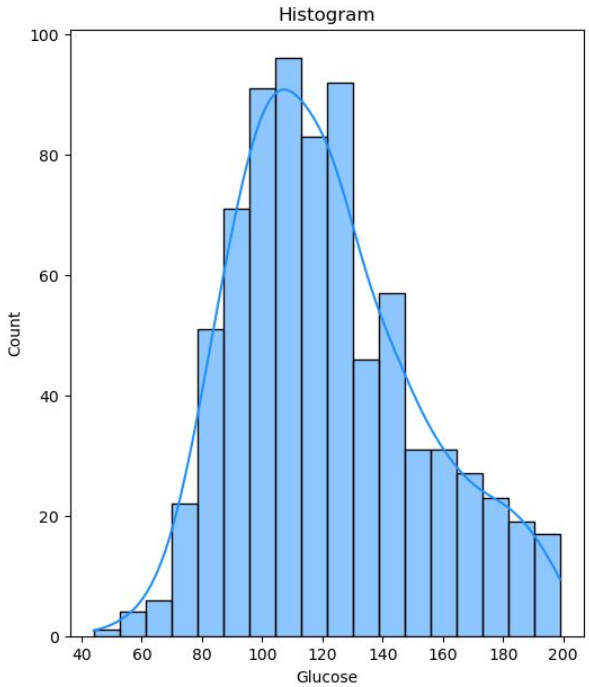


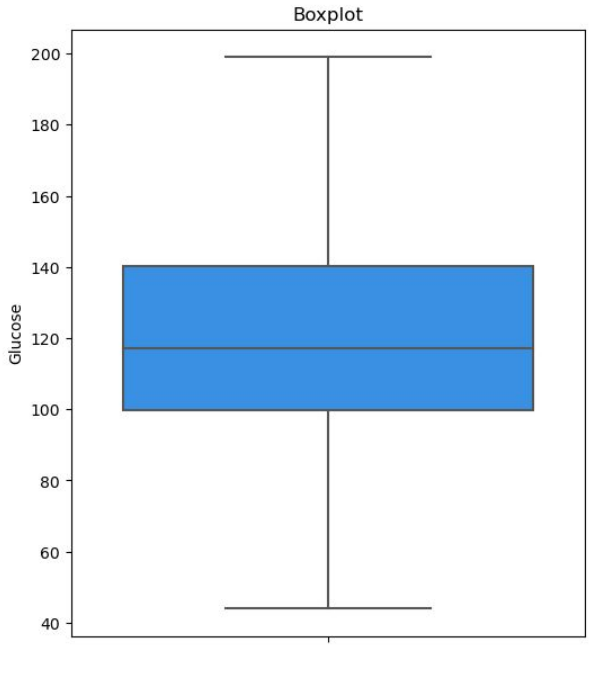


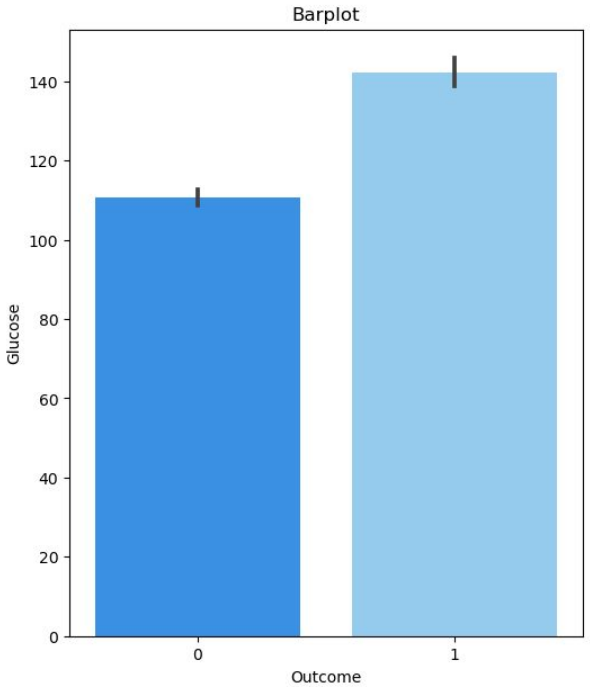


Outlier Boxplot Discussion: glucose level can't be zero

Glucose

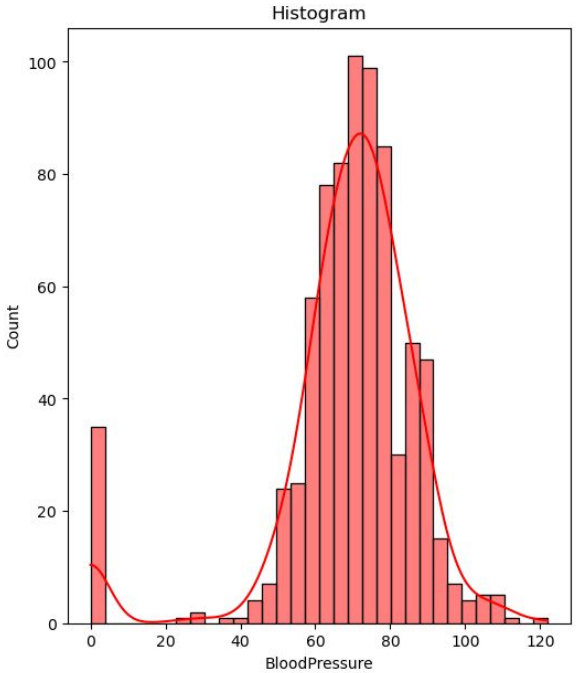


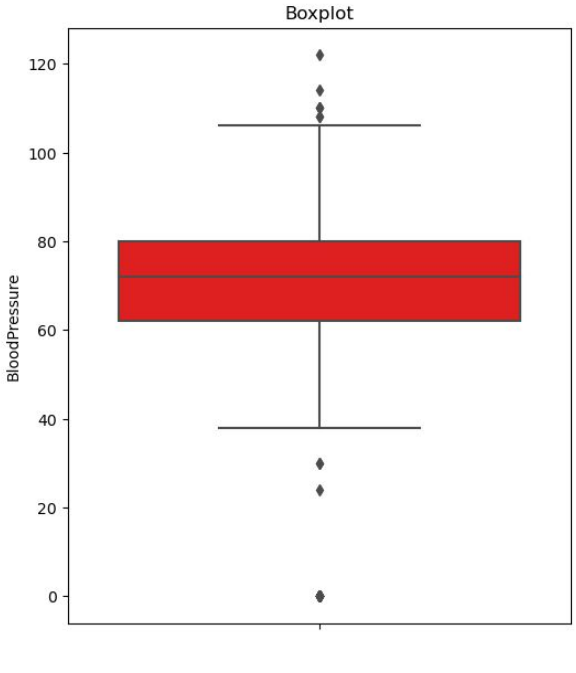


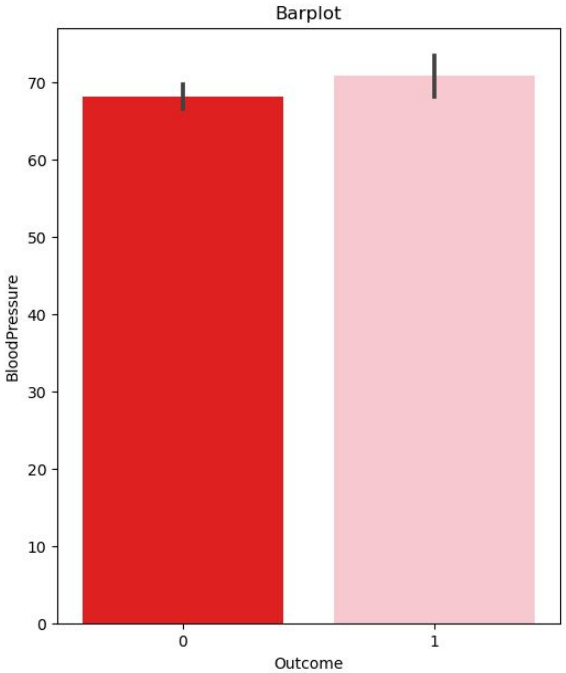


What has changed?: value zero replaced with its median

Blood Pressure

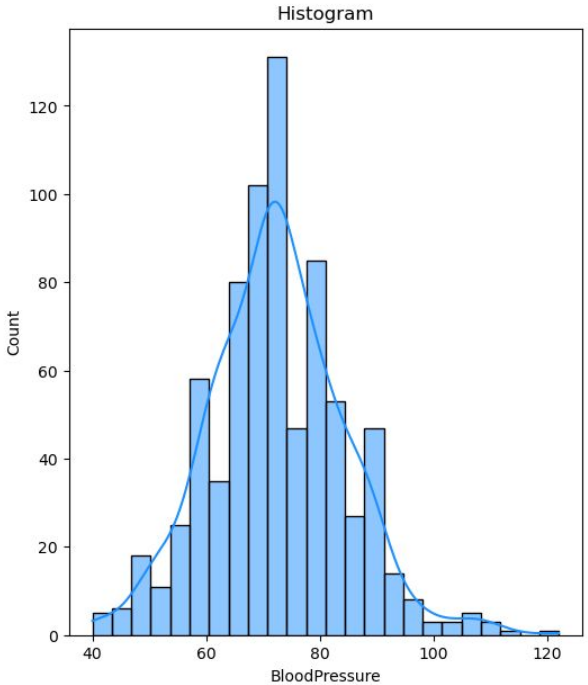


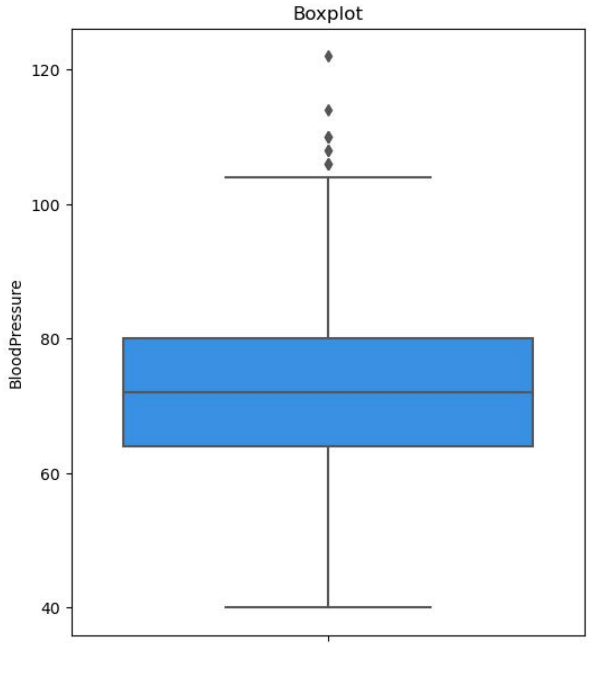


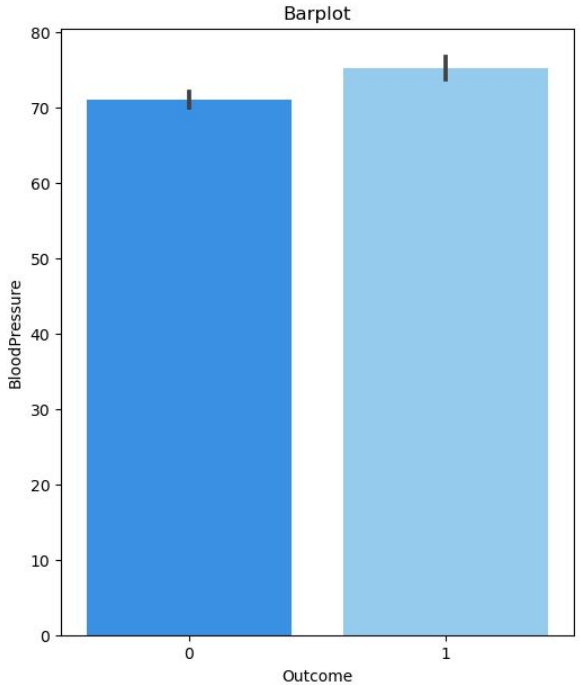


Outlier Boxplot Discussion: it's possible to have high blood pressure but not less than 40 and not equal to 0

Blood Pressure

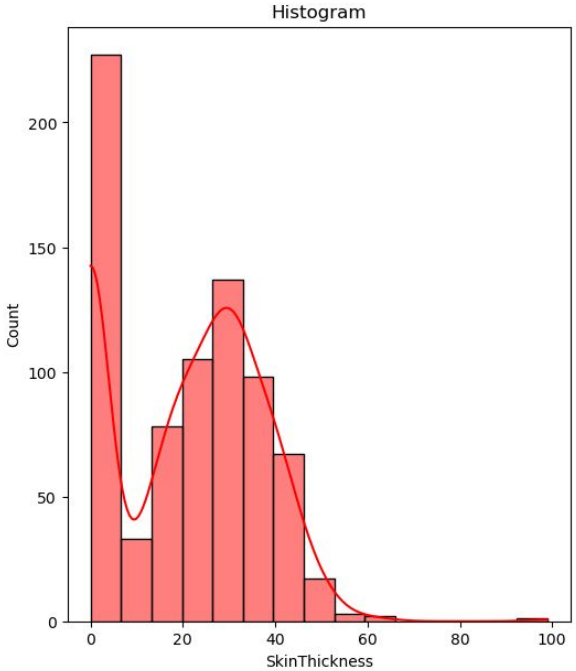


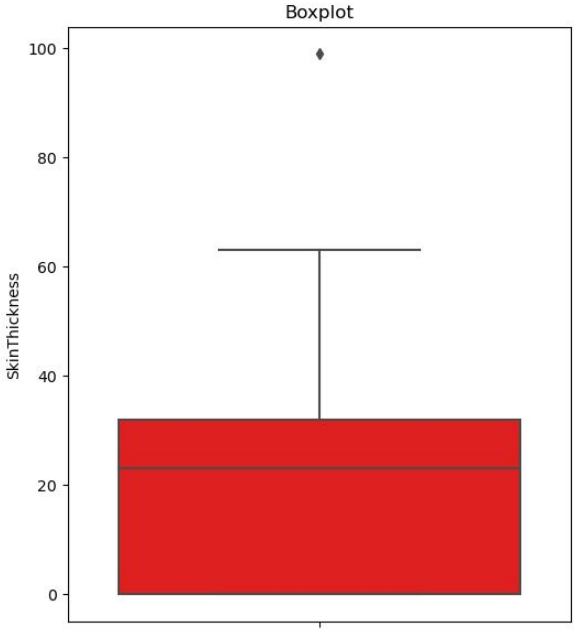


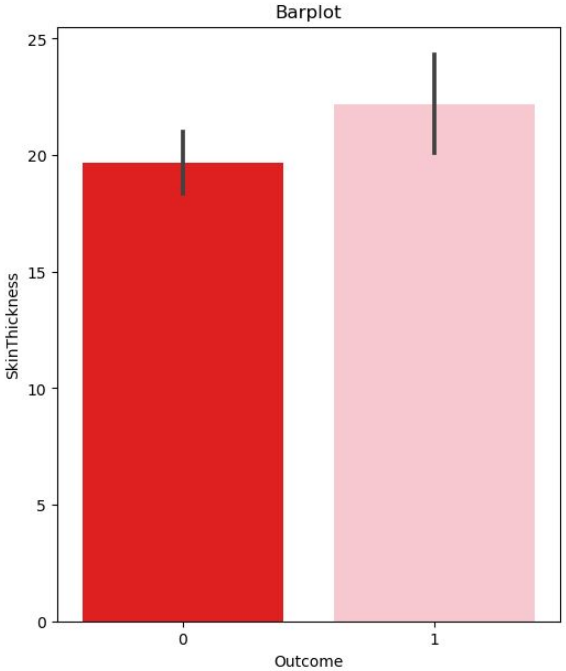


What has changed?: replaced value 0 to its median and removed outlier that lower than lower quartile

Skin Thickness

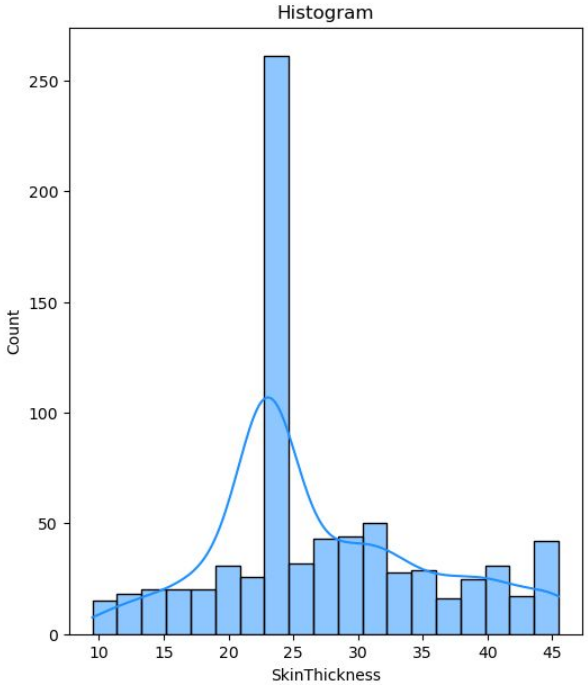


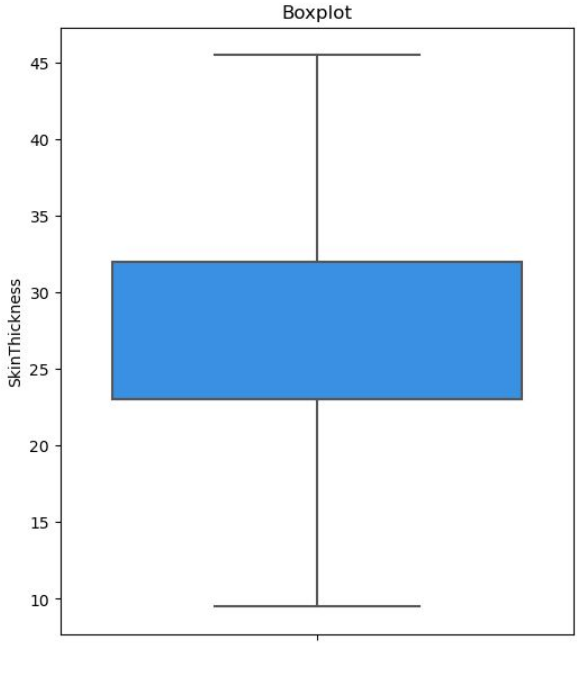


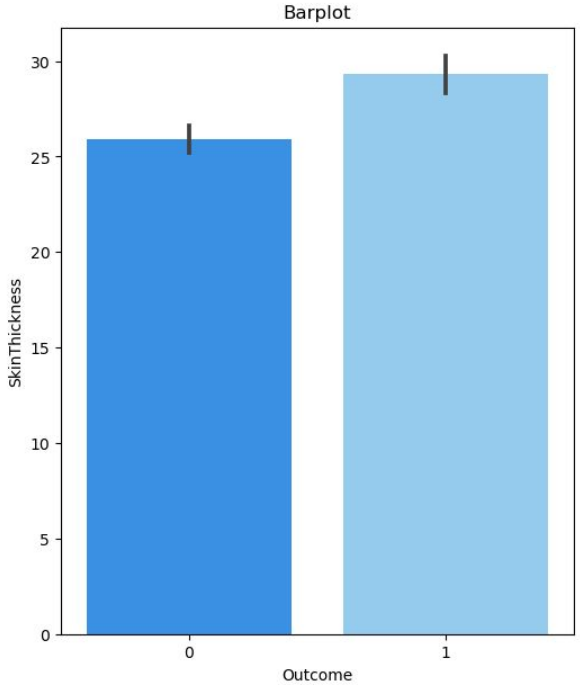


Outlier Boxplot Discussion: the outlier too far from upper quartile and can not be zero

Skin Thickness

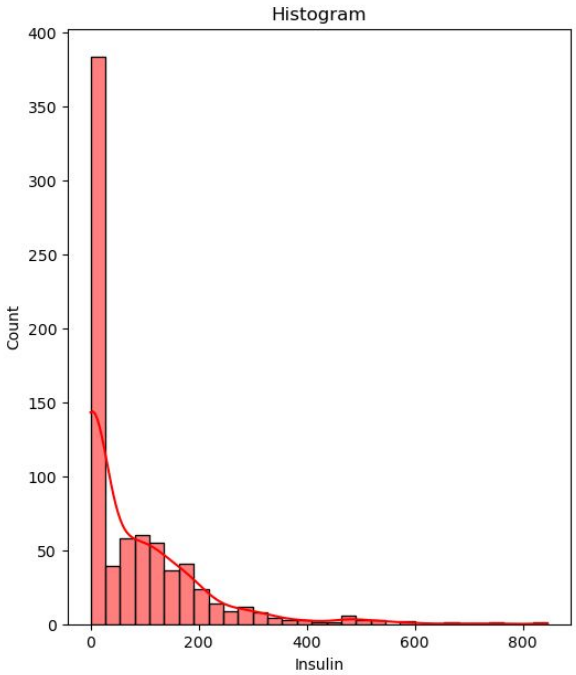


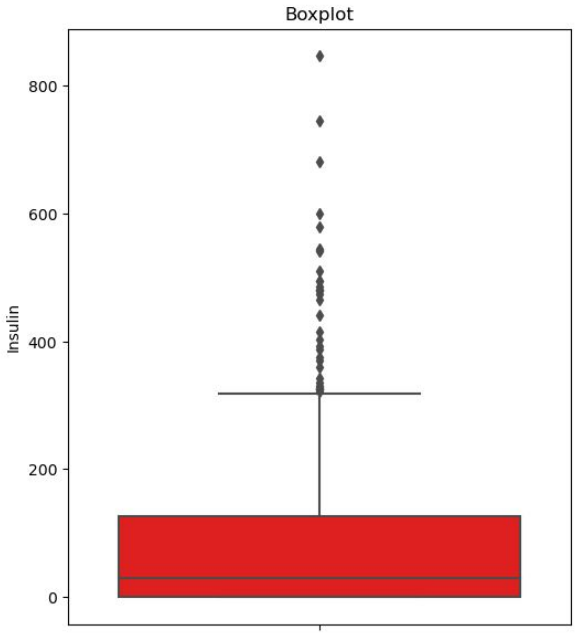


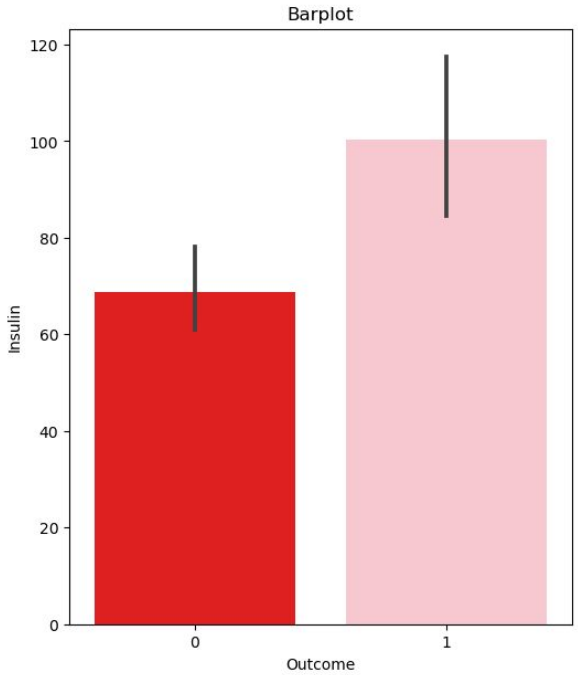


What has changed?: outliers are removed and value zero are replaced with its median

Insulin

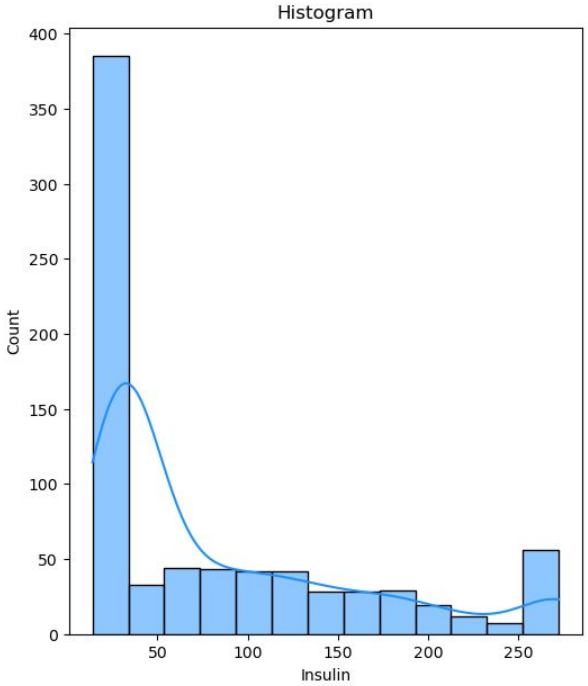


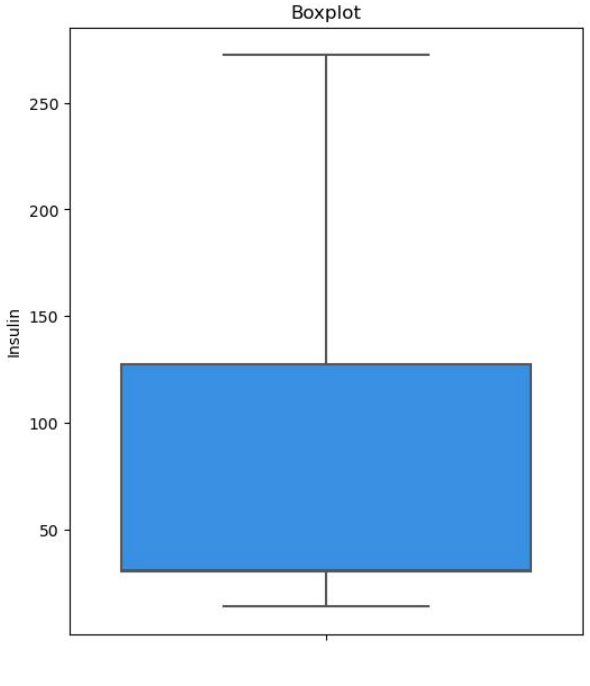


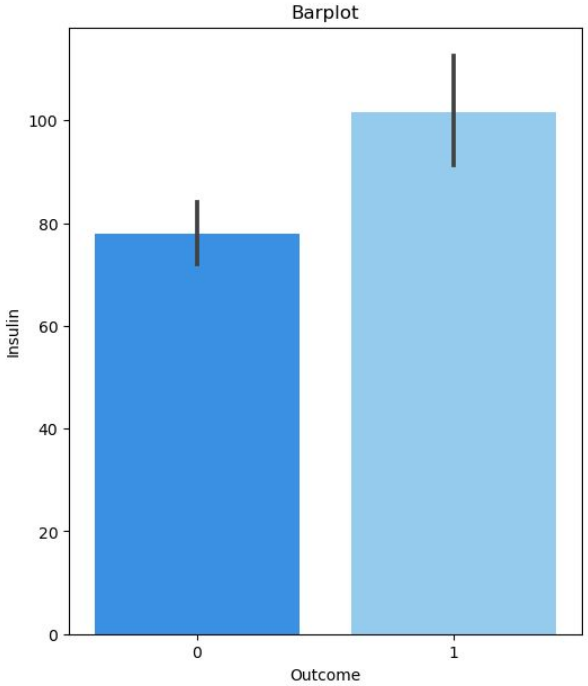


Outlier Boxplot Discussion: should be no more than 400 mIU/L and can not be zero

Insulin

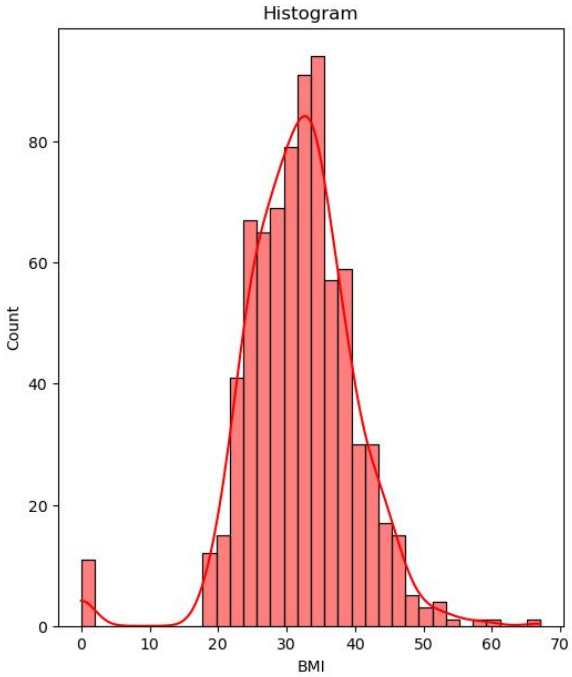


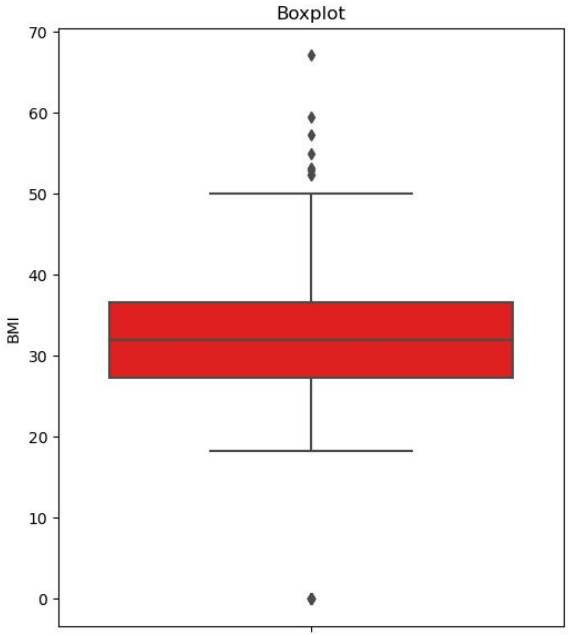


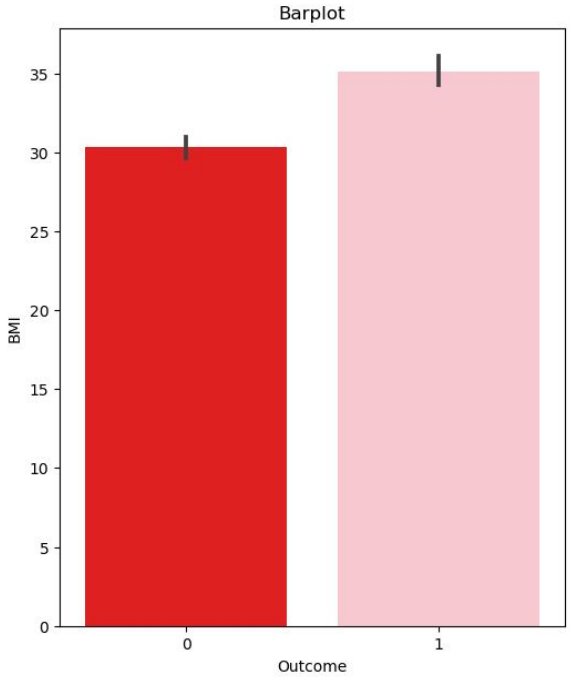


What has changed?: outliers are removed and value zero are replaced with its median

BMI

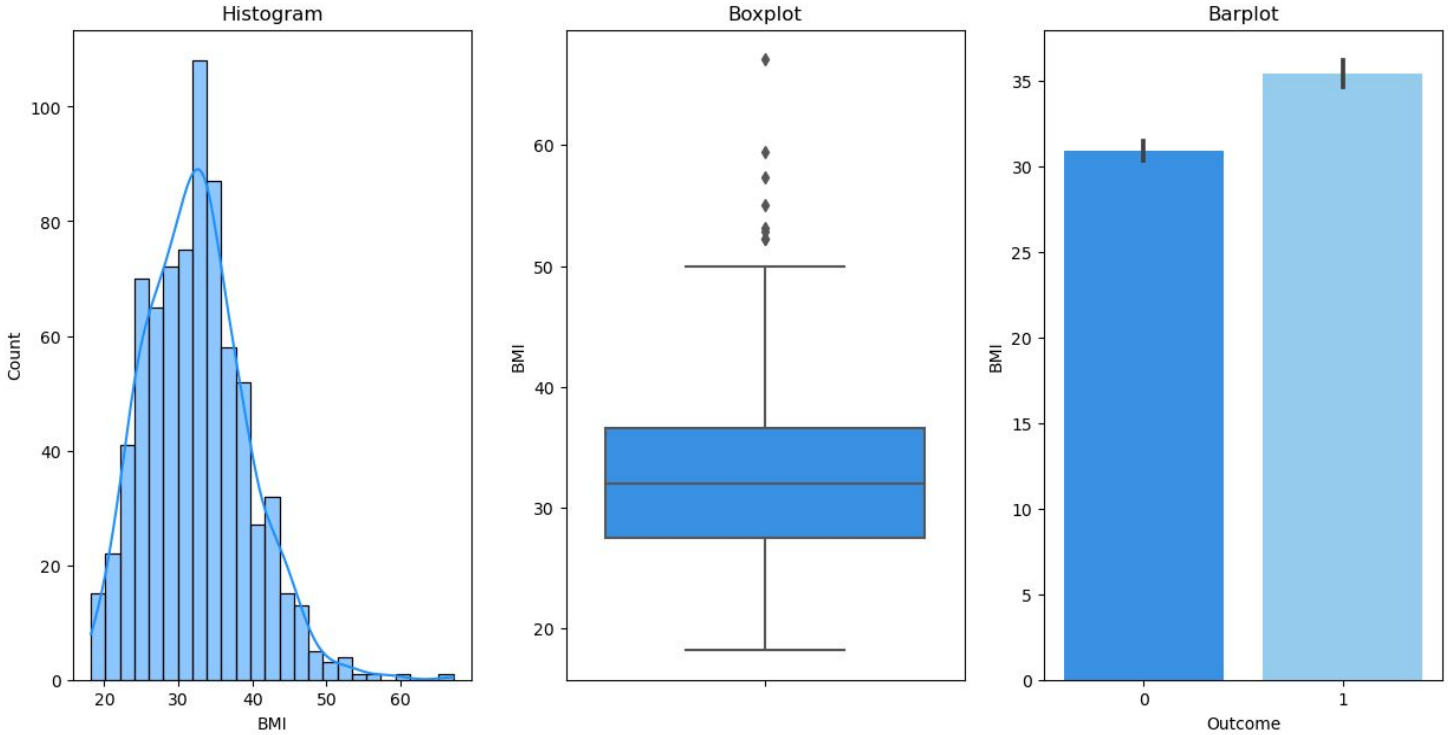






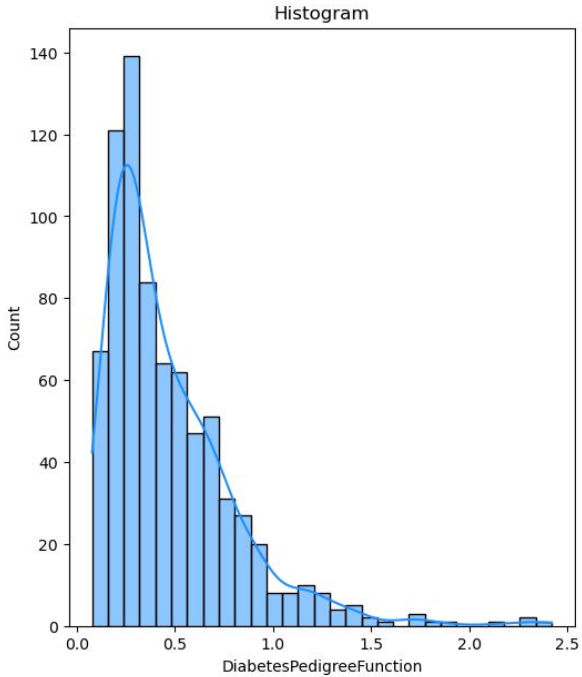
Outlier Boxplot Discussion: can not equal to zero

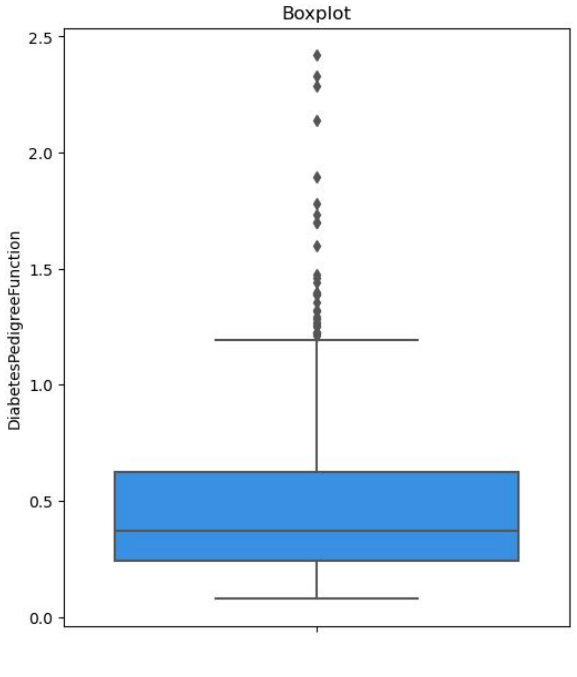
BMI

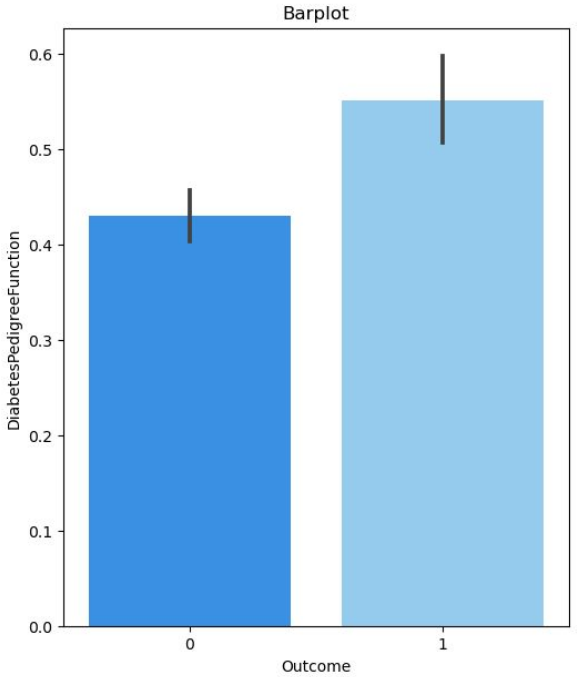


What has changed?: value zero replaced with its median

Diabetes Pedigree Function

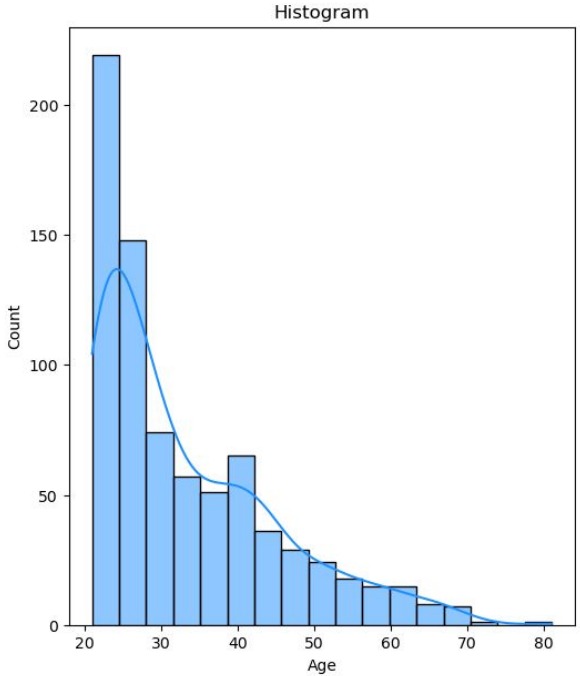


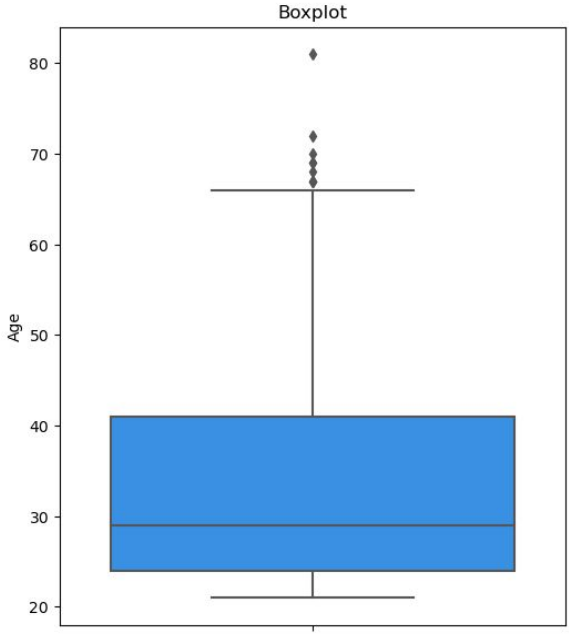


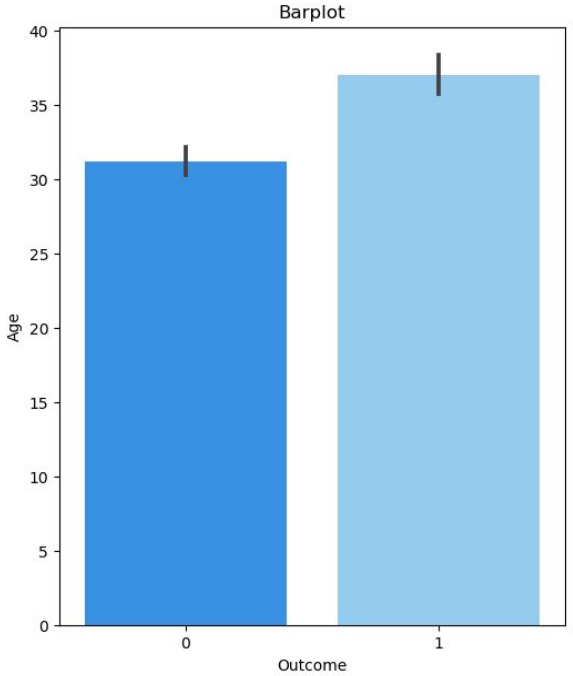


Outlier Boxplot Discussion: it depends on family background

Age

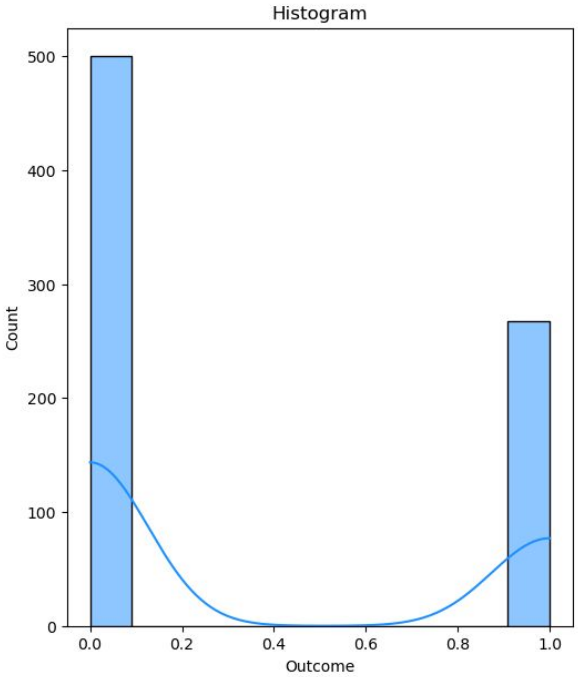


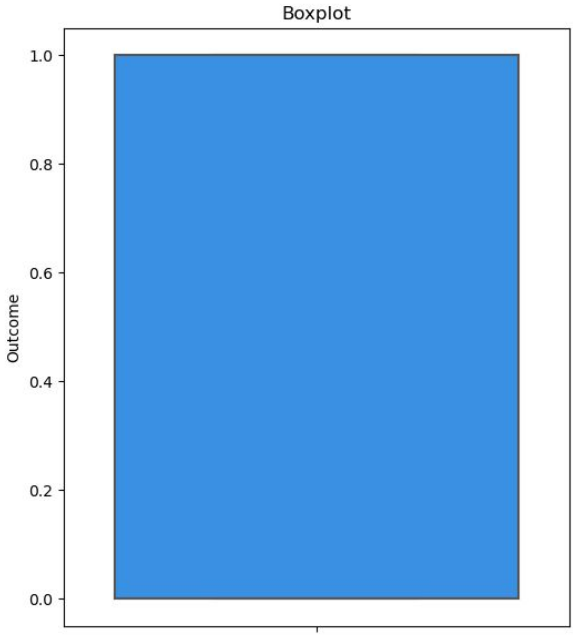


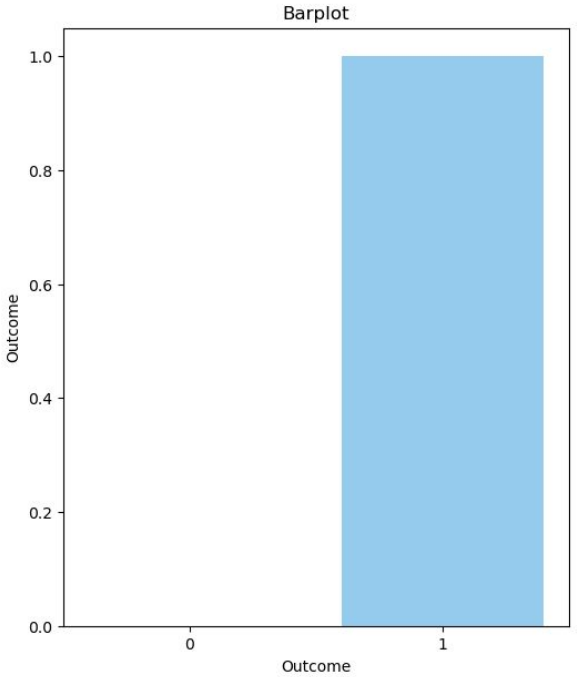


Outlier Boxplot Discussion: it is possible to age up to 80

Outcome


Histogram
Count
Outcome
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0

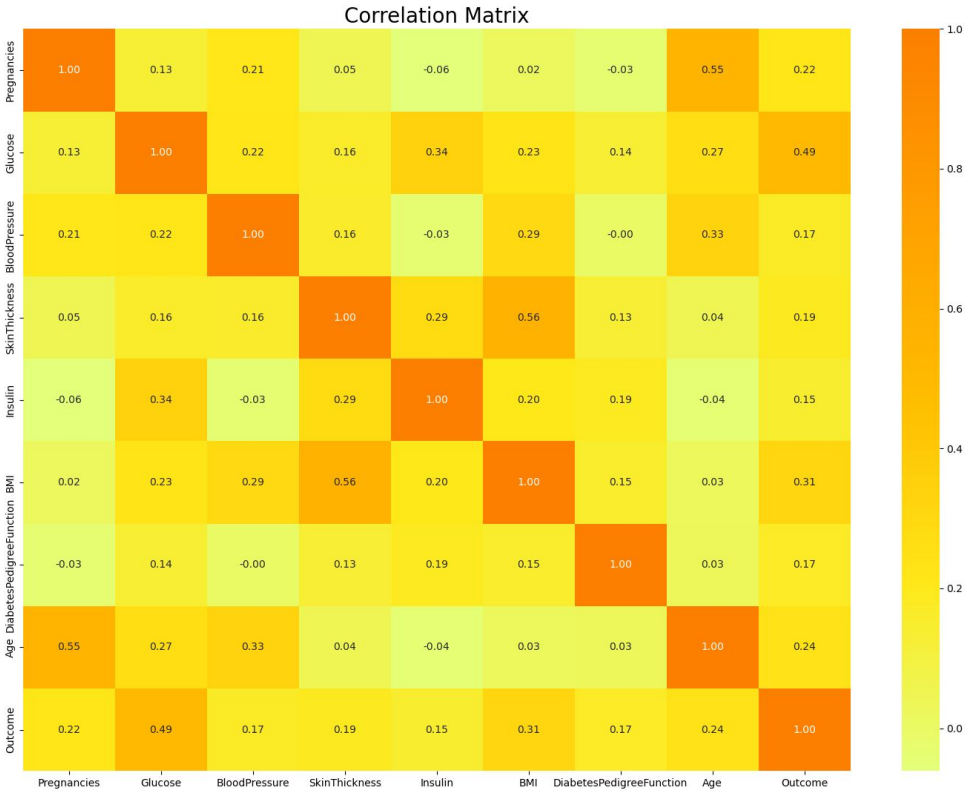

Boxplot
Outcome
0.0
0.2
0.4
0.6
0.8
1.0


Barplot
Outcome
0.0
0.2
0.4
0.6
0.8
1.0
0
1
Outcome

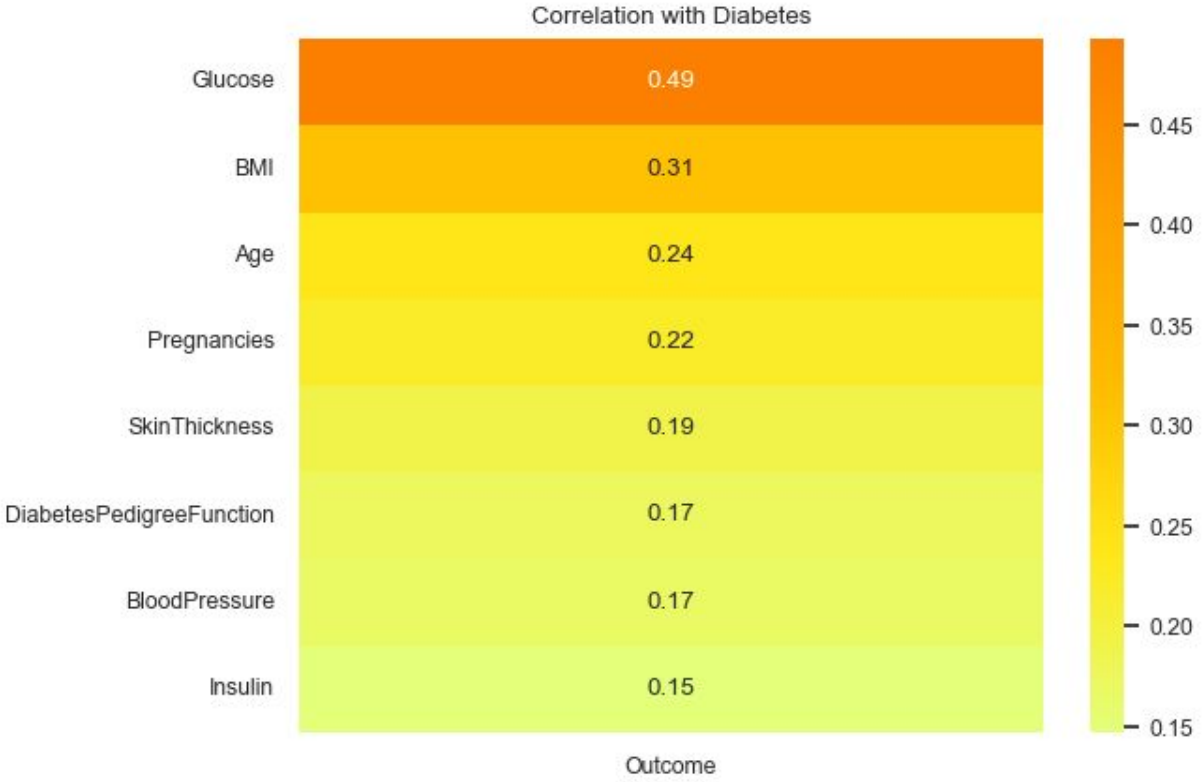
04

EXPLORATORY DATA ANALYSIS

Correlation of Features to All



Correlation of Features to the

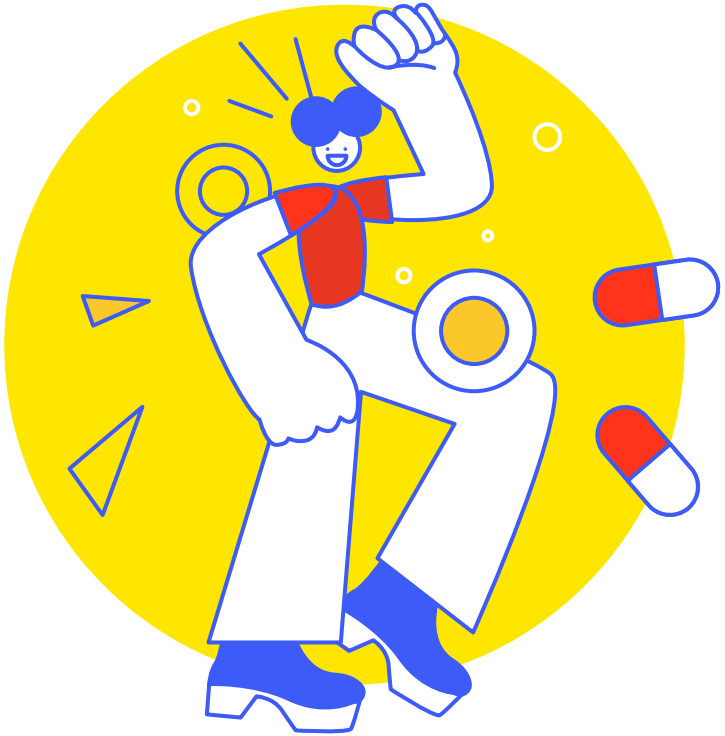


Pick and Choose Features

Glucose, BMI, Age, Pregnancies

Dep. Variable:	Outcome	**No. Observations:**	768
Model:	Logit	**Df Residuals:**	764
Method:	MLE	**Df Model:**	3
Date:	Tue, 11 Jul 2023	**Pseudo R-squ.:**	0.01631
Time:	09:26:49	**Log-Likelihood:**	-488.64
converged:	True	**LL-Null:**	-496.74
Covariance Type:	nonrobust	**LLR p-value:**	0.001031

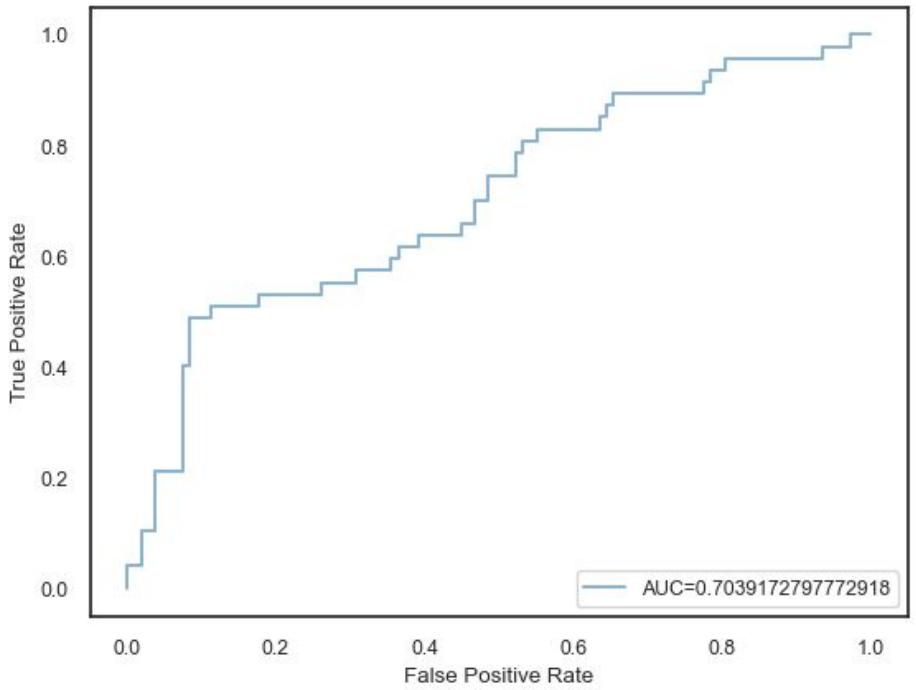
	coef	std err	z	P>|z|	[0.025	0.975]
Glucose	0.0135	0.003	5.248	0.000	0.008	0.019
BMI	-0.0484	0.009	-5.322	0.000	-0.066	-0.031
Age	-0.0303	0.008	-3.760	0.000	-0.046	-0.015
Pregnancies	0.1195	0.028	4.283	0.000	0.065	0.174



05

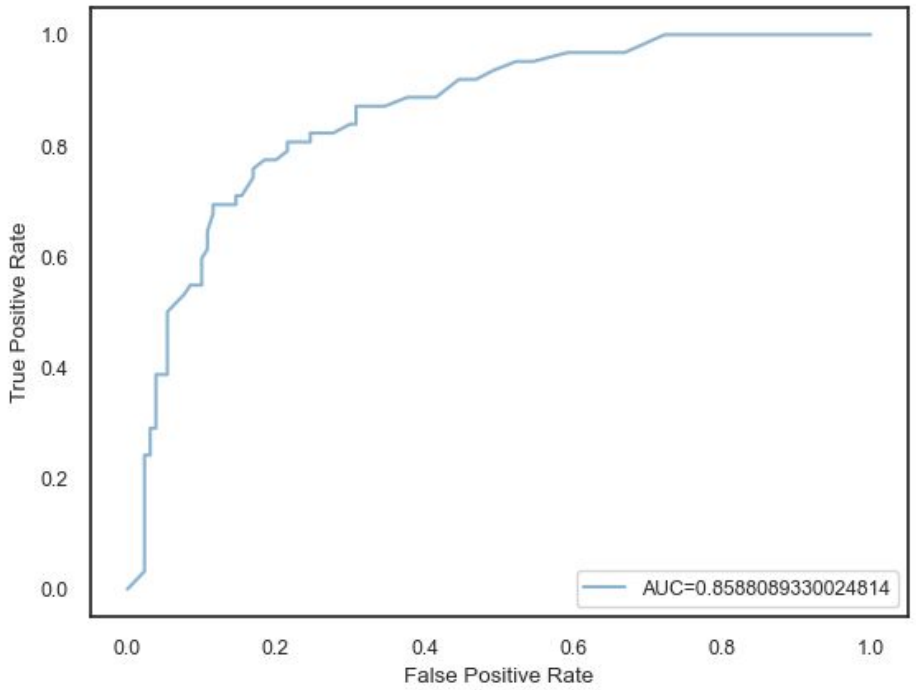
PREDICTION MODELS

Logistic Regression Model



AUC = 0.70

Random Forest Model



AUC = 0.85

06

DISCUSSION

Why Random Forest Model is Better?

Model Complexity

- Random Forest can capture nonlinear relationships better.

Feature Importance

- Random Forest automatically selects important features, while Logistic Regression treats all features equally.

Outliers

- Logistic Regression is sensitive to outliers, impacting its performance more than Random Forest.

How to enhance the prediction?

- The outcome would be more accurate with a dataset filled with accurate information.
- Many values are zero, indicating missing patient values.
- I had to decide whether to remove outliers and replace zero values with their median or remain the same, which might affect the prediction accuracy

Conclusion

This project uses only Glucose, BMI, Age and Pregnancies as the features to predict diabetes classification on female patients. Also, the preferred prediction model is Random Forest with AUC equals to 85%.