Exploratory data Analysis (EDA) and Data Visualization using ggplot

```
#Loading essential libraries
library(tidyverse)
library(ggplot2)
library(readr)
library(scales)
library(dplyr)
library(reshape2)
library(readxl)
library(corrplot)
library(Hmisc)
library(ggalt)
```

```
#Importing dataset (taken from kaggle)
> data <- read.csv("C:/Users/heena/Downloads/diabetes1.csv")
```

EDA

```
#Returns first 6 rows
> head(data)
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction
Age Outcome
1           6     148            72            35       0 33.6                    0.627  50       1
2           1      NA            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
```

```
> tail(data)
#Returns last 6 rows

    Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction
Age Outcome
763           9      89            62             0       0 22.5                    0.142  33       0
764          10     101            76            48     180 32.9                    0.171  63       0
765           2     122            70            27       0 36.8                    0.340  27       0
766           5     121            72            23     112 26.2                    0.245  30       0
767           1     126            60             0       0 30.1                    0.349  47       1
768           1      93            70            31       0 30.4                    0.315  23       0
```

```
#Returns dimensions of the dataset
> dim(data)
[1] 768   9
```

```
#Returns number of rows and columns respectively
> nrow(data)
[1] 768
```

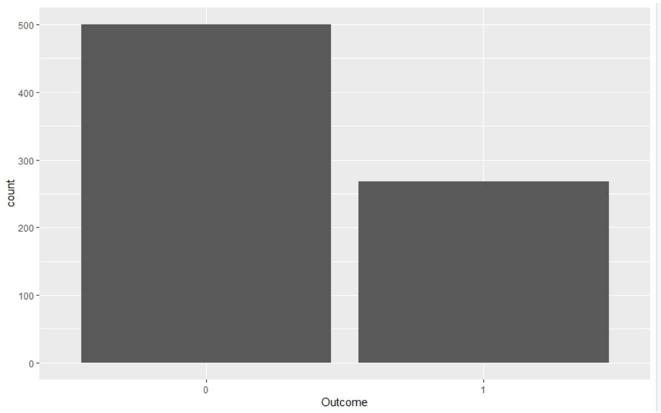
```
> ncol(data)
[1] 9

> str(data)
'data.frame':	768 obs. of  9 variables:
 $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                 : int  148 NA 183 89 137 116 78 115 197 NA ...
 $ 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 ...

 #statistics of dataset
> summary(data)
  Pregnancies       Glucose      BloodPressure    SkinThickness      Insulin          BMI
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5
Median :32.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99
 3rd Qu.: 6.000   3rd Qu.:140.8   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10
                  NA's   :2
 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

apply(data, function)
#datatype of each column as vector
> sapply(data, typeof)
             Pregnancies                  Glucose            BloodPressure            SkinThickness
               "integer"                "integer"                "integer"                "integer"
                 Insulin                      BMI DiabetesPedigreeFunction                      Age
               "integer"                 "double"                 "double"                "integer"
                 Outcome
               "integer"

#datatype of each column as list
> lapply(data, typeof)
$Pregnancies
[1] "integer"

$Glucose
[1] "integer"
```

```
$BloodPressure
[1] "integer"

$SkinThickness
[1] "integer"

$Insulin
[1] "integer"

$BMI
[1] "double"

$DiabetesPedigreeFunction
[1] "double"

$Age
[1] "integer"

$Outcome
[1] "integer"


> colnames(data)
[1] "Pregnancies"              "Glucose"                  "BloodPressure"            "SkinThickness"
[5] "Insulin"                  "BMI"                      "DiabetesPedigreeFunction" "Age"
[9] "Outcome"

#Returns NA values in each column
> colSums(is.na(data))
             Pregnancies                  Glucose            BloodPressure            SkinThickness
                       0                        2                        0                        0
                 Insulin                      BMI DiabetesPedigreeFunction                      Age
                       0                        0                        0                        0
                 Outcome
                       0

#Replacing those values with median; na.rm for removing NA values
> data$Glucose[is.na(data$Glucose)]=median(data$Glucose, na.rm=TRUE)
> colSums(is.na(data))
             Pregnancies                  Glucose            BloodPressure            SkinThickness
                       0                        0                        0                        0
                 Insulin                      BMI DiabetesPedigreeFunction                      Age
                       0                        0                        0                        0
                 Outcome
                       0
> head(data)
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction
Age Outcome
1           6     148            72            35       0 33.6                    0.627  50       1
2           1     117            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
#We see the NA values replaced with median of column(here, 117).

#Converting Outcome variable from int into factor for classification
> data$Outcome <- as.factor(data$Outcome)
> class(data$Outcome)
[1] "factor"

#frequency table for outcomes
> table(data$Outcome)

  0   1
500 268
```

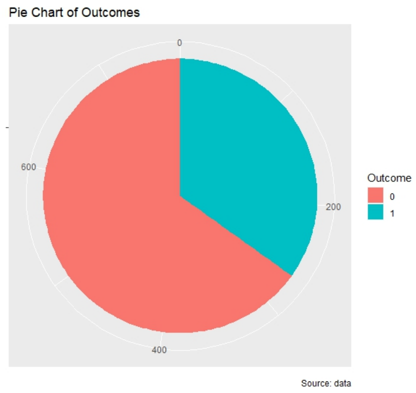
Data Visualization

```
#Bar chart for comparison of people with and without diabetes
> p1 <- ggplot(data,aes(x=Outcome)) +
+ geom_bar()
> print(p1)
```



We see an imbalance in the dataset and a bias towards people without diabetes.

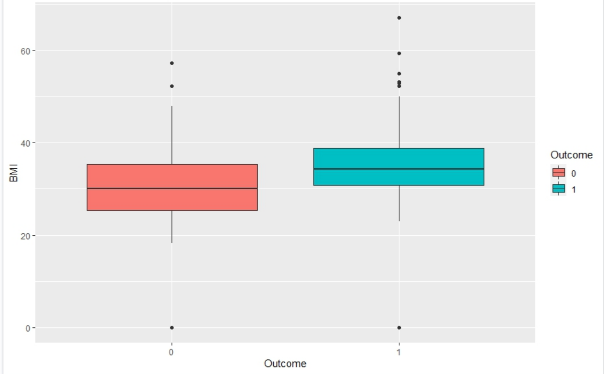
```
 #Pie chart for comparison of people with and without diabetes
> p1b <- ggplot(data, aes(x="",fill = Outcome)) +
  geom_bar() +
  labs(fill="Outcome",
       x=NULL,
       y=NULL,
       title="Pie Chart of Outcomes",
       caption="Source: data")

> p1b + coord_polar(theta = "y", start=0)
```



Ways to deal with the imbalance : Undersampling,oversampling,SMOTE, Adasyn,etc.

#Box plot for BMI of people with and without diabetes

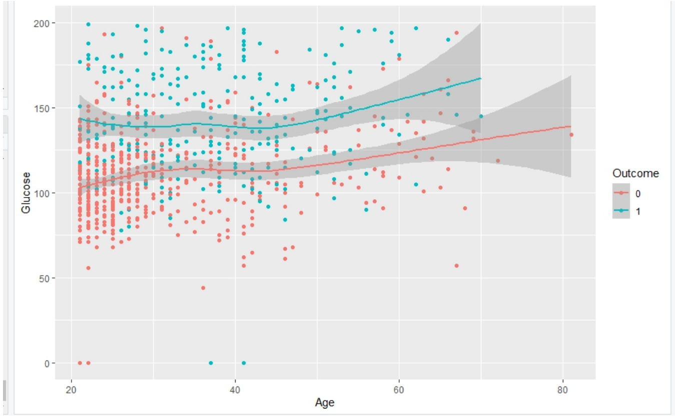
```
> p2 <- ggplot(data, aes(x=Outcome,y=BMI,fill=Outcome))+
+   geom_boxplot()
> print(p2)
```



We see BMI of non-diabetic patients is considerably lesser than those with diabetes.

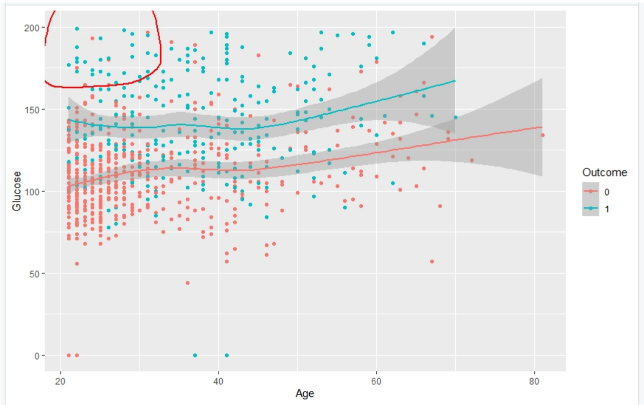
#Using scatter plot to observe correlation between age and glucose. The color of data points is based on Outcome.

```
> p3 <- ggplot(data, aes(x=Age,y=Glucose,col=Outcome))+
+   geom_point()
> p3+geom_smooth(method="loess")
```



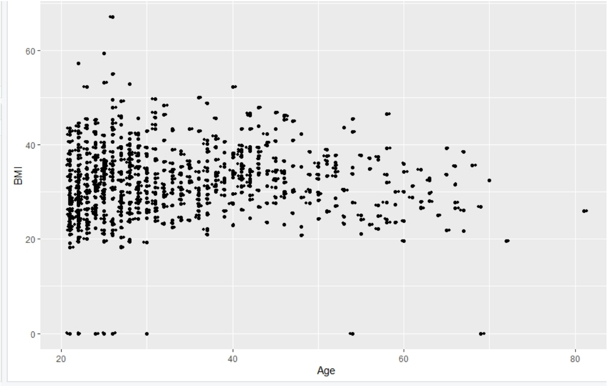
Glucose level of diabetic patients is seen to be higher. Also, as Age increases, a slight increase is seen in glucose levels.

```
#Scatterplot with Encircling
> subset <- data[data$Glucose > 175 & data$Age<30,]
> p4 <- ggplot(data, aes(x=Age,y=Glucose,col=Outcome))+
+   geom_point()
> p4+geom_smooth(method="loess")+
  geom_encircle(aes(x=Age, y=Glucose),
             data=subset,
              color="red",
              size=2,
              expand=0.08)
```



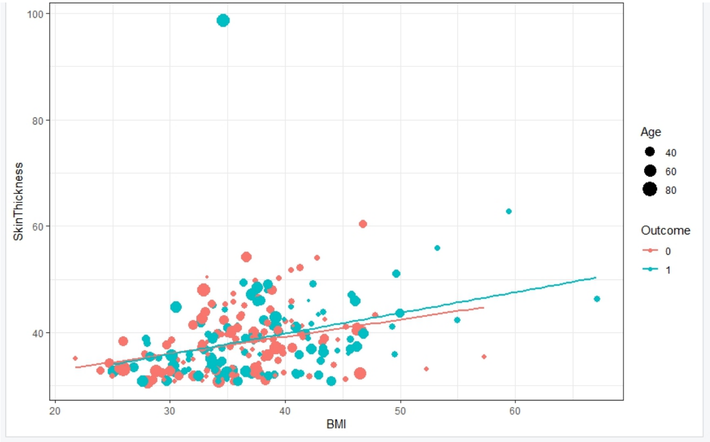
Circled area shows youngest people with the highest Glucose levels. We see that most of them do have diabetes.

```
 #Age vs BMI jitter plot
> p5 <- ggplot(data, aes(x=Age,y=BMI))+
+   geom_point()
> p5+geom_jitter(width = .4, size=1)
#p5+geom_jitter(aes(colour=Outcome))
```



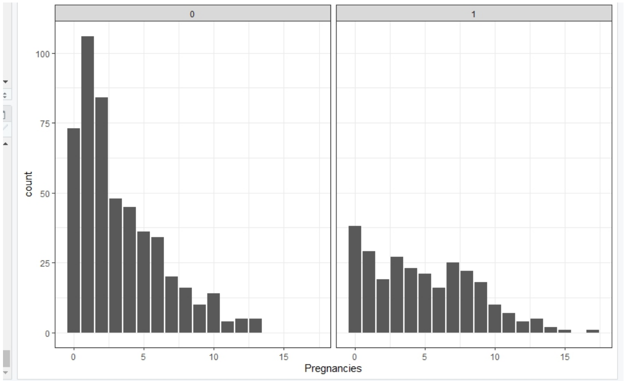
Most young people have BMI in the range of 20-35, with exception of some outliers.

```
#Bubble plot for BMI vs Skin Thickness. It also shows relationship with Outcome(color variation) and Age (by variation in size of data points).
> subset1 <- data[data$SkinThickness>30,]
> theme_set(theme_bw())  # pre-set the bw theme.
> p6 <- ggplot(subset1, aes(x=BMI,y=SkinThickness))

> p6 + geom_jitter(aes(col=Outcome, size=Age)) +
+   geom_smooth(aes(col=Outcome), method="lm", se=F)
```

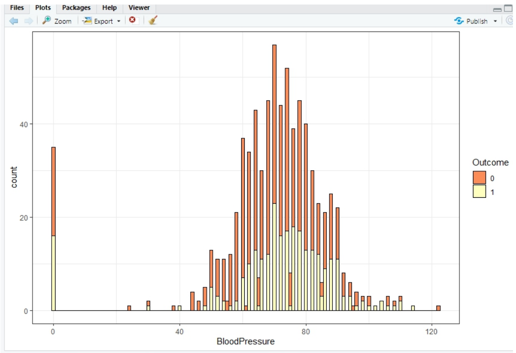


As BMI increases, so does skin thickness (linear relationship). Also, skin thickness is lesser for older people.

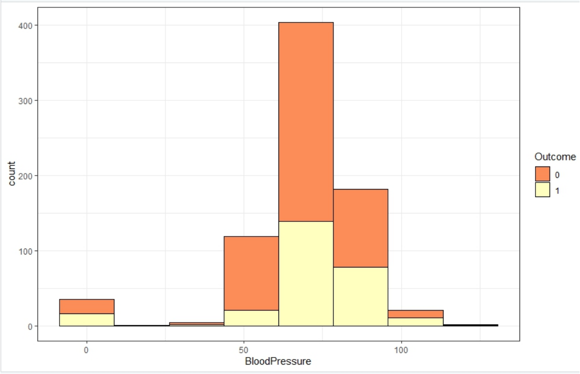
```
# Bar graph for frequency of Pregnancies with respect to Outcome
> p7 <- ggplot(data, aes(Pregnancies))
> p7 + geom_bar(aes(group=Outcome)) + facet_wrap(~Outcome)
```



```
# Histogram for Blood Pressure
> p8 <- ggplot(data, aes(BloodPressure)) + scale_fill_brewer(palette = "Spectral")
>
> p8 + geom_histogram(aes(fill=Outcome),
            binwidth = 1, #width of bars
            col="black", #color of boundary between bars
            size=.5)
```

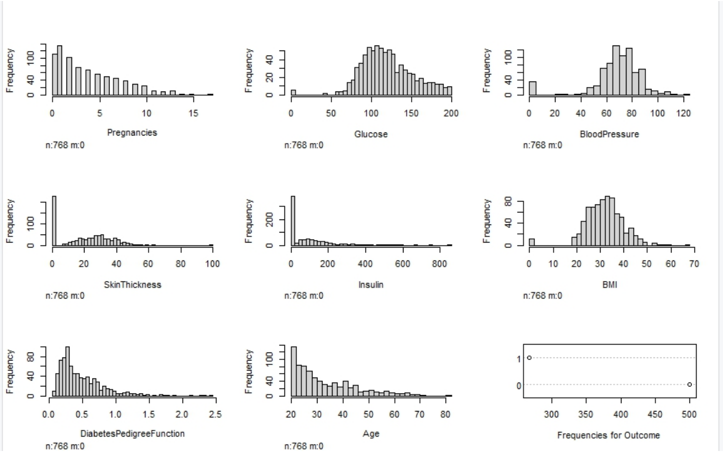


```
> p8 + geom_histogram(aes(fill=Outcome),
                  bins=8, #number of bars
                 col="black",
                 size=.5)
```



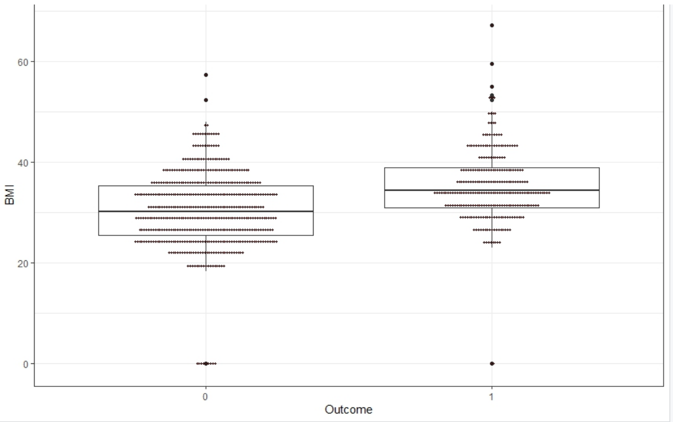
Most people had blood pressure around 70-80; and most of these people did not have diabetes.

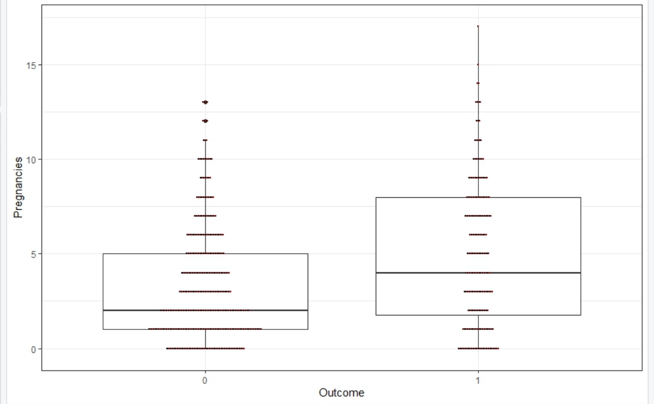
```
#Histogram for all columns
> hist.data.frame(data)
```



Can BMI, Number of pregnancies influence Outcome?

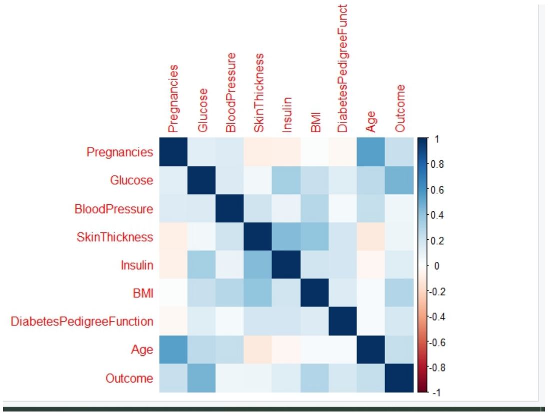
```
> p9 <- ggplot(data, aes(Outcome,BMI))
> p9 + geom_boxplot() +
  geom_dotplot(binaxis='y',
            stackdir='center',
             dotsize = .2,
             fill="red")
> p9b <- ggplot(data, aes(Outcome,Pregnancies))
> p9b + geom_boxplot() +
  geom_dotplot(binaxis='y',
             stackdir='center',
             dotsize = .1,
             fill="red")
```





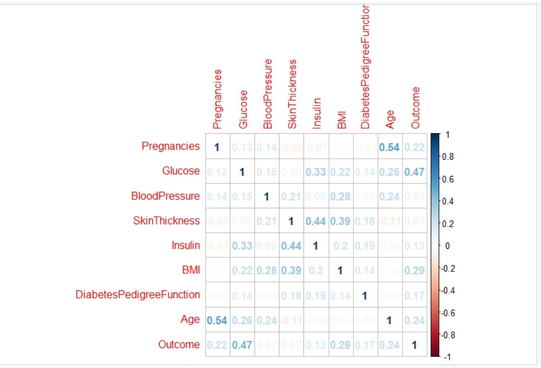
We see that people with diabetes have a higher BMI and number of pregnancies than those who tested negative. Thus, higher BMI would imply higher risk of diabetes.

```
#Correlation matrix
> dat <- read.csv("C:/Users/heena/Downloads/diabetes1.csv")
> num_vars <- unlist(lapply(dat, is.numeric))
> dia_nums <- data[ , num_vars]
>corrplot(dia_corr, method="color")
```



Highest correlation was seen between Glucose and Outcome. Also seen between age and pregnancies.

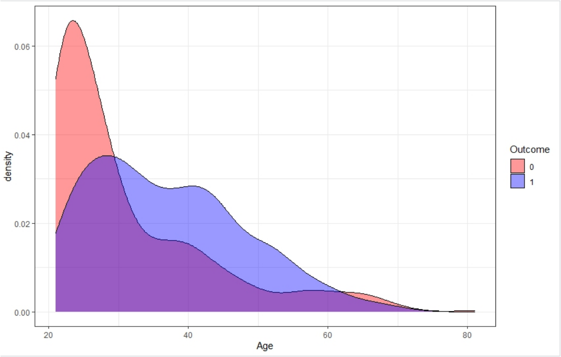
```
> #for more specific results and removing the need to rely on visual perception.
> corrplot(dia_corr, method="number")
```



Considerable relation between skin thickness and insulin.
Inverse relation seen between age and skin thickness.

#Impact of Age over the Outcome (density plot)

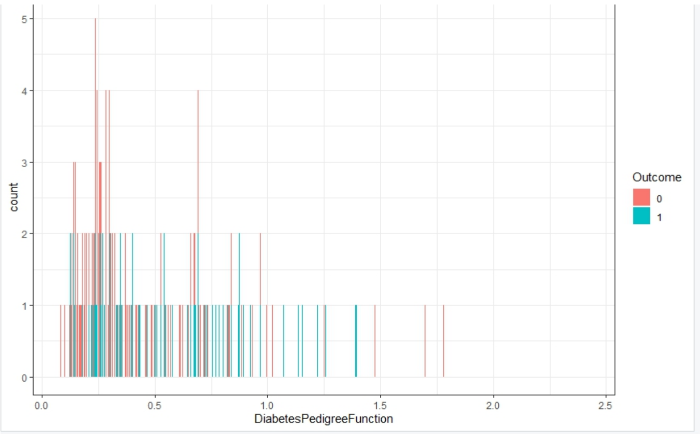
```
>
ggplot(data,aes(x=Age,fill=Outcome))+geom_density(alpha=0.4)+scale_fill_manual(values=c(
"red", "blue"))
```



People with diabetes were seen to be comparatively older.

#Is Outcome related to Diabetes Pedigree Function?

```
> p10 <- ggplot(data,aes(x=DiabetesPedigreeFunction))
> p10 + geom_bar(aes(fill=Outcome))
```



Most people who do not have diabetes were seen to have a Diabetes Pedigree Function less than 0.8.