assignment_11.3_MunjewarSheetal

Sheetal M

2023-02-25

Install and Load required packages :

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(fig.width = 12, fig.height = 10)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)

# Package names
# packages <- c("ggplot2","dplyr","tidyr","magrittr","tidyverse","purrr")
packages <- c("broom","dplyr","RWeka","class","ggplot2","caret","formatR")

# Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

# Packages loading
invisible(lapply(packages, library, character.only = TRUE))
```

```
##
## 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

## Loading required package: lattice
```

K-means algorithm

Set the working directory to the root of your DSC 520 directory

setwd("E:\Data_Science_DSC510\DSC520-Statistics\dsc520")

```
# Set the working directory to the root of your DSC 520 directory
setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")

# Load data from data/binary-classifier-data.csv
df <- read.csv("data/clustering-data.csv")
str(df)
```

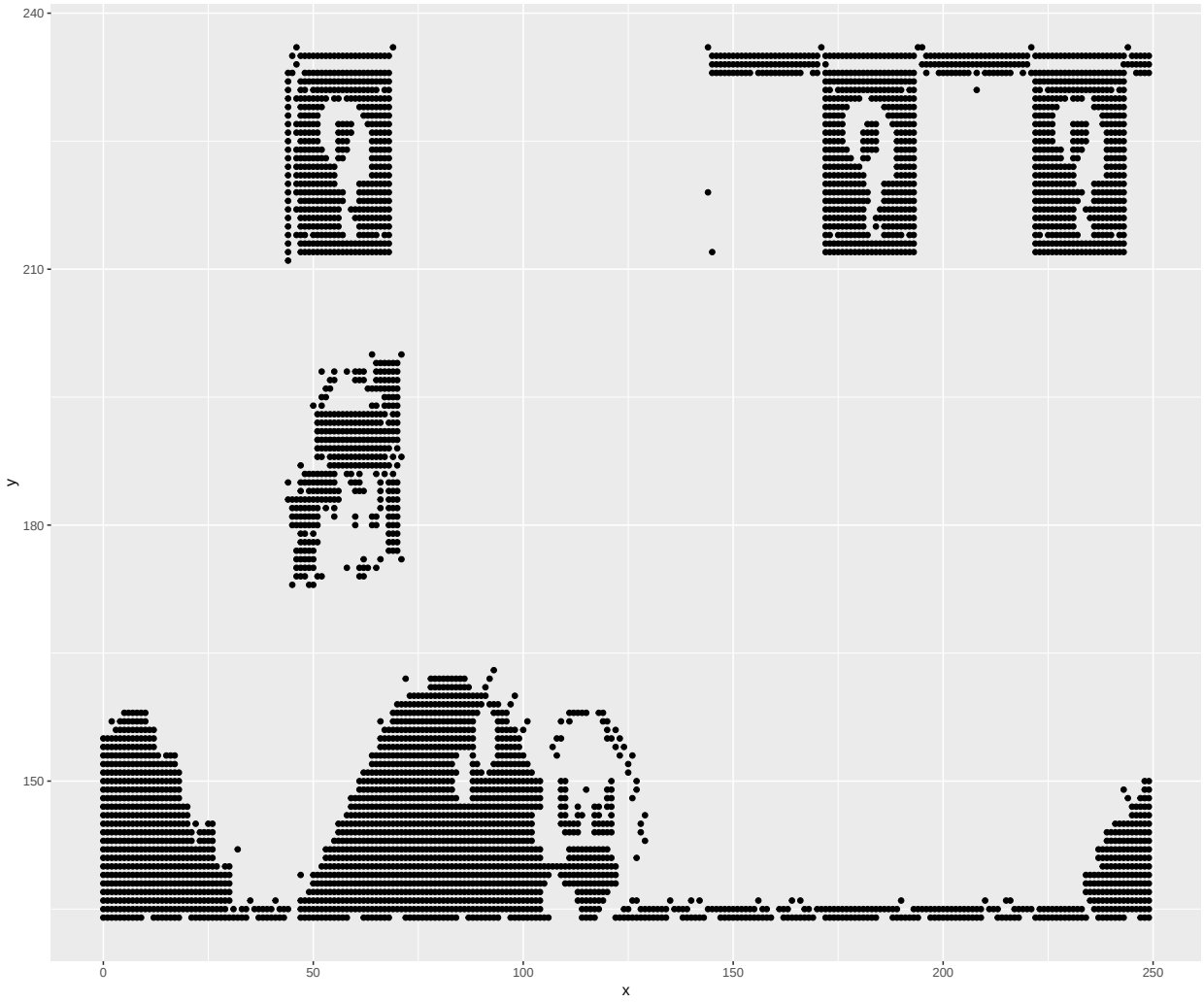
```
## 'data.frame':    4022 obs. of  2 variables:
##  $ x: int  46 69 144 171 194 195 221 244 45 47 ...
##  $ y: int  236 236 236 236 236 236 236 236 235 235 ...
```

```
nrow(df)
```

```
## [1] 4022
```

Visualize dataset - Scatter Plot

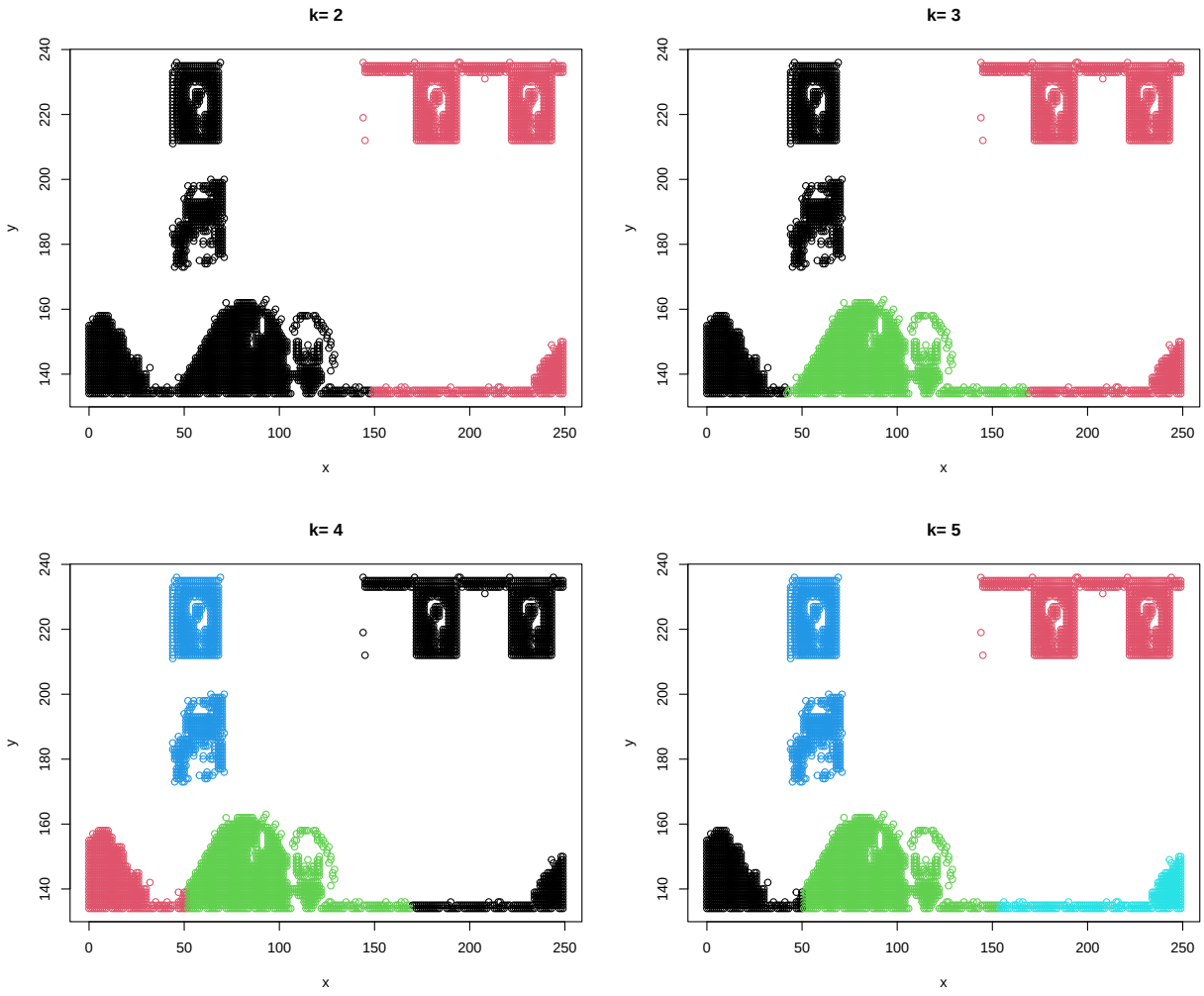
```
ggplot(data = df, aes(x, y)) + geom_point()
```

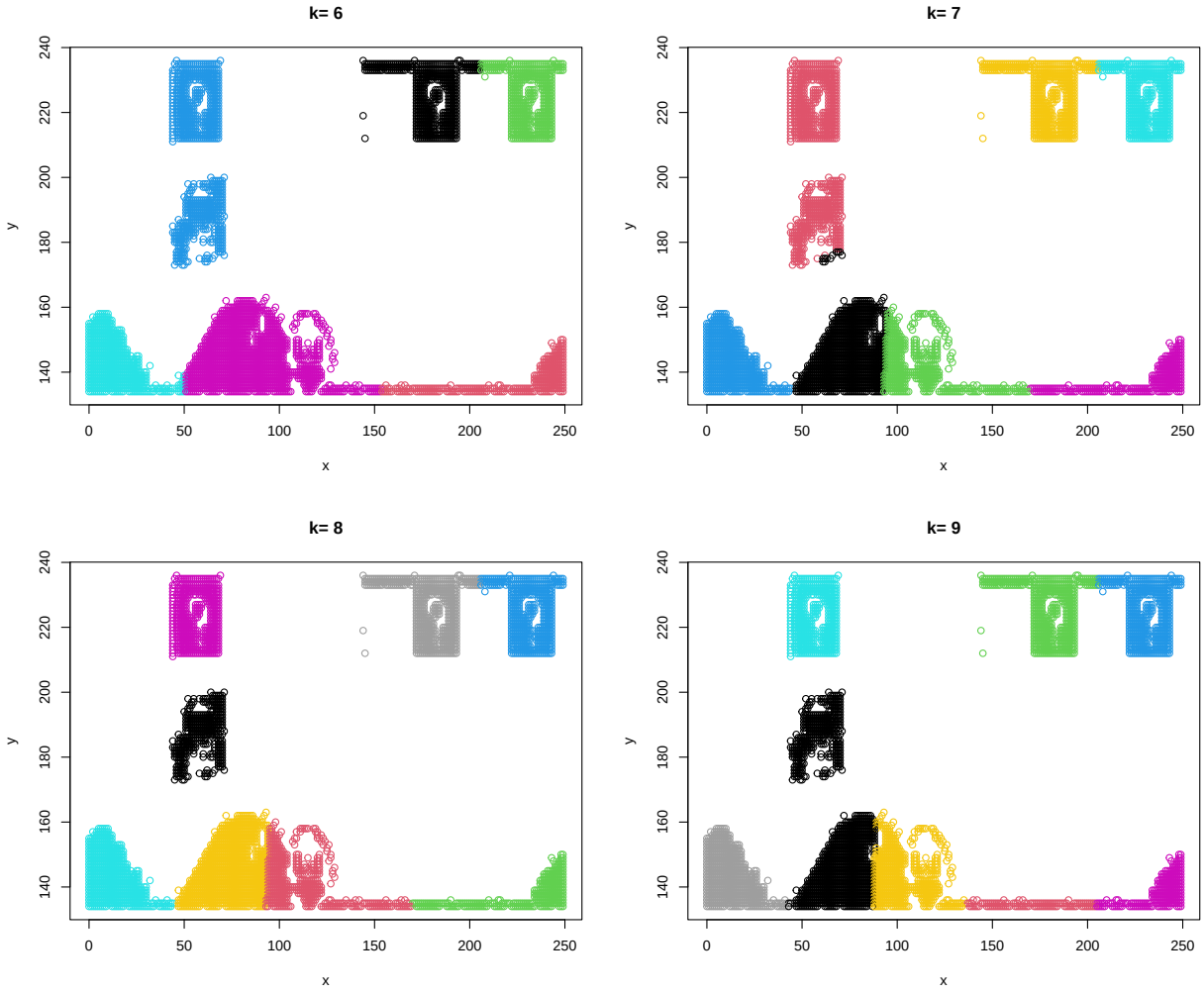


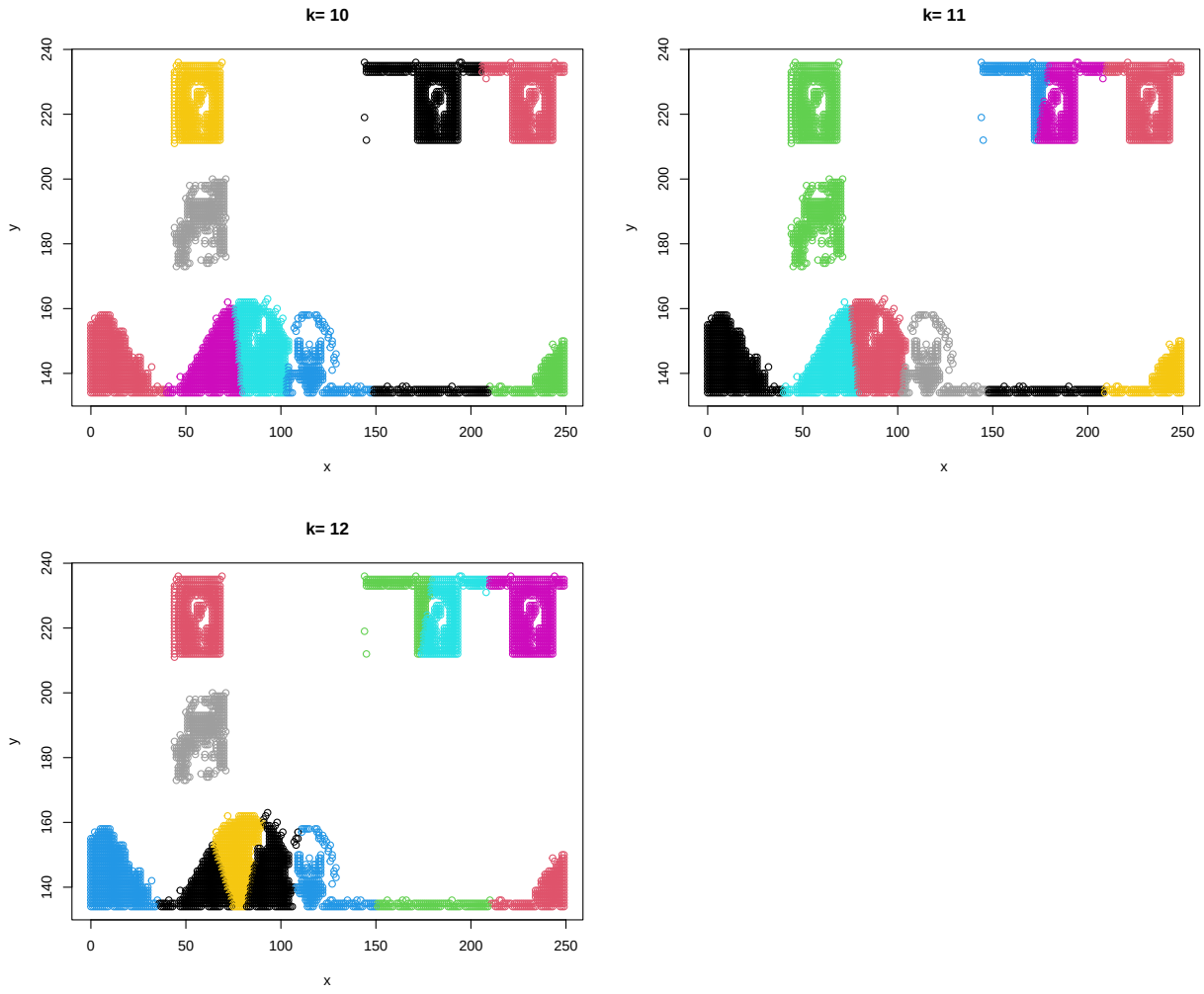
Fit the dataset using k-means from k=2 to k=12

```
par(mfrow = c(2, 2))

# Set seed
set.seed(1)
df.mean <- list()
i <- 1
for (k in 2:12) {
    km.out <- kmeans(df, centers = k, nstart = 20)
    df.distance <- data.frame()
    for (cl in 1:k) {
        cl_points <- df[km.out$cluster == cl, ]
        center_point <- km.out$centers[cl, ]
        x_dist <- (cl_points["x"] - center_point["x"])^2
        y_dist <- (cl_points["y"] - center_point["y"])^2
        distance <- sqrt(x_dist + y_dist)
        df.distance <- rbind(df.distance, distance)
    }
    df.mean[i] <- mean(df.distance$x)
    i <- i + 1
    plot(df, col = km.out$cluster, main = paste("k=", k))
    # ggplot(data = df, aes(x, y, color = km.out$cluster)) +
    # geom_point() ggplot(data = df, aes(x, y)) + geom_point()
}
```

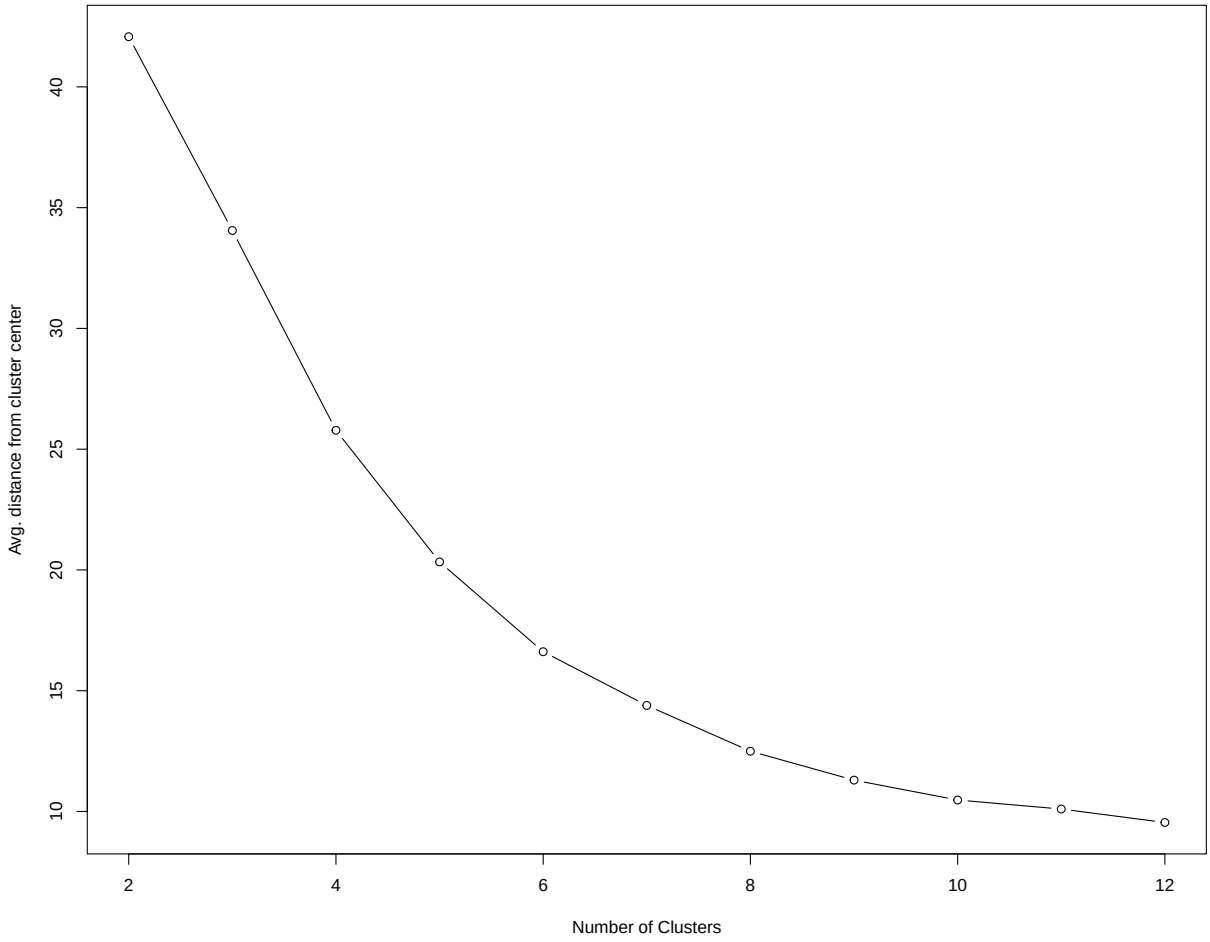

k= 2
k= 3
k= 4
k= 5
x
y
0
50
100
150
200
250
140
160
180
200
220
240


k= 6
k= 7
k= 8
k= 9
x
y
0
50
100
150
200
250
140
160
180
200
220
240



Plot with x as k and average distance as y

```
plot(2:12, df.mean, type = "b", xlab = "Number of Clusters", ylab = "Avg. distance from cluster center"
```



Elbow point.

From graph I can conclude elbow point is 6.