Module 8 R Activity

Zahlen Zbinden

2023-11-27

Errors encountered

When trying to plot the dendrogram objects there is an error that occurs when trying to add labels, I can't track down what is causing the error, nor find a way to fix it, it happens when following the example from the Activity as well.

1. Read in the auto dataset and display the first six rows of the dataset.

```
auto82 <- read_csv("D:/RepoMan/osu/data/Auto82MPGData.csv")
```

```
Rows: 31 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): ModelName
dbl (6): MPG, Cylinders, Displacement, Horsepower, Weight, Acceleration

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

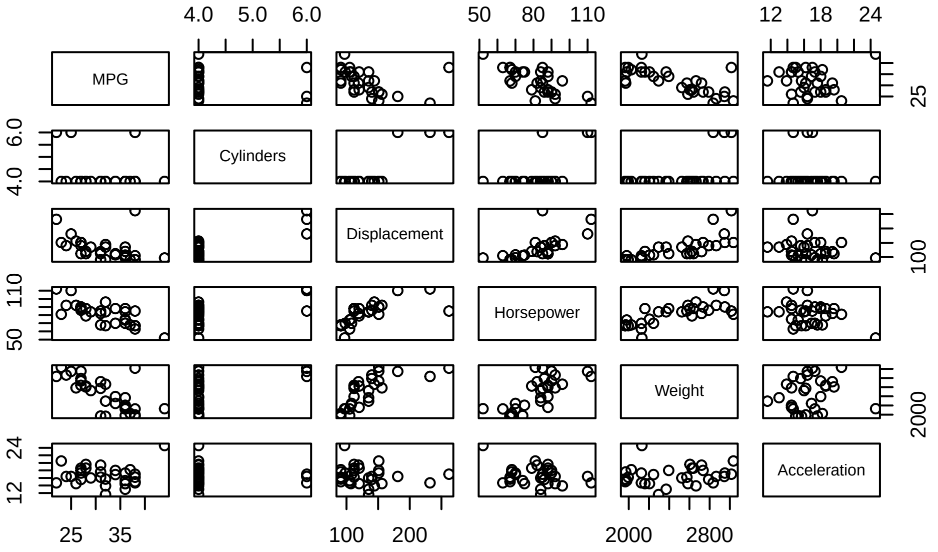
```
head(auto82, 6)
```

```
# A tibble: 6 x 7
    MPG Cylinders Displacement Horsepower Weight Acceleration ModelName
  <dbl>     <dbl>        <dbl>      <dbl>  <dbl>        <dbl> <chr>
1    28         4          112         88   2605         19.6 chevrolet cavalier
2    27         4          112         88   2640         18.6 chevrolet cavalie~
3    34         4          112         88   2395         18   chevrolet cavalie~
4    31         4          112         85   2575         16.2 pontiac j2000 se ~
5    29         4          135         84   2525         16   dodge aries se
6    27         4          151         90   2735         18   pontiac phoenix
```

2. Make a pairs plot of the 6 variables in the auto82 data set. Based on these plots, does it look like there are obvious clusters among these different car models?

There does appear to be some clustering among the car models, as we can see in the plots that the points in some of the pairs are grouped up, for example displacement vs horsepower.

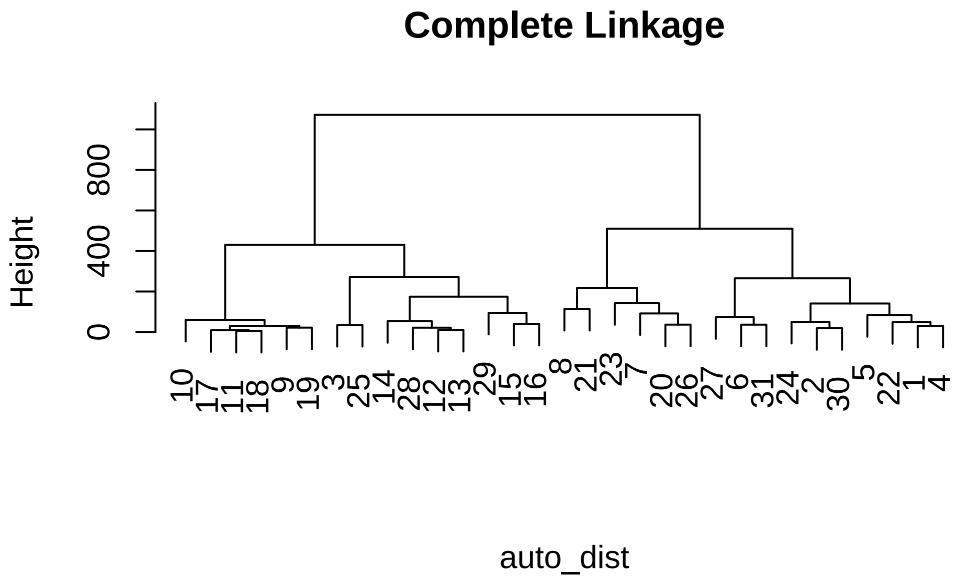
```r
pairs(auto82[1:6])
```



3. Perform heirarchical clustering using Euclidean distance on the unstandardized data, with complete linkage. Plot the resulting dendrogram with the car model names as labels. How many clusters would you say there are based on this dendrogram?

It looks like there are 4 major clusters.

```r
auto_dist <- dist(auto82[,-7]) # calculates euclidean distance between the rows
auto_hcEuc <- hclust(auto_dist, method = "complete")
# There is an error I can't figure out when trying to set the labels
# plot(auto_hcEuc, sub = "", main = "Complete Linkage", labels = auto82[, 7])
plot(auto_hcEuc, sub = "", main = "Complete Linkage")
```

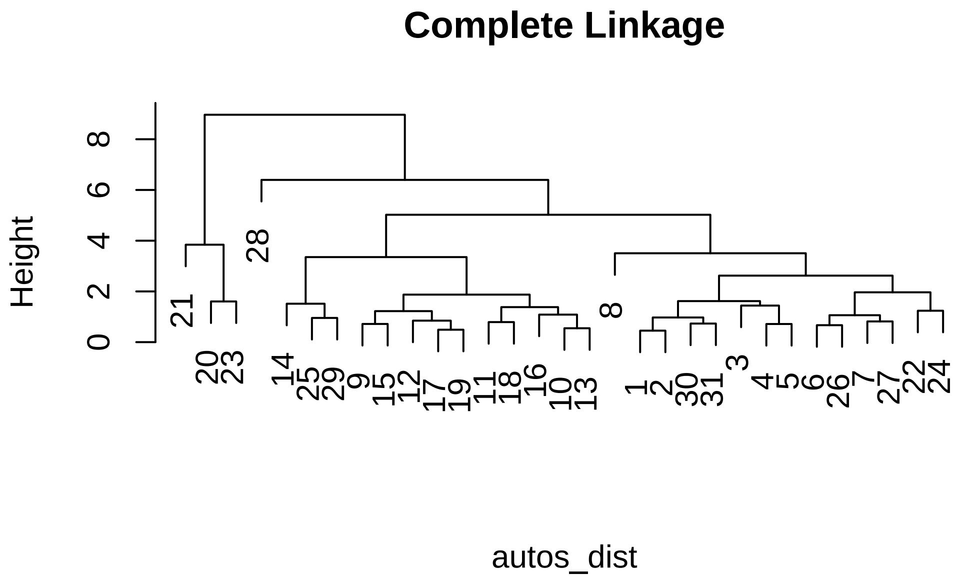


4. Perform the clustering using euclidean distance on the standardized data, with complete linkage. Does this dendrogram differ substantially from the unstandardized version?

Yes this dendrogram is completely different than the unstandardized version.

```
auto_scaled <- scale(auto82[,-7])
autos_dist <- dist(auto_scaled)
autos_hcEuc <- hclust(autos_dist, method = "complete")
```

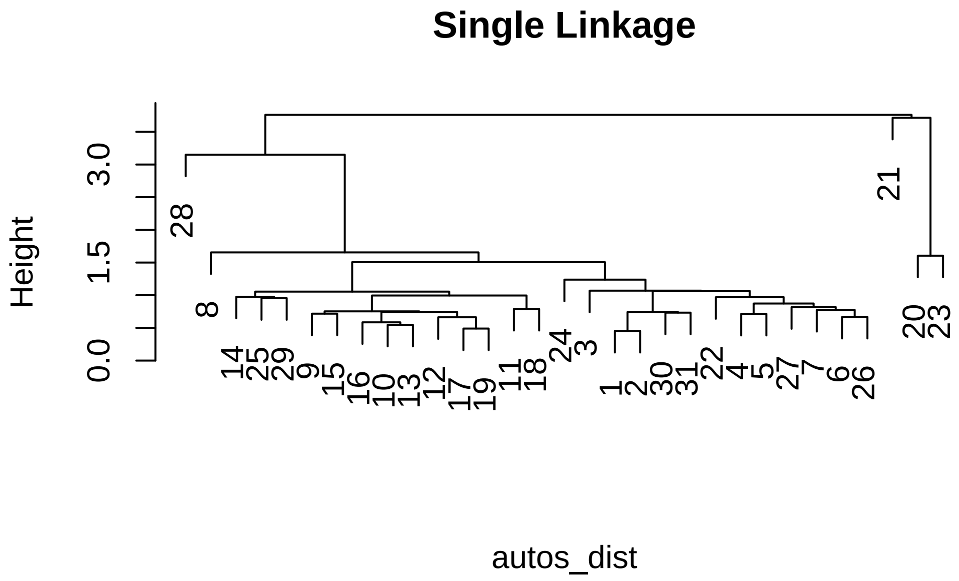
```
plot(autos_hcEuc, main = "Complete Linkage", sub = "")
```



5. Perfrom the clustering on the standardized data with single linkage. Does this dendrogram differ substantially from the version with complete linkage on the standardized data?

Yes this version differes substantially from the version with complete linkage.

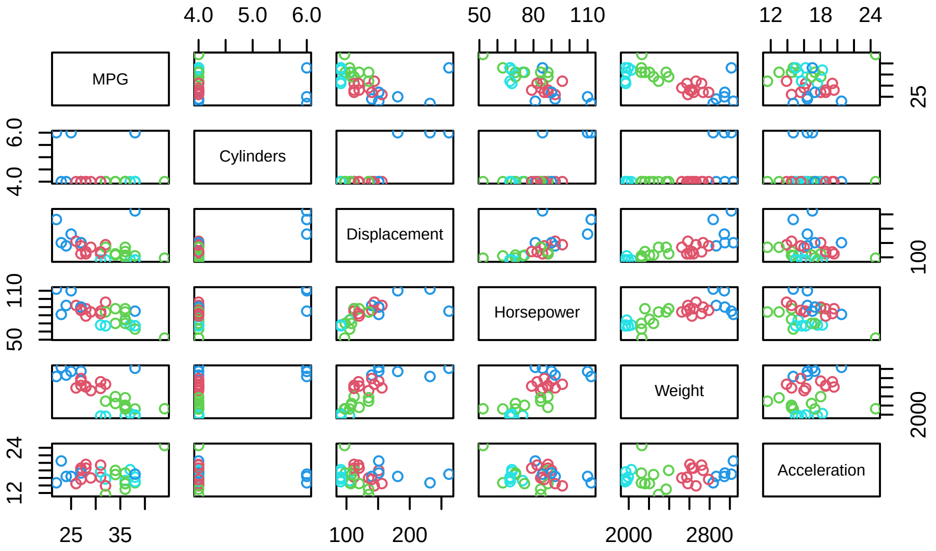
```
autos_hsEuc <- hclust(autos_dist, method = "single")
plot(autos_hsEuc, main = "Single Linkage", sub = "")
```



6. Cut the dendrogram based on the unstandardized data with euclidean distance complete linkage to produce 4 clusters. Are there any obvious patterns in these 4 clusters?

Yes there are some obvious patterns, especially with anything compared to weight.

```r
pairs(auto82[, -7], col = cutree(auto_hcEuc, k = 4) + 1)
```



7. Perform k-means clustering on the auto82 data to cluster the data into 4 clusters. Print the membership for each of the resulting clusters (note that your results may differ if you run this again, because of the random starts in the k means-algorithm)

```
auto_km4 <- kmeans(auto82[, -7], centers = 4, nstart = 10)
```

```
auto82[auto_km4$clus == 1, 7]
```

```
# A tibble: 9 x 1
  ModelName
  <chr>
1 chevrolet cavalier
2 chevrolet cavalier wagon
3 pontiac j2000 se hatchback
4 dodge aries se
5 pontiac phoenix
6 chrysler lebaron medallion
7 toyota celica gt
8 ford ranger
9 chevy s-10
```

```r
auto82[auto_km4$clus == 2, 7]
```

```
# A tibble: 10 x 1
   ModelName
   <chr>
 1 volkswagen rabbit l
 2 mazda glc custom l
 3 mazda glc custom
 4 plymouth horizon miser
 5 mercury lynx l
 6 nissan stanza xe
 7 honda civic
 8 honda civic (auto)
 9 datsun 310 gx
10 vw pickup
```

```r
auto82[auto_km4$clus == 3, 7]
```

```
# A tibble: 7 x 1
  ModelName
  <chr>
1 ford fairmont futura
2 amc concord dl
3 buick century limited
4 oldsmobile cutlass ciera (diesel)
5 ford granada l
6 chevrolet camaro
7 ford mustang gl
```

```r
auto82[auto_km4$clus == 4, 7]
```

```
# A tibble: 5 x 1
  ModelName
  <chr>
1 chevrolet cavalier 2-door
2 honda accord
3 toyota corolla
4 dodge charger 2.2
5 dodge rampage
```

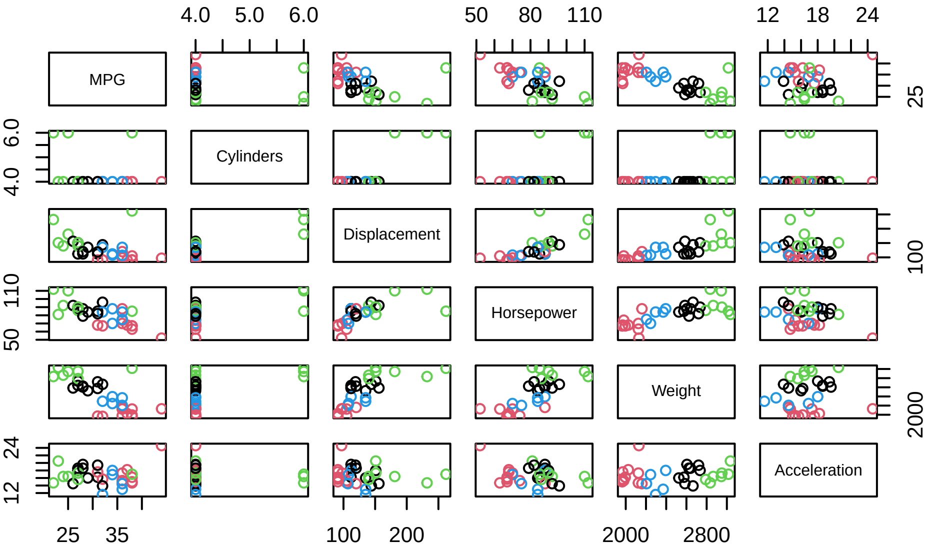
8. Make a table to compare the 4-cluster k-means clustering solution to the 4-cluster result from hierarchical clustering of the unstandardized data with Euclidean distance, complete linkage.

```
table(auto_km4$clus, cutree(auto_hcEuc, k = 4))
```

```
  1 2 3 4
1 9 0 0 0
2 0 4 0 6
3 1 0 6 0
4 0 5 0 0
```

9. Produce a pairs plot of the 6 variables for the auto82 data, with points colored based on their cluster assignment from your k-means solution.

```
pairs(auto82[, -7], col = auto_km4$clus)
```



10. Run model-based clustering on the auto82 data without specifying the number of clusters. How many clusters does the BIC select as optimal for this data?

This is set by the G argument and can be called from the object as object$G. This model chose 2 clusters.

```
auto_mc <- Mclust(auto82[,-7])
```

```
auto_mc$G
```

```
[1] 2
```

11. Run model-based clustering specifying 4 clusters, how does the 4-cluster model-based clustering solution compare to the 4-cluster k-means clustering solution?

They are fairly different, as we can see from the table, the groups that are being assigned by the model-based do not align with the groups being assigned by the k-means clustering solution.

```
auto_mc4 <- Mclust(auto82[, -7], G = 4)
```

```
table(auto_mc4$classification, auto_km4$clus)
```

```
   
    1 2 3 4
  1 7 0 4 1
  2 0 8 0 2
  3 2 1 0 2
  4 0 1 3 0
```