Modeling Assigment #4

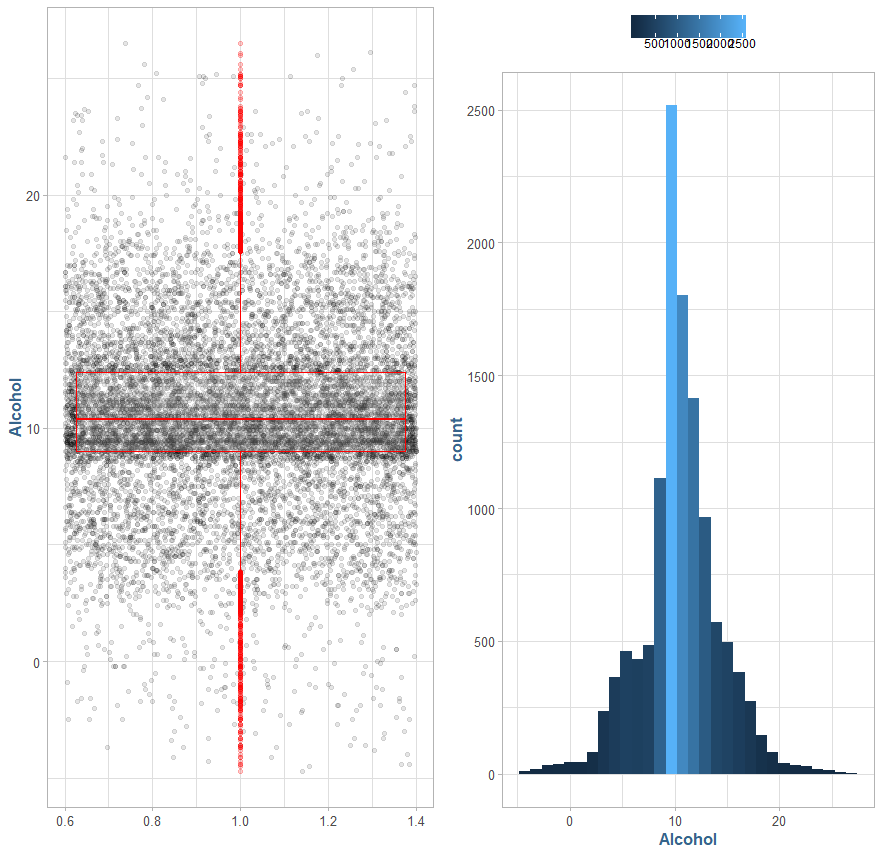
Brandon Moretz

### Introduction

In this lab we are going the explore a data set consisting of various attributes of approximately twelve thousand commercially available wines in order to predict if they will be sold, if so, how many cases will be sold, and the number of stars an expert would give this wine using a Vivino style rating system. The ultimate purpose of this data analysis and resulting model is to provide a large wine distributor client our recommendations on what kind of wines will be ordered and in what amounts so that we can be operationally prepared to manage the supply chain and resulting required logistics. Additionally, we will attempt to quantity what makes a wine ‘good’ based upon its chemical composition and characteristics.

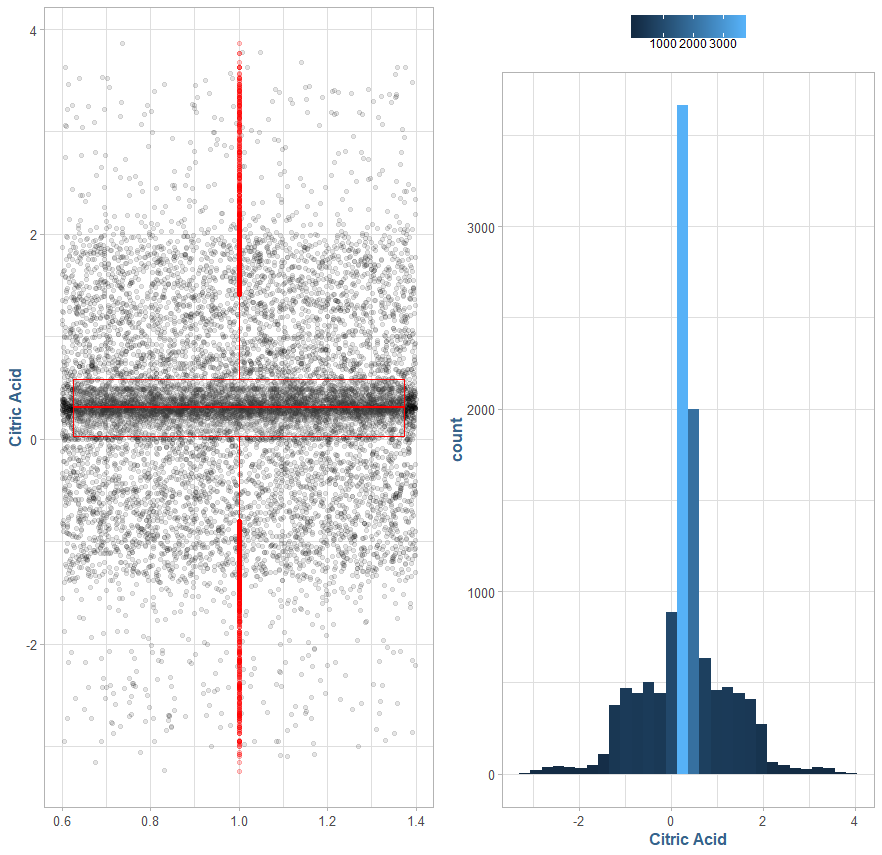
### Exploratory data analysis

The first step in this analysis project is to take a clean sweep to the data and look for any bad encodings or invalid values. Inspection of the continuous variables leads to some interesting questions on exactly what scales of representation these measurements are in. For example, the Alcohol variable appears to be in a measurement of percent volume, as we see the values lie in the range of what we would expect the alcohol content to be by volume (ABV).

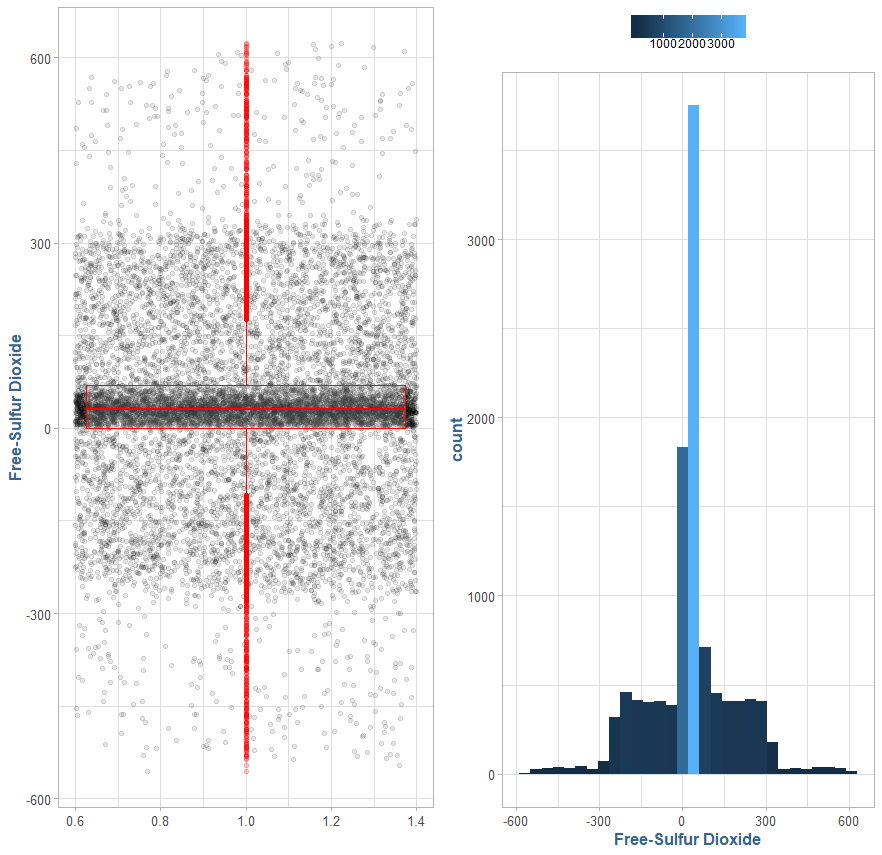


Given this measurement, we are going to remove the 771 bottles with missing or negative values.

Other continuous variables in this data set appear to be on a normalized scale (approximately -4, 4), given that they are measurements of the contents of various chemicals and the distributions are centered near zero with equal parts above and below the mean (in the positive and negative direction). The Citric Acid, Sulphates and Volatile Acidity variables for example has this quality when we look at it closely. This is the only logical explanation I can see for over half of the values being in the negative range for so many variables that represent measurements of a quantity.



Some of the other variables have similar properties (Free-Sulfur Dioxide, Chlorides and Residual Sugar), however, they appear to be on slightly different scales. We will standardize these variables and append them to our data set; this process won’t change the actual properties of the values underlying distribution, it will simply help later with interpretations having all the variables on a similar scale and value range.



The label appeal variable has the scale -2 to 2, which we will adjust by +3 so that it is on a “normal” 1-5 scale.

### Research

### Conclusion

### APPENDIX

