Working with multiple data frames

Data Science in a Box datasciencebox.org



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 \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
   8 Gladys West
                         Mathematician
##
   9 Flossie Wong-Staal Virologist and Molecular Biologist
  10 Jennifer Doudna
                         Biochemist
```



Inputs

professions

dates

works

dates

```
A tibble: 8 \times 3
##
                         birth_year death_year
     name
##
     <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
```



Inputs

professions dates works

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 elem...
                         hybrid species, biodiversity protection
## 3 Janaki Ammal
                         confim and refine theory of radioactive bet...
  4 Chien-Shiung Wu
  5 Katherine Johnson
                         calculations of orbital mechanics critical ...
  6 Vera Rubin
                         existence of dark matter
                        mathematical modeling of the shape of the E...
  7 Gladys West
## 8 Flossie Wong-Staal first scientist to clone HIV and create a m...
## 9 Jennifer Doudna
                        one of the primary developers of CRISPR, a ...
```



Desired output

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

Inputs, reminder

```
nrow(professions)
names(professions)
  [1] "name"
                                                ## [1] 10
                    "profession"
names(dates)
                                                 nrow(dates)
                    "birth_year" "death_year" ## [1] 8
## [1] "name"
names(works)
                                                 nrow(works)
                                                ## [1] 9
  [1] "name"
                   "known_for"
```

Joining data frames



Joining data frames

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

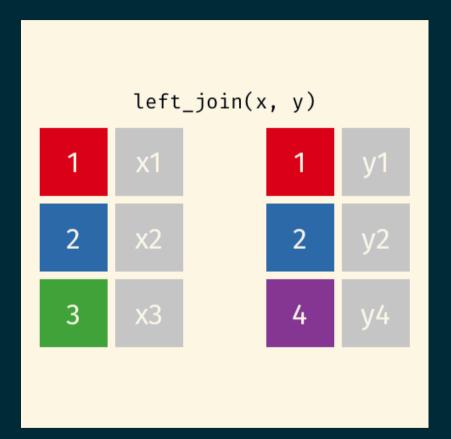
- left_join():all rows from x
- right_join():all rows from y
- full_join(): all rows from both x and y
- semi_join(): all rows from x where there are matching values in y, keeping just columns from x
- inner_join(): all rows from x where there are matching values in y, return all combination of multiple 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...

```
У
  # A tibble: 3 \times 2
                                                      # A tibble: 3 \times 2
        id value_x
                                                            id value_y
##
                                                   ##
     <dbl> <chr>
##
                                                   ##
                                                        <dbl> <chr>
                                                             1 y1
## 1
         1 x1
                                                   ## 1
## 2
      2 x2
                                                   ## 2
                                                             2 y2
## 3
         3 x3
                                                   ## 3
                                                             4 y4
```

left_join()



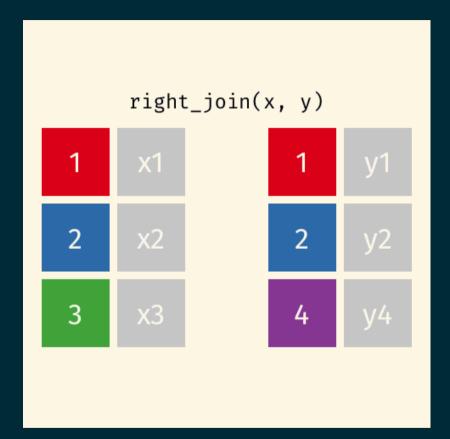
```
left_join(x, y)
```

left_join()

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

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

right_join()



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

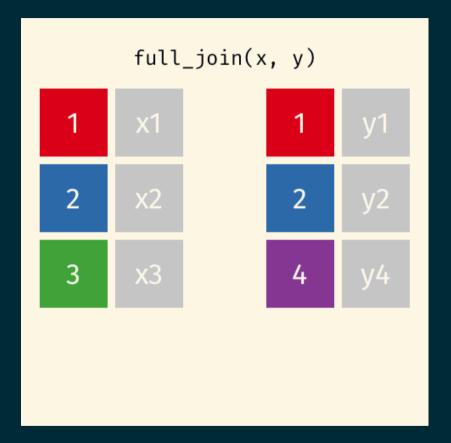
right_join()

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

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



full_join()



```
full_join(x, y)
```

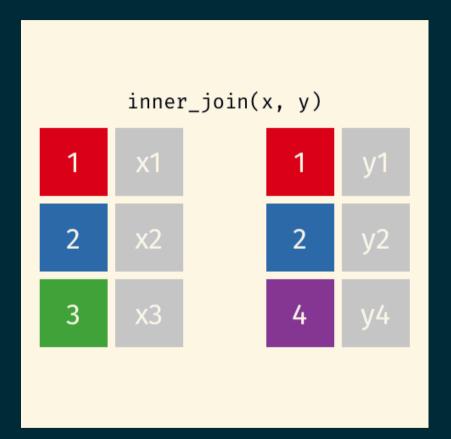
```
## # A tibble: 4 × 3
        id value_x value_y
##
    <dbl> <chr>
                  <chr>
## 1
         1 x1
                   y1
## 2
        2 x2
                   y2
## 3
        3 x3
                  <NA>
## 4
         4 <NA>
                   y4
```

full_join()

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

```
# A tibble: 10 \times 4
##
                          birth year death year known for
      name
##
      <chr>
                               <dbl>
                                           <dhl> <chr>
    1 Janaki Ammal
                                            1984 hybrid species, biod...
                                 1897
   2 Chien-Shiung Wu
                                 1912
                                             1997 confim and refine th...
##
   3 Katherine Johnson
                                 1918
##
                                            2020 calculations of orbi...
##
    4 Rosalind Franklin
                                 1920
                                            1958 <NA>
                                 1928
##
    5 Vera Rubin
                                            2016 existence of dark ma...
                                 1930
##
    6 Gladys West
                                              NA mathematical modelin...
##
    7 Flossie Wong-Staal
                                 1947
                                              NA first scientist to c...
##
   8 Jennifer Doudna
                                 1964
                                              NA one of the primary d...
##
    9 Ada Lovelace
                                              NA first computer algor...
  10 Marie Curie
                                   NA
                                              NA theory of radioactiv...
```

inner_join()



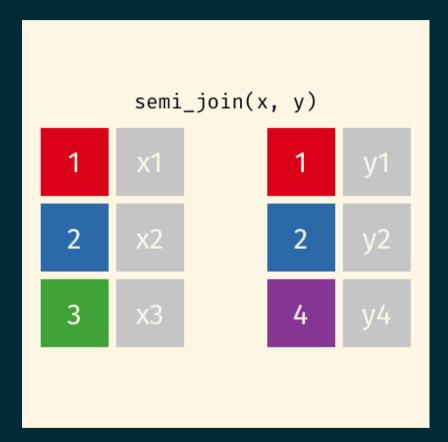
```
inner_join(x, y)
```

inner_join()

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

```
## # A tibble: 7 \times 4
##
                         birth year death year known for
     name
##
     <chr>
                               <ld><ld><
                                           <dhl> <chr>
  1 Janaki Ammal
                                            1984 hybrid species, biodi...
                                1897
  2 Chien-Shiung Wu
                                1912
                                            1997 confim and refine the...
  3 Katherine Johnson
                                1918
                                            2020 calculations of orbit...
  4 Vera Rubin
                                1928
                                            2016 existence of dark mat...
                                1930
                                              NA mathematical modeling...
## 5 Gladys West
## 6 Flossie Wong-Staal
                                              NA first scientist to cl...
                                1947
## 7 Jennifer Doudna
                                1964
                                              NA one of the primary de...
```

semi_join()



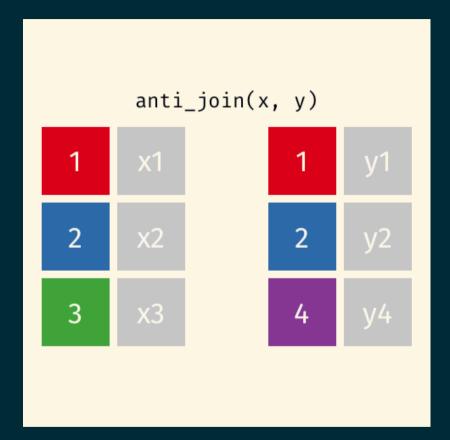
```
## # A tibble: 2 × 2
## id value_x
## 2 2 x2
```

semi_join()

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

```
## # A tibble: 7 × 3
##
                        birth_year death_year
     name
##
     <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)
## # A tibble: 1 × 2
```

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

anti_join()

dates %>%

Putting it altogether

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

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



Case study: 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

survey

In class

enrolment %>%

Survey missing

Dropped

```
left join(survey, by = "id")
  # A tibble: 3 \times 4
##
        id name.x
                           name.y
                                   username
##
    <dbl> <chr>
                           <chr>
                                   <chr>
        1 Dave Friday
                           <NA>
                                   <NA>
## 1
     2 Hermine
## 2
                           Hermine bakealongwithhermine
## 3
        3 Sura Selvarajah Sura
                                   surasbakes
```



1 Dave Friday

In class

Survey missing

Dropped

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

## # A tibble: 1 × 2
## id name
## <dbl> <chr>
```



In class

survey %>%

Survey missing

Dropped

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

prices

Grocery sales

Total revenue

purchases %>%

Revenue per customer

```
left join(prices)
## # A tibble: 5 × 3
     customer id item
##
                              price
           <dbl> <chr>
                              <dbl>
##
## 1
               1 bread
## 2
               1 milk
                               0.8
## 3
               1 banana
                               0.15
## 4
               2 milk
                               0.8
## 5
               2 toilet paper
```

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

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

Grocery sales

Total revenue

Revenue per customer

```
purchases %>%
   left join(prices)
## # A tibble: 5 × 3
     customer id item
##
                              price
           <dbl> <chr>
                              <dbl>
##
## 1
               1 bread
## 2
               1 milk
                               0.8
## 3
               1 banana
                               0.15
               2 milk
                               0.8
## 4
## 5
               2 toilet paper
```

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