Jeremy Ardhito Sulle

2440031250

Multivariate Linear Regression Assignment

Langkah-Langkah

1. Dataset

R&D Spend (x1)	Administration (x2)	Marketing Spend (x3)	State (x4)
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	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
75328.87	144135.98	134050.07	Florida
72107.6	127864.55	353183.81	New York
66051.52	182645.56	118148.2	Florida
65605.48	153032.06	107138.38	New York
61994.48	115641.28	91131.24	Florida
61136.38	152701.92	88218.23	New York
63408.86	129219.61	46085.25	California

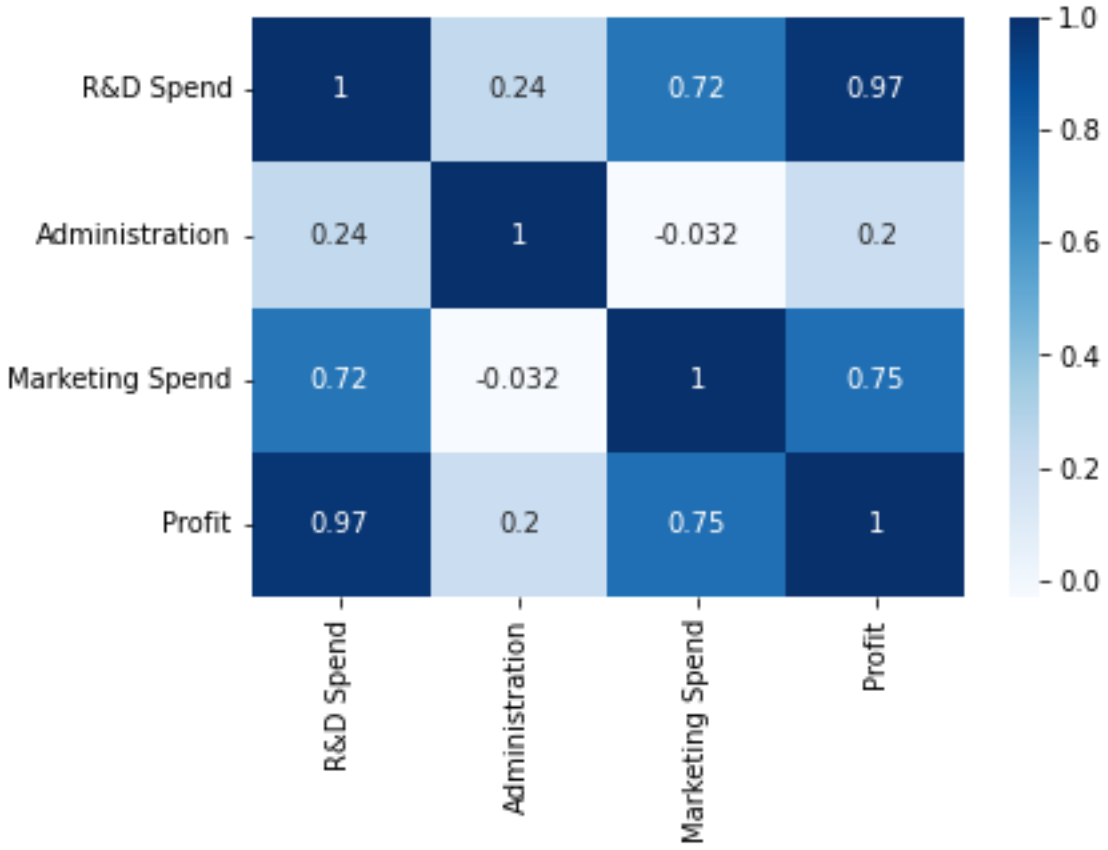
55493.95	103057.49	214634.81	Florida
46426.07	157693.92	210797.67	California
46014.02	85047.44	205517.64	New York
28663.76	127056.21	201126.82	Florida
44069.95	51283.14	197029.42	California
20229.59	65947.93	185265.1	New York
38558.51	82982.09	174999.3	California
28754.33	118546.05	172795.67	California
27892.92	84710.77	164470.71	Florida
23640.93	96189.63	148001.11	California
15505.73	127382.3	35534.17	New York
22177.74	154806.14	28334.72	California
1000.23	124153.04	1903.93	New York
1315.46	115816.21	297114.46	Florida
0	135426.92	0	California
542.05	51743.15	0	New York
0	116983.8	45173.06	California

Variabel state / x4 merupakan categorical data sehingga harus diubah ke numerik agar dapat dihitung, hal ini dapat kita lakukan dengan mengubahnya menjadi dummy variable

R&D Spend	Administration	Marketing Spend	Florida	New York
165349.2	136897.8	471784.1	0	1
162597.7	151377.59	443898.53	0	0
153441.51	101145.55	407934.54	1	0
144372.41	118671.85	383199.62	0	1
142107.34	91391.77	366168.42	1	0
131876.9	99814.71	362861.36	0	1
134615.46	147198.87	127716.82	0	0
130298.13	145530.06	323876.68	1	0
120542.52	148718.95	311613.29	0	1
123334.88	108679.17	304981.62	0	0
101913.08	110594.11	229160.95	1	0
100671.96	91790.61	249744.55	0	0
93863.75	127320.38	249839.44	1	0
91992.39	135495.07	252664.93	0	0
119943.24	156547.42	256512.92	1	0
114523.61	122616.84	261776.23	0	1
78013.11	121597.55	264346.06	0	0
94657.16	145077.58	282574.31	0	1
91749.16	114175.79	294919.57	1	0
86419.7	153514.11	0	0	1

76253.86	113867.3	298664.47	0	0
78389.47	153773.43	299737.29	0	1
73994.56	122782.75	303319.26	1	0
67532.53	105751.03	304768.73	1	0
77044.01	99281.34	140574.81	0	1
64664.71	139553.16	137962.62	0	0
75328.87	144135.98	134050.07	1	0
72107.6	127864.55	353183.81	0	1
66051.52	182645.56	118148.2	1	0
65605.48	153032.06	107138.38	0	1
61994.48	115641.28	91131.24	1	0
61136.38	152701.92	88218.23	0	1
63408.86	129219.61	46085.25	0	0
55493.95	103057.49	214634.81	1	0
46426.07	157693.92	210797.67	0	0
46014.02	85047.44	205517.64	0	1
28663.76	127056.21	201126.82	1	0
44069.95	51283.14	197029.42	0	0
20229.59	65947.93	185265.1	0	1
38558.51	82982.09	174999.3	0	0
28754.33	118546.05	172795.67	0	0
27892.92	84710.77	164470.71	1	0
23640.93	96189.63	148001.11	0	0
15505.73	127382.3	35534.17	0	1
22177.74	154806.14	28334.72	0	0
1000.23	124153.04	1903.93	0	1
1315.46	115816.21	297114.46	1	0
0	135426.92	0	0	0
542.05	51743.15	0	0	1
0	116983.8	45173.06	0	0

Selanjutnya, kita akan melihat korelasi antar fitur menggunakan seaborn



Disini dapat dilihat dua hal yang menarik, pertama Administration memiliki korelasi yang rendah terhadap profit (0.2) sehingga fitur ini dapat di drop, selanjutnya R&D Spend memiliki korelasi yang tinggi terhadap Marketing Spend sehingga mengalami multicollinearity maka salah satu dari fitur tersebut harus di drop, yang saya pilih adalah Marketing Spend karena memiliki korelasi yg lebih rendah terhadap profit daripada R&D Spend.

R&D Spend	Profit	Florida	New York
165349.2	192261.83	0	1
162597.7	191792.06	0	0
153441.51	191050.39	1	0
144372.41	182901.99	0	1
142107.34	166187.94	1	0
131876.9	156991.12	0	1
134615.46	156122.51	0	0
130298.13	155752.6	1	0
120542.52	152211.77	0	1

123334.88	149759.96	0	0
101913.08	146121.95	1	0
100671.96	144259.4	0	0
93863.75	141585.52	1	0
91992.39	134307.35	0	0
119943.24	132602.65	1	0
114523.61	129917.04	0	1
78013.11	126992.93	0	0
94657.16	125370.37	0	1
91749.16	124266.9	1	0
86419.7	122776.86	0	1
76253.86	118474.03	0	0
78389.47	111313.02	0	1
73994.56	110352.25	1	0
67532.53	108733.99	1	0
77044.01	108552.04	0	1
64664.71	107404.34	0	0
75328.87	105733.54	1	0
72107.6	105008.31	0	1
66051.52	103282.38	1	0
65605.48	101004.64	0	1
61994.48	99937.59	1	0
61136.38	97483.56	0	1
63408.86	97427.84	0	0
55493.95	96778.92	1	0
46426.07	96712.8	0	0
46014.02	96479.51	0	1
28663.76	90708.19	1	0
44069.95	89949.14	0	0
20229.59	81229.06	0	1
38558.51	81005.76	0	0
28754.33	78239.91	0	0
27892.92	77798.83	1	0
23640.93	71498.49	0	0
15505.73	69758.98	0	1
22177.74	65200.33	0	0
1000.23	64926.08	0	1
1315.46	49490.75	1	0
0	42559.73	0	0
542.05	35673.41	0	1
0	14681.4	0	0

2. Menghitung Weight dan Bias

   GRADIENT DESCENT

   - Rumus Derivative

$$\frac{\delta Error}{\delta w} = \frac{2}{N}\sum_{i=1}^{N} -x_i(y_i - (mx_i + b))$$

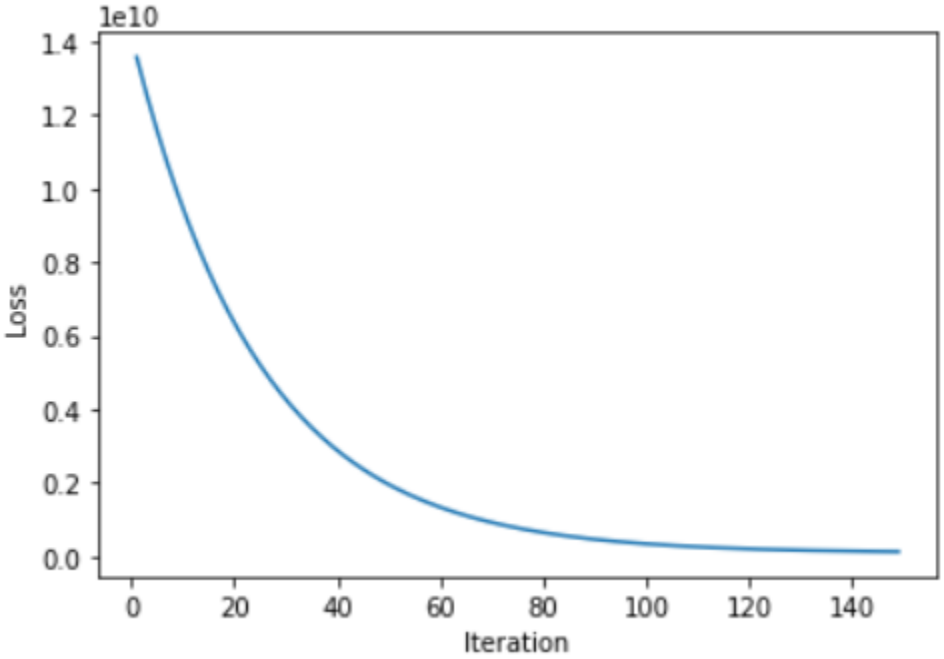
$$\frac{\delta Error}{\delta b} = \frac{2}{N}\sum_{i=1}^{N} -(y_i - (mx_i + b))$$

   - Rumus Update

$$w := w - learning\_rate \cdot \frac{\delta Error}{\delta w}$$

$$b := b - learning\_rate \cdot \frac{\delta Error}{\delta b}$$

   - Hasil Gradient Descent



Dari grafik tersebut dapat diketahui bahwa kita mencapai global optima setelah sekitar 150 iterasi/epoch

HASIL GRADIENT DESCENT DENGAN SCIKIT-LEARN

Weights yang didapatkan scikit-learn dengan test size 25%

Weights	
R&D Spend	36731.46232905
Florida	1555.08791809
New York	818.0475251

Sedangkan Bias yang didapatkan adalah **106602.87439765765**

Perbandingan Y asli dan Y predicted

Ground Truth (y)	y_predicted
192261.8	180741.6

191792.1	176790.6
191050.4	172723
182902	163785.3
166187.9	163561.2
156991.1	153684.8
156122.5	154171.5
155752.6	154015.4
152211.8	144522.8
149760	145053.1
146122	131070.8
144259.4	126733.8
141585.5	124564.2
134307.4	119717.8
132602.7	145645.2
129917	139657.5
126992.9	108417.9
125370.4	123598.8
124266.9	122854.9
122776.9	116940.1
118474	106995.8
111313	110449
110352.3	108503.3
108734	103279.8
108552	109361.5
107404.3	97627.94
105733.5	109581.8
105008.3	105371.2
103282.4	102082.6
101004.6	100115.3
99937.59	98803.19
97483.56	96502.77
97427.84	96612.79
96778.92	93548.59
96712.8	82885.01
96479.51	84278.84
90708.19	71860.82
89949.14	80980.48
81229.06	63436.39
81005.76	76525.39
78239.91	68600.34
77798.83	71237.72
71498.49	64467
69758.98	59617.93

65200.33	63284.25
64926.08	47892.64
49490.75	49754.24
42559.73	45357.22
35673.41	47522.27
14681.4	45357.22

Root Mean Squared Error (RMSE) yang didapatkan = **7832.917201655944**