g part

2025-08-18

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(tidyr)
library(caret)
```

```
## Loading required package: lattice
```

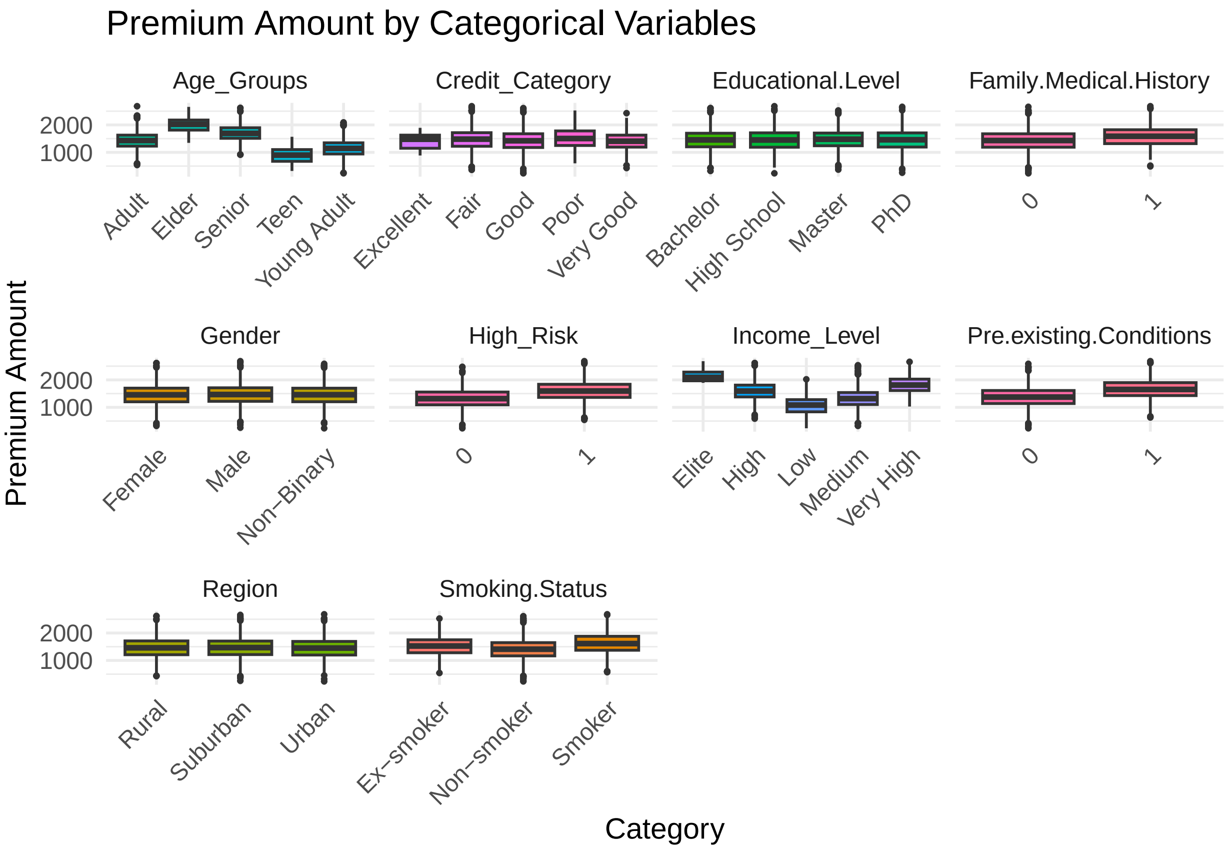
```
library(randomForest)
```

```
## randomForest 4.7-1.2

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine
```

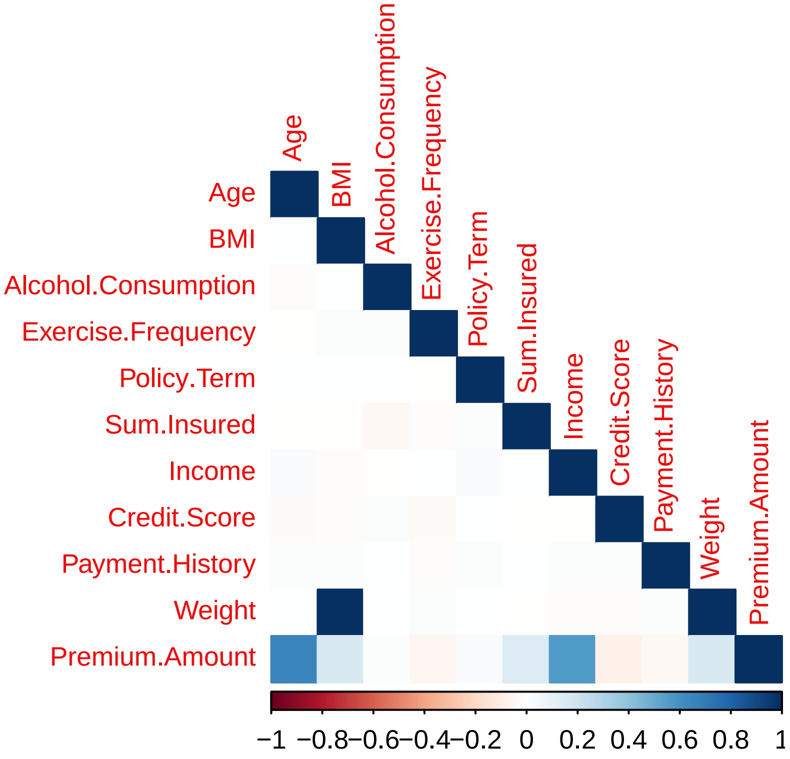
```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
rm(list = ls())

data <- read.csv("~/Desktop/MFI /Prep Project/premium_data.csv", header = TRUE)

data$Income_Level <- trimws(as.character(data$Income_Level))

data$Income_Level[grepl("^\\s*$", data$Income_Level)] <- NA

data$Income_Level <- factor(data$Income_Level)

data_clean <- na.omit(data)

data_clean$Smoking.Status <- as.factor(trimws(data_clean$Smoking.Status))

data_clean$Gender <- as.factor(data_clean$Gender)
data_clean$Region <- as.factor(data_clean$Region)
data_clean$Educational.Level <- as.factor(data_clean$Educational.Level)
data_clean$Age_Groups <- as.factor(data_clean$Age_Groups)
data_clean$Income_Level <- as.factor(data_clean$Income_Level)
data_clean$Credit_Category <- as.factor(data_clean$Credit_Category)
data_clean$Pre.existing.Conditions <- as.factor(data_clean$Pre.existing.Conditions)
data_clean$Family.Medical.History <- as.factor(data_clean$Family.Medical.History)
data_clean$High_Risk <- as.factor(data_clean$High_Risk)


cat_vars <- c("Smoking.Status", "Gender", "Region", "Educational.Level",
              "Age_Groups", "Income_Level", "Credit_Category",
              "Pre.existing.Conditions", "Family.Medical.History", "High_Risk")


cat_data <- data_clean[, c("Premium.Amount", cat_vars)]


cat_long <- cat_data %>%
  pivot_longer(cols = all_of(cat_vars),
               names_to = "Variable",
               values_to = "Category")
ggplot(cat_long, aes(x = Category, y = Premium.Amount, fill = Category)) +
  geom_boxplot(outlier.size = 0.5) +
  facet_wrap(~Variable, scales = "free_x") +
  labs(title = "Premium Amount by Categorical Variables",
       x = "Category", y = "Premium Amount") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")
```



```r
num_vars <- data_clean %>% select_if(is.numeric)

corrplot(cor(num_vars), method = "color", type = "lower", tl.cex = 0.8)
```



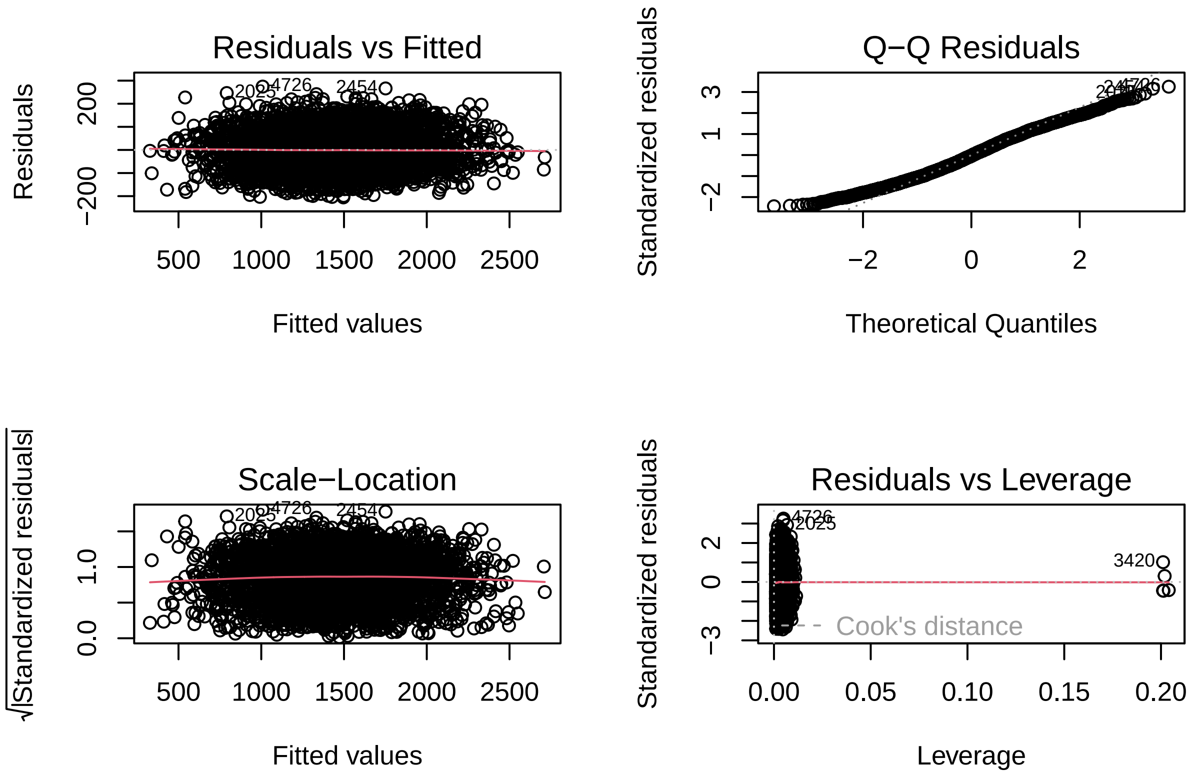
```r
model_data <- data_clean %>%
  select(Premium.Amount, Age, BMI, Credit.Score, Sum.Insured, High_Risk,
         Smoking.Status, Pre.existing.Conditions, Family.Medical.History, Income_Level)

model_data$Smoking.Status <- as.factor(model_data$Smoking.Status)
model_data$Pre.existing.Conditions <- as.factor(model_data$Pre.existing.Conditions)
model_data$Family.Medical.History <- as.factor(model_data$Family.Medical.History)
model_data$Income_Level <- as.factor(model_data$Income_Level)

str(model_data)
```

```
## 'data.frame':    5322 obs. of  10 variables:
##  $ Premium.Amount         : num  1221 1723 917 2019 1110 ...
##  $ Age                    : int  50 43 52 63 42 42 63 54 39 51 ...
##  $ BMI                    : num  23.5 30.5 22.5 28.1 30.6 ...
##  $ Credit.Score           : num  643 550 669 590 559 ...
##  $ Sum.Insured            : num  155676 185331 239240 170028 232677 ...
##  $ High_Risk              : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 2 1 1 ...
##  $ Smoking.Status         : Factor w/ 3 levels "Ex-smoker","Non-smoker",..: 2 3 2 3 2 2 1 2 2 2 ...
##  $ Pre.existing.Conditions: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 1 1 ...
##  $ Family.Medical.History : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 1 1 1 1 ...
##  $ Income_Level           : Factor w/ 5 levels "Elite","High",..: 4 2 3 4 4 2 3 2 4 4 ...
##  - attr(*, "na.action")= 'omit' Named int [1:6] 501 1476 2044 2711 3239 4079
##   ..- attr(*, "names")= chr [1:6] "501" "1476" "2044" "2711" ...
```

```r
set.seed(888)

train_index <- createDataPartition(model_data$Premium.Amount, p = 0.7, list = FALSE)
train_data <- model_data[train_index, ]
test_data <- model_data[-train_index, ]

model_lm <- lm(Premium.Amount ~ ., data = train_data)
summary(model_lm)
```

```
##
## Call:
## lm(formula = Premium.Amount ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -205.386  -65.072   -3.023   64.401  274.197
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8.410e+02  4.447e+01  18.910  < 2e-16 ***
## Age                       2.022e+01  1.160e-01 174.226  < 2e-16 ***
## BMI                       1.513e+01  3.491e-01  43.344  < 2e-16 ***
## Credit.Score             -4.486e-01  2.796e-02 -16.044  < 2e-16 ***
## Sum.Insured               1.016e-03  2.809e-05  36.164  < 2e-16 ***
## High_Risk1               -5.808e-01  6.703e+00  -0.087    0.931
## Smoking.StatusNon-smoker -1.085e+02  6.725e+00 -16.136  < 2e-16 ***
## Smoking.StatusSmoker      8.859e+01  5.395e+00  16.420  < 2e-16 ***
## Pre.existing.Conditions1  2.989e+02  5.658e+00  52.823  < 2e-16 ***
## Family.Medical.History1   1.549e+02  3.519e+00  44.030  < 2e-16 ***
## Income_LevelHigh         -5.098e+02  3.787e+01 -13.459  < 2e-16 ***
## Income_LevelLow          -1.015e+03  3.818e+01 -26.593  < 2e-16 ***
## Income_LevelMedium       -7.613e+02  3.788e+01 -20.098  < 2e-16 ***
## Income_LevelVery High    -2.486e+02  3.816e+01  -6.514 8.31e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.5 on 3712 degrees of freedom
## Multiple R-squared:  0.9475, Adjusted R-squared:  0.9474
## F-statistic:  5158 on 13 and 3712 DF,  p-value: < 2.2e-16
```

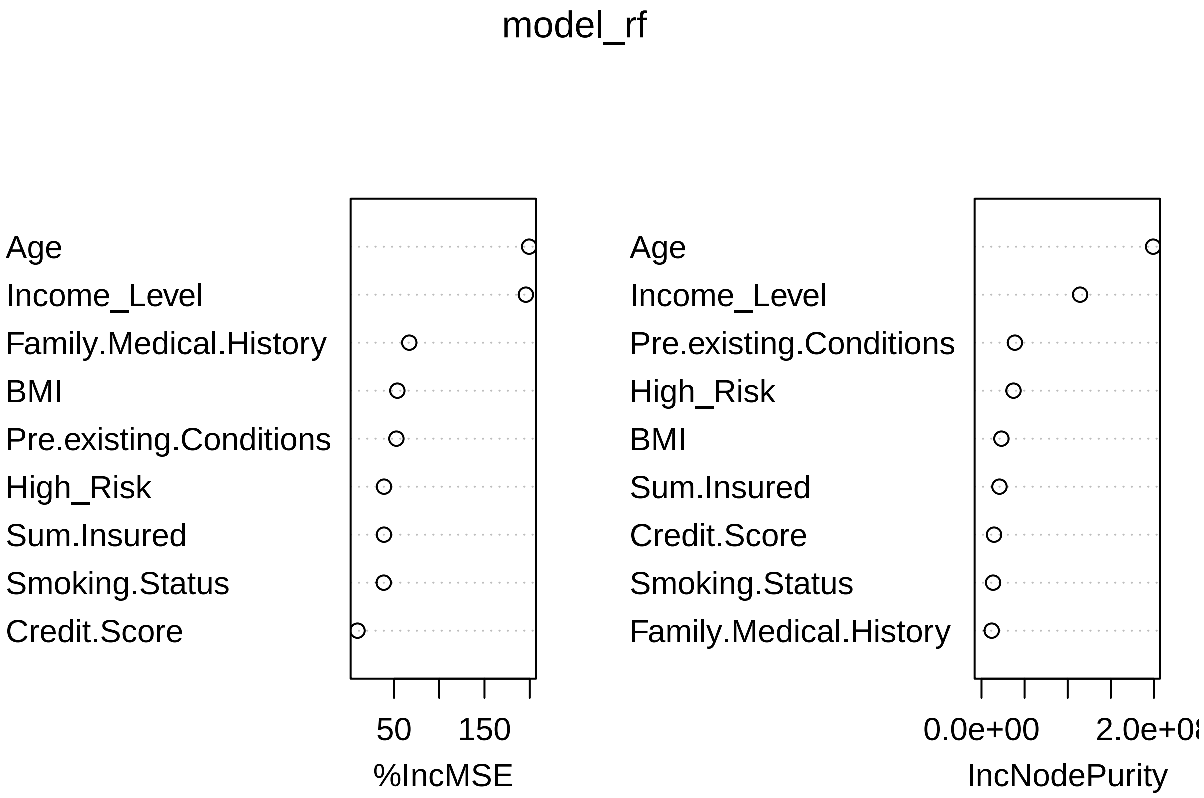
```
par(mfrow = c(2, 2))

plot(model_lm)
```



```
vif(model_lm)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## Age                     1.005467  1        1.002730
## BMI                     1.004571  1        1.002283
## Credit.Score            1.002917  1        1.001458
## Sum.Insured             1.002629  1        1.001314
## High_Risk               5.860329  1        2.420812
## Smoking.Status          3.454236  2        1.363289
## Pre.existing.Conditions 3.488755  1        1.867821
## Family.Medical.History  1.003074  1        1.001536
## Income_Level            1.010622  4        1.001322
```

```
model_rf <- randomForest(Premium.Amount ~ ., data = train_data, importance = TRUE)

varImpPlot(model_rf)
```



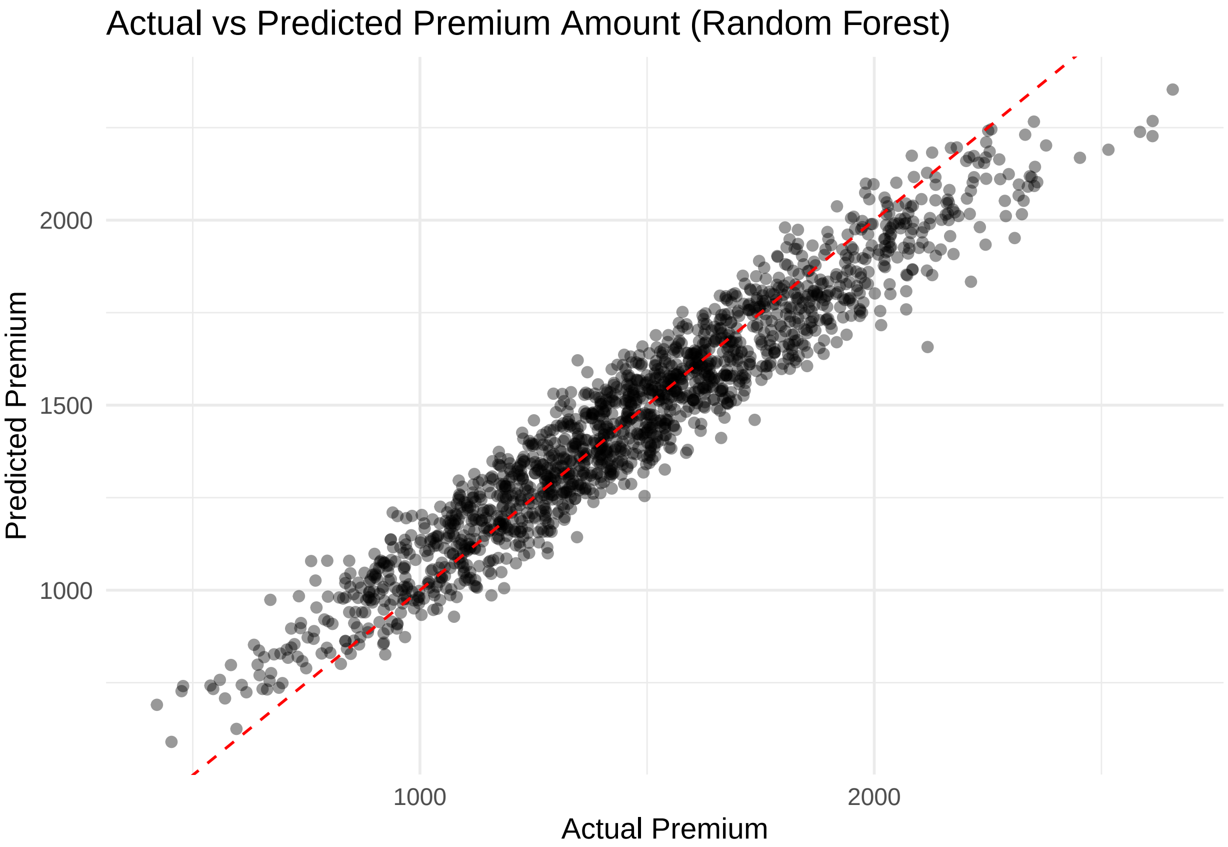
```r
pred_lm <- predict(model_lm, newdata = test_data)

rmse_lm <- sqrt(mean((pred_lm - test_data$Premium.Amount)^2))
mae_lm <- mean(abs(pred_lm - test_data$Premium.Amount))

cat("Linear Model RMSE:", rmse_lm, "\nMAE:", mae_lm, "\n")
```

```
## Linear Model RMSE: 85.28582
## MAE: 70.95451
```

```r
pred_rf <- predict(model_rf, newdata = test_data)

rmse_rf <- sqrt(mean((pred_rf - test_data$Premium.Amount)^2))
mae_rf <- mean(abs(pred_rf - test_data$Premium.Amount))

cat("Random Forest RMSE:", rmse_rf, "\nMAE:", mae_rf, "\n")
```

```
## Random Forest RMSE: 108.6852
## MAE: 87.63214
```

```r
df_compare <- data.frame(Actual = test_data$Premium.Amount, Predicted = pred_rf)

ggplot(df_compare, aes(x = Actual, y = Predicted)) +
  geom_point(alpha = 0.4) +
  geom_abline(slope = 1, intercept = 0, col = "red", linetype = "dashed") +
  labs(title = "Actual vs Predicted Premium Amount (Random Forest)",
       x = "Actual Premium", y = "Predicted Premium") +
  theme_minimal()
```



```r
results <- data.frame(
  Model = c("Linear Model", "Random Forest"),
  RMSE = c(rmse_lm, rmse_rf),
  MAE = c(mae_lm, mae_rf)
)

print(results)
```

```
##           Model      RMSE      MAE
## 1  Linear Model  85.28582 70.95451
## 2 Random Forest 108.68516 87.63214
```