

**Technological Institute of Tijuana****Academic Subdirector****Systems and Computing Department****SEMESTER:** August - December 2021**CAREER:** Computer Systems Engineer**MATTER:** Data Mining**JOB NAME:** Unit 4 - Exam**STUDENT NAME AND CONTROL NUMBER:**

Castro Cebreros Alejandro - 16211341

Márquez Millán Seashell Vanessa - 17212153

TEACHER NAME: Jose Christian Romero Hernandez**DATE OF DELIVERY:** December 12, 2021**Development**

The first part is to decide where we want to work and we put the direction

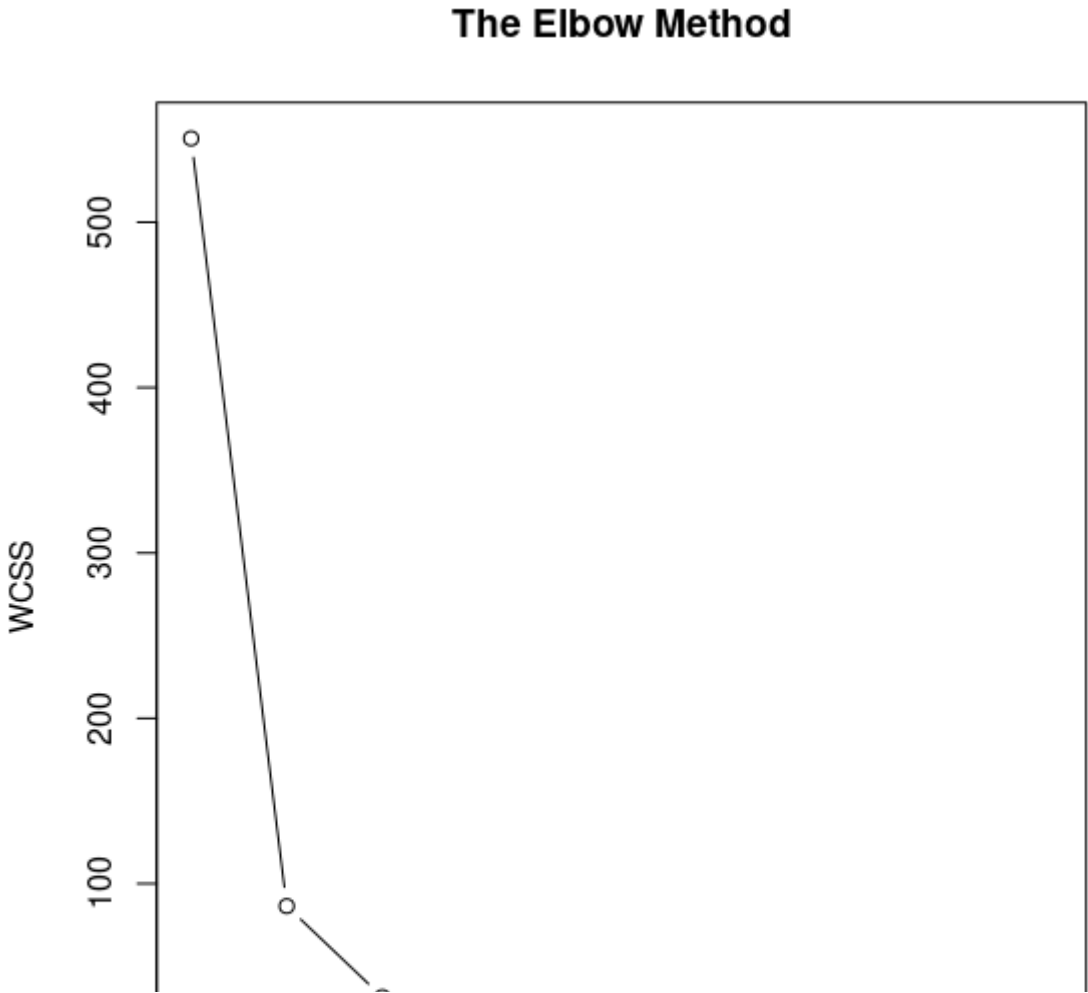
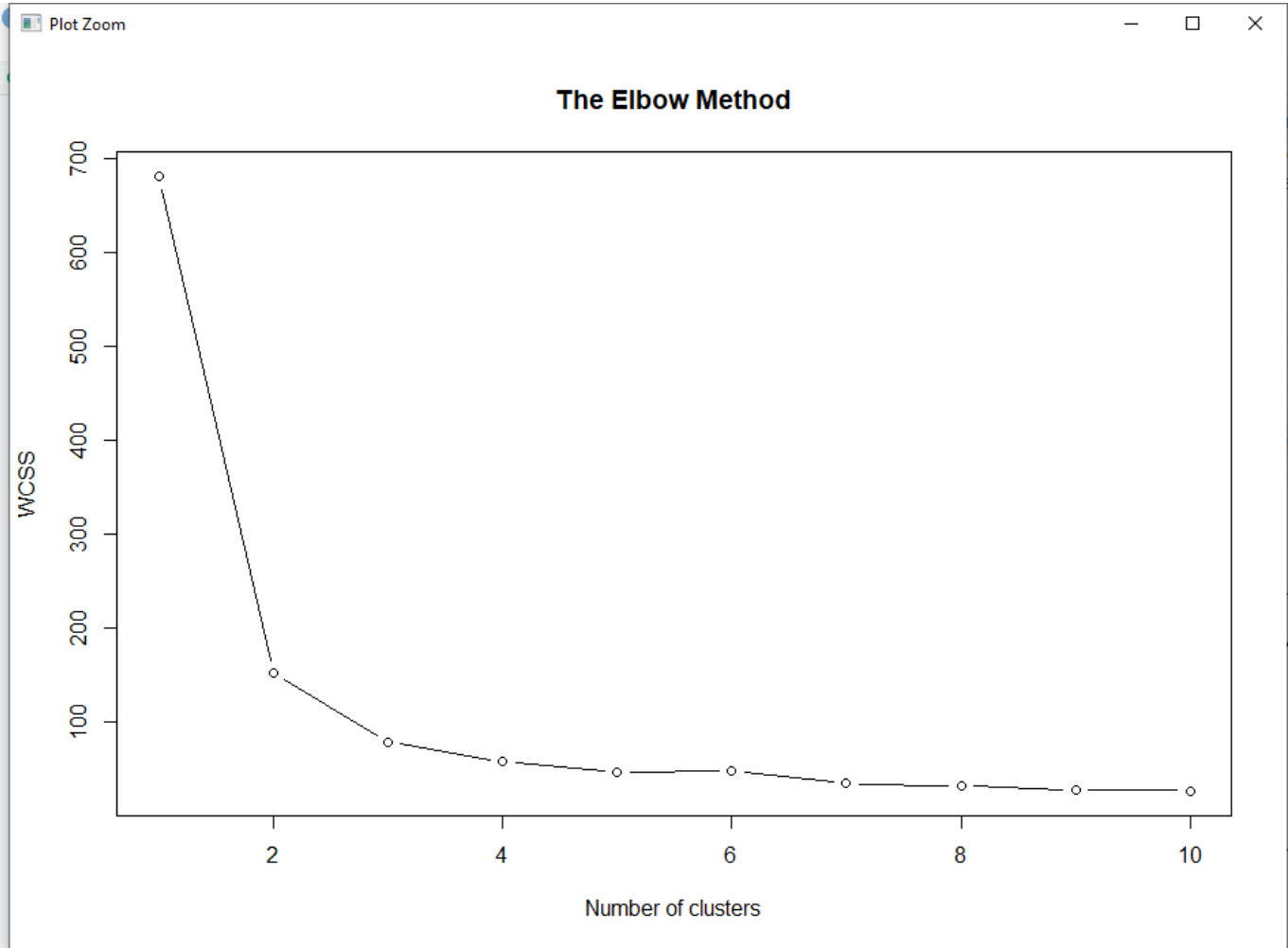
```
getwd()  
setwd("C:/Users/vanem/OneDrive/Documentos/9 SEMESTRE/Mineria/Repo  
mineria/DataMining/MachineLearning/LogisticRegression")  
getwd()
```

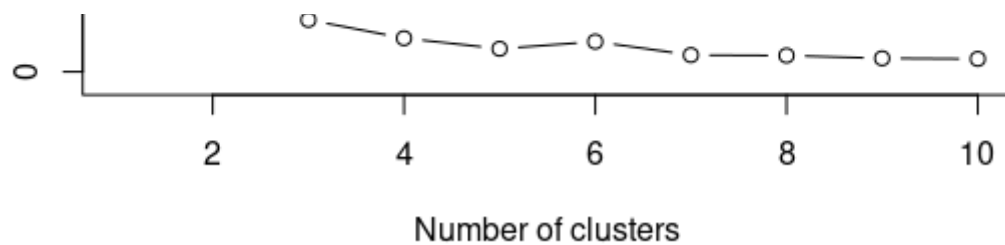
Now we import the data set and select the columns

```
dataset <- read.csv(file.choose())  
dat1 = dataset[1:4]  
dat2 = dataset[3:4]
```

Well here we used the the elbow method to find the optimal number of cluster, but these method is on the function because is more easy to visualizing the graphic, and after we call the function and send like param the data set and the resul is here.

```
TEM <- function(dataset){  
  set.seed(6)  
  wcss = vector()  
  for (i in 1:10) wcss[i] = sum(kmeans(dataset, i)$withinss)  
  plot(1:10,  
       wcss,  
       type = 'b',  
       main = paste('The Elbow Method'),  
       xlab = 'Number of clusters',  
       ylab = 'WCSS')  
}  
  
TEM(dat1)  
TEM(dat2)
```





Like we can see the breaking point is in the point number 3 so is these the number we used more later.

Here we used the model and fitting the dataset

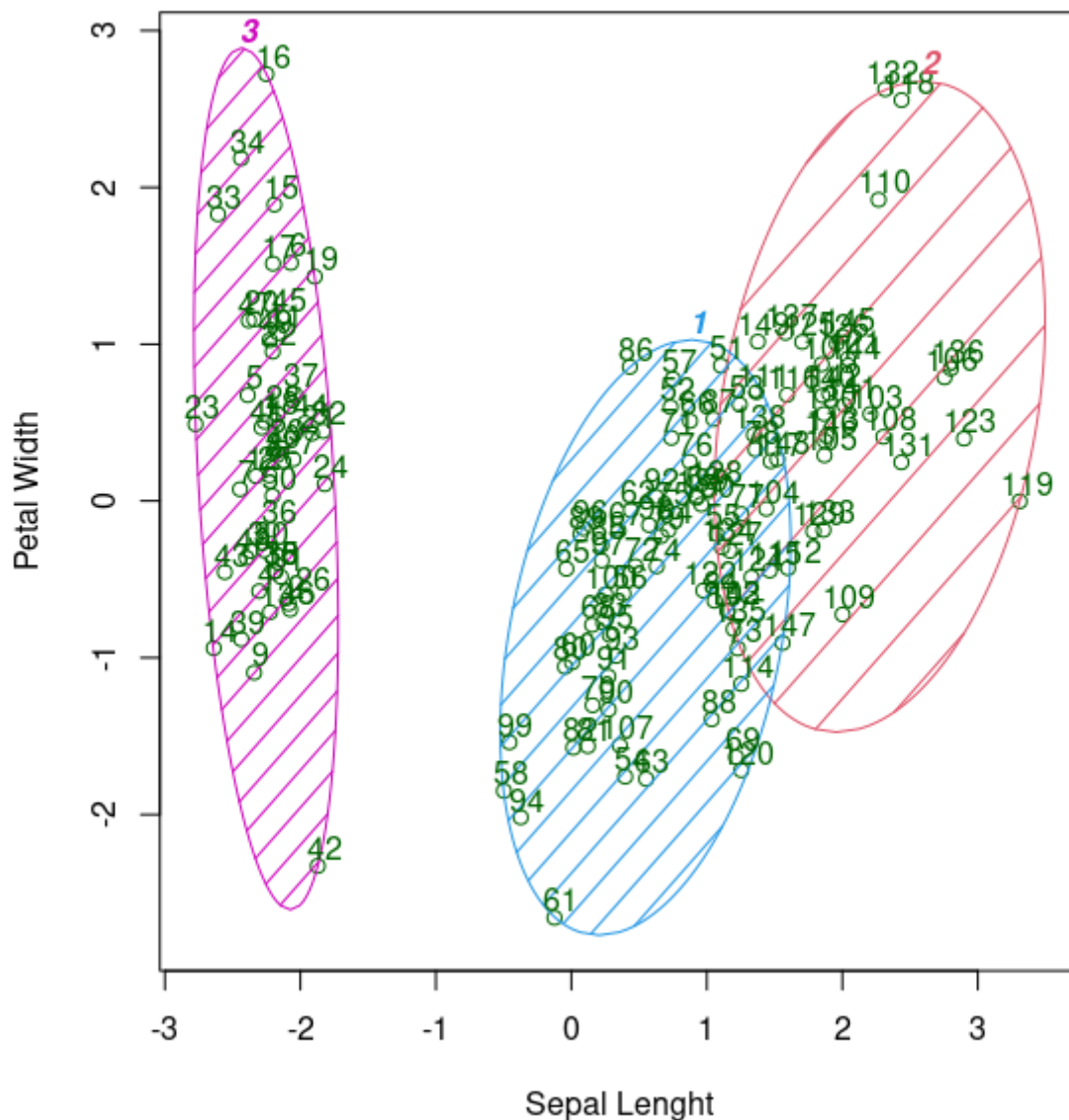
```
set.seed(29)
kmeans = kmeans(x = dataset, centers = 3)
y_kmeans = kmeans$cluster
```

Here we visualize the cluster, but first we need the cluster library in order to execute the code well

```
library(cluster)

#Sepal lenght and Petal width
clusplot(dat1,
         y_kmeans,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 2,
         plotchar = FALSE,
         span = TRUE,
         main = paste('Clusters of customers'),
         xlab = 'Annual Income',
         ylab = 'Spending Score')
```

Clusters of Iris



These two components explain 95.8 % of the point variability.

```
#Petal lenght and Petal width
clusplot(dat2,
  y_kmeans,
  lines = 0,
  shade = TRUE,
  color = TRUE,
  labels = 2,
  plotchar = FALSE,
  span = TRUE,
  main = paste('Clusters of customers'),
  xlab = 'Annual Income',
  ylab = 'Spending Score')
```

A scatter plot showing the relationship between Sepal Length (X-axis, ranging from -3 to 3) and Petal Width (Y-axis, ranging from -0.6 to 0.6). The data points are categorized into three clusters, labeled 1, 2, and 3, each enclosed by a colored ellipse with diagonal hatching. Cluster 1 (blue) is centered around (0.5, 0.1). Cluster 2 (red) is centered around (2.5, 0.1). Cluster 3 (magenta) is centered around (-2.5, 0.1). The ellipses overlap significantly, indicating some overlap between the clusters. The data points are labeled with numbers from 1 to 150.

6 / 6