In [24]:
```
import numpy as np
import pandas as pd
df = pd.read_csv("https://raw.githubusercontent.com/boosuro/profit_estimation_of_companies/master/1000_Companies.csv"
df
```

Out[24]:

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.800	471784.1000	New York	192261.83000
1	162597.70	151377.590	443898.5300	California	191792.06000
2	153441.51	101145.550	407934.5400	Florida	191050.39000
3	144372.41	118671.850	383199.6200	New York	182901.99000
4	142107.34	91391.770	366168.4200	Florida	166187.94000
...	...	...	...	...	...
995	54135.00	118451.999	173232.6695	California	95279.96251
996	134970.00	130390.080	329204.0228	California	164336.60550
997	100275.47	241926.310	227142.8200	California	413956.48000
998	128456.23	321652.140	281692.3200	California	333962.19000
999	161181.72	270939.860	295442.1700	New York	476485.43000

1000 rows × 5 columns

In [25]:
```
x = df.iloc[:,:4]
x
```

Out[25]:

	R&D Spend	Administration	Marketing Spend	State
0	165349.20	136897.800	471784.1000	New York
1	162597.70	151377.590	443898.5300	California
2	153441.51	101145.550	407934.5400	Florida
3	144372.41	118671.850	383199.6200	New York

	R&D Spend	Administration	Marketing Spend	State
4	142107.34	91391.770	366168.4200	Florida
...	...	...	...	...
995	54135.00	118451.999	173232.6695	California
996	134970.00	130390.080	329204.0228	California
997	100275.47	241926.310	227142.8200	California
998	128456.23	321652.140	281692.3200	California
999	161181.72	270939.860	295442.1700	New York

1000 rows × 4 columns

In [26]:
```
y = df.iloc[:,4]
y
```

Out[26]:
```
0      192261.83000
1      191792.06000
2      191050.39000
3      182901.99000
4      166187.94000
           ...
995     95279.96251
996    164336.60550
997    413956.48000
998    333962.19000
999    476485.43000
Name: Profit, Length: 1000, dtype: float64
```

In [27]:
```
import seaborn as sns
sns.heatmap(df.corr())
```

Out[27]:
```
<AxesSubplot:>
```

In [28]:
```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
label = LabelEncoder()
x.iloc[:,3] = label.fit_transform(x.iloc[:,3])
x
```

Out[28]:

	R&D Spend	Administration	Marketing Spend	State
0	165349.20	136897.800	471784.1000	2
1	162597.70	151377.590	443898.5300	0
2	153441.51	101145.550	407934.5400	1
3	144372.41	118671.850	383199.6200	2
4	142107.34	91391.770	366168.4200	1
...	...	...	...	...
995	54135.00	118451.999	173232.6695	0
996	134970.00	130390.080	329204.0228	0

	R&D Spend	Administration	Marketing Spend	State
997	100275.47	241926.310	227142.8200	0
998	128456.23	321652.140	281692.3200	0
999	161181.72	270939.860	295442.1700	2

1000 rows × 4 columns

In [31]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.3,random_state=5)
x_train.shape,x_test.shape
```

Out[31]:
```
((700, 4), (300, 4))
```

In [33]:
```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_train,y_train)
```

Out[33]:
```
LinearRegression()
```

In [36]:
```
y_pred = model.predict(x_test)
model.score(x_train,y_train)*100
```

Out[36]:
```
94.301670727896
```

In [53]:
```
from sklearn.metrics import r2_score
# y_test = np.array(y_test)
r2_score(y_test,y_)*100
```

Out[53]:
```
96.48909018228167
```

In [54]:
```
model.intercept_
```

Out[54]:
```
-83040.30252970719
```

In [57]:
```
coeff = pd.Series(model.coef_)
coeff
```

Out[57]:
```
0      0.573760
1      1.147883
2      0.062838
3    268.473855
dtype: float64
```

In [ ]: