STAT 306 Group Project Final Report

Ben Huckell - Jiale Wang - Grace Tu - Long (Luke) Nguyen

Introduction

The start of the COVID-19 pandemic brought unexpected changes to many individuals' financial health. The sudden loss of numerous jobs was accompanied by a surge in new investors (Lush, Fontes, Zhu, Valdes, & Mottola, 2021). Furthermore, with the transition to online platforms in both education and the corporate world and limitations to in-person interactions, the information technology sector has seen the largest growth in the stock market (Doorn, 2020). In light of this heightened interest and positive sentiment for technology stocks, our report aims to explore the company fundamentals behind some of the most prominent technology companies and develop a statistical model to predict how these factors can impact the company share price — an index of a company's value. With the data available, we chose to examine four leading technology companies, namely, Facebook, Apple, Microsoft, and Oracle (FAMO). The growth of technology companies such as FAMO have contributed significantly to the stock market and overall economy. Thus, we believe that exploring effects of company fundamentals on FAMO share prices would provide insight into better understanding the factors that contributed to the success of these top technology companies and allow us to use models to make predictions about the future share prices for FAMO companies, which can be valuable to investors.

Our data is from the CompuStat database provided by the University of Pennsylvania Wharton Research Data Services (WRDS) Center, which contains a variety of fundamental data about public companies, including FAMO. The WRDS Center records its data using annual 10-K report filings and 10-Q report filings - publicly available reports containing extensive

information on the companies' fundamentals filed by all public companies as legally mandated by the U.S. Securities and Exchange Commission (SEC).

Our dataset includes the following variables:

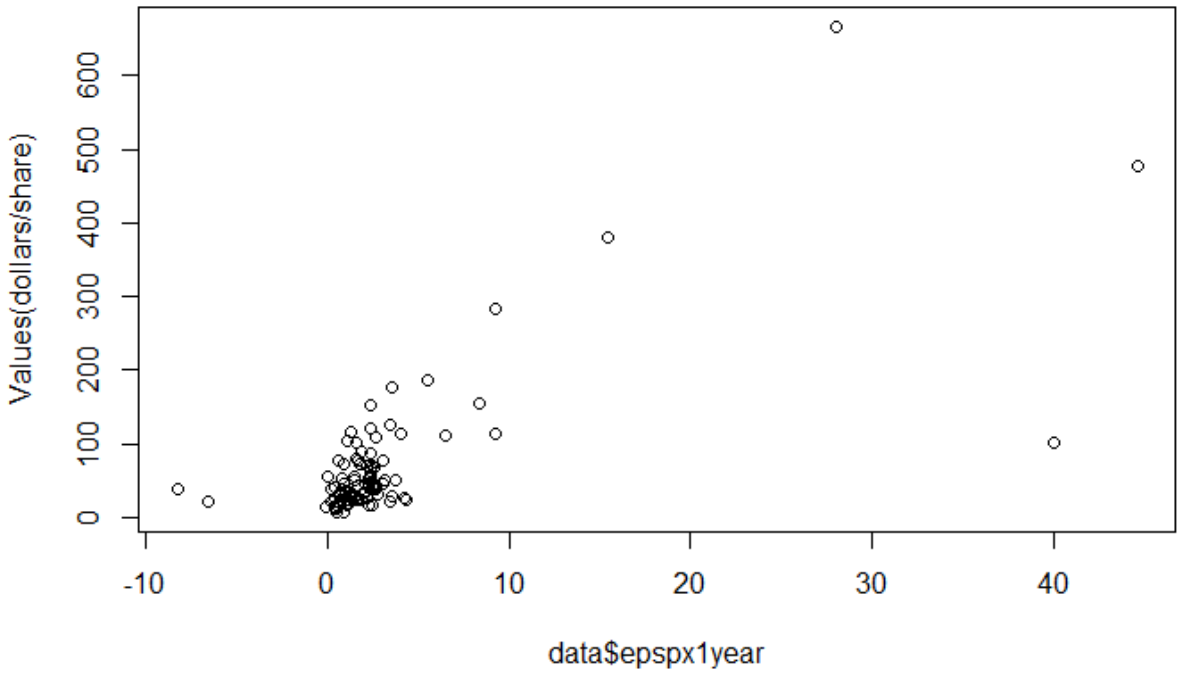
- [fyear] - the year in which we want to predict the share price; a numerical variable from 1982 to 2017.
- [tic] - the company ticker symbol (abbreviation of a company name); a categorical variable with 5 categories. "AAPL" corresponds to Apple, "ORCL" corresponds to Oracle, "MSFT" corresponds to Microsoft, and "FB" corresponds to Facebook.
- [values] - the share price; a numerical variable in USD. This is our response variable.
- [epspx1year] - the earning per share (EPS) from the previous year calculated as total profit divided by the outstanding number of shares; a numerical variable in USD/share. An EPS with growth over time and a larger value can potentially be indicative of a profitable company
- [dvt1year] - the sum of dividend payments over the previous fiscal year; a numerical variable in USD. A small or no dividend payment could indicate a growing company, while a large dividend payment could indicate more stability in the company.
- [ebitda1year] - total profit (i.e. revenue minus expenses) before applying interest, depreciation, tax, or amortization from the previous year; a numerical variable in USD which can provide a better metric for company comparisons while eliminating the effects of financing, government, or accounting decisions.
- [prcc_f1year] - the share price of the previous year; a numerical variable in USD.
- [revt1year] - total annual revenue of the previous year; a numerical variable in USD.

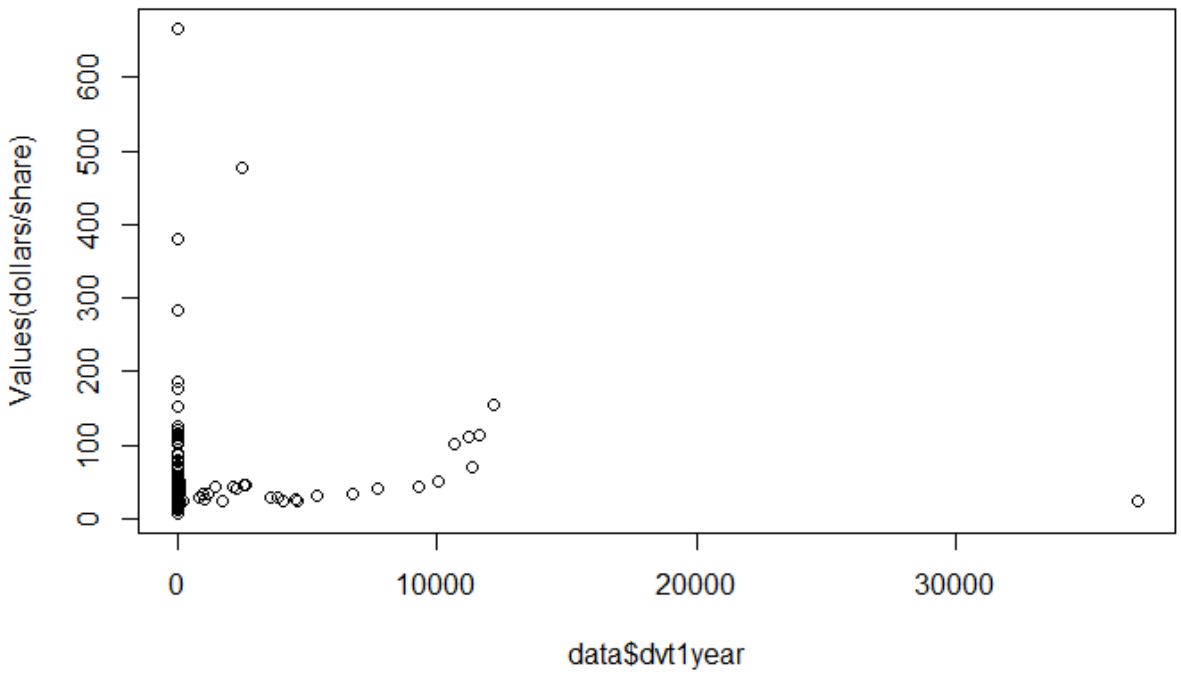
This subset includes some of the most revealing fundamental factors which aim to provide a comprehensive look into the companies' financial strength. In order to explore a causal effect on share price, all of the factors are lagged 1 year into the past. Therefore, in essence, we regress last year's annual data, with the current year's share price, for years from 1982 to 2017.

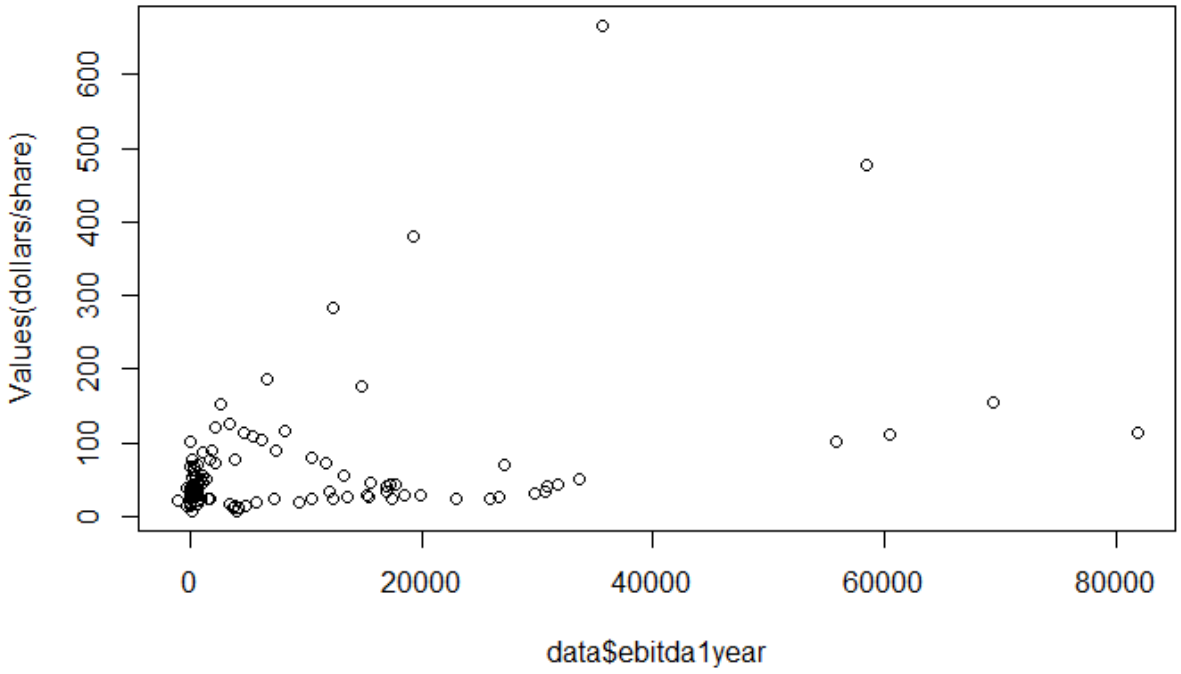
Analysis

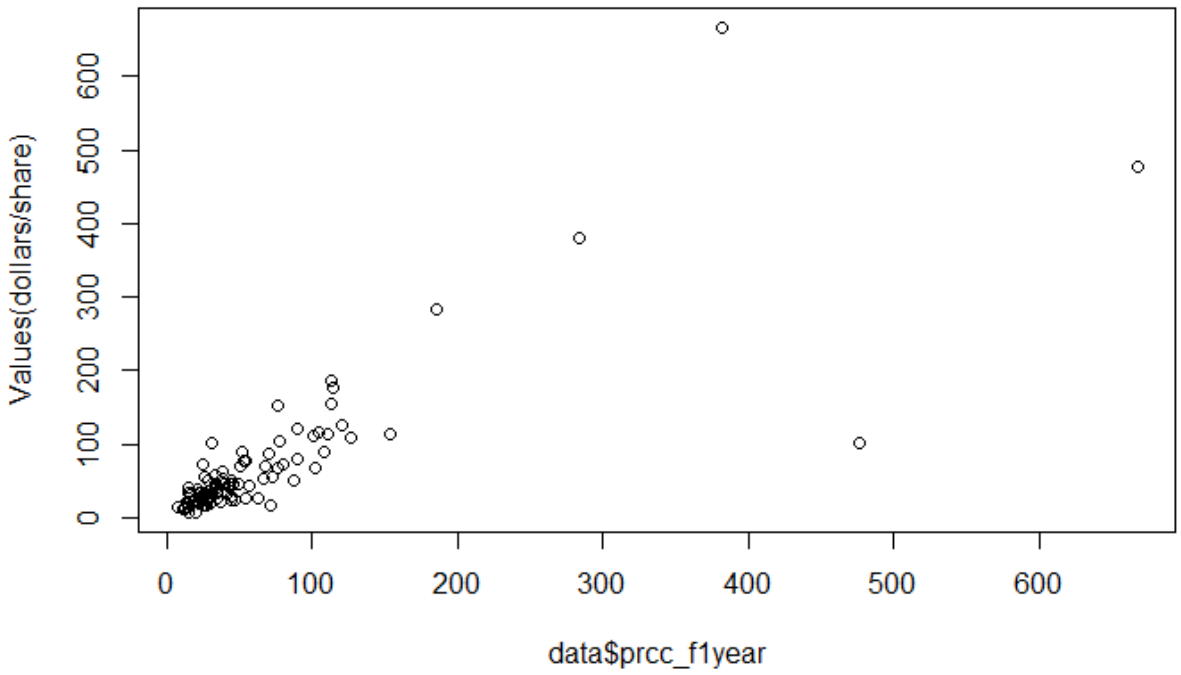
Initial visualizations for exploring the data:

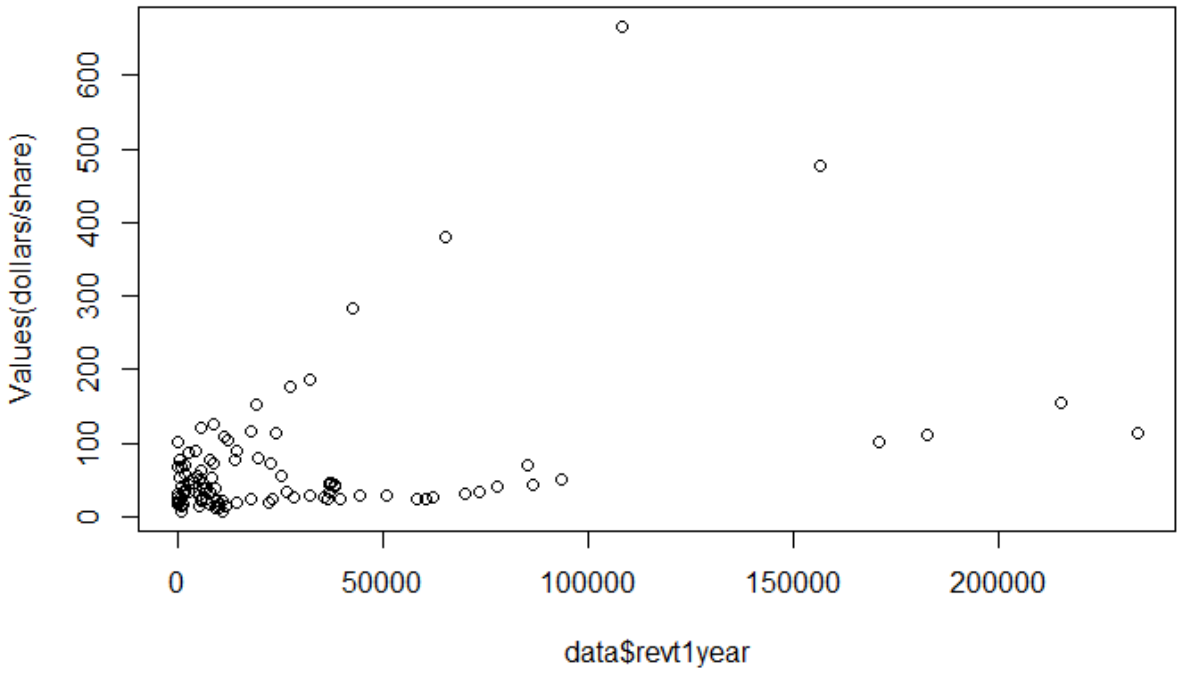
The following are scatterplots of share price (Values) against other numerical explanatory variables in the data except for fyear, as well as a boxplot of company tickers against share price values:

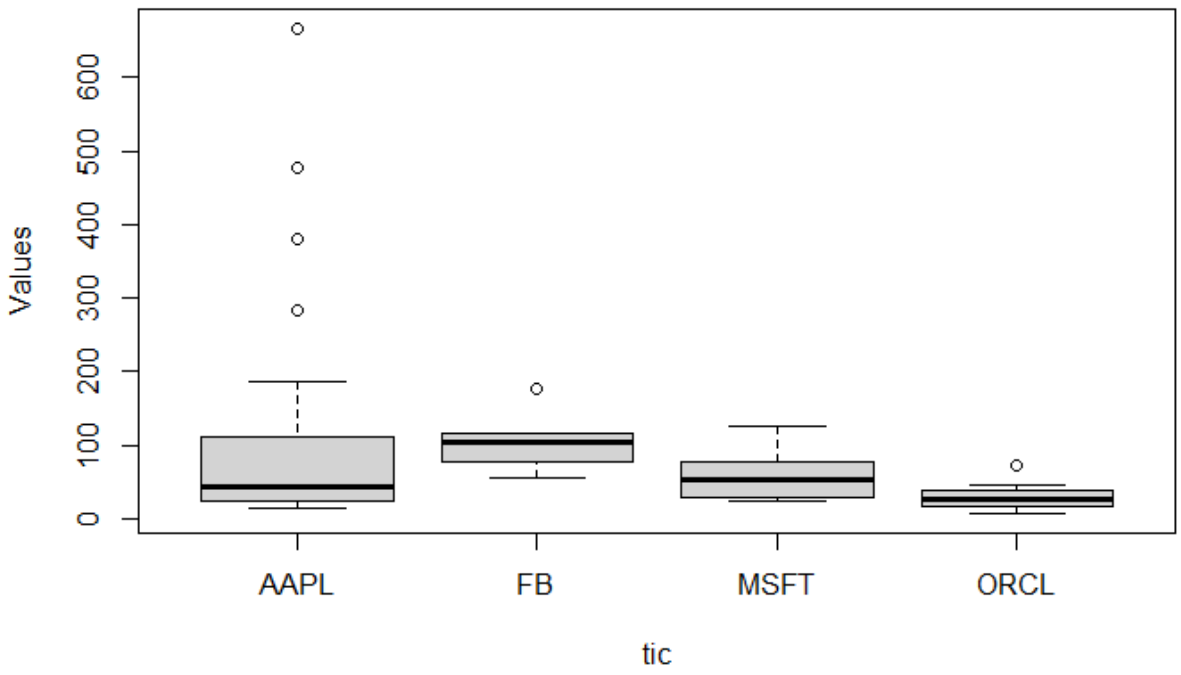



Values(dollars/share)
0
100
200
300
400
500
600
0
10000
20000
30000
data$dvt1year


Values(dollars/share)
0
100
200
300
400
500
600
0
20000
40000
60000
80000
data$ebitda1year


Values(dollars/share)
0
100
200
300
400
500
600
0
100
200
300
400
500
600
data$prcc_f1year


Values(dollars/share)
0
100
200
300
400
500
600
0
50000
100000
150000
200000
data$revt1year



We see that in a lot of cases, data points are very concentrated near the bottom left of the plots, whereas there are only a few data points that span a much larger range. At first glance, these do somewhat look like they could be outliers. However, given the high growth potential of the companies we are seeing, we do run the first of high spread in our data - which is exactly what we are seeing. This problem will be addressed later on when applying transformations to the data.

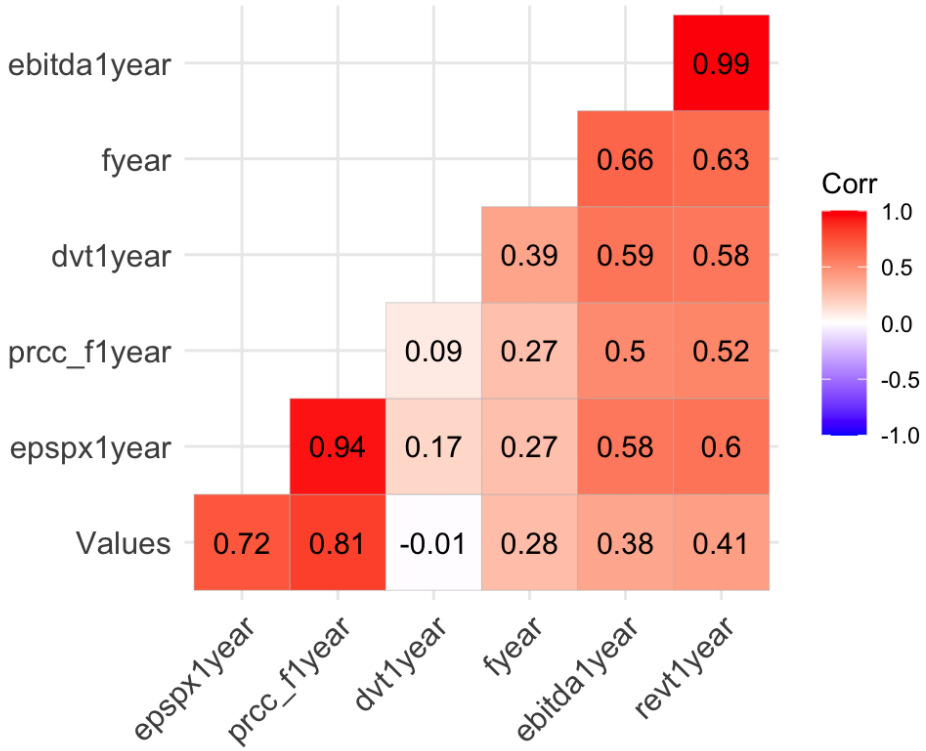
To deal with this problem we did consider applying transformations such as square root or log transformations to the explanatory variables which showed this pattern such as revenue, ebitda, and EPS. We could also attempt to square the values, just to be thorough. The one issue here (log/sqrt) is that negative values are possible for both EPS and EBITDA, therefore we apply a normalization technique to map values between 0 to 1. This is necessary to ensure that all values are positive.

Regarding dividends, we see that many of the data points are 0. This is because some of the companies chosen do not give out dividends at all. As a result, we do not expect this parameter to be at all helpful in our analysis.

Finally, we see that when we plot the boxplot of company ticker against price value, that there are differences, but the differences do not appear to be very large. In addition, we hesitate to add ticker as a categorical variable because ultimately we are trying to use a company's fundamental indicators to predict price in a sector. Changing our prediction based on an individual company does not help us achieve this.

In general, for the numerical explanatory variables graphed above, there appears to be an increase in share price value with an increase in explanatory variable value.

Correlation heatmaps were also created to assess any multicollinearity present in the explanatory variables:



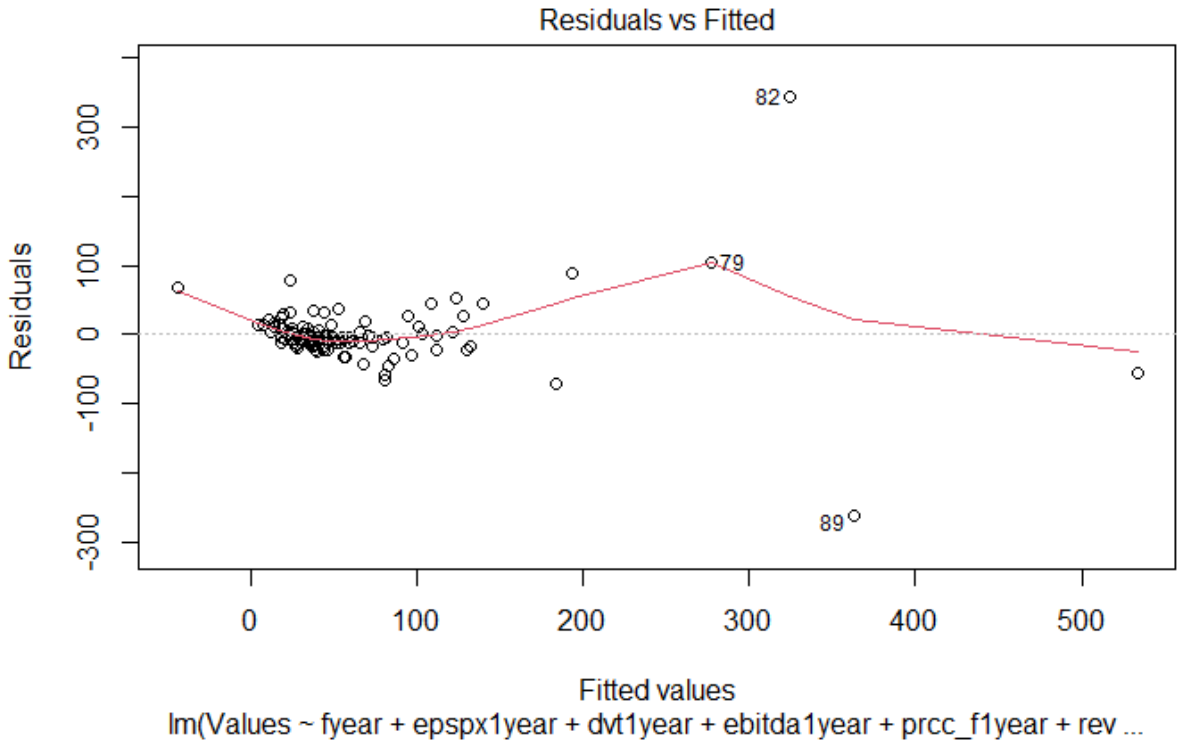
Using the above correlation heatmap, we can give a thorough look to any possible strong correlations between our explanatory variables which could give rise to issues stemming from multicollinearity. We see that some strong correlations are present such as EBITDA with revenue and EPS with the share price at the time of measurement. Otherwise, we see some correlation between many of the explanatory variables, that are worth noting.

In addition, from preliminary results, we can generate hypotheses about which factors we expect to be the most influential by looking at which explanatory variables are most correlated with "Values" or the current share price. The strongest factors are EPS with a 1 year lag and last year's price.

We do acknowledge that there are lots of factors that do show correlations, and therefore in our further analysis we make sure to not include all parameters, only those that produce the best models.

Methodologies:

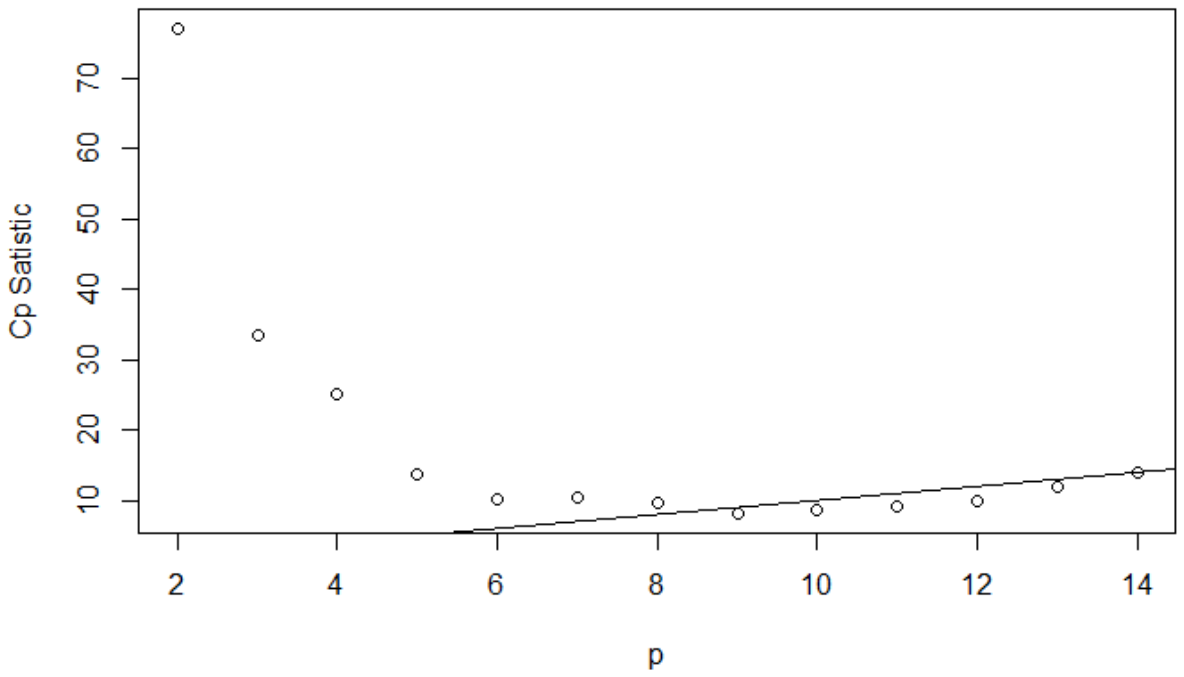
A general linear model including all explanatory variables was first fit on the data after which plots of residuals were created. Along with visualizations used to explore the data above, these residual plots were used to add and/or modify existing explanatory variables as necessary.



The residual plot still shows that the data points are concentrated towards the middle left of the residuals plot, a trend similar to the ones seen from the data visualizations. To further address this problem, the following variables were added: epspx1year^2, sqrt(epspx1year), ebitda1year^2, sqrt(ebitda1year), log(revt1year), revt1year^2, sqrt(revt1year).

A new general linear model including all explanatory variables and newly added variables was fitted. This model will be referred to as the full model.

Regsubsets was then used on the full model to determine which parameters were optimal at different model sizes. The model's respective Cp statistics were calculated and used to select the top few most accurate models. The graph of Cp statistics is shown below:



The best subset of variables at different model sizes:

	Variables included in the model									
Model p	epspx1 year^2	epspx1 year	sqrt(epsp x1year)	prcc_f 1year	revt1year^ 2	log(revt1 year)	ebitda1year	sqrt(ebitda1 year)	fyear	dvt1year
8	✓	✓	✓	✓	✓	✓			✓	
9	✓	✓	✓	✓	✓		✓	✓	✓	
10	✓	✓	✓	✓	✓		✓	✓	✓	✓
11	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Two methodologies were applied to these four models to determine which model is best.

The first method was to compute the Adjusted R-squared, AIC, and Cp values of the models:

p of the model	8	9	10	11
Adjusted R squared	0.792	0.7973	0.7985	0.7994

AIC	1078.692	1076.899	1077.18	1077.595
Cp	9.762	8.262	8.716	9.314

In terms of the Adjusted R-squared, the model with p = 11 is the best as it has the highest Adjusted R-squared. However, looking at AIC and Cp values, the model with p = 9 is the best since it has both the lowest AIC value and the lowest Cp value. Considering that two of the criterion-based methods support the nine parameter model and by the principle of parsimony, choosing the model with p = 9 seems to be a reasonable option.

Since we are also interested in prediction, we also used a second method, “leave-one-out” cross validation, to explore which model has the lowest prediction cost:

p of the model	8	9	10	11
Estimated Pred. Cost from Cross Validation	5060.674	4437.280	6517.926	6869.447

The model with p = 9 yielded the lowest estimated prediction cost, thus, we conclude that the model with p = 9 is the best model for predicting share prices in our data. The following is our best model:

E(Values) = -2.938e+03 + 1.511*fyear - 1.898e+03*epspx1year^2 + 2.495e+03*epspx1year - 9.745e+02*sqrt(epspx1year) + 1.140*prcc_f1year - 7.371e-09*revt1year^2 + 6.999e+02*ebitda1year - 4.746e+02 * sqrt(ebitda1year)

Conclusion

After examining both the criterion-based and cross-validation methods, the model with 8 variables — fyear, epspx1year, epspx1year^2, sqrt(epspx1year), prcc_f1year, revt1year^2,

ebitda1year, sqrt(ebitda1year) — produced the best overall results for maximizing explanatory power and minimizing prediction error. Our best model has an R-squared value of 0.8131, an Adjusted R-squared value of 0.7973, a Cp value of 8.262, an AIC value of 1076.899, and an estimated prediction cost from leave-one-out cross validation of 4437.280. The following is the R output of the summary of our coefficients:

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.938e+03	1.528e+03	-1.923	0.057514 .
fyear	1.511e+00	7.673e-01	1.969	0.051896 .
I(epspx1year^2)	-1.898e+03	2.458e+02	-7.719	1.17e-11 ***
epspx1year	2.495e+03	4.720e+02	5.286	7.95e-07 ***
sqrt(epspx1year)	-9.745e+02	2.438e+02	-3.997	0.000127 ***
ebitda1year	6.999e+02	2.845e+02	2.460	0.015687 *
sqrt(ebitda1year)	-4.746e+02	1.866e+02	-2.544	0.012570 *
prcc_f1year	1.140e+00	1.574e-01	7.245	1.13e-10 ***
I(revt1year^2)	-7.371e-09	2.577e-09	-2.861	0.005195 **

From analyzing the importance of the variables, we note that the variables most significant in predicting the share price value are the company's squared EPS (earnings per share), EPS, square root EPS, and the share price in the previous year. The revenue, while significant, has a coefficient of essentially zero. The year does not appear to be a significant predictor, and the intercept has a nonsignificant coefficient with a value of nearly zero. A key point to note about our model is that many of our variables are mathematically related. Given the nature of financial datasets, many of the variables present do show high levels of correlation,

which is an important caveat to note. However, these patterns are rather typical of financial datasets where strong companies tend to perform better across all fundamental metrics, rather than just a select few. Overall, we acknowledge that the presence of correlated variables does pose a risk to our modelling, but we did take sufficient steps within our analysis to ensure that our selected models were minimally affected.

In conclusion, we have determined that it is possible to predict the share price of a company one year into the future using current fundamental metrics, with a reasonable degree of accuracy. Previous year earnings per share and share price were found to be the most significant indicators for predicting share prices of the current year. Variables related to the overall revenue of the company, like EBITDA or the total annual revenue from the previous year, are not as impactful in determining the current share price value but are nonetheless significant and should be included in the model. Using these variables along with the current year, our model is able to explain about 81.31% of the variability in share price values in the data.

References

Doorn, P. (2020, June 21). Here are the best and worst stocks during the first 100 days of the coronavirus pandemic. Retrieved April 10, 2021, from https://www.marketwatch.com/story/here-are-the-best-and-worst-stocks-during-the-first-100-days-of-the-coronavirus-pandemic-2020-06-17

Lush, M., Fontes, A., Zhu, M., Valdes, O., & Mottola, G. (2021). *Investing 2020: New Accounts and the People Who Opened Them* (February ed., Consumer Insights: Money & Investing, Rep.). Chicago, IL: FINRA Investor Education Foundation. Retrieved from https://www.finrafoundation.org/sites/finrafoundation/files/investing-2020-new-accounts-and-the-people-who-opened-them_1_0.pdf