Data analysis Project

Project: Data Integration and Exploration of a Multi-Source Dataset

Análise de dados.

João Viana
jviana@ufp.edu.pt

2024/25

Universidade Fernando Pessoa

Faculdade de Ciência e Tecnologia

Introduction

In today's data-centric environment, integrating information from multiple sources is essential for comprehensive analysis and informed decision-making. **Data integration** involves combining data from different origins to provide a unified view, facilitating deeper insights and more robust conclusions.

Examples of Dataset Combinations:

- **Weather Data & Sales Data:** Analyzing how weather patterns influence consumer purchasing behavior.
- **Economic Indicators & Crime Data:** Exploring correlations between economic conditions and crime rates.
- **Healthcare Records & Demographic Data:** Investigating the impact of demographic factors on health outcomes.

Potential Data Sources:

- **Weather Data:** NOAA Climate Data Online
- **Sales Data:** Instacart Market Basket Analysis Dataset
- **Economic Indicators:** World Bank Open Data
- **Crime Data:** FBI Crime Data Explorer
- **Healthcare Data:** CDC WONDER
- **Demographic Data:** U.S. Census Bureau Data

Through effective data integration and exploration, analysts can identify patterns, correlations, and anomalies that single-source data might not reveal.

Project Details

Objective

Integrate multiple datasets to analyze and uncover meaningful relationships between variables.

Key Tasks

1. **Data Importation & Integration:**
   - Load datasets from various sources.
   - Merge data using appropriate techniques (e.g., joins) to ensure consistency.
2. **Handling Missing Data & Outliers:**
   - Identify and analyze patterns of missing data.
   - Detect outliers using statistical methods.
   - Apply suitable techniques to handle missing values and outliers, providing justification for chosen methods.
3. **Exploratory Data Analysis (EDA):**
   - Perform statistical summaries and visualizations to understand data distributions and relationships.
   - Identify trends, correlations, and anomalies.
4. **Identify Correlations & Trends Through Visualization:**
   - Utilize visual tools (e.g., scatter plots, heatmaps, time-series plots) to identify patterns and potential causal relationships.

Recommended Tools

- **Python:** `pandas`, `matplotlib`, `seaborn`, `scikit-learn`
- **R:** `dplyr`, `ggplot2`, `tidyr`
- **SQL (optional):** For efficient querying and merging of structured datasets

Grading Rubric (100%)

Grading Criteria (100%)

1. Research Question – 10%

- Clear, specific, and relevant to the datasets used.
- Demonstrates an understanding of the problem being explored.
- Aligns well with the objectives and scope of the project.

2. Data Sources – 10%

- Appropriate choice of datasets to support the analysis.
- Sources are credible and well-documented.
- Limitations or biases in the data are acknowledged and discussed.

3. Data Integration – 20%

- Proper merging of multiple datasets using correct join methods.
- Handles inconsistencies (e.g., date formats, column names) effectively.
- Integration process is clearly explained and justified.

4. Handling Missing Data & Outliers – 15%

- Identifies missing values and outliers using appropriate methods.
- Applies suitable techniques (e.g., imputation, filtering, transformation).
- Justifies the decisions made and considers the impact on results.

5. Exploratory Data Analysis (EDA) – 15%

- Includes meaningful statistical summaries and distributions.
- Identifies patterns, relationships, and anomalies.
- Demonstrates critical thinking and curiosity through the exploration.

6. Data Visualization – 15%

- Visualizations are clear, relevant, and easy to interpret.
- Follows best practices (titles, labels, color use, readability).
- Plots support the narrative and reveal key findings.

7. Interpretation & Conclusions – 10%

- Conclusions are supported by the analysis and visualizations.
- Provides insights that relate to the original research question.
- Acknowledges assumptions, limitations, and possible alternative explanations.

8. Code Quality & Reproducibility – 5%

- Code is clean, organized, and commented where needed.
- Notebook runs from start to finish without errors.
- Anyone can reproduce the analysis with the submitted notebook.

Grading Scale (0-5)

Each criterion is assigned a percentage of the total score, and will be graded from **0 (Not Submitted or No Attempt) to 5 (Excellent)**.

- **5 (Excellent):** Fully meets or exceeds expectations, with thorough execution.
- **4 (Good):** Meets most expectations, with minor issues.
- **3 (Satisfactory):** Adequate but with noticeable gaps.
- **2 (Needs Improvement):** Some effort but major issues present.
- **1 (Poor):** Minimal effort, major deficiencies.
- **0 (Not Submitted or No Attempt):** No work done in this area.

Project Deliverables

Students are required to submit:

1. **Notebook:**
   - A Jupyter Notebook (`.ipynb`) or R Notebook (`.Rmd`) containing all data integration, preprocessing, analysis, and visualization steps.
   - Code should be well-documented with clear explanations of each step.
2. **Report:**
   - A comprehensive report (`.pdf` or `.docx`) detailing the problem statement, dataset selection, methodology, challenges faced, key findings, and interpretations.
   - The report should include tables, visualisations, and discussions on limitations and potential improvements.

These project components must be submitted as a link to github through the e-learning platform by the deadline recorded in the Assignment.
At the end of the project, a live presentation / defense must be conducted on dates to be announced by the Professor(s). All presentations must take place early enough to meet the deadlines established for the release of final grades according to the current academic calendar.
Projects submitted late or not presented in person will not be considered for grading. The project may be completed in groups of up to two members.