

Relational Data: Takeaways

by Dataquest Labs, Inc. - All rights reserved © 2020

Syntax

JOINING DATAFRAMES

- Combine two dataframes using an inner join:

```
sat_results %>%  
  inner_join(class_size, by = "DBN")
```

- Combine two dataframes using a left join:

```
sat_results %>%  
  left_join(class_size, by = "DBN")
```

- Combine two dataframes using a right join:

```
sat_results %>%  
  right_join(class_size, by = "DBN")
```

- Combine two dataframes using a full join:

```
sat_results %>%  
  full_join(class_size, by = "DBN")
```

Concepts

- Relational data is data that has a relation to some data in another table.
- A key refers to the variable that connects pairs of tables.
 - Mutating joins add new variables to a dataframe based on matching observations in another dataframe.
 - Inner joins match pairs of variables in two dataframes if their values of the key are the same.
 - Outer joins keep observations that appear in at least one of the two tables you're combining. Outer joins can be divided into three types:
 - Left joins
 - Right joins
 - Full joins
- Performing a left join keeps all observations in the dataframe on the left and drops observations from the dataframe on the right that have no key match.
- Performing a right join keeps all observations in the dataframe on the right and drops observations from the dataframe on the left that have no key.
- Performing a full join keeps all observations from both dataframes and fills in missing variables with

```
NA
```

`inner_join()`

Only common rows

1	X1	Y1
2	X2	Y2

`full_join()`

Left + right
dataset rows

1	X1	Y1
2	X2	Y2
3	X3	
4		Y4

1	X1	+	1	Y1
2	X2		2	Y2
3	X3		4	Y4

1	X1	Y1
2	X2	Y2
3	X3	

Left rows + common
right rows

`left_join()`

1	X1	Y1
2	X2	Y2
4		Y4

Common left rows +
right rows

`right_join()`



DATA CLEANING



ROWS

`filter()`

Filter some of them

`group_by()`

`summarize()`

Aggregate them into new rows

`duplicate()`

`distinct()`

Remove duplicate rows



COLUMNS

`select()`

Select some of them

`mutate()`

Create new columns

`rename()`

Rename columns

`mutate()`

Check/adjust columns types



VALUES

Solve values inconsistencies

Combine/Separate values

Imputation
Dealing with missing values



DATAFRAMES

`bind_rows()`

Combine rows

`bind_columns()`

Combine columns

`pivot_longer()`

Reshape columns into rows

`pivot_wider()`

Reshape rows into columns

Resources

- [Cheat sheet for dplyr join functions](#)



Takeaways by Dataquest Labs, Inc. - All rights reserved © 2020