Assignment: Predict expected_`total_claim_amount` Using Regression.

1. Objective

You have an insurance dataset with various features such as age, policy_annual_premium, insured_hobbies, capital-gains, incident_severity, and more. Your goal is to build a regression model that predicts the `total_claim_amount` for each record.

2. Dataset Overview

Focus on identifying relevant predictors (e.g., `age`, `policy_annual_premium`, `insured_sex`, `insured_hobbies`, `capital-gains`, etc.) that can help estimate `total_claim_amount`.

1. Explore the dataset and identify features relevant to risk (e.g., `age`, `policy_annual_premium`, `insured_hobbies`, `capital-gains`, `capital-loss`, etc.).
2. Clean and preprocess the data, handling missing values, encoding categorical features, and creating any derived features that may help predict a risk score.
3. Train a regression model
4. Evaluate the model's performance using metrics such as RMSE (Root Mean Squared Error) or MAE (Mean Absolute Error) etc
5. Try to design a parameter Risk Score based on expected_total_claim_amount and other fields.
6. Good to have - Categorize policies into high, medium, and low risk using classification or clustering techniques
7. High level architecture to solve this problem with advanced LLM models like OpenAI, Gemini, DeepSeek etc

3. Deliverables

- Code and Notebook
    - A Jupyter Notebook , Google Colab or equivalent demonstrating:
        - Data loading and preprocessing steps (handling missing values, encoding, scaling).
        - Model training, evaluation, and interpretation (e.g., classification report, cluster centroids).
- Results and Visualizations
    - For classification:
        - Accuracy, precision, recall, F1-scores, and confusion matrix.
        - Optional ROC curves (if treating it as a one-vs-rest or binary problem).
    - For clustering:
        - Cluster assignment distribution (how many items in each cluster).

        - Centroid analysis or average feature values per cluster.
        - Silhouette scores or elbow plot (if applicable).
- Short Report or Presentation
    - Methodology: Explain why you chose classification or clustering (or both).
    - Key Findings: Summarize which features are most influential and how the model or clusters identify high, medium, and low risk.
    - Limitations and Next Steps: Discuss any data constraints, potential improvements, and how this approach could be integrated into a production environment.