#Load Packages

library(tibble)  
library(utils)  
library(stats)  
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(base)  
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.2.2

## Loading required package: ggplot2

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

#Read the data

data<-read.csv("OnlineRetail.csv",row.names = NULL)

#Prep the data for analysis

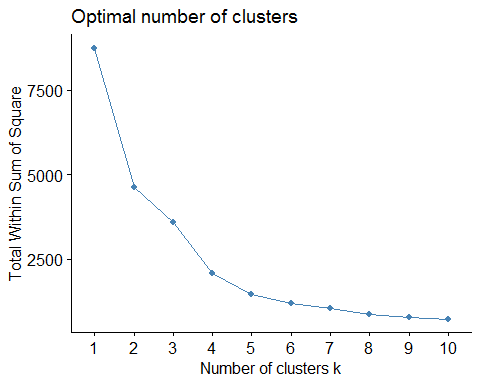
#omit null values  
data<-na.omit(data)  
  
#calculate total spent by each customer  
data.totalspent<-data%>%  
 select(Quantity,UnitPrice,CustomerID)%>%  
 mutate(TotalSpent = Quantity\*UnitPrice)%>%  
 group\_by(CustomerID)%>%  
 summarise(SumSpent=sum(TotalSpent))  
  
#calculate frequency of purchases by each customer  
data.freq<-data%>%  
 select(InvoiceNo,CustomerID)%>%  
 group\_by(CustomerID)%>%  
 count(InvoiceNo)%>%  
 summarise(freq=sum(n))  
  
summary<-merge(data.freq,data.totalspent)

#Normalize the data for analysis

summary.norm<-summary[,2:3]  
summary.norm<-sapply(summary.norm,scale)#normalization

#Choose optimal number of clusters for analysis using Elbow method

fviz\_nbclust(summary.norm,kmeans,method="wss") #results show 4 clusters is optimal



#Perform kmeans clustering and analyze the results

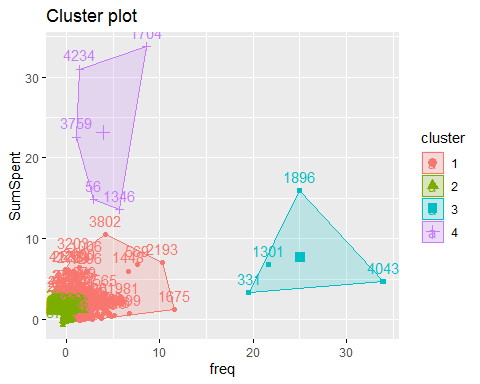
k4<-kmeans(summary.norm,centers=4,nstart=50)  
k4$centers #centroids of each cluster for each variable

## freq SumSpent  
## 1 1.6629766 1.0125886  
## 2 -0.1390071 -0.1025421  
## 3 25.0393877 7.6500186  
## 4 3.9589639 23.1411860

k4$size #size of each cluster

## [1] 270 4093 4 5

fviz\_cluster(k4,data=summary.norm) #cluster visualization



#Cluster 1 - Low frequency customers and low amount spent  
#Cluster 2 - Very low frequency customers and very low amount spent  
#Cluster 3 - Very high frequency customers and high amount spent  
#Cluster 4 - Low frequency customers and very high amount spent