# Manipulation de données avec Python
## La librairie PANDAS

Rodrique KAFANDO
Doctorant en Science de Données & IA
Email : kafando.rodrique@gmail.com

Juillet 2021

université virtuelle Burkina Faso

# SOMMAIRE

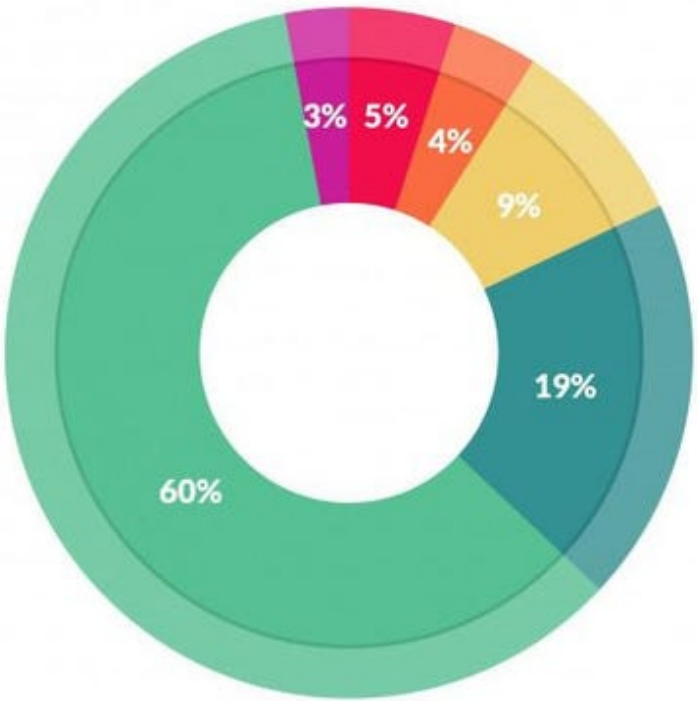
# Contexte

## À propos de ce module

- Cours : Introduction à la science de données
  - **Module 1 :** Langages et Outils de programmations appliqués à la Science de Données -> Zakaria KINDA
  - **Module 2 :** Manipulation de données -> Rodrique KAFANDO

# Contexte

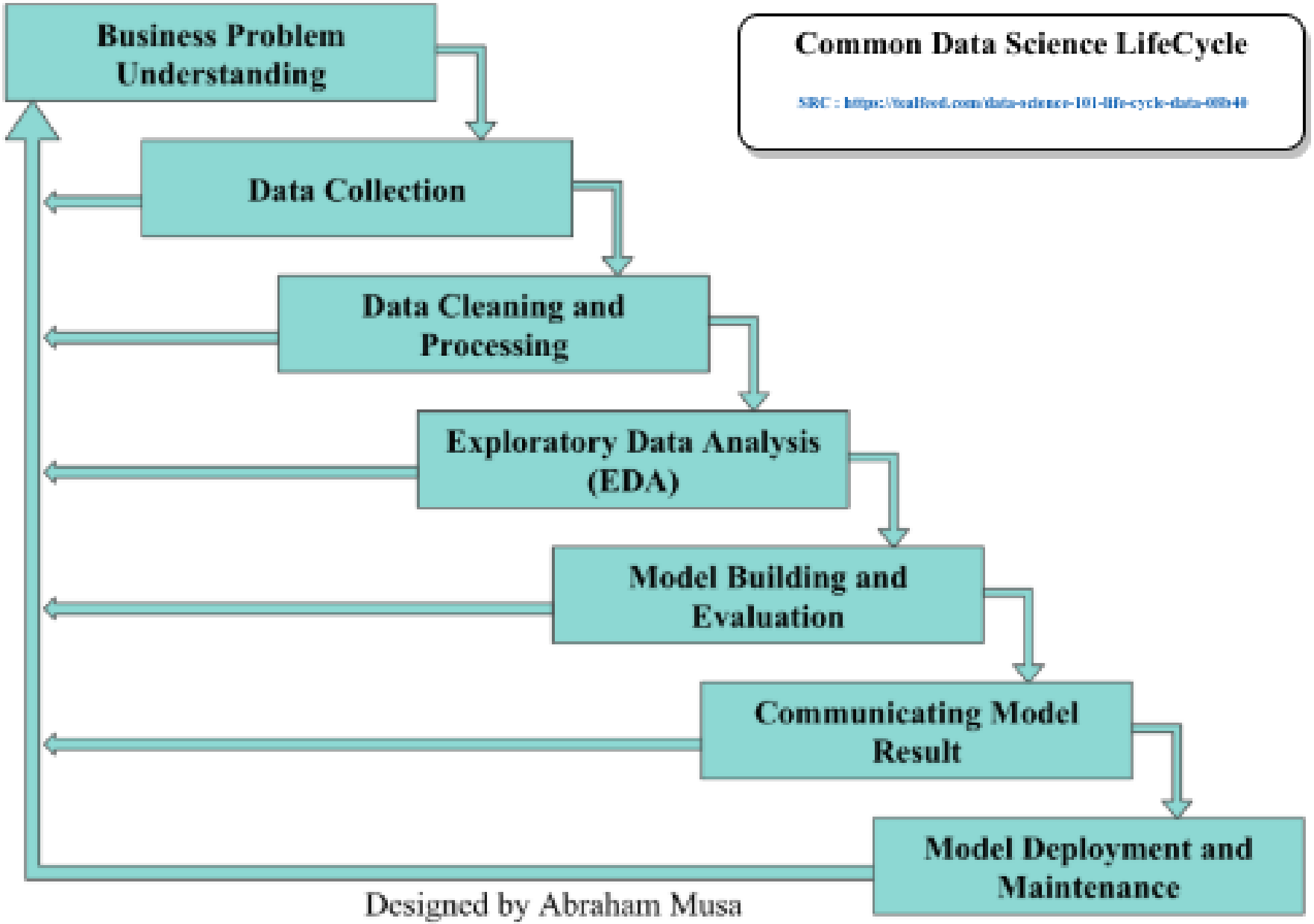
## Bon à savoir! D'après Forbes



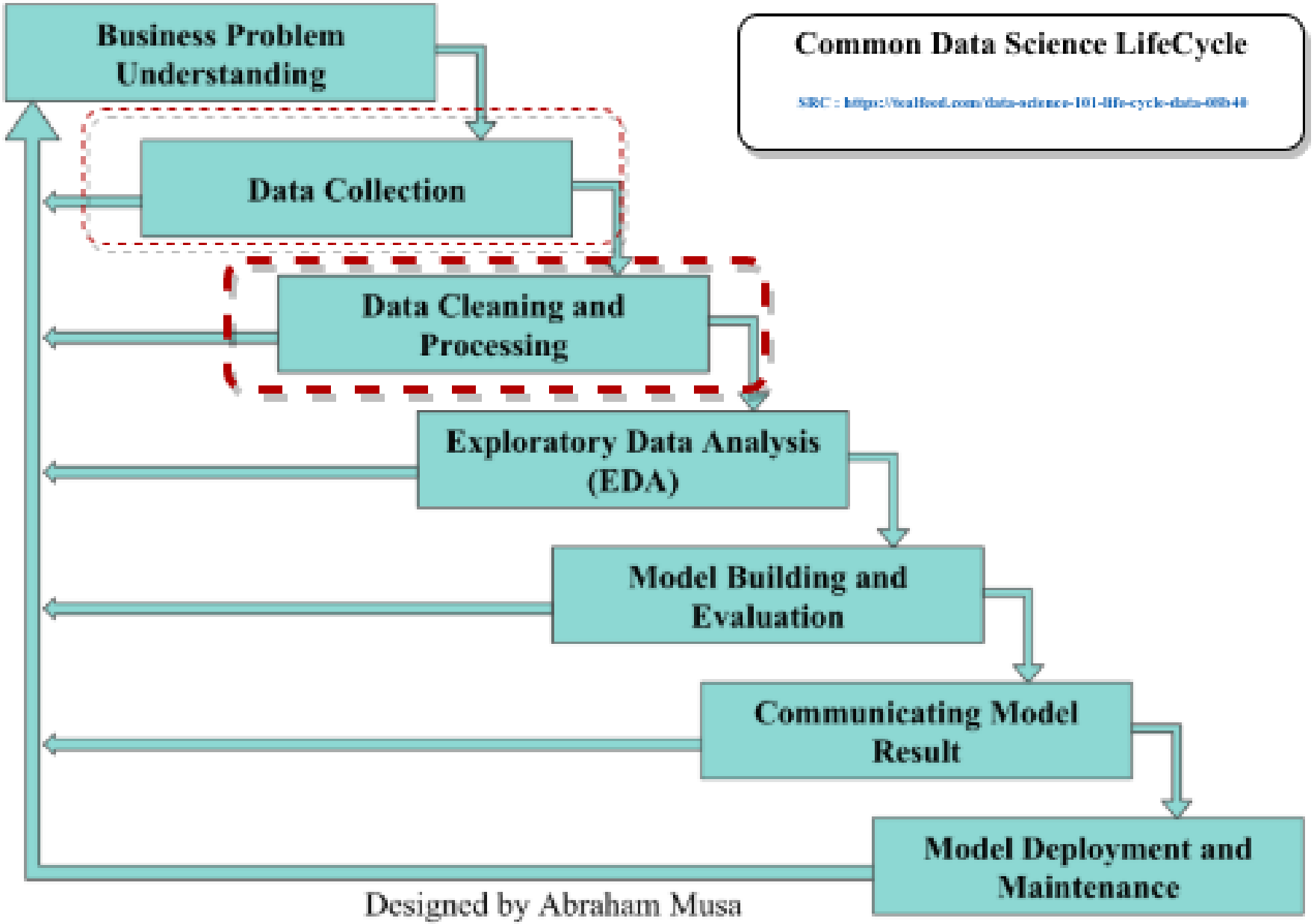

**What data scientists spend the most time doing**

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

# Contexte 3


Common Data Science LifeCycle
SRC : https://kualfood.com/data-science-101-life-cycle-data-08b48
Business Problem Understanding
Data Collection
Data Cleaning and Processing
Exploratory Data Analysis (EDA)
Model Building and Evaluation
Communicating Model Result
Model Deployment and Maintenance
Designed by Abraham Musa

# Contexte

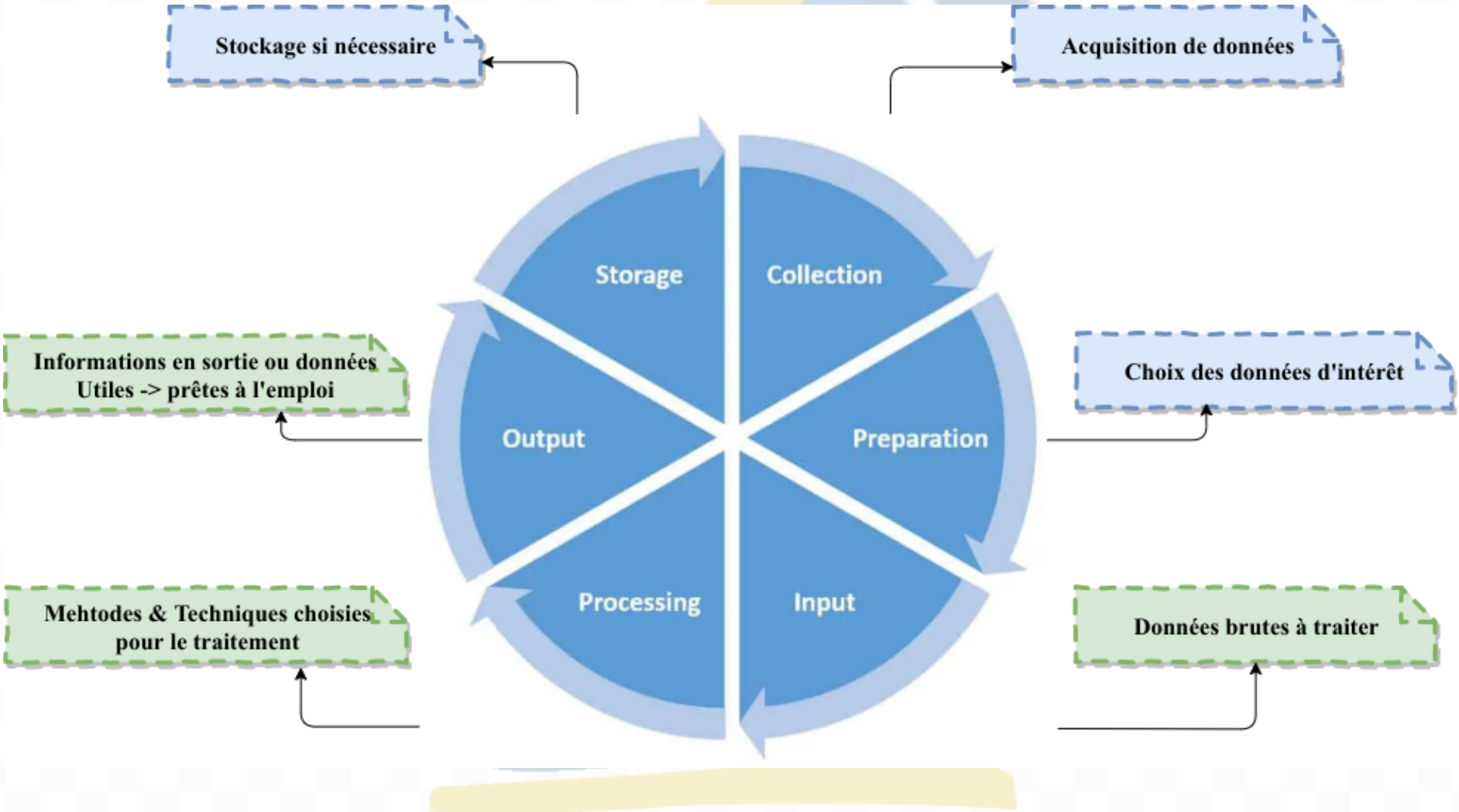

Common Data Science LifeCycle
Business Problem Understanding
Data Collection
Data Cleaning and Processing
Exploratory Data Analysis (EDA)
Model Building and Evaluation
Communicating Model Result
Model Deployment and Maintenance
Designed by Abraham Musa

# Définitions

- Data Processing Cycle : Ensemble de séquences ou le processus utilisé pour traiter les **données brutes** et les transformer en une forme **lisible/utilisable**.
- Les principaux étapes :
  - **Input :** constitue l'ensemble des données brutes obtenues après la phase de collecte
  - **Processing :** constitue l'ensemble des méthodes/techniques utilisées pour traiter les données. C'est l'une des importantes étape
  - **Output :** c'est la sortie produite par l'étape de traitement. Cette étape donne déjà des informations, qui seront utilisées pour les études spécifiques à venir dans le projet

# Définitions

- Data Processing Cycle



# Objectifs

## Objectifs & Pré-requis

- Objectif Général
  - À la fin de ce cours, vous serez capable de manipuler et de traiter correctement des données avec la librairie Pandas sous Python
- Pré-requis
  - connaissances basiques sur le tableurs connaissances basiques en python

Confère Scénario Pédagogique pour plus de détails

**NB** : Nous allons aussi introduire la notion de gestion de versions sur Github

# Séances

## Environnement de travail

- Installer Anaconda selon votre OS (recommandé)
- Installer Jupyter-Notebook
- Installer Pandas

### Introduction sur Jupyter-Notebook

- Cellules
- commentaires
- Raccourcis
- Librairies

# Séances



## Une petite révision sur Python

- Variables, Data types,
- Build In & Custom Functions
- String, Index Position & Slicing
- List, Dictionaries
- Etc.

# Séances

## Pandas Series

- Series Object from List and Dictionary
- Attributs & Methods on Series
- Series Object Parameters & Arguments
- Create Series from Dataset with pd.read_csv() method
- Head() & tail() Methods
- Sort_values(), Sort_index(), inplace parameter
- Extract Series Values by Index Label and Position
- The get() Method
- Math Methods on Series
- Find Greatest or Smallest values with idxmax or idxmin Mehthods
- Apply() and Map() Methods
- Final Test

# Séances

## DataFrames I : Introduction

- Intro to DataFrames
- Shared Methods and Attributes between Series & DataFrames
- Difference between shared Methods
- Select a given column from a DF
- Add new column to a DF
- Broadcasting Operations on DF
- Value_counts() Method
- Dropna() Method for rows with null values
- Fill in Null values with fillna()
- Use astype() to convert DF column type
- Sort_values() Methods to sort DF columns

# Séances

## DataFrames I : Introduction

- Sort_index() for DF index
- Rank method on Series values
- Final Test

# Séances

## DataFrames II : Filtering Data

- Memory optimization
- Filter a DF based on a Condition & more than one Condition (AND->, OR->|)
- The isin() Method for inclusion
- Isnull() & Notnull() Methods -> check Null
- Inclusion within a range -> Between() Method
- Check for duplacted rows in DF with duplicated() method
- Delete duplicated rows with drop_duplicated() Method

# Séances

## DataFrames III : Data Extraction

- Import Dataset with pd.read_csv() method
- Define new DF index -> set_index() and reset_index() methods
- Retrieve DF rows with loc() and iloc() accessors
- Second argument for loc() and iloc()
- Set new value for a specific cell/cells in Rows
- Set multiples values in DF
- Rename Index Labels or Columns in a DF
- Delete Rows or Columns in a DF
- Use sample() method to create a sample of Data
- nsmallest/nlargest -> get rows whith smallest/largest values
- Filetr a DF with where() and query() methods

# Séances

## DataFrames III : Data Extraction

- A review on Apply() Method on Series
- Apply method on every row of DF
- Create copy of a DF with copy() method

# Séances

## Données Textuelles

- Introduction
- Common methods for String -> lower(), upper(), len(), title()
- The str.replace() to replace all occurrences of a given character with another
- Filter a DF Rows with String Methods
- More String Methods -> strip(), lstrip(), rstrip()
- Usgae of String Methods on DF columns or Index
- Split Strings by Caracter with str.split() Method
- Split() method on Series
- The expand and n Parameters on str.split() Method

# Séances

## Multi_Index

- Intro to Multi_index
- Create Multi_index on a DF with set_index() Method
- Extract index levels values with get_level_values() method
- Change Index Level name with set_names() method
- Sort_index() method for multi_index DF
- Extract rows from MultiIndex DF
- Transpose method on DF
- The .swaplevel() method -> change multi-index level
- The .stack() & .unstack() methods
- Pivot() Method
- Pivot_table() & pd.melt() Methods on DF

# Séances

## GroupBy Object

- Intro to GroupBy Object
- Use get_group() -> retrieve a group from a Gropby Object
- Methods on GroupBy Objects and DF
- GroupBy Multiple Columns
- THe .agg() Method on GroupBy Object
- Itterating through a Groups of a GroupBy Object

# Séances

## Merging|Concatenating|Joining sur DF

- The pd.concat() Method on DFs
- The inner/outer joins
- Left joins
- The left_on & right_on Parameters
- Left_index/right_index -> merging by Indexes
- .Join() and pd.merge() Methods

# Séances

## Dates & Times

- Dates and Times Module
- Python DateTimes module - intro
- Pandas Timestamp and DateTimeIndex Object
- Pandas to_datetime Method
- .date_range() Method -> create a range of date
- Accessor for date -> .dt
- Usage of pandas-datareader library -> à installer
- Attributs & Methods for Timestamp Object
- The pd.DateOffset Object of pandas
- Offsets for Timeseries
- TimeDelta Object

## Input & Output avec Pandas

- Pass a URL to pd.read_csv() Method
- Quick Object convertion -> tolist & to_dict
- Export CSV file with to_csv() Method
- Thhe xlrd and openpyxl libraries -> for Excel files
- Read and export Excel file -> read_excel() and to_excel() Methods

# Séances

## Visualisation

- Intro to Visualization
- Plot() Method to render line chart
- Matplotlib template
- Bar Graphs to show Counts
- Pie chart to show Distributions/Proportions

# Séances

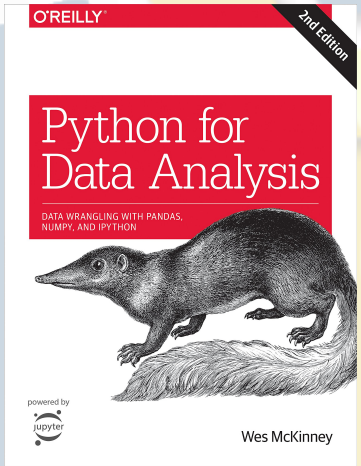


## Fichiers

- Read for specific extension
- Loop in a repertory
- Loop within more than one repertory
- Get Path, filename

# Ressources Utilisées

- Python for Data Analysis




- Online :pandas.pydata.org

# MERCI POUR VOTRE ATTENTION!