Exercise 16: Clustering

Kolekar, Ajit

November 1st, 2020

# Assignment description

**Labeled data is not always available. For these types of datasets, you can use unsupervised algorithms to extract structure. The k-means clustering algorithm and the k nearest neighbor algorithm both use the Euclidean distance between points to group data points. The difference is the k-means clustering algorithm does not use labeled data.**

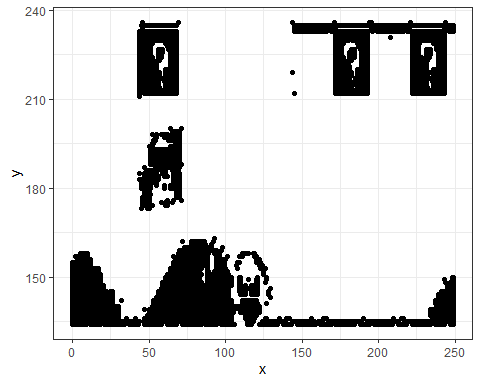
**In this problem, you will use the k-means clustering algorithm to look for patterns in an unlabeled dataset. The dataset for this problem is found at data/clustering-data.csv.**

clustering\_df <- read.csv("C:/Users/Shilp/Documents/GitHub/dsc520/data/clustering-data.csv")

## Question a

**Plot the dataset using a scatter plot.**

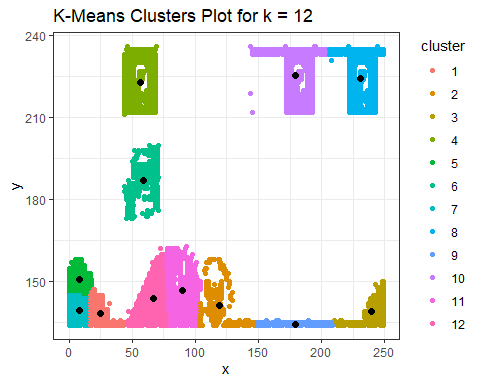
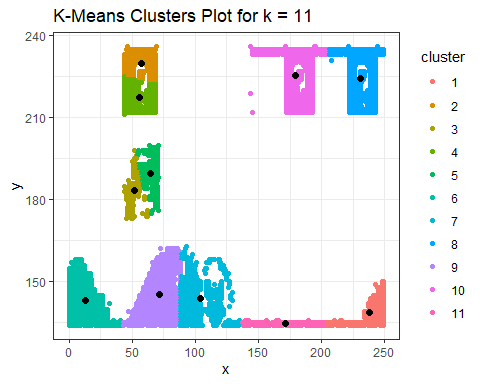
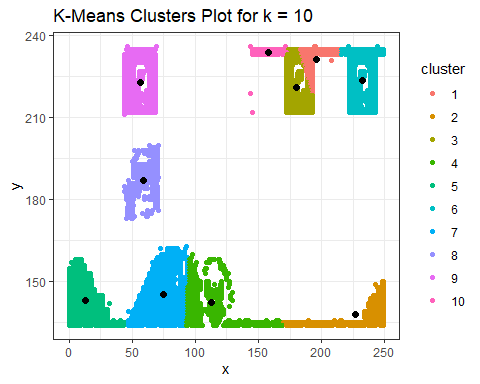
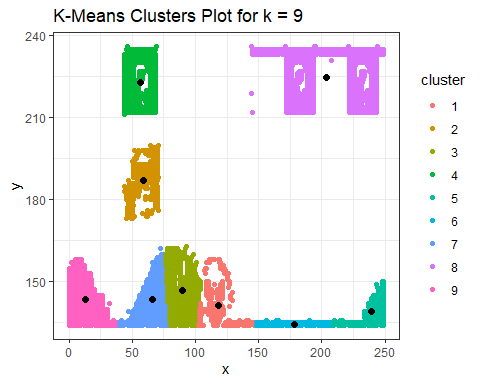
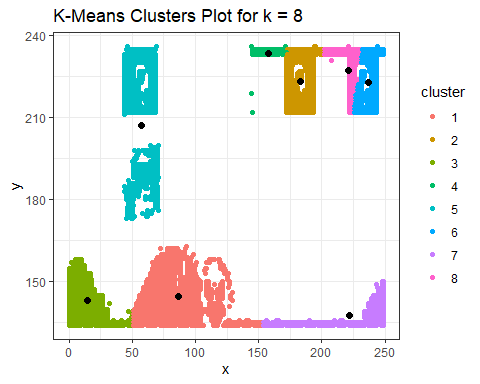
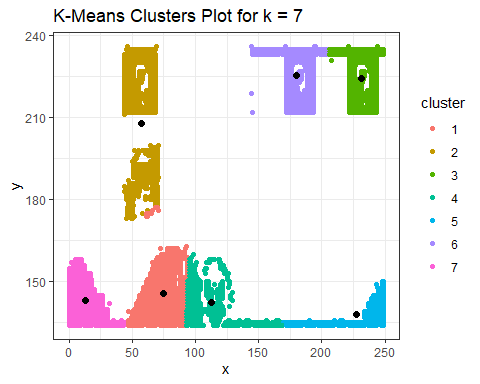
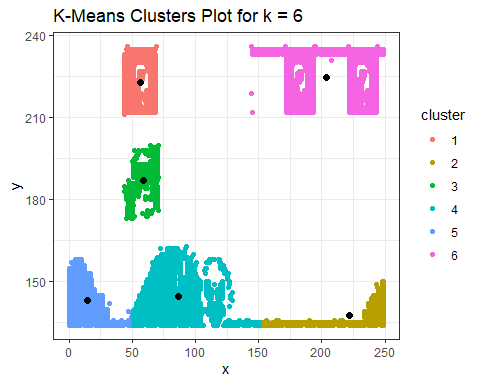
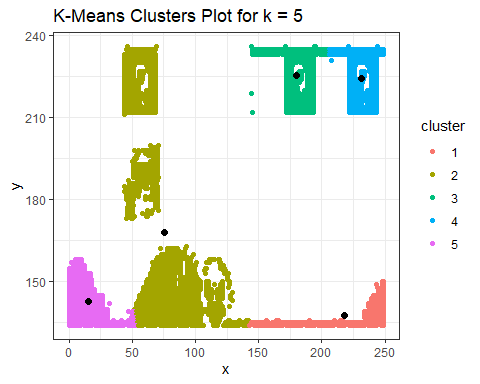
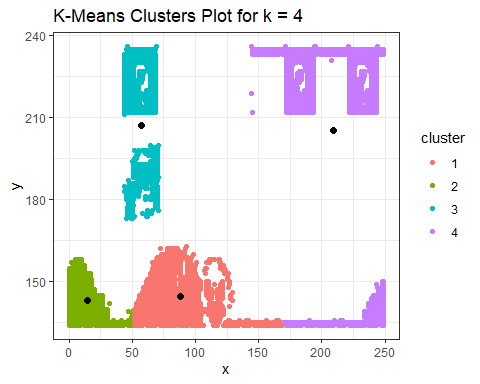
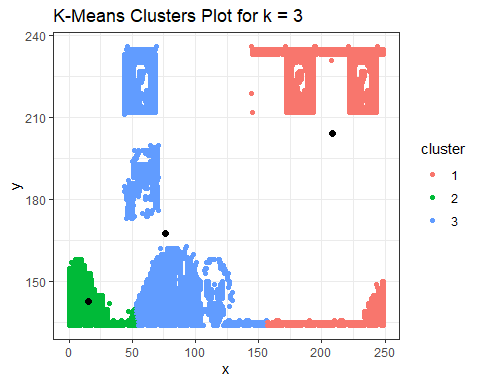
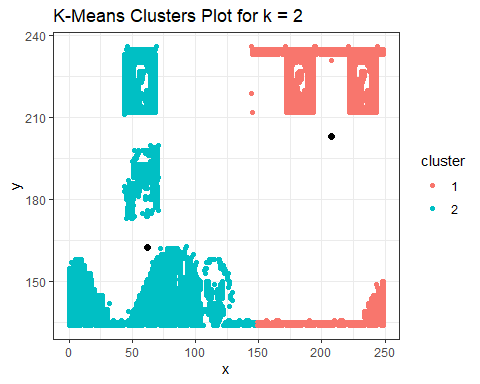
library('ggplot2')  
ggplot(clustering\_df, aes(x=x, y=y)) + geom\_point() + theme\_bw()



## Question 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.**

set.seed(20)  
  
k\_values <- c(2:12)  
total\_withinss <- NULL  
avg\_distance <- NULL  
  
  
for (i in 1:length(k\_values))  
{  
 temp\_df <- clustering\_df  
 clusters <- kmeans(temp\_df, k\_values[i])  
 temp\_df$cluster <- as.factor(clusters$cluster)  
 x = ggplot(data = temp\_df, aes(x=x, y=y, color = cluster)) + geom\_point() + theme\_bw() + geom\_point(data = as.data.frame(clusters$centers), color = "black", size = 2) + ggtitle(paste("K-Means Clusters Plot for k =", k\_values[i]))  
 print(x)  
   
 # following values are calculated for question c  
 x\_dist <- clusters$centers[temp\_df$cluster] - temp\_df$x  
 y\_dist <- clusters$centers[temp\_df$cluster] - temp\_df$y  
 tot\_dist <- sqrt((x\_dist \*\* 2) + (y\_dist \*\* 2))  
 avg\_distance <- c(avg\_distance, mean(tot\_dist))  
 total\_withinss <- c(total\_withinss, clusters$tot.withinss)  
}



## Question c

**As k-means is an unsupervised algorithm, you cannot compute the accuracy as there are no correct values to compare the output to. Instead, you will use the average distance from the center of each cluster as a measure of how well the model fits the data. To calculate this metric, simply compute the distance of each data point to the center of the cluster it is assigned to and take the average value of all of those distances.**

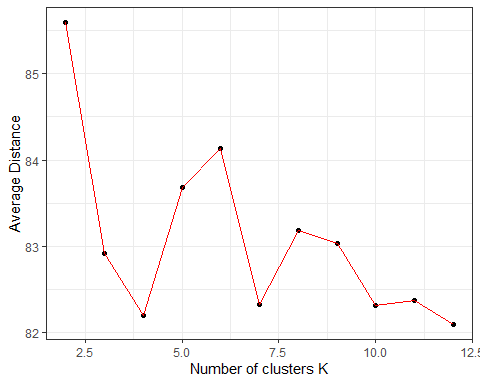
**Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.**

**One way of determining the “right” number of clusters is to look at the graph of k versus average distance and finding the “elbow point”. Looking at the graph you generated in the previous example, what is the elbow point for this dataset?**

dist\_elbow\_df <- data.frame(k\_values, avg\_distance)  
dist\_elbow\_df

## k\_values avg\_distance  
## 1 2 85.59601  
## 2 3 82.91766  
## 3 4 82.20270  
## 4 5 83.68432  
## 5 6 84.12965  
## 6 7 82.32686  
## 7 8 83.18385  
## 8 9 83.02901  
## 9 10 82.30974  
## 10 11 82.36781  
## 11 12 82.09636

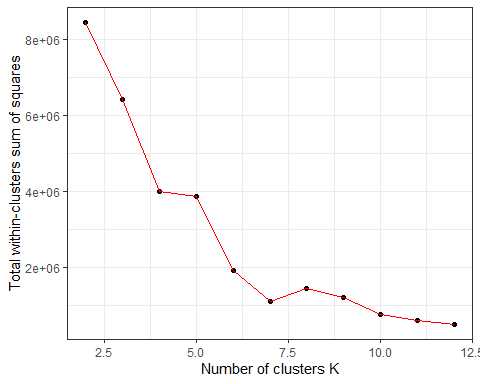
ggplot(data = dist\_elbow\_df, aes(x=k\_values, y=avg\_distance)) + xlab("Number of clusters K") + ylab("Average Distance") + theme\_bw() + geom\_point() + geom\_line(color = "red")



withinss\_elbow\_df <- data.frame(k\_values, total\_withinss)  
withinss\_elbow\_df

## k\_values total\_withinss  
## 1 2 8443681.1  
## 2 3 6411644.9  
## 3 4 4009678.4  
## 4 5 3856788.5  
## 5 6 1920708.1  
## 6 7 1102869.9  
## 7 8 1436982.5  
## 8 9 1207285.9  
## 9 10 770743.2  
## 10 11 611735.3  
## 11 12 504963.4

ggplot(data = withinss\_elbow\_df, aes(x=k\_values, y=total\_withinss)) + xlab("Number of clusters K") + ylab("Total within-clusters sum of squares") + theme\_bw() + geom\_point() + geom\_line(color = "red")



By looking at the plots for Average Distance and for Total within-clusters sum of squares, the elbow point for the dataset appears to be 7.