Identifying the Relationship between Business Operations and Success

Aaron Goh and Ritesh Sharma

Data Mining Final Project

Table Of Contents

Project Goal

The goal of this project was to analyze factors in a business' success. The following metrics were used and defined relatively as follows:

Analyze Business Success Factors:

NOE = Number of Employees - Global (Latest)
R_D = R&D Expense [LTM] ($USDmm, Historical rate)
COE = Cost Of Revenues [LTM] ($USDmm, Historical rate)
TOE = Total Operating Expenses [LTM] ($USDmm, Historical rate)

In Relation to traditional metrics of "success":

EPS = Basic EPS [LTM] ($USD, Historical rate)
TEV = Total Enterprise Value [My Setting] [Latest] ($USDmm, Historical rate)
EBITDA = EBITDA [LTM] ($USDmm, Historical rate)

In successful analysis of the fields above, it may be possible to see that fields such as research and development expenses are strongly correlated with a business' EBITDA, EPS, or TEV or whether it is negatively correlated and/or more closely related to expenses (TOE, COE).

Through running regression analysis, it can better correlate to see how certain fields may influence each other and what might relate more than others. In this way, it may be possible to find an algorithm to model a business' expense model in relation to their R&D, number of employees, and how much they should be spending for Cost of Revenues, and the Total Operating Expenses.

Data Collection

The data explored were companies from the Health, Information Technology, and Industry data which was generated and filtered off of CapIQ. In the initial run, we were hoping to get the field for marketing expenses as well however, it led to a lot of missing data. In CapIQ, we were able to return out company information as the fields provided above.

However, there was a data cleaning step and all data entries missing data in some columns were removed. This reduced the overall dataset from about 15,000 to approximately 3,000.

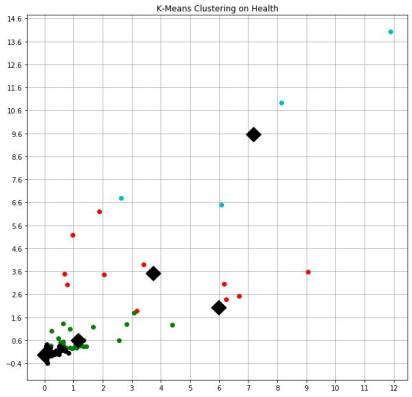
In processing the data, we ran it through the different sectors and then we ended up binning it by the Total Enterprise Value. In specifics:

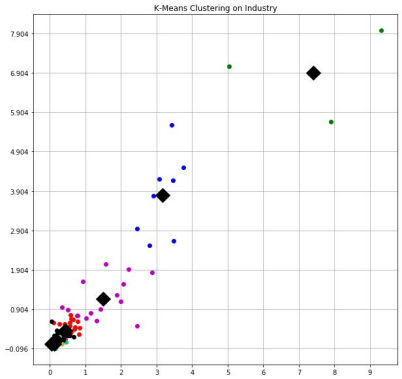
Small Cap Company is < $500,000 Total Enterprise Value
Mid Cap Company is >= $500,000 and < $2,000,000,000
Large Cap Company is >= $2,000,000,000

Analysis

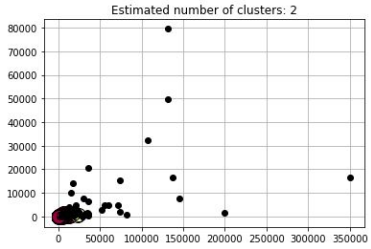
Initial Clustering

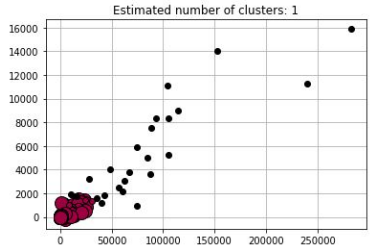
In first getting a sense of the data, clustering algorithms were run in order to try and see whether sets should be broken up or not. The initial approach of running k-means algorithm returned poor results. The data was not properly clustered and increasing k came back with visualizations that made little sense where some data points were in more sparse locations, data was split in a non-uniform manner between two centers.





Have observed a singularly dense cluster where the rest of the data was scattered, the next algorithm attempted was the DBSCAN. Differing epsilons were attempted. Using smaller epsilons ranged from ~ 5000 - 15,000, approximately two clusters were observed within each dataset's dense cluster location.





On initial glance, this suggested that there may be multiple clusters given smaller subsets of the data and in excluding "outliers" of the data, this may also improve the performance of the k-means algorithm.

The K-means algorithm and the DBSCAN were run in order observe if there were clear differences in the businesses scraped. The initial observation has demonstrated that there are a large number of outliers and due to the scale of distances observed, it is uncertain if subsets of the data may have key differences to better improve performance.

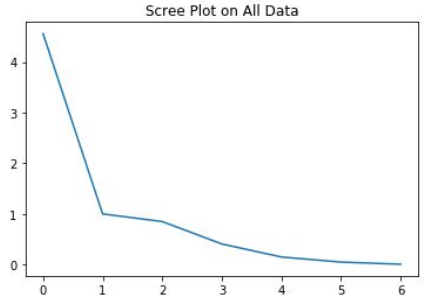
However, when running PCA next, we ended up finding that within the dense cluster, whether by running k-means or DBSCAN, it was effectively a single cluster with outliers.

PCA

From above, we decided to run PCA and see if there were components to reduce to and how the old fields projected onto the new principal components to detect if some fields correlated with each other on top of having another way to potentially filter data and gain better correlations.

From the first run on eigenvalues per dimension, we created a Scree Plot that showed the more ideal cuts for the dataset.

Scree Plot

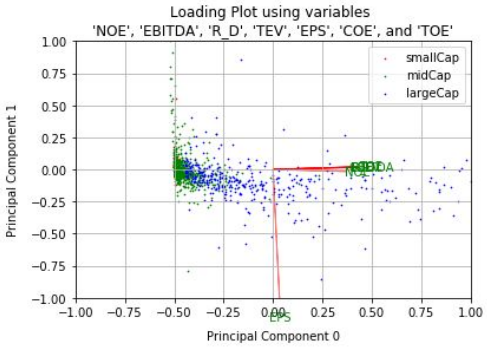


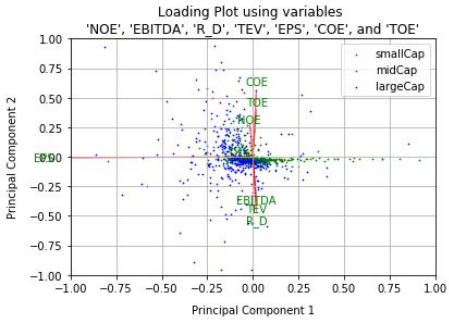
In running this analysis on the subsets, we found that there were differing results on each of the graphs for the suggested dimensional reduction. However, as explained later, 3 dimensions were chosen to maintain as much of the original data as analyzed by the explained variance.

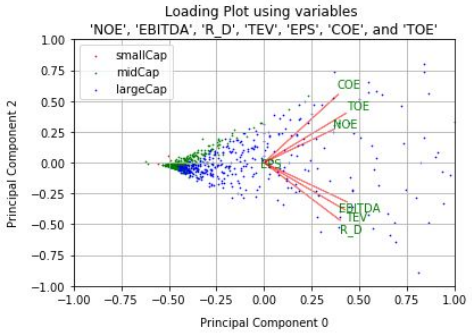
Combined Datasets

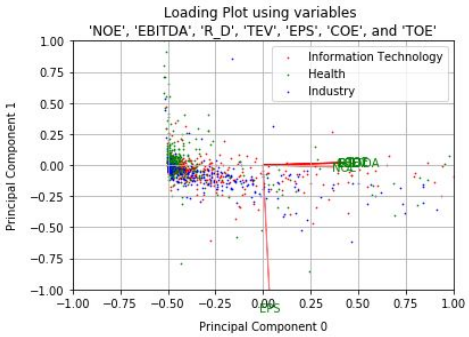
In running PCA on the overall set, we ended up seeing projecting the original fields into the new PCA components. Three dimensions were chosen as in the explained variance, that 92.5% of the original information was existent in those three dimensions.

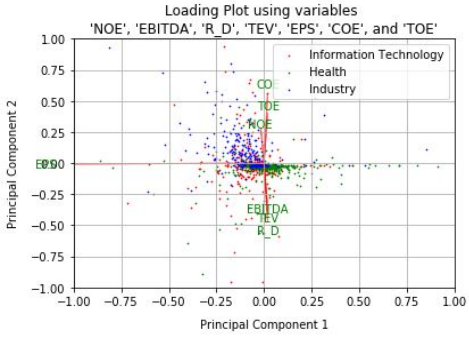
In the projections, it can be seen that the Number of Employees is generally associated with the Cost of Expenses and the Total Revenue Costs. However, the R&D Expense seems to be more directly tied with EBITDA and the Total Enterprise value.

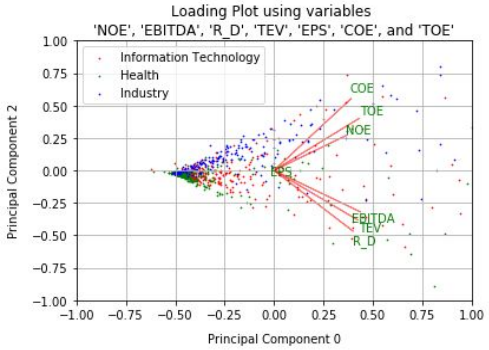










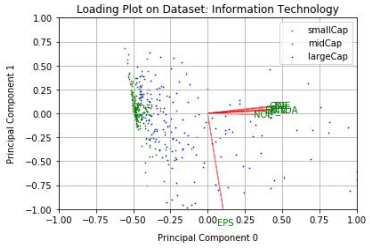


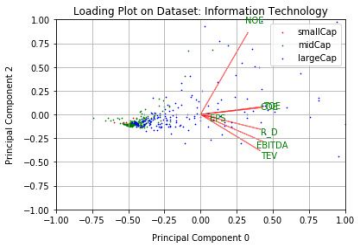
Looking at how the companies clustered as well by category, it can be seen that the industries are relatively well spread in with each other indicating decent similarity between the datasets in each dimension.

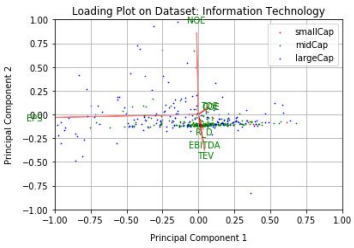
The Principal Component 1 appears to model positive growth and negative growth. In Principal Component 0, it seems that that Earnings Per Share is completely separated which might suggest that the other fields are actually not extremely relevant here or that the current dataset is not optimal to find a strong conclusion regarding EPS.

<u>IT Dataset</u>

In the IT Dataset, the following result is seen from plotting the data onto a loading graph. Once again, we choose the first three principal components as it keeps 93.2% of the original information by summing up the explained variance.







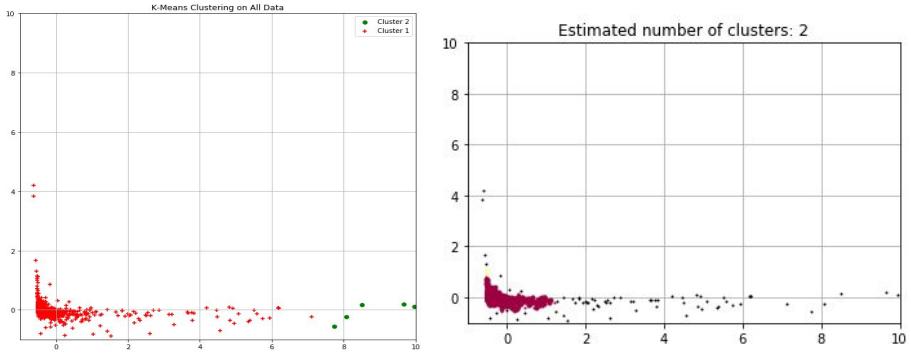
In the IT dataset, the results change a decent amount and it is seen that the number of employees doesn't seem to correlate with anything by the second principal component. The EBITDA and R&D expense seems somewhat close but not quite aligned.

Industry Dataset
Refer to Appendix

Health Dataset
Refer to Appendix

Final Clustering
After running PCA through the dataset, we use the newly reduced dimensions and then clustered the data based on the PCA coordinates. As can be seen, even after reduction, K-mean performed relatively poorly as the data still looks as it did at the start, with too many outliers.

DBSCAN Analysis



However, running DBSCAN proved decent and really served more as removing noise since the data suggested a single cluster at the dense region. In the above right graph, it can be seen that the core red cluster is of the most significance and gets rid of most data sprawling out of it.

Through this cleaning, we removed slightly more than 84 points that were identified as noise with the second cluster also negated as it happens to be barely off the main cluster.

After having done clustering on the dimensionally reduced set, we run regression back on the fields tied to the dense cluster's (red cluster) points identified in the DBSCAN algorithm.

Regression

To establish a more concrete relationship between the fields, we ran Linear regression on the datasets combined. Furthermore, we calculated the standard deviation (SD), the mean squared error (MSE), and the coefficient for each linear regression formula ran.

Initial Binning

Table 1.1(All Data Correlation)

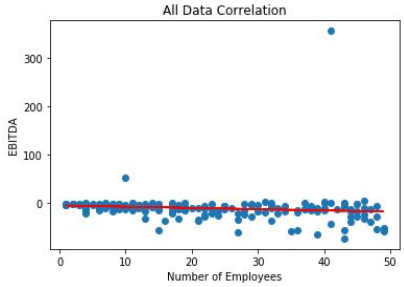
Small Cap Company ~ Small Company is < 50 employees
Mid Cap Company ~ Mid Company is >= 50 employees and < 300 employees
Large Cap Company ~ Large Company is >= 300 employees

Dataset sizes: the small cap company had ~90 datasets, medium cap company had ~210 datasets and large cap company had over 400 datasets.

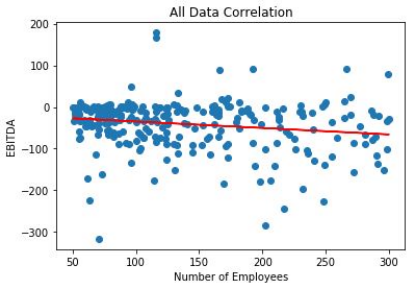
All Data Correlation:

EBITDA vs Number of Employees

Small Company Correlation



Mid Company Correlation



Large Company Correlation

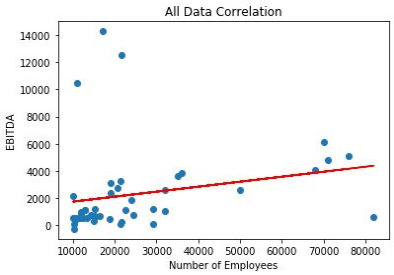


Table 1.1(Ebitda vs Number of Employees - All Data Correlation)

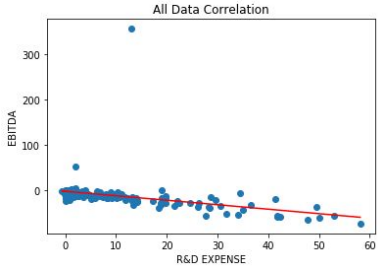
	Small	Medium	Large
Standard Deviation	22.27291905257935	42.52319712106389	251.2682535209533
Mean squared errors	96.8787800746932	338.5042852810875	9159.737782374
Coefficient	-0.25965826	-0.1552842	0.03665347

As we can see from Table 1.1, the MSE for EBITDA vs Number of Employees is smallest for small companies. Noticeably, small cap companies have fewer outliers compared to compared to medium and large companies. Since large companies have more outliers, the MSE and SD are higher especially since the range of employees goes from 300 to 50,000.
For small companies, the EBITDA and the Number of Employee has stronger correlation compared to large companies.

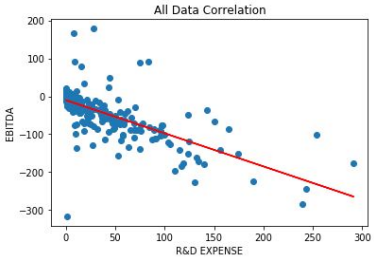
The dependent variable (Number of Employees) is most expected to increase in relation with EBITDA as a stronger factor than the other variables for small companies. However, it seems to lose meaningful correlation with larger companies.

EBITDA vs R&D Expense

Small Company Correlation



Mid Company Correlation



Large Company Correlation

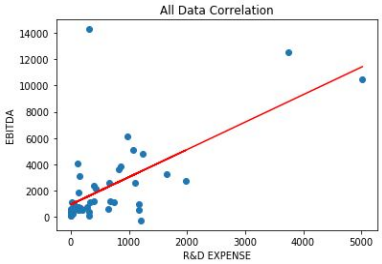


Table 1.2(EBITDA vs R&D Expenses - All Data Correlation)

	Small	Medium	Large
Standard Deviation	23.76658934878816	50.92068317954425	2605.306434307975
Mean squared errors	827.343087772335	1811.1169211317574	56184.0031118635
Coefficient	-0.98167909	-0.87341918	2.08890498

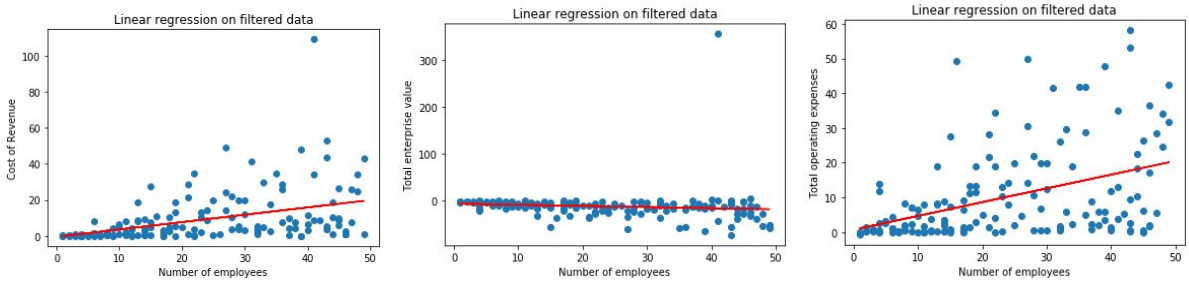
Looking at EBITDA related with R&D expenses, we see again that the SD and MSE increases from small cap to large cap company for the same reasons described for the number of employees.

However, a more interesting conclusion is that while some small and mid companies have more EBITDA, they still tend to spend less on research and development (R&D). Furthermore, The coefficients for small and mid companies are negative meaning that higher R&D investments in small/mid companies translates into a worse EBITDA. However, for large companies, the coefficient is positive which means that large companies have more EBITDA when spending more on research and development.

Total Operating Expense vs Number of Employee:

Refer to Appendix

Data filtered through DBSCAN PCA Cluster



Using the filter first established through the PCA approach, we can note that we have effectively reduced it to a merge between small cap and mid cap companies. It can be seen that the trends are still present. In the filtered data, it shows that there is overall no correlation between the number of employees and the total enterprise value.

The number of employees does increase the cost of revenue on the company and the total operating expenses has a high SD such that it is hard to state a strong correlation there as well given potentially unknown factors.

Conclusion:

From all the regression testing above, overall, we can see that predicting using linear regression is more feasible for small cap and mid cap companies compared to large cap companies. The MSE and SD is low for small and mid cap companies and high for large cap companies. Regression tests on the large cap companies suggest no correlation with number of employees and EBITDA.

For EBITDA and R&D expenses, the small and mid cap companies have small negative correlations, whereas large companies tend to have positive correlations.

Regarding Ebitda related to the number of employees, we found little to no correlation as the coefficients for the small indicated a constant trend, mid indicated a slight negative, and large indicated a slight positive.

Total Operating expenses and number of employee tends to have decent correlation for all small and large cap companies, which means more the number of employee more operating expense.

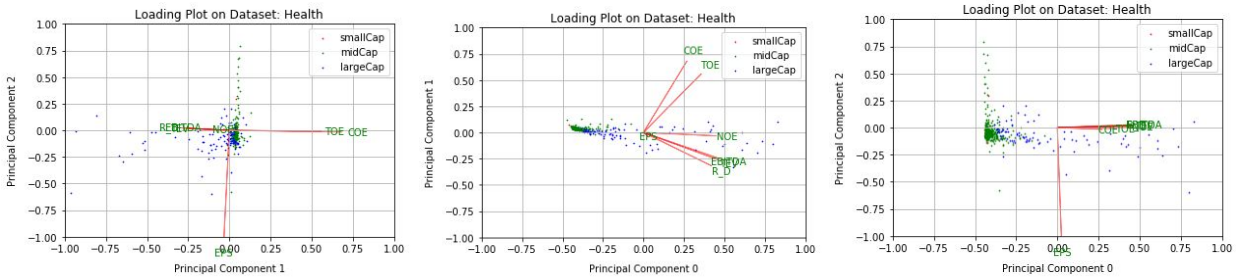
Hence we can conclude that EBITDA and R&D expenses are more negatively correlated for small and mid cap companies and not for large cap companies. The expenses are trend more heavily in relation to the number of employees.

However, the most interesting conclusion after all the analysis is that the number of employees does not correlate at all with the business EBITDA or Total Enterprise Value.

Appendix:

Health:

PCA Graphs for Health Dataset



Explained Variance Ratio
[5.88621234e-01 2.17023829e-01 1.42513863e-01 4.10004143e-02
7.78040320e-03 2.96065632e-03 9.95998694e-05]

Regression

Regression for Health and calculating mean_squared_errors, standard deviation and coefficient we obtained the following values

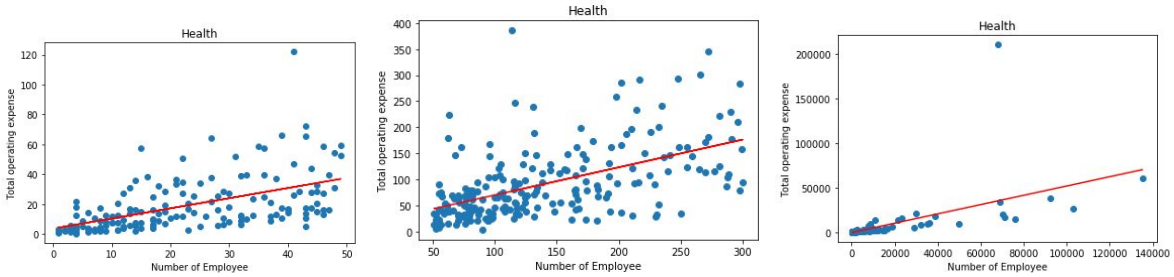


Table 1.1.3(Total Operating Expense vs Number of Employee - Health)

	Small	Medium	Large
Standard Deviation	14.67315904256822	55.23502165947591	13759.14855692272
Mean squared errors	240.9651041536837	3327.895343596862	185285.0794053
Coefficient	0.67934499	0.5319411	0.52140517

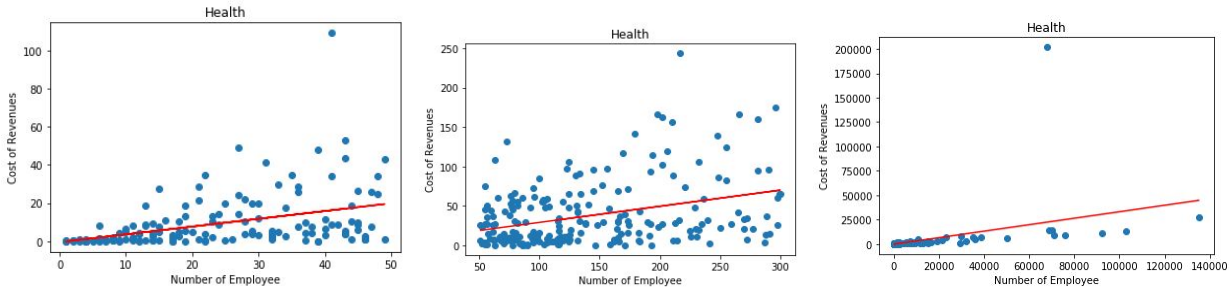
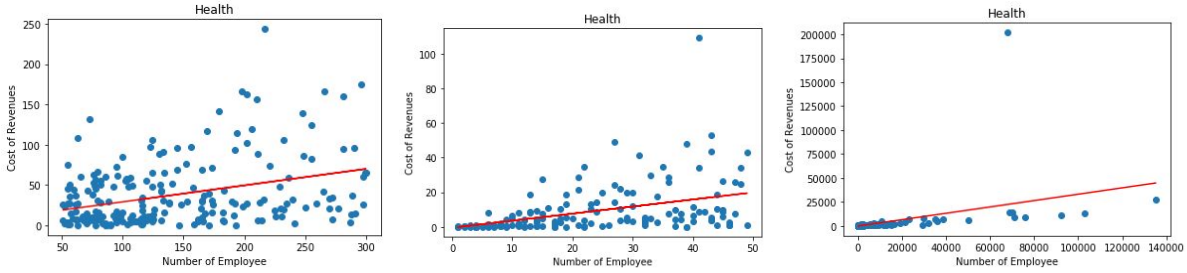


Table 1.1.4(Cost of Revenue vs Number of Employee - Health)

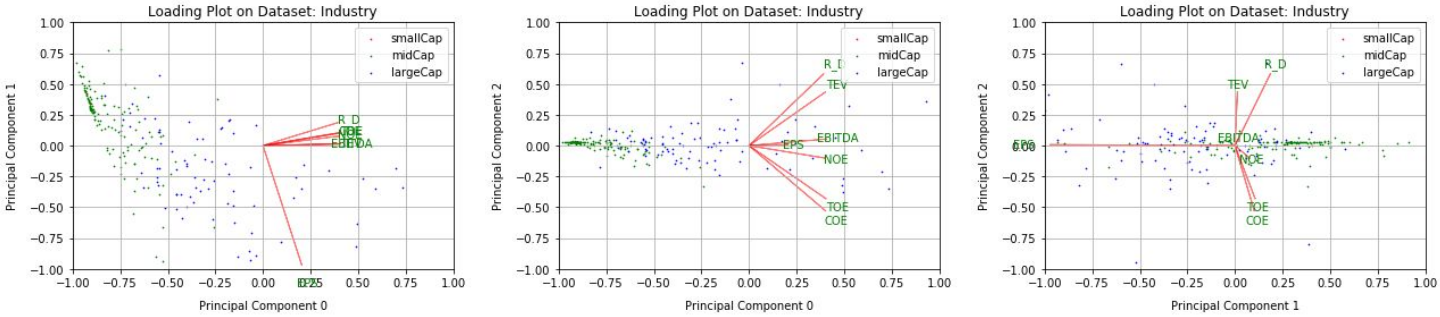
	Small	Medium	Large
Standard Deviation	10.74254472584246	30.45574289272034	161.0226687270
Mean squared errors	162.6455802919002	1445.2098557830911	188536.4364618
Coefficient	0.40727622	0.2044809	0.3327194



	Small	Medium	Large
Standard Deviation	17.12344472584246	40.251574289272034	115.3545456455
Mean squared errors	145.56456485484845	1578.6489454894	18561.548489454258
Coefficient	0.124554874	0.28945646	0.423414578

Industry:

For the industry dataset, when plotting the principal components into 2-dimensions. We see the following:



In the result above, it seems that R&D expense is still closely tied to the total revenue value. The Ebitda and the Number of Employees seem to be correlated and the fields demonstrating expenses seem to be isolated.

This may be due to the nature of employees in Industry can be tied closer to pre-made automation and labor compared to the other industries such as IT or Health. An interesting notice is that the EPS typically serves to be the 0 point in these industries.

IT:

Regression on just IT dataset and calculating MSE, SD, and coefficient we obtained the following values.

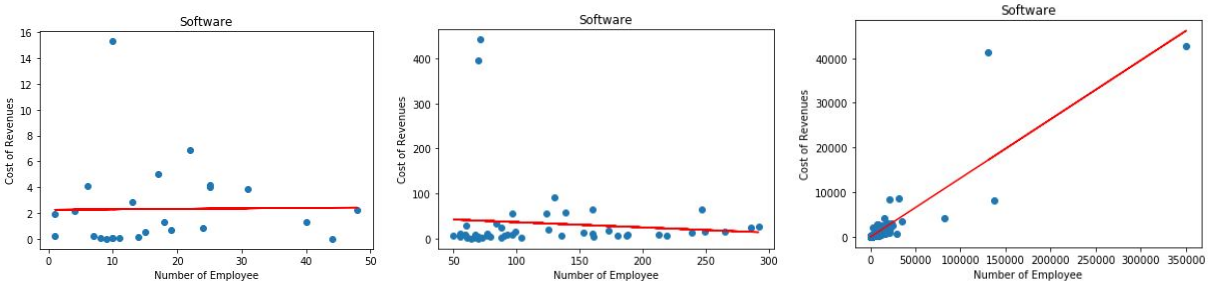


Table 1.1.1(Cost of Revenue Vs Number of Employees - Software)

	Small	Medium	Large
Standard Deviation	2.2950171788119884	58.28840243927706\	415.285049206855
Mean squared errors	10.53028450271572	6668.762756471126	41801.034368634
Coefficient	0.0035556	-0.11610375	0.13191291

Cost of Revenue and Number of Employee has more correlation for small capped companies and tend to have less correlation as the size of the company increases. Coefficient for medium

cap companies is negative which means even if the company has less number of employees the cost of revenue is high. For large cap company the cost of revenue and number of employee seems to to linear, as the number of employee increases the cost of revenue also increases. The error is small for small cap companies and more for large cap company this is because more outliers are present in large cap company, which makes it harder to get a better prediction for large cap companies while compared to small cap companies.

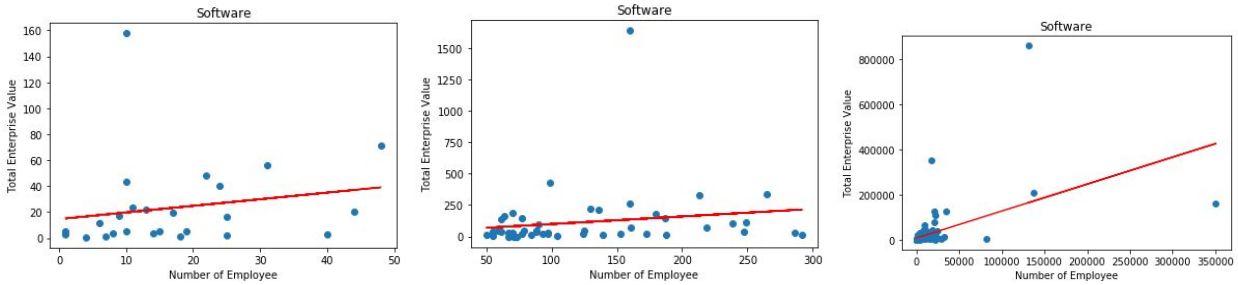
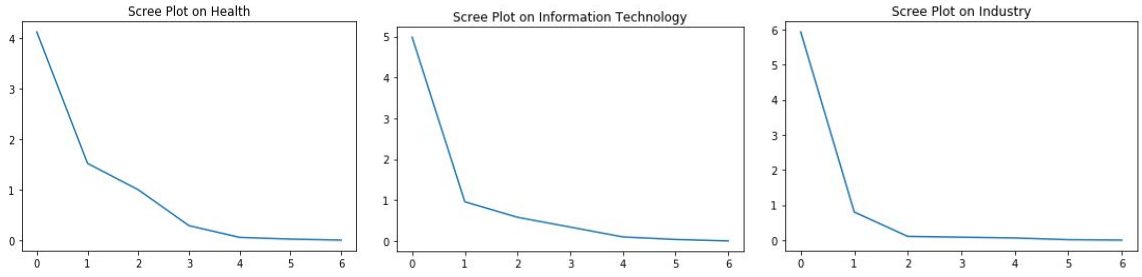


Table 1.1.2(Total Enterprise Value Vs Number of Employees - Software)

	Small	Medium	Large
Standard Deviation	24.015308287971177	173.1270994027400	5930.22075338172
Mean squared errors	1072.990834278606	5657.86452348639	4029134.158179
Coefficient	0.50925439	0.60131453	1.19374278

Scree Plots:

Total Enterprise Value and Number of Employee gives similar results to Total Enterprise Value and Number of Employees. The small cap companies has more correlation while compared to large cap companies.



From the scree plot, it seems that there is definitely some redundancy in the data which is expected as each field should denote a representation of the company's values and expenses. Loading plots demonstrating correlations of fields within principal components should show

meaning. As a sanity check in the loading plots ahead, the Cost of Revenue and the Total Operating Expenses should be closely aligned in each case.

All Data

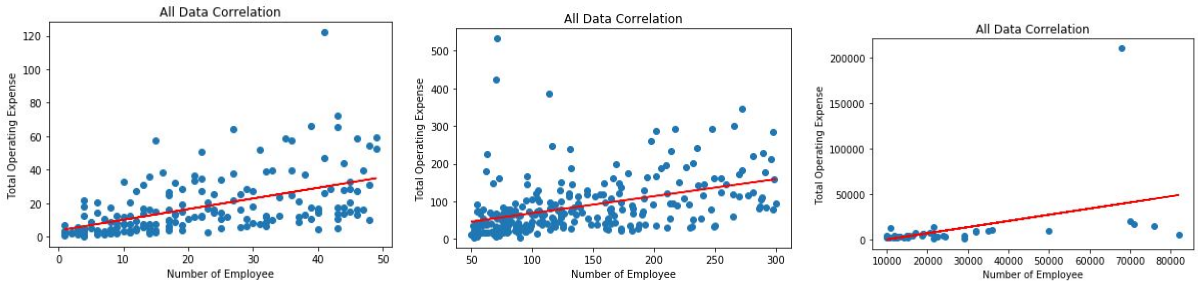


Table 1.3(Total Operating Expense vs Number of Employee - All Data Correlation)

	Small	Medium	Large
Standard Deviation	13.98738199412392	57.07994738606533	l231.827185307928
Mean squared errors	227.2944977519581	459.448189094793	76133.5171077107
Coefficient	0.63656769	0.45174431	0.67972426

Total Number of Employee and Total Operating Expenses seems to have a positive correlation. The coefficient for all small, mid and large companies are in the same range which means that as the number of employee, increases the Total Operating Expenses increases with some exceptions.

Some mid and large companies tend to have less number of employees and have more total operating expenses, this can be seen in the Table 1.3 above, the mean squared errors and standard deviation seems to be high for mid and large cap companies. This likely differs by industry as a company hiring minimum wage employees will likely have high variance compared to a company such as a software firm.