Final Project

Wine Quality Analysis

Nicole Ferreira

June 18, 2021

1) Loading Packages

```
library(dplyr)

##
## Attaching package: 'dplyr'

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

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

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   1.4.0

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ggplot2)
```

2.1) White Wine Data Wrangling

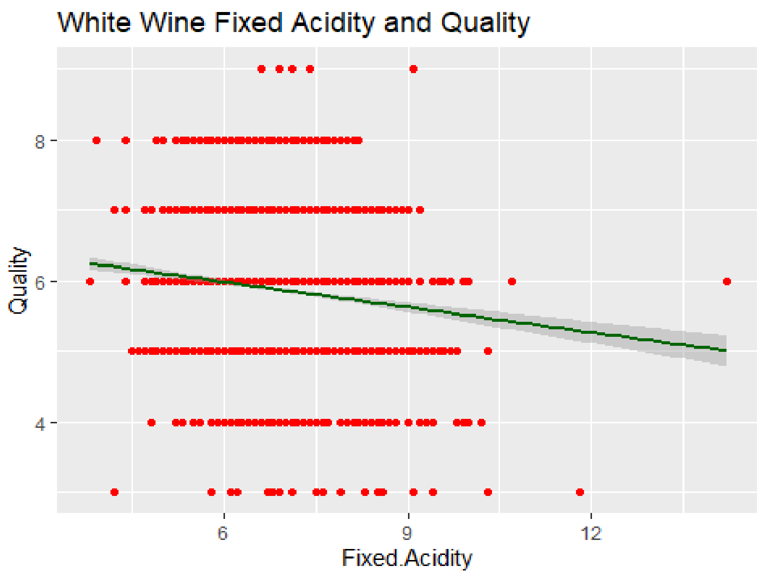
```
white.wine <- read.csv("winequality-white.csv", sep=";", header=FALSE)

names(white.wine)[names(white.wine) == "V1"] <- "Fixed.Acidity"
names(white.wine)[names(white.wine) == "V2"] <- "Volatile.Acidity"
names(white.wine)[names(white.wine) == "V3"] <- "Citric.Acid"
names(white.wine)[names(white.wine) == "V4"] <- "Residual.Sugar"
names(white.wine)[names(white.wine) == "V5"] <- "Chlorides"
names(white.wine)[names(white.wine) == "V6"] <- "Free.Sulfur.Dioxide"
names(white.wine)[names(white.wine) == "V7"] <- "Total.Sulfur.Dioxide"
names(white.wine)[names(white.wine) == "V8"] <- "Density"
names(white.wine)[names(white.wine) == "V9"] <- "pH"
names(white.wine)[names(white.wine) == "V10"] <- "Sulphates"
names(white.wine)[names(white.wine) == "V11"] <- "Alcohol"
names(white.wine)[names(white.wine) == "V12"] <- "Quality"

white.wine <- white.wine[-c(1), ]

white.wine$`Fixed.Acidity` <- as.numeric(white.wine$`Fixed.Acidity`)

#sum(is.na(white.wine))

#summary(white.wine)
str(white.wine)

## 'data.frame':    4898 obs. of  12 variables:
##  $ Fixed.Acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ Volatile.Acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3
0.22 ...
##  $ Citric.Acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34
0.43 ...
##  $ Residual.Sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ Chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045
0.045 0.049 0.044 ...
##  $ Free.Sulfur.Dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ Total.Sulfur.Dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ Density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22
...
##  $ Sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49
0.45 ...
##  $ Alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ Quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

2.2) Red Wine Data Wrangling

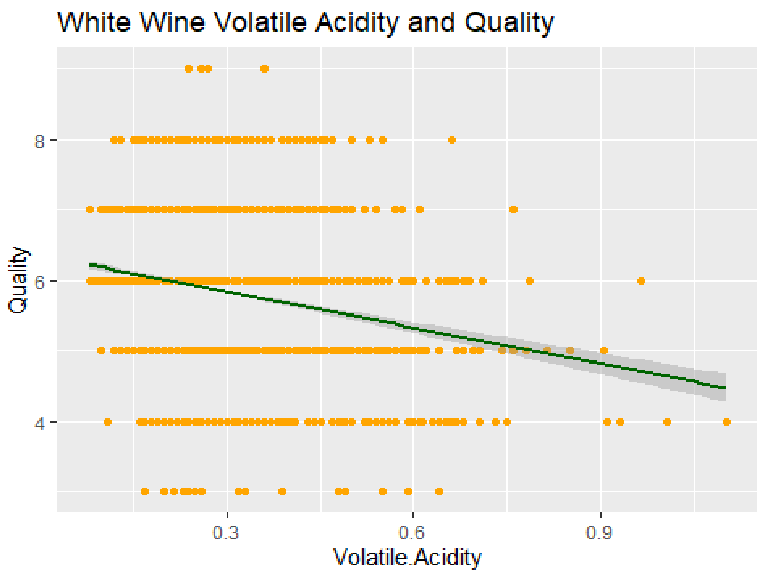
```
red.wine <- read.csv("winequality-red.csv", sep=";", header=FALSE)

names(red.wine)[names(red.wine) == "V1"] <- "Fixed.Acidity"
names(red.wine)[names(red.wine) == "V2"] <- "Volatile.Acidity"
names(red.wine)[names(red.wine) == "V3"] <- "Citric.Acid"
names(red.wine)[names(red.wine) == "V4"] <- "Residual.Sugar"
names(red.wine)[names(red.wine) == "V5"] <- "Chlorides"
names(red.wine)[names(red.wine) == "V6"] <- "Free.Sulfur.Dioxide"
names(red.wine)[names(red.wine) == "V7"] <- "Total.Sulfur.Dioxide"
names(red.wine)[names(red.wine) == "V8"] <- "Density"
names(red.wine)[names(red.wine) == "V9"] <- "pH"
names(red.wine)[names(red.wine) == "V10"] <- "Sulphates"
names(red.wine)[names(red.wine) == "V11"] <- "Alcohol"
names(red.wine)[names(red.wine) == "V12"] <- "Quality"

red.wine <- red.wine[-c(1), ]

red.wine$`Fixed.Acidity` <- as.numeric(red.wine$`Fixed.Acidity`)
red.wine$`Volatile.Acidity` <- as.numeric(red.wine$`Volatile.Acidity`)
red.wine$`Citric.Acid` <- as.numeric(red.wine$`Citric.Acid`)
red.wine$`Residual.Sugar` <- as.numeric(red.wine$`Residual.Sugar`)
red.wine$Chlorides <- as.numeric(red.wine$Chlorides)
red.wine$`Free.Sulfur.Dioxide` <- as.numeric(red.wine$`Free.Sulfur.Dioxide`)
red.wine$`Total.Sulfur.Dioxide` <-
as.numeric(red.wine$`Total.Sulfur.Dioxide`)
red.wine$Density <- as.numeric(red.wine$Density)
red.wine$pH <- as.numeric(red.wine$pH)
red.wine$Sulphates <- as.numeric(red.wine$Sulphates)
red.wine$Alcohol <- as.numeric(red.wine$Alcohol)
red.wine$Quality <- as.integer(red.wine$Quality)

#sum(is.na(red.wine))

#summary(red.wine)
str(red.wine)

## 'data.frame':    1599 obs. of  12 variables:
##  $ Fixed.Acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ Volatile.Acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58
0.5 ...
##  $ Citric.Acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ Residual.Sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ Chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069
0.065 0.073 0.071 ...
##  $ Free.Sulfur.Dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ Total.Sulfur.Dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ Density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36
3.35 ...
##  $ Sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ Alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ Quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

3.1) White Wine Exploratory Data Analysis/Scatterplots

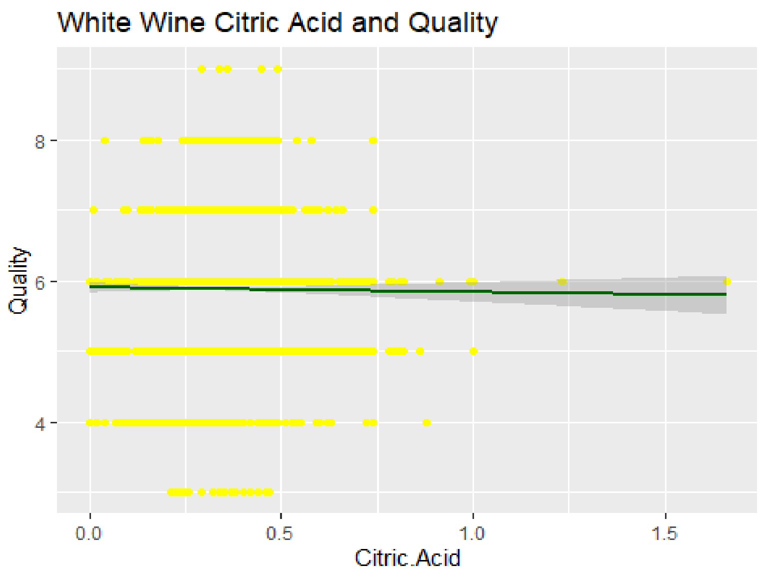
```
ggplot(white.wine, aes(x=Fixed.Acidity, y=Quality))+
  geom_point(color="red")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Fixed Acidity and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



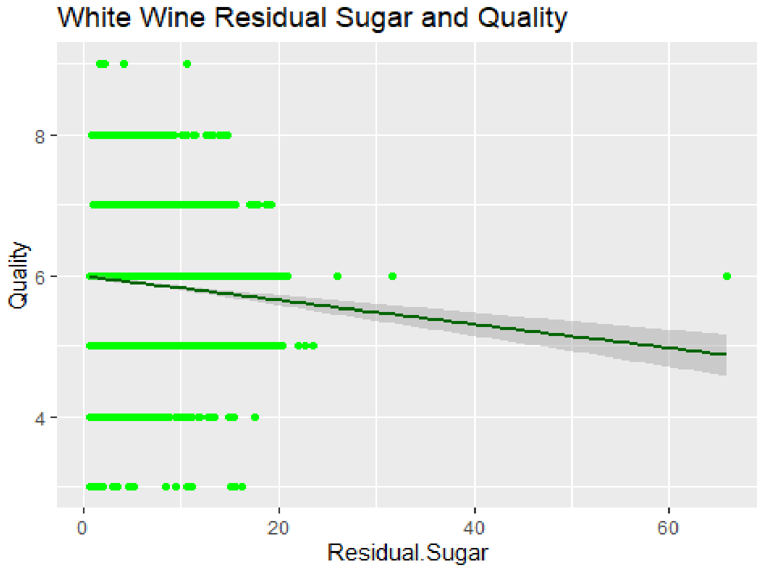
```
ggplot(white.wine, aes(x=Volatile.Acidity, y=Quality))+
  geom_point(color="orange") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Volatile Acidity and Quality") #Revealed as most
important

## `geom_smooth()` using formula 'y ~ x'
```



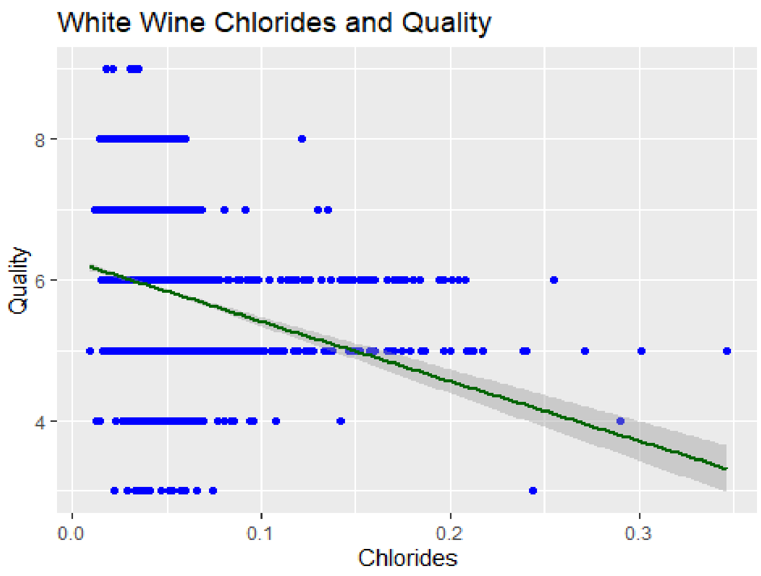
```
ggplot(white.wine, aes(x=Citric.Acid, y=Quality))+
  geom_point(color="yellow")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Citric Acid and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



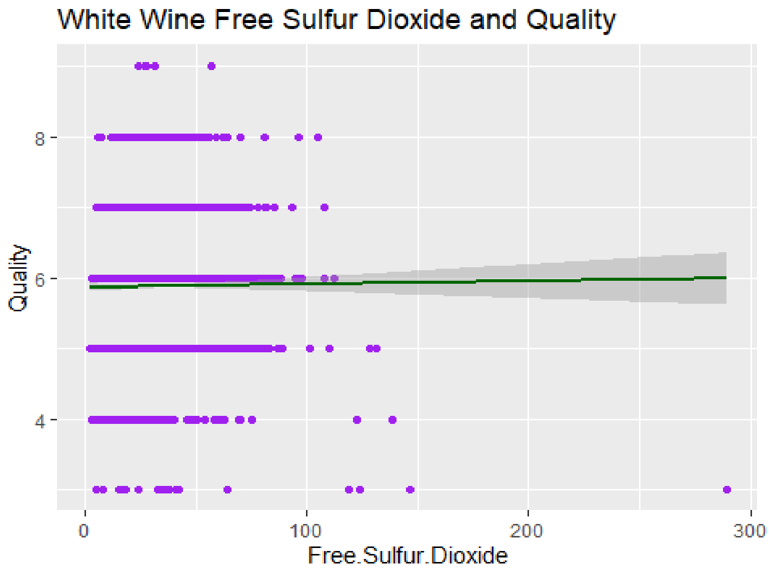
```
ggplot(white.wine, aes(x=Residual.Sugar, y=Quality))+
  geom_point(color="green")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Residual Sugar and Quality") #Revealed as most
important

## `geom_smooth()` using formula 'y ~ x'
```



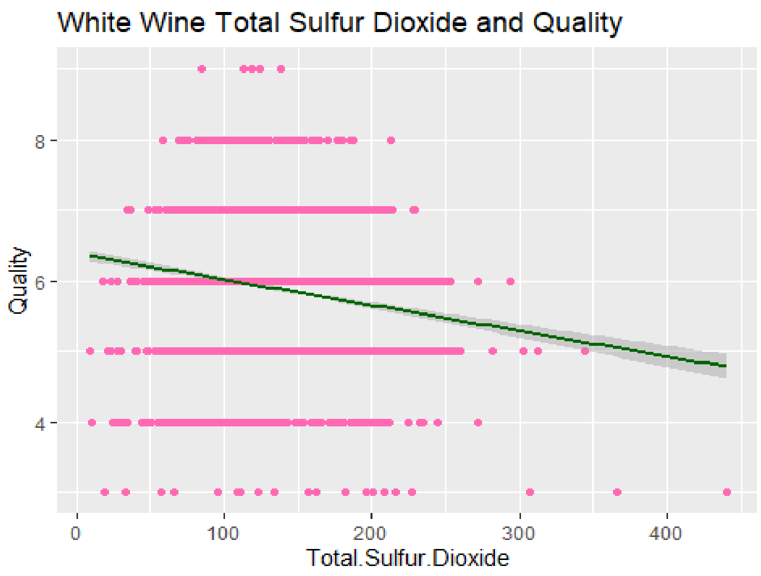
```
ggplot(white.wine, aes(x=Chlorides, y=Quality))+
  geom_point(color="blue")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Chlorides and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



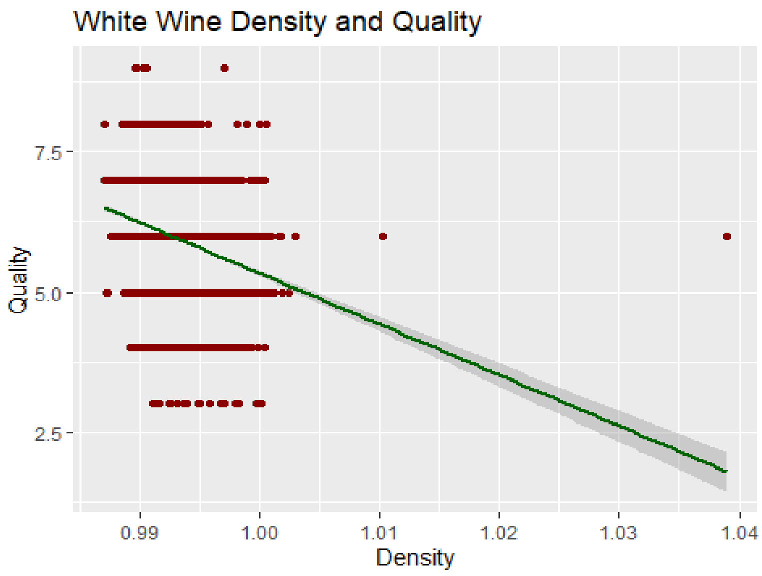
```
ggplot(white.wine, aes(x=Free.Sulfur.Dioxide, y=Quality))+
  geom_point(color="purple")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Free Sulfur Dioxide and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



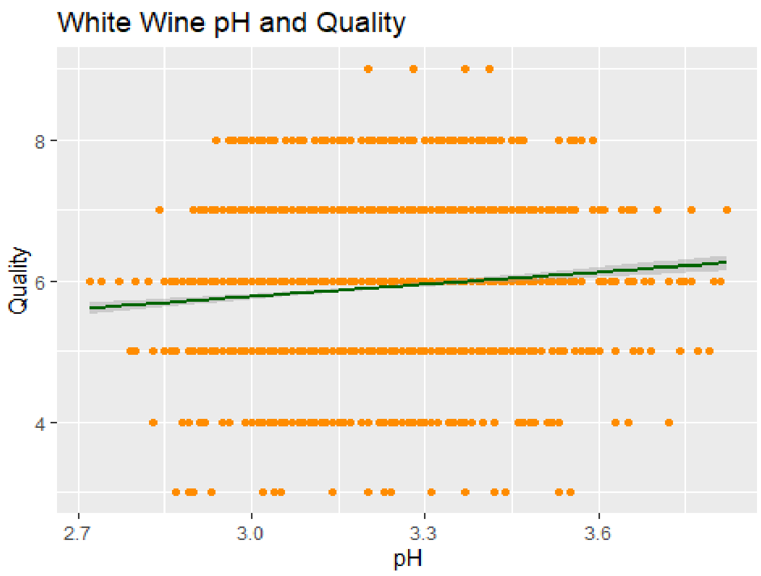
```
ggplot(white.wine, aes(x=Total.Sulfur.Dioxide, y=Quality))+
  geom_point(color="hot pink")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Total Sulfur Dioxide and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



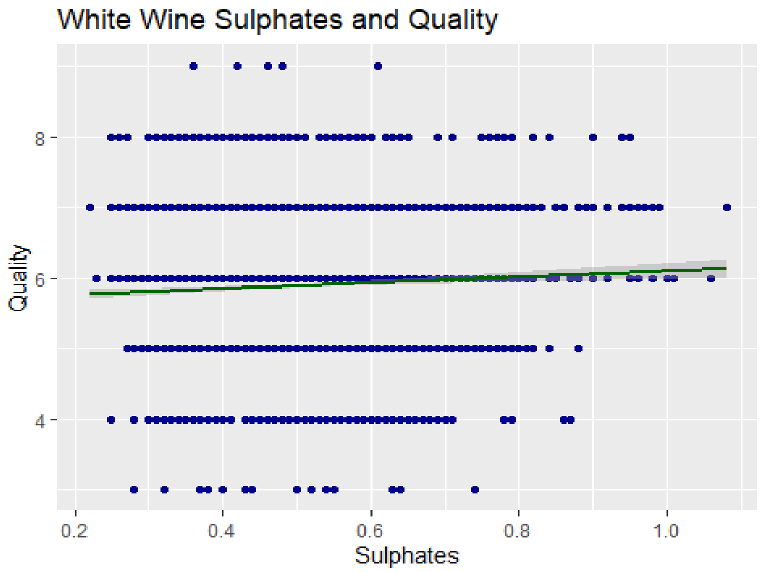
```
ggplot(white.wine, aes(x=Density, y=Quality))+
  geom_point(color="dark red")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Density and Quality") #Revealed as most important

## `geom_smooth()` using formula 'y ~ x'
```



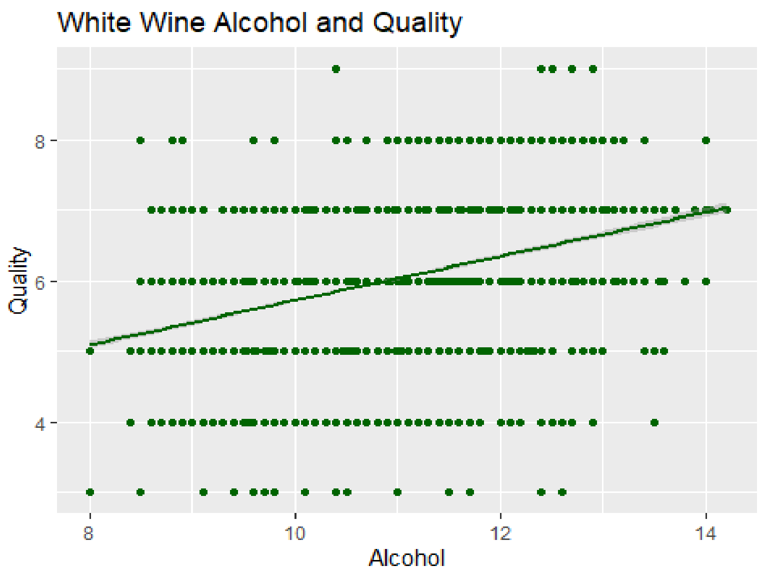
```
ggplot(white.wine, aes(x=pH, y=Quality))+
  geom_point(color="dark orange") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine pH and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(white.wine, aes(x=Sulphates, y=Quality))+
  geom_point(color="dark blue") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Sulphates and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



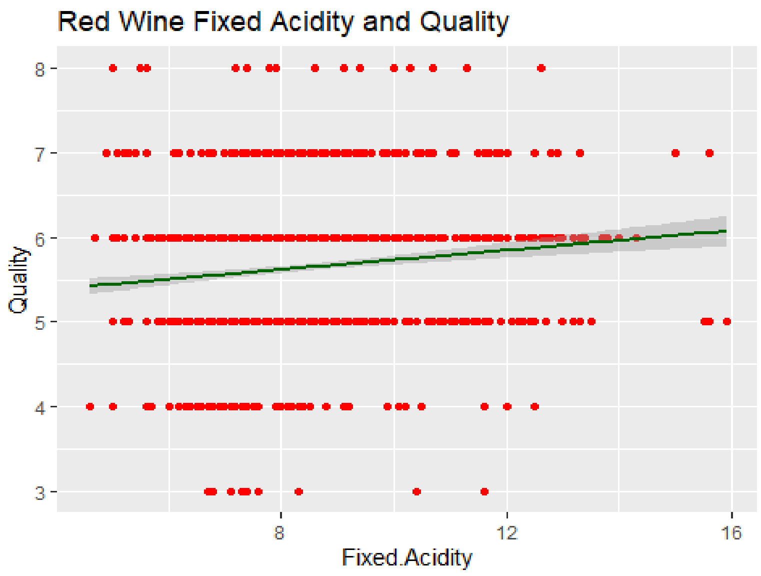
```
ggplot(white.wine, aes(x=Alcohol, y=Quality))+
  geom_point(color="dark green") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("White Wine Alcohol and Quality") #Revealed as most important

## `geom_smooth()` using formula 'y ~ x'
```



```
#pairs(white.wine,col=as.numeric(white.wine$cluster))
```

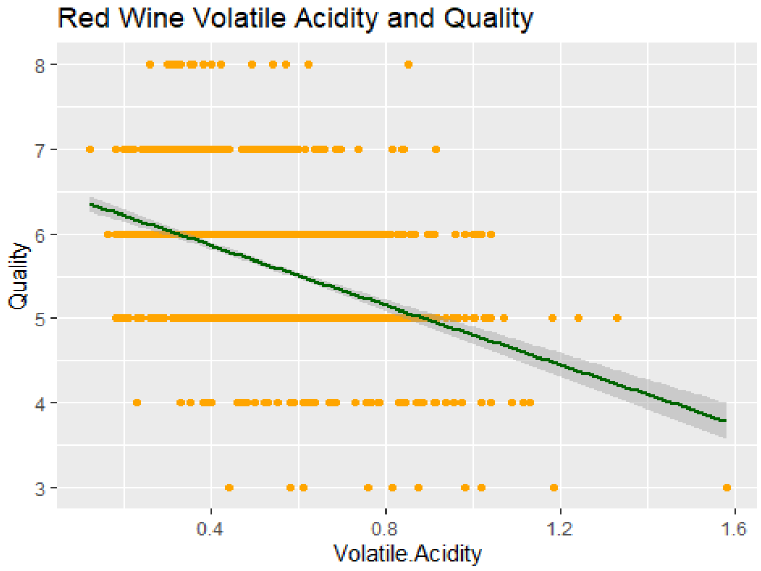
3.2) Red Wine Exploratory Data Analysis/Scatterplots

```
ggplot(red.wine, aes(x=Fixed.Acidity, y=Quality))+
  geom_point(color="red")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Fixed Acidity and Quality")

## `geom_smooth()` using formula 'y ~ x'
```

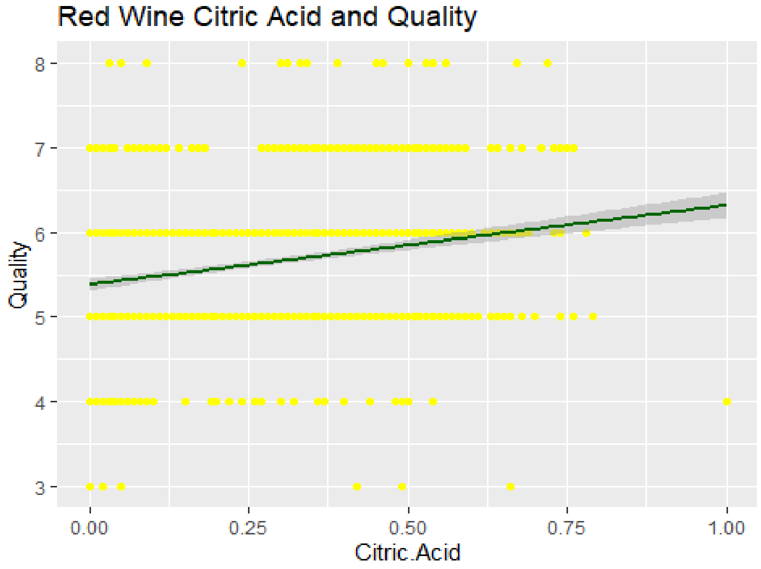


```
ggplot(red.wine, aes(x=Volatile.Acidity, y=Quality))+
  geom_point(color="orange") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Volatile Acidity and Quality") #Revealed as most
important
```

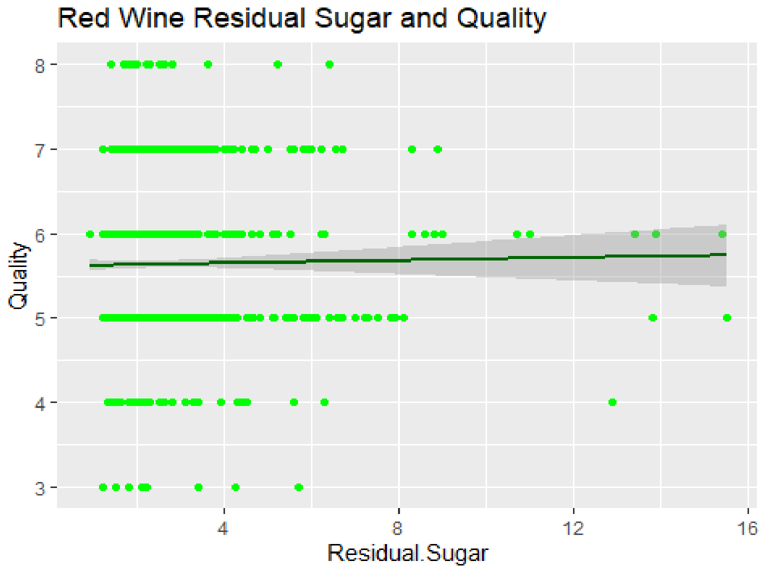
```
## `geom_smooth()` using formula 'y ~ x'
```



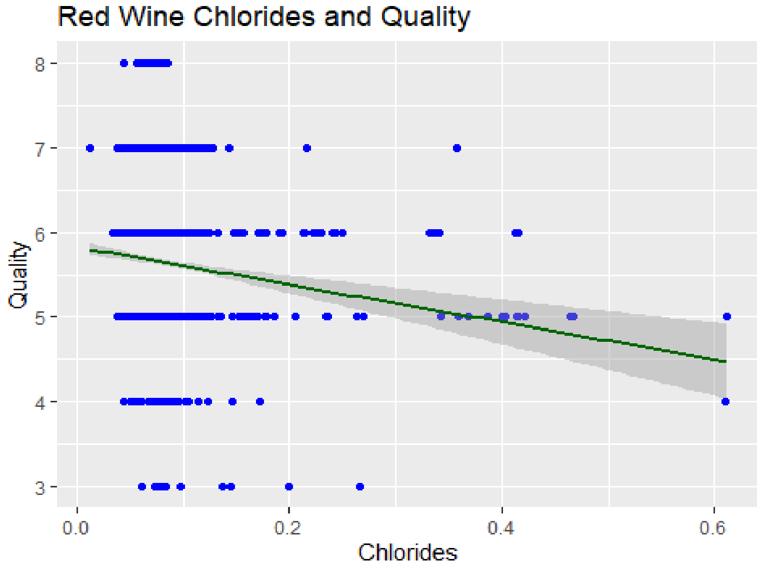
```
ggplot(red.wine, aes(x=Citric.Acid, y=Quality))+
  geom_point(color="yellow")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Citric Acid and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



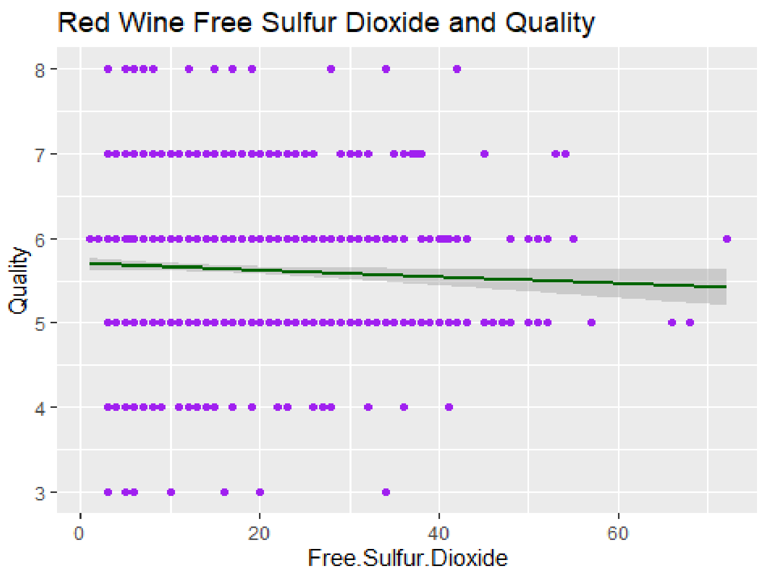
```
ggplot(red.wine, aes(x=Residual.Sugar, y=Quality))+
  geom_point(color="green")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Residual Sugar and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



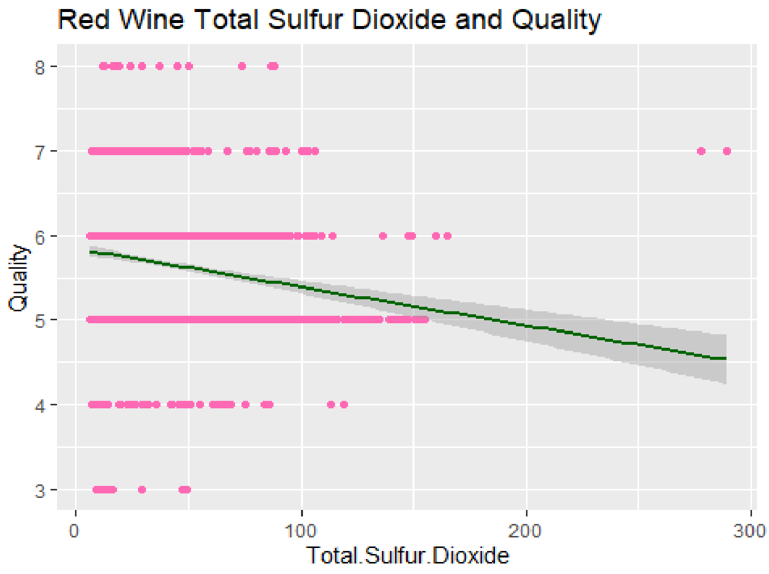
```
ggplot(red.wine, aes(x=Chlorides, y=Quality))+
  geom_point(color="blue")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Chlorides and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



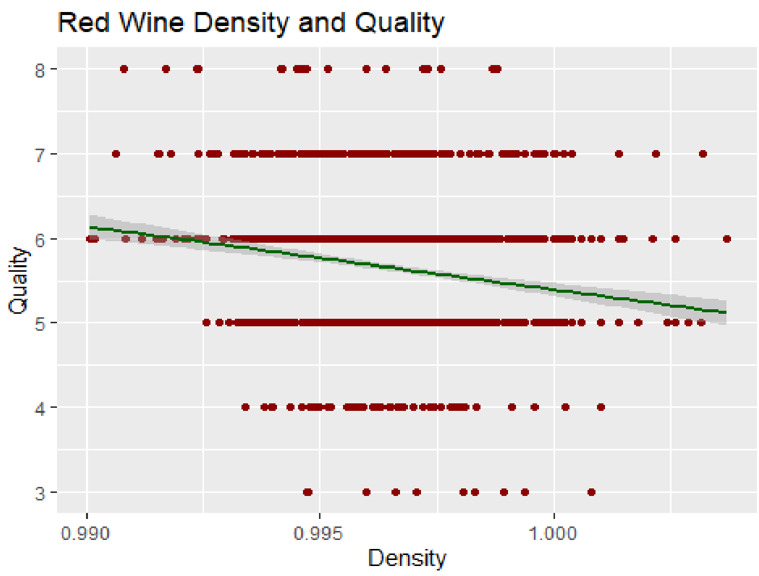
```
ggplot(red.wine, aes(x=Free.Sulfur.Dioxide, y=Quality))+
  geom_point(color="purple")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Free Sulfur Dioxide and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



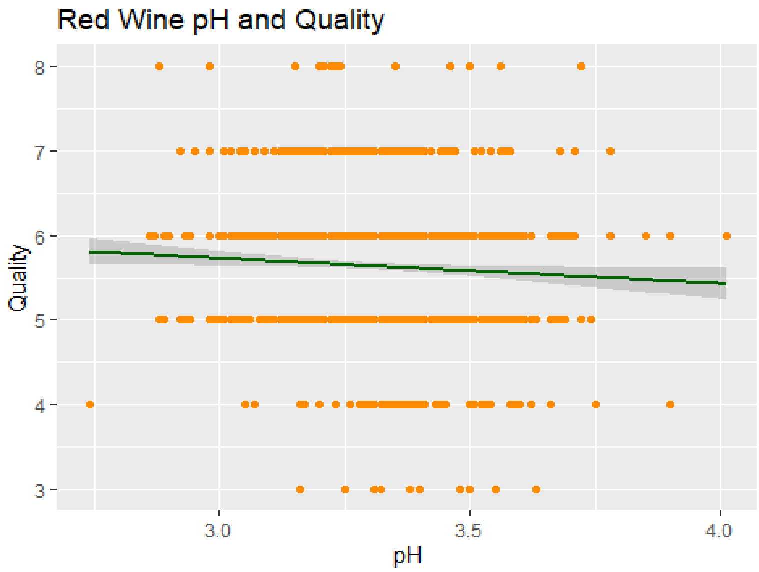
```
ggplot(red.wine, aes(x=Total.Sulfur.Dioxide, y=Quality))+
  geom_point(color="hot pink")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Total Sulfur Dioxide and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



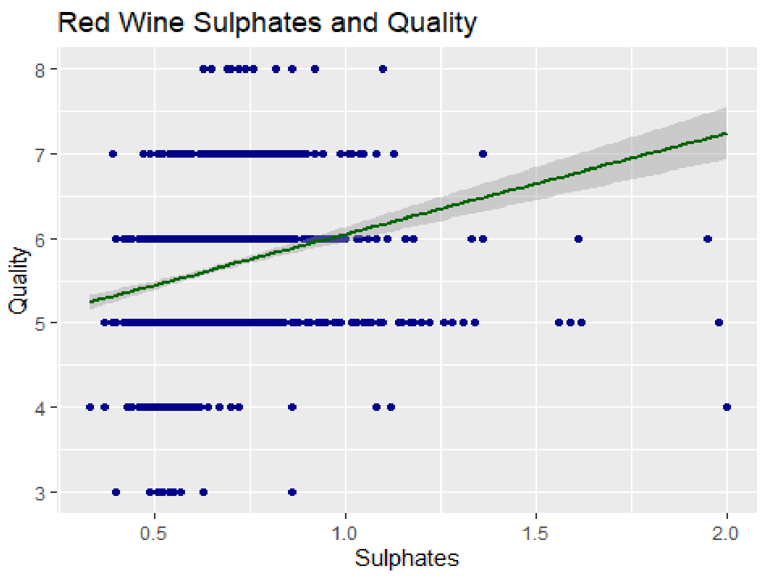
```
ggplot(red.wine, aes(x=Density, y=Quality))+
  geom_point(color="dark red")+
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Density and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



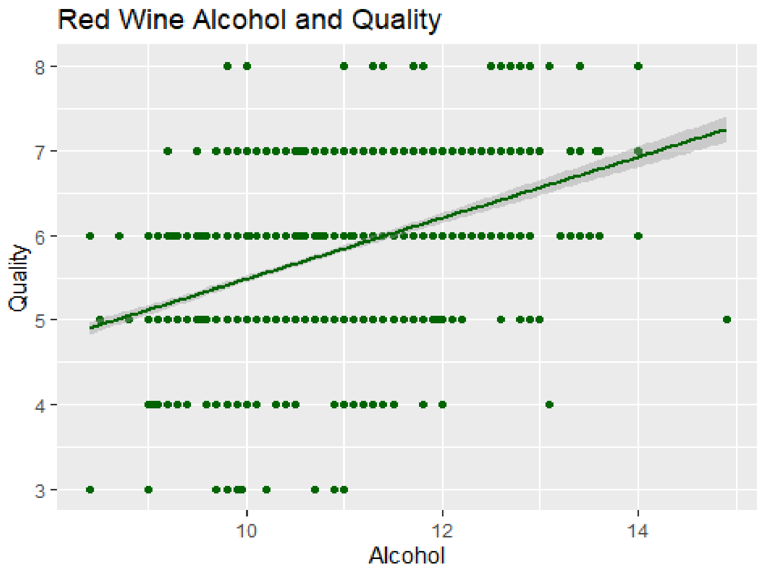
```
ggplot(red.wine, aes(x=pH, y=Quality))+
  geom_point(color="dark orange") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine pH and Quality")

## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(red.wine, aes(x=Sulphates, y=Quality))+
  geom_point(color="dark blue") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Sulphates and Quality") #Revealed as most important

## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(red.wine, aes(x=Alcohol, y=Quality))+
  geom_point(color="dark green") +
  geom_smooth(method="lm", color = 'dark green') +
  ggtitle("Red Wine Alcohol and Quality") #Revealed as most important

## `geom_smooth()` using formula 'y ~ x'
```



4.1) White Wine Multiple Linear Regression

```
white.wine.lm <- lm(Quality ~ Fixed.Acidity + Volatile.Acidity + Citric.Acid
+ Residual.Sugar + Chlorides + Free.Sulfur.Dioxide + Total.Sulfur.Dioxide +
Density + pH + Sulphates + Alcohol, data=white.wine)
summary(white.wine.lm)

##
## Call:
```

```
## lm(formula = Quality ~ Fixed.Acidity + Volatile.Acidity + Citric.Acid +
##     Residual.Sugar + Chlorides + Free.Sulfur.Dioxide +
Total.Sulfur.Dioxide +
##     Density + pH + Sulphates + Alcohol, data = white.wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
## Fixed.Acidity         6.552e-02  2.087e-02   3.139  0.00171 **
## Volatile.Acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## Citric.Acid           2.209e-02  9.577e-02   0.231  0.81759
## Residual.Sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
## Chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
## Free.Sulfur.Dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
## Total.Sulfur.Dioxide -2.857e-04  3.781e-04  -0.756  0.44979
## Density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
## Sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
## Alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

```
#plot(white.wine.lm)

white.wine.lm2 <- lm(Quality ~ Volatile.Acidity + Residual.Sugar +
Free.Sulfur.Dioxide + Density + pH + Sulphates + Alcohol, data=white.wine)
#removing fixed acidity, citric acid, chlorides, and total sulfur dioxide
summary(white.wine.lm2)
```

```
##
## Call:
## lm(formula = Quality ~ Volatile.Acidity + Residual.Sugar +
Free.Sulfur.Dioxide +
##     Density + pH + Sulphates + Alcohol, data = white.wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8107 -0.4999 -0.0375  0.4636  3.2180
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.112e+02  1.273e+01   8.734  < 2e-16 ***
## Volatile.Acidity    -1.940e+00  1.085e-01 -17.872  < 2e-16 ***
## Residual.Sugar       6.637e-02  5.358e-03  12.386  < 2e-16 ***
## Free.Sulfur.Dioxide  3.283e-03  6.770e-04   4.849 1.28e-06 ***
## Density             -1.103e+02  1.274e+01  -8.653  < 2e-16 ***
## pH                   4.619e-01  7.638e-02   6.046 1.59e-09 ***
## Sulphates            5.708e-01  9.856e-02   5.791 7.42e-09 ***
## Alcohol              2.438e-01  1.870e-02  13.035  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.752 on 4890 degrees of freedom
## Multiple R-squared:  0.2801, Adjusted R-squared:  0.2791
## F-statistic: 271.8 on 7 and 4890 DF,  p-value: < 2.2e-16
```

```
#plot(white.wine.lm2)

white.wine.lm3 <- lm(Quality ~ Volatile.Acidity + Residual.Sugar + Density +
Alcohol, data=white.wine) #removing free sulfur dioxide, pH, and sulphates
summary(white.wine.lm3)
```

```
##
## Call:
## lm(formula = Quality ~ Volatile.Acidity + Residual.Sugar + Density +
##     Alcohol, data = white.wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3401 -0.5052 -0.0317  0.4723  3.1304
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        74.224822  11.977000   6.197 6.22e-10 ***
## Volatile.Acidity   -2.059334   0.108919 -18.907  < 2e-16 ***
## Residual.Sugar      0.052299   0.004914  10.642  < 2e-16 ***
## Density           -71.546483  11.922692  -6.001 2.10e-09 ***
## Alcohol             0.286371   0.017747  16.136  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7601 on 4893 degrees of freedom
## Multiple R-squared:  0.2639, Adjusted R-squared:  0.2633
## F-statistic: 438.6 on 4 and 4893 DF,  p-value: < 2.2e-16
```

```
#plot(white.wine.lm3)

white.wine.lm4 <- lm(Quality ~ Volatile.Acidity + Residual.Sugar + Alcohol,
data=white.wine) #removing density
summary(white.wine.lm4)
```

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

```
## lm(formula = Quality ~ Volatile.Acidity + Residual.Sugar + Alcohol,
##     data = white.wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3761 -0.4943 -0.0361  0.4649  3.0356
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.355675   0.113960   20.67   <2e-16 ***
## Volatile.Acidity -2.106709   0.109020  -19.32   <2e-16 ***
## Residual.Sugar    0.026607   0.002421   10.99   <2e-16 ***
## Alcohol           0.374572   0.009982   37.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7628 on 4894 degrees of freedom
## Multiple R-squared:  0.2585, Adjusted R-squared:  0.2581
## F-statistic: 568.8 on 3 and 4894 DF,  p-value: < 2.2e-16
```

```
#plot(white.wine.lm4)
```

4.2) Red Wine Multiple Linear Regression

```
red.wine.lm <- lm(Quality ~ Fixed.Acidity + Volatile.Acidity + Citric.Acid +
Residual.Sugar + Chlorides + Free.Sulfur.Dioxide + Total.Sulfur.Dioxide +
Density + pH + Sulphates + Alcohol, data=red.wine)
summary(red.wine.lm)
```

```
##
## Call:
## lm(formula = Quality ~ Fixed.Acidity + Volatile.Acidity + Citric.Acid +
##     Residual.Sugar + Chlorides + Free.Sulfur.Dioxide +
Total.Sulfur.Dioxide +
##     Density + pH + Sulphates + Alcohol, data = red.wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002
## Fixed.Acidity         2.499e-02  2.595e-02   0.963   0.3357
## Volatile.Acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## Citric.Acid          -1.826e-01  1.472e-01  -1.240   0.2150
## Residual.Sugar        1.633e-02  1.500e-02   1.089   0.2765
## Chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## Free.Sulfur.Dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
## Total.Sulfur.Dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## Density              -1.788e+01  2.163e+01  -0.827   0.4086
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
## Sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## Alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
#plot(red.wine.lm)

red.wine.lm2 <- lm(Quality ~ Volatile.Acidity + Chlorides +
Total.Sulfur.Dioxide + Sulphates + Alcohol, data=red.wine) #removing fixed
acidity, citric acid, residual sugar, free sulfur dioxide, density, and pH
summary(red.wine.lm2)
```

```
##
## Call:
## lm(formula = Quality ~ Volatile.Acidity + Chlorides + Total.Sulfur.Dioxide
+
##     Sulphates + Alcohol, data = red.wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67443 -0.38254 -0.06368  0.44893  2.07310
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.0048920  0.2037663  14.747  < 2e-16 ***
## Volatile.Acidity     -1.1419024  0.0969400 -11.779  < 2e-16 ***
## Chlorides            -1.7047871  0.3916886  -4.352 1.43e-05 ***
## Total.Sulfur.Dioxide -0.0023096  0.0005082  -4.544 5.92e-06 ***
## Sulphates             0.9148320  0.1102702   8.296 2.26e-16 ***
## Alcohol               0.2770979  0.0164836  16.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6514 on 1593 degrees of freedom
## Multiple R-squared:  0.3515, Adjusted R-squared:  0.3495
## F-statistic: 172.7 on 5 and 1593 DF,  p-value: < 2.2e-16
```

```
#plot(red.wine.lm2)

red.wine.lm3 <- lm(Quality ~ Volatile.Acidity + Sulphates + Alcohol,
data=red.wine) #removing chlorides and total sulfur dioxide
summary(red.wine.lm3)
```

```
##
## Call:
## lm(formula = Quality ~ Volatile.Acidity + Sulphates + Alcohol,
##     data = red.wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7186 -0.3820 -0.0641  0.4746  2.1807
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.61083    0.19569  13.342  < 2e-16 ***
## Volatile.Acidity -1.22140    0.09701 -12.591  < 2e-16 ***
## Sulphates         0.67903    0.10080   6.737 2.26e-11 ***
## Alcohol           0.30922    0.01580  19.566  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6587 on 1595 degrees of freedom
## Multiple R-squared:  0.3359, Adjusted R-squared:  0.3346
## F-statistic: 268.9 on 3 and 1595 DF,  p-value: < 2.2e-16

#plot(red.wine.lm3)
```

Report

I selected the Wine Quality dataset because I have always been very interested in wine and grew up around it. Even as a kid my parents would always ask me to evaluate the taste, density, smell, color, and more of new red and white wines with them and knew that analyzing the different components of wine for this project would be very interesting. When I first looked at the Red and White Wine datasets in the Wine Quality zip I was a bit overwhelmed as my data was visually a mess. Each column of data was separated by commas and semicolons so I knew that there would be a little extra work to do before I could even start my analysis.

In my attempts to wrangle the wine data I came across many difficulties. In sections 2.1 and 2.2 I started wrangling the White and Red Wine datasets and stumbled across many errors with my first formula as I forgot to include ‘header = FALSE’ and the Environment showed that there was only 1 variable instead of 12 for each of my datasets. Thankfully, adding ‘header = FALSE’ was able to solve most of my problems but it changed the titles of each of my variables so I needed to manually change the names of each one for both red and white wine datasets. The White Wine dataset was a bit more forgiving but the Red Wine dataset classified each variable as a character instead of a numeric or integer so I needed to manually fix that as well.

The ‘Big Question’ I attempted to answer through data analysis was what component(s) of wine impacts the quality of white vs red wine. I decided to look at Quality not only because it was in the title of my data zip but because it was the only variable that was an integer and not numeric. I was provided with 11 other variables and was curious what the most important components of wine would be when looking at Quality.

In my exploratory data analysis of White Wine in section 3.1 and Red Wine in section 3.2, I decided to make multiple scatterplots with regressions for each of my 11 variables as x and Quality as y. It was very interesting to look at how each variable impacted quality for white and red wine as some results for the same type of variable was very different for each type of wine. My original hypothesis during my exploratory analysis was that the variables with a less horizontal regression line, like Fixed Acidity, Volatile Acidity, Chlorides, Density and Alcohol for White Wine and Volatile Acidity, Chlorides, Sulphates, and Alcohol for Red Wine, would have the most impact on quality.

After further analysis, I decided to use multiple linear regression for both of my datasets in sections 4.1 and 4.2 to find which variable(s) was the most significant when being compared to Quality. I decided to do multiple rounds with the multiple linear regression formula, lm, using Quality as the response and the other variables as the predictors. Each time I removed the less significant variables until I was left with 3 variables for each type of wine that were equally highly significant. I found my method of using scatterplots and multiple linear regression to be quite helpful but also a bit tedious. I found myself repeating steps for every variable multiple times and constantly wondered if there was a faster or more efficient way to accomplish the same goal. Even though my method was a bit time-consuming, I still found the answer that I was looking for and know that with more experience in R that I will be able to run my analysis more efficiently in the future.

At the end of my multiple linear regression analysis, I was thankfully able to find the answer to my ‘Big Question’. For the White Wine multiple linear regression in section 4.1 I was able to isolate Volatile Acidity, Residual Sugar, and Alcohol as the most significant variables with Density being almost as significant. After looking back at the White Wine scatterplots from section 3.1 I was able to confirm my findings in section 4.1 by looking at the regressions for those 4 variables. I found that white wines with less volatile acidity and residual sugar that have a higher alcohol content and are less dense are generally of a higher quality. For the Red Wine multiple linear regression in section 4.2 I was able to isolate Volatile Acidity and Alcohol as the most significant variables with Sulphates being almost as significant. When looking back at the Red Wine scatterplots from section 3.2 I was able to confirm my findings in section 4.2 by looking at the regressions for those 3 variables. I found that red wines with less volatile acidity, more sulphates, and that have a higher alcohol content are generally of a higher quality.

I am quite content with the analysis that I made but am also interested as to how other aspects of wine impact Quality. Location, age, bottle size/shape, and harvest time are just a few additional variables that come to mind when considering the overall quality of red and white wines.