**Ejercicio Propuesto 1 (Resuelto)**

**Utilizamos de nuevo el archivo de datos ventas\_vehiculos.sav que contiene estimaciones de ventas, listas de precios y especificaciones físicas hipotéticas de varias marcas y modelos de vehículos. Se desea hacer un estudio de mercado para poder determinar las posibles competencias para sus vehículos, para ello agrupamos las marcas de los coches según los datos disponibles, hábitos de consumo, sexo, edad, nivel de ingresos, etc. de los clientes. Las empresas de coches adaptan sus estrategias de desarrollo de productos y de marketing en función de cada grupo de consumidores para aumentar las ventas y el nivel de fidelidad a la marca.**

**Se pide:**

1. **Utilizar un análisis clúster jerárquico**
2. Método de conglomerado, el **Vecino más lejano**

**Nota: El archivo de datos ventas\_vehículos .sav contiene 157 datos y está formado por las siguientes variables:**

**Variables tipo cadena: marca (Fabricante); modelo**

**Variables tipo numérico: ventas (en miles); reventa (Valor de reventa en 4 años); tipo (Tipo de vehículo: Valores: {0, Automóvil; 1, Camión}); precio (en miles); motor (Tamaño del motor); CV (Caballos); pisada (Base de neumáticos); ancho (Anchura); largo (Longitud); peso\_neto (Peso neto); depósito (Capacidad de combustible); mpg (Consumo).**

**Solución:**

Antes de comenzar con la resolución del ejercicio vamos a introducir los datos

> setwd("C:/Users/Usuario/Desktop/Datos")   # situarnos en el directorio de trabajo

> library(foreign) # importar y exportar datos .sav

> datos<- read.spss("Ventas-vehiculos.sav",use.value.labels=TRUE, max.value.labels=TRUE, to.data.frame=TRUE, sep)

re-encoding from CP1252

En primer lugar restringiremos el archivo de datos sólo a los automóviles de los que se vendieron al menos 100.000 unidades. Para ello seleccionamos los casos que cumplan esa condición mediante la siguiente fórmula:

> datos100k <- subset(datos, subset =datos$ventas>100 & datos$tipo==0)

La función subset filtramos los datos, poniendo las condiciones de selección en el segundo argumento.

 marca modelo ventas reventa tipo precio motor CV

20 Chevrolet Cavalier 145.519 9.250 0 13.260 2.2 115

21 Chevrolet Malibu 135.126 11.225 0 16.535 3.1 170

28 Chevrolet Impala 107.995 NA 0 18.890 3.4 180

48 Ford Mustang 113.369 12.760 0 21.560 3.8 190

50 Ford Taurus 245.815 10.055 0 17.885 3.0 155

51 Ford Focus 175.670 NA 0 12.315 2.0 107

58 Honda Civic 199.685 9.850 0 12.885 1.6 106

59 Honda Accord 230.902 13.210 0 15.350 2.3 135

120 Pontiac Grand Am 131.097 10.290 0 19.720 3.4 175

137 Toyota Corolla 142.535 10.025 0 13.108 1.8 120

138 Toyota Camry 247.994 13.245 0 17.518 2.2 133

 pisada ancho largo peso\_neto deposito mpg

20 104.1 67.9 180.9 2.676 14.3 27

21 107.0 69.4 190.4 3.051 15.0 25

28 110.5 73.0 200.0 3.389 17.0 27

48 101.3 73.1 183.2 3.203 15.7 24

50 108.5 73.0 197.6 3.368 16.0 24

51 103.0 66.9 174.8 2.564 13.2 30

58 103.2 67.1 175.1 2.339 11.9 32

59 106.9 70.3 188.8 2.932 17.1 27

120 107.0 70.4 186.3 3.091 15.2 25

137 97.0 66.7 174.0 2.420 13.2 33

138 105.2 70.1 188.5 2.998 18.5 27

**a) Utilizar un análisis clúster jerárquico**

Se observan que hay variables con la marca y modelo de cada coche, estas variables deben ser eliminadas para los cálculos.

Las variables que están en medidas distintas deberán tipificarse ya que podrían producir error.

> vehiculos<-Ventasvehiculos[,-c(1,2)]

> vehiculos<-scale(vehiculos,center = TRUE,scale = TRUE)

> rownames(vehiculos)<- Ventasvehiculos$modelo

> dis<-dist(vehiculos)^2

> x<-hclust(dis)

> x$merge

 [,1] [,2]

 [1,] -34 -107

 [2,] -154 -155

 [3,] -152 -153

 [4,] -29 -37

 [5,] -52 -90

 [6,] -146 -150

 [7,] -92 -105

 [8,] -114 -124

 [9,] -24 -121

 [10,] -32 -116

 [11,] -2 -3

 [12,] -59 -138

 [13,] -21 -120

 [14,] -8 -93

 [15,] -66 11

 [16,] -97 -98

 [17,] -36 -115

 [18,] -47 -102

 [19,] -104 -139

 [20,] -12 -110

 [21,] -82 -103

 [22,] -5 3

 [23,] -151 6

 [24,] -14 -122

 [25,] -16 -17

 [26,] -72 -157

 [27,] -131 -132

 [28,] -87 -88

 …………………..

[154,] 149 153

[155,] 141 154

[156,] 152 155

> plot(x)



**b) Método de conglomerado, el Vecino más lejano**

> datos100k <- subset(datos, subset =datos$ventas>100 & datos$tipo==0)

Al leer los datos vemos que se han colado una serie de espacios en las variables “**marca**” y “**modelo**”.

Esto lo corregimos con la función gsub, que funciona sustituyendo lo que pongas en el primer argumento por lo que pones en el segundo. En este caso un espacio por nada.

> datos100k$marca= gsub(" ","",datos100k$marca)

> datos100k$modelo= gsub(" ","",datos100k$modelo)

> x <- datos100k[,c("precio","motor","CV","pisada","ancho","largo","peso\_neto","deposito","mpg" )]

Como las variables se miden en unidades diferentes vamos a normalizar la variable con Media 0 y Varianza 1. Para ello usamos la función scale. Elegimos True tanto para centrar como para escalar. También renombramos a las filas como el nombre de sus modelos.

> X.scale<- scale(x, center = T, scale = T)

> row.names(X.scale)<- datos100k$modelo

> dmx <- dist(X.scale[, colnames(x)], method = "euclidean")

> dmx

 Cavalier Malibu Impala Mustang Taurus Focus Civic Accord

Malibu 3.072816

Impala 5.274025 2.708019

Mustang 5.110737 3.084236 3.526982

Taurus 4.418552 2.122562 1.618253 3.224947

Focus 1.522351 4.323238 6.447844 6.136681 5.743645

Civic 2.518551 5.107705 7.099278 6.849459 6.472819 1.256555

Accord 2.391317 2.102429 3.297454 4.134429 2.650065 3.652378 4.447080

GrandAm 3.812895 1.274623 2.532611 2.185040 2.190572 4.956890 5.681518 2.727180

Corolla 3.063649 5.408865 7.389655 6.598459 6.831109 2.002918 1.915638 4.783256

Camry 3.030349 2.642446 3.450847 3.988278 2.915417 4.201357 4.992374 1.122524

 GrandAm Corolla

Malibu

Impala

Mustang

Taurus

Focus

Civic

Accord

GrandAm

Corolla 5.896650

Camry 2.927103 5.048393

En este caso seleccionamos el método **“complete”** en la función**hclust**.

> cc <- hclust(dmx^2, method = "complete")

> cc

Call:

hclust(d = dmx^2, method = "complete")

Cluster method : complete

Distance : euclidean

Number of objects: 11

> plot(cc)



> abline(h=27, col="red")



> cc$height
[1] 1.260060 1.578930 1.624664 2.618741 4.011680 7.333368 9.183015
[8] 12.439603 25.486274 54.606994
> cc$merge
[,1] [,2]
[1,] -8 -11
[2,] -6 -7
[3,] -2 -9
[4,] -3 -5
[5,] -10 2
[6,] 3 4
[7,] -1 1
[8,] -4 6
[9,] 5 7
[10,] 8 9

En las primeras etapas, el Historial de conglomeración para la solución de vinculación completa (vecino más lejano) es similar a la solución de vinculación única (vecino más próximo). En cambio, en las etapas finales los historiales de conglomeración son muy diferentes. Mediante el método de conglomeración del vecino más lejano se realiza una clasificación fuerte de dos o tres grupos.

La primera gran diferencia es entre las etapas 5 y 6 (6 clusters), la segunda entre 8 y 9 (3 clusters) y entre 9 y 10 (2 clusters).

La decisión de esta clasificación se refleja en el dendrograma.

* La división inicial del árbol forma dos grupos, (8, 11, 1, 6, 7, 10) y (2, 9, 3, 5, 4). El clúster primero contiene los automóviles más pequeños y el clúster segundo contiene los coches más grandes.
* El grupo de coches más pequeños se puede dividir en dos subgrupos, uno de ellos formado por los coches más pequeños y más baratos. Así la división siguiente en 3 clusters: (Accord (8), Camry (11), Cavalier (1)), (Focus (6), Civic (7), Corolla (10)), estos tres coches son más pequeños y más baratos que los tres anteriores) y (Malibú (2), Gran Am (9), Impala (3), Taurus (5), Mustang (4)).

**Resumen**

La solución de la vinculación completa (vecino más lejano) es satisfactoria debido a que sus grupos son diferentes, mientras que la solución del vecino más cercano es menos concluyente. Usando como Método de conglomeración la vinculación completa (Vecino más lejano), se puede determinar la competencia que hay entre los vehículos en la fase de diseño mediante la introducción de sus especificaciones como nuevos casos en el conjunto de datos y volver a ejecutar el análisis.