॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE

Symbiosis International (Deemed University)

(Established under section 3 of the UGC Act, 1956)

Re-accredited by NAAC with 'A' grade (3.58/4) | Awarded Category – I by UGC

Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

Assignment No. 08

Subject: Data Science Lab

Name of Student	Achyut Shukla
PRN No.	**20070122005**
Branch	CS
Class	A1
Academic Year & Semester	2023-24 _ 7th semester
Date	**4th October**
Title of Lab Assignment	**REGRESSION MODEL FOR PREDICTION**

Theory:

Linear Regression

Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The steps to perform linear regression

- *Design Matrix: This matrix contains all the independent variables. For categorical data like 'State', you'd typically use one-hot encoding to convert it into numerical values.*
- *Fitting the Model: The coefficients of the equation are estimated based on the given data. In simple terms, the aim is to find the best-fitting line (in case of multiple regressions, it's a hyperplane) that represents the relationship between the variables.*
- *Prediction: Once the model is trained, you can input the features of a new dataset and get predictions for the dependent variable (in this case, 'Profit').*

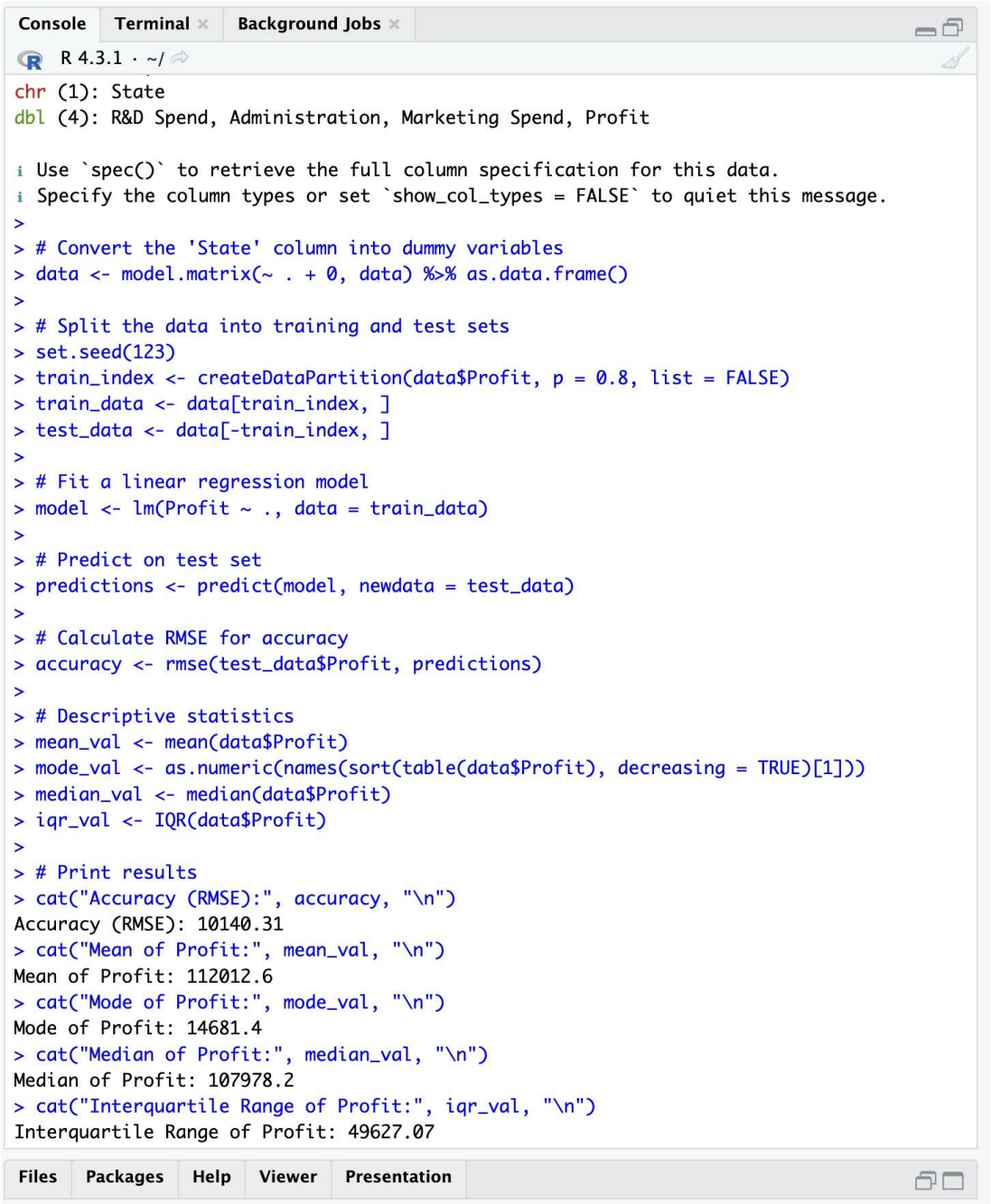
Descriptive Statistics

Descriptive statistics provide a summary or description of the main aspects of the data:

- *Mean: Represents the average value. It's calculated by summing up all the values and dividing by the count of values.*
- *Mode: Represents the value that appears most frequently in the dataset.*
- *Median: It's the middle value of a dataset when sorted in order. For an even number of data points, it's the average of the two middle numbers.*
- *Interquartile Range (IQR): Represents the range within which the middle 50% of values lie. It's calculated as the difference between the third quartile (Q3) and the first quartile (Q1).*

Answer:

 # Print results
> cat("Accuracy (RMSE):", accuracy, "\n")
Accuracy (RMSE): 10140.31
> cat("Mean of Profit:", mean_val, "\n")
Mean of Profit: 112012.6
> cat("Mode of Profit:", mode_val, "\n")
Mode of Profit: 14681.4
> cat("Median of Profit:", median_val, "\n")
Median of Profit: 107978.2
> cat("Interquartile Range of Profit:", iqr_val, "\n")
Interquartile Range of Profit: 49627.07



Console | Terminal | Background Jobs

R 4.3.1 · ~/

```
chr (1): State
dbl (4): R&D Spend, Administration, Marketing Spend, Profit

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
>
> # Convert the 'State' column into dummy variables
> data <- model.matrix(~ . + 0, data) %>% as.data.frame()
>
> # Split the data into training and test sets
> set.seed(123)
> train_index <- createDataPartition(data$Profit, p = 0.8, list = FALSE)
> train_data <- data[train_index, ]
> test_data <- data[-train_index, ]
>
> # Fit a linear regression model
> model <- lm(Profit ~ ., data = train_data)
>
> # Predict on test set
> predictions <- predict(model, newdata = test_data)
>
> # Calculate RMSE for accuracy
> accuracy <- rmse(test_data$Profit, predictions)
>
> # Descriptive statistics
> mean_val <- mean(data$Profit)
> mode_val <- as.numeric(names(sort(table(data$Profit), decreasing = TRUE)[1]))
> median_val <- median(data$Profit)
> iqr_val <- IQR(data$Profit)
>
> # Print results
> cat("Accuracy (RMSE):", accuracy, "\n")
Accuracy (RMSE): 10140.31
> cat("Mean of Profit:", mean_val, "\n")
Mean of Profit: 112012.6
> cat("Mode of Profit:", mode_val, "\n")
Mode of Profit: 14681.4
> cat("Median of Profit:", median_val, "\n")
Median of Profit: 107978.2
> cat("Interquartile Range of Profit:", iqr_val, "\n")
Interquartile Range of Profit: 49627.07
```

Files | Packages | Help | Viewer | Presentation

Code:

```
# load required libraries
install.packages (c("readr", "dplyr", "caret", "lmtest", "Metrics"))
library(readr)
library(dplyr)
library (caret)
library(lmtest)
librory (Metrics)

# Load the data
data <- read.csv("Users/Achyut/Documents/DS_Lab/Assignment_8/50_Startups.csv)

# Convert the "State' colum into dummy voriables
data <- model.matrix(~ . + 0, data) %>% as.data.frame()

# Split the doto into training ond test sets
set.seed(123)
train_index <- createDataPartition(data$Profit, p = 0.8, list - FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, J

# Prediction on test set
predictions <- predict(model, newdata = test_data)

# Calculate RXSE for occurocy
accuracy <- rmse(test_datas$Profit, predictions)

# Descriptive statistics
mean_val <- mean(data$Profit)
mode_val <- as.numeric(name(sort(table(data$Profit), decreasing = TRUE)[1]))
median_val <- median(data$Profit)
iqr_val <- IQR(data$Profit)

# Print results
cat("Accuracy (RMSE):", accuracy, "\n")
cat("Mean of Profit:", mean_val, "\n")
cat("Mode of Profit:", mode_val, "\n")
cat("Median of Profit:", median_val, "\n")
cat("Interquartile Range of Profit:", mode_val, "\n")
```

Conclusion: The objective of the study was to examine the influence of various financial approaches and operational conditions on the profitability of startup companies. Using linear regression, we assessed the correlation between profits and expenditures in various departments. Descriptive statistics were used to gain an understanding of the overall profit distribution among the startups, while the RMSE served as a measure of the model's performance.