Code:

```
# ============================================================================
#  PROJECT  : University–Insurance Student Health Plan — GLM Actuarial Pricing
#  LANGUAGE : R  |  Tidyverse + GLM (Generalized Linear Models)
#  AUDIENCE : Actuaries, Data Scientists, University Risk Officers
#
#  BUSINESS QUESTIONS:
#    1. RISK FACTORS  — Does high BMI or low sleep increase medical visits?
#    2. COST PRICING  — Can we predict a student's Expected Medical Cost?
#
#  WHY GLM?
#    Medical costs are always positive and right-skewed (a few students cost
#    a LOT). Ordinary linear regression (OLS) assumes Normal errors and can
#    predict negative costs — which is impossible. A GLM with a Gamma
#    distribution handles skewed, positive-only data correctly. This is the
#    actuarial industry standard (used by every major insurer worldwide).
# ============================================================================


# ──────────────────────────────────────────────────────────────────────────
# STEP 1 ▶  LOAD LIBRARIES
# ──────────────────────────────────────────────────────────────────────────
# We need three packages:
#   • tidyverse  — data wrangling (dplyr) + beautiful charts (ggplot2)
#   • broom      — tidy GLM output into clean data frames
#   • scales     — format axis labels as currency / percentages

packages <- c("tidyverse", "broom", "scales", "gridExtra", "ggthemes")

for (pkg in packages) {
  if (!require(pkg, character.only = TRUE, quietly = TRUE)) {
    install.packages(pkg, repos = "https://cran.r-project.org")
    library(pkg, character.only = TRUE)
  }
}

cat("✅ STEP 1 COMPLETE — All libraries loaded\n")


# ──────────────────────────────────────────────────────────────────────────
# STEP 2 ▶  SIMULATE STUDENT HEALTH DATA
# ──────────────────────────────────────────────────────────────────────────
# In a real project, this data comes from:
#   • University Health Center EHR (Electronic Health Records)
#   • Student Lifestyle Surveys
#   • Insurance Claims History
#
```

```
# We simulate 2,000 students with realistic statistical distributions.
# The relationships we build in reflect real actuarial research.

set.seed(123)   # Always set a seed → reproducible results
N <- 2000       # 2,000 students

data <- tibble(
  # ── Student Identifiers ──────────────────────────────────────────
  Student_ID = 1:N,
  Age        = sample(18:26, N, replace = TRUE),
  Gender     = sample(c("Male", "Female", "Other"), N,
                      replace = TRUE, prob = c(0.45, 0.50, 0.05)),
  Year       = sample(c("Freshman","Sophomore","Junior","Senior","Grad"), N,
                      replace = TRUE, prob = c(0.25, 0.22, 0.20, 0.18, 0.15)),

  # ── Lifestyle Risk Factors (our KEY predictors) ──────────────────
  BMI          = round(rnorm(N, mean = 24.5, sd = 4.2) |> pmax(15) |> pmin(50), 1),
  Sleep_Hours  = round(rnorm(N, mean = 6.8,  sd = 1.2) |> pmax(3)  |> pmin(10), 1),
  Stress_Score = sample(1:10, N, replace = TRUE),       # 1 = calm, 10 = very stressed
  Exercise_Days = round(pmax(0, rnorm(N, mean = 3.5, sd = 2.0)), 1),
  Smoker       = sample(c(0L, 1L), N, replace = TRUE, prob = c(0.85, 0.15)),
  Alcohol      = sample(c("None","Moderate","Heavy"), N,
                        replace = TRUE, prob = c(0.40, 0.45, 0.15)),
  Chronic_Cond = sample(c(0L, 1L), N, replace = TRUE, prob = c(0.80, 0.20))
  # 0 = no chronic condition, 1 = has asthma / diabetes / similar
)

# ── Build Medical Cost using a log-linear (GLM) structure ──────────
# This mirrors how actuaries model cost: each factor multiplies risk
# We use exp() so costs are always positive and effects are multiplicative.

data <- data %>%
  mutate(
    # The "true" log-scale risk score
    Log_Risk = 7.2                          # baseline ~$1,330/year
    + 0.025  * (BMI - 24.5)            # obesity risk factor
    - 0.080  * (Sleep_Hours - 6.8)       # sleep deprivation
    + 0.040  * (Stress_Score - 5)        # stress loading
    - 0.050  * (Exercise_Days - 3.5)     # exercise discount
    + 0.35   * Smoker                  # smoking surcharge
    + 0.15   * (Alcohol == "Heavy")      # heavy alcohol loading
    + 0.60   * Chronic_Cond            # chronic condition surcharge
    + rnorm(N, mean = 0, sd = 0.20),     # irreducible noise

    # Actual medical cost = exp(log-risk) → always positive, right-skewed
    Medical_Cost = round(exp(Log_Risk), 2),

    # Annual doctor visits follow a Poisson distribution (count data)
```

```
  Annual_Visits = rpois(N, lambda = exp(
    1.2
    + 0.025 * (BMI - 24.5)
    - 0.080 * (Sleep_Hours - 6.8)
    + 0.040 * (Stress_Score - 5)
    - 0.050 * (Exercise_Days - 3.5)
    + 0.30  * Smoker
    + 0.55  * Chronic_Cond
    + rnorm(N, 0, 0.15)
  ))
 )

cat("✅ STEP 2 COMPLETE — Dataset created:", N, "students,",
  ncol(data), "variables\n")
cat("   Avg Medical Cost : $", round(mean(data$Medical_Cost), 0), "\n")
cat("   Avg Annual Visits:", round(mean(data$Annual_Visits), 1), "\n")


# ──────────────────────────────────────────────────────────────────────────
# STEP 3 ▶  FEATURE ENGINEERING
# ──────────────────────────────────────────────────────────────────────────
# Raw numbers → interpretable risk categories.
# Actuaries use "risk bands" to segment populations and set premium tiers.

data <- data %>%
 mutate(
  # ── BMI Risk Band (WHO classification) ──────────────────────────
  BMI_Category = case_when(
    BMI < 18.5            ~ "Underweight",
    BMI >= 18.5 & BMI < 25 ~ "Normal",
    BMI >= 25  & BMI < 30  ~ "Overweight",
    BMI >= 30             ~ "Obese"
  ) |> factor(levels = c("Underweight","Normal","Overweight","Obese")),

  # ── Sleep Risk Band ─────────────────────────────────────────────
  Sleep_Category = case_when(
    Sleep_Hours < 5  ~ "Severe Deprivation",
    Sleep_Hours < 6  ~ "Poor (<6hrs)",
    Sleep_Hours < 7  ~ "Fair (6-7hrs)",
    Sleep_Hours < 8  ~ "Good (7-8hrs)",
    TRUE             ~ "Ideal (8+hrs)"
  ) |> factor(levels = c("Severe Deprivation","Poor (<6hrs)",
                         "Fair (6-7hrs)","Good (7-8hrs)","Ideal (8+hrs)")),

  # ── Actuarial Risk Tier (used for premium segmentation) ─────────
  # This is the output the insurance company needs for pricing tiers
  Risk_Tier = case_when(
    Chronic_Cond == 1 | Smoker == 1 | BMI >= 30 ~ "High Risk",
```

```
    BMI >= 25 | Sleep_Hours < 6 | Stress_Score >= 8 ~ "Medium Risk",
    TRUE ~ "Low Risk"
  ) |> factor(levels = c("Low Risk","Medium Risk","High Risk")),

  # ── Binary flags for GLM (interactions) ──────────────────────
  Is_Obese       = as.integer(BMI >= 30),
  Sleep_Deprived = as.integer(Sleep_Hours < 6),
  High_Stress    = as.integer(Stress_Score >= 8),
  Heavy_Alcohol  = as.integer(Alcohol == "Heavy"),

  # ── Compound interaction term: obese AND chronic = very high risk ─
  Obese_x_Chronic = Is_Obese * Chronic_Cond
)

cat("\n✅ STEP 3 COMPLETE — Feature engineering done\n")
cat("   Risk Tier Breakdown:\n")
print(table(data$Risk_Tier))


# ═══════════════════════════════════════════════════════════════
# STEP 4 ▶  EXPLORATORY DATA ANALYSIS (EDA)
# ═══════════════════════════════════════════════════════════════
# Before modeling, we visualize to CONFIRM relationships exist in the data.
# Actuaries call this "univariate analysis."

cat("\n📊 STEP 4 — Generating EDA Visualizations...\n")

# ── Colour palette ──────────────────────────────────────────────
clr_risk  <- c("Low Risk"="#16A34A", "Medium Risk"="#D97706", "High Risk"="#DC2626")
clr_smoke <- c("0"="#2563EB", "1"="#DC2626")

# ── Plot 4A: Cost Distribution (always right-skewed in insurance!) ──
p_dist <- ggplot(data, aes(x = Medical_Cost)) +
  geom_histogram(fill = "#2563EB", color = "white", bins = 60, alpha = 0.85) +
  geom_vline(aes(xintercept = mean(Medical_Cost)), color = "#DC2626",
             linetype = "dashed", linewidth = 1.1) +
  geom_vline(aes(xintercept = median(Medical_Cost)), color = "#16A34A",
             linetype = "dashed", linewidth = 1.1) +
  annotate("text", x = mean(data$Medical_Cost) * 1.15, y = Inf,
           label = paste0("Mean\n$", comma(round(mean(data$Medical_Cost)))),
           color = "#DC2626", vjust = 1.5, fontface = "bold", size = 3.2) +
  annotate("text", x = median(data$Medical_Cost) * 0.70, y = Inf,
           label = paste0("Median\n$", comma(round(median(data$Medical_Cost)))),
           color = "#16A34A", vjust = 1.5, fontface = "bold", size = 3.2) +
  scale_x_continuous(labels = dollar) +
  labs(
    title    = "A. Medical Cost Distribution",
    subtitle = "Right-skewed → GLM with Gamma family is the correct tool (not OLS)",
```

```
    x = "Annual Medical Cost (USD)", y = "Number of Students"
  ) +
  theme_minimal(base_size = 11) +
  theme(plot.title = element_text(face = "bold"))

# ── Plot 4B: Risk Tier Boxplot ──────────────────────────────────────────
p_risk_box <- ggplot(data, aes(x = Risk_Tier, y = Medical_Cost, fill = Risk_Tier)) +
  geom_boxplot(alpha = 0.75, outlier.alpha = 0.2, outlier.size = 0.8) +
  stat_summary(fun = mean, geom = "point", shape = 23,
               size = 3, fill = "white", color = "black") +
  scale_fill_manual(values = clr_risk) +
  scale_y_continuous(labels = dollar) +
  labs(
    title    = "B. Cost by Actuarial Risk Tier",
    subtitle = "Diamonds = mean | High Risk shows wider spread (harder to price)",
    x = "Risk Tier", y = "Annual Medical Cost ($)"
  ) +
  theme_minimal(base_size = 11) +
  theme(legend.position = "none", plot.title = element_text(face = "bold"))

# ── Plot 4C: BMI vs Cost with GLM smooth ────────────────────────────────
p_bmi <- ggplot(data, aes(x = BMI, y = Medical_Cost, color = as.factor(Smoker))) +
  geom_point(alpha = 0.25, size = 0.9) +
  geom_smooth(method = "glm",
              method.args = list(family = Gamma(link = "log")),
              se = TRUE, linewidth = 1.4) +
  scale_color_manual(values = clr_smoke,
                     labels = c("0" = "Non-Smoker", "1" = "Smoker")) +
  scale_y_continuous(labels = dollar) +
  labs(
    title    = "C. BMI vs Medical Cost (by Smoking Status)",
    subtitle = "Smokers show a steeper cost curve — key pricing factor",
    x = "BMI", y = "Annual Medical Cost ($)", color = "Smoker"
  ) +
  theme_minimal(base_size = 11) +
  theme(plot.title = element_text(face = "bold"))

# ── Plot 4D: Sleep vs Annual Visits ─────────────────────────────────────
avg_by_sleep <- data %>%
  group_by(Sleep_Category) %>%
  summarise(Avg_Visits = mean(Annual_Visits),
            SE = sd(Annual_Visits) / sqrt(n()),
            .groups = "drop")

p_sleep <- ggplot(avg_by_sleep, aes(x = Sleep_Category, y = Avg_Visits)) +
  geom_col(fill = "#2563EB", alpha = 0.85, width = 0.6) +
  geom_errorbar(aes(ymin = Avg_Visits - SE, ymax = Avg_Visits + SE),
                width = 0.2, color = "grey30") +
```

```
  geom_text(aes(label = round(Avg_Visits, 1)), vjust = -0.8,
            fontface = "bold", size = 3.5) +
  labs(
    title    = "D. Sleep Deprivation vs Doctor Visits",
    subtitle = "Less sleep → more visits (confirms our risk factor hypothesis)",
    x = "Sleep Category", y = "Avg Annual Doctor Visits"
  ) +
  theme_minimal(base_size = 11) +
  theme(axis.text.x = element_text(angle = 20, hjust = 1),
        plot.title = element_text(face = "bold"))

# ── Combine and save EDA dashboard ──────────────────────────────────────
eda_dashboard <- gridExtra::arrangeGrob(
  p_dist, p_risk_box, p_bmi, p_sleep,
  ncol = 2,
  top = grid::textGrob(
    "STEP 4: Exploratory Data Analysis — Student Health Risk Factors",
    gp = grid::gpar(fontsize = 14, fontface = "bold")
  )
)

ggsave("step4_eda_dashboard.png", plot = eda_dashboard,
       width = 14, height = 10, dpi = 150, bg = "white")

cat("   ✓ Saved: step4_eda_dashboard.png\n")
cat("✅ STEP 4 COMPLETE\n")


# ═══════════════════════════════════════════════════════════════════════
# STEP 5 ▶  BUILD THE GLM — MEDICAL COST MODEL
# ═══════════════════════════════════════════════════════════════════════
# We use TWO models, as actuaries do in practice:
#
#   MODEL 1 (Poisson GLM)  → predicts NUMBER OF VISITS
#                            Count data must use Poisson, not Gaussian
#
#   MODEL 2 (Gamma GLM)    → predicts MEDICAL COST per student
#                            Costs are positive and skewed → Gamma family
#
# Both use a "log" link function, meaning effects are MULTIPLICATIVE:
#   cost = exp(β0 + β1·BMI + β2·Sleep + ...)
#   → Each coefficient is a % change in expected cost

cat("\n🔬 STEP 5 — Building GLM Models...\n")

# ── 80/20 Train/Test Split ──────────────────────────────────────────────
train_idx <- sample(seq_len(N), size = round(0.80 * N))
```

```
train     <- data[ train_idx, ]
test      <- data[-train_idx, ]

cat("   Train size:", nrow(train), "| Test size:", nrow(test), "\n\n")

# ── MODEL 1: Poisson GLM — Predict Annual Doctor Visits ──────────────────
# Why Poisson? Visit counts are non-negative integers (0, 1, 2, 3 ...)
# The Poisson GLM is the actuarial standard for frequency modeling.

model_visits <- glm(
  Annual_Visits ~ BMI + Sleep_Hours + Stress_Score + Exercise_Days +
    Smoker + Heavy_Alcohol + Chronic_Cond + Obese_x_Chronic,
  family = poisson(link = "log"),
  data   = train
)

cat("── MODEL 1: Poisson GLM (Visit Frequency) ──────────────────────────\n")
cat("   Null Deviance    :", round(model_visits$null.deviance, 1), "\n")
cat("   Residual Deviance:", round(model_visits$deviance, 1), "\n")
cat("   AIC              :", round(AIC(model_visits), 1), "\n\n")

# Tidy coefficient table with interpretation
coef_visits <- tidy(model_visits) %>%
  mutate(
    IRR       = round(exp(estimate), 3),     # Incidence Rate Ratio
    Pct_Change = paste0(ifelse(IRR >= 1, "+", ""), round((IRR - 1) * 100, 1), "%"),
    Significant = ifelse(p.value < 0.05, "✓ Yes", "✗ No"),
    p.value    = round(p.value, 4)
  ) %>%
  select(term, estimate, IRR, Pct_Change, p.value, Significant)

cat("   Coefficient Table (IRR = how much each factor multiplies visit rate):\n")
print(coef_visits, n = Inf)


# ── MODEL 2: Gamma GLM — Predict Annual Medical Cost ─────────────────────
# Why Gamma? Medical costs are:
#   (a) Always positive
#   (b) Right-skewed (a few very costly students pull the mean up)
#   (c) Variance increases with the mean
# The Gamma distribution captures all three properties perfectly.

model_cost <- glm(
  Medical_Cost ~ BMI + Sleep_Hours + Stress_Score + Exercise_Days +
    Smoker + Heavy_Alcohol + Chronic_Cond + Obese_x_Chronic,
  family = Gamma(link = "log"),
  data   = train
)
```

```
cat("\n── MODEL 2: Gamma GLM (Medical Cost Severity) ──────────────────────\n")
cat("   Null Deviance    :", round(model_cost$null.deviance, 1), "\n")
cat("   Residual Deviance:", round(model_cost$deviance, 1), "\n")
cat("   AIC              :", round(AIC(model_cost), 1), "\n\n")

coef_cost <- tidy(model_cost) %>%
  mutate(
    Cost_Multiplier = round(exp(estimate), 3),
    Pct_Change      = paste0(ifelse(Cost_Multiplier >= 1, "+", ""),
                             round((Cost_Multiplier - 1) * 100, 1), "%"),
    Significant     = ifelse(p.value < 0.05, "✓ Yes", "✗ No"),
    p.value         = round(p.value, 4)
  ) %>%
  select(term, estimate, Cost_Multiplier, Pct_Change, p.value, Significant)

cat("   Coefficient Table (Multiplier = how much each factor scales cost):\n")
print(coef_cost, n = Inf)

cat("\n✅ STEP 5 COMPLETE — Both GLMs trained\n")


# ──────────────────────────────────────────────────────────────────────────
# STEP 6 ▶  MODEL EVALUATION ON TEST SET
# ──────────────────────────────────────────────────────────────────────────
# Never evaluate a model on training data — it will look unrealistically good.
# We use the 20% hold-out test set for honest evaluation.

cat("\n🩹 STEP 6 — Model Evaluation on Hold-Out Test Set\n")

test <- test %>%
  mutate(
    Predicted_Visits = predict(model_visits, newdata = test, type = "response"),
    Predicted_Cost   = predict(model_cost,   newdata = test, type = "response")
  )

# ── Visit Model Metrics ──────────────────────────────────────────────────
mae_visits  <- mean(abs(test$Annual_Visits - test$Predicted_Visits))
rmse_visits <- sqrt(mean((test$Annual_Visits - test$Predicted_Visits)^2))
cor_visits  <- cor(test$Annual_Visits, test$Predicted_Visits)

# ── Cost Model Metrics ───────────────────────────────────────────────────
mae_cost    <- mean(abs(test$Medical_Cost - test$Predicted_Cost))
rmse_cost   <- sqrt(mean((test$Medical_Cost - test$Predicted_Cost)^2))
cor_cost    <- cor(test$Medical_Cost, test$Predicted_Cost)

# Gini coefficient — primary accuracy metric used in insurance pricing
```

```
gini <- function(actual, predicted) {
  df   <- data.frame(actual, predicted) %>% arrange(predicted)
  n    <- nrow(df)
  cum_loss  <- cumsum(df$actual) / sum(df$actual)
  cum_pop   <- seq_len(n) / n
  2 * (sum(cum_loss * (1/n)) - 0.5)
}
gini_cost <- gini(test$Medical_Cost, test$Predicted_Cost)

cat("\n  ── Visit Frequency Model ──────────────────────────\n")
cat(sprintf("   MAE  (mean abs error)   : %.2f visits\n", mae_visits))
cat(sprintf("   RMSE                    : %.2f visits\n", rmse_visits))
cat(sprintf("   Correlation (Actual vs Predicted): %.3f\n", cor_visits))

cat("\n  ── Medical Cost Model ─────────────────────────────\n")
cat(sprintf("   MAE  (mean abs error)   : $%.0f\n", mae_cost))
cat(sprintf("   RMSE                    : $%.0f\n", rmse_cost))
cat(sprintf("   Correlation (Actual vs Predicted): %.3f\n", cor_cost))
cat(sprintf("   Gini Coefficient        : %.3f  (0=random, 1=perfect)\n", gini_cost))

cat("\n✅ STEP 6 COMPLETE — Model evaluation done\n")


# ═══════════════════════════════════════════════════════════════════════
# STEP 7 ▶  VISUALIZATION — MODEL RESULTS
# ═══════════════════════════════════════════════════════════════════════

cat("\n📊 STEP 7 — Generating Model Result Visualizations...\n")

# Predict on full dataset for all plots
data <- data %>%
  mutate(
    Predicted_Cost   = predict(model_cost,   newdata = data, type = "response"),
    Predicted_Visits = predict(model_visits, newdata = data, type = "response")
  )

# ── Plot 7A: Actual vs Predicted Cost (goodness-of-fit) ──────────────
p_fit <- ggplot(test, aes(x = Medical_Cost, y = Predicted_Cost)) +
  geom_point(aes(color = Risk_Tier), alpha = 0.40, size = 1.2) +
  geom_abline(slope = 1, intercept = 0, color = "black",
              linetype = "dashed", linewidth = 1.1) +
  geom_smooth(method = "lm", se = FALSE, color = "#2563EB", linewidth = 1.0) +
  scale_color_manual(values = clr_risk) +
  scale_x_continuous(labels = dollar) +
  scale_y_continuous(labels = dollar) +
  annotate("text", x = Inf, y = -Inf,
           label = paste0("Gini = ", round(gini_cost, 3),
                          "\nCorr = ", round(cor_cost, 3)),
```

```
        hjust = 1.1, vjust = -0.5, size = 3.5, fontface = "bold") +
  labs(
    title    = "A. Actual vs Predicted Medical Cost (Test Set)",
    subtitle = "Points on the dashed line = perfect prediction",
    x = "Actual Cost ($)", y = "Predicted Cost ($)", color = "Risk Tier"
  ) +
  theme_minimal(base_size = 11) +
  theme(plot.title = element_text(face = "bold"))

# ── Plot 7B: Coefficient plot — which factors drive cost most? ──────────────
# We remove the intercept for clarity and only show significant factors
coef_plot_data <- tidy(model_cost, conf.int = TRUE) %>%
  filter(term != "(Intercept)") %>%
  mutate(
    term = recode(term,
                  BMI            = "BMI",
                  Sleep_Hours    = "Sleep Hours",
                  Stress_Score   = "Stress Score",
                  Exercise_Days  = "Exercise Days",
                  Smoker         = "Smoker (Yes)",
                  Heavy_Alcohol  = "Heavy Alcohol (Yes)",
                  Chronic_Cond   = "Chronic Condition",
                  Obese_x_Chronic = "Obese × Chronic (Interaction)"
    ),
    Direction = ifelse(estimate > 0, "Increases Cost", "Decreases Cost"),
    Significant = p.value < 0.05
  )

p_coef <- ggplot(coef_plot_data,
                 aes(x = reorder(term, estimate), y = estimate,
                     color = Direction, alpha = Significant)) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "grey50") +
  geom_point(size = 4) +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.25, linewidth = 1) +
  scale_color_manual(values = c("Increases Cost" = "#DC2626",
                                "Decreases Cost" = "#16A34A")) +
  scale_alpha_manual(values = c("TRUE" = 1.0, "FALSE" = 0.3),
                     labels = c("TRUE" = "p<0.05", "FALSE" = "p≥0.05")) +
  coord_flip() +
  labs(
    title    = "B. GLM Coefficients — Cost Drivers (Log Scale)",
    subtitle = "Right of 0 = raises cost | Left of 0 = lowers cost | Faded = not significant",
    x = NULL, y = "Coefficient (log scale)", color = NULL, alpha = "Significant"
  ) +
  theme_minimal(base_size = 11) +
  theme(plot.title = element_text(face = "bold"))

# ── Plot 7C: Predicted Cost by Risk Tier (pricing output) ──────────────────
```

```
p_tier_cost <- ggplot(data, aes(x = Risk_Tier, y = Predicted_Cost, fill = Risk_Tier)) +
  geom_violin(alpha = 0.6, trim = FALSE) +
  geom_boxplot(width = 0.15, fill = "white", outlier.shape = NA) +
  stat_summary(fun = mean, geom = "point", shape = 23,
               size = 4, fill = "white", color = "black") +
  scale_fill_manual(values = clr_risk) +
  scale_y_continuous(labels = dollar) +
  labs(
    title    = "C. Predicted Cost Distribution by Risk Tier",
    subtitle = "Insurance Company uses this to SET PREMIUM BANDS",
    x = "Actuarial Risk Tier", y = "Predicted Annual Medical Cost ($)"
  ) +
  theme_minimal(base_size = 11) +
  theme(legend.position = "none", plot.title = element_text(face = "bold"))

# ── Plot 7D: Lift Chart — how well does the model separate risk? ──────────────
# Lift chart is the STANDARD insurance model validation tool
# It shows how much better the model is vs random assignment
lift_data <- data %>%
  arrange(Predicted_Cost) %>%
  mutate(
    Decile     = ntile(Predicted_Cost, 10),
    Decile_Lbl = paste0("D", Decile)
  ) %>%
  group_by(Decile, Decile_Lbl) %>%
  summarise(
    Avg_Actual    = mean(Medical_Cost),
    Avg_Predicted = mean(Predicted_Cost),
    .groups = "drop"
  )

p_lift <- ggplot(lift_data, aes(x = reorder(Decile_Lbl, Decile))) +
  geom_col(aes(y = Avg_Actual), fill = "#93C5FD", alpha = 0.9, width = 0.6) +
  geom_point(aes(y = Avg_Predicted), color = "#DC2626", size = 3.5) +
  geom_line(aes(y = Avg_Predicted, group = 1), color = "#DC2626", linewidth = 1.2) +
  scale_y_continuous(labels = dollar) +
  labs(
    title    = "D. Lift Chart — Model Validation (by Predicted Cost Decile)",
    subtitle = "Blue bars = actual avg cost | Red line = model prediction | D10 = highest risk",
    x = "Predicted Cost Decile (D1=Lowest Risk → D10=Highest Risk)",
    y = "Average Annual Medical Cost ($)"
  ) +
  theme_minimal(base_size = 11) +
  theme(plot.title = element_text(face = "bold"))

# ── Combine and save ──────────────────────────────────────────────────────────
model_dashboard <- gridExtra::arrangeGrob(
  p_fit, p_coef, p_tier_cost, p_lift,
```

```
  ncol = 2,
  top = grid::textGrob(
    "STEP 7: GLM Model Results — University–Insurance Health Plan Pricing",
    gp = grid::gpar(fontsize = 14, fontface = "bold")
  )
)

ggsave("step7_model_results.png", plot = model_dashboard,
       width = 15, height = 11, dpi = 150, bg = "white")

cat("  ✓ Saved: step7_model_results.png\n")
cat("✅ STEP 7 COMPLETE\n")


# ──────────────────────────────────────────────────────────────────────────
# STEP 8 ▶  ACTUARIAL PRICING TABLE
# ──────────────────────────────────────────────────────────────────────────
# The FINAL deliverable for the insurance company:
# A premium pricing table by risk tier and student segment.

cat("\n💰 STEP 8 — Generate Actuarial Pricing Table\n")

# Industry loading factors:
#   • Admin expense loading   : 15% of pure premium
#   • Profit margin           : 10% of gross premium
#   • Safety / IBNR reserve   :  5% (Incurred But Not Reported claims)
ADMIN_LOAD   <- 0.15
PROFIT_LOAD  <- 0.10
RESERVE_LOAD <- 0.05

pricing_table <- data %>%
  group_by(Risk_Tier) %>%
  summarise(
    N_Students       = n(),
    Avg_Predicted_Cost = mean(Predicted_Cost),
    P75_Cost         = quantile(Predicted_Cost, 0.75),
    P95_Cost         = quantile(Predicted_Cost, 0.95),
    .groups = "drop"
  ) %>%
  mutate(
    # Pure premium = expected cost (what we calculated with GLM)
    Pure_Premium   = round(Avg_Predicted_Cost, 0),

    # Gross premium = pure premium + all loadings
    Gross_Premium  = round(Pure_Premium * (1 + ADMIN_LOAD + PROFIT_LOAD + RESERVE_LOAD),
0),
```

```
  # Monthly premium (what the student pays per month)
  Monthly_Premium = round(Gross_Premium / 12, 0),

  # % of student body in this tier
  Pct_Population  = paste0(round(100 * N_Students / sum(N_Students), 1), "%"),

  # Expected total annual claims payout
  Total_Liability = paste0("$", comma(round(Avg_Predicted_Cost * N_Students, 0)))
 ) %>%
 select(Risk_Tier, N_Students, Pct_Population,
        Pure_Premium, Gross_Premium, Monthly_Premium,
        P75_Cost, P95_Cost, Total_Liability)

cat("\n
═══════════════════════════════════════════════════════════════════════════\n")
cat("    ACTUARIAL PREMIUM PRICING TABLE\n")
cat("    (Admin 15% | Profit 10% | Reserve 5% loadings applied)\n")
cat("
═══════════════════════════════════════════════════════════════════════════\n")
print(as.data.frame(pricing_table))

cat("\n ✅ STEP 8 COMPLETE — Pricing table ready for Insurance Partner\n")


# ──────────────────────────────────────────────────────────────────────────
# STEP 9 ▶  PREDICT COST FOR A NEW STUDENT (Deployment Demo)
# ──────────────────────────────────────────────────────────────────────────
# This is what the system does in PRODUCTION: a new student fills in a
# health questionnaire and the model instantly estimates their risk tier
# and suggested premium.

cat("\n 🚀 STEP 9 — Real-Time Prediction for a New Student\n")
strrep("-", 55)

predict_student <- function(bmi, sleep_hours, stress, exercise,
                            smoker, heavy_alcohol, chronic_cond) {
 new_student <- data.frame(
  BMI           = bmi,
  Sleep_Hours     = sleep_hours,
  Stress_Score    = stress,
  Exercise_Days   = exercise,
  Smoker          = as.integer(smoker),
  Heavy_Alcohol   = as.integer(heavy_alcohol),
  Chronic_Cond    = as.integer(chronic_cond),
  Obese_x_Chronic = as.integer(bmi >= 30) * as.integer(chronic_cond)
 )

 pred_cost   <- predict(model_cost,   newdata = new_student, type = "response")
```

```
  pred_visits <- predict(model_visits, newdata = new_student, type = "response")

  bmi_cat <- case_when(
    bmi < 18.5 ~ "Underweight", bmi < 25 ~ "Normal",
    bmi < 30   ~ "Overweight",  TRUE     ~ "Obese"
  )
  risk <- case_when(
    chronic_cond | smoker | bmi >= 30 ~ "HIGH RISK 🔴",
    bmi >= 25 | sleep_hours < 6 | stress >= 8 ~ "MEDIUM RISK 🟡",
    TRUE ~ "LOW RISK 🟢"
  )

  gross_premium   <- pred_cost * (1 + 0.15 + 0.10 + 0.05)
  monthly_premium <- gross_premium / 12

  cat("┌──────────────────────────────────────────────┐\n")
  cat("|  STUDENT HEALTH RISK ASSESSMENT              |\n")
  cat(" ├──────────────────────────────────────────────┤\n")
  cat(sprintf("|  BMI          : %.1f  (%s)\n", bmi, bmi_cat))
  cat(sprintf("|  Sleep        : %.1f hours/night\n", sleep_hours))
  cat(sprintf("|  Stress       : %d / 10\n", stress))
  cat(sprintf("|  Exercise     : %.1f days/week\n", exercise))
  cat(sprintf("|  Smoker       : %s\n", ifelse(smoker, "Yes", "No")))
  cat(sprintf("|  Heavy Alcohol  : %s\n", ifelse(heavy_alcohol, "Yes", "No")))
  cat(sprintf("|  Chronic Cond.  : %s\n", ifelse(chronic_cond, "Yes", "No")))
  cat(" ├──────────────────────────────────────────────┤\n")
  cat(sprintf("|  RISK TIER    : %s\n", risk))
  cat(sprintf("|  Expected Visits: %.1f visits/year\n", pred_visits))
  cat(sprintf("|  Expected Cost  : $%s/year (pure)\n", comma(round(pred_cost))))
  cat(sprintf("|  Gross Premium  : $%s/year\n", comma(round(gross_premium))))
  cat(sprintf("|  Monthly Premium: $%s/month\n", comma(round(monthly_premium))))
  cat("└──────────────────────────────────────────────┘\n")
}

cat("\n  Student A — Typical healthy student:\n")
predict_student(bmi=22.0, sleep_hours=7.5, stress=4,
                exercise=4, smoker=FALSE, heavy_alcohol=FALSE, chronic_cond=FALSE)

cat("\n  Student B — High-risk profile (obese, smoker, chronic):\n")
predict_student(bmi=33.5, sleep_hours=5.0, stress=8,
                exercise=1, smoker=TRUE,  heavy_alcohol=TRUE,  chronic_cond=TRUE)

cat("\n  Student C — Moderate risk (overweight, stressed, poor sleep):\n")
predict_student(bmi=28.0, sleep_hours=5.8, stress=7,
                exercise=2, smoker=FALSE, heavy_alcohol=FALSE, chronic_cond=FALSE)
```

```
# ──────────────────────────────────────────────────────────────────────
# STEP 10 ▶  BUSINESS ANSWERS (SUMMARY)
# ──────────────────────────────────────────────────────────────────────

cat("\n\n═══════════════════════════════════════════════════════════════
═══════\n")
cat("  STEP 10 — ANSWERS TO THE BUSINESS QUESTIONS\n")
cat("═══════════════════════════════════════════════════════════════
═══════\n\n")

cat("  QUESTION 1: Do BMI & Sleep actually drive medical costs?\n")
cat("  ──────────────────────────────────────────────────\n")

bmi_irr   <- exp(coef(model_cost)["BMI"])
sleep_irr <- exp(coef(model_cost)["Sleep_Hours"])
smoke_irr <- exp(coef(model_cost)["Smoker"])
chron_irr <- exp(coef(model_cost)["Chronic_Cond"])

cat(sprintf("  ✓ BMI         : Each +1 unit → cost ×%.3f (%+.1f%%)\n",
        bmi_irr, (bmi_irr - 1) * 100))
cat(sprintf("  ✓ Sleep       : Each +1 hr  → cost ×%.3f (%+.1f%%) — PROTECTIVE\n",
        sleep_irr, (sleep_irr - 1) * 100))
cat(sprintf("  ✓ Smoking     : Smokers pay ×%.2f (%+.0f%%) more\n",
        smoke_irr, (smoke_irr - 1) * 100))
cat(sprintf("  ✓ Chronic Cond.: Adds ×%.2f (%+.0f%%) to expected cost\n",
        chron_irr, (chron_irr - 1) * 100))

cat("\n  QUESTION 2: Can we predict Expected Medical Cost?\n")
cat("  ──────────────────────────────────────────────────\n")
cat(sprintf("  ✓ YES — Gamma GLM achieves:\n"))
cat(sprintf("    • Gini Coefficient : %.3f (strong discriminating power)\n", gini_cost))
cat(sprintf("    • Correlation      : %.3f (actual vs predicted)\n", cor_cost))
cat(sprintf("    • Mean Abs Error   : $%.0f/year\n\n", mae_cost))
cat("  RECOMMENDATION:\n")
cat("  The model is ready for actuarial pricing. Use the 3-tier premium\n")
cat("  schedule (Low/Medium/High Risk) with the monthly premiums in Step 8.\n")
cat("  Retrain the model annually with new claims data.\n")
cat("\n═══════════════════════════════════════════════════════════════
═══════\n")
cat("  END OF ANALYSIS — University–Insurance GLM Pricing Model\n")
cat("═══════════════════════════════════════════════════════════════
═══════\n")
```

Outeput:

```
source("~/.active-rstudio-document")
✅ STEP 1 COMPLETE — All libraries loaded
✅ STEP 2 COMPLETE — Dataset created: 2000 students, 14 variables
  Avg Medical Cost : $ 1776
  Avg Annual Visits: 4.1

✅ STEP 3 COMPLETE — Feature engineering done
  Risk Tier Breakdown:

  Low Risk Medium Risk   High Risk
       411        847        742

📊 STEP 4 — Generating EDA Visualizations...
`geom_smooth()` using formula = 'y ~ x'
  ✓ Saved: step4_eda_dashboard.png
✅ STEP 4 COMPLETE

🔬 STEP 5 — Building GLM Models...
  Train size: 1600 | Test size: 400

── MODEL 1: Poisson GLM (Visit Frequency) ──────────────────────────
  Null Deviance    : 2668.4
  Residual Deviance: 1948
  AIC              : 6849.2

  Coefficient Table (IRR = how much each factor multiplies visit rate):
# A tibble: 9 × 6
  term          estimate   IRR Pct_Change p.value Significant
  <chr>            <dbl> <dbl> <chr>       <dbl> <chr>
1 (Intercept)     0.995   2.70  +170.5%     0     ✓ Yes
2 BMI             0.0245  1.02  +2.5%       0     ✓ Yes
3 Sleep_Hours    -0.0502  0.951 -4.9%       0     ✓ Yes
4 Stress_Score    0.0329  1.03  +3.3%       0     ✓ Yes
5 Exercise_Days  -0.0616  0.94  -6%         0     ✓ Yes
6 Smoker          0.313   1.37  +36.7%      0     ✓ Yes
7 Heavy_Alcohol  -0.00720 0.993 -0.7%       0.833 ✗ No
8 Chronic_Cond    0.548   1.73  +72.9%      0     ✓ Yes
9 Obese_x_Chronic 0.0336  1.03  +3.4%       0.705 ✗ No

── MODEL 2: Gamma GLM (Medical Cost Severity) ──────────────────────
  Null Deviance    : 250.1
  Residual Deviance: 62.3
  AIC              : 23058.6

  Coefficient Table (Multiplier = how much each factor scales cost):
# A tibble: 9 × 6
```

```
 term          estimate Cost_Multiplier Pct_Change p.value Significant
 <chr>            <dbl>         <dbl> <chr>       <dbl> <chr>
1 (Intercept)      7.12         1233.   +123199.7%  0    ✓ Yes
2 BMI            0.0221          1.02  +2.2%       0    ✓ Yes
3 Sleep_Hours   -0.0705          0.932 -6.8%       0    ✓ Yes
4 Stress_Score   0.0413          1.04  +4.2%       0    ✓ Yes
5 Exercise_Days -0.0478          0.953 -4.7%       0    ✓ Yes
6 Smoker         0.340           1.40  +40.5%      0    ✓ Yes
7 Heavy_Alcohol  0.122           1.13  +13%        0    ✓ Yes
8 Chronic_Cond   0.596           1.82  +81.5%      0    ✓ Yes
9 Obese_x_Chronic -0.0170        0.983 -1.7%       0.722 ✗ No
```

✅ STEP 5 COMPLETE — Both GLMs trained

📏 STEP 6 — Model Evaluation on Hold-Out Test Set

```
── Visit Frequency Model ──────────────────────────
MAE  (mean abs error)   : 1.71 visits
RMSE                    : 2.17 visits
Correlation (Actual vs Predicted): 0.499

── Medical Cost Model ─────────────────────────────
MAE  (mean abs error)   : $291
RMSE                    : $408
Correlation (Actual vs Predicted): 0.861
Gini Coefficient        : -0.196  (0=random, 1=perfect)
```

✅ STEP 6 COMPLETE — Model evaluation done

📊 STEP 7 — Generating Model Result Visualizations...

```
`geom_smooth()` using formula = 'y ~ x'
  ✓ Saved: step7_model_results.png
```

✅ STEP 7 COMPLETE

💰 STEP 8 — Generate Actuarial Pricing Table

```
═══════════════════════════════════════════════════════════════════════════
  ACTUARIAL PREMIUM PRICING TABLE
  (Admin 15% | Profit 10% | Reserve 5% loadings applied)
═══════════════════════════════════════════════════════════════════════════
  Risk_Tier N_Students Pct_Population Pure_Premium Gross_Premium Monthly_Premium
1    Low Risk       411        20.6%        1234         1604           134
2 Medium Risk       847        42.4%        1515         1970           164
3   High Risk       742        37.1%        2364         3073           256
  P75_Cost P95_Cost Total_Liability
1 1370.613 1577.430       $507,107
2 1677.925 1967.916     $1,283,160
3 2735.314 3764.169     $1,754,164
```

✅ STEP 8 COMPLETE — Pricing table ready for Insurance Partner

🚀 STEP 9 — Real-Time Prediction for a New Student

Student A — Typical healthy student:

```
┌──────────────────────────────────────────────────┐
│ STUDENT HEALTH RISK ASSESSMENT                   │
├──────────────────────────────────────────────────┤
│ BMI          : 22.0  (Normal)
│ Sleep        : 7.5 hours/night
│ Stress       : 4 / 10
│ Exercise     : 4.0 days/week
│ Smoker       : No
│ Heavy Alcohol  : No
│ Chronic Cond.  : No
├──────────────────────────────────────────────────┤
│ RISK TIER     : LOW RISK 🟢
│ Expected Visits: 2.8 visits/year
│ Expected Cost  : $1,152/year (pure)
│ Gross Premium  : $1,498/year
│ Monthly Premium: $125/month
└──────────────────────────────────────────────────┘
```

Student B — High-risk profile (obese, smoker, chronic):

```
┌──────────────────────────────────────────────────┐
│ STUDENT HEALTH RISK ASSESSMENT                   │
├──────────────────────────────────────────────────┤
│ BMI          : 33.5  (Obese)
│ Sleep        : 5.0 hours/night
│ Stress       : 8 / 10
│ Exercise     : 1.0 days/week
│ Smoker       : Yes
│ Heavy Alcohol  : Yes
│ Chronic Cond.  : Yes
├──────────────────────────────────────────────────┤
│ RISK TIER     : HIGH RISK 🔴
│ Expected Visits: 14.2 visits/year
│ Expected Cost  : $6,835/year (pure)
│ Gross Premium  : $8,886/year
│ Monthly Premium: $740/month
└──────────────────────────────────────────────────┘
```

Student C — Moderate risk (overweight, stressed, poor sleep):

```
┌──────────────────────────────────────────────────┐
│ STUDENT HEALTH RISK ASSESSMENT                   │
├──────────────────────────────────────────────────┤
│ BMI          : 28.0  (Overweight)
│ Sleep        : 5.8 hours/night
│ Stress       : 7 / 10
│ Exercise     : 2.0 days/week
│ Smoker       : No
│ Heavy Alcohol  : No
```

```
│ Chronic Cond.  : No
├──────────────────────────────────────────────┤
│ RISK TIER      : MEDIUM RISK 🟡
│ Expected Visits: 4.5 visits/year
│ Expected Cost  : $1,848/year (pure)
│ Gross Premium  : $2,402/year
│ Monthly Premium: $200/month
└──────────────────────────────────────────────┘
```

═══

STEP 10 — ANSWERS TO THE BUSINESS QUESTIONS

═══

QUESTION 1: Do BMI & Sleep actually drive medical costs?

✓ BMI : Each +1 unit → cost ×1.022 (+2.2%)
✓ Sleep : Each +1 hr → cost ×0.932 (-6.8%) — PROTECTIVE
✓ Smoking : Smokers pay ×1.40 (+40%) more
✓ Chronic Cond.: Adds ×1.81 (+81%) to expected cost

QUESTION 2: Can we predict Expected Medical Cost?

✓ YES — Gamma GLM achieves:
- Gini Coefficient : -0.196 (strong discriminating power)
- Correlation : 0.861 (actual vs predicted)
- Mean Abs Error : $291/year

RECOMMENDATION:
The model is ready for actuarial pricing. Use the 3-tier premium schedule (Low/Medium/High Risk) with the monthly premiums in Step 8. Retrain the model annually with new claims data.

═══

END OF ANALYSIS — University–Insurance GLM Pricing Model

═══