# Homework 5

2022-11-18

Libraries

```
library(mlbench)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(mltools)
library(caret)
```

```
## Loading required package: lattice
```

```
library(modelr)
```

```
##
## Attaching package: 'modelr'

## The following objects are masked from 'package:mltools':
##
##     mse, rmse
```

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:mltools':
##
##     replace_na
```

# Part A

The Miniposter I selected is **Students Performance in Exams** by **Sarthak Khandelwal** Since we are approaching end of semester and everyone will be facing exams soon this topic seemed suitable to work on. Link to the dataset is mentioned. **Dataset**: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams Lets import the dataset

## Problem 1

```
data <- read.csv('/home/notorious/Documents/IDMP/StudentsPerformance.csv')
sum(is.na(data))
```

```
## [1] 0
```

There are no NA in the dataset lets look at the head to see if the data is in tidy format

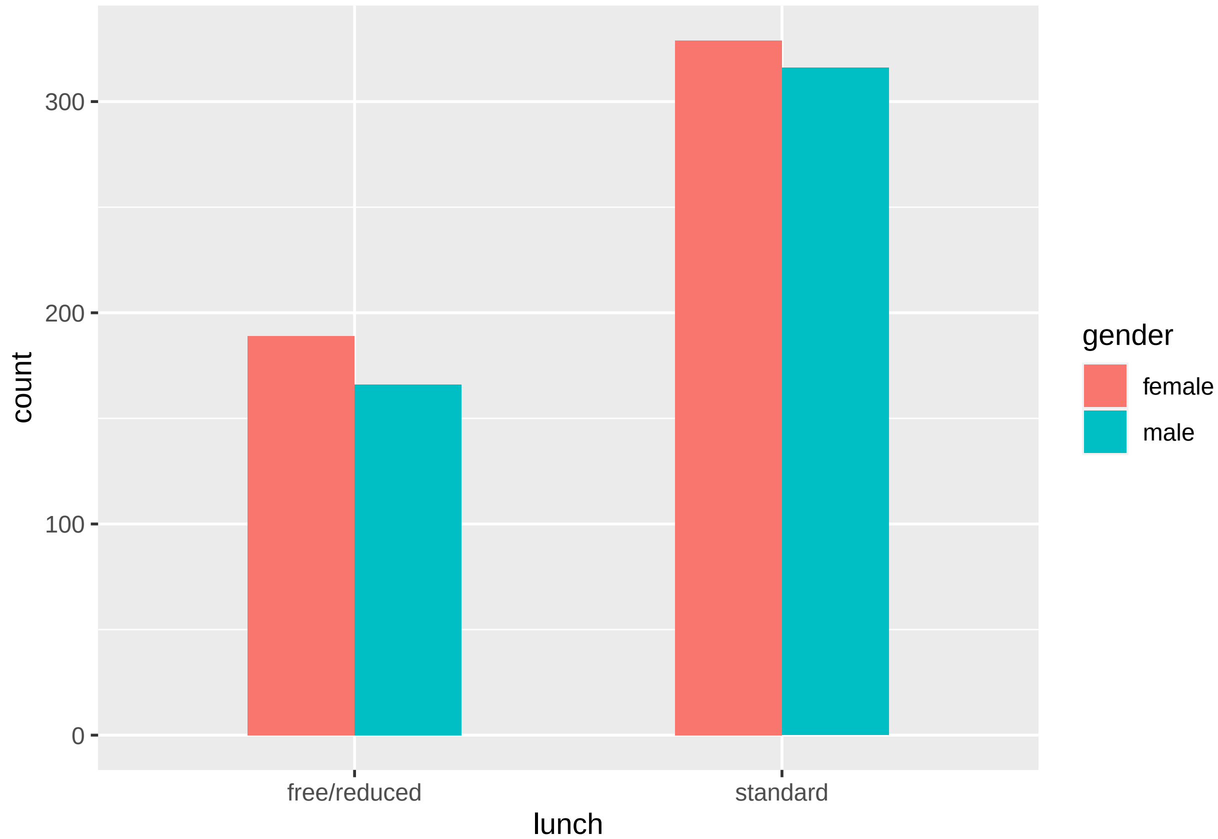
```
head(data)
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor’s degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master’s degree     standard
## 4   male        group A          associate’s degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate’s degree     standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

The data seems to be in tidy format therefore no tidying is required we can go head and reproduce the plots.

## Problem 2

```
p <- ggplot(aes(x=lunch, fill = as.factor(gender)), data=data) +
  geom_bar(position = 'dodge',width=0.5)+ scale_fill_discrete('gender')
p
```




Now let's try to reproduce the next plot

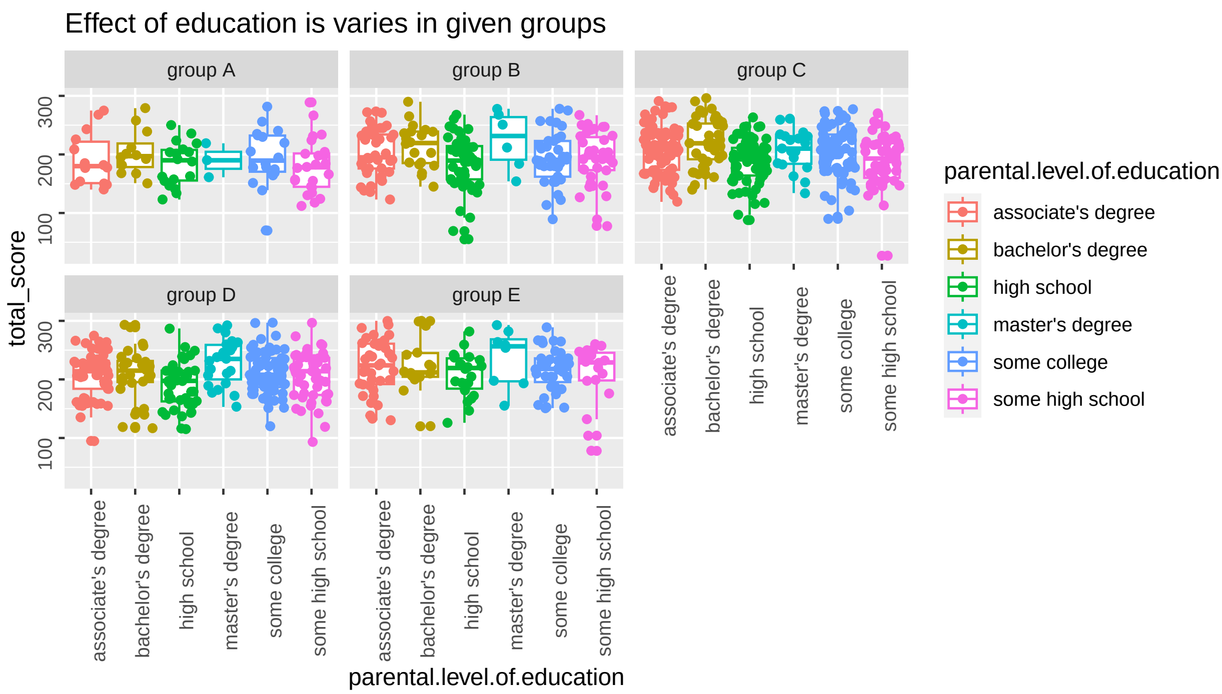
```
data$total_score = data$math.score + data$reading.score + data$writing.score

p <- ggplot(data, aes(x=gender, y=total_score, color=`test.preparation.course`)) +
geom_boxplot(width=0.3)+
facet_wrap(~`test.preparation.course`) +
geom_jitter() + labs(x='Gender',y='Total Scores',title="People tend to give exam without a course")

p
```


Finally the last plot

```
p <- ggplot(data, mapping=aes(x=`parental.level.of.education`, y=total_score, color=`parental.level.of.
p <- p+  geom_boxplot(width=0.8, height=12)
```

```
## Warning in geom_boxplot(width = 0.8, height = 12): Ignoring unknown parameters:
## ‘height‘
```

```
p <- p + geom_jitter()
p <- p + facet_wrap(~`race.ethnicity`)
p <- p+ theme(axis.text=element_text(angle=90))
p <- p +labs(title="Effect of education is varies in given groups")
p
```




All the Plots from the miniposter have been reproduced.

# Part B

## Problem 3

```
#importing the dataset
data(PimaIndiansDiabetes2)
#removing the null values
PimaIndiansDiabetes2 <- na.omit(PimaIndiansDiabetes2)
```

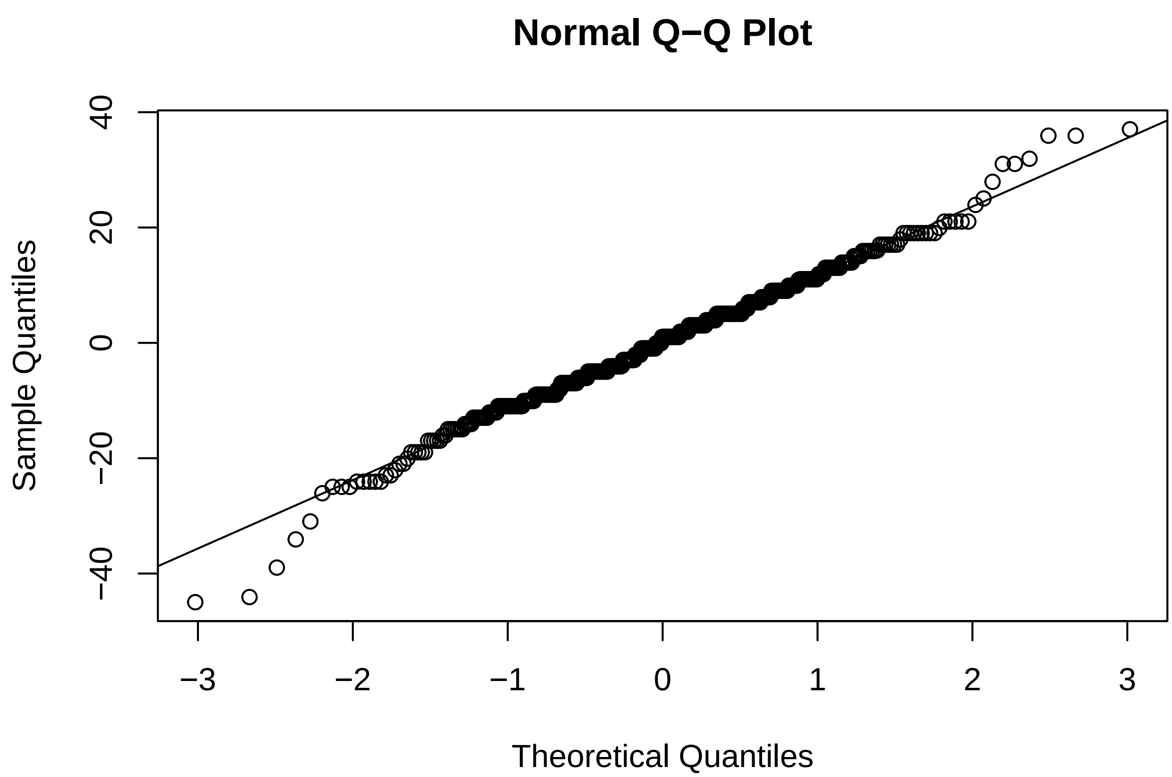
```
#fitting linear model for pressure using diabetes as exploratory variable
fit <- lm(pressure ~ diabetes , data=PimaIndiansDiabetes2)
summary(fit)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes, data = PimaIndiansDiabetes2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.969  -8.077   1.031   7.923  37.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.9695     0.7585  90.927  < 2e-16 ***
## diabetespos   5.1075     1.3172   3.878 0.000124 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.28 on 390 degrees of freedom
## Multiple R-squared:  0.03712,    Adjusted R-squared:  0.03465
## F-statistic: 15.04 on 1 and 390 DF,  p-value: 0.0001237
```

Let Perform Model Diagnosis

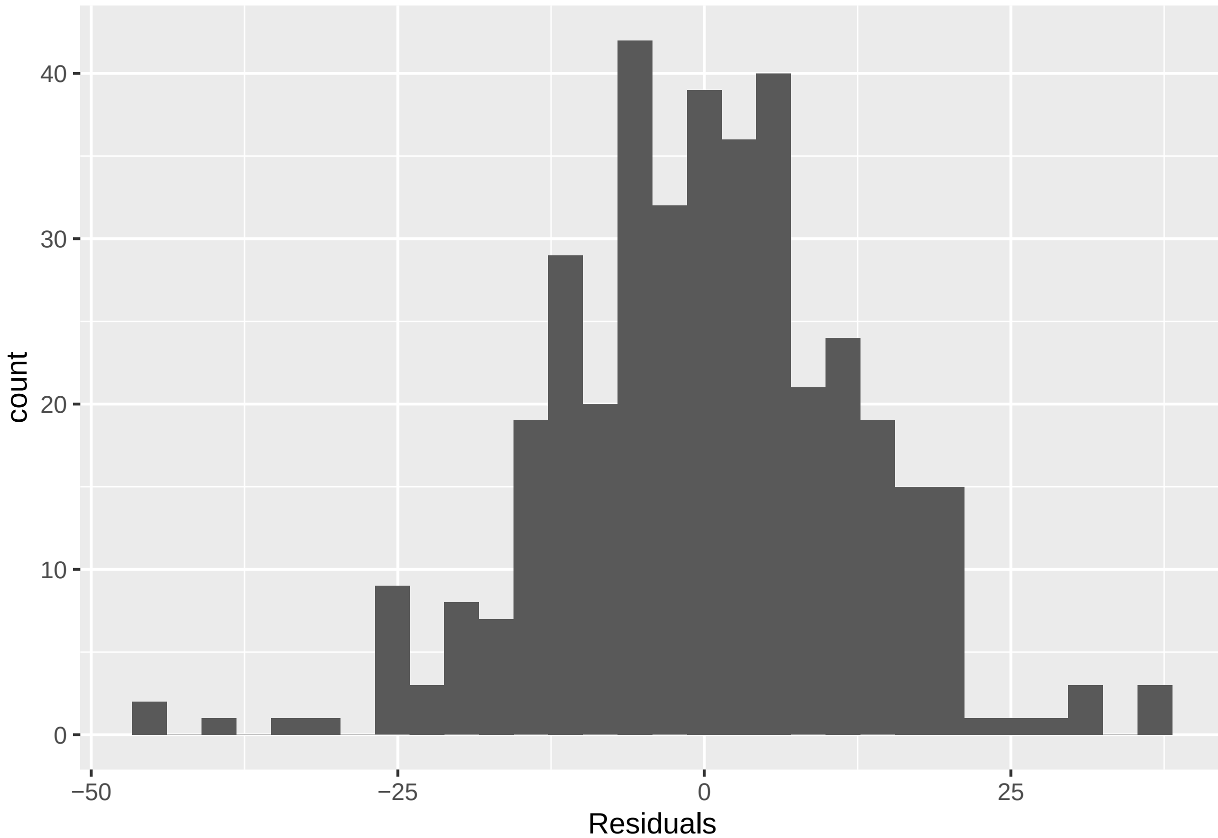
```
res<-resid(fit)
#model Diagnosis
qqnorm(res)
qqline(res)
```




From the QQ plot we can infer that there is no violation of model assumption Lets also check the distribution of residuals

```
ggplot(PimaIndiansDiabetes2, aes(x=res)) +
  geom_histogram()+ labs(x='Residuals')
```

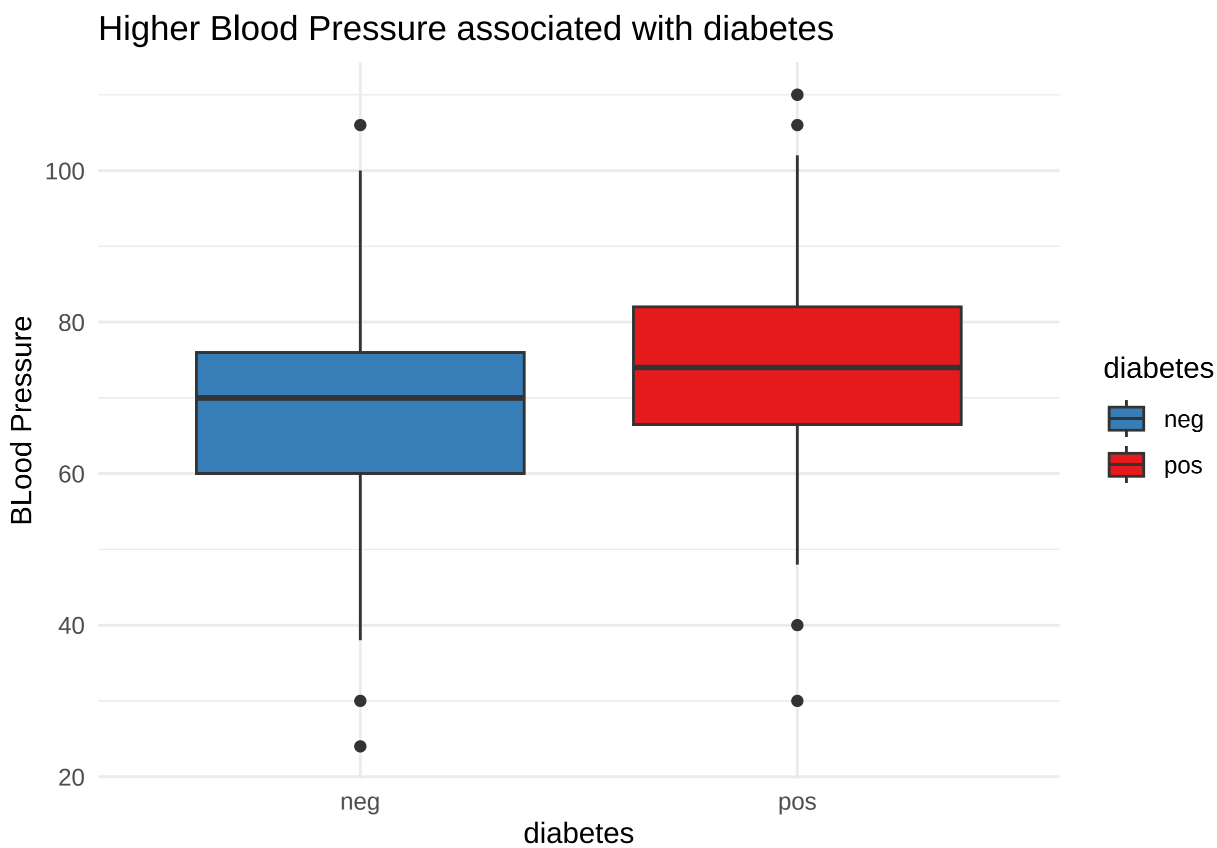
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




As we can from the histogram of residuals they are normally distributed therefore the model assumption is not violated.

Lets Visualize diabetes v/s Blood pressure

```
p <- ggplot(data = PimaIndiansDiabetes2, aes(x= diabetes, y=pressure,fill = diabetes))
p + geom_boxplot()+scale_fill_brewer(palette = 'Set1',direction=-1)+
  labs(y='BLood Pressure',title='Higher Blood Pressure associated with diabetes')+
  theme_minimal()
```




From the above plot we can see that people with Diabetes tend to have higher blood pressure compared to those with no diabetes.

Ho <- Beta = 0 (There is no difference in blood pressure between people with diabetes and ones with no diabetes)

H1 <- Beta not equal to 0 (There is a difference in blood pressure between people with diabetes and people without diabetes)

alpha = 0.05

```
fit <- lm(pressure ~ diabetes, data = PimaIndiansDiabetes2)
summary(fit)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes, data = PimaIndiansDiabetes2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.969  -8.077   1.031   7.923  37.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.9695     0.7585  90.927  < 2e-16 ***
## diabetespos   5.1075     1.3172   3.878 0.000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 12.28 on 390 degrees of freedom
## Multiple R-squared:  0.03712,    Adjusted R-squared:  0.03465
## F-statistic: 15.04 on 1 and 390 DF,  p-value: 0.0001237
```

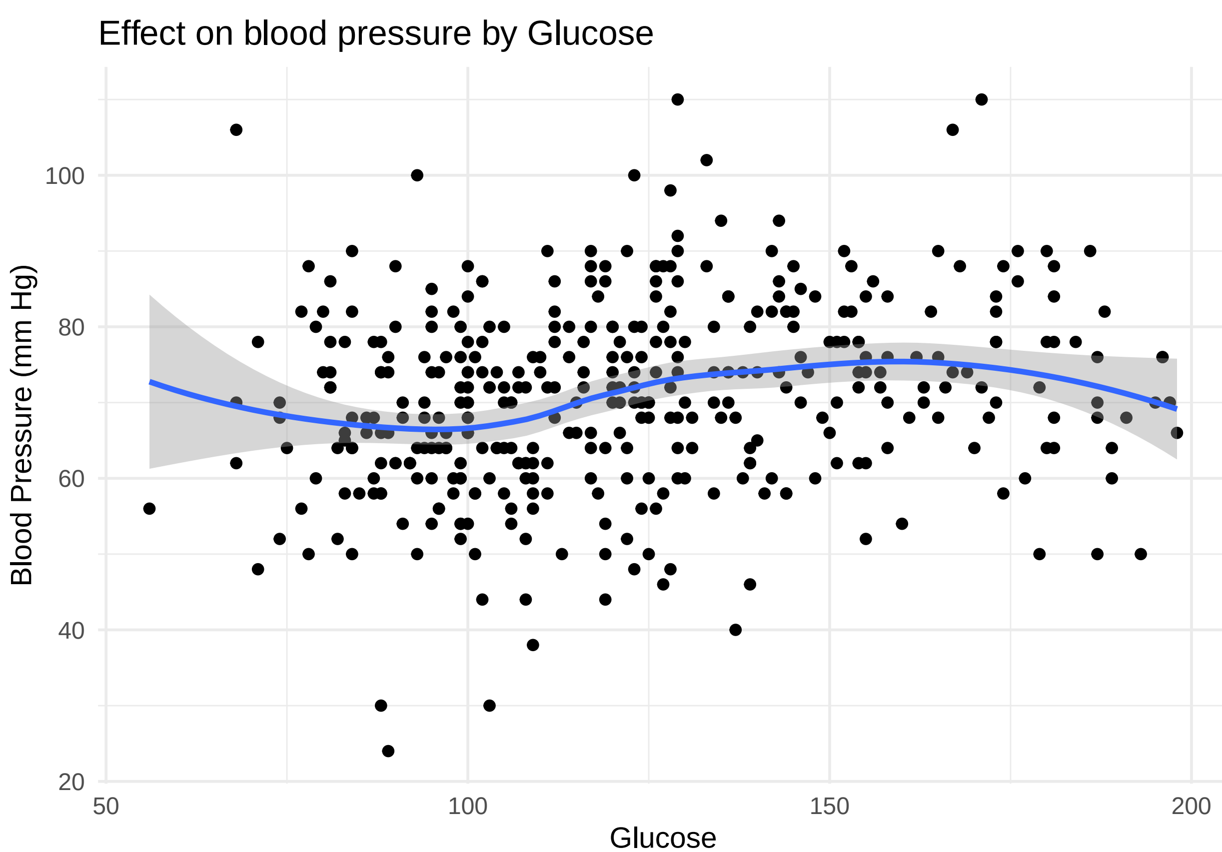
As we can see that the intercept is not equal to zero also `p-value` is 0.000124 which is very low therefore we reject the hypothesis Ho therefore, People with diabetes tend to have higher blood pressure compared to that of People without diabetes.

## Problem 4

lets visualize other features compared to blood pressure Effect of Glucose on Blood Pressure

```
p <- ggplot(data= PimaIndiansDiabetes2, aes(x=glucose,y=pressure))

p+geom_point() + labs(x='Glucose',y='Blood Pressure (mm Hg)',title='Effect on blood pressure by Glucose
  theme_minimal() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

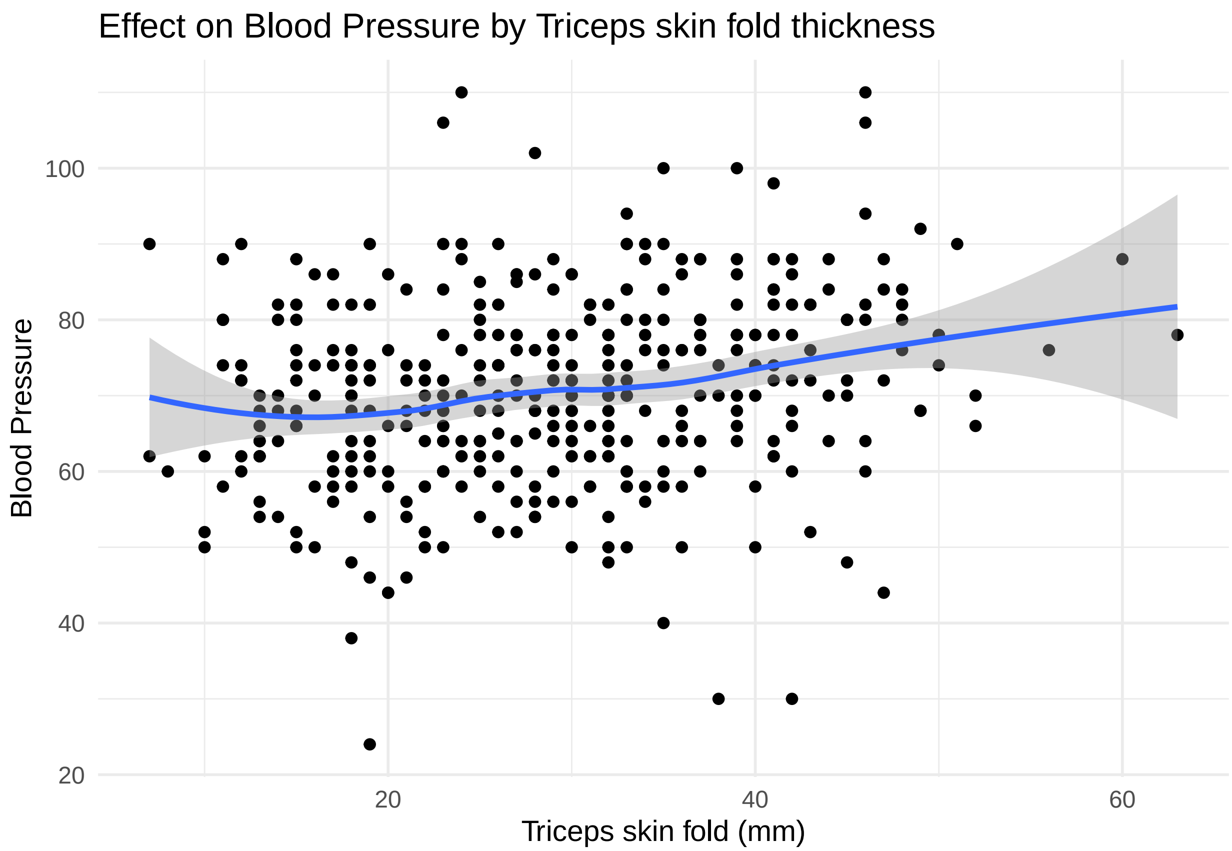



From the above plot we can infer that glucose has some effect on the Blood Pressure

Lets See effect of triceps on Pressure

```r
p <- ggplot(data= PimaIndiansDiabetes2, aes(x=triceps,y=pressure))

p+geom_point() + labs(x='Triceps skin fold (mm)',y='Blood Pressure',title='Effect on Blood Pressure by '
  theme_minimal()
```

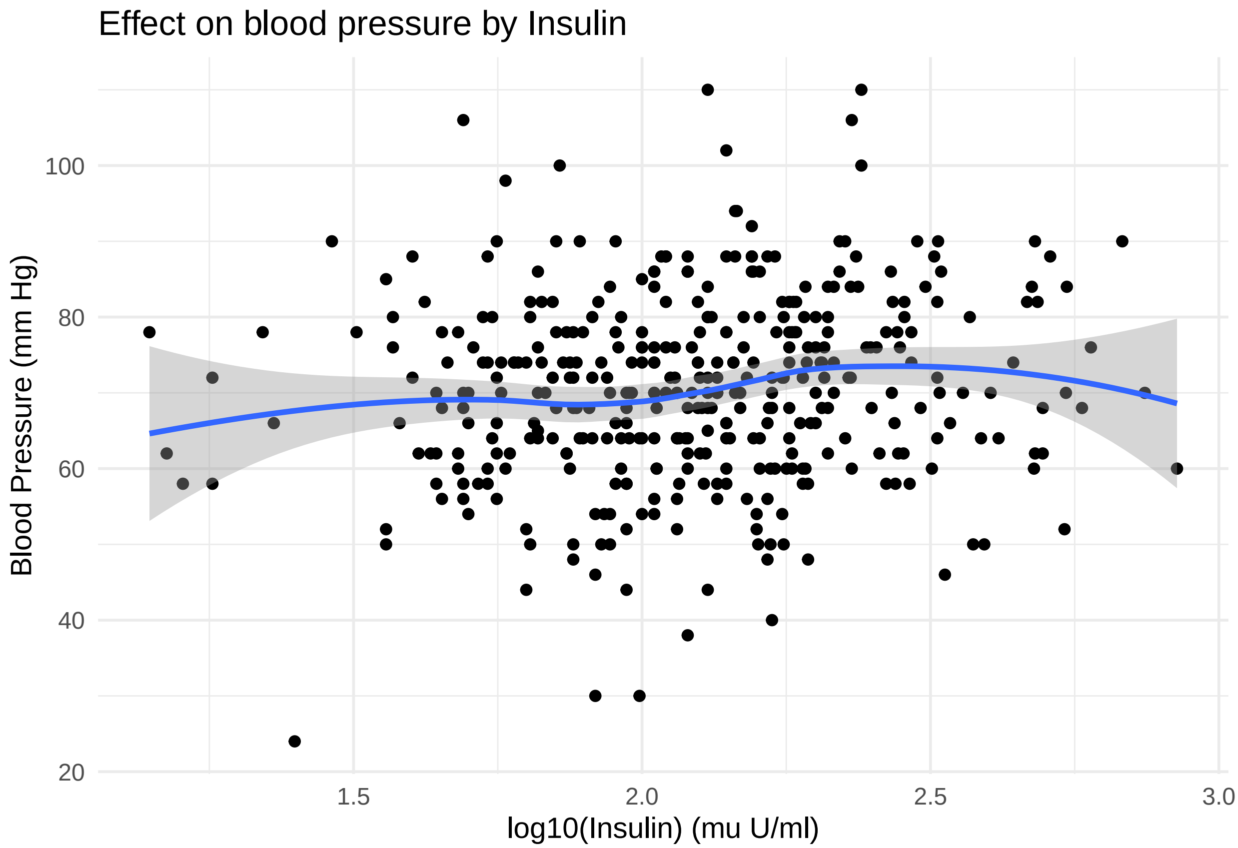
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




As we can see there is very slight between Blood Pressure and Triceps Skin fold

```r
p <- ggplot(data= PimaIndiansDiabetes2, aes(x=log10(insulin),y=pressure))

p+geom_point() + labs(x='log10(Insulin) (mu U/ml)',y='Blood Pressure (mm Hg)',title='Effect on blood pre
  theme_minimal()
```

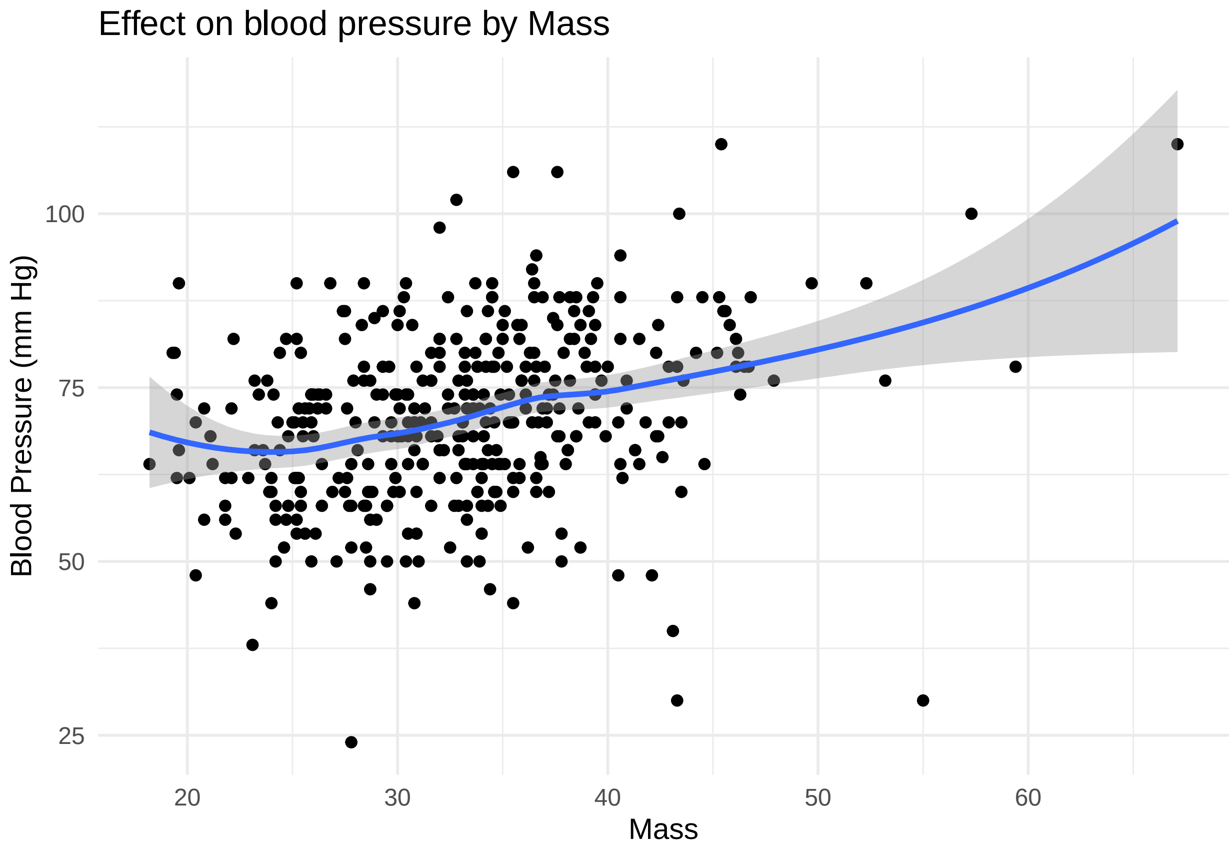
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




We can see that insulin has some effect on the Blood pressure also since the plot was very right skewed I took `log10` of Insulin

```
p <- ggplot(data= PimaIndiansDiabetes2, aes(x=mass,y=pressure))

p+geom_point() + labs(x='Mass',y='Blood Pressure (mm Hg)',title='Effect on blood pressure by Mass')+
  theme_minimal() + geom_smooth()
```

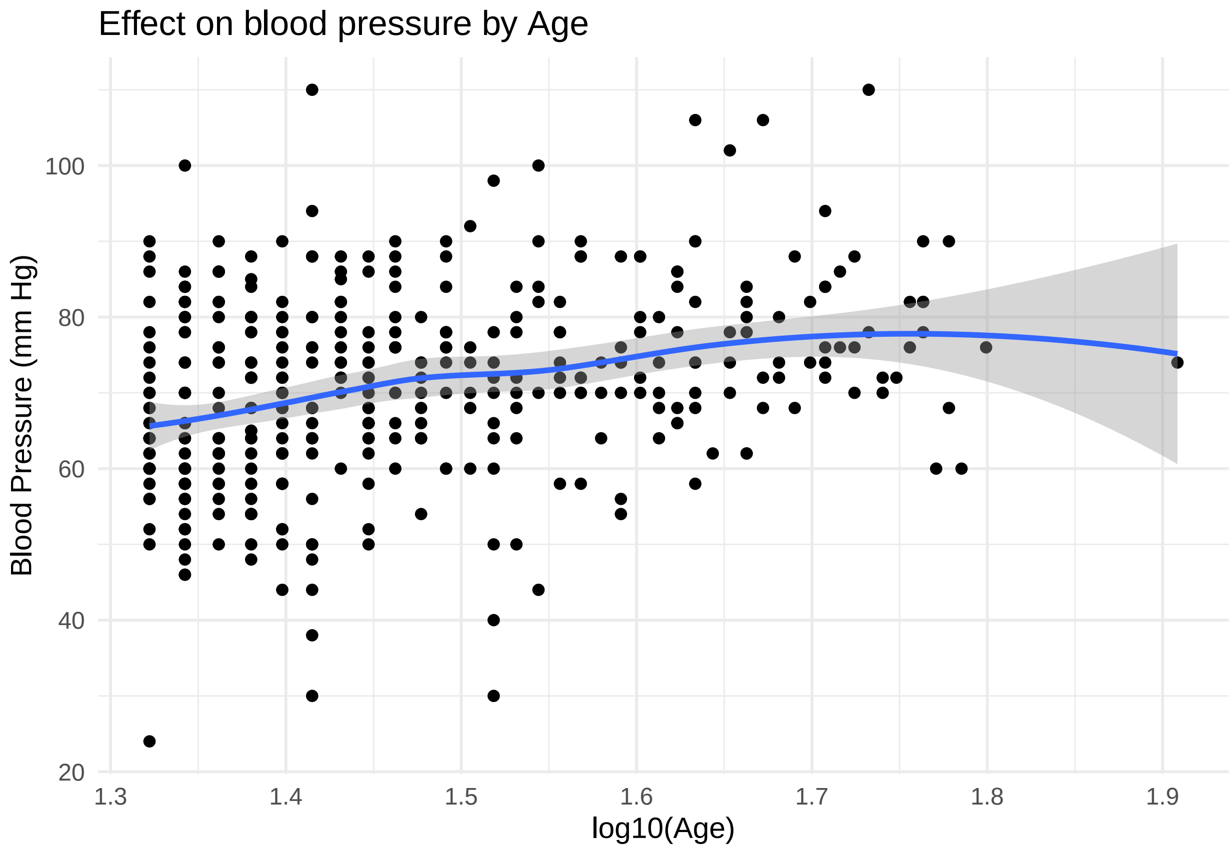
```
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
```




From the above plot we can infer that Mass has a stronger relation to blood pressure compared to other features

```
p <- ggplot(data= PimaIndiansDiabetes2, aes(x=log10(age),y=pressure))

p+geom_point() + labs(x='log10(Age)',y='Blood Pressure (mm Hg)',title='Effect on blood pressure by Age')
  theme_minimal() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




There is a very slight increase in Blood Pressure with Age also since age was also right skewed I took `log10` of the value.

Based on the above visualizations I would consider selecting Mass, log10(Age), Glucose,Tricep and log10(insulin), Diabetes.

```
mlr = lm(pressure ~ diabetes + glucose + log10(insulin) + triceps +mass + log10(age) ,data=PimaIndiansDi
summary(mlr)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes + glucose + log10(insulin) +
##     triceps + mass + log10(age), data = PimaIndiansDiabetes2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.683  -6.936  -0.499   7.792  29.554
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.83573    8.18248   1.935   0.0537 .
## diabetespos    -0.41477    1.50098  -0.276   0.7824
## glucose         0.04571    0.02626   1.740   0.0826 .
## log10(insulin) -3.29620    2.46014  -1.340   0.1811
## triceps        -0.01698    0.07452  -0.228   0.8199
## mass            0.51390    0.11256   4.565 6.72e-06 ***
## log10(age)     27.04431    5.06248   5.342 1.58e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.4 on 385 degrees of freedom
## Multiple R-squared:  0.1811, Adjusted R-squared:  0.1684
## F-statistic: 14.19 on 6 and 385 DF,  p-value: 1.274e-14
```

```
step(mlr, scope = list(lower = ~ diabetes) )
```

```
## Start:  AIC=1914.59
## pressure ~ diabetes + glucose + log10(insulin) + triceps + mass +
##     log10(age)
##
##                  Df Sum of Sq   RSS    AIC
## - triceps         1       6.7 50004 1912.7
## - log10(insulin)  1     233.1 50230 1914.4
## <none>                        49997 1914.6
## - glucose         1     393.3 50390 1915.7
## - mass            1    2706.7 52703 1933.3
## - log10(age)      1    3706.0 53703 1940.6
##
## Step:  AIC=1912.65
## pressure ~ diabetes + glucose + log10(insulin) + mass + log10(age)
##
##                  Df Sum of Sq   RSS    AIC
## - log10(insulin)  1     231.8 50235 1912.5
## <none>                        50004 1912.7
## - glucose         1     391.9 50395 1913.7
## - log10(age)      1    3735.5 53739 1938.9
## - mass            1    4286.4 54290 1942.9
##
## Step:  AIC=1912.46
## pressure ~ diabetes + glucose + mass + log10(age)
##
##              Df Sum of Sq   RSS    AIC
## - glucose     1     192.0 50427 1912.0
## <none>                    50235 1912.5
## - log10(age)  1    3657.1 53892 1938.0
## - mass        1    4079.3 54315 1941.1
##
## Step:  AIC=1911.95
## pressure ~ diabetes + mass + log10(age)
##
##              Df Sum of Sq   RSS    AIC
## <none>                    50427 1912.0
## - log10(age)  1    4167.3 54595 1941.1
## - mass        1    4271.4 54699 1941.8
```

```
##
## Call:
## lm(formula = pressure ~ diabetes + mass + log10(age), data = PimaIndiansDiabetes2)
##
## Coefficients:
```

```
## (Intercept)  diabetespos         mass   log10(age)
##     13.5586       0.3225       0.4885      27.7884
```

After performing Stepwise selection with AIC features Diabetes, mass and log10(age) are the best for predictions.

## Problem 5

Lets do Hypothesis testing for this model.

Ho <- B = 0 (There is no difference in blood pressure between people with diabetes and ones with no diabetes)

H1 <- B not equal to 0 (There is a difference in blood pressure between people with diabetes and people without diabetes)

alpha = 0.05

```r
final_mlr = lm(pressure ~ diabetes + mass + log10(age),data=PimaIndiansDiabetes2)
summary(final_mlr)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes + mass + log10(age), data = PimaIndiansDiabetes2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.068  -7.053  -0.558   7.609  28.635
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.55858    7.60398   1.783   0.0754 .
## diabetespos  0.32247    1.36541   0.236   0.8134
## mass         0.48849    0.08521   5.733 1.99e-08 ***
## log10(age)  27.78836    4.90744   5.662 2.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.4 on 388 degrees of freedom
## Multiple R-squared:  0.1741, Adjusted R-squared:  0.1677
## F-statistic: 27.26 on 3 and 388 DF,  p-value: 5.121e-16
```

From the above statistics we can see tha p-value associated with Diabetes is **0.8134** which is greater then our alpha (0.05) There we **fail to reject the Hypothesis** meaning There is no difference in blood pressure between people with diabetes and ones with no diabetes. The possible reason for this could be the fact that we are taking multiple predictors into account and not just one predictor, Taking only one predictor (diabetes) like in problem 3 would have supported the model to capture more information.