Hierarchical Clustering of the 'mtcars' Data

Michael Levine

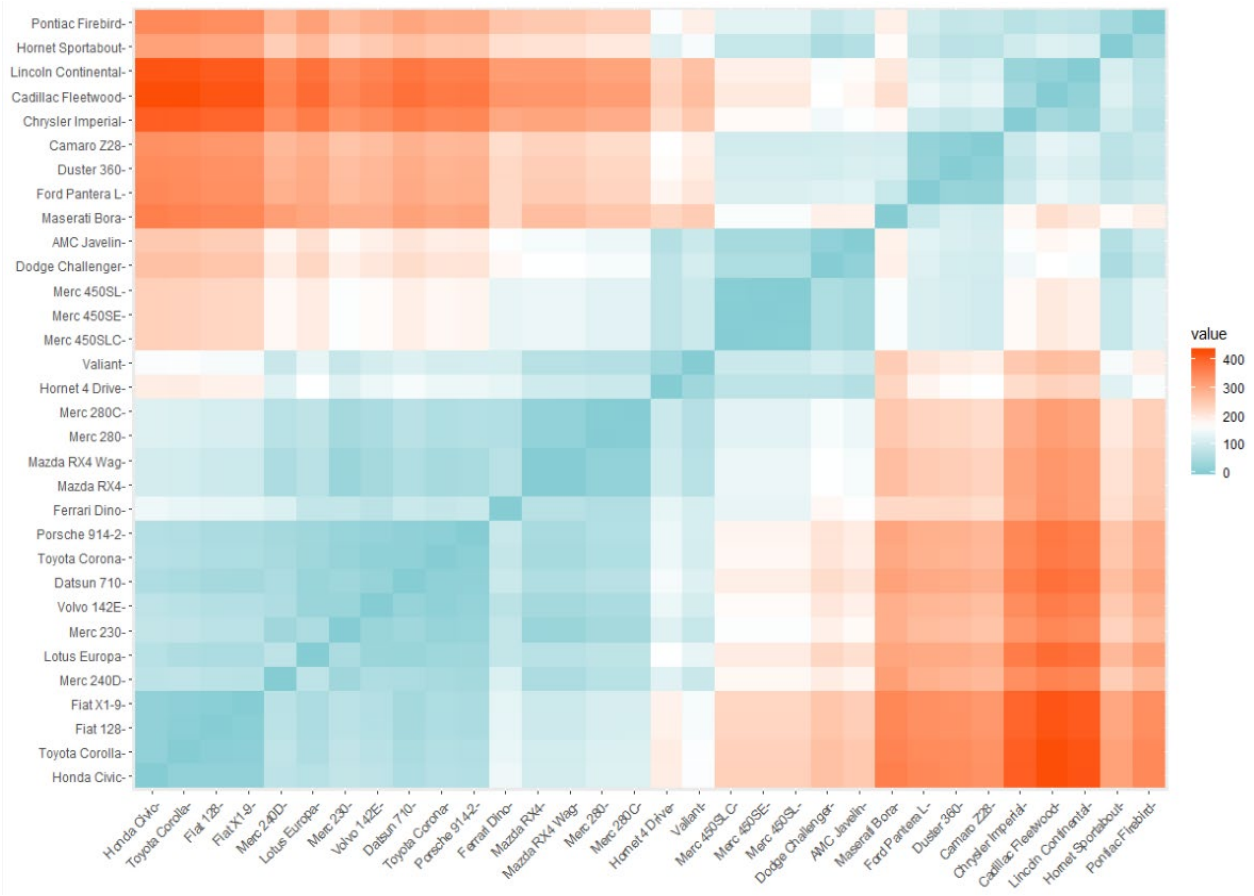
Statistical Learning and Data Mining (STAT 5603-401, Spring 2019)

The R dataset *mtcars* consists of measurements from a collection of 32 automobiles. In the dataset is 11 separate numeric variables. Using 'head', let's look at the first five rows of the dataset:

```
head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

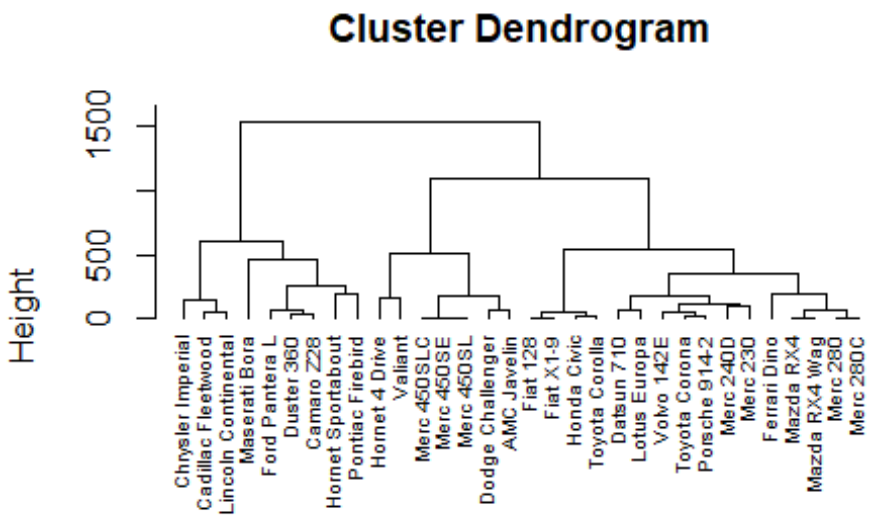
First, let's use the 'dist' function to find the distance between each car. We'll then plot the distance matrix using the *fviz_dist* function.

```
distance <- get_dist(mtcars)
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#
FC4E07"))
```

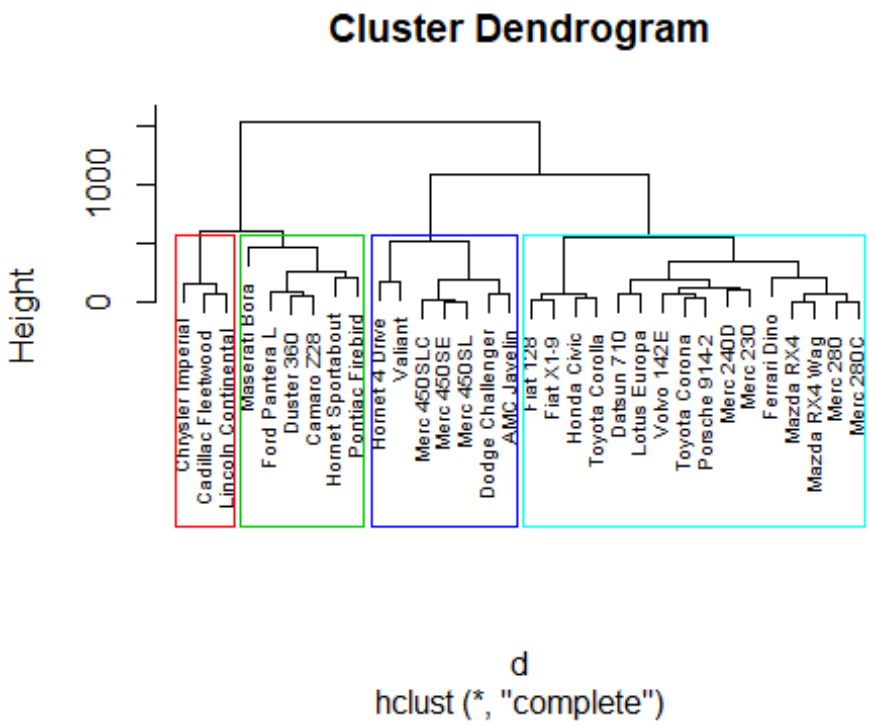


Next, let's perform a hierarchical clustering using the distance matrix that we created in the previous step. We'll then plot the dendrogram into four separate clusters.

```
df2 <- distance
df <- na.omit(df2)
df <- scale(df2)
d <- dist(df2, method = "euclidean")
hc1 <- hclust(d, method = "complete" )
plot(hc1, cex = 0.6, hang = -1)
```

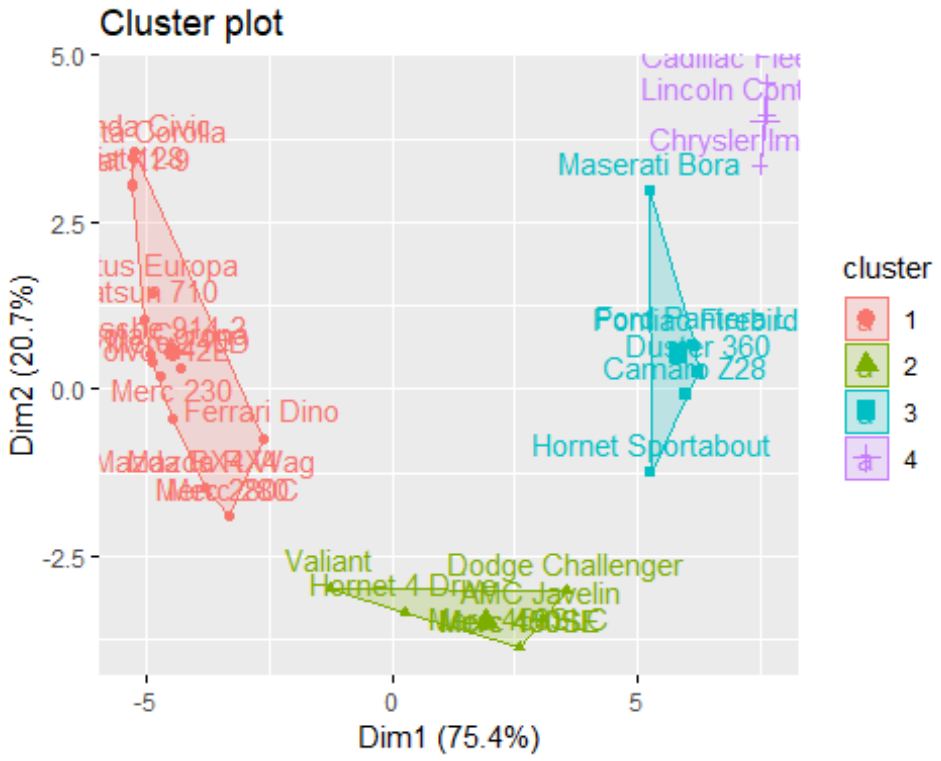



```
# Ward's method
hc5 <- hclust(d, method = "complete" )

# Cut tree into 4 groups
sub_grp <- cutree(hc5, k = 4)

# Number of members in each cluster
table(sub_grp)

## sub_grp
##  1  2  3  4
## 16  7  6  3

plot(hc5, cex = 0.6)
rect.hclust(hc5, k = 4, border = 2:5)
```



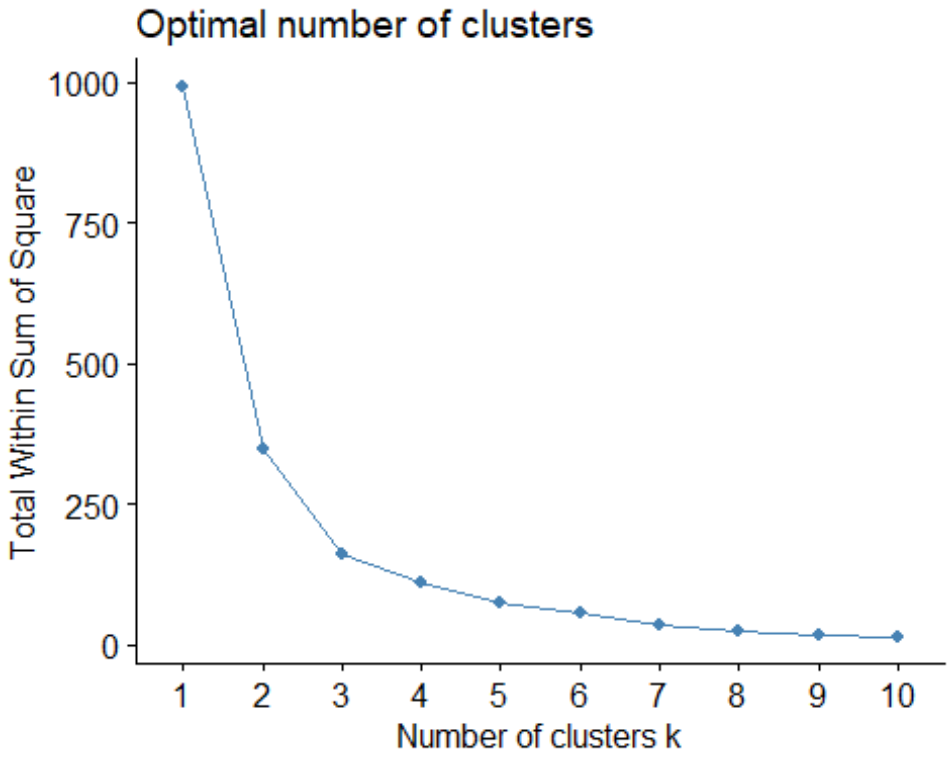
As we can see in our dendrogram, we now have four separate clusters (as we specified in our code). Now let's plot this in an x-y graph.

```
fviz_cluster(list(data = df, cluster = sub_grp))
```



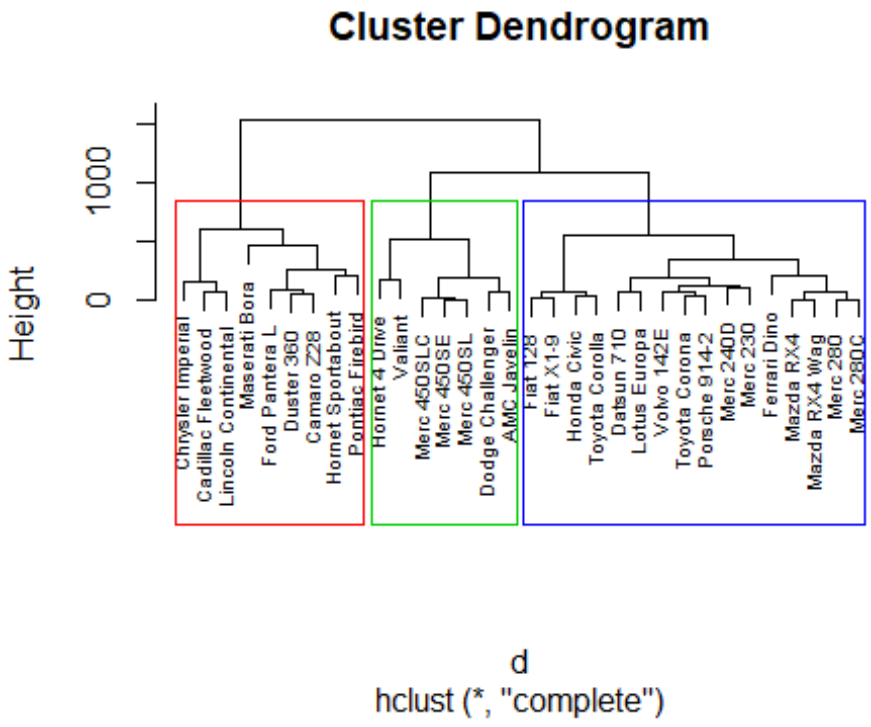
Here we can see the clusters in a two-dimensional plane. Now let's use use the Elbow Method to select the optimal number of clusters.

```
fviz_nbclust(df, FUN = hcut, method = "wss")
```



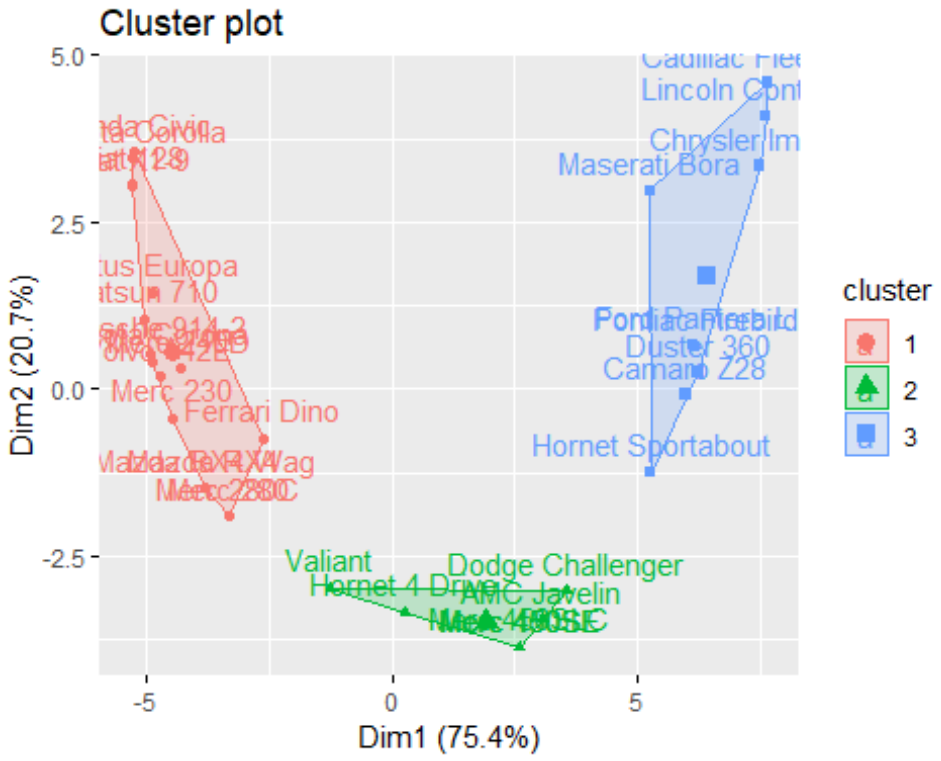
Using the Elbow Method, we can see that the optimal number of clusters is three. This is because the 'elbow' of the curve is at three along the x-axis. Now let's cut the dendrogram at three, as well.

```
hc5 <- hclust(d, method = "complete" )

sub_grp <- cutree(hc5, k = 3)

table(sub_grp)

## sub_grp
##  1  2  3
## 16  7  9

plot(hc5, cex = 0.6)
rect.hclust(hc5, k = 3, border = 2:5)
```



Now let's plot this in a two-dimensional plane.

```
fviz_cluster(list(data = df, cluster = sub_grp))
```


Cluster plot
Dim2 (20.7%)
Dim1 (75.4%)
5.0
2.5
0.0
-2.5
-5
0
5
Lincoln Cont
Chrysler Im
Maserati Bora
Duster 360
Camaro Z28
Hornet Sportabout
Merc 230
Ferrari Dino
Valiant
Dodge Challenger
AMC Javelin
cluster
1
2
3