# Healthcare Insurance Analysis – Solution Writeup

## 1. Introduction

The rising cost of healthcare is a major concern for individuals and insurance providers alike. This project aims to predict hospitalization costs and understand the key contributing factors. It involves integrating multiple datasets, cleaning and transforming data, performing statistical analysis, and building machine learning models to support strategic decisions in healthcare insurance.

## 2. Data Integration and Cleaning

The project utilizes three datasets:

- **Hospitalization details**
- **Medical examinations**
- **Names dataset**

Using Customer ID as a common identifier, we merge these datasets into a unified dataframe. Data cleaning steps included:

- **Handling Missing Values**: Removed rows with significant missing or trivial entries such as ?.
- **Type Conversion**: Converted categorical data into appropriate numerical or dummy formats.
- **Data Normalization**: Standardized numerical variables to prepare for modeling.
- **Outlier Detection**: Identified and addressed outliers in the charges variable using visual methods.

## 3. Feature Engineering

Several new features were engineered to enhance model performance:

- **Age Calculation**: Derived from the patient's date of birth.
- **Gender Identification**: Inferred from salutations in the names.
- **Categorical Transformations**:
  - Hospital and city tiers were encoded using ordinal and one-hot encoding.
  - State ID was filtered to include only R1011, R1012, and R1013 for dummy creation.
- **Surgery Count Cleaning**: Converted NumberOfMajorSurgeries to integer values.
- **Lifestyle and Health Indicators**: Smoking status, heart issues, diabetes, and BMI were all standardized.

---

## 4. Exploratory Data Analysis

To gain insights into the distribution and variance of hospitalization costs, we employed:

- **Histograms, Boxplots, and Swarm Plots**: To visualize cost distribution.

- **Radar Charts**: To show median hospitalization costs by hospital tier.
- **Stacked Bar Charts**: To illustrate the population distribution across hospital and city tiers.

We observed higher costs in tier-1 hospitals and cities, and a noticeable skew toward higher costs for smokers and patients with comorbidities.

### 5. Hypothesis Testing

We conducted hypothesis testing to evaluate:

- **Hospital Tier Differences**: ANOVA confirmed significant cost differences among hospital tiers.
- **City Tier Differences**: ANOVA indicated significant variation among city tiers.
- **Smoker vs Non-Smoker Costs**: A t-test revealed a significant cost increase among smokers.
- **Independence of Smoking and Heart Issues**: A chi-square test showed a significant association.

### 6. Machine Learning Modeling

#### Correlation Analysis

A heatmap was used to detect multicollinearity among features, allowing us to drop or combine redundant predictors.

#### Model Development

- **Linear and Ridge Regression**: Applied for baseline performance.
- **Gradient Boosting**: Implemented to capture nonlinear interactions and improve accuracy.
- **Pipelines and Cross-Validation**:
    - Used sklearn pipelines for data preprocessing.
    - Applied 5-fold stratified cross-validation for robust model evaluation.
    - Hyperparameter tuning was conducted via GridSearchCV.

#### Feature Importance

Gradient Boosting revealed that **age**, **BMI**, **hospital tier**, **diabetic status**, and **smoking** were among the top predictors.

### 7. Case Study: Cost Estimation for Ms. Jayna

Using the best-performing model, we predicted the hospitalization cost for a hypothetical patient, Ms. Jayna:

- **Profile**: 36-year-old female, BMI of 29.4, diabetic-negative, smoker, lives in a tier-1 city and hospital.

- **Prediction**: The estimated hospitalization cost was derived using the Gradient Boosting model with her encoded features.

## 8. SQL Analysis

Key SQL insights included:

- **Average Metrics**: For diabetic patients with heart problems (age, children, BMI, cost).
- **Cost Comparisons**: Across hospital and city tiers.
- **Surgical and Cancer History**: Number of patients with both.
- **Hospital Distribution**: Count of tier-1 hospitals per state.

## 9. Tableau Dashboard

A comprehensive Tableau dashboard was created to convey key insights with visual clarity. Components included:

- Hba1C and BMI Histogram
- Hospital tiers based on State
- Cost Distribution by Hospital/City Tier
- Bar Charts

## 10. Conclusion

This analysis highlighted the importance of demographic, medical, and regional factors in predicting healthcare costs. The machine learning model, backed by thorough EDA and statistical testing, provides a robust framework for healthcare cost estimation. Insurance providers can leverage these insights for premium setting, risk assessment, and strategic planning.