9.3 Clustering

Arindam Samanta

February 09 2020

#### Loading the required libraries for our analysis

library(ggplot2)  
library(dplyr)

There are 4022 observations and 2 variables. The other 2 variables are x and y.

#populating the data file to dataframe  
wd <- getwd()  
fname <- "clustering-data.csv"  
path\_to\_file <- paste(wd,'/dataset/',fname, sep = "")  
path\_to\_file  
  
my\_clusterdf <- read.csv(path\_to\_file, header = TRUE)  
  
summary(my\_clusterdf)  
head(my\_clusterdf)  
str(my\_clusterdf)

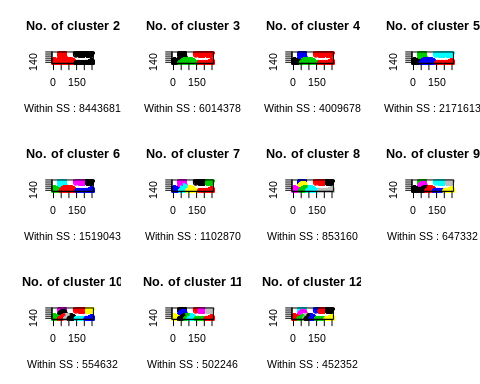
### a) Plot the data for each data set using a scatter plot.

Plotting the data to visualize the relationship between the variables. The plot shows that the values do not show any relationship.

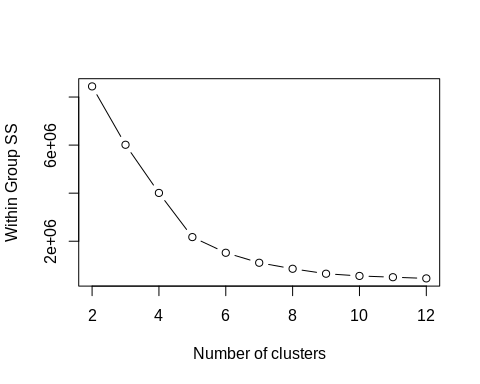


### b) Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

To start with fit a k-means model to cluster\_data using the center k = 2 to 12. Then create scatter plot for each cluster.



### c. Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.



### d. Looking at the graph you generated in the previous example, what is the elbow point for this dataset?

The elbow point looks like when k = 6