Visualization of massive data : project

NICOLAS LE HIR
nicolaslehir@gmail.com

1 DESCRIPTION OF THE PROJECT

The goal of the project is to propose visualizations of a dataset and a quantitative analysis.

1.1 Dataset constraints

You are free to choose the dataset within the following constraints :

— utf-8 encoded in a **data.csv** file

— several hundreds of lines

— at least 6 attributes (columns), the first being a unique `id`, separated by commas

— you may use some categorical (non quantitative) features.

— some fields should be correlated

It's nice if the dataset comes from a real example but you can also generate it.
Example datasets available :
`https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research`
`https://perso.telecom-paristech.fr/eagan/class/igr204/datasets`
`https://github.com/awesomedata/awesome-public-datasets`
`https://www.kaggle.com/datasets`
The processing must be made with **python** (**python** 3 **preferred**).

1.2 Visualization

Write a program that allows the user to select between at least 2 visualizations of your dataset. This means 2 different method of visualization (for instance a parallel coordinate plot and / or scatter plots / and or a hierarchical clustering).
The interface does not have to be complicated, you can also write two different scripts, one for each visualization method (the project is not about the interface, rather about the visualization method).

For instance the user could be able to select a subset of the initial dataset, and/or a subset of the features that will be taken into account in the representation. However, please propose a set a meaningful default parameters so that the user can quickly use your visualization in a relevant manner.

You may use any non-trivial visualization method that we studied during the course, and also any other relevant method.
Use relevant visualization so that they are helpful to identify **tendencies** or **structure** in the data. Please comment on this aspect in the accompanying document (see below).

1.3 Quantitative analysis

Select one or more columns of your dataset and perform one of the following analysis :

— learn a predictive model of one column as a function of another column.

— analyse the way those columns are distributed and fit a distribution to those. Explain the method used to fit the distribution and the choice of the model.

Important :
The usage of this dataset and its processing should be justified by a question of your choice. Thus, the approach should be explained and justified in a separate pdf file. The pdf file needs to contain explanations about :

— the nature of the dataset

— information on the potential correlation between variables.

— explanations on the quantitative processing and on the visualization.

— comments on the results obtained.

2 ORGANIZATION

The students can form groups with at most 3 students (they can work alone too).
The deadline for submitting the project is **May 3rd 2020** in a compressed folder or a repo containing :

— the name of the student(s)

— the csv dataset **data.csv**

Please write "Visualization session 3" in the subject of your email.
You can reach me by email if you have any question.

3 EXERCISES DONE DURING THE COURSE

The exercises we made during the class are available with correction here : `https://github.com/nlehir/Visu`.

4 LIBRARIES

You may use third-party libraries : however, if you do so, it is required that you present them in your document and describe the functions that you use from thie library, and comment on the choice of the parameters.

5 VALIDATION

The project is not mandatory and can offer two extra credits, apart from the credit based on being present in the class.