Assignment 5 Ireland Kuhn

cereals <- read.csv("C:\\Users\\Ireland\\Downloads\\Cereals.csv")  
library(readr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

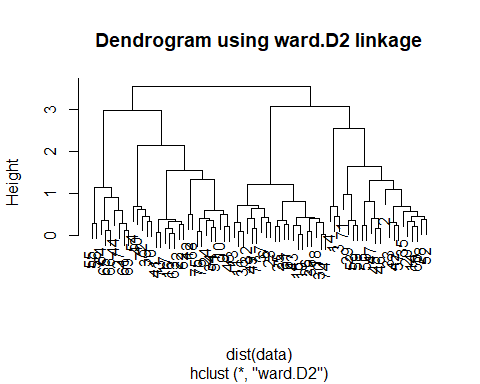
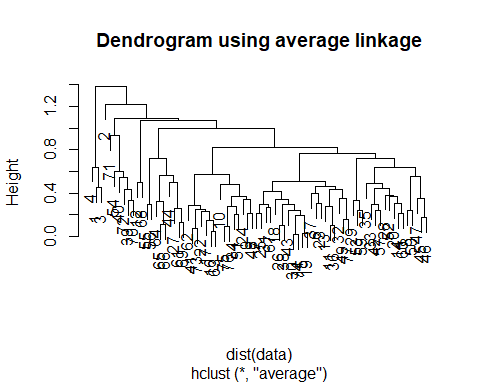
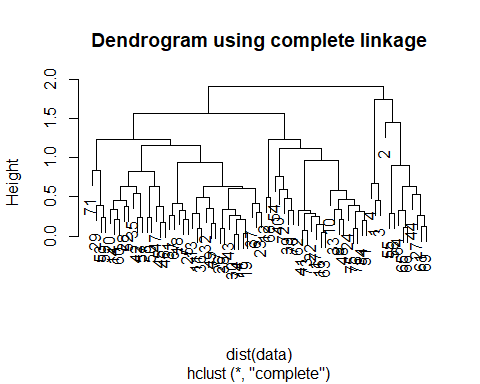
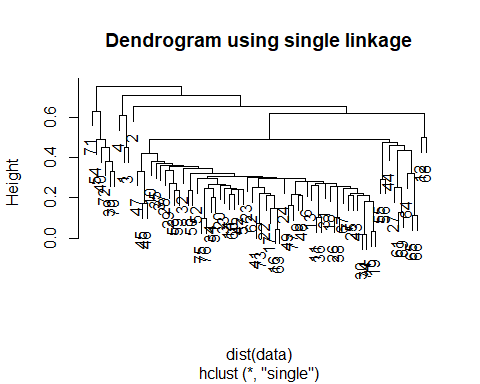
## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

library(ggplot2)  
library(cluster)

cereals\_clean <- na.omit(cereals)  
  
normalize <- function(x) {  
 return ((x - min(x)) / (max(x) - min(x)))  
}  
  
nutritional\_columns <- c("calories", "protein", "fat", "sodium", "fiber", "carbo", "sugars", "potass", "vitamins")  
cereals\_clean[nutritional\_columns] <- lapply(cereals\_clean[nutritional\_columns], normalize)

methods <- c("single", "complete", "average", "ward.D2")  
  
perform\_clustering <- function(method, data) {  
 hc <- hclust(dist(data), method = method)  
 plot(hc, main = paste("Dendrogram using", method, "linkage"))  
}  
  
for (method in methods) {  
 perform\_clustering(method, cereals\_clean[nutritional\_columns])  
}



num\_clusters <- 4  
  
clusters\_ward <- cutree(hclust(dist(cereals\_clean[nutritional\_columns]), method = "ward.D2"), k = num\_clusters)  
  
cereals\_clean$Cluster\_Ward <- clusters\_ward

hc\_full <- hclust(dist(cereals\_clean[nutritional\_columns]), method = "ward.D2")  
clusters\_full <- cutree(hc\_full, k = num\_clusters)  
cereals\_clean$Cluster\_Full <- clusters\_full  
  
set.seed(42)  
train\_indices <- sample(1:nrow(cereals\_clean), size = nrow(cereals\_clean) / 2)  
train\_data <- cereals\_clean[train\_indices, ]  
test\_data <- cereals\_clean[-train\_indices, ]

train\_centroids <- aggregate(train\_data[nutritional\_columns], by=list(Cluster=train\_data$Cluster\_Full), mean)  
  
find\_closest\_centroid <- function(record, centroids) {  
 distances <- apply(centroids[, -1], 1, function(centroid) sum((record - centroid)^2))  
 return(which.min(distances))  
}  
  
test\_data$Cluster\_Assigned <- apply(test\_data[nutritional\_columns], 1, function(x) find\_closest\_centroid(x, train\_centroids))  
  
original\_test\_clusters <- test\_data$Cluster\_Full  
  
consistency <- mean(test\_data$Cluster\_Assigned == original\_test\_clusters)  
consistency\_percentage <- consistency \* 100  
print(paste("Consistency: ", consistency\_percentage, "%"))

## [1] "Consistency: 97.2972972972973 %"

# A consistency of approximately 97.3% means that when the data was split into two parts (training and testing), and clustering was performed separately on these parts, about 97.3% of the cereals in the testing set were assigned to the same clusters as they were when the entire dataset was used for clustering. This indicates that my hierarchical clustering model using Ward's method has done a good job at reliably grouping cereals based on their nutritional profiles. This result provides strong support for the choices and recommendations one might make based on these clusters.

cluster\_summary\_ward <- aggregate(cereals\_clean[nutritional\_columns], by=list(Cluster=cereals\_clean$Cluster\_Ward), mean)  
print(cluster\_summary\_ward)

## Cluster calories protein fat sodium fiber carbo  
## 1 1 0.5909091 0.4636364 0.36363636 0.491477273 0.30519481 0.4368687  
## 2 2 0.5541126 0.1047619 0.20000000 0.526785714 0.04081633 0.4126984  
## 3 3 0.5041322 0.3363636 0.10909091 0.710227273 0.11688312 0.7525253  
## 4 4 0.2929293 0.2888889 0.02222222 0.005208333 0.15079365 0.5740741  
## sugars potass vitamins  
## 1 0.5575758 0.5050505 0.2727273  
## 2 0.7650794 0.1027967 0.2500000  
## 3 0.2424242 0.1904762 0.4204545  
## 4 0.1555556 0.2398589 0.1111111

# Cluster 4: Lower in calories, fat, sodium, and sugar, moderate protein, higher fiber content than Clusters 2 and 3.  
  
# Cluster 3: Moderate in calories, protein, sugar and fiber content, higher sodium content compared to other clusters, lower potassium and vitamin levels.  
  
# Cluster 2: Higher in calories,lower protein and fiber content, highest fat content and sugar among all clusters, moderate sodium content.  
  
# Cluster 1: Higher in calories and fat, highest protein, fiber, potassium and vitamins, moderate sodium content, moderate to high sugar content.

healthy\_cluster <- cluster\_summary\_ward[cluster\_summary\_ward$Cluster == 4, ]  
print(healthy\_cluster)

## Cluster calories protein fat sodium fiber carbo  
## 4 4 0.2929293 0.2888889 0.02222222 0.005208333 0.1507937 0.5740741  
## sugars potass vitamins  
## 4 0.1555556 0.2398589 0.1111111

# Listing cereals in the healthy cluster  
healthy\_cereals <- cereals\_clean[cereals\_clean$Cluster\_Ward == 4, ]  
print(healthy\_cereals$name)

## [1] "Frosted\_Mini-Wheats" "Maypo"   
## [3] "Puffed\_Rice" "Puffed\_Wheat"   
## [5] "Raisin\_Squares" "Shredded\_Wheat"   
## [7] "Shredded\_Wheat\_'n'Bran" "Shredded\_Wheat\_spoon\_size"  
## [9] "Strawberry\_Fruit\_Wheats"