BSB - M1 : Data Science

Lucas Morlet - lucas.morlet@eseo.fr

Introduction

In this project, you will find the answer to a problem using the raw data. As a team, you will choose a subject that interests you (and that has an exploitable dataset). Then you will work on this data (by using Python, pandas, and associated libraries) to extract information that can help you find the answer to a problem you choose to address.

Schedule

- November 19th : Project reveal, team composition, choice of the dataset and the subject, and first steps on data preparation
- November 26th : Correlation
- December 3rd : Clustering
- December 10th : Supervised classification
- To be announced : project due

Team composition

Teams are composed of 2 to 4 students
Each team should work on a different dataset

Dataset choice

With your team, you should start by choosing a subject. This subject must have enough data to make some analyzes. You can select data from some open data providers from the links below or search for data on your own.

Famous Open-Acces Repositories :

- Kaggle - https://www.kaggle.com/
- Google dataset search - https://datasetsearch.research.google.com/
- Registry of Open Data on AWS - https://registry.opendata.aws/
- UCI Machine learning Repository - https://archive.ics.uci.edu/

Scientific research :

- Harvard dataverse - https://dataverse.harvard.edu/
- Zenodo - https://zenodo.org/
- FigShare - https://figshare.com/browse
- NASA EarthData - https://www.earthdata.nasa.gov/data/catalog
- NOAA Climate Data Online - https://www.ncei.noaa.gov/cdo-web/datasets
- WHO Global Health Observatory - https://www.who.int/data/gho/info/athena-api

Cities and countries

- USA Data.gov - https://data.gov
- San Diego - https://data.sandiego.gov/datasets/
- Chicago - https://data.cityofchicago.org/
- San Francisco - https://datasf.org/opendata/
- New York - https://data.cityofnewyork.us/browse
- European Data - https://data.europa.eu/en
- Luxembourg - https://data.public.lu/fr/
- French government API - https://api.gouv.fr/
- French statistics institute : INSEE - https://www.insee.fr/fr/statistiques
- Paris - https://opendata.paris.fr/pages/home/
- Rennes métropole - https://data.rennesmetropole.fr/explore/?sort=modified

Instructions

Report

- Describe the context of your data
- Import your data inside a pandas dataset
- Prepare your data to make them usable
- Choose a global problematic you want to address
- Explain which information you need to answer your problematic
- For each information :
  - Present which algorithm can help you to find your information and why this one is the most suitable
  - Reference the part of your code that contains this algorithm
  - Insert the result (extract from the console or a figure)
  - Write some analysis and interpretation of your result
- Aggregate everything you have find during your research
- Conclude your report by drawing a global answer to your problematic, if possible

Document to return :

- An archive (.zip) containing your dataset and all the files of your code
- Your report as a PDF

Video

- A 10 to 15-minutes video focused on the results (you don't need to explain your code there, that is the purpose of your report)
- You don't need to do video-editing (but you could, if you know how to do it), a sequence of slides with a voiceover will be sufficient.
- If pertinent, you can add short videos taken from the Internet to support your position

Document to return :

Your video in a standard format (e.g. .mp4, .avi…)

Marking scheme

Project mark

To be announced

Individual mark

To be announced