Week 11 - Exercise 11.2

Ganesh Kale

May 30, 2021

K-Nearest Neighbors Algorithm

Load the required packages

```
library(dplyr)
library(ggplot2)
library(class)
library(caTools)
library(e1071)
library(factoextra)
```

Load the Data Sets

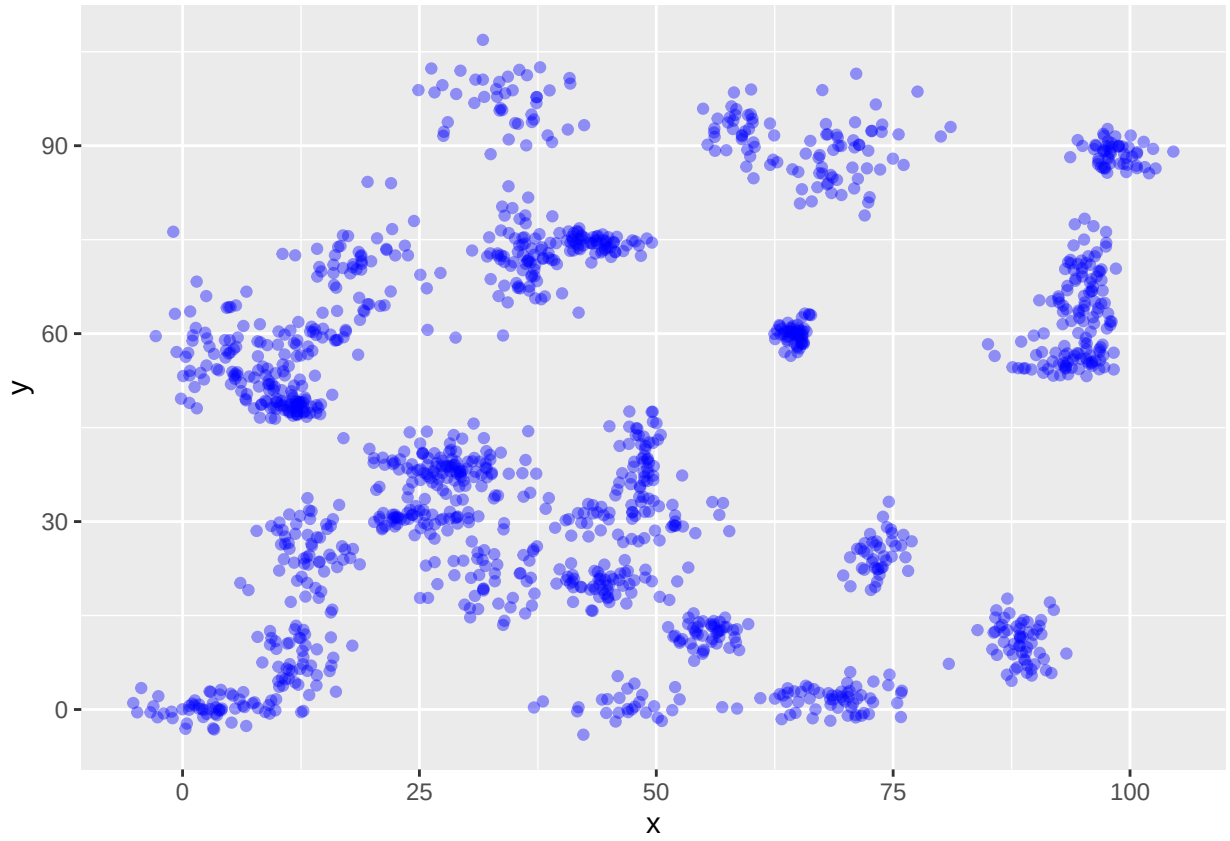
binary classiffier data:

```
##   label        x        y
## 1     0 70.88469 83.17702
## 2     0 74.97176 87.92922
## 3     0 73.78333 92.20325
## 4     0 66.40747 81.10617
## 5     0 69.07399 84.53739
```

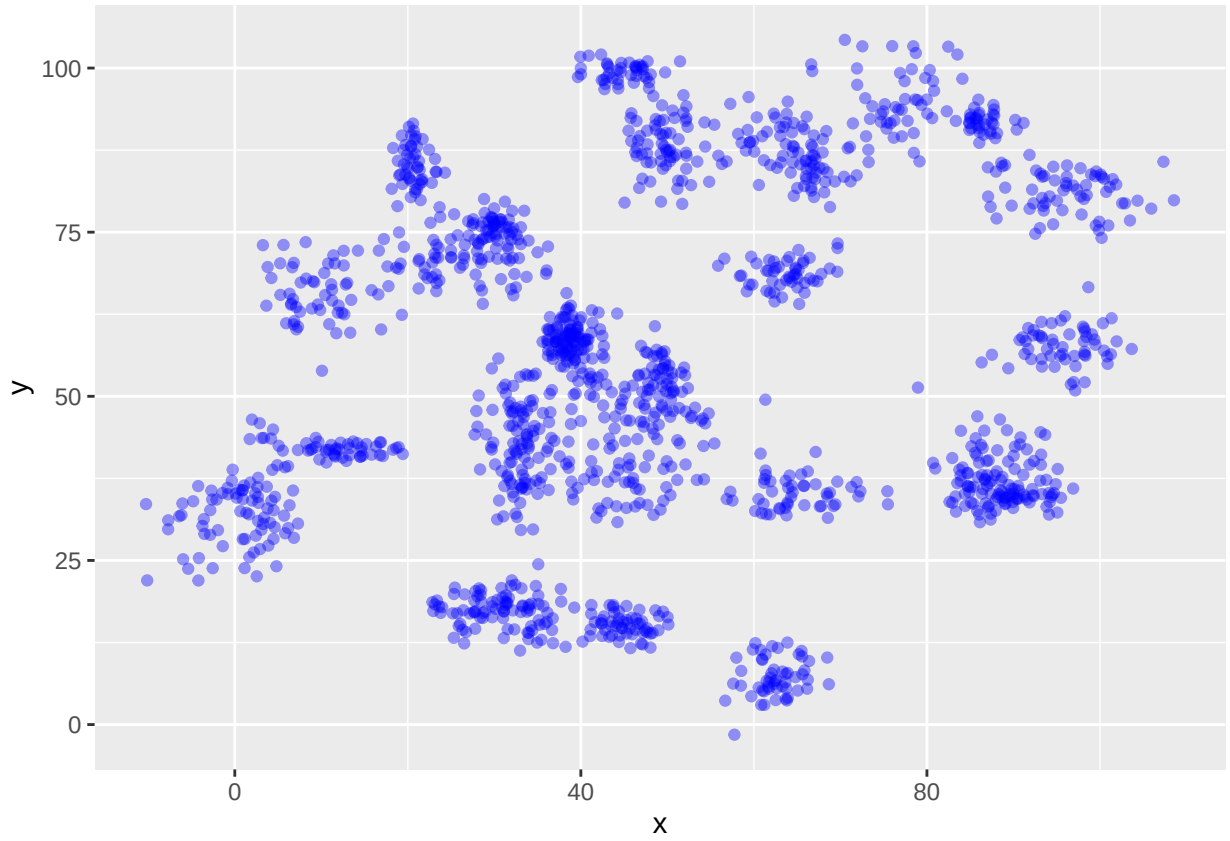
trinary classifier data:

```
##   label        x        y
## 1     0 30.08387 39.63094
## 2     0 31.27613 51.77511
## 3     0 34.12138 49.27575
## 4     0 32.58222 41.23300
## 5     0 34.65069 45.47956
```

Scatter plot - Binary dataset



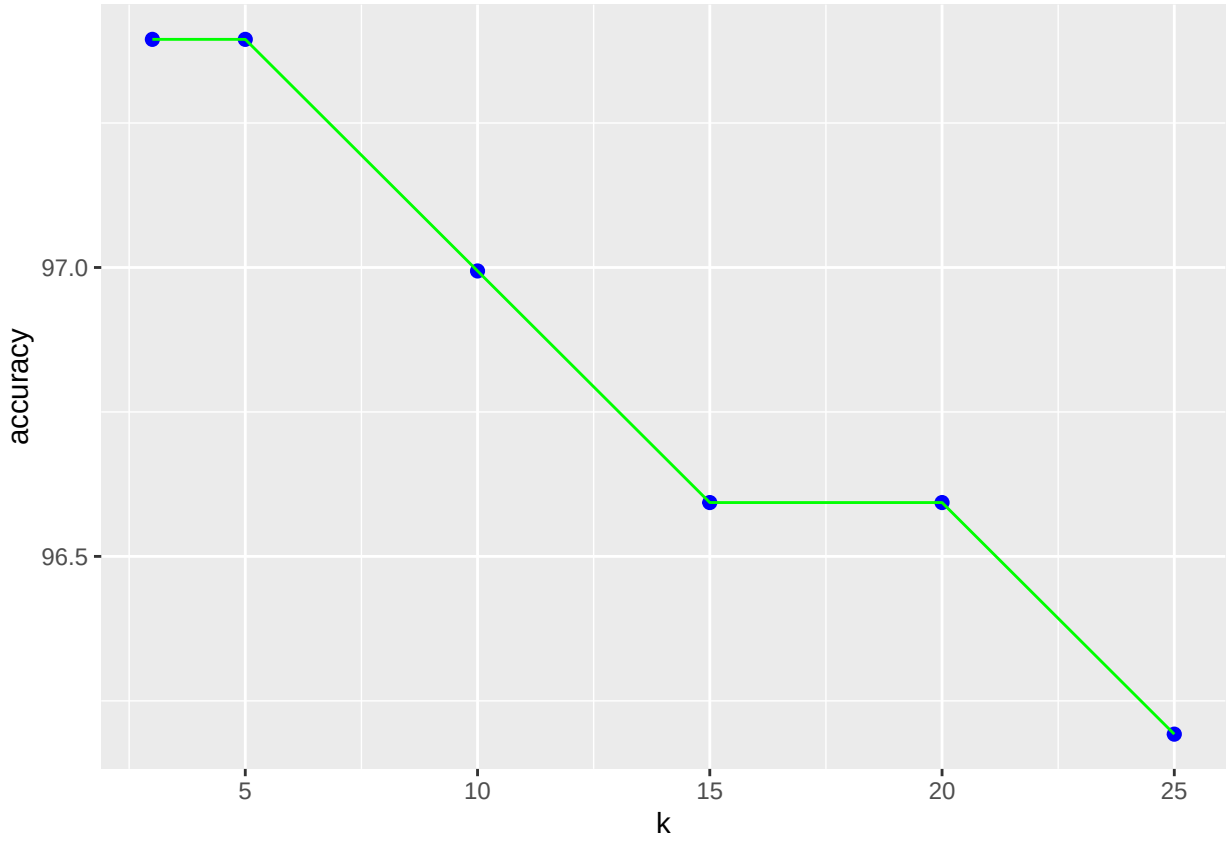
Scatter plot - Trinary dataset



fit k-nearest neighbor model to binary data set for k = 3,5,10,15,20,25:

plot the graph

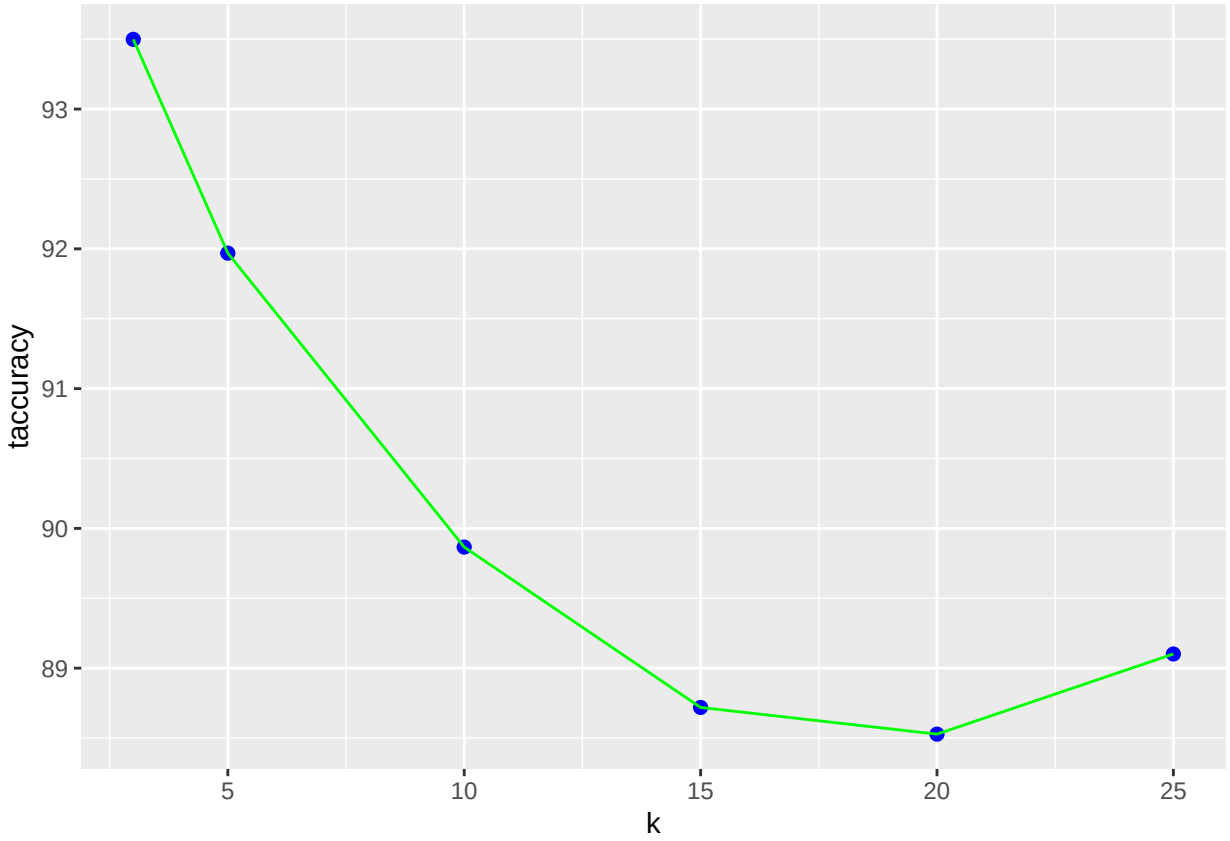
```
##    k accuracy
## 1  3 97.39479
## 2  5 97.39479
## 3 10 96.99399
## 4 15 96.59319
## 5 20 96.59319
## 6 25 96.19238
```



fit k-nearest neighbor model to trinary data set for k = 3,5,10,15,20,25:

plot the graph

```
##    k taccuracy
## 1  3  93.49904
## 2  5  91.96941
## 3 10  89.86616
## 4 15  88.71893
## 5 20  88.52772
## 6 25  89.10134
```



Looking back at the plots of the data, do you think a linear classifier would work well on these datasets?

Based on the plots of binary and trinary data, the linear classifier will not work because, linear classifiers classify data into lables based on a linear combinations of input features. These classifiers separate data using a line or plane and can be used to classify data that is linearly separable. The data in these two datasets can not be linearly separated.

How does the accuracy of your logistic regression classifier from last week compare? Why is the accuracy different between these two methods?

The accuracy score of binary data was 0.512016 or 52.00%, with KNN algorithm or ML techniques the accuracy score is 98.4% for k = 15 and k = 25. Compared to Logistic regression the accuracy score of KNN is much higher and close to 100%. The reason of different accuracy score between these two models is because KNN supports non-linear solutions while logistic only supports linear solutions.

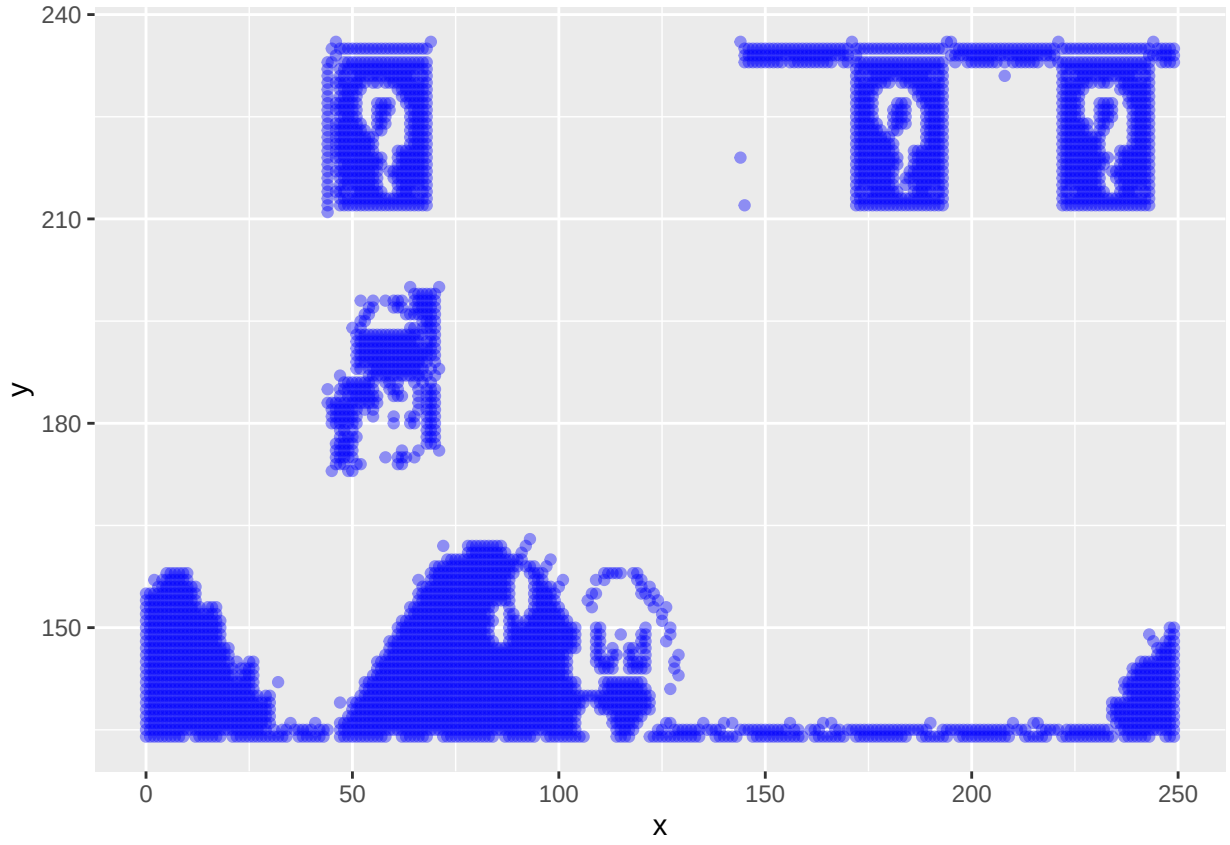
Clustering Exercise

Load the clustering dataset:

```
##     x   y
## 1  46 236
```

```
## 2  69 236
## 3 144 236
## 4 171 236
## 5 194 236
```

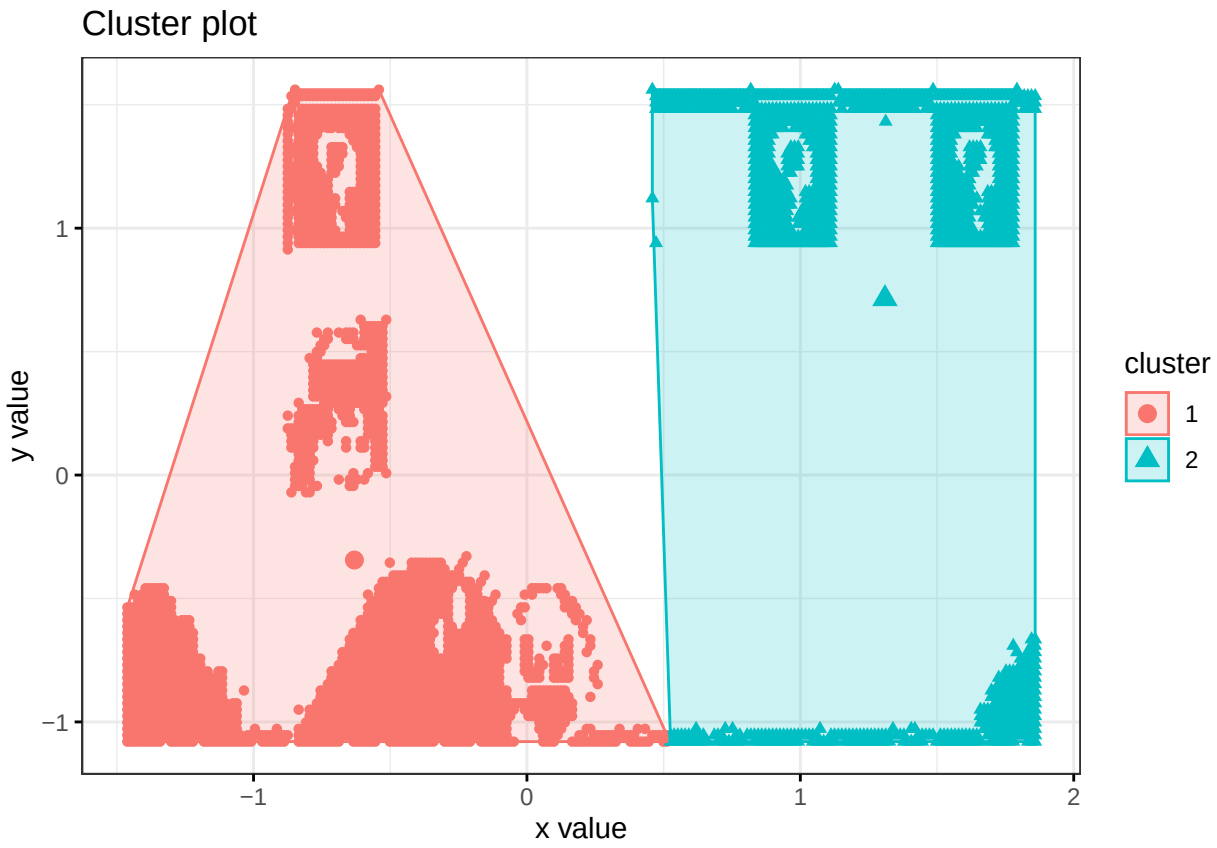
Plot the dataset using a scatter plot



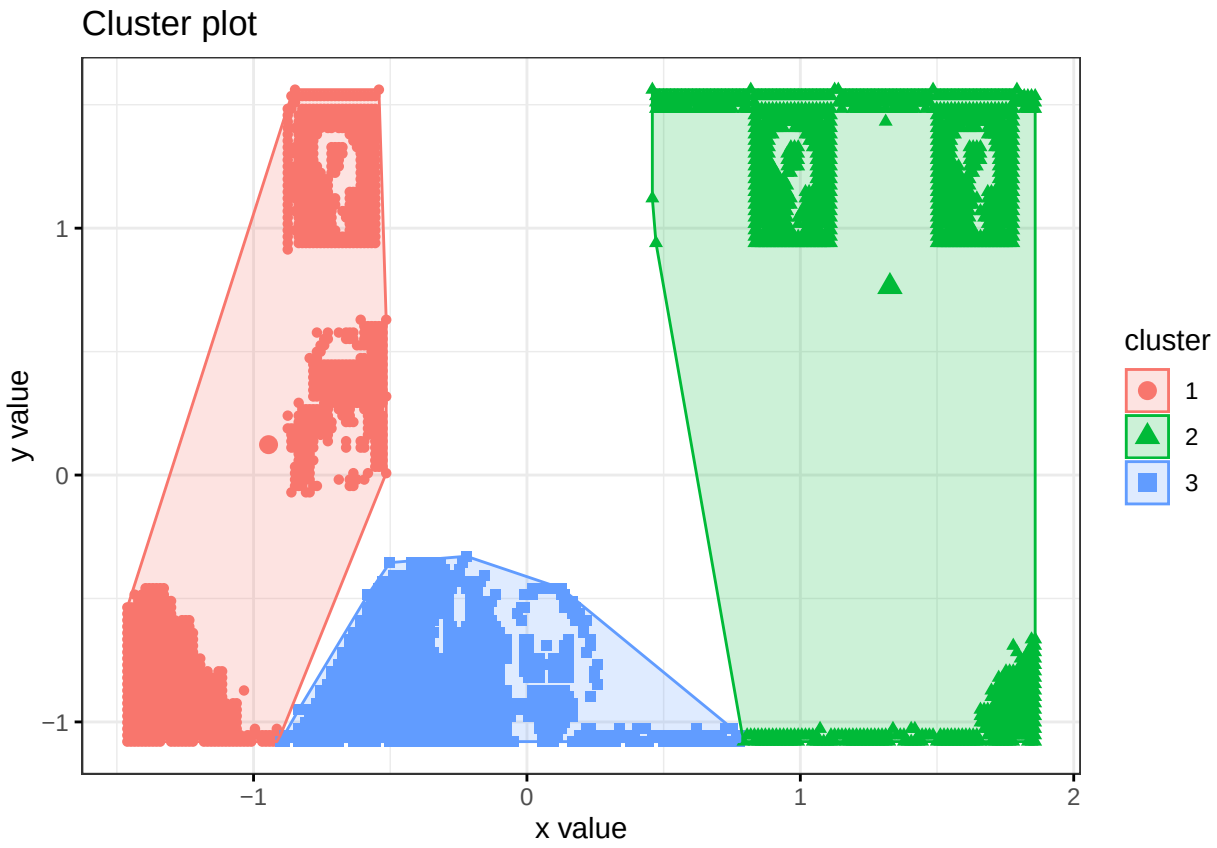
Fit the data set using the k-means algorithm from k=2 to k=12

Create a scatter plot of the resultant clusters for each value of k

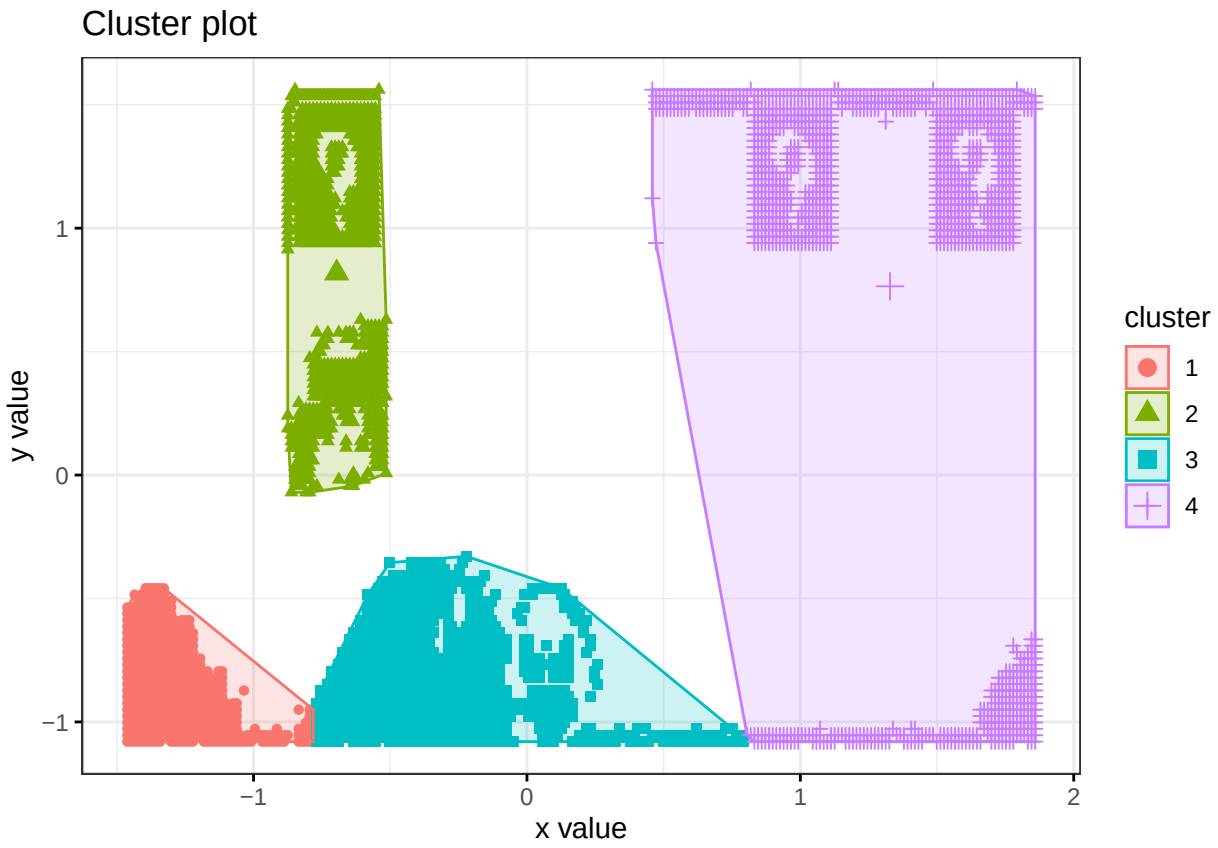
Cluster Plot for k = 2



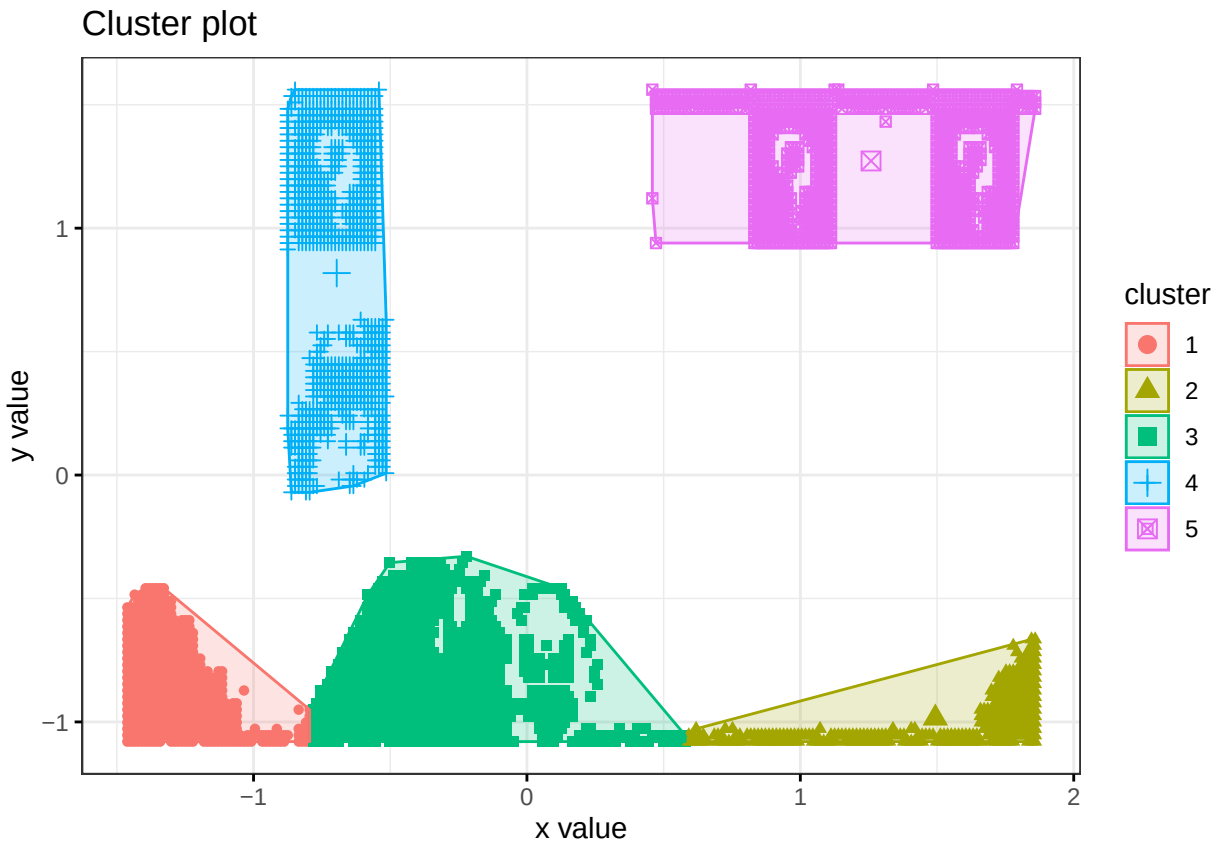
Cluster Plot for k = 3



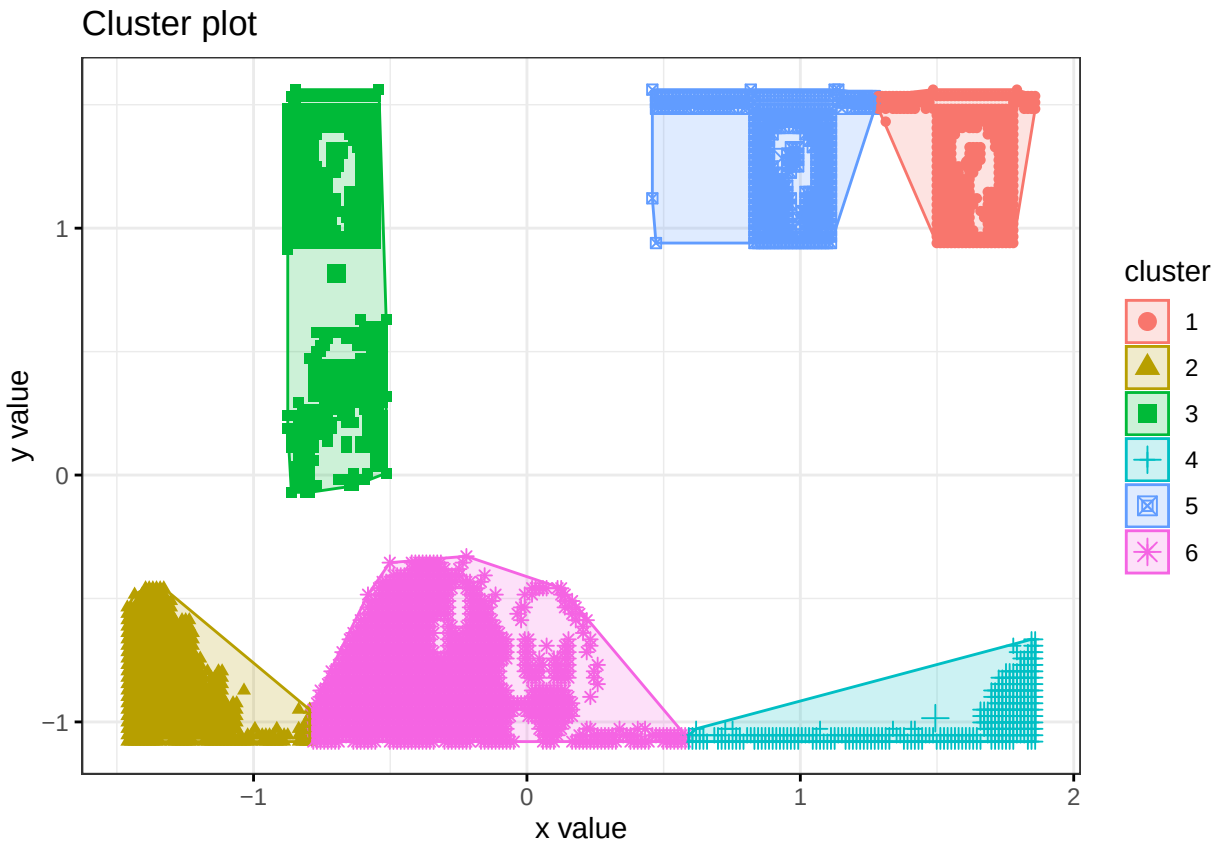
Cluster Plot for k = 4



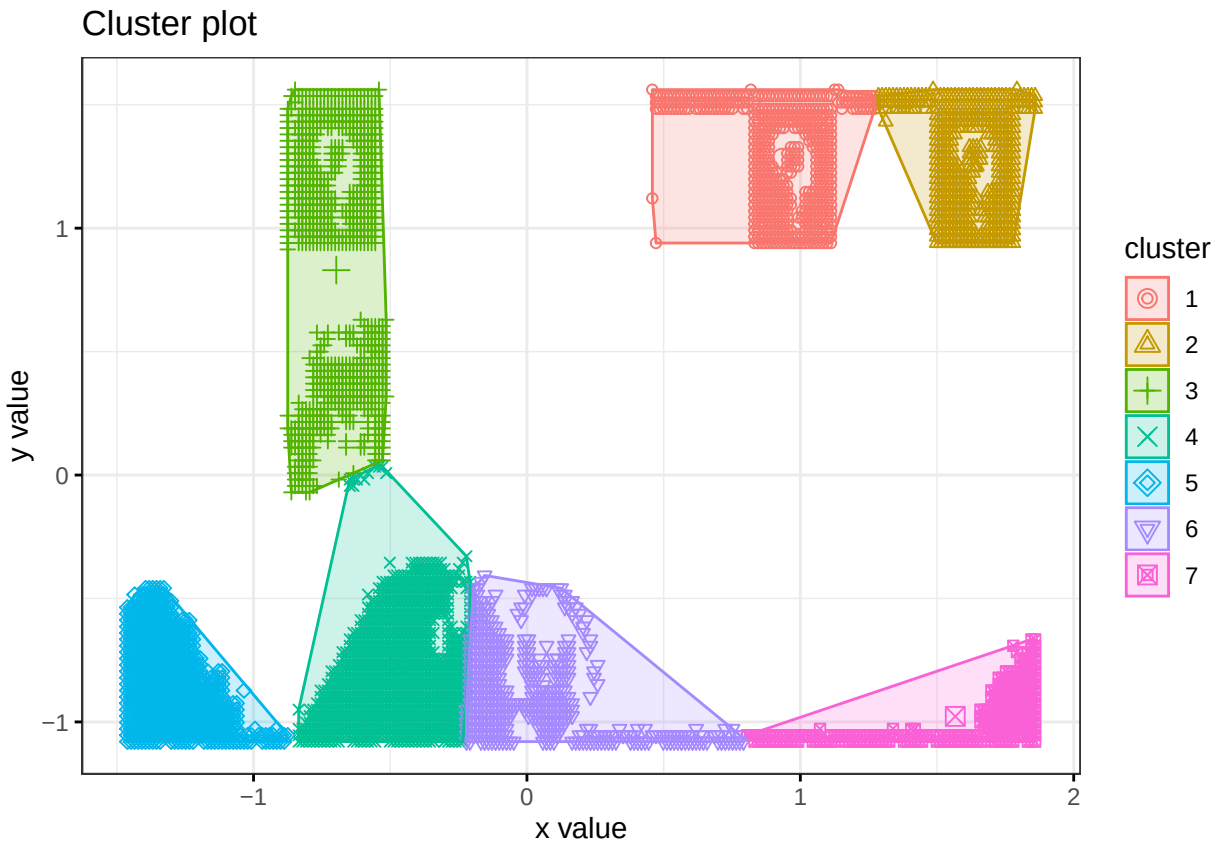
Cluster Plot for k = 5



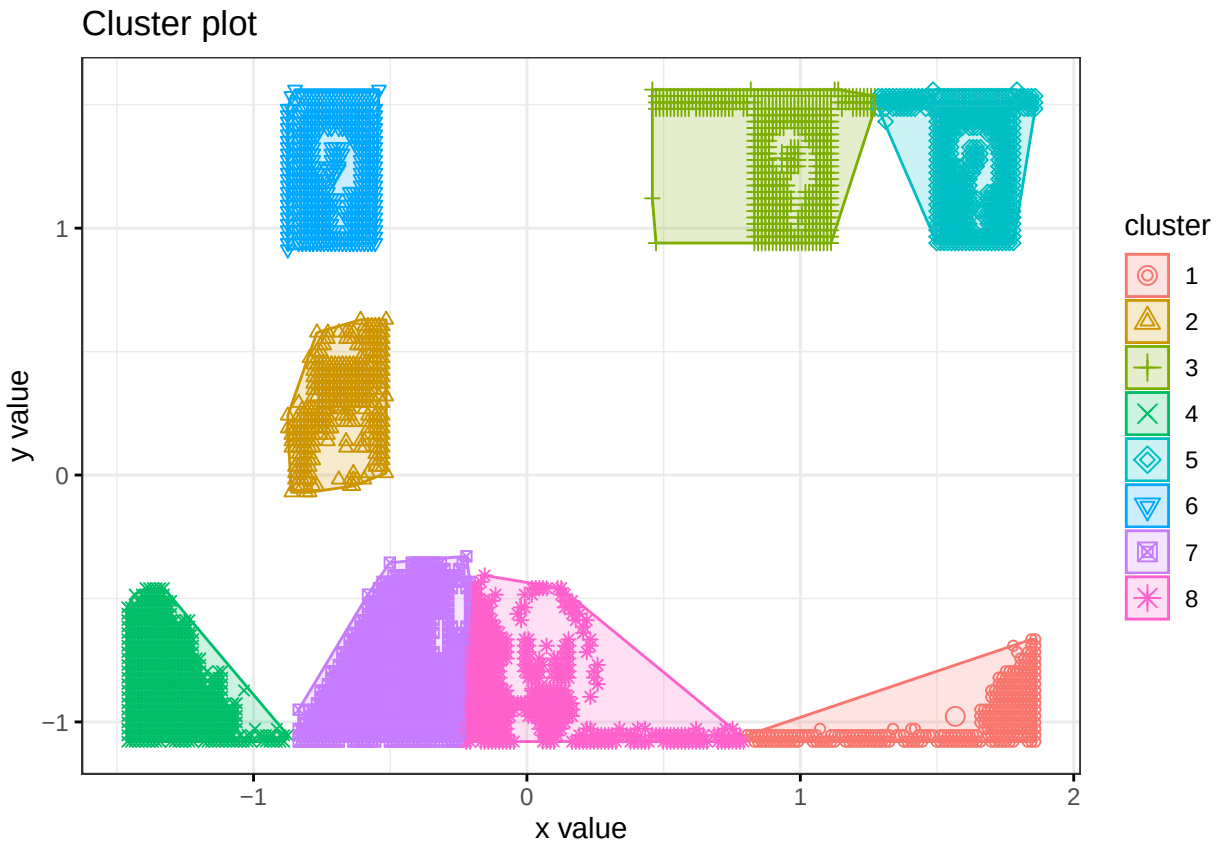
Cluster Plot for k = 6



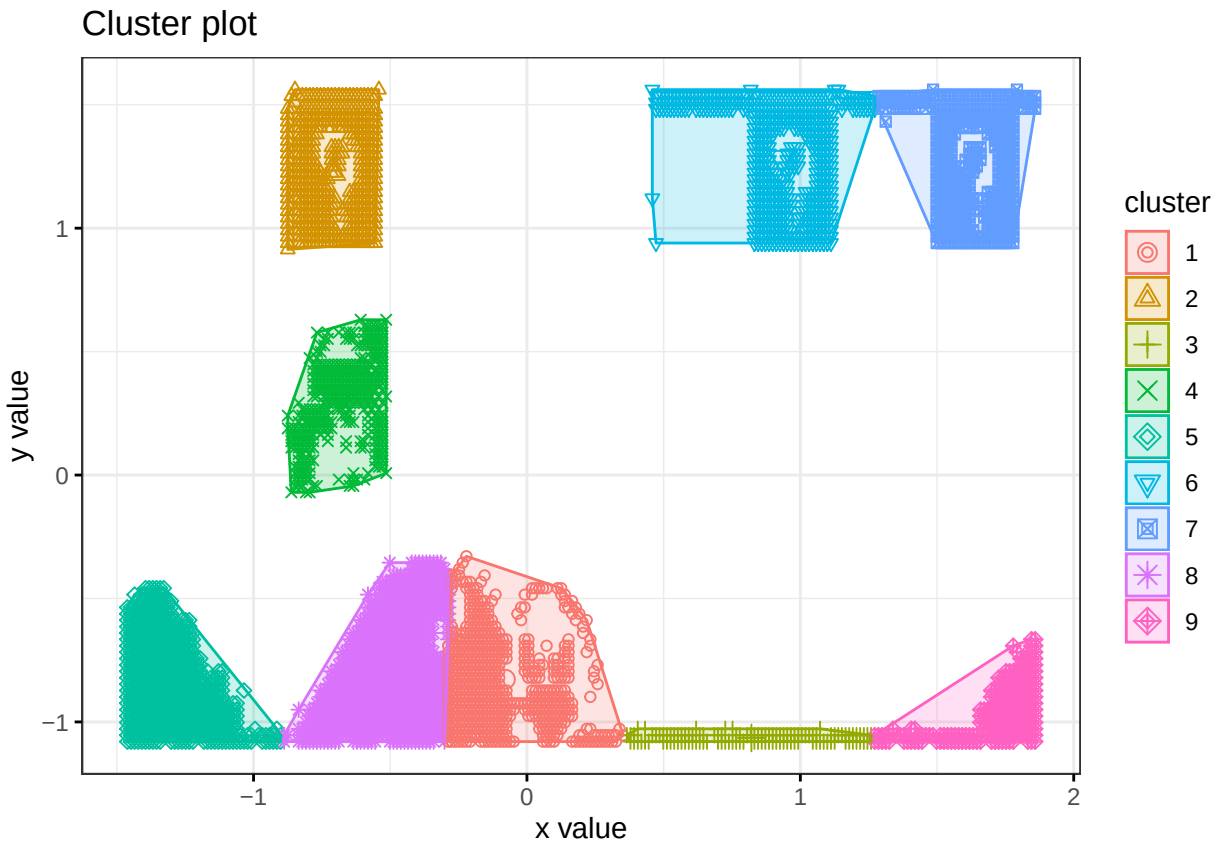
Cluster Plot for k = 7



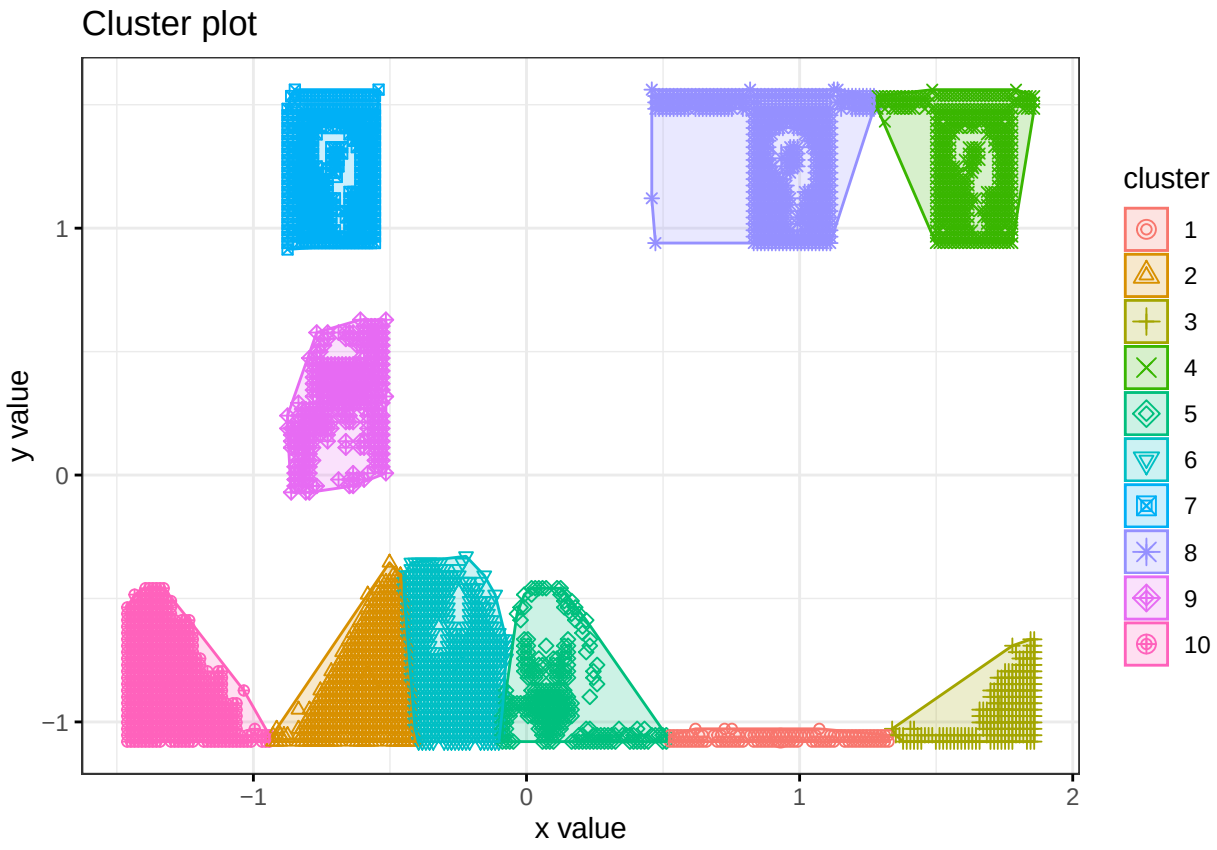
Cluster Plot for k = 8



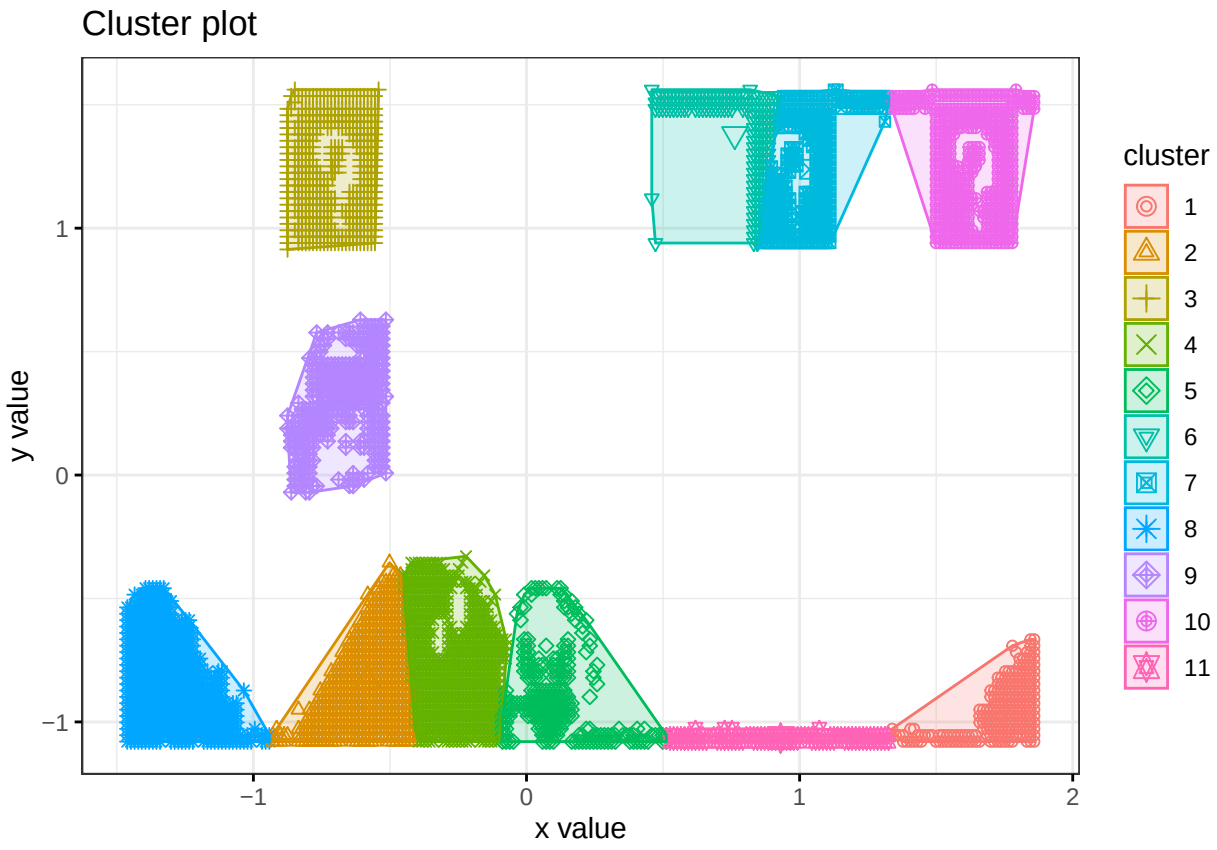
Cluster Plot for k = 9



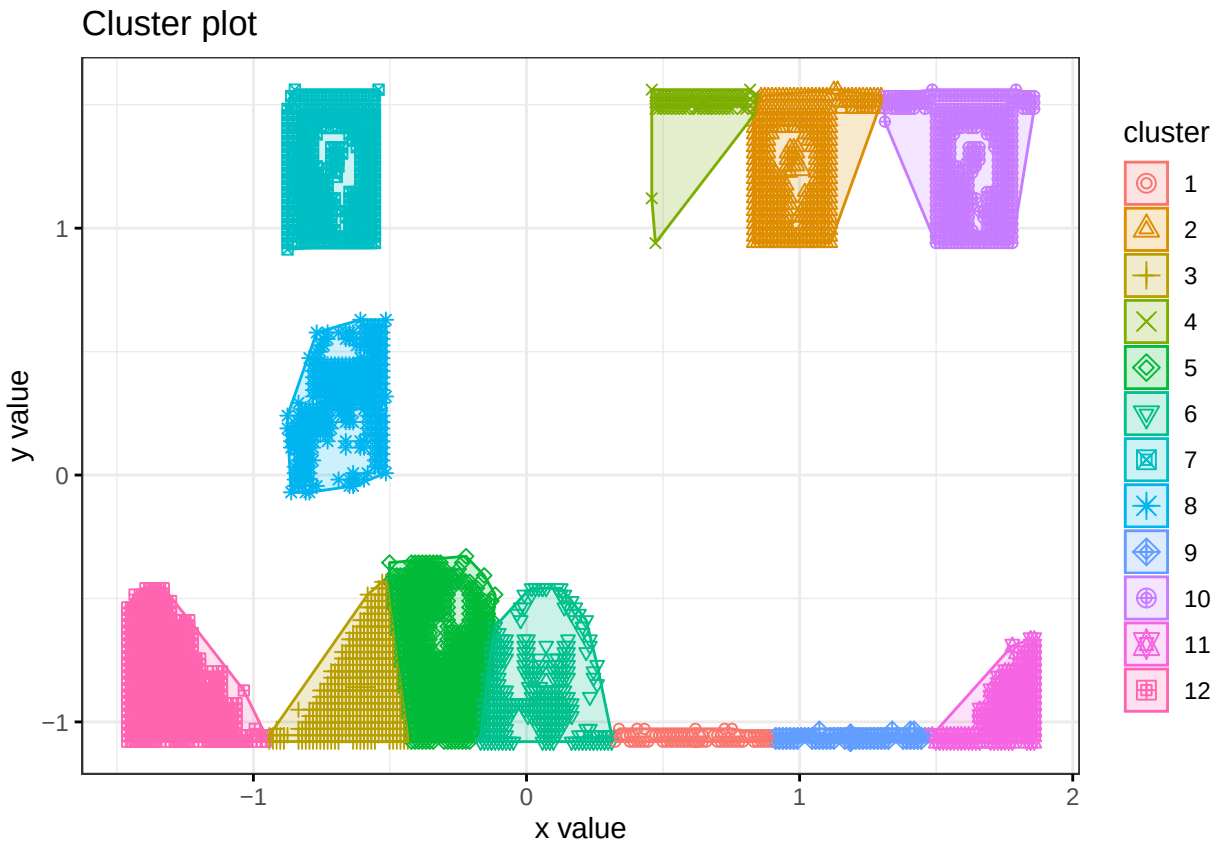
Cluster Plot for k = 10



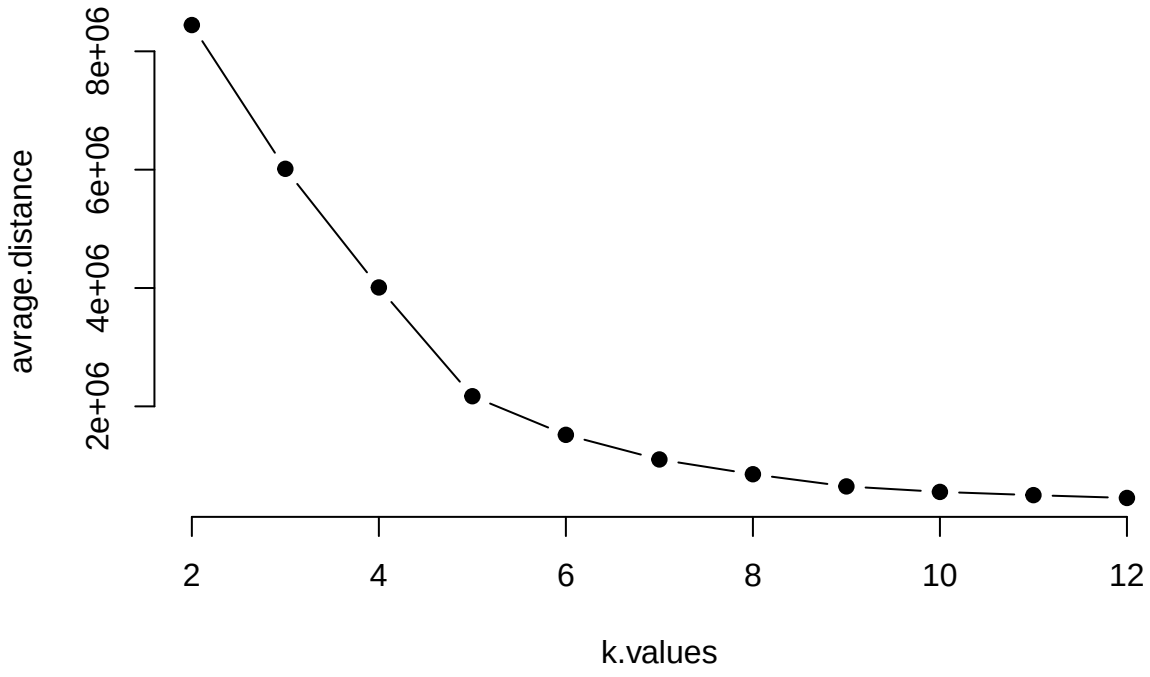
Cluster Plot for k = 11



Cluster Plot for k = 12



plot it as a line chart where k is the x-axis and the average distance is the y-axis.



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

Based on the line chart, the elbow point is at 5, which means that right number of clusters for this data set is 5.