Data Analysis - Master MLDM - 2022-23
Project: Analyzing Spotify Data

1 Project description

General objective In this project, we will evaluate your ability to describe a dataset, and to extract the information from it using the techniques we have studied during this semester. Deliverables: **source code** that we can run on our machine (in Python) + **report**, which must contains 4 sections: "*Introduction*", "*Overall Description*", "*Analysis*", and "*Conclusion*". You can use illustrations, graphs, tables etc... The report should be 6 pages long maximum.

Deadline and additional information You can work alone or in pairs. Project and source code should be sent to eduardo.brandao@univ-st-etienne.fr before January 6 (11:59pm).

2 Code

You should provide a requirements.txt file so that we can reproduce your experiments (see `https://realpython.com/lessons/using-requirement-files/`, for example). The code should be commented, and well organized. This will be taken into account for your final grade.

3 Data

Download the data The dataset is available on Claroline. An overall description of the features can be found on `https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features`.

4 Report

***Overal Description* section** In the *Overall Description* section of the report, you must describe the dataset, its size, the variables and their nature. Propose an overall analysis of the variables in the dataset: mean and total of the quantitative variables, proportion, variances of the variables of interest, etc. Do not hesitate to illustrate this part with numerous graphs (e.g. using dimensionality reduction techniques).

***Analysis* section** In the *Analysis* section of your report, you are free to provide any analysis you find interesting. You can perform a wide range of method to extract information from the data. For example, you can try to train a classifier on any variable of the dataset, try to do clustering, and see whether it fits some categorical data information (such as the genre, for example). You can also augment the dataset with additional information you find online (such as the .wav of the song as a variable, for example). Do at least one analysis, with full investigation of performance, optimal parameters, comparing different approaches, etc... If there are missing values, you may want to use imputation techniques. Examples of analysis:

- training a classifier to predict the genre from the quantitative variables
- predict the year from the variables
- regression of Valence using the quantitative variables as predictors