ALY 6000 Prof- Richard He

Final Report M6

Date: 20-02-2022

-Tanvi Jain

Introduction

I have chosen a dataset related to diabetes prediction. The survey findings were analyzed, and a data analysis approach was devised. In this project, the ability to show and comprehend facts was demonstrated. This research takes a more in-depth look at the subject. The R code and script used to make the visuals are included in this report.

Attributes:
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)
Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

Part 1

1. Importing the datasets

```
> diabetes <- read.csv('/Users/tanvi/Downloads/diabetes.csv')
```

```
> head(diabetes)
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
1           6     148            72            35       0 33.6
2           1      85            66            29       0 26.6
3           8     183            64             0       0 23.3
4           1      89            66            23      94 28.1
5           0     137            40            35     168 43.1
6           5     116            74             0       0 25.6
  DiabetesPedigreeFunction Age Outcome
1                    0.627  50       1
2                    0.351  31       0
3                    0.672  32       1
4                    0.167  21       0
5                    2.288  33       1
6                    0.201  30       0
```

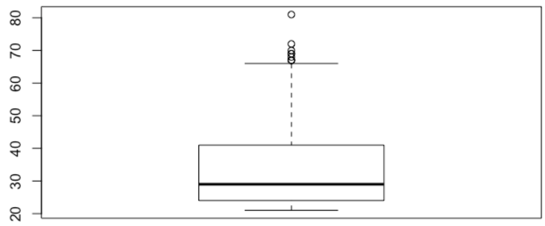
```
> tail(diabetes)
    Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
763           9      89            62         20.54    79.8 22.5
764          10     101            76         48.00   180.0 32.9
765           2     122            70         27.00    79.8 36.8
766           5     121            72         23.00   112.0 26.2
767           1     126            60         20.54    79.8 30.1
768           1      93            70         31.00    79.8 30.4
    DiabetesPedigreeFunction Age Outcome    BMI_cat Glucose_cat
763                    0.142  33       0    Healthy      Normal
764                    0.171  63       0    Obesity      Normal
765                    0.340  27       0    Obesity      Normal
766                    0.245  30       0 Overweight      Normal
767                    0.349  47       1    Obesity      Normal
768                    0.315  23       0    Obesity      Normal
    Insulin_cat
763      Normal
764    Abnormal
765      Normal
766      Normal
767      Normal
768      Normal
```

2.

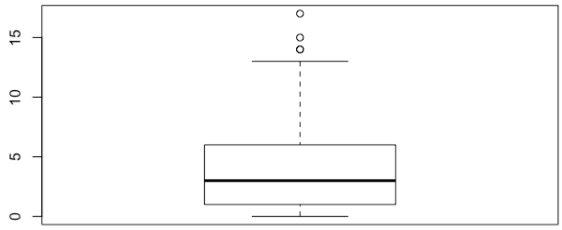
```
> str(diabetes)
'data.frame':	768 obs. of  9 variables:
 $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

3. Boxplots

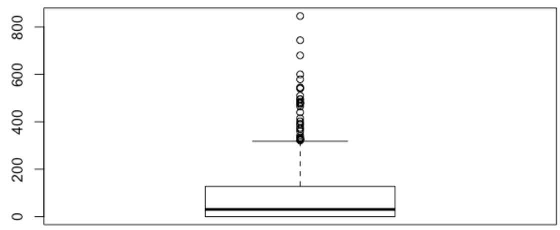
Age



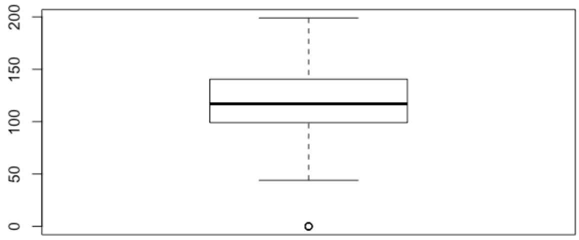
Pregnancies



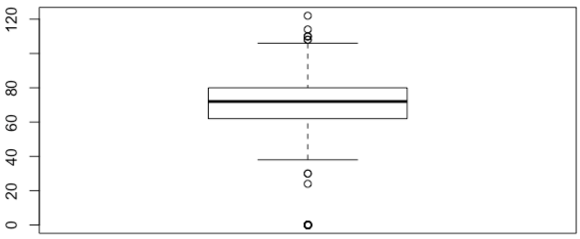
Insulin



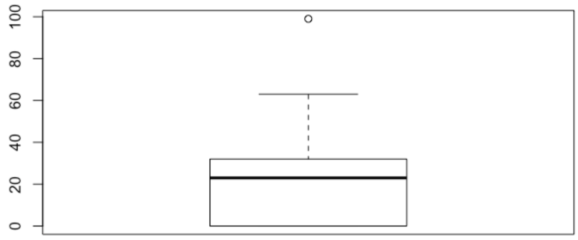
Glucose



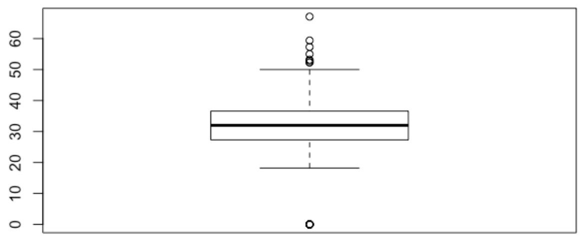
Blood Pressure



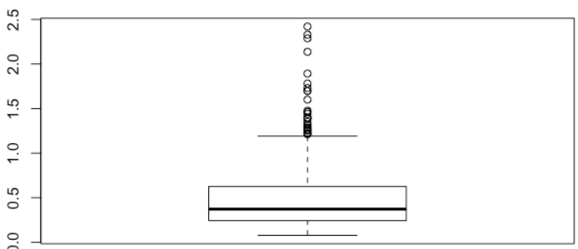
Skin Thickness



BMI



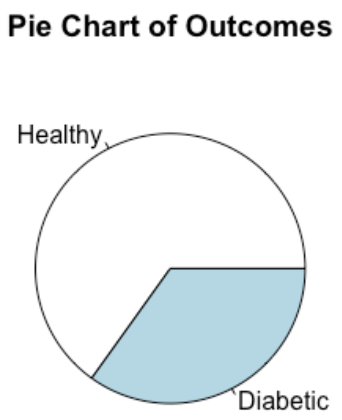
Diabetes pedigree function



We see a significant amount of data points which are greater than the 100 percentile mark for the features: Insulin and DiabetesPedigreeFunction. I have decided against removing these data points as they would possibly lead to a significant loss of information.

4.

```
> a<-as.data.frame(table(diabetes$Outcome))
> pie(a$Freq, labels = c('Healthy','Diabetic'), main="Pie Chart of Outcomes")
```



From the above pie chart we can conclude that we have an unbalanced dataset. We can see that we have a higher number of healthy people >50%.

5.

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

6. From the above summary we can see that the following features: BMI, Glucose, Insulin, SkinThickness, BloodPeessure have 0’s in them. The 0

values of these features indicate missing data as these counts can’t be zero for a human. I will be replacing the 0’s with the corresponding mean from the table.

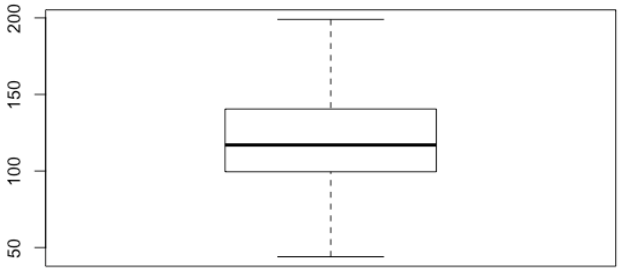
7.

```
> diabetes['Glucose'][diabetes['Glucose']==0]=120.9
> diabetes['BloodPressure'][diabetes['BloodPressure']==0]=69.11
> diabetes['SkinThickness'][diabetes['SkinThickness']==0]=20.54
> diabetes['Insulin'][diabetes['Insulin']==0]=79.8
> diabetes['BMI'][diabetes['BMI']==0]=31.99
```

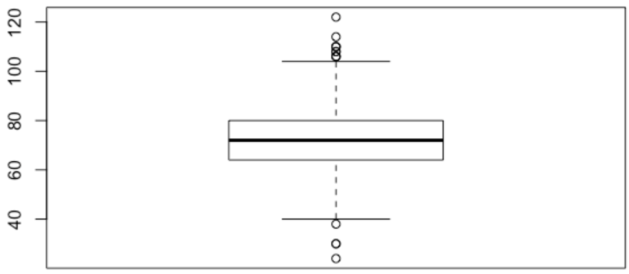
8.

Making box plots

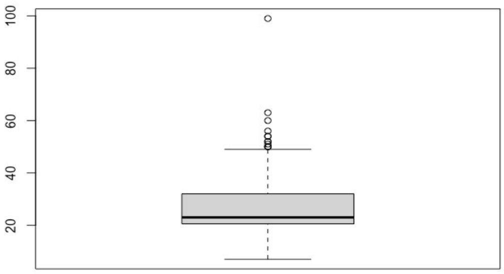
- Glucose



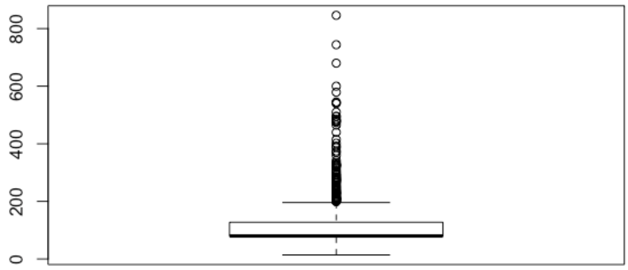
- BloodPressure



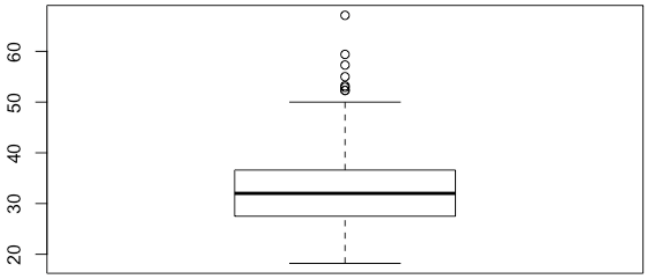
- SkinThickness



- Insulin



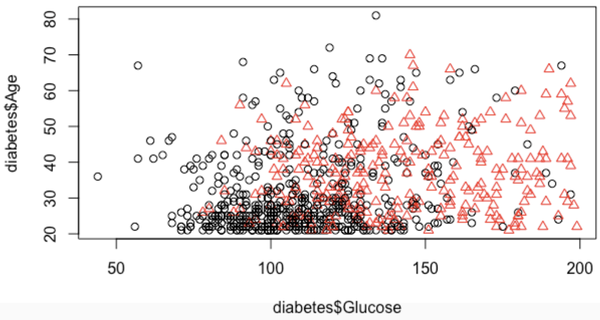
- BMI



We can see significant changes in the box plots for the following 5 features: Glucose, BloodPressure, SkinThinckness, Insulin, BMI after replacing the 0's with the corresponding means obtained from the result of summary(diabetes).

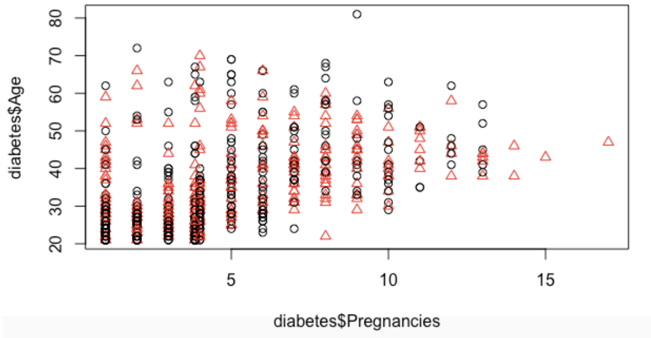
9.

```
> group <- as.factor(ifelse(diabetes$Outcome == 0, "Group 1", "Group 2"))
> plot(diabetes$Glucose, diabetes$Age, pch = as.numeric(group), col=group)
```



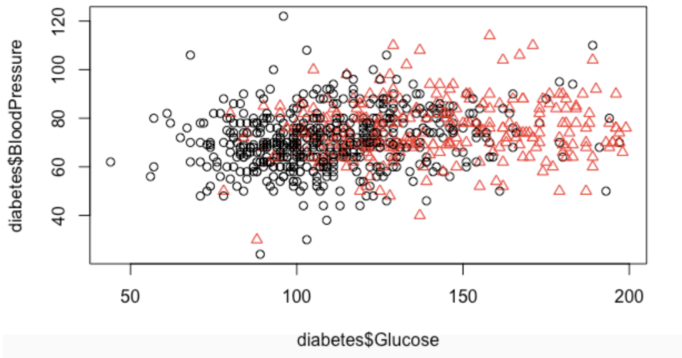
In the above graph we can see that the healthy people are concentrated <= the age of 30 years and <= glucose level of 120.
The red triangles represent people women who are diabetic.

```
> plot(diabetes$Pregnancies, diabetes$Age, pch = as.numeric(group), col=group)
```



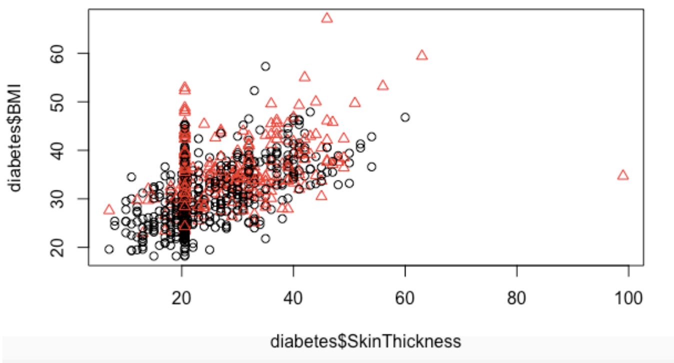
The above graph shows that the pregnancies number is <= 6 and the age <= 30 for healthy women. The red triangles represent women who are diabetic.

```
> plot(diabetes$Glucose, diabetes$BloodPressure, pch = as.numeric(group), col=group)
```



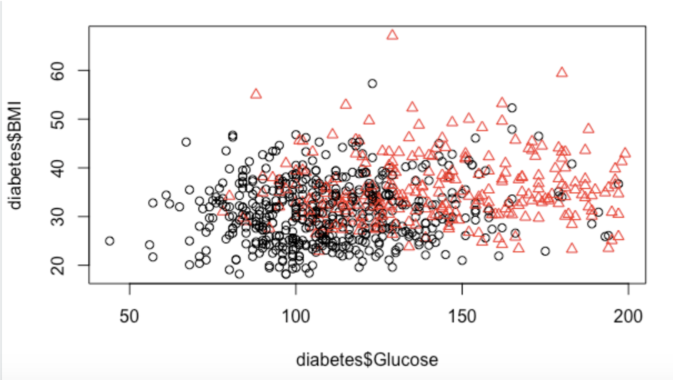
In the above graph Glucose should be <=110 and the Blood pressure should be <=80 as we can see that the black circles are concentrated in this area of the graph. The red triangles represent women who are diabetic.

```
> plot(diabetes$SkinThickness, diabetes$BMI, pch = as.numeric(group), col=group)
```



The above graph shows than BMI <=30 and skin thickness is <=20.The red triangles represent women who are diabetic.

```
> plot(diabetes$Glucose, diabetes$BMI, pch = as.numeric(group), col=group)
```



The above graph shows glucose<= 105 and BMI<=30.The red triangles represent women who are diabetic.

Part 2

Added attributes

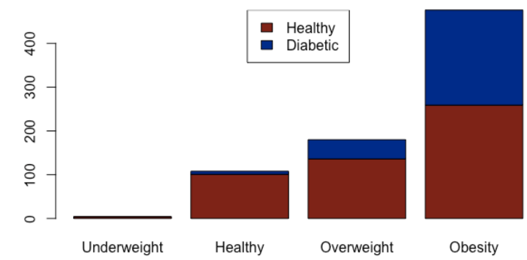
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	BMI_cat	Glucose_cat	Insulin_cat
1	6.000	148.0	72.00	35.00	79.8	33.60	0.627	50	1	Obesity	Prediabetic	Normal
2	1.000	85.0	66.00	29.00	79.8	26.60	0.351	31	0	Overweight	Normal	Normal
3	8.000	183.0	64.00	20.54	79.8	23.30	0.672	32	1	Healthy	Prediabetic	Normal
4	1.000	89.0	66.00	23.00	94.0	28.10	0.167	21	0	Overweight	Normal	Normal
5	3.845	137.0	40.00	35.00	168.0	43.10	2.288	33	1	Obesity	Normal	Abnormal
6	5.000	116.0	74.00	20.54	79.8	25.60	0.201	30	0	Overweight	Normal	Normal
7	3.000	78.0	50.00	32.00	88.0	31.00	0.248	26	1	Obesity	Normal	Normal
8	10.000	115.0	69.11	20.54	79.8	35.30	0.134	29	0	Obesity	Normal	Normal
9	2.000	197.0	70.00	45.00	543.0	30.50	0.158	53	1	Obesity	Prediabetic	Abnormal
10	8.000	125.0	96.00	20.54	79.8	31.99	0.232	54	1	Obesity	Normal	Normal
11	4.000	110.0	92.00	20.54	79.8	37.60	0.191	30	0	Obesity	Normal	Normal
12	10.000	168.0	74.00	20.54	79.8	38.00	0.537	34	1	Obesity	Prediabetic	Normal
13	10.000	139.0	80.00	20.54	79.8	27.10	1.441	57	0	Overweight	Normal	Normal
14	1.000	189.0	60.00	23.00	846.0	30.10	0.398	59	1	Obesity	Prediabetic	Abnormal

During my research about various features I found out that the following features- BMI, Glucose and Insulin have been categorised in specific categories according to well defined levels across the healthcare industry. As we wanted to see how the categorical variables will affect the distribution of data and graphs have also been plotted. I want to see how the distribution of healthy and diabetic people is in aforementioned well defined levels for each of the features individually.

a) BMI is a person’s weight in kilograms divided by the square of height in meters. A high BMI can indicate high body fatness.

- If your BMI is less than 18.5, it falls within the **underweight** range.
- If your BMI is 18.5 to <25, it falls within the **healthy weight** range.
- If your BMI is 25.0 to <30, it falls within the **overweight** range.
- If your BMI is 30.0 or higher, it falls within the **obesity** range.

```
> diabetes$BMI_cat<- cut(diabetes$BMI, breaks = c(0,18.5,25,30,67.10), labels =
 c('Underweight','Healthy','Overweight','Obesity'))
> barplot(table(diabetes$Outcome,diabetes$BMI_cat),col=c('darkred','darkblue'))
> legend("top", legend = c('Healthy','Diabetic'), fill =c('darkred','darkblue'))
```

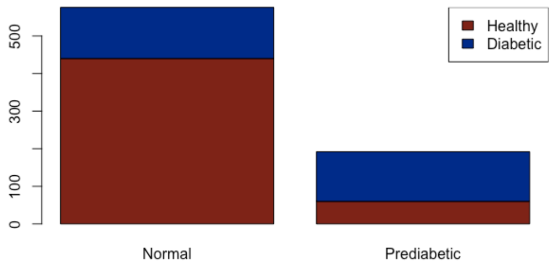


We can see most of the women in heathy and overweight BMI categories are the ones without diabetes. However the women who are obese have a 50% chance of being diabetic. We can also see that none of the women who are underweight are diabetic.

b)A blood sugar level less than 140 mg/dL (7.8 mmol/L) is considered normal.

A blood sugar level from 140 to 199 mg/dL (7.8 to 11.0 mmol/L) is considered pre-diabetes. This is sometimes referred to as impaired glucose tolerance.

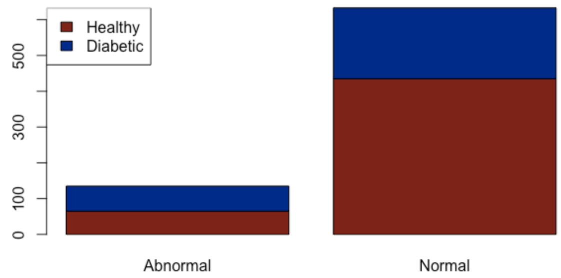
```
> diabetes$Glucose_cat<- cut(diabetes$Glucose, breaks = c(0,140,199), labels = c
('Normal','Prediabetic'))
> barplot(table(diabetes$Outcome,diabetes$Glucose_cat),col=c('darkred','darkblu
e'))
> legend("topright", legend = c('Healthy','Diabetic'), fill =c('darkred','darkbl
ue'))
```



We can see that about 75% of the women who have normal glucose levels are healthy. Whereas about 70% of women with prediabetic glucose levels are experiencing diabetes.

c) Insulin is a hormone (a chemical substance that acts as a messenger in the human body) that is secreted by an abdominal organ called the pancreas. Insulin levels in the range of 16<=x<=166 are considered to be normal.

```
> diabetes$Insulin_cat<- cut(diabetes$Insulin, breaks = c(0,16,166,846), labels
 = c('Abnormal','Normal','Abnormal'))
> barplot(table(diabetes$Outcome,diabetes$Insulin_cat),col=c('darkred','darkblu
e'))
> legend("topleft", legend = c('Healthy','Diabetic'), fill =c('darkred','darkblu
e'))
```



We can see that about 70% of the women who have normal insulin level don't have diabetes. Whereas, women who have an abnormal insulin level have a 50% chance of being diabetic.

Mean

```
> mean(diabetes$BloodPressure)
[1] 72.25501
> mean(diabetes$BMI)
[1] 32.45077
> mean(diabetes$Glucose)
[1] 121.6816
```

Median

```
> median(diabetes$Glucose)
[1] 117
> median(diabetes$BMI)
[1] 32
> median(diabetes$BloodPressure)
[1] 72
```

Part 3

Results and conclusions:

The primary goal of this study is to evaluate and analyze Diabetes Prediction in order to provide insights and a Health chart of women. I decided to choose something related to health. Diabetes is a leading cause of mortality worldwide. Early detection of diseases such as diabetes can be managed and human lives saved. To do so, this analysis looks at diabetes prediction using a variety of diabetes-related variables. I found a data set on diabetes prediction on kaggle. This dataset consists 9 attributes 768 records. The different attributes are Pregnancies, Age, Insulin, Blood Pressure etc. data set is specifically for women above the age of 21 of Pima Indian origin. Obesity and diabetes have become more prevalent in the Pima's during the last century, possibly as a result of fast cultural and nutritional changes in a people genetically predisposed to diabetes. There was also no missing values in the data set but lot of cleaning was required as lot of values were mentioned 0 which were not possible and did not make sense. I then replaced that data with the corresponding mean values of appropriate ranges. The greatest issue I ran into was that this data collection had a lot of information that I didn't need for my study. As a result, I had to first filter and compress the data to just the factors I was interested in studying.

References:

https://www.openml.org/d/37