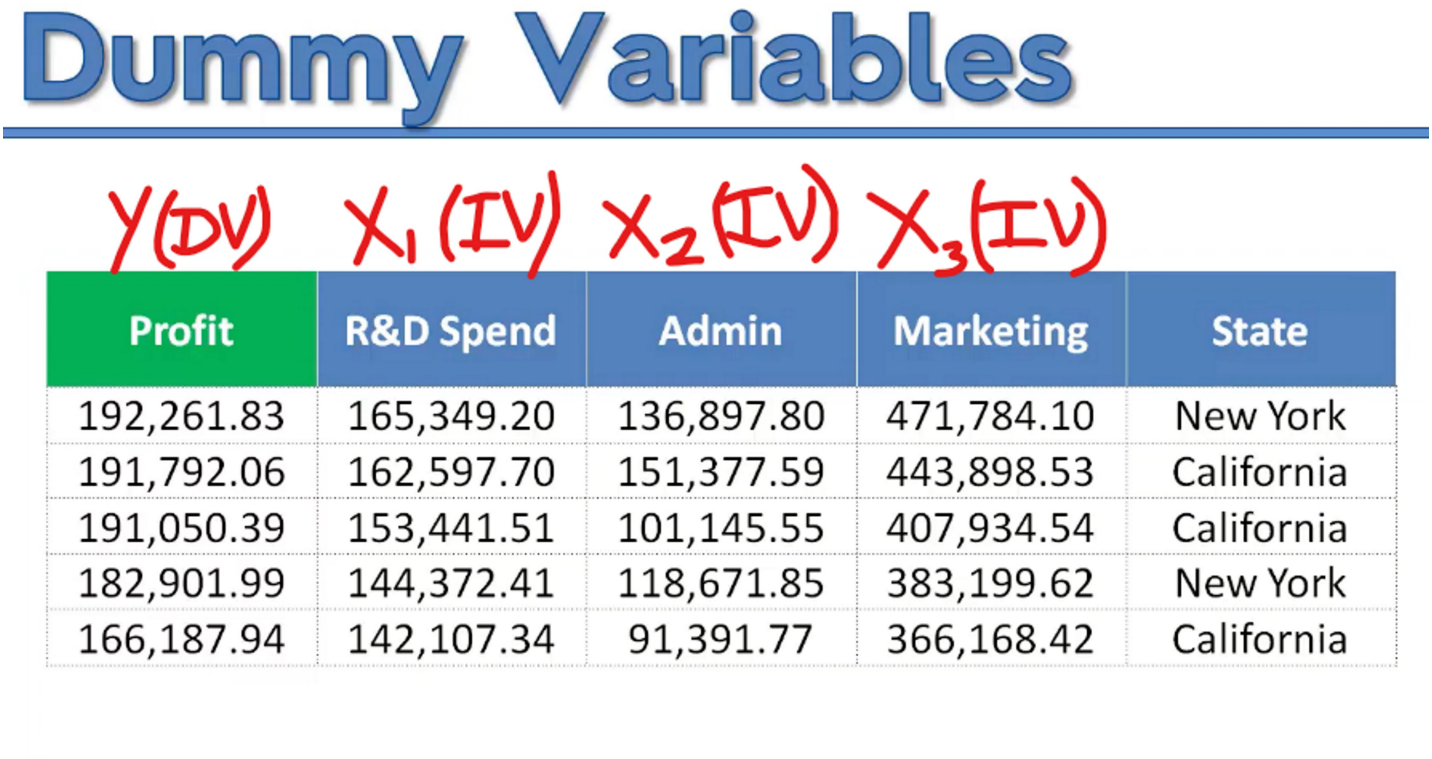
Step 3: Multiple Linear Regression Intuition

Friday, April 8, 2022 12:57 AM



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 is a Dependent Variable(DV). For Instance, how does a person's salary change with the years of experience he has.
- X_1, X_2, X_3 is an Independent Variables (IVs) is variable that, we are assuming that it is causing dependent variable to change.
- **State** is also an independent variable but it doesn't has number and it contains details of state so, it is Categorical Variable but we don't have a number here or dollar value or any other type of number to add into our equation. We can't add word to our equation(in Multiple Linear Regression).
  The approach that we need to take when we face Categorical Variables in Regression model is we need to create **Dummy Variables**. Lets see how we do that:
  - First go through the column and find the different categories we have. In this case we have two categories so, for every single category that we find, we need to create a new column.
    For New York we are going to create a column called New York and For California, we are going to create column California. We are kind of expanding our dataset and adding some additional columns into it.

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	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

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

- We transformed state into two different column. We need to find all of our rows where "State" actually says New York and for those rows we need to put a 1 in the new York Column and same goes for California as well. "1" and "0" acts as a switch.

- Building our Regression model from here is very simple. All we have to do is use the New York column and we are not going to use "State" column anymore. Basically, we add a variable which is b_4*D_1. And D1 in this case is our Dummy Variable for New York and we don't use California column either. So, as we can see here all of the information in our data is preserved if we just stick to the one New York column because we can tell right away if D1 is 1 then, it's the company that operates in New York. And if D1 is 0 then, it's the company that operates in California.

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	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

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

So, we didn't loose information by including only the New York Column.

Note: We should never include all of our Dummy Variable columns in our Regression Model.