**Análisis de Clustering Jerárquico de mtcars**

Carlos Enrique Olaya Hernandez

Iniciamos haciendo la instalación de paquetes y cargues de estos.

#Instalacion de paquetes si es necesario

#install.packages("ggplot2")

#install.packages("cluster")

#install.packages("NbClust")

#install.packages("dendextend")

#install.packages("factoextra")

# librerias

library(cluster)

library(ggplot2)

library(NbClust)

library(dendextend)

library(factoextra)

Luego cargamos el conjunto de datos de mtcars y escalamos las variables

data(mtcars)

usamos la función str para observar el tipado de variables y sus valores

str(mtcars)

'data.frame': 32 obs. of 11 variables:

$ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...

$ cyl : num 6 6 4 6 8 6 8 4 4 6 ...

$ disp: num 160 160 108 258 360 ...

$ hp : num 110 110 93 110 175 105 245 62 95 123 ...

$ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...

$ wt : num 2.62 2.88 2.32 3.21 3.44 ...

$ qsec: num 16.5 17 18.6 19.4 17 ...

$ vs : num 0 0 1 1 0 1 0 1 1 1 ...

$ am : num 1 1 1 0 0 0 0 0 0 0 ...

$ gear: num 4 4 4 3 3 3 3 4 4 4 ...

$ carb: num 4 4 1 1 2 1 4 2 2 4 ...

logramos observar que los algunos datos tienen valores muy grandes a comparación de otras variables y procedemos hacer una escalada

datos\_escalados <- scale(mtcars)

con los datos bien escalados procedemos a calcular la matriz de distancias con los datos escalados

distancias <- dist(datos\_escalados, *method* = "euclidean")

esta función mediante el método euclidean nos retorna una matriz de distancias entre las observaciones.

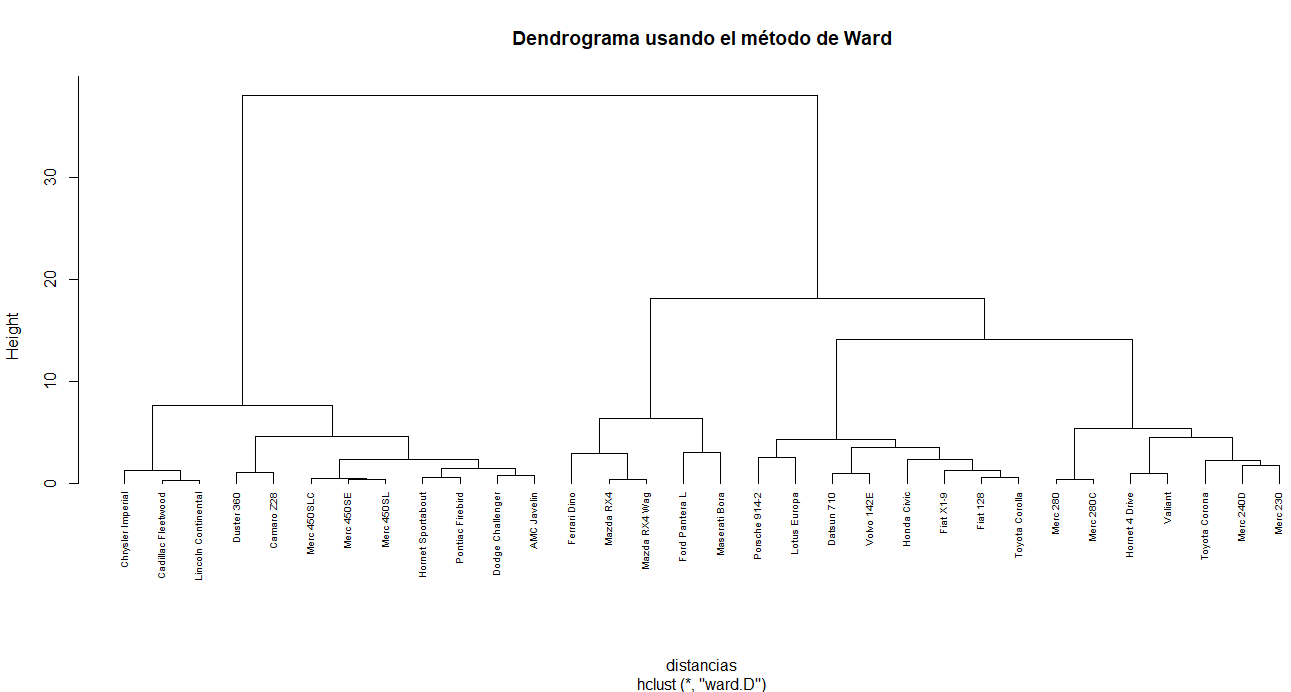
Realizamos el clustering jerárquico mediante el método de Ward con la matriz calculada

hc1 <- hclust(distancias, *method* = "ward.D")

Graficamos el dendograma del metodo de ward

plot(hc1, *cex* = 0.6, *hang* = -1, *main* = "Dendrograma usando el método de Ward")

Obtenemos el siguiente grafico que nos permite observar los clusters de los datos



Del grafico podemos analizar las líneas las cuales nos indica la distancia de unión entre los clusters de manera subjetiva podemos unir y sacar los cluster teniendo.

Para un mejor análisis nos apoyamos de la librería NbClust para determinar el número óptimo de clusters.

set.seed(123) # Asegurar reproducibilidad

res <- NbClust(datos\_escalados, *distance* = "euclidean", *min.nc* = 2, *max.nc* = 10, *method* = "ward.D", *index* = "silhouette")

observamos los resultados

> res

$All.index

2 3 4 5 6 7 8 9 10

0.3788 0.4161 0.4289 0.3989 0.4035 0.4201 0.4442 0.4644 0.4608

$Best.nc

Number\_clusters Value\_Index

9.0000 0.4644

$Best.partition

Mazda RX4 Mazda RX4 Wag Datsun 710

1 1 2

Hornet 4 Drive Hornet Sportabout Valiant

3 4 3

Duster 360 Merc 240D Merc 230

5 6 6

Merc 280 Merc 280C Merc 450SE

7 7 4

Merc 450SL Merc 450SLC Cadillac Fleetwood

4 4 8

Lincoln Continental Chrysler Imperial Fiat 128

8 8 2

Honda Civic Toyota Corolla Toyota Corona

2 2 6

Dodge Challenger AMC Javelin Camaro Z28

4 4 5

Pontiac Firebird Fiat X1-9 Porsche 914-2

4 2 2

Lotus Europa Ford Pantera L Ferrari Dino

2 9 1

Maserati Bora Volvo 142E

9 2

Esto nos indica los posibles números de cluster que podemos tener con su respectivo índice que evalúa la coherencia de los clusters donde el valor mas alto indica el cluster ideal. Procedemos a extraer el número óptimo de clusters sugerido.

optimal\_clusters <- res$Best.nc["Number\_clusters"]

como respuesta témenos 9 clusters óptimos

print(paste("Número óptimo de clusters sugerido por la mayoría de índices:", optimal\_clusters))

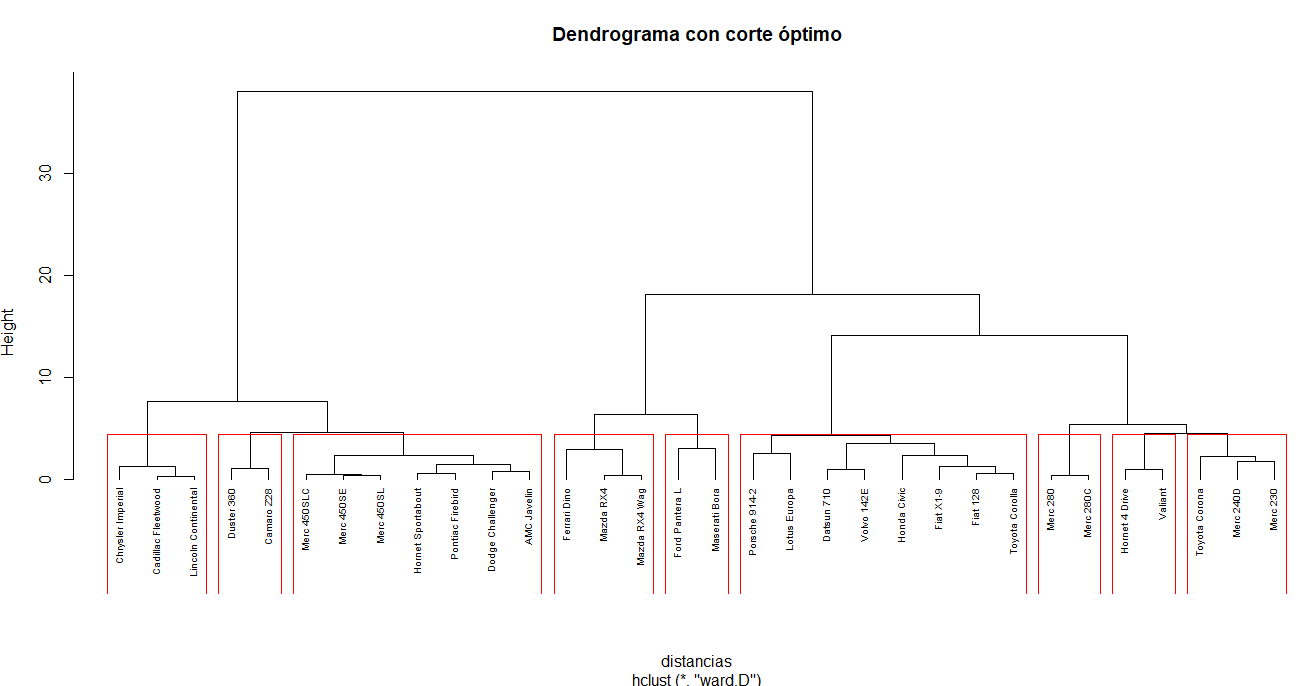
[1] "Número óptimo de clusters sugerido por la mayoría de índices: 9"

para márcalos y tener una mejor visualización usamos la siguiente función.

plot(hc1, *cex* = 0.6, *hang* = -1, *main* = "Dendrograma con corte óptimo")

rect.hclust(hc1, *k*= optimal\_clusters , *border* = "red")

obtenemos el siguiente grafico donde podemos observas de una forma mas clara los 9 clusters.



Estos rectángulos en el dendrograma nos muestran cómo se agrupan las observaciones en los cluster sugeridos.

Seguido de esto procedemos a cortar el dendrograma y obtener los resultados.

resultado\_cluster <- data.frame(datos\_escalados, *clust* = factor(clust))

Podemos observar los resultados de forma tabular para ver cuales observaciones pertenecen a cada cluster.

resultado\_cluster <- cbind( datos\_escalados , clust)

clust

Mazda RX4 1

Mazda RX4 Wag 1

Datsun 710 2

Hornet 4 Drive 3

Hornet Sportabout 4

Valiant 3

Duster 360 5

Merc 240D 6

Merc 230 6

Merc 280 7

Merc 280C 7

Merc 450SE 4

Merc 450SL 4

Merc 450SLC 4

Cadillac Fleetwood 8

Lincoln Continental 8

Chrysler Imperial 8

Fiat 128 2

Honda Civic 2

Toyota Corolla 2

Toyota Corona 6

Dodge Challenger 4

AMC Javelin 4

Camaro Z28 5

Pontiac Firebird 4

Fiat X1-9 2

Porsche 914-2 2

Lotus Europa 2

Ford Pantera L 9

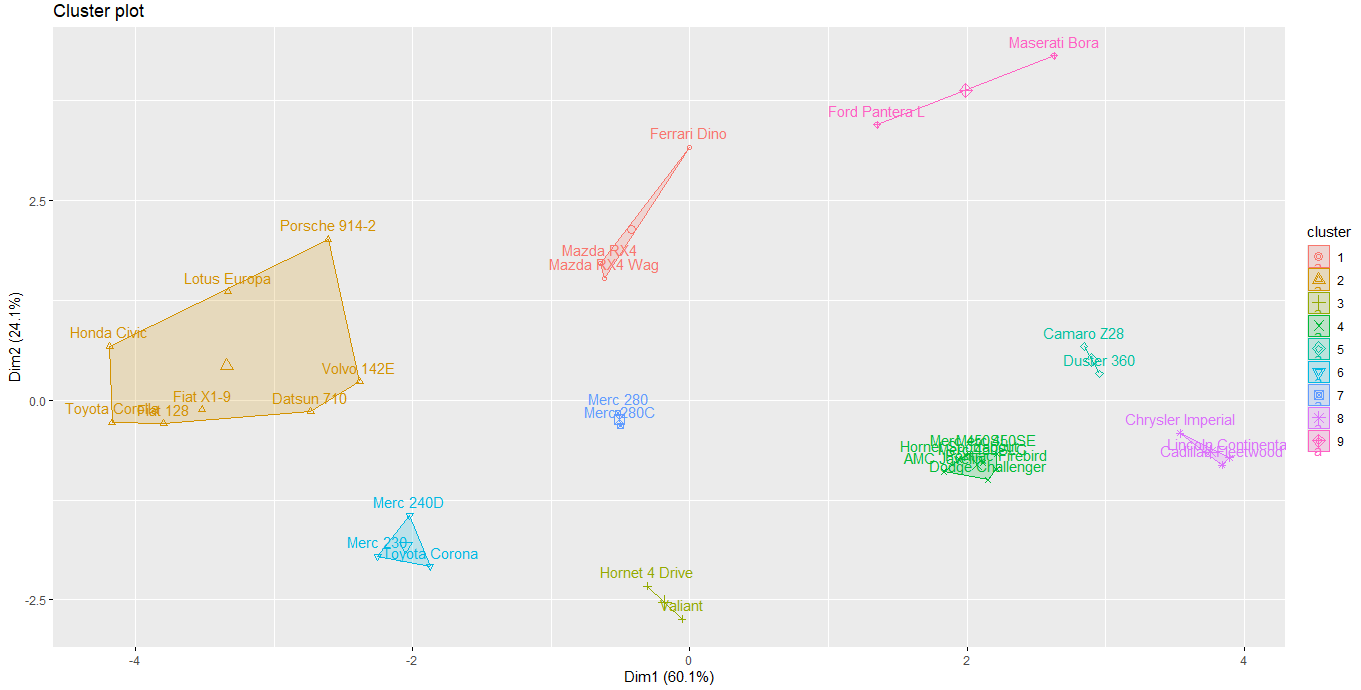
Ferrari Dino 1

Maserati Bora 9

Volvo 142E 2

También podemos tener una visualización en una columna de forma gráfica.

fviz\_cluster(*list*(*data* = datos\_escalados, *cluster* = clust))



Finalmente podemos observar los clusters y que observaciones pertenecen a cada cluster.

como ultimo obtenemos el historial de aglomeración

# Obtener el historial de aglomeración

cluster\_resumen <- data.frame(hc1$merge, *height* = hc1$height)

names(cluster\_resumen) <- c("Cluster1", "Cluster2", "Altura")

print(cluster\_resumen)

Cluster1 Cluster2 Altura

1 -15 -16 0.2956825

2 -12 -13 0.3944266

3 -1 -2 0.4075899

4 -10 -11 0.4082884

5 -14 2 0.4734669

6 -5 -25 0.5475333

7 -18 -20 0.5757917

8 -22 -23 0.7827694

9 -3 -32 0.9799181

10 -4 -6 0.9936969

11 -7 -24 1.0554323

12 -17 1 1.2158431

13 -26 7 1.2795388

14 6 8 1.4570881

15 -8 -9 1.7677287

16 -21 15 2.1967530

17 -19 13 2.3061417

18 5 14 2.3438299

19 -27 -28 2.5613776

20 -30 3 2.8828514

21 -29 -31 3.0287549

22 9 17 3.5376949

23 19 22 4.2647644

24 10 16 4.5371363

25 11 18 4.5658397

26 4 24 5.4003803

27 20 21 6.3641631

28 12 25 7.6650447

29 23 26 14.1342011

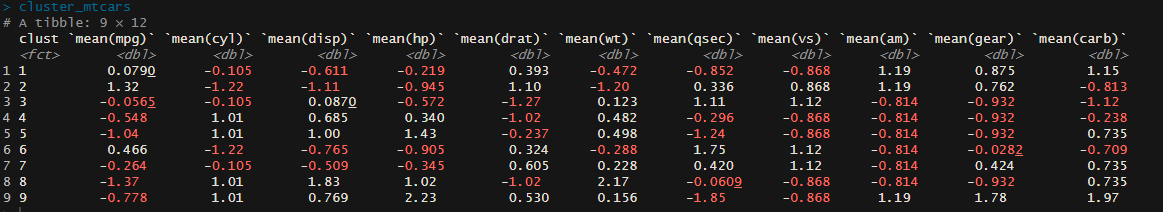
30 27 29 18.1889754

31 28 30 38.0300925

Con esto podemos observar un resumen de fusiones durante el análisis de agrupamiento jerárquico, incluyendo los cluster que se fusionan y la distancia a la que se fusionan como en el primero el -15 y -16 en la altura 0.2956 y de esa misma forma los demás.

Para finalizar agrupamos por clust y calculamos las medias de las variables deseadas.

Obtenemos los siguientes resultados



Procedemos a concluir

Lo que podemos observar son los valores de las medias de cada variable para cada cluster lo cual nos permite inferir de las agrupaciones sobre que valor de la media están, esto nos permitirá hacer campañas de marketing y marketing personalizado para algunos grupos de las observaciones de mtcars como, por ejemplo:

-El cluster 2 tiene una eficiencia de combustible muy superior a la media y su peso es muy inferior a la media, esto puede ser un factor muy fuerte para la publicidad y oferta de este grupo de vehículos en el mercado para consumidores que demanden vehículos con eficiencia de combustible y sean livianos.

-El cluster 3 por otro lado tiene una eficiencia de combustible muy inferior a la media un hp muy superior a la media, esto puede ser un factor muy fuerte para la oferta de estos vehículos en el mercado para consumidores que demanden vehículos con bastante fuerza.