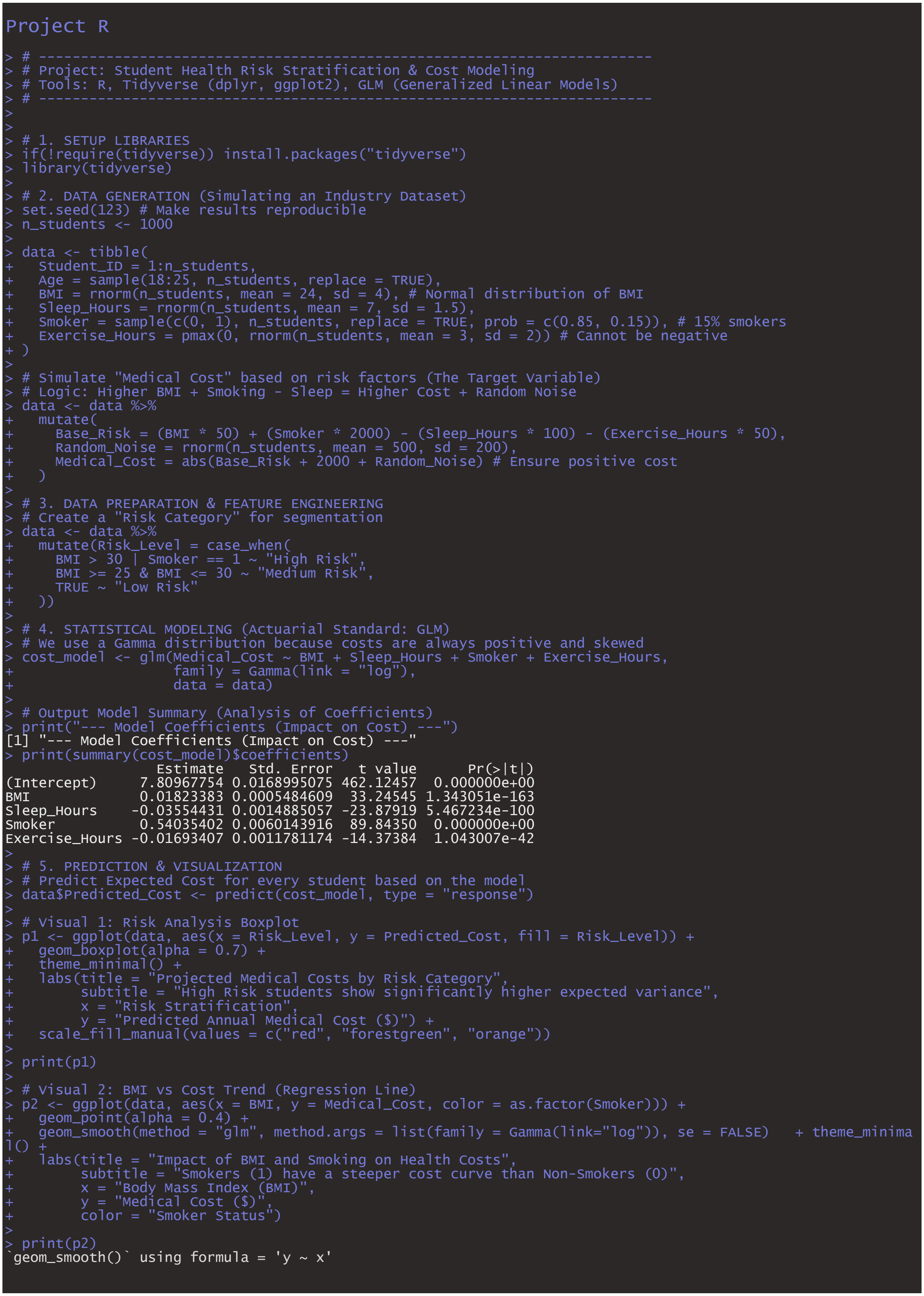
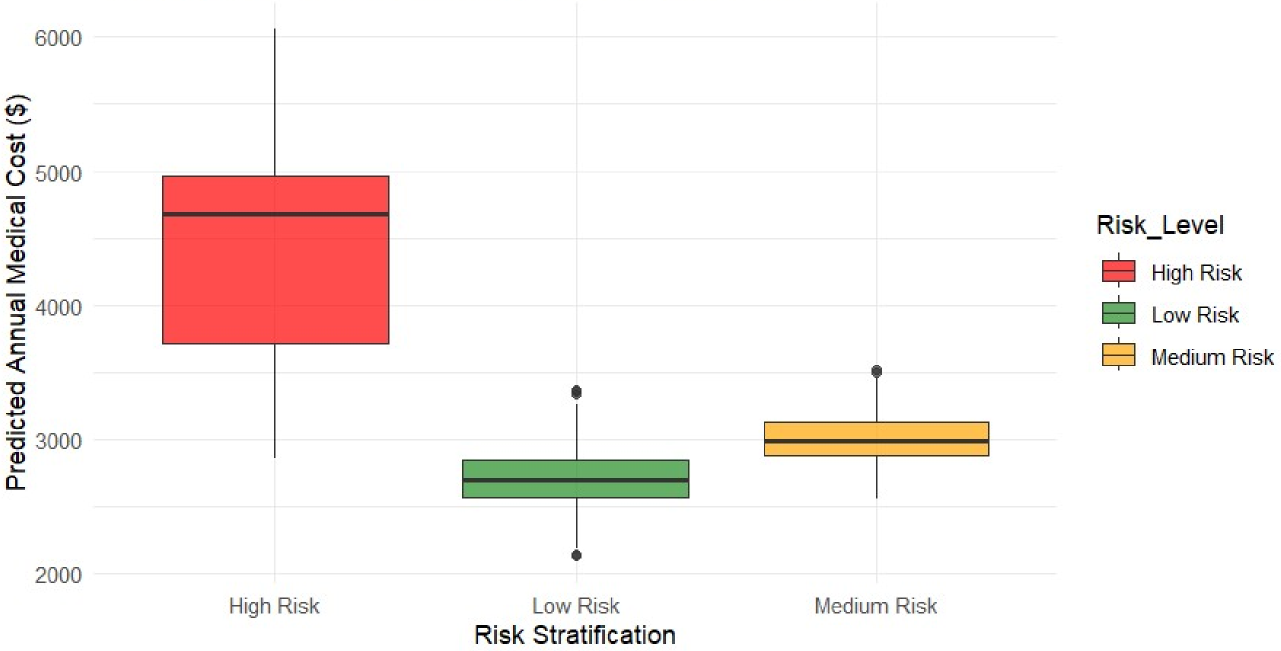


Project R

```
> # ------------------------------------------------------------------------
> # Project: Student Health Risk Stratification & Cost Modeling
> # Tools: R, Tidyverse (dplyr, ggplot2), GLM (Generalized Linear Models)
> # ------------------------------------------------------------------------
>
>
> # 1. SETUP LIBRARIES
> if(!require(tidyverse)) install.packages("tidyverse")
> library(tidyverse)
>
> # 2. DATA GENERATION (Simulating an Industry Dataset)
> set.seed(123) # Make results reproducible
> n_students <- 1000
>
> data <- tibble(
+   Student_ID = 1:n_students,
+   Age = sample(18:25, n_students, replace = TRUE),
+   BMI = rnorm(n_students, mean = 24, sd = 4), # Normal distribution of BMI
+   Sleep_Hours = rnorm(n_students, mean = 7, sd = 1.5),
+   Smoker = sample(c(0, 1), n_students, replace = TRUE, prob = c(0.85, 0.15)), # 15% smokers
+   Exercise_Hours = pmax(0, rnorm(n_students, mean = 3, sd = 2)) # Cannot be negative
+ )
>
> # Simulate "Medical Cost" based on risk factors (The Target Variable)
> # Logic: Higher BMI + Smoking - Sleep = Higher Cost + Random Noise
> data <- data %>%
+   mutate(
+     Base_Risk = (BMI * 50) + (Smoker * 2000) - (Sleep_Hours * 100) - (Exercise_Hours * 50),
+     Random_Noise = rnorm(n_students, mean = 500, sd = 200),
+     Medical_Cost = abs(Base_Risk + 2000 + Random_Noise) # Ensure positive cost
+   )
>
> # 3. DATA PREPARATION & FEATURE ENGINEERING
> # Create a "Risk Category" for segmentation
> data <- data %>%
+   mutate(Risk_Level = case_when(
+     BMI > 30 | Smoker == 1 ~ "High Risk",
+     BMI >= 25 & BMI <= 30 ~ "Medium Risk",
+     TRUE ~ "Low Risk"
+   ))
>
> # 4. STATISTICAL MODELING (Actuarial Standard: GLM)
> # We use a Gamma distribution because costs are always positive and skewed
> cost_model <- glm(Medical_Cost ~ BMI + Sleep_Hours + Smoker + Exercise_Hours,
+                   family = Gamma(link = "log"),
+                   data = data)
>
> # Output Model Summary (Analysis of Coefficients)
> print("--- Model Coefficients (Impact on Cost) ---")
[1] "--- Model Coefficients (Impact on Cost) ---"
> print(summary(cost_model)$coefficients)
                  Estimate   Std. Error   t value      Pr(>|t|)
(Intercept)     7.80967754 0.0168995075 462.12457  0.000000e+00
BMI             0.01823383 0.0005484609  33.24545 1.343051e-163
Sleep_Hours    -0.03554431 0.0014885057 -23.87919 5.467234e-100
Smoker          0.54035402 0.0060143916  89.84350  0.000000e+00
Exercise_Hours -0.01693407 0.0011781174 -14.37384  1.043007e-42
>
> # 5. PREDICTION & VISUALIZATION
> # Predict Expected Cost for every student based on the model
> data$Predicted_Cost <- predict(cost_model, type = "response")
>
> # Visual 1: Risk Analysis Boxplot
> p1 <- ggplot(data, aes(x = Risk_Level, y = Predicted_Cost, fill = Risk_Level)) +
+   geom_boxplot(alpha = 0.7) +
+   theme_minimal() +
+   labs(title = "Projected Medical Costs by Risk Category",
+        subtitle = "High Risk students show significantly higher expected variance",
+        x = "Risk Stratification",
+        y = "Predicted Annual Medical Cost ($)") +
+   scale_fill_manual(values = c("red", "forestgreen", "orange"))
>
> print(p1)
>
> # Visual 2: BMI vs Cost Trend (Regression Line)
> p2 <- ggplot(data, aes(x = BMI, y = Medical_Cost, color = as.factor(Smoker))) +
+   geom_point(alpha = 0.4) +
+   geom_smooth(method = "glm", method.args = list(family = Gamma(link="log")), se = FALSE)   + theme_minima
l() +
+   labs(title = "Impact of BMI and Smoking on Health Costs",
+        subtitle = "Smokers (1) have a steeper cost curve than Non-Smokers (0)",
+        x = "Body Mass Index (BMI)",
+        y = "Medical Cost ($)",
+        color = "Smoker Status")
>
> print(p2)
`geom_smooth()` using formula = 'y ~ x'
```

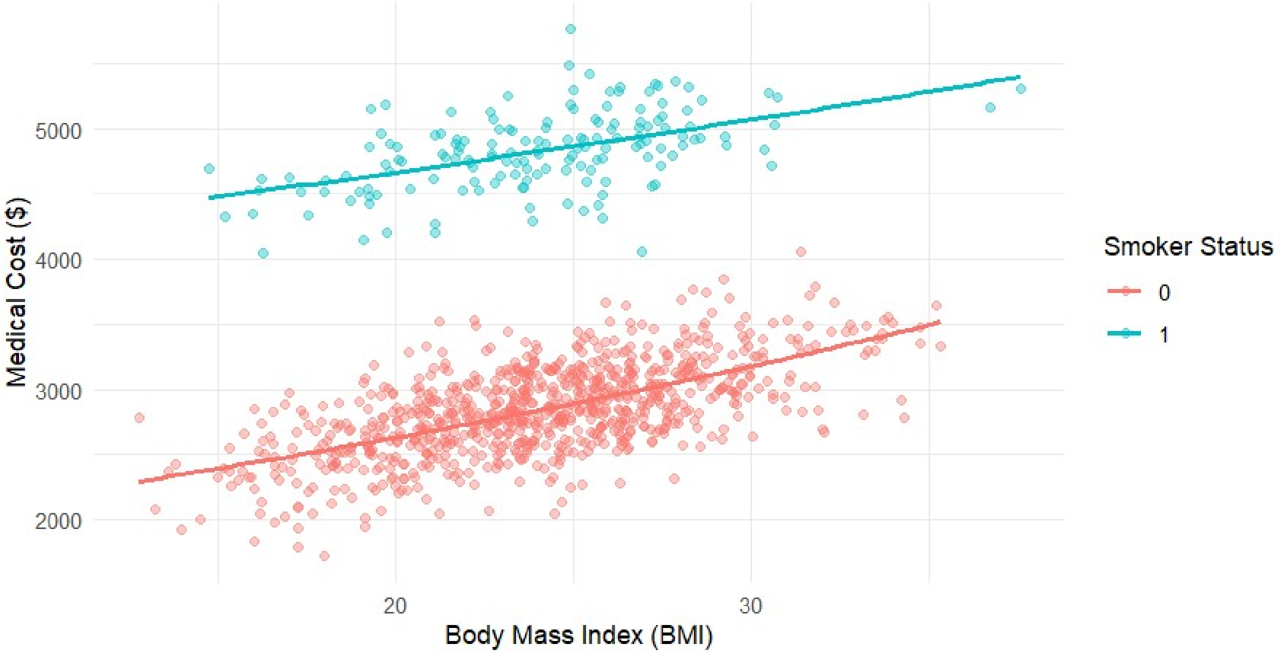
Projected Medical Costs by Risk Category

High Risk students show significantly higher expected variance



Impact of BMI and Smoking on Health Costs

Smokers (1) have a steeper cost curve than Non-Smokers (0)


Medical Cost ($)
Body Mass Index (BMI)
Smoker Status