

# Pràctica 1, Eines d'ajuda a la presa de decisions

*Oscar Galera i Alfaro*

*15 d'Octubre, 2018*

## Analisi del Data Set Auto MPG

En aquesta pràctica s'analitzar el fitxer de dades que hi ha disponible en el següent enllaç <https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data>

## Carregar les dades

El primer que cal fer és carregar el fitxer que conté les dades, per això executem la següent comanda

```
#Carregar les dades
dades <- read.csv(file="dades.csv", header=FALSE, sep = ",")
```

## Preprocessament de les dades

Si tot va bé, s'hauria de generar un nou dataset amb 398 registres i 9 variables. El tipus d'aquestes variables correspon a:

- V1. mpg: V. Quantitativa continua
- V2. cylinders: V. Quantitativa discreta
- V3. displacement: V. Quantitativa continua
- V4. horsepower: V. Quantitativa continua
- V5. weight: V. Quantitativa continua
- V6. acceleration: V. Quantitativa continua
- V7. model year: V. Quantitativa discreta
- V8. origin: V. Quantitativa discreta
- V9. car name: V. Qualitativa

Com que el fitxer de dades no conté el nom de les variables i perquè sigui més fàcil la seva interpretació, fem l'assignació de noms amb la següent comanda.

```
#Renomenar les columnes
names(dades) <- c("mpg", "cylinders", "displ", "hp", "weight", "accel", "model_year",
                  "origin", "car_name")
```

Per veure com ha interpretat les variables R, executem la següent comanda

```
#Mostrar el tipus de les variables
str(dades)
```

```
## 'data.frame': 398 obs. of 9 variables:
## $ mpg : num 18 15 18 16 17 15 14 14 15 ...
## $ cylinders : int 8 8 8 8 8 8 8 8 8 ...
## $ displ : num 307 350 318 304 302 429 454 440 455 390 ...
## $ hp : Factor w/ 94 levels "?","100.0","102.0",...: 17 35 29 29 24 42 47 46 48 40 ...
## $ weight : num 3504 3693 3436 3433 3449 ...
## $ accel : num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ model_year: int 70 70 70 70 70 70 70 70 70 70 ...
## $ origin : int 1 1 1 1 1 1 1 1 1 1 ...
## $ car_name : Factor w/ 305 levels "amc ambassador brougham",...: 50 37 232 15 162 142 55 224 242 2
```

Dels resultats obtinguts, veiem que R reconeix hp com una variable qualitativa quan sabem que es tracta d'una variable quantitativa, amb la següent comanda podem veure la distribució de valors que pren la variable

```
#Mostrar la distribució de valors que pren la variable hp
levels(dades$hp)
```

```
## [1] "?" "100.0" "102.0" "103.0" "105.0" "107.0" "108.0" "110.0"
## [9] "112.0" "113.0" "115.0" "116.0" "120.0" "122.0" "125.0" "129.0"
## [17] "130.0" "132.0" "133.0" "135.0" "137.0" "138.0" "139.0" "140.0"
## [25] "142.0" "145.0" "148.0" "149.0" "150.0" "152.0" "153.0" "155.0"
## [33] "158.0" "160.0" "165.0" "167.0" "170.0" "175.0" "180.0" "190.0"
## [41] "193.0" "198.0" "200.0" "208.0" "210.0" "215.0" "220.0" "225.0"
## [49] "230.0" "46.00" "48.00" "49.00" "52.00" "53.00" "54.00" "58.00"
## [57] "60.00" "61.00" "62.00" "63.00" "64.00" "65.00" "66.00" "67.00"
## [65] "68.00" "69.00" "70.00" "71.00" "72.00" "74.00" "75.00" "76.00"
## [73] "77.00" "78.00" "79.00" "80.00" "81.00" "82.00" "83.00" "84.00"
## [81] "85.00" "86.00" "87.00" "88.00" "89.00" "90.00" "91.00" "92.00"
## [89] "93.00" "94.00" "95.00" "96.00" "97.00" "98.00"
```

Com es pot veure, hi ha observacions amb valor abscent (?) i per aquest motiu el inferior de R no ha detectat correctament el tipus. Quantes d'aquestes observacions tenen valors abscents?

```
#Nombre d'observacions amb valor abscent (?) en la variable hp
nrow(dades[dades$hp == "?",])
```

```
## [1] 6
```

De les diferents estratègies per resoldre aquesta situació, s'obta per eliminar les observació afectades perquè es recomana complir la regla  $N > 20p$  (on  $N$  correspon al nombre d'observacions i  $p$  al nombre de variables) per a poder fer l'anàlisi sense problemes, i en aquest cas s'assoleix el valor mínim, ja que tenim  $392 \gg 180$

```
#Filtrar les observacions que tenen valor abscent en la variable hp
dades = dades[dades$hp != "?", ]
```

Fet això, ja es pot convertir la variable hp de qualitativa a quantitativa

```
#Convertir hp de qualitativa a quantitativa
dades$hp = as.numeric(dades$hp)
str(dades)
```

```
## 'data.frame': 392 obs. of 9 variables:
## $ mpg : num 18 15 18 16 17 15 14 14 15 ...
## $ cylinders : int 8 8 8 8 8 8 8 8 8 ...
## $ displ : num 307 350 318 304 302 429 454 440 455 390 ...
## $ hp : num 17 35 29 29 24 42 47 46 48 40 ...
## $ weight : num 3504 3693 3436 3433 3449 ...
## $ accel : num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ model_year: int 70 70 70 70 70 70 70 70 70 70 ...
```

```
## $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
## $ car_name    : Factor w/ 305 levels "amc ambassador brougham",...: 50 37 232 15 162 142 55 224 242 2
```

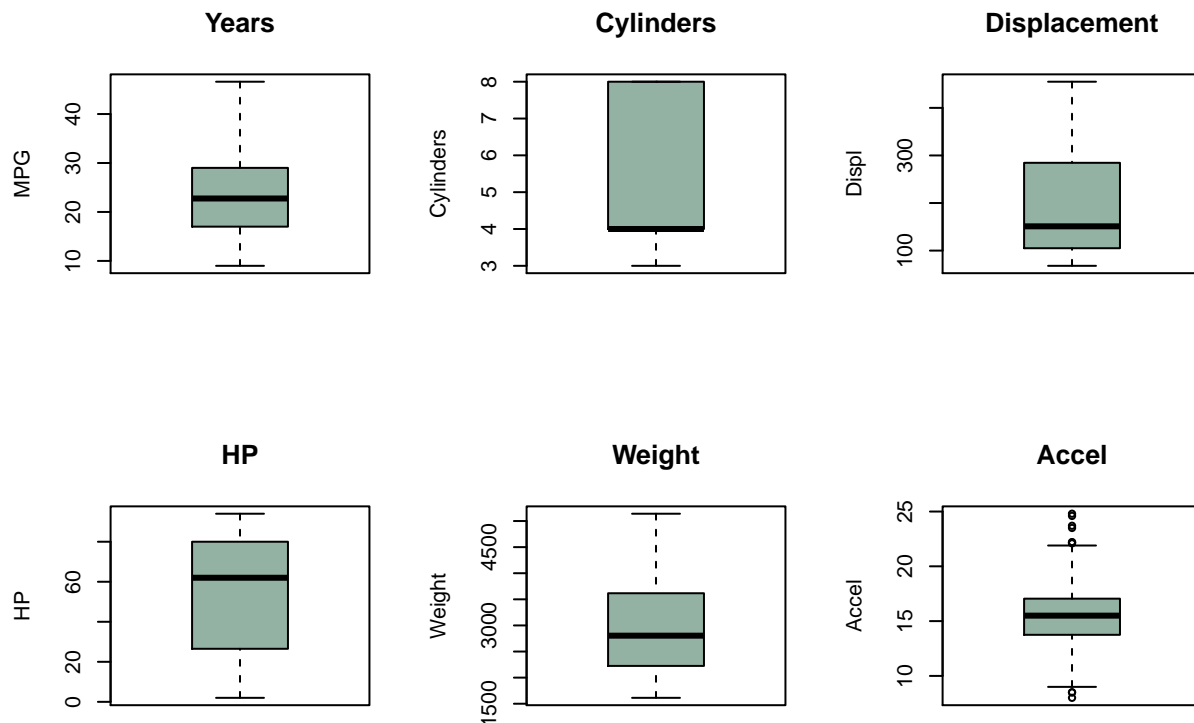
Arribats a aquest punt ja tenim les dades preparades per començar a treballar.

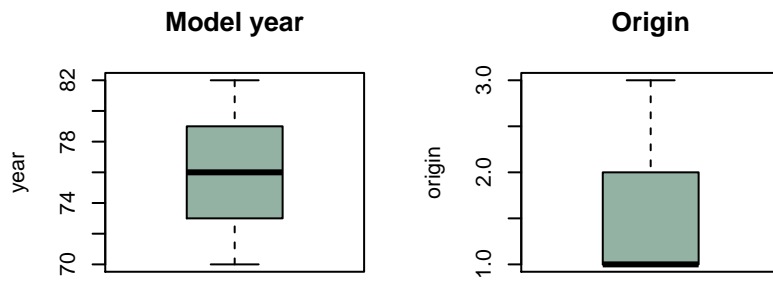
## Analisi descriptiu bàsic

```
bplotNumerics <- function(){
  #Agrupar els gràfics en tripletes
  attach(mtcars)
  par(mfrow=c(2,3))

  #Boxplots
  boxplot(dades$mpg, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="MPG", main="Years")
  boxplot(dades$cylinders, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Cylinders", main="Cylinders")
  boxplot(dades$displ, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Displ", main="Displacement")
  boxplot(dades$hp, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="HP", main="HP")
  boxplot(dades$weight, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Weight", main="Weight")
  boxplot(dades$accel, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Accel", main="Accel")
  boxplot(dades$model_year, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="year", main="Model year")
  boxplot(dades$origin, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="origin", main="Origin")
}

bplotNumerics()
```





## Es pot aplicar l'anàlisi de components principals?

L'anàlisi de components principals o ACP permet descriure un conjunt de dades, resumint-lo i reduir la seva dimensionalitat. A través d'aquest anàlisi buscarem resposta a les preguntes:

- Estudi dels vehicles: Dos vehicles es consideraran similars si tenen característiques semblants.
- Estudi de les característiques

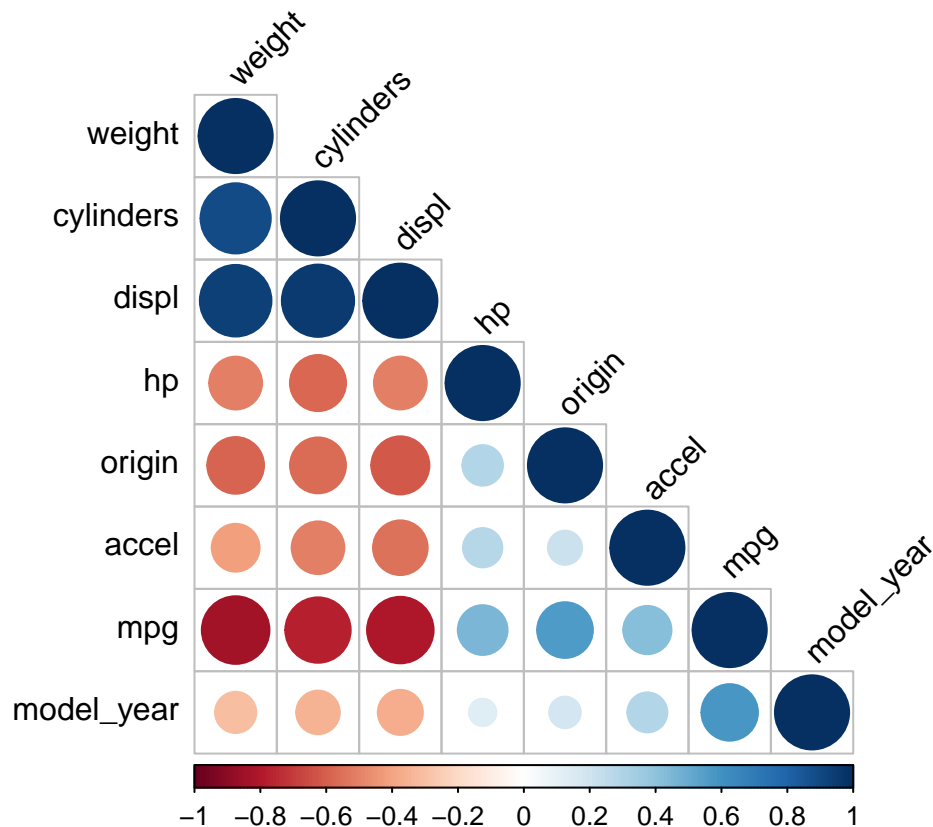
Primer però, cal comprovar que es pugui realitzar l'anàlisi de components principals. Això serà així si la correlació entre variables és significativa. En la següent matriu es pot veure la correlació entre les variables del conjunt de dades.

```
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```
cor.mat <- cor(dades[,1:8])
```

```
corrplot(cor.mat, type="lower", order="hclust", tl.col="black", tl.srt=45)
```



```
library("PerformanceAnalytics")
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
```

```
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##   as.Date, as.Date.numeric
```

```
##
```

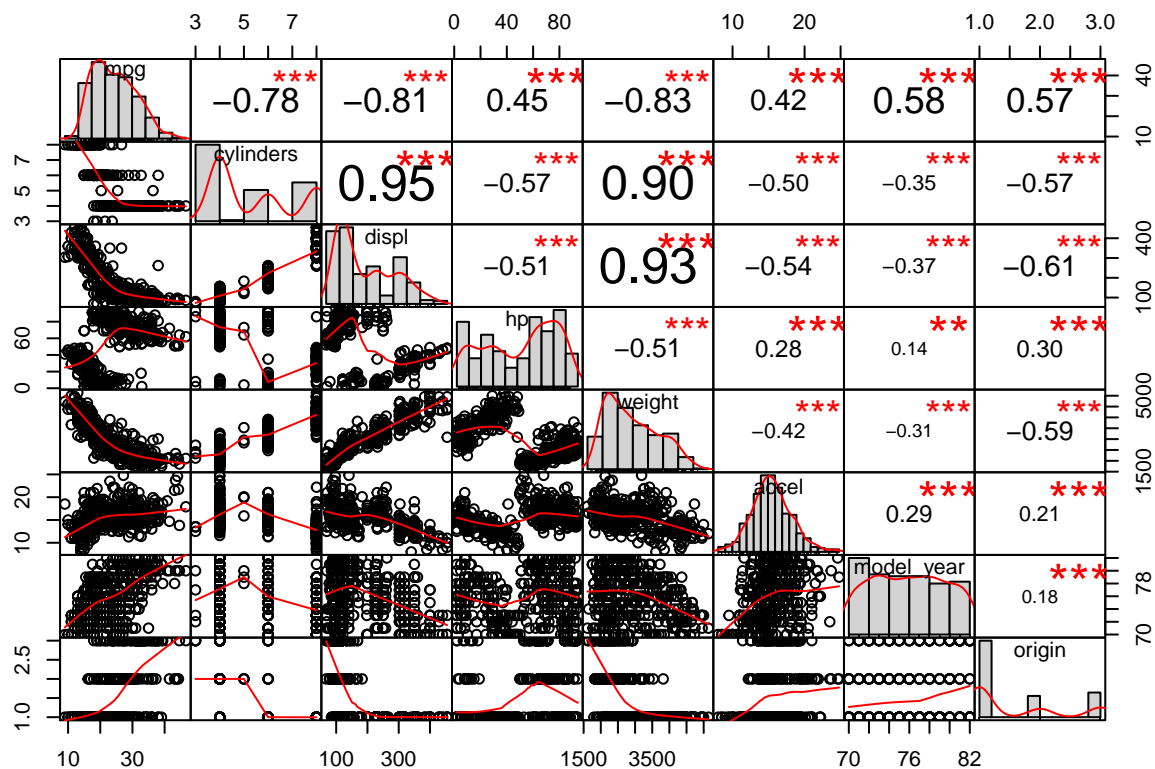
```
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
```

```
##
```

```
##   legend
```

```
chart.Correlation(dades[,1:8], histogram = TRUE, pch = 19)
```



Seguidament es fa el test d'esfericitat de Barlett

```
library(psych)
```

```
cortest.bartlett(cor.mat, n=100 )
```

```
## $chisq
```

```
## [1] 733.6116
```

```
##
```

```
## $p.value
```

```
## [1] 1.808352e-136
```

```
##
```

```
## $df
## [1] 28
```

Degut a que el p-value és molt petit (pròxim a 0) descartem la hipòtesis nula (homogeneïtat de variàncies) i acceptem com a hipòtesi alternativa que ens diu que hi ha diferència entre les variàncies de les variables (heterogeneïtat de variàncies).

Seguidament apliquem un test de *Kaiser – Mayer – Olkin*

```
library(psych)
KMO(cor.mat)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor.mat)
## Overall MSA = 0.8
## MSA for each item =
##      mpg  cylinders      displ      hp      weight      accel
##      0.80      0.85      0.76      0.89      0.79      0.79
## model_year      origin
##      0.58      0.87
```

Com obtenim un resultat de 0.8, assumim que es pot aplicar l'anàlisi de components principals.

## Aplicant l'anàlisi de components principals?

Primer seleccionem les variables que seràn actives en aquest anàlisi

```
dades.PCA <- dades[, 1:8]
library(FactoMineR)
(res <- PCA(dades.PCA, scale.unit=TRUE, ncp=8, graph=FALSE))
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 392 individuals, described by 8 variables
## *The results are available in the following objects:
##
##      name      description
## 1  "$eig"      "eigenvalues"
## 2  "$var"      "results for the variables"
## 3  "$var$coord" "coord. for the variables"
## 4  "$var$cor"   "correlations variables - dimensions"
## 5  "$var$cos2"  "cos2 for the variables"
## 6  "$var$contrib" "contributions of the variables"
## 7  "$ind"      "results for the individuals"
## 8  "$ind$coord" "coord. for the individuals"
## 9  "$ind$cos2"  "cos2 for the individuals"
## 10 "$ind$contrib" "contributions of the individuals"
## 11 "$call"      "summary statistics"
## 12 "$call$centre" "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"  "weights for the individuals"
## 15 "$call$col.w"  "weights for the variables"
```

Com que hi ha 8 variables actives, s'han trobat 8 eixos factorials. Però quins són els eixos més representatius? Segons el criteri de "Latent Root" es poden considerar tots aquells eixos amb un valor propi superior a 1, és a dir que tenen més inèrcia que qualsevol de les variables originals.

En la següent llista es mostra la inèrcia per a cada un dels eixos.

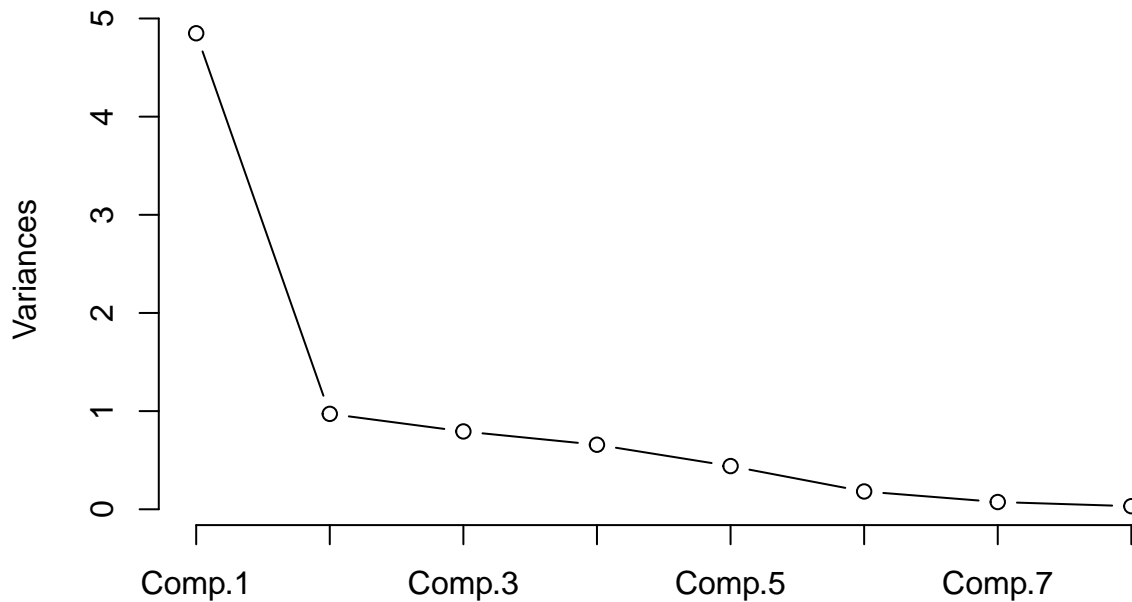
```
res$eig[,1]
```

```
##      comp 1      comp 2      comp 3      comp 4      comp 5      comp 6
## 4.84933335 0.97218078 0.79304775 0.65788136 0.44022008 0.18166134
##      comp 7      comp 8
## 0.07399022 0.03168512
```

Per veure-ho amb més perspectiva, apliquem un Scree Plot entre els components principals i el seu valor pròpi per veure on es troba la discontinuïtat.

```
screepLOT(princomp(dades.PCA,cor=TRUE),type = c("lines"))
```

**princomp(dades.PCA, cor = TRUE)**



```
library("factoextra")
```

```
## Loading required package: ggplot2
```

```
##
```

```
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
```

```
##
```

```
##      %+%, alpha
```

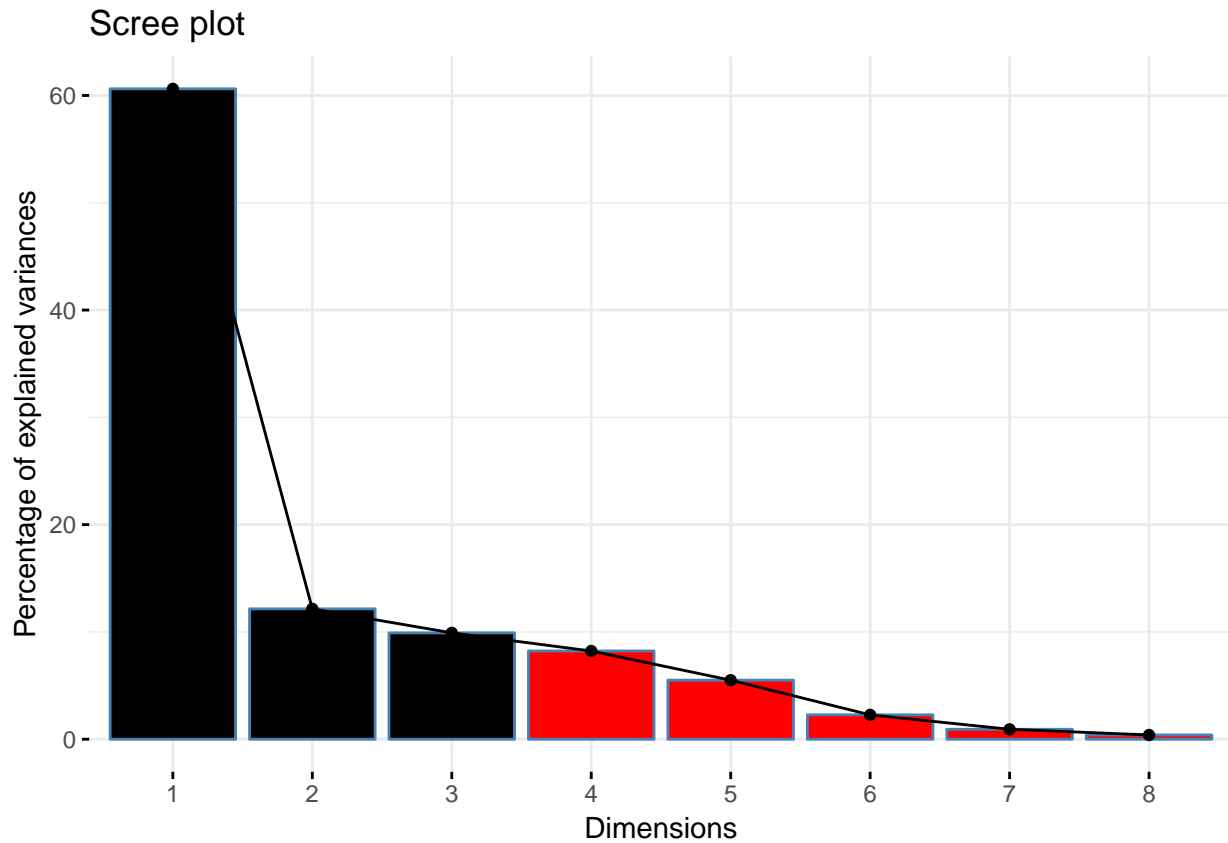
```
## The following object is masked from 'mtcars':
```

```
##
```

```
##      mpg
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_screepLOT(res, ncp=8, barfill=c(rep(1,3),rep(2,5)))
```



És fàcil veure que aquesta discontinuïtat ja es troba a partir del segon component. Només els tres primers factors contenen el 82.68% de la informació.

```
res[["eig"]]
```

##	eigenvalue	percentage of variance	cumulative percentage of variance
## comp 1	4.84933335	60.6166669	60.61667
## comp 2	0.97218078	12.1522597	72.76893
## comp 3	0.79304775	9.9130969	82.68202
## comp 4	0.65788136	8.2235170	90.90554
## comp 5	0.44022008	5.5027510	96.40829
## comp 6	0.18166134	2.2707667	98.67906
## comp 7	0.07399022	0.9248777	99.60394
## comp 8	0.03168512	0.3960640	100.00000

Quines variables han sigut les que més han influït en aquests eixos factorials?

```
res[["var"]][["contrib"]]
```

##	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
## mpg	16.576079	3.500464	4.02636297	1.14794250	1.246773
## cylinders	18.300466	1.412842	0.37178190	0.04680311	6.391609
## displ	18.976098	0.565500	0.01480703	1.89551621	5.632553
## hp	7.557651	12.074596	15.75436655	50.43598118	13.349304
## weight	17.775420	2.173861	0.65096334	0.15477995	15.558894
## accel	6.834005	8.092882	47.40171217	25.85272535	9.213599
## model_year	4.813065	64.274256	4.54250503	11.03283627	3.261315

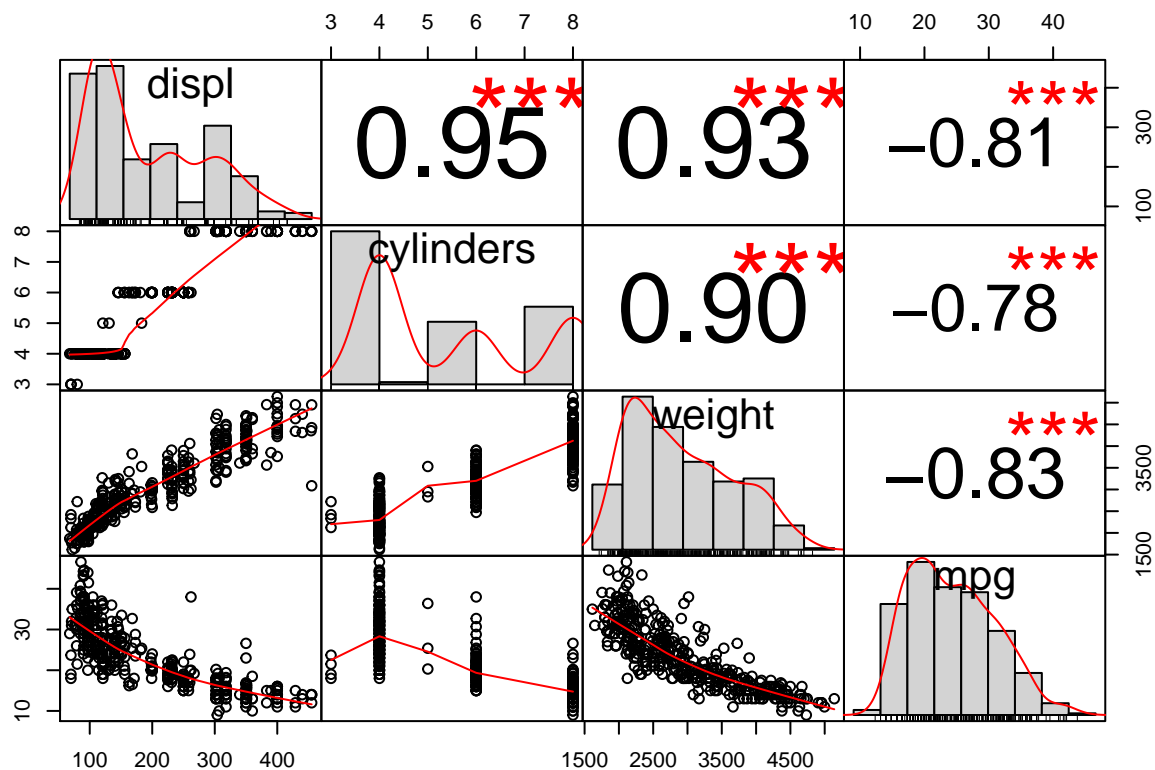


```
## origin      9.167216  7.905599 27.23750101  9.43341543 45.345953
##              Dim.6      Dim.7      Dim.8
## mpg         58.79353938 13.4795852  1.2292539
## cylinders    19.36120403 35.6983514 18.4169417
## displ        9.00596757 0.6963108 63.2132470
## hp           0.06825386 0.5620917 0.1977565
## weight       1.51577813 46.6855785 15.4847253
## accel       1.59850815 0.3249305 0.6816374
## model_year   9.28266676 2.4210501 0.3723068
## origin       0.37408211 0.1321018 0.4041315
```

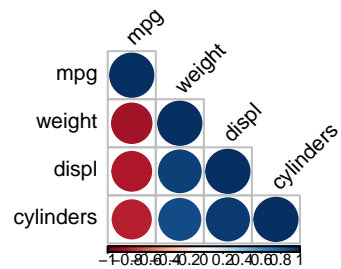
En primer eix factorial, les variables que més han influït són: displ, cylinders, weight i mpg amb 18.98%, 18.30%, 17.77 i 16.57 respectivament. Mentre que en el segon eix les variables que més han contribuït són: model\_year i hp amb 64.27% i 12.07% respectivament.

```
#Agrupar els gràfics en tripletes
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##      mpg
##
## The following objects are masked from mtcars (pos = 11):
##
##      am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
par(mfrow=c(2,3))
chart.Correlation(dades[,c("displ", "cylinders", "weight", "mpg")], histogram = TRUE, pch = 19)
```



```
corrplot(cor(dades[, c("displ", "cylinders", "weight", "mpg")]), type="lower", order="hclust", tl.col="")
```



Quines són les variables més ben representades en cada eix factorial?

```
plot.PCA(res, axes=c(1,2), choix="ind", habillage="none", col.ind="black", col.ind.sup="blue", col.quali="mag
```

