Sta 141a Final Group Project
UC Davis
Professor Debashis Paul
Red Wine Quality

Ryan Kuan
Matthew Lee
Benjamin Mok
Sooyeon Park

Part 1: Introduction

The dataset that will be used for this project is the Wine Quality Data Set from the University of California, Irvine - Machine Learning Repository. Included within the dataset are 1599 red wine samples of Portuguese Vinho Verde, a type of wine originating from the Northern region of the country. For each wine sample, 10 chemical variables were recorded in order to determine the quality. We were interested in finding out how each of these different factors affected the quality of red wines by investigating variables in the dataset through visual simulations. Our main goal was to find variables which significantly influenced the taste or quality of red wines.

Part 2: Methodology

Initially, we thought to use linear models to begin our analysis of the the independent variables. However, we concluded that this would not be a good idea since the response variable, the quality of the red wine, is actually categorical. Therefore, to figure out where to start our analysis, we searched information online from both professional viticulture experts and experienced home winemakers. For some plots, we made subsets of the dataset depending on variables, and each subset had a different sample size, so we focused on proportions of the dataset instead of actual sample sizes.

Using the “ggplot2” package, we are given the ability to compare different subsets of the data. With repetitive but reasoned data exploration, we slowly discovered interesting ways one variable would relate to another. An advantage of comparison with “ggplot” is the ability to facet in a 3rd or a 4th variable, allowing us more insight on further variables whilst retaining what we’ve already learned. With these separate methodologies we can do our best to discover the answer all students crave: What makes red-wine taste better?

Part 3: Graphs and Analysis

In order to better understand the factors considered into the quality scores of wine, our group spent a significant amount of time learning about given independent variables in the dataset. According to information provided by British Student Experiment, it stated that the amount of sugar, alcohol content, and acid content were the three most important factors in determining the quality of wine. From these findings, our group chose to analyze these variables to further investigate the relationship they had with the taste.

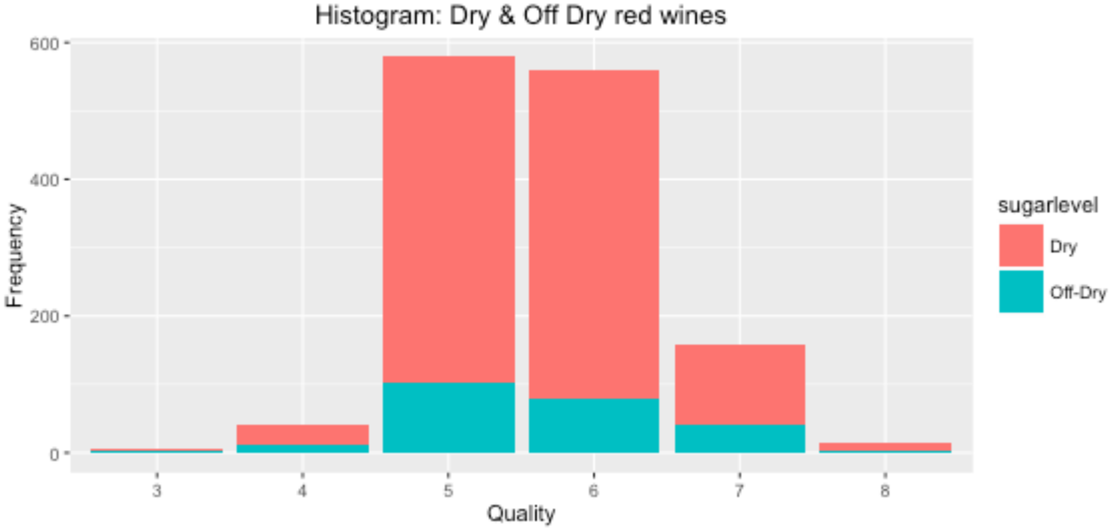
Residual Sugar vs Quality Analysis

Within the dataset, sweetness is measured as the amount of residual sugar. This variable was continuous, ranging from 0.9-15.5 g/L, with median of 2.20 g/L and mean of 2.539 g/L.

According to information from the popular online wine blog, Wine Curmudgeon, Dry Red wines contain 2-3 g/L of residual sugar, and Off-Dry wines contain 10-30 g/L Given the range from the data set, it made sense to split into dry and off-dry categories to match the sweetness spectrum. We set the parameters for dry wines to have less than or equal to 3.0 g/L of sugar and the parameter for Off-Dry to be all wines with greater than 3.0 g/L of sugar.

One of the main reasons for this was that there were only 11 wine samples that had residual sugar 10 g/L or greater; we concluded that this sample size for the Off-Dry level was too small to get meaningful analysis, thus we settled for Off-Dry being greater than or equal to 3.0 g/L. After subsetting the data, we observed there to be 1,359 Dry red wines and 240 Off-Dry red wines. Our group hypothesized that the sweeter the wine, the better quality score it would receive.

Figure 1.



The figure above contains a histogram of the Quality scores that were cut by sugar levels. It is apparent that for both Dry and Off-Dry that a score of 5 and 6 were the most popular. However, when we compare based upon sweetness, it appears that the sweeter “Off-Dry” wine does not have a trend of higher quality. Both levels of sweetness followed a similar frequency distribution with higher quality scores being uncommon. While this did not support our hypothesis, it is important to note that the dataset is specifically for red wines. Had the data comprised a compilation of both red and white wine, there may have been a different result. Similarly, dry and off-dry wines are not known for their sweetness, but the acidic flavor they produce.

Alcohol vs Quality Analysis

Alcohol content of red wine is measured by the amount of alcohol in a given volume, and it is known that the lower alcohol level the wine has, the sweeter the wine becomes. Since we did not find a clear relationship between sweetness and quality, we set out to find a possible association between sweetness and alcohol content of the wine. Subsequently, we moved on to figure out if there is any relationship between alcohol concentration and quality.

According to online reports (cited previously), the alcohol level is categorized based on ABV. Low-alcohol wines are wines with less than 10% ABV. Medium-low-alcohol wines have 10~11.5% ABV. Medium-alcohol wines have 11.5%~13.5% ABV. Medium-high-alcohol wines have 13.5~15% ABV. And lastly, high-alcohol wines above 15% ABV. Based on the alcohol levels we searched, we made four subsets (and realized there wasn't any high-alcohol wines in the sample). Then, since all each sample size of the subset varied greatly, we displayed a density plot depending on the subsetted alcohol levels to look at the proportions.

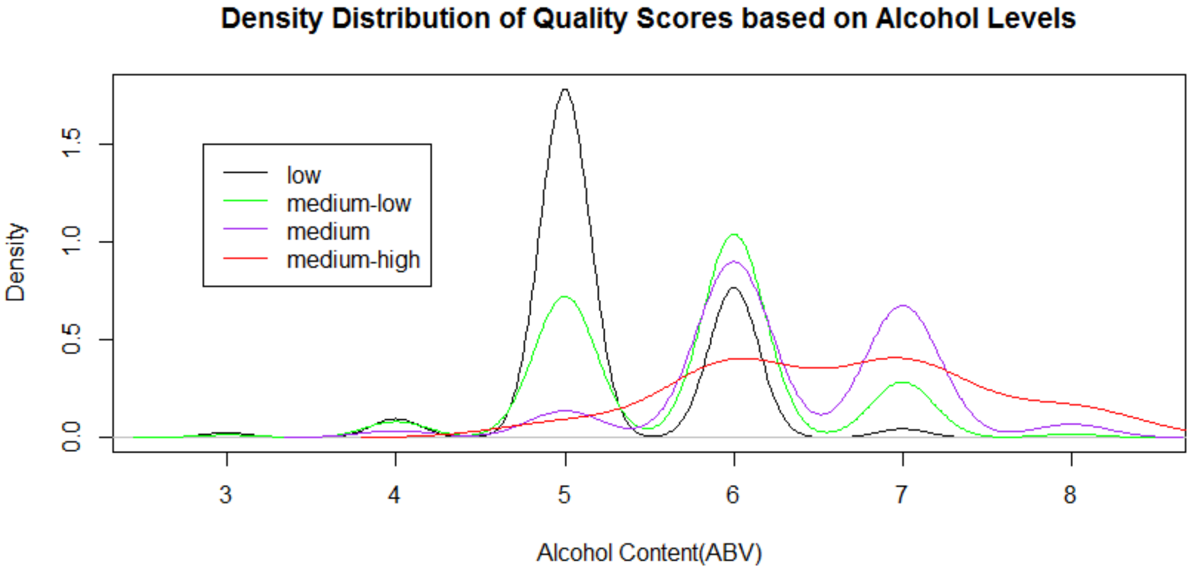
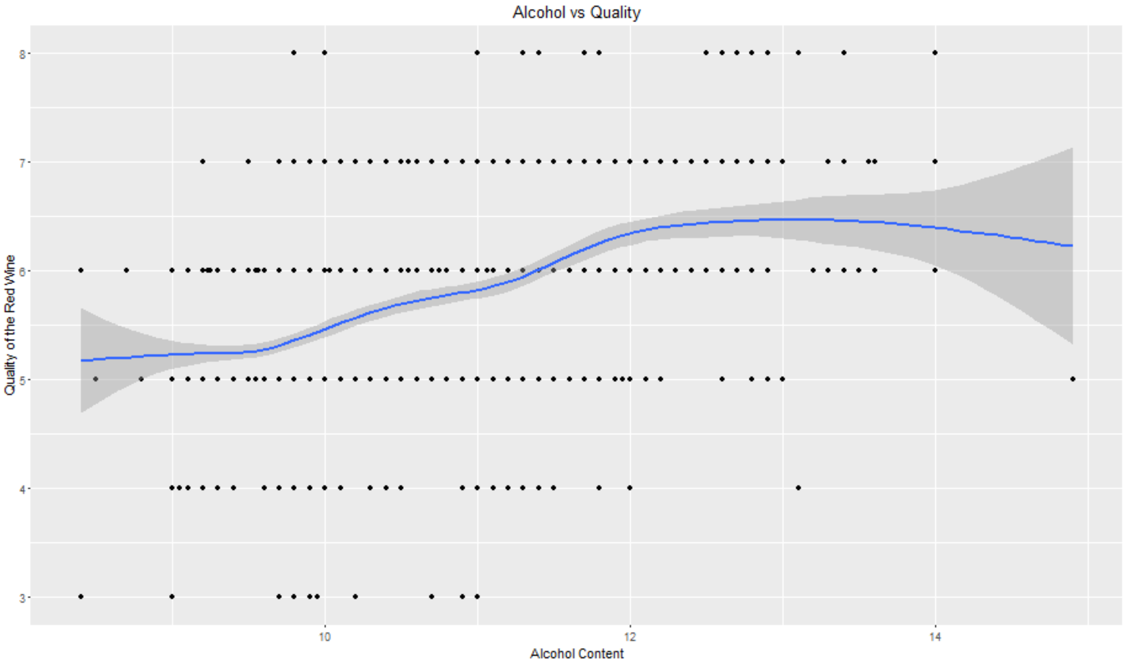


Figure 2.
According to the plot above, red wines with lower alcohol levels compared to those with higher alcohol levels have higher proportions in higher quality scores (ex. Wines with medium-high levels of alcohol have the highest proportion of score 8 than wines with medium or low alcohol levels). This indicates that there is a positive correlation between quality scores and alcohol levels of red wines. However, there is a limitation to this finding because, as we mentioned before, the dataset did not contain any high-alcohol wines.

Figure 3.



Using ggplot, we were able to show a graphical representation on how alcohol content may affect the quality of taste in red wines. As shown in the graph above, what we see is more of a curvilinear relationship. The lower alcohol levels lack in taste, but as the alcohol content steadily increases, so does the quality of taste. The curvilinear aspect comes in at around 12-13% alcohol, which should be the ideal amount of alcohol in red wine.

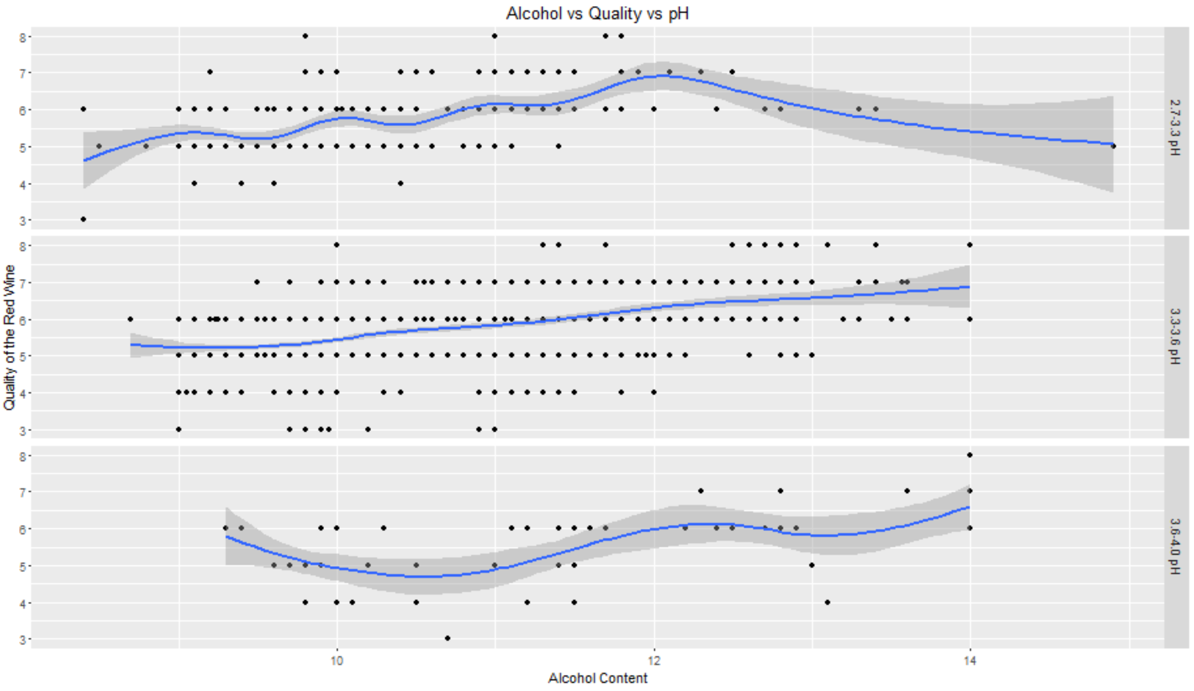


Figure 4.

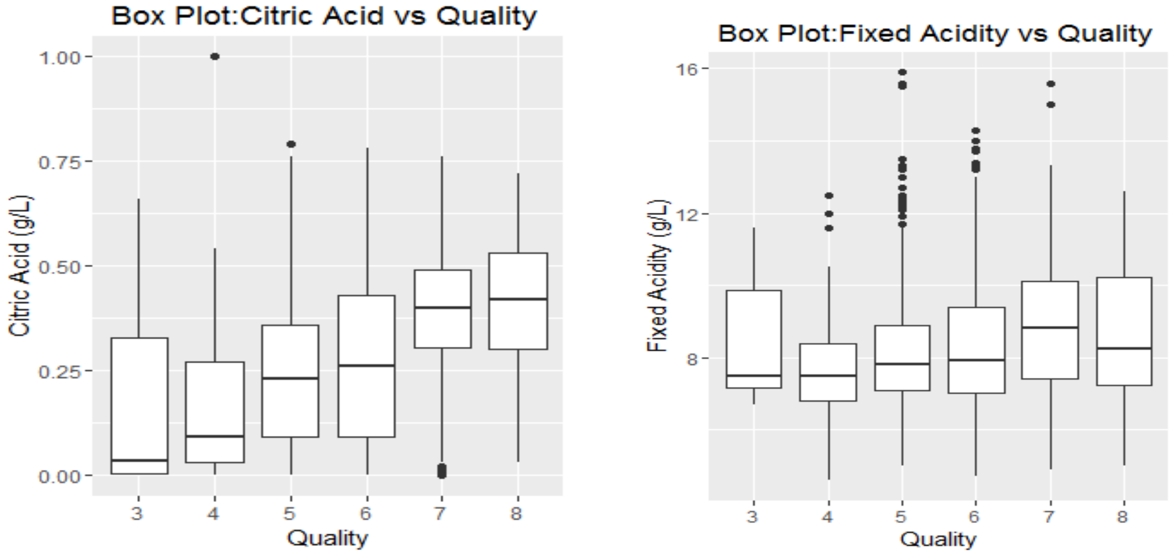
Knowing this curvilinear relationship, we wanted to see if this held true across different variables. One variable, which we are very familiar with is pH. Using the “cut” function in R, we appropriately cut “pH” into 3 separate portions. A lower pH usually indicates a more basic and less acidic taste. As shown in the previous graph, our previous curvilinear aspect of 12-13% alcohol only held true on the conditions of less acidic red wines. Seeing that Acidity is a possible factor, we set out to explore other acidic variables and how those specifically influenced taste.

Citric, Volatile and Fixed Acidity

As a group we learned that certain acids within red wines can alter the taste of the wine and thus changing its quality. From there, we wanted to analyze which type of acidity is most important in determining quality. However, there were three different variables that contained acidic information. These three were Citric Acid, Volatile Acidity, and Fixed Acidity, all measured in (g/L). In order to better understand these factors, our group did some research and learned that different acidity levels had different roles in winemaking and led to many different tastes across different red wines.

We began by analyzing the effects that citric acid and fixed acidity on the quality of red wine. In our data, fixed acidity has a with a mean of 8.31 (g/L), median of 7.9 (g/L) and a range of 4.6 (g/L) - 15.9 (g/L) . Citric acid has a mean of 0.27 (g/L), median of 0.26 (g/L) and range of 0 (g/L) - 1 (g/L). Below are the boxplots used in comparing fixed acidity and citric acid to quality. By analyzing the boxplots, it is clear that citric acid has an apparent positive correlation across the medians of the boxes with quality, meaning that with our data the higher levels of citric acid definitely have a positive effect on the wine's level of quality. However, after analyzing the boxplot for fixed acidity, it is determined that there is no clear correlation between fixed acidity and quality.

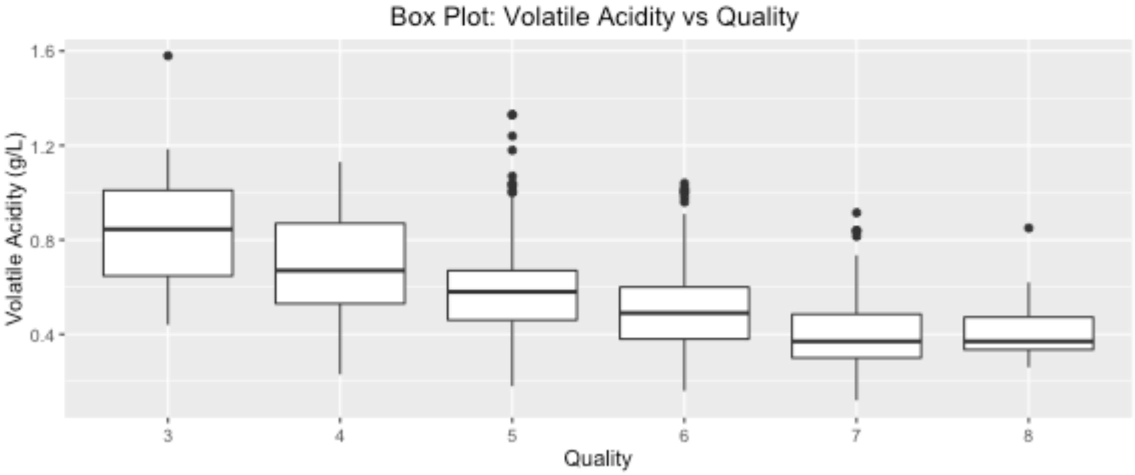
Figure 5.



Through some research we found that citric acid plays a major role in the development and fermentation of wine while also slightly influencing the taste of wine, giving it a more acidic or tart taste. Fixed acidity also plays a large role in wine fermentation, and like the citric acid, the more fixed acids there are in a wine the more acidic it will taste. It is also interesting to note that though our data clearly reveals a correlation between citric acid and the quality of red wine. However, through extensive research we found that this is not the case. Neither citric or fixed acids play a key role in affecting the quality of a wine.

Our group then set out to analyze volatile acidity and how that plays a role in determining the quality. Volatile acidity content is due to the bacteria within the wine creating acetic acid. This is extremely important in the fermentation process as acetic acid creates a distinct vinegar taste and aroma. Bacteria within the wine is difficult for a winemaker to control, as it can form at any point due to oxygen or increased temperatures. The volatile acidity levels ranged from 0.12-1.58 g/L and contained a mean of 0.527 g/L and a median of 0.520 g/L.

Figure 6.



From the plot above, there is an inverse relationship and a trend that shows as the volatile acidity levels decrease, the quality of the red wine increases. This supports our initial thought-process that having too much acidity would not bode well for the overall quality. In order to prevent high volatile acidity levels from ruining wine taste, the reduction of any air within wine containers should be emphasized. The next step would be to ensure that the wine is preserved in cooler temperature; cooler temperatures deter the growth and spread of acetic acid and bacteria.

Part 4: Conclusion

Although it was believed that sugar content would have a dramatic influence on the quality of the wine, that was not the case with our dataset. From our analysis, we determined that the red wines were predominantly “dry” or off dry, meaning that they had relatively low sugar contents. If our analysis had included white wines, which have a larger spectrum of sugar distribution, the results could potentially be different.

From our comparison of data, we found significance with alcohol content and quality, in that they are both positively correlated. And through research and analysis we were able to find the effects that different acidities have on the quality of wine, and concluded that volatile acidity is the most associated with quality. The simplicity of connecting these variables together made us question exactly which variables were connected. Despite not being able to fit linear models and determining which variables have statistical significance on the quality of wine, we were still able to have sufficient analysis.

Code Appendix:

```
# Sta 141a Project Analysis
# Red wine data set
setwd("~/Documents/Academics/Senior 2016-17/Fall 2016/STA 141a")
redWine=read.csv("winequality-red.csv",header=T)
head(redWine)

library(ggplot2)

# fixed.acidity, volatile acidity, citric.acid, residual.sugar
# chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density
# pH, sulphates, alcohol.   quality
summary(redWine)

#Codes for Alcohol Levels vs Quality
range(redWine$alcohol)
alcohol_low = subset(redWine, alcohol < 10)
alcohol_med_low = subset(redWine, alcohol >= 10 & alcohol < 11.5)
alcohol_med = subset(redWine, alcohol >= 11.5 & alcohol <13.5)
alcohol_med_high = subset(redWine, alcohol >13.5)
plot(density(alcohol_low$quality),xlab="Alcohol Content(ABV)", main="Density
Distribution of Quality Scores based on Alcohol Levels")
lines(density(alcohol_med_low$quality),col="green")
lines(density(alcohol_med$quality),col="purple")
lines(density(alcohol_med_high$quality),col="red")
legend(2.853,1.5,legend=c("low","medium-low","medium","medium-
high"),col=c("black","green","purple","red"),lty=c(1,1,1,1))

test = ggplot(data=redwine, aes(alcohol, quality))
test + geom_point() + geom_smooth()  + ggtitle("Alcohol vs Quality") + xlab("Alcohol
Content") + ylab("Quality of the Red Wine")
table(redwine$residual.sugar)
redwine$pH.groups = cut(redwine$pH, 3)
levels(redwine$pH.groups) = c("2.7-3.3 pH","3.3-3.6 pH","3.6-4.0 pH")
test = ggplot(data=redwine, aes(alcohol, quality))
test + geom_point() + geom_smooth() + facet_grid(pH.groups ~., scale="free") +
ggtitle("Alcohol vs Quality vs pH") + xlab("Alcohol Content") + ylab("Quality of the
Red Wine")

## while loop for making factors for sugar levels and also barplot for sugar levels
i = 1
while(i < length(redWine$residual.sugar)){
  if (redWine$residual.sugar[i] < 3){
    redWine$sugarlevel[i] = "dry"
    i = i + 1
    }
  else {
    redWine$sugarlevel[i] = "offdry"
    i = i + 1
    }
}
# Sugar histogram ggpplot 2
```

```
quality_resid = ggplot(redWine, aes(quality, fill =
sugarlevel))+labs(title="Histogram: Dry & Off Dry red
wines",x="Quality",y="Frequency") + geom_bar(position = "identity")
quality_resid

# bar plot using ggplot of pH levels
ph_low = subset(redWine, pH < 3.2)
ph_med = subset(redWine, pH >= 3.2 & pH < 3.65)
ph_high = subset(redWine, pH >= 3.65)
plot(density(ph_low$alcohol), xlab="alcohol concentration", col="purple",
main="Distribution of alcohol concentration based on pH levels")
lines(density(ph_med$alcohol),col=1)
lines(density(ph_high$alcohol),col="green")
legend(13,0.45,title = “”, legend=c("low","medium","high"),col =
c("purple",1,"green"), lty=c(1,1,1))
legend (13,0.45, title = “pH Levels”, legend = c(“”,””,””), bty = “n”, cex = 0.8,
title.adj = 0.15)

plot(density(ph_med$quality), xlab="Quality", main="Probability Distribution of Wine
Quality based on pH Levels")
lines(density(ph_low$quality), col = "purple")
lines(density(ph_high$quality), col = "green")
legend(6.7, 0.89, title = "",legend=c("low","medium","high"),col = c("purple", 1,
"green"), lty = c(1,1,1))
legend (6.7, 0.89,  title = "pH Levels", legend = c("","",""), bty = "n",cex = 0.8,
title.adj = 0.3)

# Boxplots for 3 acidities compared with quality
redWine$quality = as.factor(redWine$quality)
levels.qual=redWine$quality
levels(levels.qual)=c("3","4","5","6","7","8")
citric.quality=ggplot(redWine, aes(quality,citric.acid))
citric.quality + geom_boxplot() + labs(title="Box Plot: Citric Acid vs Quality ",
x="Quality",y="Citric Acid (g/L)")
fixed.quality=ggplot(redWine, aes(quality,fixed.acidity))
fixed.quality + geom_boxplot() + labs(title="Box Plot: Fixed Acidity vs Quality ",
x="Quality",y="Fixed Acidity")
volatile.quality=ggplot(redWine, aes(quality,volatile.acidity))
volatile.quality + geom_boxplot() + labs(title="Box Plot: Volatile Acidity vs Quality
", x="Quality",y="Volatile Acidity (g/L)")
```