# MScT Complex System Engineering

## Centrale - 2023-2024

## Objectives

The goal of the project is to mobilize the knowledge you have acquired through the course in order to process one or more datasets using Python.

You are going to write a data science project. Meaning you are going to

You are going to write a data science project, which means you will need to address a question or a problem using various techniques from the field of data science.

## Course of action

To this end you will hand in an .ipynb (ipython notebook) script that :

**1.** Will start with an introduction to present the analysis that will follow.

This should contain a URL to the dataset(s) used. You must deal with quantitative data at some point in your project. This quantitative data can come from your dataset, or you can compute it yourself. Datasets in tabular format are recommended, as they are easier to process.

**2.** Introduce the dataset(s) used (Where does it come from? Who created it? For what purpose? What can we say about the data? How were they obtained?)

**3.** Will address a specific problem by testing one or more hypotheses.

**4.** Describe the data briefly using statistical indicators (number of rows, columns, mean, median, standard deviation, etc.)

**5.** Process the data by performing calculations, generating graphs or maps (if needed) or predict values.

**6.** Will end with a conclusion that will answer to the initial problematic and will summarize which hypotheses have been verified.

The conclusion should also contain a critical opinion exposing the limits of this work and leading to an opening.

# Methodology

The result must an ".ipynb" notebook readable by Jupyter Lab or Jupyter Notebook.

Each of you has to process a different dataset or a combination of different datasets, which requires you to consult each other. You are allowed to use a dataset used by another student, but only if it is not your main dataset and if you need to retrieve specific data by performing a join.

The file should not be edited beforehand by excel or another software, you will have to do everything in python from the raw data as downloaded from the internet.

The use of the Pandas library for data processing is mandatory.

Spelling and layout will be taken into account: a neat script, with no or few mistakes, and appropriate titles and paragraphs will earn you 4 points out of 20. Otherwise, you may be given a penalty of up to -8 points.

Remember that you can easily give your text a correct appearance by using the Markdown language.

You may **NOT** use a dataset from the organization "AtmoSud".

In your notebook, you should **NOT** include all the intermediate work and the exploratory phase: go straight to the point.

# Submission format

The file shall be submitted on the https://moodle.centrale-marseille.fr/ platform. If, for some reason, it's not possible, then send your work to virgile.pesce@univ-amu.fr

It shall consist of a compressed archive (ZIP, RAR...) containing the .ipynb script as well as the dataset(s) used and other useful documents. If the dataset(s) used is/are too large to be sent to Ametice, please provide a way to download an exact copy of the file.

# Tips

- There are a lot of datasets available on many platforms : data.gouv, data sud, nosdonnees...
- You can open and examine the file with Visual Studio Code, or notepad++ or any other text editor but you are not allowed to modify it.
- Once finished, kill the kernel and restart your entire script to check that it works without generating errors.

# Bonus points

Some actions will generate bonus points:

- Performing a merge with another dataset.
- Use an API to fetch data.

Don't hesitate to contact me at the following address if you have any questions or difficulties, I'll be happy to answer them (if I have time): virgile.pesce@univ-amu.fr

Don't wait until the last moment to start this project.

Good luck!