Red Wine Analysis using R, Robert H Lee, May 2017

The following is conducted on the well-known UCI public wine dataset:
https://archive.ics.uci.edu/ml/datasets/wine+quality

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE)
library(ggplot2)
library(gridExtra)
```

Red Wine dataframe (redwine)

First we take a look at the head and the structure of the dataframe. The info sheet gives us the units, for instance grams per cubic decimeter, for many of the columns, % by volume for alcohol, which is a unit most of us are familiar with, density (water is ~1) and quality, apparently a unitless integer between 0 and 10. There are 13 variables. One of them is the row number (column X), so we remove that as extraneous. 'redwine$X <- NULL'

The structure of the dataframe below reveals no surprises, including that the quality rating is indeed an integer, and that the other colums are numbers (floats) with zero or one or two decimal places.

There are 1599 rows in the dataset

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlori
des
## 1 1           7.4             0.70        0.00            1.9     0.
076
## 2 2           7.8             0.88        0.00            2.6     0.
098
## 3 3           7.8             0.76        0.04            2.3     0.
092
## 4 4          11.2             0.28        0.56            1.9     0.
075
## 5 5           7.4             0.70        0.00            1.9     0.
076
## 6 6           7.4             0.66        0.00            1.8     0.
075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates al
cohol
## 1                  11                   34  0.9978 3.51      0.56
9.4
## 2                  25                   67  0.9968 3.20      0.68
9.8
## 3                  15                   54  0.9970 3.26      0.65
9.8
## 4                  17                   60  0.9980 3.16      0.58
```

```
9.8
## 5                  11                   34  0.9978 3.51      0.56
9.4
## 6                  13                   40  0.9978 3.51      0.56
9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5

## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7
.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0
.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1
...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.
069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39
3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.4
7 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.
5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...

## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7
.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0
.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1
...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.
069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39
3.36 3.35 ...
```

```
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.4
7 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.
5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Univariate Analysis

Wine experts, oenelogists, sommeliers might care about some of these scientific variables, but the average person just cares about taste and quality. So let's look at that. Here are 250 data points. Most of the values are towards the middle of th 0-10 scale. we wonder if this is becasue human judges tend to pick toward the middle of the scale and more rarely give out more extreme scores, or because of something else.

```
## Head: quality

##  [1] 5 5 5 6 5 5 5 7 7 5 5 5 5 5 5 5 7 5 4 6
```

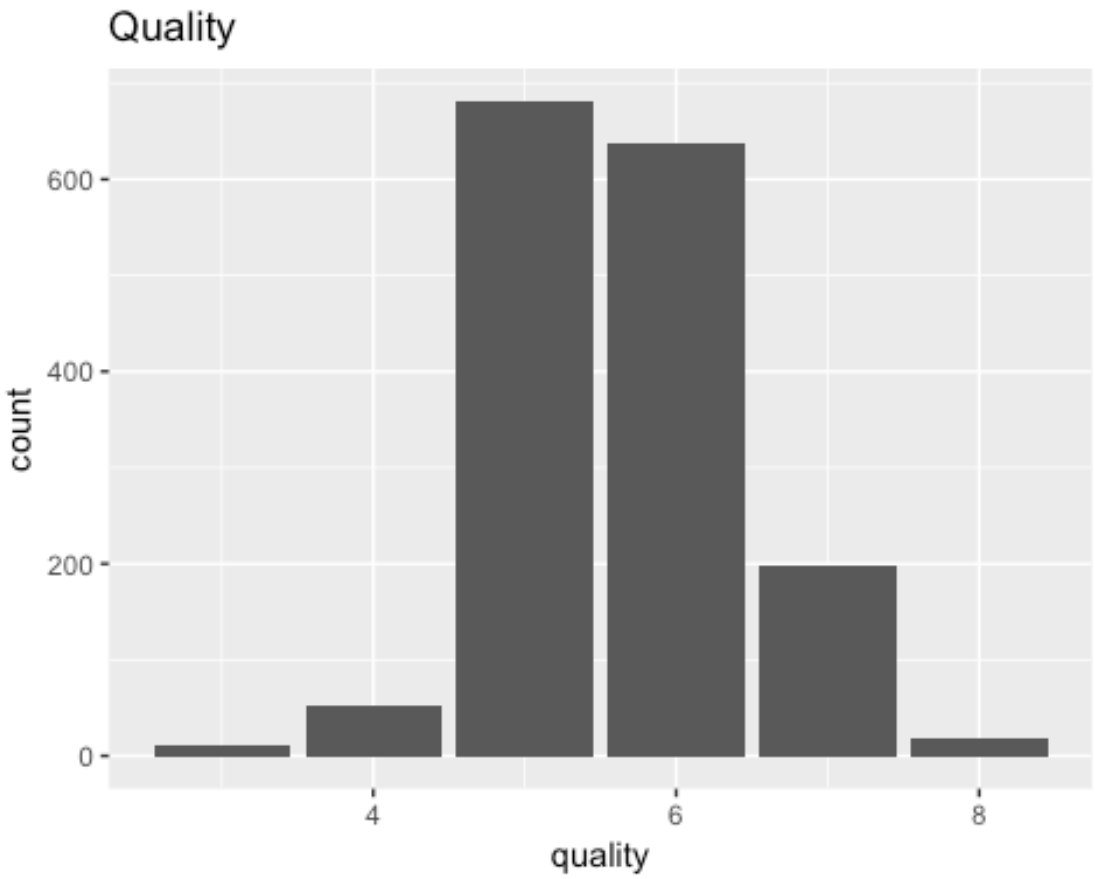
Wine Ratings 3-8

The histogram of all the data points goes only from 3-8. This is a bit surprising, but then we remember that all this data is only from one specific type of wine, which may not have any high quality wines at all. From the data overview:

"the datasets are related to red and white variants of the Portuguese "Vinho Verde" wine."

I confirm this range with the summary and the histogram. So rather than representing some kind of judging bias toward middle-tier scores, this may be an artifcat of the wine selected for analysis.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```



Acidity

Let's next look at the 3 measures of acidity. The mean and range of Fixed Acidity are about an order of magnitude greater than Volatile Acidity and Citric Acid. Unless the latter two types of acid are much stronger than the first (unlikely), then the effect the first acidity on pH should also be about an order of magnitude stronger. So we should see a stronger correlation between the first and pH, and who knows, perhaps quality, in the bivariate analysis.

3 acids

Fixed acidity and volatile acidity are fairly normal with right skew. For citric acid, gnoring the zero value (which might represent bad data or "unknown" values defaulted to zero, this variable is fairly evenly distributed.

Summary, Range, IQR, Head

Fixed acidity Volatile acidity Citric acid

```
## Fixed acidity:

## summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.60    7.10    7.90    8.32    9.20   15.90 

## range

## [1] 11.3

## IQR

## [1] 2.1

## Volatile acidity:

## summary

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

## range

## [1] 1.46

## IQR

## [1] 0.25

## Citric acid:

## summary

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.090   0.260   0.271   0.420   1.000 

## range

## [1] 1

## IQR

## [1] 0.33

## Head: fixed.acidity

##  [1]  7.4  7.8  7.8 11.2  7.4  7.4  7.9  7.3  7.8  7.5  6.7  7.5  5.
6  7.8
## [15]  8.9  8.9  8.5  8.1  7.4  7.9

## Head: volatile.acidity

##  [1] 0.700 0.880 0.760 0.280 0.700 0.660 0.600 0.650 0.580 0.500 0.5
80
## [12] 0.500 0.615 0.610 0.620 0.620 0.280 0.560 0.590 0.320

## Head: citric.acid
```

```
##  [1] 0.00 0.00 0.04 0.56 0.00 0.00 0.06 0.00 0.02 0.36 0.08 0.36 0.0
0 0.29
## [15] 0.18 0.19 0.56 0.28 0.08 0.51
```

New variable - totalacid

I decide it would be useful to add up all the acid values (they have the same units). Assuming again, that no one acid is much stronger than the others (which is reasonable), this total acid value is a simple proxy for the three separate values. At the same time, Fixed acid would be a similarly good proxy, since it is an order of magnitude greater than the others in terms of mean and range.

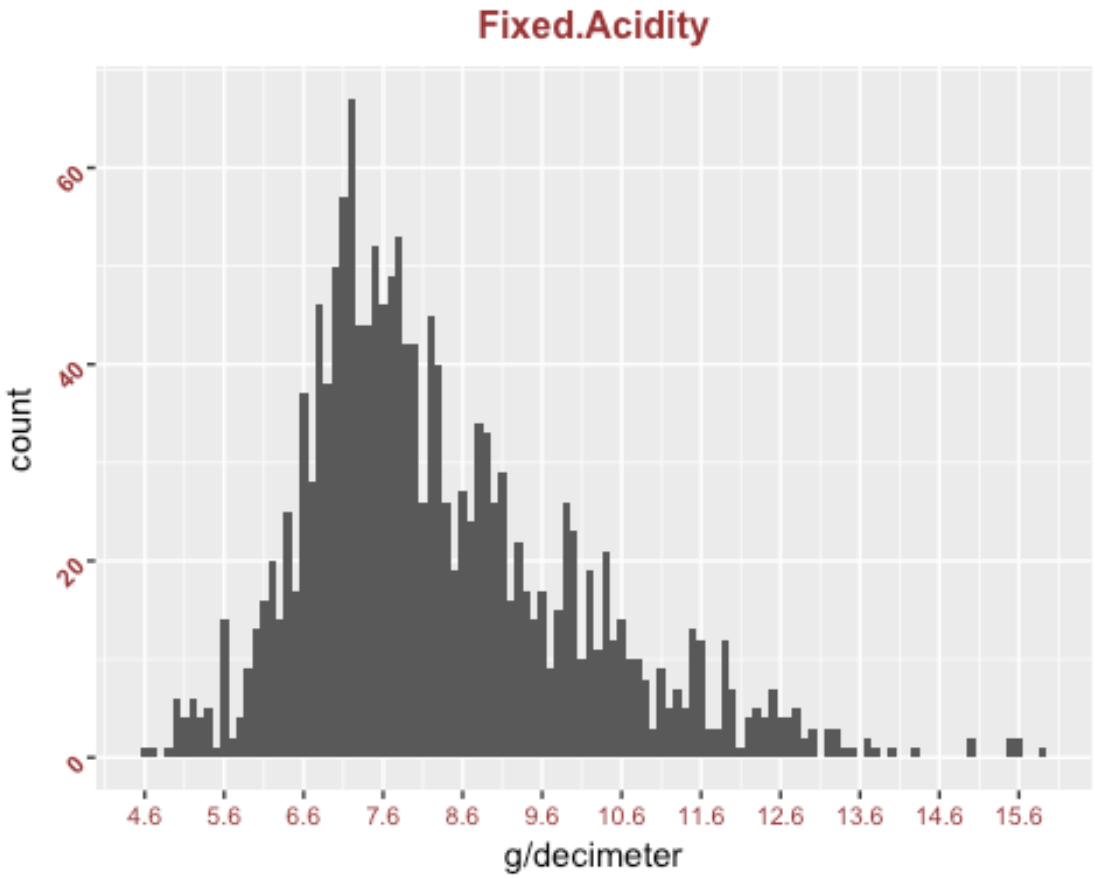
Summary, Range, IQR, Head

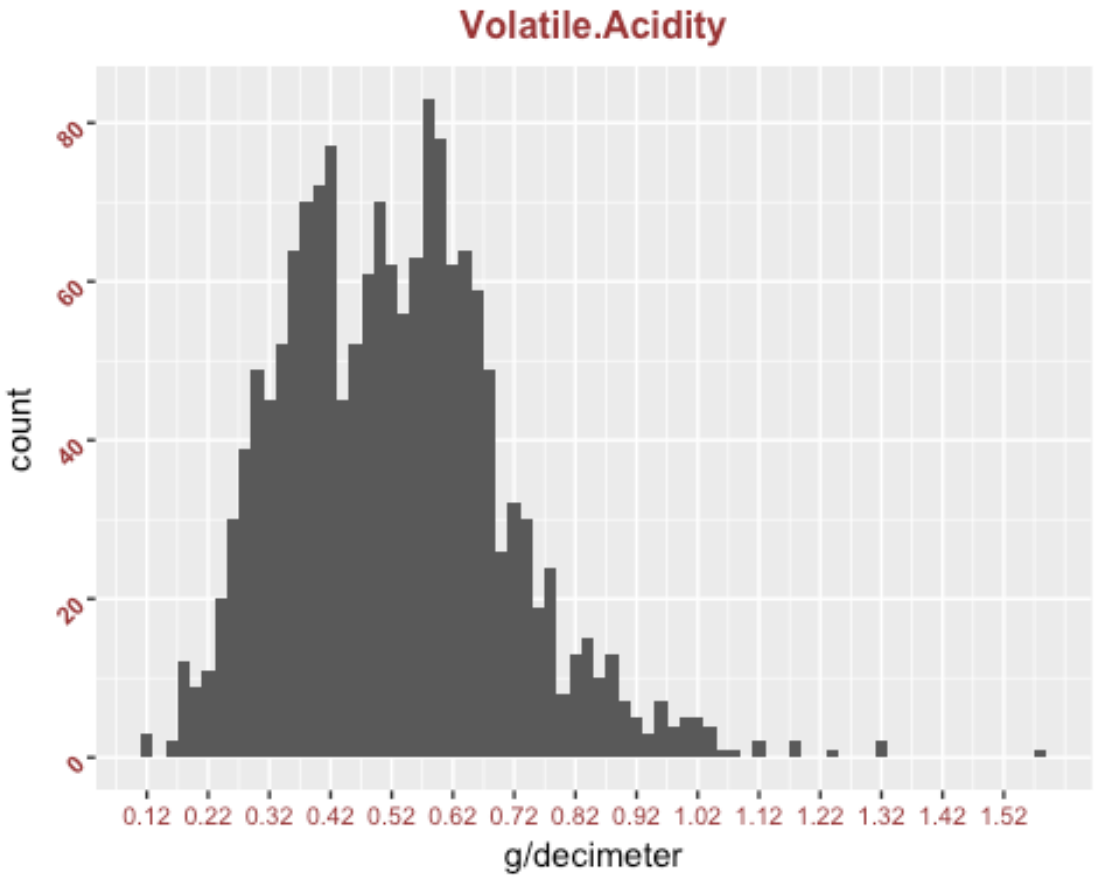
Total acid

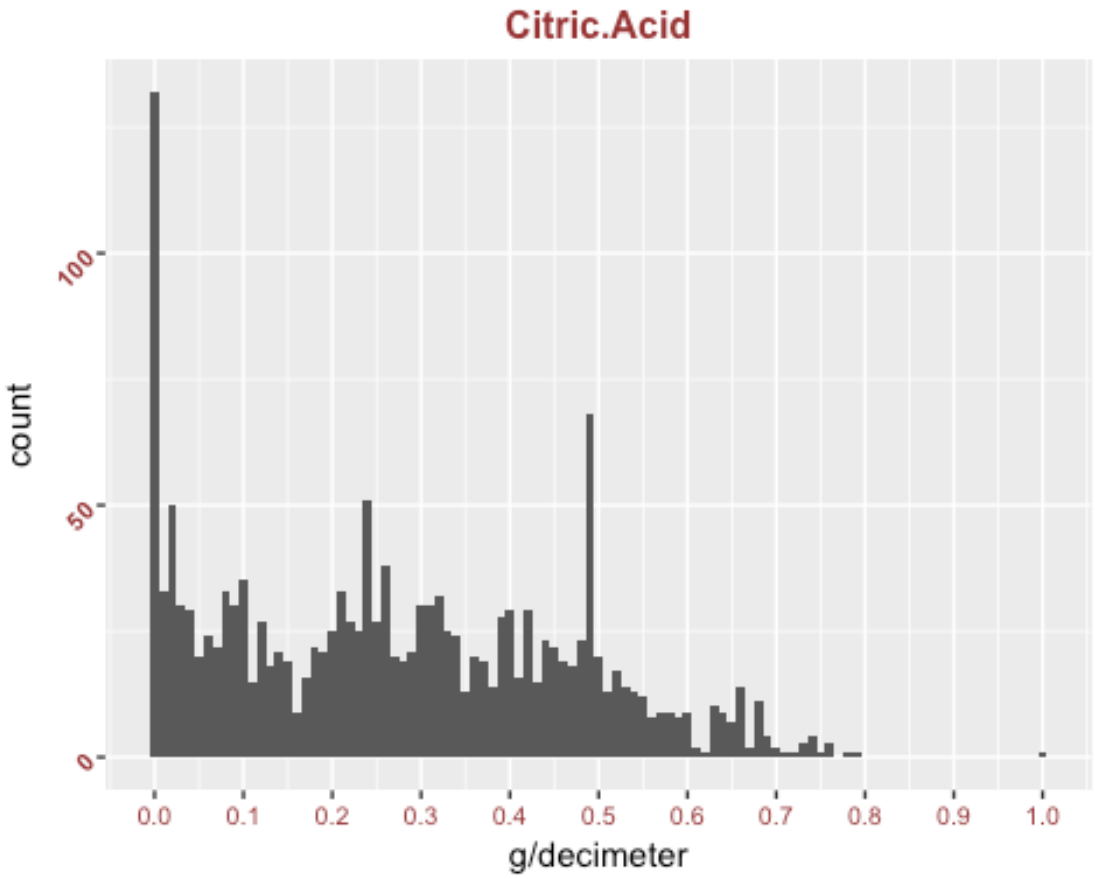
```
## Summary

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.30    7.80    8.70    9.12   10.10   17.00

## range

## [1] 11.7

## IQR

## [1] 2.3

## Head: total.acid

##  [1]  8.1  8.7  8.6 12.0  8.1  8.1  8.6  8.0  8.4  8.4  7.4  8.4  6.
2  8.7
## [15]  9.7  9.7  9.3  8.9  8.1  8.7
```

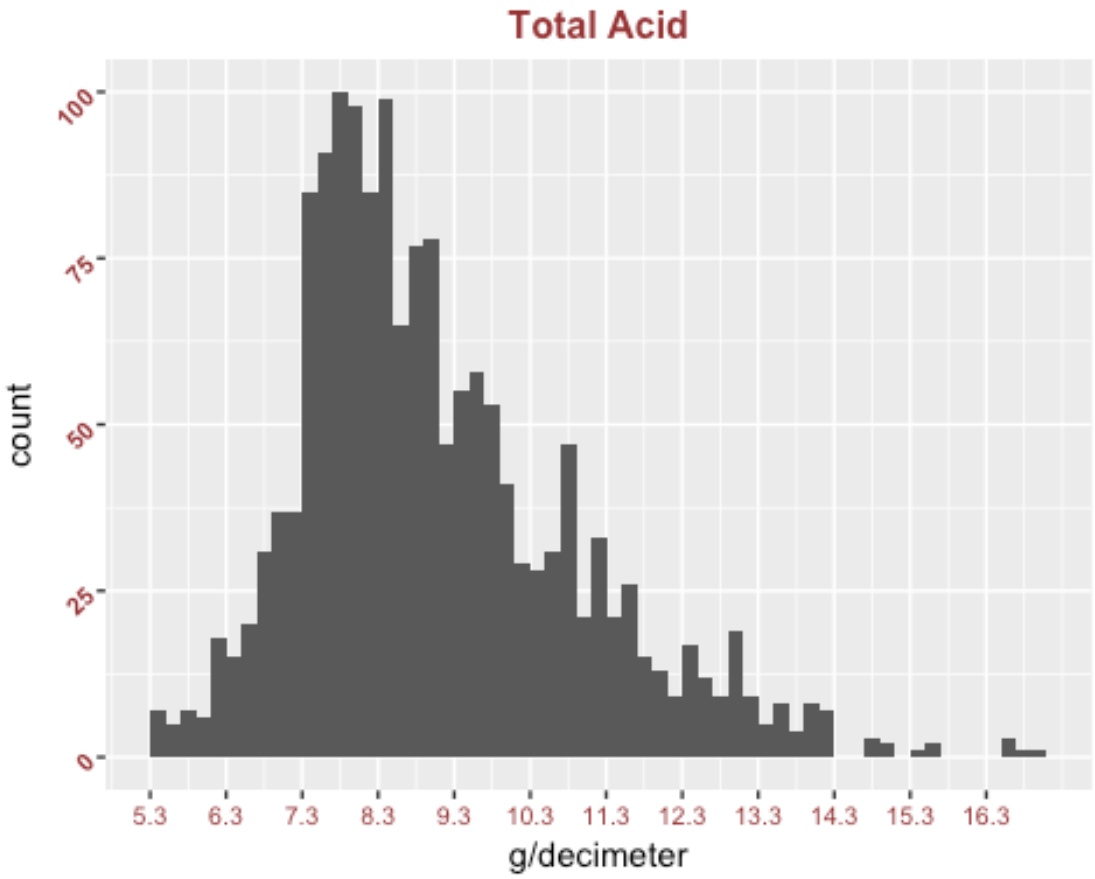
4 acids

Next we plot all four acid variables, including totalacid, which we created


Fixed.Acidity
count
60
40
20
0
4.6
5.6
6.6
7.6
8.6
9.6
10.6
11.6
12.6
13.6
14.6
15.6
g/decimeter

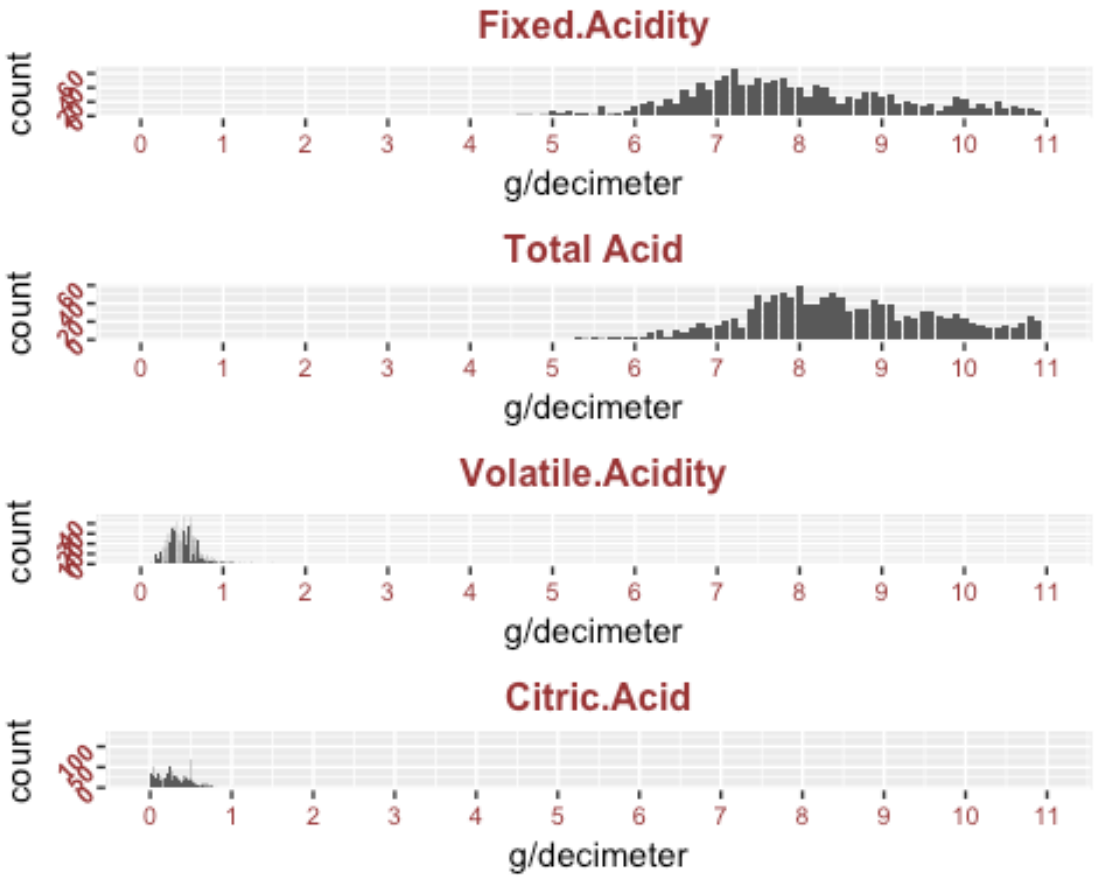

Volatile.Acidity
count
80
60
40
20
0
0.12
0.22
0.32
0.42
0.52
0.62
0.72
0.82
0.92
1.02
1.12
1.22
1.32
1.42
1.52
g/decimeter


Citric.Acid
count
100
50
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
g/decimeter



Plotting acids together

I next plot these on the same scales to get a feel for the fact that indeed, total acid and fixed acidity are similar, and one can be somewhat interchangeable for each other.



Sugar, alcohol, density

Next we look at (1) residual sugar (g/dm3) (2) alcohol (% by volume) (3) density (water ~ 1)

Summary, Range, IQR, Head

```
## Residual Sugar:

## summary

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500

## range

## [1] 14.6

## IQR

## [1] 0.7

## Residual Alcohol:

## summary
```

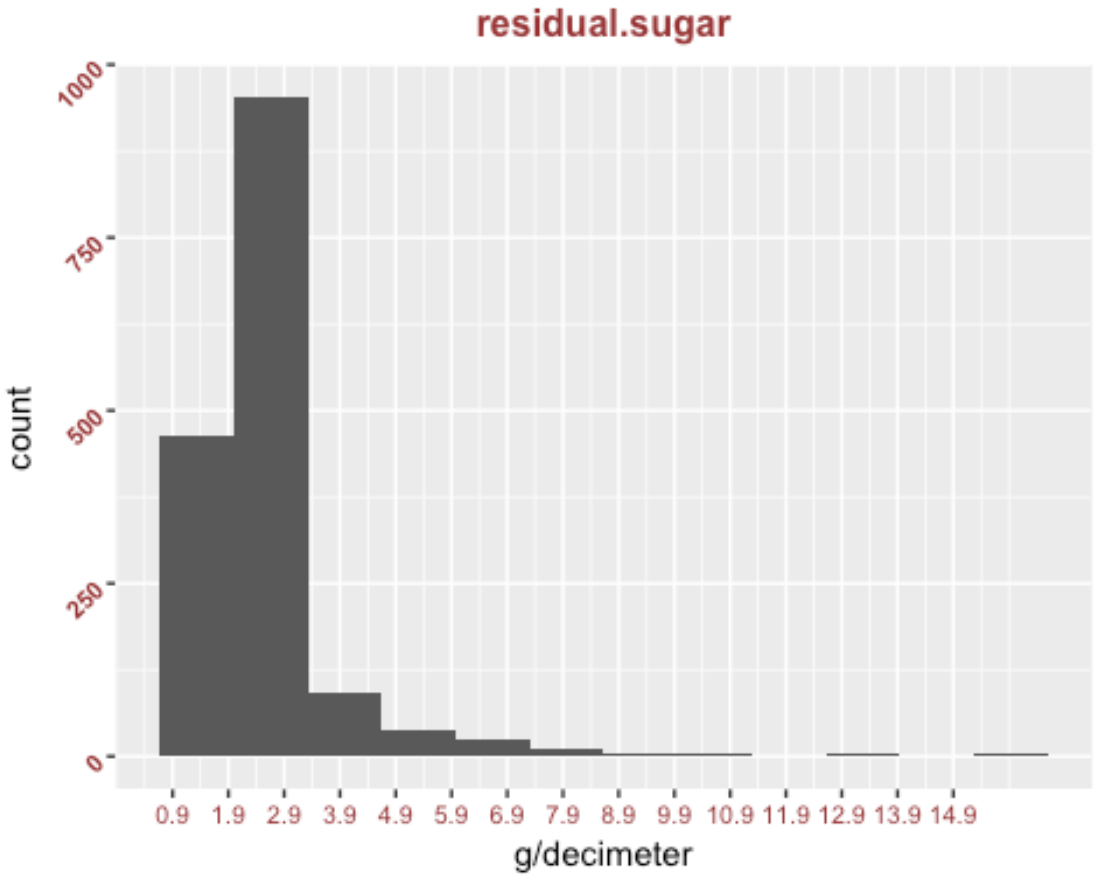
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.40    9.50   10.20   10.42   11.10   14.90 

## range

## [1] 6.5

## IQR

## [1] 1.6

## Density:

## summary

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040 

## range

## [1] 0.01362

## IQR

## [1] 0.002235

## Head: residual.sugar

##  [1] 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2.0 6.1 1.8 6.1 1.6 1.6 3.8 3.9
1.8
## [18] 1.7 4.4 1.8

## Head: alcohol

##  [1]  9.4  9.8  9.8  9.8  9.4  9.4  9.4 10.0  9.5 10.5  9.2 10.5  9.
9  9.1
## [15]  9.2  9.2 10.5  9.3  9.0  9.2

## Head: density

##  [1] 0.9978 0.9968 0.9970 0.9980 0.9978 0.9978 0.9964 0.9946 0.9968
0.9978
## [11] 0.9959 0.9978 0.9943 0.9974 0.9986 0.9986 0.9969 0.9968 0.9974
0.9969
```

Stats

The median of alcohol content is about 10%, which comports with my inutition about wine. Density, not surprisingly, is close to 1 (and generally below), as wine is mostly water.

Plot for Residual Sugar

Fairly normal, with a few highly right-skewed values.

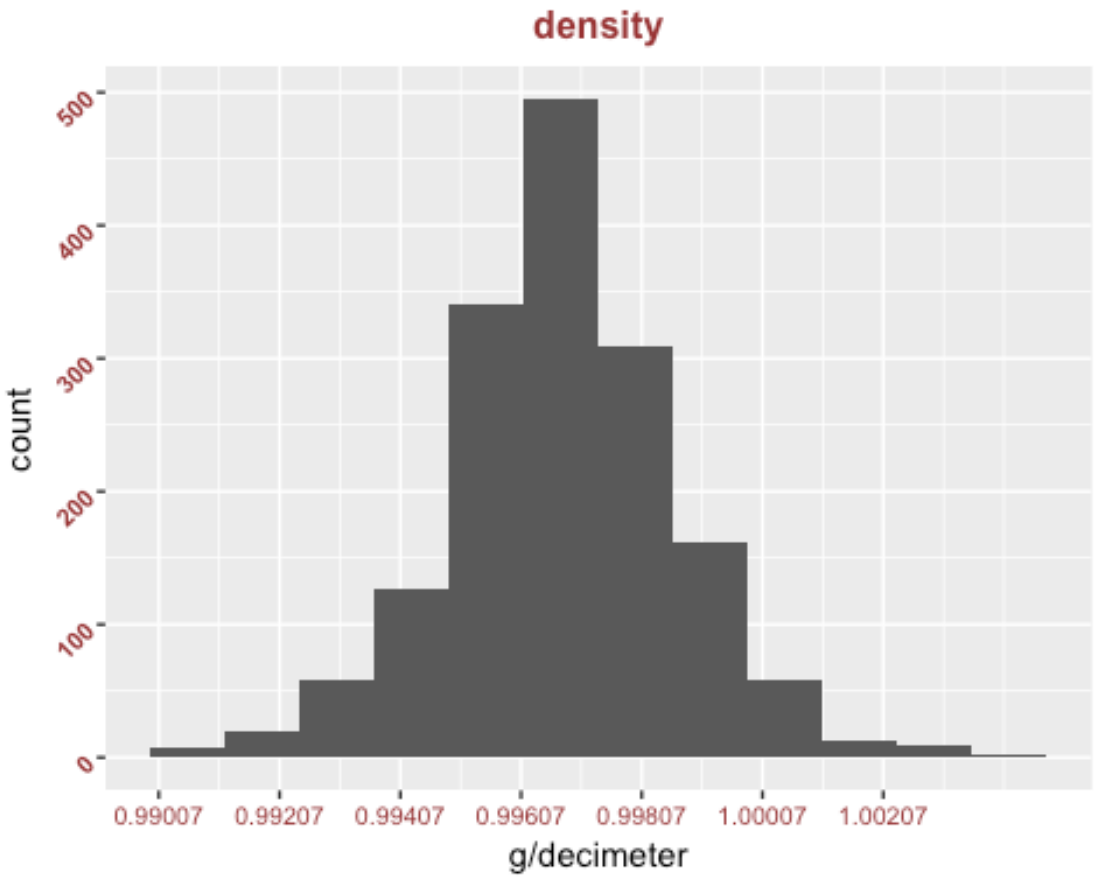


Plot for Alcohol

Right-skewed

Plot for density:

Fairly normal



chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH

Remaining Variables chlorides (sodium chloride - g / dm^3) free.sulfur.dioxide (mg / dm^3) total.sulfur.dioxide (mg / dm^3) pH (0-14)

Summary, Range, IQR, Head

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100

## [1] 0.599

## [1] 0.02

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00

## [1] 71

## [1] 14

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00

## [1] 283
```

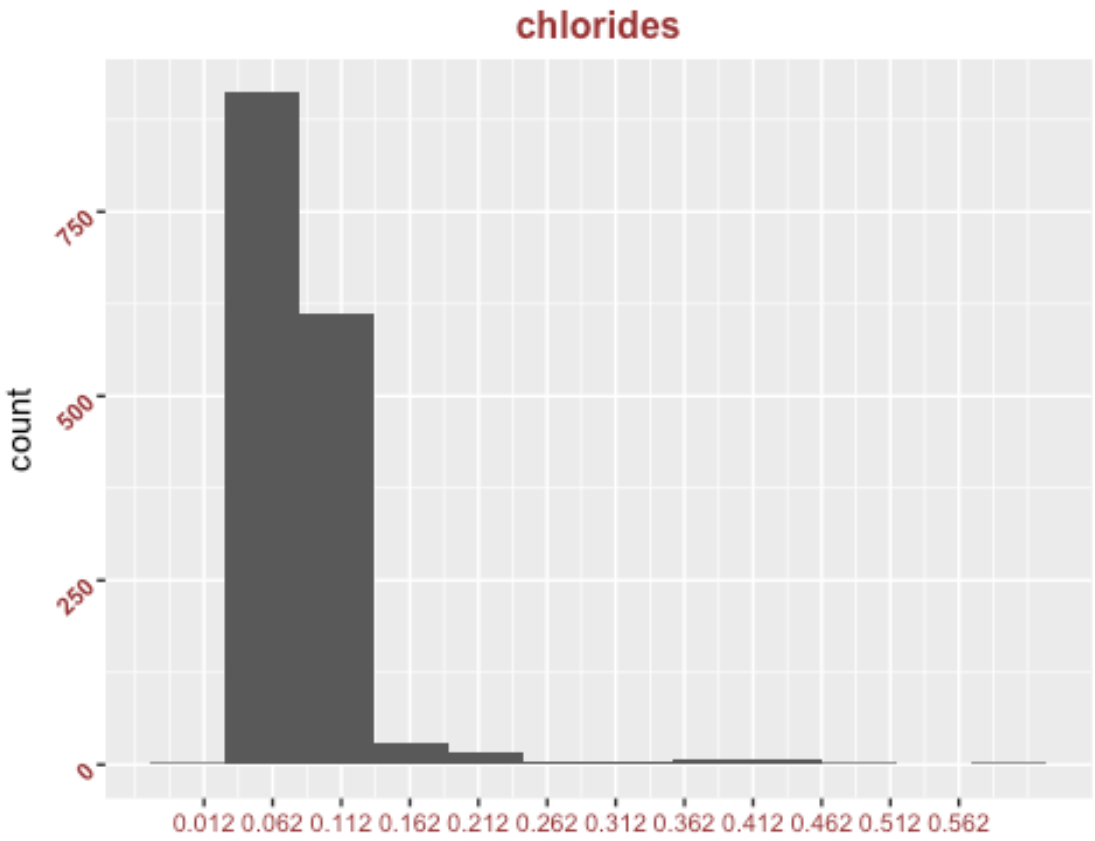
```
## [1] 40

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010

## [1] 1.27

## [1] 0.19

## Head: chlorides

##  [1] 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 0.0
97
## [12] 0.071 0.089 0.114 0.176 0.170 0.092 0.368 0.086 0.341

## Head: total.sulfur.dioxide

##  [1]  34  67  54  60  34  40  59  21  18 102  65 102  59  29 145 148
103
## [18]  56  29  56

## Head: free.sulfur.dioxide

##  [1] 11 25 15 17 11 13 15 15  9 17 15 17 16  9 52 51 35 16  6 17

## Head: pH

##  [1] 3.51 3.20 3.26 3.16 3.51 3.51 3.30 3.39 3.36 3.35 3.28 3.35 3.5
8 3.26
## [15] 3.16 3.17 3.30 3.11 3.38 3.04
```

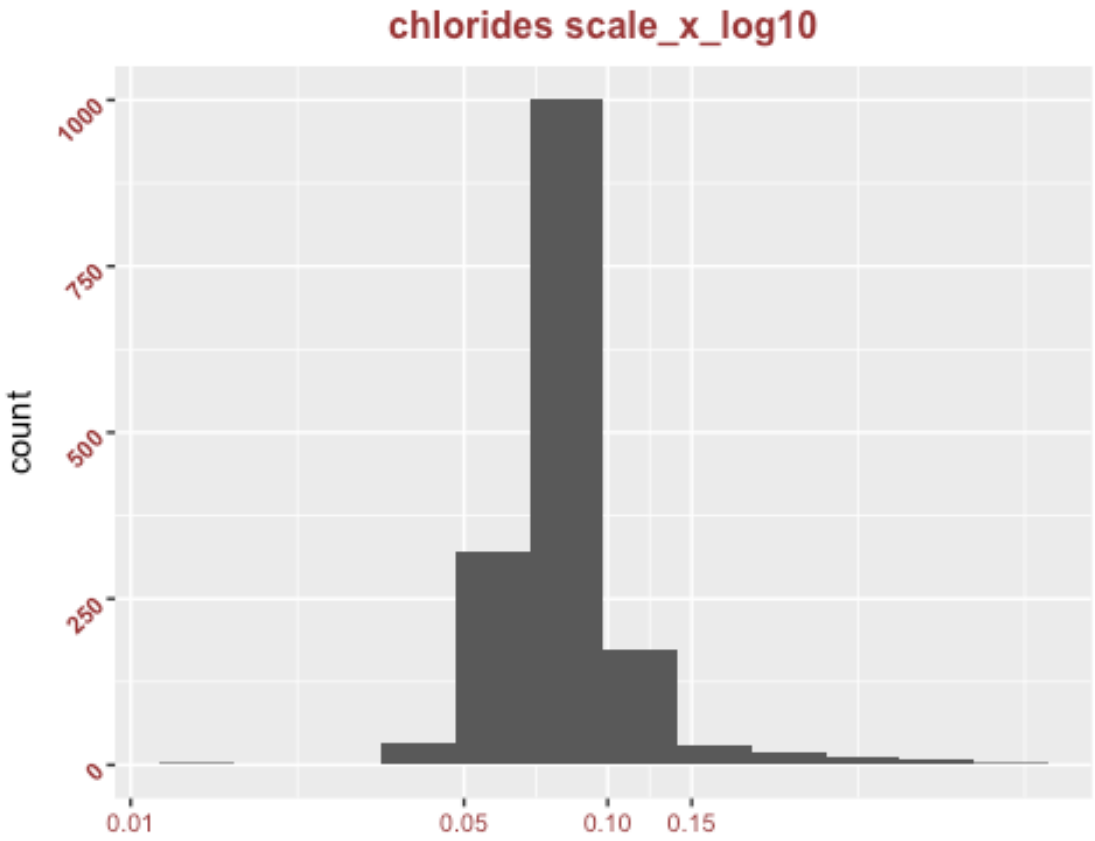
Stats

Don't know anything much about these features, so hard to make observations.

Plot for Chlorides

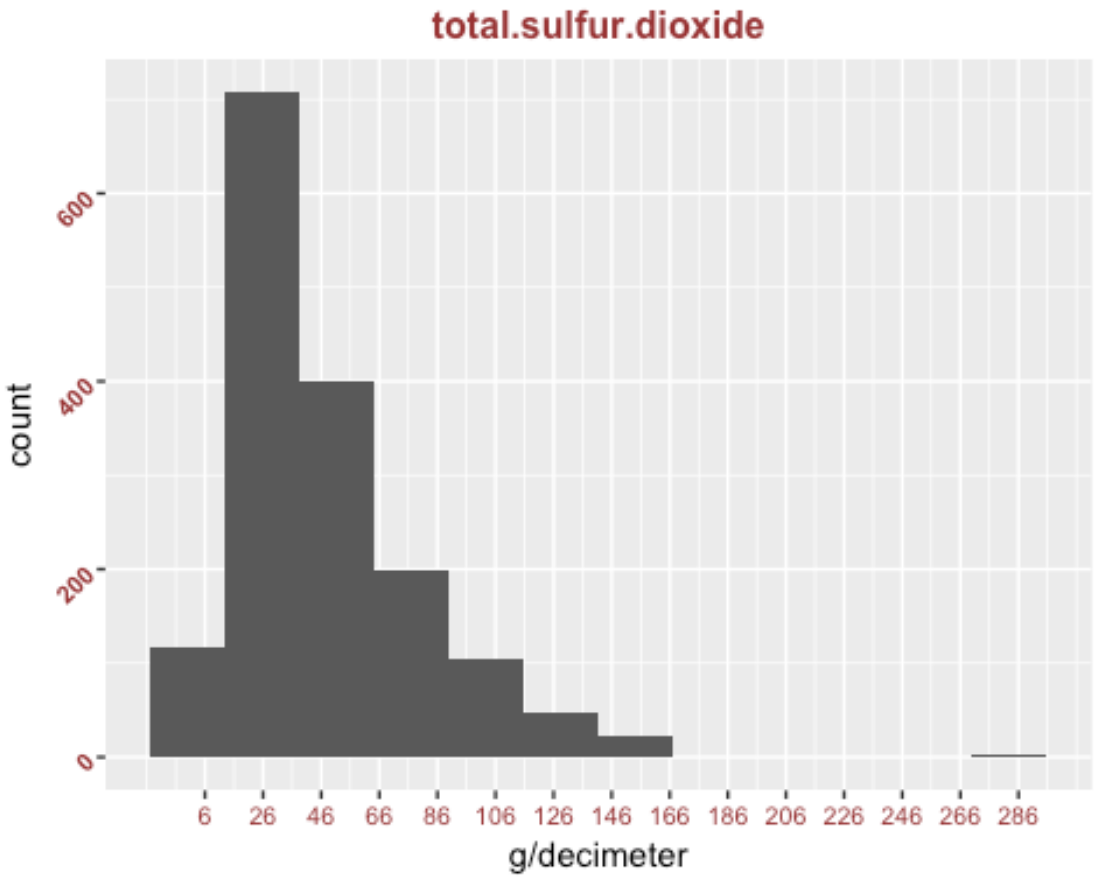
The linear plot is right-skewed so we add a log_10 transformation.

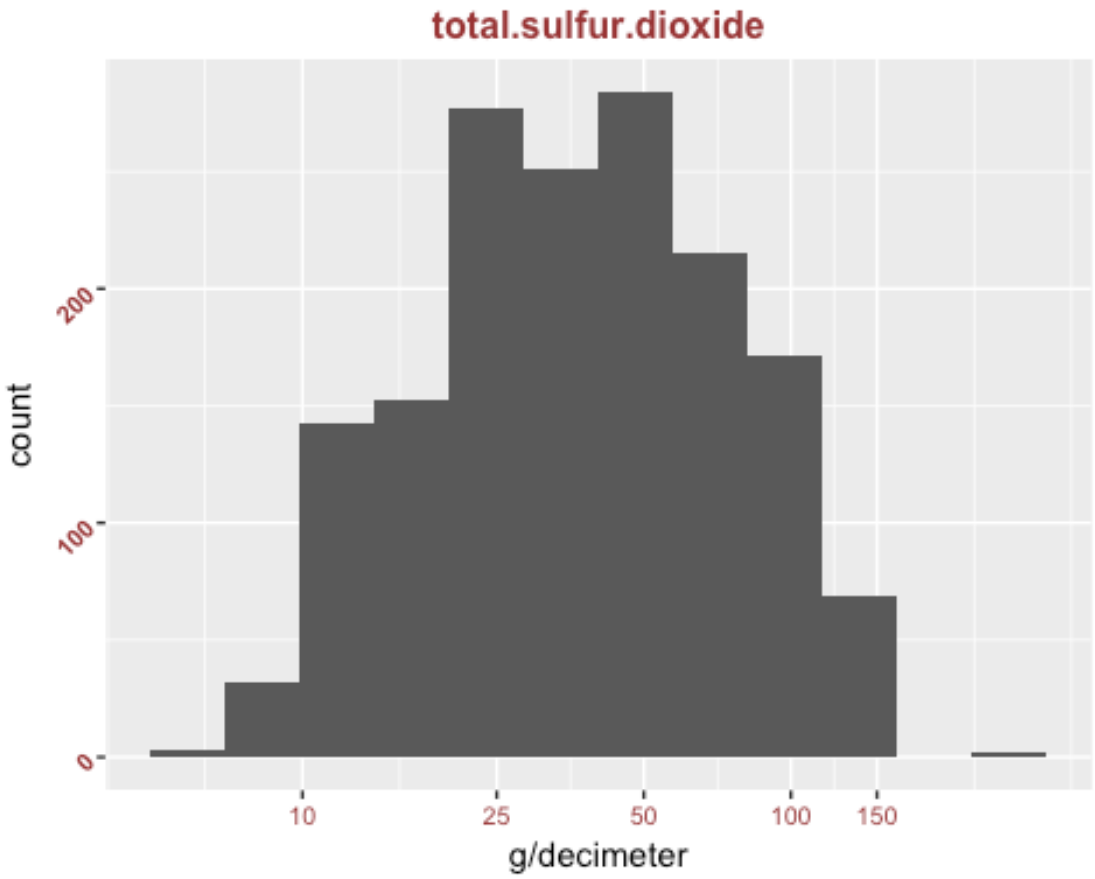

chlorides
count
750
500
250
0
0.012 0.062 0.112 0.162 0.212 0.262 0.312 0.362 0.412 0.462 0.512 0.562



Plot for total.sulfur.dioxide

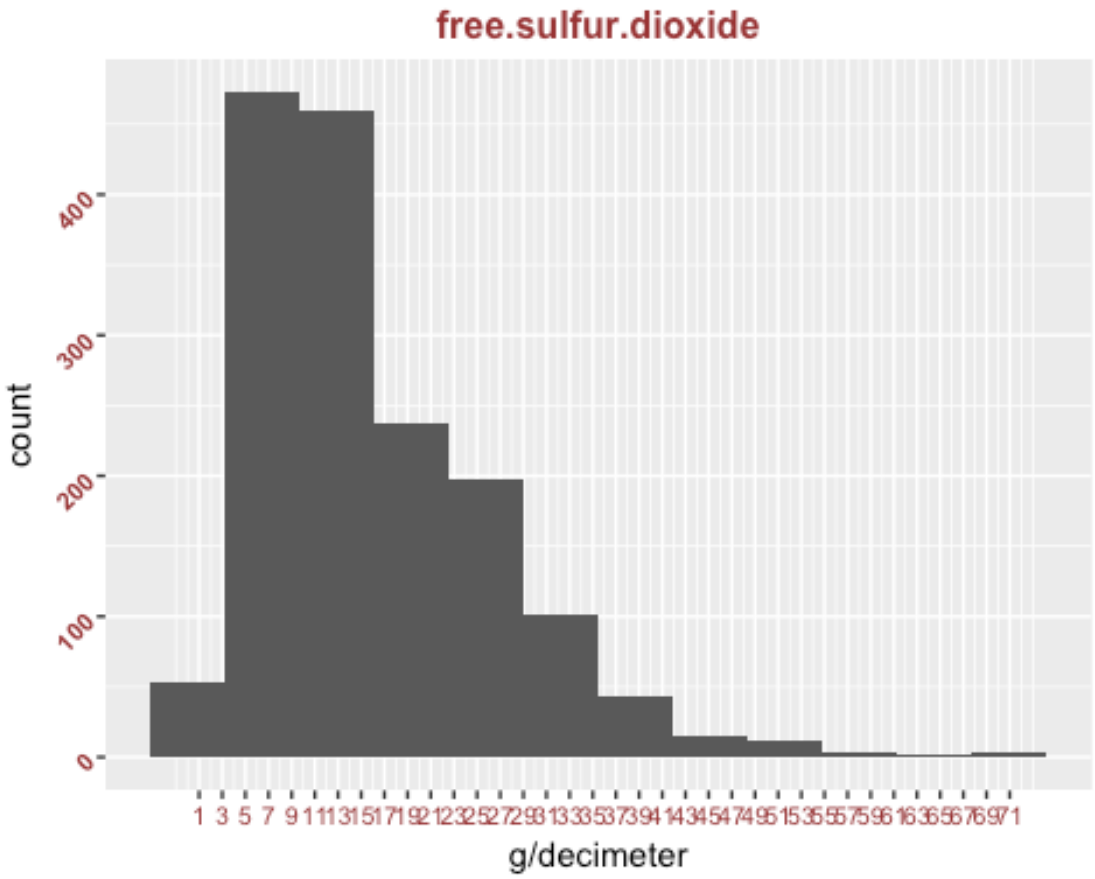
The linear plot is right-skewed so we add a log_10 transformation.

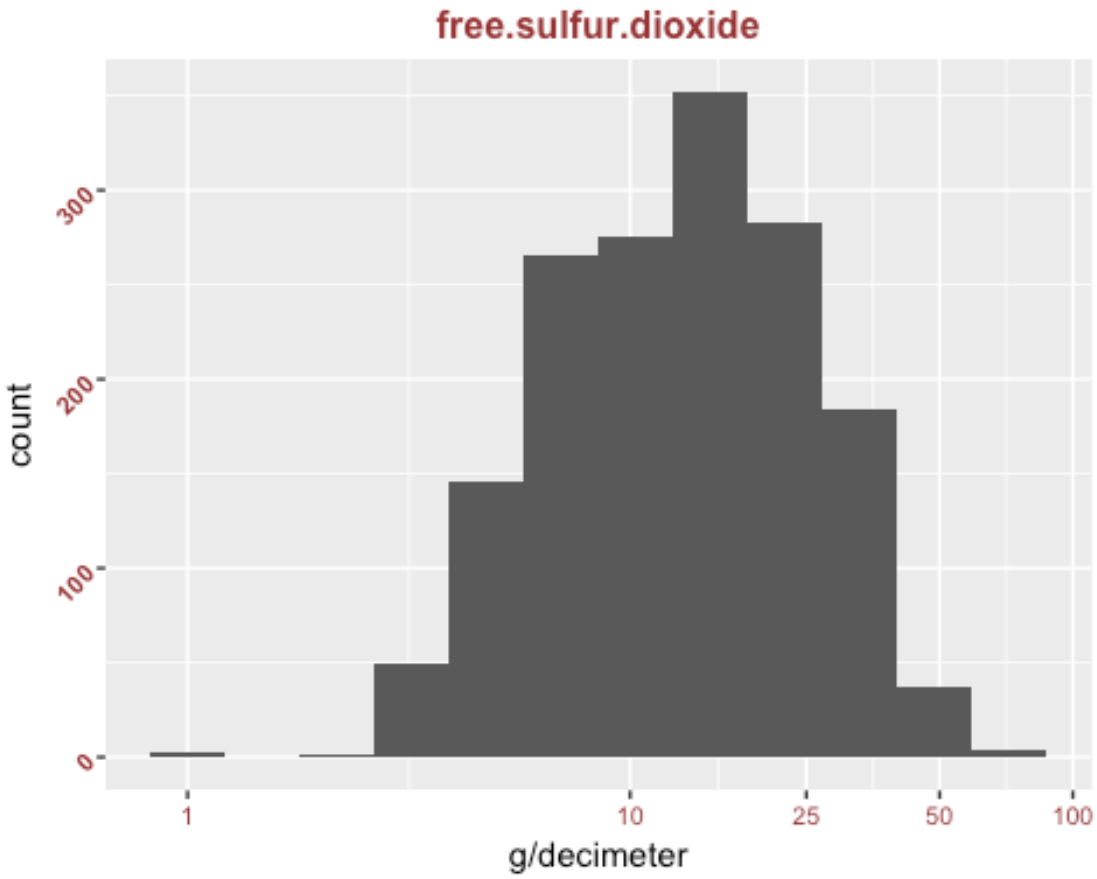

total.sulfur.dioxide
count
600
400
200
0
6
26
46
66
86
106
126
146
166
186
206
226
246
266
286
g/decimeter



Plot for free.sulfur.dioxide

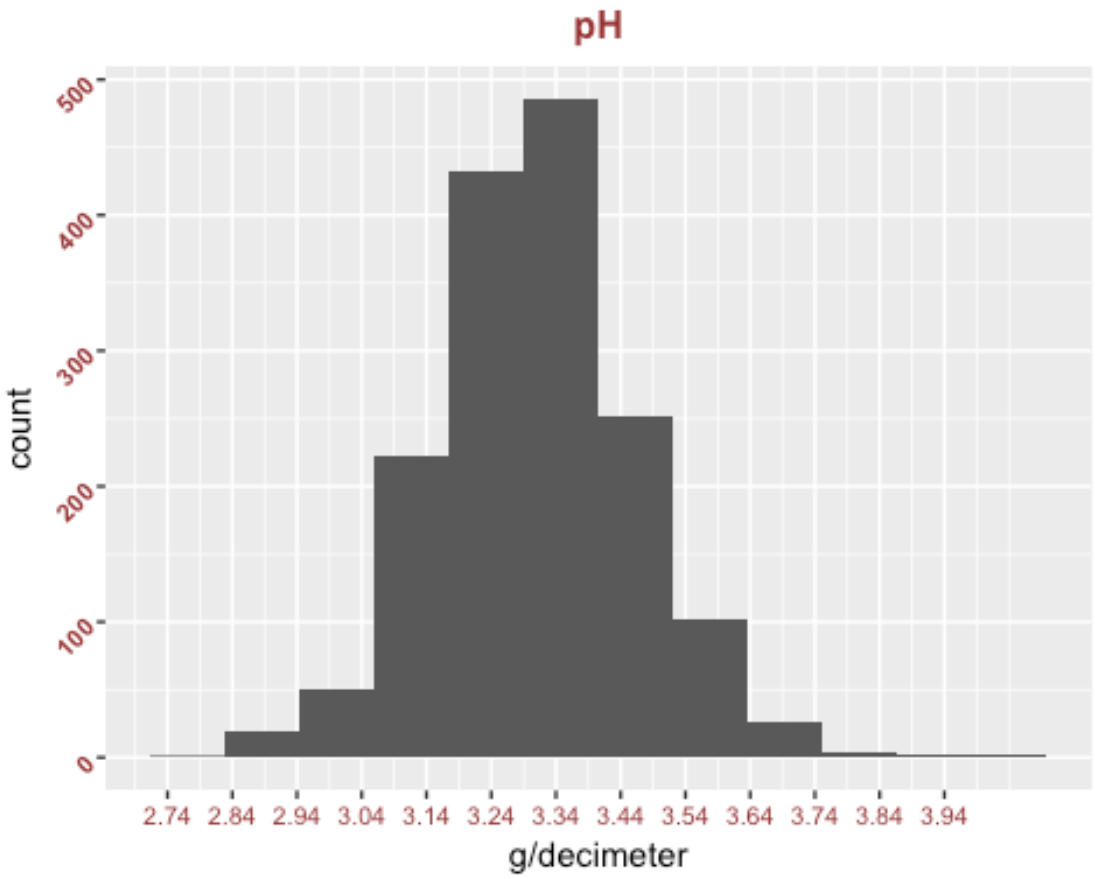
The linear plot is right-skewed so we add a log_10 transformation.


free.sulfur.dioxide
count
400
300
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71
g/decimeter



Plot for pH

Looks normal (note that pH is already log scale)



Univariate Analysis

Structure of dataset

1599 rows of values, each corresponding to a wine. The variables (columns) are:

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.diox
ide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
## [13] "totalacid"
```

Variables

I took a quick look at: "fixed.acidity"
"volatile.acidity" "citric.acid"
"residual.sugar" "density" "pH" "alcohol"

I created the variable "totalacid" which is the sum of the 3 given acid values, and is close to "fixed.acidity" since the other two have very small values.

And of course we looked at "quality", which is not a chemical measurement, like the others, but presumably reporting of quality of the wine by human judges.

Unless you know a bit about the chemistry of wines, this univariate analysis is unlikely to be that informative.

Since all the variables besides quality are chemical quantities, it's hard to have an intuition on how to think of them. But hopefully with this univariate analysis this will lead me to some interesting bivariate and mulitvariate analysis.

I created the new variable $totalacid by summing fixed.acidity + volatile.acidity + citric.acid, rounded to 0.1 I simply adjusted the scale_x_continuous limits and breaks to show all the data with appropriate labels on the graphs.

Bivariate Plots Section

I have a little knowledge of biochemistry, so we used this knowledge to try some bivariate analysis. Yeast act by fermenting sugar to alcohol, so you think these might be inversely correlated, although the amount of starting sugar would also highly effect each. Also, since the density of alcohol is about 0.79, wine density might be inversely correlated with alocohol content. Sugar might increase density a bit, as dissolved solids generally increase the density of water.

We can get a quick overview of all of the pairwise correlations with a correlations plot. Looking at quality, which is the feature that most people are most interested in (and is not an objective chemical variable), we see that nothing has a high correlation (near 1.0). This is not surprising to me, because we would assume that in a complex art like wine-making, there wouldn't be any single chemical feature that would be highly correlated with quality. We will discuss this more below.

Pairwise Correlations

Let's look at some of the pairwise correlations in detail.

1. sugar v alcohol
2. alcohol v density
3. sugar v density

Pearsons r: 1. 0.04207544 2. -0.4961798 3. 0.3552834

r is the first is almost zero, so I'm guessing initial sugar content varies a lot. The second has a negative correlation as we suspected, and it indeed is fairly strong. The third has a modest correlation.

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$residual.sugar and redwine$alcohol
## t = 1.6829, df = 1597, p-value = 0.09258
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006960058  0.090909069
## sample estimates:
##       cor
## 0.04207544

##
##  Pearson's product-moment correlation
##
## data:  redwine$alcohol and redwine$density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798

##
##  Pearson's product-moment correlation
##
## data:  redwine$residual.sugar and redwine$density
## t = 15.189, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3116908 0.3973835
## sample estimates:
##       cor
## 0.3552834
```

Acids and pH

Since pH is a measure of acidity, the acid content of the wine should show a strong correlation, at least for fixed acidity (which is most of the acid in the wine) and total acid. Note that lower pH means more acidic.

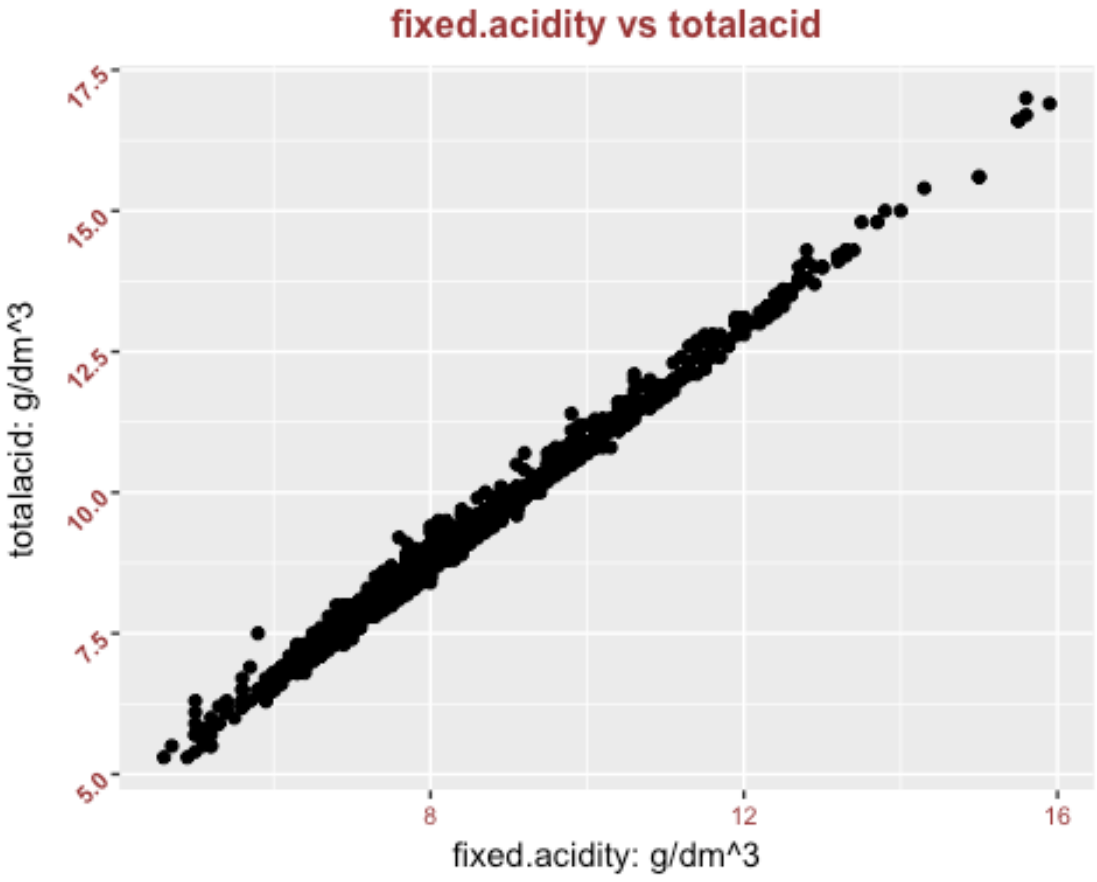
1. fixed acidity vs pH
2. volatile acidity vs pH
3. citric acid vs pH
4. total acid vs pH

Pearsons r: 1. -0.6829782 2. 0.2349373 3. -0.5419041 4. -0.6833998

And indeed 1 and the 4 are strongly negatively correlated with pH. It is a bit surprising that the third is also fairly strongly negatively correlated.

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$fixed.acidity and redwine$pH
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##       cor
## -0.6829782

##
##  Pearson's product-moment correlation
##
## data:  redwine$volatile.acidity and redwine$pH
## t = 9.659, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1880823 0.2807254
## sample estimates:
##       cor
## 0.2349373

##
##  Pearson's product-moment correlation
##
## data:  redwine$citric.acid and redwine$pH
## t = -25.767, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5756337 -0.5063336
## sample estimates:
##       cor
## -0.5419041

##
##  Pearson's product-moment correlation
##
## data:  redwine$totalacid and redwine$pH
## t = -37.409, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7086794 -0.6563676
## sample estimates:
##       cor
## -0.6833998
```

total.acid should be highly correlated to fixed.acidity, since the first is mostly composed of the second. Indeed r is 0.996198, and a scatterplot visually confirms this. There aren't many, or very strong, outliers from the line.



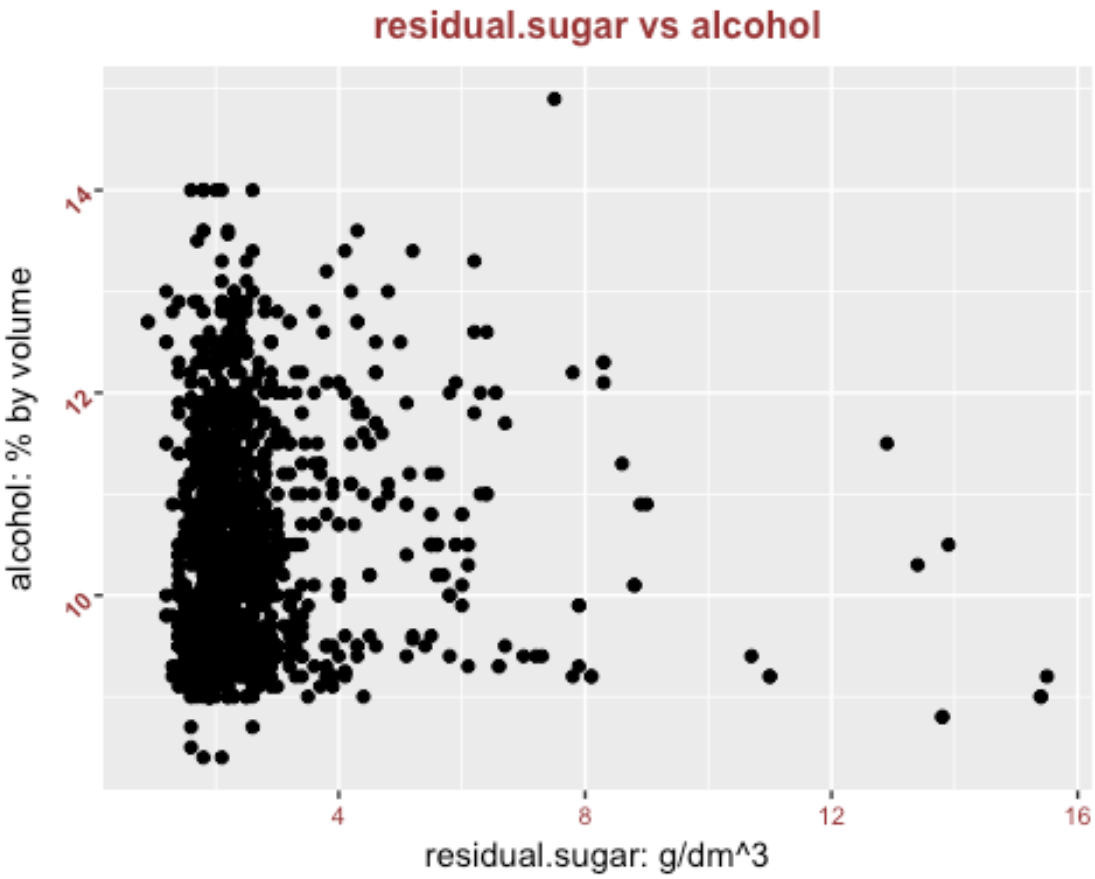

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$fixed.acidity and redwine$totalacid
## t = 456.98, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9958069 0.9965528
## sample estimates:
##      cor
## 0.996198
```

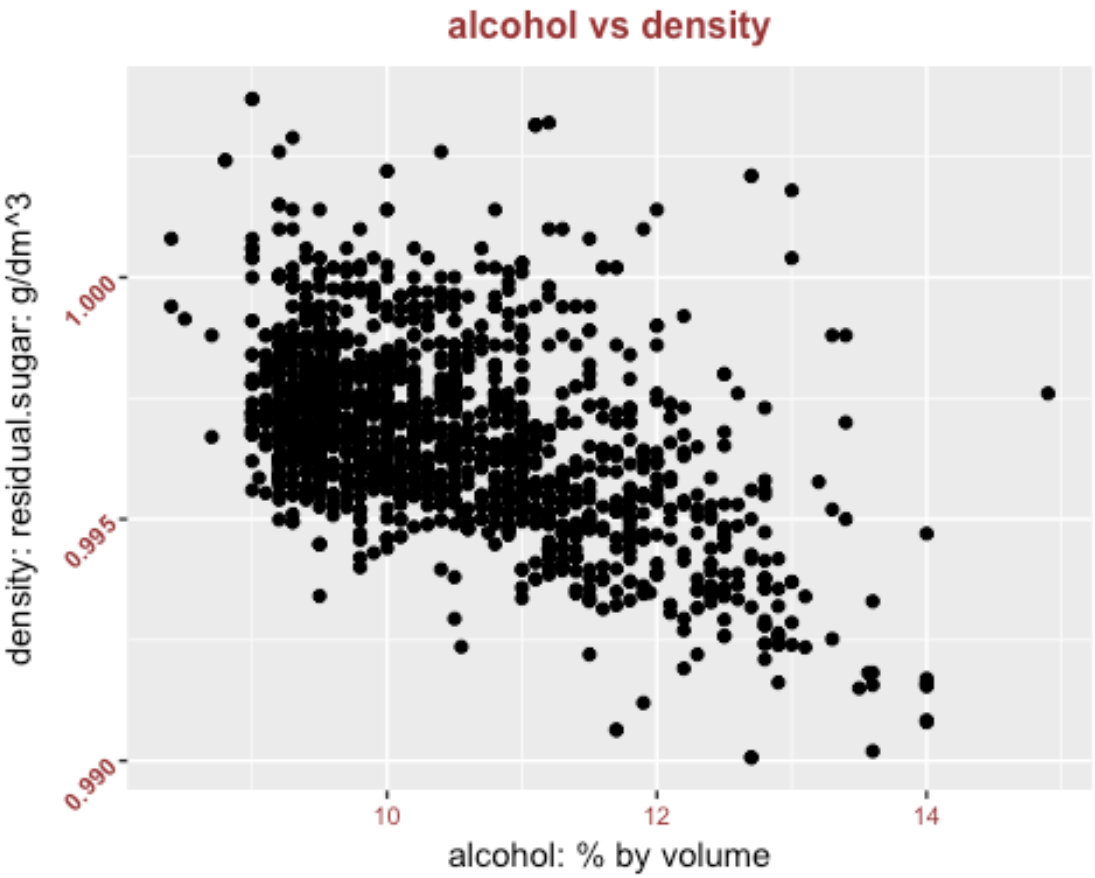
R and scatterplots

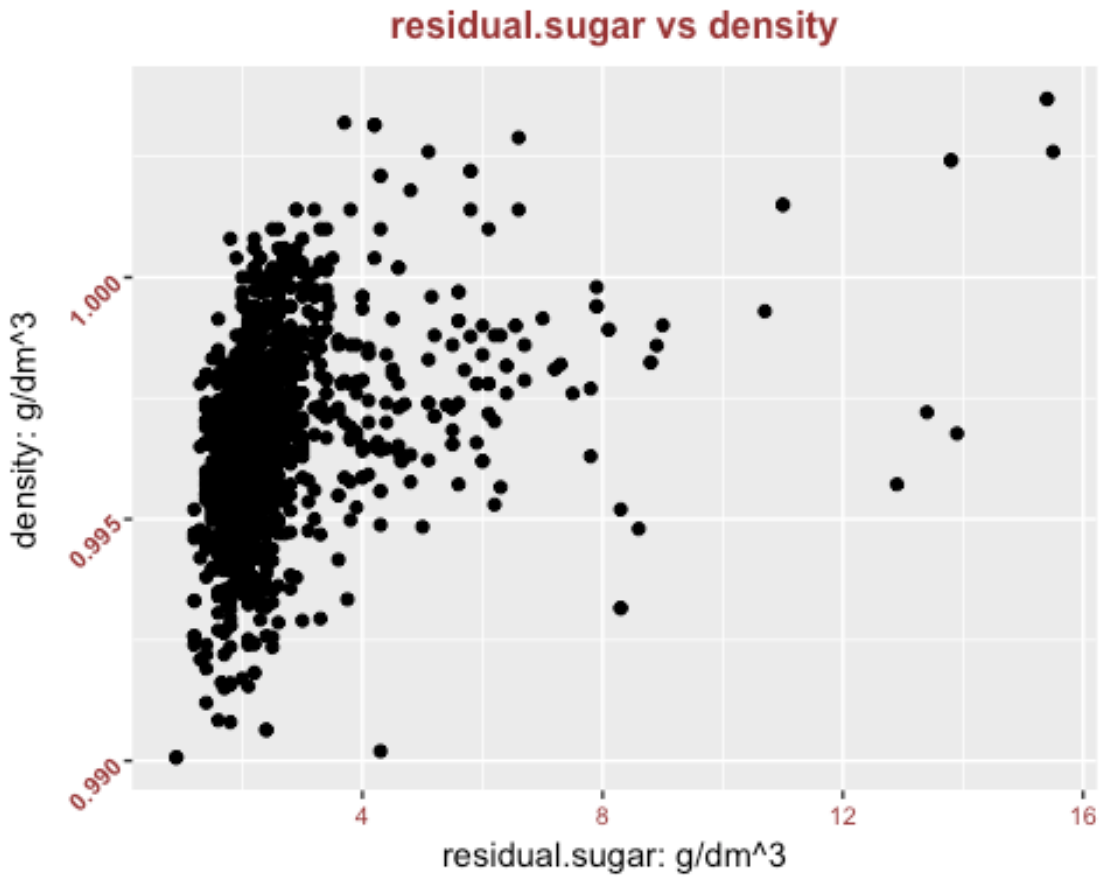
Now let's scatterplot the following variable pairs. 1. sugar v alcohol 2. alcohol v density 3. sugar v density

Pearsons r: 1. 0.04207544 2. -0.4961798 3. 0.3552834

For 1, correlation is very low, and the plot indeed looks random. For 2, we see the fairly strong negative correlation. For 3, it is hard to see the correlation, which is larger than 1, but still modest.


residual.sugar vs alcohol
alcohol: % by volume
14
12
10
4
8
12
16
residual.sugar: g/dm^3


alcohol vs density
density: residual.sugar: g/dm^3
1.000
0.995
0.990
10
12
14
alcohol: % by volume

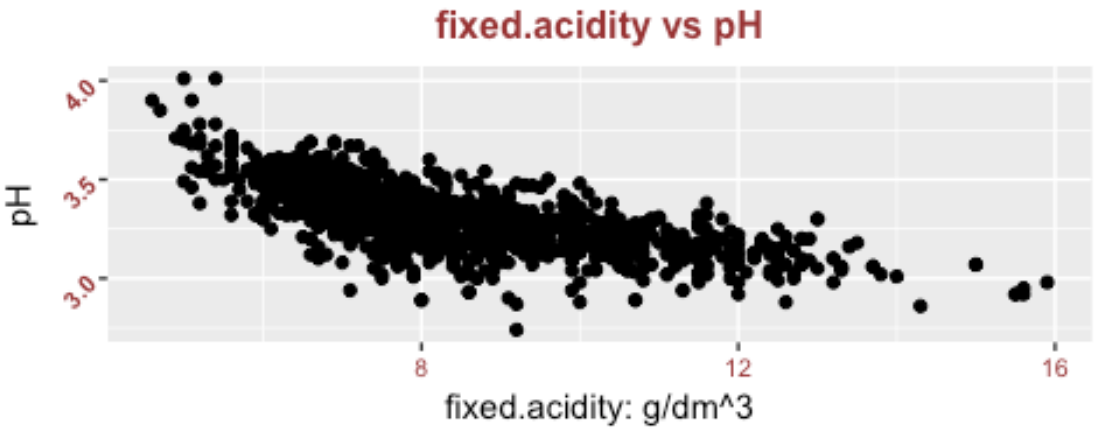


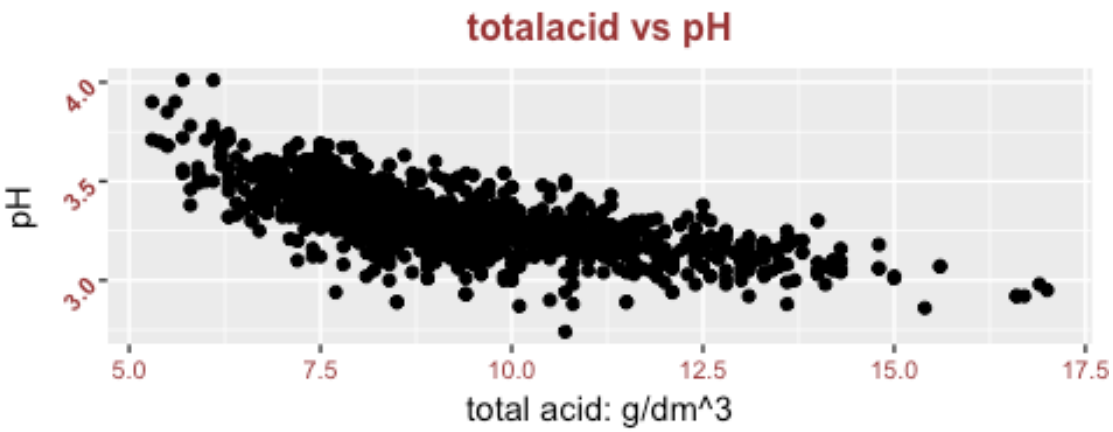
R and scatterplots

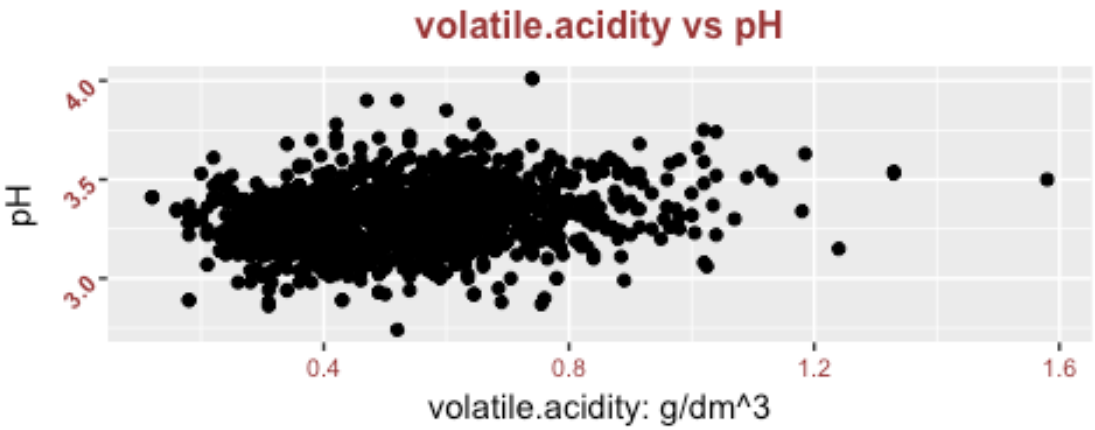
1. fixed acidity vs pH
2. volatile acidity vs pH
3. citric acid vs pH
4. total acid vs pH

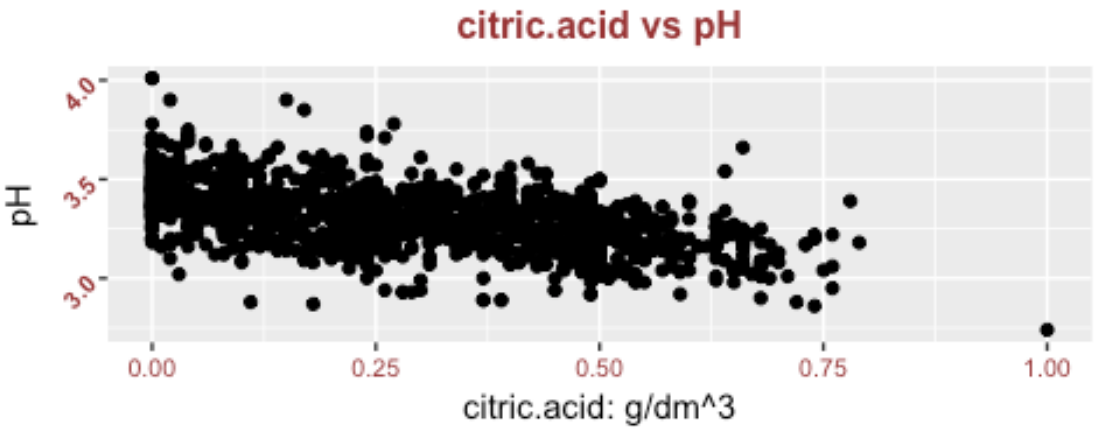
Pearsons r: 1. -0.6829782 2. 0.2349373 3. -0.5419041 4. -0.6833998

I will plot 1 and 4 together. They should look similar. Next we plot 2 and 3. 2 should look random, 3 should have a modest discernible negative slope.


fixed.acidity vs pH
pH
4.0
3.5
3.0
8
12
16
fixed.acidity: g/dm^3


totalacid vs pH
pH
4.0
3.5
3.0
5.0
7.5
10.0
12.5
15.0
17.5
total acid: g/dm^3





Bivariate Plots Section

Each variable v quality

Now we compare quality to every other variable (columns 2-12) to quality. we also compare quality to my new variable totalacid (column 14,totalacid = 2 + 3 + 4, rounded to 0.1).

I would be surprised if any of them have a very high correlation with quality. Wine making is a complex art, and if this were true, you might be able to easily improve the quality of the wine by addressing this one variable.

[2] "fixed.acidity"
[3] "volatile.acidity" [4] "citric.acid"
[5] "residual.sugar" [6] "chlorides"
[7] "free.sulfur.dioxide" [8] "total.sulfur.dioxide" [9] "density" [10] "pH"
[11] "sulphates" [12] "alcohol"
[14] "totalacid"

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.diox
```

```
ide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
## [13] "totalacid"
```

R vs each variable

2. 0.1240516
3. -0.3905578
4. 0.2263725
5. 0.01373164
6. -0.1289066
7. -0.05065606
8. -0.1851003
9. -0.1749192
10. -0.05773139
11. 0.2513971
12. 0.4761663
13. 0.1036382

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$fixed.acidity and redwine$quality
## t = 4.996, df = 1597, p-value = 6.496e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07548957 0.17202667
## sample estimates:
##       cor
## 0.1240516

##
##  Pearson's product-moment correlation
##
## data:  redwine$volatile.acidity and redwine$quality
## t = -16.954, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##        cor
## -0.3905578

##
##  Pearson's product-moment correlation
##
## data:  redwine$citric.acid and redwine$quality
## t = 9.2875, df = 1597, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
## sample estimates:
##       cor
## 0.2263725

##
##  Pearson's product-moment correlation
##
## data:  redwine$residual.sugar and redwine$quality
## t = 0.5488, df = 1597, p-value = 0.5832
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03531327  0.06271056
## sample estimates:
##        cor
## 0.01373164

##
##  Pearson's product-moment correlation
##
## data:  redwine$chlorides and redwine$quality
## t = -5.1948, df = 1597, p-value = 2.313e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17681041 -0.08039344
## sample estimates:
##        cor
## -0.1289066

##
##  Pearson's product-moment correlation
##
## data:  redwine$free.sulfur.dioxide and redwine$quality
## t = -2.0269, df = 1597, p-value = 0.04283
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.099430290 -0.001638987
## sample estimates:
##         cor
## -0.05065606

##
##  Pearson's product-moment correlation
##
## data:  redwine$total.sulfur.dioxide and redwine$quality
## t = -7.5271, df = 1597, p-value = 8.622e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2320162 -0.1373252
```

```
## sample estimates:
##        cor
## -0.1851003

##
##  Pearson's product-moment correlation
##
## data:  redwine$density and redwine$quality
## t = -7.0997, df = 1597, p-value = 1.875e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2220365 -0.1269870
## sample estimates:
##        cor
## -0.1749192

##
##  Pearson's product-moment correlation
##
## data:  redwine$pH and redwine$quality
## t = -2.3109, df = 1597, p-value = 0.02096
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.106451268 -0.008734972
## sample estimates:
##         cor
## -0.05773139

##
##  Pearson's product-moment correlation
##
## data:  redwine$sulphates and redwine$quality
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971

##
##  Pearson's product-moment correlation
##
## data:  redwine$alcohol and redwine$quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$totalacid and redwine$quality
## t = 4.1641, df = 1597, p-value = 3.293e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05489594 0.15188766
## sample estimates:
##       cor
## 0.1036382
```

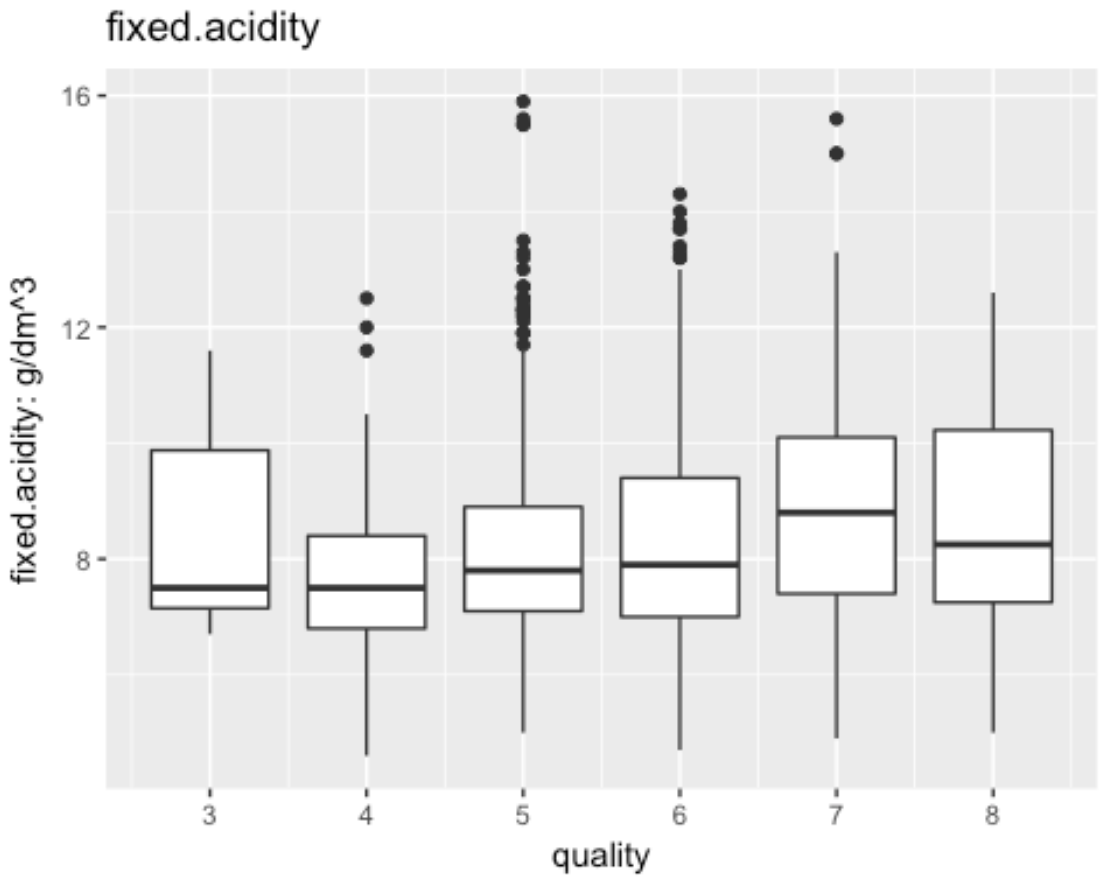
Highest correlations

3, volatile acidity, shows the highest negative corr with quality (-0.39) 12, alcohol, has the highest positive correlation with quality (0.47). But these correlations are still pretty modest.

So we decide we will create box plots for each of these variable, with the different wine ratings along the X axis. We might notice a different pattern with each one of them as between different wine ratings (though we am guessing not). We could, also calculate r within each of these wine ratings to see if there is a particularly high r value.

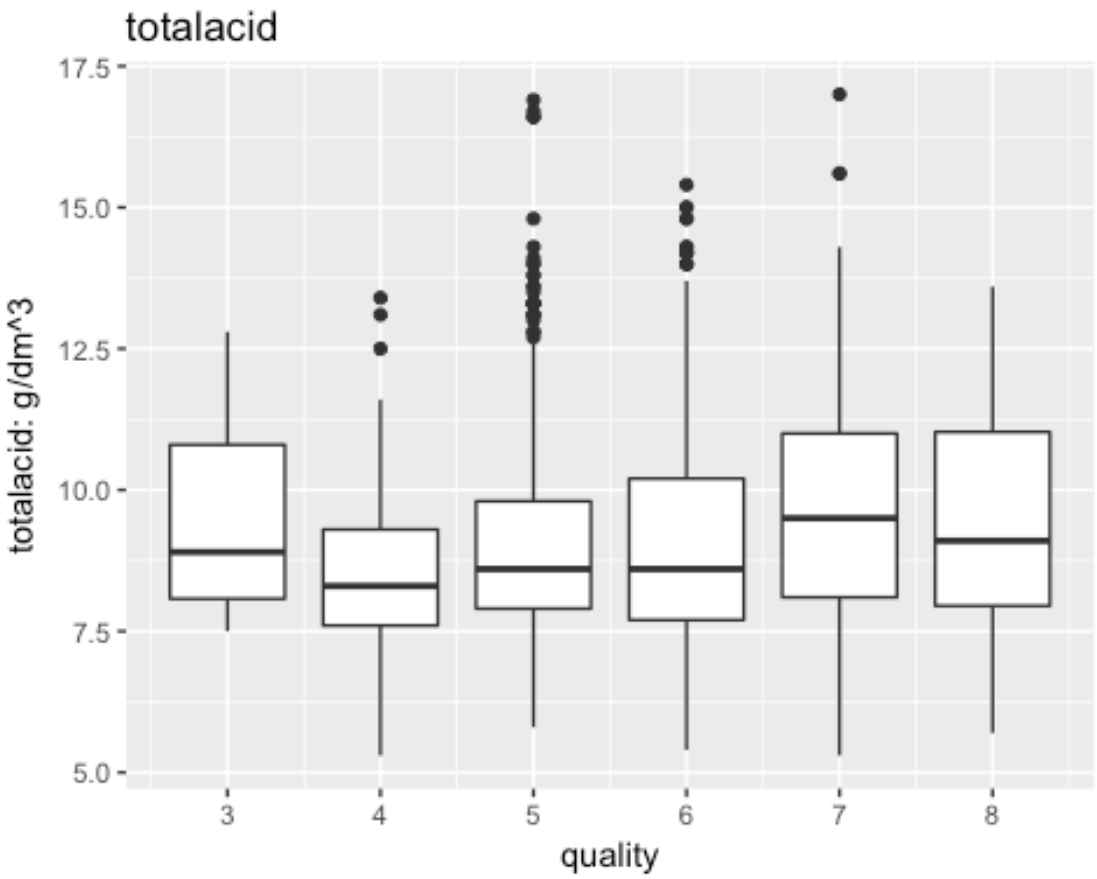
fixed.acidity

Both high (8) and low (3) wine rating have fewer outliers. we don't know if that means anything.



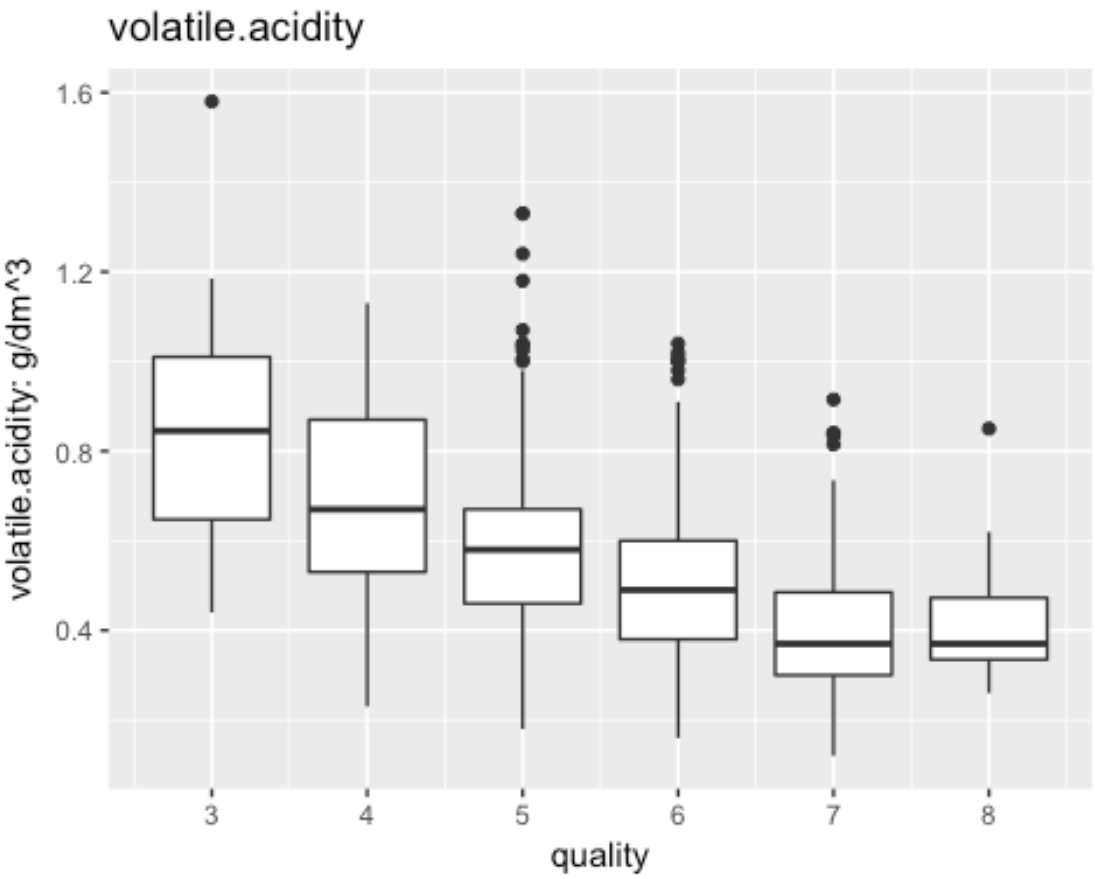
totalacid

Looks very much like fixed.acidity, not surprisingly



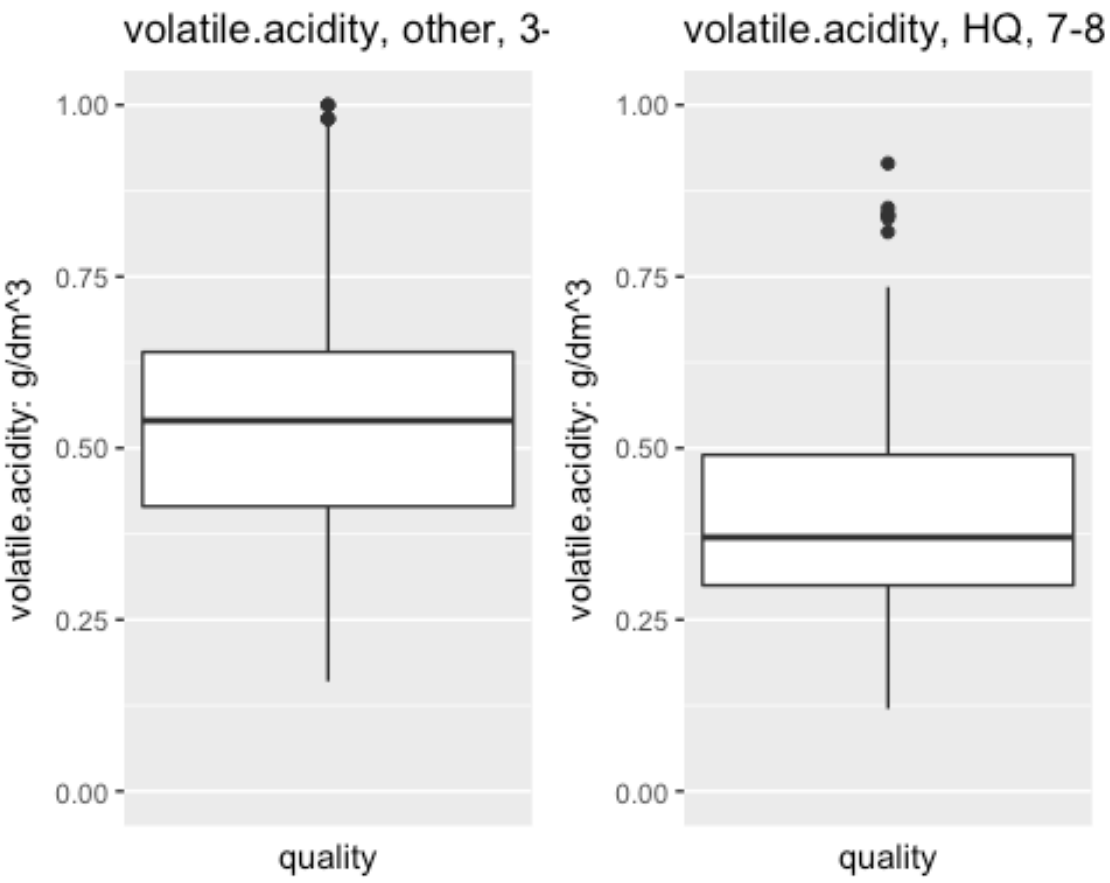
volatile.acidity

Aha!!! Volatile acidity goes seem to go down with higher quality wine.



Two box plots

Now we arbitrarily decide to bucket wines into high-quality wines (scores 7-8) and other (3-6) with box plots.



Segregated stats

Mean volatile.acidity 7-8 wines: 0.4055 3-6 wines: 0.547

I'm guessing we would find a statistically significant difference between the two groups with respect to volatile.acidity.

The r values are different too, though I'm not sure that means anything.

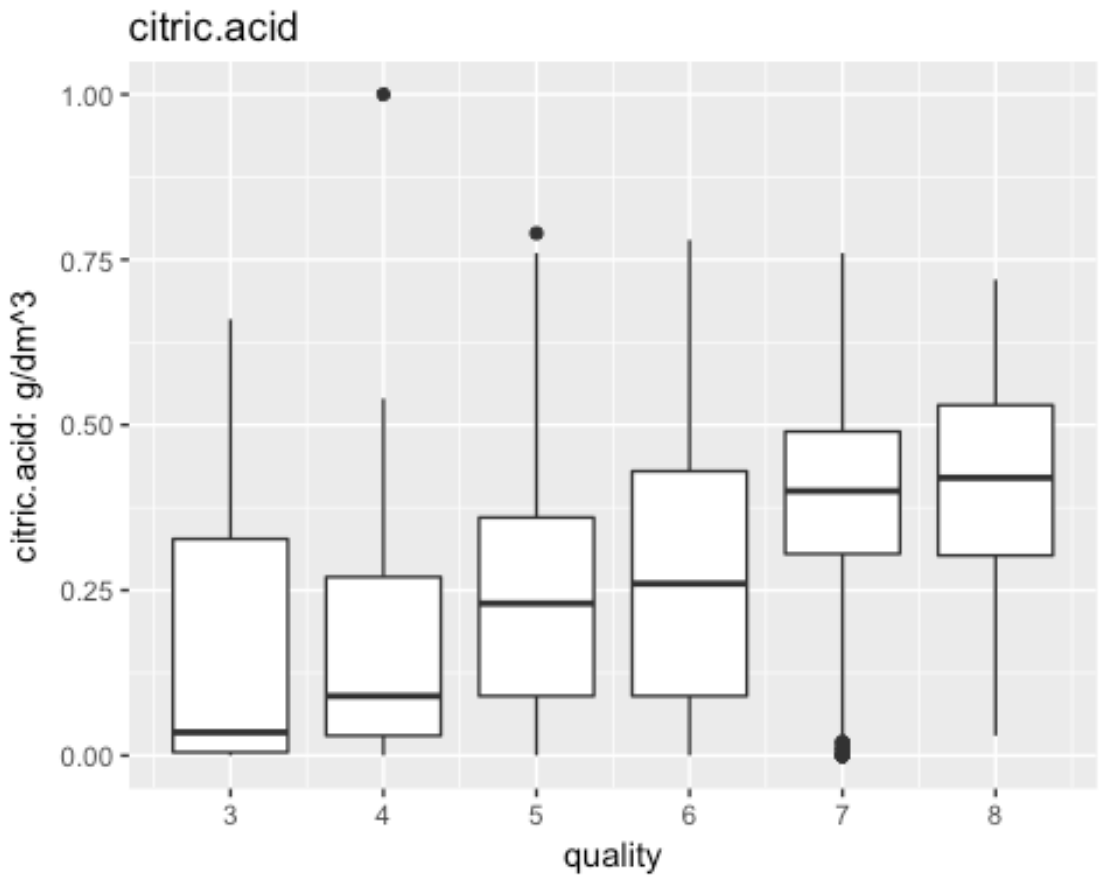
r 7-8 wines: -0.3148996 3-6 wines: 0.03702191

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.160   0.420   0.540   0.547   0.650   1.580

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4055  0.4900  0.9150

##
##  Pearson's product-moment correlation
##
## data:  subset(redwine, quality <= 6)$volatile.acidity and subset(red
```

```
wine, quality <= 6)$quality
## t = -12.325, df = 1380, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3616255 -0.2665958
## sample estimates:
##        cor
## -0.3148996

##
##  Pearson's product-moment correlation
##
## data:  subset(redwine, quality >= 7)$volatile.acidity and subset(red
wine, quality >= 7)$quality
## t = 0.54322, df = 215, p-value = 0.5875
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0966390  0.1693712
## sample estimates:
##        cor
## 0.03702191
```

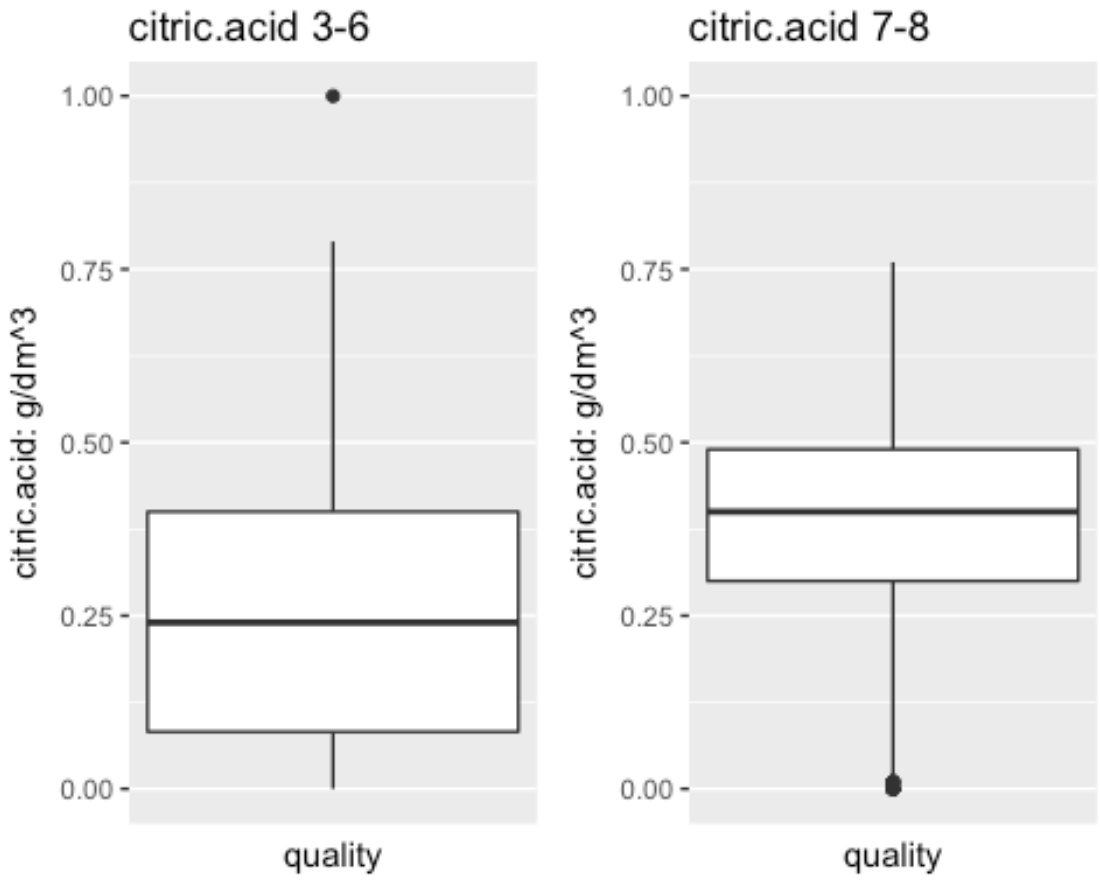
citric.acid

Not as noticeable as volatile.acid, but higher quality wines seem to have more of this fruity acid.



Two box plots

Now we arbitrarily decide to bucket wines into high-quality wines (scores 7-8) and other (3-6) with box plots.



Segreated statistics

Mean citric.acid 7-8 wines: 0.3765 3-6 wines: 0.2544

I'm guessing we would find a statistically significant difference between the two groups with respect to citric.acid.

The r values are different too, though I'm not sure that's useful. Perhaps it means that really good wines are more complex than just more or less acid.

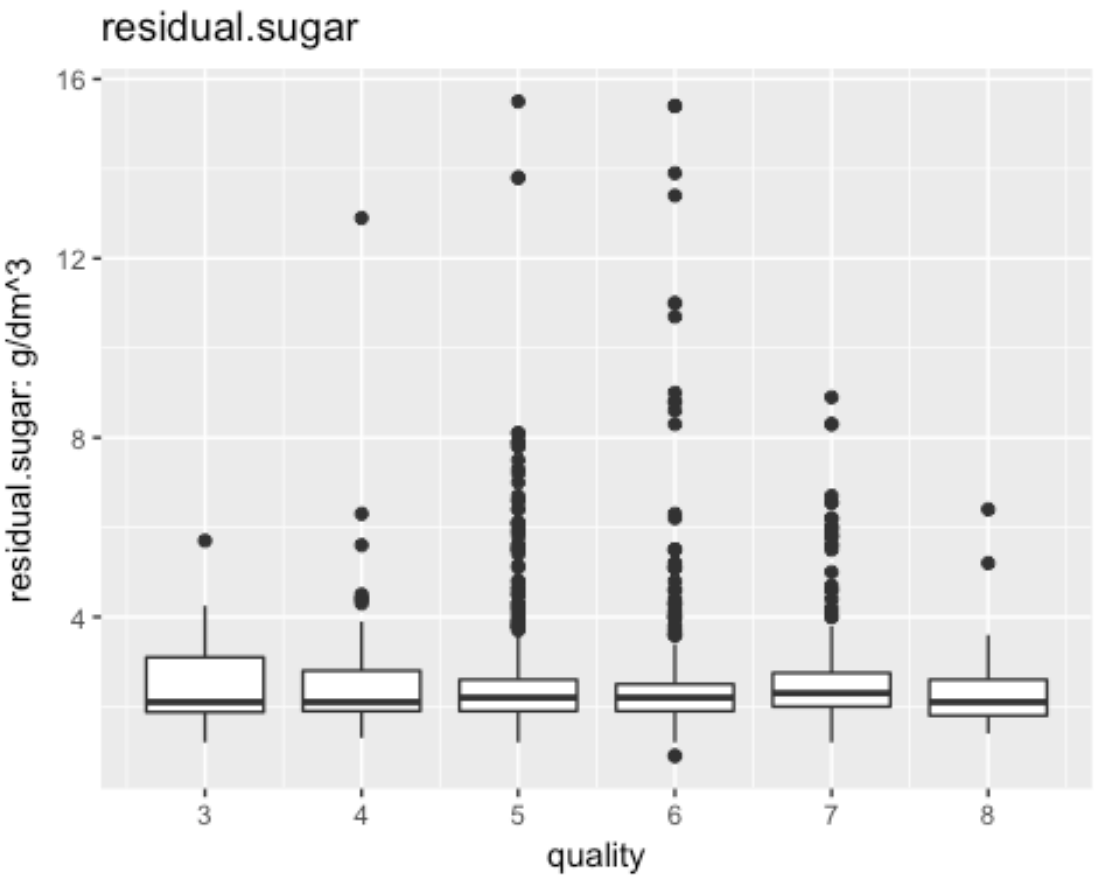
r 7-8 wines: 0.1161771 3-6 wines: 0.02265598

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0825  0.2400  0.2544  0.4000  1.0000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3000  0.4000  0.3765  0.4900  0.7600

##
##  Pearson's product-moment correlation
##
```

```
## data:  subset(redwine, quality <= 6)$citric.acid and subset(redwine,
quality <= 6)$quality
## t = 4.3452, df = 1380, p-value = 1.493e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06383753 0.16787927
## sample estimates:
##       cor
## 0.1161771

##
##  Pearson's product-moment correlation
##
## data:  subset(redwine, quality >= 7)$citric.acid and subset(redwine,
quality >= 7)$quality
## t = 0.33229, df = 215, p-value = 0.74
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1108629  0.1553716
## sample estimates:
##        cor
## 0.02265598
```

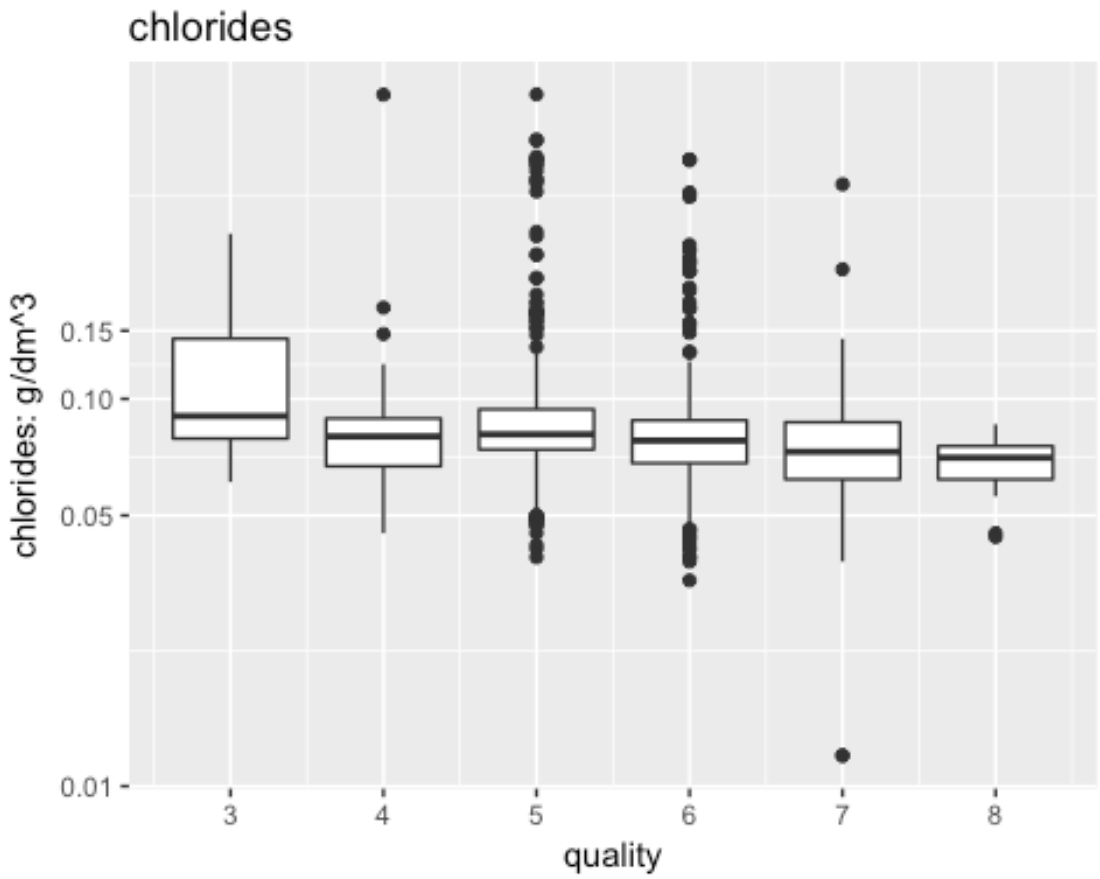
residual.sugar

Nothing jumps out



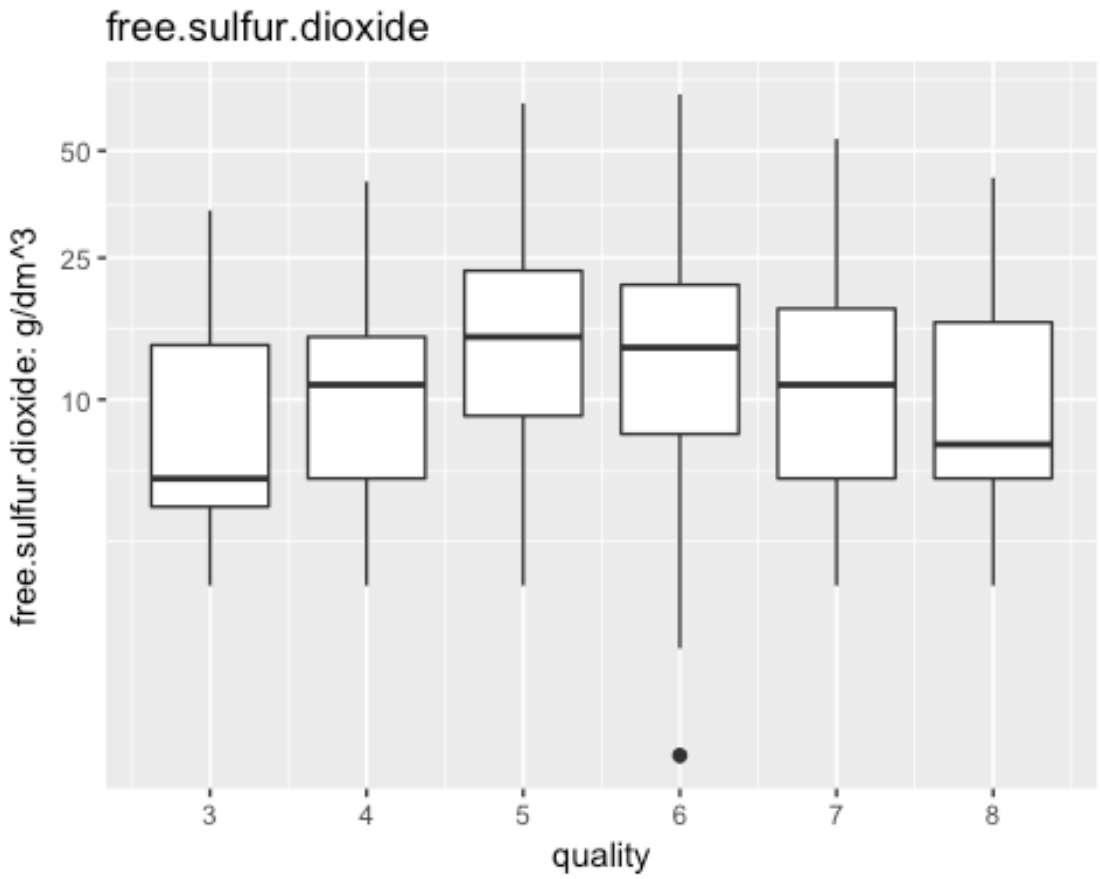
chlorides

Seems to go down with higher quality wine, but not strikingly so.



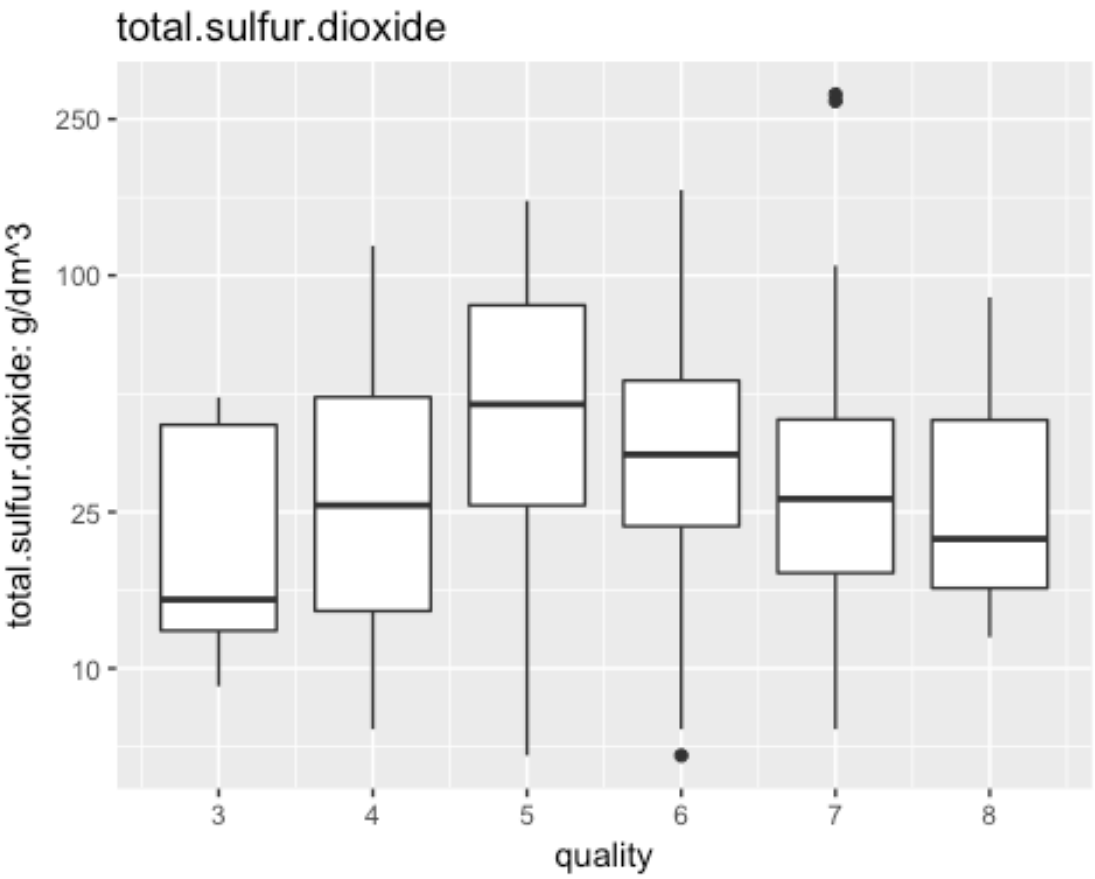
free.sulfur.dioxide

Nothing jumps out



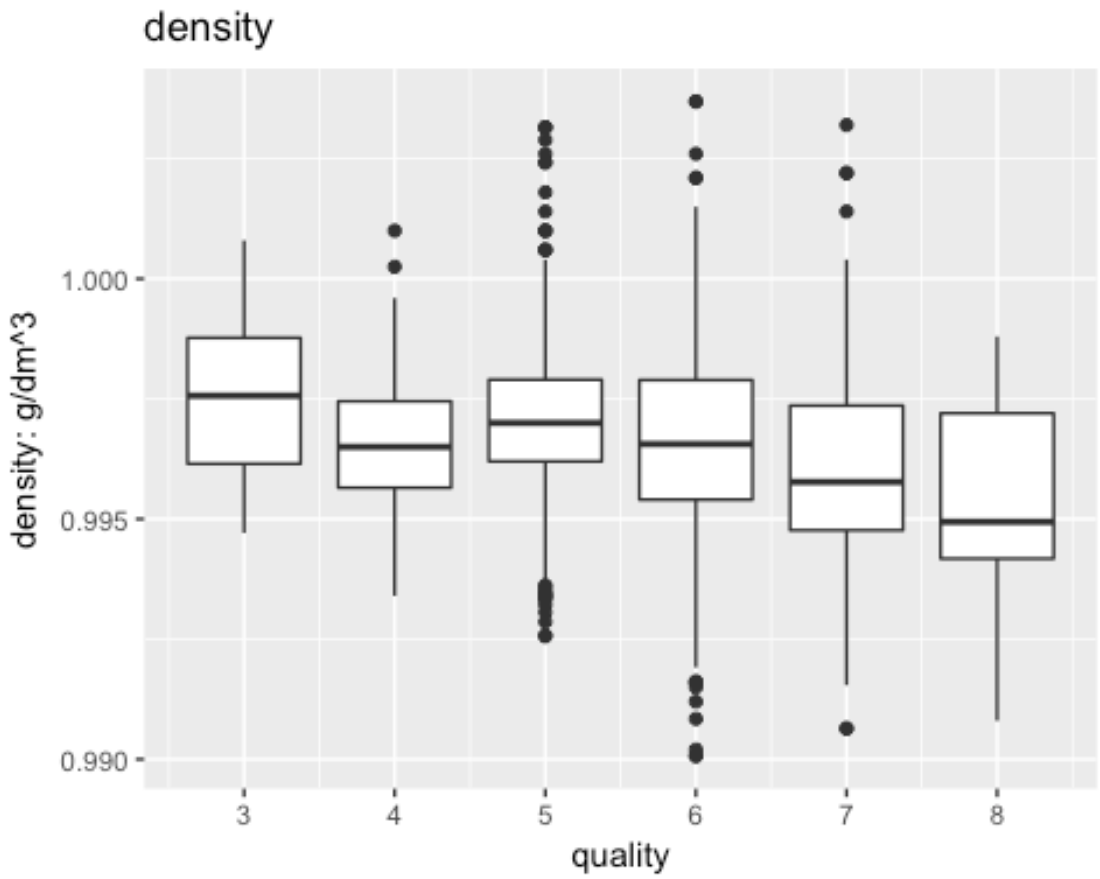
total.sulfur.dioxide

Nothing jumps out



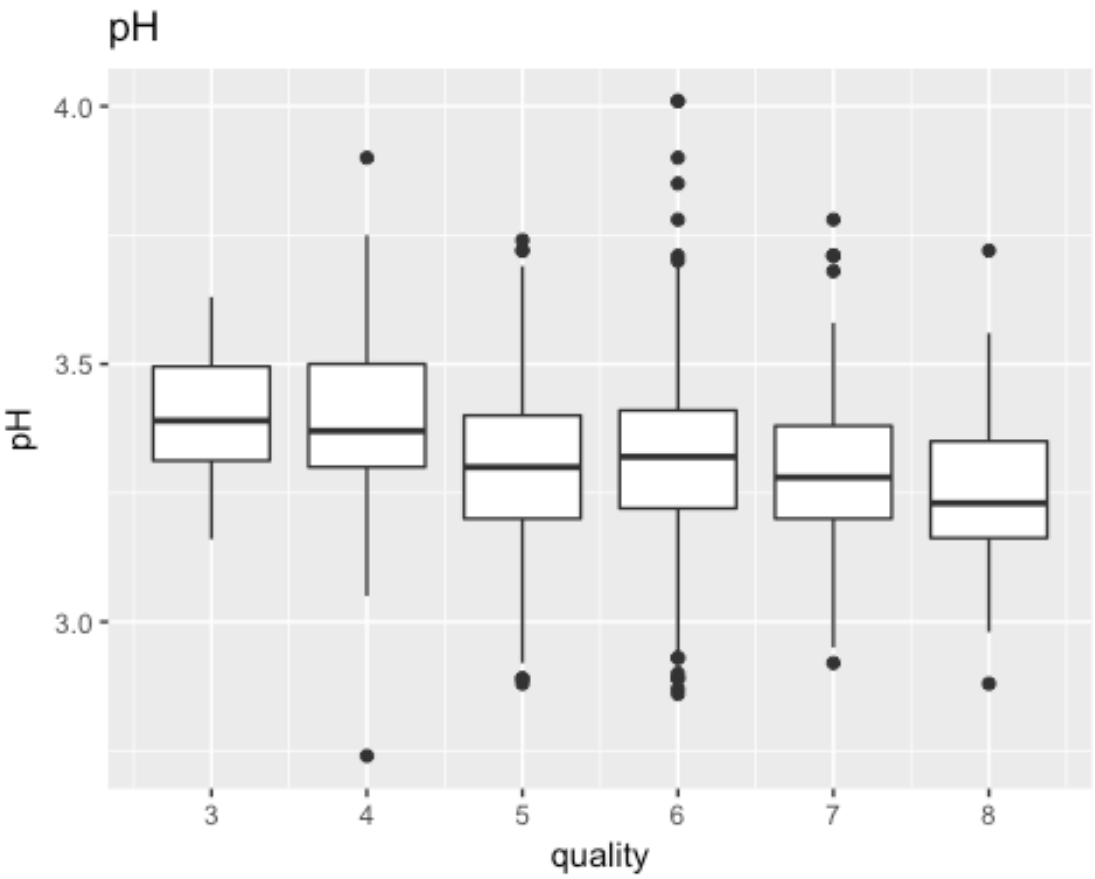
density

Nothing jumps out



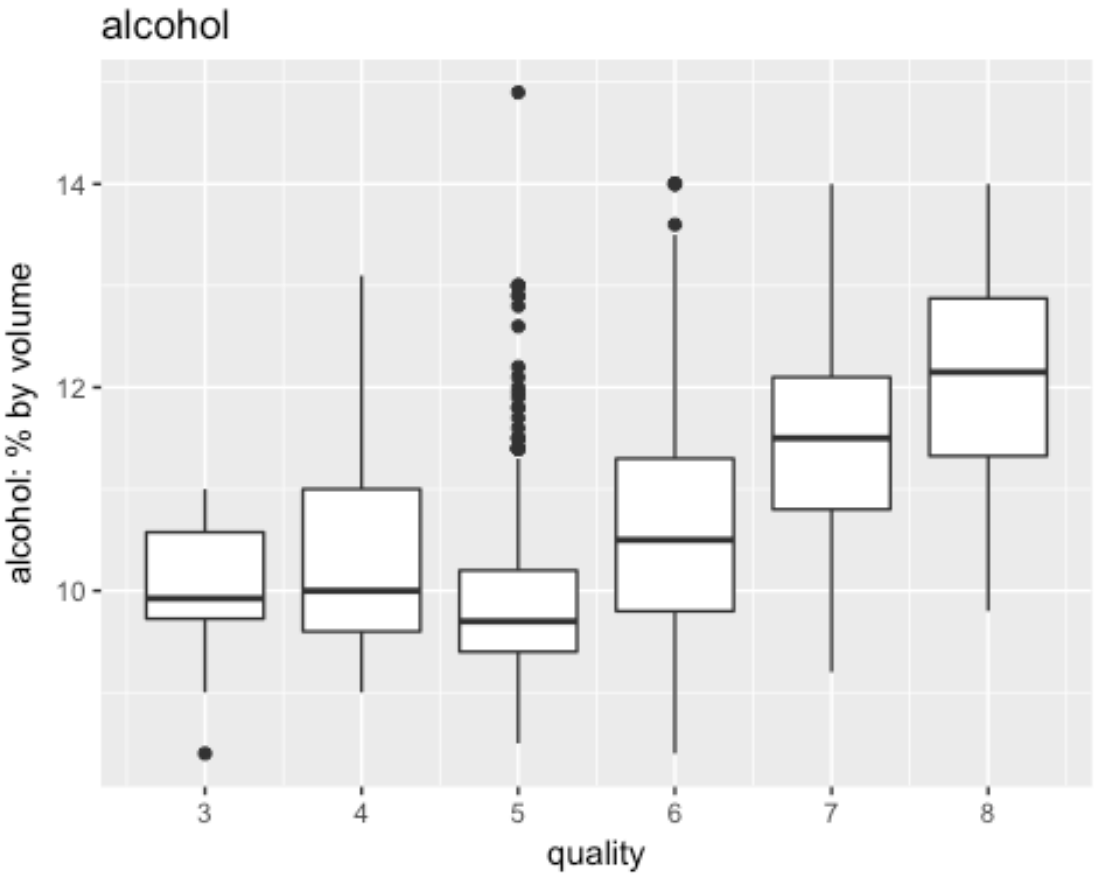
pH

Nothing jumps out



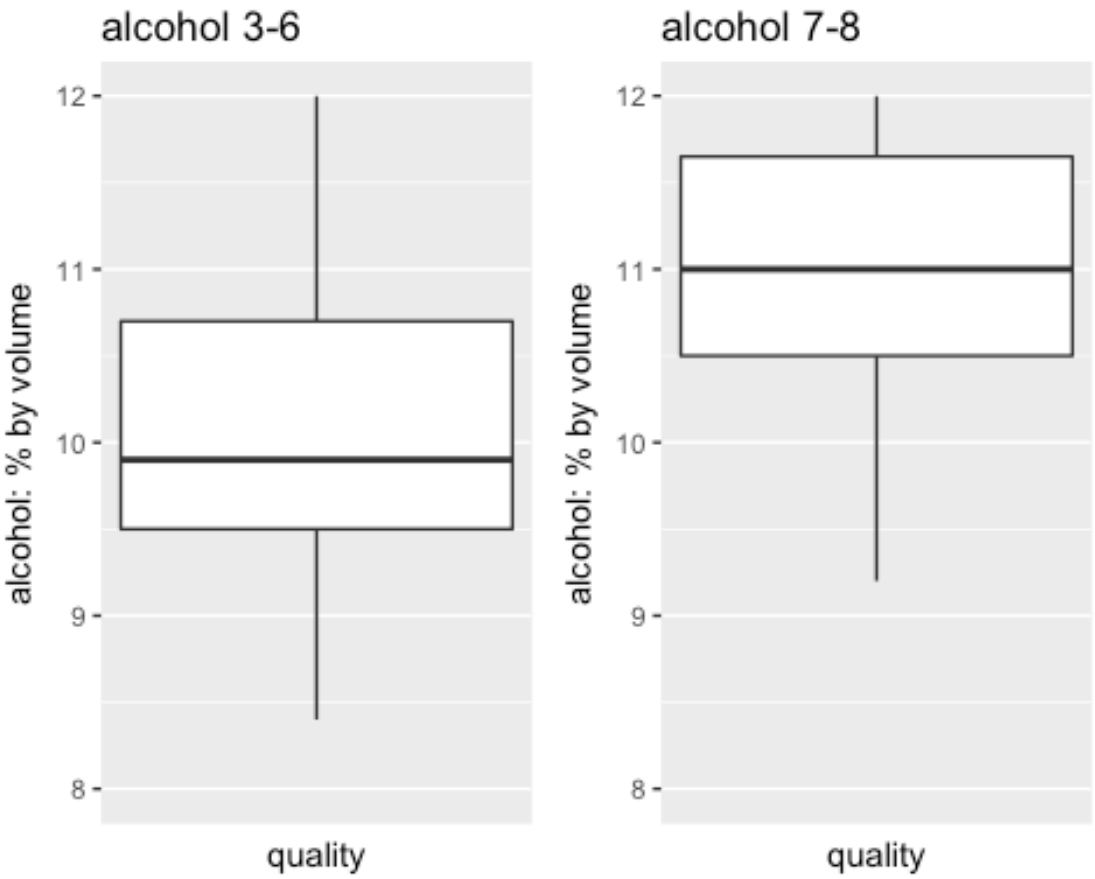
alcohol

Higher quality wines seem to have more alcohol!



Two box plots

Now we arbitrarily decide to bucket wines into high-quality wines (scores 7-8) and other (3-6) with box plots.



alcohol statistics

Mean alcohol 7-8 wines: 11.52 3-6 wines: 10.25

I'm guessing we would find a statistically significant difference between the two groups with respect to alcohol.

The r values are different too, though I'm not sure that's useful. Perhaps it means that really good wines are more complex than just more or less alcohol.

The r values are different too, though I'm not sure that's useful. Perhaps it means that really good wines are more complex than just more or less alcohol.

r 7-8 wines: 0.1740748 3-6 wines: 0.3061033

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.00   10.25   10.90   14.90

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.60   11.52   12.20   14.00
```

```
##
##  Pearson's product-moment correlation
##
## data:  subset(redwine, quality <= 6)$alcohol and subset(redwine, qua
lity <= 6)$quality
## t = 11.945, df = 1380, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2575295 0.3531340
## sample estimates:
##       cor
## 0.3061033

##
##  Pearson's product-moment correlation
##
## data:  subset(redwine, quality >= 7)$alcohol and subset(redwine, qua
lity >= 7)$quality
## t = 2.592, df = 215, p-value = 0.0102
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0418609 0.3002971
## sample estimates:
##       cor
## 0.1740748
```

Bivariate Analysis

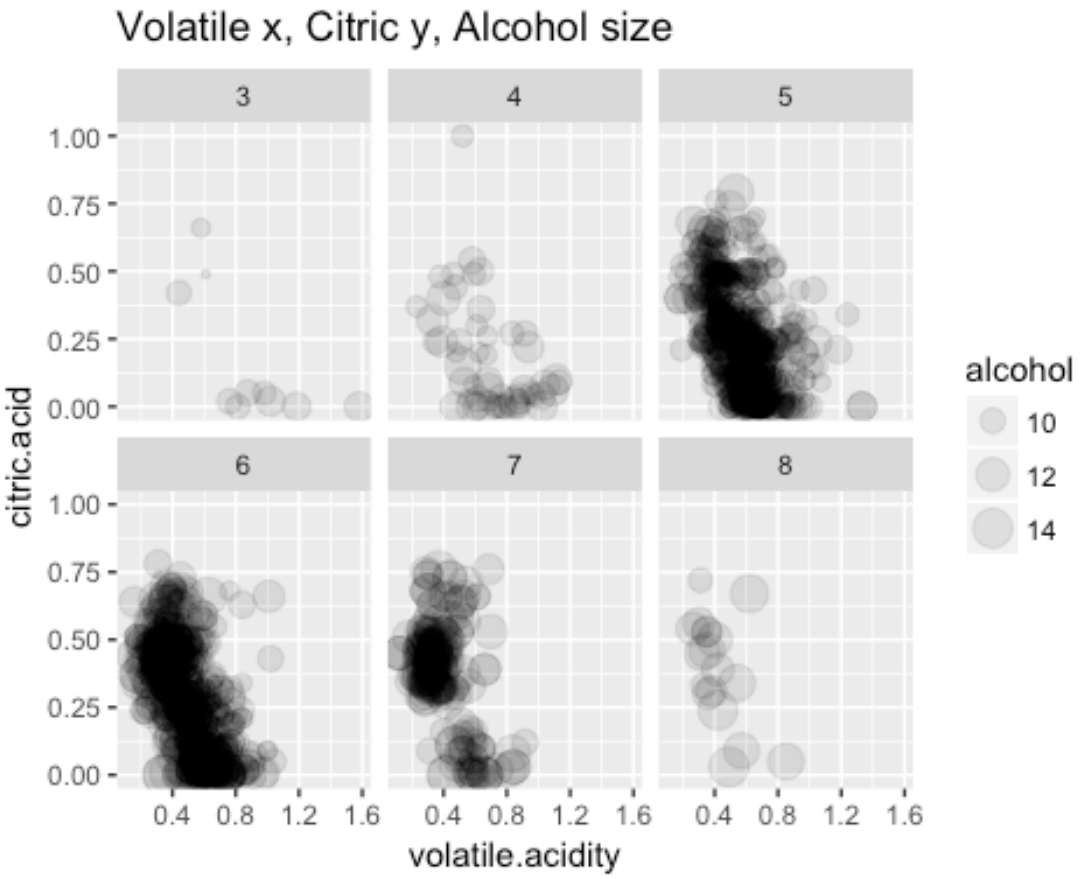
I stated above that we would be surprised if any of them have a very high correlation with quality. Wine making is a complex art, and if this were true, you might be able to easily improve the quality of the wine by addressing this one variable. This did turn out be the case, generally, but see below.

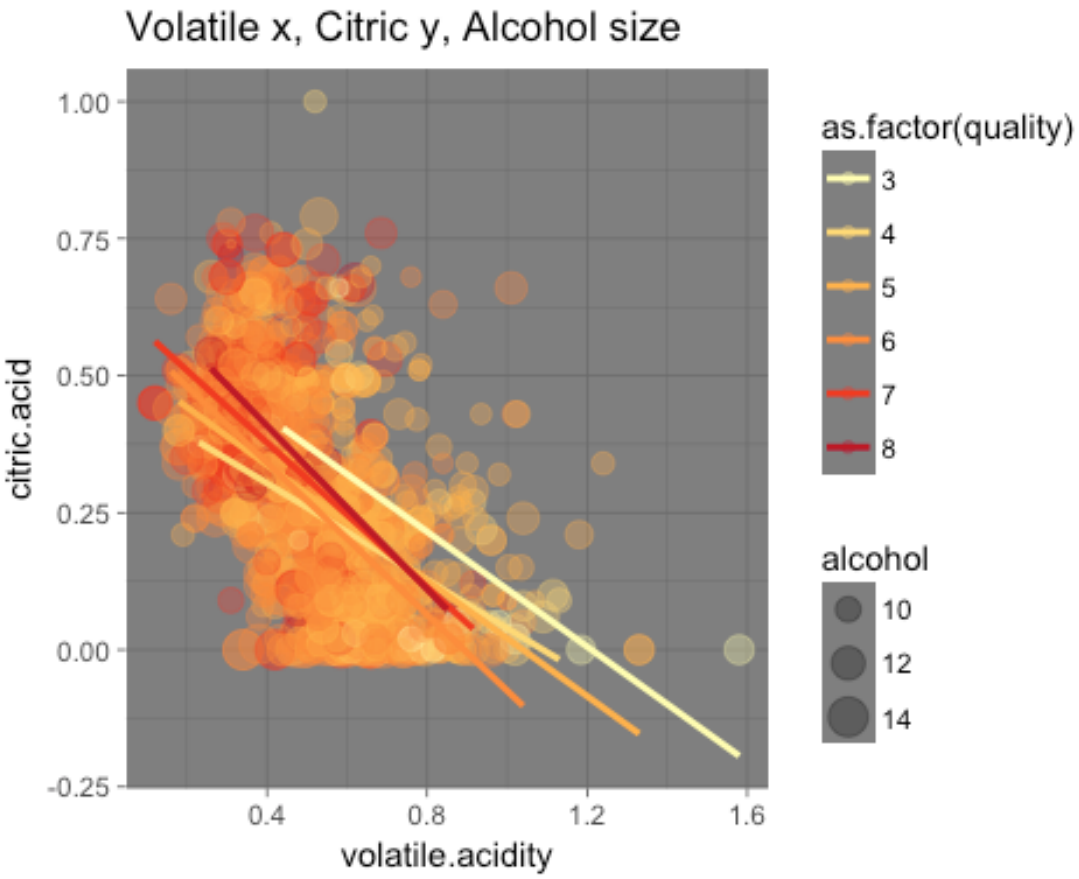
The correlation between volatile.acid and quality, citric.acid and quality, and alcohol and quality, were not particularly high. However, when we split the data points into high-quality (7-8) wine and other (3-6) wine, there was a noticeable difference in the graphs and stats. Quality vs these three measures had the strongest relationship.

Multivariate Plots Section

Multi-variate analyses

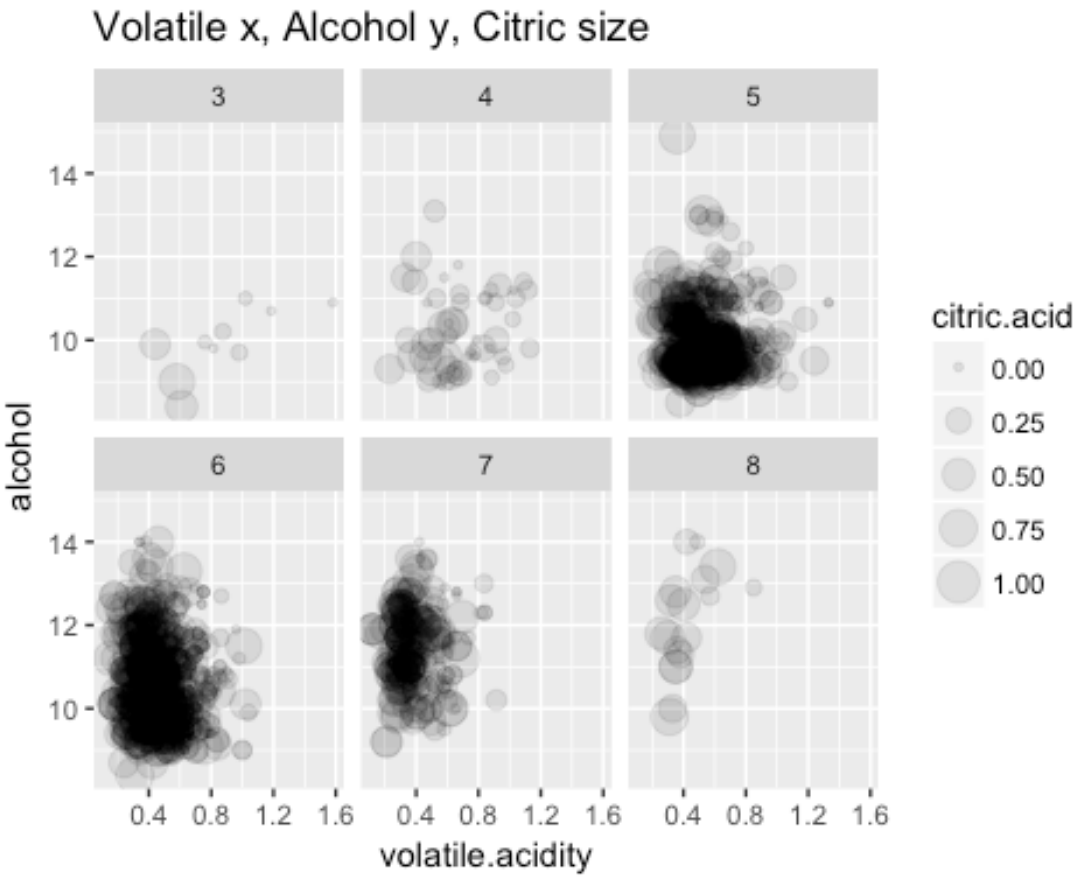
Now we take the three variables citric acid, alcohol, and volatile acidity and use the axes of X,Y,size to look at them. On top of this, we layer (1) faceting by quality and (2) overall different colors for quality. In (2) we also draw regression lines for each quality.

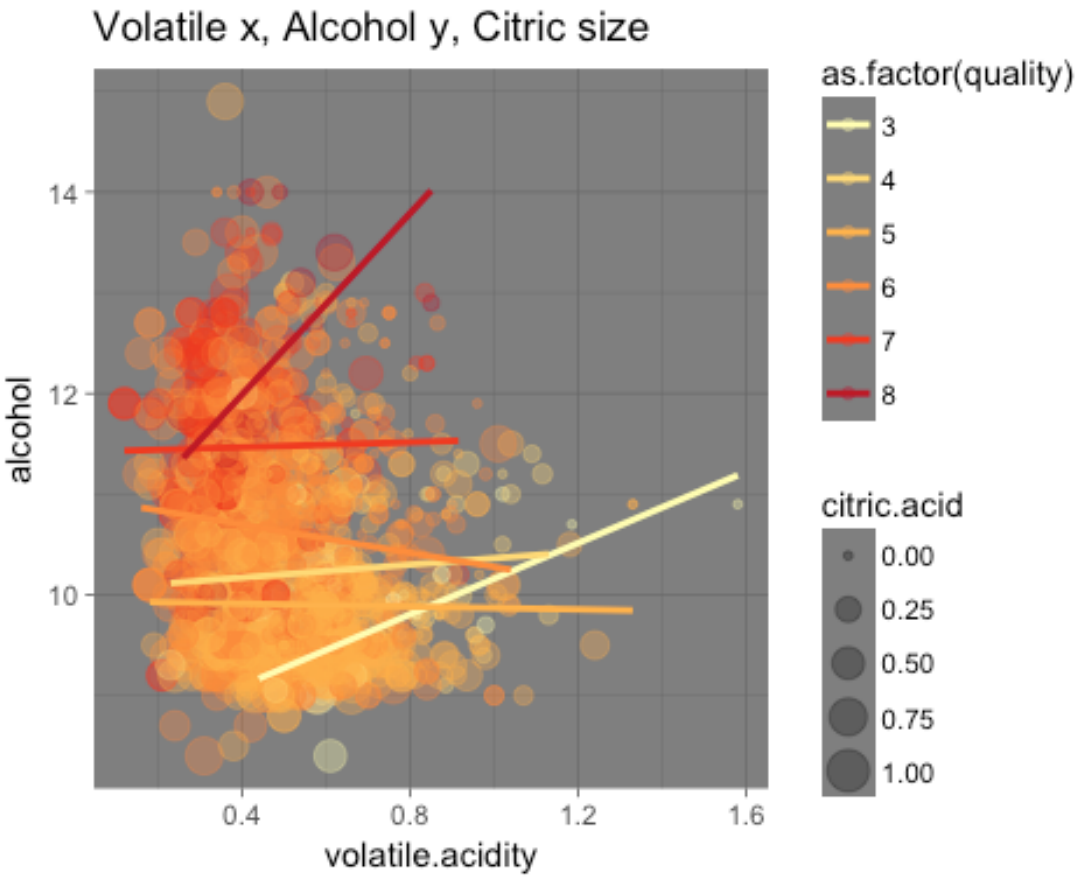

Volatile x, Citric y, Alcohol size
3
4
5
6
7
8
1.00
0.75
0.50
0.25
0.00
citric.acid
0.4
0.8
1.2
1.6
volatile.acidity
alcohol
10
12
14



Observations

I don't see much of a pattern here.3 seems to have a lot of high volatile acid point, and 8 relatively few. But 3 and 8 also have fewer data points to look at. The regression lines are all pretty similar.

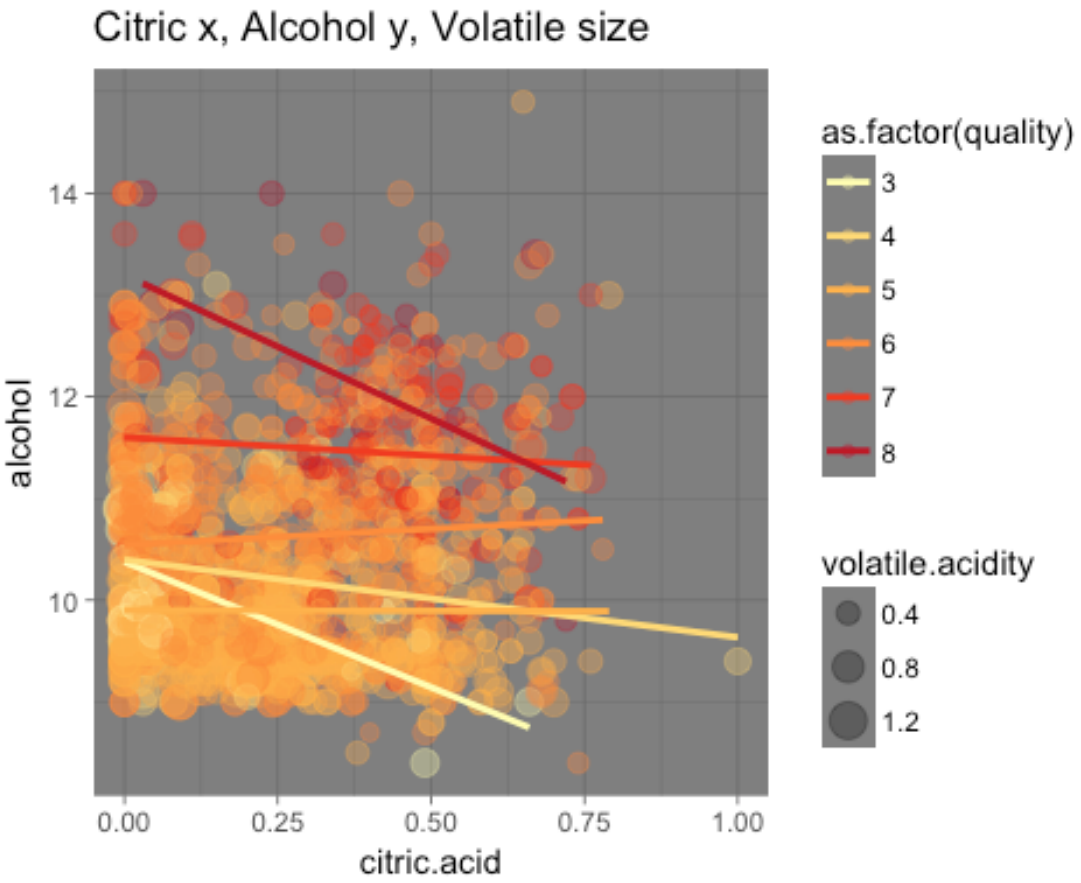

Volatile x, Alcohol y, Citric size
3
4
5
6
7
8
alcohol
14
12
10
0.4
0.8
1.2
1.6
volatile.acidity
citric.acid
0.00
0.25
0.50
0.75
1.00



Observations

Here, the facted 3 and 8 do seem different, with 3 having high volatile.acidity points and 8 having high alcohol points. The regression lines for 3 and 8 are furthest apart on the second plots. Also, interestingly, 3 and 8 have the most similar slopes.


Citrix x, Alcohol y, Volatile size
3
4
5
6
7
8
14
12
10
alcohol
0.00
0.25
0.50
0.75
1.00
citric.acid
volatile.acidity
0.4
0.8
1.2



Observations

In the two plots above, the higher alcohol and higher citric acid in 8 are noticeable here, particularly as compared to the lower alcohol and citric acid in 3, though both have small samples sizes. AThe regression lines for 3 and 8 are furthest apart on the second plots. Also, interestingly, 3 and 8 have the most similar slopes.

Multivariate Analysis

Volatile acid, citric acid, and alcohol had noticeable differences when we separated high-quality wine from the others.

Let's analyze one of the above for statistical significance

Mean volatile.acidity 7-8 wines: 0.4055 3-6 wines: 0.547 Difference: 0.1415

By subsetting these two groups of wine, we get 1382 wines for OTHER and 217 for HQ. We can treat these as independent samples, so we pool their SDs to get the standard error. The HQ wine sample has n = 217 and mean = 0.4055.

Calculations:

```
## [1] 217  13

## [1] 1382   13

## SD1: 0.176337

## SD2: 0.1449627

## Pooled SD: 0.1495447

## t-statistic: 12.95821

## df: 1597

## P statistical significance: very close to 1
```

The colored plots generally work well, in that we see the huge points right on or along the correlation lines.

Modeling of Data:

As we said earlier, it is probably not easy to create a model for good wine, otherwise it might be easy to do biochemically. At the very least, it would require a lot of features. we will use the three features from above which seem most signficant, volatile.acidity, citric.acid, alcohol. we expect these will not make a very good model, and indeed the r we achieve in only 0.3.

m1 <- lm(I(quality) ~ I(volatile.acidity), data=redwine) m2 <- update(m1,~.+citric.acid) m3 <- update(m2,~.+alcohol) mtable(m1,m2,m3)

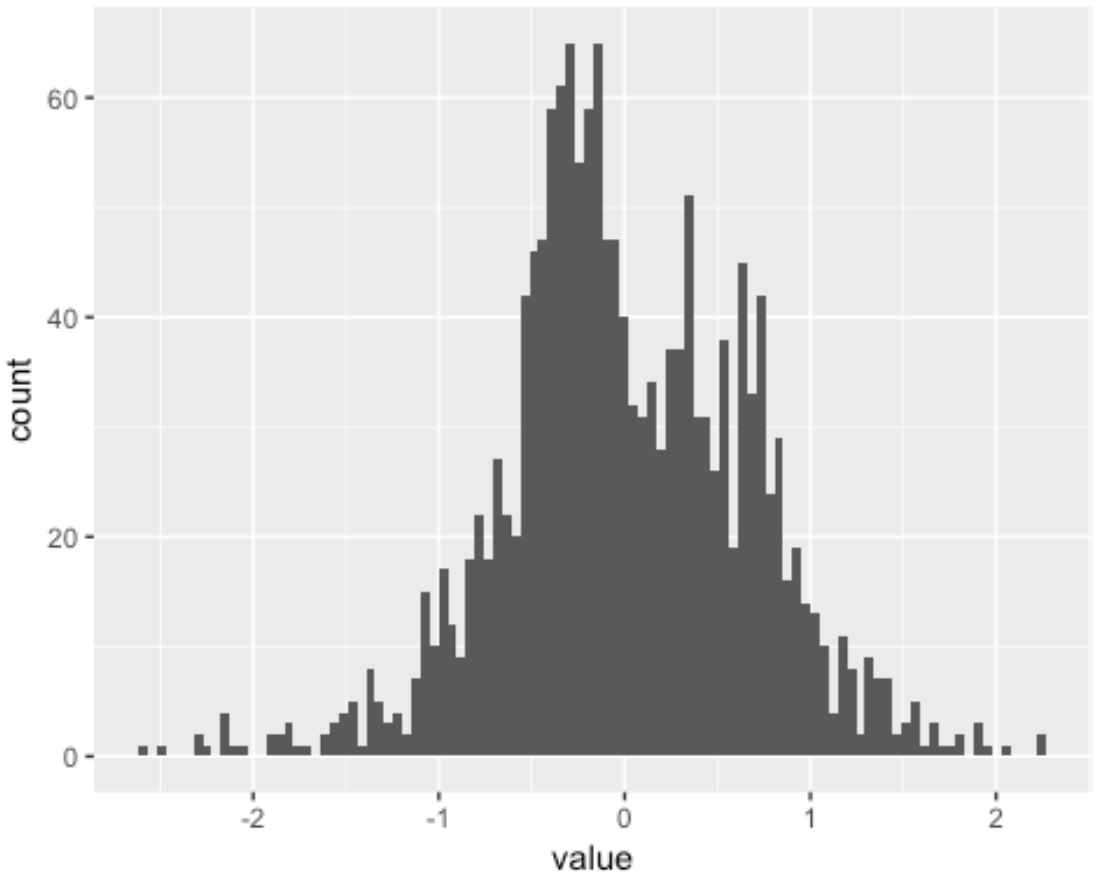
We also print some info on m3: summary(m3) names(m3) anova(m3) head(fitted(coef(m3)) head(coef(m3)) head(residuals(m3))

We plot the residuals and it comes out fairly normal, as we would expect, even though the model isn't that good.

```
##
## Calls:
## m1: lm(formula = I(quality) ~ I(volatile.acidity), data = redwine)
## m2: lm(formula = I(quality) ~ I(volatile.acidity) + citric.acid,
##     data = redwine)
## m3: lm(formula = I(quality) ~ I(volatile.acidity) + citric.acid +
##     alcohol, data = redwine)
##
## ============================================================
##                            m1         m2         m3
## ------------------------------------------------------------
##   (Intercept)           6.566***   6.529***   3.055***
##                        (0.058)    (0.089)    (0.194)
##   I(volatile.acidity)  -1.761***  -1.723***  -1.343***
##                        (0.104)    (0.125)    (0.114)
##   citric.acid                      0.063      0.068
```

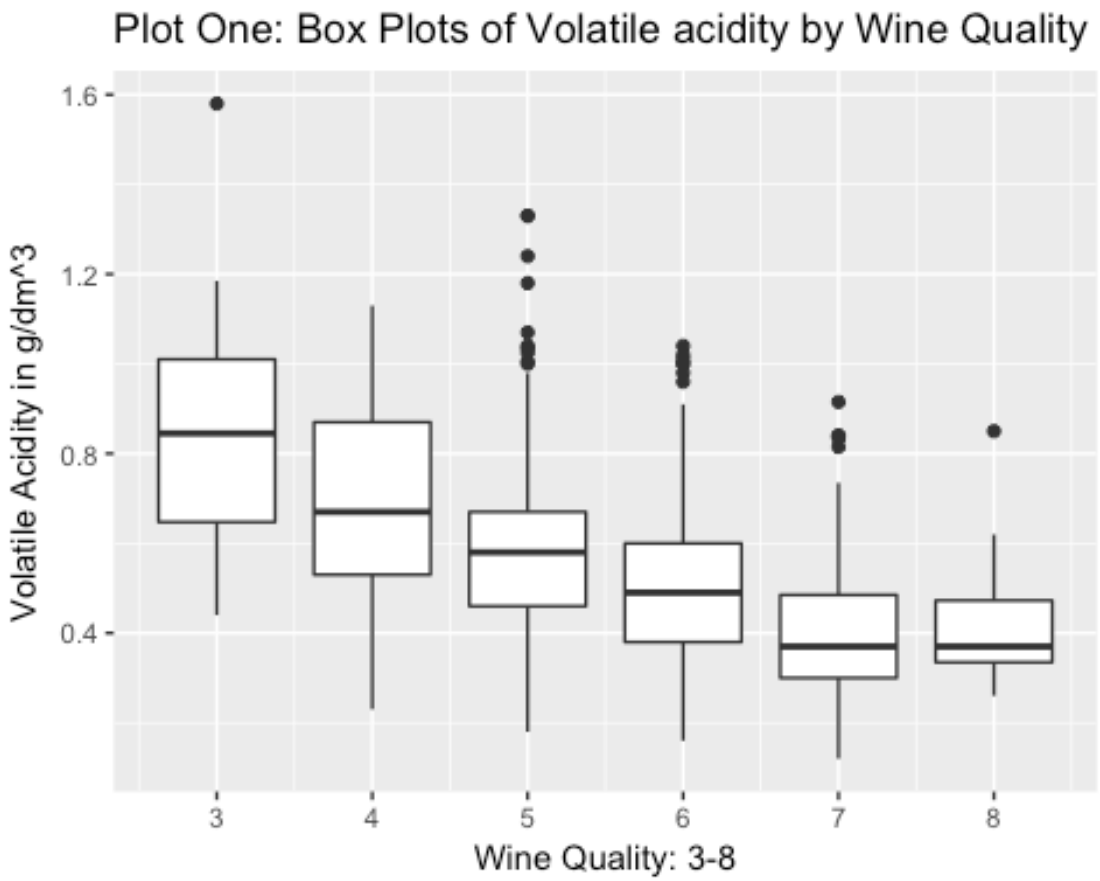
```
##                                    (0.115)    (0.103)
##   alcohol                                      0.314***
##                                               (0.016)
## ------------------------------------------------------
##   R-squared                 0.2        0.2        0.3
##   adj. R-squared            0.2        0.2        0.3
##   sigma                     0.7        0.7        0.7
##   F                       287.4      143.8      247.0
##   p                         0.0        0.0        0.0
##   Log-likelihood        -1794.3    -1794.2    -1621.6
##   Deviance                883.2      883.0      711.6
##   AIC                    3594.6     3596.3     3253.2
##   BIC                    3610.8     3617.8     3280.1
##   N                      1599       1599       1599
## ======================================================

##
## Call:
## lm(formula = I(quality) ~ I(volatile.acidity) + citric.acid +
##     alcohol, data = redwine)
##
## Coefficients:
##         (Intercept)  I(volatile.acidity)          citric.acid
##             3.05533             -1.34286              0.06779
##             alcohol
##             0.31384

##
## Call:
## lm(formula = I(quality) ~ I(volatile.acidity) + citric.acid +
##     alcohol, data = redwine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59992 -0.40354 -0.07282  0.47165  2.23655
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.05533    0.19433  15.722   <2e-16 ***
## I(volatile.acidity) -1.34286    0.11362 -11.818   <2e-16 ***
## citric.acid          0.06779    0.10291   0.659     0.51
## alcohol              0.31384    0.01601  19.602   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6679 on 1595 degrees of freedom
## Multiple R-squared:  0.3172, Adjusted R-squared:  0.3159
## F-statistic:   247 on 3 and 1595 DF,  p-value: < 2.2e-16
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"

## Analysis of Variance Table
##
## Response: I(quality)
##                       Df Sum Sq Mean Sq  F value Pr(>F)
## I(volatile.acidity)    1 158.97 158.967 356.3119 <2e-16 ***
## citric.acid            1   0.17   0.168   0.3771 0.5392
## alcohol                1 171.43 171.427 384.2403 <2e-16 ***
## Residuals           1595 711.60   0.446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##        1        2        3        4        5        6
## 5.065390 4.949210 5.113065 5.792890 5.065390 5.119104

##         (Intercept) I(volatile.acidity)         citric.acid
##          3.05532710         -1.34285820          0.06779379
##             alcohol
##          0.31383655

##           1           2           3           4           5
6
## -0.06538992  0.05078993 -0.11306480  0.20711049 -0.06538992 -0.11910
425
```


60
40
20
0
count
-2
-1
0
1
2
value

Final Plots and Summary

Plot One



Description One

Plot one above was my first surprising findings. Up till this point we was going on a fishing expedition, since we know nothing about wine.

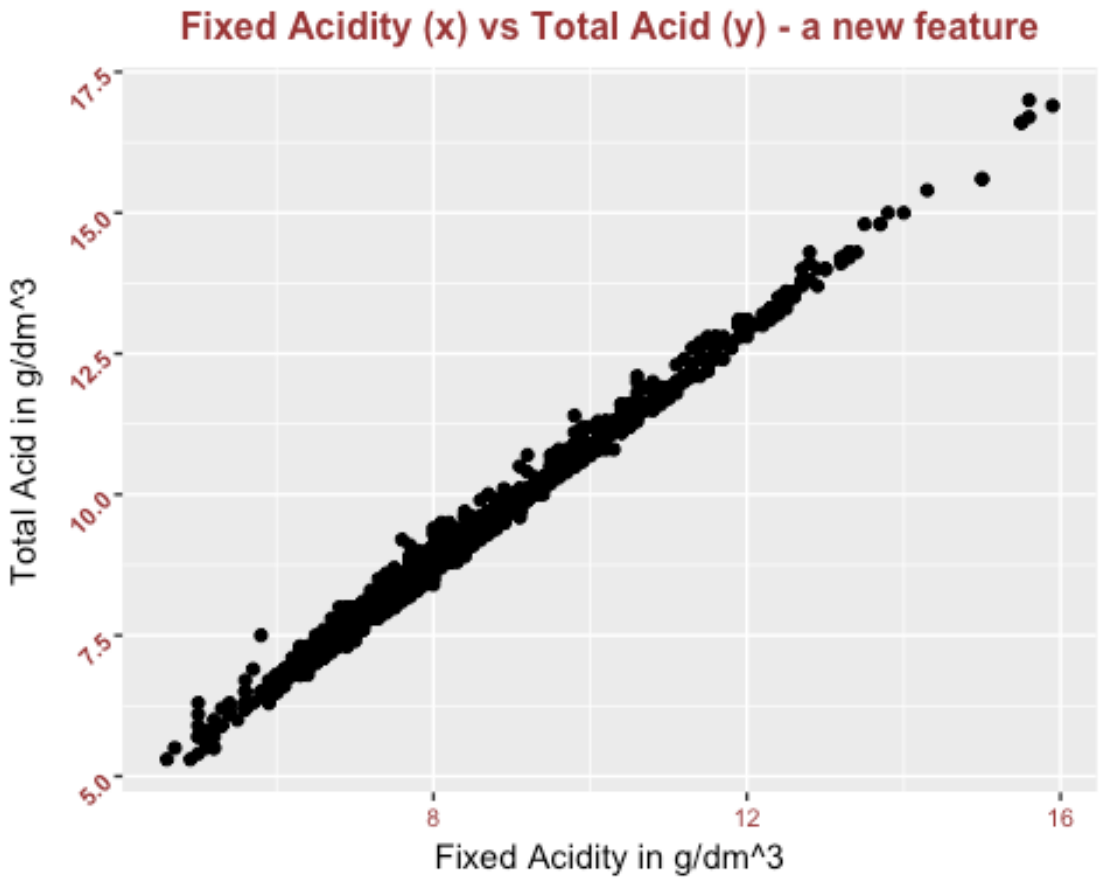
I said: "Aha!!! Volatile acidity goes seem to go down with higher quality wine."

Before seeing this plot, we thought there might not be any interesting relationships between any of the chemical features and wine uality.

Regarding this set of box plots, note the difference in the means: Mean volatile.acidity 7-8 wines: 0.4055 3-6 wines: 0.547

I guessed we would find a statistically significant difference between the two groups with respect to volatile.acidity, and we confirmed this by calculating a T-statistic and P (very close to 1) above.

Plot Two



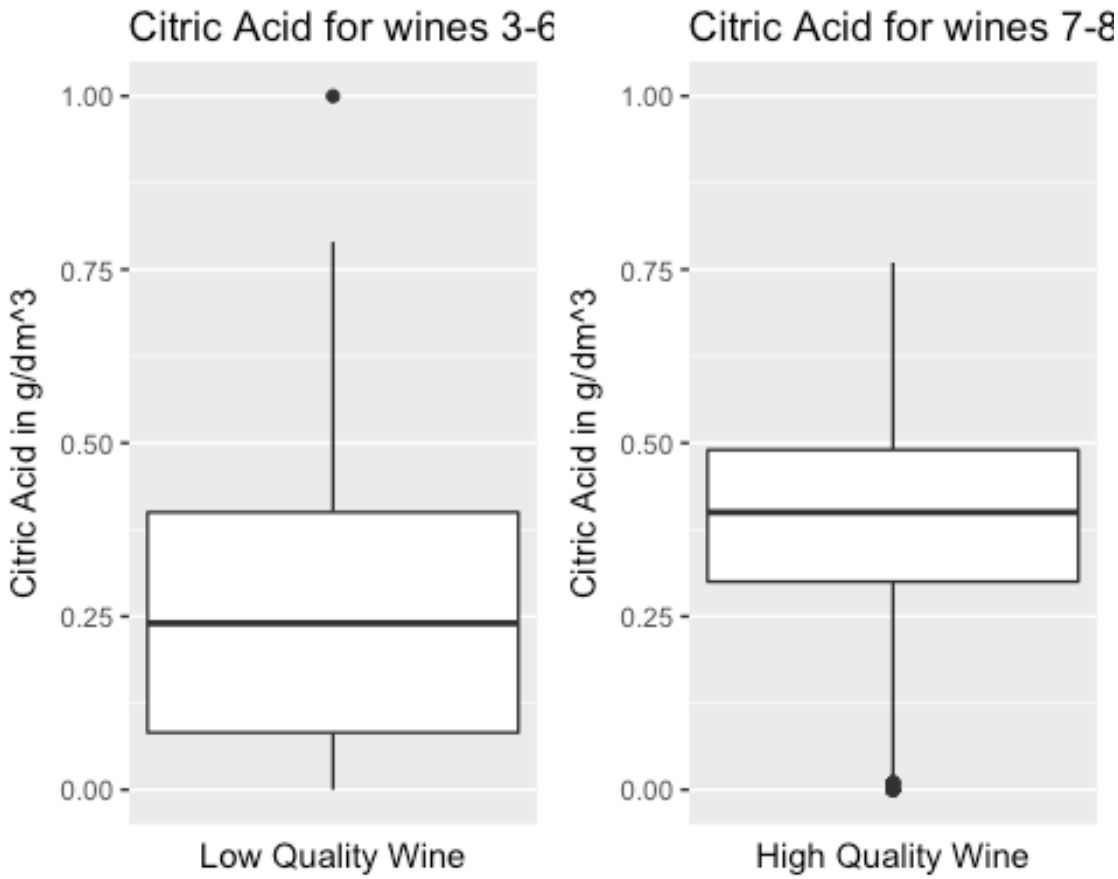
Description Two

This plot has the results that we expected. we saw that the Fixed Acidity amounts were about an order of magnitude greater in amount than the other acids (Volatile Acidity, Citric Acid). And we can see that the scatter plot points are all pretty much right around the line. r calculated below for Fixed Acidity and Total Acid is very close to 1. This means that in terms of absolute amounts of acid, we could just look at Fixed Acidity and ignore the other two.

```
##
##  Pearson's product-moment correlation
##
## data:  redwine$fixed.acidity and redwine$totalacid
## t = 456.98, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9958069 0.9965528
## sample estimates:
##      cor
## 0.996198
```

Now, if one of acids was much stronger than the others, then this conclusion in terms of "amount" of acid would not be useful in terms of the pH effect of the acid. We don't have any information on the relative strengths of these acids. But when we calculated r for Fixed Acidity vs pH and Total Acid vs Ph, we see the r values are almost identical. Therefore, for pH purposes, we can say that Fixed Acidity is essentially the same as Total Acid. -0.6829782 -0.6833998

Plot Three



Description Three

While Fixed Acidity is a good proxy for Total Acid and acid pH effect, it turns out that we cannot ignore the lower quantity acids in terms of correlation with wine quality. Here we see that high quality wines have quite a bit more Citric Acid. I'm guessing we would find a statistically significant difference between these two means.

Mean Citric Acid amounts: 7-8 wines: 0.3765 3-6 wines: 0.2544

Reflection

Since we know nothing about wine, this EDA was a bit of a starter fishing expedition, so we suspect that we would have to do quite a bit more EDA to some meaningful results. As long as this project is, we suspect we would have do to as much EDA again to some decent leads on what to analyze. Some general research on wine might help.

One thing we am beginning to suspect is that these chemical variable shave little to do with wine quality (except perhaps at the extremes). Therefore, in a way, it is weird to have a bunch of purely chemical features such as density, and then one subject human measure, quality. we suspect that it would be more interesting to compare other variables to quality, such as varietal, growing region, year, and not so much these chemical characteristics. Also, it might be more interesting to compare within subject criteria themselves. In other words, if there are other wine quality / taste measure than simply the one, there might be more interesting data patterns in that then what we have seen with this EDA.