# CPTS 475 Homework 2

Andrew Balaschak

2023-09-08

# 1

## 1.a. Read the winequality-red.csv file into a dataframe named redwine.

```
redwine <- read.csv("winequality-red.csv")
```

## 1.b. Calculate the mean quality and mean alcohol for the dataset.

```
print(paste("Mean quality: ", round(mean(redwine$quality), 2)))
```

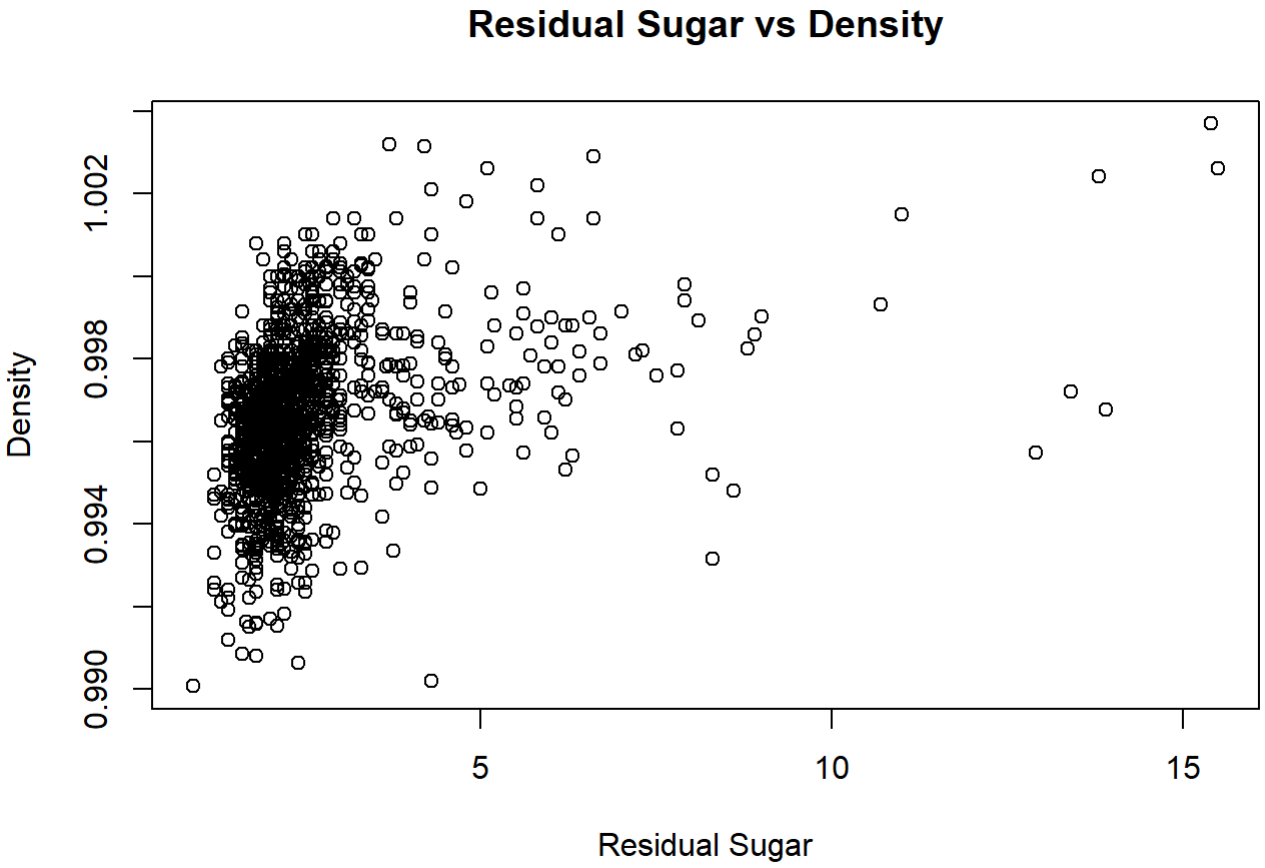
```
## [1] "Mean quality:  5.64"
```

```
print(paste("Median alcohol: ", round(median(redwine$alcohol), 2)))
```

```
## [1] "Median alcohol:  10.2"
```

## 1.c. Produce a scatterplot that shows the relationship between wine density and residual sugar.

```
plot(redwine$residual_sugar, redwine$density, xlab = "Residual Sugar", ylab = "Density", main =
"Residual Sugar vs Density")
```




Based on this plot, I don't see the residual sugar having any strong effect on the density, as there are wines with a broad range of densities for a similar amount of residual sugar.

# 1.d. Bin the alcohol variable into two categories, medium and high, based on whether or not the alcohol level exceeds 11. From there, calculate the sulphates to chrlorides ratio for each wine, and create a box plot that shows the difference between the ratio for medium and high alcohol wines.

Binning the wines based on alcohol content, and adding the column to the dataframe.

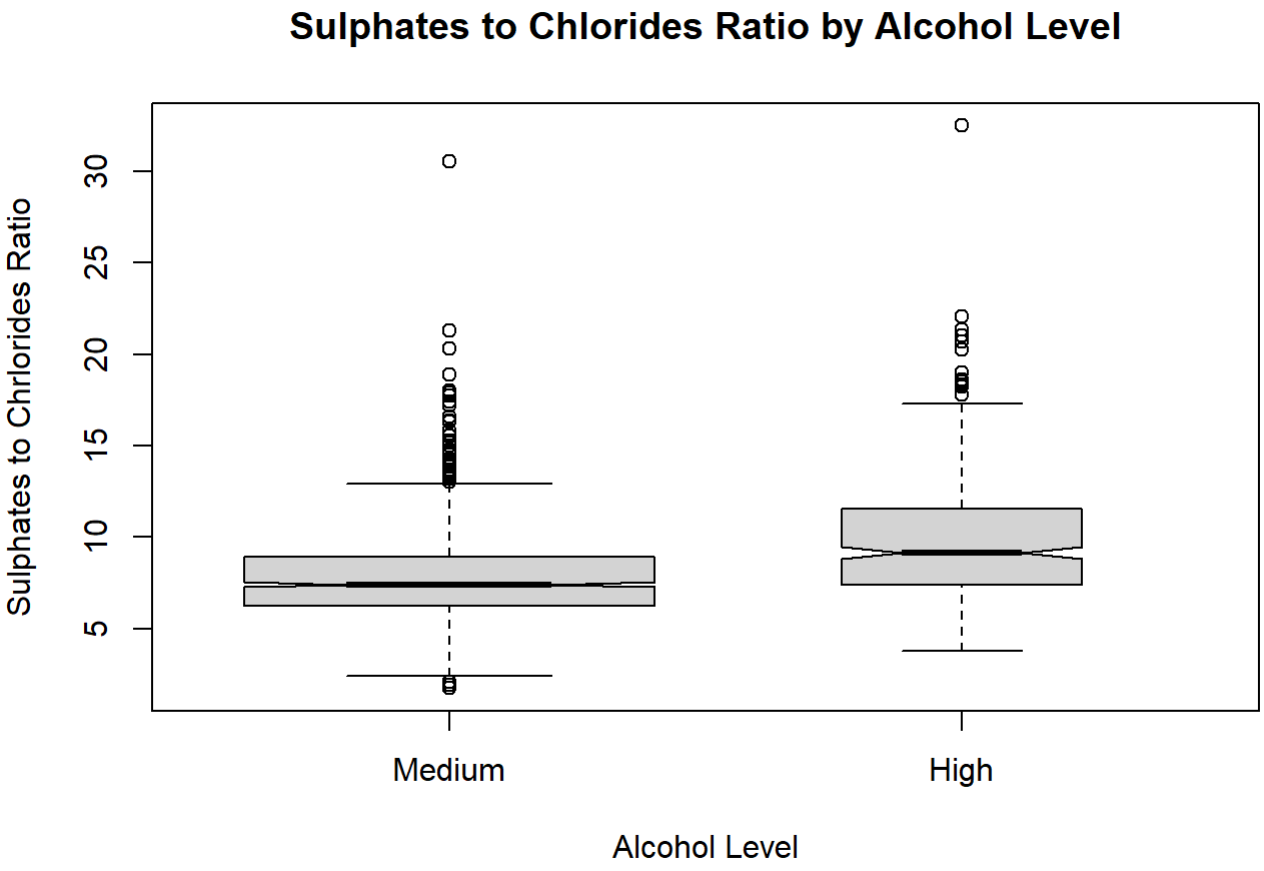
```
ALevel <- cut(redwine$alcohol, labels = c("Medium", "High"), breaks = c(0, 11, 100), include.low
est = TRUE)
redwine = cbind(redwine, ALevel)
```

Calculating the sulphates to chlorides ratio and adding the column to the dataframe.

```
sulphates_chlorides_ratio = redwine$sulphates/redwine$chlorides
redwine = cbind(redwine, sulphates_chlorides_ratio)
```

Plotting the Sulphates to Chlorides Ratio by Alcohol Level in a side-by-side boxplot.

```
boxplot(redwine$sulphates_chlorides_ratio ~ redwine$ALevel, xlab = "Alcohol Level", ylab = "Sulp
hates to Chrlorides Ratio", main = "Sulphates to Chlorides Ratio by Alcohol Level", varwidth = T
RUE, notch = TRUE)
```




```
table(redwine$ALevel)
```

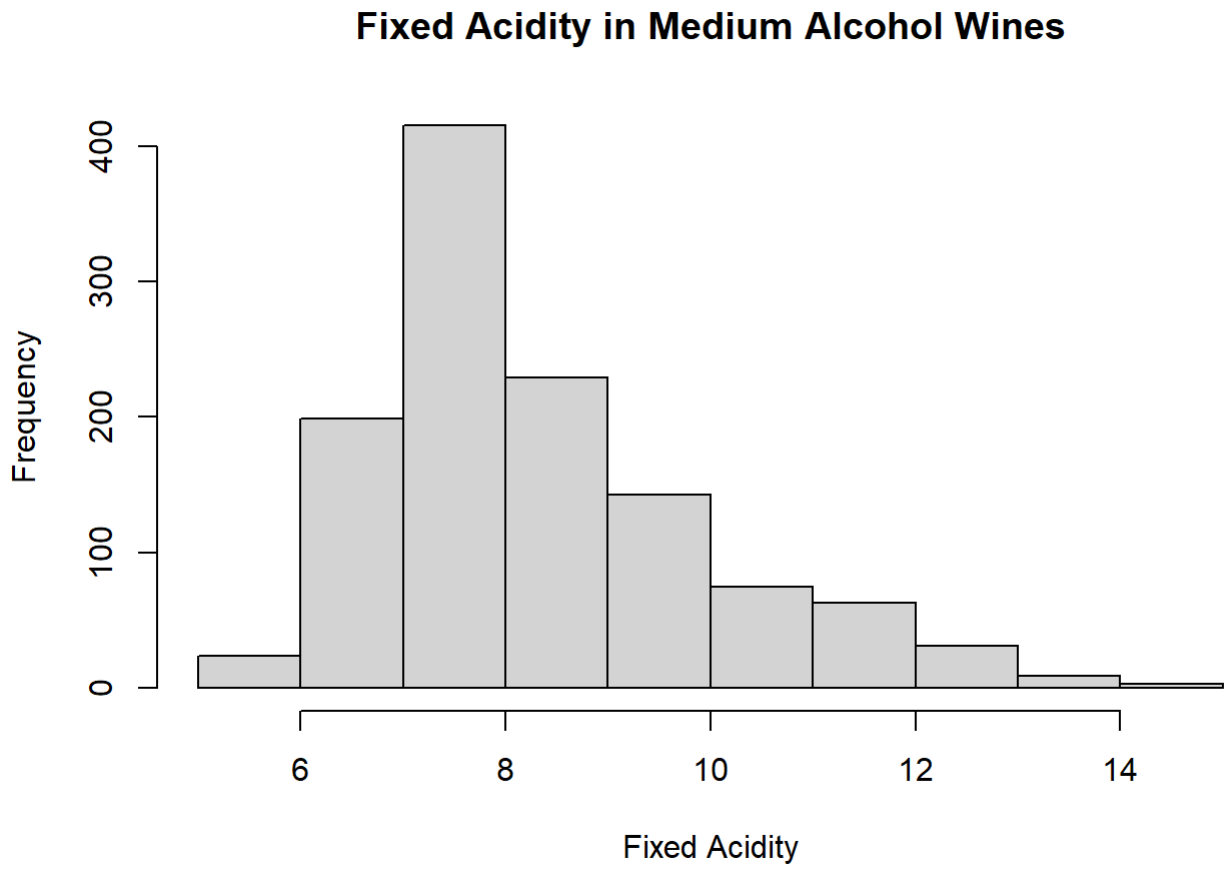
```
##
## Medium   High
##   1191    408
```

The table shows that there are 408 samples in the High Alcohol category.

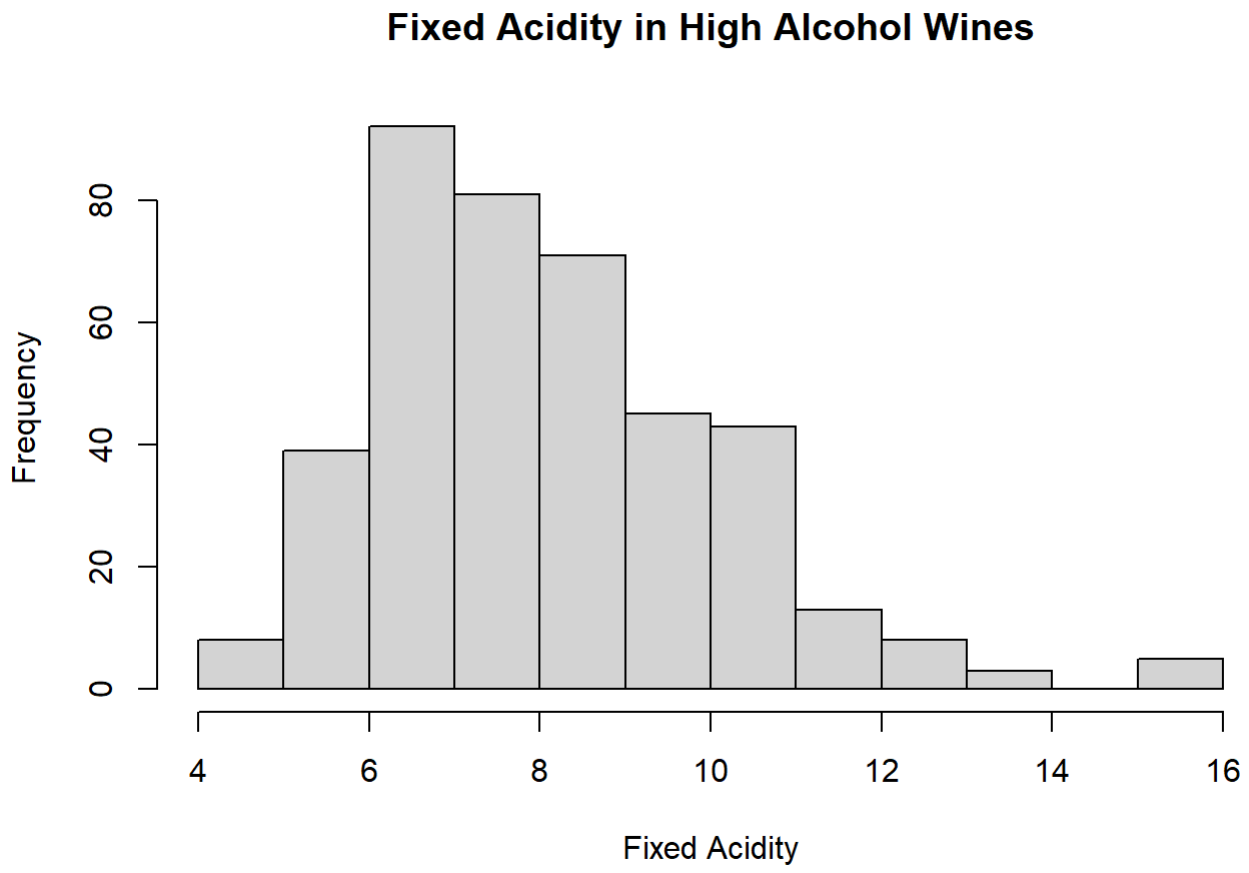
# 1.e. Produce a histogram showing the fixed acidity numbers for both High and Medium Alcohol Level

wine samples.

```
Medium_Wines <- subset(redwine, subset = ALevel == "Medium")
hist(Medium_Wines$fixed_acidity, xlab = "Fixed Acidity", main = "Fixed Acidity in Medium Alcohol
Wines")
```




```
High_Wines <- subset(redwine, subset = ALevel == "High")
hist(High_Wines$fixed_acidity, xlab = "Fixed Acidity", main = "Fixed Acidity in High Alcohol Win
es")
```



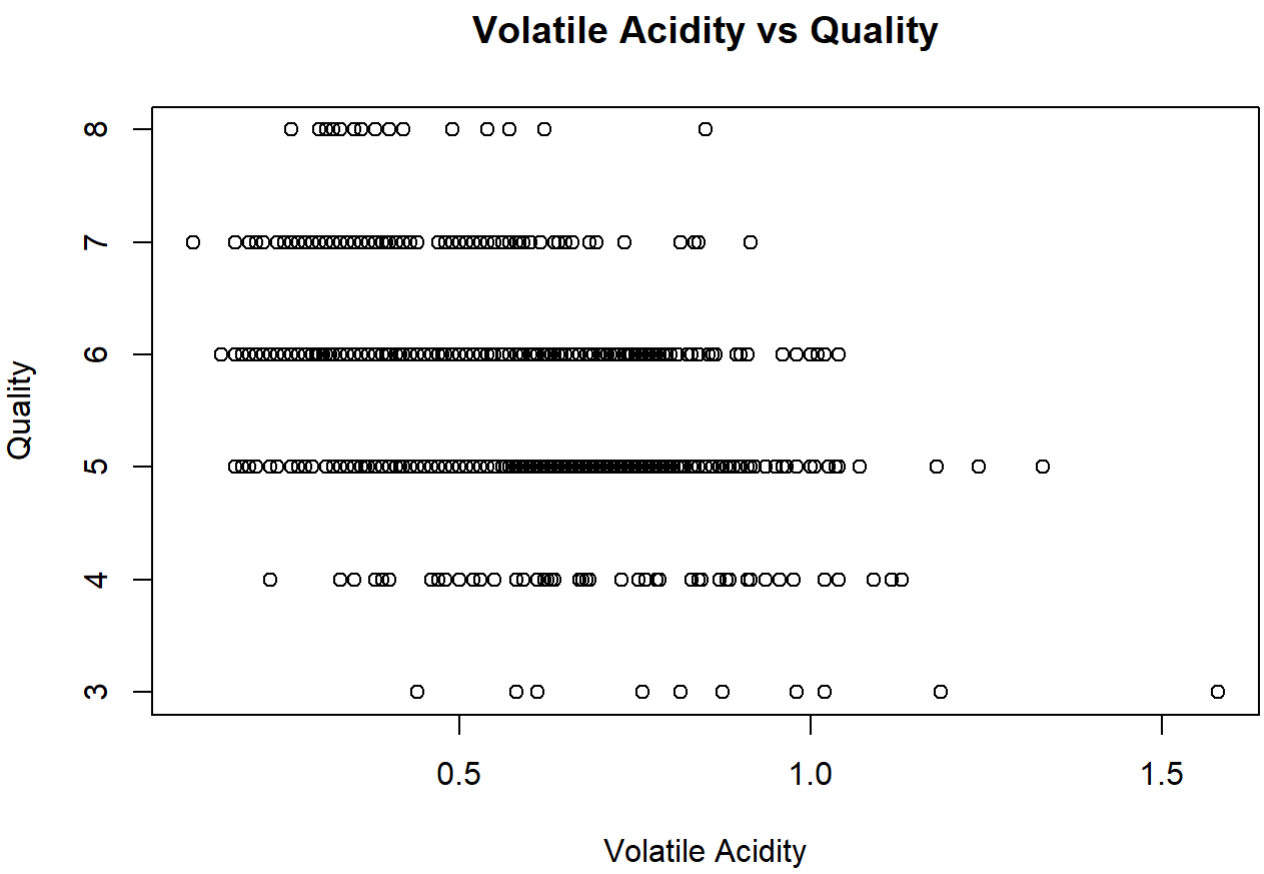

# 1.f. Produce two new plots of any type and provide a brief summary of your hypothesis and what you discover.

```
round(cor(subset(redwine, select = -c(ALevel))), 2)
```

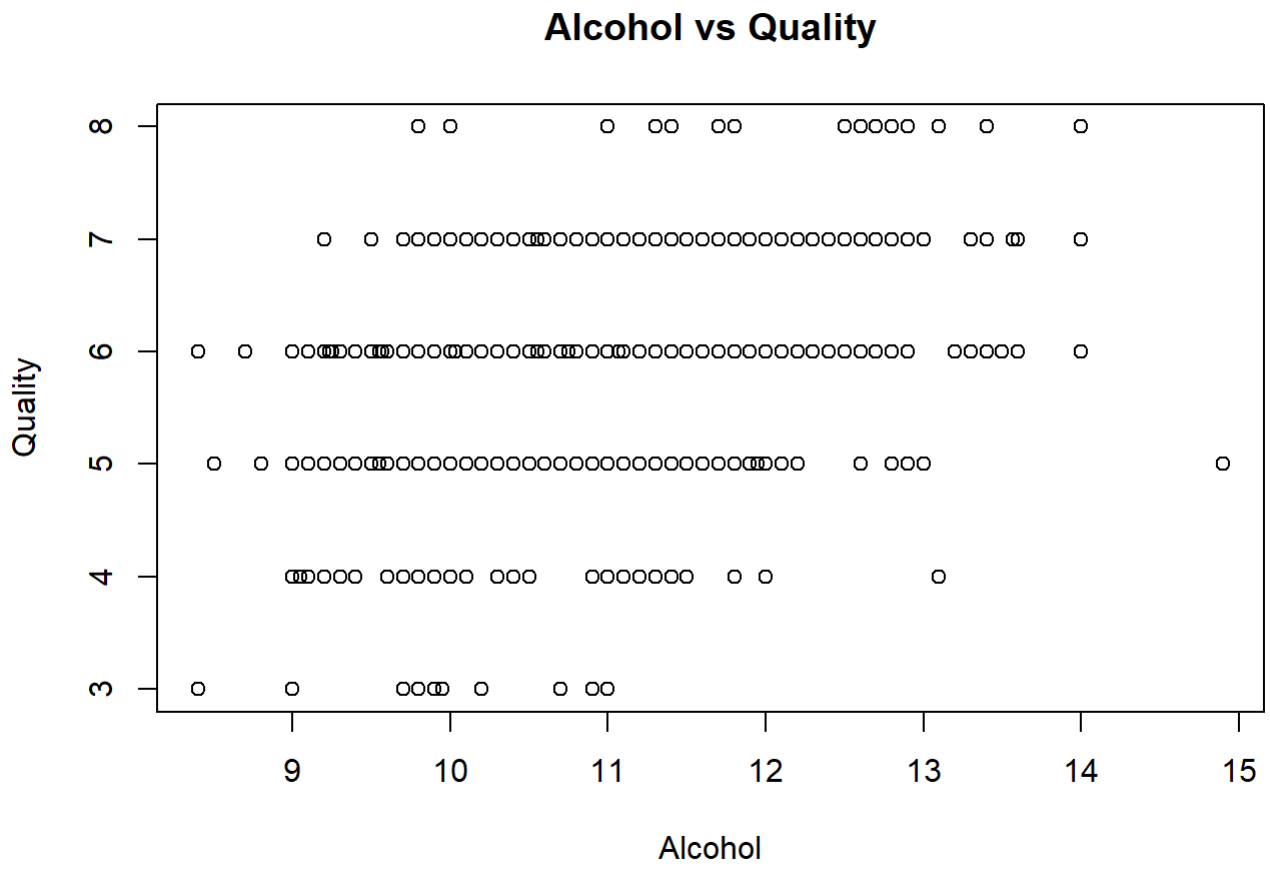
```
##                          fixed_acidity volatile_acidity citric_acid
## fixed_acidity                     1.00            -0.26        0.67
## volatile_acidity                 -0.26             1.00       -0.55
## citric_acid                       0.67            -0.55        1.00
## residual_sugar                    0.11             0.00        0.14
## chlorides                         0.09             0.06        0.20
## free_sulfur_dioxide              -0.15            -0.01       -0.06
## total_sulfur_dioxide             -0.11             0.08        0.04
## density                           0.67             0.02        0.36
## pH                               -0.68             0.23       -0.54
## sulphates                         0.18            -0.26        0.31
## alcohol                          -0.06            -0.20        0.11
## quality                           0.12            -0.39        0.23
## sulphates_chlorides_ratio        -0.06            -0.29        0.06
##                          residual_sugar chlorides free_sulfur_dioxide
## fixed_acidity                      0.11      0.09               -0.15
## volatile_acidity                   0.00      0.06               -0.01
## citric_acid                        0.14      0.20               -0.06
## residual_sugar                     1.00      0.06                0.19
## chlorides                          0.06      1.00                0.01
## free_sulfur_dioxide                0.19      0.01                1.00
## total_sulfur_dioxide               0.20      0.05                0.67
## density                            0.36      0.20               -0.02
## pH                                -0.09     -0.27                0.07
## sulphates                          0.01      0.37                0.05
## alcohol                            0.04     -0.22               -0.07
## quality                            0.01     -0.13               -0.05
## sulphates_chlorides_ratio         -0.08     -0.49                0.05
##                          total_sulfur_dioxide density    pH sulphates alcohol
## fixed_acidity                           -0.11    0.67 -0.68      0.18   -0.06
## volatile_acidity                         0.08    0.02  0.23     -0.26   -0.20
## citric_acid                              0.04    0.36 -0.54      0.31    0.11
## residual_sugar                           0.20    0.36 -0.09      0.01    0.04
## chlorides                                0.05    0.20 -0.27      0.37   -0.22
## free_sulfur_dioxide                      0.67   -0.02  0.07      0.05   -0.07
## total_sulfur_dioxide                     1.00    0.07 -0.07      0.04   -0.21
## density                                  0.07    1.00 -0.34      0.15   -0.50
## pH                                      -0.07   -0.34  1.00     -0.20    0.21
## sulphates                                0.04    0.15 -0.20      1.00    0.09
## alcohol                                 -0.21   -0.50  0.21      0.09    1.00
## quality                                 -0.19   -0.17 -0.06      0.25    0.48
## sulphates_chlorides_ratio               -0.03   -0.26  0.13      0.45    0.37
##                          quality sulphates_chlorides_ratio
## fixed_acidity               0.12                    -0.06
## volatile_acidity           -0.39                    -0.29
## citric_acid                 0.23                     0.06
## residual_sugar              0.01                    -0.08
## chlorides                  -0.13                    -0.49
## free_sulfur_dioxide        -0.05                     0.05
## total_sulfur_dioxide       -0.19                    -0.03
## density                    -0.17                    -0.26
## pH                         -0.06                     0.13
```

```
## sulphates                    0.25                    0.45
## alcohol                      0.48                    0.37
## quality                      1.00                    0.36
## sulphates_chlorides_ratio    0.36                    1.00
```

```
plot(redwine$volatile_acidity, redwine$quality, xlab = "Volatile Acidity", ylab = "Quality", mai
n = "Volatile Acidity vs Quality")
```




```
plot(redwine$alcohol, redwine$quality, xlab = "Alcohol", ylab = "Quality", main = "Alcohol vs Qu
ality")
```




Using the correlation matrix as a jumping-off point, we can see from the plots that as Volatile Acidity increases, Quality decreases. Additionally, as the Alcohol increases, so does Quality.

# 2.

Here I read in the forestfires.csv file, converting the DC column to numeric since it was not for some reason.

```
forestfires <- read.csv("forestfires.csv")
suppressWarnings(forestfires$DC <- as.numeric(forestfires$DC))
```

# 2.a. Specify which of the predictors are quantitative and which are qualitative

Month and Day are discrete, categorical variables and this qualitative. X, Y, FFMC, DMC, DC, ISI, Temp, RH, Wind, Rain, and Area are all continuous and quantitative.

# 2.b. What is the range, mean and standard deviation of each quantitative predictor? Which month has the

# highest number of fires?

```
quantitative_subset = subset(forestfires, select = -c(X.1, X.2, month, day))

suppressWarnings(round(sapply(quantitative_subset, function(x) c("Range" = (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)), "Mean" = mean(x, na.rm = TRUE), "Standard Deviation" = sd(x, na.rm = TRUE)))), 2)
```

```
##                    X Y FFMC DMC  DC ISI temp RH wind rain area
## Range              8 7   78 290 853  56   31 85    9    6 1091
## Mean               5 4   91 111 548   9   19 44    4    0   13
## Standard Deviation 2 1    6  64 248   5    6 16    2    0   64
```

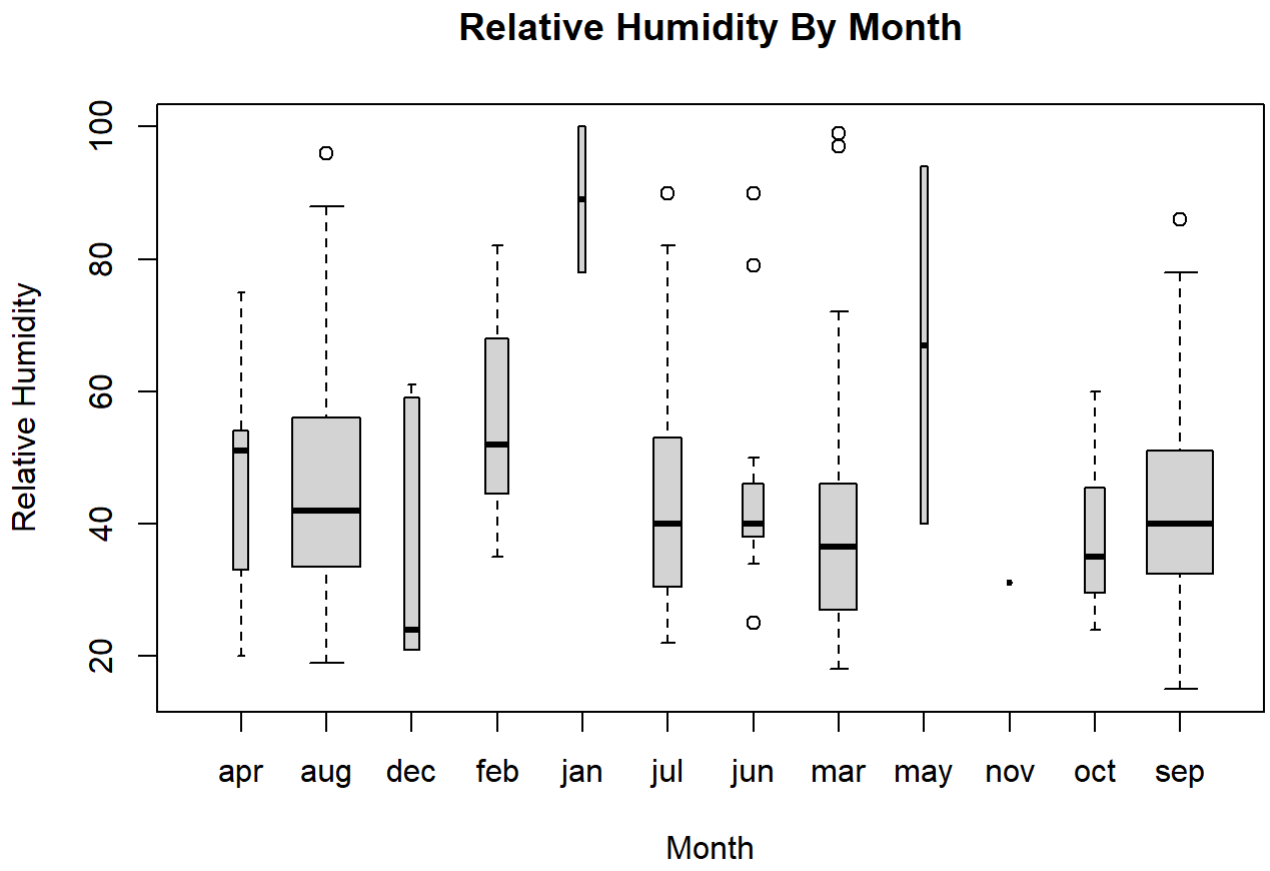
```
table(forestfires$month)
```

```
##
## apr aug dec feb jan jul jun mar may nov oct sep
##   9 184   9  20   2  32  17  54   2   1  15 172
```

From our table, August has the highest number of fires, with September as a close second.

# 2.c Produce boxplots of relative humidity (RH) by month. Your figure will have a boxplot for every month. Which month has the highest median RH value?

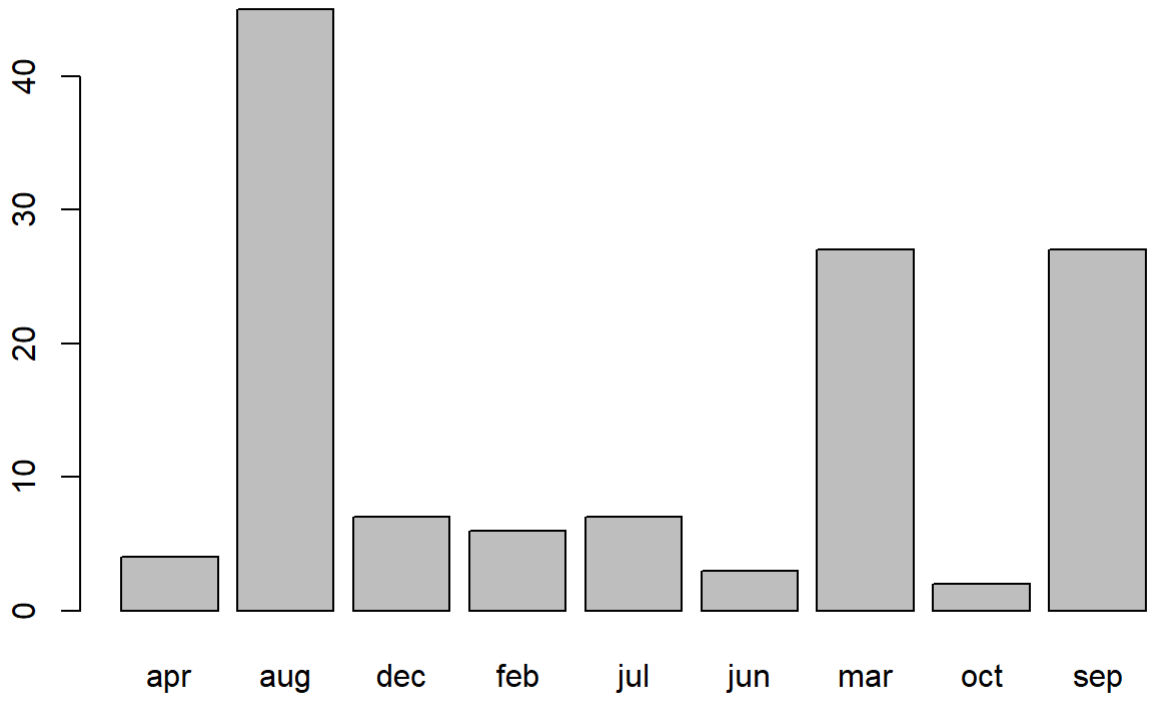
```
boxplot(forestfires$RH ~ forestfires$month, xlab = "Month", ylab = "Relative Humidity", main = "Relative Humidity By Month", varwidth = TRUE)
```




The month with the highest median Relative Humidity during a forest fire is January.

# 2.d. Produce a bar plot to show the count of forest fires in each month for which wind is greater than 4.9. During which months are high wind forest fires most common? (Hint: filter data by wind, group data by month and calculate count.)

```
barplot(table(forestfires[forestfires$wind > 4.9, ]$month))
```



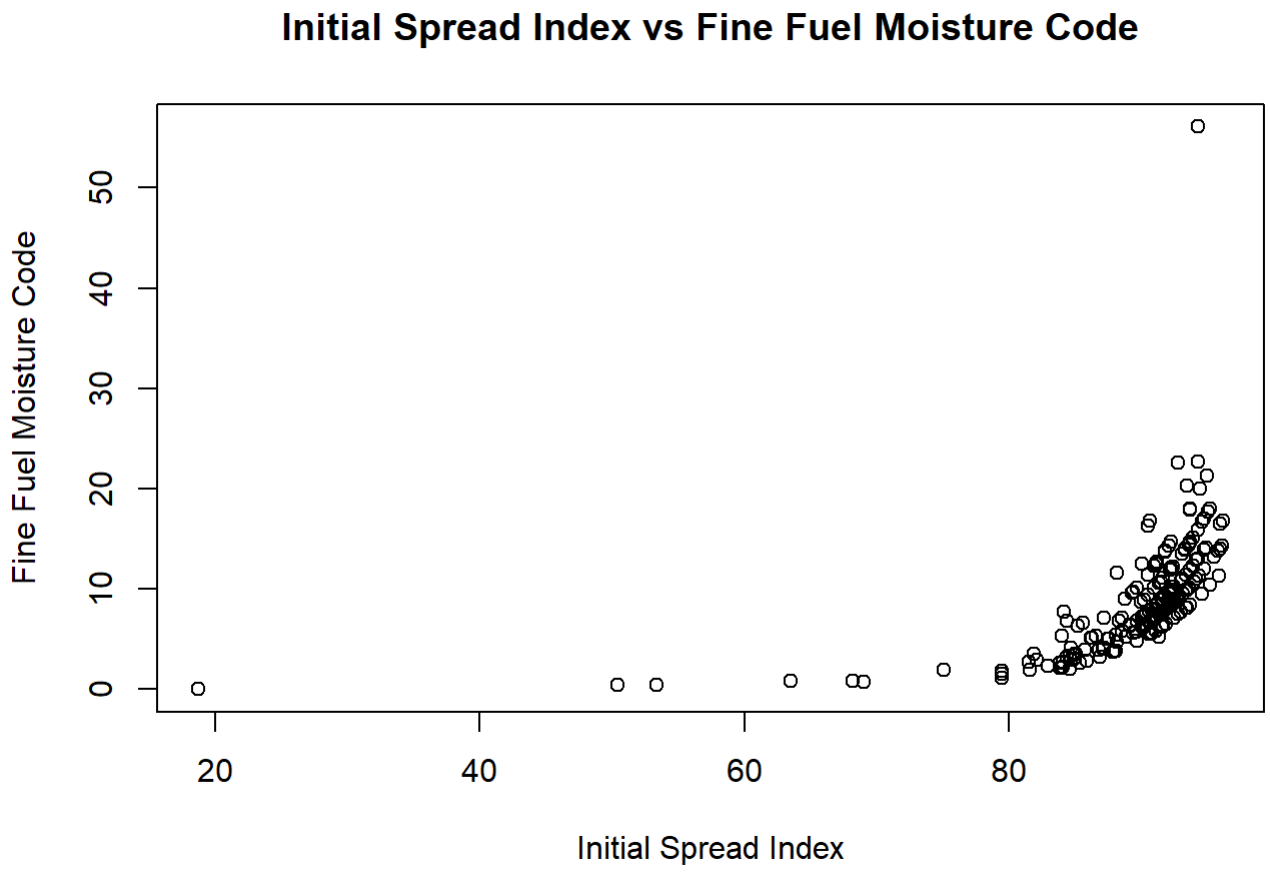

High wind forest fires are most common in August.

# 2.e. Using the full data set, investigate the predictors graphically, using scatterplots, correlation scores or other tools of your choice. Create a correlation matrix for the relevant variables.

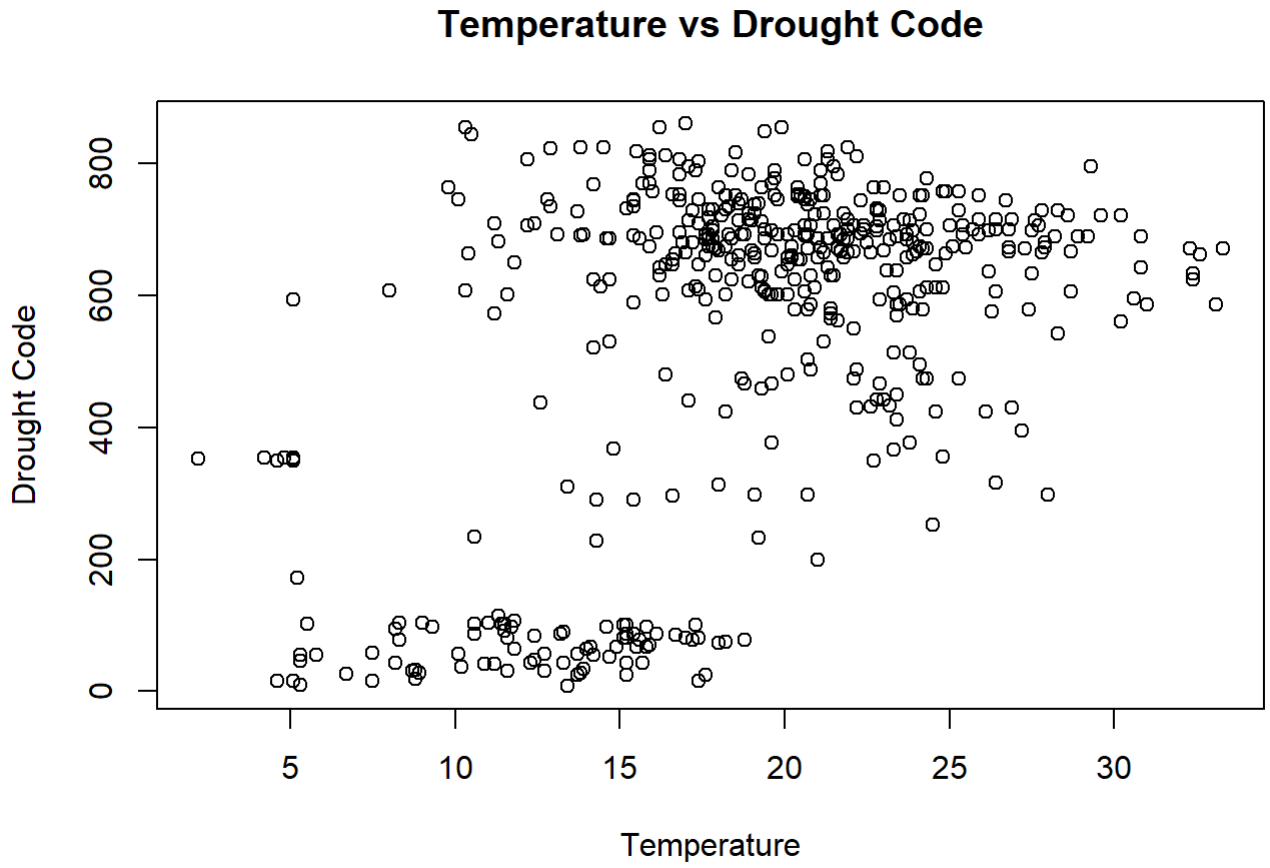
```
correlation_subset = subset(forestfires, select = c(FFMC, DMC, DC, ISI, temp, RH, wind, rain, ar
ea))
round(cor(correlation_subset, method = "pearson", use="complete.obs"), 2)
```

```
##       FFMC   DMC    DC   ISI  temp    RH  wind  rain  area
## FFMC  1.00  0.38  0.33  0.53  0.43 -0.30 -0.03  0.06  0.04
## DMC   0.38  1.00  0.68  0.30  0.47  0.07 -0.11  0.07  0.07
## DC    0.33  0.68  1.00  0.23  0.50 -0.04 -0.20  0.04  0.05
## ISI   0.53  0.30  0.23  1.00  0.39 -0.13  0.11  0.07  0.01
## temp  0.43  0.47  0.50  0.39  1.00 -0.53 -0.23  0.07  0.10
## RH   -0.30  0.07 -0.04 -0.13 -0.53  1.00  0.07  0.10 -0.08
## wind -0.03 -0.11 -0.20  0.11 -0.23  0.07  1.00  0.06  0.01
## rain  0.06  0.07  0.04  0.07  0.07  0.10  0.06  1.00 -0.01
## area  0.04  0.07  0.05  0.01  0.10 -0.08  0.01 -0.01  1.00
```

```
plot(forestfires$FFMC, forestfires$ISI, xlab = "Initial Spread Index", ylab = "Fine Fuel Moistur
e Code", main = "Initial Spread Index vs Fine Fuel Moisture Code")
```




```
plot(forestfires$temp, forestfires$DC, xlab = "Temperature", ylab = "Drought Code", main = "Temp
erature vs Drought Code")
```




From the correlation matrix, we can see that the Fine Fuel Moisture Code is correlated with the Initial Spread Index, Duff Moisture Code is correlated with Drought Code, and Temperature is negatively correlated with relative humidity.

# 2.f. Suppose that we wish to predict the Initial spread index (ISI) based on the other variables. Which, if any, of the other variables might be useful in predicting ISI? Justify your answer based on the prior correlations.

The variable most directly correlated with the Initial Spread Index is the Fine Fuel Moisture Code, followed by the Temperature as shown by the correlation matrix above. Additionally, this resource from the Canadian Government (https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system) shows that the Initial Spread Index is calculated based on Temperature, Relative Humidity, Wind, Rain, the Fine Fuel Moisture Code, as well as the previous day's Fine Fuel Moisture Code.