wine quality initial analysis

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Import to R the following file:

winequality <- read.csv(file = "http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv", header = TRUE, sep = ";")

Check data characteristics. Is there missing data?

str(winequality)

## 'data.frame': 4898 obs. of 12 variables:  
## $ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...  
## $ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...  
## $ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...  
## $ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...  
## $ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...  
## $ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...  
## $ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...  
## $ density : num 1.001 0.994 0.995 0.996 0.996 ...  
## $ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...  
## $ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...  
## $ alcohol : num 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...  
## $ quality : int 6 6 6 6 6 6 6 6 6 6 ...

summary(winequality)

## fixed.acidity volatile.acidity citric.acid residual.sugar   
## Min. : 3.800 Min. :0.0800 Min. :0.0000 Min. : 0.600   
## 1st Qu.: 6.300 1st Qu.:0.2100 1st Qu.:0.2700 1st Qu.: 1.700   
## Median : 6.800 Median :0.2600 Median :0.3200 Median : 5.200   
## Mean : 6.855 Mean :0.2782 Mean :0.3342 Mean : 6.391   
## 3rd Qu.: 7.300 3rd Qu.:0.3200 3rd Qu.:0.3900 3rd Qu.: 9.900   
## Max. :14.200 Max. :1.1000 Max. :1.6600 Max. :65.800   
## chlorides free.sulfur.dioxide total.sulfur.dioxide  
## Min. :0.00900 Min. : 2.00 Min. : 9.0   
## 1st Qu.:0.03600 1st Qu.: 23.00 1st Qu.:108.0   
## Median :0.04300 Median : 34.00 Median :134.0   
## Mean :0.04577 Mean : 35.31 Mean :138.4   
## 3rd Qu.:0.05000 3rd Qu.: 46.00 3rd Qu.:167.0   
## Max. :0.34600 Max. :289.00 Max. :440.0   
## density pH sulphates alcohol   
## Min. :0.9871 Min. :2.720 Min. :0.2200 Min. : 8.00   
## 1st Qu.:0.9917 1st Qu.:3.090 1st Qu.:0.4100 1st Qu.: 9.50   
## Median :0.9937 Median :3.180 Median :0.4700 Median :10.40   
## Mean :0.9940 Mean :3.188 Mean :0.4898 Mean :10.51   
## 3rd Qu.:0.9961 3rd Qu.:3.280 3rd Qu.:0.5500 3rd Qu.:11.40   
## Max. :1.0390 Max. :3.820 Max. :1.0800 Max. :14.20   
## quality   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.878   
## 3rd Qu.:6.000   
## Max. :9.000

sum(is.na(winequality) == TRUE)

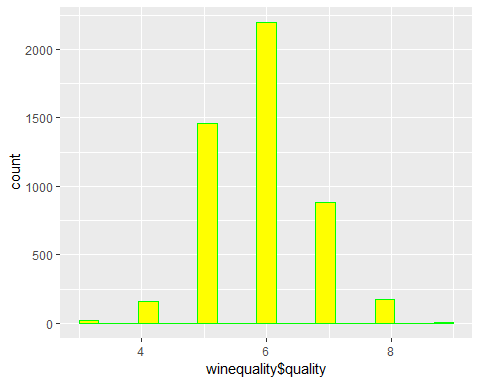
## [1] 0

There are 4898 observation and 12 variables.

There is no missing data

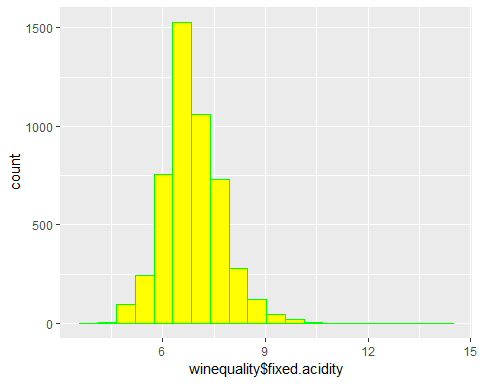
Visualize quality

library(ggplot2)  
ggplot(data = winequality,aes(x = winequality$quality)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")



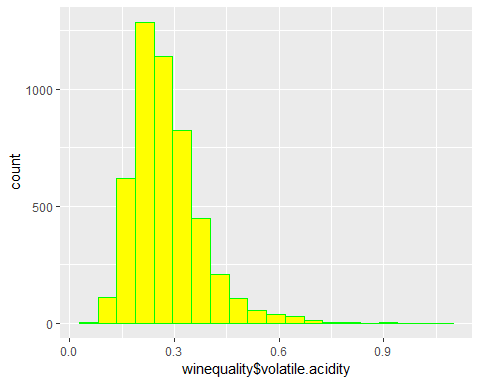
fixed.acidity

ggplot(data = winequality,aes(x = winequality$fixed.acidity)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")



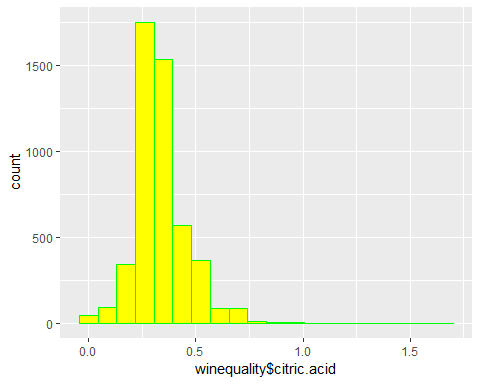
volatile.acidity

ggplot(data = winequality,aes(x = winequality$volatile.acidity)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")



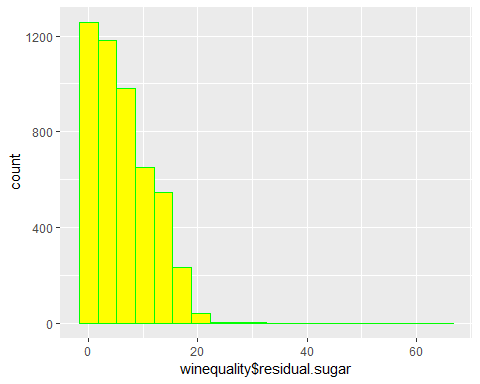
citric.acid

ggplot(data = winequality,aes(x = winequality$citric.acid)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")

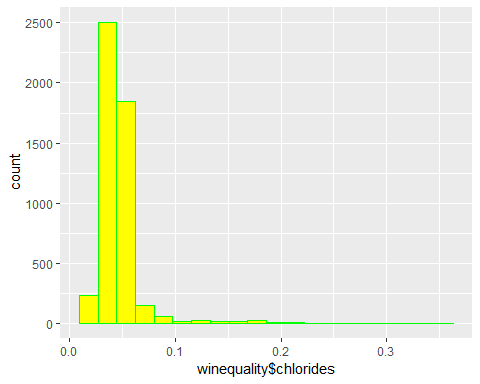


residual.sugar

ggplot(data = winequality,aes(x = winequality$residual.sugar)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")

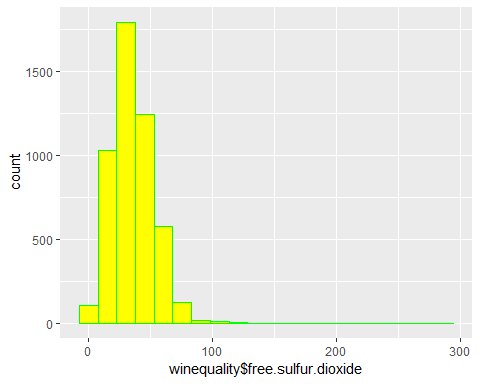
 chlorides

ggplot(data = winequality,aes(x = winequality$chlorides)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")

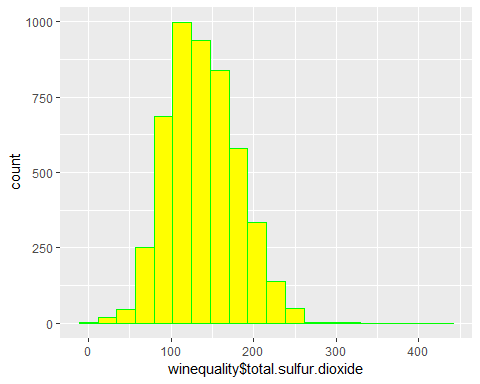


free.sulfur.dioxide

ggplot(data = winequality,aes(x = winequality$free.sulfur.dioxide)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")

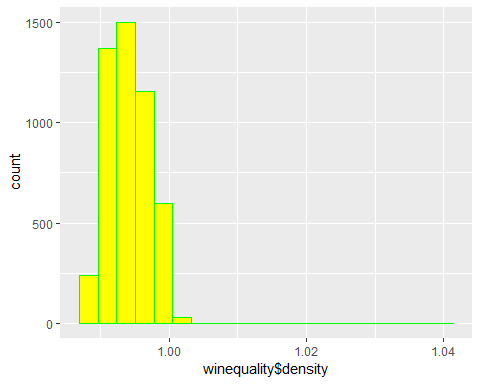
 total.sulfur.dioxide

ggplot(data = winequality,aes(x = winequality$total.sulfur.dioxide)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")



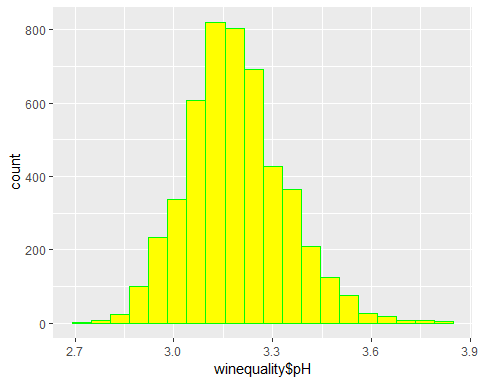
density

ggplot(data = winequality,aes(x = winequality$density)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")

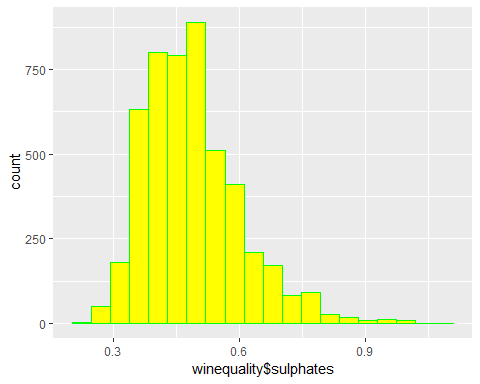


pH

ggplot(data = winequality,aes(x = winequality$pH)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")

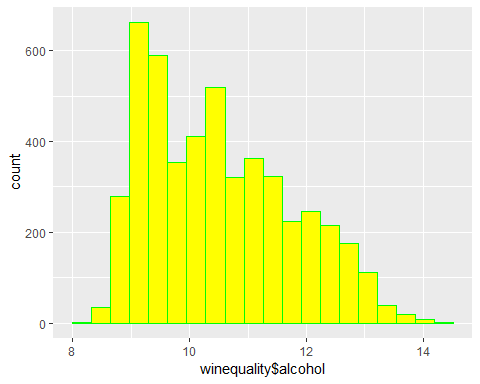
 sulphates

ggplot(data = winequality,aes(x = winequality$sulphates)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")



alcohol

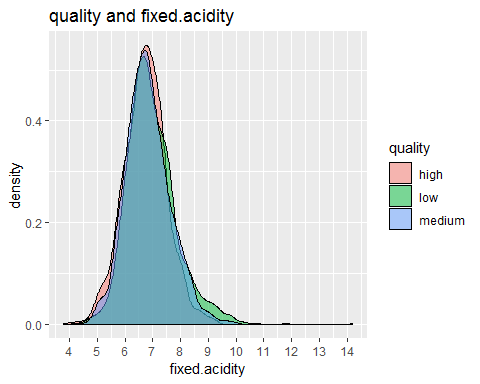
ggplot(data = winequality,aes(x = winequality$alcohol)) + geom\_histogram(bins = 20, fill = "yellow", col = "green")

 Reduce the levels of rating for quality to three levels as high, medium and low

winequality$quality[which(winequality$quality %in% c(3,4,5))] = "low"  
winequality$quality[which(winequality$quality %in% c(6))] = "medium"  
winequality$quality[which(winequality$quality %in% c(7,8,9))] = "high"  
winequality$quality <- as.factor(winequality$quality)

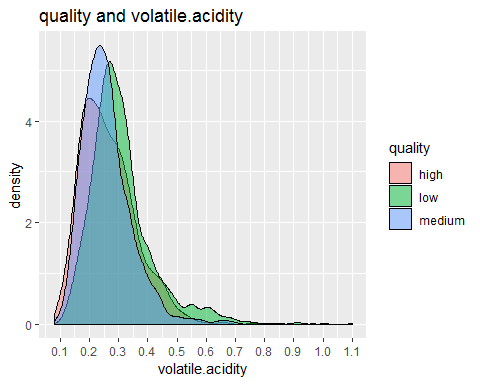
fixed.acidity Vs quality

ggplot(winequality, aes(x = fixed.acidity, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and fixed.acidity") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

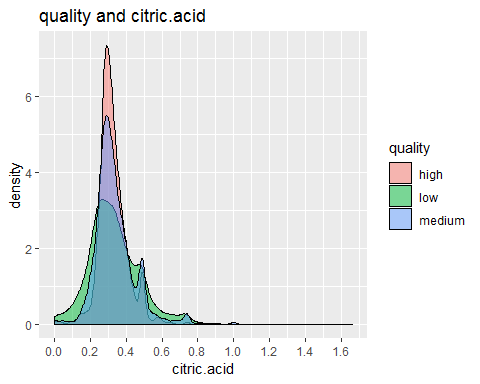


volatile.acidity Vs quality

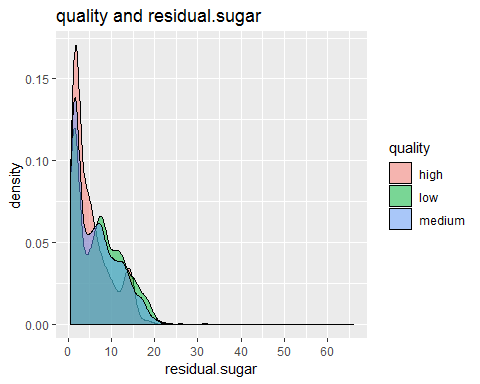
ggplot(winequality, aes(x = volatile.acidity, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and volatile.acidity") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

 This plot helps us note that high quality wines have less amount of volatile acidity. citric.acid Vs quality

ggplot(winequality, aes(x = citric.acid, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and citric.acid") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

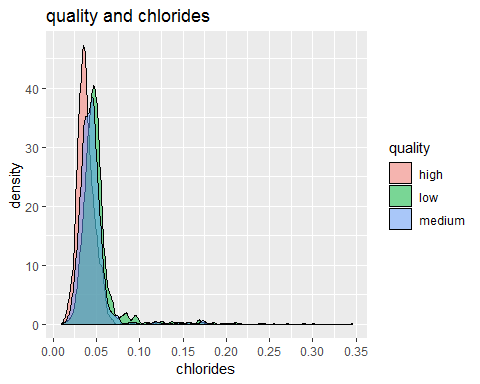
 Again, per our observation we note that higher quality wines have higher levels of citric acid residual.sugar Vs quality

ggplot(winequality, aes(x = residual.sugar, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and residual.sugar") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()



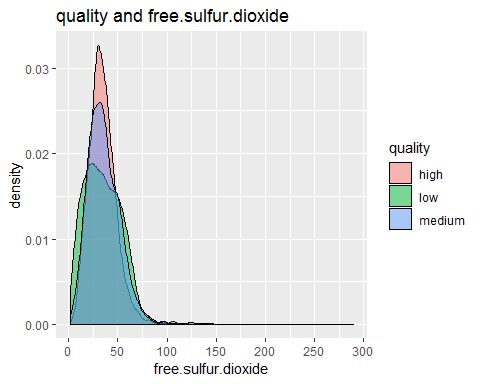
chlorides Vs quality

ggplot(winequality, aes(x = chlorides, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and chlorides") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()



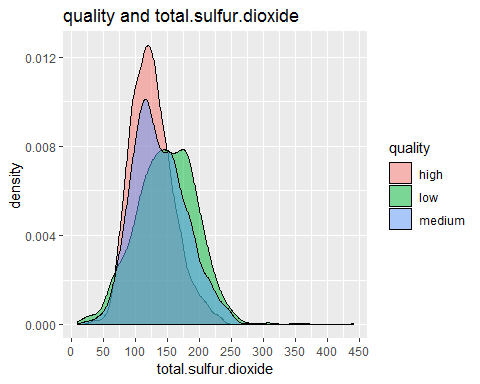
free.sulfur.dioxide Vs quality

ggplot(winequality, aes(x = free.sulfur.dioxide, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and free.sulfur.dioxide") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()



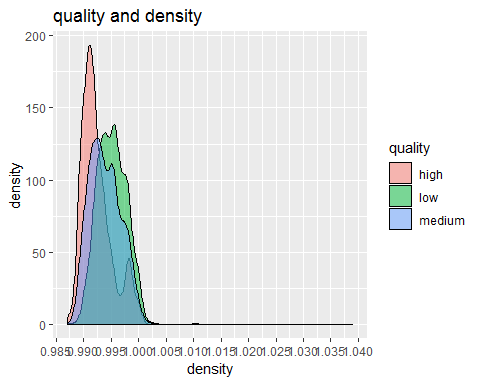
total.sulfur.dioxide Vs quality

ggplot(winequality, aes(x = total.sulfur.dioxide, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and total.sulfur.dioxide") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

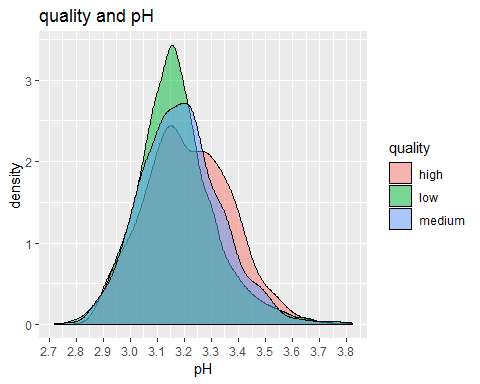


density Vs quality

ggplot(winequality, aes(x = density, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and density") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

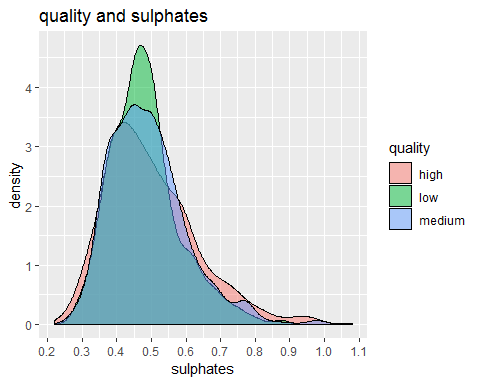
 We can see that best quality wines have lowest density pH Vs quality

ggplot(winequality, aes(x = pH, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and pH") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()



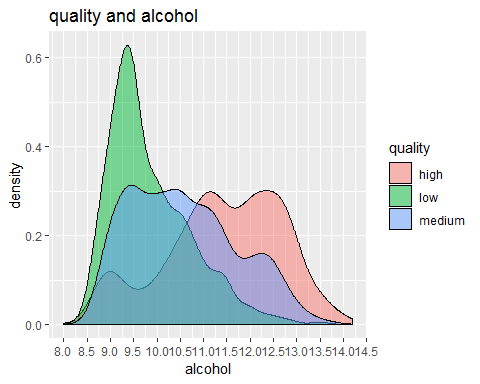
sulphates Vs quality

ggplot(winequality, aes(x = sulphates, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and sulphates") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()



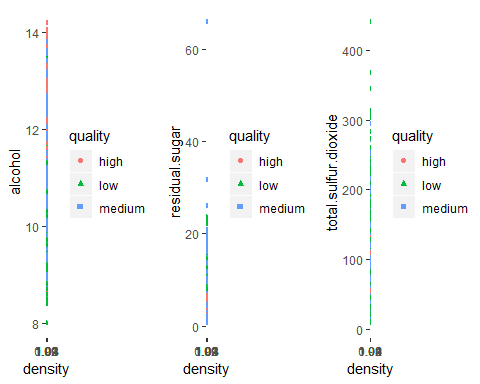
alcohol Vs quality

ggplot(winequality, aes(x = alcohol, fill = quality)) +  
geom\_density(alpha=0.5, aes(fill=factor(quality))) + labs(title="quality and alcohol") +  
scale\_x\_continuous(breaks = scales::pretty\_breaks(n = 10)) + theme\_grey()

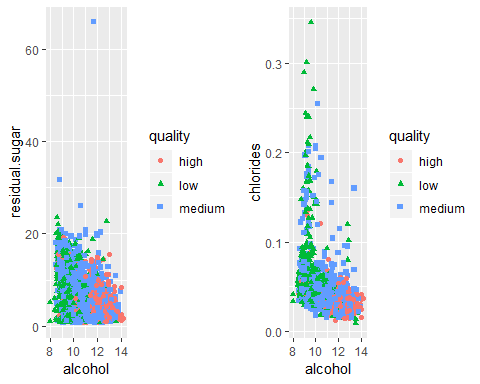
 As per our observation, good quality wines have higher levels of alcohol.

Within the same range of density, best quality wines have highest level of alcohol.

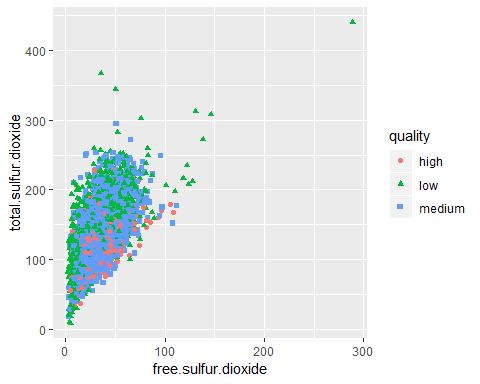
#install.packages("gridExtra")  
library(gridExtra)  
f1 <- ggplot(winequality, aes(x = density, y = alcohol, col = quality, shape = quality)) + geom\_point()  
f2 <- ggplot(winequality, aes(x = density, y = residual.sugar, col = quality, shape = quality)) + geom\_point()  
f3 <- ggplot(winequality, aes(x = density, y = total.sulfur.dioxide, col = quality, shape = quality)) + geom\_point()  
grid.arrange(f1, f2, f3, ncol=3)

 Within the same range of density, best quality wines have highest level of alcohol.In the above 3 graphs, holding density (mostly from the lower end of density), higher residual.sugar or alcohol, or total sulfur dioxide seem to have better quality respectively.

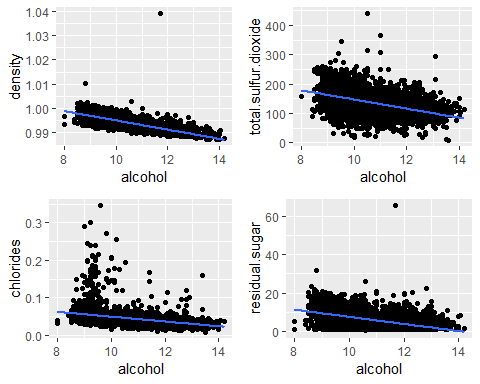
e1 <- ggplot(winequality, aes(x = alcohol, y = residual.sugar, col = quality, shape = quality)) + geom\_point()  
e2 <- ggplot(winequality, aes(x = alcohol, y = chlorides, col = quality, shape = quality)) + geom\_point()  
grid.arrange(e1, e2, ncol=2)

 In the above 2 graphs, holding residual.sugar or chlorides, higher alcohol level seem to have better quality.

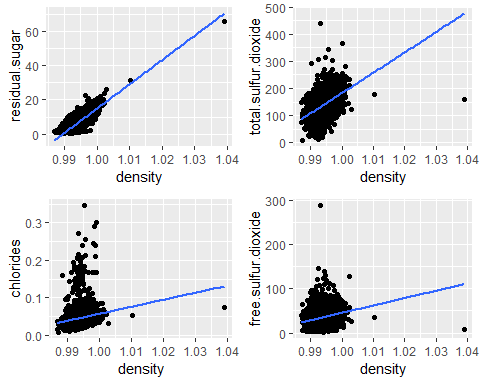
ggplot(winequality, aes(x = free.sulfur.dioxide, y = total.sulfur.dioxide, col = quality, shape = quality)) + geom\_point()

 it seems that there are more better quality wine under 150 total sulfur dioxide.

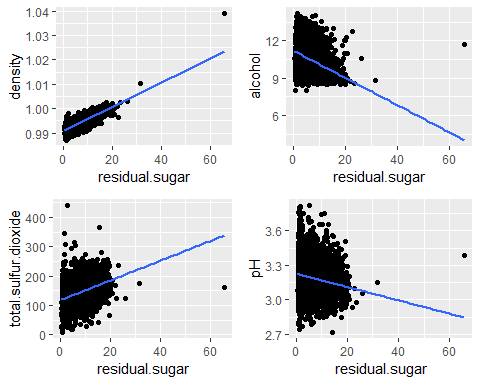
d1 <- ggplot(data = winequality, aes(y = density, x = alcohol))+geom\_point()+geom\_smooth(method = "lm", se = F)  
d2 <- ggplot(data = winequality, aes(y = total.sulfur.dioxide, x = alcohol))+geom\_point()+geom\_smooth(method = "lm", se = F)  
d3 <- ggplot(data = winequality, aes(y = chlorides, x = alcohol))+geom\_point()+geom\_smooth(method = "lm", se = F)  
d4 <- ggplot(data = winequality, aes(y = residual.sugar, x = alcohol))+geom\_point()+geom\_smooth(method = "lm", se = F)  
grid.arrange(d1, d2, d3, d4, ncol=2, nrow = 2)

 Alcohol is strongly correlated to density, and weakly correlated to residual sugar, total sulfur dioxide, and chlorides.

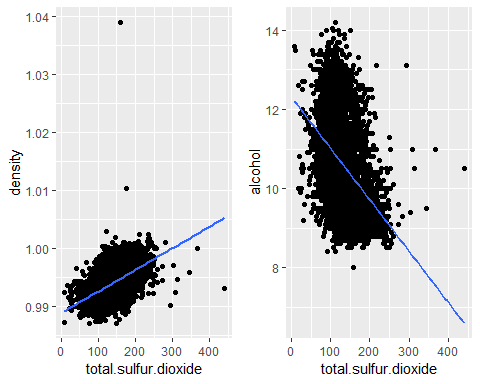
c1 <- ggplot(data = winequality, aes(y = residual.sugar, x = density))+geom\_point()+geom\_smooth(method = "lm", se = F)  
c2 <- ggplot(data = winequality, aes(y = total.sulfur.dioxide, x = density))+geom\_point()+geom\_smooth(method = "lm", se = F)  
c3 <- ggplot(data = winequality, aes(y = chlorides, x = density))+geom\_point()+geom\_smooth(method = "lm", se = F)  
c4 <- ggplot(data = winequality, aes(y = free.sulfur.dioxide, x = density))+geom\_point()+geom\_smooth(method = "lm", se = F)  
grid.arrange(c1, c2, c3, c4, ncol=2, nrow = 2)

 Density is strongly correlated to residual sugar, and weakly correlated to chlorides, total sulfur dioxide, free sulfur dioxide.

b1 <- ggplot(data = winequality, aes(y = density, x = residual.sugar))+geom\_point()+geom\_smooth(method = "lm", se = F)  
b2 <- ggplot(data = winequality, aes(y = alcohol, x = residual.sugar))+geom\_point()+geom\_smooth(method = "lm", se = F)  
b3 <- ggplot(data = winequality, aes(y = total.sulfur.dioxide, x = residual.sugar))+geom\_point()+geom\_smooth(method = "lm", se = F)  
b4 <- ggplot(data = winequality, aes(y = pH, x = residual.sugar))+geom\_point()+geom\_smooth(method = "lm", se = F)  
grid.arrange(b1, b2, b3, b4, ncol=2, nrow = 2)

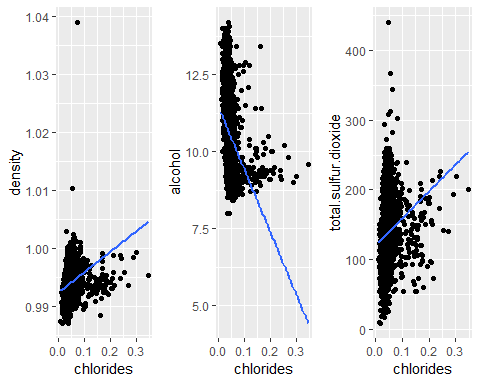
 Residual sugar level is strongly positively correlated to density, weakly negatively correlated to alcohol and positively correlated to total sulfur dioxide. Residual sugar is not correlated to pH.

z1 <- ggplot(data = winequality, aes(y = density, x = total.sulfur.dioxide))+geom\_point()+geom\_smooth(method = "lm", se = F)  
z2 <- ggplot(data = winequality, aes(y = alcohol, x = total.sulfur.dioxide))+geom\_point()+geom\_smooth(method = "lm", se = F)  
grid.arrange(z1, z2, ncol=2)



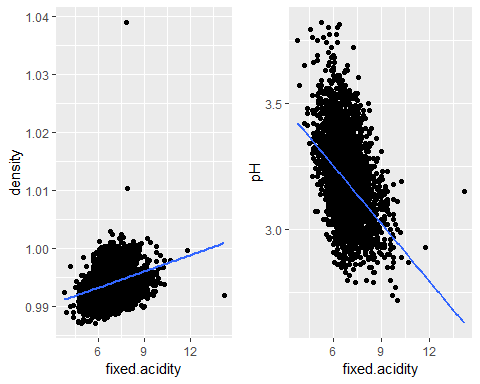
Total sulfur dioxide is strongly correlated to alcohol.

a1 <- ggplot(data = winequality, aes(y = density, x = chlorides))+geom\_point()+geom\_smooth(method = "lm", se = F)  
a2 <- ggplot(data = winequality, aes(y = alcohol, x = chlorides))+geom\_point()+geom\_smooth(method = "lm", se = F)  
a3 <-ggplot(data = winequality, aes(y = total.sulfur.dioxide, x = chlorides))+geom\_point()+geom\_smooth(method = "lm", se = F)  
grid.arrange(a1, a2, a3, ncol=3)



Chlorides is positively correlated to density although not a strong relationship. Chlorides is slightly correlated to total.sulfur.dioxide.

p1 <- ggplot(data = winequality, aes(y = density, x = fixed.acidity))+geom\_point()+geom\_smooth(method = "lm", se = F)  
p2 <- ggplot(data = winequality, aes(y = pH, x = fixed.acidity))+geom\_point()+geom\_smooth(method = "lm", se = F)  
grid.arrange(p1, p2, ncol=2)

 Fixed acidity is negatively correlated to pH and positively correlated to density weakly.