Project 4, Bus. 744, Unsupervised Learning

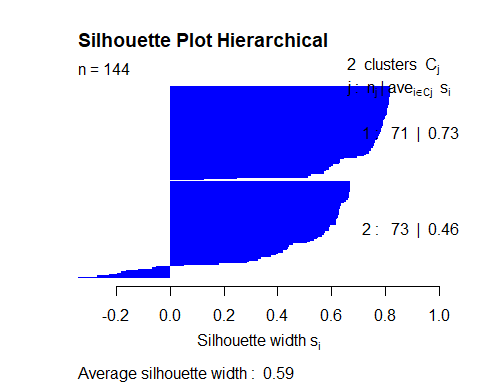
Jack Fitzgibbons

1. Select a dataset of interest and perform a cluster analysis of your data for 2, 3, 4, 5, 6, 7 clusters using either k-means or hierarchical clustering

library(cluster)  
library(datasets)  
data <- read.csv("C://Users/Fitzg/Onedrive/Documents/Airpassengers.csv")  
  
# Hierachical (2 cluster only)  
 avW <- agnes(data[,2:3], method ="ward", stand = TRUE )  
 h2 <- cutree(as.hclust(avW), k = 2) # Cut into 2 groups

1. Evaluate each partition above by creating a Silhouette plot.

# Hierachical Cluster Silhoutte Plot  
 siH2 <- silhouette (h2, daisy(data[,2:3]))  
 plot (siH2 , main ="Silhouette Plot Hierarchical", col =" blue ", border=NA)



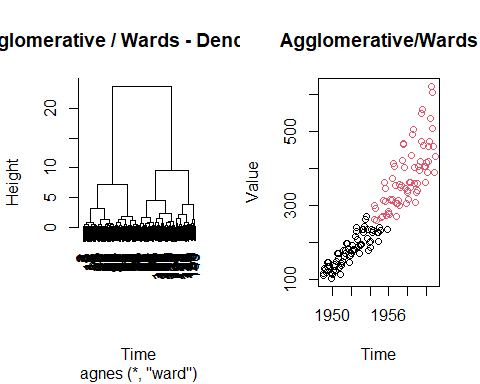
1. Choose an optimal solution based on your results in step (3).

I used the AirPassengers dataset which gives two columns that include a year, and a value for the numbers of airpassengers that traveled during that year. Using hierarchical clustering, I could see that the optimal cluster solution is optimized as a two cluster solution. Group 1 and group 2 have very positive silhouette values. There is small group of observations that seem to be, according to the silhouette plot, classified incorrectly. Comparing the two cluster solution with 3, 4, 5, 6, and 7, the silhouette plot gets progressively worse. More and more observations show an incorrect classification the more clusters that I include in the plot. Moreover, the average silhouette width gets progressively lower as clusters increase. This being a fairly simple dataset with only two variables, the two cluster solution fits best.

1. Profile each cluster from your optimal solution using graphical depictions, summary numbers, etc. and give an intuitive description of each cluster

Looking at the dendrogram, the difference between the two last clusters are very large which again indicates a two cluster solution. However, one set back I had when identifying how many clusters should be considered the solution was identifying where one cluster starts and one begins. Looking at the scatter plot, you can see that there isn’t necessarily an obvious cut off point for each cluster, which initially led me to believe there could need to be more than 2. However, the silhouette results and dendrogram ultimately helped in identifying the two cluster solution. Cluster 1 begins with the earliest data in terms of the year, and ends roughly around 1956 around a value of 250 passengers. Cluster two starts where cluster 1 left off, with little to no gap, visually, between the two, and it incorporates the rest of the values for the remaining years. Luckily, the more years that went by, the trend for passengers was almost always positive. If there was values in the later years that showed a drastic drop off in passengers, this may have needed 3 clusters.

par(mfrow=c(1,2))  
  
 pltree(avW, main="Agglomerative / Wards - Dendogram", xlab = "Time")  
   
 plot(data$time, data$value, xlab="Time", ylab="Value", main = "Agglomerative/Wards", col=as.integer(h2))



library(psych)

## Warning: package 'psych' was built under R version 4.1.3

# Hierachical  
 hierarchDat <- cbind(data,group=h2)  
 describeBy(hierarchDat, hierarchDat$group)

##   
## Descriptive statistics by group   
## group: 1  
## vars n mean sd median trimmed mad min max range skew  
## X 1 71 36.45 21.43 36.00 36.00 26.69 1 83.00 82.00 0.12  
## time 2 71 1951.95 1.79 1951.92 1951.92 2.22 1949 1955.83 6.83 0.12  
## value 3 71 179.76 43.28 180.00 179.40 56.34 104 272.00 168.00 0.11  
## group 4 71 1.00 0.00 1.00 1.00 0.00 1 1.00 0.00 NaN  
## kurtosis se  
## X -1.07 2.54  
## time -1.07 0.21  
## value -1.11 5.14  
## group NaN 0.00  
## ------------------------------------------------------------   
## group: 2  
## vars n mean sd median trimmed mad min max range skew  
## X 1 73 107.56 21.94 108.00 107.92 26.69 66.00 144.00 78.0 -0.1  
## time 2 73 1957.88 1.83 1957.92 1957.91 2.22 1954.42 1960.92 6.5 -0.1  
## value 3 73 378.08 84.27 360.00 370.22 80.06 259.00 622.00 363.0 0.8  
## group 4 73 2.00 0.00 2.00 2.00 0.00 2.00 2.00 0.0 NaN  
## kurtosis se  
## X -1.13 2.57  
## time -1.13 0.21  
## value 0.15 9.86  
## group NaN 0.00