Joining and Pivoting

Advanced data manipulation

Download the section 6 .Rmd handout to `STAT240/lecture/06-join-pivot`.

Material in this section is covered by Chapter 8 on the notes website.

Joining combines information from two dataframes (e.g. the `produce` from last section).

Two types:

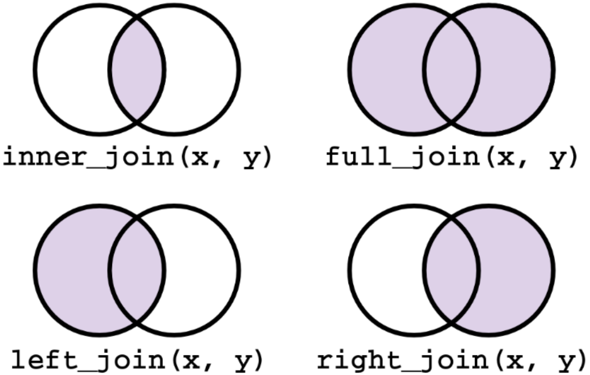
- **Mutating joins** append columns together
- **Filtering joins** keep or delete rows based on another df

Mutating join arguments:

- Two data frames
- `by =` Names of columns to join

`left_join`, `right_join`, `inner_join`, `full_join`

Which dataset is given “priority’?



from tavareshugo.github.io

`left_join(x, y)` keeps all rows in `x`, regardless of what `y` looks like.

- `x` is "nailed down"
- Then `y` columns are added
- Can induce `NA` values of `y`'s columns

`right_join(x, y)` keeps all rows in `y`.

Be mindful of column names!

- The matching column might have a different name in `x` and `y`
- If we don't provide `by`, R will try to match names

If the dataframes have no column names in common, and `by` is not given, we get an error.

`full_join` keeps all rows from both dataframes.

- Like `left_join`, but non-matching rows get added anyway
- Can induce `NA` values for columns in `x` or `y`

This is not the same as “stacking” the dataframes, which is done with `bind_rows`.

inner_join keeps only rows that are in both dataframes.

- Does not induce NA values

The order of arguments x and y does not matter for full_join or inner_join.

Predict what will happen when joining the band instruments and band members datasets.

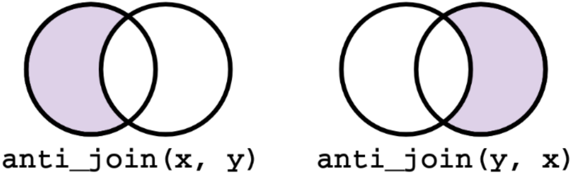
- How many rows will there be?
- How many columns will there be?
- Will there be `NA` values?

Uncomment the lines to see if you were right.

Filtering joins remove rows of the `x` dataset.

- `semi_join` returns the rows in `x` that also appear in `y`
- `anti_join` returns the rows in `x` that don't appear in `y`

No columns from `y` appear in the output.



from tavareshugo.github.io

Now predict the output when filter-joining the band instruments and band members datasets.

- How many rows will there be?
- How many columns will there be?
- Will there be `NA` values?

Uncomment the lines to see if you were right.

Pivoting changes the shape of the dataframe while retaining all of its information.

Datasets can be `long` or `wide`, depending on how we want to structure the rows.

`pivot_longer` increases rows and reduces columns.

- First argument: dataframe
- Second: existing column names

All of the specified columns will be merged into one long column, which we can optionally name.

`pivot_wider` decreases rows and increases columns.

- First argument: dataframe
- Second: column we want to split
- Third: values to population the new columns

The names of the columns come from the 2nd argument, and the values come from the 3rd argument.