Data Analysis
Final Project - Analyzing Spotify Data
Master MLDM - 2021/2022

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. You have to provide: a source code that we can run on our machine (in *Python*) and a report. It must contains 4 sections: "Introduction","Overall Description", "Analysis", "Conclusion". You can use illustrations, graphs, tables etc... The report should be 6 pages long **maximum**.

Deadline and additional information You have until **January the 7th (11:59pm)**. Please send your project to `antoine.gourru@univ-st-etienne.fr`. This is a group project, you should be maximum **2** by groups. You are allowed to work alone.

Download the data You can download the data here :
`http://antoinegourru.com/spotifyData.csv`

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

Quick description of the dataset These are the top song on Spotify by years (2009-2019). The song are described by several variables of different types. It has 603 observations and 17 variables. The variables are the following:

- ID : the ID :-)
- title: Song's title
- artist: Song's artist
- top genre: the genre of the track
- year: Song's year in the Billboard
- bpm: Beats.Per.Minute - The tempo of the song.
- nrgy: Energy- The energy of a song - the higher the value, the more energtic the song
- dnce: Danceability - The higher the value, the easier it is to dance to this song.
- dB: Loudness..dB.. - The higher the value, the louder the song
- live: Liveness - The higher the value, the more likely the song is a live recording
- val: Valence - The higher the value, the more positive mood for the song.
- dur: Length - The duration of the song.
- acous: Acousticness.. - The higher the value the more acoustic the song is.
- spch: Speechiness - The higher the value the more spoken word the song contains.
- pop: Popularity- The higher the value the more popular the song is.
- emo: Bad feeling - The higher the value, the more depressing the song is
- ins: Instrumentalness - The higher the value the more instrumental the song is.

Overall Description 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 variable, proportion, variances of the variables of interest, studies of the variables of interest by modality, etc. Do not hesitate to illustrate this part with numerous graphs (e.g. using dimensionality reduction techniques).

Analysis 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, using Adjusted Rand Index). You are free here. You can also augment the dataset with additional information you find online. You can even add the .wav of the song as a variable if you manage to get it. Do at least one analysis, with full investigation of performance, optimal parameters, comparing different approaches, etc... Warning : the dataset demonstrates missing values : you might 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.