Morganne Peak

Data Practicum II

Final Project Write-Up

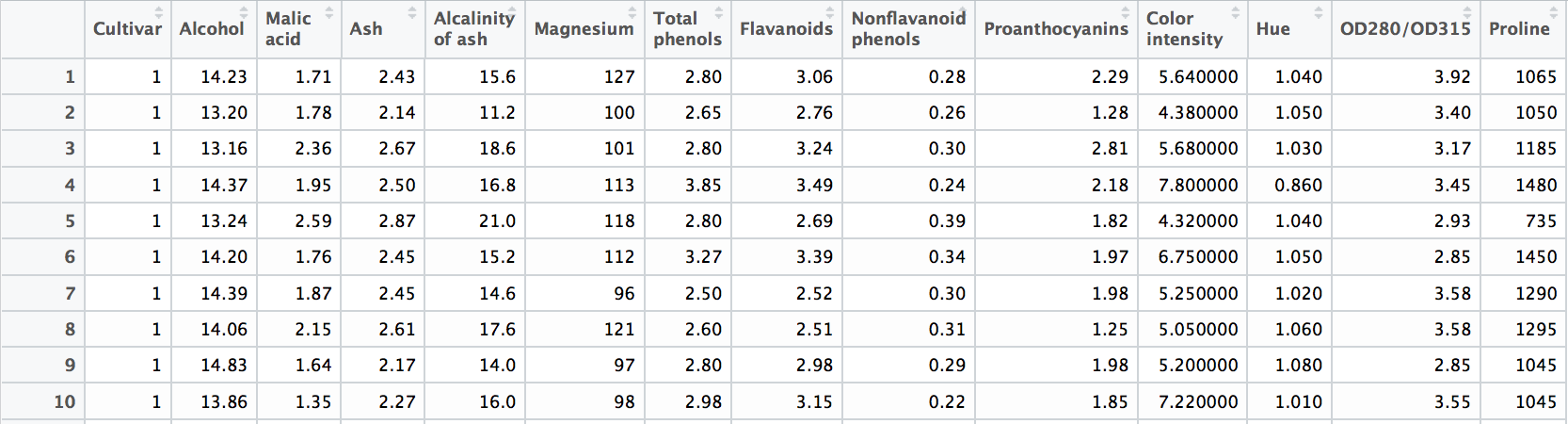
Dataset:

I retrieved this dataset from the UCI Machine Learning Repository. This is a dataset that has 13 attributes of 178 different wines and each wine is categorized into one of three cultivars. The data is derived from a chemical analysis done on different wines from Italy. The dataset was ready clean because there were not any missing values for any of the attributes and nothing else that I thought needed to be cleaned up.

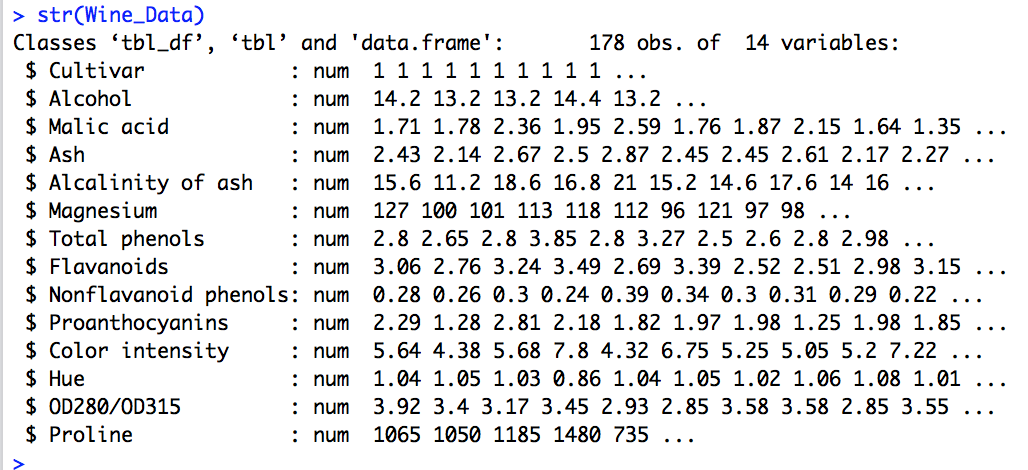
Variables:

There are not any specific decriptions of the different variables but they are listed below:

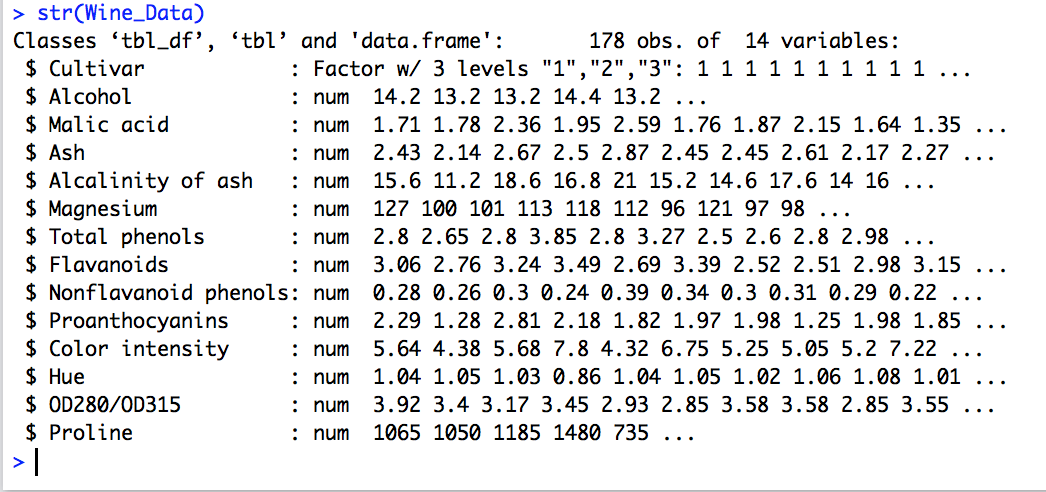
* Cultivar
* Alcohol
* Malic Acid
* Ash
* Alcalinity of Ash
* Magnesium
* Total Phenols
* Flavanoids
* NonFlavanoid Phenols
* Proanthocyanins
* Color Intensity
* Hue
* OD280/OD315
* Proline

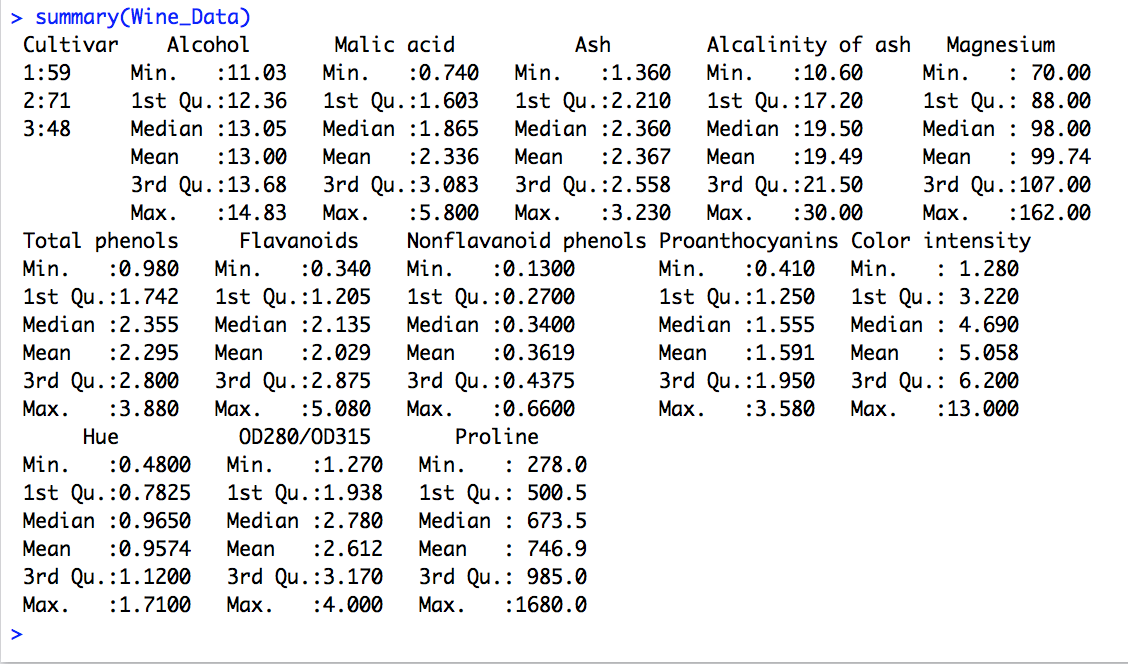


I also ran the str() command in R studio to learn more about how the variables were loaded. I also included that screen shot below:



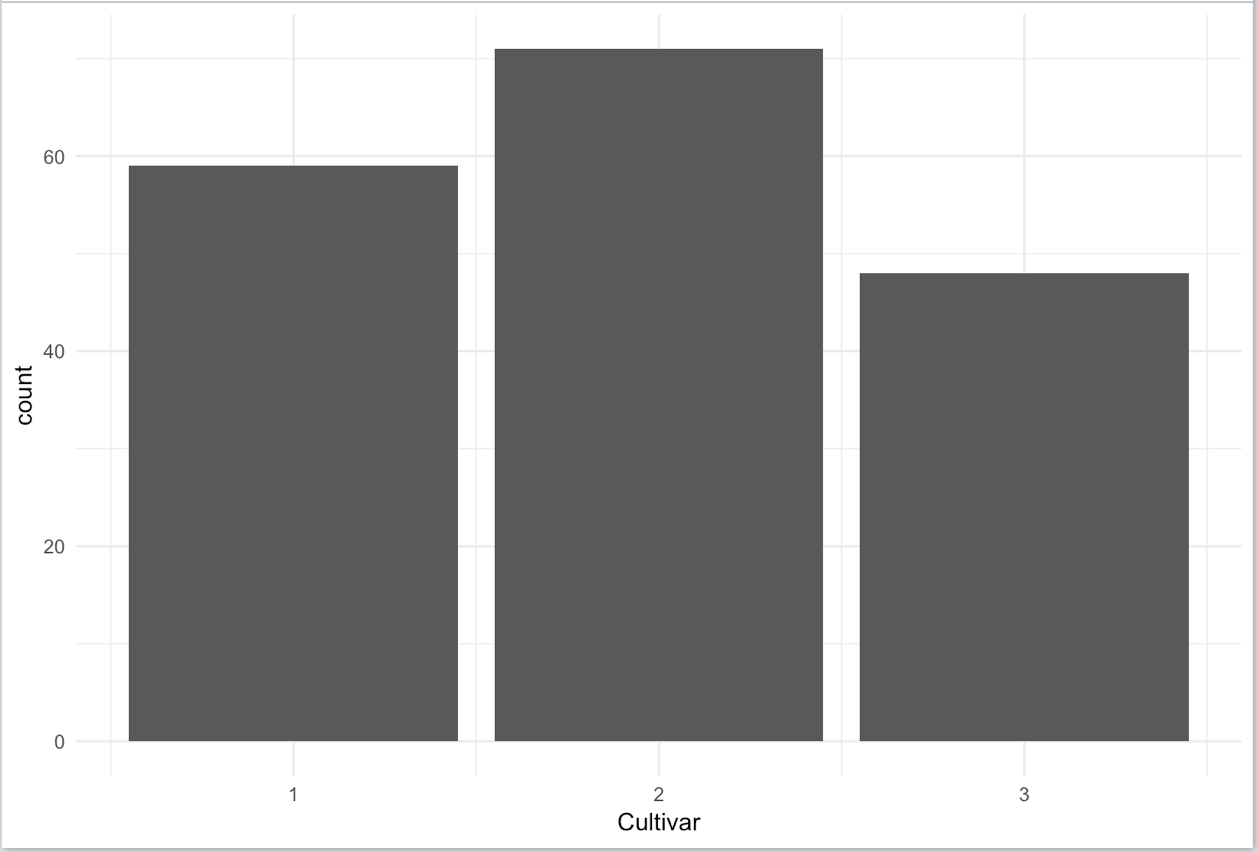
I decided to keep all the variables the same besides Cultivar which I will change to factor to make the data analysis produce better results. The str() and summary() commands after that change are shown below:



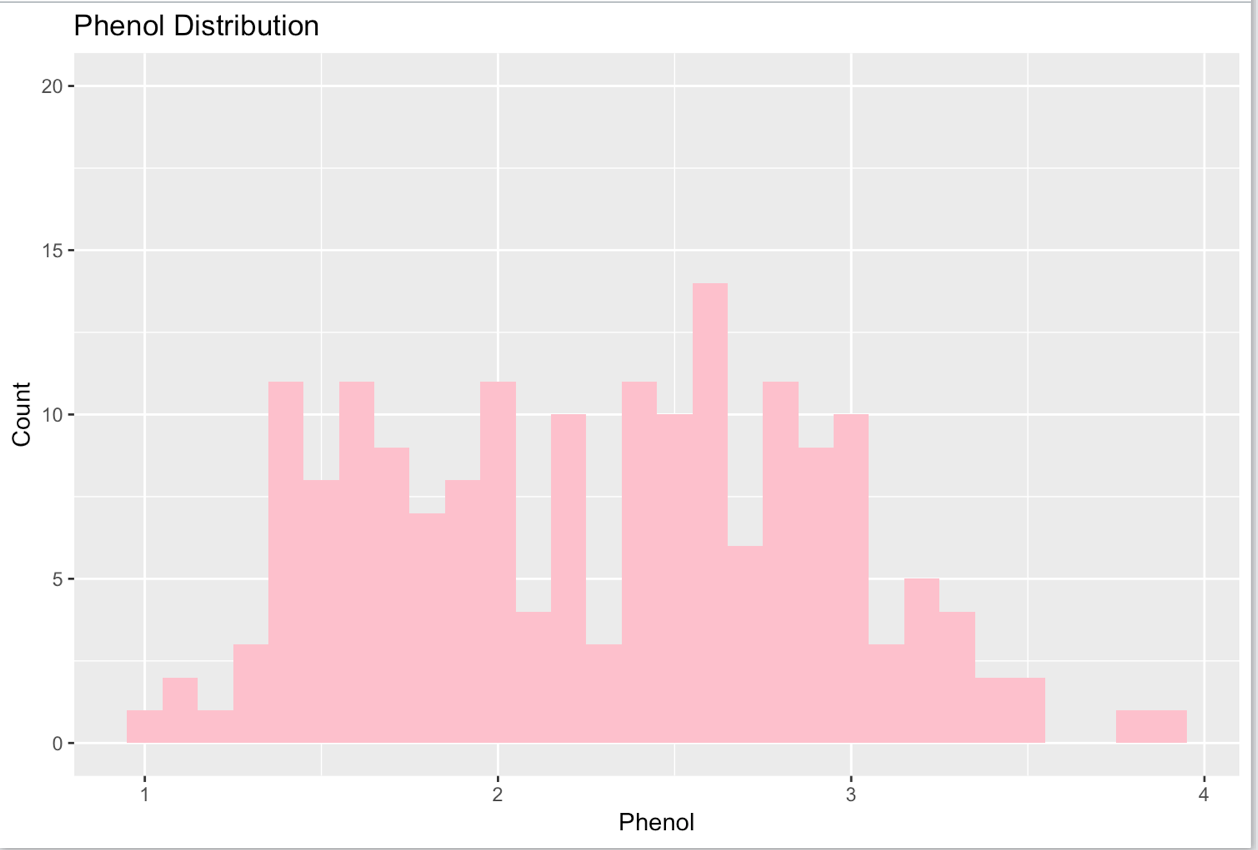
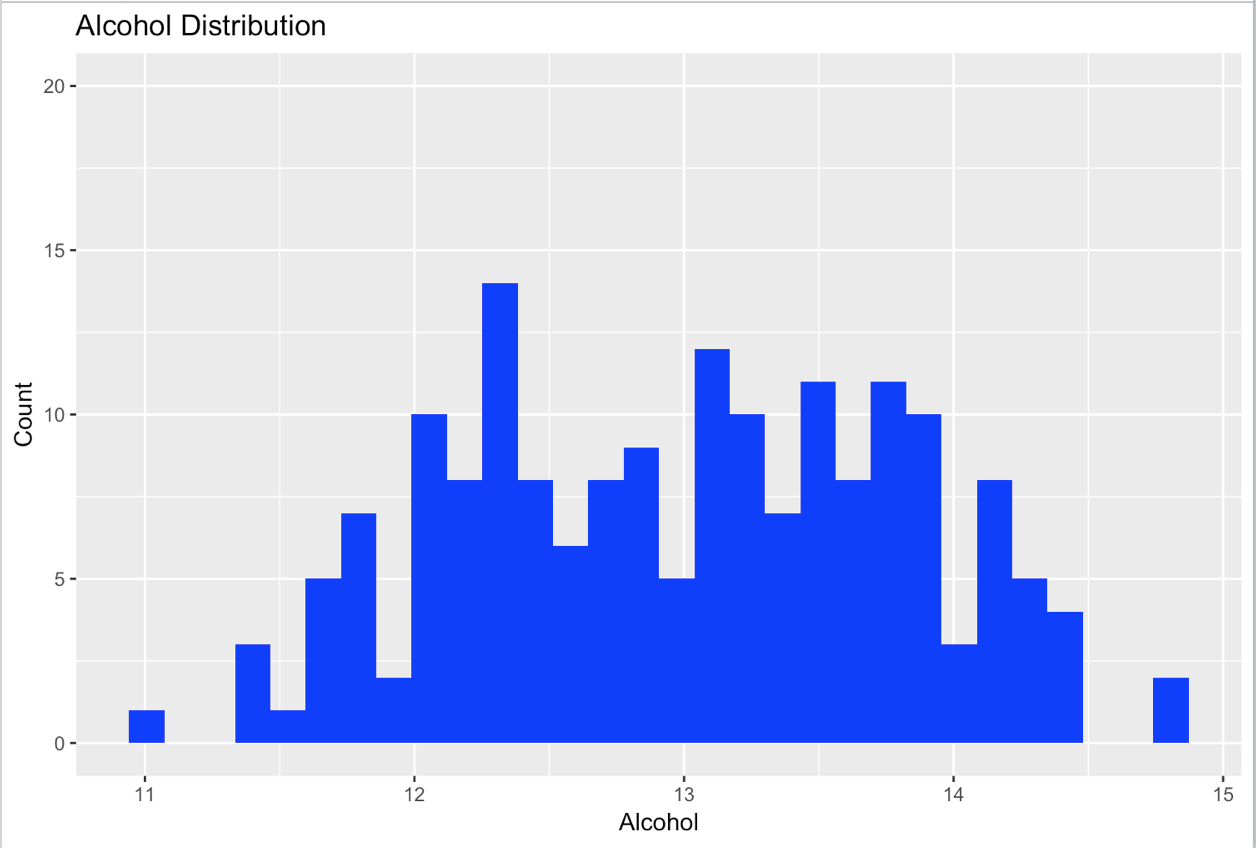


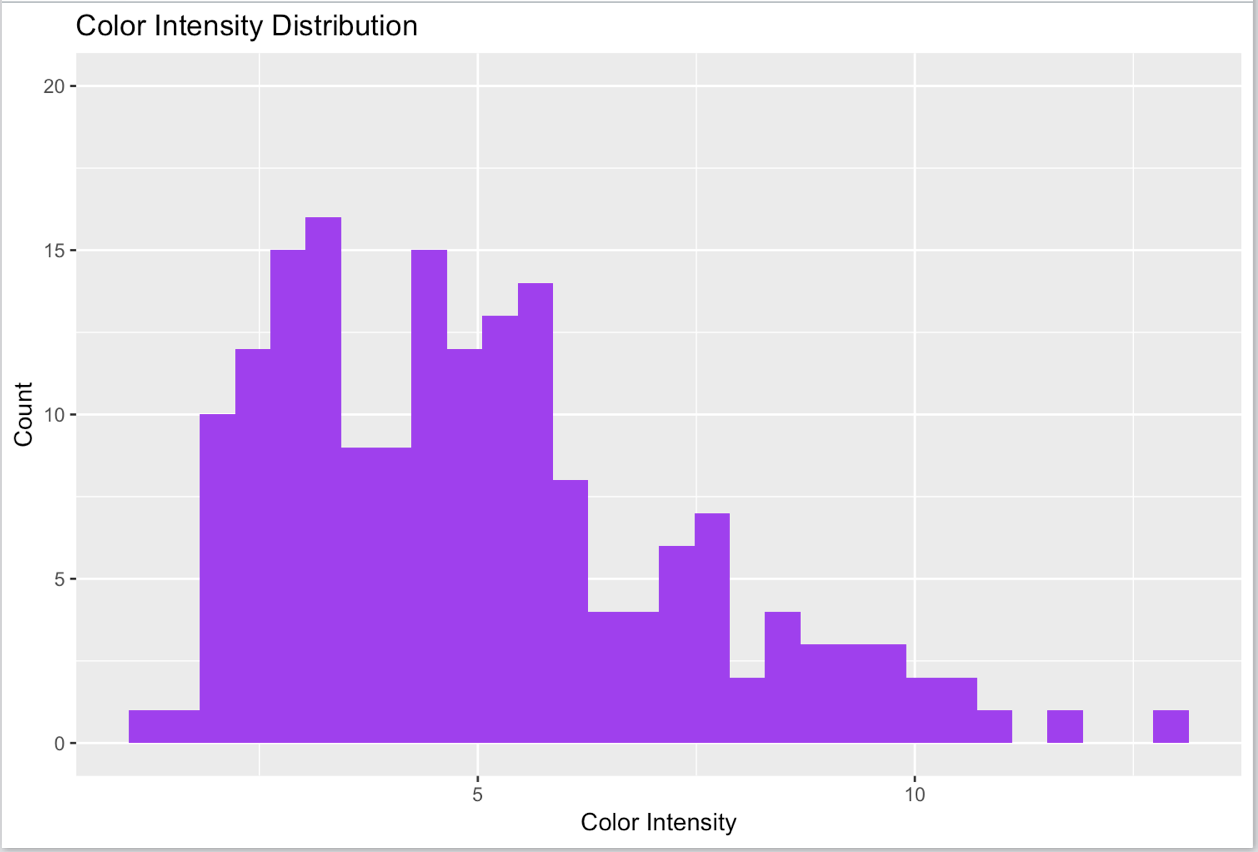
Data Visualization:

I created a few visuals for this project. The first graph I created is a bar graph for the three different Cultivars that can found in the dataset. I created this graph because I wanted to see how of many of each Cultivar there is within the data. This was important because there needs to be enough of each Cultivar to create a good model. The bar graph I created is below:



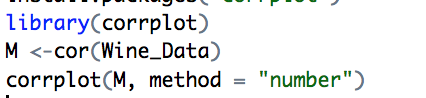
I also created some histograms of some of the other attributes just to see how they were distributed and if I would need to normalize the data, those graphs are below:





Both the distribution for Alcohol and Phenol are relatively normal. However, Alcohol is slightly skewed to the left and Phenol is slightly skewed to the right. Color intensity is clearly leaning towards the left side and not normally distribution so it may do the model some good to normalize the data first before moving forward.

I also create a heat map for this data set to see if there is any correlation between Cultivar and any of the other attributes. The commands and heatmap I created are below:

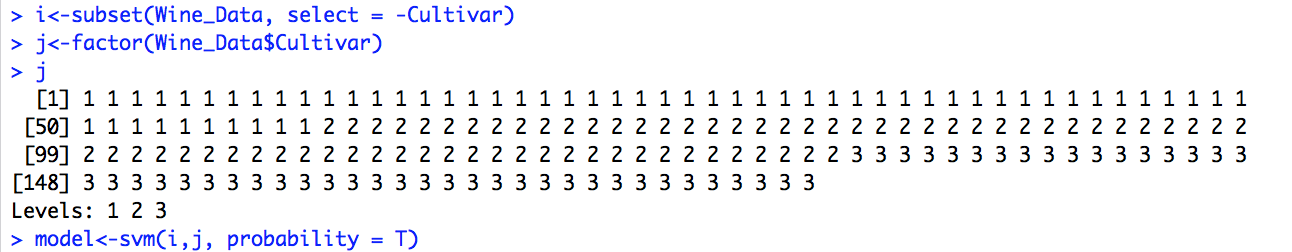


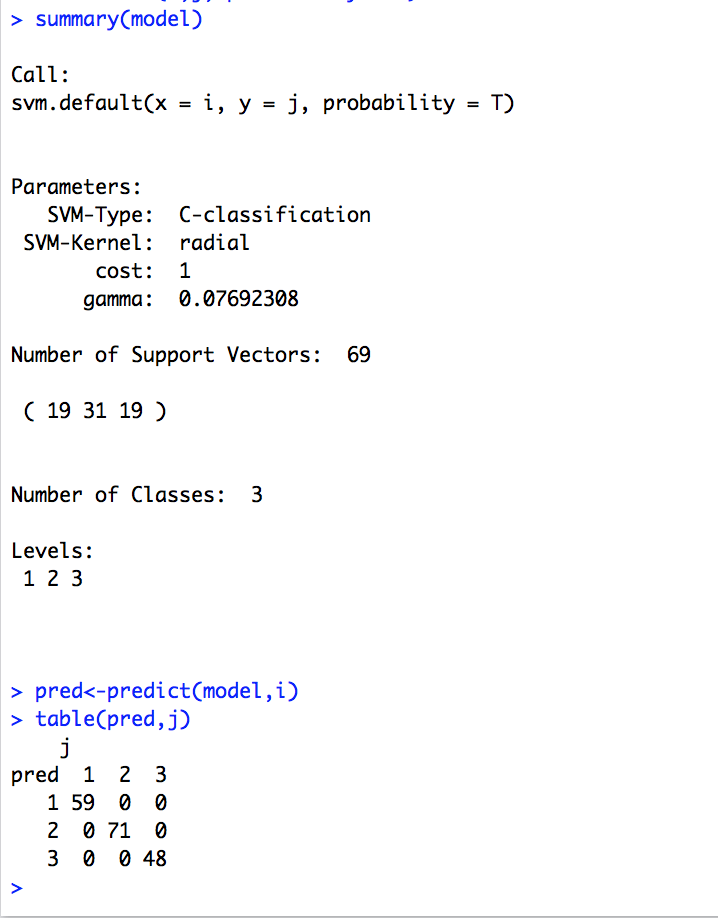


From the heatmap I can see that there are a few attributes that are correlated with the Cultivar attribute. Total phenols, Flavanoids, Hut, OD280/OD315, and Proline are the most correlated variables to the Cultivar type. These factors might come into play the most when trying to classify the different wines.

Models:

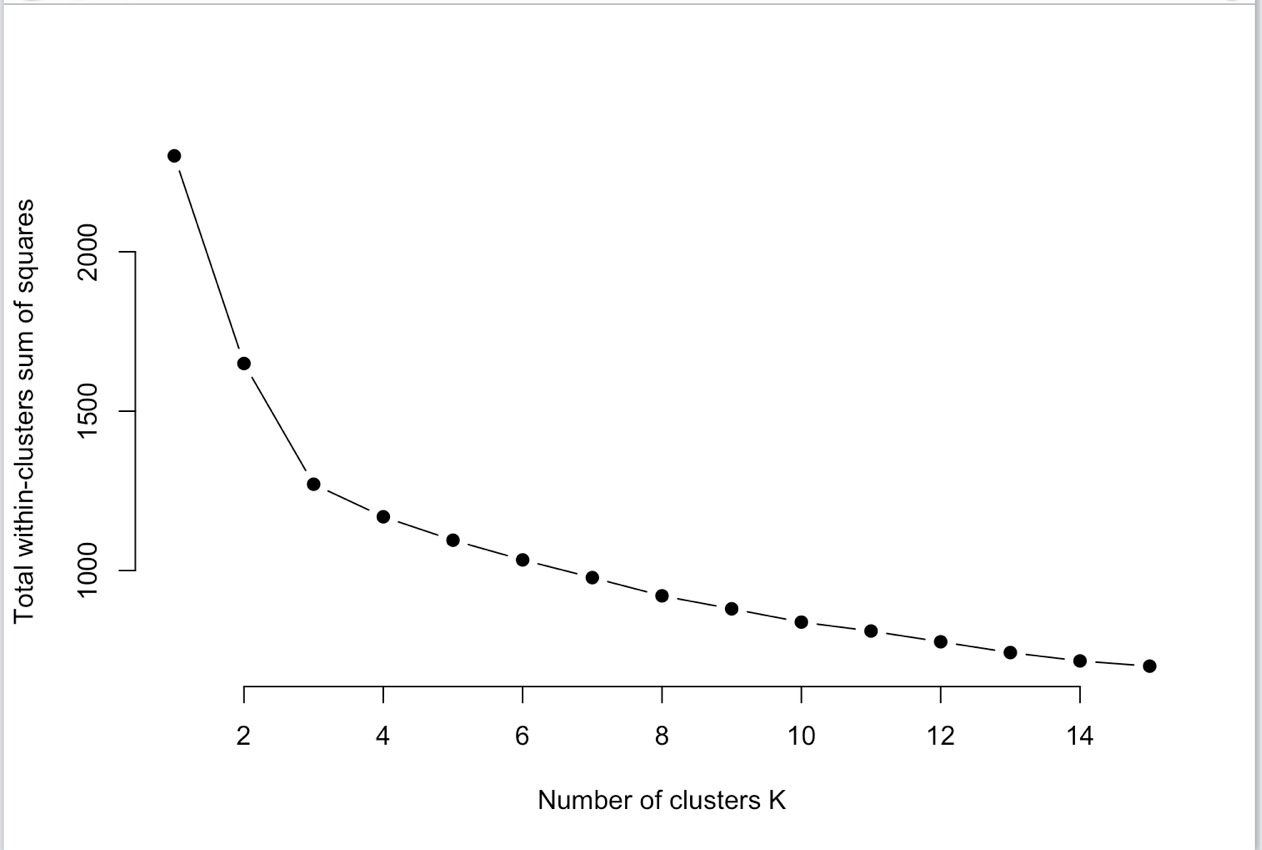
At the beginning of this assignment, I wasn’t sure about what type of model I wanted to create. I ended up doing a couple different ones. I did supervised machine learning, a decision tree, and k-means clustering. The first one I tried to was supervised machine learning using support vectors.



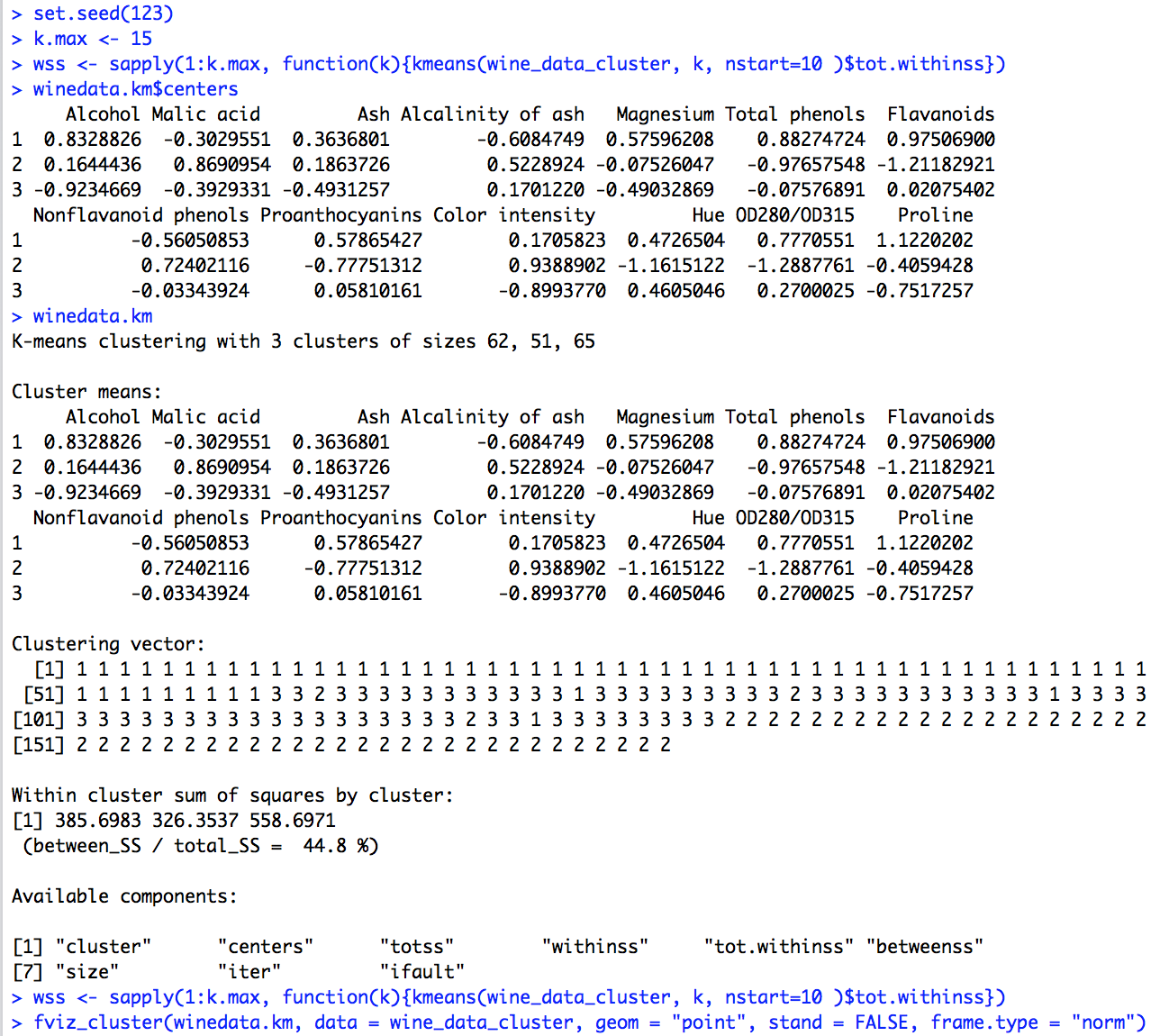


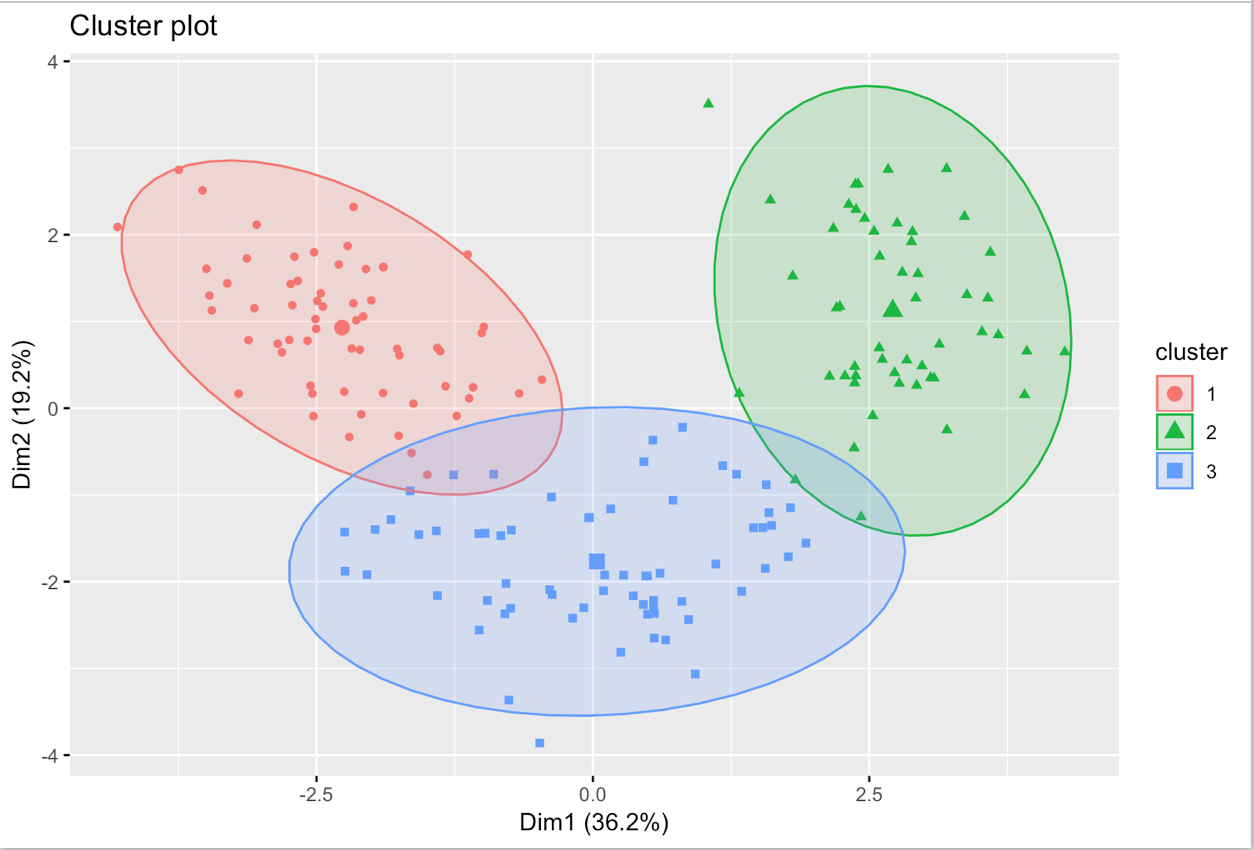
The result from my support vector machine is interesting. The accuracy is 100% each time I run it. I have not ever experienced a 100% result from this type of machine learning before and I found it to be a little odd. Perhaps this is something that happens more often than I think but I just have never seen it before. The result from this machine made me decided to try others to see if those may work better.

The next model I decided to try was kmeans clustering. In order to determine the ideal number of clusters for this data I used the elbow method. The elbow method determined that k=3 was the optimal number of clusters for this data set. That elbow graph is below:



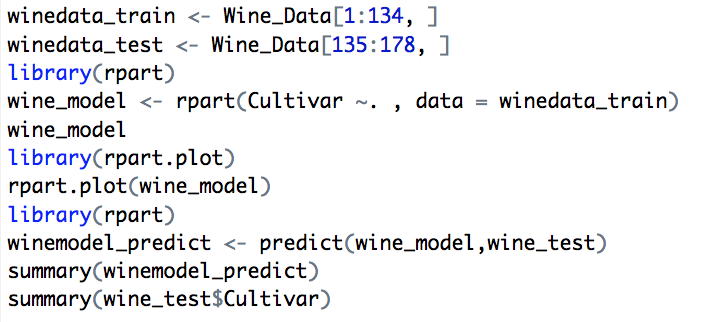
Once the optimal number of clusters is determined, I can the clusters script to see if the machine can predict the clusters that each wine belongs to. The command and the graphical display of the data is below:

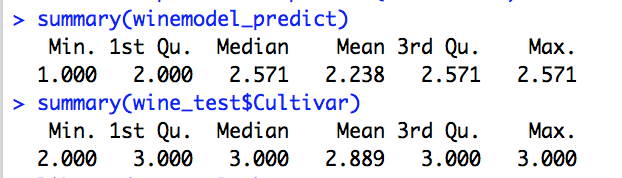


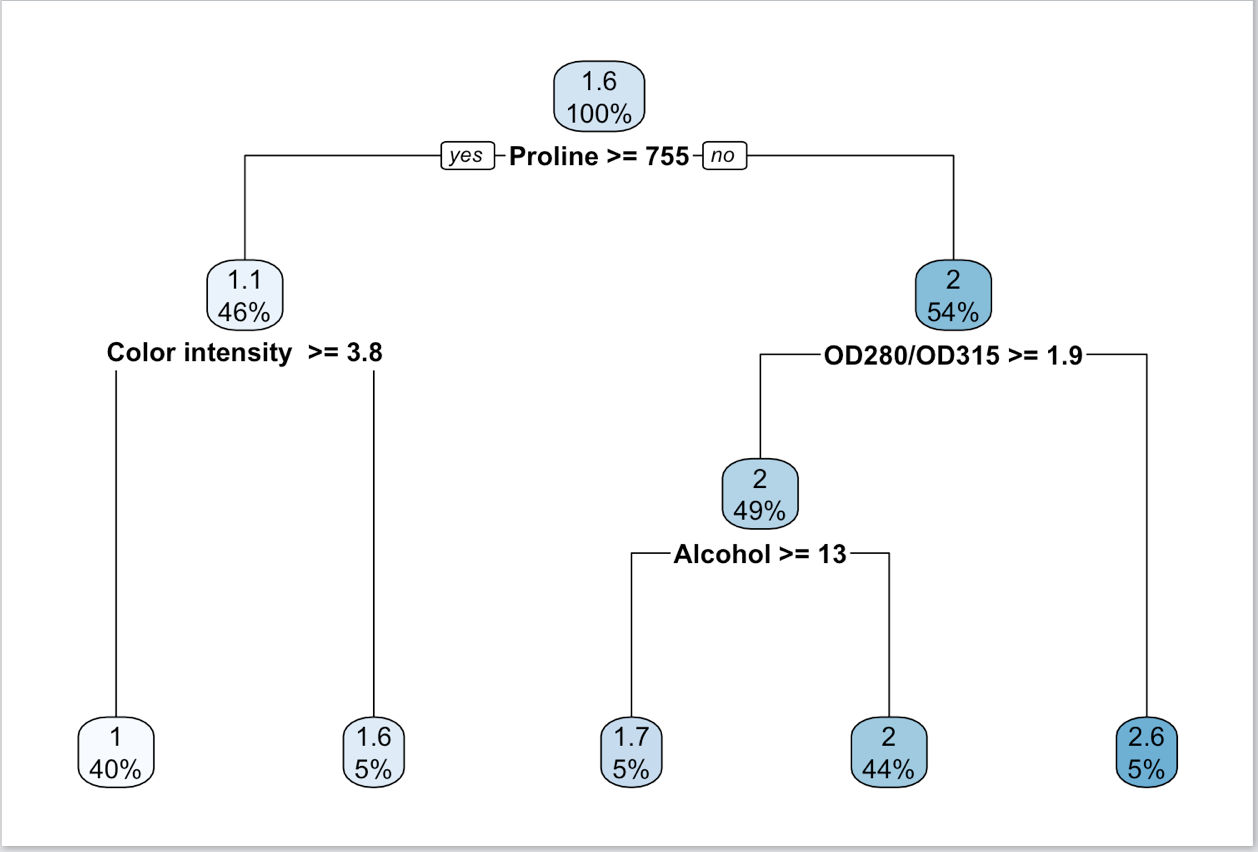


K-means clustering seemed to be pretty successful. The blue cluster had some overlap with the red and green clusters but there isn’t a lot of overlap, just a few points.

The last model I created was a decision tree. I tried to use this model for classification as well since I wasn’t certain about my other classification model. I included all of the coding and a decision tree graphic below. I decided to see how the model did by comparing the summaries of my prediction model and the test data set. It turns out that the summary of the prediction model and the summary of the test model are not too far off. That should mean that the decision tree for classification was pretty accurate.







Conclusions:

I am concluding that it is possible to see that there is a correlation between the cultivars and the other attributes. I think the best way to see that is through the heatmap that I created. You could also see that there is a relationship there because I was able to create clusters that put the test data into different clusters without too much overlap. Since those correlations exist, it is possible to determine the wine cultivar using the other attributes in the dataset.

References:

(n.d.). Retrieved from http://archive.ics.uci.edu/ml/datasets/Wine

(n.d.). Retrieved December 17, 2018, from https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

(n.d.). Retrieved from http://www.rpubs.com/jasonchanhku/wine

Determining the optimal number of clusters: 3 must known methods - Unsupervised Machine Learning. (n.d.). Retrieved from http://www.sthda.com/english/wiki/print.php?id=239

Ggplot2 barplots : Quick start guide - R software and data visualization. (n.d.). Retrieved from http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization#bar-plot-with-labels