Week 8 Final Assignment

Species Data Set

1. Using the species dataset provided to you, please develop a multiple regression model such that the model predicts width of p1.width using the remaining 4 variables as predictors. Once you have developed the model, please write a brief summary explaining your model and how you would evaluate your data model. Also include in your summary, a correlation scatter plot matrix of the four variables length, p1.length, s1.width, p1.width

Ans:

based on the scatterplot matrix, the histogram, prediction for increased width for p1.width appears to decrease as frequemcy increases but there is a steady increase in width to somewhere between 1.0 and 1.5 where it peeks at just under 1.5 and remains the average at that point with a mean width of 1.15. With increasing frequency p1.width of 1.15 seems to be a steady occurrence with outliers going up to 2.5. 1.15 in width is most commonly occurring in p1.width so I think it is safe to predict that width will remain steady at 1.15 even with an increase in frequency.

I made a comprehensive scatterplot matrix using the following code:

pairs.panels(species\_dataSet1[c("p1.width", "p1.length", "s1.width","s1.length", "variety")])

My linear model equation was as follows with p1.width coefficient of -.47314

p1 width coef = -.47314

-.47314 + s1.length + s1.width +p1.length \* species coef

2. Using the same dataset (as in 1), develop a K nearest neighbor model. Also include in your R code the function that computes the accuracy of your model. Once you have developed the model, please write a brief summary of your data model.

Ans:

I wrote the following model:

species\_dataSet2\_test\_pred <- knn(train = species\_dataSet2\_train, test = species\_dataSet2\_test,

cl = species\_dataSet2\_train\_labels, k = 12)

My k-value was 12 because that is the square root of the total number of rows. I felt that I should train and test on the same data because if we are testing on the exact same data it should always make correct predictions resulting in 100% accuracy. Unfortunately when I ran the KNN algorithm it gave me the following in the output:

Error in knn(train = species\_dataSet2\_train, test = species\_dataSet2\_test, :

NA/NaN/Inf in foreign function call (arg 6)

In addition: Warning messages:

1: In knn(train = species\_dataSet2\_train, test = species\_dataSet2\_test, :

NAs introduced by coercion

2: In knn(train = species\_dataSet2\_train, test = species\_dataSet2\_test, :

NAs introduced by coercion

> .

I'm not sure why it did this. I looked over the code and argument 6 is supposed to be in there. I tried tweeking it but got the same results every time.

3. Using the same dataset, apply k means clustering to the dataset. Remember, k means clustering is used with unlabeled data, so you will need to exclude the ‘variety’ column from the analysis. Once you have developed the model, please write a brief summary of your data model.

Ans:

The mean for p1.width is 1.19 , so looking at the cluster data, cluster1 is showing below the mean for all columns so a weak prediction for p1.width. Cluster 2 is mostly negative in all columns again showing a week prediction for p1.width. Cluster 3 is at or above the mean so a strong prediction for all columns including p1.width. Cluster 4 is below the mean and showing negative values so a weak prediction for all columns including p1.width. I'm not sure what to make of Cluster 3 being the one to stand out among the rest except that the average p1.width is seen most commonly occurring in Cluster 3.