HI 743 - Predictive Analytics in Healthcare RMarkdown Assignment Rubric

Topic: Linear_Regression_R
Name: Deepthi Doddaka
Dataset : Healthcare Dataset from Kaggle.

healthcare_dataset.csv

Section:1- Static Analysis Report

Introduction & Objective:

The aim of this analysis is to build a predictive model that estimates a patient's hospital billing amount based on various factors such as age, admission type, room number, insurance provider, and medical condition.

Data Understanding & Preparation:

1.Load the new healthcare dataset in R.

2.Perform data cleaning (handle missing values, remove duplicates, check variable types).

Model Implementation & Explanation

The target variable (Billing Amount) is continuous and numeric, which aligns well with linear regression.

Results & Interpretation:

Create visualizations such as histograms, scatterplots and calculate MSE values for both Training and Test Data.

Section 2: Follow-Up Assignment

Problem Definition & Justification:

Predicting Hospital Billing Amount Based on Patient and Admission Characteristics
Data Import, Cleaning, & Exploration

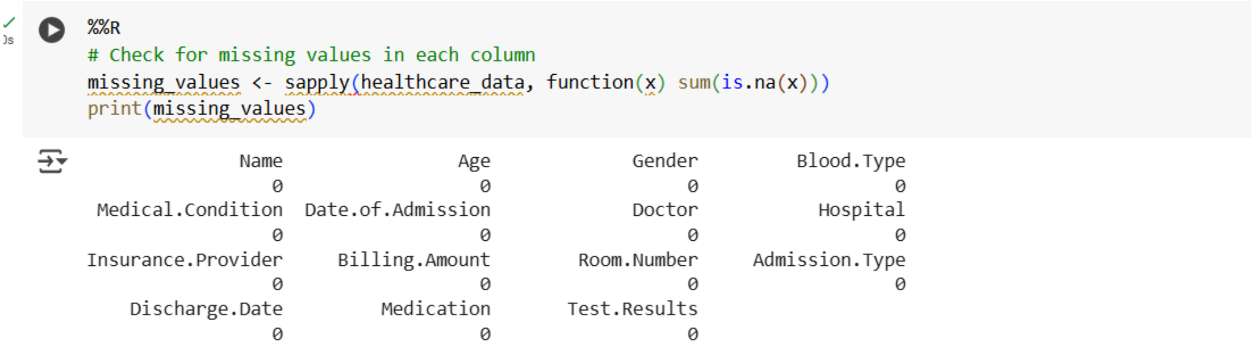
```
[48] from google.colab import files
     uploaded = files.upload()
```

Choose Files healthcare_dataset.csv

- **healthcare_dataset.csv**(text/csv) - 8399221 bytes, last modified: 3/13/2025 - 100% done

```
Saving healthcare_dataset.csv to healthcare_dataset (2).csv
```

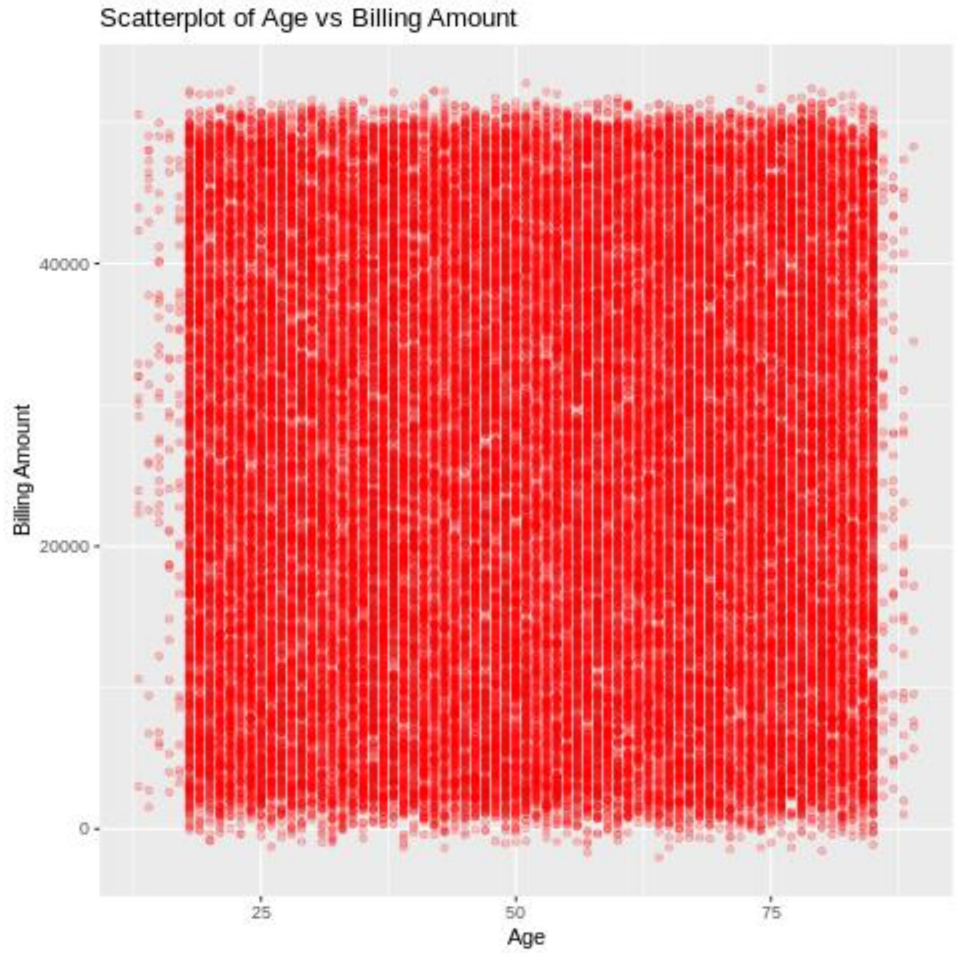
```
[49] %%R
     # Load dataset
     healthcare_data <- read.csv("/content/healthcare_dataset.csv")

     # View dataset structure
     str(healthcare_data)

     # Summary statistics
     summary(healthcare data)
```



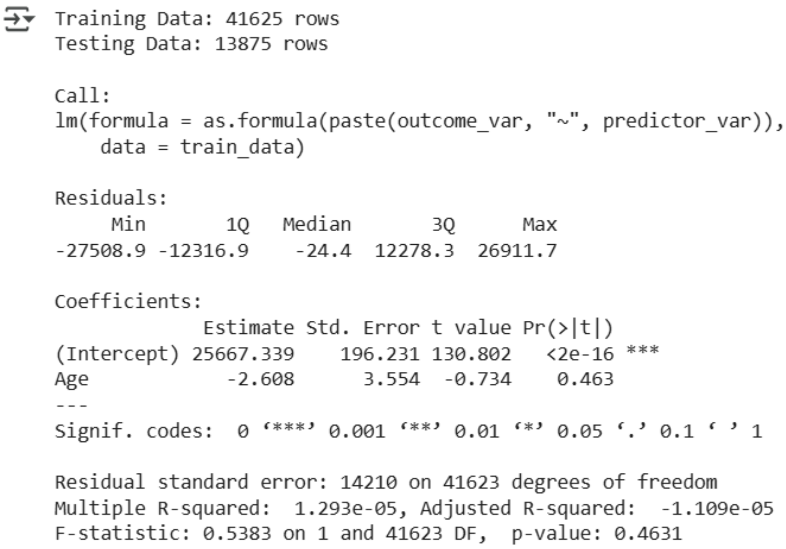
```
%%R
# Check for missing values in each column
missing_values <- sapply(healthcare_data, function(x) sum(is.na(x)))
print(missing_values)
```

```
              Name                Age             Gender         Blood.Type 
                 0                  0                  0                  0 
 Medical.Condition  Date.of.Admission             Doctor           Hospital 
                 0                  0                  0                  0 
Insurance.Provider     Billing.Amount        Room.Number     Admission.Type 
                 0                  0                  0                  0 
    Discharge.Date         Medication       Test.Results 
                 0                  0                  0 
```

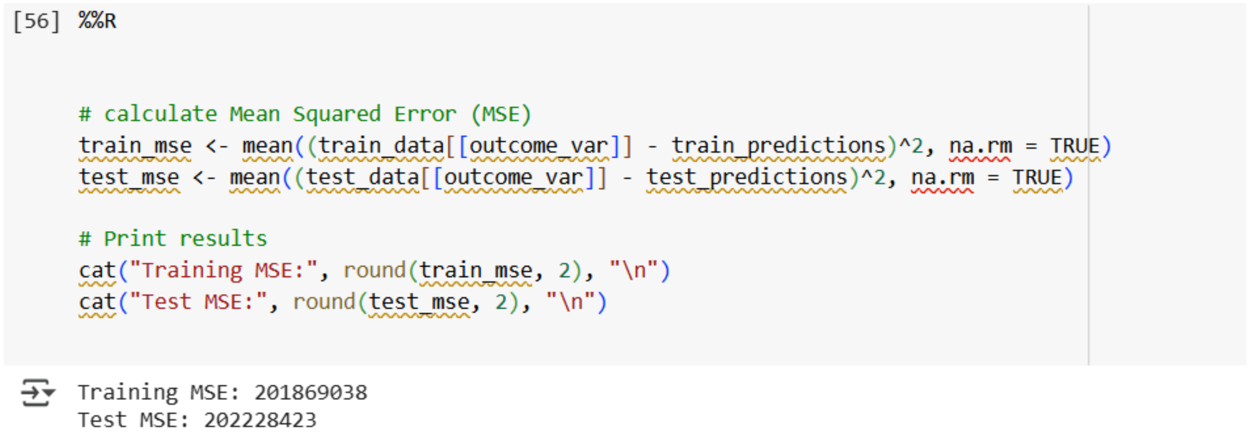
```
%%R
# Load required libraries
library(ggplot2)
library(dplyr)


outcome_var <- "Billing.Amount"
predictor_var <- "Age"

# Histogram of the outcome variable
ggplot(healthcare_data, aes_string(x = outcome_var)) +
  geom_histogram(fill = "steelblue", binwidth = 5000, color = "black") +
  labs(title = "Distribution of Billing Amount", x = "Billing Amount", y = "Count")

# Scatterplot between predictor and outcome
ggplot(healthcare_data, aes_string(x = predictor_var, y = outcome_var)) +
  geom_point(color = "red", alpha = 0.2) +
  labs(title = "Scatterplot of Age vs Billing Amount", x = "Age", y = "Billing Amount")
```



Results & Performance Evaluation:



```
Training Data: 41625 rows
Testing Data: 13875 rows

Call:
lm(formula = as.formula(paste(outcome_var, "~", predictor_var)),
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-27508.9 -12316.9    -24.4  12278.3  26911.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25667.339    196.231 130.802   <2e-16 ***
Age            -2.608      3.554  -0.734    0.463    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14210 on 41623 degrees of freedom
Multiple R-squared:  1.293e-05, Adjusted R-squared:  -1.109e-05 
F-statistic: 0.5383 on 1 and 41623 DF,  p-value: 0.4631
```



```
[56] %%R


    # calculate Mean Squared Error (MSE)
    train_mse <- mean((train_data[[outcome_var]] - train_predictions)^2, na.rm = TRUE)
    test_mse <- mean((test_data[[outcome_var]] - test_predictions)^2, na.rm = TRUE)

    # Print results
    cat("Training MSE:", round(train_mse, 2), "\n")
    cat("Test MSE:", round(test_mse, 2), "\n")
```

```
Training MSE: 201869038
Test MSE: 202228423
```

Conclusion:

After running the linear regression analysis, I found:

1.The relationship between Age and Billing Amount was weak, suggesting that Age alone is not a strong predictor.

2.Adding multiple predictors (Room Number, Insurance Provider, etc.) will improve model accuracy.

3. We applied linear regression for this dataset because:

Billing Amount is a Continuous Variable: Linear regression is ideal for predicting numerical values like medical costs.