# Load necessary libraries

library(readr) # For reading CSV files

library(ggplot2) # For creating plots

library(broom) # For tidying model output

# Step 1: Read the data

# Read the CSV file named "SimulationProject.csv"

# The file should have two columns: "calls" and "duration"

SimulationProject <- read\_csv("SimulationProject.csv")

# Print the first few rows of the data to verify it's loaded correctly

print(head(SimulationProject))

# Step 2: Explore the data

# Calculate summary statistics for both variables

summary(SimulationProject)

# Create a scatter plot to visualize the relationship between calls and duration

ggplot(SimulationProject, aes(x = calls, y = duration)) +

geom\_point() +

labs(title = "Scatter Plot of duration vs. calls",

x = "calls",

y = "duration") +

theme\_minimal()

# Step 3: Perform simple linear regression

# Fit a linear model where duration is the dependent variable and calls is the independent variable

model <- lm(duration ~ calls, data = SimulationProject)

# Step 4: Examine the model results

# Print a summary of the model, including coefficients, R-squared, and p-values

summary(model)

# Use broom to get a tidy version of the model coefficients

tidy\_model <- tidy(model)

print(tidy\_model)

# Step 5: Visualize the regression line

ggplot(SimulationProject, aes(x = calls, y = duration)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE, color = "red") +

labs(title = "Linear Regression: duration vs. calls",

x = "calls",

y = "duration") +

theme\_minimal()

# Step 6: Model diagnostics

# Create diagnostic plots to check assumptions

par(mfrow = c(2, 2)) # Set up a 2x2 plot layout

plot(model)

# Step 7: Predictions

# Create a data frame with new calls values for prediction

new\_calls <- data.frame(calls = seq(min(SimulationProject$calls), max(SimulationProject$calls), length.out = 100))

# Make predictions using the model

predictions <- predict(model, newdata = new\_calls, interval = "confidence")

# Combine predictions with new\_calls

prediction\_data <- cbind(new\_calls, predictions)

# Plot the original data, regression line, and confidence interval

ggplot() +

geom\_point(data = SimulationProject, aes(x = calls, y = duration)) +

geom\_line(data = prediction\_data, aes(x = calls, y = fit), color = "blue") +

geom\_ribbon(data = prediction\_data, aes(x = calls, ymin = lwr, ymax = upr), alpha = 0.2) +

labs(title = "Linear Regression with Confidence Interval",

x = "calls",

y = "duration") +

theme\_minimal()

# Print the R-squared value

cat("R-squared:", summary(model)$r.squared, "\n")

# Print the equation of the regression line

cat("Regression equation: duration =",

round(coef(model)[1], 3), "+",

round(coef(model)[2], 3), "\* calls\n")