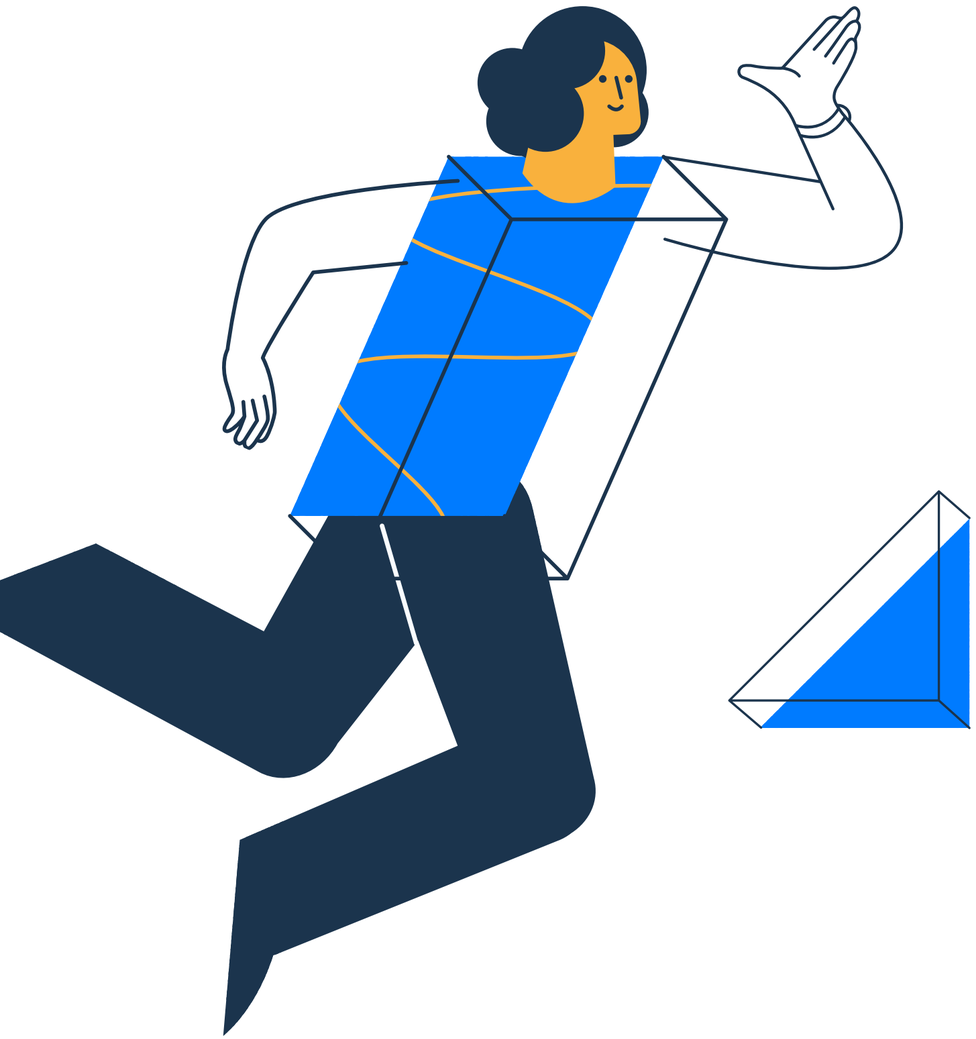
DATA SCIENCE

Federico Baiocco
baioccofede@gmail.com
3512075440



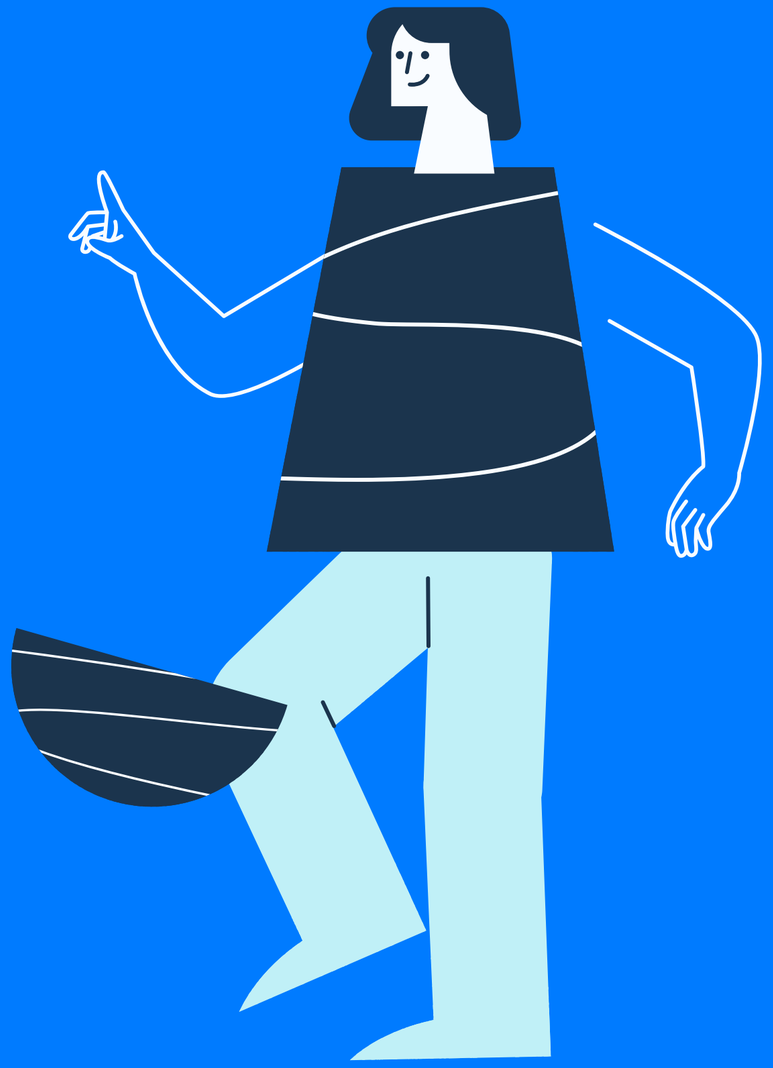
Clase 15 - Agenda

COMBINANDO DATASETS EN PANDAS

- concat
- append
- merge
- join

¿ Dudas de la clase pasada ?

¿ Todos pudieron terminar ?



Combinando datasets con pandas

- Cuando estemos trabajando con datos, en general no vamos a encontrar todo lo que necesitamos en un único dataset (aunque ojalá si).
- Pandas nos permite "combinar" datasets podemos usar simples concatenaciones (al estilo de los append que hacíamos en listas) o operaciones del tipo join a partir de una columna que tengan los datasets en común.
- Pandas nos brinda métodos para hacer todo esto de forma simple

Concat

- Pandas tiene la función **concat que nos permite concatenar 2 o más datasets.**
- **Podemos concatenar datasets de manera horizontal o vertical:**

- Horizontal

	A	B
0	A0	B0
1	A1	B1

	C	D
0	C0	D0
1	C1	D1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1

- Vertical

	A	B
0	A0	B0
1	A1	B1

	A	B
0	A2	B2
1	A3	B3

	A	B
0	A0	B0
1	A1	B1
0	A2	B2
1	A3	B3

Append

- Append nos permite hacer lo mismo que pd.concat pero de manera más simple.
- Debemos tener en cuenta que al hacer append sobre un df, se crea un nuevo df y no se modifica el original
- Es menos eficiente que pd.concat

```
df1     df2     df1.append(df2)
```

	A	B
1	A1	B1
2	A2	B2

	A	B
3	A3	B3
4	A4	B4

	A	B
1	A1	B1
2	A2	B2
3	A3	B3
4	A4	B4

Append

- Append nos permite hacer lo mismo que pd.concat pero de manera más simple.
- Debemos tener en cuenta que al hacer append sobre un df, se crea un nuevo df y no se modifica el original
- Es menos eficiente que pd.concat

```
df1     df2     df1.append(df2)
```

	A	B
1	A1	B1
2	A2	B2

	A	B
3	A3	B3
4	A4	B4

	A	B
1	A1	B1
2	A2	B2
3	A3	B3
4	A4	B4

Merge

- La función merge implementa distintos tipos de "joins" entre dataframes:
    - One to one
    - Many to one
    - Many to many

- El tipo de join que se implementa depende de la forma de los datasets

Merge

One to one

- El tipo de join más simple es el one to one.
- Es similar a la concatenación a nivel columnas que veíamos con el método concat.

`df1`

	employee	group
0	Bob	Accounting
1	Jake	Engineering
2	Lisa	Engineering
3	Sue	HR

`df2`

	employee	hire_date
0	Lisa	2004
1	Bob	2008
2	Jake	2012
3	Sue	2014

	employee	group	hire_date
0	Bob	Accounting	2008
1	Jake	Engineering	2012
2	Lisa	Engineering	2004
3	Sue	HR	2014

Merge

Many to one

- Este tipo de joins ocurre cuando uno de los df contiene duplicados en la columna que utilizamos como key.

`df3`

	employee	group	hire_date
0	Bob	Accounting	2008
1	Jake	Engineering	2012
2	Lisa	Engineering	2004
3	Sue	HR	2014

`df4`

	group	supervisor
0	Accounting	Carly
1	Engineering	Guido
2	HR	Steve

`pd.merge(df3, df4)`

	employee	group	hire_date	supervisor
0	Bob	Accounting	2008	Carly
1	Jake	Engineering	2012	Guido
2	Lisa	Engineering	2004	Guido
3	Sue	HR	2014	Steve

Merge

Many to many

- La key contiene duplicados en ambos dataframes.

`df1`

	employee	group
0	Bob	Accounting
1	Jake	Engineering
2	Lisa	Engineering
3	Sue	HR

`df5`

	group	skills
0	Accounting	math
1	Accounting	spreadsheets
2	Engineering	coding
3	Engineering	linux
4	HR	spreadsheets
5	HR	organization

`pd.merge(df1, df5)`

	employee	group	skills
0	Bob	Accounting	math
1	Bob	Accounting	spreadsheets
2	Jake	Engineering	coding
3	Jake	Engineering	linux
4	Lisa	Engineering	coding
5	Lisa	Engineering	linux
6	Sue	HR	spreadsheets
7	Sue	HR	organization

Merge

Many to many

- La key contiene duplicados en ambos dataframes.

`df1`

	employee	group
0	Bob	Accounting
1	Jake	Engineering
2	Lisa	Engineering
3	Sue	HR

`df5`

	group	skills
0	Accounting	math
1	Accounting	spreadsheets
2	Engineering	coding
3	Engineering	linux
4	HR	spreadsheets
5	HR	organization

`pd.merge(df1, df5)`

	employee	group	skills
0	Bob	Accounting	math
1	Bob	Accounting	spreadsheets
2	Jake	Engineering	coding
3	Jake	Engineering	linux
4	Lisa	Engineering	coding
5	Lisa	Engineering	linux
6	Sue	HR	spreadsheets
7	Sue	HR	organization

Join

- Cuando queremos hacer un "merge" donde las keys son los indices, podemos utilizar el método join:

`df1a`

	group
employee	
Bob	Accounting
Jake	Engineering
Lisa	Engineering
Sue	HR

`df2a`

	hire_date
employee	
Lisa	2004
Bob	2008
Jake	2012
Sue	2014

`df1a.join(df2a)`

	group	hire_date
employee		
Bob	Accounting	2008
Jake	Engineering	2012
Lisa	Engineering	2004
Sue	HR	2014

Abrimos notebook "combinando datasets"