Exploratory Data Analysis

2024-05-04

EXPLORATORY DATA ANALYSIS

We began our EDA by conducting a preliminary review of the 2 datasets in terms of the number of data points, summary statistics of each predictor, and mean values of the outcome variable 'quality' in red and white wines.

```
#Loading the data into 2 different dataframes
redwines <- read.csv('winequality-red.csv', sep = ';')
whitewines <- read.csv('winequality-white.csv', sep = ';')

#No. of rows in both datasets
nrow(redwines)
```

```
## [1] 1599
```

```
nrow(whitewines)
```

```
## [1] 4898
```

The number of rows of data available for red wines is 1599, while there are 4898 data points for white wines.

```
#Summary of all the red and white wines predictors
summary(redwines)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901  
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956  
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968  
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
##        pH          sulphates         alcohol         quality     
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000  
```

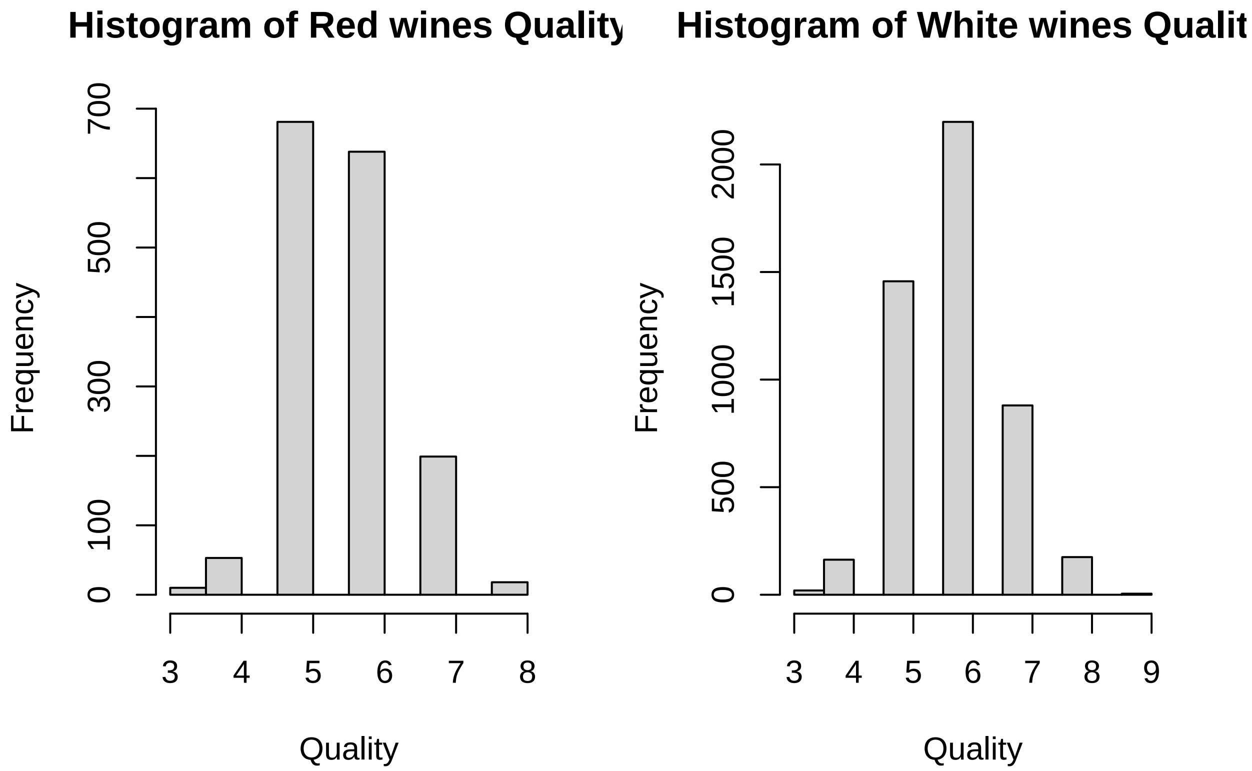
```
summary(whitewines)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar  
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

This is a summary of all the variables that make up both datasets. While it is difficult to draw any valuable conclusions from this directly, we see that the means quality of both types of wine are fairly similar - 5.636 vs. 5.878.

To further examine the distribution of the outcome variable between the 2 datasets, histograms in the following manner can be plotted:

```
par(mfrow = c(1,2))
hist(redwines$quality, xlab = 'Quality', main = 'Histogram of Red wines Quality')
hist(whitewines$quality, xlab = 'Quality', main = 'Histogram of White wines Quality')
```



Both appear to show a similar normal distribution representing a peak between 5 and 6 and tapering at the sides. There seems to be slightly more variance in the quality of white wines as they spread across more evenly, while the red wines' qualities are very closely concentrated around 5 and 6. This gives us some evidence that the 2 datasets might have different underlying true models, and should therefore be approached separately.

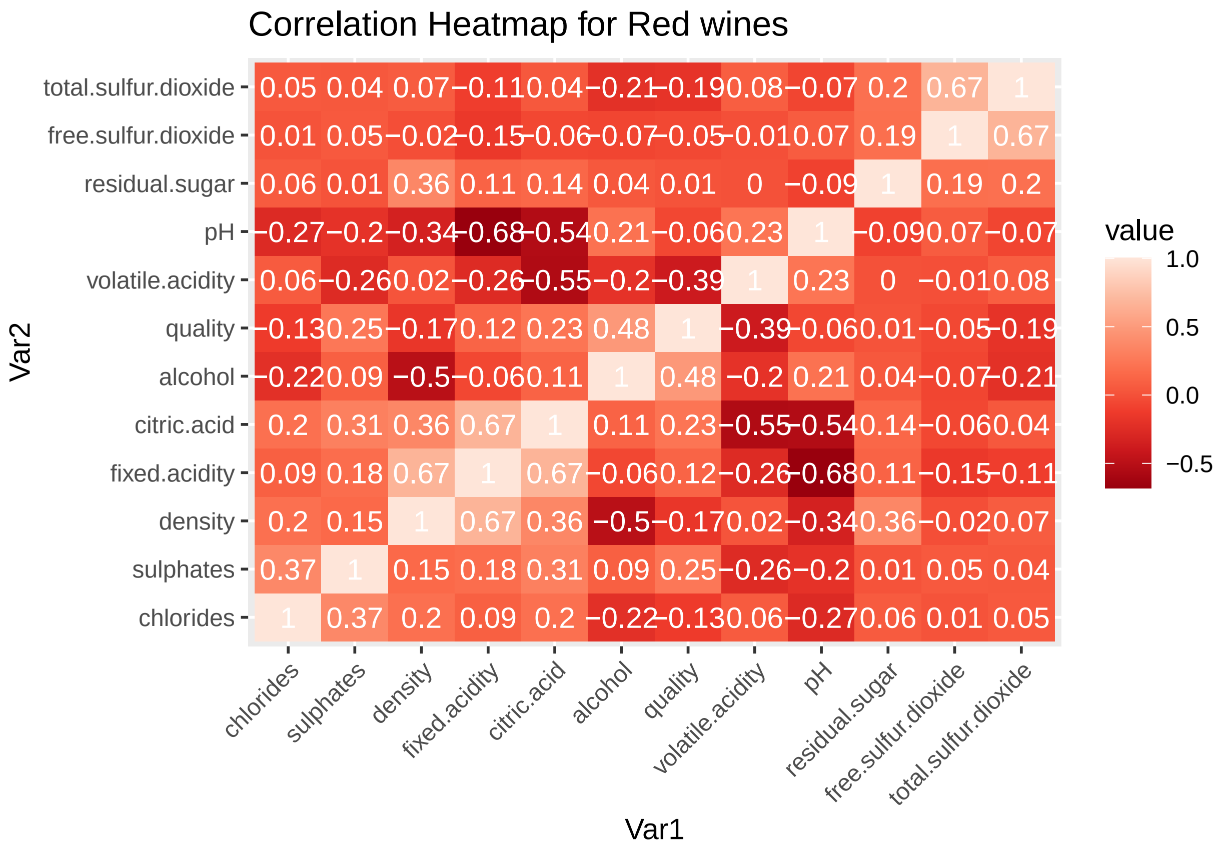
The correlations between each of the variables were then looked at, to check for any noticeable interaction effects that stood out. This would provide us insight to construct our models while accounting for any confounding interactions that would otherwise distort the fitting of any model we apply.

```r
cor_table_red <- cor(redwines)
cor_table_white <- cor(whitewines)
```

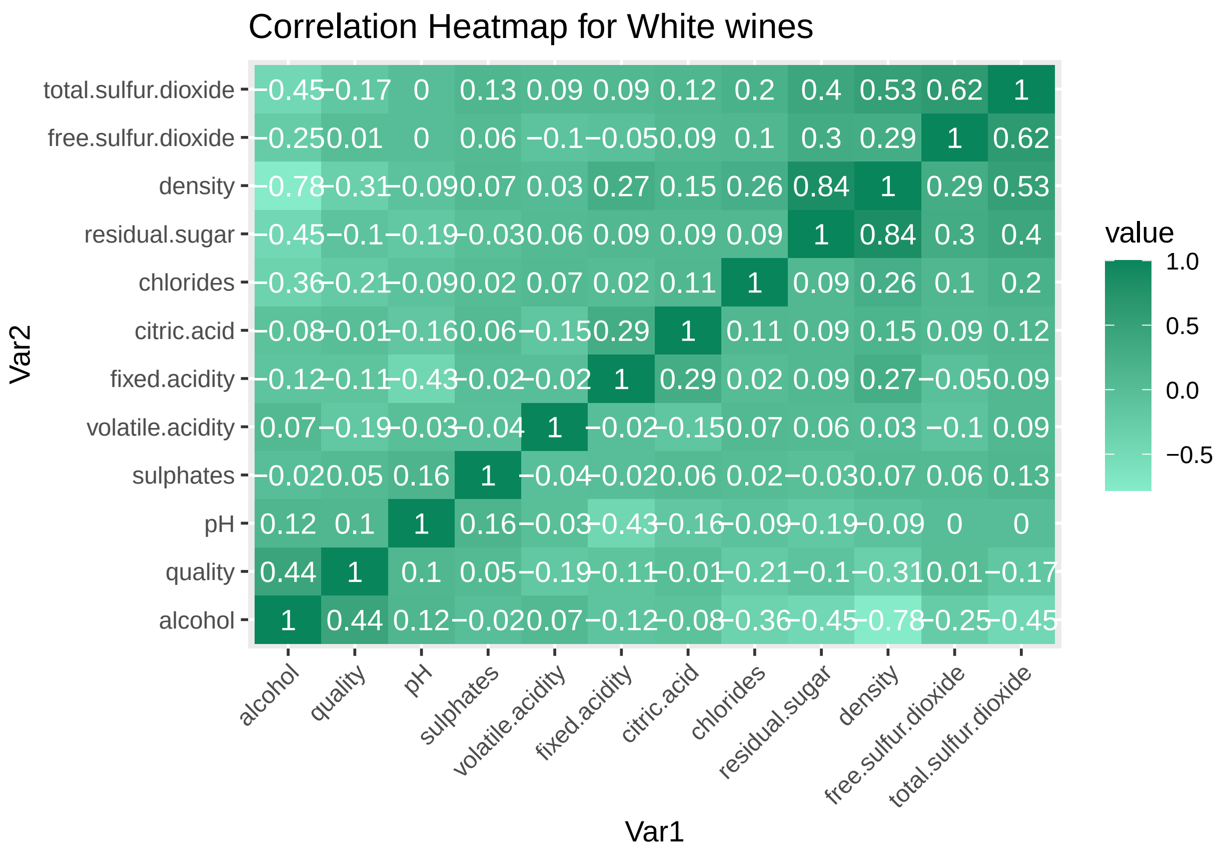
Plotting heatmaps to better visualize the interaction between variables in both datasets.

```r
library(reshape2)

corr_mat_red <- round(cor_table_red,2)
dist <- as.dist((1-corr_mat_red)/2)
hc <- hclust(dist)
corr_mat_red <-corr_mat_red[hc$order, hc$order]
melted_corr_mat_red <- melt(corr_mat_red)
library(ggplot2)
ggplot(data = melted_corr_mat_red, aes(x=Var1, y=Var2, fill=value)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_tile() +
  scale_fill_distiller(palette = "Reds") +
  ggtitle('Correlation Heatmap for Red wines') +
  geom_text(aes(Var2, Var1, label = value),
          color = "white", size = 4)
```



```
corr_mat_white <- round(cor_table_white,2)
dist <- as.dist((1-corr_mat_white)/2)
hc <- hclust(dist)
corr_mat_white <-corr_mat_white[hc$order, hc$order]
melted_corr_mat_white <- melt(corr_mat_white)
library(ggplot2)
ggplot(data = melted_corr_mat_white, aes(x=Var1, y=Var2, fill=value)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_tile() +
  scale_fill_gradient(low = "#86ebc9",
                    high = "#09855c",
                    guide = "colorbar") +
  ggtitle('Correlation Heatmap for White wines') +
  geom_text(aes(Var2, Var1, label = value),
          color = "white", size = 4)
```



The correlations between each predictor and the outcome variable were then calculated individually. The predictors and their correlations were then sorted in decreasing order.

```r
process_correlations <- function(dataset, title) {
  correlation_matrix <- cor(dataset[, sapply(dataset, is.numeric)], use="complete.obs")

  #print(correlation_matrix)

  melted_corr <- melt(correlation_matrix)

  heatmap_plot <- ggplot(data = melted_corr, aes(x = Var1, y = Var2, fill = value)) +
    geom_tile(color = "white") +
    scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0,
                         limit = c(-1, 1), space = "Lab", name = "Correlation") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
    labs(title = paste(title, "Correlation Heatmap"), x = "", y = "")

  quality_correlations <- correlation_matrix["quality",]
  sorted_correlations <- sort(abs(quality_correlations), decreasing = TRUE, na.last = NA)

  sorted_features <- quality_correlations[names(sorted_correlations[-1])]  # Exclude "quality" itself

  # Print all features against "quality" based on absolute correlation
  print(paste(title, "All Predictors Against Quality in Descending Order:"))
  print(sorted_features)
}
```

```r
process_correlations(redwines, "Red Wine")
```

```
## [1] "Red Wine All Predictors Against Quality in Descending Order:"
##              alcohol     volatile.acidity            sulphates 
##           0.47616632          -0.39055778           0.25139708 
##          citric.acid total.sulfur.dioxide              density 
##           0.22637251          -0.18510029          -0.17491923 
##            chlorides        fixed.acidity                   pH 
##          -0.12890656           0.12405165          -0.05773139 
##  free.sulfur.dioxide       residual.sugar 
##          -0.05065606           0.01373164
```

```r
process_correlations(whitewines, "White Wine")
```

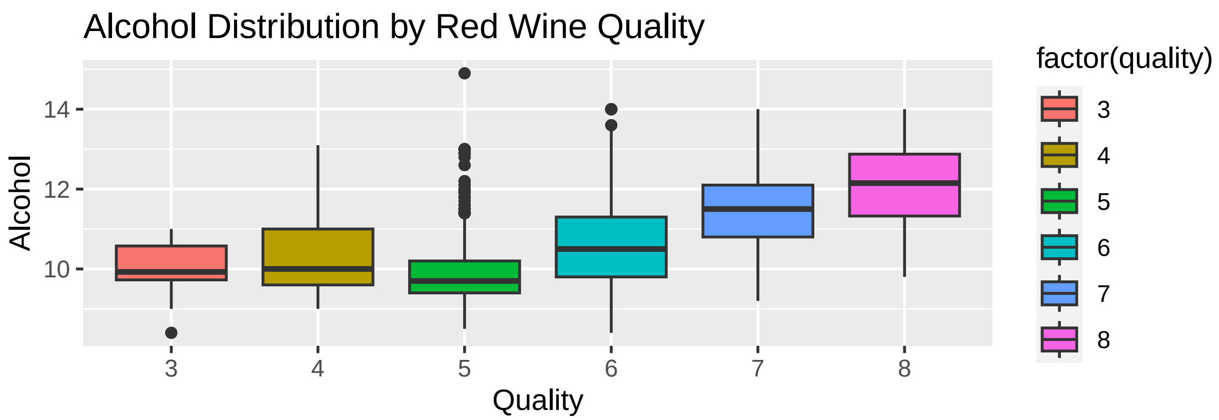
```
## [1] "White Wine All Predictors Against Quality in Descending Order:"
##              alcohol              density            chlorides 
##          0.435574715         -0.307123313         -0.209934411 
##     volatile.acidity total.sulfur.dioxide        fixed.acidity 
##         -0.194722969         -0.174737218         -0.113662831 
##                   pH       residual.sugar            sulphates 
##          0.099427246         -0.097576829          0.053677877 
##          citric.acid  free.sulfur.dioxide 
##         -0.009209091          0.008158067
```

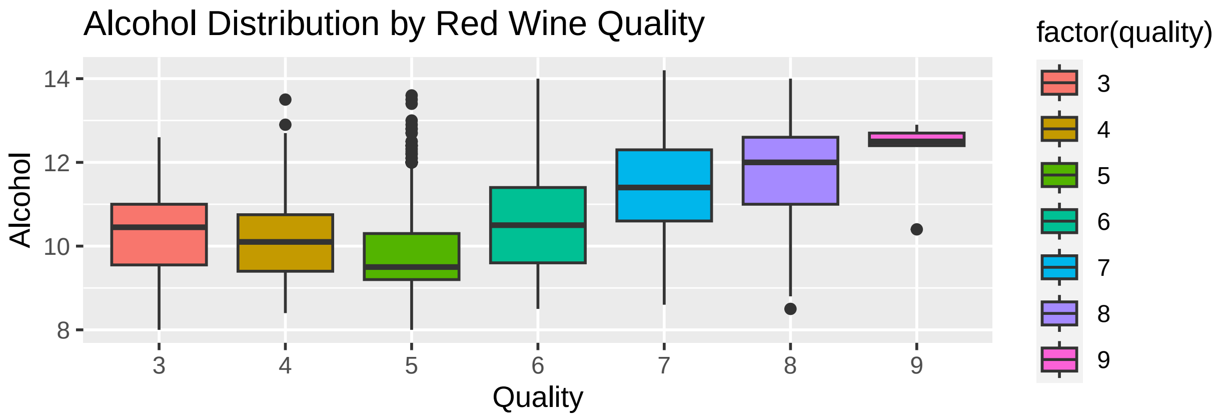
As seen above, the variable 'alcohol' seems to be the most important feature in predicting the quality of both red and white wines. Therefore, we decided to visualize its association with wine quality through 2 boxplots to investigate any obvious trends.

```r
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```r
plot1 <- ggplot(redwines, aes(x = factor(quality), y = alcohol)) +
  geom_boxplot(aes(fill = factor(quality))) +
  ggtitle("Alcohol Distribution by Red Wine Quality") +
  xlab("Quality") +
  ylab("Alcohol")
plot2 <- ggplot(whitewines, aes(x = factor(quality), y = alcohol)) +
  geom_boxplot(aes(fill = factor(quality))) +
  ggtitle("Alcohol Distribution by Red Wine Quality") +
  xlab("Quality") +
  ylab("Alcohol")
grid.arrange(plot1, plot2, nrow=2)
```





To gauge the linearity of the model, we applied preliminary linear models to look at the distribution of residuals of each predicted value. A Q-Q plot was also made to examine whether residuals followed a normal distribution.

```
prelim_linear_white <- glm(whitewines$quality ~
                       whitewines$fixed.acidity + whitewines$volatile.acidity + whitewines$citric.acid 
                       whitewines$residual.sugar + whitewines$chlorides + whitewines$free.sulfur.dioxid
                       whitewines$total.sulfur.dioxide + whitewines$density + whitewines$pH + whitewine
                       whitewines$sulphates + whitewines$alcohol,
            family = gaussian(link = "identity"), data = whitewines)

prelim_linear_red <- glm(redwines$quality ~
                       redwines$fixed.acidity + redwines$volatile.acidity + redwines$citric.acid +
                       redwines$residual.sugar + redwines$chlorides + redwines$free.sulfur.dioxide +
                       redwines$total.sulfur.dioxide + redwines$density + redwines$pH + redwines$pH +
                       redwines$sulphates + redwines$alcohol,
            family = gaussian(link = "identity"), data = redwines)

summary(prelim_linear_white)
```

```
##
## Call:
## glm(formula = whitewines$quality ~ whitewines$fixed.acidity +
##     whitewines$volatile.acidity + whitewines$citric.acid + whitewines$residual.sugar +
##     whitewines$chlorides + whitewines$free.sulfur.dioxide + whitewines$total.sulfur.dioxide +
##     whitewines$density + whitewines$pH + whitewines$pH + whitewines$sulphates +
##     whitewines$alcohol, family = gaussian(link = "identity"),
##     data = whitewines)
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      1.502e+02  1.880e+01   7.987 1.71e-15 ***
## whitewines$fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 ** 
## whitewines$volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## whitewines$citric.acid           2.209e-02  9.577e-02   0.231  0.81759    
## whitewines$residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
## whitewines$chlorides            -2.473e-01  5.465e-01  -0.452  0.65097    
## whitewines$free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
## whitewines$total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979    
## whitewines$density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## whitewines$pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
## whitewines$sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
## whitewines$alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.5645372)
## 
##     Null deviance: 3841.0  on 4897  degrees of freedom
## Residual deviance: 2758.3  on 4886  degrees of freedom
## AIC: 11113
## 
## Number of Fisher Scoring iterations: 2
```

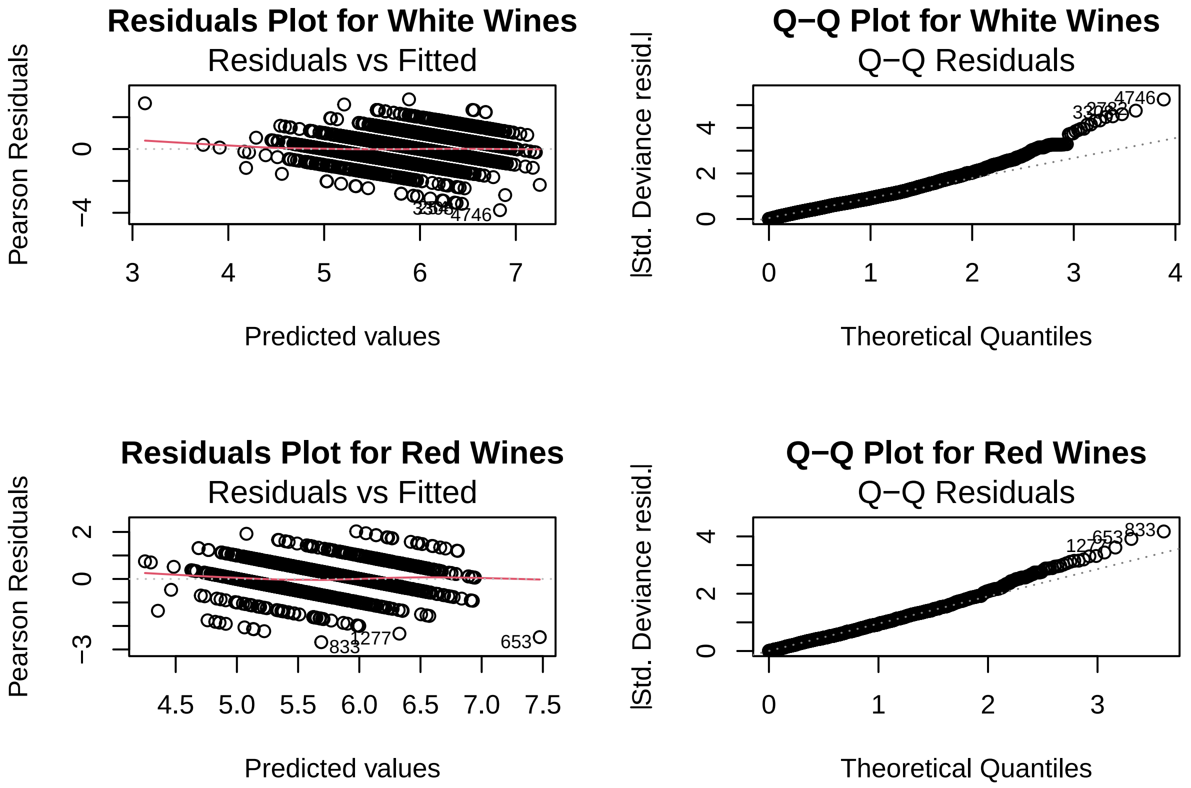
```
summary(prelim_linear_red)
```

```
## 
## Call:
## glm(formula = redwines$quality ~ redwines$fixed.acidity + redwines$volatile.acidity + 
##     redwines$citric.acid + redwines$residual.sugar + redwines$chlorides + 
##     redwines$free.sulfur.dioxide + redwines$total.sulfur.dioxide + 
##     redwines$density + redwines$pH + redwines$pH + redwines$sulphates + 
##     redwines$alcohol, family = gaussian(link = "identity"), data = redwines)
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    2.197e+01  2.119e+01   1.036   0.3002    
## redwines$fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357    
## redwines$volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## redwines$citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150    
## redwines$residual.sugar        1.633e-02  1.500e-02   1.089   0.2765    
## redwines$chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## redwines$free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *  
## redwines$total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## redwines$density              -1.788e+01  2.163e+01  -0.827   0.4086    
## redwines$pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *  
## redwines$sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## redwines$alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.4199185)
```

```
##
##     Null deviance: 1042.17  on 1598  degrees of freedom
## Residual deviance:  666.41  on 1587  degrees of freedom
## AIC: 3164.3
##
## Number of Fisher Scoring iterations: 2
```

```
par(mfrow = c(2,2))
plot(prelim_linear_white, which = 1, main = 'Residuals Plot for White Wines')
plot(prelim_linear_white, which = 2, main = 'Q-Q Plot for White Wines')

plot(prelim_linear_red, which = 1, main = 'Residuals Plot for Red Wines')
plot(prelim_linear_red, which = 2, main = 'Q-Q Plot for Red Wines')
```



Both residual plots above illustrate an extremely non-random scattering of residuals above the horizontal line, indicating a non-linear association between the predictors and outcome. This, in addition to the extreme deviations of the residuals in the Q-Q plots shows that a linear regression would not be appropriate to model either of the red or white wine datasets.