Explore and Summarize Data

Udacity NanoDegree Project 3

Coursework

- Data Analysis with R

By Joe Nyzio

Explore and Analyze Data

Udacity Nanodegree

By Joe Nyzio

Introduction

How would you make the perfect wine? I've been given a list of wines, their expert quality ratings, and their ingredients. I want to see if there is a way to predict what the experts would say about a wine before they even taste it.

Summary of the data

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

```
##        X         fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
```

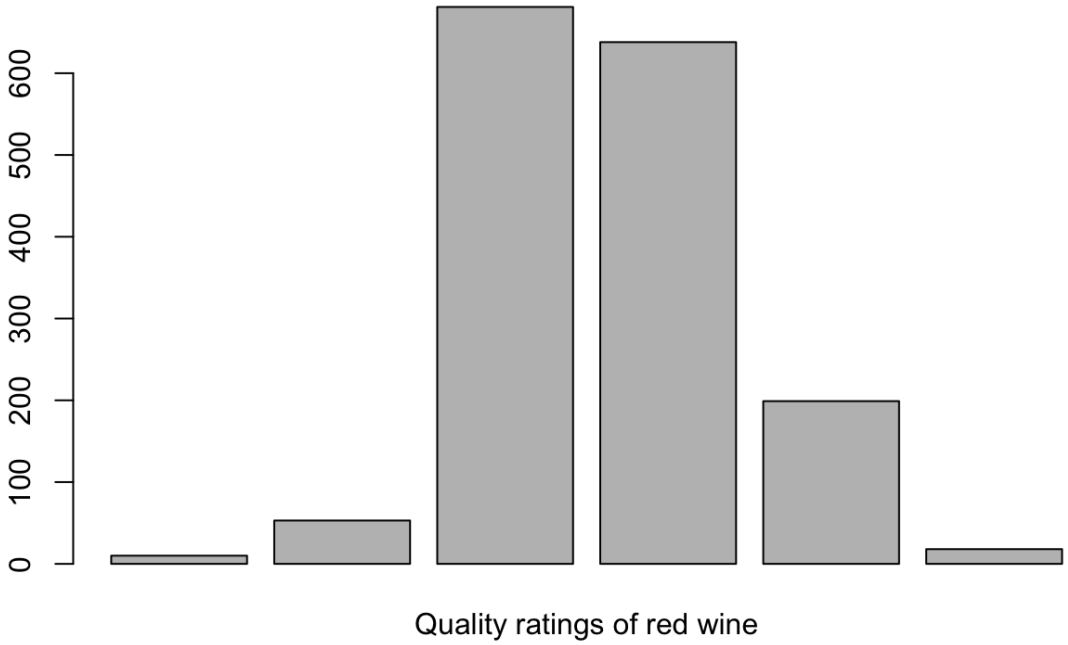
```
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000

## '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 0.47 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 Plots Section

How much of each ingredient is in our data?

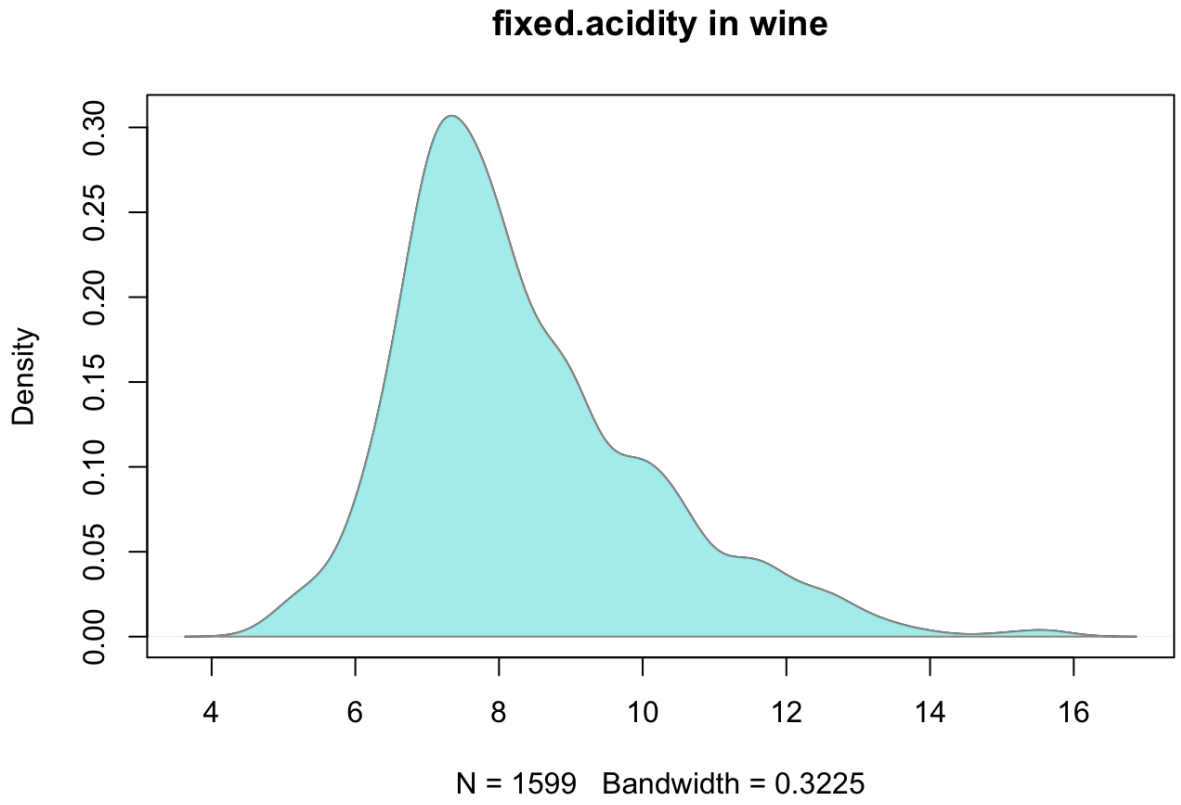
To start I'll just be making a density plot for each variable in the dataset. Nothing too exciting here but these plots will help us get an idea of the distributions of each variable.

```
##
##  3   4   5   6   7   8
##  10  53 681 638 199  18
```



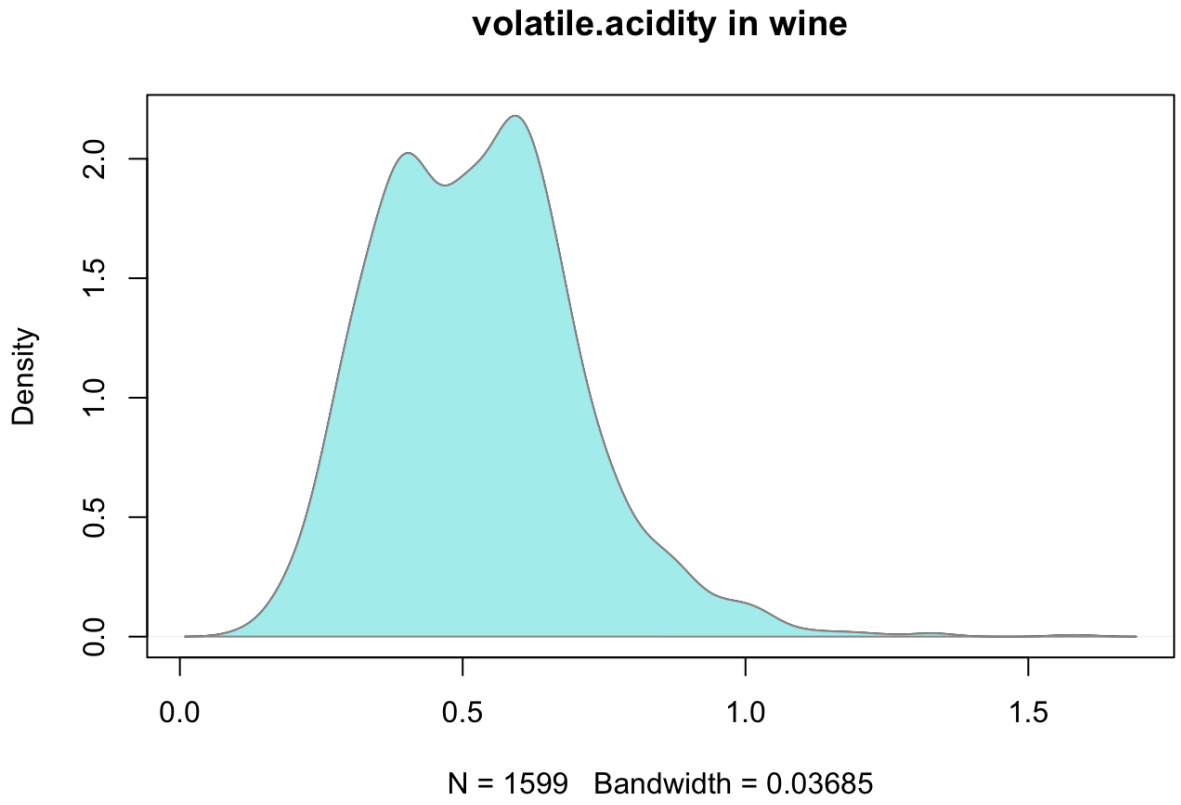

```
##
##   3   4   5   6   7   8
##  0.6  3.3 42.6 39.9 12.4  1.1
```

DESCRIPTION: A histogram that is somewhat normally distributed showing that a majority of the plots are rated 5 or 6. There is a sharp decline as the bounds extend out toward 3 and 8 with no wines being rated greater than 8 or less than 3.



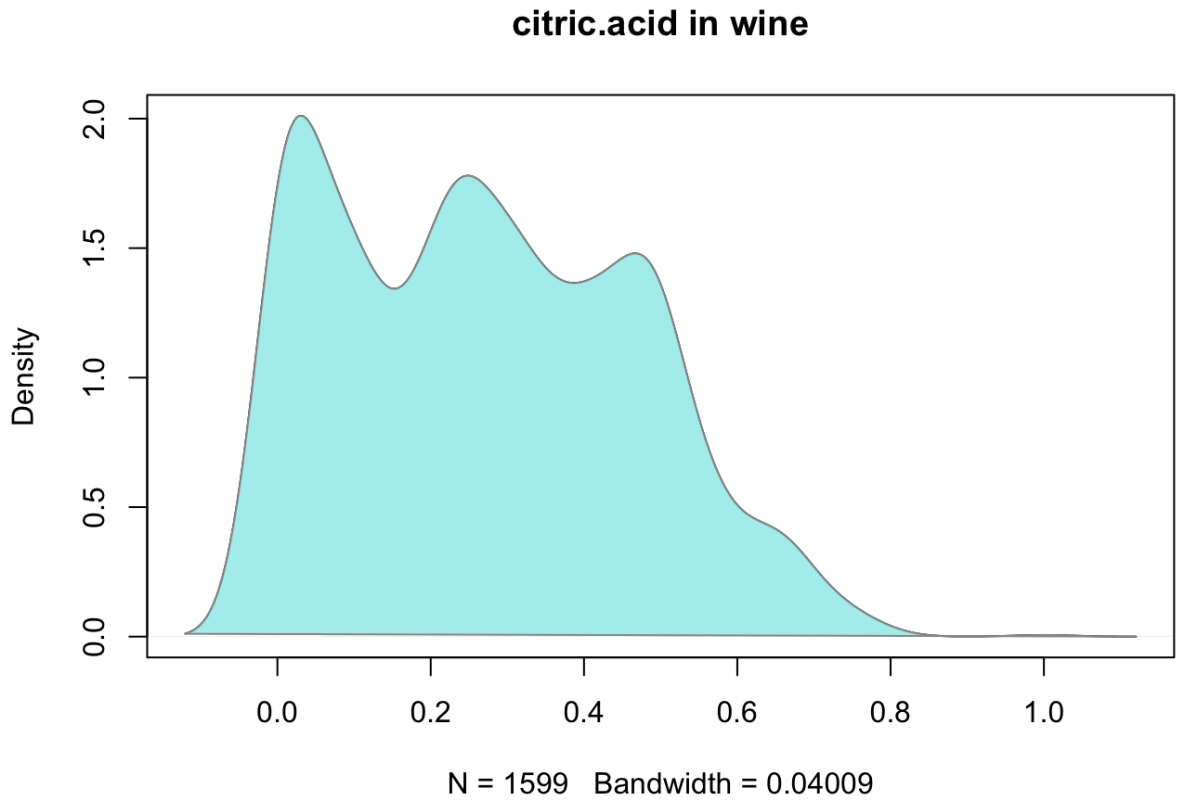

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
## [1] 8.319637
## [1] 1.741096
##  10%  50%  90%
##  6.5  7.9 10.7
```

DESCRIPTION: A a relatively normal density distribution whose x axis extends from 0 to 16 peaking just below 8 skewed slightly to the right with secondary interest around 10 and 12. The y axis reads 0 to .30 and the peak reaches this limit.



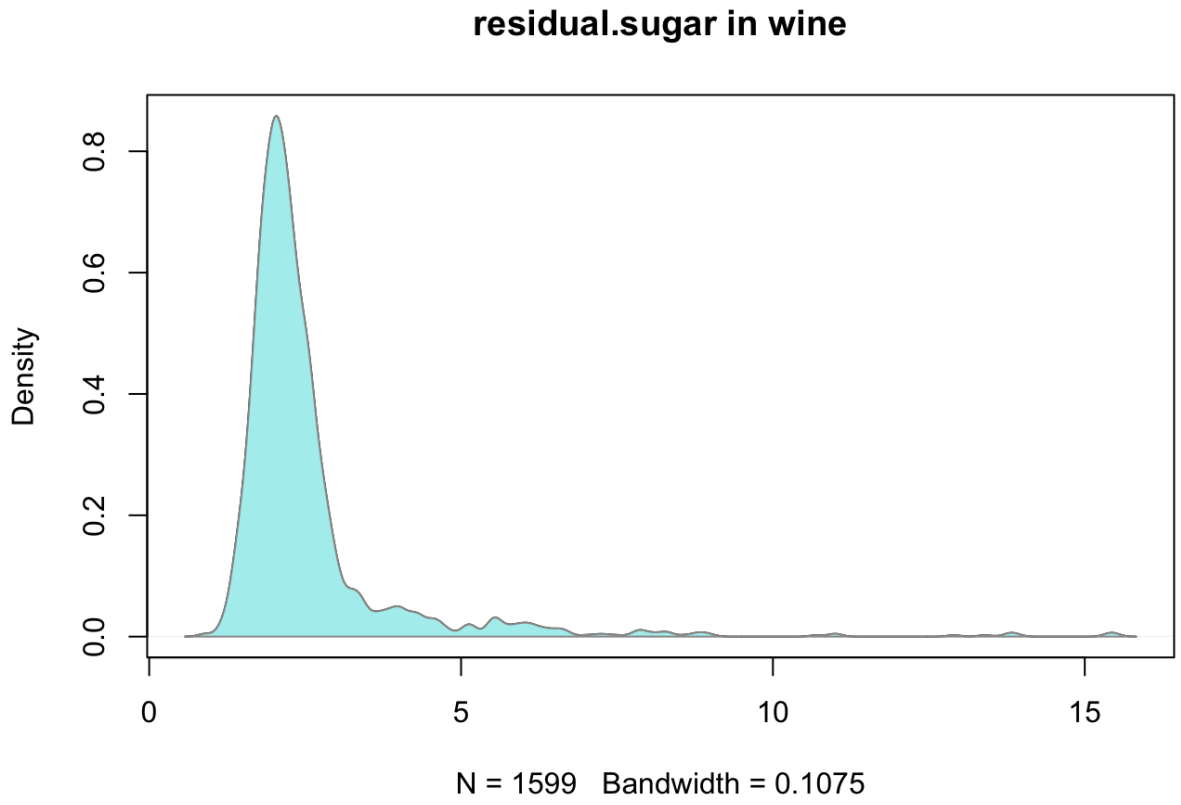

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
## [1] 0.5278205
## [1] 0.1790597
##   10%   50%   90%
## 0.310 0.520 0.745
```

DESCRIPTION: A normal distribution whose x axis extends from 0 to 1.5 with a dip around .5 creating bimodal peaks at .3 and .7 tapering off by around .2 and 1 with a slight surge at 1. The y axis extends reads from 0 to 2.0 and both peaks approcimately reach this limit with the 2nd being slightly higher.



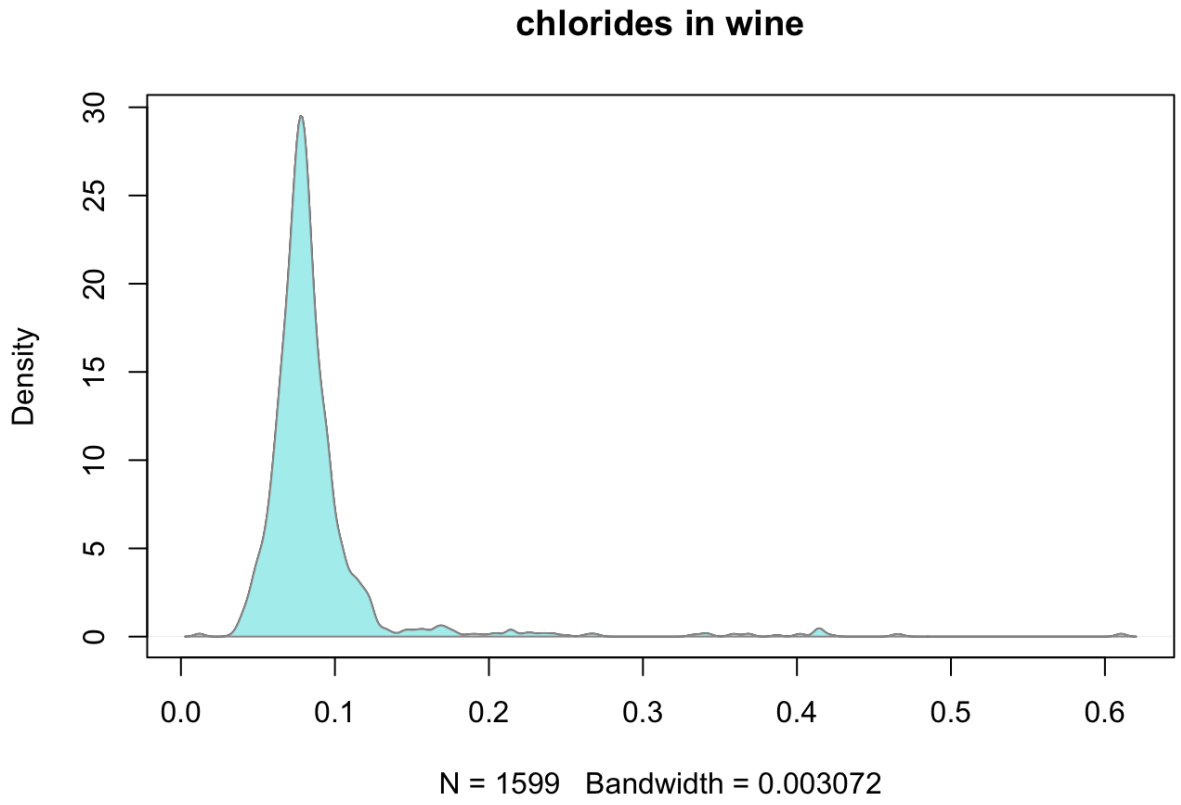

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
## [1] 0.2709756
## [1] 0.1948011
##   10%   50%   90%
## 0.010 0.260 0.522
```

DESCRIPTION: A trimodal normal distribution whose x axis extends from 0 to 1.0 with 3 peaks occuring at 0.1, 0.3, and 0.5. The y axis reads 0 to 2.0 and the peaks reach 2.0, 1.7, and 1.5 respectively.



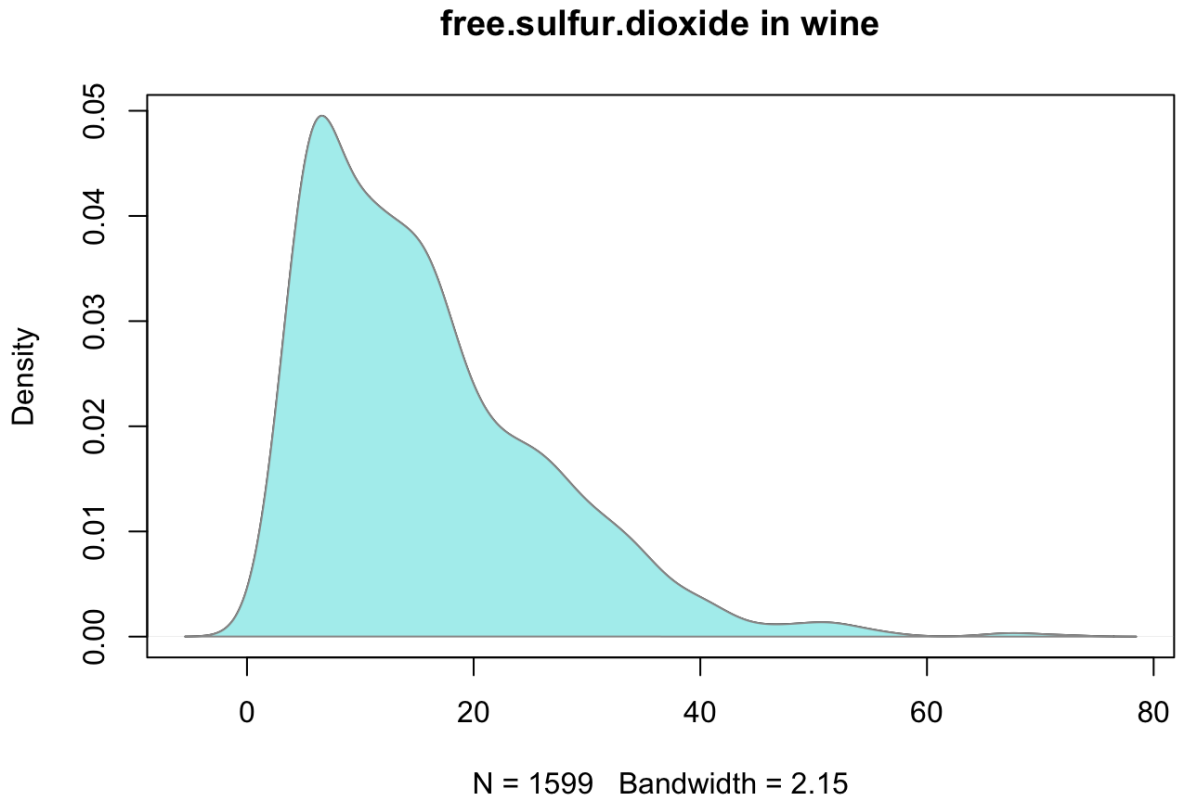

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
## [1] 2.538806
## [1] 1.409928
## 10% 50% 90%
## 1.7 2.2 3.6
```

DESCRIPTION: A normal distribution whose x axis extends from 0 to 15. There is a sharp peak at around 2.5 which extends to the limit of the y axis which is .8. It quickly tapers off by 0 and 4 with little action elsewhere.



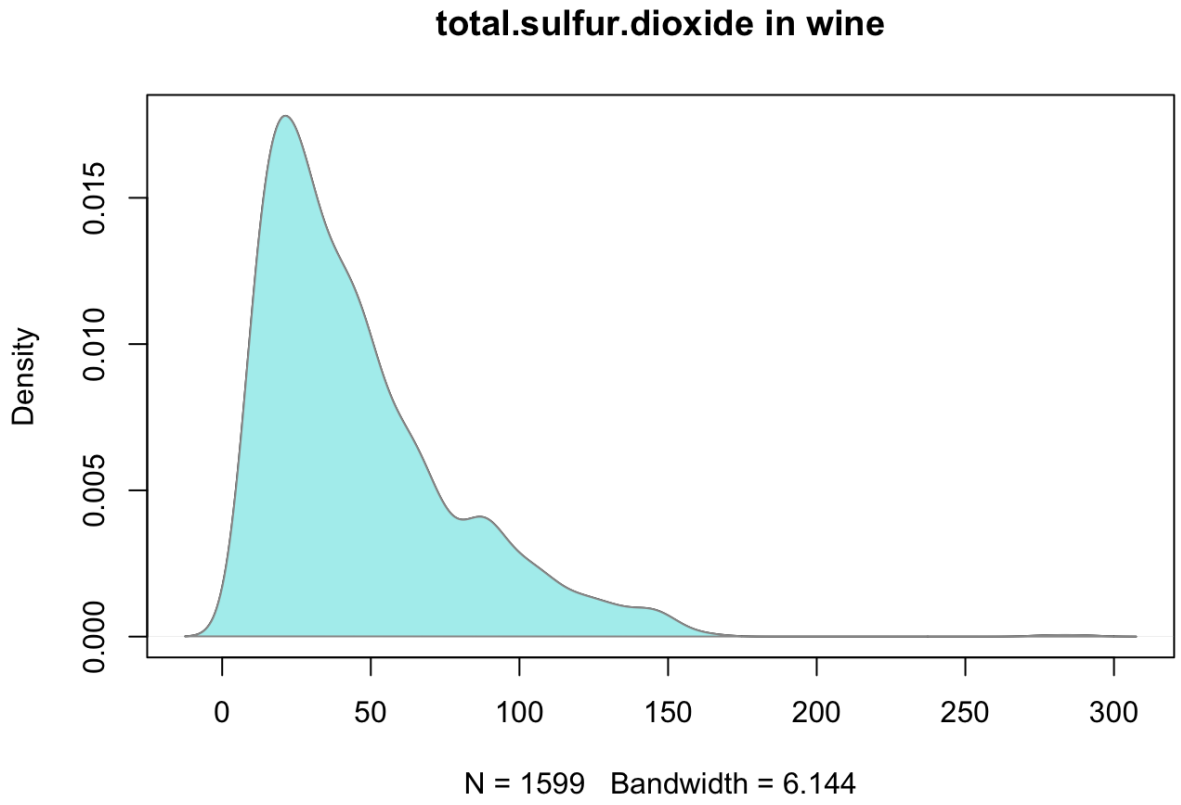

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
## [1] 0.08746654
## [1] 0.0470653
##   10%   50%   90%
## 0.060 0.079 0.109
```

DESCRIPTION: A normal distribution whose x axis extends from 0.0 to 0.6. There is a sharp peack around .09 reaching near the limit of the y axis at 30 that quickly tapers off by .05 and 1.4.



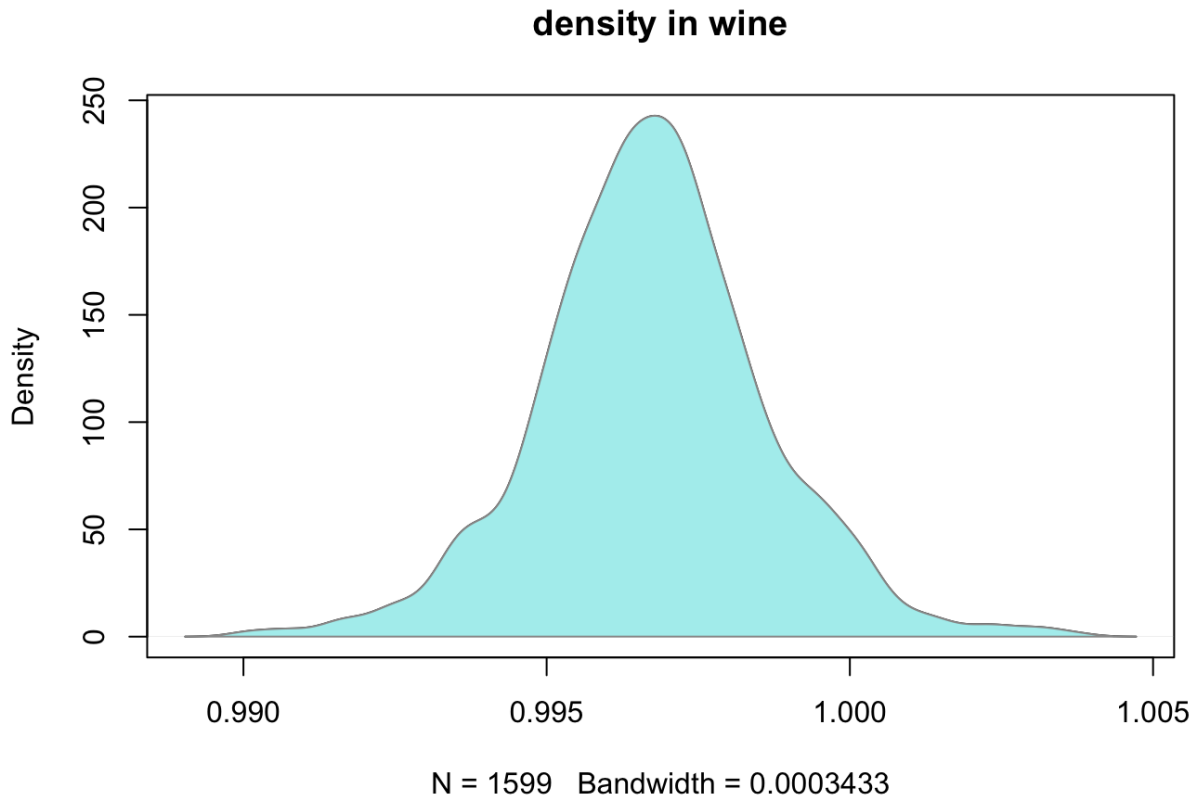

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
## [1] 15.87492
## [1] 10.46016
## 10% 50% 90%
##   5  14  31
```

DESCRIPTION: A normal distribution whose x axis extends from 0 to 80 and y axis from 0.00 to 0.05. There is a peak at 10 with subpeaks around 18 and 22 which contribute to its heavy right skew that finally tapers off at limits 0 and 60.



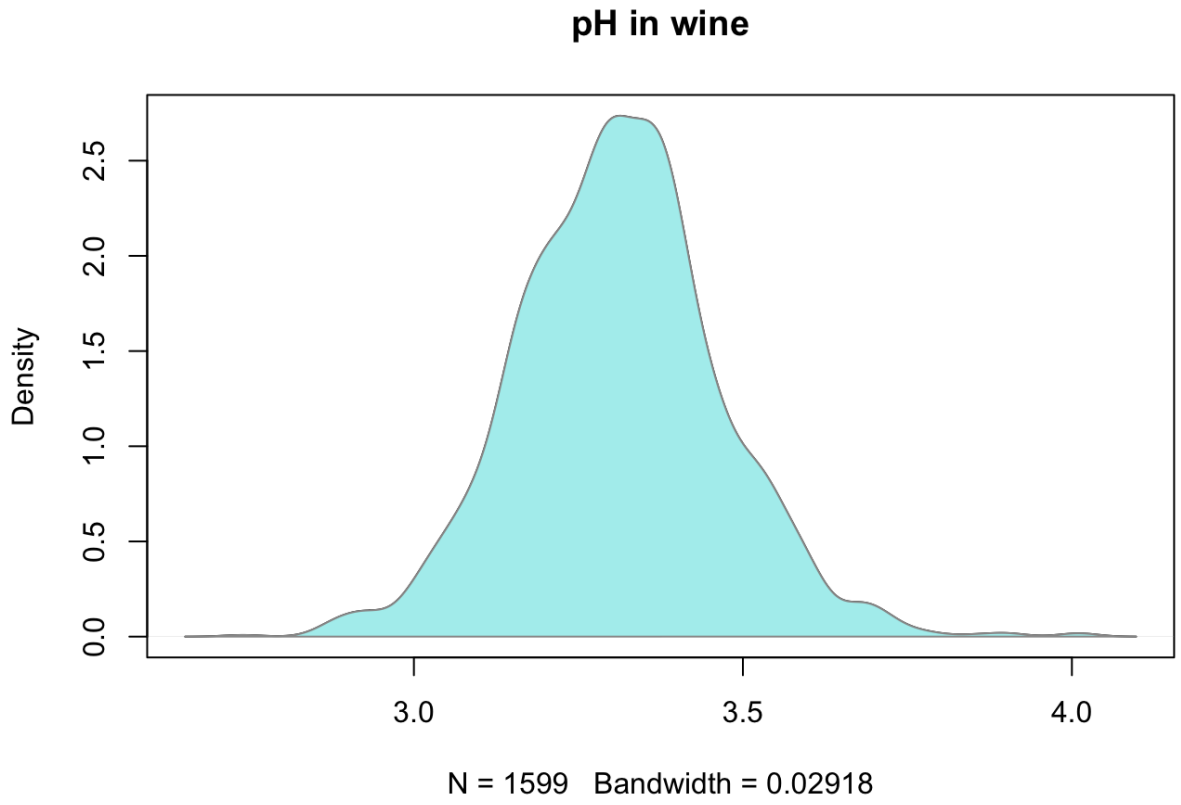

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
## [1] 46.46779
## [1] 32.89532
##  10%  50%  90%
## 14.0 38.0 93.2
```

DESCRIPTION: A normal distribution whose x axis extends from 0 to 300 and y axis from 0.015. There is a peak at 30 reaching the limits of the y axis and subpeaks around 90 and 150 that contribute to its right skew before fully tapering off at limits of 0 and 170 before becoming inactive.



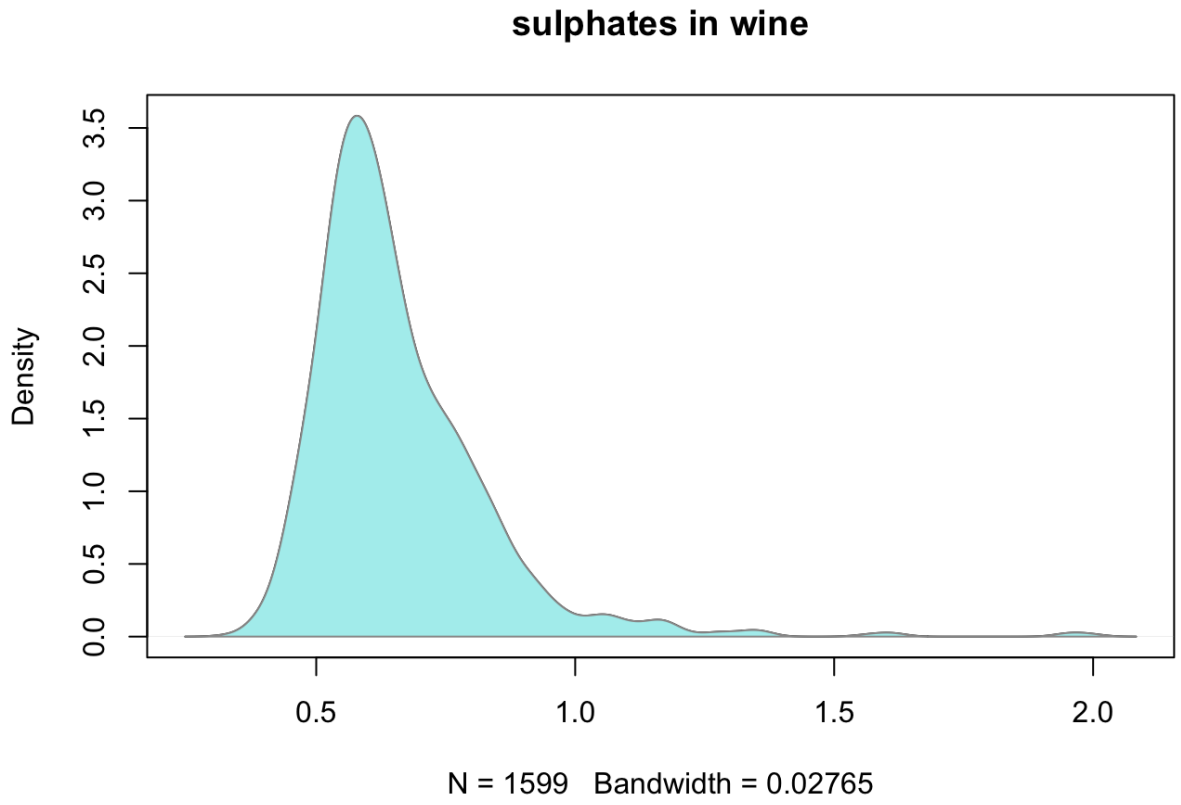

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
## [1] 0.9967467
## [1] 0.001887334
##      10%      50%      90%
## 0.994556 0.996750 0.999140
```

DESCRIPTION: Very close to a real normal distribution. The x axis extends from .990 to 1.005 and the y axis from 0 to 250. The peak occurs around .997 with bumps at .994 and .997 before tapering off to the limits of the x axis.



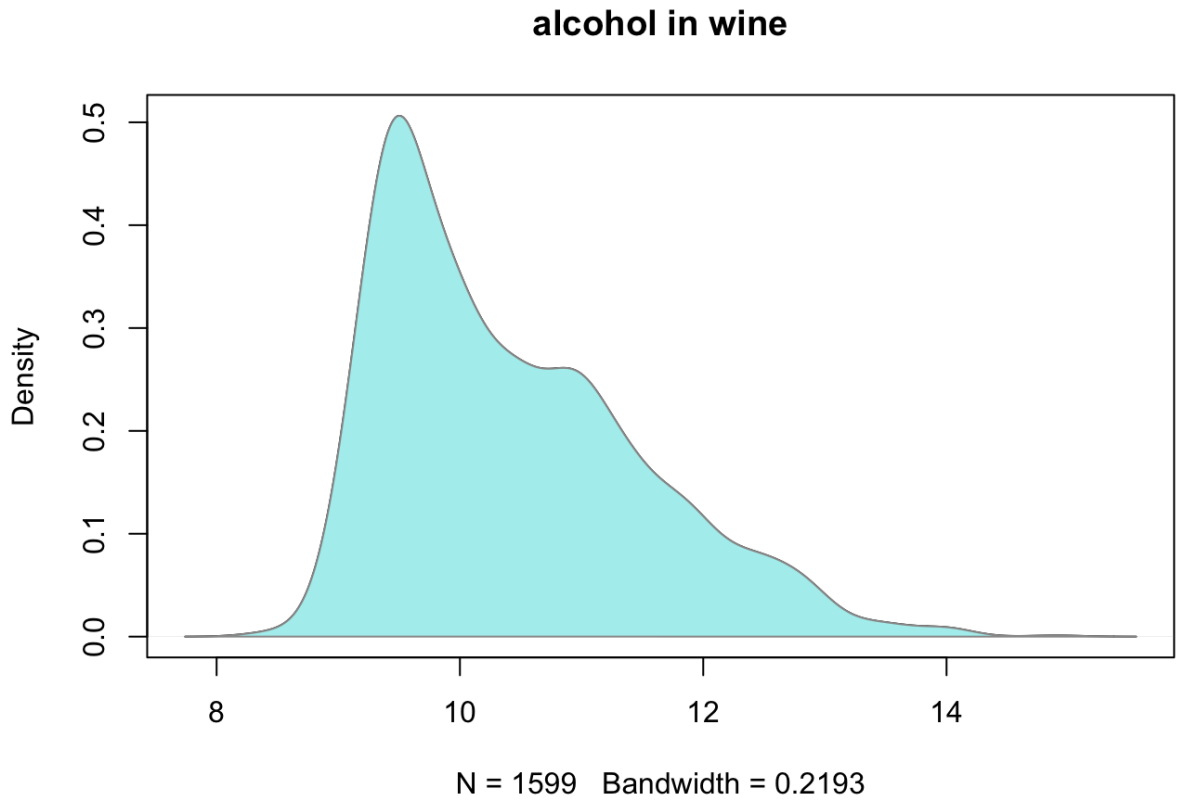

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
## [1] 3.311113
## [1] 0.1543865
##  10%  50%  90%
## 3.12 3.31 3.51
```

DESCRIPTION: Again very close to a normal distribution. The x axis extends from approximately 2.5 to 4.0 and the y axis from 0.0 to 2.5. The peak at 3.3 extends to just above the limits of the y axis and tapers off to about 2.8 and 3.8 with a subtle left skew.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
## [1] 0.6581488
## [1] 0.169507
##  10%  50%  90%
## 0.50 0.62 0.85
```

DESCRIPTION: The x axis extends from 0 to 2.0 and y axis from 0 to 3.5. This normal distribution peaks around .6 with a sharp drop off tapering off at 0 and 1.0 with a bit more action beyond 1.0. A slight bump at .8 skews this distribution slightly right.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
## [1] 10.42298
## [1] 1.065668
##  10%  50%  90%
##  9.3 10.2 12.0
```

DESCRIPTION: An x axis that extends from 8 to 14 and a y axis from 0 to .5. This distribution peaks to the limits of the y axis at approximately 9 and tapers off quickly to the left side by about 9. We've got a right skew with a secondary peak that boarders on being bimodal at 11 and reaches to approximately .25 on the y axis. This peak then tapers off to 14.

Univariate Analysis

What is the structure of your dataset?

I have a list of 1599 unnamed wines. They have all been graded on quality and 11 variables of each wine have been given. Each value given is numeric other than quality and X which are integers. The rating scale is from 1-10 but only 3-8 has been used. A majority of the wines have been rated either 5 or 6.

What is/are the main feature(s) of interest in your dataset?

The main feature I'm concerned with is the quality. I hope to find out what features can most accurately predict the quality of a wine.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

I'd like alcohol to be in the model because it's the most relatable to other people. My guess is that whatever has the strongest effect on taste will be in the final model. I've read some research that suggests wine qualities are actually being mostly judged by price, location it was made, and atmosphere but that doesn't help at all right now.

Did you create any new variables from existing variables in the dataset?

I relabled all the columns in a shorthand version to make coding quicker for me.

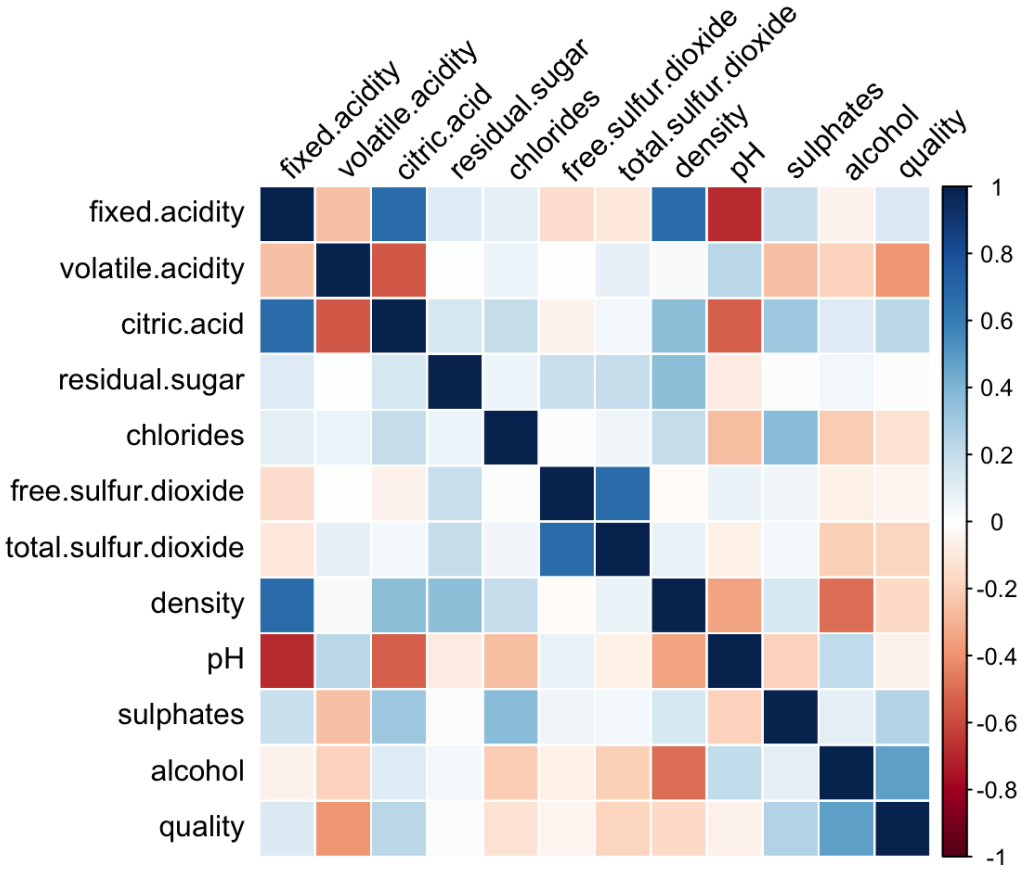
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Volatile.acidity was bimodal. Citric acid freaks me out. It's trimodal and I can tell already I don't like it. Density is the closest to a normal curve. Everything else is close to normal with either a left or right skew. I did not make any major adjustments to the structure of the data.

Bivariate Plots Section

We'll start with a birds eye view of all the associations. I'll use a correlation matrix to get an idea of what we're looking at.

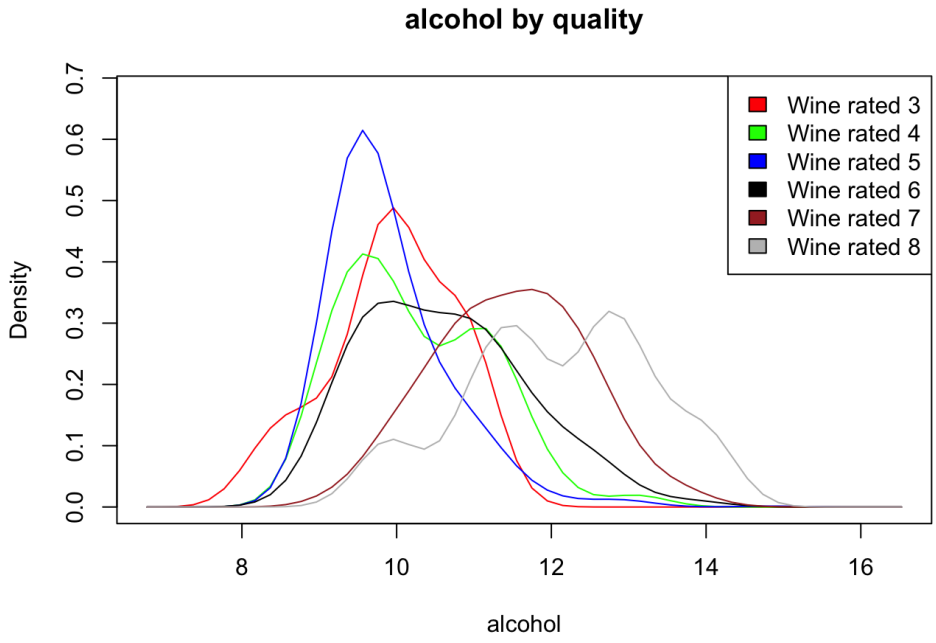
Correlation coefficient matrix of the variables in red wine.

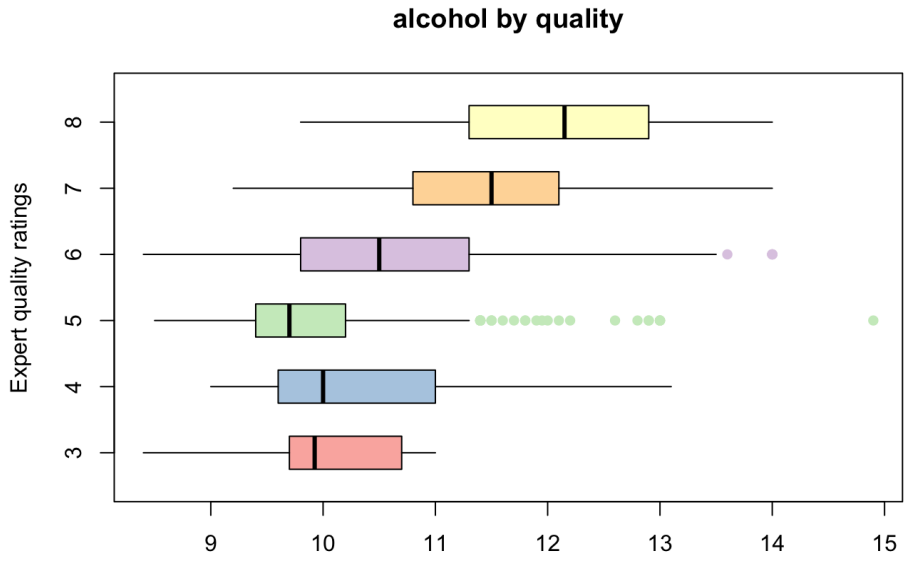


DESCRIPTION: This correlation matrix is a 12X12 matrix with each square representing the calculated value of the correlation coefficient between the 2 intersecting variables. It's gradient is measured from 1 to -1 colored from dark blue to dark red respectively. These limits fade to white as the correlation approaches zero. We can match the color of a square to its corresponding place on the legend to understand the approximate correlation of the variables in question.

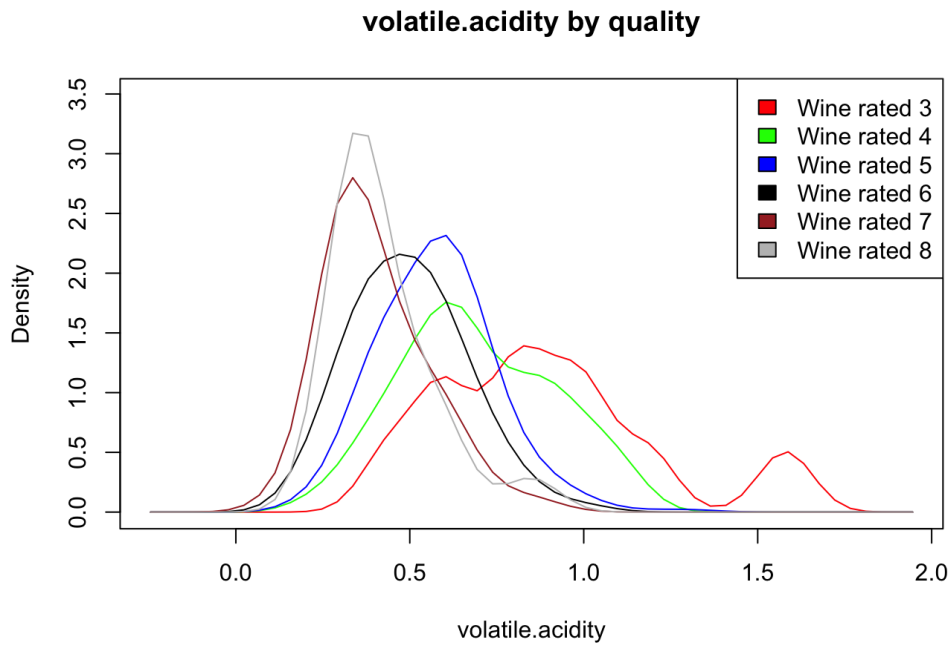
It looks like the strongest correlations to quality are with alcohol, sulphates, volatile.acidity, and maybe citric.acid.

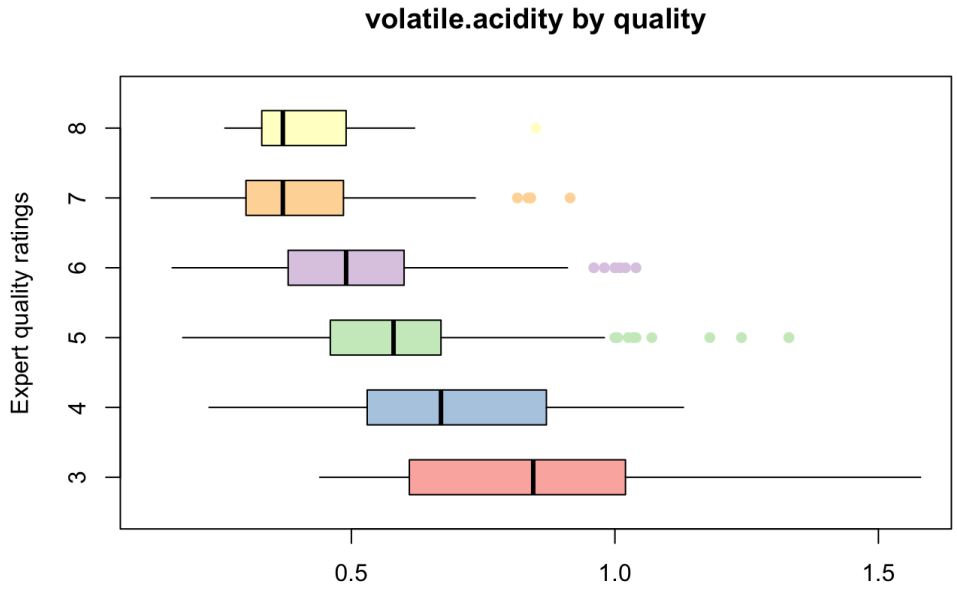
The strongest correlations on the table have nothing to do with quality and are between ingredients like pH/fixed.acidity, citric.acid/volatile.acidity, and total.sulfur.dioxide/free.sulfur.dioxide.



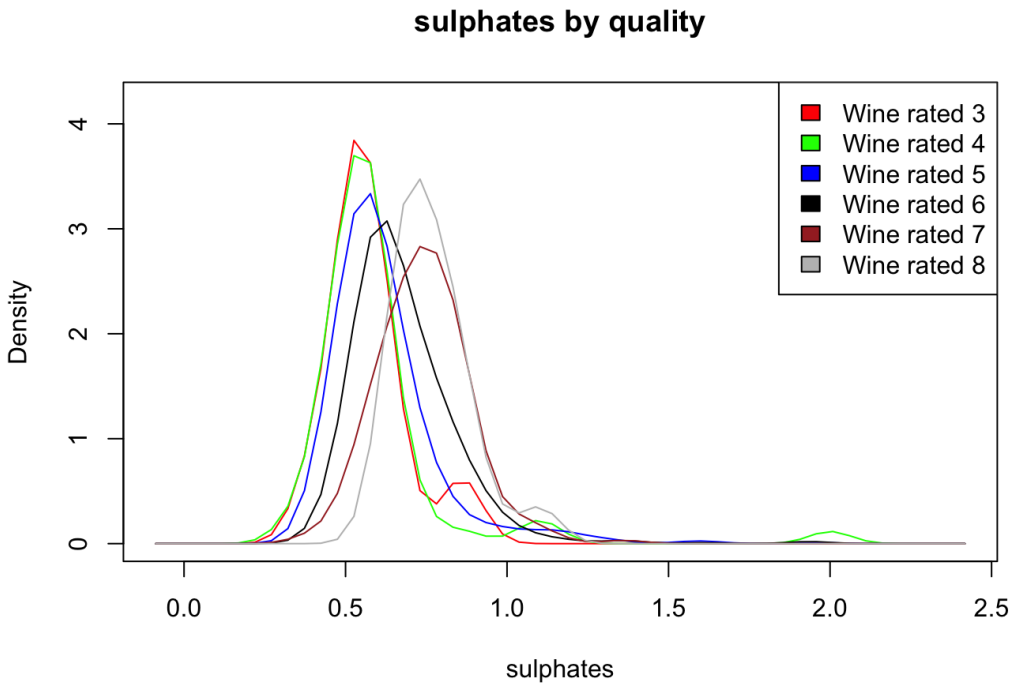


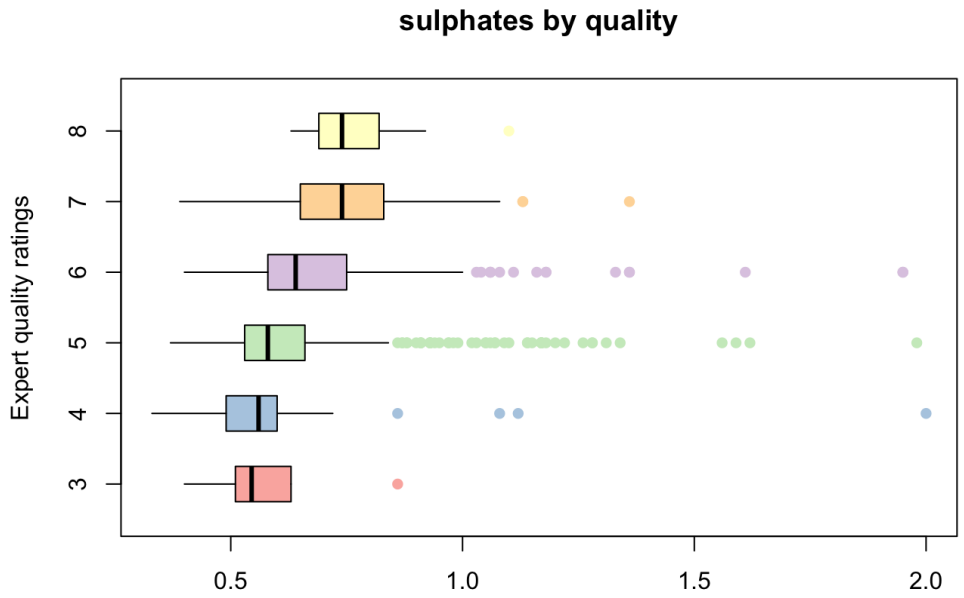
DESCRIPTION: The mean values alcohol percentage in relation to quality increase as the quality goes up. The worst wines average about 10% alcohol while the best average just above 12% alcohol. The only deviation from this trend is wines rated a 5 which tend to be lower than expected coming in around 9.7% alcohol.



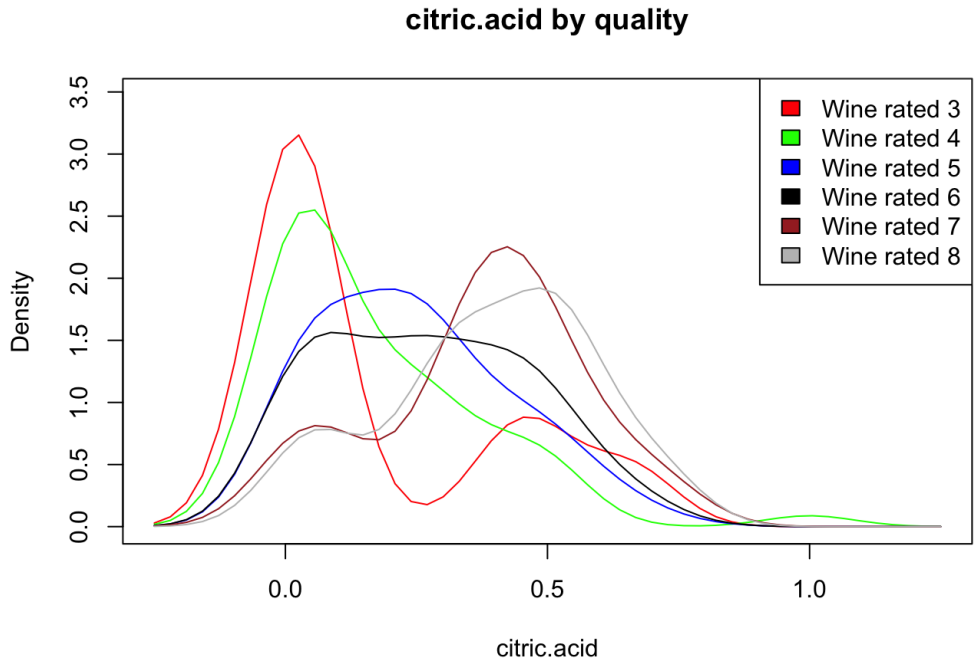


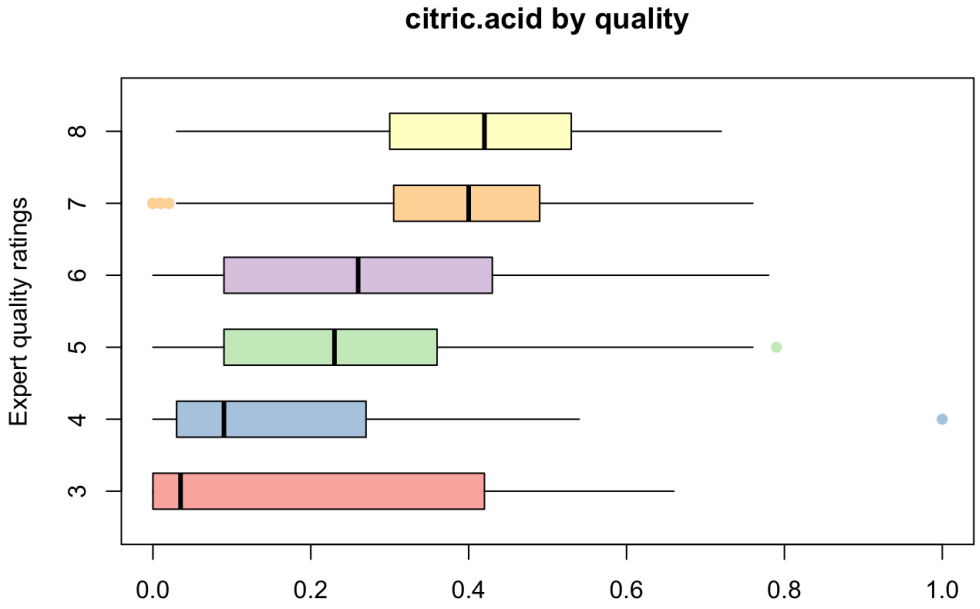
DESCRIPTION: The mean values of volatile.acidity tend to decrease as the quality goes up. This is opposite to what we just discovered about alcohol. The best wines come in at about .4 while the worst wines reach about .8. There is no deviation from this trend like alcohol had.





DESCRIPTION: The correlation matrix pins this as the next most likely candidate for the model but I am not seeing as strong of a correlation as I would hope for. As the quality of wine increases, the so do the sulphates.





DESCRIPTION: The mean value of citric acid tends to increase as the quality goes up. Something different about this trend is that rather than increasing by rating it appears to increase in sets of bad wine (3-4), average wine (5,6) and good wine (7,8). It could just as easily be represented with 3 bars since these groupings are so similar to one

another. Definite connection here but to avoid multicollinearity I'll keep this out of my model because of its strong correlation with volatile.acidity.

That's about all that seems worth looking into for making a model.

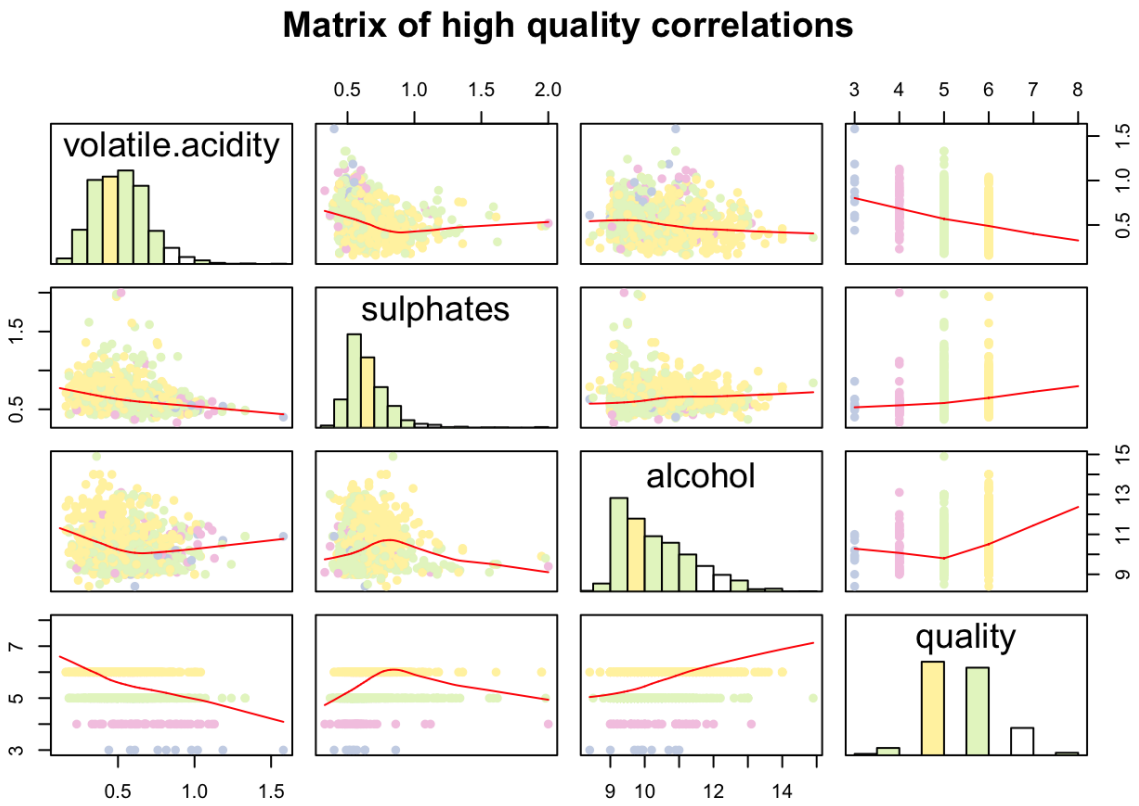
It looks like the variables with the highest correlation to quality are the ones whos distributions change the most consistently between quality ratings. I guess thats pretty obvious but now it's been confirmed visually.

Picking the model

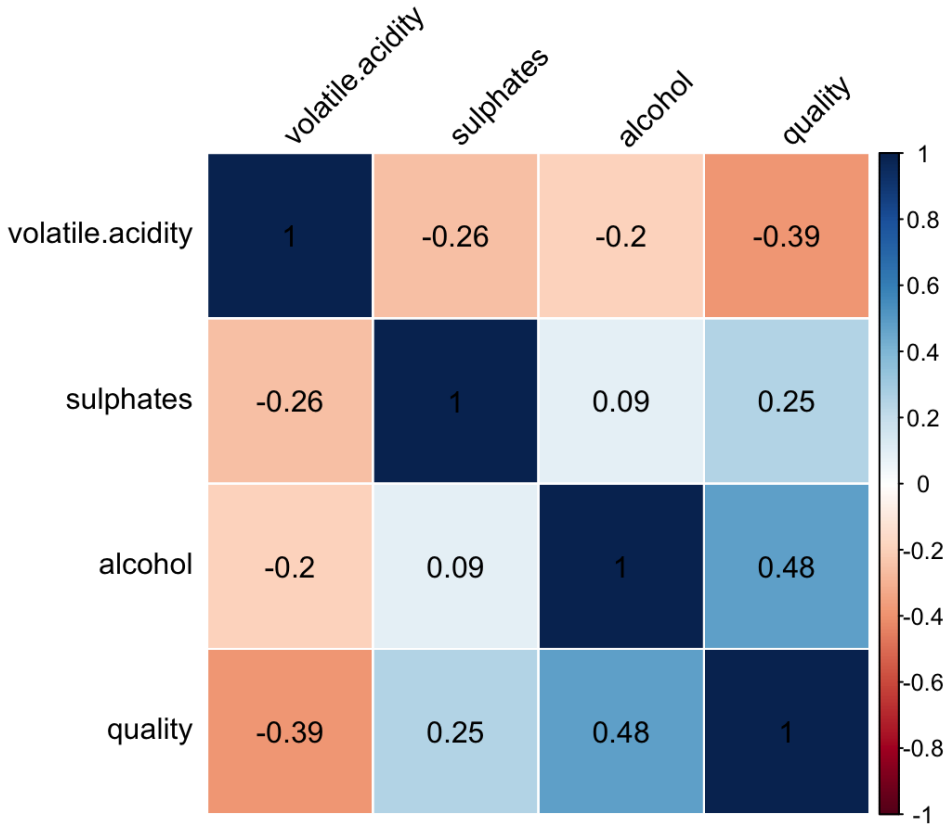
I'll take the variables that are most correlated to the quality of wine so I can start to build my model.

Based on what I learned previously I'm going to pick alcohol, sulphates, and volatile.acidity as possible candidates for my model. Let's investigate further.

```
## [1] "volatile.acidity" "sulphates"        "alcohol"
## [4] "quality"
```



DESCRIPTION: This matrix has reduced my original matrix to the variables I have found to be most relevant to a potential model; volatile.acidity, sulphates, alcohol, and quality. The diagonal is occupied by histograms with the rest of the matrices showing scatterplots indicating the relationship between the corresponding variables. The red lines indicate the overall trend of the datasets and help visualize where the correlations are coming from. I used a pastel color palete to highlight the points in terms of quality to get a general idea of where the different qualities lie in relation to these variables. This will be expanded on in the multivariate plotting section.



This is the simple matrix reduced to the same variables with the correlation coeffients included within the squares so I can see their exact values.

We're down to the final four of matrix maddness. Hopefully your bracket didn't break when sulphates pulled the upset over citric.acid. Volatile acidity, alcohol, and sulphates all go to the next round.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Quality didn't actually have the strongest associations in the list but I'll need to take what I can get. Part of the lack of correlations might have to do with the dataset being small and not well distributed. The strong correlations that do exist are between chemicals that I would expect to be highly correlated with each other. pH, fixed acidity,

and citric.acid are all strongly correlated, free.sulfur.dioxide and total.sulfur.dioxide have a strong positive correlation.

I don't understand why the ingredients I ended up are the most related to wine quality. Alcohol makes sense and I had thought that from the beginning but its strange to me that sulphates, citric acid, or volatile acidity would be particularly sensitive to changing the quality of a wine any more than the other variables do.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Features I wouldn't have expected to have strong relationships are pH and Density, alcohol and density, and fixed acidity and density. Basically I don't think I understand density because alot of things seem to be correlated to it. I wish my problem was with how many things are well correlated with quality because I don't care about density. I wouldn't call their relationships interesting but I guess I'm just jealous. Maybe all of these relationships have something to do with why it was my most normal curve in the univariate section.

What was the strongest relationship you found?

The strongest relationship I found is between pH and fixed.acidity at -.68. Free.sulfur.dioxide and total.sulfur.dioxide cam in at a close 2nd with .67.

The strongest relationship to quality was alcohol at .48

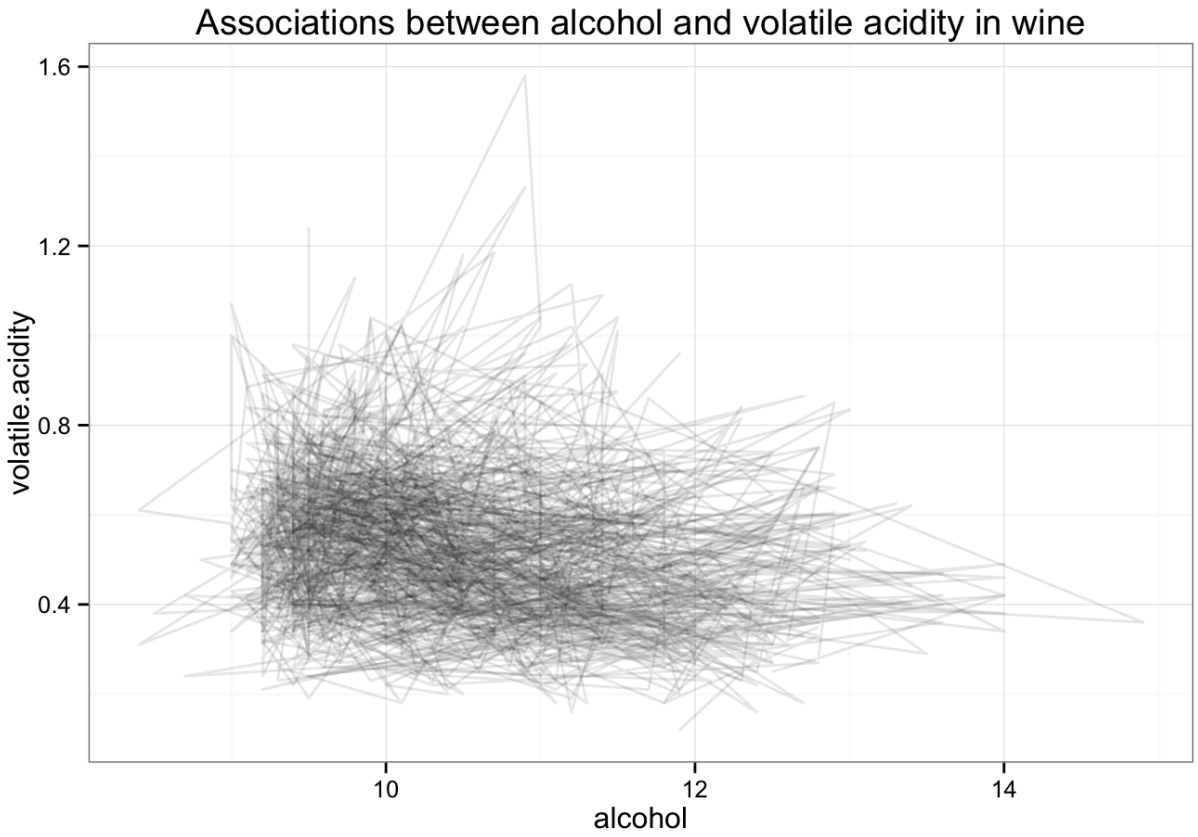
Extra information about my model

I don't like that sulphates are more related to volatile acidity than they are to quality. This is the same problem I had with citric.acid. I'll kick out the weaker of the 2 which is sulphates. This leaves my model with alcohol and volatile.acidity to be used to predict the values of quality.

Multivariate Plots Section

I'll need to explore how the relationships between variables are related to quality.

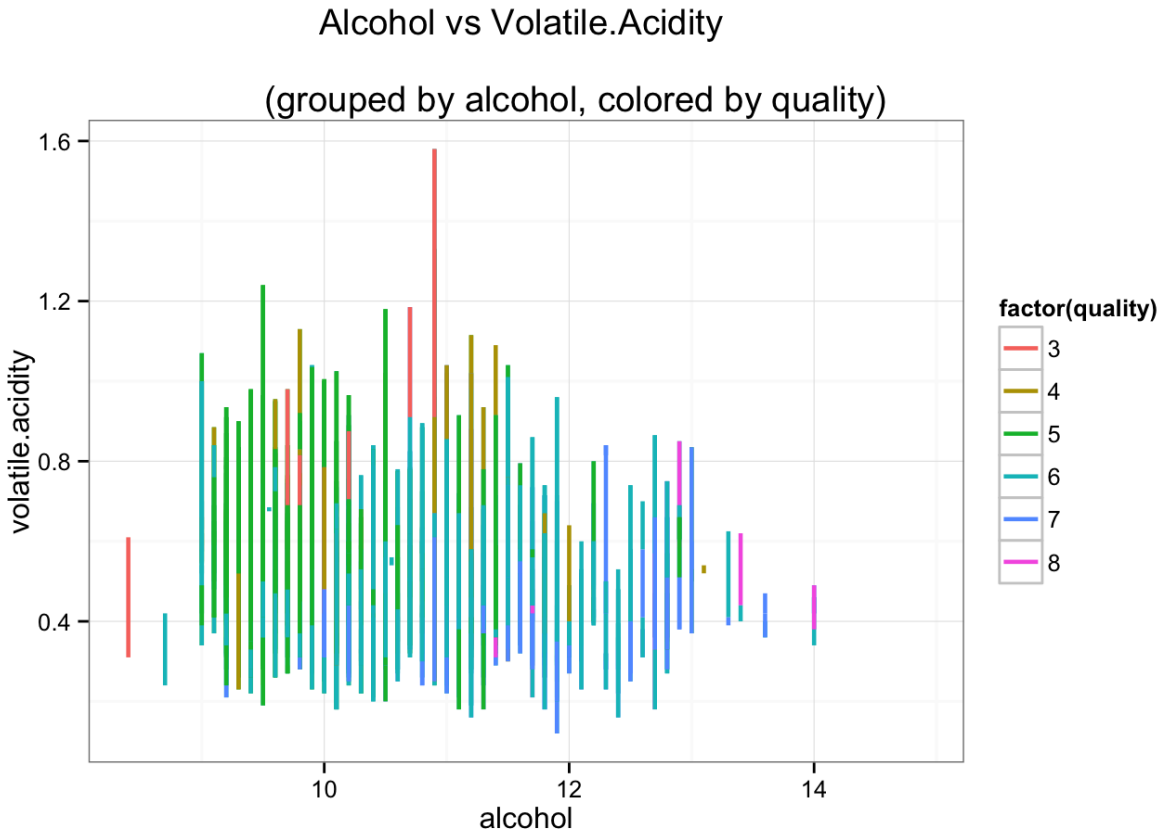
Alcohol and volatile acidity by quality



DESCRIPTION: This is a pretty high level overview of the interactions between volatile acidity, alcohol, and wine quality. I like that it feels as though the computer can visualize these values as a landscape and tried to shade the values with a pencil for me to see the basic idea what it's looking at. I can see that a bulk of the association is occurring in the lower left portion of the plot and I get a sense of a right triangle whose vertex lies at x = 7, y = .2.

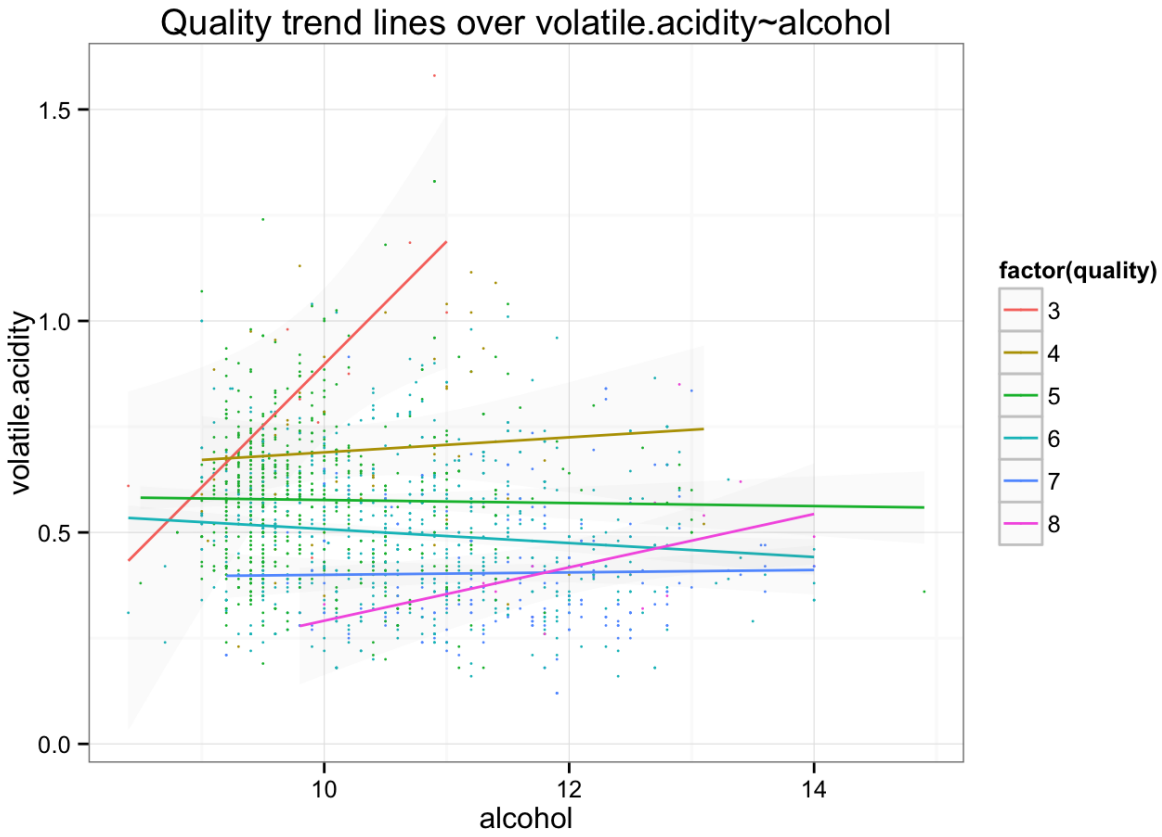
I'll need to work at it to uncover more meaningful insignts.

I'll sort the connections in terms of alcohol content and color them by quality.



I can see where that triangle came from now. In general the association between alcohol has a negative slope and the bulk of the action starts around the vertex of that original triangle. I can see overall that the reds, yellows, and greens (lower quality wines) tend to be in the upper left half while the aquas, blues, and pinks (higher quality wines) occupy the lower right half of the graph. This gives me a general sense that a wine with higher alcohol and lower volatile acidity will be rated higher.

I'd like keep the general idea of this graph but use trend lines to make things a bit clearer.



I changed the lines to a faint scatter plot occupying the background to keep a general sense of where the trend lines came from. It allows me to keep the overall idea of where the distributions are. I tried to leave a suggestion of the standard error while keeping it subtle so the trend lines can remain the focus of the image.

Explore Multivariate Model

After exploring the interaction between alcohol, volatile acidity, and quality I can see I've got something to work with. I've got a series of plots I can use to test a model that I'll be plugging these variables into.

We've got a few especially confusing plots coming up. Here's some definitions you might find useful.

Residuals vs fitted plot

What is a residual? The point of my model is to predict the values of the quality of wine. In my model there is that prediction value (predicted value) and then there is an actual value (observed value) that the wines were rated. Take any point and subtract

the predicted value from the observed value and you get a RESIDUAL. Do that for all the points and you have the RESIDUALS. This mean and sum of all the residuals are 0 so plotting it gives you a horizontal line at y = 0.

What is Fitted? Fitted values are the values of the linear regression model.

Why Residual vs Fitted plot? To spot trends in the difference between observed and predicted values in relation to the linear regression model I built.

Normal Q-Q plot

What is a normal Q-Q plot?

Take your RESIDUAL and divide it by its STANDARD deviation of the residuals and you have a STANDARDized RESIDUAL. This is on the y axis of the normal qq plot.

On the x axis is the theoretical quantiles.

The THEORETICAL QUANTILE basically tells you the distribution of the errors in the linear model that you made.

Put the standardized residuals on a plot with the theoretical quantiles and it should look like a snake trying to climb a stick around about x = y. The better it's doing the happier you can be.

Well get by saying that SCALE-LOCATION is the same as RESIDUALS vs FITTED for model diagnostics. Take the square root of that standardized error from the normal qq and put it in terms of the fitted value from the residual-fitted plot. It's a more confusing way to see if theres a trend in my residuals.

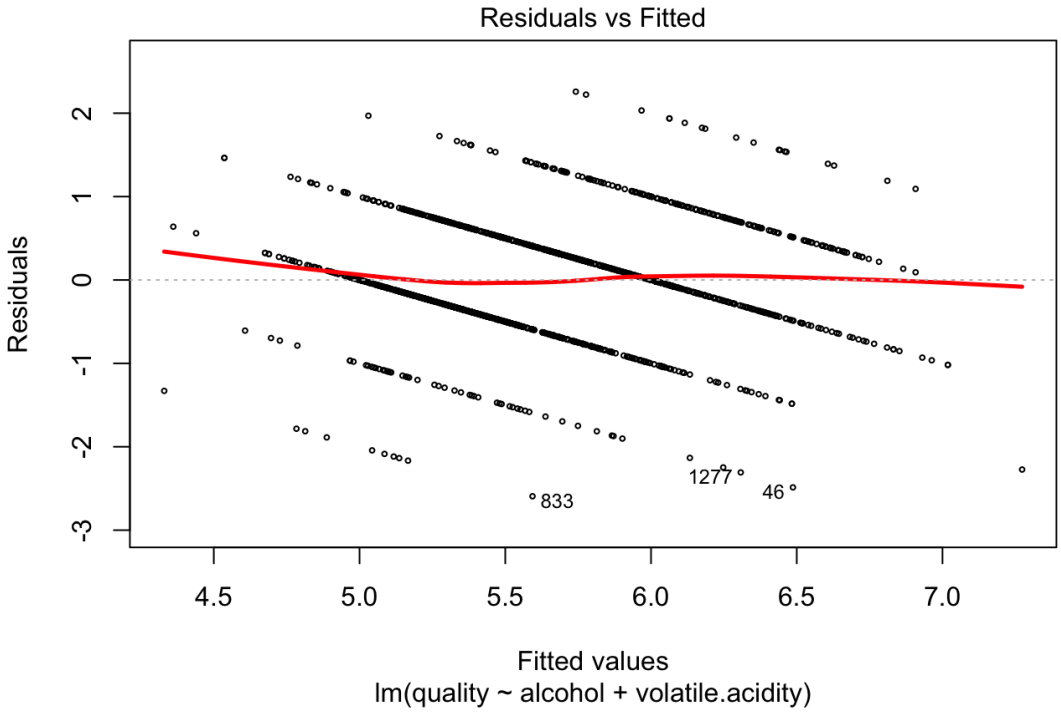
Residuals vs leverage

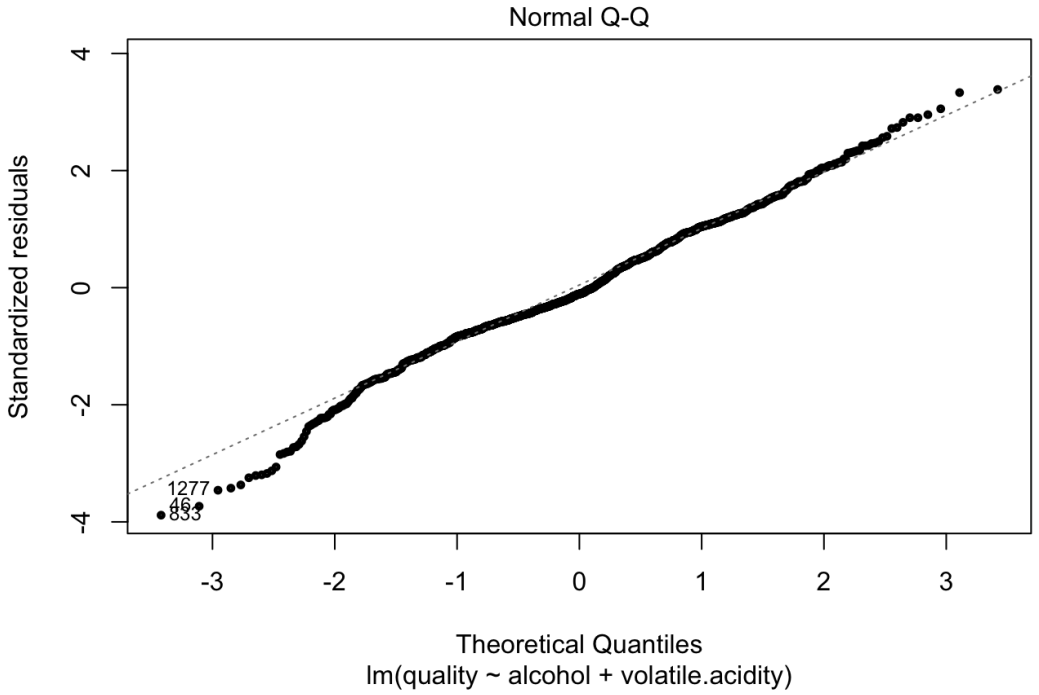
What weight does each datapoint have in terms of effecting the overall distribution of your residuals. If you've ever played the board game Leverage then you know what leverage is here. The weight of a datapoint can cause the residuals to tip in one direction or the other and this plot tells us if and where thats happening. If you've never played the board game leverage you have no chance at ever understanding this plot.

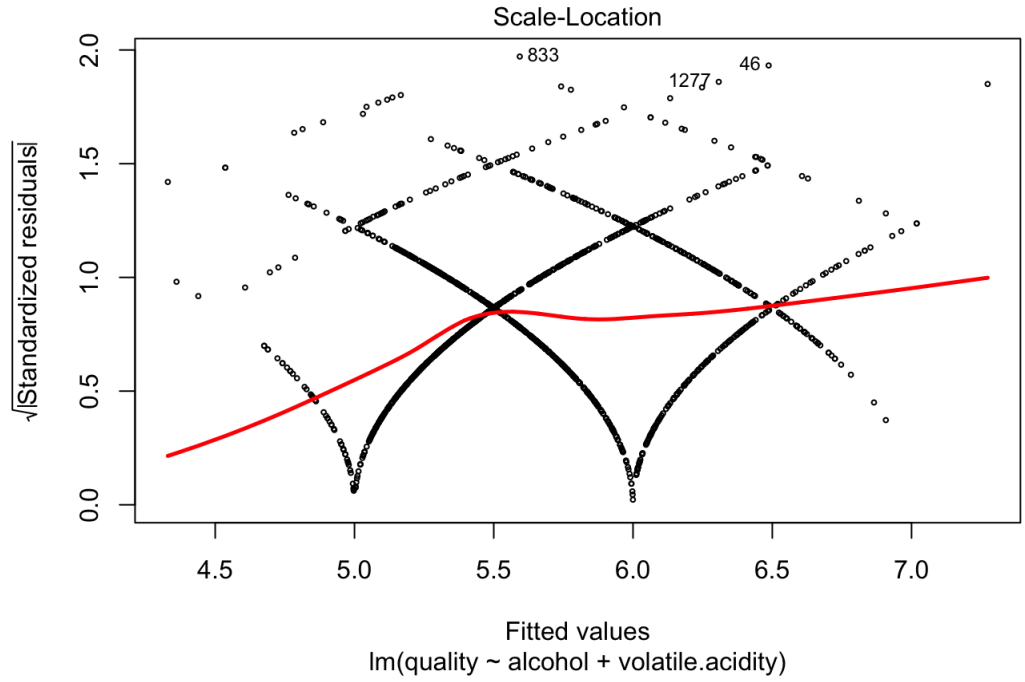
If you only remember one thing before looking at these plots make it this…

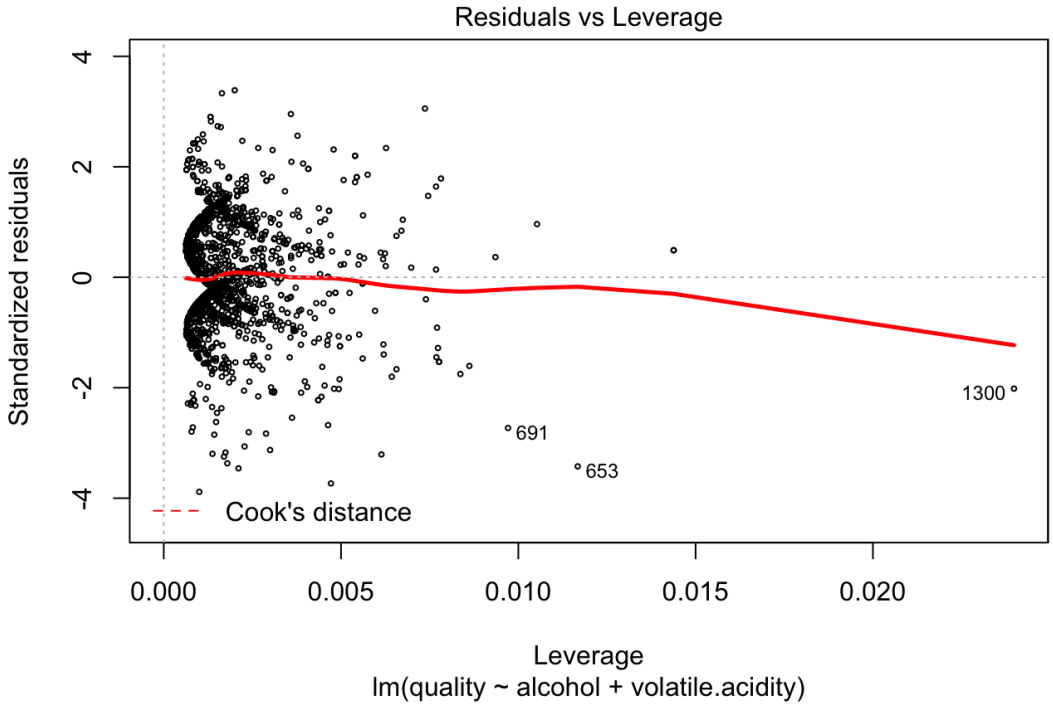
Your happiness depends on how straight these lines are.

Finally, here are the plots.


Residuals vs Fitted
Residuals
Fitted values
lm(quality ~ alcohol + volatile.acidity)
1277
46
833


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(quality ~ alcohol + volatile.acidity)
1277
46
833


Scale-Location
833
1277
46
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(quality ~ alcohol + volatile.acidity)


Residuals vs Leverage
Standardized residuals
4
2
0
-2
-4
1300
691
653
Cook's distance
0.000
0.005
0.010
0.015
0.020
Leverage
lm(quality ~ alcohol + volatile.acidity)

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity, data = w)
##
## Coefficients:
##      (Intercept)          alcohol  volatile.acidity
##           3.0955           0.3138           -1.3836
```

DESCRIPTION Residuals vs Fitted - My residuals vs fitted plot has a generally horizontal trend line with a generally random variation a points surrounding it. This tells me that the distribution of my residuals is generally unbiased and homoscedastic.

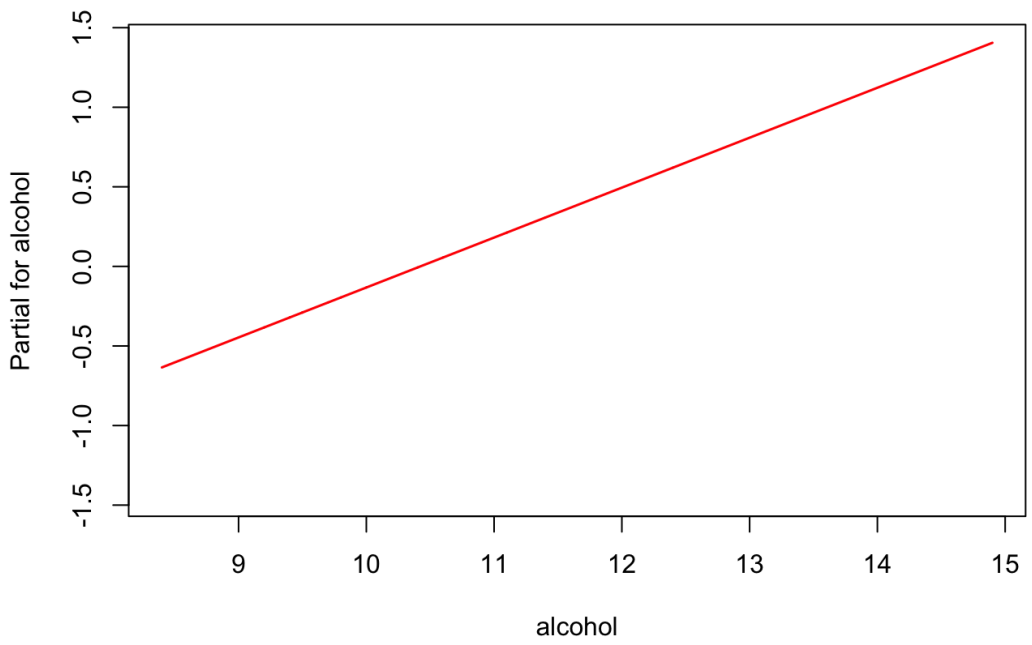
DESCRIPTION Normal QQ - The theoretical quantiles generally follow the standardized residuals which lets me conclude that the errors of my linear model are approximately normally distributed. The lack of variation about the trend line shows me that there are no real concerns of a poor model fit.

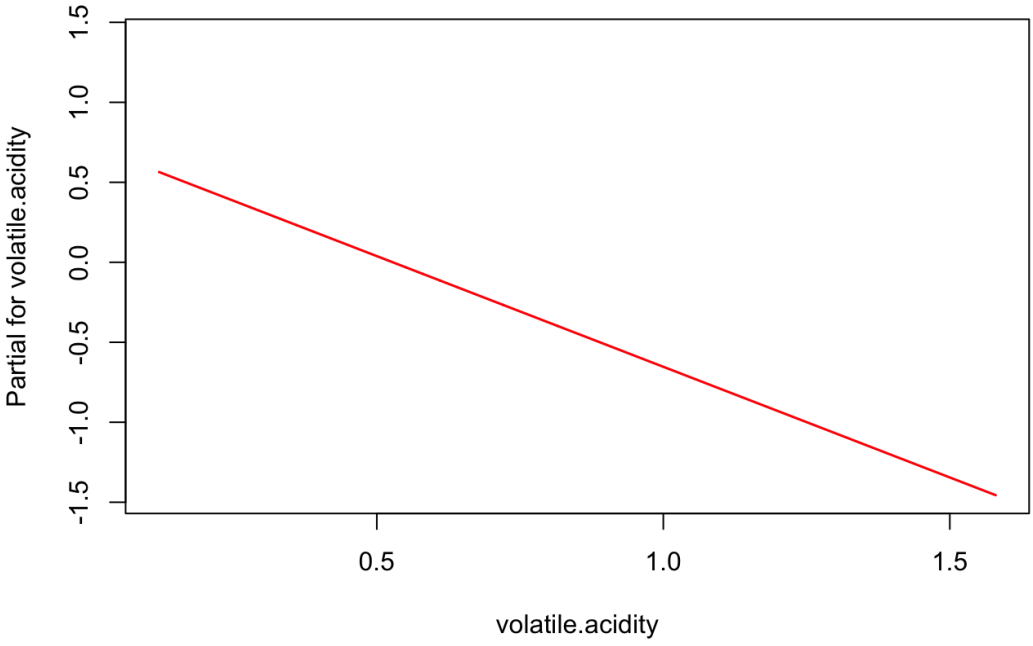
DESCRIPTION Scale-Location - This plot is a variation of the Residuals vs Fitted plot that uses the square root of the standardized residuals instead of the residuals. It also tells me that the distribution of my residuals is generally unbiased and homoscedastic in a more confusing way.

DESCRIPTION Residuals vs Leverage - This plot shows me that there is no major pulling points in the model. Point 1300 on appears to have a drastic effect on the outcome of the model because there are so few points beyond .010.

Termplots

Let’s compute the slope of the regression between our variables and quality. This will show us the effect of each variable on the regression.





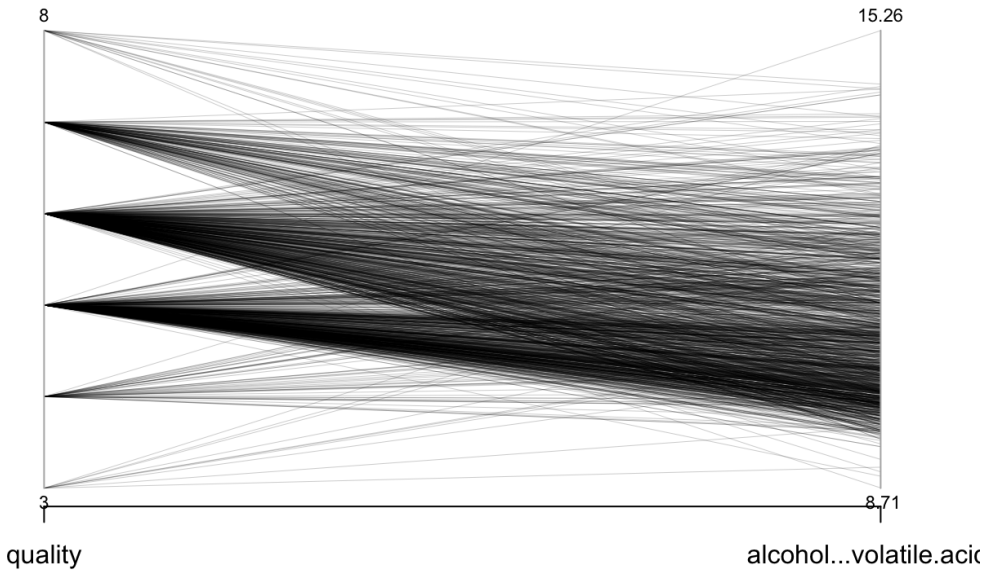
Partial for alcohol - This will tell us the effect alochol has on our linear model. We know there is a positive correlation between alcohol and quality we can expect an

increase in the partial as alcohol percentage increases. We can see that this is the case.

Partial for volatile acidity - This will tell us the effect volatile.acidity has on the linear model we built. Since we know there is a negative correlation between quality and volatile.acidity levels we can expect a decrease in the partial as the volatile acidity levels increase. We can see that this is the case.

Explain par coord

And finally I've got a paired coordinates plot. On the left side is my qualities scaled from 3 to 8. On the right I've got the relationship between alcohol and volatile acidity. Each line is a different wine from the set. The line connects these 2 characteristics. I'm looking to see if each quality appears to have more lines going up or down toward the opposing line. I find it easiest to look at how dark it is around the part of the graph closer to the qualities before they get all tangled up. This will tell you the overall trend of the relationship between alcohol and volatile acidity in terms of quality.



Looks like overall the trend of each quality is for the opposing relationship to be lower rather than higher. This is a result of the correlation of alcohol being positive and of volatile acidity being negative.

My model summary

Here are the final numbers. They're all just pretty much numerical verifications of what I learned from the plots.

```
##
##  Paired t-test
##
## data:  quality and alcohol + volatile.acidity
## t = -208.7944, df = 1598, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.364709 -5.264853
## sample estimates:
## mean of the differences
##               -5.314781

##                     2.5 %     97.5 %
## (Intercept)      2.7335882  3.4573543
## alcohol          0.2824146  0.3452105
## volatile.acidity -1.5704996 -1.1967718

##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity, data = w)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59342 -0.40416 -0.07426  0.46539  2.25809
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.09547    0.18450   16.78   <2e-16 ***
## alcohol           0.31381    0.01601   19.60   <2e-16 ***
## volatile.acidity -1.38364    0.09527  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6678 on 1596 degrees of freedom
```

Multiple R-squared: 0.317, Adjusted R-squared: 0.3161

F-statistic: 370.4 on 2 and 1596 DF, p-value: < 2.2e-16

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

The link between alcohol and volatile acidity is the strongest I can use to determine the quality of wine. I was able to weed out sulphates and citric acid because their strong relationship with volatile.acidity is likely to reduce the effectiveness of my model.

Were there any interesting or surprising interactions between features?

This section mostly was verifying what I had learned and put it in terms of quality. The most important insight of this section was that as volatile.acidity decreases, alcohol content increases, and this makes a better wine.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I made a model that uses the volatile acidity and alcohol content of wine to predict its quality.

You can see its summary below.

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity, data = w)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.59342 -0.40416 -0.07426  0.46539  2.25809
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

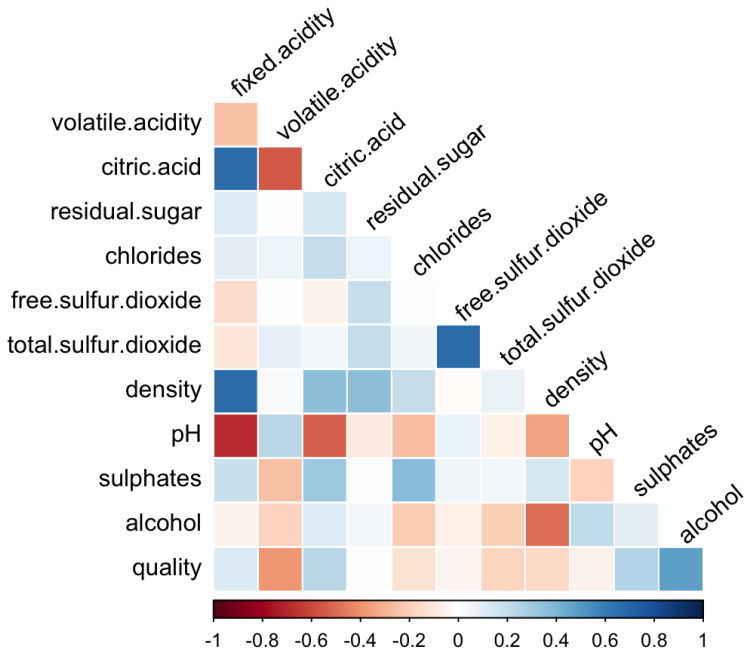
```
## (Intercept)       3.09547    0.18450   16.78   <2e-16 ***
## alcohol           0.31381    0.01601   19.60   <2e-16 ***
## volatile.acidity -1.38364    0.09527  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6678 on 1596 degrees of freedom
## Multiple R-squared:  0.317,  Adjusted R-squared:  0.3161
## F-statistic: 370.4 on 2 and 1596 DF,  p-value: < 2.2e-16
```

I think this model could be improved with a larger dataset. There are small trends I observed in the graphs that could become more prominent with a larger sample size. Having such little variation within the ratings still bothers me and I think is a major reason for not being able to build a better model.

Final Plots and Summary

Plot One

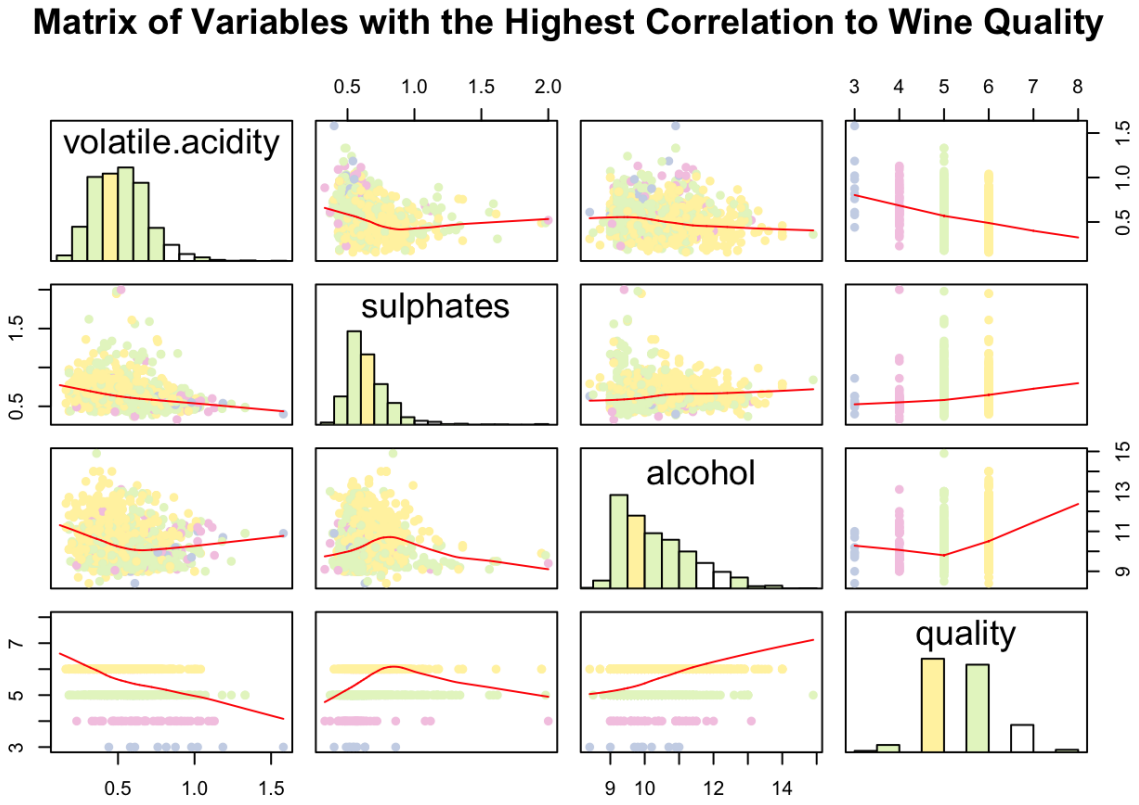
Matrix of Red Wine Variable Correlations



Description One

DESCRIPTION: This correlation matrix is a 12X12 cut off at -x = y with each square representing the calculated value of the correlation coefficient between the 2 intersecting variables. It's gradient is measured from 1 to -1 colored from dark blue to dark red respectively. These limits fade to white as the correlation approaches zero. We can match the color of a square to its corresponding place on the legend to understand the approximate correlation of the variables in question.

Plot Two



Description Two

DESCRIPTION: This matrix has reduced my original matrix to the variables I have found to be most relevant to a potential model; volatile.acidity, sulphates, alcohol, and quality. The diagonal is occupied by histograms with the rest of the matrices showing scatterplots indicating the relationship between the corresponding variables. The red lines indicate the overall trend between the variables and help visualize the correlations.

I used a pastel color palete to highlight the points in terms of quality to get a general idea of where the different qualities lie in relation to these variables. This will be expanded on in the multivariate plotting section.

Plot Three

Description Three

The primary purpose of this graph is to analyze the trend lines of the interactions between alcohol and volatile acidity as they relate to quality. I colored the lines with a corresponding legend to the right side. I tried to put the standard error in the midground by reducing its alpha value. The scatterplot points are not meant to be completely coherent but I'm hoping they can be used to see a general sense of where the trend lines came from while residing in the background.

Reflection

I chose the red wine data because it was easy to understand and left me time to focus on learning R and its associated tools rather than getting tripped up in the data. I wish I had understood how to build functions before I made my whole first draft without them.

This had to be the simplest possible dataset so I don't think I have any real concerns with the challenges of working with this dataset itself. The only troubles I had we're with my own incompetence with R. The course taught me to make graphs and build models but putting them into practice took some creativity.

I know there is plenty of room for improvement especially since completing the project for Intro to Machine learning. I don't think EDA is the time for me to try and complete regression models. In the future I'll use this section for exploring the dataset since that's what exploratory data analysis should be. If I want to make models I'll dedicate time to exploring them through machine learning techniques I've learned so I can be sure I'm picking a good fit for the data I've explored. At this point I feel more comfortable doing this work in Python so I'd do my model building in there with sklearn.

I spent too much time optimizing graphs as I went along. By my final graphs section I basically had made the graphs exactly how I had wanted them to look which should not be the case. Alot of it had to do with not knowing the graphs so I was exploring what they were capable of but in the future I know to get the quickest version for insight and move on because theres alot more to be done as a Data Scientist than mess around with clean data sets.