DATA SCRAPING AND PROCESSING

Selma Hadzic

2024

[INNER]
[LEFT]
[RIGHT]
[FULL]

COMBINE DATASETS

Selma Hadzic

2024

CONCAT

Concatenate the pandas objects (DataFrames and Series):

- either vertically (same columns but more rows): axis='index'
- or horizontally (same rows but more columns) : axis='columns'

By default, it **stacks** the pandas objects vertically.

When stacking objects horizontally (more columns), possibility to use inner join: the result dataframe has only the rows that corresponds to $A \cap B$

CONCAT VERTICALLY

df_a

id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan

df_b

id	name	country
04	Shakuntala DEVI	India
05	Annie EASLY	USA
06	Olubunmi Abidemi FADIPE-JOSEPH	Nigeria

```
vertical_df = pd.concat([df_a, df_b], axis='index')
```

id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan
04	Shakuntala DEVI	India
05	Annie EASLY	USA
06	Olubunmi Abidemi FADIPE-JOSEPH	Nigeria

CONCAT VERTICALLY

df_a

id	name	country	topic
01	Maria AGNESI	Italy	differential & integral calculus
02	Roya BEHESHTI	Iran	algebraic geometry
03	Mei-Chu CHANG	Taïwan	combinatorial number theory

df_b

id	name	country
04	Shakuntala DEVI	India
05	Annie EASLY	USA
06	Olubunmi Abidemi FADIPE-JOSEPH	Nigeria

```
vertical_df = pd.concat([df_a, df_b], axis='index')
```

id	name	country	topic
01	Maria AGNESI	Italy	differential & integral calculus
02	Roya BEHESHTI	Iran	algebraic geometry
03	Mei-Chu CHANG	Taïwan	combinatorial number theory
04	Shakuntala DEVI	India	NULL
05	Annie EASLY	USA	NULL
06	Olubunmi Abidemi FADIPE-JOSEPH	Nigeria	NULL

CONCAT HORIZONTALLY

df_a

id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan

df_b

id	topic
01	differential & integral calculus
02	algebraic geometry
03	combinatorial number theory

```
horizontal_df = pd.concat([df_a, df_b], axis='columns')
```

id	name	country	topic
01	Maria AGNESI	Italy	differential & integral calculus
02	Roya BEHESHTI	Iran	algebraic geometry
03	Mei-Chu CHANG	Taïwan	combinatorial number theory

CONCAT: AVOID NULL VALUES

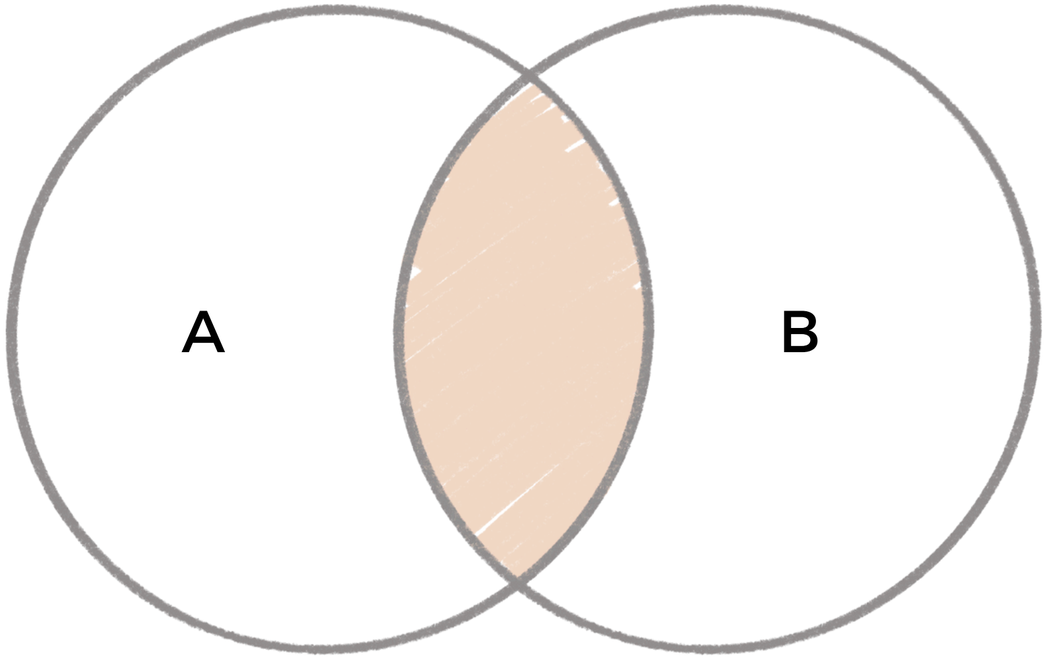
Vertically:

- Both DataFrames must have the same number of columns
- Both DataFrames must have the same column names

Horizontally:

- Both DataFrames have the same number of rows
- Both DataFrames have the same index

JOIN (SQL VOCABULARY)



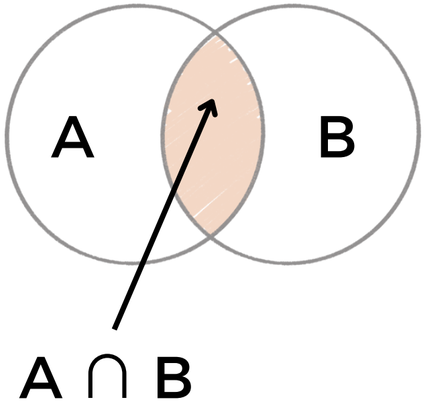
df_a

id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan

df_b

id	topic
02	algebraic geometry
03	combinatorial number theory
04	world record of mental calculation

INNER JOIN



df_a

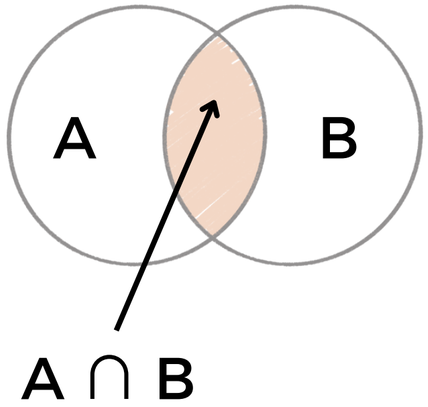
id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan

df_b

id	topic
02	algebraic geometry
03	combinatorial number theory
04	world record of mental calculation

INNER JOIN

```
inner_df = df_a.merge(df_b, how="inner", on= "id")
```



df_a

id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan

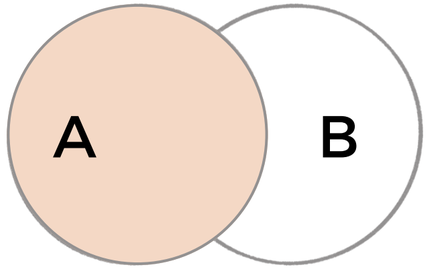
df_b

id	topic
02	algebraic geometry
03	combinatorial number theory
04	world record of mental calculation

A INNER JOIN B
ON id

id	name	country	topic
02	Roya BEHESHTI	Iran	algebraic geometry
03	Mei-Chu CHANG	Taïwan	combinatorial number theory

LEFT JOIN



A + A ∩ B

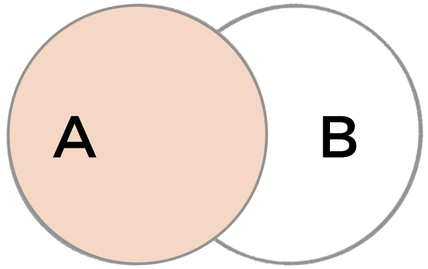
df_a

id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan

df_b

id	topic
02	algebraic geometry
03	combinatorial number theory
04	world record of mental calculation

LEFT JOIN



A + A ∩ B

```
inner_df = df_a.merge(df_b, how="left", on= "id")
```

df_a

id	name	country
01	Maria Agnesi	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu Chang	Taïwan

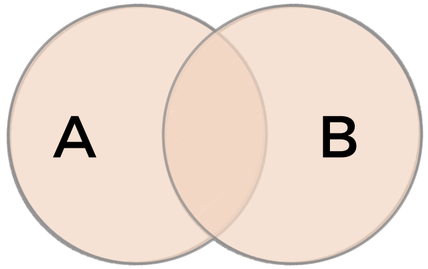
df_b

id	topic
02	algebraic geometry
03	combinatorial number theory
04	world record of mental calculation

A LEFT JOIN B
ON id

id	name	country	topic
01	Maria Agnesi	Italy	NULL
02	Roya BEHESHTI	Iran	algebraic geometry
03	Mei-Chu Chang	Taïwan	combinatorial number theory

FULL JOIN



A ∪ B

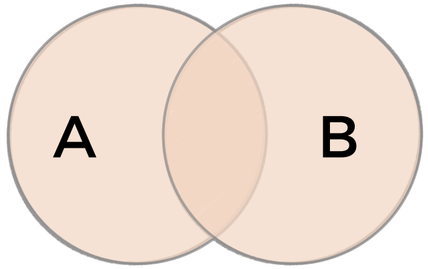
df_a

id	name	country
01	Maria AGNESI	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu CHANG	Taïwan

df_b

id	topic
02	algebraic geometry
03	combinatorial number theory
04	world record of mental calculation

FULL JOIN



A ∪ B

A FULL JOIN B
ON id

df_a

id	name	country
01	Maria Agnesi	Italy
02	Roya BEHESHTI	Iran
03	Mei-Chu Chang	Taïwan

df_b

id	topic
02	algebraic geometry
03	combinatorial number theory
04	world record of mental calculation

```
inner_df = df_a.merge(df_b, how="outer", on= "id")
```

id	name	country	topic
01	Maria Agnesi	Italy	NULL
02	Roya BEHESHTI	Iran	algebraic geometry
03	Mei-Chu Chang	Taïwan	combinatorial number theory
04	NULL	NULL	world record of mental calculation

MERGE (AND JOIN)

Join and **merge** in pandas both perform what is considered a “JOIN” in SQL.
Join uses the index to combine the datasets.
Merge uses one or a combination of columns to combine the datasets.

In pandas, it is more frequent to use **MERGE** because the **KEY** on which to combine the datasets is more often a column than an index.

To avoid NULL values:

- Avoid duplicates in both source DataFrames
- Avoid FULL JOIN (SQL)

SELMA HADZIC