Ray Beecham
Cory Singer
JP Fernandez
Athena Mills
Cade Northcutt QMST 3339 February 24, 2018

Homework #2

1 Partitional Clustering

1. Determine the numerical variables and scale them using the default scale() function in R.

sapply(wine, is.numeric) # Which Variables are numeric using logical operator. Type is False.

scWine <- scale(wine[-1]) # I am scaling the variables minus the first variable 'Type' since it is not a part of the numerical variables.

2. Using the scaled data set, utilize k-means clustering analysis for 25 randomly selected starting points, with initial seed as 1234. Answer the questions below. 2.4 pts, 0.6 pts each.

(a) Using WSS method, determine the number of clusters.

Using the Within cluster sum of squares method it was determined that there are 15 clusters but looking at the plot the best number of clusters is 3.

(b) Report the within cluster sum of squares by cluster. What is the ratio of between sum of squares and total sum of squares? Comment on its meaning, does it correspond to having high quality clusters?

Cluster sum of squares:

```
385.6983 558.6971 326.3537
```

The ratio of between sum of squares and total sum of squares is 0.4477405 and because it is further away from 1, so we would conclude that it wouldn’t correspond to having high quality clusters.

(c) What is the size of each cluster? Share the centroid means of each cluster.

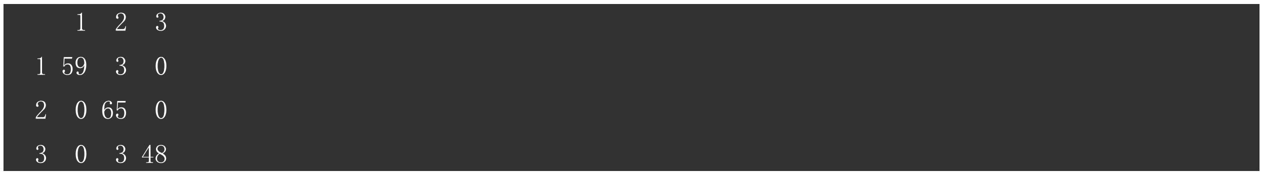
The Size of each cluster is: 62, 65, 51

Ray Beecham
Cory Singer
JP Fernandez
Athena Mills
Cade Northcutt QMST 3339 February 24, 2018

The centroid means of each cluster:

#Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids
#1 0.83 -0.30 0.36 -0.61 0.576 0.883 0.975 -0.561
#2 -0.92 -0.39 -0.49 0.17 -0.490 -0.076 0.021 -0.033
#3 0.16 0.87 0.19 0.52 -0.075 -0.977 -1.212 0.724
#Proanthocyanins Color Hue Dilution Proline
#1 0.579 0.17 0.47 0.78 1.12
#2 0.058 -0.90 0.46 0.27 -0.75
#3 -0.778 0.94 -1.16 -1.29 -0.41

(d) Report the counts of wine types in each cluster. What type of wine do you see most in the first cluster?



```
     1  2  3
  1 59  3  0
  2  0 65  0
  3  0  3 48
```

Judging by the data the first type of wine (59) is seen the most in the first cluster out of 62. 59+3 = 62

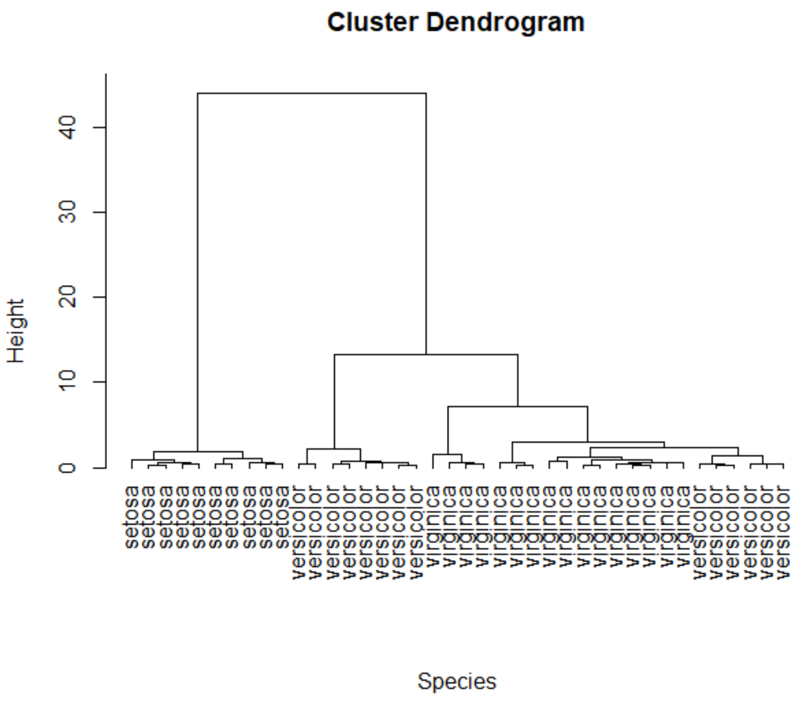
2 Hierarchical Clustering

1. Share the R code to draw 40 randomly selected records from the iris data set that only includes the numerical variables of interest. Use the seed as 123.

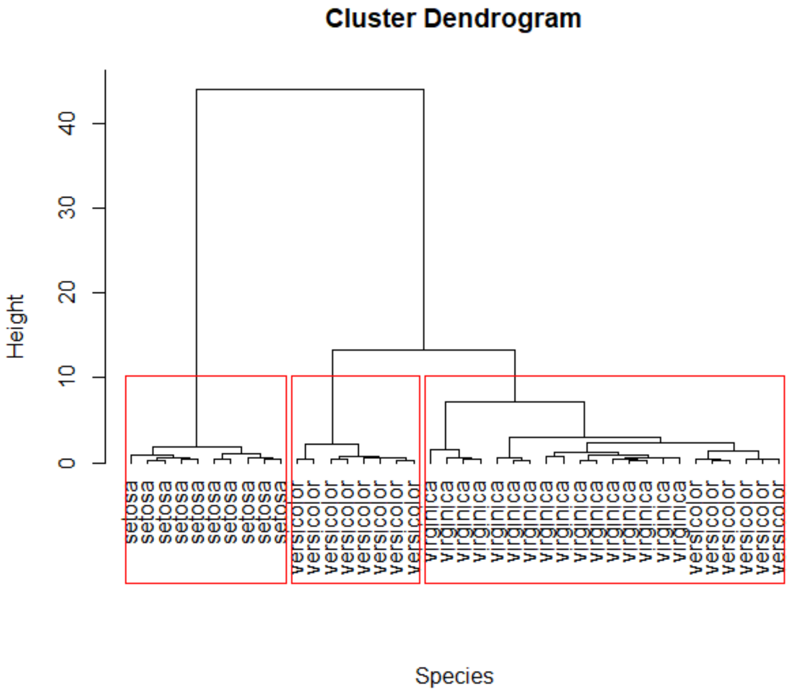
set.seed(123)
myD <- sample(1:dim(iris)[1], 40)
with40 <- iris[myD,1:5]
comp40 <- iris[myD,1:4]

2. Using the Ward‘s method, create a dendrogram. Share the dendrogram that is labelled with respect to species.

Ray Beecham
Cory Singer
JP Fernandez
Athena Mills
Cade Northcutt QMST 3339 February 24, 2018



3. Share the plot of dendrogram with three clusters.


Cluster Dendrogram
Height
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
Species