Visualization of data : project

NICOLAS LE HIR
nicolaslehir@gmail.com

TABLE DES MATIÈRES

1 PART 1 : ARTIFICIAL DATASET GENERATION

The goal of this part is to make you work with statistical notions such as mean, standard deviation, and correlation.

Write a file named **artificial_dataset.py** that generates a numerical dataset with 300 datapoints (i.e. lines) and at least 6 columns and saves it to a csv file or to a numpy array in a binary python file.

The columns must satisfy the following requirements :

— they must all have a different mean
— they must all have a different standard deviation (English for "écart type")
— at least one column should contain integers.
— at least one column should contain floats.
— one column must have a mean close to 2.5.
— some columns must be positively correlated.
— some columns must be negatively correlated.
— some columns must have a correlation close to 0.

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

2 PART 2 : DATASET ANALYSIS AND VISUALIZATION

The goal of this part is to propose visualizations of a dataset and a quantitative analysis. The dataset should not be the dataset generated in Part 1.

2.1 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

If necessary, you can tweak a dataset in order to artificially make it possible to apply analysis ans visualization techniques.

Example resources to find datasets :

— Link 1

— Link 2

— Link 2

— Link 4

2.2 Processing

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

2.2.1 *Analysis*

The file **analysis.py** presents a quick analysis of the dataset. For instance :

— Histograms of quantitative variables with a comment on important statistical aspects, such as **means** , **standard deviations** , etc.

— A study of potential **outliers**

— Correlation matrices (maybe not for all variables)

— Any interesting analysis : if you have categorical data, with categories are represented most ? To what extent ?

If the dataset is very large you may also extract a random sample of the dataset to build histogram or compute correlations. You can discuss whether the randomness of the sample has an important influence on the analysis result (this will depend on the dataset).

Note for students who did the Algo2 project : there is a similarity of the first part of the Algo2 project here, please contact me if you want to re-use the same dataset.

2.2.2 *Visualization*

Write two python programs, **visualization_1.py** and **visualization_2.py**, 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 optional 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).

2.2.3 *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, or as a function of several other columns. You are free to choose the supervised learning method (linear model, perceptron, etc.) This analysis should be performed in a file named **supervised_learning.py**.

— Perform an unsupervised algorithm on your dataset (not necessary on all columns, you can choose relevant columns). For instance : density estimation, dimensionality reduction, clustering, etc. This analysis should be performed in a file named **unsupervised_learning.py**.

Important : in either case, please provide an **evaluation** of your algorithm. For supervised learning, this could be an average squared error, coefficient of determination, etc. For unsupervised learning this could be the inertia, distorsion, KL divergence, etc. You could for instance use this evaluation to choose the parameters of your model.

2.2.4 *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.

3 ORGANIZATION

Number of students per group : 2.

Submission deadlines :

— Session 1 : **December 12 2021**

— Session 2 : **March 6th 2022**

The project must be shared through a github repo, sent by email with contributions from all students. The repo should contain :

— the pdf report. Please indicate how work was divided between students (each student must have contributions to the repository).

— the csv dataset **data.csv**

— the python scripts

— artificial_dataset.py
— analysis.py
— visualization_1.py
— visualization_2.py
— supervised_learning.py or unsupervised_learning.py

Please write "Visualization session 1", "Visualization session 2" (depending on your session) in the subject of your email.

You can reach me by email, I will answer faster if you use the gmail address rather than the Epitech address.

4 EXERCISES DONE DURING THE COURSE

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

5 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.