

Curso Data Science



Prof Me Eng Marcelo Bianchi
Data Scientist



Aula – Machine Learning

Multiple Linear Regression

Machine Learning

Multiple Linear Regression

Multiple Linear Regression

Step 1 – Data Preprocessing (if it will be necessary)

Step 2 – Multiple Linear Regression

Theory

Regressão Múltipla - Definição

Regressão Múltipla

Em um modelo de regressão múltipla, a variável dependente (Y) será determinada por mais de uma variável independente (X). Genericamente, um modelo de regressão linear múltipla com k variáveis independentes e p parâmetros ($p=k+1$) pode ser representado por:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \dots + \beta_k X_{k_i} + e_i$$

Onde:

α é o valor esperado de Y quando todas as variáveis independentes forem nulas;

β_1 é a variação esperada em Y dado um incremento unitário em X_1 , mantendo-se constantes todas as demais variáveis independentes;

...

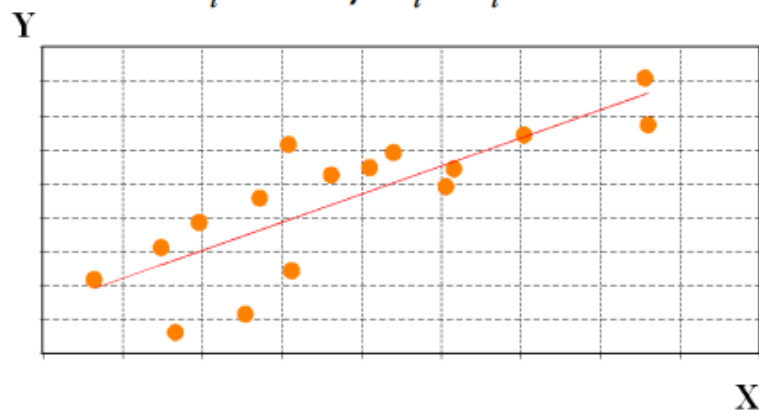
β_k é a variação esperada em Y dado um incremento unitário em X_k , mantendo-se constantes todas as demais variáveis independentes;

e_i é o erro não explicado pelo modelo;

Regressão Múltipla

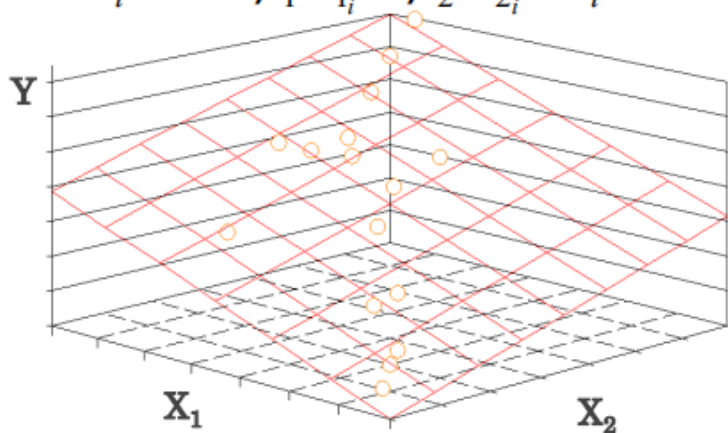
Regressão Linear Simples:

$$Y_i = \alpha + \beta X_i + e_i$$



Regressão Linear Múltipla:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$$



Regressions


**Simple
Linear
Regression**

$$y = b_0 + b_1 * x_1$$

**Multiple
Linear
Regression**

Dependent variable (DV)

Independent variables (IVs)



The diagram shows four green arrows pointing from the labels above to the variables in the equation below. One arrow points from 'Dependent variable (DV)' to 'y'. Three arrows point from 'Independent variables (IVs)' to 'x₁', 'x₂', and 'x_n' respectively.

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

Practical

Nosso Desafio de Predição:

- Baseado nas variáveis independentes: Temos 3 áreas; Spend (Despesas), Administration (Administração) e Marketing (Propaganda). De acordo com o State (Estado), então teremos o lucro.
- A nossa predição será baseada em Quanto a Empresa precisa Gastar nessas 3 áreas, e daí então a empresa terá mais lucro ?

R&D Spend	Administration	Marketing Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.77
123334.88	108679.17	304981.62	California	149759.96
101913.08	110594.11	229160.95	Florida	146121.95
100671.96	91790.61	249744.55	California	144259.4
93863.75	127320.38	249839.44	Florida	141585.52
91992.39	135495.07	252664.93	California	134307.35
119943.24	156547.42	256512.92	Florida	132602.65
114523.61	122616.84	261776.23	New York	129917.04
78013.11	121597.55	264346.06	California	126992.93
94657.16	145077.58	282574.31	New York	125370.37
91749.16	114175.79	294919.57	Florida	124266.9
86419.7	153514.11	0	New York	122776.86
76253.86	113867.3	298664.47	California	118474.03
78389.47	153773.43	299737.29	New York	111313.02
73994.56	122782.75	303319.26	Florida	110352.25
67532.53	105751.03	304768.73	Florida	108733.99

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York		
191,792.06	162,597.70	151,377.59	443,898.53	California		
191,050.39	153,441.51	101,145.55	407,934.54	California		
182,901.99	144,372.41	118,671.85	383,199.62	New York		
166,187.94	142,107.34	91,391.77	366,168.42	California		

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	
191,792.06	162,597.70	151,377.59	443,898.53	California	0	
191,050.39	153,441.51	101,145.55	407,934.54	California	0	
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	
166,187.94	142,107.34	91,391.77	366,168.42	California	0	

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Dummy Variables

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Dummy Variables

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$



Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$



Dummy Variable Trap

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1 + \underline{b_5 * D_2}$$

Dummy Variable Trap

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	$D_2 = 1 - D_1$		California
191,050.39	153,441.51			California
182,901.99	144,372.41			New York
166,187.94	142,107.34			California

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1 + \underline{b_5 * D_2}$$

Dummy Variable Trap

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$



$$+ b_4 * D_1 + \underline{b_5 * D_2}$$



Dummy Variable Trap

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

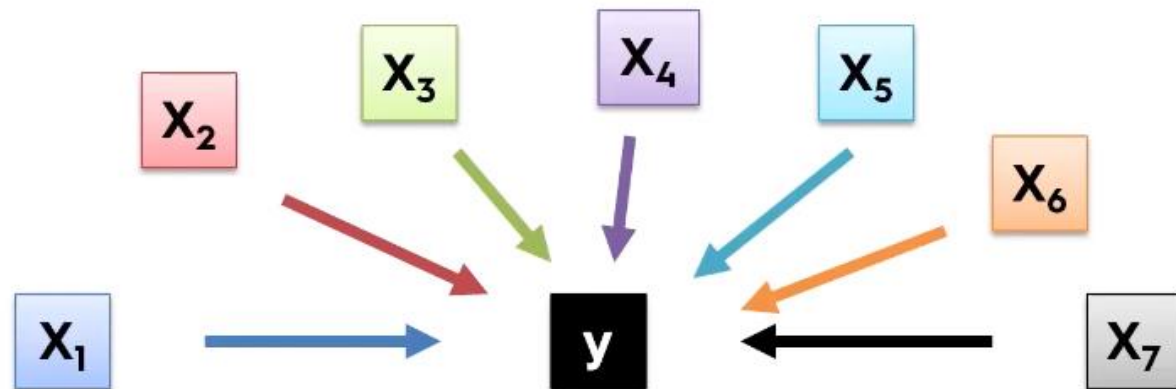
New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3$$

$$+ b_4 * D_1 + \cancel{b_5 * D_2}$$

**Always omit one
dummy variable**

Building A Model



Building A Model

Backward Elimination

STEP 1: Select a significance level to stay in the model (e.g. $SL = 0.05$)



STEP 2: Fit the full model with all possible predictors



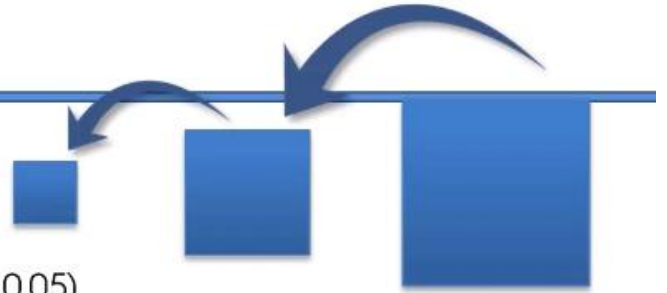
STEP 3: Consider the predictor with the highest P-value. If $P > SL$, go to STEP 4, otherwise go to FIN



STEP 4: Remove the predictor



STEP 5: Fit model without this variable*



Building A Model



Backward Elimination

STEP 1: Select a significance level to stay in the model (e.g. $SL = 0.05$)



STEP 2: Fit the full model with all possible predictors



STEP 3: Consider the predictor with the highest P-value. If $P > SL$, go to STEP 4, otherwise go to FIN



STEP 4: Remove the predictor



STEP 5: Fit model without this variable*



FIN: Your Model Is Ready

Nossa Meta:

- Precisamos determinar se haverá uma dependência linear entre as variáveis.
- Nosso modelo deverá estar apto a predizer o lucro baseado na informação da variável.

Dummy Variables

	0	1	2	3	4	5
0	0	0	1	165349	136896	471784
1	1	0	0	162598	151378	443889
2	0	1	0	153442	101146	407935
3	0	0	1	144372	110672	383200
4	0	1	0	142107	91392	366168
5	0	0	1	131877	99815	362861
6	1	0	0	134615	147199	127717
7	0	1	0	130298	145530	323877
8	0	0	1	120543	148719	311613
9	1	0	0	123335	108679	304982
10	0	1	0	101913	110594	229161
11	1	0	0	100672	91791	249745
12	0	1	0	93864	127320	240839
13	1	0	0	91992	135490	252605
14	0	1	0	110943	156547	256513
15	0	0	1	114524	122617	261776

Dummy Variables

Index	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349	136898	471784	New York	192262
1	162598	151378	443899	California	191792
2	153442	181146	487935	Florida	191858
3	144372	118672	383288	New York	182982
4	142187	91392	366168	Florida	166188
5	131877	99815	362861	New York	156991
6	134615	147199	127717	California	156123
7	138298	145538	323877	Florida	155753
8	128543	148719	311613	New York	152212
9	123335	188679	384982	California	149768
10	181913	118594	229161	Florida	146122
11	188672	91791	249745	California	144259
12	93864	127328	248839	Florida	141586
13	91992	135495	252665	California	134387
14	119943	156547	256513	Florida	132683
15	114524	122617	261776	New York	129917
16	78813	121598	264346	California	126993
17	94657	145878	282574	New York	125378
18	81749	114176	364878	Florida	126287

	0	1	2	3	4	5
0	0	1	0	165349	136898	471784
1	1	0	1	162598	151378	443899
2	0	0	0	153442	181146	487935
3	0	0	1	144372	118672	383288
4	0	1	0	142187	91392	366168
5	0	0	1	131877	99815	362861
6	1	0	0	134615	147199	127717
7	0	1	0	138298	145538	323877
8	0	0	1	128543	148719	311613
9	1	0	0	123335	188679	384982
10	0	1	0	181913	118594	229161
11	1	0	0	188672	91791	249745
12	0	1	0	93864	127328	248839
13	1	0	0	91992	135495	252665
14	0	1	0	119943	156547	256513
15	0	0	1	114524	122617	261776

Format

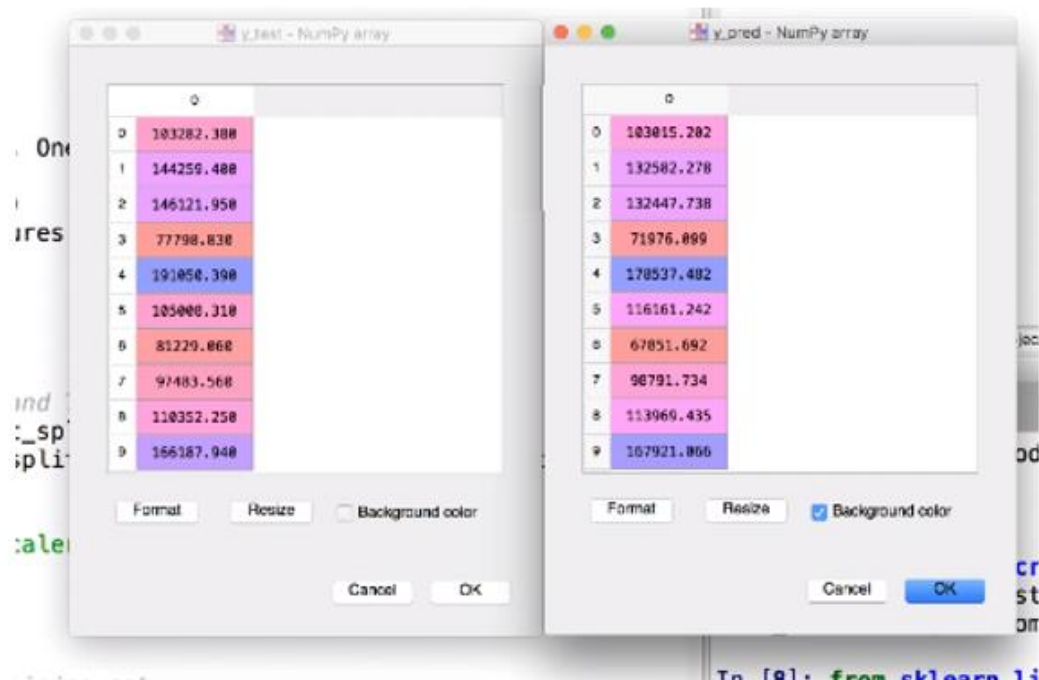
Resize

☐ Background color

Evitando a Armadilha da Dummy Variable

- Para evitar a armadilha da dummy variable, será necessário ter as colunas das dummies variables -1.
- Por exemplo, se você tiver 3, você usará somente 2.
- $N-1 \rightarrow 3-1 = 2$

Real life Profit x Profit Prediction



Accurate Predictions ! -> Good Model !

Importante: Para termos um modelo ótimo

- Precisamos construir um modelo ótimo, no qual será necessário usar o método para eliminar algumas variáveis que não tem dependência linear estatística alta.
- Apenas irá restar variáveis independentes que irão ter alta dependência estatística na variável dependente Lucro.

Modelo Ótimo

Nota:

- Quanto menor o p value, maior será a significancia que a variável independente terá. Então iremos eliminar as colunas que irão ter o p-value maior a cada rodada , até chegarmos a uma última coluna.

Técnica Backward Elimination

Iremos remover 1 por 1 restando apenas a coluna que terá alto impacto no lucro que é a nossa variável dependente.

```

# Fitting Multiple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results
y_pred = regressor.predict(X_test)

# Building the optimal model using Backward Elimination
# import statsmodels.formula.api as sm
import statsmodels.regression.linear_model as sm

#Your X_opt array has a dtype of object and this may be causing an error. Try changing it to float. For example you can use this:
X = np.append(arr = np.ones((50,1)).astype(int), values = X, axis =1)
X_opt = X[:,[0,1,2,3,4,5]]
X_opt = np.array(X_opt, dtype=float)

#X = np.append(arr = np.ones((50, 1)), values = X, axis = 1)
#X_opt = X[:, [0, 1, 2, 3, 4, 5]]
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()

X_opt = X[:, [0, 1, 3, 4, 5]]
X_opt = np.array(X_opt, dtype=float)
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()

X_opt = X[:, [0, 3, 4, 5]]
X_opt = np.array(X_opt, dtype=float)
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()

X_opt = X[:, [0, 3, 5]]
X_opt = np.array(X_opt, dtype=float)
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()

X_opt = X[:, [0, 3]]
X_opt = np.array(X_opt, dtype=float)
regressor_OLS = sm.OLS(endog = y, exog = X_opt).fit()
regressor_OLS.summary()

```

Chegamos a variável vencedora remanescente

OLS Regression Results						
=====						
Dep. Variable:	y	R-squared:	0.947			
Model:	OLS	Adj. R-squared:	0.945			
Method:	Least Squares	F-statistic:	849.8			
Date:	Sun, 07 Feb 2021	Prob (F-statistic):	3.50e-32			
Time:	22:11:56	Log-Likelihood:	-527.44			
No. Observations:	50	AIC:	1059.			
Df Residuals:	48	BIC:	1063.			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	4.903e+04	2537.897	19.320	0.000	4.39e+04	5.41e+04
x1	0.8543	0.029	29.151	0.000	0.795	0.913
=====						
Omnibus:	13.727	Durbin-Watson:	1.116			
Prob(Omnibus):	0.001	Jarque-Bera (JB):	18.536			
Skew:	-0.911	Prob(JB):	9.44e-05			
Kurtosis:	5.361	Cond. No.	1.65e+05			
=====						

R&D Spend	Administration	Marketing Spend	State	Profit
1653492	1368978	4717841	New York	19226183
1625977	15137759	44389853	California	19179206
15344151	10114555	40793454	Florida	19105039
14437241	11867185	38319962	New York	18290199
14210734	9139177	36616842	Florida	16618794
1318769	9981471	36286136	New York	15699112
13461546	14719887	12771682	California	15612251
13029813	14553006	32387668	Florida	1557526
12054252	14871895	31161329	New York	15221177
12333488	10867917	30498162	California	14975996
10191308	11059411	22916095	Florida	14612195
10067196	9179061	24974455	California	1442594
9386375	12732038	24983944	Florida	14158552
9199239	13549507	25266493	California	13430735
11994324	15654742	25651292	Florida	13260265
11452361	12261684	26177623	New York	12991704
7801311	12159755	26434606	California	12699293
9465716	14507758	28257431	New York	12537037
9174916	11417579	29491957	Florida	1242669
864197	15351411	0	New York	12277686
7625386	1138673	29866447	California	11847403
7838947	15377343	29973729	New York	11131302
7399456	12278275	30331926	Florida	11035225
6753253	10575103	30476873	Florida	10873399
7704401	9928134	14057481	New York	10855204
6466471	13955316	13796262	California	10740434
7532887	14413598	13405007	Florida	10573354
721076	12786455	35318381	New York	10500831
6605152	18264556	1181482	Florida	10328238
6560548	15303206	10713838	New York	10100464
6199448	11564128	9113124	Florida	9993759
6113638	15270192	8821823	New York	9748356
6340886	12921961	4608525	California	9742784
5549395	10305749	21463481	Florida	9677892
4642607	15769392	21079767	California	967128
4601402	8504744	20551764	New York	9647951
2866376	12705621	20112682	Florida	9070819

Coluna 3: Marketing (Coluna mais importante – modelo ótimo)

- Poderá ser aplicado agora o Simple Linear Regression visto que a coluna Marketing é a que tem mais alta significancia estatística para o Profit (lucro).
- Ou seja posso utilizar o Backward Elimination, achar a variável indepente mais importante e então aplicar o algoritmo de simple linear regression.

Muito obrigado!