Visualization of 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, or even build a hybrid dataset mixing real data and generated data. The goal of the project is to apply some methods seen in class.

Example resources to find datasets :

— Link 1

— Link 2

— Link 2

— Link 4

The processing must be made with **python 3**.

1.2 Processing

1.2.1 *Visualization*

Write two python programs, each producing a visualization your dataset. This means 2 different method of visualization (for instance a parallel coordinate plot and / or scatter plots / and or a hierarchical clustering).

These programs must allow the user to select a range of points of the datasets to be represented.

The user may also choose some other visualization parameter. For instance, the user could be able to select a subset of of the features that will be taken into account in the representation.

However, both the range of points plotted and the optionnal parameters should have default values, so that the user can quickly use the visualization.

The interface does not have to be complicated, a console input is fine. The project is not about the interface, rather about the visualization method, so you should not focus on the interface part too much.

You may use any non-trivial visualization method that we studied during the course, and also any other relevant method.

Use relevant visualizations so that they are helpful to identify **tendencies** or **structure** in the data. Please comment on this aspect in the accompanying document (see below).

1.2.2 *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, or of several other columns (supervised learning).

— 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 (unsupervised learning).

1.2.3 *Comments*

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 2 or 3 students.
You can send the project in a compressed folder or a repo containing :

— the name of the students

— the csv dataset **data.csv**

— the python scripts.

Deadlines :

— Session 1 (January 7-8 2021) : January 24th 2021

— Session 2 (January 14-15 2021) : January 31th 2021

Please write "Visualization session 1", "Visualization session 2" (depending on your session) 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 the library, and comment on the choice of the parameters.