# Executive Summary

Based off my regression analysis I was able to conclude that while alcohol and volatile acid were key factors in predicting the rating of red wine, those aren’t enough to predict an accurate rating. However, a red wine manufacturer could infer that the having a higher amount of alcohol and a lower amount of volatile acid may help earn a higher rating of red wine. But those two factors alone won’t guarantee you a good wine rating.

# Problem and Data Description

When I retrieved the data from Kaggle, I wanted to see what the key driving factors were in determining a good red wine rating. By finding what those key factors were, a person could better understand which ingredients to focus on when creating red wine to ensure a high-quality wine.

Most of the variables in the data I collected were numeric except for the quality variable which was an integer. So, I didn’t need to convert any of the raw data for the sake of creating a linear regression.

# Describing and Visualizing Data

When I first loaded the data to RStudio, I checked if there were any null values that would mess up my regression. Thankfully there weren’t any so I then checked the correlation of the variables so I could pick my target factors. To best visualize the correlations between variables, I created a correlations plot with color. The darker colors mean higher correlation. The blue represents a positive correlation, and the red represents an negative correlation.

Chart

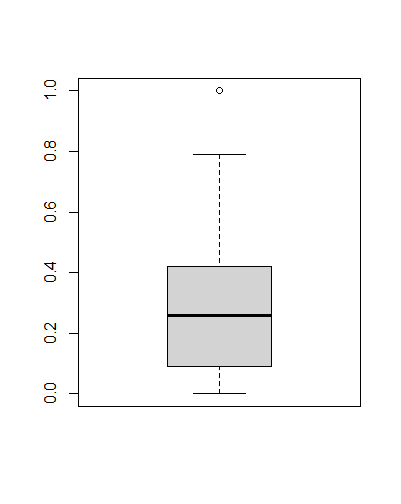
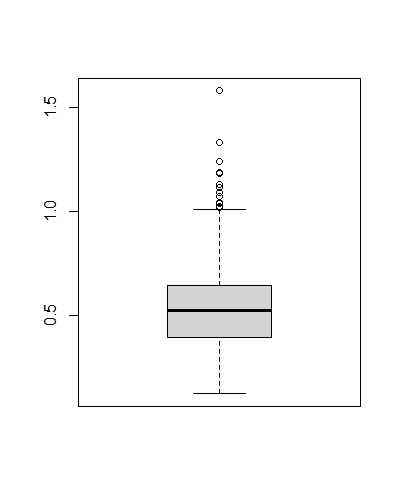
Description automatically generated with medium confidence

Based on the correlation plot, I gathered that alcohol, volatile acidity, sulphates and citric acid had the highest correlations to quality. So, I chose to use those four factors to build my model.

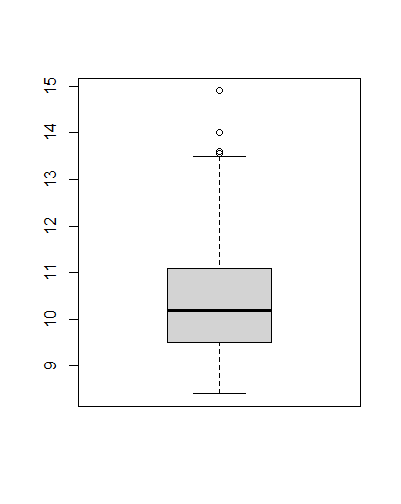
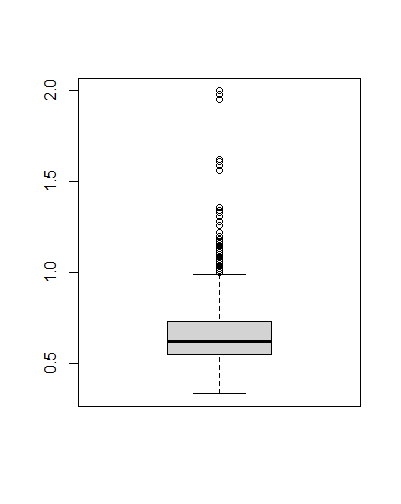
I also observed that volatile acidity and citric acid have a correlation so it would be wise to see how that relationship affects quality.

Next, I made box plots of each of my chosen variables to analyze the distribution of the variables.

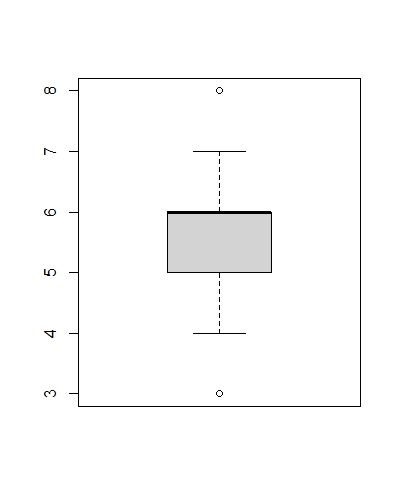
Volatile Acidity Boxplot 2 Citric Acid Boxplot



3 Sulphate Boxplot 4 Alcohol Boxplot



5 Quality Boxplot



# Analysis

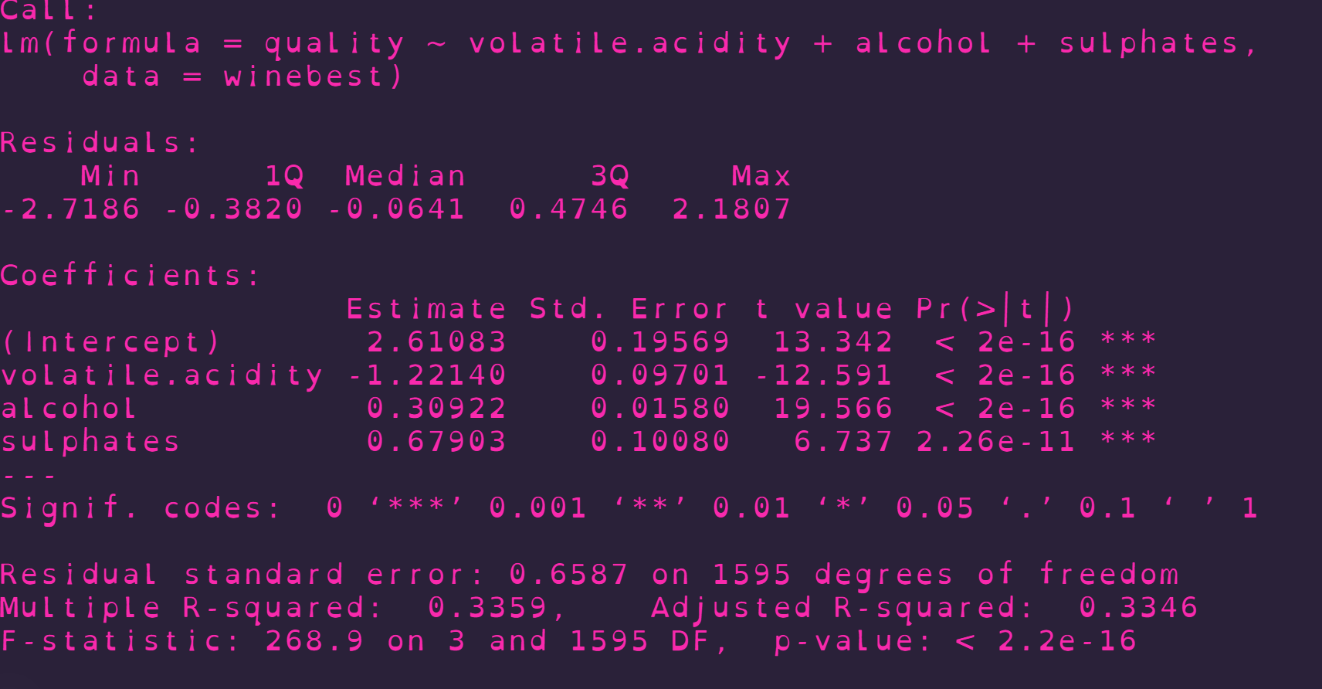
To analyze and predict the data I built 3 multiple linear regressions. The first model included alcohol, citric acid and sulphates. The second model included volatile acidity, alcohol and sulphates. Finally, the third model included all of the variables in the regression.

Text

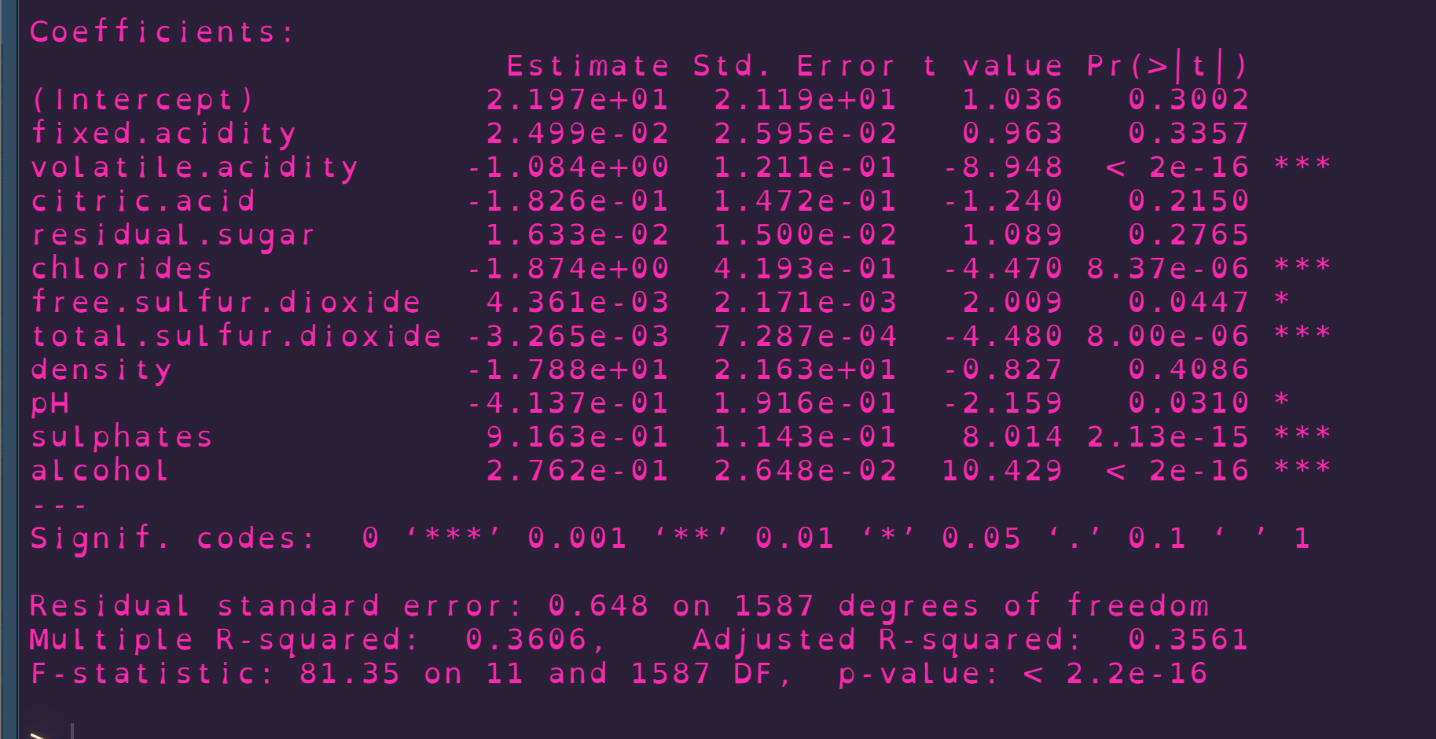
Description automatically generated

According to my model citric acid will increase the wine quality by about 0.51 when one unit is added. Sulphates will increase quality by 0.81 and alcohol will increase it by 0.33. So according to my model, sulphates are a great way to increase the quality. However, the model isn’t very significant (with RSE of about 0.68 and an F-statistic of 210.5)

The next model is model two:



In this model, the benefit of increasing the amount of alcohol and sulphates each decrease when compared to the first model, but we have new insight into how volatile acidity will affect quality. Volatile acidity will decrease quality by about 1.22 with each unit added so it perhaps isn’t beneficial to add a large amount of it when you’re trying to create a quality wine. However, this model is only slightly better than the first (F statistic is 268.9 and RSE of 0.6587)

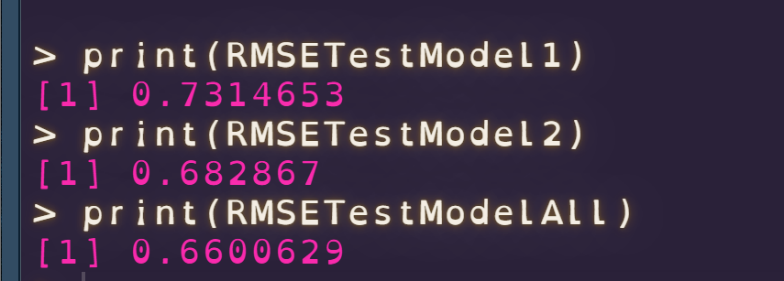
The final model uses all of the variables to predict quality:

As you can see this model performs a little better than the first two but the F-Statistic of 81.35 is concerning. However, we can see that most variables (with the exception of alcohol and volatile acidity) are pretty insignificant when it comes to their affect on quality.

After creating the models, I then created test and training sets to see how well they performed.

My evaluation on how they performed was based on my calculation of the RMSE of each model. Unfortunately, the RMSE only strengthened the conclusion that my models did not perform well in accurately predicting quality. The RMSE for my three models were 0.7314653, 0.682867, and 0.6600629 respectively. So, we can infer that the third model does the best at predicting wine quality, but it wouldn’t be very useful to rely on it to guarantee a good wine rating.

RMSE calculation for each model using the test dataset



# Summary

Based on my analysis, one can infer that you can’t really predict the quality of wine with great certainty. This is most likely because wine quality is more subjective rather than objective. However, you can infer the amounts of the various ingredients may affect the final rating. For example, based on my analysis, more alcohol content in red wine may result in a higher rating so it would not hurt for a wine company to increase alcohol content in their wine. But a higher alcohol content will not guarantee a good rating. It will only help push you towards that good rating. We can also infer that its better to focus on citric acid and not volatile acidity when making wine. This is since volatile acidity has an inverse relationship with quality (based on the dataset). So, a wine company can prioritize citric acid when creating red wine to help with its quality rating.

# Appendix

library(tidyverse)

library(ggplot2)

library(corrplot)

library(caTools)

#reading in data

winebest <- read.table("C:/Users/devin/OneDrive - Texas State University/Documents/Working D for R script/winequality-red.csv", header = TRUE, sep = ",")

#checking for null values

#colSums(is.na(wine))

Num.cols <- sapply(winebest, is.numeric)

Cor.data <- cor(winebest[, Num.cols])

str(winebest)

corrplot(Cor.data, method = 'color')

#cor(wine)

#alcohol, sulphates, citric acid, volatile acidity

# condensing data to four hightest correlated variables

#winebest <- wine[,c(2,3,10:12)]

head(winebest)

#rm(wine)

#summary of each variable

summary(winebest)

attach(winebest)

pairs(winebest)

###ANALYSING Distribution

boxplot(winebest$volatile.acidity)

#High 1 Low 0.1

boxplot(winebest$citric.acid)

#high 0.8 Low 0.0

boxplot(winebest$sulphates)

#high 0.9 Low 0.2

boxplot(winebest$alcohol)

#High 13.5 Low 8.5

boxplot(winebest$quality)

#high 7 low 4

##### Multiple Linear Model with interaction between citric acid and sulphates, and alcohol

model1 <- lm(quality ~ citric.acid +sulphates + alcohol, data = winebest)

summary(model1)

##### Regression with volatile acid, alcohol, and interaction between sulphates and volatile acid

model2 <- lm(quality ~ volatile.acidity + alcohol+ sulphates, data = winebest)

summary(model2)

### regression using all variables

ALLWineregression = lm(formula = quality~.,data = winebest)

summary(ALLWineregression)

### testing test model

set.seed(33)

wine\_split<- sample(2,nrow(winebest), replace = TRUE, prob = c(.7,.3))

wine\_train<- winebest[wine\_split == 1,]

wine\_test<- winebest[wine\_split == 2,]

### prediction using test data

ALLWineReg\_predict <- predict(ALLWineregression, newdata = wine\_test)

model1\_predict <- predict(model1, newdata = wine\_test)

model2\_predict <- predict(model2, newdata = wine\_test)

tab\_model2predict <- data.frame(True = wine\_test$quality, Quality = model2\_predict)

head(tab\_model2predict,20)

tab\_model1predict <- data.frame(True = wine\_test$quality, Quality = model1\_predict)

head(tab\_model1predict,20)

tab\_AllWineReg\_predict <- data.frame(True = wine\_test$quality, Quality = ALLWineReg\_predict)

head(tab\_AllWineReg\_predict,20)

### prediction using training data

ALLWineReg\_predicttrain <- predict(ALLWineregression, newdata = wine\_train)

model1\_predicttrain <- predict(model1, newdata = wine\_train)

model2\_predicttrain <- predict(model2, newdata = wine\_train)

summary(model1)

RMSETestModel1<- sqrt(mean((wine\_test$quality - model1\_predict)^2))

print(RMSETestModel1)

RMSETestModel2<- sqrt(mean((wine\_test$quality - model2\_predict)^2))

print(RMSETestModel2)

RMSETestModelAll<- sqrt(mean((wine\_test$quality - ALLWineReg\_predict)^2))

print(RMSETestModelAll)

RMSETrainModel1<- sqrt(mean((wine\_train$quality - model1\_predicttrain)^2))

print(RMSETrainModel1)

RMSETrainModel2<- sqrt(mean((wine\_train$quality - model2\_predicttrain)^2))

print(RMSETrainModel2)

RMSETrainModelAll<- sqrt(mean((wine\_train$quality - ALLWineReg\_predicttrain)^2))

print(RMSETrainModelAll)