BA 64060 - Assignment 5

Disney Maxwell

2025-11-22

Install packages

```
library("stats")
library("cluster")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("caret")
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

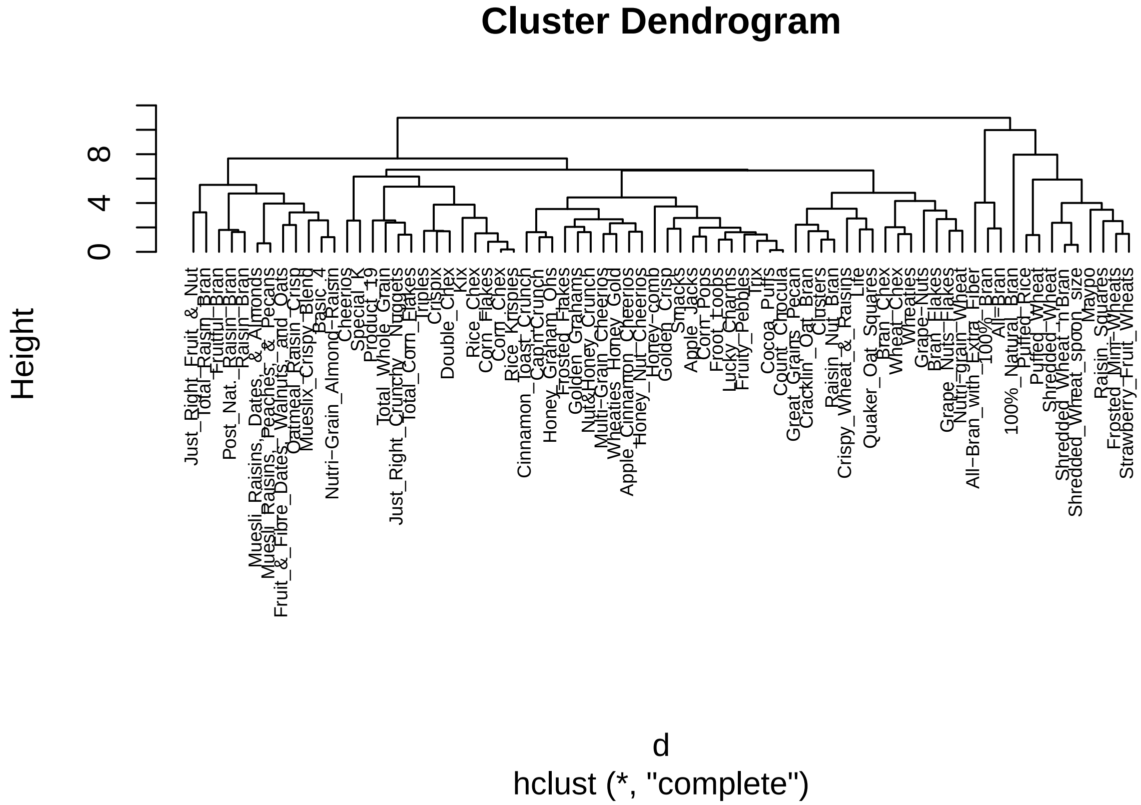
```
library("factoextra")
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
cereal_brands_init <- read.csv("Cereals.csv")

# Data Preprocessing - Remove NA (missing) values
cereal_brands_data <- na.omit(cereal_brands_init)
```

Question 1

```
set.seed(123)
# Convert the Name column to row names
rownames(cereal_brands_data) <- cereal_brands_data$name

# Remove non-numeric columns - Name, Mfr and Type
cereal_brands_data <- cereal_brands_data %>% select(-"name", -"mfr", -"type")

#Normalization
cereal_brands_norm<-scale(cereal_brands_data)

#Applying Hierarchical clustering to Data using Euclidean distance

# Dissimilarity matrix
d <- dist(cereal_brands_norm, method = "euclidean")

# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete")

# Plot the obtained dendrogram
plot(hc1, cex=0.6, hang=-1)
```



```
# Compute with Agnes and with different linkage methods
hc_single <- agnes(cereal_brands_norm, method = "single")
hc_complete <- agnes(cereal_brands_norm, method = "complete")
hc_average <- agnes(cereal_brands_norm, method = "average")
```

```r
hc_ward <- agnes(cereal_brands_norm, method = "ward")

# Compare Agglomerative coefficients
print(hc_single$ac)
```

```
## [1] 0.6067859
```

```r
print(hc_complete$ac)
```

```
## [1] 0.8353712
```

```r
print(hc_average$ac)
```

```
## [1] 0.7766075
```

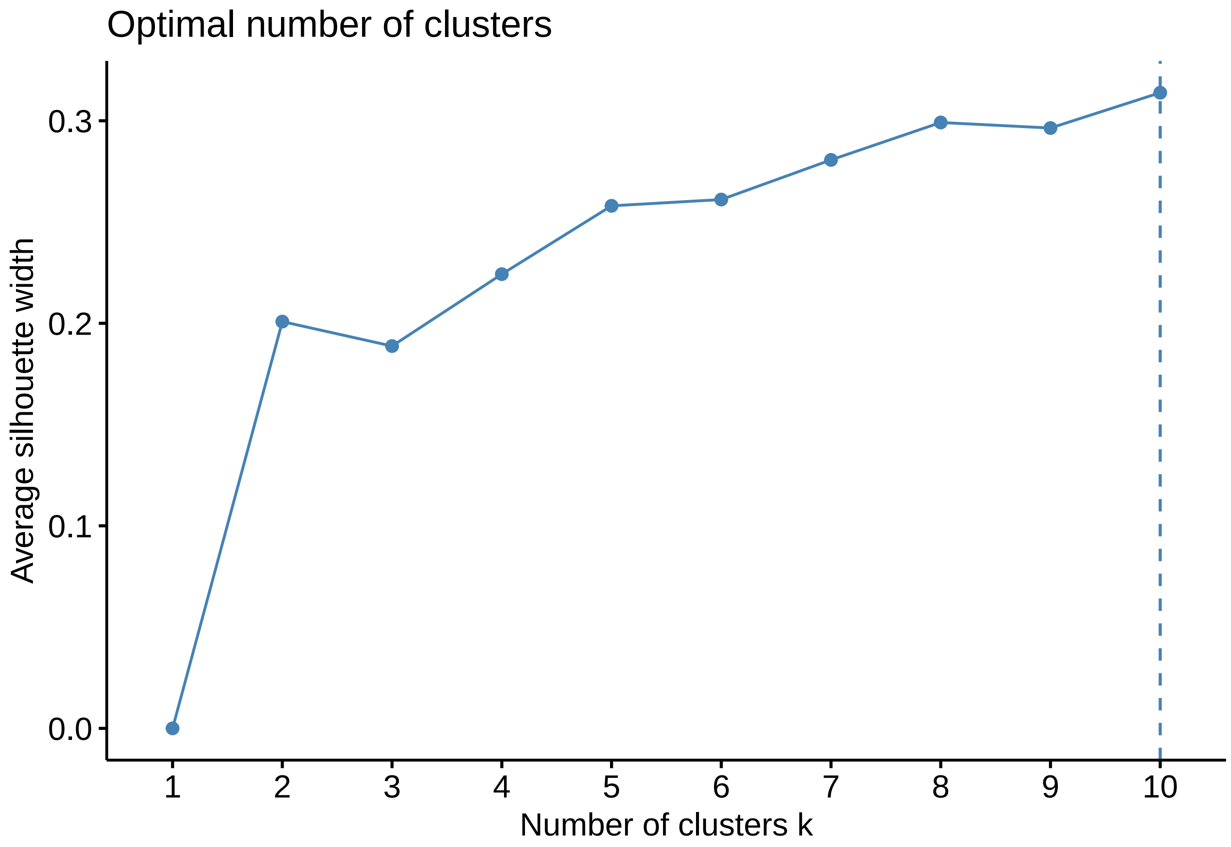
```r
print(hc_ward$ac)
```

```
## [1] 0.9046042
```

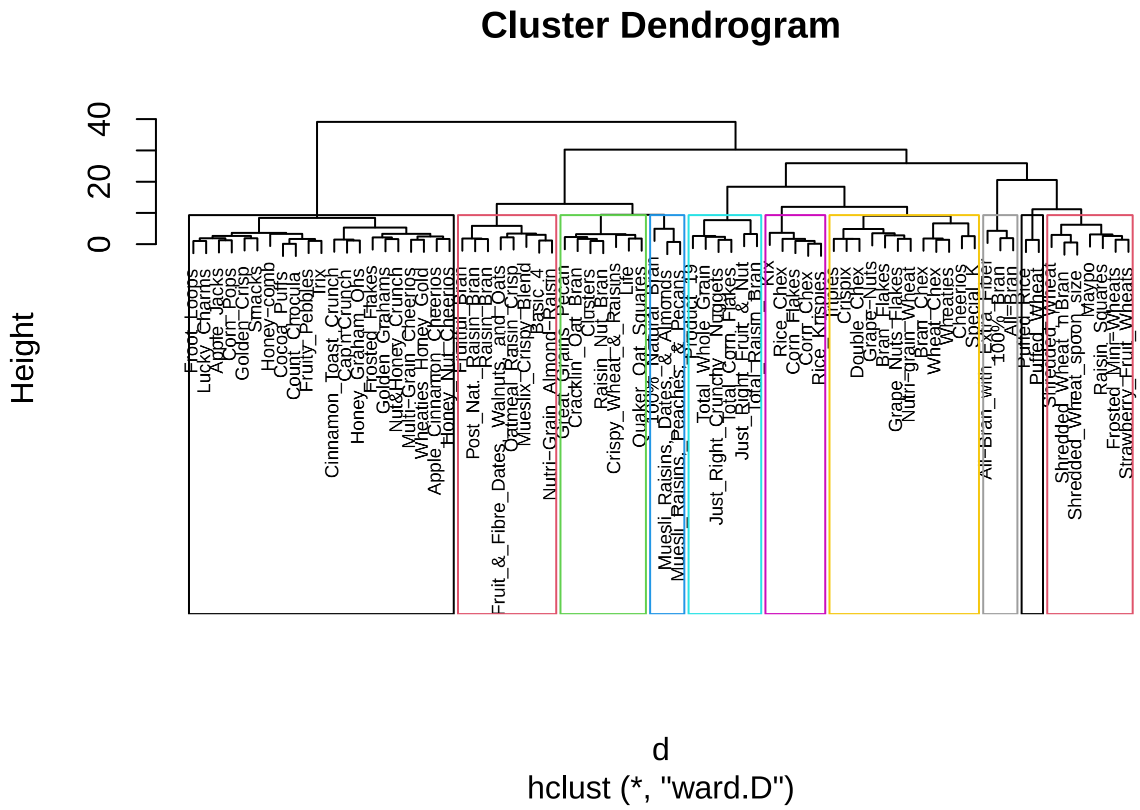
```r
# The Ward method gives the highest Agglomerative coefficient (0.9046042). Hence, this is the best link
```

Question 2 - Number of Clusters

```r
# Use Silhouette method to find optimal k
fviz_nbclust(cereal_brands_norm, hcut, method="silhouette")
```



```
# From Silhouette method, k=10 gives the ideal number of clusters.

# Compute Divisive hierarchical clustering using the Ward method
hc1_ward <- hclust(d, method = "ward.D")

# Plot dendrogram using k = 10
plot(hc1_ward, cex=0.6)
rect.hclust(hc1_ward, k=10, border = 1:10)
```



Question 3 - Comment on structure of clusters and stability

```
# Cluster 1 consists of Cereals with high sugar, similar calories and vitamins.
# Cluster 2 consists of Cereals with medium calories and high potassium.
# Cluster 3 consists of Cereals medium calories.
# Cluster 4 consists of Cereals with medium fiber.
# Cluster 5 consists of Cereals with high Vitamins.
# Cluster 6 consists of Cereals with low fiber and low sugar.
# Cluster 7 consists of Cereals with high sodium.
# Cluster 8 consists of Cereals with highest fiber and lowest calories.
# Cluster 9 consists of Cereals with low calories, low protein, low carbs, zero sugar and vitamins.
# Cluster 10 consists of Cereals with the second highest group of ratings.

# Convert to Data frame
cereal_brands_norm_df <- as.data.frame(cereal_brands_norm)

# Partition Data into two Clusters (A and B)
Train_Index <- createDataPartition(cereal_brands_norm_df$calories, p=0.5, list=FALSE)
A <- cereal_brands_norm_df[Train_Index, ]
B <- cereal_brands_norm_df[-Train_Index, ]

# Partition Cluster A using Ward method and k = 10
agnes_A <- agnes(A, method = "ward")
```

```
clusters_A <- cutree(as.hclust(agnes_A), k = 10)

# Calculate cluster centroids in A
centroids_A <- aggregate(A, by = list(cluster = clusters_A), FUN = mean)
```

Question 4 - Healthy Cereals

```
# The data does not need to be normalized for this step.
# In the Cluster analysis, it is seen that Cluster 8 which has cereals with the highest fiber and low c
# the group of Healthy Cereals. This group also has the highest average rating.
```