White Portuguese Wine Analysis by Quentin Chenevier

This analysis aims at analysis a dataset about portuguese white wines.

The dataset is related to white variant of the Portuguese "Vinho Verde" wine (http://www.vinhoverde.pt/en/).

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ 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             : Factor w/ 7 levels "3","4","5","6",..: 4 4 4 4 4 4 4 4 4 4 ...
```

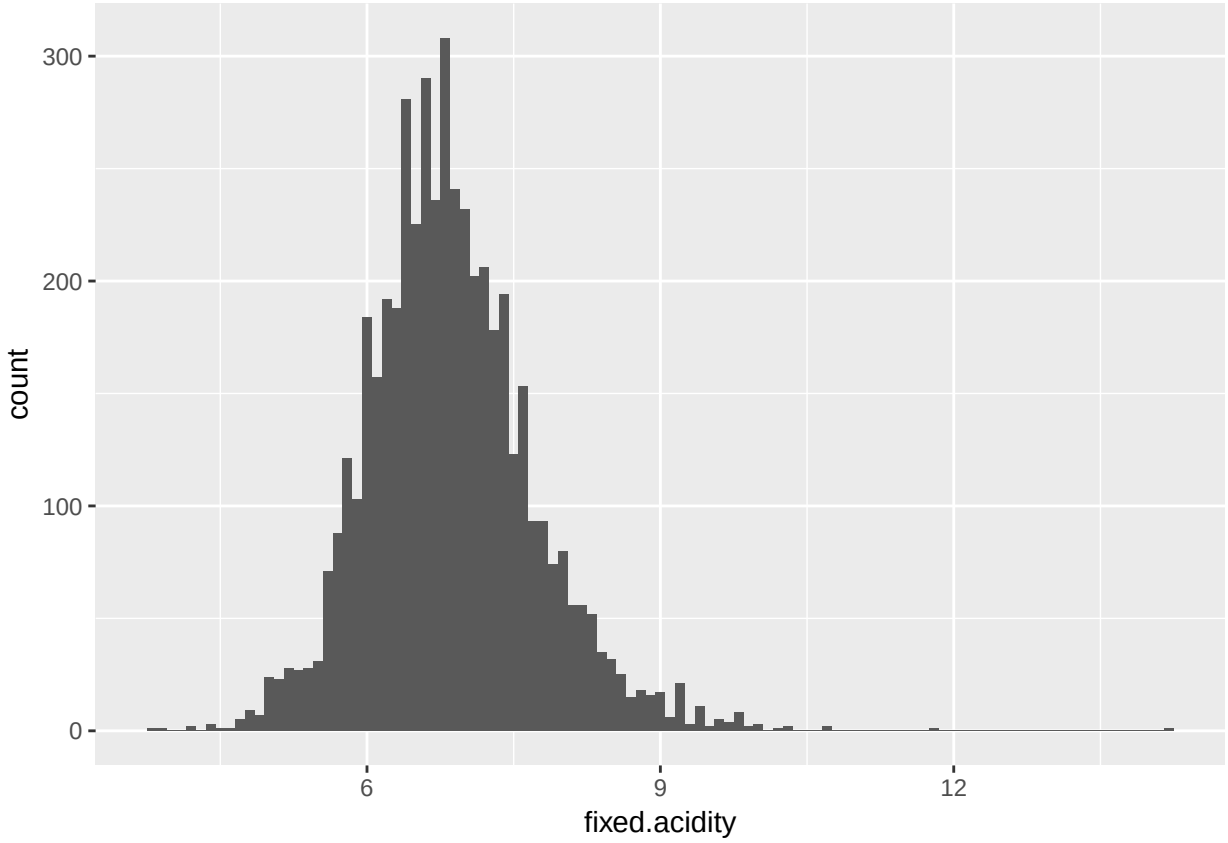
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##
##  quality
##  3:  20
##  4: 163
##  5:1457
##  6:2198
##  7: 880
##  8: 175
##  9:   5
```

The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). The goal of the analysis is to find relationships between the measured inputs variables (the

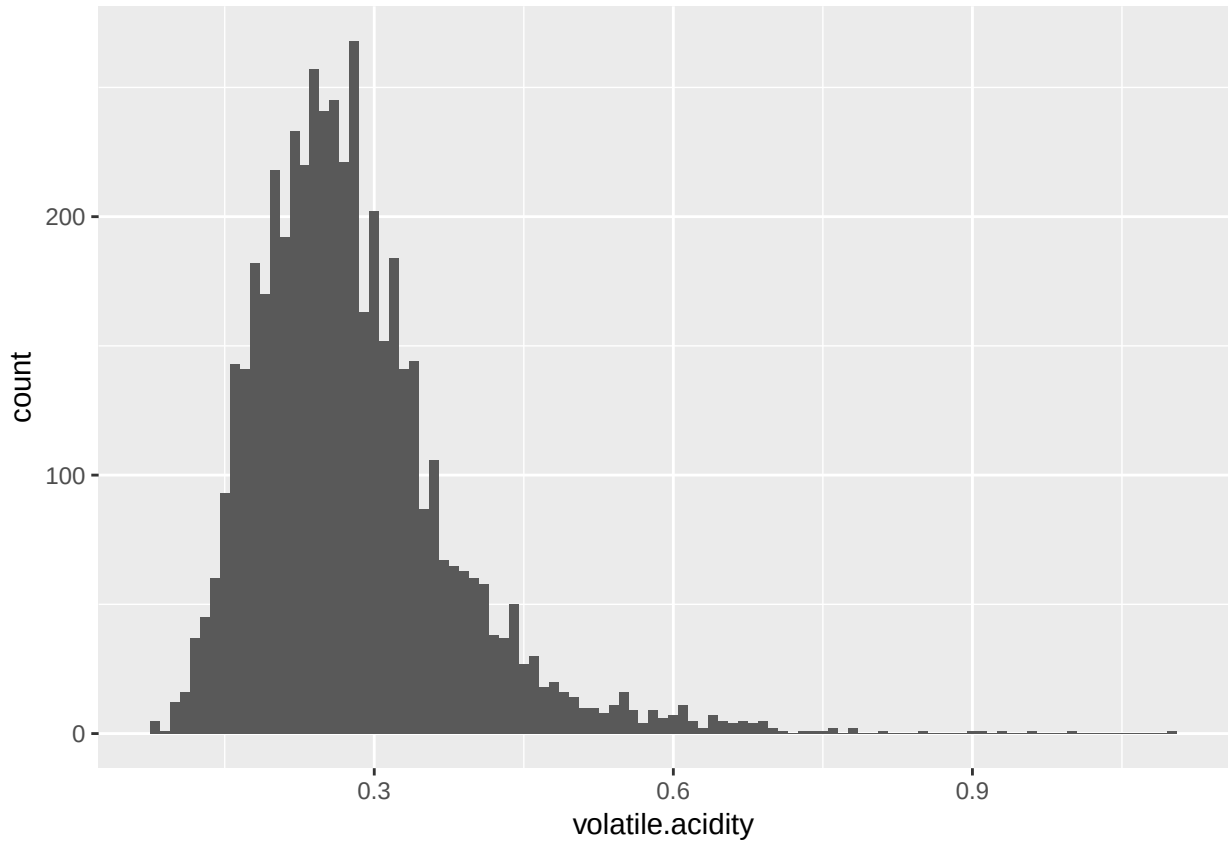
first 11 columns) and the output “quality” variable.

Univariate Plots Section

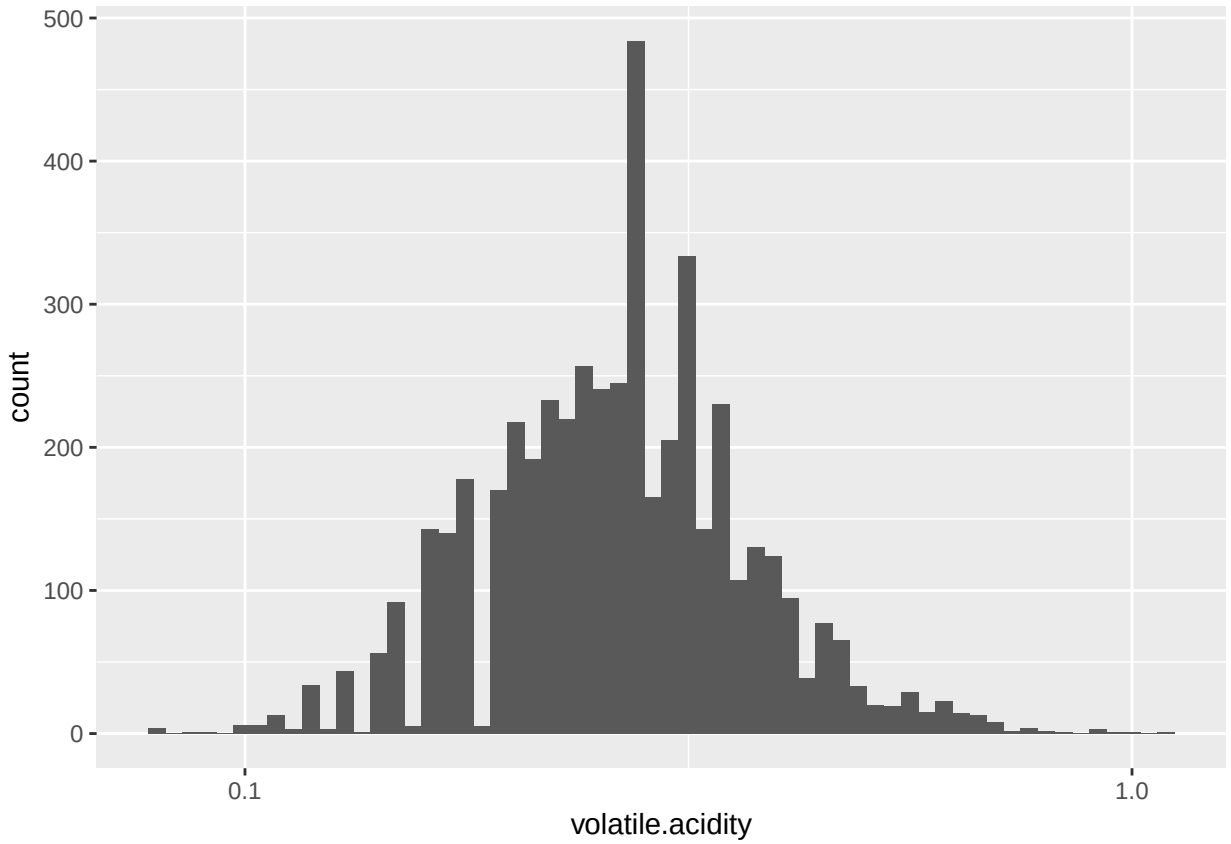
In this section, preliminary exploration on each variable.



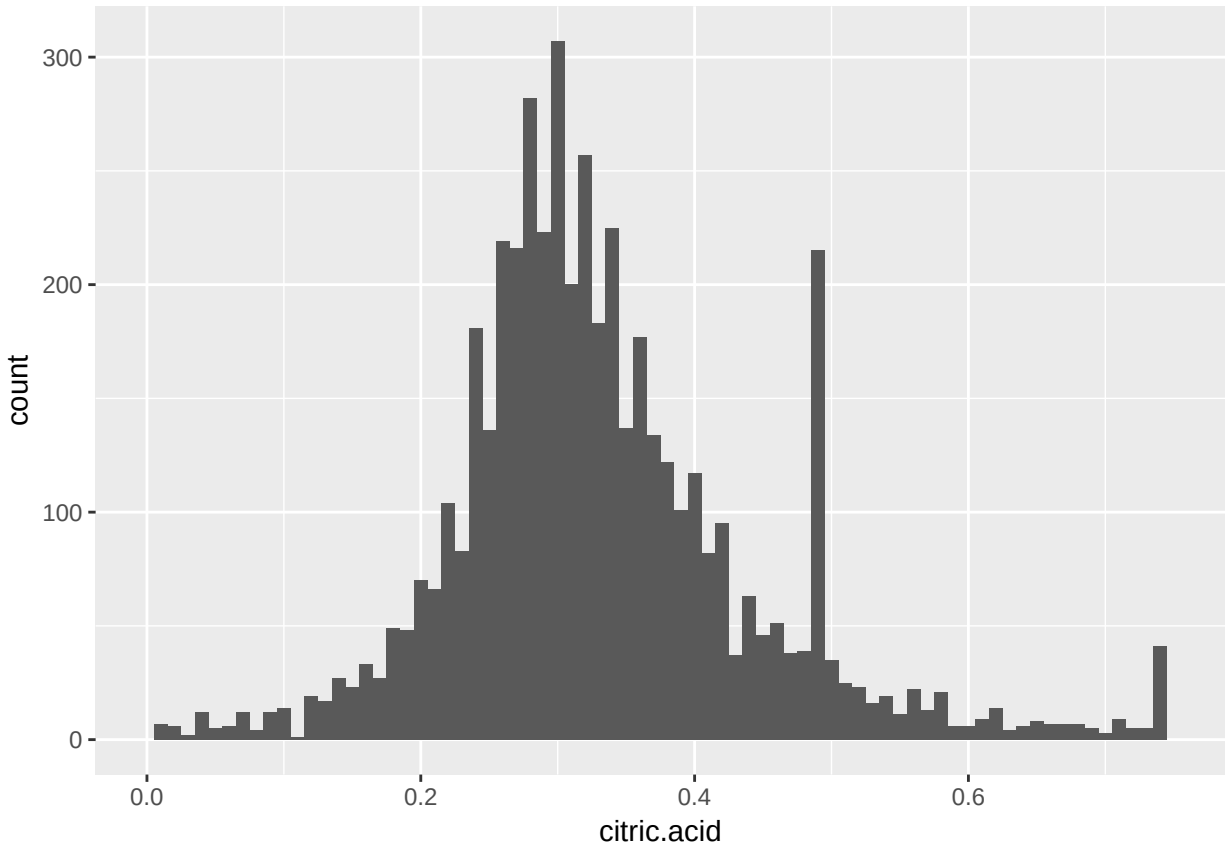
Fixed Acidity seems to be a gaussian distribution.



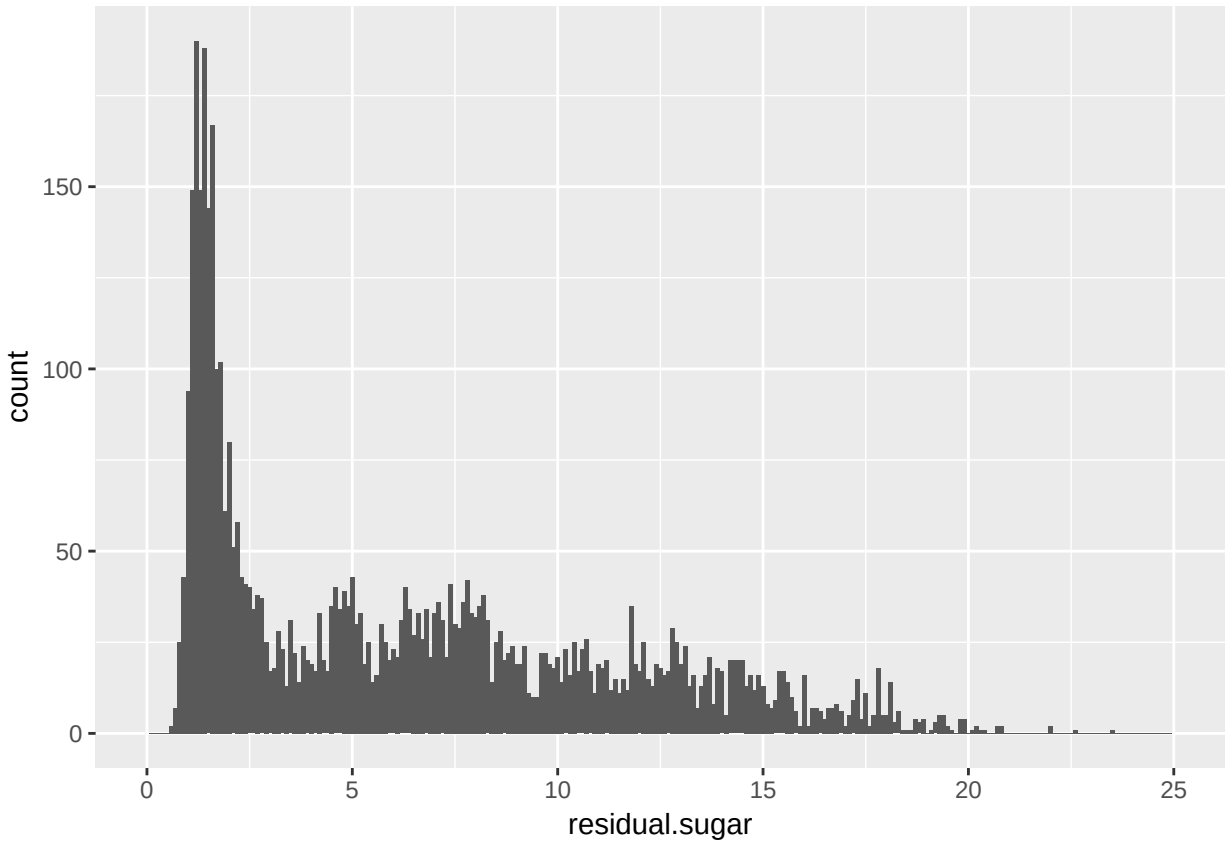
Volatile Acidity seems to be a right-skewed_distribution



Given the shape of the distribution when we transform it with a log, maybe volatile acidity has a log relationship with other normally distributed variables.

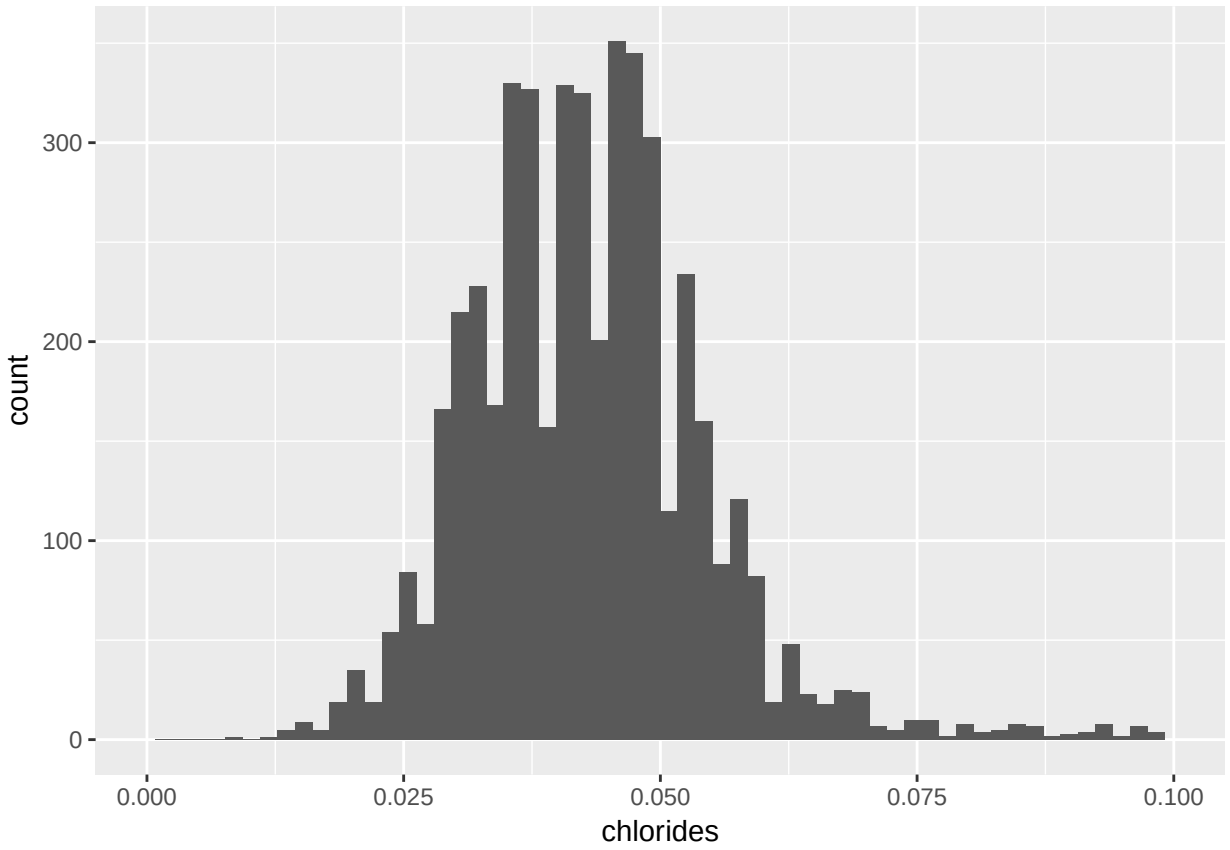


Citric Acid seems to be a slightly right-skewed_distribution. There are some outliers (above 0.75) which need to be filtered to see a proper plot.

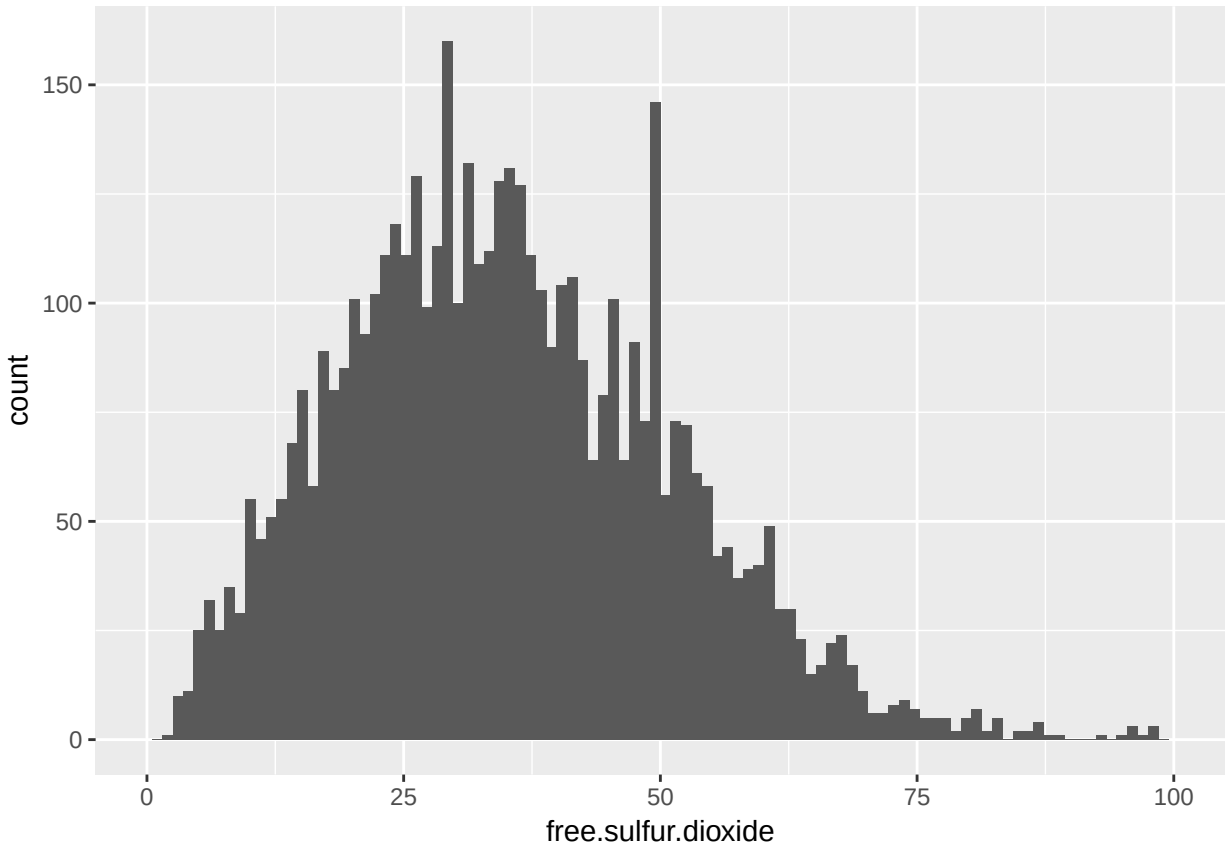


Residual sugar is clearly not a gaussian distribution. There is a big gaussian mode with a low value and a continuum of higher values spread along the sugar spectrum.

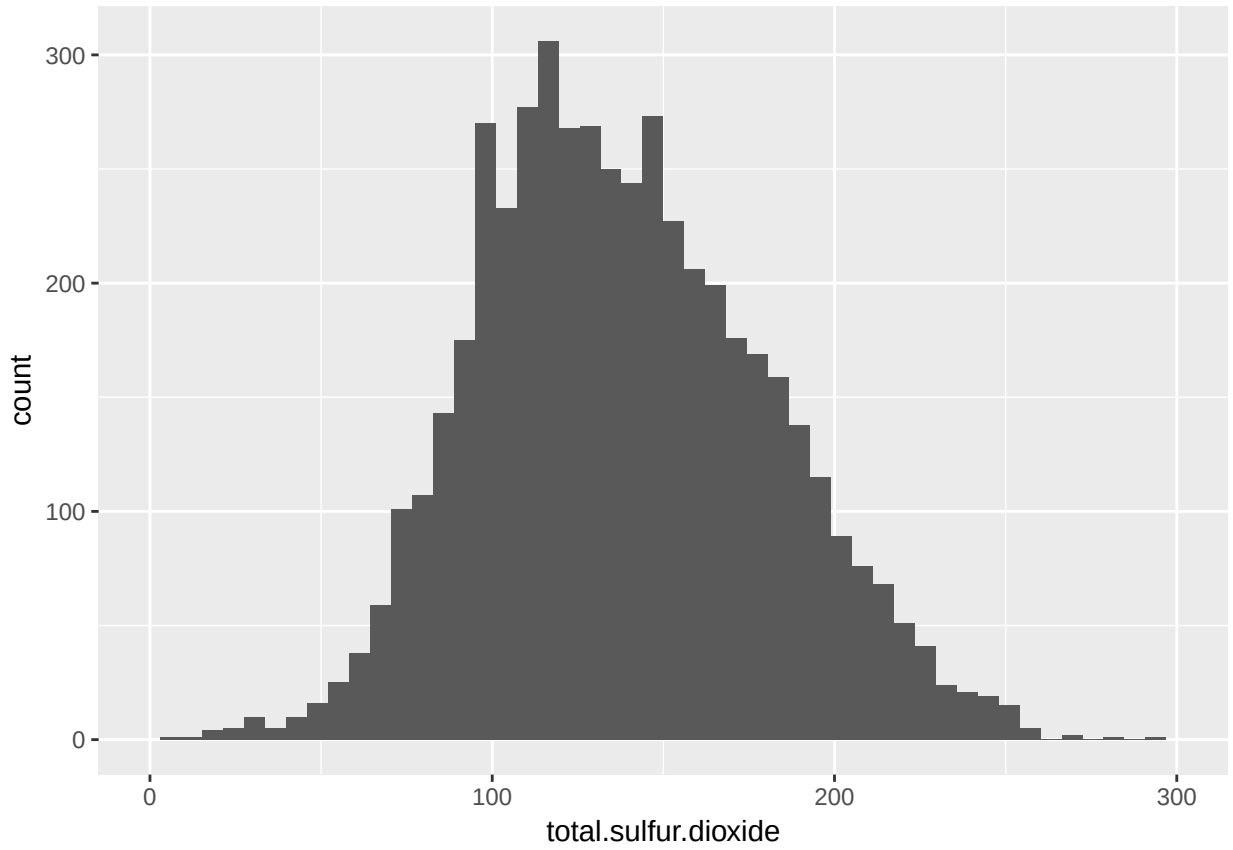
Further analysis is needed to see if this distribution can be splitted in several groups, based on a categorical value.



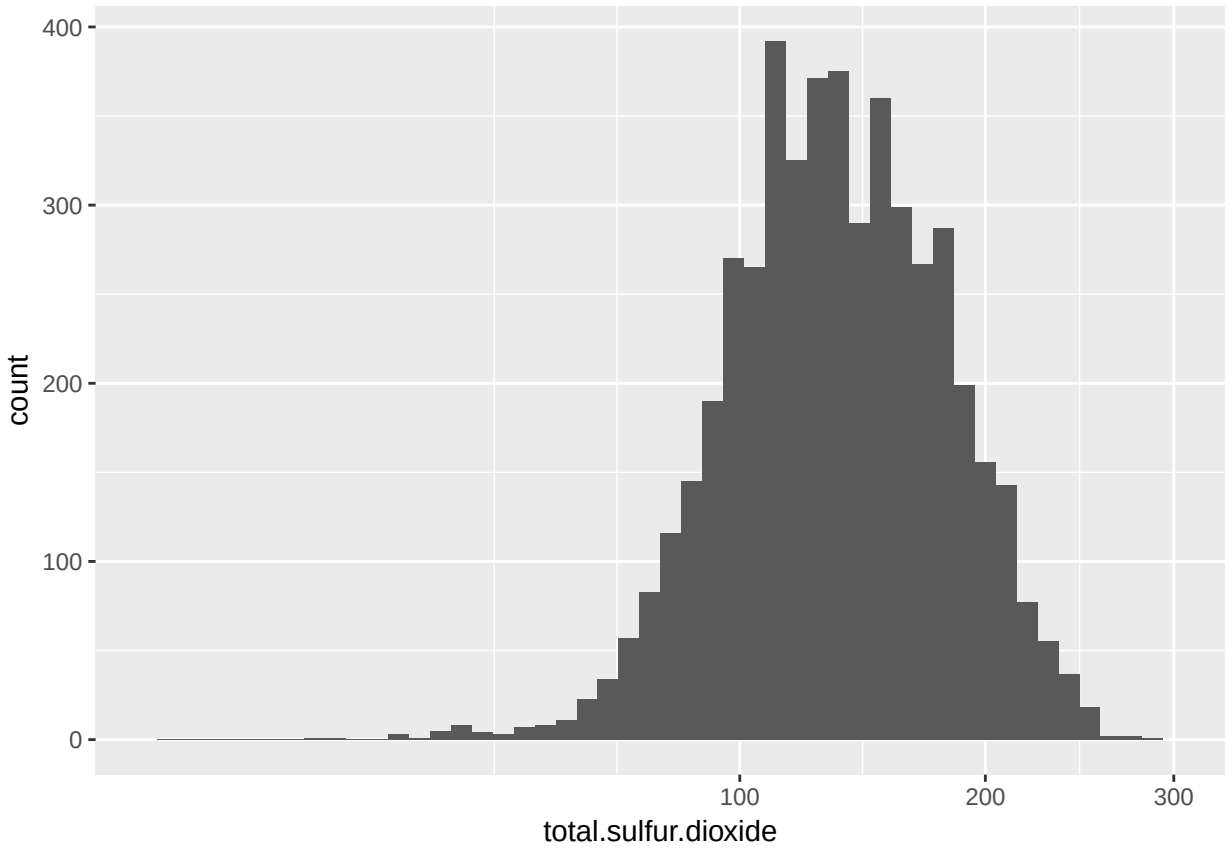
The chlorides variable seems to be a normal distribution.



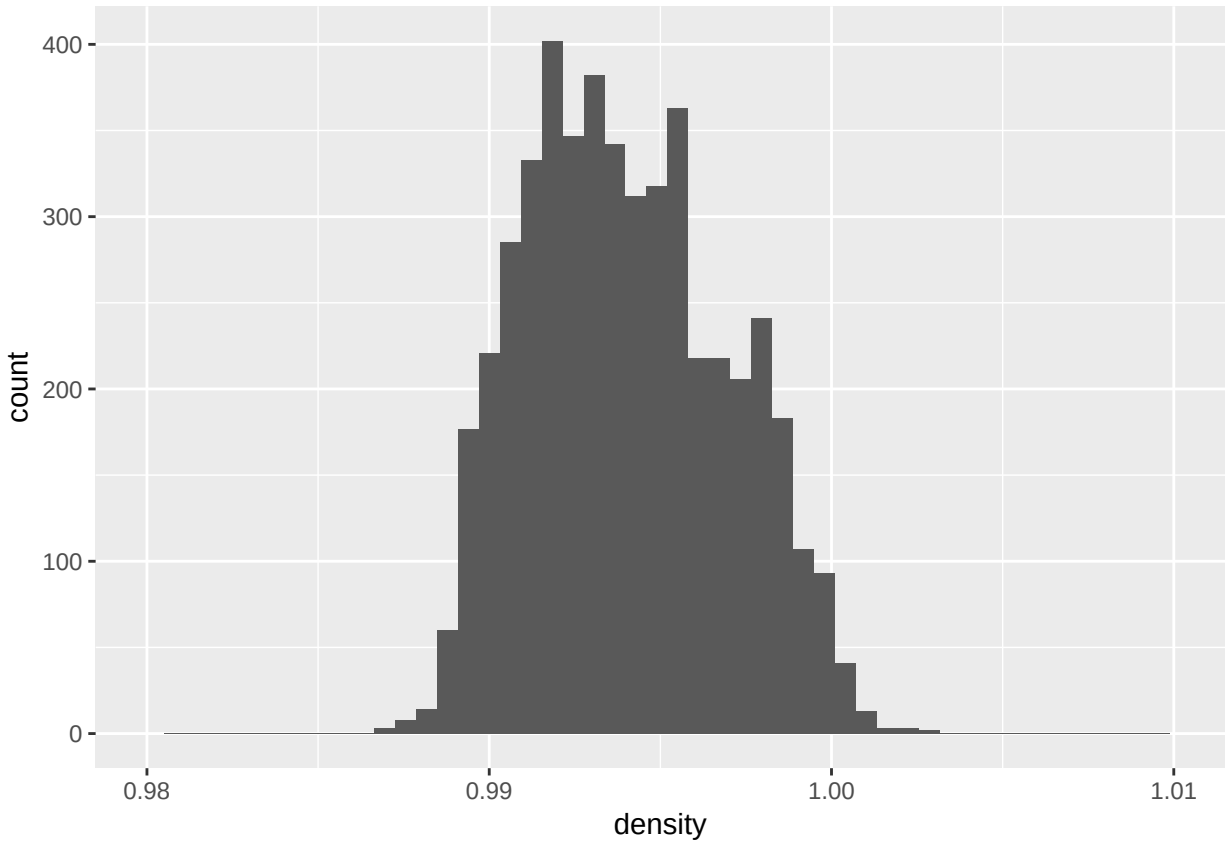
The free sulfur dioxide distribution seems to be a bit right skewed



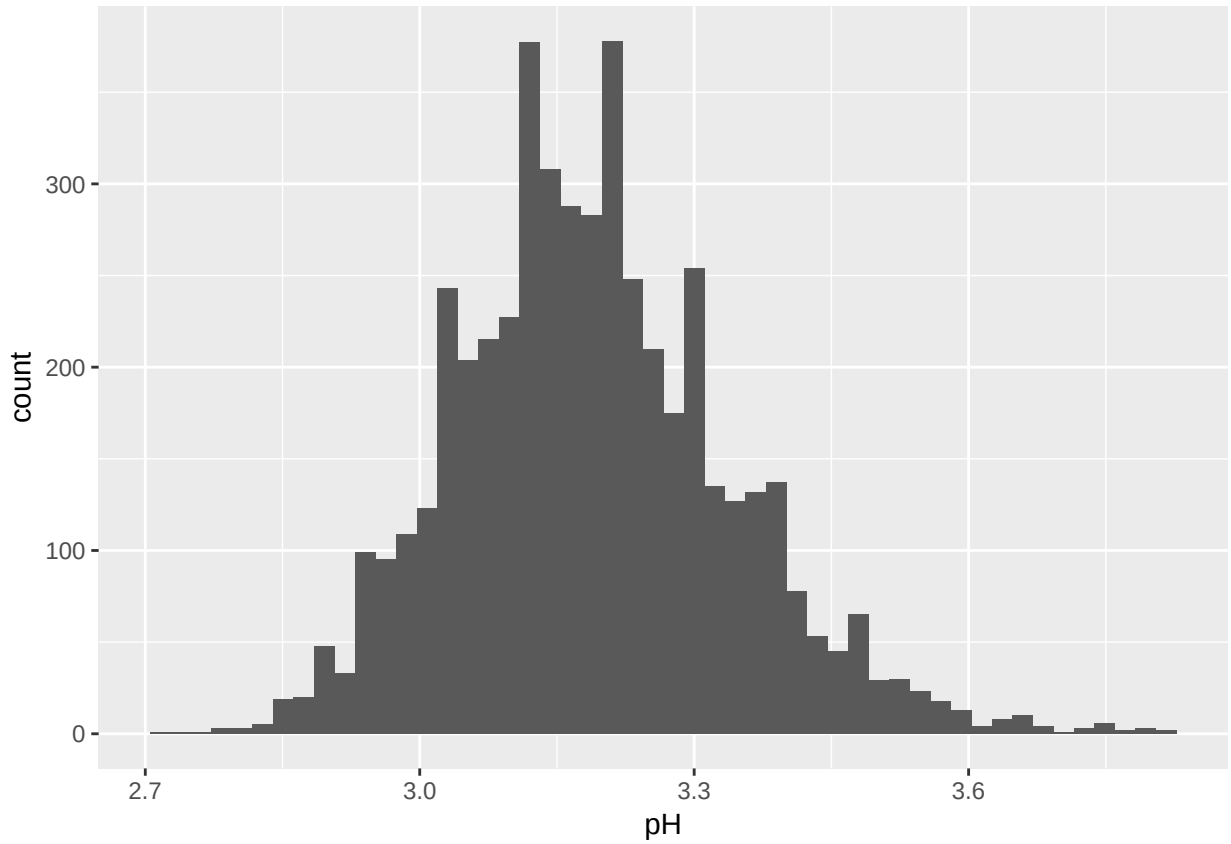
The total sulfur dioxide distribution seems to be a bit right skewed



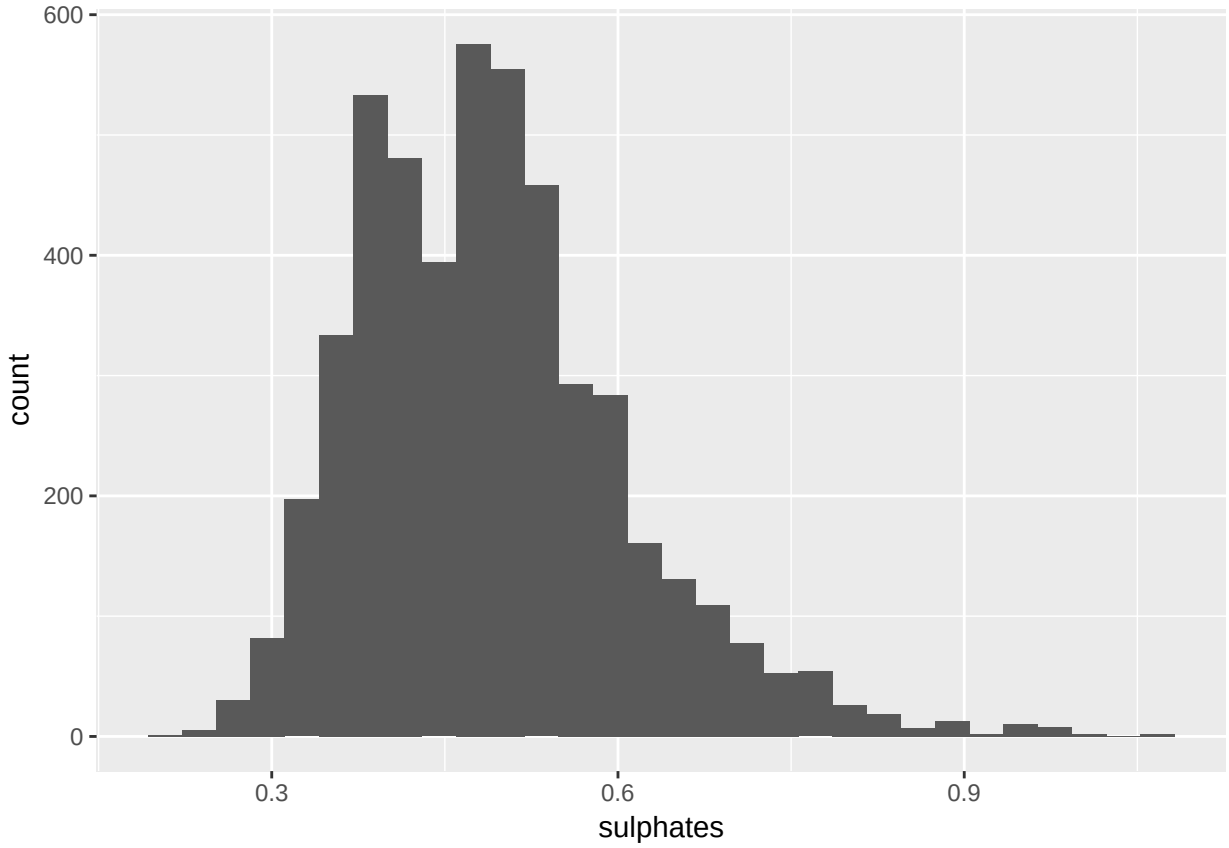
To investigate why the distribution is right skewed, we plot the square root transformation. It is suggesting that total sulfur dioxide has a square rooted gaussian distribution.



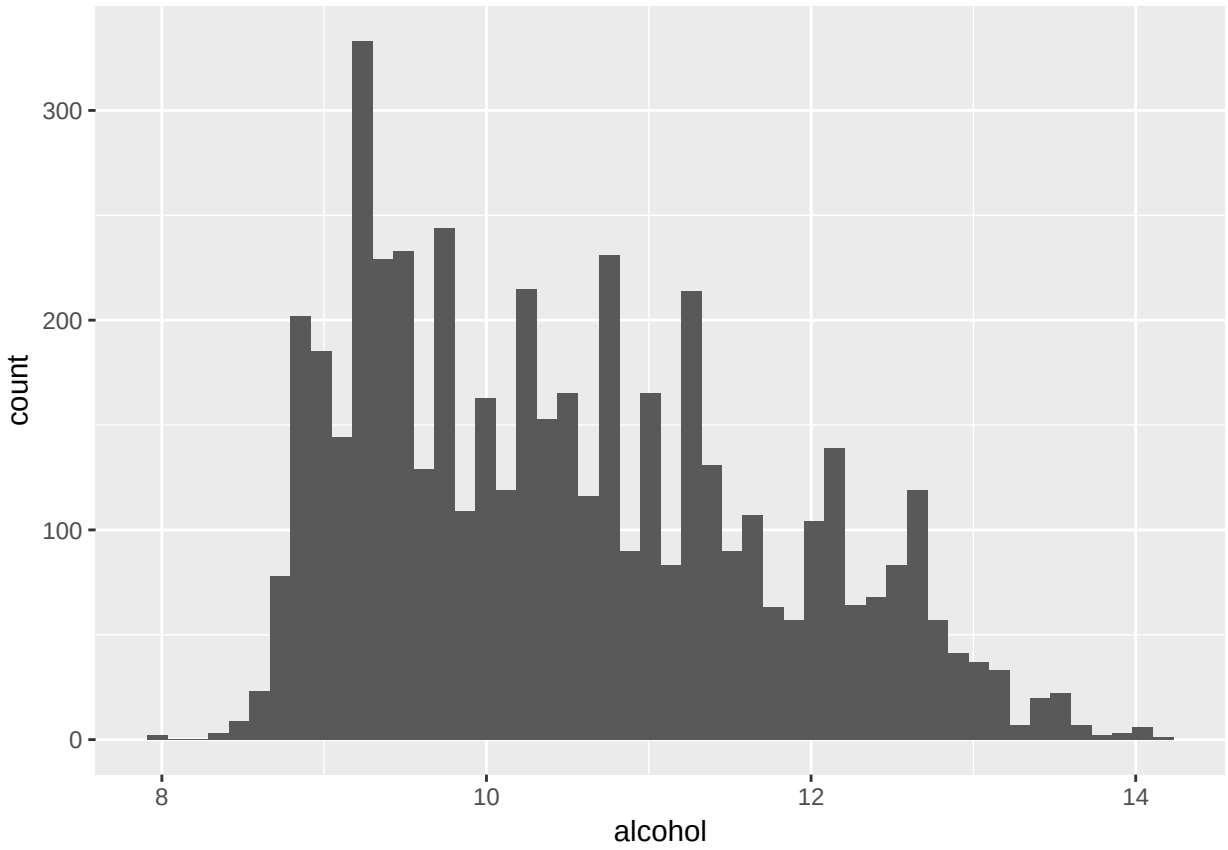
The density seems to be a gaussian distribution



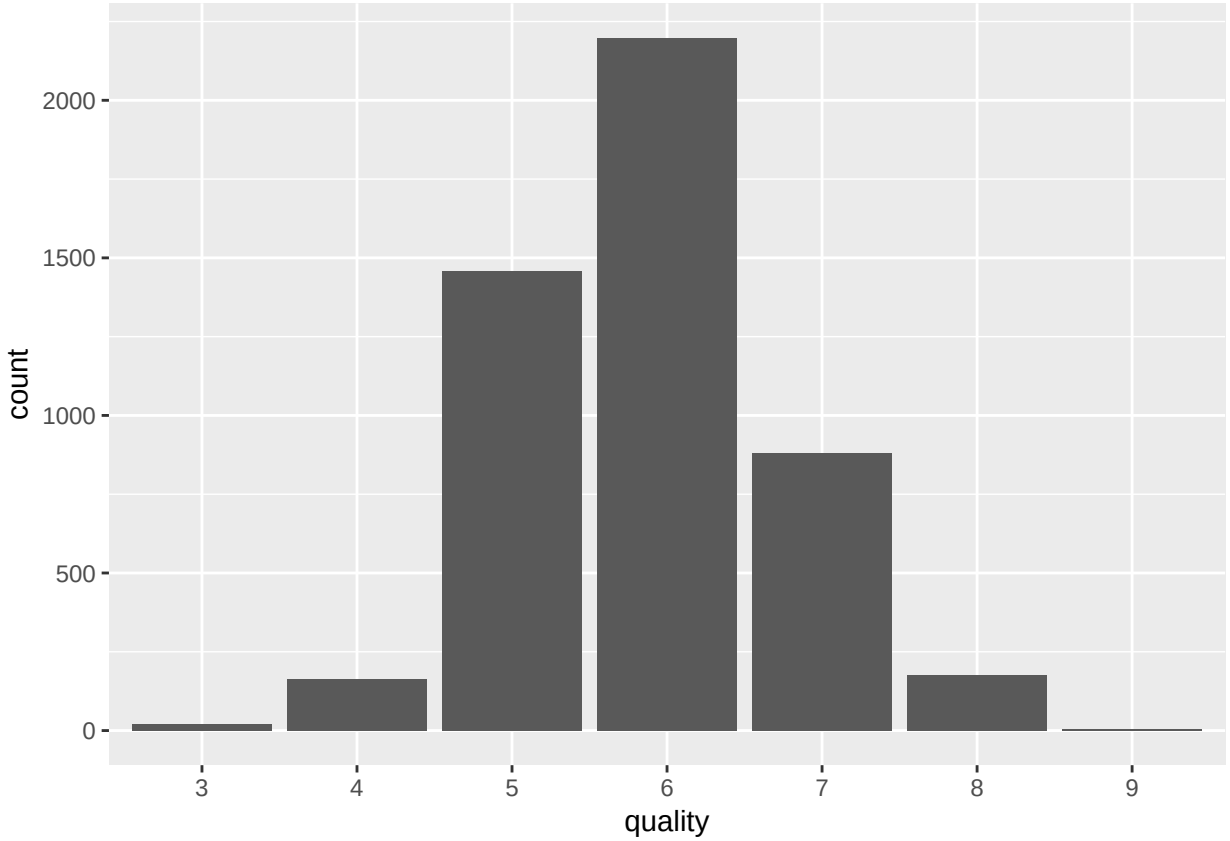
The density seems to be a symmetric distribution



The sulphates seems to be a bit right skewed.



The alcohol measure seems to be nearly evenly distributed in the range 8 to 14 degrees.



The quality measure seems to be normally distributed.

Univariate Analysis

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

The dataset is made of 11 numerical features and 1 categorical feature (wine quality).

I found that lots of variables seem to be normally distributed or to be right skewed, like a square root transform or a log transform of normal distribution. However, even with 4000+ more records, the distributions we see are noisy: there are modes (or pikes) that render the visual identification of a distribution less easy. A multivariate analysis might help to refine that.

The resiudual sugar feature seems to have a very peculiar distribution compared to the others. It might help create a new categorical feature (sweet or not sweet).

I will try to investigate the relationship between the most normally distributed variables (like acidity, chlorides, sulfur and density related features)

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

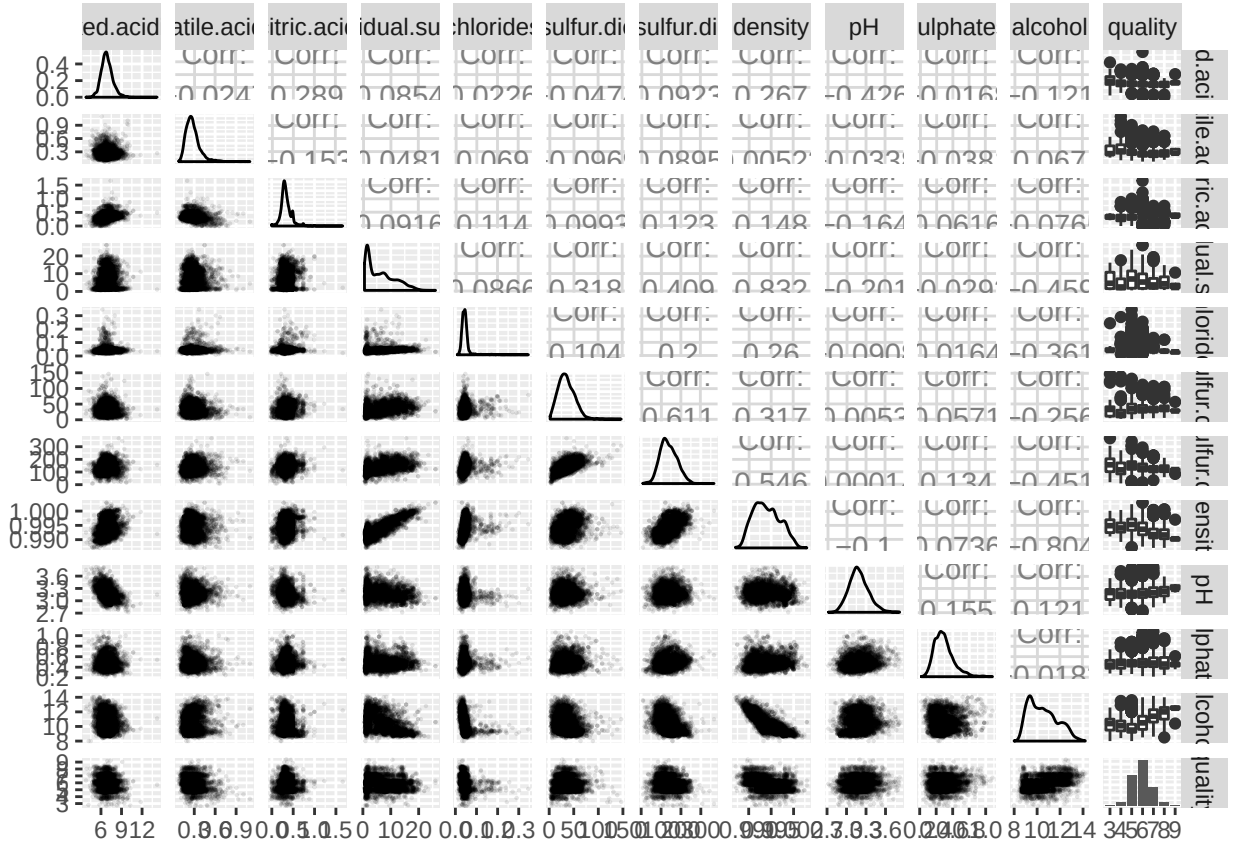
I didn't create new variables in the dataset, however I might try to do a log transform on the volatile acidity feature and square root transform on the total sulfur dioxide feature.

Of the features you investigated, were there any unusual distributions?

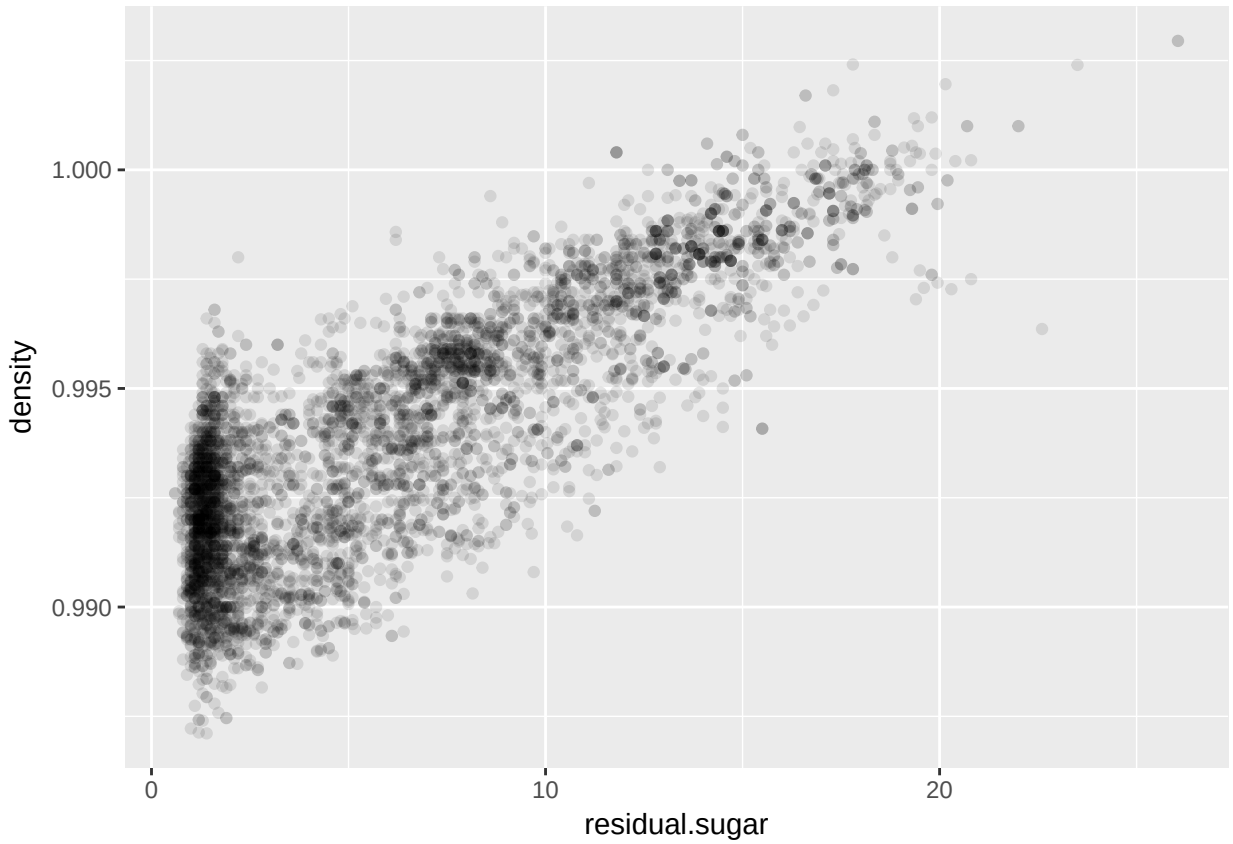
The most unusual distributions are the residual sugar one and the alcohol one. In the wine making process, those to variables are deeply intertwined and the wine makers are voluntarily altering those two. This could explain why those two variables are not normally distributed.

Bivariate Plots Section

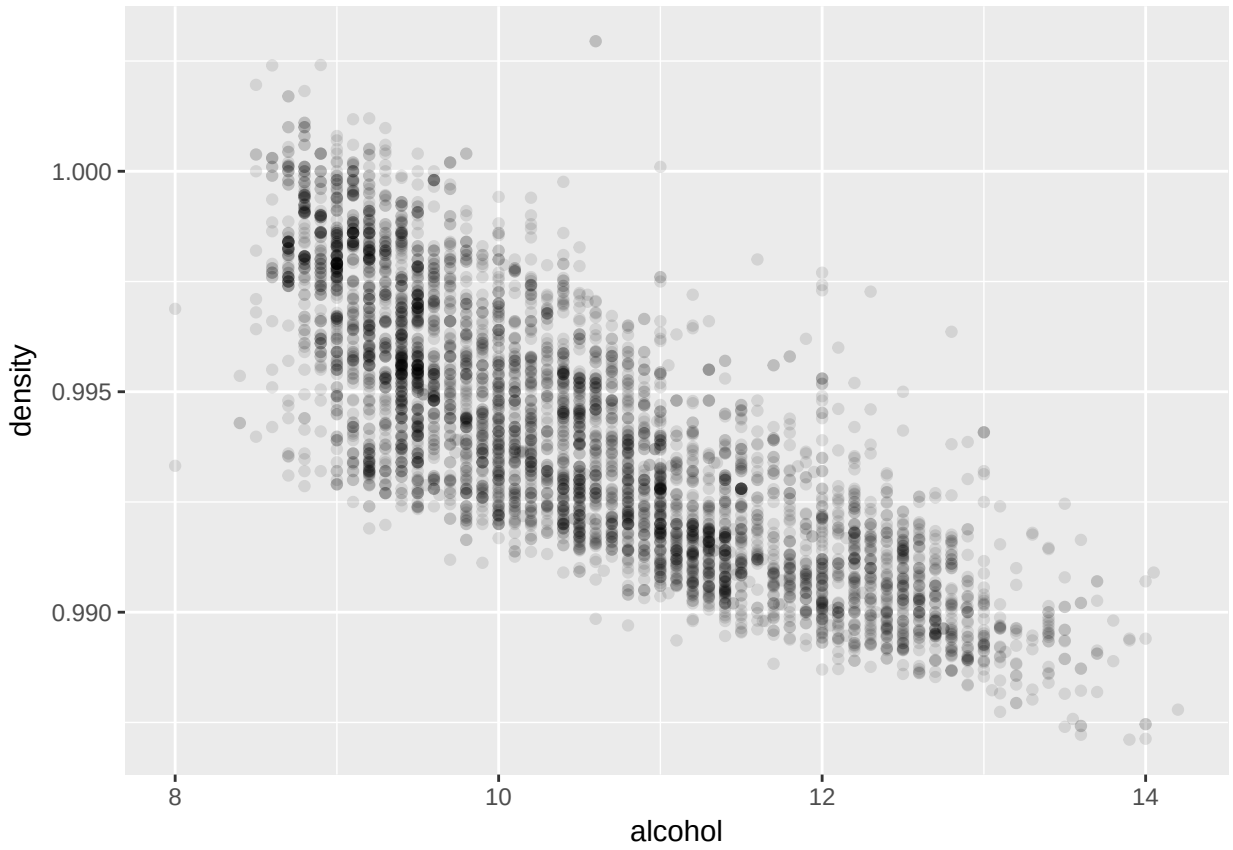
In this section, some more deeper analysis of variables with bivariate plots



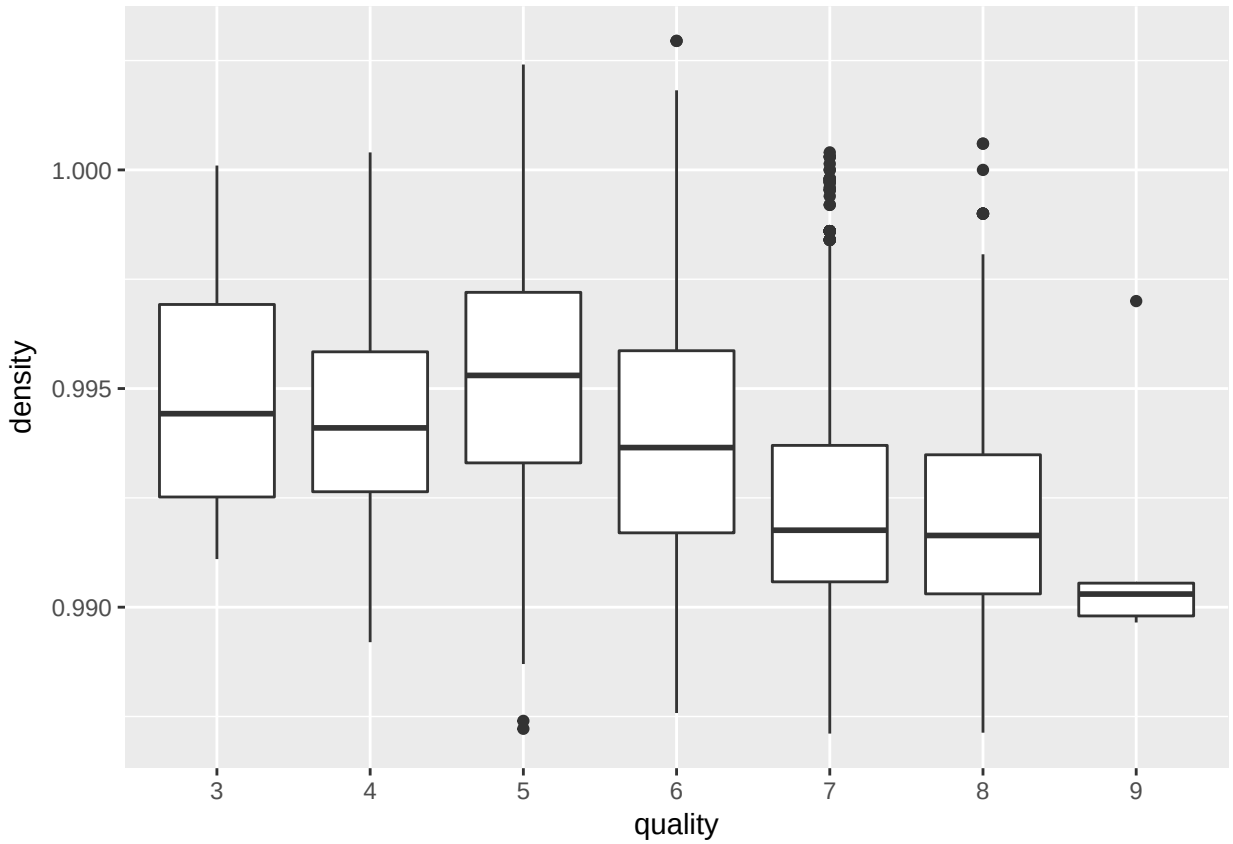
I don't see much obvious relationships between variables, appart from residual sugar, density and alcohol, which seem to be related. Density and alcohol seem to have a pretty weak relationship with quality.

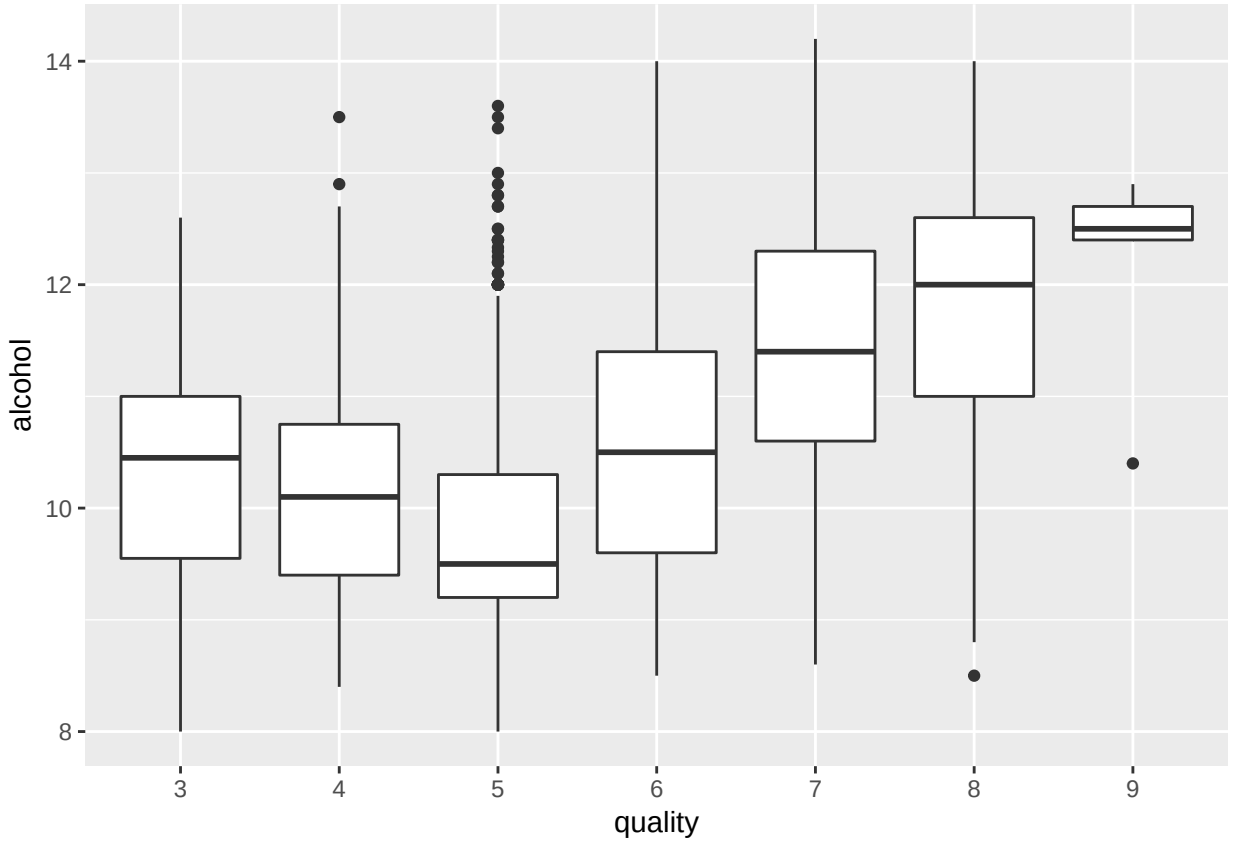


There is a clear positive relationship between residual sugar and density. Density increases with higher residual sugar concentration.

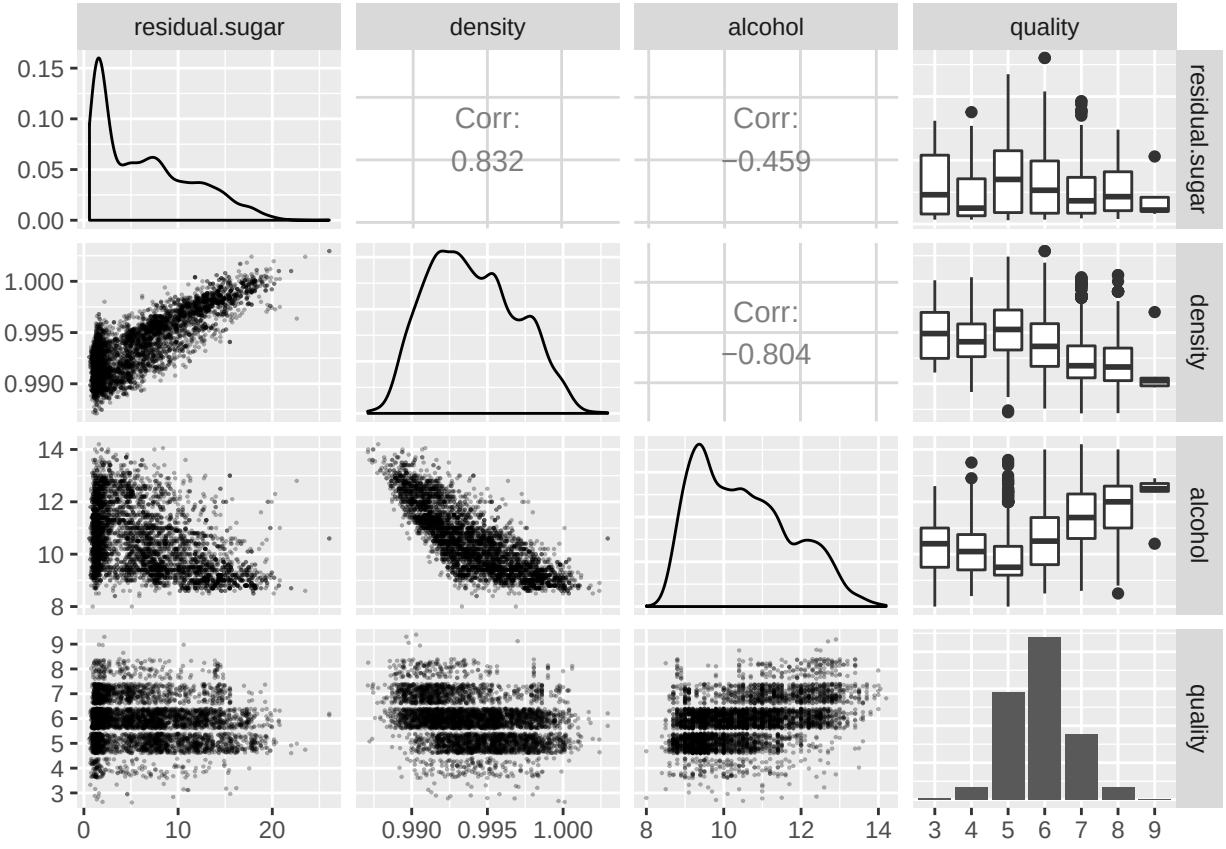


There is a clear negative relationship between alcohol and density. Density decreases with higher alcohol content.


1.000
0.995
0.990
density
3
4
5
6
7
8
9
quality


14
12
10
8
alcohol
3
4
5
6
7
8
9
quality

Those two boxplots seem to suggest that there is a relationship between quality, alcohol and density. Wines with more quality seem to all have a higher alcohol content and a lower density. It would be interesting to see if the relationship between density and quality is only due to the dependency of the density on alcohol.



Bivariate Analysis

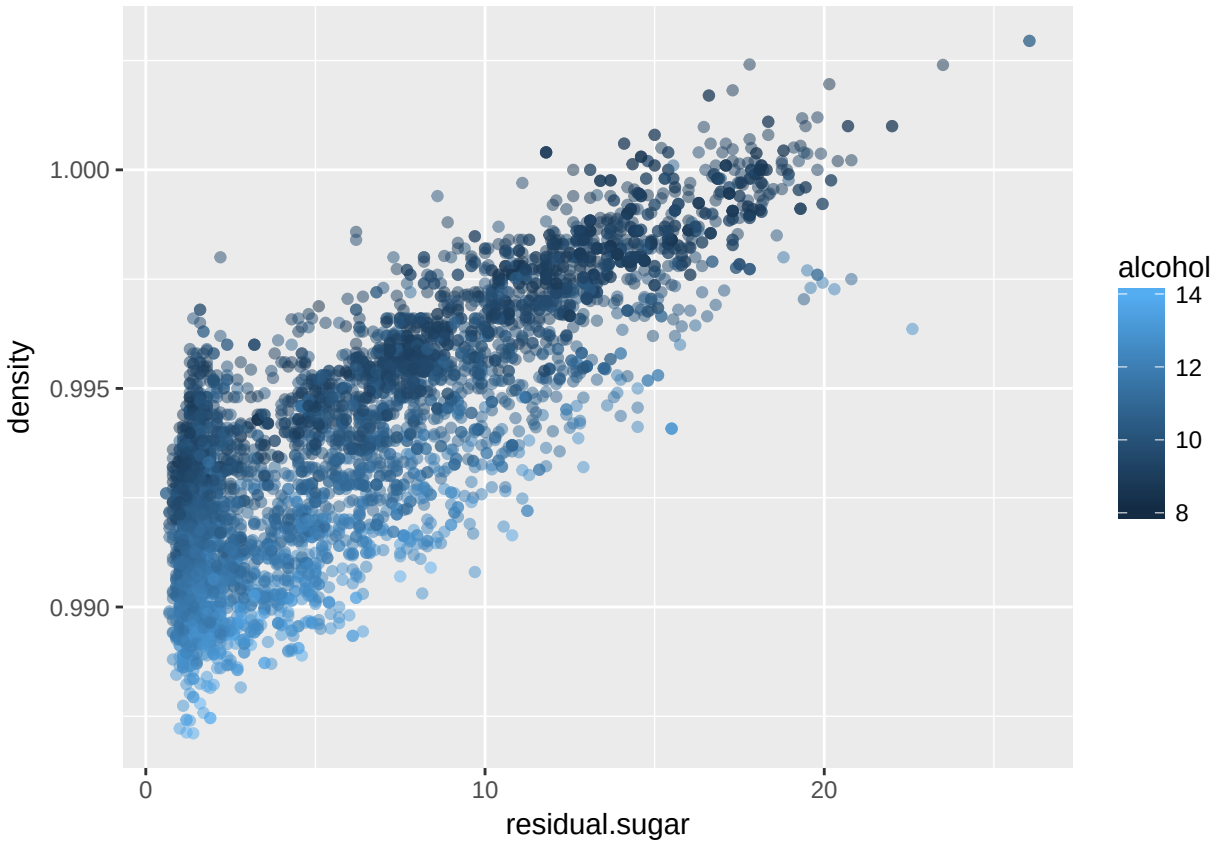
Relationships

There is a strong multivariate relationship between residual sugar, alcohol and density.

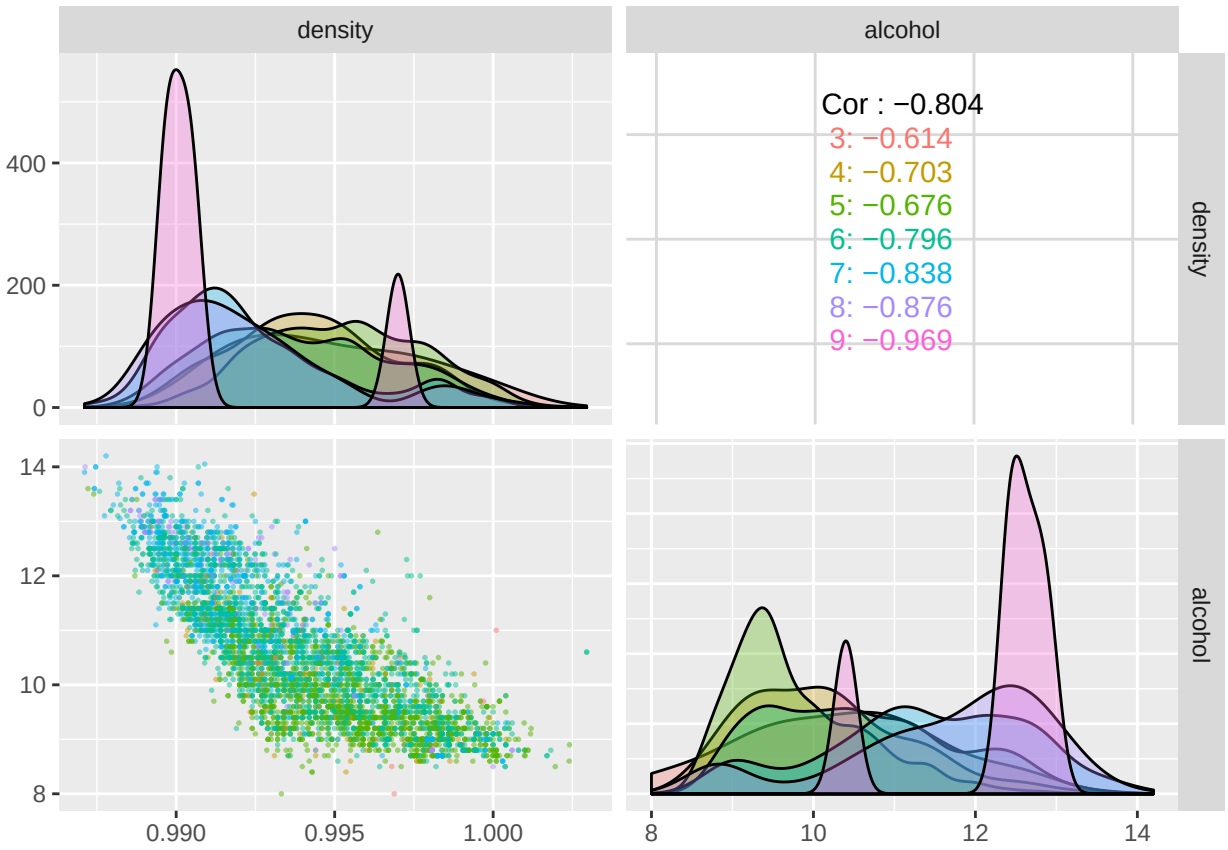
Those last two variables (alcohol, density) seem to be partially linked to quality. It would be interesting to investigate the link of the three numerical features (residual sugar, alcohol, density) with the categorical "quality" feature.

Multivariate Plots Section

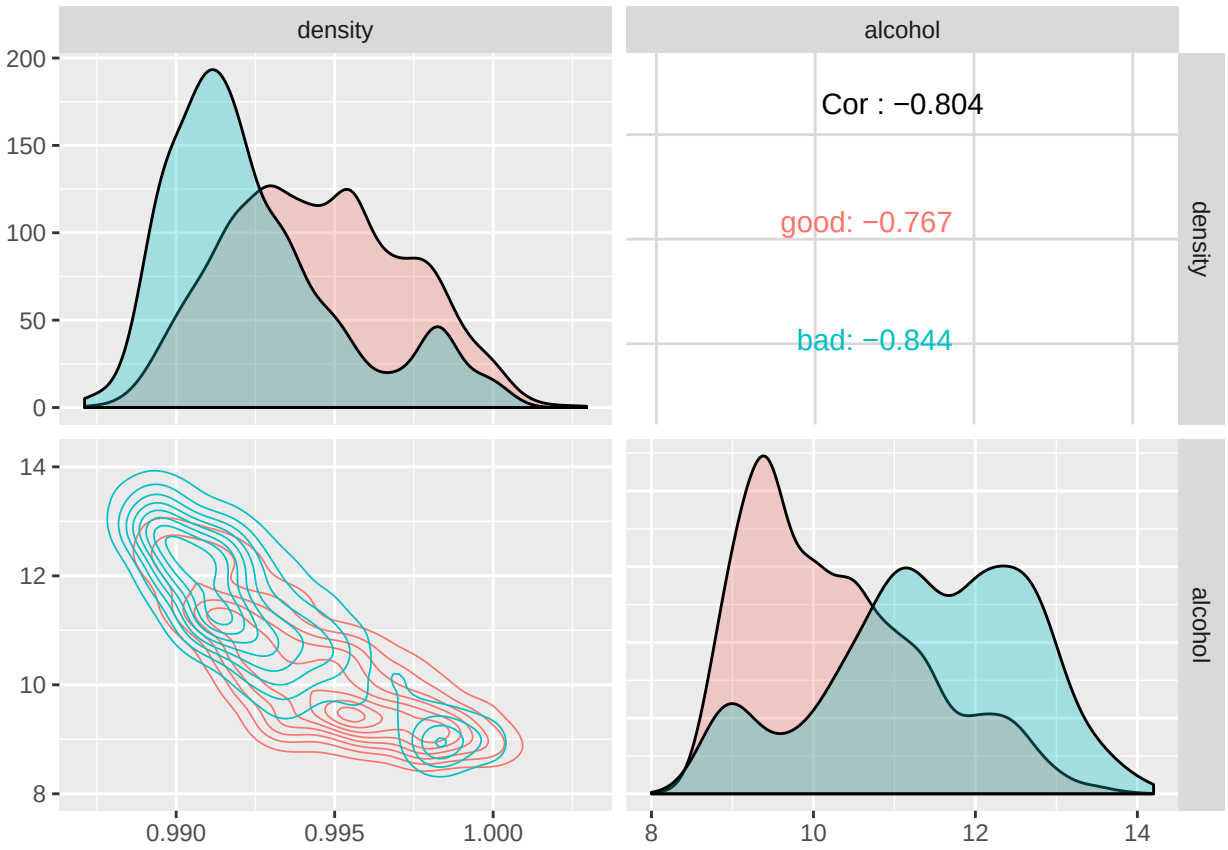
In this section, deeper analysis of variable dependencies through multivariate plots.



This plot shows pretty clearly the cumulative impact of alcohol (color) and residual sugar (x axis) on the density. It seems that there is linear relationship between those 3 features.



This plot intends to show the dependency of quality on alcohol and density. Quality is a categorical variable with 7 categories, hence it is quite messy to understand. Let's create a new categorical variable "good/bad" based on a quality above a certain threshold.



Here I chose the threshold "quality > 6" for a wine to be categorized as "good". We see a pretty strong dependency of the good wines on density and alcohol. Good wines have a low density and high alcohol.

Multivariate Analysis

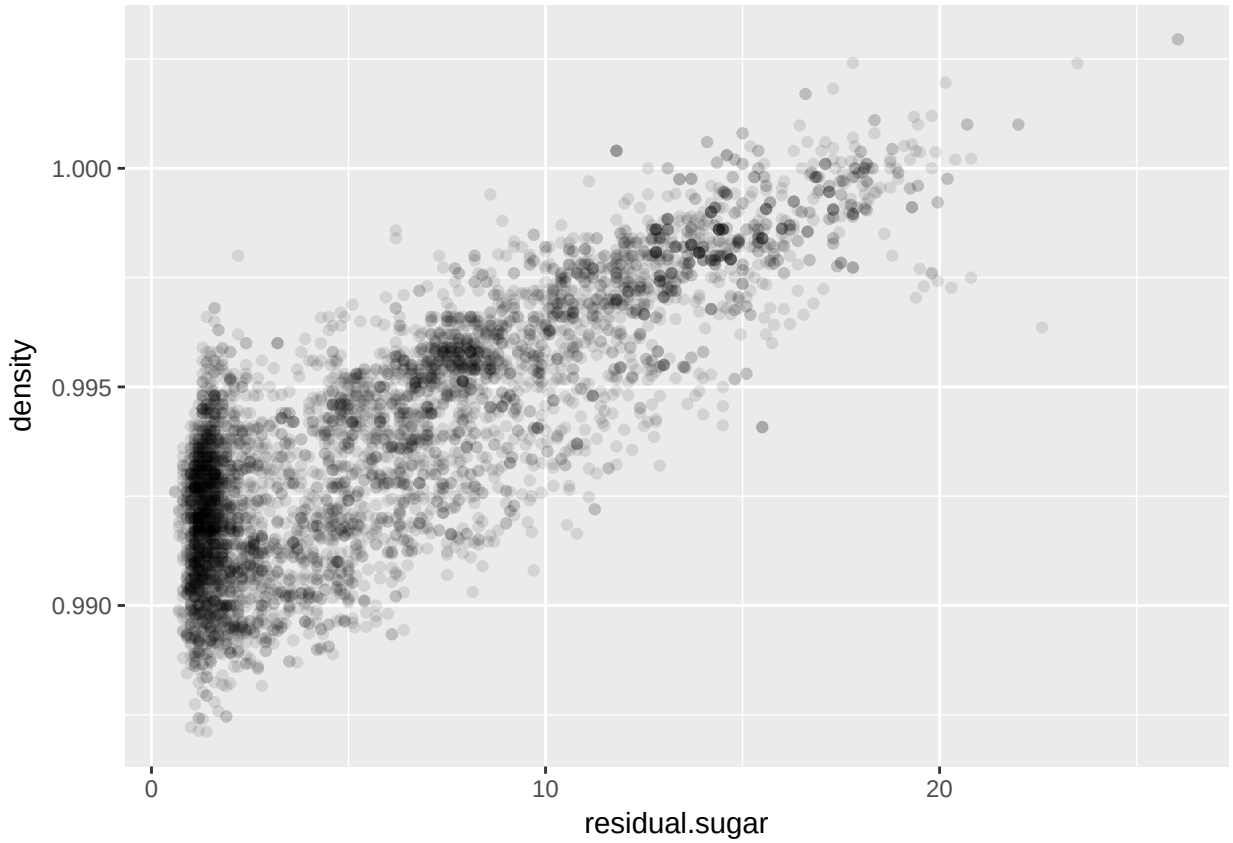
Relationships

The multivariate analysis confirms the relationship we saw between density, residual sugar and alcohol, during the bivariate analysis.

We also saw the cumulative effect of density and alcohol content on quality, by creating a binary quality feature.

Final Plots and Summary

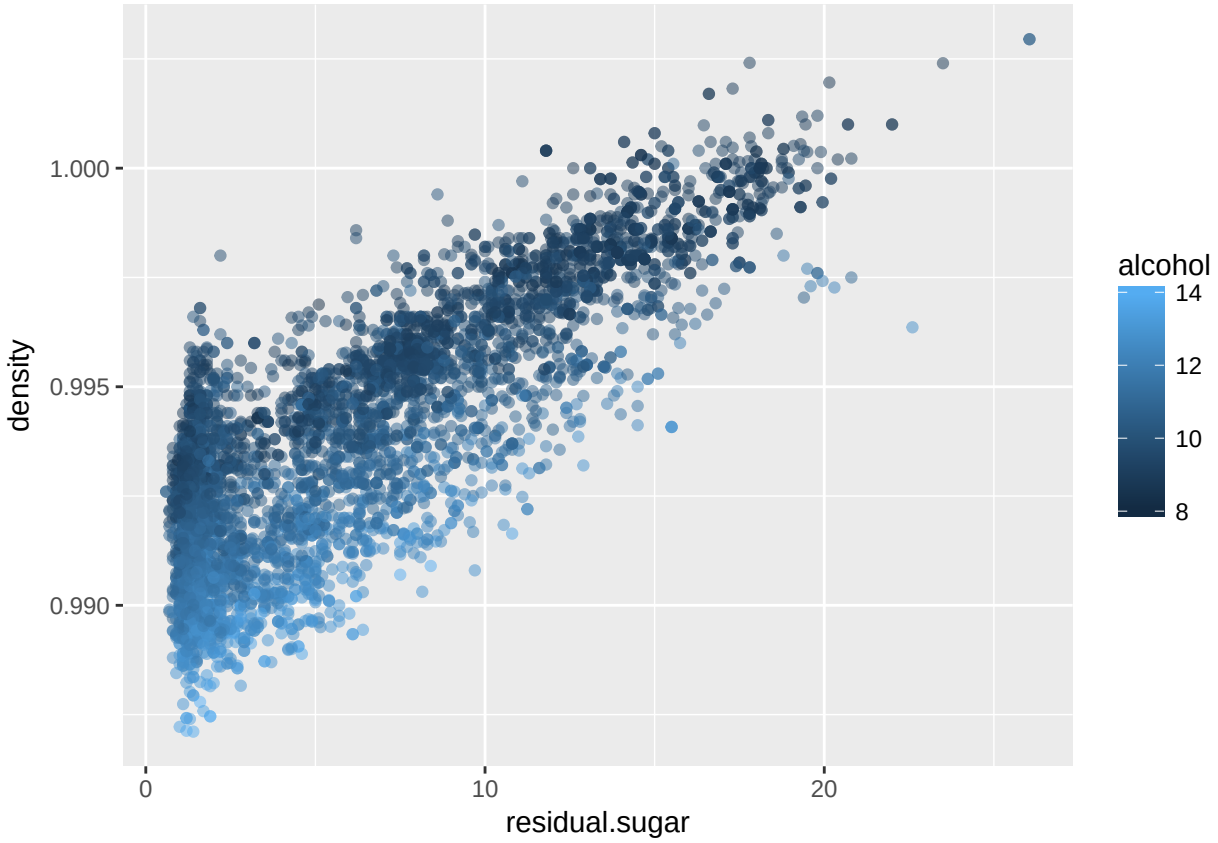
Plot One



Description One

This plot shows that there is a strong dependency between the resiudal sugar content and the density of the wine. But wait, is this the only feature influencing the density ? Shouldn't alcohol play a role ?

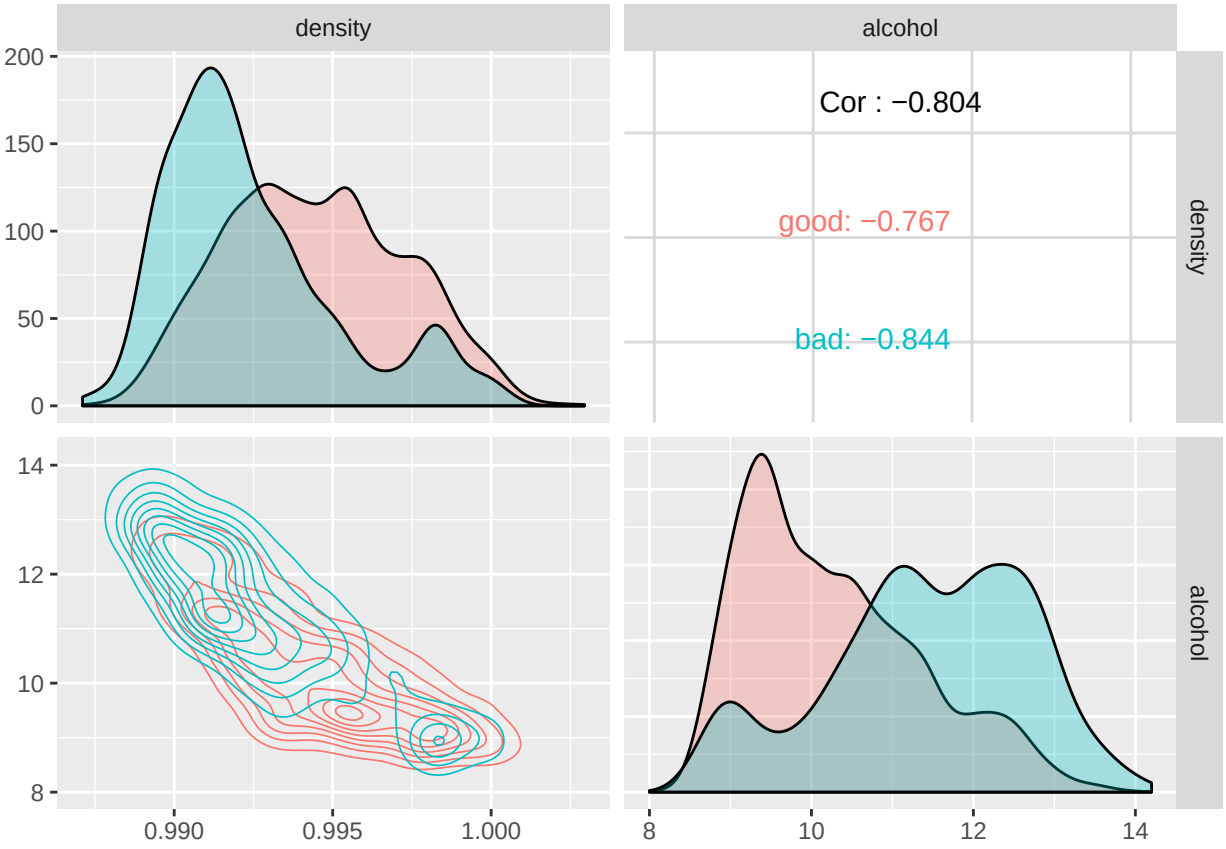
Plot Two



Description Two

This plot shows that the effect of alcohol on density are cumulative with residual sugar content. The higher the sugar content and the lower the alcohol, the higher the density.

Plot Three



Description Three

This plot shows the dependency of quality on alcohol and density. We created a new binary feature: "good" if quality > 6 and "bad" otherwise. This plot shows that good wines have a tendency to have a high alcohol content and a low density.

Reflection

It might be interesting to explore some transformation (log or square root) on some features to see if it would reinforce the relationship we investigated.

In the end, building a linear model predicting quality of the wine based on density and alcohol content seems promising.