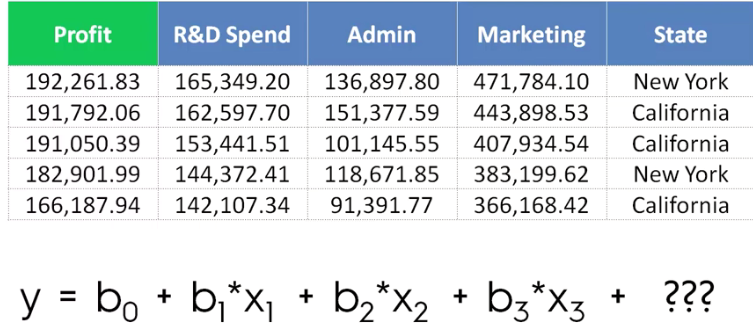
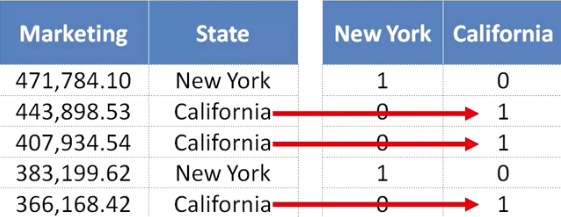
***Multiple Linear Regression***

* When dealing w/ categorical variables in linear regression, we need **dummy variables**



* Find all unique categories, and split them out into their own binary column, which will preserve all the data, but just say if it was in a category or not



* Use one of the columns (i.e. New York) in the linear regression



* Now, for CA, the coefficient here = 0 for this equation, so it’s a bit of a pointless thing at first glance
* But the linear regression model will take this dummy variable *not* in the equation and it will become the **default situation/state** for the model
* Therefore, the CA coefficient is included in the constant b0 and when D1 = 0, we would have an “alternate” equation that represents data for CA houses
* And then when D1 has a value, the model is adding the difference between NY + CA to alter the formula into the NY state (on/off switch)
* **Dummy Variable Trap**
* If we include both dummy variables above, we are then *duplicating* *a variable*
* *D2 always = 1 – D1*
* So, there’s **multicollinearity =** 1 or more independent variables predicts another
* When this happens the model cannot differentiate the difference between the effects of D1 vs. the effects of D2 + therefore won’t work properly
* Cannot have b0 and both dummy variables in it at the same time
* *Always remove ONE dummy variable* (9 states = 8 dummy variables)
* If we have 2 sets of dummy variables, apply the same rules to each set (weather type + state)
* Building a model
* Don’t necessarily need *all* predictors/features in a linear regression
* Need a simple model that is easy to explain/present, so only keep very important variables that actually predict something significantly
* Methods
* **All-in** 🡪 throw in all variables (only do if we know for sure from domain knowledge/previous knowledge that all factors are important)
* Or we may *have* to due to some framework in the company (not your decision)
* **Backwards elimination**
* 1) Select significance level for the variable to reach to STAY in the model
* 2) Fit all variables into a model
* 3) Consider predictor w/ highest p-value (worst predictor) + remove it
* 4) Fit new model w/out that variable
* 5) Repeat step 3 and 4 until all variable p-values are < significance level
* **Forward selection**
* 1) Select significance level for the variable to ENTER the model
* 2) Fit all possible SIMPLE regression models (y = b1\*x1)
* 3) Select simple linear regression model w/ lowest p-value
* 4) Fit all possible regression models w/ the variable from step 3 PLUS a 2nd predictor
* 5) Select new linear regression model w/ lowest p-value
* 6) Add 3rd, then 4th variable…. 🡪 repeat until variable added becomes > significance
* 7) Keep model prior to adding this variable that has too large of a p-value
* **Bidirectional elimination (Stepwise regression)**
* 1) Select a significance level for the variable to ENTER the model + to STAY in the model
* 2) Perform Step 2 of forward selection (fit all simple linear regression models)
* 3) Perform ALL steps of backward elimination
* 4) Repeat Step 2 and 3 until we cannot add nor remove any variables
* **Score comparison/All possible models**
* 1) Select criterion of **goodness of fit** (e.g. Akaike criterion)
* 2) Construct all possible regression models (2^n - 1 total combinations)
* 3) Select model that best meets the criterion
* *NOTE*: 10 columns = 1023 models