Lesson 3 Automating Analyses with Make Assignment: Building an Automated Data Analysis Pipeline using Make

Objective:
To create an automated data analysis pipeline using Make. This pipeline will involve multiple steps, including data preprocessing, analysis, and report generation.

Dataset Preparation:

- Prepare or choose at least two datasets for analysis. These could be CSV, JSON, TXT or any other structured data format.
- Place the datasets in a directory named data.

Script Development:

- Write two scripts:
    - A preprocessing script (e.g., preprocess.py) that cleans or prepares the data.
    - An analysis script (e.g., analyze.py) that performs some form of analysis on the data, like summarization, statistical analysis, or visualization.

Makefile Creation:

- Create a Makefile to automate the execution of the above scripts. Your Makefile should include:
    - A target for preprocessing each dataset.
    - A target for analyzing each preprocessed dataset.
    - A target ‘all’ that runs all preprocessing and analysis.
    - A target ‘clean’ that removes all generated files.

Report Generation:

- Extend the pipeline with a script (e.g., generate_report.py) that creates a report (text or HTML) summarizing the results of the analysis.
- Update the Makefile to include this step in the workflow.

Documentation:

- Document your Makefile clearly, explaining each target and its dependencies.
- Include a README.md file in your project that explains how to run the Makefile and what each script does.

Advanced (Optional):

- Add error handling in your scripts to manage cases where data may be missing or incorrect.
- Implement a feature in your Makefile that allows processing of new data files added to the data directory without modifying the Makefile.

Submission:

- Submit a link to your GitHub repository. The directory should include:
    - The data folder with datasets.
    - Your scripts for preprocessing, analysis, and report generation.
    - The Makefile.
    - The README.md file.

Evaluation Criteria:

- Correctness and efficiency of the scripts.
- Functionality and completeness of the Makefile.
- Clarity and accuracy of the documentation.
- (For advanced section) Robustness and scalability of the Makefile setup.

Deadline:

Submit the assignment by [deadline].