 # MODULE 1 : DATA EXPLORATION

DATA-MANIPULATION-1

Wrangling data to extract meaningful insights

Club Informatique & Télécom
Data Cell

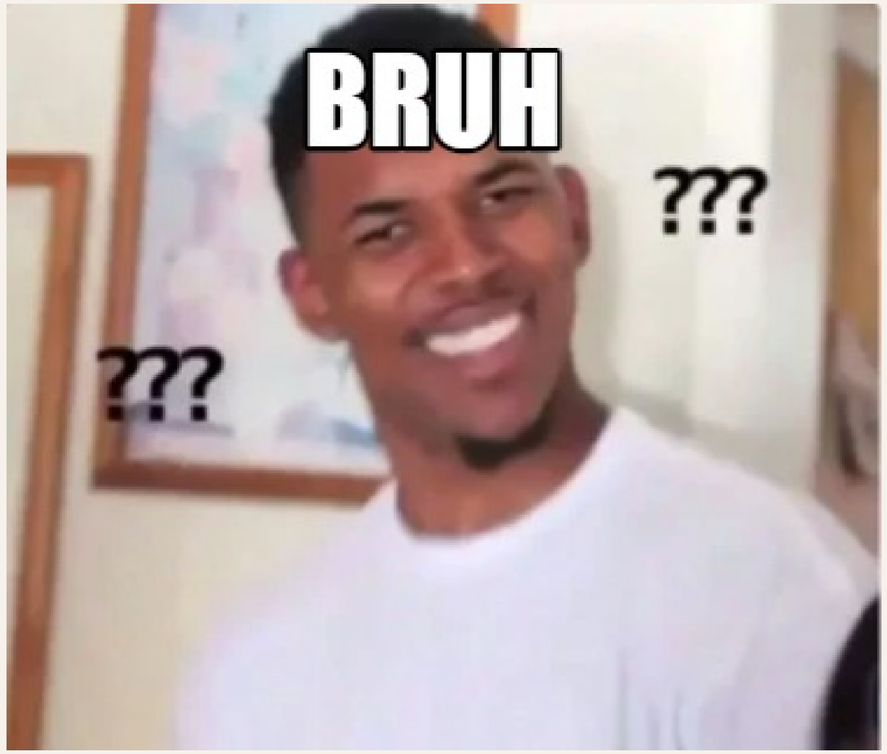
Data Manipulation - 1

- What is data wrangling ?
- Pandas data structures
- Viewing / inspecting data
- Selecting data
- Cleaning data
- Today’s lab : Olympic Data

Data Manipulation

Data wrangling is the process of transforming and structuring data from one raw form into a desired format with the intent of improving data quality and making it more consumable and useful for analytics or machine learning. It’s also sometimes called data munging.

alteryx.com


BRUH
???
???

Data Wrangling

Pandas library

Pandas Data Structures

`pd.Series` + `pd.Series` = `pd.DataFrame`

	apples
0	3
1	2
2	0
3	1

+

	oranges
0	0
1	3
2	7
3	2

=

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Viewing / inspecting data

`df.head(n)`	First n rows of the DataFrame
`df.tail(n)`	Last n rows of the DataFrame
`df.shape`	Number of rows and columns
`df.info()`	Index, Datatype and Memory information
`df.describe()`	Summary statistics for numerical columns
`s.value_counts(dropna=False)`	View unique values and counts
`df.apply(pd.Series.value_counts)`	Unique values and counts for all columns

Selecting data

`df[col]`	Returns column with label col as Series
`df[[col1, col2]]`	Returns columns as a new DataFrame
`s.iloc[0]`	Selection by position
`s.loc['index_one']`	Selection by index
`df.iloc[0,:]`	First row
`df.iloc[0,0]`	First element of first column

Data Cleaning

`pd.isnull()`	Checks for null Values, Returns Boolean Arrray
`df.dropna() / df.dropna(axis=1)`	Drop all rows / columns that contain null values
`df.fillna(x)`	Replace all null values with x
`s.fillna(s.mean())`	Replace all null values with the mean
`s.astype(float)`	Convert the datatype of the series to float
`s.replace([2,3],['two', 'three'])`	Replace all 2 with 'two' and 3 with 'three'
`df.rename(columns={'old': 'new'})`	Selective renaming
`df.set_index('column_one')`	Change the index

Time to practice !