Lab Assignment 2: Data Reading and Processing

Overview

This project is designed to test students' proficiency in handling and processing different types of data. The challenge lies in dealing with 'damaged' files, requiring students to demonstrate their skills in data cleaning, preprocessing, and integration.

Objectives

- **Read Various Data Types**: Students request to read and interpret the given data.
- **Data Cleaning and Preprocessing**: Identify and rectify issues in the 'damaged' files, which may include missing values, inconsistencies, and format-related problems.
- **Data Integration**: Combine information from all three data types to create a cohesive dataset.
- **Code Documentation**: clean code, well-commented Python code to process the data with necessary comments.

Project Tasks

1. **Data Acquisition and Consolidation**:
    - **Objective**: You will be provided with three distinct files, each encapsulating a unique type of data that is part of a fragmented and somewhat compromised dataset. Your task is to meticulously read and amalgamate these files into a single, comprehensive dataset, subsequently saving it in a CSV format.
2. **Data Preprocessing and Refinement**:
    - **Data Identification**: Engage with datasets that may not be immediately interpretable. Your responsibility includes isolating pertinent data and refining it for further analysis.
    - **Data Cleansing**: Address the challenge of disorganized data sets. For instance, in numeric (integer) data, you may encounter strings embedded with special characters. Your role is to locate and purify these data anomalies to ensure dataset readability. Additionally, identify and label any void or missing data points as 'NaN' for future processing stages.
    - **Data Standardization**: Given that the three files are segments of an originally unified dataset, you are required to restructure these files into a standardized format. This process involves aligning column names, standardizing data types across columns, excising any special characters, and eliminating null entries.
3. **Data Statistics:**
    - This section should elucidate the findings from your data analysis, emphasizing aspects **included**
        - *the aggregate data volume (Good data).*
        - *instances of missing data (Bad data).*
        - *the number of unique companies.*
        - *the company with the highest revenue from 1995 to 1998.*
        - *the company with the highest profit from 1995 to 1998.*
    - Merge these results into a Single Data frame table named “Results_Combine”, and organize them with suitable column name and index.
    - Print “Results_Combine” in a new cell.

    - Save “Results_Combine” to CSV file.

4. **Project Documentation**: Your report should be systematically structured with the following segments with **Markdown format**:
    - **Title Section**: This section should encompass the project title, a complete list of your members, and the submission date.
    - **Introduction Section**: Concisely articulate the objectives and scope of the project. This entails describing the content represented by the data.
    - **Results**: Describe your results and the methods you used to describe them to the reader.
5. **Comments**: Add comments for each line of your code.

Submission

- **Entire Code:** Submit the **entire code (.ipynb)** used for reading, cleaning, preprocessing, and integrating the data, which included clear and readable **comments** throughout your code.
- **Processed Data Set:** Submit the processed data with a CSV format.
- **Upload to GitHub:** Please upload all your files to GitHub and submit GitHub link to Canvas.