Red wine quality dataset exploration and summary by Mohamed Zeghlache

> **General information about the dataset**: This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent). The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods. Number of Instances: 1599. Number of Attributes: 11 + output attribute

Univariate Plots Section

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

as we can see, this dataset contains 1599 observations(rows) and 13 variables(attributes) but the first variable is just an index and we don't need it in our analysis,so i'll delete it from the dataset.but the rest of the variables are just as described in the introduction.

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

nice, now the dataset looks better.

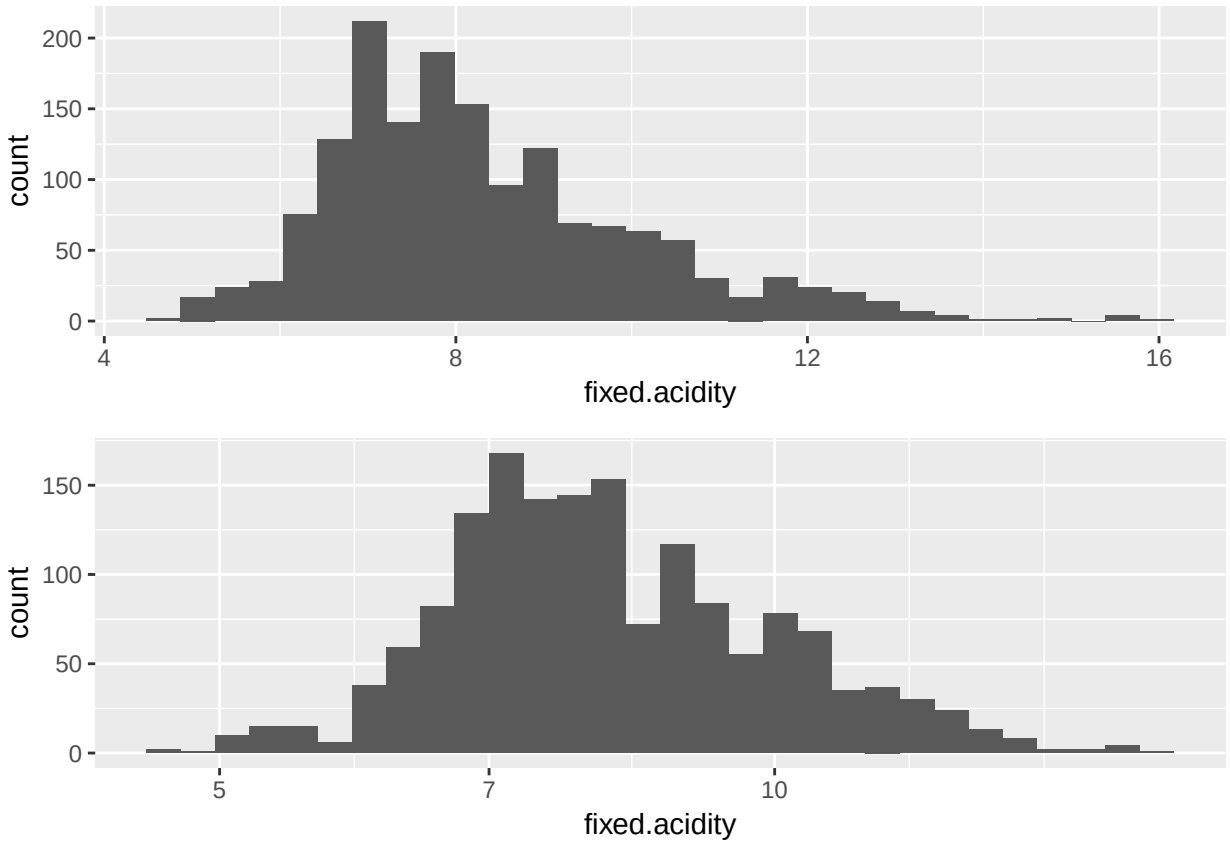
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
```

```
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

from this result, i can already see that some of these variables have outliers, but we'll get a closer look on that using plots.

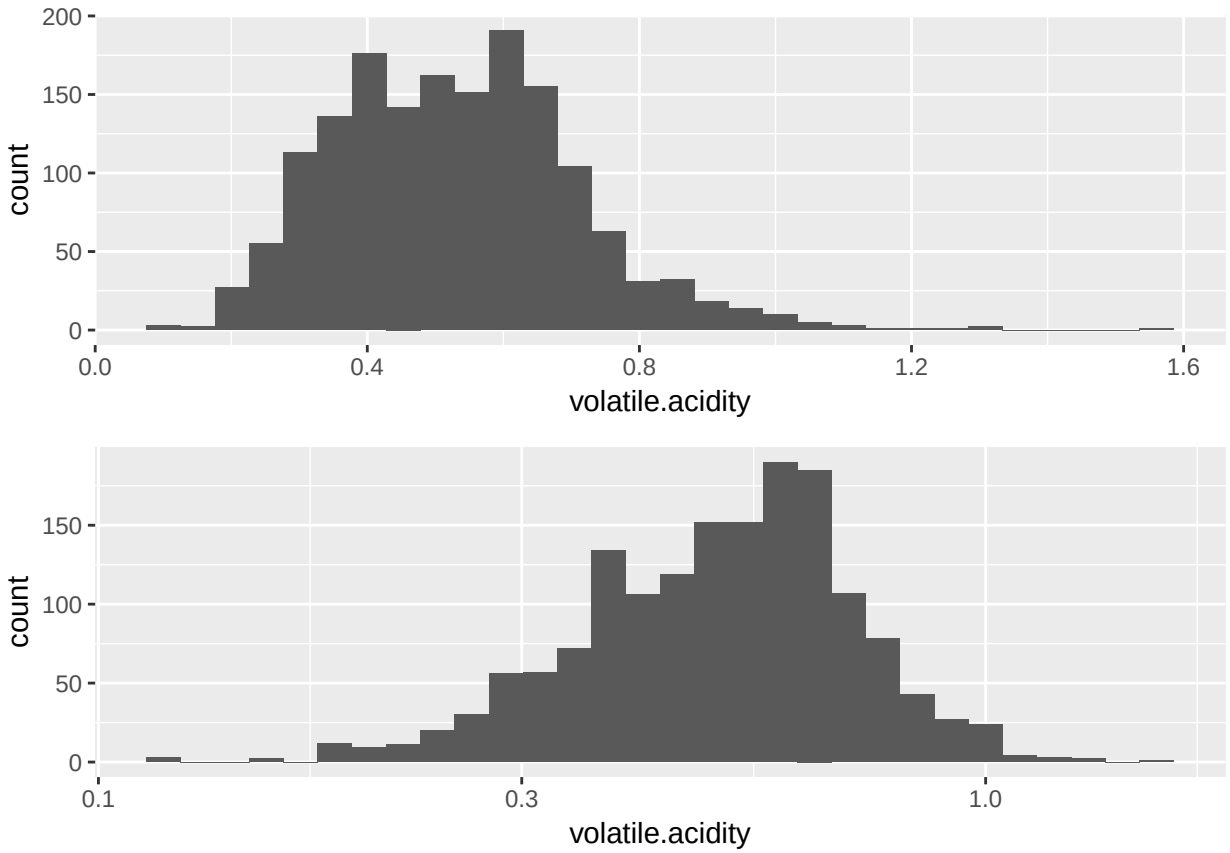
ploting

fixed.acidity



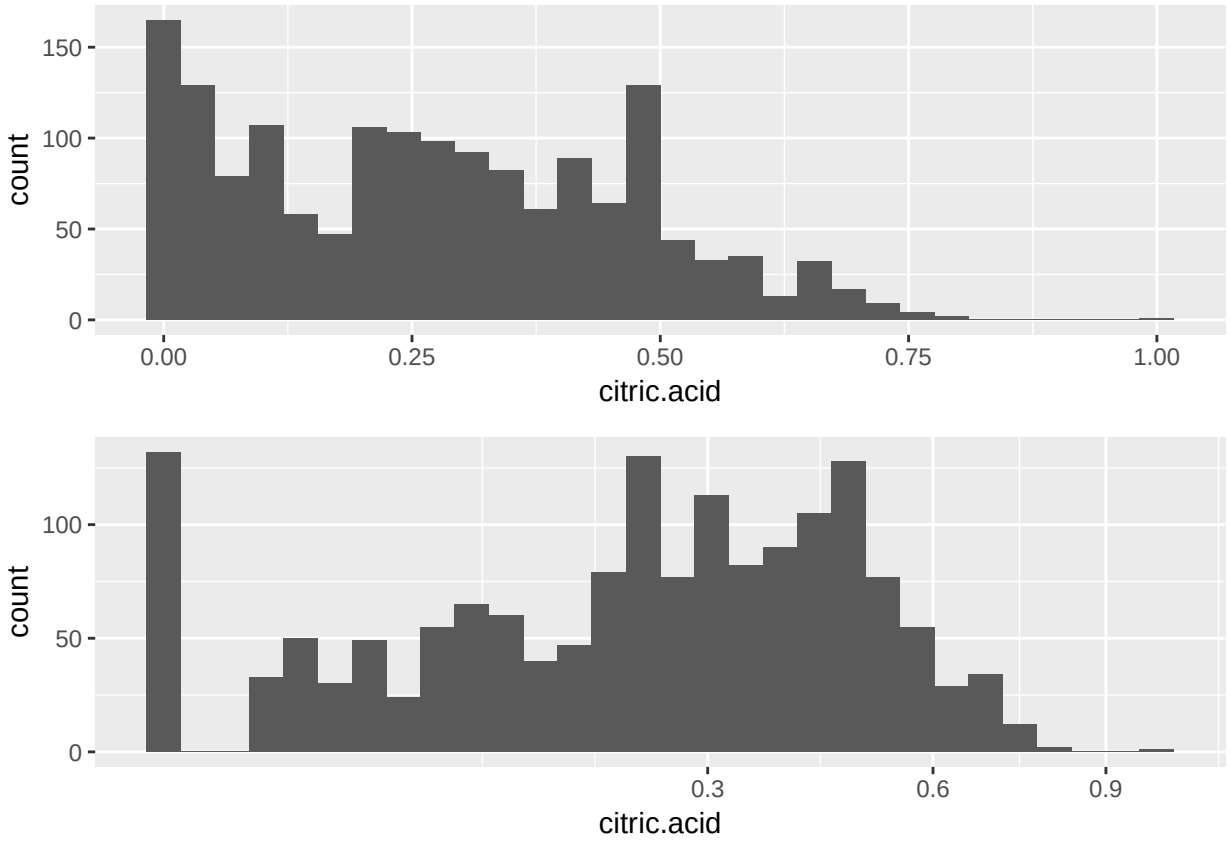
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right.so, i applied a log scale to the x axis to make the distribution look more normal(the one in the bottom).

volatile.acidity



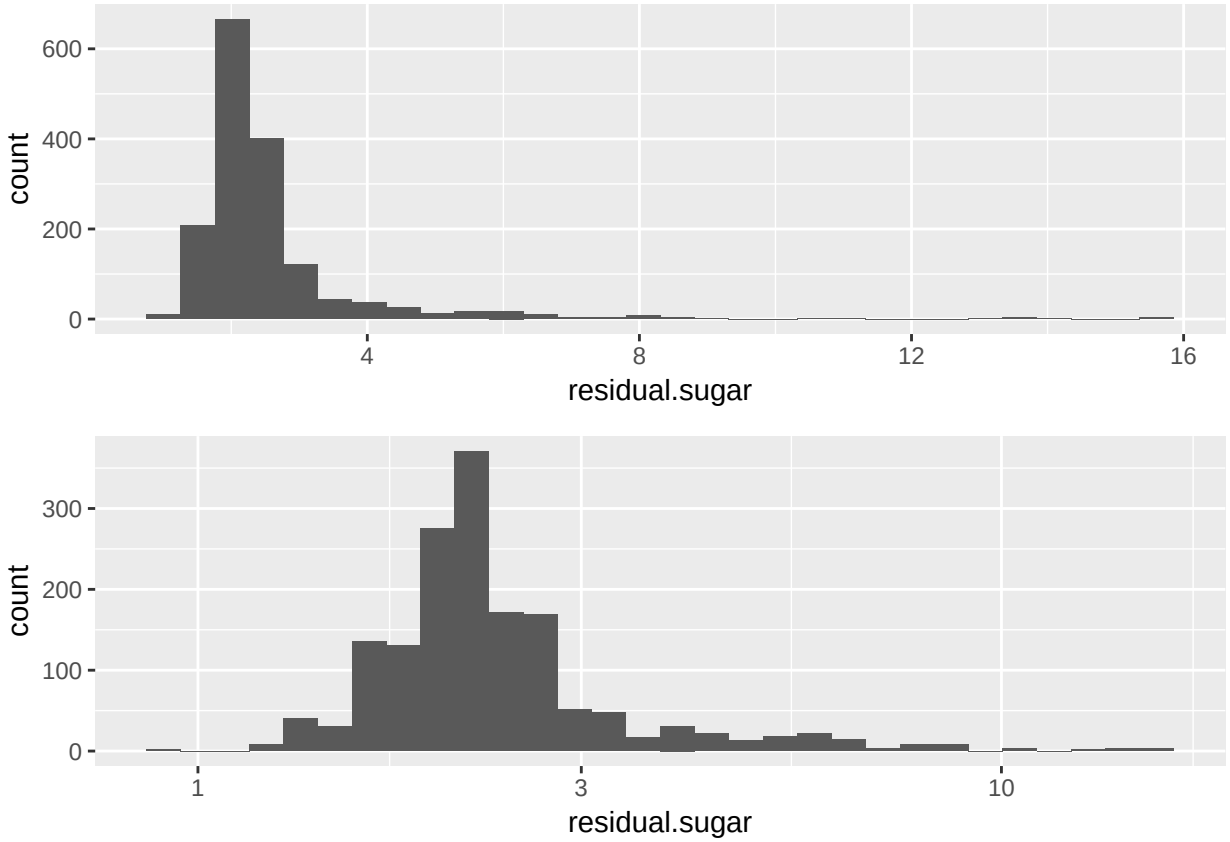
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right.so, i applied a log scale to the x axis to make the distribution look more normal(the one in the bottom).

citric.acid



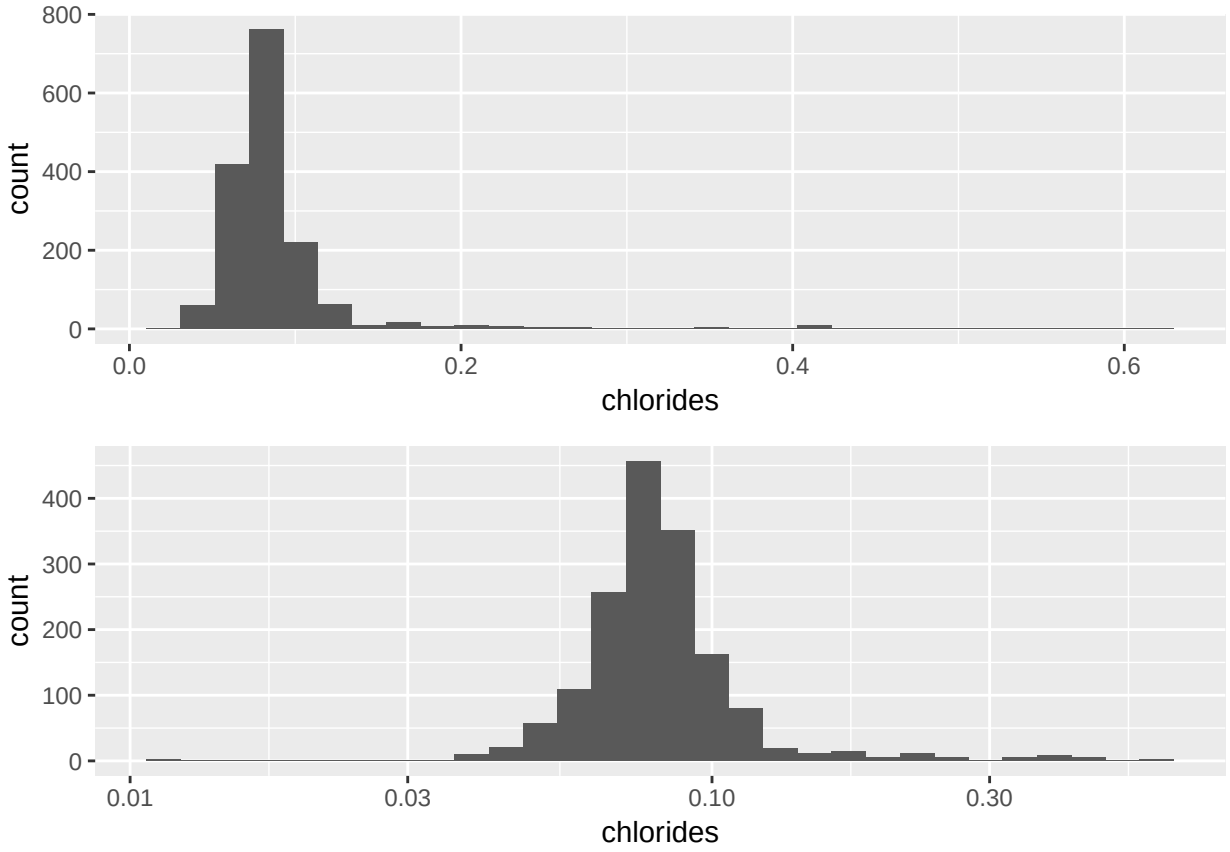
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right.so, i applied a square root scale to the x axis to make the distribution look more normal(the one in the bottom).

residual.sugar



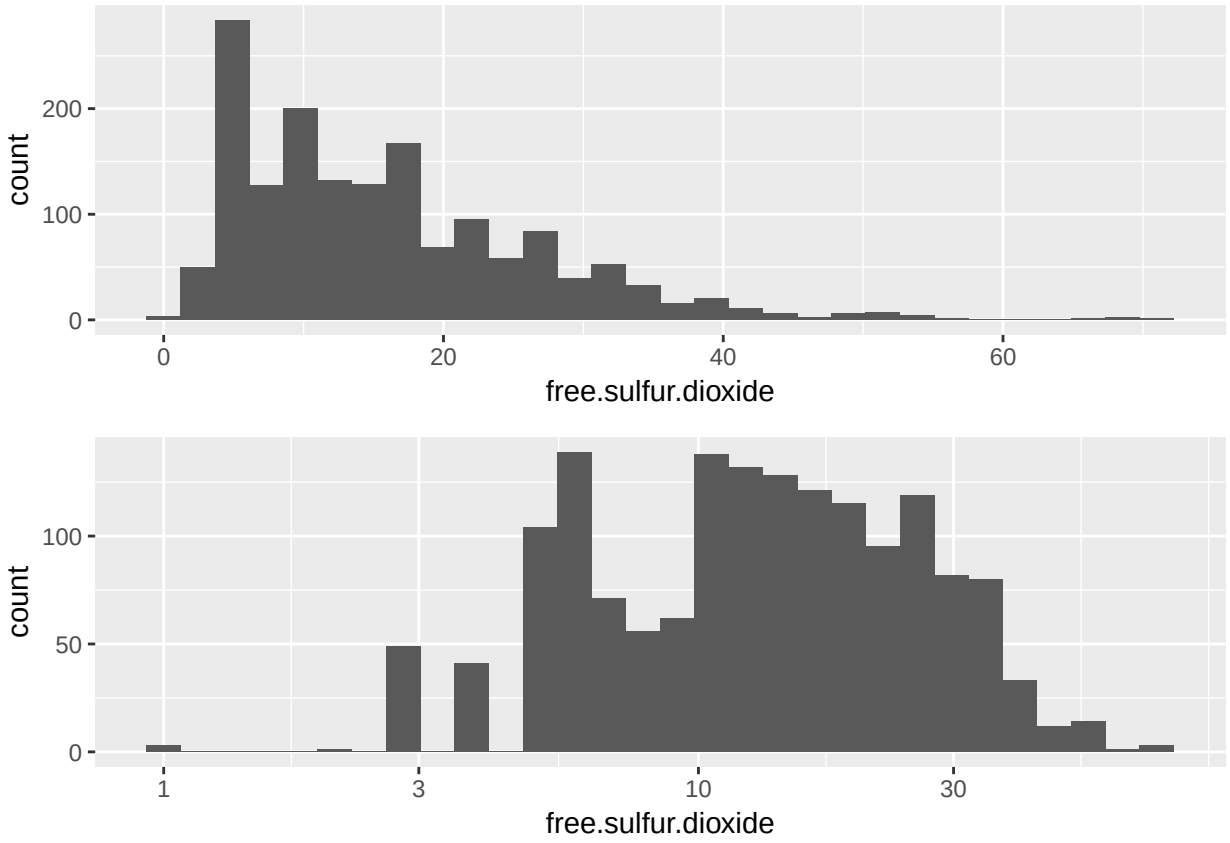
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right and has outliers.so, i applied a log10 scale to the x axis to make the distribution look more normal(the one in the bottom).

chlorides



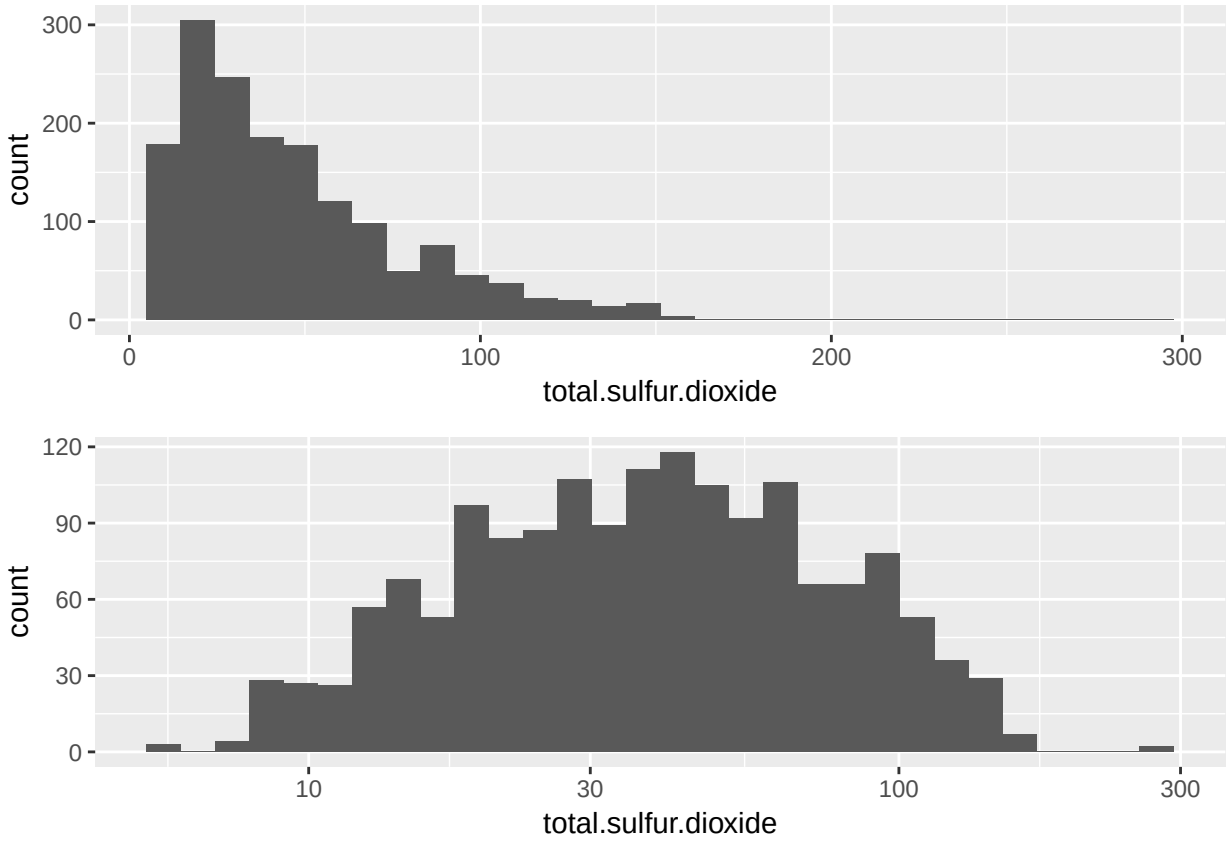
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right and has outliers.so, i applied a log10 scale to the x axis to make the distribution look normal(the one in the bottom).

free.sulfur.dioxide



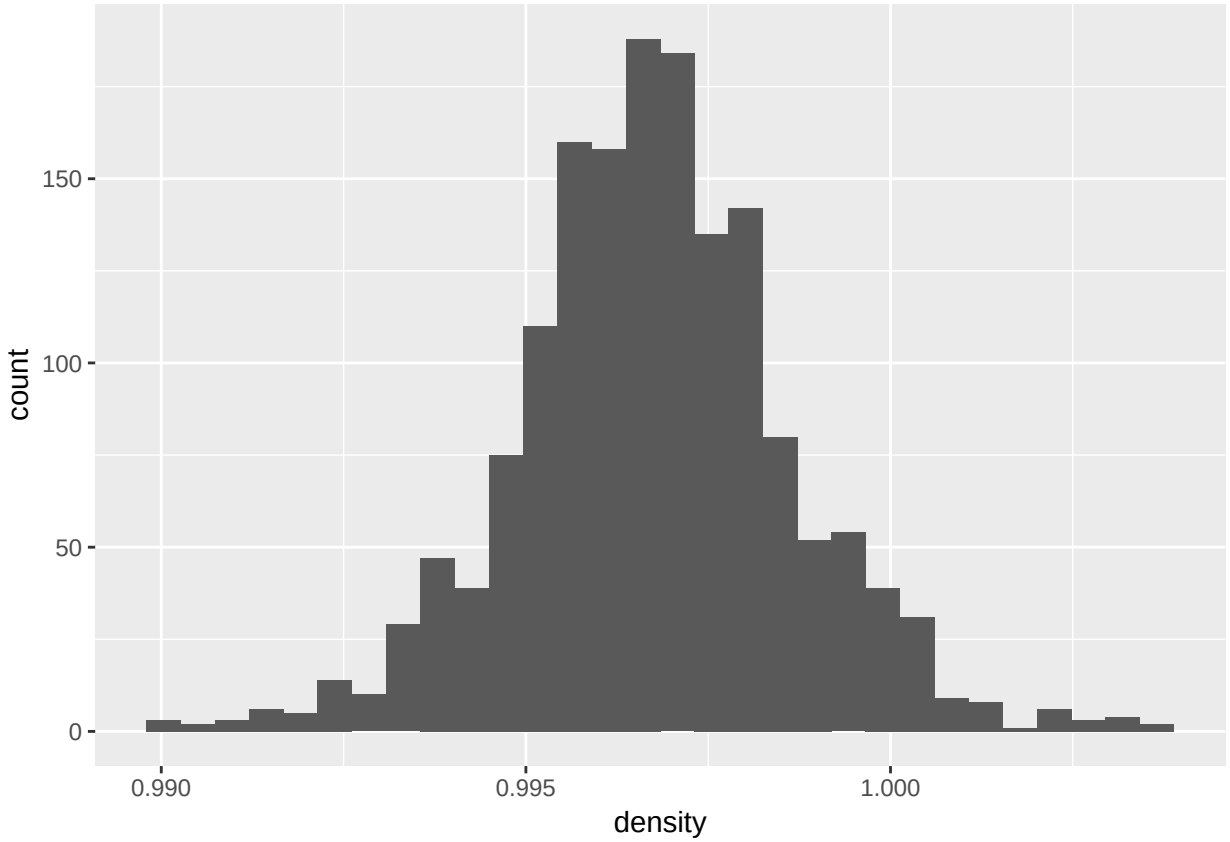
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right and has outliers.so, i applied a log10 scale to the x axis to make the distribution look more normal or bimodal(the one in the bottom).

total.sulfur.dioxide



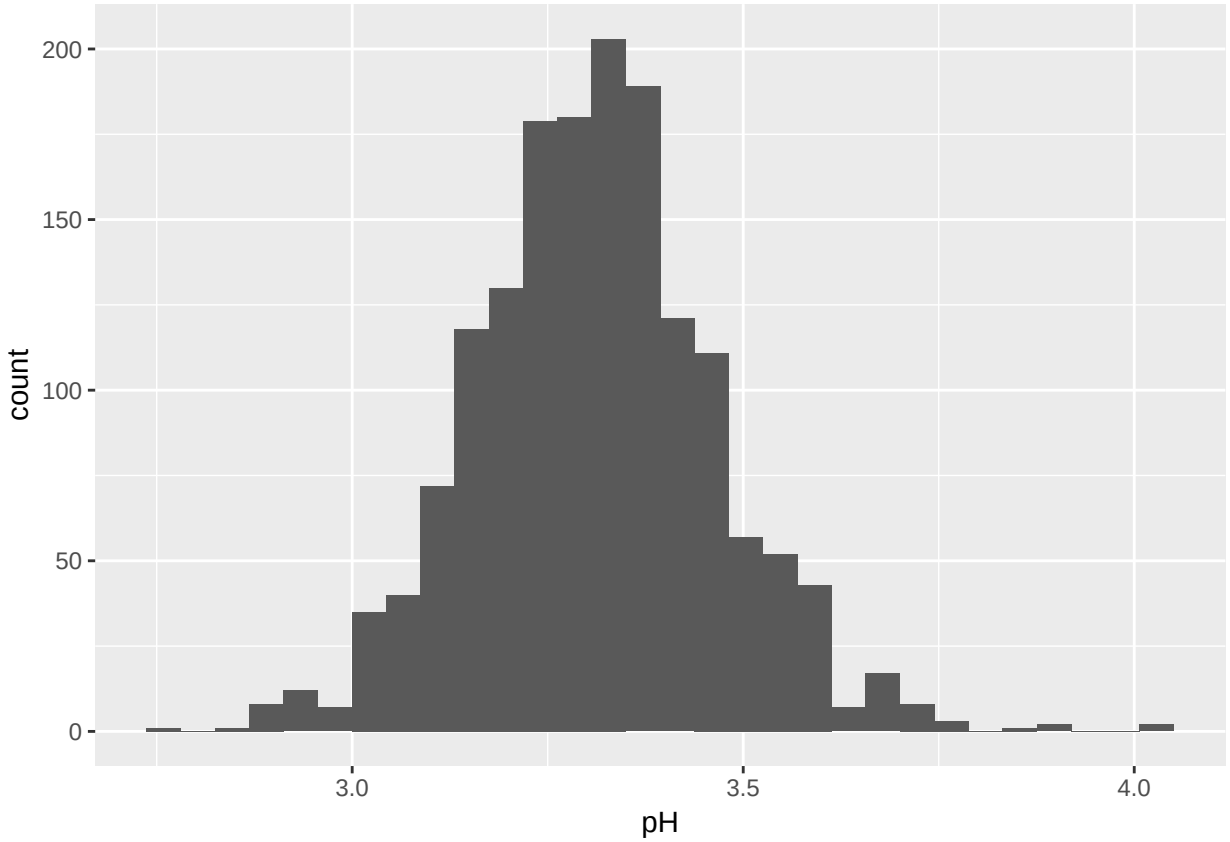
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right and has outliers.so, i applied a log10 scale to the x axis to make the distribution look normal (the one in the bottom).

density



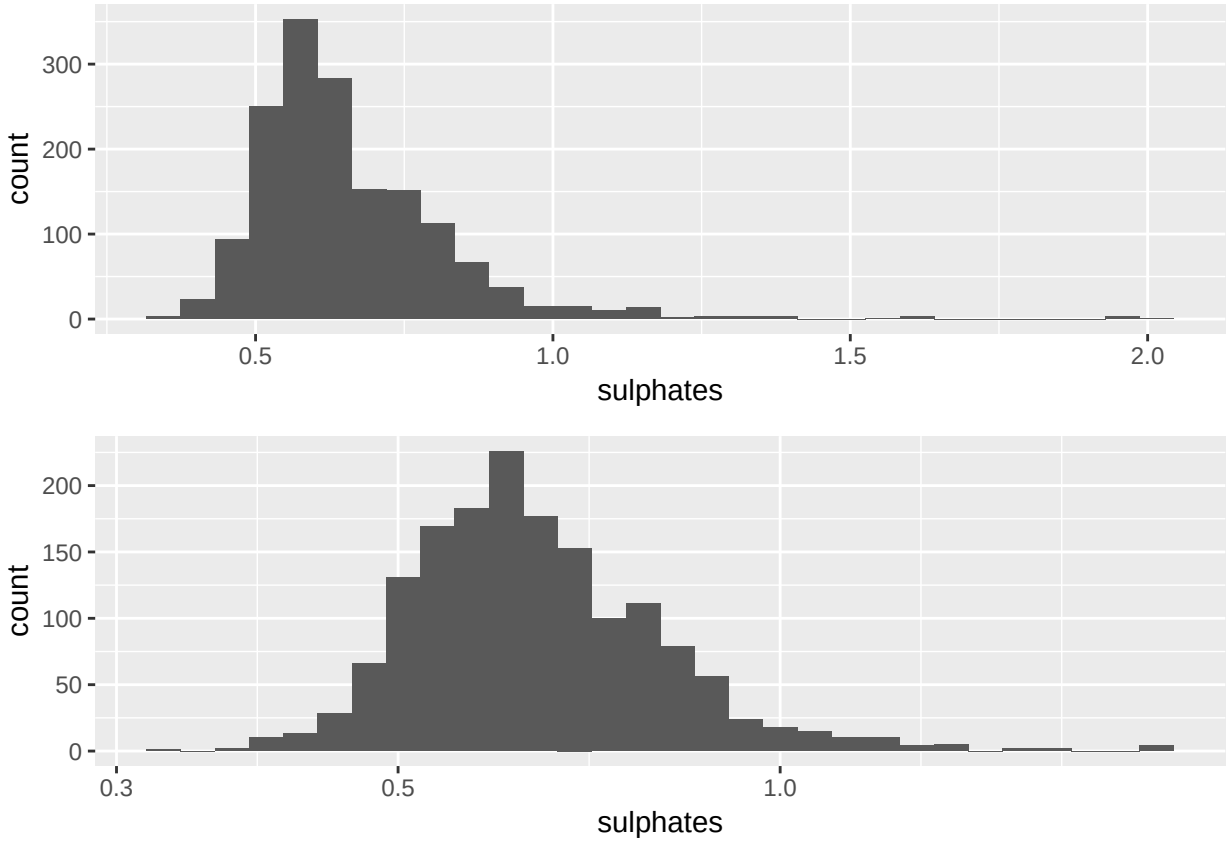
this histogram looks perfectly normaly ditributed.

pH



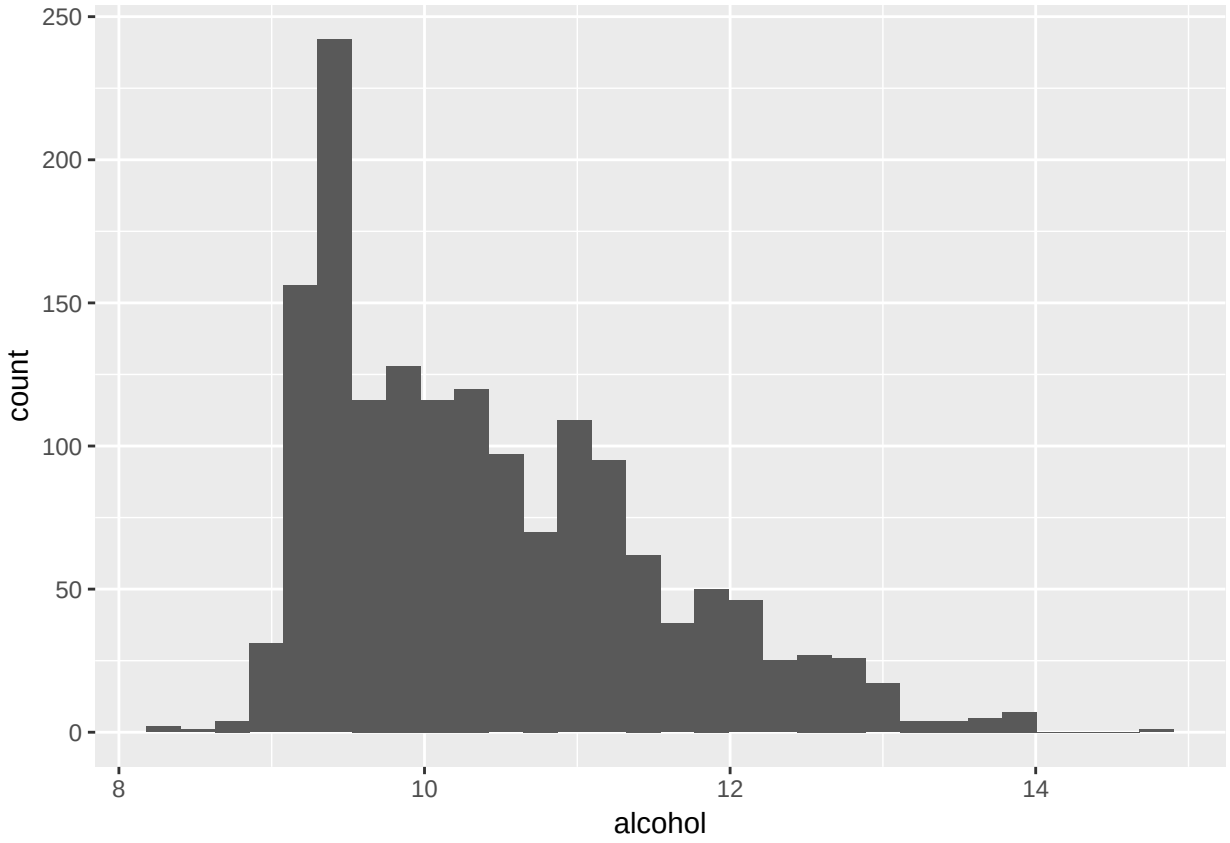
this histogram also looks very normaly ditributed, but it might contain some outliers.

sulphates



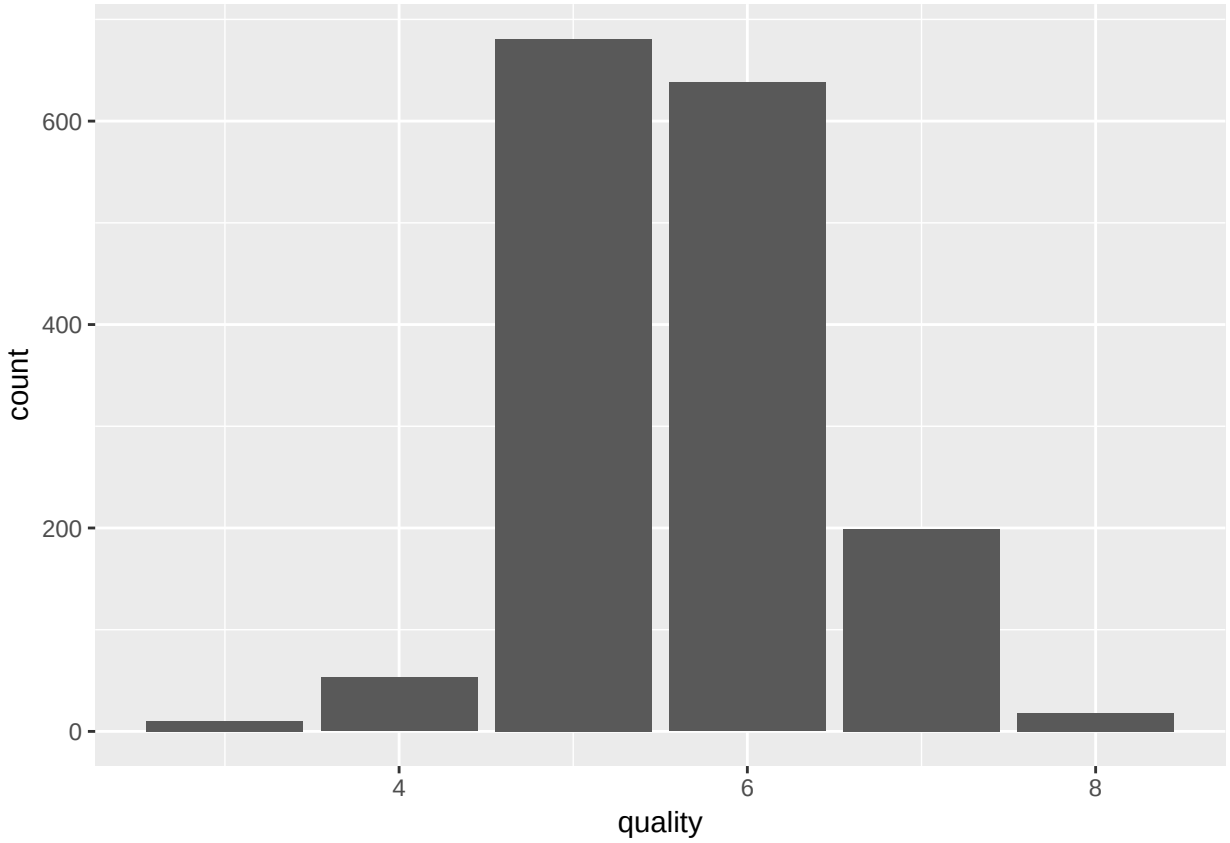
these two histograms are both for the same variable,the histogram that's in the top looks skewed to the right and has outliers.so, i applied a log10 scale to the x axis to make the distribution look normal (the one in the bottom).

alcohol



this histogram is skewed to the right.

quality



this bar plot shows the distribution of the quality values, it looks like the quality of the red wine is between 3 and 8.with most of the red wine having a quality of 5 or 6.

Univariate Analysis

What is the structure of your dataset?

the red wine dataset contains 12 attributes(variables), 11 of them are numerical values for chemical tests made on the wine, and the last attribute is the output of those chemical tests(the quality). and it consists of 1599 observations.

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

clearly, the main feature of interest in the dataset is the quality.

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

all the other 11 features may have an effect over the quality because the quality depends on them. although, they might have different levels of correlation to the quality.

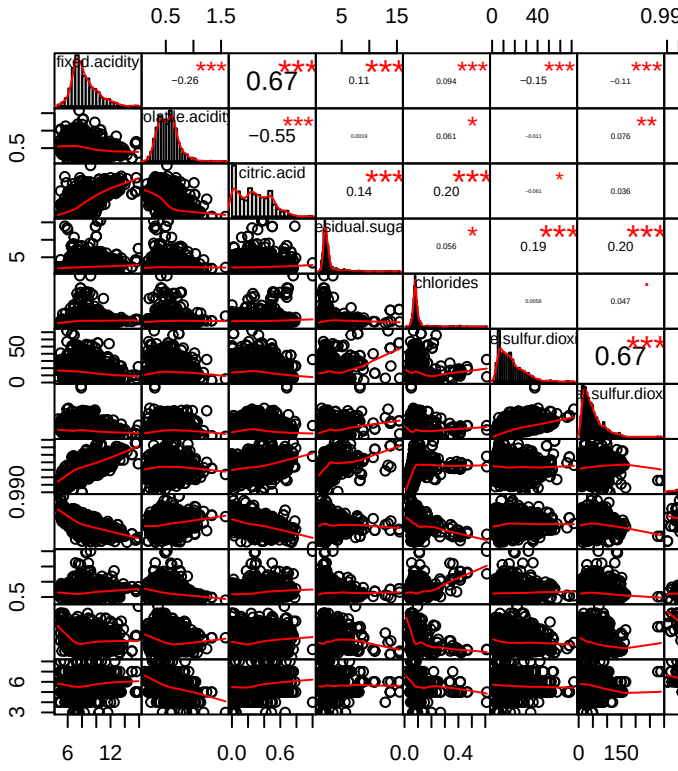
Did you create any new variables from existing variables in the dataset?

No, i did not.

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?

There were no unusual distributions. i did not perform any operations to tidy the data, it's already tidy. I've just dropped a column because it represented the index of the rows which is "redundant".

Bivariate Plots Section

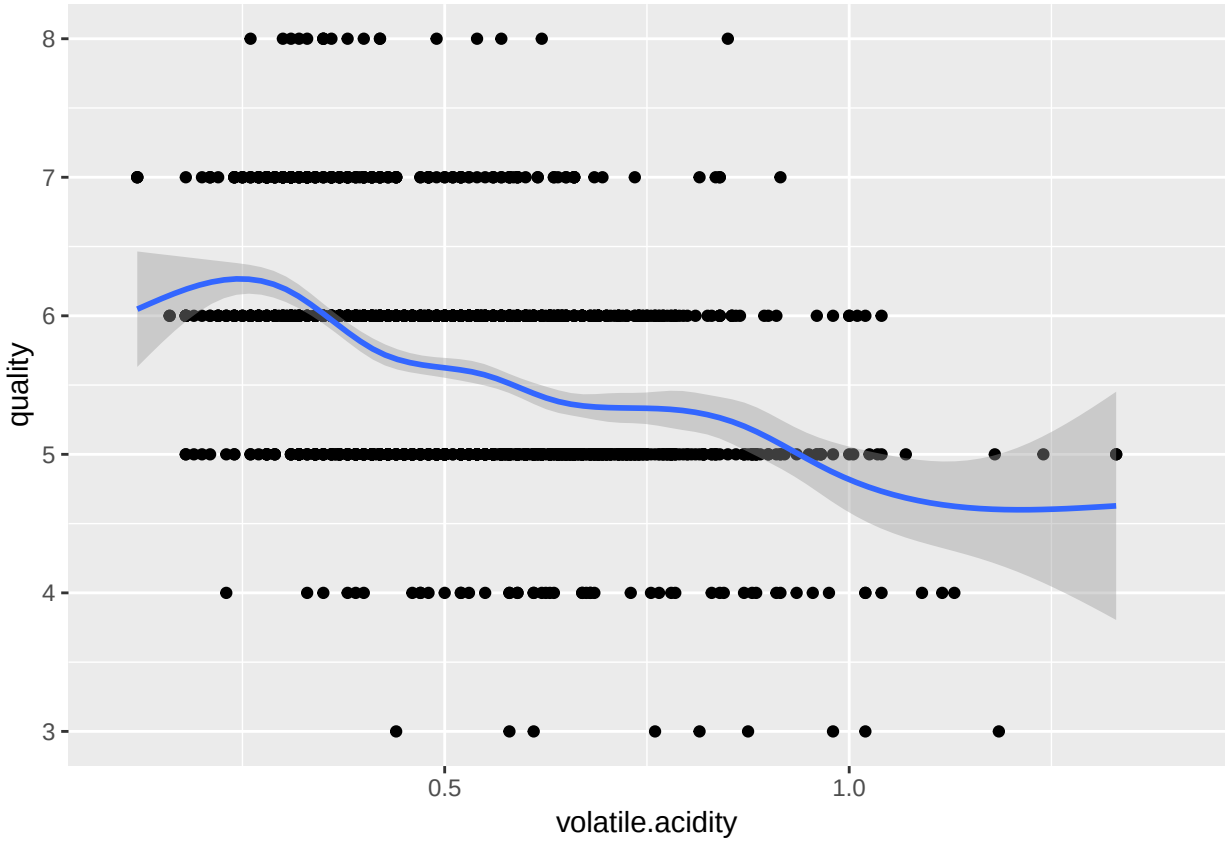


we'll first start by drawing a correlation chart of all the variables

on this plot we can see the distributions of all the variables on the diagonal, and we can also see the scatter plots of different variable below the diagram, and on top of the diagram we can see the significance and correletion between the variables.

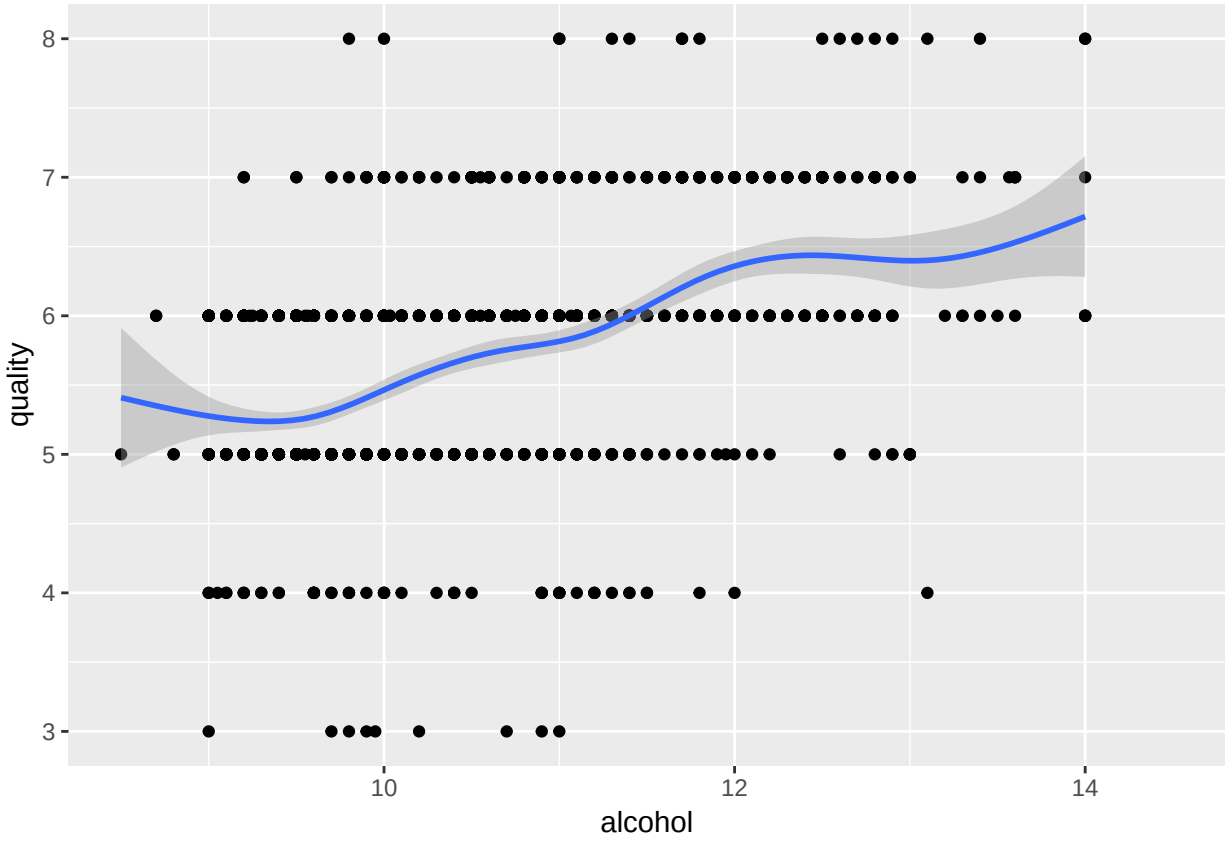
based on the scatter plots and the correlation test scores; i'll perform some biavariate plots. i will not draw many plots, because this chart has it all. i'll draw two plots of the two most correlated variables to the output variable depending on this chart, and i'll draw a couple more plots of other variables that

volatile.acidity vs quality



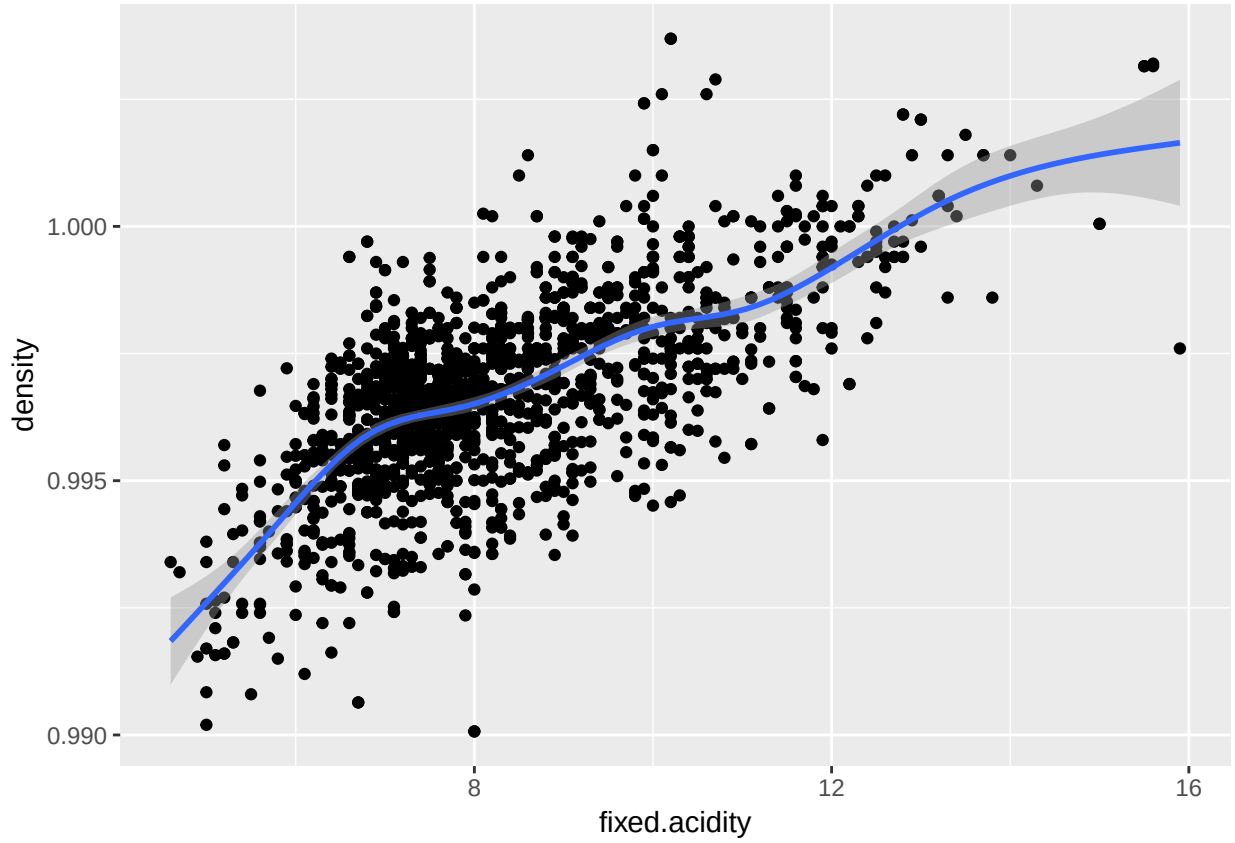
we can see that there is a small correlation between the volatile.acidity and the quality, as the volatile.acidity increases the quality decreases.with only on outlier that i have eliminated from the plot.

alcohol vs quality



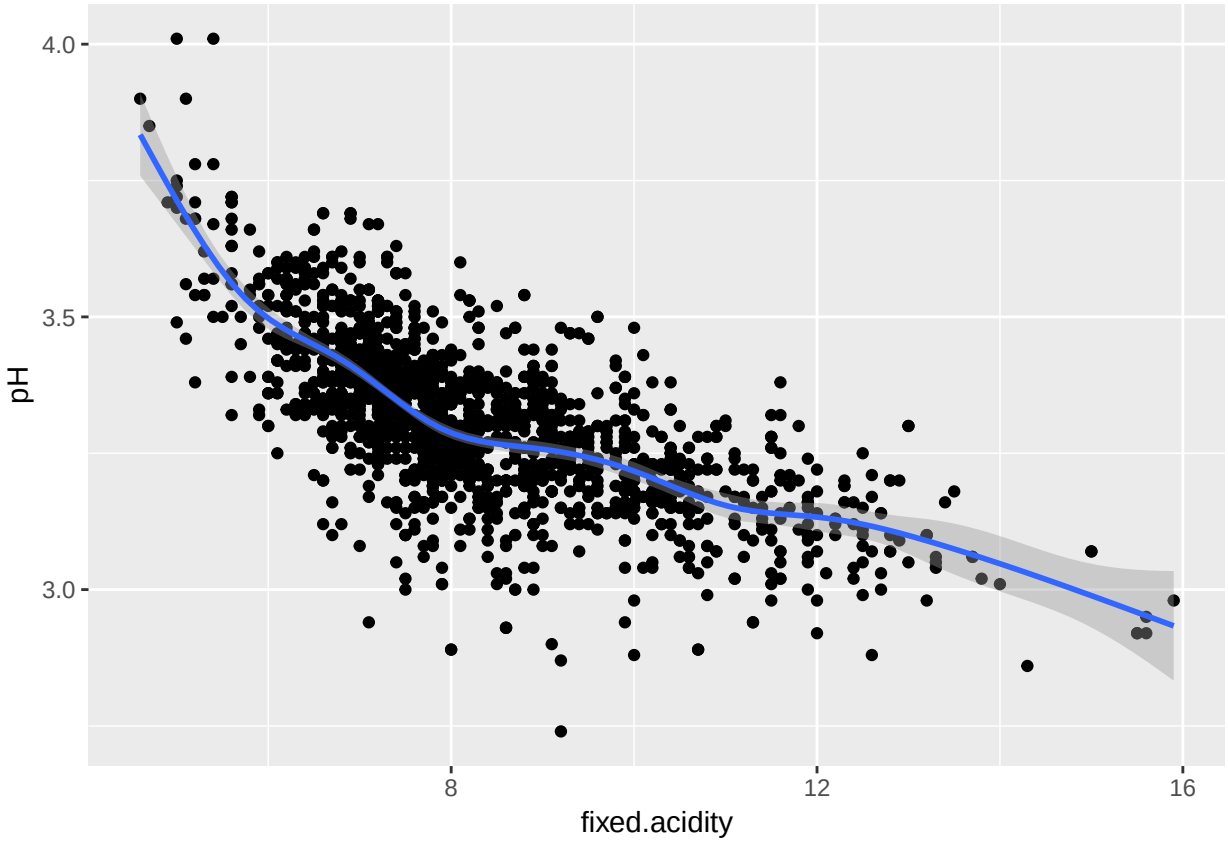
we can see that there is a small correlation between the alcohol and the quality, as the alcohol increases the quality also increases, with only on outlier that i have eliminated from the plot.

fixed.acidity vs density



we can also notice that there is a strong correlation between the fixed.acidity and the density of the wine.

fixed.acidity vs pH



we can also notice that there is a strong negative correlation between the fixed.acidity and the pH of the wine.

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?

according to the correlation chart, the strongest relationships for the feaute of interest(quality) are with the alcohol and the citric.acid

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

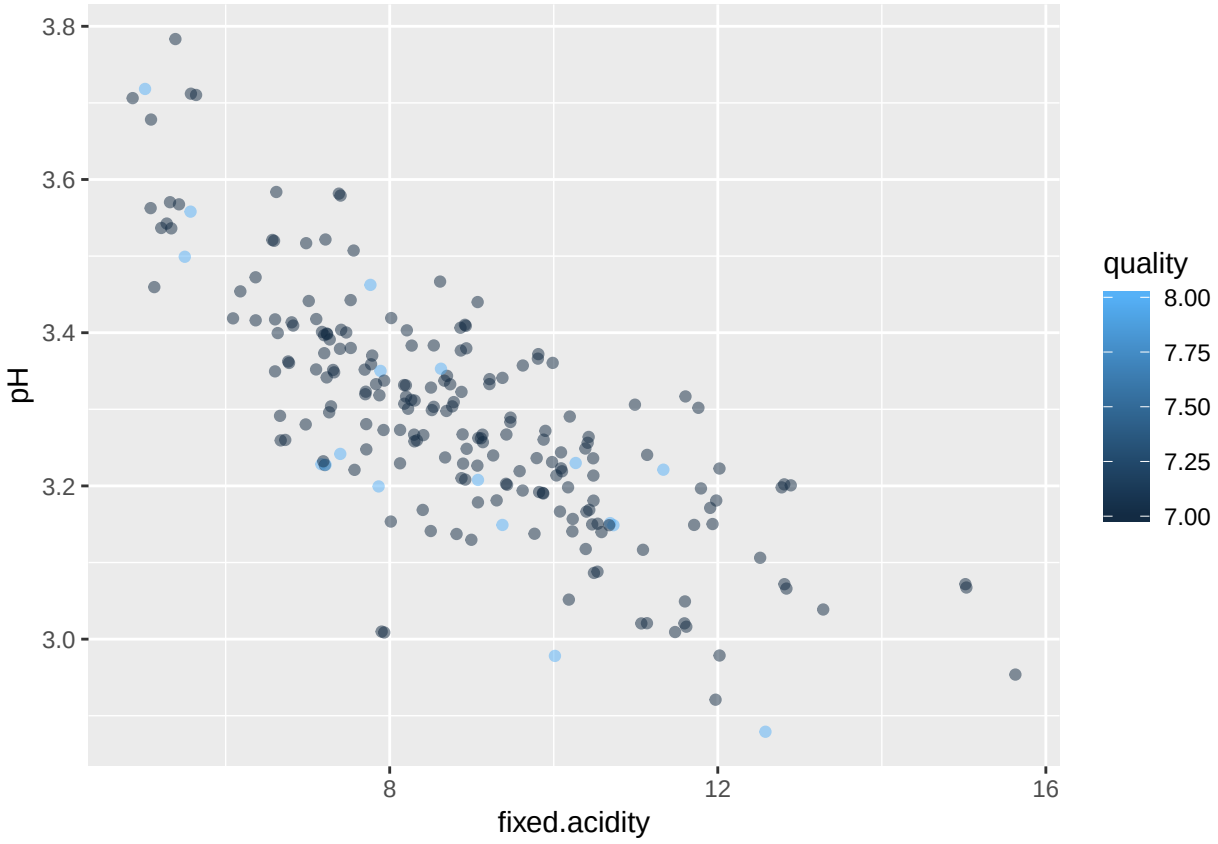
yeah, there were some strong correlation test scores between other features as (fixed.acidity vs pH) and (fixed.acidity vs density)

What was the strongest relationship you found?

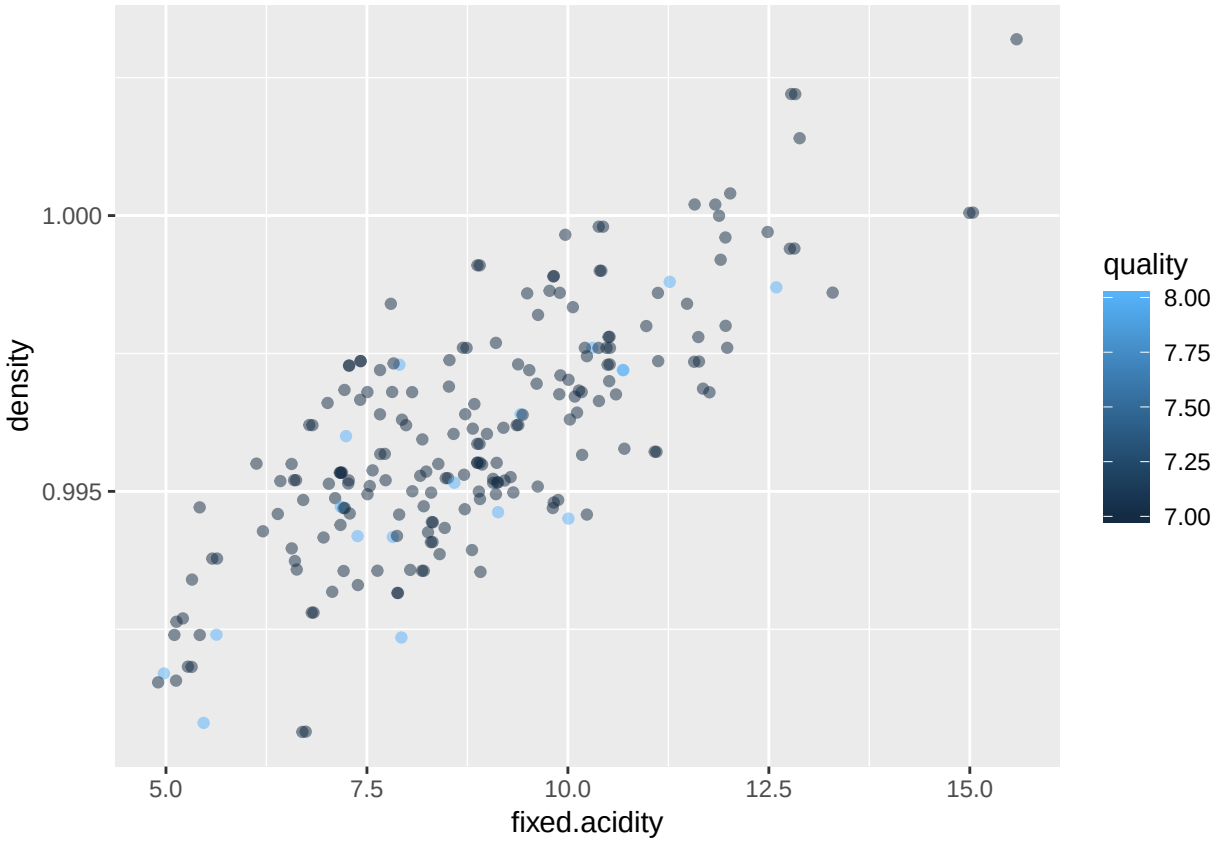
the strongest relationship i found was (fixed.acidity vs pH)

Multivariate Plots Section

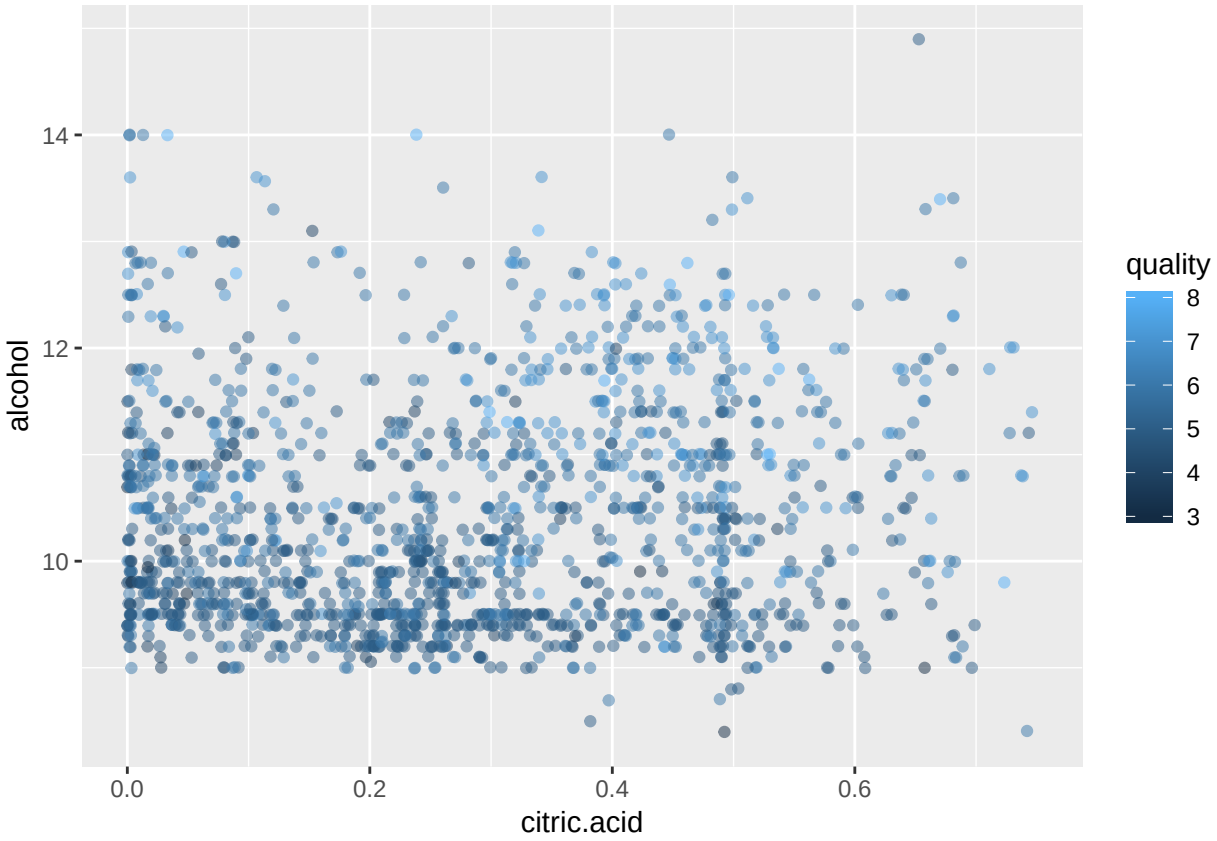
Tip: Now it's time to put everything together. Based on what you found in the bivariate plots section, create a few multivariate plots to investigate more complex interactions between variables. Make sure that the plots that you create here are justified by the plots you explored in the previous section. If you plan on creating any mathematical models, this is the section where you will do that.



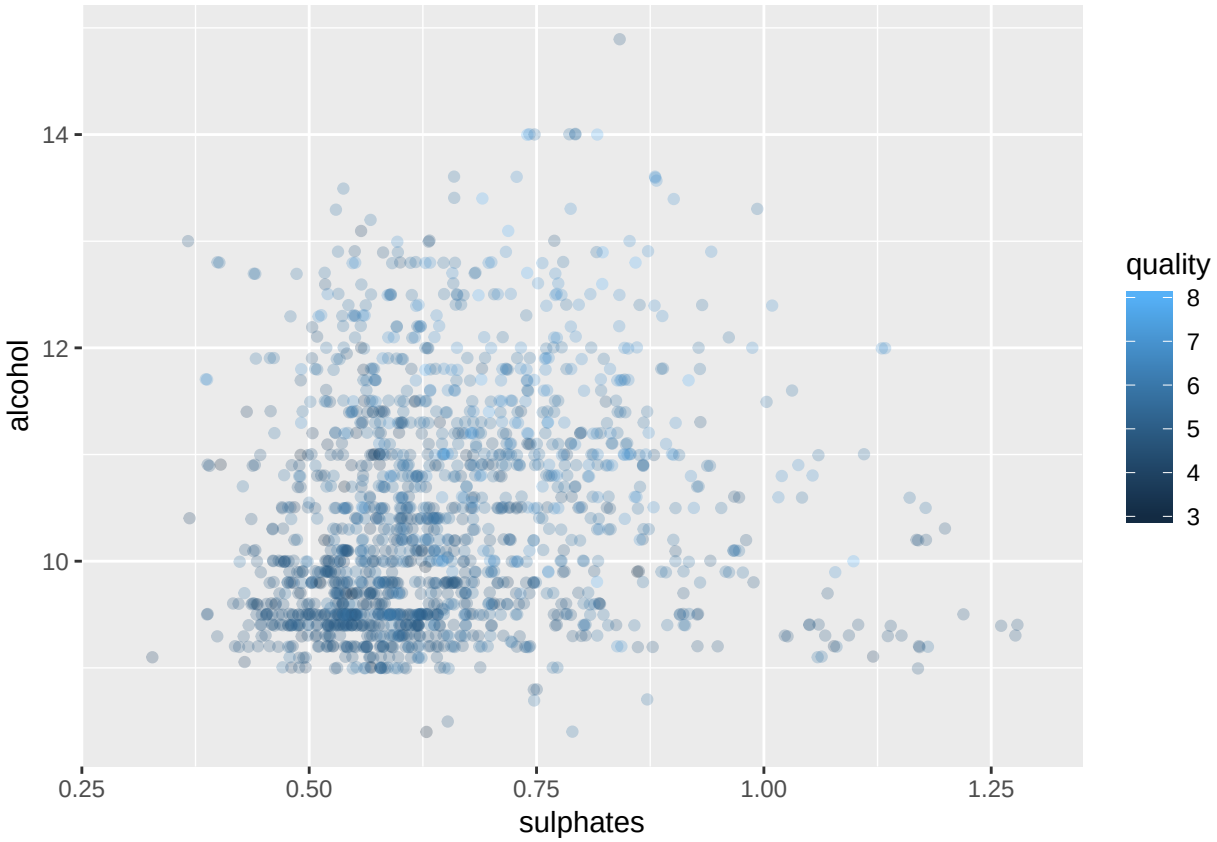
i can't see a clear pattern for the high quality wine's relationship to pH and fixed.acidity.



i can't see a clear pattern for the high quality wine's relationship to pH and fixed.acidity.



the increase of both alcohol and citric acid results in an increase in th quality of red wine.



a high quality of alcohol has a value of sulphates between (0.65 and 1.25) and a value of alcohol between (5 and 13) "on average".

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?

yeah, i think that the citric acid and the alcohol stregthen each other, and alcohol and sulphates also stregthen each other.

Were there any interesting or surprising interactions between features?

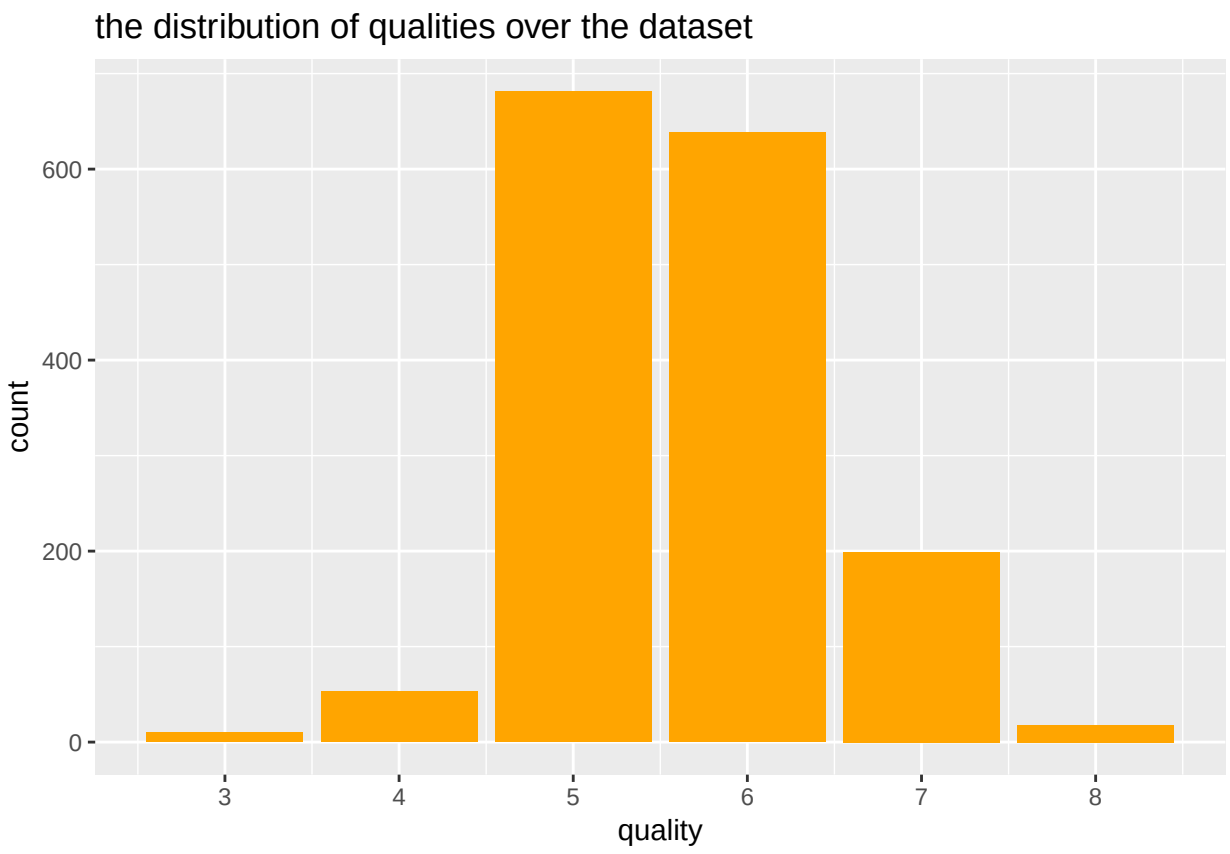
i think that alcohol has a big impact on almost all of the other variables .

Final Plots and Summary

> **Tip**: You've done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings.

Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

Plot One

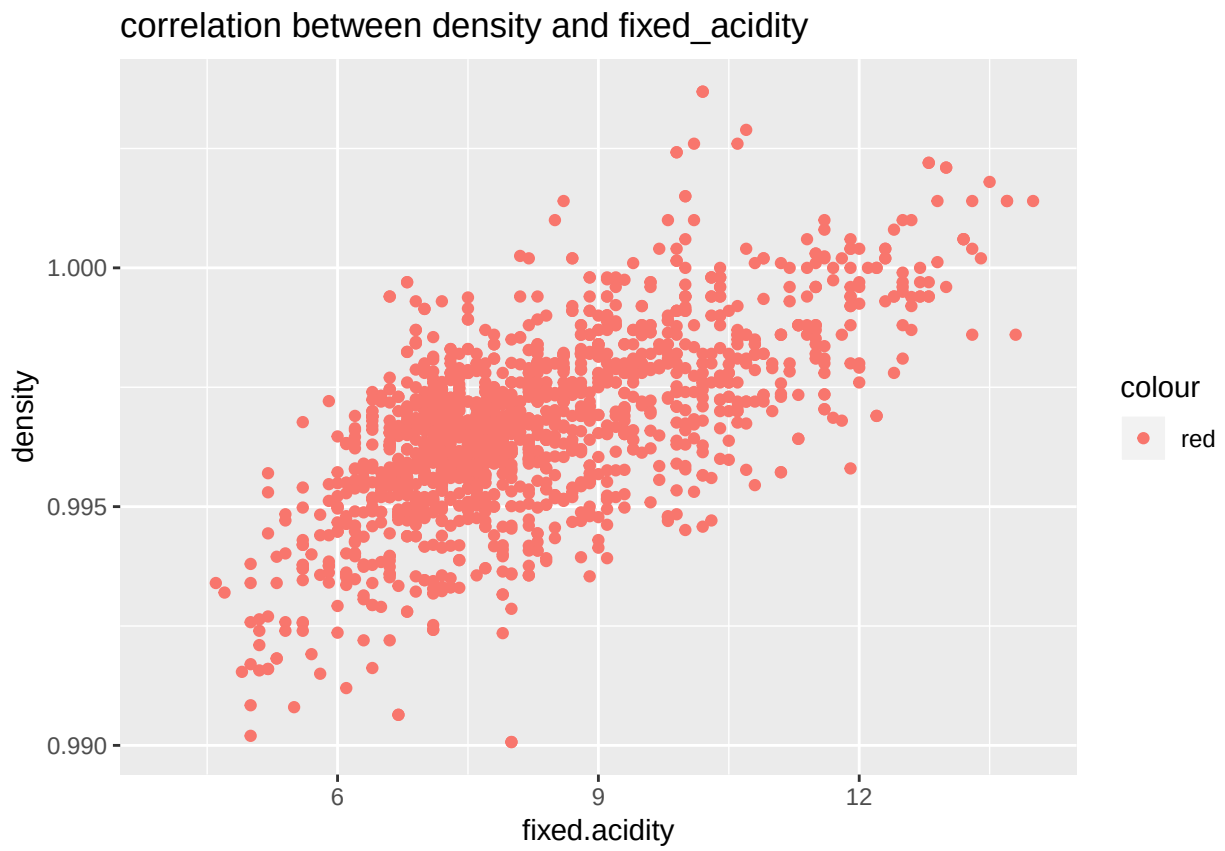


Description One

this bar plot shows that the quality of the red wine ranges between 3 and 8, but most the red wine in the dataset have a quality of 5 or 6.

Plot Two

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

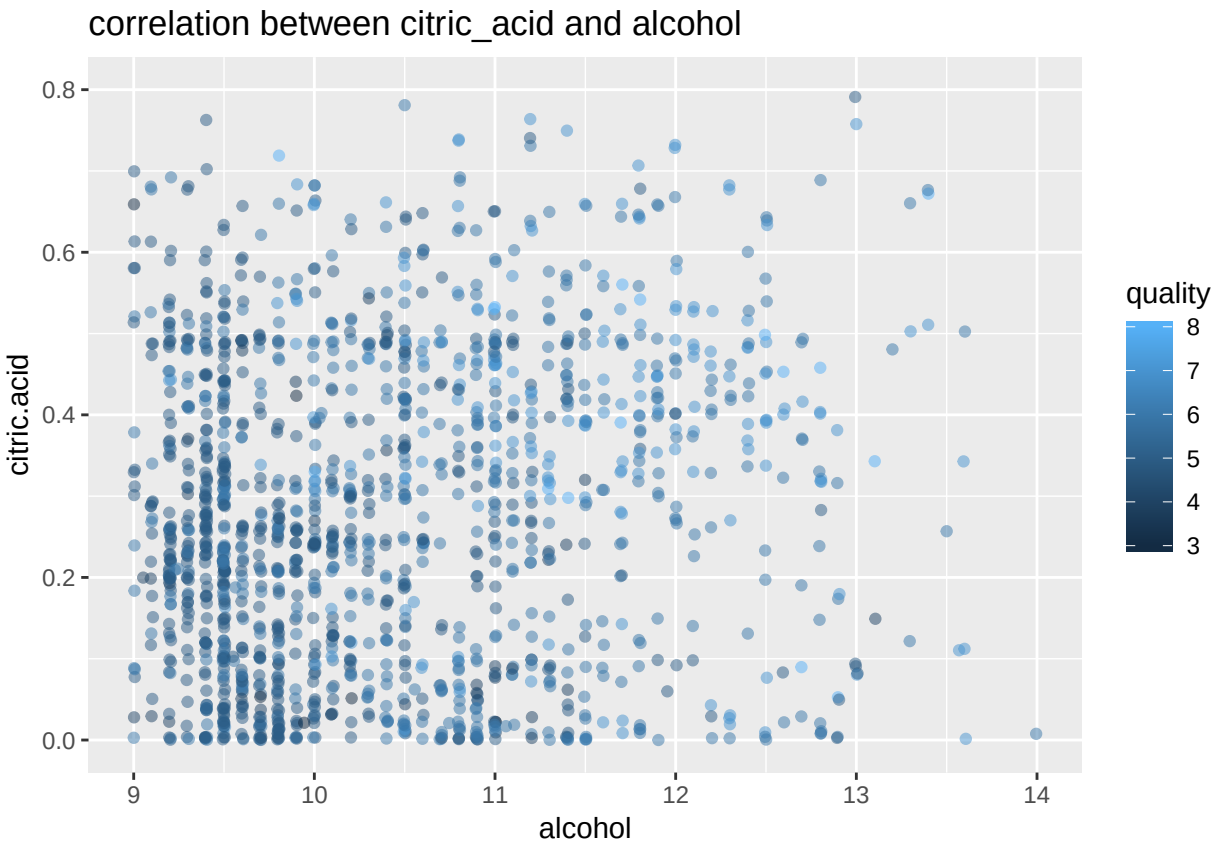


Description Two

this plot is interesting because we can see how the density and the fixed.acidity strengthen each other. The plot shows a strong positive correlation.

Plot Three

```
## Warning: Removed 86 rows containing missing values (geom_point).
```



Description Three

in this plot we can see that the increase of both alcohol and citric acid results in an increase in the quality of red wine.

Reflection

this was a very interesting dataset with many variables to explore.

one of the struggles that i went through is deciding if any variable really had a significant effect on the quality of the wine, because the dataset didn't have many high quality wine samples(unbalanced).

for me, drawing the bar chart had a very usefel impact on defining a path of the exploration of this dataset.

i think that a future work that could be done with the dataset is to give all of the wine samples the same alcohol value and see if the distribution of the quality in the dataset will have a significant change.