

# Data wrangling

Working with multiple data frames

Prof. Dr. Jan Kirenz

*The following content is based on Mine Çetinkaya-Rundel's excellent book Data Science in a Box*

We...

**have** multiple data frames

**want** to bring them together

# Data: Women in science

Information on 10 women in science who changed the world

---

name
Ada Lovelace
Marie Curie
Janaki Ammal
Chien-Shiung Wu
Katherine Johnson
Rosalind Franklin
Vera Rubin
Gladys West
Flossie Wong-Staal
Jennifer Doudna

---

Source: [Discover Magazine](#)

# Inputs

professions

dates

works

professions

```
## # A tibble: 10 x 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
```

# Desired output

```
## # A tibble: 10 x 5
##   name      profession  birth_year death_year known_for
##   <chr>      <chr>          <dbl>      <dbl> <chr>
## 1 Ada Lov... Mathematician      NA          NA first computer a...
## 2 Marie C... Physicist an...      NA          NA theory of radioa...
## 3 Janaki ... Botanist          1897        1984 hybrid species, ...
## 4 Chien-S... Physicist          1912        1997 confirm and refin...
## 5 Katheri... Mathematician      1918        2020 calculations of ...
## 6 Rosalin... Chemist           1920        1958 <NA>
## 7 Vera Ru... Astronomer         1928        2016 existence of dar...
## 8 Gladys ... Mathematician      1930          NA mathematical mod...
## 9 Flossie... Virologist a...      1947          NA first scientist ...
## 10 Jennife... Biochemist         1964          NA one of the prima...
```

# Inputs, reminder

```
names(professions)
```

```
## [1] "name"      "profession"
```

```
names(dates)
```

```
## [1] "name"      "birth_year" "death_year"
```

```
names(works)
```

```
## [1] "name"      "known_for"
```

```
nrow(professions)
```

```
## [1] 10
```

```
nrow(dates)
```

```
## [1] 8
```

```
nrow(works)
```

```
## [1] 9
```

# Joining data frames

# Joining data frames

```
something_join(x, y)
```

- `left_join()`: all rows from x
- `right_join()`: all rows from y
- `full_join()`: all rows from both x and y
- `semi_join()`: all rows from x where there are matching values in y, keeping just columns from x
- `inner_join()`: all rows from x where there are matching values in y, return all combination of multiple matches in the case of multiple matches
- `anti_join()`: return all rows from x where there are not matching values in y, never duplicate rows of x
- ...



# Setup

For the next few slides...

x

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

y

```
## # A tibble: 3 x 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

```
left_join(x, y)
```

```
## # A tibble: 3 x 3
##       id value_x value_y
##   <dbl> <chr>   <chr>
## 1     1    x1     y1
## 2     2    x2     y2
## 3     3    x3    <NA>
```

# left\_join()

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

```
## # A tibble: 10 x 4  
##   name                profession      birth_year death_year  
##   <chr>               <chr>          <dbl>      <dbl>  
## 1 Ada Lovelace        Mathematician      NA         NA  
## 2 Marie Curie         Physicist and Chemist NA         NA  
## 3 Janaki Ammal        Botanist          1897      1984  
## 4 Chien-Shiung ...    Physicist         1912      1997  
## 5 Katherine Joh...    Mathematician      1918      2020  
## 6 Rosalind Fran...    Chemist           1920      1958  
## 7 Vera Rubin          Astronomer        1928      2016  
## 8 Gladys West         Mathematician      1930         NA  
## 9 Flossie Wong-...    Virologist and Molecular... 1947         NA  
## 10 Jennifer Doud...    Biochemist        1964         NA
```

# right\_join()

right\_join(x, y)

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

```
right_join(x, y)
```

```
## # A tibble: 3 x 3
##       id value_x value_y
##   <dbl> <chr>   <chr>
## 1     1    x1     y1
## 2     2    x2     y2
## 3     4 <NA>    y4
```

# right\_join()

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

```
## # A tibble: 8 x 4  
##   name                profession      birth_year death_year  
##   <chr>              <chr>          <dbl>      <dbl>  
## 1 Janaki Ammal      Botanist        1897        1984  
## 2 Chien-Shiung ... Physicist       1912        1997  
## 3 Katherine Joh... Mathematician   1918        2020  
## 4 Rosalind Fran... Chemist        1920        1958  
## 5 Vera Rubin       Astronomer     1928        2016  
## 6 Gladys West      Mathematician   1930         NA  
## 7 Flossie Wong-... Virologist and Molecular ... 1947         NA  
## 8 Jennifer Doud... Biochemist     1964         NA
```

# full\_join()

full\_join(x, y)

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

full\_join(x, y)

```
## # A tibble: 4 x 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
```

# full\_join()

```
dates %>%  
  full_join(works)
```

```
## # A tibble: 10 x 4  
##   name      birth_year death_year known_for  
##   <chr>      <dbl>      <dbl> <chr>  
## 1 Janaki Am...    1897        1984 hybrid species, biodiversity...  
## 2 Chien-Shi...    1912        1997 confirm and refine theory of ...  
## 3 Katherine...    1918        2020 calculations of orbital mech...  
## 4 Rosalind ...    1920        1958 <NA>  
## 5 Vera Rubin     1928        2016 existence of dark matter  
## 6 Gladys We...    1930         NA mathematical modeling of the...  
## 7 Flossie W...    1947         NA first scientist to clone HIV...  
## 8 Jennifer ...    1964         NA one of the primary developer...  
## 9 Ada Lovel...     NA         NA first computer algorithm  
## 10 Marie Cur...    NA         NA theory of radioactivity, di...
```

# inner\_join()

inner\_join(x, y)

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

```
inner_join(x, y)
```

```
## # A tibble: 2 x 3
##       id value_x value_y
##   <dbl> <chr>   <chr>
## 1     1    x1     y1
## 2     2    x2     y2
```



# inner\_join()

```
dates %>%  
  inner_join(works)
```

```
## # A tibble: 7 x 4  
##   name      birth_year death_year known_for  
##   <chr>      <dbl>      <dbl> <chr>  
## 1 Janaki Amm...    1897        1984 hybrid species, biodiversity...  
## 2 Chien-Shiu...    1912        1997 confirm and refine theory of ...  
## 3 Katherine ...    1918        2020 calculations of orbital mech...  
## 4 Vera Rubin      1928        2016 existence of dark matter  
## 5 Gladys West     1930         NA mathematical modeling of the...  
## 6 Flossie Wo...    1947         NA first scientist to clone HIV...  
## 7 Jennifer D...    1964         NA one of the primary developer...
```

# semi\_join()

semi\_join(x, y)

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

```
semi_join(x, y)
```

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

# semi\_join()

```
dates %>%  
  semi_join(works)
```

```
## # A tibble: 7 x 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 Vera Rubin          1928        2016  
## 5 Gladys West         1930         NA  
## 6 Flossie Wong-Staal  1947         NA  
## 7 Jennifer Doudna     1964         NA
```

# anti\_join()

anti\_join(x, y)

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

```
anti_join(x, y)
```

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

# anti\_join()

```
dates %>%  
  anti_join(works)
```

```
## # A tibble: 1 x 3  
##   name          birth_year death_year  
##   <chr>         <dbl>     <dbl>  
## 1 Rosalind Franklin    1920      1958
```

# Putting it altogether

```
professions %>%  
  left_join(dates) %>%  
  left_join(works)
```

```
## # A tibble: 10 x 5  
##   name      profession birth_year death_year known_for  
##   <chr>    <chr>         <dbl>      <dbl> <chr>  
## 1 Ada Lov... Mathematician      NA         NA first computer a...  
## 2 Marie C... Physicist an...      NA         NA theory of radioa...  
## 3 Janaki ... Botanist         1897        1984 hybrid species, ...  
## 4 Chien-S... Physicist         1912        1997 confirm and refin...  
## 5 Katheri... Mathematician     1918        2020 calculations of ...  
## 6 Rosalin... Chemist          1920        1958 <NA>  
## 7 Vera Ru... Astronomer       1928        2016 existence of dar...  
## 8 Gladys ... Mathematician     1930         NA mathematical mod...  
## 9 Flossie... Virologist a...     1947         NA first scientist ...  
## 10 Jennife... Biochemist       1964         NA one of the prima...
```

# Case study: Student records

# Student records

- **Have:**

- `enrolment`: official university enrolment records
- `survey`: Student provided info; missing students who never filled it out; and including students who filled it out but dropped the class

- **Want:** Survey info for all enrolled in class

enrolment

```
## # A tibble: 3 x 2
##   id name
##   <dbl> <chr>
## 1     1 Dave Friday
## 2     2 Hermine
## 3     3 Sura Selvarajah
```

survey

```
## # A tibble: 4 x 3
##   id name      username
##   <dbl> <chr>    <chr>
## 1     2 Hermine bakealongwithhermine
## 2     3 Sura    surasbakes
## 3     4 Peter   peter_bakes
## 4     5 Mark    thebakingbuddha
```



# Student records

In class

Survey missing

Dropped

```
enrolment %>%  
  left_join(survey, by = "id")
```

```
## # A tibble: 3 x 4  
##       id name.x      name.y username  
##   <dbl> <chr>      <chr>    <chr>  
## 1     1  Dave Friday <NA>    <NA>  
## 2     2   Hermine Hermine bakealongwithhermine  
## 3     3 Sura Selvarajah Sura    surasbakes
```

# Case study: Grocery sales

# Grocery sales

- **Have:**

- Purchases: One row per customer per item, listing purchases they made
- Prices: One row per item in the store, listing their prices

- **Want:** Total revenue

purchases

```
## # A tibble: 5 x 2
##   customer_id item
##         <dbl> <chr>
## 1           1 bread
## 2           1 milk
## 3           1 banana
## 4           2 milk
## 5           2 toilet paper
```

prices

```
## # A tibble: 5 x 2
##   item      price
##   <chr>    <dbl>
## 1 avocado    0.5
## 2 banana    0.15
## 3 bread      1
## 4 milk      0.8
## 5 toilet paper 3
```

# Grocery sales

Total revenue

Revenue per customer

```
purchases %>%  
  left_join(prices)
```

```
## # A tibble: 5 x 3  
##   customer_id item      price  
##         <dbl> <chr>    <dbl>  
## 1           1 bread      1  
## 2           1 milk      0.8  
## 3           1 banana    0.15  
## 4           2 milk      0.8  
## 5           2 toilet paper 3
```

```
purchases %>%  
  left_join(prices) %>%  
  summarise(total_revenue = sum(price))
```

```
## # A tibble: 1 x 1  
##   total_revenue  
##         <dbl>  
## 1           5.75
```