Explore and Summarize Data

==========

Introduction and Summary Data

The data set contains information on red wines and the chemical properties that influence the quality of red wines. I selected this data set out of curiosity of red wines and was curious what insights could be gleaned. I found the following definitions and attributes from another Udacity wine project, and listed the site in the resources section. I found the attributes helpful when looking into the different variables.

Data variables and definition1

Atributes Information

1. Fixed Acidity (tartaric acid - g/dm^3)
2. Volatile acidity (acetic acid - g/dm^3)
3. Citric acid (g/dm^3)
4. Residual sugar (g/dm^3)
5. Chlorides (sodium chloride - g/dm^3)
6. Free sulfur dioxide (mg/dm^3)
7. Total sulfur dioxide (mg/dm^3)
8. Density (g/cm^3)
9. pH
10. Sulphates (potassium sulphate - g/dm^3)
11. Alcohol (% by volume)
12. Quality (score between 0 and 10)

Description Atributes

1. Fixed acidity: Most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2. Volatile acidity: The amount of acetic acid in wine
3. Citric acid: Found in small quantities, citric acid can add 'freshness' and flavor to wines
4. Residual sugar: The amount of sugar remaining after fermentation stops
5. Chlorides: the amount of salt in the wine
6. Free sulfur dioxide: The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. Total sulfur dioxide: Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. Density: The density of a substance is its mass per unite volume
9. pH: Describes how acidic or basic a substance is on a scale from 0 (very acidic) to 14 (very basic)
10. Sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
11. Alcohol: The percent alcohol content of the wine
12. Quality: Score between 0 and 10

Initial Discovery

```
head(wineData, 10)
```

```
##    fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1            7.4             0.70        0.00            1.9     0.076
## 2            7.8             0.88        0.00            2.6     0.098
## 3            7.8             0.76        0.04            2.3     0.092
## 4           11.2             0.28        0.56            1.9     0.075
## 5            7.4             0.70        0.00            1.9     0.076
## 6            7.4             0.66        0.00            1.8     0.075
## 7            7.9             0.60        0.06            1.6     0.069
## 8            7.3             0.65        0.00            1.2     0.065
## 9            7.8             0.58        0.02            2.0     0.073
## 10           7.5             0.50        0.36            6.1     0.071
##    free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 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
## 7                   15                   59  0.9964 3.30      0.46     9.4
## 8                   15                   21  0.9946 3.39      0.47    10.0
## 9                    9                   18  0.9968 3.36      0.57     9.5
## 10                  17                  102  0.9978 3.35      0.80    10.5
##    quality
## 1        5
## 2        5
## 3        5
## 4        6
## 5        5
## 6        5
## 7        5
## 8        7
## 9        7
## 10       5
```

```
str(wineData)
```

```
## '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 0.47 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 ...
```

```
sum(is.na(wineData))
```

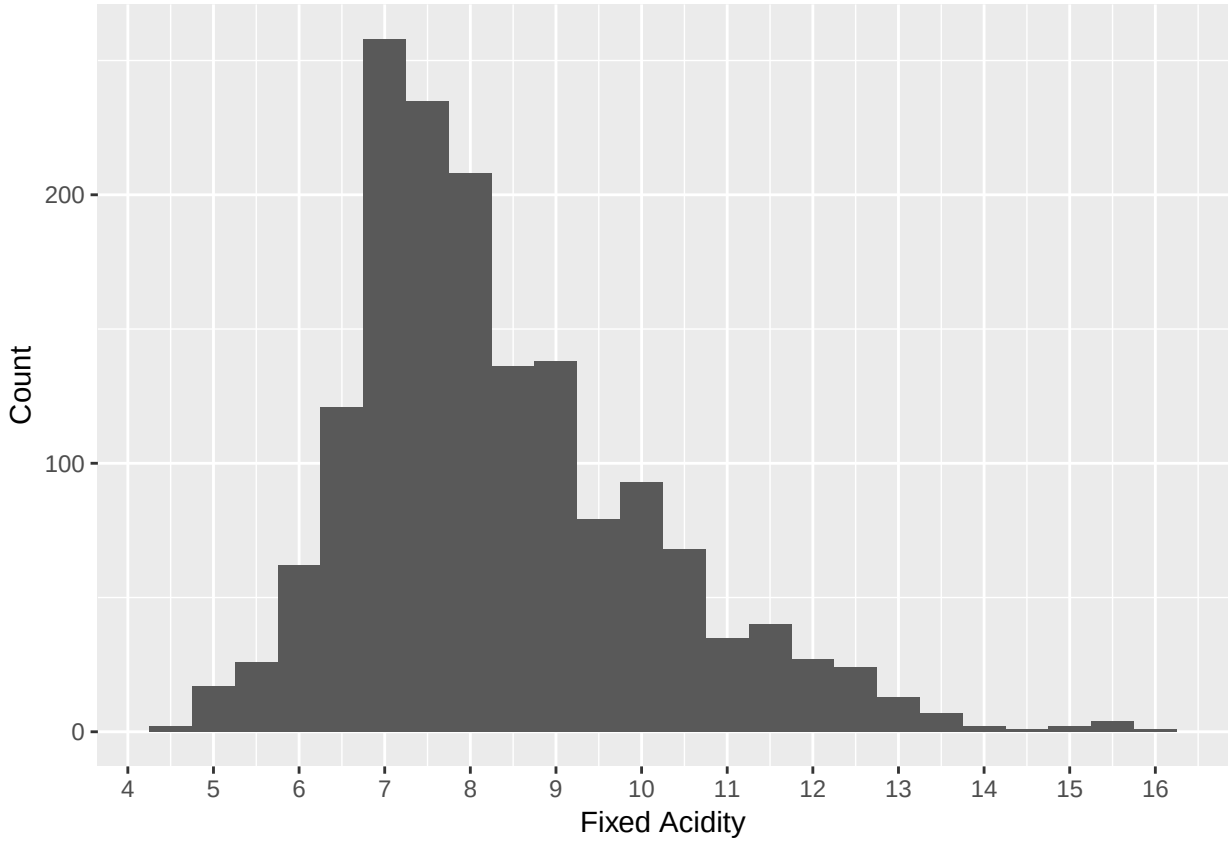
```
## [1] 0
```

I loaded the first 10 rows and ran the *str)* function to view a compact summary plus understand the overall data structure. Finally, ran *sum(is.na)* on data set to verify there aren't any missing values in the data set.

Univariate Plot Section

Fixed Acidity

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=fixed.acidity), binwidth = 0.5) +
  scale_x_continuous(breaks = seq(0,17,1)) +
  xlab("Fixed Acidity") +
  ylab("Count")
```



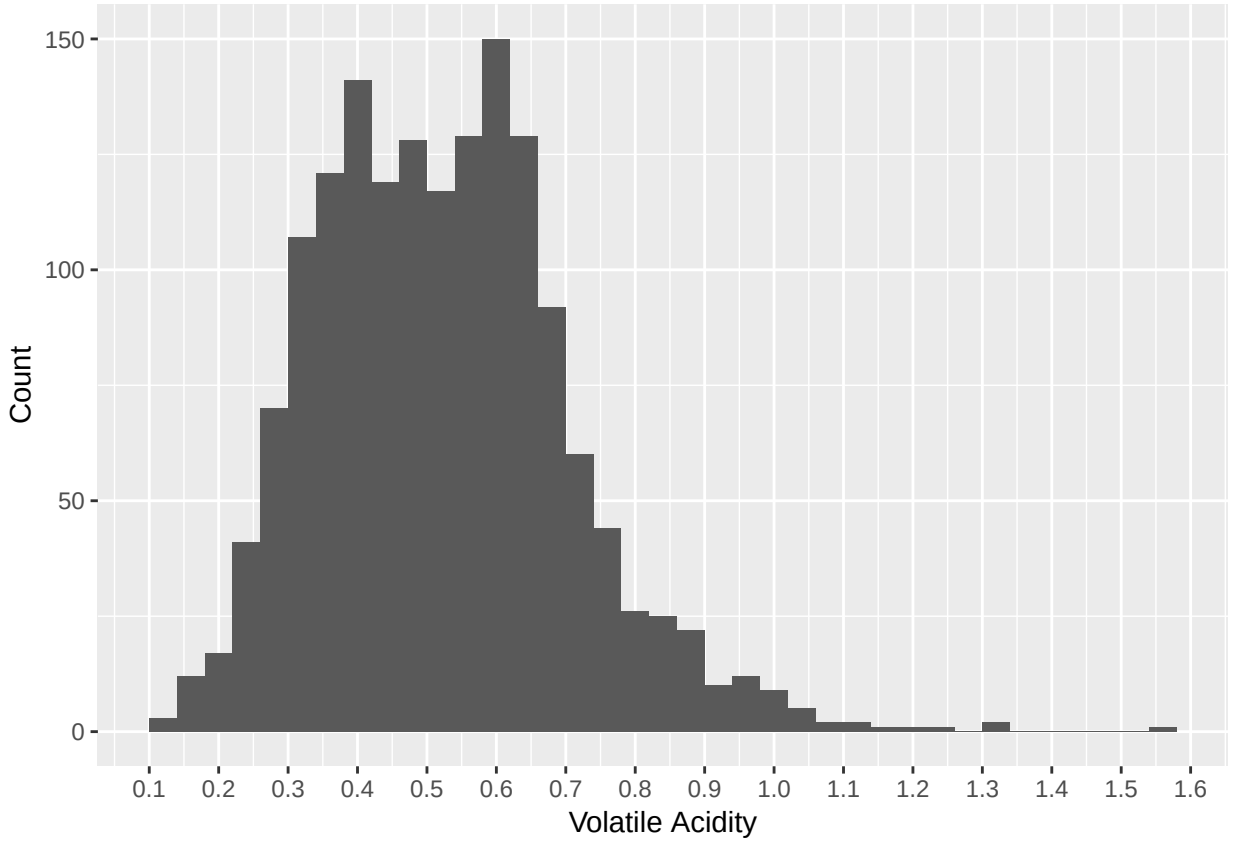

```
summary(wineData$fixed.acidity)
```

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

The histogram is skewed right. The fixed acidity in most red wines is approximately between 6.5 g/dm^3 and 7.5 g/dm^3. The median (7.90) and mean (8.32) are pulled to the left, and tail is to the right, which are all indicators of a right skewed distribution.

Volatile Acidity

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=volatile.acidity), binwidth = 0.04) +
  scale_x_continuous(breaks = seq(0,1.6,0.1)) +
  xlab("Volatile Acidity") +
  ylab("Count")
```



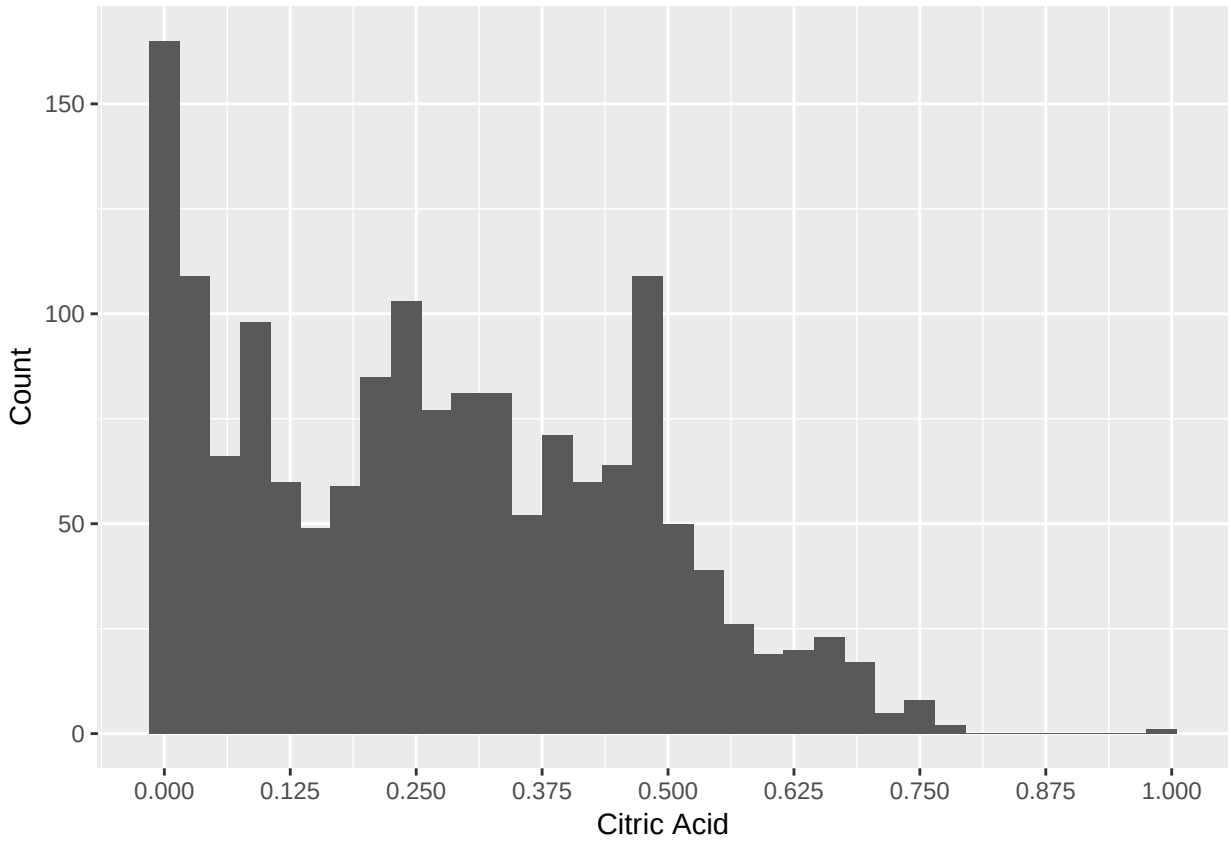

```
summary(wineData$volatile.acidity)
```

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

The histogram is skewed right. The volatile acidity in most red wines is approximately between 0.35 g/dm^3 and 0.65 g/dm^3. The median (0.52) and mean (0.53) are pulled to the left, and tail is to the right. There might be some outliers around 1.3 g/dm^3 and 1.55 g/dm^3.

Citric Acid

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=citric.acid), binwidth = 0.03) +
  scale_x_continuous(breaks = seq(0,1,0.125)) +
  xlab("Citric Acid") +
  ylab("Count")
```




```
summary(wineData$citric.acid)
```

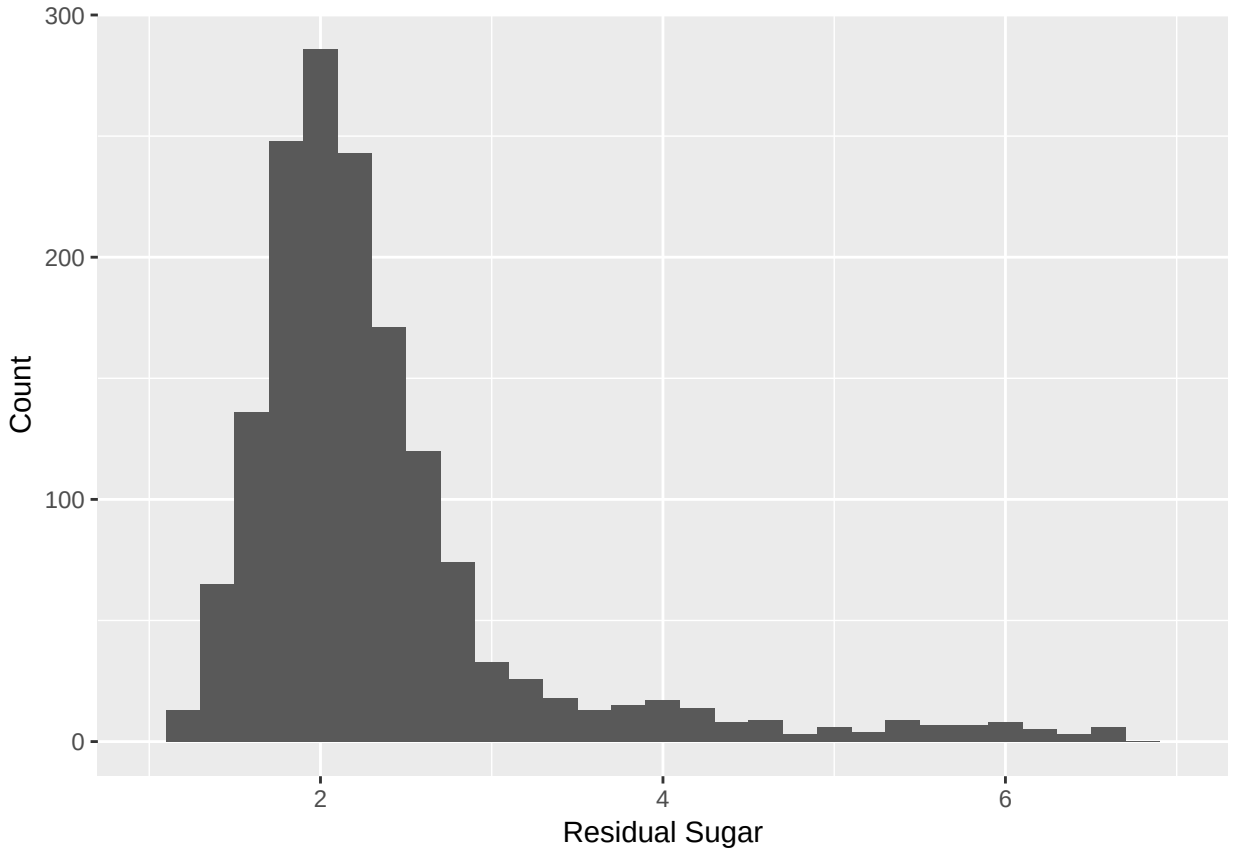
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

The histogram is skewed right. The citric acid in most red wines is approximately 0.0 g/dm^3. The median (0.26) and mean (0.27) are pulled to the left.

Residual Sugar

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=residual.sugar), binwidth = 0.2) +
  xlim(1,7) +
  xlab("Residual Sugar") +
  ylab("Count")
```

```
## Warning: Removed 31 rows containing non-finite values (stat_bin).
```




```
summary(wineData$residual.sugar)
```

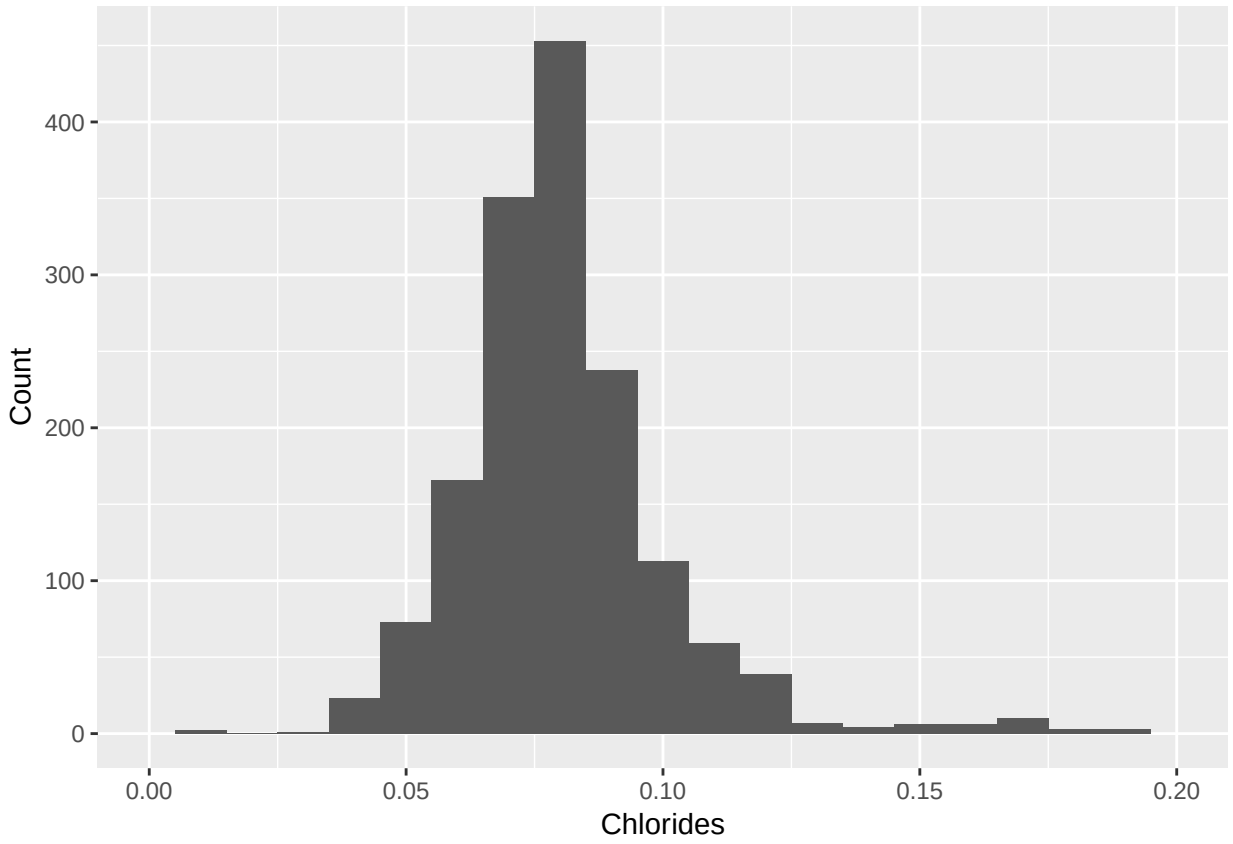
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

The histogram is skewed right. The residual sugar in most red wines is approximately 2.0 g/dm^3. The median (2.2) and mean (2.539) are pulled to the left. I also used *xlim()* to remove outliers to create a cleaner visual.

Chlorides

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=chlorides), binwidth = 0.01) +
  xlim(0,0.2) +
  xlab("Chlorides") +
  ylab("Count")
```

```
## Warning: Removed 41 rows containing non-finite values (stat_bin).
```



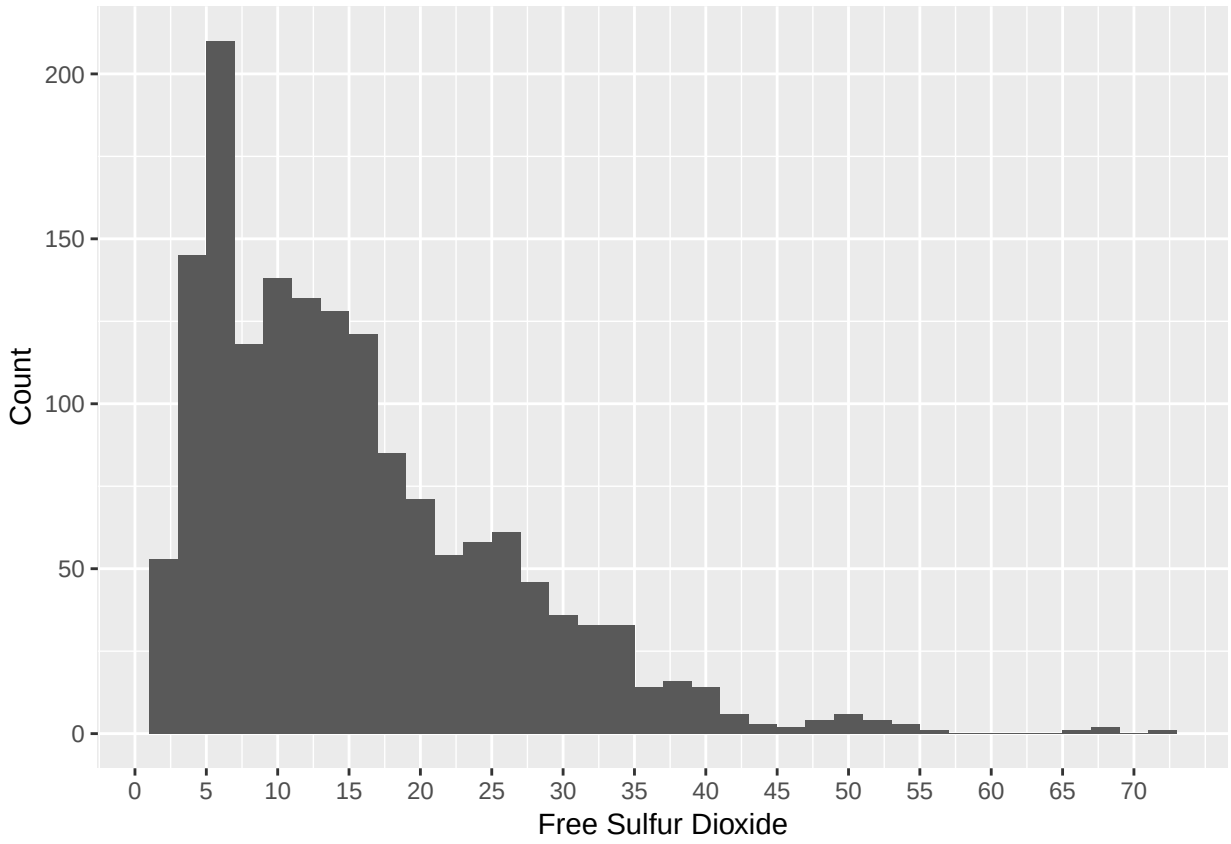

```
summary(wineData$chlorides)
```

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

The histogram is skewed right. The amount of chlorides found in most red wines is approximately 0.78 g/dm^3. The median (0.79) and mean (0.087) are pulled to the left. I also used *xlim()* to remove outliers to create a cleaner visual.

Free Sulfur Dioxide

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=free.sulfur.dioxide), binwidth = 2) +
  scale_x_continuous(breaks = seq(0,70,5)) +
  xlab("Free Sulfur Dioxide") +
  ylab("Count")
```




```
summary(wineData$free.sulfur.dioxide)
```

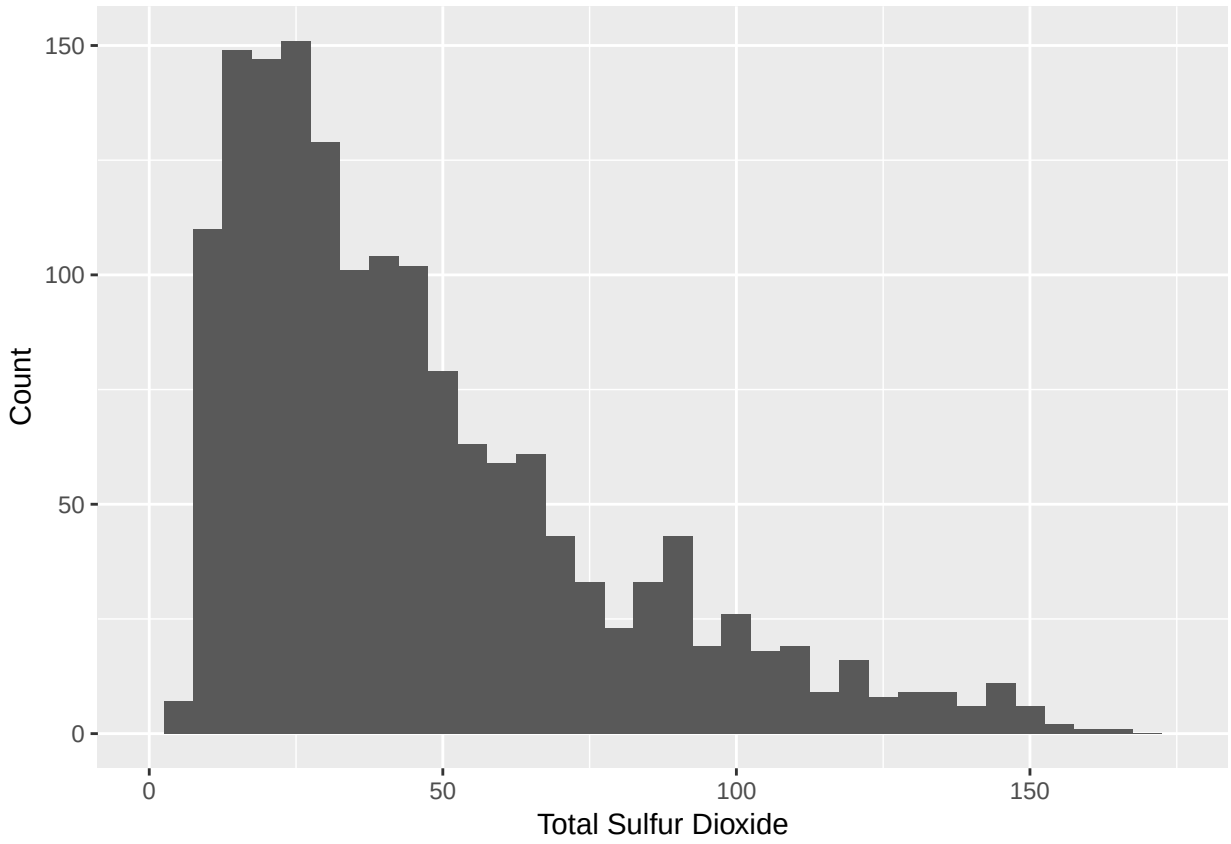
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

The histogram is skewed right. The citric acid in most red wines is approximately 6.0 mg/dm^3. The median (14.0) and mean (15.87) are pulled to the left.

Total Sulfur Dioxide

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=total.sulfur.dioxide), binwidth = 5) +
  xlim(0,175) +
  xlab("Total Sulfur Dioxide") +
  ylab("Count")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



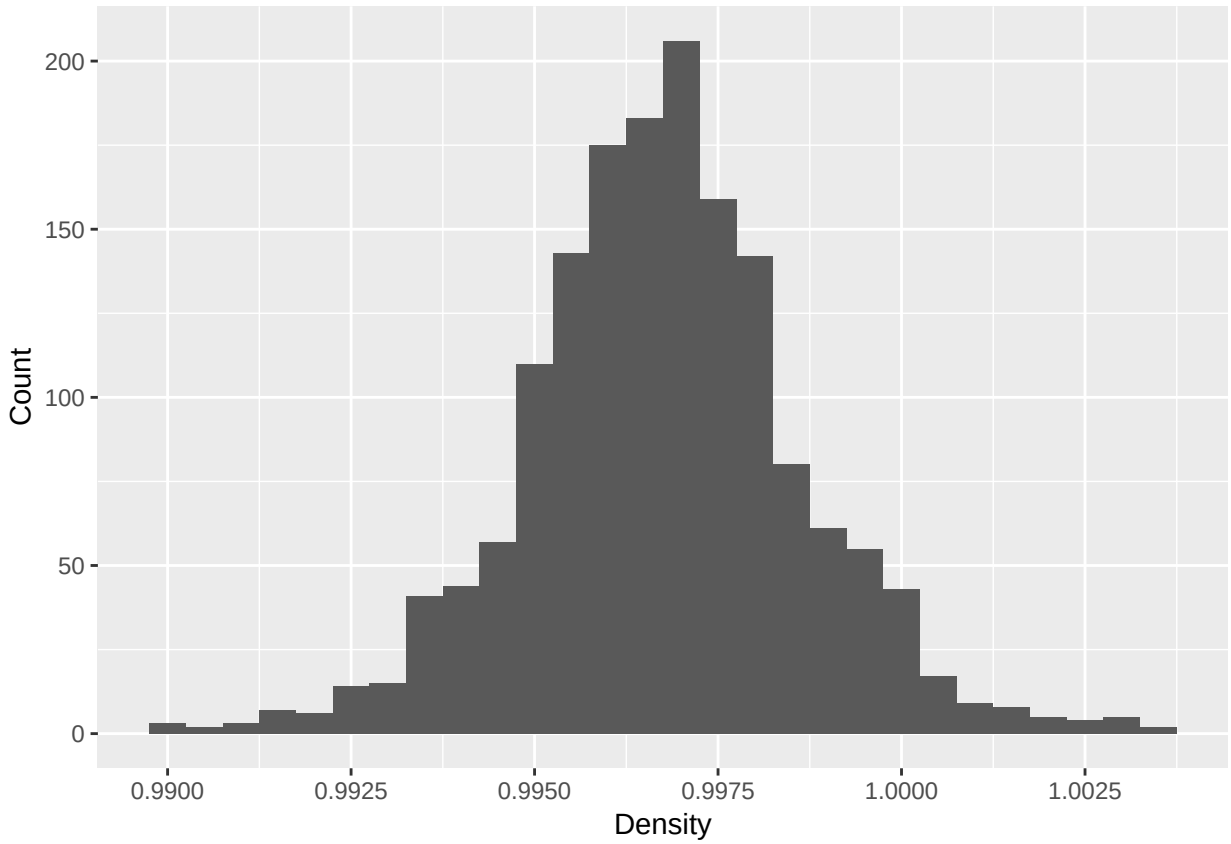

```
summary(wineData$total.sulfur.dioxide)
```

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

The histogram is skewed right. The amount of total sulfur dioxide found in most red wines is approximately 25.0 mg/dm^3. The median (38.0) and mean (46.47) are pulled to the left. I also used *xlim()* to remove outliers to create a cleaner visual.

Density

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=density), binwidth = 0.0005) +
  scale_x_continuous(breaks = seq(0.9,1.05,0.0025)) +
  xlab("Density") +
  ylab("Count")
```



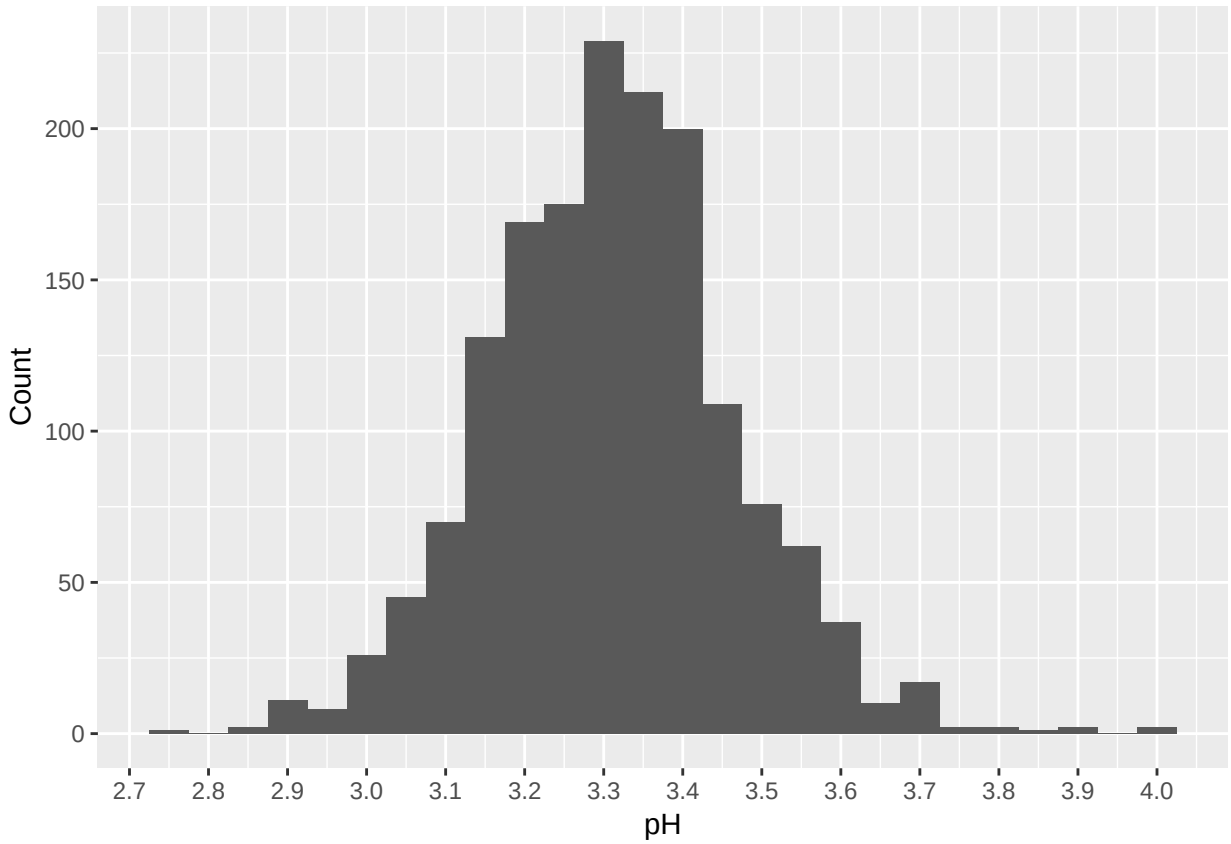

```
summary(wineData$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

The histogram is skewed right. The density of most red wines is approximately 0.996 g/cm^3. The median (0.9968) and mean (0.9967) are pulled to the left.

pH

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=pH), binwidth = 0.05) +
  scale_x_continuous(breaks = seq(0,4,0.1)) +
  xlab("pH") +
  ylab("Count")
```




```
summary(wineData$pH)
```

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

The histogram is skewed right. The pH of most red wines is approximately 3.3 pH units. The median (3.31) and mean (3.311) are pulled to the left.

Sulphates

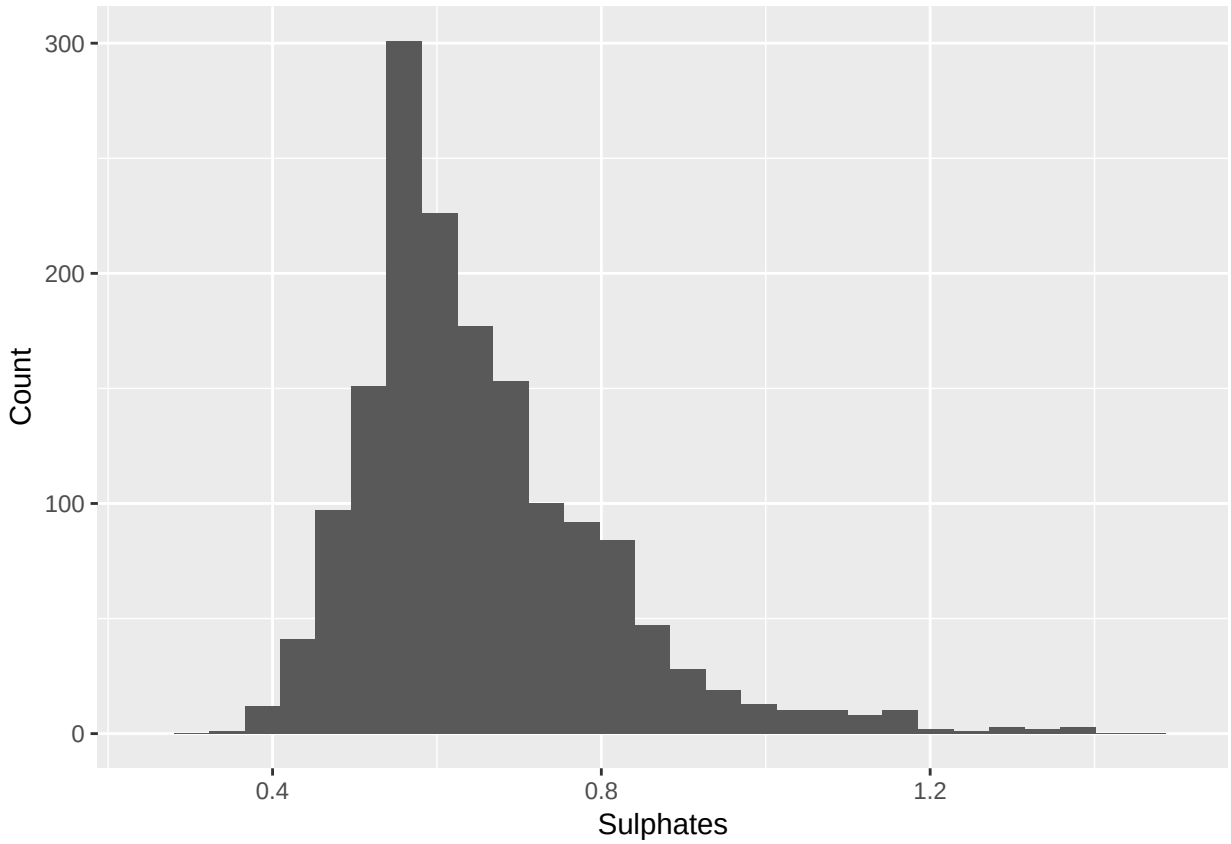
```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=sulphates), bandwidth = 0.25) +
  xlim(0.25, 1.5) +
  xlab("Sulphates") +
  ylab("Count")
```

```
## Warning: Ignoring unknown parameters: bandwidth
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 8 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



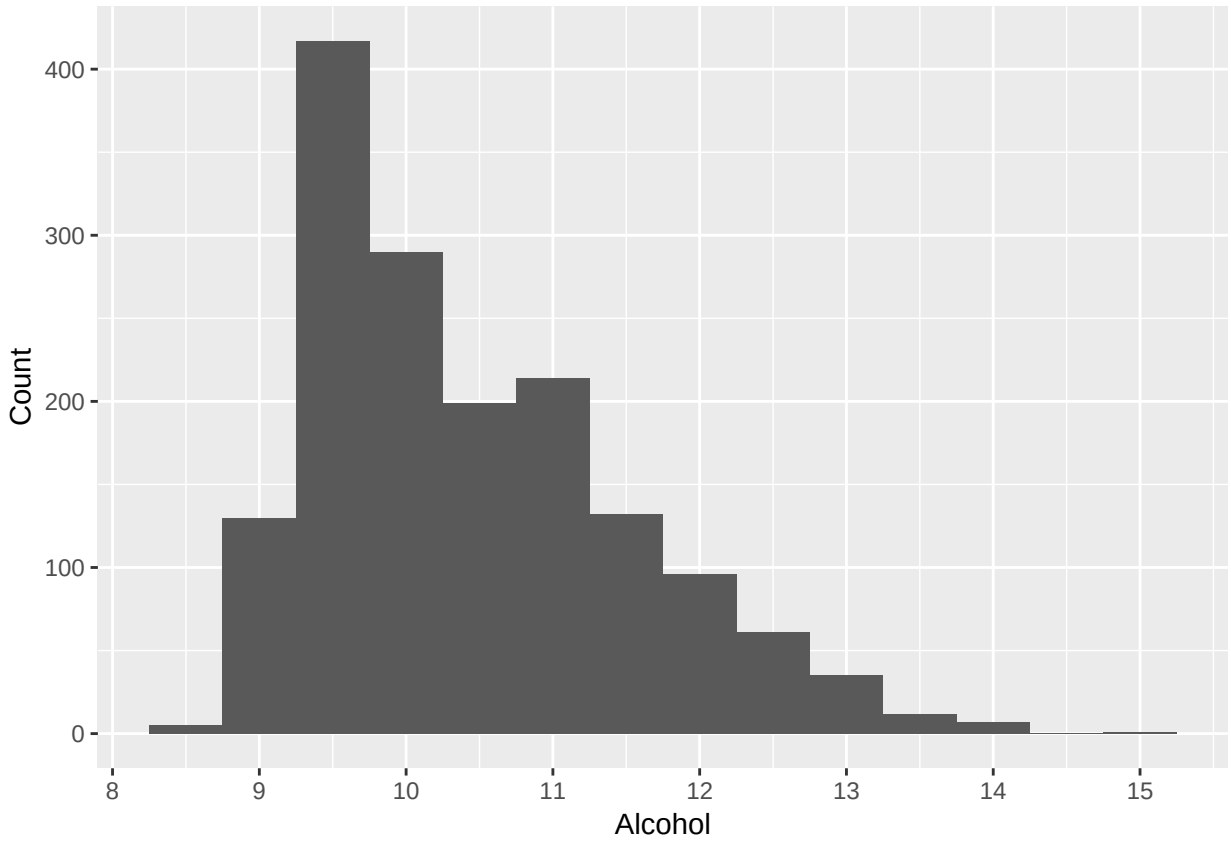

```
summary(wineData$sulphates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

The histogram is skewed right. The amount of sulphatees found in most red wines is approximately 0.55 g/dm^3. The median (0.62) and mean (0.65) are pulled to the left. I also used *xlim()* to remove outliers to create a cleaner visual.

Alcohol

```
ggplot(data = wineData) +
  geom_histogram(mapping = aes(x=alcohol), binwidth = 0.5) +
  scale_x_continuous(breaks = seq(8,15,1)) +
  xlab("Alcohol") +
  ylab("Count")
```

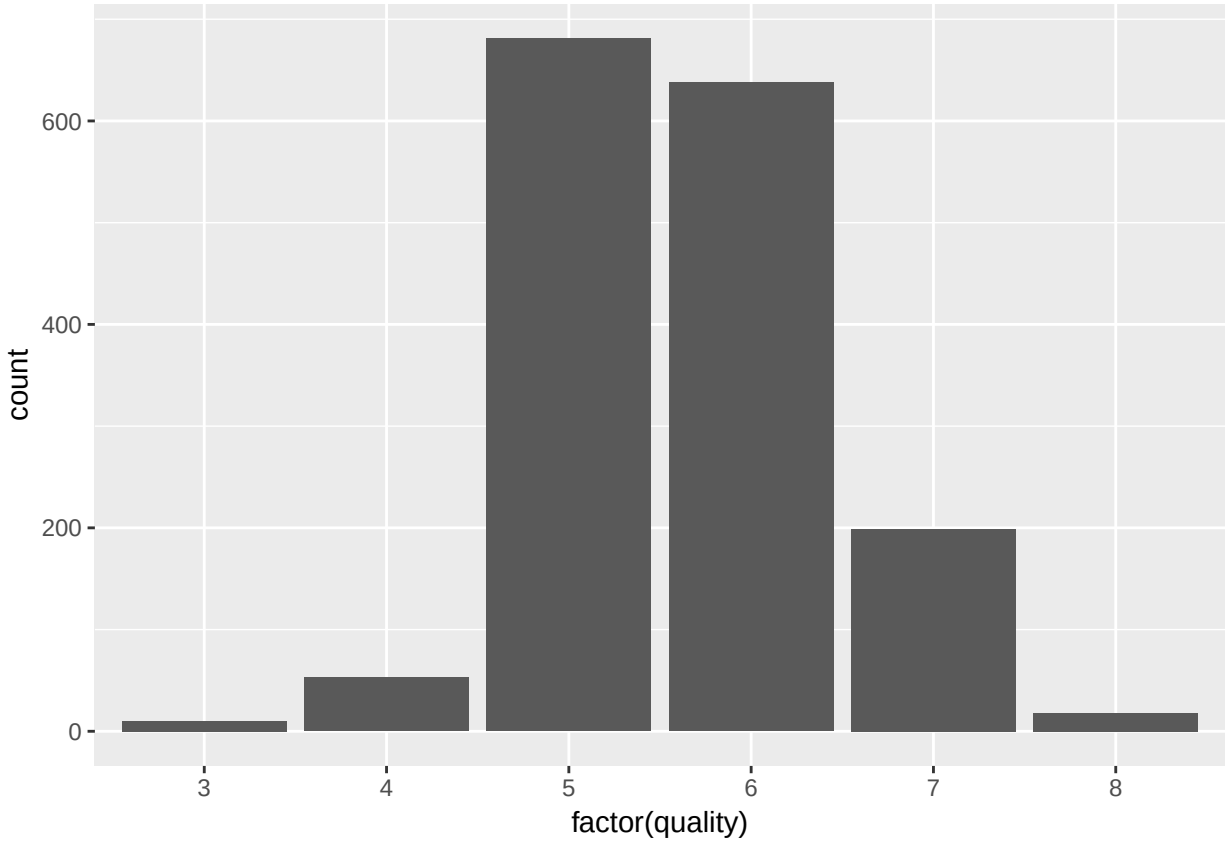



```
summary(wineData$alcohol)
```

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

The histogram is skewed right. The amount of alcohol found in most red wines is approximately 8.0% by volume. The median (10.2) and mean (10.42) are pulled to the left. I also used *xlim()* to remove outliers to create a cleaner visual.

Quality




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

The data set of red wine quality follows a normal distribution. Most wines are rated at a quality level of 5.0. When looking at other measures of central tendancy, the 1st quartile, median, 3rd quartile, and mean values are 5.0, 6.0, 6,0, and 5.636 respectively.

Univariate Analysis

What is the structure of your dataset?

The data set contains 12 variables, 1599 rows.

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

I am interested in looking at which ingredients drive quality ratings.

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

I would also like to investigate any corrleations found between acidity levels & pH, sulphates & density, or residual sugars & density.

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

No.

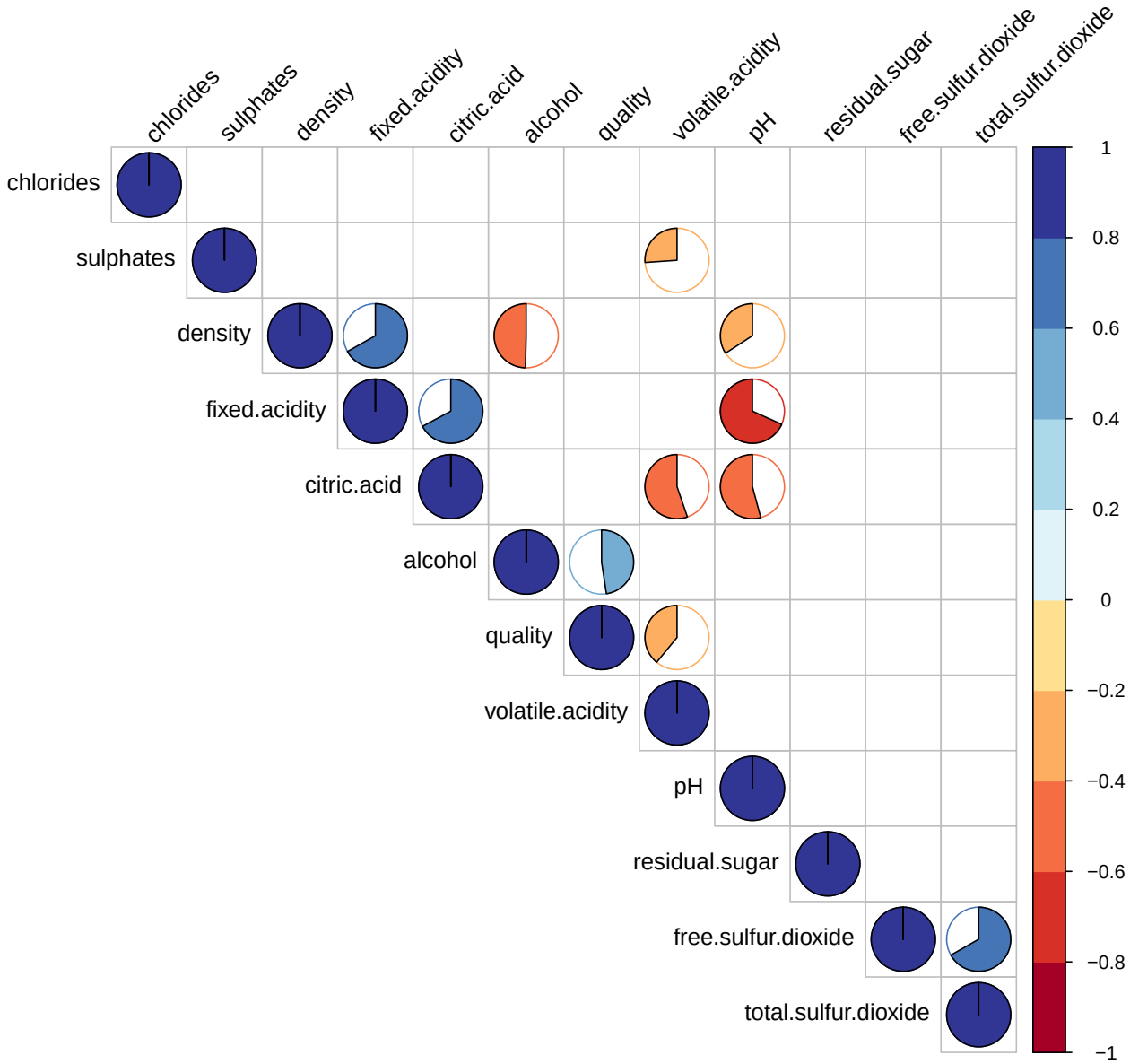
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

No, I did not perform any operations on the dataset. But, I did reduce the *xlim()* max value on a few graphs to clean up the visualizations. Also regarding the quality graph, I've never seen a 2nd and 3rd quartile value equal the same value, found this to be interesting, and would like to dig into it further.

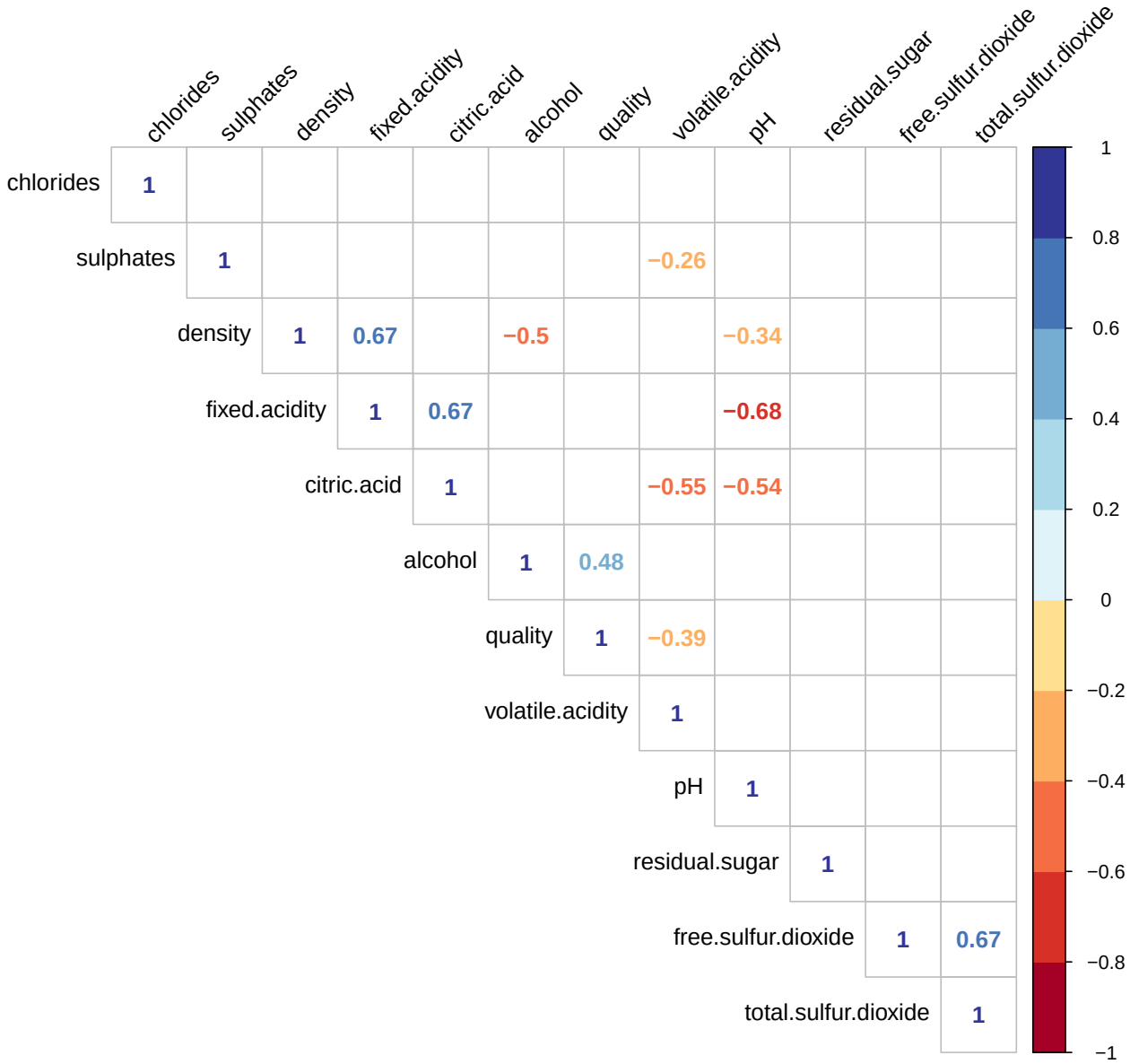
Bivariate Plots Section

As stated in the previous section I would like to investigate correlations between alcohol & quality, acidity levels & pH, sulphates & density, or residual sugars & density. To start out, I'm going to use *corrplot* [2] and other correlation matrices to determine which features have the strongest relationships.

```
wineCorrelation <- cor(wineData)
significance <- cor.mtest(wineCorrelation, conf.level = .95)
```




```
corrplot(wineCorrelation, method = "number",
         p.mat = significance$p, sig.level = .05, insig = "blank", order="hclust", tl.col = "black",type
         col=brewer.pal(n=10,name="RdYlBu"))
```



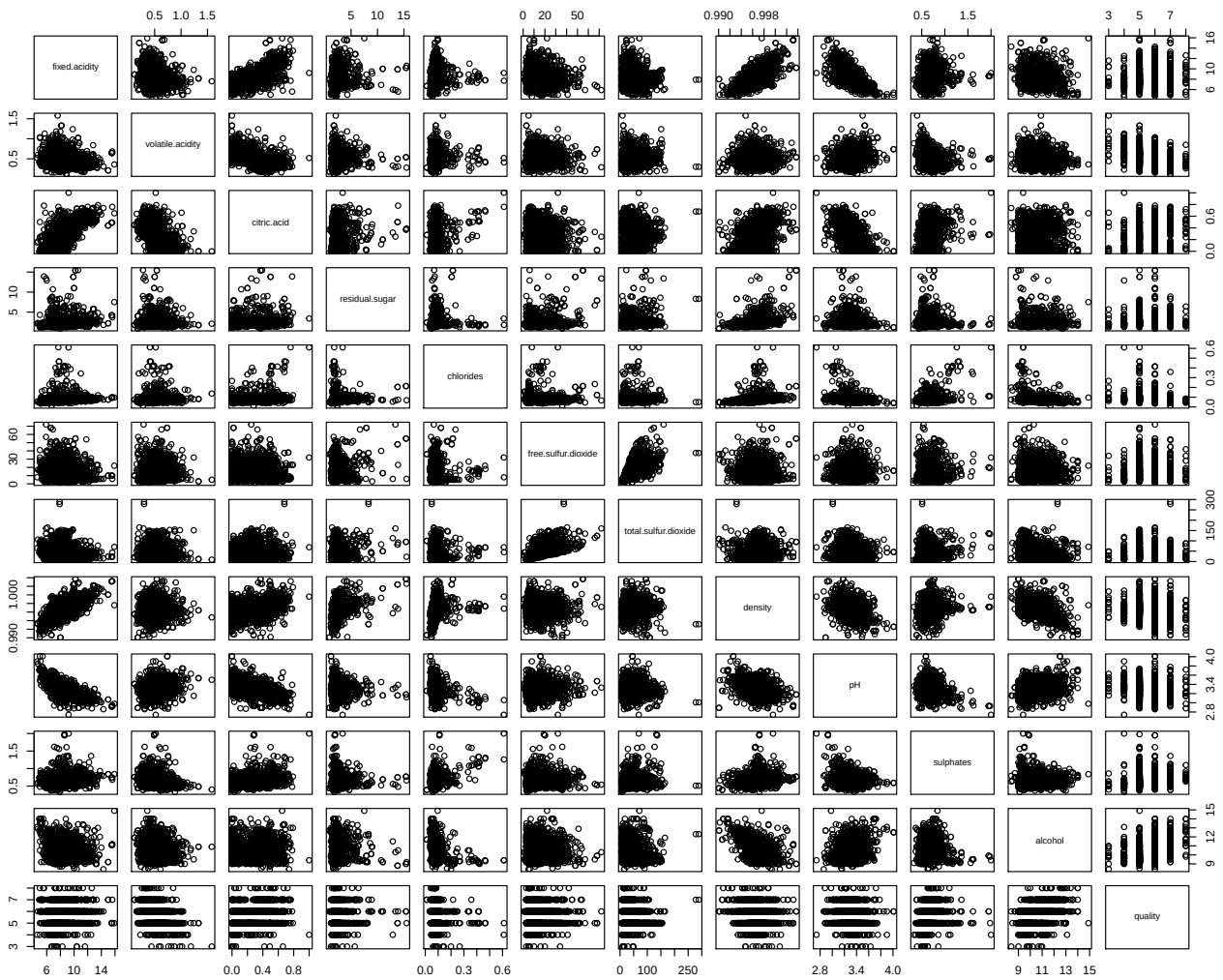
In the above figures, correlations with p-value > 0.05 are considered insignificant. In this case the correlation coefficient values are were left blank. The term insignifican comes from *corrplot*. However, I believe the terminology is slightly misleading. Another way of thinking about this is the observed statistic was closer to zero (i.e. null hypothesis) rather away from the test statistic. Therefore the values were left blank in this situation. It's not that the value was actually insignificant, but rather it did not exceed the test statistic value or fall within the area made up by the p-value which would make it statistically significant.

```
head(round(wineCorrelation, 4))
```

```
##                     fixed.acidity volatile.acidity citric.acid
## fixed.acidity              1.0000          -0.2561      0.6717
## volatile.acidity          -0.2561           1.0000     -0.5525
## citric.acid                0.6717          -0.5525      1.0000
## residual.sugar             0.1148           0.0019      0.1436
## chlorides                  0.0937           0.0613      0.2038
## free.sulfur.dioxide       -0.1538          -0.0105     -0.0610
##                     residual.sugar chlorides free.sulfur.dioxide
```

```
## fixed.acidity               0.1148    0.0937             -0.1538
## volatile.acidity            0.0019    0.0613             -0.0105
## citric.acid                 0.1436    0.2038             -0.0610
## residual.sugar              1.0000    0.0556              0.1870
## chlorides                   0.0556    1.0000              0.0056
## free.sulfur.dioxide         0.1870    0.0056              1.0000
##                      total.sulfur.dioxide density      pH sulphates alcohol
## fixed.acidity                 -0.1132  0.6680 -0.6830    0.1830 -0.0617
## volatile.acidity               0.0765  0.0220  0.2349   -0.2610 -0.2023
## citric.acid                    0.0355  0.3649 -0.5419    0.3128  0.1099
## residual.sugar                 0.2030  0.3553 -0.0857    0.0055  0.0421
## chlorides                      0.0474  0.2006 -0.2650    0.3713 -0.2211
## free.sulfur.dioxide            0.6677 -0.0219  0.0704    0.0517 -0.0694
##                      quality
## fixed.acidity         0.1241
## volatile.acidity     -0.3906
## citric.acid           0.2264
## residual.sugar        0.0137
## chlorides            -0.1289
## free.sulfur.dioxide -0.0507
```

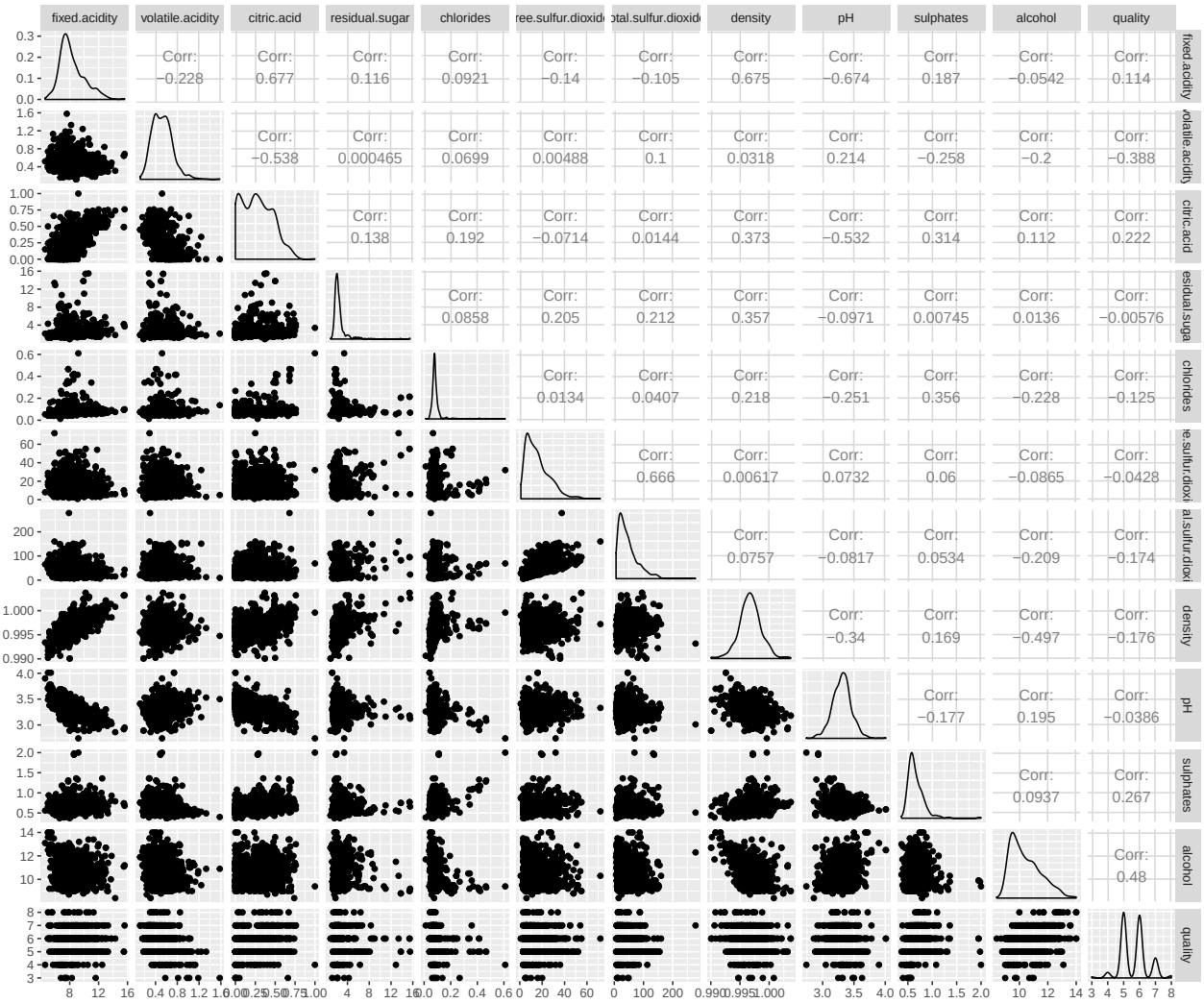
I used *round()* to make this data set more readable.




```
set.seed(1599)
wine_subset <- wineData[,c(1:12)]
names(wine_subset)
```

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

```
ggpairs(wine_subset[sample.int(nrow(wine_subset), 1000),])
```

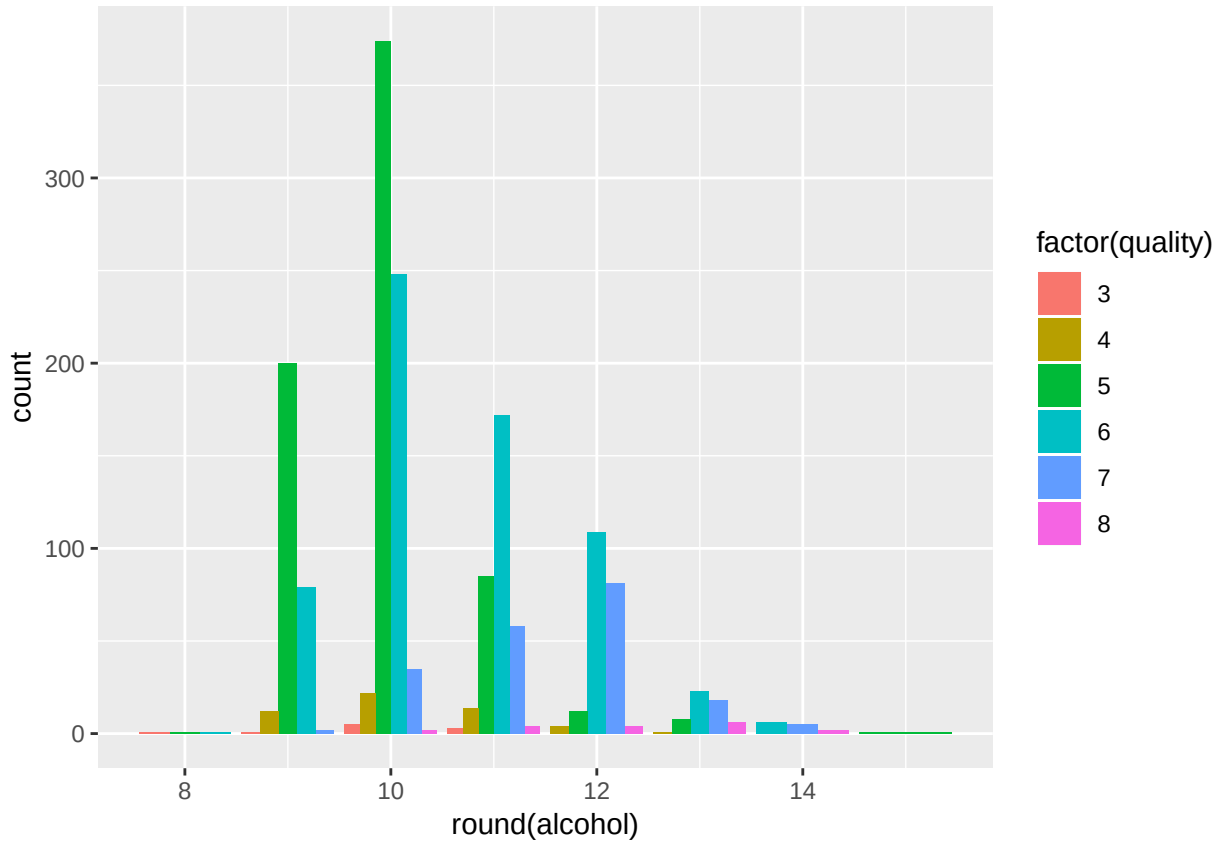


Simple scatter plot matrix[3].

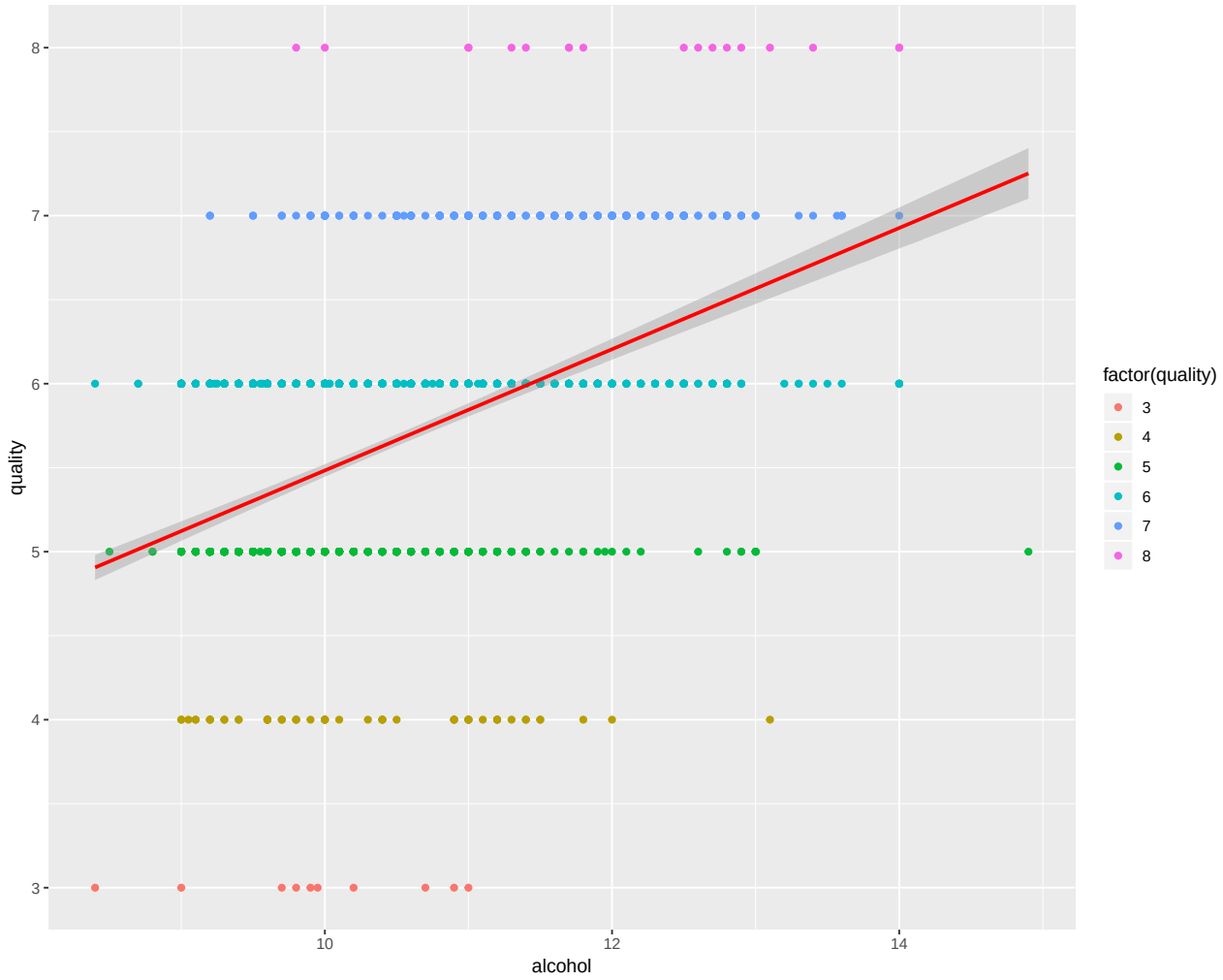
Bivariate Plots

Now that I looked at various matrices, I've determined which relationships I want to look at. Some items of interest are aclohol & quality, and fixed acidity & pH.

```
ggplot(data = wineData) +
  geom_bar(
    mapping = aes(x = round(alcohol), fill = factor(quality)),
    position = "dodge"
  )
```



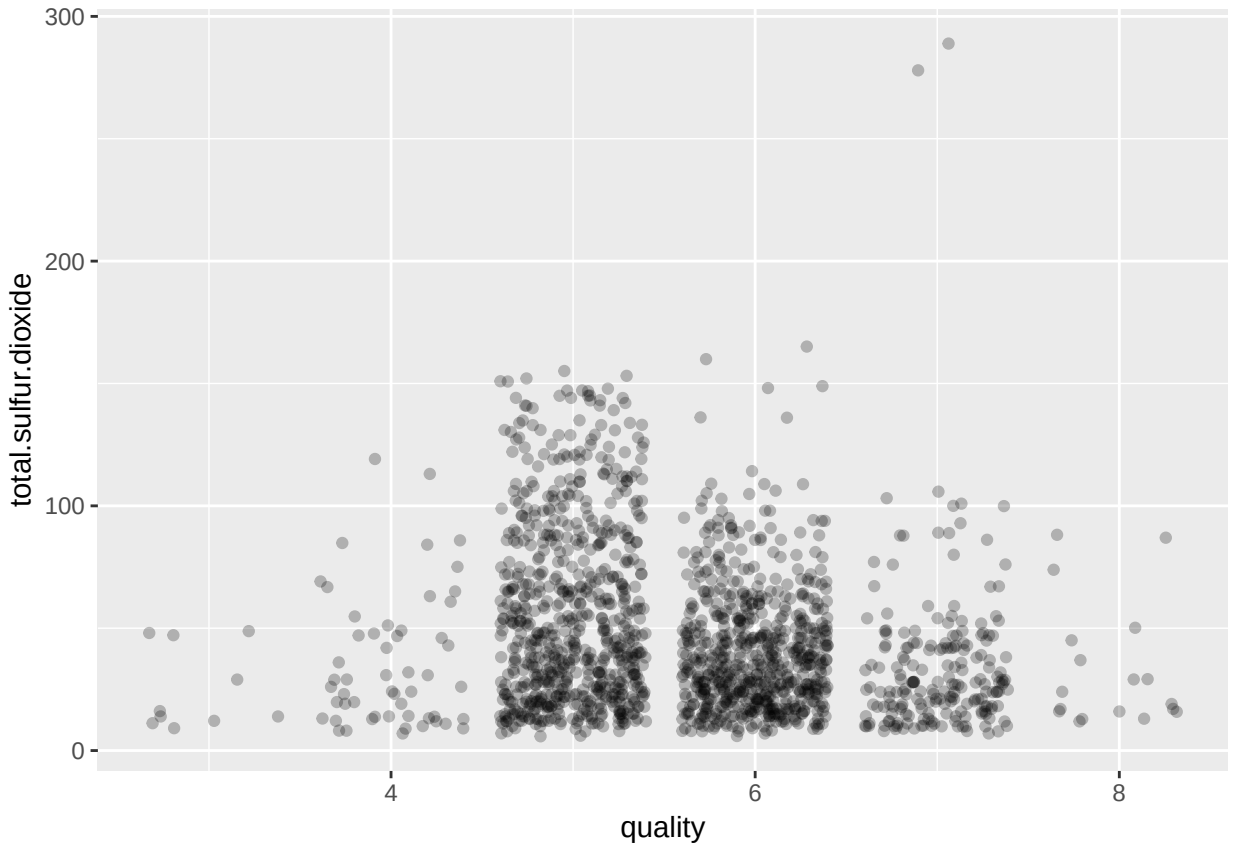

```
ggplot(data = wineData, mapping = aes(x=alcohol, y=quality)) +
  geom_point(mapping = aes(color = factor(quality))) +
  geom_smooth(color='red',method="lm")
```




```
round(cor(wineData$alcohol, wineData$quality), 2)
```

```
## [1] 0.48
```

Created grouped bar chart and scatter plot to look at alcohol and quality. There is a moderate positive correlation of approximately (0.48) between alcohol and quality.



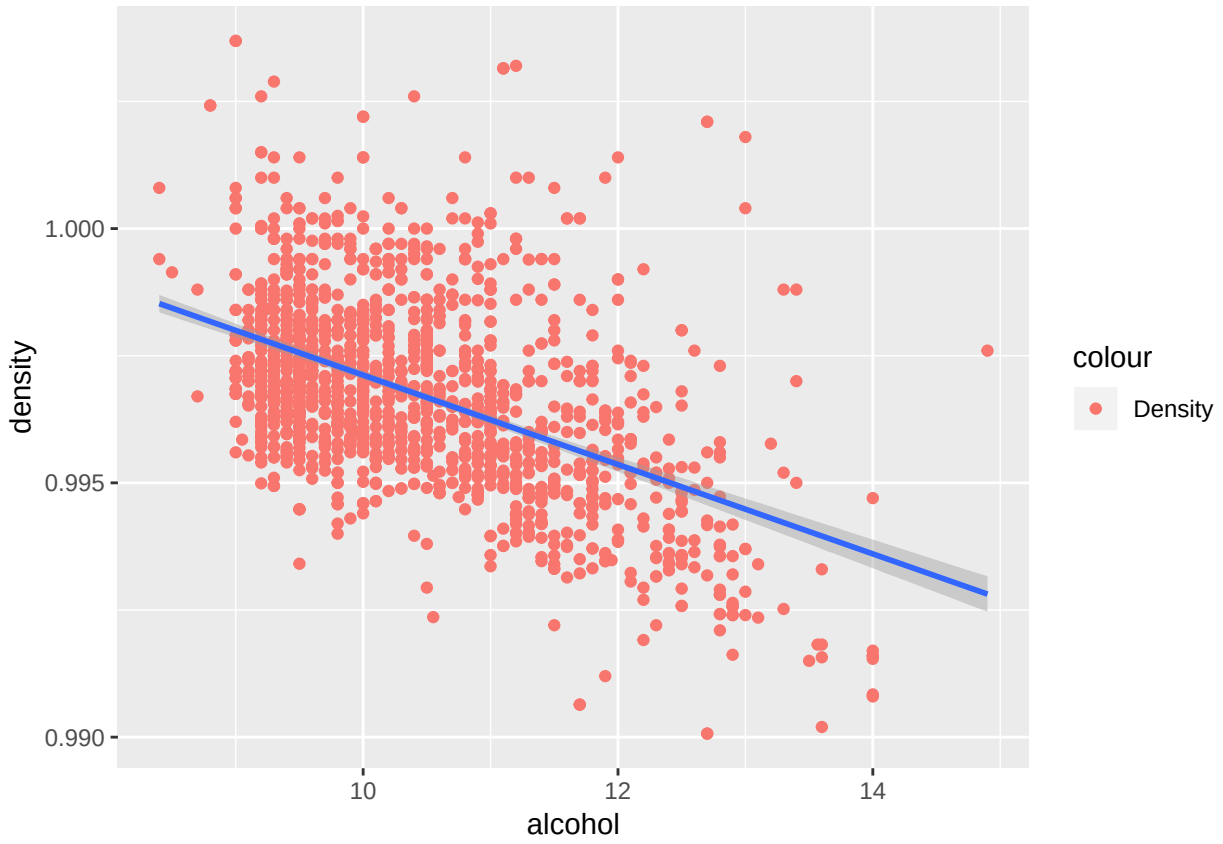

```
round(cor(wineData$quality, wineData$total.sulfur.dioxide), 2)
```

```
## [1] -0.19
```

I created a jittering scatter plot to view the relationship between quality and total sulfur dioxide. The data looks to like it is normally distributed. It has a weak correlation of apporximately -0.19.

```
ggplot(data = wineData, mapping = aes(x=alcohol, y=density)) +
  geom_point(aes(position = "jitter", color="Density")) +
  geom_smooth(method="lm")
```

```
## Warning: Ignoring unknown aesthetics: position
```



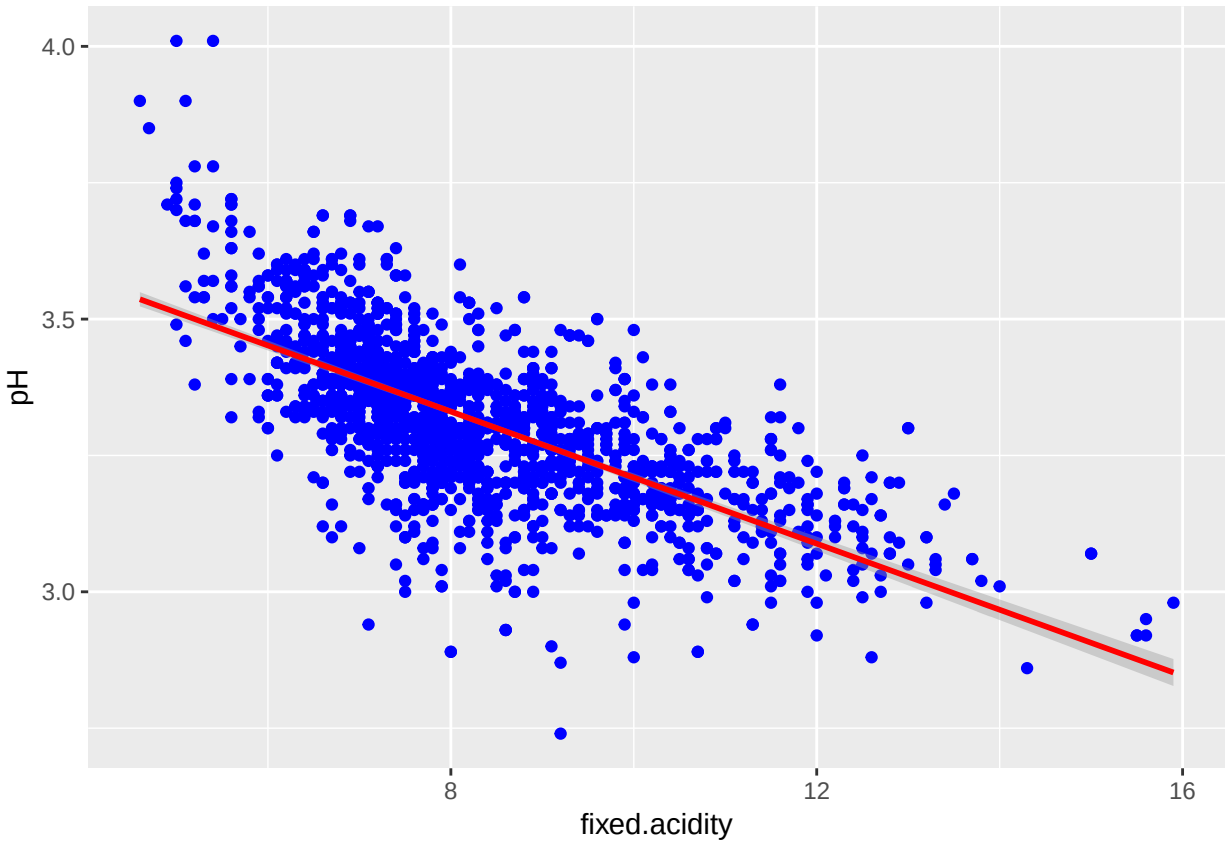

```
round(cor(wineData$alcohol, wineData$density), 2)
```

```
## [1] -0.5
```

This scatterplot looks at the relationship between alchol and density. As alcohol increases, density decreases. This inverse relationship has a moderate correlation of apporximately -0.5.

```
ggplot(data = wineData, mapping = aes(x=fixed.acidity, y=pH)) +
  geom_point(aes(position = "jitter"),color="blue") +
  geom_smooth(color='red',method="lm")
```

```
## Warning: Ignoring unknown aesthetics: position
```

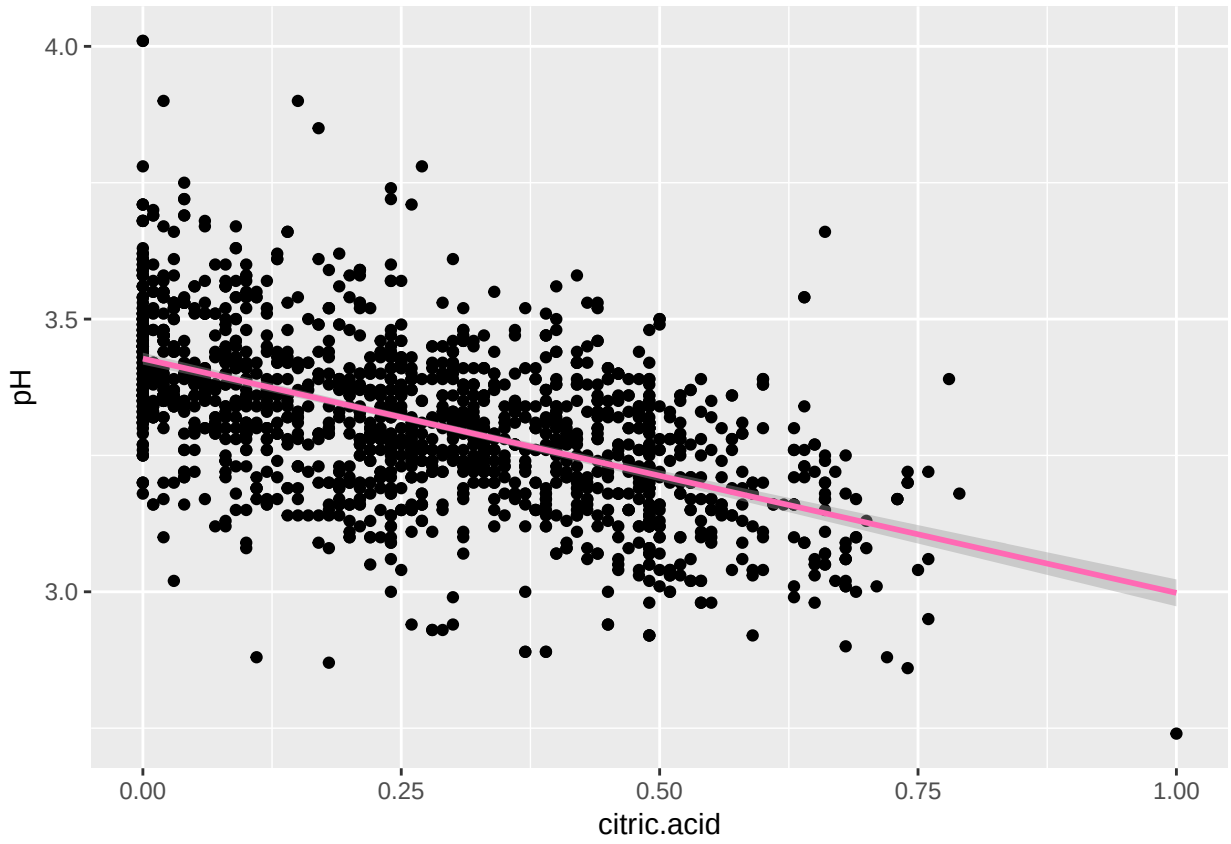



```
round(cor(wineData$fixed.acidity, wineData$pH), 2)
```

```
## [1] -0.68
```

Fixed acidity and pH also have an inverse relationship. As fixed acidity increases, pH decreases. The relationship has a strong correlation of apporximately -0.68.

```
ggplot(data = wineData, mapping = aes(x = citric.acid, y = pH)) +
  geom_point() +
  geom_smooth(color='hotpink',method="lm")
```

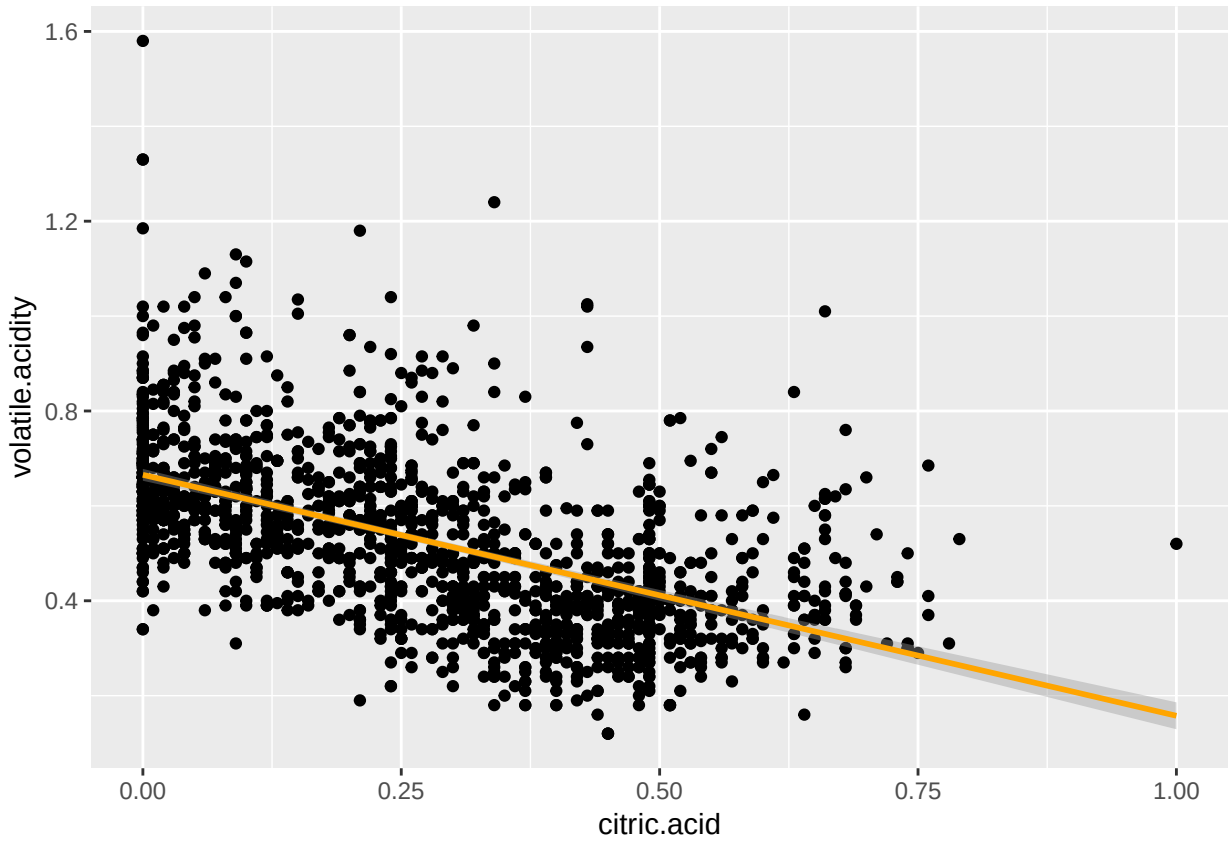



```
round(cor(wineData$citric.acid, wineData$pH), 2)
```

```
## [1] -0.54
```

As citric acid increases, pH decreases. This is an inverse relationship, with a strong correlation of apporximately -0.54. This is similar to the previous graph.

```
ggplot(data = wineData, mapping = aes(x = citric.acid, y = volatile.acidity)) +
  geom_point() +
  geom_smooth(color='orange',method="lm")
```

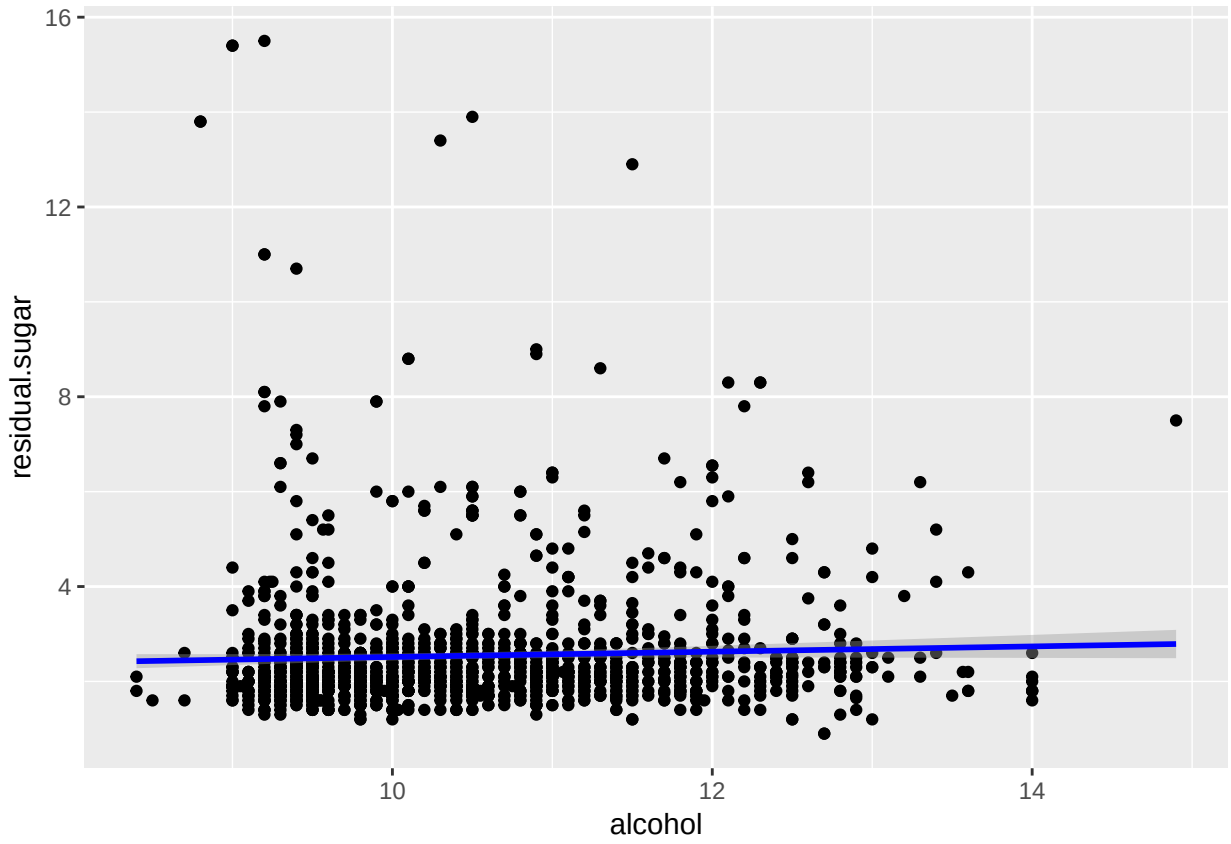



```
round(cor(wineData$citric.acid, wineData$volatile.acidity), 2)
```

```
## [1] -0.55
```

As citric acid increases, volatile acidity decreases. This is an inverse relationship, with a strong correlation of apporximately -0.55.

```
ggplot(data = wineData, mapping = aes(y=residual.sugar, x=alcohol)) +
  geom_point() +
  geom_smooth(color='blue',method="lm")
```

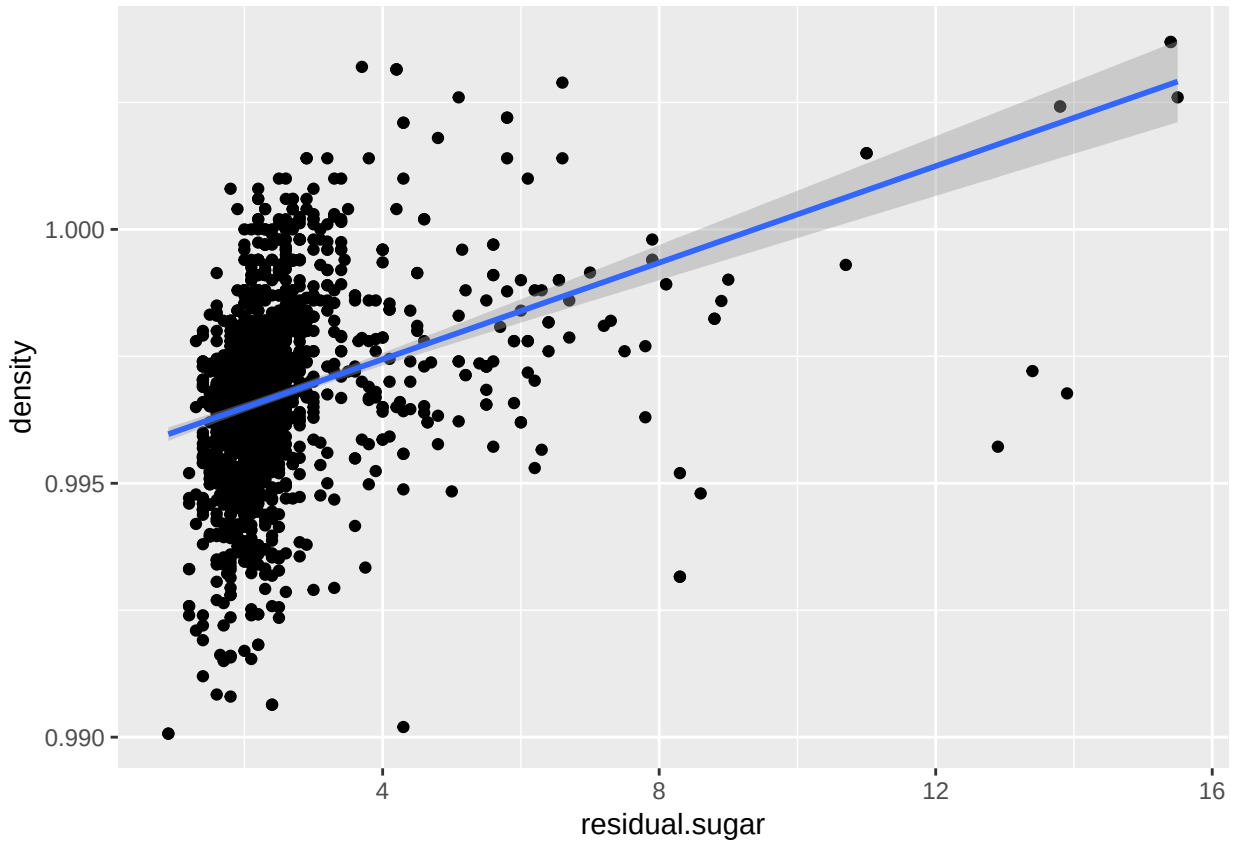



```
round(cor(wineData$alcohol, wineData$residual.sugar), 2)
```

```
## [1] 0.04
```

The percent of alcohol and residual sugar have a very weak correlation of apporximately 0.04.

```
ggplot(data = wineData, mapping = aes(x=residual.sugar, y=density)) +
  geom_point() +
  geom_smooth(method = "lm")
```



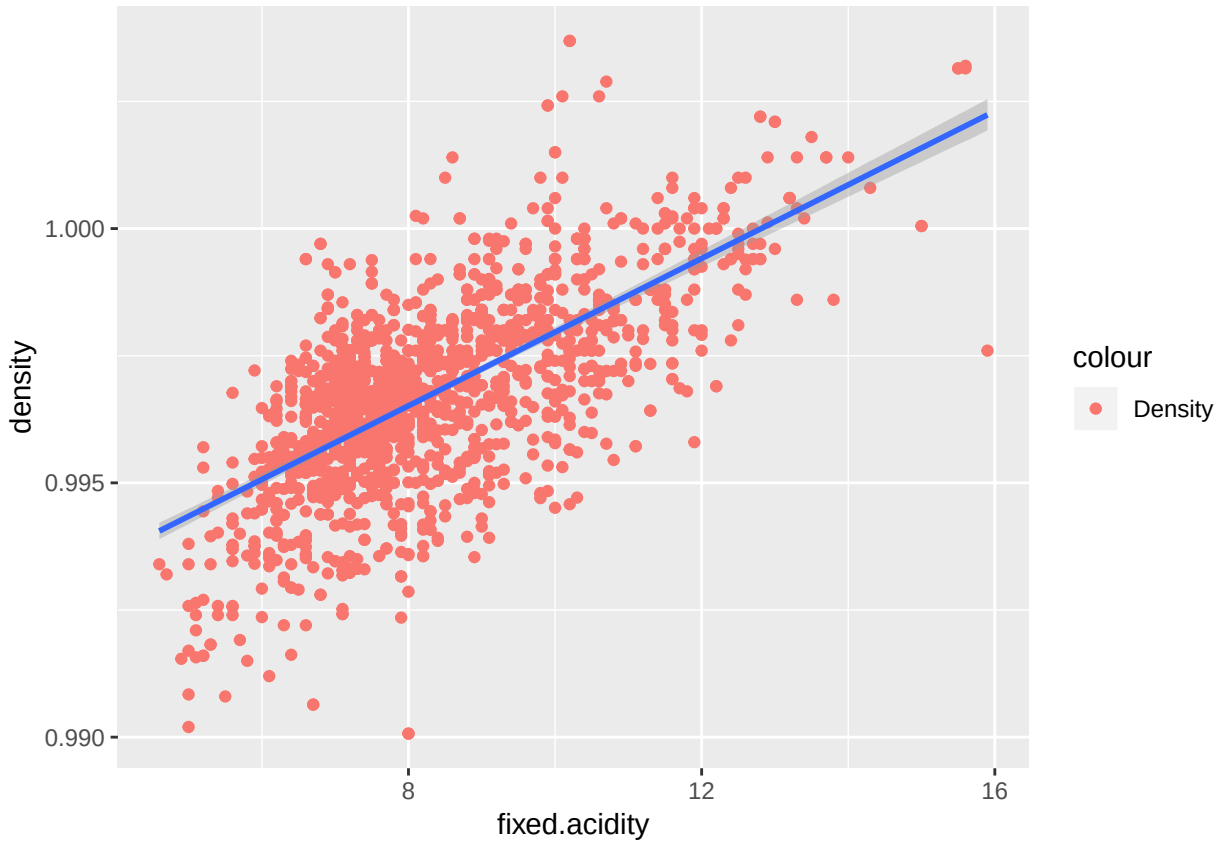

```
round(cor(wineData$residual.sugar, wineData$density), 2)
```

```
## [1] 0.36
```

As residual sugar increases density tends to increase. The correlation of this relationship is apporximately 0.36, which is considered moderate.

```
ggplot(data = wineData, mapping = aes(x=fixed.acidity, y = density)) +
  geom_point(aes(position = "jitter", color="Density")) +
  geom_smooth(method = "lm")
```

```
## Warning: Ignoring unknown aesthetics: position
```

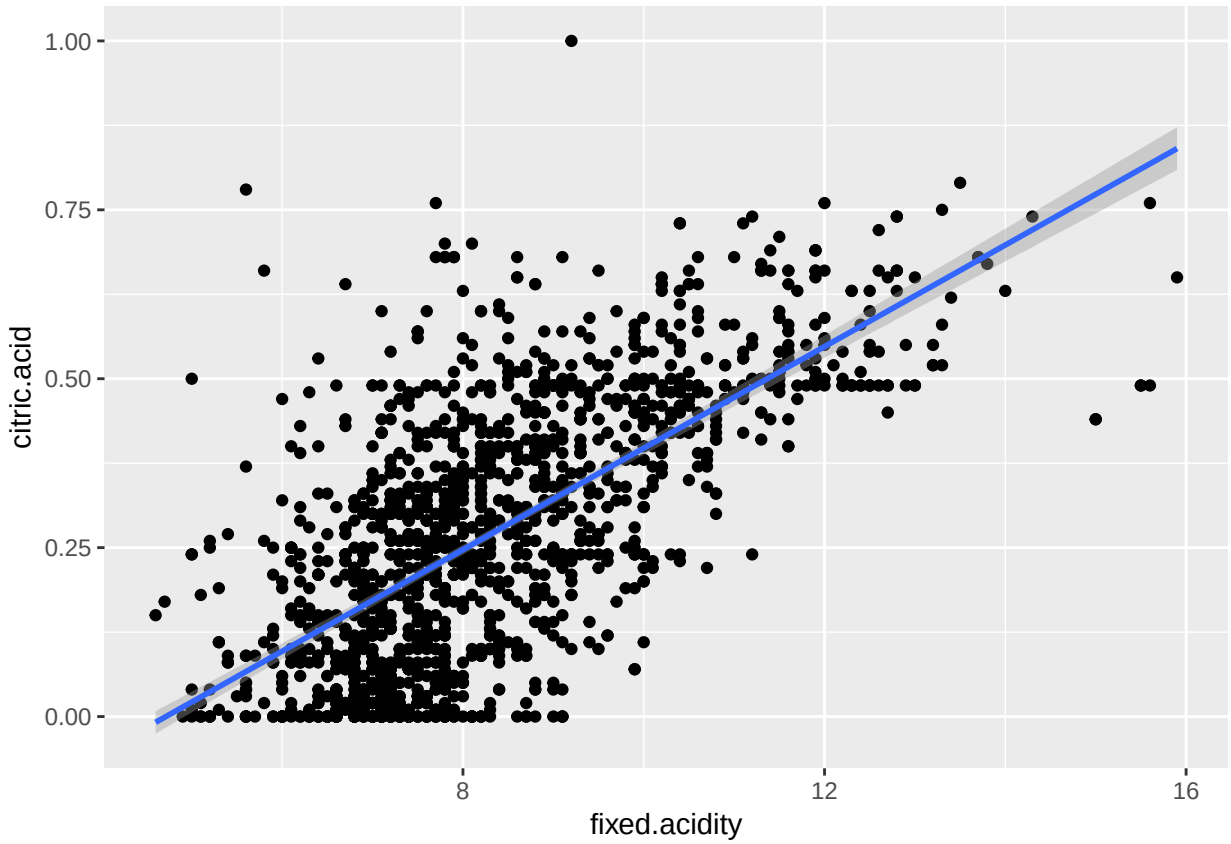



```
round(cor(wineData$fixed.acidity, wineData$density), 2)
```

```
## [1] 0.67
```

There is a direct relationship between fixed acidity and density. As fixed acidity increases, density increases. The relationship has a strong correlation of apporximately 0.67.

```
ggplot(data = wineData, mapping = aes(x=fixed.acidity, y=citric.acid)) +
  geom_point() +
  geom_smooth(method = "lm")
```

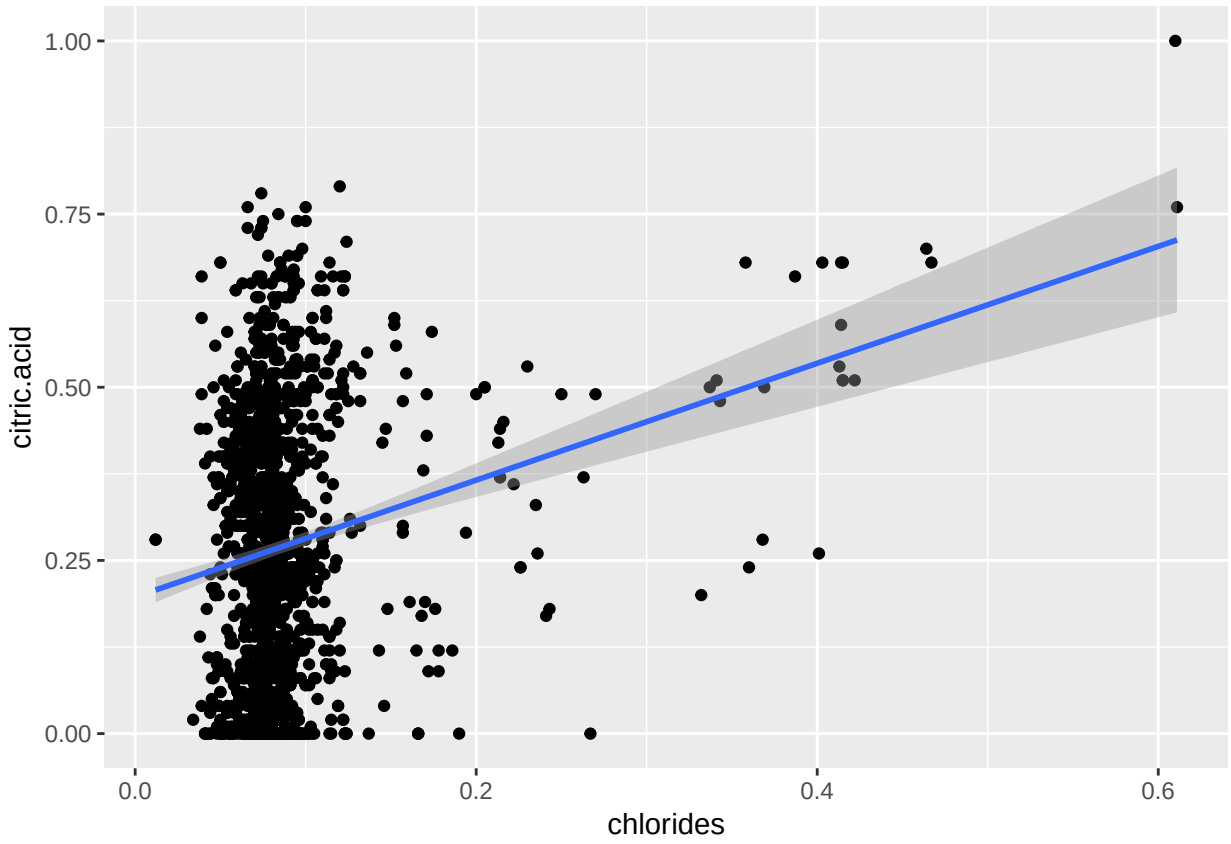



```
round(cor(wineData$fixed.acidity, wineData$citric.acid), 2)
```

```
## [1] 0.67
```

Fixed acidity and citric acid also have a direct relationship. As fixed acidity increases, citric acid increases. The relationship has a strong correlation of apporximately 0.67. One question I have is whether citric acids also fall into the category of fixed acids?

```
ggplot(data = wineData, mapping = aes(x=chlorides, y=citric.acid)) +
  geom_point() +
  geom_smooth(method = "lm")
```

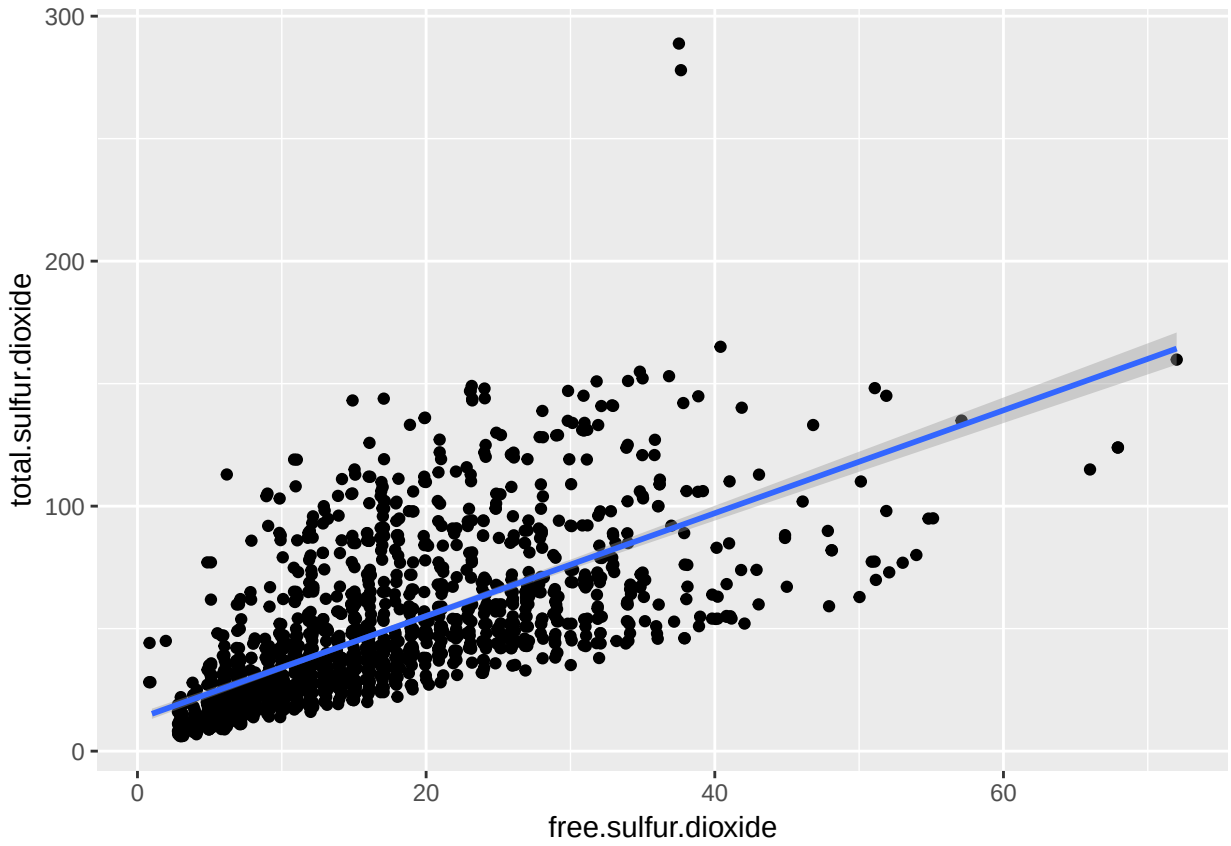



```
round(cor(wineData$chlorides, wineData$citric.acid), 2)
```

```
## [1] 0.2
```

There is a positive correlation between chlorides and citric acid. It is a weak correlation of apporximately -0.2.

```
ggplot(data = wineData, mapping = aes(x=free.sulfur.dioxide, y=total.sulfur.dioxide)) +
  geom_point(position = "jitter") +
  geom_smooth(method = "lm")
```




```
round(cor(wineData$free.sulfur.dioxide, wineData$total.sulfur.dioxide), 2)
```

```
## [1] 0.67
```

There is a direct relationship between free sulfur dioxide and total sulfur dioxide. As free sulfur dioxide increases, total sulfur dioxide increases. The relationship has a strong correlation of apporximately 0.67. This makes sense since free sulfur dioxide makes up the free forms of SO_2 existing in total sulfar dioxide.

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?

I was most interested in alcohol and quality. My initial assumption was there would be a direct relationship between alcohol and quality. Although there was a moderate positive correlation (0.48), I thought the correlation would be stronger.

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

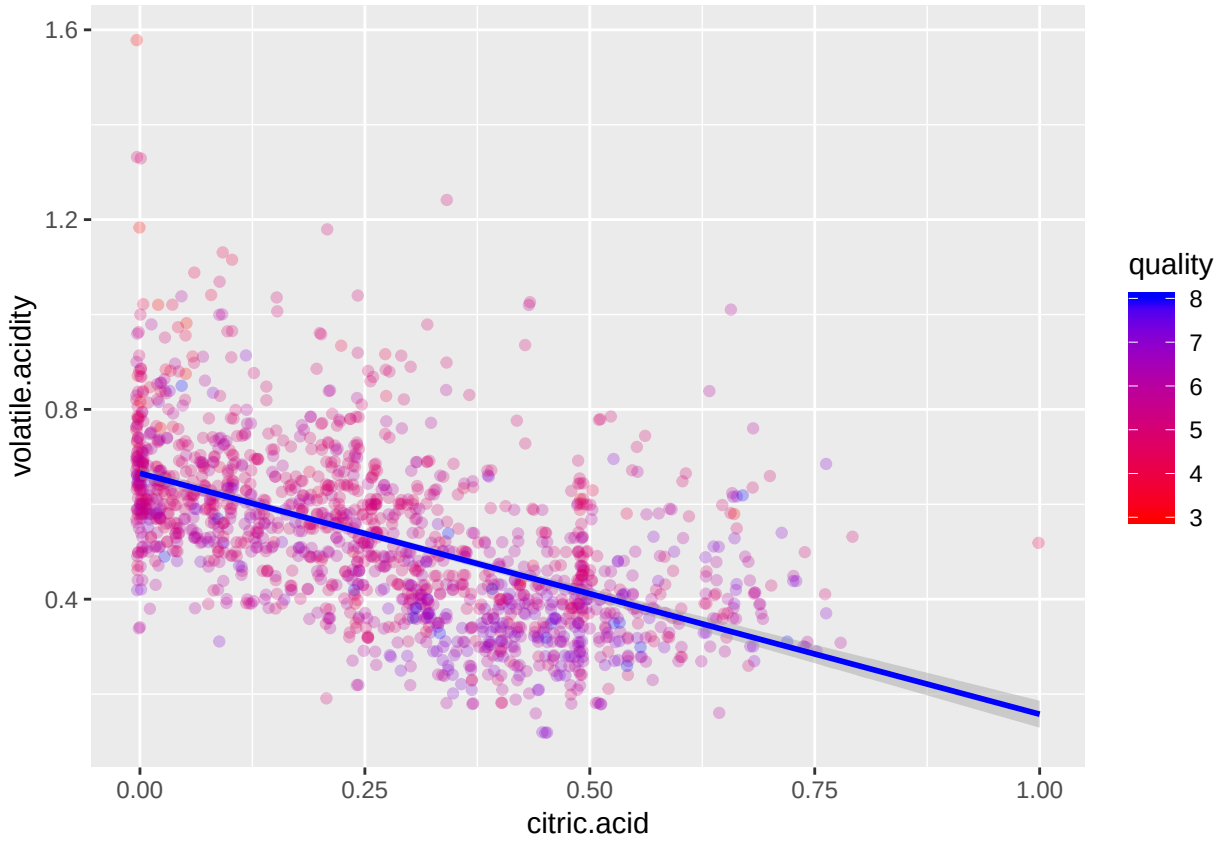
The correlation between fixed acidity and pH is approximately -0.68. As fixed acidity increases, pH decreases. This makes sense considering the pH scale. Substances are considered more acidic as their pH approaches zero. I found it interesting that the pH range of wines in this study are equivalent to the pH of orange juice, soda, and acid rain.

I also thought it was interesting that there is not a strong relationship between percent of alcohol and residual sugar. My initial thought was wine with higher alcohol content had less sugar remaining. However, the trend shows little change in residual sugar as alcholol increases. This leads me to question how much sugar is placed into the various wines prior to fermintation? Could higher alcohol content receive higher amounts of sugar? Is there a formula to follow to acheive a desired alcohol level which leads to similar amounts of residual sugar?

What was the strongest relationship you found?

The strongest correlation I found exists between fixed acidity and pH (-0.68).

Multivariate Plots Section




```
r <- cor(wineData$citric.acid, wineData$volatile.acidity)
rSquare <- r^2
r
```

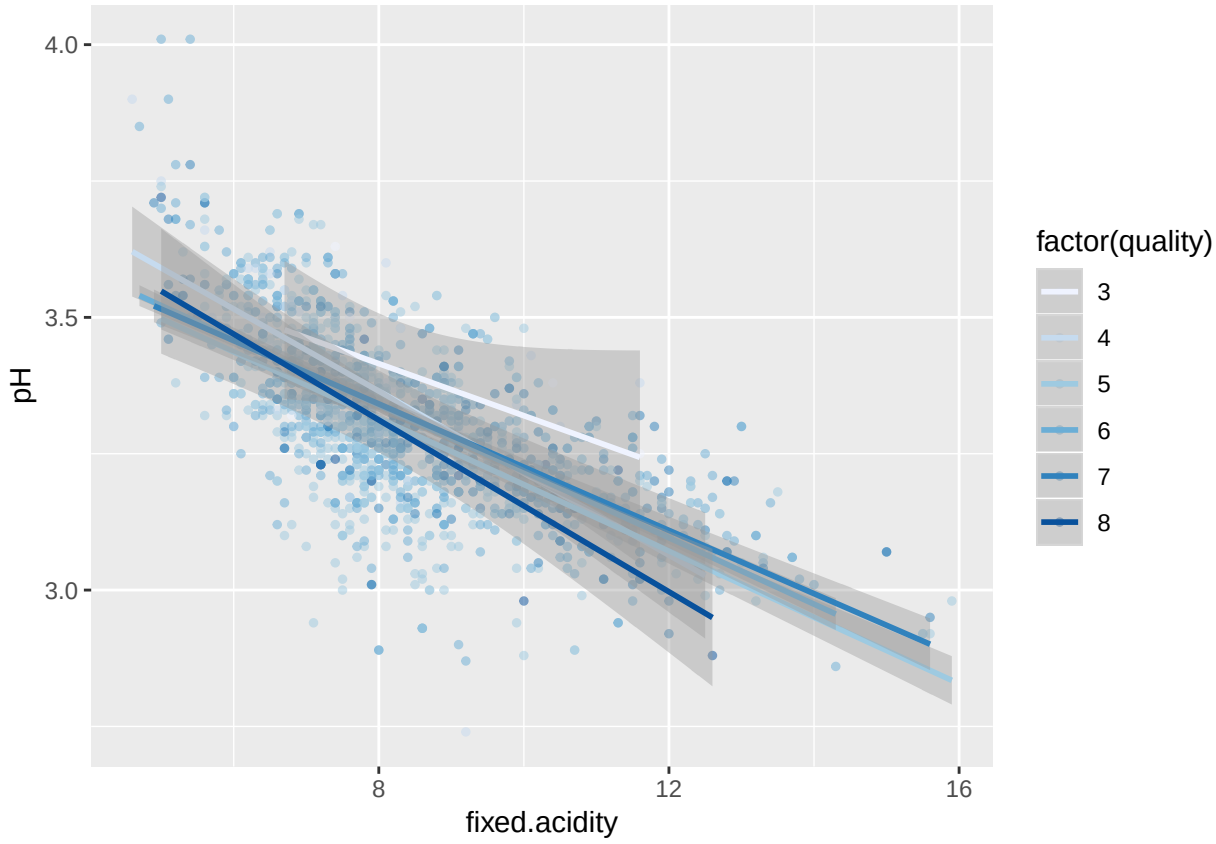
```
## [1] -0.5524957
```

```
rSquare
```

```
## [1] 0.3052515
```

I created a scatterplot to display the relationship between citric acid, volatile acidity, and quality. Since the correlation coefficient r for citric acid and volatile acidity is approximately 0.55, then its correlation of determination r^2 is approximately 0.31. If *r-squared* is 0.31 then it means 31% of variations in volatile acidity are explained by the citric acid in this model.

```
ggplot(data = wineData, mapping = aes(x=fixed.acidity, y=pH, color=factor(quality))) +
  geom_point(alpha = 0.5, size = 1) +
  geom_smooth(method = "lm") +
  scale_color_brewer(type = "seq")
```




```
r_2 <- cor(wineData$fixed.acidity, wineData$pH)
rSquare_2 <- r_2^2
r_2
```

```
## [1] -0.6829782
```

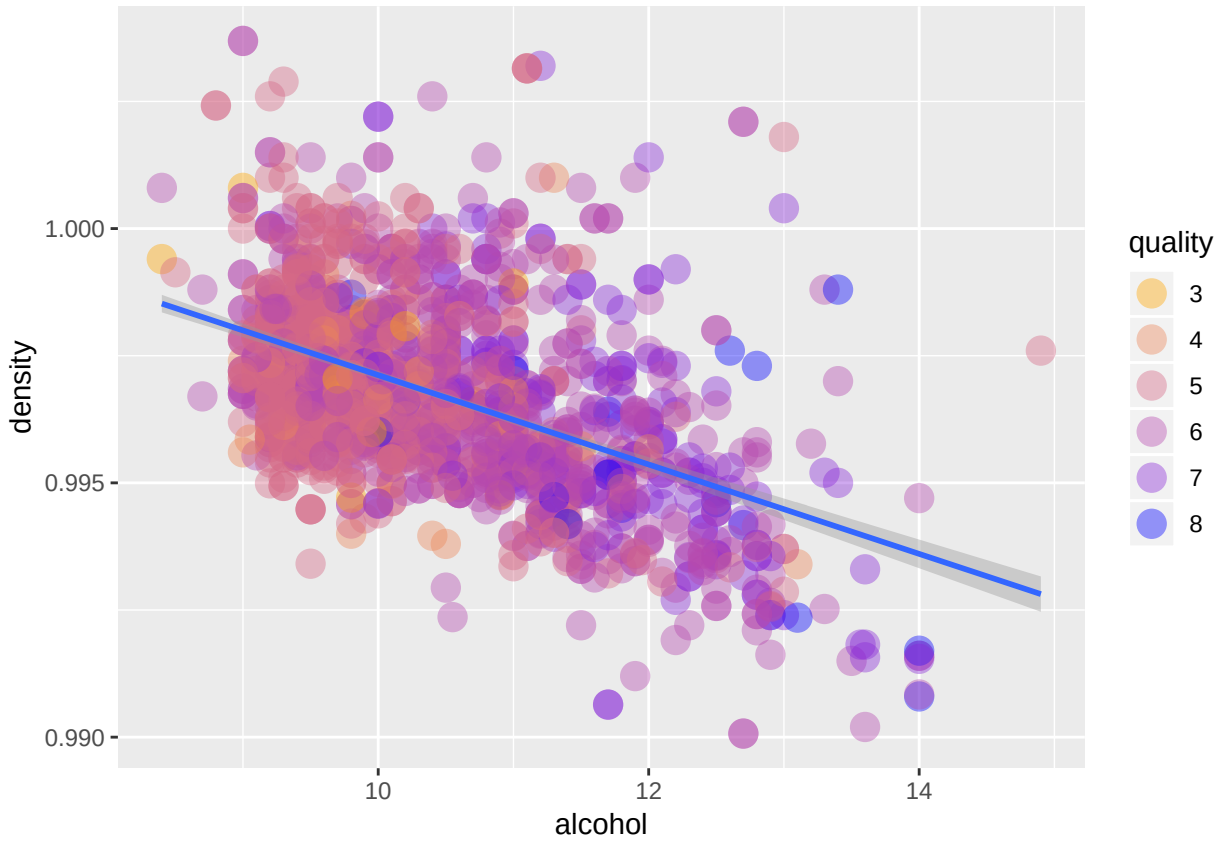
```
rSquare_2
```

```
## [1] 0.4664592
```

I created a scatter plot to look at the relationship between fixed acidity & pH, and grouped the dots by quality. Since the correlation coefficient r for fixed acidity and pH is approximately 0.68, then the correlation of determination r^2 is approximately 0.47. If *r-squared* is 0.47 then it means 47% of variations in pH are explained by the fixed acidity in this model.

```
ggplot(data = wineData, mapping = aes(x=alcohol, y=density)) +
  geom_point(mapping = aes(position = "jitter", color= quality), alpha = .4, shape = 16, size = 5) +
  guides(colour = guide_legend()) +
  scale_color_gradient(low="blue", high="orange", trans = 'reverse') +
  geom_smooth(method="lm")
```

```
## Warning: Ignoring unknown aesthetics: position
```

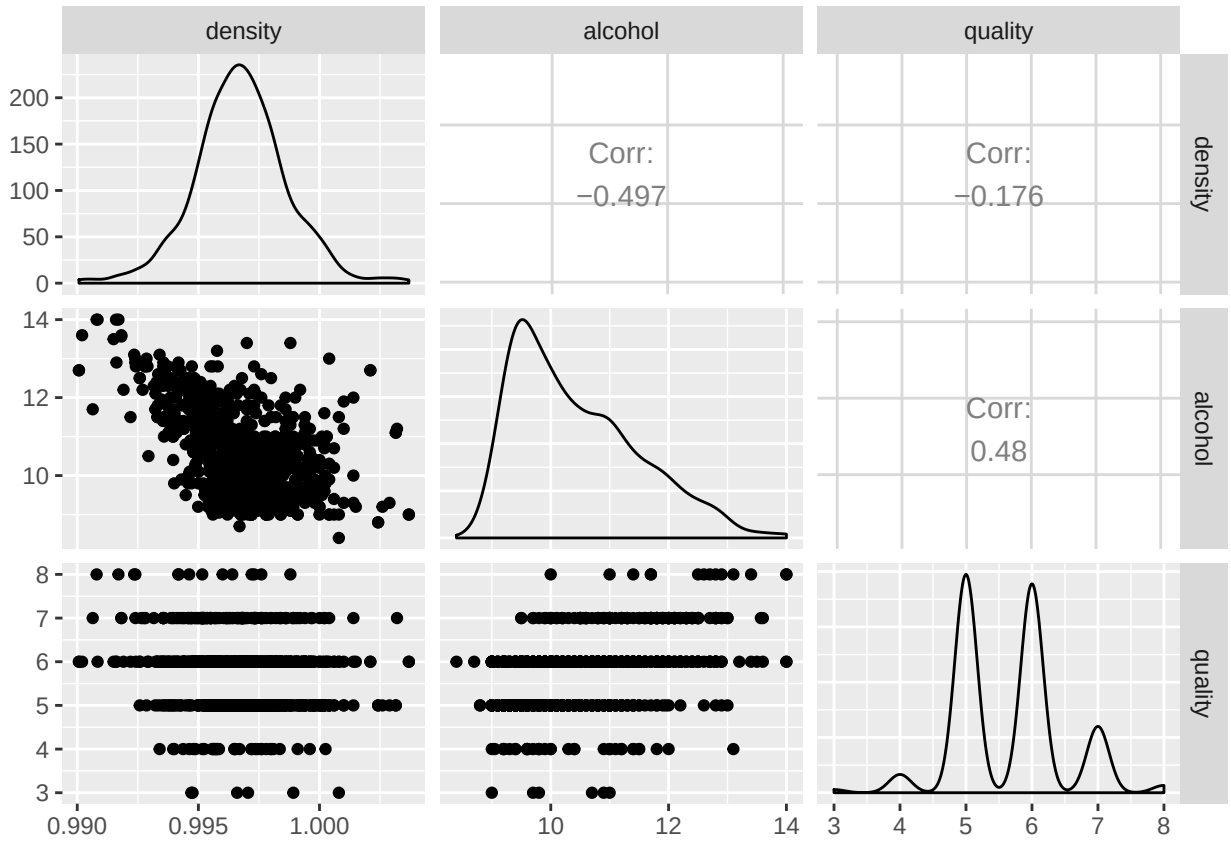


I had to set *scale_color_gradient* low to blue, and high to orange in order to get my intended result of 3 set to orange, and 8 set to blue. An oddity happened after I set *trans* to reverse, the values reversed, but the colors did too.

```
set.seed(1599)
wine_subset <- wineData[c(8,11,12)]
names(wine_subset)
```

```
## [1] "density" "alcohol" "quality"
```

```
ggpairs(wine_subset[sample.int(nrow(wine_subset), 1000),])
```




```
r_3 <- cor(wineData$alcohol, wineData$density)
rSquare_3 <- r_3^2
r_3
```

```
## [1] -0.4961798
```

```
rSquare_3
```

```
## [1] 0.2461944
```

The two visualizations above show the relationship between alcohol, density, and quality. The box plot provided a helpful visualization while *ggpairs()* provided a matrix of the plots and provided their correlation coefficients quickly.

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 scatterplot featuring citric acid, volatile acitidy, and quality shows that as citric acid increases, volatile acitidy tends to decrease, and quality tends to increase. While the scatterplot featuring fixed acidity, pH, and quality shows that as fixed acidity increases pH and quality tend to decrease.

The graphs featuring alcohol, density, and quality show that as alcohol content increases, density tends to decrease, and quality tends to increase. There is a weak correlation between density and quality (-0.159), while there are moderate correlations in both alcohol & density (-0.489), and alcohol & quality (0.472).

Were there any interesting or surprising interactions between features?

I had trouble finding a simple function that calculates r-squared. What I really wanted to do is place a regression line on a scatterplot and place the r-squared value on the graph.

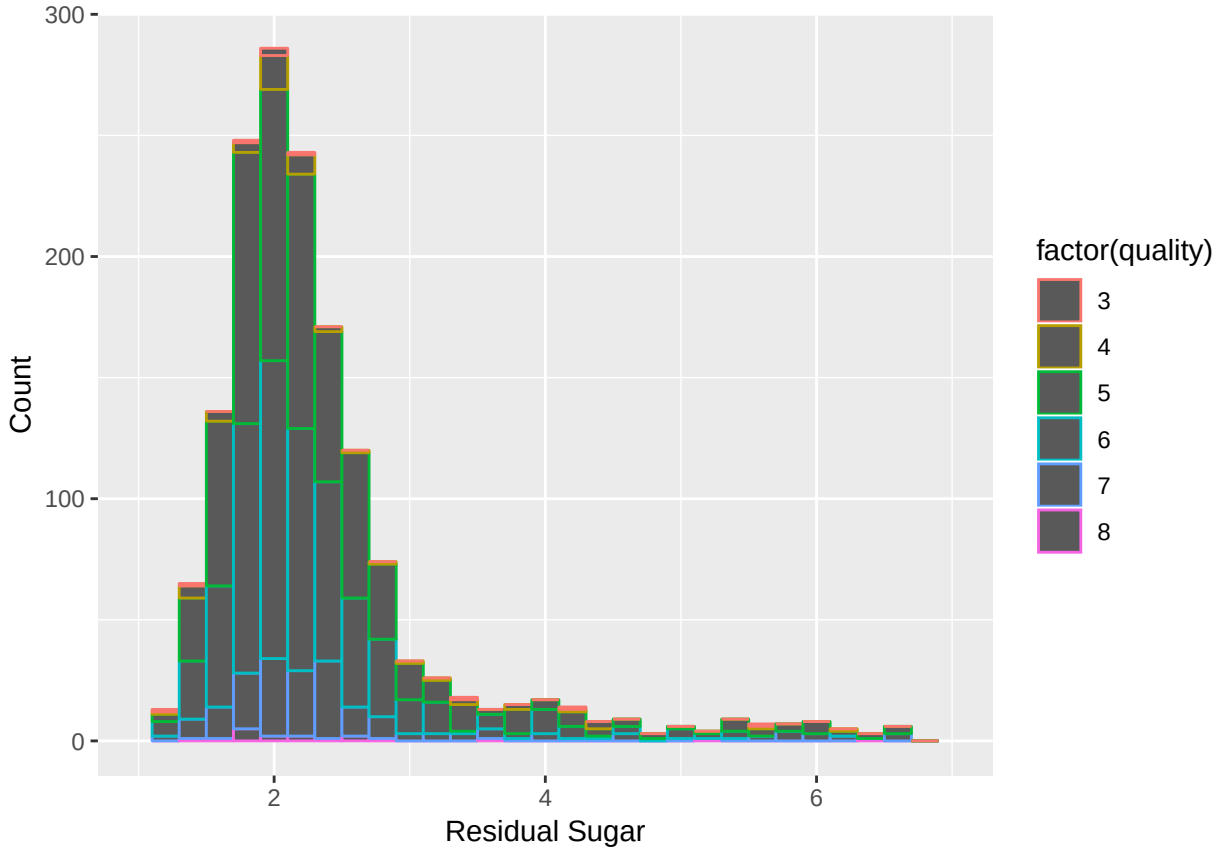
I found it interesting that while some items increased others decreased (e.g. alcohol, density, and quality). I found some additional anomalies that are noted in the final plot section under plot three.

Final Plots and Summary

Plot One

```
ggplot(data = wineData, mapping = aes(x=residual.sugar, color=factor(quality))) +
  geom_histogram(binwidth = 0.2) +
  xlim(1,7) +
  xlab("Residual Sugar") +
  ylab("Count")
```

```
## Warning: Removed 31 rows containing non-finite values (stat_bin).
```




```
summary(wineData$residual.sugar)
```

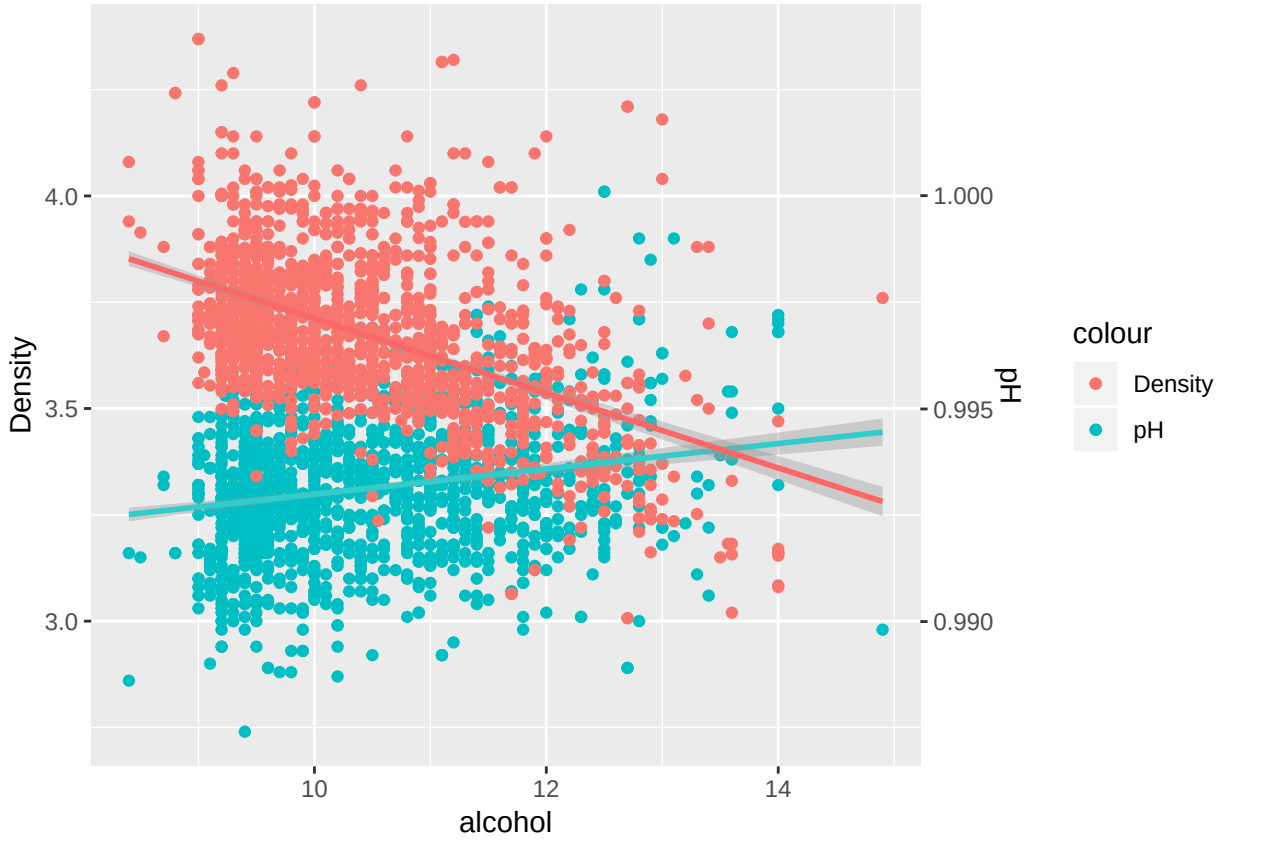
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

Description One

As stated in the *univariate plot section*, this histogram is skewed right. The residual sugar in most red wines is approximately 2.0 g/dm^3. The median (2.2) and mean (2.539) are pulled to the left. I used *xlim()* to remove outliers to create a cleaner visual. This time, I also grouped the data by quality. Since quality followed a normal distribution in the *univariate plot section*, we can see the majority of wines recieved a quailty rating of 5 or 6. The color coding in this graph helps to uncover to uncover this.

Plot Two

```
ggplot(data = wineData, mapping = aes(x=alcohol)) +
  geom_point(aes(y=pH,color="pH")) +
  geom_point(aes(y=(density-.96)*100,color="Density")) +
  geom_smooth(aes(y=pH), color="#33CCCC", method = "lm") +
  geom_smooth(aes(y=(density-.96)*100), color="#FF6666", method = "lm") +
  scale_y_continuous(name="Density",sec.axis = sec_axis(~((./100)+.96), name="pH"))
```



Description Two

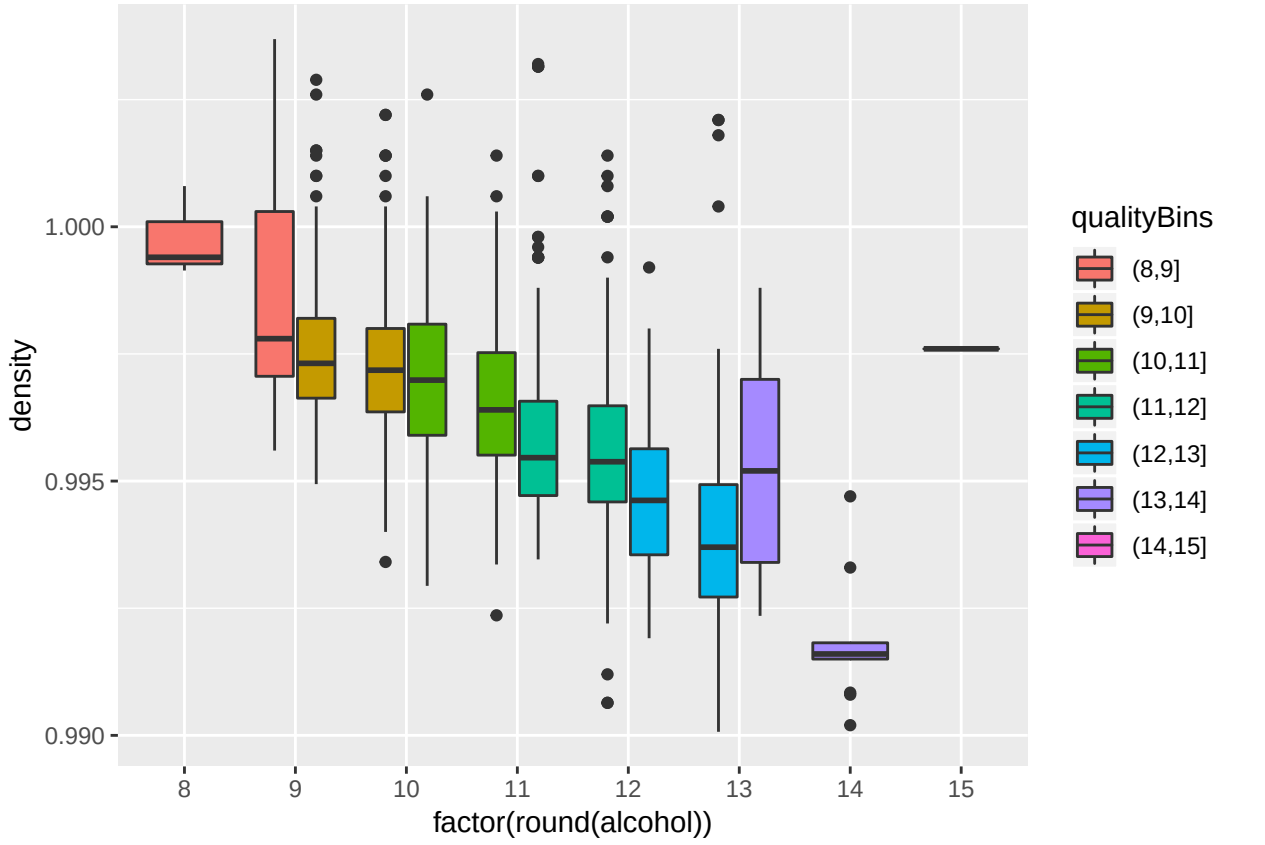
This graph displays two different relationships. A direct relationship between alcohol & density, and an inverse relationship between alcohol & pH. As alcohol increases, density decreases and pH increases. It can also be inferred that as pH increases, density decreases. This is also verified in the simple scatterplot matrix found in the *bivariate plots section*.

Throughout this entire project, I had the most fun working on this graph. The reason is I plotted a second variable (pH) along the vertical axis and it was not as simple as when working with software like excel. I had

to use the *scale_y_continuous()* function which required a specific formula (i.e. *~((./100)+.96)*) to scale the values for this axis to a desired set of numbers.

Plot Three

```
wineData$qualityBins = cut(wineData$alcohol,
                           c(8:16))
ggplot(wineData, mapping = aes(x = factor(round(alcohol)), y = density)) +
  geom_boxplot(aes(fill = qualityBins))
```



Description Three

As stated in the *multivariate plots section*, as alcohol content increases, density tends to decrease, and quality tends to increase. I wanted to look at this data set using a different type of graph and chose a box plot. I noticed there are multiple dots below and above the min and max whiskers. I have never observed this before. I found out that some of these additional dots are outliers. Outliers should carefully be considered before being removed [4]. Outliers, lie 1.5 x IQR (interquartile range) below the 1st or above the 3rd quartile ranges. Running *IQR()* on pH provides one value for the entire range, and I could easily multiply this by 1.5. However, I believe I need to calculate multiple IQR values since I grouped data on citric.acid for this visualization. This is something I would like to research further.

Reflection

I started this project with the thought that I would be most interested in determining whether or not there was a relationship between alcohol & quality. As I started working on the project, I became most interested in the relationship between fixed acidity and pH. I was able to observe the inverse relationship and directly relate the pH values to common items I typically consume. The first section dealt with simple graphs, which were not very interesting, but as the sections grew more complex, the visualizations became more interesting. One item I found helpful when working with *geom_point()* was to set position to *jitter*. This feature spread the data points slightly and created some random noise which helped clean up the visuals by preventing *overplotting*. I feel confident in creating graphs and researching the visuals I want to build, but I need to get better at looking for patterns, playing with statistical formulas, understanding and explaining the results.

Resources

1. List of Wine Attributes
2. Corrplot
3. Simple Scatterplot Matrix
4. Outliers
5. *Wickham, H., & Grolemund, G. (2017). R for data science: Import, tidy, transform, visualize and model data. Sebastopol: OReilly.*