HW7

Getong Zhong

2023-04-14

1

1 - Correlation Matrix

```
data(USArrests)
data_centered <- scale(USArrests, center = TRUE, scale = TRUE)
cor <- cor(data_centered)
cor_dis <- 1 - cor
cor_dis
```

```
##             Murder   Assault  UrbanPop      Rape
## Murder   0.0000000 0.1981267 0.9304274 0.4364212
## Assault  0.1981267 0.0000000 0.7411283 0.3347588
## UrbanPop 0.9304274 0.7411283 0.0000000 0.5886588
## Rape     0.4364212 0.3347588 0.5886588 0.0000000
```

Proportional to the squared Euclidean distance

```
euc_sqrt <- as.matrix(dist(t(data_centered), method = "euclidean", diag = TRUE, upper = TRUE))^2
euc_sqrt/cor_dis
```

```
##          Murder Assault UrbanPop Rape
## Murder      NaN      98       98   98
## Assault      98     NaN       98   98
## UrbanPop     98      98      NaN   98
## Rape         98      98       98  NaN
```

From the matrix we can see that the proportionality holds for every variable.

2

Textbook formula:

```
cov <- cov(data_centered)
ev <- eigen(cov)$values
ptv1 <- ev / sum(ev)
ptv1
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

By R function

```
pca <- prcomp(data_centered)
sdev <- pca$sdev
ptv2 <- (sdev^2) / sum(sdev^2)
ptv2
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

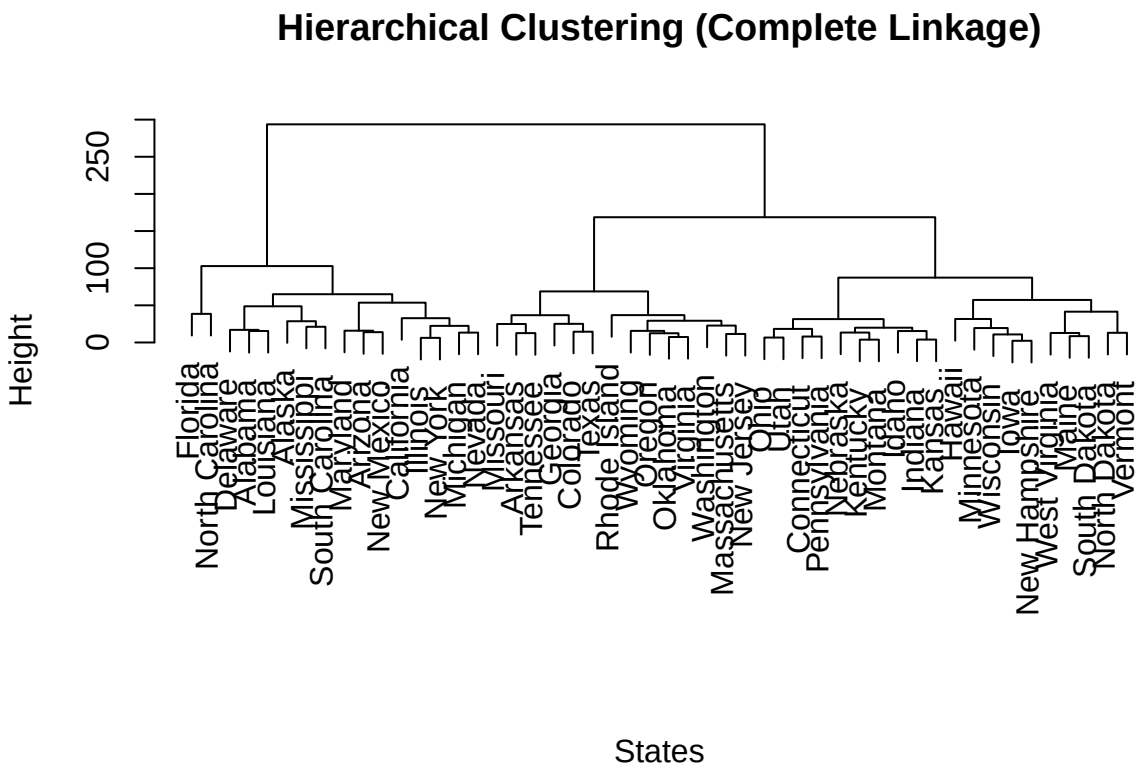
By two methods, we get the same results.

3

(a)

```
eucli_dist <- dist(USArrests, method = "euclidean")
hc <- hclust(eucli_dist, method = "complete")

# Plot the dendrogram
plot(hc, labels = row.names(USArrests), main = "Hierarchical Clustering (Complete Linkage)", xlab = "Sta
```



(b)

From the table we find that

```
clusters <- cutree(hc, k = 3)
states_clusters <- data.frame(State = row.names(USArrests), Cluster = clusters)
states_clusters
```

```
##                          State Cluster
## Alabama                Alabama       1
## Alaska                  Alaska       1
## Arizona                Arizona       1
## Arkansas              Arkansas       2
## California          California       1
## Colorado              Colorado       2
## Connecticut        Connecticut       3
## Delaware              Delaware       1
## Florida                Florida       1
## Georgia                Georgia       2
## Hawaii                  Hawaii       3
## Idaho                    Idaho       3
## Illinois              Illinois       1
## Indiana                Indiana       3
## Iowa                      Iowa       3
## Kansas                  Kansas       3
## Kentucky              Kentucky       3
## Louisiana            Louisiana       1
## Maine                    Maine       3
## Maryland              Maryland       1
## Massachusetts   Massachusetts       2
## Michigan              Michigan       1
## Minnesota            Minnesota       3
## Mississippi       Mississippi       1
## Missouri              Missouri       2
## Montana                Montana       3
## Nebraska              Nebraska       3
## Nevada                  Nevada       1
## New Hampshire   New Hampshire       3
## New Jersey          New Jersey       2
## New Mexico          New Mexico       1
## New York              New York       1
## North Carolina North Carolina       1
## North Dakota     North Dakota       3
## Ohio                     Ohio       3
## Oklahoma              Oklahoma       2
## Oregon                  Oregon       2
## Pennsylvania     Pennsylvania       3
## Rhode Island     Rhode Island       2
## South Carolina South Carolina       1
## South Dakota     South Dakota       3
## Tennessee           Tennessee       2
## Texas                   Texas       2
## Utah                     Utah       3
## Vermont                Vermont       3
```

```
## Virginia             Virginia       2
## Washington         Washington       2
## West Virginia   West Virginia       3
## Wisconsin           Wisconsin       3
## Wyoming               Wyoming       2
```

```
c1 <- states_clusters$State[states_clusters$Cluster == 1]
c2 <- states_clusters$State[states_clusters$Cluster == 2]
c3 <- states_clusters$State[states_clusters$Cluster == 3]
```

Cluster 1:

```
print(c1)
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "California"
##  [5] "Delaware"       "Florida"        "Illinois"       "Louisiana"
##  [9] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
## [13] "New Mexico"     "New York"       "North Carolina" "South Carolina"
```

Cluster 2:

```
print(c2)
```

```
##  [1] "Arkansas"      "Colorado"      "Georgia"       "Massachusetts"
##  [5] "Missouri"      "New Jersey"    "Oklahoma"      "Oregon"
##  [9] "Rhode Island"  "Tennessee"     "Texas"         "Virginia"
## [13] "Washington"    "Wyoming"
```

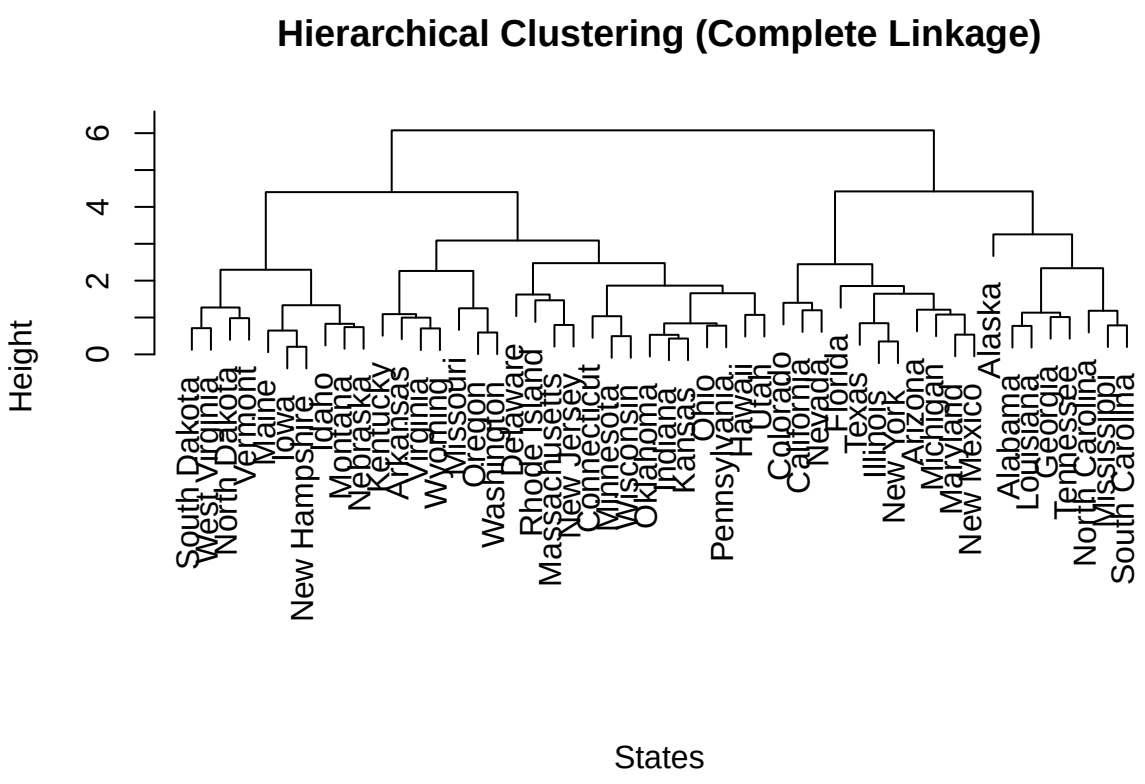
Cluster 3:

```
print(c3)
```

```
##  [1] "Connecticut"   "Hawaii"        "Idaho"         "Indiana"
##  [5] "Iowa"          "Kansas"        "Kentucky"      "Maine"
##  [9] "Minnesota"     "Montana"       "Nebraska"      "New Hampshire"
## [13] "North Dakota"  "Ohio"          "Pennsylvania"  "South Dakota"
## [17] "Utah"          "Vermont"       "West Virginia" "Wisconsin"
```

(c)

```
data_scaled <- scale(USArrests, center = FALSE, scale = apply(USArrests, 2, sd))
euc_dis <- dist(data_scaled, method = "euclidean")
hc <- hclust(euc_dis, method = "complete")
plot(hc, labels = row.names(USArrests), main = "Hierarchical Clustering (Complete Linkage)", xlab = "Sta
```



(d)

When we scale the variables to have a standard deviation of one, they are basically beening standardized, meaning that all the variables have equal importance when calculating the dissimilarity. Therefore, use scaled data for clustering can make the results more balanced and get more reasonable clustering. I think it is a good idea to scale the variables before clustering. First, we scaled the data before the clustering can make each variable equaly effective to the results, and reduce the risk that the results might affect by certain variable due the their larger scale. Second, when we scale the variable into same unit, we can make the comparison results more meaningful. However, we still need to keep in mind that the context of the problem is also important, we need to consider if scaled variable still make sense in the context before we do the scaling.