# PROFIT PREDICTION OF START-UPS

BAN 210 – Predictive Analytics
Professor – Ji Qi
Date – 15th Nov 2022


Group 09

- Srijith Leelakrishnan
- Laxmikant Mukkawar
- Henadera A Pulathis Maduranga Perera
- Ramya Tallapudi
- Keerthana Shreepal Urs

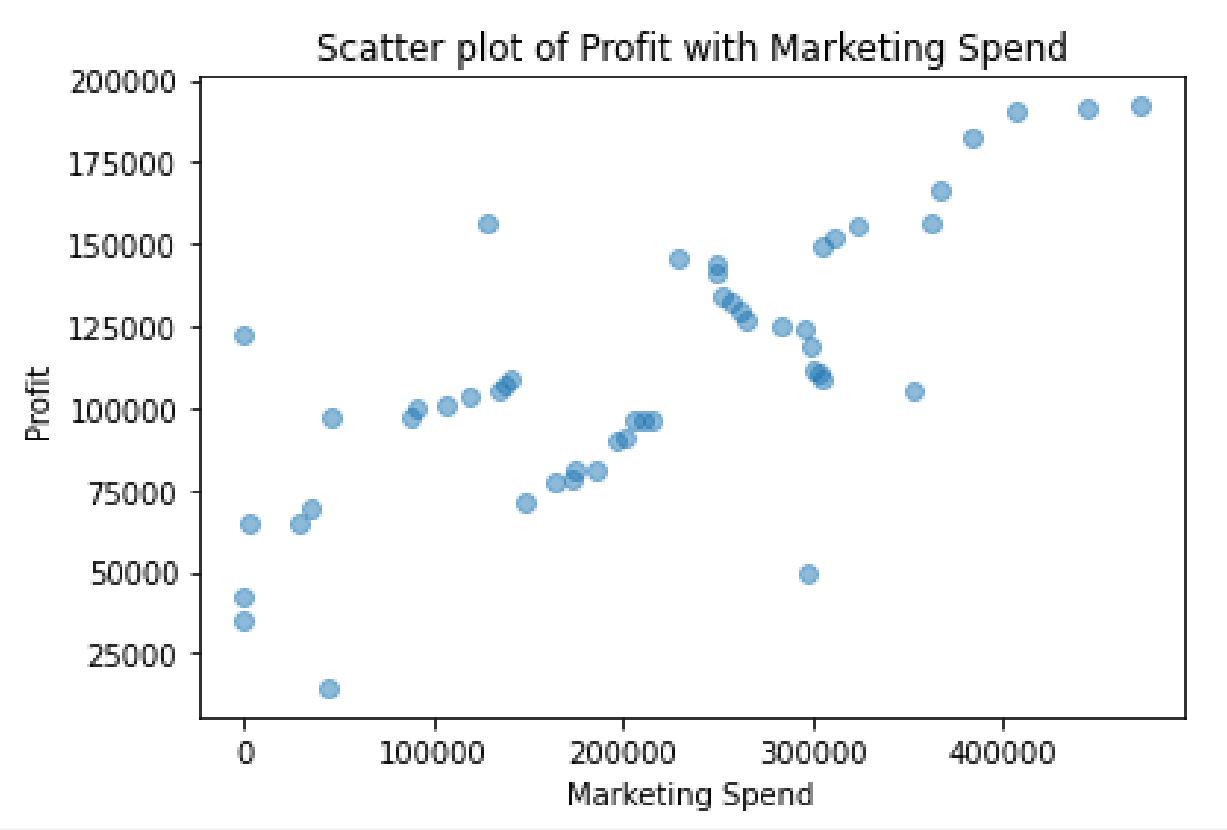
# Objective

To predict the profit made by startups based on expenses incurred and the state where they operate

# Introduction

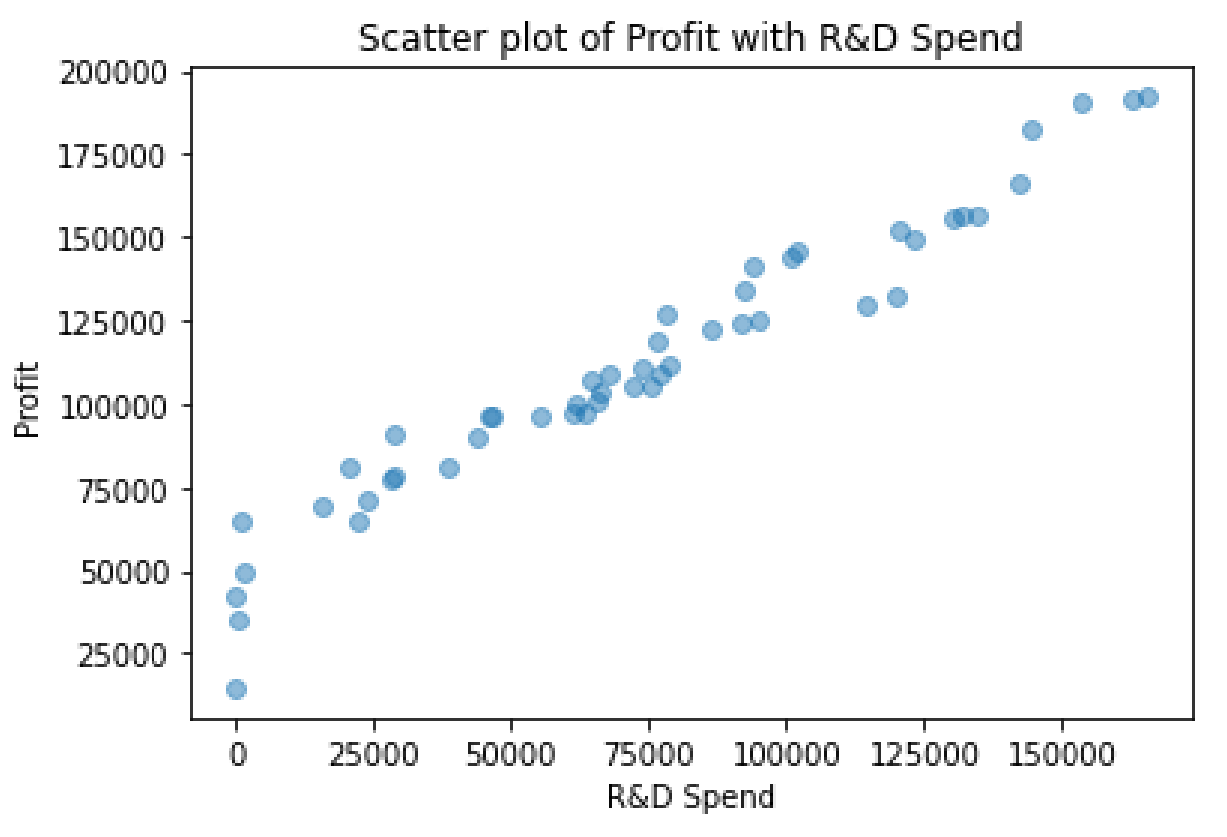
- The dataset contains information about expenses (R&D, Administration and Marketing), state of operation and profits of startups
- Profit is the dependent variable while others are independent variables
- A portion of this data will be used as a training set and the remaining as a test set through random selection
- A multiple linear regression model is used as the prediction model
- Performance of the model is tested against the test data set
- Use of scatter plots to check for possible correlation between dependent and independent variables
- Heteroscedasticity check to validate assumption of linear regression model
- Multi collinearity checked to determine correlation within independent variables

# Effect of marketing spend on profit



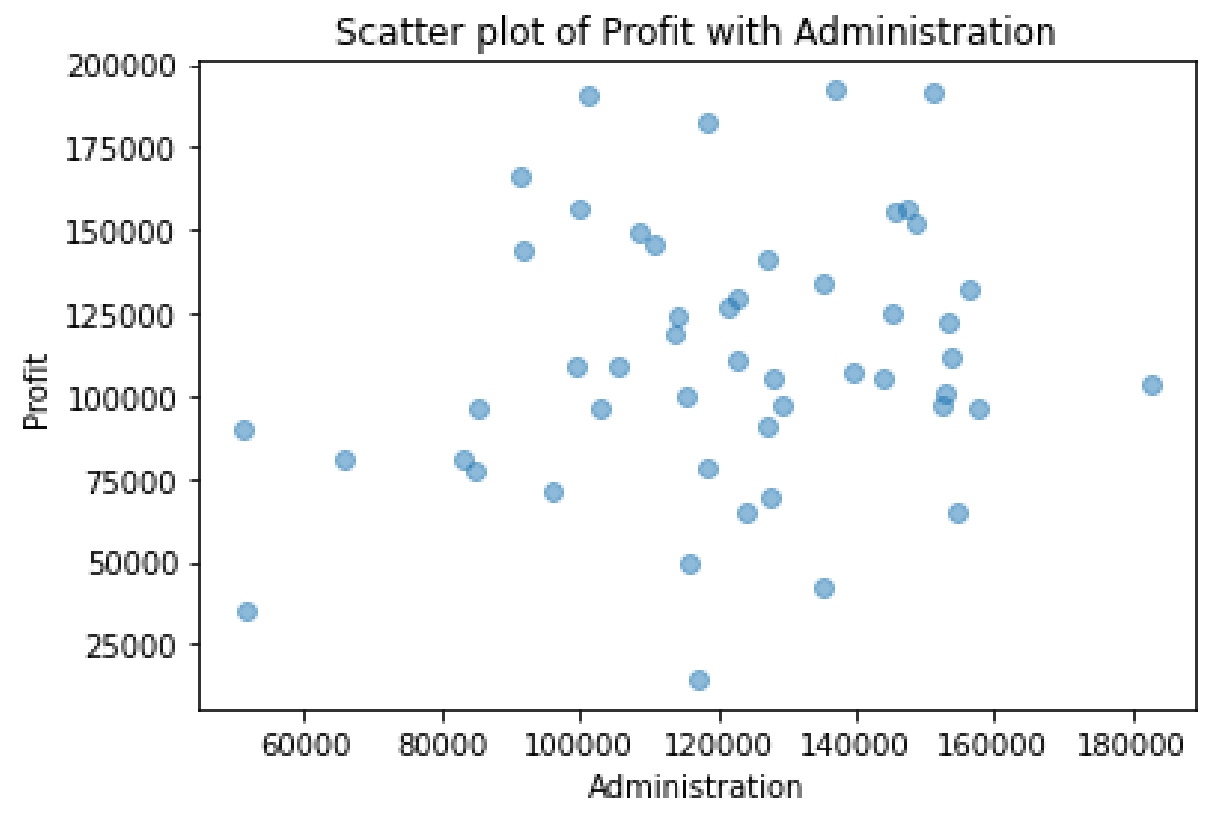

- Strong positive correlation visualized between marketing spend and profit
- Higher the spend on Marketing, higher the profit.
- Goes on to show that more marketing leads to better sales

# Effect of R&D spend on profit



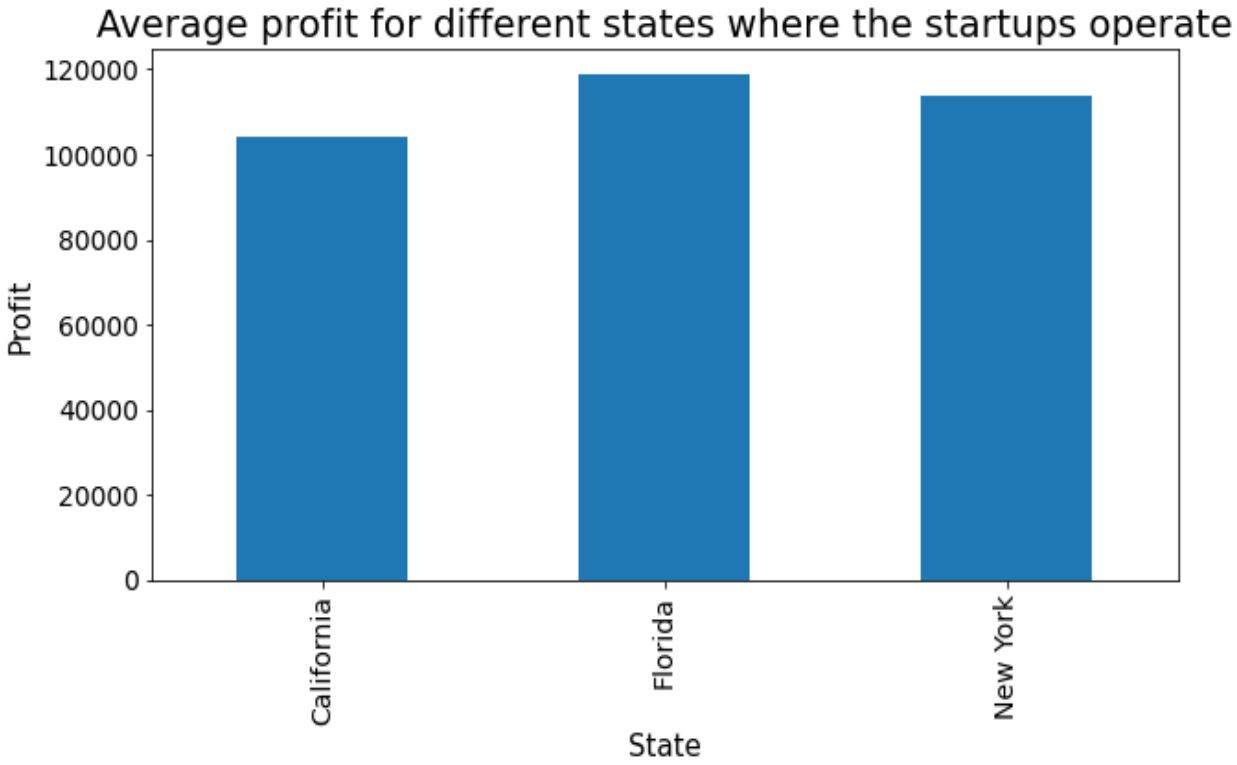

- Stronger positive correlation visualized between R&D spend and profit
- Higher the spend on R&D, higher the profits
- By spending more on research and development, startups strive to improve customer base and profits

# Effect of Administration spend on profit



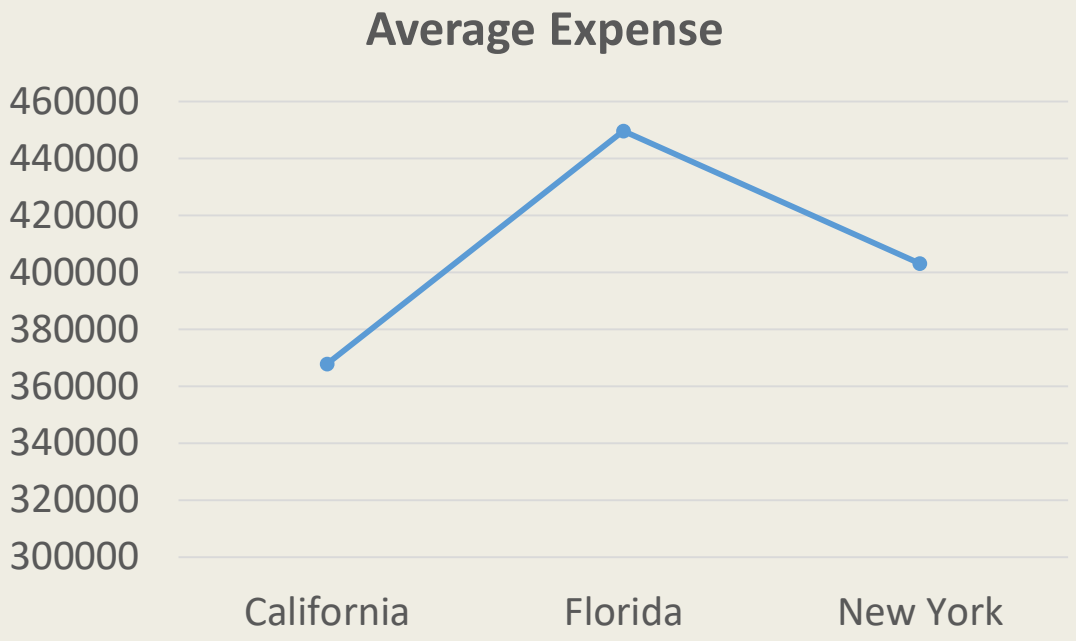

- There is no correlation visualized between Admin spend and profit
- Admin costs can be reduced without affecting profit and re-invested in either R&D or marketing

# Effect of State on profit



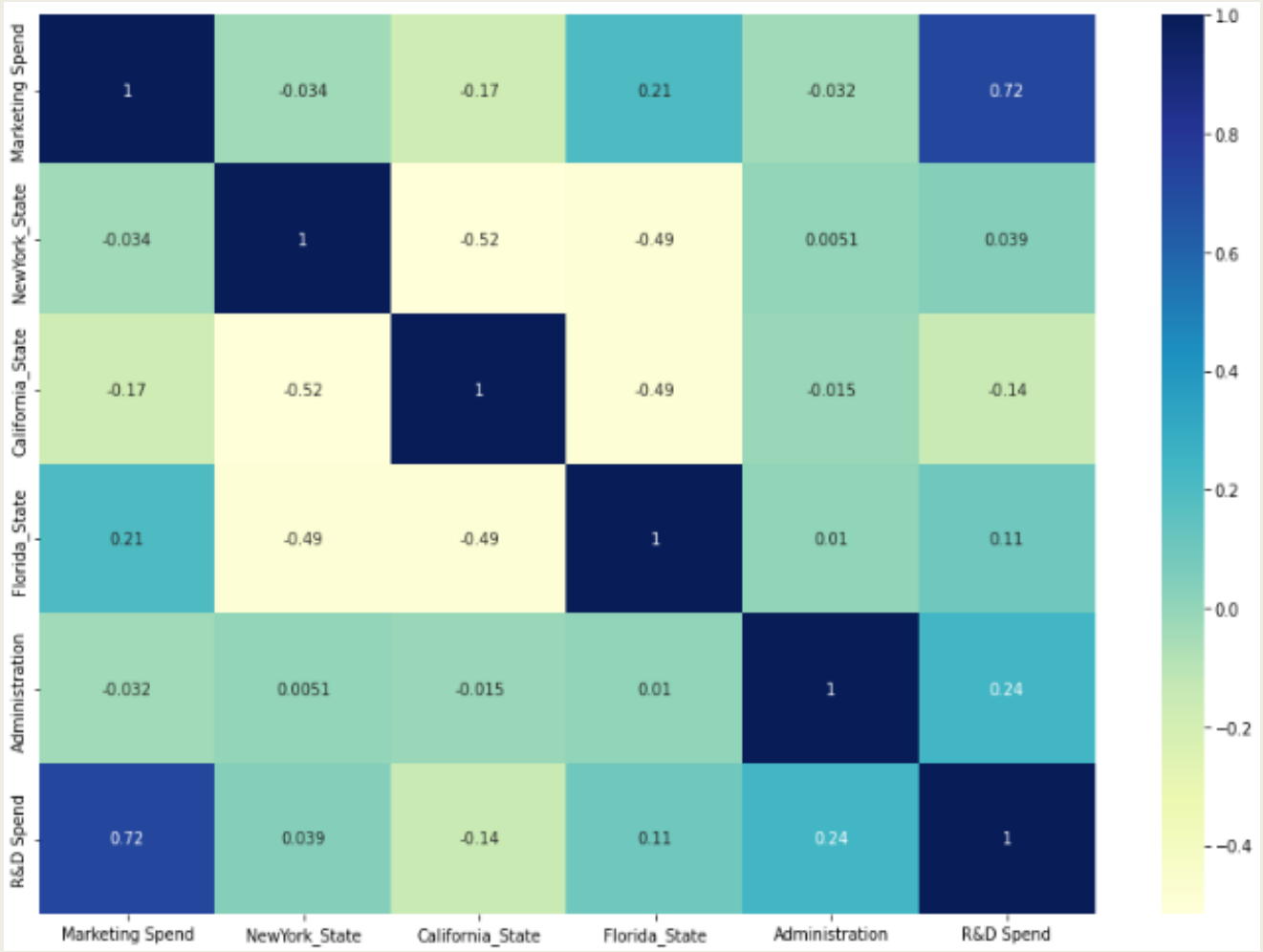

- Florida shows the highest profits
- The other two states are close behind
- There is equal distribution of data among the 3 states

# Cost analysis between states



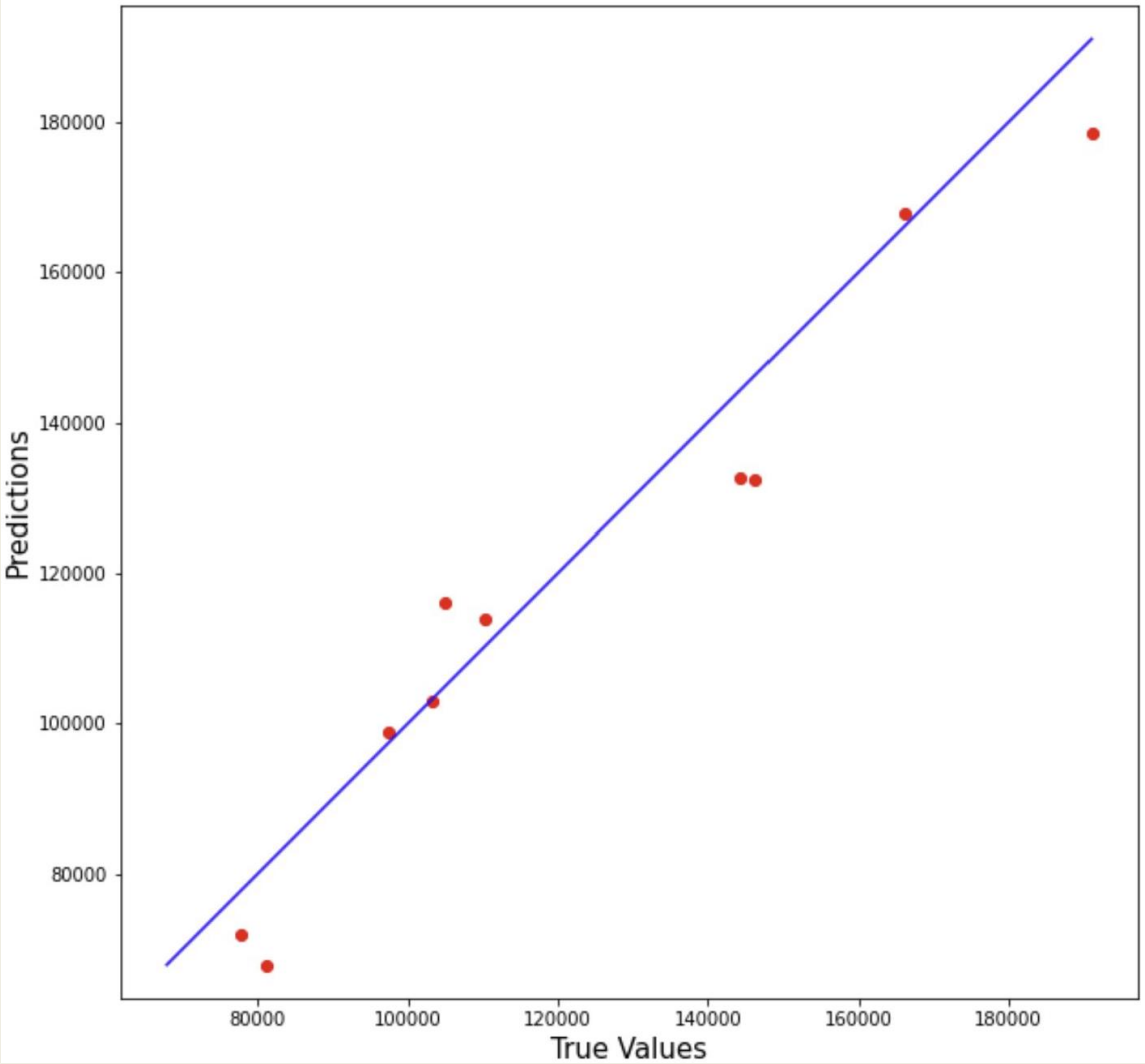

- Profit is the highest in Florida and the average costs are the lowest in California which is 22% lesser than Florida and 10% lesser than New York

# Multi-Collinearity




- R&D and Marketing costs are strongly correlated, indicating that if one cost increases, the other increases and vice versa
- This an area that requires attention and further analysis to improve profits.

# Regression Model

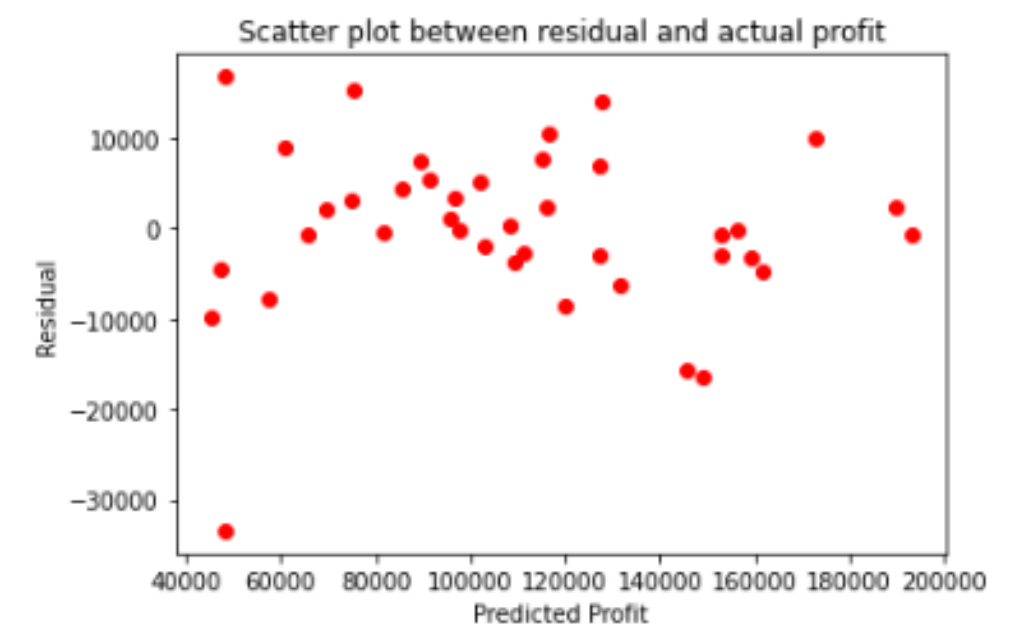



- Regression model indicates a close match between actuals and predictions
- R-square score of actual test results and predicted test results is 0.935 which means that 93.5% of the observed variation can be explained by the model.

# Regression Model - Output on Test set

| Marketing | Administration | R&D | California | Florida | New York | Actual profit | Predicted profit |
|---|---|---|---|---|---|---|---|
| 182,645.56 | 66,051.52 | 118,148.20 | 1 | 0 | 0 | 103,282.38 | 103,015.20 |
| 91,790.61 | 100,671.96 | 249,744.55 | 0 | 0 | 1 | 144,259.40 | 132,582.28 |
| 110,594.11 | 101,913.08 | 229,160.95 | 1 | 0 | 0 | 146,121.95 | 132,447.74 |
| 84,710.77 | 27,892.92 | 164,470.71 | 1 | 0 | 0 | 77,798.83 | 71,976.10 |
| 101,145.55 | 153,441.51 | 407,934.54 | 1 | 0 | 0 | 191,050.39 | 178,537.48 |
| 127,864.55 | 72,107.60 | 353,183.81 | 0 | 1 | 0 | 105,008.31 | 116,161.24 |
| 65,947.93 | 20,229.59 | 185,265.10 | 0 | 1 | 0 | 81,229.06 | 67,851.69 |
| 152,701.92 | 61,136.38 | 88,218.23 | 0 | 1 | 0 | 97,483.56 | 98,791.73 |
| 122,782.75 | 73,994.56 | 303,319.26 | 1 | 0 | 0 | 110,352.25 | 113,969.44 |
| 91,391.77 | 142,107.34 | 366,168.42 | 1 | 0 | 0 | 166,187.94 | 167,921.07 |

# Validating linear regression using Heteroscedasticity




- Check for heteroscedasticity does not indicate a noticeable pattern to prove presence of heteroscedasticity.
- Regression model contains sufficient predictor variables to explain the performance of the dependent variable
- Therefore, assumption of linear regression is valid

# Error metrics

| Statistics | Training | Test | Variance |
|---|---|---|---|
| MSE | 81,571,001.80 | 83,502,864.03 | 2% |
| RMSE | 9,031.67 | 9,137.99 | 1% |
| MAE | 6,341.54 | 7,514.29 | 18% |

- After calculating the Mean square error (mean of the difference of errors) and RMSE (Root) and MAE (Mean average error). We can conclude that the model is the best fit since the variance is not much between the test and the training datasets.

# Key Findings

- The prediction model was best fit and did not under or over fit as per the error statistics explained below.
- Profit is highly positively correlated to R&D and marketing spend.
- Administration cost has a weak correlation to profit
- Interestingly, we can see that R&D cost and Marketing cost have a high correlation, indicating that if one cost increases, the other increases and vice versa. This an area that requires attention and further analysis to improve profits.
- Profit is the highest in Florida and the average costs are the lowest in California which is 22% lesser than Florida and 10% lesser than New York