ST557: HOMEWORK 5

Brian Cervantes Alvarez
December 2, 2023

Question 1

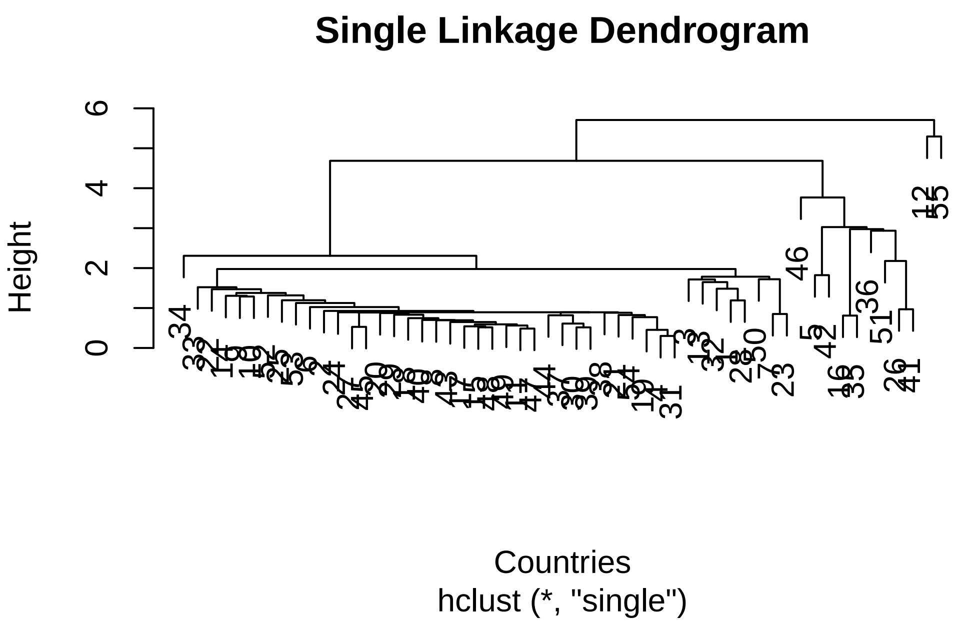
Part A

In the single linkage dendrogram, the main cluster gradually stretches vertically from left to right, hinting that points within it are getting farther apart, possibly because the algorithm is sensitive to small distances.

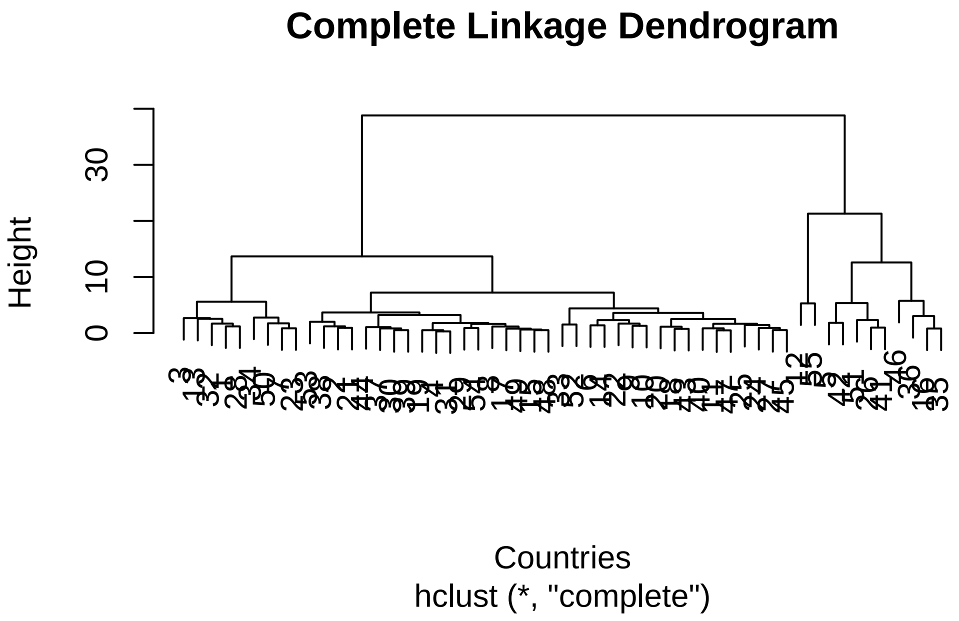
On the flip side, the complete linkage dendrogram shows three nicely leveled clusters, suggesting that it has identified solid groups. These leveled clusters mean that the dissimilarity within each cluster is consistent, and the merging stops when the distances between clusters are at a maximum.

Based on these results, the complete linkage dendrogram has successfully grouped the data into three distinct and internally similar clusters, making complete linkage a good choice; especially when dealing with the impact of outliers in this dataset.

```r
library(cluster)

trackData <- read.csv("TrackData.csv")

# Extract the columns
countryData <- trackData[, -c(1, 2)]

# Function to calculate Euclidean distances between countries
euclideanDistances <- dist(countryData, method = "euclidean")

# Hierarchical clustering using single and complete linkage
singleLinkageClusters <- hclust(euclideanDistances, method = "single")
completeLinkageClusters <- hclust(euclideanDistances, method = "complete")

# Plots
plot(singleLinkageClusters,
     main = "Single Linkage Dendrogram",
     xlab = "Countries")
```



```r
plot(completeLinkageClusters,
     main = "Complete Linkage Dendrogram",
     xlab = "Countries")
```



Part B

```r
# Perform k-means clustering for k = 2, 3, and 4
kMeans2 <- kmeans(countryData, centers = 2)
kMeans3 <- kmeans(countryData, centers = 3)
kMeans4 <- kmeans(countryData, centers = 4)


# Use this function to print out the clusters (very neat)
getClusterLists <- function(result, k) {
  clusterLists <- lapply(1:k, function(i) {
    countriesInCluster <- result$cluster == i
    return(trackData$Country[countriesInCluster])
  })
  names(clusterLists) <- paste0("Cluster", 1:k)
  return(clusterLists)
}

# Get lists of countries for each cluster
clusterLists2 <- getClusterLists(kMeans2, 2)
clusterLists3 <- getClusterLists(kMeans3, 3)
clusterLists4 <- getClusterLists(kMeans4, 4)

# Display lists of countries by cluster
print("K-means Clustering (k = 2):")
```

```
[1] "K-means Clustering (k = 2):"
```

```r
clusterLists2
```

```
$Cluster1
 [1] "Argentina"     "Australia"     "Austria"       "Belgium"
 [5] "Brazil"        "Burma"         "Canada"        "Chile"
 [9] "China"         "Columbia"      "CostaRica"     "Czechoslovakia"
[13] "Denmark"       "Finland"       "Grance"        "EastGermany"
[17] "WestGermany"   "GreatBritain"  "Greece"        "Guatemala"
[21] "Hungary"       "India"         "Ireland"       "Israel"
[25] "Italy"         "Japan"         "Kenya"         "SouthKorea"
[29] "NorthKorea"    "Luxembourg"    "Mexico"        "Netherlands"
[33] "NewZealand"    "Norway"        "Poland"        "Portugal"
[37] "Romania"       "Spain"         "Sweden"        "Switzerland"
[41] "Taiwan"        "Turkey"        "USA"           "Russia"
```

```
$Cluster2
 [1] "Bermuda"           "CookIslands"       "DominicanRepublic"
 [4] "Indonesia"         "Malaysia"          "Mauritius"
 [7] "PapuaNewGuinea"    "Philippines"       "Singapore"
[10] "Thailand"          "WestSamoa"
```

```
print("K-means Clustering (k = 3):")
```

```
[1] "K-means Clustering (k = 3):"
```

```
clusterLists3
```

```
$Cluster1
 [1] "Argentina"   "Austria"     "Bermuda"     "Burma"       "CostaRica"
 [6] "Guatemala"   "Israel"      "SouthKorea"  "Luxembourg"  "Philippines"
[11] "Taiwan"

$Cluster2
[1] "CookIslands"       "DominicanRepublic" "Indonesia"
[4] "Malaysia"          "Mauritius"         "PapuaNewGuinea"
[7] "Singapore"         "Thailand"          "WestSamoa"

$Cluster3
 [1] "Australia"      "Belgium"        "Brazil"         "Canada"
 [5] "Chile"          "China"          "Columbia"       "Czechoslovakia"
 [9] "Denmark"        "Finland"        "Grance"         "EastGermany"
[13] "WestGermany"    "GreatBritain"   "Greece"         "Hungary"
[17] "India"          "Ireland"        "Italy"          "Japan"
[21] "Kenya"          "NorthKorea"     "Mexico"         "Netherlands"
[25] "NewZealand"     "Norway"         "Poland"         "Portugal"
[29] "Romania"        "Spain"          "Sweden"         "Switzerland"
[33] "Turkey"         "USA"            "Russia"
```

```
print("K-means Clustering (k = 4):")
```

```
[1] "K-means Clustering (k = 4):"
```

```
clusterLists4
```

```
$Cluster1
[1] "Argentina"   "Bermuda"     "Burma"       "CostaRica"   "Guatemala"
[6] "Israel"      "Luxembourg"  "Philippines" "Taiwan"
```

```
$Cluster2
[1] "CookIslands"       "DominicanRepublic" "Indonesia"
[4] "Malaysia"          "Mauritius"         "PapuaNewGuinea"
[7] "Singapore"         "Thailand"          "WestSamoa"

$Cluster3
 [1] "Australia"     "Belgium"       "Canada"        "Denmark"       "Finland"
 [6] "EastGermany"   "GreatBritain"  "Italy"         "Japan"         "Kenya"
[11] "Mexico"        "Netherlands"   "NewZealand"    "Portugal"      "Sweden"
[16] "Switzerland"   "USA"           "Russia"

$Cluster4
 [1] "Austria"         "Brazil"          "Chile"           "China"
 [5] "Columbia"        "Czechoslovakia"  "Grance"          "WestGermany"
 [9] "Greece"          "Hungary"         "India"           "Ireland"
[13] "SouthKorea"      "NorthKorea"      "Norway"          "Poland"
[17] "Romania"         "Spain"           "Turkey"
```

Part C

I prefer using K-means clustering for this data because it offers a simpler and more direct way to understand how the data is grouped. To me, K-means provides clear clusters and has made it easier to interpret why certain countries are grouped together. For example, if I took the extra time to figure out relationships between the countries that are grouped, it would be more insightful than the hierarchical system.

I considered hierarchical clustering, particularly complete linkage, but K-means seemed more straightforward. The method's efficiency makes it a better choice, especially since I don't need the hierarchical relationships captured by methods like complete linkage. Though, that could just be this specific dataset, after all.

```
# Compare k-means clustering with hierarchical clustering
print("Hierarchical Clustering - Single Linkage:")
```

```
[1] "Hierarchical Clustering - Single Linkage:"
```

```
print(cutree(singleLinkageClusters, k = 2))
```

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
```

```
print(cutree(singleLinkageClusters, k = 3))
```

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3
```

```
print(cutree(singleLinkageClusters, k = 4))
```

```
 [1] 1 1 1 1 2 1 1 1 1 1 1 3 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 1 1
[39] 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 4
```

```
print("Hierarchical Clustering - Complete Linkage:")
```

```
[1] "Hierarchical Clustering - Complete Linkage:"
```

```
print(cutree(completeLinkageClusters, k = 2))
```

```
 [1] 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 1 1
[39] 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 2
```

```
print(cutree(completeLinkageClusters, k = 3))
```

```
 [1] 1 1 1 1 2 1 1 1 1 1 1 3 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 1 1
[39] 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 3
```

```
print(cutree(completeLinkageClusters, k = 4))
```

```
 [1] 1 2 1 2 3 2 1 2 2 2 2 4 1 2 2 3 2 2 2 2 2 2 1 2 2 3 2 1 2 2 2 1 2 1 3 3 2 2
[39] 2 2 3 3 2 2 2 3 2 2 2 1 3 2 2 2 4
```

```
print("K-means Clustering:")
```

```
[1] "K-means Clustering:"
```

```
print(kMeans2$cluster)
```

```
 [1] 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 1 1
[39] 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 2
```

```
print(kMeans3$cluster)
```

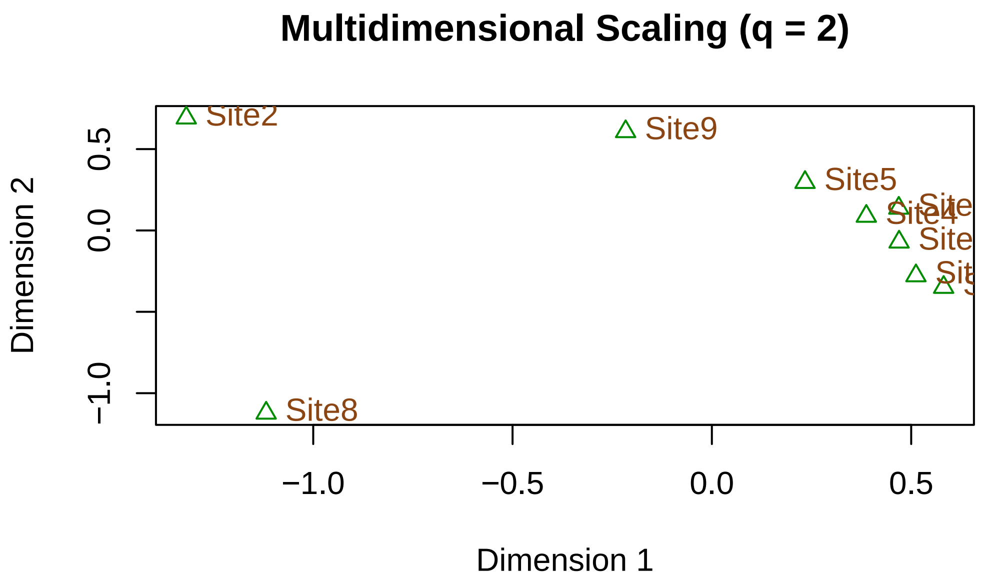
```
 [1] 1 3 1 3 1 3 1 3 3 3 3 2 1 3 3 2 3 3 3 3 3 3 1 3 3 2 3 1 3 3 3 1 3 1 2 2 3 3
[39] 3 3 2 1 3 3 3 2 3 3 3 1 2 3 3 3 2
```

```
print(kMeans4$cluster)
```

```
 [1] 1 3 4 3 1 4 1 3 4 4 4 2 1 4 3 2 3 4 3 4 3 4 1 4 4 2 4 1 3 3 3 4 4 1 2 2 3 3
[39] 3 4 2 1 4 3 4 2 4 3 3 1 2 4 3 3 2
```

Question 2

```r
archaeo <- read.csv("ArchaeoData.csv",
                 header = TRUE,
                 row.names = 1)

distances <- as.matrix(archaeo)

# Perform multidimensional scaling for q = 2
mds_result <- cmdscale(distances, k = 2)

# Plot the coordinates for q = 2
plot(mds_result, type = "n",
     xlab = "Dimension 1",
     ylab = "Dimension 2",
     main = "Multidimensional Scaling (q = 2)")

# Add points with labels (not perfect but it works)
points(mds_result, col = "green4", pch = 2)
text(mds_result, labels = rownames(mds_result),
     pos = 4, col = "chocolate4")
```



Question 3

Part A

Scenario: Twenty subjects were given each of three diets (in random order) and the subjects' blood pressures were measured at the end of each diet, so there were three blood pressure measurement associated with each subject.

Question: Did the different treatments affect the subjects' blood pressure different

- **Answer:** ANOVA.

Part B

Scenario: Two varieties of chickweed are difficult to distinguish. Measurements on four variables were obtained for chickweed plants whose variety was known.

Question: Use these observations to establish a rule for classifying a new candidate plant into one of the two varieties

- **Answer:** Discriminant Function Analysis/Linear Discriminant Analysis

Part C

Scenario: Each of 50 eight-year-old girls and 50 eight-year-old boys were given a total of 10 tests. Five of these tests had to do with language and five had to do with mathematical reasoning.

Question: Do scores differ between boys and girls?

- **Answer:** Univariate two-sample t-test

Question: Combining the boys and girls, what combination of the language tests is most associated with some combination of the math tests

- **Answer:** MANOVA

Part D

Scenario: Daily measurements of seven pollution-related variables were recorded over an extended period of time at a single location in Los Angeles.

Question: Find a low-dimensional representation for these variables that captures most of the variability.

- **Answer:** Principal Components Analysis

Question: Test whether the pollution on weekends differed from that on weekday

- **Answer:** Univariate two-sample t-test

Part E

Scenario: For each of a sample of 42 new microwaves made by a certain manufacturer, the amount of radiation emitted when the door of the microwave is closed and the amount of radiation emitted when the door of the microwave is opened are measured.

Question: Construct a confidence interval for the difference in amount of radiation emitted under these two conditions.

- **Answer:** Univariate one-sample t-test (paired differences)

Part F

Scenario: A sample of 50 married couples was obtained. The wife and the husband each answered four questions regarding their relationship on a scale of 0 to 10.

Question: Do the wife's answers tend to be similar to the husband's answers, and in what way are they most similar? That is, what combination of the wife's answers is most similar to what combination of the husband's answers

- **Answer:** Canonical Correlation Analysis

Part G

Scenario: The standardized scores for each of the ten events in the decathlon were obtained for each of 50 entrants.

Question: Can the variation in the scores be explained by three underlying athletic abilities, and how might these abilities be described

- **Answer:** Principal Component Analysis

Part H

Scenario: For 15 different species of predator fish, data were gathered on several aspects of their diet.

Question: How can these species of fish be grouped based on similarities in their diet

- **Answer:** Clustering

Part I

Scenario: Calcite content was measured at 25 equally-spaced locations along the leg bone for each of seven Tyrannosaurus Rex skeletons and also for each of five skeletons of a newly-discovered type of dinosaur.

Question: Do the calcite concentrations at these locations differ between the two dinosaur species?

- **Answer:** Hotelling's two-sample T2 test

Question: Combining the dinosaur species, is calcite concentration the same at all of the measured locations in the leg bone?

- **Answer:** Hotelling's one-sample T2 test (repeated measures)

Question: Based on these measurements, construct a rule for classifying a new bone as coming from a Tyrannosaurus Rex or from the newly-discovered specie

- **Answer:** Discriminant Function Analysis/Linear Discriminant Analysis

Part J

Scenario: Blood samples from 40 patients were obtained and each divided into six subsamples, which were sent to six different laboratories to have iron content measured.

Question: Do the six different laboratory results have the same means?

- **Answer:** ANOVA

Part K

Scenario: Measurements on six accounting and financial variables were obtained from a sample of insurance companies that were distressed (close to bankrupt) and an independent sample of insurance companies that were solvent.

Question: Establish a rule for classifying future insurance companies as solvent or distressed based on these variables

- **Answer:** Discriminant Function Analysis/Linear Discriminant Analysis

Part L

Scenario: DNA analysis was performed on hair specimens from each of 100 mummies taken from Egyptian pyramids. For each mummy, twenty variables concerning the DNA sequence were measured.

Question: Based on the measured variables, identify groups of mummies that are related to each other (have similar values of the variables).

- **Answer:** Clustering

Question: Based on the distances between these variables, construct a two-dimensional plot of the mummies to visualize the grouping

- **Answer:** Multidimensional Scaling

Part M

Scenario: SAT subject test scores are obtained for a random sample of 100 12th graders who took Math, Biology, Literature, and World History subject tests.

Question: Test whether the average score for all four tests is 500.

- **Answer:** Univariate one-sample t-test

Question: Test whether the average scores are equal for all four tests

- **Answer:** ANOVA

Part N

Scenario: A wildlife ecologist measured tail length and wing length for a sample of 45 female hook-billed kites and 45 male hook-billed kites.

Question: Are average tail length and wing length the same for female and male hook-billed kites?

- **Answer:** Univariate two-sample t-test

Part 0

Scenario: Several measurements were obtained on chief executive officers (CEO) of companies, regarding the degree to which the officers took risks. Several additional measurements were available on the success of the company under their leadership.

Question: What aspects of risk-taking propensity of the CEO are associated with which aspects of company success?

- **Answer:** Canonical Correlation Analysis

Question: What combination of risk-taking propensities displays the greatest variation between CEOs?

- **Answer:** Discriminant Function Analysis/Linear Discriminant Analysis

Part P

Scenario: The age, diameter, and height were measured for a sample of trees that contained eagle roost sites and for an independent sample of trees that did not contain eagle roost sites.

Question: Construct confidence intervals for the difference in age, difference in diameter, and difference in height between roosting trees and non-roosting trees.

- **Answer:** Bonferroni simultaneous tests

Question: Determine a rule for classifying a new tree as a likely roosting site or unlikely roosting site, based on these three variable.

- **Answer:** Discriminant Function Analysis/Linear Discriminant Analysis

Part Q

Scenario: For all of the NBA rookies who started in 2000, data were collected on their free-throw percentages each year for the first five years of their NBA careers.

Question: Does average free-throw percentage change over these five year

- **Answer:** ANOVA

Part R

Scenario: Twelve measurements were taken on fossilized skull measurements from 20 kinds of squirrels. The goal of the analysis was to order the 20 squirrels chronologically, on the basis of the similarities between the skull measurements for different squirrels.

Question: Find a one-dimensional representation of the 20 squirrels that best captures the differences between the measured variables

- **Answer:** Principal Components Analysis

Part S

Scenario: The protein, fat content, calories, and Vitamin A content were measured for each of ten brands of hot dogs.

Question: Group the brands of hot dogs based on their nutritional content.

- **Answer:** Clustering

Question: What combination of these nutritional measurements captures the greatest difference between the hot dog brands?

- **Answer:** Principal Components Analysis

Part T

Scenario: Measurements were obtained on five pre-college predictor variables and four college performance variables for each of several hundred students.

Question: What combination of pre-college variables is most associated with a combination of college performance?

- **Answer:** Canonical Correlation Analysis

Question: Combining the two variable sets, are there a few underlying abilities that explain the pre-college and college performance?

- **Answer:** Factor Analysis