ESILV LÉONARD DE VINCI | ÉCOLE D'INGÉNIEURS PARIS-LA DÉFENSE

Programming for Data Science (MESIIN471625)

Sarah MALAEB, Constantin TESTU, Hugo ALATRISTA-SALAS

Final project

1. Introduction

Many phenomena can be modeled using computational systems. The modeling process begins by identifying the objects that comprise the overall system. Next, the attributes or characteristics of objects that simplify reality are described. These objects and their attributes are stored in data repositories such as databases, data warehouses, data lakes, and spreadsheets. Later, instances of these objects are created, populating the database. Data is essential for extracting information that improves our understanding of the phenomena under study and informs data-driven decision-making.

In this context, in this final project, the techniques learned in this course will be used to extract and visualize helpful information from different data sources.

2. Project goal

The aim of this project is to implement a visualization prototype showing four indicators extracted from the available dataset. The Knowledge Discovery in Databases (KDD)[1] process can be used to inspire the extraction of useful information from a real dataset.Additionally, your proposed solution should be encapsulated, dividing the main problem into small, independent tasks. This will clarify your code and enable reuse. Thus, each method or technique in your solution should be implemented as a function using *def():* and include any necessary parameters. Of course, you can reuse or improve methods implemented in previous labs. Finally, all the functions created must be called from a *__main__* function implemented at the end of the project, which also launches the visualization prototype.

Note: For inspiration on writing your notebook, see the code box at the end of this PDF.

3. About the dataset

The invoice dataset is a synthetic dataset created with the Python Faker library, designed to replicate the structure of data typically gathered from an online store. It includes multiple fields such as customer details (first name, last name, email), transaction information (product ID, quantity, amount, invoice date), and additional attributes like address, city, and stock code. For more details, visit `https://www.kaggle.com/datasets/cankatsrc/invoices`.

4. Important information

This work must be completed by groups of at **least 4 (four)** and at **most 6 (six)** students. Each team must elect a leader who will upload the deliverables for this final project. The deadline for uploading

[1] `https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf`

the project files is **December 15, 2025**. Presentation dates will be announced in the coming weeks. The presentation will include a demo and a presentation of additional information and main findings with the help of a support tool, such as slides. Since this is a team grade, all group members must be present during the oral presentation. If a group member cannot answer a question or present the project, the group will automatically receive a penalty.

5. Deliverables

1. [*15 points*] The notebook in HTML format[2]: The groups must show how they carried out the steps described above (see Section 2) in their notebooks. The grade for this work will depend on how creative they are and how well they follow the instructions given earlier. Comments describing key lines of code are also essential in this submission. This file should be named as *familyNameLeader_nameDataset.ipynb.html*

2. [*5 points*] A presentation containing no more than 8 slides in PDF format, outlining the key points of your project, should be submitted. This second file called *familyNameLeader_nameSlides.pdf* should contain additional information to that presented in the notebook.

6. Tasks and its evaluation

This project should evaluate two main capabilities: 1) technical capabilities (Python implementation), and 2) the presentation of results (oral presentation).

6.1. Evaluation of technical capabilities (15 points)

The evaluation of this capability is divided into positive and negative points. The following list describes how the points should be given, considering the main tasks to be implemented in this project.

1. [*2 points*] Data Collection: Create a notebook, upload the data and store it in a structure that allows data manipulation. Explore the data using the data summarization methods provided by Python to fully understand it. You can use the statements we have learned in the course. At this stage, you should at least show the number of columns and data types in each column, the number of rows and columns in the dataset, the number of missing data points in each column, and the ranking between variables. You can use the *markdown* boxes to describe the dataset at your disposal and some visual components to help you understand your data. This last result can be used in the next step.

2. [*9 points*] Apply different techniques to extract information from the dataset. Explain what each technique represents in detail and visualize the results. The results of this step should include:

   - [*1 points*] A query using grouping queries from the dataset, for example, "the maximal number of ..." or "the sum of the ..."
   - [*2 points*] A result that uses data transformation methods such as normalization, discretization, frequent pattern mining, or others learned in our course. You may also include results from techniques learned in other courses, such as machine learning.
   - [*6 points*] Two results using temporal and/or spatial data features, one of each if the dataset allows. Apply the forecasting and/or spatial clustering techniques learned in this course. You may also use techniques learned in other courses.

   Note. Explain all applied methods and results, including how they are calculated, what they represent, and how they will be interpreted. Use the *Markdown* boxes to explain your findings.

[2] If you are having problems converting your notebook into an HTML file, you can also send us the *.ipynb* file or in a ZIP format.

3. [*4 points*] A Visualization Dashboard: Use Python Dash (`https://dash.plotly.com/`) to visualize a dashboard containing the four elements built in the previous step. The dashboard should include the names of the team members and the name of the dataset used for this project and the description of the project objective. Remember that a goal should be concise, achievable, and tangible.

4. [*1 extra point*] External dataset: You can enrich your results by using an external dataset. You can also use the additional dataset to interpret the information you obtained.

If all the previous tasks are completed correctly, the team will score 16 points. However, some mistakes can penalize the final punctuation. The most common errors are:

1. [*-2 points*] Failure to organize all the code into functions.

2. [*-2 points*] If at least one of the functions does not include comments about its name, input, and output (see the example in the code box). Remember to add comments to make your code more understandable and to facilitate correction of the notebook. It is worth noting that if you decide to remove a feature, you must explain why. Describe all decisions in your proposal.

3. [*-2 points*] If the dashboard does not work or there is no evidence that it works.

4. [*-2 points*] Each indicator should be described in human terms in a text cell (*markdown*). The penalty for each indicator description is -2 points.

6.2. Evaluation of the oral presentation (5 points)

Projects must be presented orally. Each group will have a maximum of 8 minutes to present their project in two stages: 1) a presentation of the project, including the objectives, description of the data, techniques used on the analytical step, and results obtained; and 2) a demo of their code and dashboard. The instructor may ask questions about the code and the slide presentation. The grade depends on how well the group presents its work and answers questions.

Some mistakes may result in a lower grade for the oral presentation. The penalties are as follows:

- [*-5 points*] If at least one member is not present when the project is presented.
- [*-3 points*] If the demo and the dashboard does not work.
- [*-2 points*] If the group does not have support materials (e.g., slides).
- [*-1 points*] If a member fails to respond to a question and other member answer for him (accumulative penalty).

May the force be with you !!!

familyNameResponsable_nameDataset.ipynb

```
# Team members and dataset
Dataset: earthquakes
FirstName, FamilyName (team head)
FirstName, FamilyName (team member)
...

# Import libraries
import pandas as pd
import numpy as ny
...

# Function definitions

# Function for reading a CSV file
# Input: the path to the csv file
# Output: a dataframe
def load_data(file_path):
      bla bla bla # Try to comment on each line
      bla bla bla # especially if this line helps us to understand your project
      ...

# Function to bla bla bla
# Input: bla bla bla
# Output: bla bla bla
def function_1(df):
      bla bla bla
      ...

def function_2(df):
      bla bla bla
      ...

# Main block
if __name__ == "__main__":
      # Step 1: Data collection
      file_path = "myfile.csv"
      data = load_data(file_path)
      ...

      #Step 2: Indicators construction
      result = function_1(df) # this is only an example
      ...
```