

# SEMINARI 3.2. *Data Massaging* (Respostes)

## 1. OBJECTIUS

Aquest seminari introdueix al *Data Massaging*.

## 2. PART 1. *Data Massaging*

En aquesta primera part del seminari anem a fer alguns canvis en el dataset *starwars* que ja vam explorar en el seminari 3 de la llibreria tidyverse. Si heu obert R de nou, recordeu carregar la llibreria tidyverse.

Starwars és un dataset on cada fila és una observació i cada columna és una variable

```
> starwars
# A tibble: 87 x 14
  name          height mass hair_color skin_color eye_color birth_year sex gender homeworld species films vehicles starships
<chr>          <dbl> <dbl> <chr>    <chr>    <chr>    <dbl> <chr> <chr> <chr>    <chr> <list> <list> <list>
1 Luke Skywalker 172    77 blond    fair     blue     19    male masculine Tatooine Human <chr [5]> <chr [2]> <chr [2]>
2 C-3PO         167    75 <NA>    gold     yellow  112   none masculine Tatooine Droid  <chr [6]> <chr [0]> <chr [0]>
3 R2-D2         96     32 <NA>    white, blue red     33    none masculine Naboo   Droid  <chr [7]> <chr [0]> <chr [0]>
4 Darth Vader   202   136 none     white     yellow  41.9  male masculine Tatooine Human <chr [4]> <chr [0]> <chr [1]>
5 Leia Organa   150    49 brown    light    blue     19    female feminine Alderaan Human <chr [5]> <chr [1]> <chr [0]>
6 Owen Lars     178   120 brown, grey light    blue     52    male masculine Tatooine Human <chr [3]> <chr [0]> <chr [0]>
7 Beru Whitesun lars 165    75 brown    light    blue     47    female feminine Tatooine Human <chr [3]> <chr [0]> <chr [0]>
8 R5-D4         97     32 <NA>    white, red red     NA     none masculine Tatooine Droid  <chr [1]> <chr [0]> <chr [0]>
9 Biggs Darklighter 183    84 black     light    brown    24    male masculine Tatooine Human <chr [1]> <chr [0]> <chr [1]>
10 Obi-Wan Kenobi 182    77 auburn, white fair     blue-gray 57    male masculine Stewjon Human <chr [6]> <chr [1]> <chr [5]>
# ... with 77 more rows
> |
```

Primer ens familiaritzarem amb algunes funcions de la llibreria *dyplr* com:

- `select`: Seleccionar columnes
- `filter`: Filtrar
- `arrange`: Ordenar un conjunt de dades
- `group_by`: Agrupar per alguna variable
- `summarize/summarise`: Especificar algunes funcions d'agregats:
  - `n()` : comptar;
  - `sum()` : sumar variables numèriques;
  - `mean()` : la mitjana de variables numèriques entre altres
- `mutate`: modificar, transformar o agregar variables del conjunt de dades
- pipes `%>%`: combinar operacions

## EXERCICIS:

En primer lloc, carreguem la llibreria *dyplr*

```
> library(tidyverse)
```

```
> library(dplyr)
```

### 1.- Seleccionem només el nom i gènere del conjunt de dades *starwars*

Primer cridem les llibreries i després fem us de la funció `select()`, ja que estem seleccionant variables/columnes de la taula.



**select(.data, ...)**  
Extract columns as a table. Also **select\_if()**.  
*select(iris, Sepal.Length, Species)*

Use these helpers with **select()**,  
e.g. *select(iris, starts\_with("Sepal"))*

**contains(match)**      **num\_range(prefix, range)**    ;, e.g. mpg:cyl  
**ends\_with(match)**    **one\_of(...)**                               -, e.g. -Species  
**matches(match)**      **starts\_with(match)**

```
> select(starwars, name, gender)
# A tibble: 87 x 2
  name                gender
  <chr>              <chr>
1 Luke Skywalker    masculine
2 C-3PO              masculine
3 R2-D2              masculine
4 Darth Vader        masculine
5 Leia Organa        feminine
6 Owen Lars          masculine
7 Beru Whitesun lars feminine
8 R5-D4              masculine
9 Biggs Darklighter masculine
10 Obi-Wan Kenobi     masculine
# ... with 77 more rows
> |
```

## 2.- Seleccionen els personatges que són humans. Després seleccionen els que no

En la columna *species*, trobem algunes files que són "Human". Per tant, seleccionem les observacions/files amb la funció *filter* com hem vist a classe

## Manipulate Cases

### EXTRACT CASES

Row functions return a subset of rows as a new table.



**filter(.data, ...)** Extract rows that meet logical  
criteria. *filter(iris, Sepal.Length > 7)*

```
> filter(starwars, species=="Human")
> #or also:
> starwars %>% filter(species=="Human")
```

Al primer seminari vam veure que les cadenes de caràcters podien escriure's entre "Human" o 'Human'.

```
> starwars %>% filter(species=="Human")
# A tibble: 35 x 14
  name             height mass hair_color skin_color eye_color birth_year sex gender homeworld species films vehicles starships
  <chr>            <dbl> <dbl> <chr>    <chr>    <chr>    <dbl> <chr> <chr>    <chr>    <list>    <list>    <list>
1 Luke Skywalker   172    77 blond    fair     blue     19    male masculine Tatooine Human <chr [5]> <chr [2]> <chr [2]>
2 Darth Vader      202   136 none     white    yellow   41.9  male masculine Tatooine Human <chr [4]> <chr [0]> <chr [1]>
3 Leia Organa      150    49 brown    light    brown    19    female feminine Alderaan Human <chr [5]> <chr [1]> <chr [0]>
4 Owen Lars        178   120 brown, grey light    blue     52    male masculine Tatooine Human <chr [3]> <chr [0]> <chr [0]>
5 Beru Whitesun lars 165    75 brown    light    blue     47    female feminine Tatooine Human <chr [3]> <chr [0]> <chr [0]>
6 Biggs Darklighter 183    84 black    light    brown    24    male masculine Tatooine Human <chr [1]> <chr [0]> <chr [1]>
7 Obi-Wan Kenobi    182    77 auburn, white fair     blue-gray 57    male masculine Stewjon Human <chr [6]> <chr [1]> <chr [5]>
8 Anakin Skywalker  188    84 blond    fair     blue     41.9  male masculine Tatooine Human <chr [3]> <chr [2]> <chr [3]>
9 Wilhuff Tarkin    180    NA auburn, grey fair     blue     64    male masculine Eriadu Human <chr [2]> <chr [0]> <chr [0]>
10 Han Solo         180    80 brown    fair     brown    29    male masculine Corellia Human <chr [4]> <chr [0]> <chr [2]>
# ... with 25 more rows
> |
```

Per veure els que *no*, utilitzem la negació en R !

#### Logical and boolean operators to use with filter()

```
<      <=      is.na()  %in%      |      xor()
>      >=      !is.na()  !      &
See ?base::Logic and ?Comparison for help.
```

```
> filter(starwars, !species=="Human")
#or:
> starwars %>% filter(!species=="Human") #alternatively
#or:
> starwars %>% filter(species!="Human")
#or:
> filter(starwars, species!="Human")
```

```
> starwars %>% filter(!species=="Human")
# A tibble: 48 x 14
  name             height mass hair_color skin_color eye_color birth_year sex gender homeworld species films vehicles starships
  <chr>            <dbl> <dbl> <chr>    <chr>    <chr>    <dbl> <chr> <chr>    <chr>    <list>    <list>    <list>
1 C-3PO           167    75 <NA>    gold     yellow   112 none masculine Tatooine Droid <chr [6]> <chr [0]> <chr [0]>
2 R2-D2           96    32 <NA>    white, blue red     33 none masculine Naboo Droid <chr [7]> <chr [0]> <chr [0]>
3 R5-D4           97    32 <NA>    white, red  red     NA none masculine Tatooine Droid <chr [1]> <chr [0]> <chr [0]>
4 Chewbacca       228   112 brown    unknown    blue     200 male masculine Kashyyyk Wookiee <chr [5]> <chr [1]> <chr [2]>
5 Greedo          173    74 <NA>    green     black     44 male masculine Rodia Rodian <chr [1]> <chr [0]> <chr [0]>
6 Jabba Desilijic Tiure 175 1358 <NA>    green-tan, brown orange 600 hermaphroditic masculine Nal Hutta Hutt <chr [3]> <chr [0]> <chr [0]>
7 Yoda            66    17 white    green     brown    896 male masculine <NA> Yoda's species <chr [5]> <chr [0]> <chr [0]>
8 IG-88           200   140 none     metal     red     15 none masculine <NA> Droid <chr [1]> <chr [0]> <chr [0]>
9 Bossk           190   113 none     green     red     53 male masculine Trandoshia Trandoshan <chr [1]> <chr [0]> <chr [0]>
10 Ackbar          180    83 none     brown mottle orange 41 male masculine Mon Cala Mon Calamari <chr [2]> <chr [0]> <chr [0]>
# ... with 38 more rows
> |
```

### 3.- Seleccionem només el nom i gènere dels personatges amb gènere femení

- Pas a pas: Assignem primer a una variable temporal (per exemple de nom *dadesub*) la selecció feta en l'exercici 1. Després seleccionem del subconjunt *dadesub* les observacions/files que corresponen al gènere femení i assignem el subconjunt resultant a una altra variable (per exemple de nom *femenins*)
- Feu ús de les *pipes* que hem vist a classe, per tal de combinar les dues operacions fetes en l'apartat a. Assignem el subconjunt resultant a una variable (per exemple de nom *femeninsub*) i verifiquem que correspon al mateix subconjunt assignat a *femenins* en l'apartat a). **NOTA:** Sempre que puguem utilitzarem *pipes* com en aquest apartat, és a dir, no crearem variables temporals extres com em fet en l'apartat a).

a) Assignem a la variable temporal `dadesub` el subconjunt de dades resultant de la selecció feta en l'exercici 1. Després seleccionem les observacions/files amb la funció `filter` com hem vist a l'exercici 2. Fixem-nos que `filter` selecciona files/observacions que satisfan un cert criteri lògic (en aquest cas seria com pensar `if gender='feminine'`)

```
> dadesub <- select(starwars, name, gender)
> femenins <- filter(dadesub, gender=='feminine')
> femenins
```

```
> femenins <- filter(dadesub, gender=='feminine')
> femenins
# A tibble: 17 x 2
  name                gender
  <chr>              <chr>
1 Leia Organa        feminine
2 Beru Whitesun lars feminine
3 Mon Mothma         feminine
4 Shmi Skywalker     feminine
5 Ayla Secura        feminine
6 Adi Gallia         feminine
7 Cordé              feminine
8 Luminara Unduli    feminine
9 Barriss Offee      feminine
10 Dormé              feminine
11 Zam Wesell         feminine
12 Taun We            feminine
13 Jocasta Nu         feminine
14 R4-P17             feminine
15 Shaak Ti           feminine
16 Rey                feminine
17 Padmé Amidala      feminine
> |
```

b. Hem vist que podem seleccionar variables (columnes) i observacions (o files) fent us de *pipes* següent:

**dplyr** functions work with pipes and expect **tidy data**. In tidy data:



Each **variable** is in its own **column**

&



Each **observation**, or **case**, is in its own **row**



**pipes**

`x %>% f(y)`  
becomes `f(x, y)`

Per tant combinem la selecció i el filtratge utilitzant *pipes* de la següent manera:

```
> feminisub <- starwars %>% filter(gender=='feminine')
%>%select(name,gender)
> feminisub
```

En aquest cas, podem intercanviar el filtratge i la selecció obtenint el mateix resultat:

```
> starwars %>%select(name,gender)%>%filter(gender=='feminine')
```

```
> feminisub <- starwars %>% filter(gender=="feminine") %>% select(name, gender)
> feminisub
# A tibble: 17 x 2
  name                gender
  <chr>               <chr>
1 Leia Organa        feminine
2 Beru Whitesun lars feminine
3 Mon Mothma         feminine
4 Shmi Skywalker     feminine
5 Ayla Secura        feminine
6 Adi Gallia         feminine
7 Cordé              feminine
8 Luminara Unduli    feminine
9 Barriss Offee      feminine
10 Dormé              feminine
11 Zam Wesell         feminine
12 Taun We            feminine
13 Jocasta Nu        feminine
14 R4-P17             feminine
15 Shaak Ti           feminine
16 Rey                feminine
17 Padmé Amidala     feminine
> |
```

4.- Ordeneu els noms dels personatges femenins en ordre descendent. Podríeu fer-ho sense fer us de les variables temporals creades en exercici anterior (és a dir, sense fer ús del subgrup)?

Hem vist :

#### ARRANGE CASES



**arrange(.data, ...)** Order rows by values of a column or columns (low to high), use with **desc()** to order from high to low.  
 arrange(mtcars, mpg)  
 arrange(mtcars, desc(mpg))

Per tant:

```
> arrange(feminisub, desc(name)) #partint del subgrup creat

> arrange(feminisub, desc(name))
# A tibble: 17 x 2
  name                gender
  <chr>               <chr>
1 Zam Wesell         feminine
2 Taun We            feminine
3 Shmi Skywalker     feminine
4 Shaak Ti           feminine
5 Rey                feminine
6 R4-P17             feminine
7 Padmé Amidala     feminine
8 Mon Mothma         feminine
9 Luminara Unduli    feminine
10 Leia Organa        feminine
11 Jocasta Nu        feminine
12 Dormé              feminine
13 Cordé              feminine
14 Beru Whitesun lars feminine
15 Barriss Offee      feminine
16 Ayla Secura        feminine
17 Adi Gallia         feminine
> |
```

Per tal de no fer us del subgrup *feminisub*, altre cop hauríem d'utilitzar les *pipes*:

```
> ordered_fem <- starwars %>%filter(gender=="feminine")
%>%select(name,gender) %>%arrange(desc(name))
> ordered_fem

> ordered_fem <- starwars %>% filter(gender=="feminine") %>%select(name,gender) %>%arrange(desc(name))
> ordered_fem
# A tibble: 17 x 2
   name                gender
  <chr>              <chr>
1 Zam Wesell        feminine
2 Taun We           feminine
3 Shmi Skywalker    feminine
4 Shaak Ti          feminine
5 Rey               feminine
6 R4-P17            feminine
7 Padmé Amidala     feminine
8 Mon Mothma        feminine
9 Luminara Unduli   feminine
10 Leia Organa       feminine
11 Jocasta Nu        feminine
12 Dormé            feminine
13 Cordé            feminine
14 Beru Whitesun lars feminine
15 Barriss Offee     feminine
16 Ayla Secura       feminine
17 Adi Gallia        feminine
> |
```

Judit Chamorro Servent

Bellaterra, Març 2025