Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1

In [3]:
```
# Importing the Dataset
dataset = pd.read_csv('50_Startups.csv')
X = dataset.iloc[:, :-1].values
```

In [4]:
```
#Splitting the Dataset into the Training set and Test Set
from sklearn.model_selection import train_test_split
```

In [5]:
```
dataset
```

Out[5]:

	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

Berdasarkan informasi tersebut, yang merupakan :

- Variable yang independent adalah :
  - ✓ R&D Spend

Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1

- ✓ Administration
- ✓ Marketing Spend
- ✓ State

- variable yang dependent adalah :
  - ✓ Profit

```
In [8]: X = dataset.iloc[:, :-1].values
        y = dataset.iloc[:, 4].values
```

Dalam X sebagai data dari variable yang dependen , sedangkan y sebagai variabel yang independent.

Maksud dari ":" pada baris pertama berfungsi untuk memilih semua urutan pada index yang kita miliki sedangkan :-1 berfungsi untuk mengambil semua kolom kecuali kolom terakhir, sedangkan iloc() untuk memilih data berdasarkan index.

Sedangkan maksud dari 4 pada baris kedua berfungsi untuk mengambil kolom index ke 4.

```
In [9]: X
Out[9]: array([[165349.2, 136897.8, 471784.1, 'New York'],
               [162597.7, 151377.59, 443898.53, 'California'],
               [153441.51, 101145.55, 407934.54, 'Florida'],
               [144372.41, 118671.85, 383199.62, 'New York'],
               [142107.34, 91391.77, 366168.42, 'Florida'],
               [131876.9, 99814.71, 362861.36, 'New York'],
               [134615.46, 147198.87, 127716.82, 'California'],
               [130298.13, 145530.06, 323876.68, 'Florida'],
               [120542.52, 148718.95, 311613.29, 'New York'],
               [123334.88, 108679.17, 304981.62, 'California'],
               [101913.08, 110594.11, 229160.95, 'Florida'],
               [100671.96, 91790.61, 249744.55, 'California'],
               [93863.75, 127320.38, 249839.44, 'Florida'],
               [91992.39, 135495.07, 252664.93, 'California'],
               [119943.24, 156547.42, 256512.92, 'Florida'],
               [114523.61, 122616.84, 261776.23, 'New York'],
               [78013.11, 121597.55, 264346.06, 'California'],
               [94657.16, 145077.58, 282574.31, 'New York'],
               [91749.16, 114175.79, 294919.57, 'Florida'],
               [86419.7, 153514.11, 0.0, 'New York'],
               [76253.86, 113867.3, 298664.47, 'California'],
               [78389.47, 153773.43, 299737.29, 'New York'],
               [73994.56, 122782.75, 303319.26, 'Florida'],
               [67532.53, 105751.03, 304768.73, 'Florida'],
               [77044.01, 99281.34, 140574.81, 'New York'],
               [64664.71, 139553.16, 137962.62, 'California'],
```

Dalam hal ini, Matrix X dibuat dalam tipe Object agar saat pemrosesan data,

Karena perlu kita ketahui bahwa kita memiliki 1 kolom yang bersifat Categorical yaitu kolom 'State' , dan 3 kolom lainnya yaitu Numerical (R&D Spend, Administration, Marketing Spend, Profit)

Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1
maka semua tipe yang ada seperti categorical dan numerical tersebut akan dapat dibaca dalam tipe yang sama yaitu Object.

Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1

Encode anda menjadi 0 digit di belakang koma

```
In [12]: X

Out[12]: 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, 1.0191388e+05,
                 1.1059411e+05, 2.2916095e+05],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 1.0067196e+05,
                 9.1790510e+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,
                 1.3549587e+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, 7.8813110e+04,
                 1.2159755e+05, 2.6434686e+05],
                [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 9.4657160e+04,
                 1.4587758e+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,
                 1.1386738e+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, 7.7044810e+04,
                 9.9281340e+04, 1.4857481e+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, 6.6851520e+04,
                 1.8264556e+05, 1.1814820e+05],
                [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 6.5605480e+04,
                 1.5303206e+05, 1.0713838e+05],
                [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 6.1994480e+04,
                 1.1564128e+05, 9.1131240e+04],
                [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 6.1136380e+04,
                 1.5270192e+05, 8.8218230e+04],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 6.3408860e+04,
                 1.2921961e+05, 4.6885250e+04],
                [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 5.5493950e+04,
                 1.0305749e+05, 2.1463481e+05],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 4.6426870e+04,
                 1.5769392e+05, 2.1079767e+05],
                [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 4.6814020e+04,
                 8.5047440e+04, 2.0551764e+05],
                [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 2.8663760e+04,
                 1.2705621e+05, 2.0112682e+05],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 4.4069950e+04,
                 5.1283140e+04, 1.9702942e+05],
                [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 2.0229590e+04,
                 6.5947930e+04, 1.8526510e+05],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 3.8558510e+04,
                 8.2982050e+04, 1.7499930e+05],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 2.8754330e+04,
                 1.1854605e+05, 1.7279567e+05],
                [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 2.7892920e+04,
                 8.4710770e+04, 1.6447871e+05],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 2.3640930e+04,
                 9.6189630e+04, 1.4880111e+05],
                [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.5505730e+04,
                 1.2738230e+05, 3.5534170e+04],
                [1.0000000e+00, 0.0000000e+00, 0.0000000e+00, 2.2177740e+04,
                 1.5480614e+05, 2.8334720e+04],
                [0.0000000e+00, 0.0000000e+00, 1.0000000e+00, 1.0002300e+03,
                 1.2415304e+05, 1.9039380e+03],
                [0.0000000e+00, 1.0000000e+00, 0.0000000e+00, 1.3154600e+03,
```

Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1

Pada proses tersebut, dapat melakukan Proses konversi berhasil mengonversi data kategori menjadi angka.

In [11]:
```
# Encoding categorical data
# Encoding the independent variable

from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder_X = LabelEncoder()
X[:,3] = labelencoder_X.fit_transform(X[:,3])
onehotencoder = OneHotEncoder(categorical_features = [3])
X = onehotencoder.fit_transform(X).toarray()
```

Disini, kita melakukan proses encode untuk data kategori. Fungsi dari encode ini sendiri adalah untuk melakukan konversi menjadi sebuah format angka agar bisa dilakukan proses yang secara matematis, karena machine learning sendiri merupakan suatu proses yang matematis.

Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1

UKURAN TEST SET SEBESAR 20% TEST SET

```
In [4]: #Splitting the Dataset into the Training set and Test Set
        from sklearn.model_selection import train_test_split
```

```
In [5]: dataset
```

Out[5]:

	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

Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1

UKURAN TEST SET SEBESAR 30% TEST SET

In [15]:
```
#Splitting the Dataset into the Training set and Test Set
from sklearn.model_selection import train_test_split
```

In [16]:
```
dataset
```

Out[16]:

	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

Nama : Triska Paskah Uli Pangaribuan
NIM : 11S17011
Kelas : 13 IF 1

FITTING MULTIPLE LINEAR REGRESSION TO THE TRAINING SET

```
In [18]: regressor = LinearRegression()

In [19]: regressor.fit(X_train, y_train)

Out[19]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

PREDICTING TEST SET RESULT

```
In [20]: #Predicting Test Set Result
         y_predict=regressor.predict(X_test)
```

Karena X merupakan independent , dan y itu bergantung pada X, sehingga X diinisialisasikan sebagai method predict.

```
In [21]: y_predict

Out[21]: array([103992.71212542, 132640.62967132, 133560.64863724,  72248.59734202,
                179599.66427339, 114641.51252941,  66433.77429985,  98398.72462281,
                114230.68725969, 169025.35824983,  96204.48176884,  88035.23539024,
                110615.4719743 ,  90632.65428793, 127718.94471095])
```

Karena nilai data yang ada dalam data set berbeda tidak signifikan, maka menghasilkan data yang akurat.