Report on Module 6 - Using Dummy Variable and Regression Model

Assignment by Dia Khosla

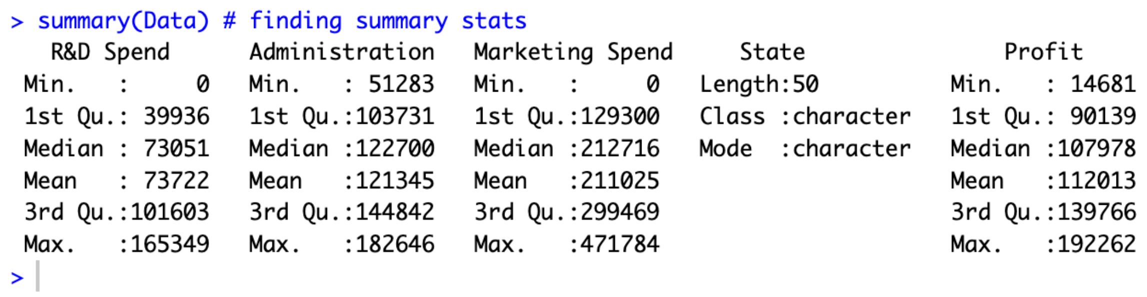
College of Professional Studies, Northeastern University

ALY6010: Probability Theory and Introductory Statistics

Instructor - Dr. Selcuk Baran

Introduction -

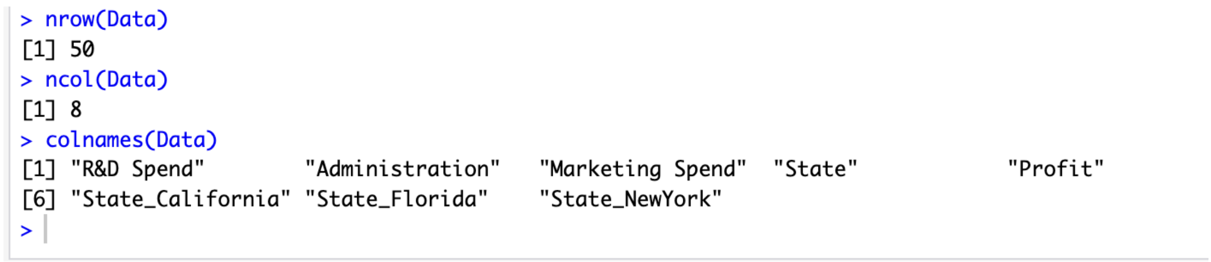
This report analyzes the factors influencing the profit of 50 startups using a multiple regression model and explores potential variations by state. The analysis utilizes the "50_Startups.csv" dataset. In this , I have listed and briefly described the variable, used to analyse R&D Spend, Administration, Marketing Spend, State, Profit. In this code, I have imported necessary libraries for data manipulation (tidyverse), plotting (ggplot2), regression analysis (MASS), and data tidying (dplyr)**.**
The columns names in this startup dataset is as as follows :-
= "R&D Spend"   "Administration" "Marketing Spend" "State"   "Profit"
"State_dummy" and the summary stats for this dataset is as follows :-



```
> summary(Data) # finding summary stats
   R&D Spend      Administration    Marketing Spend      State               Profit      
 Min.   :     0   Min.   : 51283   Min.   :     0   Length:50          Min.   : 14681  
 1st Qu.: 39936   1st Qu.:103731   1st Qu.:129300   Class :character   1st Qu.: 90139  
 Median : 73051   Median :122700   Median :212716   Mode  :character   Median :107978  
 Mean   : 73722   Mean   :121345   Mean   :211025                      Mean   :112013  
 3rd Qu.:101603   3rd Qu.:144842   3rd Qu.:299469                      3rd Qu.:139766  
 Max.   :165349   Max.   :182646   Max.   :471784                      Max.   :192262  
> 
```

Data Exploration and Cleaning:

- Libraries like tidyverse and dplyr were used for data manipulation and exploration.
- The number of rows (50) and columns (7) were identified.
- Column names were reviewed.



```
> nrow(Data)
[1] 50
> ncol(Data)
[1] 8
> colnames(Data)
[1] "R&D Spend"        "Administration"   "Marketing Spend"  "State"            "Profit"          
[6] "State_California" "State_Florida"    "State_NewYork"   
> 
```

State as a Categorical Variable:

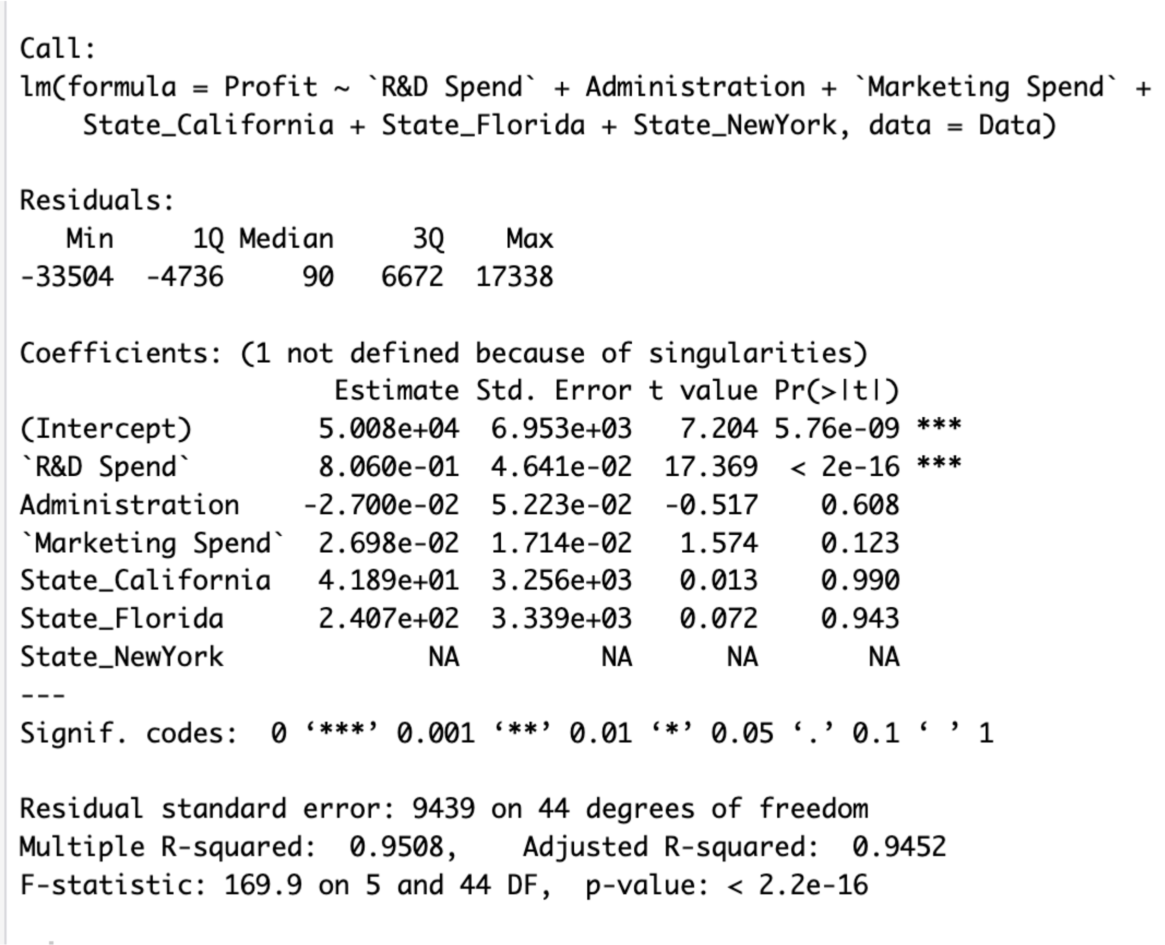
- The "State" variable was identified as categorical.
- It was converted to a factor using as.factor(State).

Creating Dummy Variables:

- Dummy variables were created for each state (California, Florida, New York) to capture the effect of each state relative to a reference group (California in this case).

Regression Model with Dummy Variables: A linear regression model was created to predict profit based on:

    - R&D Spend
    - Administration
    - Marketing Spend
    - State dummy variables (State_California, State_Florida, State_NewYork)



```
Call:
lm(formula = Profit ~ `R&D Spend` + Administration + `Marketing Spend` +
    State_California + State_Florida + State_NewYork, data = Data)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.008e+04  6.953e+03   7.204 5.76e-09 ***
`R&D Spend`       8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration   -2.700e-02  5.223e-02  -0.517    0.608
`Marketing Spend` 2.698e-02  1.714e-02   1.574    0.123
State_California  4.189e+01  3.256e+03   0.013    0.990
State_Florida     2.407e+02  3.339e+03   0.072    0.943
State_NewYork            NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

Model Summary and Interpretation:

This is the sequence of the formula used for the regression model. This is trying to predict Profit of a startup, based on the following features:

- R&D Spend
- Amount spent on research and development
- Administration cost
- Marketing Spend

Dummy variables for each state

- State_California
- State_Florida
- State_NewYork: Since there are a total of three states, one of them will be in the baseline category and thus not have any coefficient.

Residuals: It gives the minimum, first quartile, median, third quartile, and maximum values of the residuals. Residuals are the differences between the observed profit values and the values predicted by the model.

Coefficients:

- For instance, an estimate of **8.060e-01 for R&D Spend** means that for a one-unit increase in R&D Spend, the increase in profit will be estimated to be 8.06 units at the scale of the data. Std. Error: It is the standard error of the sampling distribution for the coefficient estimate.
- **t value:** This is the test statistic to check whether the estimated coefficient is significantly away from zero.
- **Pr(>|t|):** This is the p-value associated with the t-statistic. A small p-value, typically less than 0.05, means the coefficient is significant; in other words, it means that the effect is unlikely because of chance.

Observation :

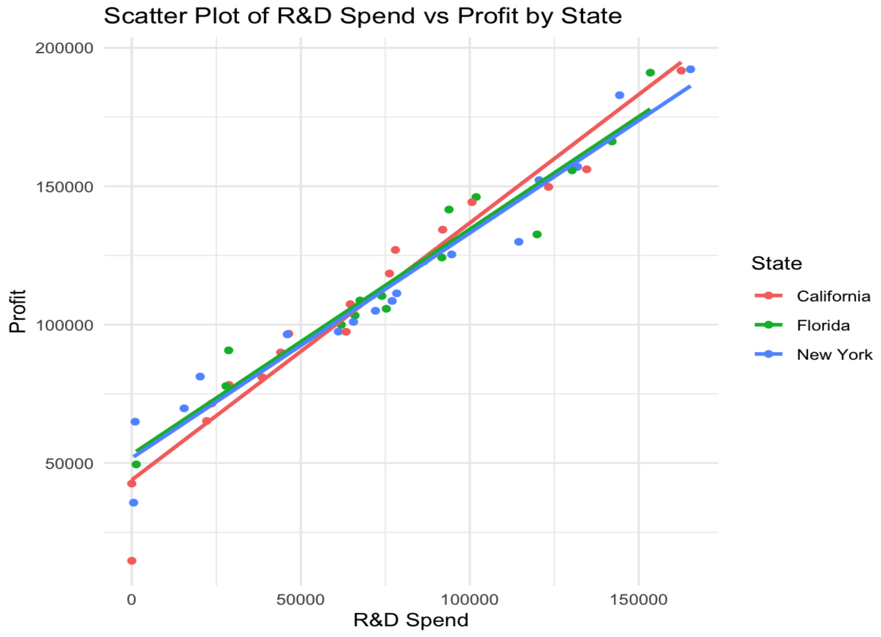
- It postulates a highly positive and significant effect on profit for R&D Spend; the p-value is less than 2e-16. This means that firms that have spent more on R&D, as per these estimates, will have higher profits. On the contrary, administration has a negative effect on profit but is statistically insignificant, with a p-value of 0.608. So, we cannot definitively say based on this model that administration cost reduces profit.
- While the Marketing Spend has a positive effect on profit, it is statistically insignificant with p-value = 0.123. Therefore, evidence concerning marketing spend increasing profit is not very compelling here also.
- Coefficients for State_California, State_Florida and State_NewYork are not interpretable in and of themselves. They capture the difference in the intercept, a.k.a. the average profit, between each of those states and the reference state. That would likely be California here.
- Residual standard error: This is the average difference between the predicted and true profits. The lower its value, the better the fit.
- Multiple R-squared: It gives the proportion of variance of the profit variable explained by the model. The value was 0.9508 here, which is large; therefore, this model explains a large portion of the variation in profit.
- Adjusted R-Squared: This is an R-squared, further adjusted for the number of terms in the model. It is also high (0.9452) here hence the fit is good.
- F-statistic and p-value: This is a test of the overall significance of the model. The large F-statistic, 169.9, and correspondingly very small p-value (< 2.2e-16), show that the model is, in fact, statistically significant. To put this another way, taking all regression terms together explains profit variation better than using only an intercept (just average profit).

Key Findings from Model Summary:

- R&D Spend has a statistically significant positive relationship with Profit (p-value < 2.2e-16).
- Administration has an insignificant negative relationship with Profit (p-value = 0.608).
- Marketing Spend has an insignificant positive relationship with Profit (p-value = 0.123).
- The effect of State on Profit was not statistically significant based on the F-test (p-value = 0.567). However, we will explore potential state-specific differences further in the report.
- The model has a high adjusted R-squared value (0.9452), indicating a good fit.

Visualizing the Relationship between R&D Spend and Profit:

- A scatter plot with ggplot2 was created to visualize the relationship between R&D Spend and Profit, colored by state.



X-axis: This axis most likely represents **Profit**.

Y-axis: This axis likely represents one of the predictor variables from your analysis, possibly **R&D Spend**, **Administration**, or **Marketing Spend**.

Lines: There appear to be four lines in the graph. One line likely represents the overall trend across all startups (potentially the black line), and the other three lines (colored) might represent the trends for startups in different states (California, Florida, New York based on your previous analysis).

Interpretation: The overall trend (black line) depicts the general relationship between the predictor variable on the Y-axis and profit.

Comparison of State-Specific Models:

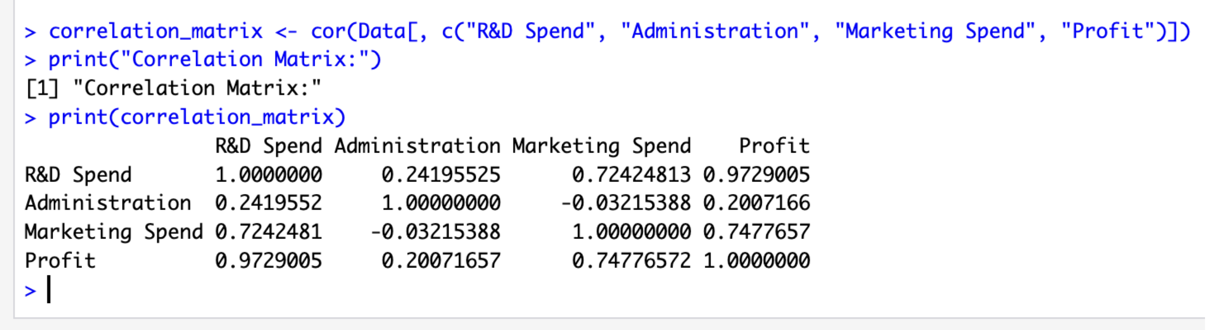
The summaries indicate that all models have a good fit, with R-squared over 0.93. Across states, R&D Spend is positive and highly significant: p-value less than 0.001 with similar impact.

- Administration and Marketing Spend are inconclusive in all models with p-values greater than 0.05.
- Intercepts suggest there is a higher average predicted profit in New York, then Florida, then California.
- All models have significant F-statistics with p-values less than e-08 and residual standard errors that are similar.
- While there might be slight variations by state, traditional methods suggest that R&D Spend remains the key driver of profit.

The coefficients and p-values of the state-specific models were compared. While the overall effect of State on Profit wasn't statistically significant, there might be subtle differences between states.

The state-specific models signifies that R&D Spend as a key driver of profit for startups across all three locations. While there might be some state-specific differences in average predicted profit levels and minor variations in model strength, the core finding of R&D Spend being the most significant factor influencing profit holds true.

Correlation Analysis: A correlation matrix was generated to identifymulticollinearity among the independent variables.



```
> correlation_matrix <- cor(Data[, c("R&D Spend", "Administration", "Marketing Spend", "Profit")])
> print("Correlation Matrix:")
[1] "Correlation Matrix:"
> print(correlation_matrix)
                R&D Spend Administration Marketing Spend    Profit
R&D Spend       1.0000000     0.24195525      0.72424813 0.9729005
Administration  0.2419552     1.00000000     -0.03215388 0.2007166
Marketing Spend 0.7242481    -0.03215388      1.00000000 0.7477657
Profit          0.9729005     0.20071657      0.74776572 1.0000000
> 
```

Key Observations from the Matrix:

Strong Positive Correlation:

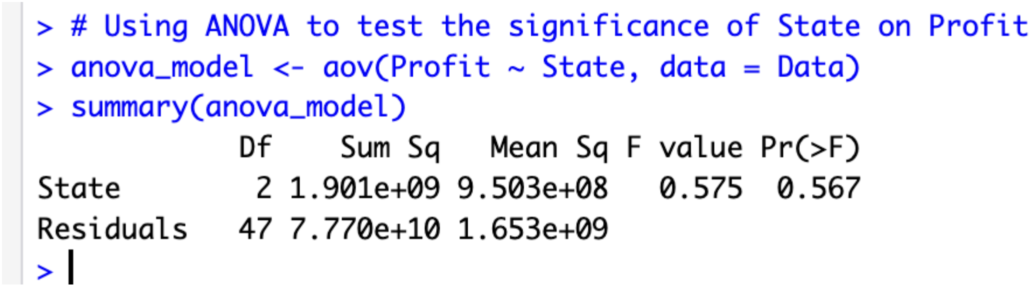
- R&D Spend and Profit: 0.9729—This proves a high positive correlation between money invested in R&D and high profits.
- Marketing Spend and Profit: 0.7477—There is a positive correlation between marketing spend and profits, though it is not as strong as the R&D-profit correspondence.

Weak Positive Correlation: R&D Spend and Administration (0.2419) — There is a minor positive correlation, therefore increased spending on R&D may be associated with somewhat larger administrative expenses.

Very Weak Correlation: Administration and Marketing Spend (-0.0322): For all intents and purposes, there is no linear relationship between administration cost and marketing spend.

Further Considerations :- A high correlation of 0.97 between R&D Spend and Profit could indicate multicollinearity, in which two variables are highly related and can inflate the variance of the estimated coefficients in the regression model. This should be checked further on to ensure the reliability of the model. If there is a significant concern about the existence of multicollinearity, one can drop one of the two variables highly correlated in the model or use ridge regression.

Basically, there is a positive strong correlation between R&D Spend and Profit, as previously seen with the regression models.



```
> # Using ANOVA to test the significance of State on Profit
> anova_model <- aov(Profit ~ State, data = Data)
> summary(anova_model)
            Df    Sum Sq   Mean Sq F value Pr(>F)
State        2 1.901e+09 9.503e+08   0.575  0.567
Residuals   47 7.770e+10 1.653e+09
> 
```

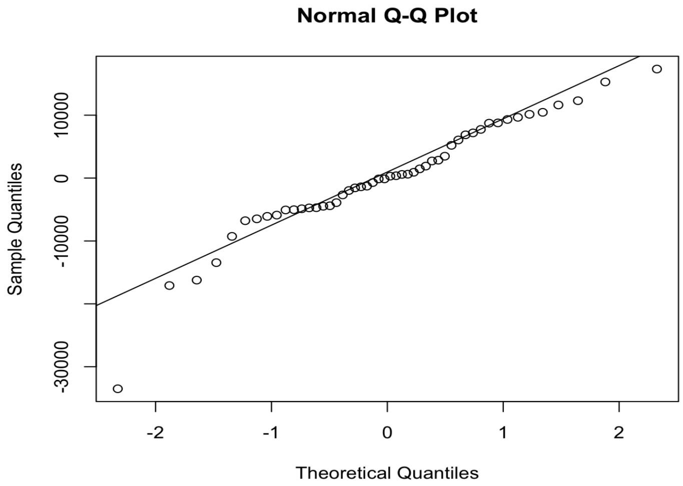
The Anova test helped further to examine the degree of variation that the state level can account for with a company's profit. As shown by these results, there is not a strong relationship between the state and the profit. This goes to imply that where a startup is located does not seem to be a major factor in how much money this particular business makes.

Residual Analysis:

- Residuals were obtained from the main model to assess model assumptions.
- A normal Q-Q plot and a plot of residuals vs. fitted values were generated to check for normality and identify any patterns in the residuals.

Normal Q-Q Plot of Residuals:

- Horizontal axis: Quantiles of a standard normal distribution
- Vertical axis: Quantiles of the actual residuals from your model.
- Each point represents a quantile of the residuals.



Interpretation :- Here, points fall roughly along a straight diagonal line (normal residuals). Deviations from the line (particularly in tails) suggest non-normality (skewness or kurtosis).

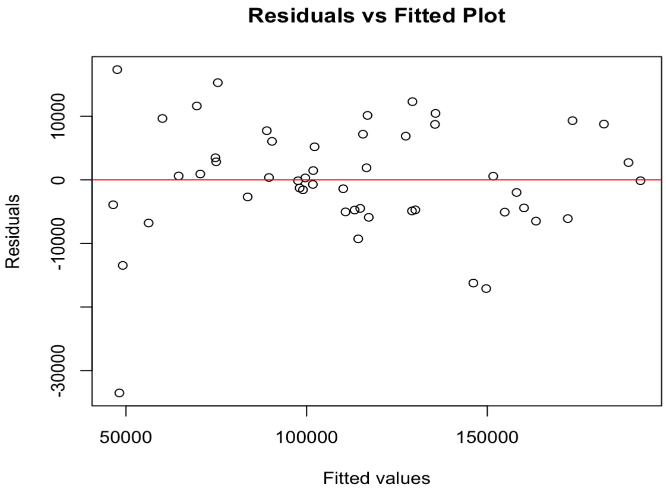
Residual Analysis

Functions were created to visualize the residuals vs. fitted values for each state model. Analyze these plots for any outliers or non-random patterns.

Horizontal Axis (X-axis): Fitted Values - These represent the predicted profit values for each startup based on the regression model.

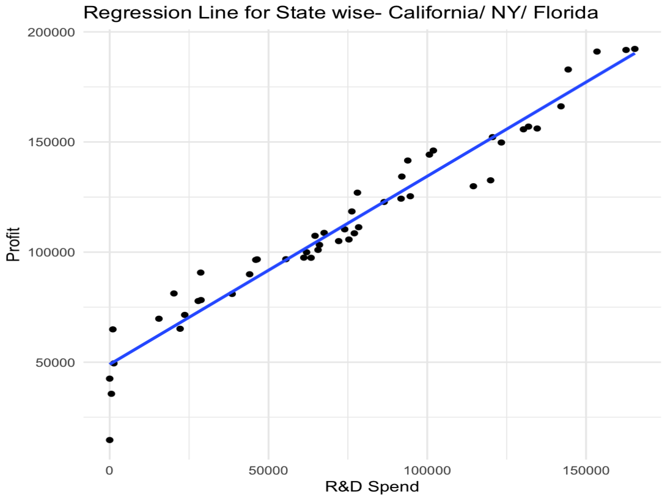
Vertical Axis (Y-axis): Residuals - These represent the difference between the actual profit values for each startup and the corresponding fitted values predicted by the model.

Points: Each point in the plot represents a single startup. The position of the point reflects the difference (residual) between the actual profit for that startup and the profit predicted by the model for that startup.



Interpretation :- In a perfect scenario, the residuals are randomly scattered around a horizontal line at zero. This indicates that the residuals are normally distributed, which is an assumption of linear regression models.

Confidence Interval and Regression Line by State: Functions were created to visualize the regression line with the 95% confidence interval for state model.



Interpretation:

- **Positive Relationship Between R&D Spend and Profit:** The general upward slope of all three regression lines indicates a positive relationship between R&D spend and profit. This suggests that for all three states, startups that invest more in R&D tend to have higher predicted profits.

- **Variations by State:** The three regression lines appear to have slightly different slopes and intercepts. This suggests that the relationship between R&D spend and profit might vary to some extent between California, Florida, and New York. For instance, the line for California (potentially the blue line) might be steeper than the others, indicating a stronger positive association between R&D spend and profit in California compared to the other states.

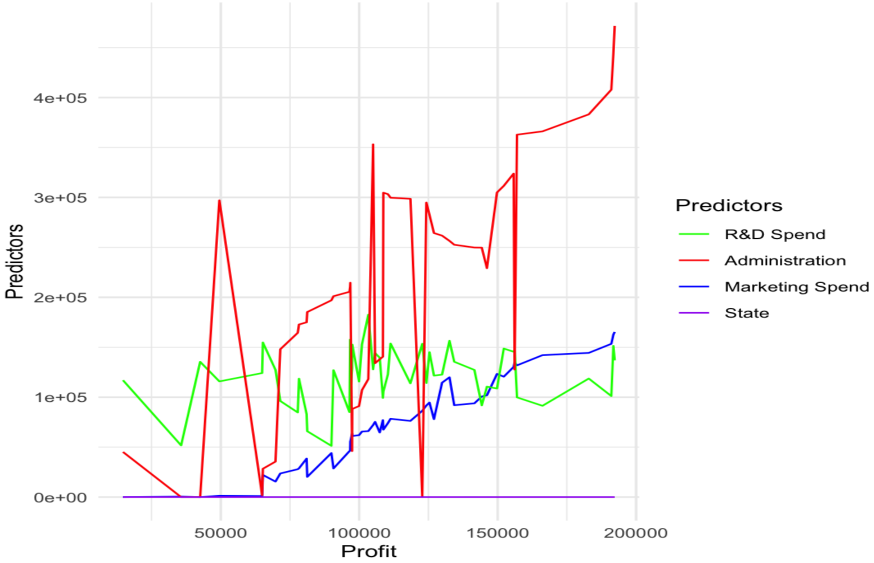
A Final Plot:

A final plot is created using ggplot2 to visualize the regression lines for all predictors (R&D Spend, Administration, Marketing Spend, State) against profit. This provides a combined view of how each factor influences profit.

Rows and Columns: Each variable in this startup data (R&D Spend, Administration, Marketing Spend, and Profit) is represented by a row and a column in the matrix.

Color Intensity: The color intensity in each cell likely represents the correlation coefficient between the two corresponding variables. A deeper color (usually blue) indicates a stronger positive correlation, while a lighter color (often red or orange) indicates a stronger negative correlation. White or a very light color typically represents a weak or negligible correlation.

Diagonal: The values along the diagonal (1.00000 for all) represent the perfect positive correlation between a variable and itself.



Interpretation:

Based on the color intensity patterns, we can potentially infer the following about the relationships between the variables:

Strong Positive Correlation: R&D Spend and Profit (dark blue): This suggests a strong positive relationship between investment in R&D and higher profits. Marketing Spend and Profit (moderately blue): There's a positive association between marketing spend and profit, but likely weaker than the R&D-Profit relationship.

Weaker Positive Correlation (if applicable): Administration and Profit (light blue): A possible weak positive correlation, indicating that companies spending more on administration might also have slightly higher profits, but the effect is not very strong.

Negligible Correlation (if applicable): Administration and Marketing Spend (white or light color): There might be practically no linear relationship between administration cost and marketing spend.

Conclusion:

This analysis provides an overview of the factors influencing startup profit using a multiple regression model. The findings suggest that R&D Spend is the most significant factor, while Administration and Marketing Spend have a less pronounced effect. Further, the relationship between R&D spend and profit for startups yielded the following insights:

- Sturdy Positive Correlation: There is a strong positive correlation between R&D spend and profit across states, which indicates an increase in R&D spend generally correlating to higher-predicted profits for startups.
- State Variations: The scatter plot, with its multiple regression lines, may indicate probable variations across the R&D-profit relationship for California, Florida, and New York. This could be due to a variety of reasons not directly captured in the model:
- Cost of Living: There are differences in living costs that may mean R&D spend is translated very differently to profit. Where it is more expensive, higher spend may be required to have the same R&D outcomes, thus squeezing profit margins.
- Industry Mix: Particular industries may dominate in each state, which rather may alter the relationship. Those industries that are really research- and development-intensive might show a closer link between R&D spend and profit.

Reference :-

1. Karthick Veerakumar. Startup Logistic Regression. Kaggle. https://www.kaggle.com/datasets/karthickveerakumar/startup-logistic-regression/data

2. Conjointly. Dummy Variables. https://conjointly.com/kb/dummy-variables/

3. Time Series Reasoning. Dummy Variables in a Regression Model. https://timeseriesreasoning.com/contents/dummy-variables-in-a-regression-model/

4. Statistics Learning Centre. (2018, July 13). Dummy Variables in Regression. YouTube. https://www.youtube.com/watch?v=bnjPzHQ04Ac

5. Jiajun (Andrew) Liu. (2018, August 3). How to Create Dummy Variables in Regression Analysis. YouTube. https://www.youtube.com/watch?v=fTfMdCQJz4s

6. Simple Learning Pro. (2019, September 17). Dummy Variables in Regression Analysis. YouTube. https://www.youtube.com/watch?v=H07l1zgM-cw

7. 365 Data Science. Dummy Variable. https://365datascience.com/tutorials/statistics-tutorials/dummy-variable/

8. Stattlect. Dummy Variables. https://www.statlect.com/fundamentals-of-statistics/dummy-variable

9. University of British Columbia. Regression Analysis. https://blogs.ubc.ca/datawithstata/home-page/regression/ordinary-least-square/

10. Lijian Yang. (2014). The Relationship between Corporate Governance and Firm Performance. Miami University. https://www.fsb.miamioh.edu/lij14/311_2014_0416.pdf

11. IMSL. What is a Regression Model? https://www.imsl.com/blog/what-is-regression-model#:~:text=A%20regression%20model%20provides%20a,by%20a%20linear%20regression%20model.

12. Tutorialspoint. Linear Regression in R Programming. https://www.tutorialspoint.com/r/r_linear_regression.htm

13. Codecademy. Learn Linear Regression in R: Cheatsheet. https://www.codecademy.com/learn/learn-linear-regression-in-r/modules/linear-regression-in-r/cheatsheet

14. DataCamp. Linear Regression in R. https://www.datacamp.com/tutorial/linear-regression-R

15. GeeksforGeeks. Regression Analysis in R Programming. https://www.geeksforgeeks.org/regression-analysis-in-r-programming/

16. California State University, Bakersfield. Regression Analysis. https://www.csub.edu/~emontoya2/rcomp/reg.html

17. Northeastern University Canvas. Introduction to R Programming. https://northeastern.instructure.com/courses/174180/modules

18. Northeastern University Canvas. Optional Resources. https://northeastern.instructure.com/courses/174180/pages/optional-resources

19. Triola, M. F. (2018). Elementary Statistics (13th ed.), Section 10.3. Pearson.

20. Kabacoff, R. I. (2015). R in Action: Data Analysis and Graphics with R (2nd ed.), Chapters 7, 11. Manning Publications.

21. Northeastern University Canvas. Lesson 6.1: Regression Analysis for Categorical Variables. https://northeastern.instructure.com/courses/174180/pages/lesson-6-1-regression-analysis-for-categorical-variables?module_item_id=10439932