Predicting Wine Quality using Physicochemical Properties

Pat McCornack

2022-12-09

Dataset

I chose the wine quality dataset from the UCI Machine Learning repository for this analysis (https://archive.ics.uci.edu/ml/datasets/Wine+Quality). The data had already been cleaned prior to being distributed and had no quality issues that required pre-processing. After some research it seems that while sulphates are not used in wine-making, sulphites are widely used so I assumed there was a typo and changed the column name.

```
library(ggplot2)
library(dplyr)
r.wine <- read.csv("./wine-data/winequality-red.csv", sep = ";")
colnames(r.wine)[10] = "sulphites"
```

Big Question

The question I was interested in answering using this data set was whether the physicochemical properties of a wine could be used to predict the quality. In order to answer this question I chose to initially work with the red wine data and implement a regression model to attempt to predict a wines quality. The type of regression model was determined after some initial exploratory data analysis.

Before beginning to work with the data I did preliminary research on each of the parameters in order to get an idea of possible correlations between variables and which variables may be strong predictors. Based on this research I suspected that volatile acidity, alcohol content, and residual sugar content would have the biggest impact on quality. I also hypothesized that there would be a negative correlation between volatile acidity and citric acid as well as sulphites.

My belief that volatile acidity would be a strong negative predictor was due to its association with acetic acid content. High acetic acid levels indicated the wine had undesirable micro-organisms growing in it during fermentation. These micro-organisms compete for sugar with the yeast and produce acetic acid. Acetic acid will make a wine taste and smell of vinegar and therefore lower its quality.

The idea that alcohol and residual sugar content may also be predictors mostly came from personal observations that cheap wines typically have lower alcohol content and higher sugar content.

```
names(r.wine)
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphites"            "alcohol"              "quality"
```

Exploratory Data Analsis

Before attempting to create models, I wanted to understand the relationships within the data better. This was accomplished using data summaries and visualizations.

Wine Quality Summary

A high level overview of the data shows that we have 1,599 samples that are described using 12 parameters. Both volatile acidity and residual sugar content each have maximum values far beyond the central measures. These outliers will be further explored later on. Alcohol content seems to have a much more even spread. Most important is the observation that the dataset is very imbalanced with the vast majority of observations being rated either 5 or 6. We will keep this imbalance in mind throughout the analysis.

Note that while this summarization was performed with quality as a factor variable, the regression models were built by modeling quality as a continuous number.

```
r.wine.qual <- r.wine
r.wine.qual$quality <- as.factor(r.wine$quality)

str(r.wine.qual)
```

```
## '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 ...
##  $ sulphites           : 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             : Factor w/ 6 levels "3","4","5","6",..: 3 3 3 4 3 3 3 5 5 3 ...
```

```
summary(r.wine.qual)
```

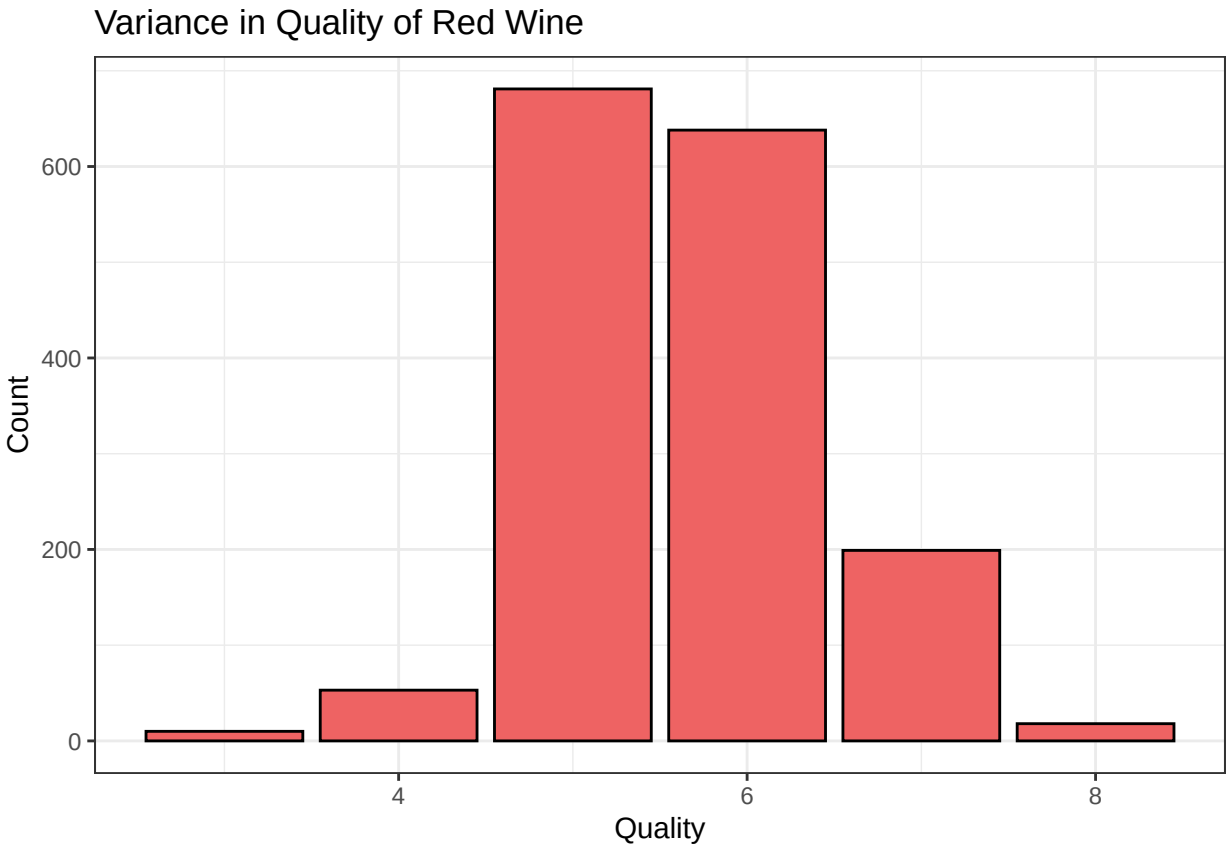
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :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          sulphites         alcohol      quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   3: 10
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   4: 53
##  Median :3.310   Median :0.6200   Median :10.20   5:681
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   6:638
```

```
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   7:199
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   8: 18
```

Checking variance of parameters of interest

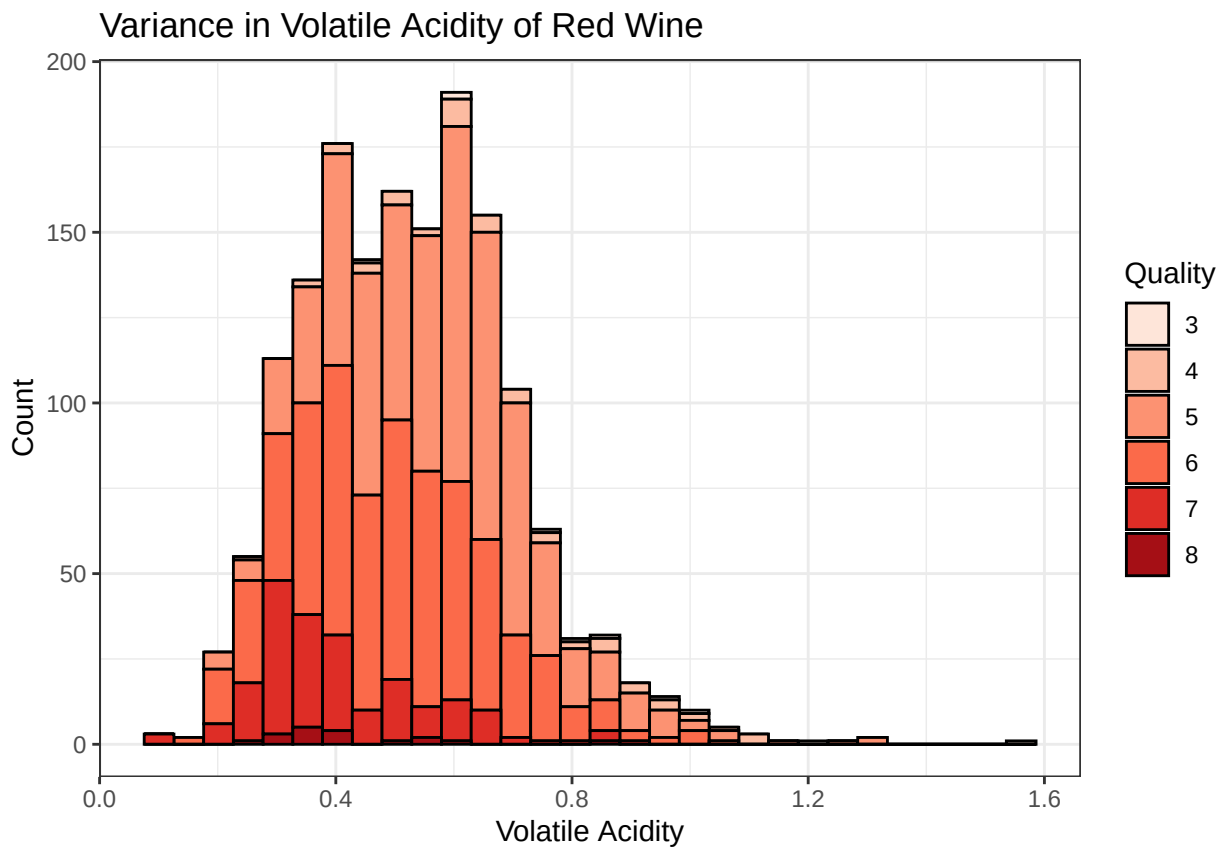
As noted there are far more samples rated 5 and 6 than the higher and lower scores.

```
ggplot(data = r.wine) +
  geom_bar(aes(x=quality), fill = "indianred2", color = "black") +
  labs(title = "Variance in Quality of Red Wine",
       x = "Quality",
       y = "Count") +
  theme_bw()
```



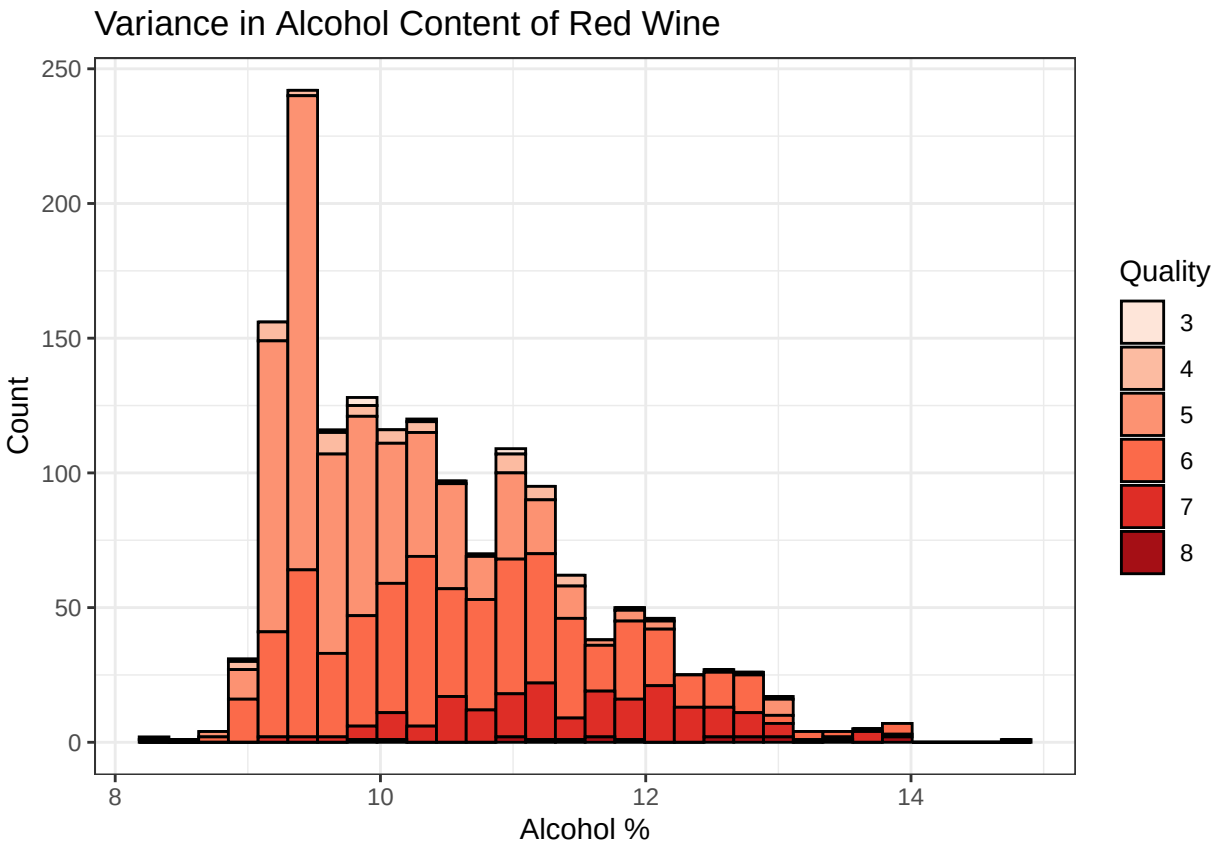
The histogram of volatile acidity shows that the values are fairly well distributed about the center, but there is a tail to the right. We also note the outlier we noticed earlier that has a value of around 1.6. The distribution of the quality does seem to be biased towards wines with lower levels of volatile acidity which supports our hypothesis

```
ggplot(data = r.wine) +
  geom_histogram(aes(x=volatile.acidity, fill = r.wine.qual$quality), color = "black", bins = 30) +
  labs(title = "Variance in Volatile Acidity of Red Wine",
       x = "Volatile Acidity",
       y = "Count") +
  theme_bw() +
  scale_fill_brewer(palette="Reds") +
  guides(fill=guide_legend("Quality"))
```



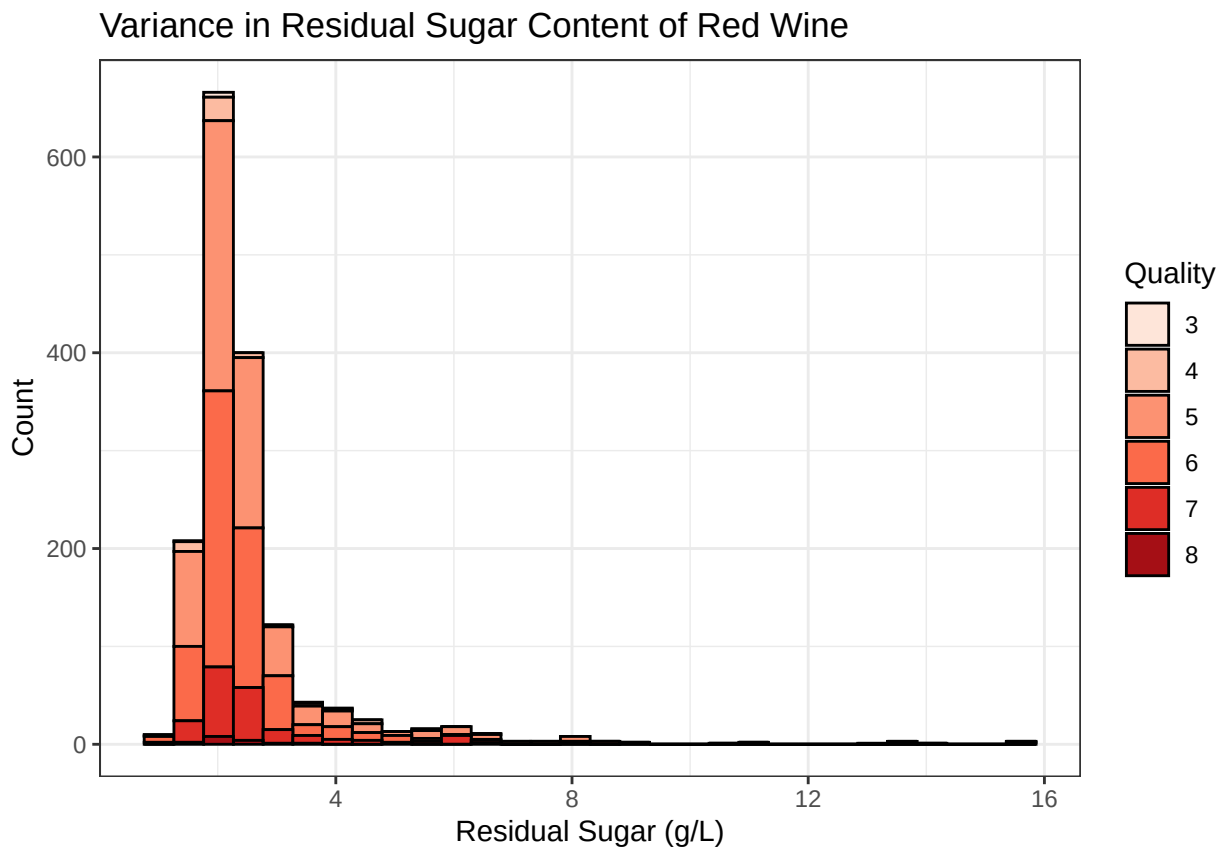
The alcohol content of red wines in this dataset is skewed to the right with one notable outlier. Interestingly the distribution of the wines rated 7 or 8 do not match the overall distribution and occupy a disproportionate number of the wines with higher percentages. Similarly, the majority of the lower quality wine with ratings of 3 or 4 occupy the left side of the distribution. This suggests that alcohol may be used as a strong predictor of quality.

```
ggplot(data = r.wine) +
  geom_histogram(aes(x=alcohol, fill = r.wine.qual$quality), color = "black", bins = 30) +
  labs(title = "Variance in Alcohol Content of Red Wine",
       x = "Alcohol %",
       y = "Count") +
  theme_bw() +
  scale_fill_brewer(palette="Reds") +
  guides(fill=guide_legend("Quality"))
```



The residual sugar content is also right-tailed, but only a small percentage of wines had residual sugar contents over 4 g/L. The distribution of each quality does seem to match the overall distribution, suggesting that the residual sugar content is not a strong predictor. However, this could also be due to the fact that the dataset is so imbalanced.

```
ggplot(data = r.wine) +
  geom_histogram(aes(x=residual.sugar, fill = r.wine.qual$quality), color = "black") +
  labs(title = "Variance in Residual Sugar Content of Red Wine",
       x = "Residual Sugar (g/L)",
       y = "Count") +
  theme_bw() +
  scale_fill_brewer(palette="Reds") +
  guides(fill=guide_legend("Quality"))
```



Checking for covariance within the dataset

For a quick overview of linear correlation among variables, we create a correlation matrix. As suspected, alcohol and volatile acidity are relatively highly correlated with quality. This also confirms that residual sugar is not highly correlated with quality, as was suggested by the histogram. Some other parameters that this suggests may be of interest are citric acid and sulphites. Both of these are used to control the chemical environment of the wine and reduce the growth of undesirable micro-organisms, so I suspected that there was collinearity between volatile acidity and citric acid as well as with sulphite content. This was checked visually using scatterplots to find any potential relationships.

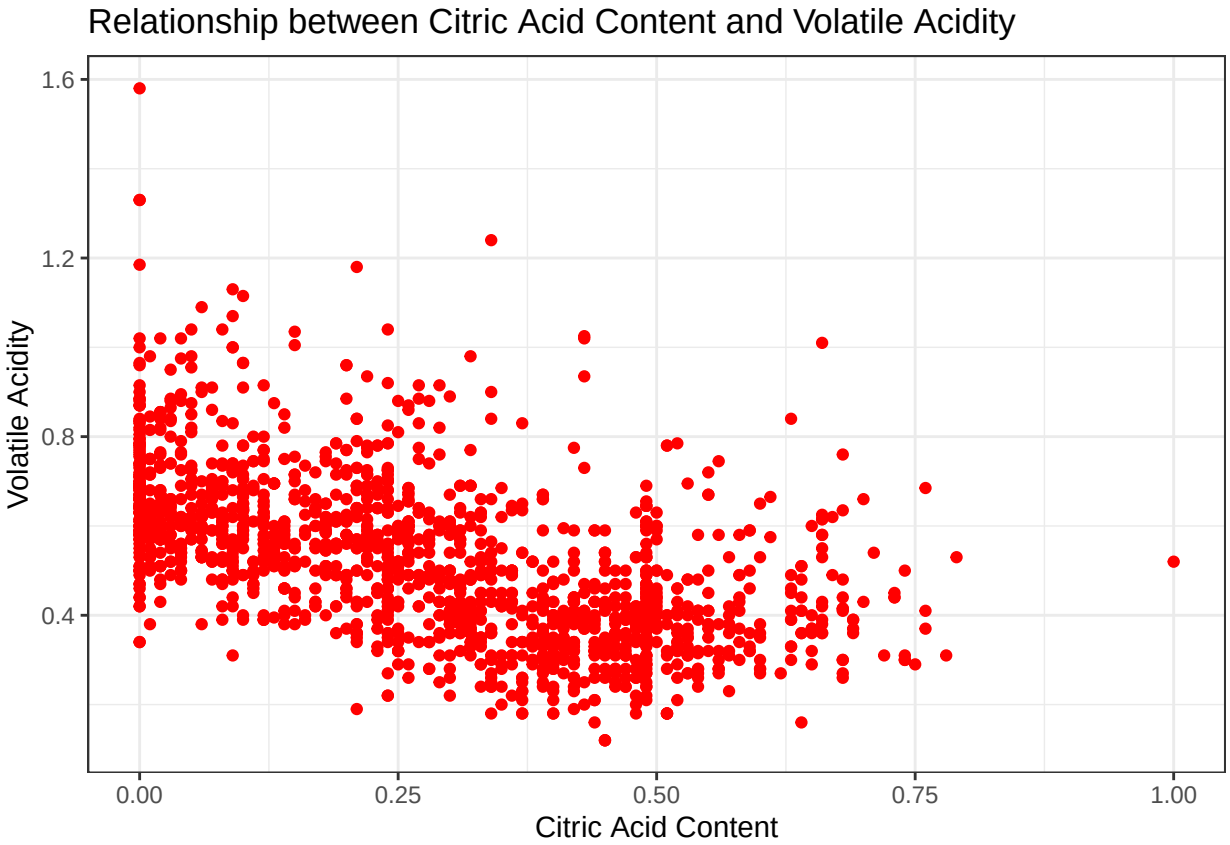
```
cor(r.wine)
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724
## volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000
## chlorides               0.09370519      0.061297772  0.20382291    0.055609535
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882
## density                 0.66804729      0.022026232  0.36494718    0.355283371
## pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphites               0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437
## quality                 0.12405165     -0.390557780  0.22637251    0.013731637
##                       chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.093705186        -0.153794193          -0.11318144
```

```
## volatile.acidity      0.061297772        -0.010503827           0.07647000
## citric.acid           0.203822914        -0.060978129           0.03553302
## residual.sugar        0.055609535         0.187048995           0.20302788
## chlorides             1.000000000         0.005562147           0.04740047
## free.sulfur.dioxide   0.005562147         1.000000000           0.66766645
## total.sulfur.dioxide  0.047400468         0.667666450           1.00000000
## density               0.200632327        -0.021945831           0.07126948
## pH                   -0.265026131         0.070377499          -0.06649456
## sulphites             0.371260481         0.051657572           0.04294684
## alcohol              -0.221140545        -0.069408354          -0.20565394
## quality              -0.128906560        -0.050656057          -0.18510029
##                         density          pH    sulphites     alcohol
## fixed.acidity         0.66804729 -0.68297819  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.23493729 -0.260986685 -0.20228803
## citric.acid           0.36494718 -0.54190414  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.08565242  0.005527121  0.04207544
## chlorides             0.20063233 -0.26502613  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.07037750  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.06649456  0.042946836 -0.20565394
## density               1.00000000 -0.34169933  0.148506412 -0.49617977
## pH                   -0.34169933  1.00000000 -0.196647602  0.20563251
## sulphites             0.14850641 -0.19664760  1.000000000  0.09359475
## alcohol              -0.49617977  0.20563251  0.093594750  1.00000000
## quality              -0.17491923 -0.05773139  0.251397079  0.47616632
##                         quality
## fixed.acidity         0.12405165
## volatile.acidity     -0.39055778
## citric.acid           0.22637251
## residual.sugar        0.01373164
## chlorides            -0.12890656
## free.sulfur.dioxide  -0.05065606
## total.sulfur.dioxide -0.18510029
## density              -0.17491923
## pH                   -0.05773139
## sulphites             0.25139708
## alcohol               0.47616632
## quality               1.00000000
```

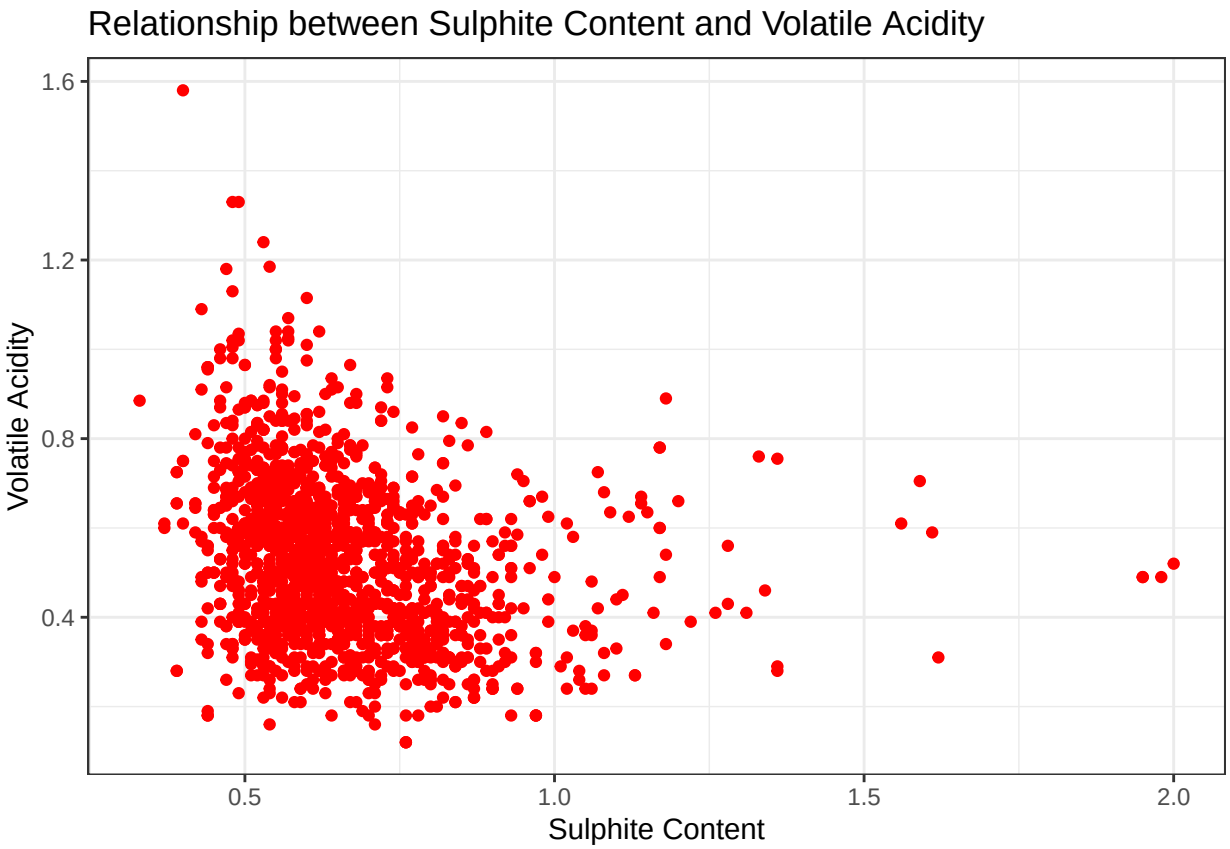
Using a scatterplot to examine the relationship between citric acid and volatile acidity reveals a strong negative linear trend between them. This is unsurprising given the function of citric acid in wine-making.

```
ggplot(data = r.wine.qual) +
  geom_point(aes(x = citric.acid, y = volatile.acidity), color = "Red") +
  labs(title = "Relationship between Citric Acid Content and Volatile Acidity",
       x = "Citric Acid Content",
       y = "Volatile Acidity") +
  theme_bw()
```



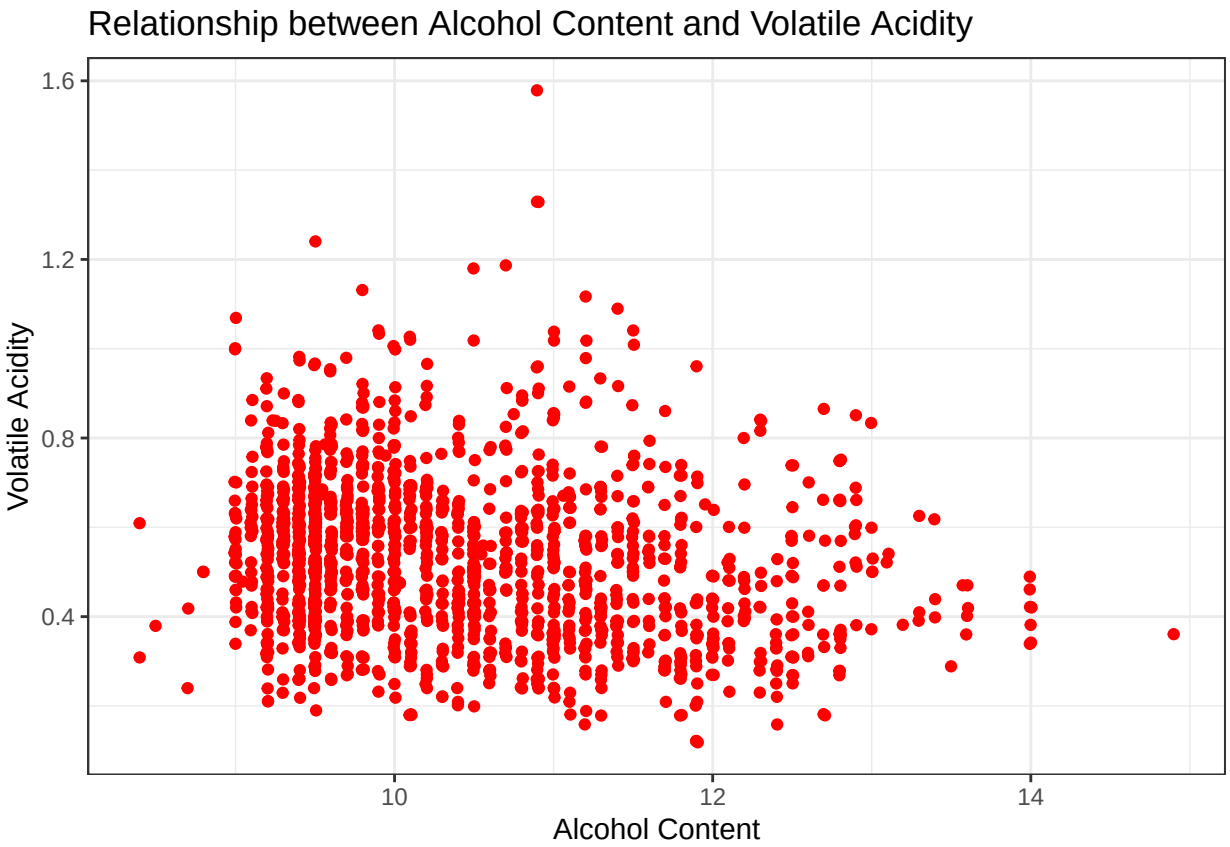
The method reveals what may be a slight negative trend between sulphite content and volatile acidity. However, the trend may not be linear and the points towards the right of the graph caused me to investigate the relationship between sulphite content and quality further using a histogram.

```
ggplot(data = r.wine.qual) +
  geom_point(aes(x = sulphites, y = volatile.acidity), color = "Red") +
  labs(title = "Relationship between Sulphite Content and Volatile Acidity",
       x = "Sulphite Content",
       y = "Volatile Acidity") +
  theme_bw()
```



Lastly, since at this point it appears that alcohol content and volatile acidity are the strongest predictors we want to check whether there is colinearity between them. The chart shows there may be a slight downward trend, but this is more likely due to the high concentration of observations with alcohol contents of 9-10%. Overall there doesn't seem to be colinearity between the alcohol content and volatile acidity.

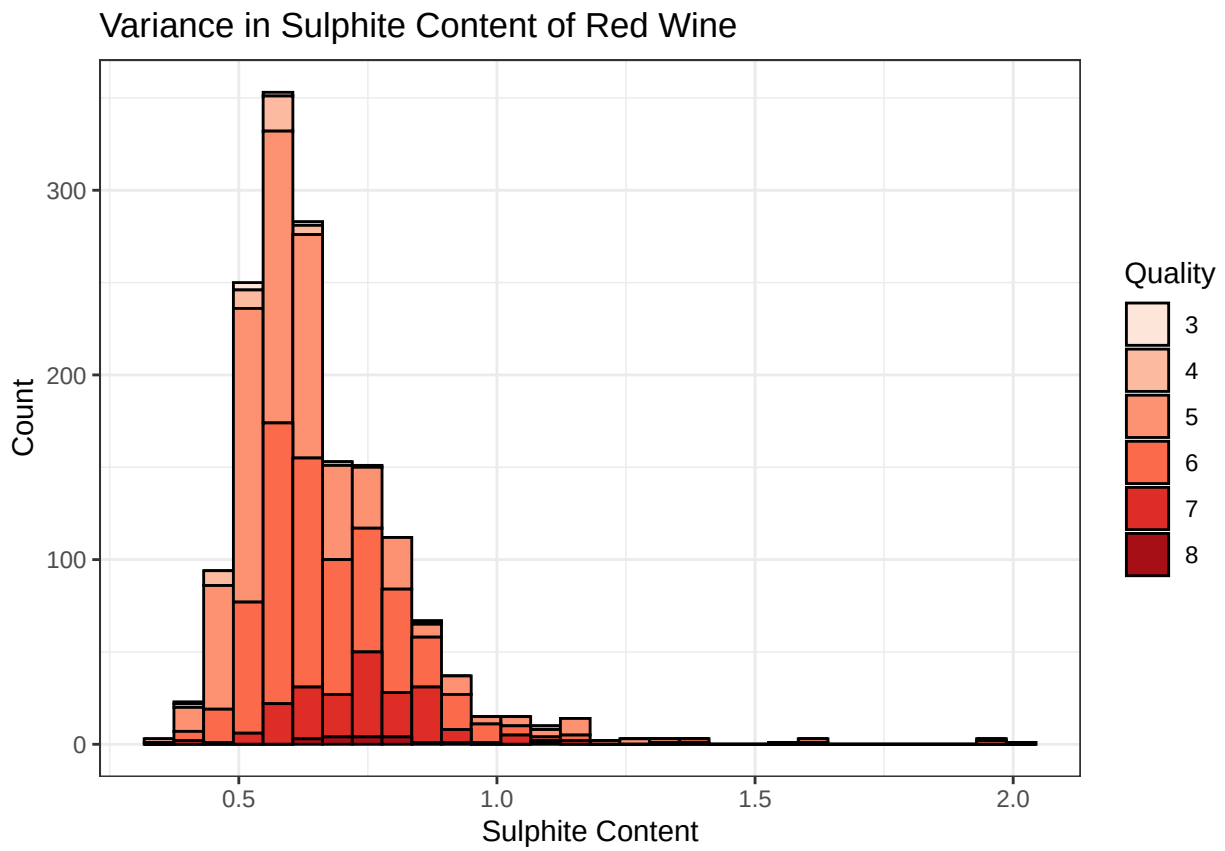
```
ggplot(data = r.wine.qual) +
  geom_jitter(aes(x = alcohol, y = volatile.acidity), color = "Red") +
  labs(title = "Relationship between Alcohol Content and Volatile Acidity",
       x = "Alcohol Content",
       y = "Volatile Acidity") +
  theme_bw()
```



Further Investigation of Sulphite Content

The distribution of the higher quality wines does fully match the overall distribution of sulphites vs. wine. This suggests there may be some predictive power in using sulphites to predict quality, but this could be due to the colinearity between sulphite content and volatile acidity shown above. Investigating the outlying observations where the sulphite content was greater than 1 does not yield any interesting conclusions.

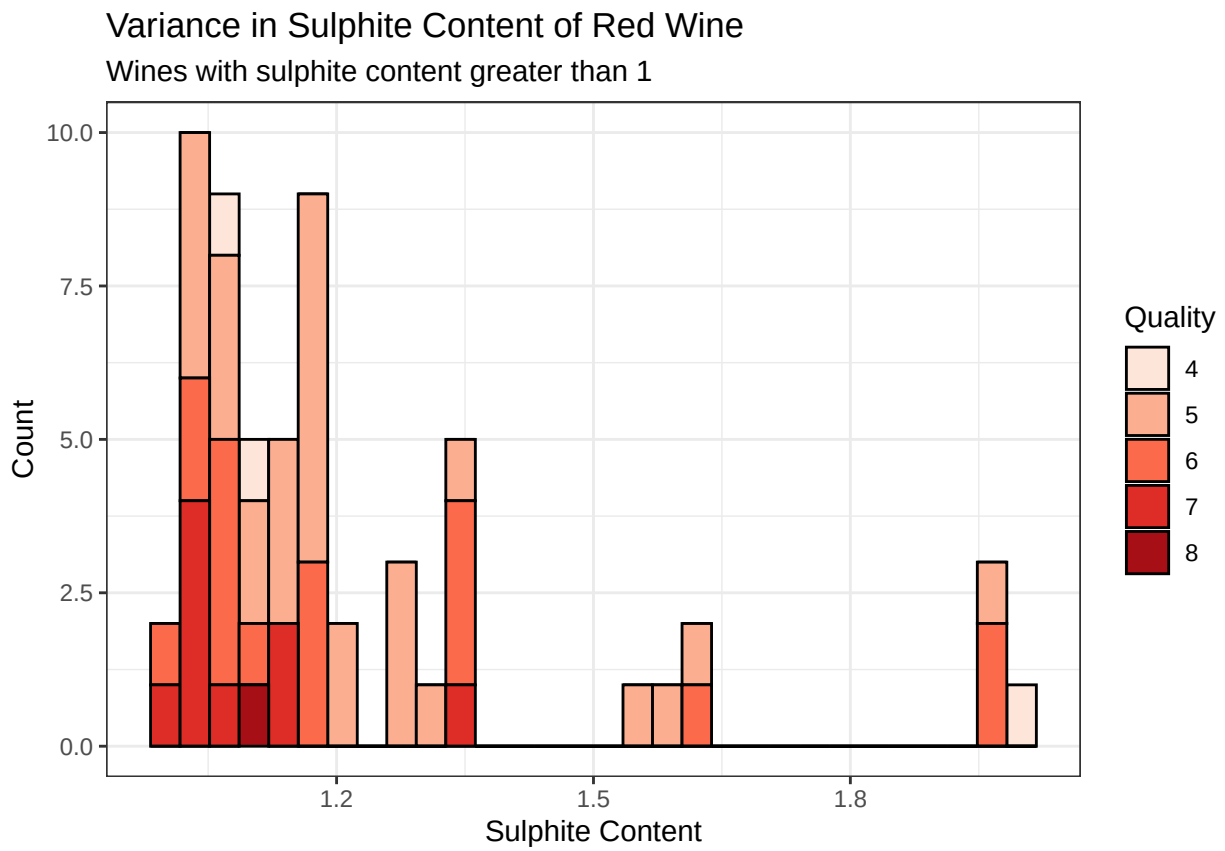
```
ggplot(data = r.wine) +
  geom_histogram(aes(x=sulphites, fill = r.wine.qual$quality), color = "black") +
  labs(title = "Variance in Sulphite Content of Red Wine",
       x = "Sulphite Content",
       y = "Count") +
  theme_bw() +
  scale_fill_brewer(palette="Reds") +
  guides(fill=guide_legend("Quality"))
```




```
high.sulphite <- r.wine.qual[r.wine.qual$sulphites >= 1, ]
high.sulphite$quality <- as.factor(high.sulphite$quality)
dim(high.sulphite)
```

```
## [1] 59 12
```

```
ggplot(data = high.sulphite) +
  geom_histogram(aes(x=sulphites, fill = quality), color = "black") +
  labs(title = "Variance in Sulphite Content of Red Wine",
       subtitle = "Wines with sulphite content greater than 1",
       x = "Sulphite Content",
       y = "Count") +
  theme_bw() +
  scale_fill_brewer(palette="Reds") +
  guides(fill=guide_legend("Quality"))
```



Modeling Wine Quality

At this point I suspect that alcohol content and volatile acidity will be the strongest predictors of wine quality. To investigate their predictive power, I created a simple linear regression model for each predictor and evaluated them using cross-validation. I then created a multiple linear regression model using both predictors and evaluated its performance using cross-validation and compared the results against those of a model using every parameter in the dataset as a predictor. While linear regression was used, the prediction for each observation was rounded to the nearest whole number for comparison.

Simple Linear Regression

Cross validation was performed using 5 folds and models were evaluated using accuracy and compared using mean-squared error (MSE).

Cross-validate the models:

```
set.seed(1234)  # For reproducibility

k <- 5
ncv <- ceiling(nrow(r.wine)/k)
cv.ind <- rep(1:k, ncv)
cv.ind.rand <- sample(cv.ind, nrow(r.wine))

MSE.cv.alc <- c()
MSE.cv.vol <- c()
```

```
for(j in 1:k){
  train <- r.wine[cv.ind.rand != j, ]  # Training predictor
  model.alc <- lm(quality ~ alcohol, train)
  model.vol <- lm(quality ~ volatile.acidity, train)

  test <- r.wine[cv.ind.rand == j, ]  # Test predictor

  cv.pred.alc <- predict(model.alc, newdata = test)
  cv.pred.alc <- round(cv.pred.alc, 0)

  cv.pred.vol <- predict(model.vol, newdata = test)
  cv.pred.vol <- round(cv.pred.vol, 0)

  MSE.cv.alc[j] <- sum((test$quality - cv.pred.alc)^2) / nrow(test)
  MSE.cv.vol[j] <- sum((test$quality - cv.pred.vol)^2) / nrow(test)
}
```

Analyze the Results:

Using a confusion matrix for each model, we find that the model that just used alcohol to predict quality was correct 57% of the time while the model that just used volatile acidity was correct 52% of the time. The MSE of the alcohol model was also lower than that of the volatile acidity model. While the alcohol model did perform marginally better, neither performed exceptionally well. This was likely due to bias introduced by most of the samples being rated either a 5 or 6. The visualizations reveal that for the most part, the model just guessed that each sample was either 5 or 6.

```
mean(MSE.cv.alc)
```

```
## [1] 0.6053625
```

```
mean(MSE.cv.vol)
```

```
## [1] 0.648509
```

```
table(cv.pred.alc, test$quality)
```

```
##
## cv.pred.alc  3  4  5  6  7  8
##           5  1  1 98 46  6  2
##           6  0  2 41 83 28  0
##           7  0  0  2  5  2  2
```

```
table(cv.pred.vol, test$quality)
```

```
##
## cv.pred.vol   3   4   5   6   7   8
##           4   0   0   1   0   0   0
##           5   1   0  64  32   3   2
##           6   0   3  76 102  33   2
```

```
correct.alc = (98 + 83 + 2) / nrow(test)
correct.vol = (64 + 102) / nrow(test)

correct.alc
```

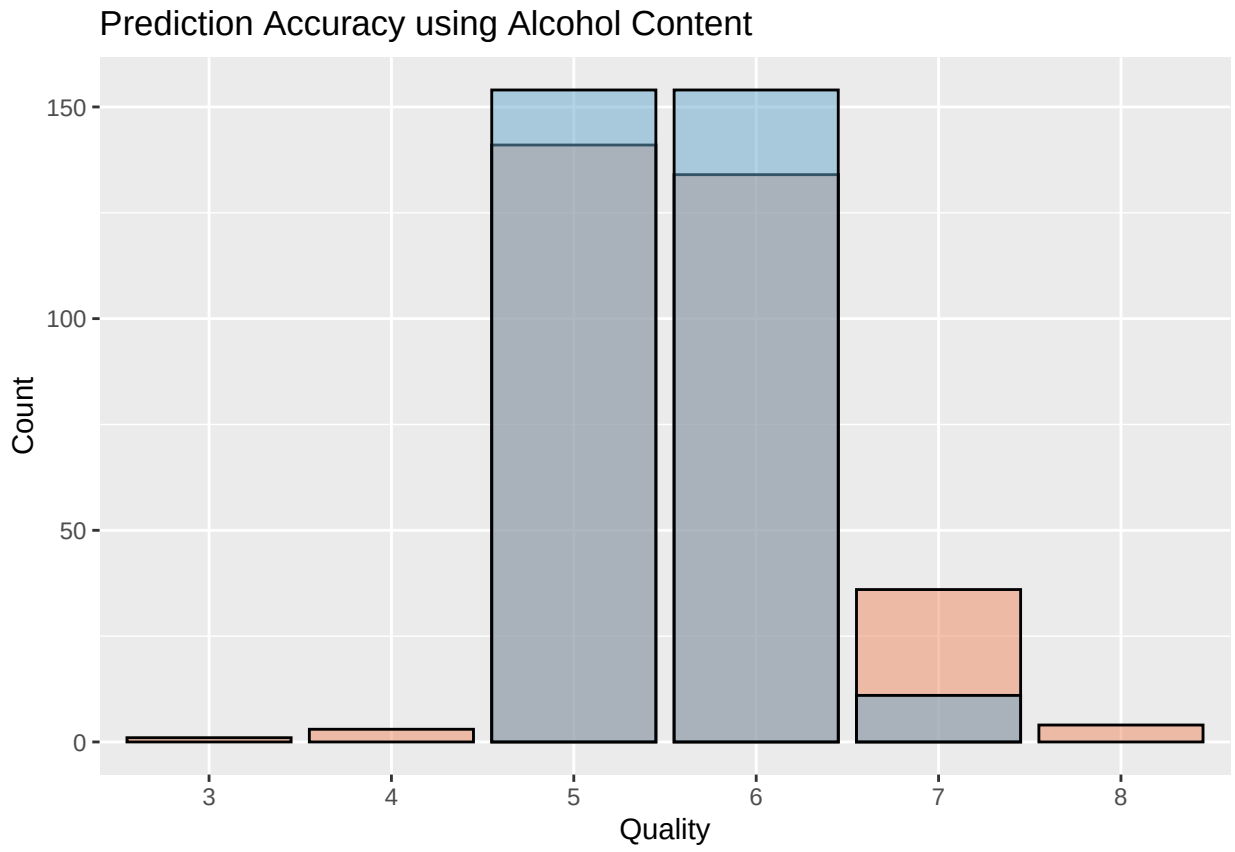
```
## [1] 0.5736677
```

```
correct.vol
```

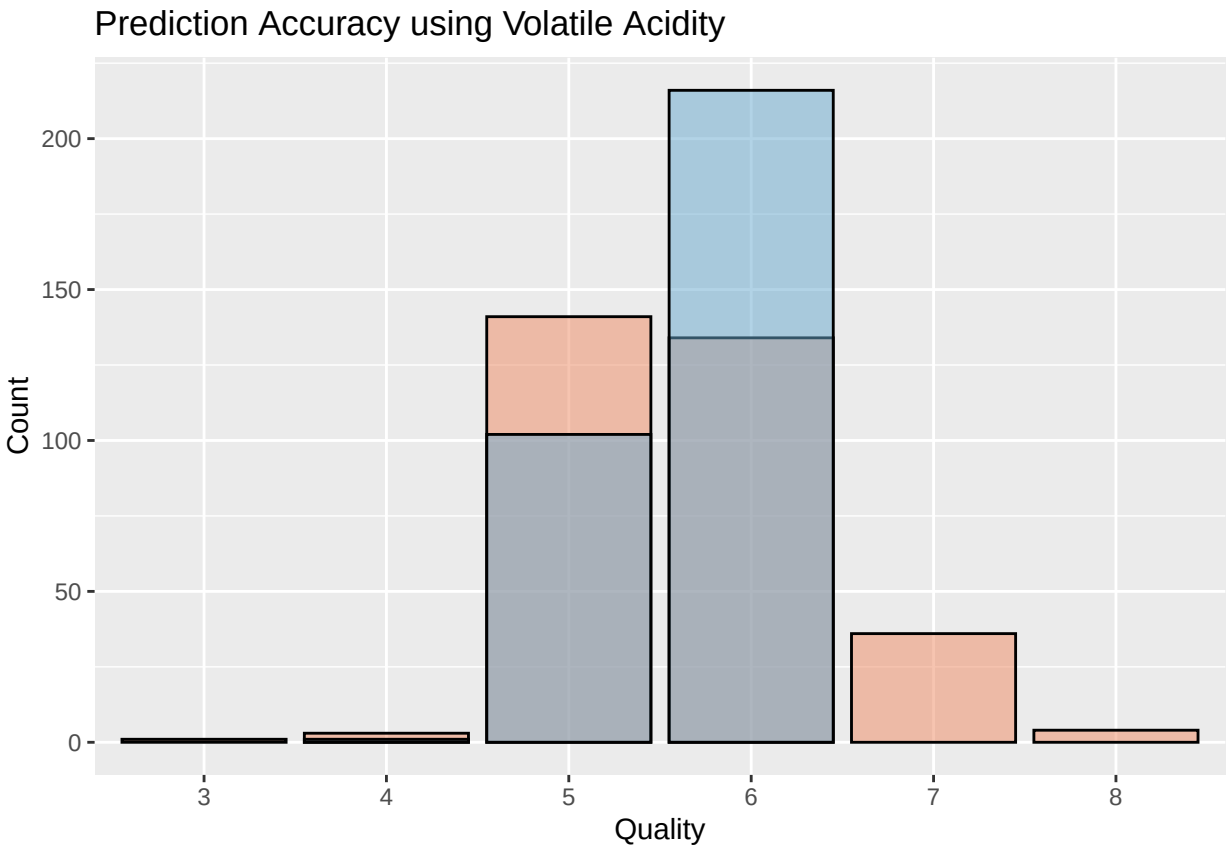
```
## [1] 0.5203762
```

```
# Make quality a factor to analyze the results:
test$quality <- as.factor(test$quality)
cv.pred.alc <- as.factor(cv.pred.alc)
cv.pred.vol <- as.factor(cv.pred.vol)

ggplot(data = test) +
  geom_bar(aes(x = quality), color = "Black", fill = "#ef8a62", alpha = 0.5) +
  geom_bar(aes(x = cv.pred.alc), color = "Black", fill = "#67a9cf", alpha = 0.5) +
  labs(title = "Prediction Accuracy using Alcohol Content",
       x = "Quality",
       y = "Count")
```




```
ggplot(data = test) +
  geom_bar(aes(x = quality), color = "Black", fill = "#ef8a62", alpha = 0.5) +
  geom_bar(aes(x = cv.pred.vol), color = "Black", fill = "#67a9cf", alpha = 0.5) +
  labs(title = "Prediction Accuracy using Volatile Acidity",
       x = "Quality",
       y = "Count")
```



Multiple Linear Regression

Once again 5-fold cross validation was used to evaluate each model. One model was created using just volatile acidity and alcohol, while the other used every parameter in the dataset as a predictor. Models were evaluated by their accuracy and compared using mean-squared-error (MSE).

Cross-Validate the Models:

```
k <- 5
ncv <- ceiling(nrow(r.wine)/k)
cv.ind <- rep(1:k, ncv)
cv.ind.rand <- sample(cv.ind, nrow(r.wine))

MSE.cv.1 <- c()
MSE.cv.2 <- c()

for(j in 1:k){
  train <- r.wine[cv.ind.rand != j, ]  # Training predictor
  model.1 <- lm(quality ~ volatile.acidity + alcohol, train)
  model.2 <- lm(quality ~ ., train)

  test <- r.wine[cv.ind.rand == j, ]  # Test predictor

  cv.pred.1 <- predict(model.1, newdata = test)
  cv.pred.1 <- round(cv.pred.1, 0)

  cv.pred.2 <- predict(model.2, newdata = test)
```

```
  cv.pred.2 <- round(cv.pred.2, 0)

  MSE.cv.1[j] <- sum((test$quality - cv.pred.1)^2) / nrow(test)
  MSE.cv.2[j] <- sum((test$quality - cv.pred.2)^2) / nrow(test)

}
```

Analyze the Results:

The model using alcohol content and volatile acidity to predict quality was correct 59% of the time and had a slightly higher MSE than the model using all predictors which was correct 60% of the time. Based on the p-value of the second model, many of the predictors were not statistically significant.

While the model using all the predictors was slightly more accurate, the added complexity from using all predictors vs. using just two makes the model using just alcohol and volatile acidity better. Within this model, the weight of volatile acidity is higher than the weight of alcohol which means that it was a stronger predictor.

```
mean(MSE.cv.1)
```

```
## [1] 0.5309365
```

```
mean(MSE.cv.2)
```

```
## [1] 0.5059267
```

```
table(cv.pred.1, test$quality)
```

```
##
## cv.pred.1  3  4  5  6  7  8
##         4  0  0  1  0  0  0
##         5  1  9 94 35  1  0
##         6  0  3 45 86 25  4
##         7  0  0  0  6  7  2
```

```
table(cv.pred.2, test$quality)
```

```
##
## cv.pred.2  3  4  5  6  7  8
##         4  0  0  1  0  0  0
##         5  1  9 96 35  1  0
##         6  0  3 43 86 22  4
##         7  0  0  0  6 10  2
```

```
correct.1 = (94 + 86 + 7) / nrow(test)
correct.2 = (96 + 86 + 10) / nrow(test)

correct.1
```

```
## [1] 0.5862069
```

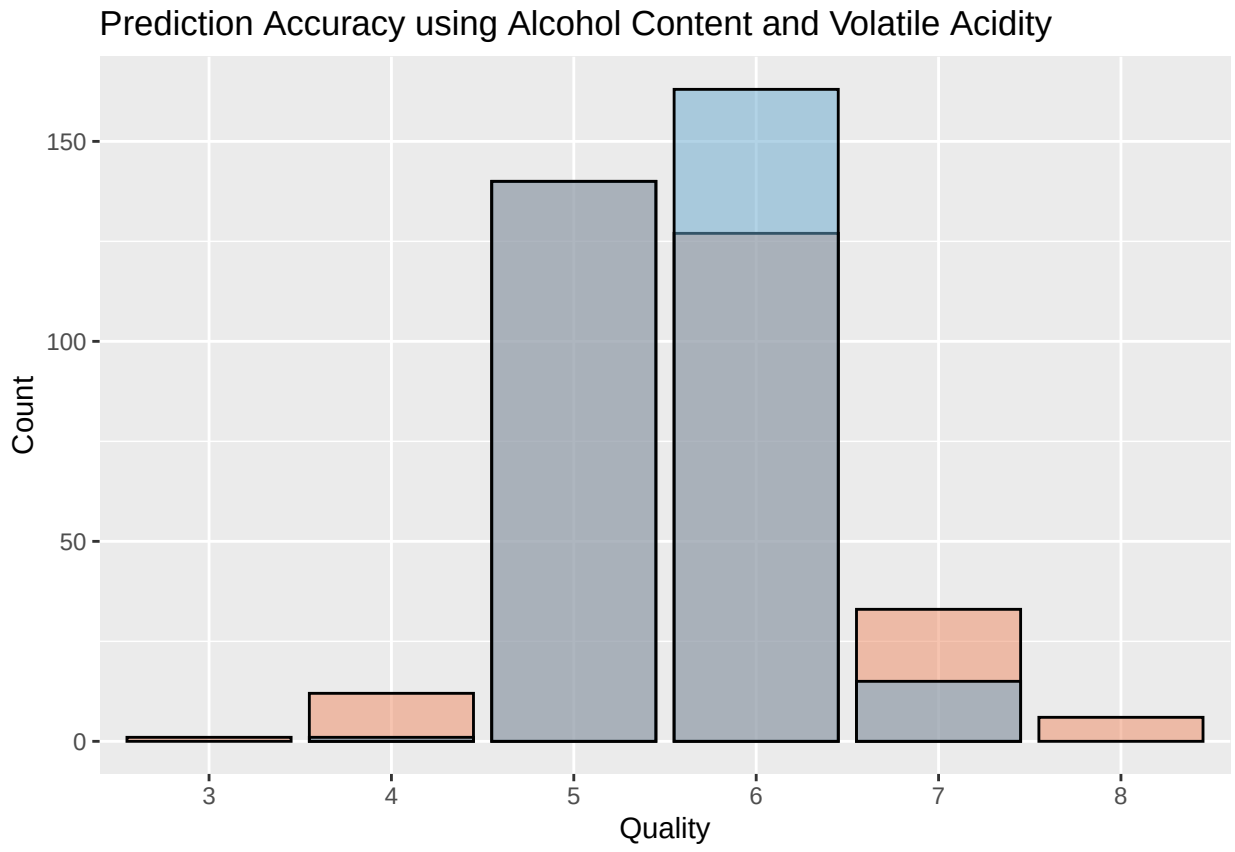
```
correct.2
```

```
## [1] 0.6018809
```

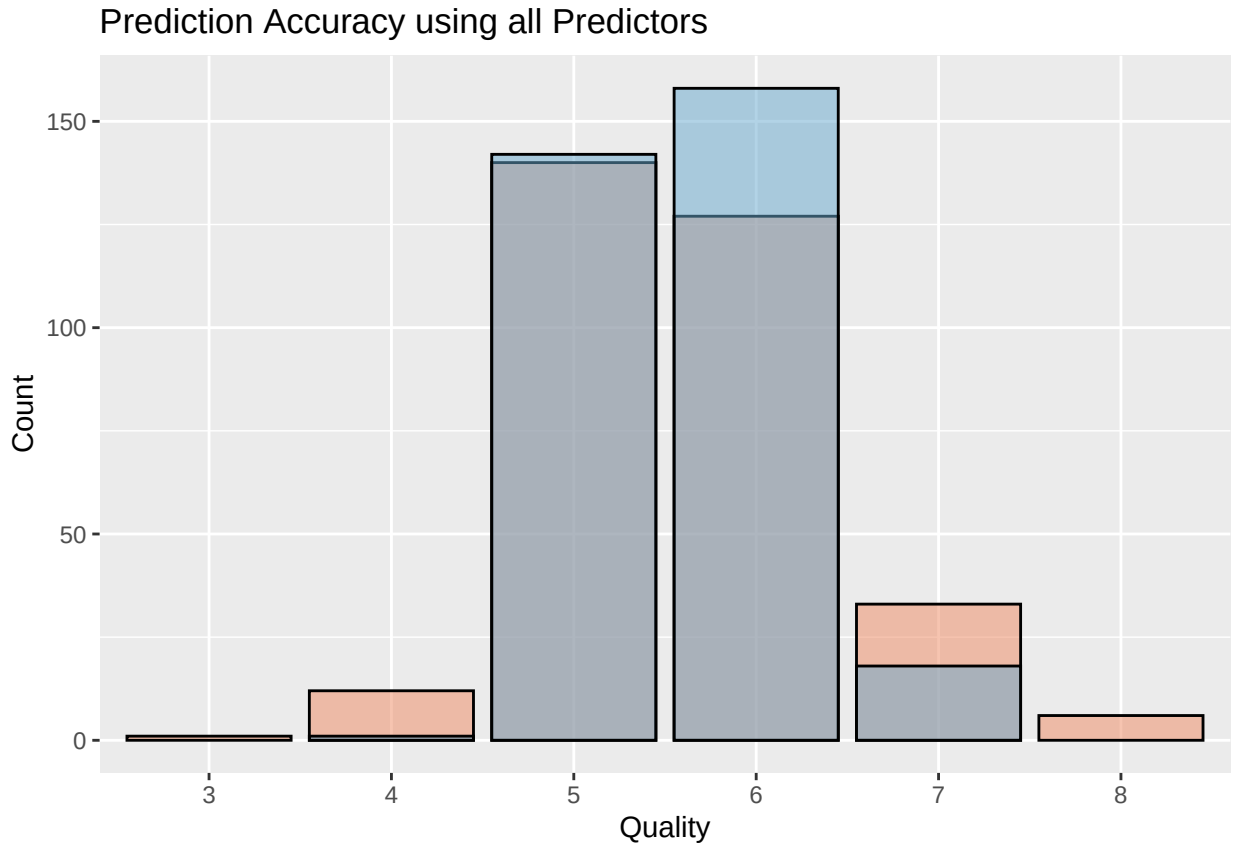
```
test$quality <- as.factor(test$quality)
cv.pred.1 <- as.factor(cv.pred.1)
cv.pred.2 <- as.factor(cv.pred.2)

ggplot(data = test) +
```

```
  geom_bar(aes(x = quality), color = "Black", fill = "#ef8a62", alpha = 0.5) +
  geom_bar(aes(x = cv.pred.1), color = "Black", fill = "#67a9cf", alpha = 0.5) +
  labs(title = "Prediction Accuracy using Alcohol Content and Volatile Acidity",
       x = "Quality",
       y = "Count")
```




```
ggplot(data = test) +
  geom_bar(aes(x = quality), color = "Black", fill = "#ef8a62", alpha = 0.5) +
  geom_bar(aes(x = cv.pred.2), color = "Black", fill = "#67a9cf", alpha = 0.5) +
  labs(title = "Prediction Accuracy using all Predictors",
       x = "Quality",
       y = "Count")
```




```
summary(model.1)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + alcohol, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59730 -0.39978 -0.06972  0.48116  2.25819
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.10454    0.20806   14.92   <2e-16 ***
## volatile.acidity -1.35147    0.10865  -12.44   <2e-16 ***
## alcohol           0.31186    0.01801   17.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6728 on 1277 degrees of freedom
## Multiple R-squared:  0.3074, Adjusted R-squared:  0.3063
## F-statistic: 283.4 on 2 and 1277 DF,  p-value: < 2.2e-16
```

```
summary(model.2)
```

```
##
## Call:
## lm(formula = quality ~ ., data = train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74040 -0.36338 -0.05723  0.46016  1.99646
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.143e+01  2.378e+01   0.901 0.367667
## fixed.acidity         3.393e-02  2.943e-02   1.153 0.249246
## volatile.acidity     -1.065e+00  1.391e-01  -7.653 3.87e-14 ***
## citric.acid          -1.355e-01  1.669e-01  -0.812 0.417145
## residual.sugar        5.685e-03  1.706e-02   0.333 0.738976
## chlorides            -1.847e+00  4.636e-01  -3.984 7.18e-05 ***
## free.sulfur.dioxide   4.515e-03  2.465e-03   1.831 0.067275 .
## total.sulfur.dioxide -2.990e-03  8.237e-04  -3.630 0.000295 ***
## density              -1.797e+01  2.430e+01  -0.739 0.459790
## pH                   -2.315e-01  2.183e-01  -1.060 0.289171
## sulphites             8.761e-01  1.272e-01   6.887 8.96e-12 ***
## alcohol               2.725e-01  2.973e-02   9.168  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6551 on 1268 degrees of freedom
## Multiple R-squared:  0.3481, Adjusted R-squared:  0.3424
## F-statistic: 61.55 on 11 and 1268 DF,  p-value: < 2.2e-16
```

Conclusion

While I was able to predict the quality of the wine 59% of the time using multiple linear regression with alcohol content and volatile acidity as the predictors, it is likely that the model is biased. The original dataset is heavily imbalanced with wines rated 5 or 6 making up the majority of the population. The model is likely overfit on this data and would almost always predict that the wine was rated 5 or 6 given any dataset. In retrospect, I should have over-sampled the data for ratings 3,4,7, and 8 in order to balanced the dataset before running the analysis. While 59% is a non-trivial accuracy rate for a problem with 6 classifications, the results of my analysis were inconclusive.

However, I did find that there likely is an association between quality and volatile acidity as well as with alcohol content. An analysis would need to be performed on a balanced dataset to confirm this however.

In addition to perform this analysis with a balanced dataset, I would also like to look at the differences between white wine and red wine and determine if they have similar predictors with similar weights for quality prediction. It may also be interesting to perform k-means clustering and see whether red wine and white wine could be separated into clusters.

Overall the most important thing I learned from this analysis is the importance of using balanced data in statistical modeling.