CptS 575_Assignment 02

Sajjad Uddin Mahmud | WSU ID: 011789534

2022-09-09

Solution of Problem 01

1(a) Reading .csv file

```
# Reading red wine quality dataset in a dataframe named Redwine
Redwine <- read.csv("winequality-red.csv")
```

1(b) Calculating median quality and mean alcohol level

```
# Finding median
Median_Quality <- median(Redwine$quality)

# Finding mean
Mean_Alcohol_Level <- mean(Redwine$alcohol)
Mean_Alcohol_Level <- format(round(Mean_Alcohol_Level,2)) # Formatting decimal places
```

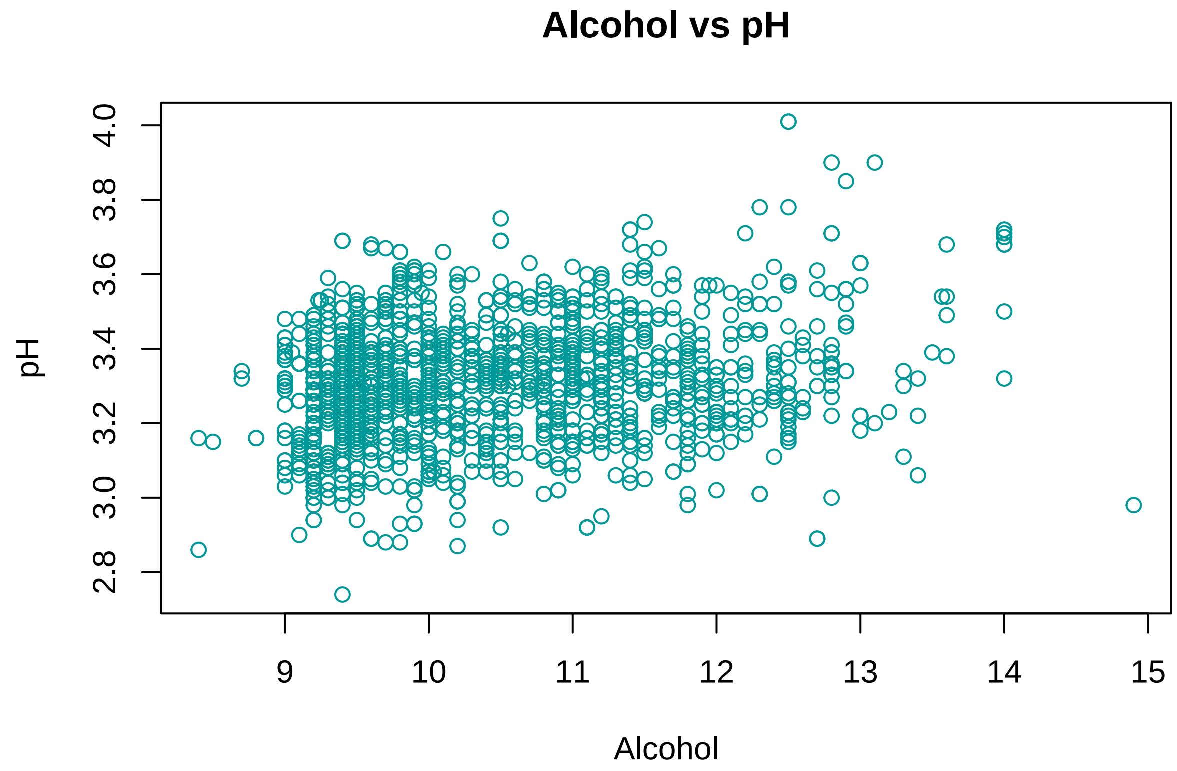
The median quality of all the wine samples is **6**.

The mean alcohol level of all the wine samples is **10.42**.

1(c) Producing scatter plot

```
# Getting the input parameters
x <- Redwine$alcohol
y <- Redwine$pH

# Creating the plot
plot(x, y, xlab="Alcohol", ylab="pH", main="Alcohol vs pH", col=c("#009999"))
```



1(d) Producing box plot

```
# IFELSE: Alcohol level separation
ALevel <- ifelse(Redwine$alcohol > 10.2, "High", "Medium")
table(ALevel)
```

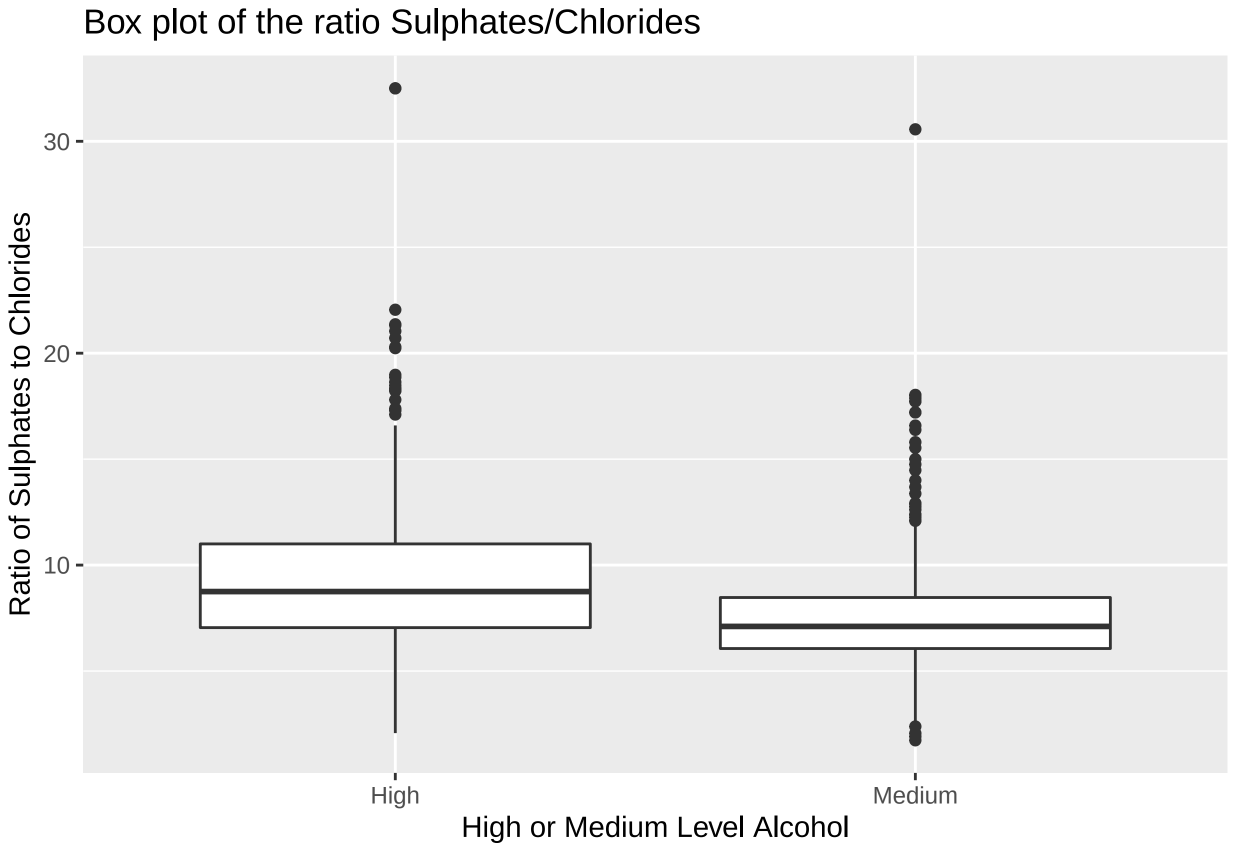
```
## ALevel
##   High Medium
##    757    842
```

```
Ratio_Sulphates_Chlorides <- Redwine$sulphate/Redwine$chlorides

Redwine_New <- cbind(Redwine,ALevel,Ratio_Sulphates_Chlorides)

# Creating box plots
library(ggplot2)

ggplot(data=Redwine_New,aes(x=ALevel,y=Ratio_Sulphates_Chlorides)) +
  geom_boxplot() +
  labs(title = "Box plot of the ratio Sulphates/Chlorides",
       x = "High or Medium Level Alcohol",
       y = "Ratio of Sulphates to Chlorides")
```

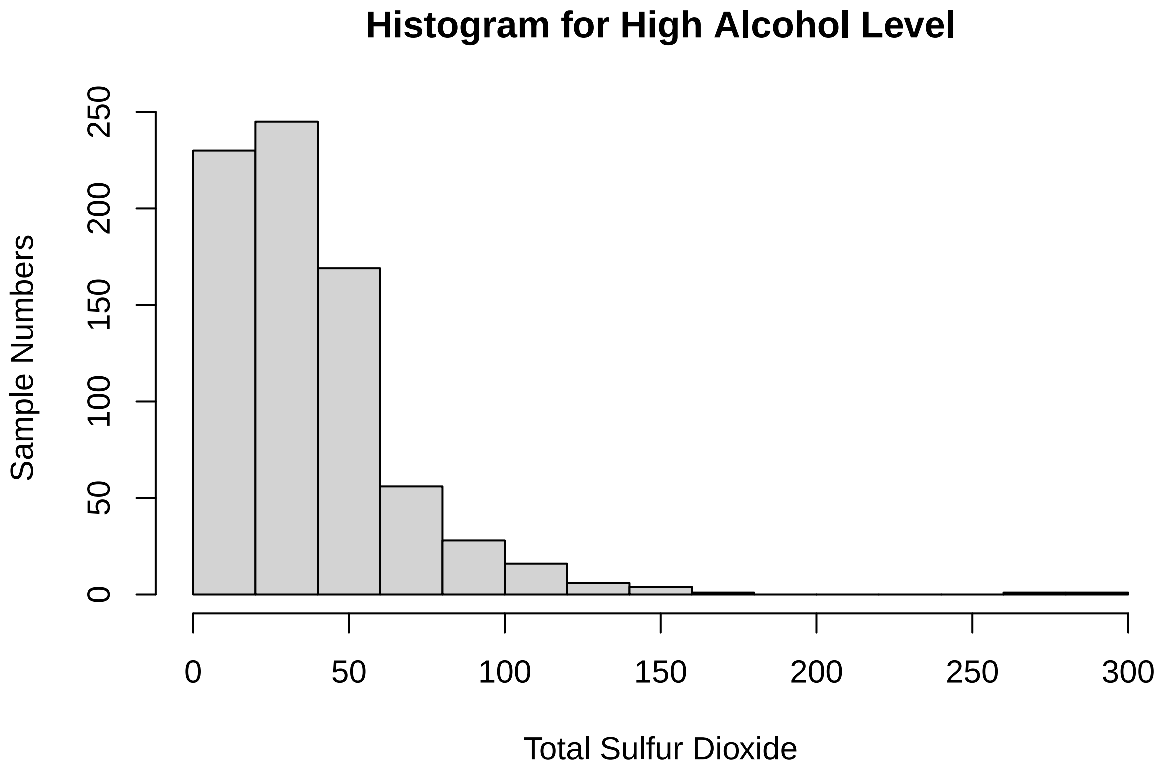


The number of samples in the High category is **757**

1(e) Producing histogram

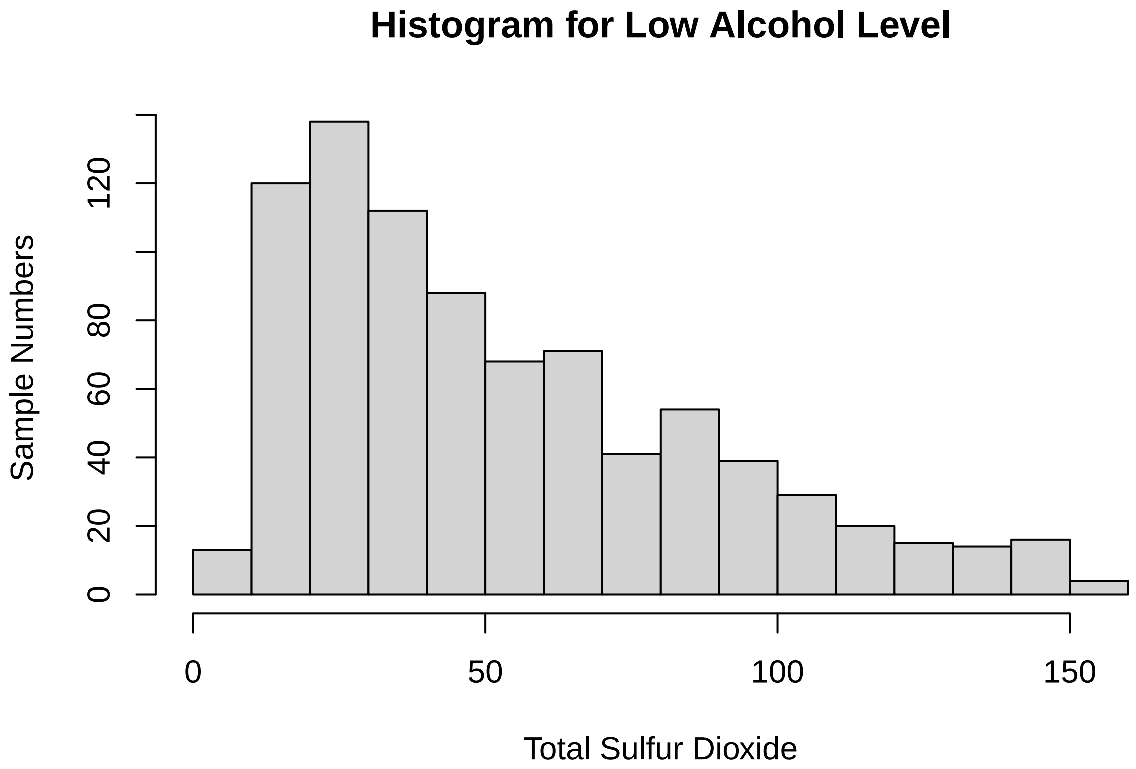
```r
# Creating data frame of high alcohol level
ALevel_High <- Redwine[which(Redwine$alcohol>10.2),]

# Plotting histogram
hist(ALevel_High$total_sulfur_dioxide, main="Histogram for High Alcohol Level",
     xlab="Total Sulfur Dioxide", ylab="Sample Numbers")
```



```
# Creating data frame of low alcohol level
ALevel_Low <- Redwine[which(Redwine$alcohol<=10.2),]

# Plotting histogram
hist(ALevel_Low$total_sulfur_dioxide, main="Histogram for Low Alcohol Level",
    xlab="Total Sulfur Dioxide", ylab="Sample Numbers")
```



1(f) Hypothesis

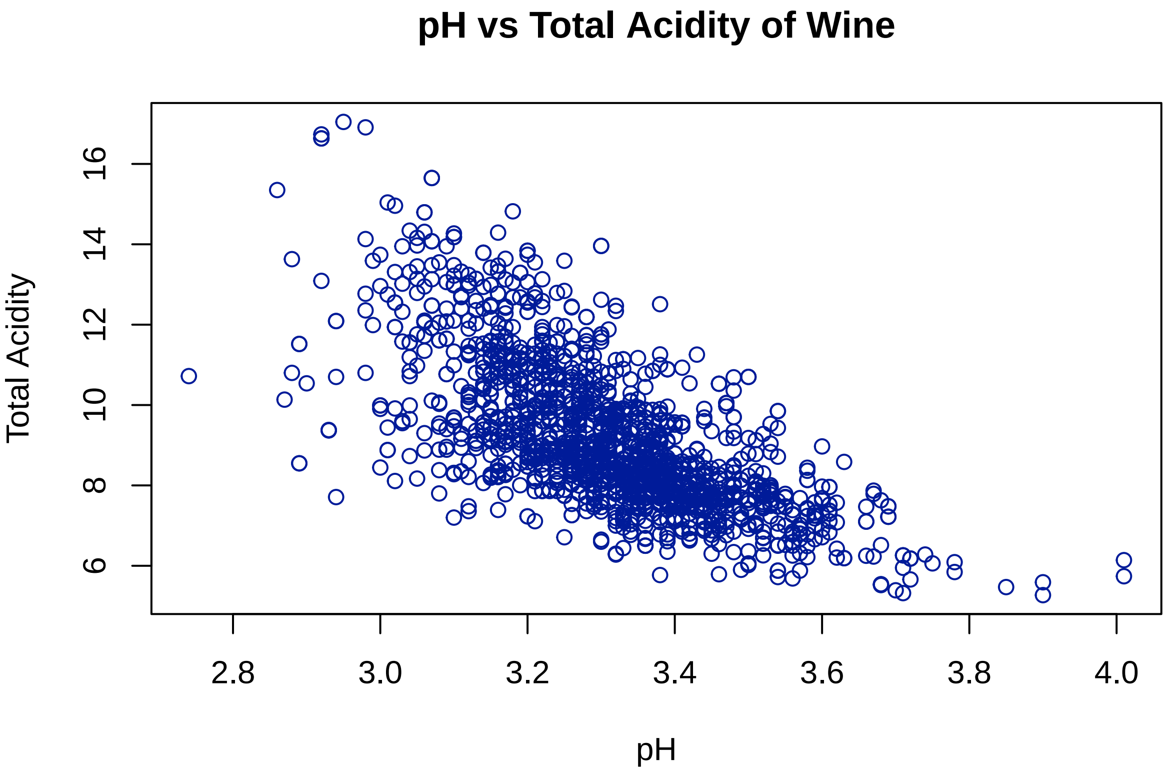
Plot 1: pH vs Acidity of Wine

We all know that less value of pH means more acidic. The pH vs total acidity of wine has been plotted below. To get the total acidity, I have taken the sum of fixed acidity, volatile acidity and citric acid.

```
# Measuring total acidity
Total_Acid <- Redwine$fixed_acidity + Redwine$volatile_acidity + Redwine$citric_acid

# Adding a new column in the data frame for total acidity
Redwine_Acid <- Redwine
Redwine_Acid$Total_Acidity <- Total_Acid

# Plotting the relation
plot(x=Redwine_Acid$pH, y=Redwine_Acid$Total_Acidity,
     xlab="pH", ylab="Total Acidity",
     main="pH vs Total Acidity of Wine", col=c("#001B99"))
```

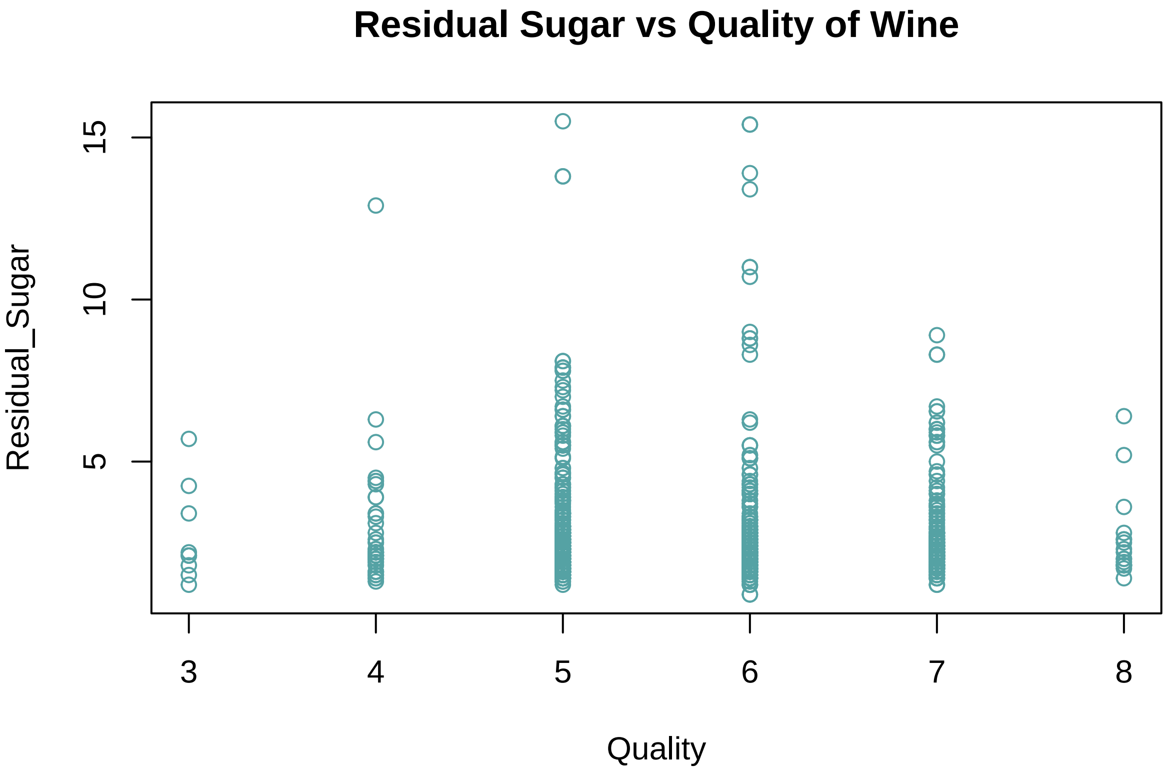


As can be seen in the preceding graph, wine with higher pH values are less acidic than those with lower values. However, there is a low acidic value when pH is less than 2. This could be an error in the data set or some other ingredients may affect the acidity for that particular wine case.

Plot 2: Sugar vs Wine Quality

Now we want to see whether the better quality wines are also sweeter.

```
# Plotting the relation
plot(x=Redwine_Acid$quality, y=Redwine_Acid$residual_sugar,
     xlab="Quality", ylab="Residual_Sugar",
     main="Residual Sugar vs Quality of Wine", col=c("#55A2A4"))
```



From the graph it can be seen than the finer quality wine does not necessarily mean that it will be sweeter. Most of the sweet wines are belong in the average quality of wine as per this data set.

Solution of Problem 02

2(a) Quantitative and qualitative predictors

Quantitative predictors:

- FFMC
- DMC
- DC
- ISI
- temp
- RH
- wind
- rain and
- area

Qualitative predictors:

- X
- Y
- month
- day

2(b) Range, mean and standard deviation

```r
library(tidyverse)

# Reading data set
ForestFires <- read.csv("forestfires.csv")

# Declaring a function that will calculate range, mean and standard deviation
Range_Mean_SD <- function(Range_Mean_SD){

  # Variable initialization
  Range_value <- vector()
  Mean_Value <- vector()
  SD_Value <- vector()
  Answer_Table <- data.frame()

  # Selecting quantitative predictors
  ForestFires_Quantitative <- subset(Range_Mean_SD, select = c(FFMC:area))

  # Loop: Calculating range, mean and SD
  for (name in colnames(ForestFires_Quantitative)){
    Range_value <- range(ForestFires_Quantitative[[name]])
    Mean_Value <- format(round(mean(ForestFires_Quantitative[[name]]),2))
    SD_Value <- format(round(sd(ForestFires_Quantitative[[name]]),2))

    Answer_Table <- rbind(Answer_Table, c(Range_value, Mean_Value, SD_Value))
  }
  # Giving row and column names
  rownames(Answer_Table) <- c("FFMC", "DMC", "DC", "ISI", "temp", "RH", "wind", "rain", "area")
  colnames(Answer_Table) <- c("Range_L", "Range_U", "Mean", "SD")

  # Displaying table
  print(Answer_Table)
}

# Calling the function to calculate range, mean and SD
Range_Mean_SD(ForestFires)
```

```
##      Range_L Range_U   Mean    SD
## FFMC    18.7    96.2  90.64  5.52
## DMC      1.1   291.3 110.87 64.05
## DC         ?    97.8     NA    NA
## ISI        0    56.1   9.02  4.56
## temp     2.2    33.3  18.89  5.81
## RH        15     100  44.29 16.32
## wind     0.4     9.4   4.02  1.79
## rain       0     6.4   0.02   0.3
## area       0 1090.84  12.85 63.66
```

Here, Range_L is the lower bound of the range, Range_U is the upper bound of the range and SD is the standard deviation.

For DC (Drought code index) there is a data missing in the column and that is why the lower bound of the range, mean and standard deviation of DC cannot be calculated.

```
# Determining which day in the week has the highest number of fire
Fire_Day <- data.frame()

# Loop: Fire day
for (value in 1:length(ForestFires$area)){
  if (ForestFires$area[value]!=0){
    Fire_Day <- ForestFires$day[value]
  }
}
Max_Fire_Day <- names(which.max(table(Fire_Day)))
```

To determine which day in the week has the highest number of fire, at first the 0 areas (i.e. the measured area affected by fire is 0) are ignored and from the rest of the area, the days are observed.

Hence, the highest number of fire occurred in **sunday**.

2(c) Range, mean and standard deviation after removing the 40th through 80th (inclusive) observations

```
# Removing 40th through 80th (inclusive) observations
ForestFires_New <- ForestFires[-c(40:80),]

# Calling the function that will calculate range, mean and standard deviation
Range_Mean_SD(ForestFires_New)
```

```
##      Range_L Range_U   Mean    SD
## FFMC    18.7    96.2  90.66  5.68
## DMC      1.1   291.3 113.47 65.05
## DC         ?    97.8     NA    NA
## ISI        0    56.1   9.07  4.63
## temp     2.2    33.3  19.01  5.85
## RH        15     100  44.47 16.42
## wind     0.4     9.4   4.01   1.8
## rain       0     6.4   0.02  0.31
## area       0 1090.84  13.95 66.23
```

Here, Range_L is the lower bound of the range, Range_U is the upper bound of the range and SD is the standard deviation.

For DC (Drought code index) there is a data missing in the column and that is why the lower bound of the range, mean and standard deviation of DC cannot be calculated.

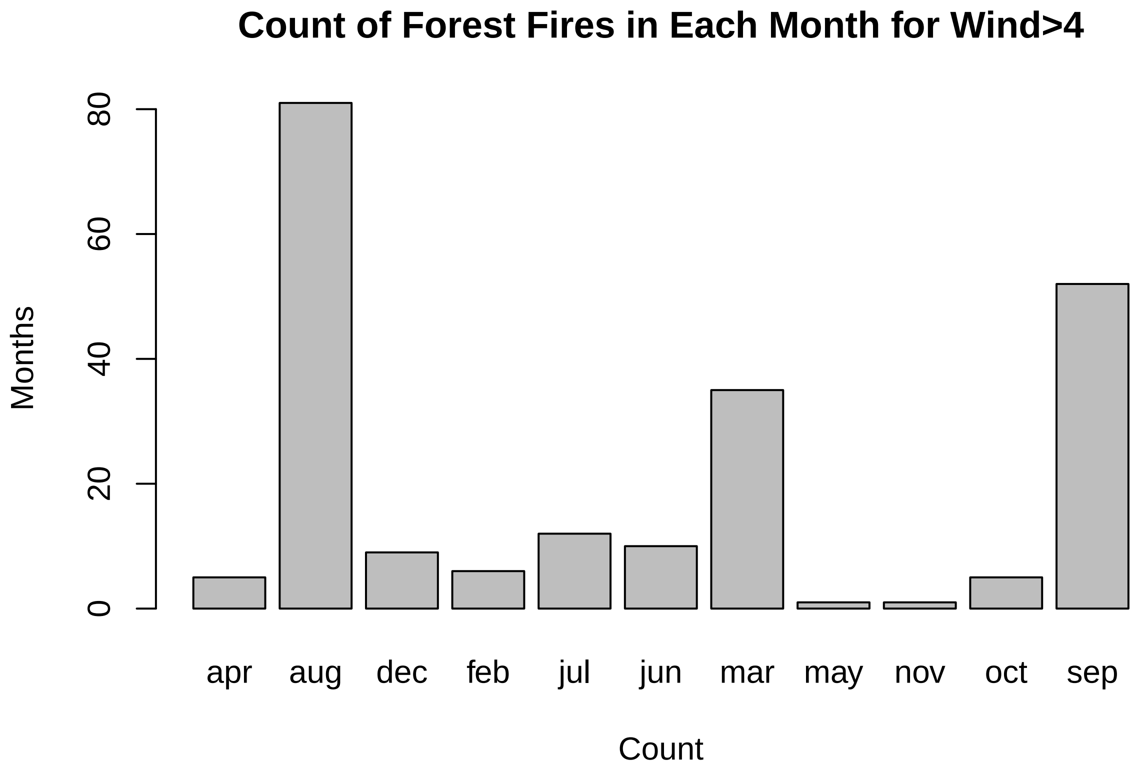
2(d) Bar plots of month vs forest fires

```
#
library(plyr)
library(dplyr)

# Creating dataframe for wind>4
ForestFires_New2 <- ForestFires[which(ForestFires$wind>4),]

# Counting months
Fire_Month <- count(ForestFires_New2, vars="month")

# Creating Bar plot
barplot(Fire_Month$freq, names.arg=Fire_Month$month,
        main="Count of Forest Fires in Each Month for Wind>4",
        xlab="Count", ylab="Months")
```



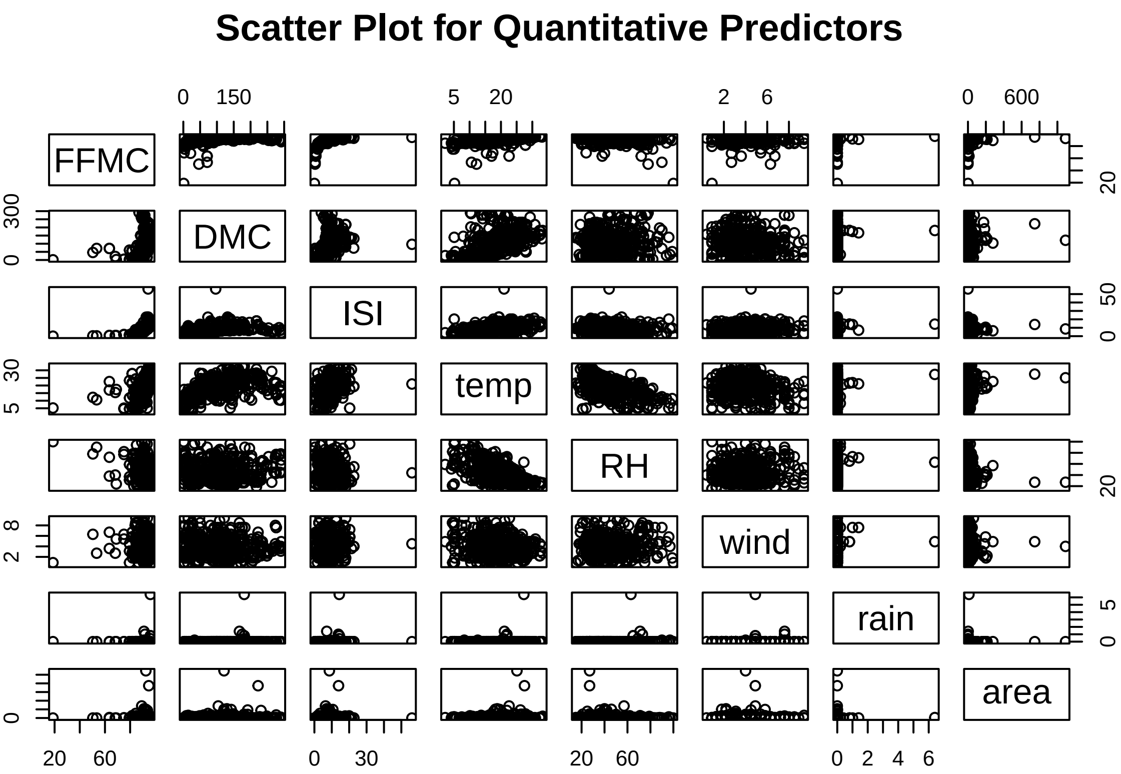
From the above bar plot, we can see that during **August** month, high wind forest fires are the most common.

2(e) Data set investigation

Scatter plots has been done only on quantitative predictors. However, as DC has a missing value, hence it is ignored in the scatter plot.

```
# Creating quantitative predictor data frame
ForestFires_Quantitative <- subset(ForestFires, select = c(FFMC,DMC,ISI,temp,RH,wind,rain,area))

# Creating scatter plot matrix
pairs(ForestFires_Quantitative, main="Scatter Plot for Quantitative Predictors")
```



```r
# Correlation matrix
Correlation_Matrix <- format(round(cor(ForestFires_Quantitative),3))
Correlation_Matrix
```

```
##      FFMC     DMC      ISI      temp     RH       wind     rain     area
## FFMC " 1.000" " 0.383" " 0.532" " 0.432" "-0.301" "-0.028" " 0.057" " 0.040"
## DMC  " 0.383" " 1.000" " 0.305" " 0.470" " 0.074" "-0.105" " 0.075" " 0.073"
## ISI  " 0.532" " 0.305" " 1.000" " 0.394" "-0.133" " 0.107" " 0.068" " 0.008"
## temp " 0.432" " 0.470" " 0.394" " 1.000" "-0.527" "-0.227" " 0.069" " 0.098"
## RH   "-0.301" " 0.074" "-0.133" "-0.527" " 1.000" " 0.069" " 0.100" "-0.076"
## wind "-0.028" "-0.105" " 0.107" "-0.227" " 0.069" " 1.000" " 0.061" " 0.012"
## rain " 0.057" " 0.075" " 0.068" " 0.069" " 0.100" " 0.061" " 1.000" "-0.007"
## area " 0.040" " 0.073" " 0.008" " 0.098" "-0.076" " 0.012" "-0.007" " 1.000"
```

2(f) Wind speed prediction

According to the findings of the Pearson Correlation analysis, the degree of correlation that exists between two variables can be positive, negative, or even zero, and its range is from -1 to +1. According to the correlation matrix that was just presented, we can see that the absolute value between wind and temperature is the largest, despite the fact that it has a negative correlation (-0.227). This indicates that there is an inverse relationship between wind speed and temperature. In addition, it is a well-known fact that when there is a strong breeze, the temperature drops. Therefore, if we want to forecast the wind, temperature can be a suitable indicator to use.