Data Engineering

Assignment 1: Healthcare Analytics

Overview:

This assignment is designed to assess your practical skills and problem-solving abilities in the field of data engineering. This assignment will test your understanding of data processing, transformation, and integration. You will be required to ingest, transform and analyze healthcare data.

Data Source:

You will be using the "**Health Insurance Marketplace Analysis**" from Kaggle. The dataset represents individual medical costs billed by health insurance and includes the following fields: age, sex, bmi, children, smoker, region, and charges. You can download the dataset from this **link**.

Tasks:

1. **Ingestion and ETL:** Load the dataset. Create an ETL job using the framework of your preference, such as AWS Glue, Databricks, Apache Airflow, Talend, or any other tool you are comfortable with. Perform necessary transformations, and store the transformed data. Choose an appropriate data storage service.
   Bonus: Develop unit tests for your ETL job and data analysis queries.
2. **Data Quality Checks:** Implement data quality checks like checking for null values or verifying the transformation logic output.
3. **Data Analysis:** Use SQL queries to analyze the data.
4. **Output:**

   Understand the Market Dynamics for Health Insurance Plans & develop a Targeted Marketing Strategy

   Understand the distribution and characteristics of health insurance plans across different states and age groups to identify potential markets to target or areas where the company could improve its offerings. Understand the value proposition of different plans and recommend the top 5 avenues where marketing effort should be spent by plan, age, and state.

Submission Guidelines:

1. **Code Submission:** Submit your well-commented code files via repository link. Include all necessary files and instructions to run the code.
2. **Documentation:** Include a README file explaining your approach, technologies used, and any challenges faced. Describe how to run the code and any additional setup required.

Evaluation Criteria:

Your assignment will be evaluated based on the following criteria:

1. **Data Accuracy:** The accuracy of the processed data after extraction, transformation, and integration.
2. **Code Quality:** The readability, organization, and efficiency of your code.
3. **Error Handling:** How well you handle errors and edge cases during data engineering processes.