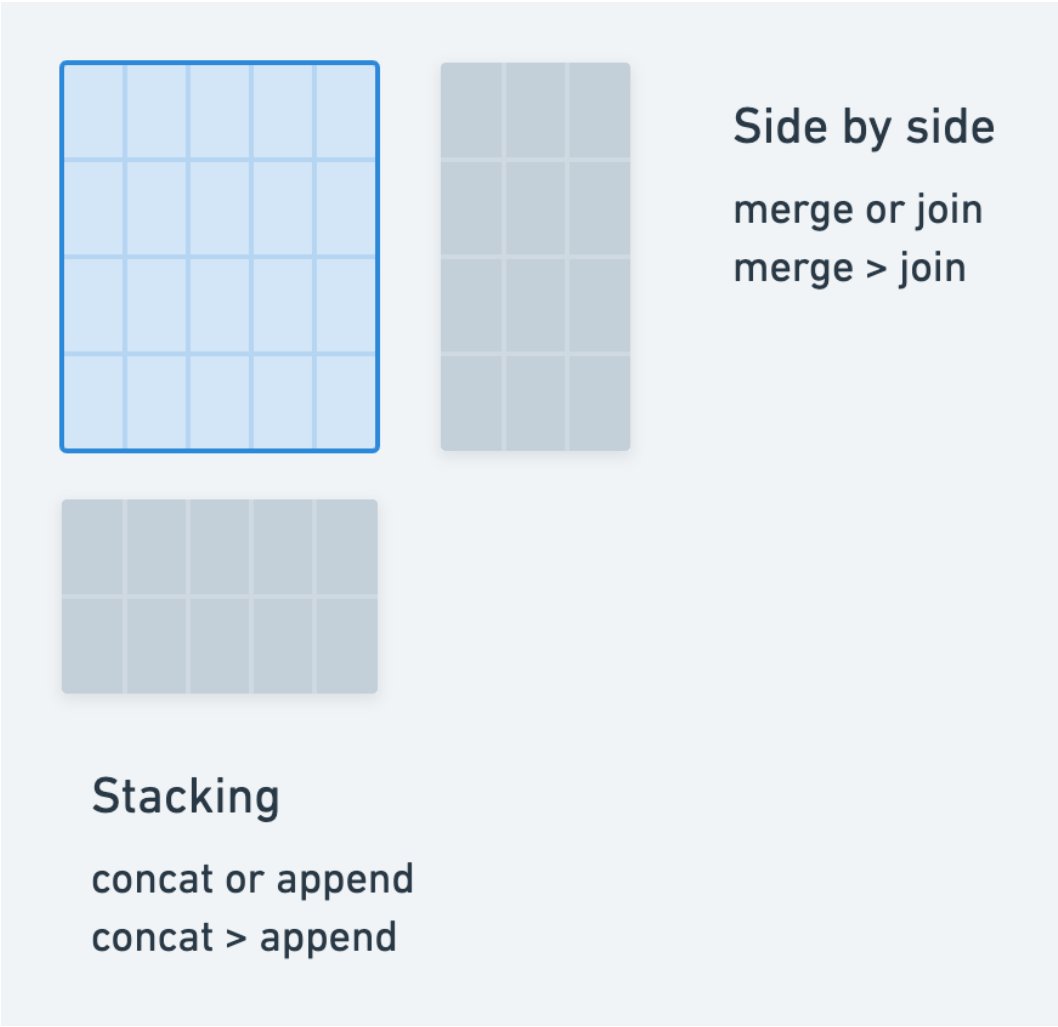
Data Wrangling

Agenda

- How to combine data in pandas
- Pandas datatypes and changing data types in data frames
- How to sort your data
- Pandas limitations and when not to use it
- Time to try out things out yourself

1. Combining Data



1.1 Concat

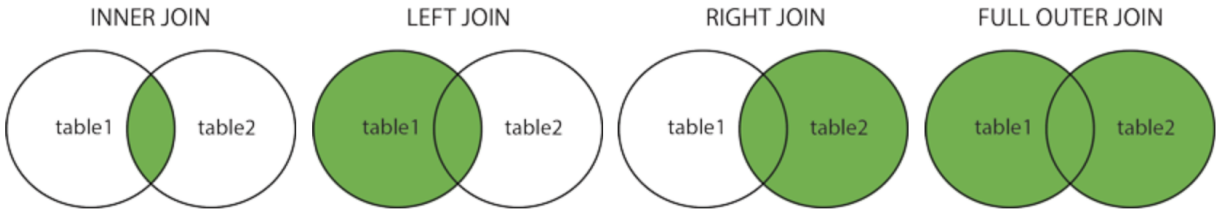
Mostly used to combine together observations with the same information (same columns) that are stored separately. While it can technically be used to add columns, however, we use merge for that.

Stock Price

time	name	close_price	open_price
12/01/1987	microsoft	120	124
13/01/1987	google	131	132
....	...		

1.2 Merge

Used to join together data side by side whether for the same observations or different ones.



MERGE Inner/Outer

df1

Spice	In_stock	Ordered
Cinnamon	0	5
Anise	2	2
Nutmeg	3	0
Clove	7	0

df2

Spice	Price
Cinnamon	3.50
Anis	4.00
Saffron	9.90
Pepper	3.20
Turmeric	6.75

pd.merge(df1, df2, how= 'inner', on='Spice')

Spice	In_stock	Ordered	Price
Cinnamon	0	5	3.50
Anise	2	2	4.00

pd.merge(df1, df2, how= 'outer', on='Spice')

Spice	In_stock	Ordered	Price
Cinnamon	0	5	3.50
Anis	2	2	4.00
Nutmeg	3	0	NaN
Clove	7	0	NaN
Saffron	NaN	NaN	9.90
Pepper	NaN	NaN	3.20
Turmeric	NaN	NaN	6.75

MERGE left/right

df1

Spice	In_stock	Ordered
Cinnamon	0	5
Anise	2	2
Nutmeg	3	0
Clove	7	0

df2

Spice	Price
Cinnamon	3.50
Anis	4.00
Saffron	9.90
Pepper	3.20
Turmeric	6.75

pd.merge(left=df1, right=df2, how= 'left', on='Spice')

Spice	In_stock	Ordered	Price
Cinnamon	0	5	3.50
Anise	2	2	4.00
Nutmeg	3	0	NaN
Clove	7	0	NaN

pd.merge(df1, df2, how= 'right', on='Spice')

Spice	Price	In_stock	Ordered
Cinnamon	3.50	0	5
Anis	4.00	2	2
Saffron	9.90	NaN	NaN
Pepper	3.20	NaN	NaN
Turmeric	6.75	NaN	NaN

Merges dataframes on specific column/index with options left, right, inner, outer.

Tips:

- Be explicit about the `how=` and the `on=` attributes and avoid letting merge decide this for you

2. Data types

Pandas Datatypes

Dtype	Tags
object	text
int64 or int32	integer numbers
float64 or float32	floating point numbers
bool	True/False values
category	Finite list of text values, which could also be ordered
datetime64	Date and time values
timedelta	Difference between date time values
NaN	Nothing

It is important to add the correct data type to your data to optimise for both memory usage and cpu usage when performing operations like group by for example.

3. Sorting

Go to notebook

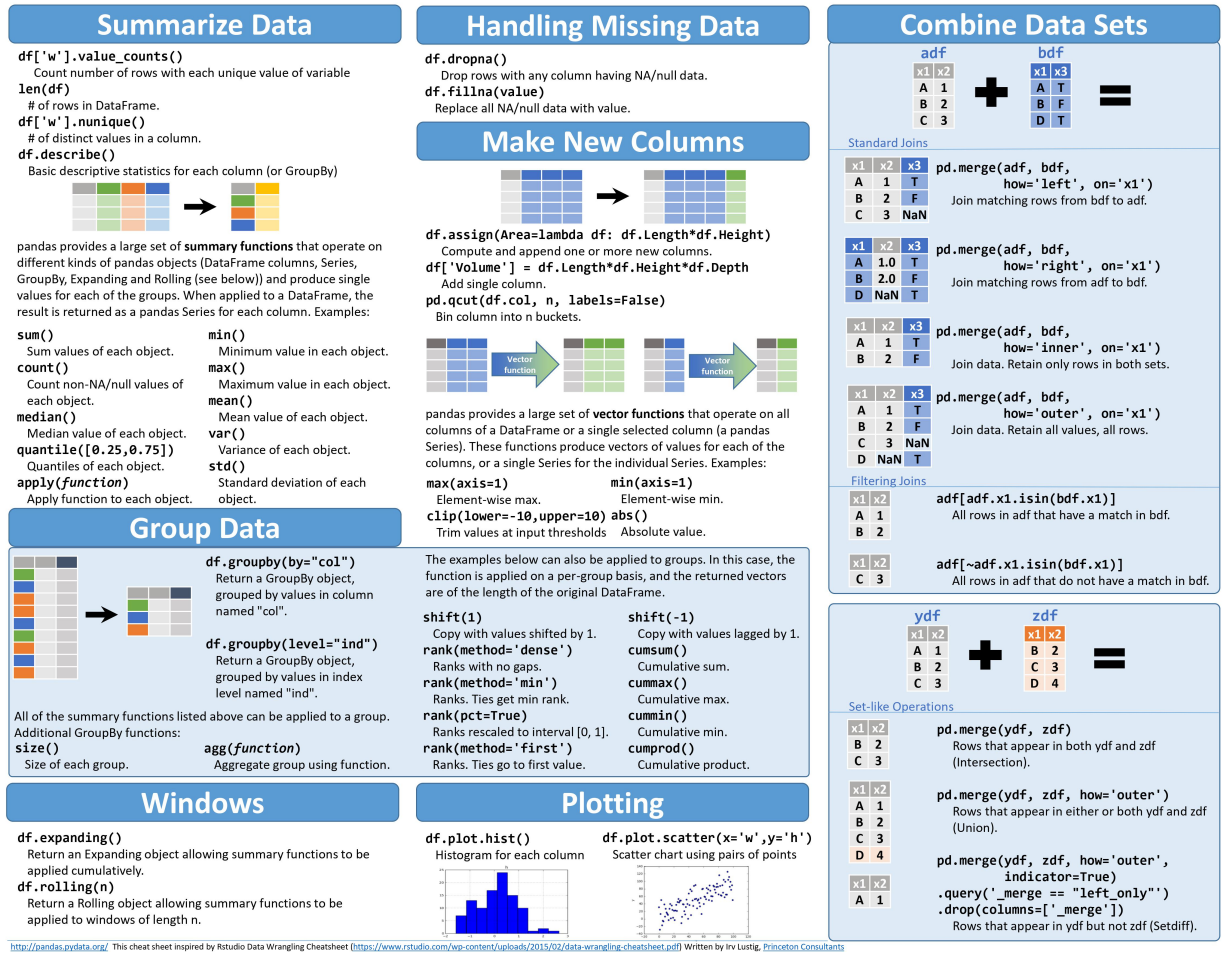
4. Notes

- in-memory pandas vs. databases uses
- Dask

  https://dask.org/

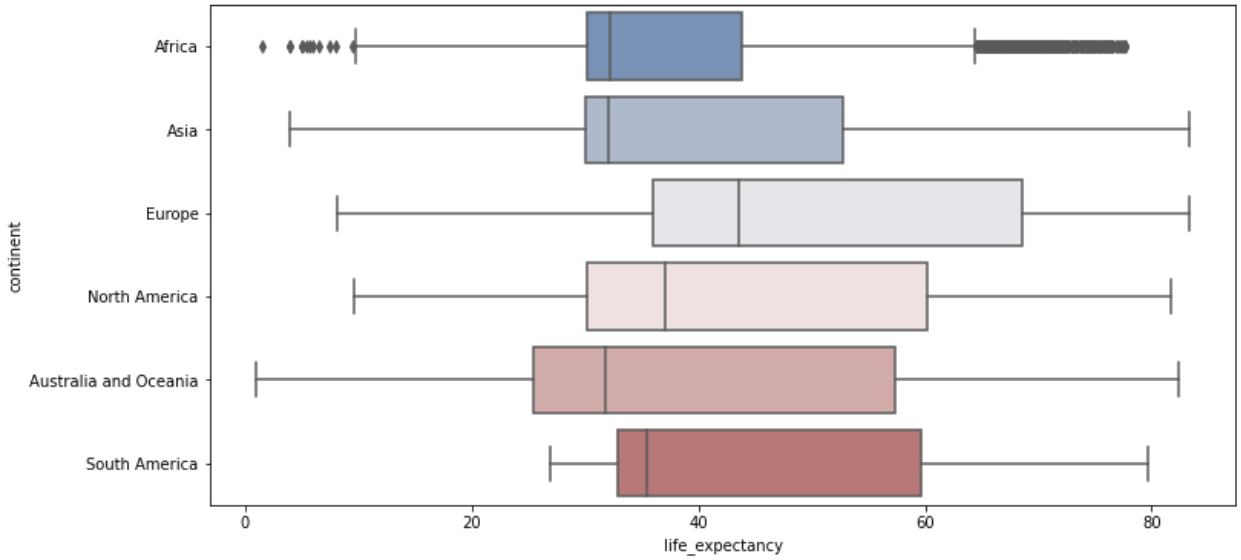
https://stories.dask.org/en/latest/

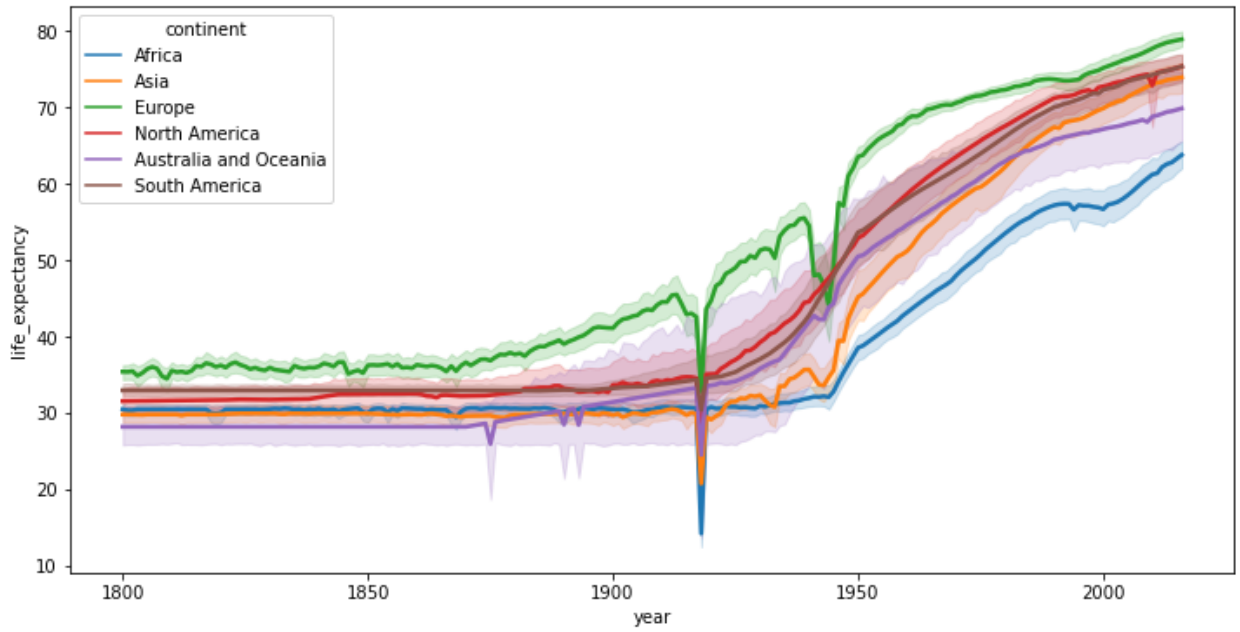
- Cheat Sheet



5. Exercises

Do the gap minder challenge and the results we expect should look something like this:


Africa
Asia
Europe
North America
Australia and Oceania
South America
continent
0
20
40
60
80
life_expectancy


continent
Africa
Asia
Europe
North America
Australia and Oceania
South America
life_expectancy
10
20
30
40
50
60
70
80
1800
1850
1900
1950
2000
year