Lab in Data Analysis (62010)

Final Project: Comprehensive Exploratory Data Analysis (EDA)

Objective

The goal is to apply the principles of EDA to a real-world dataset, showcasing skills in data cleaning, transformation, visualization, and statistical summarization using Python.

Requirements

1. Dataset Selection

- **Source**: Students must choose a publicly available dataset (e.g., Kaggle, UCI Machine Learning Repository, government open data portals) or one provided by the instructor.
- **Size**: The dataset should have:
    - At least 10,000 rows and 10 columns.
    - Mixed data types (e.g., numerical, categorical, and potentially missing values).
- **Approval**: Submit a short description of the dataset and its context for approval.

2. Data Understanding

- Provide a **detailed introduction** to the dataset:
    - Source and context of the data.
    - Description of columns (features) and their meanings.
    - Identification of potential target variables, if applicable.

3. Data Cleaning and Preparation

- Handle missing values:
    - Explain missing data patterns and apply appropriate strategies.
- Handle duplicates and irrelevant data.
- Identify and handle outliers where necessary.
- Standardize column names for consistency.
- Convert data types where applicable (e.g., date strings to datetime).

4. Data Transformation

- Perform transformations such as:
    - Encoding categorical variables (one-hot, label encoding, etc.).
    - Normalizing or scaling numerical data.
    - Feature engineering (creating new features from existing ones).

5. Data Merging and Integration

- If applicable, merge or integrate the chosen dataset with another dataset:
    - Demonstrate the process of joining datasets using merge() or other techniques in Pandas.
    - Ensure alignment and consistency between datasets.

6. Data Exploration

- Provide **visual and statistical insights** for all variables:
    - Summary statistics: mean, median, variance, etc.
    - Distribution plots for numerical variables (histograms, KDE).
    - Bar plots for categorical variables.
    - Correlation matrix and heatmap.
- Identify patterns, trends, and relationships between variables.
- Highlight at least **three key findings** from the data.

7. Data Visualization

- Use Matplotlib and Seaborn (or other libraries) to create:
    - A variety of plots (scatter, bar, line, box, violin, etc.).
    - Interactive visualizations (optional but encouraged).
- Ensure plots are clear, properly labeled, and include legends where necessary.

8. Aggregation and Grouping

- Perform data aggregation using groupby() or similar methods:
    - Summarize data based on relevant groupings.
    - Create pivot tables where applicable.

9. Insights and Conclusion

- Summarize key insights from the analysis.
- Reflect on the data's potential use cases and limitations.

10. Code and Documentation

- Submit a Jupyter Notebook with:
    - Well-commented code.
    - Clear and organized markdown cells explaining each step.
    - A professional and concise summary of findings at the end.

11. Presentation

- Prepare a **15-20-minute presentation** to showcase:

- Dataset introduction and rationale for selection.
- Key steps of the EDA process.
- Major findings and visualizations.
- Challenges faced and how they were addressed.

12. Submission Requirements

- **Jupyter Notebook**: Fully functional and well-documented notebook.
- **Presentation Slides**: Clear and concise slides summarizing the project.
- **Deadline**: 23/01/2025.
- **Presentations Dates**: 26/01/2025, 02/02/2025.

Evaluation Criteria

Criteria	Weightage
Dataset Selection	10%
Data Cleaning and Preparation	20%
Data Exploration and Insights	20%
Visualizations	20%
Documentation and Code Quality	15%
Presentation	15%

Good Luck!