**Business Problem Statement**

A health insurance provider aims to better understand the key factors influencing medical insurance costs to improve pricing accuracy, customer satisfaction, and profitability. The management team has observed significant variations in charges among different demographic and lifestyle groups — particularly across age, BMI, smoking status, and family size.

They are interested in uncovering **how customer characteristics contribute to higher medical costs** and **which risk factors drive premium adjustments or claims**. The ultimate goal is to use data insights to support more equitable pricing models, identify cost-reduction opportunities, and enhance risk management strategies.

You are tasked with analyzing the company’s insurance dataset to answer the following overarching business question:

**“How can data on demographics, lifestyle, and regional factors be used to understand and manage the drivers of medical insurance costs?”**

**Deliverables**

1. **Data Preparation & Modeling (Python):**  
   Clean and transform the raw Kaggle insurance dataset for analysis, including creation of categorical features such as age group and BMI group.
2. **Data Analysis (MySQL):**  
   Load the cleaned data into a MySQL database and execute advanced SQL queries to extract insights on cost distribution, risk factors, and customer segments.
3. **Visualization & Insights (Power BI):**  
   Build an interactive dashboard highlighting key drivers of charges, regional cost variations, and potential savings scenarios (e.g., smoker reduction).
4. **Report and Presentation:**  
   Summarize findings, visual insights, and actionable business recommendations for optimizing insurance pricing and health policy decisions.
5. **GitHub Repository:**  
   Include all Python scripts, SQL queries, Power BI files, and documentation in a structured repository for transparency and reproducibility.