AIG100 – Project 1: Real-World Dataset Exploration

Overview

This project will involve performing an exploratory data analysis (EDA) on a real-world dataset of your choice (e.g., a dataset from Kaggle or any public data repository). The goal is to apply statistical inference methods to draw meaningful insights from the data. The project will leverage Python libraries such as Pandas for data manipulation, Matplotlib and Seaborn for data visualization, and SciPy for statistical tests. These libraries are only as examples of what you can use. You may choose to work with any package/library of your choice.

Objectives

- To apply data manipulation and cleaning techniques to prepare a dataset for analysis.
- To perform exploratory data analysis to uncover trends, patterns, and anomalies in the data.
- To utilize statistical tests to make data-driven inferences.
- To practice presenting findings in a clear, understandable manner using visualizations.

Tasks

1. Dataset Selection and Preliminary Research
    a. Select a dataset of interest from a public data repository.
    b. Conduct preliminary research to understand the dataset's context and potential questions to explore.
2. Data Cleaning and Preprocessing
    a. Use Pandas (or anything else you wish) to clean the dataset, handling missing values, outliers, and incorrect data types.
    b. Perform necessary data transformations to prepare the dataset for analysis.
3. Exploratory Data Analysis
    a. Conduct an EDA to uncover trends and patterns in the dataset. This may include:
        i. Distribution of key variables using histograms.
        ii. Relationships between variables using scatter plots and correlation matrices.
        iii. Group comparisons using box plots and bar charts.
4. Statistical Inference
    a. Formulate one or two hypotheses based on the EDA findings.
    b. Use SciPy (or anything else you wish) to conduct appropriate statistical analysis. You could explore SciPy *t-tests*, *chi-square tests*, etc.
    c. Interpret the results of the statistical tests to draw conclusions about the data.
5. Visualization and Presentation of Findings
    a. Create visualizations to present the findings from the EDA and statistical tests.
    b. Use Matplotlib and Seaborn to generate clear, informative graphs. Or use whatever visualization library you wish.

6. Report Writing
    a. Document the analysis process, findings, and conclusions in a comprehensive report.
    b. The report should include an introduction to the dataset, a summary of the data cleaning/preprocessing steps, key findings from the EDA, results of the statistical tests, and conclusions drawn from the analysis.
7. Reflection
    a. Reflect on the analysis process, discussing any challenges encountered and how they were overcome.
    b. Propose further questions for exploration or additional analyses that could be performed with more advanced techniques.

Submission

Implement everything in a Jupyter Notebook file. Make sure that codes are written in a *code cell* and textual explanations are written in a *Markdown cell*. You need to run your notebook so that visualizations are saved in the submitted notebook. Submit a link to your completed Jupyter Notebook file hosted on your private GitHub repository through the submission link in Blackboard.