1000 Biotech Startups

**Problem Statement**

Biotech companies are very competitive. This is a field that thrives on innovative products that help people worldwide. A huge part of that innovation is the research and development department. How much should companies invest in this department? What makes a successful Biotech company? How should their money be allocated to maximize profits?

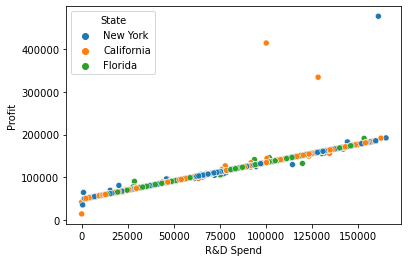
**Data Wrangling**

The raw data was selected through Kaggle and was relatively clean and complete. There were 1000 rows and 5 columns. The data includes Research and Development Spending(R&D), Administrative Spending (Administration), Market Spending, Profit, and State.

**Exploratory Data Analysis**

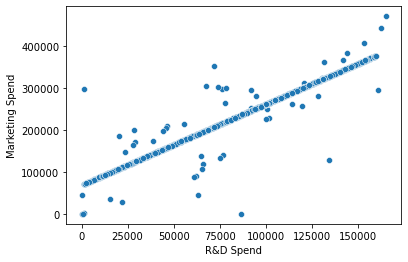
You need to understand your data to properly analyze it. The methods I use to get a feel for the data is the correlation(.corr) and heatmap function. The correlation function gives me a numerical representation of the data, while the heatmap is a more visual aid. From there you select data points that have similarities in the way they behave.

The main objective in any business is to maximize profits. So, I decided to take a look at R&D spending and Profit. I used a scatterplot to check their relationship.

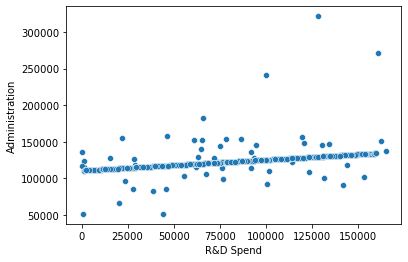


From the scatterplot, we can see that R&D spending and Profit are very linear, with very distinct outliers. The outliers only appear when the R&D spending reach around $100,000 and are only in New York/California.

Let’s look at other variables to see how they compare to R&D spending. The first we will be looking at is the allocation of Marketing and Administrative Spending compared to R&D.



For the most part the scatterplot is linear, but there are a lot more outliers. Low R&D spending seem to also have very low market spending, we can assume this is due to low capital for newer Biotech companies.



The administrative Spending appears to be very consistent, hovering from $100,000-$150,000 no matter how much R&D spending.

We would recommend R&D spending to have a budget of $80,000-$150,000. Administrative Spending shouldn’t be anymore than $125,000. Marketing expenses should be around $75,000.

**Model Selection**

**Linear Regression**

We tested the parameters with profit as our dependent variable while R&D Marketing, and Administrative Spending were predictor variables. The mindset of the 1000 Biotech companies should be to increase profit. According to the exploratory analysis, many of the predictor are very linear so Linear Regression should be an ideal model. About 90%of the variance can be explained by the predictor variables.

**Random Forest Regression**

The Random Forest model had an initial accuracy of around 67%. When we tuned the max depth from 1 to 2 the accuracy increased by 24% to around 91%. Our data has very high outliers, so there might be a snowballing effect when we increase the sample.

**Random Forest Classifier**

The Biotech companies were pooled from three different states, and the outliers only appeared in two states. I decided to set up a random forest classifier to test how accurately the predictors could predict states. The predictors had a hard time identifying the states with about 30% of variance explained.

**Conclusion**

Random Forest Regression can explain more variance than Linear Regression by a miniscule amount. A classifier to determine state has very low accuracy and should not be used. The Random Forest model is the best due to the flexibility of the hyperparameters and very similar accuracy to Linear Regression.