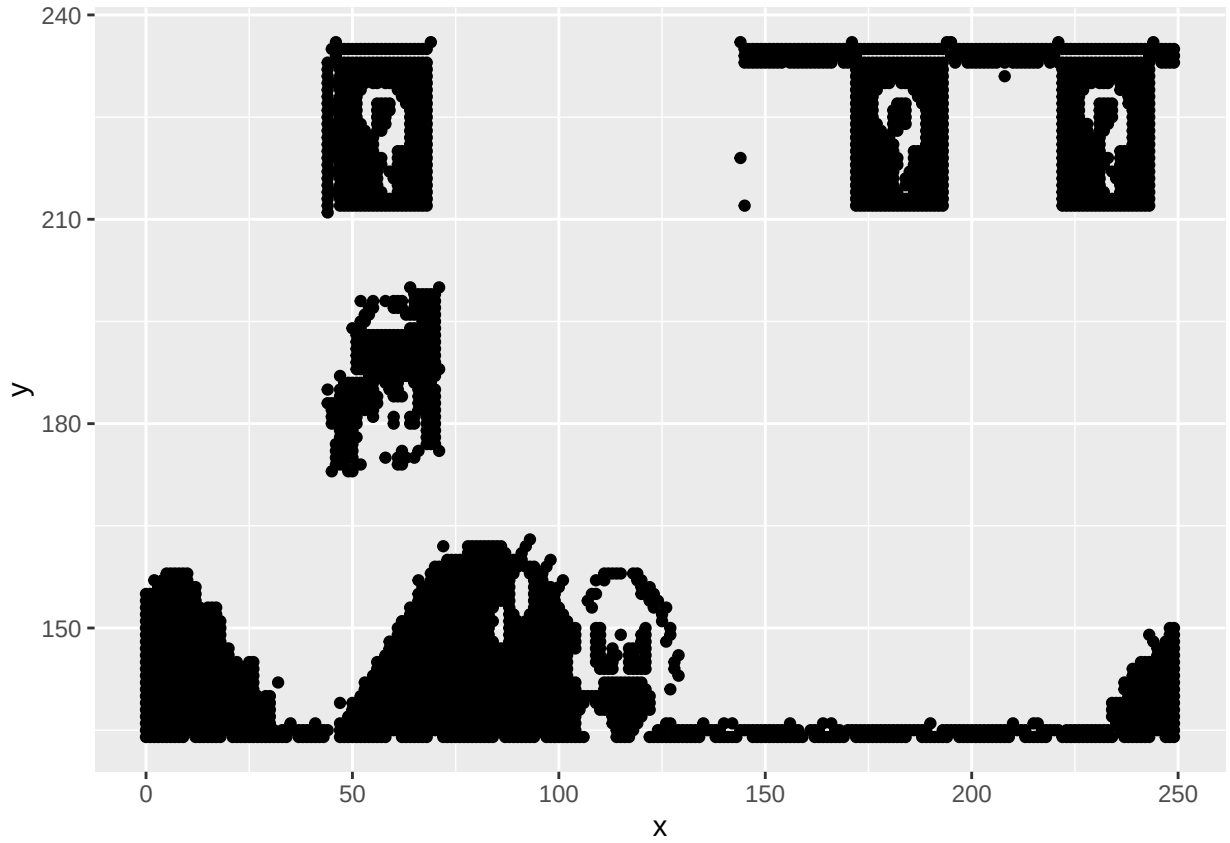
Assignment 11.2.2

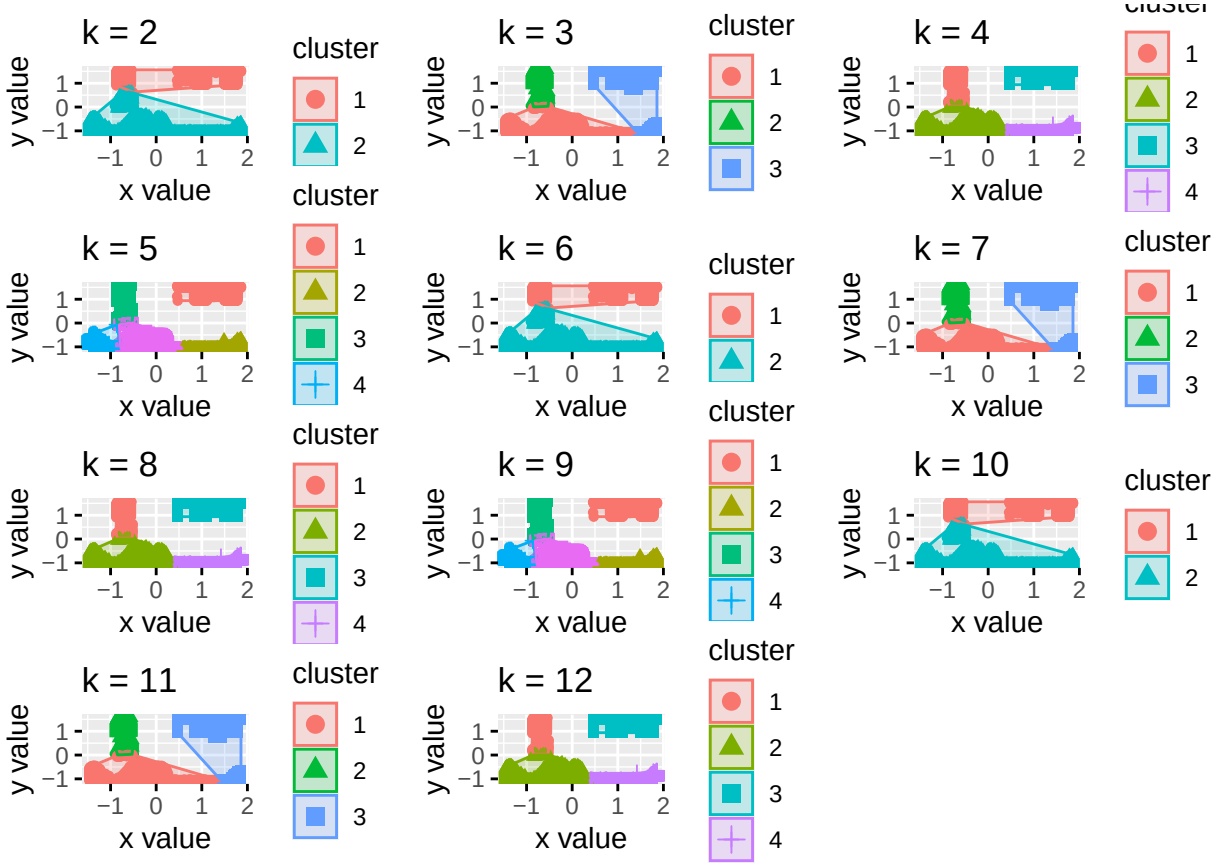
Kiran Komati

2021-06-05

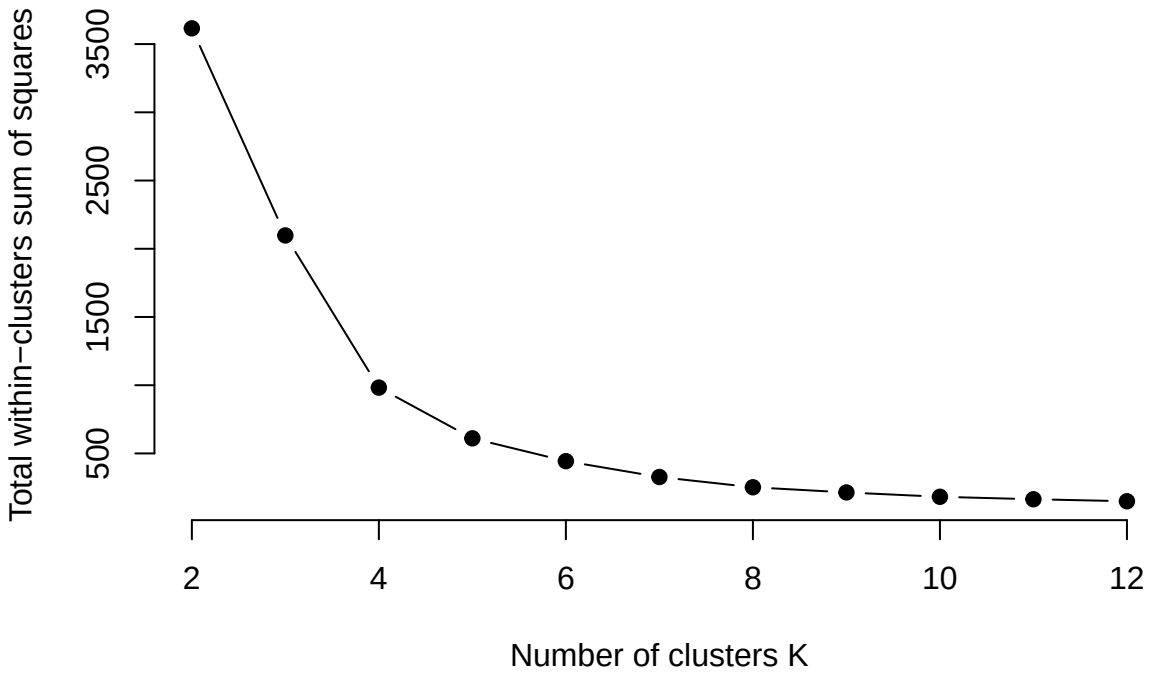
```
setwd("C:/Users/kiran/dsc520/data")
library(factoextra)
library(gridExtra)
library(cluster)
library(tidyverse)
library(ggplot2)

# Plot the dataset using a scatter plot.

cd <- read.csv('clustering-data.csv')
ggplot(cd,aes(x=x,y=y)) + geom_point()
```



```r
cd_scaled <- as.data.frame(scale(cd))
k2<- kmeans(cd_scaled, centers=2,nstart=25)
k3<- kmeans(cd_scaled, centers=3,nstart=25)
k4<- kmeans(cd_scaled, centers=4,nstart=25)
k5<- kmeans(cd_scaled, centers=5,nstart=25)
k6<- kmeans(cd_scaled, centers=6,nstart=25)
k7<- kmeans(cd_scaled, centers=7,nstart=25)
k8<- kmeans(cd_scaled, centers=8,nstart=25)
k9<- kmeans(cd_scaled, centers=9,nstart=25)
k10<- kmeans(cd_scaled, centers=10,nstart=25)
k11<- kmeans(cd_scaled, centers=11,nstart=25)
k12<- kmeans(cd_scaled, centers=12,nstart=25)

# Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant
p1 <- fviz_cluster(k2, geom = "point", data = cd_scaled) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = cd_scaled) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = cd_scaled) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = cd_scaled) + ggtitle("k = 5")
p5 <- fviz_cluster(k2, geom = "point", data = cd_scaled) + ggtitle("k = 6")
p6 <- fviz_cluster(k3, geom = "point",  data = cd_scaled) + ggtitle("k = 7")
p7 <- fviz_cluster(k4, geom = "point",  data = cd_scaled) + ggtitle("k = 8")
p8 <- fviz_cluster(k5, geom = "point",  data = cd_scaled) + ggtitle("k = 9")
p9 <- fviz_cluster(k2, geom = "point", data = cd_scaled) + ggtitle("k = 10")
p10 <- fviz_cluster(k3, geom = "point",  data = cd_scaled) + ggtitle("k = 11")
p11 <- fviz_cluster(k4, geom = "point",  data = cd_scaled) + ggtitle("k = 12")


grid.arrange(p1, p2, p3, p4,p5,p6,p7,p8,p9,p10,p11, nrow = 4)
```




```
set.seed(123)

# function to compute total within-cluster sum of square
wss <- function(k) {
    kmeans(cd_scaled, k, nstart = 10 )$tot.withinss
}

# Compute and plot wss for k = 2 to k = 12
k.values <- 2:12

# extract wss for 2-12 clusters
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```




```
# Concluding Remarks
```

Based on the plot, we can conclude that the optimal no. of cluster is 4 which is the elbow point.