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.

```
#Renombrar 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 14 15 ...
##  $ cylinders : int  8 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 aquesta variable

```r
#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 inferidor de R no ha detectat correctament el tipus. Quantes d'aquestes observacions tenen valor abscent?

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

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

De les diferents estartègies 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'assoleig el valor mínim, ja que tenim $392 > 180$

```r
#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

```r
#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 14 15 ...
##  $ cylinders : int  8 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
```

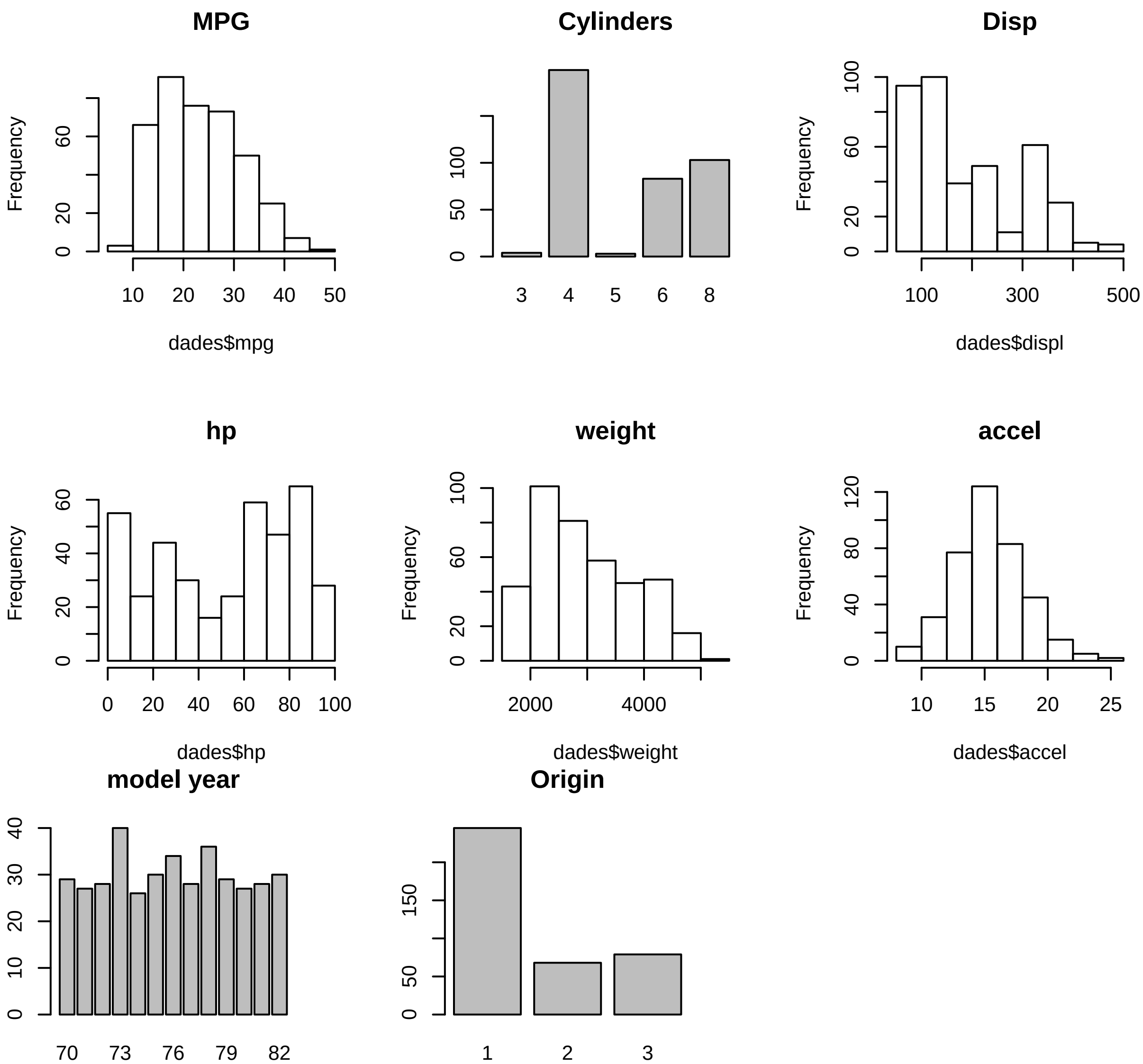
Arribats a aquest punt ja tenim les dades preparades per començar a treballar.

Analisi descriptiu bàsic

Per aquest analisi bàsic, es vol veure els valors de centralitat i dispersió per cada variable. En els següents diagrames es mostra la distribució de valors per les variables.

Histogrames

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

  hist(x = dades$mpg, main="MPG")
  barplot(table(dades$cylinders), main="Cylinders")
  hist(x = dades$displ, main = "Disp")
  hist(x = dades$hp, main="hp")
  hist(x = dades$weight, main="weight")
  hist(x = dades$accel, main="accel")
  barplot(table(dades$model_year), main="model year")
  barplot(table(dades$origin), main="Origin")
}
hNumerics()
```



Es destaca que les variables "Accel" i "Model year" tenen un alt grau de simetria, i en el cas de "Model year" la seva distribució és força uniforme.

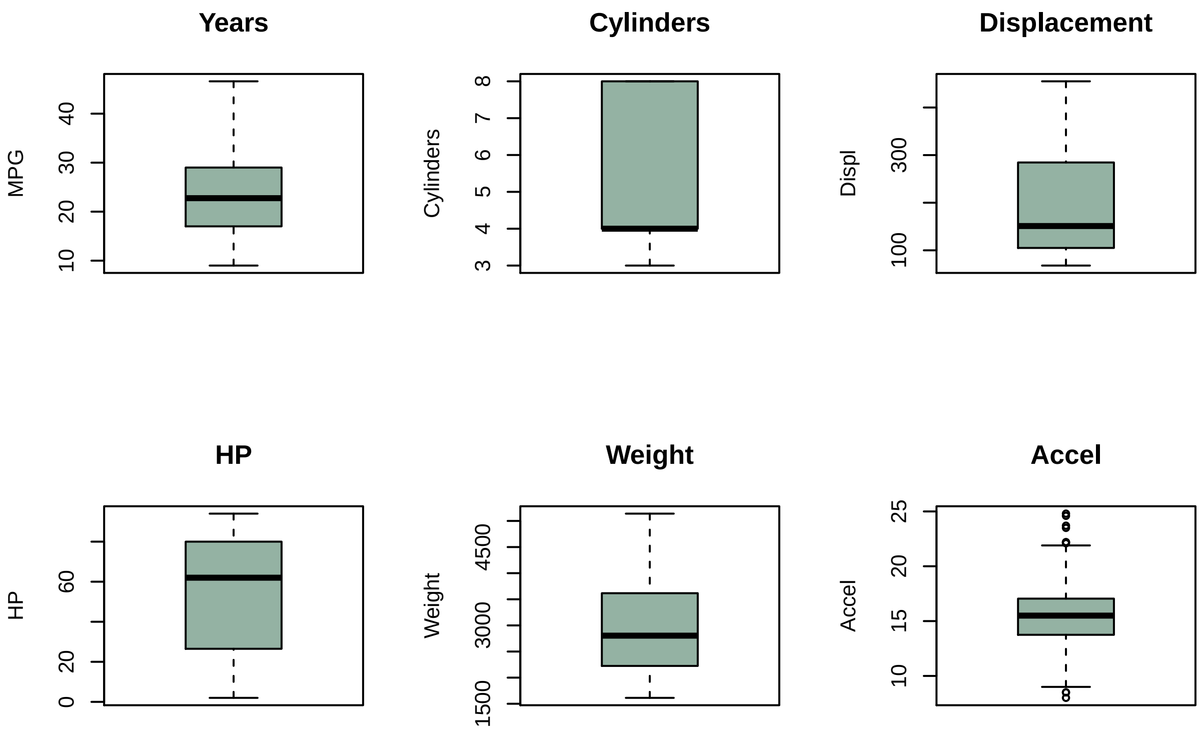
Diagrames de caixa

Un altre tipus de gràfic molt útil per aquesta finalitat és el diagrama de caixa.

```
bplotNumerics <- function(){
  attach(mtcars)
  par(mfrow=c(2,3))

  #Bloxplots
  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")
```

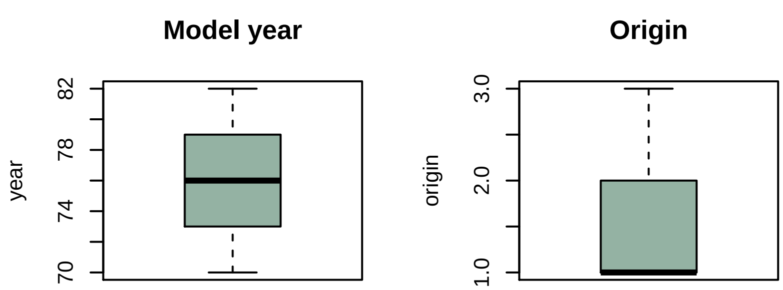
```r
  boxplot(dades$model_year, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="year", main="Model year")
  boxplot(dades$origin, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="origin", main="Origin")
}
bplotNumerics()
```

```
## The following objects are masked from mtcars (pos = 3):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```




```r
print("Accel")
```

```
## [1] "Accel"
```



Aquí es pot tornar a veure com hi ha un clar biaix a la dreta per la variable "Cylinders" i "Origin", també es pot observar com les variables "Accel" i "Model Year" tenen certa simetria (com s'ha dit anteriorment) i com hi ha dades atípiques per la variable "Accel".

Les dades atípies inferiors són aquelles que tenen un valor menor a $1Q - 1.5RIQ$ i les superiors aquelles amb un valor major a $3Q + 1.5RIQ$.

```r
#Valors atípics per la variable Accel.
#Inferiors
dades[dades$accel <= 13.78 - 1.5 * IQR(dades$accel), ][c("accel", "model_year", "car_name")]
```

```
##    accel model_year           car_name
## 8    8.5         70  plymouth fury iii
## 10   8.5         70 amc ambassador dpl
## 12   8.0         70 plymouth 'cuda 340
```

```
#Superiors
dades[dades$accel >= 17.02 + 1.5 * IQR(dades$accel), ][c("accel", "model_year", "car_name")]
```

```
##     accel model_year                          car_name
## 60   23.5         72                 volkswagen type 3
## 196  22.2         76                chevrolet chevette
## 197  22.1         76                   chevrolet woody
## 210  21.9         76                       peugeot 504
## 300  24.8         79                       peugeot 504
## 301  22.2         79 oldsmobile cutlass salon brougham
## 327  23.7         80                vw dasher (diesel)
## 395  24.6         82                         vw pickup
```

Resum de les variables

En les següents taules es mostren els estadístics de centralistat i dispersió classificats com a robustos i no robustos

```
#install.packages("psych")
library(knitr)
library(kableExtra)
taulaResum <- function(dades, nom, fw = TRUE){
  library(psych)
  #Estadístics de centralitat
  dades <- na.omit(dades)
  mitja <- mean(dades)
  mediana <- median(dades)
  mitjana_retallada_05 <- mean(dades, trim=0.5, na.rm = TRUE)

  if(fw){
    mitjana_winsoritzada_05 <- winsor.mean(dades, trim=0.5, na.rm=TRUE)
  }else{
    mitjana_winsoritzada_05 <- NA
  }

  #Estadístics de dispersió
  sd <- sd(dades)
  iqr <- IQR(dades)
  mad <- max(dades)

  df <- data.frame("Estadistic" =
                    c("Mediana", "Mit. Retallada", "Mit. Winsoritzada", "Mitjana"
                      , "RIC", "MAD", "D. Estàndard"), "Valor" =
                    c(mediana, mitjana_winsoritzada_05,
                      mitjana_winsoritzada_05, mitja, iqr, mad, sd))
  kable(df, caption = paste("Variable ",nom)) %>%
    kable_styling("striped", full_width = F) %>%
    group_rows("C. robustos", 1, 4) %>%
    group_rows("C. NO robustos", 4, 4) %>%
```

Table 1: Variable mpg

Estadistic	Valor
C. robustos	
Mediana	22.750000
Mit. Retallada	22.750000
Mit. Winsoritzada	22.750000
C. NO robustos	
Mitjana	23.445918
D. robustos	
RIC	12.000000
MAD	46.600000
D. NO robustos	
D. Estàndard	7.805008

Table 2: Variable disp

Estadistic	Valor
C. robustos	
Mediana	151.000
Mit. Retallada	151.000
Mit. Winsoritzada	151.000
C. NO robustos	
Mitjana	194.412
D. robustos	
RIC	170.750
MAD	455.000
D. NO robustos	
D. Estàndard	104.644

```
    group_rows("D. robustos", 5, 6) %>%
    group_rows("D. NO robustos", 7, 7)
}
```

```
taulaResum(dades$mpg, "mpg")
```

```
taulaResum(dades$displ, "disp")
```

```
taulaResum(dades$hp, "hp")
```

```
taulaResum(dades$weight, "weight")
```

```
taulaResum(dades$accel, "accel")
```

Es pot aplicar l'analisi de components principals?

L'anàlisi de components principals o ACP permet descriure un conjunt de dades, resumint-lo i reduir la seva dimensionalitat. En aquest cas, s'usarà per interpretar la relació que hi ha entre les variables.

Del conjunt de variables disponibles, s'ha deccidit utilitzar la variable "car_name" com a variable suplementaria en la representació i per tant queda exclosa de les variables actives.

Table 3: Variable hp

Estadistic	Valor
C. robustos	
Mediana	62.00000
Mit. Retallada	62.00000
Mit. Winsoritzada	62.00000
C. NO robustos	
Mitjana	52.16071
D. robustos	
RIC	53.25000
MAD	94.00000
D. NO robustos	
D. Estàndard	29.49805

Table 4: Variable weight

Estadistic	Valor
C. robustos	
Mediana	2803.5000
Mit. Retallada	2803.5000
Mit. Winsoritzada	2803.5000
C. NO robustos	
Mitjana	2977.5842
D. robustos	
RIC	1389.5000
MAD	5140.0000
D. NO robustos	
D. Estàndard	849.4026

Table 5: Variable accel

Estadistic	Valor
C. robustos	
Mediana	15.500000
Mit. Retallada	15.500000
Mit. Winsoritzada	15.500000
C. NO robustos	
Mitjana	15.541327
D. robustos	
RIC	3.250000
MAD	24.800000
D. NO robustos	
D. Estàndard	2.758864

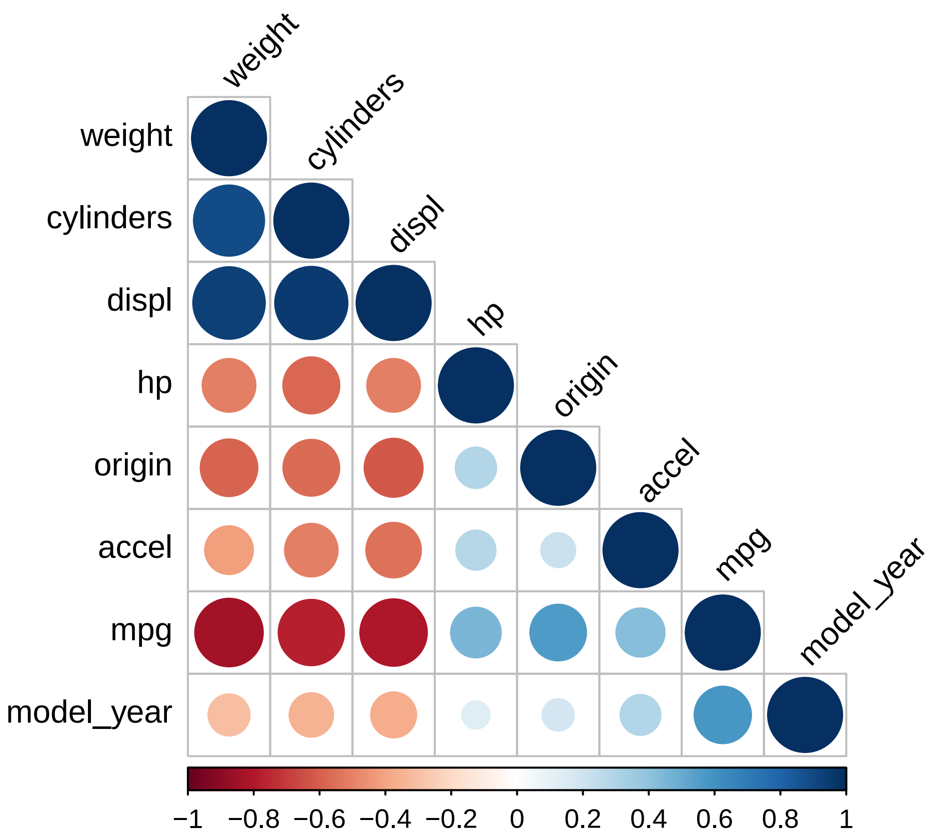
```
#Seleccionar varialbes actives
dades.PCA <- dades[, c("mpg", "cylinders", "displ", "hp", "weight", "accel", "model_year",
                     "origin")]

names(dades.PCA)
```

```
## [1] "mpg"        "cylinders"  "displ"      "hp"         "weight"
## [6] "accel"      "model_year" "origin"
```

Abans de fer l'anàlisi però, cal comprovar que aquest es pugui realitzar. Això serà així si la correlació entre variables és significativa (hi ha heterocedasticitat).

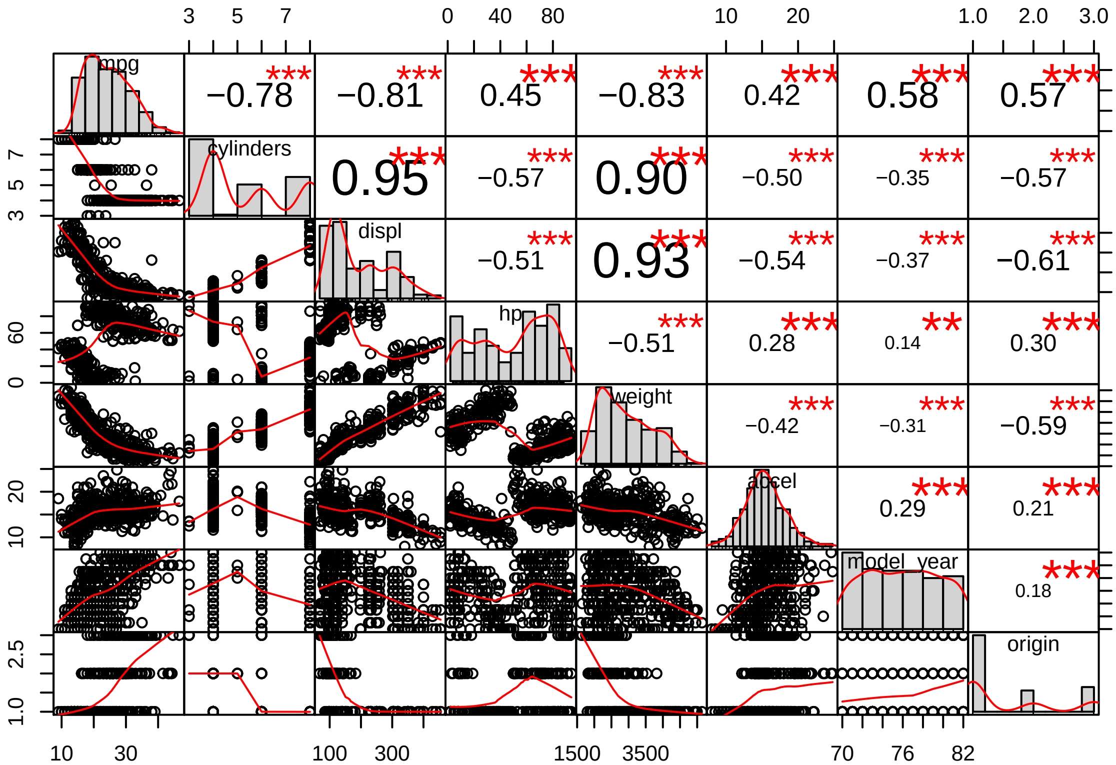
Matriu esquemàtica de correlació entre variables.

```
library("corrplot")
cor.mat <- cor(dades.PCA)
corrplot(cor.mat, type="lower", order="hclust", tl.col="black", tl.srt=45)
```



Matriu de gràfics bivariants i coeficient de correlació de Pearson.

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



Seguidament es fa el test d'esfericitat de Barlett, que comprova si hi ha com a mínim dues de les variables de treball tenen diferent variancia, és a dir, que aplica el següent contrast d'hipòtesis:

$$H_0 : s_1^2 = s_2^2 \wedge s_3^2 ... \wedge s_k^2$$

$$H_1 : s_1^2 \neq s_2^2 \vee s_2^2 \neq s_3^2 ... \vee \neq s_k^2$$

On k correspon al nombre de variables.

```r
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) rebutgem la hipòtesis nul·la (homocedasticitat) i acceptem amb un nivell de confiança superior a 99% que com a mínim hi ha una variable amb una desviació estàndard diferent al de la resta.

I també apliquem un test de $Kaiser - Mayer - Olkin$

```r
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 el test és positiu i que es pot aplicar l'anàlisi de components principals.

Aplicant l'analisi de components principals

Com que ja sabem que hi ha heterocedasticitat, apliquem la tècnica de l'anàlisi de components principals amb la funció PCA del paquet $FactoMineR$.

```r
library(FactoMineR)
res <- PCA(dades.PCA, scale.unit=TRUE, ncp=5, graph=FALSE)
```

Eixos factorials

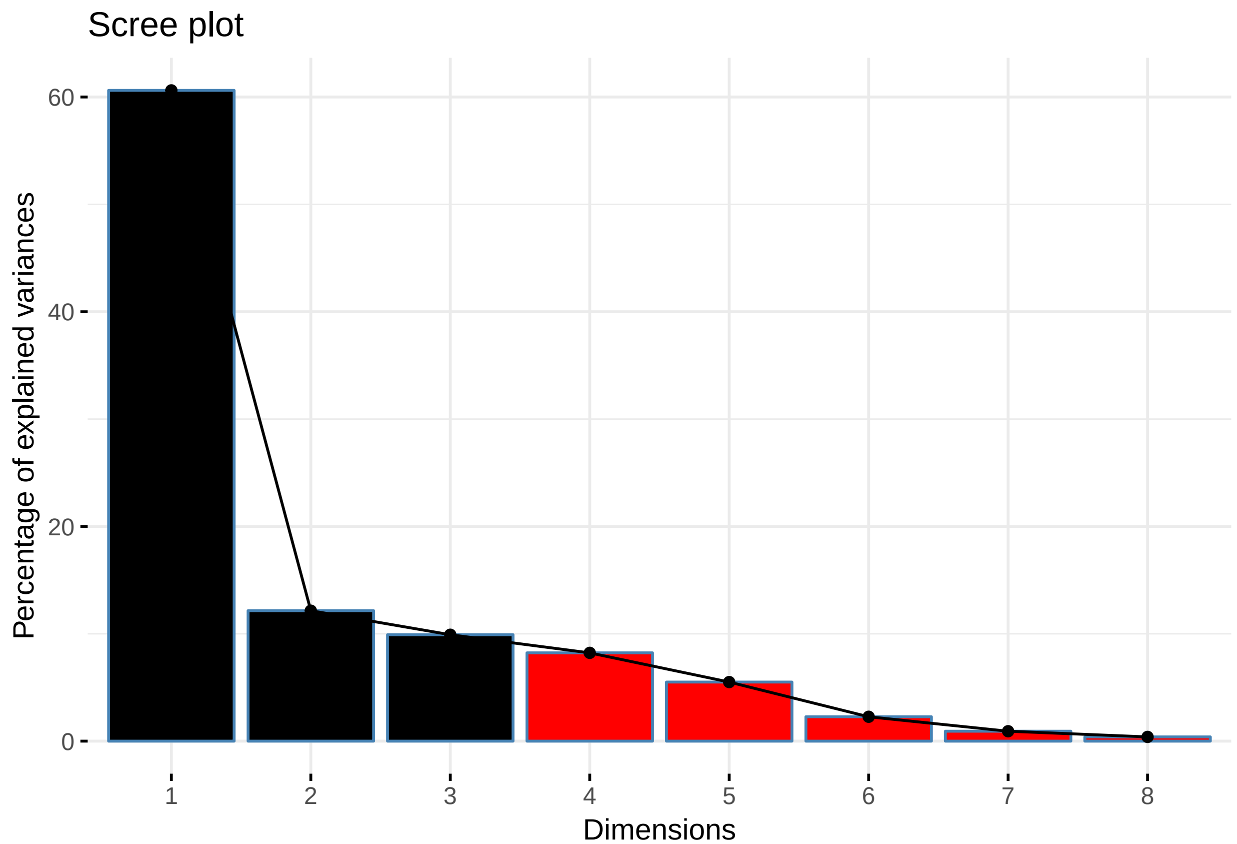
Com que hi ha buit variables actives, s'han trobat buit eixos factorials. En la següent llista es mostra la inèrcia que conté cada un d'aquests eixos, aquesta inèrica ens ve representada pels valors pròpis de la matriu de variancies.

```r
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
```

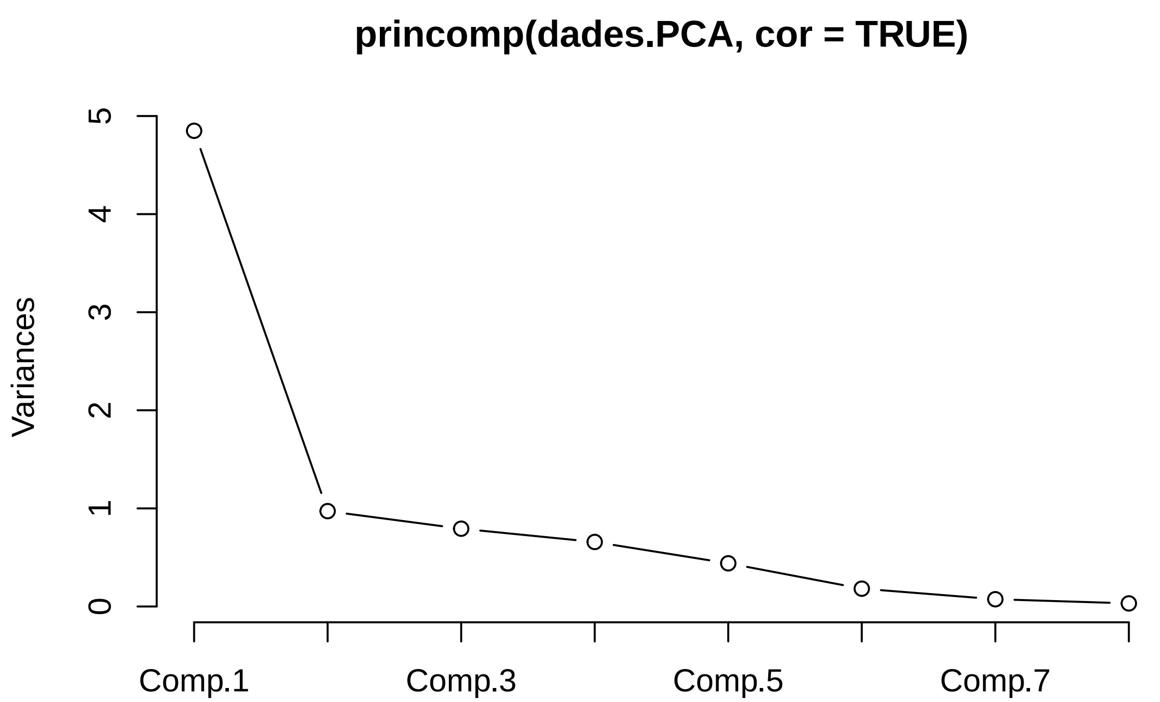
Però és clar, no serveix de gaire utilitzar-los tots, i així doncs, quins són els eixos més representatius? Segons el criteri de "Latent Root" es poden considerar tots aquells eixos amb un valor propi superior a 1, és a dir, que tenen més inèrcia que qualsevol de les variables originals.

```r
library("factoextra")
fviz_screeplot(res, ncp=8, barfill=c(rep(1,3),rep(2,5)))
```



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"))
```



En aquest cas, tot i que un 60,61% de la variabilitat ve explicada pel primer eix, s'ha decidit agafar els tres primers per tenir més marge en els exemples. Així doncs, amb aquests tres eixos tenim un 82,68% de la variabilitat de les dades.

També podem veure quines variables han tingut un major impacte alhora de determinar els eixos, això ve dictat pel factor "contribution". Pel primer eix, aquestes variables són: *displ*, *cylinders*, *weight* i *mpg*, i les de menor impacte són: $model_year$, *accel*.

```r
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
```

Les variables més ben representades tenen un major valor en el camp "cos2", pel primer eix són: *displ*, *cylinders*, *weight* i *mpg*, i les menys ben representades: $model_year$, *accel* i *origin*. És important veure que en aquest cas les variables més ben representades coincideixen amb les que han tingui un major impacte per elegir l'eix factorial, però no té perquè ser sempre així (CORRECTE???).

```r
res[["var"]][["cos2"]]
```

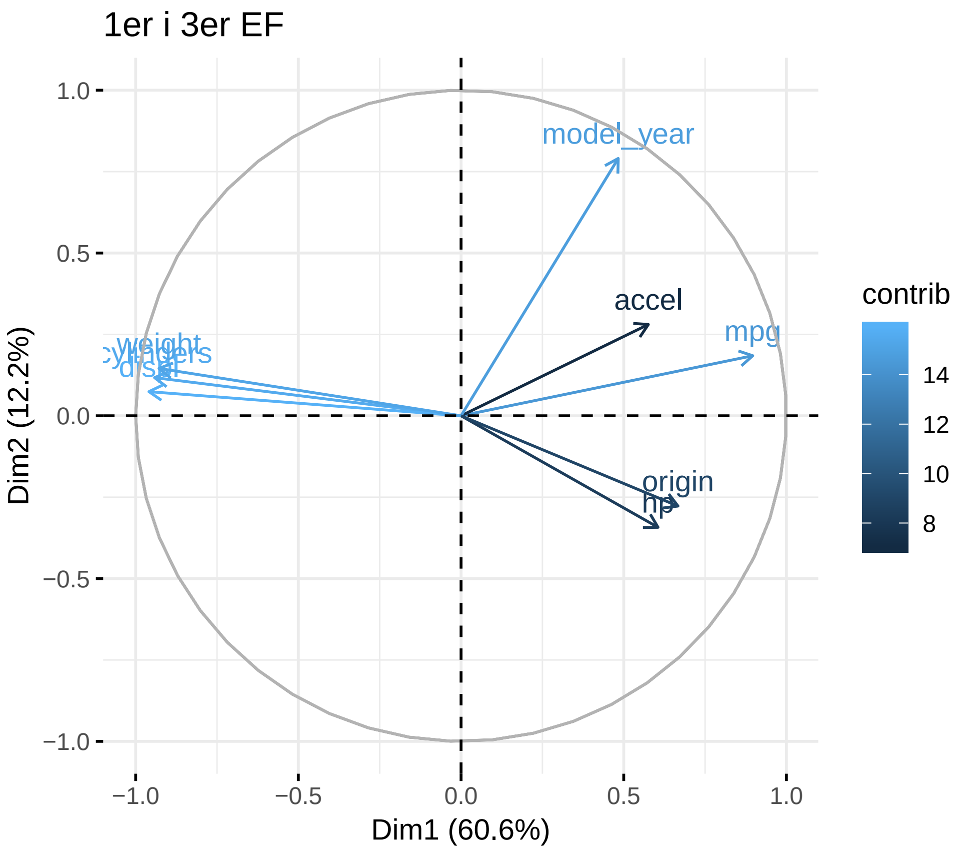
```
##                Dim.1       Dim.2        Dim.3        Dim.4       Dim.5
## mpg        0.8038293 0.034030839 0.0319309812 0.0075520997 0.005488544
## cylinders  0.8874506 0.013735380 0.0029484080 0.0003079089 0.028137148
## displ      0.9202143 0.005497682 0.0001174269 0.0124702478 0.024795631
## hp         0.3664957 0.117386898 0.1249396502 0.3318089175 0.058766317
## weight     0.8619893 0.021133861 0.0051624502 0.0010182685 0.068493376
## accel      0.3314037 0.078677447 0.3759182141 0.1700802604 0.040560111
## model_year 0.2334015 0.624861960 0.0360242341 0.0725829730 0.014356962
## origin     0.4445489 0.076856712 0.2160063902 0.0620606815 0.199621992
```

Ara per veure millor la representació de les variables sobre els eixos factorials, podem plasmar-les sobre en una cirfunferència de radi u, on els eixos de coordenades corresponen als dos eixos factorials amb més inèrica i la tonalitat de blau al nivell de contribució.

```r
#Coordenades
res[["var"]][["cor"]]
```

```
##                 Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
## mpg         0.8965653  0.18447449 -0.17869242  0.08690282 -0.07408471
## cylinders  -0.9420460  0.11719804 -0.05429924  0.01754733  0.16774131
## displ      -0.9592780  0.07414635 -0.01083637  0.11167026  0.15746629
## hp          0.6053889 -0.34261771  0.35346803  0.57602857  0.24241765
## weight     -0.9284338  0.14537490  0.07185019  0.03191032  0.26171239
## accel       0.5756767  0.28049500  0.61312170 -0.41240788  0.20139541
## model_year  0.4831165  0.79048211 -0.18980051  0.26941227  0.11982054
## origin      0.6667450 -0.27723043 -0.46476488 -0.24911981  0.44679077
```

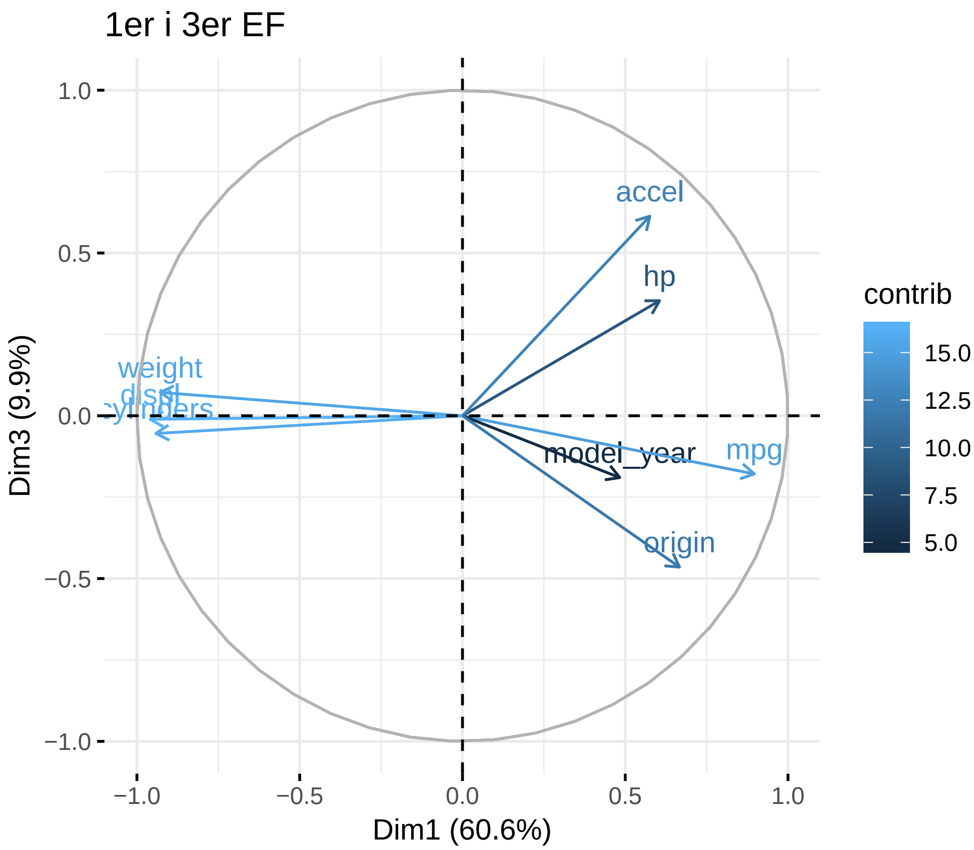
```r
#Gràfic
fviz_pca_var(res, axes = c(1, 2), col.var="contrib", title="1er i 3er EF")
```



En el primer eix es mostren els vehicles més pesats (weight) que solen tenir una major cilindrada (cylinder) i que tendeixen a consumir més combustible per milla recorreguda (mpg). També es pot veure com la variable "disp" ens diu que els cotxes menys pes i cilindrada necessiten menys espai "displacement" per frenar. El segon eix ens diu que a mesura que augmenta l'any del model, hi ha tendència a disminuir el pes i cilindrada.

Com que les variables *hp*, *accel* i *origin* no estan ben representades en cap dels dos primers eixos, però si ho estan millor en el tercer eix. Es torna a dibuixar la circumferència però aquesta vegada amb el primer i tercer eix.

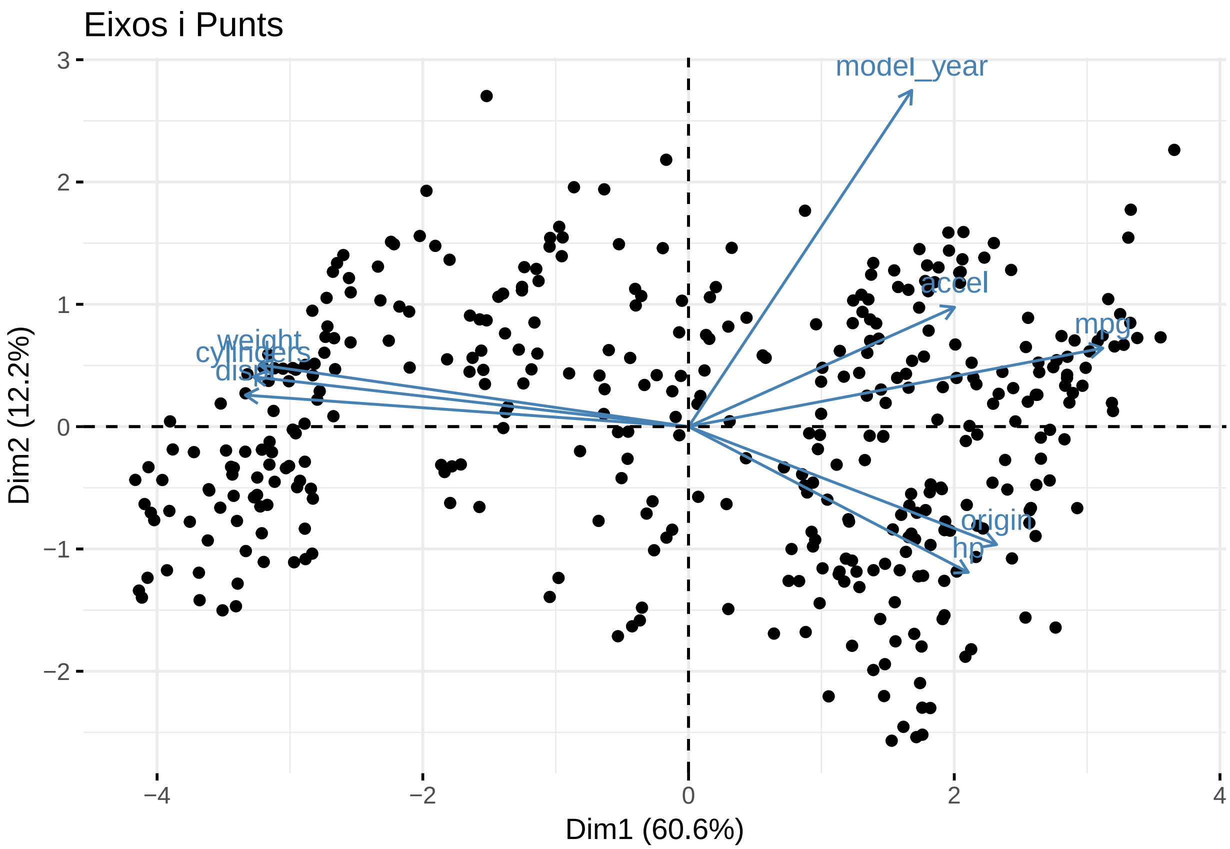
```
#Gràfic
fviz_pca_var(res, axes = c(1, 3), col.var="contrib", title="1er i 3er EF")
```



En el tercer eix es pot intuir que a mesura que augmenta la poténcia “hp”, augmenta l’acceleració “accel”.

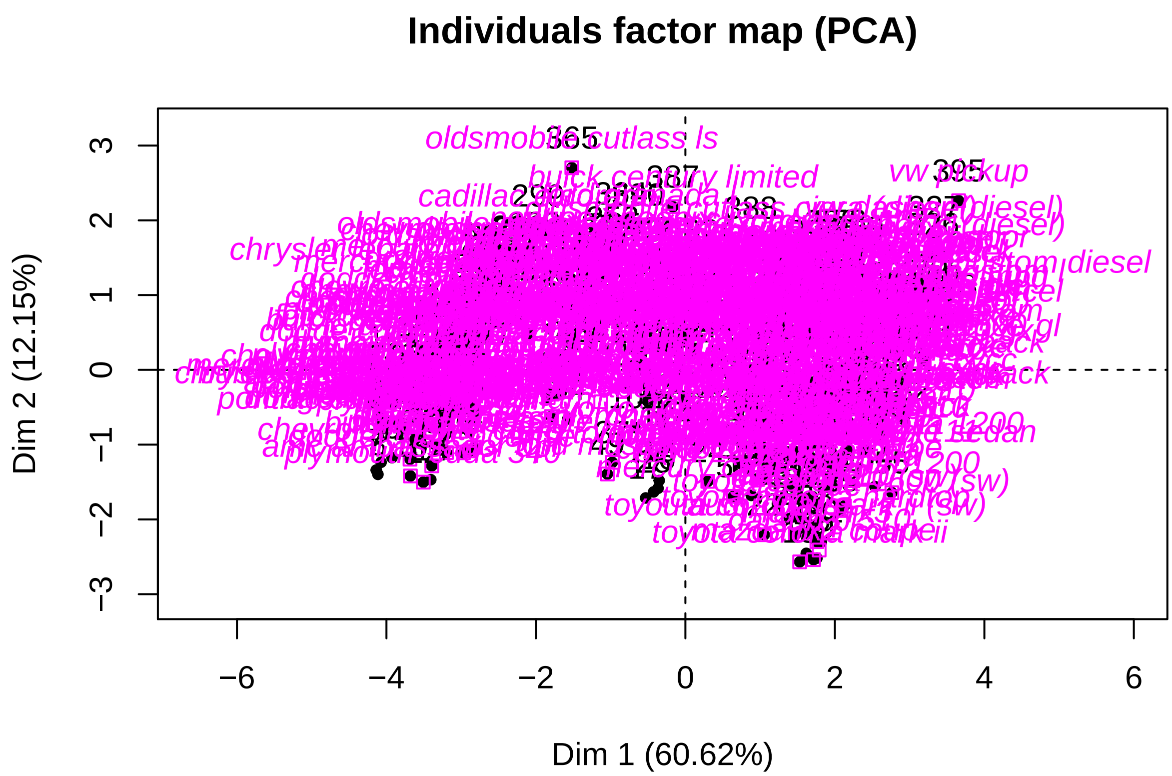
Ara toca representar les dades sobre els eixos factorials

```
fviz_pca_biplot(res, axes = c(1, 2), geom="point", title="Eixos i Punts")
```



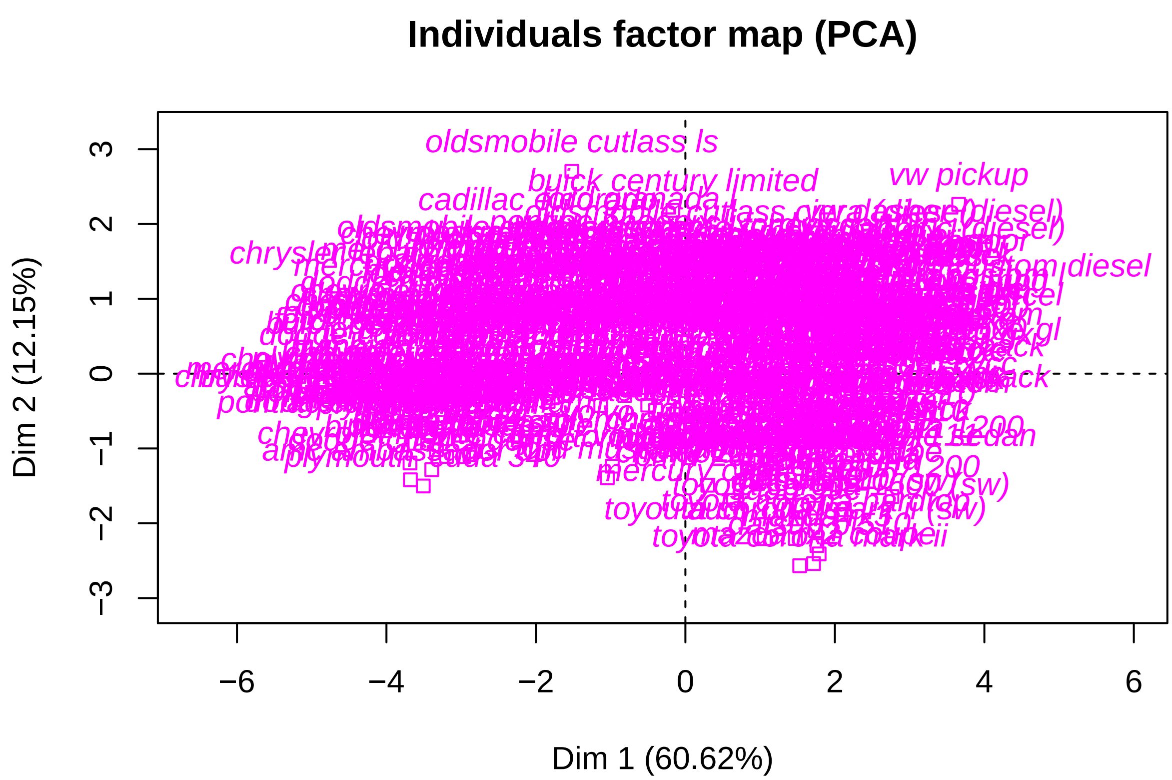
Ara projectarem la variable car_name que és suplementaria (no ha influit en els calculs dels eixos factorials).

```
remove(dades.PCA)
dades.PCA = dades[, c("mpg", "cylinders", "displ", "hp", "weight", "accel", "model_year",
                      "origin", "car_name")]
res<-PCA(dades.PCA , scale.unit=TRUE, ncp=5, quali.sup=c(9:9), graph = FALSE)
plot.PCA(res, axes=c(1, 2), choix="ind", habillage="none", col.ind="black",
            col.ind.sup="blue", col.quali="magenta", label=c("ind", "ind.sup", "quali"),new.plot=TRUE)
```



Com es pot veure, al lectura de les dades és impossible degut a la gran quantitat d'aquestes que s'estan representant. Per això mostrarem 4 (FALTA!!!)

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



Analisi predictiu

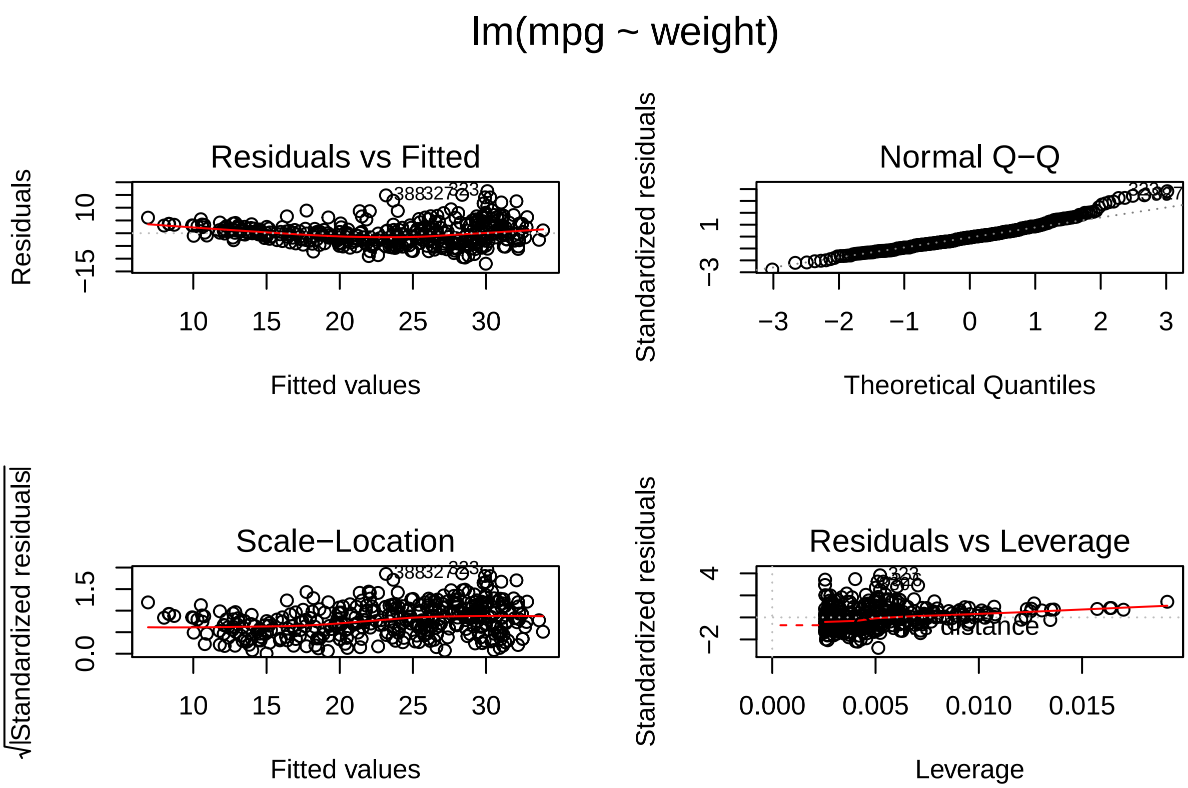
Per fer aquest anàlisi primer farem una regressió simple on s'intentarà preedir el valor de la variable MPG a partir de la variable *weight*

Regressió simple

Per aquest models de regressió lineal s'han de complir les següents propietats 1. Independència, s'interpreta que la observacions de la mostra s'han extret amb independencia entre elles. 2. Linealitat, s'assumeix que la variable MPG es pot modelar linealment a partir de la variable *weigth*. 3. Normalitat en els errors, on els errors segueixen una distribució normal. 4. Homocedasticitat, on la variancia és constant.

Per comprobar el tercer i quart punt es mostren les gràfiques.

```
res <- lm(mpg ~ weight, data = dades)
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(res)
```




```
par(oldpar)
```

El gràfic "Residuals vs Fitted" mostra la homocedasticitat (punt 4) mentre que el gràfic "Normal Q-Q" mostra la normalitat dels errors (Punt 3).

Per assegurar que es compleix l'homocedasticitat, es pot aplicar un test *Breuch Pagane* on la hipòtesi nul · la és l'homocedasticitat.

```
install.packages("lmtest")
library(lmtest)
bptest(mpg~weight,data=dades)
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  mpg ~ weight
## BP = 22.456, df = 1, p-value = 2.15e-06
```

Com que el p-value del test és molt inferior al nivell de significació, amb un nivell de confiança superior al 99% es pot rebutjar la hipòtesi nul · la i considerar que no hi ha homocedasticitat (hi ha hterocedasticitat).

Per comprobar la constància en els errors, es pot aplicar un test de $Shapiro\ Wilk$ i un altre d' $Anderson - Daling$ on en tots dos casos, la hipòtesi nul · la representa normalitat en els residus.

```r
shapiro.test(residuals(res))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(res)
## W = 0.96938, p-value = 2.525e-07
```

```r
install.packages("nortest")
library(nortest)
ad.test(residuals(res))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(res)
## A = 2.6013, p-value = 1.428e-06
```

En el test de $Shapiro\ Wilk$ s'ha obtingut un p-value de 2.525e-7 i en el test d'$Anderson - Daling$ un p-value de 1.428e-6, i per tant en tots dos casos podem descartar la hipòtesi nul · la amb un nivell de confiança superior al 99% i dir que els residus no segueixen una distribució normal.

Com que hem demostrat que en aquest cas no es compleixen el 3er i 4rt punt anunciats, la regressió que s'obtindria seria de baixa qualitat i per aquest motiu no es continua.