RedWine

Laura Larregui

12/10/2019

# Load and view the variables in data  
readURL <- function(inputURL) #Begin function named readURL that takes a URL  
{  
 csvFile <- read.csv(url(inputURL), sep = ';') #assign the results of the URL call as a csv file to a dataframe named csvFile. Added sep = ';' to seperate the data into columns  
 return(csvFile) # return the dataframe  
}  
  
data <- readURL("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv")  
  
  
# Determine whether there are any 'NA' values in the dataset  
  
data <- na.omit(data)  
  
# The resulting dataframe is same size, so there are no NA values  
  
  
  
# Now transform quality into an ordinal variable   
  
data$quality.cat <- ordered(data$quality)  
  
str(data$quality.cat)

## Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...

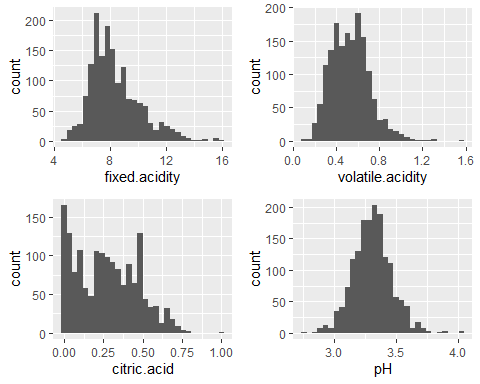
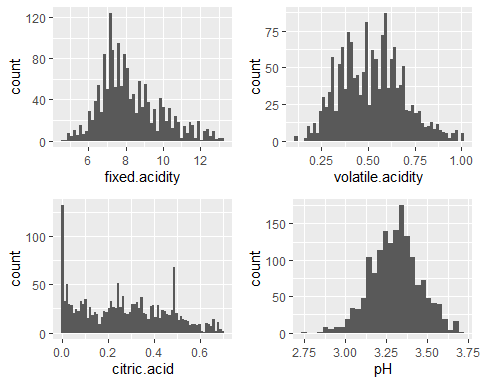
# Function that calculates the maximum / median for any variable  
  
maxmedianratio = function(x)max(x)/median(x)  
  
  
  
# Apply the function to the dataset  
  
apply(subset(data,select = -c(quality.cat)),2,maxmedianratio)

## fixed.acidity volatile.acidity citric.acid   
## 2.012658 3.038462 3.846154   
## residual.sugar chlorides free.sulfur.dioxide   
## 7.045455 7.734177 5.142857   
## total.sulfur.dioxide density pH   
## 7.605263 1.006963 1.211480   
## sulphates alcohol quality   
## 3.225806 1.460784 1.333333

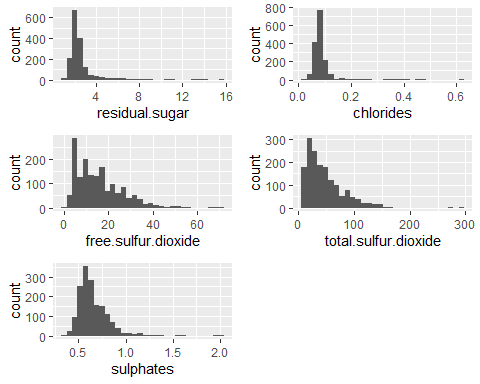
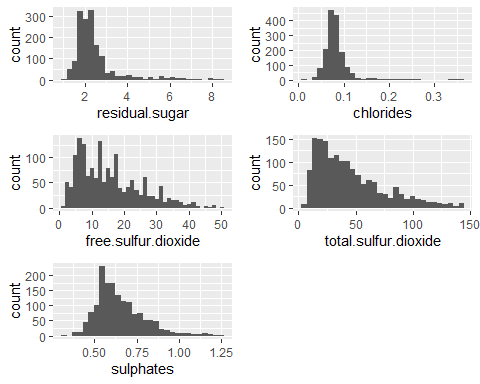
##There is a fair amount of variance within variables.Density has the lowest ratio.

#Univariate Plots and Analysis Section ##Rather than simply output 12 histograms, I will group the 12 properties into 3 different categories, and look at each category in turn. Since pH is a measure of acidity, I will group pH together with the graphs showing the 3 acid levels (fixed.acidity, volatile.acidity, and citric.acid). Next, I will group together the 5 remaining concentration measurements (residual.sugar, chlorides,free.sulfur.dioxide, total.sulfur.dioxide, and sulphates). Finally, I will group together alcohol, density and quality.

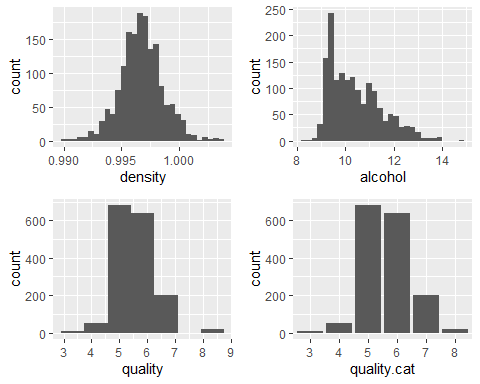
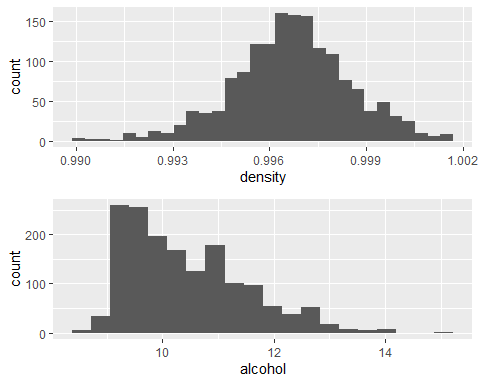
### “Acidity” Related Histograms:

 ## These four parameters all look reasonably normally distributed. In all four cases, there is some positive skewing, as can be judged by the long extension on the right hand side of the graph, with very low ‘count’ values for the higher x-axis values. As I get deeper into the analysis, it might make sense to exclude the upper most quantile (e.g. 1%) of each of these parameters, to remove this skewing, which appears to impact only a small number of wines (as judged by the very small count values).  ##Once the top 1% of each parameter is excluded, it is easier to see the shape of the bulk of the data. All four parameters appear to be approximately normally distributed. There are two interesting ‘spikes’ in the citric acid profile, one near the median and a second smaller one near a value of 0.5.

# Other Concentration Related Histograms:

 ##As was seen with the four “acid” related parameters, the five graphs above also exhibit positive skew.  ##Once the top 1% of each parameter is excluded, it is easier to see the shape of the bulk of the data. Most parameters appear to be approximately normally distributed here, with the exception of residual.sugar.

# “Other” Variables Histograms

###(Note: a bar chart is used in the case of ‘quality.cat’, since it is categorical):  ## The quality rating appears to be normally distributed, with the bulk of assessments in the middle bins. Density appears normal too, but with some positive skew. The alcohol content looks interesting.  ##Density looks fairly normally distributed, whereas alcohol content does not.

# Create New Variables:

### Reference:<http://beerandwinejournal.com/chloride-and-sulfate/>

#### The chlorides to sulphates ratio might be a far more important measure of quality than the individual levels of either ion. Thus, I will create a chlorides-to-sulphate ratio variable.

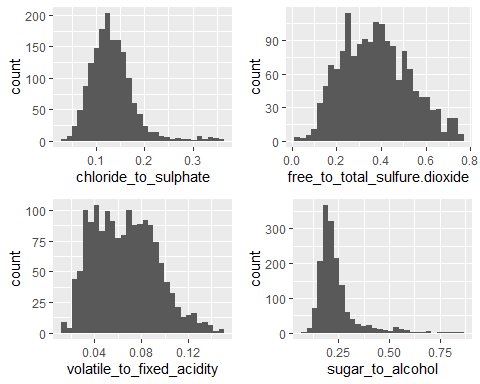
# Create and add four new variables to the dataframe:  
data$chloride\_to\_sulphate <-with(data,chlorides / sulphates)  
data$free\_to\_total\_sulfure.dioxide <-with(data,free.sulfur.dioxide / total.sulfur.dioxide)  
data$volatile\_to\_fixed\_acidity <-with(data,volatile.acidity / fixed.acidity)  
data$sugar\_to\_alcohol <-with(data,residual.sugar / alcohol)  
# Output summary data on the new variables:  
str(subset(data,select = c(chloride\_to\_sulphate,free\_to\_total\_sulfure.dioxide,volatile\_to\_fixed\_acidity,sugar\_to\_alcohol)))

## 'data.frame': 1599 obs. of 4 variables:  
## $ chloride\_to\_sulphate : num 0.136 0.144 0.142 0.129 0.136 ...  
## $ free\_to\_total\_sulfure.dioxide: num 0.324 0.373 0.278 0.283 0.324 ...  
## $ volatile\_to\_fixed\_acidity : num 0.0946 0.1128 0.0974 0.025 0.0946 ...  
## $ sugar\_to\_alcohol : num 0.202 0.265 0.235 0.194 0.202 ...

summary(subset(data,select = c(chloride\_to\_sulphate,free\_to\_total\_sulfure.dioxide,volatile\_to\_fixed\_acidity,sugar\_to\_alcohol)))

## chloride\_to\_sulphate free\_to\_total\_sulfure.dioxide  
## Min. :0.03077 Min. :0.02273   
## 1st Qu.:0.10455 1st Qu.:0.25926   
## Median :0.12833 Median :0.37500   
## Mean :0.13572 Mean :0.38231   
## 3rd Qu.:0.15581 3rd Qu.:0.48485   
## Max. :0.57761 Max. :0.85714   
## volatile\_to\_fixed\_acidity sugar\_to\_alcohol   
## Min. :0.01348 Min. :0.07087   
## 1st Qu.:0.04405 1st Qu.:0.18306   
## Median :0.06569 Median :0.21111   
## Mean :0.06706 Mean :0.24550   
## 3rd Qu.:0.08581 3rd Qu.:0.25481   
## Max. :0.20800 Max. :1.71111

# Plot the new parameters as a group:

 ## The free:total sulfur dioxide graph looks normally distributed. The chloride:sulphate, volatile:fixed acidity and sugar:alcohol graphs look positively skewed.

# Looking at the relationship between the various parameters.

# Bivariate Plots and Analysis Section

## Correlation

## fixed.acidity volatile.acidity citric.acid  
## fixed.acidity 1.00000000 -0.256130895 0.671703435  
## volatile.acidity -0.25613089 1.000000000 -0.552495685  
## citric.acid 0.67170343 -0.552495685 1.000000000  
## residual.sugar 0.11477672 0.001917882 0.143577162  
## chlorides 0.09370519 0.061297772 0.203822914  
## free.sulfur.dioxide -0.15379419 -0.010503827 -0.060978129  
## total.sulfur.dioxide -0.11318144 0.076470005 0.035533024  
## density 0.66804729 0.022026232 0.364947175  
## pH -0.68297819 0.234937294 -0.541904145  
## sulphates 0.18300566 -0.260986685 0.312770044  
## alcohol -0.06166827 -0.202288027 0.109903247  
## quality 0.12405165 -0.390557780 0.226372514  
## chloride\_to\_sulphate 0.01226842 0.249628439 0.001380406  
## free\_to\_total\_sulfure.dioxide -0.13081236 -0.072618561 -0.166938888  
## volatile\_to\_fixed\_acidity -0.60921527 0.896837551 -0.715165969  
## sugar\_to\_alcohol 0.11300710 0.037540873 0.110316549  
## residual.sugar chlorides  
## fixed.acidity 0.114776724 0.093705186  
## volatile.acidity 0.001917882 0.061297772  
## citric.acid 0.143577162 0.203822914  
## residual.sugar 1.000000000 0.055609535  
## chlorides 0.055609535 1.000000000  
## free.sulfur.dioxide 0.187048995 0.005562147  
## total.sulfur.dioxide 0.203027882 0.047400468  
## density 0.355283371 0.200632327  
## pH -0.085652422 -0.265026131  
## sulphates 0.005527121 0.371260481  
## alcohol 0.042075437 -0.221140545  
## quality 0.013731637 -0.128906560  
## chloride\_to\_sulphate 0.074675759 0.778543362  
## free\_to\_total\_sulfure.dioxide -0.070626080 -0.105156413  
## volatile\_to\_fixed\_acidity -0.049511157 -0.010464756  
## sugar\_to\_alcohol 0.978845044 0.094452464  
## free.sulfur.dioxide total.sulfur.dioxide  
## fixed.acidity -0.153794193 -0.11318144  
## volatile.acidity -0.010503827 0.07647000  
## citric.acid -0.060978129 0.03553302  
## residual.sugar 0.187048995 0.20302788  
## chlorides 0.005562147 0.04740047  
## free.sulfur.dioxide 1.000000000 0.66766645  
## total.sulfur.dioxide 0.667666450 1.00000000  
## density -0.021945831 0.07126948  
## pH 0.070377499 -0.06649456  
## sulphates 0.051657572 0.04294684  
## alcohol -0.069408354 -0.20565394  
## quality -0.050656057 -0.18510029  
## chloride\_to\_sulphate -0.025987885 0.03752217  
## free\_to\_total\_sulfure.dioxide 0.327240869 -0.37143493  
## volatile\_to\_fixed\_acidity 0.039823346 0.07699049  
## sugar\_to\_alcohol 0.213431359 0.23705590  
## density pH sulphates  
## fixed.acidity 0.66804729 -0.68297819 0.183005664  
## volatile.acidity 0.02202623 0.23493729 -0.260986685  
## citric.acid 0.36494718 -0.54190414 0.312770044  
## residual.sugar 0.35528337 -0.08565242 0.005527121  
## chlorides 0.20063233 -0.26502613 0.371260481  
## free.sulfur.dioxide -0.02194583 0.07037750 0.051657572  
## total.sulfur.dioxide 0.07126948 -0.06649456 0.042946836  
## density 1.00000000 -0.34169933 0.148506412  
## pH -0.34169933 1.00000000 -0.196647602  
## sulphates 0.14850641 -0.19664760 1.000000000  
## alcohol -0.49617977 0.20563251 0.093594750  
## quality -0.17491923 -0.05773139 0.251397079  
## chloride\_to\_sulphate 0.16924175 -0.17646945 -0.204935858  
## free\_to\_total\_sulfure.dioxide -0.26497991 0.18489507 -0.010459139  
## volatile\_to\_fixed\_acidity -0.27912029 0.51305534 -0.274887351  
## sugar\_to\_alcohol 0.42710842 -0.11019992 -0.009753611  
## alcohol quality chloride\_to\_sulphate  
## fixed.acidity -0.06166827 0.12405165 0.012268424  
## volatile.acidity -0.20228803 -0.39055778 0.249628439  
## citric.acid 0.10990325 0.22637251 0.001380406  
## residual.sugar 0.04207544 0.01373164 0.074675759  
## chlorides -0.22114054 -0.12890656 0.778543362  
## free.sulfur.dioxide -0.06940835 -0.05065606 -0.025987885  
## total.sulfur.dioxide -0.20565394 -0.18510029 0.037522169  
## density -0.49617977 -0.17491923 0.169241747  
## pH 0.20563251 -0.05773139 -0.176469449  
## sulphates 0.09359475 0.25139708 -0.204935858  
## alcohol 1.00000000 0.47616632 -0.318336914  
## quality 0.47616632 1.00000000 -0.308693076  
## chloride\_to\_sulphate -0.31833691 -0.30869308 1.000000000  
## free\_to\_total\_sulfure.dioxide 0.24627450 0.19411335 -0.124891371  
## volatile\_to\_fixed\_acidity -0.08778639 -0.34390466 0.168469523  
## sugar\_to\_alcohol -0.13224271 -0.06537353 0.129780930  
## free\_to\_total\_sulfure.dioxide  
## fixed.acidity -0.13081236  
## volatile.acidity -0.07261856  
## citric.acid -0.16693889  
## residual.sugar -0.07062608  
## chlorides -0.10515641  
## free.sulfur.dioxide 0.32724087  
## total.sulfur.dioxide -0.37143493  
## density -0.26497991  
## pH 0.18489507  
## sulphates -0.01045914  
## alcohol 0.24627450  
## quality 0.19411335  
## chloride\_to\_sulphate -0.12489137  
## free\_to\_total\_sulfure.dioxide 1.00000000  
## volatile\_to\_fixed\_acidity 0.01727629  
## sugar\_to\_alcohol -0.10047853  
## volatile\_to\_fixed\_acidity sugar\_to\_alcohol  
## fixed.acidity -0.60921527 0.113007097  
## volatile.acidity 0.89683755 0.037540873  
## citric.acid -0.71516597 0.110316549  
## residual.sugar -0.04951116 0.978845044  
## chlorides -0.01046476 0.094452464  
## free.sulfur.dioxide 0.03982335 0.213431359  
## total.sulfur.dioxide 0.07699049 0.237055905  
## density -0.27912029 0.427108421  
## pH 0.51305534 -0.110199917  
## sulphates -0.27488735 -0.009753611  
## alcohol -0.08778639 -0.132242714  
## quality -0.34390466 -0.065373533  
## chloride\_to\_sulphate 0.16846952 0.129780930  
## free\_to\_total\_sulfure.dioxide 0.01727629 -0.100478534  
## volatile\_to\_fixed\_acidity 1.00000000 -0.027891595  
## sugar\_to\_alcohol -0.02789160 1.000000000

##Based on the correlations, it appears several chemicals negatively impact quality (correlations are shown in parentheses below): #\* volatile.acidity (-0.39)  
#*chloride\_to\_sulphate (-0.31) #*volatile\_to\_fixed\_acidity (-0.34)

### Let’s create a new variable, ‘bad\_solids’ that adds them together

data$bad\_solids <-with(data,volatile.acidity,chloride\_to\_sulphate,volatile\_to\_fixed\_acidity)  
  
str(data$bad\_solids)

## num [1:1599] 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...

summary(data$bad\_solids)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1200 0.3900 0.5200 0.5278 0.6400 1.5800

# Determine the correlation of bad\_solids with quality

# Determine the correlation of bad\_solids with quality  
  
cor(data$bad\_solids,data$quality)

## [1] -0.3905578

# This new variable does not have a particularly strong correlation with quality. Its correlation coefficient (-0.39).

## variables with a modest correlation with quality:

###\* volatile.acidity (-0.39)

###\* chlorides (-0.12)

###\* total.sulfur.dioxide (-0.18)

###\* density (-0.17)

###\* sulphates (0.23)

###\* alcohol (0.47)

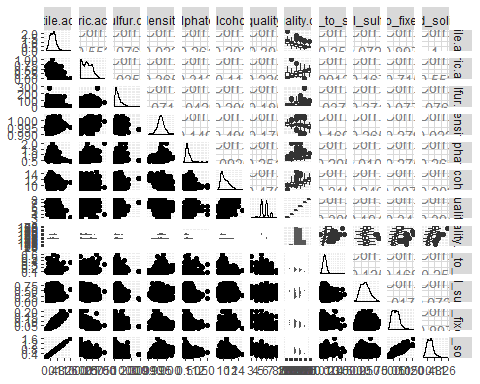
###\* chloride\_to\_sulphate (-0.30)

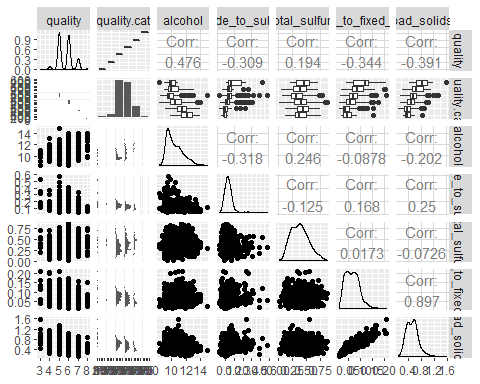
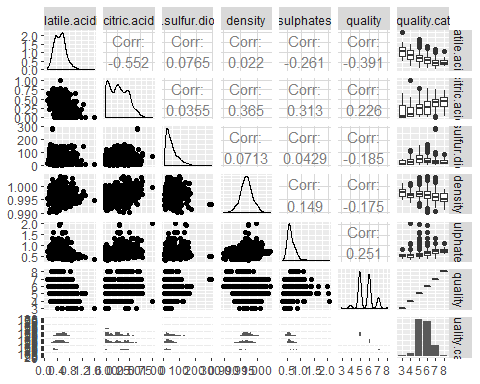
###\* free\_to\_total\_sulfure.dioxide (0.20)

###\* volatile\_to\_fixed\_acidity (-0.34)

###\* bad solids (-0.39)

# Scatterplot matrix

 # Although there appear to be some trends, the plot is too dense for any meaningful analysis, so I will split it up a bit. First, I will generate two scatterplot matrices that involve the primary feature of interest (quality):



#The most interesting observations I gleam from the scatterplots are as follows:

##\* There appears to be a meaningful, positive relationship between quality and alcohol percent (0.48 correlation).

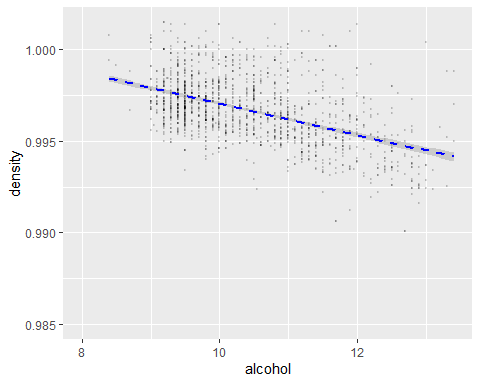
##\* There appears to be a meaningful, negative relationship between quality and volatile acidity (-0.39 correlation).

##\* There appears to be a possibly linear relationship between density and alcohol percent, with a relatively strong correlation (-0.50).

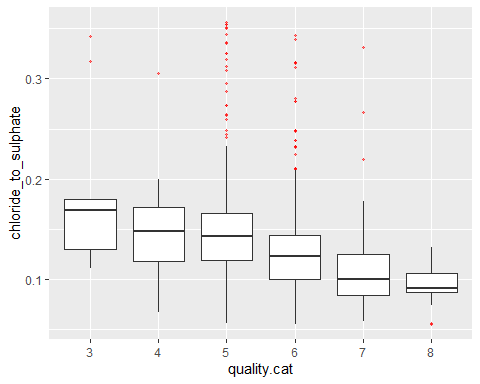
##\* Citric acid and density have an positive correlation (0.37)

# Bivariate pairs

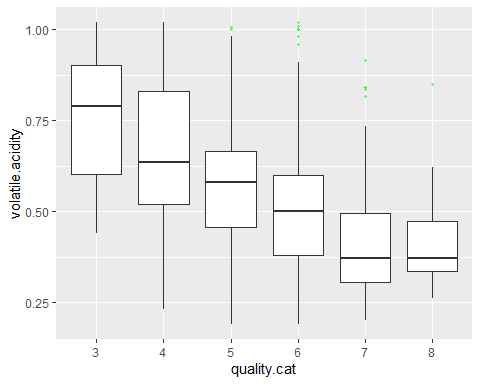
##Density and Alcohol Content:

 ##This inverse relationship is apparent from the graph, although at any given alcohol level, there is a fair amount of variability in the density value.

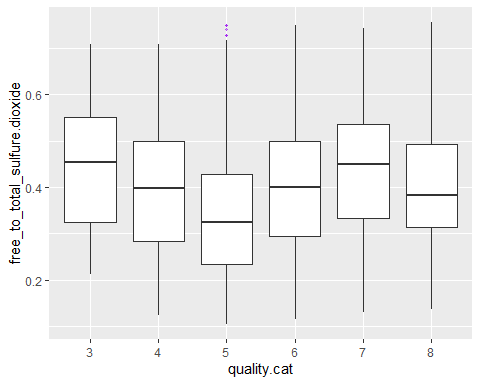
##2a. Quality and the Chloride:Sulphate Ratio

 ##It appears that higher quality wines have lower chloride:sulphate ratios

##2b. Quality and Volatile Acidity

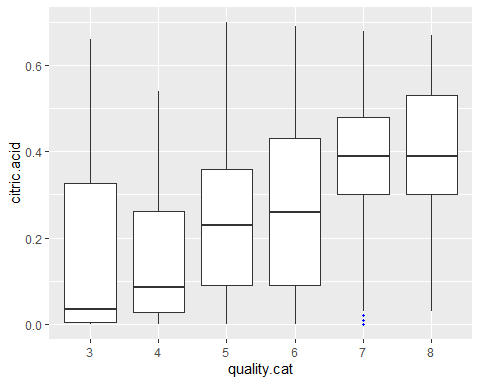
 ##It appears that higher quality wines have lower volatile acidity

##2c. Quality and the Free:Total Sulfur Dioxide Ratio

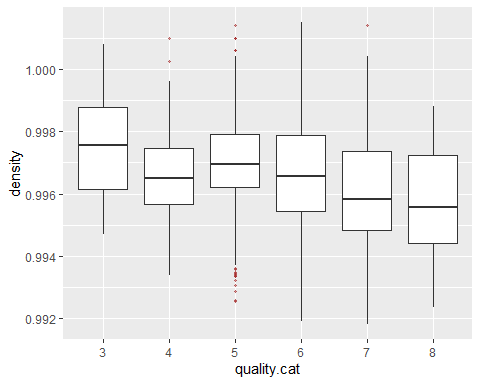


## No promising trend appears.

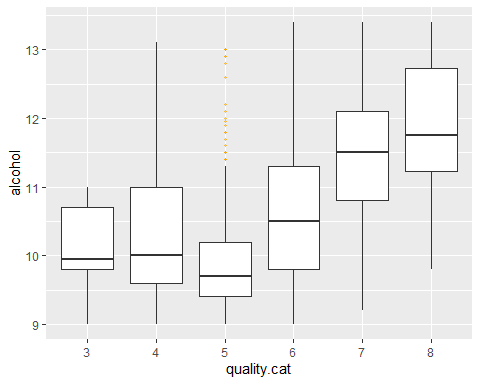
##2d. Quality and citric.acid

 ##It appears that in general, higher quality wines have higher citric acid levels.

##2e. Quality and Density

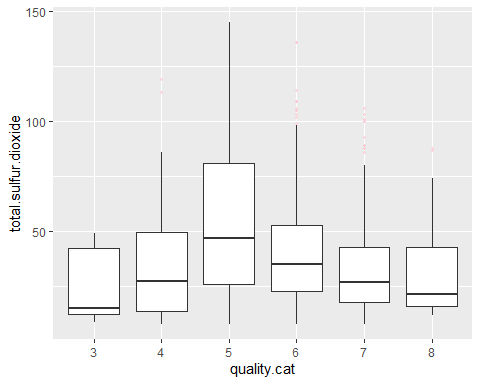
 # It is hard to discern any clear trend between the density and a wine’s quality, given that the median values move up and down as the quality improves.

##2f. Quality and Alcohol Content



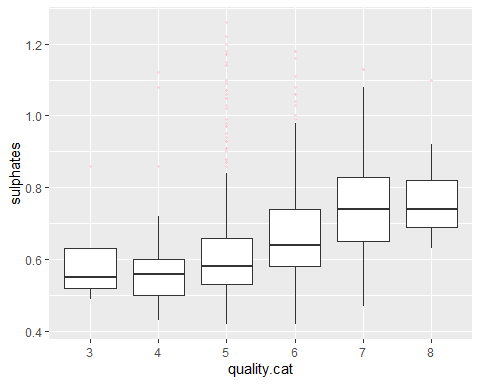
##The relationship between alcohol content and quality appears potentially promising, particularly at the higher end of the quality scale, where there is a clear upwards trend in quality (from levels 6 through 8).

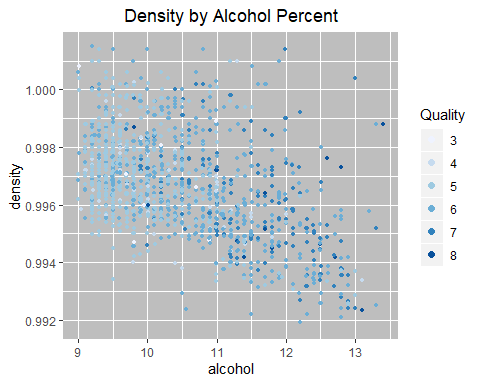
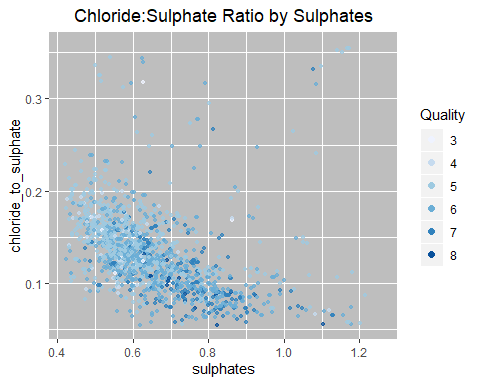
##2g. Quality and Total Sulphur Dioxide

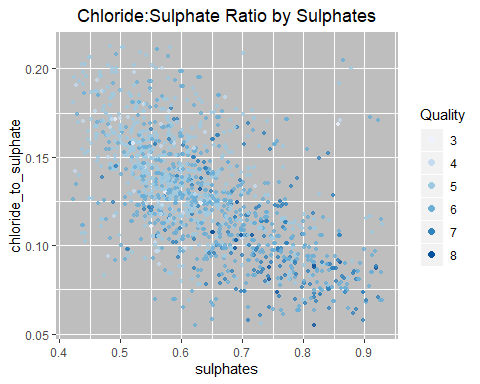


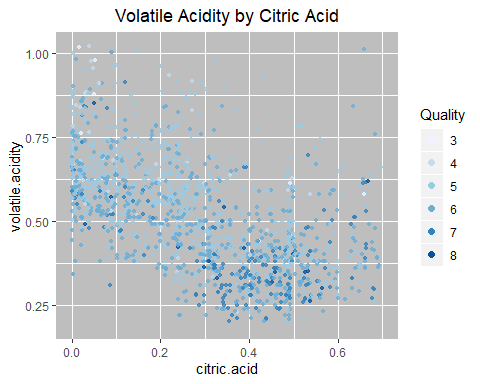
##It is hard to discern any clear trend between the total sulphus dioxide and a wine’s quality.

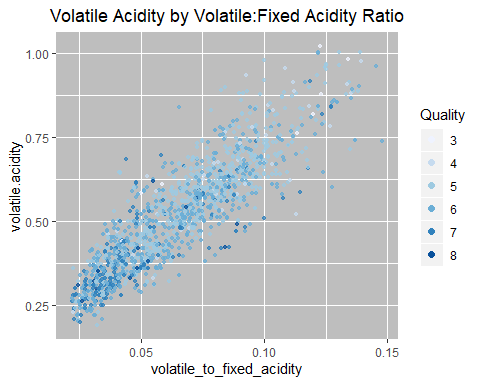
##2h. Quality and Sulphates

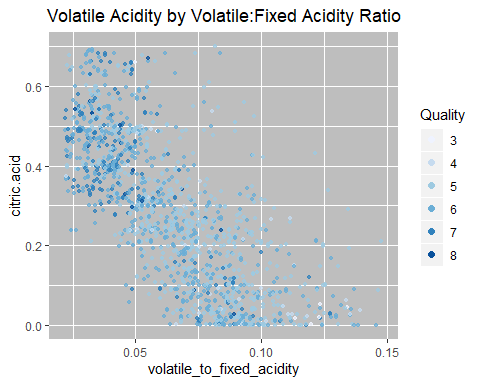
 ##It appears that in general, higher quality wines have higher sulphate levels.  
# Multivariate Plots and Analysis Section

##I will now consider the interaction of multiple variables. First, it was observed in the bivariate analysis that there is a relatively strong inverse relationship between density and the alcohol content (correlation coefficient of -0.50).   # It appears there might be a tendency for high quality wines to be high sulphate levels and low chloride:sulphate ratio. Let’s zoom in on the lower left portion of the graph, which contains most of the data points, by truncating out the top 5% quantile for each variable:

 ## There does indeed appear to be a tendency for the higher quality wines to be higher in sulphates and lower chloride:sulphate ratio, given that the quality 7-8 wines have tended to cluster in the lower right portion of the graph, whereas the quality 3-5 wines are more in the upper left portion.

 # There is no strong pattern regarding where the higher versus lower quality wines fall on the graph. The quality points are dispersed throughout, even though there might be some weak relationships in terms of where they tend to fall.

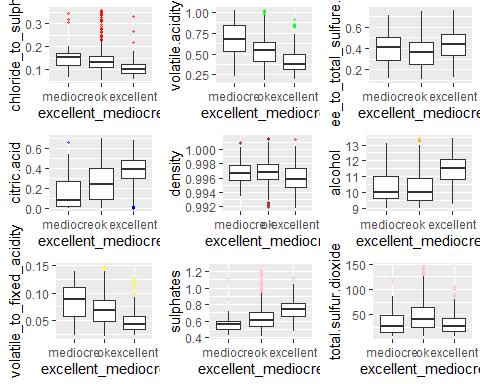
 ## There does indeed appear to be a tendency for the higher quality wines to be lower in volatile acidity and volatile:fixed acidity ratio, given that the quality 7-8 wines have tended to cluster in the lower left portion of the graph, whereas the quality 3-5 wines are more in the upper right portion.

 ## There does indeed appear to be a tendency for the higher quality wines to be higher in citric acid levels and lower volatile:fixed acidity ratio, given that the quality 7-8 wines have tended to cluster in the upper left portion of the graph, whereas the quality 3-5 wines are more in the lower right portion.

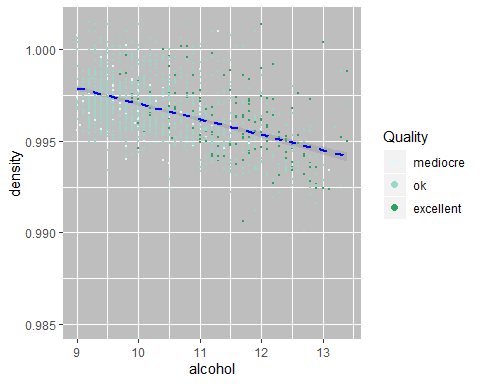
# One Final Data Transformation

##Lets consider any wine with a 3-4 rating as ‘mediocre’, a wine with a 5-6 rating as ‘ok’ and a wine with a 7-8 rating as ‘excellent’.

## mediocre ok excellent   
## 63 1319 217

 #The following variables correlate inversely with quality (i.e. quality decreases as these variables increase in value): ##\* chloride:sulphate ratio  
##\* volatile acidity  
##\* Volatile:fixed acidity  
##\* density  
#The following variables correlate with quality (i.e. quality increases as these variables increase in value): ##\* alcohol content  
##\* citric acid ##\* sulphates

# the new quality categories on the density vs. alcohol content graph:



The categories split quite well: good wines tend to have higher alcohol content and lower density levels.

### Predictive Models

## Call:  
## polr(formula = quality.cat ~ alcohol + density + sulphates +   
## citric.acid + volatile.acidity + total.sulfur.dioxide + chloride\_to\_sulphate +   
## volatile\_to\_fixed\_acidity + free\_to\_total\_sulfure.dioxide,   
## data = data\_subset, Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## alcohol 0.869203 0.059471 14.616  
## density -9.426778 0.427938 -22.028  
## sulphates 2.070837 0.331911 6.239  
## citric.acid -0.628833 0.423761 -1.484  
## volatile.acidity -0.841046 0.720169 -1.168  
## total.sulfur.dioxide -0.005829 0.001742 -3.346  
## chloride\_to\_sulphate -4.259009 1.093675 -3.894  
## volatile\_to\_fixed\_acidity -20.428885 5.339168 -3.826  
## free\_to\_total\_sulfure.dioxide 1.041803 0.365607 2.850  
##   
## Intercepts:  
## Value Std. Error t value   
## 3|4 -7.3751 0.4681 -15.7550  
## 4|5 -5.4386 0.4695 -11.5847  
## 5|6 -1.7003 0.4804 -3.5396  
## 6|7 1.1474 0.5160 2.2238  
## 7|8 4.1439 0.5785 7.1634  
##   
## Residual Deviance: 3073.782   
## AIC: 3101.782

## [1] "Confidence Levels:"

## 2.5 % 97.5 %  
## alcohol 0.752641665 0.985764951  
## density -10.265521765 -8.588034596  
## sulphates 1.420303554 2.721370147  
## citric.acid -1.459389427 0.201723332  
## volatile.acidity -2.252550543 0.570458515  
## total.sulfur.dioxide -0.009242937 -0.002414235  
## chloride\_to\_sulphate -6.402572413 -2.115445487  
## volatile\_to\_fixed\_acidity -30.893461066 -9.964308536  
## free\_to\_total\_sulfure.dioxide 0.325227053 1.758379079

# The model can also be built for the scenario where the ‘transformed’ quality categories of ‘mediocre’, ‘ok’, and ‘excellent’ are the desired prediction outcome, and those modeling results are as follows:

## Call:  
## polr(formula = excellent\_mediocre ~ alcohol + density + sulphates +   
## citric.acid + volatile.acidity + total.sulfur.dioxide + chloride\_to\_sulphate +   
## volatile\_to\_fixed\_acidity + free\_to\_total\_sulfure.dioxide,   
## data = data\_subset, Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## alcohol 6.968e-01 0.078473 8.8790  
## density -1.515e+02 0.552851 -274.0592  
## sulphates 2.026e+00 0.450141 4.5015  
## citric.acid 6.508e-01 0.608723 1.0691  
## volatile.acidity 1.583e-01 1.043737 0.1516  
## total.sulfur.dioxide -1.173e-03 0.002376 -0.4937  
## chloride\_to\_sulphate -5.147e+00 1.521959 -3.3819  
## volatile\_to\_fixed\_acidity -2.938e+01 7.774591 -3.7788  
## free\_to\_total\_sulfure.dioxide 6.140e-01 0.520361 1.1800  
##   
## Intercepts:  
## Value Std. Error t value   
## mediocre|ok -148.7238 0.6197 -240.0107  
## ok|excellent -142.0578 0.6863 -206.9808  
##   
## Residual Deviance: 1378.793   
## AIC: 1400.793

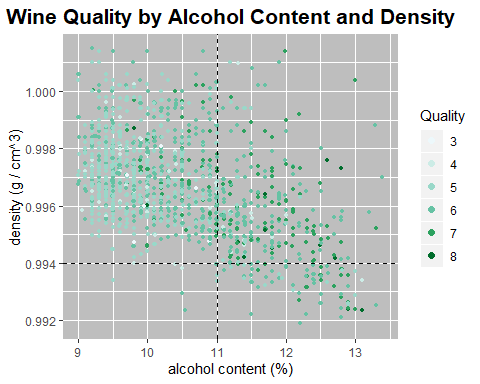
## [1] "Confidence Levels:"

## 2.5 % 97.5 %  
## alcohol 5.429597e-01 8.505697e-01  
## density -1.525976e+02 -1.504304e+02  
## sulphates 1.144070e+00 2.908590e+00  
## citric.acid -5.422662e-01 1.843883e+00  
## volatile.acidity -1.887416e+00 2.203958e+00  
## total.sulfur.dioxide -5.829032e-03 3.483229e-03  
## chloride\_to\_sulphate -8.130066e+00 -2.164096e+00  
## volatile\_to\_fixed\_acidity -4.461688e+01 -1.414105e+01  
## free\_to\_total\_sulfure.dioxide -4.058795e-01 1.633900e+00

Both models appear to fit the data well, with the estimated value to standard error ratio (i.e. the t-value) exceeding 2.9 for all parameters. The parameter estimate for alcohol content and density had the highest t-values for both models, which is not surprising given the trends observed in the multivariate graphs, where these two properties were key predictors of a given wine’s quality score.

Both models have limitiations however. First, they are only valid for the quality range exhibited in the dataset. Since the dataset only contained wines in the 3-9 quality range, these models would be unreliable at identifying wines outside of this range. Second, the models are only valid for the particular wine under consideration here (i.e. Portuguese "Vinho Verde" wines). A new model would likely be needed for each wine variety, or at the very least, this model would need to be validated against a new set of data before one could make any claims about its applicability beyond this particular dataset and wine variety.

## Final Plots and Summary

In this section, three particularly interesting graphs that help summarize the key findings from the EDA are presented. ### Plot One  #This plot demonstrates that in general, the high quality wines (quality 7-8) tend to have high alcohol content and low density, as shown by the preponderance of green shaded points in the lower right quadrant of the graph. Conversely, the poor quality wines (quality 3-4) tend to have low alcohol content and high density, dominating the two left side quadrants.

# Reflection

##One major struggle in this EDA was getting my head around all the variables and their impact on quality. It wasn’t just the large number of variables that was problematic, but the fact that many of these variables are likely related.