Wine Quality Data Analysis

Sarah

Wine Quality

The following code will evaluate the basic descriptive statistics of the Wine Quality variable of interest/ variable to be tested in the machine learning algorithms.

```
#import and view dataset
library(readxl)
winequalityred <- read_excel("C:/Users/sarah/Desktop/info8000/winequalityred.xlsx",
    col_types = c("numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric"))
QualityMax<-max(winequalityred$quality)
QualityMin<-min(winequalityred$quality)
QualityAverage<-mean(winequalityred$quality)
QualityMedian<-median(winequalityred$quality)
QualityMax
```

```
## [1] 8
```

```
QualityMin
```

```
## [1] 3
```

```
QualityAverage
```

```
## [1] 5.636023
```

```
QualityMedian
```

```
## [1] 6
```

```
QualityMode <- names(which.max(table(winequalityred$quality)))
QualityMode
```

```
## [1] "5"
```

```
QualityStandardDeviation <- sd(winequalityred$quality)
QualityStandardDeviation
```

```
## [1] 0.8075694
```

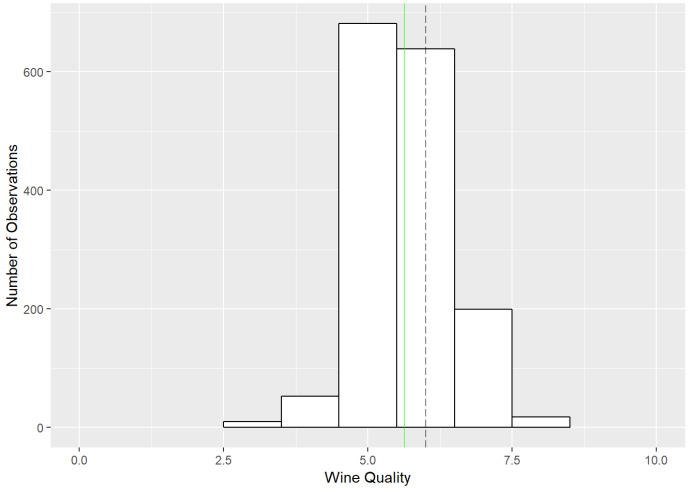
```
#as was stated in the UCI machine learning respository, there aren't many extreme values (very good/very bad wines). most of the wines are "Normal".
```

##Wine Quality Distribution Graph

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
ggplot(data=winequalityred, aes(x=quality)) +
    geom_histogram(binwidth=1, color='black', fill='white') +
    coord_cartesian(xlim=c(0,10)) +
    geom_vline(xintercept = QualityMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = QualityAverage, linetype=1, color='green', alpha=.5) +
    xlab("Wine Quality") +
    ylab("Number of Observations")
```



##Independent Variable Descripive Statistics #Fixed Acidity

```
#fixedac
fixedacMax<-max(winequalityred$fixedacidity)
fixedacMin<-min(winequalityred$fixedacidity)
fixedacAverage<-mean(winequalityred$fixedacidity)
fixedacMedian<-median(winequalityred$fixedacidity)
fixedacMax
```

```
## [1] 15.9
```

```
fixedacMin
```

```
## [1] 4.6
```

```
fixedacAverage
```

```
## [1] 8.319637
```

```
fixedacMedian
```

```
## [1] 7.9
```

```
fixedacMode <- names(which.max(table(winequalityred$fixedacidity)))
fixedacMode
```

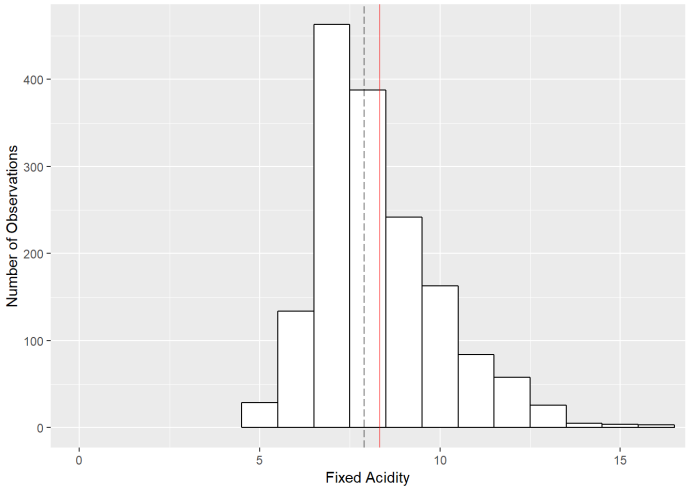
```
## [1] "7.2"
```

```
fixedacStandardDeviation <- sd(winequalityred$fixedacidity)
fixedacStandardDeviation
```

```
## [1] 1.741096
```

#Fixed Acidity Distribution

```
ggplot(data=winequalityred, aes(x=fixedacidity)) +
    geom_histogram(binwidth=1, color='black', fill='white') +
    coord_cartesian(xlim=c(0,16)) +
    geom_vline(xintercept = fixedacMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = fixedacAverage, linetype=1, color='red', alpha=.5) +
    xlab("Fixed Acidity") +
    ylab("Number of Observations")
```



#Volatile Acidity

```
volatileacMax<-max(winequalityred$volatileacidity)
volatileacMin<-min(winequalityred$volatileacidity)
volatileacAverage<-mean(winequalityred$volatileacidity)
volatileacMedian<-median(winequalityred$volatileacidity)
volatileacMax
```

```
## [1] 1.58
```

```
volatileacMin
```

```
## [1] 0.12
```

```
volatileacAverage
```

```
## [1] 0.5278205
```

```
volatileacMedian
```

```
## [1] 0.52
```

```
volatileacMode <- names(which.max(table(winequalityred$volatileacidity)))
volatileacMode
```

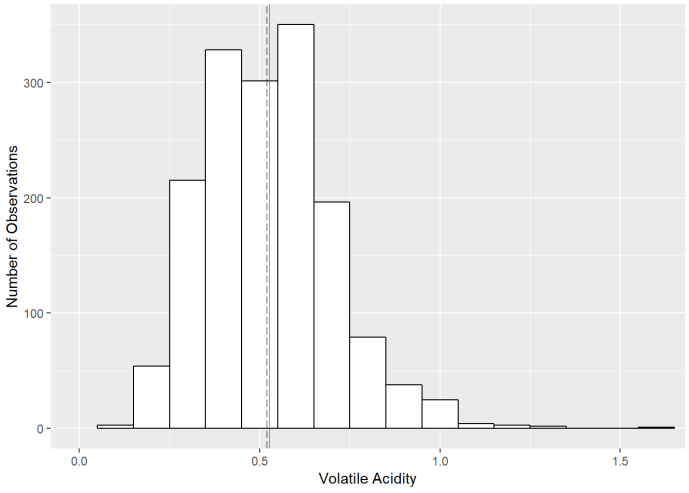
```
## [1] "0.6"
```

```
volatileacStandardDeviation <- sd(winequalityred$volatileacidity)
volatileacStandardDeviation
```

```
## [1] 0.1790597
```

##Volatile Acidity Distribution

```
ggplot(data=winequalityred, aes(x=volatileacidity)) +
    geom_histogram(binwidth=.1, color='black', fill='white') +
    coord_cartesian(xlim=c(0,1.6)) +
    geom_vline(xintercept = volatileacMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = volatileacAverage, linetype=1, color='red', alpha=.5) +
    xlab("Volatile Acidity") +
    ylab("Number of Observations")
```



#Citric Acid

```
citricMax<-max(winequalityred$citricacid)
citricMin<-min(winequalityred$citricacid)
citricAverage<-mean(winequalityred$citricacid)
citricMedian<-median(winequalityred$citricacid)
citricMax
```

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

```
citricMin
```

```
## [1] 0
```

```
citricAverage
```

```
## [1] 0.2709756
```

```
citricMedian
```

```
## [1] 0.26
```

```
citricMode <- names(which.max(table(winequalityred$citricacid)))
citricMode
```

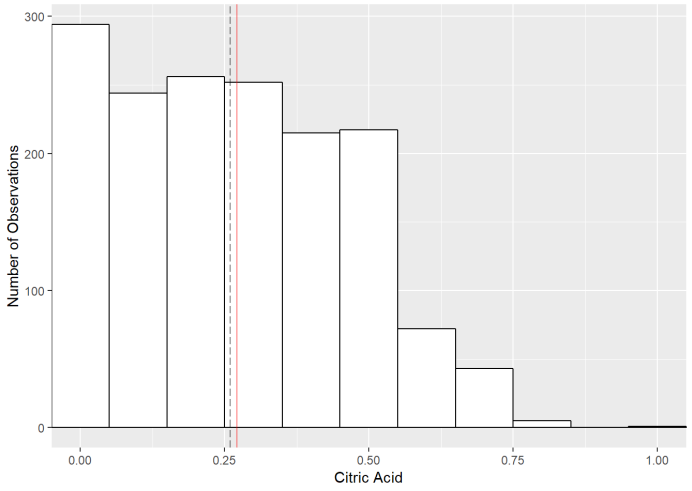
```
## [1] "0"
```

```
citricStandardDeviation <- sd(winequalityred$citricacid)
citricStandardDeviation
```

```
## [1] 0.1948011
```

Citric acid distribution

```
ggplot(data=winequalityred, aes(x=citricacid)) +
    geom_histogram(binwidth=.1, color='black', fill='white') +
    coord_cartesian(xlim=c(0,1)) +
    geom_vline(xintercept = citricMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = citricAverage, linetype=1, color='red', alpha=.5) +
    xlab("Citric Acid") +
    ylab("Number of Observations")
```



#Residual Sugar

```
sugarMax<-max(winequalityred$residualsugar)
sugarMin<-min(winequalityred$residualsugar)
sugarAverage<-mean(winequalityred$residualsugar)
sugarMedian<-median(winequalityred$residualsugar)
sugarMax
```

```
## [1] 15.5
```

```
sugarMin
```

```
## [1] 0.9
```

```
sugarAverage
```

```
## [1] 2.538806
```

```
sugarMedian
```

```
## [1] 2.2
```

```
sugarMode <- names(which.max(table(winequalityred$residualsugar)))
sugarMode
```

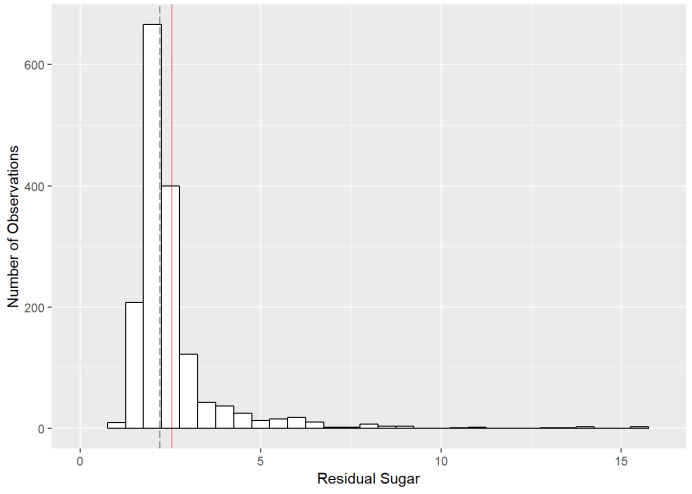
```
## [1] "2"
```

```
sugarStandardDeviation <- sd(winequalityred$residualsugar)
sugarStandardDeviation
```

```
## [1] 1.409928
```

#sugar distribution

```
ggplot(data=winequalityred, aes(x=residualsugar)) +
    geom_histogram(binwidth=.5, color='black', fill='white') +
    coord_cartesian(xlim=c(0,16)) +
    geom_vline(xintercept = sugarMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = sugarAverage, linetype=1, color='red', alpha=.5) +
    xlab("Residual Sugar") +
    ylab("Number of Observations")
```



#Chlorides

```
chloridesMax<-max(winequalityred$chlorides)
chloridesMin<-min(winequalityred$chlorides)
chloridesAverage<-mean(winequalityred$chlorides)
chloridesMedian<-median(winequalityred$chlorides)
chloridesMax
```

```
## [1] 0.611
```

```
chloridesMin
```

```
## [1] 0.012
```

```
chloridesAverage
```

```
## [1] 0.08746654
```

```
chloridesMedian
```

```
## [1] 0.079
```

```
chloridesMode <- names(which.max(table(winequalityred$chlorides)))
chloridesMode
```

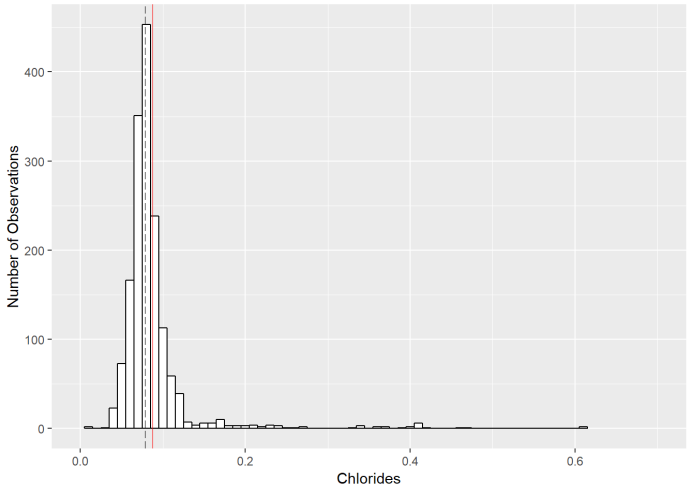
```
## [1] "0.08"
```

```
chloridesStandardDeviation <- sd(winequalityred$chlorides)
chloridesStandardDeviation
```

```
## [1] 0.0470653
```

##Chlorides Distribution

```
ggplot(data=winequalityred, aes(x=chlorides)) +
    geom_histogram(binwidth=.01, color='black', fill='white') +
    coord_cartesian(xlim=c(0,.7)) +
    geom_vline(xintercept = chloridesMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = chloridesAverage, linetype=1, color='red', alpha=.5) +
    xlab("Chlorides") +
    ylab("Number of Observations")
```



#Free SO2

```
freesulfurdioxideMax<-max(winequalityred$freesulfurdioxide)
freesulfurdioxideMin<-min(winequalityred$freesulfurdioxide)
freesulfurdioxideAverage<-mean(winequalityred$freesulfurdioxide)
freesulfurdioxideMedian<-median(winequalityred$freesulfurdioxide)
freesulfurdioxideMax
```

```
## [1] 72
```

```
freesulfurdioxideMin
```

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

```
freesulfurdioxideAverage
```

```
## [1] 15.87492
```

```
freesulfurdioxideMedian
```

```
## [1] 14
```

```
freesulfurdioxideMode <- names(which.max(table(winequalityred$freesulfurdioxide)))
freesulfurdioxideMode
```

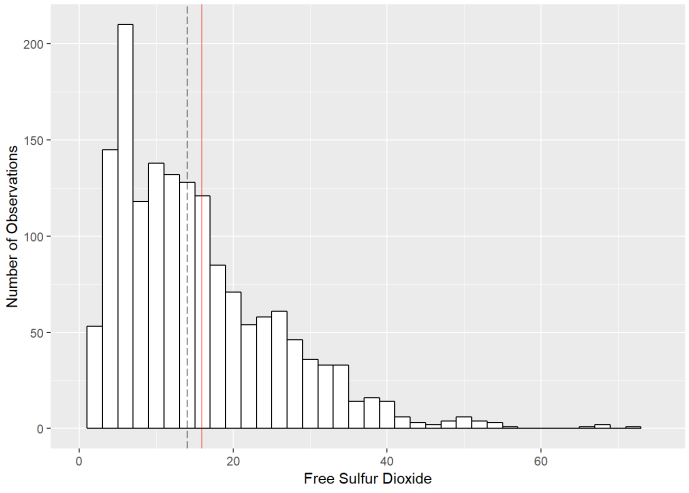
```
## [1] "6"
```

```
freesulfurdioxideStandardDeviation <- sd(winequalityred$freesulfurdioxide)
freesulfurdioxideStandardDeviation
```

```
## [1] 10.46016
```

##Free SO2 Distribution

```
ggplot(data=winequalityred, aes(x=freesulfurdioxide)) +
    geom_histogram(binwidth=2, color='black', fill='white') +
    coord_cartesian(xlim=c(0,75)) +
    geom_vline(xintercept = freesulfurdioxideMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = freesulfurdioxideAverage, linetype=1, color='red', alpha=.5) +
    xlab("Free Sulfur Dioxide") +
    ylab("Number of Observations")
```



#Total SO2

```
totalsulfurdioxideMax<-max(winequalityred$totalsulfurdioxide)
totalsulfurdioxideMin<-min(winequalityred$totalsulfurdioxide)
totalsulfurdioxideAverage<-mean(winequalityred$totalsulfurdioxide)
totalsulfurdioxideMedian<-median(winequalityred$totalsulfurdioxide)
totalsulfurdioxideMax
```

```
## [1] 289
```

```
totalsulfurdioxideMin
```

```
## [1] 6
```

```
totalsulfurdioxideAverage
```

```
## [1] 46.46779
```

```
totalsulfurdioxideMedian
```

```
## [1] 38
```

```
totalsulfurdioxideMode <- names(which.max(table(winequalityred$totalsulfurdioxide)))
totalsulfurdioxideMode
```

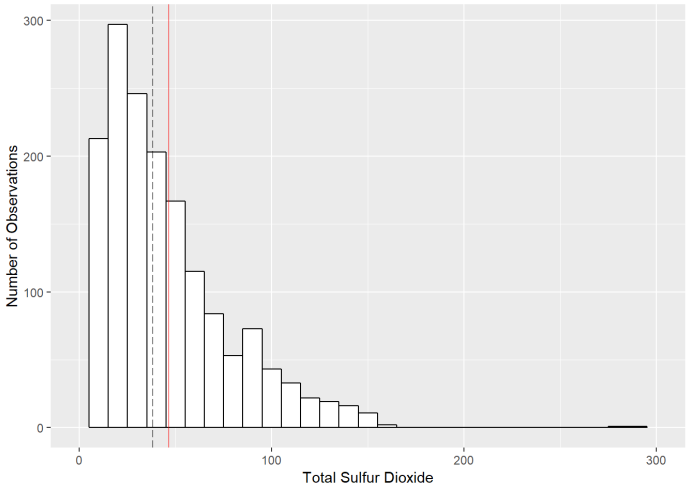
```
## [1] "28"
```

```
totalsulfurdioxideStandardDeviation <- sd(winequalityred$totalsulfurdioxide)
totalsulfurdioxideStandardDeviation
```

```
## [1] 32.89532
```

##Total Sulfur Dioxide Distribution

```
ggplot(data=winequalityred, aes(x=totalsulfurdioxide)) +
    geom_histogram(binwidth=10, color='black', fill='white') +
    coord_cartesian(xlim=c(0,300)) +
    geom_vline(xintercept = totalsulfurdioxideMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = totalsulfurdioxideAverage, linetype=1, color='red', alpha=.5) +
    xlab("Total Sulfur Dioxide") +
    ylab("Number of Observations")
```



#Density

```
densityMax<-max(winequalityred$density)
densityMin<-min(winequalityred$density)
densityAverage<-mean(winequalityred$density)
densityMedian<-median(winequalityred$density)
densityMax
```

```
## [1] 1.00369
```

```
densityMin
```

```
## [1] 0.99007
```

```
densityAverage
```

```
## [1] 0.9967467
```

```
densityMedian
```

```
## [1] 0.99675
```

```
densityMode <- names(which.max(table(winequalityred$density)))
densityMode
```

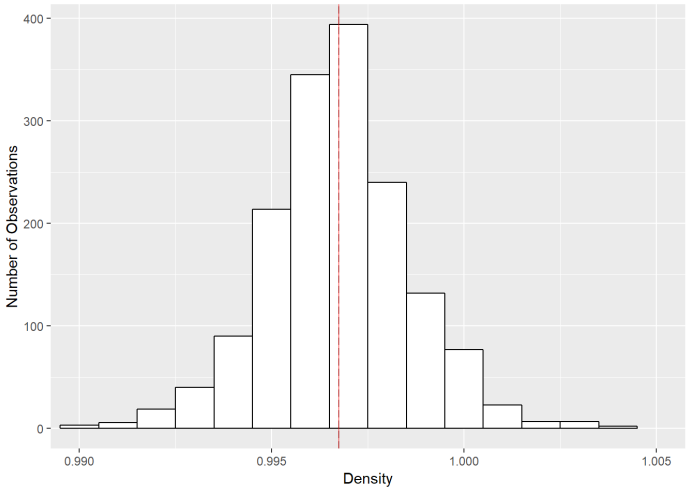
```
## [1] "0.9972"
```

```
densityStandardDeviation <- sd(winequalityred$density)
densityStandardDeviation
```

```
## [1] 0.001887334
```

##Density Distribution

```
ggplot(data=winequalityred, aes(x=density)) +
    geom_histogram(binwidth=.001, color='black', fill='white') +
    coord_cartesian(xlim=c(.99,1.005)) +
    geom_vline(xintercept = densityMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = densityAverage, linetype=1, color='red', alpha=.5) +
    xlab("Density") +
    ylab("Number of Observations")
```



#PH

```
pHMax<-max(winequalityred$pH)
pHMin<-min(winequalityred$pH)
pHAverage<-mean(winequalityred$pH)
pHMedian<-median(winequalityred$pH)
pHMax
```

```
## [1] 4.01
```

```
pHMin
```

```
## [1] 2.74
```

```
pHAverage
```

```
## [1] 3.311113
```

```
pHMedian
```

```
## [1] 3.31
```

```
pHMode <- names(which.max(table(winequalityred$pH)))
pHMode
```

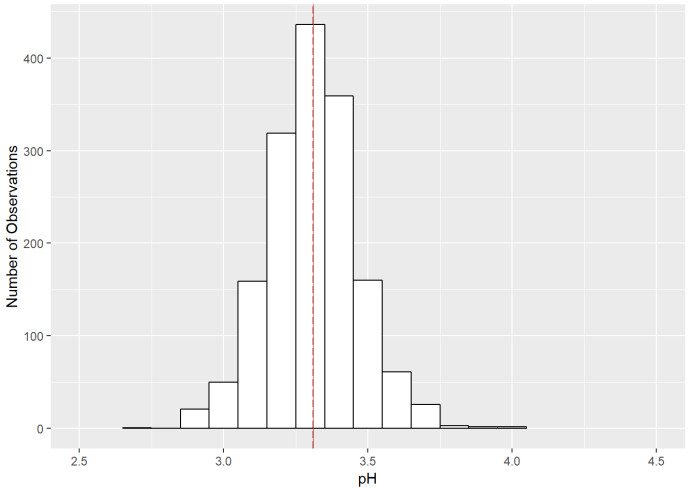
```
## [1] "3.3"
```

```
pHStandardDeviation <- sd(winequalityred$pH)
pHStandardDeviation
```

```
## [1] 0.1543865
```

#pH Distribution

```
ggplot(data=winequalityred, aes(x=pH)) +
    geom_histogram(binwidth=.1, color='black', fill='white') +
    coord_cartesian(xlim=c(2.5,4.5)) +
    geom_vline(xintercept = pHMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = pHAverage, linetype=1, color='red', alpha=.5) +
    xlab("pH") +
    ylab("Number of Observations")
```



#Sulfates

```
sulphatesMax<-max(winequalityred$sulphates)
sulphatesMin<-min(winequalityred$sulphates)
sulphatesAverage<-mean(winequalityred$sulphates)
sulphatesMedian<-median(winequalityred$sulphates)
sulphatesMax
```

```
## [1] 2
```

```
sulphatesMin
```

```
## [1] 0.33
```

```
sulphatesAverage
```

```
## [1] 0.6581488
```

```
sulphatesMedian
```

```
## [1] 0.62
```

```
sulphatesMode <- names(which.max(table(winequalityred$sulphates)))
sulphatesMode
```

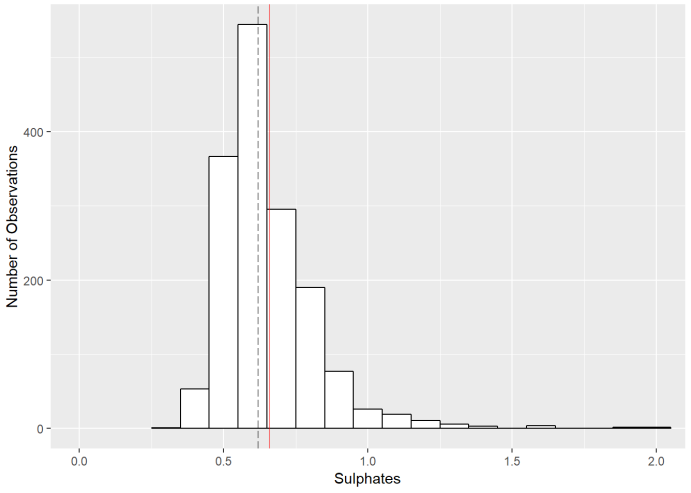
```
## [1] "0.6"
```

```
sulphatesStandardDeviation <- sd(winequalityred$sulphates)
sulphatesStandardDeviation
```

```
## [1] 0.169507
```

##Sulphates Distribution

```
ggplot(data=winequalityred, aes(x=sulphates)) +
    geom_histogram(binwidth=.1, color='black', fill='white') +
    coord_cartesian(xlim=c(0,2)) +
    geom_vline(xintercept = sulphatesMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = sulphatesAverage, linetype=1, color='red', alpha=.5) +
    xlab("Sulphates") +
    ylab("Number of Observations")
```



#Alcohol

```
alcoholMax<-max(winequalityred$alcohol)
alcoholMin<-min(winequalityred$alcohol)
alcoholAverage<-mean(winequalityred$alcohol)
alcoholMedian<-median(winequalityred$alcohol)
alcoholMax
```

```
## [1] 14.9
```

```
alcoholMin
```

```
## [1] 8.4
```

```
alcoholAverage
```

```
## [1] 10.42298
```

```
alcoholMedian
```

```
## [1] 10.2
```

```
alcoholMode <- names(which.max(table(winequalityred$alcohol)))
alcoholMode
```

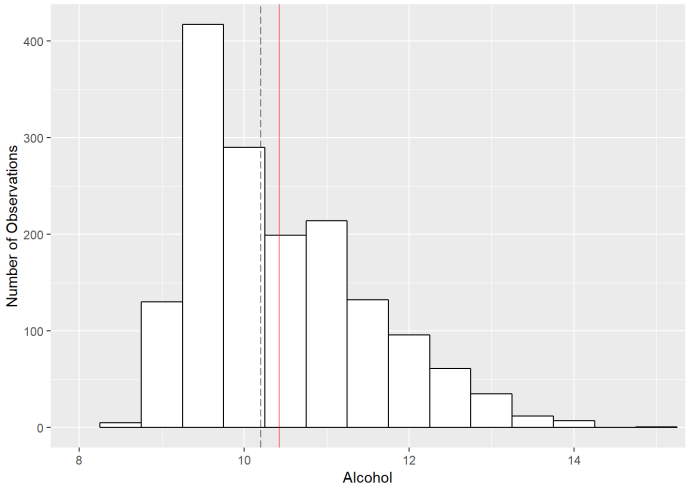
```
## [1] "9.5"
```

```
alcoholStandardDeviation <- sd(winequalityred$alcohol)
alcoholStandardDeviation
```

```
## [1] 1.065668
```

##Alcohol Distribution

```
ggplot(data=winequalityred, aes(x=alcohol)) +
    geom_histogram(binwidth=.5, color='black', fill='white') +
    coord_cartesian(xlim=c(8,15)) +
    geom_vline(xintercept = alcoholMedian, linetype='longdash', alpha=.5) +
    geom_vline(xintercept = alcoholAverage, linetype=1, color='red', alpha=.5) +
    xlab("Alcohol") +
    ylab("Number of Observations")
```



##Variable Correlations

```
wine.cor = cor(winequalityred)
library("Hmisc")
```

```
## Warning: package 'Hmisc' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 3.5.2
```

```
##
## Attaching package: 'Hmisc'
```

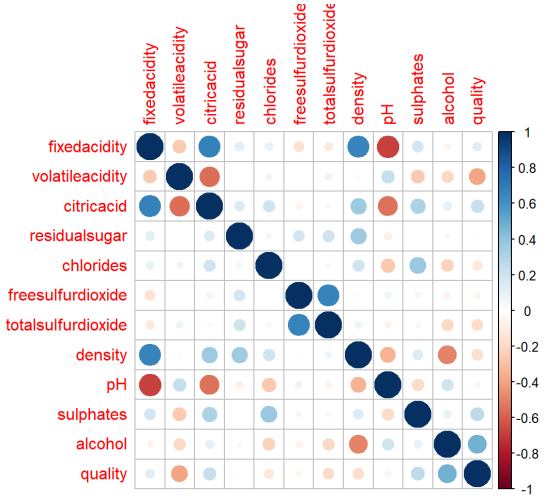
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
wine.rcorr = rcorr(as.matrix(winequalityred))
wine.coeff = wine.rcorr$r
wine.p = wine.rcorr$P
library(corrplot)
```

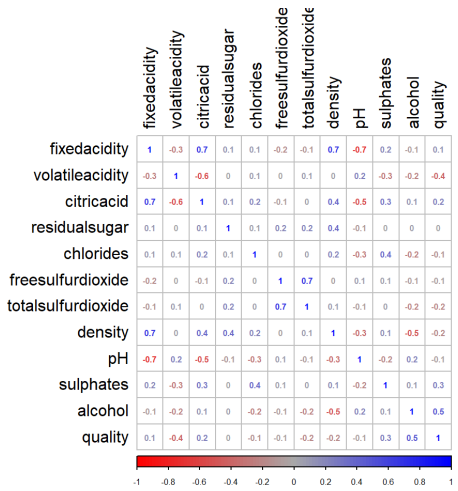
```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```
corrplot(wine.cor)
```




```
corrplot(wine.cor, tl.cex = 1, tl.col = "black", method = "number",
         addCoef.col = "black", number.digits = 1, number.cex = .5,
         cl.pos = 'b', cl.cex = .5, addrect = .5, rect.lwd = 1,
col = colorRampPalette(c("red", "darkgrey","blue"))(100))
```



#correlation cont'd

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.5.3
```

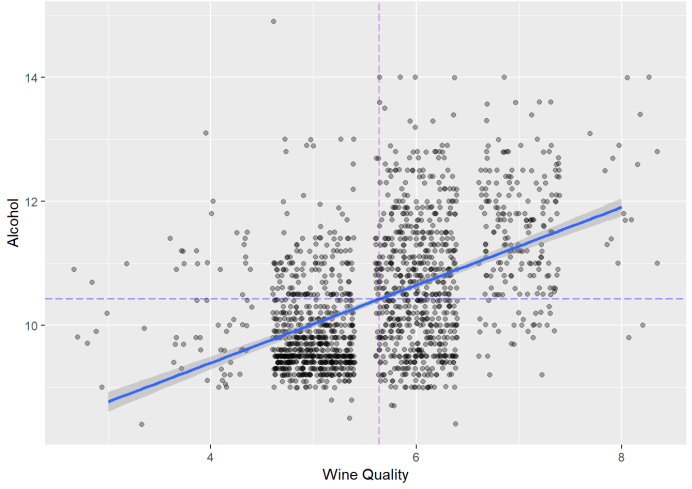
```
kable(wine.cor, caption = "Wine quality feature correlations")
```

wine quality feature correlations

	fixedacidity	volatileacidity	citricacid	residualsugar	chlorides	freesulfurdioxide	totalsulfurdioxide	density	pH	sulphates	alcohol	
fixedacidity	1.0000000	-0.2561309	0.6717034	0.1147767	0.0937052	-0.1537942	-0.1131814	0.6680473	-0.6829782	0.1830057	-0.0616683	0.1
volatileacidity	-0.2561309	1.0000000	-0.5524957	0.0019179	0.0612978	-0.0105038	0.0764700	0.0220262	0.2349373	-0.2609867	-0.2022880	-0.3
citricacid	0.6717034	-0.5524957	1.0000000	0.1435772	0.2038229	-0.0609781	0.0355330	0.3649472	-0.5419041	0.3127700	0.1099032	0.2
residualsugar	0.1147767	0.0019179	0.1435772	1.0000000	0.0556095	0.1870490	0.2030279	0.3552834	-0.0856524	0.0055271	0.0420754	0.0
chlorides	0.0937052	0.0612978	0.2038229	0.0556095	1.0000000	0.0055621	0.0474005	0.2006323	-0.2650261	0.3712605	-0.2211405	-0.1
freesulfurdioxide	-0.1537942	-0.0105038	-0.0609781	0.1870490	0.0055621	1.0000000	0.6676665	-0.0219458	0.0703775	0.0516576	-0.0694084	-0.0
totalsulfurdioxide	-0.1131814	0.0764700	0.0355330	0.2030279	0.0474005	0.6676665	1.0000000	0.0712695	-0.0664946	0.0429468	-0.2056539	-0.1
density	0.6680473	0.0220262	0.3649472	0.3552834	0.2006323	-0.0219458	0.0712695	1.0000000	-0.3416993	0.1485064	-0.4961798	-0.1
pH	-0.6829782	0.2349373	-0.5419041	-0.0856524	-0.2650261	0.0703775	-0.0664946	-0.3416993	1.0000000	-0.1966476	0.2056325	-0.0
sulphates	0.1830057	-0.2609867	0.3127700	0.0055271	0.3712605	0.0516576	0.0429468	0.1485064	-0.1966476	1.0000000	0.0935948	0.2
alcohol	-0.0616683	-0.2022880	0.1099032	0.0420754	-0.2211405	-0.0694084	-0.2056539	-0.4961798	0.2056325	0.0935948	1.0000000	0.4
quality	0.1240516	-0.3905578	0.2263725	0.0137316	-0.1289066	-0.0506561	-0.1851003	-0.1749192	-0.0577314	0.2513971	0.4761663	1.0

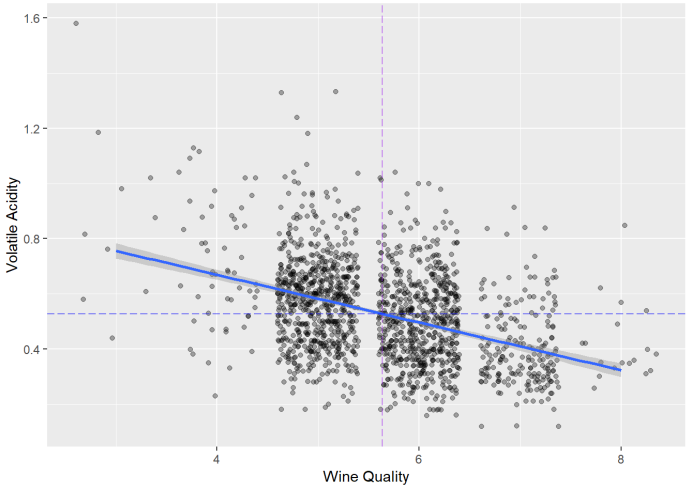
#Alcohol Vs. Quality

```
ggplot(data=winequalityred, aes(x=as.numeric(quality), y=alcohol)) +
    geom_jitter(alpha=1/3) +
    geom_smooth(method='lm', aes(group = 1))+
    geom_hline(yintercept=alcoholAverage, linetype='longdash', alpha=.5, color='blue') +
    geom_vline(xintercept = QualityAverage, linetype='longdash', color='purple', alpha=.5) +
    xlab("Wine Quality") +
    ylab("Alcohol")
```



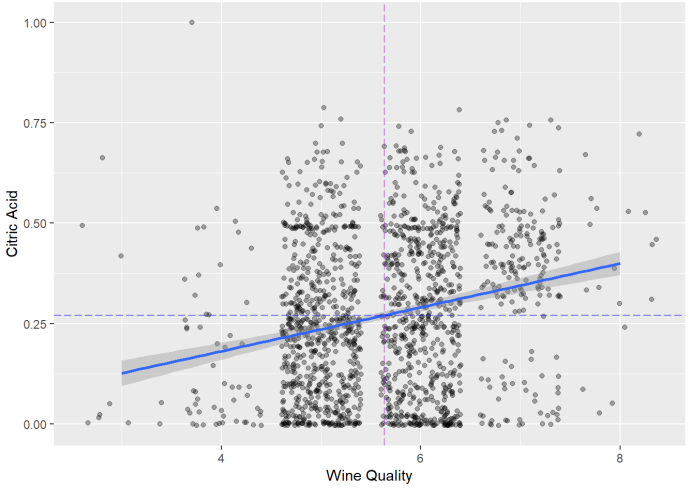
#Volatile Acidity Vs. Quality:

```
ggplot(data=winequalityred, aes(x=as.numeric(quality), y=volatileacidity)) +
    geom_jitter(alpha=1/3) +
    geom_smooth(method='lm', aes(group = 1))+
    geom_hline(yintercept=volatileacAverage, linetype='longdash', alpha=.5, color='blue') +
    geom_vline(xintercept = QualityAverage, linetype='longdash', color='purple', alpha=.5) +
    xlab("Wine Quality") +
    ylab("Volatile Acidity")
```



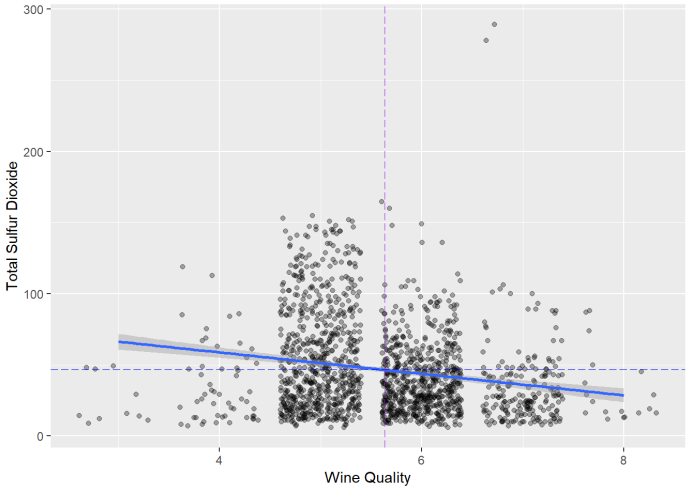
#Citric Acid Vs. Quality:

```
ggplot(data=winequalityred, aes(x=as.numeric(quality), y=citricacid)) +
    geom_jitter(alpha=1/3) +
    geom_smooth(method='lm', aes(group = 1))+
    geom_hline(yintercept=citricAverage, linetype='longdash', alpha=.5, color='blue') +
    geom_vline(xintercept = QualityAverage, linetype='longdash', color='purple', alpha=.5) +
    xlab("Wine Quality") +
    ylab("Citric Acid")
```



#Total Sulfur Vs. Quality:

```
ggplot(data=winequalityred, aes(x=as.numeric(quality), y=totalsulfurdioxide)) +
    geom_jitter(alpha=1/3) +
    geom_smooth(method='lm', aes(group = 1))+
    geom_hline(yintercept=totalsulfurdioxideAverage, linetype='longdash', alpha=.5, color='blue') +
    geom_vline(xintercept = QualityAverage, linetype='longdash', color='purple', alpha=.5) +
    xlab("Wine Quality") +
    ylab("Total Sulfur Dioxide")
```



#Density Vs. Quality:

```
ggplot(data=winequalityred, aes(x=as.numeric(quality), y=density)) +
    geom_jitter(alpha=1/3) +
    geom_smooth(method='lm', aes(group = 1))+
    geom_hline(yintercept=densityAverage, linetype='longdash', alpha=.5, color='blue') +
    geom_vline(xintercept = QualityAverage, linetype='longdash', color='purple', alpha=.5) +
    xlab("Wine Quality") +
    ylab("Density")
```