CENG463 Assignment 0

Spring 2024-2025

Deadline
02.05.2025, 23.55

Assignment Instructions

- Students are required to form groups consisting of **two or three members**. Each group must register by completing the shared Google Sheets document available at the following link: https://docs.google.com/spreadsheets/d/19eFg47Quj3YbOAaqiFsSCGH1MFQGrFpQp8yb4tFdWFA
- The submitted Python source file must be named according to the following convention, listing student numbers in **ascending numerical order**:
  CENG463_Studentnum1_Studentnum2_Studentnum3.py

Rules and Restrictions

- Collaboration across groups, the use of publicly available code, and the generation of code via artificial intelligence tools are **forbidden.**
- You are **only allowed to use** the following libraries: **Scikit-learn, NumPy, Pandas, Matplotlib**, and **Python's standard library**. Should you have any further inquiries, please feel free to contact me. If you believe another library is necessary, please request approval in advance via email at: **cerensozeri@iyte.edu.tr**

Tasks

Using the provided CSV dataset, your group is required to develop a Python script that performs the following tasks:

1. **Data Quality Report**

   Your implementation should generate two distinct **Data Quality Reports** based on the nature of the features in the dataset:

   - One for **categorical features**: output_DQR_Categorical.csv
   - One for **continuous features**: output_DQR_Continuous.csv

2. **Feature Visualization**

   For all **descriptive features**, you are required to visualize their relationship with the **target feature**.The type of plot should be selected **appropriately**, based on the characteristics of **each feature.** Do not rely solely on the first row to infer the nature of a feature; instead, examine the entire column to determine the most suitable visualization method.

For the provided input.csv file, your implementation must generate **three distinct outputs.** These include **two CSV files named** output_DQR_Continuous.csv and output_DQR_Categorical.csv, and a set of visualizations illustrating **the relationship between each feature and the target feature.** The script must accept **two command-line arguments**: the first specifying the path to the input CSV file, and the second indicating **the directory** in which all visualization **outputs** should be saved. For **continuous-continuous** feature pairs, use **scatter plots**; for **continuous-categorical** pairs, use **multiple histograms**; and for **categorical-categorical** pairs, use **multiple bar plots**. In scatter plots, the target variable must always be placed on the **y-axis**. When constructing **histograms**, ensure that **binning** is performed thoughtfully to allow meaningful comparisons.