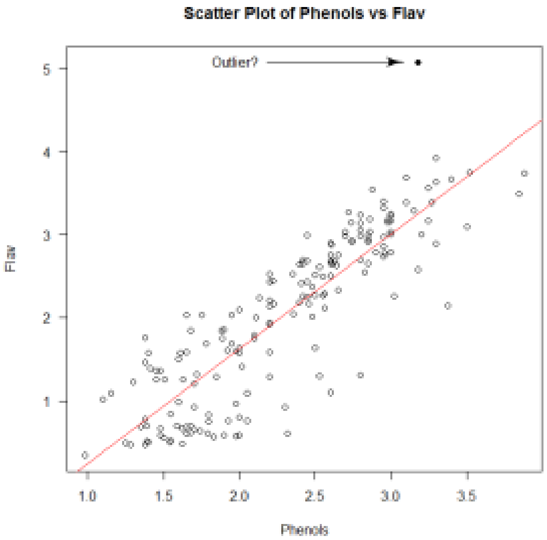
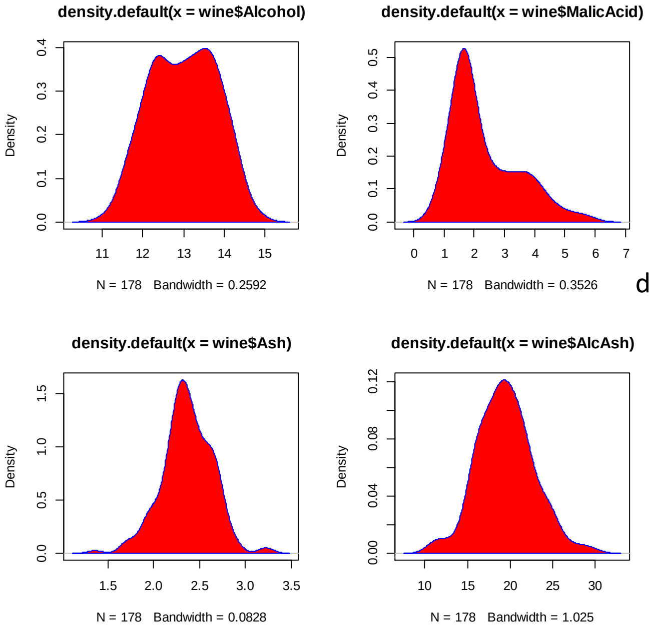
Daniel Prusinski
Predict 412

Programming Assignment 1: Statistical Graphics in R

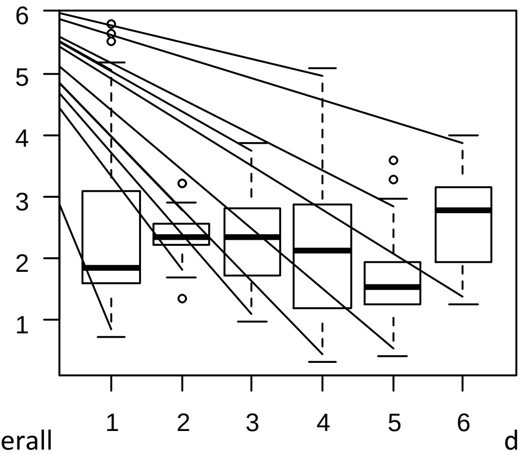
A well rounded Exploratory Data Analysis (EDA) contains data visualizations along with numerical analyses. This EDA will showcase a few of the graphical functions in R using the wine dataset. Before delving into visualizations, parametric analysis is helpful for context. The wine dataset contains 178 observations and 15 variables. Thirteen of the variables are numerical, and two are class variables. While I could list the Minimum, Mean, and Max, visualizations are better. Utilizing the "plot" function (see appendix one) produces a scatter plot of each of the variables with one another. From this graphic, it can be seen that a few of the variables appear correlated with one another including Flavonoid and Phenols. This individual scatter plot to the right highlights scatter plots in R, and I would like to call out the title, axis' label, regression line, and the outlier. All of these functions where carried out with the basic R graphic Package.





The code I wrote shows the density plots filled in with red and are graphics 5-8. Ash, Alcohol Ash, Mg, and Proa follow a relatively normal distribution. Alcohol, Malic Acid, Phenols, Flav, Color, Hue, and Proline are positively skewed in their distributions. Phenols, NonFlavPhenols, and OD are negatively skewed. For the skewed variables, a data transformation such as logarithmic transformation might modify the distribution such that the data followed a more normal distribution. Graphic 3 in the appendix demonstrates the density of the variable Flav, and one can see that the distribution is rather positively skewed with a slight increase at the end from the one outlier data point. Graphic four in the appendix highlights the scatter plot matrix capability of basic R, which is very helpful for analyzing multiple variables at once. Beyond the scatter plot, which reveals how variables interact with each other, the density plot is very helpful for analyzing the distribution of each variable.

Given that thirteen variables are continuous and two are ternary, one could build both a linear regression model as well as logistic regression model depending on what one wanted to predict. In addition, one could build decision trees for a machine learning approach. In order to showcase more graphics in R, the lattice package offers unique data visualizations. Box plots are very helpful for analyzing a variable or groups of variables in respect to data distribution and outliers. In the graphic to the left, the numbers represent the following variables, 1= Malic, 2= Ash, 3= Phenols, 4= Flav, 5= Proa, 6= OD. From this visualization one can see that Malic, Ash, and Proa have a few outliers that may be skewing the overall distribution. All the variable box plots can be seen in graphics 9 and 10 in the appendix. From theses few visualizations, one can assess linearity, density/distribution, and potential outliers. These rudimentary graphics are essential for linear modeling, and assessing the generalizations from key metrics. Knowing the outliers is also important for classification modeling. From this first assignment, I feel like I have very few graphics to show for the amount of time I invested (15 hours), and I wish I had more time to continue exploring. In order to successfully complete future assignments, I will need to scope the projects before modeling and stick to key functions. Given all the packages in R, it is very easy to explore without generating results.


6
5
4
3
2
1
1
2
3
4
5
6