Data Science Final Project Assignment

Project Objective:

The goal is for each group to analyze publicly available datasets to uncover insights on societal trends (e.g., health, economic development, environmental impacts, etc.). Each group will gather, transform, and analyze data, culminating in a presentation and report that outlines their findings, ETL process, and any cloud storage solutions used.

Project Breakdown

1. Data Selection and Exploration (20%)

Objective: Each group selects a topic and finds at least two datasets from open data sources (Kaggle, Google Dataset Search, government portals, etc.).

Tasks:

- - Identify datasets relevant to the chosen topic.
  - Explore datasets to understand their structure, variables, and potential value.
  - Document the rationale for choosing the datasets, including the expected insights.

2. ETL Setup (15%)

Objective: Design and document an ETL pipeline to clean, transform, and store the data in a suitable format for analysis.

Tasks:

- - Define ETL steps: extraction (loading from sources), transformation (cleaning, filtering, structuring), and loading (MySQL/MongoDB).
  - Provide a flowchart or diagram of the ETL pipeline.
  - Discuss data storage considerations and any cloud storage requirements.

3. ETL Implementation (20%)

Objective: Develop the ETL pipeline in Python.

Tasks:

- - Code the ETL steps, loading data from source(s) and storing it in a MySQL or MongoDB database.
  - Ensure the ETL script can handle updates to data sources and is designed for reproducibility.
  - Use comments and structure the code clearly for readability.

4. Data Analysis (25%)

Objective: Analyze the cleaned and transformed data to extract meaningful insights.

Tasks:

- - Develop Python scripts for exploratory data analysis, visualizations, and basic statistical analysis.
  - Create visualizations that effectively communicate insights (e.g., trends, distributions).
  - Write a summary of findings, supported by visualizations and statistics.

5. Cloud Storage and Documentation (10%)

Objective: Store transformed data in Google Cloud and document the cloud storage setup.

Tasks:

- - Set up Google Cloud storage for the transformed data.
  - Document the process, including credentials management and access control.
  - Ensure data is accessible for the analysis step.

6. Reflection Paper (5%)

Objective: Each group writes a 2 to 4-page reflection paper on the project experience.

Tasks:

- - Reflect on the challenges faced during data selection, ETL setup and implementation, analysis, and cloud storage.
  - Discuss lessons learned, particularly focusing on technical challenges, team coordination, and any improvements for future projects.
  - Summarize skills gained and areas for further development.

7. Presentation (5%)

Objective: Each group presents their project, with a focus on challenges and lessons learned in both the process and insights.

Tasks:

- - Include sections on data selection, ETL process, analysis, and cloud storage.
  - Highlight the main challenges encountered and how they were addressed.
  - Share key insights and takeaways, emphasizing both technical and analytical lessons learned.
  - Ensure equal participation from all group members.