# Patrones K-Means en R

12/5/2022


- A01245418 Andrés Sarellano Acevedo
- A00829837 Axel Amós Hernández Cárdenas
- A01281371 Izel María Ávila Rodríguez
- A01383422 Melissa Elvia Salazar Carrillo
- A00573134 Macías Romero Jorge Humberto


## 1. Cargando el dataframe

```
setwd('C:/users/hplax/Desktop/ITESM/CuartoSemestre')
```

## 2. Filtra y pre-procesa tus datos de acuerdo a lo visto en clase

**Matriz Escala**

```
df1_data_scale <- scale(df1[,-1])
head(df1_data_scale)
```

```
##        Alcohol  Malic.acid        Ash Alcalinity.of.ash  Magnesium
## [1,] 1.5143408 -0.56066822  0.2313998        -1.1663032 1.90852151
## [2,] 0.2455968 -0.49800856 -0.8256672        -2.4838405 0.01809398
## [3,] 0.1963252  0.02117152  1.1062139        -0.2679823 0.08810981
## [4,] 1.6867914 -0.34583508  0.4865539        -0.8069748 0.92829983
## [5,] 0.2948684  0.22705328  1.8352256         0.4506745 1.27837900
## [6,] 1.4773871 -0.51591132  0.3043010        -1.2860793 0.85828399
##      Total.phenols Flavanoids Nonflavanoid.phenols Proanthocyanin
## [1,]     0.8067217  1.0319081           -0.6577078      1.2214385
## [2,]     0.5670481  0.7315653           -0.8184106     -0.5431887
## [3,]     0.8067217  1.2121137           -0.4970050      2.1299594
## [4,]     2.4844372  1.4623994           -0.9791134      1.0292513
## [5,]     0.8067217  0.6614853            0.2261576      0.4002753
## [6,]     1.5576991  1.3622851           -0.1755994      0.6623487
##      Color.intensity        Hue OD280.OD315.of.diluted.wines     Proline
## [1,]       0.2510088  0.3611585                    1.8427215  1.01015939
## [2,]      -0.2924962  0.4049085                    1.1103172  0.96252635
## [3,]       0.2682629  0.3174085                    0.7863692  1.39122370
## [4,]       1.1827317 -0.4263410                    1.1807407  2.32800680
## [5,]      -0.3183774  0.3611585                    0.4483365 -0.03776747
## [6,]       0.7298108  0.4049085                    0.3356589  2.23274072
```

**Observando los rangos de cada variable de la matriz escala**

```
apply(df1_data_scale, 2, range)
```

```
##        Alcohol Malic.acid       Ash Alcalinity.of.ash Magnesium Total.phenols
## [1,] -2.427388  -1.428952 -3.668813         -2.663505 -2.082381     -2.101318
## [2,]  2.253415   3.100446  3.147447          3.145637  4.359076      2.532372
##      Flavanoids Nonflavanoid.phenols Proanthocyanin Color.intensity       Hue
## [1,]  -1.691200            -1.862979      -2.063214       -1.629691 -2.088840
## [2,]   3.054216             2.395645       3.475269        3.425768  3.292407
##      OD280.OD315.of.diluted.wines   Proline
## [1,]                    -1.889723 -1.488987
## [2,]                     1.955399  2.963114
```

**Obtención de la matriz de distancias con el método Euclidean**

```
df1_dist_mat = dist(df1_data_scale, method = "euclidean")
as.matrix(df2)[1:13,1:13]
```

```
##       Alcohol Malic.acid  Ash Alcalinity.of.ash Magnesium Total.phenols
##  [1,]   14.23       1.71 2.43              15.6       127          2.80
##  [2,]   13.20       1.78 2.14              11.2       100          2.65
##  [3,]   13.16       2.36 2.67              18.6       101          2.80
##  [4,]   14.37       1.95 2.50              16.8       113          3.85
##  [5,]   13.24       2.59 2.87              21.0       118          2.80
##  [6,]   14.20       1.76 2.45              15.2       112          3.27
##  [7,]   14.39       1.87 2.45              14.6        96          2.50
##  [8,]   14.06       2.15 2.61              17.6       121          2.60
##  [9,]   14.83       1.64 2.17              14.0        97          2.80
## [10,]   13.86       1.35 2.27              16.0        98          2.98
## [11,]   14.10       2.16 2.30              18.0       105          2.95
## [12,]   14.12       1.48 2.32              16.8        95          2.20
## [13,]   13.75       1.73 2.41              16.0        89          2.60
##       Flavanoids Nonflavanoid.phenols Proanthocyanin Color.intensity  Hue
##  [1,]       3.06                0.28           2.29            5.64 1.04
##  [2,]       2.76                0.26           1.28            4.38 1.05
##  [3,]       3.24                0.30           2.81            5.68 1.03
##  [4,]       3.49                0.24           2.18            7.80 0.86
##  [5,]       2.69                0.39           1.82            4.32 1.04
##  [6,]       3.39                0.34           1.97            6.75 1.05
##  [7,]       2.52                0.30           1.98            5.25 1.02
##  [8,]       2.51                0.31           1.25            5.05 1.06
##  [9,]       2.98                0.29           1.98            5.20 1.08
## [10,]       3.15                0.22           1.85            7.22 1.01
## [11,]       3.32                0.22           2.38            5.75 1.25
## [12,]       2.43                0.26           1.57            5.00 1.17
## [13,]       2.76                0.29           1.81            5.60 1.15
##       OD280.OD315.of.diluted.wines Proline
##  [1,]                         3.92    1065
##  [2,]                         3.40    1050
```

```
##  [3,]                        3.17    1185
##  [4,]                        3.45    1480
##  [5,]                        2.93     735
##  [6,]                        2.85    1450
##  [7,]                        3.58    1290
##  [8,]                        3.58    1295
##  [9,]                        2.85    1045
## [10,]                        3.55    1045
## [11,]                        3.17    1510
## [12,]                        2.82    1280
## [13,]                        2.90    1320
```

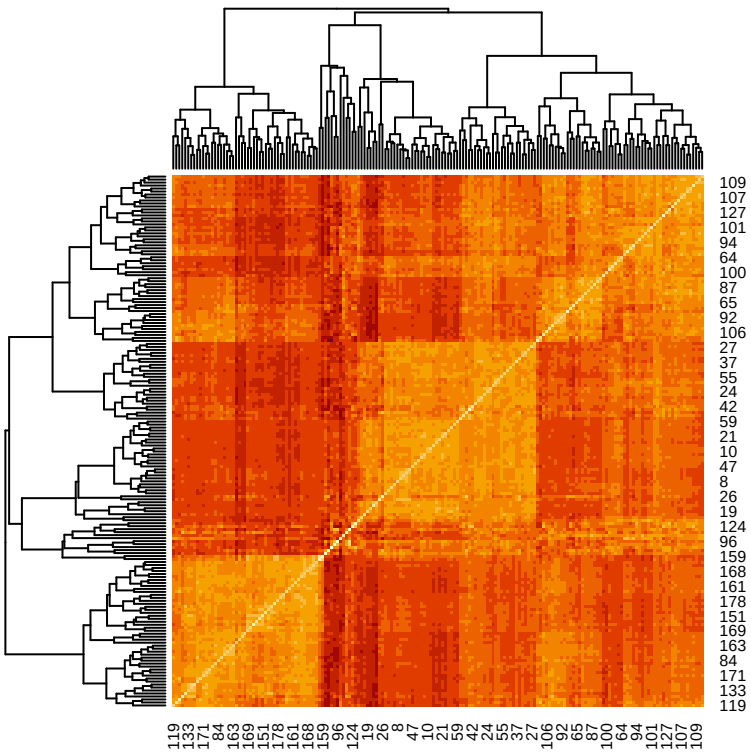
### Dimensiones de la matriz de distancias

```
dim(as.matrix(df1_dist_mat))
```

```
## [1] 178 178
```

### Creando el Heatmap de la matriz de distancias

```
heatmap(as.matrix(df1_dist_mat))
```



**Cambiando el nombre de las variables**

```
rownames(df1_data_scale) = paste(df1$Cultivar, seq(1,100), sep="_Cultivo")
head(df1_data_scale)
```

```
##              Alcohol  Malic.acid        Ash Alcalinity.of.ash  Magnesium
## 1_Cultivo1 1.5143408 -0.56066822  0.2313998        -1.1663032 1.90852151
## 1_Cultivo2 0.2455968 -0.49800856 -0.8256672        -2.4838405 0.01809398
## 1_Cultivo3 0.1963252  0.02117152  1.1062139        -0.2679823 0.08810981
## 1_Cultivo4 1.6867914 -0.34583508  0.4865539        -0.8069748 0.92829983
## 1_Cultivo5 0.2948684  0.22705328  1.8352256         0.4506745 1.27837900
## 1_Cultivo6 1.4773871 -0.51591132  0.3043010        -1.2860793 0.85828399
##            Total.phenols Flavanoids Nonflavanoid.phenols Proanthocyanin
## 1_Cultivo1     0.8067217  1.0319081           -0.6577078      1.2214385
## 1_Cultivo2     0.5670481  0.7315653           -0.8184106     -0.5431887
## 1_Cultivo3     0.8067217  1.2121137           -0.4970050      2.1299594
## 1_Cultivo4     2.4844372  1.4623994           -0.9791134      1.0292513
## 1_Cultivo5     0.8067217  0.6614853            0.2261576      0.4002753
## 1_Cultivo6     1.5576991  1.3622851           -0.1755994      0.6623487
##            Color.intensity        Hue OD280.OD315.of.diluted.wines     Proline
## 1_Cultivo1       0.2510088  0.3611585                    1.8427215  1.01015939
## 1_Cultivo2      -0.2924962  0.4049085                    1.1103172  0.96252635
## 1_Cultivo3       0.2682629  0.3174085                    0.7863692  1.39122370
## 1_Cultivo4       1.1827317 -0.4263410                    1.1807407  2.32800680
## 1_Cultivo5      -0.3183774  0.3611585                    0.4483365 -0.03776747
## 1_Cultivo6       0.7298108  0.4049085                    0.3356589  2.23274072
```

**Mostrando en forma de matriz el mapa de calor con el metodo de euclidian**

```
df1_dist_mat = dist(df1_data_scale, method="euclidian")
as.matrix(df1_dist_mat)[1:5,1:5]
```

```
##            1_Cultivo1 1_Cultivo2 1_Cultivo3 1_Cultivo4 1_Cultivo5
## 1_Cultivo1   0.000000   3.487697   3.018094   2.834509   3.556821
## 1_Cultivo2   3.487697   0.000000   4.131258   4.348349   4.614454
## 1_Cultivo3   3.018094   4.131258   0.000000   3.237354   2.972721
## 1_Cultivo4   2.834509   4.348349   3.237354   0.000000   4.483310
## 1_Cultivo5   3.556821   4.614454   2.972721   4.483310   0.000000
```

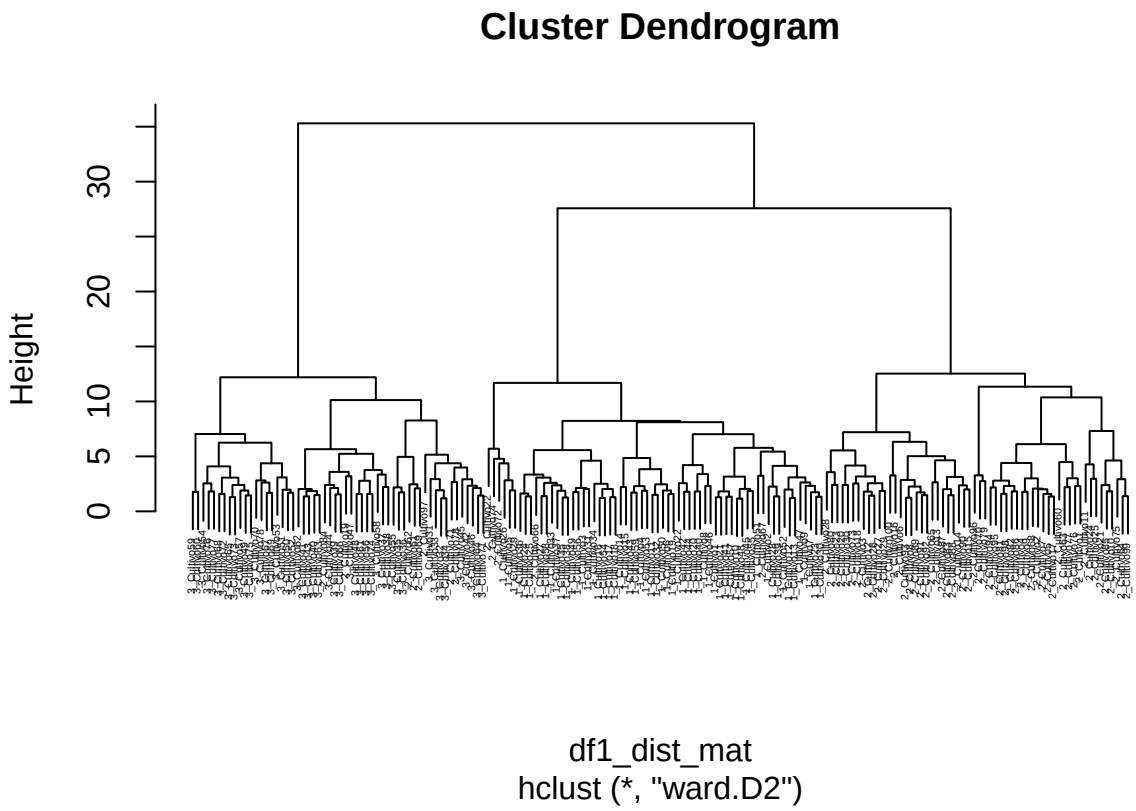
**Creación de los hclust**

```
hclust(df1_dist_mat, method = "ward.D2")
```

```
##
## Call:
## hclust(d = df1_dist_mat, method = "ward.D2")
##
## Cluster method   : ward.D2
## Distance         : euclidean
## Number of objects: 178
```

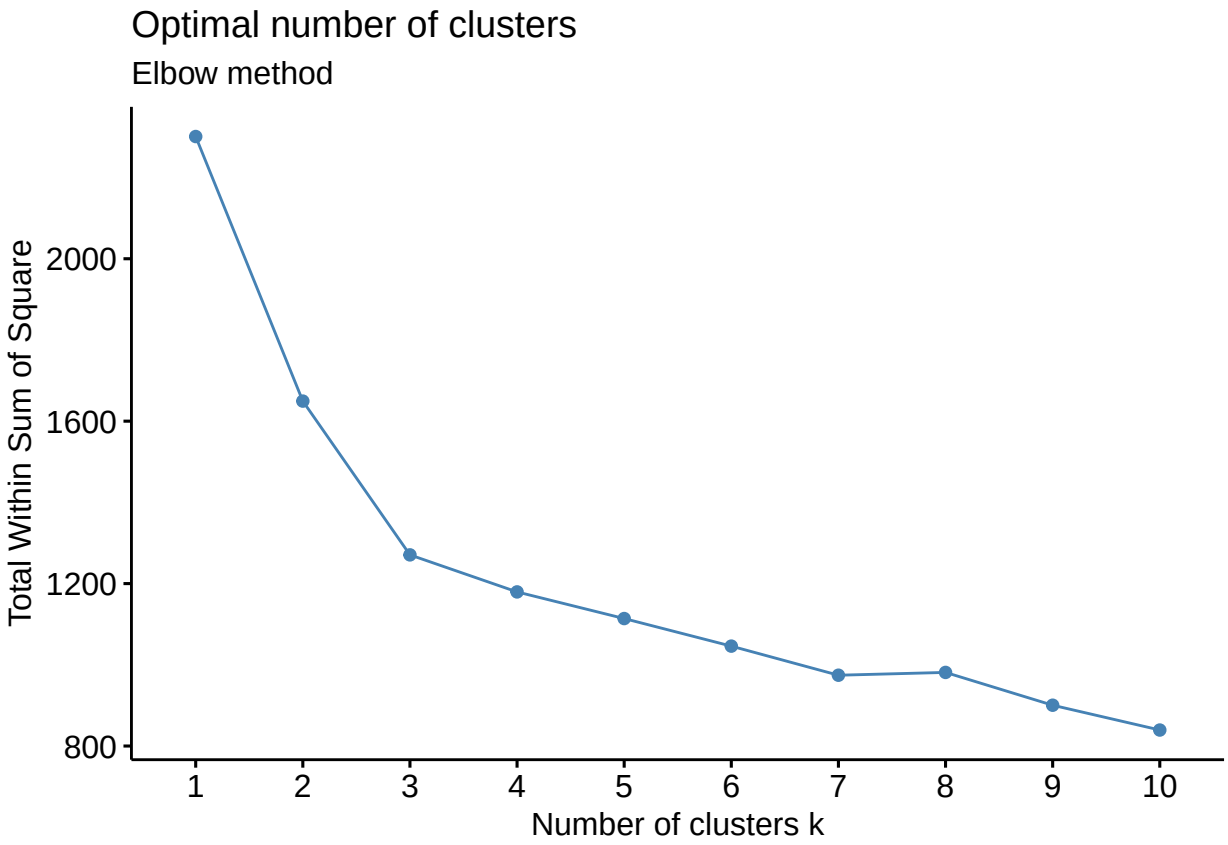
**Grafica de Dendograma**

```
plot(hclust(df1_dist_mat, method = "ward.D2"), cex=0.3)
```




**Gráfica del Metodo de Elbow**

```
fviz_nbclust(df1_data_scale, kmeans, method="wss")+labs(subtitle = "Elbow method")
```




Se puede obtener que para el kmeans, center = 3 y nstart = 1300 aproximadamente.

**Obtencion del K-means**

```
km.out = kmeans(df1_data_scale, center=3, nstart=1300)
km.out
```

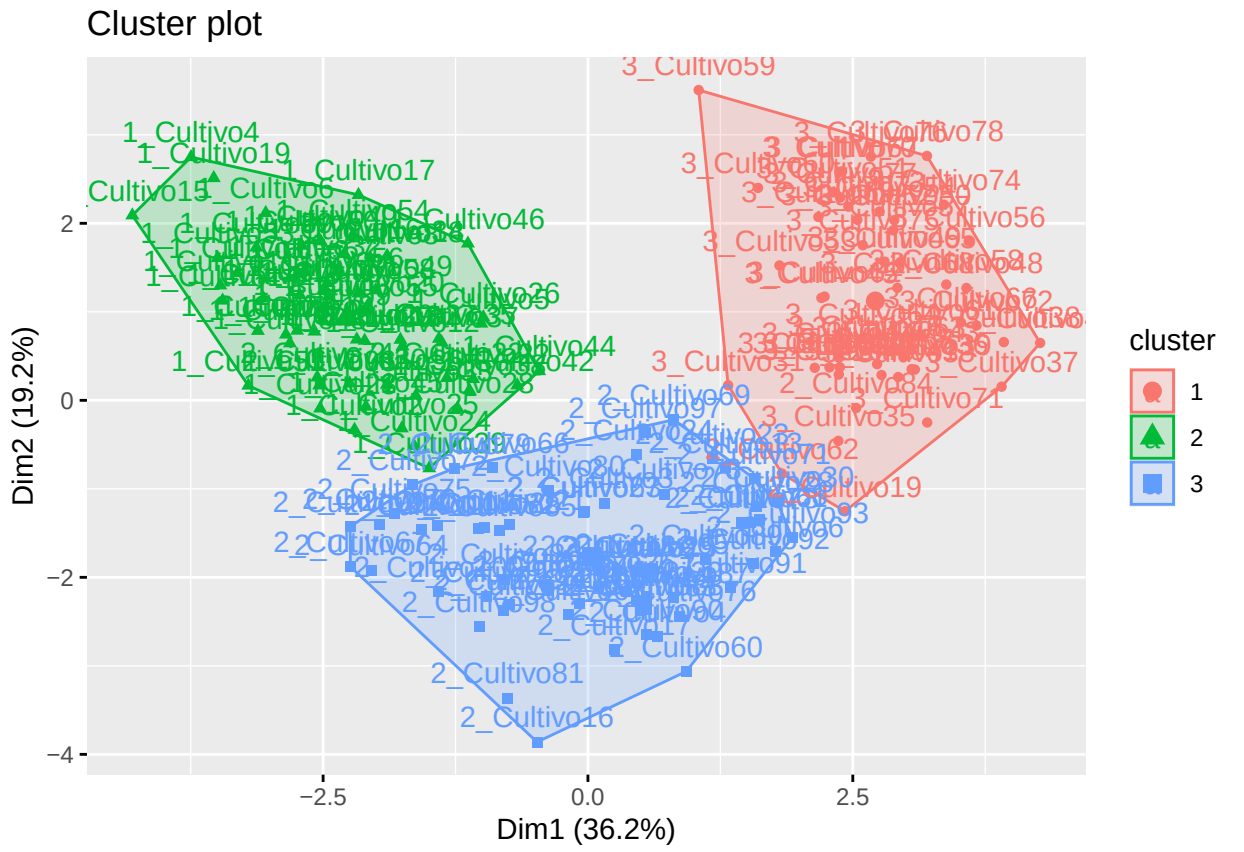
```
## K-means clustering with 3 clusters of sizes 51, 62, 65
##
## Cluster means:
##      Alcohol Malic.acid        Ash Alcalinity.of.ash   Magnesium Total.phenols
## 1  0.1644436  0.8690954  0.1863726         0.5228924 -0.07526047   -0.97657548
## 2  0.8328826 -0.3029551  0.3636801        -0.6084749  0.57596208    0.88274724
## 3 -0.9234669 -0.3929331 -0.4931257         0.1701220 -0.49032869   -0.07576891
##    Flavanoids Nonflavanoid.phenols Proanthocyanin Color.intensity        Hue
## 1 -1.21182921           0.72402116    -0.77751312       0.9388902 -1.1615122
## 2  0.97506900          -0.56050853     0.57865427       0.1705823  0.4726504
## 3  0.02075402          -0.03343924     0.05810161      -0.8993770  0.4605046
##   OD280.OD315.of.diluted.wines    Proline
## 1                   -1.2887761 -0.4059428
## 2                    0.7770551  1.1220202
## 3                    0.2700025 -0.7517257
##
## Clustering vector:
```

```
##   1_Cultivo1   1_Cultivo2   1_Cultivo3   1_Cultivo4   1_Cultivo5   1_Cultivo6
##            2            2            2            2            2            2
##   1_Cultivo7   1_Cultivo8   1_Cultivo9  1_Cultivo10  1_Cultivo11  1_Cultivo12
##            2            2            2            2            2            2
##  1_Cultivo13  1_Cultivo14  1_Cultivo15  1_Cultivo16  1_Cultivo17  1_Cultivo18
##            2            2            2            2            2            2
##  1_Cultivo19  1_Cultivo20  1_Cultivo21  1_Cultivo22  1_Cultivo23  1_Cultivo24
##            2            2            2            2            2            2
##  1_Cultivo25  1_Cultivo26  1_Cultivo27  1_Cultivo28  1_Cultivo29  1_Cultivo30
##            2            2            2            2            2            2
##  1_Cultivo31  1_Cultivo32  1_Cultivo33  1_Cultivo34  1_Cultivo35  1_Cultivo36
##            2            2            2            2            2            2
##  1_Cultivo37  1_Cultivo38  1_Cultivo39  1_Cultivo40  1_Cultivo41  1_Cultivo42
##            2            2            2            2            2            2
##  1_Cultivo43  1_Cultivo44  1_Cultivo45  1_Cultivo46  1_Cultivo47  1_Cultivo48
##            2            2            2            2            2            2
##  1_Cultivo49  1_Cultivo50  1_Cultivo51  1_Cultivo52  1_Cultivo53  1_Cultivo54
##            2            2            2            2            2            2
##  1_Cultivo55  1_Cultivo56  1_Cultivo57  1_Cultivo58  1_Cultivo59  2_Cultivo60
##            2            2            2            2            2            3
##  2_Cultivo61  2_Cultivo62  2_Cultivo63  2_Cultivo64  2_Cultivo65  2_Cultivo66
##            3            1            3            3            3            3
##  2_Cultivo67  2_Cultivo68  2_Cultivo69  2_Cultivo70  2_Cultivo71  2_Cultivo72
##            3            3            3            3            3            3
##  2_Cultivo73  2_Cultivo74  2_Cultivo75  2_Cultivo76  2_Cultivo77  2_Cultivo78
##            3            2            3            3            3            3
##  2_Cultivo79  2_Cultivo80  2_Cultivo81  2_Cultivo82  2_Cultivo83  2_Cultivo84
##            3            3            3            3            3            1
##  2_Cultivo85  2_Cultivo86  2_Cultivo87  2_Cultivo88  2_Cultivo89  2_Cultivo90
##            3            3            3            3            3            3
##  2_Cultivo91  2_Cultivo92  2_Cultivo93  2_Cultivo94  2_Cultivo95  2_Cultivo96
##            3            3            3            3            3            2
##  2_Cultivo97  2_Cultivo98  2_Cultivo99 2_Cultivo100   2_Cultivo1   2_Cultivo2
##            3            3            3            3            3            3
##   2_Cultivo3   2_Cultivo4   2_Cultivo5   2_Cultivo6   2_Cultivo7   2_Cultivo8
##            3            3            3            3            3            3
##   2_Cultivo9  2_Cultivo10  2_Cultivo11  2_Cultivo12  2_Cultivo13  2_Cultivo14
##            3            3            3            3            3            3
##  2_Cultivo15  2_Cultivo16  2_Cultivo17  2_Cultivo18  2_Cultivo19  2_Cultivo20
##            3            3            3            3            1            3
##  2_Cultivo21  2_Cultivo22  2_Cultivo23  2_Cultivo24  2_Cultivo25  2_Cultivo26
##            3            2            3            3            3            3
##  2_Cultivo27  2_Cultivo28  2_Cultivo29  2_Cultivo30  3_Cultivo31  3_Cultivo32
##            3            3            3            3            1            1
##  3_Cultivo33  3_Cultivo34  3_Cultivo35  3_Cultivo36  3_Cultivo37  3_Cultivo38
##            1            1            1            1            1            1
##  3_Cultivo39  3_Cultivo40  3_Cultivo41  3_Cultivo42  3_Cultivo43  3_Cultivo44
##            1            1            1            1            1            1
##  3_Cultivo45  3_Cultivo46  3_Cultivo47  3_Cultivo48  3_Cultivo49  3_Cultivo50
##            1            1            1            1            1            1
##  3_Cultivo51  3_Cultivo52  3_Cultivo53  3_Cultivo54  3_Cultivo55  3_Cultivo56
##            1            1            1            1            1            1
##  3_Cultivo57  3_Cultivo58  3_Cultivo59  3_Cultivo60  3_Cultivo61  3_Cultivo62
##            1            1            1            1            1            1
```

```
##  3_Cultivo63  3_Cultivo64  3_Cultivo65  3_Cultivo66  3_Cultivo67  3_Cultivo68
##            1            1            1            1            1            1
##  3_Cultivo69  3_Cultivo70  3_Cultivo71  3_Cultivo72  3_Cultivo73  3_Cultivo74
##            1            1            1            1            1            1
##  3_Cultivo75  3_Cultivo76  3_Cultivo77  3_Cultivo78
##            1            1            1            1
##
## Within cluster sum of squares by cluster:
## [1] 326.3537 385.6983 558.6971
##  (between_SS / total_SS =  44.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

**Gráfica de clusters a partir del K-means**

```
km.clusters = km.out$cluster
fviz_cluster(list(data=df1_data_scale, cluster=km.clusters))
```




**Preguntas**

**1. ¿Por qué es necesario escalar los datos?**

Es imprescindible escalar los datos ya que cuando se realiza esta acción se normalizan (comprimen o extender valores de la variable para que esté en un rango definido) los datos de forma que no se dé prioridad a una característica concreta. El escalado es muy importante en los algoritmos basados en la distancia.

**2. ¿Qué información nos da la gráfica "Elbow plot"?**

Esta gráfica nos muestra la cantidad de clusters que serán generados con los datos de nuestra base de datos, por lo que para poder generar nuestro valor de kmeans, gracias al punto de inflexión que nuestra función en la gráfica de codo proporciona, podemos concluir que nuestro valor k que será la entrada de la función de kmeans para generar las agrupaciones, será de 3 clusters, y la cantidad de agrupaciones aleatorias será alrededor de 1300 agrupaciones.

**3. Basándote en las matrices de similitud obtenidas, cuáles pares de datos son los más parecidos? ¿Cuáles son los más diferentes? ¿Salen los mismos pares (más parecidos y más diferentes) utilizando otra métrica de distancia?**

Omitiendo los datos con valor de 0.0, el dato más cercano (parecido) a cero cuenta con un valor de 1.16084, mientras que el más alejado (diferente) tiene un valor de 2.212 (maximo valor permitido por el print de la consola). Tomando en cuenta que el heatmap tambien puede representar una matriz de similitud, se puede observar que la mayoría de los datos muestran una diferencia entre si.

Por otra parte, es posible que realizar otro método de distancia, como lo es el método de Manhattan, arroje diferentes valores de distancia entre los puntos, haciendolos más parecidos o más diferentes comparandolo con el metodo Euclidian.

**4. ¿Hay datos atípicos?**

En la gráfica de cluster no se pueden observar datos atípicos. Sin embargo, si se encuentran datos muy alejado de las agrupaciones centrales de cada clusters, sin embargo, estas aun delimitan al cluster por lo que no se considerarán atipicos.

**5. ¿Crees que tus resultados serían diferentes si eliminamos variables?**

Dependiendo de qué variables sean si pueden cambiar hasta un cierto punto el resultado, ya que independientemente de cuál variable se remueva va afectar en clustering porque se van a agrupar menos datos y hay que tener cuidado en que tantas variables también removemos porque si tenemos una muestra con pocas variables no hay mucho que se va a poder analizar. Pero intenté remover dos variables y no cambió tanto en total sum of clusters solo en la suma y el mapa de calor con las matrices si cambió mucho.

**6. ¿Qué es el clustering jerárquico?**

El Clustering Jerárquico va agrupando de poco en poco dependiendo de la distancia entre cada uno de los datos y buscando que los clústers sean los más similares entre sí. La representación es a través de un dendrograma. Lo malo de este tipo de mapa es que no son buenos para grandes cantidades de datos y lo más importante es la medición de la distancia entre datos. (Duk2, 2019)

## Referencia

Duk2. (2019). Algoritmos de data mining para agrupar datos – Clustering Jerárquico. ESTRATEGIAS DE TRADING. https://estrategiastrading.com/clustering-jerarquico/