Tecnológico Nacional de México
Instituto Tecnológico de Tijuana

Subdirección Académica

Departamento de Sistemas y Computación

SEMESTRE:
Febrero-Julio 2021

CARRERA:
Ingeniería en Sistemas Computacionales e Ingeniería en Tecnologías de la Información y Comunicaciones

MATERIA Y SERIE:
Minería de Datos
BDD-1703 TI9A

UNIDAD A EVALUAR:
Unidad IV

NOMBRE DEL TRABAJO:
Práctica #1
Documentación del programa K-MEANS en R

NOMBRE Y NÚMERO DE CONTROL DE LOS INTEGRANTES:
Rodriguez Medrano Marco Antonio 17210635

NOMBRE DEL DOCENTE:
José Christian Romero Hernández

16/06/2021

Instrucciones: El docente le pidió al alumno que describiera el siguiente código así como los resultados obtenidos y ver el grupo de persona o empresas en las que es recomendable hacer algún tipo de publicidad o promoción en compras..

Primero insertamos nuestra carpeta de trabajo o repositorio utilizando los comandos getwd() y setwd()., esto nos asegura que toda la información del programa no se pierda y esté contenida únicamente en nuestro repositorio.

```
# K-Means Clustering
# Set our workspace
getwd()
setwd("/Users/DELL/Desktop/DataMining/MachineLearning/K-Means")
getwd()
```

Aquí importamos el archivo csv con el cual vamos a trabajar el cual lo guardaremos en la variable dataset, pero una vez guardada la información y la revisamos notaremos que hay columnas que no queremos utilizar ya sea por el tipo de los datos o por que no nos sirven dichos datos.

Entonces para limitar los datos y así quedarnos con los que vamos a utilizar utilizamos el comando dataset = dataset[4:5] en donde le indicaremos qué columnas vamos a utilizar, en este caso la columna 4 y 5 serán las que utilizaremos.

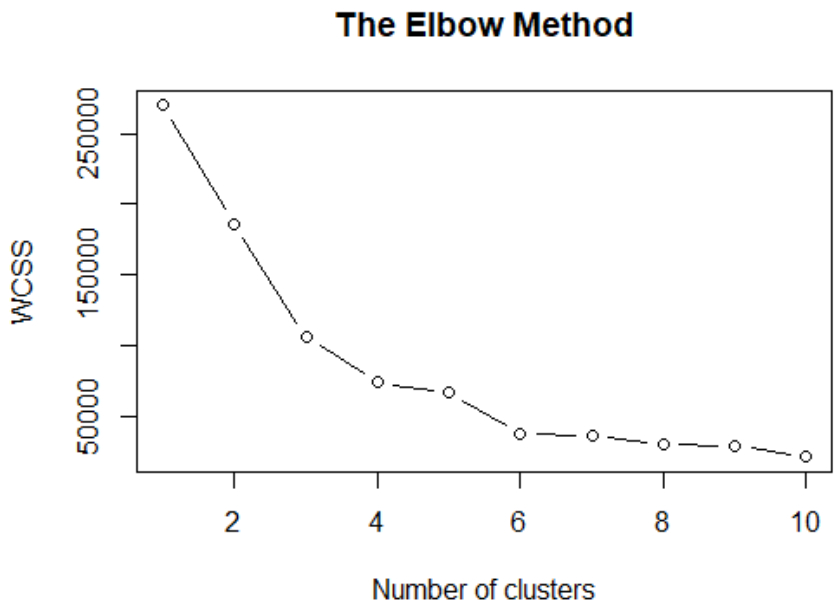
```
# Importing the dataset
dataset = read.csv('Mall_Customers.csv')
dataset = dataset[4:5]
```

En esta parte utilizando la semilla o seed vamos a encontrar todos los clusters que necesitaremos y crearemos también la variable wcss a la que le asignaremos el valor de vector, para ello utilizaremos un ‘ciclo for’, el cual nos ayudará a encontrar los cluster.

Le indicaremos desde donde va a empezar y finalizar así como asignando una variable al ciclo for, bien cada vez que el ciclo for avance guardaremos la posición en el vector que se guardó en la variable wcss y asuves le agregaremos la sumatoria de los kmeans.

```
# Using the elbow method to find the optimal number of clusters
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')
```

Resultado.



En esta gráfica podemos observar que los en cada cluster van mejorando al punto de que en el cluster 6 podemos observar ya una estabilidad a la que llamamos ‘codo’.

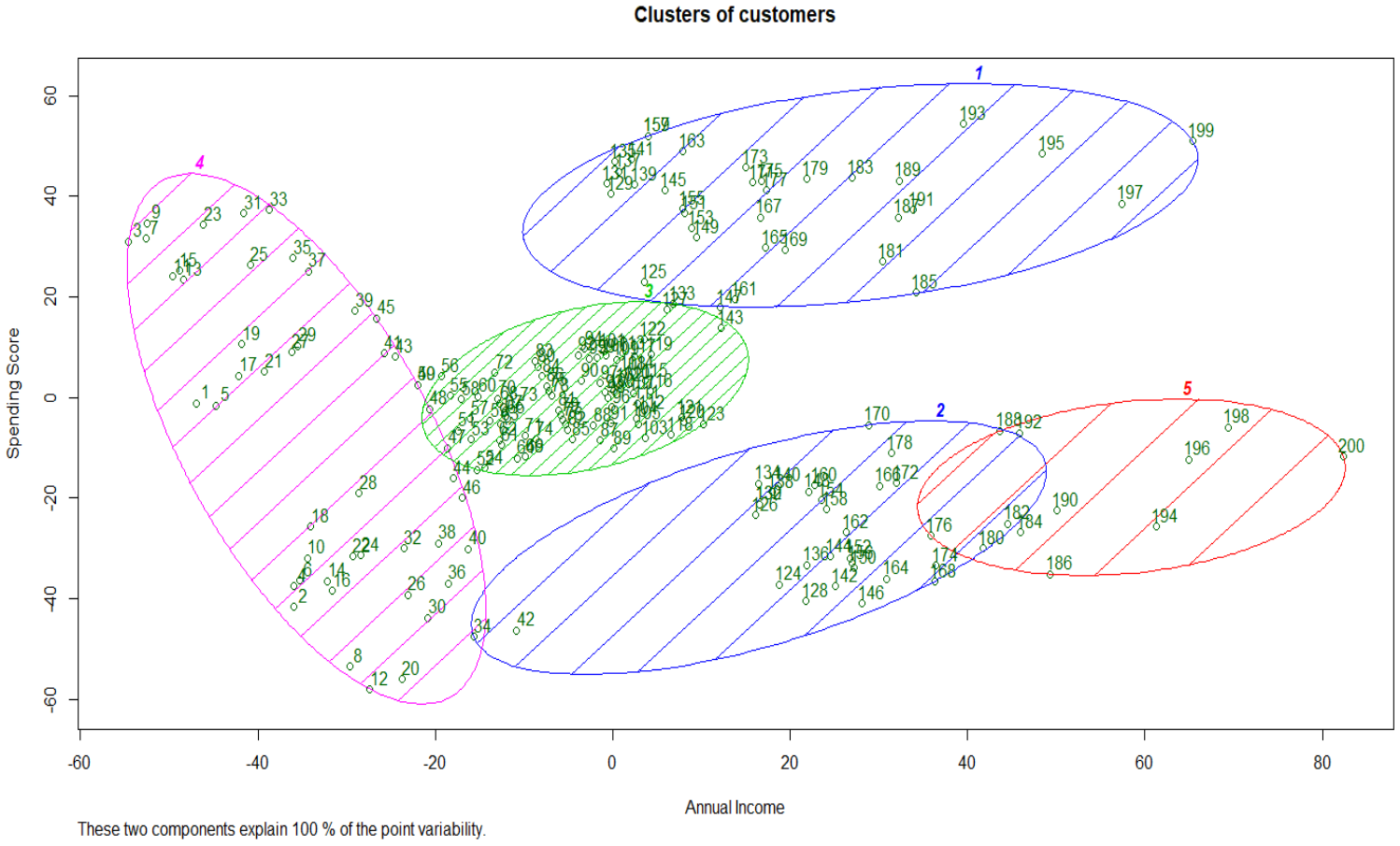
En la primera tabla podemos ver el número de cluster, pero ahora nos interesa saber a qué grupo de personas es conveniente hacer publicidad para ventas. Lo que haremos será volver a crear una semilla "seed', crearemos una variable llamada kmeans a la le asignaremos la función kmeans la cual nos servirá para hacer el centroide el cual nos ayudará a clasificar los puntos o datos, también nos ayudaremos de los cluster para hacer la gráfica.

```
# Fitting K-Means to the dataset
set.seed(29)
kmeans = kmeans(x = dataset, centers = 5)
y_kmeans = kmeans$cluster
```

Una vez hecho esto cargamos la librería cluster, la cual nos permite crear un plot (gráfica) de los cluster, asu vez indicaremos varios valores para la gráfica por nombrar algunos son la variable y_kmeans que es donde están los cluster, las líneas que en este caso queremos que sean 0, shade, color labels que en este caso sea de dos dimensiones el título de la grafica asi como el de los label "X y Y".

```
# Visualising the clusters
#install.packages('cluster')
library(cluster)
clusplot(dataset,
         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')
```

Resultado.



En esta gráfica podemos observar que el tercer grupo de empresas o personas son las idóneas para hacerles promoción de compras, ya que donde se concentra más los usuarios que tienden a ganar y gastar más dinero.