

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")
2 nobel

# A tibble: 935 × 26
   id firstname  surname  year category affiliation city  country born_date
  <dbl> <chr>      <chr>   <dbl> <chr>      <chr>    <chr> <chr>   <date>
1     1 Wilhelm Co... Röntgen  1901 Physics  Munich Uni... Muni... Germany 1845-03-27
2     2 Hendrik A.   Lorentz  1902 Physics  Leiden Uni... Leid... Nether... 1853-07-18
3     3 Pieter       Zeeman   1902 Physics  Amsterdam ... Amst... Nether... 1865-05-25
4     4 Henri       Becque... 1903 Physics  École Poly... Paris France 1852-12-15
5     5 Pierre      Curie     1903 Physics  École muni... Paris France 1859-05-15
6     6 Marie      Curie     1903 Physics  <NA>         <NA>   <NA>     1867-11-07
7     6 Marie      Curie     1911 Chemist... Sorbonne U... Paris France 1867-11-07
8     8 Lord       Raylei... 1904 Physics  Royal Inst... Lond... United... 1842-11-12
9     9 Philipp    Lenard   1905 Physics  Kiel Unive... Kiel  Germany 1862-06-07
10    10 J.J.       Thomson  1906 Physics  University... Camb... United... 1856-12-18

# 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 variables

Le nom des variables n'est pas toujours optimal.

```
1 edibnb_badnames <- read_csv("data/edibnb-badnames.csv")
2 names(edibnb_badnames)
```

```
[1] "ID"                "Price"                "neighbourhood"
[4] "accommodates"      "Number of bathrooms"  "Number of Bedrooms"
[7] "n beds"            "Review Scores Rating" "Number of reviews"
[10] "listing_url"
```

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

```
1 ggplot(edibnb_badnames, aes(x = Number of bathrooms, y = Price)) +
2   geom_point()
```

```
Error: <text>:1:40: symbole inattendu
1: ggplot(edibnb_badnames, aes(x = Number of
                                     ^
```

Option 1: Spécifier les noms des variables

```
1 edibnb_col_names <- read_csv("data/edibnb-badnames.csv",
2                               col_names = c("id", "price",
3                                              "neighbourhood", "accommodates",
4                                              "bathroom", "bedroom",
5                                              "bed", "review_scores_rating",
6                                              "n_reviews", "url"))
7 names(edibnb_col_names)
```

[1]	"id"	"price"	"neighbourhood"
[4]	"accommodates"	"bathroom"	"bedroom"
[7]	"bed"	"review_scores_rating"	"n_reviews"
[10]	"url"		

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()`

```
1 edibnb_clean_names <- read_csv("data/edibnb-badnames.csv") %>%  
2   janitor::clean_names()  
3 names(edibnb_clean_names)
```

```
[1] "id"           "price"         "neighbourhood"  
[4] "accommodates" "number_of_bathrooms" "number_of_bedrooms"  
[7] "n_beds"       "review_scores_rating" "number_of_reviews"  
[10] "listing_url"
```

Gestion des types de données

x	y	z
1	a	hi
NA	b	hello
3	Not applicable	9999
4	d	ola
5	e	hola
.	f	whatup
7	g	wassup
8	h	sup
9	i	

```
1 read_csv("data/df-na.csv")
# A tibble: 9 × 3
  x     y     z
<chr> <chr> <chr>
1 1     a     hi
2 <NA>  b     hello
3 3     Not applicable 9999
4 4     d     ola
5 5     e     hola
6 .     f     whatup
7 7     g     wassup
8 8     h     sup
9 9     i     <NA>
```

Spécification des NAs

```
1 read_csv("data/df-na.csv",
2           na = c("", "NA", ".", "9999", "Not applicable"))

# A tibble: 9 × 3
#       x y      z
#   <dbl> <chr> <chr>
1     1 a     hi
2    NA b     hello
3     3 <NA> <NA>
4     4 d     ola
5     5 e     hola
6    NA f     whatup
7     7 g     wassup
8     8 h     sup
9     9 i     <NA>
```

Spécification des types de chaque colonne

```
1 read_csv("data/df-na.csv", col_types = list(col_double(),
2                                             col_character(),
3                                             col_character()))
```

A tibble: 9 × 3

	x	y	z
	<dbl>	<chr>	<chr>
1	1	a	hi
2	NA	b	hello
3	3	Not applicable	9999
4	4	d	ola
5	5	e	hola
6	NA	f	whatup
7	7	g	wassup
8	8	h	sup
9	9	i	<NA>

Les types de colonnes

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

```
1 professions <- read_csv("data/scientists/professions.csv")
2 professions

# A tibble: 10 × 2
   name                profession
   <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
8 Gladys West         Mathematician
9 Flossie Wong-Staal  Virologist and Molecular Biologist
10 Jennifer Doudna     Biochemist
```

dates.csv

```
1 dates <- read_csv("data/scientists/dates.csv")
2 dates
```

```
# A tibble: 8 × 3
  name          birth_year death_year
  <chr>          <dbl>     <dbl>
1 Janaki Ammal    1897       1984
2 Chien-Shiung Wu  1912       1997
3 Katherine Johnson 1918       2020
4 Rosalind Franklin 1920       1958
5 Vera Rubin      1928       2016
6 Gladys West      1930        NA
7 Flossie Wong-Staal 1947        NA
8 Jennifer Doudna   1964        NA
```

works.csv

```
1 works <- read_csv("data/scientists/works.csv")
2 works
```

A tibble: 9 × 2

	name	known_for
	<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	confirm and refine theory of radioactive beta decay, Wu expe...
5	Katherine Johnson	calculations of orbital mechanics critical to sending the ...
6	Vera Rubin	existence of dark matter
7	Gladys West	mathematical modeling of the shape of the Earth which serv...
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 <chr>	profession <chr>	birth_year <dbl>	death_year <dbl>	known_for <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>
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)

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Data Science in a Box

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

left_join()

```
1 professions %>%
2   left_join(dates)
```

```
# A tibble: 10 × 4
```

	name <chr>	profession <chr>	birth_year <dbl>	death_year <dbl>
1	Ada Lovelace	Mathematician	NA	NA
2	Marie Curie	Physicist and Chemist	NA	NA
3	Janaki Ammal	Botanist	1897	1984
4	Chien-Shiung Wu	Physicist	1912	1997
5	Katherine Johnson	Mathematician	1918	2020
6	Rosalind Franklin	Chemist	1920	1958
7	Vera Rubin	Astronomer	1928	2016
8	Gladys West	Mathematician	1930	NA
9	Flossie Wong-Staal	Virologist and Molecular Biologist	1947	NA
10	Jennifer Doudna	Biochemist	1964	NA

right_join()

right_join(x, y)

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Data Science in a Box

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

right_join()

```
1 professions %>%
2   right_join(dates)
```

A tibble: 8 × 4

	name	profession	birth_year	death_year
	<chr>	<chr>	<dbl>	<dbl>
1	Janaki Ammal	Botanist	1897	1984
2	Chien-Shiung Wu	Physicist	1912	1997
3	Katherine Johnson	Mathematician	1918	2020
4	Rosalind Franklin	Chemist	1920	1958
5	Vera Rubin	Astronomer	1928	2016
6	Gladys West	Mathematician	1930	NA
7	Flossie Wong-Staal	Virologist and Molecular Biologist	1947	NA
8	Jennifer Doudna	Biochemist	1964	NA

full_join()

full_join(x, y)

1	x1	1	y1
2	x2	2	y2
3	x3		
		4	y4

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

Data Science in a Box

full_join()

```
1 professions %>%
2   full_join(works)
```

A tibble: 10 × 3

	name <chr>	profession <chr>	known_for <chr>
1	Ada Lovelace	Mathematician	first computer algorit...
2	Marie Curie	Physicist and Chemist	theory of radioactivit...
3	Janaki Ammal	Botanist	hybrid species, biodiv...
4	Chien-Shiung Wu	Physicist	confim and refine theo...
5	Katherine Johnson	Mathematician	calculations of orbita...
6	Rosalind Franklin	Chemist	<NA>
7	Vera Rubin	Astronomer	existence of dark matt...
8	Gladys West	Mathematician	mathematical modeling ...
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)

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Data Science in a Box

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

inner_join()

```
1 professions %>%
2   inner_join(dates)
```

A tibble: 8 × 4

	name	profession	birth_year	death_year
	<chr>	<chr>	<dbl>	<dbl>
1	Janaki Ammal	Botanist	1897	1984
2	Chien-Shiung Wu	Physicist	1912	1997
3	Katherine Johnson	Mathematician	1918	2020
4	Rosalind Franklin	Chemist	1920	1958
5	Vera Rubin	Astronomer	1928	2016
6	Gladys West	Mathematician	1930	NA
7	Flossie Wong-Staal	Virologist and Molecular Biologist	1947	NA
8	Jennifer Doudna	Biochemist	1964	NA

Si on reprend...

```

1 professions %>%
2   left_join(dates) %>%
3   left_join(works)
# A tibble: 10 × 5
   name                profession      birth_year death_year known_for
  <chr>                <chr>          <dbl>        <dbl> <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 confir a...
5 Katherine Johnson    Mathematician      1918          2020 calculat...
6 Rosalind Franklin    Chemist            1920          1958 <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...

```

Références