Merging Data

1405 Instructor: Ruiqing (Sam) Cao

Merging Data from Different Sources

Suppose we need to merge two DataFrames `df1` and `df2` by matching the foreign key in `df1` to the primary key in `df2`

There are three possible match outcomes

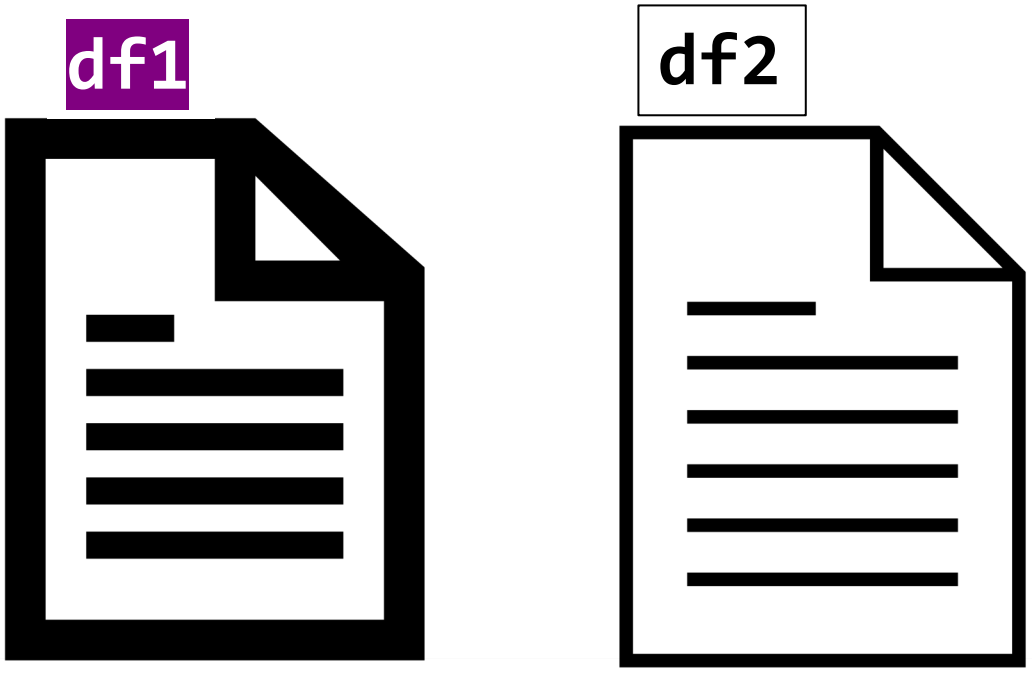
1. primary key value exists in df2 but no foreign key in df1 is matched to it
2. foreign key value exists in df2 but finds no match in df1
3. foreign key value exists in df1 that is matched to a primary key value in df2

Merging Two DataFrames

Suppose we need to merge two DataFrames `df1` and `df2` by matching the foreign key in `df1` to the primary key in `df2`

`df1` : The DataFrame on the ***left***

`df2` : The DataFrame on the ***right***



Three Types of Joins

Join type	Match outcomes to keep	Match outcomes to drop
Inner	3 (matched)	1 (left only) & 2 (right only)
Left	1 (left only) and 3 (matched)	2 (right only)
Right	2 (right only) and 3 (matched)	1 (left only)
Outer	1 (left only), 2(right only), and 3 (matched)	N/A

- ***Left*** join preserves all information in the left-hand-side data (**`df1`**)
- ***Right*** join preserves all information in the right-hand-side data (**`df2`**)
- ***Outer*** join preserves all information both sides (**`df1`** and **`df2`**)

Merging Two DataFrames

	`df1.merge(df2, how, left_on, right_on)`
Arguments	`how`: one of “inner”, “left”, “right”, or “outer” `left_on`: list of column names in `df1` as foreign key `right_on`: list of column names in `df2` as primary key
Returns	The resulting DataFrame from merging `df1` (*left*) with `df2` (*right*)

Merging Two DataFrames: an Example

Foreign keys

Primary keys

	year	month
0	2011	02
1	2011	03
2	2011	04
3	2011	05
4	2011	06
...	...	...
87	2018	05
88	2018	06
89	2018	07
90	2018	08
91	2018	09

`months` (*left*)

	year	month	count1	count0
0	2011	02	0	1
1	2011	11	0	1
2	2012	01	0	1
3	2012	02	0	2
4	2012	04	0	1
...	...	...	...	...
76	2018	04	1	355
77	2018	05	1	393
78	2018	06	0	430
79	2018	07	0	558
80	2018	08	0	1

`data` (*right*)

Left join

	year	month	count1	count0
0	2011	02	0.000	1
1	2011	03	NaN	0
2	2011	04	NaN	0
3	2011	05	NaN	0
4	2011	06	NaN	0
...	...	...	...	...
87	2018	05	1.000	393
88	2018	06	0.000	430
89	2018	07	0.000	558
90	2018	08	0.000	1
91	2018	09	NaN	0

Notes on *Right* Join

```
df1.merge(df2, how='right', left_on, right_on)
```

- If you think that you need to perform right join, you should often switch the order of *df1* and *df2* and perform left join instead

- It's easier to think about merging a *more complex* data object (df2) onto a *simpler* one (df1), as implied by `df1.merge(df2)`