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 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
```

Inputs

professions dates works

```
dates
```

```
## # A tibble: 8 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 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 x 2
##   name               known_for
##   <chr>              <chr>
## 1 Ada Lovelace       first computer algorithm
## 2 Marie Curie        theory of radioactivity,  discovery of elem~
## 3 Janaki Ammal       hybrid species, biodiversity protection
## 4 Chien-Shiung Wu    confim and refine theory of radioactive bet~
## 5 Katherine Johnson  calculations of orbital mechanics critical ~
## 6 Vera Rubin         existence of dark matter
## 7 Gladys West        mathematical modeling of the shape of the E~
## 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 x 5
##    name               profession          birth~1 death~2 known~3
##    <chr>              <chr>                 <dbl>   <dbl> <chr>
##  1 Ada Lovelace       Mathematician            NA      NA first ~
##  2 Marie Curie        Physicist and Chem~      NA      NA theory~
##  3 Janaki Ammal       Botanist               1897    1984 hybrid~
##  4 Chien-Shiung Wu    Physicist              1912    1997 confim~
##  5 Katherine Johnson  Mathematician          1918    2020 calcul~
##  6 Rosalind Franklin  Chemist                1920    1958 <NA>
##  7 Vera Rubin         Astronomer             1928    2016 existe~
##  8 Gladys West        Mathematician          1930      NA mathem~
##  9 Flossie Wong-Staal Virologist and Mol~    1947      NA first ~
## 10 Jennifer Doudna    Biochemist             1964      NA one of~
## # ... with abbreviated variable names 1: birth_year,
## #   2: death_year, 3: known_for
```

Inputs, reminder

```r
names(professions)
```

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

```r
names(dates)
```

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

```r
names(works)
```

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

```r
nrow(professions)
```

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

```r
nrow(dates)
```

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

```r
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...

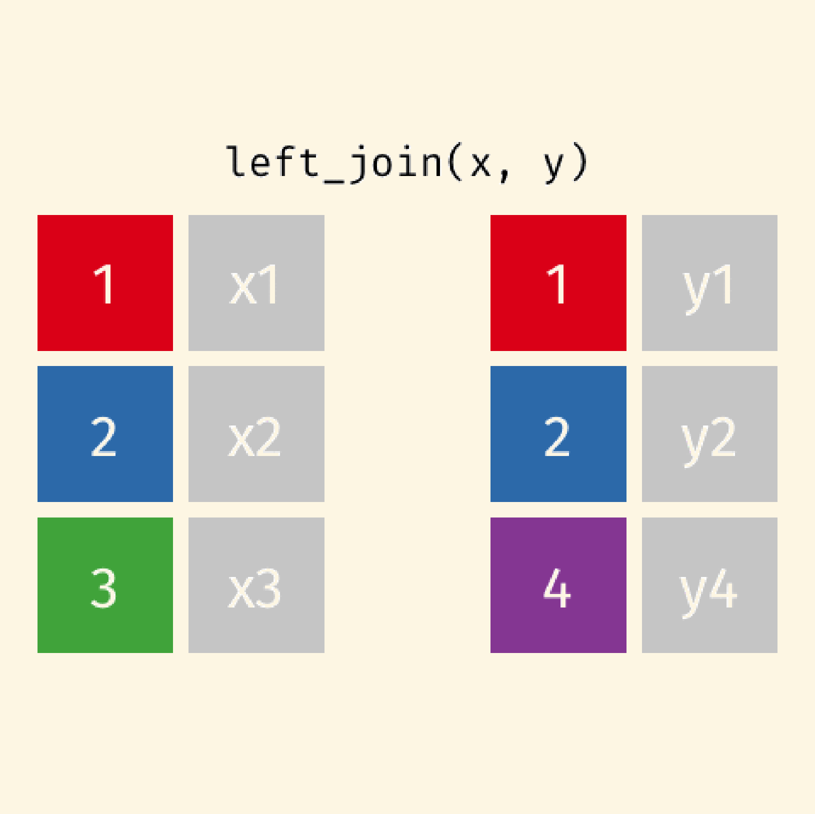
```
x
```

```
## # 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)
```

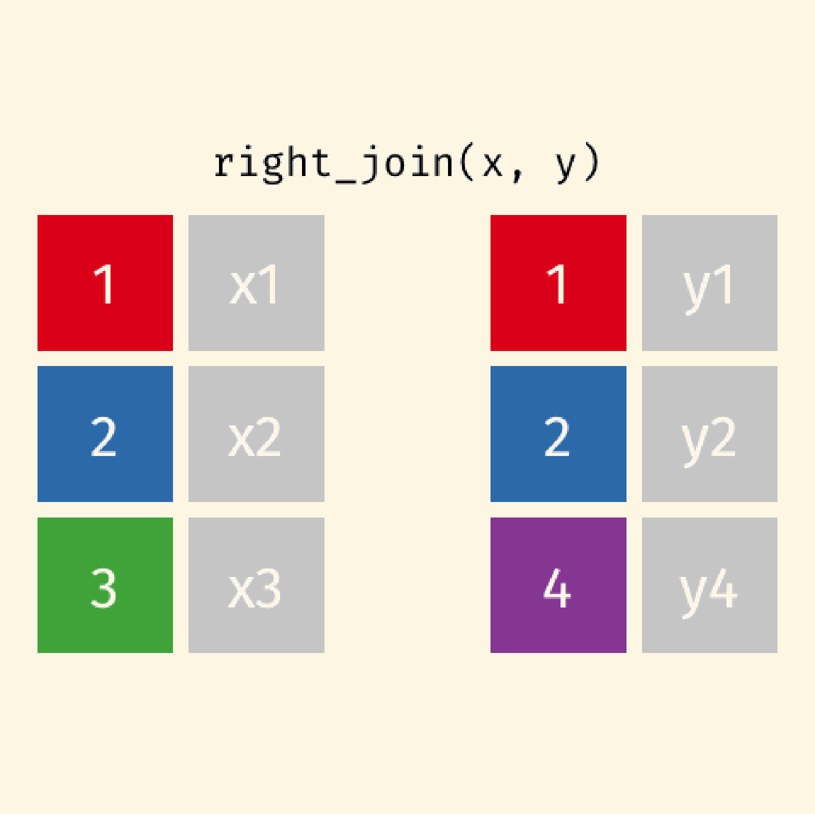
```
## # A tibble: 3 x 3
##      id value_x value_y
##   <dbl> <chr>   <chr>  
## 1     1 x1      y1     
## 2     2 x2      y2     
## 3     3 x3      <NA>   
```

left_join()

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

```
## # A tibble: 10 x 4
##    name               profession                   birth~1 death~2
##    <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 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 B~     1947      NA
## 10 Jennifer Doudna    Biochemist                      1964      NA
## # ... with abbreviated variable names 1: birth_year,
## #   2: death_year
```

right_join()



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

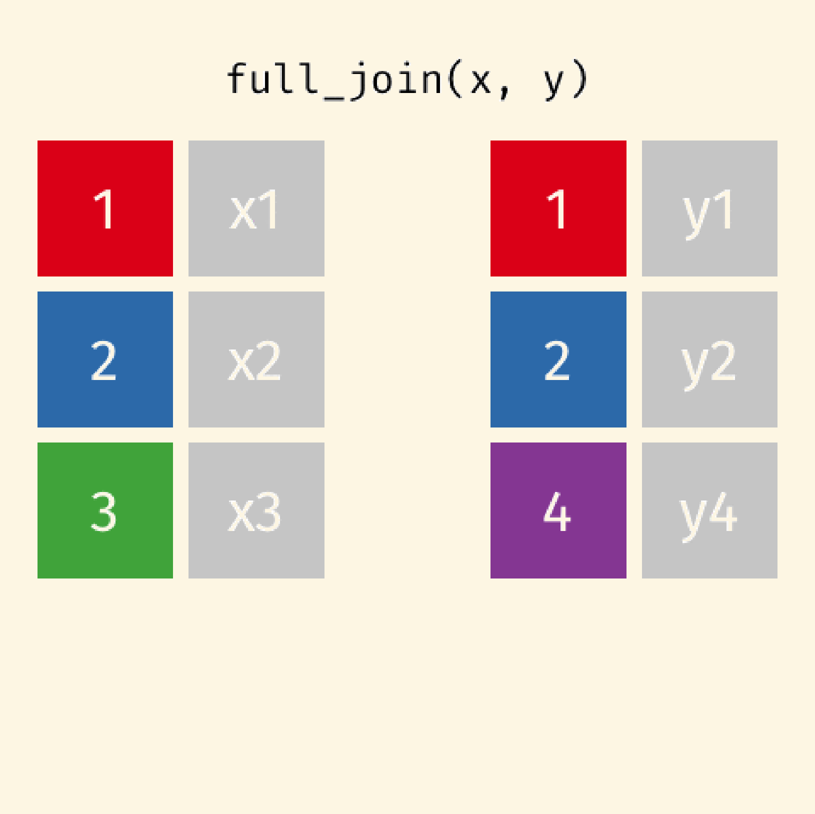
```
## # A tibble: 3 x 3
##      id value_x value_y
##   <dbl> <chr>   <chr>  
## 1     1 x1      y1     
## 2     2 x2      y2     
## 3     4 <NA>    y4     
```

right_join()

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

```
## # A tibble: 8 x 4
##   name               profession                   birth~1 death~2
##   <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 Bi~    1947      NA
## 8 Jennifer Doudna    Biochemist                      1964      NA
## # ... with abbreviated variable names 1: birth_year,
## #   2: death_year
```

full_join()



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

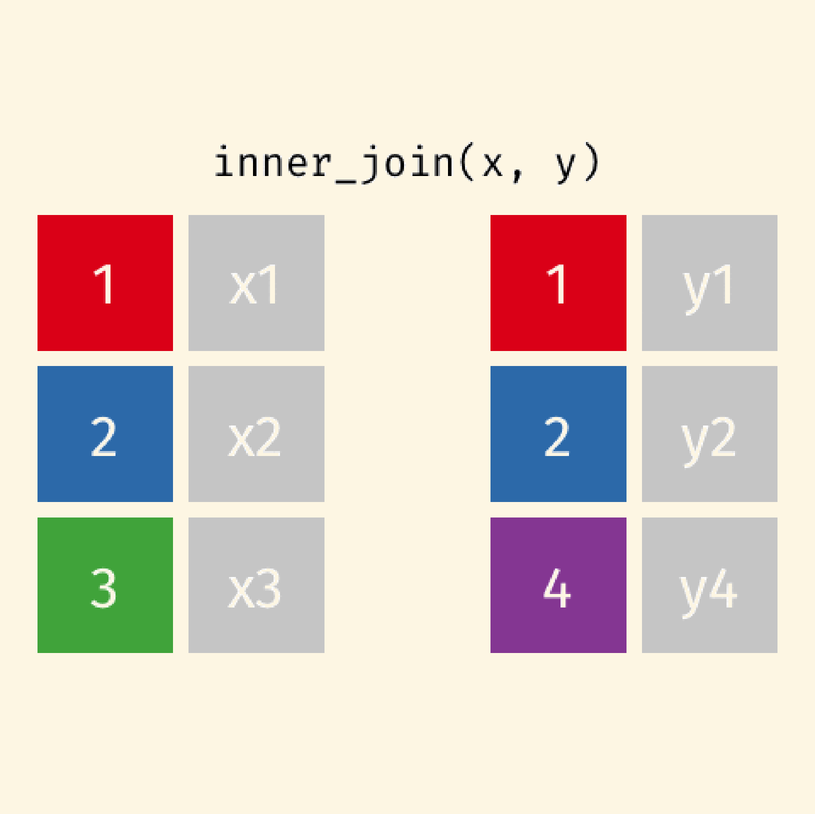
```
## # A tibble: 4 x 3
##      id value_x value_y
##   <dbl> <chr>   <chr>
## 1     1 x1      y1
## 2     2 x2      y2
## 3     3 x3      <NA>
## 4     4 <NA>    y4
```

full_join()

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

```
## # A tibble: 10 x 4
##    name                birth_year death_year known_for
##    <chr>                    <dbl>      <dbl> <chr>
##  1 Janaki Ammal              1897       1984 hybrid species, biod~
##  2 Chien-Shiung Wu           1912       1997 confim and refine th~
##  3 Katherine Johnson         1918       2020 calculations of orbi~
##  4 Rosalind Franklin         1920       1958 <NA>
##  5 Vera Rubin                1928       2016 existence of dark ma~
##  6 Gladys West               1930         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         NA first computer algor~
## 10 Marie Curie                 NA         NA theory of radioactiv~
```

inner_join()



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

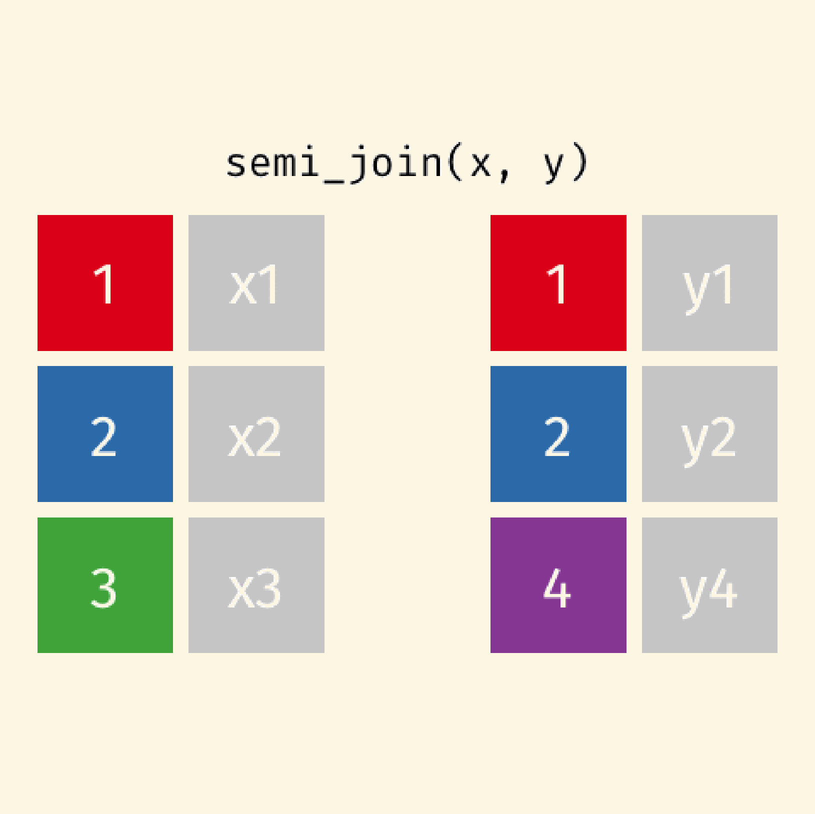
```
## # A tibble: 2 x 3
##      id value_x value_y
##   <dbl> <chr>   <chr>  
## 1     1 x1      y1     
## 2     2 x2      y2     
```

inner_join()

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

```
## # A tibble: 7 x 4
##   name               birth_year death_year known_for
##   <chr>                   <dbl>      <dbl> <chr>
## 1 Janaki Ammal             1897       1984 hybrid species, biodi~
## 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~
## 5 Gladys West              1930         NA mathematical modeling~
## 6 Flossie Wong-Staal       1947         NA first scientist to cl~
## 7 Jennifer Doudna          1964         NA one of the primary de~
```

semi_join()



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

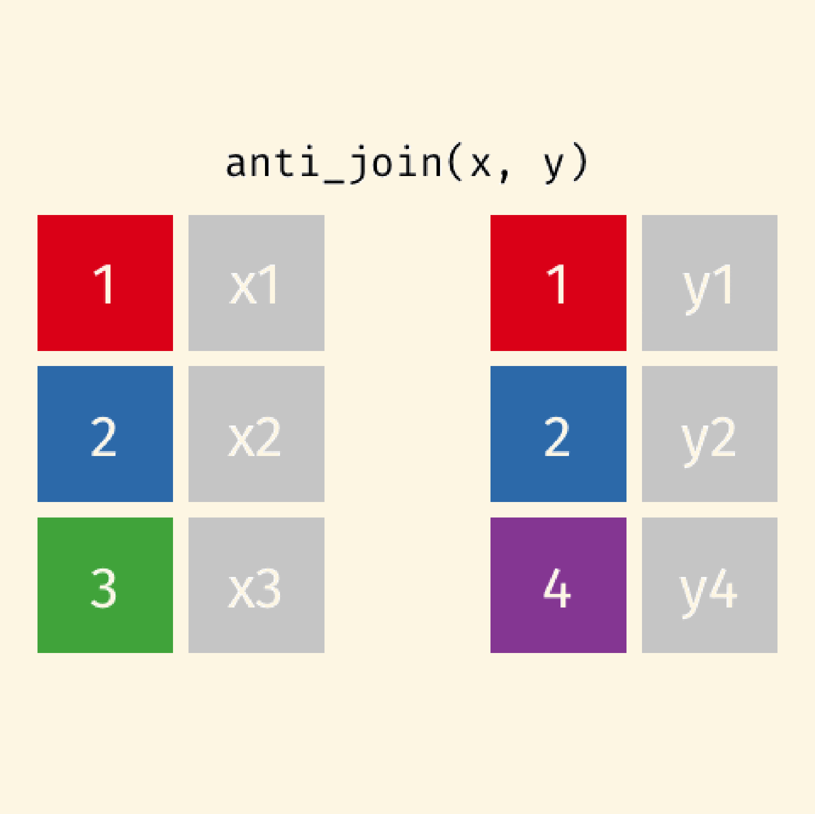
```
## # A tibble: 2 x 2
##      id value_x
##   <dbl> <chr>  
## 1     1 x1     
## 2     2 x2     
```

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()



1	x1
2	x2
3	x3

1	y1
2	y2
4	y4

```r
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~1 death~2 known~3
##    <chr>              <chr>                 <dbl>   <dbl> <chr>
##  1 Ada Lovelace       Mathematician            NA      NA first ~
##  2 Marie Curie        Physicist and Chem~      NA      NA theory~
##  3 Janaki Ammal       Botanist               1897    1984 hybrid~
##  4 Chien-Shiung Wu    Physicist              1912    1997 confim~
##  5 Katherine Johnson  Mathematician          1918    2020 calcul~
##  6 Rosalind Franklin  Chemist                1920    1958 <NA>
##  7 Vera Rubin         Astronomer             1928    2016 existe~
##  8 Gladys West        Mathematician          1930      NA mathem~
##  9 Flossie Wong-Staal Virologist and Mol~    1947      NA first ~
## 10 Jennifer Doudna    Biochemist             1964      NA one of~
## # ... with abbreviated variable names 1: birth_year,
## #   2: death_year, 3: known_for
```

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

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

```r
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
```

Student records

In class

Survey missing

Dropped

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

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

Student records

In class   Survey missing   **Dropped**

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

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

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

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

```r
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
```

```r
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
```

Grocery sales

Total revenue

Revenue per customer

```r
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   
```

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

```
## # A tibble: 2 x 2
##   customer_id total_revenue
##         <dbl>         <dbl>
## 1           1          1.95
## 2           2          3.8 
```