In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
data = pd.read_csv("Startups.csv")
print(data)
```

```
    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            0.00             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
```

In [3]:
```
print(data.describe())
```

```
          R&D Spend  Administration  Marketing Spend        Profit
count     50.000000       50.000000        50.000000     50.000000
mean   73721.615600   123344.639600    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 [4]:
```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
                                        # converting categorical to numerical by lable encoding
label_encoder = LabelEncoder()
data['state_encoded'] = label_encoder.fit_transform(data['State'])
data
```

Out[4]:

	R&D Spend	Administration	Marketing Spend	State	Profit	state_encoded
0	165349.20	136897.80	471784.10	New York	192261.83	2
1	162597.70	151377.59	443898.53	California	191792.06	0
2	153441.51	101145.55	407934.54	Florida	191050.39	1
3	144372.41	118671.85	383199.62	New York	182901.99	2
4	142107.34	91391.77	366168.42	Florida	166187.94	1
5	131876.90	99814.71	362861.36	New York	156991.12	2
6	134615.46	147198.87	127716.82	California	156122.51	0
7	130298.13	145530.06	323876.68	Florida	155752.60	1
8	120542.52	148718.95	311613.29	New York	152211.77	2
9	123334.88	108679.17	304981.62	California	149759.96	0
10	101913.08	110594.11	229160.95	Florida	146121.95	1
11	100671.96	91790.61	249744.55	California	144259.40	0
12	93863.75	127320.38	249839.44	Florida	141585.52	1
13	91992.39	135495.07	252664.93	California	134307.35	0
14	119943.24	156547.42	256512.92	Florida	132602.65	1
15	114523.61	122616.84	261776.23	New York	129917.04	2
16	78013.11	121597.55	264346.06	California	126992.93	0
17	94657.16	145077.58	282574.31	New York	125370.37	2
18	91749.16	114175.79	294919.57	Florida	124266.90	1
19	86419.70	153514.11	0.00	New York	122776.86	2
20	76253.86	113867.30	298664.47	California	118474.03	0
21	78389.47	153773.43	299737.29	New York	111313.02	2
22	73994.56	122782.75	303319.26	Florida	110352.25	1
23	67532.53	105751.03	304768.73	Florida	108733.99	1
24	77044.01	99281.34	140574.81	New York	108552.04	2
25	64664.71	139553.16	137962.62	California	107404.34	0
26	75328.87	144135.98	134050.07	Florida	105733.54	1
27	72107.60	127864.55	353183.81	New York	105008.31	2
28	66051.52	182645.56	118148.20	Florida	103282.38	1
29	65605.48	153032.06	107138.38	New York	101004.64	2
30	61994.48	115641.28	91131.24	Florida	99937.59	1
31	61136.38	152701.92	88218.23	New York	97483.56	2
32	63408.86	129219.61	46085.25	California	97427.84	0
33	55493.95	103057.49	214634.81	Florida	96778.92	1
34	46426.07	157693.92	210797.67	California	96712.80	0
35	46014.02	85047.44	205517.64	New York	96479.51	2
36	28663.76	127056.21	201126.82	Florida	90708.19	1
37	44069.95	51283.14	197029.42	California	89949.14	0
38	20229.59	65947.93	185265.10	New York	81229.06	2
39	38558.51	82982.09	174999.30	California	81005.76	0
40	28754.33	118546.05	172795.67	California	78239.91	0
41	27892.92	84710.77	164470.71	Florida	77798.83	1
42	23640.93	96189.63	148001.11	California	71498.49	0
43	15505.73	127382.30	35534.17	New York	69758.98	2
44	22177.74	154806.14	28334.72	California	65200.33	0
45	1000.23	124153.04	1903.93	New York	64926.08	2
46	1315.46	115816.21	297114.46	Florida	49490.75	1
47	0.00	135426.92	0.00	California	42559.73	0
48	542.05	51743.15	0.00	New York	35673.41	2
49	0.00	116983.80	45173.06	California	14681.40	0

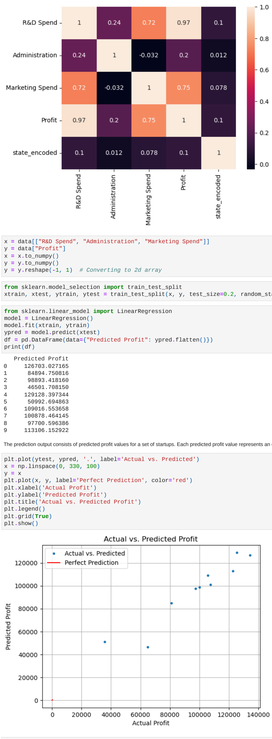
In [5]:
```
data.drop('State',axis =1,inplace = True)
```

In [6]:
```
data
```

Out[6]:

	R&D Spend	Administration	Marketing Spend	Profit	state_encoded
0	165349.20	136897.80	471784.10	192261.83	2
1	162597.70	151377.59	443898.53	191792.06	0
2	153441.51	101145.55	407934.54	191050.39	1
3	144372.41	118671.85	383199.62	182901.99	2
4	142107.34	91391.77	366168.42	166187.94	1
5	131876.90	99814.71	362861.36	156991.12	2
6	134615.46	147198.87	127716.82	156122.51	0
7	130298.13	145530.06	323876.68	155752.60	1
8	120542.52	148718.95	311613.29	152211.77	2
9	123334.88	108679.17	304981.62	149759.96	0
10	101913.08	110594.11	229160.95	146121.95	1
11	100671.96	91790.61	249744.55	144259.40	0
12	93863.75	127320.38	249839.44	141585.52	1
13	91992.39	135495.07	252664.93	134307.35	0
14	119943.24	156547.42	256512.92	132602.65	1
15	114523.61	122616.84	261776.23	129917.04	2
16	78013.11	121597.55	264346.06	126992.93	0
17	94657.16	145077.58	282574.31	125370.37	2
18	91749.16	114175.79	294919.57	124266.90	1
19	86419.70	153514.11	0.00	122776.86	2
20	76253.86	113867.30	298664.47	118474.03	0
21	78389.47	153773.43	299737.29	111313.02	2
22	73994.56	122782.75	303319.26	110352.25	1
23	67532.53	105751.03	304768.73	108733.99	1
24	77044.01	99281.34	140574.81	108552.04	2
25	64664.71	139553.16	137962.62	107404.34	0
26	75328.87	144135.98	134050.07	105733.54	1
27	72107.60	127864.55	353183.81	105008.31	2
28	66051.52	182645.56	118148.20	103282.38	1
29	65605.48	153032.06	107138.38	101004.64	2
30	61994.48	115641.28	91131.24	99937.59	1
31	61136.38	152701.92	88218.23	97483.56	2
32	63408.86	129219.61	46085.25	97427.84	0
33	55493.95	103057.49	214634.81	96778.92	1
34	46426.07	157693.92	210797.67	96712.80	0
35	46014.02	85047.44	205517.64	96479.51	2
36	28663.76	127056.21	201126.82	90708.19	1
37	44069.95	51283.14	197029.42	89949.14	0
38	20229.59	65947.93	185265.10	81229.06	2
39	38558.51	82982.09	174999.30	81005.76	0
40	28754.33	118546.05	172795.67	78239.91	0
41	27892.92	84710.77	164470.71	77798.83	1
42	23640.93	96189.63	148001.11	71498.49	0
43	15505.73	127382.30	35534.17	69758.98	2
44	22177.74	154806.14	28334.72	65200.33	0
45	1000.23	124153.04	1903.93	64926.08	2
46	1315.46	115816.21	297114.46	49490.75	1
47	0.00	135426.92	0.00	42559.73	0
48	542.05	51743.15	0.00	35673.41	2
49	0.00	116983.80	45173.06	14681.40	0

In [7]:
```
sns.heatmap(data.corr(), annot=True)
plt.show()
```



In [8]:
```
x = data[["R&D Spend", "Administration", "Marketing Spend"]]
y = data["Profit"]
x = x.to_numpy()
y = y.to_numpy()
y = y.reshape(-1, 1)  # Converting to 2d array
```

In [9]:
```
from sklearn.model_selection import train_test_split
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size=0.2, random_state=42)
```

In [10]:
```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(xtrain, ytrain)
ypred = model.predict(xtest)
df = pd.DataFrame(data={"Predicted Profit": ypred.flatten()})
print(df)
```

```
   Predicted Profit
0    126703.027165
1     84894.750816
2     98893.418160
3     46581.708150
4    129128.397344
5     59992.694863
6    109016.553658
7    100878.464145
8     97700.596386
9    113106.152922
```

The prediction output consists of predicted profit values for a set of startups. Each predicted profit value represents an estimate of the profit that a startup is expected to generate based on the features provided to the linear regression model.

In [11]:
```
plt.plot(ytest, ypred, '.', label='Actual vs. Predicted')
x = np.linspace(0, 330, 100)
y = x
plt.plot(x, y, label='Perfect Prediction', color='red')
plt.xlabel('Actual Profit')
plt.ylabel('Predicted Profit')
plt.title('Actual vs. Predicted Profit')
plt.legend()
plt.grid(True)
plt.show()
```


In [ ]:
```
```