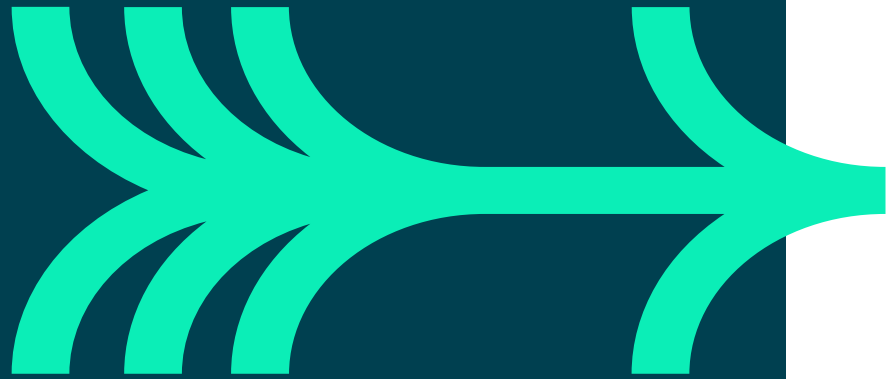
INTRODUCTION TO NUMPY AND PANDAS



The Topic: What?

•An introduction to python libraries for data analysis

•How and when to use Pandas

Applications: Why?

•To understand what libraries already exist for Python that are relevant for data analysis

•To be able to manipulate data in Python so that processes can be later automated

Expectations: Who?

•It is assumed you have used Python previously. This module will focus on how ready-written code can be applied.

•It is assumed you have worked with data previously. It is advantageous if you have previously used SQL.

OBJECTIVES

- Import Numpy and Pandas
- Understand Numpy Arrays
- Create Pandas Data Frames
- Carry out Basic Operations on Data Frames

A SELECTION OF PYTHON LIBRARIES FOR DATA ANALYSIS

- NumPy
  - Arrays
- Pandas
  - provides single-machine DataFrames
- Seaborn & matplotlib
  - visualization
- Spark
  - query over distributed file systems
- plotly
  - interactive visuals
- scipy, sklearn, tensorflow, pytorch, statsmodels
  - scientific & statistical programming

WHAT IS NUMPY?

- NumPy arrays have a fixed size at creation
- The elements in a NumPy array are all required to be of the same data type, and thus will be the same size in memory
- NumPy arrays facilitate advanced mathematical and other types of operations on large numbers of data
- A growing plethora of scientific and mathematical Python-based packages are using NumPy arrays

Why is it fast?

WHAT IS PANDAS?

- Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks.
- It is one of the packages that is built upon Numpy Arrays
- One of the most popular “Data Wrangling” packages in Python and therefore works well with many other Data Science packages in the Python Ecosystem.
- Numpy Arrays → Pandas DataFrames

```
In [33]: data
```

Out[33]:

	Area Abbreviation	Area Code	Area	Item Code	Item	Element Code	Element	Unit	latitude	longitude	...	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009
0	AF	2	Afghanistan	2511	Wheat and products	5142	Food	1000 tonnes	33.94	67.71	...	3249.0	3486.0	3704.0	4164.0	4252.0	4538.0
1	AF	2	Afghanistan	2805	Rice (Milled Equivalent)	5142	Food	1000 tonnes	33.94	67.71	...	419.0	445.0	546.0	455.0	490.0	415.0
2	AF	2	Afghanistan	2513	Barley and products	5521	Feed	1000 tonnes	33.94	67.71	...	58.0	236.0	262.0	263.0	230.0	379.0
3	AF	2	Afghanistan	2513	Barley and products	5142	Food	1000 tonnes	33.94	67.71	...	185.0	43.0	44.0	48.0	62.0	55.0
4	AF	2	Afghanistan	2514	Maize and products	5521	Feed	1000 tonnes	33.94	67.71	...	120.0	208.0	233.0	249.0	247.0	195.0

WHAT CAN PANDAS BE USED FOR?

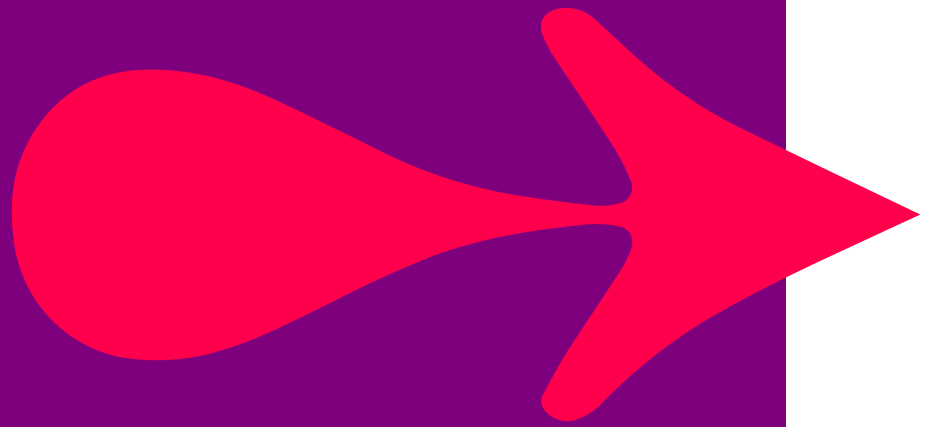
Pandas makes it simple to do many of the time consuming, repetitive tasks associated with working with data, including:

- Data cleansing
- Data fill
- Data normalization
- Merges and joins
- Data visualization
- Statistical analysis
- Data inspection
- Loading and saving data
- And much more

DEMO AND EXERCISES

Open the file: Introduction_to_Pandasipynb

Trainer Demo and Exercises



END OF MODULE

- What advantages do Numpy Arrays have over Python Lists?
- Are there any differences you should take into account when using them?
- How are Numpy Arrays implemented in Pandas?