Joining Datasets – Part 3

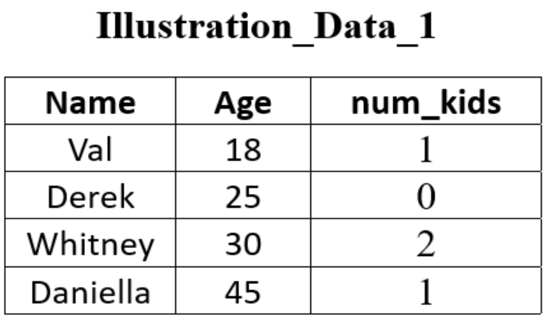
Full Join

Full Join: includes all rows from both datasets. Columns from the dataset with missing values will be filled with NA where there's no match.

Command Illustration

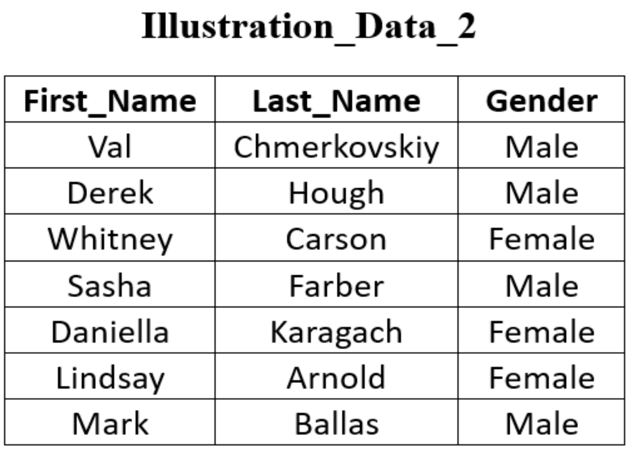
```
new_dataframe_name <- dataframe_1 %>%
     full_join(dataframe_2, c("colname_1" = "colname_2"))
```

For Example 1, 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: Do a full join between Illustration_Data_1 & Illustration_Data_2.

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


For Example 2, 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: Do a full join between Illustration_Data_1 & Illustration_Data_3.

```
example_2 <- Illustration_Data_1 %>%
  full_join(Illustration_Data_3, by = c("First_Name" = "Name"))
```


Join by Multiply Columns

Note: You will not be tested on this material

Join by multiple columns: The `by` argument specifies the column(s) that should be used for matching. These join functions work well when the datasets have a shared column containing the same type of data (e.g., IDs or keys). You can use multiple column names to define the matching conditions.

Command Illustration

```
new_dataframe <- name_dataframe_x %>%
         inner_join(name_dataframe_y, by=c(“x1” = “y1” ,  “x2” =
“y2”))
```

Example 3: We will manually create two datafames in R. Merge the following two dataframes (emp_df & dept_df) by “dept_id” & “dept_branch_id”.

```
example_3 <- emp_df %>%
     inner_join(dept_df,  by  =  c(“dept_id”  =  “dept_id”,
“dept_branch_id” = “dept_branch_id”))
```

emp_df

emp_id	name	superior_emp_id	dept_id	dept_branch_id
1	Smith	-1	10	101
2	Rose	1	20	102
3	Williams	1	10	101
4	Jones	2	10	101
5	Brown	2	40	104
6	Brown	2	50	105

dept_df

dept_id	dept_branch_id	dept_name
10	101	Finance
20	102	Marketing
30	103	Sales
40	104	IT

When you merge these two dataframes, your output will look like:

emp_id	name	superior_emp_id	dept_id	dept_branch_id	dept_name
1	Smith	-1	10	101	Finance
2	Rose	1	20	102	Marketing
3	Williams	1	10	101	Finance
4	Jones	2	10	101	Finance
5	Brown	2	40	104	IT