# 05 - Importation et jointure de données

PRO1036 - Analyse de données scientifiques en R

Tim Bollé

October 7, 2024



# Importation de données

#### Présentations







#### readr

- read\_csv() fichiers CSV
  - CSV: comma-separated values valeurs séparées par des virgules
- read\_csv2() fichiers CSV mais où le séparateur est un point-virgule
  - Commun dans les pays où le séparateur décimal est la virgule
- read tsv() fichiers TSV
  - TSV: tab-separated values valeurs séparées par des tabulations
- read\_delim() fichiers avec un délimiteur spécifique
  - On peut spécifier le délimiteur avec l'argument delim

• ...



# readxl

• read\_excel() - Permet de lire des fichiers Excel



#### Lecture de fichiers

```
1 nobel <- read csv(file = "data/nobel.csv")</pre>
  2 nobel
\# A tibble: 935 \times 26
      id firstname
                     surname year category affiliation city country born date
   <dbl> <chr>
                             <dbl> <dhr>
                                             <chr>
                                                         <chr> <chr>
                     <chr>
                                                                        <date>
      1 Wilhelm Co... Röntgen 1901 Physics Munich Uni... Muni... Germany 1845-03-27
      2 Hendrik A. Lorentz 1902 Physics Leiden Uni... Leid... Nether... 1853-07-18
      3 Pieter
                     Zeeman 1902 Physics Amsterdam ... Amst... Nether... 1865-05-25
      4 Henri
                     Becque... 1903 Physics École Poly... Paris France 1852-12-15
                     Curie 1903 Physics École muni... Paris France 1859-05-15
      5 Pierre
                     Curie 1903 Physics <NA>
      6 Marie
                                                         <NA> <NA>
                                                                       1867-11-07
                     Curie 1911 Chemist... Sorbonne U... Paris France 1867-11-07
      6 Marie
      8 Lord
                     Raylei... 1904 Physics Royal Inst... Lond... United... 1842-11-12
       9 Philipp
                     Lenard 1905 Physics Kiel Unive... Kiel Germany 1862-06-07
                     Thomson 1906 Physics University... Camb... United... 1856-12-18
10
      10 J.J.
# i 925 more rows
# i 17 more variables: died date <date>, gender <chr>, born city <chr>,
   born country <chr>, born country code <chr>, died city <chr>,
   died country <chr>, died country code <chr>, overall motivation <chr>,
   share <dbl>, motivation <chr>, born country original <chr>,
   born city original <chr>, died country original <chr>,
    died city original <chr>, city original <chr>, country original <chr>
```

#### Noms des varaibles

#### Le nom des variables n'est pas toujours optimal.

#### Et R n'aime pas les noms de variables avec des espaces.



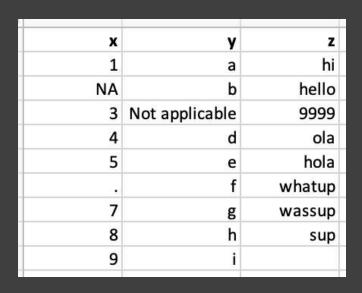
# Option 1: Spécifier les noms des variables

# Option 2: Utiliser le format snake\_case

- Les espaces sont remplacés par des underscores
- Les lettres sont en minuscules

#### Nous pouvons utiliser la fonction janitor::clean\_names()

# Gestion des types de données





# Spécification des NAs

# Spécification des types de chaque colonne

```
read csv("data/df-na.csv", col types = list(col double(),
                                                 col character(),
                                                 col character()))
# A tibble: 9 \times 3
 <dbl> <chr>
                   <chr>
                       hi
    NA b
                      hello
    3 Not applicable 9999
     4 d
                       ola
                      hola
    NA f
                      whatup
                      wassup
                       sup
                       <NA>
```

# Les types de colonnes

Fonction	Types de données
<pre>col_character()</pre>	Chaine de caractères
col_date()	Date
col_datetime()	POSIXct (date-time)
col_double()	Double (numeric)
col_factor()	Factor
col_guess()	Laisse readr deviner ( par défaut)
col_integer()	Entier
col_logical()	Logique
col_number()	Nombre et texte mélangés
col_numeric()	Double ou entier
col_skip()	Ne pas lire la colonne
col_time()	Temps



# Jointure de données



#### Kesako

Lorsque nous avons des données dans plusieurs fichiers/tables, il est souvent nécessaire de les combiner.

#### Données: Les femmes dans la science

Nous avons des informations sur 10 femmes qui ont changé le monde. Les informations sont réparties dans trois fichiers:

- professions.csv: Information sur la profession de chacune
- dates.csv: date de naissance et de décès de chacune
- works.csv: Ce qu'elles ont fait pour changer le monde

### professions.csv

```
professions <- read csv("data/scientists/professions.csv")</pre>
  2 professions
\# A tibble: 10 \times 2
                      profession
   name
                      <chr>
   <chr>
1 Ada Lovelace
                      Mathematician
2 Marie Curie
                      Physicist and Chemist
 3 Janaki Ammal
                      Botanist
 4 Chien-Shiung Wu
                      Physicist
 5 Katherine Johnson Mathematician
 6 Rosalind Franklin Chemist
 7 Vera Rubin
                      Astronomer
                      Mathematician
8 Gladys West
 9 Flossie Wong-Staal Virologist and Molecular Biologist
10 Jennifer Doudna
                      Biochemist
```



#### dates.csv

```
1 dates <- read csv("data/scientists/dates.csv")</pre>
  2 dates
# A tibble: 8 \times 3
                      birth year death year
  name
  <chr>
                           <dbl>
                                       <dbl>
1 Janaki Ammal
                                       1984
                            1897
2 Chien-Shiung Wu
                                       1997
                            1912
3 Katherine Johnson
                            1918
                                       2020
                                       1958
4 Rosalind Franklin
                            1920
5 Vera Rubin
                            1928
                                        2016
6 Gladys West
                            1930
7 Flossie Wong-Staal
                            1947
                                          NA
8 Jennifer Doudna
                            1964
                                          NA
```

#### works.csv

```
1 works <- read csv("data/scientists/works.csv")</pre>
  2 works
# A tibble: 9 \times 2
                     known for
  name
  <chr>
                     <chr>
1 Ada Lovelace
                     first computer algorithm
2 Marie Curie
                     theory of radioactivity, discovery of elements polonium a...
3 Janaki Ammal
                     hybrid species, biodiversity protection
4 Chien-Shiung Wu
                     confim and refine theory of radioactive beta decy, Wu expe...
5 Katherine Johnson calculations of orbital mechanics critical to sending the ...
6 Vera Rubin
                     existence of dark matter
                     mathematical modeling of the shape of the Earth which serv...
7 Gladys West
8 Flossie Wong-Staal first scientist to clone HIV and create a map of its genes...
9 Jennifer Doudna
                     one of the primary developers of CRISPR, a ground-breaking...
```



# Ce que nous voulons comme output

# A tibble: 10 × 5						
	name	profession		birth_year	death_year	known_for
	<chr></chr>	<chr></chr>		<dbl></dbl>	<dbl></dbl>	<chr></chr>
1	Ada Lovelace	Mathematician		NA	NA	first co
2	Marie Curie	Physicist and	Chemist	NA	NA	theory o
3	Janaki Ammal	Botanist		1897	1984	hybrid s
4	Chien-Shiung Wu	Physicist		1912	1997	confim a
5	Katherine Johnson	Mathematician		1918	2020	calculat
6	Rosalind Franklin	Chemist		1920	1958	<na></na>
7	Vera Rubin	Astronomer		1928	2016	existenc
8	Gladys West	Mathematician		1930	NA	mathemat
9	Flossie Wong-Staal	Virologist and	Molecular	1947	NA	first sc
10	Jennifer Doudna	Biochemist		1964	NA	one of t



# Types de jointures

- left\_join(): Retourne toutes les lignes de la première table et les lignes correspondantes de la deuxième table
- right\_join(): Retourne toutes les lignes de la deuxième table et les lignes correspondantes de la première table
- inner\_join(): Retourne les lignes qui ont une correspondance dans les deux tables
- full\_join(): Retourne toutes les lignes des deux tables
- semi\_join(): Retourne toutes les lignes de la première table qui ont une correspondance dans la deuxième table
- anti\_join(): Retourne toutes les lignes de la première table qui n'ont pas de correspondance dans la deuxième table



# Pour l'exemple...

```
1 x
# A tibble: 3 × 2
    id value_x
    <dbl> <chr>
1    1 x1
2    2 x2
3    3 x3
```

```
1 y
# A tibble: 3 × 2
    id value_y
    <dbl> <chr>
1     1 y1
2     2 y2
3     4 y4
```



# left\_join()

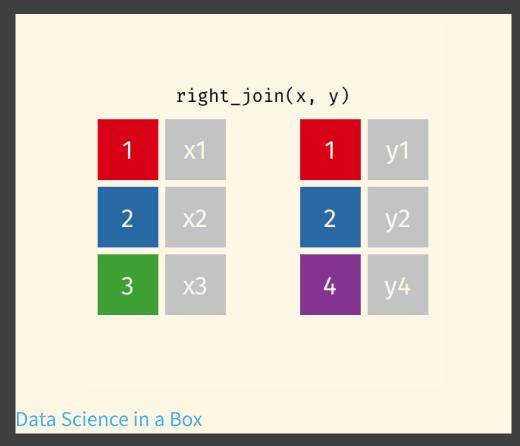
```
left_join(x, y)
          3
                                4
Data Science in a Box
```



# left\_join()

```
professions %>%
      left join(dates)
# A tibble: 10 \times 4
                      profession
                                                          birth year death year
   name
                      <chr>
   <chr>
                                                                <dbl>
                                                                           <dbl>
                      Mathematician
1 Ada Lovelace
                                                                              NA
2 Marie Curie
                      Physicist and Chemist
 3 Janaki Ammal
                                                                 1897
                                                                            1984
                      Botanist
 4 Chien-Shiung Wu
                      Physicist
                                                                1912
                                                                            1997
                     Mathematician
                                                                1918
                                                                            2020
 5 Katherine Johnson
 6 Rosalind Franklin Chemist
                                                                1920
                                                                            1958
 7 Vera Rubin
                                                                1928
                                                                            2016
                      Astronomer
                      Mathematician
8 Gladys West
                                                                1930
 9 Flossie Wong-Staal Virologist and Molecular Biologist
                                                                1947
                                                                              NA
10 Jennifer Doudna
                      Biochemist
                                                                1964
                                                                              NA
```

# right\_join()



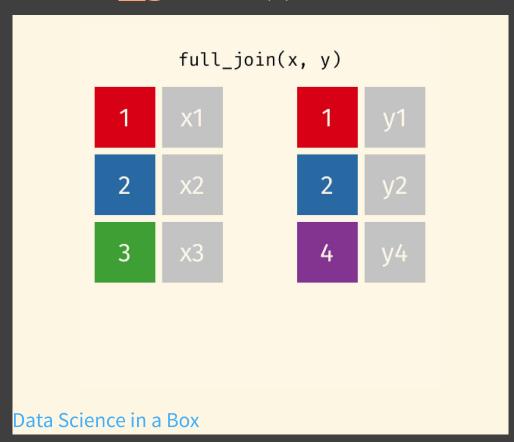


# right\_join()

```
professions %>%
      right join(dates)
# A tibble: 8 \times 4
                     profession
                                                         birth year death year
  name
                     <chr>
                                                              <dbl>
                                                                         <dbl>
  <chr>
1 Janaki Ammal
                                                               1897
                                                                          1984
                     Botanist
                                                               1912
2 Chien-Shiung Wu
                     Physicist
                                                                          1997
3 Katherine Johnson Mathematician
                                                               1918
                                                                          2020
4 Rosalind Franklin Chemist
                                                               1920
                                                                          1958
                                                               1928
5 Vera Rubin
                                                                          2016
                     Astronomer
                     Mathematician
6 Gladys West
                                                               1930
7 Flossie Wong-Staal Virologist and Molecular Biologist
                                                               1947
                                                                            NA
8 Jennifer Doudna
                     Biochemist
                                                               1964
```



# full\_join()



```
1 full_join(x, y)
# A tibble: 4 × 3
        id value_x value_y
        <dbl> <chr>
        1      1      x1           y1
2         2      x2           y2
3         3      x3           <NA>
4         4 <NA>           y4
```

# full\_join()

```
professions %>%
      full join (works)
# A tibble: 10 \times 3
                       profession
                                                            known for
   name
   <chr>
                       <chr>
                                                            <chr>
                                                            first computer algorit ...
1 Ada Lovelace
                       Mathematician
                       Physicist and Chemist
2 Marie Curie
                                                            theory of radioactivit...
 3 Janaki Ammal
                                                            hybrid species, biodiv...
                       Botanist
 4 Chien-Shiung Wu
                       Physicist
                                                            confim and refine theo ...
 5 Katherine Johnson
                      Mathematician
                                                            calculations of orbita...
 6 Rosalind Franklin Chemist
                                                            <NA>
 7 Vera Rubin
                                                            existence of dark matt...
                       Astronomer
                       Mathematician
                                                            mathematical modeling ...
8 Gladys West
 9 Flossie Wong-Staal Virologist and Molecular Biologist first scientist to clo...
10 Jennifer Doudna
                       Biochemist
                                                            one of the primary dev...
```



# inner\_join()

```
inner_join(x, y)
          3
                               4
Data Science in a Box
```

# inner\_join()

```
professions %>%
      inner join (dates)
# A tibble: 8 \times 4
                     profession
                                                         birth year death year
  name
                     <chr>
                                                              <dbl>
                                                                         <dbl>
  <chr>
1 Janaki Ammal
                                                               1897
                                                                          1984
                     Botanist
                                                               1912
2 Chien-Shiung Wu
                     Physicist
                                                                          1997
3 Katherine Johnson Mathematician
                                                               1918
                                                                          2020
4 Rosalind Franklin Chemist
                                                               1920
                                                                          1958
                                                               1928
5 Vera Rubin
                                                                          2016
                     Astronomer
                     Mathematician
6 Gladys West
                                                               1930
7 Flossie Wong-Staal Virologist and Molecular Biologist
                                                               1947
                                                                            NA
8 Jennifer Doudna
                     Biochemist
                                                               1964
```

## Si on reprend...

```
professions %>%
      left join(dates) %>%
      left join(works)
\# A tibble: 10 \times 5
                       profession
                                                   birth year death year known for
   name
   <chr>
                       <chr>
                                                         <db1>
                                                                     <dbl> <chr>
1 Ada Lovelace
                       Mathematician
                                                                        NA first co...
2 Marie Curie
                       Physicist and Chemist
                                                                        NA theory o...
                       Botanist
                                                                      1984 hybrid s...
 3 Janaki Ammal
                                                          1897
                                                                      1997 confim a...
4 Chien-Shiung Wu
                       Physicist
                                                          1912
5 Katherine Johnson Mathematician
                                                          1918
                                                                      2020 calculat...
 6 Rosalind Franklin Chemist
                                                          1920
                                                                      1958 <NA>
                                                          1928
 7 Vera Rubin
                                                                      2016 existenc...
                       Astronomer
                                                          1930
                                                                        NA mathemat...
8 Gladys West
                       Mathematician
9 Flossie Wong-Staal Virologist and Molecular ...
                                                                        NA first sc...
                                                          1947
10 Jennifer Doudna
                       Biochemist
                                                          1964
                                                                        NA one of t...
```

# Références

