Joining Datasets – Part 1

Example:

Dataset With Scores

ID	Score
15672	800
16892	"Issue"
56749	650
85413	200

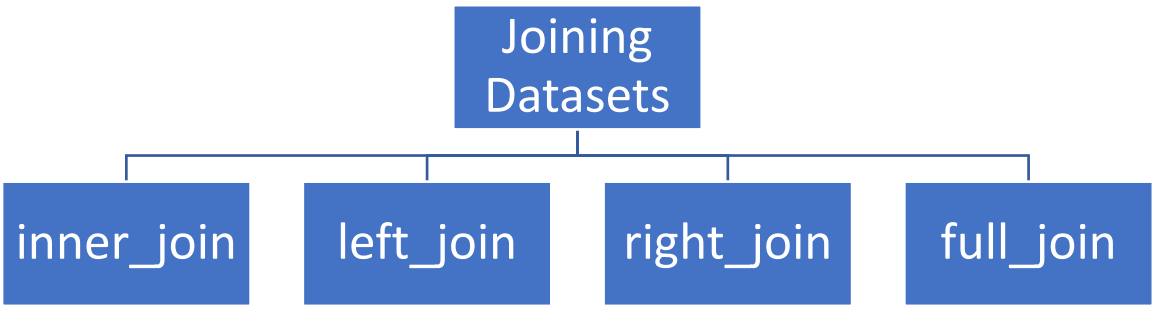
Dataset With Names

ID	Name
15672	Deborah H.
16892	John D.
56749	Errol M.
85413	Juan O.

ID	Name	Score
15672	Deborah H.	800
16892	John D.	"Issue"
56749	Errol M.	650
85413	Juan O.	200

Joins: In the `tidyverse`, joining two datasets together is a way to combine data from different sources based on common variables.

The `dplyr` package within the tidyverse provides functions to perform these joins.



Inner Join: An inner join combines rows from two datasets where there's a match between the specified variables.

- Rows with no matching values are excluded.
- Inner joins return results if the keys are matched in BOTH tables.

Command Illustration

```
new_dataframe_name <- dataframe_name_1 %>%
     inner_join(dataframe_name_2, c("colname_1" = "colname_2"))
```

For the illustration examples, assume the two following dataframes:

Illustration_Data_1

Name	Age	num_kids
Val	18	1
Derek	25	0
Whitney	30	2
Daniella	45	1

Illustration_Data_2

First_Name	Last_Name	Gender
Val	Chmerkovskiy	Male
Derek	Hough	Male
Whitney	Carson	Female
Sasha	Farber	Male
Daniella	Karagach	Female
Lindsay	Arnold	Female
Mark	Ballas	Male

Example 1: Perform an inner join between Illustration_Data_1 and Illustration_Data_2

```
example_1 <- Illustration_Data_1 %>%
  inner_join(Illustration_Data_2, c("Name" = "First_Name"))
```


For the illustration examples, assume the two following dataframes:

Illustration_Data_1

Name	Age	num_kids
Val	18	1
Derek	25	0
Whitney	30	2
Daniella	45	1

Illustration_Data_3

Name	Last_Name	Car
Val	Chmerkovskiy	Mercedes
Val	Chmerkovskiy	Tesla
Val	Chmerkovskiy	Audi
Derek	Hough	Ferrari
Lindsay	Arnold	Tesla
Mark	Ballas	BMW

Example 2: Perform an inner join between Illustration_Data_1 and Illustration_Data_3

```
example_2 <- Illustration_Data_1 %>%
  inner_join(Illustration_Data_3, c("Name" = "Name"))
```


It is always a good idea to carefully check that the number of rows returned by a join operation is what you expected. In particular, you should carefully check for rows in one table that matched to more than one row in the other table.

- Inspect the column by which you are joining.

```
nrow(Illustration_Data_1)
```
Output:

```
n_distinct(Illustration_Data_1$Name)
```
Output:

```
nrow(Illustration_Data_2)
```
Output:

```
n_distinct(Illustration_Data_2$First_Name)
```
Output:

- Check how many data values from one dataset are in the other dataset.

```
table(Illustration_Data_1$Name %in% Illustration_Data_2$First_Name)
```

Output:

```
table(Illustration_Data_2$First_Name %in% Illustration_Data_1$Name)
```

Output: