In [72]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [73]:
```
ds=pd.read_csv('startup.csv')
```

In [38]:
```
ds.head()
```

Out[38]:

	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

In [4]:
```
ds.shape
```

Out[4]:
```
(50, 5)
```

In [5]:
```
ds.describe()
```

Out[5]:

	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

In [76]:
```
X_mod=ds.iloc[:,:-1]
X_mod
```

Out[76]:

	R&D Spend	Administration	Marketing Spend	State
0	165349.20	136897.80	471784.10	New York
1	162597.70	151377.59	443898.53	California
2	153441.51	101145.55	407934.54	Florida
3	144372.41	118671.85	383199.62	New York
4	142107.34	91391.77	366168.42	Florida
5	131876.90	99814.71	362861.36	New York
6	134615.46	147198.87	127716.82	California
7	130298.13	145530.06	323876.68	Florida
8	120542.52	148718.95	311613.29	New York
9	123334.88	108679.17	304981.62	California
10	101913.08	110594.11	229160.95	Florida
11	100671.96	91790.61	249744.55	California
12	93863.75	127320.38	249839.44	Florida
13	91992.39	135495.07	252664.93	California
14	119943.24	156547.42	256512.92	Florida
15	114523.61	122616.84	261776.23	New York
16	78013.11	121597.55	264346.06	California
17	94657.16	145077.58	282574.31	New York
18	91749.16	114175.79	294919.57	Florida
19	86419.70	153514.11	0.00	New York
20	76253.86	113867.30	298664.47	California
21	78389.47	153773.43	299737.29	New York
22	73994.56	122782.75	303319.26	Florida

	R&D Spend	Administration	Marketing Spend	State
23	67532.53	105751.03	304768.73	Florida
24	77044.01	99281.34	140574.81	New York
25	64664.71	139553.16	137962.62	California
26	75328.87	144135.98	134050.07	Florida
27	72107.60	127864.55	353183.81	New York
28	66051.52	182645.56	118148.20	Florida
29	65605.48	153032.06	107138.38	New York
30	61994.48	115641.28	91131.24	Florida
31	61136.38	152701.92	88218.23	New York
32	63408.86	129219.61	46085.25	California
33	55493.95	103057.49	214634.81	Florida
34	46426.07	157693.92	210797.67	California
35	46014.02	85047.44	205517.64	New York
36	28663.76	127056.21	201126.82	Florida
37	44069.95	51283.14	197029.42	California
38	20229.59	65947.93	185265.10	New York
39	38558.51	82982.09	174999.30	California
40	28754.33	118546.05	172795.67	California
41	27892.92	84710.77	164470.71	Florida
42	23640.93	96189.63	148001.11	California
43	15505.73	127382.30	35534.17	New York
44	22177.74	154806.14	28334.72	California
45	1000.23	124153.04	1903.93	New York
46	1315.46	115816.21	297114.46	Florida
47	0.00	135426.92	0.00	California
48	542.05	51743.15	0.00	New York

	R&D Spend	Administration	Marketing Spend	State
49	0.00	116983.80	45173.06	California

In [89]:
```
#One Hot Encoding on Categoricaal data
X_mod['State'].astype('category')
newX =pd.get_dummies(X_mod['State'],drop_first=True)

X=pd.concat([X_mod,newX],axis=1)
X
X=X.drop('State',axis=1)
X
```

Out[89]:

	R&D Spend	Administration	Marketing Spend	Florida	New York
0	165349.20	136897.80	471784.10	0	1
1	162597.70	151377.59	443898.53	0	0
2	153441.51	101145.55	407934.54	1	0
3	144372.41	118671.85	383199.62	0	1
4	142107.34	91391.77	366168.42	1	0
5	131876.90	99814.71	362861.36	0	1
6	134615.46	147198.87	127716.82	0	0
7	130298.13	145530.06	323876.68	1	0
8	120542.52	148718.95	311613.29	0	1
9	123334.88	108679.17	304981.62	0	0
10	101913.08	110594.11	229160.95	1	0
11	100671.96	91790.61	249744.55	0	0
12	93863.75	127320.38	249839.44	1	0
13	91992.39	135495.07	252664.93	0	0
14	119943.24	156547.42	256512.92	1	0
15	114523.61	122616.84	261776.23	0	1
16	78013.11	121597.55	264346.06	0	0
17	94657.16	145077.58	282574.31	0	1
18	91749.16	114175.79	294919.57	1	0

	R&D Spend	Administration	Marketing Spend	Florida	New York
19	86419.70	153514.11	0.00	0	1
20	76253.86	113867.30	298664.47	0	0
21	78389.47	153773.43	299737.29	0	1
22	73994.56	122782.75	303319.26	1	0
23	67532.53	105751.03	304768.73	1	0
24	77044.01	99281.34	140574.81	0	1
25	64664.71	139553.16	137962.62	0	0
26	75328.87	144135.98	134050.07	1	0
27	72107.60	127864.55	353183.81	0	1
28	66051.52	182645.56	118148.20	1	0
29	65605.48	153032.06	107138.38	0	1
30	61994.48	115641.28	91131.24	1	0
31	61136.38	152701.92	88218.23	0	1
32	63408.86	129219.61	46085.25	0	0
33	55493.95	103057.49	214634.81	1	0
34	46426.07	157693.92	210797.67	0	0
35	46014.02	85047.44	205517.64	0	1
36	28663.76	127056.21	201126.82	1	0
37	44069.95	51283.14	197029.42	0	0
38	20229.59	65947.93	185265.10	0	1
39	38558.51	82982.09	174999.30	0	0
40	28754.33	118546.05	172795.67	0	0
41	27892.92	84710.77	164470.71	1	0
42	23640.93	96189.63	148001.11	0	0
43	15505.73	127382.30	35534.17	0	1
44	22177.74	154806.14	28334.72	0	0

	R&D Spend	Administration	Marketing Spend	Florida	New York
45	1000.23	124153.04	1903.93	0	1
46	1315.46	115816.21	297114.46	1	0
47	0.00	135426.92	0.00	0	0
48	542.05	51743.15	0.00	0	1
49	0.00	116983.80	45173.06	0	0

In [90]:
```
y=ds.iloc[:,4]
```

In [12]:
```
y
```

Out[12]:
```
0     192261.83
1     191792.06
2     191050.39
3     182901.99
4     166187.94
5     156991.12
6     156122.51
7     155752.60
8     152211.77
9     149759.96
10    146121.95
11    144259.40
12    141585.52
13    134307.35
14    132602.65
15    129917.04
16    126992.93
17    125370.37
18    124266.90
19    122776.86
20    118474.03
21    111313.02
22    110352.25
23    108733.99
24    108552.04
25    107404.34
26    105733.54
27    105008.31
28    103282.38
29    101004.64
30     99937.59
31     97483.56
32     97427.84
33     96778.92
34     96712.80
35     96479.51
36     90708.19
37     89949.14
38     81229.06
```

```
39     81005.76
40     78239.91
41     77798.83
42     71498.49
43     69758.98
44     65200.33
45     64926.08
46     49490.75
47     42559.73
48     35673.41
49     14681.40
Name: Profit, dtype: float64
```

In [91]:
```
from  sklearn.model_selection import train_test_split
```

In [92]:
```
X_train,X_test,y_train,y_test= train_test_split(X,y,test_size=0.2,random_state=0)
```

In [93]:
```
from sklearn.linear_model import LinearRegression
ml = LinearRegression()
```

In [94]:
```
p = ml.fit(X_train,y_train)
```

In [99]:
```
#Prediction
y_pred = ml.predict(X_test)
```

In [100]:
```
y_pred
```

Out[100]:
```
array([103015.20159796, 132582.27760816, 132447.73845174,  71976.09851258,
       178537.48221055, 116161.24230165,  67851.69209676,  98791.73374687,
       113969.43533012, 167921.0656955 ])
```

In [101]:
```
y_test
```

Out[101]:
```
array([103015.20159796, 132582.27760816, 132447.73845174,  71976.09851258,
       178537.48221055, 116161.24230165,  67851.69209676,  98791.73374687,
       113969.43533012, 167921.0656955 ])
```

In [102]:
```
#R squared to check the model accuracy
from sklearn.metrics import r2_score
score =r2_score(y_test,y_pred)
```

In [103]:
```
score
```

Out[103]:
```
1.0
```

In [ ]:
```
#R Square shows 1 mean the model is best.
```