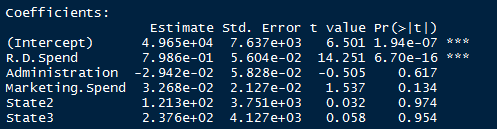
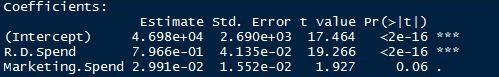
**Business problem:** Identify what variables (R.D.Spend, Administration spend, Marketing spend and the State the company is in) have an influence on the profit a company makes.

**Explaining the model**

1. Import the dataset, which is in a csv format.
2. Need do transform the categorical data (from the State column, we have New York, California and Florida) into numerical values (1, 2 and 3, respectively).
3. After installing the package caTools, I’ve decided to split the sets on 80% – 20% proportion for the training and test sets, respectively. We have more entries this time, which allows us to increase the number of entries in the training set, while at the same time maintain a good number of entries in the test set.
4. Next, the regressor is created. We want to see how the profit is impacted by the 4 independent variables. The data we want to use in our regressor is our Training Set, because we’re training our model.
5. When we analyze our regressor, we can see there are different coefficients for the independent variables, with only R.D.Spend affecting significantly the dependent variable. By now, we could be transforming this into a linear regression model, and only use the R.D.Spend as an independent variable.

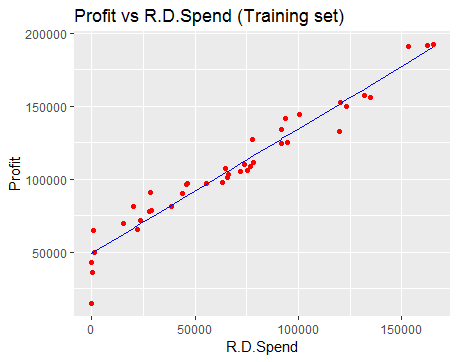


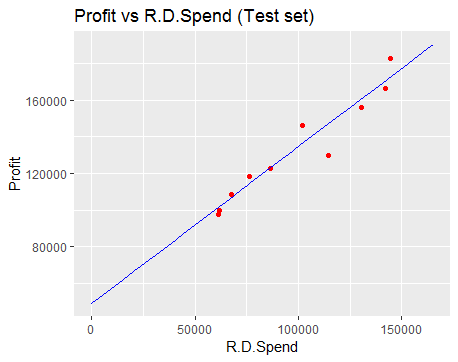
1. However, to have a more optimized model, a Backward Elimination process can be implemented, which basically means removing one the least significant independent variable at a time while seeing the effect it has on the others. This process should be performed on the whole dataset, as it allows a more broader view of the impact of each independent variable on the dependent variable.
2. After removing the state (dummy) variables, we can see that we still have R.D.Spend as the variable with more impact on the dependent variable, but Marketing Spend as now increased its impact. Still, since it’s still above 0.05, I’ve decided to remove it from the model.



1. After this, it’s time to see the results of our model

**Analyzing the results**





The red dots are the real observations, and the line represents our model. Again, the model shows the tendency of the data, allowing us to conclude that the money companies spend on R.D. have a positive correlation with the profit they make.