What dictates the quality of Red Wine by Matthew Morris

In order to understand the wine variables, I will plot distribution of each variable. This will give me an idea if there is normal distrubution, bi-modal, skews or outliers that will need to be addressed.

References Wine Understanding

Quality wine lables https://www.winespectator.com/articles/scoring-scale (https://www.winespectator.com/articles/scoring-scale)

Acidity in wine (fixed.acidity, volatile.acidity,ph) https://winefolly.com/deep-dive/understanding-acidity-in-wine/ (https://winefolly.com/deep-dive/understanding-acidity-in-wine/)

Sugar levels and wine http://www.naijawinelovers.com/what-is-residual-sugar-in-wine/ (http://www.naijawinelovers.com/what-is-residual-sugar-in-wine/)

Alcohol levels and wine https://winefolly.com/tips/the-lightest-to-the-strongest-wine/ (https://winefolly.com/tips/the-lightest-to-the-strongest-wine/)

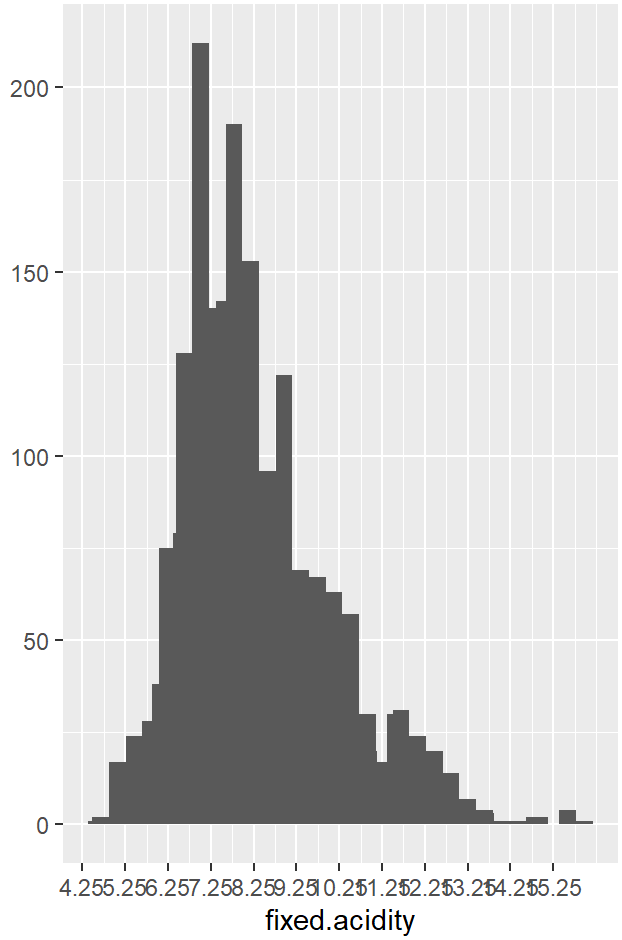
pH https://winemakermag.com/technique/1650-monitoring-adjusting-ph (https://winemakermag.com/technique/1650-monitoring-adjusting-ph)

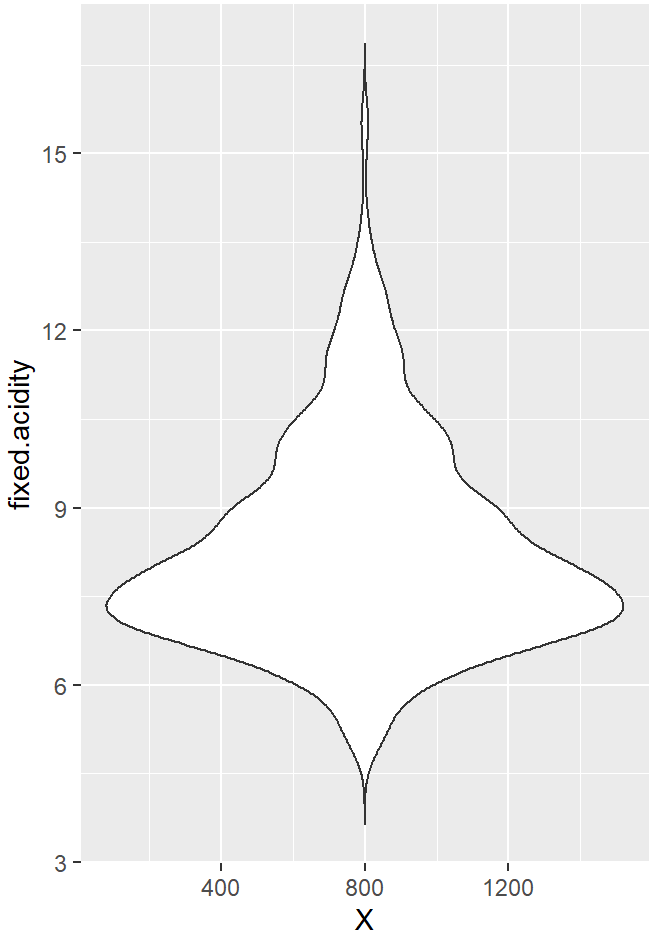
R references

https://mgimond.github.io/ES218/Week09a.html (https://mgimond.github.io/ES218/Week09a.html) https://rpubs.com/Mentors_Ubiqum/scale_x_continuous (https://rpubs.com/Mentors_Ubiqum/scale_x_continuous) http://www.sthda.com/english/wiki/ggally-r-package-extension-to-ggplot2-for-correlation-matrix-and-survival-plots-r-software-and-data-visualization (http://www.sthda.com/english/wiki/ggally-r-package-extension-to-ggplot2-for-correlation-matrix-and-survival-plots-r-software-and-data-visualization) http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs (http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs) http://www.sthda.com/english/wiki/qplot-quick-plot-with-ggplot2-r-software-and-data-visualization (http://www.sthda.com/english/wiki/qplot-quick-plot-with-ggplot2-r-software-and-data-visualization) https://stackoverflow.com/questions/38446804/coloring-points-based-on-variable-with-r-ggpairs (https://stackoverflow.com/questions/38446804/coloring-points-based-on-variable-with-r-ggpairs)

Univariate Plots Section

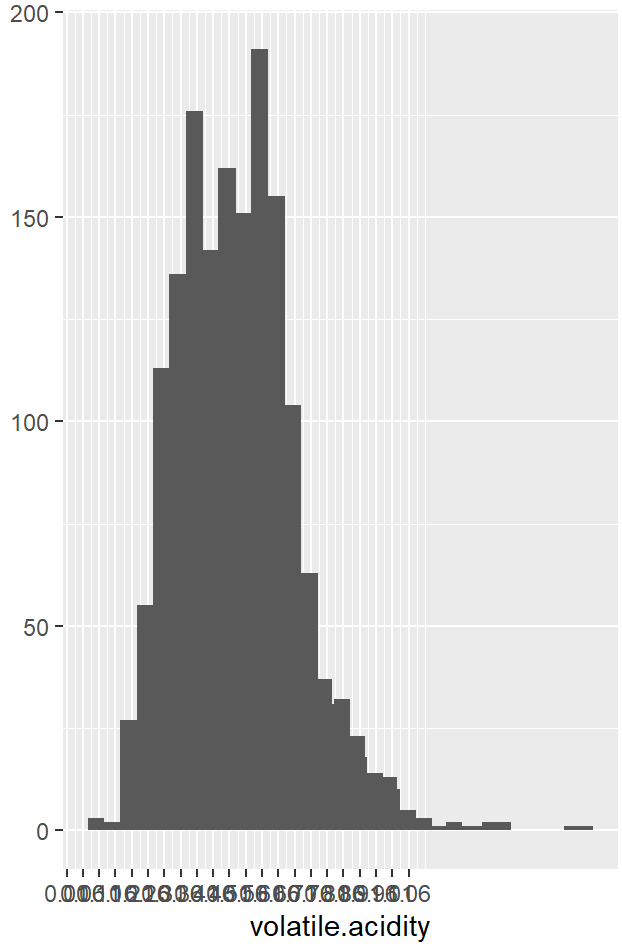
Initial impressions of distribution

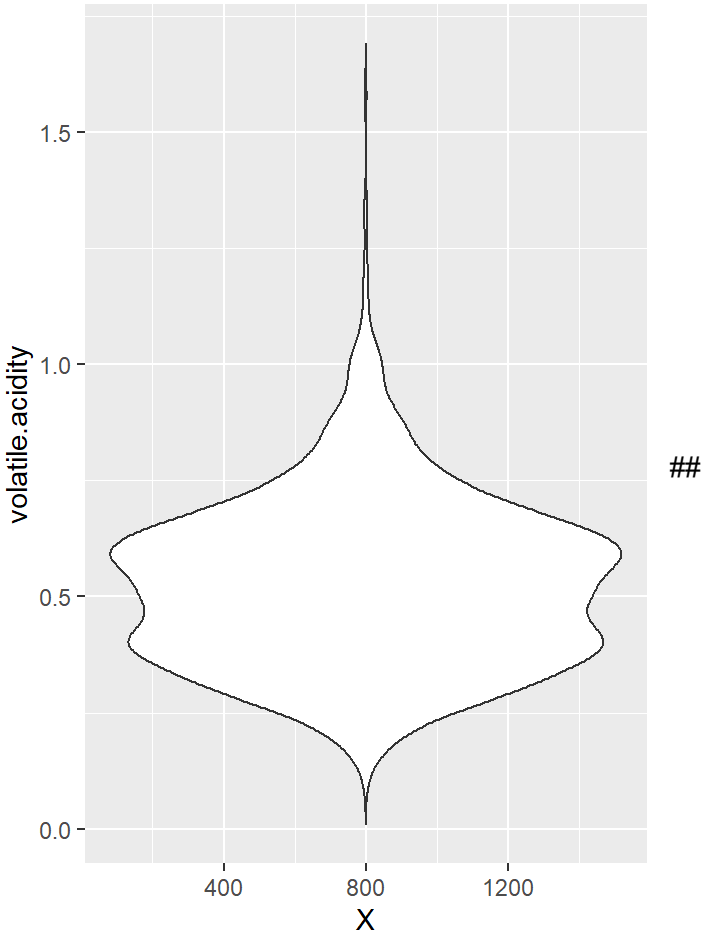




##

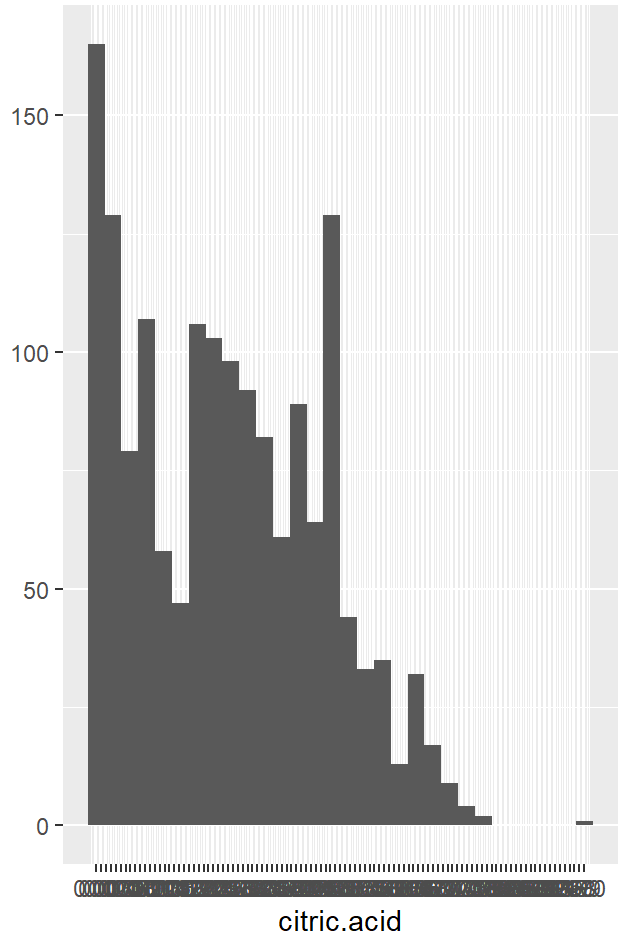
discovery - majority of fixed acidity population between 5.75 and 10.75

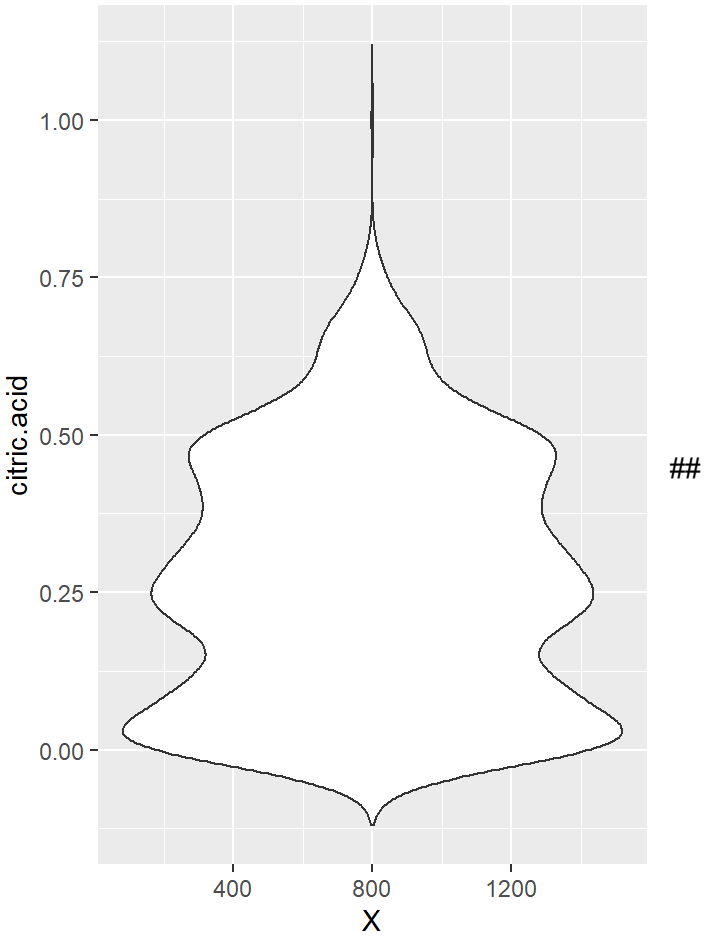




##

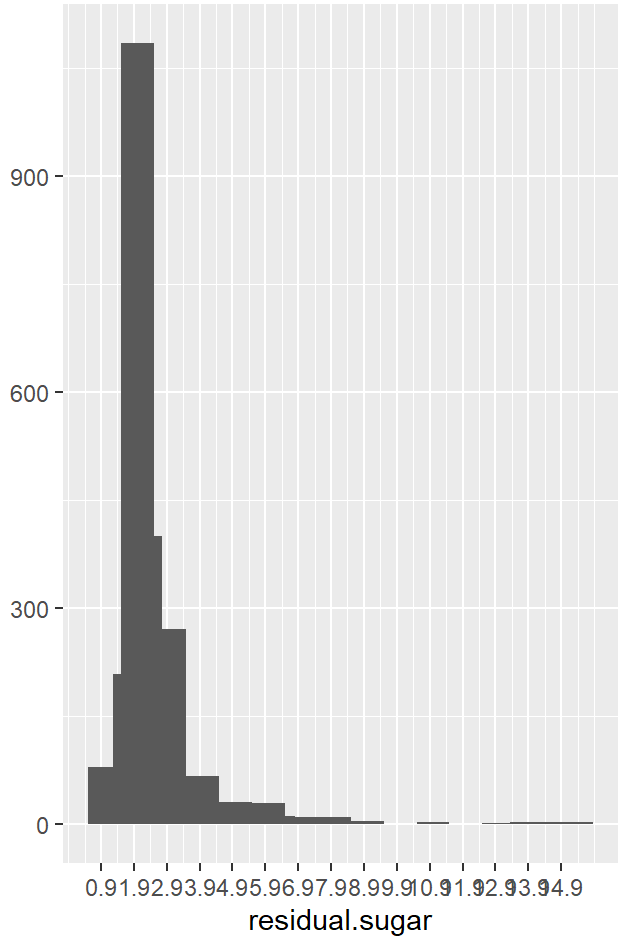
discovery - majority of volatile population between .25 and .75

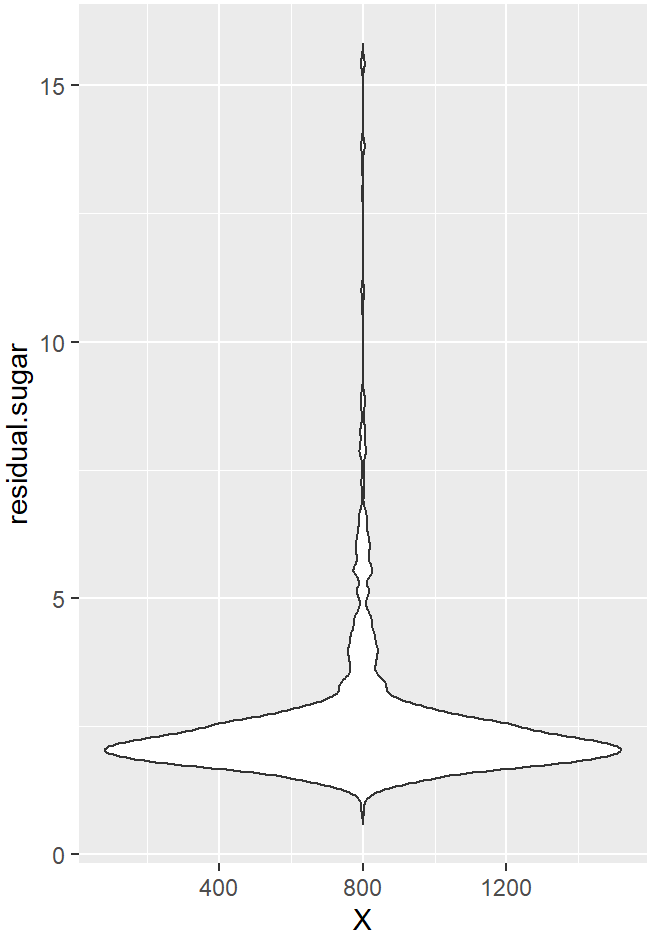




##

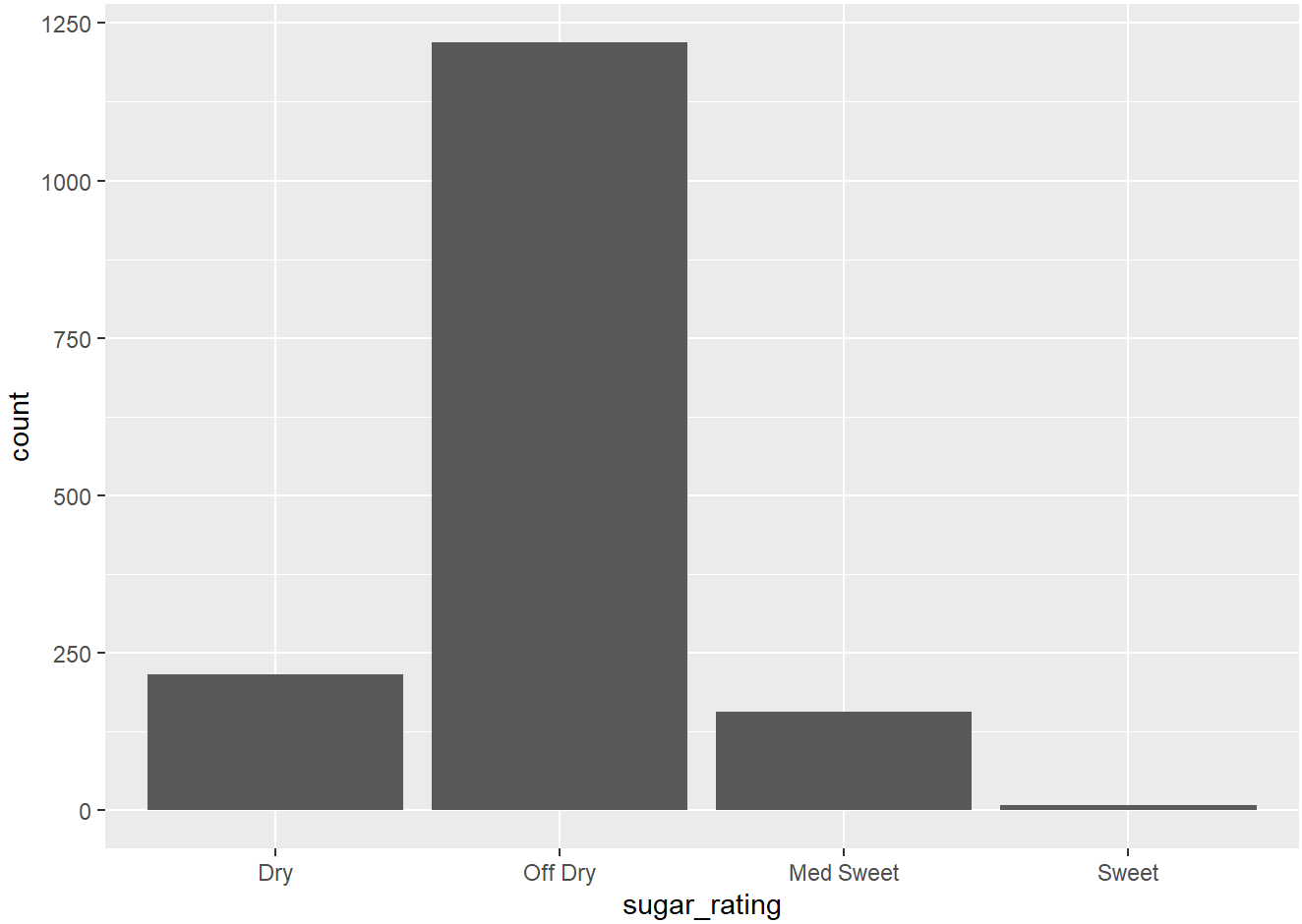
discovery - citric acid has a slight bimodal distribution between .0 and .25 and .35 and .5





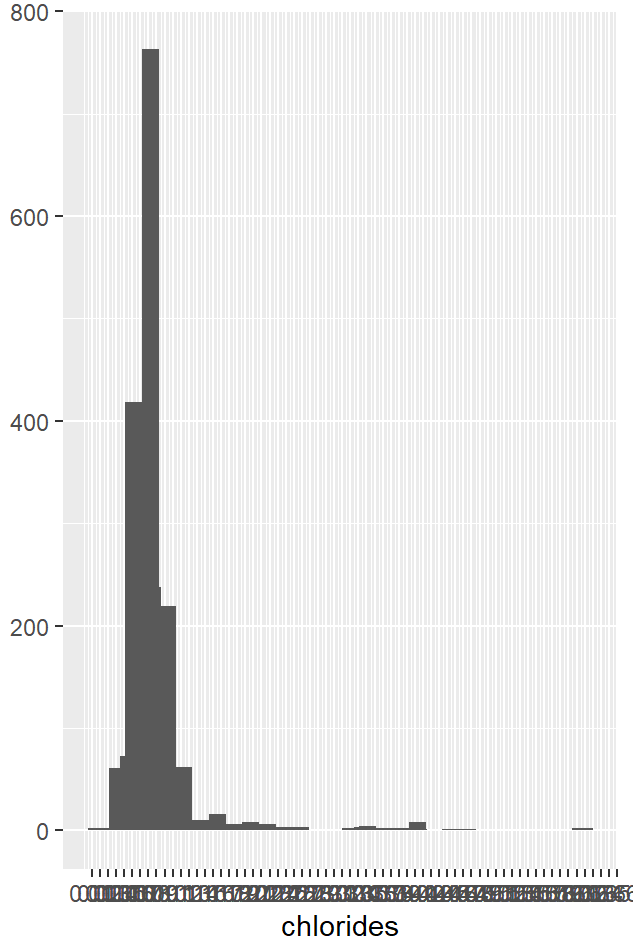
##

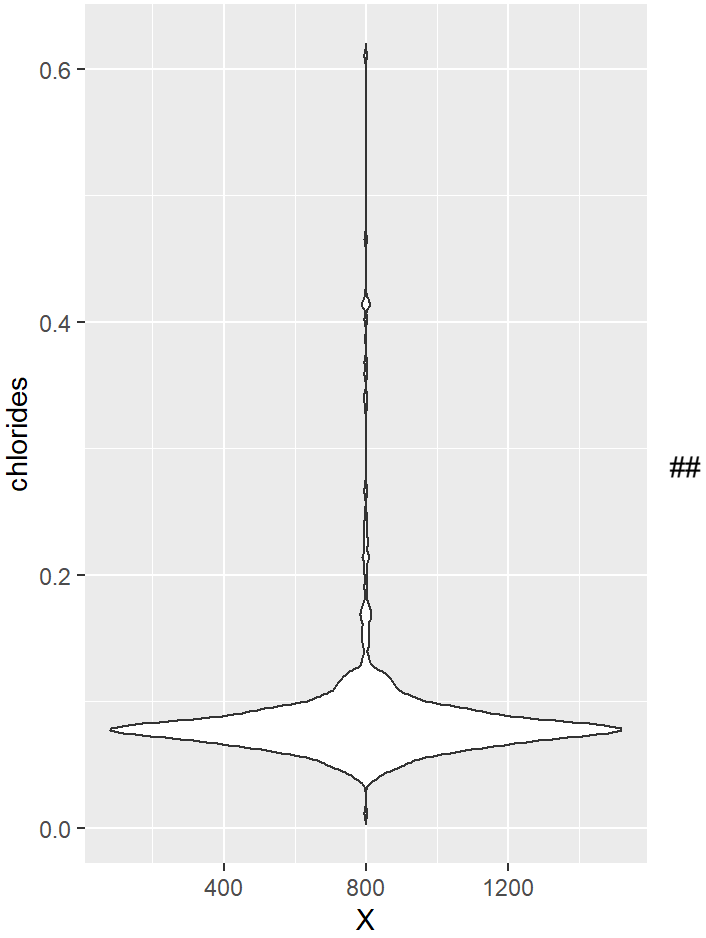
discovery - residual.sugar is skewed right with outliers population between 1 and 3



##

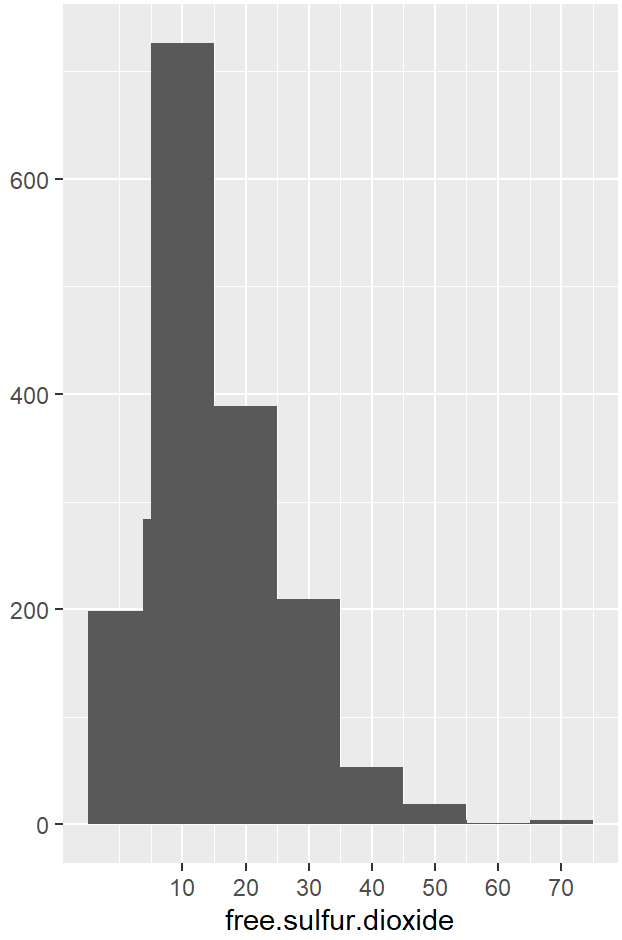
Data set is dominated by Off Dry wines

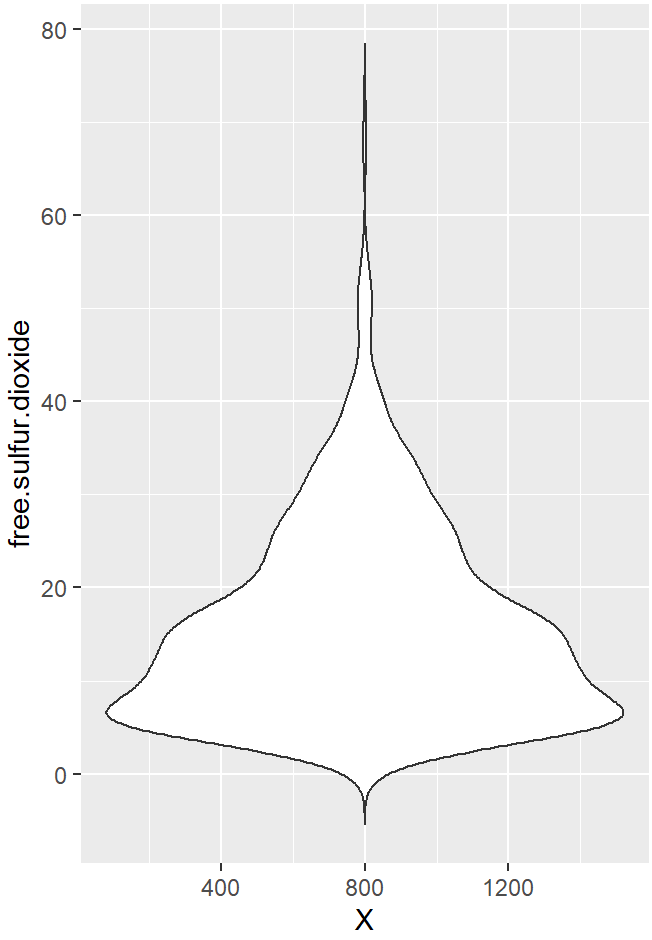




##

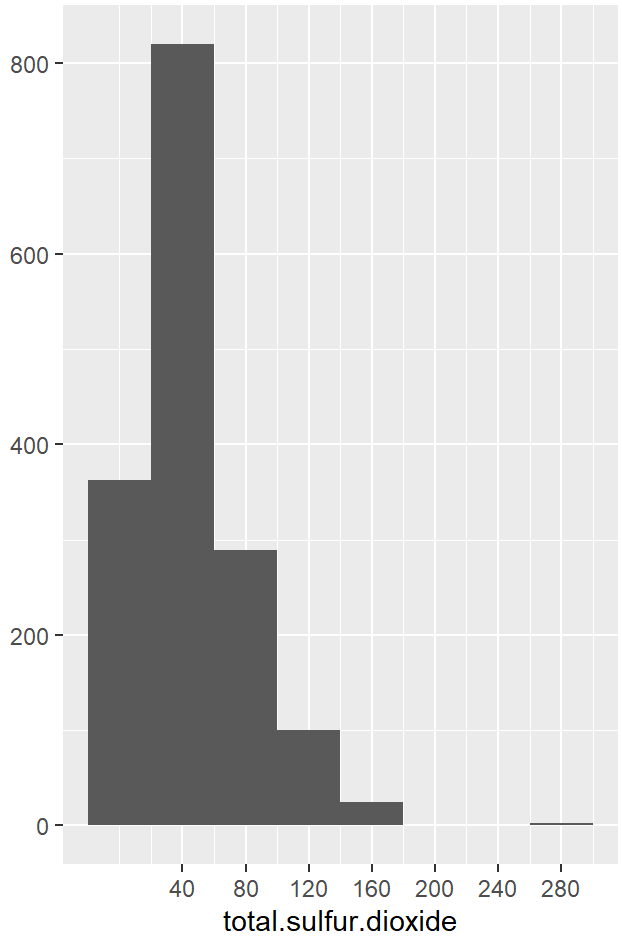
discovery - chlorides are skewed right with outliers

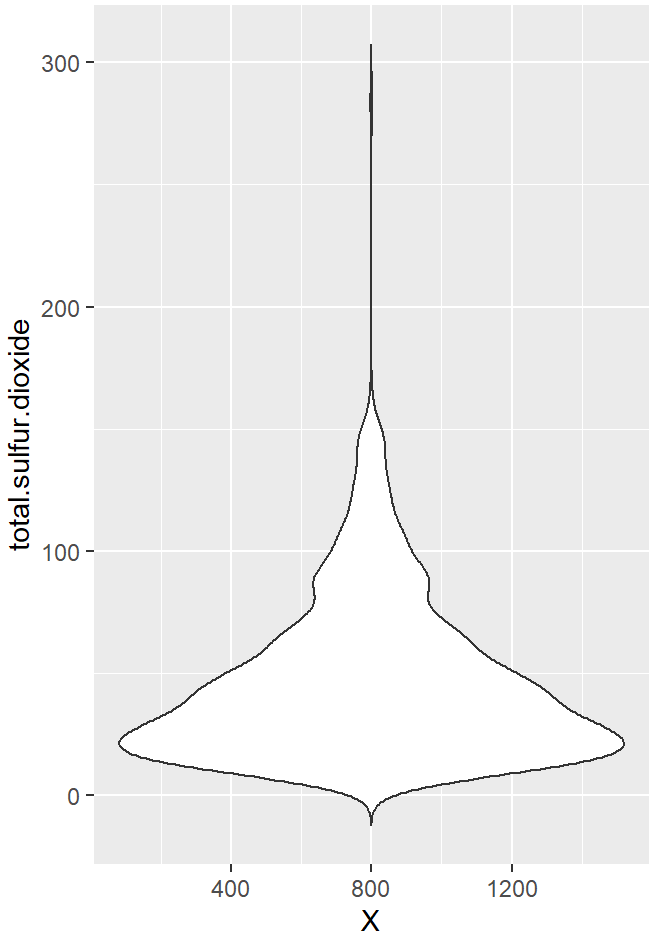




##

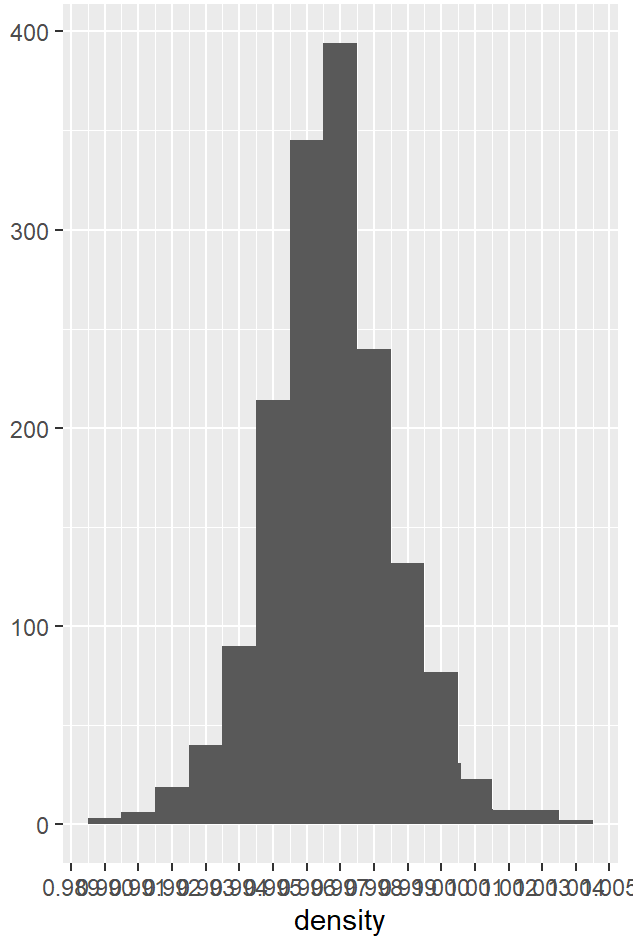
discovery - skewed right with outliers

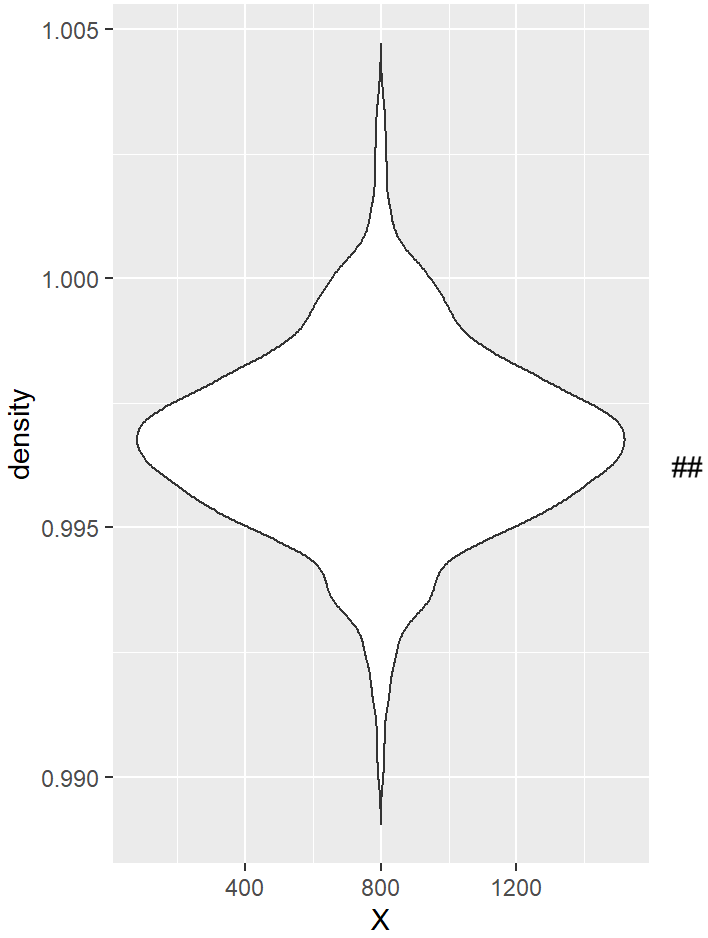




##

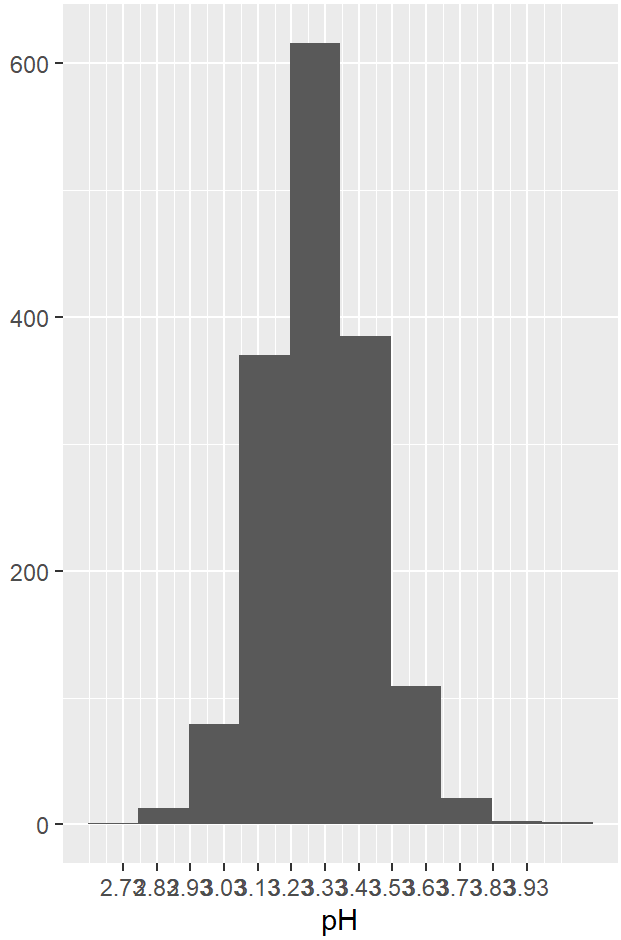
discovery - total.sulfer.dioxide skewed right with outliers

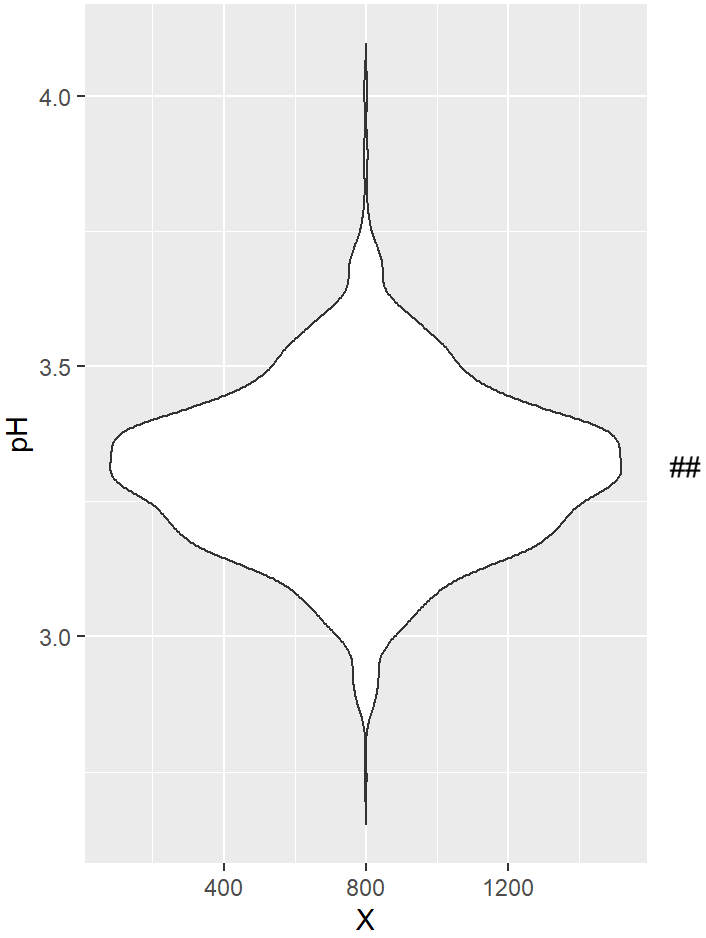




##

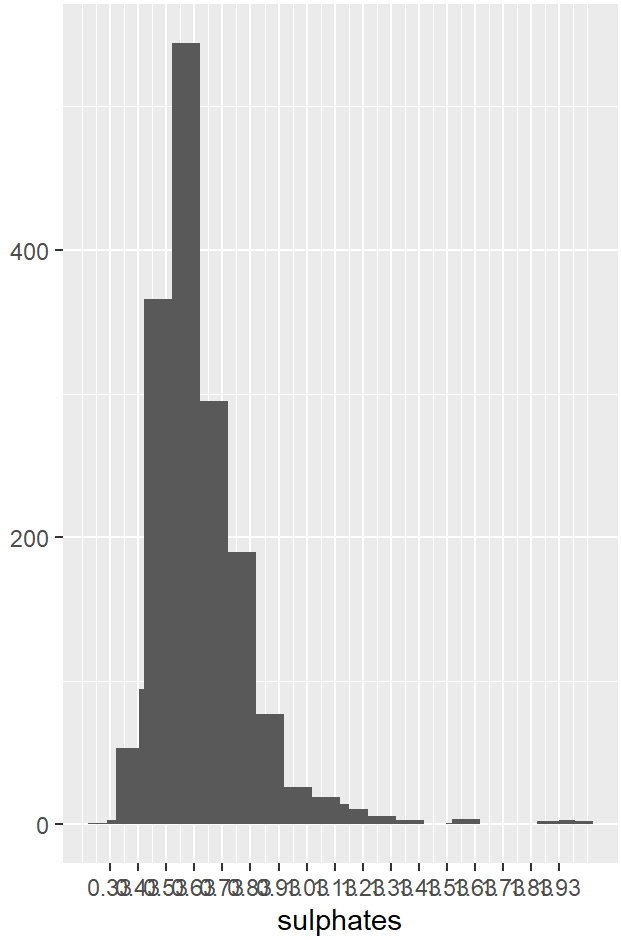
discovery - fairly normal distribution for density

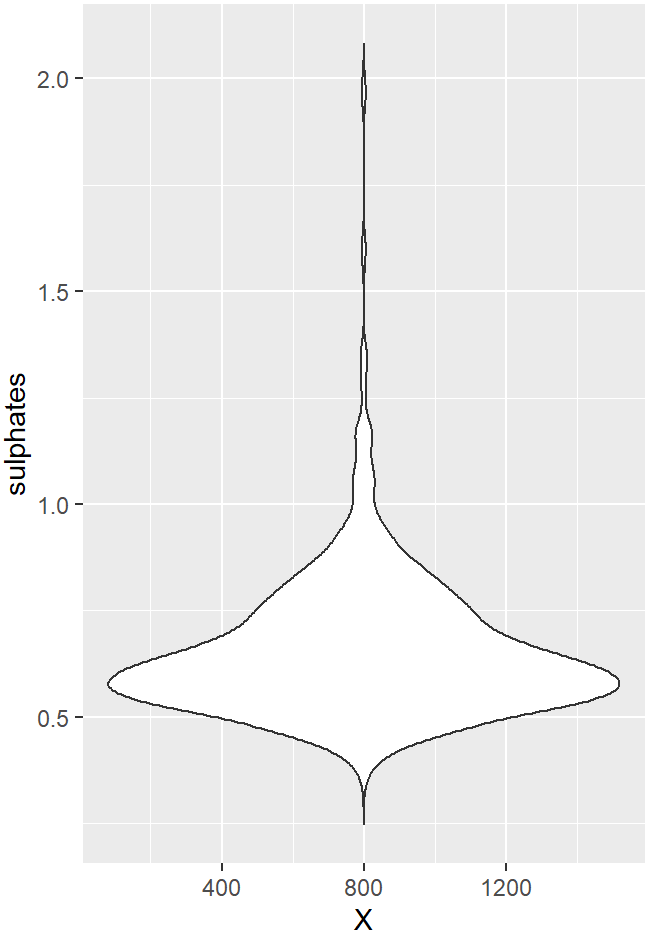




##

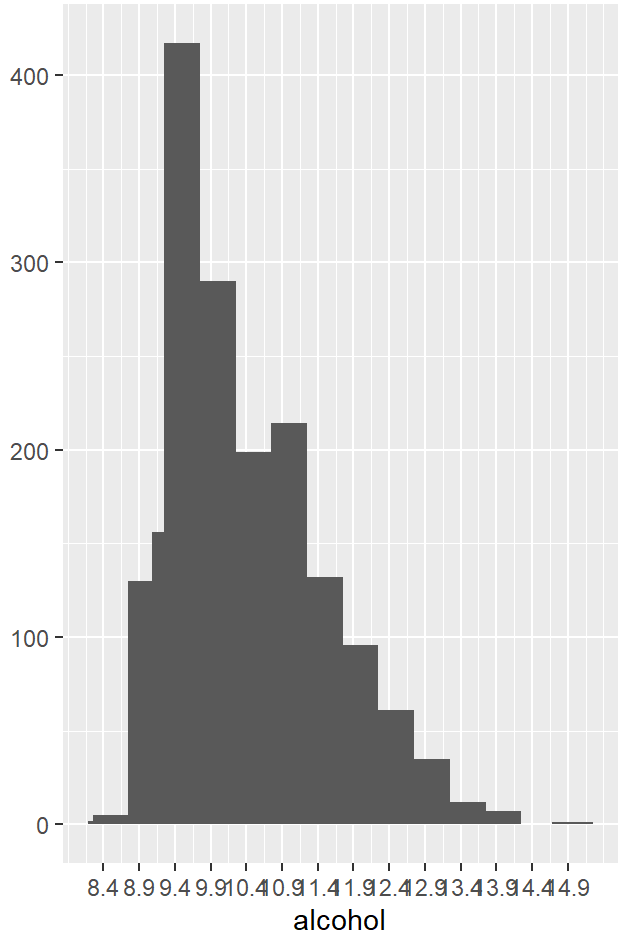
discovery - pH has a normal distribution

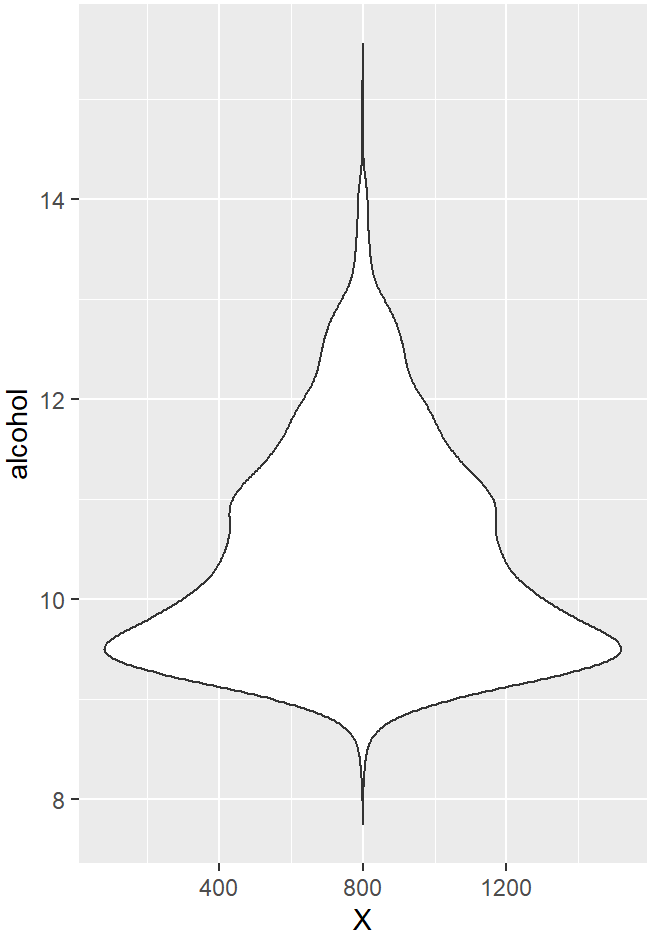

400
200
0
sulphates


2.0
1.5
1.0
0.5
sulphates
400
800
1200
X

discovery - sulphates skewed right with outliers

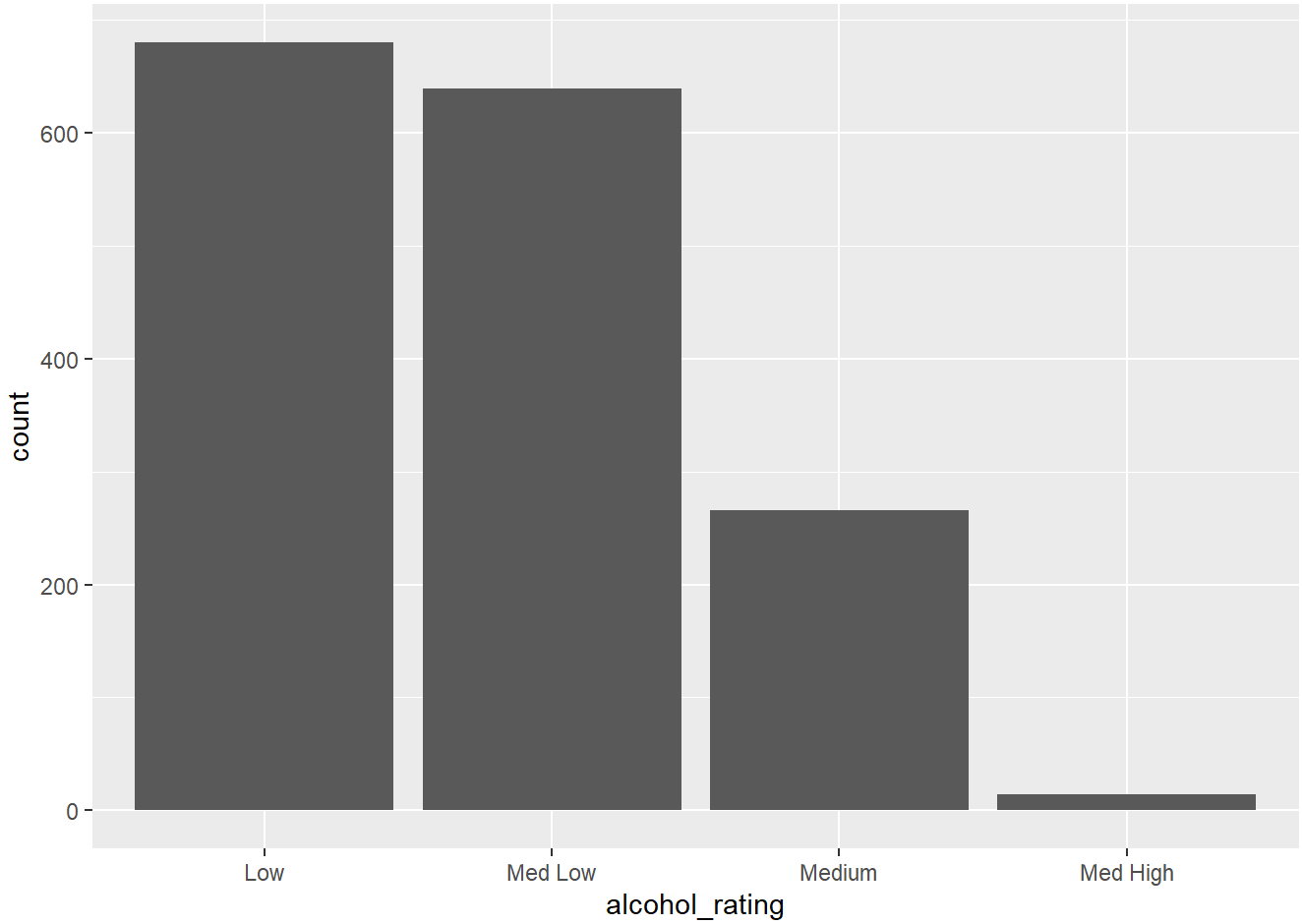
fairly normal distribution between .33 and .93





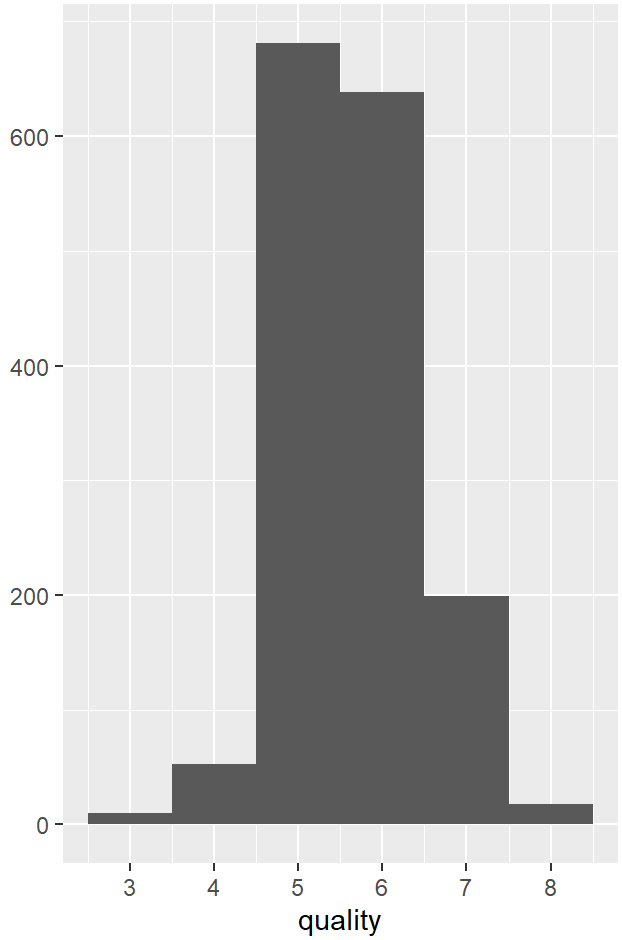
##

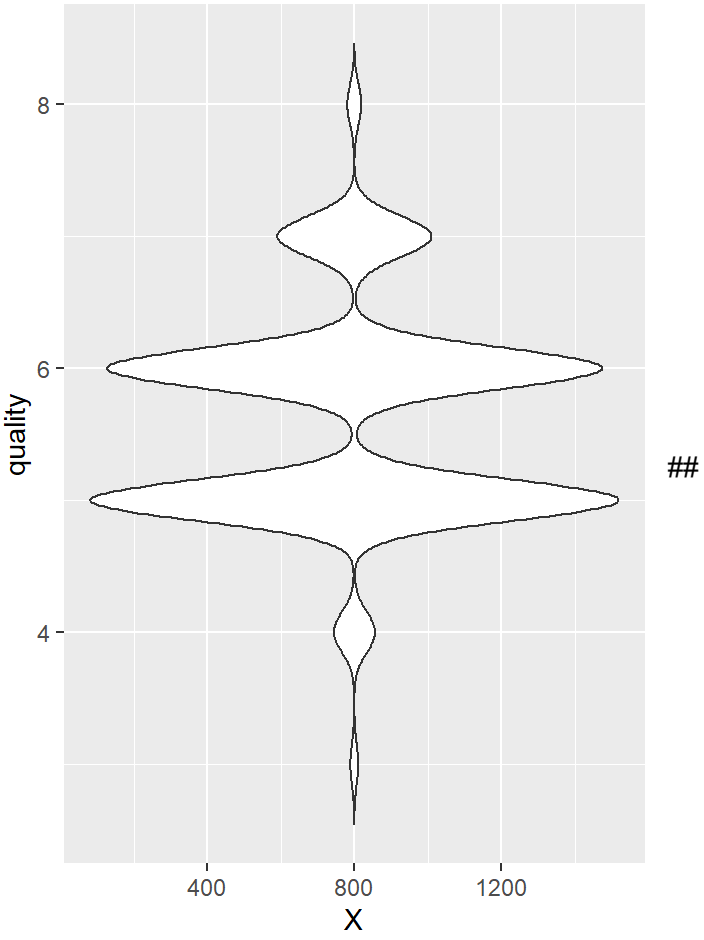
discovery - alcohol us skewed right with outliers ## potentially outliers to right and left



##

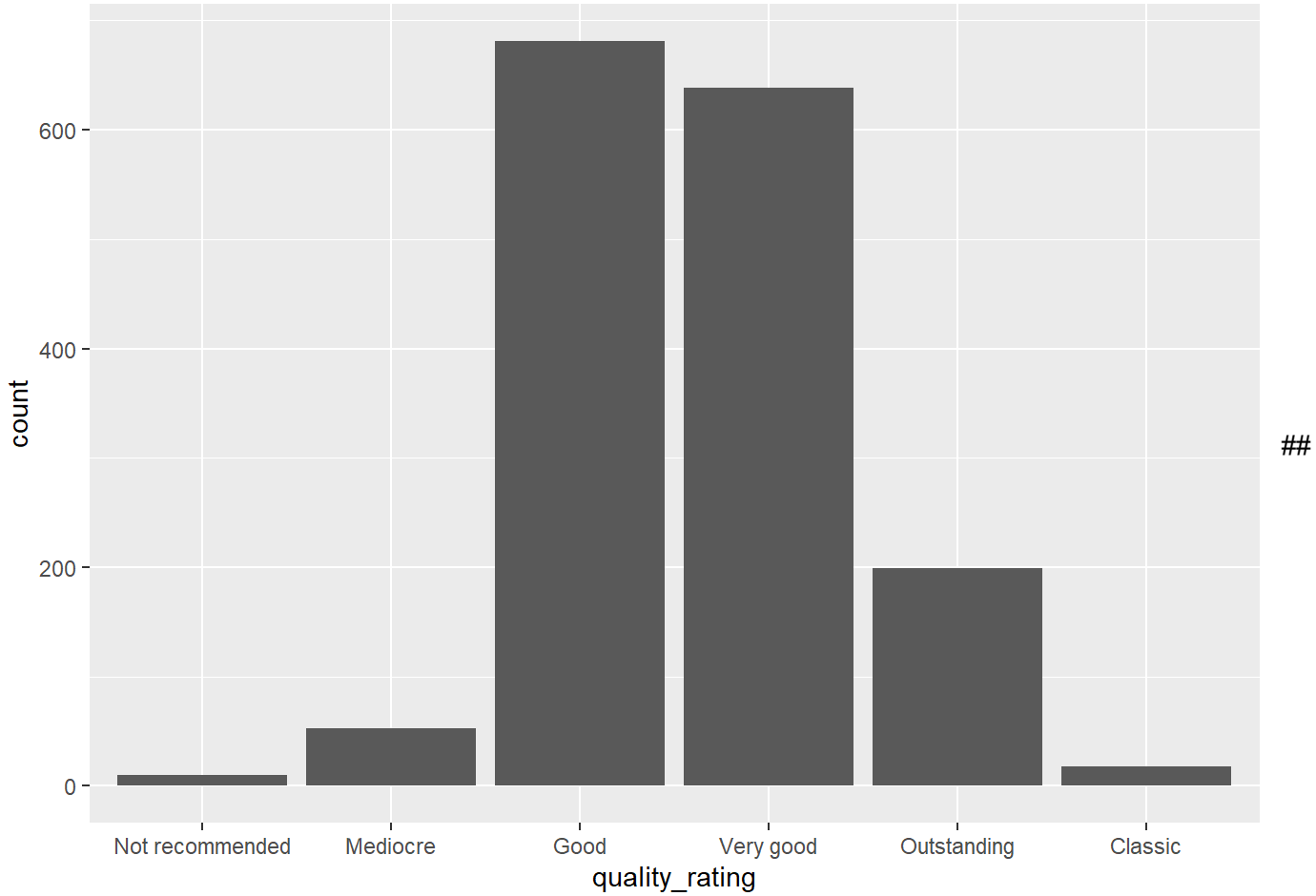
The data set is primarly low to med low alcohol rating





##

discovery - quality has fairly normal distribution I will need to create these levels as factors as this will be my category to compare features to.



The data set is mostly Good and Very Good Wines

Univariate Analysis

What is the structure of your dataset?

After reviewing all of the variables it became apparent that it would quality may be predicted by the other variables in the data set.

After doing some research online I found that navigating the relationships of acidity, Chlorides, sulfates and pH is fairly complex balancing act. Most of these factors are right skew and while I would be tempted to remove outliers from this I feel I would need to research more on how all of the variations of acidity, chlorides, sulfates and pH play a factor in quality. While I may be able to find positive or negative corrleations I would not feel confident in reporting on those knowing that other factors may balances those out ie lurking variables.

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

Because of the complexity mentioned above, my initial focus will be on Sugars and alcohol and how they play a part in Quality rating.

Essentially does sweeter or more boozier wines tend to have a higher quality in this data set. Becuase most wines are limited to Good and Very Good this could pose a challenge.

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest?

As I have time I would like to explore more of the chemistry and associations between acidity, chlorides, sulfates and pH. I have dones some research online on density but have not found much useful at this stage.

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

I created 3 new factor variables to label; quality, residual.sugar, and alcohol. These are; quality_rating, sugar_rating and alcohol_rating. I used cited reference to create the rating system.

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? Holestically looking at the distribution if they are skewed it is always to the right. Factors created show major dominance in just 1 to 2 areas rather than spread throughout. Sugar.residual == Off Dry alcohol == Low, Med Low Quality == Good, Very Good This tells me we are mostly looking at 1 or 2 varietals in the data set. With enough time a model could be created wherein you could predict what varietal the wine is based on the variables in this data set.

volatile.acidity:

Distribution: somewhat normal, skew right Majority population: between .25 and .75

citric.acid

Distribution: bimodal distribution, slight skew right Majority of population: Between .0 and .25 and .35 and .5 #### residual.sugar Distribution: Skewed Right Majority of population: Between 1 and 3

chlorides

Distribution: skewed right with outliers Majority of population: need more exploration detail #### free.sulfer.dioxide Distribution: Skewed Right Majority of population: Between 5 and 35

total.sulfer.dioxide

Distribution: Skewed Right Majority of population: Between 40 and 80

Density

Distribution: Normal Majority of population: Between .994 and 1.0

pH

Distribution: Normal Majority of population: Between 3.03 and 3.63 #### Sulphates Distribution: Skewed Right Majority of population: Between .33 and .93

Alcohol

Distribution: Skewed Right Majority of population: Between 8.8 and 11.9

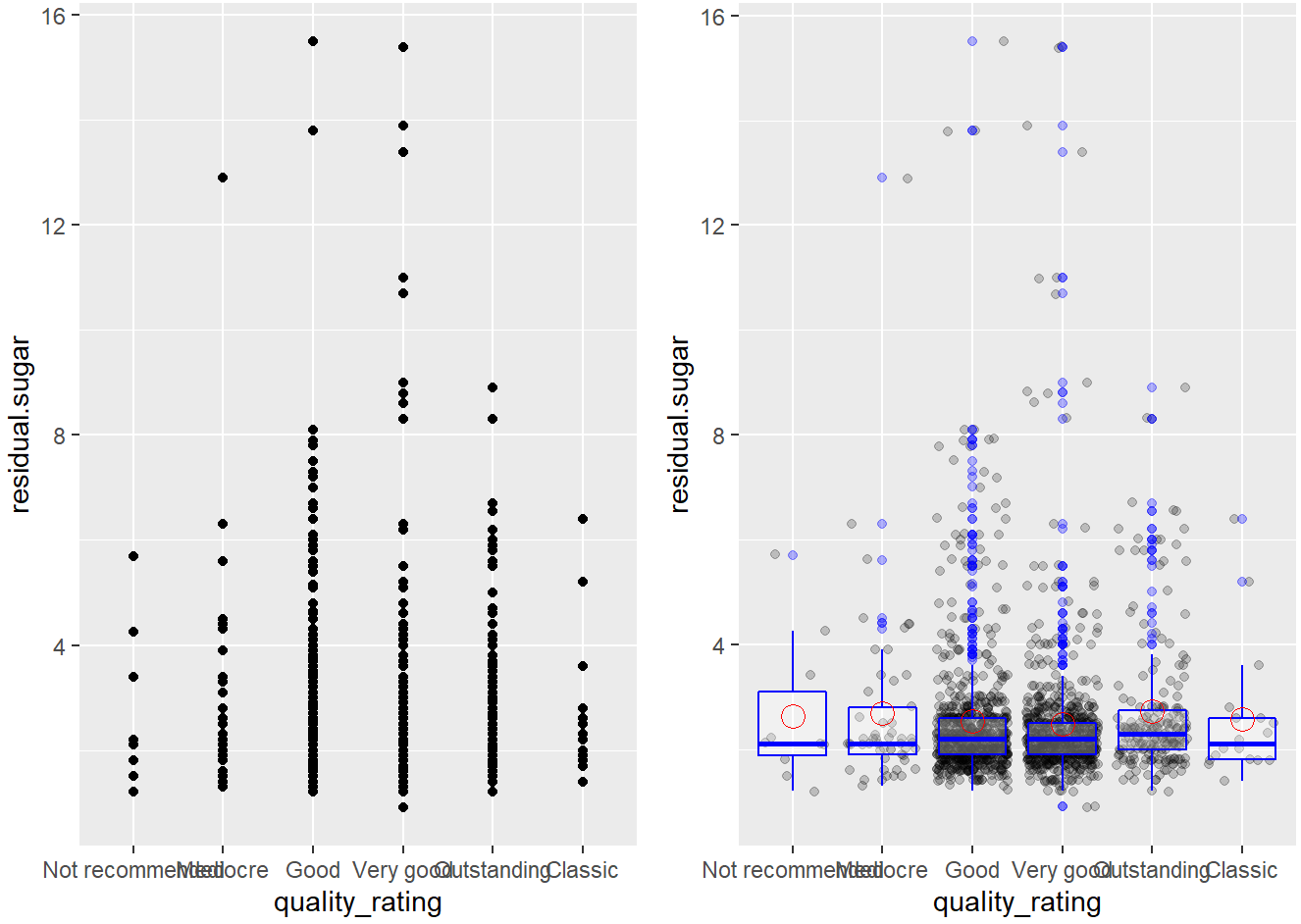
Quality

Distribution: categorical and focused primarily on 2 Majority of population: Category 5, 6

Bivariate Plots Section

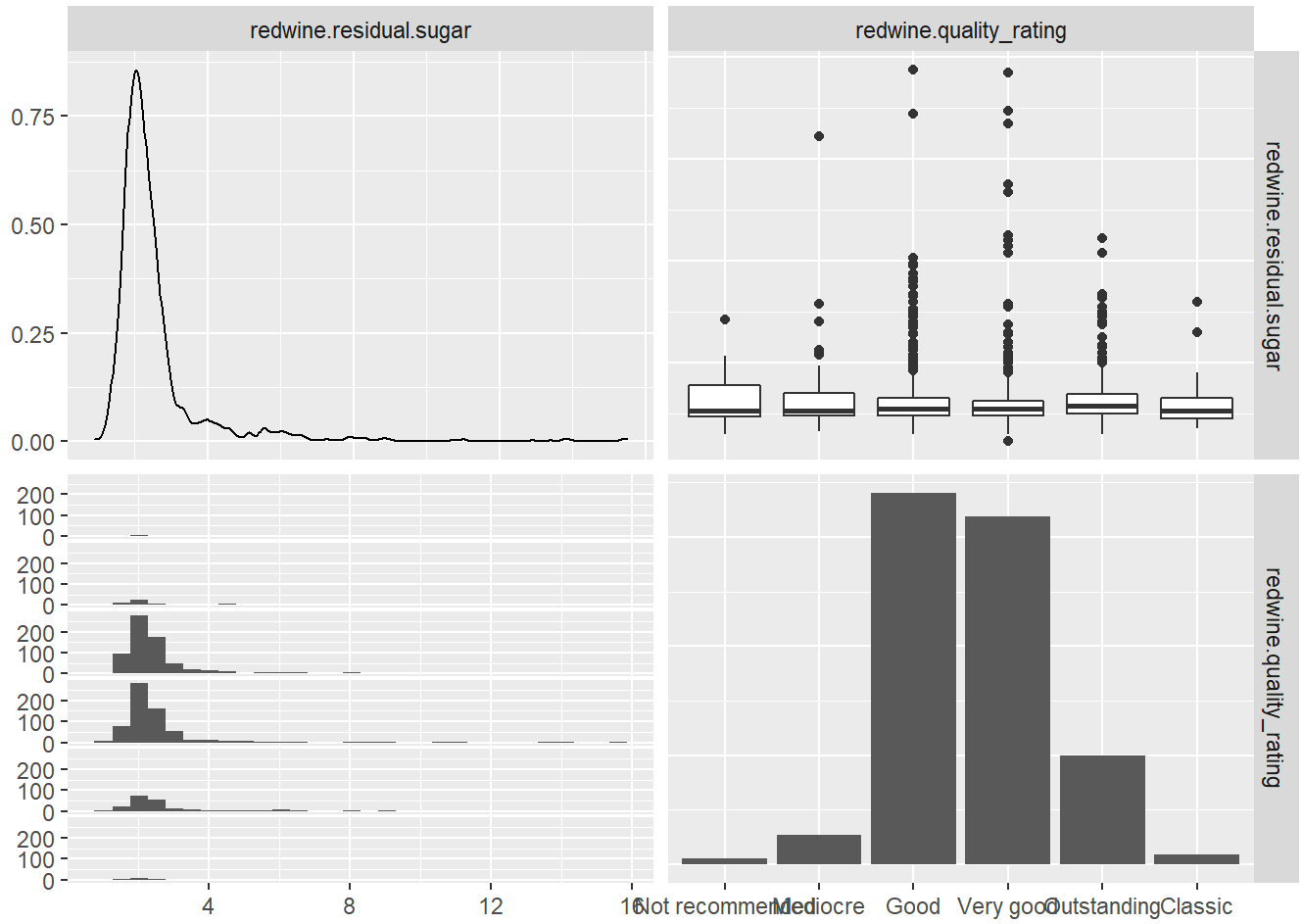
It would be interesting to see how sugar and alcohol relate to quality. I would also like to see how sugar relates to alcohol.

Residual.sugar to Quality_rating

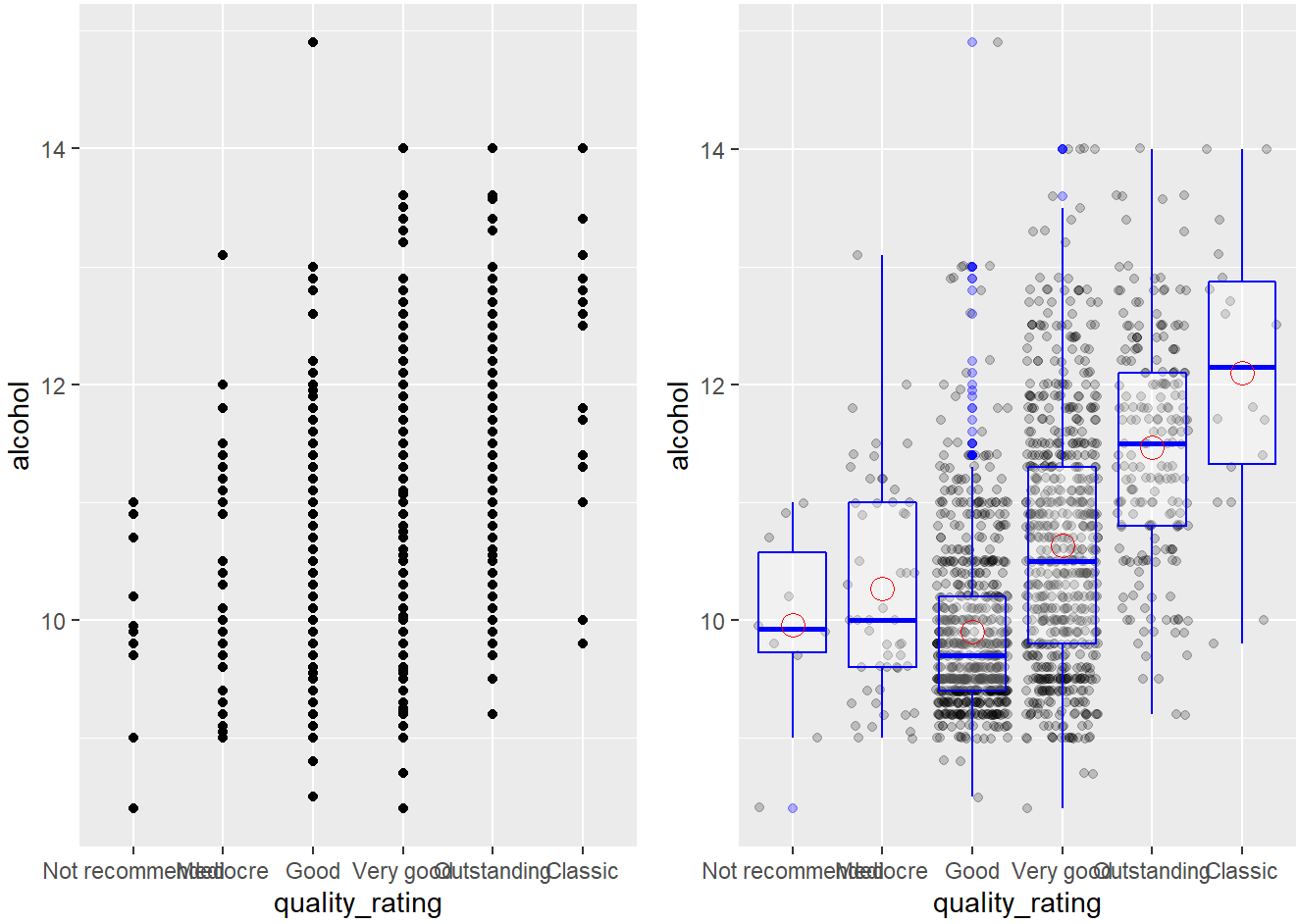


Closer look into relationships between quality and

alcohol.

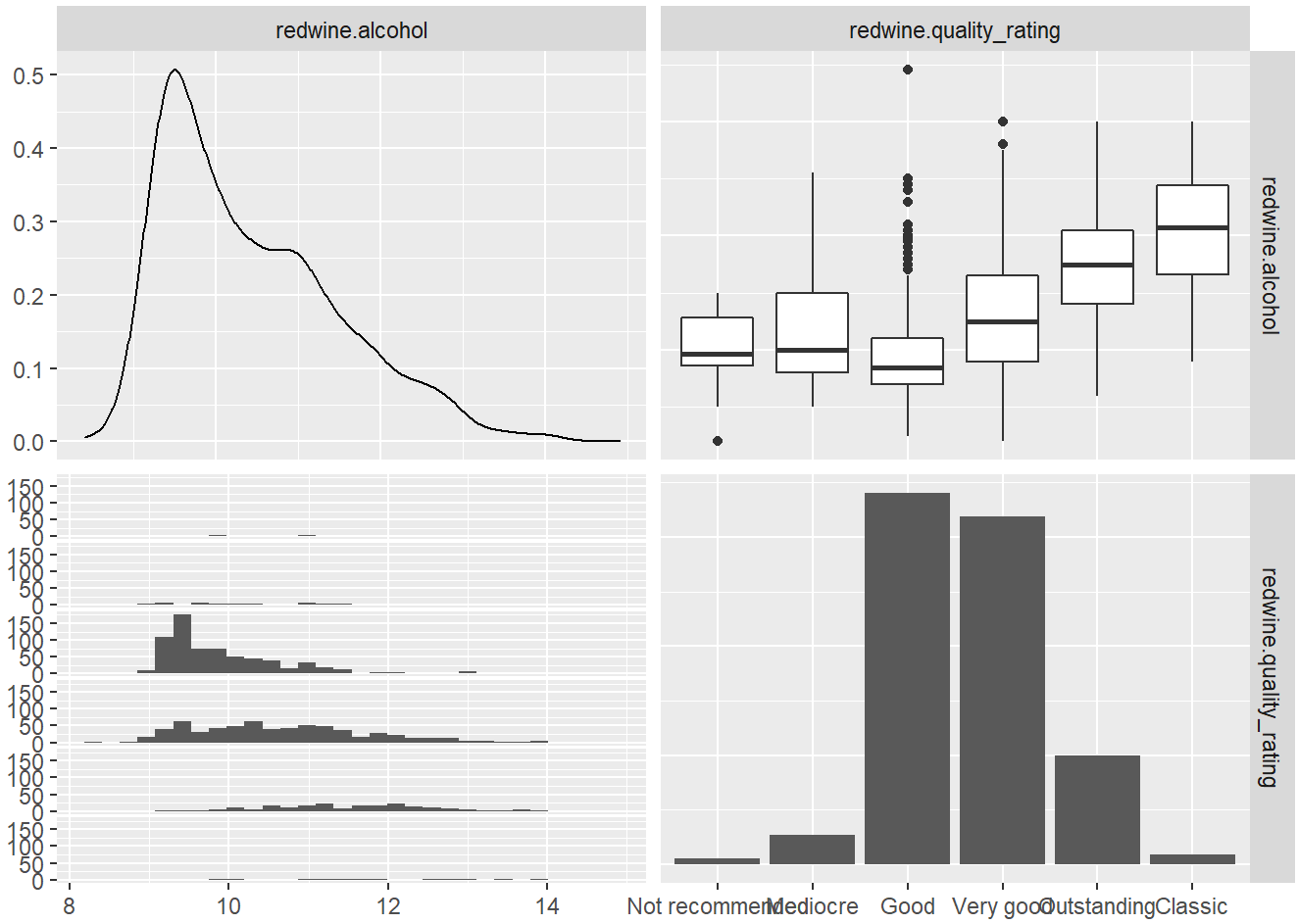


Good and Very Good Wines tend to have more sugar in them as opposed to; Not recommended, Mediocre, Outstanding and Classic. This still could be due to lack of observations in the data set. It does appear that there is some relationship between more sugar and quality.

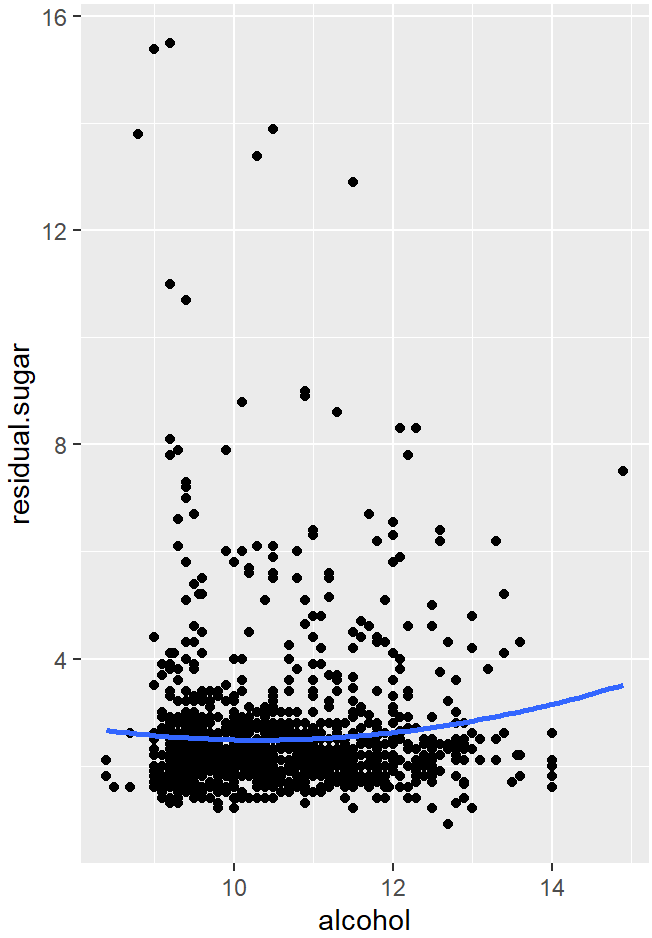


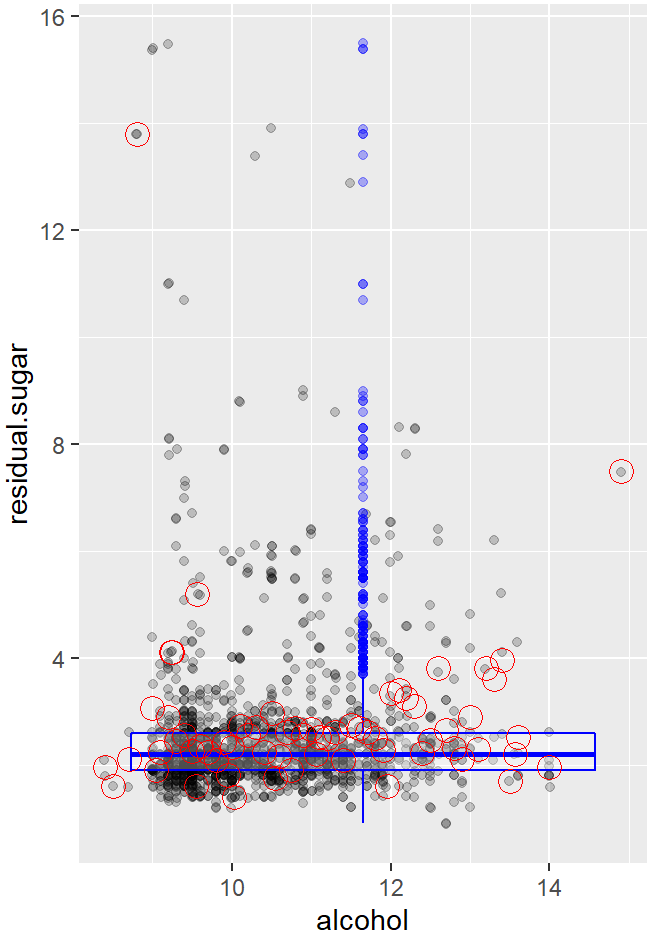
Closer look into relationships between quality and

alcohol.



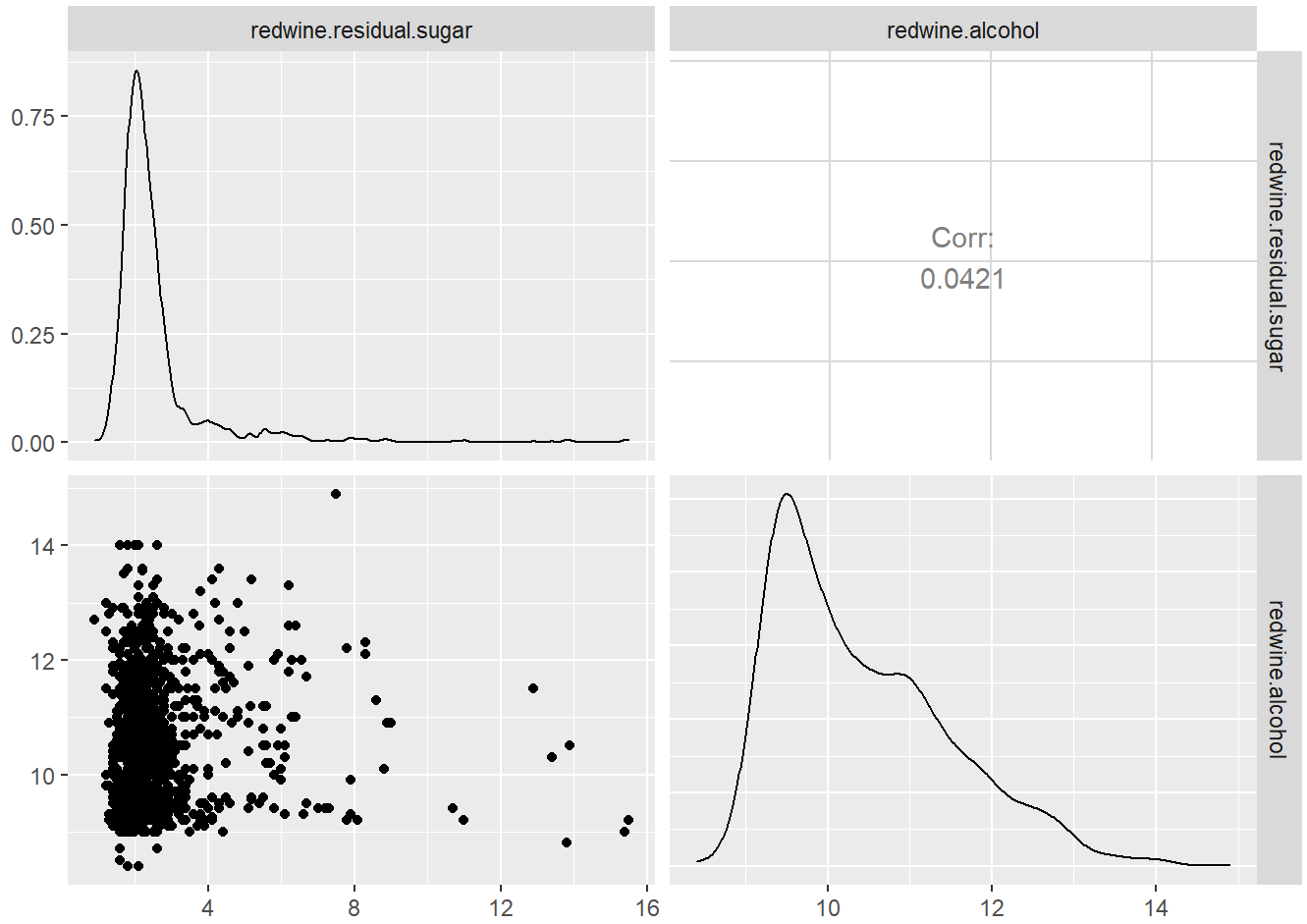
There appears to be a relationship between alcohol and quality of wine. The data here also looks to be more evenly distriubuted than residual.sugar.





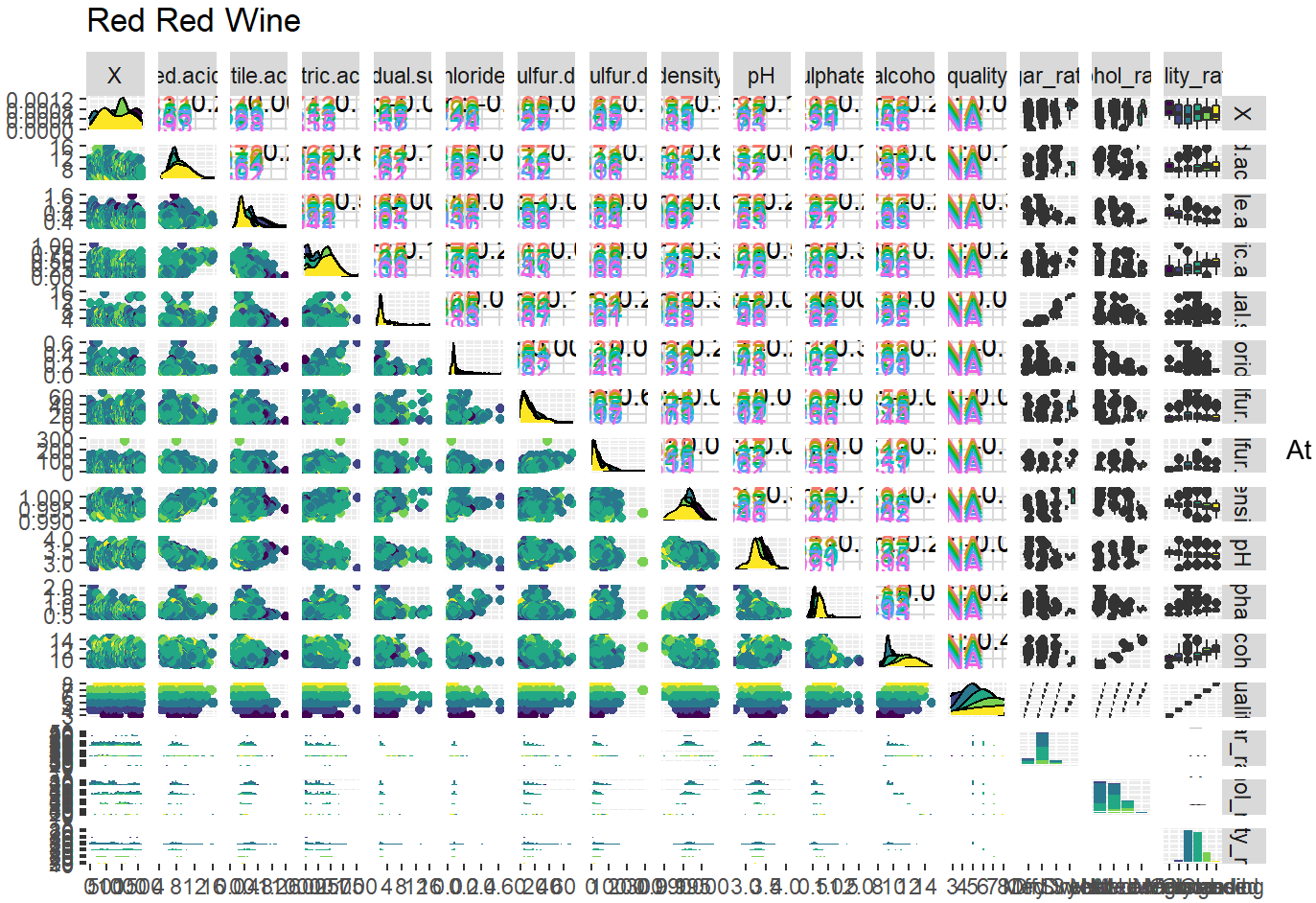
Closer look into relationships between residual.sugar,

alcohol, and quality.



In this dataset there is more sugars in lower alcohol wines. Supporting that alcohol may indeed play a factor in quality.

Other features review



first review it would look as though alcohol has the strongest relationship to quality. There are other features that have significant relationships that I believe are worth reviewing. While I still need a deeper business understanding of acidity, chlorides, sulfates, pH, and density I believe it is worth at least reviewing for now and compare to alcohols relationship to quality.

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
## log10.residual.sugar      log10.chlordies  free.sulfur.dioxide
##           0.02353331          -0.17613996          -0.05065606
## total.sulfur.dioxide              density                   pH
##          -0.18510029          -0.17491923          -0.05773139
##      log10.sulphates              alcohol
##           0.30864193           0.47616632
```

Alcohol still seems to have the highest relationship on quality, followed by volatile.acidity. # 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? My features of interest were residual, sugar, alcohol and quality_rating. My analysis shows that there is a lack of evenly distributed variety of observations which I believe is hindering some of the analysis. As for now,

Alchol has a stronger correlation to quality than sugar and while there may be some relationship to residual.sugar and alcohol it is low with a correlation score of .042

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

Relationship features of interest or significance are defined here as having a correlation score of -+ .63 and higher/lower respective to stronger r scores. summary(redwine) fixed.acidity to citric.acid rscore .67 fixed.acidity to density rscore .67 fixed.acidity to pH -.68

The top 3 correlations to Quality are alcohol to quality rscore .48 volatile.acidity to quality rscore .39 sulphate to quality rscore .25

feature of note sugar, I was reviweing this and found the correlation is not significant wtih an rscore of .014

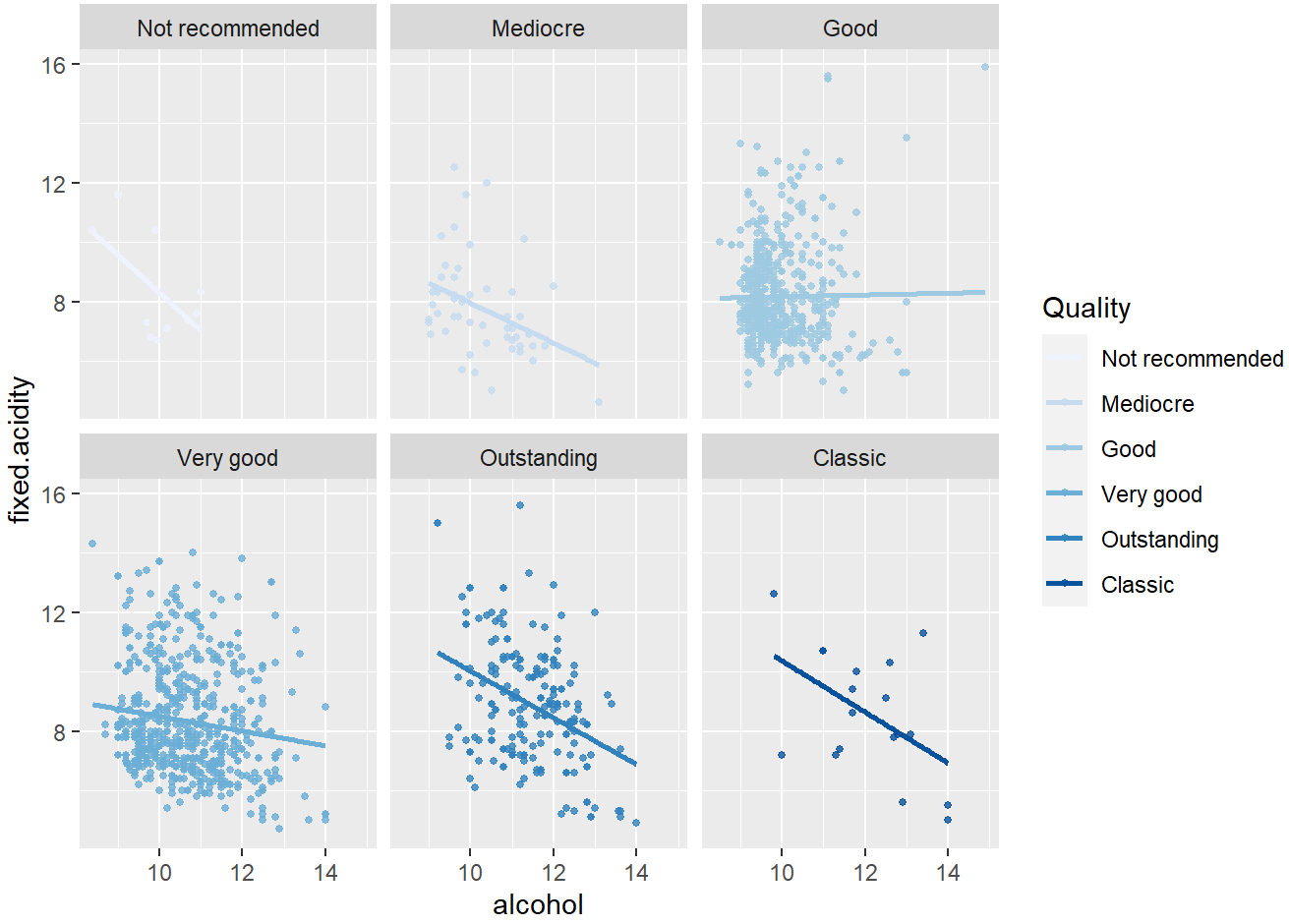
I would expect acidity to have a strong relationship with pH. I expect there is more to be learned about how, chlorides, acidity, sulphates and density work together to build or lower quality. I'm still theorizing that within even 1 variatal sugar and alcohol together play a strong role in quality.

What was the strongest relationship you found?

The strongest relationship in the dataset is pH. to fixed acidity. The strongest relationship to quality is alcohol

Multivariate Plots Section

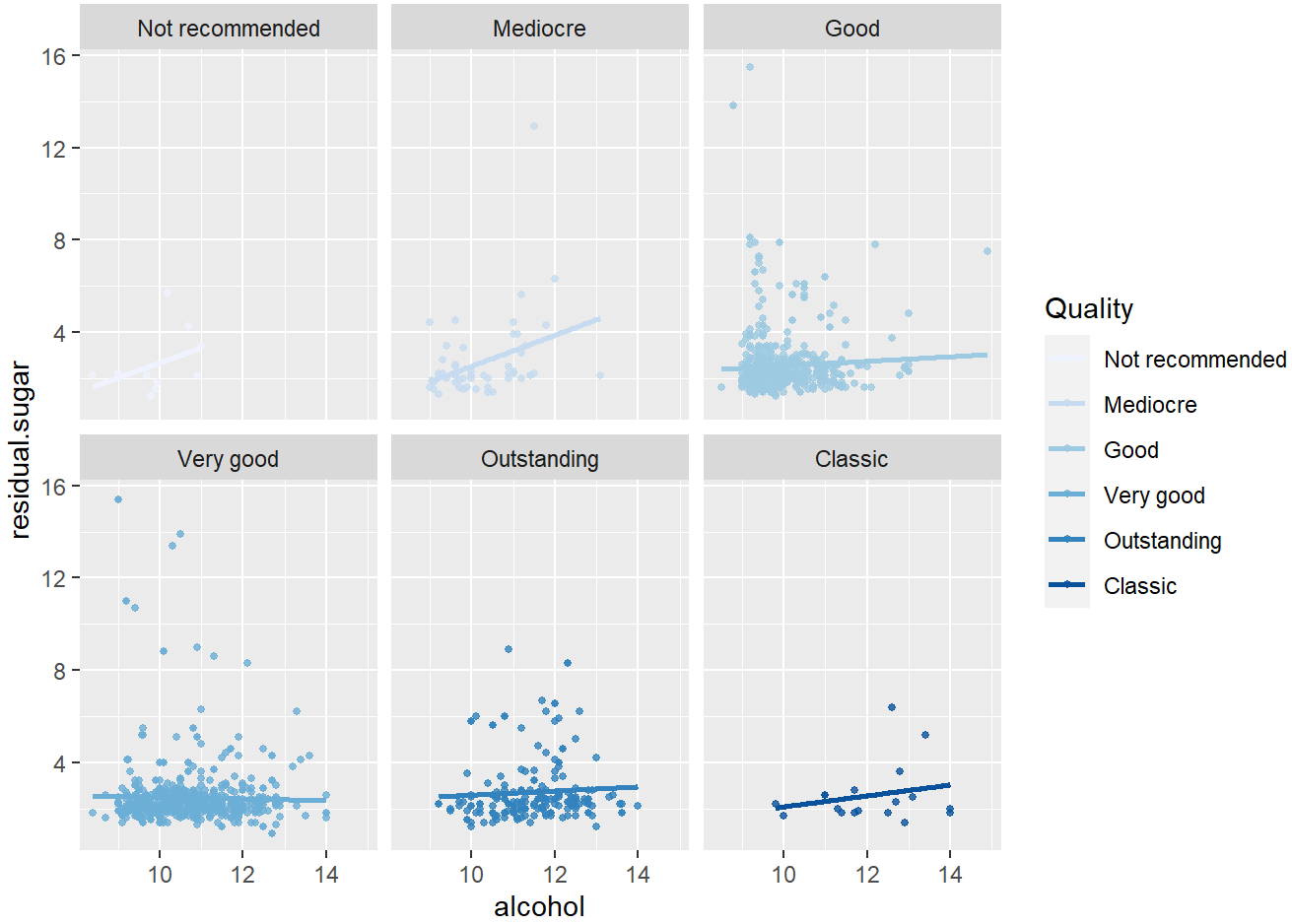
> Knowing that Alcohol is a major contributer to overall quality, I would like to see what combines with alcohol building a stronger relationship to quality. My initial hunch is we will see a relationship with citric.acidity and or sugar.



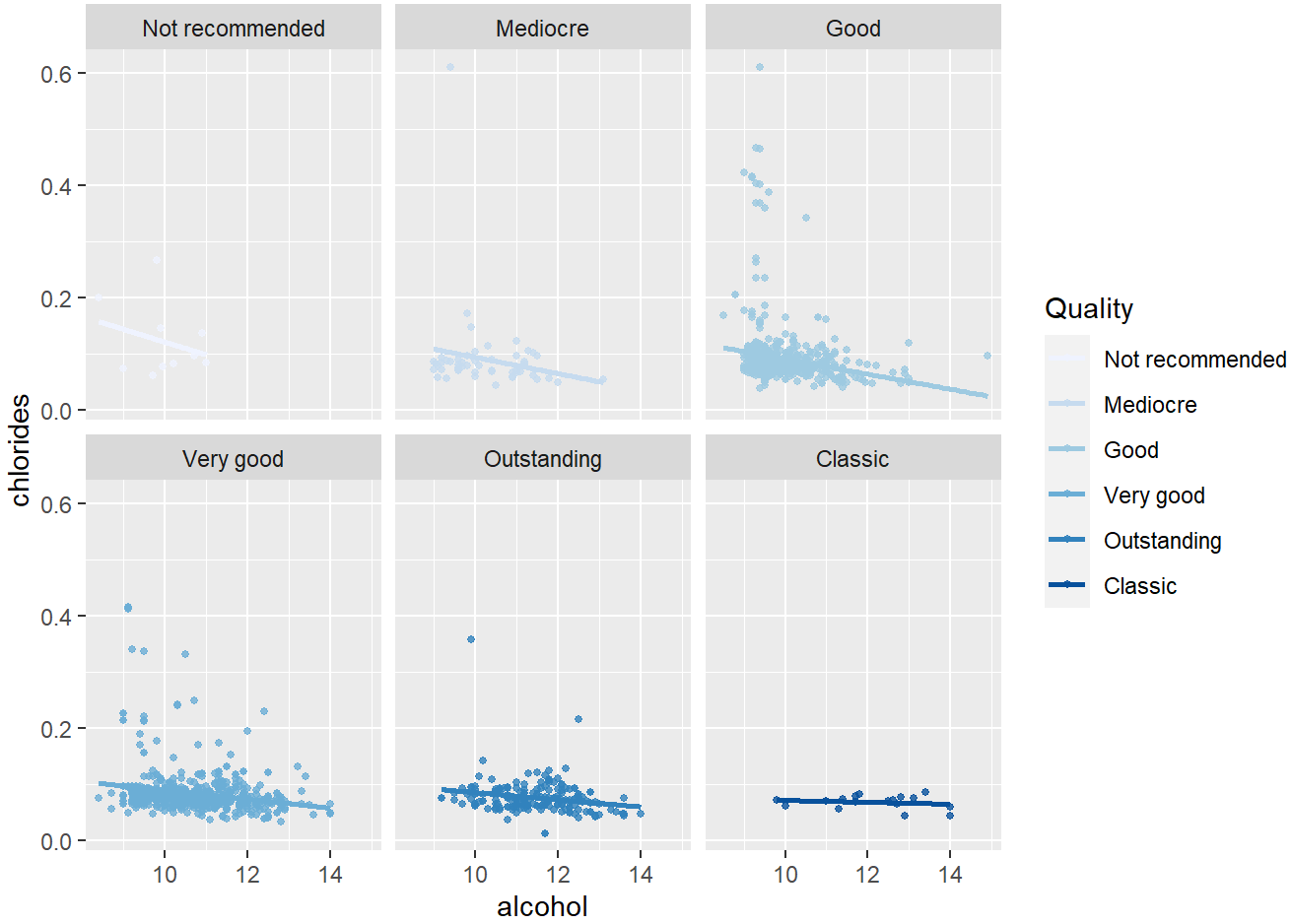
Fixed.acidity and alcohol has some slight positive coorelations in Good wines. Overall the relationship is negative.

volatile.acidity and alcohol do not have much of a correlation in quality of wines, there looks to be a positive correlation in Classic wines.

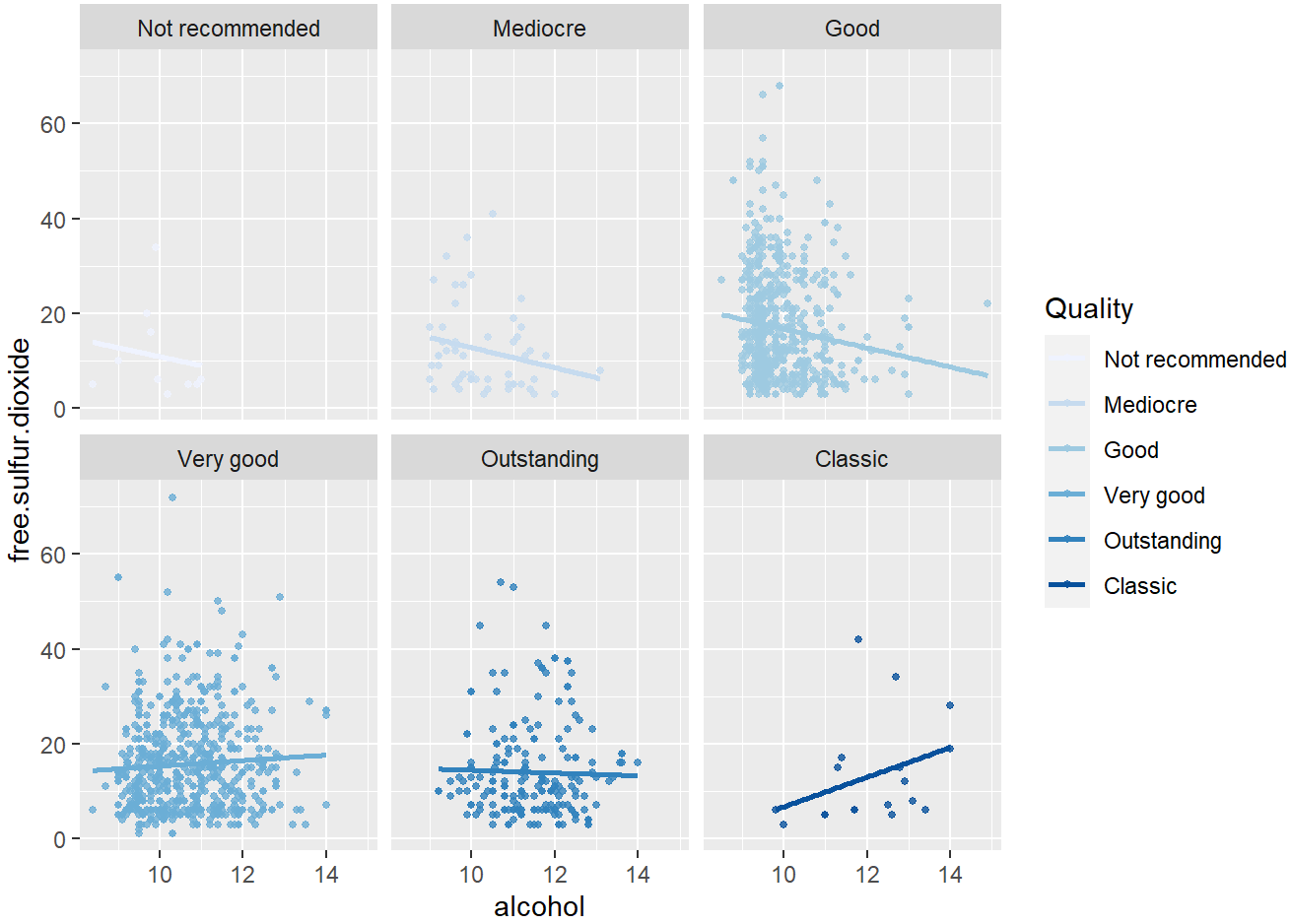
Citric.acid and alcohol has some slight positive coorelations in Very Good wines. Overall the relationship is negative. There is a strong negative Correlations With classic wines.



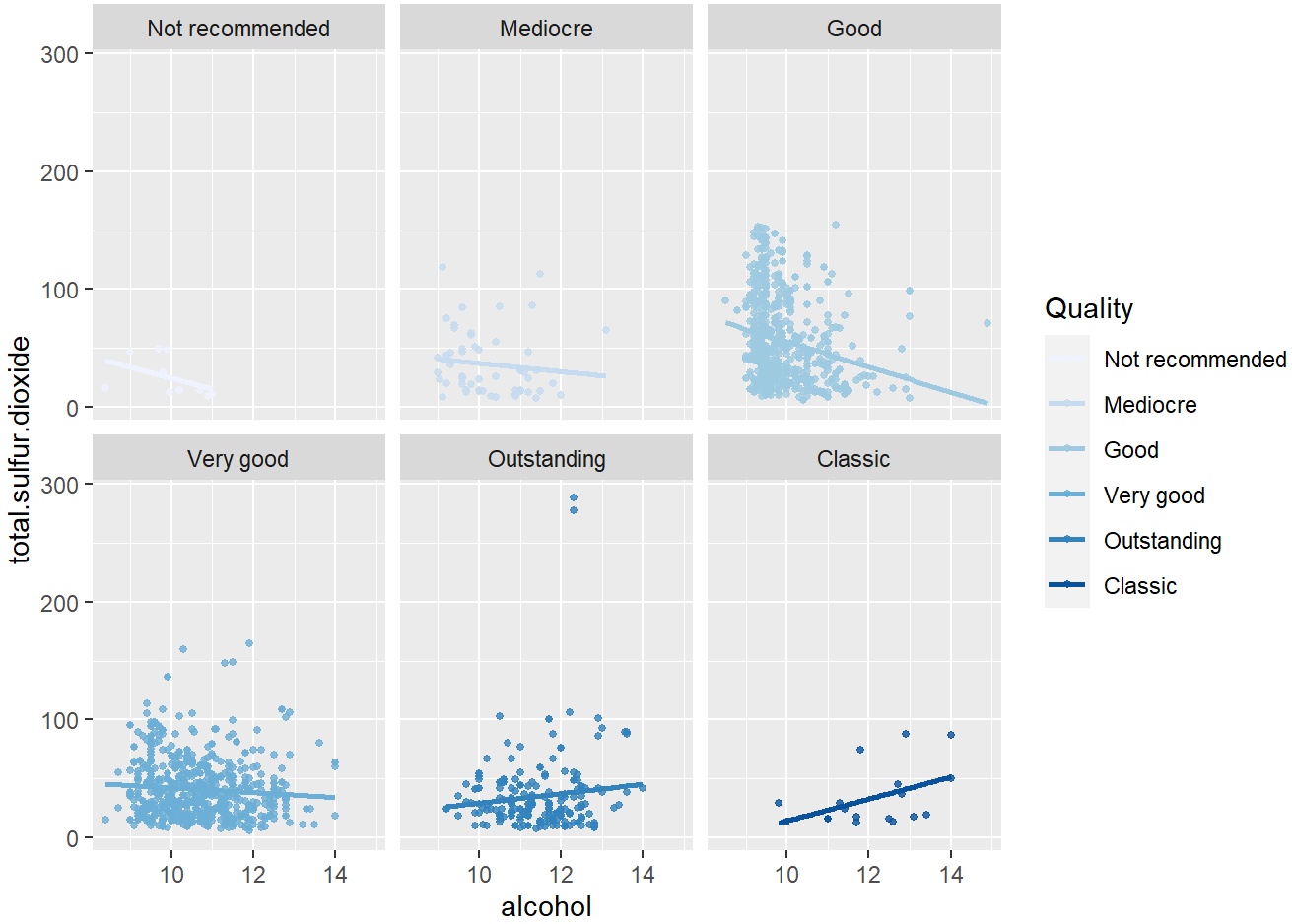
While there is a slight positive correlation with residual.sugar and Alcohol with Quality of wine it is strongest with Mediocre and Not recommended wines.



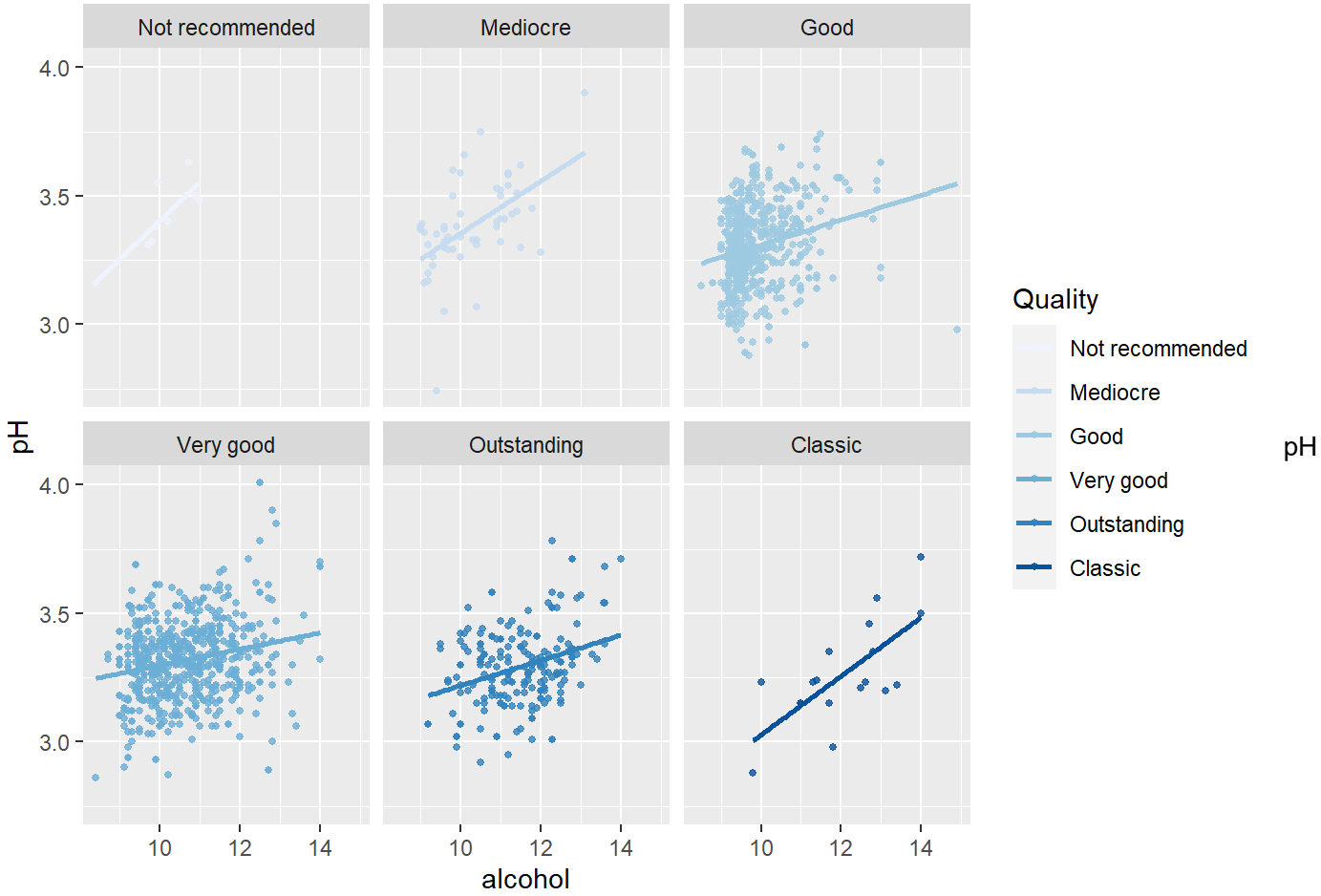
Chlorides and Alcohol overall have a negative correlation with quality of wine.



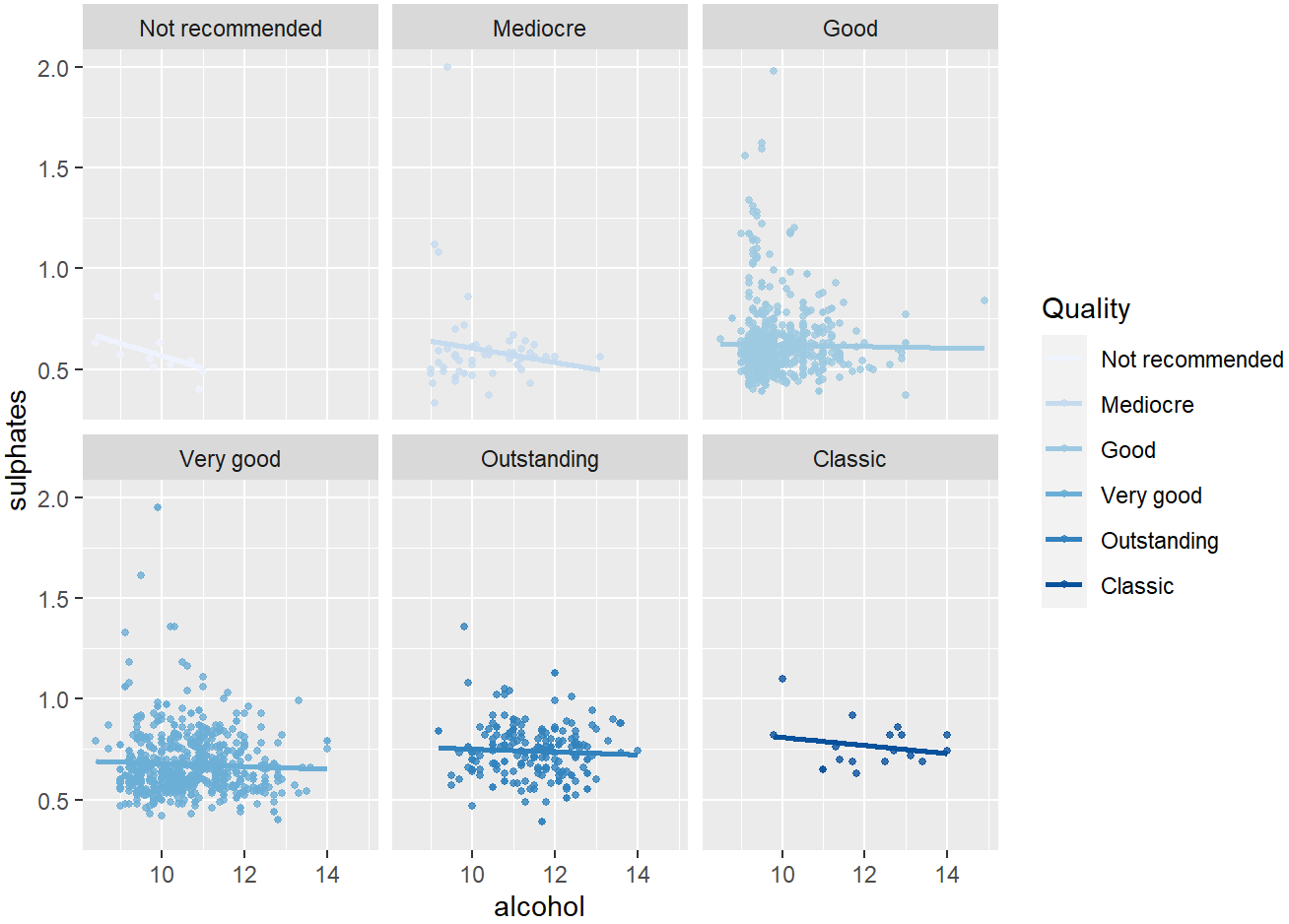
Here again, free.sulfur.dioxide and alcohol have overal negative correlations and then a positve one in classic quality rating.



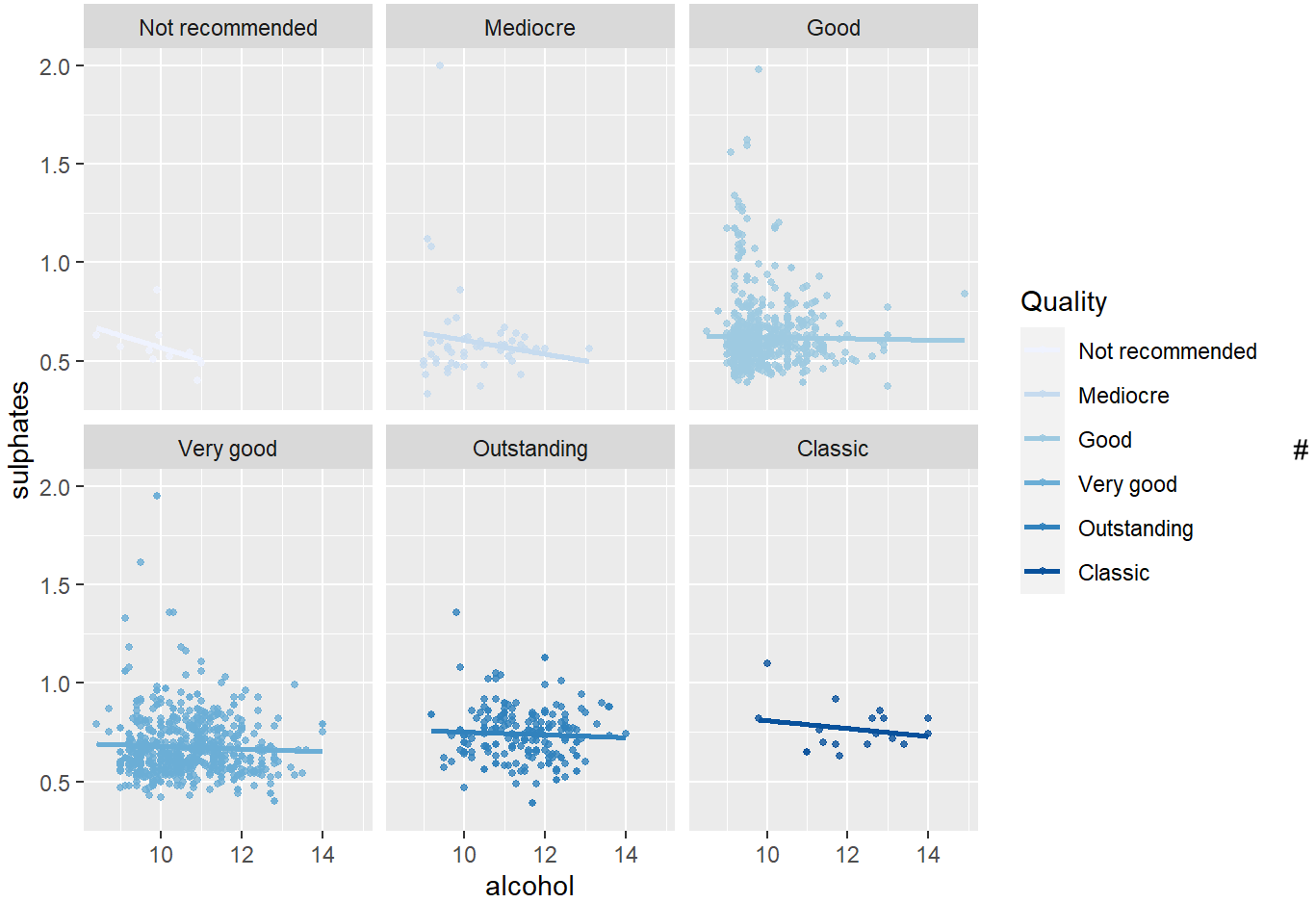
Total.sulfer.dioxide appears to have an overall correlation to quality of wine By rating you can observe the trend going from negative to positive as quality goes up. summary(redwine)



and alcohol have a positive correlation at every quality level and a very strong correlation at the classic quality level.



There does not appear to be a correlation between sulphates, alcohol and Quality.



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? Sugar and Alcohol were not strengthened. pH and Total.sulfer.Dioxide (TSD) are of interest as they show positive correlations. While TSD begins negative you can observe it becoming more positive through the ratings.

Were there any interesting or surprising interactions between features?

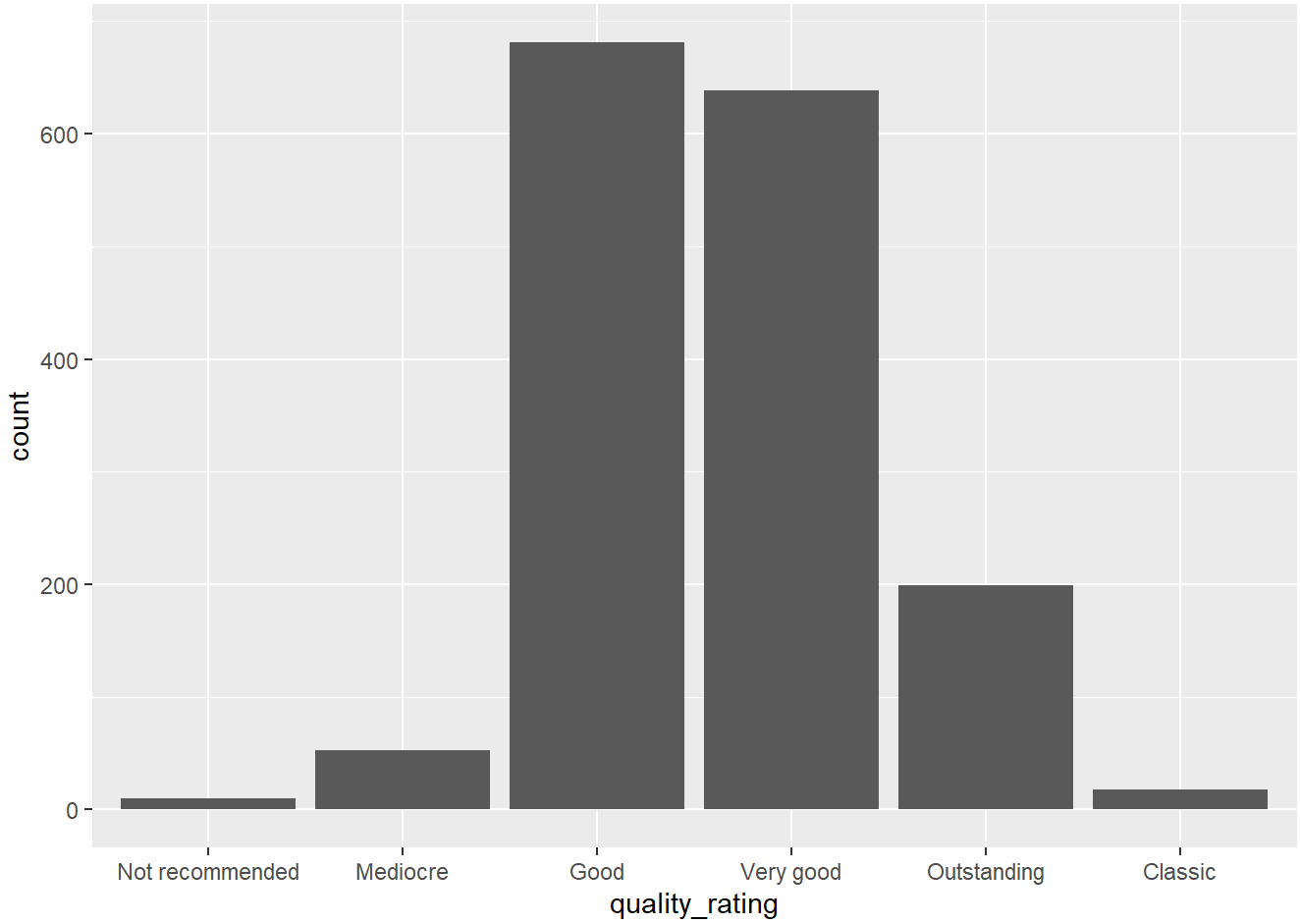
I found it interesting that what in some features there were positive correlations in lower rated wines but as you skipped to the highest rating those would then become positive. This lends me to believe there are combinations of features still be be explored.

Final Plots and Summary

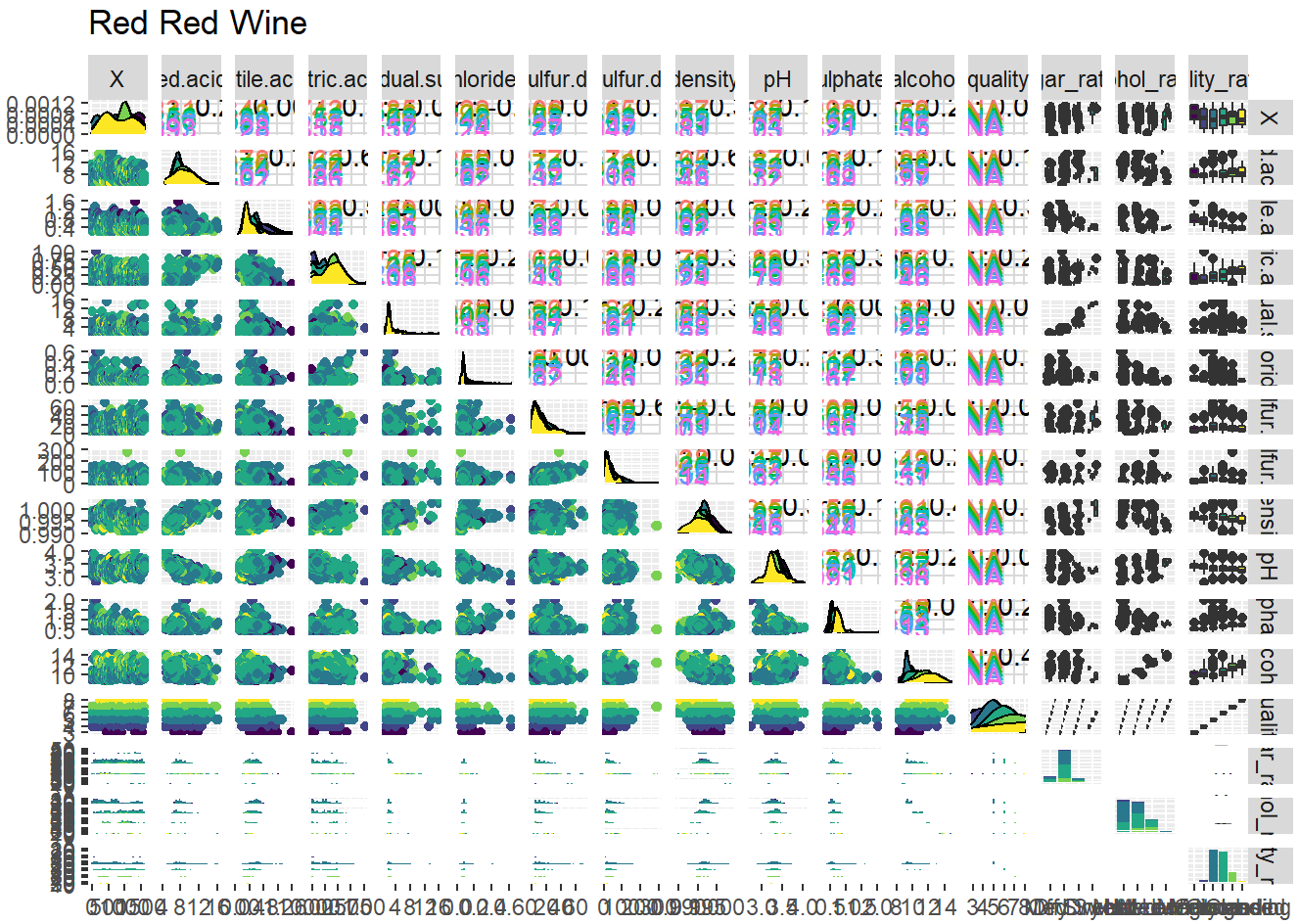
1. Alcohol and pH seems to produce better wines.
2. Residual.Sugar is correlated in wines that are Not recommended, Mediocre and Classic.

I wanted to stayed with what I knew about wine which was sugar and alcohol. While very simple I thought it was a good starting point.

Plot One



My initial exploration revealed that the data set is limited to mostly Good and Very Good Wines. This to me will influence finding correlations with Quality.

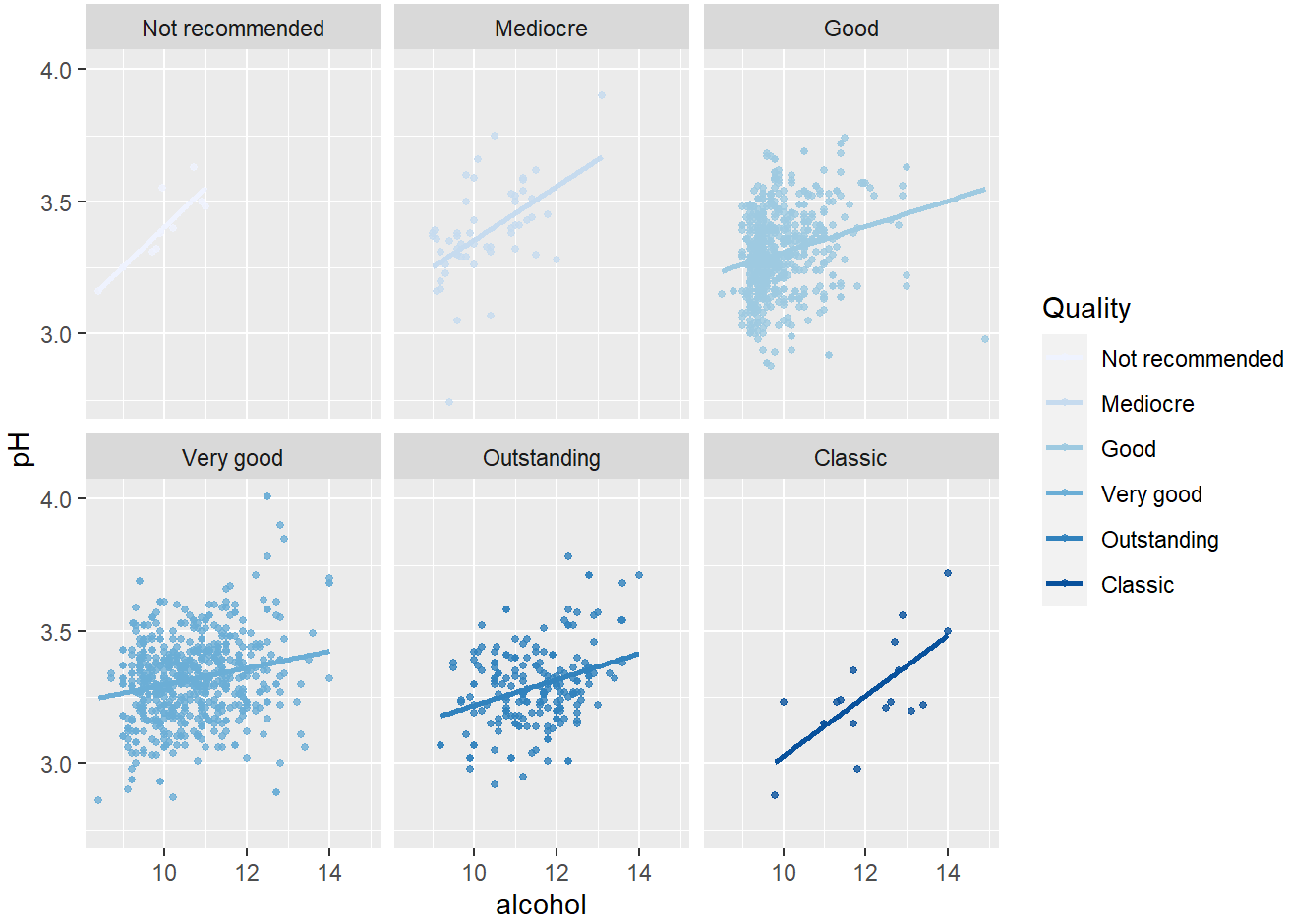


Description Two

Luckily I was able to find a correlation with alcohol but not so much with sugar. This led me to review all features.

Alcohol still seems to have the highest relationship on quality, followed by volatile.acidity. At this stage it is apparent that other features will need to be reviewed.

Plot Three



Description Three

While this has a positive Correlation that is not necessarily a good thing. After a little review I found that the Higher pH wines will taste flat and lack freshness. Bacteria also thrives at a higher pH. So why is there a positive Correlation with higher pH levels and quality of wine. As wine ages the pH level rises and older wines tend to have higher quality. There is a deeper science to this and on further research I believe taking a wine data set of 1 variatel all from the same year would yeild better overal analysis. ——

Reflections

This data set was focuse primarly on Good and Very Good wines. With the category of Quality you are drawn to look for relationships in what would drive quality. However when your population is relegated to 2 main categories a lot more cleaning would need to be done.

To build better analysis I would want to normalize each quality level by indexing wines based on ratio.

Upon further reflection I would need to spend more time with my business sponsor to better understand how pH, Acidity, Chlorides and sulfates work together. I believe once I understood this more I would have more confidence in isolating outliers and removing appropriate ones rather than chopping of a set standard quartile.

I would like to have a dataset that incorporated age, price, variatel. This would lend itself to a great classification project.