**Project Summary: Life Insurance Claim Risk Modeling using GLMs in R**

**Objective:**  
To build a predictive model that estimates the probability of a life insurance claim based on policyholder characteristics, using logistic regression techniques inspired by CS1 actuarial concepts.

**Dataset:**

* Source: Kaggle Medical Cost Personal Dataset
* Size: 1338 records
* Key features: age, bmi, smoker, charges, region, children, sex
* Target variable created: claim\_made (1 if charges > median, else 0)

**Methodology:**

1. **Data Cleaning & Preparation:**
   * Loaded and cleaned data using R and Excel
   * Created binary target variable claim\_made to indicate high-risk policyholders
2. **Model Building:**
   * Used **Generalized Linear Model (GLM)** with logistic regression in R
   * Formula: claim\_made ~ age + smoker + bmi
   * Interpreted coefficients to evaluate impact of variables on claim risk
3. **Evaluation Metrics:**
   * Accuracy: ~78%
   * AUC (Area Under Curve): ~0.85
   * ROC curve plotted to assess classification performance
4. **Visualization:**
   * Created risk plots to visualize predicted claim probabilities by age and smoker status
   * Saved key visuals (claim\_risk\_plot.png, roc\_curve.png) for communication and documentation

**Tools Used:**

* **R:** glm(), ggplot2, pROC, dplyr
* **Excel:** Initial data filtering and exploration
* **(Optional)** SQL: Simulated queries using sqldf package in R

**Key Findings:**

* **Smoker status** and **age** were the most significant predictors of high claim risk
* The model successfully classified high-risk individuals for potential pricing or underwriting decisions
* Visualization highlighted clear risk differentiation between smoker and non-smoker groups