MA6101 R Peoject

```r
# 10 December 2020
# Student Name: Tom Meehan
# Student ID: 18220975

setwd("C:/Users/meeha/OneDrive/College/Statistics/Project")

library(tidyverse)
library(dplyr)
library(ggplot2)
library(car)
```

Question 1

```r
# Question 1 - Load the pima Indians dataset into R

data = read.csv(file = "MA6101data.csv",header = TRUE)
```

Question 2

```r
#Question 2 - Set the random number seed in R as your UL student number

set.seed(18220975)
```

Question 3

```r
#Question 3 - Extract a random sample of 320 rows

mydata = sample_n(data,320)
```

Question 4

A)

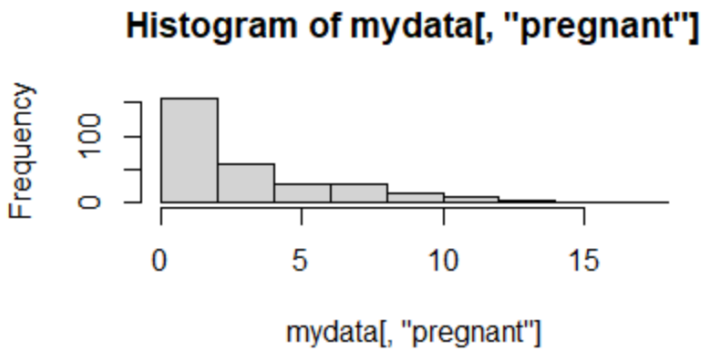
```r
#Question 4 - Explore the mydata dataset

#a) Summarise each variable and provide appropriate statistics and confidence intervals

# Summarises the entire data set
summary(mydata)
```

```
> summary(mydata)
       X            pregnant         glucose     
 Min.   :  4.0   Min.   : 0.000   Min.   : 56.0  
 1st Qu.:198.2   1st Qu.: 1.000   1st Qu.: 99.0  
 Median :375.5   Median : 2.000   Median :119.0  
 Mean   :378.9   Mean   : 3.247   Mean   :121.8  
 3rd Qu.:555.2   3rd Qu.: 5.000   3rd Qu.:142.2  
 Max.   :766.0   Max.   :17.000   Max.   :197.0  
    pressure         triceps          insulin     
 Min.   : 24.00   Min.   : 7.00   Min.   : 14.00  
 1st Qu.: 62.00   1st Qu.:20.75   1st Qu.: 76.75  
 Median : 70.00   Median :29.00   Median :125.50  
 Mean   : 70.48   Mean   :29.13   Mean   :156.97  
 3rd Qu.: 78.50   3rd Qu.:37.00   3rd Qu.:188.50  
 Max.   :110.00   Max.   :63.00   Max.   :846.00  
      mass             age          diabetes        
 Min.   :18.20   Min.   :21.00   Length:320         
 1st Qu.:28.07   1st Qu.:23.00   Class :character   
 Median :33.25   Median :27.00   Mode  :character   
 Mean   :33.12   Mean   :30.66                      
 3rd Qu.:36.95   3rd Qu.:36.00                      
```

```
      Risk       
 Min.   : 85.32  
 1st Qu.:246.52  
 Median :297.13  
 Mean   :311.06  
 3rd Qu.:365.36  
 Max.   :878.14  
```

```r
# Summarises Pregnancies
mean(mydata$pregnant)
median(mydata$pregnant)
var(mydata$pregnant)
quantile(mydata$pregnant,c(0.25,0.75))
hist(mydata[ ,"pregnant"])
```

```
> # Summarises Pregnancies
> mean(mydata$pregnant)
[1] 3.246875
> median(mydata$pregnant)
[1] 2
> var(mydata$pregnant)
[1] 9.929457
> quantile(mydata$pregnant,c(0.25,0.75))
25% 75% 
  1   5 
> hist(mydata[ ,"pregnant"])
```

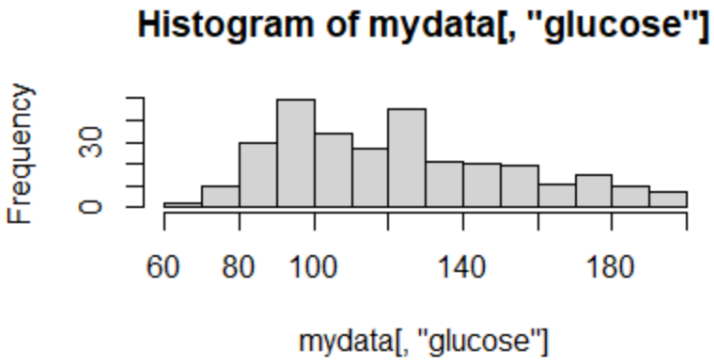



```r
#Pregnancy 95% confidence intervals

s1 <- sqrt(var(mydata[,2]))
x_bar1 <- mean(mydata[,2])
x_bar1+c(-1.96,1.96)*s1/sqrt(320)
```

```
> s1 <- sqrt(var(mydata[,2]))
> x_bar1 <- mean(mydata[,2])
> x_bar1+c(-1.96,1.96)*s1/sqrt(320)
[1] 2.901617 3.592133
```

```
> mean(mydata$glucose)
[1] 121.7656
> median(mydata$glucose)
[1] 119
> var(mydata$glucose)
[1] 939.7474
> quantile(mydata$glucose,c(0.25,0.75))
   25%    75%
 99.00 142.25
> hist(mydata[ ,"glucose"])
```

```r
#Summarises glucose levels
mean(mydata$glucose)
median(mydata$glucose)
var(mydata$glucose)
quantile(mydata$glucose,c(0.25,0.75))
hist(mydata[ ,"glucose"])
```

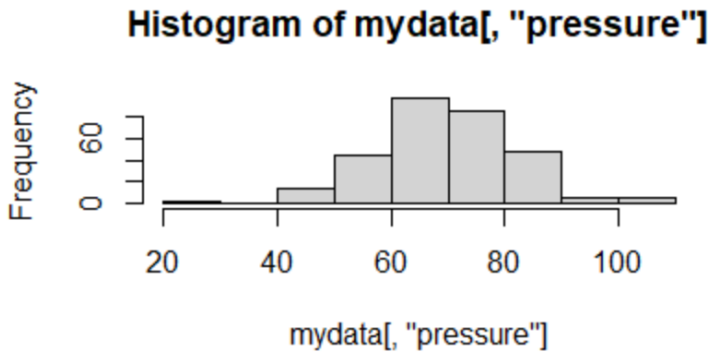



```r
#Glucose 95% confidence intervals

s2 <- sqrt(var(mydata[,3]))
x_bar2 <- mean(mydata[,3])
x_bar2+c(-1.96,1.96)*s2/sqrt(320)
```

```
> s2 <- sqrt(var(mydata[,3]))
> x_bar2 <- mean(mydata[,3])
> x_bar2+c(-1.96,1.96)*s2/sqrt(320)
[1] 118.4068 125.1244
```

```r
#Summarises blood pressure
mean(mydata$pressure)
median(mydata$pressure)
var(mydata$pressure)
quantile(mydata$pressure,c(0.25,0.75))
hist(mydata[ ,"pressure"])
```

```
> #Summarises blood pressure
> mean(mydata$pressure)
[1] 70.48125
> median(mydata$pressure)
[1] 70
> var(mydata$pressure)
[1] 162.5326
> quantile(mydata$pressure,c(0.25,0.75))
 25%  75%
62.0 78.5
> hist(mydata[ ,"pressure"])
```

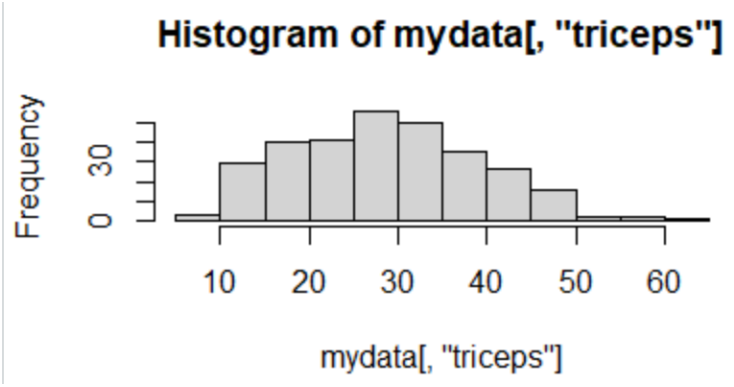



```r
#Blood Pressure 95% confidence intervals

s3 <- sqrt(var(mydata[,4]))
x_bar3 <- mean(mydata[,4])
x_bar3+c(-1.96,1.96)*s3/sqrt(320)
```

```
> hist(mydata[ ,"pressure"])
> s3 <- sqrt(var(mydata[,4]))
> x_bar3 <- mean(mydata[,4])
> x_bar3+c(-1.96,1.96)*s3/sqrt(320)
[1] 69.0844 71.8781
```

```r
#Summarises triceps
mean(mydata$triceps)
median(mydata$triceps)
var(mydata$triceps)
quantile(mydata$triceps,c(0.25,0.75))
hist(mydata[ ,"triceps"])
```

```
> #Summarises triceps
> mean(mydata$triceps)
[1] 29.13125
> median(mydata$triceps)
[1] 29
> var(mydata$triceps)
[1] 112.9608
> quantile(mydata$triceps,c(0.25,0.75))
  25%   75%
20.75 37.00
> hist(mydata[ ,"triceps"])
```

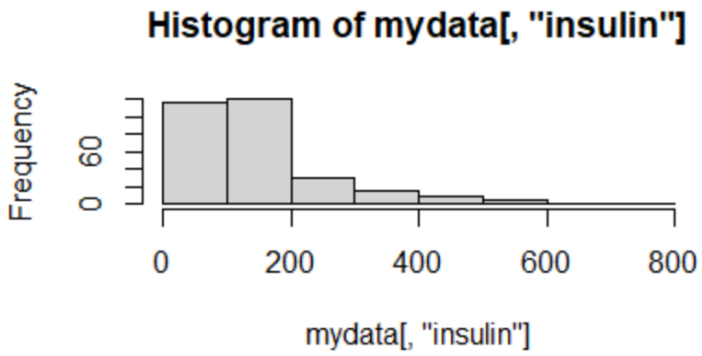



```r
#Triceps 95% confidence intervals

s4 <- sqrt(var(mydata[,5]))
x_bar4 <- mean(mydata[,5])
x_bar4+c(-1.96,1.96)*s5/sqrt(320)
```

```
> s4 <- sqrt(var(mydata[,5]))
> x_bar4 <- mean(mydata[,5])
> x_bar4+c(-1.96,1.96)*s5/sqrt(320)
[1] 15.30602 42.95648
```

```r
#Summarises Insulin level
mean(mydata$insulin)
median(mydata$insulin)
var(mydata$insulin)
quantile(mydata$insulin,c(0.25,0.75))
hist(mydata[ ,"insulin"])
```

```
> #Summarises Insulin level
> mean(mydata$insulin)
[1] 156.9719
> median(mydata$insulin)
[1] 125.5
> var(mydata$insulin)
[1] 15052.07
> quantile(mydata$insulin,c(0.25,0.75))
   25%    75%
 76.75 188.50
> hist(mydata[ ,"insulin"])
```

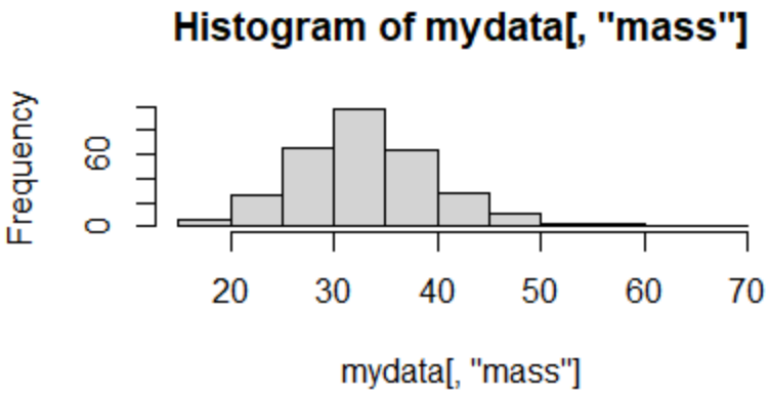



```
90  #Insulin 95% confidence intervals
91
92  s5 <- sqrt(var(mydata[,6]))
93  x_bar5 <- mean(mydata[,6])
94  x_bar5+c(-1.96,1.96)*s5/sqrt(320)
```

```
> s5 <- sqrt(var(mydata[,6]))
> x_bar5 <- mean(mydata[,6])
> x_bar5+c(-1.96,1.96)*s5/sqrt(320)
[1] 143.5294 170.4143
```

```
 96  #Summarises BMI
 97  mean(mydata$mass)
 98  median(mydata$mass)
 99  var(mydata$mass)
100  quantile(mydata$mass,c(0.25,0.75))
101  hist(mydata[ ,"mass"])
```

```
> #Summarises BMI
> mean(mydata$mass)
[1] 33.11844
> median(mydata$mass)
[1] 33.25
> var(mydata$mass)
[1] 51.28032
> quantile(mydata$mass,c(0.25,0.75))
   25%    75%
28.075 36.950
> hist(mydata[ ,"mass"])
```

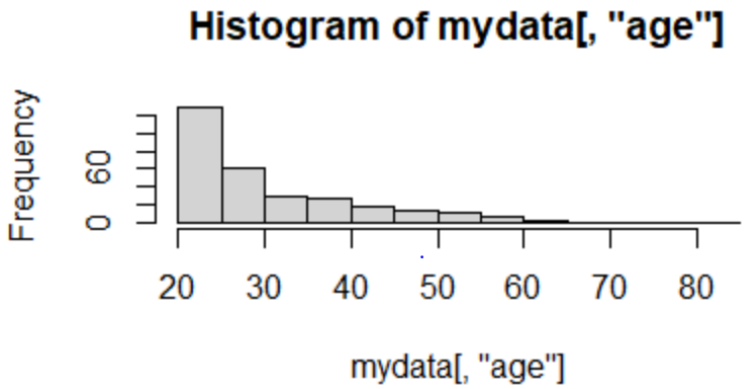



```
#Mass 95% confidence intervals

s6 <- sqrt(var(mydata[,7]))
x_bar6 <- mean(mydata[,7])
x_bar6+c(-1.96,1.96)*s6/sqrt(320)
```

```
> s6 <- sqrt(var(mydata[,7]))
> x_bar6 <- mean(mydata[,7])
> x_bar6+c(-1.96,1.96)*s6/sqrt(320)
[1] 32.33382 33.90305
```

```
#Summarises age
mean(mydata$age)
median(mydata$age)
var(mydata$age)
quantile(mydata$age,c(0.25,0.75))
hist(mydata[ ,"age"])
```

```
> #Summarises age
> mean(mydata$age)
[1] 30.65625
> median(mydata$age)
[1] 27
> var(mydata$age)
[1] 102.1072
> quantile(mydata$age,c(0.25,0.75))
25% 75%
 23  36
> hist(mydata[ ,"age"])
```




```
#Age function 95% confidence intervals

s7 <- sqrt(var(mydata[,8]))
x_bar7 <- mean(mydata[,8])
x_bar7+c(-1.96,1.96)*s7/sqrt(320)
```

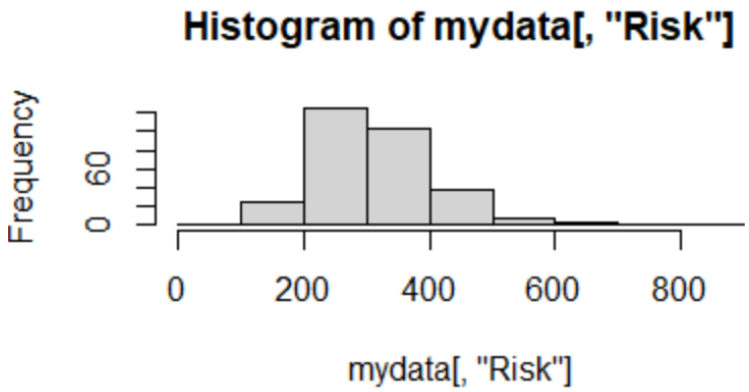
```
> s7 <- sqrt(var(mydata[,8]))
> x_bar7 <- mean(mydata[,8])
> x_bar7+c(-1.96,1.96)*s7/sqrt(320)
[1] 29.54909 31.76341
```

```
#Summarises diabetes

# Create a function for finding the mode.
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

summary(mydata$diabetes)
getmode(mydata$diabetes)
```

```
> #Summarises diabetes
>
> # Create a function for finding the mode.
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
>
> summary(mydata$diabetes)
   Length     Class      Mode
      320 character character
> getmode(mydata$diabetes)
[1] "neg"
```

```
#Summarises risk
mean(mydata$Risk)
median(mydata$Risk)
var(mydata$Risk)
quantile(mydata$Risk,c(0.25,0.75))
hist(mydata[ ,"Risk"])
```

```
> #Summarises risk
> mean(mydata$Risk)
[1] 311.0589
> median(mydata$Risk)
[1] 297.1339
> var(mydata$Risk)
[1] 9373.772
> quantile(mydata$Risk,c(0.25,0.75))
     25%      75%
246.5213 365.3615
> hist(mydata[ ,"Risk"])
```

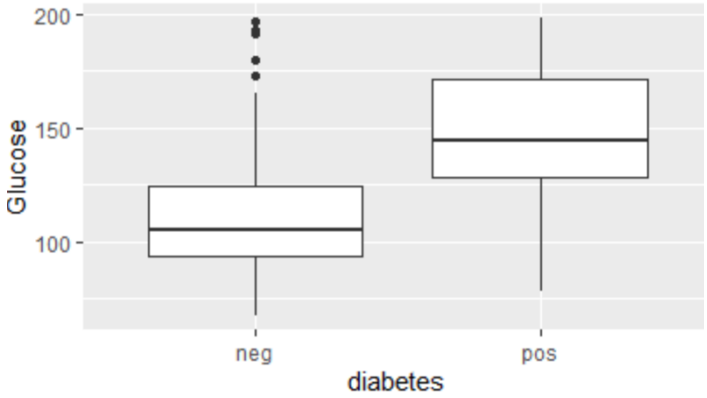



```
#Risk 95% confidence intervals

s9 <- sqrt(var(mydata[,10]))
x_bar9 <- mean(mydata[,10])
x_bar9+c(-1.96,1.96)*s9/sqrt(320)
```

```
> s9 <- sqrt(var(mydata[,10]))
> x_bar9 <- mean(mydata[,10])
> x_bar9+c(-1.96,1.96)*s9/sqrt(320)
[1] 300.4508 321.6671
```

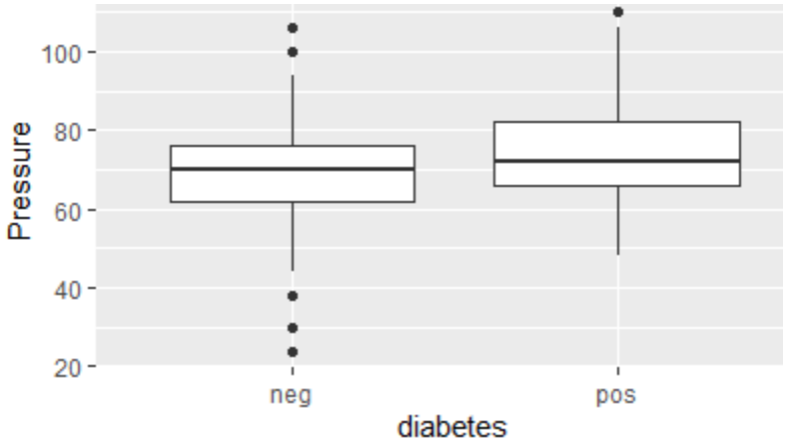
```
147 #b)
148 #Compare the distribution for people with and without diabetes for the following variables:
149 #glucose
150 #pressure
151 #insulin
```

```
153 #Comparing Glucose and diabetes
154
155 ggplot(aes(y=glucose,x = diabetes),data=mydata) + geom_boxplot()+ylab("Glucose")
```



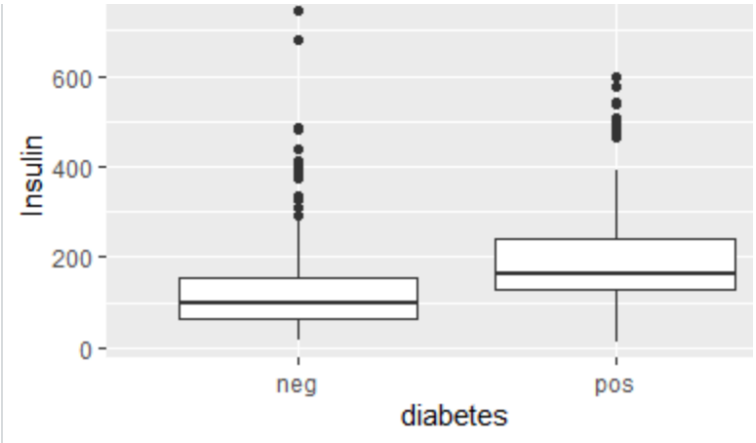
- For people with negative diabetes the median glucose level is approximately 110 with an upper quartile of 150 and a lower quartile of 90.
- For people with positive diabetes the median glucose level is approximately 145 with an upper quartile of 172 and a lower quartile of 127.
- From this it can be observed that on average the people with higher glucose levels tend to have more of a chance of testing positive for diabetes.
- The graphs also show that glucose level and negative diabetes is normally distributed while glucose level and positive diabetes is positively skewed.

```
157 #Comparing pressure and diabetes
158
159 ggplot(aes(y=pressure,x = diabetes),data=mydata) + geom_boxplot()+ylab("Pressure")
```



- For people with negative diabetes the median pressure is approximately 70 with an upper quartile of 77 and a lower quartile of 62.
- For people with positive diabetes the median glucose level is approximately 72 with an upper quartile of 82 and a lower quartile of 76.
- From this it can be observed that on average the people with higher pressure tend to have more of a chance of testing positive for diabetes although there is not much variation in the data for this to be conclusive.
- The graphs also show that pressure and negative diabetes is negatively skewed while glucose level and positive diabetes is positively skewed.

```
161 #Comparing Insulin and diabetes
162
163 ggplot(aes(y=insulin,x = diabetes),data=mydata) + geom_boxplot()+ylab("Insulin")
```



- For people with negative diabetes the median insulin level is approximately 100 with an upper quartile of 165 and a lower quartile of 90.
- For people with positive diabetes the median glucose level is approximately 170 with an upper quartile of 220 and a lower quartile of 167.

- From this it can be observed that on average the people with insulin levels tend to have more of a chance of testing positive for diabetes.
- The graphs also show that insulin level and negative diabetes is normally distributed while insulin and positive diabetes is positively skewed.

C)

```
#C) Test the significance of the mean glucose for those with and without diabetes

# Creates data for all negative and positive results.
mydata_neg<-mydata[mydata$diabetes == "neg",]
mydata_pos<-mydata[mydata$diabetes == "pos",]

# H0 : m1 - m2 = 0, Ha : m1 - m2 != 0
t.test(mydata_neg$glucose, mydata_pos$glucose,
       alternative = c("two.sided"),
       var.equal=TRUE,
       mu = 0)
```

```
        Two Sample t-test

data:  mydata_neg$glucose and mydata_pos$glucose
t = -10.247, df = 318, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal
 to 0
95 percent confidence interval:
 -39.00720 -26.44072
sample estimates:
mean of x mean of y
 111.3349  144.0588
```

As 0 is not within the 95% confidence interval we reject the null hypothesis at the 5% significance level.

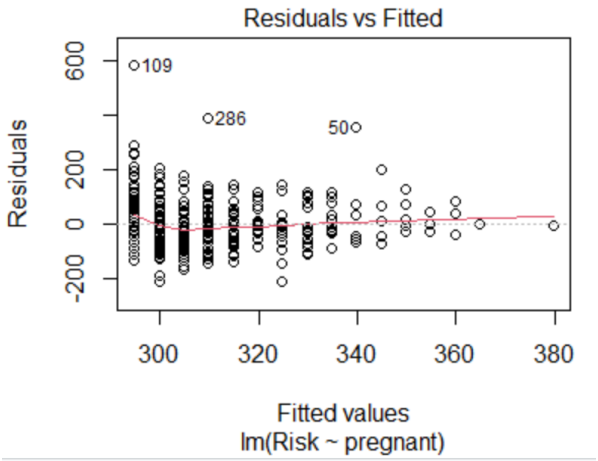
From this we can conclude that there is evidence to suggest that there is a statically significant difference in the mean of those who test positive for diabetes and those who test negative.

D)

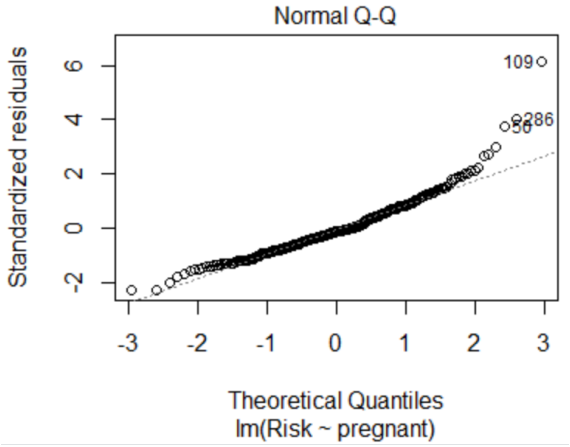
```
#D) Use scatter plots and correlation to predict risk score.

# Risk V Pregnancy
model1 <- lm(Risk ~pregnant,data = mydata)
summary(model1)

plot(model1)
hist(model1$residuals)


plot(mydata$Risk,mydata$pregnant)
abline(model1,col="blue")
```

```
> model1 <- lm(Risk ~pregnant,data = mydata)
> summary(model1)

Call:
lm(formula = Risk ~ pregnant, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-216.00  -63.49  -10.40   52.40  583.30

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  294.843      7.686  38.360  < 2e-16 ***
pregnant       4.994      1.700   2.938  0.00355 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 95.68 on 318 degrees of freedom
Multiple R-squared:  0.02642,   Adjusted R-squared:  0.02336
F-statistic: 8.631 on 1 and 318 DF,  p-value: 0.003547
```

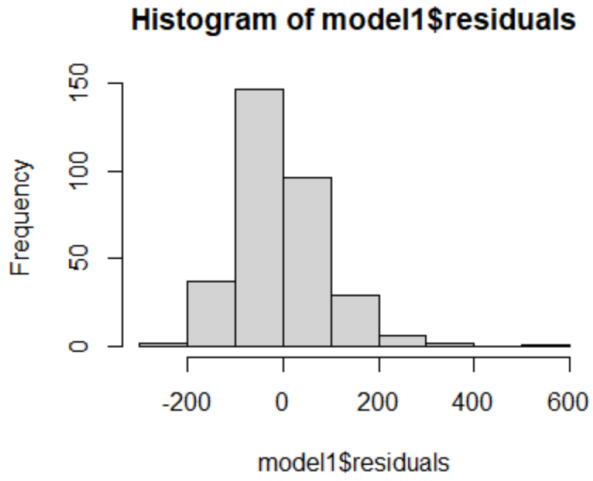
The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.02336. This means 2.336% of the variability in risk is explained by a linear relationship with pregnancies.



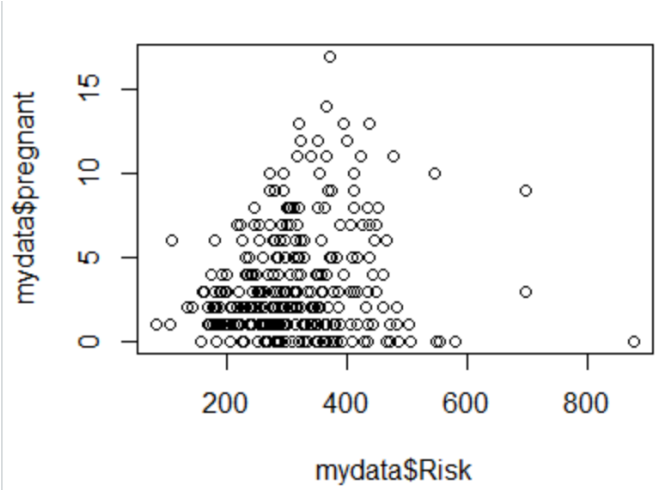
The variabilities of residuals appear to increase for increasing risk score.



The qqplot show that the residuals seem to follow a normal distribution.



The histogram indicates that the residuals do follow a normal distribution.



As the data isn't very linear the variability is hard to predict.

```
190  #Predictions
191
192  predict(model1, newdata = list(pregnant = 3), interval = "prediction",level=0.95)
```

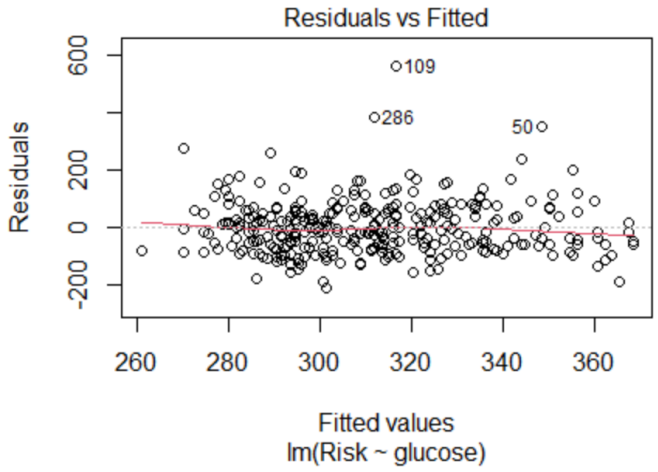
```
> predict(model1, newdata = list(pregnant = 3), interval = "prediction",level=0.95)
       fit      lwr      upr
1 309.8259 121.2832 498.3687
```

For a predicted pregnancy of 3 the risk level is 309.83 with 95% prediction interval of 121.28 and 498.37.

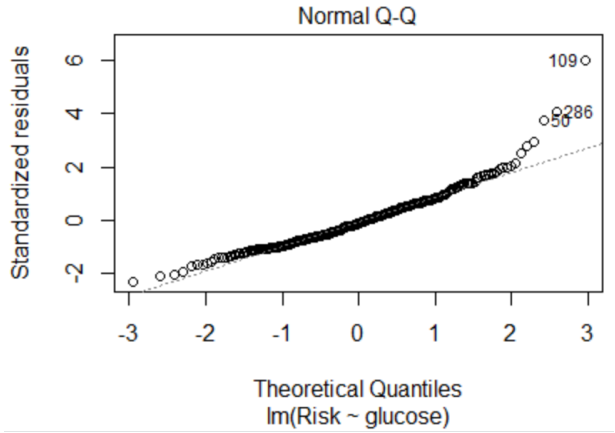
```
194  # Risk V Glucose
195  model2 <- lm(Risk ~glucose,data = mydata)
196  summary(model2)
197
198  plot(model2)
199  hist(model2$residuals)
200
201
202  plot(mydata$Risk,mydata$glucose)
203  abline(model2,col="blue")
```

```
> model2 <- lm(Risk ~glucose,data = mydata)
> summary(model2)

Call:
lm(formula = Risk ~ glucose, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-216.01  -63.81  -10.86   52.56  561.57

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 218.3277    21.5804  10.117  < 2e-16 ***
glucose       0.7616     0.1719   4.431 1.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.11 on 318 degrees of freedom
Multiple R-squared:  0.05814,   Adjusted R-squared:  0.05518
F-statistic: 19.63 on 1 and 318 DF,  p-value: 1.294e-05
```

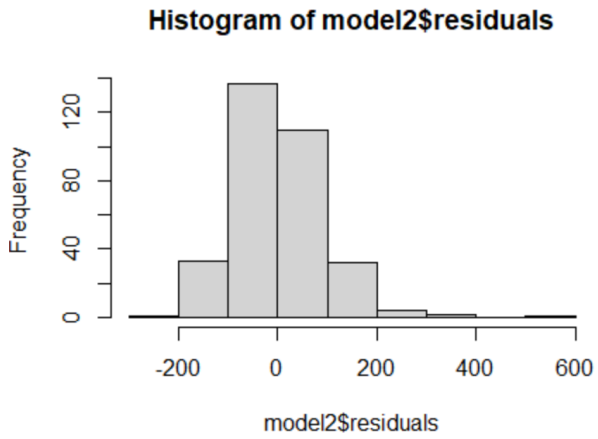
The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.05518. This means 5.5518% of the variability in risk is explained by a linear relationship with glucose.



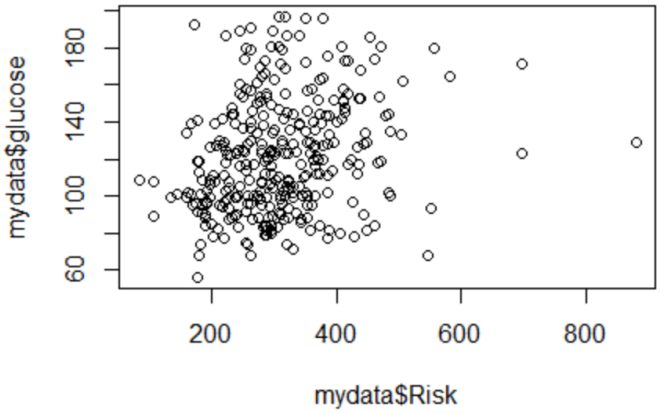
The residuals appear to remain relatively constant as the level of glucose increases.



The qqplot show that the residuals seem to follow a normal distribution.



The histogram indicates that the residuals do follow a normal distribution.



As the data isn't very linear the variability is hard to predict.

```
#Predictions

predict(model2, newdata = list(glucose = 122), interval = "prediction",level=0.95)
```

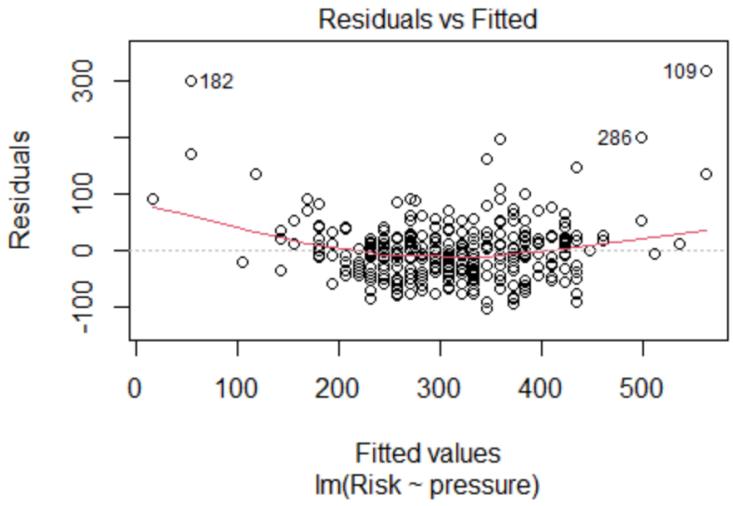
```
> predict(model2, newdata = list(glucose = 122), interval = "prediction",level=0.95)
       fit      lwr      upr
1 311.2374 125.7933 496.6816
```

For a predicted glucose level of 122 the risk level is 311.23 with 95% prediction interval of 125.78 and 496.68.

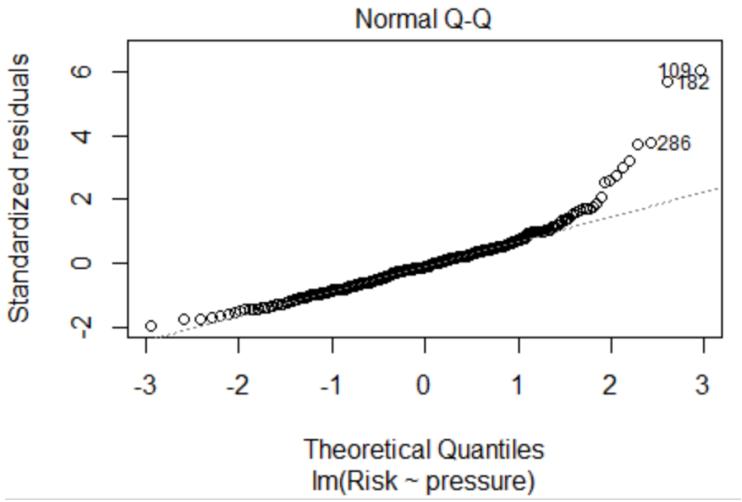
```
# Risk V Pressure
model3 <- lm(Risk ~pressure,data = mydata)
summary(model3)

plot(model3)
hist(model3$residuals)


plot(mydata$Risk,mydata$pressure)
abline(model3,col="blue")
```

```
> model3 <- lm(Risk ~pressure,data = mydata)
> summary(model3)

Call:
lm(formula = Risk ~ pressure, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-104.371  -33.415   -6.428   22.516  316.163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -136.4500    16.7338  -8.154 8.21e-15 ***
pressure       6.3493     0.2336  27.175  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.2 on 318 degrees of freedom
Multiple R-squared:  0.699,     Adjusted R-squared:  0.6981
F-statistic: 738.5 on 1 and 318 DF,  p-value: < 2.2e-16
```

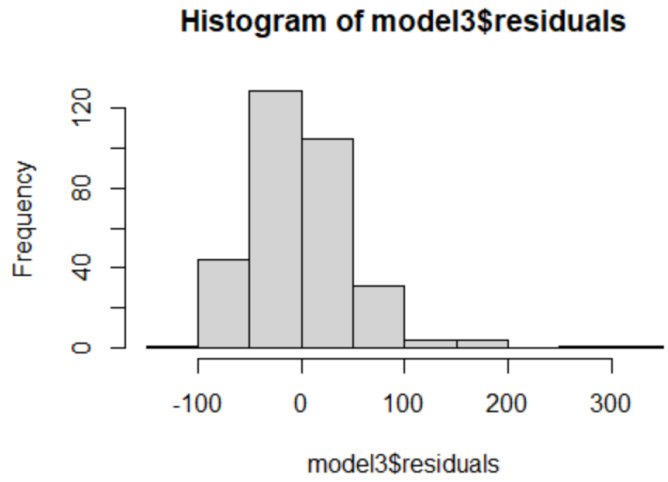
The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.6981. This means 69.81% of the variability in risk is explained by a linear relationship with pressure.



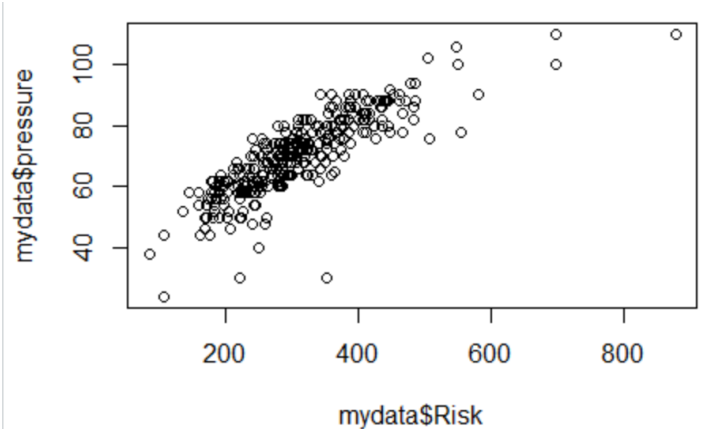
The residuals appear to decrease as pressure increases.



The qqplot show that the residuals seem to follow a normal distribution.



The histogram indicates that the residuals do follow a normal distribution.



The graph shows that the pressure and the risk level follow a linear relationship.

```
220  #Predictions
221
222  predict(model3, newdata = list(pressure = 71), interval = "prediction",level=0.95)
```

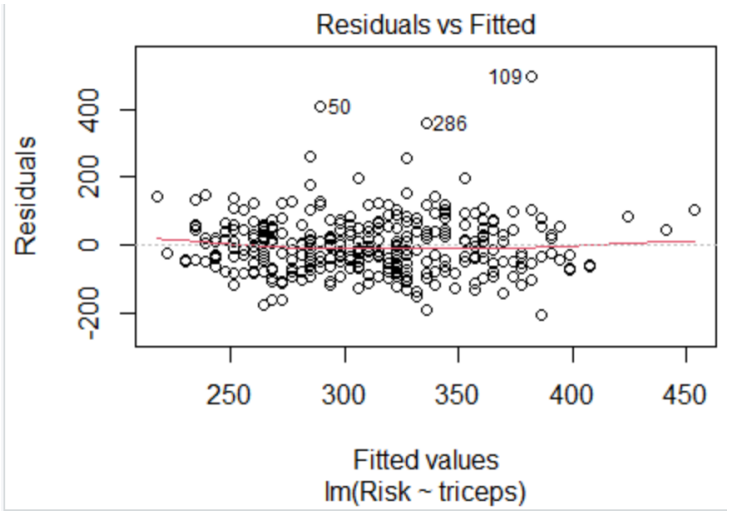
```
> predict(model3, newdata = list(pressure = 71), interval = "prediction",level=0.95)
       fit      lwr      upr
1 314.3527 209.5194 419.1859
```

For a predicted pressure level of 71 the risk level is 314.35 with 95% prediction interval of 209.52 and 419.19.

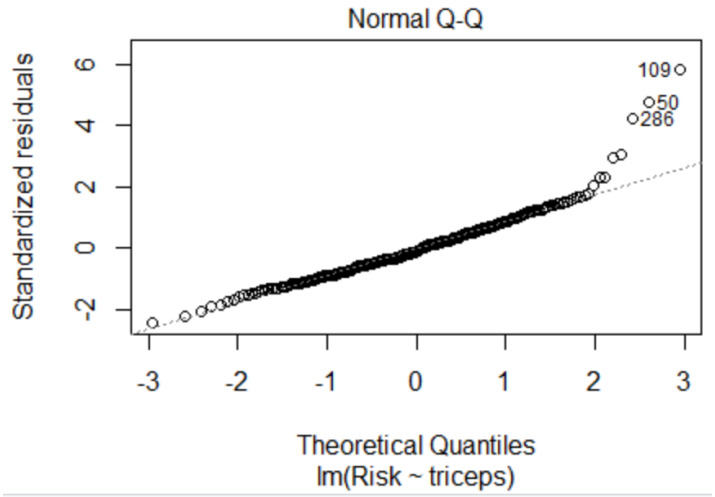
```
224  # Risk V Triceps
225  model4 <- lm(Risk ~triceps,data = mydata)
226  summary(model4)
227
228  plot(model4)
229  hist(model4$residuals)
230
231  plot(mydata$Risk,mydata$triceps)
232  abline(model1,col="blue")
```

```
> model4 <- lm(Risk ~triceps,data = mydata)
> summary(model4)

Call:
lm(formula = Risk ~ triceps, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-208.68  -53.72  -10.71   47.62  496.04

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  188.378     14.044  13.414   <2e-16 ***
triceps        4.211      0.453   9.297   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.99 on 318 degrees of freedom
Multiple R-squared:  0.2137,    Adjusted R-squared:  0.2113
F-statistic: 86.44 on 1 and 318 DF,  p-value: < 2.2e-16
```

The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.0.2113. This means 2.113% of the variability in risk is explained by a linear relationship with triceps thickness.

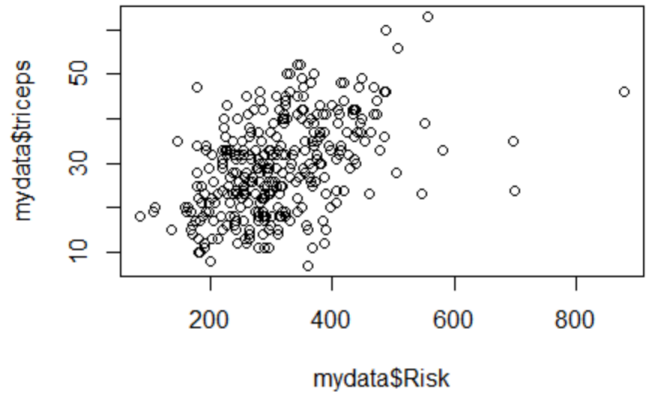


The residuals appear to remain constant as tricep thickness increases.



The qqplot show that the residuals seem to follow a normal distribution.

The histogram indicates that the residuals do follow a normal distribution.



The graph shows that risk level and tricep thickness have a small linear relationship.

```
234  #Predictions
235
236  predict(model4, newdata = list(triceps = 29), interval = "prediction",level=0.95)
```

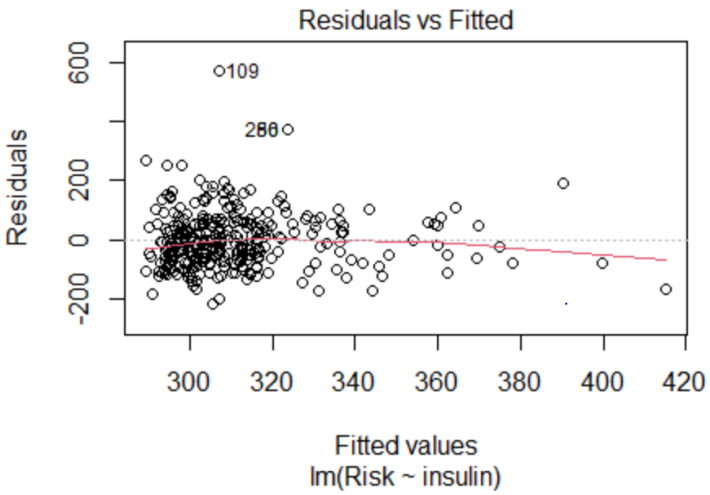
```
> predict(model4, newdata = list(triceps = 29), interval = "prediction",level=0.95)
       fit      lwr      upr
1 310.5062 141.0691 479.9433
```

For a predicted tricep thickness of 29 the risk level is 310.5 with 95% prediction interval of 141.07 and 479.94.

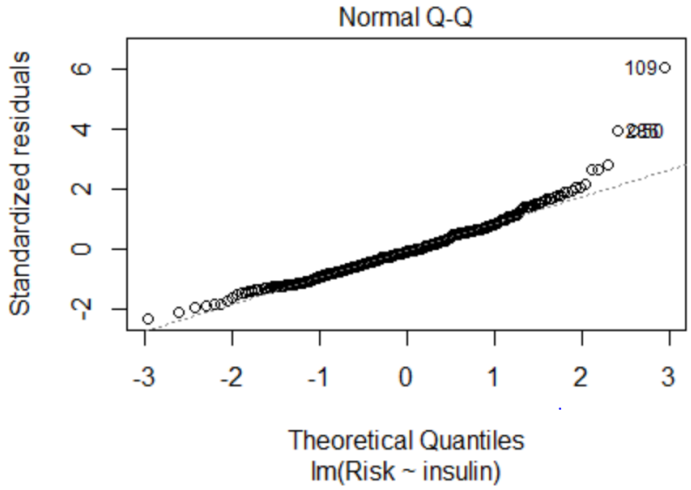
```
238  # Risk V Insulin
239  model5 <- lm(Risk ~insulin,data = mydata)
240  summary(model5)
241
242  plot(model5)
243  hist(model5$residuals)
244
245  plot(mydata$Risk,mydata$insulin)
246  abline(model1,col="blue")
```

```
> model5 <- lm(Risk ~insulin,data = mydata)
> summary(model5)

Call:
lm(formula = Risk ~ insulin, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-220.14  -62.35   -8.45   51.67  571.16

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 287.32233    8.64804  33.224  < 2e-16 ***
insulin       0.15122    0.04343   3.482 0.000568 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 95.17 on 318 degrees of freedom
Multiple R-squared:  0.03672,   Adjusted R-squared:  0.03369
F-statistic: 12.12 on 1 and 318 DF,  p-value: 0.0005682
```

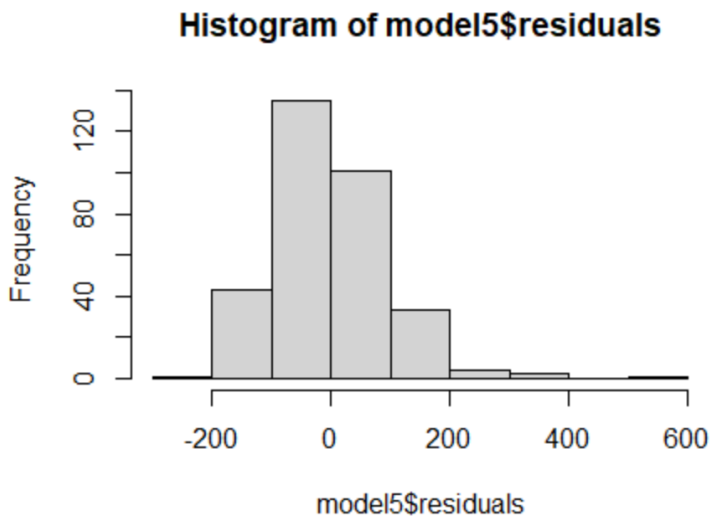
The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.03369. This means 3.3369% of the variability in risk is explained by a linear relationship with insulin.



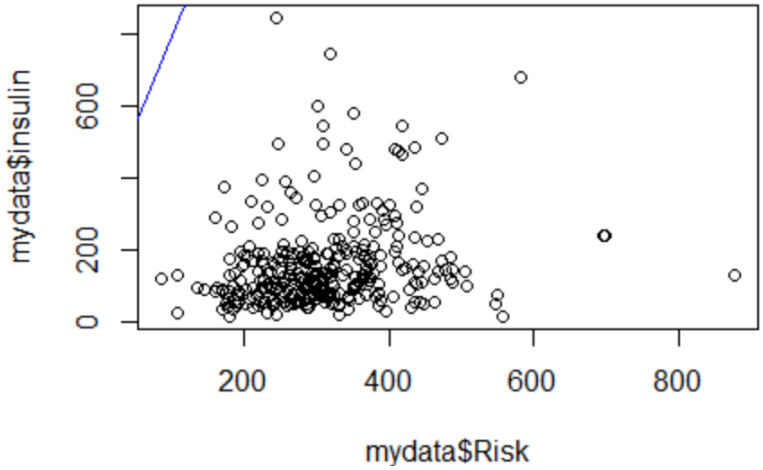
The residuals appear to decrease as insulin levels increase.



The qqplot show that the residuals seem to follow a normal distribution.



The histogram indicates that the residuals do follow a normal distribution.



The following graph shows a non-linear relationship between risk level and insulin.

```
248  #Predictions
249
250  predict(model5, newdata = list(insulin = 157), interval = "prediction",level=0.95)
```

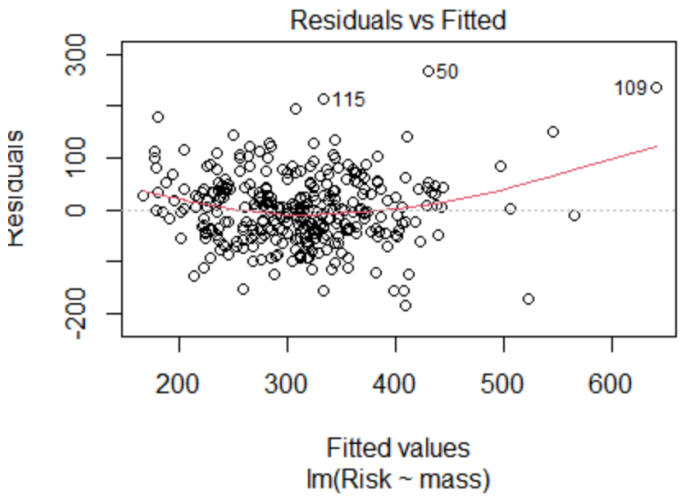
```
> predict(model5, newdata = list(insulin = 157), interval = "prediction",level=0.95)
       fit      lwr      upr
1 311.0632 123.5217 498.6047
```

For a predicted insulin level of 157 the risk level is 311.06 with 95% prediction interval of 123.52 and 498.6.

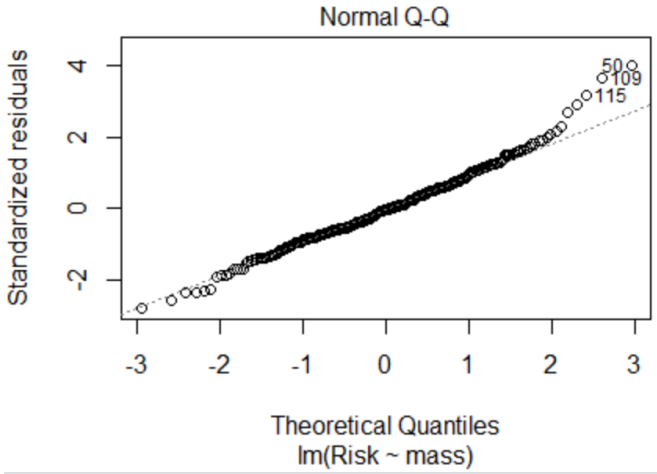
```
252  # Risk V mass
253  model6 <- lm(Risk ~mass,data = mydata)
254  summary(model6)
255
256  plot(model6)
257  hist(model6$residuals)
```

```
> model6 <- lm(Risk ~mass,data = mydata)
> summary(model6)

Call:
lm(formula = Risk ~ mass, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-186.891  -43.265   -2.252   40.074  267.007

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.5043    17.8760  -0.588    0.557
mass          9.7095     0.5276  18.403   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.48 on 318 degrees of freedom
Multiple R-squared:  0.5157,    Adjusted R-squared:  0.5142
F-statistic: 338.7 on 1 and 318 DF,  p-value: < 2.2e-16
```

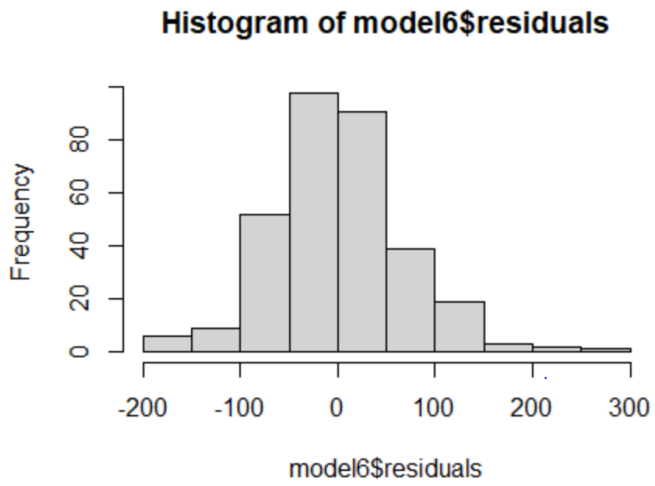
The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.0.5142. This means 51.42% of the variability in risk is explained by a linear relationship with mass.



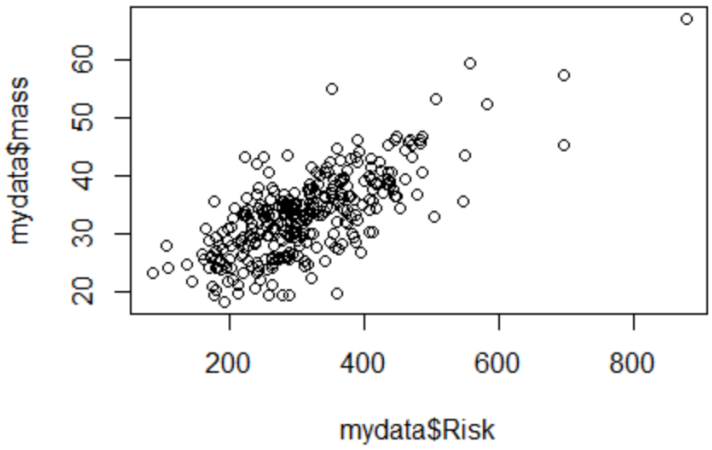
The residuals appear to increase as mass increases.



The qqplot show that the residuals seem to follow a normal distribution.



The histogram indicates that the residuals do follow a normal distribution.



The following graph shows a linear relationship between risk level and mass.

```
263  #Predictions
264
265  predict(model6, newdata = list(mass = 33), interval = "prediction",level=0.95)
```

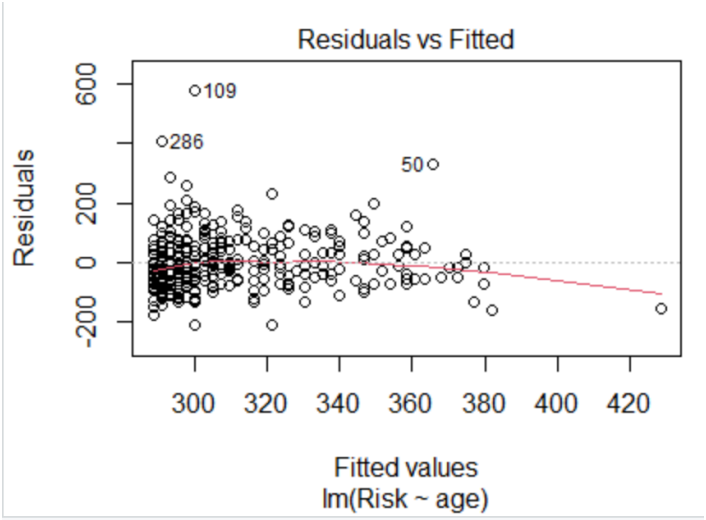
```
> predict(model6, newdata = list(mass = 33), interval = "prediction",level=0.95)
      fit      lwr      upr
1 309.909 176.9368 442.8811
```

For a predicted mass of 33 the risk level is 309.91 with 95% prediction interval of 176.94 and 442.88.

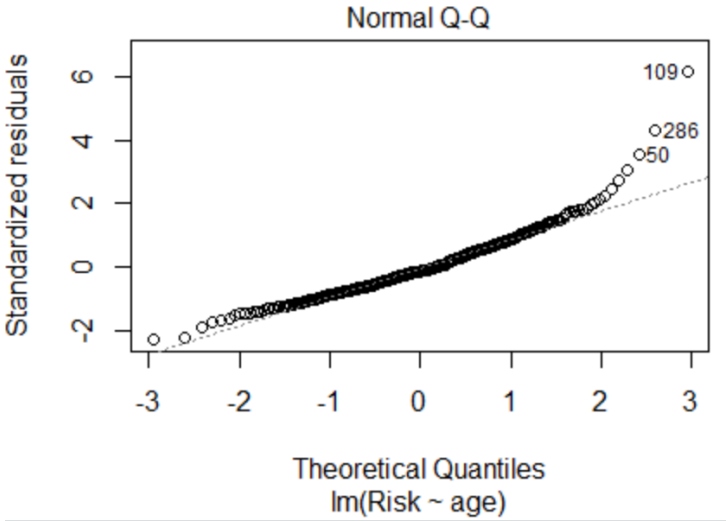
```
267  # Risk V Age
268  model7 <- lm(Risk ~age,data = mydata)
269  summary(model7)
270
271  plot(model7)
272  hist(model7$residuals)
273
274
275  plot(mydata$Risk,mydata$age)
276  abline(model7,col="blue")
```

```
> model7 <- lm(Risk ~age,data = mydata)
> summary(model7)

Call:
lm(formula = Risk ~ age, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-214.87  -62.03  -12.52   52.26  577.94

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 239.5441    16.8188  14.243  < 2e-16 ***
age           2.3328     0.5211   4.476 1.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.05 on 318 degrees of freedom
Multiple R-squared:  0.05928,   Adjusted R-squared:  0.05632
F-statistic: 20.04 on 1 and 318 DF,  p-value: 1.059e-05
```

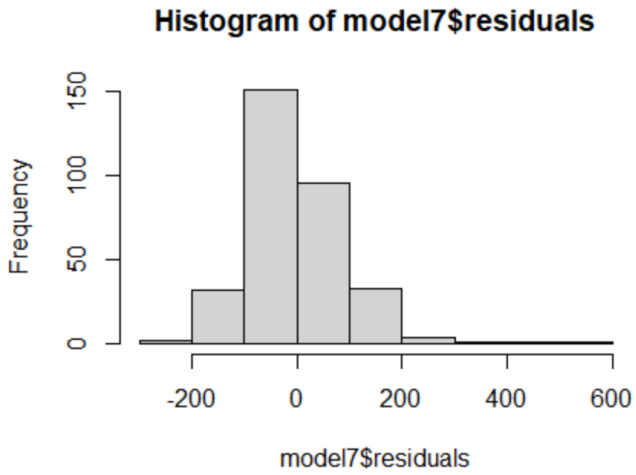
The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.05632. This means 5.632% of the variability in risk is explained by a linear relationship with age.



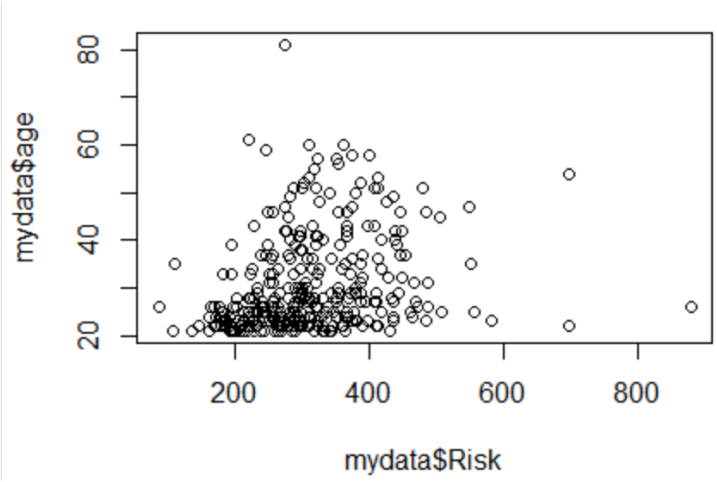
The graph shows that the residuals appear to decrease as age increases.



The qqplot show that the residuals seem to follow a normal distribution.



The histogram indicates that the residuals do follow a normal distribution.



The graph shows that there is a non-linear relationship between risk level and age.

```
278  #Predictions
279
280  predict(model7, newdata = list(age = 31), interval = "prediction",level=0.95)
```

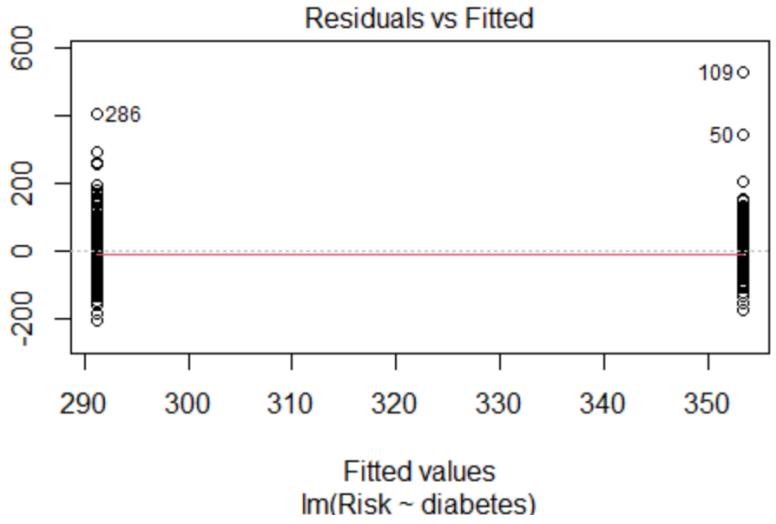
```
> predict(model7, newdata = list(age = 31), interval = "prediction",level=0.95)
       fit      lwr      upr
1 311.8608 126.5282 497.1935
```

For a predicted age of 31 the risk level is 311.86 with 95% prediction interval of 126.52 and 497.19.

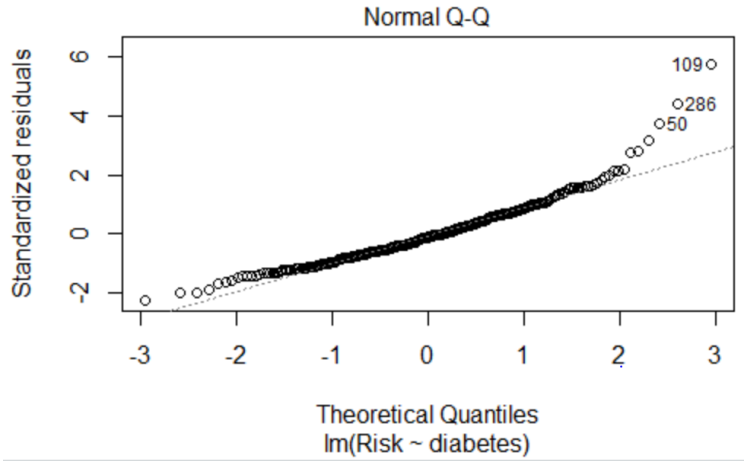
```
282  # Risk V Diabetes
283  model8 <- lm(Risk ~diabetes,data = mydata)
284  summary(model8)
285
286  plot(model8)
287  hist(model8$residuals)
```

```
> model8 <- lm(Risk ~diabetes,data = mydata)
> summary(model8)

Call:
lm(formula = Risk ~ diabetes, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-205.92  -63.01   -9.96   54.13  524.73

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  291.246      6.266  46.481  < 2e-16 ***
diabetespos   62.158     11.098   5.601 4.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.51 on 318 degrees of freedom
Multiple R-squared:  0.08978,   Adjusted R-squared:  0.08692
F-statistic: 31.37 on 1 and 318 DF,  p-value: 4.623e-08
```

The model output shows that the β_0 and β_1 values are statistically significant. The R-squared value is 0.08692. This means 8.692% of the variability in risk is explained by a linear relationship with diabetes.



The graph shows that the residuals appear to remain constant as the rate of diabetes increases.



The qqplot show that the residuals seem to follow a normal distribution.

The histogram indicates that the residuals do follow a normal distribution.

```
289  #Predictions
290
291  predict(model8, newdata = list(diabetes = "pos"), interval = "prediction",level=0.95)
292  predict(model8, newdata = list(diabetes = "neg"), interval = "prediction",level=0.95)
```

```
> predict(model8, newdata = list(diabetes = "pos"), interval = "prediction",level=0.95)
       fit     lwr      upr
1 353.4044 170.496 536.3128
```

For a predicted positive diabetes, the risk level is 353.4 with 95% prediction interval of 170.5 and 536.31.

```
> predict(model8, newdata = list(diabetes = "neg"), interval = "prediction",level=0.95)
       fit      lwr      upr
1 291.2459 108.8106 473.6813
```

For a predicted negative diabetes, the risk level is 291.25 with 95% prediction interval of 108.81 and 473.68.