Stat561 Homework1

Group 2: Syeda Mariah Banu, Clara Cherotich Chelibei, Shanmukha Sai Reddy Manukonda, and Sagar Kalauni

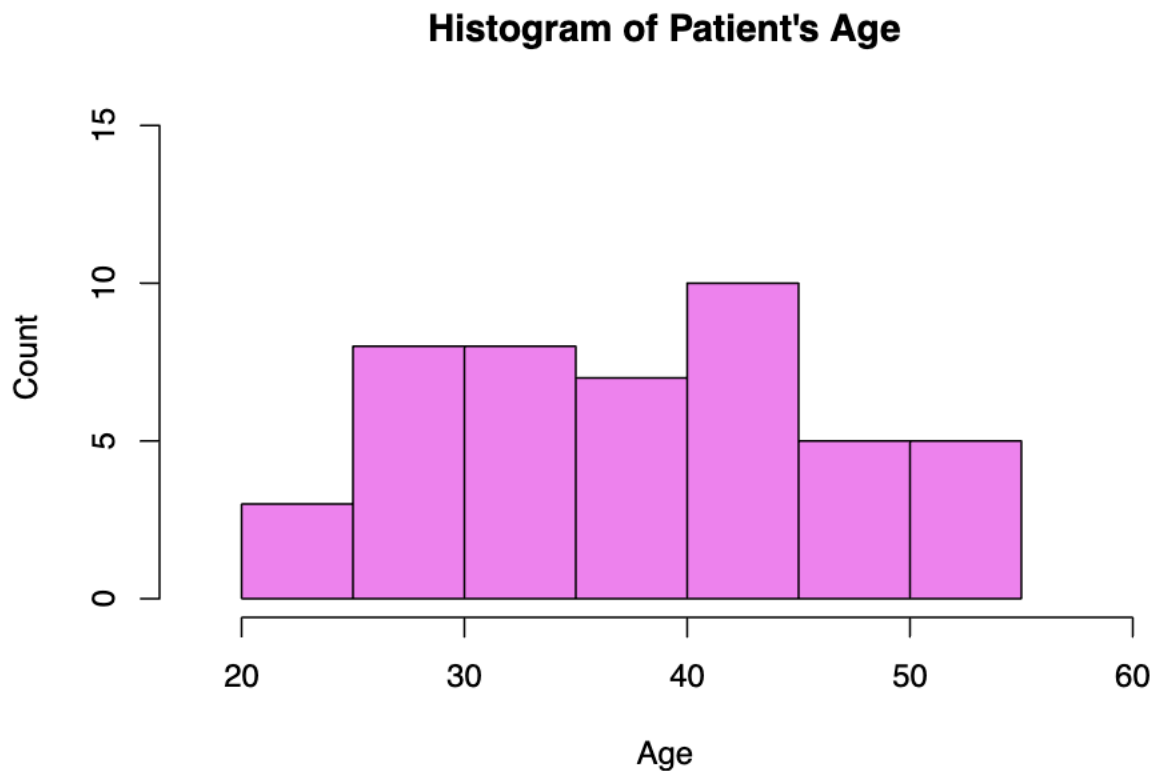
```r
# Load the data:
pat_sat <- read.table("pat_sat.txt", header = TRUE)
head(pat_sat)
```

```
##   pat_sat pat_age severity anxiety
## 1      48      50       51     2.3
## 2      57      36       46     2.3
## 3      66      40       48     2.2
## 4      70      41       44     1.8
## 5      89      28       43     1.8
## 6      36      49       54     2.9
```

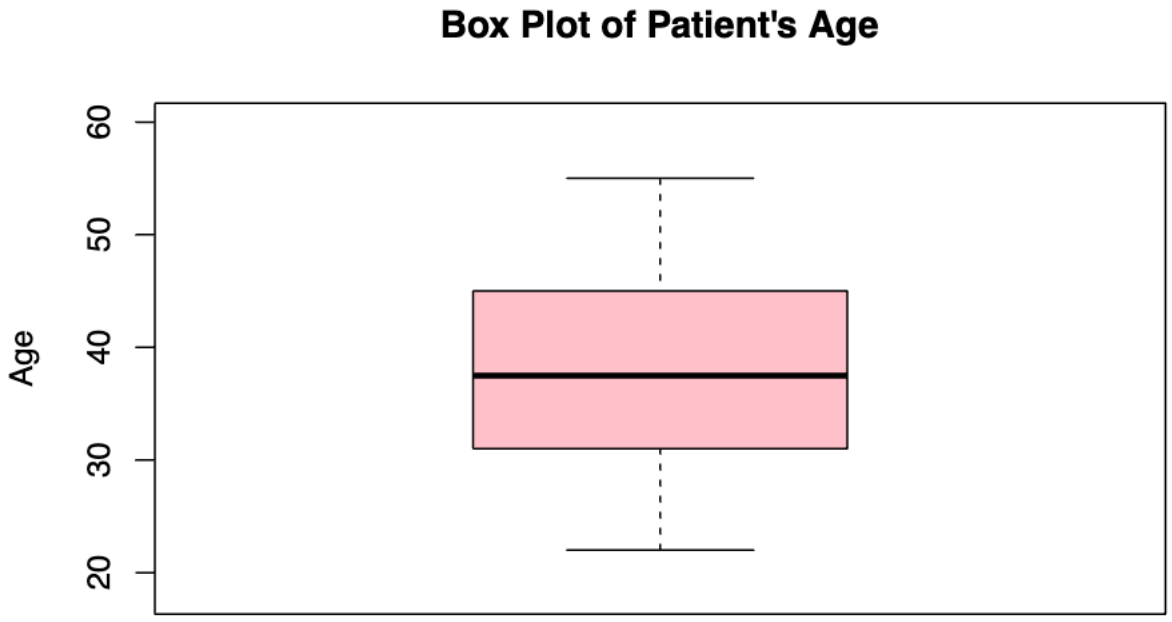
```r
str(pat_sat)
```

```
## 'data.frame':    46 obs. of  4 variables:
##  $ pat_sat : int  48 57 66 70 89 36 46 54 26 77 ...
##  $ pat_age : int  50 36 40 41 28 49 42 45 52 29 ...
##  $ severity: int  51 46 48 44 43 54 50 48 62 50 ...
##  $ anxiety : num  2.3 2.3 2.2 1.8 1.8 2.9 2.2 2.4 2.9 2.1 ...
```

```r
# Analyze the Predictors:

# Patient Age Analysis
hist(pat_sat$pat_age, main = "Histogram of Patient's Age", xlab = "Age", ylab= " Count",
     xlim = c(18,60),
     ylim= c(0,15),
     col= c("violet"))
```



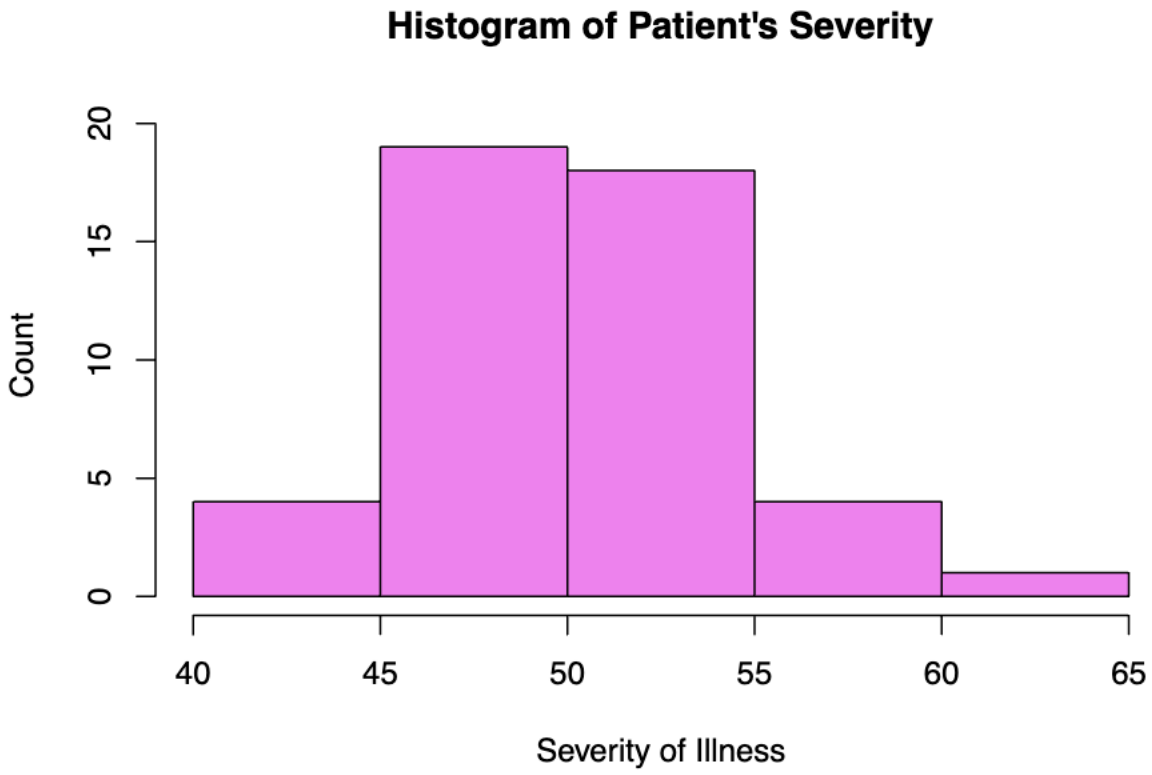

```
boxplot(pat_sat$pat_age, main = "Box Plot of Patient's Age", ylab = "Age",  ylim= c(18,60), col=c( "pin
```



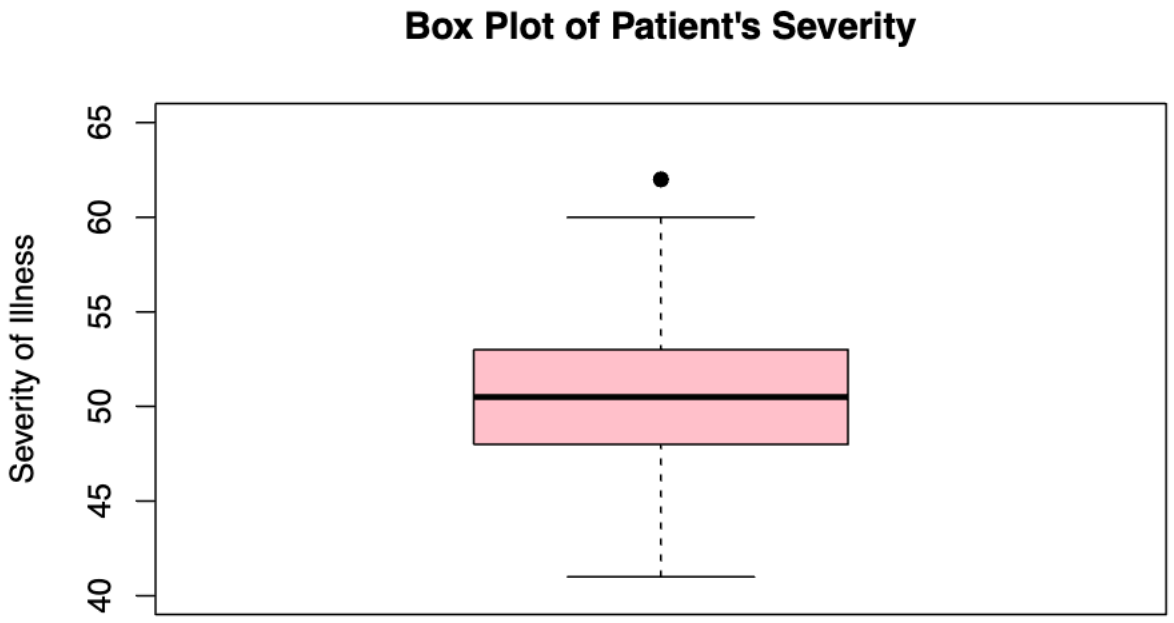

```
summary(pat_sat$pat_age)
```

```
# Severity Analysis
hist(pat_sat$severity, main = "Histogram of Patient's Severity", xlab = "Severity of Illness",ylab= " C
      xlim = c(40,65),
      ylim= c(0,20),
```

```
      col= c("violet"))
```



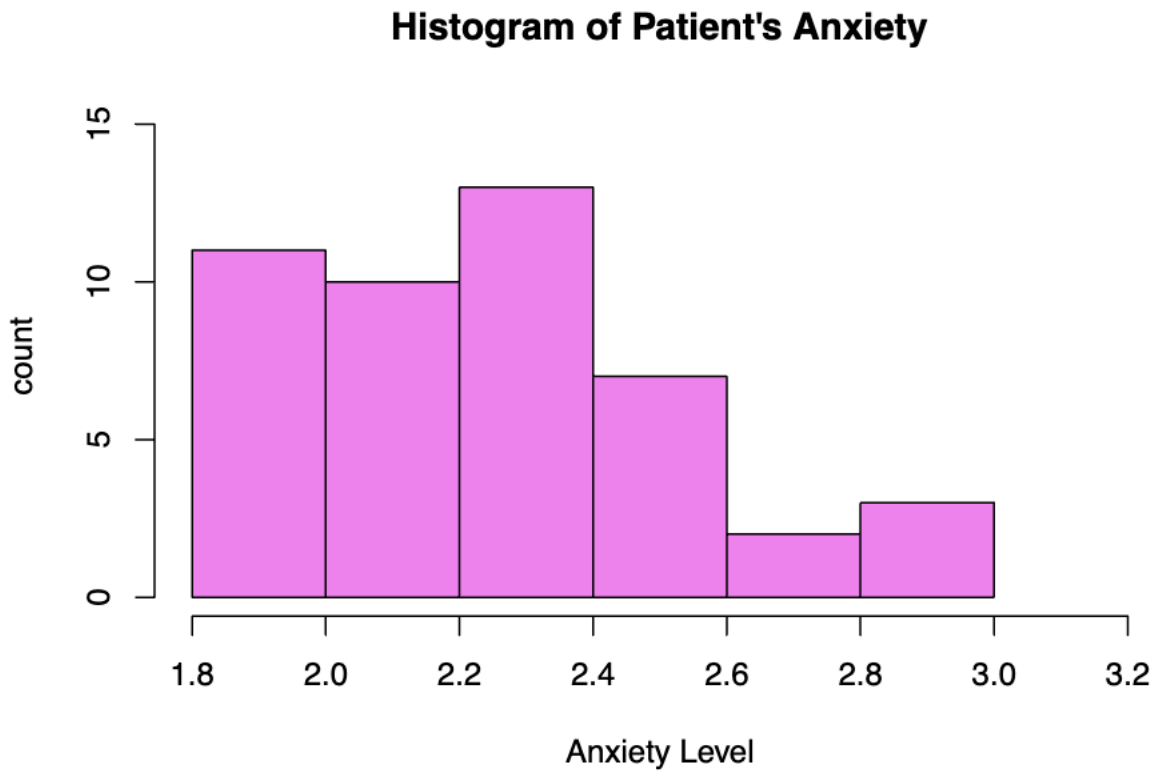

```
boxplot(pat_sat$severity, main = "Box Plot of Patient's Severity", ylab = "Severity of Illness", ylim=
```



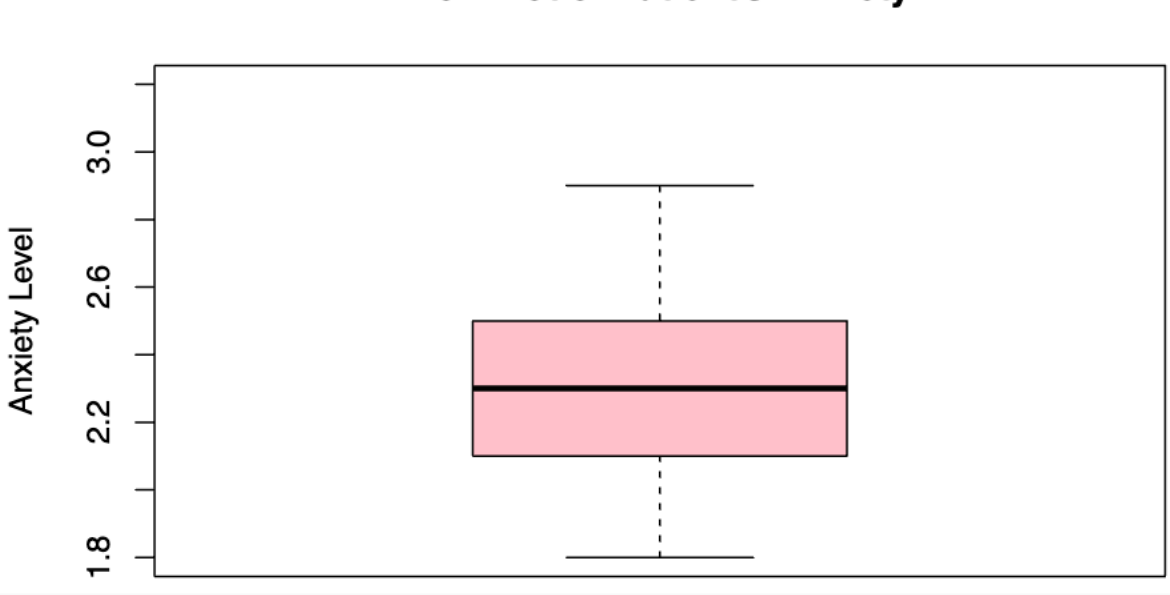

```
summary(pat_sat$severity)
```

```
# Anxiety Analysis
hist(pat_sat$anxiety, main = "Histogram of Patient's Anxiety", xlab = "Anxiety Level", ylab= " count",
```

```
      xlim = c(1.8,3.2),
      ylim= c(0,15),
      col= c("violet"))
```




```
boxplot(pat_sat$anxiety, main = "Box Plot of Patient's Anxiety", ylab = "Anxiety Level",ylim= c(1.8,3.2
```



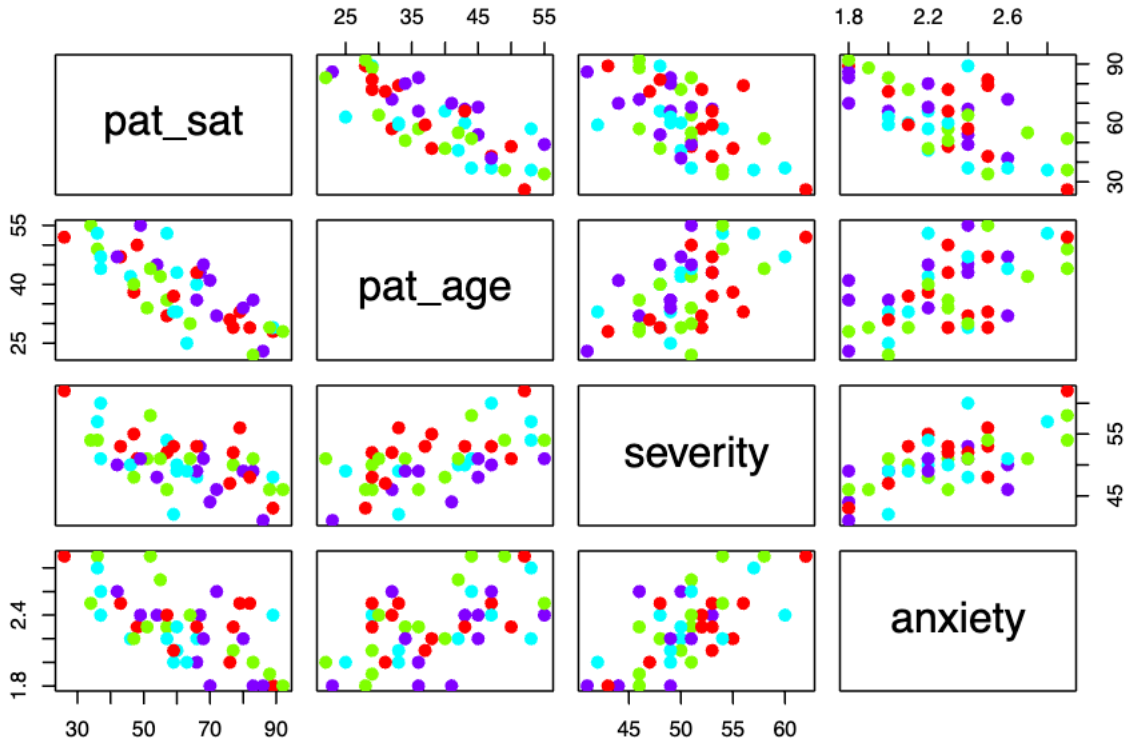

```
summary(pat_sat$anxiety)
```

Question 1.a:

Observations:

- While the multiple different ages and anxiety levels were explored in the experiment, the Severity of Illness tended to mostly center heavily around the range of 45-55. This can be noticed in both the histogram's shape as well as the smaller length of the box plot.
- The mean and median of the age of the people in this experiment were also along the late 30s, with a maximum of 55 and a minimum of 22.
- Looking at the boxplot of the severity of illness, We can see one outlier. Severity of illness data looks some how normally distributed, with mean and median almost equal
- It was interesting to see that no one reported an anxiety level of 3 or 0 and that the minimum reported level was 1.8. This meant that everyone had anxiety on some level.

```
#Scatter Plot Matrix:
n= dim(pat_sat)[2]
pairs(pat_sat, pch=19, cex=1.2,col=rainbow(n))
```




```
#Correlation Matrix:
cor_matrix <- cor(pat_sat)
print(cor_matrix)
```

```
##              pat_sat    pat_age   severity    anxiety
## pat_sat    1.0000000 -0.7867555 -0.6029417 -0.6445910
## pat_age   -0.7867555  1.0000000  0.5679505  0.5696775
## severity -0.6029417  0.5679505  1.0000000  0.6705287
## anxiety  -0.6445910  0.5696775  0.6705287  1.0000000
```

Question 1.b:

Interpretation:

- patient age and patient satisfication seem to be the most correlated pair, with negative correlation as the correlation matrix displays a higher value and the scatter plot seem to show a more linear relationship.
- However, the other two pairs age and severity/anxiety show a lower correlation as the plots are less linear and values in the matrix are lower.
- The age and anxiety scatter plot is the most scattered showing the least correlation between them.
- It can be noted that there is only negative correlations between patient satisfaction and all given variables like patient age, severity of illness, anxiety level, which kind of make sense also because these are the things that reduces the patient's satisfication level.

```r
# Fit the model
model <- lm(pat_sat ~ pat_age + severity + anxiety, data = pat_sat)

# Display the summary of the model to see the estimated coefficients
summary(model)
```

```
##
## Call:
## lm(formula = pat_sat ~ pat_age + severity + anxiety, data = pat_sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## pat_age      -1.1416     0.2148  -5.315 3.81e-06 ***
## severity     -0.4420     0.4920  -0.898   0.3741
## anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

Question 1.c:

($\hat{\beta}_2$) Interpretation:

- The estimated regression coefficient for severity ($\hat{\beta}_2$) is -0.4420.
- Therefore, when other variables are kept constant (e.g., age and anxiety), unit increase in the severity of illness of patient decreases the patient satisfaction level by 0.4420.
- The negative sign highlights an inverse relationship between severity and satisfaction. So when severity increases, patient satisfaction is predicted to decrease.

Question 1.d:

Hypotheses:

- Null Hypothesis H_0: All the coefficients are equal to zero, which implies that none of the predictors have an effect on the response.(i.e Model is not significant)
- Alternative Hypothesis H_1: At least one of the coefficient is not zero, which implies that at least one predictor affects the response.(i.e Model is significant)
- p-value: 1.542e-10 Since p-value is too small $1.542e-10$ compared to level of significance $\alpha = 0.05$, we REJECT null hypothesis H_0.And hence we conclude that the model is significant
- Conclusion: At least one of the predictors (age, severity, anxiety) has an effect on the response patient satisfaction.

```
#Confidence Interval:
confint(model,level=0.9)
```

```
##                    5 %        95 %
## (Intercept) 128.004370 188.9781330
## pat_age      -1.502893  -0.7803305
## severity     -1.269467   0.3854587
## anxiety     -25.411454  -1.5288719
```

Question 1.e:

We can interpret the following results with a 90% confidence:

- Age: The negative values indicate that there is an inverse relationship between Age and Patient Satisfaction. Thus elder people are less likely to be satisfied. An increase in age may cause a decrease in satisfaction between 0.780 and 1.503.
- Severity: As there are both positive and negative values the relationship is not clear between severity and satisfaction. Infact it my not be that significant.
- Anxiety: The negative values again indicate that there is an inverse relationship between Anxiety and Patient Satisfaction. An increase in anxiety may cause a decrease in satisfaction between 1.529 and 25.411.

Question 1.f:

R2 value:

- The value is 0.6822.

-The values of R^2 is 0.6822, and this means that With this model(having all given predictor variables) we can explain almost 68.22% of total variation of the response variable patient satification.

- This indicates that the model was able to fit the data well as an R2 value of 1 indicates a good fit and 0 a bad fit. The model is able to capture the relationships and various aspects of the data.

-Note: We are saying model is good enough to explain almost ~79% variation in the response variable. This does not means model is best.

```r
# New data for prediction
new_data <- data.frame(pat_age=35, severity=45, anxiety=2.2)

# Predict patient satisfaction with prediction intervals
predicted_values <- predict(model, newdata = new_data, interval = "prediction", level = 0.9)

# Display the predicted values
predicted_values
```

```
##        fit      lwr      upr
## 1 69.01029 51.50965 86.51092
```

Question 1.g:

Prediction:

- The model we fit predicts a satisfaction score of 69.01 for the given age, severity, and anxiety.
- With 90% confidence we can say that, for the patient with given age, given level of a anxiety and given severity of illness, the value of patient satisfication score will be within (51.51, 86.51) this interval.

```r
# Full model for backward selection
full_model <- lm(pat_sat ~ ., data = pat_sat)

# Null model for forward selection
null_model <- lm(pat_sat ~ 1, data = pat_sat)

# Backward selection
backward_model <- step(full_model, direction = "backward", trace = 1)
```

```
## Start:  AIC=216.18
## pat_sat ~ pat_age + severity + anxiety
##
##            Df Sum of Sq    RSS    AIC
## - severity  1     81.66 4330.5 215.06
## <none>                  4248.8 216.19
## - anxiety   1    364.16 4613.0 217.97
## - pat_age   1   2857.55 7106.4 237.84
##
## Step:  AIC=215.06
## pat_sat ~ pat_age + anxiety
##
##           Df Sum of Sq    RSS    AIC
## <none>                 4330.5 215.06
## - anxiety  1     763.4 5093.9 220.53
## - pat_age  1    3483.9 7814.4 240.21
```

```r
# Forward selection
forward_model <- step(null_model, scope = list(lower = null_model, upper = full_model), direction = "fo
```

```
## Start:  AIC=262.92
## pat_sat ~ 1
##
##            Df Sum of Sq     RSS    AIC
```

```
## + pat_age   1    8275.4  5093.9 220.53
## + anxiety   1    5554.9  7814.4 240.21
## + severity  1    4860.3  8509.0 244.13
## <none>                  13369.3 262.92
##
## Step:  AIC=220.53
## pat_sat ~ pat_age
##
##            Df Sum of Sq    RSS    AIC
## + anxiety   1    763.42 4330.5 215.06
## + severity  1    480.92 4613.0 217.97
## <none>                  5093.9 220.53
##
## Step:  AIC=215.06
## pat_sat ~ pat_age + anxiety
##
##            Df Sum of Sq    RSS    AIC
## <none>                  4330.5 215.06
## + severity  1    81.659 4248.8 216.19
```

Question 1.h:

Forward and Backward Selection:

- Evidently, both forward and backward selection have produced the same model.

```
# Load the muscle mass dataset
muscle_mass <- read.table("muscle_mass.txt", header = TRUE)

# Calculate the correlation between age and muscle mass
correlation <- cor(muscle_mass$age, muscle_mass$mmass)

# Print the correlation
correlation
```

```
## [1] -0.866064
```

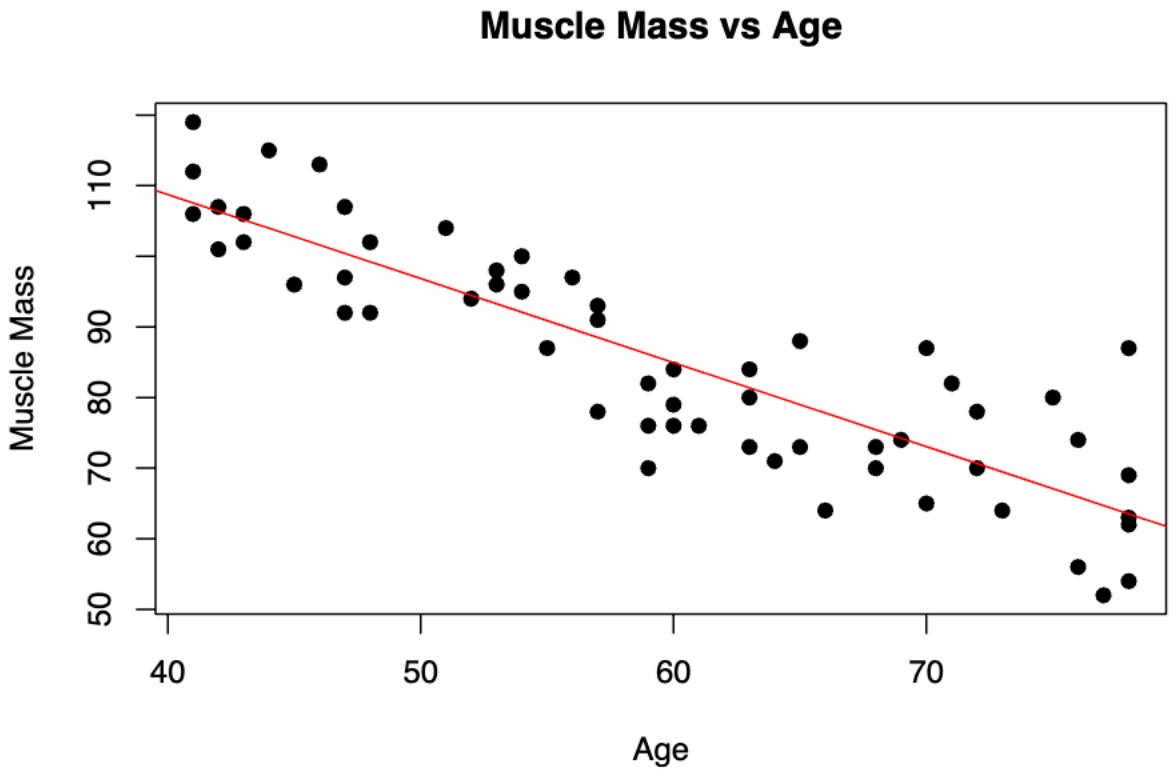
Question 2.a:

Correlation coefficient:

- The coefficient is -0.866064. As this is a negative value, it indicates an inverse relationship between age and muscle mass in women.
- As the value is close to -1 it indicates a strong reverse relationship between age and muscle mass.

```
# Fit the linear model
model <- lm(mmass ~ age, data = muscle_mass)

# Plot the data
plot(muscle_mass$age, muscle_mass$mmass, main = "Muscle Mass vs Age",
     xlab = "Age", ylab = "Muscle Mass", pch = 19)
```

```r
# Add the regression line
abline(model, col = "red")
```




```r
# Print the summary of the model to get R^2
summary(model)
```

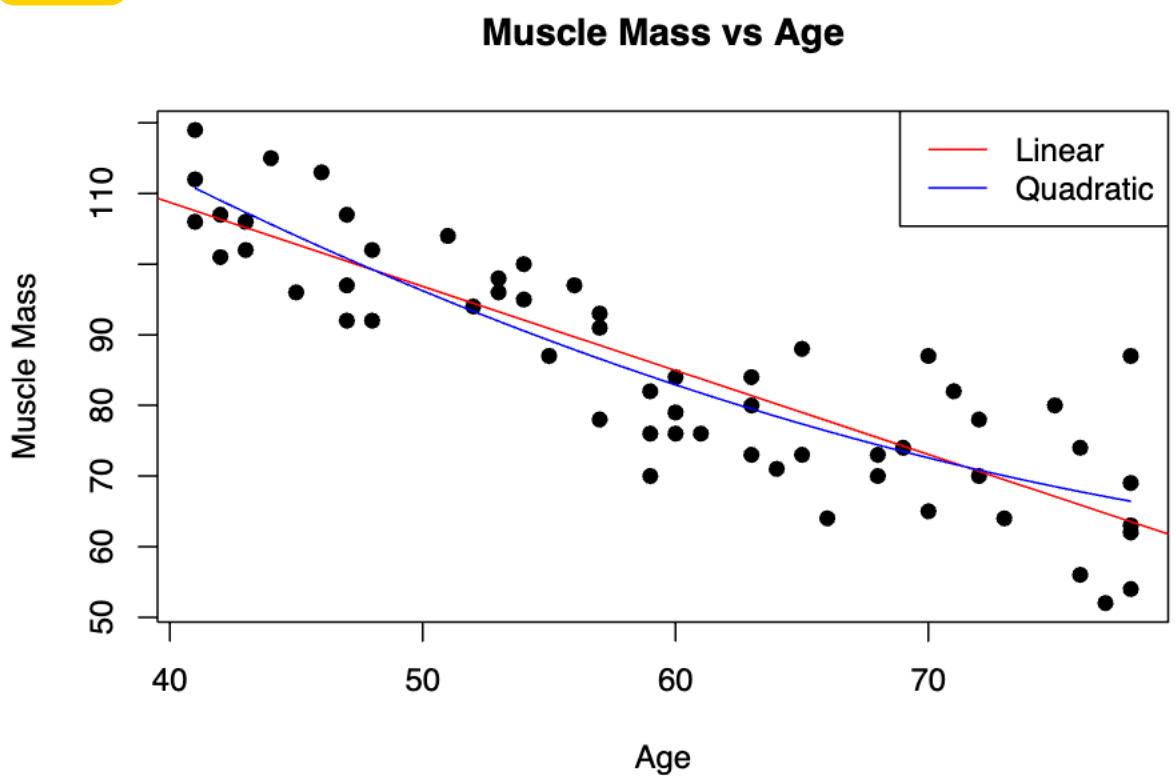
```
##
## Call:
## lm(formula = mmass ~ age, data = muscle_mass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

Question 2.b:

Linear Fit:

- The regression function does appear to be a good fit to the data.
- Although there are some outliers, the negative linearity of the data (a constant decrease in muscle mass as age increases) does seem to be captured by the function.
- The R^2 value is 0.7501, which means 75.01% of the variation in the response variable mmass can be explained by this model and it also testifies that the model fits the data well.

```r
# Fit the second order model
model_second <- lm(mmass ~ age + I(age^2), data = muscle_mass)

# Plot the original data
plot(muscle_mass$age, muscle_mass$mmass, main = "Muscle Mass vs Age",
     xlab = "Age", ylab = "Muscle Mass", pch = 19)

# Add the linear regression line (in red)
abline(model, col = "red")

# Add the quadratic regression curve (in blue)
curve(coef(model_second)[1] + coef(model_second)[2]*x + coef(model_second)[3]*x^2, add = TRUE, col = "b

# Add a legend to the plot
legend("topright", legend = c("Linear", "Quadratic"), col = c("red", "blue"), lty = 1)
```



Question 2.c.d:

Second order Fit:

- From the plot, it seems that the Second Order Regression model (quadratic) captures the data more

closely and fits better.

Question 2.e:

H_0: None of the predictor variable is Significant H_1: At least one of the predictor variable is significant

P-Value:

- Since the p-value: 2.2e-16 is small enough, we REJECT the null hypothesis H_0 and hence we can conclude that there is a significant regression relation for the model i.e at least one of the predictor variable is significant.

Question 2.f:

Hypotheses:

- Null Hypothesis H_0: $\beta_{11} = 0$ i.e The quadratic term does not contribute to the model(NOT significant).
- Alternative Hypothesis H_1: $\beta_{11} \neq 0$ i.e The quadratic term does contribute to the model(significant).
- p-value: 0.08109 -As we can see the p-value associated with β_{11} is 0.0811 which is quite high as compared to level of significance α =0.05. So, we fail to reject null hypothesis and hence conclude that quadratic term is not that significant and hence we can drop it out.
- Conclusion: The quadratic term is not required and the linear term is sufficient.

```
# Fit the third-order polynomial model
model_third <- lm(mmass ~ age + I(age^2) + I(age^3), data = muscle_mass)
```

```
# Get the summary of the model
summary(model_third)
```

```
##
## Call:
## lm(formula = mmass ~ age + I(age^2) + I(age^3), data = muscle_mass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3671  -5.8483  -0.6755   6.1376  20.0637
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.404e+02  1.877e+02   0.748    0.458
## age          5.648e-01  9.822e+00   0.058    0.954
## I(age^2)    -4.559e-02  1.675e-01  -0.272    0.786
## I(age^3)     3.369e-04  9.327e-04   0.361    0.719
##
## Residual standard error: 8.087 on 56 degrees of freedom
## Multiple R-squared:  0.7637, Adjusted R-squared:  0.7511
## F-statistic: 60.34 on 3 and 56 DF,  p-value: < 2.2e-16
```

Question 2.g:

H_0: $\beta_{111} = 0$ i.e cubic term is not significant
H_1: $\beta_{111} \neq 0$ i.e cubic term is significant

Third Order Model P-Value:

- p-value: 0.08109

As we can see the p-value associated with β_{111} is 0.7193 which is quite high as compared to level of significance α =0.05. So, we fail to reject null hypothesis and hence conclude that cubic term is not that significant and hence we can drop it out.

- This shows that a simple linear model is the best function to define the data and adding further complexity does not improve the model.

```r
# Step 1: Load the Data
data <- read.table("cdi.txt", header = TRUE, sep = "", stringsAsFactors = TRUE)

# Step 2: Prepare the Data
# Convert geographic_region to a factor and then to dummy variables
data$geographic_region <- as.factor(data$geographic_region)
data <- cbind(data, model.matrix(~ geographic_region - 1, data=data))

# Step 3: Fit the Regression Model
model <- lm(number_active_physicians ~ total_population + total_personal_income_millions + geographic_r

# Step 4: Extract Results
summary(model)
```

```
##
## Call:
## lm(formula = number_active_physicians ~ total_population + total_personal_income_millions +
##     geographic_region, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -5.848e+01  5.882e+01  -0.994   0.3207
## total_population               5.515e-04  2.835e-04   1.945   0.0524 .
## total_personal_income_millions 1.070e-01  1.325e-02   8.073  6.8e-15 ***
## geographic_region2            -3.493e+00  7.881e+01  -0.044   0.9647
## geographic_region3             4.220e+01  7.402e+01   0.570   0.5689
## geographic_region4            -1.490e+02  8.683e+01  -1.716   0.0868 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

Question 3.a:

Multiple Linear Regression Model:

Encoding

$$X_3 = \begin{cases} 0, & \text{Person Does not live in Region 2} \\ 1, & \text{Person Lives in Region 2} \end{cases}$$

$$X_4 = \begin{cases} 0, & \text{Person Does not live in Region 3} \\ 1, & \text{Person Lives in Region 3} \end{cases}$$

$$X_5 = \begin{cases} 0, & \text{Person Does not live in Region 4} \\ 1, & \text{Person Lives in Region 4} \end{cases}$$

Regression Equation is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Simplified:

$$\hat{y} = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3, & \text{Person lives in Region 2} \\ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4, , & \text{Person Lives in Region 3} \\ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5, & \text{Person Lives in Region 4} \\ \beta_0 + \beta_1 X_1 + \beta_2 X_2, & \text{Person Lives in Region 1} \end{cases}$$

$$X_3 = \begin{cases} -58.48 + 0.0005515 X_1 + 0.107 X_2 - 3.493, & \text{Person lives in Region 2} \\ -58.48 + 0.0005515 X_1 + 0.107 X_2 + 42.2, & \text{Person Lives in Region 3} \\ -58.48 + 0.0005515 X_1 + 0.107 X_2 - 149, & \text{Person Lives in Region 4} \\ -58.48 + 0.0005515 X_1 + 0.107 X_2, & \text{Person Lives in Region 1} \end{cases}$$

$$X_3 = \begin{cases} -61.9735 + 0.0005515 X_1 + 0.107 X_2, & \text{Person lives in Region 2} \\ -16.3735 + 0.0005515 X_1 + 0.107 X_2, & \text{Person Lives in Region 3} \\ -207.48 + 0.0005515 X_1 + 0.107 X_2, & \text{Person Lives in Region 4} \\ -58.48 + 0.0005515 X_1 + 0.107 X_2, & \text{Person Lives in Region 1} \end{cases}$$

- X3, X4, and X5, are geographic_region2, geographic_region3, and geographic_region4 variables, respectively.
- These are dummy variables used to encode categorical data representing different geographic regions.

Question 3.b:

- β_2 (Total Personal Income in Millions): This shows how the income level affects the number of active physicians. Higher income would result in higher number of physicians.
- Statistically, we can say that for every 1 million dollar increase in total personal income, the number of active physicians will increase by 0.0107 units, given that all other variables are kept constant.
- β_3 (Geographic Region 2): This tells if Region 2 has more or fewer active physicians compared to the base region (Region 1), given that all other variables are kept constant.

- Statistically, we can say that number of active physicians living in region 2 are on average 3.493 less then that of the number of active physicians living in base region (Region 1).

Alternative approach[Not for grading, does this work??]

Multiple Linear Regression Model:

equation will be $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \varepsilon_i$

- X3, X4, and X5, are geographic_region2, geographic_region3, and geographic_region4 variables, respectively.
- These are dummy variables used to encode categorical data representing different geographic regions.

Stat561 Homework2

Group 2: Syeda Mariah Banu, Clara Cherotich Chelibei, Shanmukha Sai Reddy Manukonda, and Sagar Kalauni

1. Absenteesim data In this work, we tackle the challenge of absenteeism in today's dynamic work environment using logistic regression. Leveraging a detailed dataset which was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil, we aim to predict and understand the patterns of absenteeism. The features to be used are: Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Education, Son, Pet, Weight, Height, Body mass index. Be sure to make the categorical variables factors in R. Your task is to develop a multinomial logistic regression model to predict the level of absenteeism (Low, Moderate, High) of an employee based on various predictors. The absenteeism time in hours should be categorized into three groups: Low: $0 - 20$ hours, Moderate: $21 - 40$ hours, and High Absenteeism: > 40 hours.

```
absent= read_excel("Absenteeism_at_work.xls")
names(absent) #column names
```

```
##  [1] "ID"                              "Reason for absence"
##  [3] "Month_of_absence"                "Day_of_the_week"
##  [5] "Seasons"                         "Transportation_expense"
##  [7] "Distance_from_Residence_to_Work" "Service_time"
##  [9] "Age"                             "Work_load_Average_in_days"
## [11] "Hit_ target"                     "Disciplinary_failure"
## [13] "Education"                       "Son"
## [15] "Social_drinker"                  "Social_smoker"
## [17] "Pet"                             "Weight"
## [19] "Height"                          "Body_mass_index"
## [21] "Absenteeism_time_in_hours"
```

```
head(absent)
```

```
## # A tibble: 6 x 21
##      ID `Reason for absence` Month_of_absence Day_of_the_week Seasons
##   <dbl>                <dbl>            <dbl>           <dbl>   <dbl>
## 1    11                   26                7               3       1
## 2    36                    0                7               3       1
## 3     3                   23                7               4       1
## 4     7                    7                7               5       1
## 5    11                   23                7               5       1
## 6     3                   23                7               6       1
## # i 16 more variables: Transportation_expense <dbl>,
## #   Distance_from_Residence_to_Work <dbl>, Service_time <dbl>, Age <dbl>,
## #   Work_load_Average_in_days <dbl>, `Hit_ target` <dbl>,
## #   Disciplinary_failure <dbl>, Education <dbl>, Son <dbl>,
## #   Social_drinker <dbl>, Social_smoker <dbl>, Pet <dbl>, Weight <dbl>,
## #   Height <dbl>, Body_mass_index <dbl>, Absenteeism_time_in_hours <dbl>
```

```
sum(is.na(absent))  # No null values
```

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

```
#recode absenteeism
absent <- absent %>%
  mutate(Absenteeism = case_when(
```

```
    Absenteeism_time_in_hours <= 20 ~ "Low",
    Absenteeism_time_in_hours <= 40 ~ "Moderate",
    Absenteeism_time_in_hours > 40 ~ "High"
  ))
#View(absent)
```

```
# Making proper data structure by formatting proper data type by looking the meta data of the data.

absent$Month_of_absence = as.factor(absent$Month_of_absence)
absent$Day_of_the_week = as.factor(absent$Day_of_the_week)
absent$Seasons = as.factor(absent$Seasons)
absent$Education = as.factor(absent$Education)
absent$Absenteeism = as.factor(absent$Absenteeism)
```

```
str(absent)
```

```
## tibble [740 x 22] (S3: tbl_df/tbl/data.frame)
##  $ ID                               : num [1:740] 11 36 3 7 11 3 10 20 14 1 ...
##  $ Reason for absence               : num [1:740] 26 0 23 7 23 23 22 23 19 22 ...
##  $ Month_of_absence                 : Factor w/ 13 levels "0","1","2","3",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ Day_of_the_week                  : Factor w/ 5 levels "2","3","4","5",..: 2 2 3 4 4 5 5 5 1 1 ...
##  $ Seasons                          : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Transportation_expense           : num [1:740] 289 118 179 279 289 179 361 260 155 235 ...
##  $ Distance_from_Residence_to_Work: num [1:740] 36 13 51 5 36 51 52 50 12 11 ...
##  $ Service_time                     : num [1:740] 13 18 18 14 13 18 3 11 14 14 ...
##  $ Age                              : num [1:740] 33 50 38 39 33 38 28 36 34 37 ...
##  $ Work_load_Average_in_days        : num [1:740] 239554 239554 239554 239554 239554 ...
##  $ Hit_ target                      : num [1:740] 97 97 97 97 97 97 97 97 97 97 ...
##  $ Disciplinary_failure             : num [1:740] 0 1 0 0 0 0 0 0 0 0 ...
##  $ Education                        : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...
##  $ Son                              : num [1:740] 2 1 0 2 2 0 1 4 2 1 ...
##  $ Social_drinker                   : num [1:740] 1 1 1 1 1 1 1 1 1 0 ...
##  $ Social_smoker                    : num [1:740] 0 0 0 1 0 0 0 0 0 0 ...
##  $ Pet                              : num [1:740] 1 0 0 0 1 0 4 0 0 1 ...
##  $ Weight                           : num [1:740] 90 98 89 68 90 89 80 65 95 88 ...
##  $ Height                           : num [1:740] 172 178 170 168 172 170 172 168 196 172 ...
##  $ Body_mass_index                  : num [1:740] 30 31 31 24 30 31 27 23 25 29 ...
##  $ Absenteeism_time_in_hours        : num [1:740] 4 0 2 4 2 2 8 4 40 8 ...
##  $ Absenteeism                      : Factor w/ 3 levels "High","Low","Moderate": 2 2 2 2 2 2 2 2 3 2
```

Exploratory data analysis (EDA)

Use R to create any of the following to help answer the following questions: histograms, box plots, scatter plots, correlations and or correlation matrices, bar graphs, summary(), etc.

Question 1.a

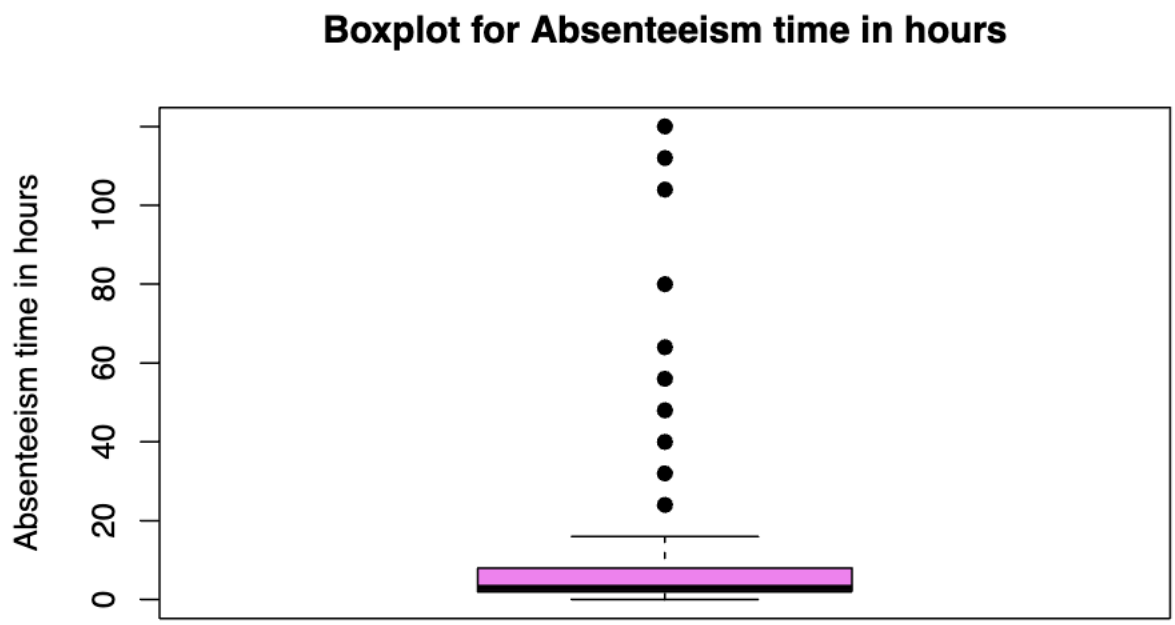
(a) How is the 'Absenteeism time in hours' distributed? Are there any noticeable patterns or outliers? ANS:- To look at the distribution of Absenteeism time in hours and for any possible outlier in it, we can look for summary and boxplot for it.

```r
summary(absent$Absenteeism_time_in_hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   6.924   8.000 120.000
```

Looking for outlier

```r
boxplot(absent$Absenteeism_time_in_hours, ylab= "Absenteeism time in hours", pch=19, col = "violet", ma
```




```r
# looking for the outlier from the boxplot
outliers_1 = boxplot.stats(absent$Absenteeism_time_in_hours)$out
unique(outliers_1)
```

```
##  [1]  40  32  24  64  56  80 120 112 104  48
```

Observation

- The data distribution for Absenteeism time in hours is: min=0, max=120, median=2 and mean=6
- The Absenteeism time in hours data consist of outliers also (10 unique outliers).
- The mean is drawn up from the dedian due to several higher values: Positively skewed or skewed to the right.
- Most of the points lie in the low absenteeism range of 0-20.
- In fact the median and mean are 3 an 6.924 respectively which both also lie within this range. The higher mean indicates a skew in the data towards the higher side.

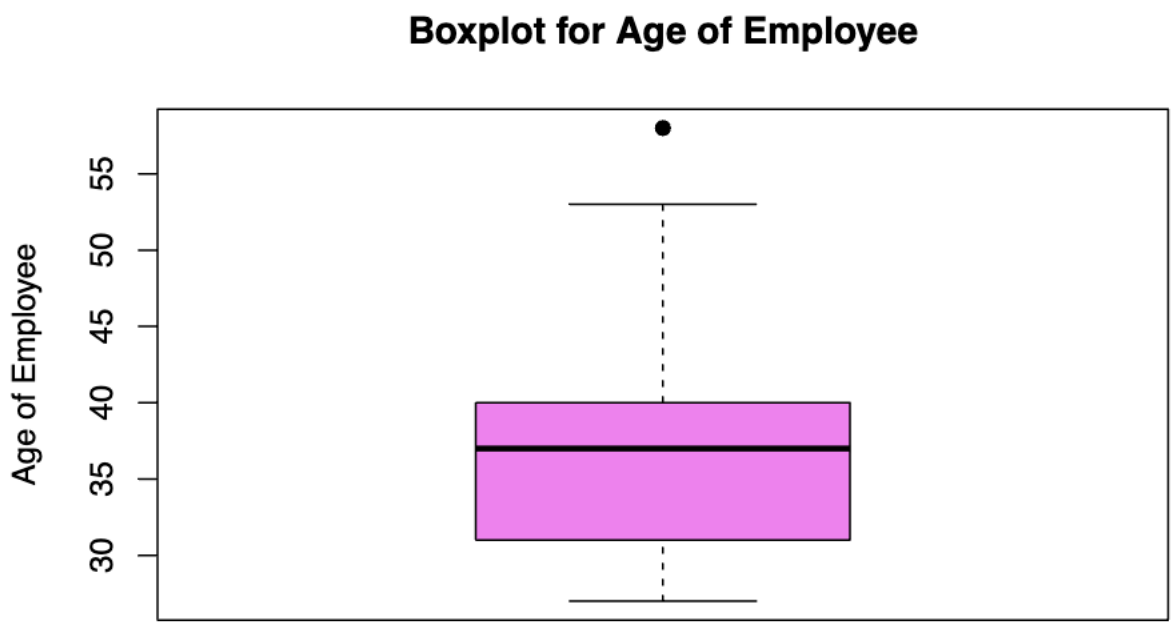
Question 1.b:

(b) What is the distribution of ages among the employees? Are certain age groups more prevalent? ANS:- To look at the distribution of Absenteeism time in hours and for any possible outlier in it, we can look for summary and boxplot for it.

```r
summary(absent$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.00   31.00   37.00   36.45   40.00   58.00
```

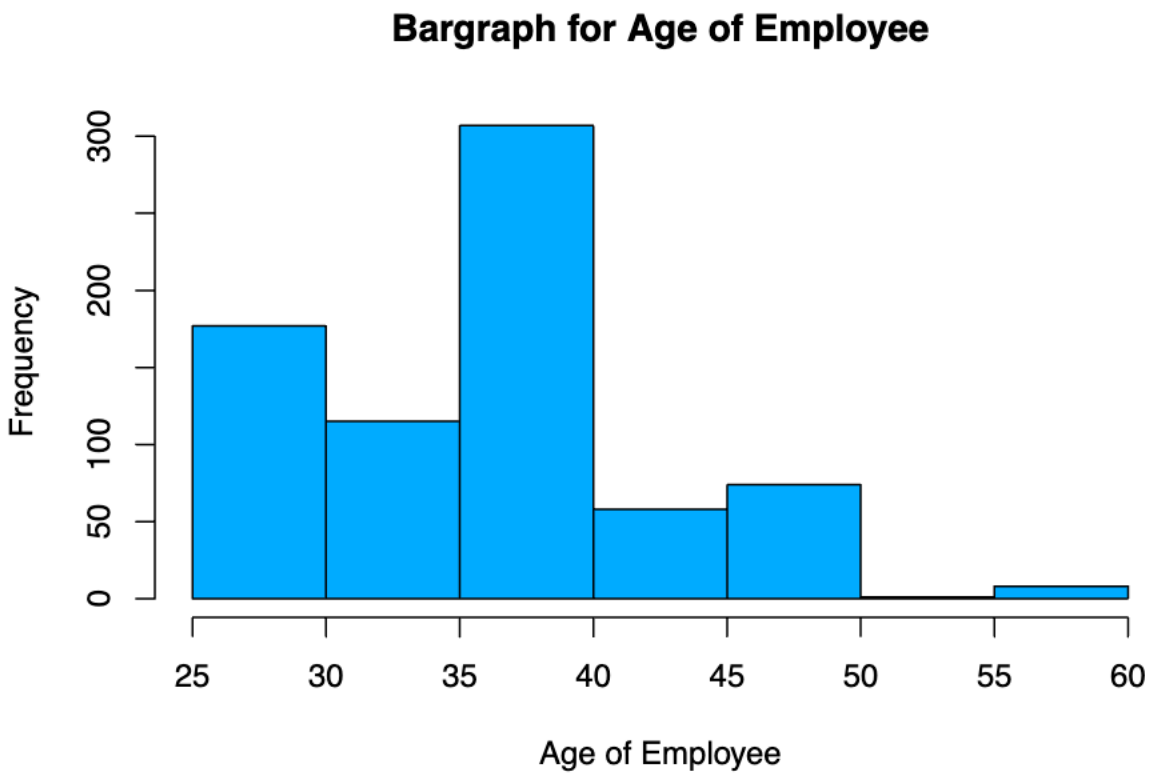
```r
boxplot(absent$Age, ylab= "Age of Employee", pch=19, col = "violet", main ="Boxplot for Age of Employee
```




```r
# looking for the outlier from the boxplot
outliers_2 = boxplot.stats(absent$Age)$out
unique(outliers_2)
```

```
## [1] 58
```

```r
# looking for the histogram for the ages
hist(absent$Age, main = "Bargraph for Age of Employee",
        xlab = "Age of Employee", col = "#00abff", freq = TRUE)
```




```
table(absent$Age)
```

```
##
##  27  28  29  30  31  32  33  34  36  37  38  39  40  41  43  46  47  48  49  50
##   7 117   7  46  22  13  51  29  50  78 113   8  58  34  24   2  24   6   5  37
##  53  58
##   1   8
```

```
unique(absent$Age)
```

```
##  [1] 33 50 38 39 28 36 34 37 41 47 29 48 32 27 43 40 31 30 49 58 46 53
```

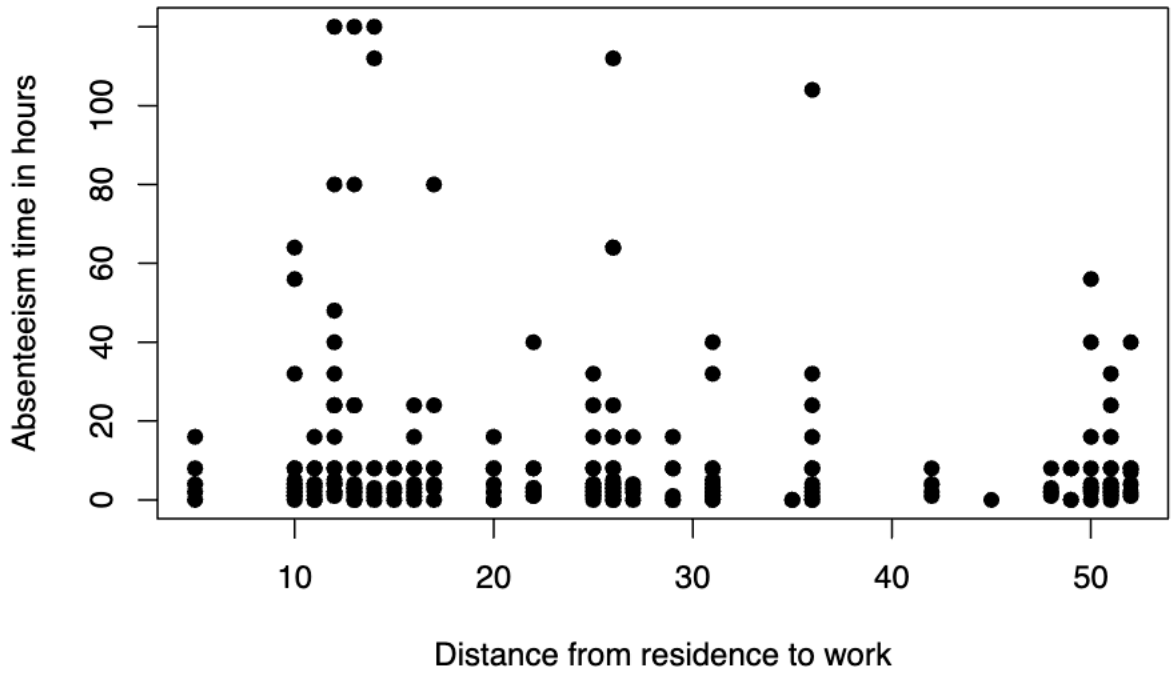
Observation

- The age is not evenly distributed either but is closer to a slight bell curve shape.
- The data distribution for Age of the employee is: min=27, max=58, median=37 and mean=36
- The Age of the employee data consist of outlier. (1 unique outlier).
- Age group of 35-40 is the most repeated in the whole data set.
- In fact the median and mean are 37 an 36.45 respectively which both also lie within this range. This indicates that the average employees are in their mid 30s (more prevalent).[113 employees]
- Data are not normally distributed

Question 1.c:

(c) Is there a correlation between the distance from residence to work and absenteeism time? ANS:- We will look for the following things over here

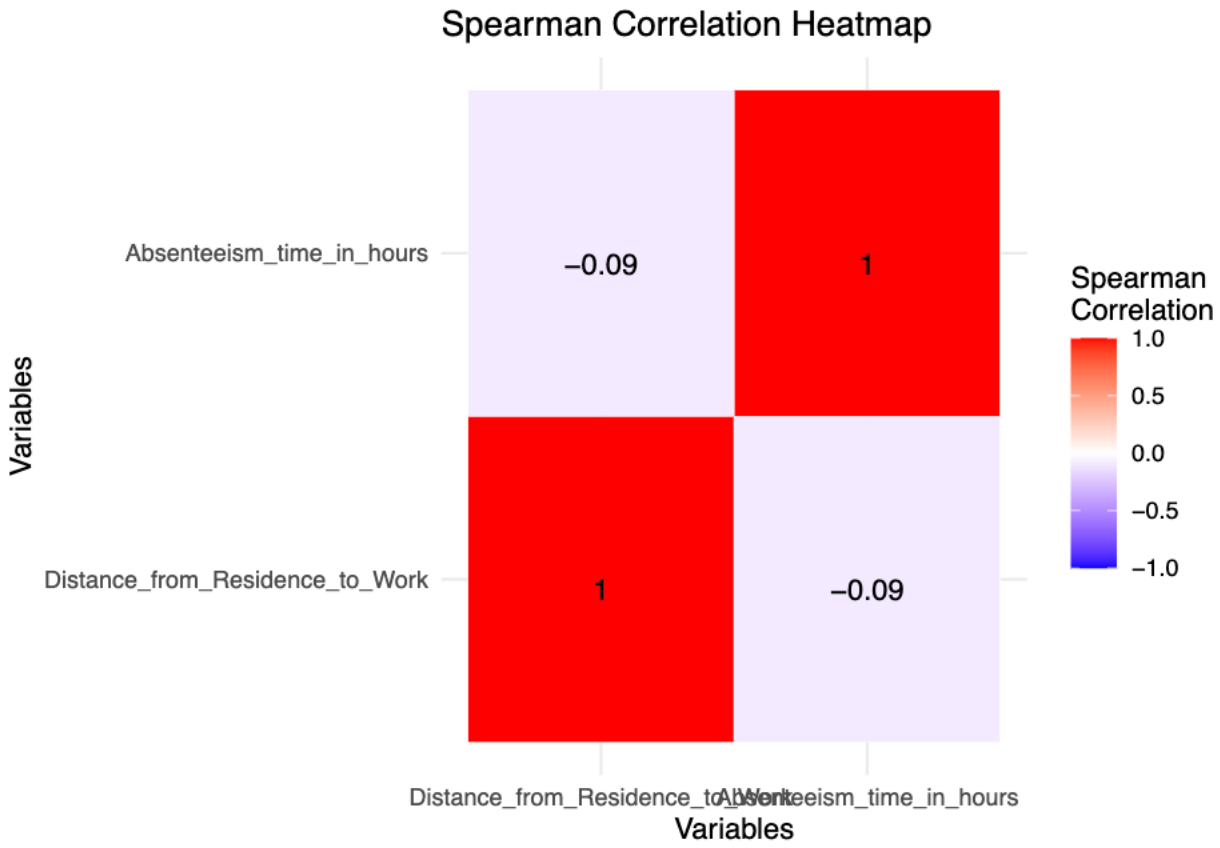
```
plot(Distance_from_Residence_to_Work, Absenteeism_time_in_hours, pch=19, xlab="Distance from residence
```




```r
# Calculate Spearman correlation matrix
correlation <- cor(absent[c('Distance_from_Residence_to_Work', 'Absenteeism_time_in_hours')])

cor_matrix=round(correlation, 5)
cor_matrix
```

```
##                                Distance_from_Residence_to_Work
## Distance_from_Residence_to_Work                         1.00000
## Absenteeism_time_in_hours                              -0.08836
##                                Absenteeism_time_in_hours
## Distance_from_Residence_to_Work                  -0.08836
## Absenteeism_time_in_hours                         1.00000
```

```r
library(ggplot2)

# Melt the correlation matrix for ggplot
library(reshape2)
cor_melted <- melt(cor_matrix)

# Create a small heatmap with values
ggplot(cor_melted, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1,1), space =
  theme_minimal() +
  geom_text(aes(Var1, Var2, label = round(value, 2)), vjust = 1) +
  labs(title = "Spearman Correlation Heatmap",
       x = "Variables",
       y = "Variables")
```



Observation

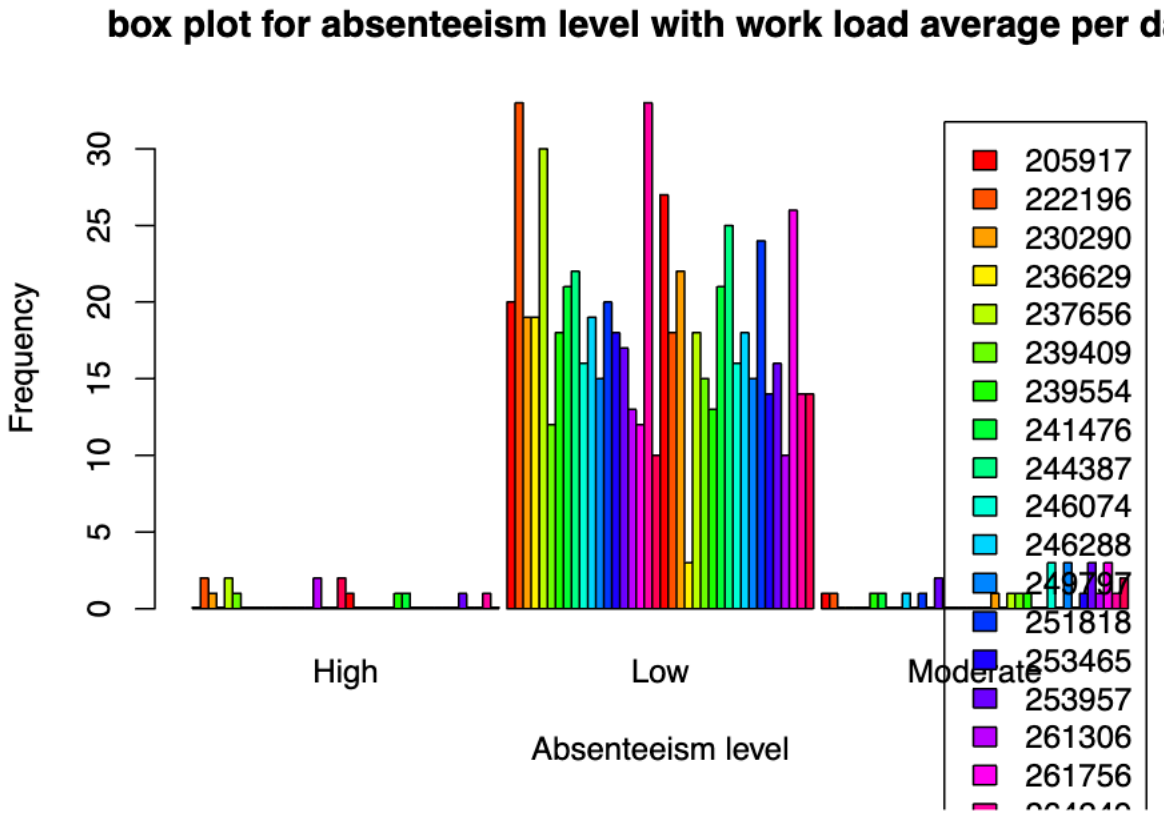
- The correlation value is -0.08836282.
- As it is negative it does indicate an inverse relation between the distance to work and absenteeism.
- HOWEVER, since the value is really close to zero the relationship is not strong and doesn't seem to be significant enough. We can claim a negligible correlation between the two.
- Looking at the scatter plot, it seems the distance from home to work doesn't really show a clear pattern with the absenteeism time of employee. The numbers in the correlation matrix suggest a very weak connection – almost close to no connection at all. So, it's safe to say that how far someone lives from work doesn't seem to be a big reason for employees incresed or decreased absenteeism time.

Question 1.d:

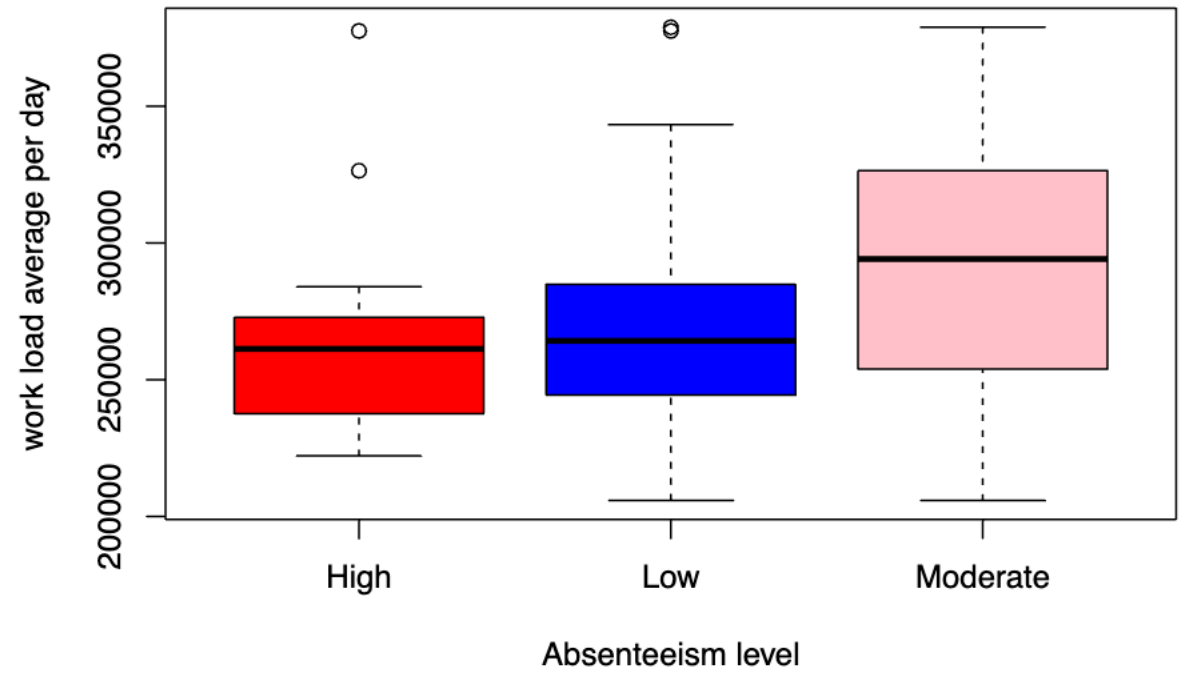
(d) How does the work load average per day relate to absenteeism? Are higher workloads associated with more or less absenteeism?

```
#how to create a 2x2 confusion matrix
count_1 = table(absent$Work_load_Average_in_days, absent$Absenteeism)
n= length(unique(absent$Absenteeism_time_in_hours))

barplot(count_1,
        beside = T,
        legend.text = T,
        xlab = "Absenteeism level",
        ylab = "Frequency",
```

```
        main = "box plot for absenteeism level with work load average per day ",
        col = rainbow(n))
```

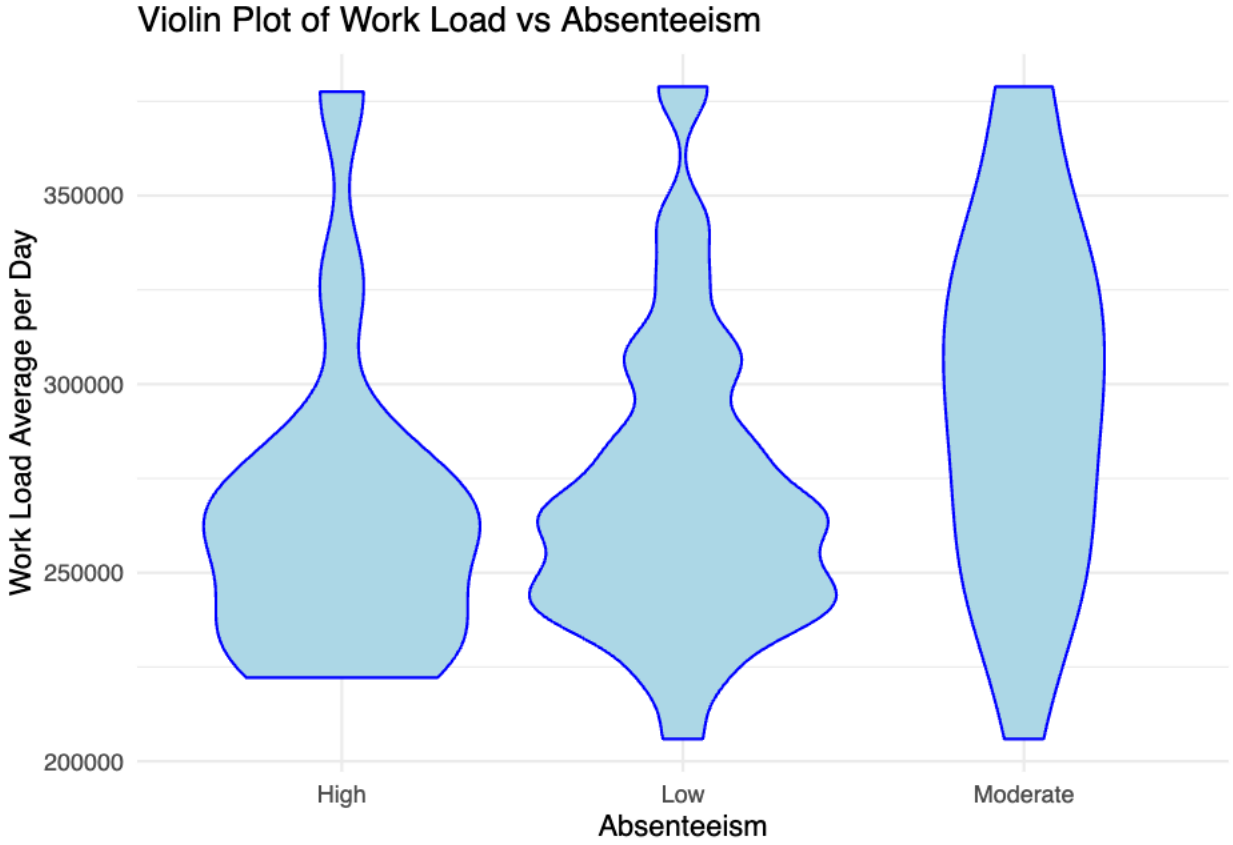



```
boxplot(absent$Work_load_Average_in_days  ~  absent$Absenteeism,ylab="work load average per day",xlab=
```

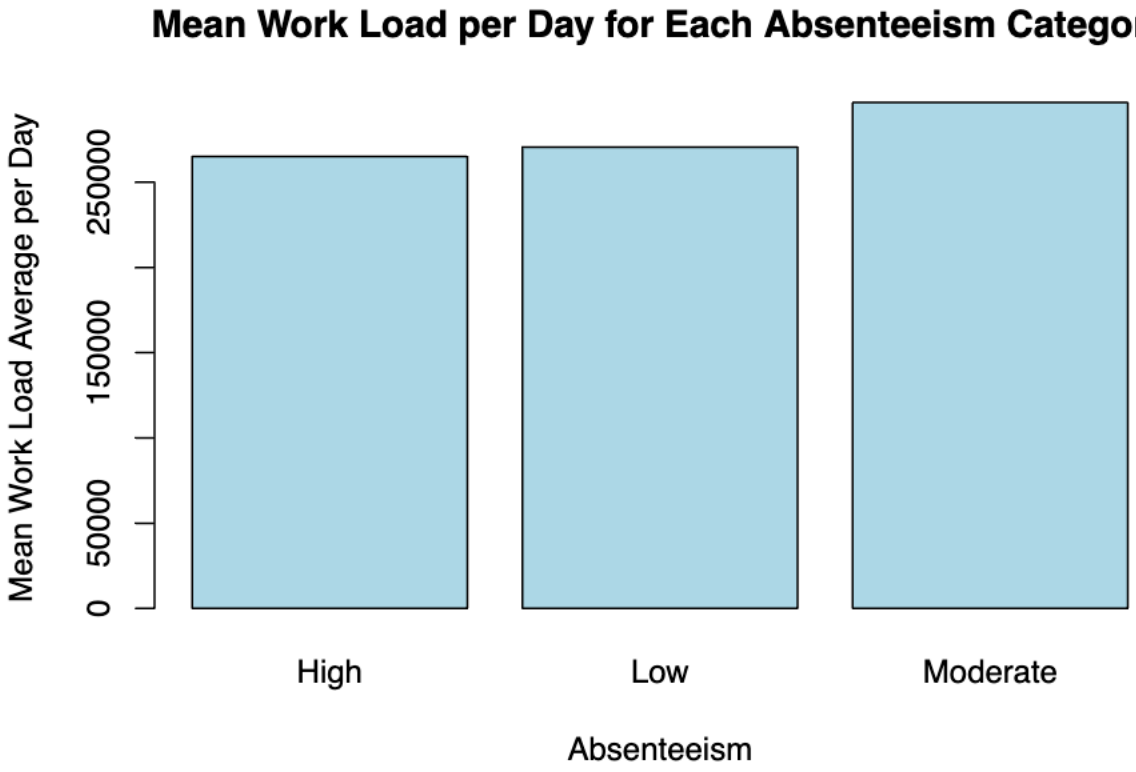



```
# Violin Plot
library(ggplot2)
```

```r
ggplot(absent, aes(x = Absenteeism, y = Work_load_Average_in_days)) +
  geom_violin(fill = "lightblue", color = "blue") +
  labs(title = "Violin Plot of Work Load vs Absenteeism", y = "Work Load Average per Day", x = "Absente
  theme_minimal()
```




```r
# Bar Plot
barplot(tapply(absent$Work_load_Average_in_days, absent$Absenteeism, mean),
        col = "lightblue",
        main = "Mean Work Load per Day for Each Absenteeism Category",
        ylab = "Mean Work Load Average per Day",
        xlab = "Absenteeism")
```




```
# Calculate the correlation between work load average per day and absenteeism time
# Calculate the correlation
correlation_1 <- cor(absent$Work_load_Average_in_days, absent$Absenteeism_time_in_hours, use = "complet
print(correlation_1)
```

```
## [1] 0.0247489
```

Observation

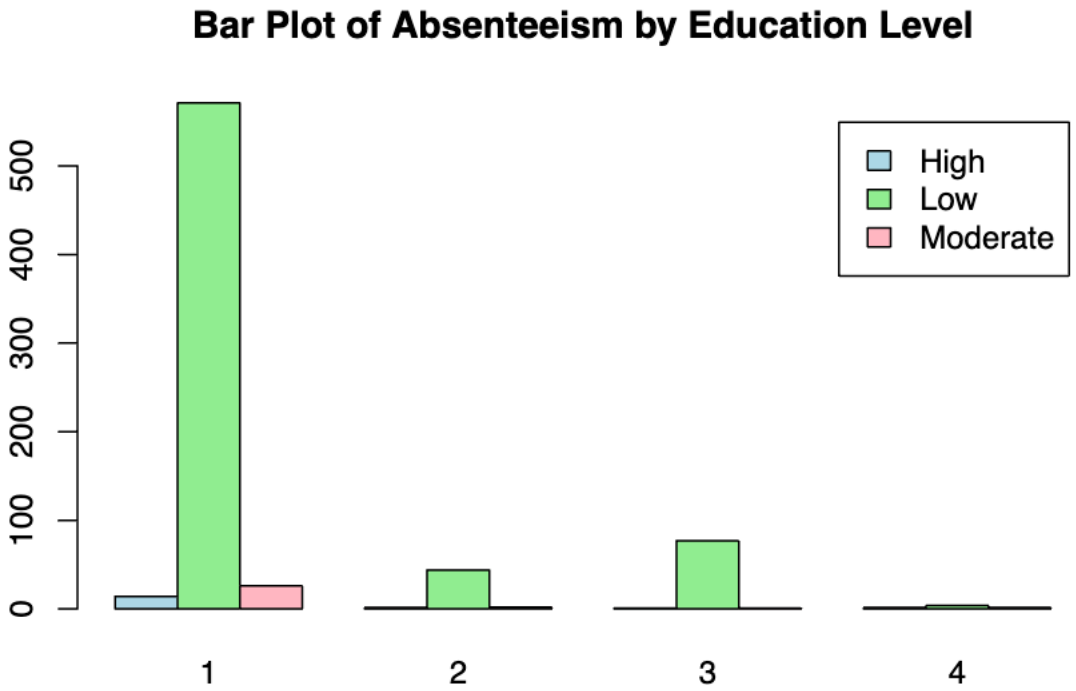
Correlation between workload and absenteeism:

- The correlation value is 0.0247489.
- Since the value is really close to zero the relationship is not strong and doesn't seem to be significant enough. We can claim a negligible correlation between the two.
- Higher workloads might not necessarily result in higher absenteeism, although there is a very slight positive correlation.
- If we take a look at the boxplot, median average workload per day for employee who are absent moderate is high compared to others at the same time the variation in work load average per day is also high for the employee who absent moderate. Since work load average for employee having both high or low absent rate is similar, so this mean there is no linear relation with workload average per day and employee being absent. so We could conclude that workload is not the factors that made the employee absent. Usually higher workload leads to more stress for the employee, but based on this data(violin plot) higher workload does not affecting their mental health.

Question 1.e:

(e) Analyze the absenteeism based on education levels. Do certain education levels correlate with higher or lower absenteeism?

```r
# Bar Plot
barplot(table(absent$Absenteeism, absent$Education),
        col = c("lightblue", "lightgreen", "lightpink"),
        beside = TRUE,
        main = "Bar Plot of Absenteeism by Education Level",
        legend.text = TRUE )
```

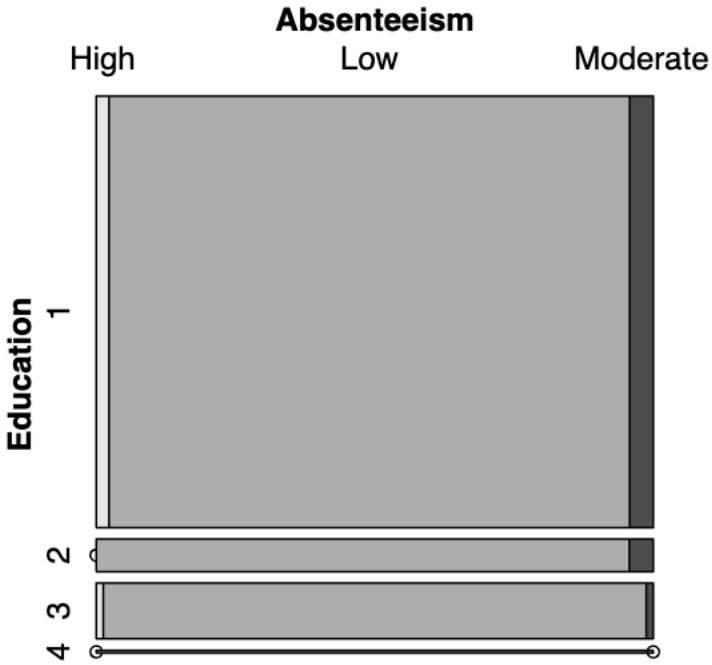



```r
# Mosaic Plot
library(vcd)
```

```
## Loading required package: grid
```

```r
mosaic(Absenteeism ~ Education, data = absent, main = "Mosaic Plot of Absenteeism by Education Level",
```

Mosaic Plot of Absenteeism by Education Level



observation

-Most employees have a high school education, and a larger number of absences are seen in this group. The mosaic plot suggests that there is no linear relationship between education level and absenteeism, it is not the case that higher education make absenteeism lower, looking given data.

- Highest number of absenteeism if for the employee who have only high school of education.

(if Education variable is taken numeric) - The correlation value is -0.04623522.

- As it is negative it does indicate an inverse relation between the Education Level and absenteeism.
- HOWEVER, since the value is really close to zero the relationship is not strong and doesn't seem to be significant enough. We can claim a negligible correlation between the two.

Question 1.f:

(f) Which variables show the strongest correlation with absenteeism time in hours? How might these influence your logistic regression model?

```
# Select only numeric variables
numeric_absent <- absent[, sapply(absent, is.numeric)]

# Calculate correlation matrix among numeric variables
cor_matrix <- cor(numeric_absent)

# Display the correlation matrix
print(cor_matrix)
```

```
##                                      ID Reason for absence
## ID                         1.000000000      -0.0642446728
```

```
## Reason for absence              -0.064244673       1.0000000000
## Transportation_expense          -0.224162785      -0.1193814506
## Distance_from_Residence_to_Work -0.486160317       0.1618307843
## Service_time                    -0.272703782       0.0484251154
## Age                              0.040899097      -0.0786080013
## Work_load_Average_in_days        0.092456917      -0.1234723206
## Hit_ target                      0.018789221       0.0889431495
## Disciplinary_failure             0.004502156      -0.5450539489
## Son                              0.002766785      -0.0553639437
## Social_drinker                  -0.451337911       0.0654408504
## Social_smoker                   -0.010825827      -0.1157016863
## Pet                             -0.041417969      -0.0559109724
## Weight                          -0.254221676      -0.0002692878
## Height                           0.076363379      -0.0792670646
## Body_mass_index                 -0.306924255       0.0372049359
## Absenteeism_time_in_hours       -0.017996594      -0.1731155615
##                                 Transportation_expense
## ID                                         -0.224162785
## Reason for absence                         -0.119381451
## Transportation_expense                      1.000000000
## Distance_from_Residence_to_Work             0.262183111
## Service_time                               -0.349887036
## Age                                        -0.227542434
## Work_load_Average_in_days                   0.005438065
## Hit_ target                                -0.080193040
## Disciplinary_failure                        0.109221690
## Son                                         0.383001191
## Social_drinker                              0.145117454
## Social_smoker                               0.044356144
## Pet                                         0.400080301
## Weight                                     -0.207434941
## Height                                     -0.194495956
## Body_mass_index                            -0.136516573
## Absenteeism_time_in_hours                   0.027584631
##                                 Distance_from_Residence_to_Work  Service_time
## ID                                                  -0.48616032 -0.2727037819
## Reason for absence                                   0.16183078  0.0484251154
## Transportation_expense                               0.26218311 -0.3498870362
## Distance_from_Residence_to_Work                      1.00000000  0.1317303037
## Service_time                                         0.13173030  1.0000000000
## Age                                                 -0.14588637  0.6709789169
## Work_load_Average_in_days                           -0.06867696 -0.0006684910
## Hit_ target                                         -0.01386463 -0.0078400347
## Disciplinary_failure                                -0.05652710 -0.0002211603
## Son                                                  0.05423039 -0.0471284116
## Social_drinker                                       0.45219570  0.3531406086
## Social_smoker                                       -0.07536879  0.0724243087
## Pet                                                  0.20594058 -0.4403006671
## Weight                                              -0.04785909  0.4559748045
## Height                                              -0.35337218 -0.0531345133
## Body_mass_index                                      0.11377164  0.4997179504
## Absenteeism_time_in_hours                           -0.08836282  0.0190292614
##                                        Age Work_load_Average_in_days
## ID                                0.04089910               0.092456917
```

```
## Reason for absence              -0.07860800              -0.123472321
## Transportation_expense          -0.22754243               0.005438065
## Distance_from_Residence_to_Work -0.14588637              -0.068676958
## Service_time                     0.67097892              -0.000668491
## Age                              1.00000000              -0.039425176
## Work_load_Average_in_days       -0.03942518               1.000000000
## Hit_ target                     -0.03922431              -0.089444956
## Disciplinary_failure             0.10430385               0.029025751
## Son                              0.05698412               0.027820236
## Social_drinker                   0.21318262              -0.033712619
## Social_smoker                    0.12173839               0.030968324
## Pet                             -0.23122600               0.007114198
## Weight                           0.41873046              -0.038521570
## Height                          -0.06299658               0.103314799
## Body_mass_index                  0.47068802              -0.090709281
## Absenteeism_time_in_hours        0.06575970               0.024748900
##                                  Hit_ target Disciplinary_failure          Son
## ID                               0.018789221         0.0045021559  0.002766785
## Reason for absence               0.088943150        -0.5450539489 -0.055363944
## Transportation_expense          -0.080193040         0.1092216899  0.383001191
## Distance_from_Residence_to_Work -0.013864625        -0.0565270951  0.054230393
## Service_time                    -0.007840035        -0.0002211603 -0.047128412
## Age                             -0.039224314         0.1043038505  0.056984121
## Work_load_Average_in_days       -0.089444956         0.0290257512  0.027820236
## Hit_ target                      1.000000000        -0.1479708344 -0.014090600
## Disciplinary_failure            -0.147970834         1.0000000000  0.072096347
## Son                             -0.014090600         0.0720963471  1.000000000
## Social_drinker                  -0.102479904         0.0518380278  0.206375968
## Social_smoker                    0.051253895         0.1167481188  0.156087556
## Pet                              0.007201276         0.0188813529  0.108917279
## Weight                          -0.044947425         0.0722245804 -0.139552437
## Height                           0.093266929        -0.0104983751 -0.014208322
## Body_mass_index                 -0.088939452         0.0794281798 -0.144150249
## Absenteeism_time_in_hours        0.026695065        -0.1242479841  0.113756496
##                                 Social_drinker Social_smoker          Pet
## ID                                -0.45133791  -0.010825827 -0.041417969
## Reason for absence                 0.06544085  -0.115701686 -0.055910972
## Transportation_expense             0.14511745   0.044356144  0.400080301
## Distance_from_Residence_to_Work    0.45219570  -0.075368792  0.205940581
## Service_time                       0.35314061   0.072424309 -0.440300667
## Age                                0.21318262   0.121738391 -0.231225999
## Work_load_Average_in_days         -0.03371262   0.030968324  0.007114198
## Hit_ target                       -0.10247990   0.051253895  0.007201276
## Disciplinary_failure               0.05183803   0.116748119  0.018881353
## Son                                0.20637597   0.156087556  0.108917279
## Social_drinker                     1.00000000  -0.111678005 -0.122780216
## Social_smoker                     -0.11167800   1.000000000  0.105378757
## Pet                               -0.12278022   0.105378757  1.000000000
## Weight                             0.37866361  -0.198511210 -0.103770410
## Height                             0.16995108   0.003271391 -0.103143350
## Body_mass_index                    0.32397779  -0.196006483 -0.076102946
## Absenteeism_time_in_hours          0.06506734  -0.008936423 -0.028276589
##                                       Weight       Height Body_mass_index
## ID                               -0.2542216757  0.076363379     -0.30692425
```

```
## Reason for absence              -0.0002692878 -0.079267065      0.03720494
## Transportation_expense          -0.2074349415 -0.194495956     -0.13651657
## Distance_from_Residence_to_Work -0.0478590935 -0.353372180      0.11377164
## Service_time                     0.4559748045 -0.053134513      0.49971795
## Age                              0.4187304592 -0.062996583      0.47068802
## Work_load_Average_in_days       -0.0385215697  0.103314799     -0.09070928
## Hit_ target                     -0.0449474247  0.093266929     -0.08893945
## Disciplinary_failure             0.0722245804 -0.010498375      0.07942818
## Son                             -0.1395524370 -0.014208322     -0.14415025
## Social_drinker                   0.3786636067  0.169951078      0.32397779
## Social_smoker                   -0.1985112098  0.003271391     -0.19600648
## Pet                             -0.1037704099 -0.103143350     -0.07610295
## Weight                           1.0000000000  0.306801835      0.90411690
## Height                           0.3068018348  1.000000000     -0.12104878
## Body_mass_index                  0.9041169006 -0.121048777      1.00000000
## Absenteeism_time_in_hours        0.0157891765  0.144420476     -0.04971948
##                                 Absenteeism_time_in_hours
## ID                                          -0.017996594
## Reason for absence                          -0.173115561
## Transportation_expense                       0.027584631
## Distance_from_Residence_to_Work             -0.088362822
## Service_time                                 0.019029261
## Age                                          0.065759701
## Work_load_Average_in_days                    0.024748900
## Hit_ target                                  0.026695065
## Disciplinary_failure                        -0.124247984
## Son                                          0.113756496
## Social_drinker                               0.065067343
## Social_smoker                               -0.008936423
## Pet                                         -0.028276589
## Weight                                       0.015789177
## Height                                       0.144420476
## Body_mass_index                             -0.049719479
## Absenteeism_time_in_hours                    1.000000000
```

Observation

- These are the variables which shows comparatively more strong relation then others to the variable Absenteeism_time_in_hours

1) Reason for absence (-)
2) Day_of_the_week (-)
3) Disciplinary_failure (-)
4) Son (+)
5) Height (+)

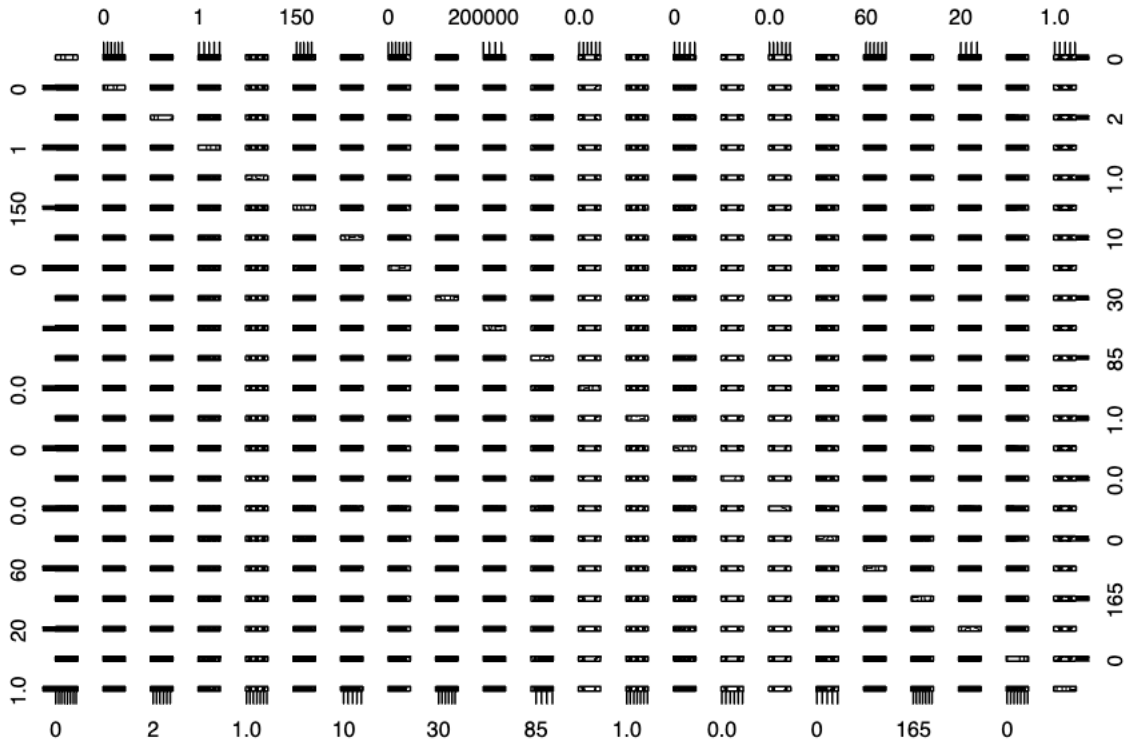
-Most of these variables are not significant for our model but they do have as given above correlation with our Absenteeism_time_in_hours variable

- Height seems to have the strongest positive correlation with absenteeism and reason for absence seems to have the strongest negative correlation.
- Secondly, Son also have a strong positive correlation indicating that people with children seem to have higher absenteeism.
- Variables with stronger correlations are highly relevant as they have a higher impact on the performance of the logistic regression model.

Question 1.g:

(g) Are there any unexpected correlations or findings that challenge common assumptions about workplace absenteeism?

```
pairs(absent)
```




```
# Select only numeric variables
numeric_absent <- absent[, sapply(absent, is.numeric)]

# Calculate correlation matrix among numeric variables
cor_matrix <- cor(numeric_absent)

# Display the correlation matrix
print(cor_matrix)
```

```
##                                          ID Reason for absence
## ID                              1.000000000      -0.0642446728
## Reason for absence             -0.064244673       1.0000000000
## Transportation_expense         -0.224162785      -0.1193814506
## Distance_from_Residence_to_Work -0.486160317       0.1618307843
## Service_time                   -0.272703782       0.0484251154
## Age                             0.040899097      -0.0786080013
## Work_load_Average_in_days       0.092456917      -0.1234723206
## Hit_ target                     0.018789221       0.0889431495
## Disciplinary_failure            0.004502156      -0.5450539489
## Son                             0.002766785      -0.0553639437
## Social_drinker                 -0.451337911       0.0654408504
## Social_smoker                  -0.010825827      -0.1157016863
## Pet                            -0.041417969      -0.0559109724
## Weight                         -0.254221676      -0.0002692878
## Height                          0.076363379      -0.0792670646
```

```
## Body_mass_index                -0.306924255       0.0372049359
## Absenteeism_time_in_hours      -0.017996594      -0.1731155615
##                               Transportation_expense
## ID                                    -0.224162785
## Reason for absence                    -0.119381451
## Transportation_expense                 1.000000000
## Distance_from_Residence_to_Work        0.262183111
## Service_time                          -0.349887036
## Age                                   -0.227542434
## Work_load_Average_in_days              0.005438065
## Hit_ target                           -0.080193040
## Disciplinary_failure                   0.109221690
## Son                                    0.383001191
## Social_drinker                         0.145117454
## Social_smoker                          0.044356144
## Pet                                    0.400080301
## Weight                                -0.207434941
## Height                                -0.194495956
## Body_mass_index                       -0.136516573
## Absenteeism_time_in_hours              0.027584631
##                               Distance_from_Residence_to_Work  Service_time
## ID                                                -0.48616032 -0.2727037819
## Reason for absence                                 0.16183078  0.0484251154
## Transportation_expense                             0.26218311 -0.3498870362
## Distance_from_Residence_to_Work                    1.00000000  0.1317303037
## Service_time                                       0.13173030  1.0000000000
## Age                                               -0.14588637  0.6709789169
## Work_load_Average_in_days                         -0.06867696 -0.0006684910
## Hit_ target                                       -0.01386463 -0.0078400347
## Disciplinary_failure                              -0.05652710 -0.0002211603
## Son                                                0.05423039 -0.0471284116
## Social_drinker                                     0.45219570  0.3531406086
## Social_smoker                                     -0.07536879  0.0724243087
## Pet                                                0.20594058 -0.4403006671
## Weight                                            -0.04785909  0.4559748045
## Height                                            -0.35337218 -0.0531345133
## Body_mass_index                                    0.11377164  0.4997179504
## Absenteeism_time_in_hours                         -0.08836282  0.0190292614
##                                      Age Work_load_Average_in_days
## ID                            0.04089910                0.092456917
## Reason for absence           -0.07860800               -0.123472321
## Transportation_expense       -0.22754243                0.005438065
## Distance_from_Residence_to_Work -0.14588637             -0.068676958
## Service_time                  0.67097892               -0.000668491
## Age                           1.00000000               -0.039425176
## Work_load_Average_in_days    -0.03942518                1.000000000
## Hit_ target                  -0.03922431               -0.089444956
## Disciplinary_failure          0.10430385                0.029025751
## Son                           0.05698412                0.027820236
## Social_drinker                0.21318262               -0.033712619
## Social_smoker                 0.12173839                0.030968324
## Pet                          -0.23122600                0.007114198
## Weight                        0.41873046               -0.038521570
## Height                       -0.06299658                0.103314799
```

```
## Body_mass_index                  0.47068802              -0.090709281
## Absenteeism_time_in_hours        0.06575970               0.024748900
##                                  Hit_ target Disciplinary_failure          Son
## ID                               0.018789221         0.0045021559  0.002766785
## Reason for absence               0.088943150        -0.5450539489 -0.055363944
## Transportation_expense          -0.080193040         0.1092216899  0.383001191
## Distance_from_Residence_to_Work -0.013864625        -0.0565270951  0.054230393
## Service_time                    -0.007840035        -0.0002211603 -0.047128412
## Age                             -0.039224314         0.1043038505  0.056984121
## Work_load_Average_in_days       -0.089444956         0.0290257512  0.027820236
## Hit_ target                      1.000000000        -0.1479708344 -0.014090600
## Disciplinary_failure            -0.147970834         1.0000000000  0.072096347
## Son                             -0.014090600         0.0720963471  1.000000000
## Social_drinker                  -0.102479904         0.0518380278  0.206375968
## Social_smoker                    0.051253895         0.1167481188  0.156087556
## Pet                              0.007201276         0.0188813529  0.108917279
## Weight                          -0.044947425         0.0722245804 -0.139552437
## Height                           0.093266929        -0.0104983751 -0.014208322
## Body_mass_index                 -0.088939452         0.0794281798 -0.144150249
## Absenteeism_time_in_hours        0.026695065        -0.1242479841  0.113756496
##                                 Social_drinker Social_smoker          Pet
## ID                                -0.45133791  -0.010825827 -0.041417969
## Reason for absence                 0.06544085  -0.115701686 -0.055910972
## Transportation_expense             0.14511745   0.044356144  0.400080301
## Distance_from_Residence_to_Work    0.45219570  -0.075368792  0.205940581
## Service_time                       0.35314061   0.072424309 -0.440300667
## Age                                0.21318262   0.121738391 -0.231225999
## Work_load_Average_in_days         -0.03371262   0.030968324  0.007114198
## Hit_ target                       -0.10247990   0.051253895  0.007201276
## Disciplinary_failure               0.05183803   0.116748119  0.018881353
## Son                                0.20637597   0.156087556  0.108917279
## Social_drinker                     1.00000000  -0.111678005 -0.122780216
## Social_smoker                     -0.11167800   1.000000000  0.105378757
## Pet                               -0.12278022   0.105378757  1.000000000
## Weight                             0.37866361  -0.198511210 -0.103770410
## Height                             0.16995108   0.003271391 -0.103143350
## Body_mass_index                    0.32397779  -0.196006483 -0.076102946
## Absenteeism_time_in_hours          0.06506734  -0.008936423 -0.028276589
##                                       Weight       Height Body_mass_index
## ID                               -0.2542216757  0.076363379     -0.30692425
## Reason for absence               -0.0002692878 -0.079267065      0.03720494
## Transportation_expense           -0.2074349415 -0.194495956     -0.13651657
## Distance_from_Residence_to_Work  -0.0478590935 -0.353372180      0.11377164
## Service_time                      0.4559748045 -0.053134513      0.49971795
## Age                               0.4187304592 -0.062996583      0.47068802
## Work_load_Average_in_days        -0.0385215697  0.103314799     -0.09070928
## Hit_ target                      -0.0449474247  0.093266929     -0.08893945
## Disciplinary_failure              0.0722245804 -0.010498375      0.07942818
## Son                              -0.1395524370 -0.014208322     -0.14415025
## Social_drinker                    0.3786636067  0.169951078      0.32397779
## Social_smoker                    -0.1985112098  0.003271391     -0.19600648
## Pet                              -0.1037704099 -0.103143350     -0.07610295
## Weight                            1.0000000000  0.306801835      0.90411690
## Height                            0.3068018348  1.000000000     -0.12104878
```

```
## Body_mass_index                  0.9041169006 -0.121048777      1.00000000
## Absenteeism_time_in_hours        0.0157891765  0.144420476     -0.04971948
##                                 Absenteeism_time_in_hours
## ID                                          -0.017996594
## Reason for absence                          -0.173115561
## Transportation_expense                       0.027584631
## Distance_from_Residence_to_Work             -0.088362822
## Service_time                                 0.019029261
## Age                                          0.065759701
## Work_load_Average_in_days                    0.024748900
## Hit_ target                                  0.026695065
## Disciplinary_failure                        -0.124247984
## Son                                          0.113756496
## Social_drinker                               0.065067343
## Social_smoker                               -0.008936423
## Pet                                         -0.028276589
## Weight                                       0.015789177
## Height                                       0.144420476
## Body_mass_index                             -0.049719479
## Absenteeism_time_in_hours                    1.000000000
```

Observation

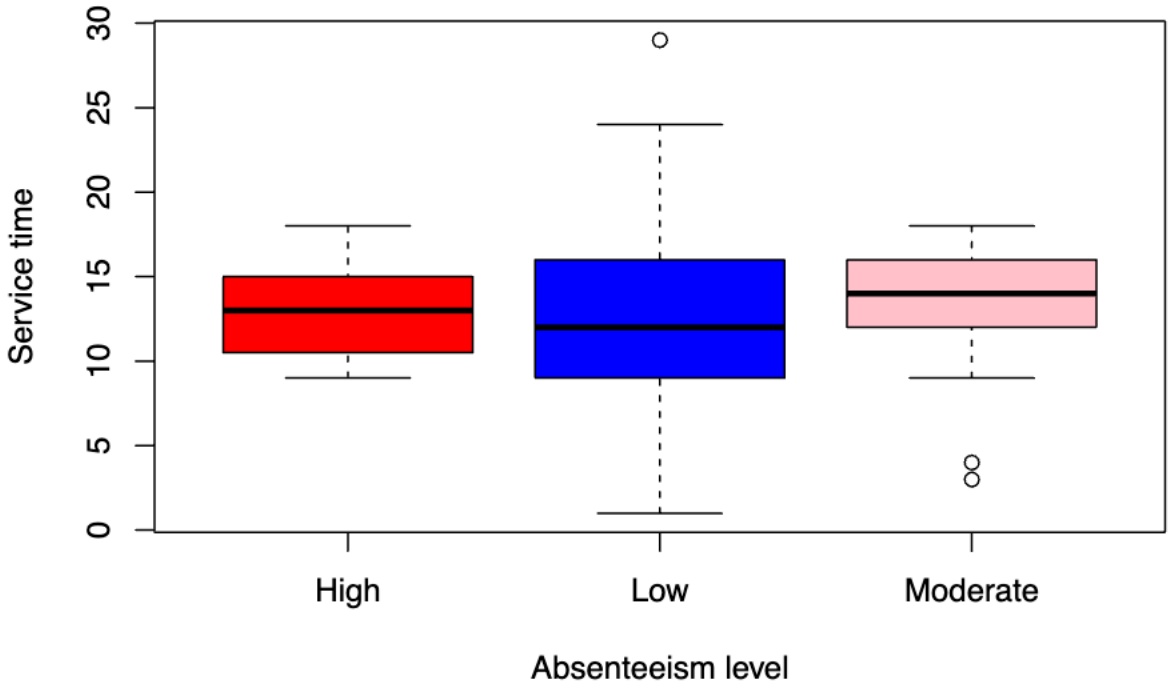
Unexpected Correlations:

- Height was the most surprising strong correlation. Physical factors like heights go against the common workplace assumptions as it has nothing to do with how one performs.
- Seasons having a very weak correlation was quite unexpected as well. One would assume that the heavy rainy season would result in more absenteeism or even the winter causing people to stay at home.
- Social_drinker having a positive correlation also goes against the common assumption that it is harmless and doesn't affect one's work.
- One important findings that challenge common assumptions about workplace absenteeism is that the distance from the workplace does not matter a lot for the absenteeism, which is not what I was accepting but clearly data show a very weak negative correlation between these two variables, which was interesting and surprising for me to know.

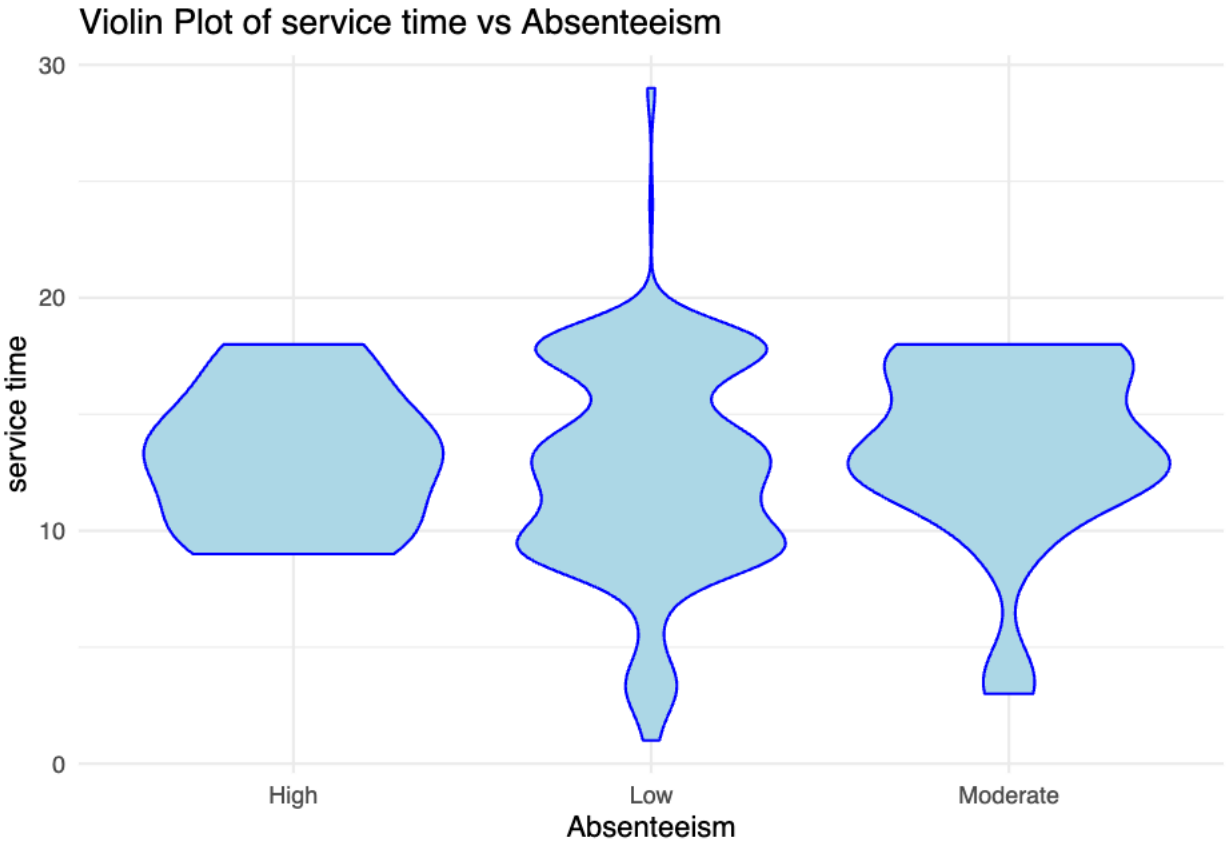
Question 1.h:

(h) Does service time (duration of service in the company) have any impact on the absenteeism rate?

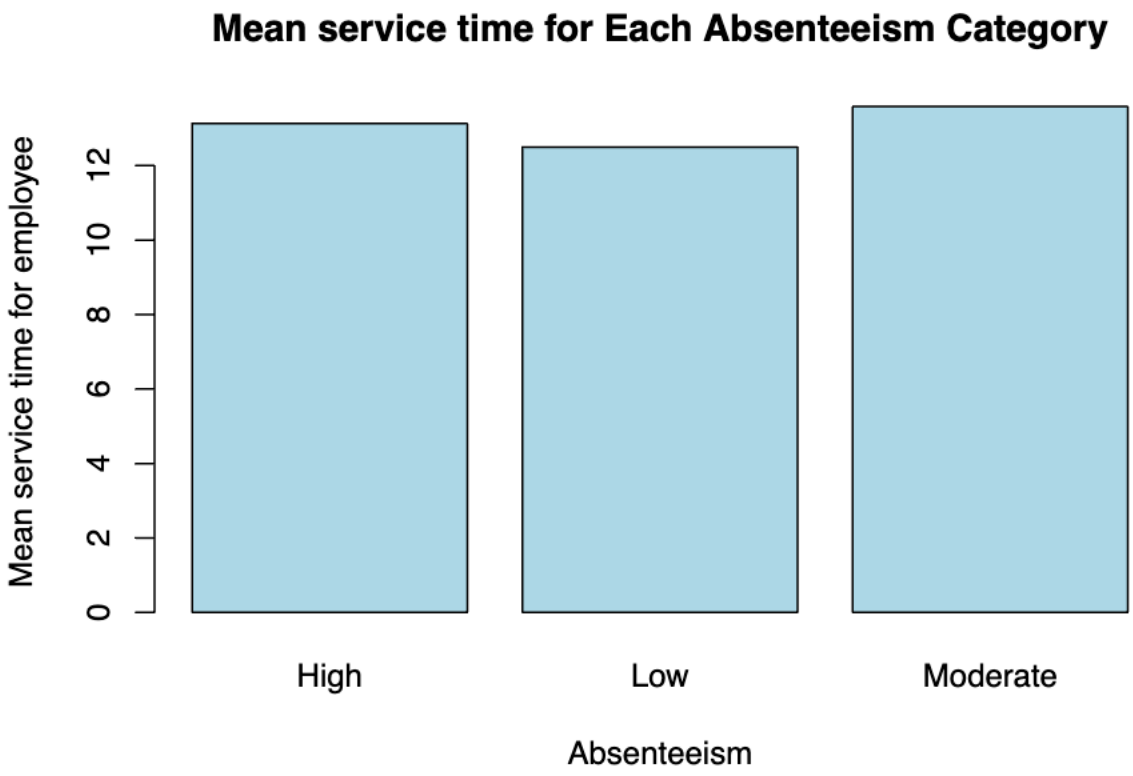
```
boxplot(Service_time  ~   Absenteeism,ylab="Service time",xlab="Absenteeism level", col=c("Red","blue",
```




```r
# Violin Plot
library(ggplot2)
ggplot(absent, aes(x = Absenteeism, y = Service_time)) +
  geom_violin(fill = "lightblue", color = "blue") +
  labs(title = "Violin Plot of service time vs Absenteeism", y = "service time", x = "Absenteeism") +
  theme_minimal()
```




```
# Bar Plot
barplot(tapply(Service_time, Absenteeism, mean),
        col = "lightblue",
        main = "Mean service time for Each Absenteeism Category",
        ylab = "Mean service time for employee",
        xlab = "Absenteeism")
```




```
# Calculate the correlation between service time and absenteeism time in hours
correlation_3 <- cor(absent$Service_time, absent$Absenteeism_time_in_hours, use = "complete.obs")
print(correlation_3)
```

```
## [1] 0.01902926
```

Observation

Correlation between Service Time and absenteeism:

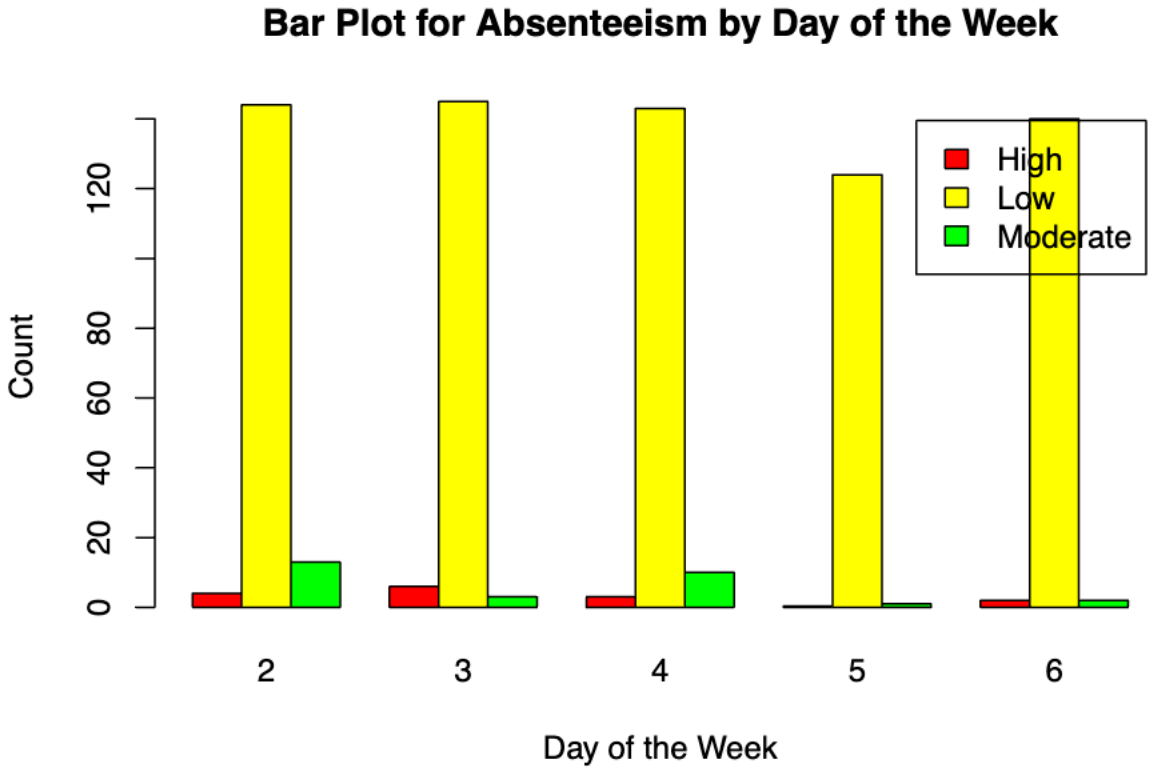
- The correlation value is 0.01902926.
- One might assume that those who have worked longer are more loyal/disciplined and would have lower absenteeism, however Service time has a positive correlation.
- Since the value is really close to zero the relationship is not strong and doesn't seem to be significant enough. We can claim a negligible correlation between the two.
- From above plots, Mean service time is highest among the employee who take moderate absent from work. It seems that service time does not have linear relationship with absenteeism level or have weak negative linear relationship.

Question 1.i:

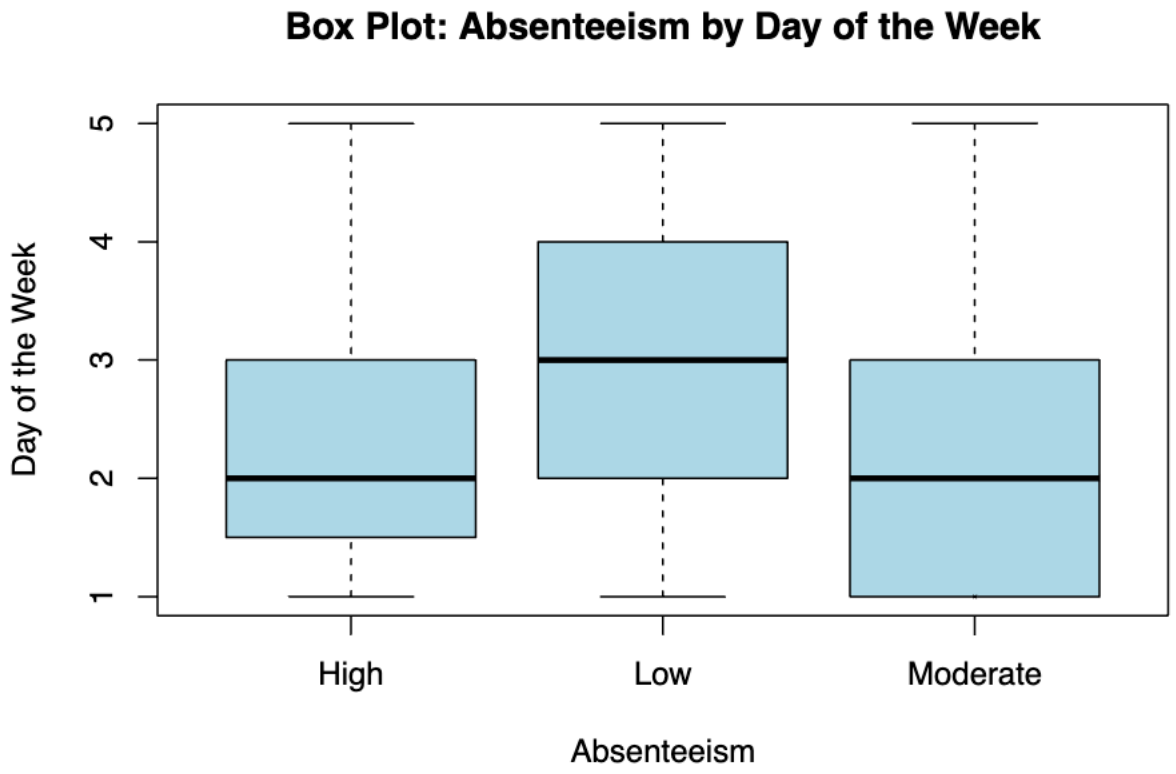
(i) Examine if day of the week has any influence on absenteeism – are certain days more prone to absences?

```
barplot(table(Absenteeism,Day_of_the_week ),
        beside = TRUE,
        legend.text = TRUE,
        col=c("red", "yellow", "green"),
        main = "Bar Plot for Absenteeism by Day of the Week",
```

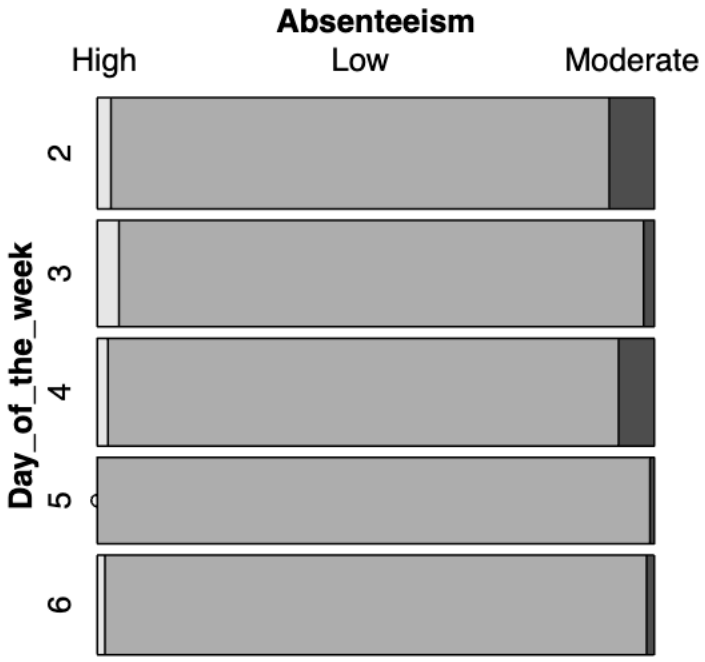
```
        xlab = "Day of the Week",
        ylab = "Count")
```




```
boxplot(Day_of_the_week ~ Absenteeism,
        col = "lightblue",
        main = "Box Plot: Absenteeism by Day of the Week",
        ylab = "Day of the Week",
        xlab = "Absenteeism")
```




```r
# Mosaic Plot
library(vcd)
mosaic(Absenteeism ~ Day_of_the_week, data = absent, main = "Mosaic Plot of Absenteeism by Days of the
```



Observation

- There seems to be a higher absenteeism on average on the 2nd day of the week.
- Yes from the data it seems that days of the weeks does have impact in the absenteeism, we can see in Tuesday, high absenteeism but, no high absenteeism is seen in Thursday but again we can see some high absenteeism in Friday indicating non linear relationship. On Monday a lot of employee had moderate absenteeism compared to other days.
- Least absenteeism happens in the 5th day of the week i.e in Thursday. Employee seems to absent less in Thursday indicating may be it is almost end of the week and need to work more during that period or may be they have scheduled all their important meeting in the Thursday (may be).

Question 1.j:

(j) Identify any outliers in the data set. What could be the reasons for these anomalies, and how might they affect the analysis?

```
outlier_check <- function(data) {
  data <- sort(data)
  q <- quantile(data, c(0.25, 0.75))
  iqr <- q[2] - q[1]
  lower_fence <- q[1] - (1.5 * iqr)
  upper_fence <- q[2] + (1.5 * iqr)

  outliers <- data[data < lower_fence | data > upper_fence]
  return(outliers)
}
```

```
outlier_check(absent$Transportation_expense)
```

```
## [1] 388 388 388
```

```
outlier_check(absent$Service_time)
```

```
## [1] 29 29 29 29 29
```

```
outlier_check(absent$Age)
```

```
## [1] 58 58 58 58 58 58 58 58
```

```
outlier_check(absent$`Hit_ target`)
```

```
##  [1] 81 81 81 81 81 81 81 81 81 81 81 81 81 81 81 81 81 81 81 81
```

```
outlier_check(absent$Disciplinary_failure)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [39] 1 1
```

```
outlier_check(absent$Height)
```

```
##   [1] 163 163 163 163 163 163 178 178 178 178 178 178 178 178 178 178 178 178 178
##  [19] 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178
##  [37] 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178 178
##  [55] 178 178 178 178 178 178 178 178 178 182 182 182 182 182 182 182 182 182 182
##  [73] 182 182 182 182 182 182 182 182 182 182 182 185 185 185 185 185 185 185 185
##  [91] 196 196 196 196 196 196 196 196 196 196 196 196 196 196 196 196 196 196 196
## [109] 196 196 196 196 196 196 196 196 196 196 196
```

```r
outlier_check(absent$Absenteeism_time_in_hours)
```

```
##  [1]  24  24  24  24  24  24  24  24  24  24  24  24  24  24  24  24  32  32  32
## [20]  32  32  32  40  40  40  40  40  40  40  48  56  56  64  64  64  80  80  80
## [39] 104 112 112 120 120 120
```

```r
# Select numeric columns from the dataset
numeric_cols <- sapply(absent, is.numeric)

# Extract the numeric part of the dataset
numeric_data <- absent[, numeric_cols]

# Calculate Mahalanobis distances for numeric variables
mean_vector <- colMeans(numeric_data)
cov_matrix <- cov(numeric_data)
m_dis <- mahalanobis(numeric_data, center = mean_vector, cov = cov_matrix)

# Set a threshold for Mahalanobis distance to identify outliers
threshold <- qchisq(0.995, sum(numeric_cols))

# Identify joint outliers in the original dataset
joint_outliers <- absent[m_dis > threshold, ]

# Display the joint outliers
joint_outliers
```

```
## # A tibble: 62 x 22
##       ID `Reason for absence` Month_of_absence Day_of_the_week Seasons
##    <dbl>                <dbl> <fct>            <fct>           <fct>
##  1     7                    7 7                5               1
##  2    30                   28 8                4               1
##  3     2                   18 8                5               1
##  4     2                   18 8                2               1
##  5     2                   28 8                6               1
##  6    13                    0 9                4               4
##  7    31                   11 2                2               2
##  8    31                    1 2                3               2
##  9    30                   19 3                3               2
## 10     2                    0 4                2               3
## # i 52 more rows
## # i 17 more variables: Transportation_expense <dbl>,
## #   Distance_from_Residence_to_Work <dbl>, Service_time <dbl>, Age <dbl>,
## #   Work_load_Average_in_days <dbl>, `Hit_ target` <dbl>,
## #   Disciplinary_failure <dbl>, Education <fct>, Son <dbl>,
## #   Social_drinker <dbl>, Social_smoker <dbl>, Pet <dbl>, Weight <dbl>,
## #   Height <dbl>, Body_mass_index <dbl>, Absenteeism_time_in_hours <dbl>, ...
```

Observation

- We do have outlier in few variables if taken individually. These outlier may disturb the model and make our model prediction incorrect or miss-leading.
- If we do multivariate data analysis and look the data jointly then we can see there are 60 observation out of 740, which act as outlier which is almost 8% of our dataset.

Logistic Regression Analysis Instructions

Question a

(a) Build a logistic regression model using the recoded absenteeism categories (Low, Moderate, High) as the response variable. Ensure that categorical variables are appropriately handled.

```
library(stats)
library(dplyr)
library(nnet)

# Fit the model without including Absenteeism_time_in_hours as a predictor
model <- multinom(Absenteeism ~ . -Absenteeism_time_in_hours, data = absent, trace = FALSE)
```

```
coef(model)
```

```
##          (Intercept)          ID `Reason for absence` Month_of_absence1
## Low        -42.57712 -0.05416046           0.14917324          -2.28031
## Moderate    18.11326 -0.06112384           0.01454161          13.25420
##          Month_of_absence2 Month_of_absence3 Month_of_absence4
## Low               1.846773        -10.312036        -10.974013
## Moderate         16.025695          4.879751          4.794439
##          Month_of_absence5 Month_of_absence6 Month_of_absence7
## Low              -9.573949        -11.348082        -19.498305
## Moderate          4.737455          3.888736         -3.878796
##          Month_of_absence8 Month_of_absence9 Month_of_absence10
## Low              -5.186395         -4.789579         -10.956382
## Moderate         10.342054          9.880989          -7.278377
##          Month_of_absence11 Month_of_absence12 Day_of_the_week3
## Low              -10.295057         -12.778949       -0.7121388
## Moderate           4.156168           1.795961       -2.0451808
##          Day_of_the_week4 Day_of_the_week5 Day_of_the_week6  Seasons2  Seasons3
## Low             0.4593683         8.195485         0.401326 -8.560371 -9.657234
## Moderate        0.1460565         5.704492        -1.379872 -7.083600 -8.029441
##           Seasons4 Transportation_expense Distance_from_Residence_to_Work
## Low      -8.837197            -0.01230570                      0.07844500
## Moderate -6.730758            -0.01005798                      0.07842651
##          Service_time        Age Work_load_Average_in_days `Hit_ target`
## Low         0.2860973 -0.1124851              4.604454e-06    -0.1305756
## Moderate    0.3174566 -0.1330298              7.121480e-06    -0.2654964
##          Disciplinary_failure Education2 Education3 Education4        Son
## Low                11.3232964   19.28550  0.7974196 10.1411790 -0.6592586
## Moderate            0.6581683   20.49184  1.0153700 -0.3689871 -0.7510511
##          Social_drinker Social_smoker       Pet     Weight     Height
## Low          0.02379041     -0.559771 0.5769838 -0.6191323 0.44056965
## Moderate     2.42773533     -1.120573 0.3004755 -0.1585080 0.05896315
##          Body_mass_index
## Low            1.9194817
## Moderate       0.5173202
```

Question b

(b) Interpret the coefficients of the variables son and weight. Answer:- Interpret Coefficients of Son and Weight:

- Son: The coefficient for "Low" absenteeism is -0.6592586, which means that employees with more sons are less likely to have "Low" absenteeism compared to "High" absenteeism.
- For "Moderate" absenteeism, the coefficient is -0.7510511, which indicates the same. -Thus people more sons are more likely to have "High" absenteeism in comparison.
- Weight: The coefficient for "Low" absenteeism is -0.6191323, which means that higher weight is associated with a higher likelihood of "High" absenteeism compared to "Low".
- For "Moderate" absenteeism, the coefficient is -0.1585080, similarly suggesting "High" absenteeism more likely.

Question c

(c) Use backward selection to decide which predictor variables enter should be kept in the regression model.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
```

```
# Perform backward selection using stepAIC
best_model <- stepAIC(model, direction = "backward")
```

```
## Start:  AIC=396.02
## Absenteeism ~ (ID + `Reason for absence` + Month_of_absence +
##     Day_of_the_week + Seasons + Transportation_expense + Distance_from_Residence_to_Work +
##     Service_time + Age + Work_load_Average_in_days + `Hit_ target` +
##     Disciplinary_failure + Education + Son + Social_drinker +
##     Social_smoker + Pet + Weight + Height + Body_mass_index +
##     Absenteeism_time_in_hours) - Absenteeism_time_in_hours
##
##                                    Df    AIC
## - Month_of_absence                 24 369.34
## - Seasons                           6 387.68
## - Education                         6 389.73
## - Social_smoker                     2 392.07
## - Work_load_Average_in_days         2 392.11
## - ID                                2 392.69
## - Transportation_expense            2 392.72
## - Service_time                      2 392.82
## - Pet                               2 393.07
## - Distance_from_Residence_to_Work   2 393.34
## - Son                               2 393.42
## - Age                               2 393.47
## - `Hit_ target`                     2 394.28
## - Height                            2 394.62
## - Weight                            2 395.29
```

```
## - Body_mass_index                  2 395.67
## <none>                               396.02
## - Social_drinker                   2 396.90
## - Day_of_the_week                  8 397.61
## - Disciplinary_failure             2 411.33
## - `Reason for absence`             2 419.68
##
## Step:  AIC=369.34
## Absenteeism ~ ID + `Reason for absence` + Day_of_the_week + Seasons +
##     Transportation_expense + Distance_from_Residence_to_Work +
##     Service_time + Age + Work_load_Average_in_days + `Hit_ target` +
##     Disciplinary_failure + Education + Son + Social_drinker +
##     Social_smoker + Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - Seasons                          6 362.27
## - Education                        6 363.12
## - Transportation_expense           2 365.50
## - ID                               2 365.64
## - Service_time                     2 366.09
## - Social_smoker                    2 366.09
## - Age                              2 366.12
## - Work_load_Average_in_days        2 366.40
## - Pet                              2 367.20
## - `Hit_ target`                    2 368.14
## - Height                           2 369.03
## <none>                               369.34
## - Distance_from_Residence_to_Work  2 369.39
## - Son                              2 369.49
## - Weight                           2 369.54
## - Body_mass_index                  2 370.05
## - Social_drinker                   2 372.72
## - Day_of_the_week                  8 373.63
## - Disciplinary_failure             2 388.27
## - `Reason for absence`             2 392.40
##
## Step:  AIC=362.27
## Absenteeism ~ ID + `Reason for absence` + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Age + Work_load_Average_in_days +
##     `Hit_ target` + Disciplinary_failure + Education + Son +
##     Social_drinker + Social_smoker + Pet + Weight + Height +
##     Body_mass_index
##
##                                   Df    AIC
## - Education                        6 355.19
## - Transportation_expense           2 358.44
## - Age                              2 358.70
## - Service_time                     2 358.71
## - Social_smoker                    2 358.88
## - ID                               2 358.94
## - Pet                              2 359.58
## - `Hit_ target`                    2 359.77
## - Height                           2 361.42
## - Weight                           2 361.74
```

```
## - Body_mass_index                  2 362.25
## <none>                               362.27
## - Work_load_Average_in_days        2 362.60
## - Distance_from_Residence_to_Work  2 362.92
## - Son                              2 363.16
## - Day_of_the_week                  8 365.80
## - Social_drinker                   2 366.06
## - Disciplinary_failure             2 379.06
## - `Reason for absence`             2 385.32
##
## Step:  AIC=355.19
## Absenteeism ~ ID + `Reason for absence` + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Age + Work_load_Average_in_days +
##     `Hit_ target` + Disciplinary_failure + Son + Social_drinker +
##     Social_smoker + Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - Social_smoker                    2 351.28
## - ID                               2 351.84
## - Service_time                     2 351.94
## - Age                              2 352.07
## - Transportation_expense           2 352.33
## - `Hit_ target`                    2 352.50
## - Pet                              2 352.83
## - Height                           2 353.61
## - Weight                           2 354.02
## - Body_mass_index                  2 354.47
## <none>                               355.19
## - Work_load_Average_in_days        2 355.31
## - Distance_from_Residence_to_Work  2 355.99
## - Son                              2 357.35
## - Day_of_the_week                  8 358.94
## - Social_drinker                   2 361.79
## - Disciplinary_failure             2 371.39
## - `Reason for absence`             2 379.18
##
## Step:  AIC=351.28
## Absenteeism ~ ID + `Reason for absence` + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Age + Work_load_Average_in_days +
##     `Hit_ target` + Disciplinary_failure + Son + Social_drinker +
##     Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - ID                               2 347.92
## - Service_time                     2 348.20
## - Age                              2 348.23
## - Transportation_expense           2 348.47
## - `Hit_ target`                    2 348.64
## - Pet                              2 349.08
## - Height                           2 349.93
## - Weight                           2 350.42
## - Body_mass_index                  2 350.83
## <none>                               351.28
## - Work_load_Average_in_days        2 351.42
```

```
## - Distance_from_Residence_to_Work  2 352.12
## - Son                              2 353.51
## - Day_of_the_week                  8 355.54
## - Social_drinker                   2 358.19
## - Disciplinary_failure             2 368.49
## - `Reason for absence`             2 375.25
##
## Step:  AIC=347.92
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Age + Work_load_Average_in_days +
##     `Hit_ target` + Disciplinary_failure + Son + Social_drinker +
##     Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - Age                              2 345.38
## - `Hit_ target`                    2 345.41
## - Service_time                     2 345.73
## - Pet                              2 346.04
## - Transportation_expense           2 346.51
## - Height                           2 346.78
## - Weight                           2 347.19
## - Body_mass_index                  2 347.59
## <none>                               347.92
## - Work_load_Average_in_days        2 348.18
## - Distance_from_Residence_to_Work  2 348.69
## - Son                              2 349.57
## - Day_of_the_week                  8 352.11
## - Social_drinker                   2 355.76
## - Disciplinary_failure             2 364.98
## - `Reason for absence`             2 371.76
##
## Step:  AIC=345.38
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Work_load_Average_in_days +
##     `Hit_ target` + Disciplinary_failure + Son + Social_drinker +
##     Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - Service_time                     2 341.98
## - `Hit_ target`                    2 342.98
## - Pet                              2 343.63
## - Transportation_expense           2 343.94
## - Height                           2 344.26
## - Weight                           2 344.61
## - Body_mass_index                  2 345.00
## <none>                               345.38
## - Work_load_Average_in_days        2 345.86
## - Son                              2 346.88
## - Distance_from_Residence_to_Work  2 347.00
## - Day_of_the_week                  8 350.25
## - Social_drinker                   2 352.75
## - Disciplinary_failure             2 362.75
## - `Reason for absence`             2 370.07
##
```

```
## Step:  AIC=341.98
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Work_load_Average_in_days +
##     `Hit_ target` + Disciplinary_failure + Son + Social_drinker +
##     Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - `Hit_ target`                    2 339.45
## - Transportation_expense           2 340.20
## - Height                           2 340.66
## - Pet                              2 340.75
## - Weight                           2 341.09
## - Body_mass_index                  2 341.45
## <none>                               341.98
## - Work_load_Average_in_days        2 342.47
## - Son                              2 342.90
## - Distance_from_Residence_to_Work  2 343.07
## - Day_of_the_week                  8 346.63
## - Social_drinker                   2 350.05
## - Disciplinary_failure             2 359.07
## - `Reason for absence`             2 366.55
##
## Step:  AIC=339.45
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Work_load_Average_in_days +
##     Disciplinary_failure + Son + Social_drinker + Pet + Weight +
##     Height + Body_mass_index
##
##                                   Df    AIC
## - Transportation_expense           2 337.92
## - Height                           2 338.25
## - Pet                              2 338.58
## - Weight                           2 338.66
## - Body_mass_index                  2 339.03
## <none>                               339.45
## - Distance_from_Residence_to_Work  2 340.50
## - Work_load_Average_in_days        2 340.82
## - Son                              2 341.02
## - Day_of_the_week                  8 344.91
## - Social_drinker                   2 348.42
## - Disciplinary_failure             2 356.49
## - `Reason for absence`             2 365.29
##
## Step:  AIC=337.92
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Distance_from_Residence_to_Work +
##     Work_load_Average_in_days + Disciplinary_failure + Son +
##     Social_drinker + Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - Pet                              2 336.04
## - Height                           2 336.95
## - Weight                           2 337.27
## - Body_mass_index                  2 337.67
## <none>                               337.92
```

```
## - Son                               2 338.24
## - Work_load_Average_in_days         2 339.16
## - Distance_from_Residence_to_Work   2 339.23
## - Day_of_the_week                   8 343.45
## - Social_drinker                    2 347.37
## - Disciplinary_failure              2 354.45
## - `Reason for absence`              2 364.72
##
## Step:  AIC=336.04
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Distance_from_Residence_to_Work +
##     Work_load_Average_in_days + Disciplinary_failure + Son +
##     Social_drinker + Weight + Height + Body_mass_index
##
##                                    Df    AIC
## - Height                            2 334.29
## - Weight                            2 334.54
## - Body_mass_index                   2 334.92
## <none>                                336.04
## - Son                               2 336.22
## - Work_load_Average_in_days         2 338.05
## - Distance_from_Residence_to_Work   2 338.47
## - Day_of_the_week                   8 341.15
## - Social_drinker                    2 347.61
## - Disciplinary_failure              2 352.60
## - `Reason for absence`              2 362.51
##
## Step:  AIC=334.29
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Distance_from_Residence_to_Work +
##     Work_load_Average_in_days + Disciplinary_failure + Son +
##     Social_drinker + Weight + Body_mass_index
##
##                                    Df    AIC
## - Weight                            2 330.72
## - Body_mass_index                   2 332.19
## - Son                               2 333.87
## <none>                                334.29
## - Distance_from_Residence_to_Work   2 335.97
## - Work_load_Average_in_days         2 336.74
## - Day_of_the_week                   8 338.19
## - Social_drinker                    2 344.61
## - Disciplinary_failure              2 350.57
## - `Reason for absence`              2 361.23
##
## Step:  AIC=330.72
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Distance_from_Residence_to_Work +
##     Work_load_Average_in_days + Disciplinary_failure + Son +
##     Social_drinker + Body_mass_index
##
##                                    Df    AIC
## - Body_mass_index                   2 328.72
## - Son                               2 330.15
## <none>                                330.72
## - Work_load_Average_in_days         2 333.18
## - Day_of_the_week                   8 335.69
```

```
## - Distance_from_Residence_to_Work  2 337.97
## - Social_drinker                   2 345.76
## - Disciplinary_failure             2 348.27
## - `Reason for absence`             2 357.49
##
## Step:  AIC=328.72
## Absenteeism ~ `Reason for absence` + Day_of_the_week + Distance_from_Residence_to_Work +
##     Work_load_Average_in_days + Disciplinary_failure + Son +
##     Social_drinker
##
##                                   Df    AIC
## <none>                               328.72
## - Son                              2 329.52
## - Work_load_Average_in_days        2 331.24
## - Day_of_the_week                  8 333.20
## - Distance_from_Residence_to_Work  2 335.41
## - Social_drinker                   2 342.50
## - Disciplinary_failure             2 347.48
## - `Reason for absence`             2 357.47
```

```
# Display the summary of the selected model
summary(best_model)
```

```
## Call:
## multinom(formula = Absenteeism ~ `Reason for absence` + Day_of_the_week +
##     Distance_from_Residence_to_Work + Work_load_Average_in_days +
##     Disciplinary_failure + Son + Social_drinker, data = absent,
##     trace = FALSE)
##
## Coefficients:
##          (Intercept) `Reason for absence` Day_of_the_week3 Day_of_the_week4
## Low        -0.867036          0.118271675       -0.6826615       0.08530926
## Moderate   -4.929203         -0.008347173       -2.1494579      -0.41411907
##          Day_of_the_week5 Day_of_the_week6 Distance_from_Residence_to_Work
## Low             12.143758        0.1281333                      0.07359401
## Moderate         9.785527       -1.5220814                      0.05782726
##          Work_load_Average_in_days Disciplinary_failure        Son
## Low                   7.661789e-06            14.368243 -0.6341192
## Moderate              1.779414e-05             1.387935 -0.6924736
##          Social_drinker
## Low          -0.6815554
## Moderate      1.5167742
##
## Std. Errors:
##           (Intercept) `Reason for absence` Day_of_the_week3 Day_of_the_week4
## Low      4.118637e-12         5.985716e-11     2.508075e-12     5.183101e-13
## Moderate 4.035634e-12         5.231664e-11     5.383900e-13     1.318598e-12
##          Day_of_the_week5 Day_of_the_week6 Distance_from_Residence_to_Work
## Low          1.352491e-13     7.871675e-13                    5.804442e-11
## Moderate     1.352578e-13     3.168371e-13                    1.106083e-10
##          Work_load_Average_in_days Disciplinary_failure          Son
## Low                   1.023668e-06         1.897444e-18 7.628945e-12
## Moderate              1.198831e-06         5.194919e-18 4.901323e-12
##          Social_drinker
## Low        1.674507e-12
```

```
## Moderate   3.422302e-12
##
## Residual Deviance: 284.7211
## AIC: 328.7211
```

Logistic Regression Analysis (c,d):

Backward Selection, Interpretation, Recommendation:

- The variables that produce the best model determined by backward selection are: `Reason for absence`, Day_of_the_week, Distance_from_Residence_to_Work, Work_load_Average_in_days, Disciplinary_failure, Son, and Social_drinker.
- Interpretation: Practically in a real work space, we would expect employees with more children to have greater responsibilities, which could in turn affect absenteeism.
- Interpretation: Distance from work is also crucial as long commute can be exhausting which in turn can affect absenteeism as well.
- Recommendation: Make life easier for parents. Provide good maternal/paternal leaves and introduce assisting programs such as free day care.
- Recommendation: Offer spaces to live near the office at reasonable costs.

Flu shots data

A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$, and a client who did not receive a flu shot was coded $Y = 0$. In addition, data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2), for which higher values indicate greater awareness. Also included in the data was client sex, where males were coded $X_3 = 1$ and females were coded $X_3 = 0$.

```r
# Load the data:
flu_shot <- read.table("flu_shot.txt", header = TRUE)
head(flu_shot)
```

```
##   flu_shot age awareness sex
## 1        0  59        52   0
## 2        0  61        55   1
## 3        1  82        51   0
## 4        0  51        70   0
## 5        0  53        70   0
## 6        0  62        49   1
```

```r
str(flu_shot)
```

```
## 'data.frame':    159 obs. of  4 variables:
##  $ flu_shot : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ age      : int  59 61 82 51 53 62 51 70 71 55 ...
##  $ awareness: int  52 55 51 70 70 49 69 54 65 58 ...
##  $ sex      : int  0 1 0 0 0 1 1 1 1 1 ...
```

```r
table(flu_shot$sex)
```

```
##
##  0  1
## 81 78
```

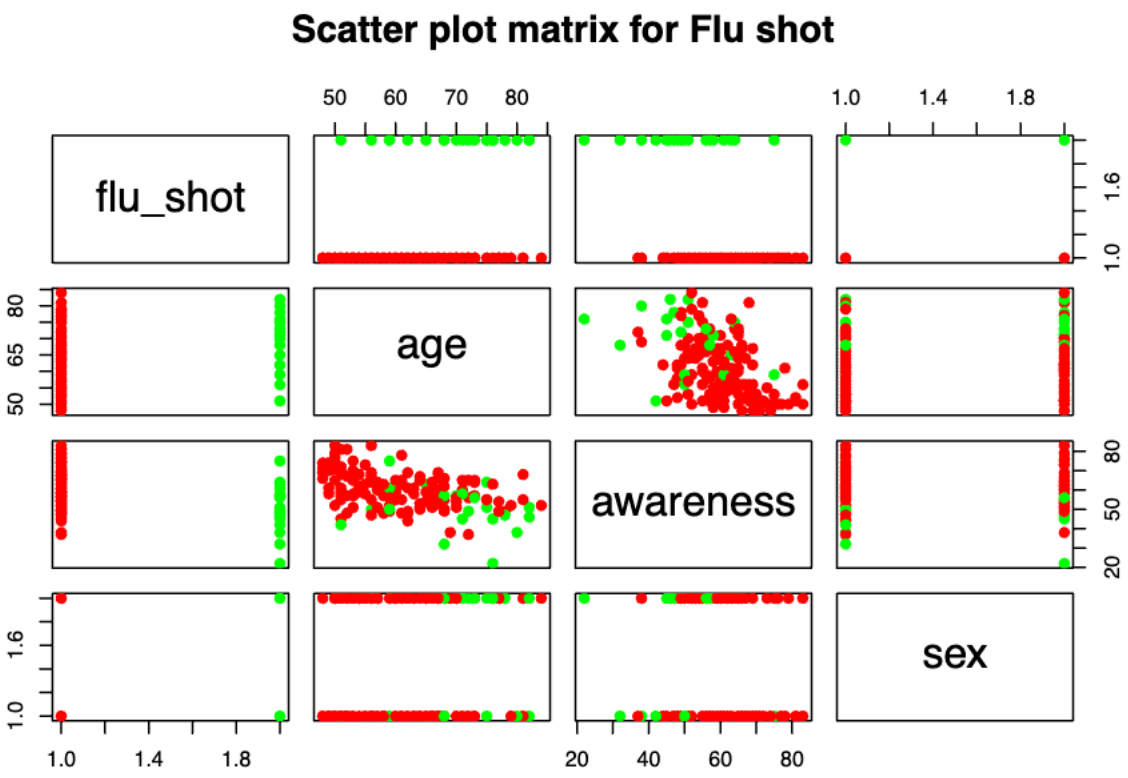
```
# Data structure
flu_shot$flu_shot = as.factor(flu_shot$flu_shot)
flu_shot$sex = as.factor(flu_shot$sex)
str(flu_shot)
```

```
## 'data.frame':    159 obs. of  4 variables:
##  $ flu_shot : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ age      : int  59 61 82 51 53 62 51 70 71 55 ...
##  $ awareness: int  52 55 51 70 70 49 69 54 65 58 ...
##  $ sex      : Factor w/ 2 levels "0","1": 1 2 1 1 1 2 2 2 2 2 ...
```

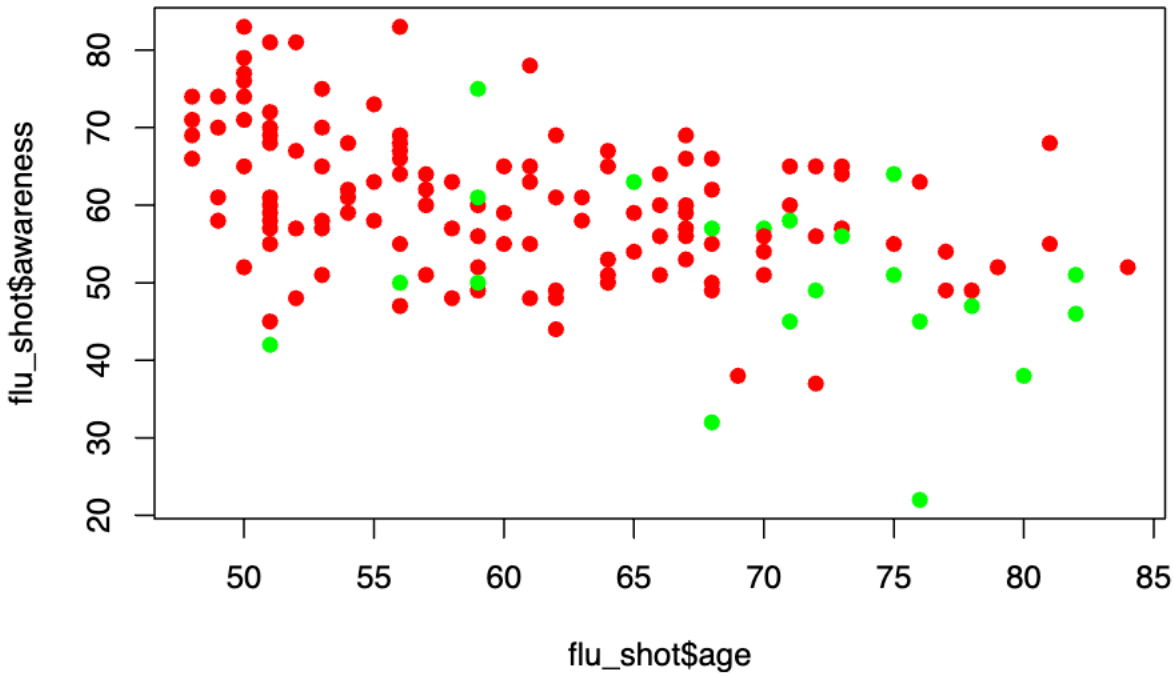
Question a

(a) Create a scatterplot matrix of the data. What are your observations?

```
flu_shot$color <- ifelse(flu_shot$flu_shot == 1, "green", "Red")

pairs(flu_shot[, c(1:4)],col= flu_shot$color,pch = 19, main = "Scatter plot matrix for Flu shot")
```




```
plot(flu_shot$age, flu_shot$awareness, col= flu_shot$color, pch=19)
plot(flu_shot$age, flu_shot$awareness, col= flu_shot$color, pch=19)
```




```
cor(flu_shot$age, flu_shot$awareness)
```

```
## [1] -0.4735178
```

Observations

- There are more people from the age 50-70 which are not vaccinated compared to the vaccinated people of that age group.

-The scatter plot data indicates a pattern showing a reduced frequency of flu vaccination among individuals aged 50-65 with heightened levels of awareness i.e People with more awareness but with less age (<65) are seems to have less vaccinated as compared to people with people with more age but less awareness.

- As age of the people increases, awareness seems to decrease

Question b

(b) Fit a multiple logistic regression to the data with the three predictors in first order terms. Fitting the model:

```
# Fit a multiple logistic regression model with age, health awareness, and sex as predictors
model_flu <- glm(flu_shot ~ age + awareness + sex, family = binomial, data = flu_shot)


# Display the model summary to check coefficients and model fit
summary(model_flu)
```

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial,
##     data = flu_shot)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## age          0.07279    0.03038   2.396  0.01658 *
## awareness   -0.09899    0.03348  -2.957  0.00311 **
## sex1         0.43397    0.52179   0.832  0.40558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

- Fitted regression equation: $\log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

(c) State the fitted regression equation. Therefore, the fitted response function is:

$$P(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}$$

By plugging in $\beta_0 = -1.18$, $\beta_1 = 0.07$, $\beta_2 = -0.10$, $\beta_3 = 0.43$, we get the response function:

$$P(X) = \frac{e^{-1.18+0.07X_1-0.1X_2+0.43X_3}}{1+e^{-1.18+0.07X_1-0.1X_2+0.43X_3}}$$

(d) Obtain $\exp(\hat{\beta}_1)$, $\exp(\hat{\beta}_2)$, and $\exp(\hat{\beta}_3)$ and interpret these numbers. ANS:-

-The $\exp(\hat{\beta}_1) = 1.0755025$, which means that for every one unit increase in age, the odds ratio of client with flu shot (Y=1) versus without flu shot (Y=0) increase by 1.0755025 times.

-The $\exp(\hat{\beta}_2) = 0.9057549$, which means that for every one unit increase in health awareness, the odds ratio of client with flu shot (Y=1) versus without flu shot (Y=0) decrease by 0.9057549 times.

-The $\exp(\hat{\beta}_3) = 1.5433801$, which means that for every one unit increase in x_3 (no.of males), the odds ratio of client with flu shot (Y=1) versus without flu shot (Y=0) increase by 1.5433801 times.

Alternatively

```r
b <- coef(model_flu)
exp(b)
```

```
## (Intercept)         age   awareness        sex1
##   0.3081529   1.0755025   0.9057549   1.5433801
```

-With each additional year of age, there is a roughly 7.6% increase in the odds of getting the flu shot, assuming that all other factors remain constant.

-With every one-unit rise in awareness, the likelihood of receiving the flu shot decreases by approximately 9.4%, under the condition that all other variables remain unchanged.

-Males, in comparison to females, have roughly 54.3% higher odds of getting the flu shot when all other variables are kept constant.

(e) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? ANS:- We have the fitted equation:

$$P(X) = \frac{e^{-1.18+0.07X_1-0.1X_2+0.43X_3}}{1+e^{-1.18+0.07X_1-0.1X_2+0.43X_3}}$$

Replacing x_3 = 1 for male x_1 = 55 for age and x_2 = 60 for awareness

By coding

```
new_data<- data.frame(age=55,awareness=60,sex= "1")

predict=predict(model_flu, newdata = new_data, type="response")

predict
```

```
##          1
## 0.06422197
```

(f) Using the Wald test, determine whether X3 , client gender, can be dropped from the regression model; use $\alpha = 0.05$. ANS:- The Wald test is a statistical test used to assess the significance of individual parameters in a statistical model. It is commonly employed in the context of hypothesis testing for parameters estimated in regression models.

```
summary(model_flu)
```

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial,
##     data = flu_shot)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## age          0.07279    0.03038   2.396  0.01658 *
## awareness   -0.09899    0.03348  -2.957  0.00311 **
## sex1         0.43397    0.52179   0.832  0.40558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

Wald Test: **Hypothesis Test:**

$$H_0 : \beta_3 = 0$$
$$H_a : \beta_3 \neq 0$$

- Sex (X3): p-value = 0.40558. The p-value for gender is greater than 0.05, indicating that sex is not statistically significant according to Wald Test and thus we can drop it out from our model.

(g) Use forward selection to decide which predictor variables enter should be kept in the regression model.

```r
#define intercept-only model
intercept_only= glm(flu_shot ~ 1, family = binomial, data = flu_shot)

#define model with all predictors
full_model <- glm(flu_shot ~ age + awareness + sex, family = binomial, data = flu_shot)

#perform forward selection
forward = step(intercept_only, direction='forward',
               scope=formula(full_model), trace=1)
```

```
## Start:  AIC=136.94
## flu_shot ~ 1
##
##             Df Deviance    AIC
## + awareness  1   113.20 117.20
## + age        1   116.27 120.27
## + sex        1   132.88 136.88
## <none>           134.94 136.94
##
## Step:  AIC=117.2
## flu_shot ~ awareness
##
##        Df Deviance    AIC
## + age   1   105.80 111.80
## + sex   1   111.19 117.19
## <none>      113.20 117.20
##
## Step:  AIC=111.8
## flu_shot ~ awareness + age
##
##        Df Deviance    AIC
## <none>      105.80 111.80
## + sex   1   105.09 113.09
```

```r
#view final model
forward$coefficients
```

```
## (Intercept)   awareness         age
## -1.45778309 -0.09547230  0.07787235
```

Observation

- predictor variables awareness and age should be kept in the regression model.

(h) Use backward selection to decide which predictor variables enter should be kept in the regression model. How does this compare to your results in part (f)?

```r
# Backward selection
#define model with all predictors
full_model <- glm(flu_shot ~ age + awareness + sex, family = binomial, data = flu_shot)

#perform backward selection
backward = step(full_model, direction='backward',
                scope=formula(full_model), trace=1)
```

```
## Start:  AIC=113.09
```

```
## flu_shot ~ age + awareness + sex
##
##             Df Deviance    AIC
## - sex        1   105.80 111.80
## <none>           105.09 113.09
## - age        1   111.19 117.19
## - awareness  1   115.80 121.80
##
## Step:  AIC=111.8
## flu_shot ~ age + awareness
##
##             Df Deviance    AIC
## <none>           105.80 111.80
## - age        1   113.20 117.20
## - awareness  1   116.27 120.27
```

```
#view final model
backward$coefficients
```

```
## (Intercept)         age   awareness
## -1.45778309  0.07787235 -0.09547230
```

Observation

- predictor variables awareness and age should be kept in the regression model.

Forward and Backward Selection:

- Both forward and backward selection models include only Age and Awareness.
- As we found out in (f), the Sex is not important and can be removed to obtain the best model.
- This outcome also matches to the outcome of the Wald test that sex is not significant and hence the variable sex is not included in both forward and backward selections.

NOTE[not for grading]

the Akaike Information Criterion (AIC) is related to the amount of information lost when a model is used to approximate the true underlying process that generated the data

(i)How would you interpret $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_3$?

```
summary(full_model)
```

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial,
##     data = flu_shot)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## age          0.07279    0.03038   2.396  0.01658 *
## awareness   -0.09899    0.03348  -2.957  0.00311 **
## sex1         0.43397    0.52179   0.832  0.40558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

Observation

Interpretation for $\hat{\beta}_0$: -The intercept $\hat{\beta}_0$ is -1.17716, suggesting that for an individual with average age, awareness, and male gender, the odds of getting the flu shot would be approximately exp(-1.17716)= 0.3081529.

Interpretation for $\hat{\beta}_1$: -The odds of someone receiving the flu shot increase by a factor of $\exp(\hat{\beta}_1) = \exp(0.07279) = 1.0755025$ with each additional year of age (X_1), holding awareness and gender constant.

Interpretation for $\hat{\beta}_1$: -Given that the coefficient for X_3 (gender) is 0.43397, it implies that being male is linked to an increase in the odds of getting the flu shot compared to being female. Consequently, the odds of males receiving the flu shot are approximately $\exp(0.43397) = 1.5433801$ times the odds for females. This indicates a significantly higher likelihood of males getting the flu shot compared to females.

——————The End——————————————————————————————

Stat561 Homework-3

Group 2: Syeda Mariah Banu, Clara Cherotich Chelibei, Shanmukha Sai Reddy Manukonda, and Sagar Kalauni

Student performance data

1) The Student Performance Data set is synthetically created dataset which includes 10,000 student profiles. The variables in the data are:

- Hours Studied: The total number of hours spent studying by each student.
- Previous Scores: The scores obtained by students in previous tests.
- Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).
- Sleep Hours: The average number of hours of sleep the student had per day.
- Sample Question Papers Practiced: The number of sample question papers the student practiced.
- Performance Index: A measure of the overall performance of each student.

The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

```
library(readr)
Student_data = read.csv("Student_Performance.csv")
```

```
head(Student_data)
```

```
##   Hours.Studied Previous.Scores Extracurricular.Activities Sleep.Hours
## 1             7              99                         Yes           9
## 2             4              82                          No           4
## 3             8              51                         Yes           7
## 4             5              52                         Yes           5
## 5             7              75                          No           8
## 6             3              78                          No           9
##   Sample.Question.Papers.Practiced Performance.Index
## 1                                1                91
## 2                                2                65
## 3                                2                45
## 4                                2                36
## 5                                5                66
## 6                                6                61
```

```
Student_data$Extracurricular.Activities = as.factor(Student_data$Extracurricular.Activities)
str(Student_data)
```

```
## 'data.frame':    10000 obs. of  6 variables:
##  $ Hours.Studied                  : int  7 4 8 5 7 3 7 8 5 4 ...
##  $ Previous.Scores                : int  99 82 51 52 75 78 73 45 77 89 ...
##  $ Extracurricular.Activities     : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 2 1 1 ...
##  $ Sleep.Hours                    : int  9 4 7 5 8 9 5 4 8 4 ...
##  $ Sample.Question.Papers.Practiced: int  1 2 2 2 5 6 6 6 2 0 ...
##  $ Performance.Index              : num  91 65 45 36 66 61 63 42 61 69 ...
```

```r
attach(Student_data)
```

```r
colnames(Student_data)
```

```
## [1] "Hours.Studied"                    "Previous.Scores"
## [3] "Extracurricular.Activities"       "Sleep.Hours"
## [5] "Sample.Question.Papers.Practiced" "Performance.Index"
```

EDA: (without interpertation)

For numerical variable

```
##
## 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
```

```
## # A tibble: 2 x 2
##   Extracurricular.Activities     n
##   <fct>                      <int>
## 1 No                          2959
## 2 Yes                         2950
```

Table 1: Data summary

Name	yes_Extracurricular
Number of rows	4948
Number of columns	6
Column type frequency:	
factor	1
numeric	5
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Extracurricular.Activities	0	1	FALSE	1	Yes: 4948, No: 0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Hours.Studied	0	1	5.00	2.57	1	3.00	5	7	9
Previous.Scores	0	1	69.59	17.36	40	54.75	70	85	99
Sleep.Hours	0	1	6.49	1.69	4	5.00	6	8	9

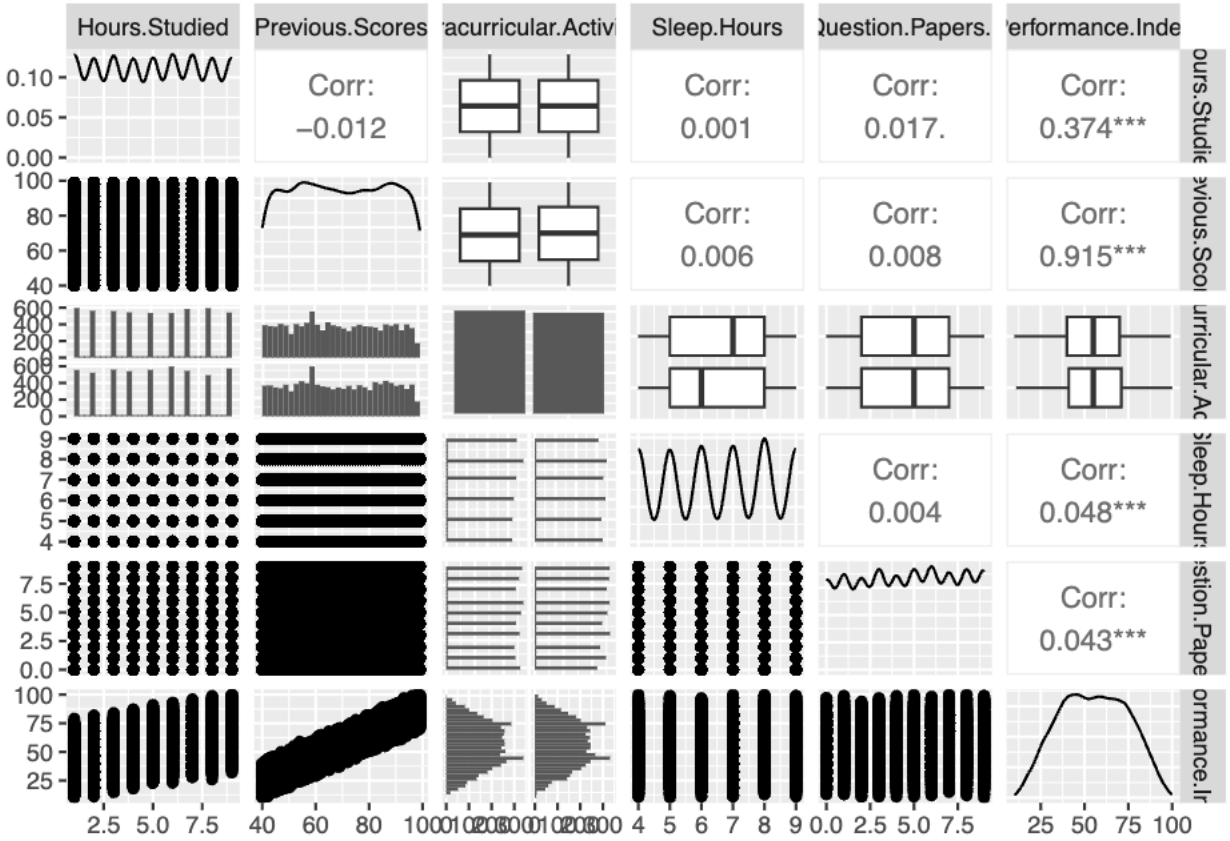
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Sample.Question.Papers.Practiced	0	1	4.62	2.85	0	2.00	5	7	9
Performance.Index	0	1	55.70	19.26	11	41.00	55	71	100

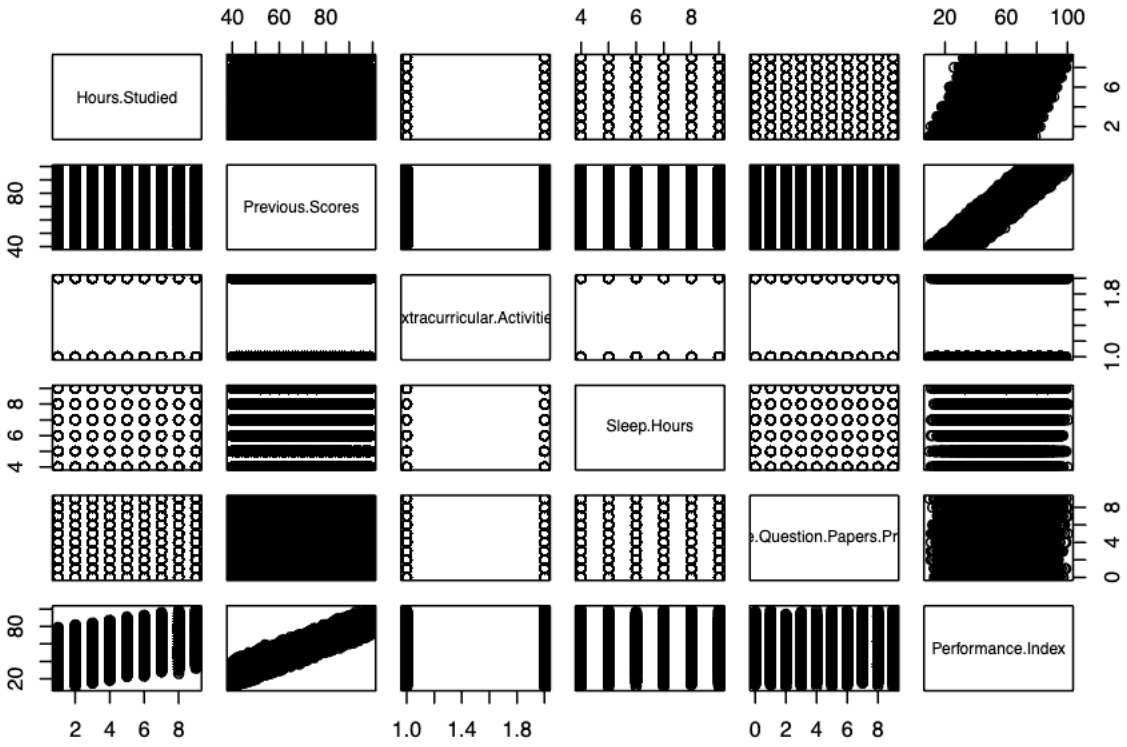
```
##  Hours.Studied   Previous.Scores Extracurricular.Activities  Sleep.Hours
##  Min.   :1.000   Min.   :40.00   No :5052                   Min.   :4.000
##  1st Qu.:3.000   1st Qu.:54.00   Yes:4948                   1st Qu.:5.000
##  Median :5.000   Median :69.00                              Median :7.000
##  Mean   :4.993   Mean   :69.45                              Mean   :6.531
##  3rd Qu.:7.000   3rd Qu.:85.00                              3rd Qu.:8.000
##  Max.   :9.000   Max.   :99.00                              Max.   :9.000
##  Sample.Question.Papers.Practiced Performance.Index
##  Min.   :0.000                    Min.   : 10.00
##  1st Qu.:2.000                    1st Qu.: 40.00
##  Median :5.000                    Median : 55.00
##  Mean   :4.583                    Mean   : 55.22
##  3rd Qu.:7.000                    3rd Qu.: 71.00
##  Max.   :9.000                    Max.   :100.00
```

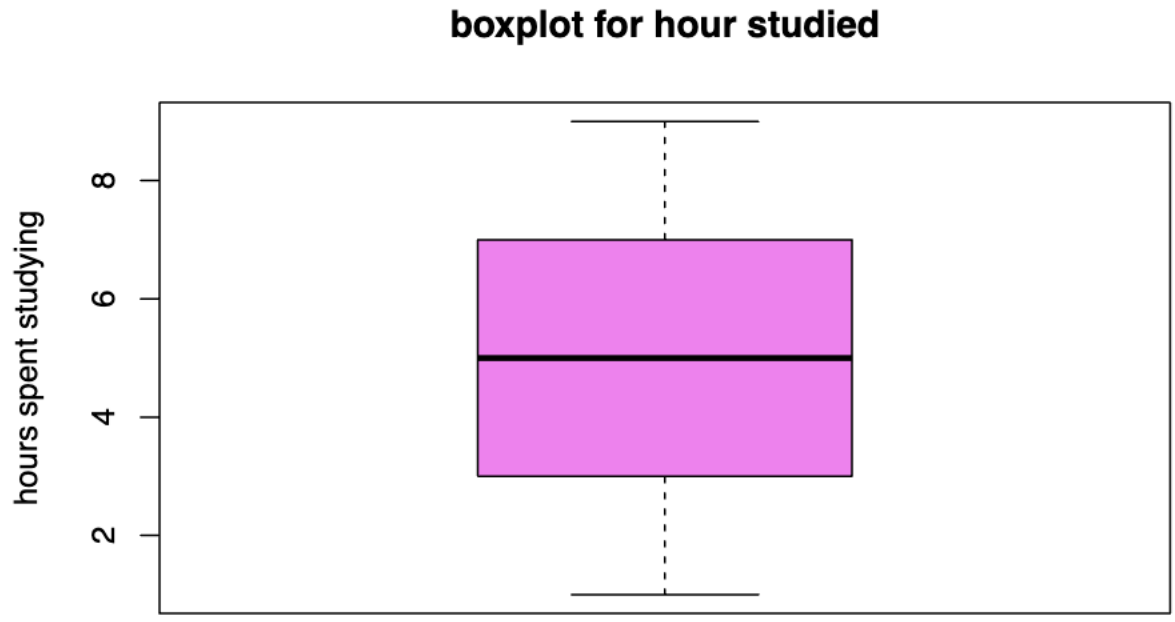
```
## Loading required package: ggplot2
```

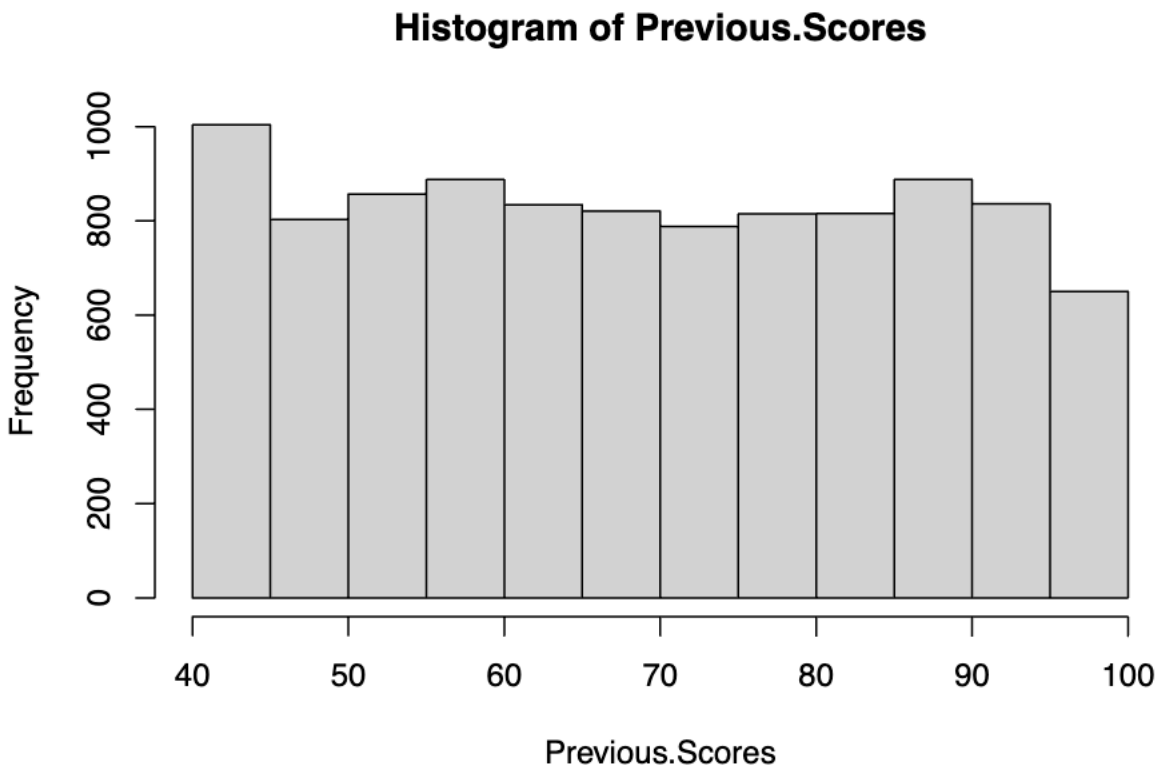
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

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

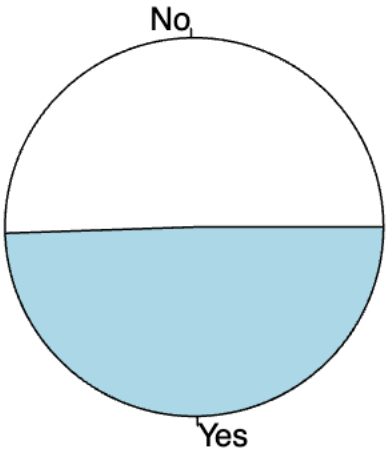

Hours.Studied
Previous.Scores
acurricular.Activi
Sleep.Hours
Question.Papers.
erformance.Inde
Corr:
−0.012
Corr:
0.001
Corr:
0.017.
Corr:
0.374***
Corr:
0.006
Corr:
0.008
Corr:
0.915***
Corr:
0.004
Corr:
0.048***
Corr:
0.043***


Hours.Studied
Previous.Scores
xtracurricular.Activitie
Sleep.Hours
.Question.Papers.Pr
Performance.Index

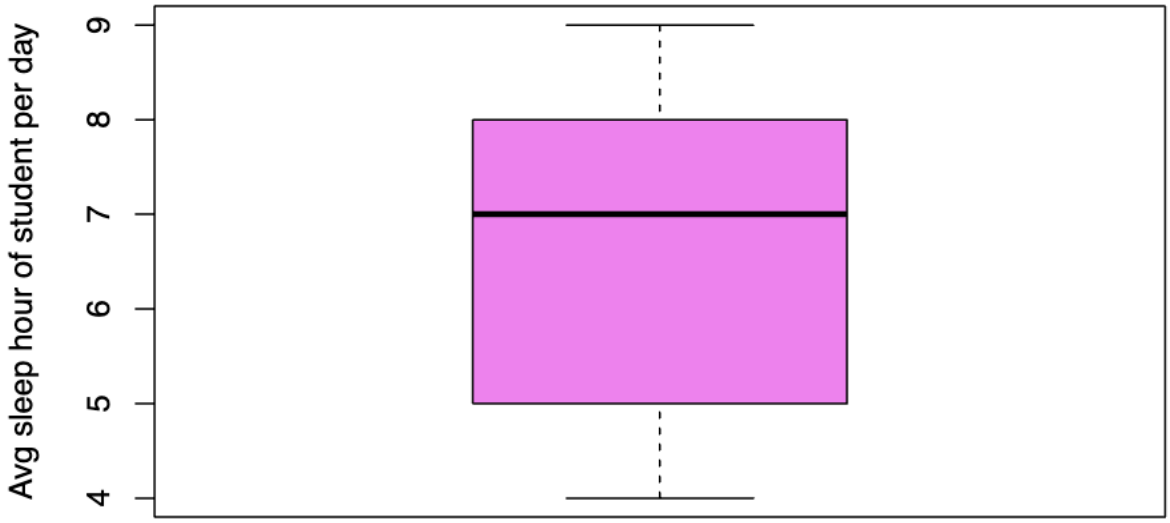

boxplot for hour studied
hours spent studying
2
4
6
8

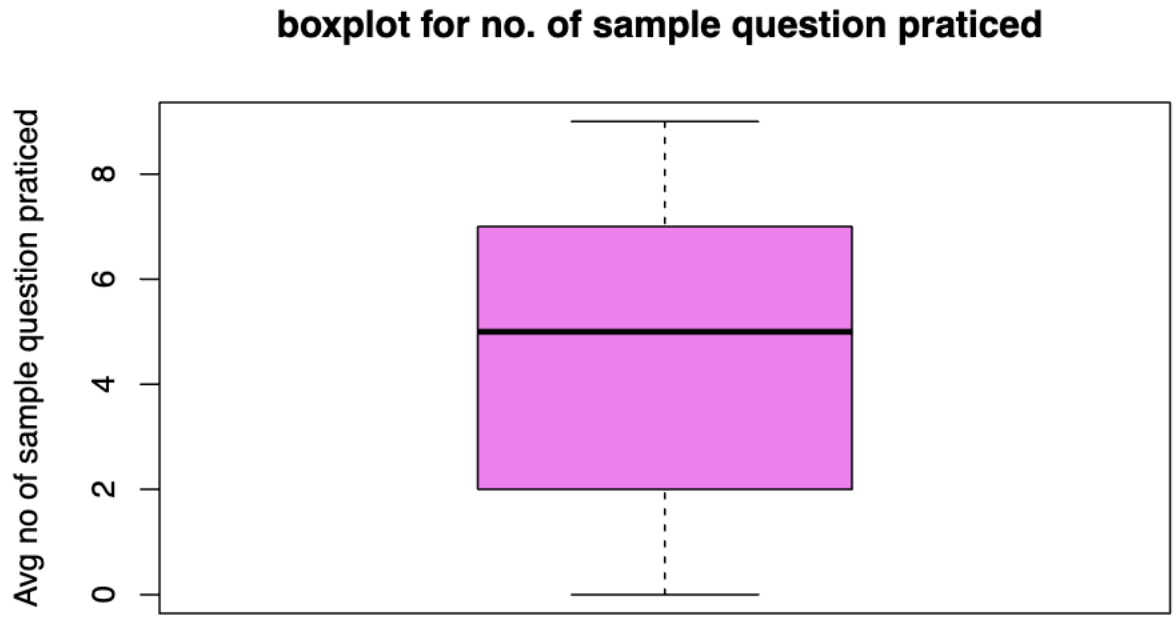

Histogram of Previous.Scores
Frequency
0
200
400
600
800
1000
40
50
60
70
80
90
100
Previous.Scores

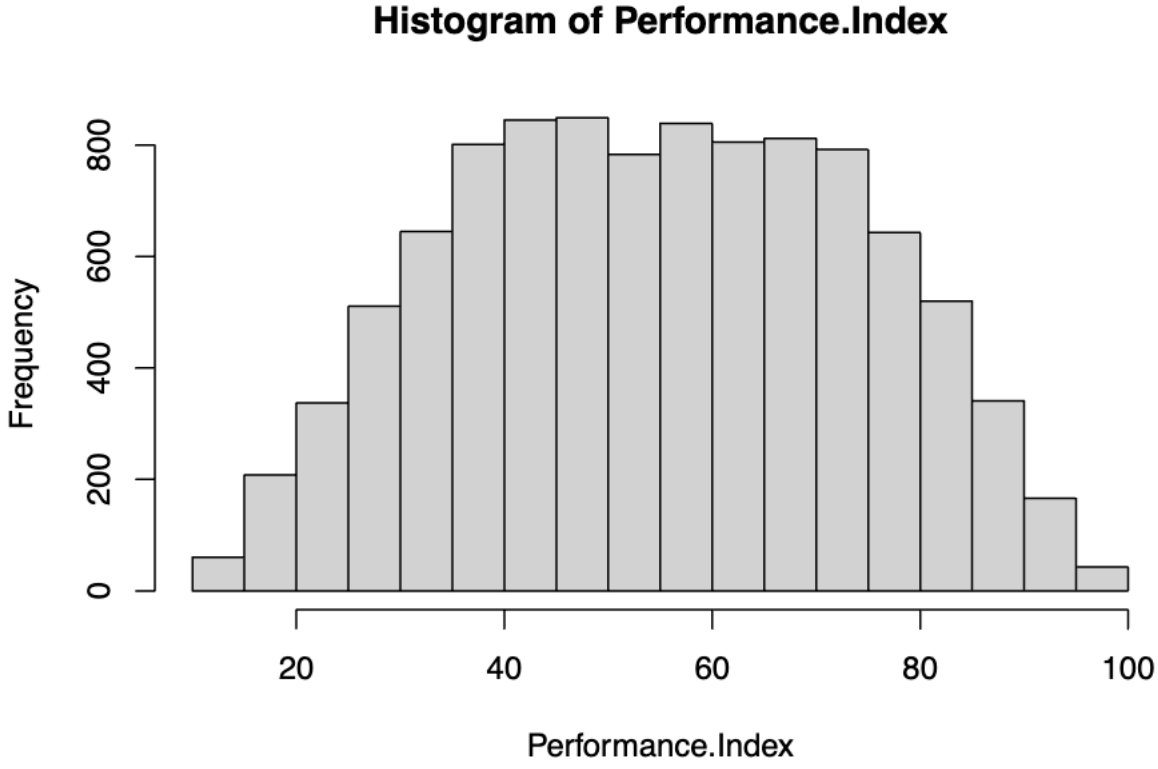
pie chart distribution student involvement in extracurricular activity

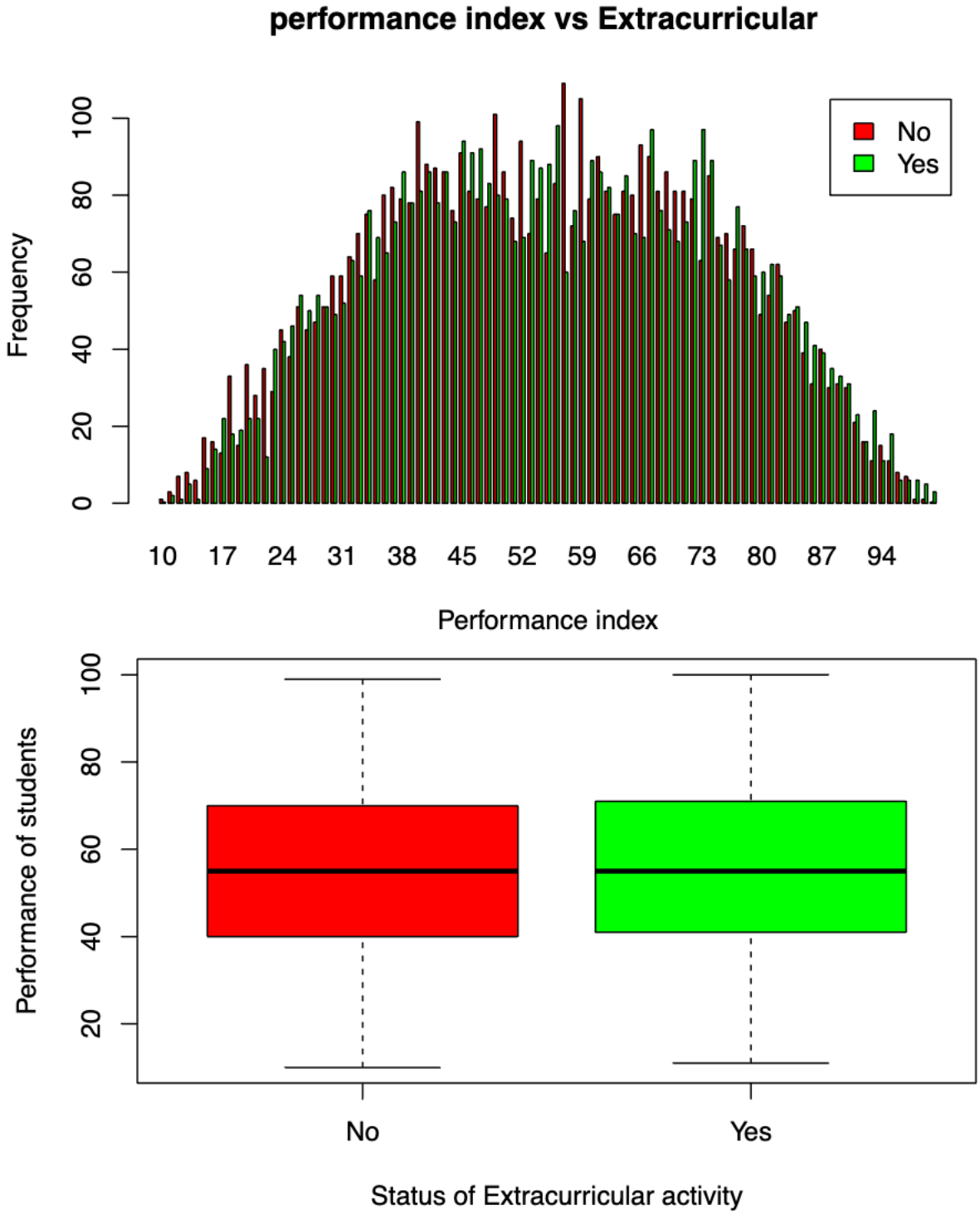


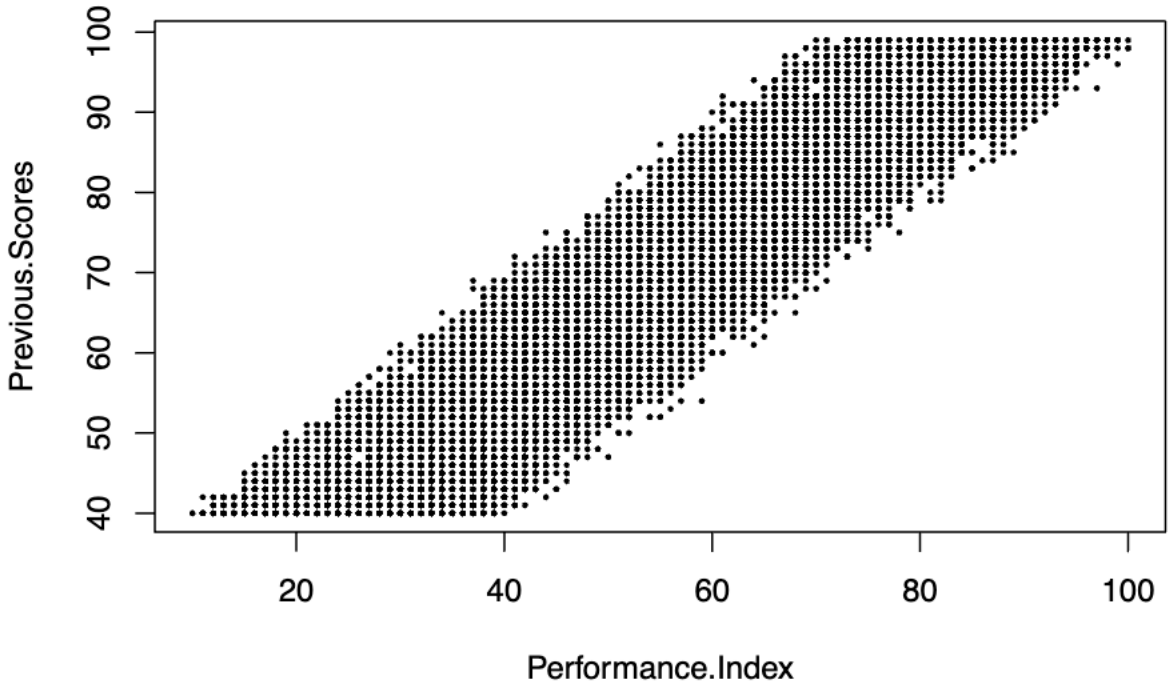
boxplot avg sleep hour




boxplot for no. of sample question praticed
Avg no of sample question praticed
0
2
4
6
8


Histogram of Performance.Index
Frequency
0
200
400
600
800
20
40
60
80
100
Performance.Index


performance index vs Extracurricular
Frequency
0
20
40
60
80
100
No
Yes
10
17
24
31
38
45
52
59
66
73
80
87
94
Performance index
Performance of students
20
40
60
80
100
No
Yes
Status of Extracurricular activity



Question 1.a

a) Formulate a question that you can use this data to answer. You would use multiple linear regression to answer this question. State clearly what your predictor variables would be and what your target variable would be.

ANSWER:- - "How do factors like sleep, preparation, previous record, and activeness affect a student's academic performance?"

- Predictor variables:
  1. Hours Studied
  2. Previous Scores
  3. Extracurricular Activities
  4. Sleep Hours
  5. Sample Question Papers Practiced
- Target variable:
  1. Performance Index

```
lm_model = lm(Performance.Index~., data = Student_data)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Performance.Index ~ ., data = Student_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6333 -1.3684 -0.0311  1.3556  8.7932
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -34.075588   0.127143 -268.01   <2e-16 ***
```

```
## Hours.Studied                     2.852982   0.007873  362.35   <2e-16 ***
## Previous.Scores                   1.018434   0.001175  866.45   <2e-16 ***
## Extracurricular.ActivitiesYes     0.612898   0.040781   15.03   <2e-16 ***
## Sleep.Hours                       0.480560   0.012022   39.97   <2e-16 ***
## Sample.Question.Papers.Practiced  0.193802   0.007110   27.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.038 on 9994 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9887
## F-statistic: 1.757e+05 on 5 and 9994 DF,  p-value: < 2.2e-16
```

In predictive modeling, one of the primary foci is on the model's performance with new data. This is crucial because evaluating error metrics on new data provides insights into the model's ability to generalize to unseen situations. The Root Mean Square Error (RMSE) serves as a measure of the model's average prediction error. For instance, an RMSE value of 6.5 indicates that the model's predictions are, on average, about 6 to 7 units off from the actual values. It is essential to select models that demonstrate strong performance on new data. The process involves: Fitting the model on the training sample to learn the patterns in the data. Making predictions on a new, unseen sample to test the model's learned patterns. Validating the model's performance on this new sample to ensure its predictions are accurate and reliable.

Question 1.b

b) Split your data into a training (70%) and test set (30% test).

```
n= nrow(Student_data)
train.index = sample(1:n, size = n*0.7, replace = FALSE)

train_data= Student_data[train.index,]
test_data = Student_data[-train.index, ]
```

```
dim(train_data)
```

```
## [1] 7000    6
```

```
dim(test_data)
```

```
## [1] 3000    6
```

Thus our split of the data is: Training Set (70%): 7000 – as train_data Testing Set (30%): 3000 – as test_data

Question 1.c

c) Train a multiple linear regression model using the training set and a 10-fold cross validation. Use the train function from the caret package, specifying method = "lm" for linear regression. How would you interpret any 2 of the regression coefficients in the context of the student performance data? Provide the fitted equation.

```
set.seed(123)
library(caret)
```

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

```
# Set up cross-validation settings
train_control = trainControl(method = "cv", number = 10,verboseIter = TRUE)
```

```r
# fitting the lm model
lm_model.cv = train(Performance.Index ~. , data = train_data,
                method = "lm",
                trControl = train_control)
```

```
## + Fold01: intercept=TRUE
## - Fold01: intercept=TRUE
## + Fold02: intercept=TRUE
## - Fold02: intercept=TRUE
## + Fold03: intercept=TRUE
## - Fold03: intercept=TRUE
## + Fold04: intercept=TRUE
## - Fold04: intercept=TRUE
## + Fold05: intercept=TRUE
## - Fold05: intercept=TRUE
## + Fold06: intercept=TRUE
## - Fold06: intercept=TRUE
## + Fold07: intercept=TRUE
## - Fold07: intercept=TRUE
## + Fold08: intercept=TRUE
## - Fold08: intercept=TRUE
## + Fold09: intercept=TRUE
## - Fold09: intercept=TRUE
## + Fold10: intercept=TRUE
## - Fold10: intercept=TRUE
## Aggregating results
## Fitting final model on full training set
```

```r
summary(lm_model.cv)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.540 -1.362 -0.040  1.351  8.813
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -34.026292   0.151016 -225.32   <2e-16 ***
## Hours.Studied                     2.839279   0.009449  300.47   <2e-16 ***
## Previous.Scores                   1.018381   0.001402  726.30   <2e-16 ***
## Extracurricular.ActivitiesYes     0.623458   0.048841   12.77   <2e-16 ***
## Sleep.Hours                       0.482937   0.014367   33.61   <2e-16 ***
## Sample.Question.Papers.Practiced  0.194401   0.008506   22.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.042 on 6994 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9888
## F-statistic: 1.236e+05 on 5 and 6994 DF,  p-value: < 2.2e-16
```

Hours.Studied (β_1):

The coefficient for Hours.Studied is approximately 2.8508. This means that for every additional hour spent studying, we can expect the Performance Index to increase by approximately 2.8508 points, assuming all other variables (such as Previous Scores, Extracurricular Activities, Sleep Hours, and Sample Question Papers Practiced) remain constant.

Previous.Scores (β_2):

The coefficient for Previous.Scores is approximately 1.0180. This suggests that for every additional point obtained in the Previous Scores, we can expect the Performance Index to increase by approximately 1.0180 points, assuming all other variables remain constant.

Thus, Spending more time studying (Hours.Studied) and having higher scores in previous tests (Previous.Scores) are associated with higher Performance Index scores.

Fitted Equation:

$$\text{Performance.Index} = -34.047166 + 2.850817 \times \text{Hours.Studied} + 1.018021 \times \text{Previous.Scores} + 0.571091 \times \text{Extracurricular.Activitie}$$

Question 1.d

d) Evaluate the performance of your regression model on the test set. Use metrics such as R-squared and Root Mean Squared Error (RMSE) to assess how well the model predicts the target variable. Comment on this.

```
# Predict on test set
prediction = predict(lm_model.cv,test_data)

# Compute errors
errors=prediction-test_data$Performance.Index

# Calculate RMSE
RMSE= sqrt(mean(errors^2))

R_squared <- summary(lm_model.cv)$r.squared

# computing the metric
cat('RMSE:', RMSE, '\n')
```

```
## RMSE: 2.031037
```

```
cat('R_squared:', R_squared, '\n')
```

```
## R_squared: 0.9888102
```

Interpertation

RMSE (Root Mean Squared Error): The RMSE value of approximately 1.995 indicates the average difference between the observed values (actual Performance.Index) and the values predicted by the model. Specifically, on average, the model's predictions deviate from the actual values by approximately 1.995 points on the scale of Performance.Index.

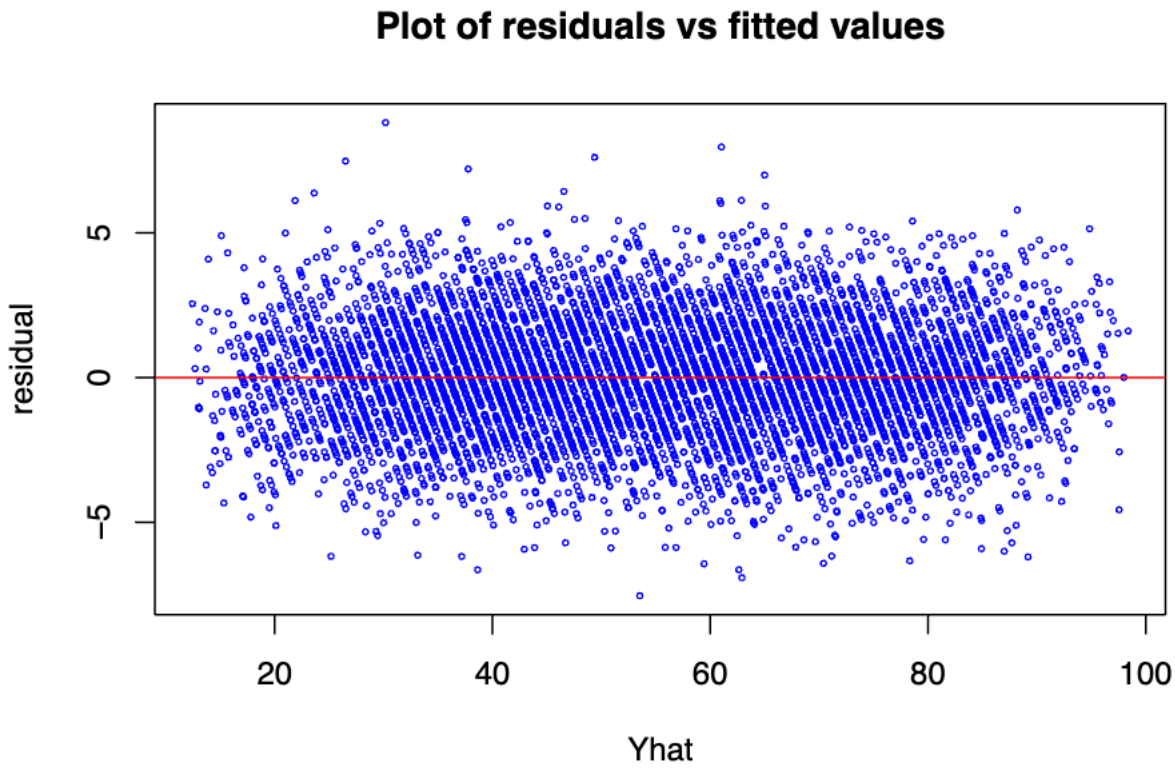
R-squared: The R-squared value of approximately 0.989 indicates the proportion of the variance in the dependent variable (Performance.Index) that is explained by the independent variables (Hours.Studied,

Previous.Scores, Extracurricular Activities, Sleep Hours, and Sample Question Papers Practiced) included in the model. In this case, around 98.93% of the variability in the Performance.Index can be explained by the independent variables in the model.

Question 1.e

e) Investigate the residuals from your regression model in c) to check model if model assumptions are satisfied. What do these diagnostics tell you about the model's assumptions and its suitability for the data?

```
#residual values
residual=resid(lm_model.cv)

#fitted values
fitted.vals=fitted(lm_model.cv)

#2. Plot of residuals against Yhat
plot(fitted.vals,residual,col="blue",ylab="residual",xlab="Yhat",
     main="Plot of residuals vs fitted values",cex=0.4)
abline(h=0,col="red")
```

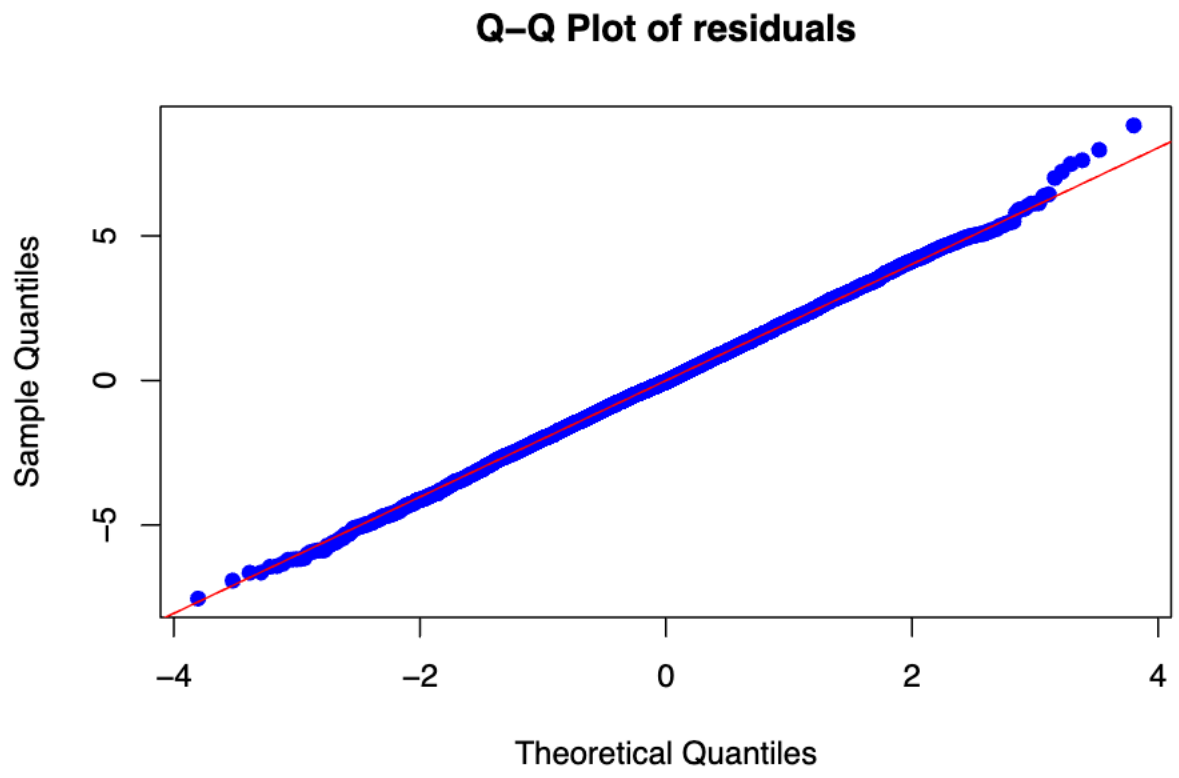


Model

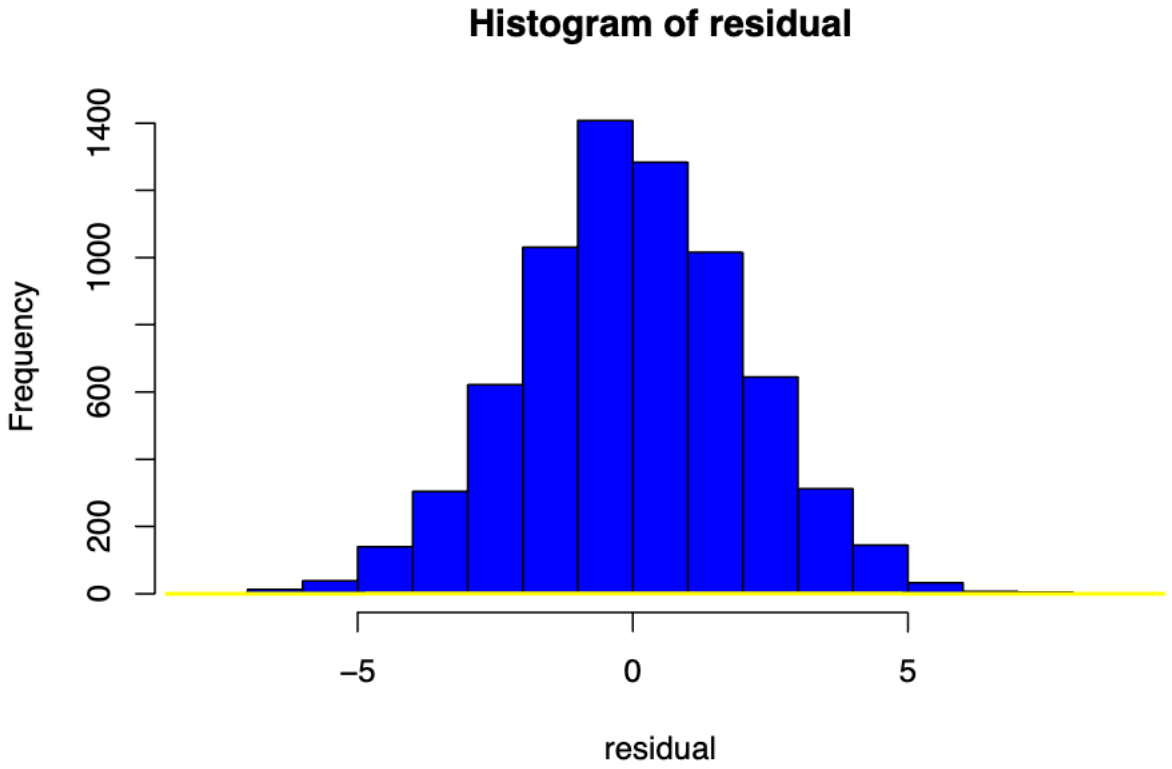
Diagnostics:

- The model assumptions of linearity and homoscedasticity are satisfied as the plot does not show any patterns and is evenly spread out.

```
#3. Normal probability plot of residuals
qqnorm(residual,col="blue",pch=19,main="Q-Q Plot of residuals")
qqline(residual,col="red",pch=19)
```




```
hist(residual,col="blue")
dens <- density(residual)# Compute the density
lines(dens, col="yellow", lwd=2)# Overlay the density curve
```



Question 1.f

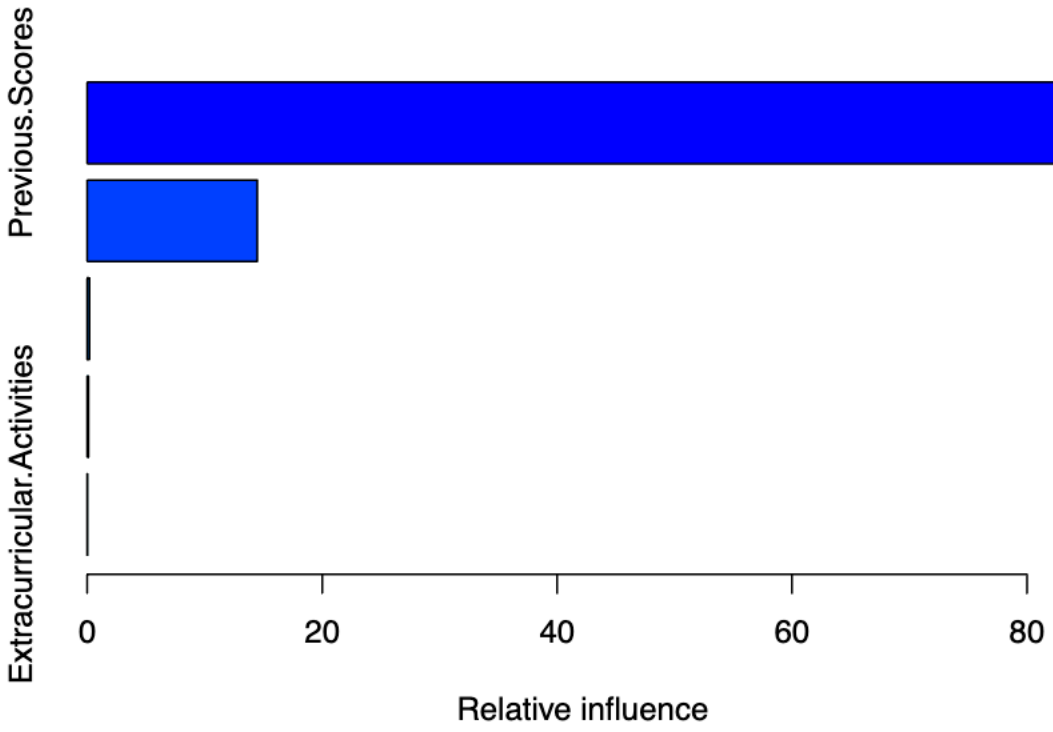
f) Based on the insights gained from your analysis, propose interventions that could potentially improve student performance. Are there any violations of assumptions? Suggest remedies for these violations.

```
library(gbm)
```

```
## Loaded gbm 2.1.9
```

```
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.c
```

```
# Fit the GBM model
boost.model <- gbm(Performance.Index ~ .,
                   data = train_data,
                   distribution = "gaussian",  # Adjusted for regression
                   n.trees = 100,
                   interaction.depth = 5,
                   shrinkage = 0.1)

# Summarize the model
summary(boost.model)
```




```
##                                                           var    rel.inf
## Previous.Scores                                 Previous.Scores 85.1794430
## Hours.Studied                                     Hours.Studied 14.4826960
## Sleep.Hours                                         Sleep.Hours  0.2028021
## Sample.Question.Papers.Practiced Sample.Question.Papers.Practiced  0.1024729
## Extracurricular.Activities             Extracurricular.Activities  0.0325859
```

Improving Student Performance:

- Students should be encouraged to put more efforts into each and every test as the previous test score matter alot in students performance in new test i.e students should aim to consistently perform well as previous scores affect performance.

- Students should be encouraged to put more efforts into preparation by studying for more hours.
- And lastly they should sleep well to do good in exam, which can help alot in improving students performance.
- Violations of assumptions: One would assume that practicing previous question papers would have a significant impact on a student's performance, at least more than whether or not they participate in extracurricular activities but instead this is violated and it has the least positive correlation to performance!
- Remedies to fix this should include maintaining a pattern the students can learn when they solve previous papers which helps them improve performance.

Question 2

Question 2 In this study, we are going to determine whether individuals can get a loan or not based on the following predictor variables:

- Loan ID: A unique loan ID.
- Gender: Either male or female.
- Married: Weather Married(yes) or Not Married(No).
- Dependents: Number of persons depending on the client.
- Education: Applicant Education(Graduate or Undergraduate).
- Self Employed: Self-employed (Yes/No).
- Applicant Income: Applicant income.
- Co-applicant Income: Co-applicant income.
- Loan Amount: Loan amount in thousands.
- Loan Amount Term: Terms of the loan in months.
- Credit History: Credit history meets guidelines.
- Property Area: Applicants are living either Urban, Semi-Urban or Rural.

Target variable: Loan Status: Loan approved (Y/N).

```
loan_data = read.csv("loan_data.csv")
```

```
head(loan_data)
```

```
##    Loan_ID Gender Married Dependents    Education Self_Employed ApplicantIncome
## 1 LP001003   Male     Yes          1     Graduate            No            4583
## 2 LP001005   Male     Yes          0     Graduate           Yes            3000
## 3 LP001006   Male     Yes          0 Not Graduate            No            2583
## 4 LP001008   Male      No          0     Graduate            No            6000
## 5 LP001013   Male     Yes          0 Not Graduate            No            2333
## 6 LP001024   Male     Yes          2     Graduate            No            3200
##   CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area
## 1              1508        128              360              1         Rural
## 2                 0         66              360              1         Urban
## 3              2358        120              360              1         Urban
## 4                 0        141              360              1         Urban
## 5              1516         95              360              1         Urban
## 6               700         70              360              1         Urban
```

```
##   Loan_Status
## 1           N
## 2           Y
## 3           Y
## 4           Y
## 5           Y
## 6           Y
```

```
attach(loan_data)
loan_data$Gender = as.factor(loan_data$Gender)
loan_data$Married = as.factor(loan_data$Married)
loan_data$Dependents = as.factor(loan_data$Dependents)
loan_data$Education = as.factor(loan_data$Education)
loan_data$Self_Employed = as.factor(loan_data$Self_Employed)
loan_data$Property_Area = as.factor(loan_data$Property_Area)
loan_data$Loan_Status = as.factor(loan_data$Loan_Status)
str(loan_data)
```

```
## 'data.frame':    381 obs. of  13 variables:
##  $ Loan_ID          : chr  "LP001003" "LP001005" "LP001006" "LP001008" ...
##  $ Gender           : Factor w/ 3 levels "","Female","Male": 3 3 3 3 3 3 3 3 3 3 ...
##  $ Married          : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 1 2 1 ...
##  $ Dependents       : Factor w/ 5 levels "","0","1","2",..: 3 2 2 2 2 4 4 2 4 2 ...
##  $ Education        : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 2 1 2 1 1 1 1 1 ...
##  $ Self_Employed    : Factor w/ 3 levels "","No","Yes": 2 3 2 2 2 2 1 2 2 2 ...
##  $ ApplicantIncome  : int  4583 3000 2583 6000 2333 3200 2500 1853 1299 4950 ...
##  $ CoapplicantIncome: num  1508 0 2358 0 1516 ...
##  $ LoanAmount       : int  128 66 120 141 95 70 109 114 17 125 ...
##  $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 120 360 ...
##  $ Credit_History   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Property_Area    : Factor w/ 3 levels "Rural","Semiurban",..: 1 3 3 3 3 3 3 1 3 3 ...
##  $ Loan_Status      : Factor w/ 2 levels "N","Y": 1 2 2 2 2 2 2 1 2 2 ...
```

```
sum(is.na(loan_data)) #tells you number of rows with missing data
```

```
## [1] 41
```

```
loan_data=na.omit(loan_data)#remove the missing observations
sum(is.na(loan_data))# use this to check if you have removed missing values
```

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

```
colnames(loan_data)
```

```
##  [1] "Loan_ID"           "Gender"            "Married"
##  [4] "Dependents"        "Education"         "Self_Employed"
##  [7] "ApplicantIncome"   "CoapplicantIncome" "LoanAmount"
## [10] "Loan_Amount_Term"  "Credit_History"    "Property_Area"
## [13] "Loan_Status"
```

Question 2.a

(a) Split the data set into a training set and a test set (80% training, 30% test).

```
train.index = sample(1:nrow(loan_data), size = nrow(loan_data)*0.8, replace = FALSE)
```

```
train_data_loan= loan_data[train.index,]
test_data_loan = loan_data[-train.index, ]
```

```
dim(loan_data)
```

```
## [1] 340  13
```

```
dim(train_data_loan)
```

```
## [1] 272  13
```

```
dim(test_data_loan)
```

```
## [1] 68 13
```

Thus our split of the data is: Training Set (80%): 272 – as train_data_loan Testing Set (20%): 68 – as test_data_loan

Question 2.b

(b) Train the logistic regression model using the training set with a 10-fold cross-validation to optimize model parameters (Use the caret package).

```
# Load the caret package
library(caret)

# Set up cross-validation settings
train_control = trainControl(method = "cv", number = 10)

# Set up the logistic regression model using the logit link
set.seed(123)
lm_model_loan = train(Loan_Status ~., data = train_data_loan[,-1],
               method = "glm", family = binomial(link = logit),
               trControl = train_control)

# Output the results
print(lm_model_loan)
```

```
## Generalized Linear Model
##
## 272 samples
##  11 predictor
##   2 classes: 'N', 'Y'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 244, 244, 245, 245, 245, 245, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8304335  0.496235
```

```
summary(lm_model_loan)
```

```
##
## Call:
```

```
## NULL
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.1871813  2.6209271  -2.742  0.00610 **
## GenderFemale              1.7931119  1.5303169   1.172  0.24131
## GenderMale                1.9681587  1.5214375   1.294  0.19580
## MarriedYes                0.0095197  0.4425306   0.022  0.98284
## Dependents0               0.1208865  1.1651001   0.104  0.91736
## Dependents1              -0.0667508  1.2155553  -0.055  0.95621
## Dependents2               0.1933691  1.2309378   0.157  0.87517
## `Dependents3+`            0.3069463  1.3200245   0.233  0.81613
## `EducationNot Graduate` -0.3671895  0.4139913  -0.887  0.37511
## Self_EmployedNo          -0.1791270  0.7885594  -0.227  0.82030
## Self_EmployedYes         -0.8915940  0.9609942  -0.928  0.35352
## ApplicantIncome           0.0002090  0.0001634   1.279  0.20087
## CoapplicantIncome         0.0003371  0.0001742   1.935  0.05303 .
## LoanAmount                0.0004413  0.0073701   0.060  0.95226
## Loan_Amount_Term          0.0010650  0.0028288   0.376  0.70656
## Credit_History            4.9389313  0.8511905   5.802 6.54e-09 ***
## Property_AreaSemiurban    1.5329727  0.4778751   3.208  0.00134 **
## Property_AreaUrban        0.3958365  0.4437533   0.892  0.37238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 318.40  on 271  degrees of freedom
## Residual deviance: 209.31  on 254  degrees of freedom
## AIC: 245.31
##
## Number of Fisher Scoring iterations: 5
```

-The accuracy in the training data is = 0.8418498 i.e 84.18%

Question 2.c

(c) Evaluate the model on the test set using appropriate metrics ; Accuracy, Sensitivity, Specificity, and AUC (Area Under the ROC Curve). Analyze the confusion matrix to understand the model's performance in predicting an individual's loan status. Interpret each of these metrics in terms of the data and the model.

```r
# Predict class probabilities using the model
predict_prob=predict(lm_model_loan, newdata = test_data_loan[,-1], type = "prob")
# Predict actual categories using the model
predictions = predict(lm_model_loan, newdata = test_data_loan[,-1], type = "raw")
```

```r
# Create the confusion matrix
conf_matrix = confusionMatrix(predictions, test_data_loan$Loan_Status,
                              positive="Y")

# Print the confusion matrix
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  N  Y
##          N 15  2
##          Y 12 39
##
##                Accuracy : 0.7941
##                  95% CI : (0.6788, 0.8826)
##     No Information Rate : 0.6029
##     P-Value [Acc > NIR] : 0.0006442
##
##                   Kappa : 0.541
##
##  Mcnemar's Test P-Value : 0.0161569
##
##             Sensitivity : 0.9512
##             Specificity : 0.5556
##          Pos Pred Value : 0.7647
##          Neg Pred Value : 0.8824
##              Prevalence : 0.6029
##          Detection Rate : 0.5735
##    Detection Prevalence : 0.7500
##       Balanced Accuracy : 0.7534
##
##        'Positive' Class : Y
##
```

```
# Extracting the metrics
accuracy = conf_matrix$overall['Accuracy']
precision= conf_matrix$byClass['Precision']
recall = conf_matrix$byClass['Recall']
F1 = conf_matrix$byClass['F1']
```

```
# Print the metrics
print(paste("Accuracy:", accuracy))
```

```
## [1] "Accuracy: 0.794117647058823"
```

```
print(paste("Precision:", precision))
```

```
## [1] "Precision: 0.764705882352941"
```

```
print(paste("Recall:", recall))
```

```
## [1] "Recall: 0.951219512195122"
```

```
print(paste("F1 Score:", F1))
```

```
## [1] "F1 Score: 0.847826086956522"
```

```
sensitivity( test_data_loan$Loan_Status,predictions,positive="Y")
```

```
## [1] 0.7647059
```

```
specificity( test_data_loan$Loan_Status,predictions,negative="N")
```

```
## [1] 0.8823529
```

-The accuracy in the test data is = 0.764705882352941 i.e 76.47%

```
####ROC and AUC
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

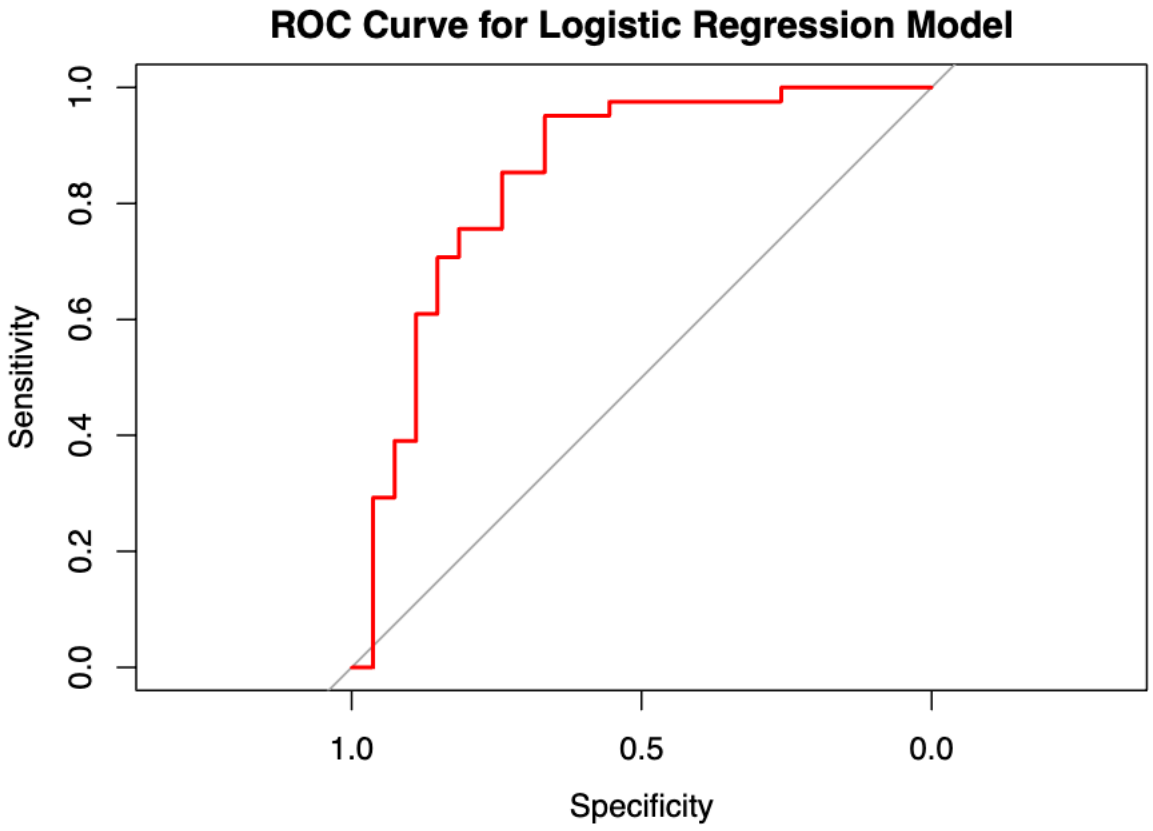
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
#ROC
# Predict probabilities on the test dataset
probabilities = predict(lm_model_loan, newdata = test_data_loan[,-1], type = "prob")[,2]
task_roc= roc(test_data_loan$Loan_Status,probabilities)
```

```
## Setting levels: control = N, case = Y
```

```
## Setting direction: controls < cases
```

```
plot(task_roc, main = "ROC Curve for Logistic Regression Model",
     col="red")
```




```
#AUC
auc(task_roc)
```

```
## Area under the curve: 0.8473
```

Stat561 Homework 4

Group 2: Syeda Mariah Banu, Clara Cherotich Chelibei, Shanmukha Sai Reddy Manukonda, and Sagar Kalauni

Using Absenteesim data from HW 2

For the questions in 1-9 below,

A. Phase 1: Use multiple linear regression with the response variable being "Absenteeism in hours" variable.

B. Phase 2: Use logistic regression with the Absenteeism categorization done in Homework 2 as your response.

The features are: Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Education, Son, Pet, Weight, Height, Body mass index. Be sure to make the categorical variables factors in R.

```
library(readxl)
Absenteeism_data = read_excel("Absenteeism_at_work.xls")
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.1
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```
# Making proper data structure by formatting proper data type by looking the meta data of the data.

Absenteeism_data$Month_of_absence = as.factor(Absenteeism_data$Month_of_absence)
Absenteeism_data$Day_of_the_week = as.factor(Absenteeism_data$Day_of_the_week)
Absenteeism_data$Seasons = as.factor(Absenteeism_data$Seasons)
Absenteeism_data$Education = as.factor(Absenteeism_data$Education)

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
```

```
Absenteeism_data <- Absenteeism_data %>%
  mutate(Absenteeism = case_when(
    Absenteeism_time_in_hours <= 20 ~ "Low",
    Absenteeism_time_in_hours <= 40 ~ "Moderate",
    Absenteeism_time_in_hours > 40 ~ "High"
  ))
```

```
Absenteeism_data$Absenteeism = as.factor(Absenteeism_data$Absenteeism)
Absenteeism_data_glm = Absenteeism_data[, c(-16,-1)] #ID column is not required,
Absenteeism_data_lm = Absenteeism_data[, c(-17,-1)]
```

- Absenteeism_data_glm model has Absenteeism as final column
- Absenteeism_data_lm model has Absenteeism_time_in_hours as final column.

1. Split the data into training and test set. How did you do your data split?

```
# Splitting the data into training and test sets
set.seed(123) # Setting a seed for reproducibility
splitIndex = sample(1:nrow(Absenteeism_data_lm), 0.8 * nrow(Absenteeism_data_lm))
trainData <- Absenteeism_data_lm[splitIndex,]
testData <- Absenteeism_data_lm[-splitIndex,]

# Prepare matrix of predictors and response variable for glmnet
x_train <- model.matrix(Absenteeism_time_in_hours ~ ., data = trainData)[, -1] # -1 to exclude the inte
y_train <- trainData$Absenteeism_time_in_hours

x_test <- model.matrix(Absenteeism_time_in_hours ~ ., data = testData)[, -1]
y_test <- testData$Absenteeism_time_in_hours
```

```
dim(Absenteeism_data_lm)
```

```
## [1] 740  15
```

```
dim(trainData)
```

```
## [1] 592  15
```

```
dim(testData)
```

```
## [1] 148  15
```

Question 1

So we have done the 80% – 20% split. Where 80% is for training set and remaining 20% for the test data set.

2. Fit a Lasso regression model in R using the glmnet package using one choice of alpha. Report the test error.

```
set.seed(123)
# Fit Lasso model
lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = 5)
predictions_lasso <- predict(lasso_model, s = 5, newx = x_test)

test_error_lasso <- mean((predictions_lasso - y_test) ^ 2)
test_error_lasso
```

```
## [1] 160.1174
```

Phase 1: Question 2

Lasso Regression Test Error: 160.1174

3. Perform Ridge regression on the same dataset using one choice of alpha. Report the test error.

```
set.seed(123)
# Fit Ridge model
ridge_model <- glmnet(x_train, y_train, alpha = 0, lambda = 5)
predictions_ridge <- predict(ridge_model, s = 5, newx = x_test)
test_error_ridge <- mean((predictions_ridge - y_test) ^ 2)
test_error_ridge
```

```
## [1] 159.1962
```

Phase 1: Question 3

Ridge regression Test Error: 159.1962

4. Now fit an Elastic Net model to the data using your own choice of hyper parameters. Report the test error.

```
set.seed(123)
# Fit Elastic Net model
elastic_net_model <- glmnet(x_train, y_train, alpha = 0.5, lambda =5) #in with out cv model using null
predictions_elastic_net <- predict(elastic_net_model, s =5, newx = x_test)
test_error_elastic_net <- mean((predictions_elastic_net - y_test) ^ 2)
test_error_elastic_net
```

```
## [1] 160.1174
```

Phase 1: Question 4

Elastic Net model Test Error: 160.1174

5. Use cross-validation to select optimal values of alpha and or lambda in each of the methods in 2-4. Report the optimal hyper parameter values you used for these methods in a Table.

```
set.seed(123)
##########     RIDGE REGRESSION    ####################
cv.out <- cv.glmnet(x_train,  y_train, alpha = 0,
                    nfolds=10,nlambda=100)
#value of lambda that results in smallest cross validation error
bestlam_rgd <- cv.out$lambda.min
bestlam_rgd
```

```
## [1] 10.80299
```

```
#Fit model with best lambda
ridge.mod <- glmnet(x_train, y_train, alpha = 0,
                    lambda = bestlam_rgd)
#make predictions on test set
ridge.pred <- predict(ridge.mod, s = bestlam_rgd,
                      newx = x_test)
# calculate test error
ridge.error=mean((ridge.pred - y_test )^2)
ridge.error
```

```
## [1] 158.1629
```

```
set.seed(123)
############   LASSO REGRESSION    ##################
###cross validation to get the test error
cv.out <- cv.glmnet(x_train, y_train, alpha = 1,nlambda=100)
bestlam_lasso <- cv.out$lambda.min
bestlam_lasso
```

```
## [1] 0.212067
```

```
lasso.mod <- glmnet(x_train,  y_train, alpha = 1,
                    lambda = bestlam_lasso)
lasso.pred <- predict(lasso.mod, s = bestlam_lasso,
                      newx = x_test)
lasso.error=mean((lasso.pred - y_test)^2)
lasso.error
```

```
## [1] 160.0181
```

```
set.seed(123)
# Define the grid for alpha and lambda
alpha_grid <- seq(0, 1, by = 0.001)
lambda_grid <- 10^seq(10, -2, length = 100)
cv_errors <- matrix(NA, length(alpha_grid), 2)

# Loop over alpha values
set.seed(1)
for(i in 1:length(alpha_grid)){
  alpha <- alpha_grid[i]

  # Perform cross-validation
  cv_model <- cv.glmnet(x_train, y_train, alpha=alpha, lambda=lambda_grid)

  # Store the best lambda and corresponding CV error for this alpha
  cv_errors[i,] <- c(cv_model$lambda.min, min(cv_model$cvm))
}

# Identify the best alpha based on CV error
best_index <- which.min(cv_errors[,2])
best_alpha <- alpha_grid[best_index]
best_lambda <- cv_errors[best_index, 1]
best_lambda
```

```
## [1] 0.2848036
```

```
elastic.mod<- glmnet(x_train, y_train, alpha=best_alpha,
                     lambda=best_lambda)
elastic.pred <- predict(elastic.mod, s = best_lambda,
                      newx = x_test)
elastic.error=mean((elastic.pred - y_test)^2)#test error
elastic.error
```

```
## [1] 159.8299
```

```
Methods = c("Lasso Regression", "Ridge Regression", "Elastic Net")
alpha = c(1, 0, best_alpha)
lambda = c(bestlam_lasso, bestlam_rgd, best_lambda)
```

```
hyper_parameter_table = data.frame(Methods, alpha, lambda)
kable(hyper_parameter_table, caption = "Table for hyperparameter")
```

Table 1: Table for hyperparameter

Methods	alpha	lambda
Lasso Regression	1.00	0.2120670
Ridge Regression	0.00	10.8029850
Elastic Net	0.77	0.2848036

6. Tabulate the test error for each of these models in 5 and compare with their corresponding models you fit in 2,3,4. What does this tell you about the model's performance?

```
set.seed(123)
library(knitr)
Models = c("Lasso Regression", "Ridge Regression", "Elastic Net")
orginal_test_error = c(test_error_lasso, test_error_ridge, test_error_elastic_net)
cv_test_error = c(lasso.error, ridge.error, elastic.error)
error_table = data.frame(Models, orginal_test_error, cv_test_error)
kable(error_table, caption = "Table for test error")
```

Table 2: Table for test error

Models	orginal_test_error	cv_test_error
Lasso Regression	160.1174	160.0181
Ridge Regression	159.1962	158.1629
Elastic Net	160.1174	159.8299

Phase 1: Question 6

Model Performance:

- The models tend to perform slightly better after optimization.

7. For the models in 5, which one will you choose as the final model based on the test errors?

Phase 1: Question 7

Model Selection:

- Ridge Regression: As it has the lowest optimized test error, it would be the best choice.

8. Describe your next steps in the modeling process now that you have selected your final model from 7.

Phase 1: Question 8

Model Selection:

final model is the one that is fitted with optimal lambda and optimal alpha, in glmnet() function. Once we choose a particular model as best model or final model, we can fit that model with full data, because we have chosen that model as best predictor model.

9. Based on the final model's output, which factors are most predictive of absenteeism in the workplace? How did you decide on those features?

```
# Prepare matrix of predictors and response variable for glmnet
x <- model.matrix(Absenteeism_time_in_hours ~ ., data = Absenteeism_data_lm)[,-1]
y <- Absenteeism_data_lm$Absenteeism_time_in_hours

#Final_model
final_model <- glmnet(x, y, alpha=0, lambda=bestlam_rgd) # final model is done in the whole data set wi

# Extract coefficients at the optimal lambda
optimal_coefs <- coef(final_model)

# Extract feature names
feature_names <- rownames(optimal_coefs)[-1]  # Exclude the intercept

# Extract coefficients
coefficients <- as.vector(optimal_coefs[-1])  # Exclude the intercept

# Create a data frame to store feature names and coefficients
feature_importance <- data.frame(
  Feature = feature_names,
  Coefficient = coefficients
)

# Sort features by absolute coefficient magnitude
feature_importance <- feature_importance[order(abs(feature_importance$Coefficient), decreasing = TRUE),

# Print feature importance
print(feature_importance)
```

```
##                          Feature   Coefficient
## 7                Month_of_absence7  2.249761e+00
## 15                Day_of_the_week5 -1.742784e+00
## 27                     Education4 -1.719652e+00
## 16                Day_of_the_week6 -1.540963e+00
## 1                Month_of_absence1 -1.308155e+00
## 12              Month_of_absence12  1.279654e+00
## 2                Month_of_absence2 -1.084448e+00
## 3                Month_of_absence3  1.057482e+00
## 5                Month_of_absence5 -1.022993e+00
## 10              Month_of_absence10 -1.011096e+00
## 25                      Education2 -9.382761e-01
## 26                      Education3 -7.984942e-01
## 4                Month_of_absence4  7.722891e-01
## 8                Month_of_absence8 -7.663129e-01
## 9                Month_of_absence9 -7.353450e-01
## 28                            Son  6.432886e-01
## 18                      Seasons3  5.146051e-01
## 11              Month_of_absence11  4.900636e-01
## 6                Month_of_absence6  2.572385e-01
```

```
## 19                          Seasons4 -2.447431e-01
## 13                 Day_of_the_week3  2.118682e-01
## 17                          Seasons2 -1.833652e-01
## 31                            Height  1.591602e-01
## 29                               Pet -1.181306e-01
## 14                 Day_of_the_week4 -1.148302e-01
## 32                  Body_mass_index -1.064389e-01
## 23                               Age  7.412588e-02
## 22                      Service_time  3.836869e-02
## 21 Distance_from_Residence_to_Work -2.987886e-02
## 20           Transportation_expense  4.624162e-03
## 30                            Weight  1.860871e-05
## 24       Work_load_Average_in_days  3.931953e-06
```

Phase 1: Question 9

Factors that affect Absenteeism:

- We can conclude that the Month of Absence, Day of Week, and Education are some important factors that influence Absenteeism.
- These features were decided based on the coefficients of the chosen final model.

10. Phase 1 only: Discuss the potential implications of your findings for the management of the company.

Phase 1: Question 10

Potential Implications of Findings:

- Month of Year: As certain months of the year are more correlated to absenteeism, management can explore the root cause of this and provide aids to overcome issues faced during those months.
- Day of the Week: The management can look into ways to reduce absenteeism in the beginning of the week by providing incentives that encourage people to go to work.

11. Phase 1 only: How might the company use the insights from your final model to reduce absenteeism rates?

Phase 1: Question 11

Company Usage of Insights:

- The company can either choose to aid the employees using this information by overcoming the obstacles that they face.
- Or the company can choose to encourage further attendance by provide incentives.

12. Phase 2 only: For the regularization methods with logistic regression, you can compare these 3 models under AUC, F1 score, etc... Tabulate these metrics under the 3 models and comment on which one you would choose based on these metrics.

```
# Splitting the data into training and test sets
set.seed(123) # Setting a seed for reproducibility
splitIndex = sample(1:nrow(Absenteeism_data_lm), 0.8 * nrow(Absenteeism_data_glm))
```

```
trainData_glm <- Absenteeism_data_glm[splitIndex,]
testData_glm <- Absenteeism_data_glm[-splitIndex,]

# Prepare matrix of predictors and response variable for glmnet
x_train_glm <- model.matrix(Absenteeism ~ ., data = trainData_glm)[, -1] # -1 to exclude the intercept
y_train_glm <- trainData_glm$Absenteeism

x_test_glm <- model.matrix(Absenteeism ~ ., data = testData_glm)[, -1]
y_test_glm <- testData_glm$Absenteeism
```

Phase 2: Question 2. 1

We did the spiliting of the data as 80% for the training of the data and remaining 20% for the testing of the data

```
set.seed(123)
# Fit Lasso model
lasso_model_glm <- glmnet(x_train_glm, y_train_glm, alpha = 1, lambda = 10, family = "multinomial", sta
predictions_lasso_glm <- predict(lasso_model_glm, s = 5, newx = x_test_glm,  type = "class")

test_error_lasso_glm <- mean(predictions_lasso_glm!=y_test_glm)
test_error_lasso_glm
```

```
## [1] 0.06081081
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

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

```
conf_matrix = confusionMatrix(factor(predictions_lasso_glm), factor(y_test_glm))
conf_matrix$table
```

```
##            Reference
## Prediction High Low Moderate
##   High        0   0        0
##   Low         3 139        6
##   Moderate    0   0        0
```

```
# Extracting the metrics
accuracy_lasso = conf_matrix$overall['Accuracy']

# Print the metrics
print(paste("Accuracy:", accuracy_lasso))
```

```
## [1] "Accuracy: 0.939189189189189"
```

Phase 2: Question 2. 2

We found the test error for Lasso model as: 0.06081081

```
set.seed(123)
# Fit Ridge model
```

```r
ridge_model_glm <- glmnet(x_train_glm, y_train_glm, alpha = 0, lambda = 10, family = "multinomial")
predictions_ridge_glm <- predict(ridge_model_glm, s = 10, newx = x_test_glm, type = "class")
test_error_ridge_glm <- mean(predictions_ridge_glm!= y_test_glm)
test_error_ridge_glm
```

```
## [1] 0.06081081
```

```r
conf_matrix = confusionMatrix(factor(predictions_ridge_glm), factor(y_test_glm))
conf_matrix$table
```

```
##            Reference
## Prediction High Low Moderate
##   High        0   0        0
##   Low         3 139        6
##   Moderate    0   0        0
```

Phase 2: Question 2. 3

We found the test error for Ridge model as: 0.09459459

```r
set.seed(123)
# Fit Elastic Net model
elastic_net_model_glm <- glmnet(x_train_glm, y_train_glm, alpha = 0.5, lambda = 10, family = "multinomi
predictions_elastic_net_glm <- predict(elastic_net_model_glm, s = 10, newx = x_test_glm, type = "class"
test_error_elastic_net_glm <- mean(predictions_elastic_net_glm != y_test_glm)
test_error_elastic_net_glm
```

```
## [1] 0.06081081
```

```r
conf_matrix_lasso = confusionMatrix(factor(predictions_elastic_net_glm), factor(y_test_glm))
conf_matrix_lasso$table
```

```
##            Reference
## Prediction High Low Moderate
##   High        0   0        0
##   Low         3 139        6
##   Moderate    0   0        0
```

- Here if we put lambda=0 then we will get different error, but putting lambda=0 means not doing regularization on the dataset, which is worthless as we have to do regularization. So when we put any other value of lambda it will generate the same error in all the models. this is not the problem of the models, the problem is with the dataset itself. Since the dataset is the imbalance dataset so models are mostly trying to predict everything as low when done regularization. (Hence I believe I will get the same test error after cross-validation also)

Phase 2: Question 2. 4

We found the test error for Elastic net model as: 0.06081081

```r
set.seed(123)
##########     RIDGE REGRESSION    ####################
cv.out_glm <- cv.glmnet(x_train_glm,  y_train_glm, alpha = 0,
                    nfolds=10,nlambda=1000,  family = "multinomial")
#value of lambda that results in smallest cross validation error
```

```
bestlam_rgd_glm <- cv.out_glm$lambda.min
bestlam_rgd_glm
```

```
## [1] 0.05056404
```

```
#Fit model with best lambda
ridge.mod_glm <- glmnet(x_train_glm, y_train_glm, alpha = 0,
                    lambda = bestlam_rgd_glm, family = "multinomial")
#make predictions on test set
ridge.pred_glm <- predict(ridge.mod_glm, s = bestlam_rgd,
                      newx = x_test_glm, type = "class")
ridge.pred <- predict(ridge.mod_glm, s = bestlam_rgd,
                      newx = x_test_glm, type = "class")
# calculate test error
ridge.error_glm=mean(ridge.pred_glm != y_test_glm )
ridge.error_glm
```

```
## [1] 0.06081081
```

```
conf_matrix = confusionMatrix(factor(ridge.pred), factor(y_test_glm))
conf_matrix$table
```

```
##           Reference
## Prediction High Low Moderate
##   High        0   0        0
##   Low         3 139        6
##   Moderate    0   0        0
```

```
set.seed(123)
############   LASSO REGRESSION    ##################
###cross validation to get the test error
cv.out_lasso_glm <- cv.glmnet(x_train_glm, y_train_glm, alpha = 1,nlambda=100, family = "multinomial")
bestlam_lasso_glm <- cv.out_lasso_glm$lambda.min
bestlam_lasso_glm
```

```
## [1] 0.006196006
```

```
lasso.mod_glm <- glmnet(x_train_glm,  y_train_glm, alpha = 1,
                    lambda = bestlam_lasso_glm,  family = "multinomial" )
lasso.pred_glm <- predict(lasso.mod_glm, s = bestlam_lasso,
                      newx = x_test_glm, type = "class")
lasso.error_glm=mean(lasso.pred_glm !=y_test_glm)
lasso.error_glm
```

```
## [1] 0.06081081
```

```
conf_matrix = confusionMatrix(factor(lasso.pred_glm), factor(y_test_glm))
conf_matrix$table
```

```
##           Reference
## Prediction High Low Moderate
##   High        0   0        0
##   Low         3 139        6
##   Moderate    0   0        0
```

```
set.seed(123)
# Define the grid for alpha and lambda
alpha_grid <- seq(0, 1, by = 0.01)  # I reduce the sequence as my computer was not working for larger s
```

```
lambda_grid <- 10^seq(10, -2, length = 3)
cv_errors <- matrix(NA, length(alpha_grid), 2)

# Loop over alpha values
set.seed(1)
for(i in 1:length(alpha_grid)){
  alpha <- alpha_grid[i]

  # Perform cross-validation
  cv_model <- cv.glmnet(x_train_glm, y_train_glm, alpha=alpha, lambda=lambda_grid, family = "multinomia

  # Store the best lambda and corresponding CV error for this alpha
  cv_errors[i,] <- c(cv_model$lambda.min, min(cv_model$cvm))
}

# Identify the best alpha based on CV error
best_index <- which.min(cv_errors[,2])
best_alpha <- alpha_grid[best_index]
best_lambda <- cv_errors[best_index, 1]
best_lambda
```

```
## [1] 0.01
```

```
elastic.mod<- glmnet(x_train_glm, y_train_glm, alpha=best_alpha,
                     lambda=best_lambda, family = "multinomial")
elastic.pred <- predict(elastic.mod, s = best_lambda,
                        newx = x_test_glm, type = "class")
elastic.error=mean(elastic.pred !=y_test_glm)#test error
elastic.error
```

```
## [1] 0.06081081
```

```
conf_matrix = confusionMatrix(factor(elastic.pred), factor(y_test_glm))
conf_matrix$table
```

```
##            Reference
## Prediction High Low Moderate
##   High        0   0        0
##   Low         3 139        6
##   Moderate    0   0        0
```

```
Methods = c("Lasso Regression", "Ridge Regression", "Elastic Net")
alpha = c(1, 0, best_alpha)
lambda_values = c(bestlam_lasso_glm, bestlam_rgd_glm, best_lambda)
hyper_parameter_table = data.frame(Methods, alpha, lambda_values)
kable(hyper_parameter_table, caption = "Table for Hyperparameters")
```

Table 3: Table for Hyperparameters

Methods	alpha	lambda_values
Lasso Regression	1.00	0.006196
Ridge Regression	0.00	0.050564
Elastic Net	0.58	0.010000

Phase 2: Question 2. 5

The hyperparameter table is:

```
set.seed(123)
library(knitr)
Models = c("Lasso Regression", "Ridge Regression", "Elastic Net")
orginal_test_error = c(test_error_lasso_glm, test_error_ridge_glm, test_error_elastic_net_glm)
cv_test_error = c(lasso.error_glm, ridge.error_glm, elastic.error)
error_table = data.frame(Models, orginal_test_error, cv_test_error)
kable(error_table, caption = "Test Error Table")
```

Table 4: Test Error Table

Models	orginal_test_error	cv_test_error
Lasso Regression	0.0608108	0.0608108
Ridge Regression	0.0608108	0.0608108
Elastic Net	0.0608108	0.0608108

- Here we go, we got the same error after cross-validation indication dataset is imbalanced and model are mostly predicting low for every data point when done any regularization.

Phase 2: Question 2 5 7

- Since the data set is imbalanced, we are getting same test error rate from all the models. So with this dataset any model you choose from given is same. but if dataset not not imbalance we can say something about which model will be much better. I am choosing lasso as my model, because mostly lasso works best for this type of dataset.

Observation

If we look at the confusion matrix for our regression models, they all are exactly same and we calculate precision, recall, accuracy and test error for the data from the confusion matrix so they all are going to be the same for all model, this is happening because of Imbalance data. Same will be the case for F1. we can check this by following table

```
library(caret)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(glmnet)

actual <- factor(y_test_glm)

# Convert predictions to factors
predictions_lasso <- factor(predictions_lasso_glm, levels = levels(actual))
```

```r
predictions_ridge <- factor(predictions_ridge_glm, levels = levels(actual))
predictions_elastic_net <- factor(predictions_elastic_net_glm, levels = levels(actual))

library(caret)
library(glmnet)

# Modify the calculateF1Scores function to also calculate accuracy
calculateMetrics <- function(actual, predicted) {
  conf_mat <- confusionMatrix(predicted, actual)
  f1_scores <- sapply(1:nrow(conf_mat$table), function(i) {
    precision <- conf_mat$table[i, i] / sum(conf_mat$table[, i])
    recall <- conf_mat$table[i, i] / sum(conf_mat$table[i, ])
    f1_score <- 2 * ((precision * recall) / (precision + recall))
    return(f1_score)
  })
  accuracy <- sum(diag(conf_mat$table)) / sum(conf_mat$table)
  list(macroF1 = mean(f1_scores, na.rm = TRUE), microF1 = sum(diag(conf_mat$table)) / sum(conf_mat$table
}

# Calculate metrics
metrics_lasso <- calculateMetrics(actual, predictions_lasso)
metrics_ridge <- calculateMetrics(actual, predictions_ridge)
metrics_elastic_net <- calculateMetrics(actual, predictions_elastic_net)

# Create a table to summarize the results, now including accuracy
results_table <- data.frame(
  Model = c("Lasso", "Ridge", "Elastic Net"),
  F1_Score_Macro = c(metrics_lasso$macroF1, metrics_ridge$macroF1, metrics_elastic_net$macroF1),
  F1_Score_Micro = c(metrics_lasso$microF1, metrics_ridge$microF1, metrics_elastic_net$microF1),
  Accuracy = c(metrics_lasso$accuracy, metrics_ridge$accuracy, metrics_elastic_net$accuracy)
)
print(results_table)
```

```
##         Model F1_Score_Macro F1_Score_Micro  Accuracy
## 1       Lasso      0.9686411      0.9391892 0.9391892
## 2       Ridge      0.9686411      0.9391892 0.9391892
## 3 Elastic Net      0.9686411      0.9391892 0.9391892
```

```r
# Prepare matrix of predictors and response variable for glmnet
x <- model.matrix(Absenteeism ~ ., data = Absenteeism_data_glm)[,-1]
y <- Absenteeism_data_glm$Absenteeism



#Final_model
final_model_glm <- glmnet(x, y, alpha=1, lambda=bestlam_lasso_glm, family = "multinomial") # final mode

# Extract coefficients at the optimal lambda
optimal_coefs <- coef(final_model_glm )[-1]
optimal_coefs
```

```
## $Low
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                          s0
```

```
##                                  8.943770976
## Month_of_absence1                0.174977663
## Month_of_absence2                0.405016689
## Month_of_absence3                .
## Month_of_absence4               -0.282638346
## Month_of_absence5                .
## Month_of_absence6                .
## Month_of_absence7                .
## Month_of_absence8                .
## Month_of_absence9                .
## Month_of_absence10               0.183628328
## Month_of_absence11               .
## Month_of_absence12               .
## Day_of_the_week3                 .
## Day_of_the_week4                 .
## Day_of_the_week5                 1.380573936
## Day_of_the_week6                 0.632945499
## Seasons2                         .
## Seasons3                        -0.140719481
## Seasons4                         .
## Transportation_expense           .
## Distance_from_Residence_to_Work  .
## Service_time                    -0.004766046
## Age                             -0.018216325
## Work_load_Average_in_days        .
## Education2                       .
## Education3                       0.054605763
## Education4                       .
## Son                             -0.016340660
## Pet                              0.105736147
## Weight                           .
## Height                          -0.047280187
## Body_mass_index                  0.005877702
##
## $Moderate
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                           s0
##                                 -5.501480e+00
## Month_of_absence1                .
## Month_of_absence2                .
## Month_of_absence3                .
## Month_of_absence4                .
## Month_of_absence5               -3.154814e-01
## Month_of_absence6                .
## Month_of_absence7                .
## Month_of_absence8                .
## Month_of_absence9                4.816432e-01
## Month_of_absence10              -3.927922e-01
## Month_of_absence11               5.117170e-02
## Month_of_absence12               .
## Day_of_the_week3                -6.811407e-01
## Day_of_the_week4                 .
## Day_of_the_week5                 .
## Day_of_the_week6                -1.683153e-01
```

```
## Seasons2                          .
## Seasons3                          .
## Seasons4                          .
## Transportation_expense            1.898018e-03
## Distance_from_Residence_to_Work   .
## Service_time                      .
## Age                               .
## Work_load_Average_in_days         9.205260e-06
## Education2                        .
## Education3                        .
## Education4                        .
## Son                               .
## Pet                               .
## Weight                            .
## Height                            .
## Body_mass_index                   .
```

Phase 2: Question 9

Factors that affect Absenteeism:

- We can conclude from the final model that low absenteeism is mainly because of month of absence, days of week, (and mildly because of seasons, service time, age, education, son, pet, height and body_mass_index) and Moderate absenteeism is mainly because of month of year, days of week, transportation expense and work_load_average_in_days, taking high absenteeism as the base line. So we can still say that, Days of the week, months of year and Education are some important factors that influence Absenteeism.
- These features were decided based on the coefficients of the chosen final model.