# Clustering Techniques

## By: Tyler Bier - Medical data set using R & K-means model

### Part 1

Loading our various packages into R.

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)

## Warning: package 'ggplot2' was built under R version 4.3.3

library(factoextra)

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

library(cluster)

## Warning: package 'cluster' was built under R version 4.3.3

To start, we will import our medical data set CSV file into R, verify, and then select/ scale our variables: Income and Initial\_days.Summary comparison for both variable before/ after scaling shown below.

medical\_clean <- read.csv("C:/Users/Tyler Bier/OneDrive/Desktop/medical\_clean.csv", header=TRUE, sep=',')   
View(medical\_clean)  
summary(medical\_clean$Income)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 154.1 19598.8 33768.4 40490.5 54296.4 207249.1

summary(medical\_clean$Initial\_days)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.002 7.896 35.836 34.455 61.161 71.981

variables\_medclean <- medical\_clean %>% select(c(17,40))  
scaled\_data <- as.data.frame(scale(variables\_medclean))  
summary(scaled\_data$Income)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.4143 -0.7325 -0.2357 0.0000 0.4841 5.8468

summary(scaled\_data$Initial\_days)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.27154 -1.00949 0.05249 0.00000 1.01507 1.42634

Saving prepared data set as a CSV file for Question C4.

write.csv(scaled\_data, file="PreparedDataD212Task1.CSV", row.names = TRUE)

### Part 2

Now determining the optimal number of clusters for our data set. Will be creating two plots using the Elbow method and the Silhouette method as demonstrated in the visualizations below.

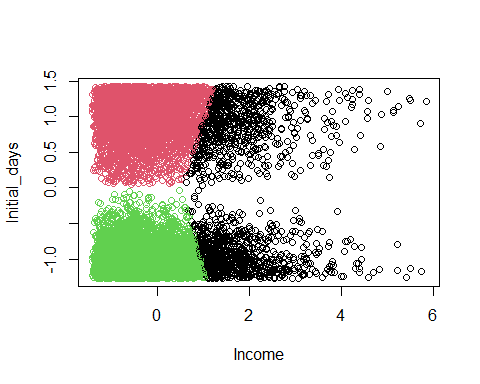
{r} set.seed(1) fviz\_nbclust(scaled\_data, kmeans, method =“wss”) + labs(subtitle = “Elbow Method”)

fviz\_nbclust(scaled\_data, kmeans, method = “silhouette”) + labs(subtitle = “Silhouette Method”)

### Part 3

We will now begin creating our K-means model, as well as creating some evaluation metrics below.

set.seed(1)  
k3\_model <- kmeans(scaled\_data, centers = 3)  
clusters\_k3 <- as.factor(k3\_model$cluster)  
cluster1 <- scaled\_data[ which(clusters\_k3==1),]  
cluster2 <- scaled\_data[ which(clusters\_k3==2),]  
cluster3 <- scaled\_data[ which(clusters\_k3==3),]  
plot(scaled\_data, col= clusters\_k3)



print('Datapoints per cluster')

## [1] "Datapoints per cluster"

summary(clusters\_k3)

## 1 2 3   
## 1500 4303 4197

print('Cluster 1 summary')

## [1] "Cluster 1 summary"

summary(cluster1)

## Income Initial\_days   
## Min. :0.6309 Min. :-1.27015   
## 1st Qu.:1.2432 1st Qu.:-1.00512   
## Median :1.5911 Median :-0.52952   
## Mean :1.8349 Mean :-0.08954   
## 3rd Qu.:2.1911 3rd Qu.: 0.88639   
## Max. :5.8468 Max. : 1.42586

print('Cluster 2 summary')

## [1] "Cluster 2 summary"

summary(cluster2)

## Income Initial\_days   
## Min. :-1.4091 Min. :0.0422   
## 1st Qu.:-0.7932 1st Qu.:0.7514   
## Median :-0.3909 Median :1.0294   
## Mean :-0.3132 Mean :0.9714   
## 3rd Qu.: 0.1257 3rd Qu.:1.2355   
## Max. : 1.1983 Max. :1.4263

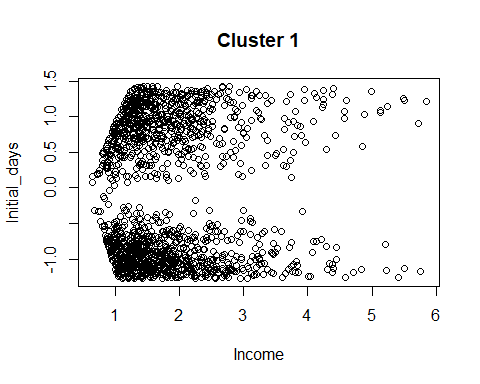
print('Cluster 3 summary')

## [1] "Cluster 3 summary"

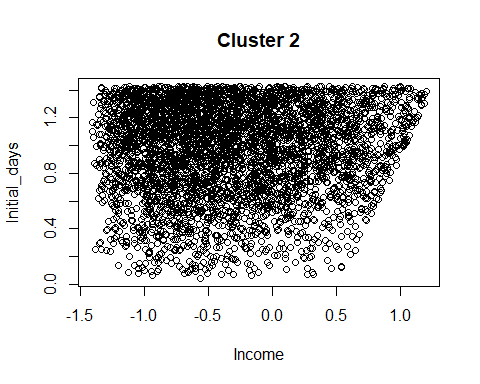
summary(cluster3)

## Income Initial\_days   
## Min. :-1.4143 Min. :-1.27154   
## 1st Qu.:-0.8059 1st Qu.:-1.14729   
## Median :-0.4117 Median :-1.01379   
## Mean :-0.3347 Mean :-0.96398   
## 3rd Qu.: 0.1122 3rd Qu.:-0.82327   
## Max. : 1.0258 Max. :-0.03265

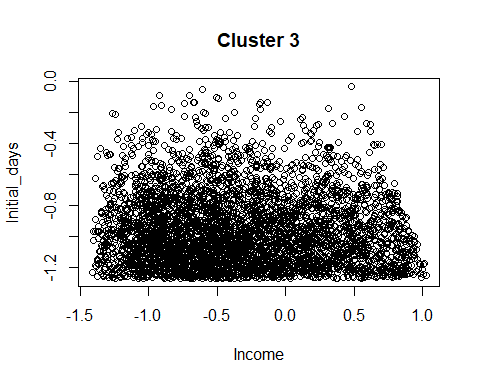
plot(cluster1, main= "Cluster 1")



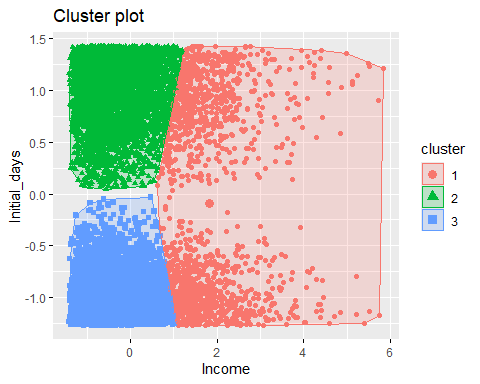
plot(cluster2, main= 'Cluster 2')



plot(cluster3, main= 'Cluster 3')



fviz\_cluster(k3\_model, data = scaled\_data, geom = "point", ellipse = TRUE)



silhouette\_score <- silhouette(k3\_model$cluster, dist(scaled\_data))  
fviz\_silhouette(silhouette\_score)

## cluster size ave.sil.width  
## 1 1 1500 0.25  
## 2 2 4303 0.57  
## 3 3 4197 0.61

