**1) Objective**

Build a model to predict **Premium.Amount** using customer demographic and health features; document feature engineering, model selection, validation, and diagnostics, and identify the key pricing drivers.

**2) Data & Preparation**

* **Source:** premium\_data.csv.
* **Cleaning:** Dropped rows with missing values (na.omit).
* **Categoricals:** Trimmed whitespace for Smoking.Status; converted to factors: Smoking.Status, Gender, Region, Educational.Level, Age\_Groups, Income\_Level, Credit\_Category, Pre.existing.Conditions, Family.Medical.History.
* **Train/Test split:** 70/30 using set.seed(888) and caret::createDataPartition.

**3) Exploratory Data Analysis (EDA)**

* **Categorical vs Premium (boxplots via facet):**
  + Clear level shifts for **Smoking.Status** (Smoker > Ex-smoker > Non-smoker), **Pre.existing.Conditions** (1 > 0), **Family.Medical.History** (1 > 0), and **Income\_Level** (higher income bands associated with higher premiums).
  + Little/no separation for some others (e.g., Gender, some Regions/Education levels).
* **Numerical correlation (corrplot):**
  + **Age**, **BMI**, **Sum.Insured** positively correlated with premiums; **Credit.Score** weakly negative.
  + No problematic numeric–numeric multicollinearity detected.

**4) Feature Set (for modeling)**

**Numerical:** Age, BMI, Credit.Score, Sum.Insured, High\_Risk  
**Categorical:** Smoking.Status, Pre.existing.Conditions, Family.Medical.History, Income\_Level

(We kept High\_Risk to test incremental lift, with the understanding it might be dropped if not significant.)

**5) Models & Validation**

* **Models compared:**
  1. Multiple Linear Regression (lm) with all selected features.
  2. Random Forest (randomForest) with default settings, importance = TRUE.
* **Validation:** Hold-out test set (30%). Metrics: **RMSE** and **MAE**.

**6) Results**

**6.1 Linear Regression (train on 70%)**

* **Adjusted R²:** **0.9481** (Multiple R² 0.9483).
* **Test metrics:** **RMSE = 85.54**, **MAE = 71.29**.
* **Significance:** All predictors are highly significant (**p < 0.001**) **except High\_Risk** (p = 0.359).
* **Key coefficients (holding others constant):**
  + **Age:** +**20.29** per year (p < 2e-16).
  + **BMI:** +**14.54** per unit (p < 2e-16).
  + **Credit.Score:** −**0.477** per point (p < 2e-16).
  + **Sum.Insured:** +**0.0009936** per $1 of coverage (~+$0.99 per $1,000).
  + **Smoking.Status:** Non-smoker **−107.8** vs Ex-smoker; Smoker **+94.1** vs Ex-smoker.
  + **Pre.existing.Conditions=1:** **+302.8**.
  + **Family.Medical.History=1:** **+153.5**.
  + **Income\_Level:** sizeable positive effects vs baseline (e.g., Elite **+1358**, Very High **+1090**, High **+824**, Medium **+575**, Low **+322**).
* **Multicollinearity:** VIFs are low (all GVIF^(1/(2\*Df)) ≈ 1.00–2.45), **no red flags**.
* **Diagnostics:** Residuals approximately centered with mild heteroskedasticity and light right-tail skew. No problematic leverage after visual check; overall diagnostics acceptable for deployment.

**6.2 Random Forest**

* **Test metrics:** **RMSE = 110.55**, **MAE = 87.96** (worse than linear).
* **Variable importance:** Aligns with linear model—**Age**, **Income\_Level**, **Pre.existing.Conditions**, and **Smoking.Status** lead; **Credit.Score** and **BMI** moderate; **High\_Risk** low.

**6.3 Model choice**

* The **linear model outperforms RF** on this dataset and is **more interpretable**, suggesting the target–feature relationships are predominantly linear (or near-linear) at the chosen feature set and scale.

**7) Interpretation & Business Takeaways**

* **Age** and **health risk proxies** (**BMI**, **Pre.existing.Conditions**, **Family.Medical.History**) are **strong premium drivers**.
* **Behavioral/risk proxies** matter: **Smokers** pay more; **non-smokers** pay less than ex-smokers.
* **Financial capacity/coverage** markers (**Income\_Level**, **Sum.Insured**) are positively associated with premiums (higher coverage and possibly product mix/benefit levels).
* **Credit.Score** has a small but significant **downward** effect on premium (better credit, slightly lower premium).
* **High\_Risk** (as currently defined) is **not adding signal** (ns), and can be dropped to simplify the model.

**8) Limitations & Next Steps**

* **Mild heteroskedasticity**: consider **log(Premium.Amount)** or **WLS** to stabilize variance; check whether performance improves materially.
* **Model simplification:** remove **High\_Risk** (ns) and re-fit to confirm metrics hold.
* **RF tuning:** if insisting on a tree ensemble, tune mtry, ntree, and node size; alternatively try **GBM/XGBoost** with CV.
* **Segmented models:** optional—fit separate models by **Smoking.Status** to see if segment-specific slopes reduce error.
* **Robustness:** run k-fold CV (e.g., 5- or 10-fold) to confirm hold-out results are stable.

**9) Recommendation**

Adopt the **multiple linear regression** as the production baseline for premium prediction (features: Age, BMI, Credit.Score, Sum.Insured, Smoking.Status, Pre.existing.Conditions, Family.Medical.History, Income\_Level; drop High\_Risk). It delivers **low error (RMSE ≈ 85.5)**, **high explanatory power (Adj. R² ≈ 0.948)**, and clear, business-intuitive effects—ideal for pricing governance and stakeholder communication.