Yachay Tech University, November 16, 2021.

MSA: Final Project

Group 3

Anthony Salazar, Emilio Ibáñez

*School of Mathematical and Computational Sciences

Yachay Tech University - Urcuqui, Ecuador

1 Introduction

Data analysis is a very important branch of mathematics, it is specially useful when trying to draw conclusions from certain data. In this project we will discuss an example of this analysis using costumer data in order to come up with conclusions about their activities and behavior. A few statistical methods will be used to cluster the data and give the researchers information about these figures.

2 Materials and Methods

2.1 Data Description

The data used in this example corresponds to mall costumer's information. This information relates to their yearly income, and their score from 1-100. This score can tell us how likely a costumer is to spend part of their salary in the mall, 1 being the lowest and 100 being the highest.

2.2 Statistical Method

K-Means

K-means is a clustering or segmentation algorithm, which, like its name, has the objective of grouping/segmenting series of points. Initially you have a set of points with different characteristics/location and what the algorithm is looking for is to search the way that made us be able to join those points that are similar in different groups, which are known as "cluster". To achieve the objective of the algorithm it has to perform different steps. Step 1: Choose the number "K" of clusters; this will determine how many groups you want to separate the initial series of points. Step 2: Randomly select K points that will be the barycenter of each cluster (these do not necessarily have to be taken from the dataset). Step 3: Assign each of the points of the dataset to the nearest barycenter and thus form the K clusters. Step 4: With the clusters created in the previous step, a new barycenter has to be calculated and assigned to each cluster. To calculate the new barycenter we use each point of the cluster in which we are working to obtain one with more precision and that resembles more closely to the points that are close to it. (The original barycenter that was chosen randomly is recalculated to be associated with all the points in the different clusters). Step 5: Each point is reassigned to its nearest barycenter (as in step 3) and if there are any reassignments (i.e a point changes place) step 4 is repeated until any point is reassigned from one cluster to another and the model is finished.

Hierarchical Clustering
In order to understand hierarchical clustering it is necessary to understand k-means, because although they are different algorithms, they have the same objective and it is possible to start from one to get to the other. In addition, it is important to know that in spite of they are very similar and often the results obtained are exactly the same, the approaches in the process of both are totally different. For hierarchical clustering there exist two types of approaches, which are: the agglomerative which has a bottom-up approach (start from the elements and put them together in clusters) and the divisive which is top-down (start with the created clusters that will be divided into unitary elements). However, for this work we are going to use the agglomerative, so we will talk specifically about it, which has to follow a series of steps to reach its goal. Step 1: Make each point its own cluster, with this, we will have N clusters (then we would start with a number N of clusters, which is equal to the number of points in our dataset). Step 2: Choose the two points closest to each other and join them together to create a single cluster to obtain N-1 clusters. Step 3: With the clusters created in the previous step, choose the two closest clusters and join them together to create a new cluster and thus obtain N-2 clusters. Step 4: Repeat STEP 3 until a single cluster of the most similar elements is obtained and the model is finished. It is important to clarify the distance between two clusters. This can be seen from different points of view, which yields different options that are, the distance that exists between the two closest points of a cluster and another, the farthest points, the average distance or the distance between their barycenters. This will change the result but it depends on what approach you want to give it.

3 Results and Discussions

In this section, we will present our results based on the the properties of the cluster analysis in which we will use the next clustering models: K-means and the Hierarchical method.
Being our example a group of customers that are going to be in a mall we will have the variables of the Annual income and their Spending score which have been shown previously.
First we are going to apply the elbow method in order to determine the optimal number of clusters to be used.

So as we can see in the Figure 1, the function stabilizes at point five; thus it is our optimal number of clusters to use in our clustering process, k-means in this case.
As we are going to see below in the next graphs, in the Figure 2 and Figure 3, we can appreciate the clustering performed by the K-means. algorithm in which we can differentiate five groups and at the same time the number of each person, but we can show one in which it just shows the label of each group. So what we have here is the graphic representation of the k-means method in which we can see the cluster of the clients with similar profiles based on the incomes and their scores (1-100). Something we can see here is that the values are normalized in a score from -60 to 60.

So in the Figure 2 and Figure 3 we can see the five groups in which each one have been classified in the groups of the clients that have high incomes but low spending points meaning that they don't spend much, clients that have high incomes and high spending points, clients that have a regular amount of income and spending points which can be considered as regular clients, clients that have a low income and low spending points and finally clients with low income but have high spending points which means that they have a risky life since they spend a lot of money knowing they don't earn much.

And then, we did a hierarchical clustering to confirm the results obtained with partitioning, from which below we can see the dendogram from the 200 clients and thanks to this we can see how the individuals come together, their unions are pretty low, cause of their small distance and how the successive unions grow a larger distance.

Next we are going to adjust the the Hierarchical clustering to our dataset that we can see in Figure 5.

4 Code of the project

4.1 Clustering with K-means

```
# Clustering con K–means

# Importar los datos
dataset = read.csv("Mall_Customers.csv")
X = dataset[, 4:5]

# Aplicamos el Metodo del codo
set.seed(6)
wcss = vector()
for (i in 1:10){
  wcss[i] <– sum(kmeans(X, i)$withinss)
}
plot(1:10, wcss, type = 'b', main = "Metodo del codo",
     xlab = "Numero de clusters (k)", ylab = "WCSS(k)")

# Aplicar el algoritmo de k–means con k optimo
set.seed(29)
kmeans <– kmeans(X, 5, iter.max = 300, nstart = 10)

kmeans

#Visualizacionn de los clusters
install.packages("cluster")
library(cluster)
clusplot(X,
         kmeans$cluster,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 3,
         plotchar = FALSE,
         span = TRUE,
         main = "Clustering de clientes",
         xlab = "Ingresos anuales",
         ylab = "Puntuacion (1-100)"
         )
```

4.2 Hierarchical clustering

```
# Clustering Jerarquico

# Importar los datos del centro comercial
dataset = read.csv("Mall_Customers.csv")
X = dataset[, 4:5]

# Utilizar el dendrograma para encontrar el numero optimo de clusters
dendrogram = hclust(dist(X, method = "euclidean"),
                    method = "ward.D")
plot(dendrogram,
     main = "Dendrograma",
     xlab = "Clientes del centro comercial",
     ylab = "Distancia Euclidea")

# Ajustar el clustering jerarquico a nuestro dataset
hc = hclust(dist(X, method = "euclidean"),
                    method = "ward.D")
y_hc = cutree(hc, k=5)

# Visualizar los clusters
#install.packages("cluster")
library(cluster)
clusplot(X,
         y_hc,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 2,
         plotchar = FALSE,
         span = TRUE,
         main = "Clustering de clientes",
         xlab = "Ingresos anuales",
        ylab = "Puntuacion (1-100)"
)
```

5 Conclusion

So as we have seen the clustering analysis used in this has been very use full because it has helped us be able to have a visual interpretation of the data of the customers, as we saw in the Figure 5, we can see the five groups created by the K-means method, this let's us see in a case where a salesman wants to make a marketing campaign in order to see what customers are more optimal to try and sell their products, that way making a profit. In this way by dividing them in those clusters makes it more easier to see what exact customers will be their main objective.

6 References

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.
Hsu, C. C., Chen, C. L., Su, Y. W. (2007). Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences, 177(20), 4474-4492.

7 Anexes

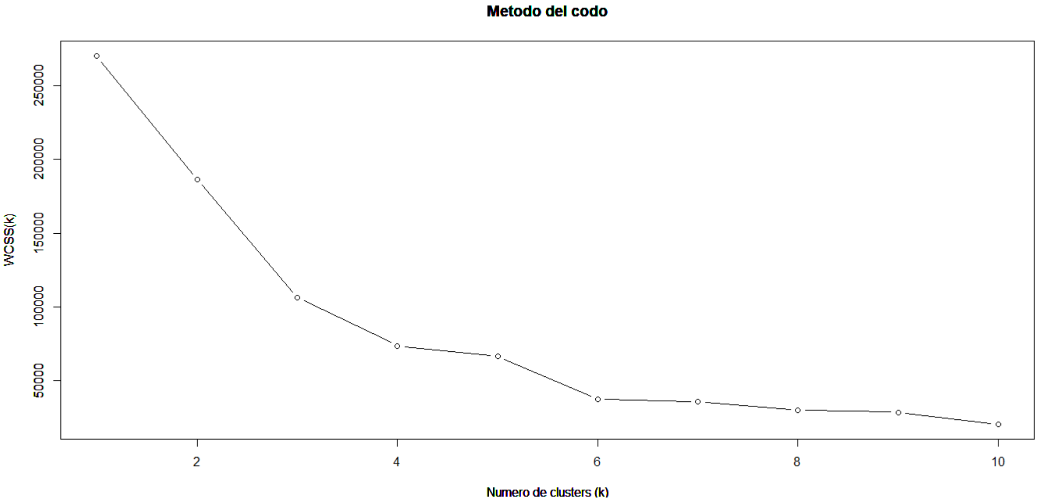

Figure 1: Elbow method for selection of an optimal number of clusters

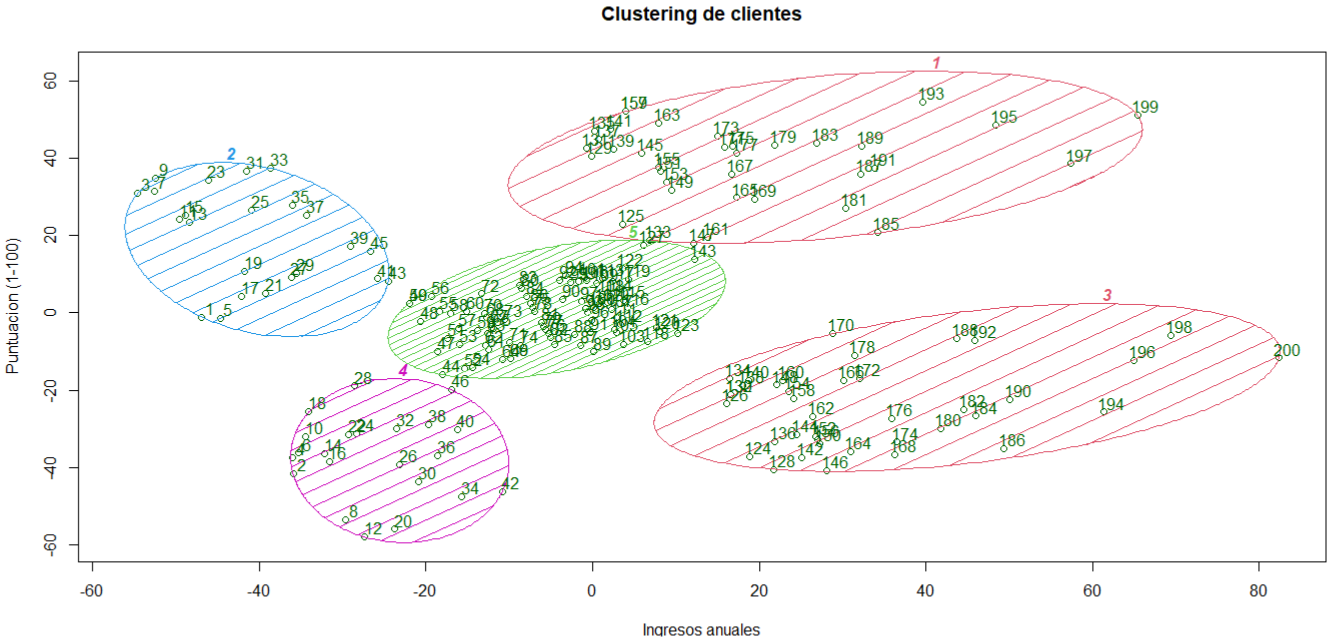

Figure 2: K-means with number of each client

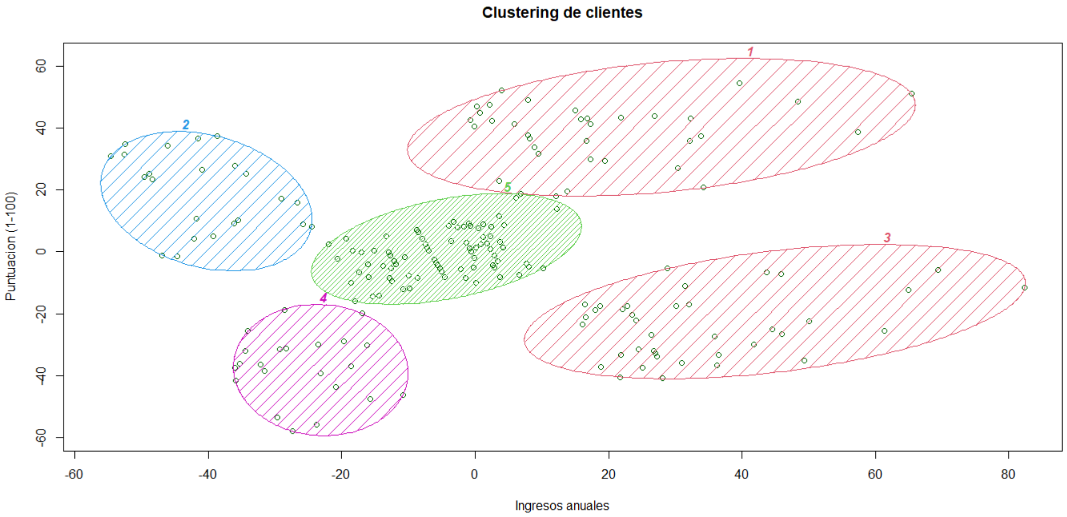

Figure 3: K-means

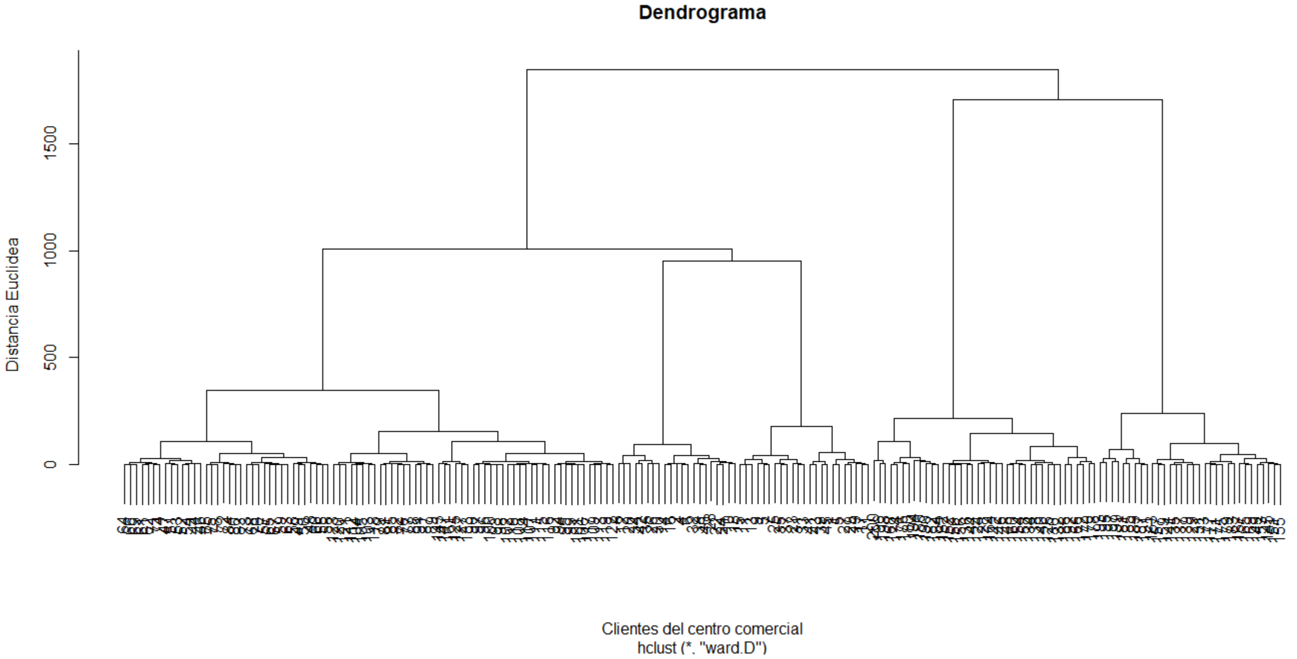

Figure 4: Cut Hierarchical clustering

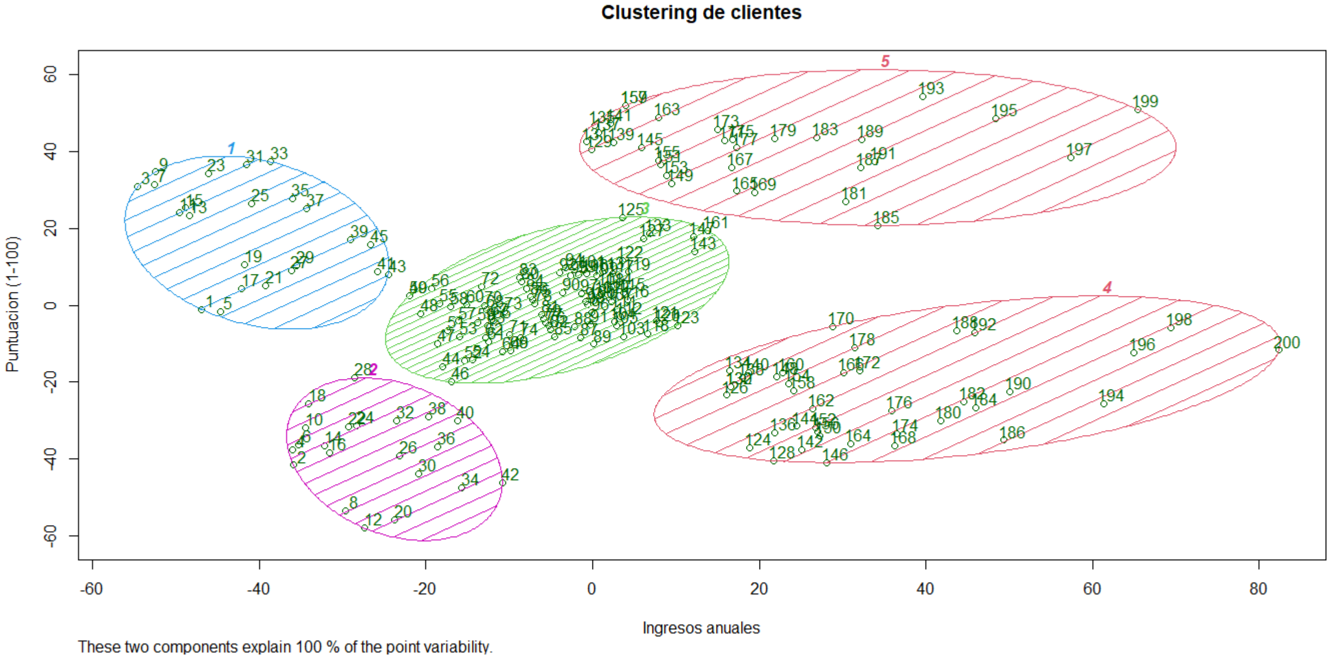

Figure 5: Adjusted K-means