Bryson Cook
ISYE 6501, Spring 2018
Homework 2

3.1(a)
Using a 10-group cross validation for the k-nearest neighbor model, a final accuracy of 85.01% with k = 10 was found when compared to the responses in "credit_card_data.txt". In last week's homework the max accuracy of 85.32% occurred at a k-values of 12 and 15, so there is a change in the k-value choice and drop in accuracy when using the cross-validation, which is expected. Since the training data set was different and the model was validated against a separate subset of the data.

3.1(b)
I used rotation to put the data into training, validation, and test sets. There is a risk of introducing bias into the data this way, but I think the risk of a pattern in the data is small when using credit card application history. The k-nearest neighbor model was again used. The code sweeps through the validation data line by line and for values of k= 1 through 50. K=6 is chosen as the best. When the model is then applied to the test data, a final accuracy of 81.53% is obtained. When compared to the HW1 and cross-validation data, those accuracies were optimistic.

4.1
I work in the Power industry for a gas turbine OEM. I believe that clustering could be useful in grouping customers by how they operate to make a good guess at the important to them so that sales pitches can be tailored more effectively. If efficiency is determined to be more important vs. total power output, then we can tailor our pitch to products that will maximize their savings, and thus interest, which should help us make the sale more frequently.

The predictors could be:
- operating hours per year: constantly operating or only at peak times?
- average sustained load: when they are running, do they usually run at part load, base load, or peak?
- Current fuel price
- Current emissions limits for their turbines (NOx, CO, etc)

4.2
At first I ran a clustering model on all four parameters, iterating from 1 to 10 clusters. One cluster isn't really applicable, but it helps with the for loop. Using all 4 parameters doesn't really seem to get much, nor does simply using more clusters. The mean distance to the cluster centers do get smaller but that doesn't mean each cluster is better grouped or more accurate than the previous. However, since we do have real data, plotting out the data vs the known responses shows that the Sepal measurements do not provide a good grouping, but the Petal measurements are excellent as shown below.

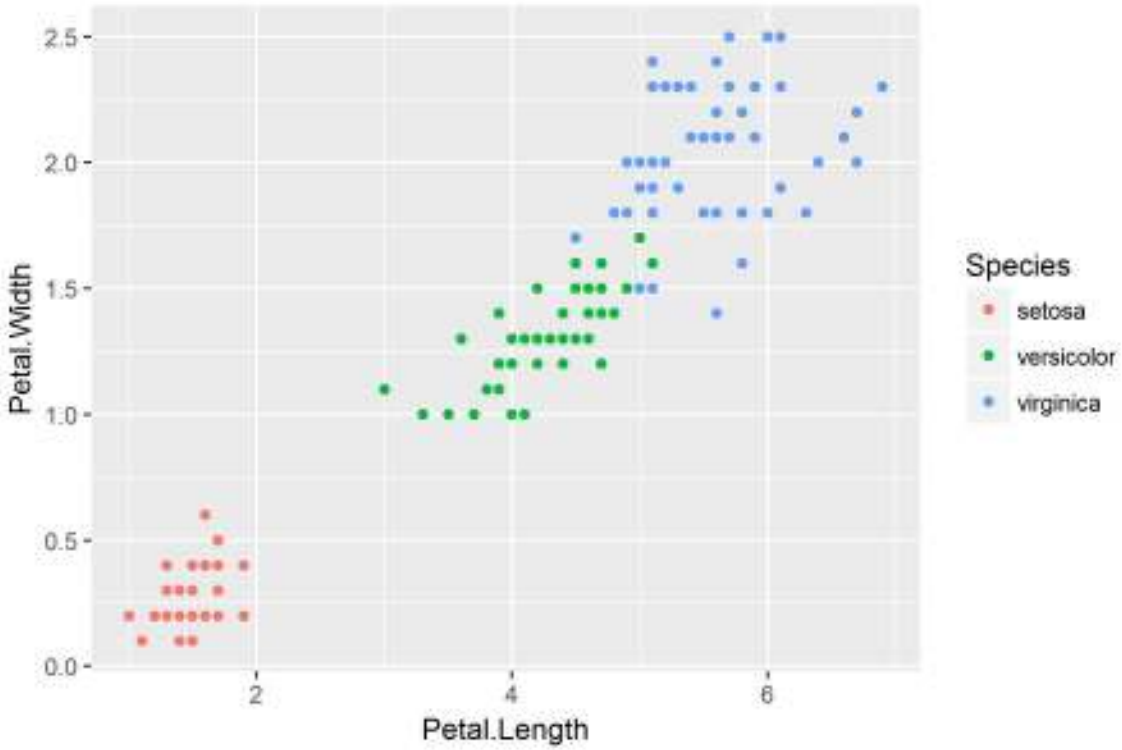

Figure 1. Actual Data Responses

Therefore, the previous model creation should be repeated with only these two predictors. Also, from looking at the grouping of the data, only 2 large groupings are seen so going up to 10 clusters is superfluous and the iterations can be limited to 5 clusters. The below plots show the results:

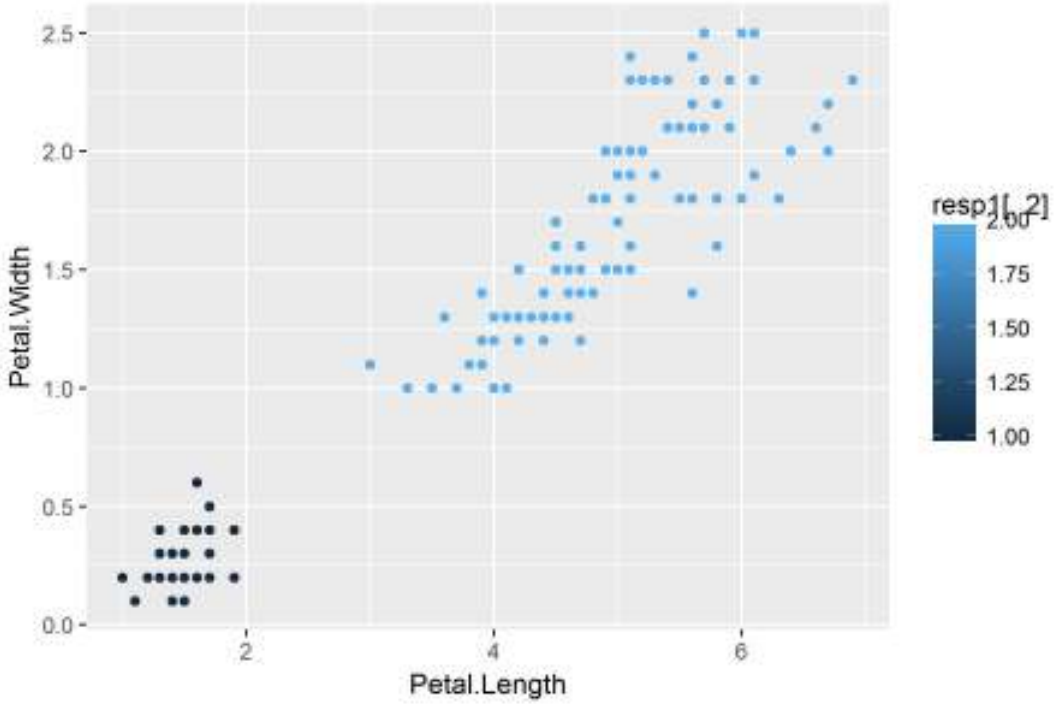

Figure 2. 2 Cluster Solution

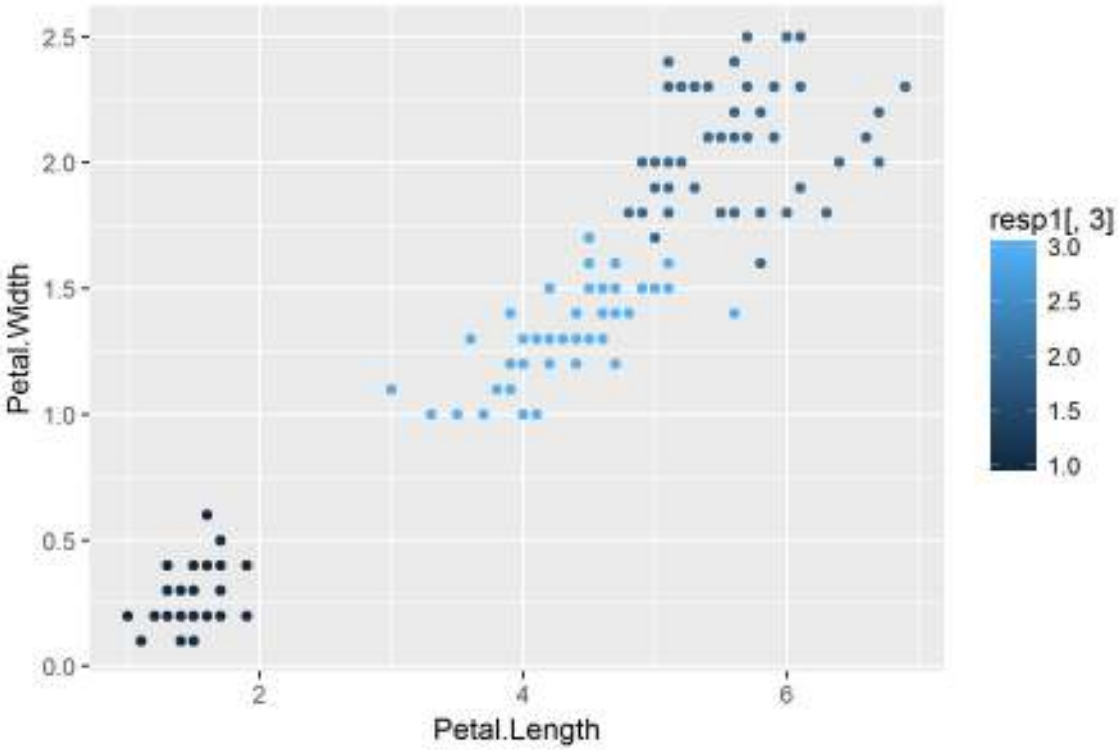

Figure 3. 3 Cluster Solution

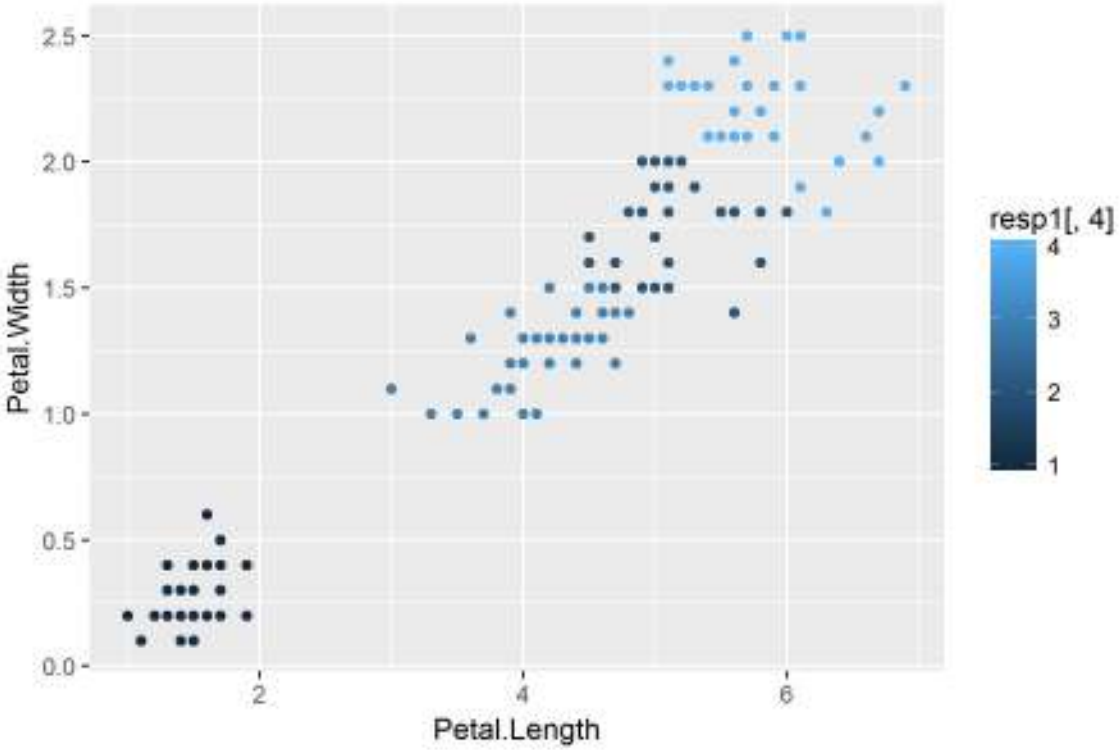

Figure 4. 4 Cluster Solution

In each case, without knowing the actual responses, you could make an argument that the clustering fits the data well and that the final choice could probably depend on understanding the problem deeper. However, since we do know that there are 3 species we can compare the 3 cluster solution to the actual data. In the comparison we see that 2 veriscolor flowers and 4 viginica flowers are mis-classified.

```
   setosa versicolor virginica
1      50          0         0
2       0          2        46
3       0         48         4
```