GIVE ME MORE WINE

by NISHCHAY CHAWLA

==========

Introduction

The dataset is related to white variant of the Portuguese “Vinho Verde” wine. The dataset points out few of many factors which inflence world’s all time favourite beverage (WINE!!).Type of grapes, yeast strain used, age, temperature, aging process, are generally the terms understood by enthusiasts. Although the dataset include variables based on physicochemical tests only, it will be very intersting to draw patterns in the technicalities of wines judged by the experts.

Citation

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 (http://dx.doi.org/10.1016/j.dss.2009.05.016) [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf (http://www3.dsi.uminho.pt/pcortez/winequality09.pdf) [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib (http://www3.dsi.uminho.pt/pcortez/dss09.bib)

Univariate Plots Section

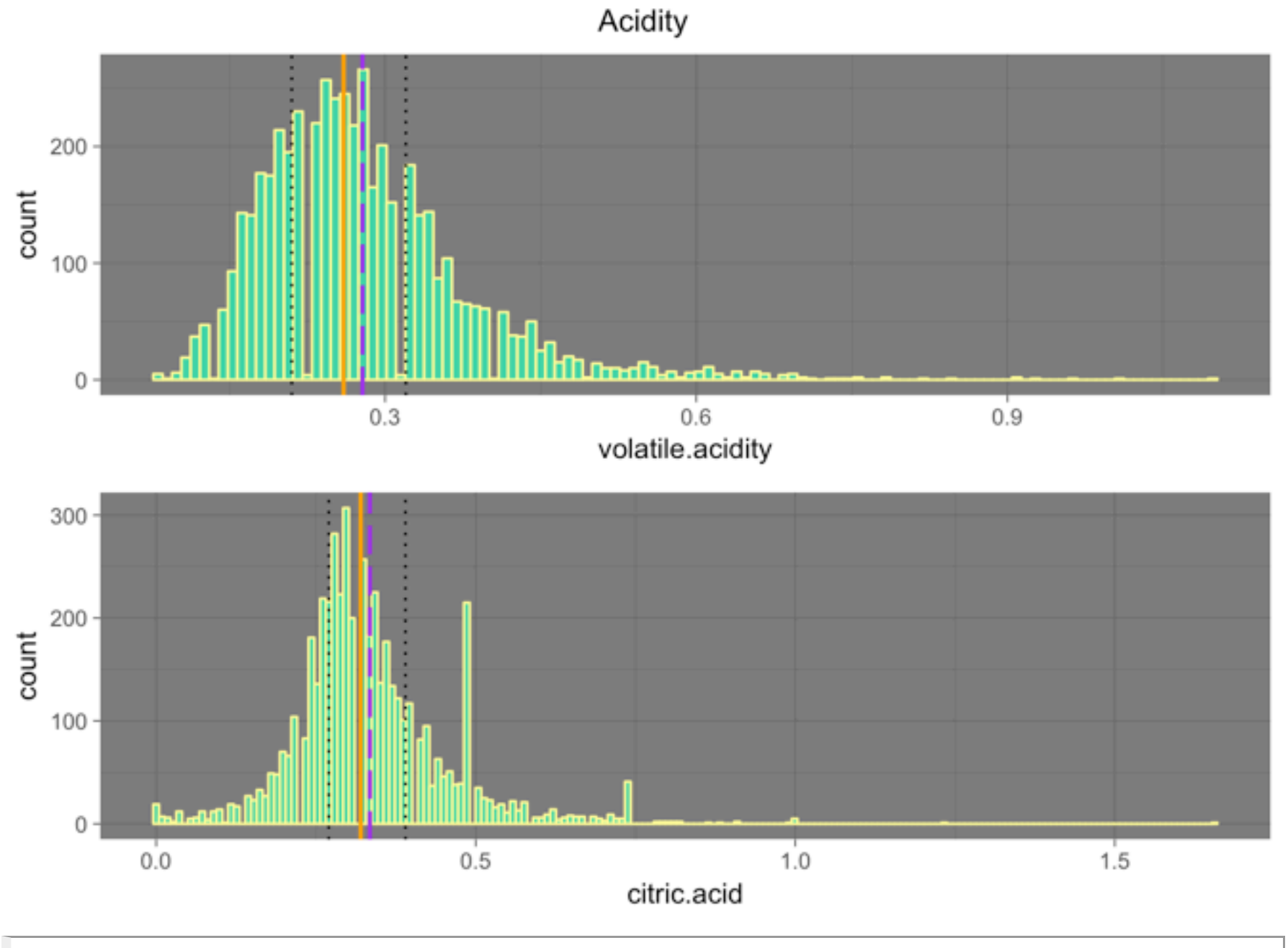
Structure Of Data:

```
## '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 0.0
49 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             : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4 4 4 4
4 4 4 4 ...
```

The orignal dataset has quality as numeric variable but it is converted to an ordered variable for the sake of exploration.

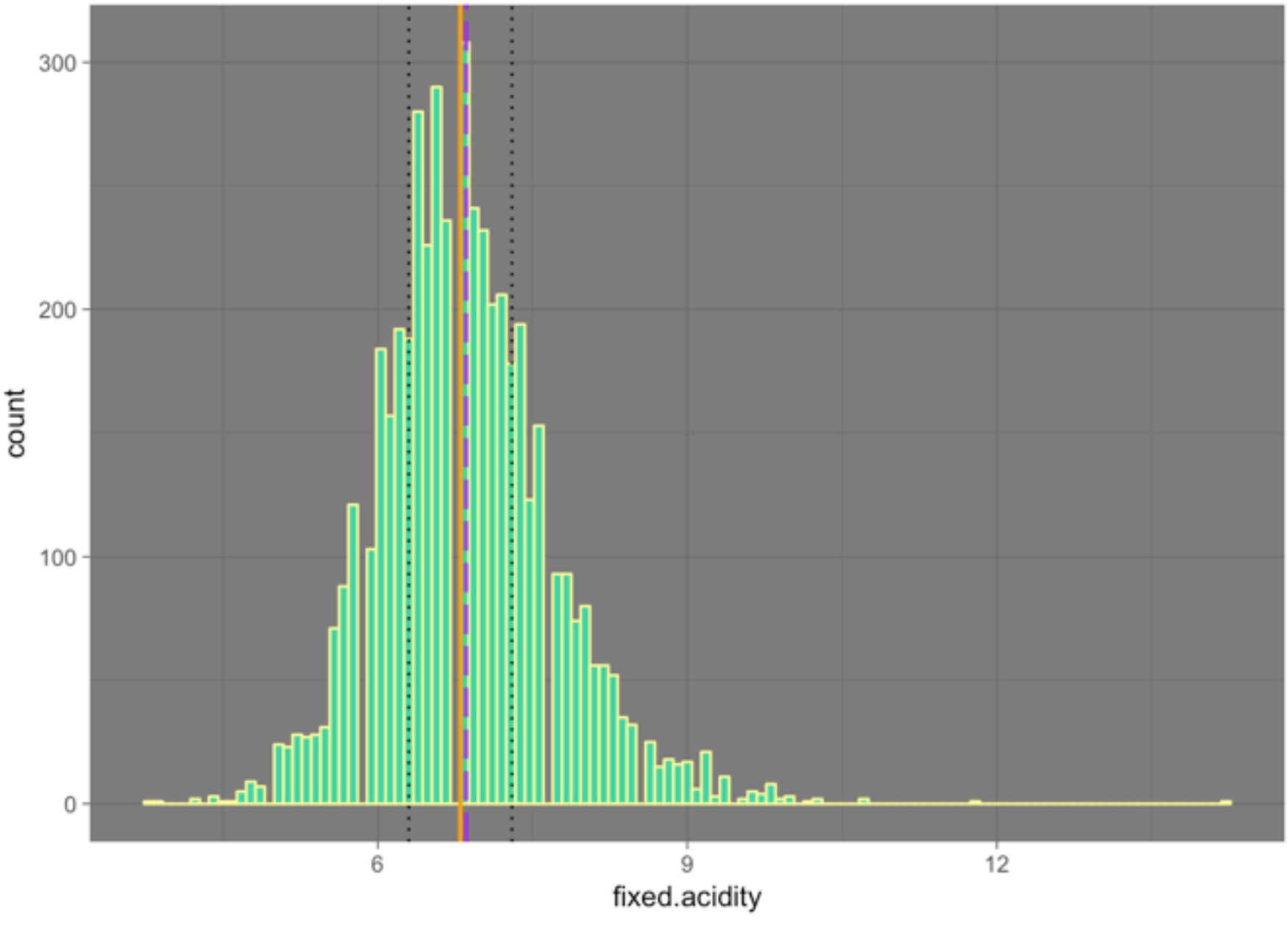
Summary Of Data

```
##  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
```

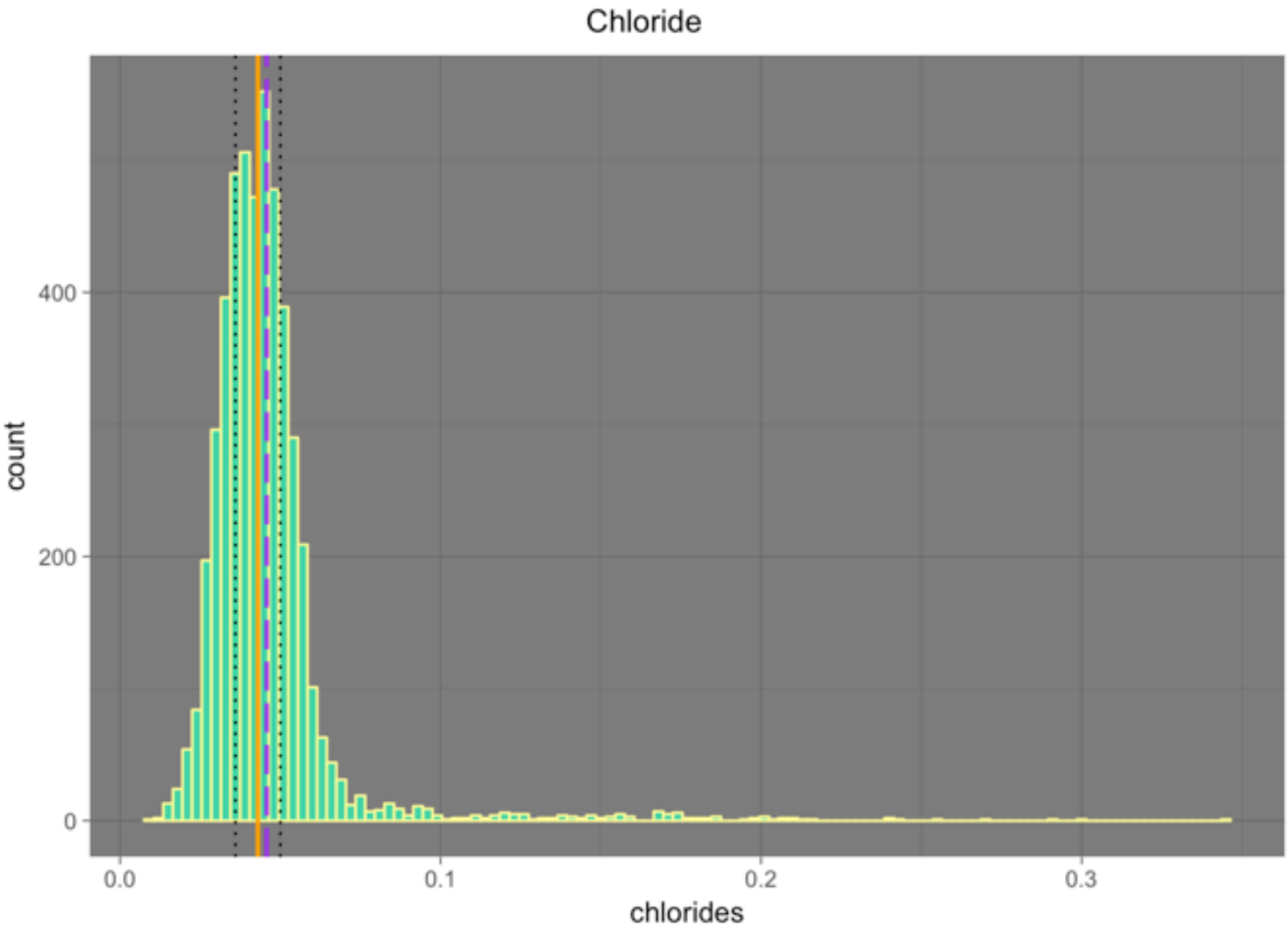


Purple-Dashed line indicates the mean, orange indicates the median, while the dotted lines are for quartiles. 1 dm^3 = 1 Liter

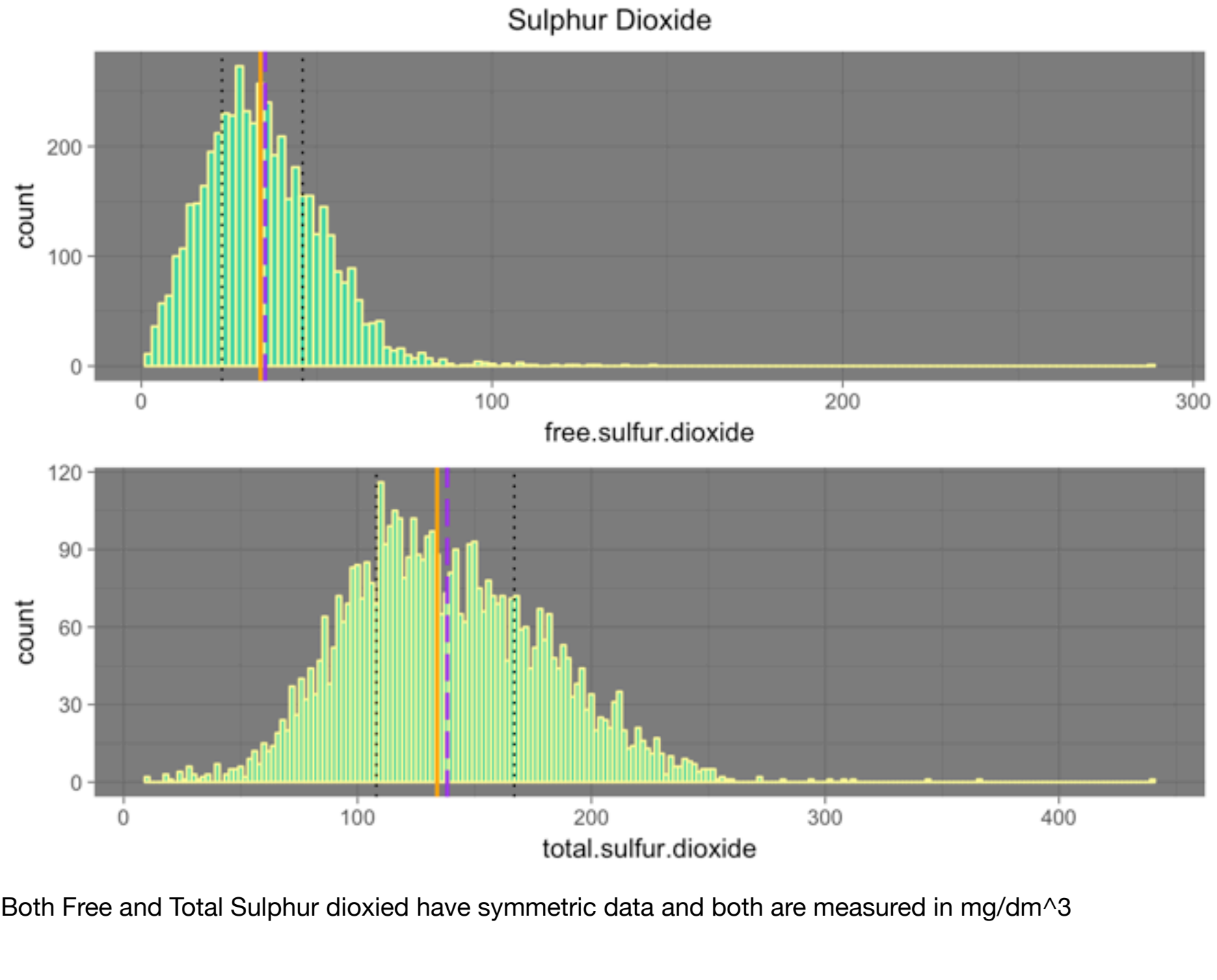
Both Citric and Acetic acid(indicated by volatile acid) are measured in gm/dm^3 and have an almost normal distribution. Citric acid have an unusual peak at 0.49 gm/dm^3, this may be because of persence of many wines from one paticular winemaker or because of any regulations.



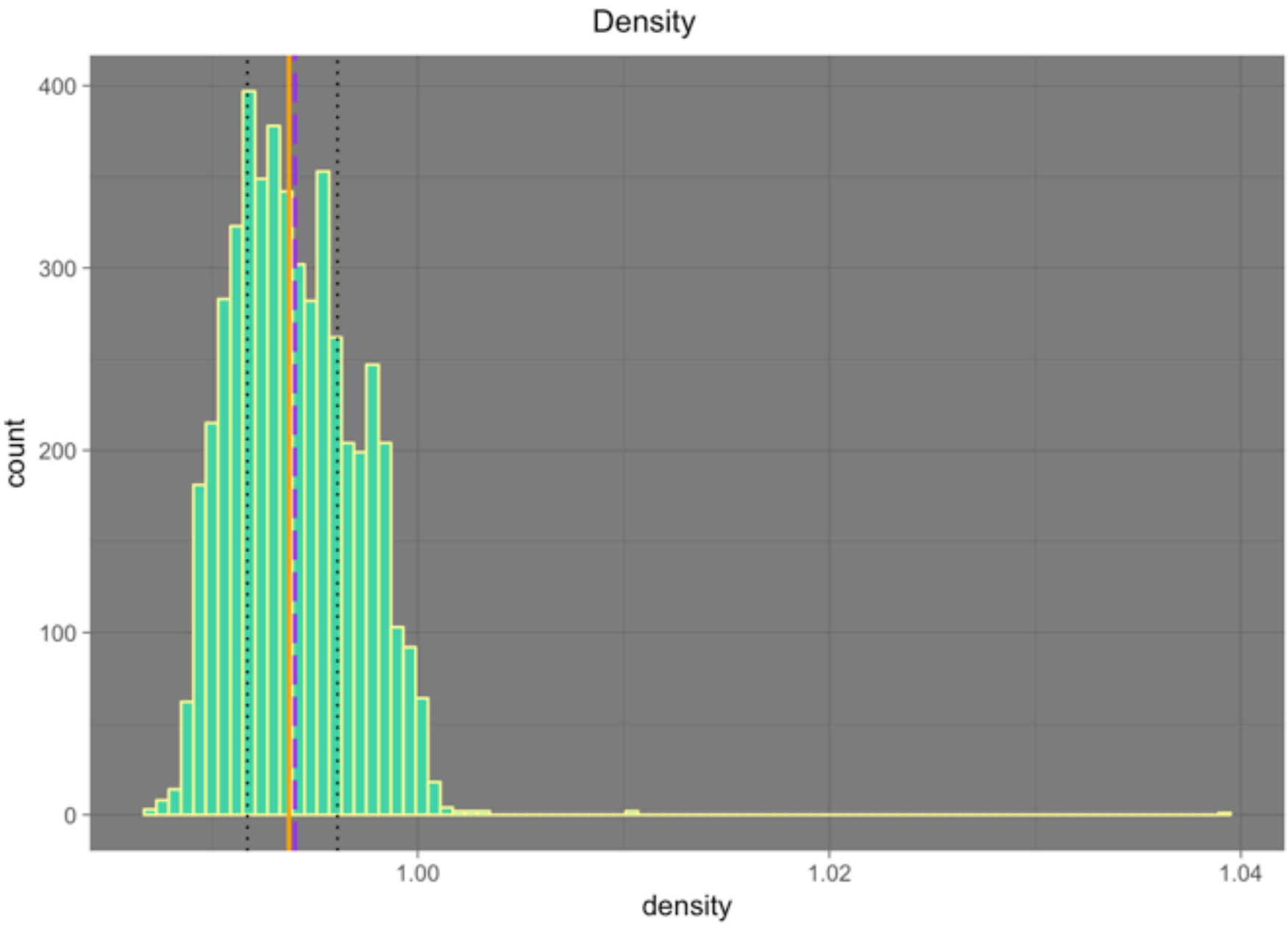
Fixed acidity indicates tartaric acid measured in gm/dm^3 is also almost normally distributed.



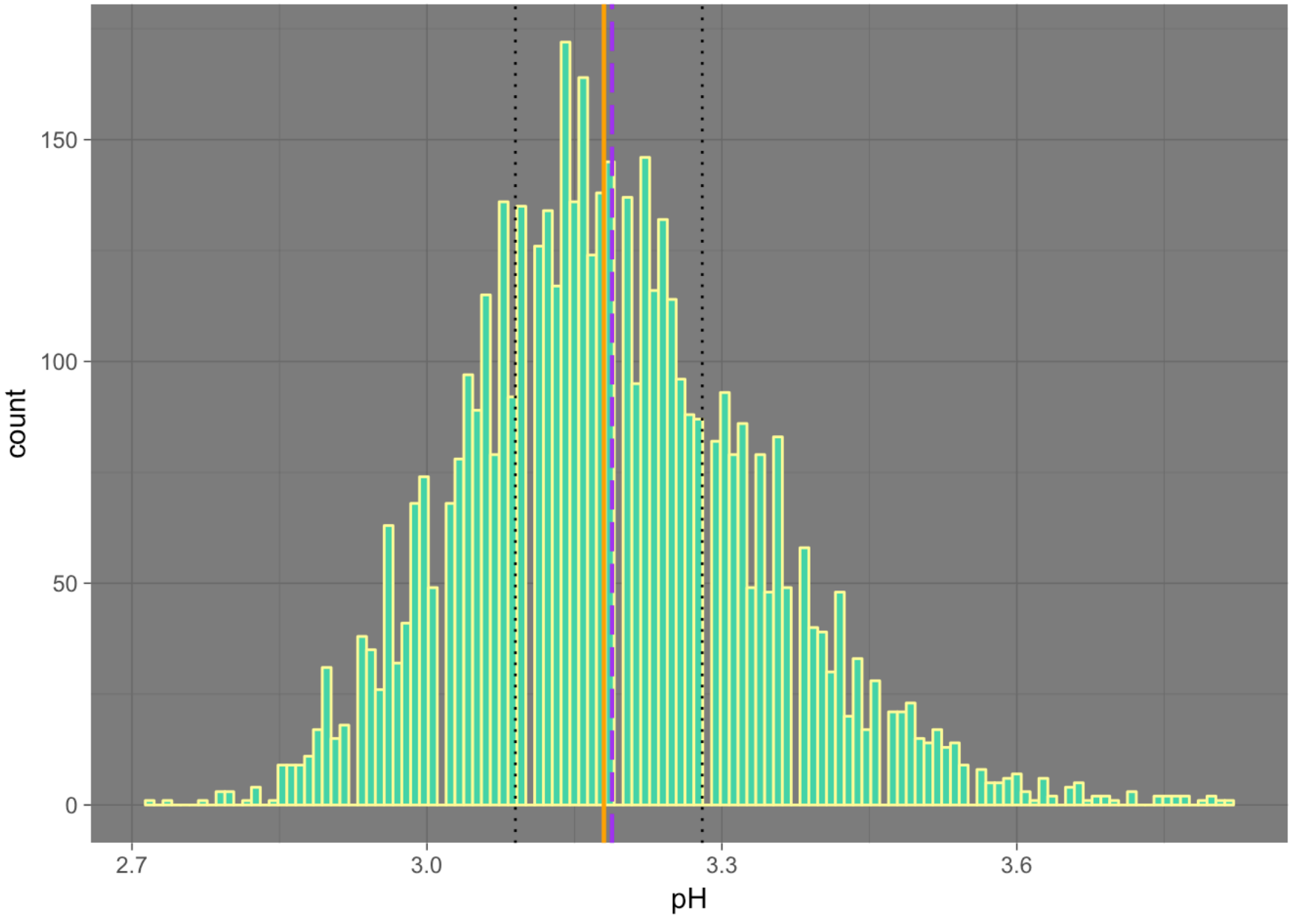
The chloride data looks very lepokurtic as most data lies in 0.1 range. Infact 3rd Quartile at 0.05 indicates 75% data lies between 0.009 an 0.05 (0.009 being min. value). Although data goes upto .34, very very few value lies above .1.



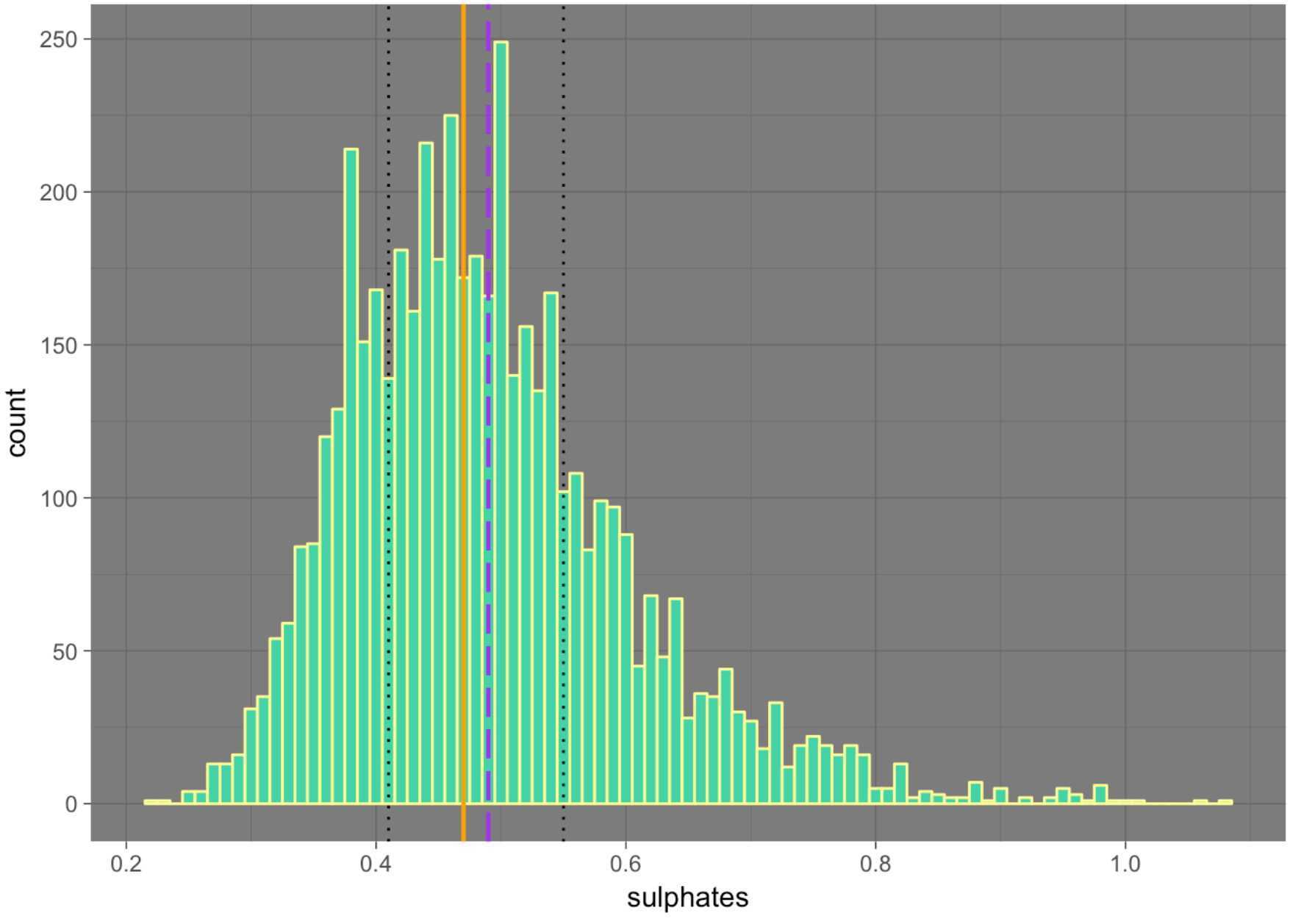
Both Free and Total Sulphur dioxied have symmetric data and both are measured in mg/dm^3



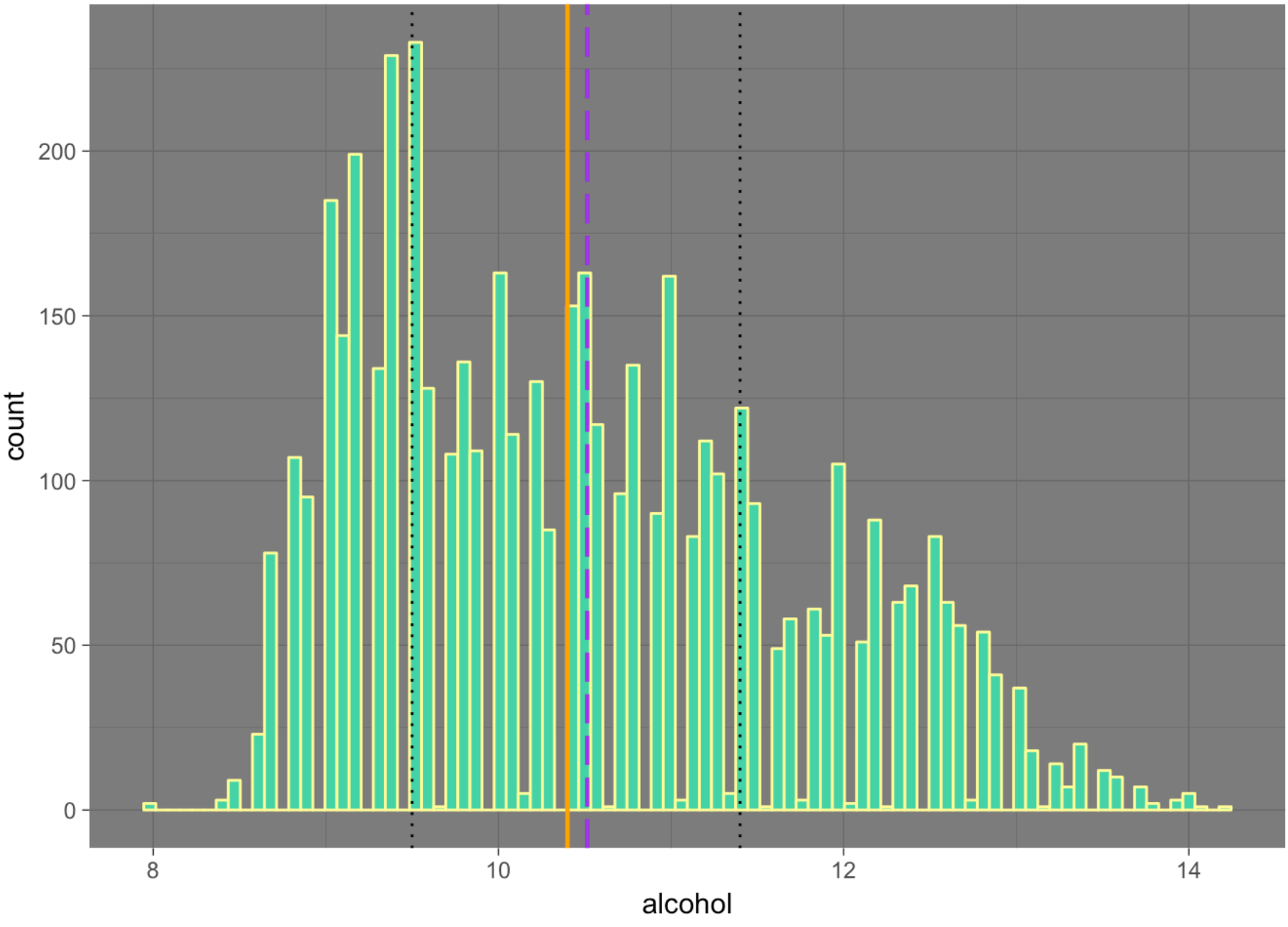
The density data is present in extremely narrow range. As indicated from summary and the visualisation above min value is .9871 and 3 Qt is at .9961. This is an very interesting observation because with almost 4900 observations this small range can help making a generalised statement like wines have density of .99 gm/cm^3. This value can be used for doing mathematical calculations.



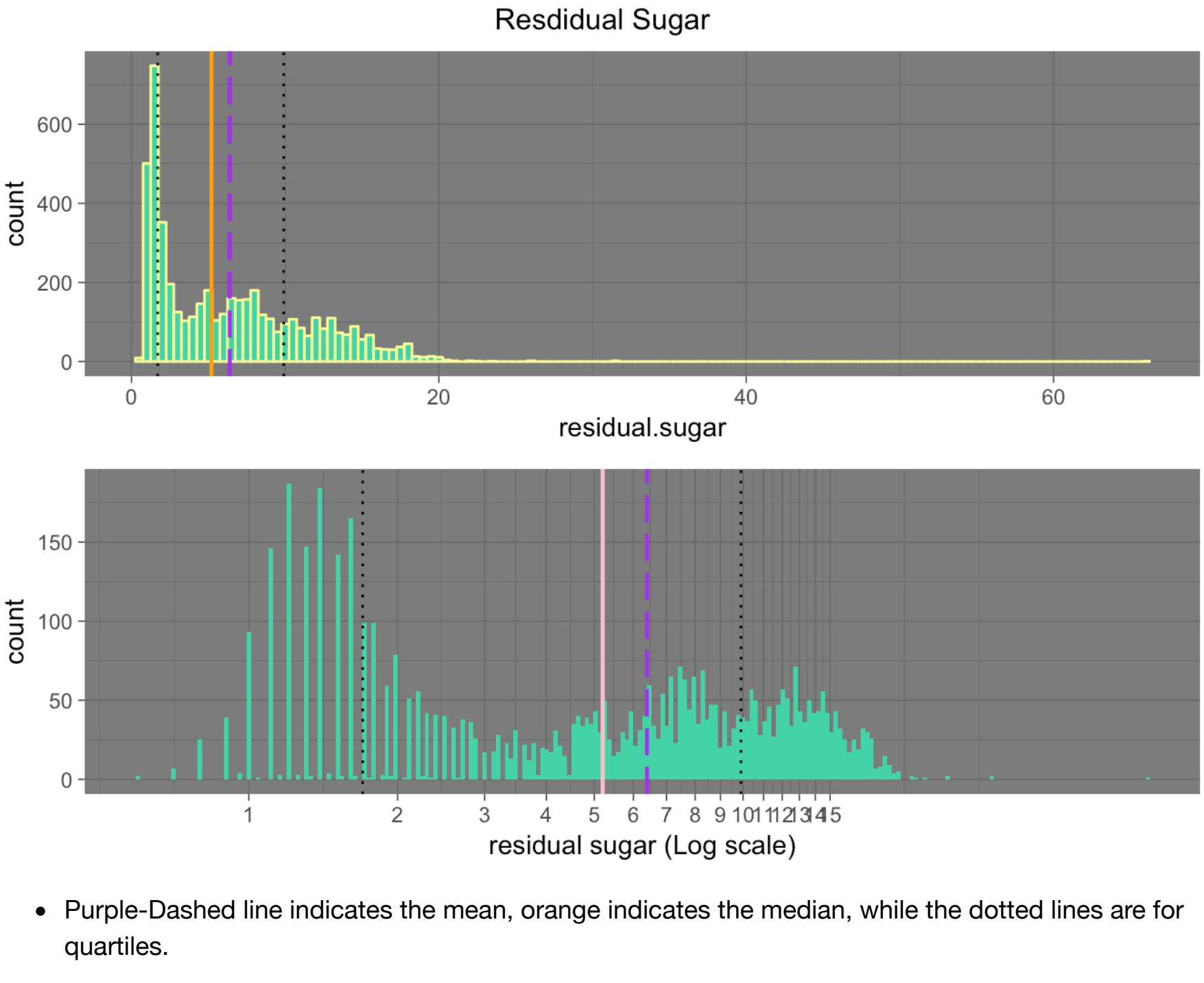
pH range of wine in our datset 2.7 and 3.8, with average of 3.18. pH is basically a measure of acidity with 0 being most acidic , 7, neutral and 14 being highly basic.



Sulphates are also measured in gm/dm^3 and the data have some peaks.

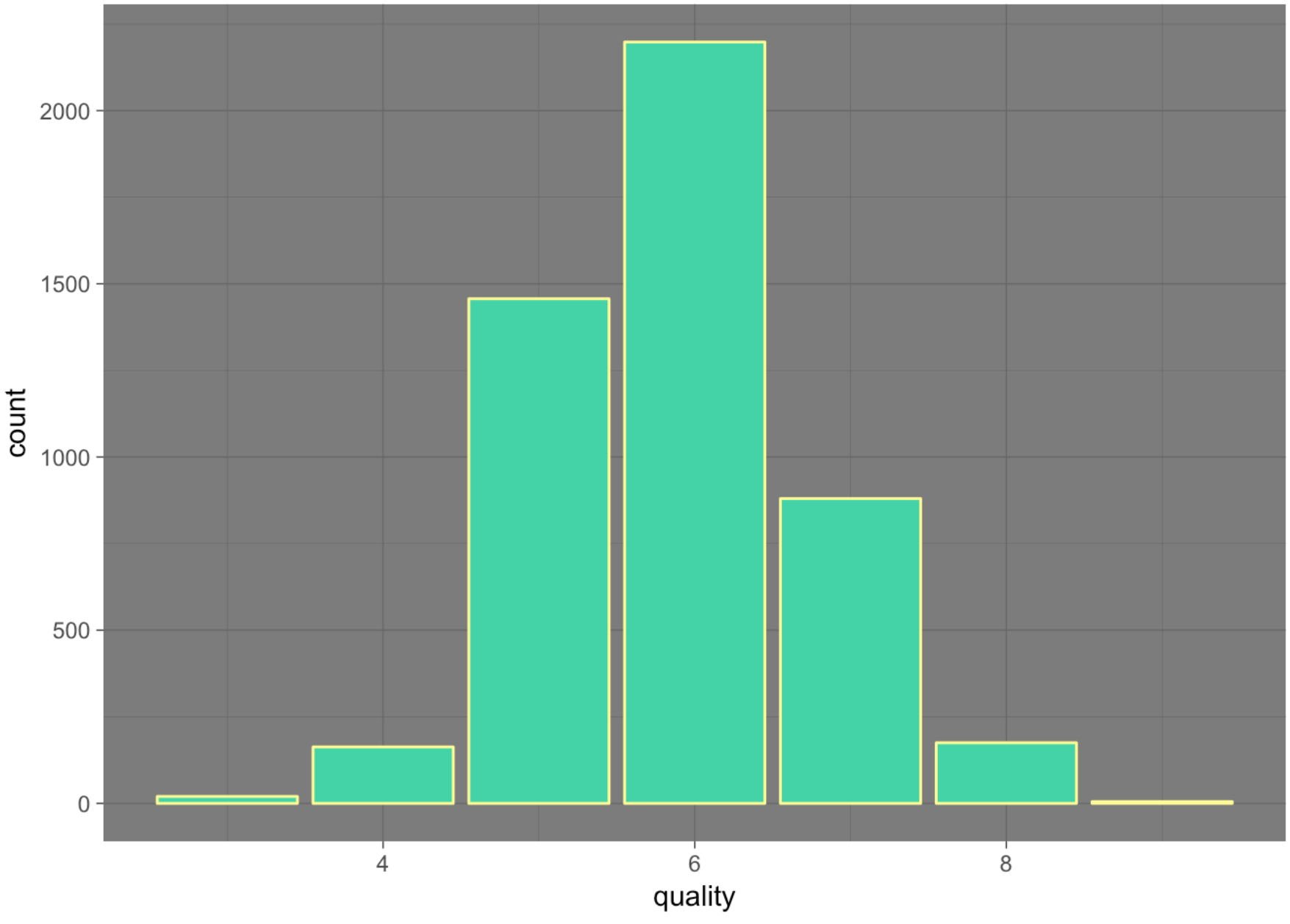


Alcohol measured in % volume dosen’t show much symmetry and have values between 8 and 14 with 75% lying below 11.4. As visible we have most no. of wines with alcohol percent between 9.5 and 9.6.



- Purple-Dashed line indicates the mean, orange indicates the median, while the dotted lines are for quartiles.

Residual Sugar, or RS for short, refers to any natural grape sugars that is leftover after fermentation ceases. When plotted on log scale showed much better visualisation. Half of the wines have RS between 0.5 and 5 gm/liter while the other half lies between 5 and 20 gm/liter. The data seems to have a normal distribution betweem 0 and 5 gm/liter on log scale.



Quality : score between 0 and 10

```
##    3    4    5    6    7    8    9 
##   20  163 1457 2198  880  175    5 
```

Most wines were rated 5 or 6, only 5 were rated 9

Univariate Analysis

What is the structure of your dataset?

The data has **11 parameters** and expert rating of **4898 wines**.

Input variables :

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: 1. Quality (score between 0 and 10) # Missing Attribute Values: None

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

1. Most of the wine are rated 5 or 6 by the experts, this might be an intersting point because generally human judgement tend to follow a normal distribution, i.e. most of us stays in between and avoid extremities. Although experts, judge on strict parameters but when close to 4900 wine are judged, a normal-ish pattern emerged.
2. Yeast feasts on sugar to produce alcohol in fermentation process, also alcohol and residual sugar dont follow the general trend of normal distribution, this may reveal some new patterns.

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

1. pH values conveys basic detials about acidity and acidity inturn plays an important role in providing wine its taste, so it might be an intersting factor.
2. Production of alcohol also varies densities.
3. Sulfur dioxide is an integral part of wine making, so it can’t be ignored here.

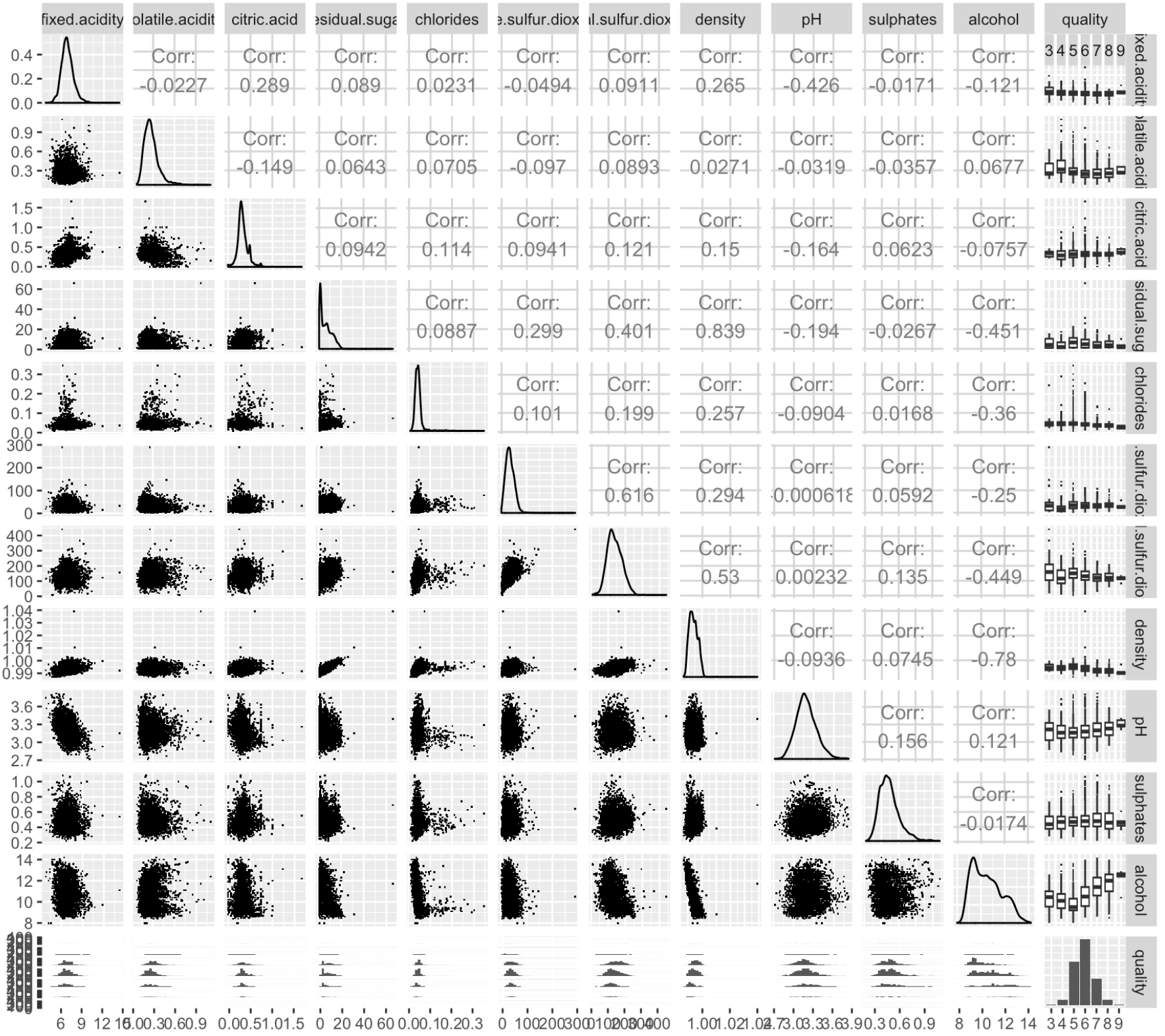
Did you create any new variables from existing variables in the dataset?

No new variables were created

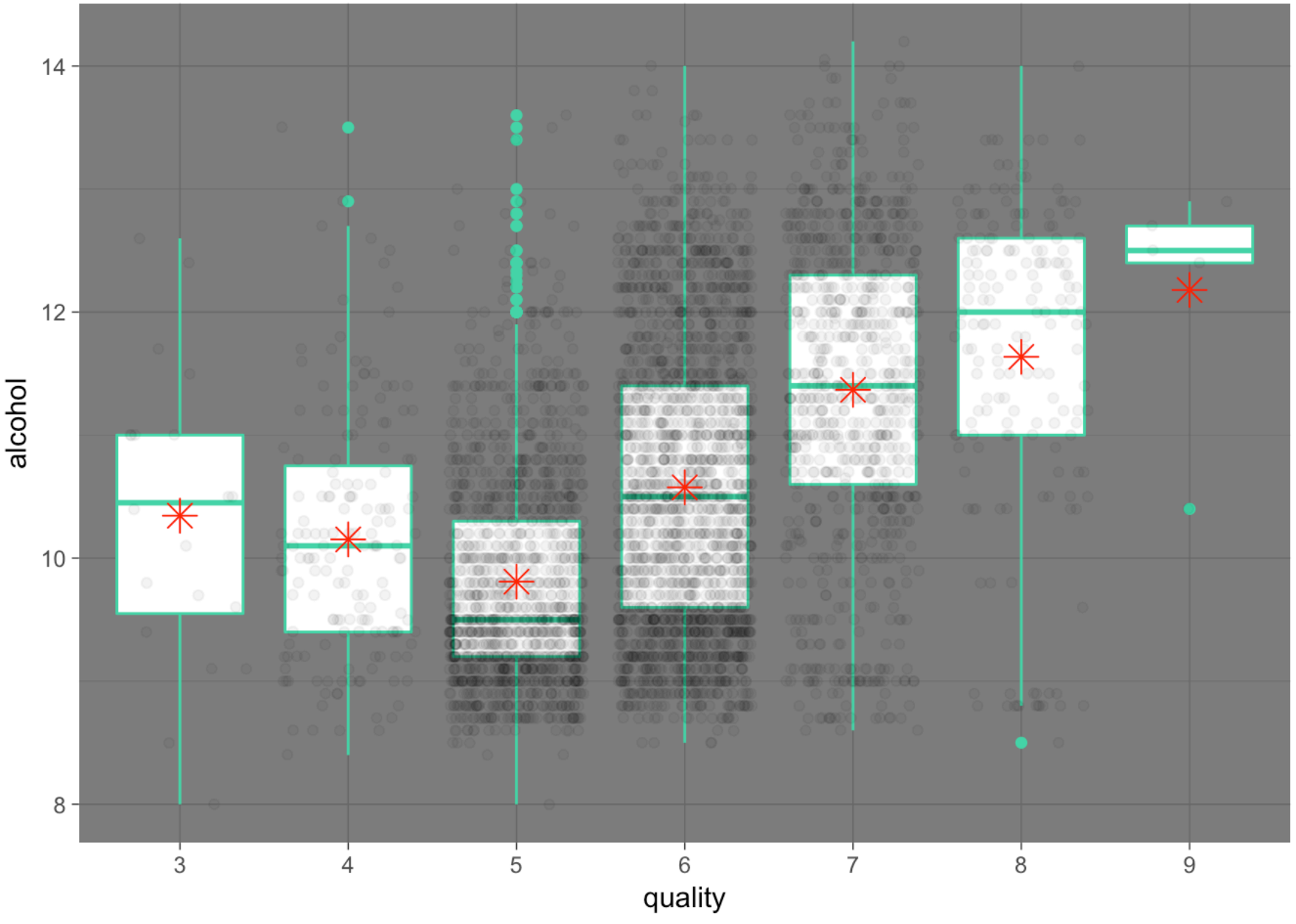
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

1. When observed closely on log scale residual sugar seemed to follow normal distribution in intervals. Log scale was chosen because the data showed skewness and most of the values were near 1 and 10
2. Quality variable was present as numeric variable initially, it was converted to ordered factor to make it more like a ranking system.

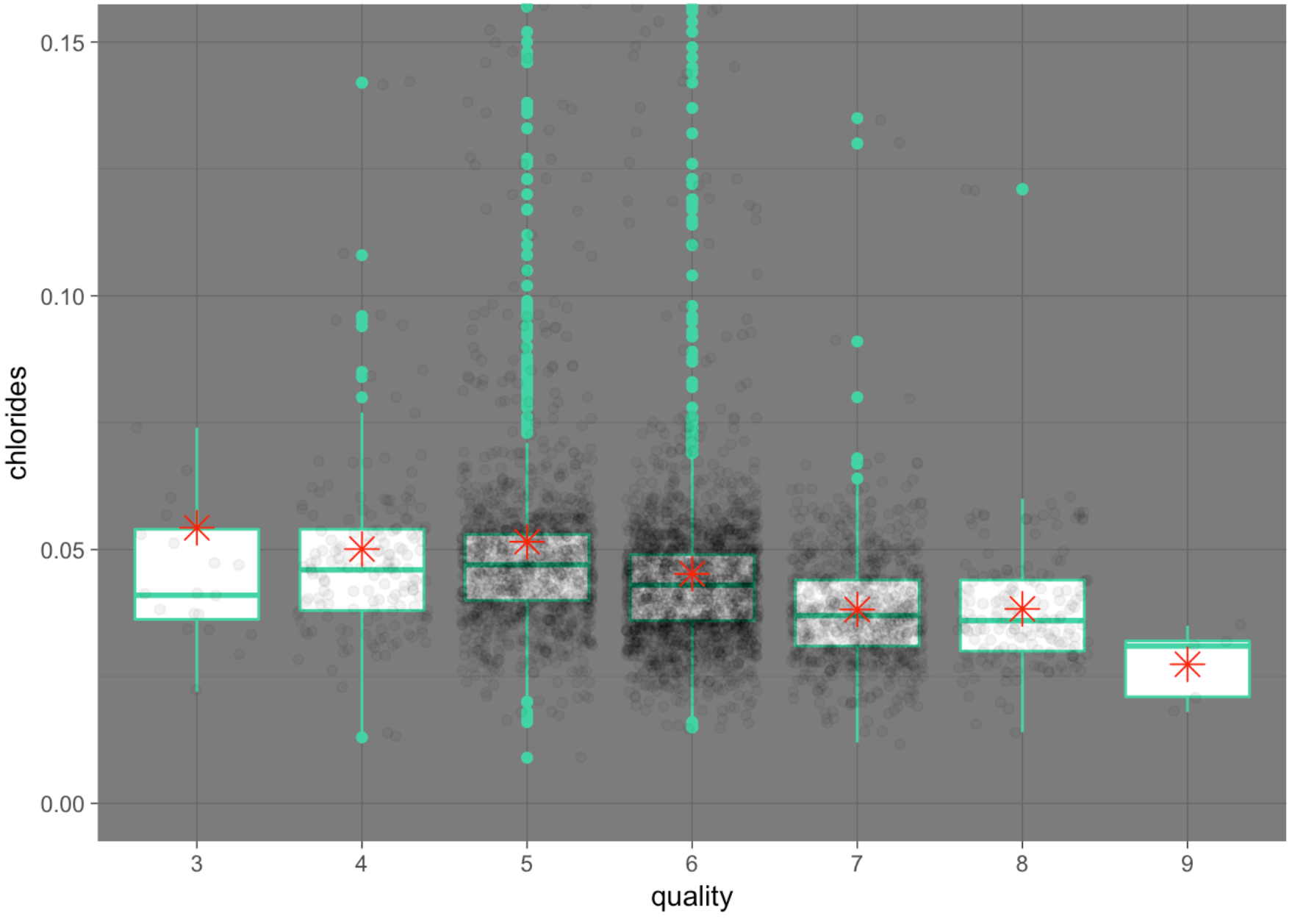
Bivariate Plots Section

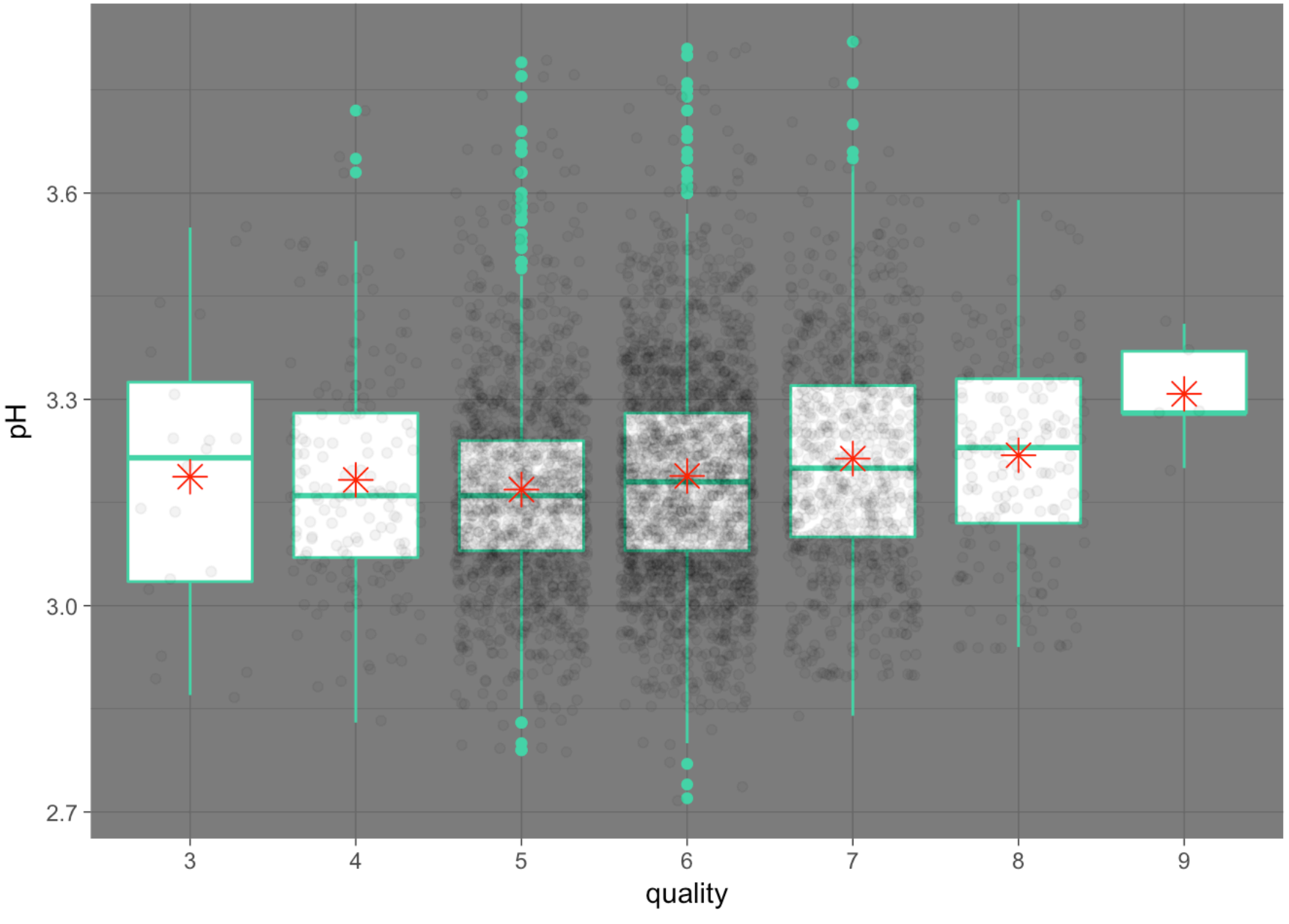


The upper triangle show corelation and lower triangle has the scatterplots between all the quantitative variables of dataset. Such matrix makes it very easy to pinpoint variables with noticeable pattern

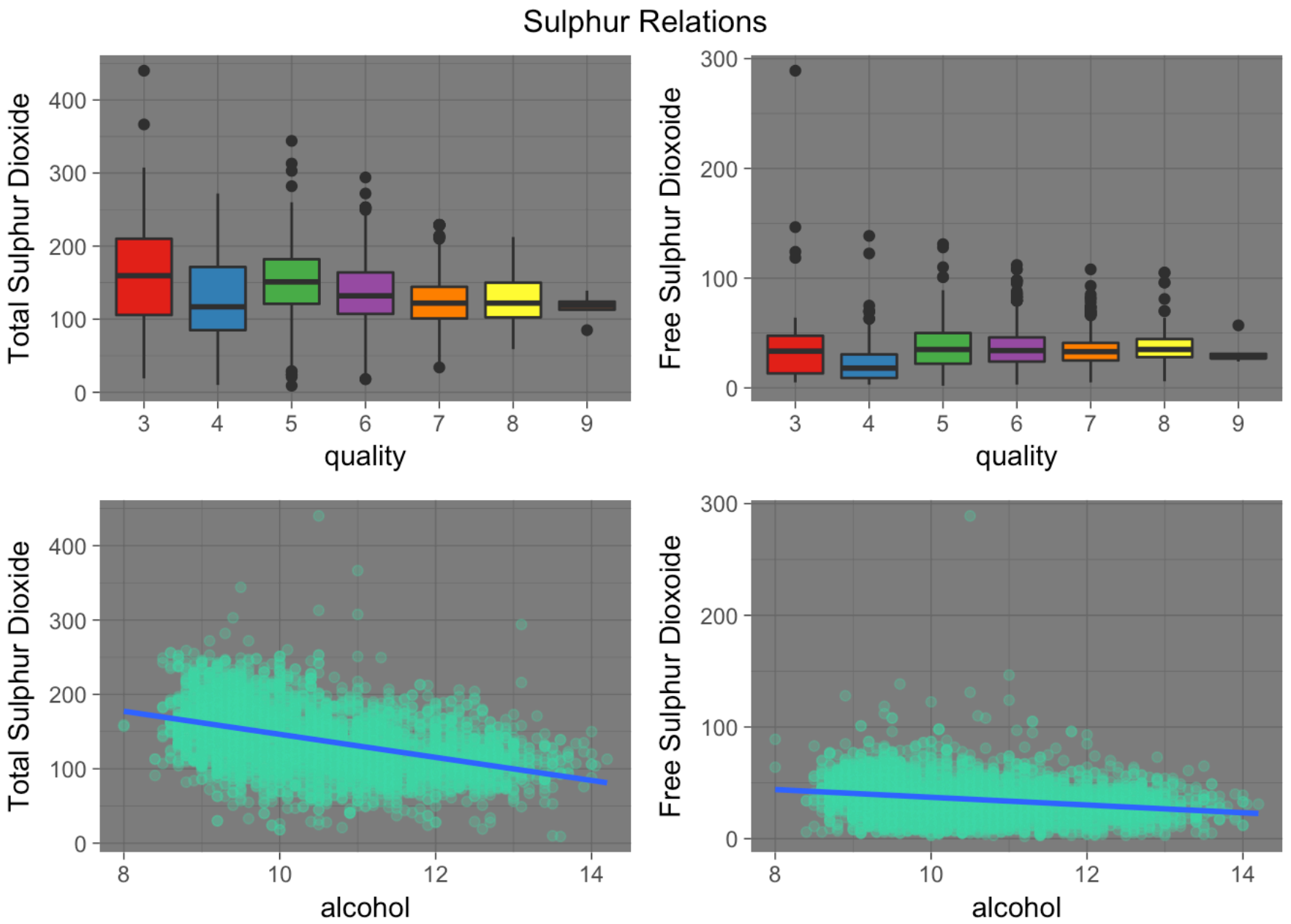


A box plot is an efficient way of visualising continous data against categorical data. Upper line of box indicates 3qt, lower indicates 1Qt while solid line inside the box indicates the median. A trend is clearly visible in the above plot.


chlorides
0.15
0.10
0.05
0.00
3
4
5
6
7
8
9
quality

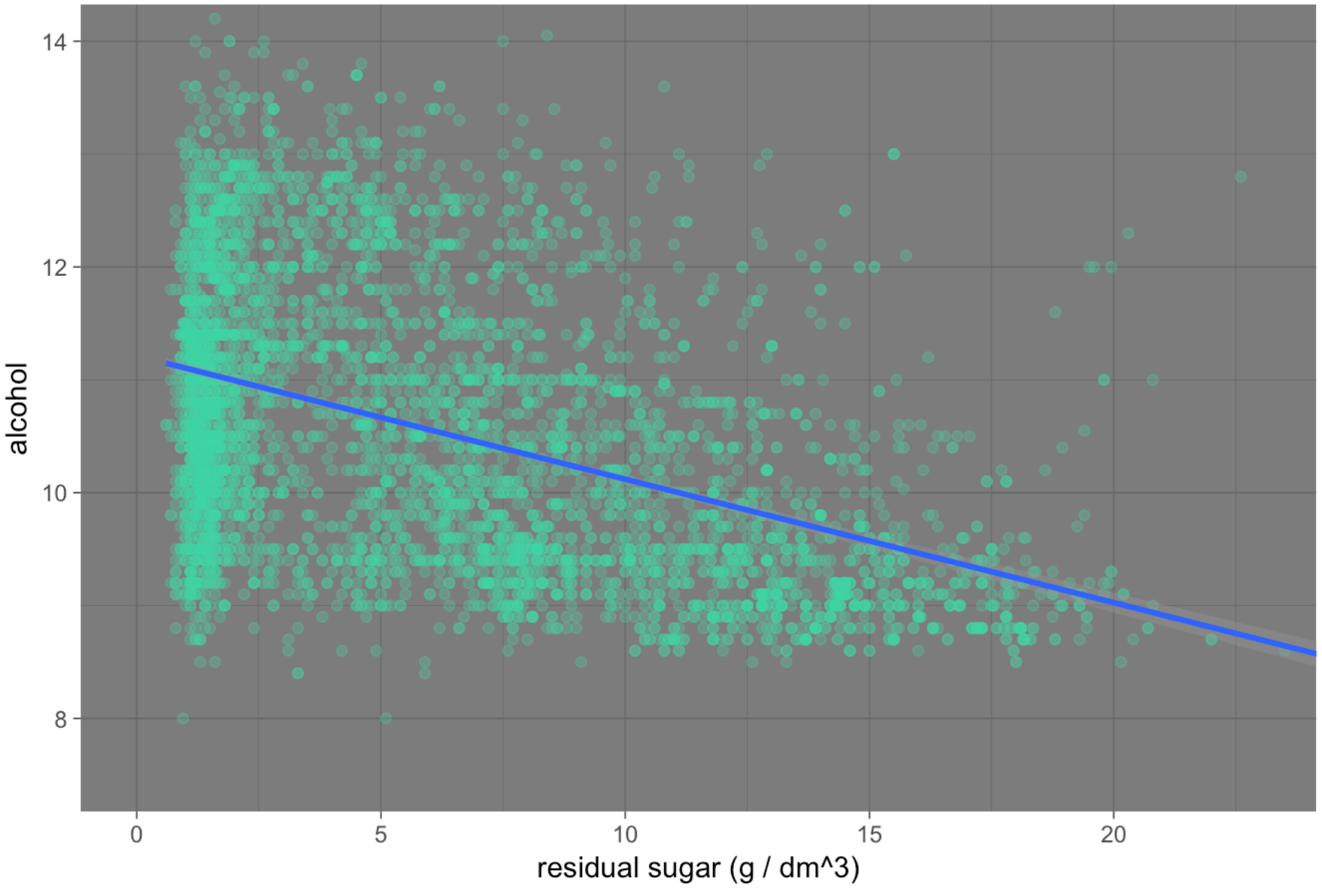

pH
3.6
3.3
3.0
2.7
3
4
5
6
7
8
9
quality

White wines of all ratings have pH in the range 3 to 3.3, except for very few. The " * " indicates the mean value of variables in different qualities.

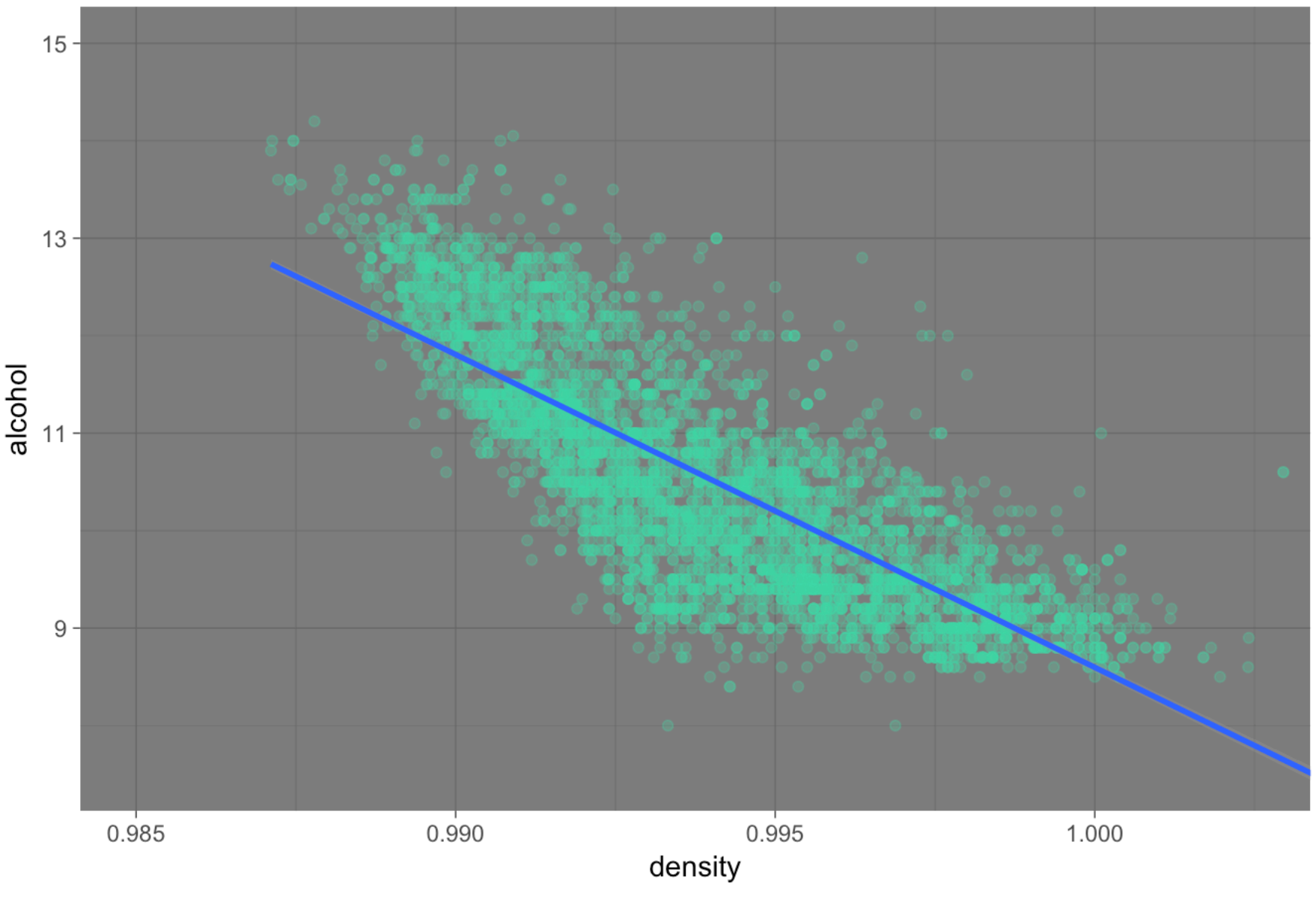


No paticular pattern emerged from the above graphs.

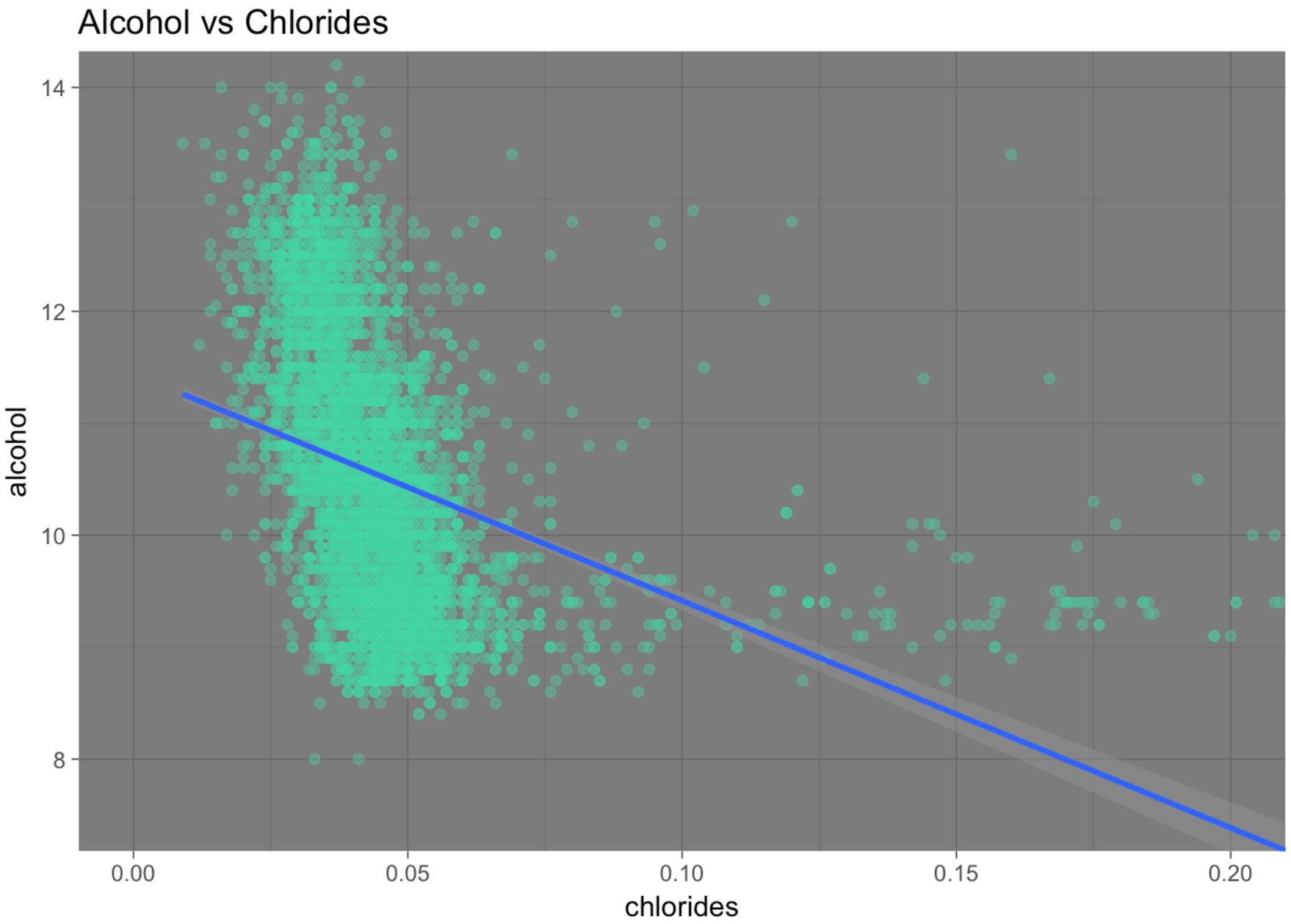
Alcohol vs Residual Sugar



Alcohol vs Density



Alcohol have negative correlatiion with Residual sugar and density, and a similar trend is observed in the scatterplots.



The majority of wines have chlorides between .01 and .075.

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?

Some negative correlations were observed in mathematical and visual form. The negative correlation between alcohol and residual sugar made complete sense, because yeast consume sugar to produce alcohol, and the sugar which is left after the fermentation ceases is the residual sugar. Infact one of the reason for fermentation reaction to stop is the increase in percentage of alcohol, yeast generally don't survive once 12%-14% concerntration is achieved and this is also visible in our data as wine having max alcohol has 14.2% .

Another interesting trend was wine rated higher have more alcohol percentage.

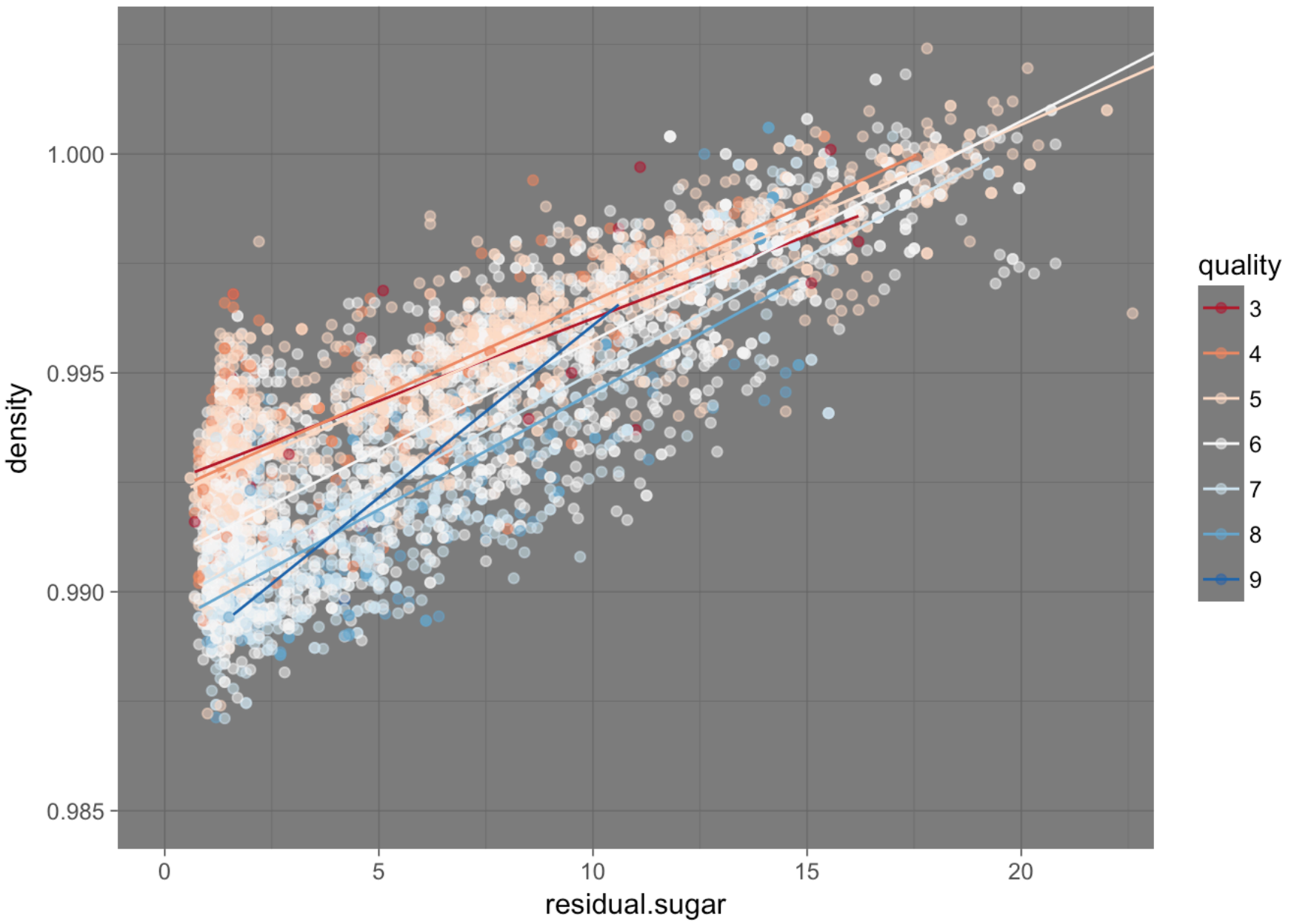
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Variation of density with alcohol percent. The density decreases as alcohol increases.

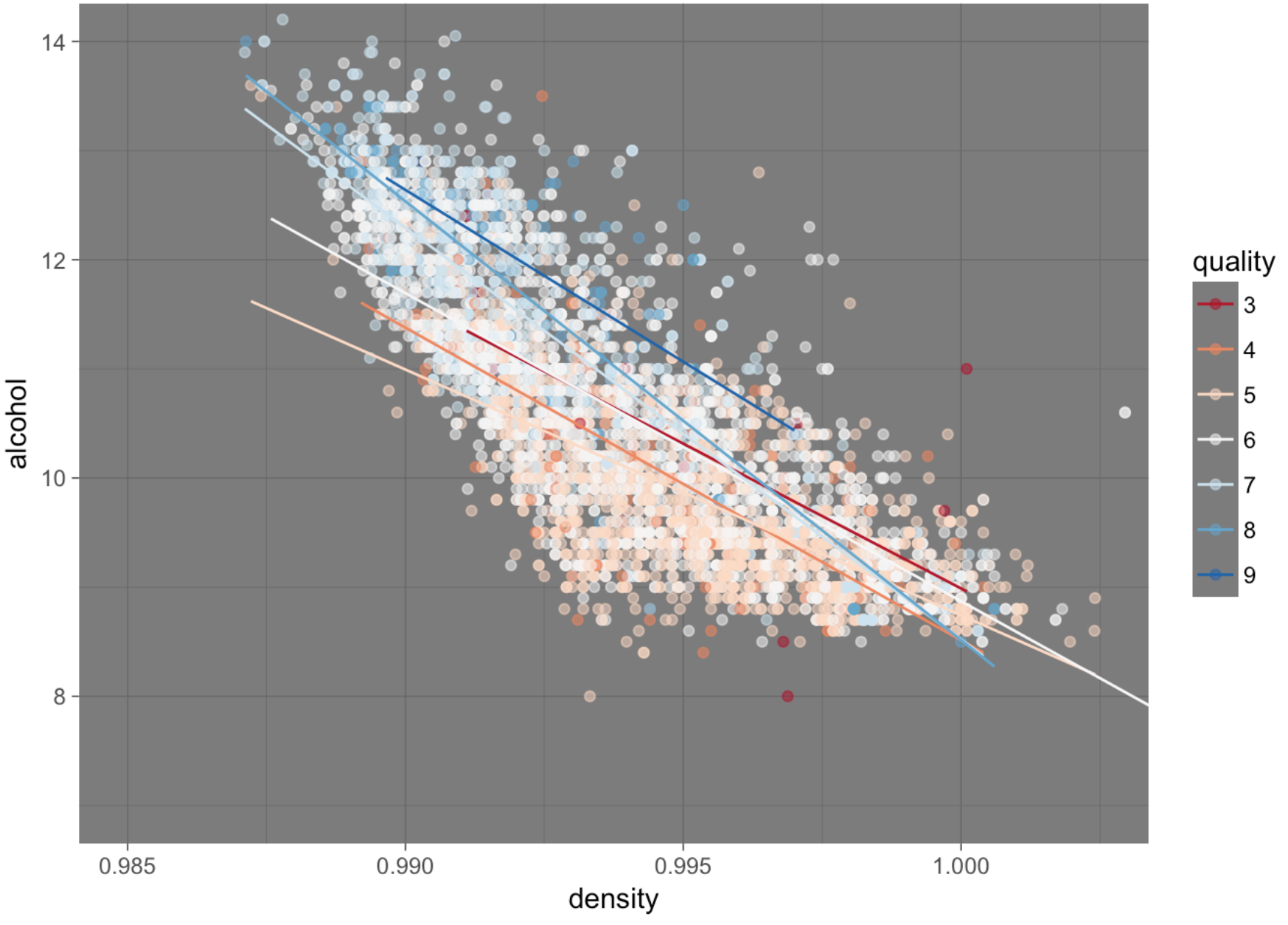
What was the strongest relationship you found?

1. The ranking of higher alcohol percentage wines was clearly more.
2. ALcohol and density had a strong negative corelation.

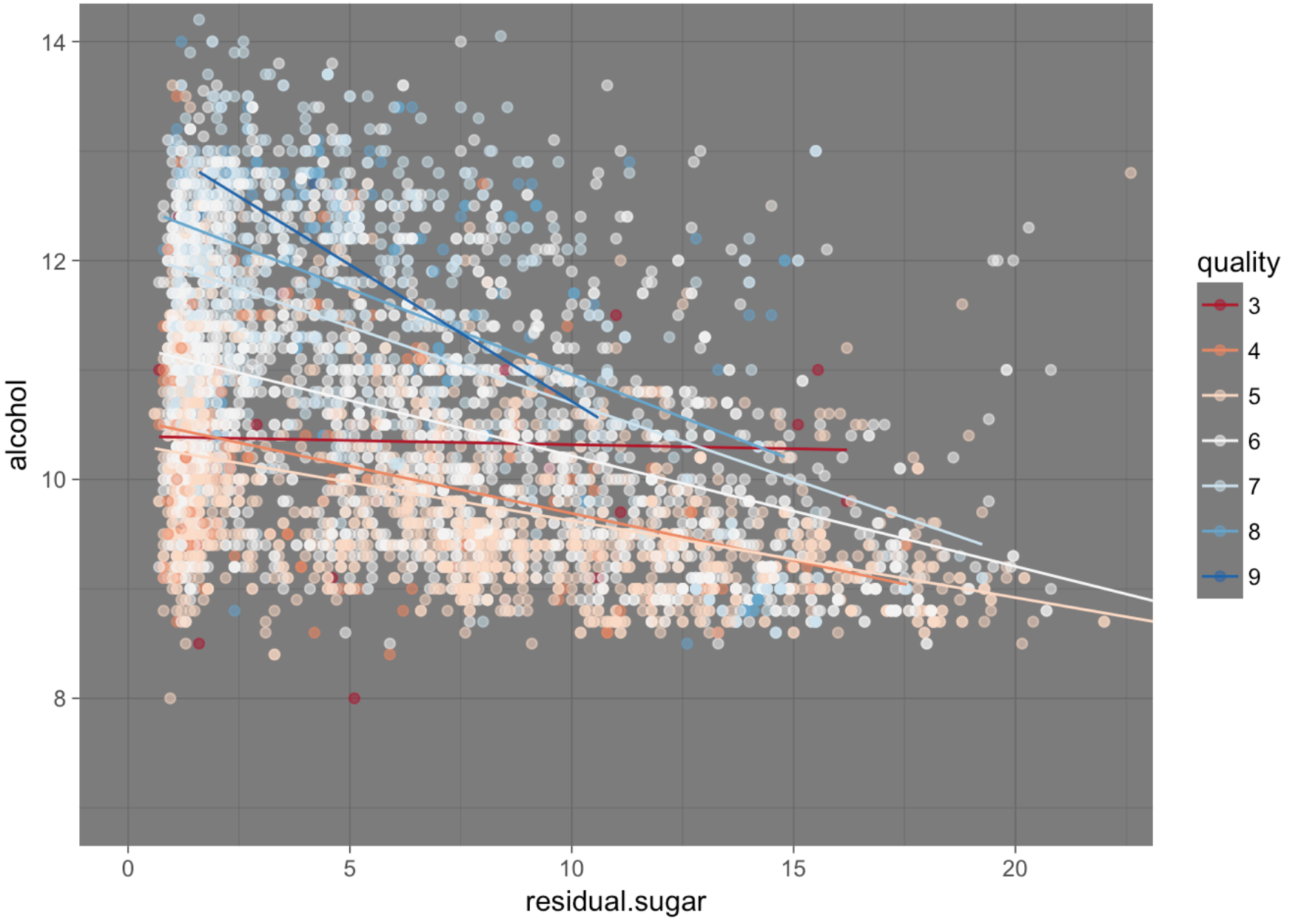
Multivariate Plots Section



Low density and low residual sugar i.e bottom left part of graph sees overcrowding of higher rated wines. The Lines are plotted according to a linear regression model.



The graph shows two clusters one with rating 7 and obove with density .990 and second one with rating 6 and below with density .995-1.00.



The above three graphs shows that higher rated wine have density and residual sugar on the lower side and alcohol on the higher as side, the Lines are plotted according to a linear regression model, which makes it even more clearer to visualise that.

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?

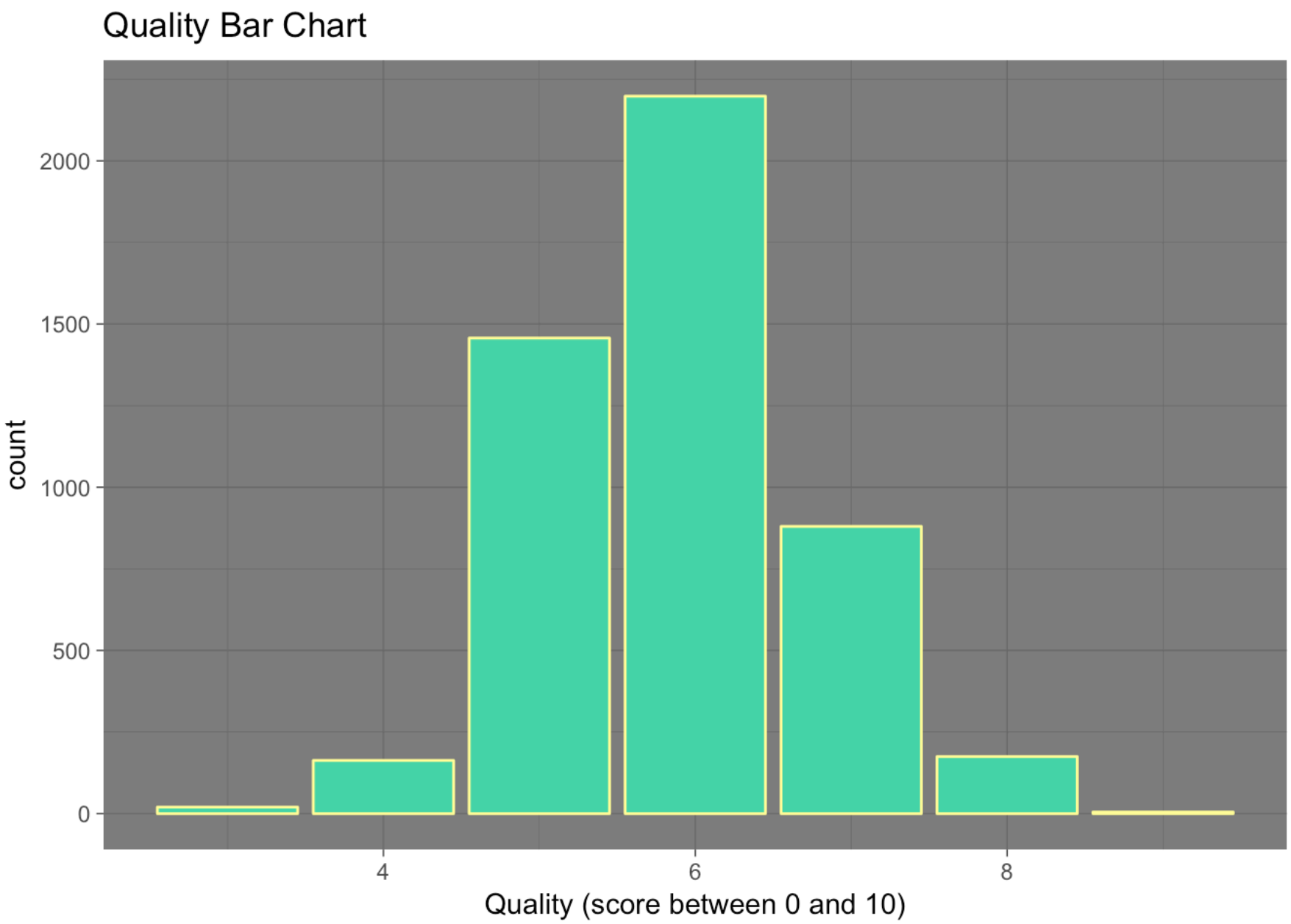
As observed earlier higher rated wines tend to have higher alcohol percentage which inturn is negatively corelated with residual sugar and density, so when all these variables were plotted together, the whole concept made more sense

Were there any interesting or surprising interactions between features?

The plotted variables interacted as expected.

Final Plots and Summary

Plot One

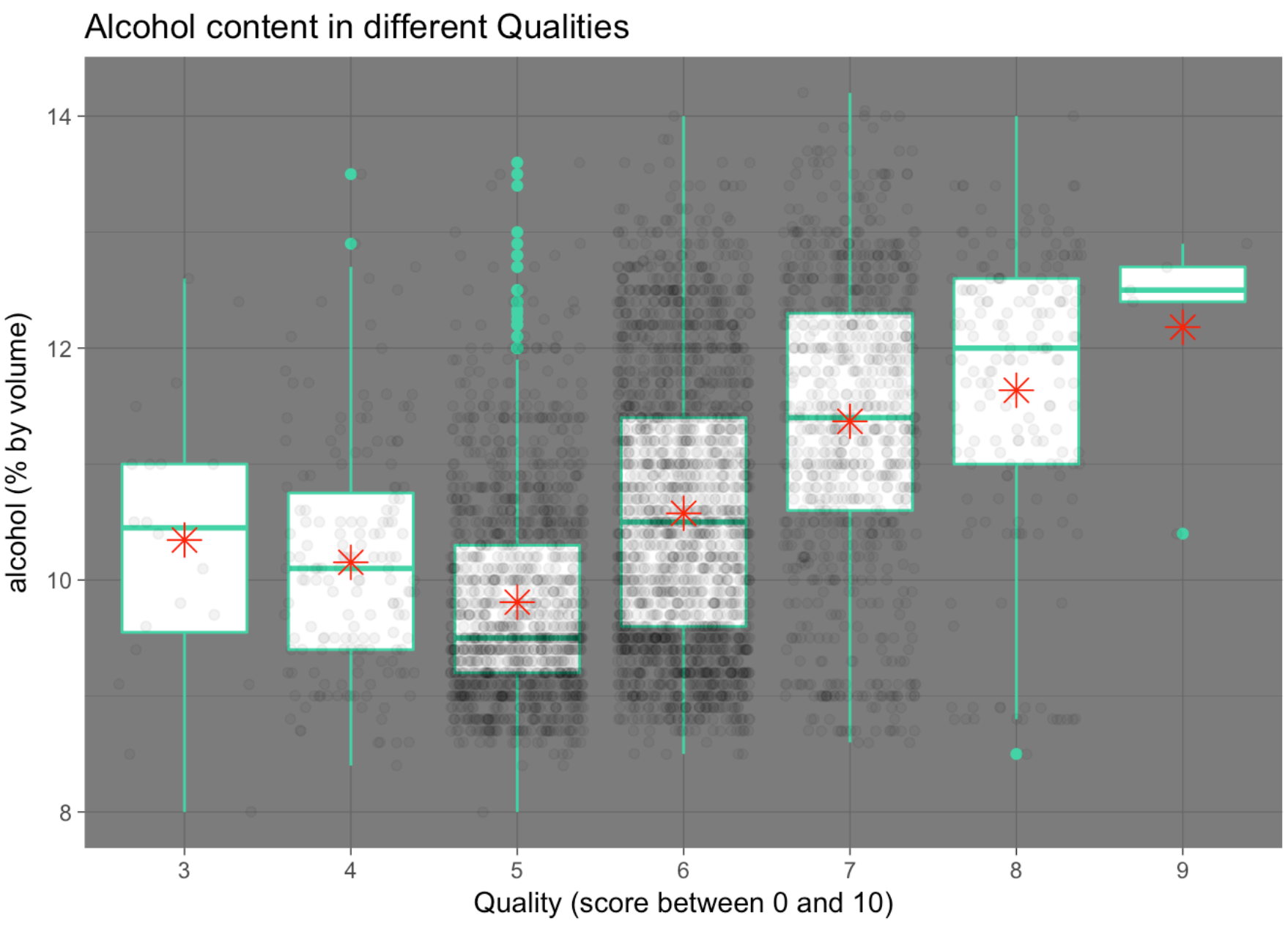


> “Accept what life offers you and try to drink from every cup. All wines should be tasted; some should only be sipped, but with others, drink the whole bottle.” — Paulo Coelho, Brida

Description One

This one simple graph is conveying a lot more than just ranking, there is a high possibility it is defining human behaviour. A simple search on google about about wine making provides methods to make your own wine. Apart from some basic framework the wine making process is very flexible, and we generally try to make thing just right. Not too spicy. Not too sweet. Just right. Most of the things we buy or make have to fit in that middle ground. The whole process from selecting grapes to fermentation to ageging is relying on the different definitions of just right, maybe this is the reason there are very few wines with rating 3 or 9. That is why the presence of normal distribution in most of the parameters makes a lot of sense. Although, it goes without saying that what might be 3 for one will be a 9 for other, but that also brings a lot of other factor in picture, anyhow in general human behaviour follows a normal distribution.

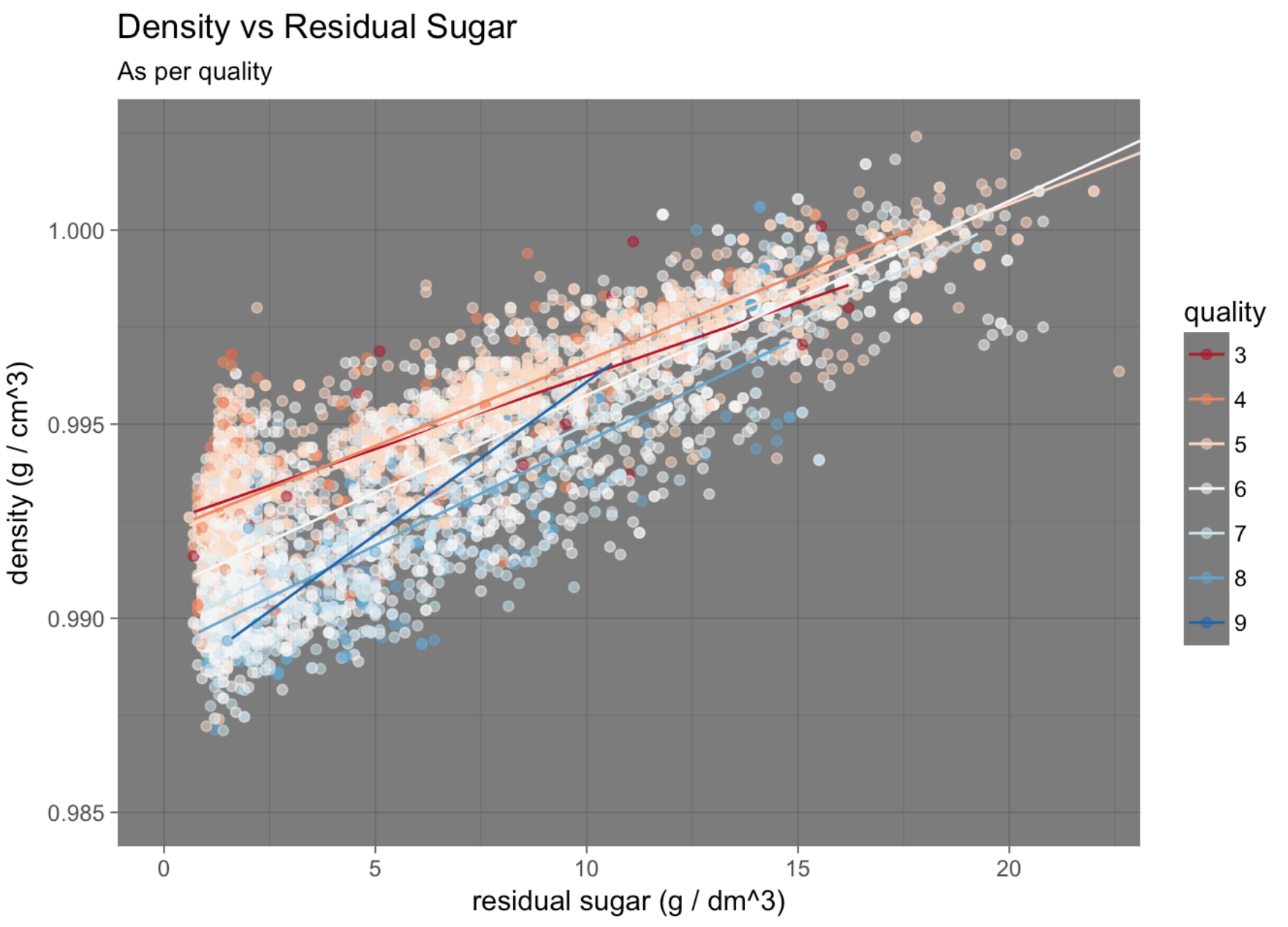
Plot Two



Description Two

The graph showed a significant pattern that more alcohol content returns better rating. This might also be an insight to rating sysytem because ~1% change in median alcohol value in wines with rating 5 and above, improved the rating by one unit.

Plot Three



Description Three

This graph shows all the key observations made among the parameters. Higher rated alcohol has low sugar and low density. Fermentation ceases once alcohol percentage is in between 12-14 (remember alcohol and residual sugar are negatively co-related), thus reaching that levels is like getting the maximum out of yeast. Fermentation also depend on other factors like temperature, acidity etc and mantaining those at higher level of alcohol(low residual sugar) can be even more challenging, thus managing to get those higher level is clearly rewarded by experts. ——

Reflection

I started by studying some basic fundamentals about wine making to get a better understanding of data like the role of sulphur dioxide, acidity etc. Then proceeded to make some random graphs matrix which helped me selcting some key variables like alcohol percentage, residual sugar. Working on this dataset

The persence of normal distribution among most values facinated me alot but that also became a challenge, because one can point out abnormalities and express those in words. With normal distributions only few options like mean, median, quartile range, come to mind initially but on close observations and patience data started making sense. Another challenge was making a clean code which was overcomed by use of custom functions

In future I hope to create a model for predicting the quality of wine, also innclusion of price, age of the wine and adding more regions could make this dataset even more exciting.

“Either give me more wine or leave me alone.” — Jalaluddin Rumi