**Problem:**

Startups (or any business in general) face the difficult decision on how to invest the capital available to them. It is easy to feel overwhelmed and it feel like every area of the business is a priority. However, not every investment will affect the growth of the business in the same way.

**Impact:**

How to manage the investment of a company is one of the most (if not the most) crucial decision that any Executive team needs to handle. The company growth and overall survival depends on it.

**Solution:**

The data collected is from [Kaggle](https://www.kaggle.com/karthickveerakumar/startup-logistic-regression) and it is based on real example from startups in the United States , specifically in California (17), New York (17) and Florida (16). Although the data set itself is not very extensive (50 samples), it is possible to clearly answer our problem.

Our dataset has the following features:

* R&D budget (quantitative)
* Marketing budget (quantitative)
* Administration budget (quantitative)
* State (qualitative): California, New York and Florida
* Profit (quantitative) 🡪 **Outcome**

After checking for NAs in the data, here is a first analysis:

A picture containing text, outdoor

Description automatically generated

From this table is possible to see that in all regions, the R&D budget is the smallest and that the Marketing budget is the biggest investment. However, this doesn’t answer our main question: what budget does it have the most significant impact on profits?

Just in regards to profits per state, we can see the following:

Chart, box and whisker chart

Description automatically generated

Where it is obvious that Florida has the highest median, all states are relatively close to each other.

How does the relationship between each budget and the profits?

Chart, scatter chart

Description automatically generated

Here is clearer that the Marketing budget is the largest but it does not necessarily reflect on better results. In fact, we could even say that at some point additional Marketing budget does not change the Profits.

This is an interesting finding so let’s dig into it:

Chart, bar chart

Description automatically generated

Now it becomes clear that the strongest correlation is between R&D spend and Profits (Marketing does also have an impact but not as strong as R&D).

Would it be possible to quantify this? Yes, we can model these relationships through a multi-linear model.

Our model output is:

Text

Description automatically generated

How to interpret these results:

* The R^2 output shows that 95% of our data fits our model (very high!)
* In order to consider the other values, we need to verify:
  + Noise is Gaussian with expectation zero
    - Histogram:

­­Chart, histogram

Description automatically generated

* + - Shapiro Test: p-value = 0.1166. Thus, it is Gaussian
    - From plotting the model:

Chart, scatter chart

Description automatically generated

* + Noise variance is constant (homoscedasticity)
    - From plotting the model

Chart, scatter chart

Description automatically generated

* + - Since it isn’t a straight line, we better double check with Breusch-Pagan test whose result is p-value:0.4483. Thus, we fail to reject H0, then we have homoscedasticity
* After this validation, we can conclude that our model is (with a confidence interval of 95%):
  + Profit = 50,740 + 0.8018 \* R&D Budget
  + The Marketing budget has very little effect of factor of 0.02676
* Validating the model with predictions:

Chart, scatter chart

Description automatically generated

**Conclusions**

* **Business**
  + Contrary to what perhaps is expected in many startups, R&D Budget is the most important budget to increase the company’s profits.
  + There is not apparent relevance based on the startup state
* **Data science**
  + Modelling a problem without validating the noise assumptions reduces significantly the interpretation from the results since the coefficients could not be used
  + This is a small dataset and the analysis so far has been done in the complete dataset which does not test the model against “unseen” data. An alternative is to use the “caret” package from R which will allow to split the data several times (25 by default with caret) by using the bootstrap method and applying it to a new subset of the data.
  + The R squared results are not so different:
    - Full data: 0.9554
    - Caret approach: 0.9421

It makes total sense to have a lower result with Caret as it actually tests the models with “unseen” data and then averages all the R squared results

* + Considering the limited samples, the model parameters may change based on new data but the conclusions remain in regards to the most dominant feature to predict the final outcome