

# Practical Exercise

## 1. OBJECTIUS

Veurem com utilitzar una eina de reducció de dades de les que hem vist a classe i com fer-ne la seva visualització: l'anàlisi de components principals (PCA).

Recordeu carregar les llibreries necessàries com sempre.

## 2. PART 1. Data Processing II

Per aquesta primera part, farem servir el *dataframe* de R: *mtcars*, que vam utilitzar també en un seminari anterior. El *dataframe* conté 32 observacions i 11 variables sobre cotxes. Torneu-vos a familiaritzar amb el *dataframe* (`str(mtcars)`).

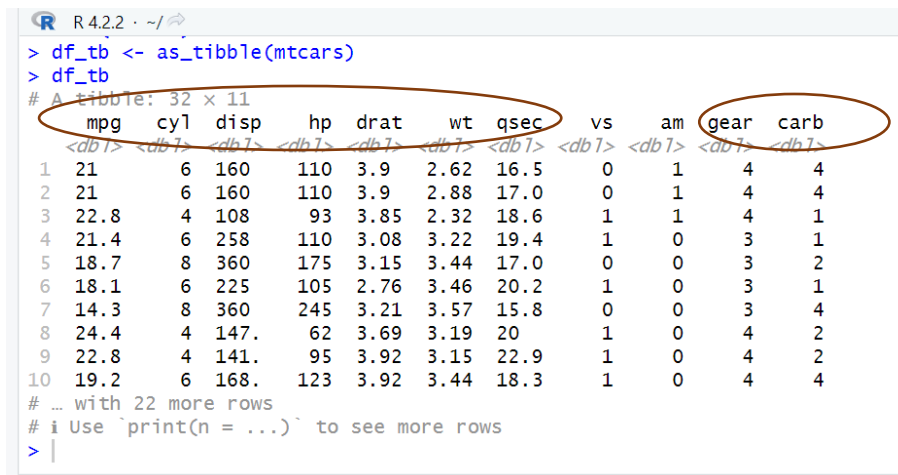
**1.- Feu una visualització reduint la dimensionalitat amb un PCA. Recordem que el PCA funciona bé amb dades numèriques, però no funciona amb dades categòriques. Per tant:**

- a) Feu primer una *tibble* que anomenarem *new2* on inclourem totes les variables de *mtcars* i les seves respectives observacions, excloent les variables que no ens van bé per fer el PCA

Excloem les variables 'vs' i 'am' del *dataframe* original (és a dir, ens quedem amb totes les variables excepte aquestes dues). 'vs' i 'am' són variables lògiques que prenen dos valors 0 i 1 per diferenciar dues categories, respectivament:

```
> df_tb <- as_tibble(mtcars)
```

```
> df_tb
```



```
R 4.2.2 · ~/
> df_tb <- as_tibble(mtcars)
> df_tb
# A tibble: 32 × 11
   mpg   cyl  disp    hp  drat    wt  qsec    vs  am  gear  carb
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1    21     6   160   110   3.9   2.62  16.5     0     1     4     4
2    21     6   160   110   3.9   2.88  17.0     0     1     4     4
3   22.8     4   108    93   3.85   2.32  18.6     1     1     4     1
4   21.4     6   258   110   3.08   3.22  19.4     1     0     3     1
5   18.7     8   360   175   3.15   3.44  17.0     0     0     3     2
6   18.1     6   225   105   2.76   3.46  20.2     1     0     3     1
7   14.3     8   360   245   3.21   3.57  15.8     0     0     3     4
8   24.4     4   147    62   3.69   3.19   20      1     0     4     2
9   22.8     4   141    95   3.92   3.15  22.9     1     0     4     2
10  19.2     6   168   123   3.92   3.44  18.3     1     0     4     4
# ... with 22 more rows
# i Use `print(n = ...)` to see more rows
> |
```

Per tant ens quedem amb les columnes de la 1 a la 7, la 11 i la 12. Una possibilitat seria:

```
> new2 <- df_tb[, c(1:7,10,11)]
```

- b) Assigneu a una nova variable anomenada *mtcars\_pca* la matriu resultant d'aplicar la funció `prcomp()` vista en la primera part de la classe d'avui. Centreu les variables a zero i escaleu-les per a que tinguin variances igual a 1.

**Nota: Hem vist com fer-ho a les diapositives d'avui però recordeu que si voleu obtenir més ajuda de com utilitzar els arguments, podeu cridar l'ajuda amb ? prcomp.**

Se us demana que centreu les variables a zero, per tant utilitzareu l'argument 'center=TRUE' que hem vist a les diapositives d'avui i escaleu-les per a que tinguin variança igual a 1, fent us de l'argument 'scale.=TRUE'.

**> mtcars\_pca<-prcomp(new2, center=TRUE, scale.=TRUE)**

```
> mtcars_pca
Standard deviations (1, ..., p=9):
[1] 2.3782219 1.4429485 0.7100809 0.5148082 0.4279704 0.3518426 0.3241326
[8] 0.2418962 0.1489644

Rotation (n x k) = (9 x 9):
      PC1      PC2      PC3      PC4      PC5      PC6
mpg -0.3931477 0.02753861 -0.22119309 -0.006126378 -0.3207620 0.72015586
cyl  0.4025537 0.01570975 -0.25231615 0.040700251 0.1171397 0.22432550
disp 0.3973528 -0.08888469 -0.07825139 0.339493732 -0.4867849 -0.01967516
hp   0.3670814 0.26941371 -0.01721159 0.068300993 -0.2947317 0.35394225
drat -0.3118165 0.34165268 0.14995507 0.845658485 0.1619259 -0.01536794
wt   0.3734771 -0.17194306 0.45373418 0.191260029 -0.1874822 -0.08377237
qsec -0.2243508 -0.48404435 0.62812782 -0.030329127 -0.1482495 0.25752940
gear -0.2094749 0.55078264 0.20658376 -0.282381831 -0.5624860 -0.32298239
carb 0.2445807 0.48431310 0.46412069 -0.214492216 0.3997820 0.35706914
      PC7      PC8      PC9
mpg -0.38138068 -0.12465987 0.11492862
cyl -0.15893251 0.81032177 0.16266295
disp -0.18233095 -0.06416707 -0.66190812
hp   0.69620751 -0.16573993 0.25177306
drat 0.04767957 0.13505066 0.03809096
wt   -0.42777608 -0.19839375 0.56918844
qsec 0.27622581 0.35613350 -0.16873731
gear -0.08555707 0.31636479 0.04719694
carb -0.20604210 -0.10832772 -0.32045892
>
```

**c) Quantes components principals obteniu? I com és la proporció de variança de cada component? Amb quantes us quedaríeu per tenir una bona representació de les vostres dades?**

Fem summary, com hem vist a les diapositives per veure la variança de cada component

```
> summary(mtcars_pca)
Importance of components:
      PC1      PC2      PC3      PC4      PC5      PC6      PC7
Standard deviation 2.3782 1.4429 0.71008 0.51481 0.42797 0.35184 0.32413
Proportion of variance 0.6284 0.2313 0.05602 0.02945 0.02035 0.01375 0.01167
Cumulative Proportion 0.6284 0.8598 0.91581 0.94525 0.96560 0.97936 0.99103
      PC8      PC9
Standard deviation 0.2419 0.14896
Proportion of variance 0.0065 0.00247
Cumulative Proportion 0.9975 1.00000
> |
```

Obtenim 9 components principals, anomenades PC1-PC9. Per tant:

- PC1 correspon al 63% de la variança total. És a dir, gairebé dos terços de la informació del conjunt de dades (9 variables) pot ser encapsulada només per aquesta component principal.
- PC2 correspon al 23% de la variança total.

Per tant, coneixent la posició d'una mostra en relació amb (només) les components PC1 i PC2, podeu obtenir una visió molt precisa de la seva ubicació en relació amb altres

mostres. Això és degut a que només PC1 i PC2 poden explicar gairebé el 86% de la varianza.

Si feu `str(mtcars_pca)`:

```
> str(mtcars_pca)
List of 5
 $ sdev      : num [1:9] 2.378 1.443 0.71 0.515 0.428 ...
 $ rotation: num [1:9, 1:9] -0.393 0.403 0.397 0.367 -0.312 ...
 .. attr(*, "dimnames")=List of 2
 .. ..$ : chr [1:9] "mpg" "cyl" "disp" "hp" ...
 .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
 $ center   : Named num [1:9] 20.09 6.19 230.72 146.69 3.6 ...
 .. attr(*, "names")= chr [1:9] "mpg" "cyl" "disp" "hp" ...
 $ scale    : Named num [1:9] 6.027 1.786 123.939 68.563 0.535 ...
 .. attr(*, "names")= chr [1:9] "mpg" "cyl" "disp" "hp" ...
 $ x        : num [1:32, 1:9] -0.664 -0.637 -2.3 -0.215 1.587 ...
 .. attr(*, "dimnames")=List of 2
 .. ..$ : NULL
 .. ..$ : chr [1:9] "PC1" "PC2" "PC3" "PC4" ...
 - attr(*, "class")= chr "prcomp"
> |
```

El vostre objecte PCA conté la informació següent:

- El punt central (`$ center`)
- L'escala (`$ scale`)
- La desviació estàndard (`$ sdev`) de cada component principal
- La relació (correlació o anticorrelació, etc) entre les variables inicials i les components principals (`$ rotation`)
- Els valors de cada mostra en termes de les components principals (`$ x`)

Però amb la informació de `summary (mtcars_pca)` ja en teniu prou

**d) Ja teniu tot preparat i és hora de visualitzar el vostre PCA. Per això feu servir les funcions que hem vist a les diapositives de la classe d'avui.**

**d.1) Instal·leu i carregueu les llibreries que necessiteu (si no ho heu fet abans).**

Per instal·lar-ho feu:

```
> install.packages("devtools")
```

```
> install.packages("factoextra")
```

Carregueu les llibreries:

```
> library(devtools)
```

```
> library("factoextra") # Ojo, si us dona error carregueu la llibreria ggplot2 o tidyverse
```

**d.2) Quan ja teniu el paquet i les llibreries que necessiteu, creeu grups segons la procedència dels cotxes. Els dividirem en una de les tres categories, segons la procedència dels vehicles: 1) nord-americans, 2) japonesos i 3) europeus. Per això us donem la comanda:**

```
> mtcars.country <- c(rep("Japan", 3), rep("US",4), rep("Europe", 7),rep("US",3),  
"Europe", rep("Japan", 3), rep("US",4), rep("Europe", 3), "US", rep("Europe", 3))
```

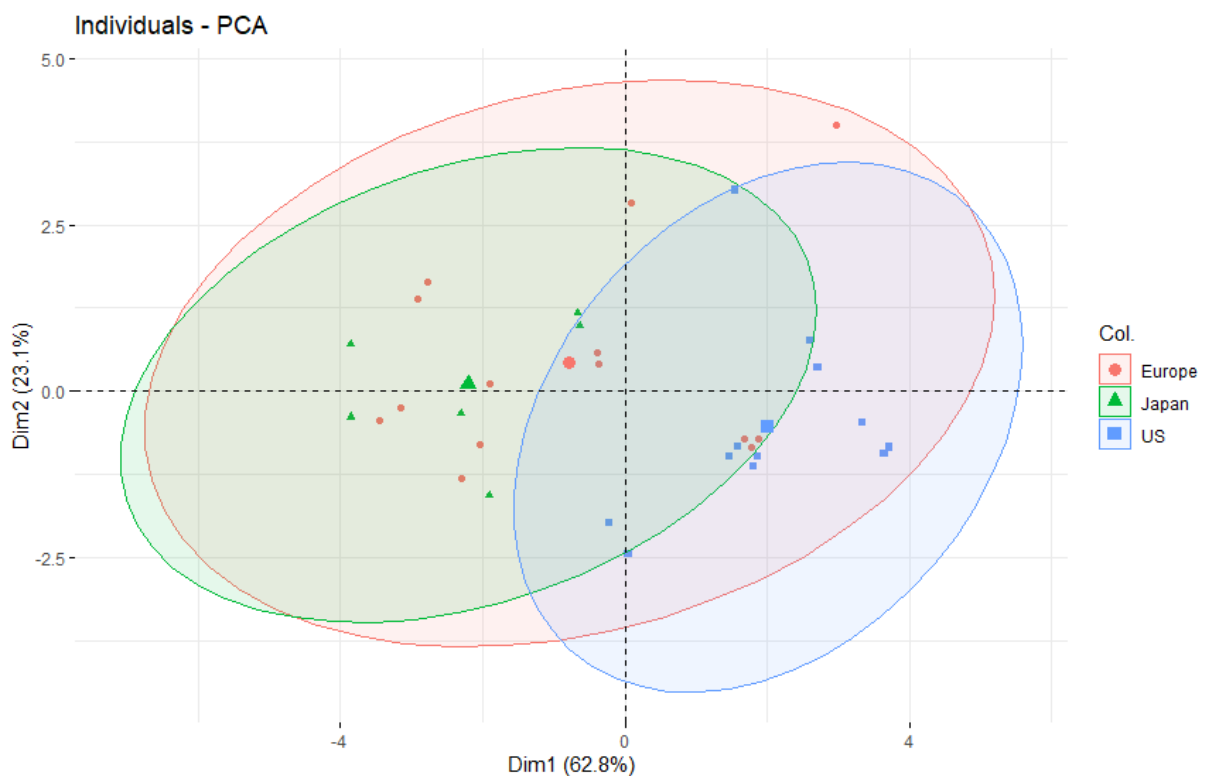
d.3) Fent ús de l'ajuda i de l'exemple que se us ha posat a les diapositives del principi de la classe. Utilitzeu la funció `fviz_pca_ind()` per aquest exemple:

- utilitzeu un `col.ind` (color) que vingui donat pel país del cotxe (`mtcars.country`)
- afegiu el·lipses i un títol de la llegenda "Groups" (ja que aquesta mostrarà els grups creats segons l'origen del cotxe).

Quines conclusions extraieu?

```
> fviz_pca_ind (mtcars_pca, geom.ind="point", col.ind=mtcars.country, addEllipses = TRUE, legend.title="Groups")
```

**Exemple de conclusions:** Els cotxes americans (US) formen un clúster més diferenciats a la dreta. Els cotxes europeus estan més al centre i tenen una el·lipse major, pel que sembla que estan menys agrupats (o més dispersos) que els dels altres dos grups. A més, els cotxes Japonesos omplen bona part de la part esquerra del clúster d'Europa.



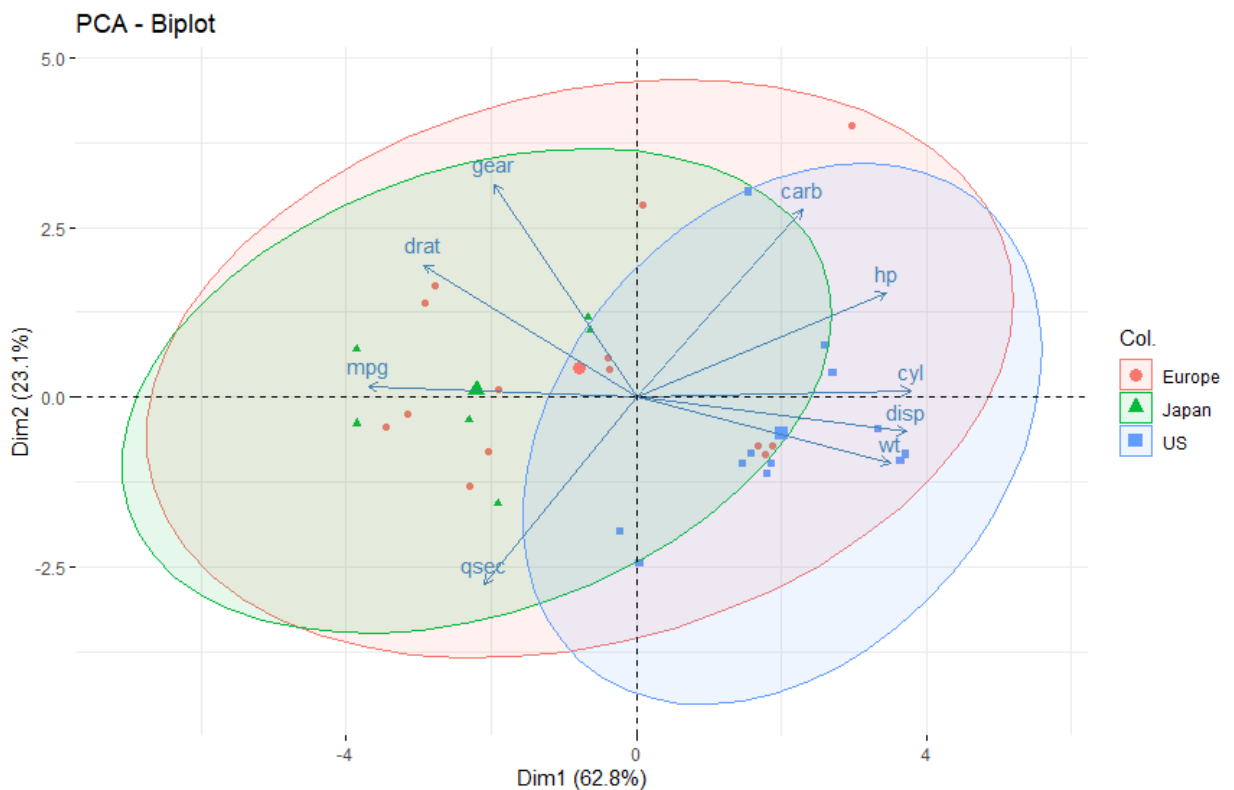
e) Feu un *biplo*t, que inclogui la posició de cada mostra en termes de PC1 i PC2. El *biplo*t, a més de veure les variables individualment, us mostrarà com es relacionen les variables inicials. Per fer dit *biplo*t, utilitzeu la funció `fviz_pca_biplo`t de manera similar a com heu utilitzat el `fviz_pca_ind` de

l'exercici anterior però afegiu una fletxa i el text referent a les variables inicials tot fent ús de l'argument `geom.var`. Podeu extreure conclusions noves?

Com sempre podeu fer ús de l'ajuda `?fviz_pca_biplot`

**NOTA:** Un *biplot* és un tipus de gràfic que us permet visualitzar la relació de les mostres entre sí dins del PCA. Ens ajuda a veure quines mostres són similars i quines són diferents, i revela simultàniament com cada variable contribueix a cada component principal.

```
> fviz_pca_biplot (mtcars_pca, geom.ind="point", geom.var = c("arrow", "text"),
col.ind=mtcars.country, addEllipses = TRUE, legend.title="Groups")
```

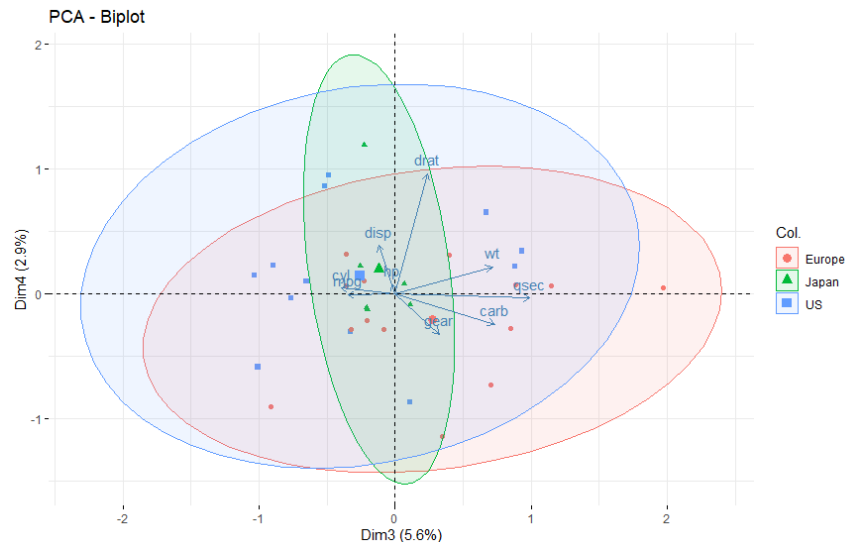


**Exemple de conclusions:** Observant els eixos, veieu que els cotxes americans (US) es caracteritzen per tenir valors elevats de *cyl*, *disp* i *wt*. Els cotxes japonesos, en canvi, es caracteritzen per tenir un *mpg* elevat. Finalment els Europeus tenen més característiques, pel que podem entendre perquè no estaven tant agrupats com els de Japó o US.

f) Per descomptat, teniu disponibles moltes components principals, cadascuna de les quals es correlaciona de manera diferent amb les variables originals. També podeu demanar a `fviz_pca_biplot` que dibuixi aquestes altres components fent ús de l'argument `axes`.

Fent ús de l'argument `axes`, feu ara la visualització segons la tercera i quarta components principals (PC3 i PC4). Creieu que podeu veure molta informació? Per què?

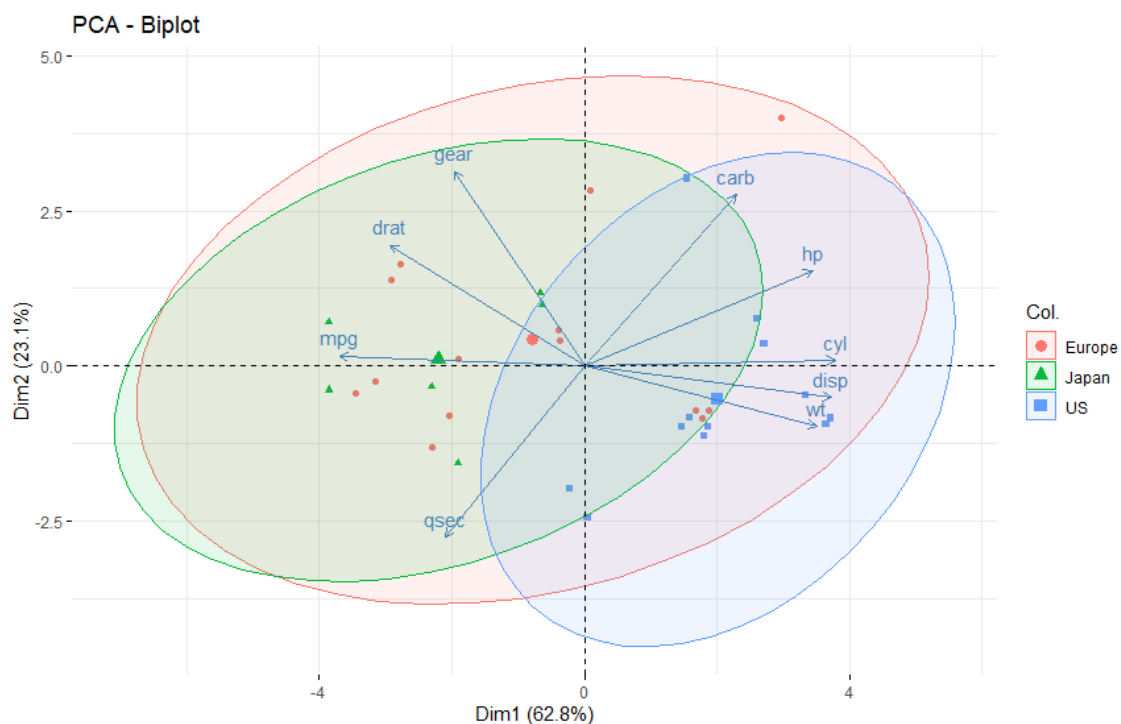
```
> fviz_pca_biplot (mtcars_pca, axes=c(3,4), geom.ind="point", geom.var = c("arrow",
"text"), col.ind=mtcars.country, addEllipses = TRUE, legend.title="Groups")
```



Com podeu observar ens han canviat els eixos (ara es mostren PC3 i PC4). Aquí no hi veieu molt, però això no és massa sorprenent. PC3 i PC4 expliquen percentatges molt reduïts de la variança total, de manera que seria sorprenent que trobéssiu que eren molt informatius i separessin els grups o revelessin patrons aparents.

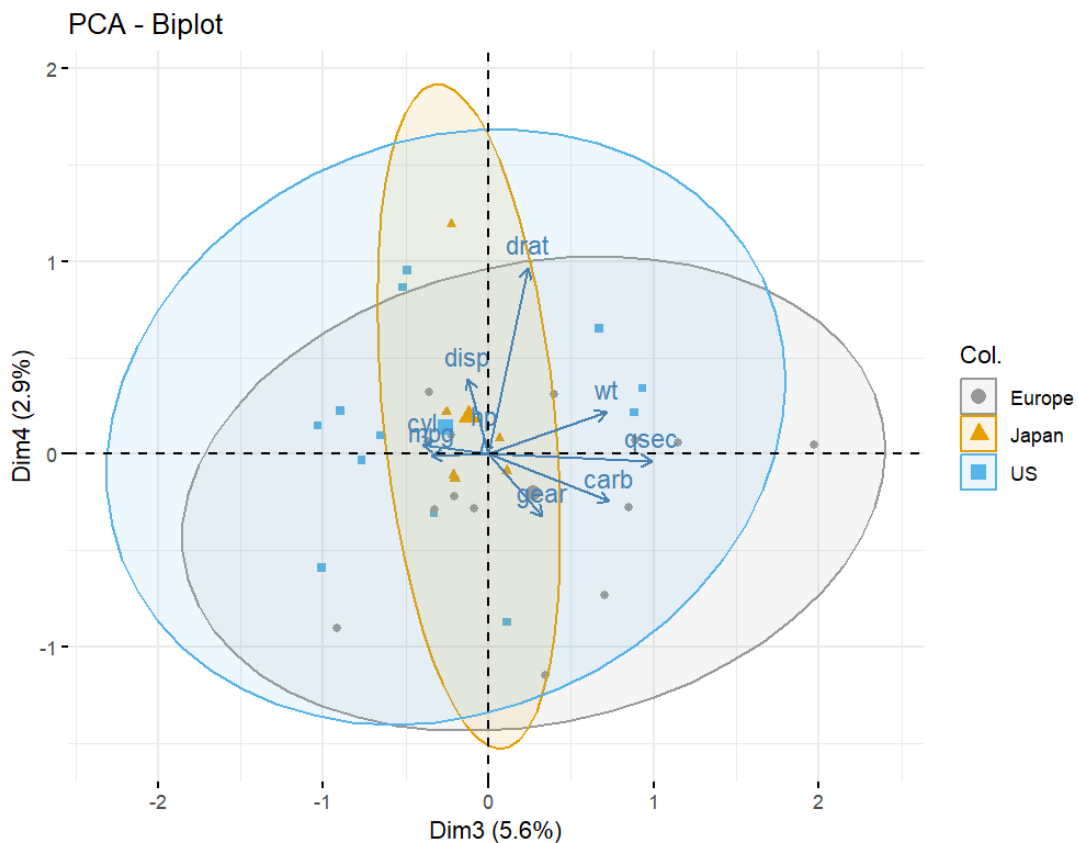
**g) I llavors quina informació ens ha donat aquest anàlisi PCA? Podeu fer un resum del que heu trobat**

**RESUM possible:** Tot realitzant un anàlisi PCA mitjançant el conjunt de dades mtcars, podem veure una clara separació entre els cotxes americans i japonesos al llarg d'una component principal que està estretament correlacionada amb *cyl*, *disp*, *wt* i *mpg*. Això ens proporciona algunes pistes per a futurs anàlisis; **si intentéssim construir un model de classificació per identificar l'origen d'un cotxe, aquestes variables podrien ser útils.**



h) Intenteu canviar la paleta de color, com teniu 3 grups (Europe, Japan, US) necessiteu un color per cada. Això ho podeu fer posant `palette = c("#999999", "#E69F00", "#56B4E9")` o alguna altra que us agradi

```
>fviz_pca_biplot(mtcars_pca, axes=c(3,4), geom.ind="point", geom.var = c("arrow", "text"),
col.ind=mtcars.country, addEllipses = TRUE, legend.title="Groups",
palette = c("#999999", "#E69F00", "#56B4E9"))
```



Judit Chamorro Servent  
Bellaterra, Març 2025