WHITE_WINE_QUALITY by JOLENE YAO

The white wine data set includes 4898 observations and 13 variables. The variables include wine properties that we will explore.

Univariate Plots Section

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
## 3 3           8.1             0.28        0.40            6.9     0.050
## 4 4           7.2             0.23        0.32            8.5     0.058
## 5 5           7.2             0.23        0.32            8.5     0.058
## 6 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

See third link in references at end of document to read about what each of the variables mean.

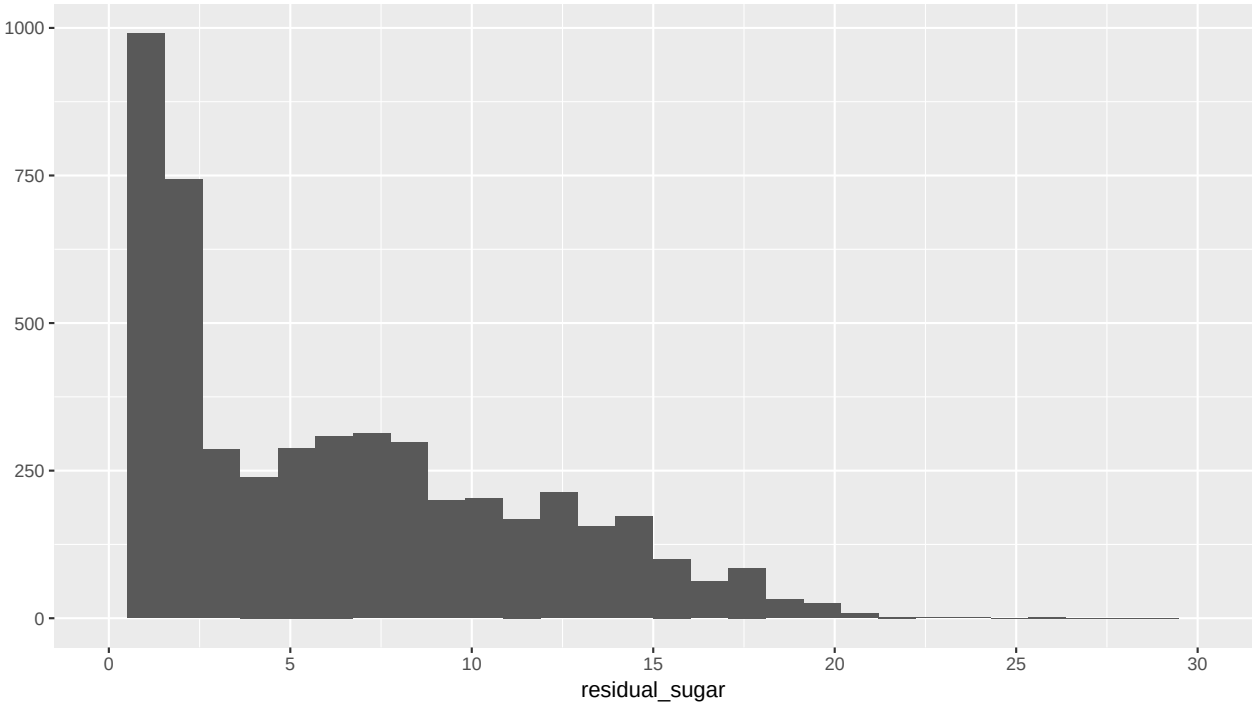
Univariate Analysis

What is the structure of your dataset?

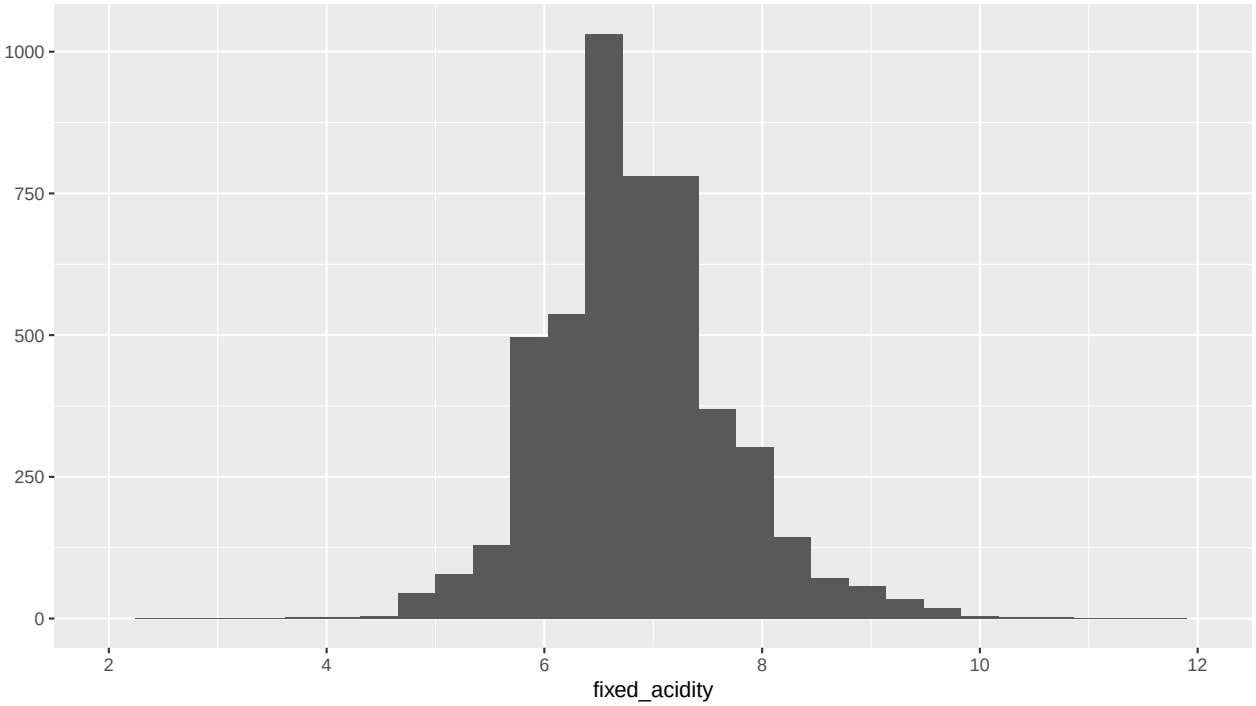
Our dataset includes 4898 observations with 13 different variables.

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

The main features of interest in our dataset includes acidity, sugar, pH, density, and overall quality.

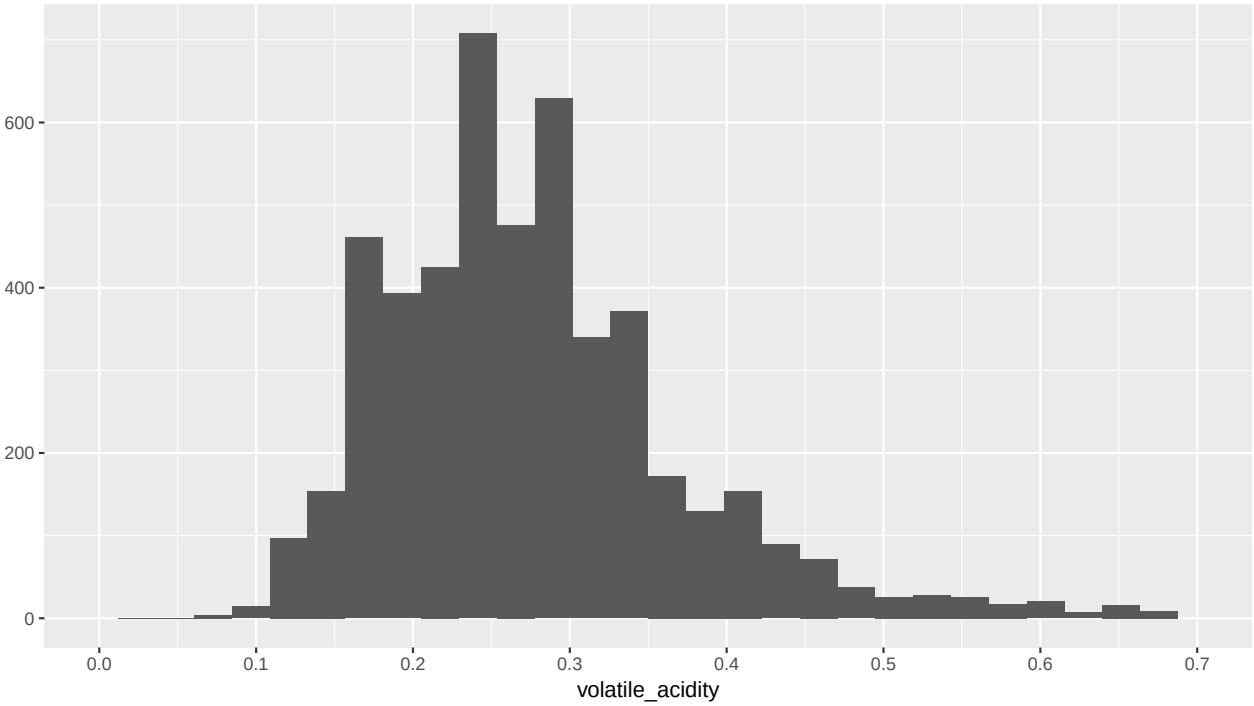


The majority of white wines do not have a high sugar content. They stay at 0-3g/L.

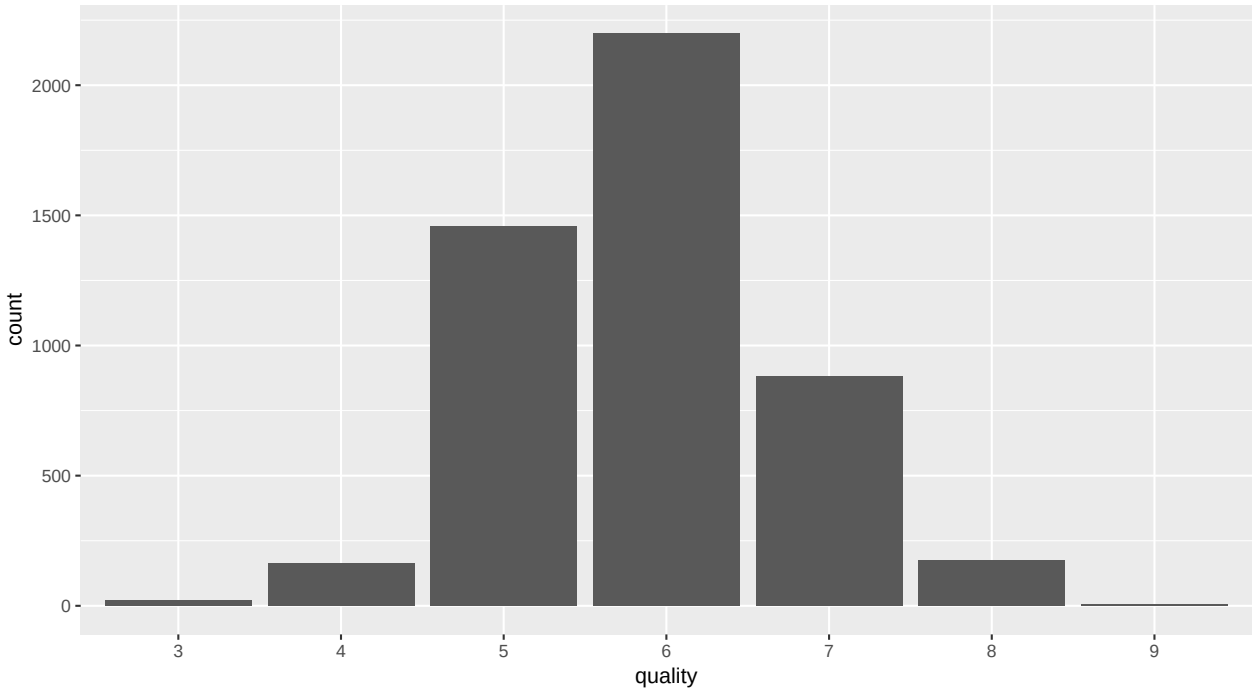


This histogram shows a normal distribution of fixed acidity. This means that most wines will have a fixed acidity between 6 and 8 g/dmˆ3 of tartaric acid - average quality of white wines will have fixed acidity of

6-8g/dmˆ3 .



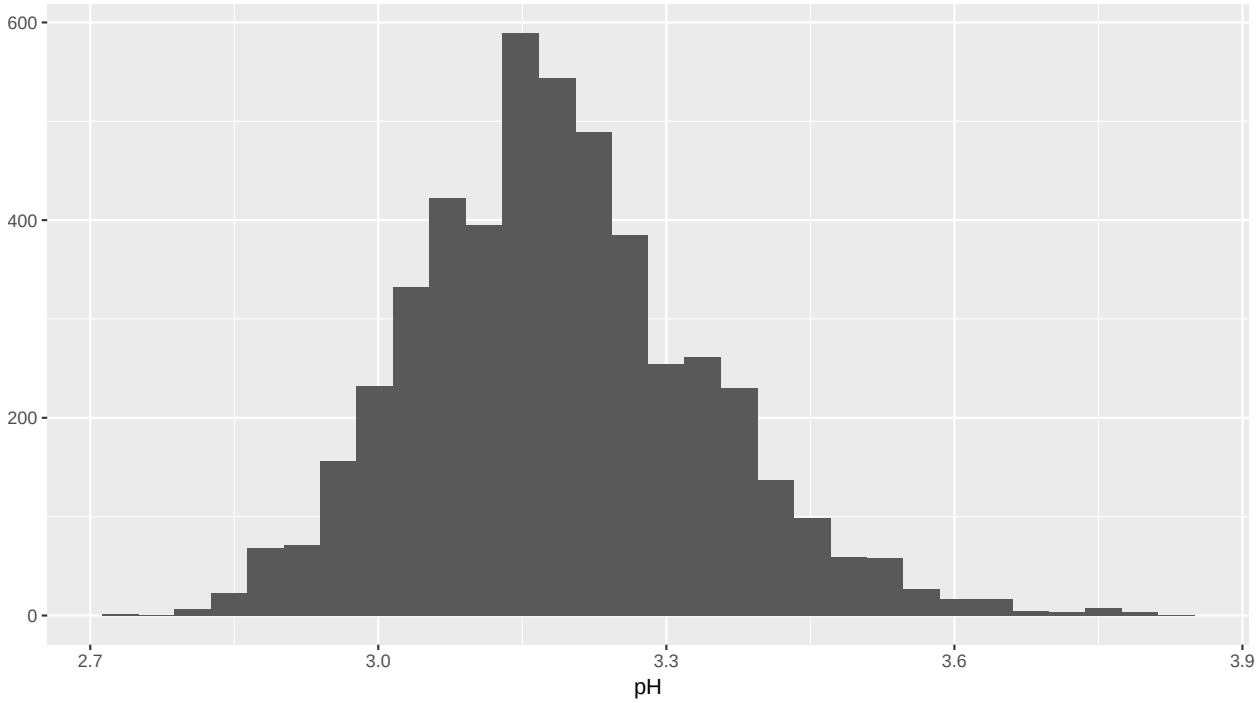
This histogram shows a normal distribution of volatile acidity, meaning that the average quality white wines have a volatile acidity between 0.2-0.3g/dmˆ3 of acetic acid. This is favorable since too high of a volatile acidity will lead to a vinegar aftertaste.




```
##   Var1 Freq
## 1    3   20
```

```
## 2    4  163
## 3    5 1457
## 4    6 2198
## 5    7  880
## 6    8  175
## 7    9    5
```

Most white wines fall at a quality of 6. This follows a normal distribution.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

Most wines have a high acidity with a mean of 3.188.

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

Other variables not mentioned, such as sulphates and alcohol content, will help us determine quality of white wine. See bivariate plots section for analyses of these variables.

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

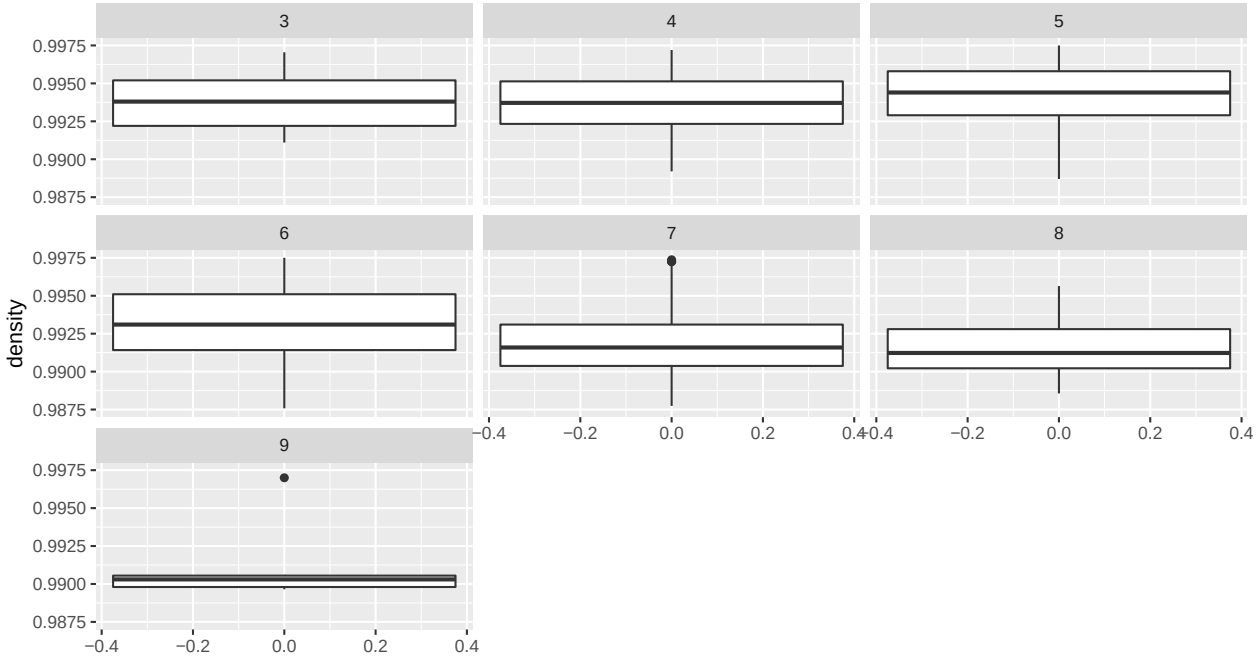
No new variables were created from existing variables.

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?

Yes, I found that as quality increases, sulphate content increases but decreases when it hits a quality rating of 9. This could be due to the very low count of 9 ratings in the dataset. There are only 5 white wines that

have a quality rating of 9.

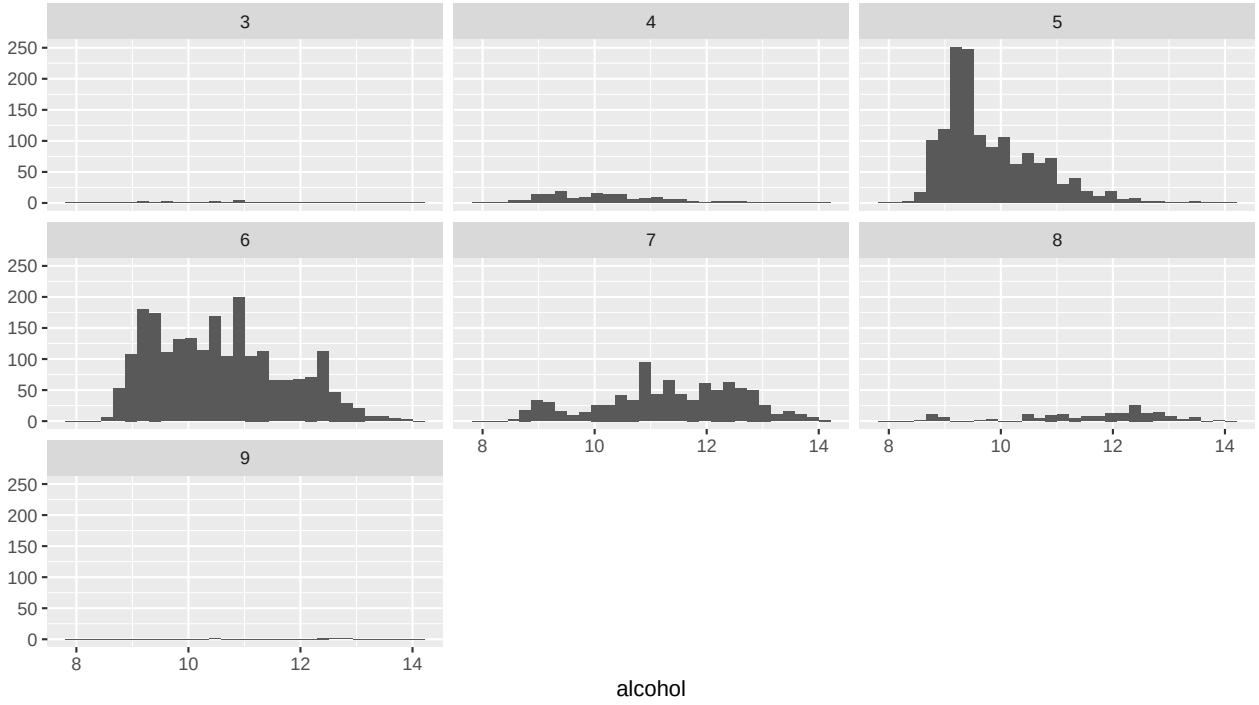
Bivariate Plots Section




```
## wineQualityWhites$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9911  0.9925  0.9944  0.9949  0.9969  1.0001
## ------------------------------------------------------------
## wineQualityWhites$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9892  0.9926  0.9941  0.9943  0.9958  1.0004
## ------------------------------------------------------------
## wineQualityWhites$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9933  0.9953  0.9953  0.9972  1.0024
## ------------------------------------------------------------
## wineQualityWhites$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
## ------------------------------------------------------------
## wineQualityWhites$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9906  0.9918  0.9925  0.9937  1.0004
## ------------------------------------------------------------
## wineQualityWhites$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9903  0.9916  0.9922  0.9935  1.0006
## ------------------------------------------------------------
## wineQualityWhites$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  0.9897  0.9898  0.9903  0.9915  0.9906  0.9970
```

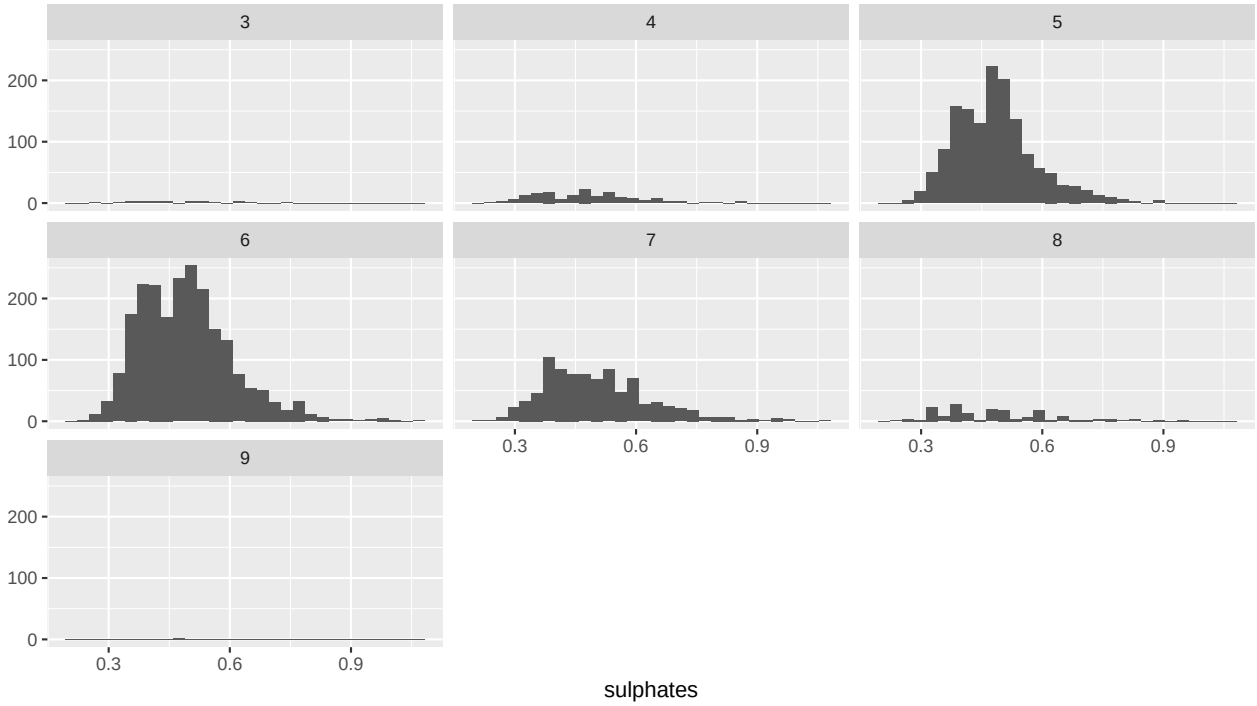
The density of water is about 1g/cm^3 and by looking at the graph, most white wines fall at 0.9940. White wine and water are about the same density. Investigating further, the summary statistics shows that there is a decreasing trend of density as quality increases.




```
## wineQualityWhites$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.55   10.45   10.35   11.00   12.60
## ------------------------------------------------------------
## wineQualityWhites$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.40   10.10   10.15   10.75   13.50
## ------------------------------------------------------------
## wineQualityWhites$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.000   9.200   9.500   9.809  10.300  13.600
## ------------------------------------------------------------
## wineQualityWhites$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## wineQualityWhites$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.60   10.60   11.40   11.37   12.30   14.20
## ------------------------------------------------------------
## wineQualityWhites$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   11.00   12.00   11.64   12.60   14.00
## ------------------------------------------------------------
## wineQualityWhites$quality: 9
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   12.40   12.50   12.18   12.70   12.90
```

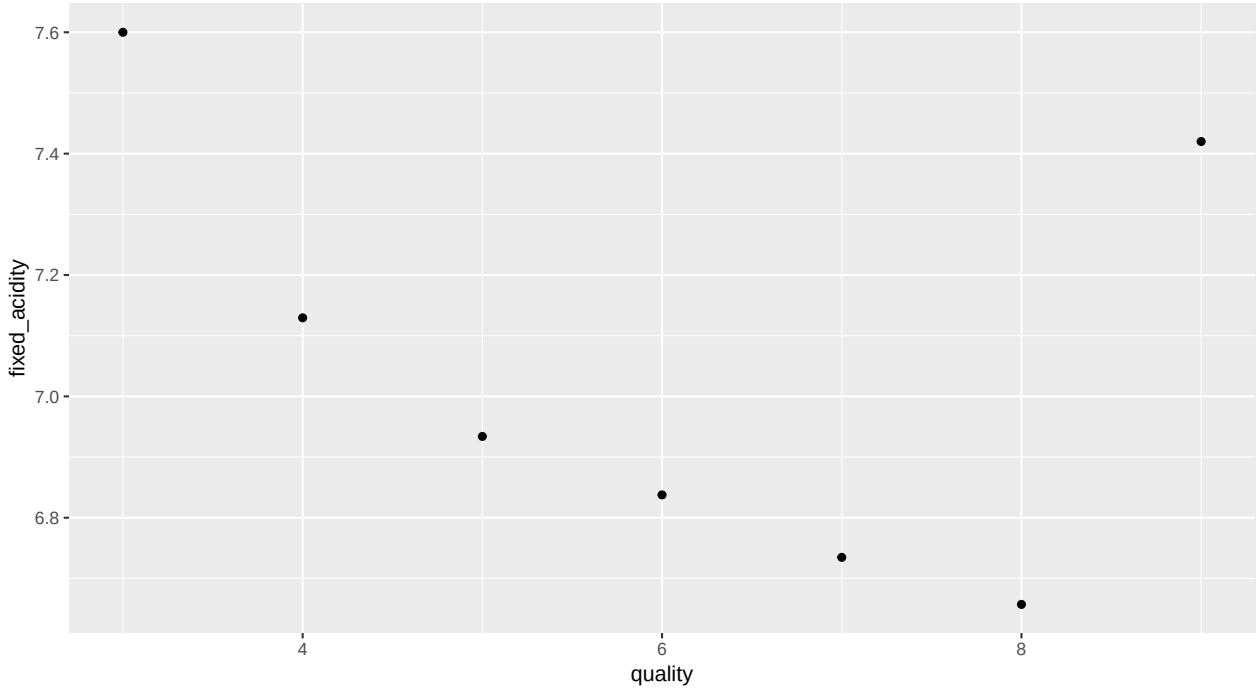
A higher quality wine has a higher alcohol content (% alcohol by volume)



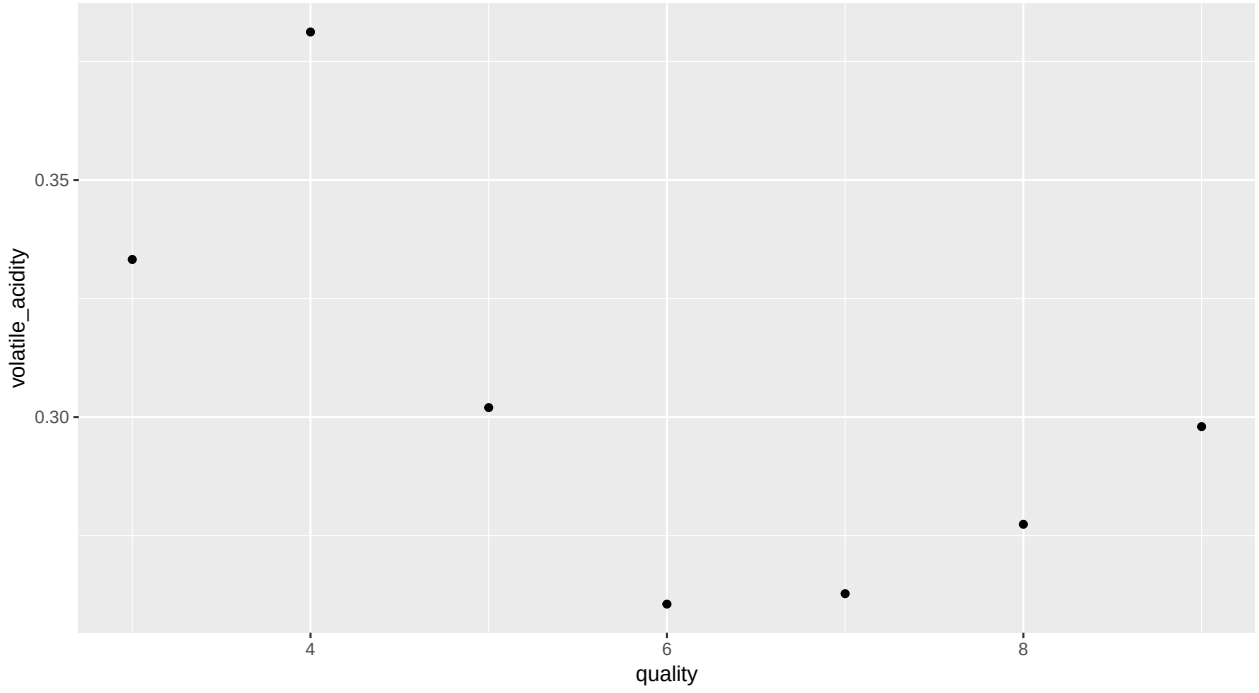

```
## wineQualityWhites$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2800  0.3800  0.4400  0.4745  0.5425  0.7400
## ------------------------------------------------------------
## wineQualityWhites$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2500  0.3800  0.4700  0.4761  0.5400  0.8700
## ------------------------------------------------------------
## wineQualityWhites$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2700  0.4200  0.4700  0.4822  0.5300  0.8800
## ------------------------------------------------------------
## wineQualityWhites$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2300  0.4100  0.4800  0.4911  0.5500  1.0600
## ------------------------------------------------------------
## wineQualityWhites$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4800  0.5031  0.5800  1.0800
## ------------------------------------------------------------
## wineQualityWhites$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2500  0.3800  0.4600  0.4862  0.5850  0.9500
## ------------------------------------------------------------
## wineQualityWhites$quality: 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   0.360   0.420   0.460   0.466   0.480   0.610
```

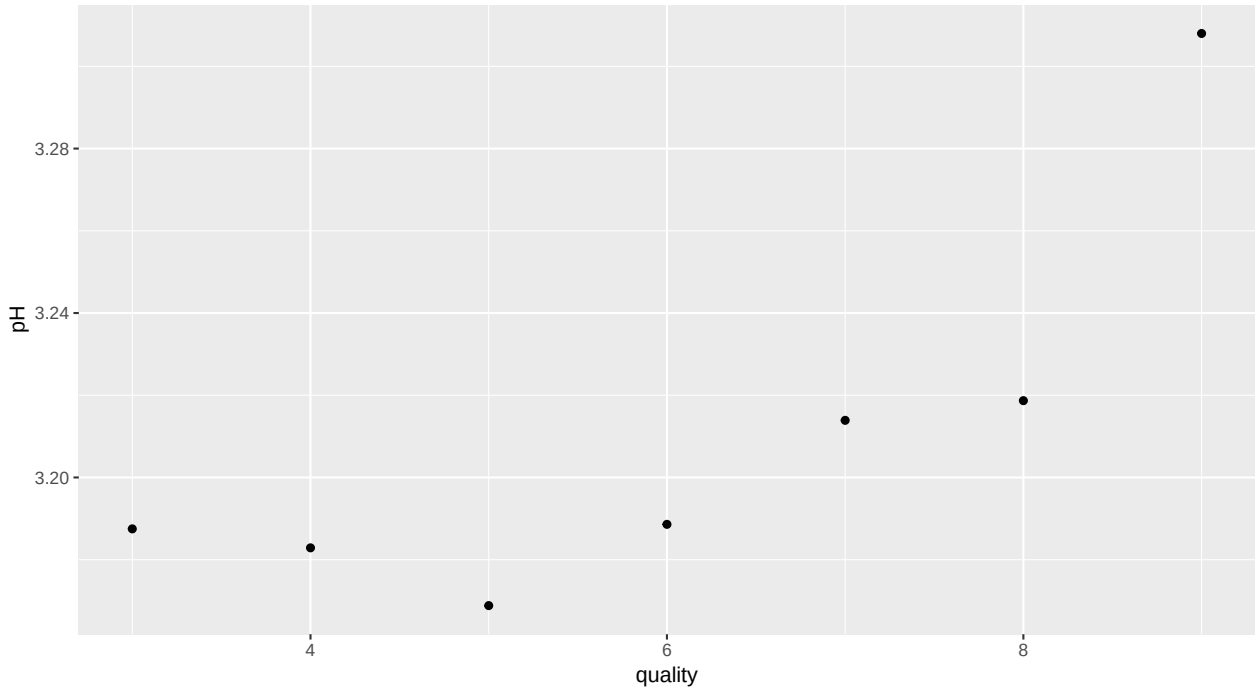
The sulphate content increases as quality increases. However, there is an unexpected decrease in sulphates at 9 quality. This could be due to the very low count of wines at 9 quality rating.



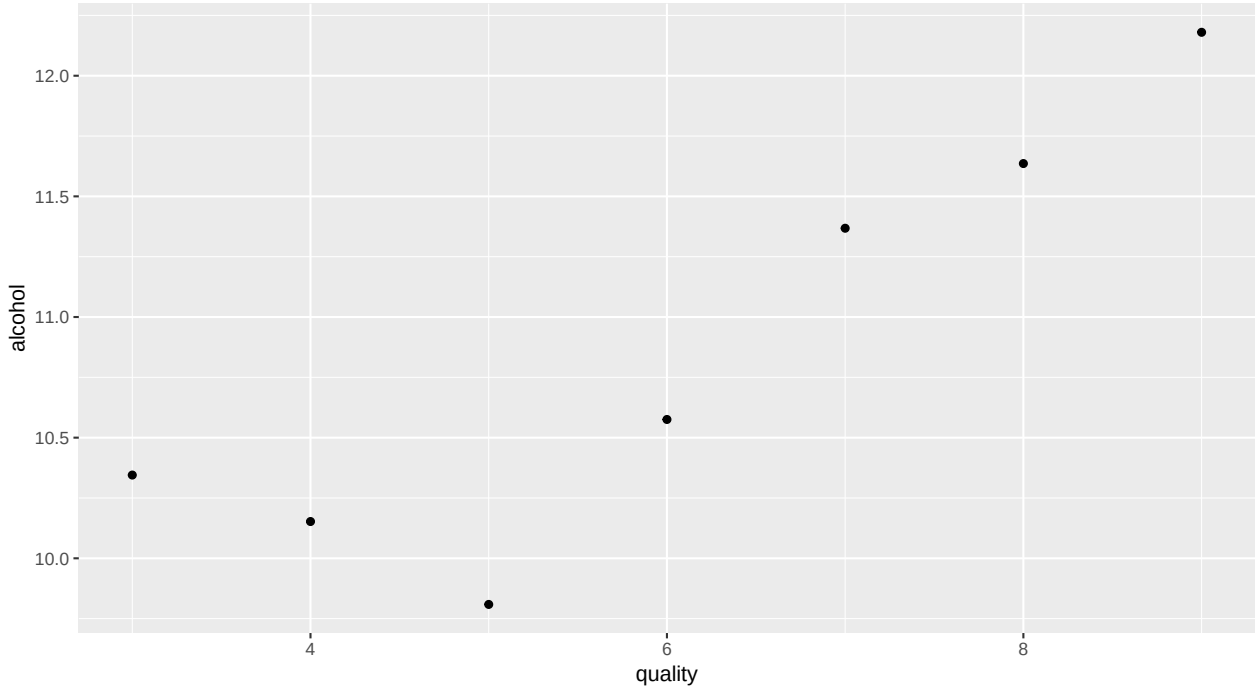
We can see that as quality increases, acidity decreases. With a lower pH, there must be a better quality. The plot uses mean to show the trend line clearly, whether it is increasing or decreasing. In this case, it is clearly decreasing. Each scatterplot following this one will show the means of the variable against quality.



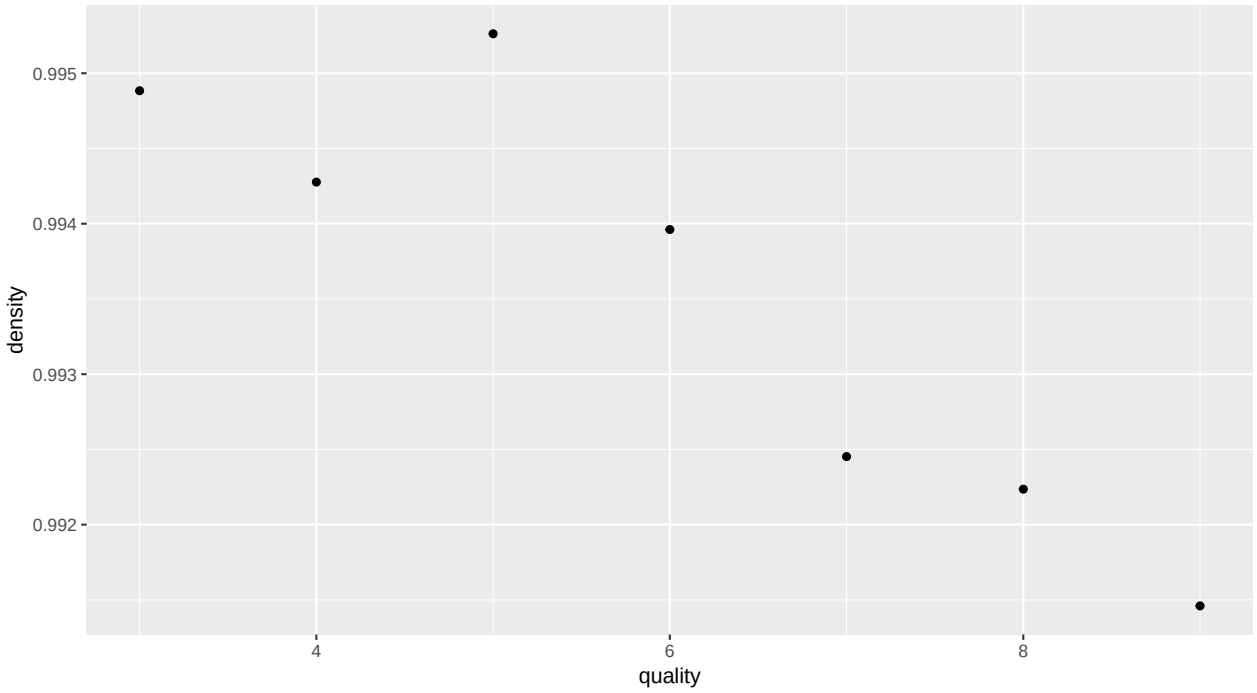
If we ignore the extreme ends of the graph (ignore quality rating of 3 and 9), then the trend is a decreasing volatile acidity will increase the quality of white wines.



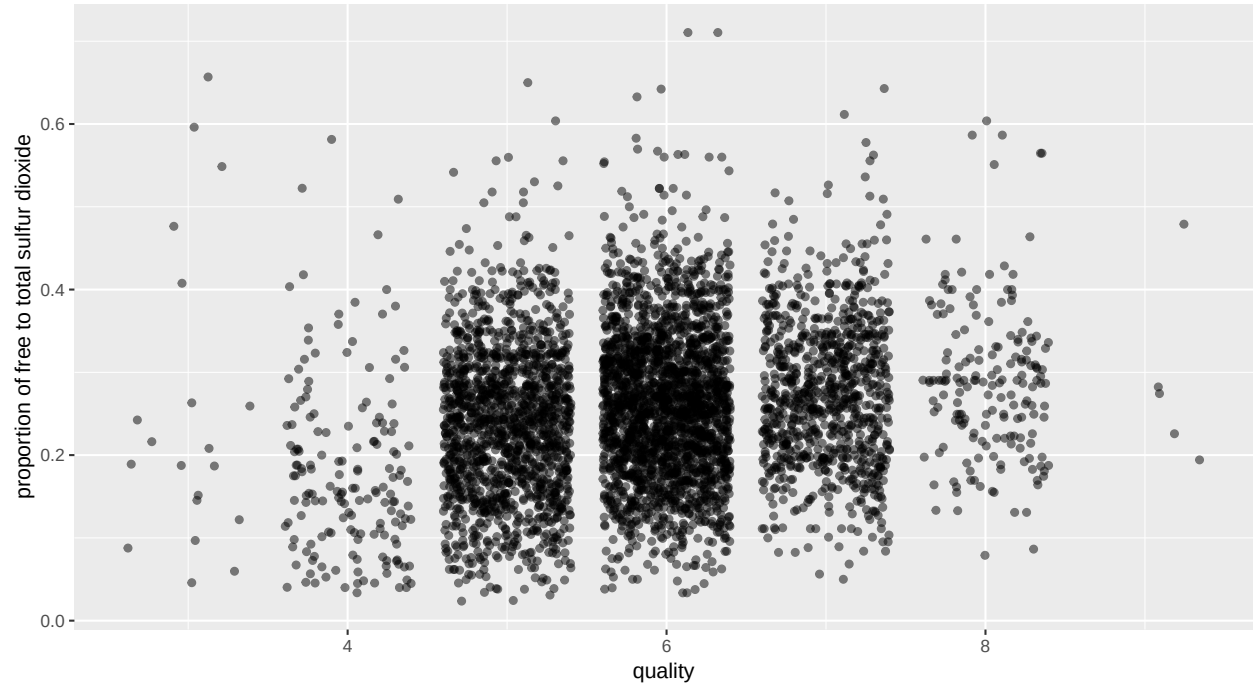
With higher quality wines, there will be a higher acidity. Based on the previous plots, we saw that most wines will have a higher fixed_acidity.



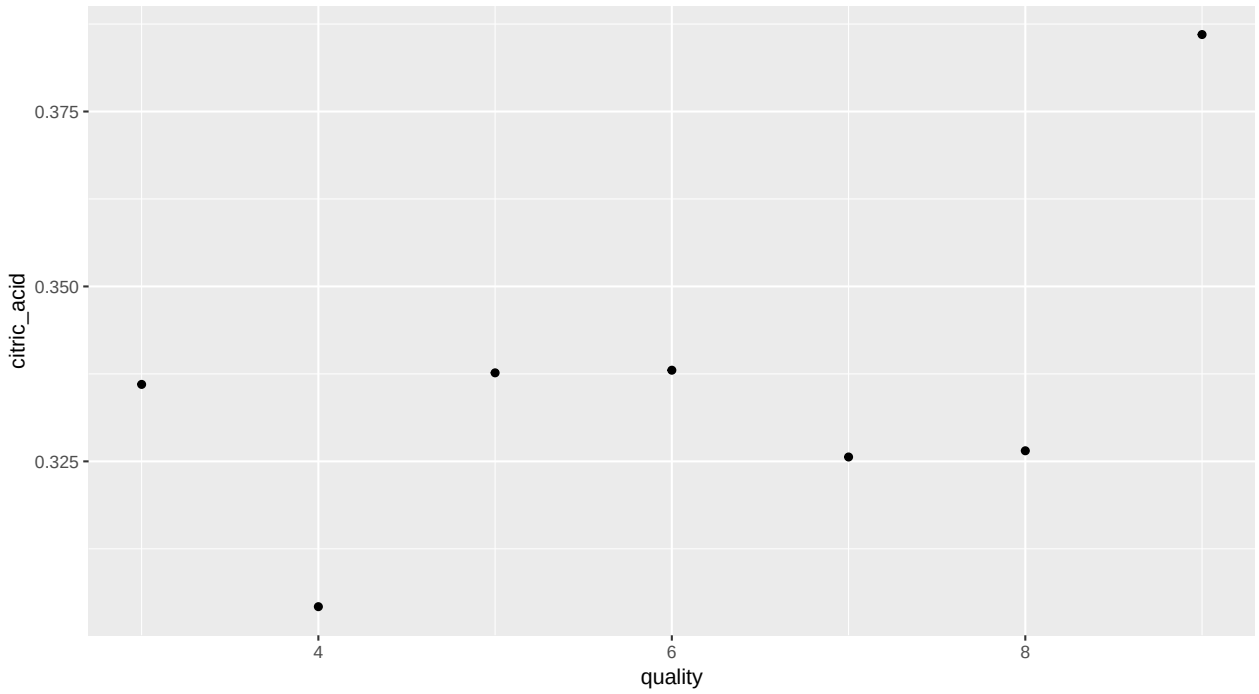
So far, we have determined that higher quality wines, 8 or above, will have higher pH, higher fixed_acidity, low volatile acidity, and higher alcohol content.



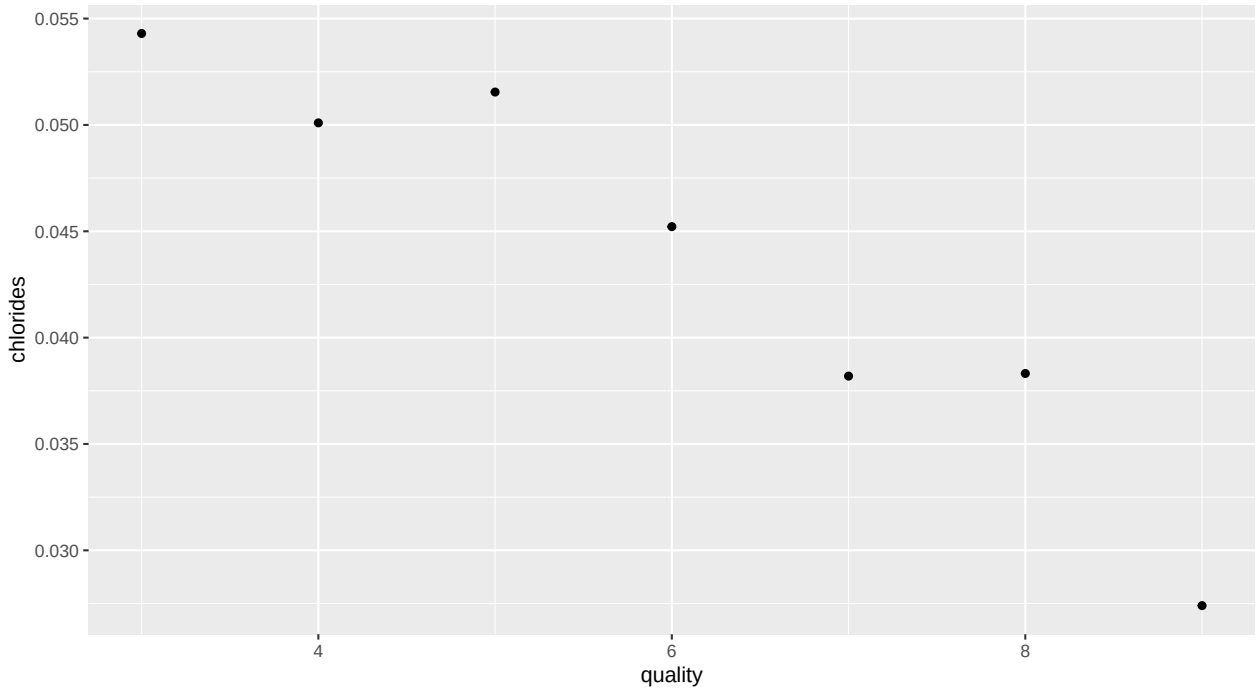
White wines with a higher quality of 8 or more will have lower density.



As quality increases, so does the proportion of free SO2 in the wine.



Higher quality wines will have a higher citric acid content.



There is a strong negative relationship between quality and chloride content.

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?

Higher quality wines will have a high alcohol content, high sulfate dioxide, high citric acid, low chlorides/salt, lower density than water, low volatile acidity, and low fixed acididty.

Alcohol has approximately 0.7893g/cm^3 density, which is lower than the density of water. So a higher alcohol content by volume in wine will also have a lower density. A higher sulfate dioxide content is a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant. So a higher amount of any of the variables sulphates, free sulfur dioxide, or total sulfur dioxide will result in a higher quality wine. Citric acid is found in small quantities and can add 'freshness' and flavor to wines. It is logical that a higher citric acid content will increase the quality of white wines. There is a strong, indirect relationship between chlorides and white wines. Lastly, too high of a volatile acidity in white wines will lead to an unpleasant vinegar taste.

According to an article by winemakermag, optimal pH levels of white wine is 3.2-3.5. This may explain why there is an increase in pH and a lower acidity favored for higher quality, white wines.

The fixed acidity graph shows that there is a dip and then an increase for quality, white wines. However, according to the variable attribute description, it states that higher quality wines will have a higher fixed acidity due to its nonvolatile properties. The increase in fixed acidity follows the description of its variables.

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

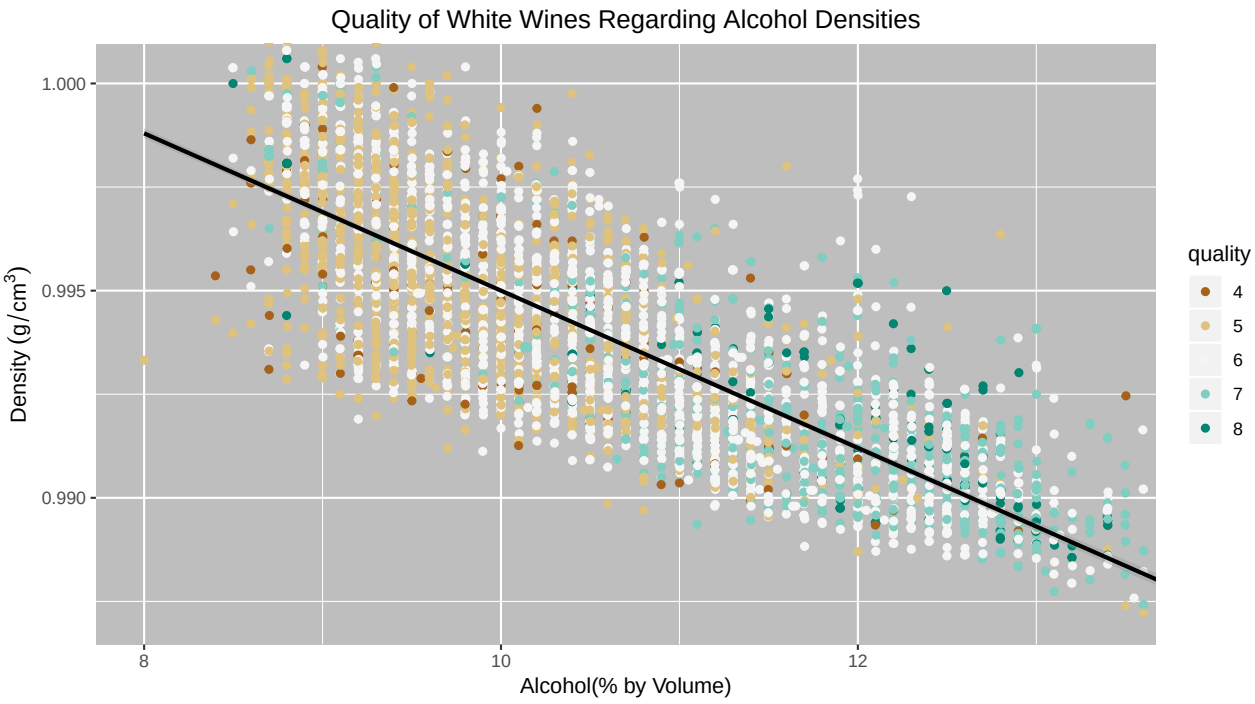
Nope

What was the strongest relationship you found?

The strongest relationship I found was that higher quality wines will have lower chloride, or salt, content.

Multivariate Plots Section

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed_acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile_acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric_acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual_sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free_sulfur_dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total_sulfur_dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : chr  "6" "6" "6" "6" ...
```




```
##
## Call:
## lm(formula = quality ~ alcohol + free_sulfur_dioxide + density,
##     data = wineQualityWhites)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4344 -0.5226 -0.0043  0.4925  3.1622
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.387e+01  6.191e+00  -2.240  0.02513 *
## alcohol              3.649e-01  1.467e-02  24.878  < 2e-16 ***
## free_sulfur_dioxide  6.211e-03  6.947e-04   8.941  < 2e-16 ***
## density              1.579e+01  6.113e+00   2.582  0.00984 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7897 on 4894 degrees of freedom
## Multiple R-squared:  0.2054, Adjusted R-squared:  0.2049
## F-statistic: 421.8 on 3 and 4894 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?

I chose the two variables density and alcohol because density is heavily affected by alcohol. Alcohol has a low density and it makes sense that with a greater alcohol content, quality of the wine will increase; the teal

colors of higher wine quality is at the lower right-hand corner of the plot. Other variables that affect density is free sulfur dioxide, total sulfur dioxide and sulphates which means that more of these will increase density slightly and increase quality. However, I did not include these in my plots since I did not want to confuse the readers of this project. A lower density of white wine should generally be higher quality wine. Though, not all dense additives to wine will increase quality. Sometimes these additivies will increase density while increasing quality.

Were there any interesting or surprising interactions between features?

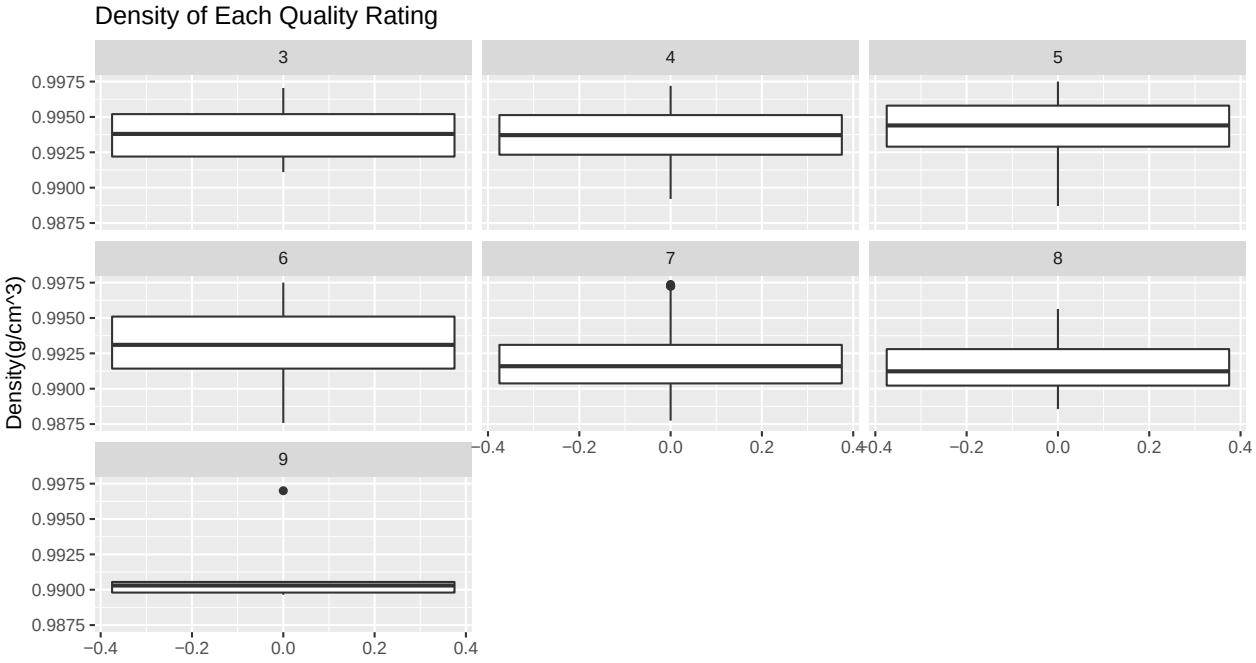
Like I stated above, some aspects of the wine will increase and decrease density. Generally, a lower density wine will have a higher quality.

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

I created a linear regression model using the variables free sulfur dioxide, alcohol, and density. It showed that the r^2 came out to be 0.21 which means that the three variables accounts for 21% of the quality of white wines. There are a lot more other variables that account for the quality of white wines.

Final Plots and Summary

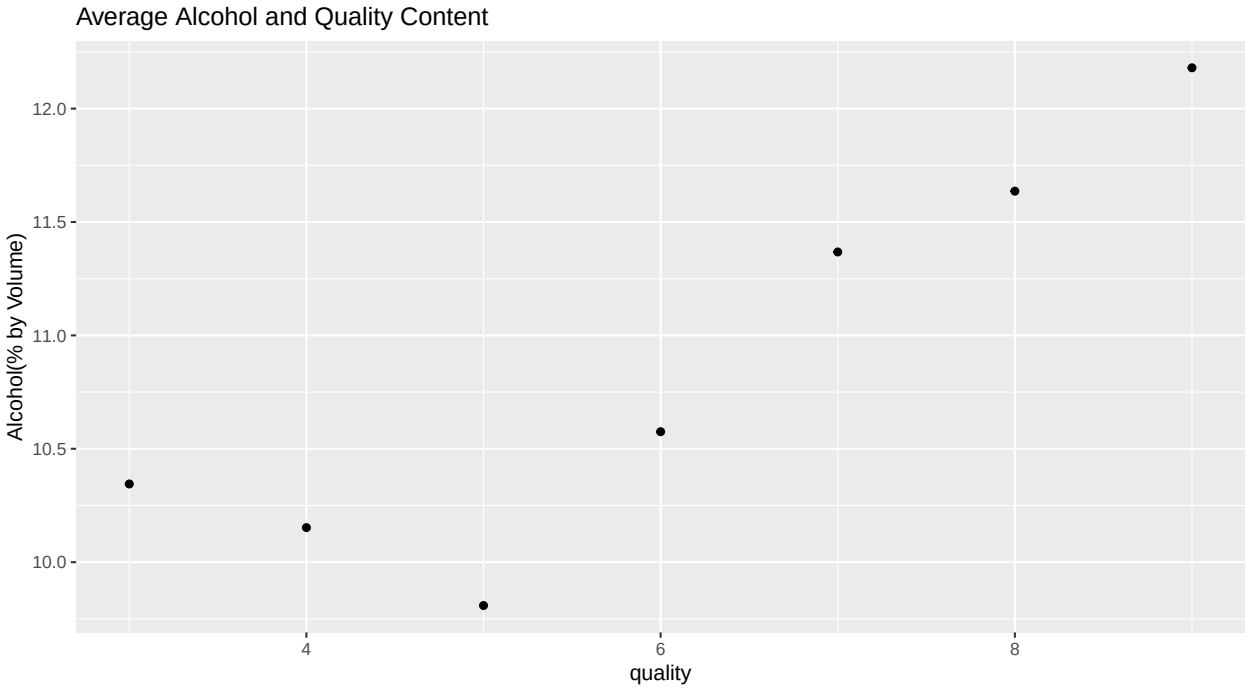
Plot One



Description One

These boxplots show that there is a decrease in density as quality of white wine increases. With a 3 quality rating, the density reaches 0.9950g/cm^3. With a 9 quality rating, it is just above 0.9900g/cm^3.

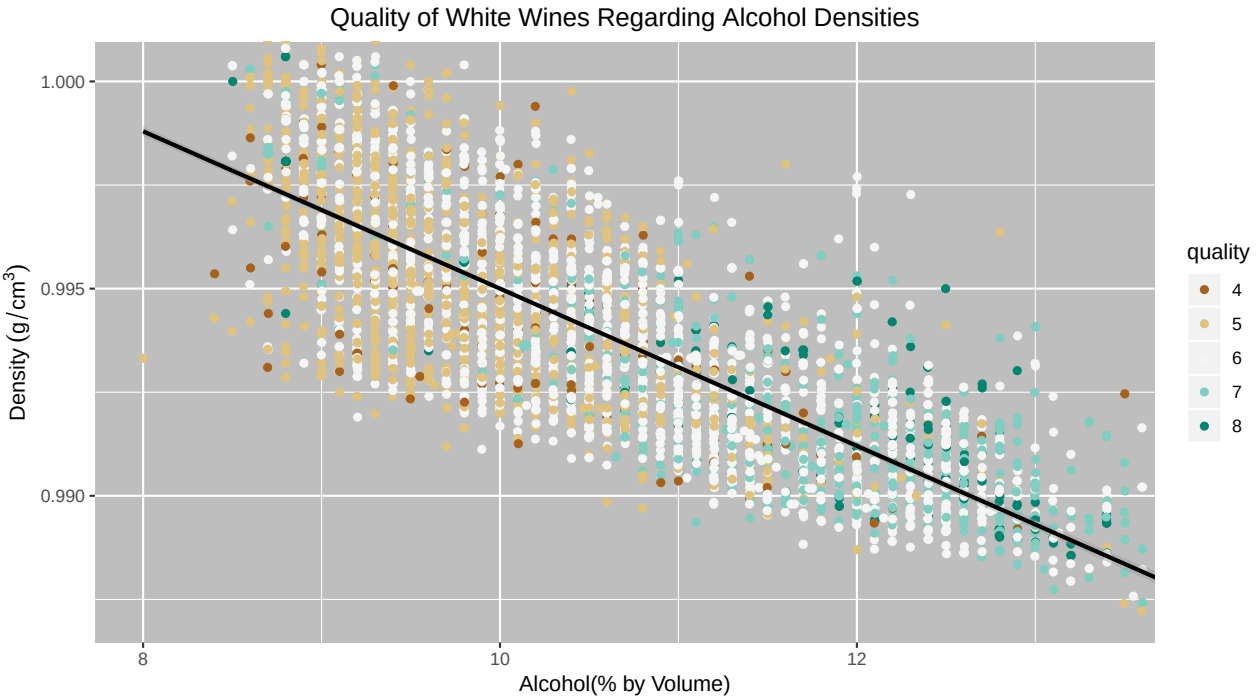
Plot Two



Description Two

As quality increases, the percentage of alcohol increases.

Plot Three



Description Three

Building upon plot 1 and 2 findings, we can see the general trend that quality wine will have a higher alcohol percentage and thus, a lower density than water.

Reflection

This reflection paragraph will sum up what went well and what didnt from doing this exploratory data analysis project. What did not go well is that the dataset included mostly all integer variables and no categorical variables. Thus, it was hard finding a way to color the graphs so that it looks presentable and interesting. Another aspect that did not go well is that the very extreme ends of the quality scale did not follow the trend of increasing or decreasing. By cutting off the quality rating of 3 and 9, the graphs followed a nice trend line. What went well is that the data was all clean, which I liked very much. The trend lines followed what was expected. The only thing that I had to clean up was the variable names which had a "." instead of a "_" to represent the spaces. For future improvements on this dataset, I recommend including more categorical variables so that data analysts can have a variety of data types to work with. Maybe "color" of the white wine can be included or a categorical pH variable(i.e. very low acidity, low acidity, acidic, neutral, basic, etc..). References:

https://winemakermag.com/article/1013-creating-a-balanced-must

https://www.google.com/search?ei=MiRrXNKGBsuOsQX_v4HgBg&q=alcohol+density+g%2Fcm3&oq=alcohol+density+g%2Fcm3&gs_l=psy-ab.3..0j0i22i30l5.131711.137190..137322...4.0..0.138.1631.24j1.....0....1..gws-wiz.....0i71j0i67j0i131j0i10j0i13j0i13i10j0i13i30j0i13i10i30.bbmQOcjTQiM

https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt