Sardar Patel Institute of Technology, Mumbai
Department of Electronics and Telecommunication Engineering
B.E. Sem-VII (2022-2023)
OEIT6 - Data Analytics

Experiment: Exploratory Data Analysis in SAS Studio

Name: Pushkar Sutar **Roll No. 2019110060**

Objective : Performing exploratory data analysis in SAS Studio.

Dataset Description:
This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.2 From the data set in the (.csv) File We can find several variables, some of them are independent (several medical predictor variables) and only one target dependent variable (Outcome).
About Dataset :
Pregnancies: To express the Number of pregnancies
Glucose: To express the Glucose level in blood
BloodPressure: To express the Blood pressure measurement
SkinThickness: To express the thickness of the skin
Insulin: To express the Insulin level in blood
BMI: To express the Body mass index
DiabetesPedigreeFunction: To express the Diabetes percentage
Age: To express the age
Outcome: To express the final result 1 is Yes and 0 is No

Code and Output:

Importing the Data -

```
* importing the data;
proc import datafile = '/home/u62322946/EDA/diabetes.csv'
 out = work.mydata
 dbms = CSV;
run;
```

Print the data -

```
*printing;
proc print
    data=work.mydata;
run;
```

Obs	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	0.248	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	0.232	54	1
11	4	110	92	0	0	37.6	0.191	.	0
12	10	168	74	0	0	38	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

Checking basic statistical parameters of the data -

```
proc means data=work.mydata
mean median mode std var min max;
```

Variable	Mean	Median	Mode	Std Dev	Variance	Minimum	Maximum
Pregnancies	3.8450521	3.0000000	1.0000000	3.3695781	11.3540563	0	17.0000000
Glucose	120.9126467	117.0000000	99.0000000	31.9895370	1023.33	0	199.0000000
BloodPressure	69.1054688	72.0000000	70.0000000	19.3558072	374.6472712	0	122.0000000
SkinThickness	20.5364583	23.0000000	0	15.9522176	254.4732453	0	99.0000000
Insulin	79.7994792	30.5000000	0	115.2440024	13281.18	0	846.0000000
BMI	31.9925781	32.0000000	32.0000000	7.8841603	62.1599840	0	67.1000000
DiabetesPedigreeFunction	0.4718763	0.3725000	0.2540000	0.3313286	0.1097786	0.0780000	2.4200000
Age	33.2451108	29.0000000	22.0000000	11.7673221	138.4698684	21.0000000	81.0000000
Outcome	0.3489583	0	0	0.4769514	0.2274826	0	1.0000000

As we can observe that the mean of pregnancies is 3.8, this study thus aims to understand the gestational diabetes that happens in women after pregnancies. The attribute insulin shows a large deviation from mean value, suggesting no clustering and many unique values. The mean age of women considered is 33.2 with minimum age of 21 years.

Cleaning the data -

```
proc means data=work.mydata nmiss;
```

```
data cleanData;
 set mydata;
 if Age = . or Glucose = . then delete;
run;

proc means data=cleanData nmiss;
```

Variable	N Miss
Pregnancies	0
Glucose	1
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	1
Outcome	0

Variable	N Miss
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

The attributes glucose and age had one null value each which are dropped by dropping the entire row that contains a null value.

Number of unique values -

```
proc sql;
select count(distinct 'Pregnancies'n) as 'Pregnancies'n,
       count(distinct Age) as Age,
       count(distinct Outcome) as Outcome,
       count(distinct BMI) as BMI
  from cleanData;
```

Pregnancies	Age	Outcome	BMI
17	52	2	248

There are a lot of unique values of BMI. There are a total 17 unique values for pregnancies which are quite surprising and may be an outlier.

Normalisation -

```
proc stdize data=cleanData out=normalized_data;
   var _numeric_;
run;
```

All the numeric values are normally distributed after execution of the procedure.

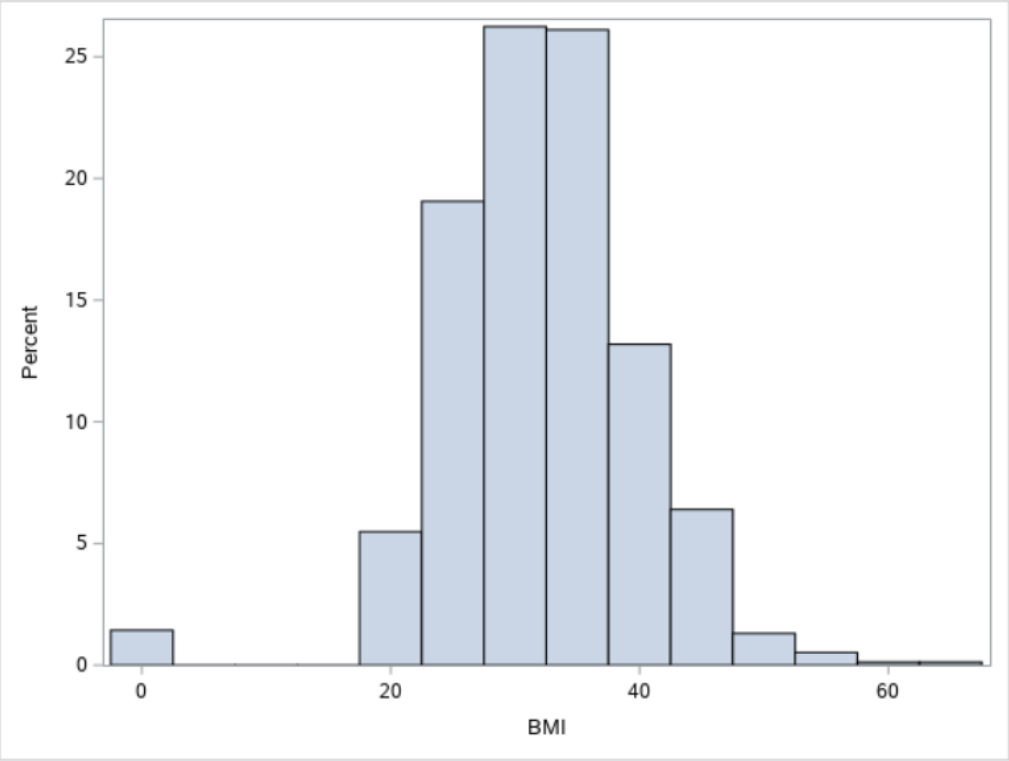
Total rows: 766 Total columns: 9 Rows 1-100

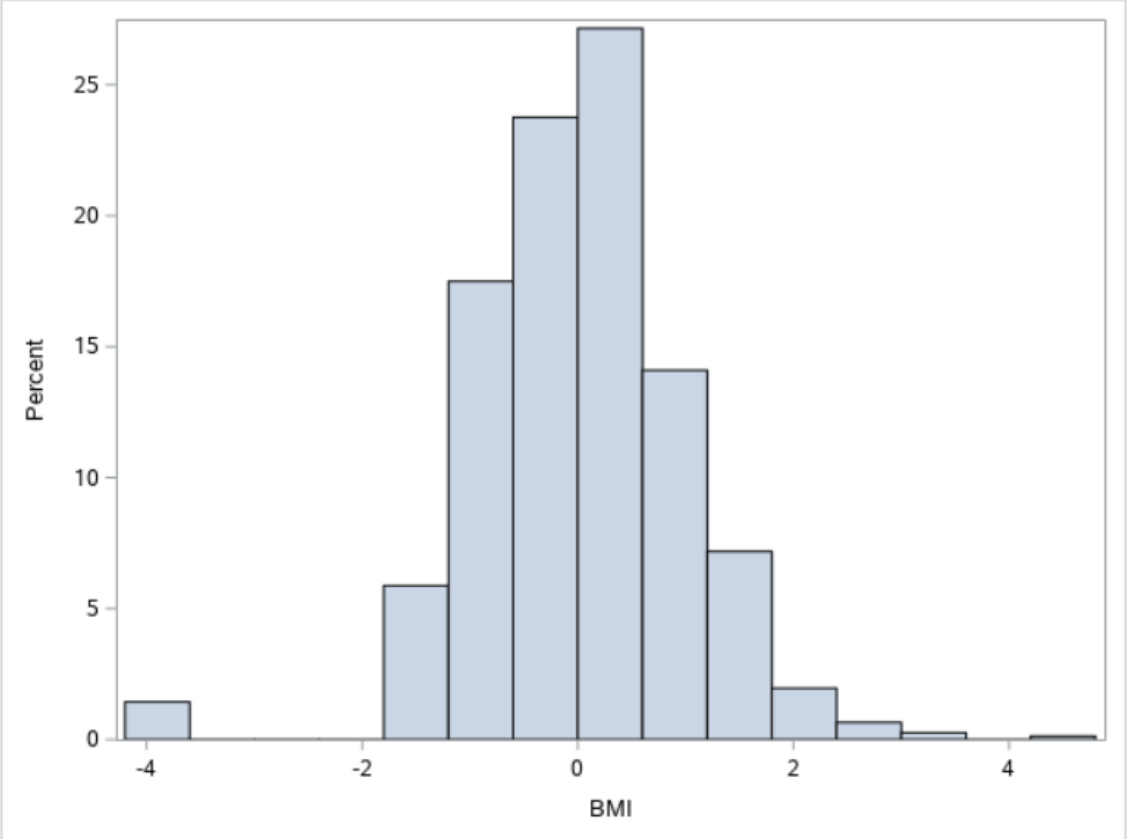
	Pregnancies	Glucose	BloodPressure	SkinThic
1	0.6403442911	0.845823044	0.151364727	0.90410
2	-0.84243602	-1.122434669	-0.158511569	0.52765
3	1.2334564157	1.9392995513	-0.261803668	-1.2918
4	-0.84243602	-0.997465926	-0.158511569	0.1512
5	-1.138992083	0.5021589988	-1.501308854	0.90410
6	0.3427882289	0.153926906	0.2546568259	1.2918

```
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=cleanData;
    histogram 'BMI'n /;

ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=normalized_data;
    histogram 'BMI'n /;
```

We can observe the difference by seeing the histogram plot of BMI before and after normalisation.





Univariate Analysis of data -

```
ods graphics on;
proc Univariate data=cleanData;
var Glucose;
run;
quit;
```

Variable: Glucose

Moments			
N	766	Sum Weights	766
Mean	120.926893	Sum Observations	92630
Std Deviation	32.0080036	Variance	1024.5123
Skewness	0.17074961	Kurtosis	0.63344063
Uncorrected SS	11985210	Corrected SS	783751.906
Coeff Variation	26.4688878	Std Error Mean	1.15649618

Basic Statistical Measures			
Location		Variability	
Mean	120.9269	Std Deviation	32.00800
Median	117.0000	Variance	1025
Mode	99.0000	Range	199.00000
		Interquartile Range	42.00000

Note: The mode displayed is the smallest of 2 modes with a count of 17

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	104.5632	Pr > |t|	<.0001
Sign	M	380.5	Pr >= |M|	<.0001
Signed Rank	S	144970.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	199
99%	196
95%	181
90%	167
75% Q3	141
50% Median	117
25% Q1	99
10%	85
5%	79
1%	57
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	501	197	227
0	348	197	407
0	341	197	578
0	181	198	560
0	74	199	660

We can see that the mean value of glucose is 120.9mg/dl which falls in the prediabetic range. A standard deviation of 32, which suggests around 68% of people are in the range of 88.9-152.9. We can see that the highest observation of 197md/dl is recorded which is quite serious and might be treated as an outlier. Mode of 99mg/dl suggests that the majority of women are having normal sugar levels.

Bivariate Analysis -

```
proc freq data=cleanData;
table Age*Outcome;
run;

ods graphics on;
```

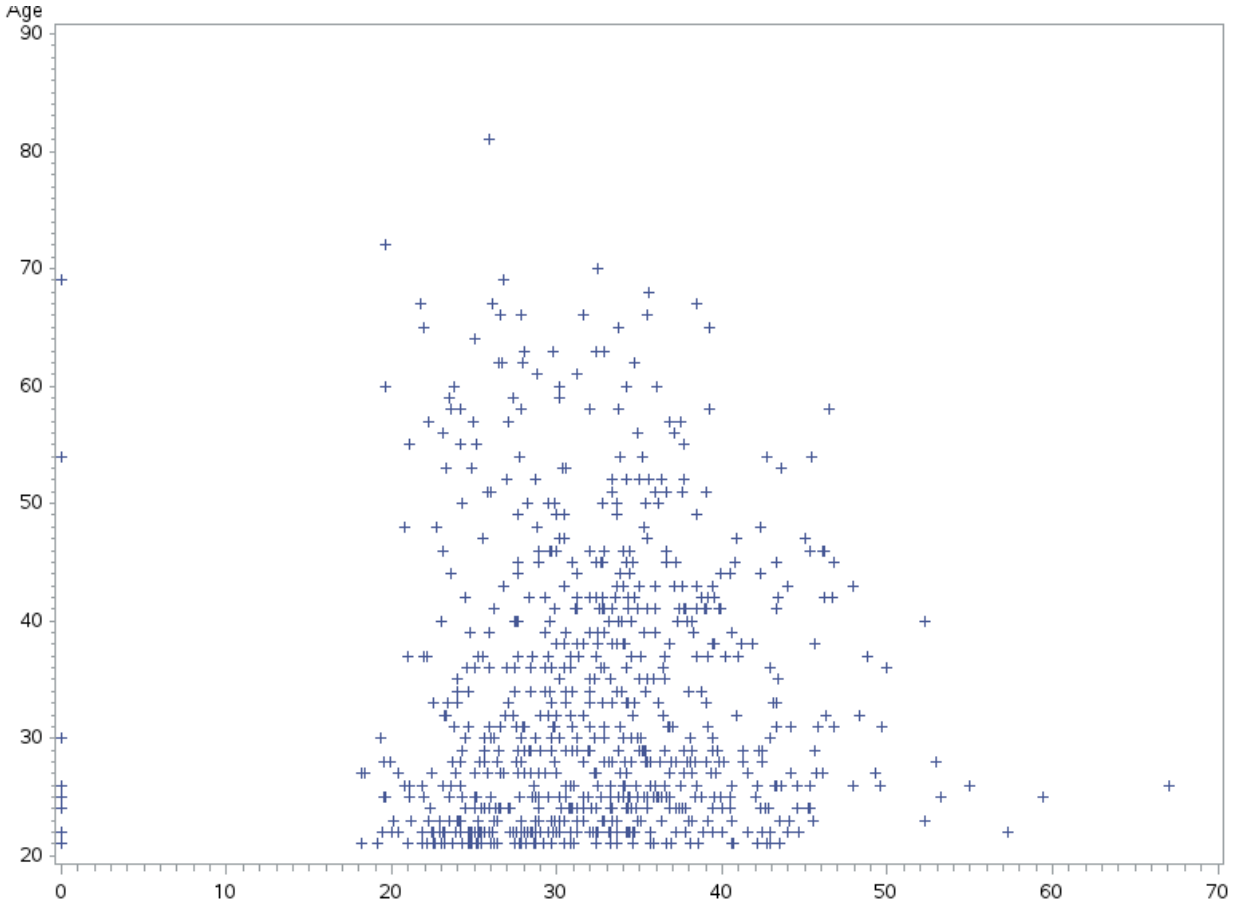
```
proc gplot data=cleanData;
plot Age*BMI;
run;
quit;
```

Frequency
Percent
Row Pct
Col Pct

Table of Age by Outcome

Age	Outcome 0	Outcome 1	Total
21	58 7.57 92.06 11.62	5 0.65 7.94 1.87	63 8.22
22	61 7.96 84.72 12.22	11 1.44 15.28 4.12	72 9.40
23	31 4.05 81.58 6.21	7 0.91 18.42 2.62	38 4.96
24	38 4.96 82.61 7.62	8 1.04 17.39 3.00	46 6.01
25	34 4.44 70.83 6.81	14 1.83 29.17 5.24	48 6.27
26	25 3.26 75.76 5.01	8 1.04 24.24 3.00	33 4.31
27	24 3.13 75.00 4.81	8 1.04 25.00 3.00	32 4.18
28	25 3.26 71.43	10 1.31 28.57	35 4.57

From the whole table it can be clearly seen that the age group of 30-40 years are at the most risk of diabetes due to pregnancy. There is little to no risk of diabetes in the age group below 25 years. There are not enough observations in the age group after 50 years.



From the scatterplot we can observe that there is no significant correlation between age and BMI. However there are a few data points which have 0 BMI which clearly indicates that there are some random errors as BMI of 0 is not possible. Such observations should be discarded.

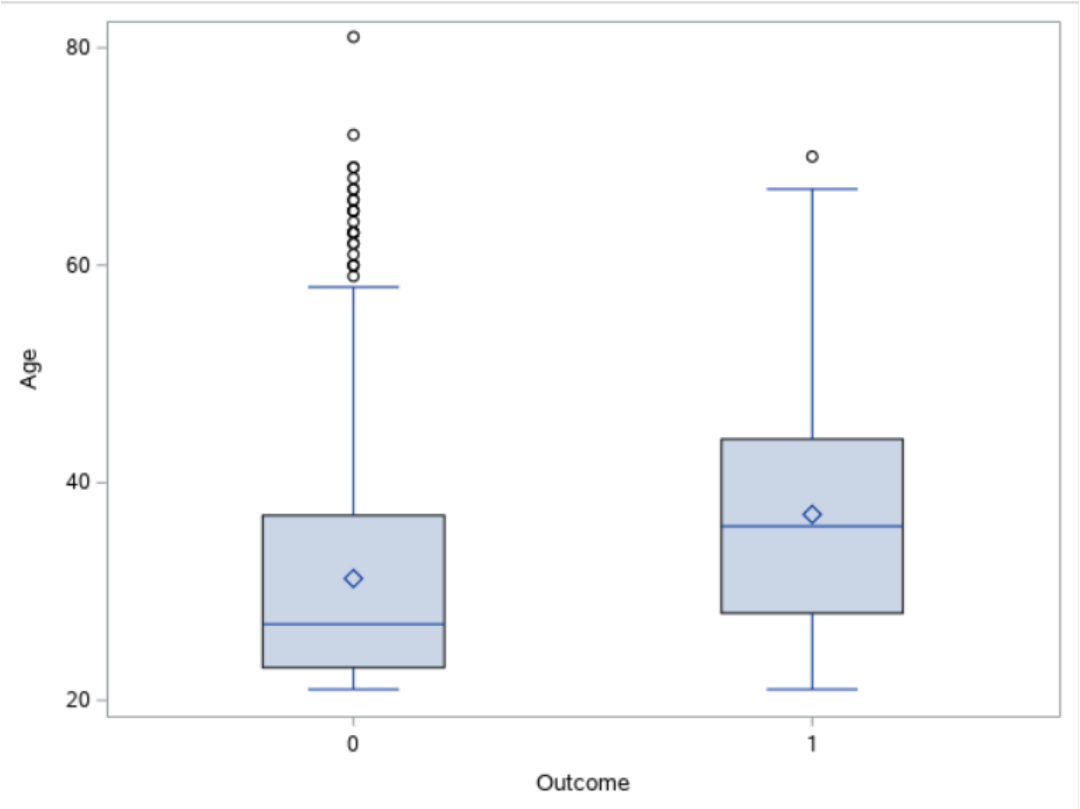
Multivariate Analysis -

```
PROC MEANS DATA=cleanData;
  CLASS Outcome;
  VAR Glucose Age BMI BloodPressure;
RUN;
```

Outcome	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	499	Glucose	499	109.9799599	26.1674328	0	197.0000000
		Age	499	31.1923848	11.6792415	21.0000000	81.0000000
		BMI	499	30.2895792	7.6906120	0	57.3000000
		BloodPressure	499	68.1362725	18.0496155	0	122.0000000
1	267	Glucose	267	141.3857678	31.9303322	0	199.0000000
		Age	267	37.0898876	10.9825293	21.0000000	70.0000000
		BMI	267	35.1632959	7.2686370	0	67.1000000
		BloodPressure	267	70.8127341	21.5312884	0	114.0000000

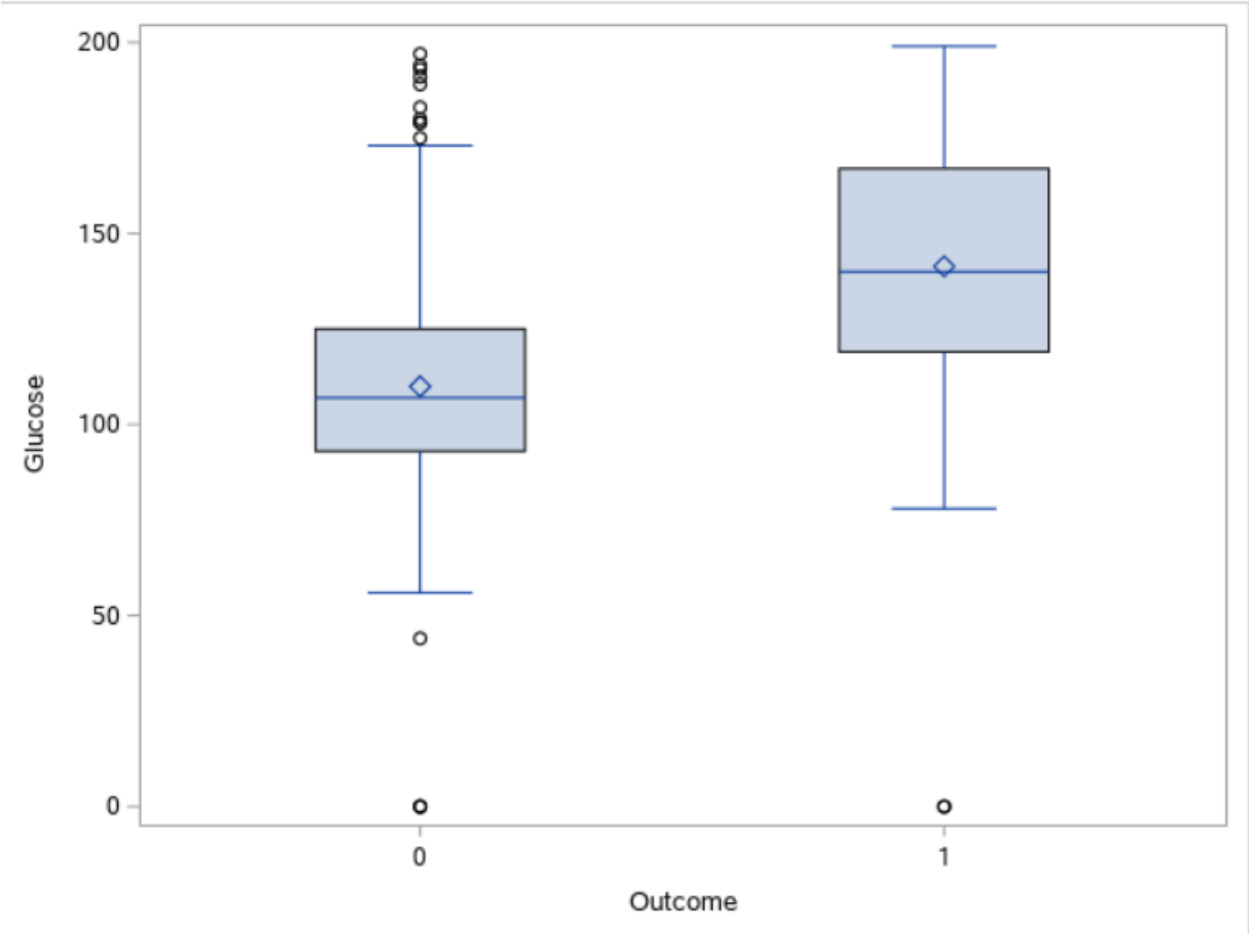
We can see that 35% of the women in study were found to be diabetic. The mean values of blood pressure are almost similar for both the groups, which might mean that this attribute does not affect the target significantly. The mean values of Glucose and BMI are higher for diabetic groups indicating that an unhealthy lifestyle might be the reason behind the condition.

```
PROC SGPLOT DATA=cleanData;
  VBOX Age / CATEGORY=Outcome;
RUN;
```

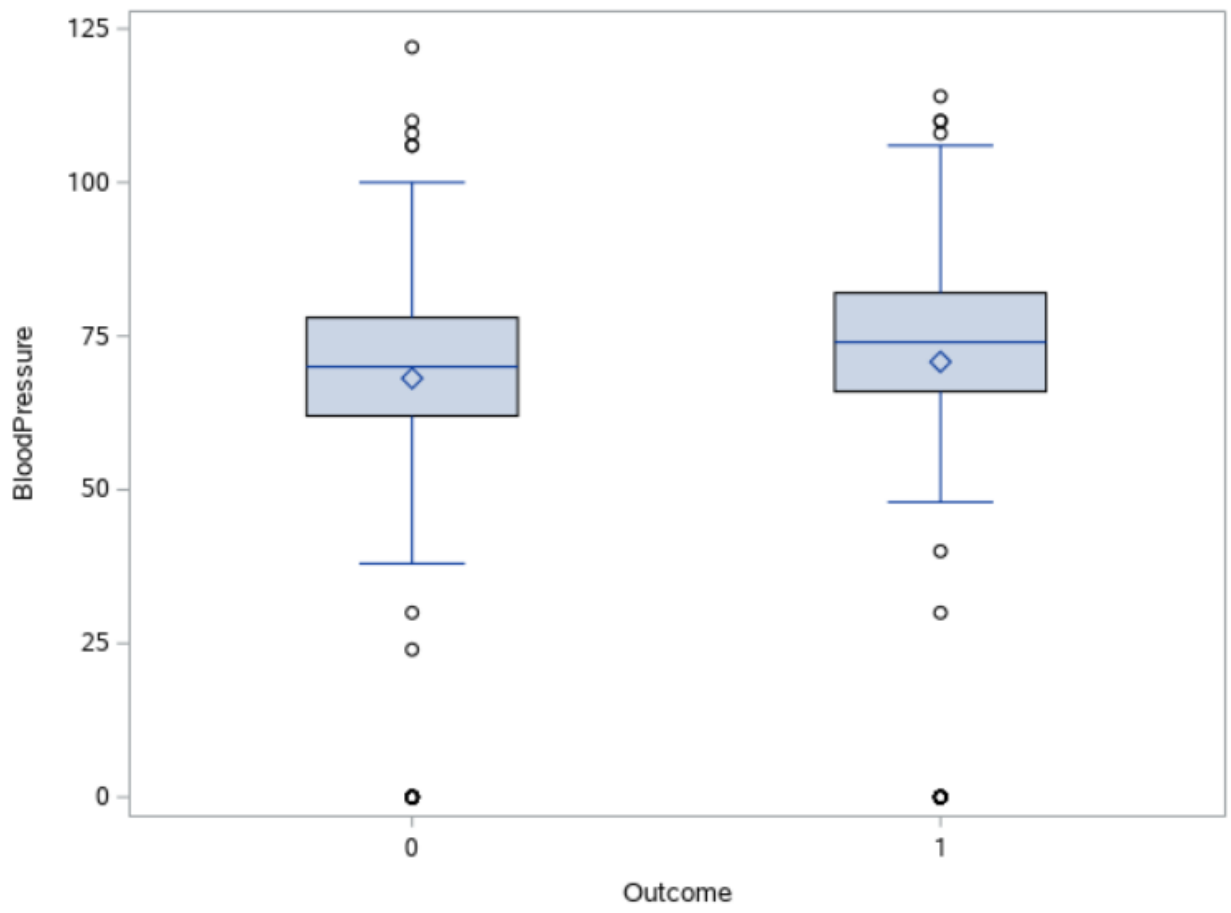


The mean age of non-diabetic group is much less, further supporting the claim that the age group 30-40 are at higher risk. There are a lot of outliers after age 50 years in non-diabetic groups.

```
PROC SGPLOT DATA=cleanData;
  VBOX Glucose / CATEGORY=Outcome;
RUN;
```



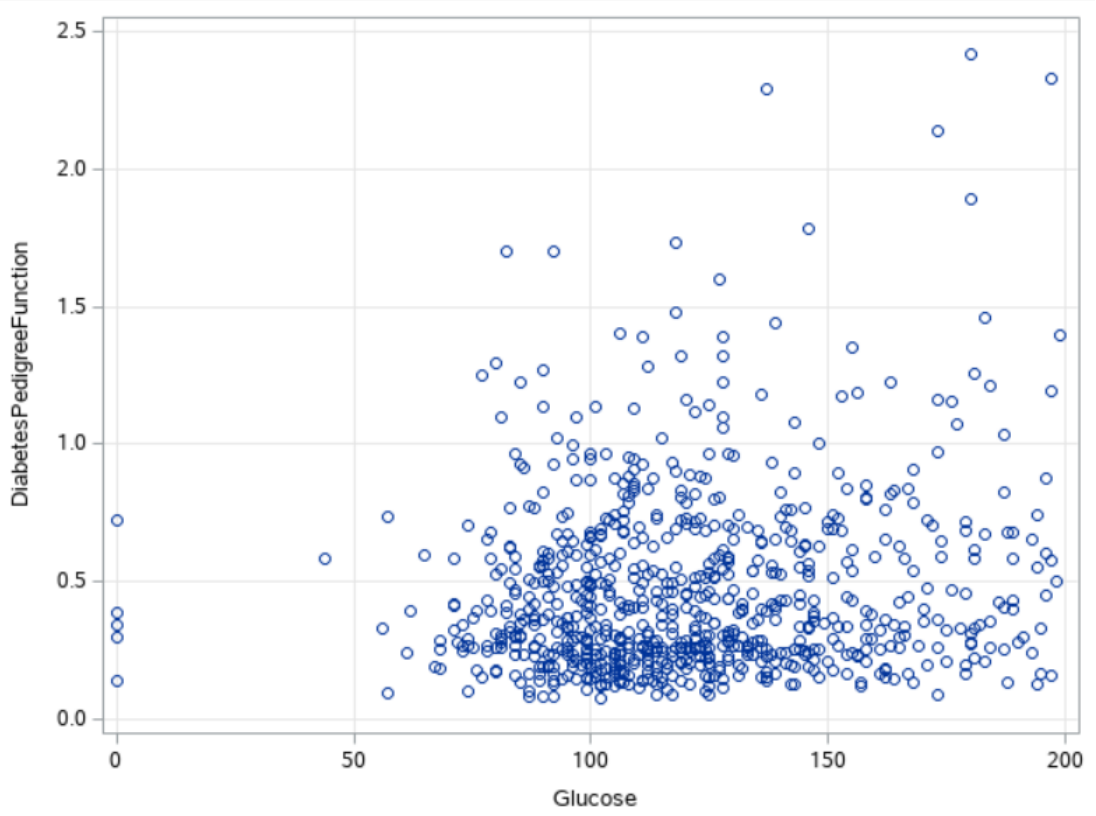

```
PROC SGPLOT DATA=cleanData;
  VBOX BloodPressure / CATEGORY=Outcome;
RUN;
```



The glucose levels of diabetic patients are higher which is as expected. Although there is no significant difference in blood pressures of diabetic and non-diabetic groups.

```
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=cleanData;
    scatter x='Glucose'n y='DiabetesPedigreeFunction'n /;
    xaxis grid;
    yaxis grid;

ods graphics / reset;
```



Diabetes Pedigree Function: indicates the function which scores likelihood of diabetes based on family history. We can observe a lot of points corresponding to around 100mg/dl glucose and DPF of below 0.5. Which supports the theory that the likelihood of diabetes is largely dependent on the family history. We can find higher values of DPF on the right hand side of the scatter plot having high glucose levels.

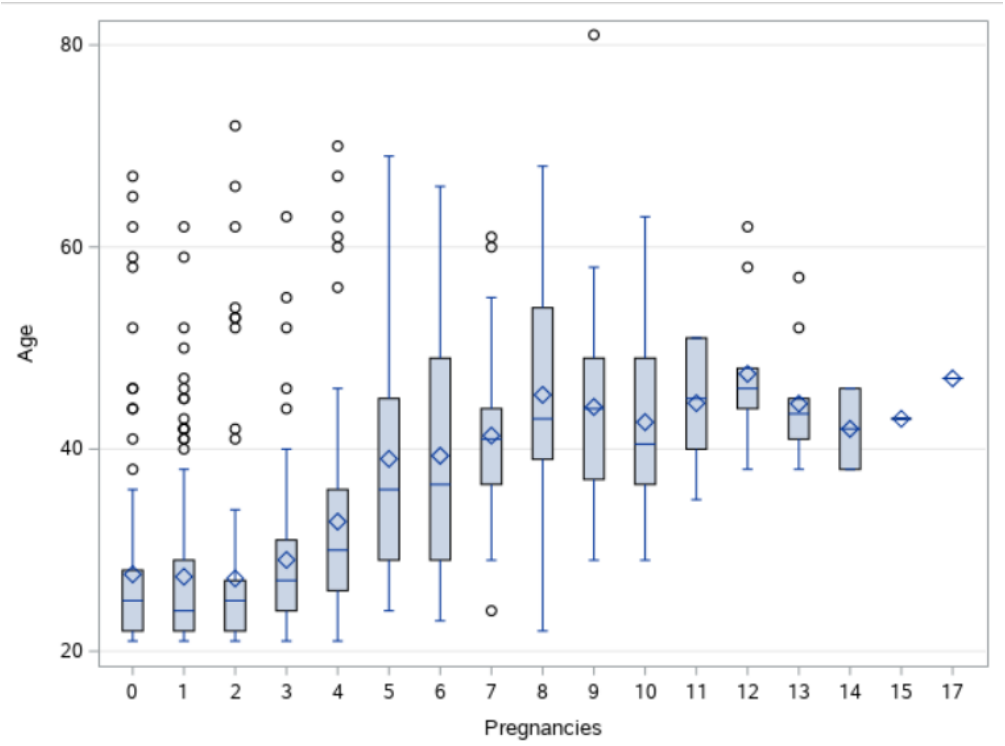
```
ods noproctitle;
ods graphics / imagemap=on;
proc corr data=cleanData pearson nosimple noprob plots=none;
    var 'Glucose'n 'Age'n 'BMI'n 'Insulin'n 'BloodPressure'n
'SkinThickness'n;
run;
```

7 Variables:	Glucose Age BMI Insulin BloodPressure SkinThickness Outcome

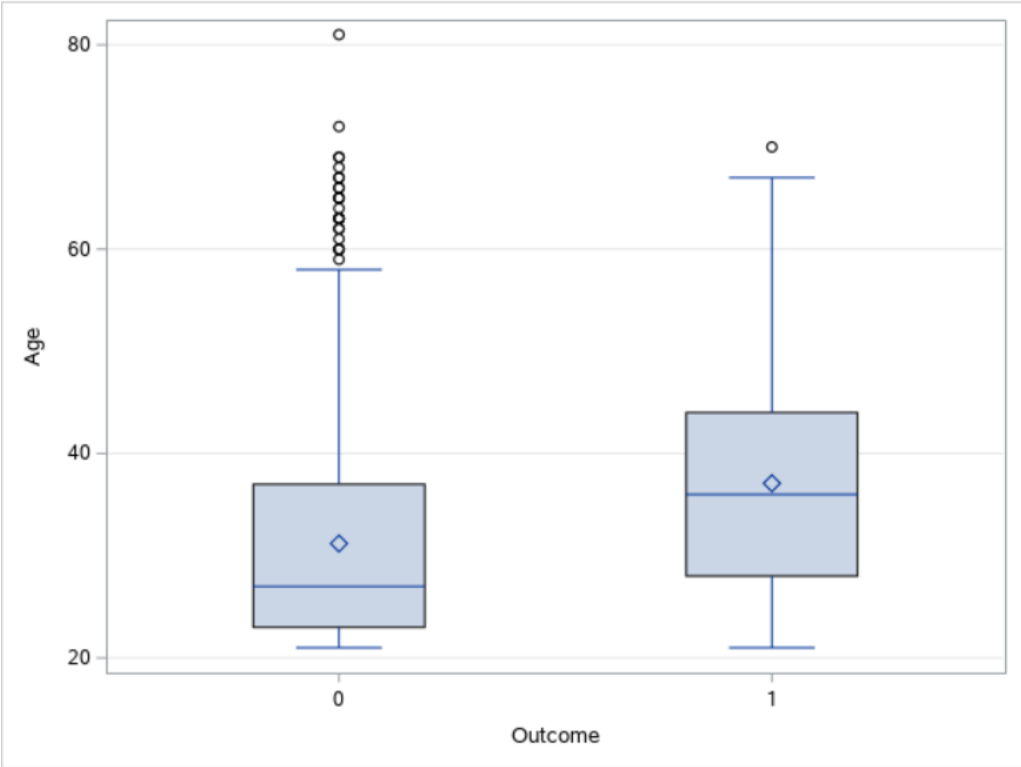
Pearson Correlation Coefficients, N = 766							
	Glucose	Age	BMI	Insulin	BloodPressure	SkinThickness	Outcome
Glucose	1.00000	0.26336	0.22135	0.33093	0.15344	0.05616	0.46786
Age	0.26336	1.00000	0.03644	-0.04261	0.24026	-0.11501	0.23882
BMI	0.22135	0.03644	1.00000	0.19843	0.28119	0.39427	0.29449
Insulin	0.33093	-0.04261	0.19843	1.00000	0.09038	0.43567	0.13141
BloodPressure	0.15344	0.24026	0.28119	0.09038	1.00000	0.21044	0.06591
SkinThickness	0.05616	-0.11501	0.39427	0.43567	0.21044	1.00000	0.07610
Outcome	0.46786	0.23882	0.29449	0.13141	0.06591	0.07610	1.00000

We can see that glucose has the most correlation with the outcome. Blood pressure and skin thickness are not very significant for prediction of diabetes. Factors like age, BMI, Insulin, PDF, glucose are the most important.

```
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=cleanData;
    vbox 'Age'n  / category='Pregnancies'n;
    yaxis grid;
run;
ods graphics / reset;
```

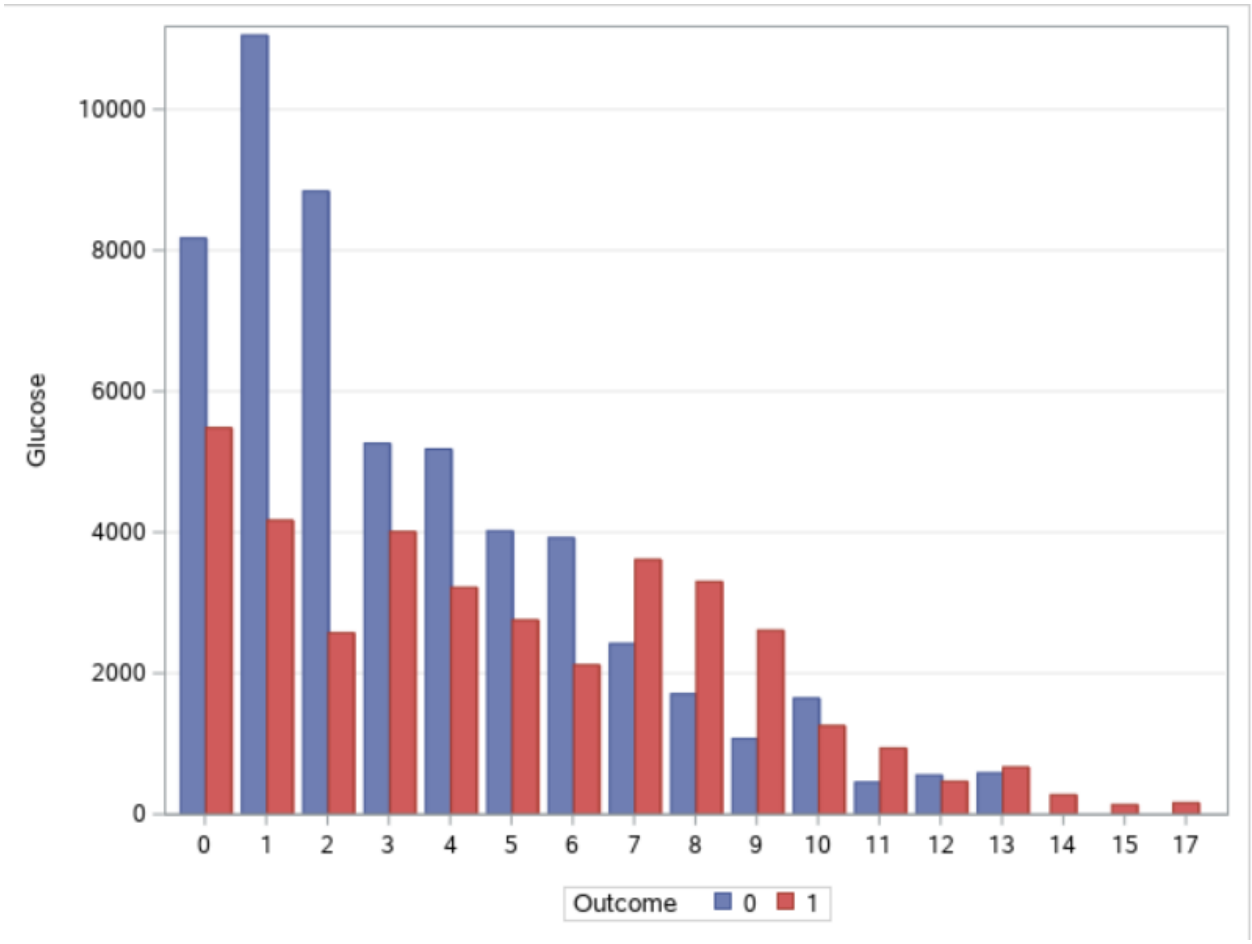



```
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=cleanData;
    vbox 'Age'n  / category='Outcome'n;
    yaxis grid;
run;
ods graphics / reset;
```



People with higher age are more likely to suffer from diabetes.

```
proc sgplot data=cleanData;
  vbar Pregnancies / response=Glucose stat=sum group=Outcome nostatlabel
         groupdisplay=cluster;
  xaxis display=(nolabel);
  yaxis grid;
  run;
```



Women with more children, more than 3 are more likely to be diabetic. This may be due to gestational diabetes after pregnancy. Women with single or no child tend to be healthier. Women with more than 13 children are almost guaranteed to suffer from gestational diabetes.

Conclusions :

- A higher BMI, reduced insulin secretion and action, a family history of diabetes, blood pressure, glucose status, and pregnancy status are common risk factors for type 2 diabetes.
- Five main predictors for diabetes are frequency of pregnancies in women, glucose, pedigree, BMI, and age.
- Blood pressure, skin thickness and insulin have no significant correlation with diabetes.
- People with higher glucose are more likely to develop diabetes. It may be because glucose is associated with insulin response.
- A higher BMI results in obesity, which could increase the fat content of the pancreas and might affect the function of pancreatic cells. Obesity could also lead to insulin resistance.

- Age is a risk factor for the onset of diabetes. Pancreatic cells lead to the decline of glucose sensitivity and impaired insulin secretion with ageing.
- We should also consider other factors such as genetic traits, gender, socio-economic status, physical activities, smoking, health information and attitude, food consumption, and spending to predict diabetes in a more generalised population.