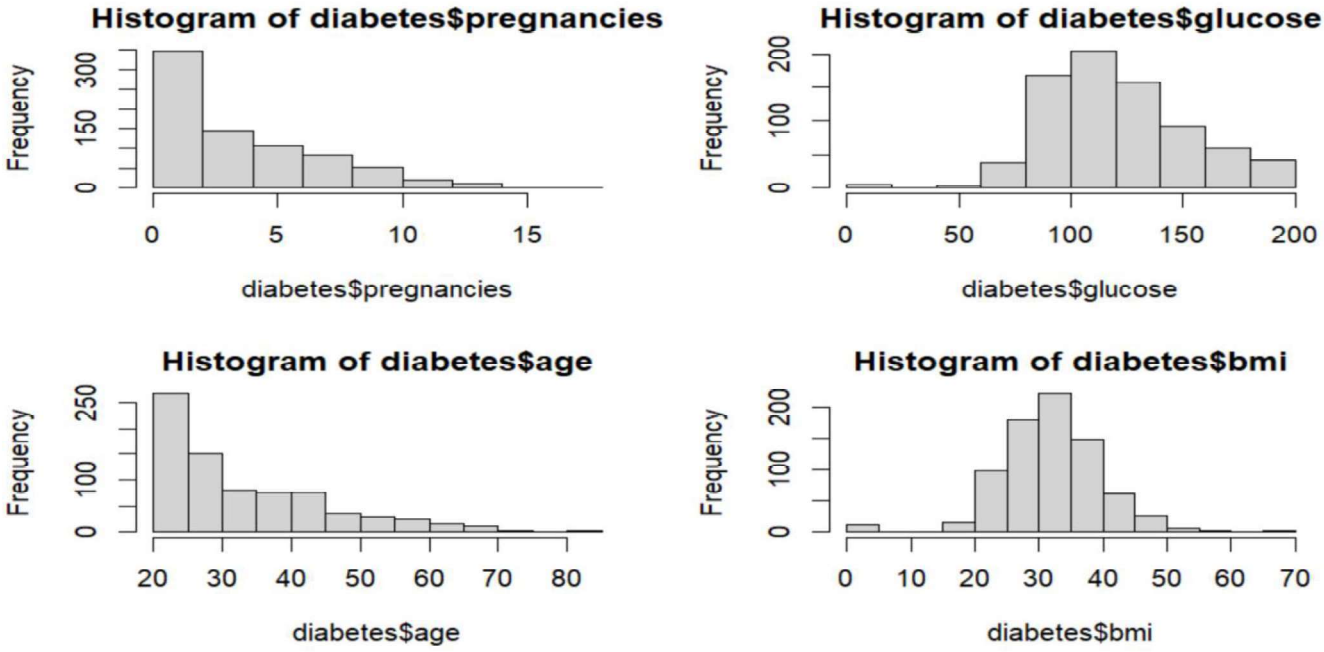
RESULTS AND ANALYSIS

a. Univariate analysis

```
{r}
par(mfrow=c(2,2))
hist(diabetes$pregnancies)
hist(diabetes$glucose)
hist(diabetes$age)
hist(diabetes$bmi)
```



By univariate analysis of variables,

Age and number of times pregnant are not in normal distributions as expected since the underlying population should not be normally distributed either.

Glucose level and BMI are following a normal distribution.

b. The shapiro.test() function in R

It is used to perform the Shapiro-Wilk test for normality. The Shapiro-Wilk test is a statistical test used to assess whether a sample of data comes from a normally distributed population.

```
{r}
shapiro.test(diabetes$pregnancies)
```

```
        Shapiro-Wilk normality test

data:  diabetes$pregnancies
W = 0.90428, p-value < 2.2e-16
```

```
{r}
shapiro.test(diabetes$glucose)
```

```
        Shapiro-Wilk normality test

data:  diabetes$glucose
W = 0.9701, p-value = 1.986e-11
```

```
{r}
shapiro.test(diabetes$bmi)
```

```
        Shapiro-Wilk normality test

data:  diabetes$bmi
W = 0.94999, p-value = 1.842e-15
```

```
{r}
shapiro.test(diabetes$age)
```

```
        Shapiro-Wilk normality test

data:  diabetes$age
W = 0.87477, p-value < 2.2e-16
```

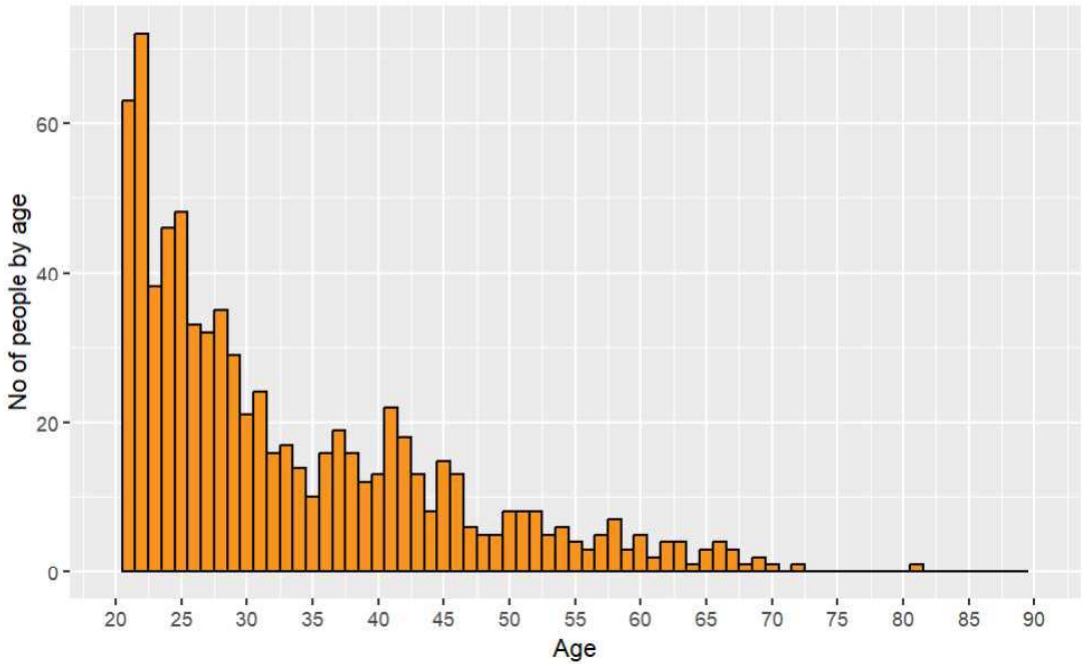
Here Shapiro.test() also shows that BMI and Glucose data originated from normal distribution

P values is greater than 0.05.

c. Categorize the age and plot the graph to know about which age category is suffered from diabetes

```
{r}
#age category column

diabetes$Age_Cat <- ifelse(diabetes$age < 21, "<21",
                   ifelse((diabetes$age>=21) & (diabetes$age<=25), "21-25",
                   ifelse((diabetes$age>25) & (diabetes$age<=30), "25-30",
                   ifelse((diabetes$age>30) & (diabetes$age<=35), "30-35",
                   ifelse((diabetes$age>35) & (diabetes$age<=40), "35-40",
                   ifelse((diabetes$age>40) & (diabetes$age<=50), "40-50",
                   ifelse((diabetes$age>50) & (diabetes$age<=60), "50-60",">60")))))))
diabetes$age_Cat <- factor(diabetes$Age_Cat, levels = c('<21','21-25','25-30','30-35','35-40','40-50'
,'50-60','>60'))
table(diabetes$age_Cat)

# Histogram of Age
library(ggplot2)

ggplot(aes(x = age), data=diabetes) +
        geom_histogram(binwidth=1, color='black', fill = "#F79420") +
        scale_x_continuous(limits=c(20,90), breaks=seq(20,90,5)) +
        xlab("Age") +
        ylab("No of people by age")
```

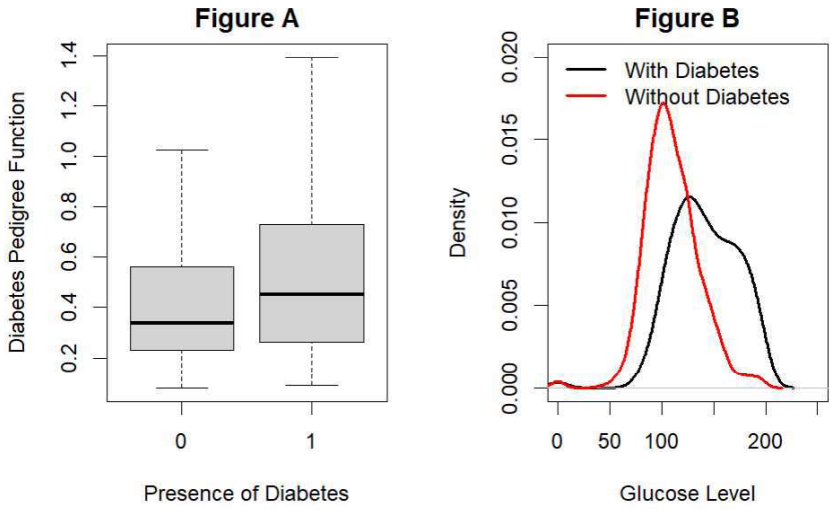
```
<21 21-25 25-30 30-35 35-40 40-50 50-60   >60
  0   267   150    81    76   113    54    27
```



- Below 21 age , no person suffered from diabetes.
- Age between 21 to 25, 267 people may have diabetes or not.
- Age between 25 to 30 , 150 people may have diabetes or not.
- Age between 30 to 35 , 81 people may have diabetes or not.
- Age between 35 to 40 , 76 people may have diabetes or not.
- Age between 40 -50 , 113 people may have diabetes or not.

d. **Other plot such as boxplot or density plot can also be used to look at the difference in values of the variables between those with diabetes and those without.**

```
{r}
 par(mfrow = c(1, 2))

# boxplot
with(diabetes, boxplot(dpf ~ outcome,
                       ylab = "Diabetes Pedigree Function",
                       xlab = "Presence of Diabetes",
                       main = "Figure A",
                       outline = FALSE))

# subsetting based on response
with <- diabetes[diabetes$outcome == 1, ]
without <- diabetes[diabetes$outcome == 0, ]

# density plot
plot(density(with$glucose),
     xlim = c(0, 250),
     ylim = c(0.00, 0.02),
     xlab = "Glucose Level",
     main = "Figure B",
     lwd = 2)
lines(density(without$glucose),
      col = "red",
      lwd = 2)
legend("topleft",
       col = c("black", "red"),
       legend = c("With Diabetes", "Without Diabetes"),
       lwd = 2,
       bty = "n")

# simple two sample t-test with unequal variance
t.test(with$dpf, without$dpf)
```



We can see from Figure A that Diabetes can be possible by prevalence of relatives in family.

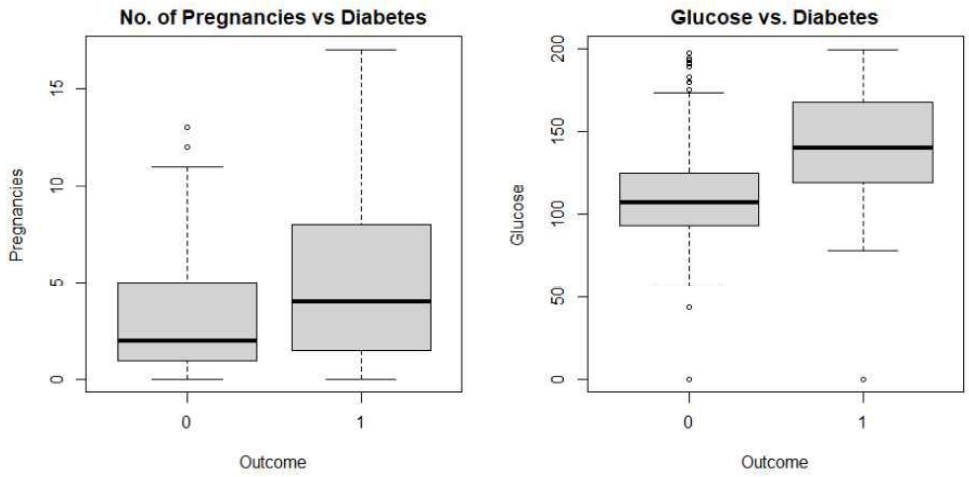
We can see from Figure B that the distribution to shifted towards the left for those without diabetes. This means those without diabetes generally have a lower blood glucose level.

```
        Welch Two Sample t-test

data:  with$dpf and without$dpf
t = 4.5768, df = 454.51, p-value = 6.1e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06891135 0.17262065
sample estimates:
mean of x mean of y
 0.550500  0.429734
```

p-value is > critical values of 0.05, so we accept the null hypothesis for the alternate hypothesis. We can say that we are, 95 % confident, that the dpf causing diabetes is < the people without causing dibetes from dpf .

e. Plotting the different variables to show how they affecting on diabetes using Boxplot().

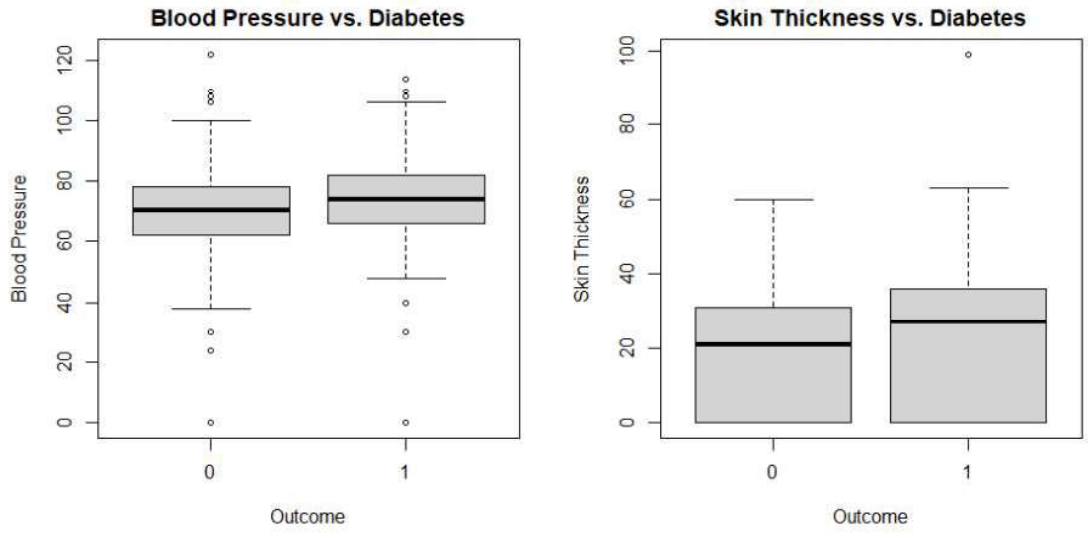
```
{r}
par(mfrow=c(2,4))
boxplot(diabetes$pregnancies~diabetes$outcome,
        main="No. of Pregnancies vs Diabetes", xlab="Outcome",         ylab="Pregnancies")

boxplot(diabetes$glucose~diabetes$outcome, main="Glucose vs. Diabetes", xlab="Outcome", ylab="Glucose")

boxplot(diabetes$bloodpressure~diabetes$outcome, main="Blood Pressure vs. Diabetes", xlab="Outcome",
ylab="Blood Pressure")

boxplot(diabetes$skinthickness~diabetes$outcome, main="Skin Thickness vs. Diabetes",xlab="Outcome",
ylab="Skin Thickness")

boxplot(diabetes$insulin~diabetes$outcome, main="Insulin vs. Diabetes", xlab="Outcome", ylab="Insulin")

boxplot(diabetes$bmi~diabetes$outcome, main="BMI vs. Diabetes", xlab="Outcome", ylab="BMI")

boxplot(diabetes$dpf~diabetes$outcome, main="Diabetes Pedigree Function vs. Diabetes", xlab="Outcome",
ylab = "DiabetesPedigreeFunction")

boxplot(diabetes$age~diabetes$outcome, main="Age vs. Diabetes", xlab="Outcome", ylab="Age")
```

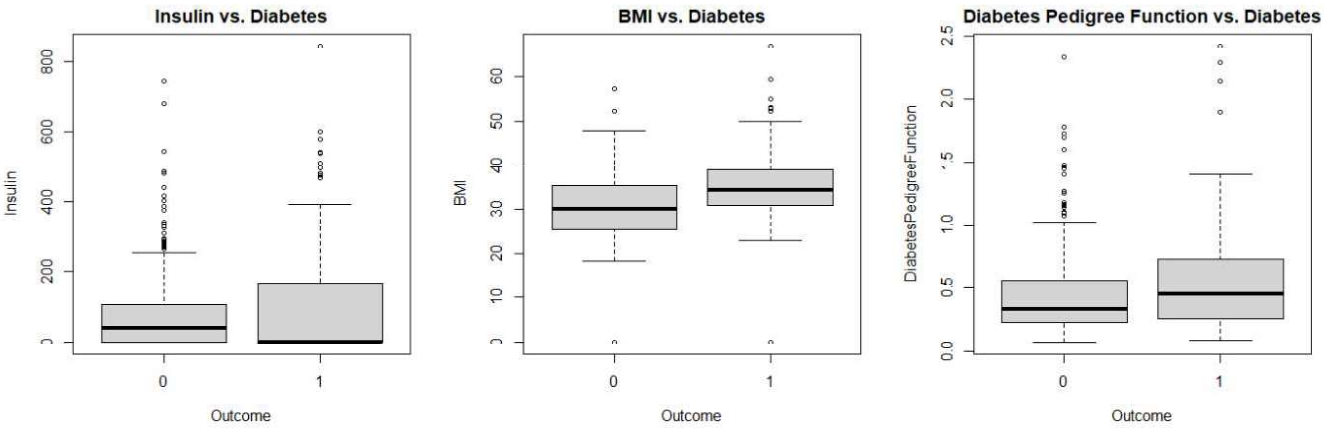


1 -> Person with diabetes.

0 -> Person without diabetes.

By above plots woman have diabetes caused by number of times pregnancy happens increase the chance of diabetes and also glucose level in body also affects on diabetes .





By above plots its shows that,

- people have high blood pressure occurrence of diabetes also more
- People's skin thickness is does not plays a main role of occurring the diabetes.
- People's having high insulin level possibility of low diabetes.
- If weight and height (BMI) is more , occurrence of diabetes is also more.

f. **To know how age factor influence on blood pressure of blood , we perform t-test**

```
{r}
t.test(diabetes$age,diabetes$bloodpressure)
```

```
        Welch Two Sample t-test

data:  diabetes$age and diabetes$bloodpressure
t = -43.884, df = 1265.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -37.46791 -34.26126
sample estimates:
mean of x mean of y
 33.24089  69.10547
```

p-value is < critical values of 0.05, so we reject the null hypothesis for the alternate hypothesis. We can say that we are, 95 % confident, that the blood pressure is depend on age .

g. **Generate the linear model to know how blood pressure varies on glucose level of body**

```
{r}
model1<-lm(bloodpressure~glucose,data = diabetes)
summary(model1)
```

```
Call:
lm(formula = bloodpressure ~ glucose, data = diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-74.843  -5.914   2.333  10.676  55.194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.93775    2.70316  21.433  < 2e-16 ***
glucose      0.09238    0.02162   4.273 2.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.14 on 766 degrees of freedom
Multiple R-squared:  0.02328,   Adjusted R-squared:  0.02201
F-statistic: 18.26 on 1 and 766 DF,  p-value: 2.17e-05
```

The dependent variable (blood pressure) is being modeled as a linear function of the independent variable (glucose).

"Residuals" section shows the difference between the observed values of the dependent variable and the values predicted by the model

"Coefficients" section shows the estimated parameters of the model.

The "Residual standard error" is a measure of how well the model fits the data, and is calculated as the square root of the mean squared error.

The "Multiple R-squared" value is a measure of the proportion of variance in the dependent variable that is explained by the independent variable, while the "Adjusted R-squared" takes into account the number of predictors in the model.

The "F-statistic" and its corresponding p-value are used to test the null hypothesis that all coefficients in the model are equal to zero.

A low p-value indicates that the **null hypothesis can be rejected**, and that the independent variable is a significant predictor of the dependent variable.

```
{r}
anova(model1,test="Chisq")
```

```
Analysis of Variance Table

Response: bloodpressure
           Df Sum Sq Mean Sq F value   Pr(>F)
glucose     1   6691  6690.6   18.26 2.17e-05 ***
Residuals 766 280664   366.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is an Analysis of Variance (ANOVA) table, which is used to determine whether the independent variable (glucose) has a significant effect on the dependent variable (bloodpressure).

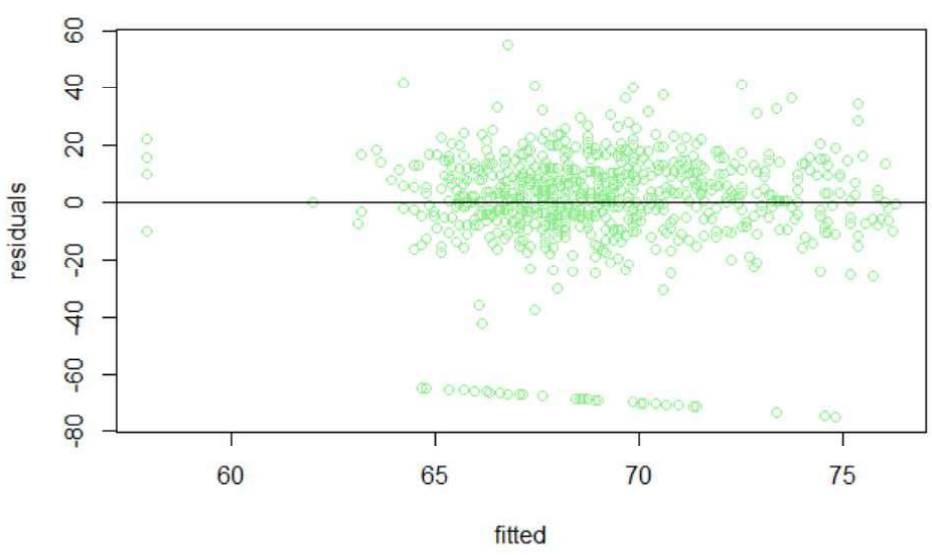
"Df" column shows the degrees of freedom for each term in the model.

"Sum Sq" column gives the sum of squared differences between the observed values and the predicted values for each term, and the "Mean Sq" column gives the mean squared difference for each term.

The "F value" is the ratio of the mean squared difference for the independent variable (glucose) to the mean squared difference for the residuals, and the "Pr(>F)" column gives the p-value associated with the F-test.

A low p-value indicates that the independent variable is a significant predictor of the dependent variable. In this case, the p-value is 2.17e-05, which is less than 0.001, and thus the independent variable is considered to be a significant predictor of the dependent variable. The "Signif. codes" give the significance levels of the p-values, with "***" ***indicating p < 0.001,*** "**" *indicating p < 0.01,* "*" indicating p < 0.05, and "." indicating p < 0.1.

```r
{r}
residuals<-residuals(model1)
fitted<-fitted(model1)
#create the residuals vs fitted
plot(fitted,residuals,type = "p",col="light green")
abline(h=0)
```



residuals of the model are calculated using the residuals() function, and the fitted values of the model are calculated using the fitted() function.

The **abline(h=0)** function is used to add a horizontal line at zero to the plot, which helps to determine if the residuals are randomly scattered around zero, which is a desirable property for a well-fitting linear regression model. In a well-fitting model, the residuals should be randomly scattered around zero, with no systematic pattern.

If there is a systematic pattern in the residuals, it suggests that the model is not capturing some important aspect of the relationship between the independent and dependent variables, and that a better model may be needed.

h. Correlation matrix

```r
# correlation matrix
library(reshape2)
cor_melt <- melt(cor(diabetes[, 1:8]))
cor_melt <- cor_melt[which(cor_melt$value > 0.5 & cor_melt$value != 1), ]
cor_melt <- cor_melt[1:2, ]
cor_melt
```

Description: df [2 × 3]

	Var1 <fctr>	Var2 <fctr>	value <dbl>
8	age	pregnancies	0.5443412
57	pregnancies	age	0.5443412

2 rows

cor() function is used to calculate the correlation matrix of the first 8 columns of the "diabetes" data set.

The **melt()** function from the reshape2 library is then used to reshape the correlation matrix into a format that is easier to visualize.

The resulting data frame, **cor_melt**, is then filtered to include only correlations with a value greater than 0.5 and not equal to 1, as these are considered to be strong correlations.

Finally, the first 2 columns of the filtered data frame are selected to display the correlations of interest. The resulting data frame shows the variables that are strongly correlated with each other.

i. Fitting a logistic regression model.

```r
model2<-glm(outcome~pregnancies+glucose+bloodpressure+skinthickness+insulin+bmi+dpf,family=binomial
,data = diabetes)
summary(model2)
```

```
Call:
glm(formula = outcome ~ pregnancies + glucose + bloodpressure +
    skinthickness + insulin + bmi + dpf, family = binomial, data = diabetes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5732  -0.7359  -0.4230   0.7425   2.9851

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.1736736  0.6963139 -11.738  < 2e-16 ***
pregnancies    0.1484580  0.0281044   5.282 1.28e-07 ***
glucose        0.0363516  0.0036544   9.947  < 2e-16 ***
bloodpressure -0.0118207  0.0051409  -2.299  0.02148 *
skinthickness -0.0003750  0.0068389  -0.055  0.95628
insulin       -0.0012783  0.0008963  -1.426  0.15383
bmi            0.0880609  0.0150195   5.863 4.54e-09 ***
dpf            0.9675342  0.2984012   3.242  0.00119 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 725.97  on 760  degrees of freedom
AIC: 741.97

Number of Fisher Scoring iterations: 5
```

The dependent variable is "outcome" and the independent variables are "pregnancies", "glucose", "bloodpressure", "skinthickness", "insulin", "bmi", and "dpf".

The **family** argument is set to "binomial", indicating that the response variable is binary and that a logistic regression model is to be fit. The data used to fit the model is specified using the **data** argument and is taken from the "diabetes" data set.

The coefficients for each independent variable provide information on the relationship between the predictor and the outcome.

the coefficient for glucose (0.0363516) suggests that for every unit increase in glucose, the log odds of having diabetes (outcome=1) increases by 0.0363516. The coefficient for bloodpressure (-0.0118207) suggests that for every unit increase in bloodpressure, the log odds of having diabetes decreases by 0.0118207.

The null deviance and residual deviance indicate how well the model fits the data.

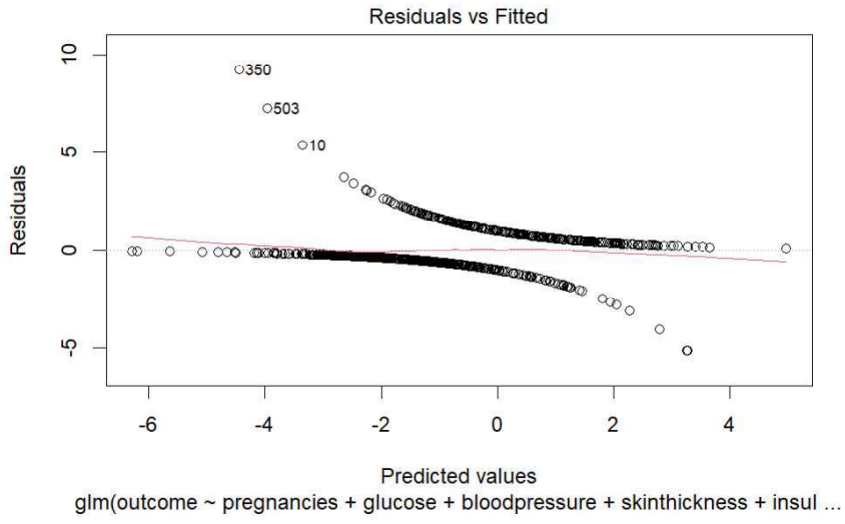
The null deviance (993.48) is the deviance when all predictors are set to zero.

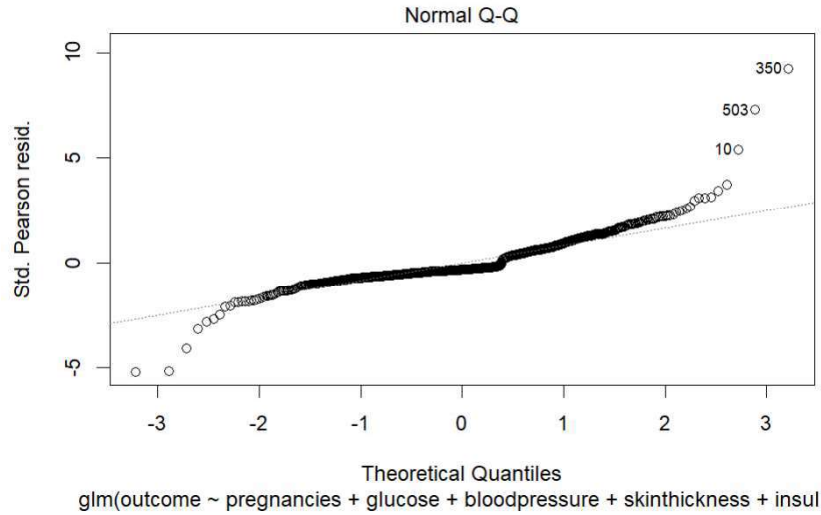
The residual deviance (725.97) is the deviance when the model is fit to the data.

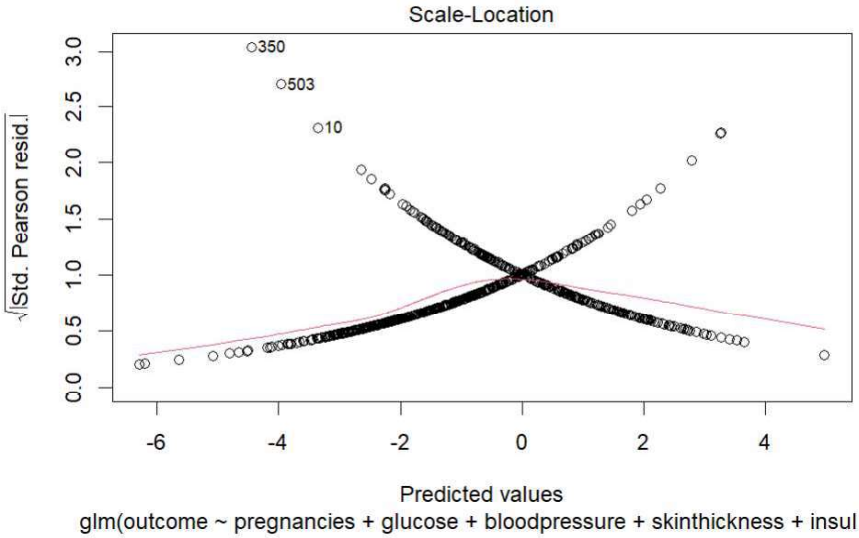
The residual deviance is smaller than the null deviance, indicating that the **model fits the data better than the null model.**

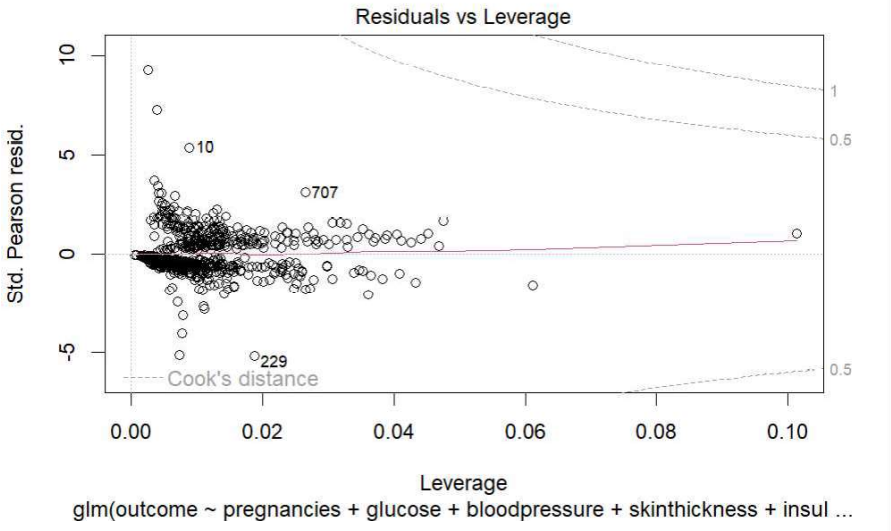
The AIC (Akaike Information Criterion) provides a measure of model fit and is used for model selection. **The lower the AIC, the better the model fit. In this case, the AIC for the model is 741.97.**

```{r}
plot(model2)
```




Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm(outcome ~ pregnancies + glucose + bloodpressure + skinthickness + insul ...
350
503
10

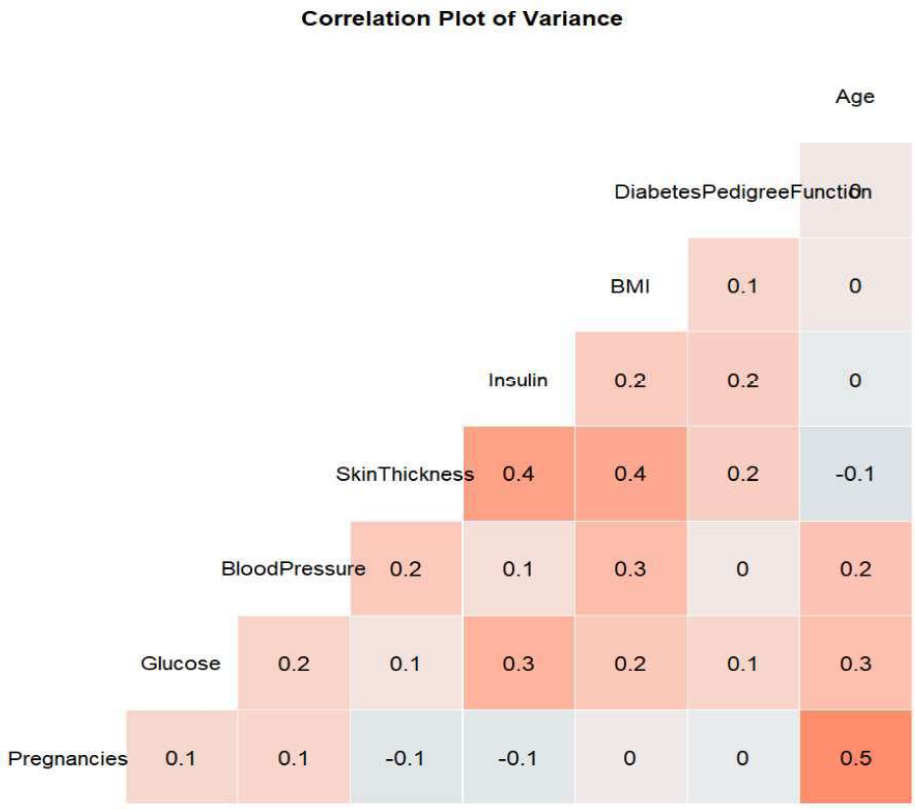

Scale-Location
√|Std. Pearson resid.|
Predicted values
glm(outcome ~ pregnancies + glucose + bloodpressure + skinthickness + insul ...
350
503
10


Residuals vs Leverage
Std. Pearson resid.
Leverage
glm(outcome ~ pregnancies + glucose + bloodpressure + skinthickness + insul ...
Cook's distance
10
707
229

j. **Correlation between each variable**

Scatter matrix of all columns

```r
{r}

#install.packages("GGally")
library(GGally)
library(ggplot2)

ggcorr(diabetes[,-9],name="corr",label = TRUE)+theme(legend.position = "none")+labs(title="Correlation
Plot of Variance")+theme(plot.title = element_text(face = 'bold',color = 'black',hjust = 0.5,size = 12
))
```



Pregnancy, Age, Insulin, skin thickness are having higher correlation.

k. Binary logistic regression model to predict Diabetes.

The response variable is "Outcome" and the predictor variables are all the remaining variables in the "diabetes" data set.

In below code, two columns, "BloodPressure" and "SkinThickness", are removed from the "diabetes" data set.

Then, the data set is split into two parts:

- a training set "train" containing the first 540 rows of the data.
- a test set "test" containing the remaining rows.
- Finally, the logistic regression model is fit using the "glm" function, with the response variable as "Outcome" and all other variables as predictor variables.

  The link function is specified as "logit", which is appropriate for binary logistic regression.

```{r}
diabetes$BloodPressure <- NULL
diabetes$SkinThickness <- NULL
train <- diabetes[1:540,]
test <- diabetes[541:768,]
model <-glm(Outcome ~.,family=binomial(link='logit'),data=train)
summary(model)
```

```
Call:
glm(formula = Outcome ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4366  -0.7741  -0.4312   0.8021   2.7310

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.3461752  0.8157916 -10.231  < 2e-16 ***
Pregnancies              0.1246856  0.0373214   3.341 0.000835 ***
Glucose                  0.0315778  0.0042497   7.431 1.08e-13 ***
Insulin                 -0.0013400  0.0009441  -1.419 0.155781
BMI                      0.0881521  0.0164090   5.372 7.78e-08 ***
DiabetesPedigreeFunction 0.9642132  0.3430094   2.811 0.004938 **
Age                      0.0018904  0.0107225   0.176 0.860053
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 700.47  on 539  degrees of freedom
Residual deviance: 526.56  on 533  degrees of freedom
AIC: 540.56

Number of Fisher Scoring iterations: 5
```

```
{r}
anova(model, test="Chisq")
```

```
Call:
glm(formula = Outcome ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4366  -0.7741  -0.4312   0.8021   2.7310

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.3461752  0.8157916 -10.231  < 2e-16 ***
Pregnancies               0.1246856  0.0373214   3.341 0.000835 ***
Glucose                   0.0315778  0.0042497   7.431 1.08e-13 ***
Insulin                  -0.0013400  0.0009441  -1.419 0.155781
BMI                       0.0881521  0.0164090   5.372 7.78e-08 ***
DiabetesPedigreeFunction  0.9642132  0.3430094   2.811 0.004938 **
Age                       0.0018904  0.0107225   0.176 0.860053
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 700.47  on 539  degrees of freedom
Residual deviance: 526.56  on 533  degrees of freedom
AIC: 540.56

Number of Fisher Scoring iterations: 5
```

Accuracy of model Binary regression to predict diabetes:

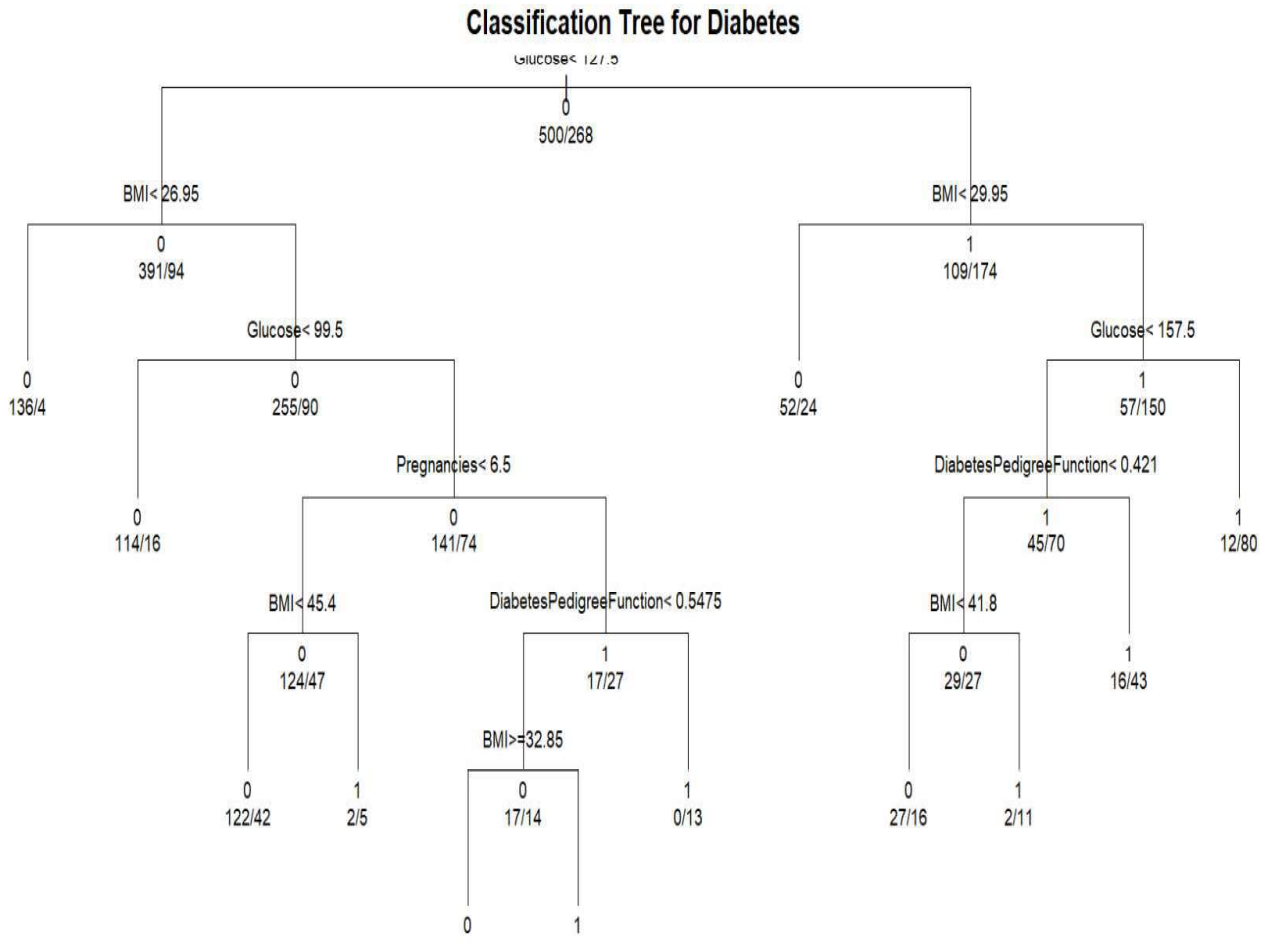
```
{r}
fitted.results <- predict(model,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$Outcome)
print(paste('Accuracy',1-misClasificError))
```

```
[1] "Accuracy 0.789473684210526"
```

1. **Decision trees create using rpart() to predict Diabetes.**

```r
{r}
library(rpart)
model2 <- rpart(Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction, data=diabetes,method
="class")
plot(model2, uniform=TRUE, main="Classification Tree for Diabetes")
text(model2, use.n=TRUE, all=TRUE, cex=.8)
```

creates a decision tree using the **rpart** library and fits the model using the variables **Pregnancies**, **Glucose**, **BMI**, and **DiabetesPedigreeFunction** as predictors for the target variable **Outcome** (0 or 1 indicating the presence or absence of diabetes).



Accuracy of model Decision trees to predict Diabetes:

```
{r}
treePred <- predict(model2,test, type = 'class')
table(treePred,test$Outcome)
mean(treePred==test$Outcome)
```

```
 treePred   0   1
        0 139  31
        1  11  47
 [1] 0.8157895
```

By the two models Binary Logistic Regression and Decision trees Decision trees model accuracy is more to predict the Diabetes .