Segmenting Shopping Mall Customers

Business Understanding

Our goal is to segment customers based on Income and SpendingScore

library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

mallcustomers <- read\_csv("Student/Data/mallcustomers.csv")

## Parsed with column specification:  
## cols(  
## CustomerID = col\_double(),  
## Gender = col\_character(),  
## Age = col\_double(),  
## Income = col\_character(),  
## SpendingScore = col\_double()  
## )

glimpse(mallcustomers)

## Rows: 200  
## Columns: 5  
## $ CustomerID <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, …  
## $ Gender <chr> "Male", "Male", "Female", "Female", "Female", "Female",…  
## $ Age <dbl> 19, 21, 20, 23, 31, 22, 35, 23, 64, 30, 67, 35, 58, 24,…  
## $ Income <chr> "15,000 USD", "15,000 USD", "16,000 USD", "16,000 USD",…  
## $ SpendingScore <dbl> 39, 81, 6, 77, 40, 76, 6, 94, 3, 72, 14, 99, 15, 77, 13…

Data Understanding

Convert income feature to a string

library(stringr)  
  
mallcustomers <- mallcustomers %>%  
 mutate(Income = str\_replace\_all(Income, " USD", "")) %>%  
 mutate(Income = str\_replace\_all(Income, ",", "")) %>%  
 mutate(Income = as.numeric(Income))  
  
summary(mallcustomers)

## CustomerID Gender Age Income   
## Min. : 1.00 Length:200 Min. :18.00 Min. : 15000   
## 1st Qu.: 50.75 Class :character 1st Qu.:28.75 1st Qu.: 41500   
## Median :100.50 Mode :character Median :36.00 Median : 61500   
## Mean :100.50 Mean :38.85 Mean : 60560   
## 3rd Qu.:150.25 3rd Qu.:49.00 3rd Qu.: 78000   
## Max. :200.00 Max. :70.00 Max. :137000   
## SpendingScore   
## Min. : 1.00   
## 1st Qu.:34.75   
## Median :50.00   
## Mean :50.20   
## 3rd Qu.:73.00   
## Max. :99.00

Data Preparation

Normalize Income and SpendingScore features - z-score

mallcustomers\_scaled <- mallcustomers %>%  
 select(-CustomerID, -Gender, -Age) %>%  
 scale()  
  
summary(mallcustomers\_scaled)

## Income SpendingScore   
## Min. :-1.73465 Min. :-1.905240   
## 1st Qu.:-0.72569 1st Qu.:-0.598292   
## Median : 0.03579 Median :-0.007745   
## Mean : 0.00000 Mean : 0.000000   
## 3rd Qu.: 0.66401 3rd Qu.: 0.882916   
## Max. : 2.91037 Max. : 1.889750

Modeling

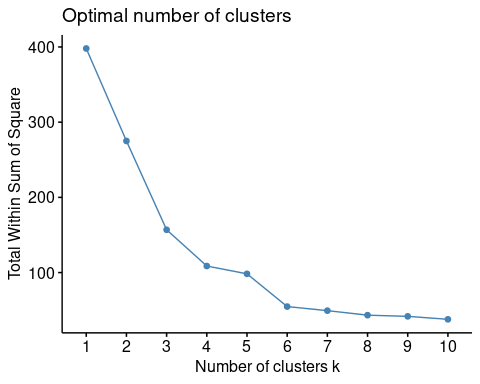
Obtain recommended value of (k)

Elbow Method

library(factoextra)

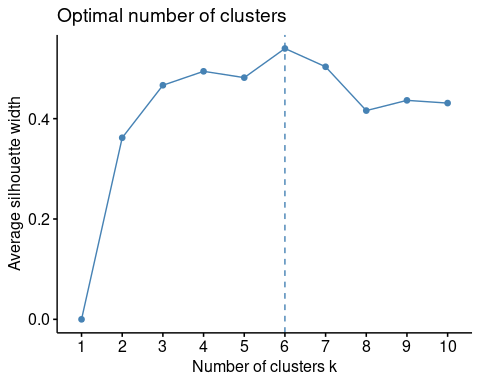
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

fviz\_nbclust(mallcustomers\_scaled, kmeans, method = "wss")

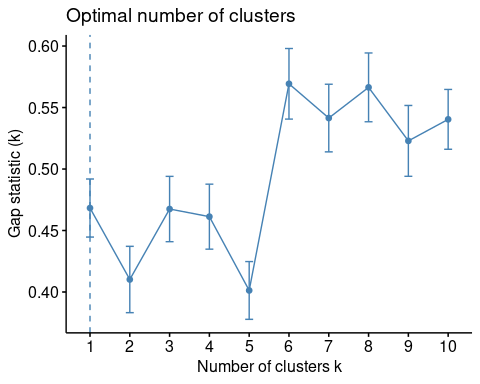


Silhouette Method

fviz\_nbclust(mallcustomers\_scaled, kmeans, method = "silhouette")

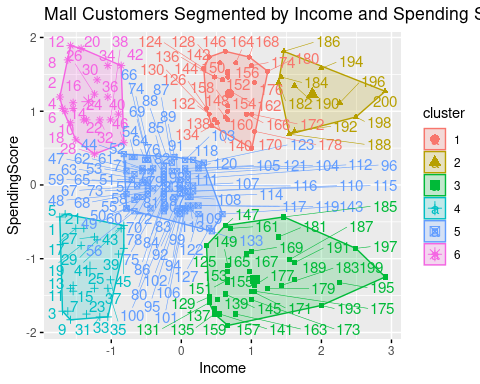
 Gap Statistic

fviz\_nbclust(mallcustomers\_scaled, kmeans, method = "gap\_stat")



Building The Model

set.seed(1234)  
k\_clust <- kmeans(mallcustomers\_scaled, centers = 6, nstart = 25)  
  
fviz\_cluster(  
 k\_clust,  
 data = mallcustomers\_scaled,  
 main = "Mall Customers Segmented by Income and Spending Score",  
 repel = TRUE  
)



Evaluating The Model

mallcustomers %>%  
 mutate(cluster = k\_clust$cluster) %>%  
 mutate(Male = ifelse(Gender == "Male", 1, 0)) %>%  
 mutate(Female = ifelse(Gender == "Female", 1, 0)) %>%  
 select(cluster, Male, Female, Age) %>%  
 group\_by(cluster) %>%  
 summarise\_all("mean")

## # A tibble: 6 x 4  
## cluster Male Female Age  
## <int> <dbl> <dbl> <dbl>  
## 1 1 0.483 0.517 32.9  
## 2 2 0.4 0.6 32.2  
## 3 3 0.543 0.457 41.1  
## 4 4 0.391 0.609 45.2  
## 5 5 0.407 0.593 42.7  
## 6 6 0.409 0.591 25.3