Exploratory Data Analysis on White Wines dataset

Eric Shell

11 April 2018

Overview of the goals of this investigation

This report will investigate the possible physical and chemical factors that may contribute to the wine's judge quality.

The dataset is available here (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityWhites.csv)

In this investigation we will look into these questions:

1. We want to look at the distribution of the raw variables. Are they skewed, or instead are they roughly symmetric and normal or bell shaped?
2. Can we say that increases or decreases in the level of a variable are associated with an increase or decrease in quality?
If so, is there a strong or orderly relationship? Does our picture we get in our charts match the picture we get from correlation coefficients? Finally, is the relationship monotic, or is it instead peaked with a neighborhood having highest quality levels?
3. When looking at how the variables relate to each other, can we see some variables moderating the effect that another variable has on quality?
4. We want to know if our intuitions reached from investigating the above questions are born out in a linear regression model. Do the variables that appear to be the most related to quality (graphically, in plots) also have a relatively high amount of variance explained in the linear regression model? And do the variables that we think interact in the ways indicated above also have a correspondingly important interaction term in the regression model?

Brief overview of the dataset

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 ...
```

The dataset has these variables seen from the summary, in their order of appearance above:

x: an index which is ignored

fixed acidity (tartaric acid - g / dm^3): most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

volatile acidity (acetic acid - g / dm^3): the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

citric acid (g / dm^3): found in small quantities, citric acid can add 'freshness' and flavor to wines

residual sugar (g / dm^3): the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

chlorides (sodium chloride - g / dm^3: the amount of salt in the wine

free sulfur dioxide (mg / dm^3): the free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

total sulfur dioxide (mg / dm^3): amount of free and bound forms of SO_2; in low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine

density (g / cm^3): the density of water is close to that of water depending on the percent alcohol and sugar content

pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

sulphates (potassium sulphate - g / dm3): a wine additive which can contribute to sulfur dioxide gas (SO_2) levels, which acts as an antimicrobial and antioxidant

alcohol (% by volume): the percent alcohol content of the wine

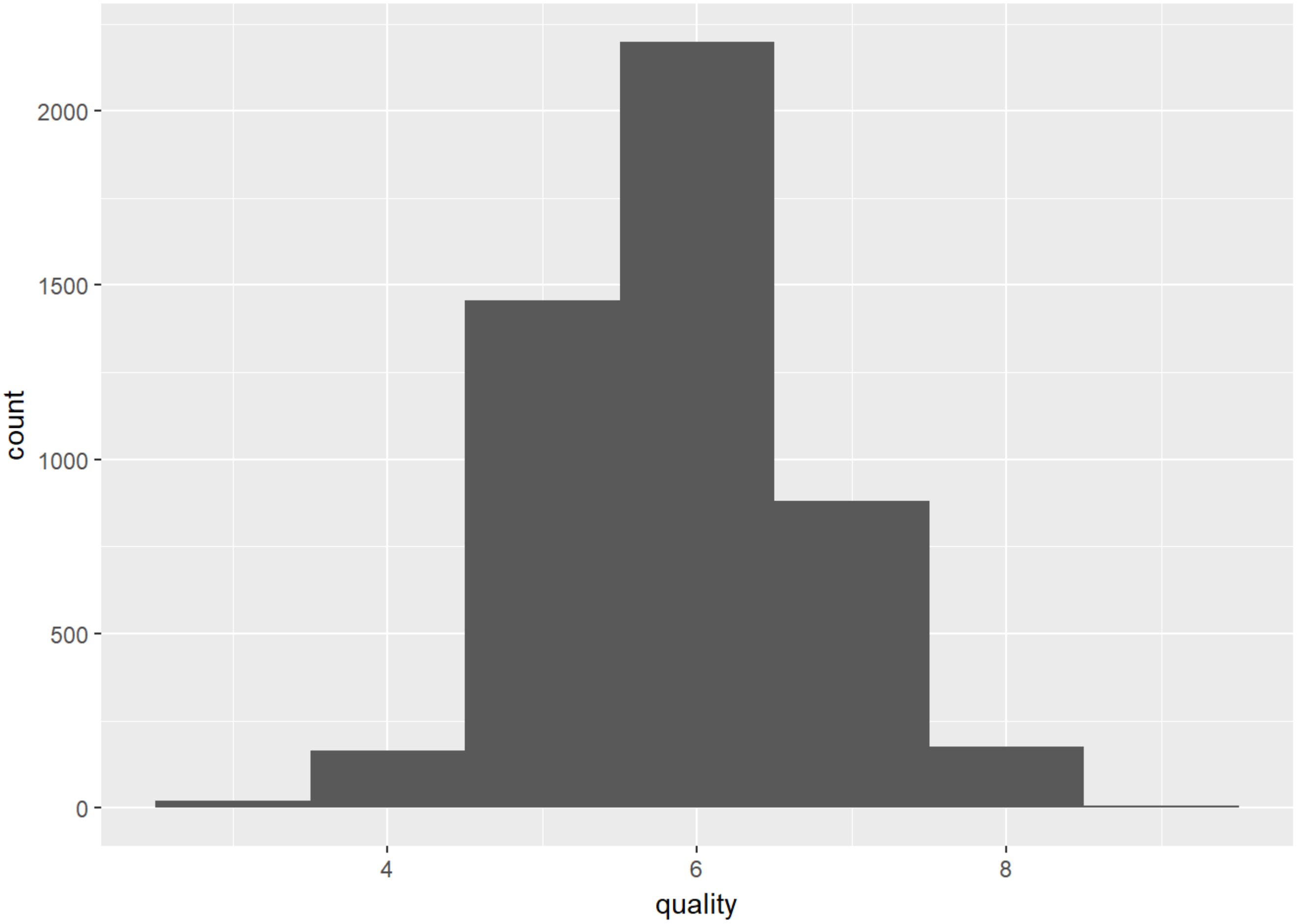
quality: score between 0 and 10

As can be seen above, all these variables of interest are numerical measures of chemical properties of the wine, aside from the quality rating itself which is a subjective 0 - 10 rating.

Univariate Plots Section

Here we look at the histograms and distribution characteristics for the different variables. We start off by looking at the main variable of interest, quality.

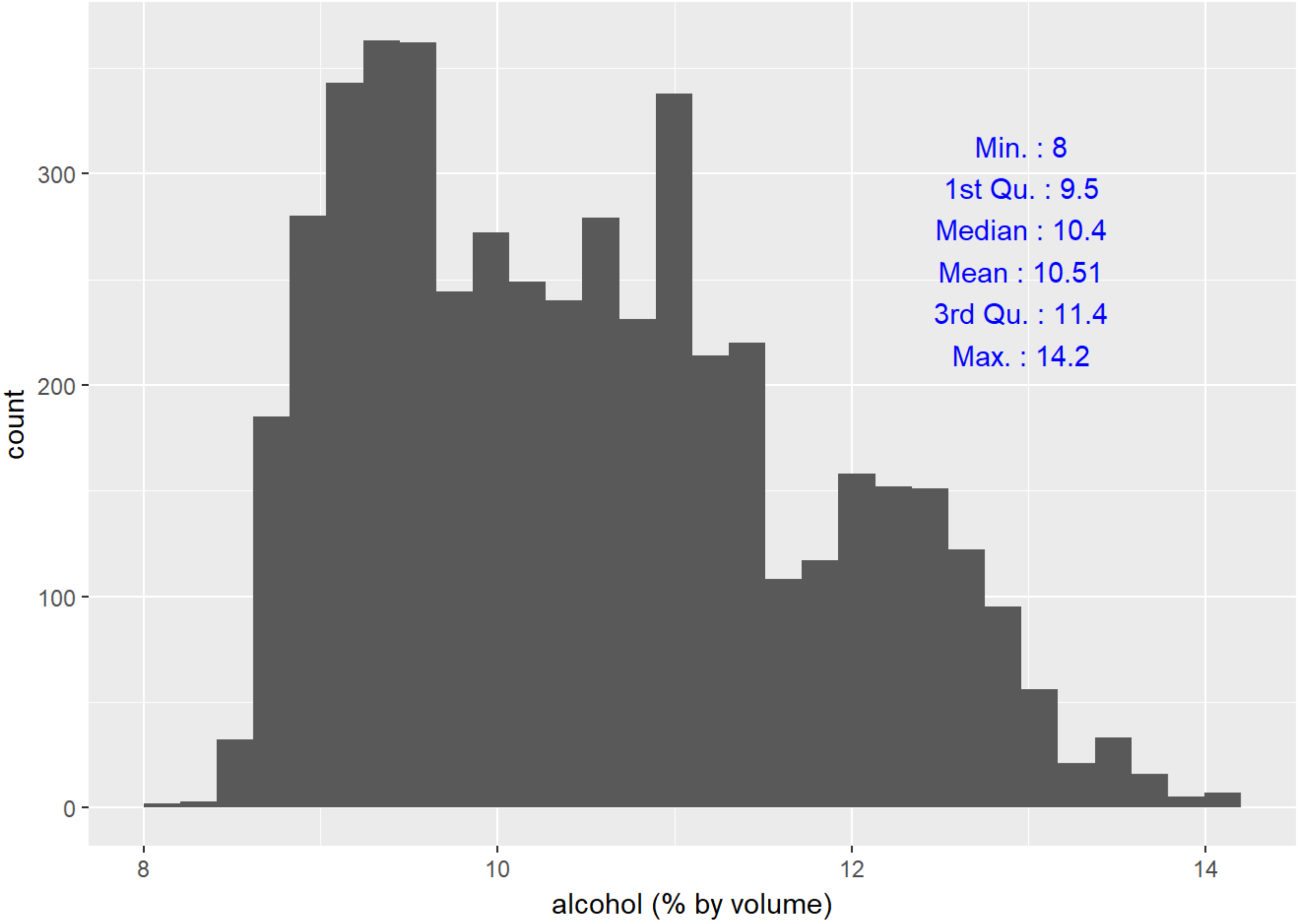
We can see below that the distribution of quality ratings is roughly symmetric, and the typical values range between 5, 6, and 7, with less rated 4 or 8 and a few outliers rated at 3 or 9. It’s a little curious that there are so few rated above 8 or below 4, and with none rated as 0, 1, or 2.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.000   5.000   6.000   5.878   6.000   9.000
```

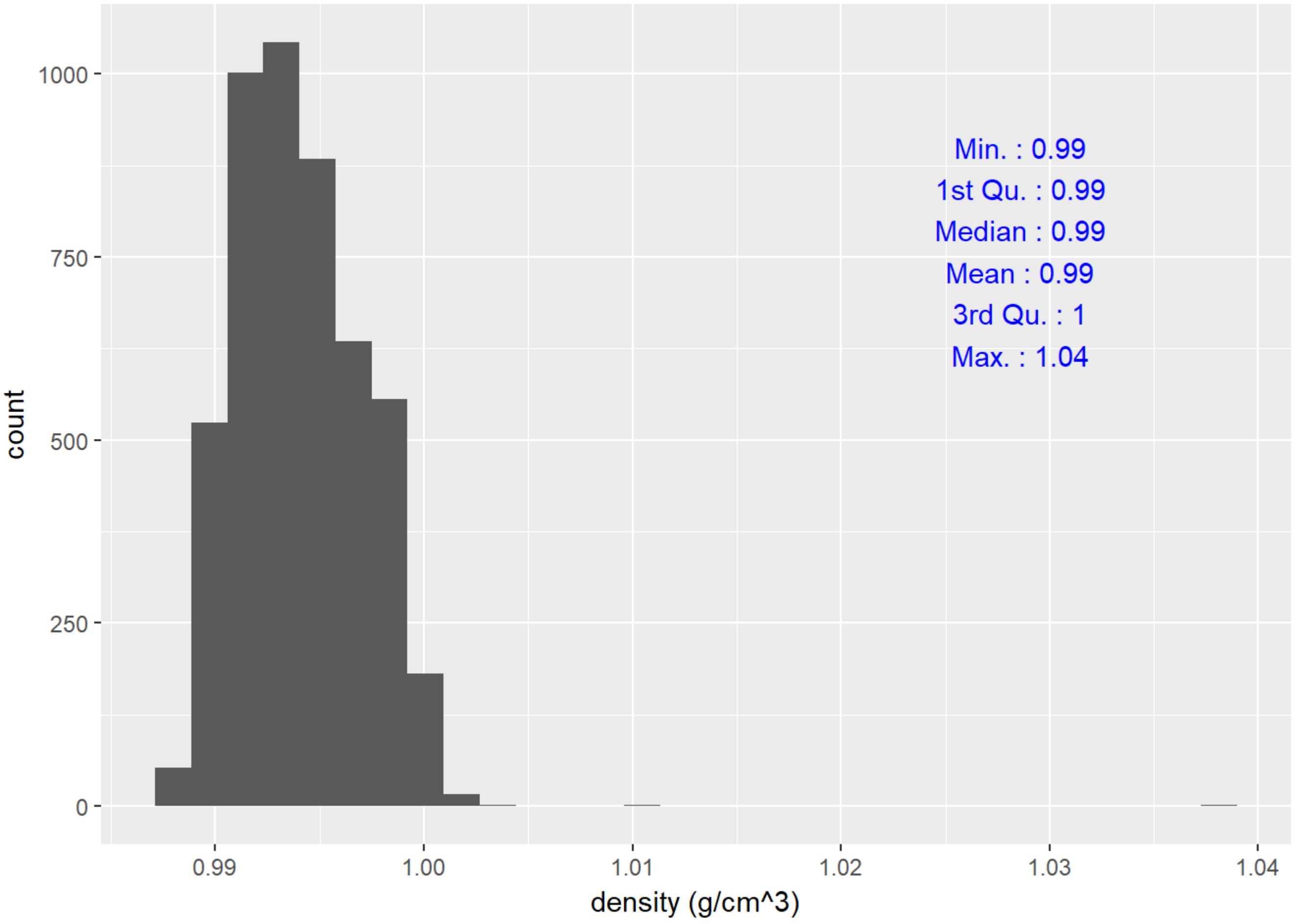
```
## 
##    3    4    5    6    7    8    9 
##   20  163 1457 2198  880  175    5
```

Next, we look at the alcohol level, which again is percent alcohol of the wine.



We see a moderately right-skewed distribution for the alcohol content, with a spike at around 8.5 to 9.5%, with the frequencies falling off after that. There are hints of a bi-modal distribution with another mini-spike at around 12%, possibly reflecting a recognizable type of white wine (say, of white zinfandel) that typically contains more alcohol. (Note, however, that according to the data source, the types of white wine like this are not included in the analysis due to copyright reasons.)

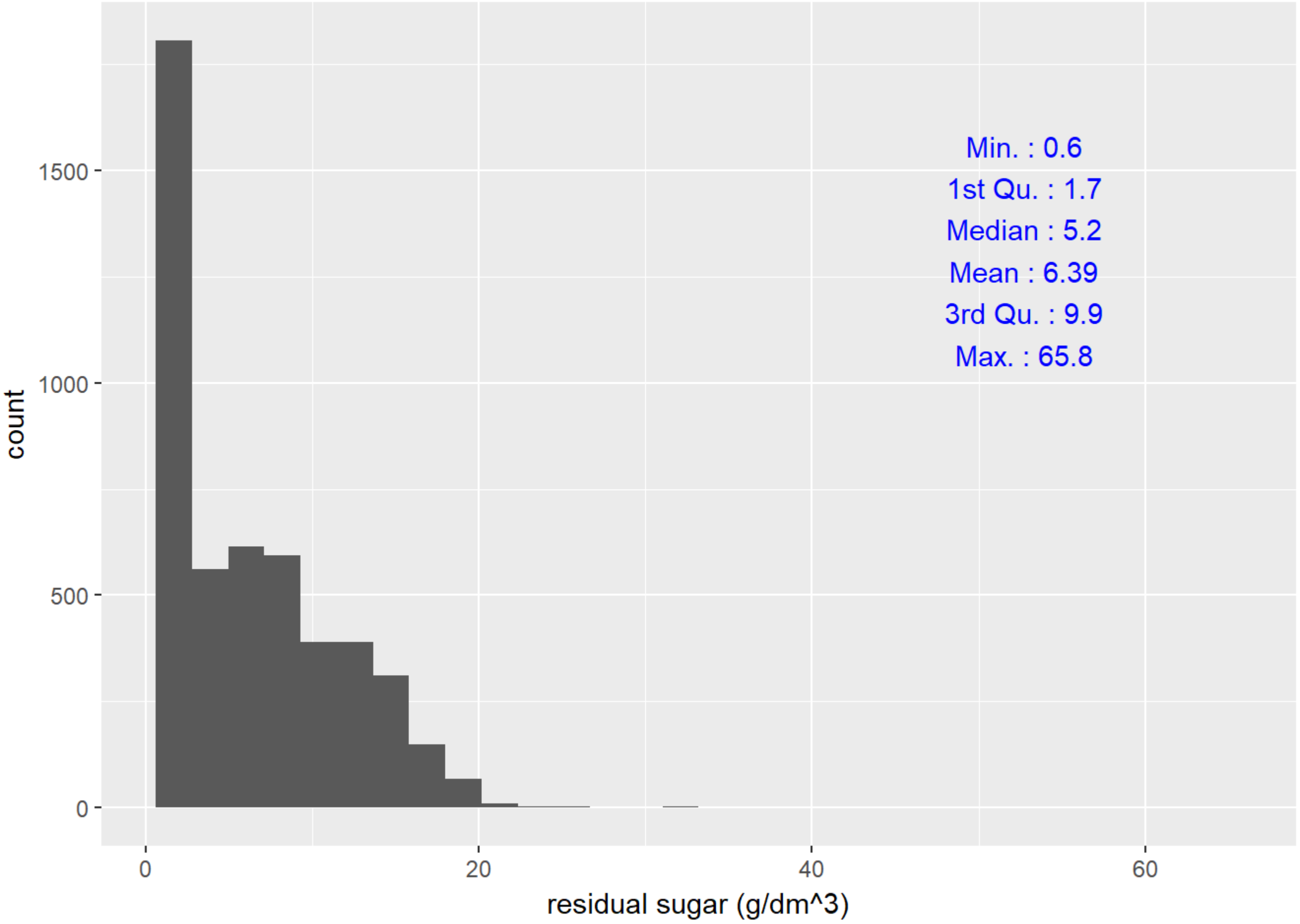
Next, we take a look at the density.



```
## 
## [0.985,0.995]     (0.995,1]      (1,1.01]   (1.01,1.02]   (1.02,1.03] 
##          3098          1797             2             0             0 
##   (1.03,1.04] 
##             1 
```

The density has a symmetric, normal or bell-shaped distribution, with just 3 outliers past the density of 1.0 which we take a note of here.

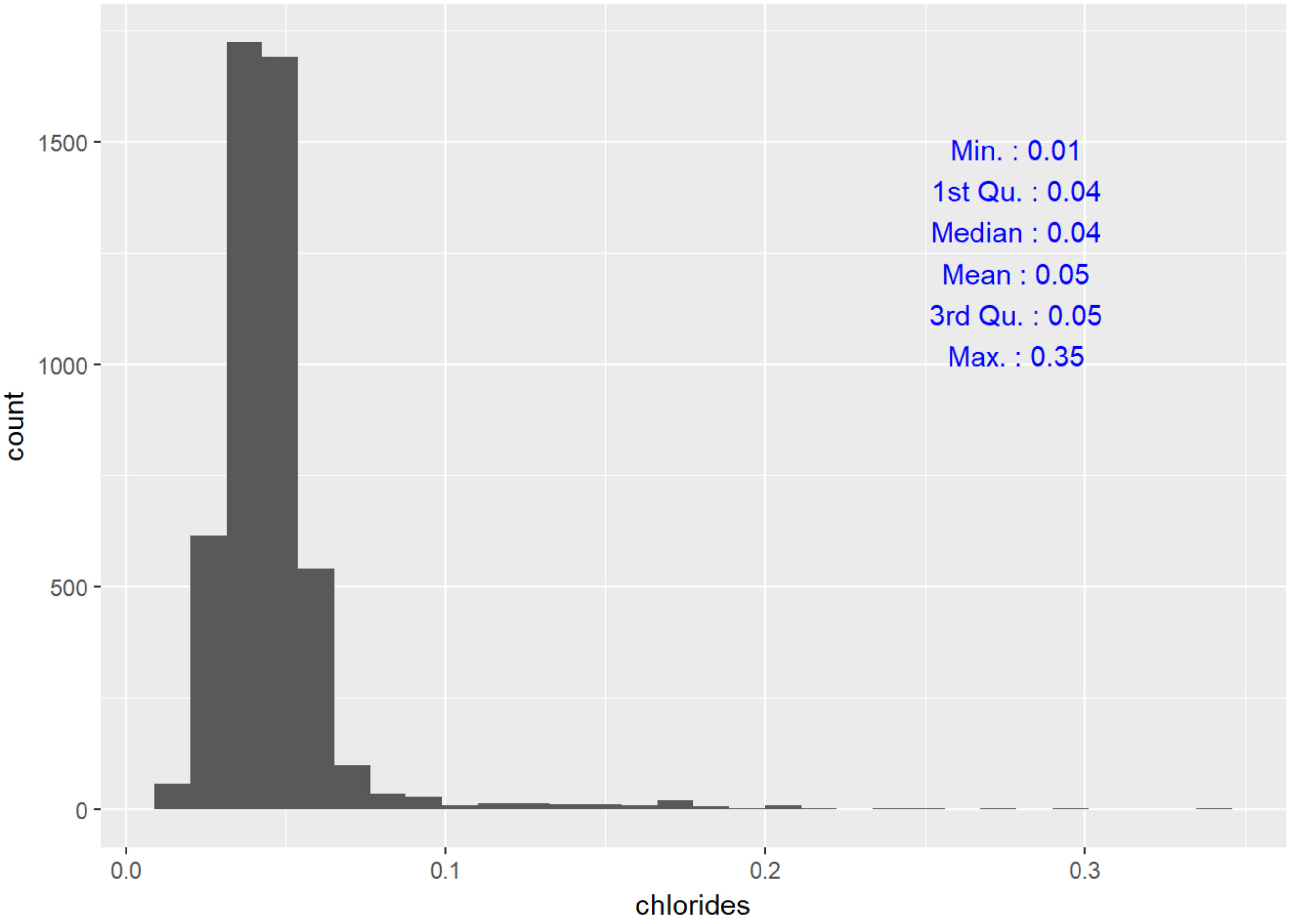
We now look at residual sugar level.



```
##
##  [-5,5]  (5,15] (15,25] (25,35] (35,45] (45,55] (55,65] (65,75]
##    2410    2167     316       4       0       0       0       1
```

Similar to density, the distribution for residual sugar is also heavily right skewed with just 5 outliers past the typical value of 25 g/dm^3. Rather than being almost symmetric, this distribution looks like the frequency is a decreasing function of the sugar level, with the most common level being very low.

Next, we look at chlorides

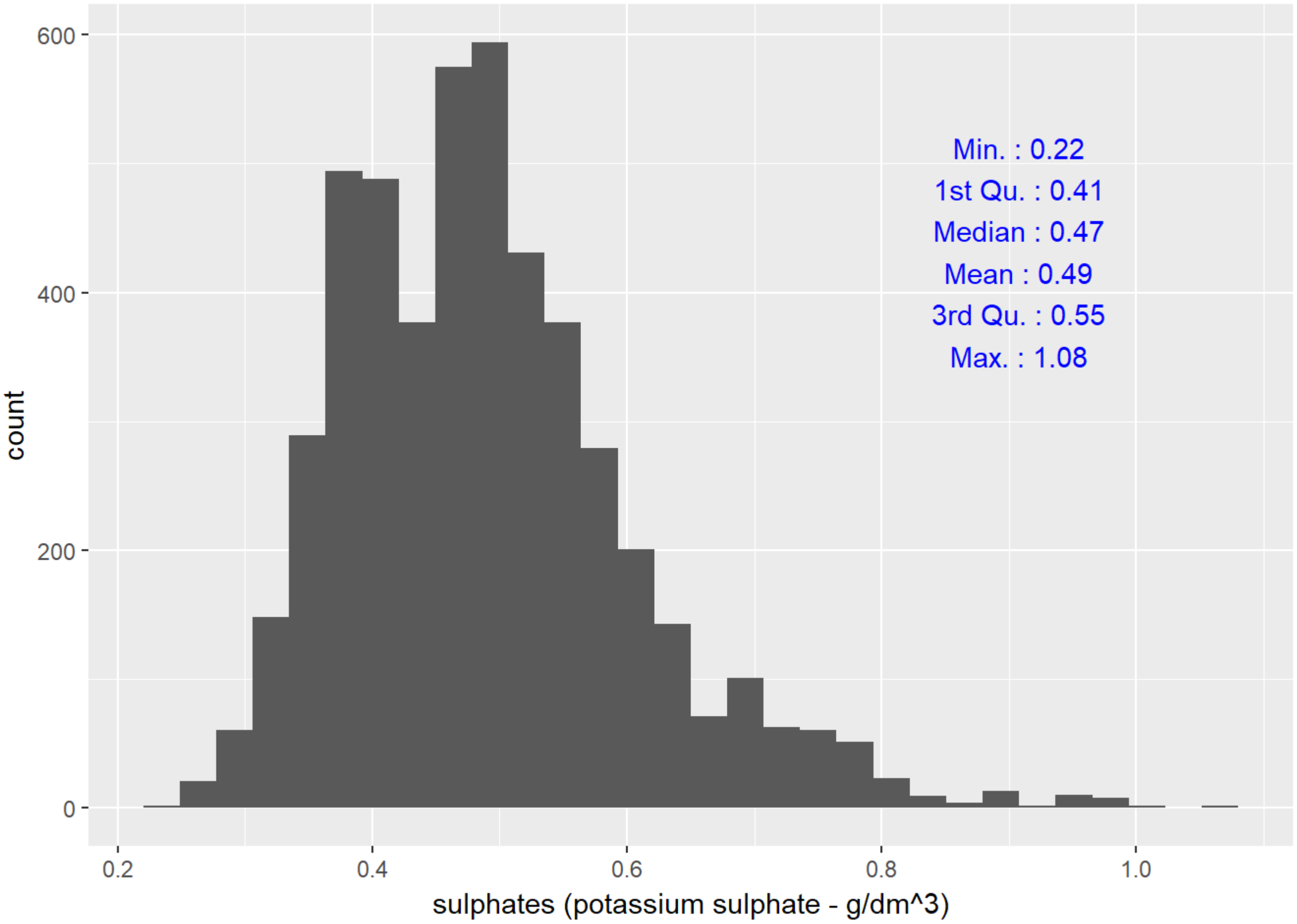


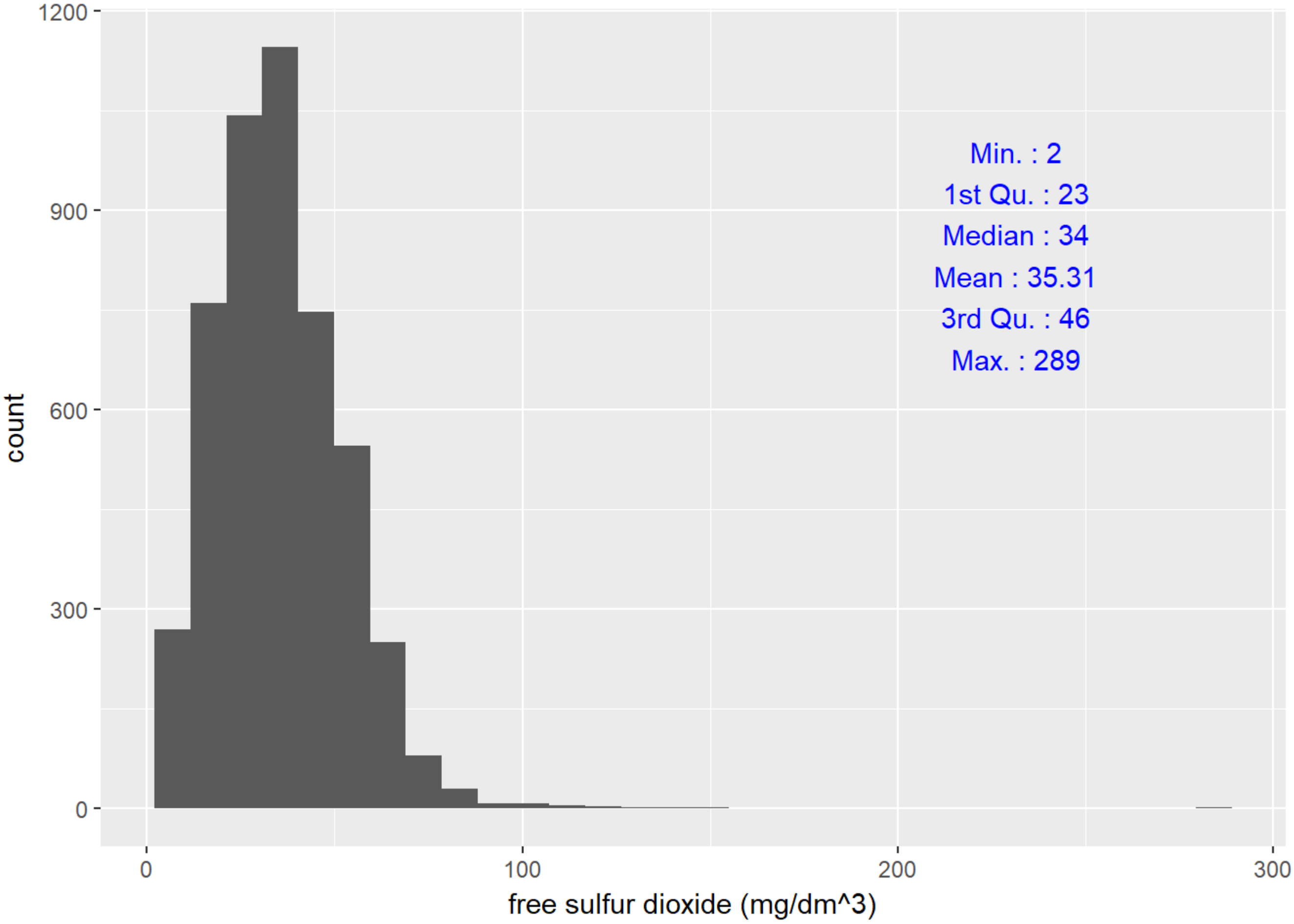
```
##
## [-0.025,0.025]  (0.025,0.075]  (0.075,0.125]  (0.125,0.175]  (0.175,0.225]
##            178           4536            100             54             22
##  (0.225,0.275]  (0.275,0.325]  (0.325,0.375]
##              5              2              1
```

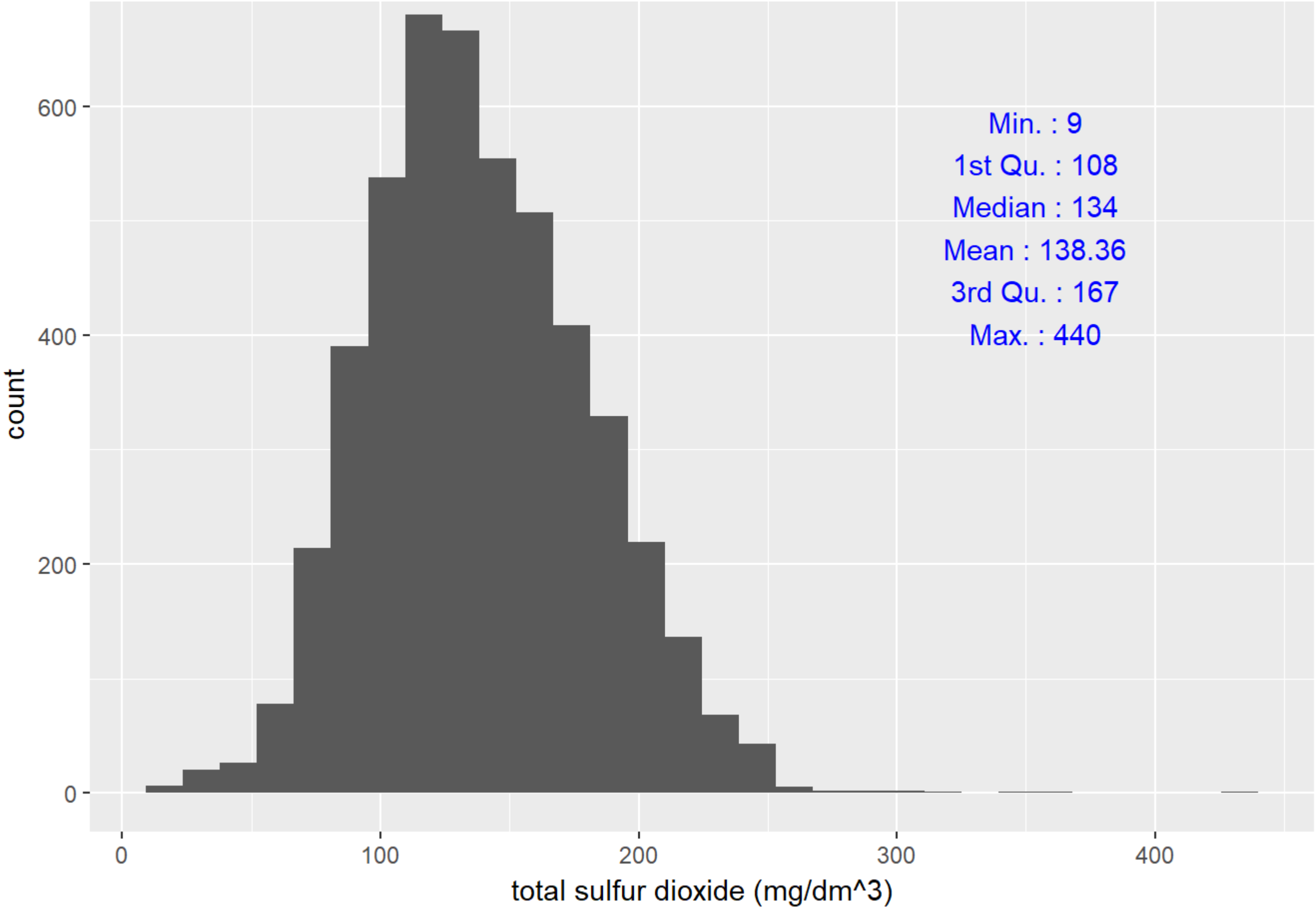
The chlorides frequency distribution is roughly symmetrical and normal shaped, with a heavy right tail having frequencies gradually dropping off as the chlorides rate increases, as opposed to just a few outliers after a cut-off point like was seen with density and sugar level.

To simplify, we look at the variables relating to sulfates together below.

As mentioned above, sulfates are added for quality control to reduce microbial growth, but in too large of quantities it can be noticeable to the taste and smell.

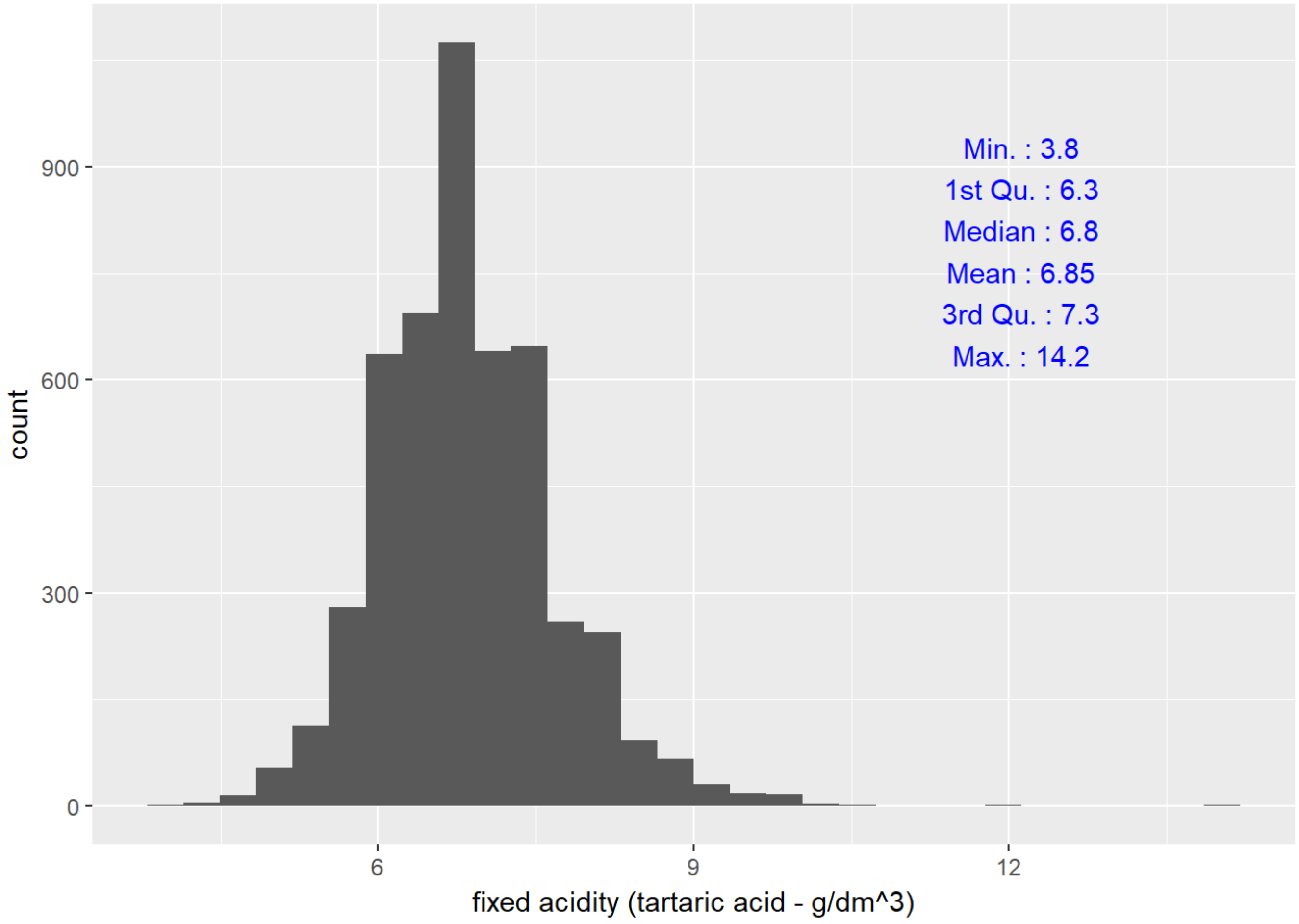

Min. : 0.22
1st Qu. : 0.41
Median : 0.47
Mean : 0.49
3rd Qu. : 0.55
Max. : 1.08
count
600
400
200
0
0.2
0.4
0.6
0.8
1.0
sulphates (potassium sulphate - g/dm^3)

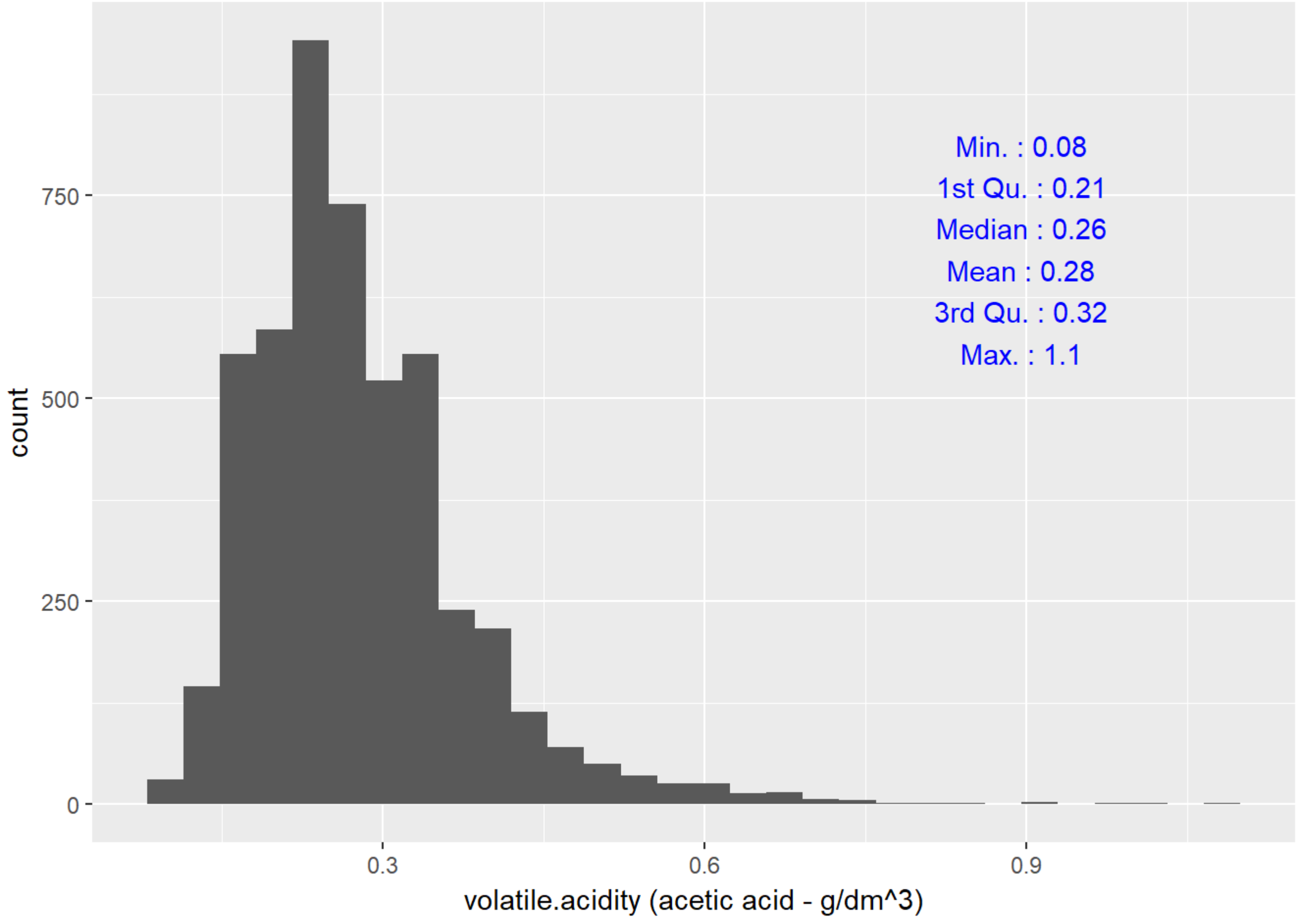

Min. : 2
1st Qu. : 23
Median : 34
Mean : 35.31
3rd Qu. : 46
Max. : 289
count
1200
900
600
300
0
0
100
200
300
free sulfur dioxide (mg/dm^3)



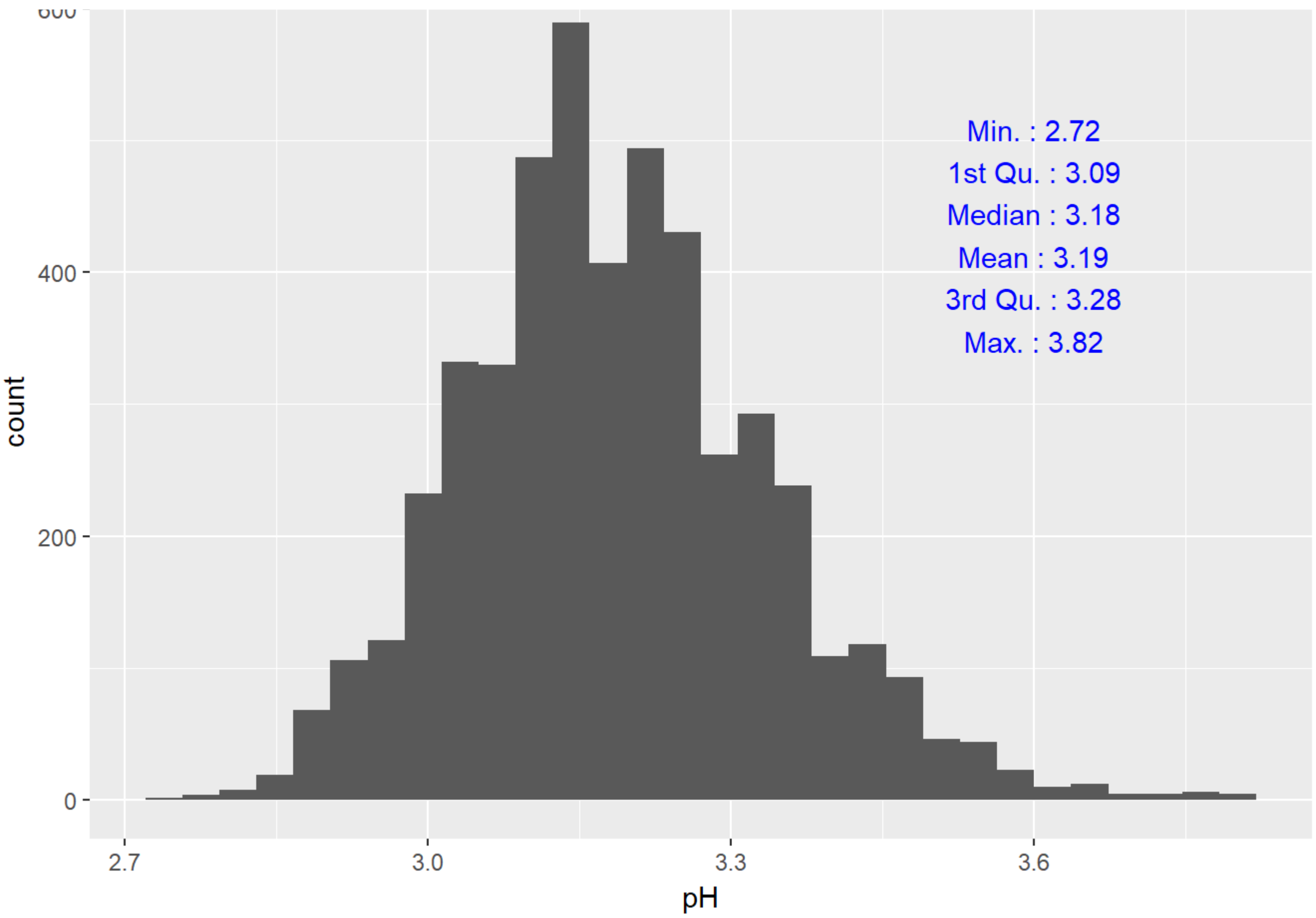
These distributions all look relatively symmetric and normal (bell-shaped).
There appear to be outliers well outside the typical range for free sulfur dioxide, but not so much for total sulfur dioxide, and no significant amount of outliers for the sulfates level itself.

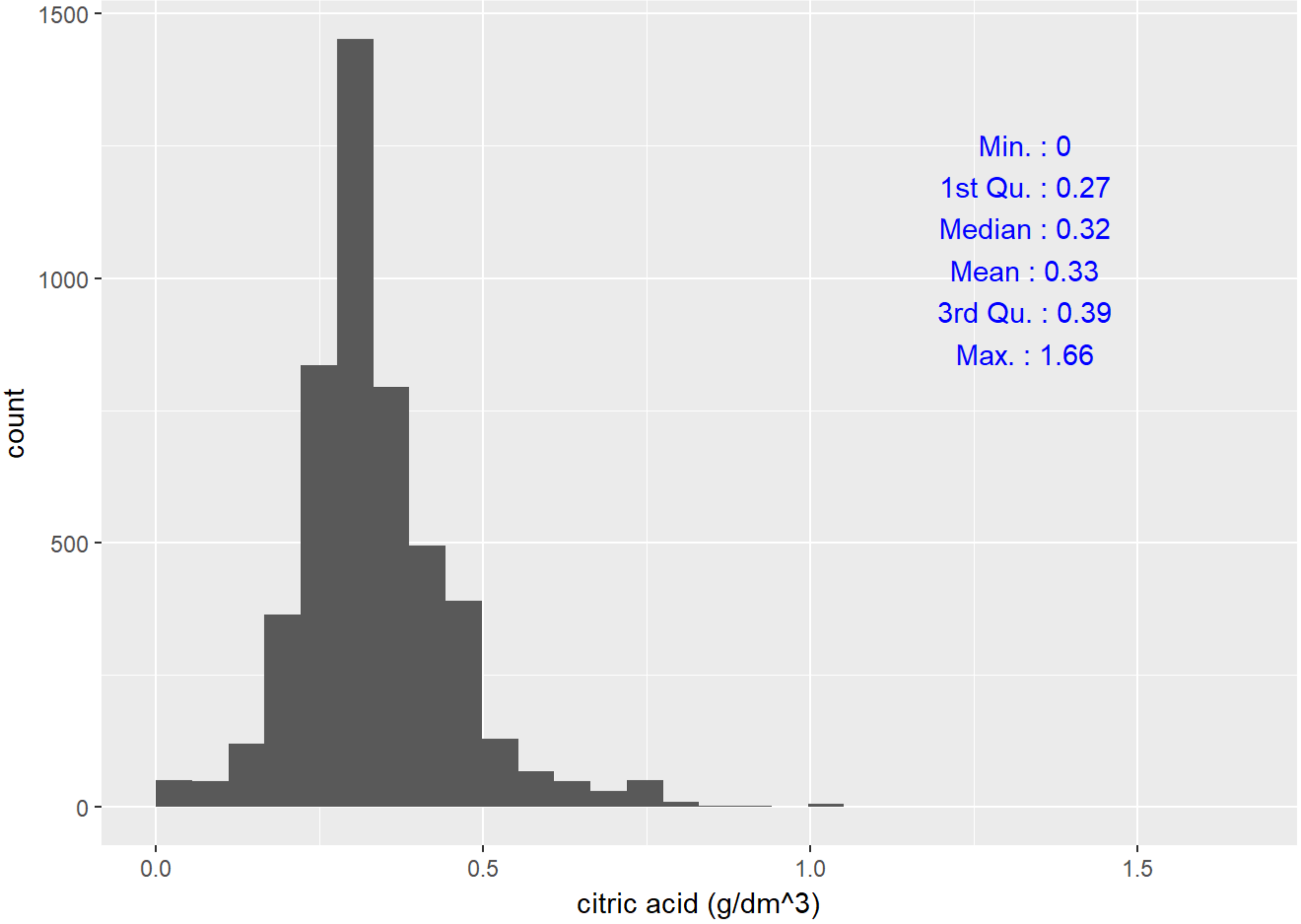
We now look at the variables associated with acidity together, again to simplify the analysis.


Min. : 3.8
1st Qu. : 6.3
Median : 6.8
Mean : 6.85
3rd Qu. : 7.3
Max. : 14.2
count
900
600
300
0
6
9
12
fixed acidity (tartaric acid - g/dm^3)


Min. : 0.08
1st Qu. : 0.21
Median : 0.26
Mean : 0.28
3rd Qu. : 0.32
Max. : 1.1
count
750
500
250
0
0.3
0.6
0.9
volatile.acidity (acetic acid - g/dm^3)

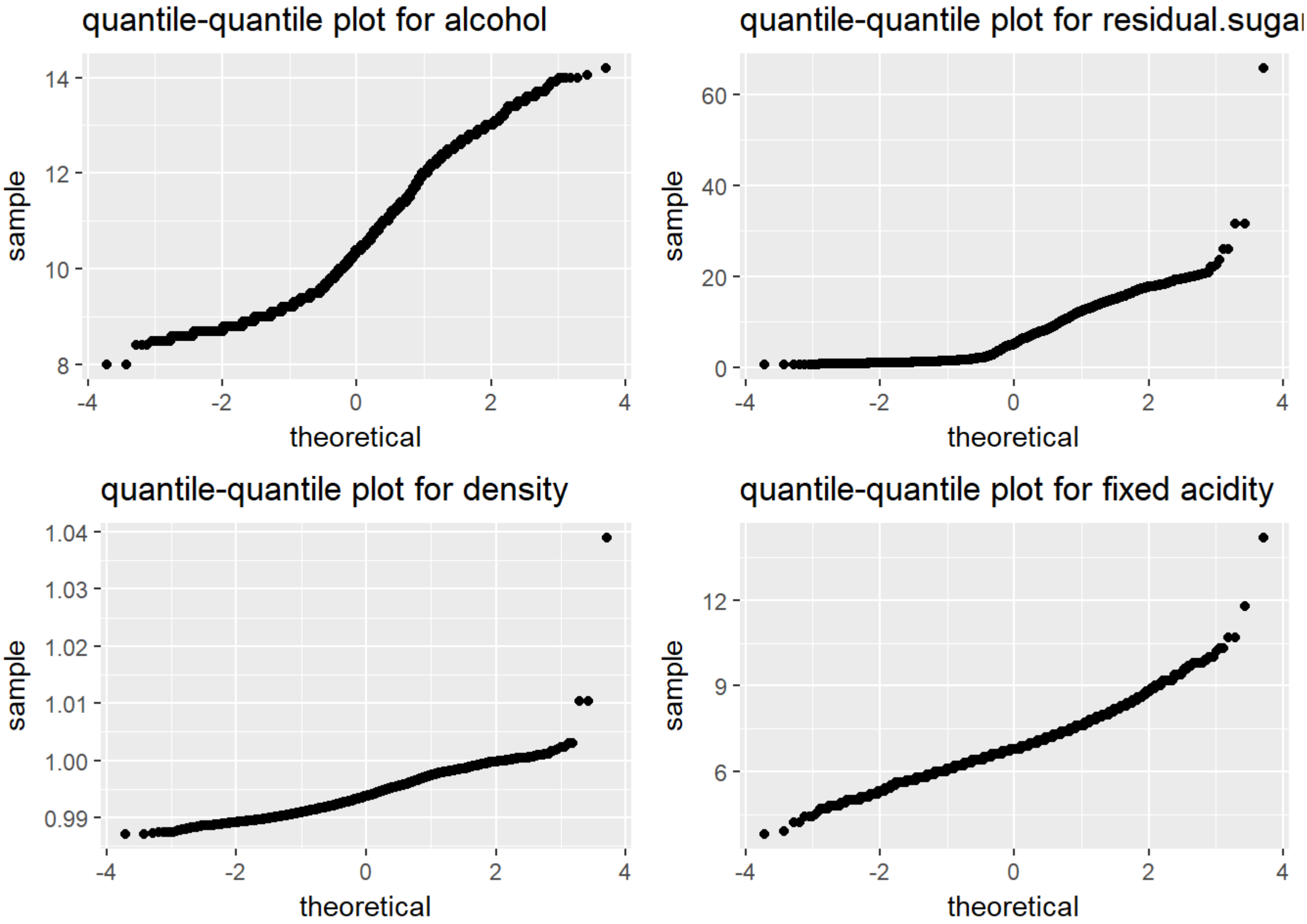
600


600
400
200
0
count
2.7
3.0
3.3
3.6
pH
Min. : 2.72
1st Qu. : 3.09
Median : 3.18
Mean : 3.19
3rd Qu. : 3.28
Max. : 3.82


1500
1000
500
0
count
0.0
0.5
1.0
1.5
citric acid (g/dm^3)
Min. : 0
1st Qu. : 0.27
Median : 0.32
Mean : 0.33
3rd Qu. : 0.39
Max. : 1.66

All of the frequency distributions for the acid-related variables are roughly symmetric and normal shaped. The means are fairly close to the median, although not as much with volatile acidity, which shows a hint of right-skewedness. All variables except pH appear to have outliers lying outside the typical range, with the outlier regions having roughly half of the range on the x-axis.

Again, it looks like with a couple exceptions, most variables have a somewhat normal distribution. To test this impression, we pick two variables that look like they have a non-normal distribution, and two variables that look like they have a normal distribution, and then we do quantile-quantile plots to verify the intuitions.



This looks like we have the upper two quantile-quantile plots showing less of a straight line than the lower two quantile-quantile plots. The plot for alcohol looks sigmoid-like (S shaped), while that for sugar level has an abrupt transition where the slope changes. The two bottom plots, on the other hand, look as straight as can be expected.

Univariate Analysis

Quality looks like a roughly symmetric, bell shaped distribution, ranging from ratings of 3 to 9, with the vast majority of quality ratings being either 5, 6, or 7, with a mean of about 5.9 and just a small number of wines rated as 3 or 9. It would seem like there are not a lot of oustanding or poor wines in the sample, at least according to the quality rating.

All variables had histograms that were symmetric or right skewed. Many had outliers on the right tail but not the left, as their levels could take high levels but the levels were bound to 0 on the left tail. The exceptions to this were the variables for sulphates, density, alcohol, and pH, which did not have as many outliers on the right tail.

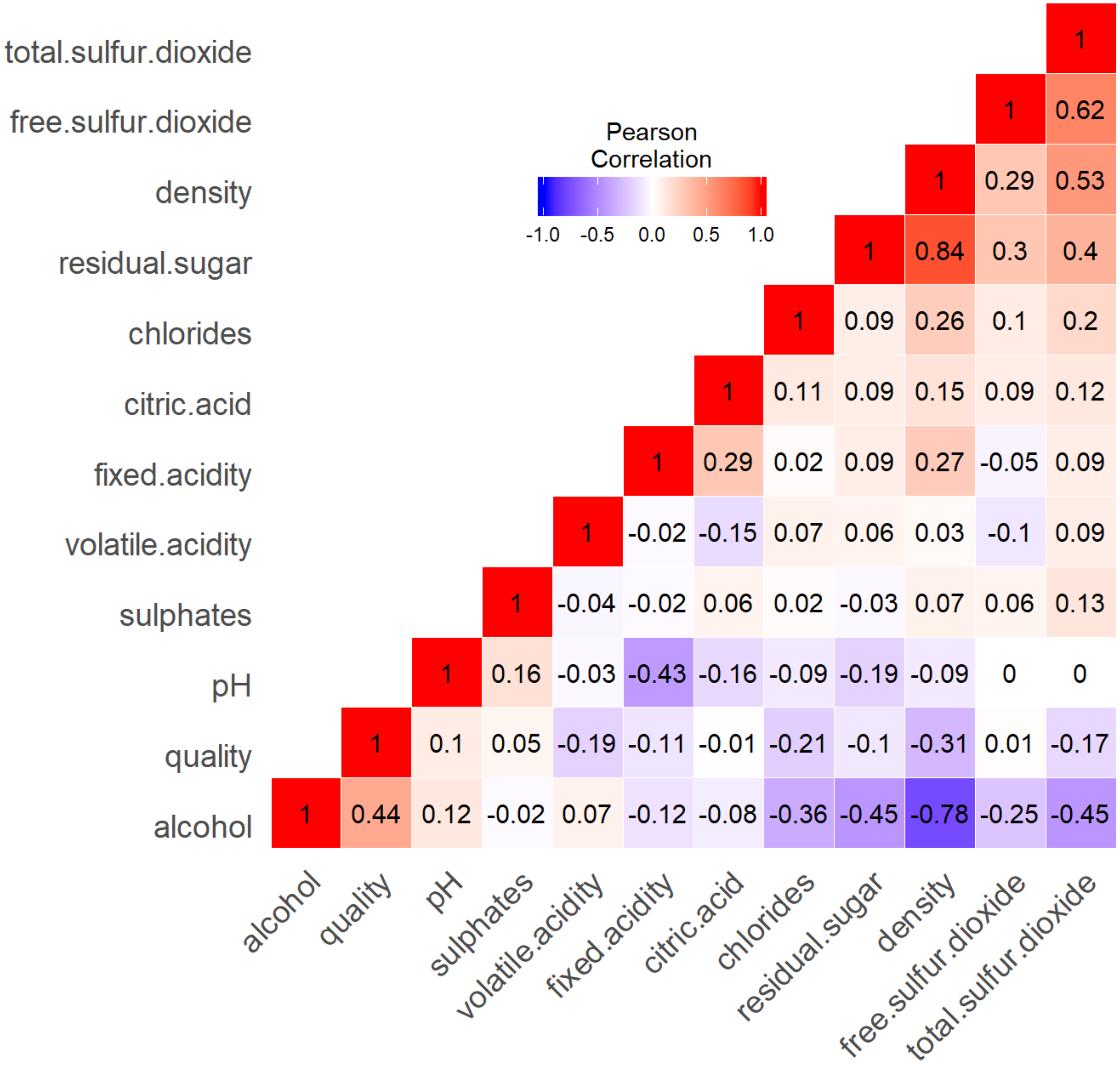
With the exception of alcohol and residual sugar level, the histograms looked essentially normal or bell shaped. We checked this impression by doing some quick quantile-quantile plotting, and indeed the q-q plots for residual sugar and alcohol levels were less straight then those for two variables which appeared normally distributed, density and fixed acidity.

Bivariate Plots Section

We now take a look at the relationships between the variables and quality.

Again, we want to know if there is a regular, well behaved relationship, if this relationship is uniform or instead it has both increasing and decreasing quality in areas, and if these relationship trends are observed for the majority of the distribution, excluding the outliers. And finally, we want to know if these relationships to quality match what we see looking at the Pearson’s correlation coefficients.

I found it cumbersome to cross correlate all the variables as seen in a ggpairs plot, so I created this correlation heatmap from code I found here (http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization)

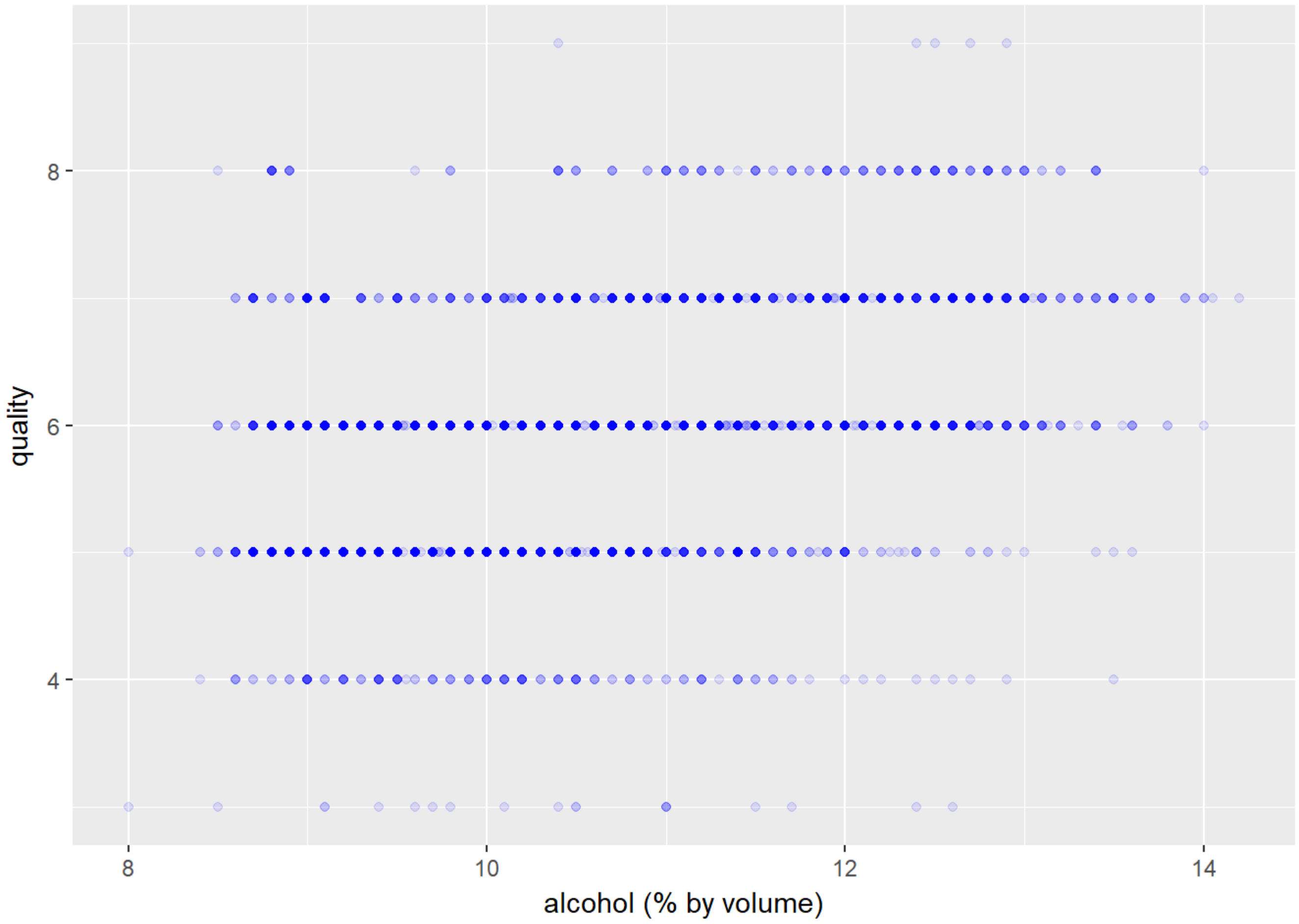


Looking at the correlation heatmap, we see just a few variables that appear to be moderately associated with quality. *Alcohol* has the most significant correlation at 0.44, followed by *density* at -0.31, *chlorides* at -0.21, *volatile acidity* at -0.19 and *total sulfur dioxide* at -0.17. There are other variables with a lesser but perhaps important degree of correlation, but we can make a cut-off to focus more closely on these variables as quality factors, keeping in mind that other variables might play a role in interacting with the variables just mentioned.

We make one more observation, noting that the variables alcohol and density, both of which may have a significant relationship to quality, themselves appear to be correlated with each other, as does density and residual sugar. This is not surprising, considering that density is affected by both alcohol and sugar level.

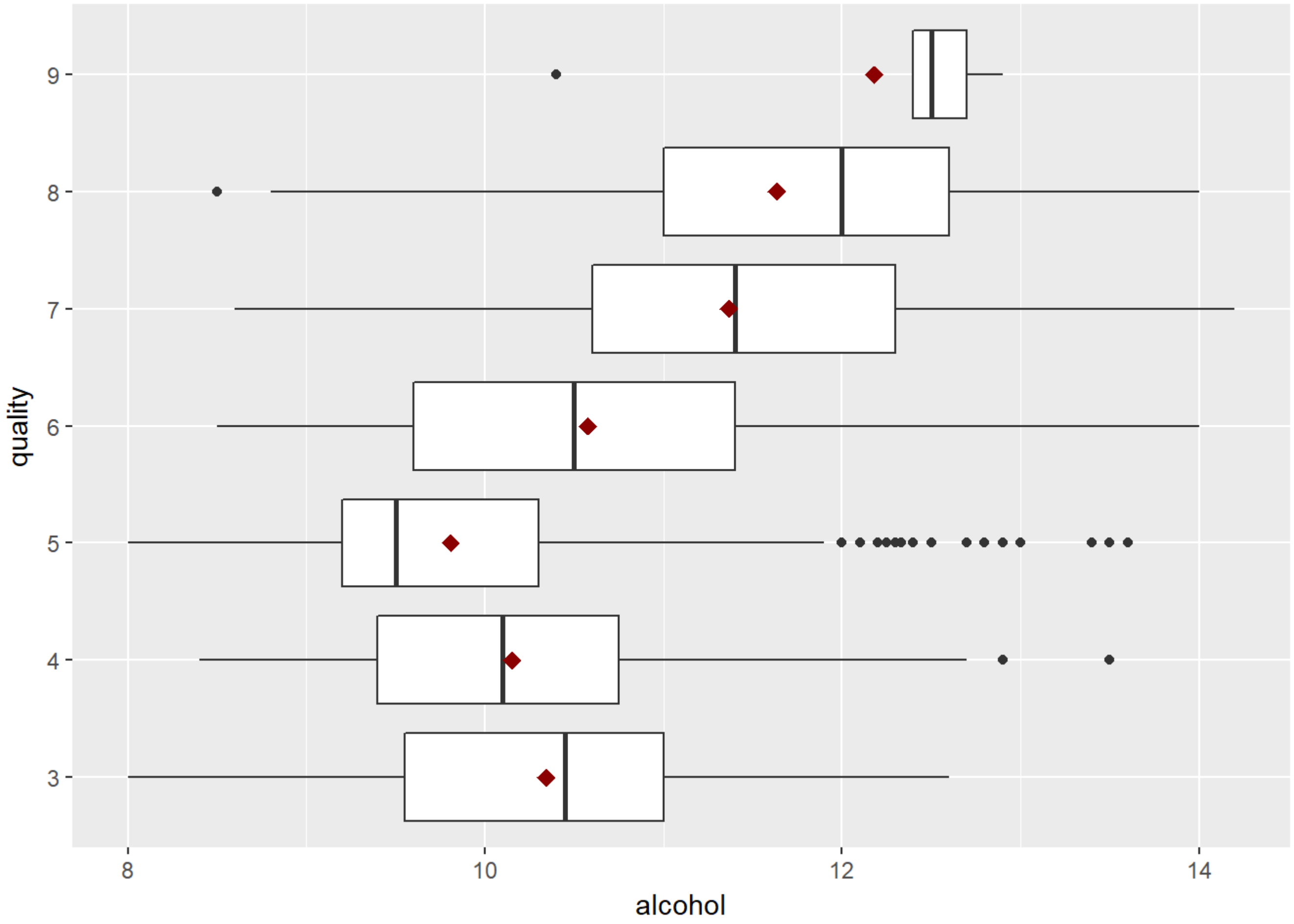
We will now take a look plotting each variables’ association with quality, focusing on the variables identified above having a possibly significant correlation with quality.

Before we do any aggregations or transformations, it’s best to look at the raw data, and the most straightforward way to do this is with the scatterplots of the variable versus the quality. We do this below with alcohol.



If we try, we can somewhat get an understanding here of how alcohol content relates to quality. For instance, by comparing the wines with an 8 quality rating against the wines with a 4 or 5 rating, we see that the higher quality wines have a higher average alcohol content on average than the lower quality wines. But this is a rather indirect inference, and we don’t see, for instance, what the average alcohol content is for each quality level.

So we try using a boxplot with the mean alcohol content to see if this gives us a better, more direct picture showing the means and how they change with quality.

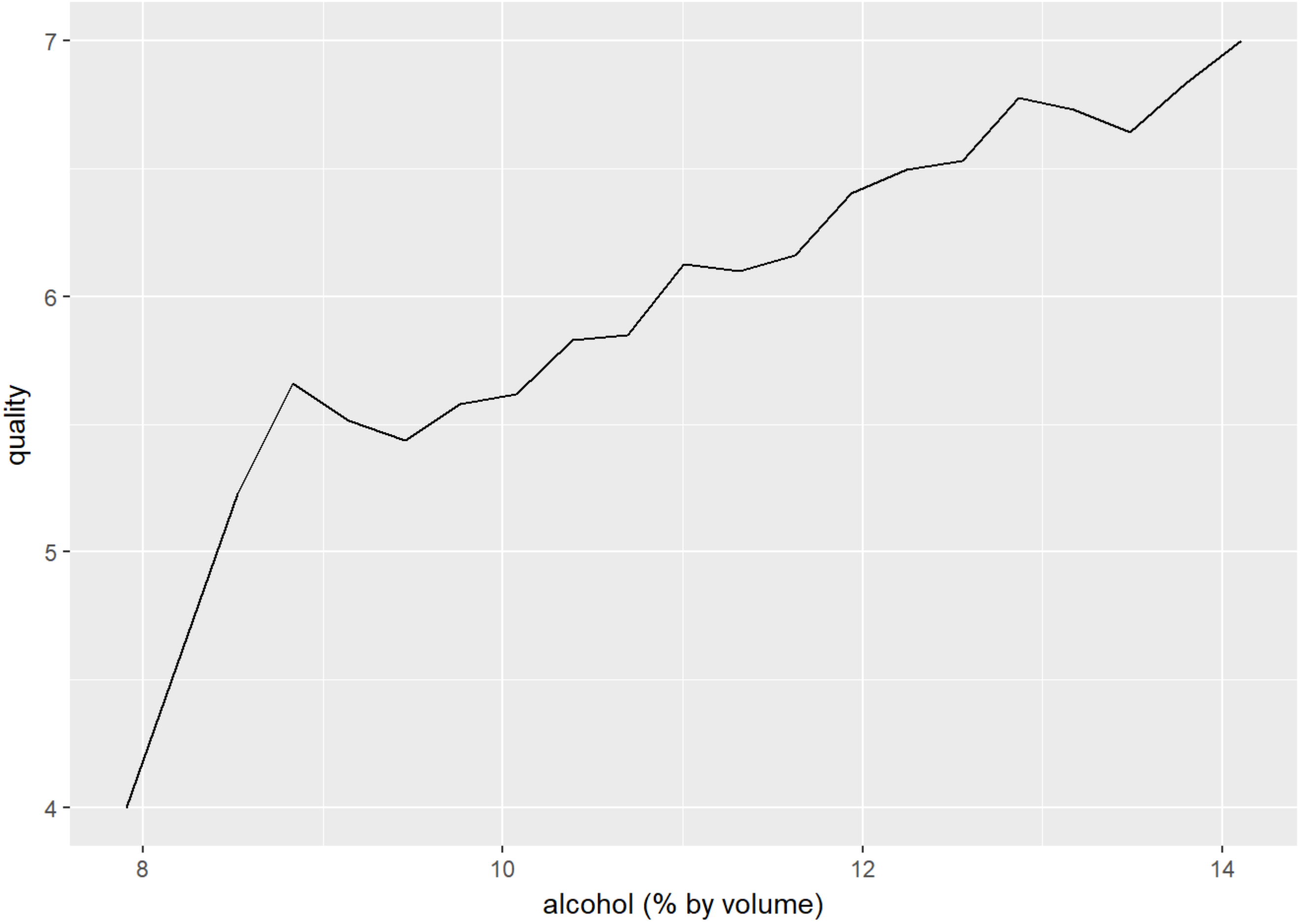


So this gives us roughly the same information as we got from the scatterplot, and we get a more direct comparison of how the mean alcohol levels change across the different quality levels. For example, we clearly see the mean alcohol level (as red diamonds) increasing for wines with qualities 5 and above. This graphically illustrates quality increasing with increasing alcohol content, which is what we want to show.

But by the same token, it misleads just a bit because it shows the alcohol content means *decreasing* as the quality increases from 3 to 5. When we look at the scatterplot we can immediately see that there are relatively few wines having these qualities and so can de-emphasize this trend, but we don't get such an indication here with the box plots; the wines with qualities 3 and 4 are given equal “weight” since they are included in the plot. We conclude from this that the scatterplot gives a more informative picture.

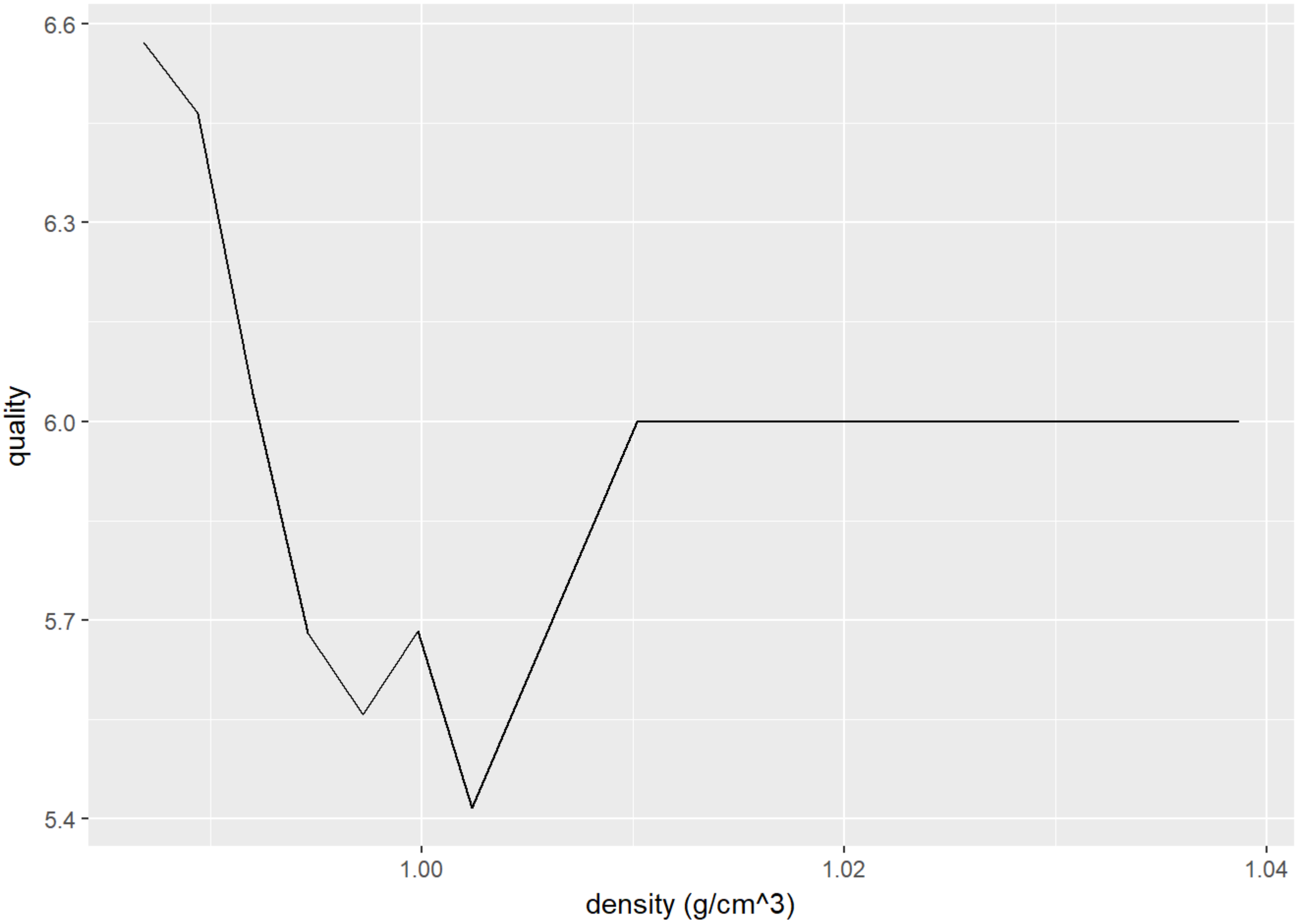
But we are still not getting the effect we want from the plots, which is to show directly how the quality varies over changes in the variable(s) of interest. We extrapolate with the scatterplots and boxplots how quality *should* be increasing or decreasing with the variable, without actually figuring out what the average quality is for a given range of the variable of interest.

So we try working more directly with the average quality, and see how that varies with the changes in the variable(s). Now we calculate the mean quality for a range (bin) of alcohol levels.



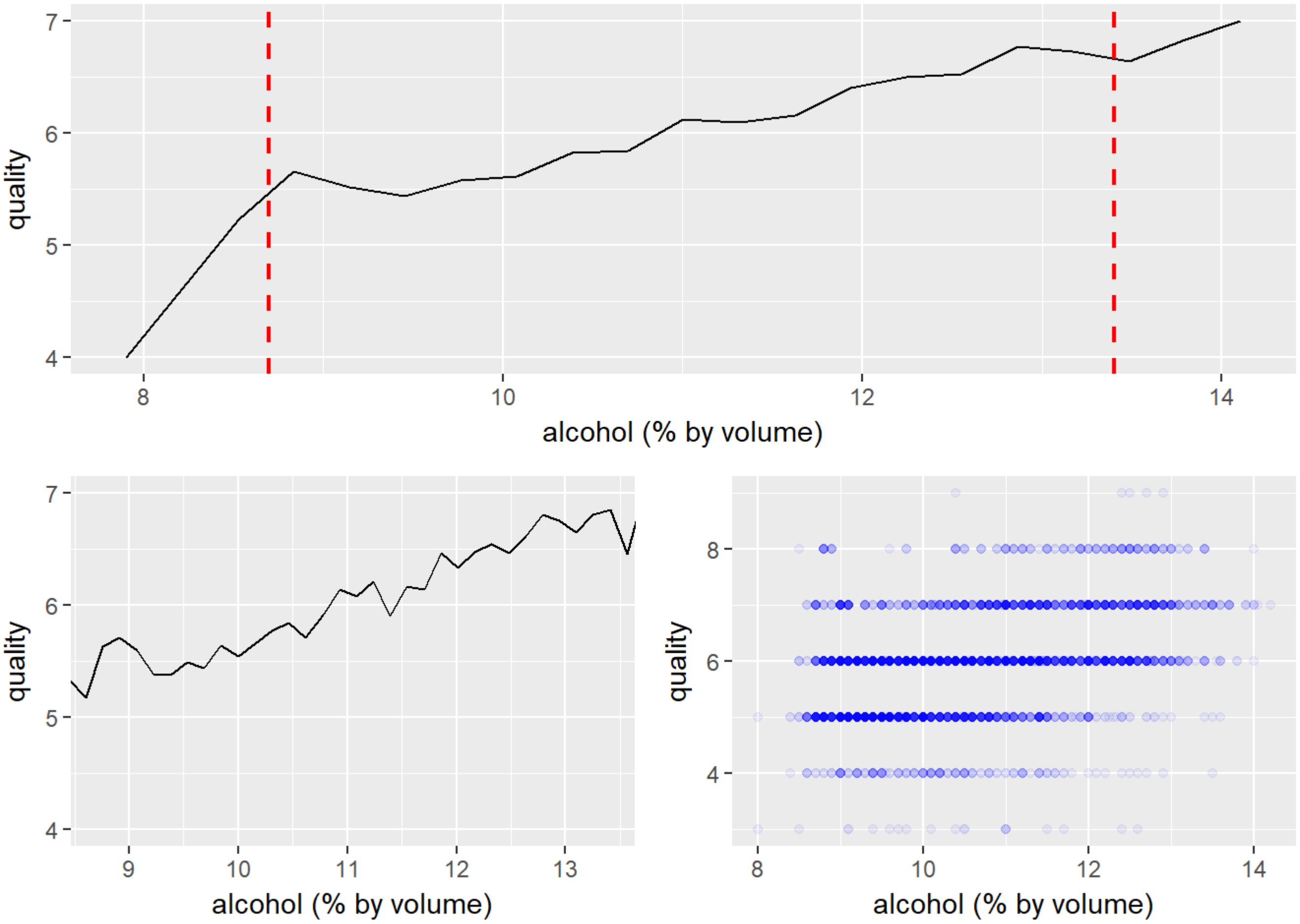
Here we see a nice, straight-ish line demonstrating quality increasing with increasing levels of alcohol. ***Finally*, something that looks like a relationship, with changes in one thing associated with changes in quality**, which is what we are ultimately interested in!

We do the same for density (below), and we run into a new problem.

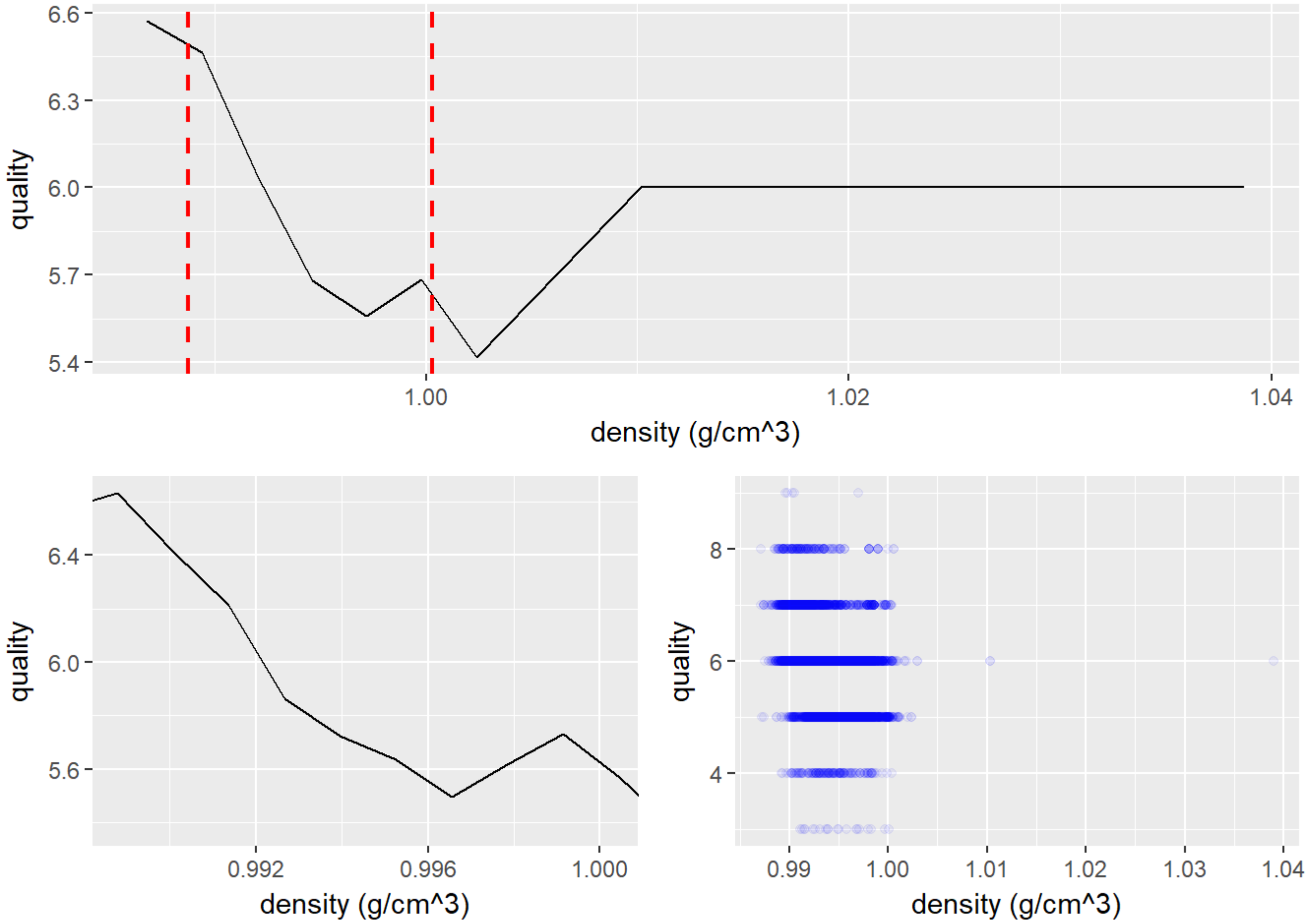


The problem is that this plot includes the outliers, taking up almost half of the plot space, without an indication this is the case. But if we just take out the outliers, we also might be leaving out information that could be useful later. So a compromise would be to include the outliers and put indications where they lie. In the following plots we do this by using vertical bars to mark off the top and bottom 1% quantiles.

With these improvements in hand, we now show the plots for all variables against quality.

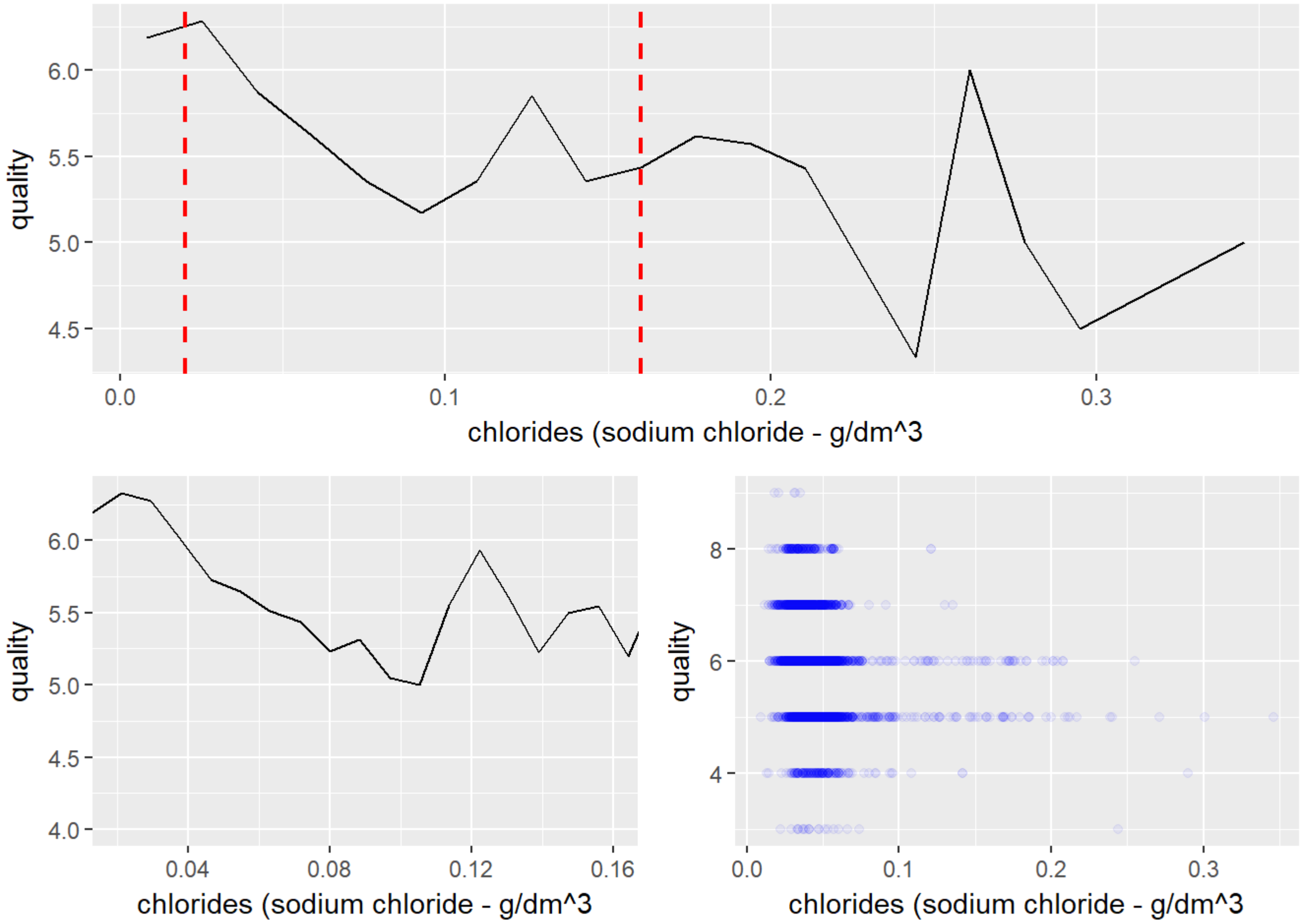


As indicated earlier, we see a clear linear relationship here between *alcohol* level and quality, and moreover this persists throughout the high and low ends of alcohol level. This is in agreement with the relatively high (0.44) correlation coefficient between alcohol and quality.

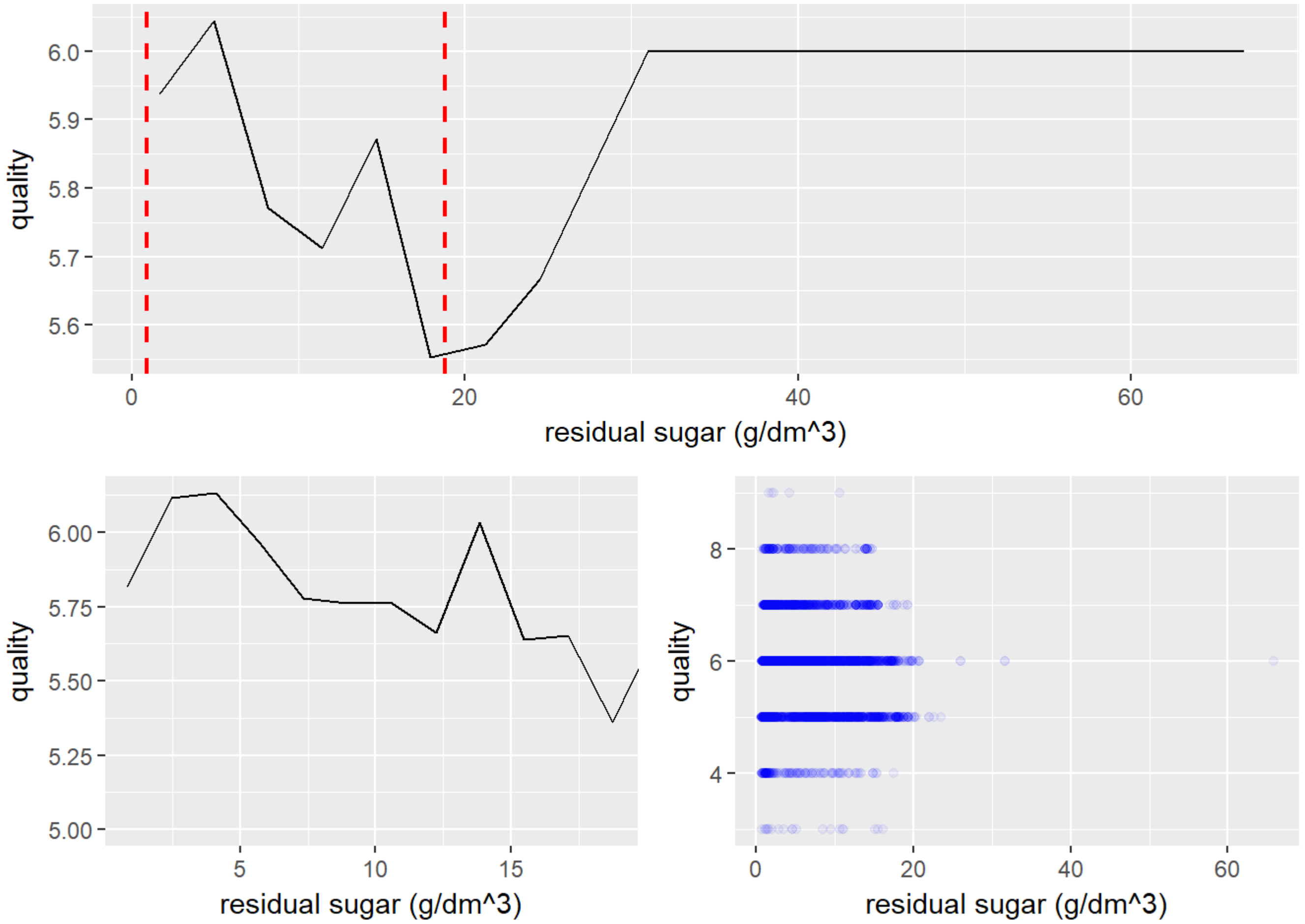


Looking at *density*, we see a mostly decreasing relationship between density and quality, in agreement with the negative correlation coefficient (-0.31) between it and quality. Its notable that so small a region of the plot is actually occupied by the area where most densities lie. Above the density of 1.0 the relationship becomes less regular.

As density has a less regular relationship with quality than alcohol, it’s possible that the correlation with quality could come about mainly through its correlation with alcohol, which we will explore later.

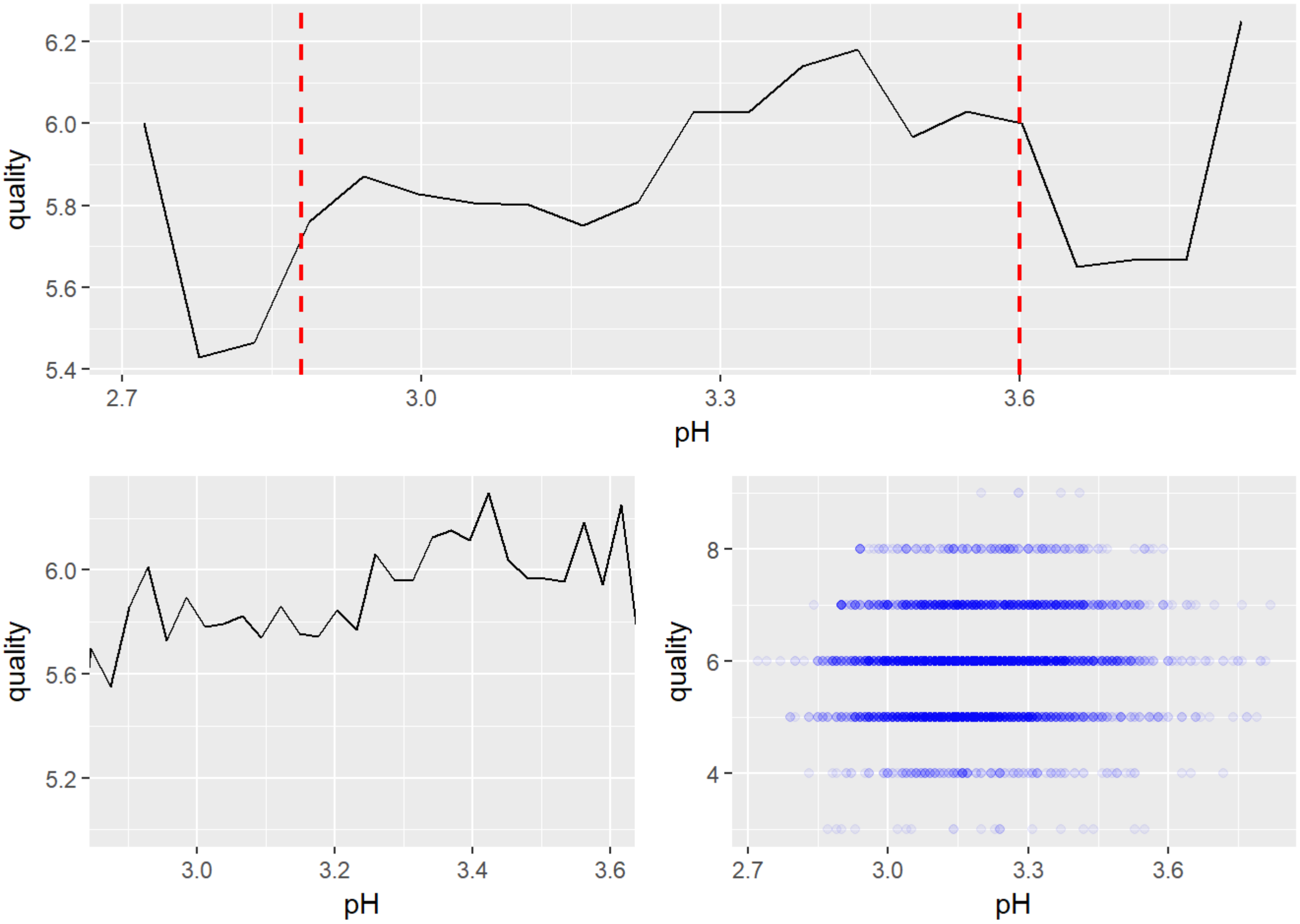


The plot between *chlorides* level and quality does not show a regular, uniform relationship. We can certainly ignore for now the wildly fluctuating mean quality values within the outlier regions, but even ignoring these we do not any regular trend that can be seen. It may be that there is no monotonic or even regular relationship to be captured here. This is curious since chlorides has the third highest correlation coefficient among all the variables.



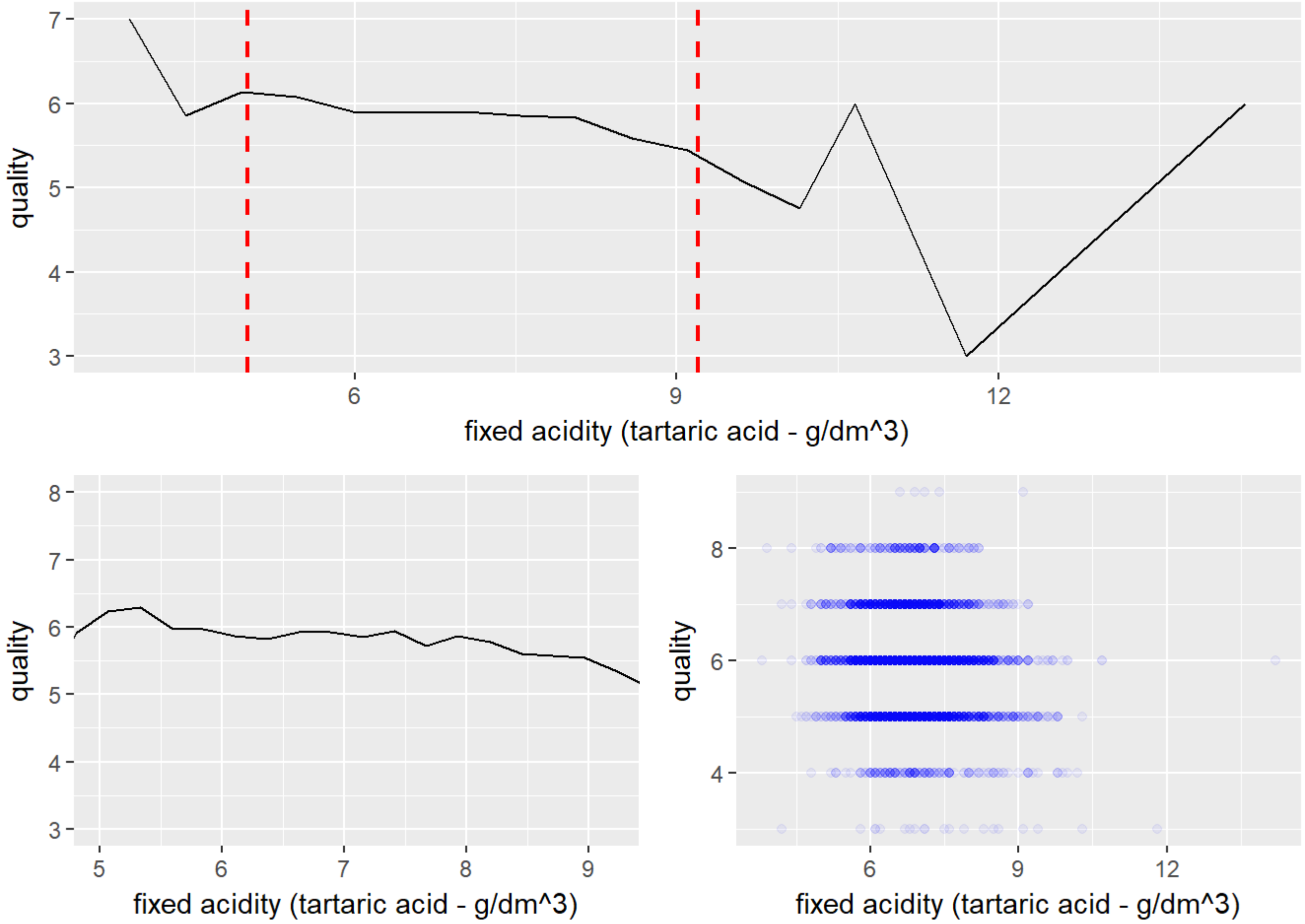
The plot for *residual sugar* has much in common with the plot for density, in that there is an irregular up and down pattern throughout the essentially decreasing relationship. And as with density, the trend does not persist in the outlier regions.

Since sugar is so strongly correlated with density and sugar has a much lower correlation to quality than density, we may be just seeing the reflected relationship of density to quality here.

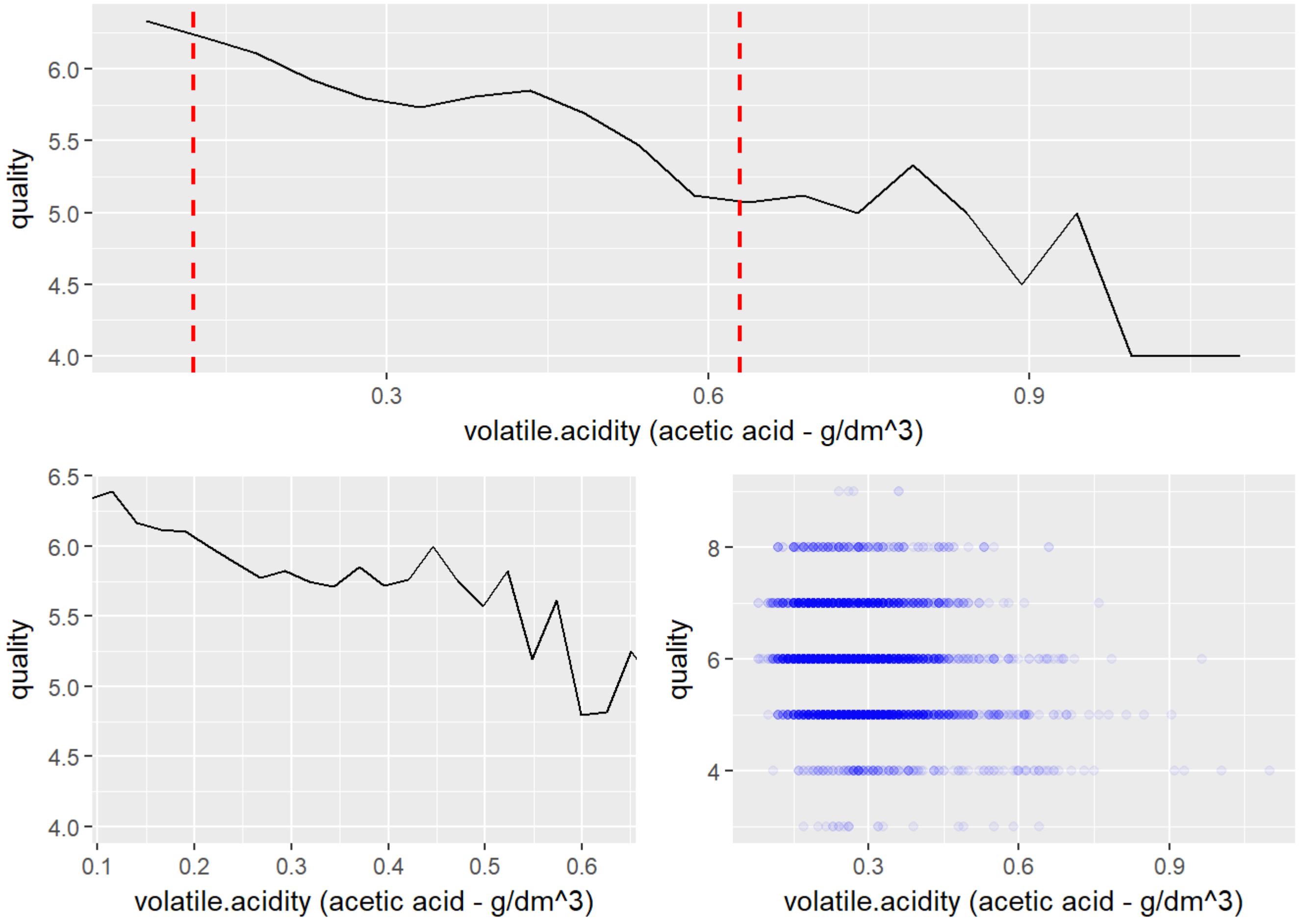


The plot for *pH* versus quality is somewhat more orderly. We see some hints of a peaked relationship, suggesting a possible near-optimal pH level for quality around 3.4. It is an interesting possibility that pH could have an orderly relationship with quality but that the effects of the upper and lower regions of pH values cancel to obscure the relationship for a correlation coefficient and a linear model. If this were the case, the small 0.1 correlation with quality would be misleading.

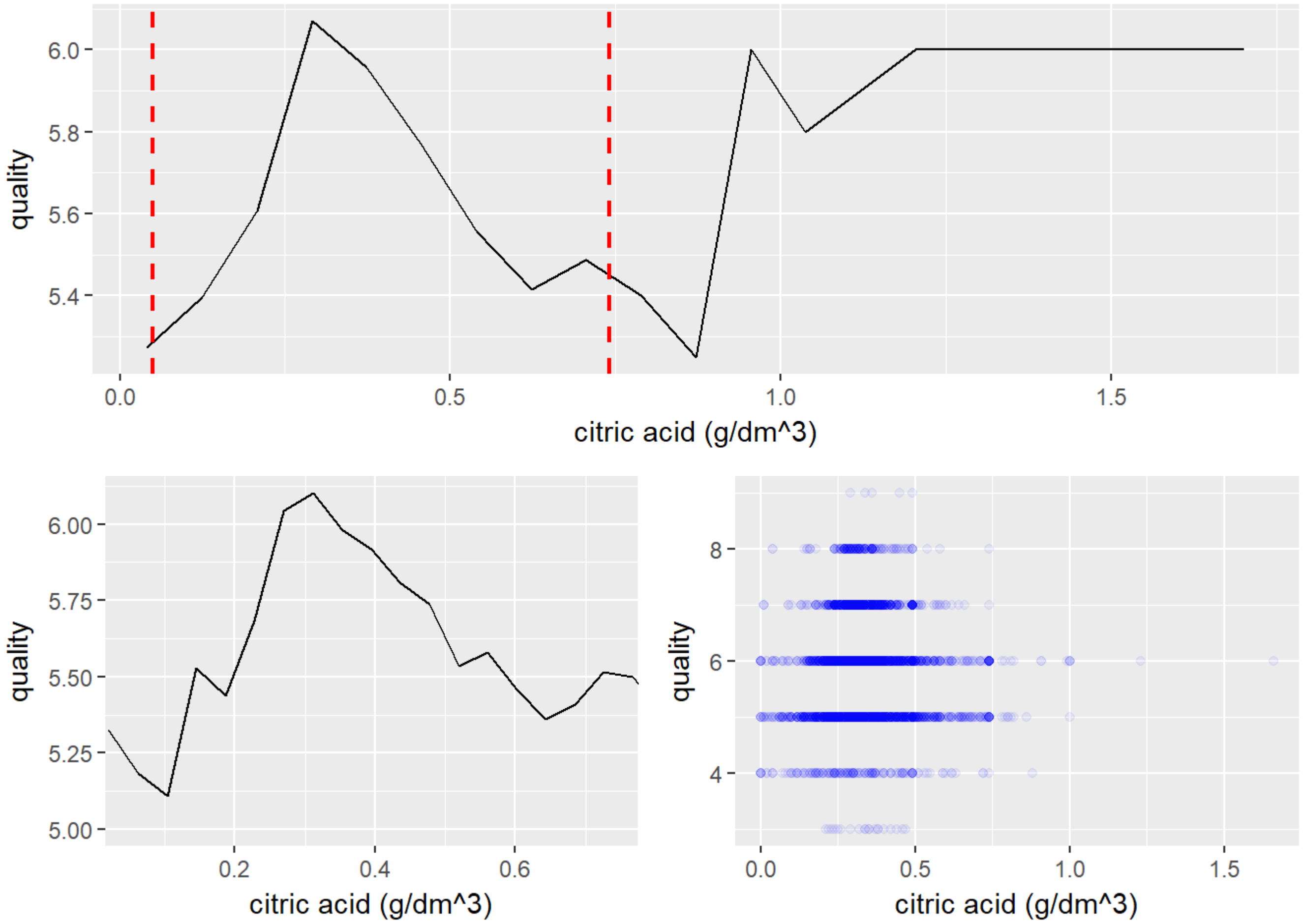
The irregular shapes in the outlier regions for pH may be an artifact of having fewer samples for the outliers.



Fixed acidity shows a very slight decreasing relationship between it and quality, but the plot is mostly flat, which suggests we can mostly ignore its effect on quality.



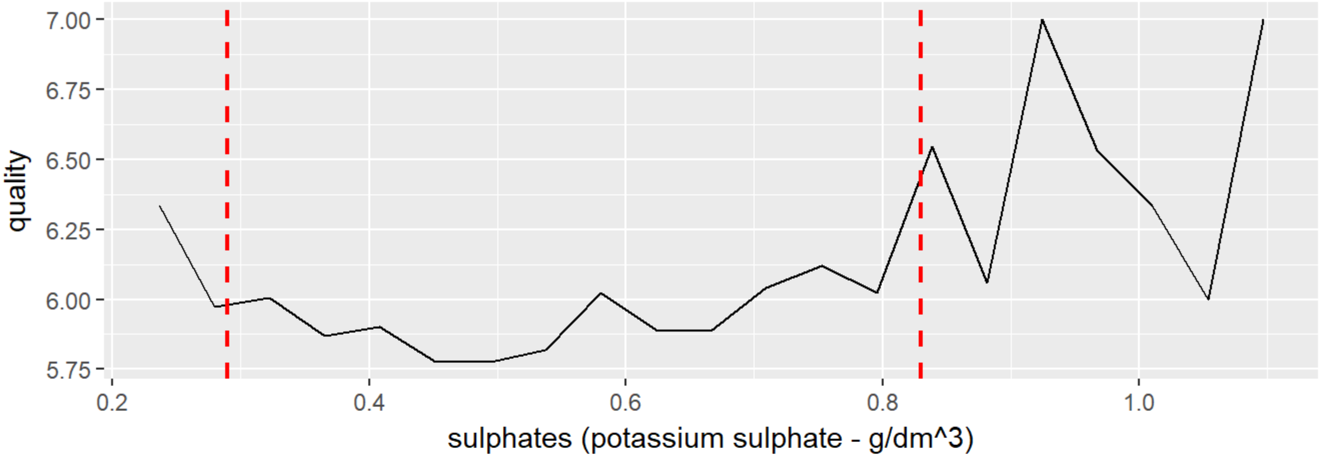
In contrast, *volatile acidity* shows a regular, mostly linear and decreasing relationship between it and quality. This relationship persists in the outlier regions.

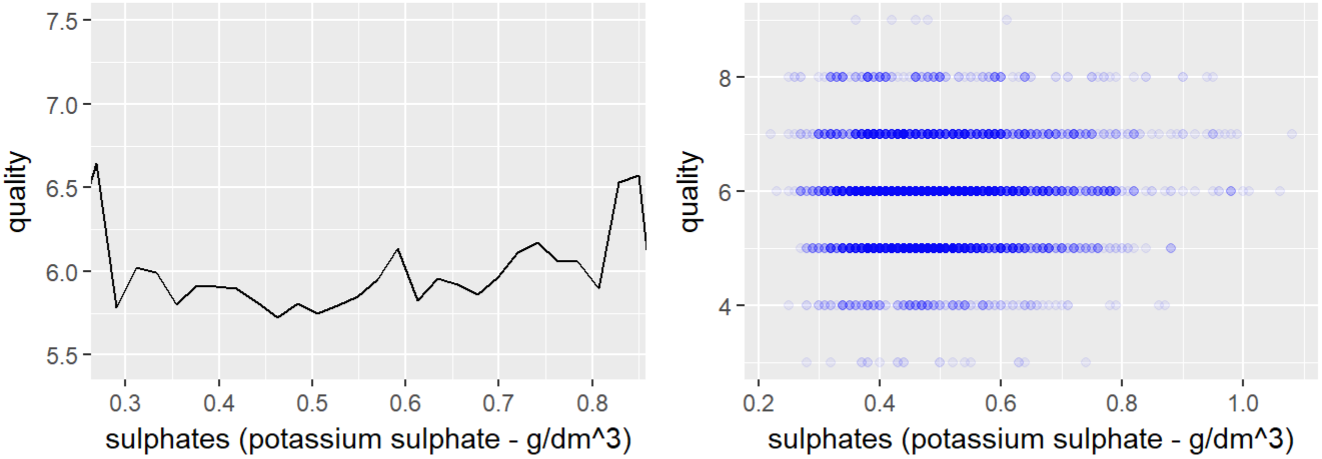


Citric acid has a very peaked relationship with quality, suggestive of a possible optimal region in terms of quality. But from the description of the variables given, there is no suggestion that too much citric acid causes bad side effects, and it is usually only seen in small quantities, so perhaps there is an unappreciated side effect of excess citric acid, or the appearance of an optimal level is illusory.

We will investigate this further when we look more closely at the variables with peak quality neighborhoods.

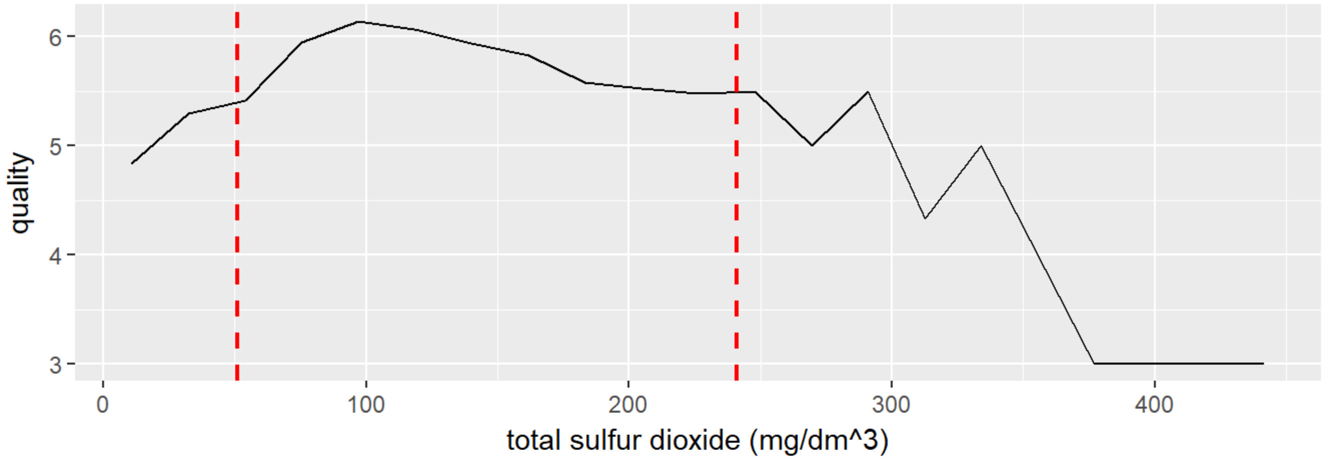
We now look at the sulfur-related variables, starting with the *sulphates* measure itself.

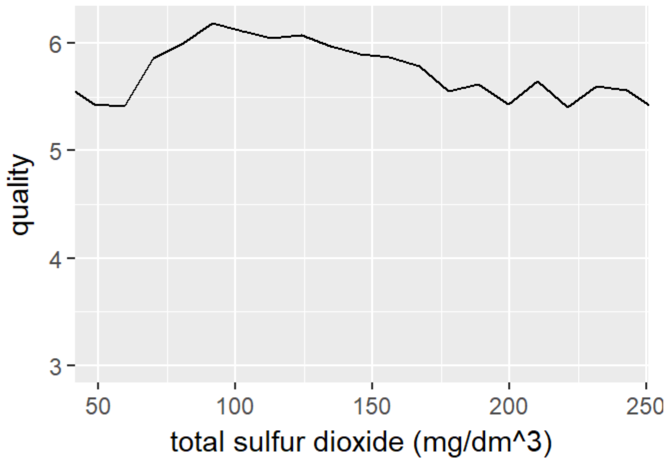


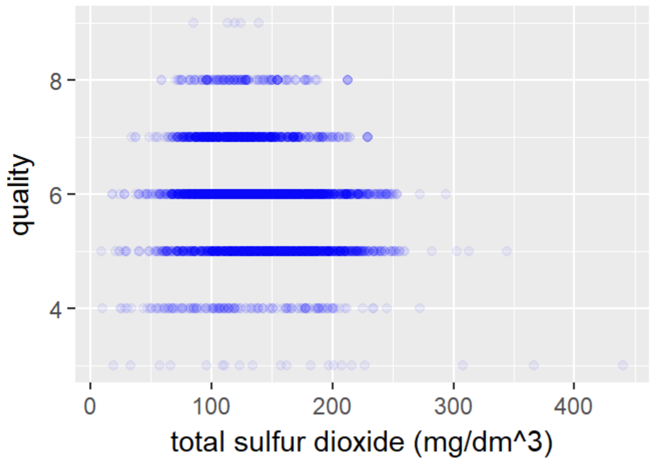


Looking at the *sulphates* levels (*above*) it’s hard to tell if there is low point quality-wise for *sulphates* here, or if instead it is an essentially flat relationship. But a low point in the middle of the range does not make as much sense, since we could expect a point of not enough sulfur risking microbial growth, and a point of too much sulfur when the smell and taste becomes noticeable, so if anything there should be a *peak* of quality at a “sweet spot”.

And this looks like what we see below with *total sulfur dioxide*.

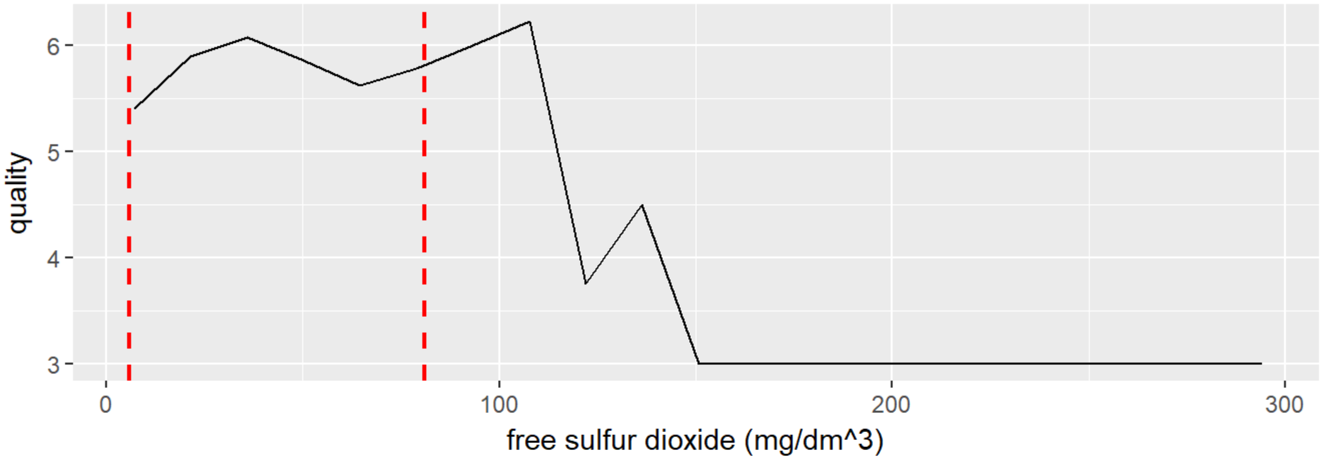


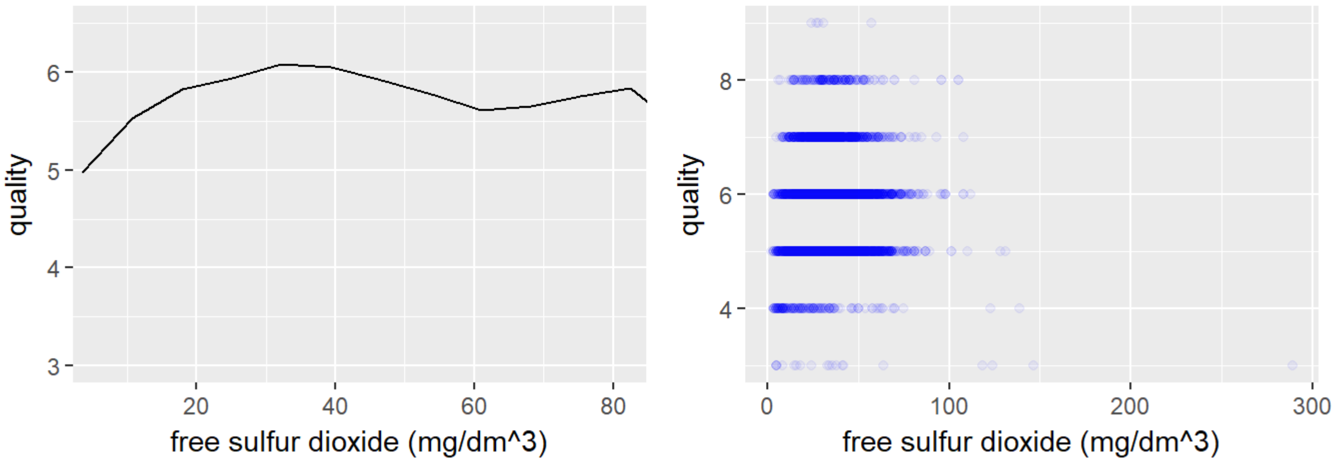




This is suggestive of the aforementioned “sweet spot” of not too little microbial effect and not too much sulfur taste and smell. The quality increases and peaks at around 100, and then drops off.

The trend looks solid, as there is not a whole lot of wigglyness on the line before and after the putative quality peak, and the decreasing quality trend after the peak continues in the outlier region.



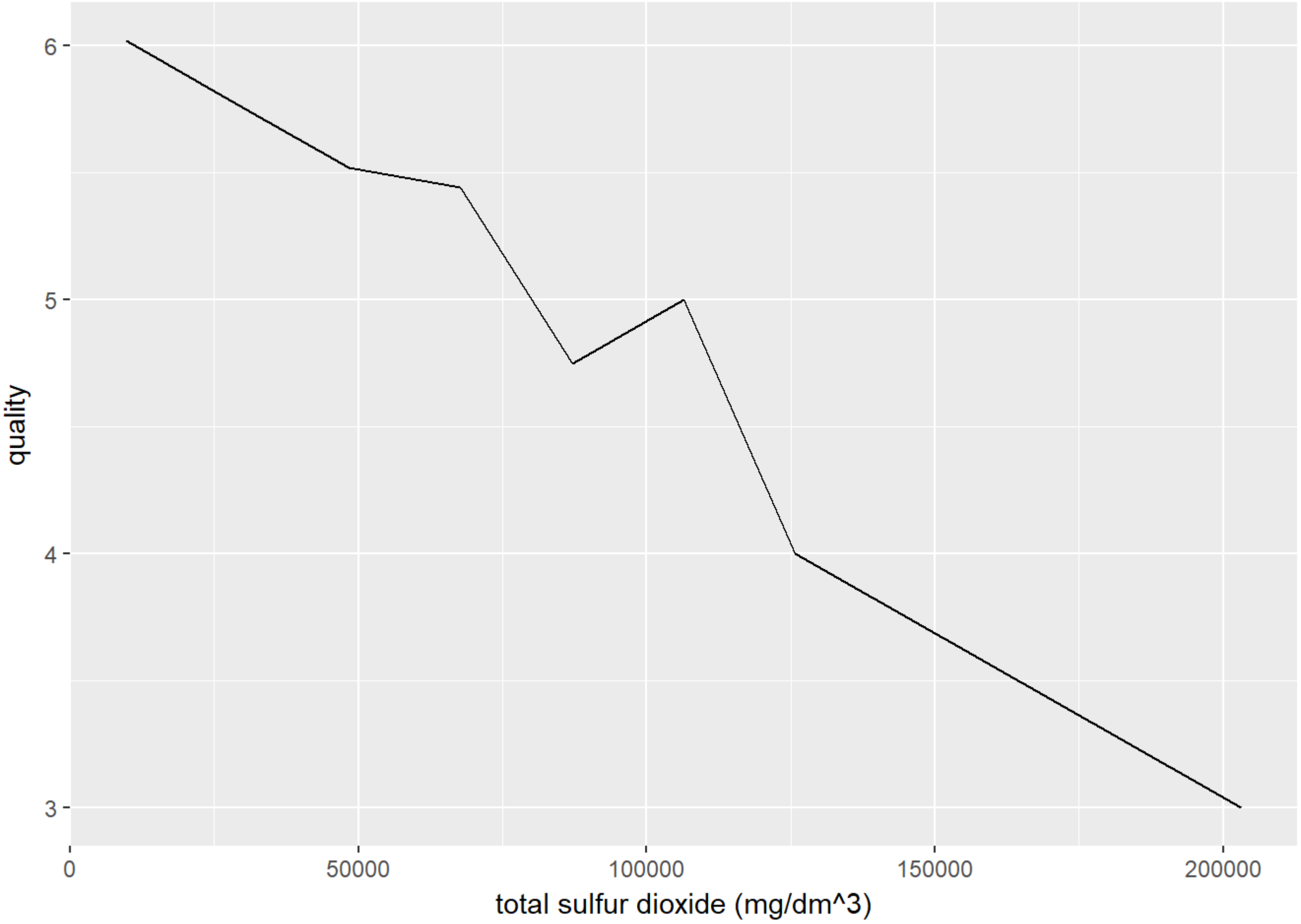


There *could* be a peak quality neighborhood for *free sulfur dioxide*, but it does not inspire as much confidence as does total sulfur dioxide, as there could be a second time where quality increases, but it is hard to know if this is largely due to an errant high quality point for a few outliers.

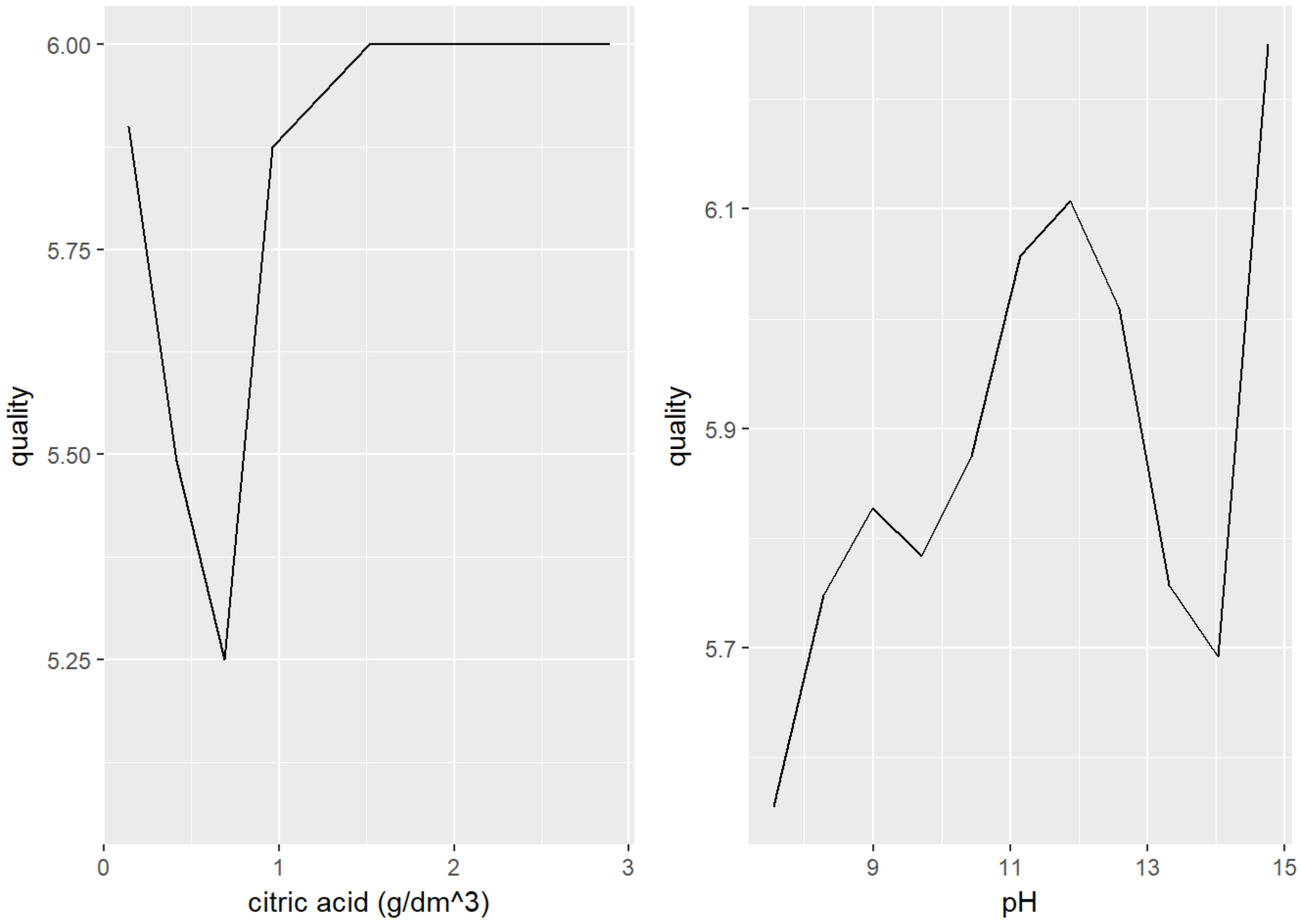
The peaked relationships to quality are examples of a *nonlinear relationship*. We take a closer look at some of the variables to see if a suitable transformation makes the plots against mean quality appear more linear.

We noted before that pH, total sulfur dioxide, and citric acids may have areas of peak quality. We try out a transformation with the squared value, as a peak could correspond roughly to the maximum of a parabola. We rule out looking at a log or exponential or square root transformation, as these would not be expected to produce an increasing and then decreasing value.

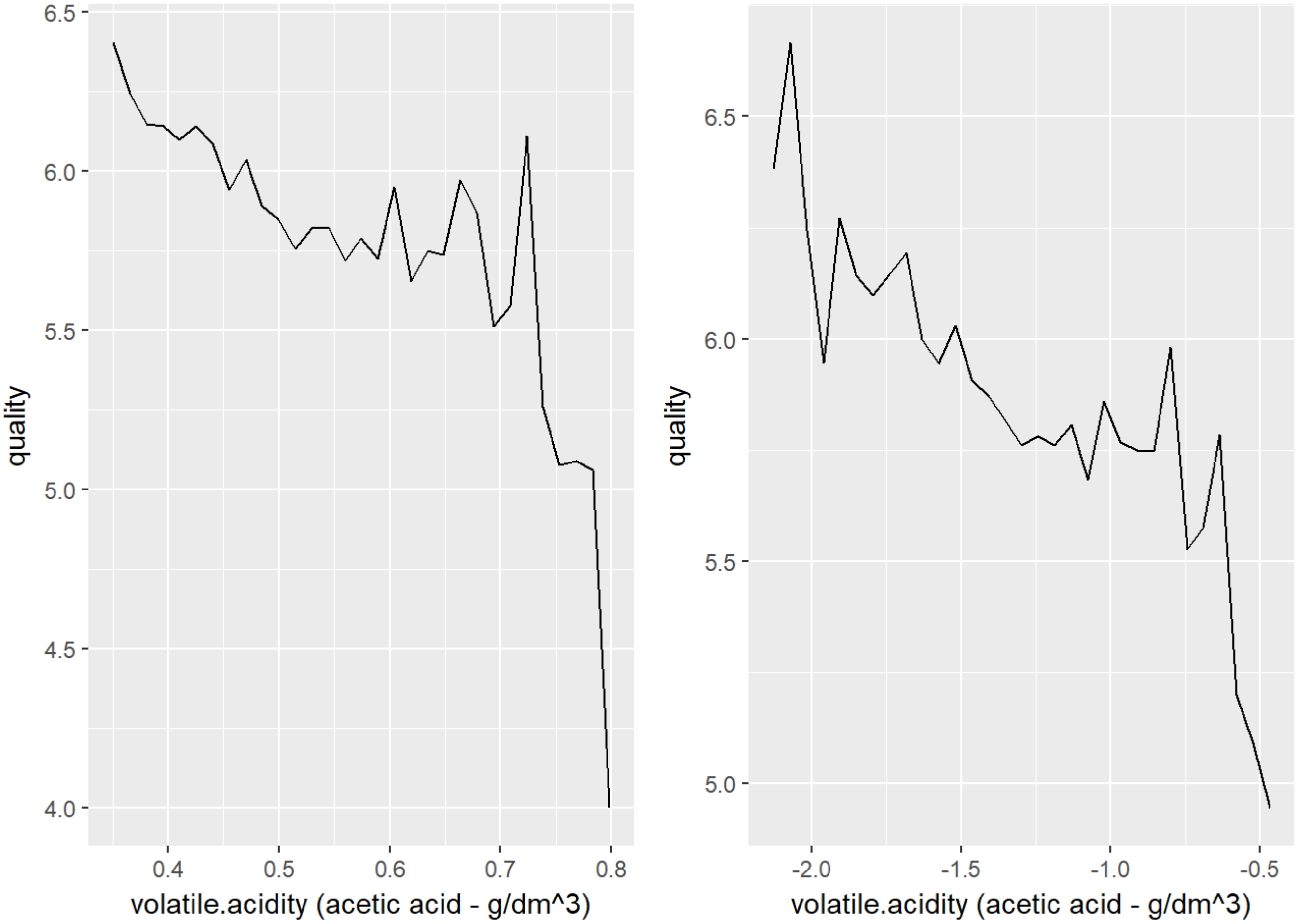
As can be seen below, *there was some improvement in the straightness after transforming the square of total sulfur dioxide.*



But the peaks remained for mean quality versus pH and citric acid level after transforming by the squared value (*below*) and using multiple possible bin sizes.



We note in passing that for *volatile acidity* a square root and/or a log transform resulted in a *fairly* linear relationship (i.e. constant slope) when plotted against mean quality for the majority of values that volatile acidity takes (i.e. outside the outliers) (*see below*).



Bivariate Analysis

Based on the orderliness of the plots and also on the correlations (Pearson’s coefficients) with quality, we only see a few variables that are likely to have a very significant association with quality. These include *alcohol* level, *density*, *volatile acidity* level and *total* SO_2 level. Missing from this list is the *chlorides* level, which has a the third highest correlation coefficient association with alcohol but does not show much regularity in how its levels vary with average quality levels.

Looking a little more closely at the shape of the variable-versus-quality plots, we some possible peaks or optimal neighborhoods regarding quality. Total SO_2, pH, and chlorides all have areas where quality uniformly rises and then falls, although as mentioned with chlorides this relationship may be too irregular to be important. The other variables as a whole have average quality uniformly rise or fall along with the measure of the variable in question. We found also that many variables had outliers occupying a significant portion of the plot, and compensated for this in future plots by removing the outliers.

Finally we found that the plots of *total* SO_2 against quality were somewhat straitened by taking the *square* of the measure of SO_2 level, and the corresponding plots for *volatile acidity* level were improved by taking the *log* or *square root* of the volatile acidity level.

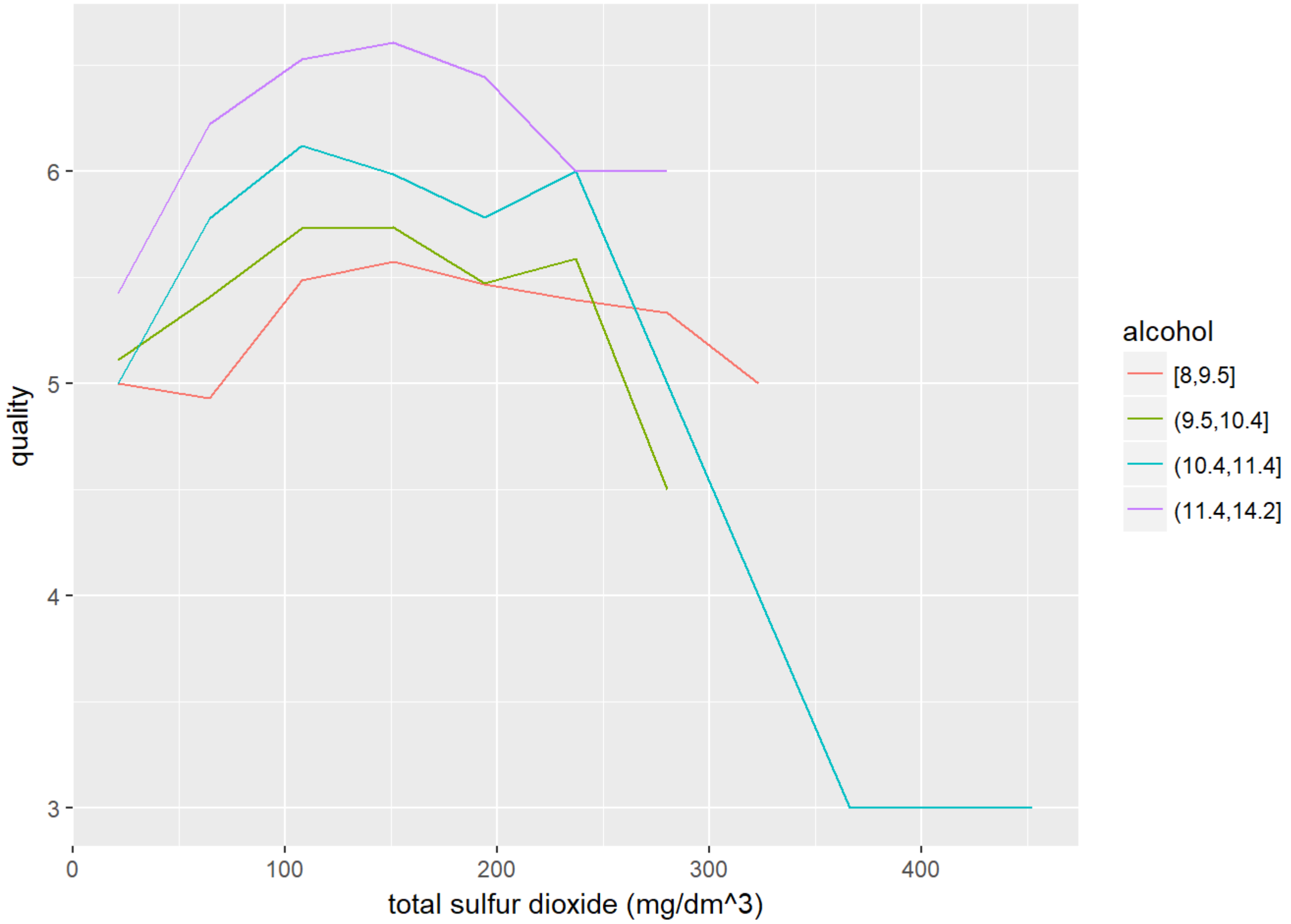
Multivariate Plots Section

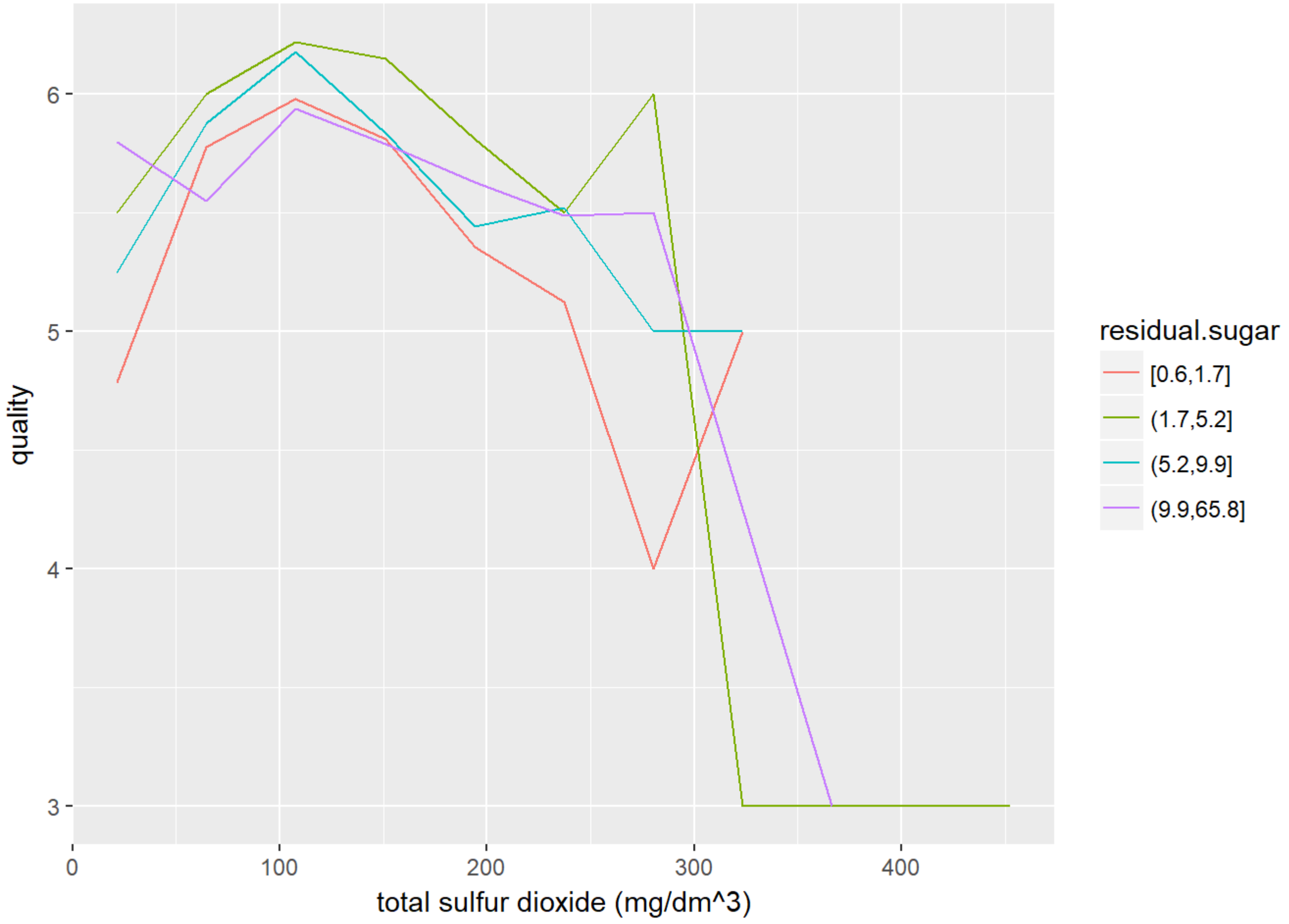
We now take a look how the variables we have looked at might interact in order to shape the quality.

We ask two questions here: 1) For variables that might have an optimal level for quality, can other variables affect this optimal level? 2) Do any variables related to alcohol level moderate the effect of alcohol on quality?

We start off taking a look at two variables that we have seen with peaks, total sulfur dioxide and pH.

First we look at the total sulfur dioxide (SO_2) We test a few variables unrelated to sulfur with moderate correlations to SO_2, picking out alcohol and residual sugar.


quality
6
5
4
3
0
100
200
300
400
total sulfur dioxide (mg/dm^3)
alcohol
[8,9.5]
(9.5,10.4]
(10.4,11.4]
(11.4,14.2]

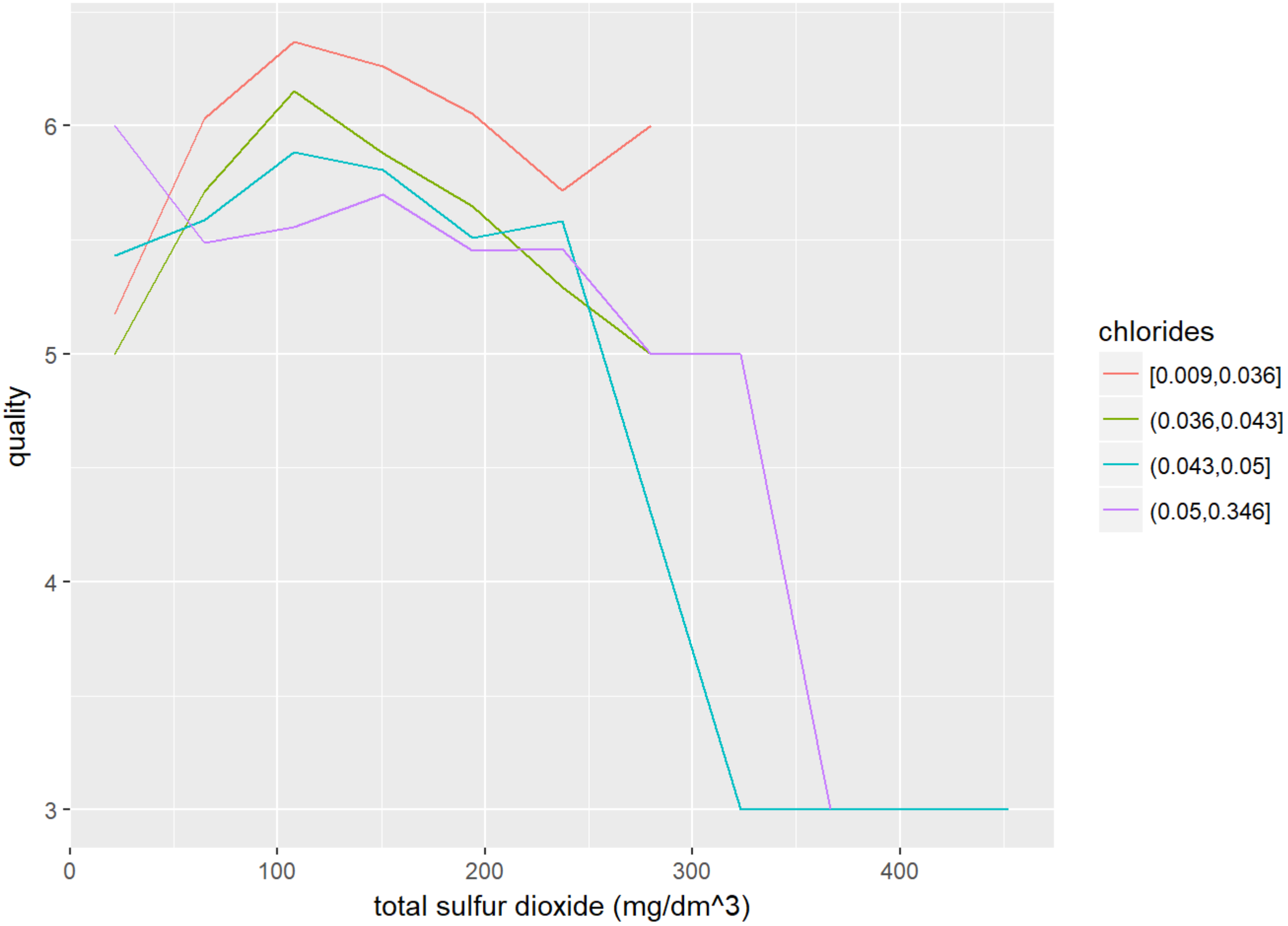

quality
6
5
4
3
0
100
200
300
400
total sulfur dioxide (mg/dm^3)
residual.sugar
[0.6,1.7]
(1.7,5.2]
(5.2,9.9]
(9.9,65.8]

These plots are interesting, because they show what looks like a definite, nonrandom relationship between the sugar and alcohol levels and the optimum total SO_2 level. Each individual line which groups the sugar or alcohol levels faithfully repeats the rising then falling relationship towards quality, peaking at almost the exact same level.

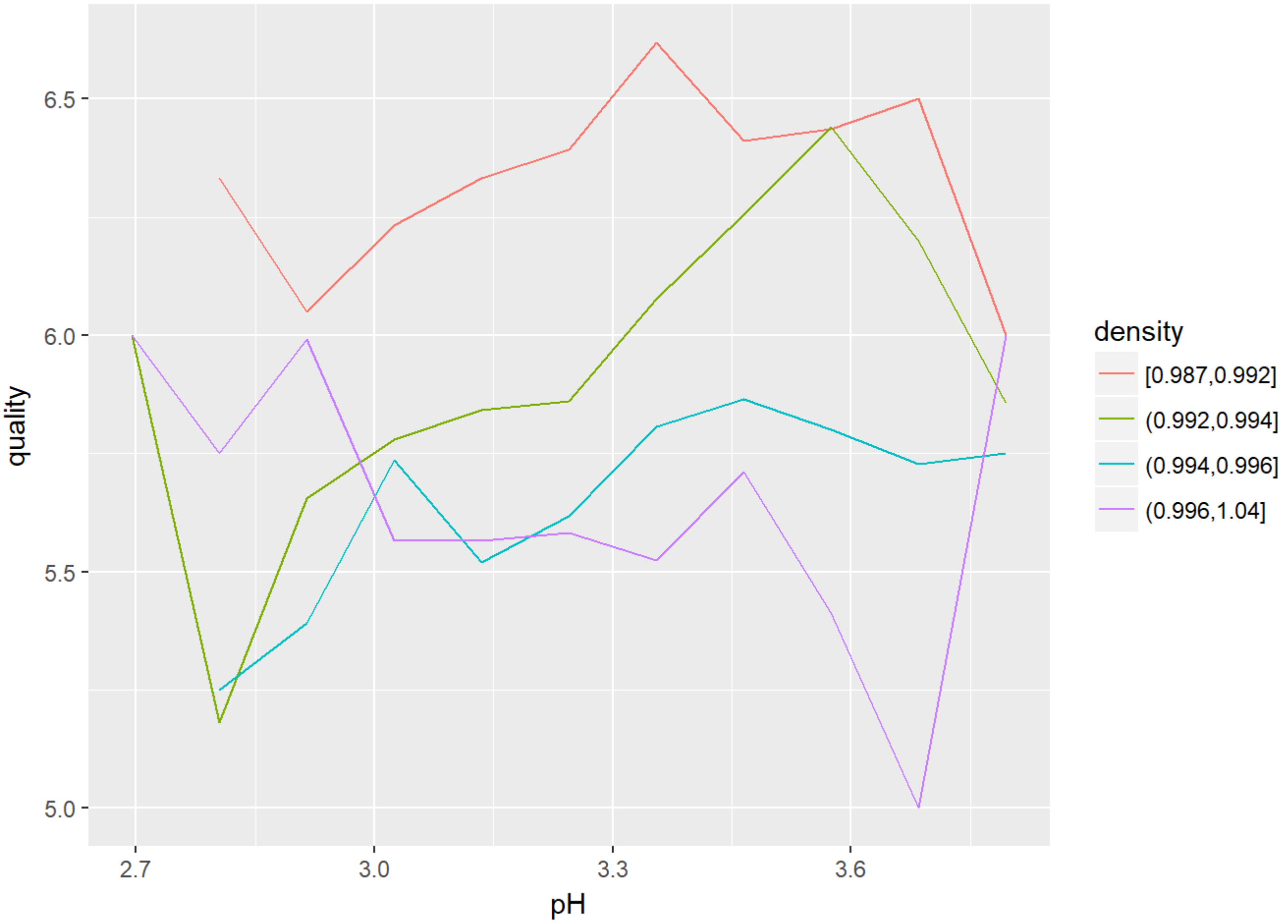
Intriguingly, there are hints that alcohol level may affect the optimum total SO_2 point, because at higher alcohol levels the quality peaks around 150, a slightly higher SO_2 level than the peaks just over 100, which is the peak level overall.

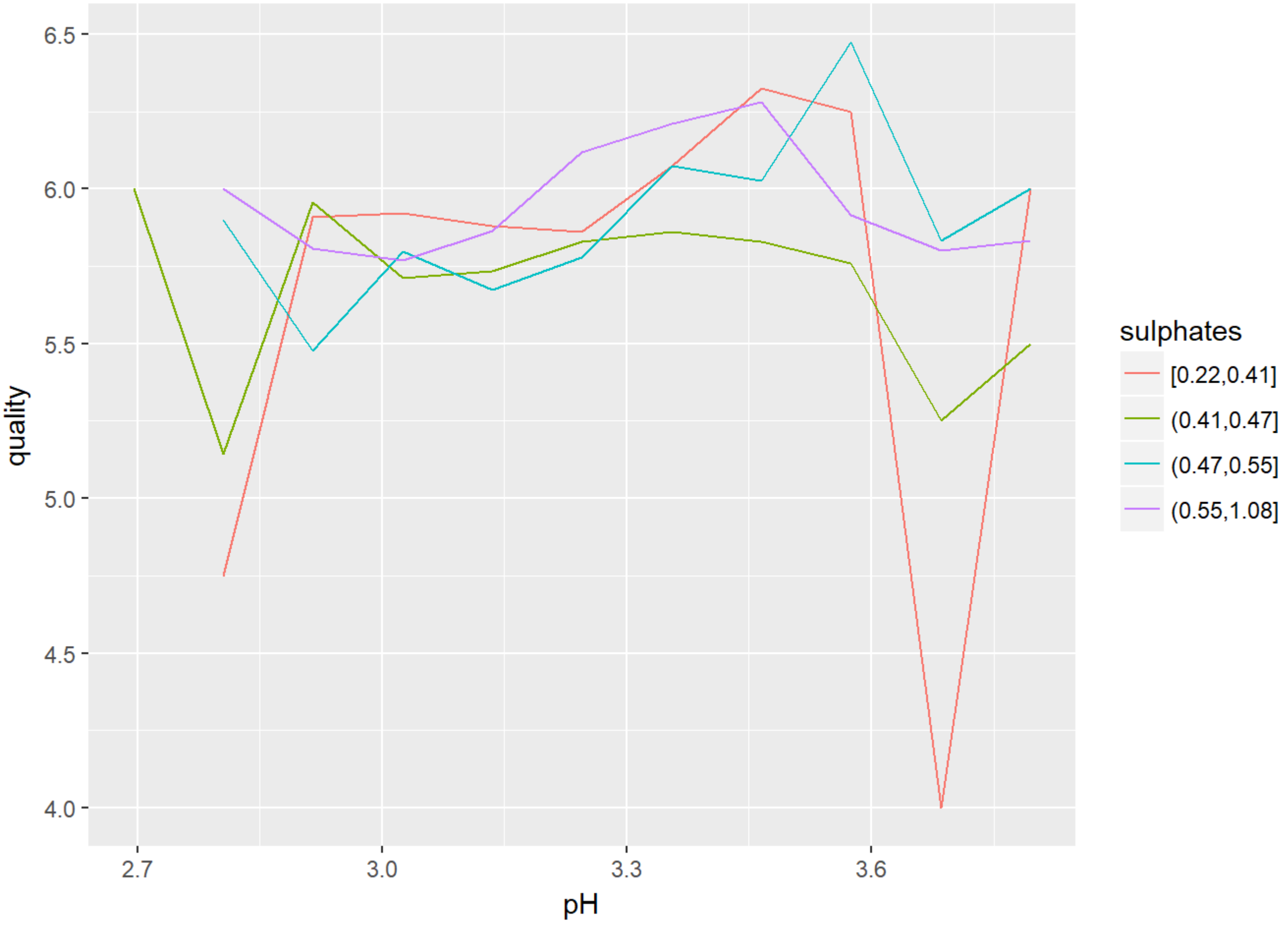
Looking at the *residual sugar* level, however, shows no such effect of sugar level on the location of the peak quality. For every level of residual sugar, the peak is at exactly the same place. This reinforces the idea that there is an actual optimum level for total SO_2 at around 110 mg/dm^3 corresponding to a “sweet spot” of not too much nor too little sulfur. Indeed, all other variables (with the one exception noted below) preserved the same peak or optimum SO_2 level across all levels of the variable.

The one exception to this was a possible effect of *chlorides* shifting the optimal total SO_2 level towards 150 mg/dm^3 at the highest chlorides level, shown below. This could be due to higher salt levels masking the sulfur taste and smell.



We now take a similar look at whether there is an optimal *pH* level that can be affected by other variables. But **a few plots are enough to show that there is no such well defined optimal pH level** which may or may not vary across different levels of another variable. This is illustrated by looking at *density* and *sulphates* below.

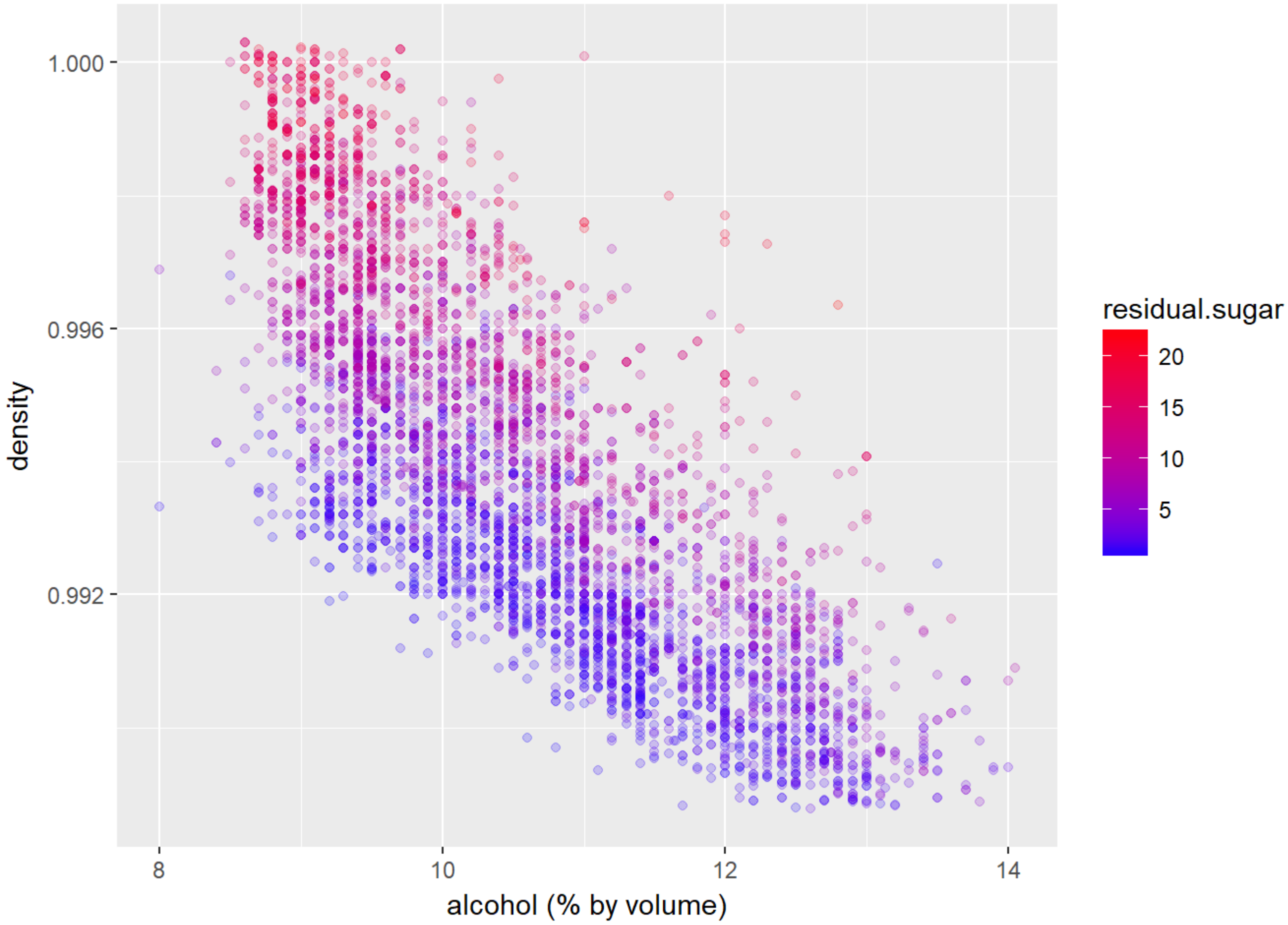

6.5
6.0
5.5
5.0
quality
2.7
3.0
3.3
3.6
pH
density
[0.987,0.992]
(0.992,0.994]
(0.994,0.996]
(0.996,1.04]


6.5
6.0
5.5
5.0
4.5
4.0
quality
2.7
3.0
3.3
3.6
pH
sulphates
[0.22,0.41]
(0.41,0.47]
(0.47,0.55]
(0.55,1.08]

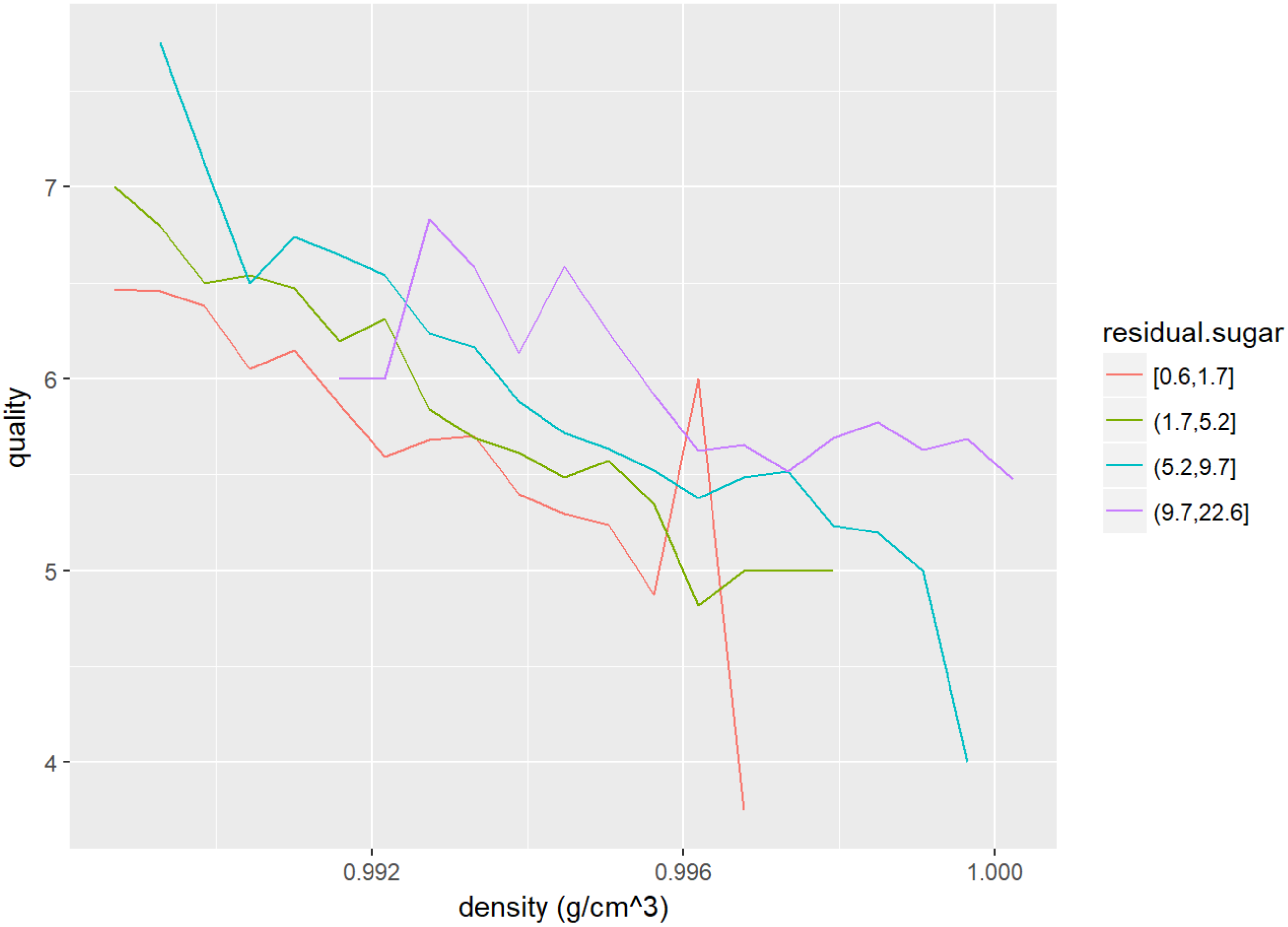
There are certainly some hints of a rough neighborhood with optimal pH as we can see above with the plots for density and sulphates, but they are not as well defined as they were in the case of total SO_2 and most importantly there is no real trend in changes of peak quality along with changes in the variables in question (density and total SO_2).

We now take a look at the interactions of alcohol level, sugar level, and density*, as these variables are correlated and alcohol plays the largest role, as we have seen, in effecting the quality rating.*

We start off taking a look *below* at the raw data points relating the three variables. This plot shows exactly the relationships posited in the initial description of the variables - higher density associated with higher sugar level but lower alcohol levels. Along a given vertical strip tracking the density level, sugar level is uniformly increasing (more red) as the density increases; likewise, we see the density decrease as the alcohol increases, as shown in the negative slope of the figure of plots taken as a whole.

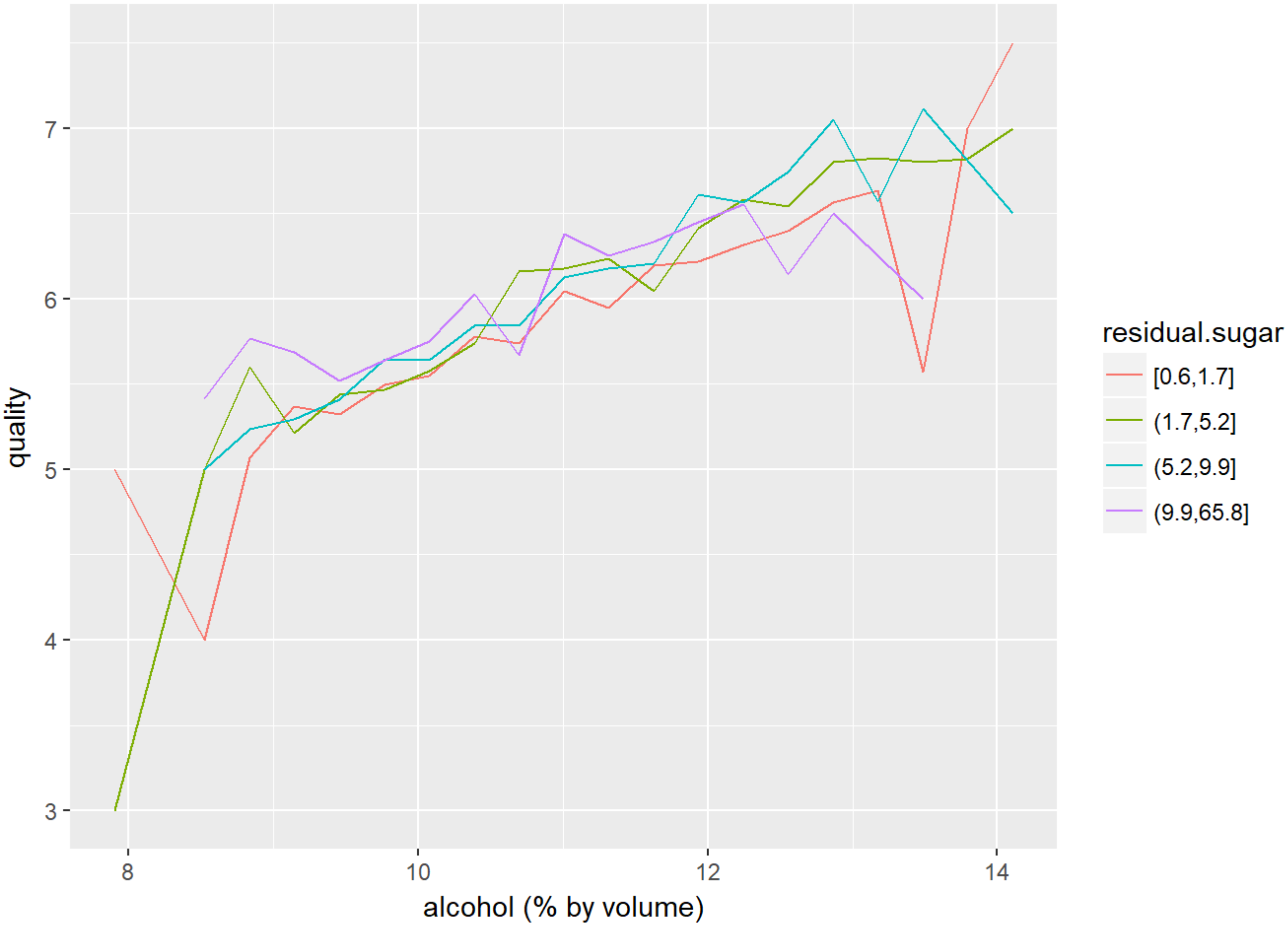


Now, we look at plots relating the variables as well as quality, starting with the below plot looking at quality versus density grouped by sugar level.

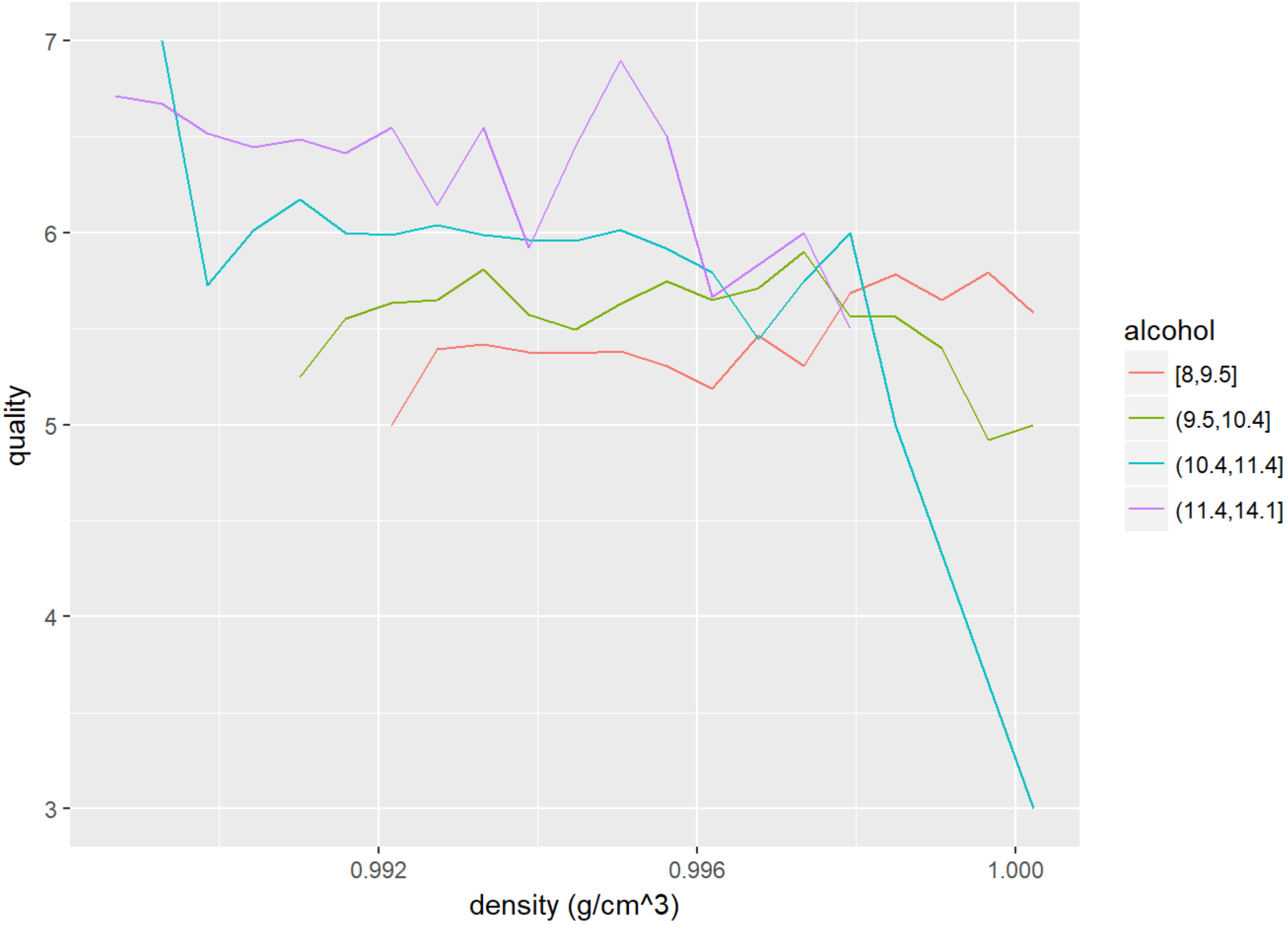


The above plot is interesting in how it shows what looks like roughly parallel lines for each group of sugar level. This seems to indicate that there is no interaction here between sugar and density in determining the quality - perhaps because only the density or only the sugar level matters.

We do a similar chart *below*, this time plotting the quality versus the alcohol level grouped by sugar level, and again see strikingly parallel lines.



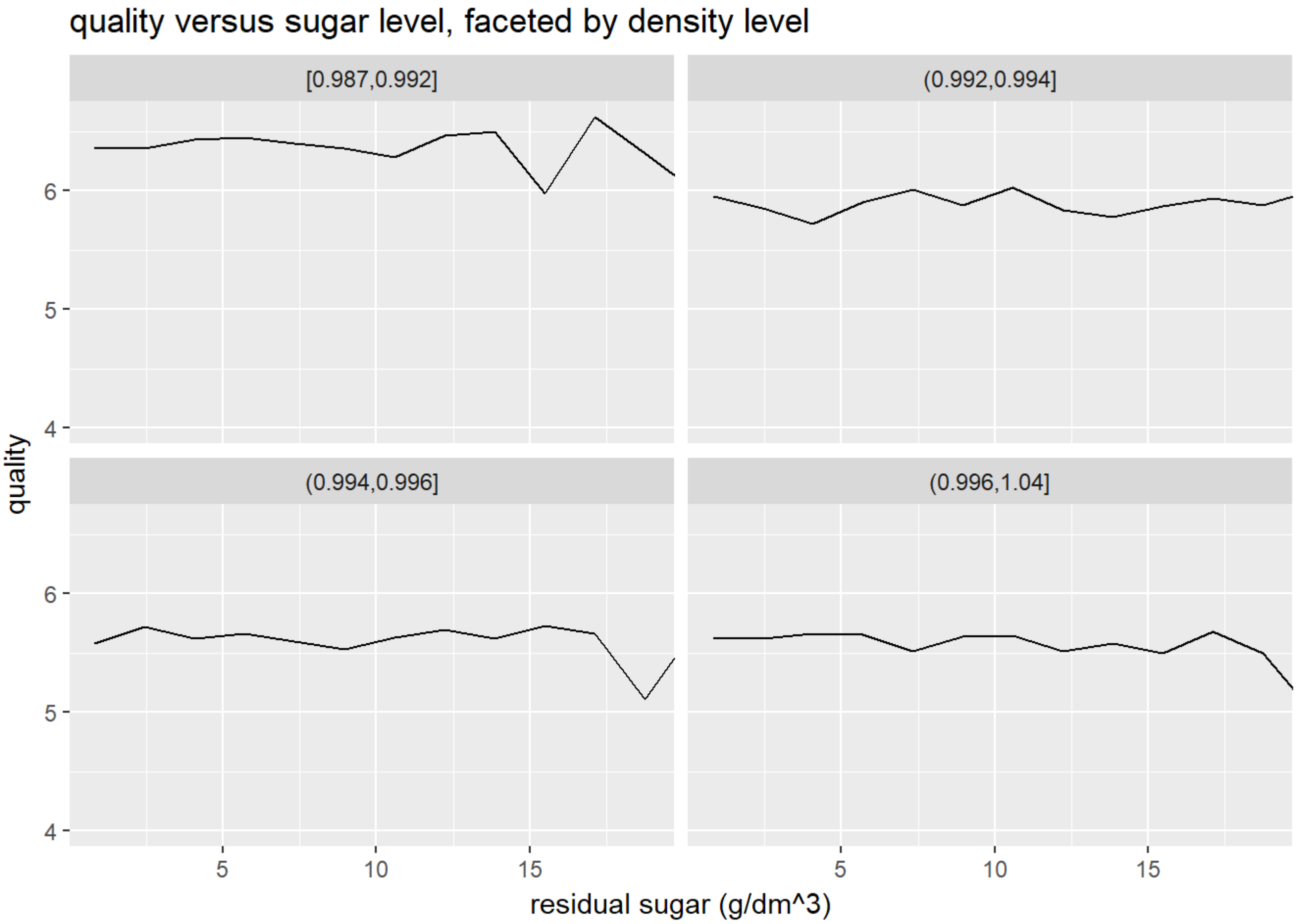
We now look at quality versus density colored by alcohol level *below*, and this time end up seeing something different that with the two previous plots.

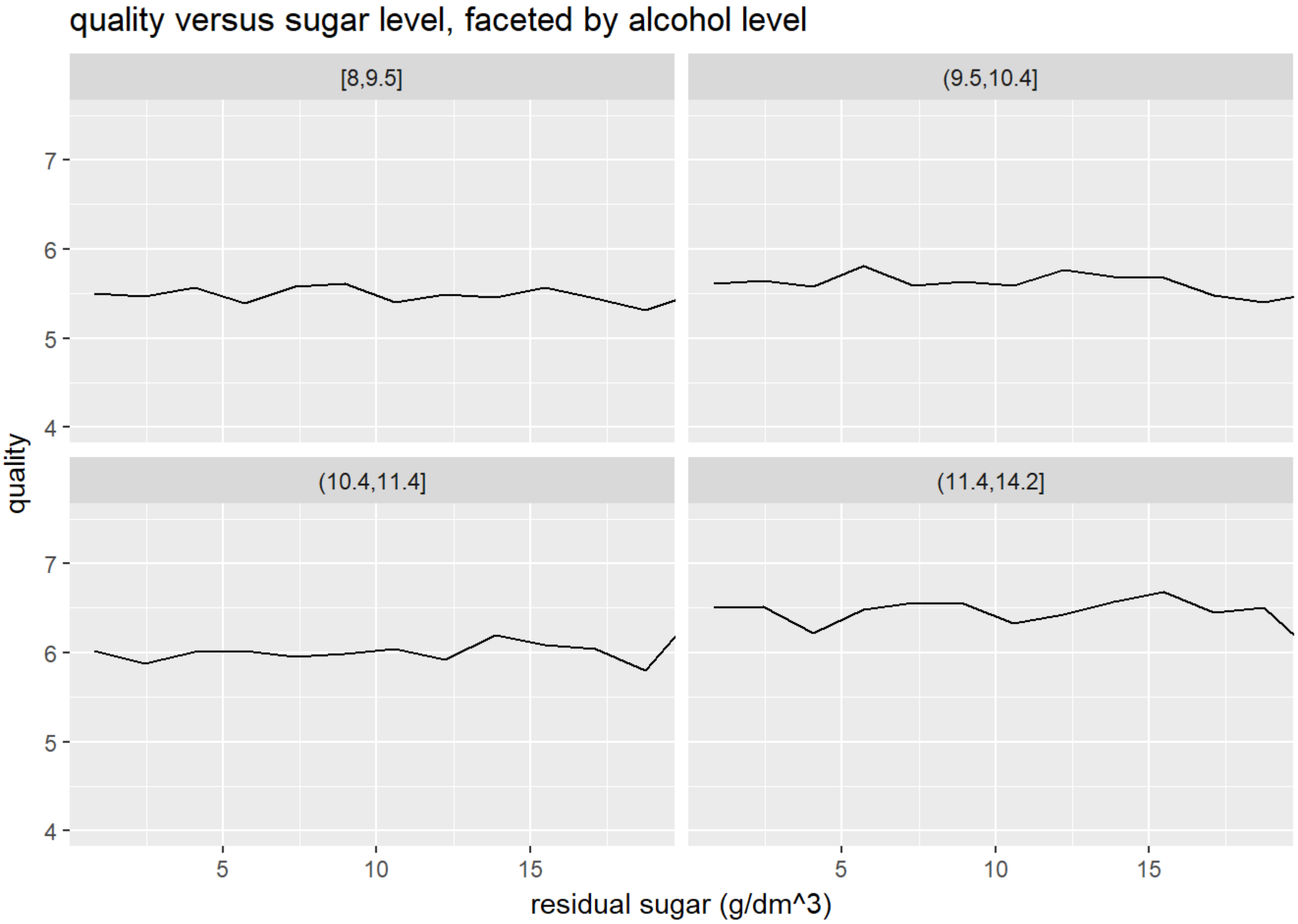


The plot above is interesting because it shows a definite change in the effect of alcohol level on quality which depends on density. It looks like there a critical density point (about .998) at which we go have the vertical order of the alcohol group lines reversing; that is, below the critical density, the the quality ordering was by increasing alcohol level, but after this density, it was by decreasing alcohol level.

So this is the opposite of the situation in the two previous charts, where we had parallel lines indicating a non-interaction of factors. Indeed we suspect now that alcohol and density interact in their effect on quality.

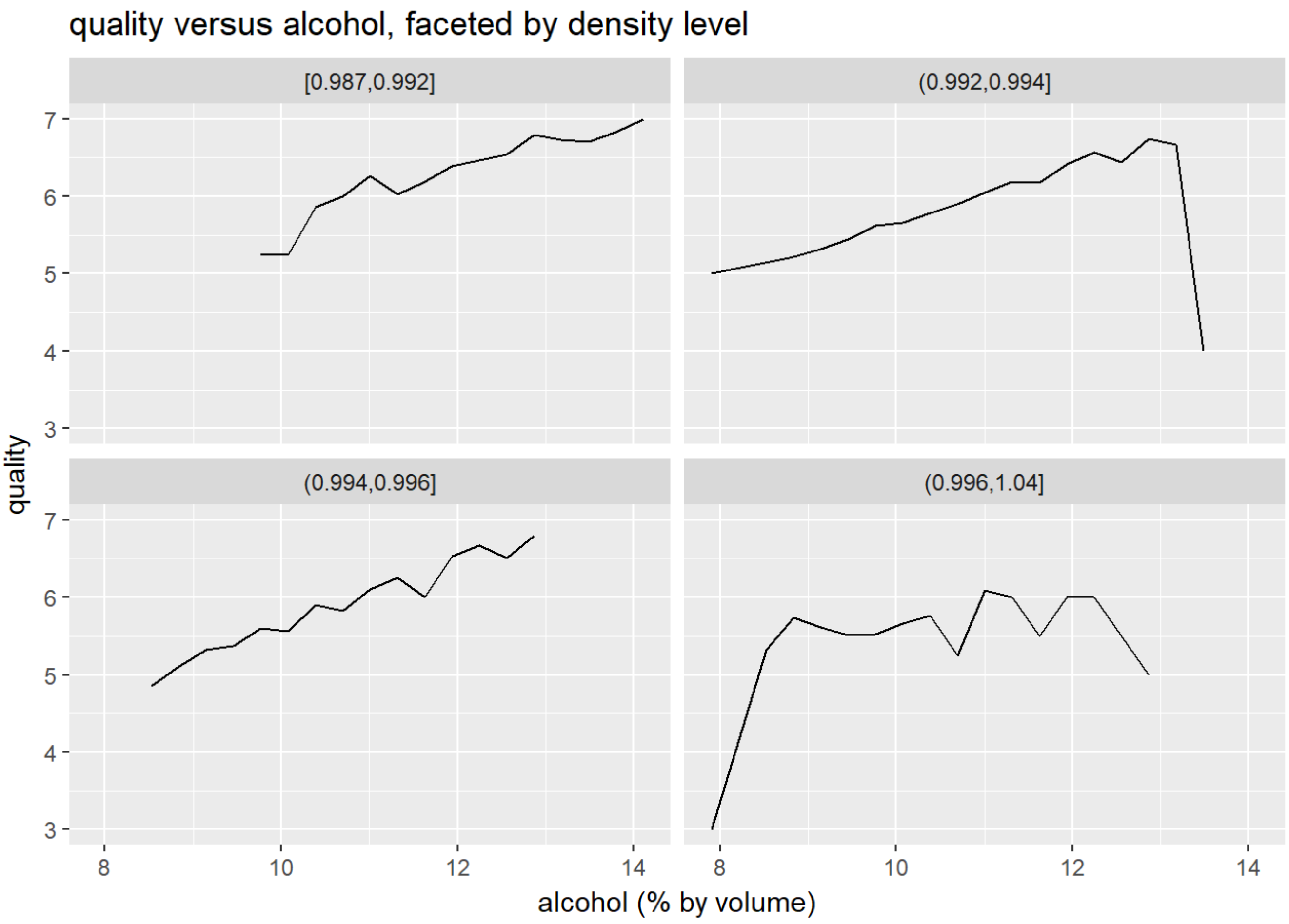
We see if these intuitions are still born out with the facet-based plots below. First, do we see the expected non-interactions of sugar level on the density and alcohol levels in regards to the later variables’ effect on quality? These plots are shown below.


quality versus sugar level, faceted by density level
[0.987,0.992]
(0.992,0.994]
(0.994,0.996]
(0.996,1.04]
quality
4
5
6
residual sugar (g/dm^3)
5
10
15


quality versus sugar level, faceted by alcohol level
[8,9.5]
(9.5,10.4]
(10.4,11.4]
(11.4,14.2]
quality
4
5
6
7
residual sugar (g/dm^3)
5
10
15

As expected, we see no interaction between sugar level and the other variables. In fact, once these variables are controlled for we see **no** effect of sugar level on density, as the plots are all flat! *So it looks like it would be safe to ignore the sugar level.*

Now we repeat with looking again at how density could affect or moderate alcohol, *below*.



And here we see our expected interaction above. At the lower density levels, we have the typical relationship where increasing alcohol is associated with increasing quality, but at the highest density level, quality does not go up with alcohol level.

Multivariate Analysis

We set out to see how some of the variables interacted together to help determine the quality ratings. We did this by looking at how variables affected any peak or optimal pH or total SO_2 measure in terms of quality rating. Is there an optimal or peak neighborhood or value, and can other variables shift this peak? We found a very well defined optimal total SO_2 level which appeared to be resistant to any peak shifts by other variables with the exception of alcohol level. It appeared that at higher alcohol levels, there is a peak shift or a higher optimal total SO_2 level. When doing the same analysis for pH, we found some hints of optimal pH but not such a well defined peak as in total SO_2, and we did not see any systematic shifting of the peak by interaction with other variables.

Based on this, it would seem there is an optimal level of total SO_2 that reflects how a minimal amount is needed to ward off microbial growth but having too large an amount starts to affect the smell and taste. It would seem that higher alcohol level favors a higher optimal SO_2 level, perhaps because with more alcohol we notice the sulfur less, or perhaps we need more in order to avoid the microbial growth. In retrospect, it is not too surprising there is not *one* optimal pH level, given that there might be different wine types with their own pH profile.

We also tried to separate out the effects of alcohol, density and sugar level, as alcohol and density appear to affect quality and these three variables are all correlated. Looking closer, we saw the correlations were the predictable sort where greater density is associated with greater sugar level and lesser alcohol level. We did not see any effect of sugar level on quality after controlling for density or alcohol level, so we narrowed in on alcohol and density in the analysis. Looking further we saw that density moderates the alcohol - at higher density levels, alcohol has less an effect on quality, and in fact it goes from being positively associated (more alcohol associated with more quality) to negatively associated at a certain critical density level.

These plots looking at the different variables in combination were interesting because we could see some very well defined and tractable relationships which were just mentioned. We got a better feel for how some variables may or may not a “sweet spot” and whether this sweet spot depends on other variables. We also saw in further detail how alcohol and density played their part in helping to determine the quality level.

We now take the intuitions gained from these investigations to see if they born out in a linear model.

```
> m1 <- lm(quality ~ alcohol, data = wines)
> m2 <- update(m1, ~ . + residual.sugar)
> m3 <- update(m2, ~ . + density)
> m4 <- update(m3, ~ . + total.sulfur.dioxide)
> m5 <- update(m4, ~ . + volatile.acidity )
> m6 <- update(m5, ~ . + I((volatile.acidity)^(1/2)))
> m7 <- update(m6, ~ . + I((total.sulfur.dioxide)^(2)))
> m8 <- update(m7, ~ . + pH * alcohol)
> m9 <- update(m8, ~ . + alcohol * residual.sugar * density)
> m10 <- update(m9, ~ . + pH * sulphates)
> mtable(m1, m2, m3, m4, m5)

Calls:
m1: lm(formula = quality ~ alcohol, data = wines)
m2: lm(formula = quality ~ alcohol + residual.sugar, data = wines)
m3: lm(formula = quality ~ alcohol + residual.sugar + density, data = wines)
m4: lm(formula = quality ~ alcohol + residual.sugar + density + total.sulfur.dioxide,
    data = wines)
m5: lm(formula = quality ~ alcohol + residual.sugar + density + total.sulfur.dioxide +
    volatile.acidity, data = wines)

==============================================================================================
                             m1            m2            m3            m4            m5
----------------------------------------------------------------------------------------------
  (Intercept)               2.582***      2.021***     90.313***     93.661***     82.245***
                           (0.098)       (0.117)      (12.374)      (12.665)      (12.231)
  alcohol                   0.313***      0.354***      0.246***      0.246***      0.288***
                           (0.009)       (0.010)       (0.018)       (0.018)       (0.018)
  residual.sugar                          0.022***      0.053***      0.054***      0.053***
                                         (0.002)       (0.005)       (0.005)       (0.005)
  density                                             -87.886***    -91.314***    -79.761***
                                                      (12.317)      (12.623)      (12.191)
  total.sulfur.dioxide                                                0.000         0.001**
                                                                     (0.000)       (0.000)
  volatile.acidity                                                                 -2.093***
                                                                                   (0.109)
----------------------------------------------------------------------------------------------
  R-squared                 0.190         0.202         0.210         0.210         0.265
  adj. R-squared            0.190         0.202         0.210         0.210         0.265
  sigma                     0.797         0.791         0.787         0.787         0.759
  F                      1146.395       619.354       434.085       325.984       353.570
  p                         0.000         0.000         0.000         0.000         0.000
  Log-likelihood        -5839.391     -5802.158     -5776.812     -5776.042     -5599.111
  Deviance               3112.257      3065.298      3033.737      3032.783      2821.404
  AIC                   11684.782     11612.317     11563.624     11564.084     11212.223
  BIC                   11704.272     11638.303     11596.107     11603.064     11257.699
  N                      4898          4898          4898          4898          4898
==============================================================================================
> mtable(m6, m7, m8, m9, m10)

Calls:
m6: lm(formula = quality ~ alcohol + residual.sugar + density + total.sulfur.dioxide +
    volatile.acidity + I((volatile.acidity)^(1/2)), data = wines)
```

```
m7: lm(formula = quality ~ alcohol + residual.sugar + density + total.sulfur.dioxide + 
    volatile.acidity + I((volatile.acidity)^(1/2)) + I((total.sulfur.dioxide)^(2)), 
    data = wines)
m8: lm(formula = quality ~ alcohol + residual.sugar + density + total.sulfur.dioxide + 
    volatile.acidity + I((volatile.acidity)^(1/2)) + I((total.sulfur.dioxide)^(2)) + 
    pH + alcohol:pH, data = wines)
m9: lm(formula = quality ~ alcohol + residual.sugar + density + total.sulfur.dioxide + 
    volatile.acidity + I((volatile.acidity)^(1/2)) + I((total.sulfur.dioxide)^(2)) + 
    pH + alcohol:pH + alcohol:residual.sugar + alcohol:density + 
    residual.sugar:density + alcohol:residual.sugar:density, 
    data = wines)
m10: lm(formula = quality ~ alcohol + residual.sugar + density + total.sulfur.dioxide + 
    volatile.acidity + I((volatile.acidity)^(1/2)) + I((total.sulfur.dioxide)^(2)) + 
    pH + sulphates + alcohol:pH + alcohol:residual.sugar + alcohol:density + 
    residual.sugar:density + pH:sulphates + alcohol:residual.sugar:density, 
    data = wines)

================================================================================================
============
                                           m6            m7            m8            m9         
  m10     
------------------------------------------------------------------------------------------------
------------
  (Intercept)                            93.455***     94.190***    117.777***   -157.465*     -
149.384*  
                                        (12.383)      (12.244)      (12.573)      (70.064)      
 (69.745)   
  alcohol                                 0.281***      0.281***     -0.922***     26.350***    
 28.108***  
                                         (0.018)       (0.018)       (0.196)       (6.325)      
 (6.301)    
  residual.sugar                          0.057***      0.057***      0.063***     -8.288       
 -7.990     
                                         (0.005)       (0.005)       (0.005)       (4.513)      
 (4.491)    
  density                               -89.738***    -91.209***   -103.692***    174.198*      
169.157*    
                                        (12.305)      (12.168)      (12.596)      (70.682)      
 (70.342)   
  total.sulfur.dioxide                    0.001***      0.013***      0.012***      0.013***    
  0.013***  
                                         (0.000)       (0.001)       (0.001)       (0.001)      
 (0.001)    
  volatile.acidity                        2.031*        2.422**       2.093**       2.101**     
  2.116**   
                                         (0.792)       (0.784)       (0.780)       (0.785)      
 (0.782)    
  I((volatile.acidity)^(1/2))            -4.680***     -5.038***     -4.653***     -4.781***    
 -4.760***  
                                         (0.890)       (0.881)       (0.876)       (0.882)      
 (0.878)    
  I((total.sulfur.dioxide)^(2))                        -0.000***     -0.000***     -0.000***    
 -0.000***  
                                                       (0.000)       (0.000)       (0.000)      
 (0.000)
  pH                                                                 -3.474***     -3.361***
 -4.303***
                                                                     (0.657)       (0.670)
 (0.721)
  alcohol x pH                                                        0.372***      0.362***
  0.344***
                                                                     (0.062)       (0.063)
 (0.063)
  alcohol x residual.sugar                                                          0.664
  0.630
                                                                                   (0.391)
 (0.389)
  alcohol x density                                                               -27.534***
 -29.283***
                                                                                   (6.381)
 (6.356)
  residual.sugar x density                                                          8.230
  7.924
                                                                                   (4.524)
 (4.502)
  alcohol x residual.sugar x density                                               -0.652
 -0.616
                                                                                   (0.392)
 (0.390)
  sulphates
 -6.470***

 (1.870)
  pH x sulphates
  2.210***

 (0.585)
------------------------------------------------------------------------------------------------
------------
  R-squared                               0.270         0.286         0.296         0.302
  0.309
  adj. R-squared                          0.269         0.285         0.295         0.300
  0.307
  sigma                                   0.757         0.749         0.744         0.741
  0.737
  F                                     300.849       279.847       228.885       162.502
145.587
  p                                       0.000         0.000         0.000         0.000
  0.000
  Log-likelihood                    -5585.318     -5529.548     -5493.388     -5474.324     -5
449.192
  Deviance                           2805.558      2742.390      2702.196      2681.242      2
653.868
  AIC                               11186.637     11077.095     11008.776     10978.648     10
932.385
  BIC                               11238.610     11135.565     11080.239     11076.096     11
042.827
  N                                    4898          4898          4898          4898          4
```

```
898
================================================================================================
============
```

Looking at the results of the regression model, we see that the previous observations about the relative association to quality among variables was mostly born out in the regression model. The big caveat was that this linear regression model had a very low R^2, which is a measure of how much variance in quality can be explained by the linear model, so it would probably not be very useful in making predictions. The question then is whether it is useful in confirming our intuitions that we get from the plots. The answer is a qualified Yes. The variables that had the more well behaved relationships to quality were also the variables with the higher correlations. This was the case for the variables alcohol, density, volatile acidity and free sulfur dioxide. Notably, chlorides did not have any benefit to the model, as noted before, despite being the variable with the third highest correlation with quality. Chlorides as well as many of the other variables under investigation were tried out in the model and created only a minimal benefit, increasing the R^2 but at most .002 (not shown), and so were not included in the final model. It is somewhat encouraging that the picture obtained from the plots against quality might have been more informative in this regard than the correlation coefficient.

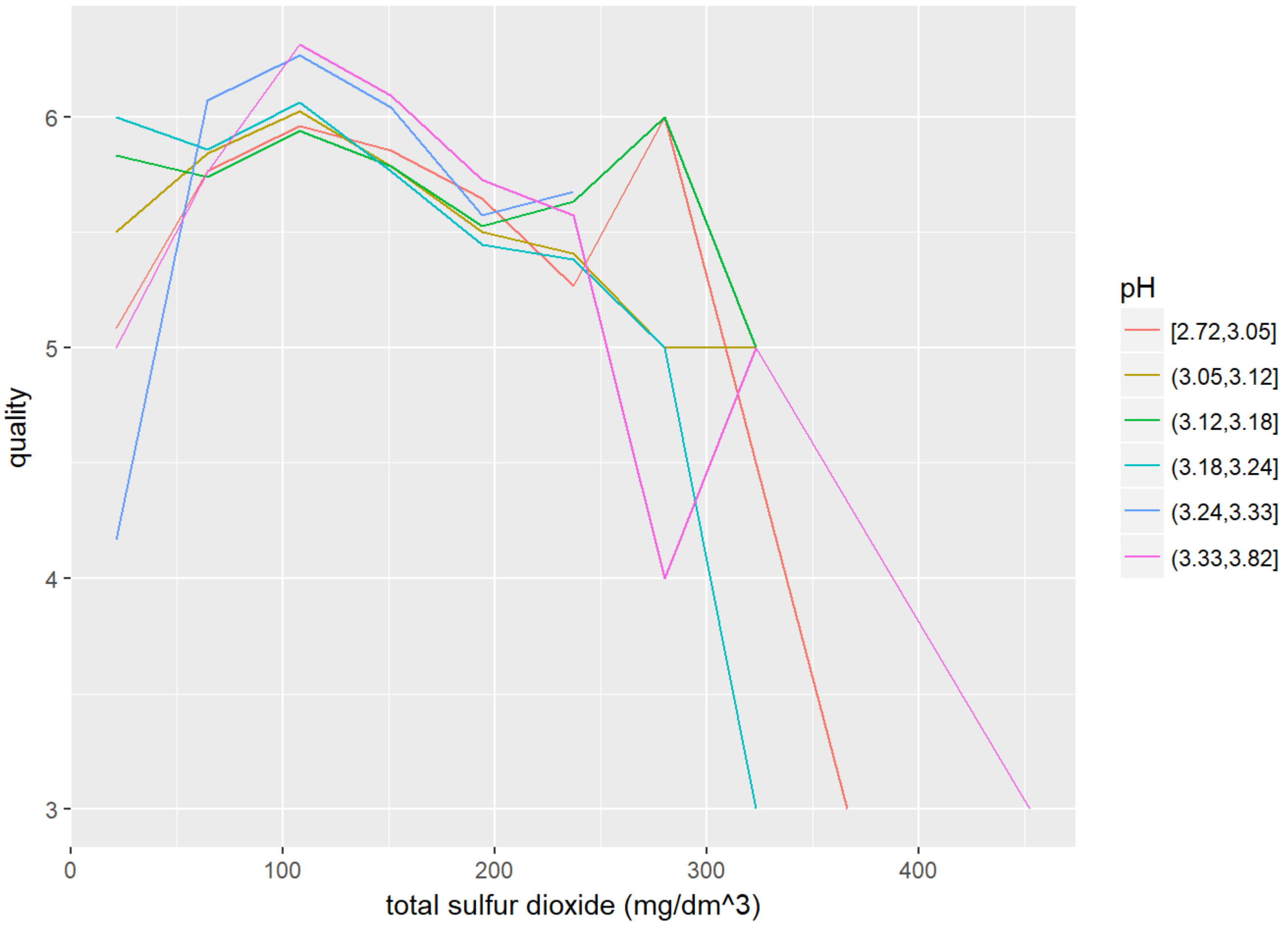
The variables which were suitably transformed also were more impactful in the model in their transformed mode than in their original mode, or at least, we needed to include the transformed mode to get any benefit from them in the model. Total sulfur was predicted to have it’s squared value proportional to quality based on the transformed plots, and in agreement with this, its squared value was also more significant than the untransformed value in the regression model, while using both the squared and untransformed terms provided the biggest benefit to the model. Similarly, volatile acidity had a more linear plot against quality after a log or square-root transformation, and these terms were also provided a bigger benefit to the regression model compared to the term for the untransformed volatile acidity value.

Unfortunately the interactions glimpsed in the plots had only a (relatively) minor effect in the regression model. There was no benefit to include the nice, well behaved alcohol by total sulfur dioxide interaction, for example. In contrast, the alcohol by chlorides interaction was about as significant as any other interaction, despite not seeing a meaningful relationship in any previous plots with these variables together.

Final Plots and Summary

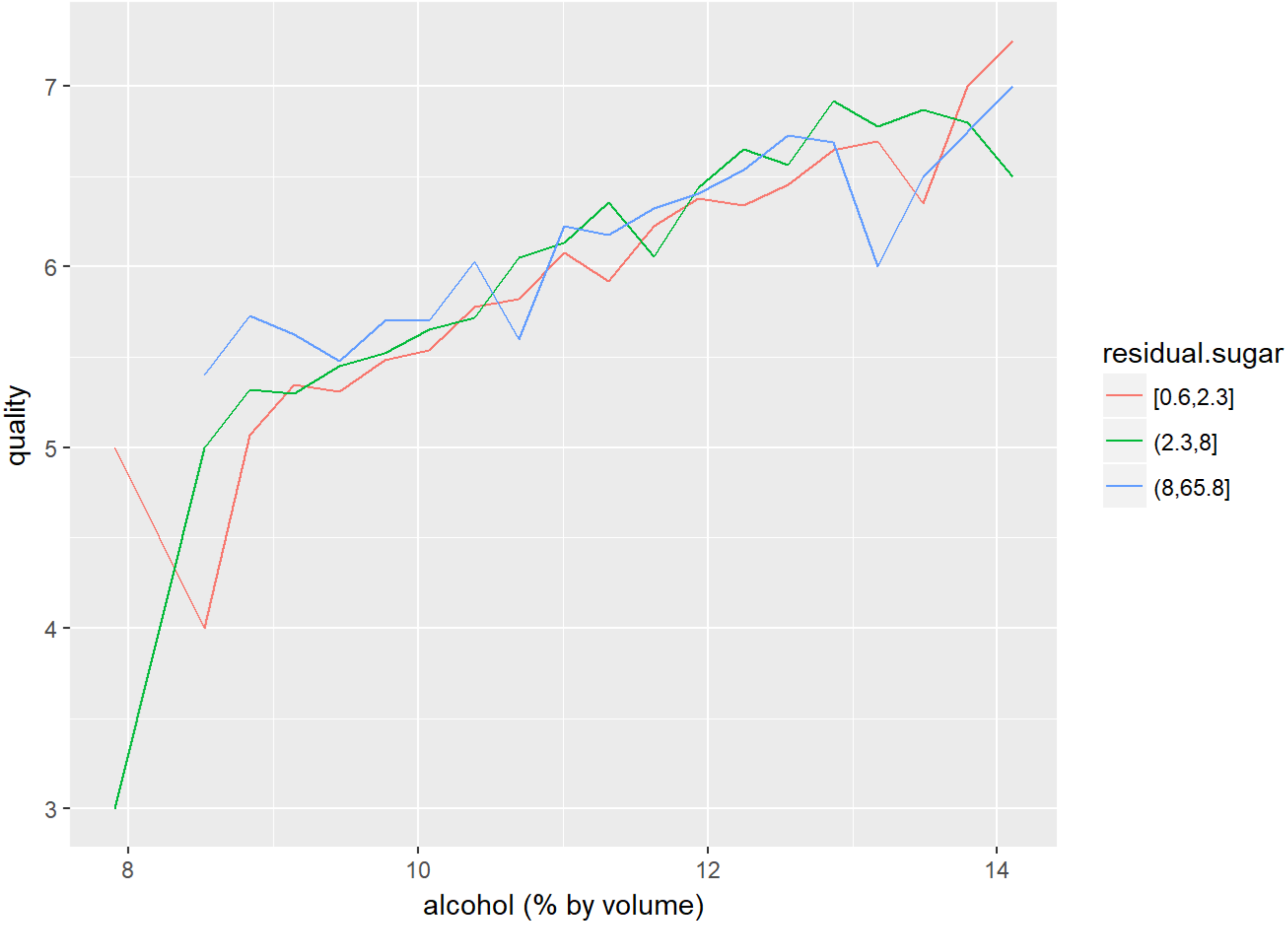
We now summarize some of the main findings with three plots.

Plot One



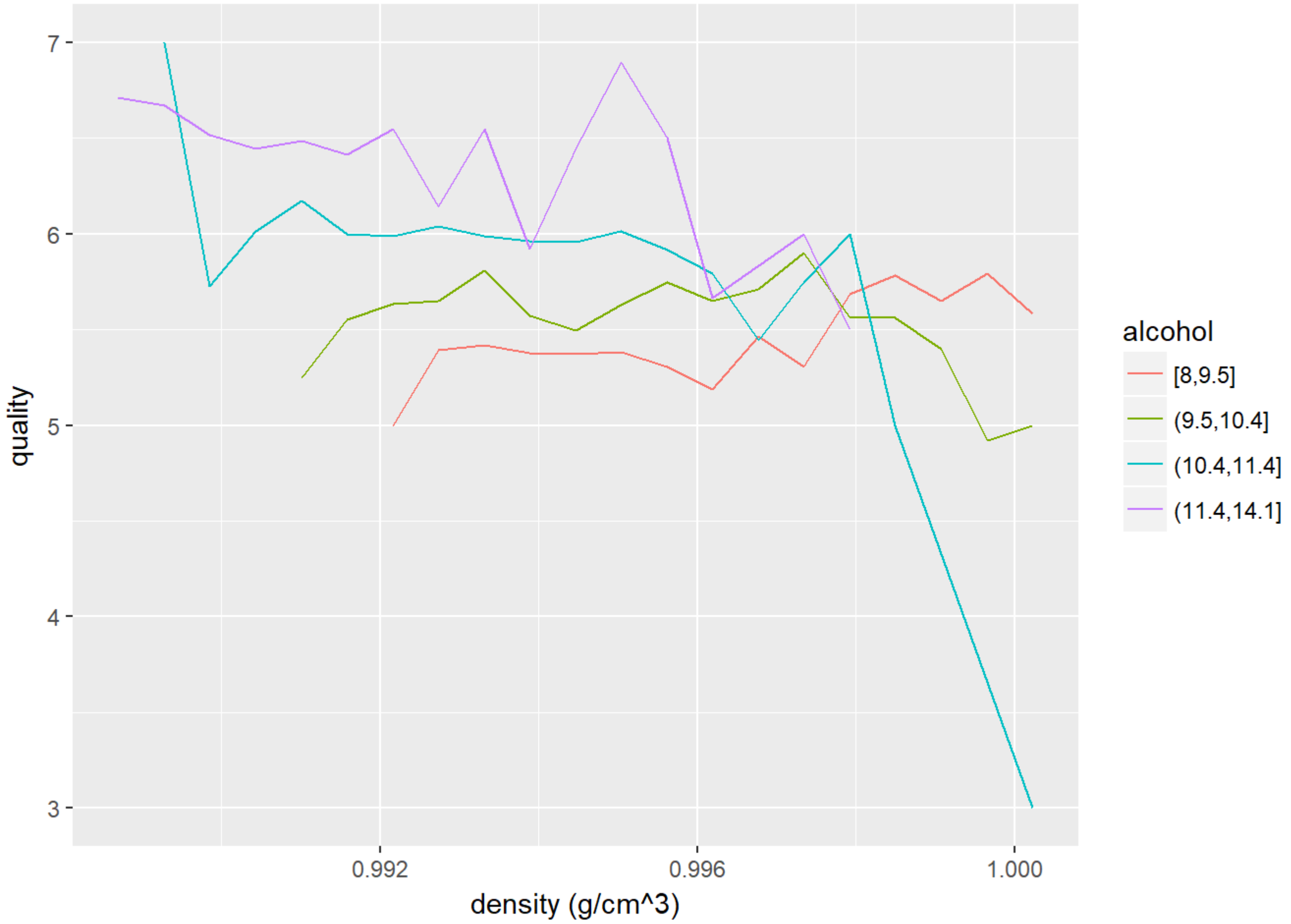
Above, we see a plot total SO_2 grouped by pH level plotted against quality. *This plot is notable because it shows a well defined an optimal point* where we have a “sweet spot” for total SO_2. This is interesting because it looks like it is quite a narrow region of SO_2 level, and moreover it appears to be **totally unaffected by the pH level.** Despite the lack of solid predictors for quality besides alcohol level (based on Pearson’s correlation coefficient), we apparently have a very solid relationship between the total SO_2 and quality, because there is no other way to explain the orderliness in this plot.

Plot Two



Next we see a very orderly relationship, but of a different nature. This *above* is a plot of the relationship of quality to the alcohol level grouped by the (residual) sugar level. It is notable how parallel the colored lines are, reflecting a lack of effect of sugar level on the effect that alcohol has on the quality. This was born out with other plots shown earlier in which we see no effect of sugar level on quality when you control for the alcohol level.

Plot Three



And finally we do see a significant interaction here among the variables affecting quality. Below a density level of about .998, we have increases in quality ratings associated with *increases* in alcohol levels, whereas above this density, increases in quality are associated with *decreasing* alcohol levels.

To summarize, we do see some very solid, tractable relationships of some variables to the quality ratings. Given this, it is somewhat surprising that we do not get many variables with stronger correlations to quality, or an effective linear regression model using these relationships and interactions.

Reflections

The most encouraging result of this investigation was that we were able to verify that there were nontrivial, tractable relationships between some of the variables analyzed and the quality, in spite of the rather meek correlation coefficients associated with the quality ratings. We saw regular, predictable line plots in some cases when plotting the variables' levels against average quality, and this persisted in some cases even when grouping by other variables.
Also, we were able to get some "straitened" plots of some variables by transforming with square, square root, or log transformations, and these turned out to be as or more important than their untransformed measures in the linear model we created.

There were some definite limitations to these investigations. We could not show any tractable, predictable trend of chlorine level versus quality, despite the chlorine level having the third highest (Pearson's) correlation with quality. Also, the linear model as a whole was not terribly useful aside from verifying a few intuitions, as it had a very low R^2 amount of variance in quality explained by the model.

One potential improvement to the dataset would have been to use something like wine *types*, such as white zinfandel. It is possible that the physical variables can only have so much explanatory power, given that there could in theory be a different physical profile (say, higher or lower pH or alcohol content) of some wine types. ***Or perhaps, as the saying goes, "There is no accounting for taste!"***