James Taylor

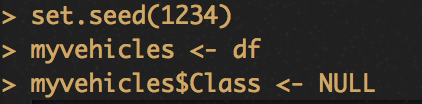
Exercise 7

1) Introduction

I expect the k-means clustering to be accurate at predicting which class a car belongs in. There should be enough differences in the classes for the unsupervised algorithm to discover.

2) Pre-processing

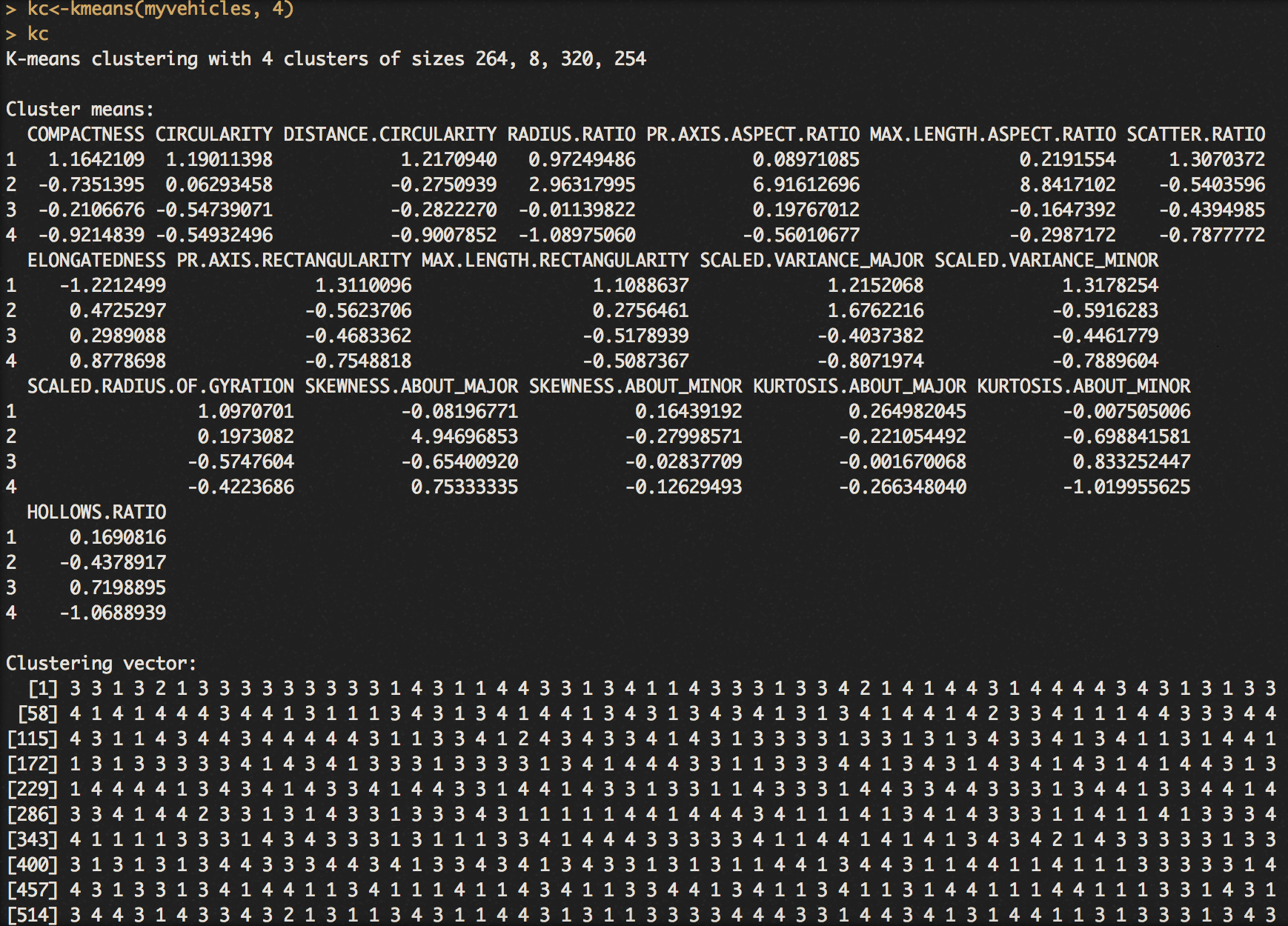
Setting the seed allows for others to recreate the situation from the script. It will allow the same outputs to occur for the same inputs.



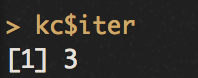
Scaling the variables normalizes the variables. It helps the model recognize deviations in data while not having to handle different scales. The class variable was removed because it is an unsupervised model, meant to recognize patterns in data without training its accuracy on labeled observation.

../../Desktop/2.png

3) K-means

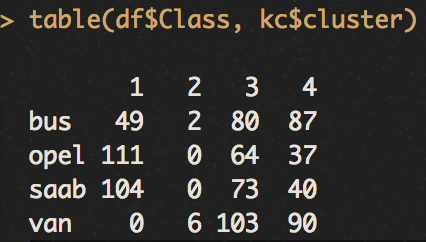


The above is the models output. There are 264, 8, 320 and 254 instances in each respective cluster. The cluster means section describes the shape of the cluster by which observations fall into. The clustering vector shows which observations fall into which cluster. The sum of squares by cluster is the distance between an instance and the closest center point of a cluster.



The above is the number of iterations completed until no data points are reassigned to a different cluster. Each iteration, a cluster’s center point is recalculated. This is based on the location of data points that are are closest to it. Once it has been recalculated, the data points are assigned clusters again. This process repeats until no data points are assigned a different cluster, ending the looping of iterations.

4) Cluster Evaluation



The dominant class for each cluster was opel and van for the tree other three. It did not do a great job of separating the classes. Obviously van should not dominate 3 of the 4 clusters.

5) Plot



The plot shows the 4 clusters that categorizes the data. The second plot has the smallest observations by far, while the others have about the same amount.

6) Experiment with different amount of clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| k | Number of instances in each cluster | Between clusters sum of squares | Within clusters sum of squares | Number of iterations |
| 4 | 264, 8, 320, 254 | 9235 | 5974 | 3 |
| 6 | 8, 228, 150, 157, 148, 155 | 10280 | 4929 | 5 |
| 8 | 201, 90, 29, 138, 8, 133, 110, 137 | 10896 | 4313 | 8 |
| 10 | 81, 81, 121, 8, 92, 96, 162, 29, 66, 110 | 11364 | 3845 | 4 |

7) Summary

As k increases, between cluster sum of squares increases and within cluster sum of squares decreases. This is the goal of k-means models, but it becomes less intuitive because there will be more clusters than categories. This is the case for our set, where the best number of ‘k’s’ tried was 6 more categories than actually existed in the data. There is no right answer for this. The best option is the one that can best separate the labeled categories you wish to