HW3 Attempt 2

Taylor Thul

February 14, 2017

White Wine Data Report: Data Exploration and Multiple Linear Regression Model Fitting

## Loading required package: splines

## Loading required package: RcmdrMisc

## Loading required package: car

## Loading required package: sandwich

## The Commander GUI is launched only in interactive sessions

Run some summary statistics:

> numSummary(White\_wines[,c("alcohol", "chlorides", "citric.acid", "density",   
+ "fixed.acidity", "free.sulfur.dioxide", "pH", "quality", "residual.sugar",   
+ "sulphates", "total.sulfur.dioxide", "volatile.acidity")],   
+ statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))

mean sd IQR 0%  
alcohol 10.51426705 1.230620568 1.9000000 8.00000  
chlorides 0.04577236 0.021847968 0.0140000 0.00900  
citric.acid 0.33419151 0.121019804 0.1200000 0.00000  
density 0.99402738 0.002990907 0.0043775 0.98711  
fixed.acidity 6.85478767 0.843868228 1.0000000 3.80000  
free.sulfur.dioxide 35.30808493 17.007137325 23.0000000 2.00000  
pH 3.18826664 0.151000600 0.1900000 2.72000  
quality 5.87790935 0.885638575 1.0000000 3.00000  
residual.sugar 6.39141486 5.072057784 8.2000000 0.60000  
sulphates 0.48984688 0.114125834 0.1400000 0.22000  
total.sulfur.dioxide 138.36065741 42.498064554 59.0000000 9.00000  
volatile.acidity 0.27824112 0.100794548 0.1100000 0.08000  
 25% 50% 75% 100% n  
alcohol 9.5000000 10.40000 11.4000 14.20000 4898  
chlorides 0.0360000 0.04300 0.0500 0.34600 4898  
citric.acid 0.2700000 0.32000 0.3900 1.66000 4898  
density 0.9917225 0.99374 0.9961 1.03898 4898  
fixed.acidity 6.3000000 6.80000 7.3000 14.20000 4898  
free.sulfur.dioxide 23.0000000 34.00000 46.0000 289.00000 4898  
pH 3.0900000 3.18000 3.2800 3.82000 4898  
quality 5.0000000 6.00000 6.0000 9.00000 4898  
residual.sugar 1.7000000 5.20000 9.9000 65.80000 4898  
sulphates 0.4100000 0.47000 0.5500 1.08000 4898  
total.sulfur.dioxide 108.0000000 134.00000 167.0000 440.00000 4898  
volatile.acidity 0.2100000 0.26000 0.3200 1.10000 4898

So the main variables we are looking at include alcohol, chlorides, citric acid, density, fixed acidity, free sulfur dioxide, pH, residual sugar, sulfates, total sulfur dioxide, and volatile acidity. We want to know how each of these variables influence the variable 'quality.'

Make sure no data is missing:

> sapply(White\_wines, function(x)(sum(is.na(x))))

fixed.acidity volatile.acidity citric.acid   
 0 0 0   
 residual.sugar chlorides free.sulfur.dioxide   
 0 0 0   
total.sulfur.dioxide density pH   
 0 0 0   
 sulphates alcohol quality   
 0 0 0

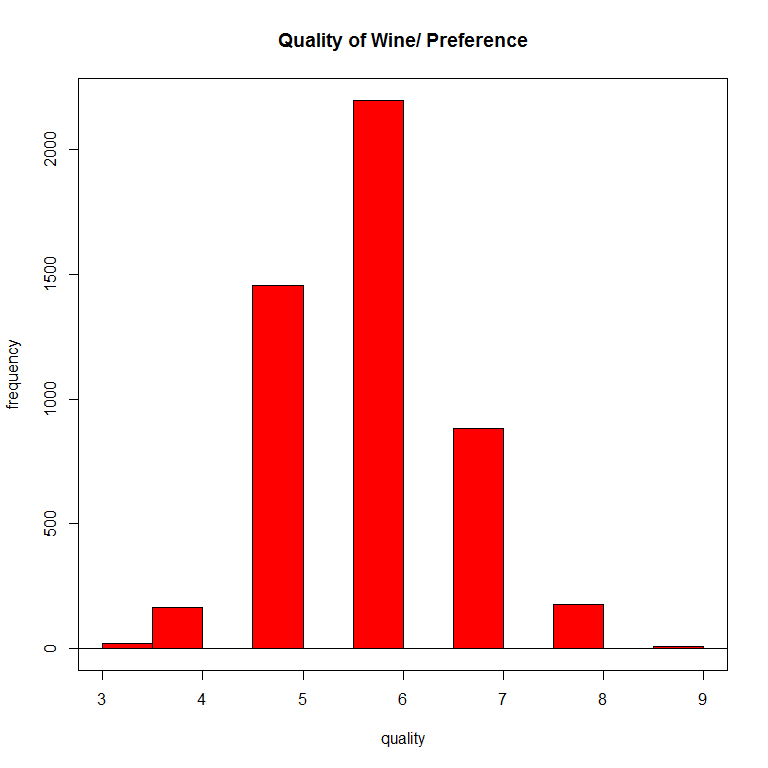
> # NA counts

No missing data! Hooray!

Now to explore each variable individually.

Firstly let us look at the variable of interest, quality:

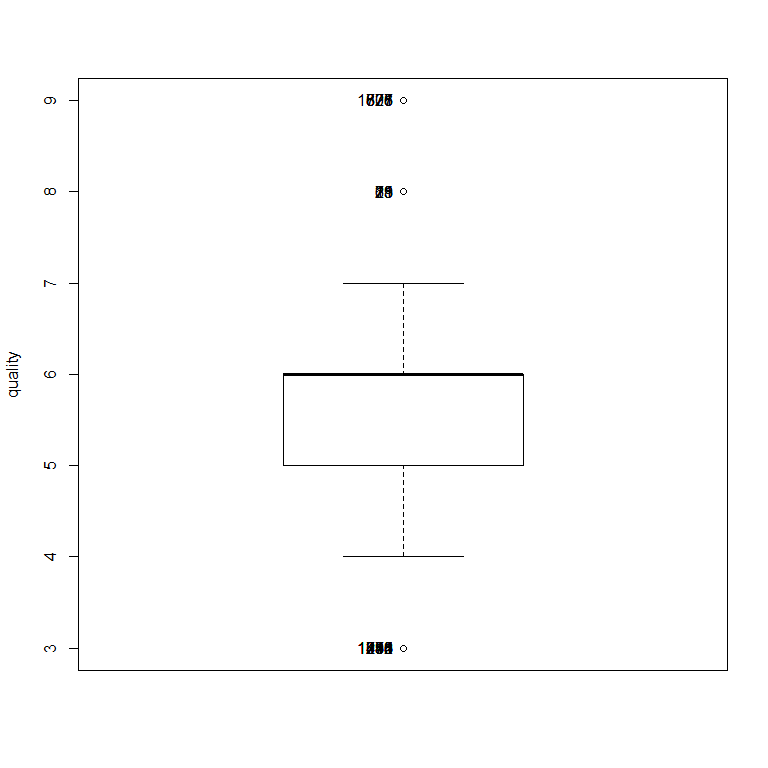
> with(White\_wines, Hist(quality, scale="frequency", breaks="Sturges", col="red"))  
> #add lables  
> title (main= "Quality of Wine/ Preference")



Based on this histogram the data appears fairly normally distributed.

Since this is the variable of interest we will also use other visual means of exploring the data:

> Boxplot( ~ quality, data=White\_wines, id.method="y")

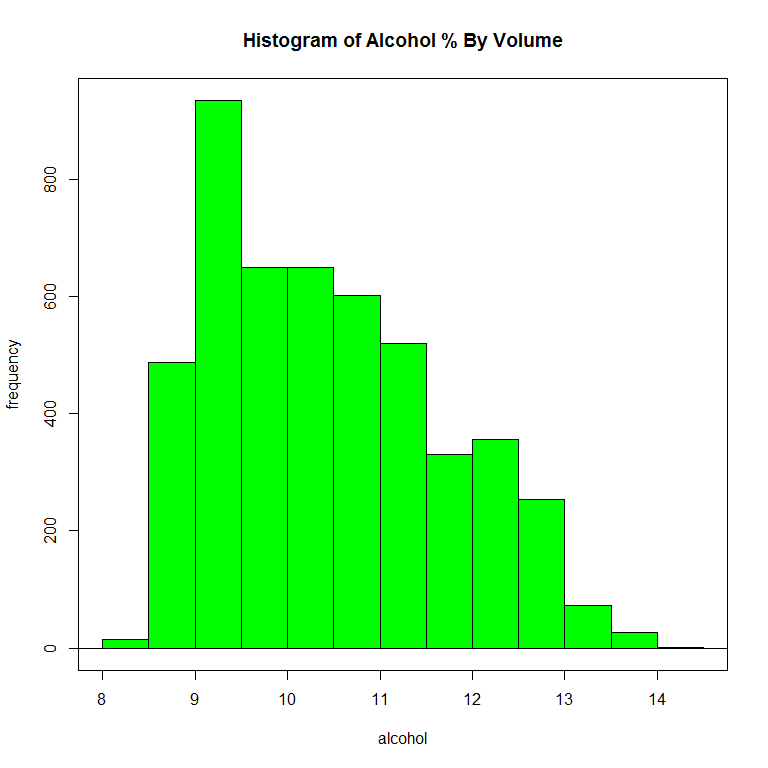


[1] "252" "254" "295" "446" "741" "874" "1035" "1230" "1418" "1485"  
[11] "775" "821" "828" "877" "1606" "18" "21" "23" "69" "75"

From the box plot we can see there are a few outliers which will need to be kept in mind during analysis.

Now look at alcohol:

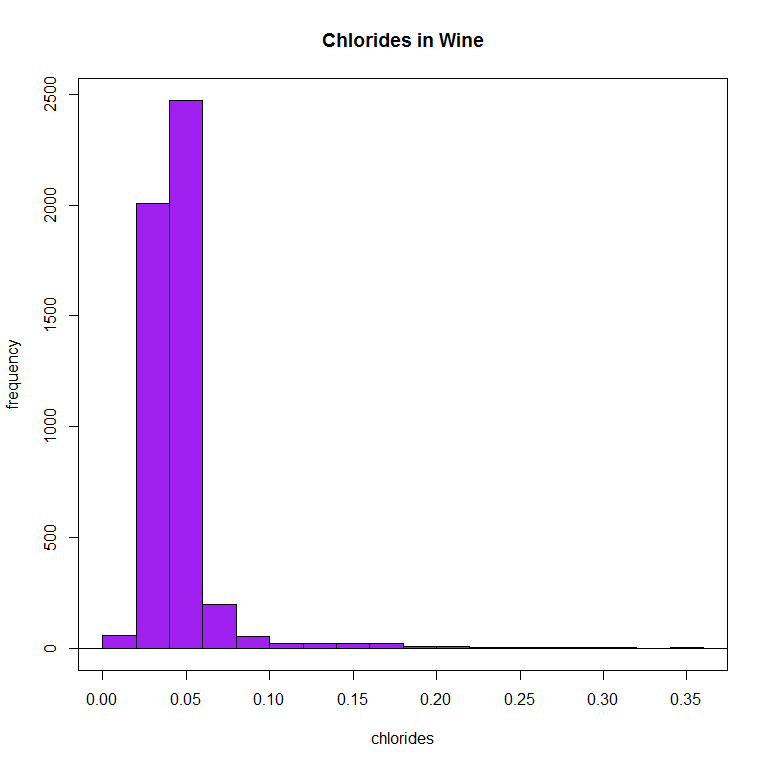
> with(White\_wines, Hist(alcohol, scale="frequency", breaks="Sturges", col="green"))  
>   
> #add lables  
> title(main="Histogram of Alcohol % By Volume")



Fairly normal distribution with few outliers.

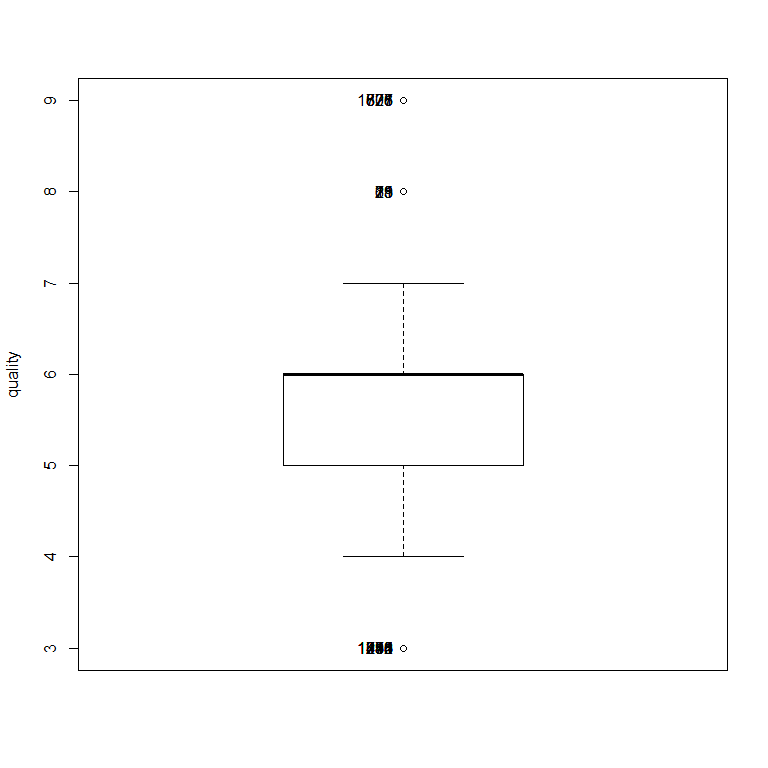
Chlorides:

> with(White\_wines, Hist(chlorides, scale="frequency", breaks="Sturges", col="purple"))  
> title(main="Chlorides in Wine")



This appears to be skewed and have some outliers so I will do a box plot

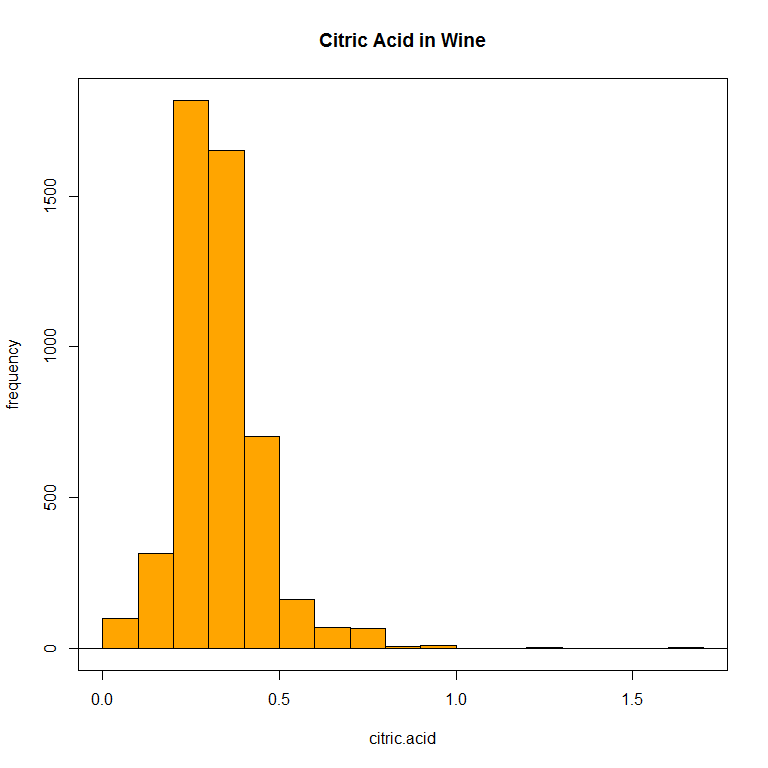
> Boxplot( ~ quality, data=White\_wines, id.method="y")



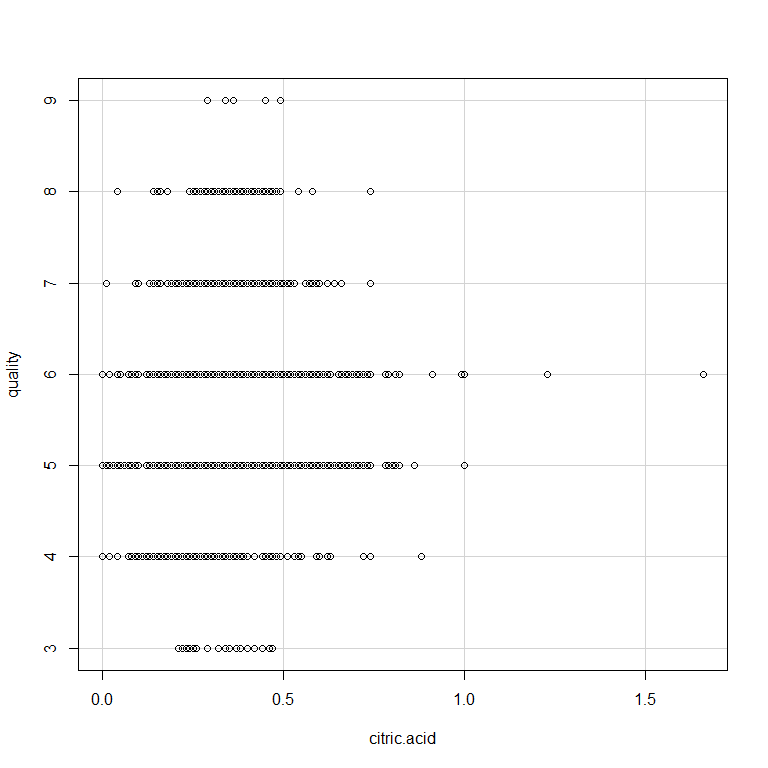
[1] "252" "254" "295" "446" "741" "874" "1035" "1230" "1418" "1485"  
[11] "775" "821" "828" "877" "1606" "18" "21" "23" "69" "75"

Citric Acid:

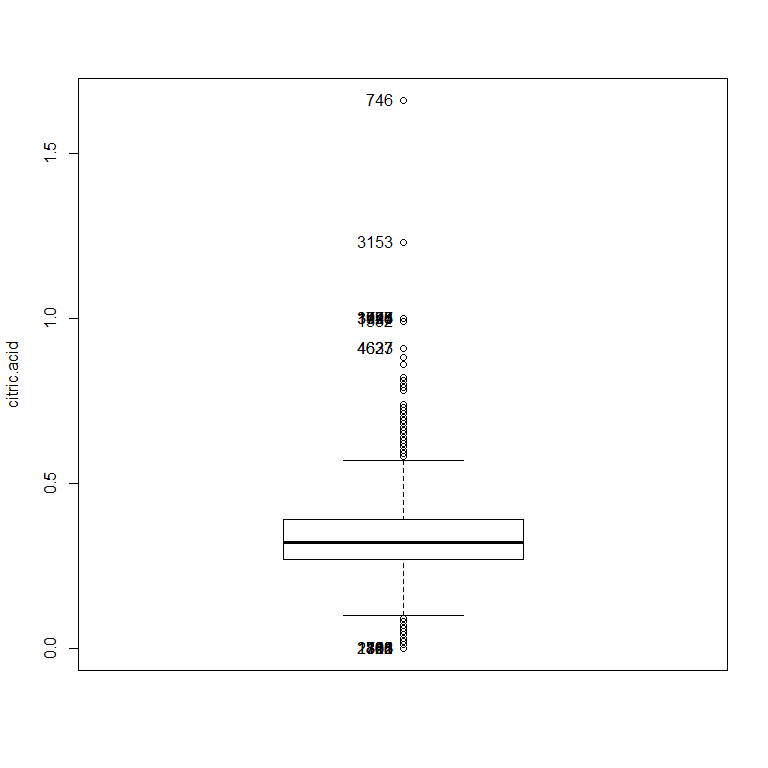
> with(White\_wines, Hist(citric.acid, scale="frequency", breaks="Sturges", col="orange"))  
> title (main= "Citric Acid in Wine")



> scatterplot(quality~citric.acid, reg.line=FALSE, smooth=FALSE, spread=FALSE, boxplots=FALSE,   
+ span=0.5, ellipse=FALSE, levels=c(.5, .9), data=White\_wines)



> Boxplot( ~ citric.acid, data=White\_wines, id.method="y")

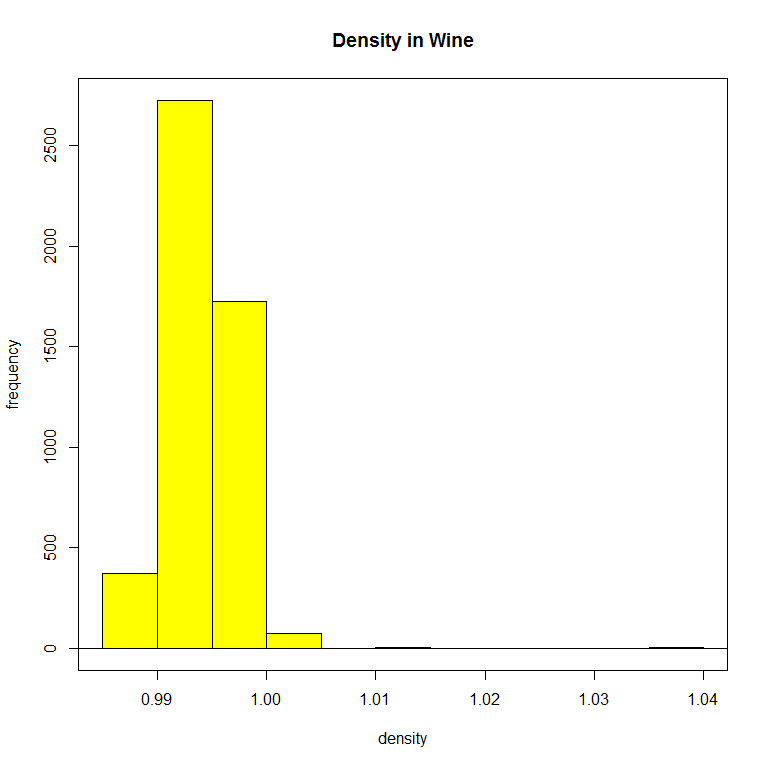


[1] "116" "301" "303" "781" "863" "865" "891" "1153" "1818" "2322"  
[11] "746" "3153" "947" "1723" "1776" "3044" "3498" "1552" "4627" "4633"

Citric acid also appears to have quite a few outliers and some newness.

Density

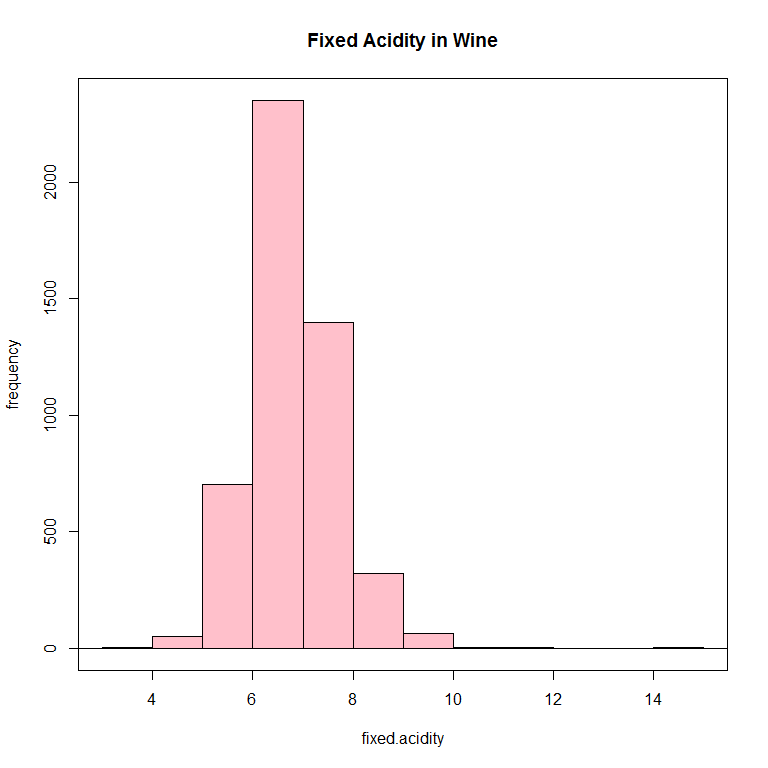
> with(White\_wines, Hist(density, scale="frequency", breaks="Sturges", col="yellow"))  
> title(main= "Density in Wine")



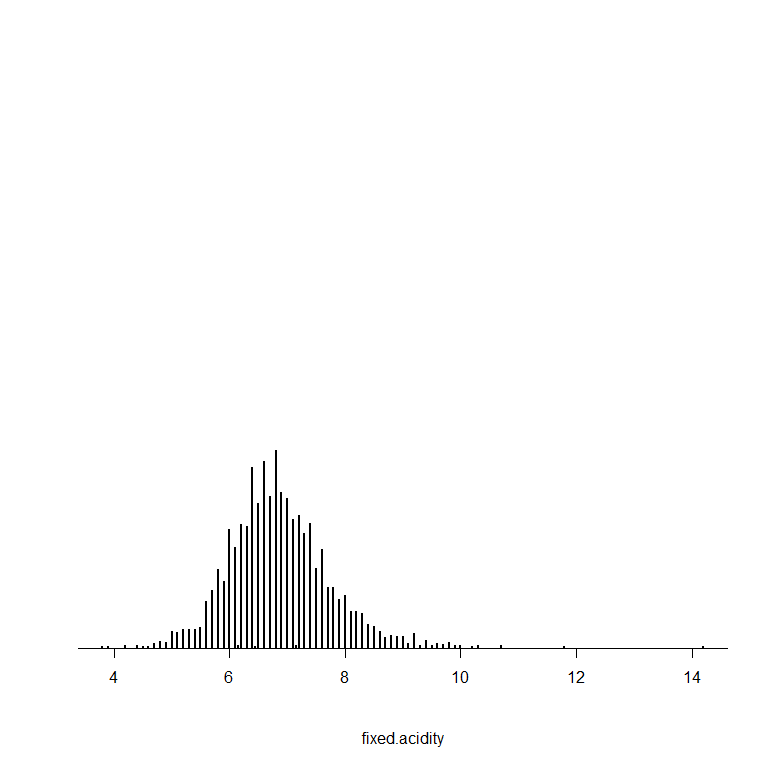
Density also appears to have a skew with outliers to the right. I'm noting most of the skewness is to the left so I am now wondering if perhaps these variables are correlated. I will explore this later after the individual variables.

Fixed acidity

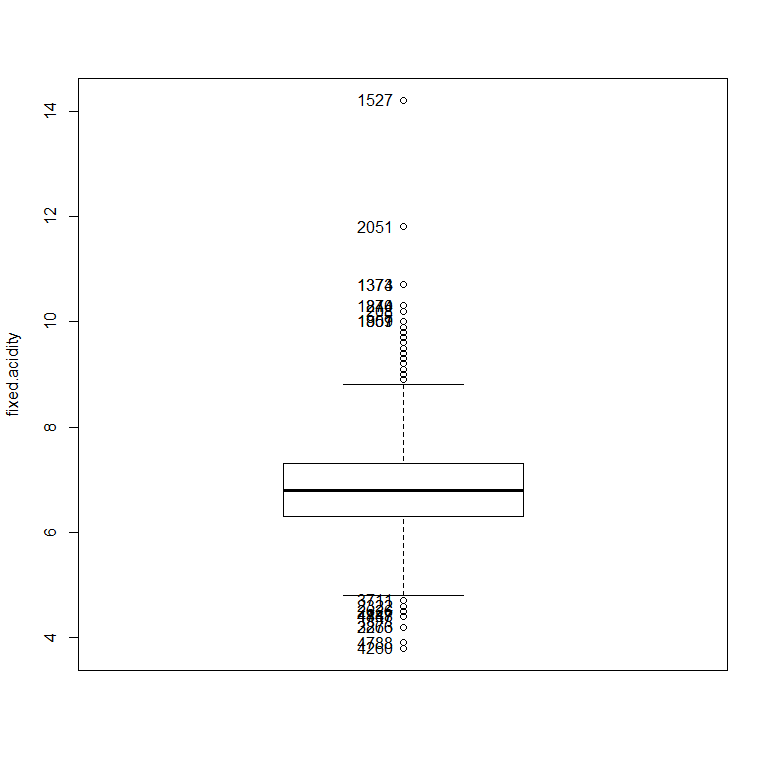
> with(White\_wines, Hist(fixed.acidity, scale="frequency", breaks="Sturges", col="pink"))  
> title(main="Fixed Acidity in Wine")



> #using a dotplot  
> with(White\_wines, Dotplot(fixed.acidity, bin=FALSE))



> Boxplot( ~ fixed.acidity, data=White\_wines, id.method="y")

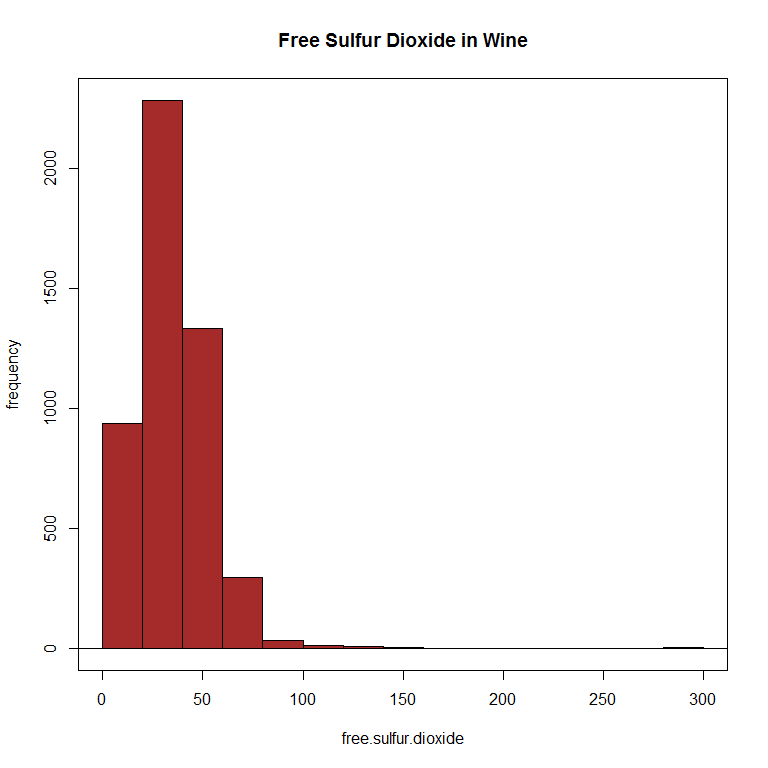


[1] "4260" "4788" "2873" "3266" "4447" "4787" "4848" "2626" "2322" "3711"  
[11] "1527" "2051" "1373" "1374" "874" "1240" "208" "359" "1857" "1901"

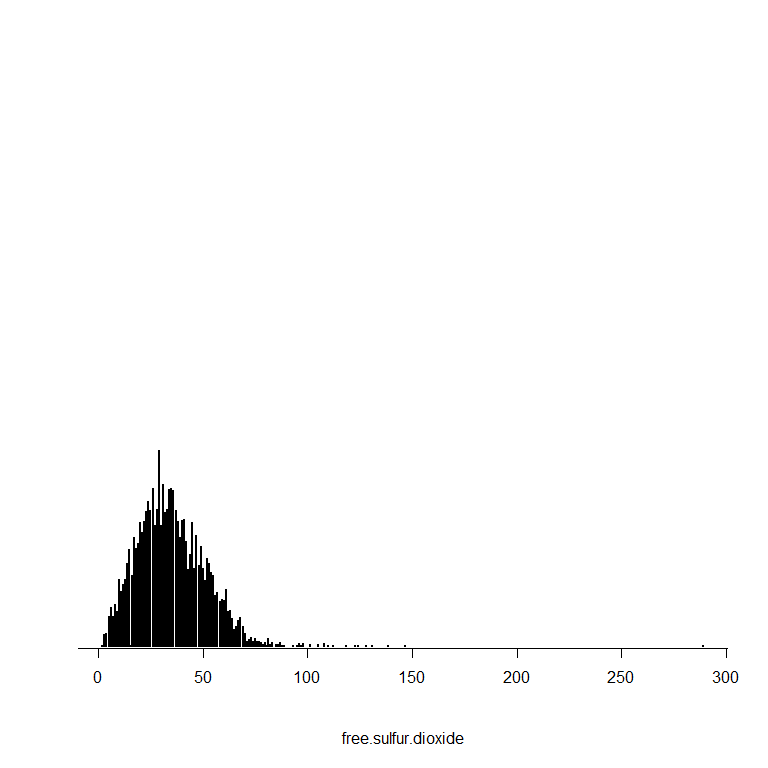
Again outliers noted so a box plot reviewed.

Free sulfur dioxide:

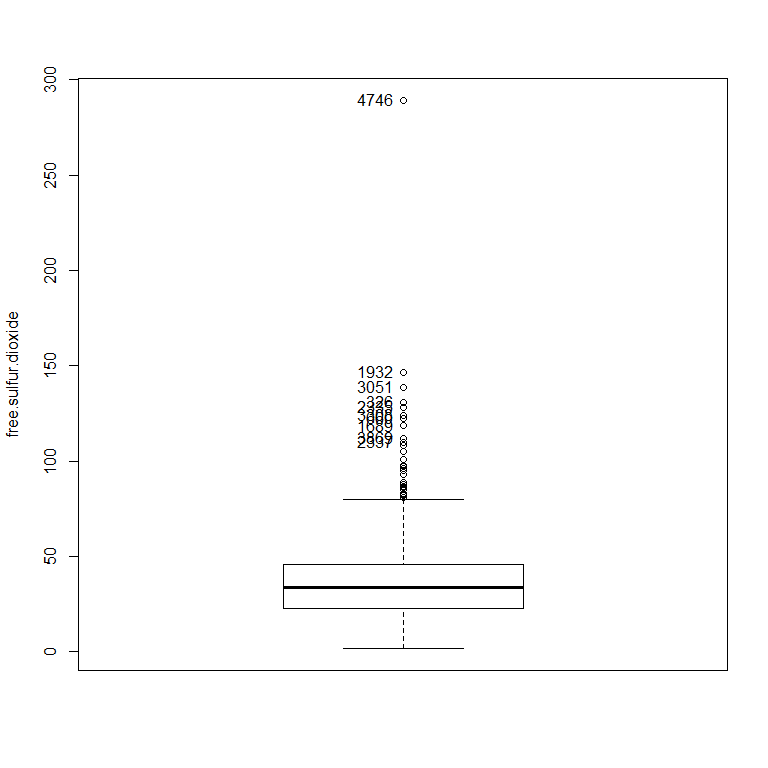
> with(White\_wines, Hist(free.sulfur.dioxide, scale="frequency", breaks="Sturges", col="brown"))  
> title(main="Free Sulfur Dioxide in Wine")



> #using a dotplot  
> with(White\_wines, Dotplot(free.sulfur.dioxide, bin=FALSE))

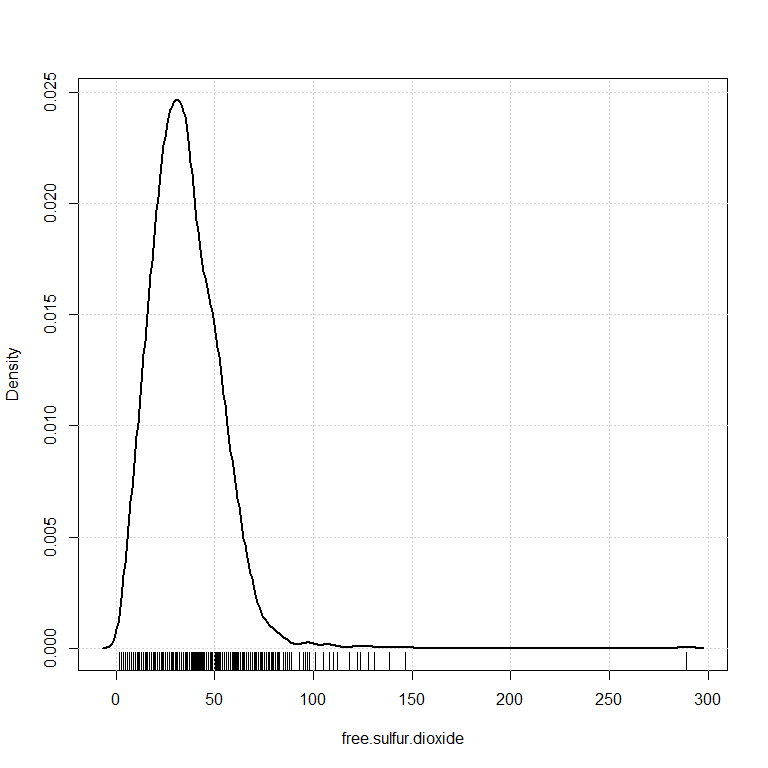


> Boxplot( ~ free.sulfur.dioxide, data=White\_wines, id.method="y")



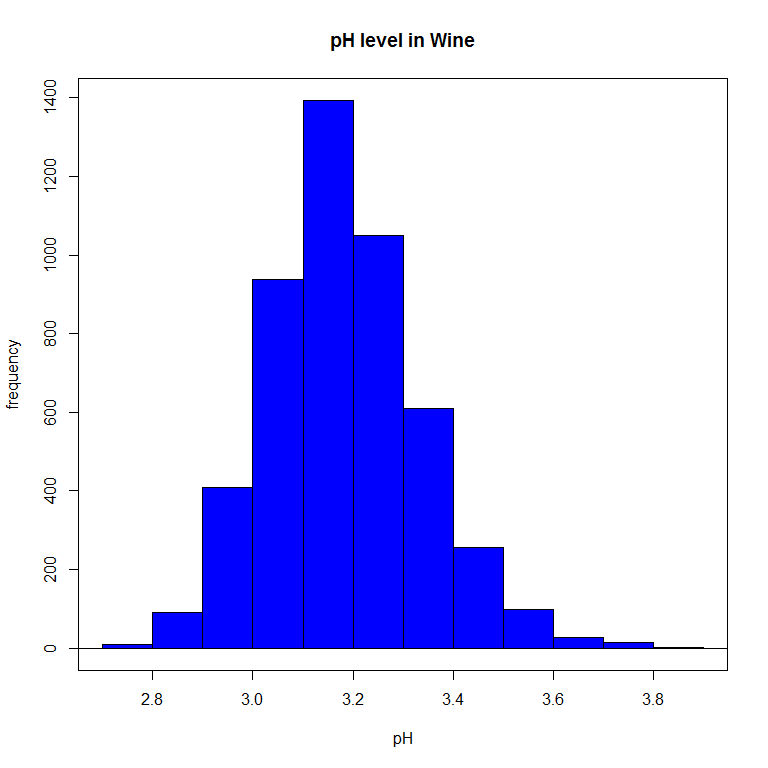
[1] "4746" "1932" "3051" "326" "2335" "3308" "660" "1689" "3869" "2337"

> #trying a density plot  
> densityPlot( ~ free.sulfur.dioxide, data=White\_wines, bw="SJ", adjust=1, kernel="gaussian")

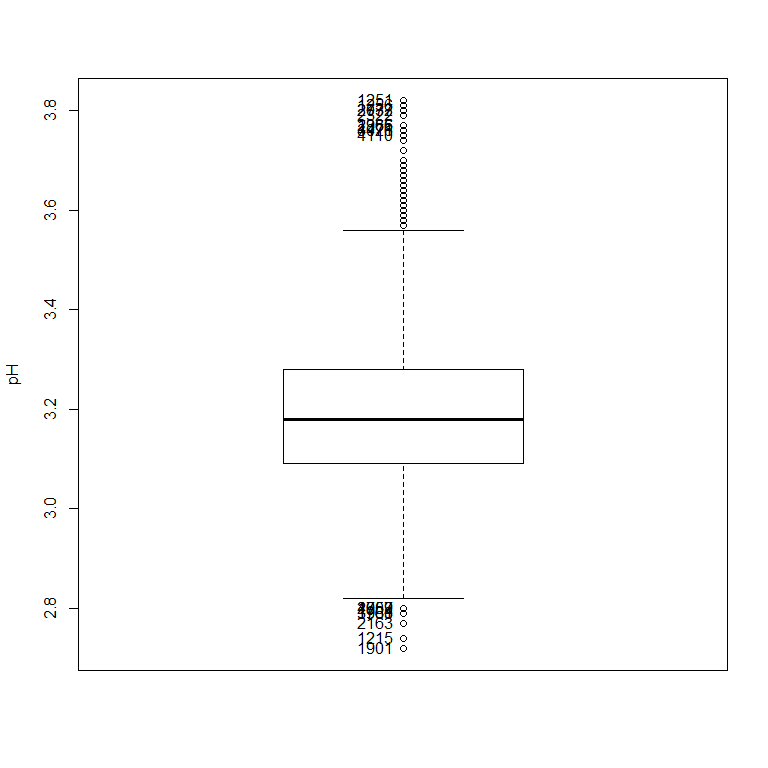


pH:

> with(White\_wines, Hist(pH, scale="frequency", breaks="Sturges", col="blue"))  
> title(main="pH level in Wine")



> Boxplot( ~ pH, data=White\_wines, id.method="y")

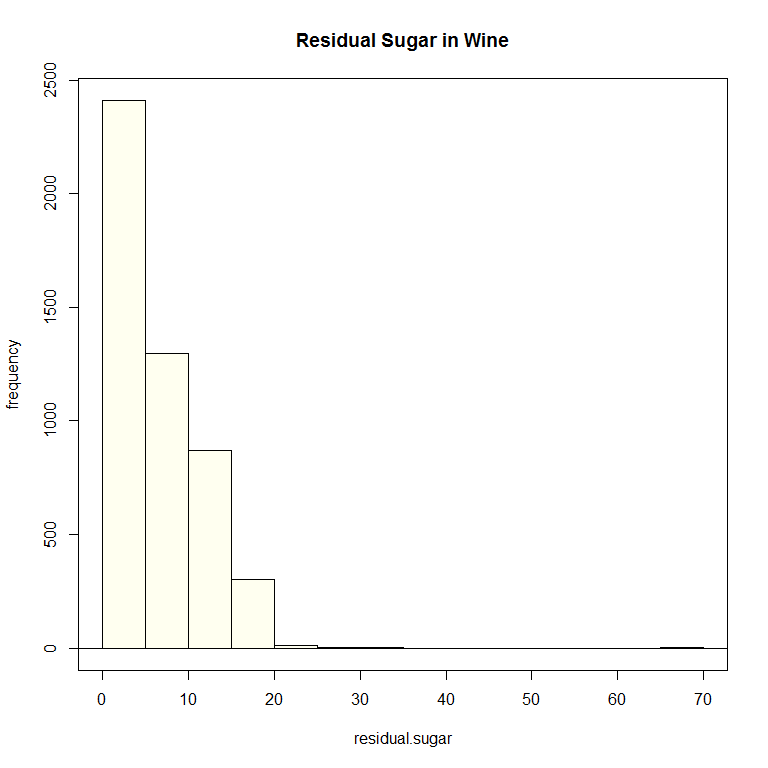


[1] "1215" "1759" "1901" "1960" "1961" "2163" "2957" "3763" "4602" "1251"  
[11] "1256" "2037" "2772" "2322" "1386" "2965" "3026" "4471" "4110"

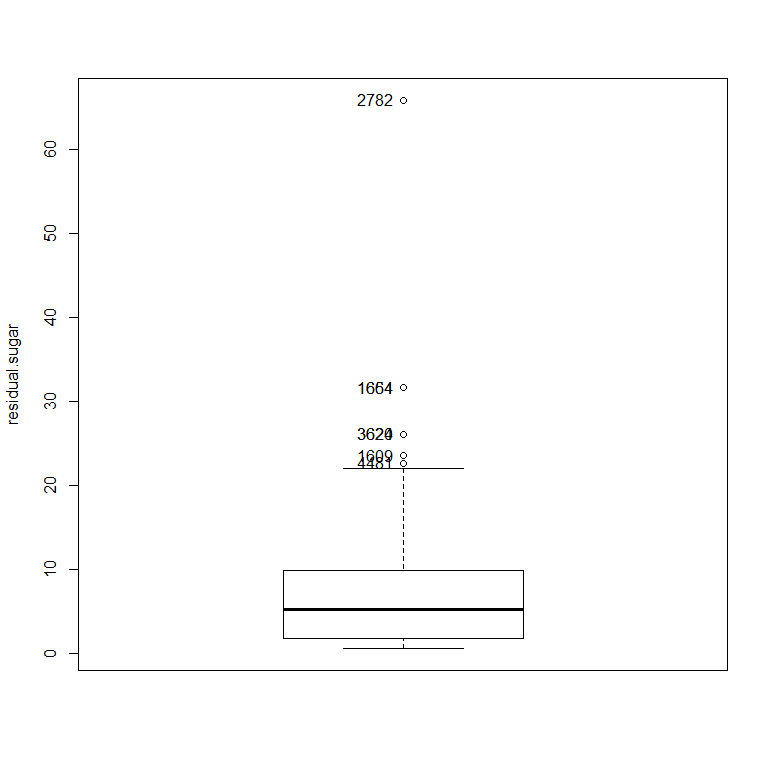
pH appears to have less outliers and overall more even distribution than the variables previous.

Residual sugar:

> with(White\_wines, Hist(residual.sugar, scale="frequency", breaks="Sturges", col="ivory"))  
> title(main="Residual Sugar in Wine")



> Boxplot( ~ residual.sugar, data=White\_wines, id.method="y")

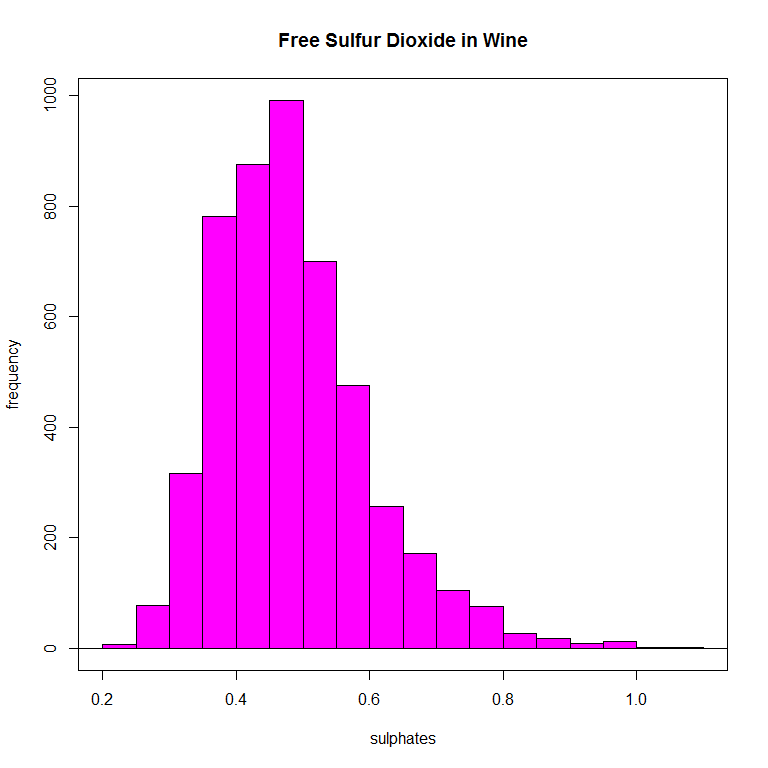


[1] "1609" "1654" "1664" "2782" "3620" "3624" "4481"

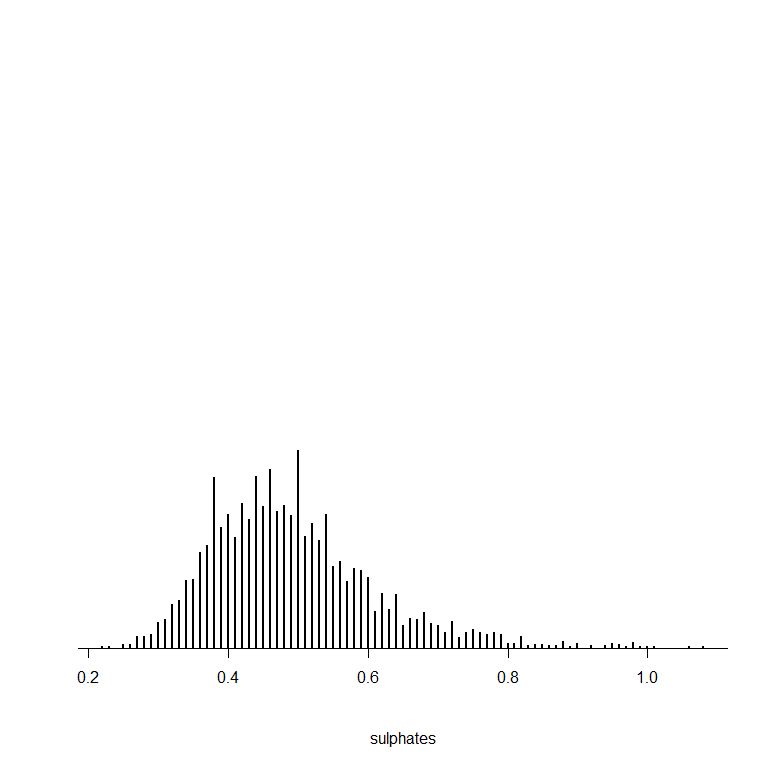
Continuing the trend of large outliers.

Sulphates: I am interested especially in this variable as I have read the level of sulphates influences the taste of wine. I will see if the data supports that in this case.

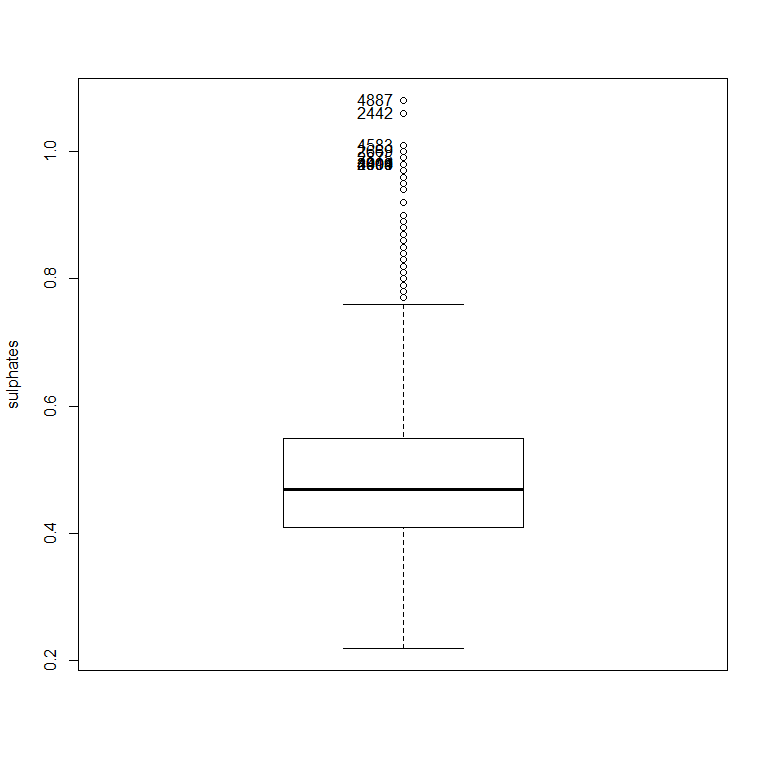
> with(White\_wines, Hist(sulphates, scale="frequency", breaks="Sturges", col="magenta"))  
> title(main="Free Sulfur Dioxide in Wine")



> #using a dotplot  
> with(White\_wines, Dotplot(sulphates, bin=FALSE))



> Boxplot( ~ sulphates, data=White\_wines, id.method="y")

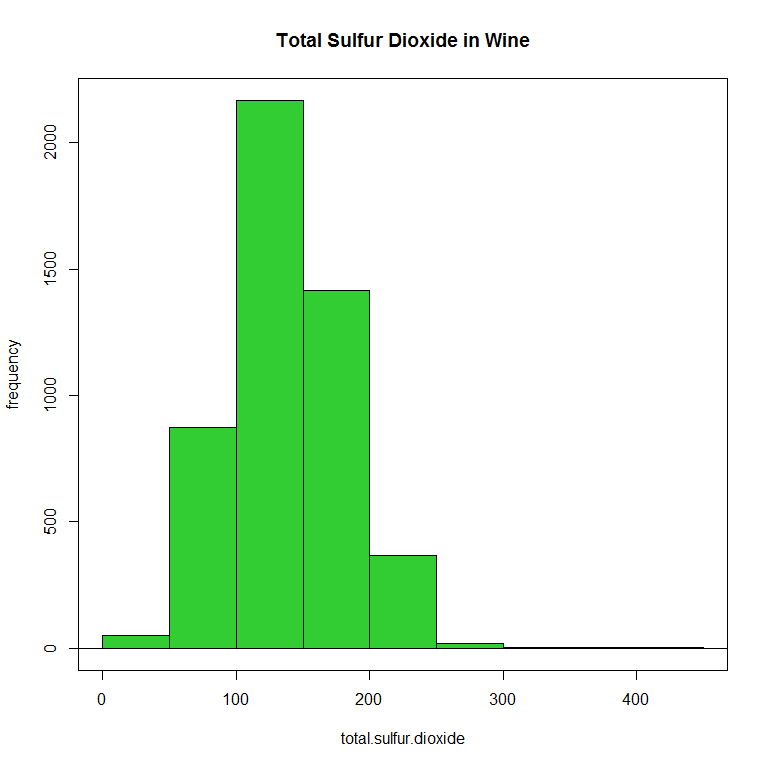


[1] "4887" "2442" "4583" "2669" "2875" "2404" "3999" "4000" "4001" "4013"

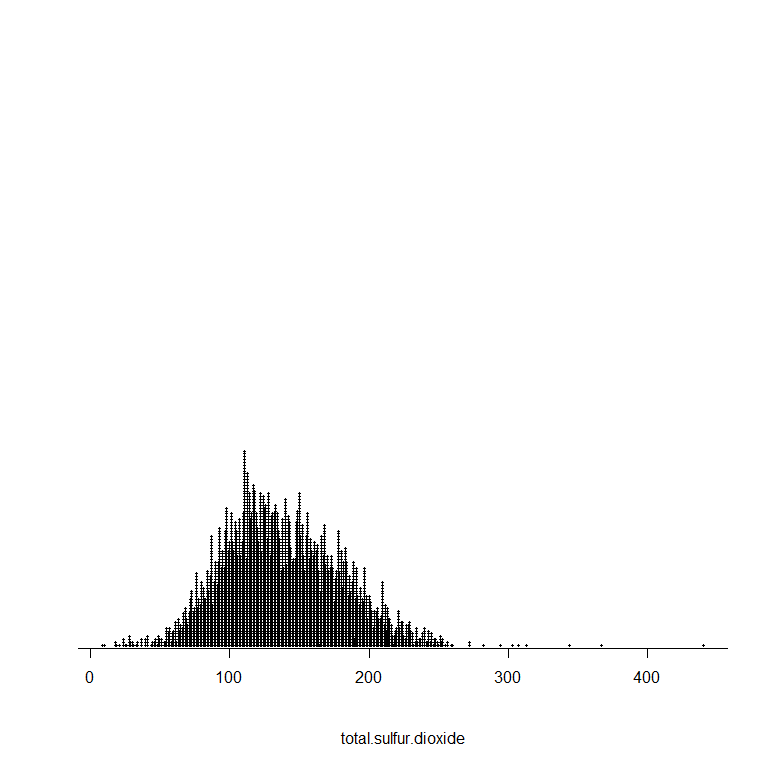
Still some skew and outliers to the right but appears more evenly distributed.

Total sulfur dixoide: I anticipate this may be related to free sulfur dioxide so will keep in mind to test later.

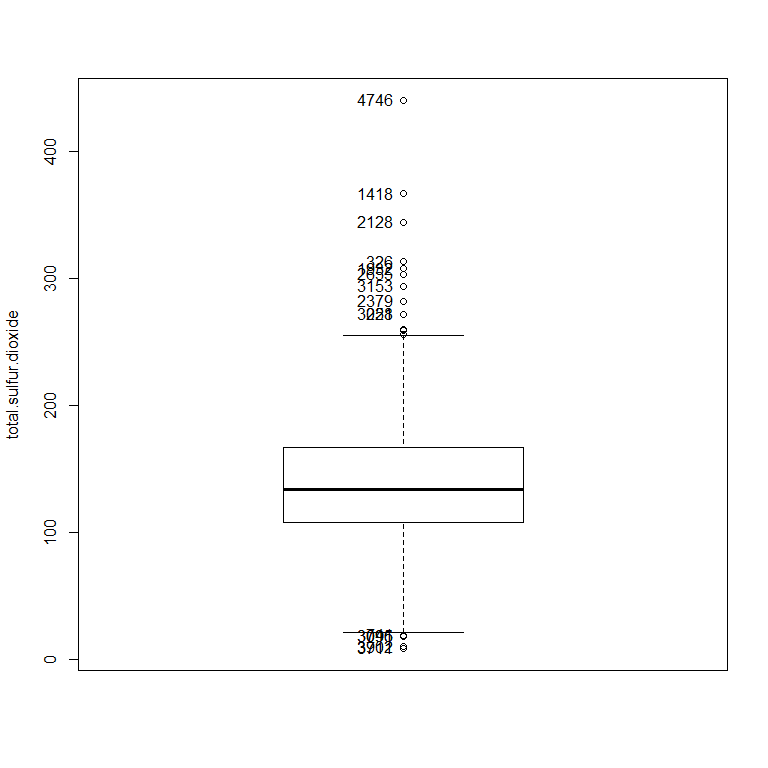
> with(White\_wines, Hist(total.sulfur.dioxide, scale="frequency", breaks="Sturges", col="limegreen"))  
> title(main="Total Sulfur Dioxide in Wine")



> #using a dotplot  
> with(White\_wines, Dotplot(total.sulfur.dioxide, bin=FALSE))



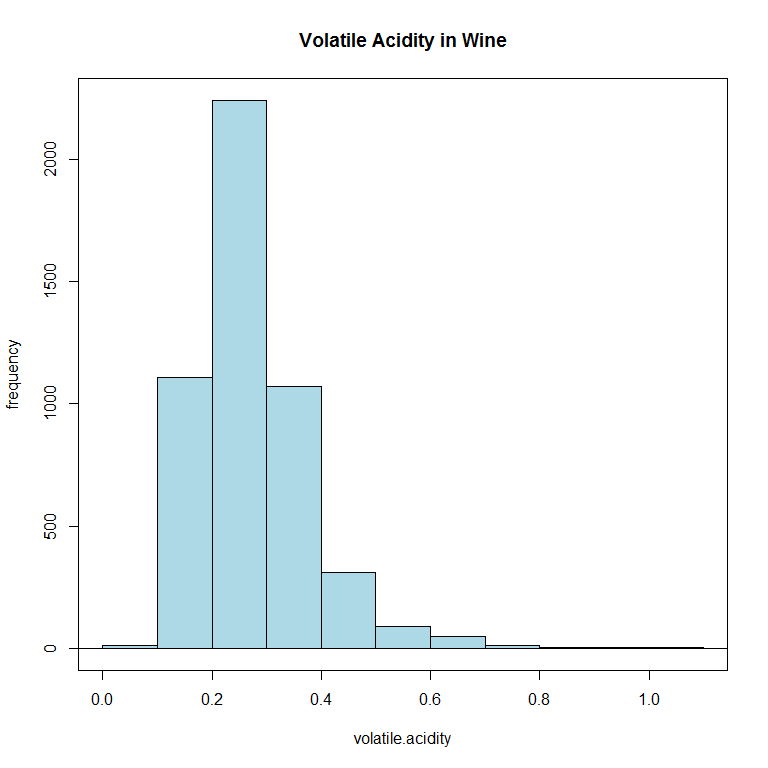
> Boxplot( ~ total.sulfur.dioxide, data=White\_wines, id.method="y")



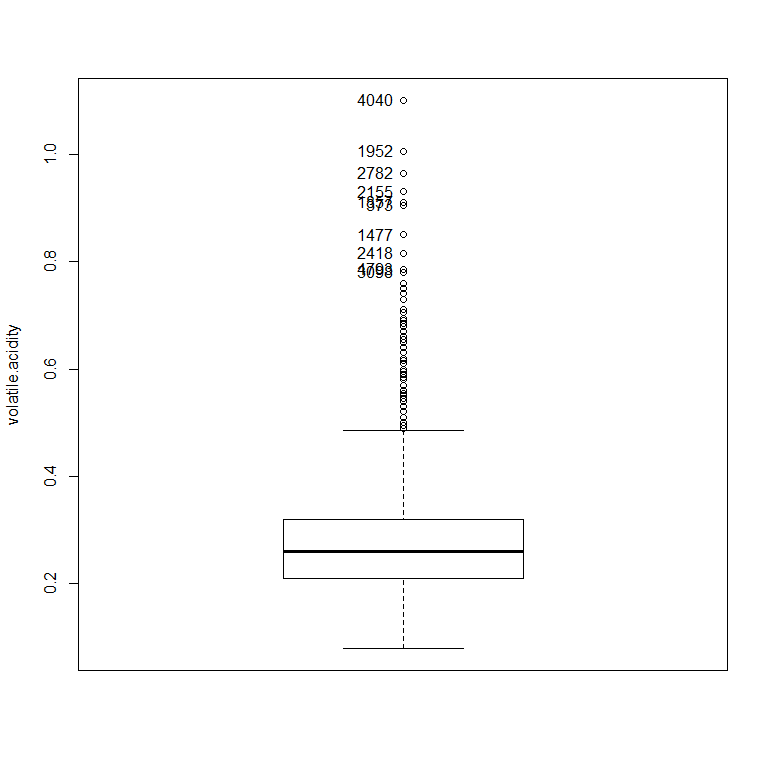
[1] "741" "3095" "3096" "3711" "3902" "4746" "1418" "2128" "326" "1932"  
[11] "2655" "3153" "2379" "228" "3051"

Volatile acidity:

> with(White\_wines, Hist(volatile.acidity, scale="frequency", breaks="Sturges", col="lightblue"))  
> title(main="Volatile Acidity in Wine")

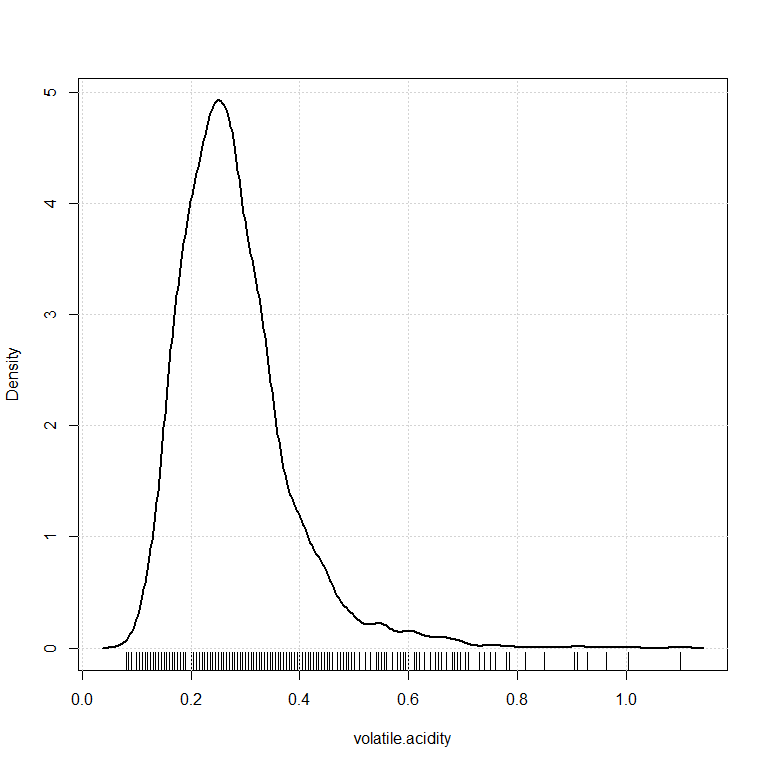


> Boxplot( ~ volatile.acidity, data=White\_wines, id.method="y")



[1] "4040" "1952" "2782" "2155" "1857" "373" "1477" "2418" "4793" "3098"

> densityPlot( ~ volatile.acidity, data=White\_wines, bw="SJ", adjust=1, kernel="gaussian")

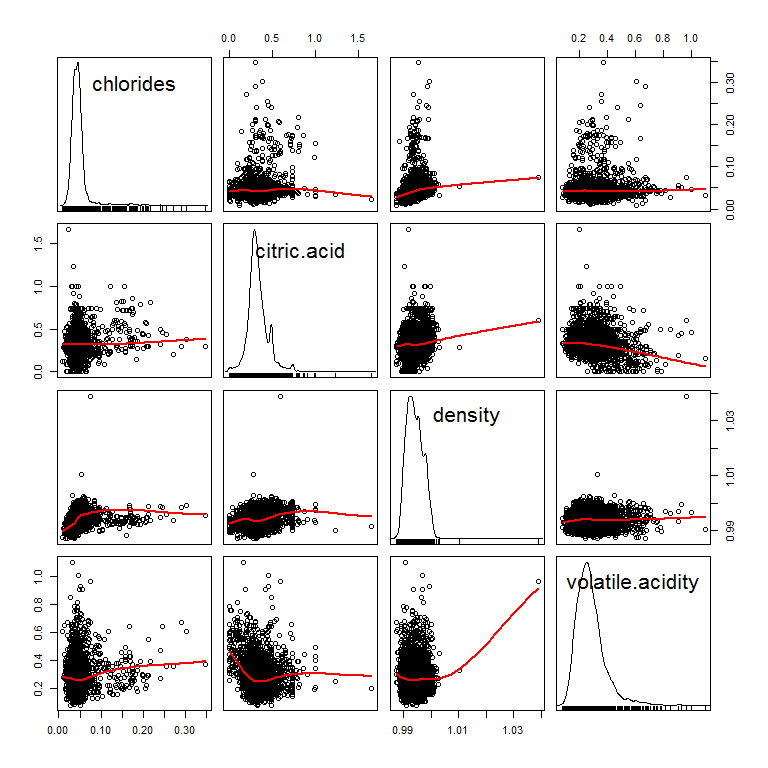


The trend continues...

Now I want to look at how these variables might be related to each other.

Compare chlorides, citric acid, density, and volatile acidity to see any relationship:

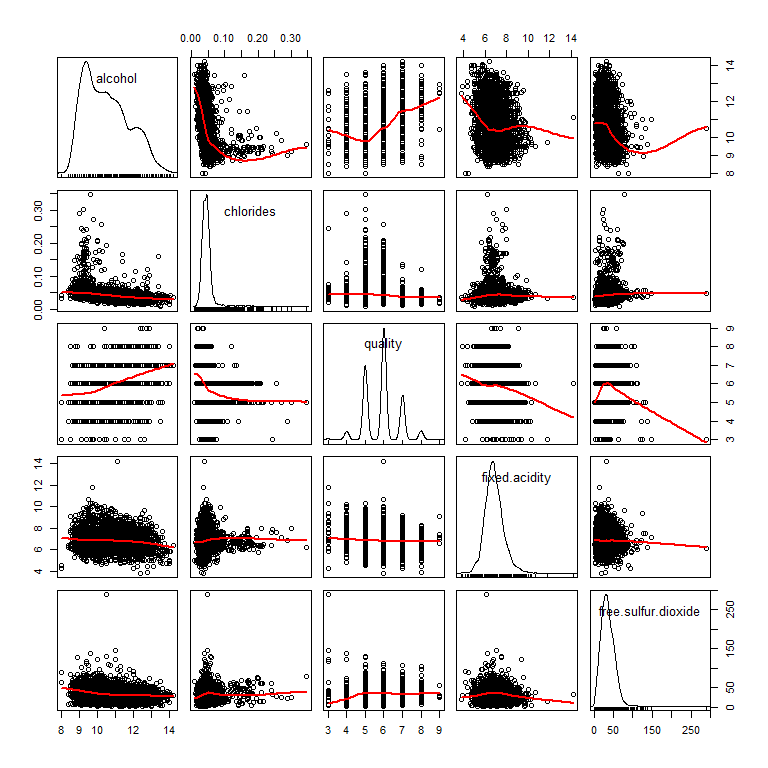
> scatterplotMatrix(~chlorides+citric.acid+density+volatile.acidity, reg.line=FALSE,   
+ smooth=TRUE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9), id.n=0, diagonal =   
+ 'density', data=White\_wines)



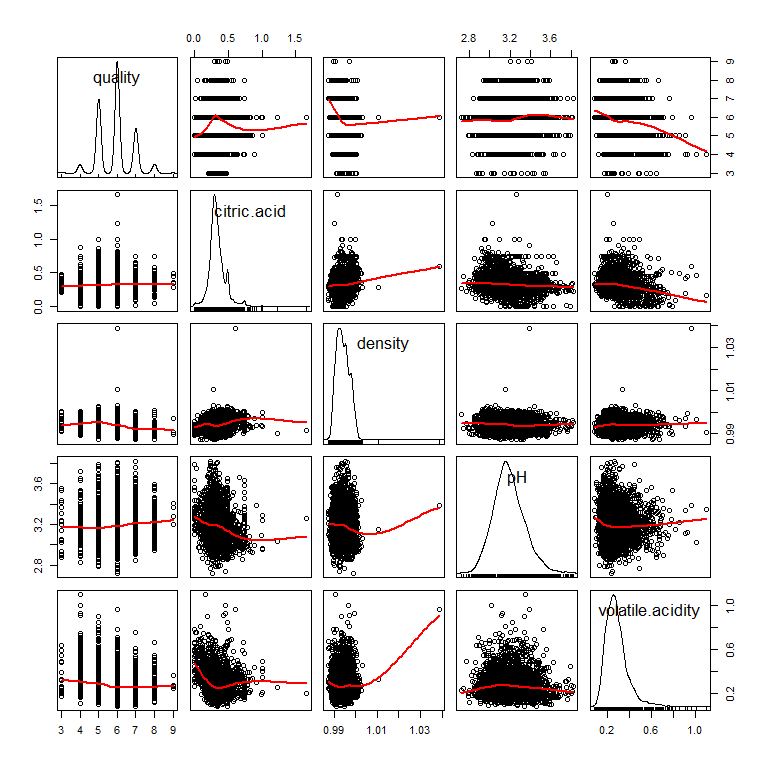
There don't seem to be any associations standing out from this comparison. This makes sense because the distribution of the data in all these variables was quite similar.

Look at how some variables may be associated with the variable of interest, quality.

> scatterplotMatrix(~alcohol+chlorides+quality+fixed.acidity+free.sulfur.dioxide, reg.line=FALSE, smooth=TRUE,  
+ spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9), id.n=0, diagonal = 'density',   
+ data=White\_wines)

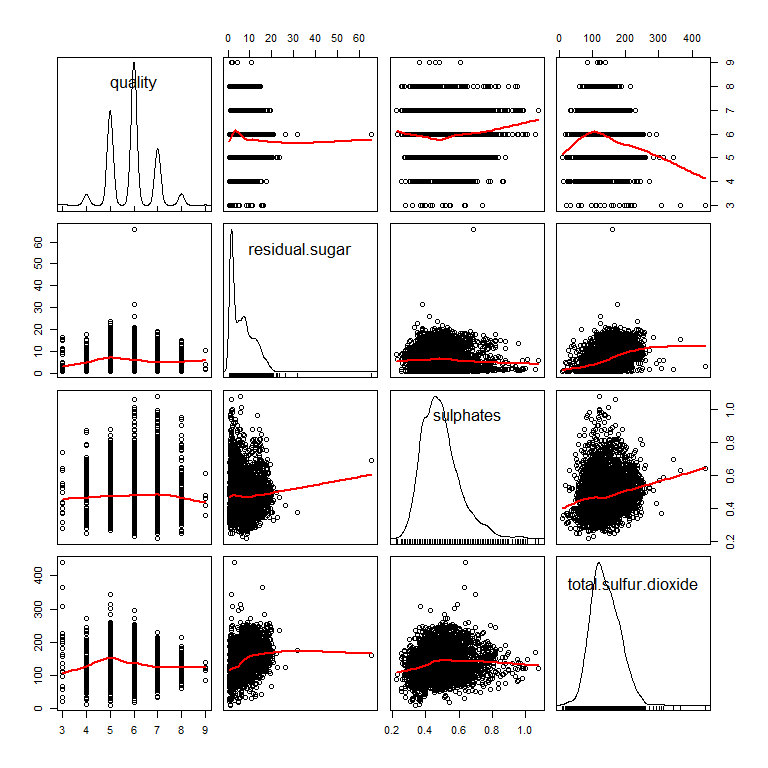
 There appear to be some weak associations between quality and fixed acidity as well as free sulfur dioxide.

> scatterplotMatrix(~quality+citric.acid+density+pH+volatile.acidity, reg.line=FALSE,   
+ smooth=TRUE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9), id.n=0, diagonal =   
+ 'density', data=White\_wines)

 No strong associations stand out except a small negative between quality and volatile acidity and a posity between density and volatile acidity.

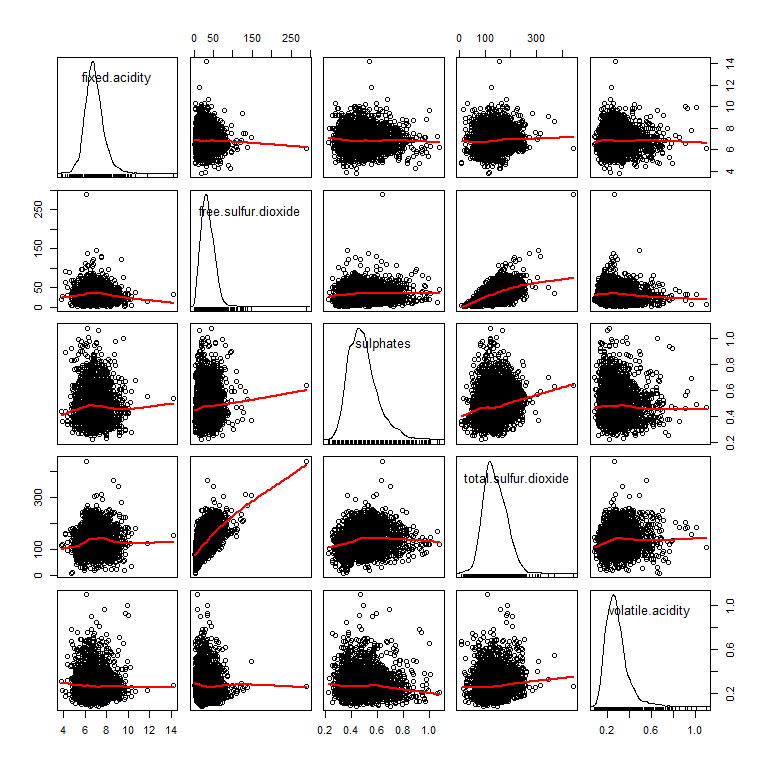
One more scatter plot matrix to capture all the variables.

> scatterplotMatrix(~quality+residual.sugar+sulphates+total.sulfur.dioxide, reg.line=FALSE,   
+ smooth=TRUE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9), id.n=0, diagonal =   
+ 'density', data=White\_wines)

 A slight negative association is seen between quality and total sulfur dioxide as well as a slight positive relationship between sulphates and total sulfar dioxide, which makes sense.

A quick exploration between variables with similar names to see any relationship before further analysis

> scatterplotMatrix(~fixed.acidity+free.sulfur.dioxide+sulphates+total.sulfur.dioxide+volatile.acidity,  
+ reg.line=FALSE, smooth=TRUE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),   
+ id.n=0, diagonal = 'density', data=White\_wines)



Some positive association are seen indicating they may be highly related.

Now lets look at how all the might go together to predict preference ('quality'). First I will split the data into a training and test data set to ensure the model fits well.

> # divide the dataset into a training and a testing set based on a random uniform number on fixed seed, which in this case we are using the date  
> # this step is also creating a new variable and adding it to the data set which is a distribution of random numbers from 0 to 1   
>   
> set.seed(20170214)  
> White\_wines$group <- runif(length(White\_wines$quality), min = 0, max = 1)  
>   
> #what random forests do is this process over and over again and makes the aggregate which might be called bootstrapping?  
>   
> White\_wines.train <- subset(White\_wines, group <= 0.90)  
> White\_wines.test <- subset(White\_wines, group > 0.90)  
>   
> #see if it worked  
> summary(White\_wines.train)

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 : 5.100   
 Mean : 6.851 Mean :0.2784 Mean :0.3337 Mean : 6.342   
 3rd Qu.: 7.300 3rd Qu.:0.3200 3rd Qu.:0.3900 3rd Qu.: 9.800   
 Max. :14.200 Max. :1.1000 Max. :1.6600 Max. :65.800   
 chlorides free.sulfur.dioxide total.sulfur.dioxide  
 Min. :0.00900 Min. : 3.00 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 :0.04574 Mean : 35.28 Mean :138.3   
 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. :2.72 Min. :0.2200 Min. : 8.00   
 1st Qu.:0.9917 1st Qu.:3.09 1st Qu.:0.4100 1st Qu.: 9.50   
 Median :0.9937 Median :3.18 Median :0.4700 Median :10.40   
 Mean :0.9940 Mean :3.19 Mean :0.4892 Mean :10.52   
 3rd Qu.:0.9960 3rd Qu.:3.28 3rd Qu.:0.5500 3rd Qu.:11.40   
 Max. :1.0390 Max. :3.82 Max. :1.0800 Max. :14.20   
 quality group   
 Min. :3.000 Min. :0.0002833   
 1st Qu.:5.000 1st Qu.:0.2285282   
 Median :6.000 Median :0.4596618   
 Mean :5.879 Mean :0.4570277   
 3rd Qu.:6.000 3rd Qu.:0.6859608   
 Max. :9.000 Max. :0.8998507

> summary(White\_wines.test)

fixed.acidity volatile.acidity citric.acid residual.sugar   
 Min. : 5.000 Min. :0.0800 Min. :0.0000 Min. : 0.800   
 1st Qu.: 6.400 1st Qu.:0.2175 1st Qu.:0.2600 1st Qu.: 2.100   
 Median : 6.800 Median :0.2600 Median :0.3200 Median : 6.300   
 Mean : 6.889 Mean :0.2766 Mean :0.3387 Mean : 6.866   
 3rd Qu.: 7.300 3rd Qu.:0.3200 3rd Qu.:0.3900 3rd Qu.:10.400   
 Max. :10.200 Max. :1.0050 Max. :0.8800 Max. :22.000   
 chlorides free.sulfur.dioxide total.sulfur.dioxide  
 Min. :0.01400 Min. : 2.0 Min. : 24.0   
 1st Qu.:0.03675 1st Qu.: 23.0 1st Qu.:108.0   
 Median :0.04300 Median : 35.0 Median :135.0   
 Mean :0.04612 Mean : 35.6 Mean :139.4   
 3rd Qu.:0.05000 3rd Qu.: 47.0 3rd Qu.:170.2   
 Max. :0.20400 Max. :124.0 Max. :260.0   
 density pH sulphates alcohol   
 Min. :0.9877 Min. :2.770 Min. :0.280 Min. : 8.40   
 1st Qu.:0.9918 1st Qu.:3.080 1st Qu.:0.400 1st Qu.: 9.40   
 Median :0.9941 Median :3.170 Median :0.480 Median :10.20   
 Mean :0.9942 Mean :3.174 Mean :0.496 Mean :10.45   
 3rd Qu.:0.9964 3rd Qu.:3.260 3rd Qu.:0.560 3rd Qu.:11.30   
 Max. :1.0010 Max. :3.690 Max. :1.010 Max. :13.90   
 quality group   
 Min. :3.000 Min. :0.9001   
 1st Qu.:5.000 1st Qu.:0.9229   
 Median :6.000 Median :0.9528   
 Mean :5.872 Mean :0.9506   
 3rd Qu.:6.000 3rd Qu.:0.9758   
 Max. :8.000 Max. :0.9993

> #I think it worked!

I will use the train data to fit the model and the test data to analysis the fit.

> RegModel.1 <-   
+ lm(quality~alcohol+chlorides+citric.acid+density+fixed.acidity+free.sulfur.dioxide+pH+residual.sugar+sulphates+total.sulfur.dioxide+volatile.acidity,  
+ data=White\_wines.train)  
> summary(RegModel.1)

Call:  
lm(formula = quality ~ alcohol + chlorides + citric.acid + density +   
 fixed.acidity + free.sulfur.dioxide + pH + residual.sugar +   
 sulphates + total.sulfur.dioxide + volatile.acidity, data = White\_wines.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.8642 -0.4973 -0.0362 0.4704 3.0782   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.552e+02 1.937e+01 8.013 1.42e-15 \*\*\*  
alcohol 1.885e-01 2.510e-02 7.510 7.10e-14 \*\*\*  
chlorides -2.444e-01 5.701e-01 -0.429 0.668114   
citric.acid 4.294e-02 1.010e-01 0.425 0.670887   
density -1.555e+02 1.965e+01 -7.917 3.06e-15 \*\*\*  
fixed.acidity 8.103e-02 2.176e-02 3.724 0.000199 \*\*\*  
free.sulfur.dioxide 4.064e-03 8.870e-04 4.581 4.74e-06 \*\*\*  
pH 7.268e-01 1.099e-01 6.614 4.19e-11 \*\*\*  
residual.sugar 8.492e-02 7.816e-03 10.865 < 2e-16 \*\*\*  
sulphates 6.578e-01 1.068e-01 6.156 8.10e-10 \*\*\*  
total.sulfur.dioxide -4.434e-04 3.963e-04 -1.119 0.263311   
volatile.acidity -1.822e+00 1.199e-01 -15.199 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7538 on 4426 degrees of freedom  
Multiple R-squared: 0.2805, Adjusted R-squared: 0.2787   
F-statistic: 156.9 on 11 and 4426 DF, p-value: < 2.2e-16

So it looks like alcohol, chlorides, density, fixed acidity, pH, residual. sugar, sulphates, and volatile acidity are all significantly associated with predicting quality.  
Run the model again taking out the non-significant variables:

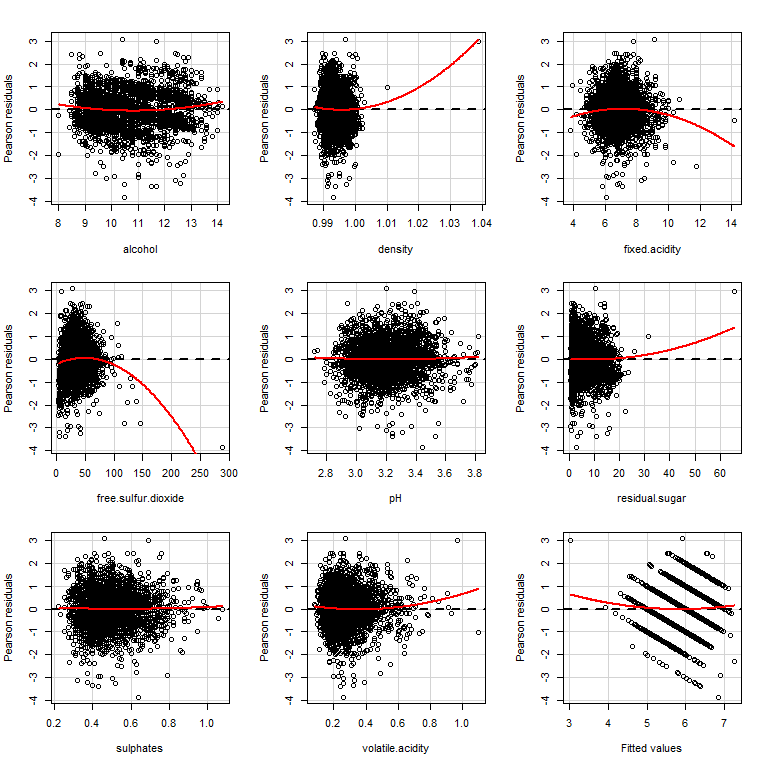
> RegModel.2 <-   
+ lm(quality~alcohol+density+fixed.acidity+free.sulfur.dioxide+pH+residual.sugar+sulphates++volatile.acidity,  
+ data=White\_wines.train)  
> summary(RegModel.2)

Call:  
lm(formula = quality ~ alcohol + density + fixed.acidity + free.sulfur.dioxide +   
 pH + residual.sugar + sulphates + +volatile.acidity, data = White\_wines.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.8536 -0.4930 -0.0388 0.4675 3.0889   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.598e+02 1.872e+01 8.535 < 2e-16 \*\*\*  
alcohol 1.888e-01 2.495e-02 7.566 4.64e-14 \*\*\*  
density -1.603e+02 1.898e+01 -8.445 < 2e-16 \*\*\*  
fixed.acidity 8.386e-02 2.133e-02 3.931 8.58e-05 \*\*\*  
free.sulfur.dioxide 3.487e-03 7.137e-04 4.885 1.07e-06 \*\*\*  
pH 7.325e-01 1.078e-01 6.792 1.25e-11 \*\*\*  
residual.sugar 8.639e-02 7.594e-03 11.377 < 2e-16 \*\*\*  
sulphates 6.524e-01 1.064e-01 6.130 9.57e-10 \*\*\*  
volatile.acidity -1.861e+00 1.152e-01 -16.150 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7537 on 4429 degrees of freedom  
Multiple R-squared: 0.2802, Adjusted R-squared: 0.2789   
F-statistic: 215.5 on 8 and 4429 DF, p-value: < 2.2e-16

In this model the adjusted R-squared is 0.28 (indicating 28% of the variability in average quality rating is due to the variables included), exactly the same as the model containing all variables and the F-statistic is highly significant.

Some diagnostic to ensure fit:

> residualPlots(RegModel.2)



Test stat Pr(>|t|)  
alcohol 5.191 0.000  
density 5.552 0.000  
fixed.acidity -4.163 0.000  
free.sulfur.dioxide -10.160 0.000  
pH 0.880 0.379  
residual.sugar 2.520 0.012  
sulphates 0.729 0.466  
volatile.acidity 3.184 0.001  
Tukey test 2.551 0.011

Now there are two additional variables that have a non-zero trend, pH and sulphates. I will investigate what happens when they are removed from the model.

> RegModel.3 <-   
+ lm(quality~alcohol+density+fixed.acidity+free.sulfur.dioxide+residual.sugar+volatile.acidity,  
+ data=White\_wines.train)  
> summary(RegModel.3)

Call:  
lm(formula = quality ~ alcohol + density + fixed.acidity + free.sulfur.dioxide +   
 residual.sugar + volatile.acidity, data = White\_wines.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.8254 -0.5078 -0.0256 0.4653 3.2095   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.566e+01 1.430e+01 3.893 0.000100 \*\*\*  
alcohol 3.157e-01 2.006e-02 15.738 < 2e-16 \*\*\*  
density -5.308e+01 1.429e+01 -3.714 0.000206 \*\*\*  
fixed.acidity -2.921e-02 1.575e-02 -1.855 0.063709 .   
free.sulfur.dioxide 3.940e-03 7.178e-04 5.489 4.27e-08 \*\*\*  
residual.sugar 4.381e-02 5.732e-03 7.643 2.58e-14 \*\*\*  
volatile.acidity -1.995e+00 1.151e-01 -17.327 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7603 on 4431 degrees of freedom  
Multiple R-squared: 0.2673, Adjusted R-squared: 0.2663   
F-statistic: 269.3 on 6 and 4431 DF, p-value: < 2.2e-16

In this new model, 27% of the variation is explained by the variables included and of note, fixed acidity does not reach significance. The F-statistic remains significant, however. What happens when fixed acidity is removed?

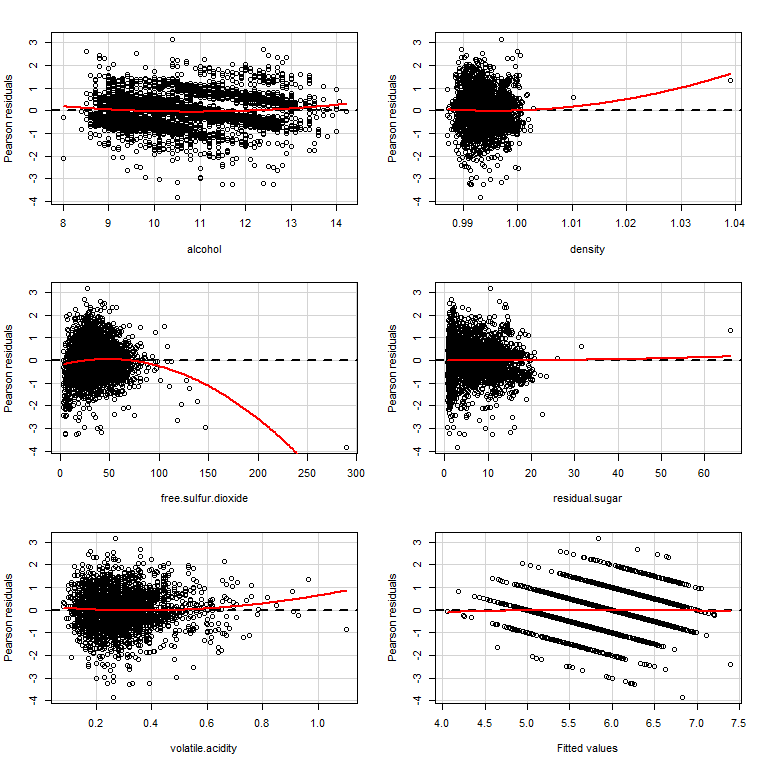
> RegModel.4 <-   
+ lm(quality~alcohol+density+free.sulfur.dioxide+residual.sugar+volatile.acidity,  
+ data=White\_wines.train)  
> summary(RegModel.4)

Call:  
lm(formula = quality ~ alcohol + density + free.sulfur.dioxide +   
 residual.sugar + volatile.acidity, data = White\_wines.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.8289 -0.5071 -0.0354 0.4676 3.1637   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 6.867e+01 1.246e+01 5.512 3.76e-08 \*\*\*  
alcohol 3.016e-01 1.857e-02 16.242 < 2e-16 \*\*\*  
density -6.626e+01 1.240e+01 -5.343 9.60e-08 \*\*\*  
free.sulfur.dioxide 4.055e-03 7.153e-04 5.669 1.53e-08 \*\*\*  
residual.sugar 4.825e-02 5.211e-03 9.260 < 2e-16 \*\*\*  
volatile.acidity -1.979e+00 1.148e-01 -17.231 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7605 on 4432 degrees of freedom  
Multiple R-squared: 0.2667, Adjusted R-squared: 0.2659   
F-statistic: 322.4 on 5 and 4432 DF, p-value: < 2.2e-16

Now all the variables included are significant, the F-statistic remains significant and the R-squared is only lowered a fraction of a percent.

Now I will run diagnostics for the updated model.

> residualPlots(RegModel.4)



Test stat Pr(>|t|)  
alcohol 4.382 0.000  
density 2.609 0.009  
free.sulfur.dioxide -10.280 0.000  
residual.sugar 0.284 0.776  
volatile.acidity 3.068 0.002  
Tukey test -0.808 0.419

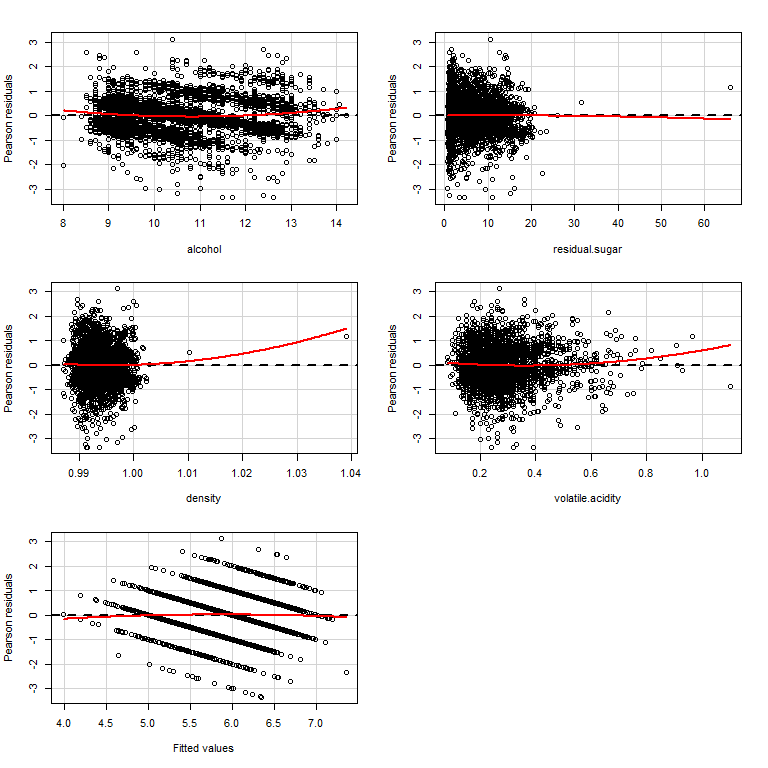
Now free sulfur dioxide has a non-zero trend. Can they be removed from the model as well?

> RegModel.5 <-   
+ lm(quality~alcohol+residual.sugar+density+volatile.acidity,  
+ data=White\_wines.train)  
> summary(RegModel.5)

Call:  
lm(formula = quality ~ alcohol + residual.sugar + density + volatile.acidity,   
 data = White\_wines.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.3451 -0.5064 -0.0326 0.4750 3.1222   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 72.074031 12.488790 5.771 8.41e-09 \*\*\*  
alcohol 0.290592 0.018531 15.681 < 2e-16 \*\*\*  
residual.sugar 0.052831 0.005166 10.227 < 2e-16 \*\*\*  
density -69.433087 12.432670 -5.585 2.48e-08 \*\*\*  
volatile.acidity -2.050559 0.114546 -17.902 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.7632 on 4433 degrees of freedom  
Multiple R-squared: 0.2614, Adjusted R-squared: 0.2607   
F-statistic: 392.2 on 4 and 4433 DF, p-value: < 2.2e-16

It seems as if it can. The model without density and free sulfur dioxide has an multiple R-squared of 0.26, slightly lower but nominally so. The F-statistic and all variables included remain significant. Now to check the residuals...

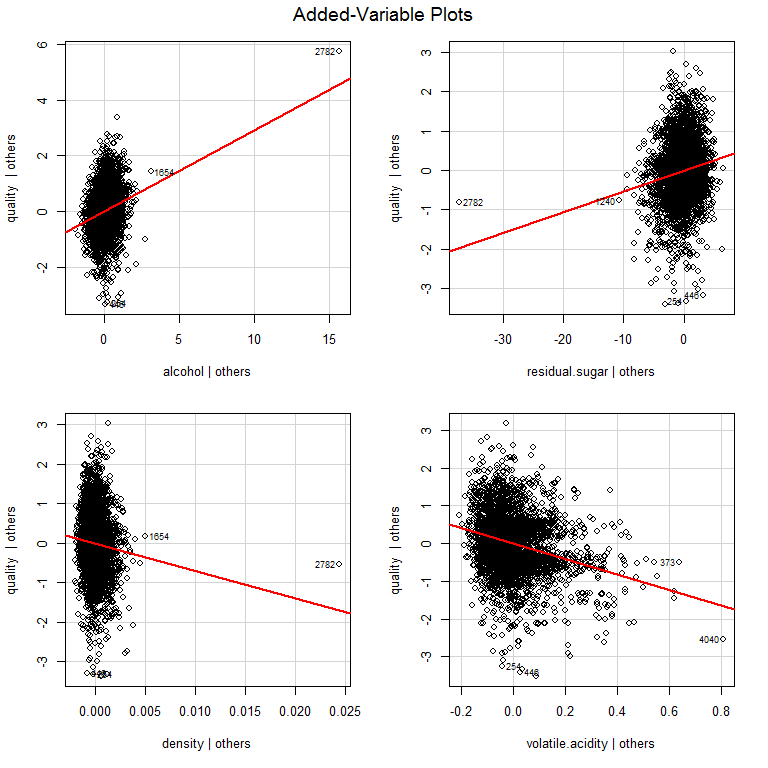
> residualPlots(RegModel.5)



Test stat Pr(>|t|)  
alcohol 4.649 0.000  
residual.sugar -0.273 0.785  
density 2.394 0.017  
volatile.acidity 2.884 0.004  
Tukey test -1.352 0.177

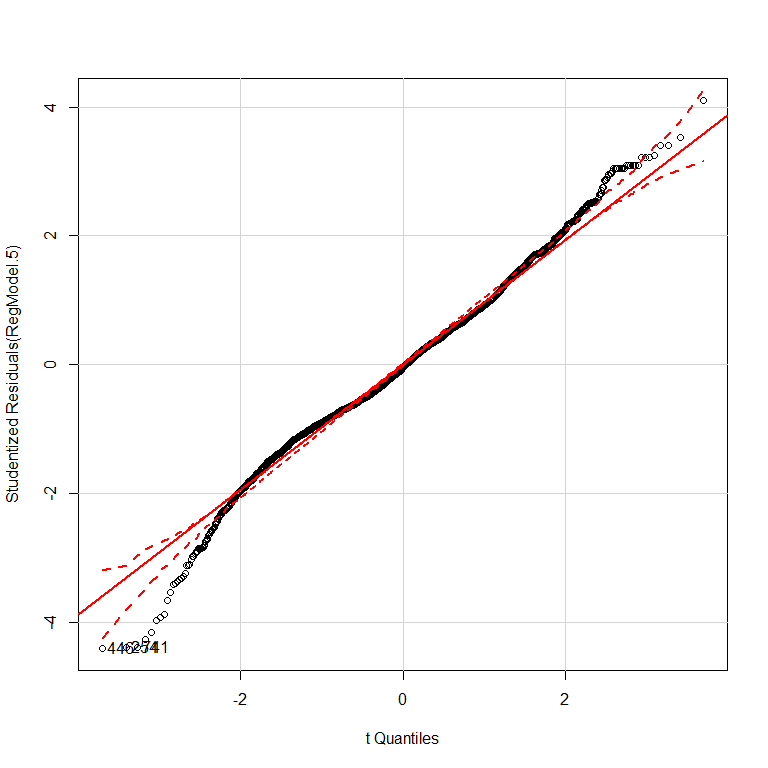
This seems to fit well. Lets continue with assessing fit.

> #added variable plots  
> avPlots(RegModel.5, id.n=2, id.cex=0.7)



> #id.n - identify n most influential observations  
> #id.cex - controls the size of the dot

> # run the qq-plot  
> qqPlot(RegModel.5, id.n=3)



446 254 741   
 1 2 3

> # here, id.n identifies the n observations with the largest residuals in absolute value

The distribution may have a slight left skew but is overall normal. Hooray!

> #run Bonferroni test for outliers  
> outlierTest(RegModel.5)

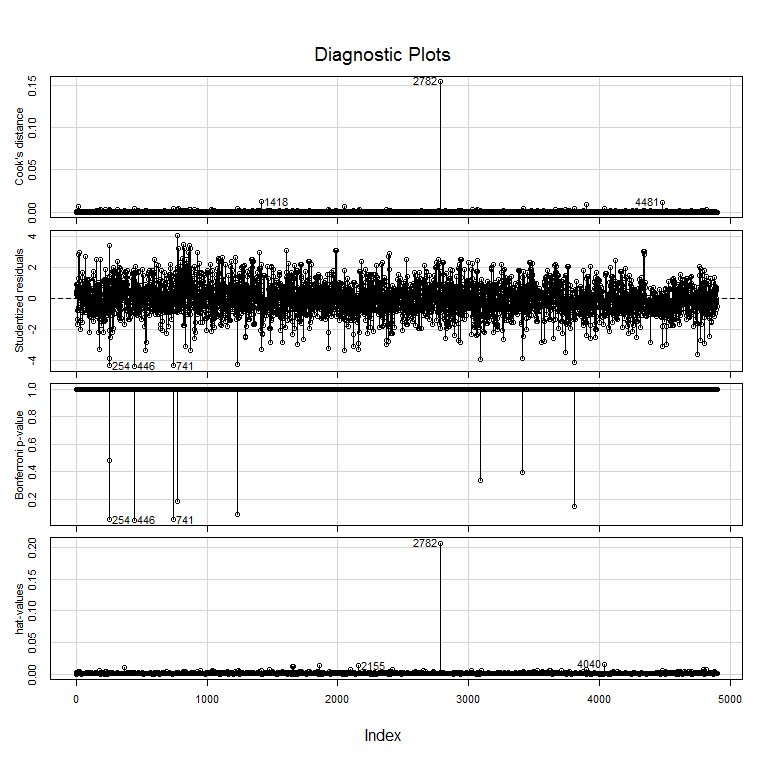
No Studentized residuals with Bonferonni p < 0.05  
Largest |rstudent|:  
 rstudent unadjusted p-value Bonferonni p  
446 -4.394467 1.1364e-05 0.050434

> outlierTest(RegModel.4)

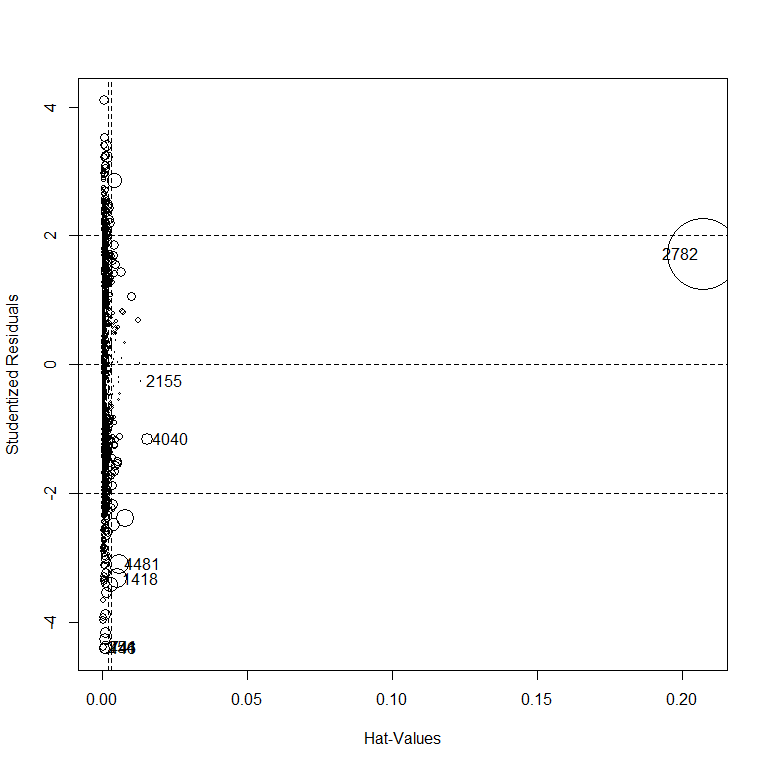
rstudent unadjusted p-value Bonferonni p  
4746 -5.204274 2.0356e-07 0.00090341

The Bonferonni p-value remains significant, indicating the model is sound even taking into account the significant error of using multiple predictor variables.

> #identify highly influential points  
> influenceIndexPlot(RegModel.5, id.n=3)

 There do appear to be some influential outliers, especially 2782 based on a quick review.

> #make influence plot  
> influencePlot(RegModel.5, id.n=3)



StudRes Hat CookD  
254 -4.3835228 0.0004920295 0.0018840852  
446 -4.3944671 0.0010092450 0.0038858618  
741 -4.3827147 0.0012170709 0.0046620936  
1418 -3.3202553 0.0053967693 0.0119364728  
2155 -0.2558756 0.0135328514 0.0001796742  
2782 1.7193861 0.2071723881 0.1544323593  
4040 -1.1558691 0.0155981988 0.0042336643  
4481 -3.0996214 0.0060284283 0.0116314800

> # diameter is related to cooks distance

Again observation 2782 appears to be a largely influential outlier. I would consider taking this point out and seeing if the model still fits.

It does appear that removing observation 2782 does make a difference when I took it out. IT was placed back in as to not affect the model. Now there are more outliers but the significance of the point is removed without greatly affecting the F-statistic and p-value.

Now I will use a Breusch–Pagan test for heteroscedasticity.

> #test for heteroskedasticity  
> ncvTest(RegModel.5)

Non-constant Variance Score Test   
Variance formula: ~ fitted.values   
Chisquare = 31.97253 Df = 1 p = 1.563686e-08

This indicates (I think) that overall the data has heteroskedasticity.

Finally I will look at the degree of multicollinearity present among variables. I was initially hypothesizing there was at least some multicollinearity between some variables simple based on their names and the distribution of data. I will first run Model #2 to see how much was present when all significant variables were included in the model.

> vif(RegModel.2)

alcohol density fixed.acidity   
 7.337053 25.278419 2.545179   
free.sulfur.dioxide pH residual.sugar   
 1.151922 2.083100 11.620123   
 sulphates volatile.acidity   
 1.126483 1.060061

Hypothesis confirmed. There was defiantly multicolleniarty issues in the first model. Now let me look at the final model.

> vif(RegModel.5)

alcohol residual.sugar density volatile.acidity   
 3.946303 5.244723 10.583613 1.021998

Better! There is still a sight issues but on the whole they seem to be independent variables.

So overall the model building process has showing we can use the variables percent alcohol, the amount of residual sugar, wine density and wine volatile acidity to predict the level of quality (or for our purposes preference) that will be given to a certain wine by a drinker. I chose these specific variables through comparing multiple models for the best fit while maintaining parsimony.