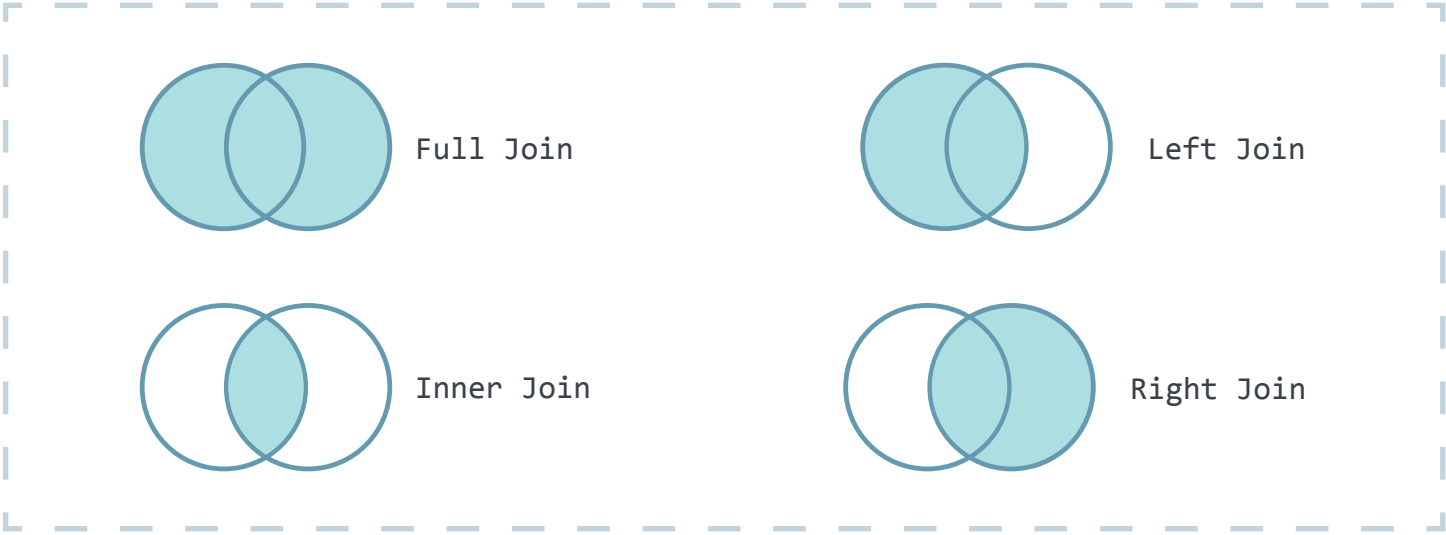
TOOL

Using Join Functions



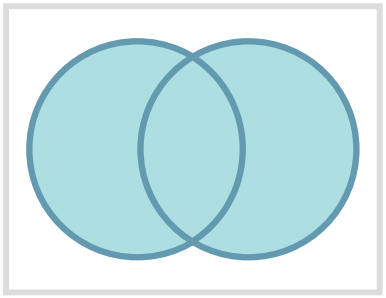
Use this tool to help you decide which join function to use to combine two data sets. Join functions are helpful since you often need information from multiple data sets to perform your analysis. For each of the join functions presented below, there are three inputs: the names of the two data frames we are joining and the name of the variable by which we're joining.

In all of the examples below, we join the **`langDiv`** (language diversity) and **`gdp`** (gross domestic product) data sets according to the variable Country. To set the data up for use with this tool, use the following code:

```
library(tidyverse)

langDiv <- read.csv("language_diversity.csv")

gdp <- read.csv("country_gdp_data.csv", header = TRUE)
```



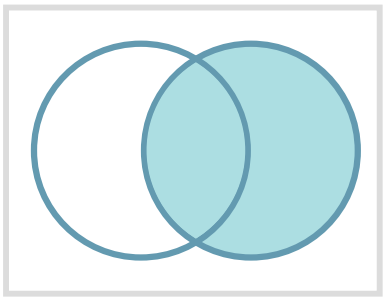
Full Join

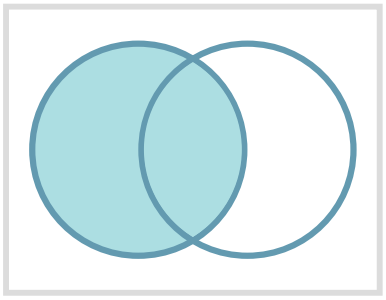
The `full_join()` function keeps:

- All of the rows from both data sets.
- All of the columns from both data sets.

Here we join the `langDiv` and `gdp` data frames by the Country variable, which means the rows of the data sets are combined when they describe the same country:

```
full_join(x = langDiv, y = gdp, by = "Country")
```





Right Join and Left Join

The `right_join()` function keeps:

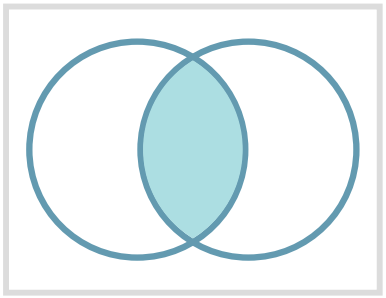
- All of the rows from the "right" data set (the one listed to the right within the join function).
- All of the columns from both data sets.

The `left_join()` function keeps:

- All of the rows from the "left" data set (the one listed to the left within the join function).
- All of the columns from both data sets.

For example, if we want to keep the countries for which we have GDP information (and language diversity is of secondary importance), then we could use `right_join()` with GDP as the rightmost data set. This will keep all of the rows (countries) that have GDP information and will join any available language diversity information to the appropriate rows.

```
right_join(x = langDiv, y = gdp, by = "Country")
```



Inner Join

The `inner_join()` function keeps:

- All of the rows that have matching values in the two data sets.
- All of the columns from both data sets.

`inner_join()` is useful if we want to keep only those rows (countries) for which we have information in both data sets (i.e., countries for which we have both language diversity and GDP information).

```
inner_join(x = langDiv, y = gdp, by = "Country")
```