```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv("https://raw.githubusercontent.com/arib168/data/main/50_Startups.csv")
df
```

	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
5	131876.90	99814.71	362861.36	New York	156991.12
6	134615.46	147198.87	127716.82	California	156122.51
7	130298.13	145530.06	323876.68	Florida	155752.60
8	120542.52	148718.95	311613.29	New York	152211.77
9	123334.88	108679.17	304981.62	California	149759.96
10	101913.08	110594.11	229160.95	Florida	146121.95
11	100671.96	91790.61	249744.55	California	144259.40
12	93863.75	127320.38	249839.44	Florida	141585.52
13	91992.39	135495.07	252664.93	California	134307.35
14	119943.24	156547.42	256512.92	Florida	132602.65
15	114523.61	122616.84	261776.23	New York	129917.04
16	78013.11	121597.55	264346.06	California	126992.93
17	94657.16	145077.58	282574.31	New York	125370.37
18	91749.16	114175.79	294919.57	Florida	124266.90
19	86419.70	153514.11	0.00	New York	122776.86
20	76253.86	113867.30	298664.47	California	118474.03
21	78389.47	153773.43	299737.29	New York	111313.02

```
df.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

```
df.tail()
```

	R&D Spend	Administration	Marketing Spend	State	Profit
45	1000.23	124153.04	1903.93	New York	64926.08
46	1315.46	115816.21	297114.46	Florida	49490.75
47	0.00	135426.92	0.00	California	42559.73

```
df.size
```

```
250
```

```
df.isna().sum()
```

```
R&D Spend          0
Administration     0
Marketing Spend    0
State              0
Profit             0
dtype: int64
```

46 1315.46 115816.21 297114.46 Florida 49490.75

```
df.columns
```

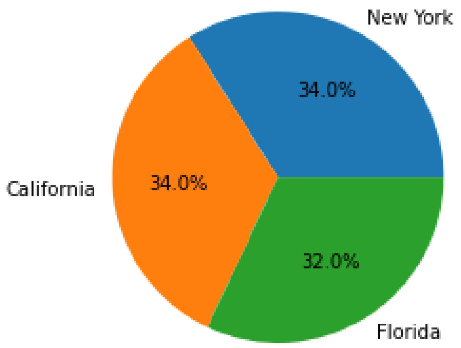
```
Index(['R&D Spend', 'Administration', 'Marketing Spend', 'State', 'Profit'],
dtype='object')
```

```
x=df['State'].value_counts()
x
```

```
New York      17
California    17
Florida       16
Name: State, dtype: int64
```

```
plt.pie(x,labels=x.index, autopct = '%1.1f%%')
plt.show
```

```
<function matplotlib.pyplot.show(*args, **kw)>
```



```
x=df.iloc[:,:4]
y=df.iloc[:,4]
y
```

```
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
```

```
x
```

	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
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

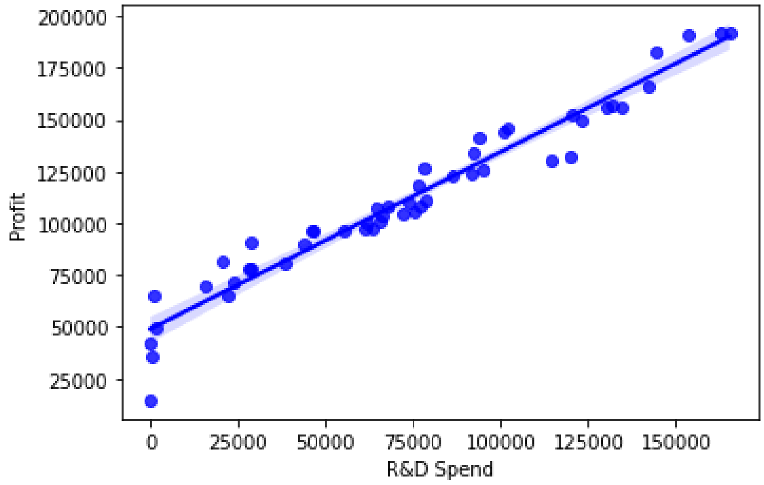
```
df.columns
```

```
Index(['R&D Spend', 'Administration', 'Marketing Spend', 'State', 'Profit'],
dtype='object')
```

49	0.00	116983.80	45173.06	California

```
sns.regplot(x=df['R&D Spend'],y=y,color="blue")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3cc37186a0>
```



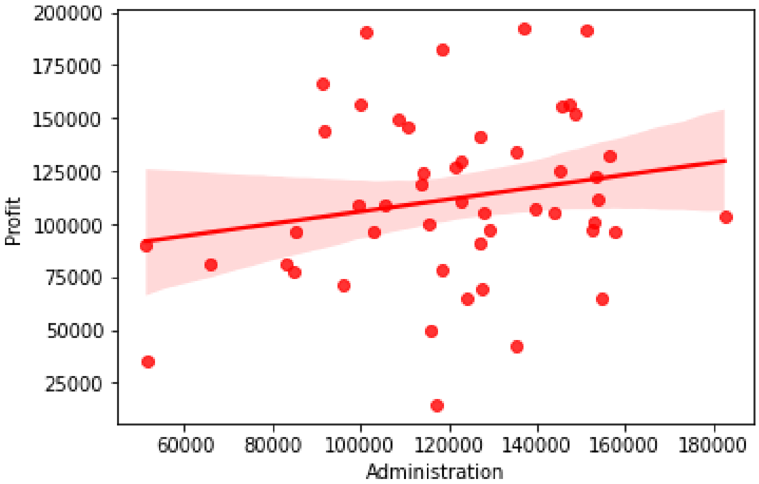
```
df.info
```

```
<bound method DataFrame.info of     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
5   131876.90        99814.71        362861.36    New York  156991.12
6   134615.46       147198.87        127716.82  California  156122.51
7   130298.13       145530.06        323876.68     Florida  155752.60
8   120542.52       148718.95        311613.29    New York  152211.77
9   123334.88       108679.17        304981.62  California  149759.96
10  101913.08       110594.11        229160.95     Florida  146121.95
```

```
11  100671.96        91790.61        249744.55  California  144259.40
12   93863.75       127320.38        249839.44     Florida  141585.52
13   91992.39       135495.07        252664.93  California  134307.35
14  119943.24       156547.42        256512.92     Florida  132602.65
15  114523.61       122616.84        261776.23    New York  129917.04
16   78013.11       121597.55        264346.06  California  126992.93
17   94657.16       145077.58        282574.31    New York  125370.37
18   91749.16       114175.79        294919.57     Florida  124266.90
19   86419.70       153514.11             0.00    New York  122776.86
20   76253.86       113867.30        298664.47  California  118474.03
21   78389.47       153773.43        299737.29    New York  111313.02
22   73994.56       122782.75        303319.26     Florida  110352.25
23   67532.53       105751.03        304768.73     Florida  108733.99
24   77044.01        99281.34        140574.81    New York  108552.04
25   64664.71       139553.16        137962.62  California  107404.34
26   75328.87       144135.98        134050.07     Florida  105733.54
27   72107.60       127864.55        353183.81    New York  105008.31
28   66051.52       182645.56        118148.20     Florida  103282.38
29   65605.48       153032.06        107138.38    New York  101004.64
30   61994.48       115641.28         91131.24     Florida   99937.59
31   61136.38       152701.92         88218.23    New York   97483.56
32   63408.86       129219.61         46085.25  California   97427.84
33   55493.95       103057.49        214634.81     Florida   96778.92
34   46426.07       157693.92        210797.67  California   96712.80
35   46014.02        85047.44        205517.64    New York   96479.51
36   28663.76       127056.21        201126.82     Florida   90708.19
37   44069.95        51283.14        197029.42  California   89949.14
38   20229.59        65947.93        185265.10    New York   81229.06
39   38558.51        82982.09        174999.30  California   81005.76
40   28754.33       118546.05        172795.67  California   78239.91
41   27892.92        84710.77        164470.71     Florida   77798.83
42   23640.93        96189.63        148001.11  California   71498.49
43   15505.73       127382.30         35534.17    New York   69758.98
44   22177.74       154806.14         28334.72  California   65200.33
45    1000.23       124153.04          1903.93    New York   64926.08
46    1315.46       115816.21        297114.46     Florida   49490.75
47       0.00       135426.92             0.00  California   42559.73
48     542.05        51743.15             0.00    New York   35673.41
49       0.00       116983.80         45173.06  California   14681.40>
```

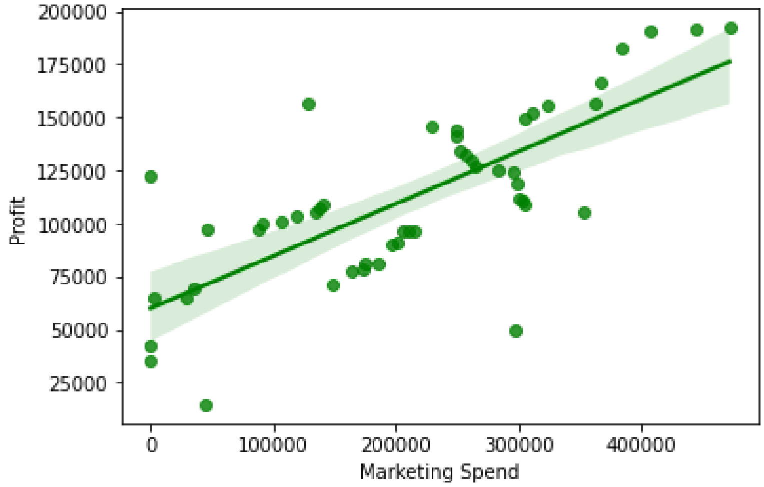
```
sns.regplot(x=df['Administration'],y=y,color="red")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3cc3703ee0>
```



```
sns.regplot(x=df['Marketing Spend'],y=y,color="green")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3cc36121f0>
```



```
# encoding technique        one_hot_encoding


from sklearn.compose import make_column_transformer
from sklearn.preprocessing import OneHotEncoder
col_transfer=make_column_transformer((OneHotEncoder(handle_unknown='ignore'),['State'])
x=col_transfer.fit_transform(x)
x
```

```
array([[0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.6534920e+05,
        1.3689780e+05, 4.7178410e+05],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 1.6259770e+05,
        1.5137759e+05, 4.4389853e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.5344151e+05,
        1.0114555e+05, 4.0793454e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.4437241e+05,
        1.1867185e+05, 3.8319962e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.4210734e+05,
        9.1391770e+04, 3.6616842e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.3187690e+05,
        9.9814710e+04, 3.6286136e+05],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 1.3461546e+05,
        1.4719887e+05, 1.2771682e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.3029813e+05,
        1.4553006e+05, 3.2387668e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.2054252e+05,
        1.4871895e+05, 3.1161329e+05],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 1.2333488e+05,
        1.0867917e+05, 3.0498162e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.0191308e+05,
        1.1059411e+05, 2.2916095e+05],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 1.0067196e+05,
        9.1790610e+04, 2.4974455e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 9.3863750e+04,
        1.2732038e+05, 2.4983944e+05],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 9.1992390e+04,
        1.3549507e+05, 2.5266493e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.1994324e+05,
        1.5654742e+05, 2.5651292e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.1452361e+05,
        1.2261684e+05, 2.6177623e+05],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 7.8013110e+04,
        1.2159755e+05, 2.6434606e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 9.4657160e+04,
        1.4507758e+05, 2.8257431e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 9.1749160e+04,
        1.1417579e+05, 2.9491957e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 8.6419700e+04,
        1.5351411e+05, 0.0000000e+00],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 7.6253860e+04,
        1.1386730e+05, 2.9866447e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 7.8389470e+04,
        1.5377343e+05, 2.9973729e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 7.3994560e+04,
        1.2278275e+05, 3.0331926e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 6.7532530e+04,
        1.0575103e+05, 3.0476873e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 7.7044010e+04,
        9.9281340e+04, 1.4057481e+05],
       [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 6.4664710e+04,
        1.3955316e+05, 1.3796262e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 7.5328870e+04,
        1.4413598e+05, 1.3405007e+05],
       [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 7.2107600e+04,
        1.2786455e+05, 3.5318381e+05],
       [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 6.6051520e+04,
        1.8264556e+05, 1.1814820e+05]
```

```
from sklearn.model_selection import train_test_split
xtrain,xtest,ytrain,ytest=train_test_split(x,y,test_size=.30,random_state=1)
```

```
from sklearn.linear_model import LinearRegression
l=LinearRegression()
l.fit(xtrain,ytrain)
ypred=l.predict(xtest)
ypred
```

```
    array([115325.09875888,  90638.08603376,  76019.13126601,  70325.43761815,
           179659.7398274 , 172204.16410706,  48850.65280981, 101321.43054263,
            58316.95833315,  97217.64504548,  98129.20007849,  84156.44747448,
           117923.69116313,  75866.34008182, 113595.93339165])
```

```
df1=pd.DataFrame({'Actual_value':ytest,'Predicted_value':ypred,'Difference':ytest-ypred
df1
```

	Actual_value	Predicted_value	Difference
27	105008.31	115325.098759	-10316.788759
35	96479.51	90638.086034	5841.423966
40	78239.91	76019.131266	2220.778734
38	81229.06	70325.437618	10903.622382
2	191050.39	179659.739827	11390.650173

```
print("Slope = ",l.coef_)
list(zip(x,l.coef_))
```

```
Slope =  [ 4.21046246e+02 -5.35781864e+02  1.14735618e+02  7.70711613e-01
 -1.41653527e-02  3.50988115e-02]
[(array([0.000000e+00, 0.000000e+00, 1.000000e+00, 1.653492e+05,
         1.368978e+05, 4.717841e+05]), 421.04624582177297),
 (array([1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 1.6259770e+05,
         1.5137759e+05, 4.4389853e+05]), -535.7818635797845),
 (array([0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.5344151e+05,
         1.0114555e+05, 4.0793454e+05]), 114.73561775949355),
 (array([0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.4437241e+05,
         1.1867185e+05, 3.8319962e+05]), 0.770711612652832),
 (array([0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.4210734e+05,
         9.1391770e+04, 3.6616842e+05]), -0.014165352675245657),
 (array([0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.3187690e+05,
         9.9814710e+04, 3.6286136e+05]), 0.035098811509271854)]
```

```
from sklearn.metrics import mean_absolute_percentage_error
print("Percentage",mean_absolute_percentage_error(ytest,ypred))
```

```
Percentage 0.08913280081237054
```

```
from sklearn.metrics import mean_absolute_error
print("MAE :",mean_absolute_error(ytest,ypred))
```

```
MAE : 7229.516119284178
```

```
from sklearn.metrics import r2_score
print("R2 score is ",r2_score(ytest,ypred))
```

```
R2 score is  0.9529676095424967
```

```
from sklearn.metrics import mean_squared_error
print("mean squared error =",mean_squared_error(ytest,ypred))
```

```
mean squared error = 74598131.69470714
```

```
print("Root mean squared error =",np.sqrt(mean_squared_error(ytest,ypred)))
```

```
Root mean squared error = 8637.020996541987
```

✓ 0s completed at 09:30 ● ×