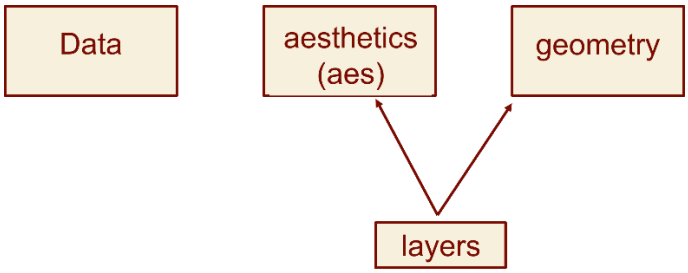
SEMINARI 1. *R* / *ggplot*. Introducció

1. OBJECTIUS

Aquest seminari serveix per familiaritzar-se amb l'ús de ggplot2 i els seus passos successius.



Si no l'heu instal·lat encara, instal·leu i carregueu la llibreria tidyverse.

```
> install.packages("tidyverse")
> library(tidyverse)
```

NOTA: En aquest seminari, l'únic objectiu és familiaritzar-se amb l'ús i l'estructura de `ggplot` i veure les diferents "point shapes" segons el tipus de dades que tenim. Treballarem només amb `geom_point()`. Per tant, les visualitzacions que farem en aquest seminari NO seran les més adequades pel tipus de dades, però això ho anirem veient amb els següents seminaris, on, un cop ja familiaritzats amb les eines, sí que farem un especial èmfasis en aquest segon aspecte.

2. PART 1. Com és el nostre dataset? Quin tipus de variables hi tenim?

El conjunt de dades *mtcars* conté informació de 32 cotxes. És un conjunt de dades petit que conté una varietat de variables contínues i categòriques i ens permetrà familiaritzar-nos amb ggplot2. Podeu utilitzar `str()` per explorar la estructura d'aquest dataset.

Si obriu R de nou, primer de tot recordeu que heu de tornar a carregar la llibreria tidyverse.

```
> library (tidyverse)
```

Podeu utilitzar `str()` per explorar la estructura d'aquest *dataset*.

```
> 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 ...
```

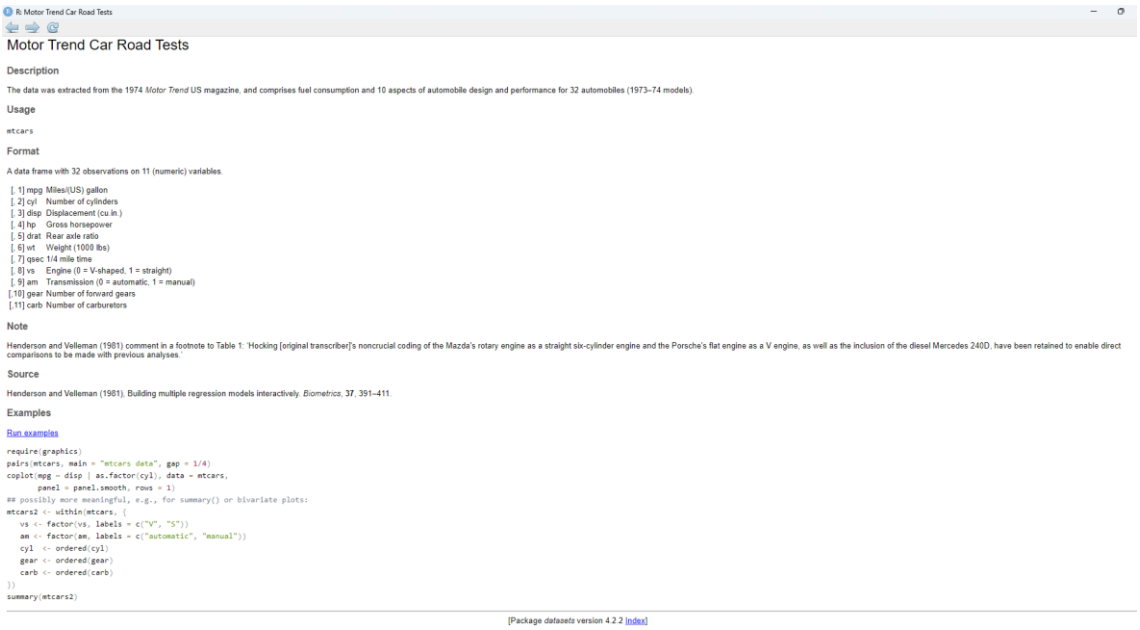
I per saber la definició de cada variable utilitzeu

```
> ?mtcars
```

O :

```
> help (mtcars)
```

Se us obrirà una pantalla nova:



R: Motor Trend Car Road Tests

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 (numeric) variables.

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

Note

Henderson and Velleman (1981) comment in a footnote to Table 1: 'Hocking [original transcriber]'s noncrucial coding of the Mazda's rotary engine as a straight six-cylinder engine and the Porsche's flat engine as a V engine, as well as the inclusion of the diesel Mercedes 240D, have been retained to enable direct comparisons to be made with previous analyses.'

Source

Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, **37**, 391–411.

Examples

Run examples

```
require(graphics)
pairs(mtcars, main = "mtcars data", gap = 1/4)
coplot(mpg ~ disp | as.factor(cyl), data = mtcars,
       panel = panel.smooth, rows = 1)
## possibly more meaningful, e.g., for summary() or bivariate plots:
mtcars2 <- within(mtcars, {
   vs <- factor(vs, labels = c("V", "S"))
   am <- factor(am, labels = c("automatic", "manual"))
   cyl  <- ordered(cyl)
   gear <- ordered(gear)
   carb <- ordered(carb)
})
summary(mtcars2)
```

[Package *datasets* version 4.2.2 Index]

EXERCICIS:

1.- Utilitzeu *ggplot* per dibuixar una gràfica on l'eix x correspongui a la variable 'cyl' (cilindres) i l'eix y a la variable 'mpg' (km de galó). Utilitzeu `geom_point()`.

2.- Utilitzeu la funció ggplot, però ara categoritzeu la variable 'cyl' ordinal. Per això utilitzeu la funció factor. Quina informació podeu extreure'n d'aquesta gràfica?

NOTA: Primer escriviu `?factor` per a que R us digui com especificar que 'cyl' és una variable ordinal que ens està diferenciant en tres grups/nivells de cotxes (els que tenen 4, 6 o 8 cilindres). És el que abans quan veiem els tipus de variables em anomenat factors.

3.- Afegiu un color segons els cilindres que tingui el cotxe. Ens aporta alguna informació nova? Per què?

4.- Seguint amb la mateixa gràfica (on l'eix x correspongui a la variable 'cyl' i l'eix y a la variable 'mpg'). Afegiu ara un color al motor del cotxe (*engine*), per això primer recordeu

mirar com és la variable 'vs' i feu els ajustos necessaris. Un cop tenim la gràfica, ens aporta alguna informació nova respecte la gràfica de l'exercici 2? Per què? Quina és aquesta informació?

Utilitzeu `?scale_x_discrete` , `?scale_x_continuous` i `?scale_color_discrete` per posar el nom als eixos i a la llegenda de colors amb *scale*

5.- Afegiu ara un color a la variable *Displacement* de cada cotxe i poseu les llegendes adients. És fàcil de veure el que ens aporta aquesta nova informació? Per què? Podeu millorar la visualització de la gràfica d'una manera simple? Quina informació diríeu que en podeu extreure al veure les dades gràficament?

6.- Seguint amb la mateixa gràfica (on l'eix x correspongui a la variable 'cyl' i l'eix y a la variable 'mpg') de l'exercici 4. Intenteu utilitzar *shape* en `aes()` per posar una forma segons cada desplaçament Què creieu que passa? Podeu utilitzar *shape* amb alguna variable? Quina per exemple?

Ara que hem vist com d'important és primer familiaritzar-nos amb el nostre dataset i el tipus de variables que hi tenim per tal d'extreure'n la màxima informació de forma visual, veiem què volem mostrar a la part 2 del seminari.

3. PART 2. Què volem mostrar? Per què és important visualitzar les dades?

Anem a utilitzar el conjunt de dades `anscombe`, del que ja us han parlat a la classe de teoria. Escrivint `anscombe` a R podeu visualitzar-lo. També podeu explorar la seva estructura amb `str(anscombe)`. O accedir a la seva ajuda escrivint `?anscombe`

```
> str(anscombe)
'data.frame':   11 obs. of  8 variables:
 $ x1: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x2: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x3: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x4: num  8 8 8 8 8 8 8 19 8 8 ...
 $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
 $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26 ...
 $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
 $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.91 ...
>
```

Primer construirem 4 grups de datasets. Els anomenarem `g1data,...,g4data`. I cada un contindrà els valors (x_i,y_i), amb i=1,..4. Després en veure'm les seves respectives mitjanes i desviació estàndards, omplint la taula següent. Tot seguit plotejarem cada grup `g1data,...,g4data` per separat utilitzant `ggplot()` amb `geom_point()`. Què està passant?

	mean(gidata$xVal)	mean(gidata$yVal)	sd(gidata$xVal)	sd(gidata$yVal)
g1data				
g2data				
g3data				
g4data				

NOTA: `g1data=with(anscombe,data.frame(xVal=c(x1),yVal=c(y1)))` **us crearà el primer grup. Per fer les respectives mitjanes farem servir les comandes que hem posat en cada columna de la taula anterior on i=1,2,3,4 respectivament:**

```
> mean(g1data$xVal)
> mean(g1data$yVal)
> sd(g1data$xVal)
> sd(g1data$yVal)
```