# Python For Data Analysis

## Introduction/Numpy

Imen Ouled Dlala

imen.ouled_dlala@devinci.fr

September 8, 2022

# Teachers

- Managing Teacher:
  - Imen OULED DLALA
  - Email: imen.ouled_dlala@devinci.fr
- Teachers:
  - Abdellah Sabry
  - Email: sabry.abdellah@gmail.com
  - Yann Kervella
  - Email: yann.kervella.pro@gmail.com

# Schedule

- 6 face to face lectures of 1h30
- Pratical Works (PW) of 1h30 (10 online and 8 face to face ) Following sequence:
  - Lecture1, PW1
  - Lecture 2, PWO1, PW2
  - Lecture 3, PWO2, PWO3, PWO4, PW3
  - Lecture 4, PWO5, PW4
  - Lecture 5, PWO6, PW5
  - Lecture 6, PWO7, PW6, PWO8, PW7, PWO9, PW8, PWO10

# General Plan

1. Introduction
2. Numpy library
3. Pandas library
4. Data analysis and visualization
   - Seaborn, Matplotlib, Bokeh
5. Webscrapping
6. Machine learning and Datasets
   - Scikit-learn
7. API Django / Flask

# Overview

# Overview

# Introduction

What is data analysis ?
A process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making.

Definition by Wikipedia

# Introduction

What is data analysis ?
A process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making.

Definition by Wikipedia

# Introduction

What is data analysis ?
A process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making.

Definition by Wikipedia

# Introduction

Data Analysis Tools

Auto-managed closed tools


Programming languages

# Introduction

Data Analysis Tools

| Auto-managed closed tools | Programming languages |
| --- | --- |
| 👎 Closed Source 🤵 | 👍 Open Source 🤩 |
| 👎 Expensive 💸 | 👍 Free (or very cheap) 🤑 |
| 👎 Limited 😩 | 👎 Extremely Powerful 💪 |
| 👍 Easy to learn 👶 | 👎 Steep learning curve 🧗 |

# Introduction

Why Python for Data Analysis ? why would we choose python over R or Julia?

- Very simple and intuitive to learn
- "correct" language
- powerful libraries (not just for data analysis)
- free and open source
- amazing community, docs and conferences

# Introduction

When to choose R ? Python, sadly, is not always the answer

- When R studio is needed
- When dealing with advanced statistical methods
- When extreme performance is needed

# Introduction

## The data analysis process


- SQL
- Scrapping
- File Formats
  - CSV
  - JSON
  - XML
- Consulting APIs
- Buying Data
- Distributed Databases

- Missing values and empty data
- Data imputation
- Incorrect types
- Incorrect or invalid values
- Outliers and non relevant data
- Statistical sanitization

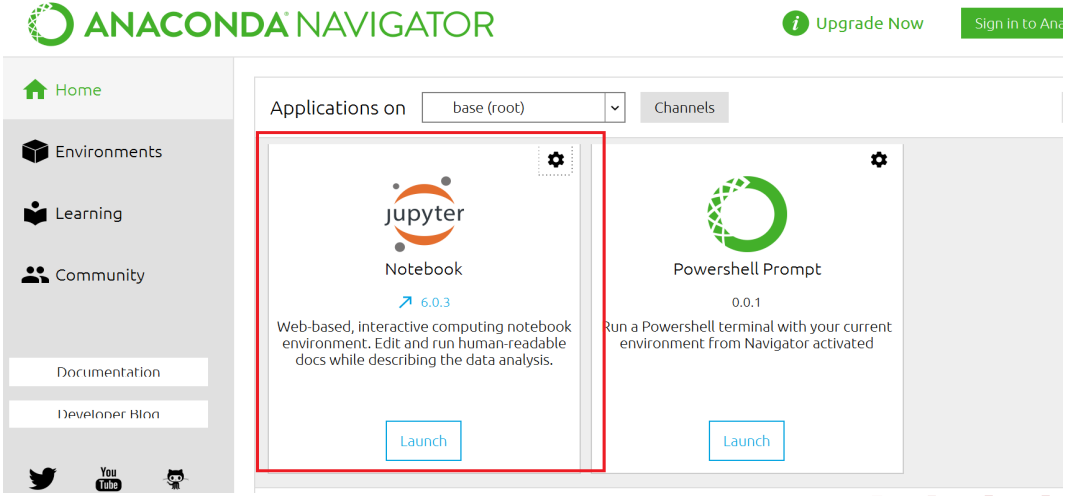
- Hierarchical Data
- Handling categorical data
- Reshaping and transforming structures
- Indexing data for quick access
- Merging, combining and joining data

- Exploration
- Building statistical models
- Visualization and representations
- Correlation vs Causation analysis
- Hypothesis testing
- Statistical analysis
- Reporting

- Building Machine Learning Models
- Feature Engineering
- Moving ML into production
- Building ETL pipelines
- Live dashboard and reporting
- Decision making and real-life tests

# Introduction

## How to use jupyter notebooks



# Overview

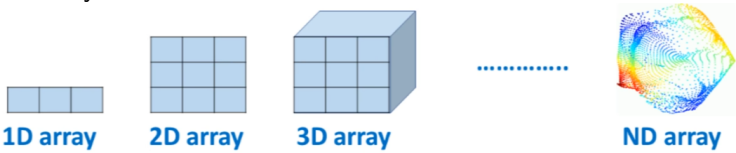
# Introduction to Numpy

Numpy: Numeric computing library

NumPy (Numerical Python) is one of the core packages for numerical computing in Python. Pandas, Matplotlib, Statmodels and many other Scientific libraries rely on NumPy..

It is a library that provides a high-performance multidimensional array object, and tools for working with these arrays.

ND Array



# Introduction to Numpy

## 1Darray

List =

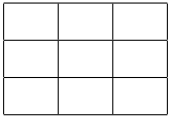
| **index** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|

ndarray =

| **index** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|

ndarray is much faster and better for scientific calculation.

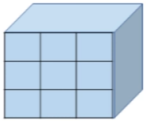
## 2Darray



| | A | B | C | D |
|---|---|---|---|---|
| 1 | Segment | Country | Product | Discount Band |
| 3 | Government | Canada | Carretera | None |
| 4 | Government | Germany | Carretera | None |
| 5 | Midmarket | France | Carretera | None |
| 6 | Midmarket | Germany | Carretera | None |
| 7 | Midmarket | Mexico | Carretera | None |
| 8 | Government | Germany | Carretera | None |
| 9 | Midmarket | Germany | Montana | None |
| 10 | Channel Partners | Canada | Montana | None |
| 11 | Government | France | Montana | None |
| 12 | Channel Partners | Germany | Montana | None |
| 13 | Midmarket | Mexico | Montana | None |
| 14 | Enterprise | Canada | Montana | None |

# Introduction to Numpy

## 3Darray

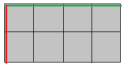


# Introduction to Numpy

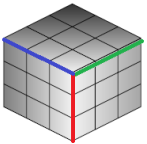
## Dimensions and shapes

ndarray.shape

ndim = 1
shape=(2,)



ndim =2
shape=(2, 4)



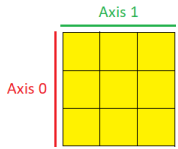
ndim= 3
shape=(3, 3, 3)

shape is a tuple !

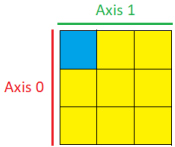
shape[0]=2
shape[1]=4

# Introduction to Numpy

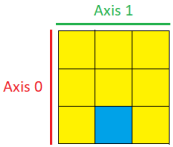
## Indexing and Slicing




A[line,column]



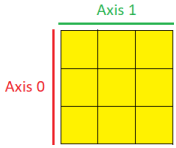

A[0,0]



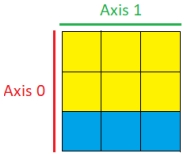

A[2,1]

# Introduction to Numpy

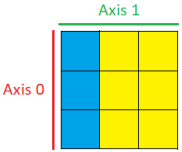
## Indexing and Slicing



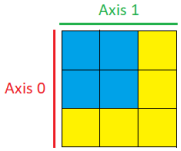

A[start:end,start:end]



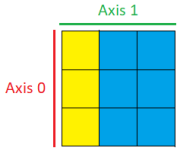

A[2,:]
A[2]




A[:,0]



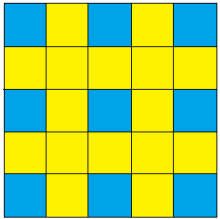

A[0:2,0:2]




A[:,1:]
A[:,-2:]

# Introduction to Numpy

## Indexing and Slicing



A[start:end:step,start:end:step]
A[::2,::2]

# Introduction to Numpy

## Boolean Indexing

$A =$

| | | |
|---|---|---|
| 5 | 4 | 5 |
| 4 | 4 | 4 |
| 5 | 4 | 5 |

$A < 5 \rightarrow$

| | | |
|---|---|---|
| F | T | F |
| T | T | T |
| F | T | F |

$A[A < 5] = 10 \rightarrow$

| | | |
|---|---|---|
| 5 | 10 | 5 |
| 10 | 10 | 10 |
| 5 | 10 | 5 |

# Introduction to Numpy

## Statistics/Mathematical functions

Statistics:

https://docs.scipy.org/doc/numpy-1.13.0/reference/routines.statistics.html

Mathematical functions:

https://docs.scipy.org/doc/numpy-1.13.0/reference/routines.math.html

# End

Good Lecture!