

Data Science Final Project-Group4

108753205 資科碩一 蔡宗諺
108753208 資科碩一 葉冠宏
107354018 統計碩二 蔡承軒



Table of Contents

- **Dataset**
- **Modeling**
- **Performance**
- **Demo**



Goal

- The objective of the dataset is to diagnostically predict whether or not a patient has diabetes.
- 預測患者是否有糖尿病



Dataset



Dataset


- **Pima Indians Diabetes Database-Predict the onset of diabetes based on diagnostic measures**







Datasource

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>


 Dataset



 1373

Pima Indians Diabetes Database

Predict the onset of diabetes based on diagnostic measures

 UCI Machine Learning • updated 4 years ago (Version 1)



INTRODUCTION

- **Datasets is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.**
- **All patients here are females at least 21 years old of Pima Indian heritage.**



Data format

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30.0	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1



Data Features

Variables	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)

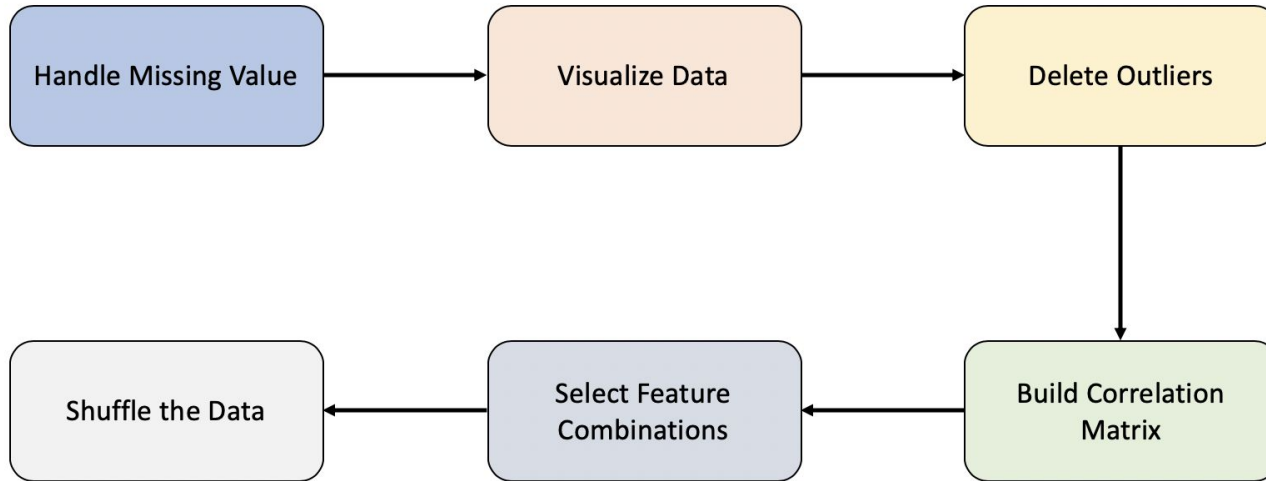


Str()

```
$ Pregnancies      : num [1:768] 6 1 8 1 0 5 3 10 2 8 ...
$ Glucose          : num [1:768] 148 85 183 89 137 116 78 115 197 125 ...
$ BloodPressure    : num [1:768] 72 66 64 66 40 74 50 0 70 96 ...
$ SkinThickness    : num [1:768] 35 29 0 23 35 0 32 0 45 0 ...
$ Insulin          : num [1:768] 0 0 0 94 168 0 88 0 543 0 ...
$ BMI              : num [1:768] 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
$ DiabetesPedigreeFunction: num [1:768] 0.627 0.351 0.672 0.167 2.288 ...
$ Age              : num [1:768] 50 31 32 21 33 30 26 29 53 54 ...
$ Outcome          : num [1:768] 1 0 1 0 1 0 1 0 1 1 ...
```



Preprocessing





Missing Value

```
> sum(is.na(diabetes))  
[1] 0
```

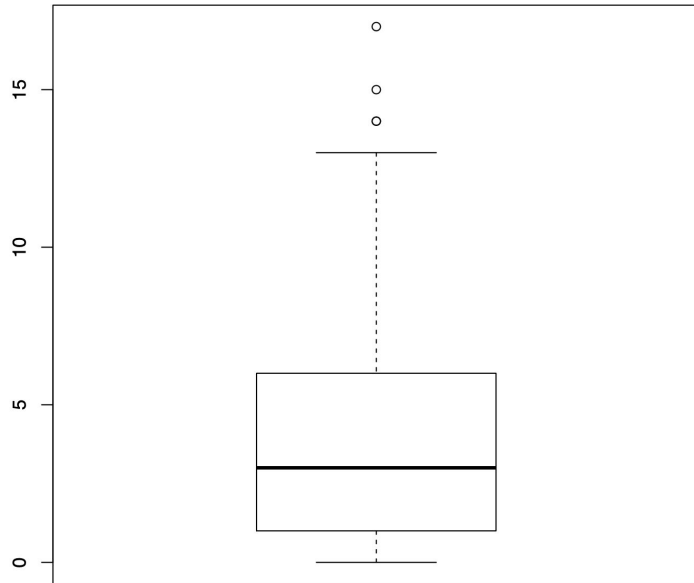


Summary()

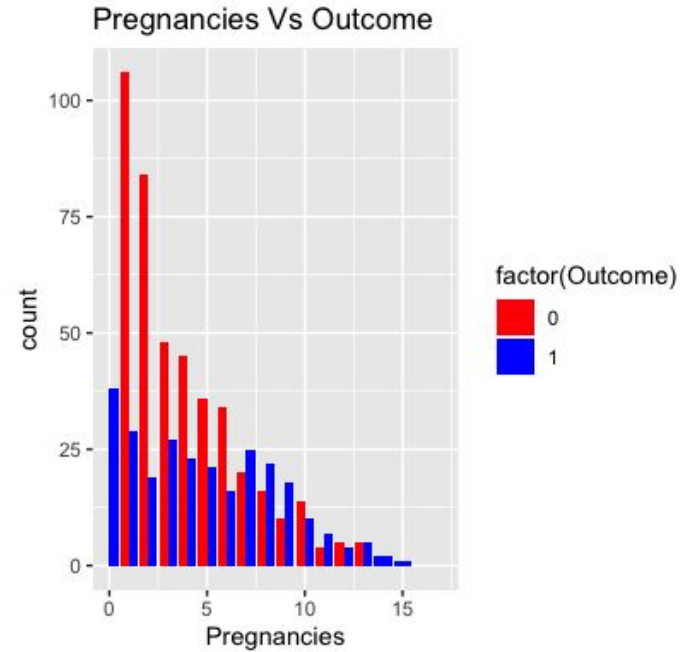
Pregnancies	Glucose	BloodPressure
Min. : 0.000	Min. : 0.0	Min. : 0.00
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 62.00
Median : 3.000	Median :117.0	Median : 72.00
Mean : 3.845	Mean :120.9	Mean : 69.11
3rd Qu.: 6.000	3rd Qu.:140.2	3rd Qu.: 80.00
Max. :17.000	Max. :199.0	Max. :122.00
SkinThickness	Insulin	BMI
Min. : 0.00	Min. : 0.0	Min. : 0.00
1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.:27.30
Median :23.00	Median : 30.5	Median :32.00
Mean :20.54	Mean : 79.8	Mean :31.99
3rd Qu.:32.00	3rd Qu.:127.2	3rd Qu.:36.60
Max. :99.00	Max. :846.0	Max. :67.10
DiabetesPedigreeFunction	Age	Outcome
Min. :0.0780	Min. :21.00	Min. :0.000
1st Qu.:0.2437	1st Qu.:24.00	1st Qu.:0.000
Median :0.3725	Median :29.00	Median :0.000
Mean :0.4719	Mean :33.24	Mean :0.349
3rd Qu.:0.6262	3rd Qu.:41.00	3rd Qu.:1.000
Max. :2.4200	Max. :81.00	Max. :1.000



Box-Plot



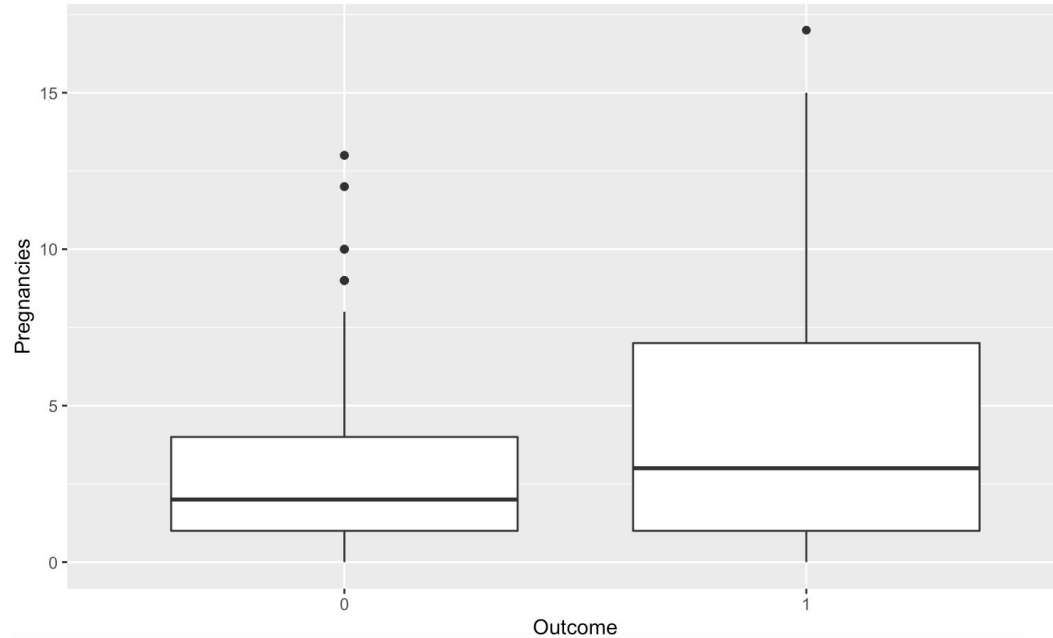
Pregnancies





Box-Plot

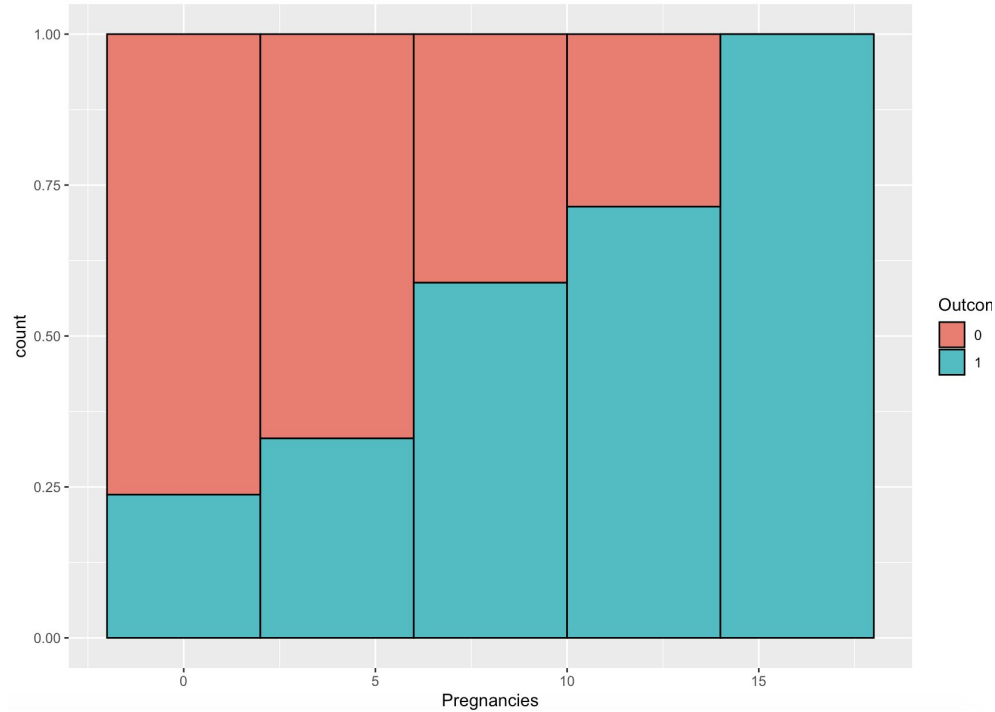
Pregnancies





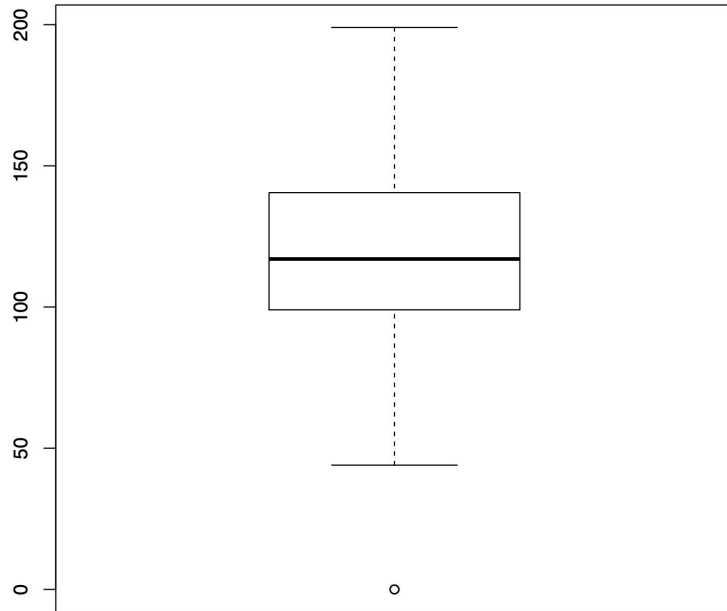
Box-Plot

Pregnancies

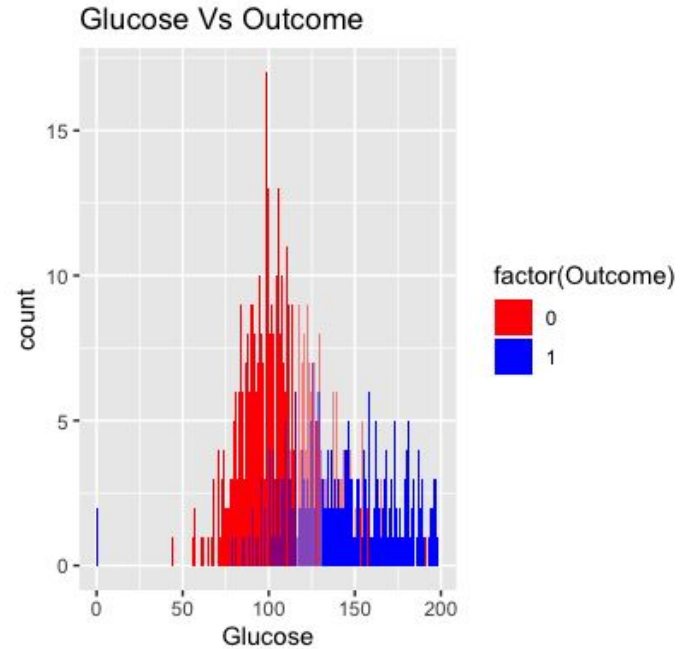




Box-Plot



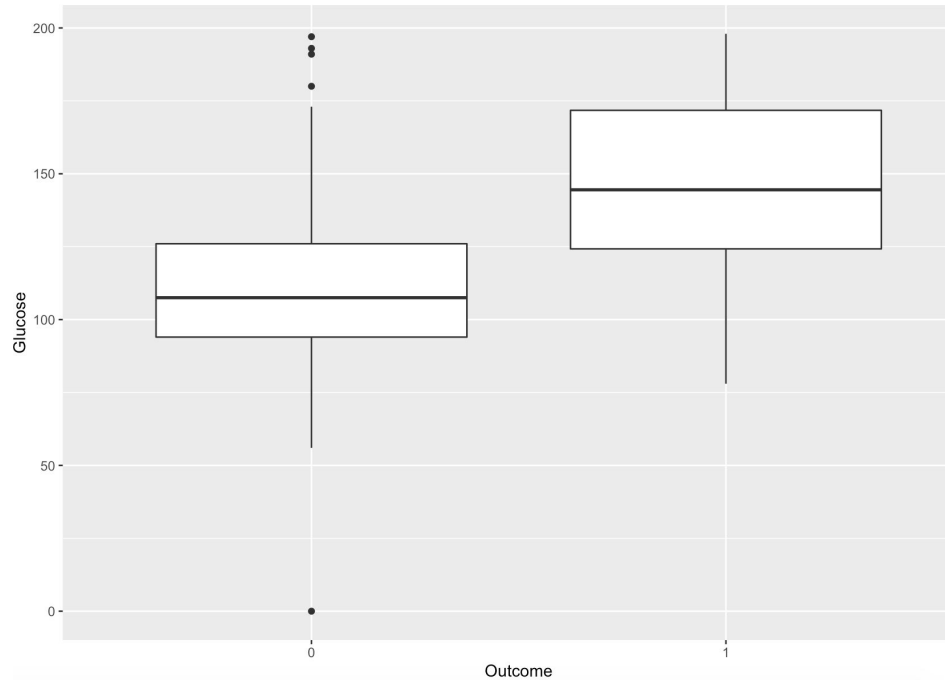
Glucose





Box-Plot

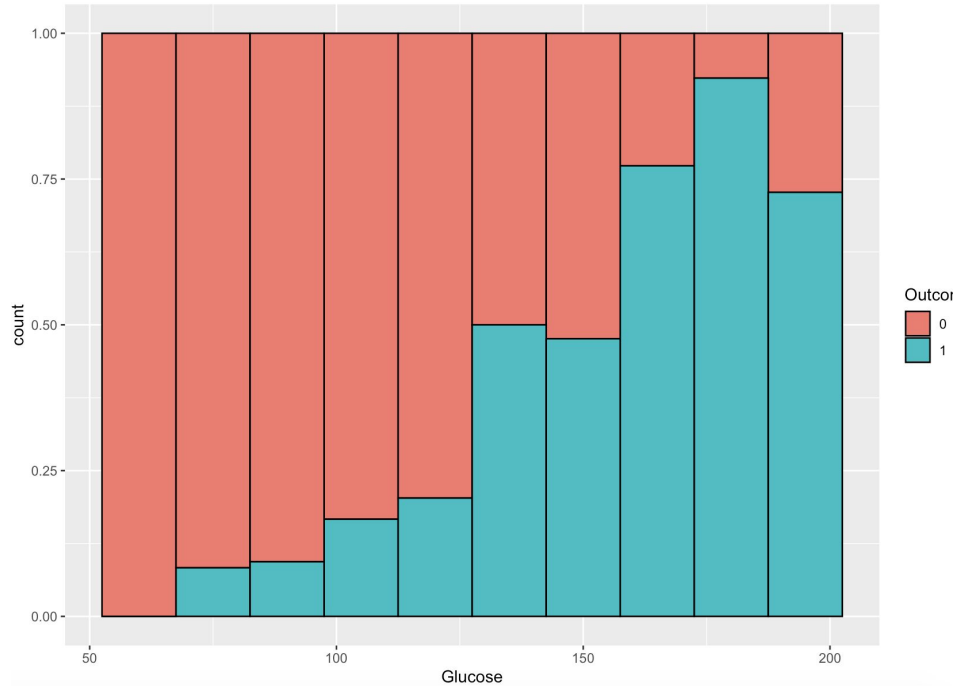
Glucose





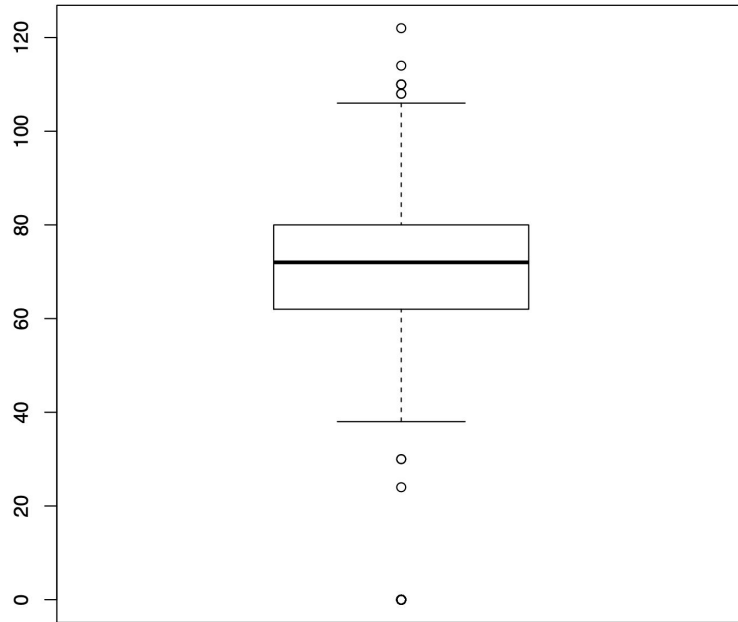
Box-Plot

Glucose

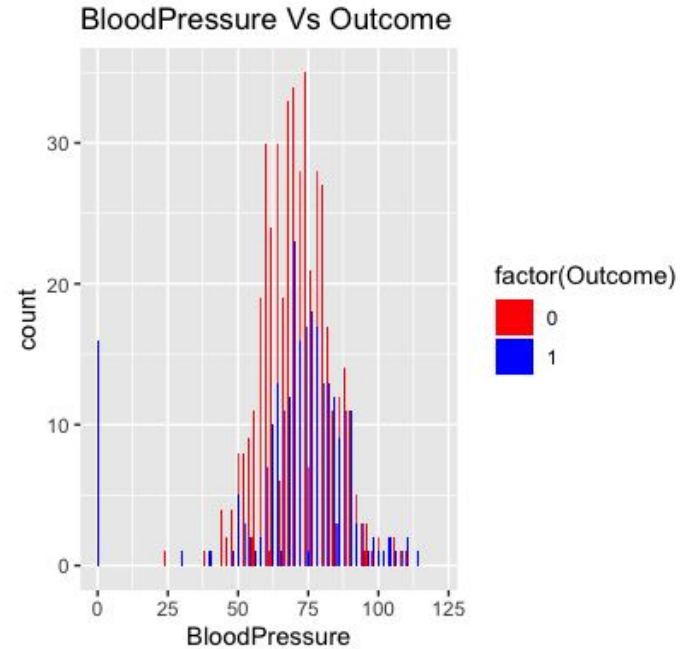




Box-Plot



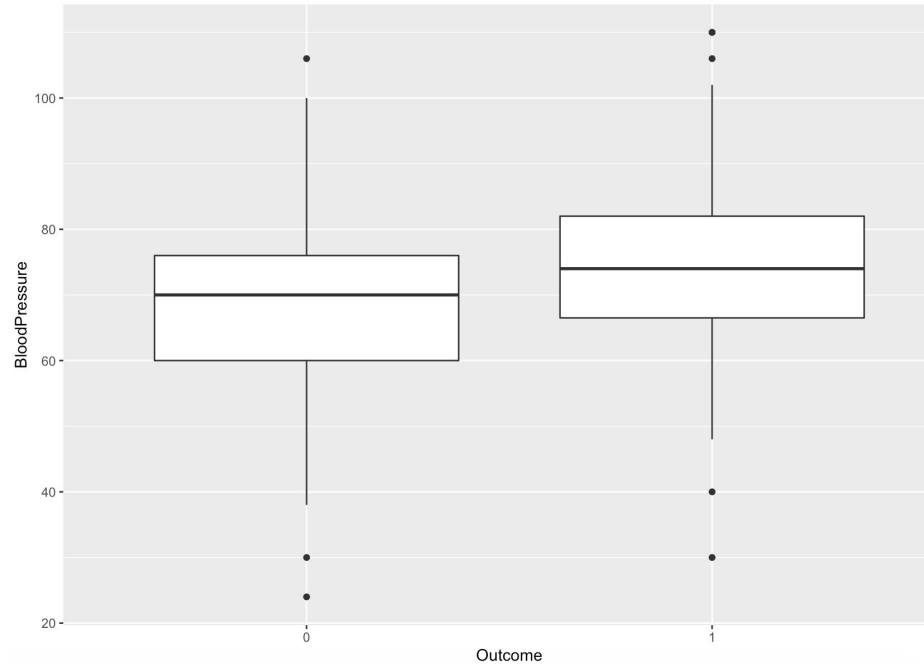
Blood Pressure





Box-Plot

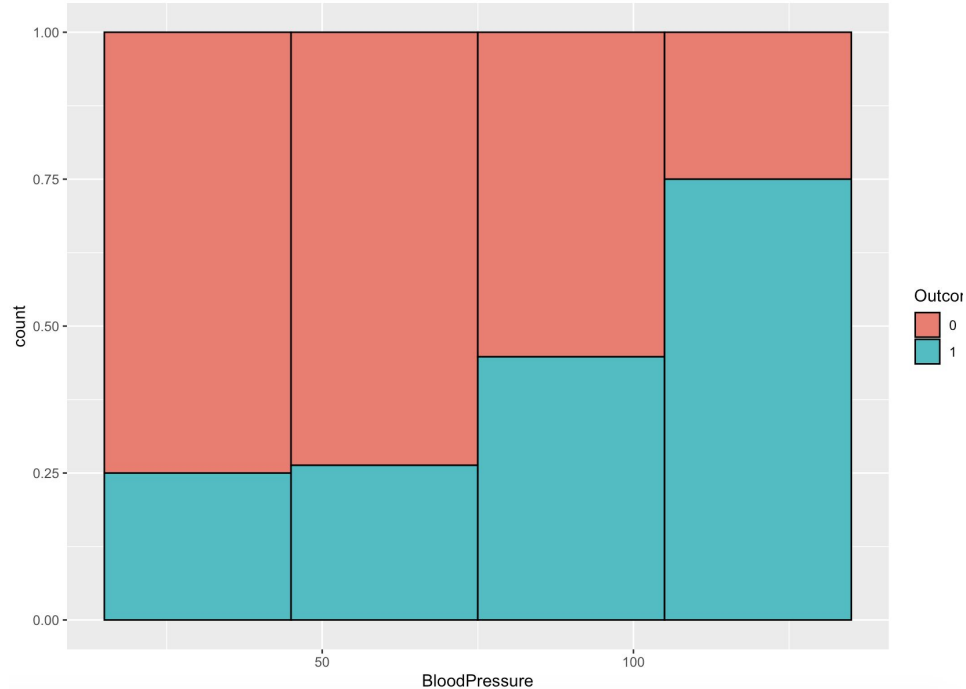
Blood Pressure





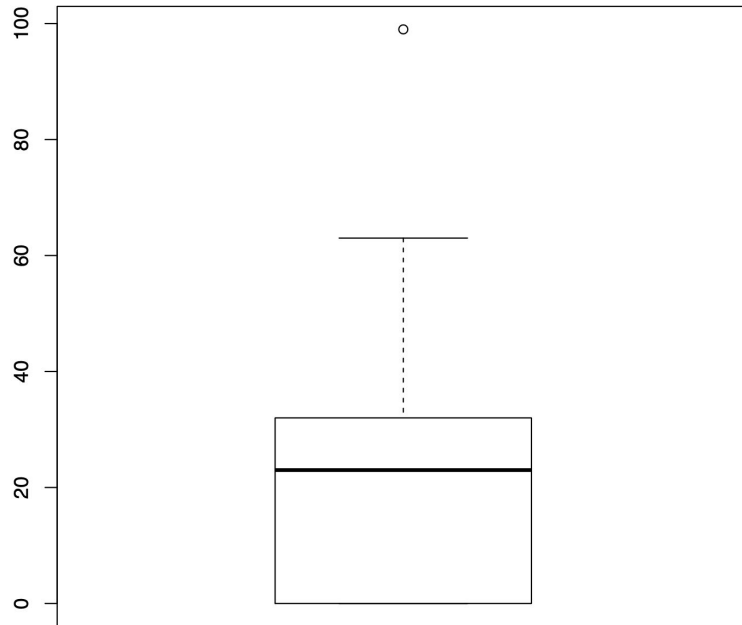
Box-Plot

Blood Pressure



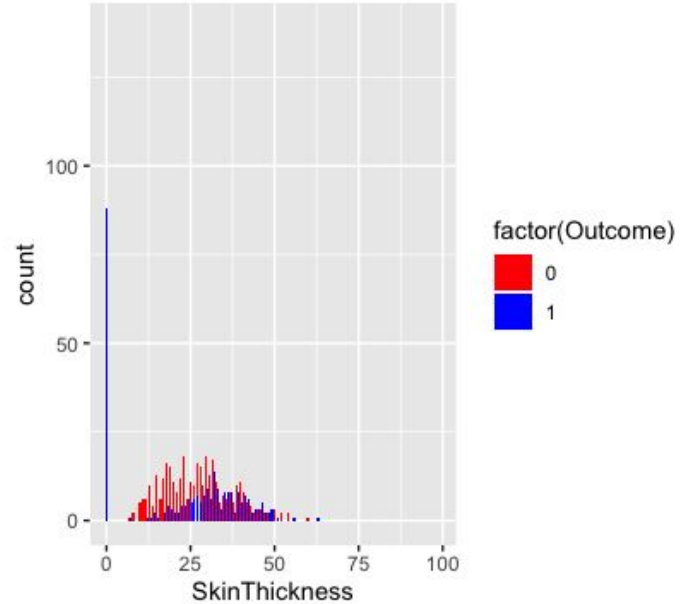


Box-Plot



SkinThickness

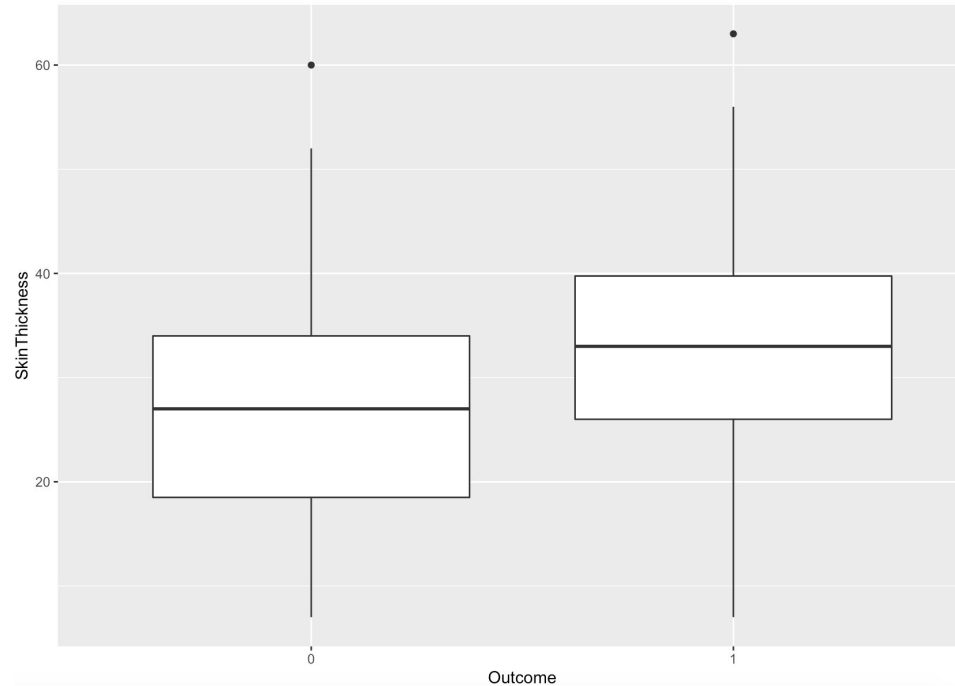
SkinThickness Vs Outcome





Box-Plot

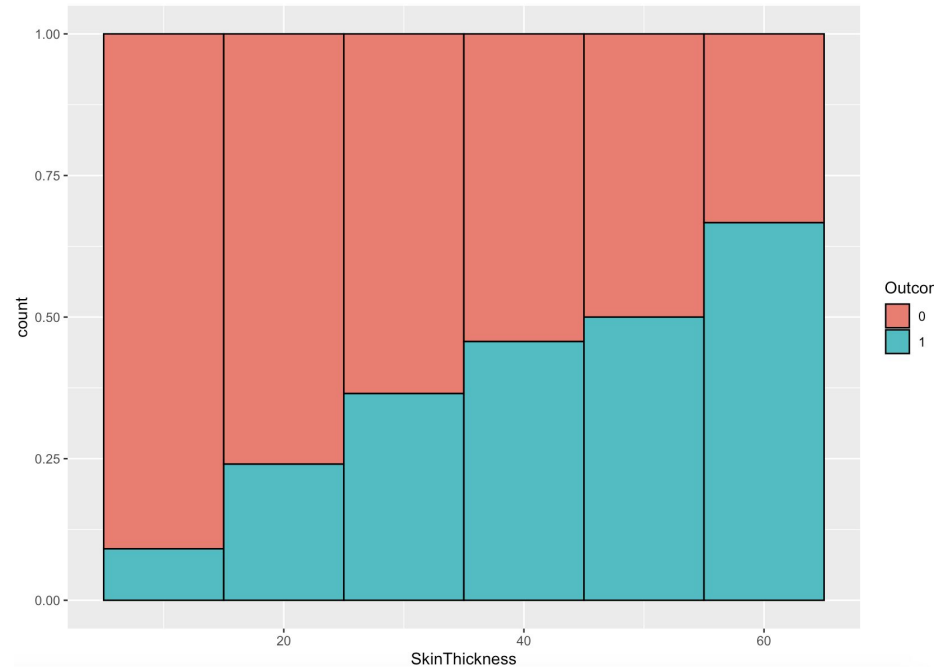
SkinThickness





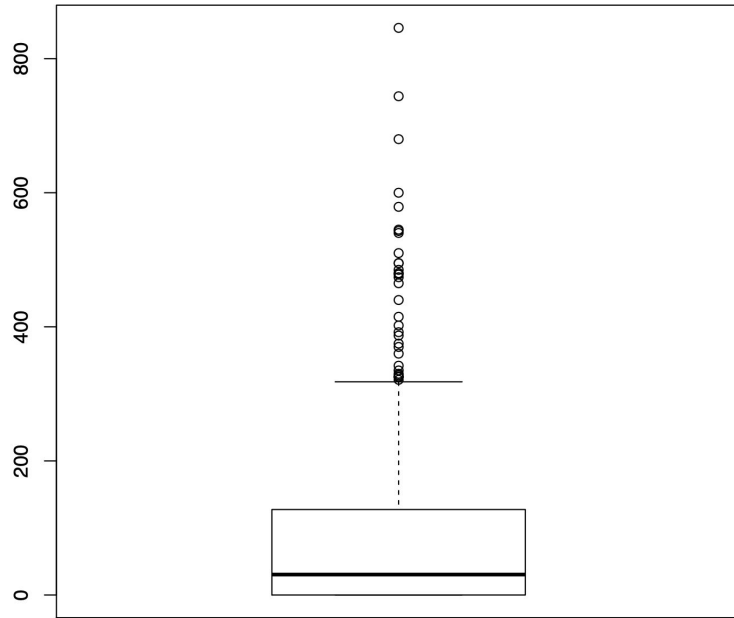
Box-Plot

SkinThickness

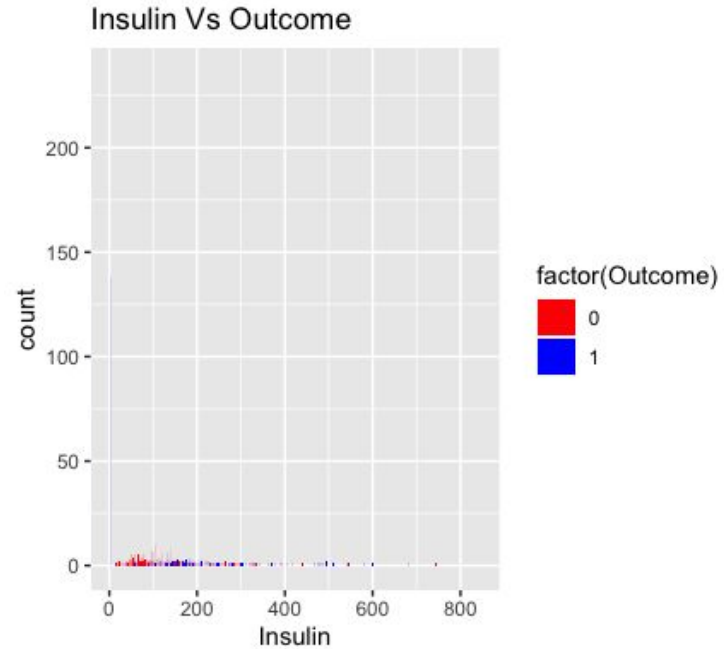




Box-Plot



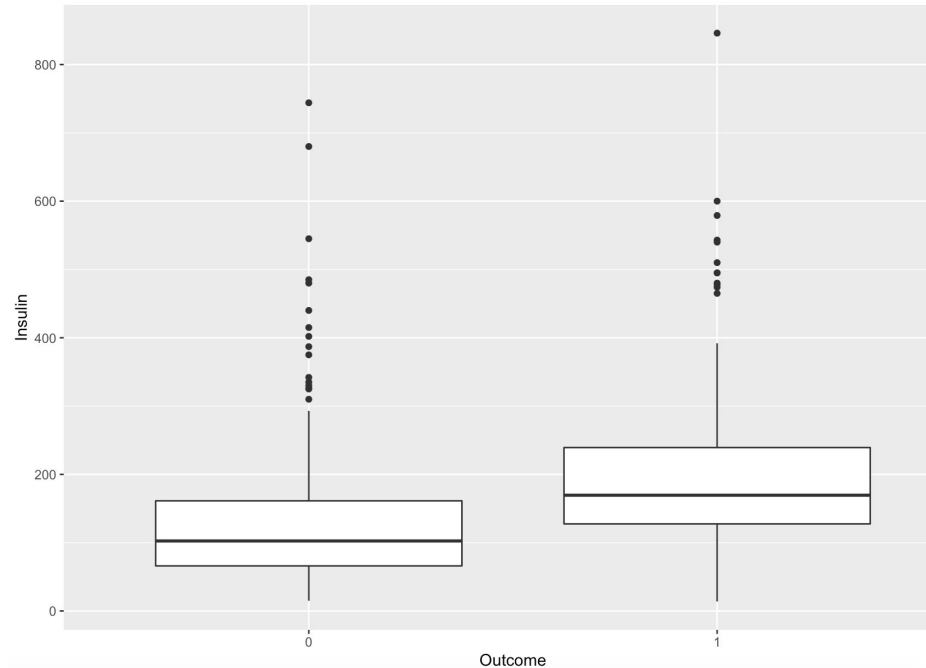
Insulin





Box-Plot

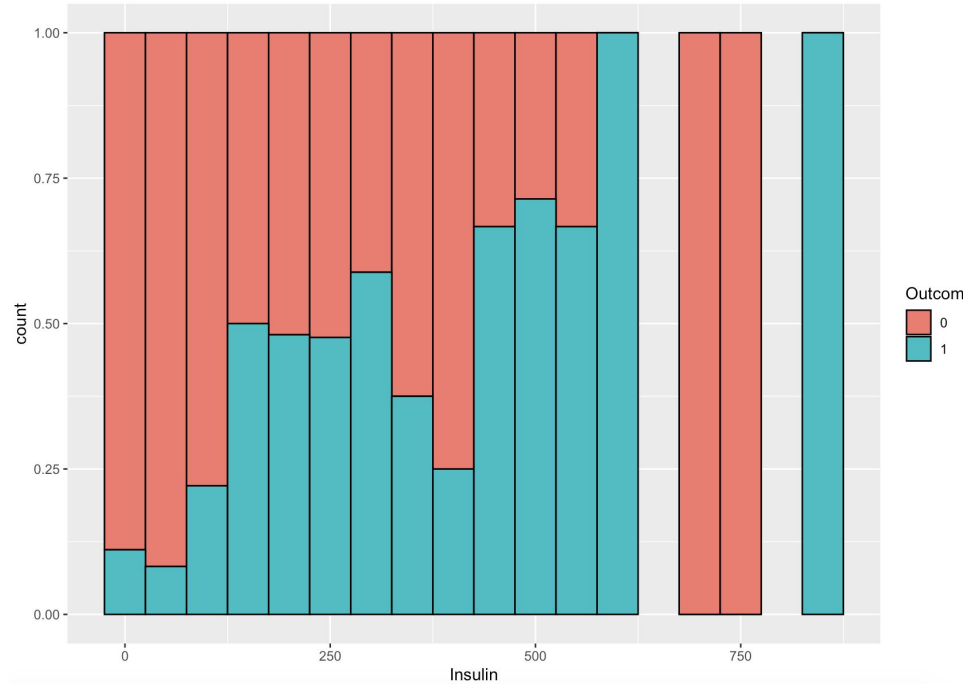
Insulin





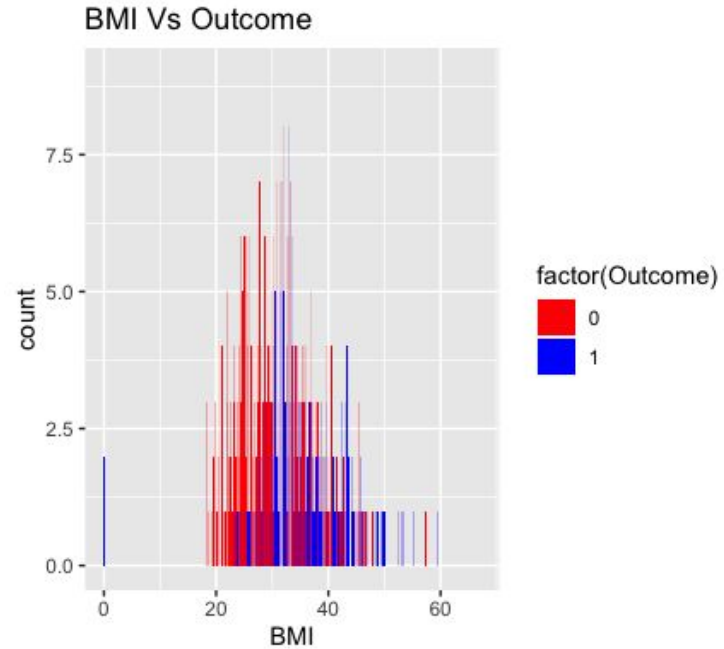
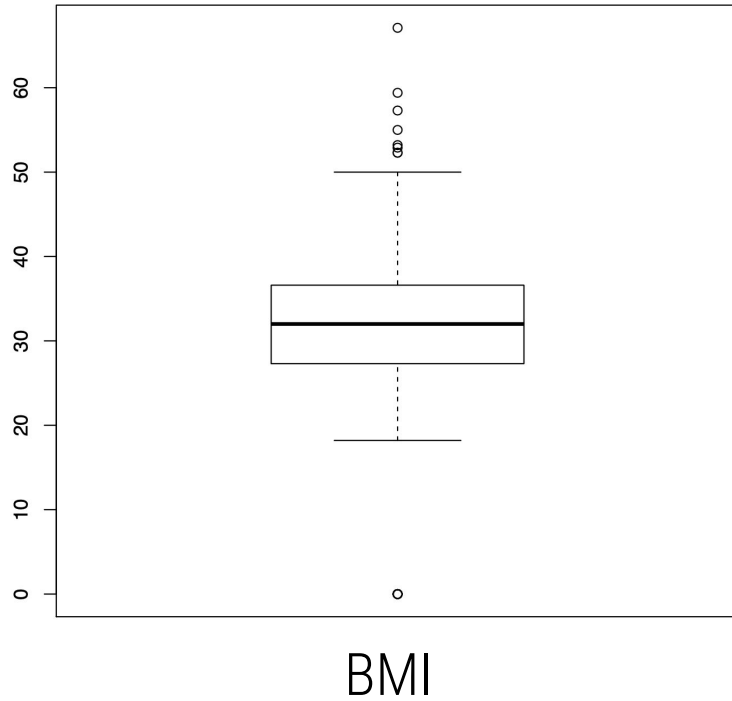
Box-Plot

Insulin





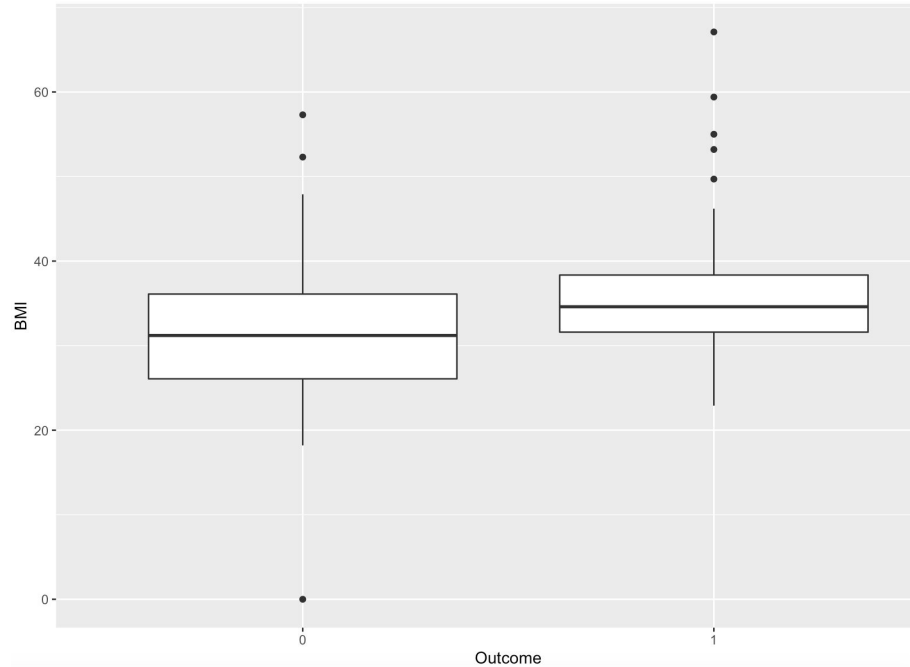
Box-Plot





Box-Plot

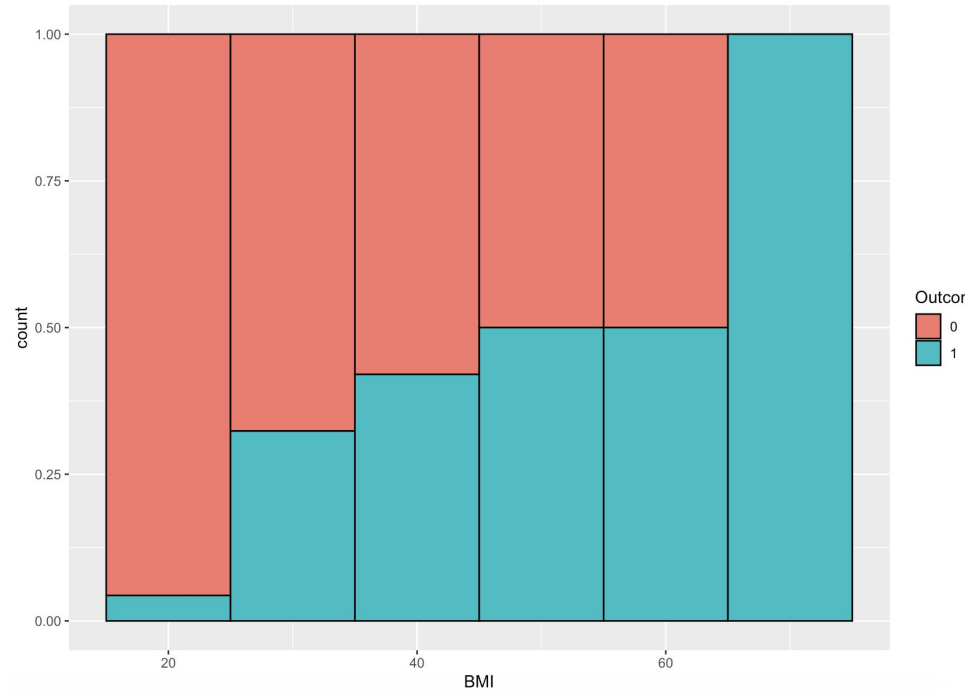
BMI





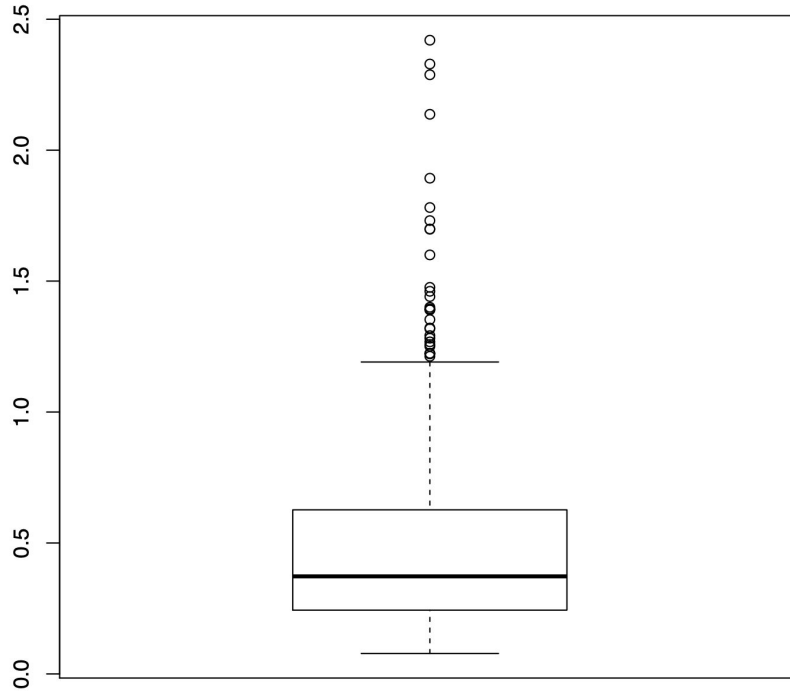
Box-Plot

BMI

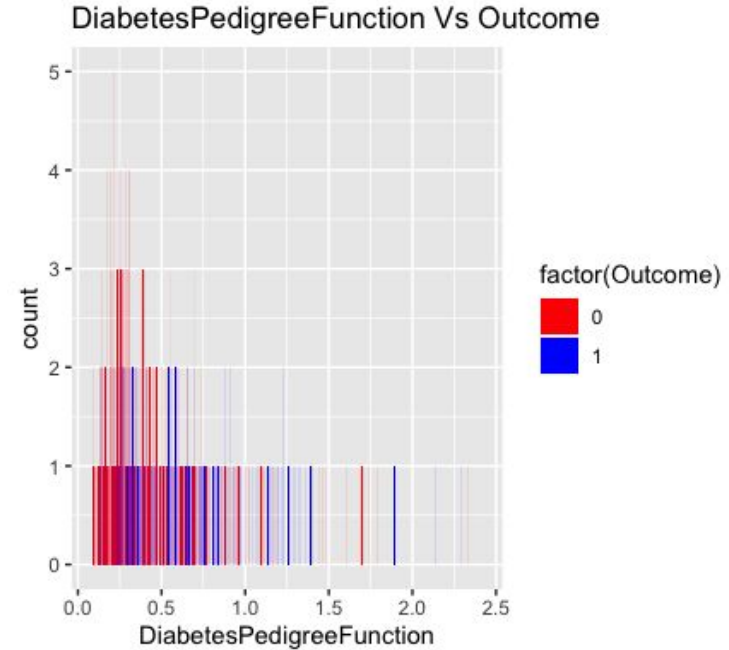




Box-Plot



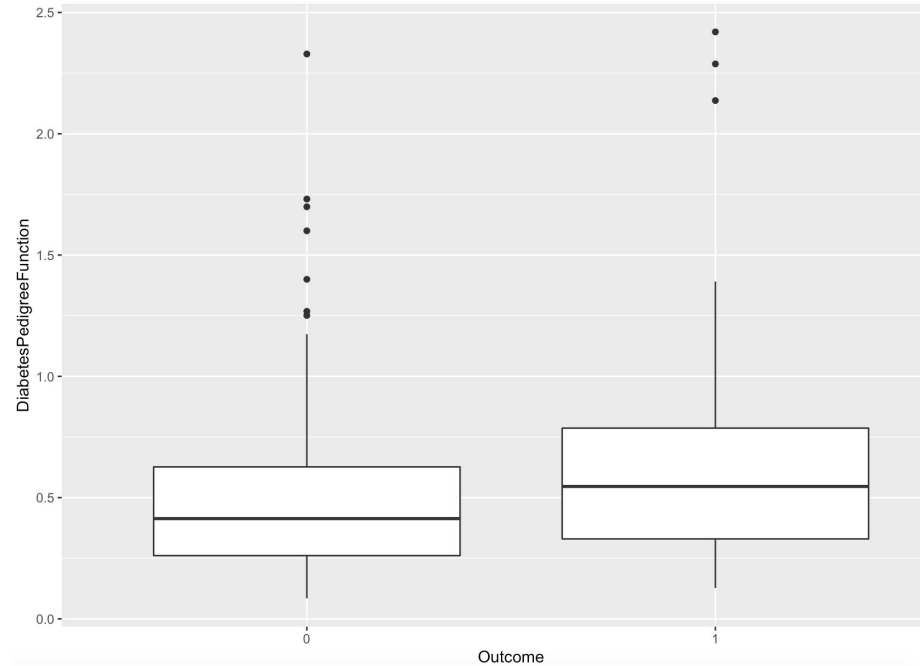
DiabetesPedigreeFunction





Box-Plot

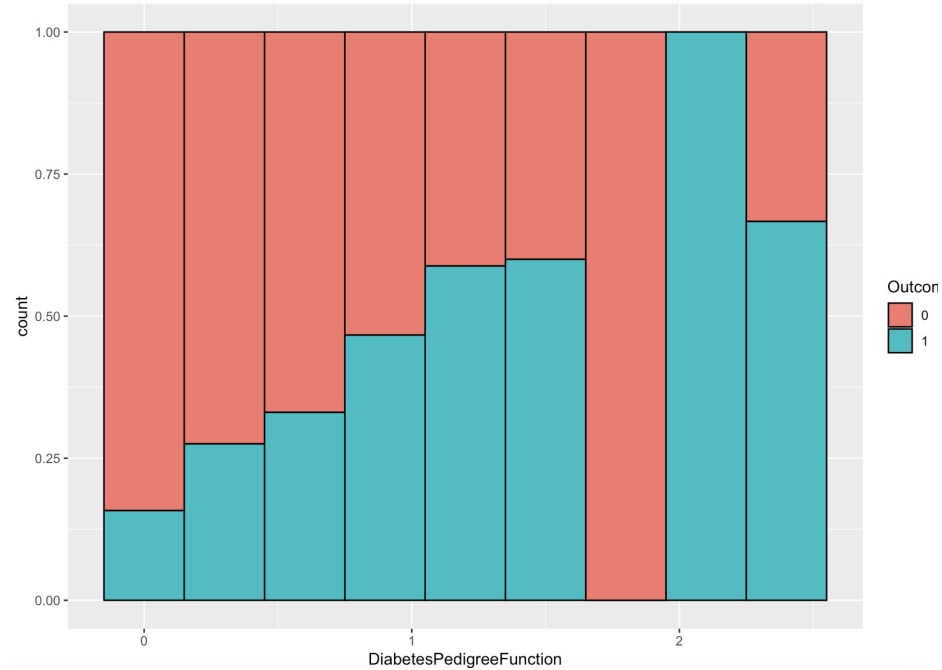
DiabetesPedigree
Function





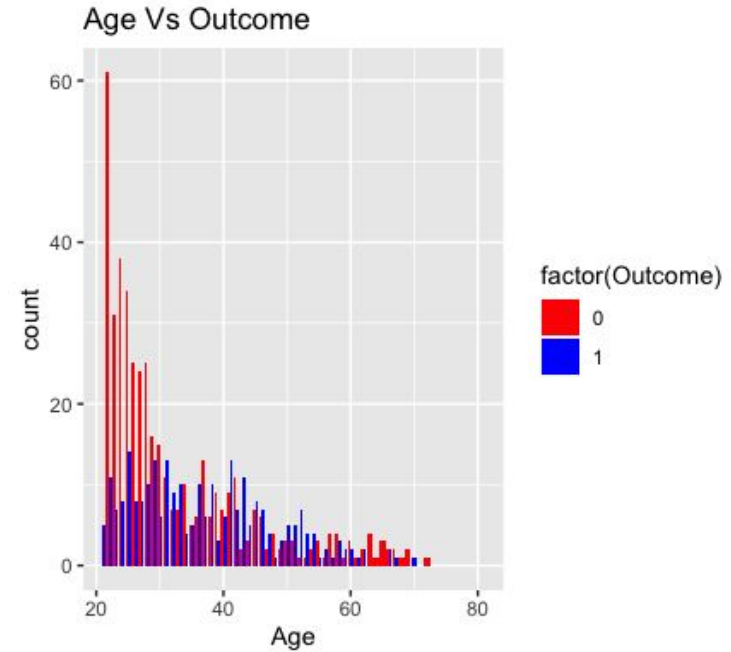
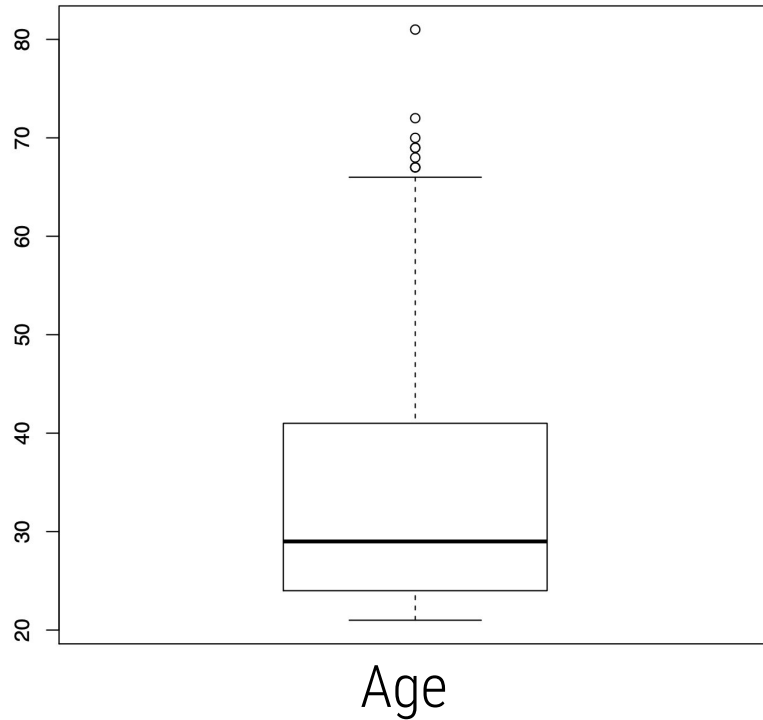
Box-Plot

DiabetesPedigree
Function





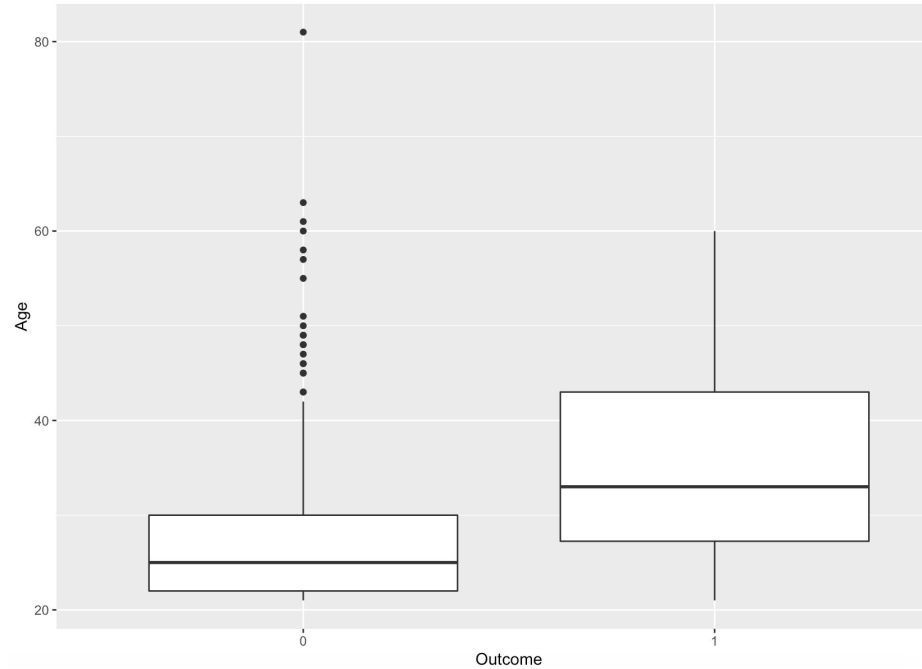
Box-Plot





Box-Plot

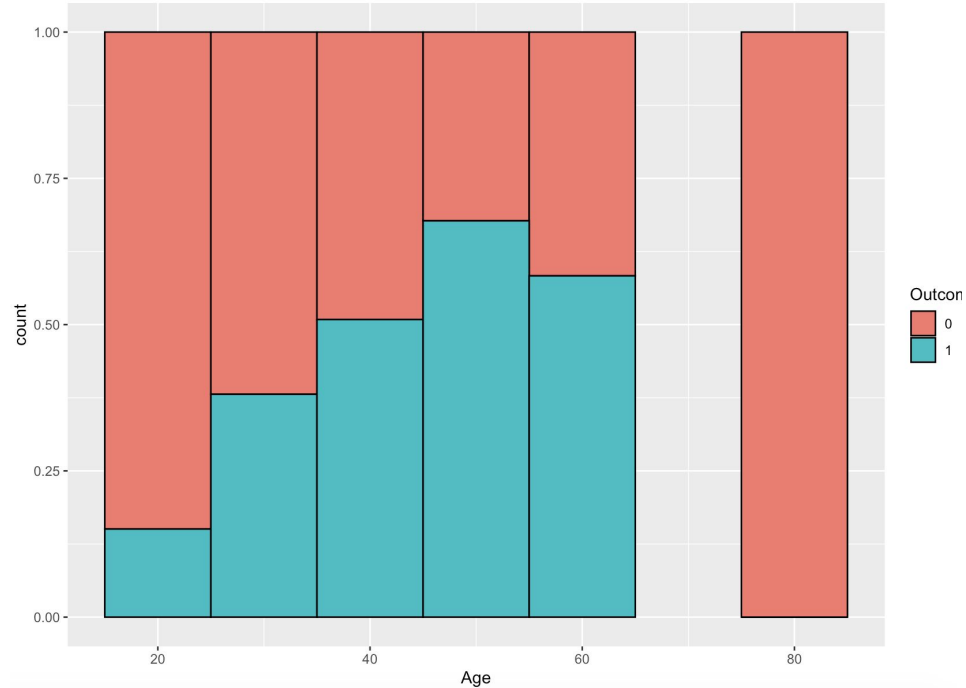
Age





Box-Plot

Age





Data Preprocessing

```
df <- df[df$Pregnancies <= 14.5,]  
df <- df[df$Glucose >= 10,]  
df <- df[df$BloodPressure >= 10,]  
df <- df[df$SkinThickness <= 80 & df$SkinThickness > 0,]  
df <- df[df$Insulin <= 600 ,]  
df <- df[df$BMI <= 60 & df$BMI >= 10,]  
df <- df[df$DiabetesPedigreeFunction <= 65,]  
df <- df[df$Age <= 75,]
```



Features Correlation Matrix



x1: Pregnancies

x2: Glucose

x3: BloodPressure

x4: SkinThickness

x5: Insulin

x6: BMI

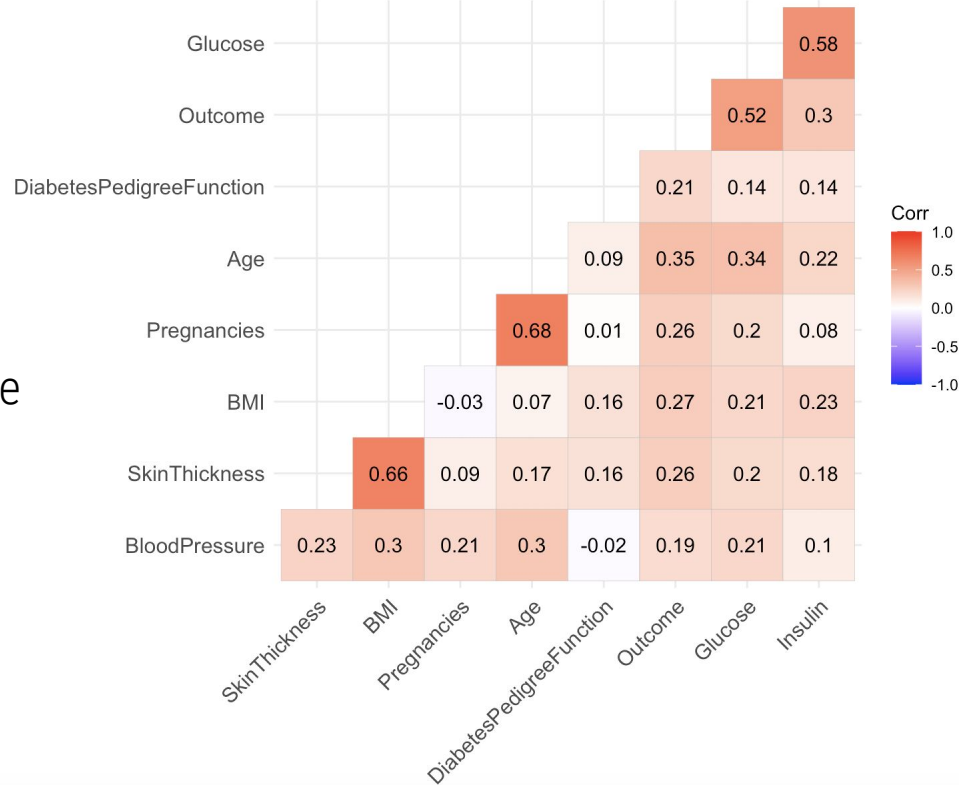
x7: DiabetesPedigreeFunction

x8: Age



Features Correlation Matrix

Correlation matrix
without missing value





Feature Selection

- All Features
- Without Insulin/DiabetesPedigreeFunction
- Without Insulin/DiabetesPedigreeFunction/Age)
- Without Insulin/DiabetesPedigreeFunction/Pregnancies)



Shuffle the Data

```
set.seed(9850)
g<-runif(nrow(df))
data<-df[order(g),]
```

```
RR<-dim(data)[1]
CC<-dim(data)[2]
```

Modeling

The background features a large, dark blue trapezoidal shape on the left side. To its right, a white diagonal line separates the blue area from a light blue background. At the bottom, a thick orange horizontal bar is partially visible, with a small dark blue triangle pointing upwards from its left end.



Methods

- Decision Tree
- Logistic Regression
- Null Model



Performance



Performance

Null Model:

0	1
500	268

Accuracy : 65%



Performance

Average Accuracy	Training	Validation	Testing
All Features	0.81	0.78	0.71
Without x5/x7	0.78	0.77	0.75
Without x1/x5/x7	0.75	0.74	0.85
Without x1/x5/x8	0.78	0.78	0.74



Performance

	Training	Validation	Testing
fold=1			
fold=2			
fold=3	0.76	0.76	0.78
fold=4	0.77	0.75	0.79
fold=5	0.75	0.74	0.85
fold=6	0.76	0.75	0.86
fold=7	0.79	0.76	0.83
fold=8	0.76	0.75	0.89
fold=9	0.76	0.76	0.88
fold=10	0.76	0.75	0.86
Ave.	0.76	0.75	0.84



Performance

Average Accuracy	Training	Validation	Testing
All Features	0.83	0.81	0.67
Without x5/x7	0.77	0.78	0.78
Without x1/x5/x7	0.78	0.76	0.78
Without x5/x7/x8	0.78	0.77	0.80



Performance

	Training	Validation	Testing
fold=1			
fold=2			
fold=3	0.78	0.76	0.77
fold=4	0.78	0.75	0.78
fold=5	0.78	0.77	0.80
fold=6	0.78	0.78	0.78
fold=7	0.78	0.77	0.76
fold=8	0.78	0.78	0.74
fold=9	0.79	0.78	0.71
fold=10	0.79	0.78	0.71
Ave.	0.78	0.77	0.76



Results

Is your improvement significant?

- As the result shows, if we put all features to train our model , no matter the logistic regression or decision tree, the training accuracy are the highest but the testing accuracy are lowest.
- We thought that putting all features to train model would cause model be overfitted. Thus we took away the features that have high correlation to avoid the collinearity.
- Eventually, we constructed a model which achieved **80%** testing accuracy only trained with five features. The improvement is significantly higher than the null model.

The background features a large, dark blue trapezoidal shape on the left side. To its right, there is a white area with a diagonal cutout. At the bottom, a thick orange horizontal bar is positioned, with a small blue trapezoidal shape overlapping its left end. The top right corner is white.

Demo



Demo

Demo

Performance

```
Rscript code/rproject.R --fold n --train data/diabetes.csv --report performance.csv --predict predict.csv
```

Reference



Packages we use

- `library(rpart)`
- `library(caret)`
- `library(party)`
- `library(varhandle)`
- `library(ggplot)`
- `library(corrplot)`
- `library(Hmisc)`
- `library(ggcorrplot)`



REFERENCE

- <https://www.kaggle.com/devisangeetha/which-factor-causes-diabetes>
- <https://medium.com/datainpoint/r-essentials-ggplot2-visualizations-817d2416b83e>
- <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- <http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2>



THANKS!