Introduction

In this project, we will be considering red vinho verde wine samples, a unique product from the Minho (northwest) region of Portugal. In particular, the red wine quality is of interest, and this project will focus on exploratory data analysis (EDA) to explore the relationships in one variable to multiple variables.

Moving on, more details on the wine data will be given. Then, exploratory data analysis (EDA) on the data will be performed.

Dataset

As mentioned before, the data is about red wine samples (vinho verde) from Portugal. The data was collected from May 2004 to February 2007 using only protected designation of origin samples that were tested at the official certification entity.

```
# Load the Data
Wine <- read.csv("C:/Users/user/Desktop/R project/wineQualityReds.csv")
Wine<-Wine[,-1]
# Save the original data before making any changes
Wine.orig = Wine
# Change quality to categorical
#Wine$quality = as.ordered(Wine$quality)
## Split wine quality into good wine or bad wine
#WineBinaryQuality = Wine.orig
#WineBinaryQuality$quality =
as.factor(ifelse(Wine.orig$quality>5.5,"Good","Bad"))
#glimpse(Wine)
```

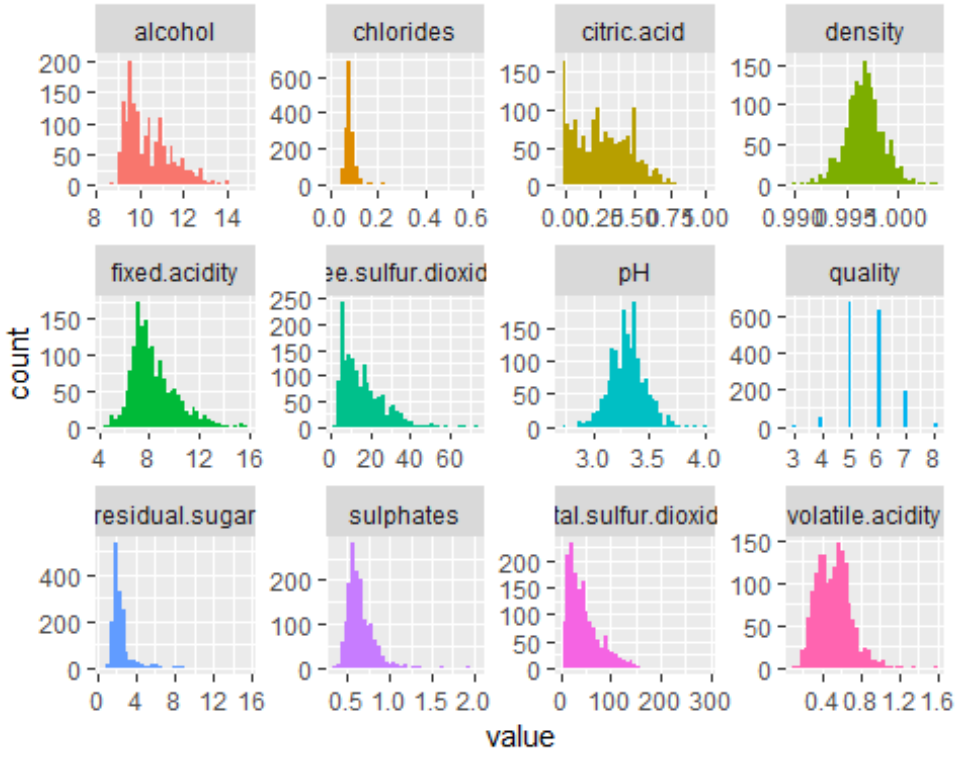
The data contains 1599 observations (wine samples) and 12 attributes or variables related to the wine. The 12 attributes and a description of each variable (attribute) are provided in the table below:

Variable	Description
Fixed acidity	Amount of tartaric acid (in grams) per decimeter cubed of wine (dm^3) [g/dm^3]
Volatile acidity	Amount of acetic acid (in grams) per decimeter cubed of wine (dm^3) [g/dm^3]
Citric acid	Amount of citric acid (in grams) per decimeter cubed of wine (dm^3) [g/dm^3]
Residual sugar	Amount of residual sugar (in grams) per decimeter cubed of wine (dm^3) [g/dm^3]
Chlorides	Amount of sodium chloride (in grams) per decimeter cubed of wine (dm^3) [g/dm^3]

Free sulfur dioxide	Amount of free sulfur dioxide (in milligrams) per decimeter cubed of wine (dm^3) [mg/dm^3]
Total sulfur dioxide	Amount of total sulfur dioxide (in milligrams) per decimeter cubed of wine (dm^3) [mg/dm^3]
Density	Density of wine [g/cm^3]
pH	Acidity/Alkalinity of the wine
Sulphates	Amount of potassium sulphate (in grams) per decimeter cubed of wine (dm^3) [g/dm^3]
Alcohol	Amount of Alcohol by percent volume
Quality	A score between 0 and 10 based on Sensory Data

Univariate Plots Section

```
Wine %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value,fill=key)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram(bins=sqrt(nrow(Wine))) +
  theme(legend.position="none")
```



From Figure 1, one can see that density and pH seem to be symmetric with no heavy tails. This points to the distribution visually looking like a normal distribution. Volatile acidity

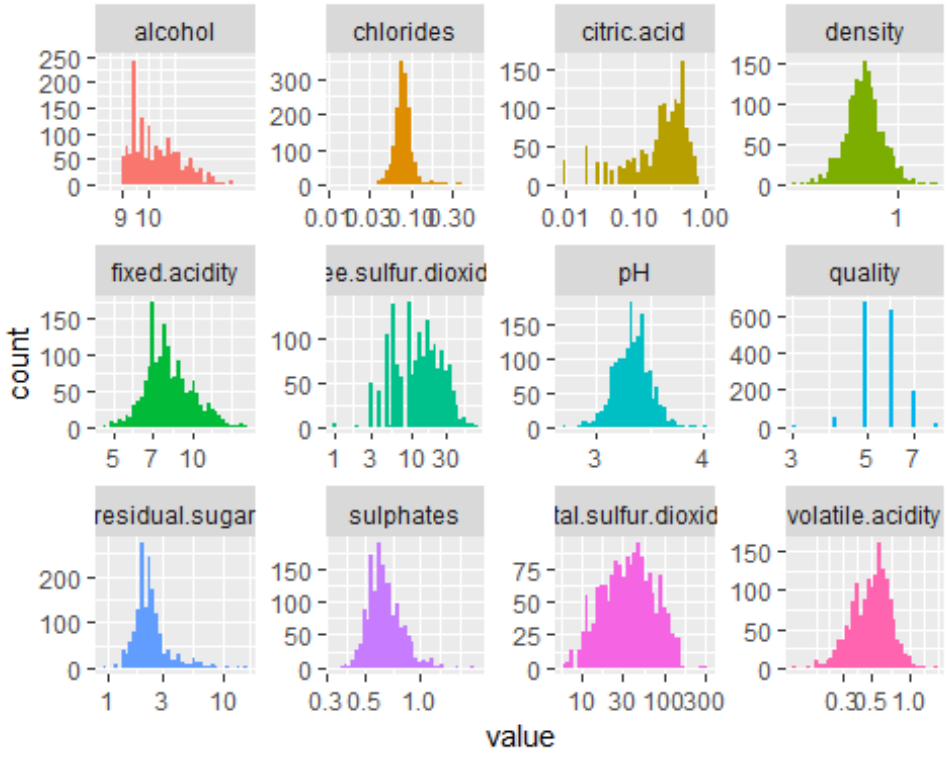
also shows a somewhat normal distribution. Looking at the remaining variables, one can also see that a number of distributions are skewed (with a right tail), like chlorides, fixed acidity, residual sugar, sulphates among others. This means these distributions have mostly lower end values with a few wines with relatively higher values. For example, looking at residual sugar, one can see a majority of the values fall between 0 and 4 grams of residual sugar per decimeter cubed of wine; however, there are a few values that are above 4 and even some that approach 16.

If a skewed distribution is undesirable, one method is to log transform the variable. We can show this effect by making the x axis on a log scale. The same plots are given below, but with the x axis on a log scale.

```
Wine %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value,fill=key)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins=sqrt(nrow(Wine))) +
    theme(legend.position="none") +
    scale_x_continuous(trans='log10')
```

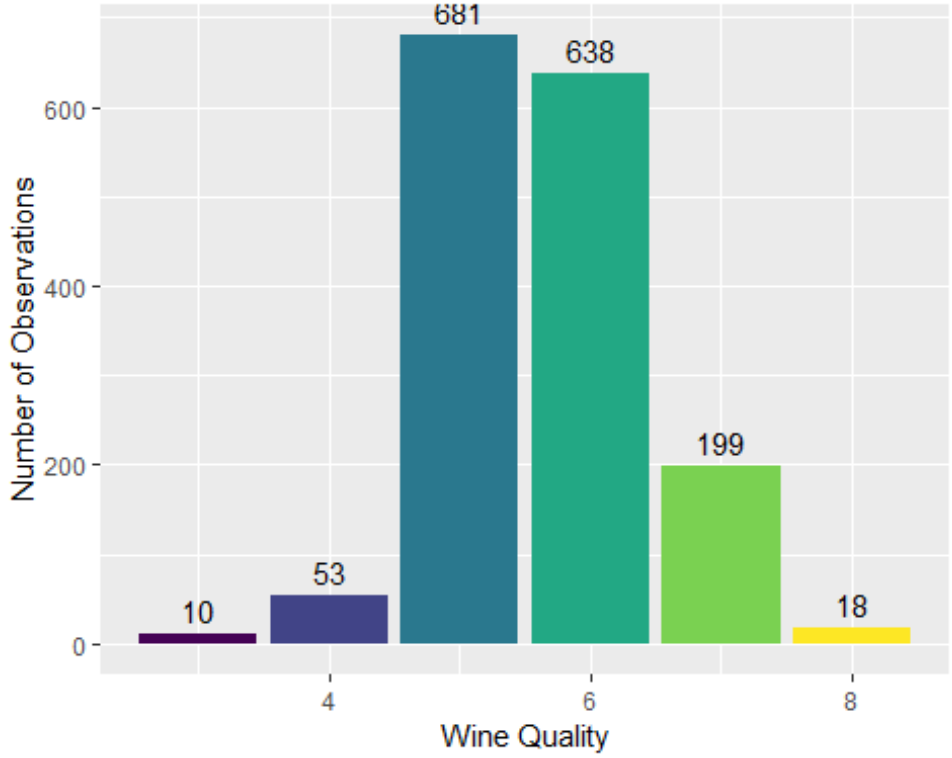
```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 132 rows containing non-finite values (stat_bin).
```




```
Wine$oquality = as.ordered(Wine$quality)
ggplot(Wine, aes(x=quality, fill = oquality)) +
```

```
  geom_bar(stat="count") +
  geom_text(position = "stack", stat='count',aes(label=..count..), vjust = -
0.5)+
  labs(y="Number of Observations", x="Wine Quality") +
  theme(legend.position="none")
```



From Figure 3, one can see quality is not balanced across its entire range of 0-10. Most of the numbers are around 5 or 6. In other words, there are much more normal wines than very excellent or poor ones. This may make it harder to determine what makes an excellent or poor wine. One could just split the wines into a good vs bad wine quality to help alleviate this issue.

With 5 and 6 having a relatively even number of wines and having the majority of the wines overall, it would make logical sense to split the wines into bad and good wines by 5 or below and 6 or above if one wants to make a binary wine quality variable that is somewhat evenly distributed. This will be explored at later. For now, let us continue with the wine qualities as they are.

Univariate Analysis

```
summary(Wine)
```

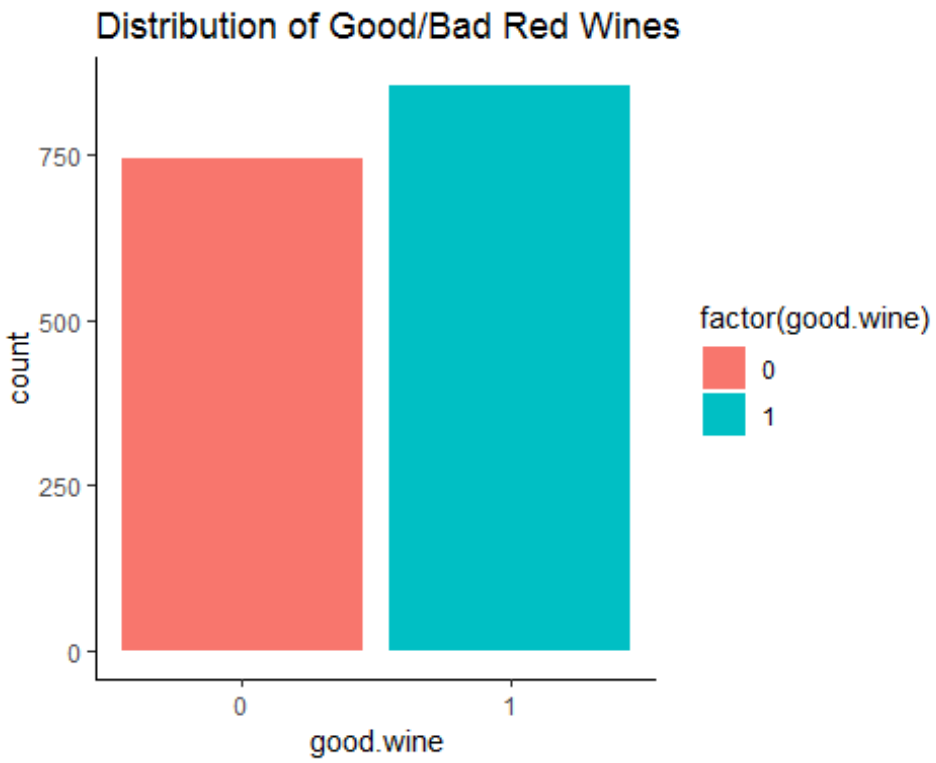
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
```

```
##  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
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality      oquality
##  Min.   :3.000   3: 10
##  1st Qu.:5.000   4: 53
##  Median :6.000   5:681
##  Mean   :5.636   6:638
##  3rd Qu.:6.000   7:199
##  Max.   :8.000   8: 18
```

The features that characterize the red wine are related to acidity (fixed.acidity, volatile.acidity, citric.acid, pH), sugar (residual.sugar), content of sulfur dioxide which is the substance that (free.sulfur.dioxide, total.sulfur.dioxide).

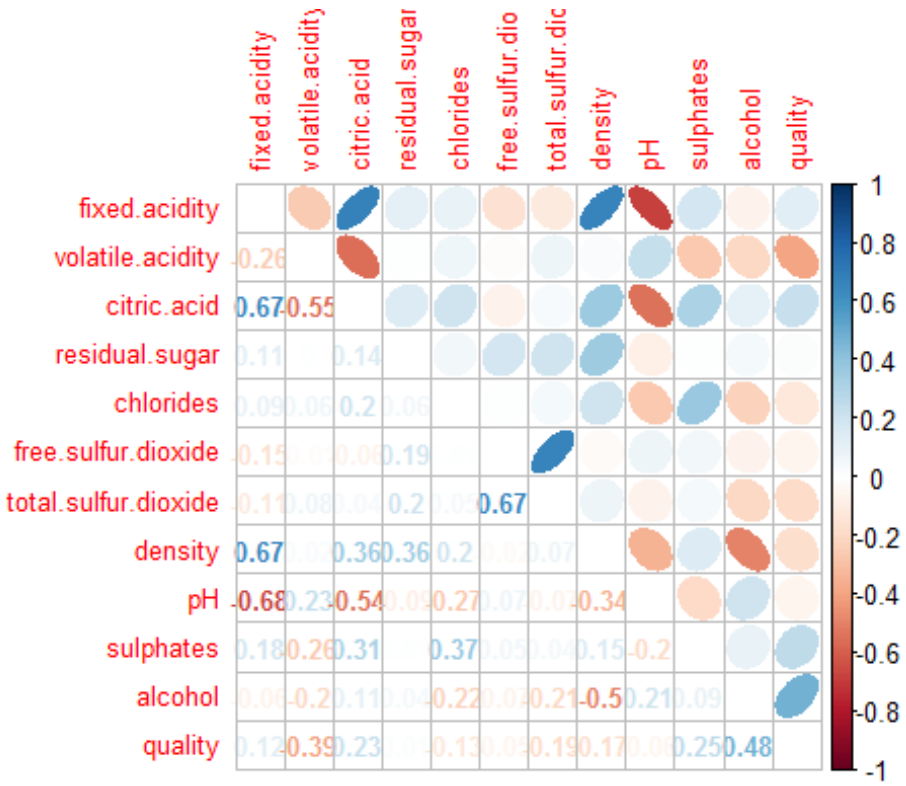
Considering what we have inferred in the previous section, we have created a new variable that separates good wines from bad ones. The following histogram shows the count of this variable

```
Wine$good.wine<-ifelse(Wine$quality>5.5,1,0)

ggplot(Wine,aes(x=good.wine,fill=factor(good.wine)))+geom_bar(stat =
"count",position = "dodge")+
  scale_x_continuous(breaks = seq(0,1,1))+
  ggtitle("Distribution of Good/Bad Red Wines")+
  theme_classic()
```



Bivariate Plots Section

```
Wine.orig %>% cor() %>% corrplot.mixed(upper = "ellipse", tl.cex=.8, tl.pos =
'lt', number.cex = .8)
```



The following dimmensions are relatively highly correlated:

total.sulfur.dioxide with **free.sulfur.dioxide**; **fixed.acidity** with density and **citric.acid**;

The following dimmensions are relatively correlated:

alcohol with **quality** (this might be a candidate for drop, since might be a leak);

The following dimmensions are relativelly highly inverse correlated:

fixed.acidity with **pH**;

The following dimmensions are relatively inverse correlated:

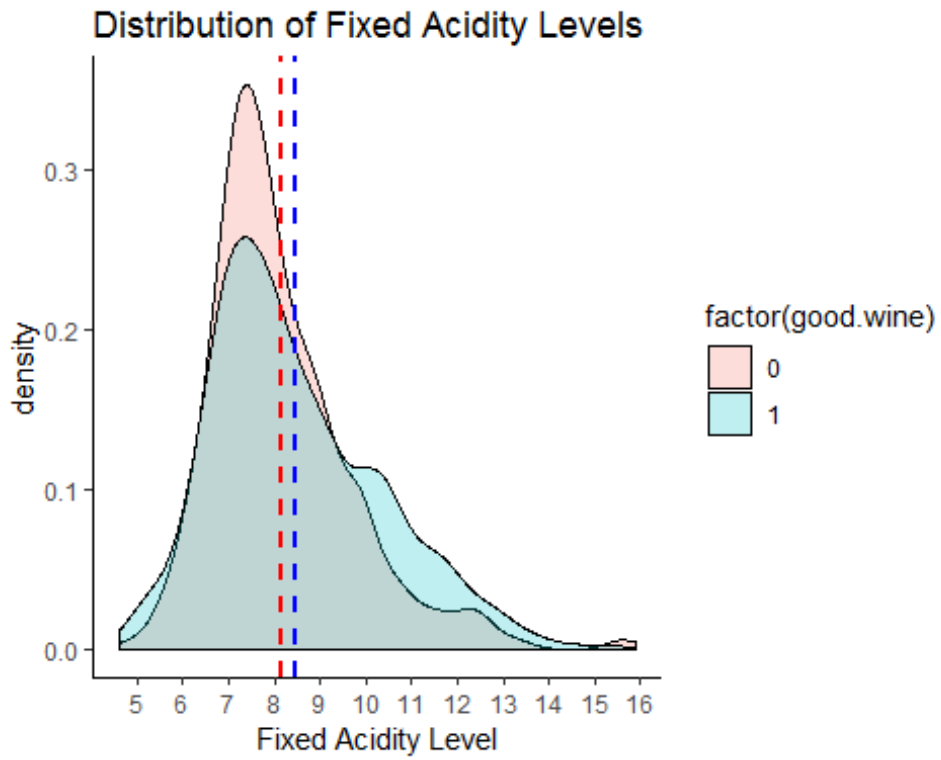
citric.acid with **pH** and volatile.acidity;

We're interested in wine's quality, so we care about the final 2 columns/rows in order to know which among the variables has the strongest relationship with wine quality. As the heatmap suggests, alcohol has the strongest correlation with wine quality.

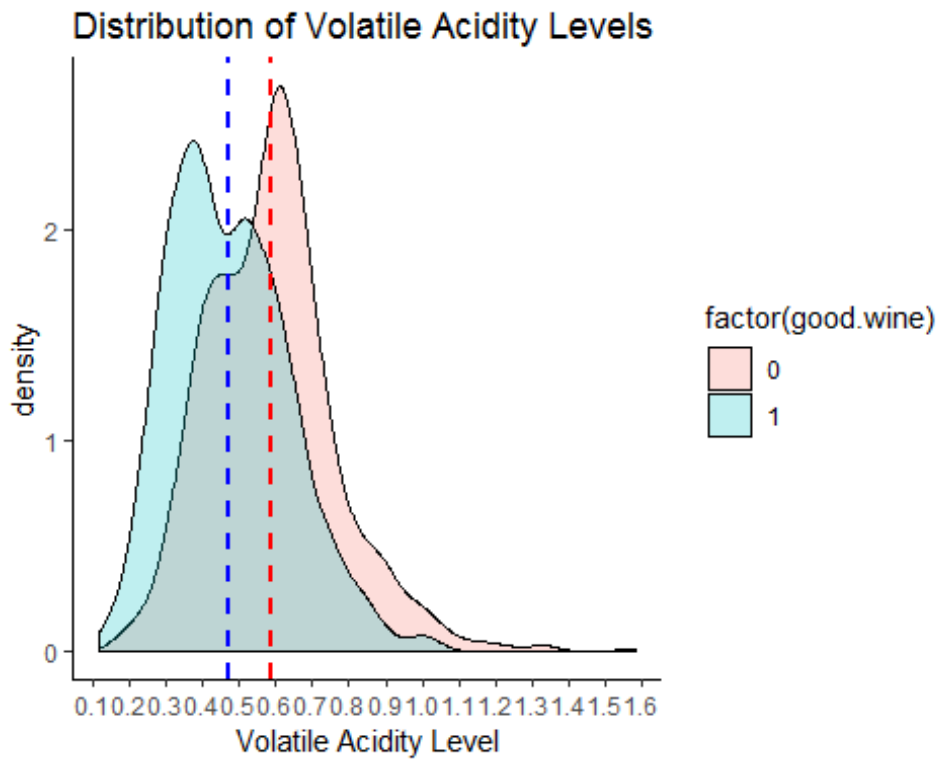
Let us see how each variable is related to wine quality. A method to visualize the relationships between variables is with the pairs plot as shown below

```
ggplot(Wine,aes(x=fixed.acidity,fill=factor(good.wine)))+geom_density(alpha=0
.25)+

geom_vline(aes(xintercept=mean(fixed.acidity[good.wine==0],na.rm=T)),color="r
ed",linetype="dashed",lwd=1)+
```

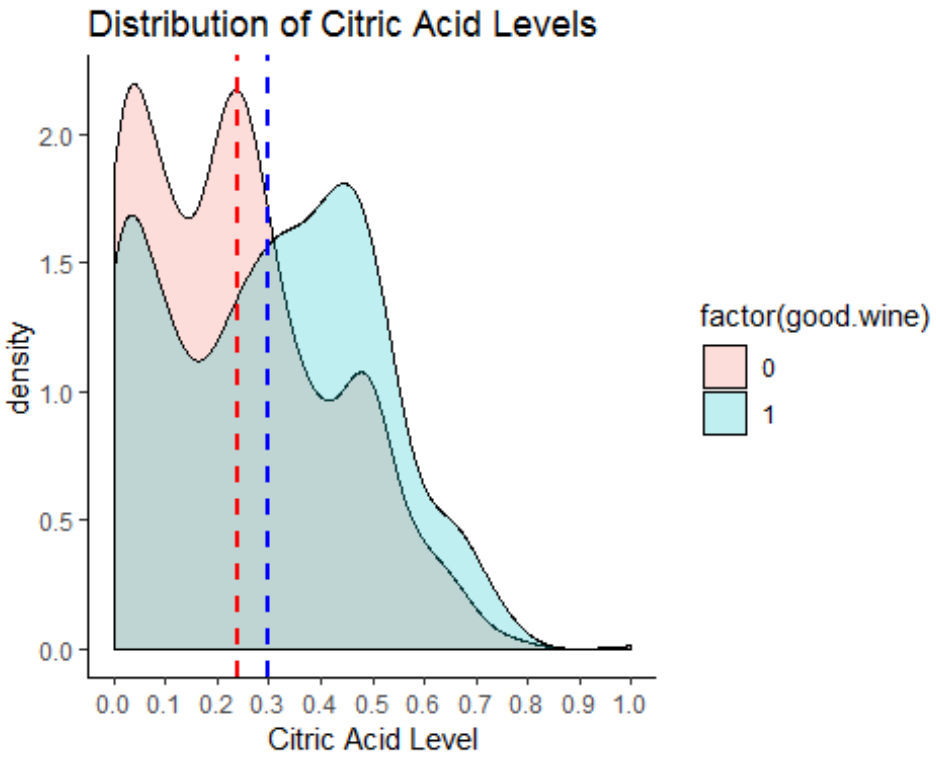
```
geom_vline(aes(xintercept=mean(fixed.acidity[good.wine==1],na.rm=T)),color="b
lue",linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(4,16,1))+
  xlab(label = "Fixed Acidity Level")+
  ggtitle("Distribution of Fixed Acidity Levels")+
  theme_classic()
```



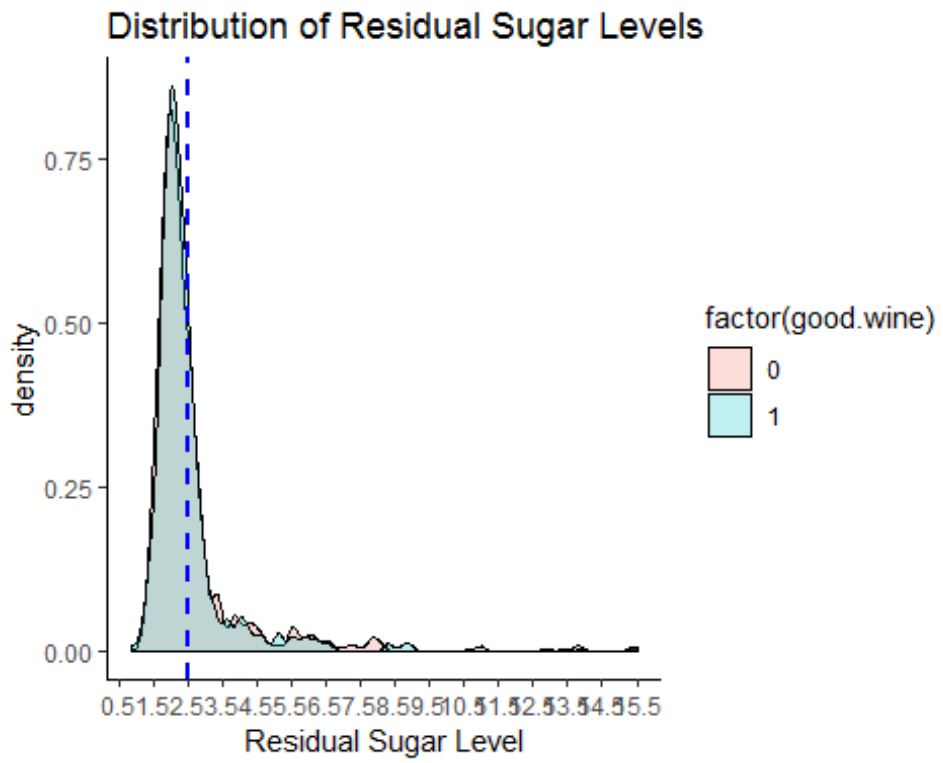

```
#Volatile Acidity and Wine Quality
ggplot(Wine,aes(x=volatile.acidity,fill=factor(good.wine)))+geom_density(alph
a=0.25)+

geom_vline(aes(xintercept=mean(volatile.acidity[good.wine==0],na.rm=T)),color
="red",linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(volatile.acidity[good.wine==1],na.rm=T)),color
="blue",linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0,1.6,0.1))+
  xlab(label = "Volatile Acidity Level")+
  ggtitle("Distribution of Volatile Acidity Levels")+
  theme_classic()
```



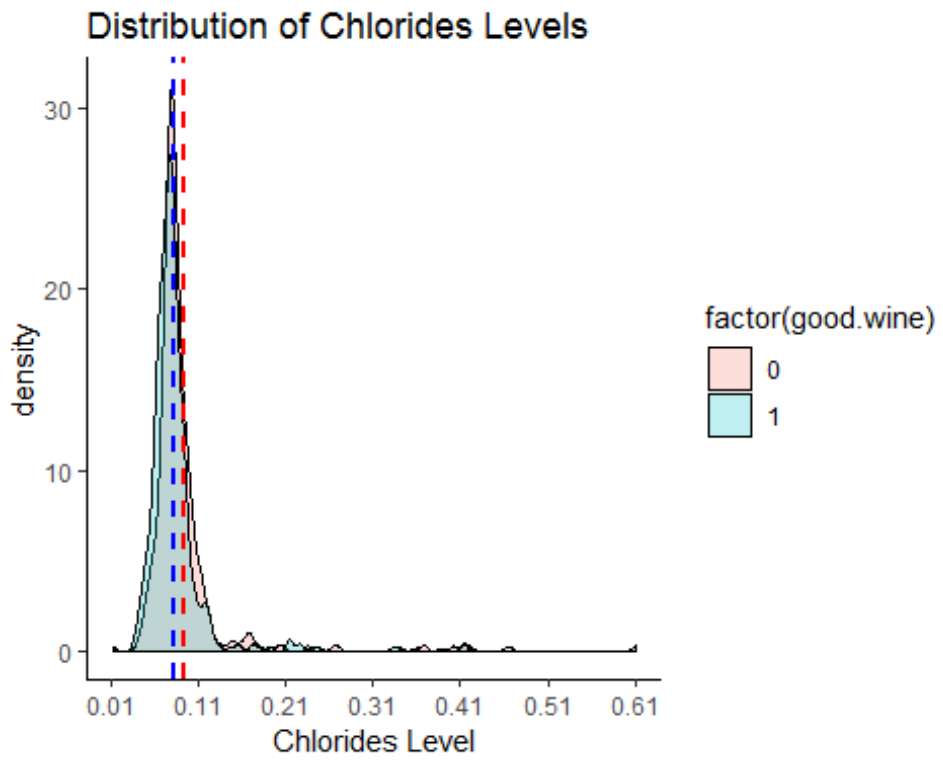

```r
#Citric Acid and Wine Quality
ggplot(Wine,aes(x=citric.acid,fill=factor(good.wine)))+geom_density(alpha=0.25)+

geom_vline(aes(xintercept=mean(citric.acid[good.wine==0],na.rm=T)),color="red",linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(citric.acid[good.wine==1],na.rm=T)),color="blue",linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0,1,0.1))+
  xlab(label = "Citric Acid Level")+
  ggtitle("Distribution of Citric Acid Levels")+
  theme_classic()
```



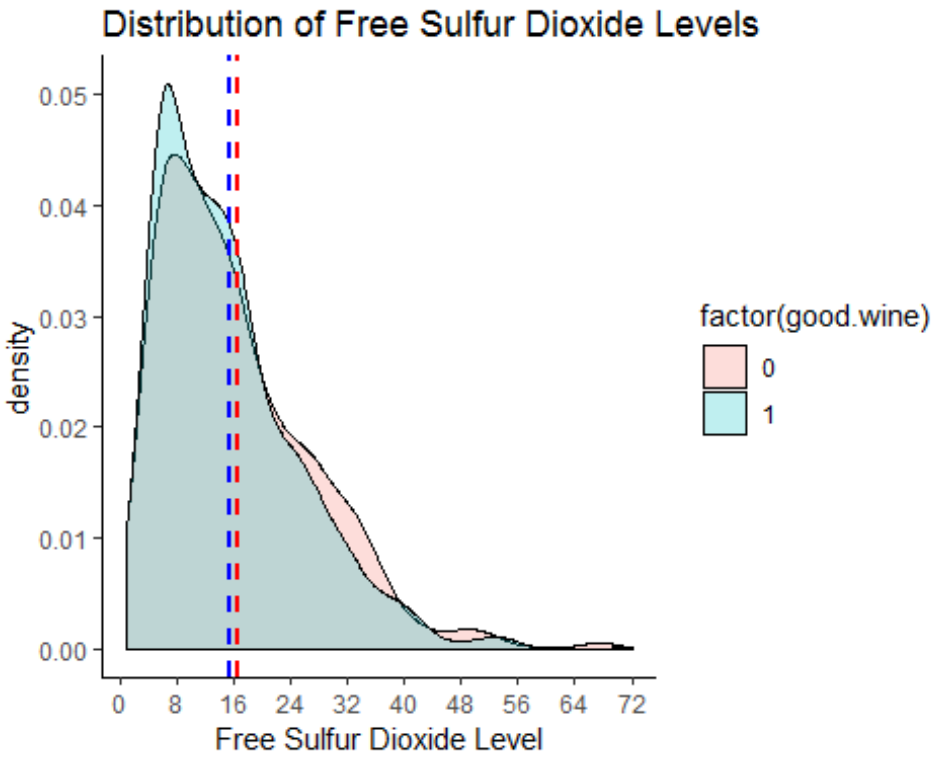

```
#Residual Sugar and Wine Quality
ggplot(Wine,aes(x=residual.sugar,fill=factor(good.wine)))+geom_density(alpha=
0.25)+

geom_vline(aes(xintercept=mean(residual.sugar[good.wine==0],na.rm=T)),color="
red",linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(residual.sugar[good.wine==1],na.rm=T)),color="
blue",linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0.5,15.5,1))+
  xlab(label = "Residual Sugar Level")+
  ggtitle("Distribution of Residual Sugar Levels")+
  theme_classic()
```



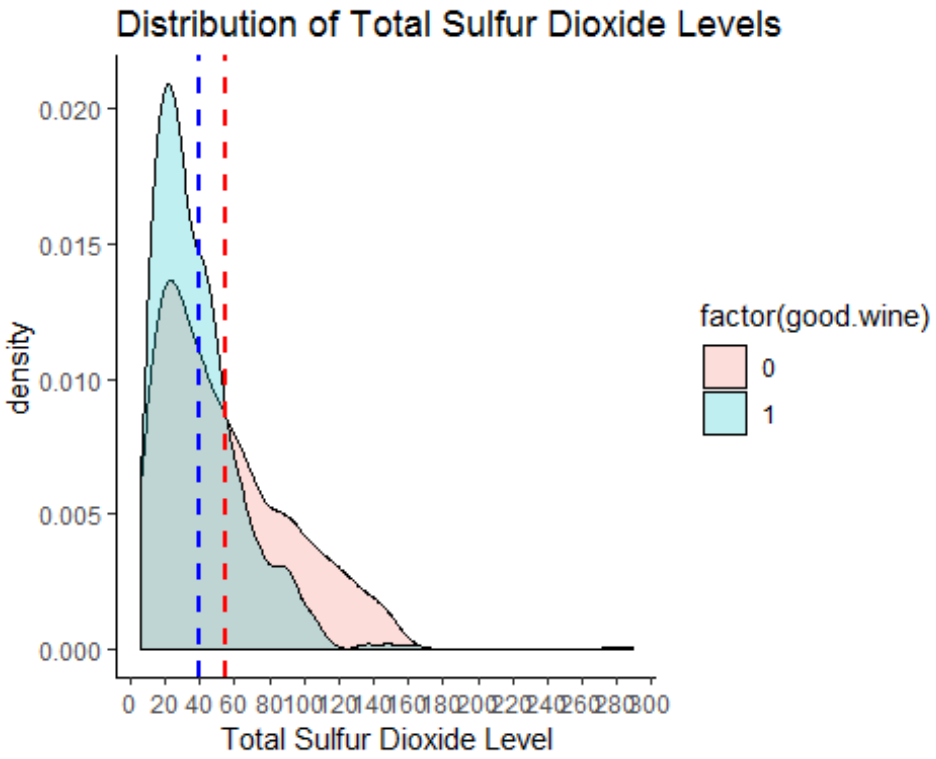

```
#Chlorides and Wine Quality
ggplot(Wine,aes(x=chlorides,fill=factor(good.wine)))+geom_density(alpha=0.25)
+

geom_vline(aes(xintercept=mean(chlorides[good.wine==0],na.rm=T)),color="red",
linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(chlorides[good.wine==1],na.rm=T)),color="blue"
,linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0.01,0.62,0.1))+
  xlab(label = "Chlorides Level")+
  ggtitle("Distribution of Chlorides Levels")+
  theme_classic()
```



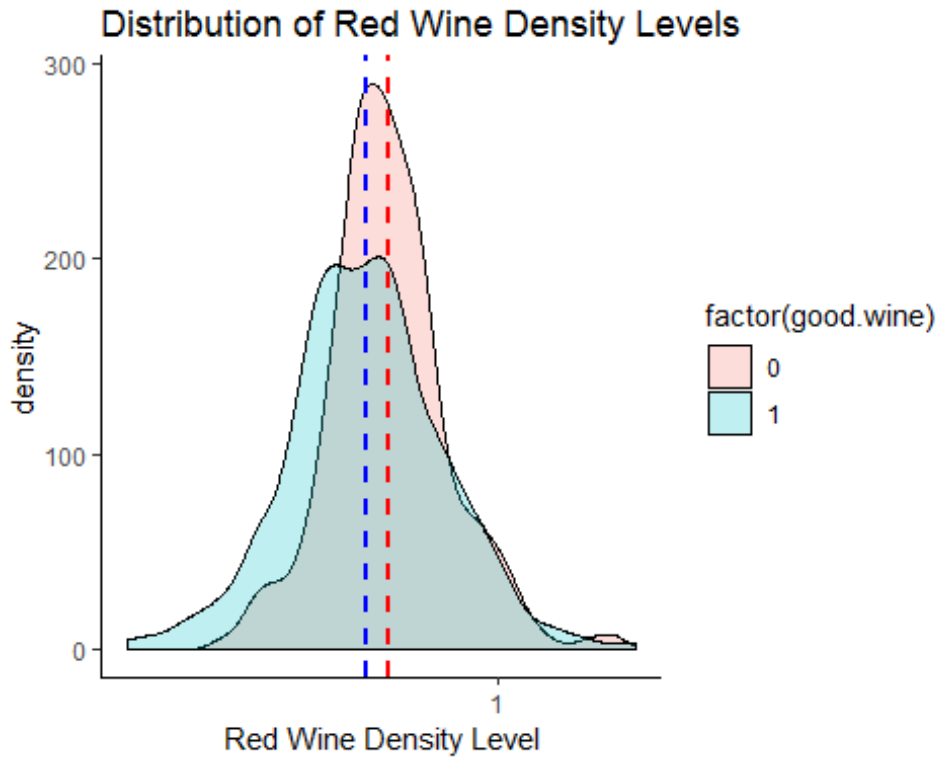

```
#Free Sulfur Dioxide and Wine Quality
ggplot(Wine,aes(x=free.sulfur.dioxide,fill=factor(good.wine)))+geom_density(a
lpha=0.25)+

geom_vline(aes(xintercept=mean(free.sulfur.dioxide[good.wine==0],na.rm=T)),co
lor="red",linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(free.sulfur.dioxide[good.wine==1],na.rm=T)),co
lor="blue",linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0,72,8))+
  xlab(label = "Free Sulfur Dioxide Level")+
  ggtitle("Distribution of Free Sulfur Dioxide Levels")+
  theme_classic()
```



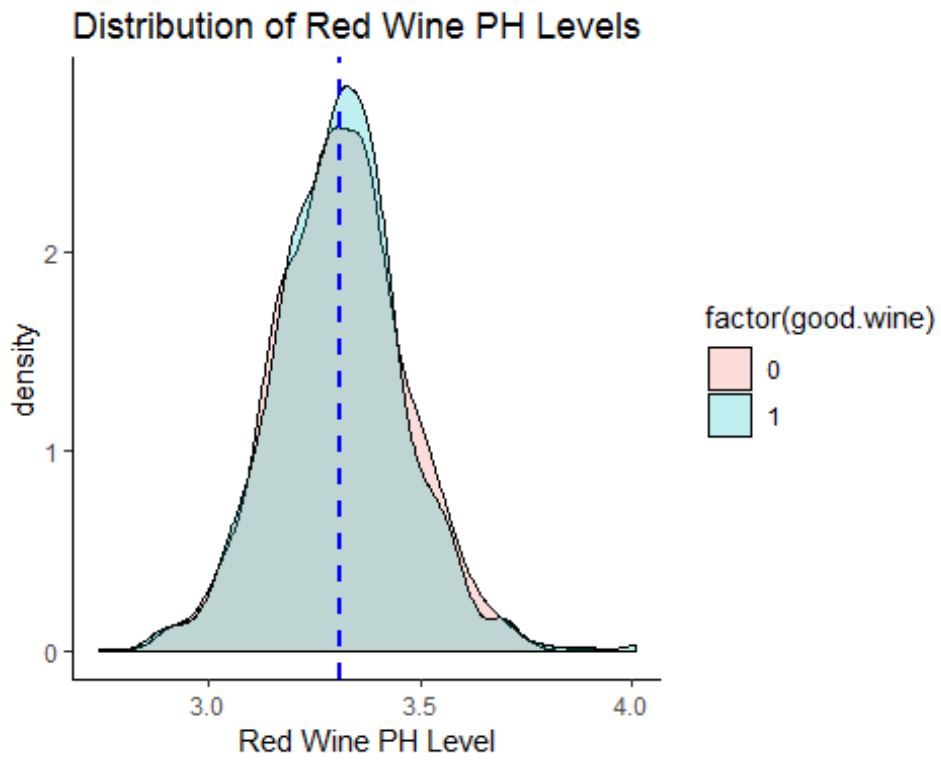

```
#Total Sulfur Dioxide and Wine Quality
ggplot(Wine,aes(x=total.sulfur.dioxide,fill=factor(good.wine)))+geom_density(
alpha=0.25)+

geom_vline(aes(xintercept=mean(total.sulfur.dioxide[good.wine==0],na.rm=T)),c
olor="red",linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(total.sulfur.dioxide[good.wine==1],na.rm=T)),c
olor="blue",linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0,300,20))+
  xlab(label = "Total Sulfur Dioxide Level")+
  ggtitle("Distribution of Total Sulfur Dioxide Levels")+
  theme_classic()
```



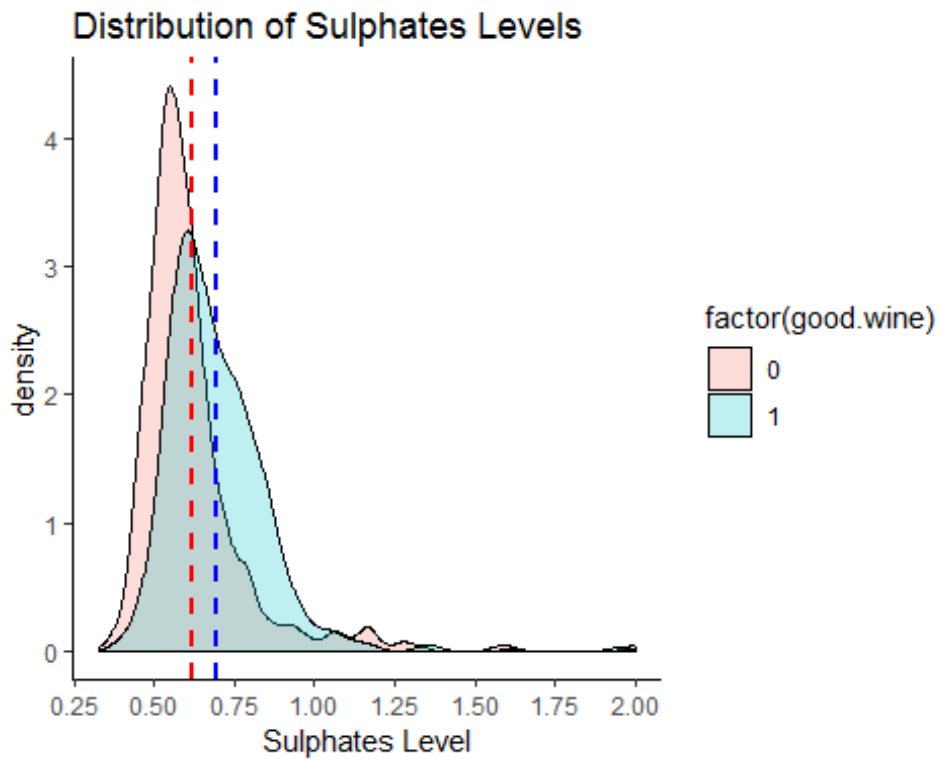

```r
#Density and Wine Quality
ggplot(Wine,aes(x=density,fill=factor(good.wine)))+geom_density(alpha=0.25)+

geom_vline(aes(xintercept=mean(density[good.wine==0],na.rm=T)),color="red",li
netype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(density[good.wine==1],na.rm=T)),color="blue",l
inetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0.9,1.1,0.05))+
  xlab(label = "Red Wine Density Level")+
  ggtitle("Distribution of Red Wine Density Levels")+
  theme_classic()
```



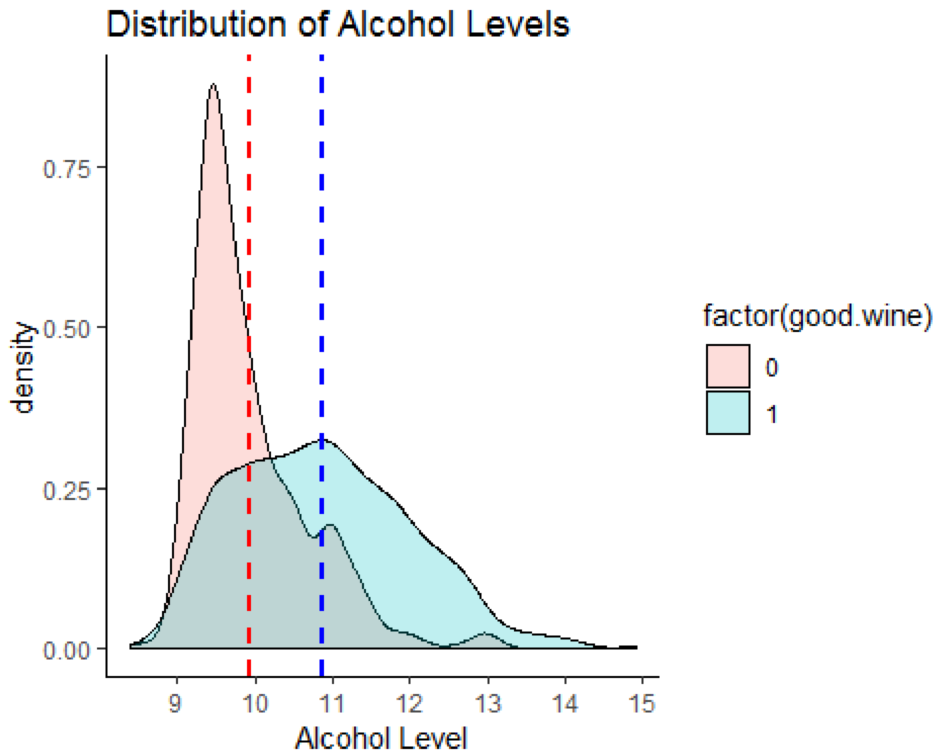

```
#PH and Wine Quality
ggplot(Wine,aes(x=pH,fill=factor(good.wine)))+geom_density(alpha=0.25)+

geom_vline(aes(xintercept=mean(pH[good.wine==0],na.rm=T)),color="red",linetyp
e="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(pH[good.wine==1],na.rm=T)),color="blue",linety
pe="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(2.5,5,0.5))+
  xlab(label = "Red Wine PH Level")+
  ggtitle("Distribution of Red Wine PH Levels")+
  theme_classic()
```




```
#Sulphates and Wine Quality
ggplot(Wine,aes(x=sulphates,fill=factor(good.wine)))+geom_density(alpha=0.25)
+

geom_vline(aes(xintercept=mean(sulphates[good.wine==0],na.rm=T)),color="red",
linetype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(sulphates[good.wine==1],na.rm=T)),color="blue"
,linetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(0,2,0.25))+
  xlab(label = "Sulphates Level")+
  ggtitle("Distribution of Sulphates Levels")+
  theme_classic()
```




```r
#Alcohol and Wine Quality
ggplot(Wine,aes(x=alcohol,fill=factor(good.wine)))+geom_density(alpha=0.25)+

geom_vline(aes(xintercept=mean(alcohol[good.wine==0],na.rm=T)),color="red",li
netype="dashed",lwd=1)+

geom_vline(aes(xintercept=mean(alcohol[good.wine==1],na.rm=T)),color="blue",l
inetype="dashed",lwd=1)+
  scale_x_continuous(breaks = seq(8,15,1))+
  xlab(label = "Alcohol Level")+
  ggtitle("Distribution of Alcohol Levels")+
  theme_classic()
```

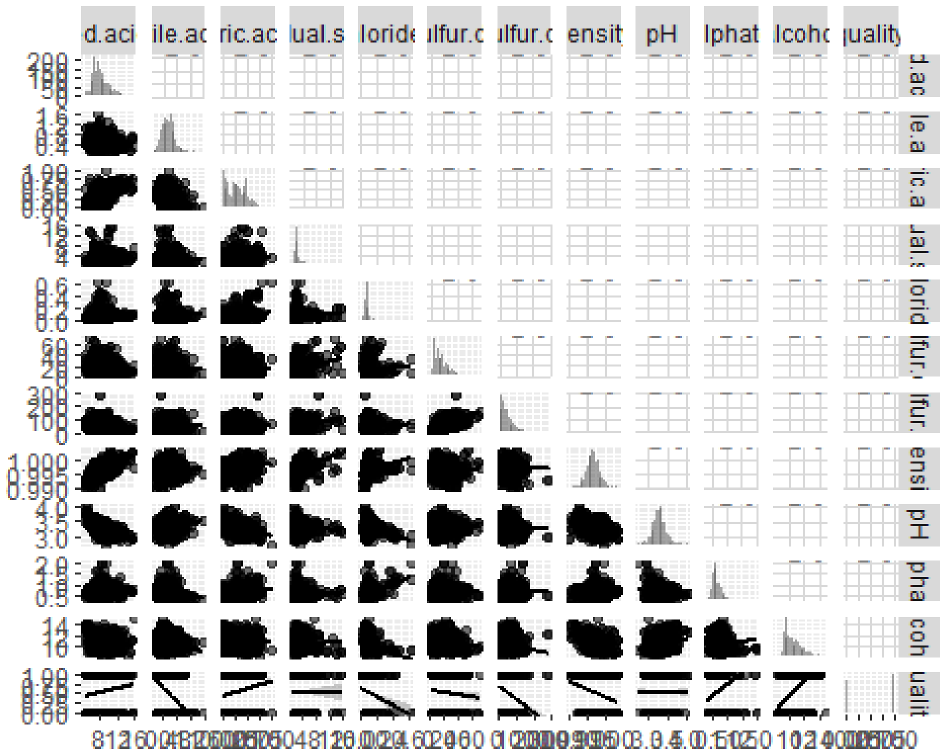


Bivariate Analysis

from the previous pair plots, we can notice that residual_sugar, free_sulfur_dioxide, total_sulfur_dioxide, and chlorides do not have significant different across difference quality. Thus, volatile_acidity, sulphates, and alcohol have more significate difference across different qualities. The rest of the variables are showing some difference; however, the difference among different qualities is not significant enough to determine all the quality directly by using one variable.

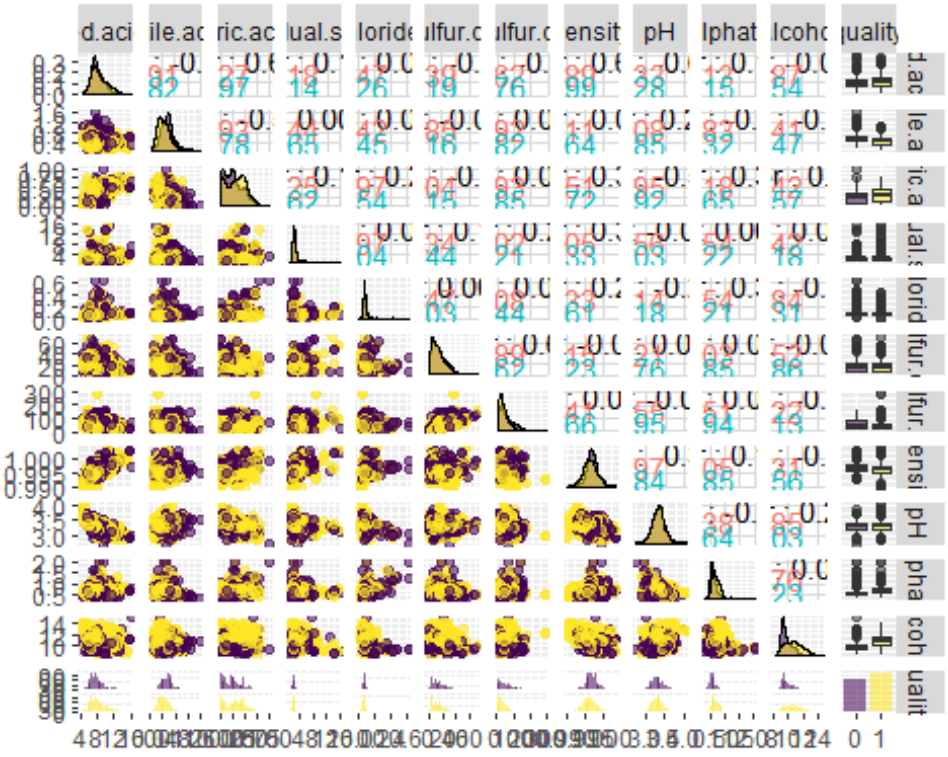
Multivariate Plots Section

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
Winet$quality = as.ordered(Winet$quality)
ggpairs(Winet, aes(colour = quality, alpha = 0.4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Multivariate Analysis

We can observe the following, for each feature:

fixed acidity - besides poor quality, mean value and variance increases with quality;

volatile acidity - smaller means and smaller variance results in increasing quality;

citric acid - quality increases with the mean value;

residual sugar - highest quality has small mean, variance and less outliers;

chlorides - highest quality has smaller mean, variance and less outliers;

free sulfur dioxide - smaller mean and variance are for both small (3) and high quality (8);

total sulfur dioxide - smaller mean and variance are for both small (3) and high quality (8);

density - smaller mean, larger variance for higher quality;

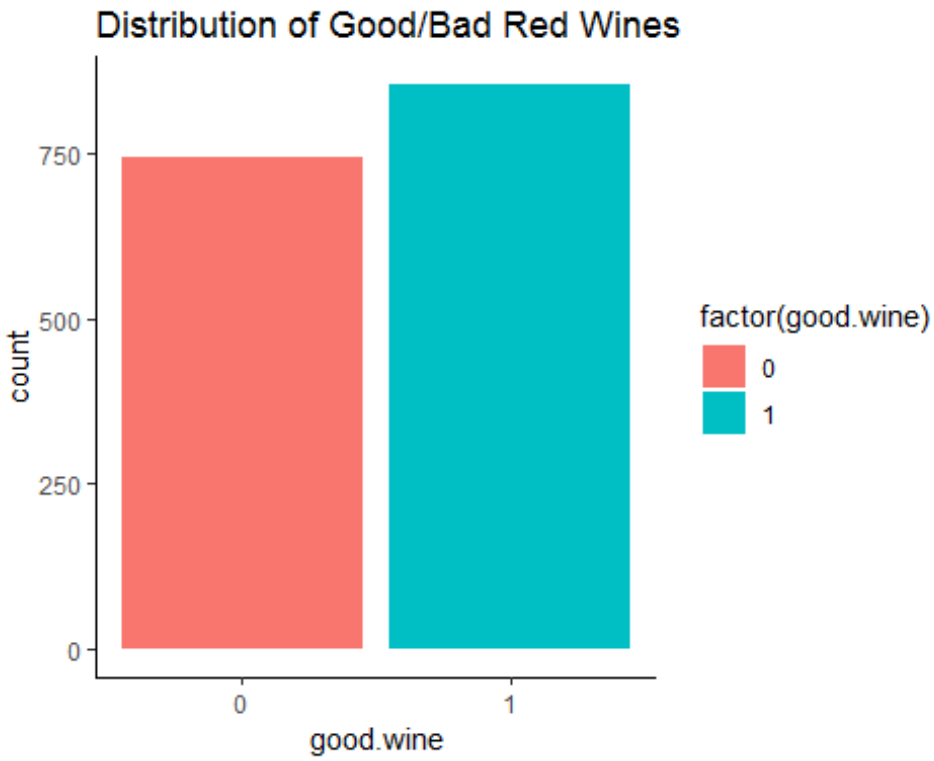
pH - smaller values for higher quality;

sulphates - higher mean, smaller variance, less outliersfor higher quality;

alcohol - higher mean values, larger variance, less outliers for higher quality;

Final Plots and Summary

Plot One

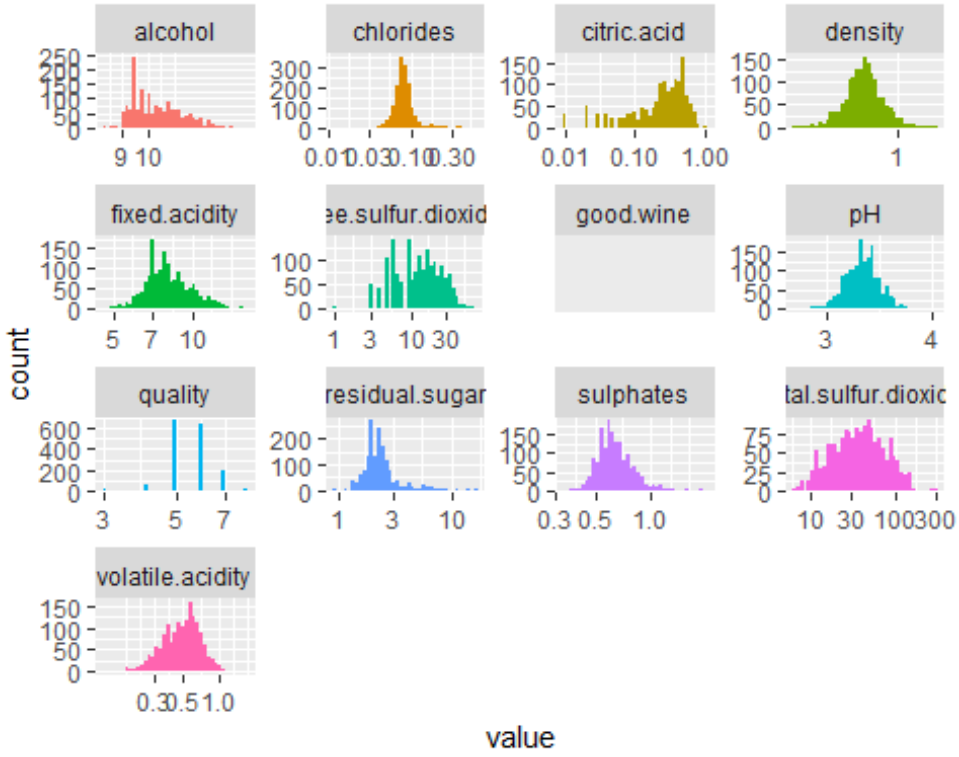


Above plot shows what we have inferred previously, that good wines were not outnumbered by bad wines by a large margin. Most wines were mediocre (rated 5 or 6), but we could also see that there are some poor wines (3 or 4). A vast majority of good wines has a quality rating of 7.

Description One

Plot Two

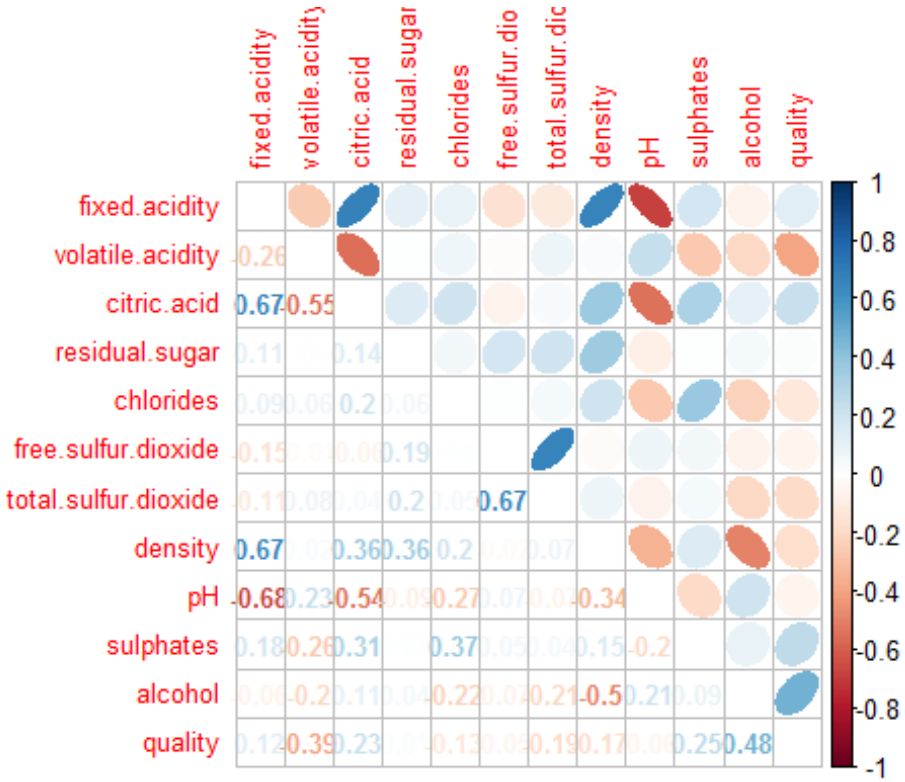
```
## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 876 rows containing non-finite values (stat_bin).

## Warning: Computation failed in `stat_bin()`:
## `binwidth` must be positive
```



Description Two

In classification model, skewed distributions are undesirable, one method is to log transform the variable. We can show this effect by making the x axis on a log scale. The same plots are given above, but with the x axis on a log scale.

Plot Three



Description Three

In feature selection, one method that we use is the correlation matrix and we choose the variables that are highly correlated with our interest variable

Reflection

To wrap it all up, this has been a good investigation of red wine quality. Further analysis can still be made with the data and classification models can also be built (eg. Logistic model, decision trees or SVM). But for this project, we only performed a DEA which the convenient step to perform before building any statistical analysis.