Is there any correlation between the top 500 grossing US companies?

Oskar Aaron Oramus

December 2022

1 Introduction

This paper aims to investigate if there is any correlation between the size, industry, location and other factors of the top 500 grossing companies in the US and their revenue or profit. The methods used to answer these questions will be lasso and ridge regression, as well as polynomial regression, to test the feasibility of the hypothesis. The results of this study will provide insight into the factors that influence the success of companies, and whether it can be narrowed down to a few (or potentially one) factors.

The data used in this study was collected from "data.world" list called Fortune 500[1] and was collected during year 2017. This dataset can be used to perform exploratory data analysis to gain insights about the companies. This involves creating visualisations to understand the distribution of the companies across different industries, examining the relationship between different variables like (revenue and employee count), but also identifying trends and patterns in the data.

We can potentially use the data to develop a predictive model capable of forecasting the company's future. If companies in the same industry and/or location are doing very well, the model could use that to predict the growth of the company.

A likely challenge when using this dataset is that some data may be incomplete, inaccurate or missing. This could impact the accuracy of the data analysis or predictions. To address these issues data will be cleaned up, preprocessed and graphed to quickly spot any outliers. Additionally, it might be useful to get more information on the companies in the dataset to help the model understand it better.

2 Methodology

Features of the data are shown in the table shown in table 1.

Rank	Rank based on the total profit in the year
Title	Name of the company
Website	Full URL of the company
Employees	Number of employees
Sector	General group of the company
Industry	Specific group of the company
Hqlocation, Hqaddr, Hqcity, Hqstate, Hqzip, Hqtel	Location of the head quarters divided up
Ceo, Ceo-title, Address	CEO details
Ticker	Symbols under which the company trades, representing the company's different types of shares.
Fullname	The registration name of the company
Revenues, Revchange	Revenue and revenue change since last year respectively
Profits, Prftchange	Profits and profit change since last year respectively
Assets	Number of company assets
Totshequity	Owner's claim after subtracting total liabilities from total assets

Table 1: Data Features in the data set

The data was then cleaned and formatted to ensure accuracy and consistency. For example, the machine learning model does not need to know about the CEO's address or title to extract meaningful information. Finally, the data were normalized to ensure that the results were not biased by outliers. This was done by removing extreme values and standardizing the data. This allowed the model to focus on the most important features and ignore the outliers. There are many ways to standardize the data, one would be to take a logarithm of everything when dealing with high numbers (like revenue & profits here), it could also be normalizing it so it's between 0-1, this is useful when we have a big spread of data, however, in this case, it's not needed.

3 Results

3.1 All Plots

After the data cleanup, we can plot all of the variables against each other to see if any obvious pattern emerges. In figure 1 we can see an obvious perfect correlation along the diagonal which is expected because the data is being compared to itself. It also reassures us that the graphs are plotted correctly. The range of the data is quite big, so the plotted variables here are scaled logarithmically (otherwise we end up with big clusters).

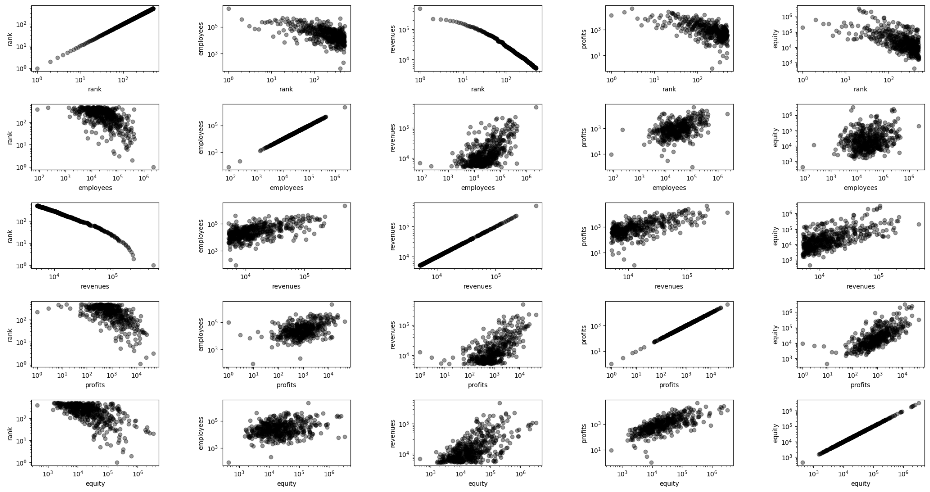


Figure 1: All variables plotted against each other

Industry and State data points are strings, if we convert them to numbers by calculating the frequency and then sorting (highest to lowest), then we change strings to the corresponding index. When plotted on the graph we get the result shown in Figure 2. It looks like state and industry are quite spread out, so when dealing with this small dataset it wouldn't help us much.

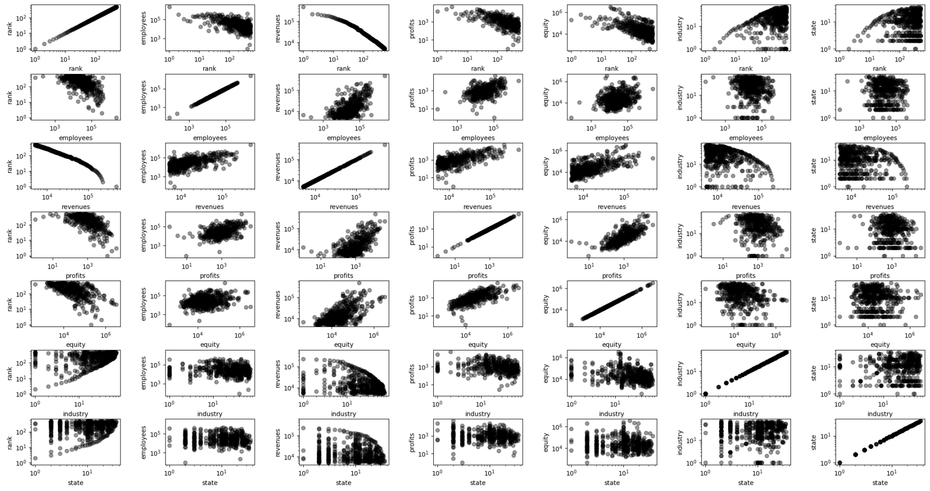


Figure 2: All variables plotted against each other including industry and state

3.2 Employees vs Profits

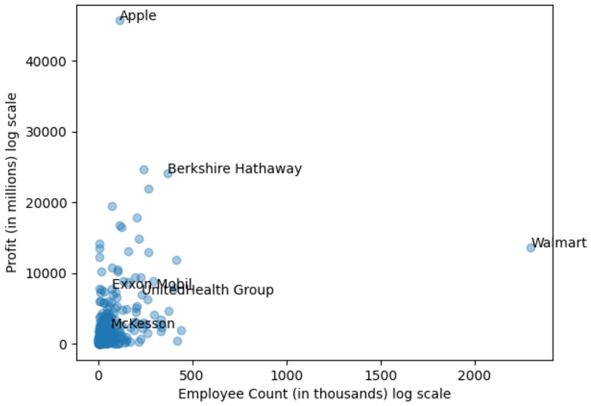


Figure 3: Employee Count vs Profits ($)

If we plot the number of employees against profits we can see that Walmart and Apple are on two extremes. In figure 3 we can see a plot of the raw data representing employee count against the profits of the company. It is very noticeable from the graph that most of the companies are cluttered up and two companies are on extremes. Apple has a high profit compared to employee count as they mainly operate through phone vendors and rely a lot on other companies selling their products. Whereas Walmart doesn't make as much money in comparison but has a lot of employees due to its physical stores and the need for more employees to manage them.

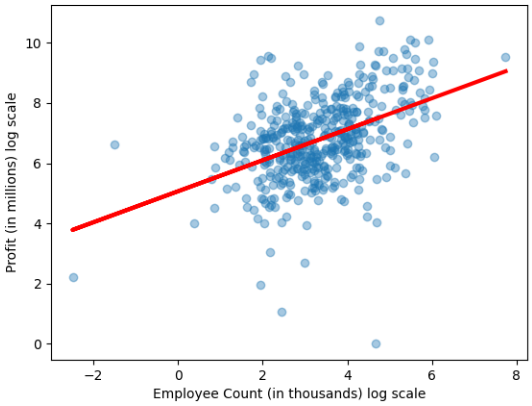
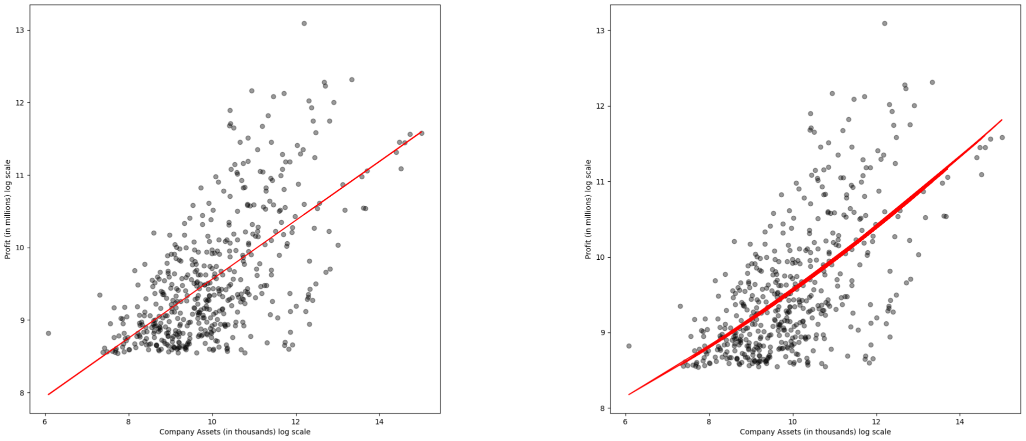


Figure 4: Employee Count vs Profits ($) (log scale)

Figure 4 shows a log scale of the same data, this is to help us visualize the data better as the values are quite large. It also lets us see the correlation a bit more clearly, there is a clearer cluster in the middle with few outliers. We can see that there is some positive correlation between employee count and the profit that the company earns, which makes sense as the bigger the company the more employees it can afford. The red line shows the linear regression model of the data. We cannot take negative values in a logarithm so all companies that had negative profits were completely removed.

3.3 Assets vs Profits



(a) Assets vs Profit (Linear)

(b) Assets vs Profit (Degree=2)

Figure 3.3 shows a plot of the raw data representing assets against the profits of the company. We can see that as the assets increase, the predicted profits of the company also increase. The shape seems to resemble some kind of polynomial. Figure 3.3 shows a comparison between two methods, one uses linear regression another uses polynomial regression with degree two. The value of r^2 we get here is 0.405 and 0.466 respectively. Using LASSO regression with degree 2 polynomial features gives us 0.406, and the same configuration with Ridge gives us 0.466. This value means that 40% of the variability is explained in the model.

3.4 Revenue vs Equity

At a first glance in the graph 2 we get a somewhat positive correlation when plotting revenue against equity. We create a model to test if there is any promising correlation. By performing simple linear regression we end up with figure (c). We can see that the variance is quite high and a linear model isn't really able to express that.

Type	r^2
Simple Linear Regression	0.4051
Lasso ($\lambda = 0.001$)	0.3912
Ridge ($\lambda = 0.001$)	0.3836
Lasso ($\lambda = 0.1$)	0.0
Ridge ($\lambda = 0.1$)	0.3915

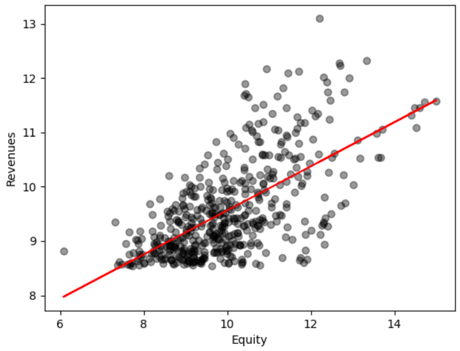
Table 2: Comparison between different hyperparameters

It looks like Lasso is giving a flat line with high values of λ. This could be because data is not well-suited for linear regression or the model is overfitting.

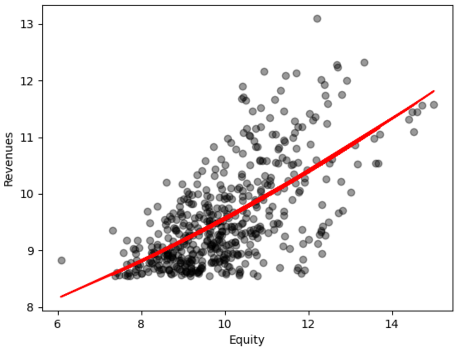
Type	r^2
Simple Linear Regression	0.4066
Lasso ($\lambda = 0.001$)	0.4065
Ridge ($\lambda = 0.001$)	0.4066
Lasso ($\lambda = 0.1$)	0.4060
Ridge ($\lambda = 0.1$)	0.4066

Table 3: Comparison between different hyperparameters on polynomial features

Lasso did much better on polynomial features, but it's still behind Ridge and a simple linear regression model.



(c) Revenue vs Equity linear features



(d) Revenue vs Equity polynomial features

3.5 Equity vs Profits

In figure 5 we can see a graph of equity compared to profits. There is quite a strong correlation shown by the red line.

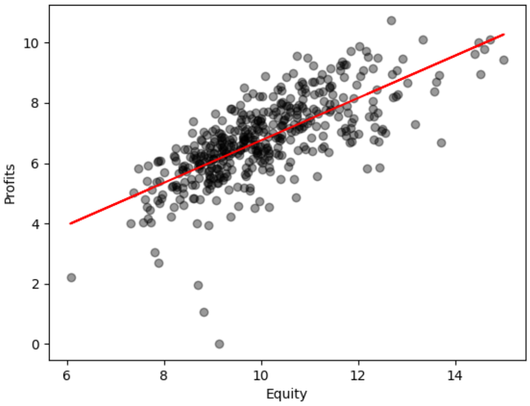


Figure 5: Equity vs Profits

Type	r^2
Simple Linear Regression	0.5117
Lasso ($\lambda = 0.001$)	0.5117
Ridge ($\lambda = 0.001$)	0.5117
Lasso ($\lambda = 0.1$)	0.5091
Ridge ($\lambda = 0.1$)	0.5117

Table 4: Comparison between different hyperparameters

There may be a moderate positive correlation between equity and profits, with most of the used methods calculating the r^2 value to be ≈0.51. This means that about 51% of the variation in the profitability of a company can be explained by the amount of equity it has. This relationship suggests that companies with higher equity tend to be more profitable, however, the relationship is not perfect so many other factors can affect it. This analysis highlights the need to consider other factors fully understand a company's success.

3.6 Everything Together

Even though there may not be many connections between the variables individually, computers excel at finding patterns in datasets consisting of thousands of dimensions.

I have used employee count, sector, state of the headquarters, revenue, revenue change, profit change, assets and total shares of equity to predict the profit. The dataset was split 70/30, training and testing respectively.

Regression Type	Scoring Method	Score
Simple	r^2	0.65065
	MSE	4 784 458.10
	RMSE	2187.34
	MAPE	5.53%
Lasso ($\lambda = 1000$)	r^2	0.64812
	MSE	4 819 025.20
	RMSE	2195.23
	MAPE	5.13%
Ridge ($\lambda = 1000$)	r^2	0.65061
	MSE	4 784 979.61
	RMSE	2187.46
	MAPE	5.48%

Table 5: Comparison between different regression types and their scores

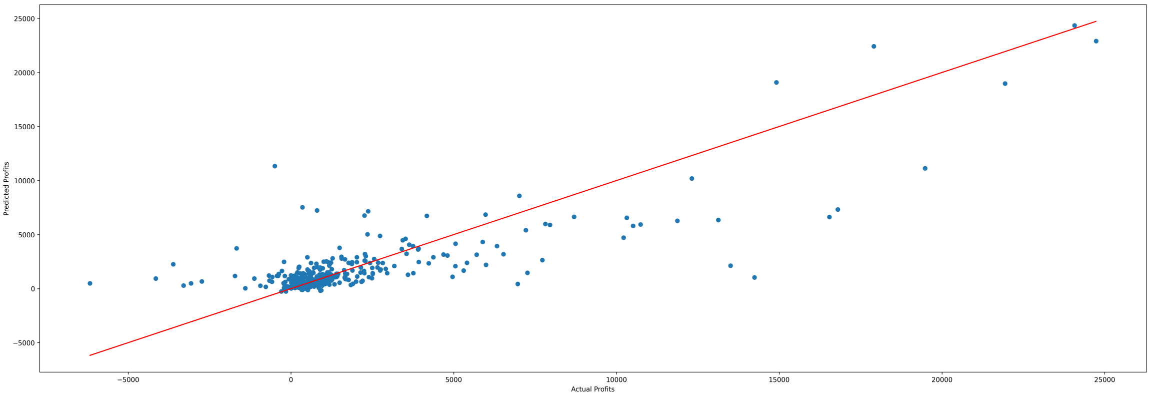


Figure 6: Graph showcasing actual profits vs predicted profits

In figure 6 we can see the result of training done on 498 companies. Walmart and Apple were removed due to their extreme values which heavily changed the values in the model. Before removing the outliers model, intercept $c = -279.67$, after $c = -244.96$. The dataset is quite small, so an outlier can heavily influence the model prediction process. This further emphasises the necessity to clean up and pre-process data to achieve more accurate results. A perfect model would map the blue points onto the red line. From the graph, we can see that the model has low bias and high variance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{1}$$

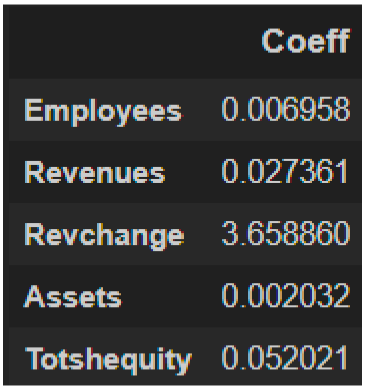
$$RMSE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i| \tag{2}$$

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{3}$$

Table 5 shows the results of training a model. The r^2 indicates that 65% of the variance can be explained by the independent variables, and 35% comes from other variables we don't have. The second row is Mean Squared Error (MSE) (eq. 1), it's the average error made by the model when predicting the variable. It might seem high but we are dealing with very large values so squaring them will make that look worse, because of that we can use the third row which is a root mean square error (RMSE) (eq. 2), which tells us the average error is around 2187.

The mean average percentage error (MAPE) (eq. 3) is used to calculate the difference between actual and predicted values as a percentage. The lower the MAPE value the better the prediction is. Even though our MAPE value is about 5.5% which is considered ”acceptably accurate” [2], the MAPE value in our case is not a good indicator of accuracy. A lot of points are clustered near ”0” and the average is heavily skewed towards that, meaning any points outside of that cluster will not move the error much.

Lasso and Ridge, λ values have been set to 1000 as any other value below yields identical values as the simple regression. Lasso performs a feature selection by setting some coefficients to 0, so it can be used to identify the most important features in the data, whereas Ridge will use all of the features given.



	Coeff
Employees	0.006958
Revenues	0.027361
Revchange	3.658860
Assets	0.002032
Totshequity	0.052021

Figure 7: Coefficients

Figure 7 shows how the model is influenced based on the input. It tells us that change in revenue heavily influences the expected profit which makes a lot of sense.

4 Discussion

After analyzing the data, it was found that there was no overwhelming correlation between the data. This suggests that the number of employees, amount of assets, level of revenue and amount of equity cannot reliably predict a company's success. These findings highlight the need for companies to carefully analyze their operations and closely examine the factors that drive profitability. It may be necessary to add more variables and use more advanced modelling techniques to more accurately predict a company's success. The dataset was also quite limited (500 entries and only 1 year), which is not enough to accurately predict anything.

This analysis provides a useful starting point for further research and emphasises the importance of carefully examining the variables and their relationship to understand the important factors.

References

[1] Aurielle Perlmann. `https://data.world/aurielle/fortune-500-2017`.

[2] David A Swanson. “On the relationship among values of the same summary measure of error when used across multiple characteristics at the same point in time: an examination of MALPE and MAPE”. In: *Review of Economics and Finance* 5.1 (2015).