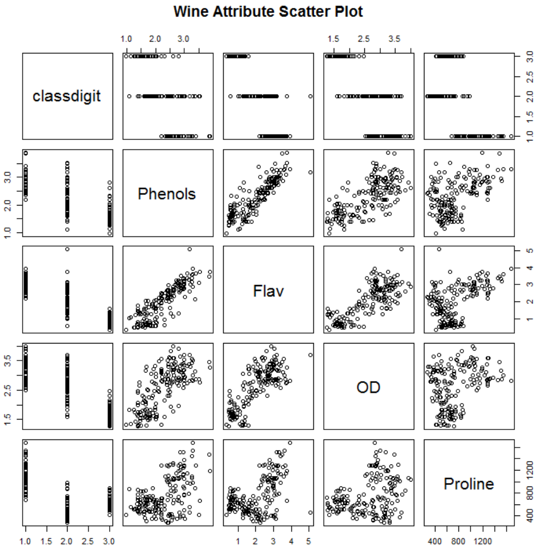
Melissa Welle
Predict 412
Statistical Graphics in R

The following paper is a detailed exploration of the wine dataset. Our overarching goal is to use attributes of the wine to predict which class they belong to: Barbera, Barolo, or Grignolino. The following are the variables of interest:

Independent: Alcohol,MalicAcid,Ash,AlcAsh,Mg,Phenols,Flav,NonFlavPhenols,Proa,Color,Hue,OD,Proline

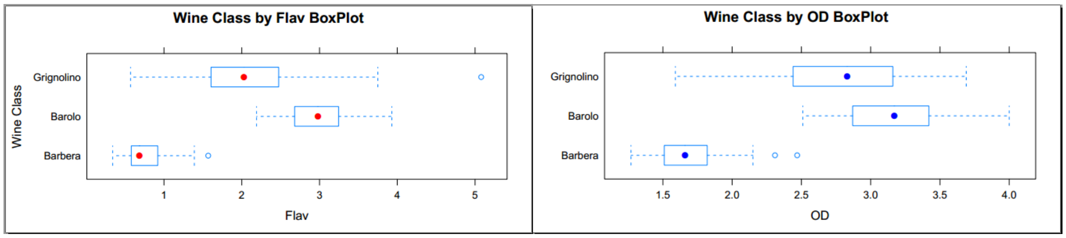
Dependent: classdigit/class

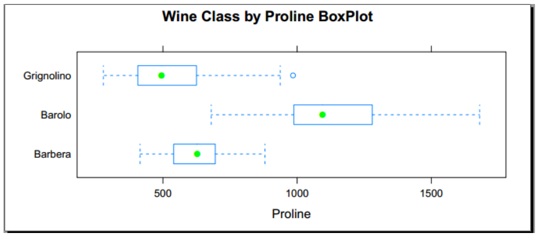
The first step is to look at the pearson correlations and/or scatterplots between all of the variables to determine which variables are more predictive in nature of the wine class. Using the cor function a .6 as the threshold for interesting variables we are left with Phenols, Flav, Hue, OD, and Proline. The Following scatter plot diagram produced by the pairs function will provide more visual:

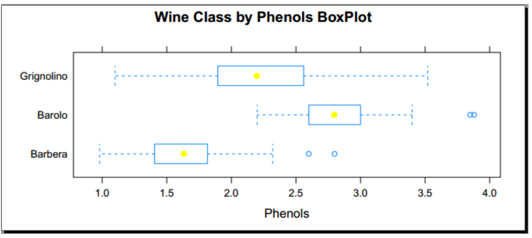


The distinct groupings for the class digit display the correlation described above. However, we see strong correlations with our predictor variables as well which is suggesting that a model will suffer from multi-collinearity. Phenols in particular appear to have a strong relationship with Flav, OD, and even Proline. Next we will look at box plots (bwplots from the lattice package) for classdigit and each of the potential independent variables.

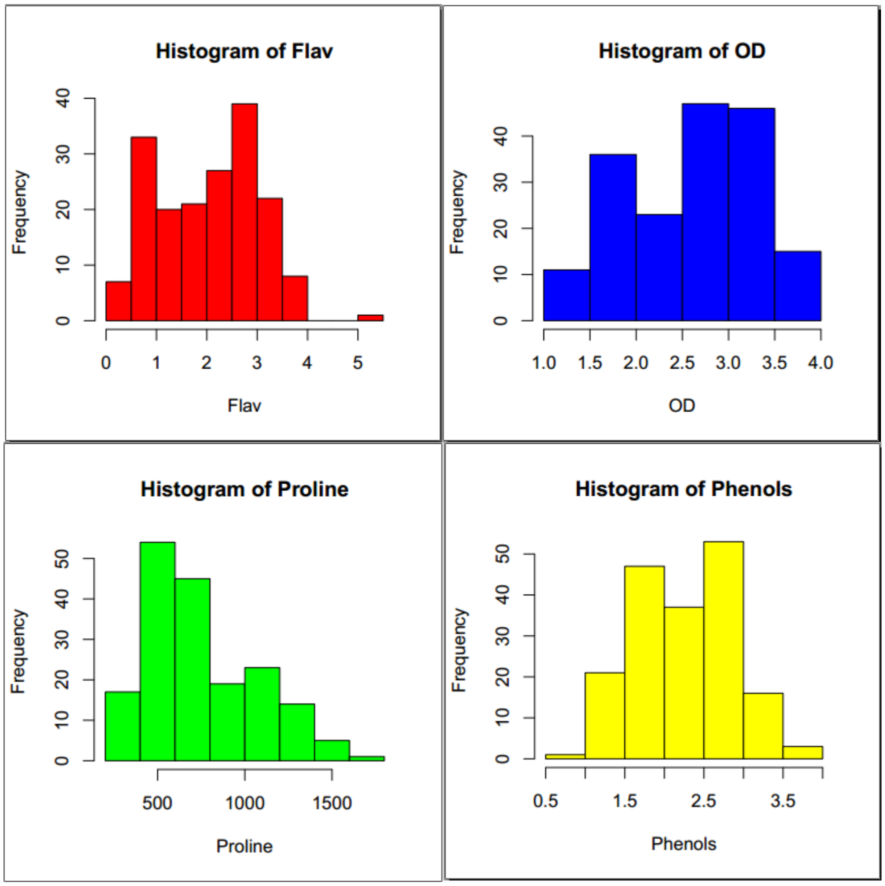
Box Plots:







The box plots allow for a quicker visual understanding of the distribution of each variable along with potential outliers. R allows for many customizations to the graphics. Displayed above are titles, labels, and color coding differentiation. Finally we will plot each density to see the individual distributions. We use the hist function for this and once again provide color differentiation.



Above we can conclude that Phenols has the most ‘normal’ looking distribution. We see that Flav may have an outlier out at 5, and we quickly can see the range that each variable takes on. Through R’s statistical capabilities and data visualization a large dataset can become significantly easier to manage and understand. It is essential to do thorough exploratory data analysis on any dataset that you want to define, model, or explain.