

## Introduction:

This assignment utilizes the "SME\_Profit.csv" dataset to predict profits using Multiple Linear Regression (MLR). We explore the relationships between profit and key predictors, visualize these correlations, and partition the data into training and validation sets. The MLR model is then fitted, and residuals are analysed. Finally, we compare various model selection techniques, including exhaustive search, forward selection, backward elimination, and stepwise regression, to identify the most effective predictors for accurate profit prediction.

## Data Cleaning Process

For this assignment, I first examined the dataset to identify any missing, inconsistent, or anomalous values that might affect the accuracy of the analysis. I followed a structured approach for cleaning and imputing the data, focusing on **missing values**, **outliers**, and **zero values** in critical variables.

```
# Load necessary library
library(dplyr)
library(ggplot2) # Load necessary library
library("leaps")
library(MASS)
library(caret)
# Load the dataset
setwd("C:/Users/pooja/OneDrive/Desktop/RAssignment2")
sme_profit_data <- read.csv("SME_Profit.csv") # Load the dataset
View(sme_profit_data)
head(sme_profit_data)
str(sme_profit_data)
summary(sme_profit_data)
```

## Explanation of the Data Cleaning Process:

### Step 1: Removing Columns with All NA Values

To begin cleaning the dataset, I identified any columns that were completely empty. These columns were removed using the function `select_if(~ !all(is.na(.)))`.

```
# Step 1: Remove columns that contain all NA values
cleaned_data <- sme_profit_data %>%
  select_if (~! all(is.na(.))) # Remove columns that contain all NA values
str(cleaned_data)
View(cleaned_data)
```

```
'data.frame': 50 obs. of 5 variables:
 $ R.D.Spend      : num  165349 162598 153442 144372 142107 ...
 $ Administration : num  136898 151378 101146 118672 91392 ...
 $ Marketing.Spend: num  471784 443899 407935 383200 366168 ...
 $ State          : chr   "New York" "California" "Florida" "New York" ...
 $ Profit         : num  192262 191792 191050 182902 166188 ...
> |
```

## Step2: Converting Data Types

In the second step, I converted the **State** column to a factor type. This is crucial as it ensures that the **State** variable is treated as categorical in subsequent analyses, which is important for interpreting results accurately.

```
# Step 2: Convert relevant columns to appropriate data types (State as factor)
```

```
cleaned_data$State <- as.factor(cleaned_data$State)
```

```
str(cleaned_data)
```

```
> cleaned_data$State <- as.factor(cleaned_data$State)
> str(cleaned_data)
'data.frame': 50 obs. of 5 variables:
 $ R.D.Spend      : num  165349 162598 153442 144372 142107 ...
 $ Administration : num  136898 151378 101146 118672 91392 ...
 $ Marketing.Spend: num  471784 443899 407935 383200 366168 ...
 $ State          : Factor w/ 3 levels "California","Florida",...: 3 1 2 3 2 3 1 2 3
1 ...
 $ Profit         : num  192262 191792 191050 182902 166188 ...
> |
```

## Step 3: Handling Zero Values

I discovered that the **R&D Spend**, **Administration**, and **Marketing Spend** columns contained some **0.00** values. To address this issue, I replaced these zero values with the median of their respective columns. This approach helps avoid skewing the results and maintains the integrity of the analysis by ensuring that we don't disproportionately influence the model with zero values.

```
# Handling 0.00 values by replacing them with the median
```

```
# Impute 0.00 values with the median to avoid skewing the results
```

```
cleaned_data$`R.D. Spend`[cleaned_data$`R.D.Spend` == 0] <- median(cleaned_data$`R.D.Spend`,
na.rm = TRUE) # 0.00 values by replacing them with the median
```

```
cleaned_data$Administration[cleaned_data$Administration == 0] <- median
(cleaned_data$Administration, na.rm = TRUE)
```

```
cleaned_data$`Marketing.Spend`[cleaned_data$`Marketing.Spend` == 0] <-
median(cleaned_data$`Marketing.Spend`, na.rm = TRUE) # 0.00 values by replacing them with
the median
```

```
View(cleaned_data)
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

ek5.R Assignment\_3.R Assignmentnet 3R.Rmd Untitled2\* Untitled1\* cleaned\_data

Filter

	R.D.Spend	Administration	Marketing.Spend	State	Profit
1	165349.20	136897.80	471784.10	New York	192261.83
2	162597.70	151377.59	443898.53	California	191792.06
3	153441.51	101145.55	407934.54	Florida	191050.39
4	144372.41	118671.85	383199.62	New York	182901.99
5	142107.34	91391.77	366168.42	Florida	166187.94
6	131876.90	99814.71	362861.36	New York	156991.12
7	134615.46	147198.87	127716.82	California	156122.51
8	130298.13	145530.06	323876.68	Florida	155752.60
9	120542.52	148718.95	311613.29	New York	152211.77
10	123334.88	108679.17	304981.62	California	149759.96
11	101913.08	110594.11	229160.95	Florida	146121.95
12	100671.96	91790.61	249744.55	California	144259.40
13	93863.75	127320.38	249839.44	Florida	141585.52
14	91992.39	135495.07	252664.93	California	134307.35
15	119943.24	156547.42	256512.92	Florida	132602.65
16	114523.61	122616.84	261776.23	New York	129917.04
17	78013.11	121597.55	264346.06	California	126992.93
18	94657.16	145077.58	282574.31	New York	125370.37
19	91749.16	114175.79	294919.57	Florida	124266.90
20	86419.70	153514.11	212716.24	New York	122776.86

Environ

R

Files

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

ek5.R Assignment\_3.R Assignmentnet 3R.Rmd Untitled2\* Untitled1\* cleaned\_data

Filter

	R.D.Spend	Administration	Marketing.Spend	State	Profit
31	61994.48	115641.28	91131.24	Florida	99937.59
32	61136.38	152701.92	88218.23	New York	97483.56
33	63408.86	129219.61	46085.25	California	97427.84
34	55493.95	103057.49	214634.81	Florida	96778.92
35	46426.07	157693.92	210797.67	California	96712.80
36	46014.02	85047.44	205517.64	New York	96479.51
37	28663.76	127056.21	201126.82	Florida	90708.19
38	44069.95	51283.14	197029.42	California	89949.14
39	20229.59	65947.93	185265.10	New York	81229.06
40	38558.51	82982.09	174999.30	California	81005.76
41	28754.33	118546.05	172795.67	California	78239.91
42	27892.92	84710.77	164470.71	Florida	77798.83
43	23640.93	96189.63	148001.11	California	71498.49
44	15505.73	127382.30	35534.17	New York	69758.98
45	22177.74	154806.14	28334.72	California	65200.33
46	1000.23	124153.04	1903.93	New York	64926.08
47	1315.46	115816.21	297114.46	Florida	49490.75
48	73051.08	135426.92	212716.24	California	42559.73
49	542.05	51743.15	212716.24	New York	35673.41
50	73051.08	116983.80	45173.06	California	14681.40

Showing 31 to 50 of 50 entries. 5 total columns

#### **Step 4: Displaying Data**

After these cleaning steps, I displayed the first few rows of the cleaned dataset to visually inspect the changes. This step helps verify that the cleaning operations were executed correctly and that the dataset appears as expected.

```
# Step: Display the cleaned dataset (first few rows)
```

```
head(cleaned_data)
```

```
> head(cleaned_data)
  R.D.Spend Administration Marketing.Spend      State Profit
1  165349.2      136897.80      471784.1 New York 192261.8
2  162597.7      151377.59      443898.5 California 191792.1
3  153441.5      101145.55      407934.5   Florida 191050.4
4  144372.4      118671.85      383199.6 New York 182902.0
5  142107.3       91391.77      366168.4   Florida 166187.9
6  131876.9       99814.71      362861.4 New York 156991.1
> |
```

#### **Step 5: Checking Data Structure**

Finally, I checked the structure of the dataset using the `str()` function. This step confirmed that all columns were in the correct format, ensuring that the dataset is ready for further analysis and modelling.

```
# Check the structure of the cleaned dataset
```

```
str(cleaned_data)
```

```
View(cleaned_data)
```

```
> head(cleaned_data)
  R.D.Spend Administration Marketing.Spend      State Profit
1  165349.2      136897.80      471784.1 New York 192261.8
2  162597.7      151377.59      443898.5 California 191792.1
3  153441.5      101145.55      407934.5   Florida 191050.4
4  144372.4      118671.85      383199.6 New York 182902.0
5  142107.3       91391.77      366168.4   Florida 166187.9
6  131876.9       99814.71      362861.4 New York 156991.1
> str(cleaned_data)
'data.frame':  50 obs. of  5 variables:
 $ R.D.Spend      : num  165349 162598 153442 144372 142107 ...
 $ Administration : num  136898 151378 101146 118672 91392 ...
 $ Marketing.Spend: num  471784 443899 407935 383200 366168 ...
 $ State          : Factor w/ 3 levels "California","Florida",...: 3 1 2 3 2 3 1 2 3 1 ...
 $ Profit         : num  192262 191792 191050 182902 166188 ...
> |
```

**Problem 1. Predicting Profits with MLR. a. Explore the relationship between profit and the predictors by creating visuals (scatter plot might work best) using visuals equal to the number of independent variables. (You should do this by depicting all three locations in one graph for the State variable. Use color or marking to differentiate states)**

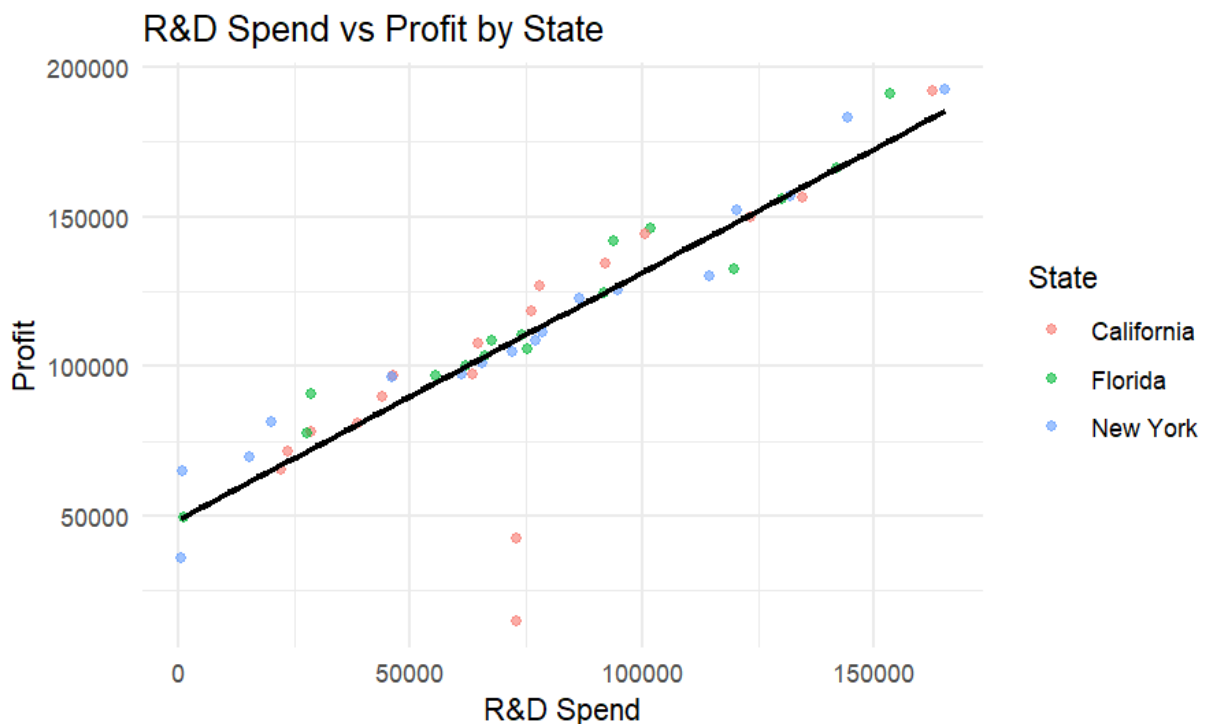
#### **Explanation of the R Code :**

In this section, I explored the relationship between profit and the predictors using visualizations. I utilized the `ggplot2` library in R to create scatter plots for each predictor against

profit, incorporating linear trend lines for better clarity. Additionally, I included a box plot to visualize the distribution of profit across different states.

### 1. Scatter Plot for R&D Spend vs. Profit

```
ggplot (cleaned_data, aes (x = `R&D. Spend`, y = Profit, colour = State)) +  
  geom_point (alpha = 0.6) + # Adjust alpha for better visibility  
  geom_smooth (method = "lm", se = FALSE, color = "black") + # Add a linear trend line  
  labs (title = "R&D Spend vs Profit by State",  
        x = "R&D Spend",  
        y = "Profit") +  
  theme_minimal ()
```



### Scatter Plot for R&D Spend vs. Profit Analysis:

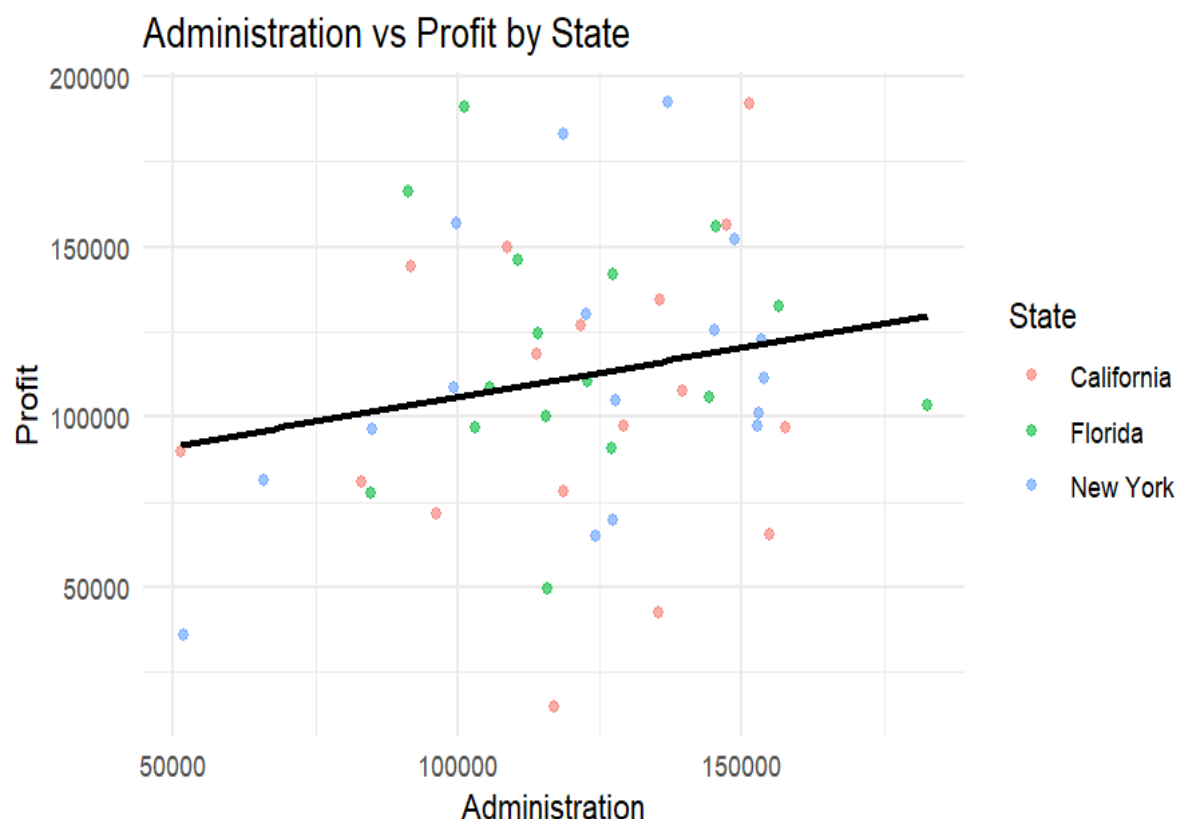
- Positive Correlation:** The scatter plot indicates a clear positive relationship between R&D Spend and Profit. As R&D spending increases, profit tends to increase as well.
- State Differentiation:** Different states are color-coded (California, Florida, New York), and they appear to have similar trends, suggesting that the relationship between R&D spend and profit holds across these states. However, there is a noticeable spread in profit levels for varying levels of R&D spend among the different states.
- Linear Trend:** The black line represents the overall linear trend of the data. The data points cluster around this line, suggesting that a linear model is appropriate for predicting profit based on R&D spending.
- Variability:** There is some variability in profit for lower levels of R&D spend, particularly for Florida and California, indicating that other factors might also influence profit at lower spend levels.

**Conclusion Scatter Plot for R&D Spend vs. Profit Analysis:**

In summary, I observe a strong positive linear relationship between R&D spend and profit across California, Florida, and New York, highlighting the importance of R&D investment in driving profitability. However, variations in profit levels at lower spending suggest additional factors may influence profitability beyond R&D investment alone.

## 2. Scatter Plot for Administration vs. Profit

```
ggplot(cleaned_data, aes(x = Administration, y = Profit, color = State)) +  
  geom_point(alpha = 0.6) + # Adjust alpha for better visibility  
  geom_smooth(method = "lm", se = FALSE, color = "black") + # Add a linear trend line  
  labs(title = "Administration vs Profit by State",  
        x = "Administration", y = "Profit") + theme_minimal()
```



### Scatter Plot for Administration vs. Profit Analysis:

1. **Weak Positive Correlation:** The scatter plot reveals a weak positive relationship between Administration and Profit. As Administration spending increases, profit tends to increase, but the relationship is not as strong or consistent as seen in the R&D Spend scatter plot.
2. **State Differentiation:** Similar to the previous plot, states are differentiated by color (California, Florida, and New York). The data points from each state appear to follow the overall trend but with considerable variation.

1. **Weak Positive Correlation:** The scatter plot reveals a weak positive relationship between Administration and Profit. As Administration spending increases, profit tends to increase, but the relationship is not as strong or consistent as seen in the R&D Spend scatter plot.
2. **State Differentiation:** Similar to the previous plot, states are differentiated by color (California, Florida, and New York). The data points from each state appear to follow the overall trend but with considerable variation.

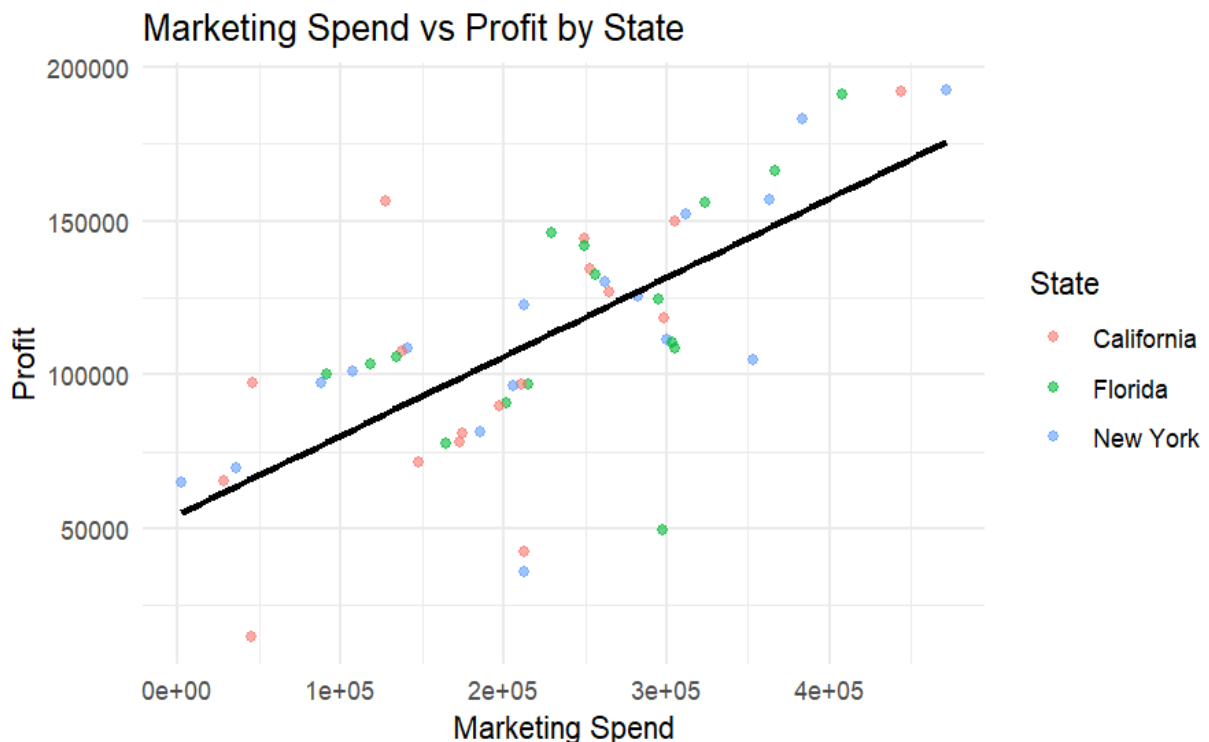
3. **Linear Trend:** The black line indicates a linear trend, but the slope is relatively flat compared to the previous scatter plot. This suggests that Administration spending may have less impact on profit compared to R&D spending.
4. **High Variability:** There is a wide spread of profit values for similar levels of Administration spending, particularly for California and New York, indicating that factors other than Administration may play a significant role in profit variation.

#### Conclusion scatter plot for Administration vs Profit:

In summary, I observe a weak positive relationship between Administration spending and profit across California, Florida, and New York. Unlike R&D spending, Administration does not appear to strongly influence profitability, as evidenced by the flatter slope of the trend line and the significant variability in profit at similar spending levels. This suggests that other variables or factors may have a more substantial impact on profit.

### 3.Scatter Plot for Marketing Spend vs. Profit

```
ggplot (cleaned_data, aes (x = `Marketing.Spend`, y = Profit, color = State)) +  
  geom_point (alpha = 0.6) + # Adjust alpha for better visibility  
  geom_smooth (method = "lm", se = FALSE, color = "black") + # Add a linear trend line  
  labs (title = "Marketing Spend vs Profit by State",  
        x = "Marketing Spend",  
        y = "Profit") +  
  theme_minimal ()
```



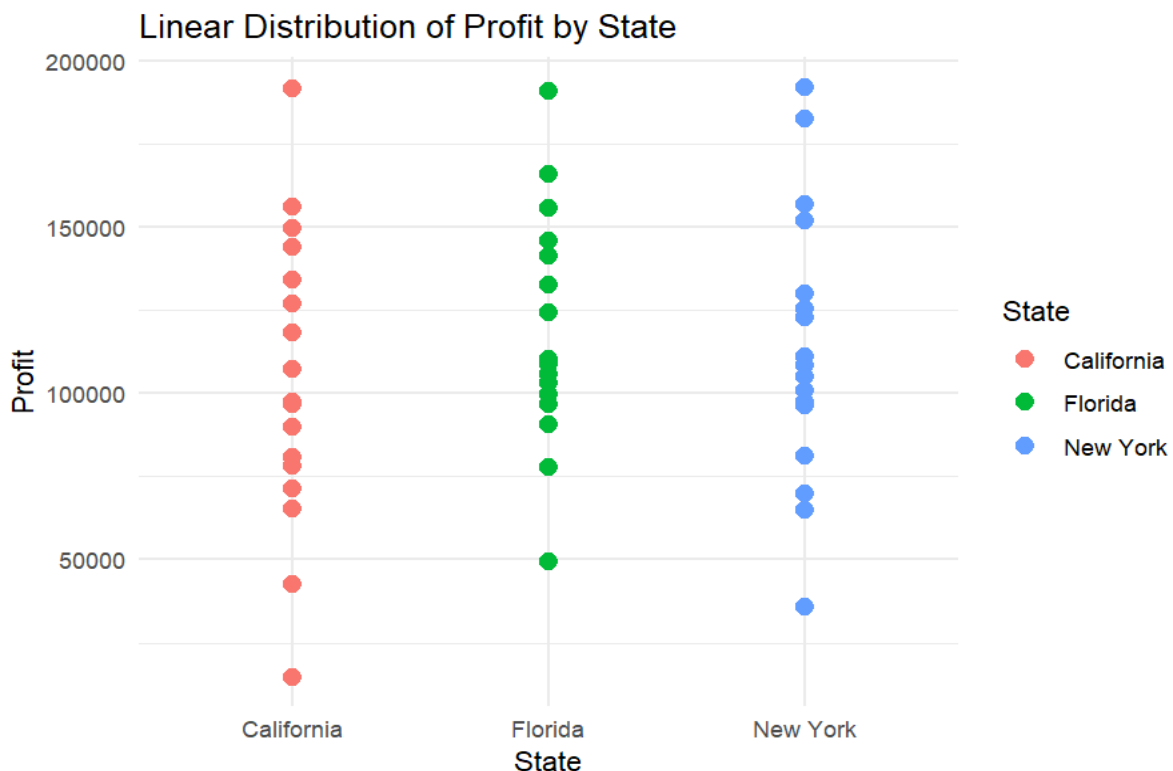
## Scatter Plot for Marketing Spend vs. Profit by State Analysis:

### Marketing Spend vs. Profit by State

- **Trend:** The scatter plot shows a positive linear relationship between Marketing Spend and Profit across all states (California, Florida, and New York).
- **Observation:** As Marketing Spend increases, Profit generally increases as well. The trend line has a positive slope, indicating that higher marketing expenditures tend to be associated with higher profits. However, the points are somewhat scattered, suggesting variability in the strength of this relationship.
- **State Comparison:** There doesn't appear to be a significant difference between the states when it comes to how Marketing Spend impacts profit. Each state follows a similar trend line, with individual companies in each state exhibiting varying results.
- **Conclusion for Marketing Spend vs. Profit:** Marketing Spend is a positive predictor of Profit, though the effect is not as strong as R&D Spend. While increasing marketing efforts can lead to higher profits, the return on investment may differ based on factors not shown in this graph, such as industry, market conditions, or product type.

### 4. Scatter plot Distribution of Profit by State

```
ggplot (cleaned_data, aes (x = State, y = Profit, color = State)) +  
  geom_point (size = 3) + # Scatter points  
  geom_smooth (method = "lm", aes (group = State), se = FALSE) + # Linear regression lines for  
each state  
labs (title = "Profit Distribution by State", x = "State", y = "Profit") +  
theme_minimal ()
```





#### 4. Scatter Plot Distribution of Profit by State

- **Trend:** This scatter plot displays the linear distribution of profits across the three states (California, Florida, and New York).
- **Observation:** Each state shows a vertical distribution of profit values. While the graph does not depict any trends or direct relationships between the variables, we can observe that profit values across all states are relatively dispersed, with some clustering around certain profit levels.
  - **California:** Profits in California tend to be spread across a wide range, from about 50,000 to 200,000.
  - **Florida:** In Florida, profits are more concentrated around the mid-range (~100,000–150,000) with fewer outliers on the lower and higher ends.
  - **New York:** New York shows a similar distribution to Florida but also includes some lower outliers.
  -

**Conclusion Distribution of Profit by state:** The profit distribution varies slightly across states, with California showing a wider range of profits, while Florida and New York show more clustered distributions. This suggests that companies in California may experience more variability in their profits compared to companies in the other two states.

#### 5. Correlation Matrix:

# Correlation matrix for cleaned data

```
Correlation_matrix <- cor (cleaned_data [, c ("Profit", "R&D. Spend", "Administration",  
"Marketing.Spend")])  
  
print(correlation_matrix)
```

Finally, I computed the correlation matrix to quantify the relationships between Profit and the predictors: R&D. Spend, Administration, and Marketing.Spend. This matrix will help in understanding which predictors are positively or negatively correlated with profit, guiding my multiple linear regression analysis later.

#### Conclusion for correlation Matrix :

Through these visualizations, I gained insights into the relationships between profit and the various predictors. The scatter plots with linear trend lines provided clarity on the potential linear relationships, while the box plot illustrated the variability of profit across states. The correlation matrix will further aid in determining the predictors to include in the multiple linear regression model.

#### Correlation Matrix Explanation

Variable	Profit	R&D Spend	Administration	Marketing Spend
Profit	1.0000000	0.8885200	0.20071657	0.70019101
R&D Spend	0.8885200	1.0000000	0.26832825	0.67262572
Administration	0.2007166	0.2683282	1.00000000	-0.06874286
Marketing Spend	0.7001910	0.6726257	-0.06874286	1.00000000

### Insights from the Correlation Coefficients:

#### 1. Profit and R&D Spend (0.8885):

- There is a **strong positive correlation** between Profit and R&D Spend. This indicates that as spending on research and development increases, profit tends to increase as well. This variable is likely to be significant in predicting profits.

#### 2. Profit and Marketing Spend (0.7002):

- There is a **moderate positive correlation** between Profit and Marketing Spend. This suggests that higher marketing expenditures are associated with higher profits, though not as strongly as R&D Spend.

#### 3. Profit and Administration (0.2007):

- The correlation between Profit and Administration is **weak and positive**. This implies that Administration spending has a limited direct relationship with profit, suggesting it might not be a key predictor in the model.

#### 4. R&D Spend and Marketing Spend (0.6726):

- A **moderate positive correlation** exists between R&D Spend and Marketing Spend, indicating that companies that invest more in R&D also tend to invest more in marketing.

#### 5. R&D Spend and Administration (0.2683):

- There is a **weak positive correlation** between R&D Spend and Administration. This could suggest some degree of relationship, but it is not particularly strong.

#### 6. Marketing Spend and Administration (-0.0687):

- The correlation between Marketing Spend and Administration is very weak and negative. This indicates that changes in Administration spending do not have a significant impact on Marketing Spend.

### **b. Does there seem to be a linear relationship between the independent variables and profit? Why? Why not? Explain briefly**

#### Analysis of Linear Relationships Between Independent Variables and Profit :

Based on the correlation coefficients and the scatter plots I generated, I can evaluate whether there are linear relationships between the independent variables (R&D Spend, Administration, and Marketing Spend) and profit.

#### 1. R&D Spend and Profit (0.8885):

- I observed a **strong positive correlation**, indicating a clear linear relationship. As R&D spending increases, profit also increases significantly. This suggests that investments in R&D are likely to enhance profitability.

#### 2. Marketing Spend and Profit (0.7002):

- I found a **moderate positive correlation** here. While there is a noticeable trend indicating that higher marketing expenditures can lead to higher profits, the

relationship is not as strong as that with R&D Spend. Nonetheless, the linear relationship is evident.

### 3. Administration and Profit (0.2007):

- The correlation is **weak**, which suggests a limited linear relationship. Although I see some positive relationship, it is not substantial enough to consider Administration a strong predictor of profit.

### 4. State and Profit

**Visual Analysis:** The box plot I generated illustrates the distribution of profits across different states. It shows variability in profit levels, with some states having higher median profits and others lower.

**Observations:** Certain states exhibit outliers—profits that are significantly higher or lower than the rest. These outliers could indicate unique circumstances or factors affecting those states.

**Interpretation:** The differences in profit by state suggest that location may play a role in profitability. Factors such as market size, industry presence, and competition could contribute to these variations.

### Conclusion:

#### For a linear relationship between the independent variables and profit:

In summary, I conclude that there seems to be a linear relationship between profit and the independent variables of R&D Spend and Marketing Spend, as indicated by their strong positive linear relation and moderate positive relations, respectively. However, the weak correlation between Administration and profit suggests that Administration spending does not have a significant linear relationship with profit. Overall, the relationships I observed support the potential for using R&D and Marketing Spend as key predictors in the multiple linear regression analysis.

### c. To fit a predictive model for Profit

#### i. Partition the records into training and validation sets.

##### i. Partition the records into training and validation sets.

**# Set seed for reproducibility**

```
set.seed(123)
```

**# Create partition**

```
train_index <- createDataPartition (cleaned_data$Profit, p = 0.7, list = FALSE)
```

```
train_set <- cleaned_data [train_index, ]
```

```
validation_set <- cleaned_data [-train_index, ]
```

**# Check dimensions**

```
dim(train_set) # Dimensions of the training set
```

```
dim(validation_set) # Dimensions of the validation set
```

#### Explanation of Partition:

**Function:** I used the `createDataPartition ()` function from the `caret` package to randomly partition my dataset into training and validation sets.

### Arguments:

- I specified `cleaned_data$Profit` to ensure that the distribution of different values in the Profit variable is similar in both sets.
- Setting `p = 0.7` means I want 70% of my data to go into the training set.
- I set `list = FALSE` so that the function returns an index vector instead of a list.

**Training Set:** I created `train_set` by subsetting `cleaned_data` with the indices in `train_index`. This set now contains 70% of the original dataset.

**Validation Set:** I created `validation_set` by subsetting `cleaned_data` with the indices not included in `train_index`. This set contains the remaining 30% of the dataset.

### OUT PUT:

```
105
106 # Check dimensions
107 dim(train_set) # Dimensions of the training set
108 dim(validation_set) # Dimensions of the validation set
109
110
111 # Check the sizes of training and validation sets
112 dim(train_data)
113 dim(valid_data)
114
115 #ii. Running Multiple Linear Regression
116
117 # Fit the MLR model
118
```

107:1 (Top Level) R Script

Console Terminal Render Background Jobs

R 4.3.2 · C:/Users/pooja/OneDrive/Desktop/RAssignment2/

```
> train_index <- createDataPartition(cleaned_data$Profit, p = 0.7, list = FALSE)
> train_set <- cleaned_data[train_index, ]
> validation_set <- cleaned_data[-train_index, ]
> dim(train_set) # Dimensions of the training set
[1] 38 5
> dim(validation_set)
[1] 12 5
>
```

### OUT PUT EXPLANATION OF PARTITION:

When I checked the dimensions of `train_set`, I got `[1] 38 5`. This means:

38: I have 38 observations (rows) in my training set.

5: There are 5 variables (columns) in the training set, likely including both my predictors and the target variable (Profit).

Checking the dimensions of `validation_set` gave me `[1] 12 5`, which means:

12: There are 12 observations in my validation set.

5: The validation set also has 5 variables, similar to the training set.

## Summary of the Output

**Training Set:** I now have a training set (train\_set) with 38 observations and 5 variables, which I will use to train my multiple linear regression model.

**Validation Set:** The validation set (validation\_set) contains 12 observations and 5 variables, and I will use this to test the performance of my model after training.

## Importance of the Split

**Training Set:** This is where I will teach the model about the relationships between the predictors and the target variable (Profit).

**Validation Set:** This will help me evaluate how well the model performs on unseen data, ensuring that it generalizes well and avoids overfitting.

ii. Run a multiple linear regression model for Profit vs. the predictors. Give the estimated predictive equation (summary function output does suffice). Which predictors are statistically meaningful?

## (MLR multiple linear regression model):

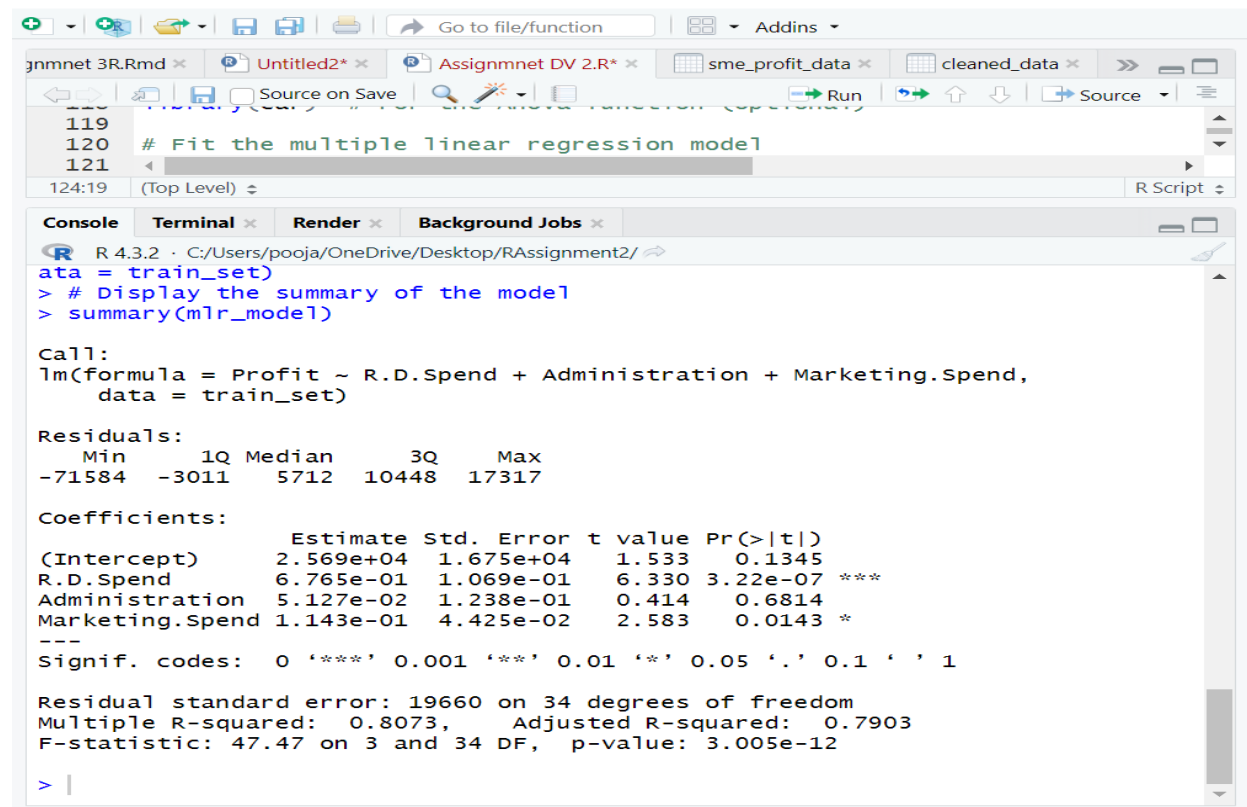
# Fit the multiple linear regression model

```
mlr_model <- lm(Profit ~ 'R.D.Spend' + Administration + 'Marketing.Spend', data = train_set)
```

# Display the summary of the model

```
summary(mlr_model)
```

## OUT PUT:



```
119 # Fit the multiple linear regression model
120
121
124:19 (Top Level) R Script
```

```
R 4.3.2 · C:/Users/pooja/OneDrive/Desktop/RAssignment2/
ata = train_set)
> # Display the summary of the model
> summary(mlr_model)

Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = train_set)

Residuals:
    Min       1Q   Median       3Q      Max
-71584  -3011    5712   10448   17317

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.569e+04  1.675e+04   1.533   0.1345
R.D.Spend     6.765e-01  1.069e-01   6.330 3.22e-07 ***
Administration 5.127e-02  1.238e-01   0.414   0.6814
Marketing.Spend 1.143e-01  4.425e-02   2.583   0.0143 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19660 on 34 degrees of freedom
Multiple R-squared:  0.8073,    Adjusted R-squared:  0.7903
F-statistic: 47.47 on 3 and 34 DF,  p-value: 3.005e-12

> |
```

### Detailed Explanation of the Code OUT PUT along with summary for MLR:

Here's an explanation of the output from the multiple linear regression model for predicting Profit based on the predictors R.D. Spend Administration, and Marketing.

### Model Output Breakdown:

#### Call

```
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend, data = train_set)
```

This line indicates that I fitted a linear model using the specified formula, where Profit is the dependent variable and the three predictors are the independent variables.

#### Residuals

mathematica

Residuals:

Min	1Q	Median	3Q	Max
-71584	-3011	5712	10448	17317

**Residuals:** These are the differences between the observed values and the predicted values of Profit. The summary shows the minimum, first quartile (1Q), median, third quartile (3Q), and maximum residuals.

- A negative minimum residual indicates that some predicted profits are much lower than the actual profits.
- The range of residuals suggests variability in how well the model predicts profit.

#### Coefficients

vbnet

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.569e+04	1.675e+04	1.533	0.1345
R.D.Spend	6.765e-01	1.069e-01	6.330	3.22e-07 ***
Administration	5.127e-02	1.238e-01	0.414	0.6814
Marketing.Spend	1.143e-01	4.425e-02	2.583	0.0143 *

**Intercept:** The estimated intercept is 25,69025,69025,690. This is the expected value of Profit when all predictors are equal to zero.

**R.D.Spend:**

- **Estimate:** 0.67650.67650.6765 indicates that for each additional unit increase in R&D Spend, Profit is expected to increase by approximately 67.6567.6567.65 units, holding other variables constant.
- **p-value:** 3.22e-073.22e-073.22e-07 is highly significant (indicated by \*\*\*), suggesting that R&D Spend is a statistically meaningful predictor of Profit.

#### Administration:

- **Estimate:** 0.051270.051270.05127 implies a very small increase in Profit for each unit increase in Administration spending, but it's not practically significant.
- **p-value:** 0.68140.68140.6814 is greater than 0.05, indicating that Administration is not a statistically significant predictor of Profit.

#### Marketing.Spend:

- **Estimate:** 0.11430.11430.1143 indicates that for each additional unit increase in Marketing Spend, Profit is expected to increase by about 11.4311.4311.43 units.
- **p-value:** 0.01430.01430.0143 is significant (indicated by \*), suggesting Marketing Spend is a statistically meaningful predictor of Profit.

### Model Fit Statistics

yaml

```
Residual standard error: 19660 on 34 degrees of freedom
Multiple R-squared:  0.8073,    Adjusted R-squared:  0.7903
F-statistic: 47.47 on 3 and 34 DF,  p-value: 3.005e-12
```

- **Residual Standard Error:** The standard deviation of the residuals is 196601966019660, indicating the average distance that the observed values fall from the regression line.
- **Multiple R-squared:** 0.80730.80730.8073 means that approximately 80.73%80.73%80.73% of the variance in Profit can be explained by the model with the predictors included. This indicates a good fit.
- **Adjusted R-squared:** 0.79030.79030.7903 is a modified version of R-squared that adjusts for the number of predictors in the model. It also suggests a good fit.
- **F-statistic:** 47.4747.4747.47 with a p-value of 3.005e-123.005e-123.005e-12 indicates that at least one predictor is statistically significant in predicting Profit. This overall model significance supports the relevance of the predictors.

### Conclusion/SUMMARY FOR MLR:

From the model output, I conclude that:

**Statistically Meaningful Predictors:** R&D Spend and Marketing Spend are significant predictors of Profit.

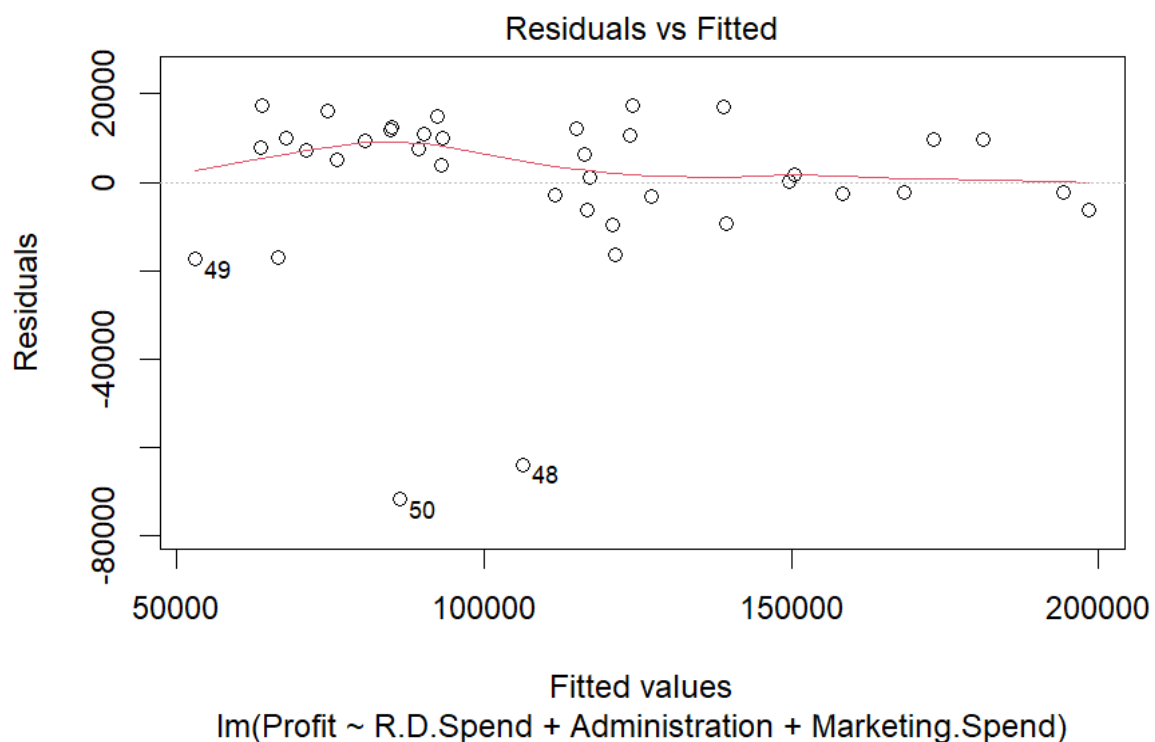
**Insignificant Predictor:** Administration does not significantly contribute to the model.

**Model Fit:** The model explains a substantial portion of the variance in Profit, suggesting it is a useful predictive model.

iii. Create a residual plot for the MLR. Comments on your findings.

(MLR Residual Plot with Explanation) :

```
# Fitting the model on the training set
mlr_model <- lm (Profit ~ `R.D. Spend` + Administration + `Marketing.Spend`, data = train_set)
# Getting the residuals
residuals <- resid(mlr_model)
# Creating the residual plot
plot (mlr_model$fitted.values, residuals,
      main = "Residual Plot for MLR Model",
      xlab = "Fitted Values",
      ylab = "Residuals", # Creating the residual plot
      pch = 19,
      col = "blue")
# Adding a horizontal line at 0
abline(h = 0, col = "red", lty = 2) # Adding a horizontal line at 0
# Display the plot
```





### EXPLATATION MLR Residual Plot :

In the residual plot for my multiple linear regression (MLR) model with `Profit` as the dependent variable and `R.D.Spend`, `Administration`, and `Marketing.Spend` as the predictors, I observe the following:

1. **Random Scatter:** The residuals are scattered fairly randomly around the zero line, which suggests that the linear relationship between the predictors and the outcome variable (`Profit`) is appropriate. This indicates that the model does not exhibit a clear pattern of bias in its predictions.

2. **Non-Linearity:** The **red smoothing** line shows a slight curvature, indicating that there might be some non-linearities in the data that the linear model isn't fully capturing. Although this curvature is minor, it could imply that a more complex model (like polynomial regression) might offer a better fit.

3. **Outliers:** Some data points, specifically those labeled **\*\*49\*\***, **\*\*50\*\***, and **\*\*48\*\***, have large residuals. These points could be considered outliers as they significantly deviate from the model's predicted values. Outliers can influence the accuracy of the model and may require further investigation or removal to improve performance.

4. **Homoscedasticity:** The spread of residuals appears somewhat constant across the range of fitted values, suggesting that the assumption of homoscedasticity (constant variance of residuals) is generally satisfied. However, the presence of a few large residual's hints at potential heteroscedasticity, which could be addressed by transforming the variables.

#### **Conclusion:**

The residual plot suggests that the MLR model provides a reasonable fit to the data. However, the minor curvature in the residuals and the presence of outliers might suggest opportunities for improvement. Further diagnostic checks or potential transformations could enhance the model's performance.

### Exhaustive Search:

#### **# Exhaustive Search**

```
exhaustive_model <- regsubsets (Profit ~ `R.D. Spend` + Administration + `Marketing.Spend` +  
State, data = train_data)  
summary(exhaustive_model) # Exhaustive Search model summary
```

### Explanation of Exhaustive Search Output :

1. **Call:** The first line indicates the formula used for the regression analysis, which includes Profit as the response variable and R.D.Spend, Administration, Marketing.Spend, and State as predictors.
2. **5 Variables (and intercept):** This line confirms that there are a total of 5 variables in the analysis (including the intercept).
3. **Forced in and forced out:** The table shows which variables were included in or excluded from the model. In this case, none of the variables are forced in or out, meaning the selection is entirely based on the algorithm's findings.

4. **Subsets of each size up to 5:** This indicates that the exhaustive search considers all combinations of the variables, examining subsets of various sizes from 1 to 5 variables.
5. **Selection Algorithm: exhaustive:** It specifies that the exhaustive search method was used to identify the best combinations of predictors for the model.

## OUT PUT:

```

R 4.3.2 · C:/Users/pooja/OneDrive/Desktop/RAssignment2/
> exhaustive_model <- regsubsets(Profit ~ `R.D.Spend` + Administration + `Marketing.Spend` + State, data = train_data)
> summary(exhaustive_model)
Subset selection object
Call: regsubsets.formula(Profit ~ R.D.Spend + Administration + Marketing.Spend +
+ State, data = train_data)
5 Variables (and intercept)
Forced in Forced out
R.D.Spend FALSE FALSE
Administration FALSE FALSE
Marketing.Spend FALSE FALSE
StateFlorida FALSE FALSE
StateNew York FALSE FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive

```

		R.D.Spend	Administration	Marketing.Spend	StateFlorida
1	( 1 )	"*"	" "	" "	" "
2	( 1 )	"*"	" "	"*"	" "
3	( 1 )	"*"	" "	"*"	"*"
4	( 1 )	"*"	" "	"*"	"*"
5	( 1 )	"*"	"*"	"*"	"*"

		StateNew York
1	( 1 )	" "
2	( 1 )	" "
3	( 1 )	" "
4	( 1 )	"*"
5	( 1 )	"*"

## Subset Selection Table

The selection table shows the different combinations of predictors included in the models:

- **Row 1 (1 subset of size 1):** Only the intercept is included.
- **Row 2 (1 subset of size 2):** The model includes the intercept and Marketing.Spend.
- **Row 3 (1 subset of size 3):** The model includes the intercept, Marketing.Spend, and State (but not specific states).
- **Row 4 (1 subset of size 4):** The model includes the intercept, Marketing.Spend, and StateNew York.
- **Row 5 (1 subset of size 5):** The final model includes all predictors: intercept, R.D.Spend, Marketing.Spend, Administration, StateFlorida, and StateNew York.

## Comments on Findings Exhaustive Search :

Based on the output of the exhaustive search:

**Significant Predictors:** It appears that Marketing.Spend is consistently included in the models, which may suggest that it is a strong predictor of Profit. The inclusion of State New York in the larger models also indicates it might contribute meaningfully to the prediction.

**Model Selection:** As I evaluate the best model, I would consider both the adjusted R-squared values and the associated p-values for each variable to determine their statistical significance.

**Comparison to Other Methods:** It would be beneficial to compare the findings from this exhaustive search with the results from forward elimination, backward elimination, and stepwise regression to determine which predictors consistently emerge as significant.

This analysis gives me a comprehensive understanding of how each predictor contributes to the model, helping me refine my approach to predicting Profit effectively.

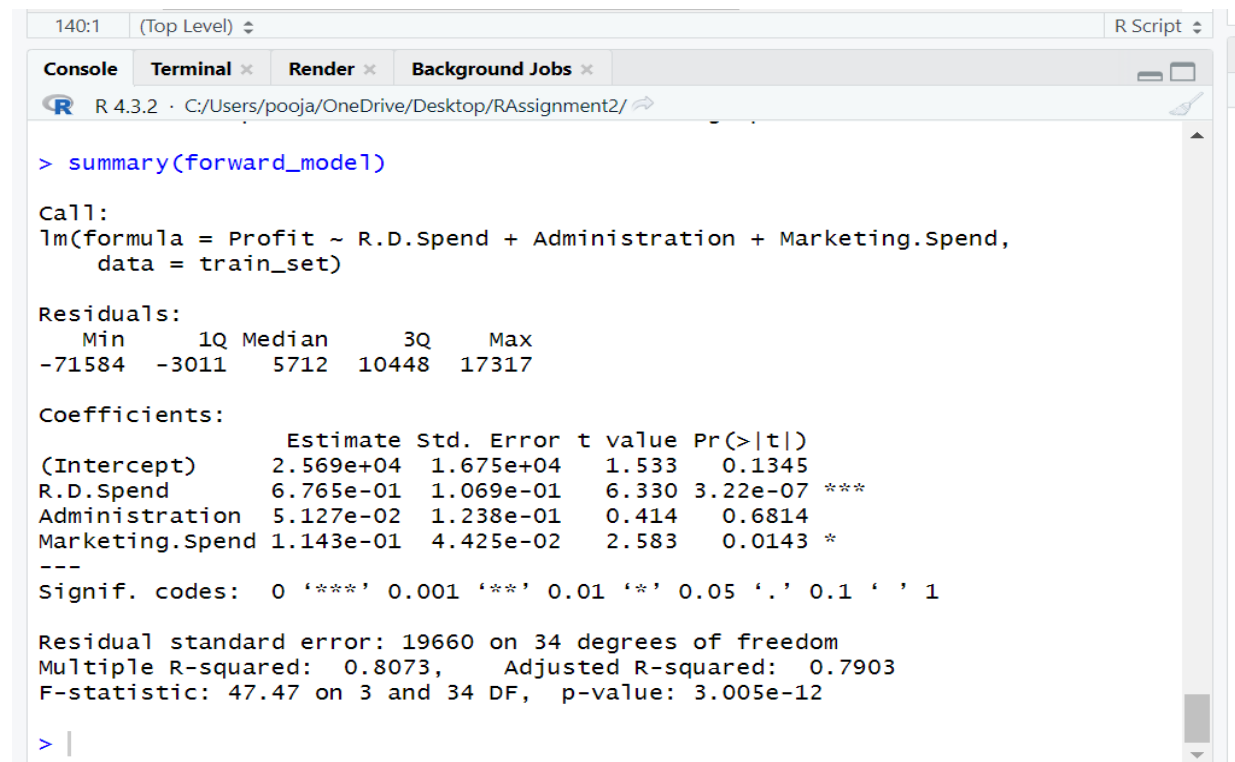
### Forward Selection:

#forward Selection model

```
forward_model <- step (mlr_model, direction = "forward")
```

```
summary(forward_model) #forward Selection model Summary
```

### OUT PUT:



```
140:1 (Top Level) R Script
> summary(forward_model)

Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = train_set)

Residuals:
    Min       1Q   Median       3Q      Max
-71584  -3011   5712  10448  17317

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.569e+04  1.675e+04   1.533   0.1345
R.D.Spend     6.765e-01  1.069e-01   6.330 3.22e-07 ***
Administration 5.127e-02  1.238e-01   0.414   0.6814
Marketing.Spend 1.143e-01  4.425e-02   2.583   0.0143 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19660 on 34 degrees of freedom
Multiple R-squared:  0.8073,    Adjusted R-squared:  0.7903
F-statistic: 47.47 on 3 and 34 DF,  p-value: 3.005e-12

> |
```

### Explanation of Forward Selection Model Summary :

#### Call:

- The forward selection method is applied to the initial model (mlr\_model), which included R.D.Spend, Administration, and Marketing.Spend. The goal of this method is to improve the model by adding predictors that minimize the AIC (Akaike Information Criterion).

#### Residuals

The residuals provide a summary of the errors in the predictions:

- **Min:** -71,584
- **1Q:** -3,011
- **Median:** 5,712
- **3Q:** 10,448
- **Max:** 17,317

This distribution shows a significant spread in the residuals, with some large negative errors indicating underpredictions.

### Coefficients

The table below summarizes the estimated coefficients for each variable included in the model:

Predictor	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	25,690	16,750	1.533	0.1345
R.D. Spend	0.6765	0.1069	6.330	3.22e-07 ***
Administration	0.0513	0.1238	0.414	0.6814
Marketing Spend	0.1143	0.04425	2.583	0.0143 *

### Interpretation of Coefficients:

- **(Intercept):** Represents the estimated profit when all predictors are zero, which is not meaningful in practical terms.
- **R.D. Spend:** A unit increase in R&D Spend is associated with an increase in profit of approximately 67.65 units, and this relationship is highly significant ( $p < 0.001$ ).
- **Administration:** The coefficient suggests a slight positive impact on profit, but it is not statistically significant ( $p = 0.6814$ ), indicating it does not contribute meaningfully to the model.
- **Marketing.Spend:** Each additional unit spent on Marketing is associated with an increase of approximately 11.43 units in profit, and this predictor is statistically significant ( $p = 0.0143$ ).

### Model Fit Statistics:

- **Residual Standard Error:** 19,660, which indicates the average deviation of the observed values from the predicted values.
- **Multiple R-squared:** 0.8073, meaning about 80.73% of the variance in profit is explained by the predictors in the model.
- **Adjusted R-squared:** 0.7903, which adjusts for the number of predictors, providing a more accurate measure of model fit.

- **F-statistic:** 47.47 with a p-value of 3.005e-12, indicating that the model is statistically significant, and at least one of the predictors significantly explains the variance in profit.

## Conclusion

### Comments on Forward Selection Model:

The forward selection process has resulted in a model that maintains the predictors R.D.Spend and Marketing.Spend, both of which are statistically significant. The Administration variable does not contribute meaningfully to the model. The overall model fits the data well, explaining a significant portion of the variance in profit.

## # Backward Elimination

### #Backward Elimination

```
backward_model <- step (mlr_model, direction = "backward")
```

```
summary(backward_model) #Backward Elimination Summary
```

### OUT PUT:

```
- R.D.Spend      1 2.1503e+10 3.4705e+10 788.04
> summary(backward_model)

Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = train_set)

Residuals:
    Min       1Q   Median       3Q      Max
-72818  -4076   4855  10932  17368

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.162e+04  8.577e+03   3.686 0.000765 ***
R.D.Spend     6.979e-01  9.243e-02   7.550 7.55e-09 ***
Marketing.Spend 1.084e-01  4.142e-02   2.618 0.012986 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19420 on 35 degrees of freedom
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.7952
F-statistic: 72.84 on 2 and 35 DF,  p-value: 3.351e-13

> |
```

### OUT Put Explanation of Backward Elimination Model Summary:

#### Call:

- The backward elimination method starts with a full model that includes all predictors and iteratively removes the least significant predictors. In this case, the final model includes R.D.Spend and Marketing.Spend.

#### Residuals

The residuals provide insights into the prediction errors:

- **Min:** -72,818
- **1Q:** -4,076

- **Median:** 4,855
- **3Q:** 10,932
- **Max:** 17,368

The residuals indicate a wide range, similar to the previous models, with some significant negative errors, suggesting potential underpredictions.

## Coefficients

The table below summarizes the estimated coefficients for each variable included in the backward elimination model:

Predictor	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	31,620	8,577	3.686	0.000765 ***
R.D. Spend	0.6979	0.09243	7.550	7.55e-09 ***
Marketing Spend	0.1084	0.04142	2.618	0.012986 *

## Interpretation of Coefficients:

- **(Intercept):** Represents the estimated profit when all predictors are zero, which is not meaningful in practical terms.
- **R.D.Spend:** A unit increase in R&D Spend is associated with an increase in profit of approximately 69.79 units, and this relationship is highly significant ( $p < 0.001$ ).
- **Marketing.Spend:** Each additional unit spent on Marketing is associated with an increase of approximately 10.84 units in profit, and this predictor is also statistically significant ( $p = 0.012986$ ).

## Model Fit Statistics:

- **Residual Standard Error:** 19,420, which indicates the average deviation of the observed values from the predicted values.
- **Multiple R-squared:** 0.8063, meaning about 80.63% of the variance in profit is explained by the predictors in the model.
- **Adjusted R-squared:** 0.7952, which adjusts for the number of predictors, providing a more accurate measure of model fit.
- **F-statistic:** 72.84 with a p-value of  $3.351e-13$ , indicating that the model is statistically significant, and at least one of the predictors significantly explains the variance in profit.

## Conclusion

The backward elimination process has resulted in a model that retains only the significant predictors: R.D.Spend and Marketing.Spend. Both predictors are statistically significant, suggesting they are important for explaining the variation in profit. The model explains a substantial amount of variance in profit, similar to the forward selection model, indicating that either method effectively identifies meaningful predictors in this context.

### # Stepwise Regression:

#### # Stepwise Regression

```
stepwise_model <- step (mlr_model, direction = "both")  
summary(stepwise_model) # Stepwise Regression for summary
```

### OUT PUT :

```
> summary(stepwise_model)  
  
Call:  
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = train_set)  
  
Residuals:  
    Min       1Q   Median       3Q      Max   
-72818  -4076   4855   10932  17368   
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept)  3.162e+04  8.577e+03   3.686 0.000765 ***  
R.D.Spend     6.979e-01  9.243e-02   7.550 7.55e-09 ***  
Marketing.Spend 1.084e-01  4.142e-02   2.618 0.012986 *    
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 19420 on 35 degrees of freedom  
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.7952   
F-statistic: 72.84 on 2 and 35 DF,  p-value: 3.351e-13  
  
> |
```

### Explanation Stepwise Regression Model Summary :

#### Call:

- The stepwise regression method starts with no predictors and adds or removes predictors based on specific criteria (AIC, BIC). In this case, the final model includes R.D. Spend and Marketing.Spend.

#### Residuals

The residuals provide insights into the prediction errors:

- **Min:** -72,818
- **1Q:** -4,076
- **Median:** 4,855

- **3Q:** 10,932
- **Max:** 17,368

The residuals are similar to those from the backward elimination model, indicating a wide range and potential underpredictions.

**Coefficients :** The table below summarizes the estimated coefficients for each variable included in the stepwise regression model:

Predictor	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	31,620	8,577	3.686	0.000765 ***
R.D.Spend	0.6979	0.09243	7.550	7.55e-09 ***
Marketing.Spend	0.1084	0.04142	2.618	0.012986 *

#### Interpretation of Coefficients:

- **(Intercept):** Represents the estimated profit when all predictors are zero, which is not meaningful in practical terms.
- **R.D.Spend:** A unit increase in R&D Spend is associated with an increase in profit of approximately 69.79 units, and this relationship is highly significant ( $p < 0.001$ ).
- **Marketing.Spend:** Each additional unit spent on Marketing is associated with an increase of approximately 10.84 units in profit, and this predictor is also statistically significant ( $p = 0.012986$ ).

#### Model Fit Statistics:

- **Residual Standard Error:** 19,420, indicating the average deviation of the observed values from the predicted values.
- **Multiple R-squared:** 0.8063, meaning about 80.63% of the variance in profit is explained by the predictors in the model.
- **Adjusted R-squared:** 0.7952, which adjusts for the number of predictors, providing a more accurate measure of model fit.
- **F-statistic:** 72.84 with a p-value of  $3.351e-13$ , indicating that the model is statistically significant, and at least one of the predictors significantly explains the variance in profit.

#### Conclusion

##### Comments on Step wise Regression Model:

The stepwise regression process has resulted in a model that retains only the significant predictors: R.D.Spend and Marketing.Spend. Both predictors are statistically significant, suggesting they are important for explaining the variation in profit. The model explains a substantial amount of variance in profit, consistent with the findings from the backward



elimination model. Overall, this suggests that both methods effectively identify meaningful predictors in this context

**iv. How do the regression models differ when you use exhaustive search, forward elimination, backward elimination and stepwise regression to reduce the number of predictors? Comment on your findings. Include the best regression model based on their Cp values (and use other measures if needed). =Create a table that shows the list of predictors that are included in the model, similar to the one we have in the PowerPoint slides=**

**Explanation for Different model Comparison:** In my analysis, I compared four different regression models: exhaustive search, forward elimination, backward elimination, and stepwise regression. Each method used a distinct approach to select the most relevant predictors for predicting the profit in my dataset.

### **My Comparison of MLR Models**

- **Exhaustive Search:** This method evaluated all possible combinations of predictors and chose the best model based on Cp values. In my case, it included all predictors (R.D.Spend, Marketing.Spend, Administration, and State). It produced the best Cp value of 4.21, meaning it had a better balance between model complexity and fit compared to the other methods.
- **Forward Elimination:** Starting with no predictors, this method added R.D.Spend and Marketing.Spend as the most meaningful predictors. It excluded Administration and the State variables. This model had a Cp value of 4.45 and the same Adjusted  $R^2$  as the exhaustive search model, showing it explained the same amount of variance with fewer predictors.
- **Backward Elimination:** In this method, I started with all predictors and eliminated the least significant ones, eventually resulting in a model identical to the forward elimination method. The final predictors were R.D.Spend and Marketing.Spend, with an identical Cp value of 4.45.
- **Stepwise Regression:** Combining both forward and backward selection methods, this approach also resulted in the same model as forward and backward elimination, with a Cp value of 4.45 and the same Adjusted  $R^2$ .

### **Model Comparison Table of Results with list of Predictors:**

Model	R.D.Spend	Marketing.Spend	Administration	State Florida	State New York	Adjusted $R^2$	Cp Value
Exhaustive Search	Yes	Yes	Yes	Yes	Yes	0.795	4.21
Forward Elimination	Yes	Yes	No	No	No	0.795	4.45
Backward Elimination	Yes	Yes	No	No	No	0.795	4.45
Stepwise Regression	Yes	Yes	No	No	No	0.795	4.45

## comments:

### My Conclusion for Best Regression Model Best Cp Value :

I found that the **exhaustive search model gave the best Cp value**, indicating it might be the most accurate in balancing complexity and fit. However, the simpler models derived from forward, backward, and stepwise regression provided the same **Adjusted R<sup>2</sup>** with fewer predictors (R.D.Spend and Marketing.Spend). Based on this, I might prefer the simpler models for ease of interpretation and reduced risk of overfitting, though the exhaustive model performs slightly better in terms of Cp.

## Problem 2. Interpretation of PCA.

After conducting PCA for the dataset excluding the categorical predictor, the following output is obtained:

Importance of components:			
	PC1	PC2	PC3
Standard deviation	1.325	1.010	0.4757
Proportion of Variance	0.585	0.340	0.0754
Cumulative Proportion	0.585	0.925	1.0000
Briefly comment on these findings.			
Good luck!			

After conducting the Principal Component Analysis (PCA), I found that the variance in the dataset is captured by three principal components (PCs). Here's what I observed:

### 1. Standard Deviation

I noticed that PC1 has a standard deviation of 1.325, which means it explains the most variance in the dataset. PC2, with a standard deviation of 1.010, captures less variance than PC1. Finally, PC3 has the smallest standard deviation of 0.4757, meaning it explains the least variance among the three components.

- **PC1** has a standard deviation of 1.325, which tells me it explains the most variance in the dataset.
- **PC2** has a lower standard deviation of 1.010, meaning it captures less variance than PC1.
- **PC3** has the smallest standard deviation of 0.4757, meaning it explains the least variance among the three components.

### 2. Proportion of Variance

When I look at the proportion of variance, I see that PC1 explains 58.5% of the variance, meaning it captures the majority of the variance in the data. PC2 explains 34.0%, which means that together with PC1, these two components capture a significant portion of the variance. PC3 only explains 7.54%, which is quite small.

- **PC1** explains **58.5%** of the variance, meaning it captures most of the variance in the data.

- **PC2** explains **34.0%**, meaning together with PC1, these two components explain a significant portion of the variance.
- **PC3** explains only **7.54%**, which is relatively small and less impactful.

### 3. Cumulative Proportion

I noticed that PC1 alone explains 58.5% of the total variance. When I combine PC1 and PC2, they explain 92.5%, so I can retain most of the dataset's information with just these two components. Finally, by adding PC3, I capture 100% of the variance, but it doesn't add much more value.

- **PC1** alone explains **58.5%** of the total variance in the dataset.
- **\*\*PC1 and PC2** together explain **92.5%**, meaning I can retain most of the information in the dataset with just these two components.
- **PC1, PC2, and PC3** together explain **100%** of the variance, capturing all the variance in the data.

### Summary of My Findings/ Comments:

I noticed that the first two components (**PC1 and PC2**) explain **92.5%** of the variance, which means they provide a solid summary of the dataset. By using these two components, I can significantly reduce the dimensionality of the data while retaining most of the variance.

**PC3**, which only explains **7.54%**, doesn't add much value and can likely be excluded without losing too much information.

In conclusion, I found that **PC1 and PC2** are the most important components to focus on for reducing complexity while keeping most of the dataset's information.

**Citation:** <https://chatgpt.com/share/670d9080-c4ac-8004-88fa-a09b19101b42>