White Wine Quality Exploration by Shirley Chen

> **Dataset Introduction**: The White Wine Quality dataset is a public dataset that was created by Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009. This tidy data set contains 4,898 white wines with 11 variables on quantifying the chemical properties of each wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

Univariate Plots Section

Dataset overview

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

The dataset contains 4898 obs. of 13 variables; these variables are either in numeric or integar format.

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
```

```
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

Summary command provides a quick look of the structure of each variable.

Closer looks at uni-variables

create_hist function

As taking closer looks at each variables I will need to create multiple plots, it will be great to define plotting function to reduce repetitive works. In below chunk I define a function which takes in variable name along with plot title, and outputs histogram.

```
create_hist <- function(variable, title) {
  return(ggplot(data = ww, aes_string(x = variable)) +
                  geom_histogram(color = 'black', fill = '#7faeff') +
                  ggtitle(title) +
                  theme(plot.title = element_text(hjust = 0.5)))
}
```

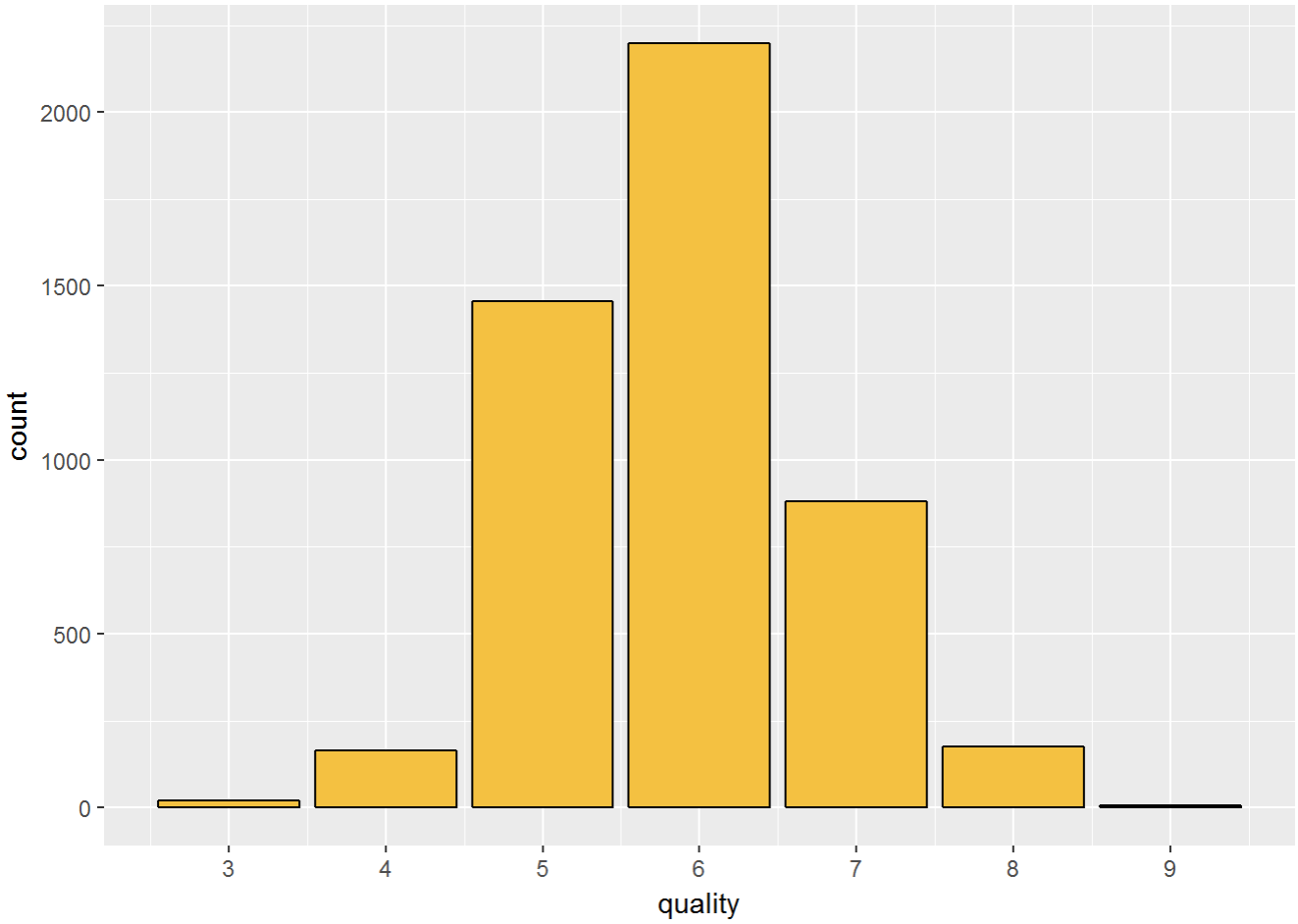
zoom_hist function

Also, considering there's a possibility to zoom in histogram, I create a function that takes axis limits and breaks for future use.

```
zoom_hist <- function(variable, binwidth, xlim_start, xlim_end, br_start, br_end, br_gap, tit
le) {
  if (missing(xlim_start)) {return(ggplot(data = ww, aes_string(x = variable)) +
            geom_histogram(color = 'black', fill = '#7faeff', binwidth = binwidth) +
            scale_x_continuous(breaks = seq(br_start, br_end, br_gap)) +
            theme(plot.title = element_text(hjust = 0.5)) +
            ggtitle(title))}
  else {
  return(ggplot(data = ww, aes_string(x = variable)) +
            geom_histogram(color = 'black', fill = '#7faeff', binwidth = binwidth) +
            coord_cartesian(xlim = c(xlim_start, xlim_end)) +
            scale_x_continuous(breaks = seq(br_start, br_end, br_gap)) +
            theme(plot.title = element_text(hjust = 0.5)) +
            ggtitle(title))
}}
```

Quality

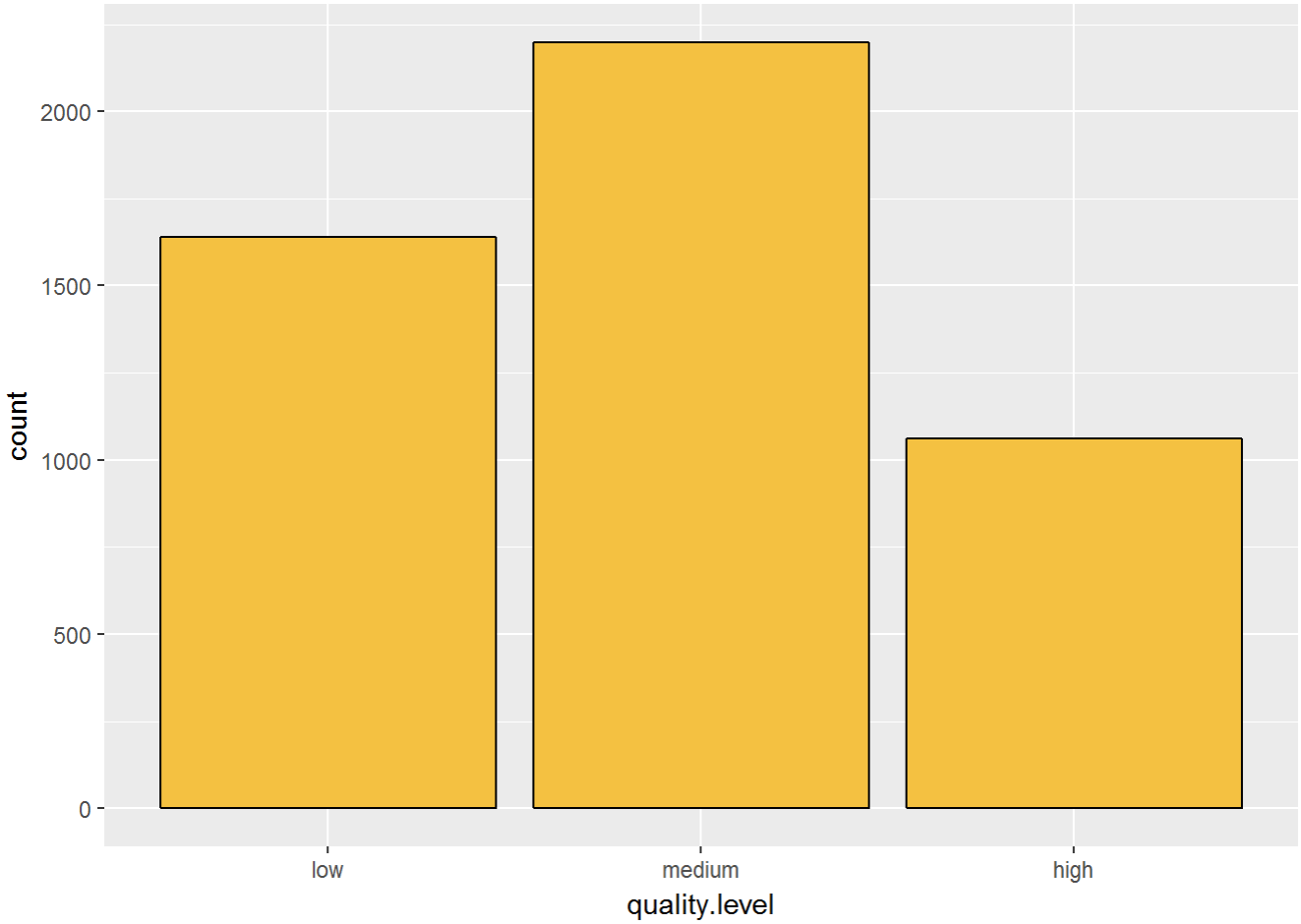
Since the main concern for this dataset is the quality of white wine, it would be a good idea to see how wines are rated at different ratings. As plot in below shows, there are no wines being rated at 0, 1, 2, and 10 points, while there are 2,000+ records are rated at 6.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

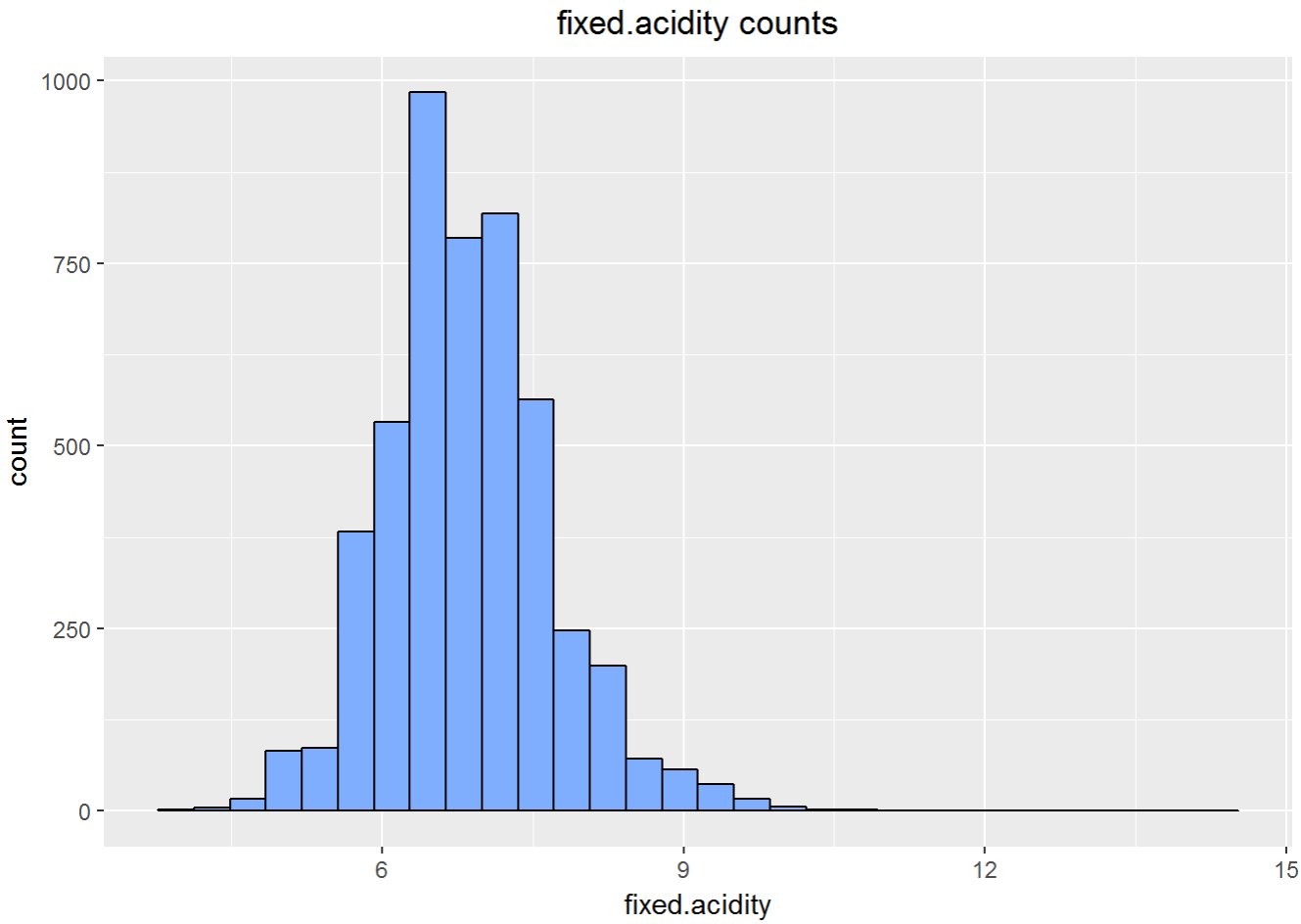
Looking at the quantile of white wine quality, it may be a good idea to cluster quality into three groups (high:7-9, low:3-5, medium:6) as quality.level for future analysis convenience. By clustering, the original

```
##
##    low medium   high
##   1640   2198   1060
```



fixed.acidity

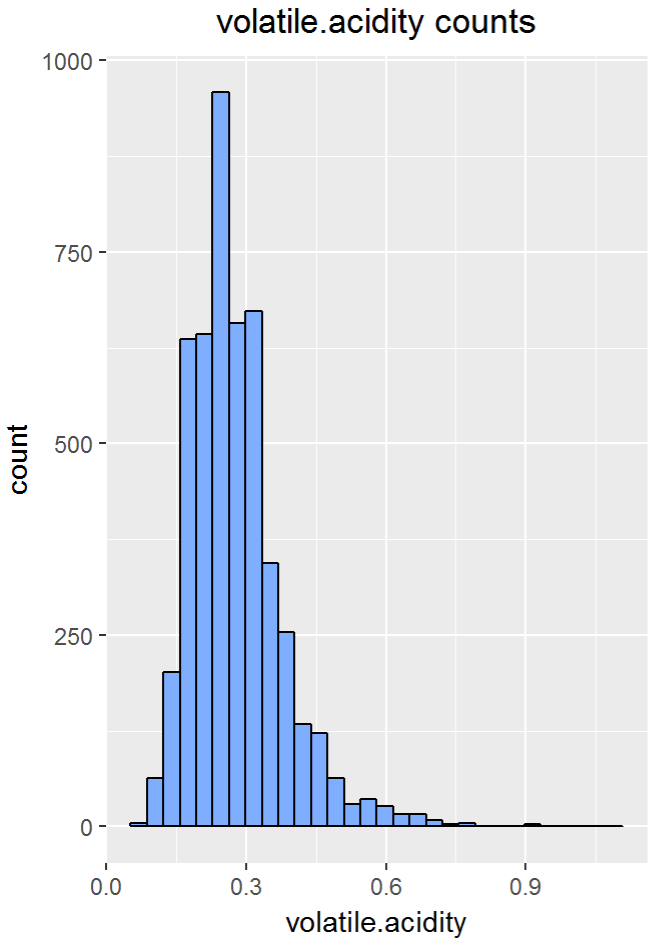
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

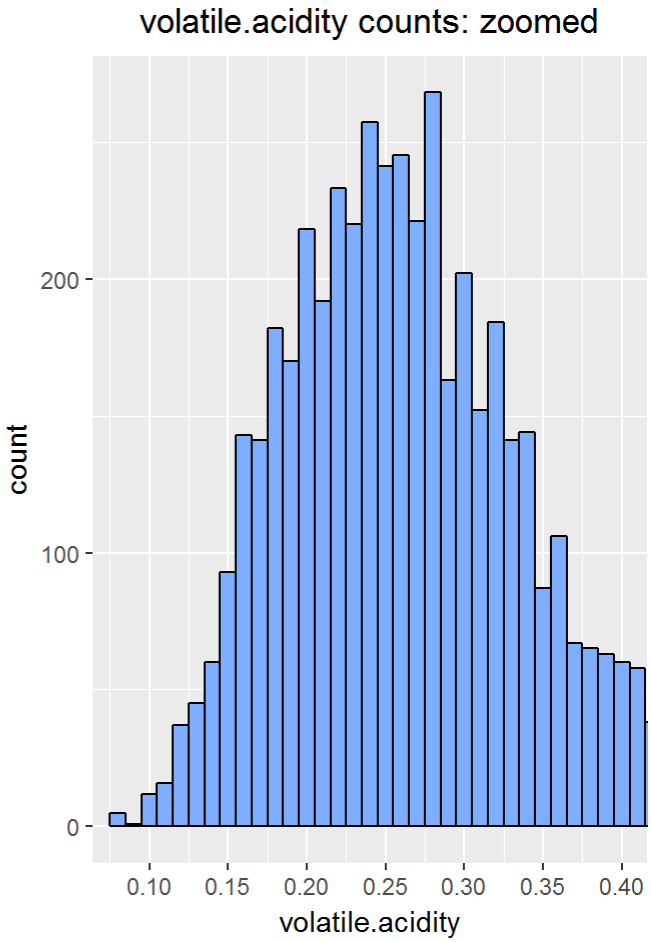


Looking at the plot and summary, we can see that majority for wines have fixed.acidity between 6.3 and 7.3.

volatile.acidity

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```



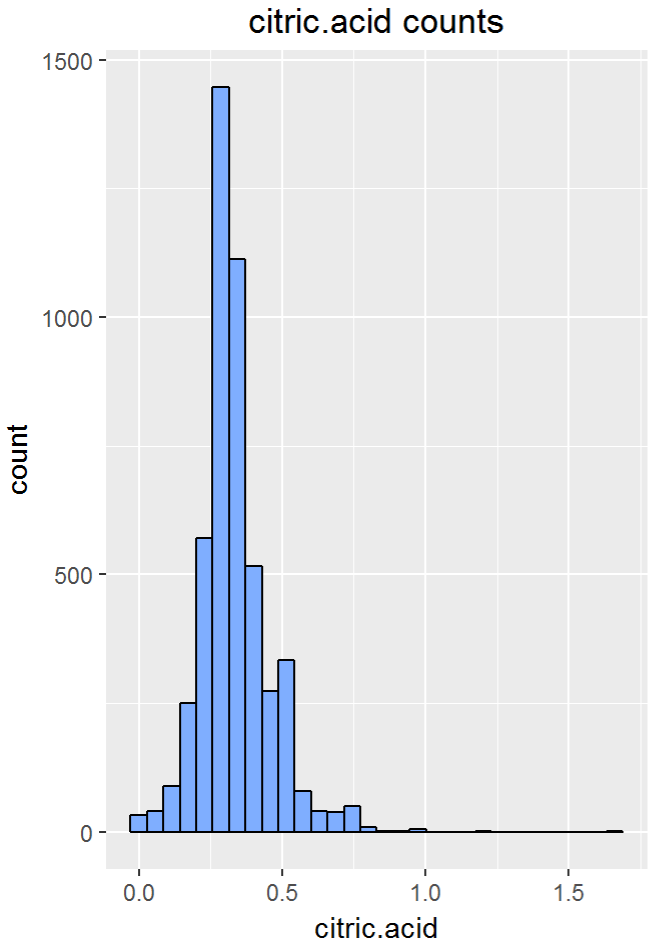


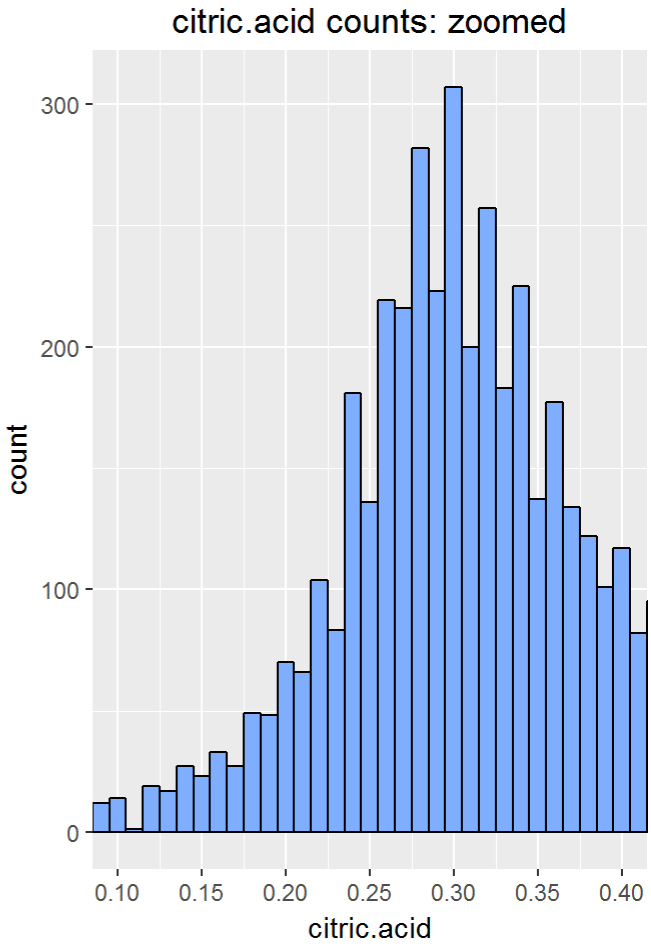
(left plot) Most of the volatile.acidity fall between 0.21 and 0.32, while there is a peak around 0.26: more than 975 wines have volatile.acidity at this rate.

(right plot) Zooming in to most data are at by adding breaks and adjusting binwidte to see if I can find anything different. Looks like there's not a volatile.acidity that is with significantly more wines in particular.

citric.acid

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```



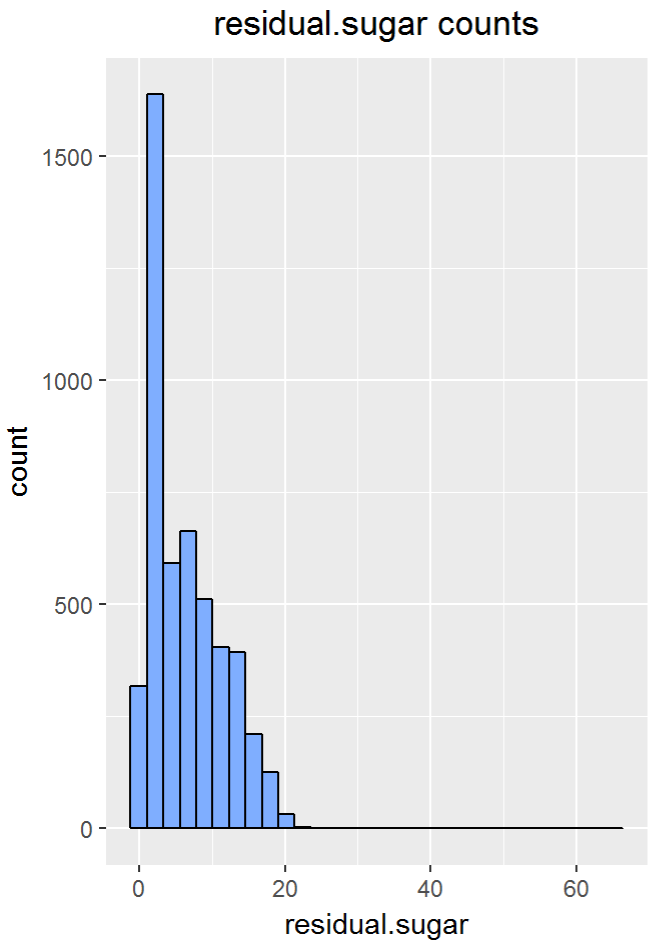


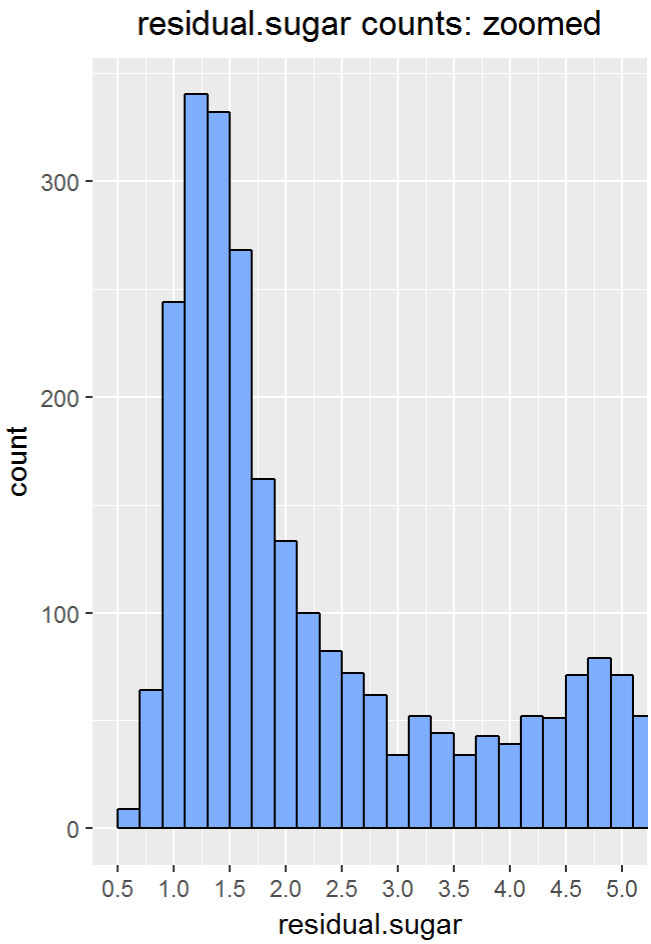
(left plot) Similar to volatile.acidity, there's also a peak for citric.acid. Also we can see citric.acid is basically bell-shaped-distributed.

(right plot) Zoomed in to see if there's something different: When changed binwidth, we can see there are still two peaks (0.28 & 0.3) for citric.acid.

residual.sugar

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

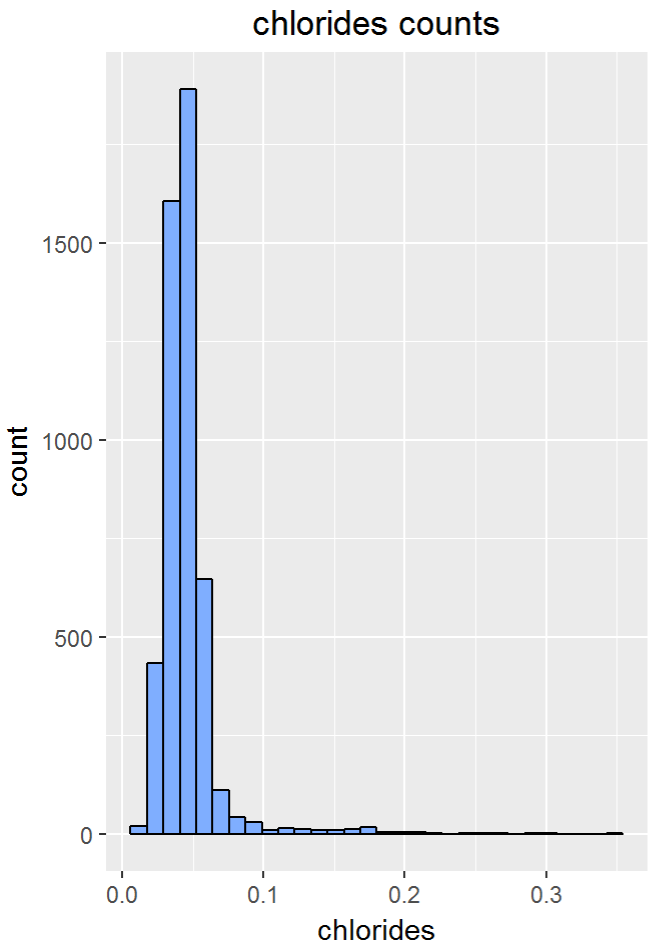


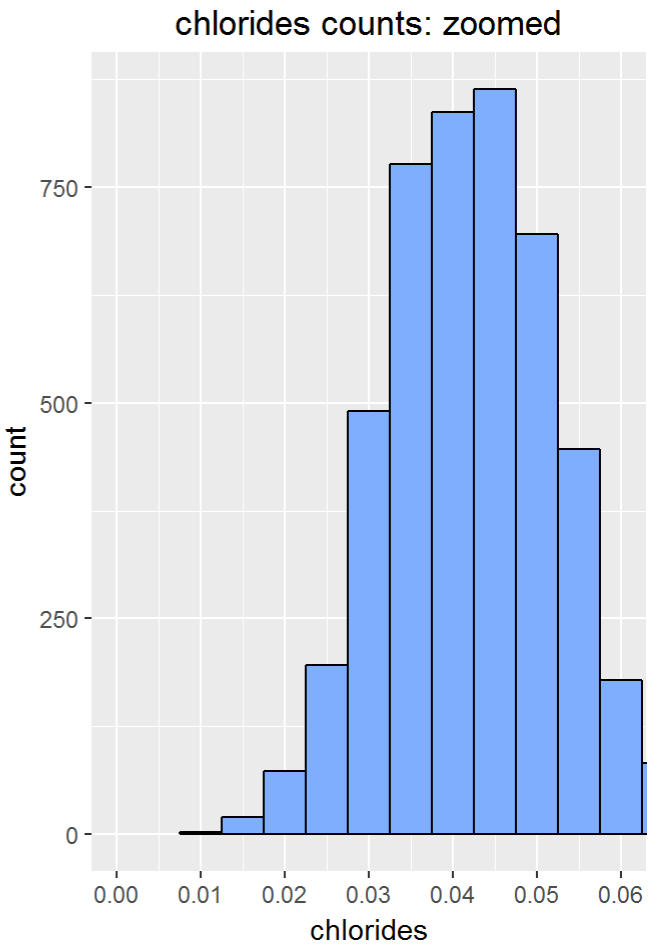


From plots above it is found that 1.2 & 1.4 are the two peaks of residual.sugar.

chlorides

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

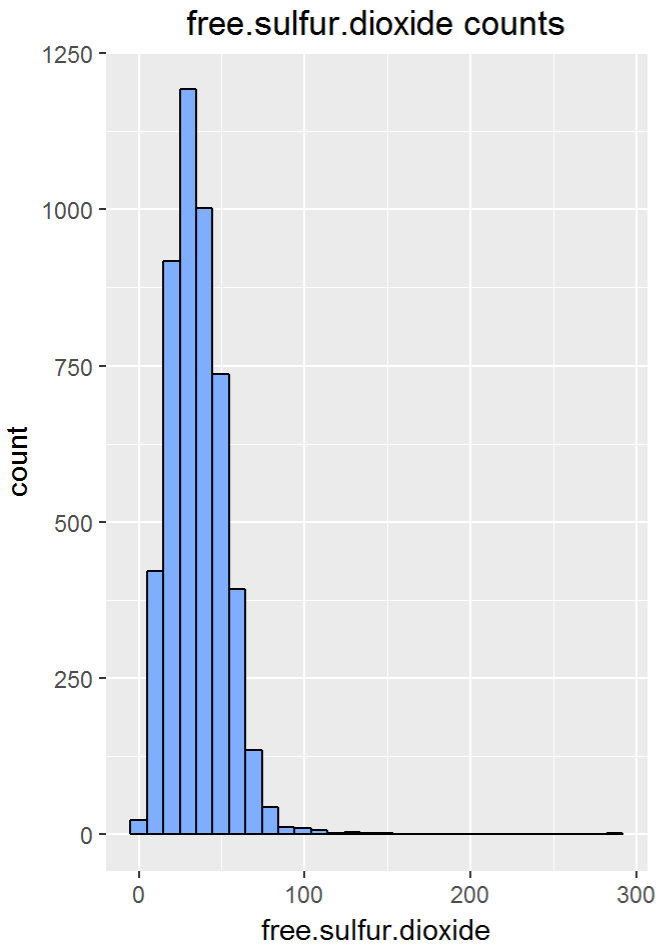


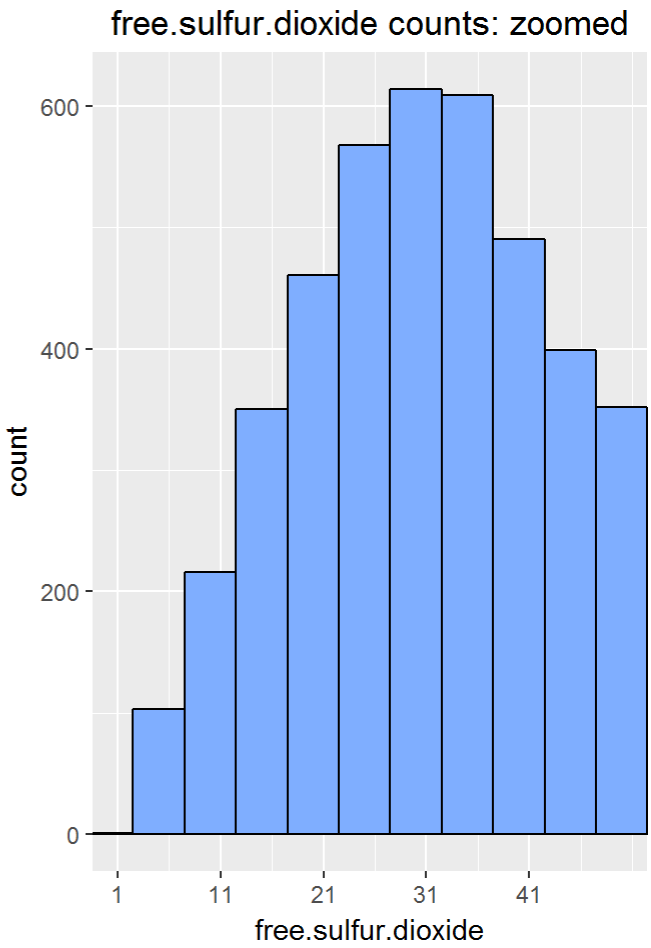


Looks like the values of chlorides are centered around 0.035~0.05, which can also be seen in summary table above.

free.sulfur.dioxide

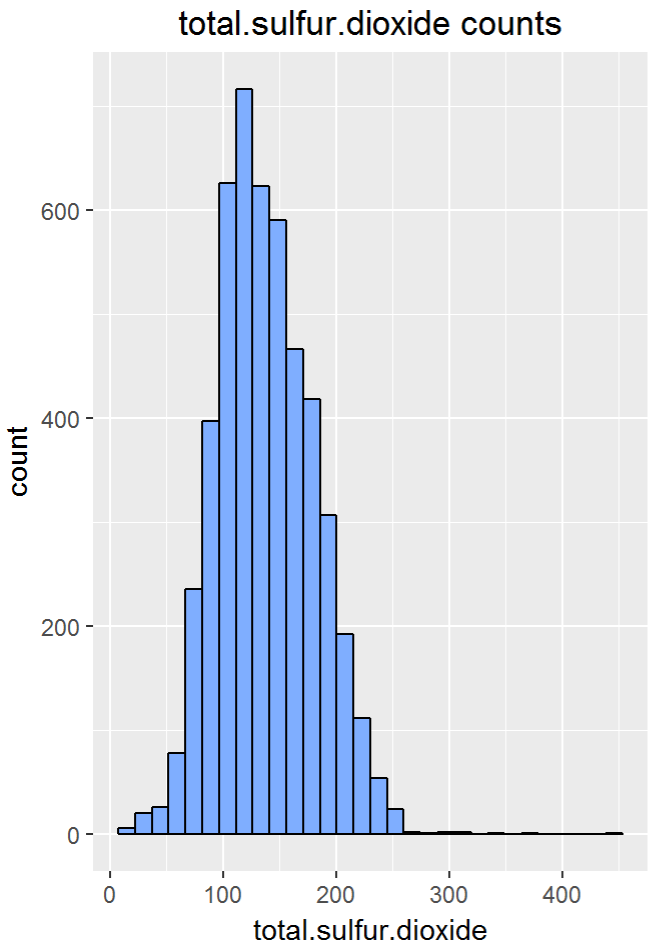
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

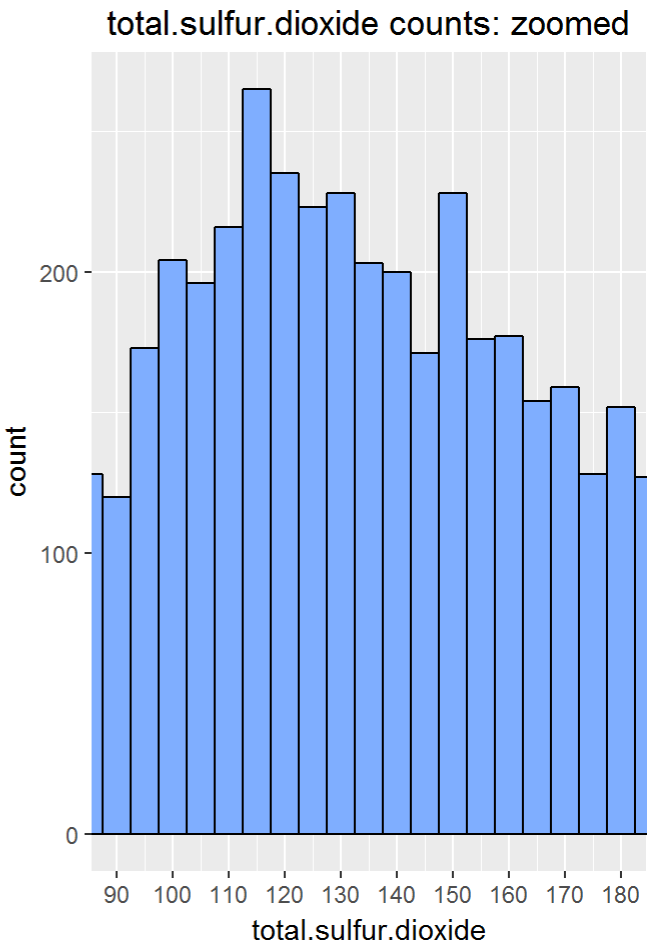




total.sulfur.dioxide

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

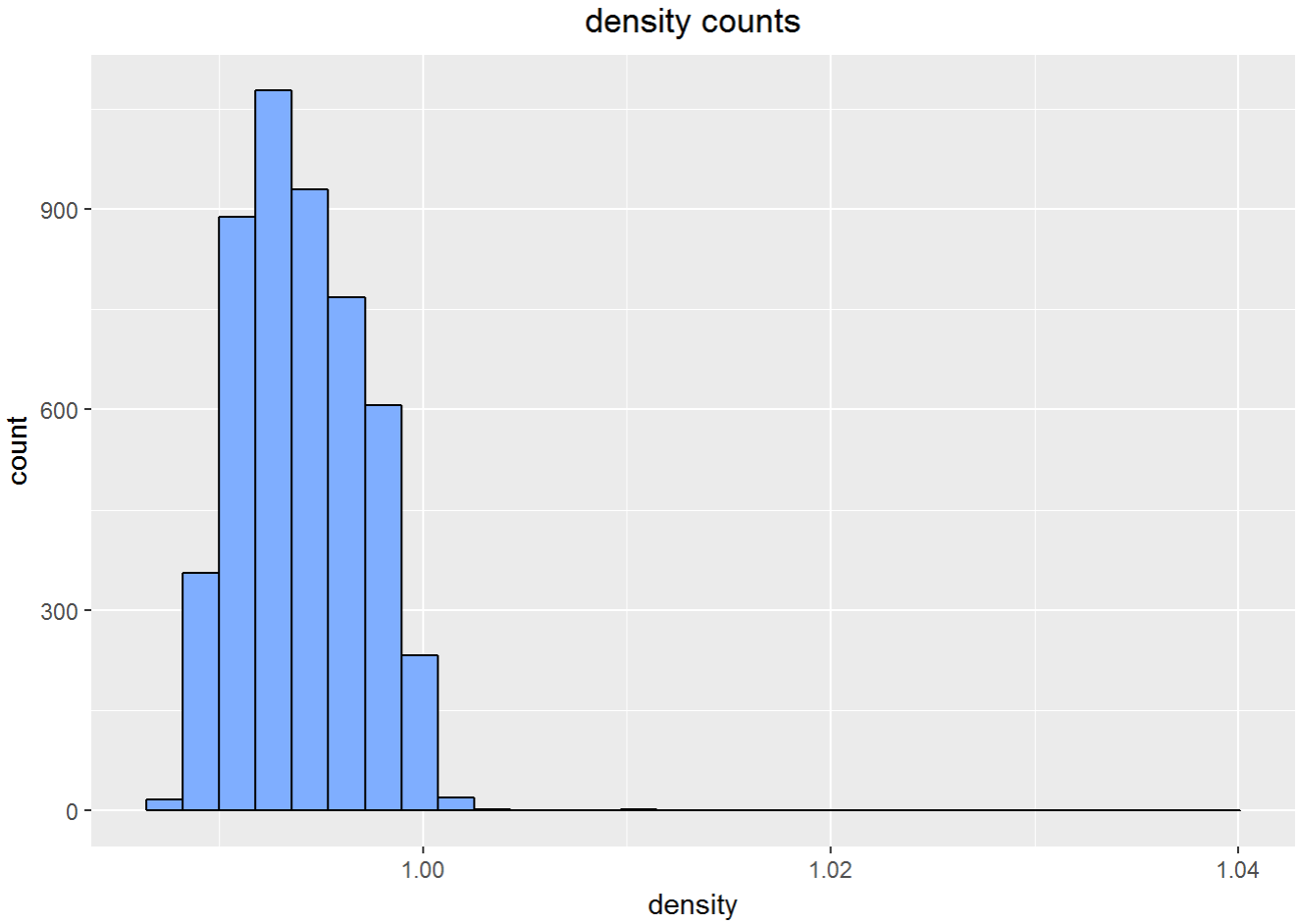




Seems there's a peak around 130. From the zoomed graph above we can see that there'not a particular total.sulfur.dioxide level that is with significantly more wines than others.

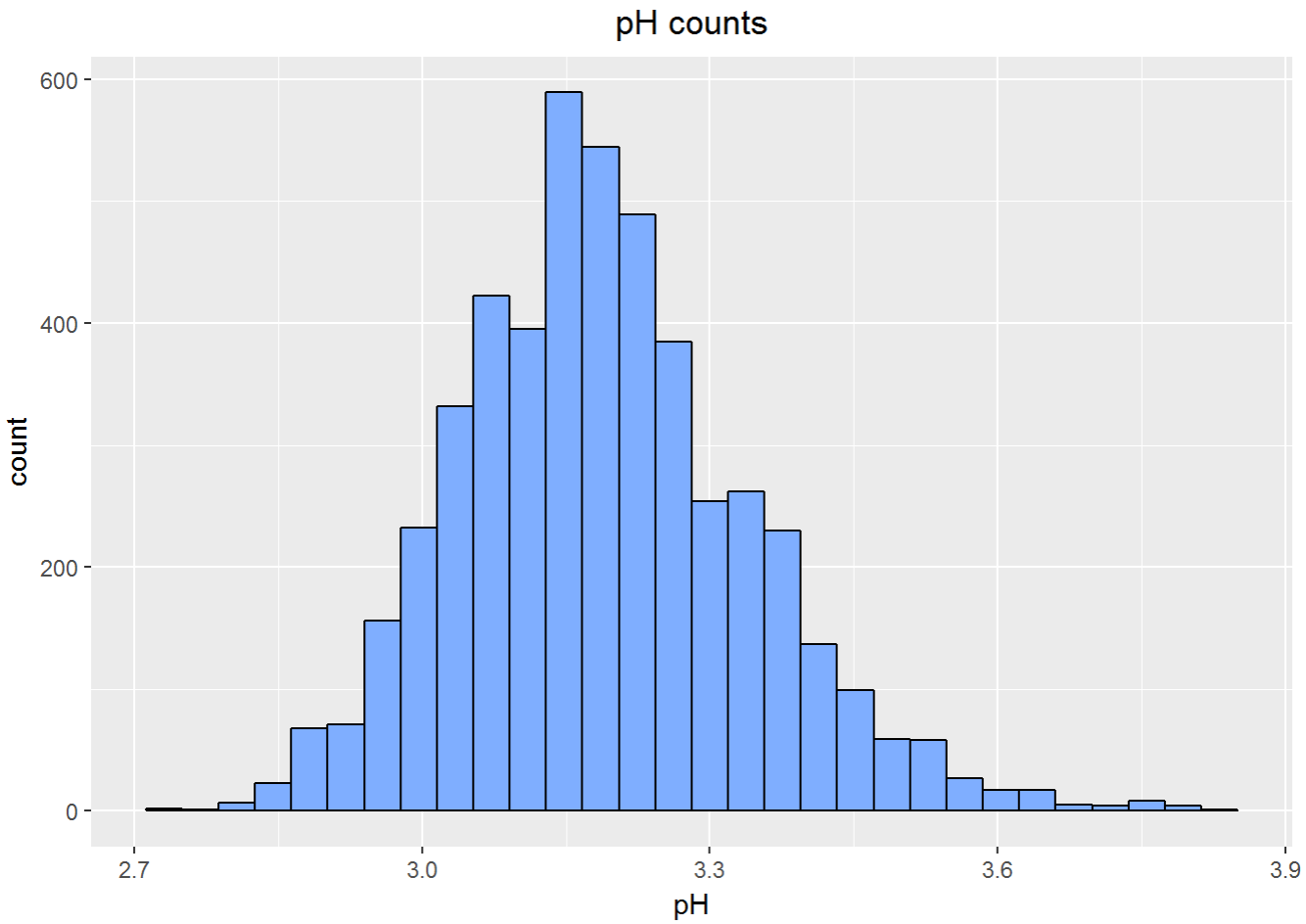
density

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```



pH

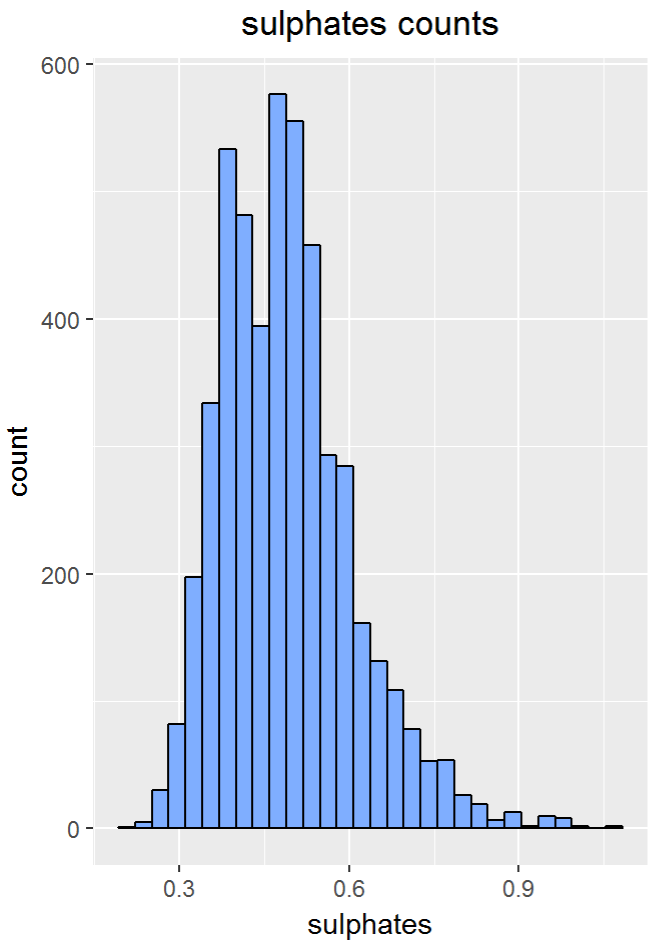
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

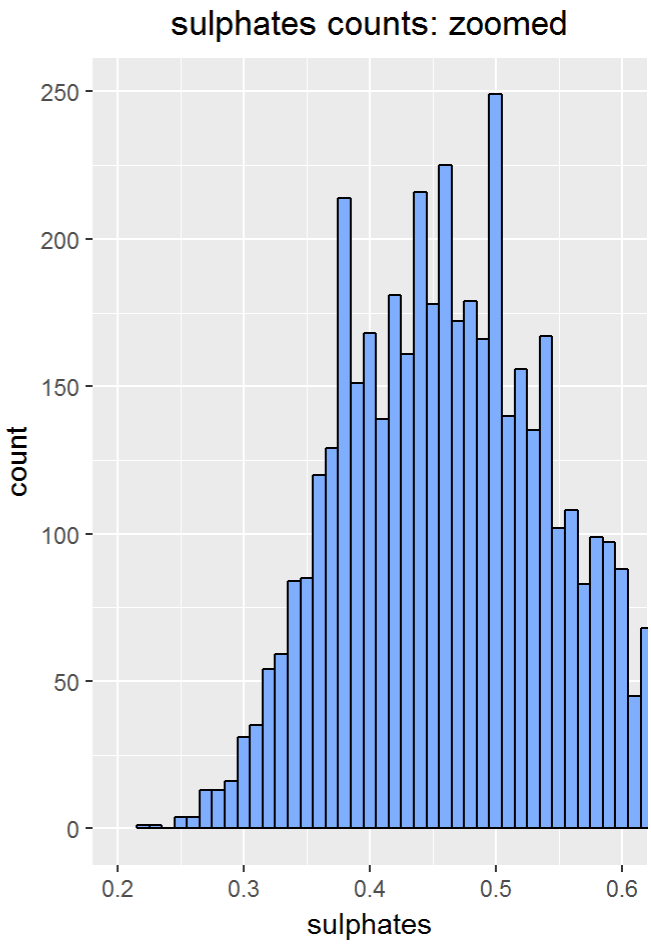


total.sulfur.dioxide, density and pH are all bell-shaped distributed.

sulphates

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```



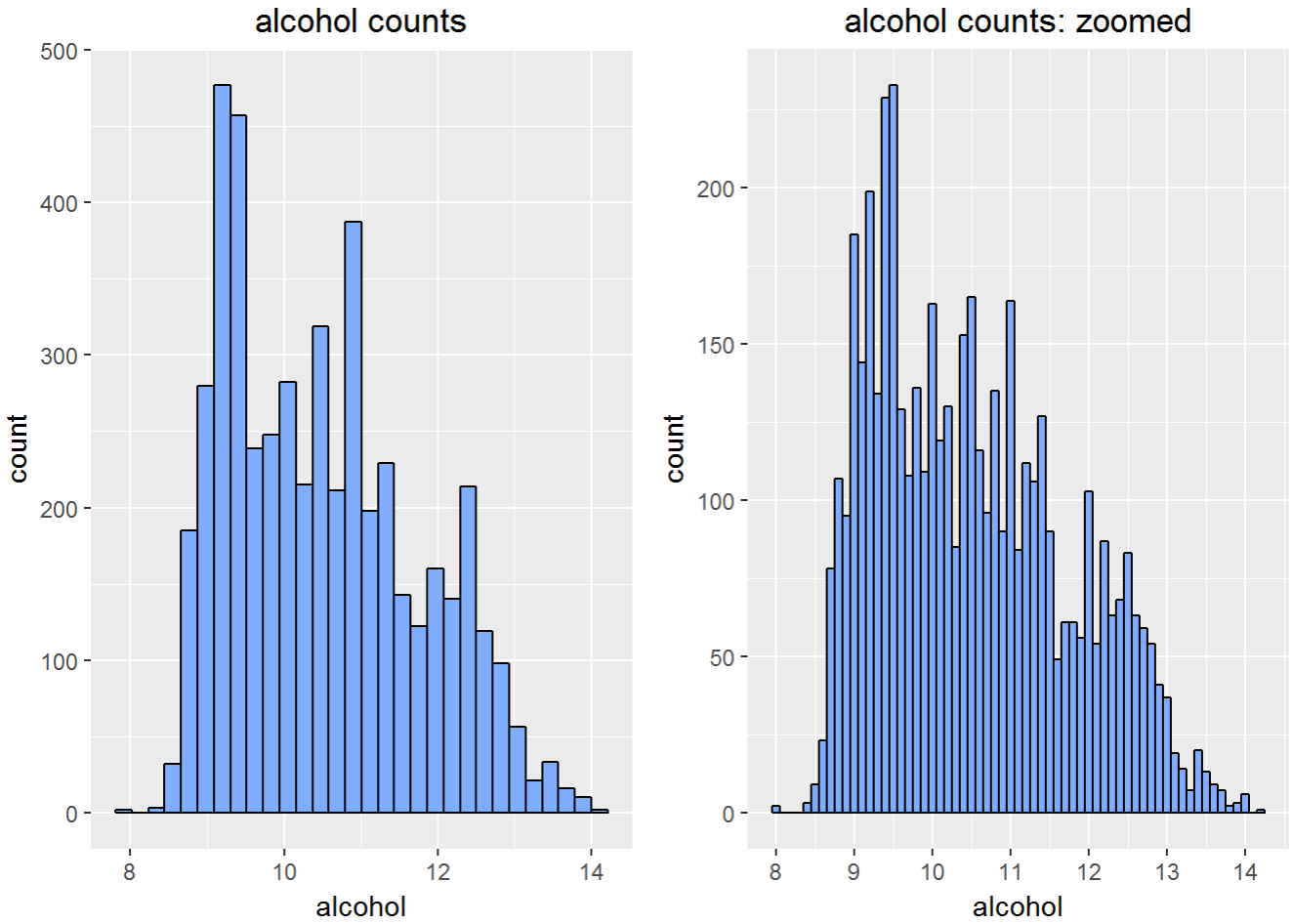


There are two higher frequencies for sulphates. This seems to correlate to the log10-transformed residual.sugar - we can look these two variables together later on.

From the zoomed plot above we can see the 0.5 level has slightly more wines than other levels.

alcohol

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

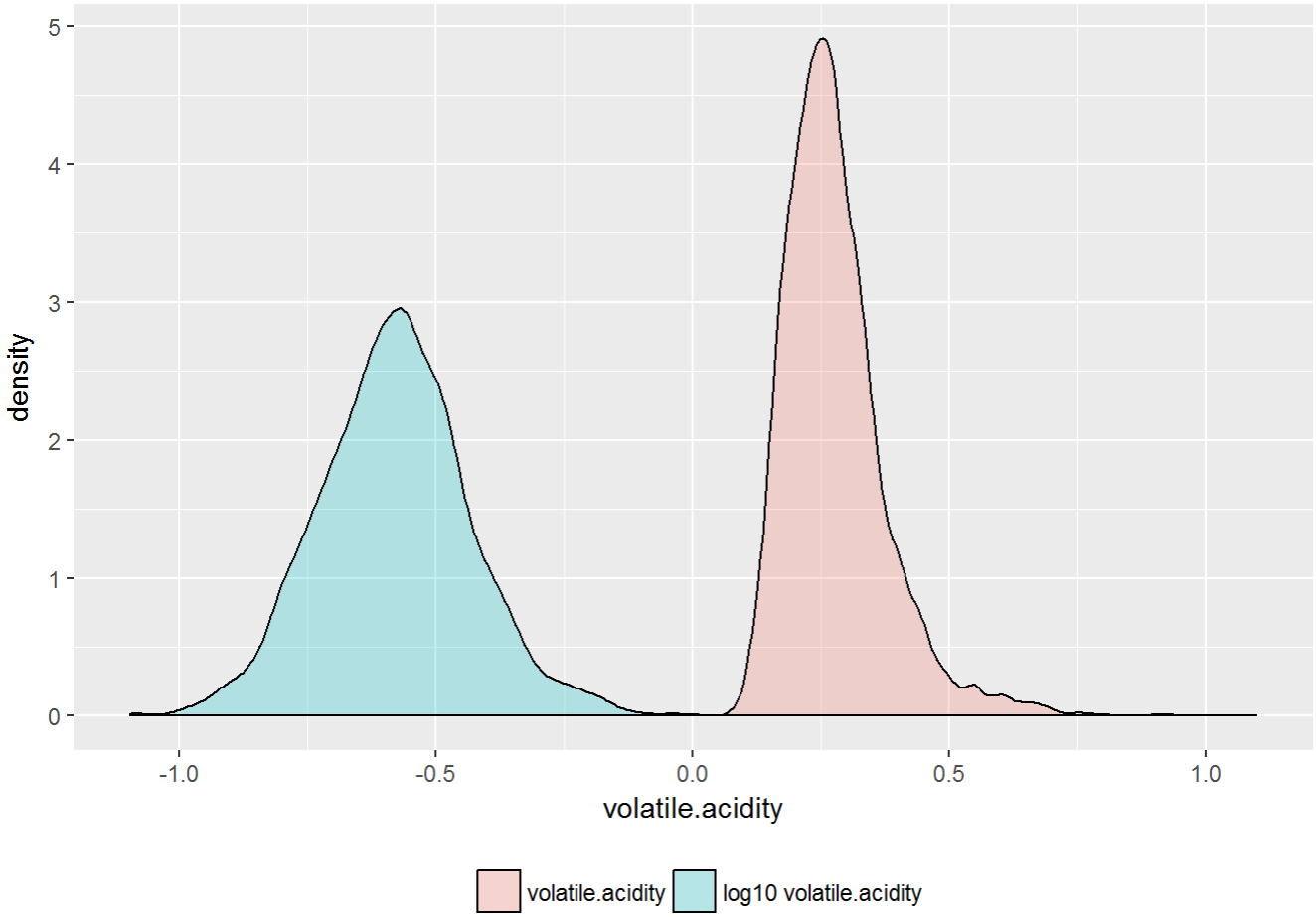


(Left plot) There doesn't seem to be a evident distribution for alcohol at the first glance, so I try adjusting binwidth to see if there's more findings. (Right plot) 9.4 & 9.5 have more wines than other levels.

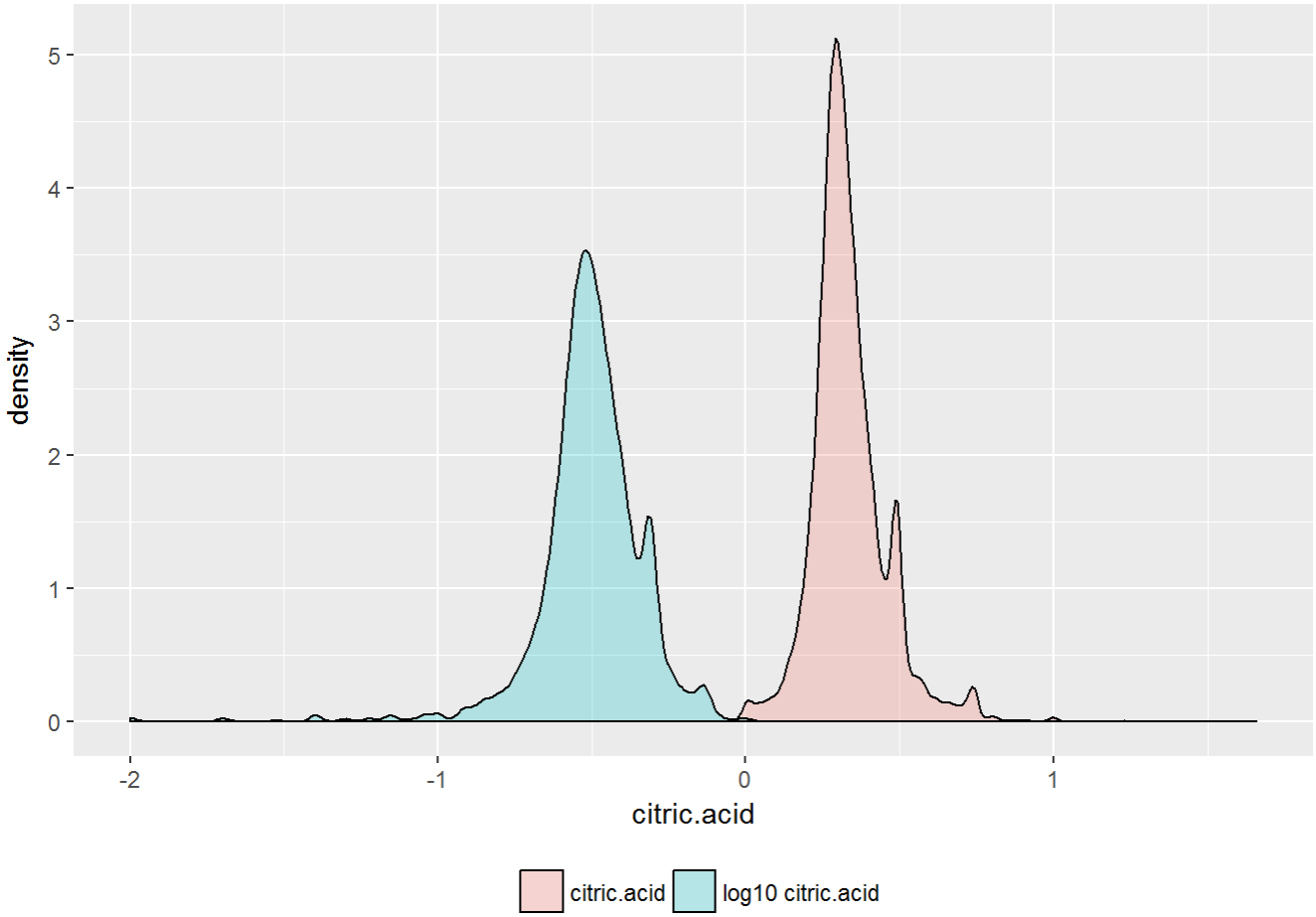
Additional plots for univariables

During the plotting process, I found that a) volatile.acidity b) citric.acid c) residual.sugar d) chlorides e) free.sulfur.dioxide are all skewed to the right, with some outliers at the right of x axis. I am curious of how these variables will look like when they are adjusted by log10, so I plot the below. For variables that cannot be observed clearly in overlaid desity plot, I will create additional desity plots separately to look into.

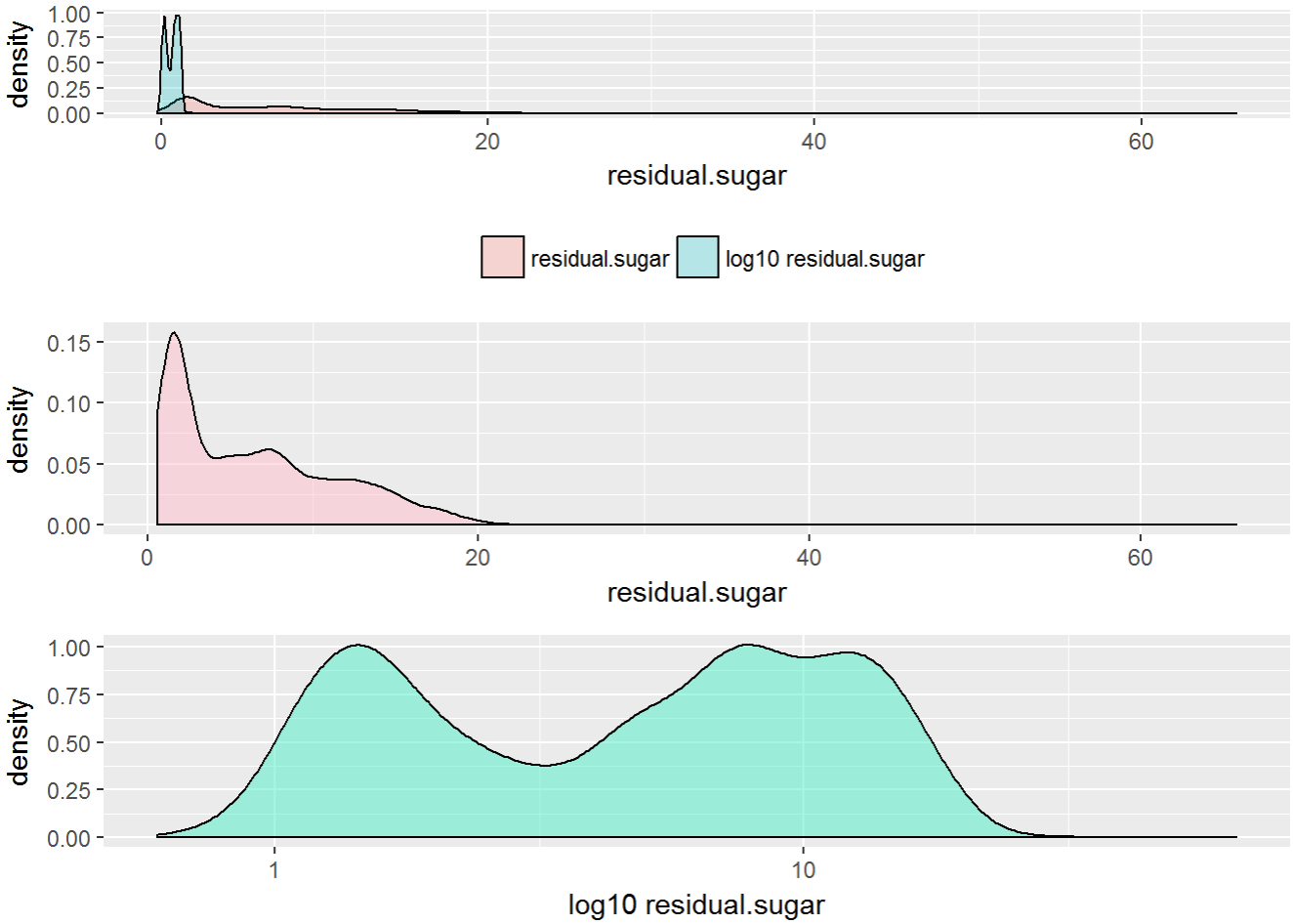
a) volatile.acidity



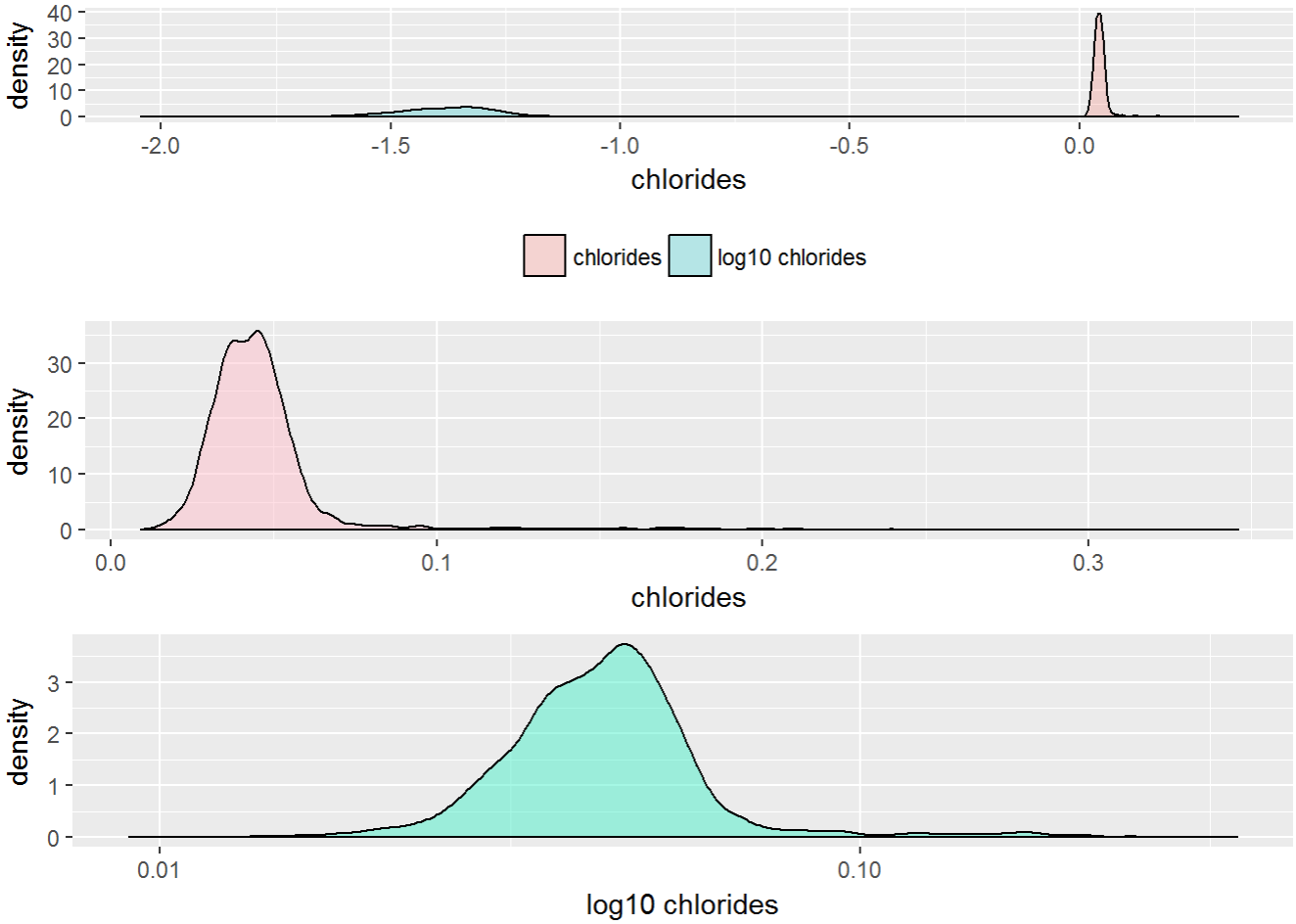
b) citric.acid



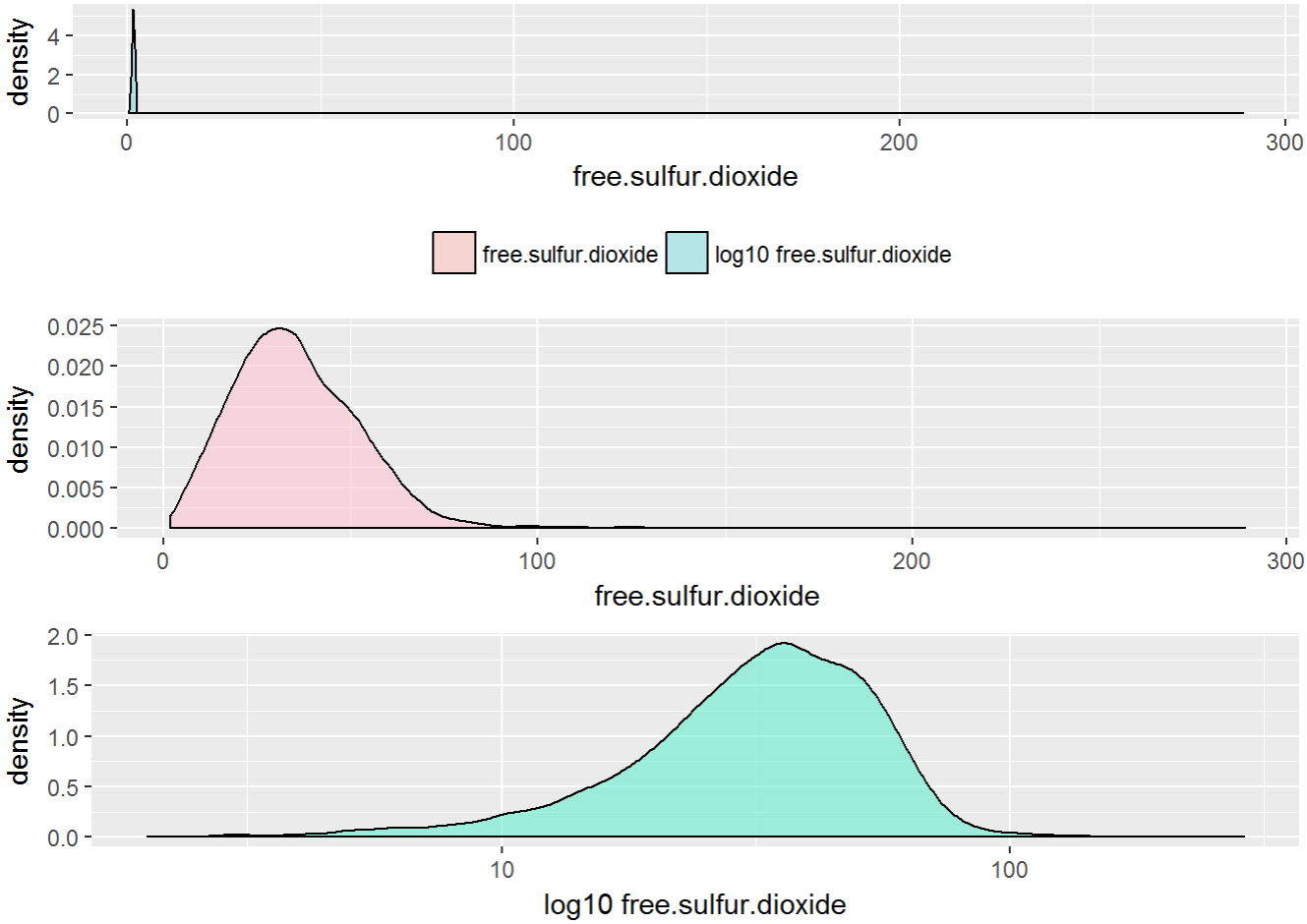
c) residual.sugar


density
1.00
0.75
0.50
0.25
0.00
0
20
40
60
residual.sugar
residual.sugar
log10 residual.sugar
density
0.15
0.10
0.05
0.00
0
20
40
60
residual.sugar
density
1.00
0.75
0.50
0.25
0.00
1
10
log10 residual.sugar

d) chlorides



e) free.sulfur.dioxidec) residual.sugar



Looking at the few plots above, it is interesting to found that the distribution shape for residual.sugar looks quite different after taking log10 - it is transformed from one-peak bell to two-peak shape.

Univariate Analysis

What is the structure of your dataset?

The dataset contains 4898 observation with 11 input and 1 output variables:

Input variables (based on physicochemical tests): 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 / dm3) 11. alcohol (% by volume)

Output variable (based on sensory data): 12. quality (score between 0 and 10)

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

The main feature of interest in this dataset is quality. I am curious about what is the key contributor(s) (what input variables) to a wine's quality score.

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

To my understanding, the eleven variables will support the investigation of white wine quality. However at this stage of

data exploration, there doesn't seem to be an evident clue on which variable have a more reliability with the quality.

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

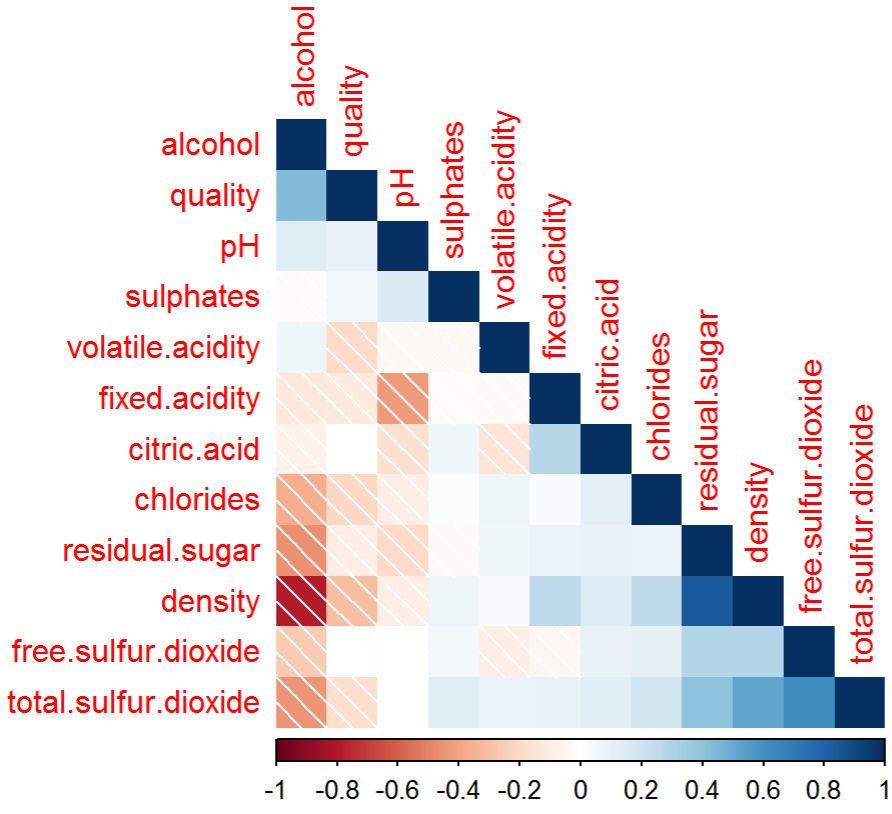
I created quality.level to cluster the main feature, quality, into three groups, so that converting quality field from numeric to factor. This new variable may come handy in the following analsis.

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?

As I commented in captions above, some of the variables are skewed to the right, so I performed log10 on x axis to see if there's any more findings. It turns out that residual.sugar was transformed to have two peaks instead of one peak that is observed in the skewed distribution. For most of variables, I did a zoom in to the peak of distribution by adding a coord_cartesian and adjusting binwidth to closely see if there's any certain level that really have more wines fall into.

Bivariate Plots Section

To see how different variables correlates with each other and whether a variable is a input or output variable, it's a good idea to plot a correlation matrix. We don't need to see correlation of some variables, such as 'X' and 'quality.level', so I omit them from the correlation matrix.



Running codes above we can find the top and bottom pairs of variables that is more related to each others. It seems residual.sugar and sulphates don't correlates as I expected earlier.

density & residual.sugar: 0.83896645

quality & alcohol: 0.435574715

total.sulfur.dioxide & residual.sugar: 0.40143931

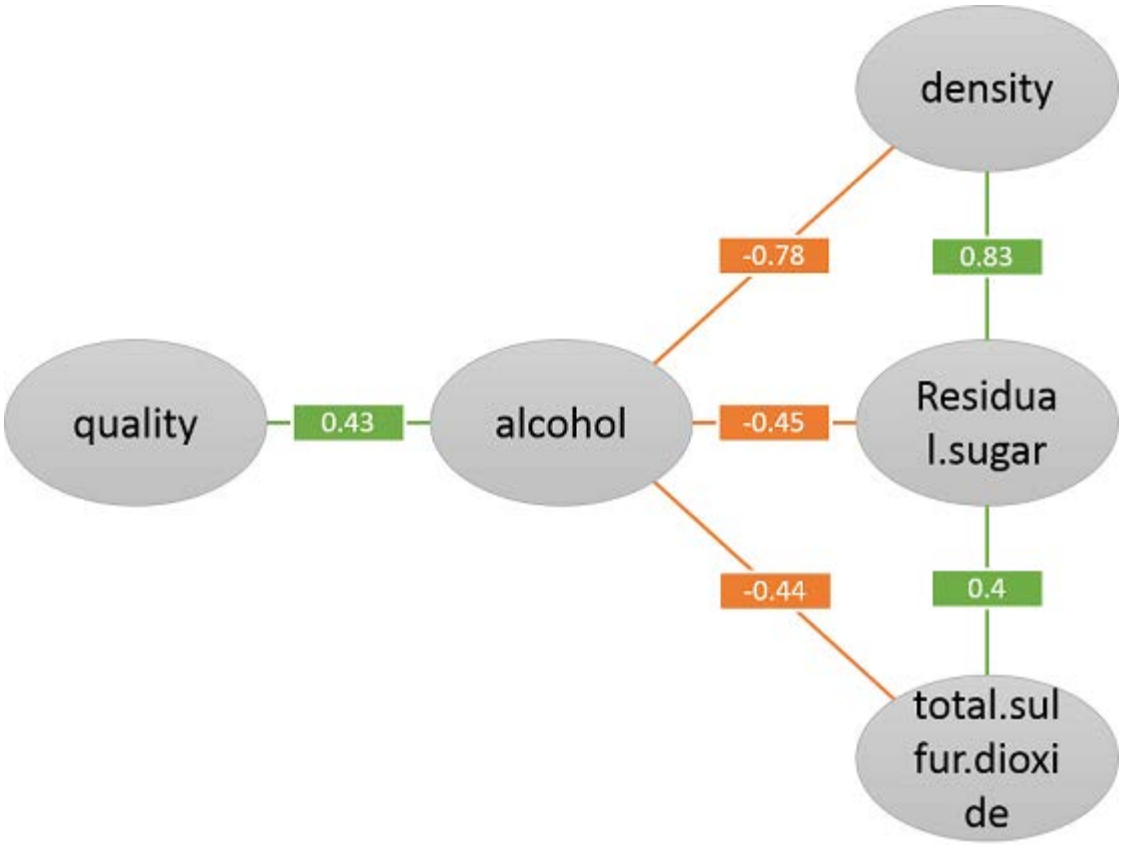
pH & fixed.acidity: -0.425858291

total.sulfur.dioxide & alcohol: -0.4488921

residual.sugar & alcohol: -0.45063122

density & alcohol: -0.78013762

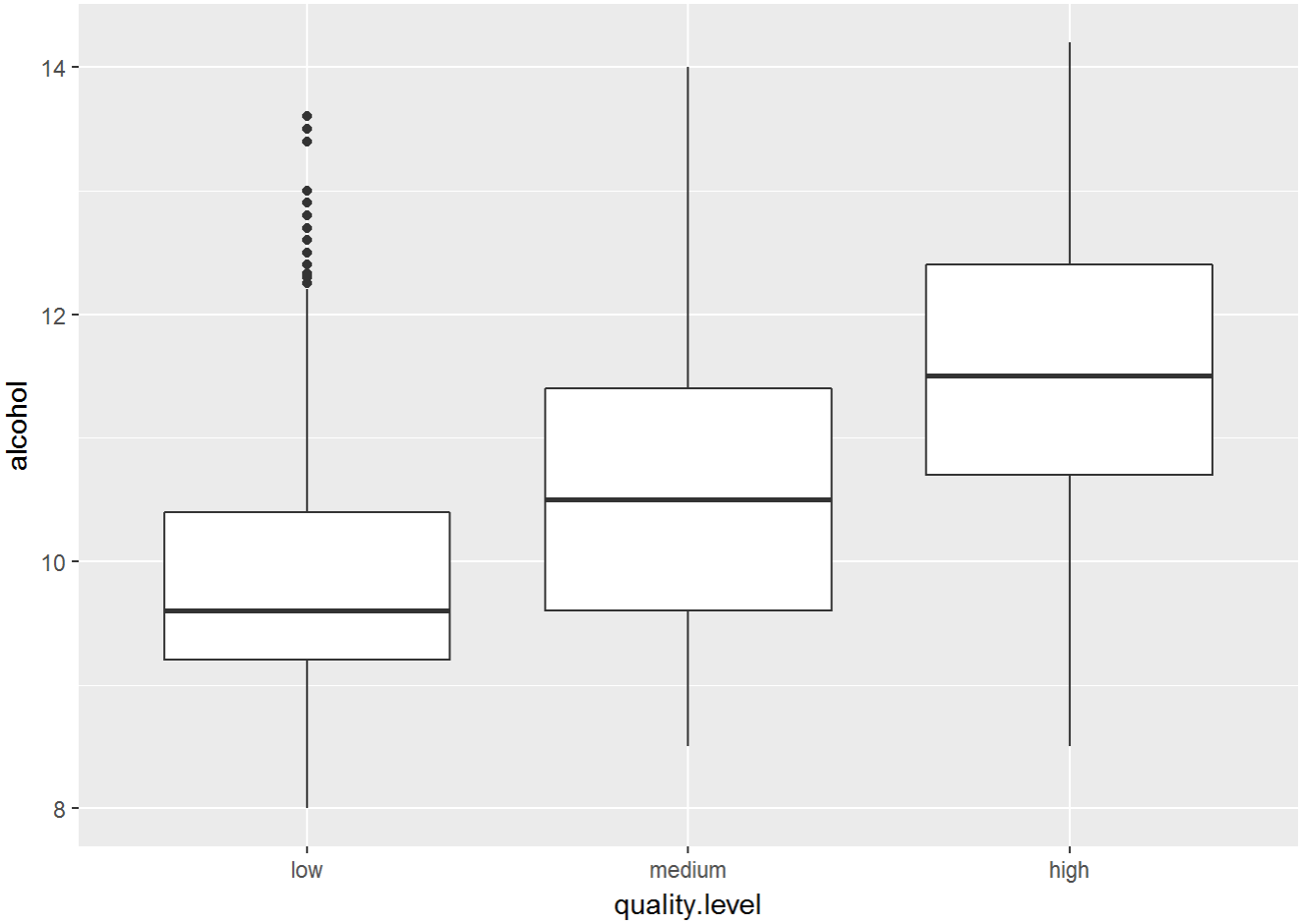
Looking at the correlations, I found that alcohol seems to positively affect the quality of wines, while there are three other variables(density, residual.sugar, total.sulfur.dioxide) that negatively affect alcohol. Interestingly, the three variables seems to positively correlate to each other, as shown in graph below.



Correlation between varibles (part)

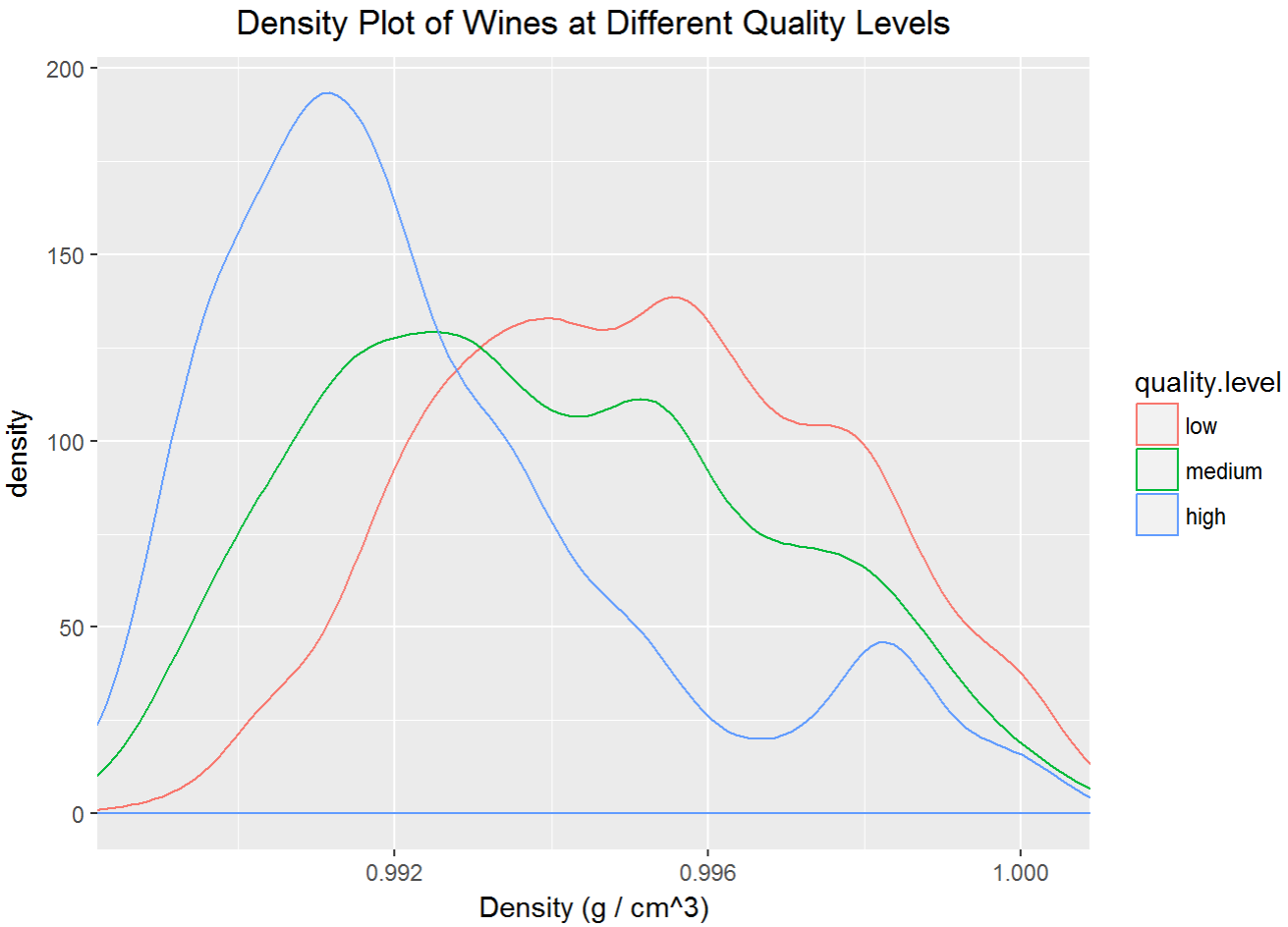
alcohol & quality

As quality is the main feature we want to explore, let's start with plotting relationship between 1) alcohol and quality, and 2) density and quality, as they are the two variables that have relatively higher correlation with quality. This is the time when quality.level comes into use.



From plot above we can see for higher quality wines, the alcohol percentage is generally higher (the box is moving higher the y-axis).

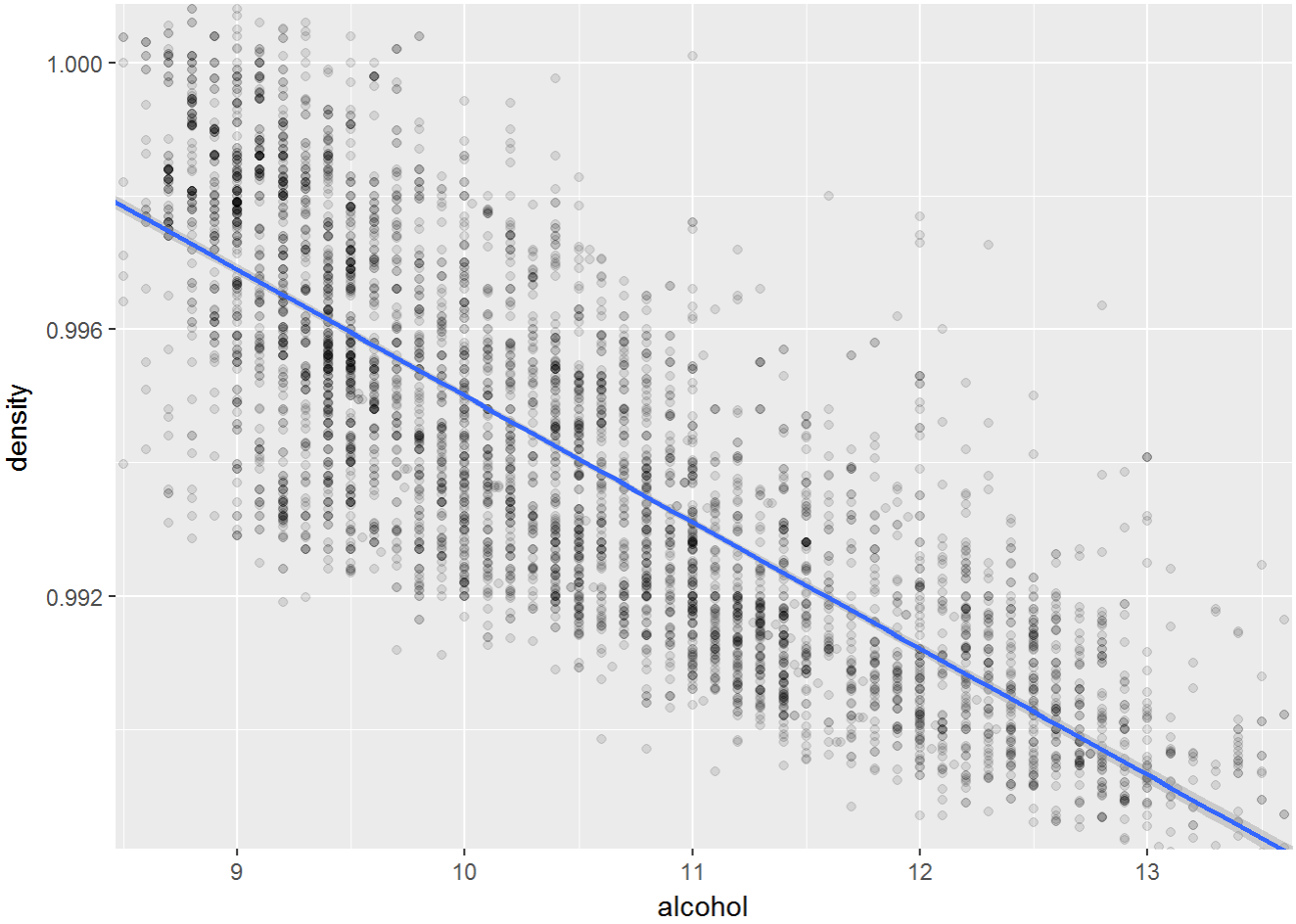
density & quality



Seeing plot above, we can find that for higher quality wines, the density tend to be lower - there seemes to be a negative relationship between density and quality, same as what is observed in correstion matrix.

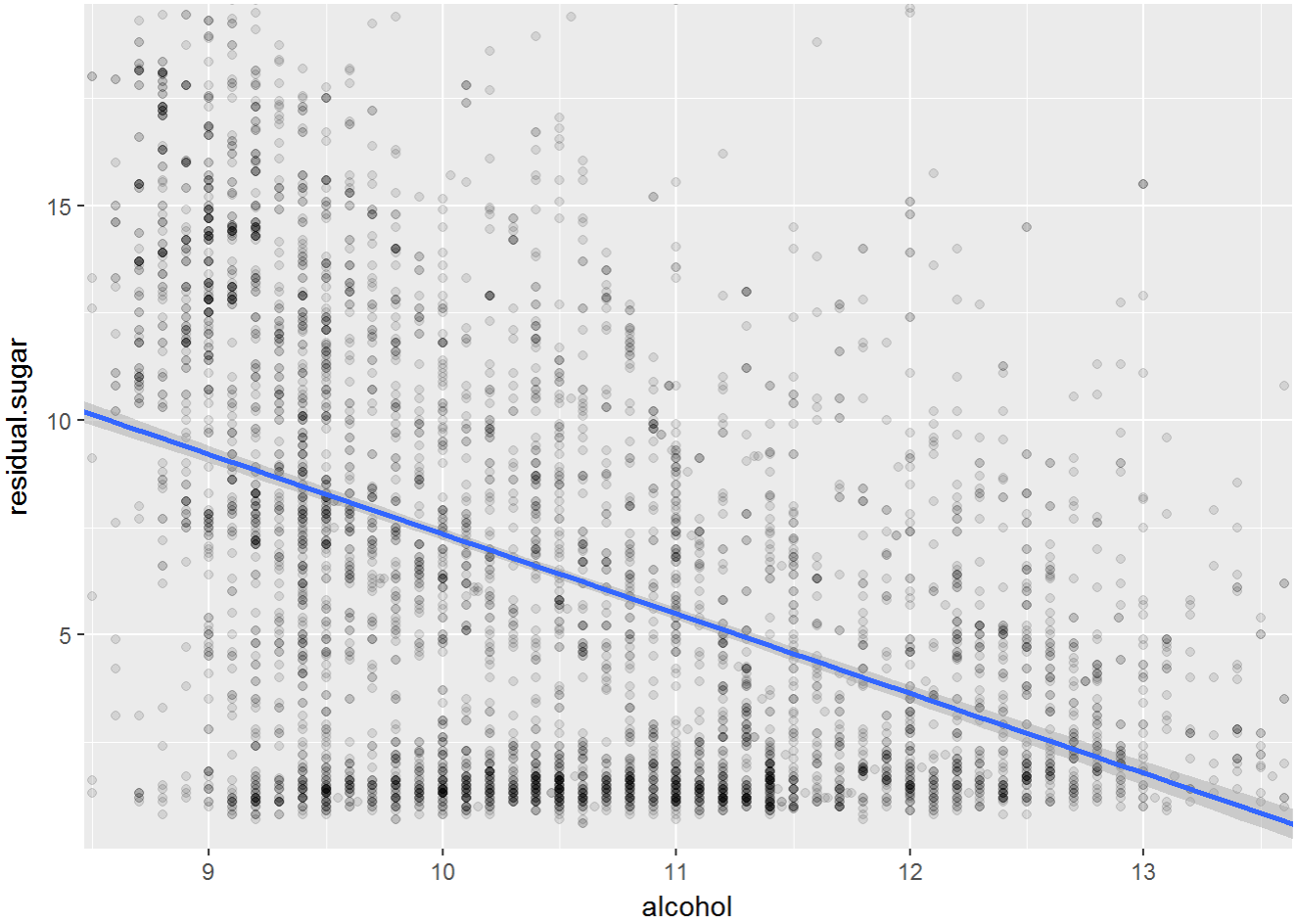
Next, since quality are correlated to alcohol, let's look at the relationships of the three variables that we found negatively-related to alcohol (density, residual.sugar, total.sulfur.dioxide) and alcohol.

alcohol & density



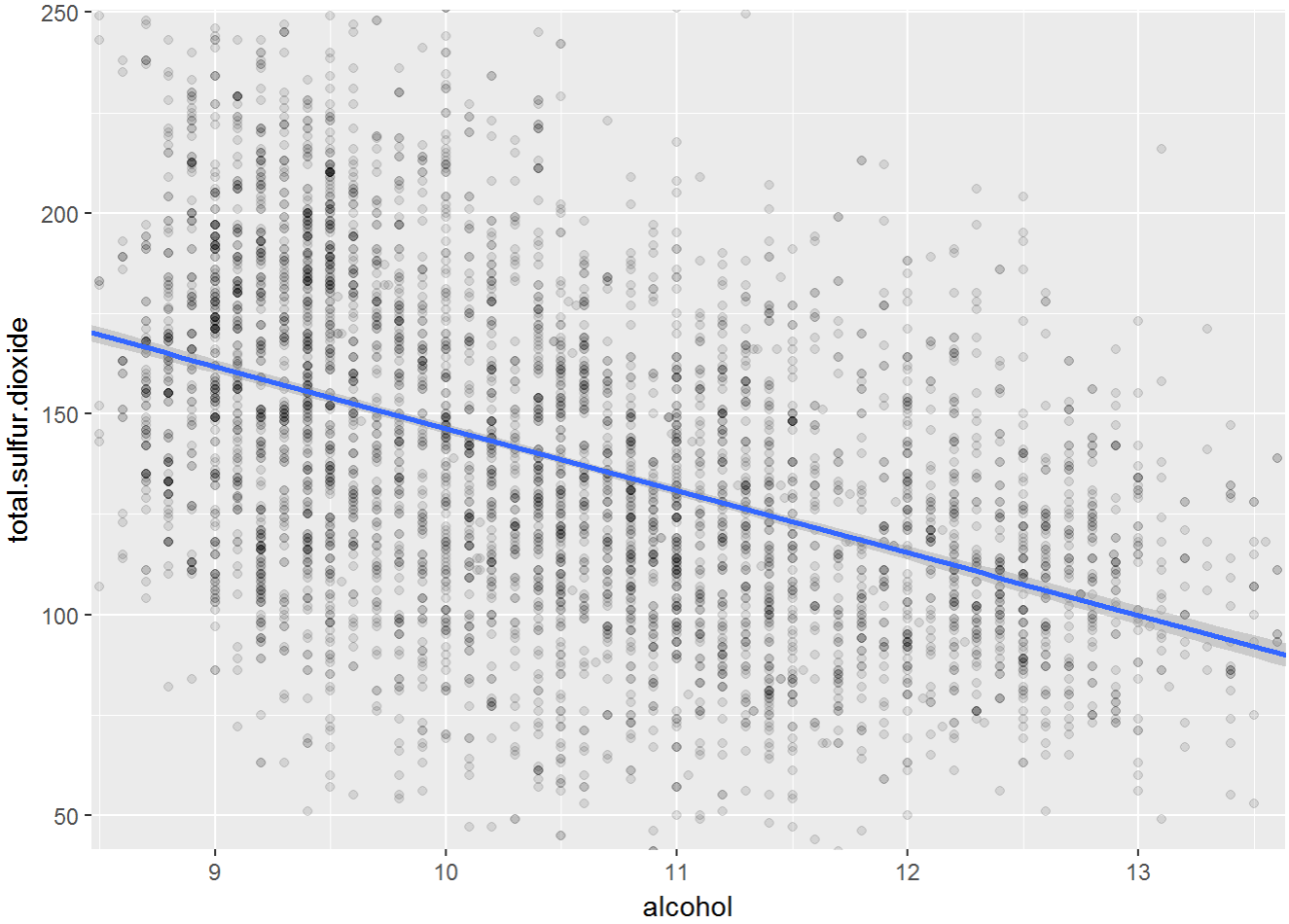
The negative relationship between alcohol and density is quite evident in plot above.

alcohol & residual.sugar



Though residual.sugar and alcohol are negatively-correlated, there are many data points that are with less than 2.5 residual.sugar.

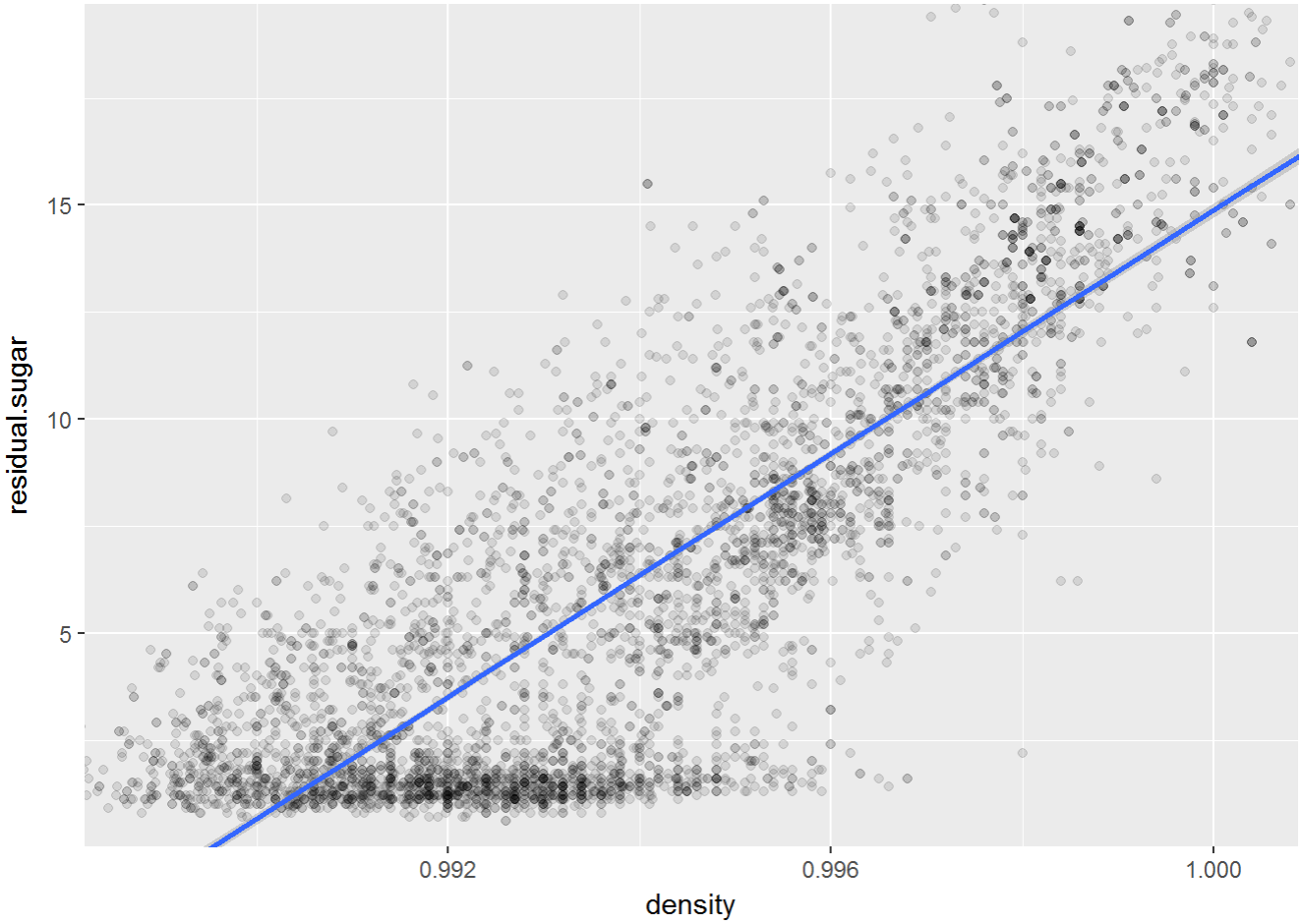
alcohol & total.sulfur.dioxide



The plot above proves the negative relationship between alcohol and total.sulfur.dioxide.

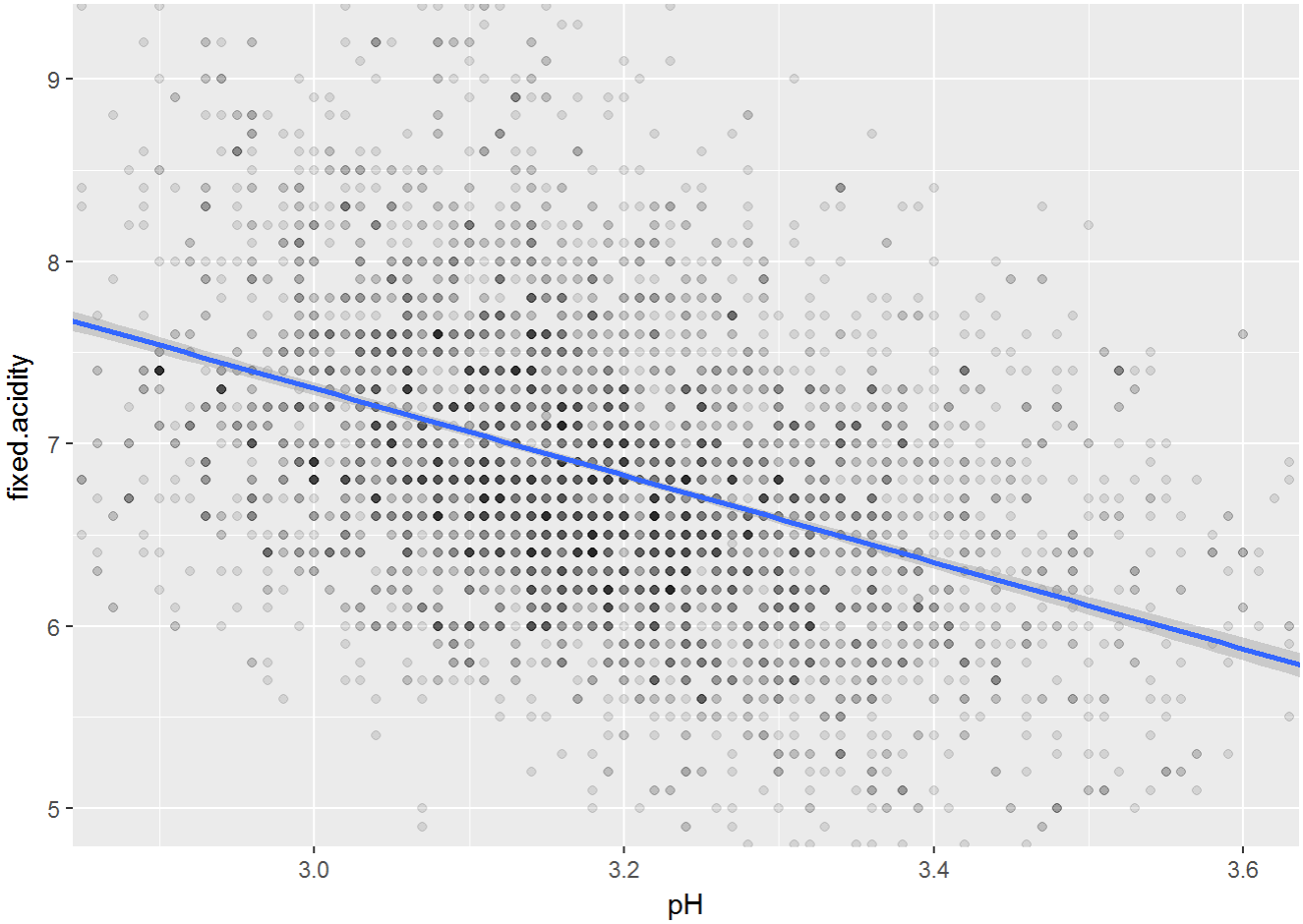
density & residual.sugar

From correlation matrix plotted earlier, we see that density and residual.sugar have close, positive relationship with each other. I am curious how that relationship will look like on a scatterplot.



pH & fixed.acidity

pH & fixed.acidity are also in the list of pairs that are closely-related variables, let's plot the two variables on a scatterplot.



The distribution of points looks funny; it's because fixed.acidity is not continuous in the dataset.

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?

The feature of interest (quality) only evidently related to alcohol from the first glance of correlation matrix. I also found that the relationships among alcohol and density, residual.sugar, total.sulfur.dioxide are interesting - they seem to be connected to each other in some way.

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

Other than the relationships of alcohol, density, residual.sugar, and total.sulfur.dioxide, pH & fixed.acidity also have negative relationships.

What was the strongest relationship you found?

The strongest relationship I found among these variables is density and residual.sugar, they have a 0.83896645 correlation. Density also have strong negative relationship with alcohol, the correlation is -0.78013762.

Multivariate Plots Section

From investigation in sections before, I found that: a) Alcohol and density are negatively related b) Density and residual.sugar are negatively related and c) Quality and alcohol are positively related

In this section I would like to blend quality into findings a) and b) above to see if there's any new, complex findings.

alcohol & density, residual.sugar, total.sulfur.dioxide

With the findings in previous section, I want to see how the three variables affect alcohol - a 3D scatter plot might help.

The plot above proves that alcohol has a negative relationship with density, residual.sugar, and total.sulfur.dioxide: the lighter points means wines with higher alcohol, and they are concentrated to corner where the three other variables are lower.

I am hence curious of how these variables interact with each other - maybe it's a good idea to build a model to see how alcohol, density, residual.sugar and total.sulfur.dioxide predicts quality.

building model with alcohol, density, residual.sugar, total.sulfur.dioxide

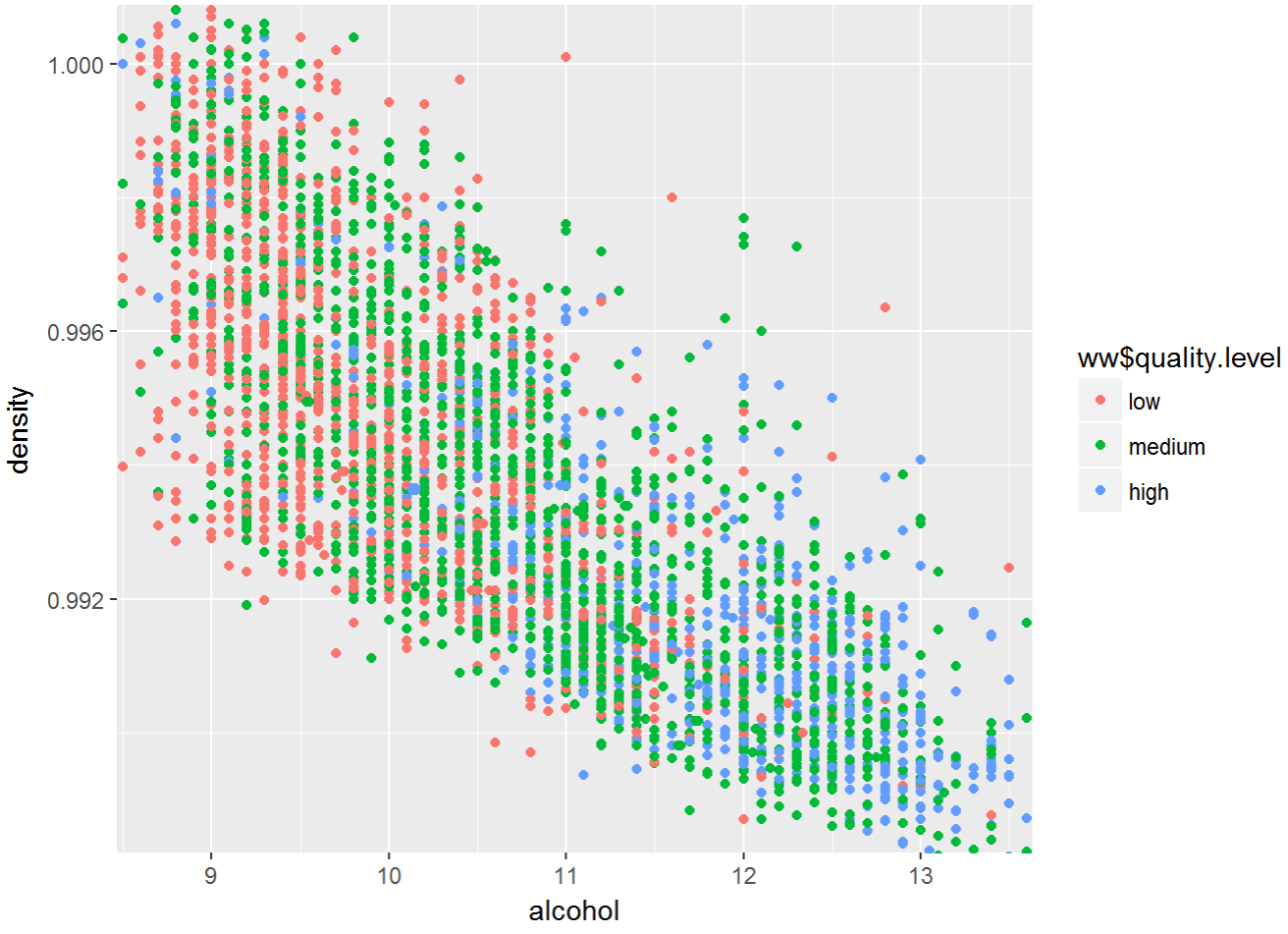
```
##
## Call:
## lm(formula = quality ~ alcohol + density + residual.sugar + total.sulfur.dioxide,
```

```
##     data = ww)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4795 -0.5377 -0.0107  0.4720  3.2011
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.366e+01  1.266e+01   7.395 1.65e-13 ***
## alcohol               2.462e-01  1.825e-02  13.491  < 2e-16 ***
## density              -9.131e+01  1.262e+01  -7.234 5.41e-13 ***
## residual.sugar        5.375e-02  5.101e-03  10.536  < 2e-16 ***
## total.sulfur.dioxide  3.888e-04  3.136e-04   1.240    0.215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7873 on 4893 degrees of freedom
## Multiple R-squared:  0.2104, Adjusted R-squared:  0.2098
## F-statistic:   326 on 4 and 4893 DF,  p-value: < 2.2e-16
```

Judgingfrom the low r-square (less than 0.22), I would not see this model an appropriate one to perdict quality.

relationship among quality, density and alcohol

As the route of building models of the four variables to predict quality doesn't seem to work, let's turn our eyes to look at how quality, density and alcohol interacts with each other.



It is discovered that the higher quality wines are centered in the high-alcohol, low-density corner of the graph.

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?

In this part of investigation I created a 3d-scatterplot to include in four variables to validate the observation that alcohol is negatively correlated to density, residual.sugar and total.sulfur.dioxide; the plot proves the relation- ship to be true. The density-alcohol scatterplot colored with quality level strengthened each other on the negative relationship with quality.

Were there any interesting or surprising interactions between features?

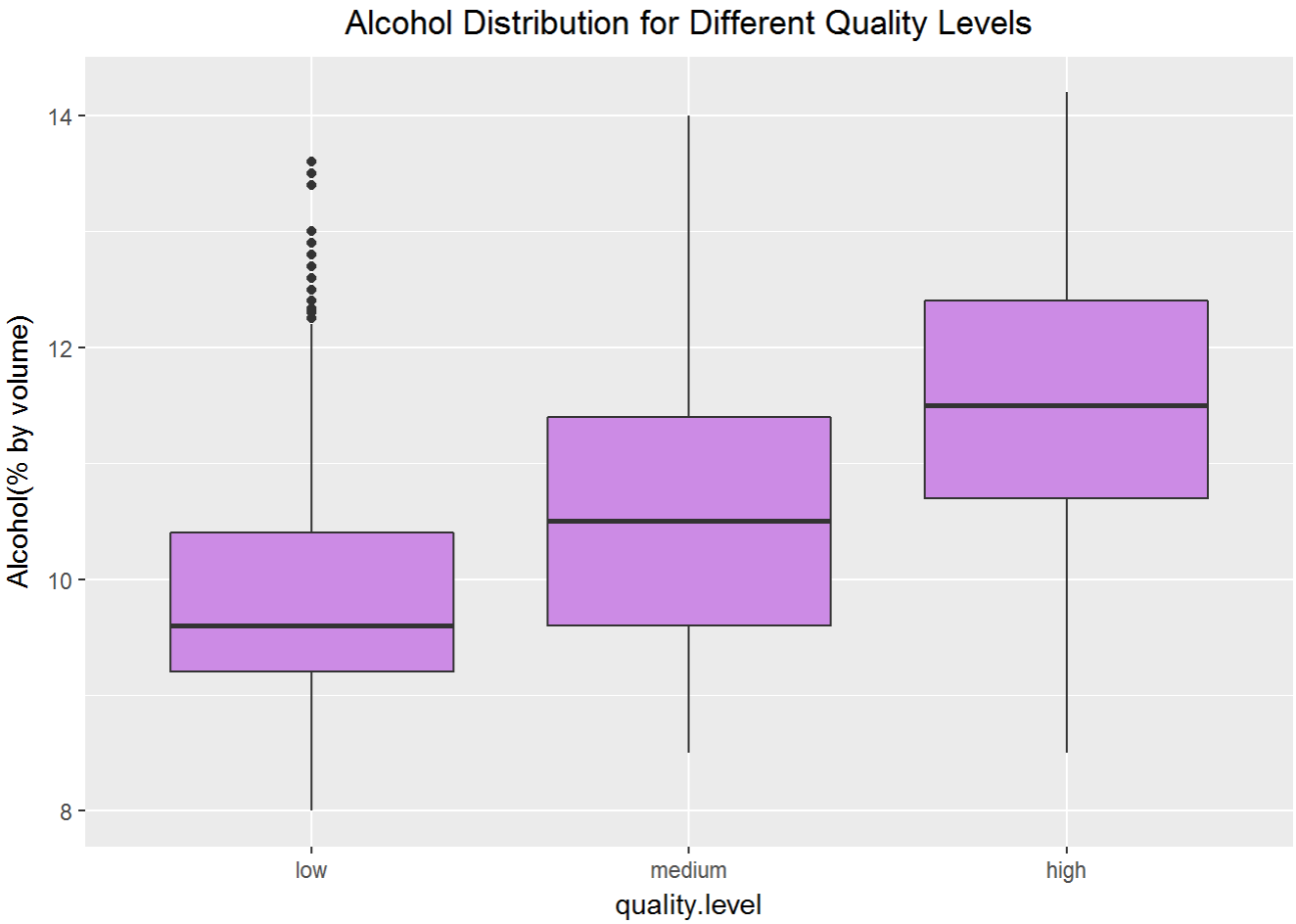
I am a bit surprised to find that the model build does not fit my initial assumption that plugging in some variables that are correlated to each other would output a not-bad prediction.

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

I created a model with alcohol, density, residual.sugar and total.sulfur.dioxide as input variable to predict wine quality. The result is not satisfactory as r-square of the model is less than 0.22. In my perspective the limitation probably comes from the low correlation of these variables with quality.

Final Plots and Summary

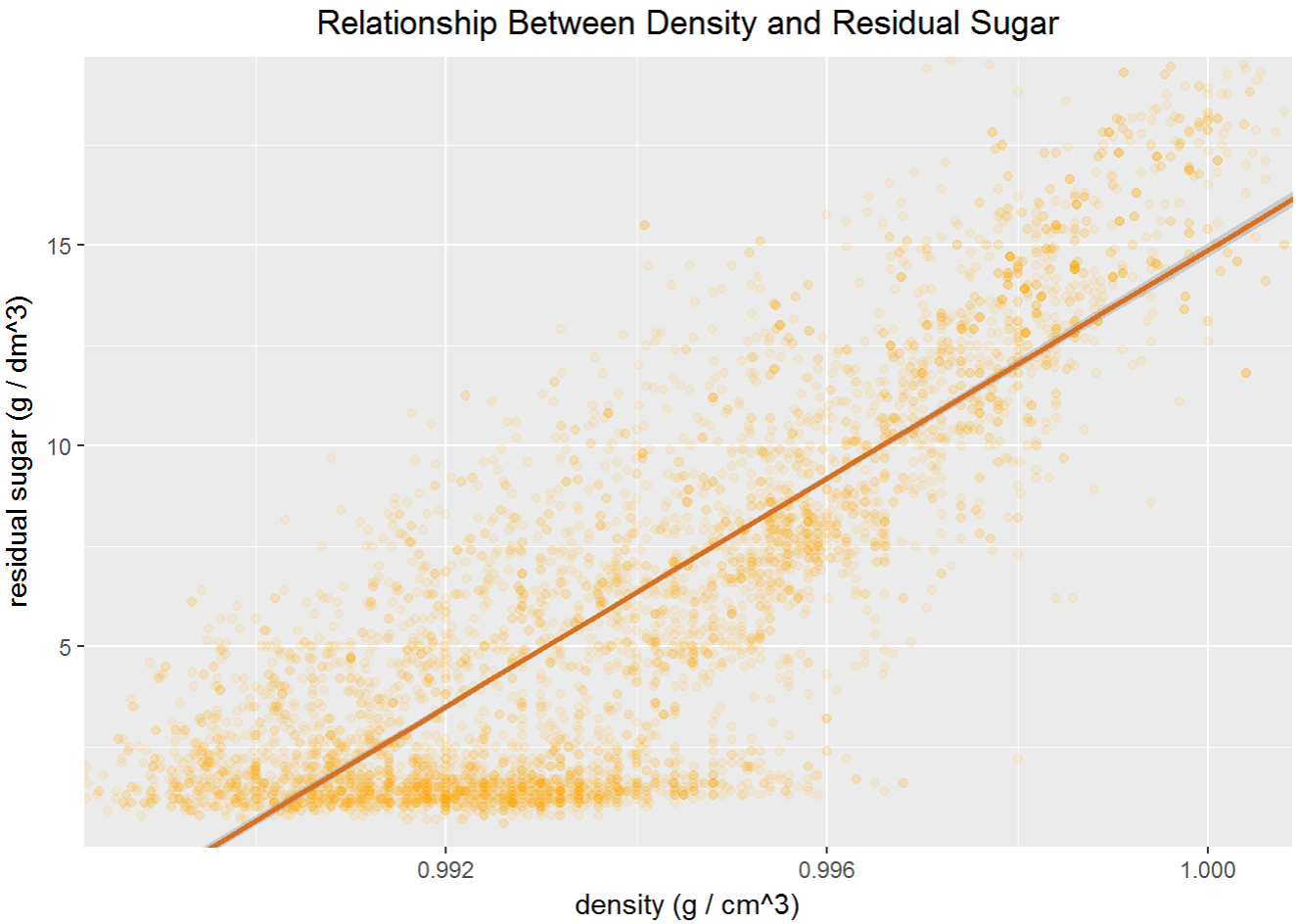
Plot One



Description One

Grouping quality into three levels (low, medium, high) and display their alcohol level in boxplots respectively, we can see that higher quality wines tend to have higer alcohol.

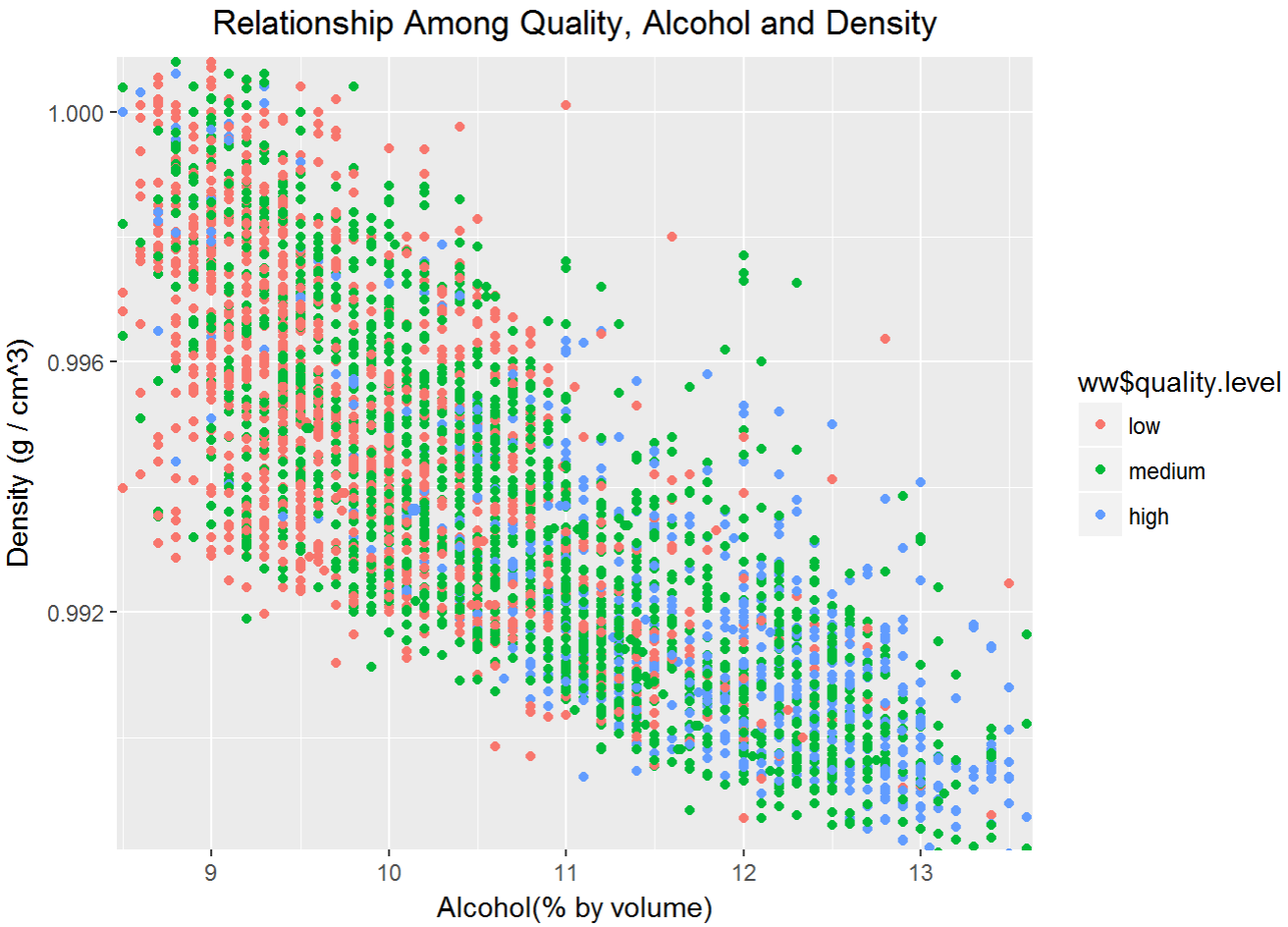
Plot Two



Description Two

Though the main feature, quality is not included in this plot, but I found it interesting that density actually has a high correlation with residual sugar in white wines - this could be observed in the relatively steep slope-smoother and the datapoints distribution in plot above.

Plot Three



Description Three

Apparently there's more factors affecting quality than just alcohol. Since density is the second evident single variable that correlates to quality, I've maken this plot to see the relationship among these three variables. From the plot we can see higher quality white wines generally have lower density and higher alcohol.

Reflection

During the process of analyzing this dataset, I found myself struggling with plotting numeric variables: the plots looks funny and I could not find any insights from these plots. I spent lots of time trying differnt plot types and force myself to come up with thoughts interpreting these plots but in vain. After several hours of struggle I try referring to how others process datasets and found that creating new variables could turn numeric inputs into factors - this helped me a lot and I can progress further by clustering quality into three levels. I believe what I did right is reach out for reference.

Another lesson learned is that I shouldn't have been stubborn looking at the relationships of alcohol, density, residual.sugar and total.sulfur.dioxide - it took much more time than expected to find they are not really contributing much to the feature variable.

To make the analysis better, in the future I would consider obeserving how some other variables distribute differently among the three quality levels.