Final Report

**Introduction:**

The wine market is an extremely expensive and popular market that is globally worth over 354.7 billion US dollars as of 2018. By 2023, the worth of the market is expected to grow to 429 billion US dollars. The purpose of our research project is to determine a model that can predict the quality of wine produced based on the physicochemical properties of wine that are measured during the wine making process. The importance of developing a model that is capable of predicting the quality of a batch of wine allows for the wine industry to determine the value of the wine prior to the completion of the wine. Furthermore, to be able to have a model based on physicochemical properties allows the vineyard to develop a means of making wine fit for the average person who enjoys and has the means of drinking a low cost bottle of wine or the vineyard can appease the sophisticated pallet of those who like to indulge on more quality and expensive wines. By understanding how the quality of wine is affected by different physicochemical properties, the wine market can more accurately control their wine making.

**Methods:**

The data was provided by the University of California, Irvine Machine Learning Repository. The data was created from Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis from The University of Minho and their commission from the Vinho Verde Vineyard. Vinho Verde produces a unique product of Minho (NorthWest) region of Portugal. Their product is renowned for their unique medium alcohol level and refreshing taste in the summer.

The dataset that we are using analyzes both red and white wine from Vinho Verde. For our research project we are using the quality of the wine as the response variable. The wine quality is measured between 0 and 10 which is based on sensory data from expert wine tasters. The predictors for our model are physicochemical properties of wine that were gathered through lab experiments. The predictors are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. For this dataset it was decided that in order to find the models that best fit the data, model selection using Criterion Based Selections would be our method of approach. Specifically the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) are used to find models for the data. The AIC is superior to BIC in instances where the main objective is to develop a predictive model. When calculating the AIC the first term is dependent on the RSS (residual sums of squares). So as the model better fits the data, the RSS is minimized. The second term of AIC is dependent on the number of predictors in the model. By reducing the number of predictors in the model, the AIC also improves. Therefore, AIC is efficient by “naturally providing a balance between fit and simplicity in model selection” (Linear Models with R, 154).

On the other hand, BIC “penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC” (Linear Models with R, 154). This means that BIC will penalize additional parameters more than AIC, hence the preference to smaller models. Using BIC might give us a better model than AIC because it will be harsher on free parameters and this is our goal, to create a better reduced model. However, our model is quite large which can affect our BIC result negatively. Overall, we are going to perform both AIC and BIC and compare the results.

The first model chosen for the red and white wine is the full model which contains the response (quality) and all of the predictors. The full model is used to develop a new reduced model that is a reference for the models developed using the AIC and BIC. The reference model is developed by checking the p-values of the predictors in the full model. If the p-value of the predictor is greater than 0.05, the predictor is said to be insignificant and is removed from the model. The Reference model only uses predictors that are significant and therefore have a p-value less than 0.05.

Utilizing AIC, we can plot the Cp statistic versus the number of predictors. This graph will show which number of predictors are on the line or below the line of Cp = p. The Cp Statistic or Mallow’s Cp is an estimate of the average mean square error of prediction (Faraway 156). By reducing the RSS of the Cp statistic, for a model with p predictors, the average mean square error of prediction decreases. This ultimately results in Cp approximately equal to the number of predictors. Therefore, if the Cp statistic is close to p or less than p, the model is a good fit for the data. However, if the Cp is greater than p, then it can be concluded that the model is not a good fit. Meaning, if our model has a number of predictors close to or on the Cp = p line, then our model is a good fit. Otherwise our model is not a good fit. Using the step() function in R and inputting our full model we can get the AIC forward model. By also including distraction = “backward” , you can get the AIC backward. Similarly to the AIC, BIC utilizes the step function. However, for the BIC forward the k needs to be determined which is the log(n) where n is the number of rows in the data set. By adding the ‘k’ into the AIC forward and backward you will similarly get BIC forward and backward.

After passing the full model through the model selection, it is important to check the distributions of the predictors to determine if transformations are required. To check the distributions of the predictors a strip chart is used. By looking at the stripchart, it can be clearly seen which predictors may not be normal and may require a transformation.

Lastly, In order to quantitatively determine how well the models developed fit the data, the residuals versus fitted values plot will be analyzed.

**Results:**

After producing the linear model for the predictors and quality for red wine, it is obvious that some predictors are not significant. The predictors that aren’t significant are fixed acidity, citric acid, residual sugar, and density. We eliminated these predictors and created a reduced linear model as a reference. AIC forward and AIC backward were both conducted, which both produced the same model we had created by looking at the summary. The AIC models were further proven by plotting the Cp statistic versus the number of predictors. Cp = p line shows that there is a negligible difference between 7 and 8 predictors in the best fit model (seen in Figure 1.1). We then performed BIC forward and backward which both produced the same model but had one less predictor than our reduced model. The additional predictor that was removed from the model was free sulfur dioxide. This was expected because of how harsh BIC is on the number of predictors. Although free sulfur dioxide was significant in the reference model, the value was not close enough to 0 in order to be included in the model.

After developing the models, it is important to check the predictors and determine if they individually follow the normal distribution. The distributions were checked using a stripchart. The stripchart, see Figure 1.2, shows that volatile acidity, chlorides, free sulfur dioxide, pH, and sulphates can undergo a transformation for the model to better fit the data.

**Full Model for Red Wine:**

quality ~ fixedAcidity + volatileAcidity + citricAcid + residualSugar + chlorides + freeSulfurDioxide + totalSulfurDioxide + densityRed + pH + sulphates + alcohol

**Reference Model for Red Wine:**

quality ~ volatileAcidity + chlorides + freeSulfurDioxide + totalSulfurDioxide + pH + sulphates + alcohol

**AIC Forward Model for Red Wine:**

quality ~ volatileAcidity + chlorides + freeSulfurDioxide + totalSulfurDioxide + pH + sulphates + alcohol

**AIC Backward Model for Red Wine:**

quality ~ volatileAcidity + chlorides + freeSulfurDioxide + totalSulfurDioxide + pH + sulphates + alcohol

**BIC Forward Model for Red Wine:**

quality ~ volatileAcidity + chlorides + totalSulfurDioxide + pH + sulphates + alcohol

**BIC Backward Model for Red Wine:**

quality ~ volatileAcidity + chlorides + totalSulfurDioxide + pH + sulphates + alcohol

We did the same statistical analysis to white wine as we did to red by creating our own reduced linear model and comparing it to AIC and BIC models. When producing the linear model for the predictors and quality for white wine, it was clear that there were some predictors that weren’t significant. We eliminated citric acid, chlorides, and total sulfur dioxide from the model and created a reduced linear model. Next we conducted an AIC forward and AIC backward which resulted in the same reduced linear model we created. The AIC models were further proven by plotting the Cp statistic versus the Number of predictors. Cp = p line shows that the best model is one with 8 predictors (shown in Figure 2.1). Then we created a BIC backward and forward model which also resulted in the same reduced linear model we created. Overall, the quality of white wine is dependent on fixed acidity, free sulfur dioxide, sulphates, ph, alcohol, density, residual sugar, and volatile acidity.

After the models were developed, it was important to also determine if the predictors were normally distributed. A stripchart visualizes the data points for the predictors and visualizes the distributions (Figure 2.2). It can be seen that the predictor volatile acidity and free sulfur dioxide can undergo transformation for the model to better fit the data.

**Full Model for White Wine:**

quality ~ fixedAcidity + volatileAcidity + citricAcid + residualSugar + chlorides + freeSulfurDioxide + totalSulfurDioxide + densityRed + pH + sulphates + alcohol

**Reference Model for White Wine:**

quality ~ fixedAcidity + volatileAcidity + residualSugar + freeSulfurDioxide + densityWhite + pH + sulphates + alcohol

**AIC Forward Model for White Wine:**

quality ~ fixedAcidity + volatileAcidity + residualSugar + freeSulfurDioxide + densityWhite + pH + sulphates + alcohol

**AIC Backward Model for White Wine:**

quality ~ fixedAcidity + volatileAcidity + residualSugar + freeSulfurDioxide + densityWhite + pH + sulphates + alcohol

**BIC Forward Model for White Wine:**

quality ~ fixedAcidity + volatileAcidity + residualSugar + freeSulfurDioxide + densityWhite + pH + sulphates + alcohol

**BIC Backward Model for White Wine:**

quality ~ fixedAcidity + volatileAcidity + residualSugar + freeSulfurDioxide + densityWhite + pH + sulphates + alcohol

Lastly, when comparing the residuals versus fitted values plot of the full model and reduced model for red wine, it can be observed that there is not a significant change in the models. The plots are seen to be fairly straight and linear lines that are close to 0 (See Figure 3.1 and Figure 3.2). Similarly to red wine, white wine’s full model and reduced model do not have a noticeable change in the residuals versus fitted values plots (See Figures 4.1 and 4.2). The Cp statistic versus number of predictors plot demonstrated that the reduced model and full model would both be acceptable models for wine quality.

**Discussion and Analysis:**

The purpose of this study was to determine a reduced model that could predict the quality of wine based on physicochemical properties of wine. Through model selection and analysis, it was determined that the full model and reduced model do not have a significant difference between them. However, for determining the quality of wine the reduced model is preferred. This is due to the fact that by reducing the amount of physicochemical properties that need to be tested, the efficiency for predicting the quality of wine increases.

We determined that red and white wine shared many common characteristics and many varying differences. The similarities of the red and white wine were that they shared five key properties. These properties were volatile acidity, free sulphur dioxide, pH, sulphates, and alcohol. The pH, sulphates, and alcohol are very crucial in wine making. Sulphates prevent the wine from oxidizing, prevents yeast and bacteria from growing, and keeps the wine fresh. The pH of wine is very important based on how acidic or basic the wine is meant to be. And lastly, the alcohol. Alcohol is the most important part of an alcoholic beverage. The most interesting results came from the difference in predictors. White wine had the predictor residual sugar in its reduced model, which makes sense sence white wines tend to be sweeter than red wine.

The questions that remain after this project and would be future topics to explore are “Why are chlorides significant in red wine and not white?”, “Would a model made of the similar predictors be more valuable than individual models for red and white wine?”, “What are the exact distribution transformations needed to increase the effectiveness of the models?”, and “Would cleansing the data increase the effectiveness of the models?”

In conclusion, the model selection methods chosen proved to be effective in determining a better and more reduced model for predicting the quality of wine. It was shown that Akaike Information Criterion (AIC) was superior to the Bayes Information Criterion (BIC) due to AIC being better at handling large data sets. The Cp statistic was shown to be a quick and easy test to see which number of predictors would be best for the model. However, there were many questions that were left unanswered and new skill sets would be required to answer these questions.

Figure 1.1 Cp Statistic as a function of the Number of Parameters where Cp = p for Red Wine.

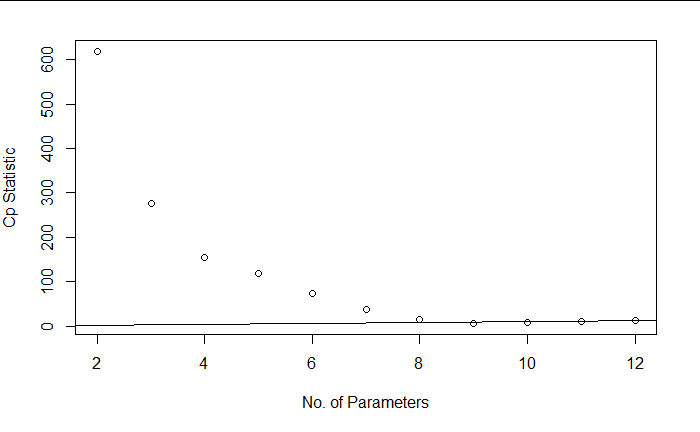
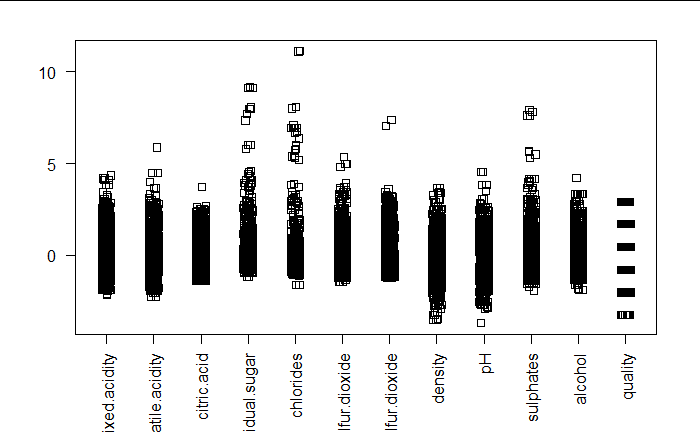
Figure 1.2 Stripchart visualizing the distributions of the predictors for Red Wine.

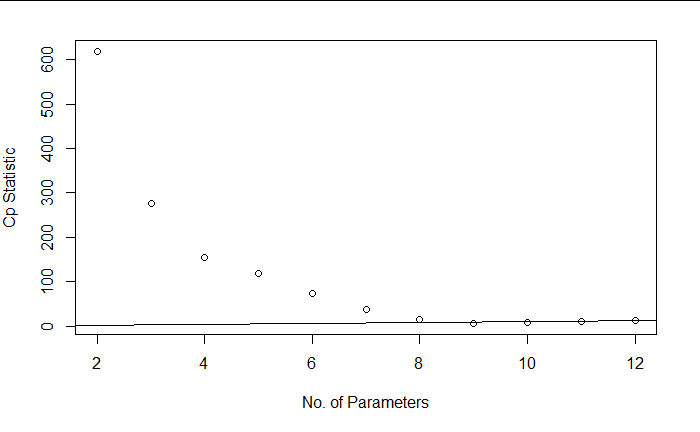
Figure 2.1 Cp Statistic as a function of the Number of Parameters where Cp = p for White Wine.

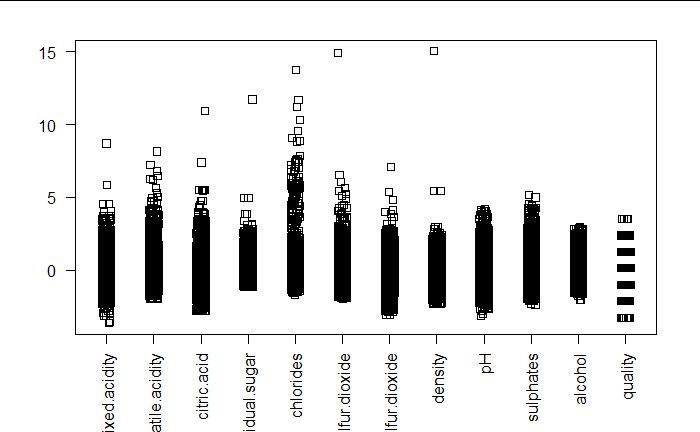
Figure 2.2 Stripchart visualizing the distributions of the predictors for White Wine.

Figure 3.1 Full Model (Red Wine) - Residuals versus Fitted Values.

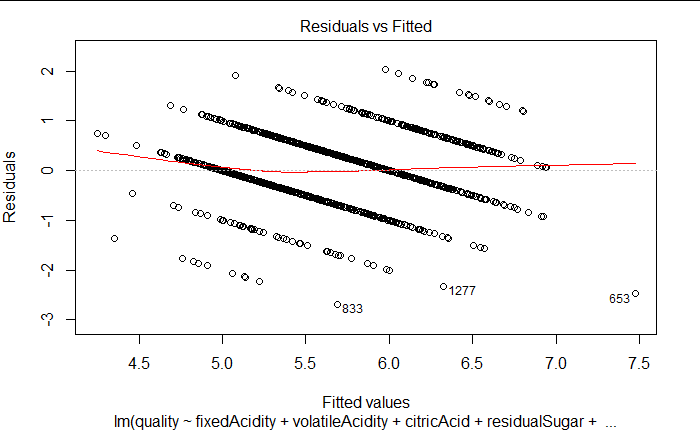


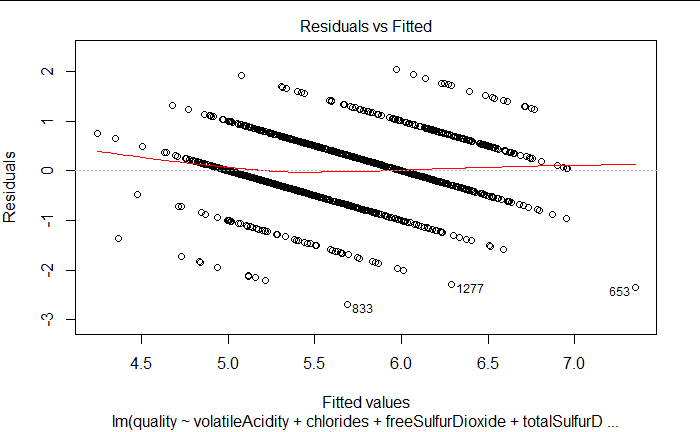
Figure 3.2 Reduced Model (Red Wine) - Residuals versus Fitted Values.

Figure 4.1 Full Model (White Wine) - Residuals versus Fitted Values.

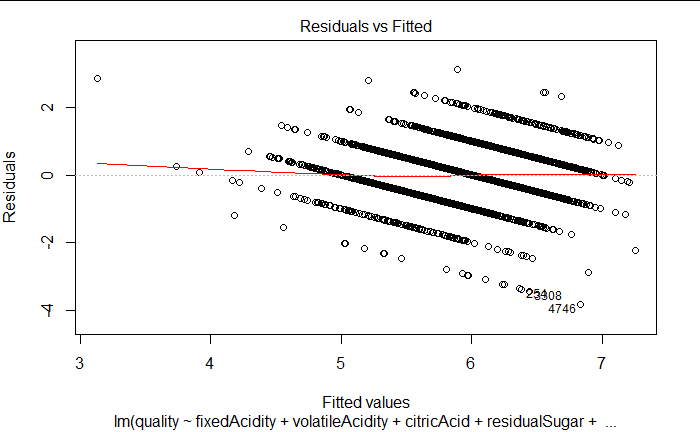
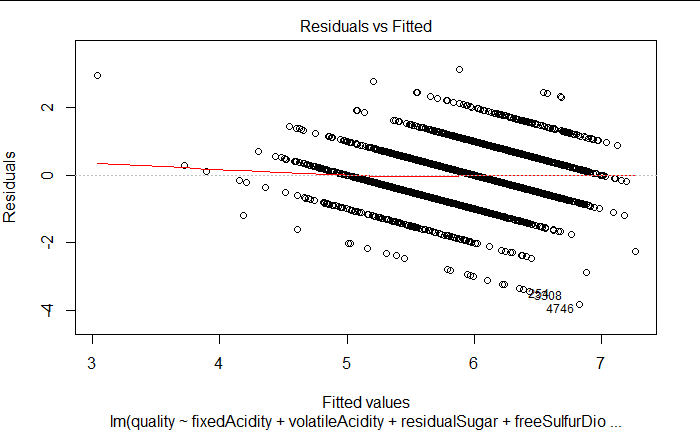


Figure 4.2 Reduced Model (White Wine) - Residuals versus Fitted Values.

**Supplementary Materials Red Wine (R code):**

wine\_red <- read.csv("~/Data 467/winequality-red.csv",sep = ’;’)

fixedAcidity = wine\_red$fixed.acidity

volatileAcidity = wine\_red$volatile.acidity

citricAcid = wine\_red$citric.acid

residualSugar = wine\_red$residual.sugar

chlorides = wine\_red$chlorides

freeSulfurDioxide = wine\_red$free.sulfur.dioxide

totalSulfurDioxide = wine\_red$total.sulfur.dioxide

densityRed = wine\_red$density

pH = wine\_red$pH

sulphates = wine\_red$sulphates

alcohol = wine\_red$alcohol

quality = wine\_red$quality

redWine.lm <-lm(quality~fixedAcidity+volatileAcidity+citricAcid+residualSugar+

chlorides+freeSulfurDioxide+totalSulfurDioxide+densityRed+pH+sulphates+alcohol)

summary(redWine.lm)

plot(redWine.lm)

redWine2.lm <- lm(quality~volatileAcidity+chlorides+freeSulfurDioxide+

totalSulfurDioxide+pH+sulphates+alcohol)

summary(redWine2.lm)

require(leaps)

b <- regsubsets(wine\_red$quality ~ ., wine\_red, nvmax = 11)

rs <- summary(b)

rs$which

AIC <- 1591\*log(rs$rss/1591) + (2:12) \* 2

plot(AIC ~ I(1:11), ylab = "AIC", xlab = "Number of Predictors")

plot(2:12,rs$adjr2,xlab=’No. of Paramters’, ylab = ’Adjusted R-square’)

which.max(rs$adjr2)

plot(2:12, rs$cp, xlab="No. of Parameters", ylab = "Cp Statistic")

abline(0,1)

step(redWine.lm)

step(redWine.lm, distraction = "backward")

stripchart(data.frame(scale(wine\_red)),method = "jitter", las=2,vertical= TRUE)

n = length(resid(redWine.lm))

BIC\_backward = step(redWine.lm,distraction = "backward", k = log(n))

n = length(resid(redWine.lm))

BIC\_backward = step(redWine.lm, k = log(n))

plot(lm(formula = quality ~ volatileAcidity + chlorides + freeSulfurDioxide +

totalSulfurDioxide + pH + sulphates + alcohol))

**Supplementary Materials White Wine (R code):**

wine\_white <- read.csv("~/Data 467/winequality-white.csv",sep = ';')

fixedAcidity = wine\_white$fixed.acidity

volatileAcidity = wine\_white$volatile.acidity

citricAcid = wine\_white$citric.acid

residualSugar = wine\_white$residual.sugar

chlorides = wine\_white$chlorides

freeSulfurDioxide = wine\_white$free.sulfur.dioxide

totalSulfurDioxide = wine\_white$total.sulfur.dioxide

densityWhite = wine\_white$density

pH = wine\_white$pH

sulphates = wine\_white$sulphates

alcohol = wine\_white$alcohol

quality = wine\_white$quality

whiteWine.lm <-lm(quality~fixedAcidity+volatileAcidity+citricAcid+residualSugar+chlorides+freeSulfurDioxide+totalSulfurDioxide+densityWhite+pH+sulphates+alcohol)

summary(whiteWine.lm)

plot(whiteWine.lm)

whiteWine2.lm <- lm(quality~fixedAcidity+volatileAcidity+residualSugar+freeSulfurDioxide+densityWhite+pH+sulphates+alcohol)

summary(whiteWine2.lm)

require(leaps)

b <- regsubsets(wine\_white$quality ~ ., wine\_white, nvmax = 11)

rs <- summary(b)

rs$which

AIC <- 1591\*log(rs$rss/1591) + (2:12) \* 2

plot(AIC ~ I(1:11), ylab = "AIC", xlab = "Number of Predictors")

plot(2:12,rs$adjr2,xlab='No. of Paramters', ylab = 'Adjusted R-square')

which.max(rs$adjr2)

plot(2:12, rs$cp, xlab="No. of Parameters", ylab = "Cp Statistic")

abline(0,1)

step(whiteWine.lm)

step(whiteWine.lm, distraction = "backward")

stripchart(data.frame(scale(wine\_white)),method = "jitter", las=2,vertical= TRUE)

n = length(resid(whiteWine.lm))

BIC\_backward = step(whiteWine.lm,distraction = "backward", k = log(n))

n = length(resid(whiteWine.lm))

BIC\_backward = step(whiteWine.lm, k = log(n))

plot(lm(formula = quality ~ fixedAcidity + volatileAcidity + residualSugar +

freeSulfurDioxide + densityWhite + pH + sulphates + alcohol))

**Supplementary Material (data sets):**

http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

**References:**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009. (archive.ics.uci.edu/ml/datasets/Wine+Quality)

Faraway, Julian James. *Linear Models With R, Second Edition*. Taylor & Francis., 2014.

Faraway, Julian James. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press, 2016.

*Wine Market Size, Share, Trends, Analysis (2020-2025)*, www.mordorintelligence.com/industry-reports/wine-market.