# high value customer identification

ecom <- read.csv(file.choose(), stringsAsFactors = FALSE)

View(ecom)

#structure of dataset

str(ecom)

#descriptive stats

summary(ecom)

#customer Id's missing, delete observations

dim(ecom)

#remove redundant variable

ecom$X <- NULL

is.na(ecom$CustomerID)

ecom\_1 <- na.omit(ecom)

dim(ecom\_1)

#aggregate the data at customer Id level

library(dplyr)

library(lubridate)

ecom\_1$InvoiceDate <- dmy(ecom\_1$InvoiceDate)

#find date of last order

#max value of invoice date

max(ecom\_1$InvoiceDate)

# 2017-12-07

#difference between last order date and transaction date

ecom\_1$days\_passed <- as.Date('2017-12-07') - ecom\_1$InvoiceDate

View(ecom\_1)

#remove those rows where quantity was returned

ecom\_2 <- ecom\_1 %>% filter(Quantity>=0)

#number of transactions

ecom\_3 <- ecom\_2 %>% group\_by(CustomerID) %>% summarise(no\_of\_trans=as.numeric(n\_distinct(InvoiceNo)),

no\_of\_dist\_items=n\_distinct(StockCode),

total\_quant=sum(Quantity),

total\_value=sum(Quantity\*UnitPrice),

latest\_order=as.numeric(min(days\_passed)))

View(ecom\_3)

#cluster data

#unsupervised - no need to divide into test and training data

#normalize data

#no need for customer ID

ecom\_sc <- scale(ecom\_3[-1])

View(ecom\_sc)

#to check if it has been scaled or not

# mean=0 for each column

apply(ecom\_sc,2,mean)

# std deviation

apply(ecom\_sc,2,sd)

#distnace between two points of cluster

distance <- dist(ecom\_sc,method = 'euclidean')

#linkage algorithm

clus\_model <- hclust(distance,method = 'ward.D')

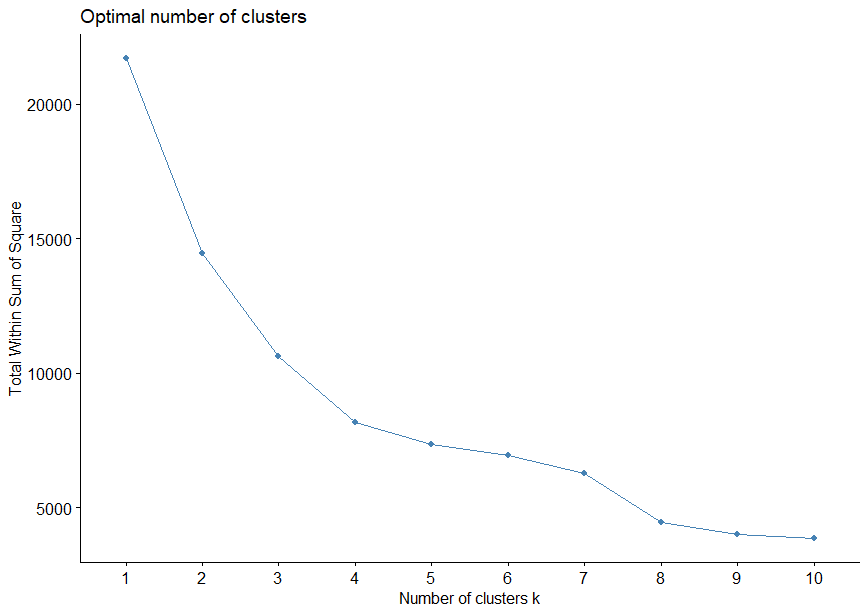
#to know optimum number of cluster in dataset

#elbowcurve

library('factoextra')

fviz\_nbclust(ecom\_sc,kmeans,method = 'wss')

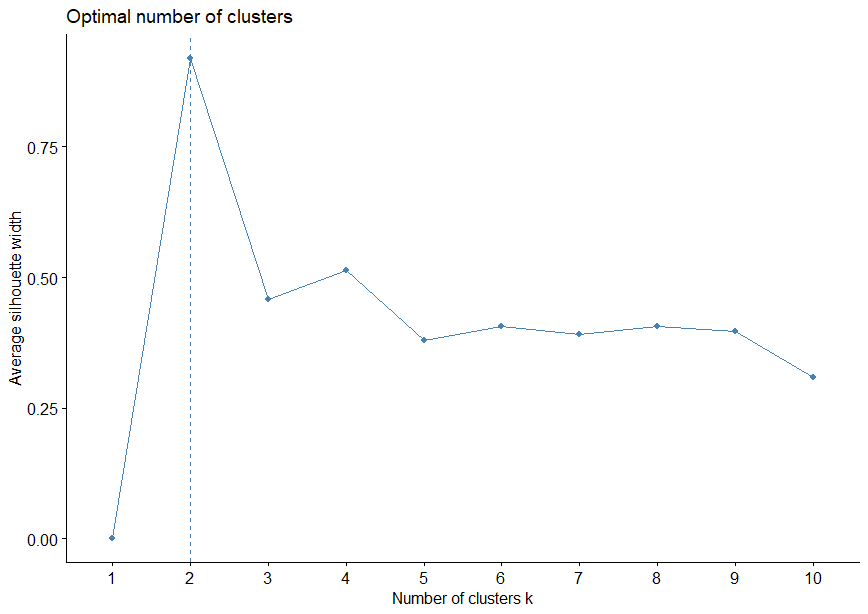
# from elbow curve we get the optimum number of cluster=4



#silhoette metric - if the points are correctly assigned

# -1 to +1

fviz\_nbclust(ecom\_sc,kmeans,method = 'silhouette')



#lets take optimum as n=4

#kmeans clustering

clus\_model <- kmeans(ecom\_sc,4,iter.max = 10)

#assign each point to the cluster from 1-4

ecom\_3$cluster <- clus\_model$cluster

View(ecom\_3)

#cluster demographics

colnames(ecom\_3)

clus\_demo <- ecom\_3 %>% group\_by(cluster) %>% summarise(mean(no\_of\_trans),mean(no\_of\_dist\_items),mean(total\_quant),mean(total\_value),

mean(latest\_order), n())

View(clus\_demo)

# to further break it down

#filter by cluster 2 and apply same methodology