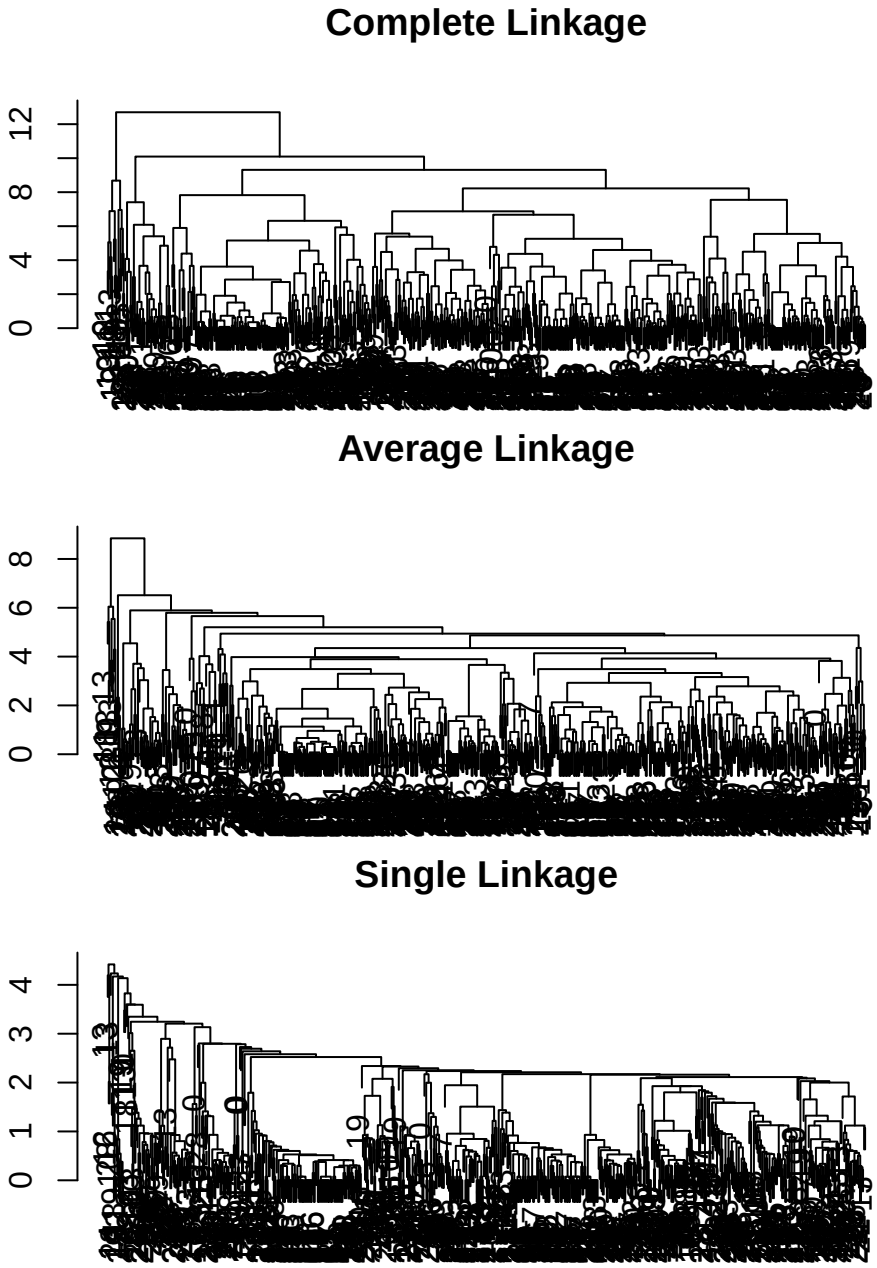
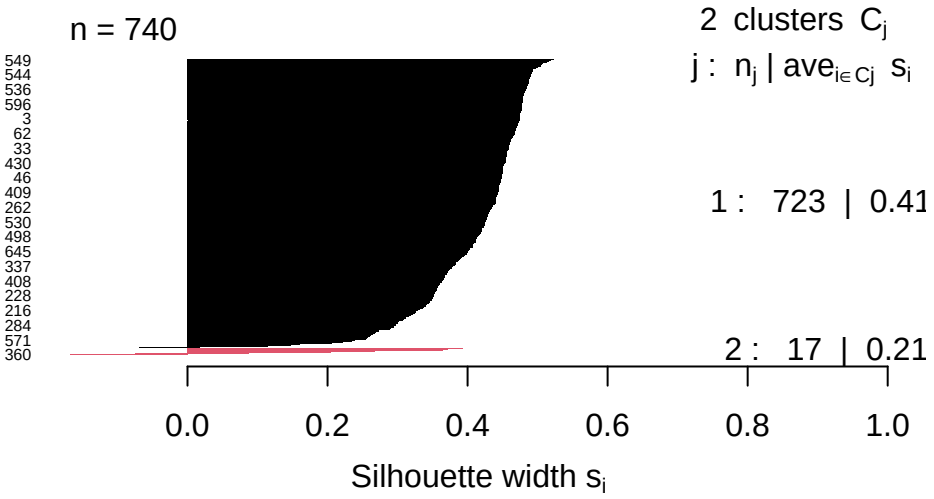
STATS/CSE 780
Assignment 3

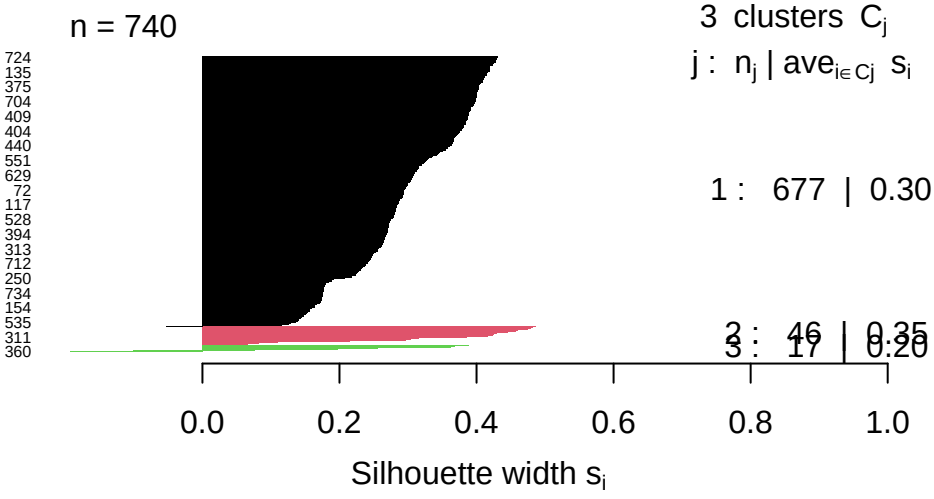
Pao Zhu Vivian Hsu (Student Number: 400547994)

2023-11-06

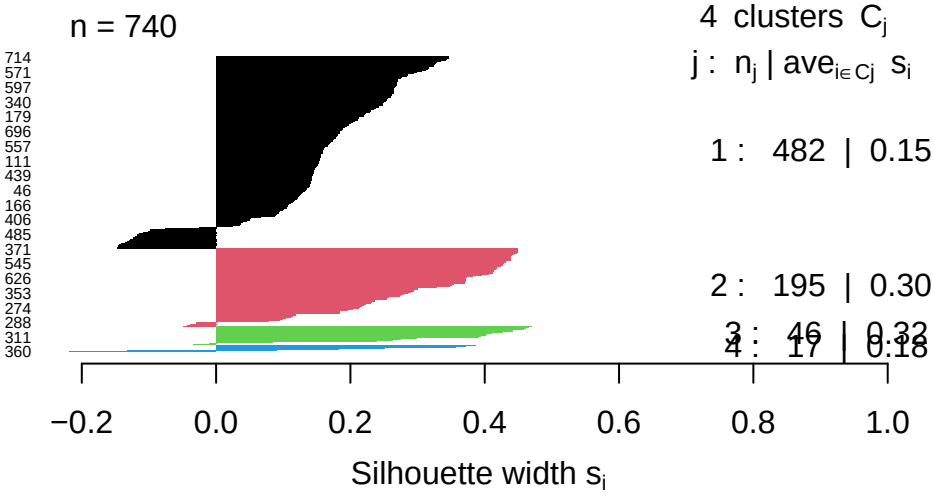

Complete Linkage
0 4 8 12
Average Linkage
0 2 4 6 8
Single Linkage
0 1 2 3 4



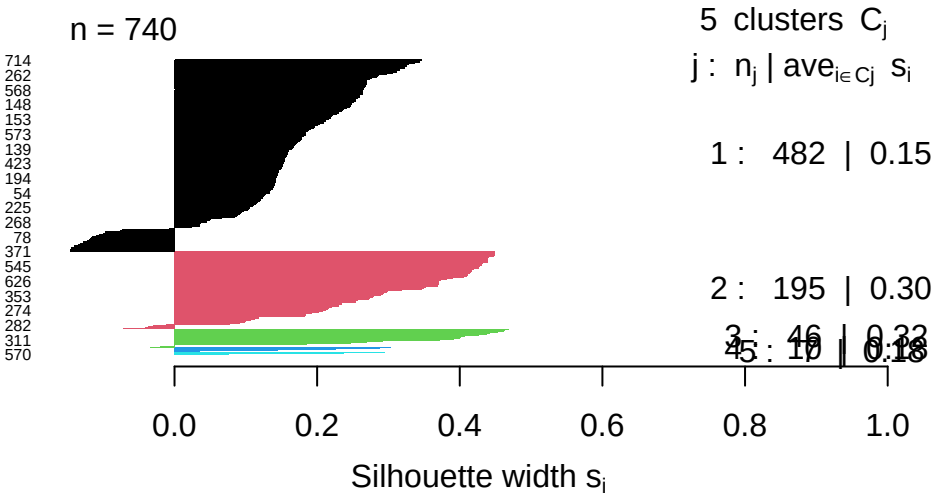
Average silhouette width : 0.41



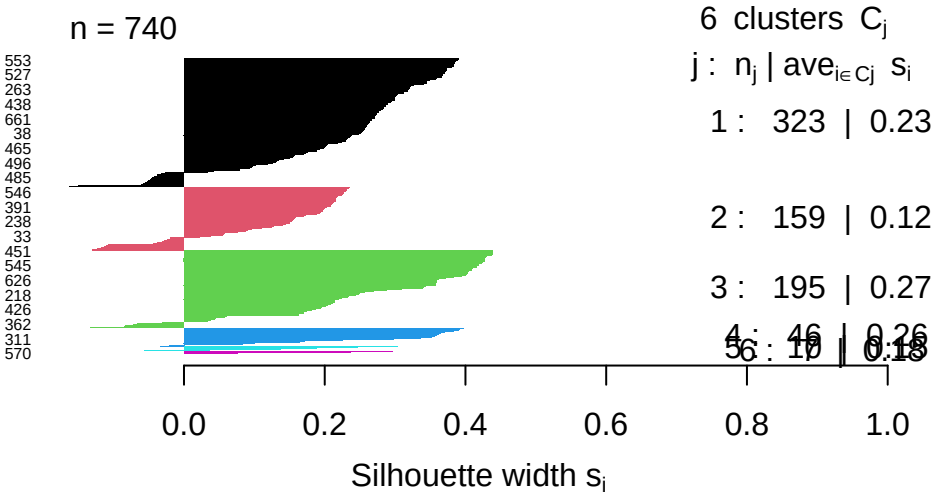
Average silhouette width : 0.3



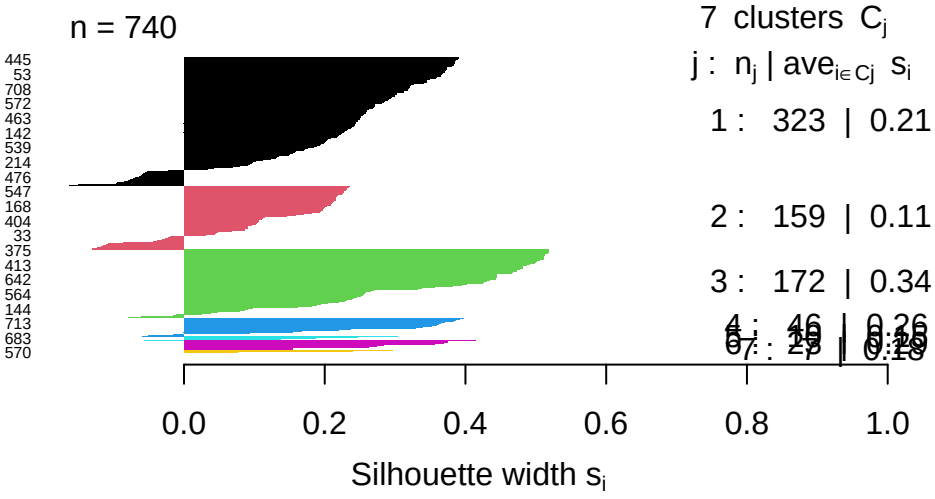
Average silhouette width : 0.2



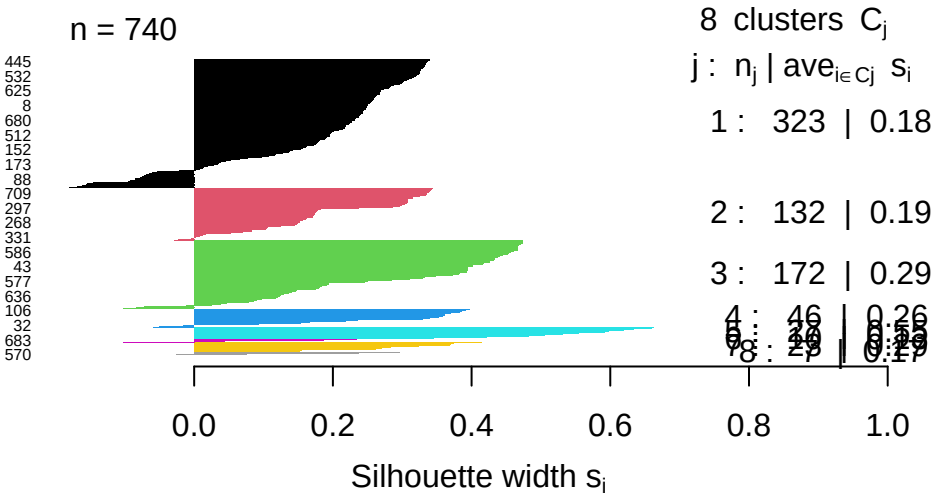
Average silhouette width : 0.2



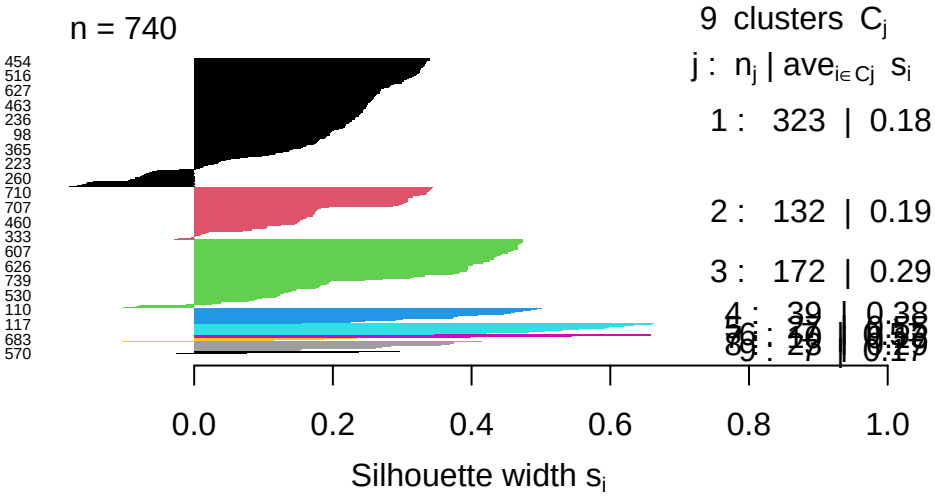
Average silhouette width : 0.22



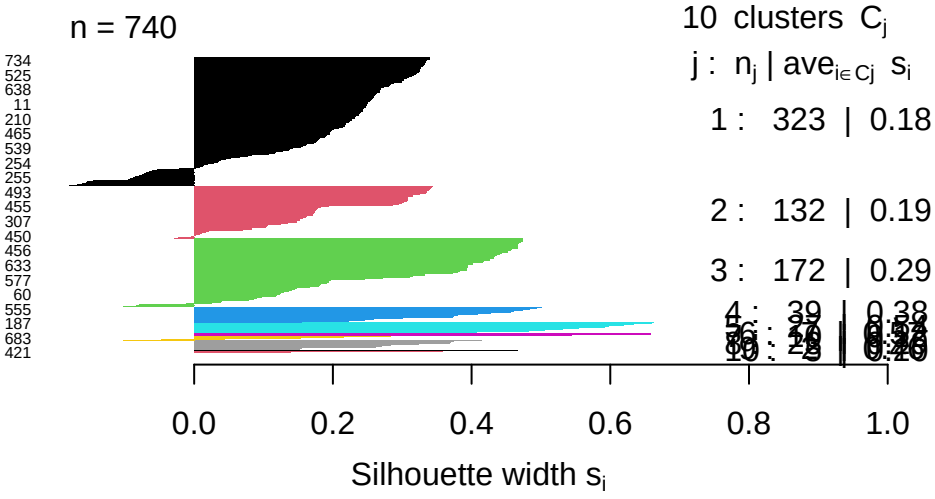
Average silhouette width : 0.22



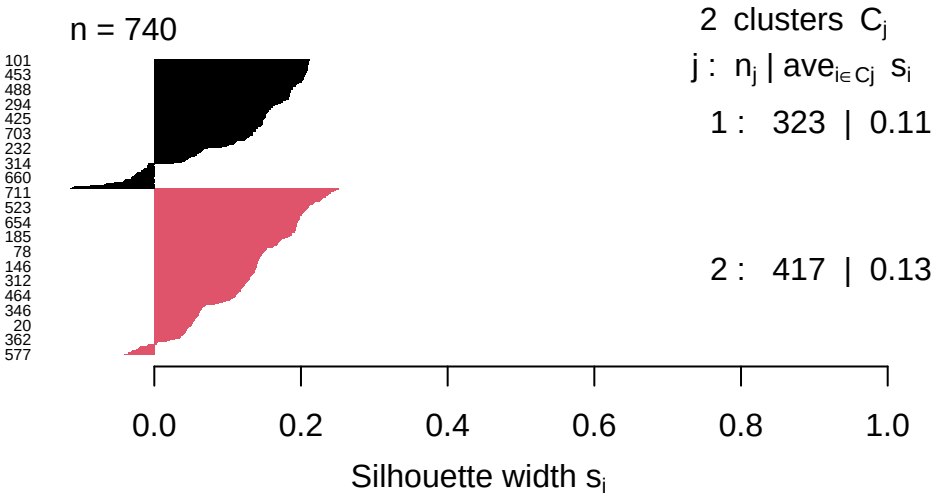
Average silhouette width : 0.23



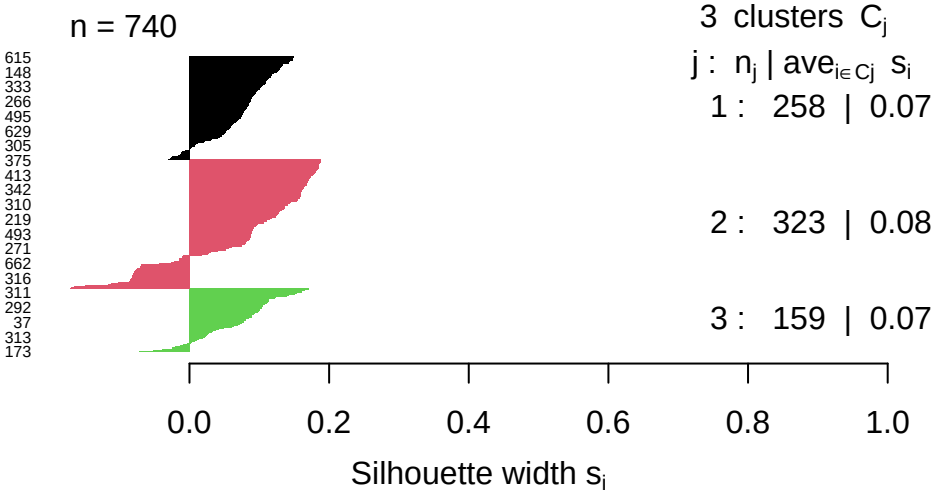
Average silhouette width : 0.24



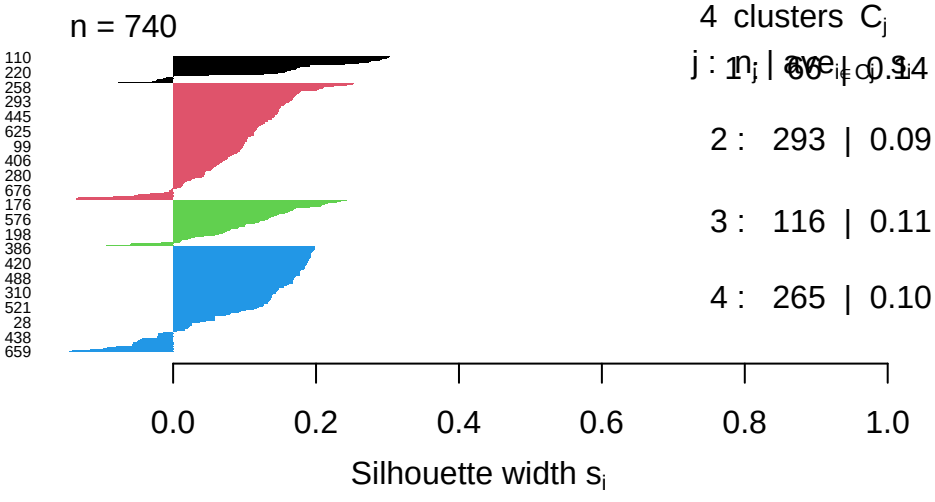
Average silhouette width : 0.24



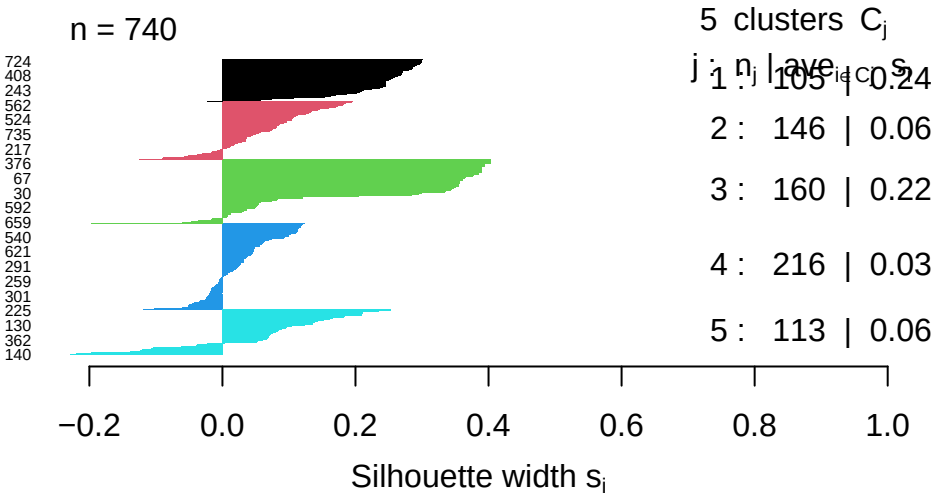
Average silhouette width : 0.12



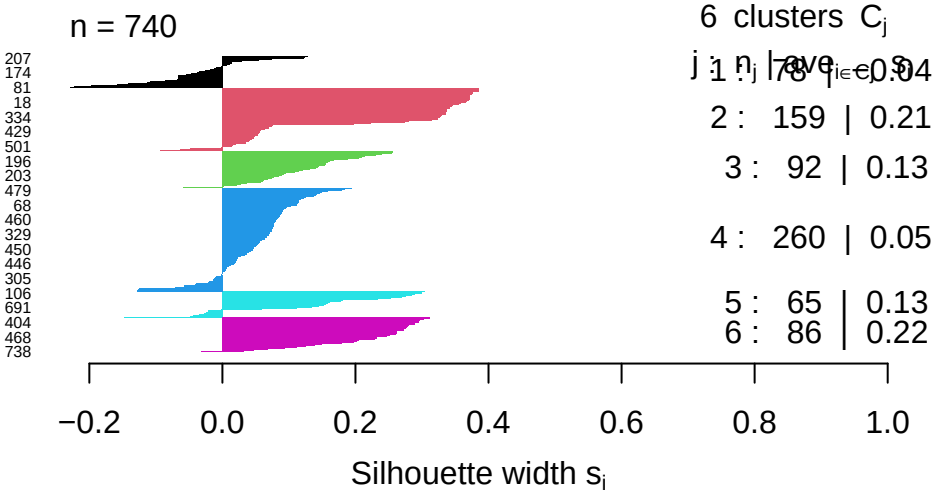
Average silhouette width : 0.07



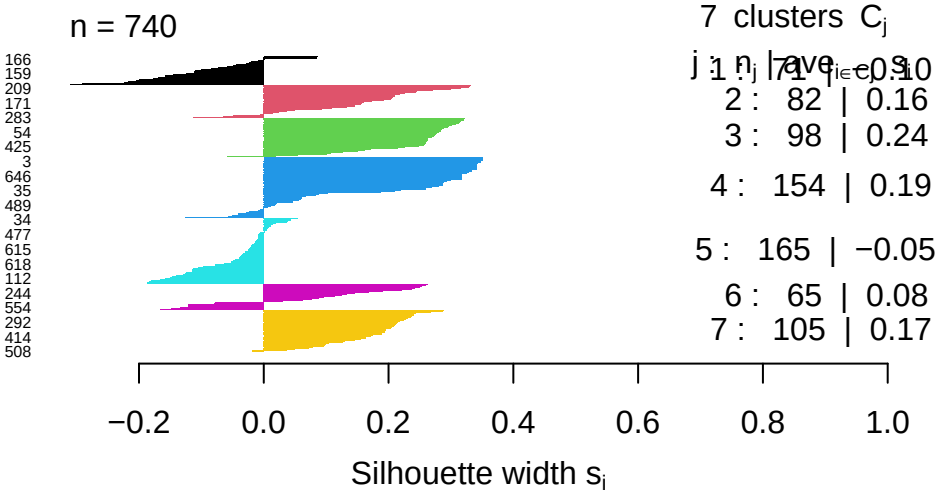
Average silhouette width : 0.1



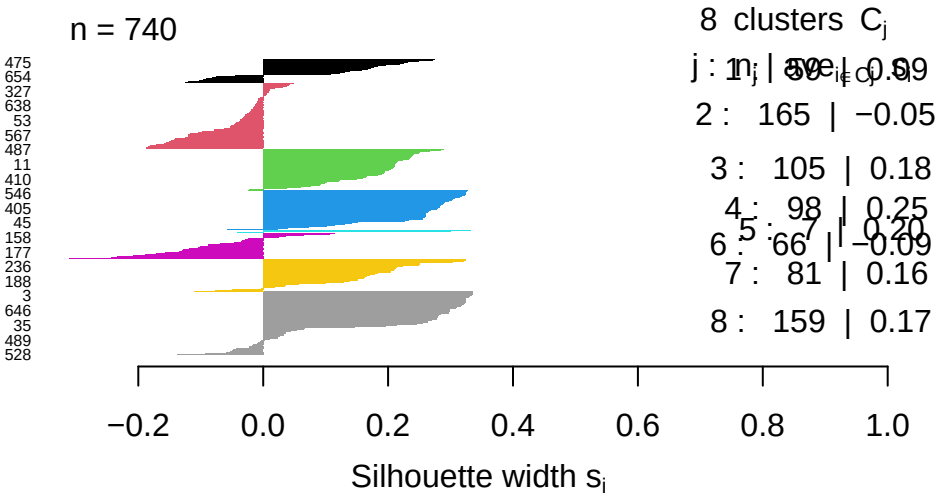
Average silhouette width : 0.11



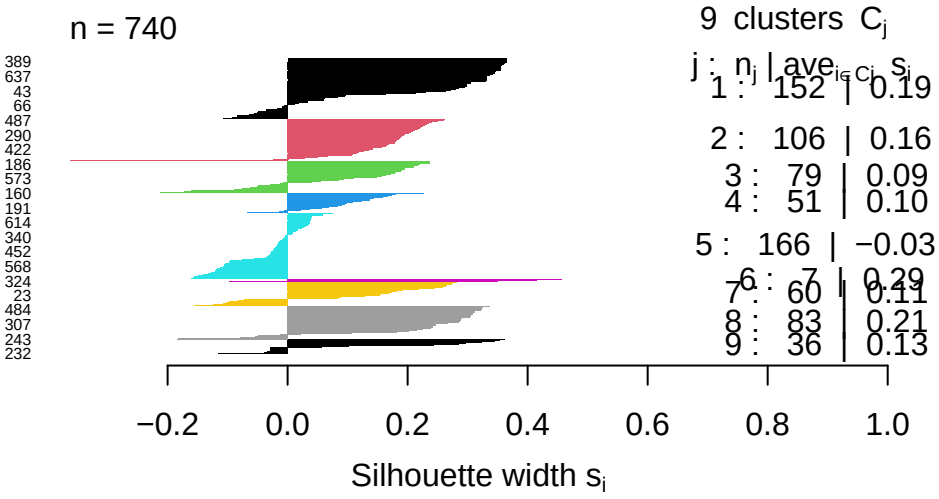
Average silhouette width : 0.11



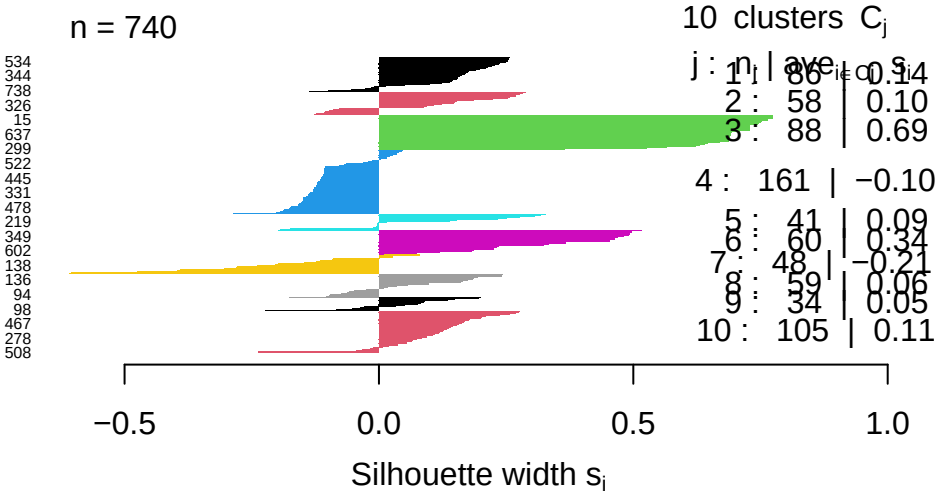
Average silhouette width : 0.1



Average silhouette width : 0.1



Average silhouette width : 0.11



Average silhouette width : 0.13

```
[1] 0.5040522
```

```
[1] 0.6137717
```

Introduction

Methods

Discussion

Supplementary material

```
# ----- SETUP ----- #
# Load packages
library(tidyverse)
library(ggplot2)
library(cluster)
library(fossil)

# Read data and extract labels
absentData_raw <- read.csv("Absenteeism_at_work.csv", sep = ";")
absentData_lab <- absentData_raw$`Reason.for.absence`

# Keep only quantitative variables because we are interested in k-means clustering
absentData <- absentData_raw %>%
  select(-c("Reason.for.absence","ID","Month.of.absence","Day.of.the.week","Seasons",
            "Disciplinary.failure","Education","Son","Social.drinker","Social.smoker"))

# ----- AGGLOMERATIVE HIERARCHICAL CLUSTERING ----- #
# Compare linkage types
absentData_sd <- scale(absentData)
absentData_dist <- dist(absentData_sd)
plot(hclust(absentData_dist), xlab = "", sub = "", ylab = "",
     labels = absentData_lab, main = "Complete Linkage")
plot(hclust(absentData_dist, method = "average"),
     labels = absentData_lab, main = "Average Linkage",
     xlab = "", sub = "", ylab = "")
plot(hclust(absentData_dist, method = "single"),
     labels = absentData_lab, main = "Single Linkage",
     xlab = "", sub = "", ylab = "")

# Choose k using goodness-of-clustering
```

```r
set.seed(780)
plotHeirSilK <- function(k){
  hc_out <- hclust(dist(absentData_sd))
  hc_clusters <- cutree(hc_out, k)
  #plot(hc_out, labels = absentData_lab)
  sil <- silhouette(hc_clusters, dist(absentData_sd))
  plot(sil, nmax= 1000, cex.names=0.5, main = "", col=1:k, border=NA)
}
plotHeirSilK(2)
plotHeirSilK(3)
plotHeirSilK(4)
plotHeirSilK(5)
plotHeirSilK(6)
plotHeirSilK(7)
plotHeirSilK(8)
plotHeirSilK(9)
plotHeirSilK(10)

# ----- K-MEANS CLUSTERING ----- #
# Choose k using goodness-of-clustering
set.seed(780)
plotSilK <- function(k){
  x_k <- kmeans(absentData, k, nstart = 20)
  sil <- silhouette(x_k$cluster, dist(absentData_sd))
  plot(sil, nmax= 1000, cex.names=0.5, main = "", col=1:k, border=NA)
}
plotSilK(2)
plotSilK(3)
plotSilK(4)
plotSilK(5)
plotSilK(6)
```

```r
plotSilK(7)
plotSilK(8)
plotSilK(9)
plotSilK(10)

# Perform k-means clustering with k=2
set.seed(780)
km_out <- kmeans(absentData, 2, nstart = 20)
km_clusters <- km_out$cluster

# Compare the k-means cell clusters with the given labels. Compute the rand index between gi
rand.index(km_clusters, as.numeric(as.factor(absentData_lab)))
adj.rand.index(km_clusters, as.numeric(as.factor(absentData_lab)))


# ----- K-MEANS CLUSTERING AFTER PCA ----- #
```

References

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/