Problem Statement

Multiple linear regression to predict Profit based on Administration spend, Marketing spends, R&D spend and State. Also, understand the trend and correlation.

Code &Outputs: Read Data and Descriptive analysis.

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('white')
```

```
# Importing the dataset

dataset = pd.read_csv('50_Startups.csv')

#Understanding the data
print(dataset.head()) # top 5 rows of the data set
print(dataset.info()) # general information about the dataset. Fill missing values.
print(dataset.shape()) #find out rows and columns in the dataset
print(dataset.isnull().sum()) #count of null values
print(dataset.describe()) # summary statistics
```

Output: Dataset.head ()

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

Output: Dataset.Info ()

```
RangeIndex: 50 entries, 0 to 49
Data columns (total 5 columns):
```

Code &Outputs: Exploratory Data Analysis and Normalization.

Output: Count of null values and data type

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   R&D Spend        50 non-null     float64
 1   Administration   50 non-null     float64
 2   Marketing Spend  50 non-null     float64
 3   State            50 non-null     object
 4   Profit           50 non-null     float64
dtypes: float64(4), object(1)
memory usage: 2.1+ KB
```

Output: Summary Statistics

	R&D Spend	Administration	Marketing Spend	Profit
count	50.000000	50.000000	50.000000	50.000000
mean	73721.615600	121344.639600	211025.097800	112012.639200
std	45902.256482	28017.802755	122290.310726	40306.180338
min	0.000000	51283.140000	0.000000	14681.400000
25%	39936.370000	103730.875000	129300.132500	90138.902500
50%	73051.080000	122699.795000	212716.240000	107978.190000
75%	101602.800000	144842.180000	299469.085000	139765.977500
max	165349.200000	182645.560000	471784.100000	192261.830000

```
#Exploratory Data Analysis
Profitperstate = dataset.groupby('State')['Profit'].mean().sort_values(ascending=False)
print(Profitperstate.head())
```

```
#assigning x & y for regression
X = dataset.iloc[:, :-1] #all values from all rows  and all columns excluding the
y = dataset.iloc[:, 4] #all values from rows in the last column
#Convert the column into categorical columns

states=pd.get_dummies(X['State'],drop_first=True)

# Drop the state coulmn
X=X.drop('State',axis=1)

# concat the dummy variables
X=pd.concat([X,states],axis=1)
```

Output:

X = (50,4)

Y = (50,)

Output: Profit Grouped by State

```
New York      113756.446471
California    103905.175294
Name: Profit, dtype: float64
```

Get dummy values (1 if value exists and 0 if it doesn't) to assign values to the state. Once in this format, the state column is dropped, and these columns are concatenated

Code &Outputs: Test & Train the model. Analyze Results.

```
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

# Fitting Multiple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results i.e. Predicting Profit based on Administration spend, Marketing spend, R&D spend and State.
y_pred = regressor.predict(X_test)

from sklearn.metrics import r2_score
score=r2_score(y_test,y_pred)
print(score)
```

Output: R2 score

```
0.9347068473282423
```

y_pred uses 20% of the dataset for training (9 data points). y_test for testing utilizes the remaining 80% of datapoints.

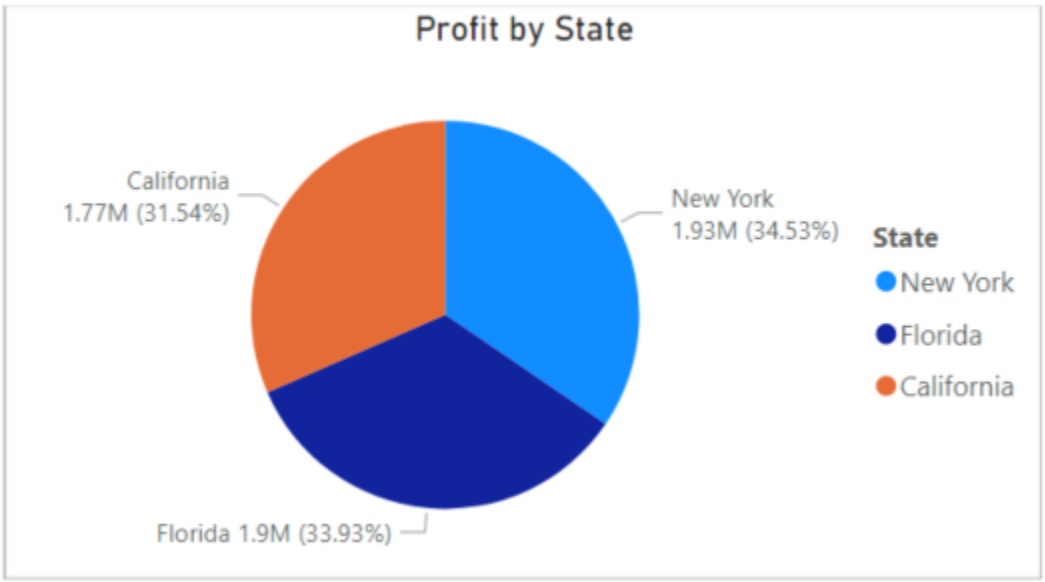
The (Co-efficient of Determination) R^2 is 93% which determines that only 7% of the data points aren't closely fitted towards the best fit line. We can conclude that, variance in 93% of the data points behavior are explained by the solution and have a strong positive correlation with Profit (y).

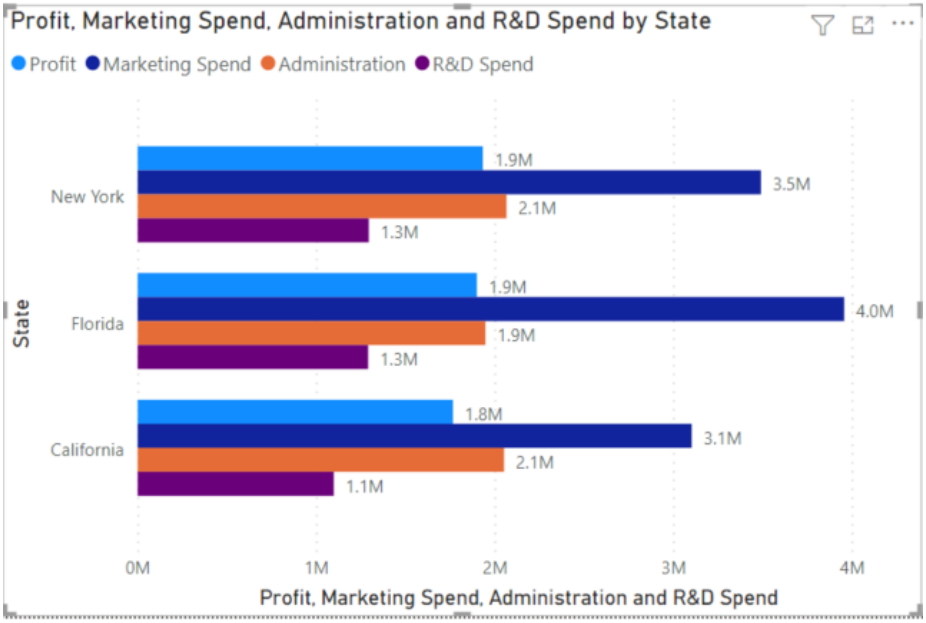
Linear regression equation: Y= Mx = c, where Y is dependent variable the to be predicted, M is the slope (y2-y1/x2-x1), C is a constant which is equal to Y, when X = 0 and X is the independent variable.

For multiple linear regression, there are independent variables: where x1 = Administration spend, Marketing spends, R&D spend and State (converted to a categorical variable).

This use case: Y(Profit) = (M1) *(x1) + (M2) *(x2) + (M3) *(x3) + (M4) *(x4) …. + c

Data Visualization





1. Between New York(NY) and Florida (FL), there is similar R&D and Profit. However, Marketing spends in NY is significantly greater (500k) which could partly increase the Administration spend as shown.
2. Between California (CA) and NY, similar admin spend and profits are seen. However CA has the least marketing and R&D spend.