Spring
2024

Exploratory Data Analysis on Stock Ratio Data

ERIC S. SCHNEIDER

I: Introduction:

In order to ensure that a large enough dataset is present, and that the most current ratio data is found, I elected to create my own dataset for this analysis and overall project. The main component of the dataset was made possible by a list of ticker symbols for companies whose stock is sold on the New York Stock exchange. After randomly removing a small number of ticker symbols to ensure that a simple random sample was used, the program took each of the remaining ticker symbols, used the yfinance package in python to pull the 2023 financial data for the company, and used said financial data to calculate the given ratios for the predictive model. From there, the ratios were appended to a list which was then propagated to the stock ratios dataset. Because of the fact that this dataset was created using a program of this type, there is little cleaning, which is needed, outside of removing individual values which hold no data. After this is done, a dataset with just over six thousand cases is yielded for use in the predictive model. This size is well over the minimum for an effective predictive model, and measures can be taken in the revision phase of the model to nearly double the size of the initial dataset.

II: Dataset Description:

Each case in the dataset represents a single stock which is sold on the New York Stock Exchange, as well as its corresponding ratio data for 2023. The variables which will be used in the predictive model are contemporary ratios meant to show the current financial strength of a company and aid in predicting future income and stock price. These ratios in particular were chosen for the fact that they each represent a given sector of ratio analysis to minimize the possible bias due to interaction of variables. The ratios to be used in the predictive model are as follows:

- *Return on Equity:* This ratio measures a company's net income divided by its shareholders' equity. It is used to help gauge the profitability of a company, and how efficiently they are able to generate profits using invested capital.
- *Current Ratio:* This Ratio helps show the general short-term liquidity of a company. It is calculated by dividing current assets by current liabilities. In the short term, the current ratio shows how well the given company can finance their short-term debt with their available liquid capital.
- *Asset Turnover Ratio:* This ratio is found by dividing net sales by the total assets of a company. It is meant to show the long-term profitability of the company relative to the value of its assets. This ratio

also shows the ability of a company to effectively use its assets for their revenue generating purpose, as abnormally large depreciation and impairments on assets can dramatically decrease asset turnover.

- *Weighted Average Cost of Capital:* This ratio represents the average after-tax cost of capital from equity and debt sources. In a stock valuation analysis, the WACC stands in as the implicit interest rate for the company's future operating income. This ratio is rarely used in contemporary ratio analysis, but as its main function is to gauge the volatility of stock value from changes in earnings and capital, it is considered very useful for this model.
- *Dividend Yield:* This ratio is found by dividing the total dividend issuance of a company by its stock price. This is important because, while the presence of dividends may shock demand at times, it actually has a stagnating effect on long-term stock prices. The inclusion of the dividend yield helps remove its bias from the regression model, thus eliminating it as a possible confounding variable.
- *Earnings Yield:* This ratio is found by dividing the earnings per share of a company by its share price. This ratio acts as a baseline for percentage yield, as when supply and demand factors are removed, the value of a stock is equal to the present value of all future years' earnings per share. Therefore, the presence of a single year's earnings yield helps give a baseline for what a stock price will become without influence from other ratios.
- *Percent Change in Revenue:* Revenue growth is the baseline of predictors for stock value. Therefore, its year-wide percentage change is used to help gauge the effects of the actual baseline earnings of a company.
- *Percentage Change in Net Income:* Similar to the change in revenue, Percentage change in Net income represents the net change in money retained by a company in a given year. Rather than measuring the ability of the company to effectively gain revenues, however, the net income is meant to help measure how well a company is at minimizing its expenses in order to maximize its net earnings regardless of what revenues actually were.

Each ratio was chosen to fit a specific section of a contemporary ratio analysis and will be utilized in the predictive model to calculate the percentage change in stock value for a company whose stock is sold on the New York Stock exchange. The predictive variable which will be trained and tested against is the percentage change in the stock price over the course of the given year.

III: Dataset Summary Statistics:

There is a large amount of variation in the values for percentage change in net income, especially when compared to the percentage change in revenue. The standard deviation for percentage change in revenue seems to be a factor less than one, meaning that revenue tends to fluctuate very little. There are however a few outliers with rather large values, such as one outlier with a revenue change of six hundred. While these are only individual values, their ratios should work to remove the bias of the outlier and help explain why this large difference occurred.

In terms of net income differentials, they seem to have a larger standard deviation of about fifty. This points to large changes in net income, which could lead to even larger fluctuations in stock prices. The standard deviation seems be about five for the dataset, meaning that there is relatively larger amount of change in net income when compared to revenue. This points to the increasing volatility of expenses in a company, meaning that net income would most likely correlate more with stock price than the variable for revenue will.

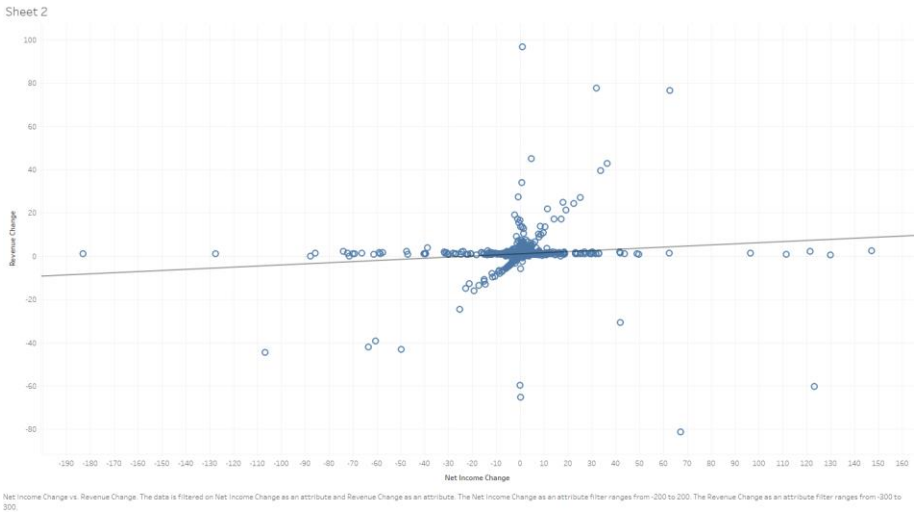
In terms of the contemporary ratios, the mean of each tends to be either one or zero, depending on the context of the ratio. Apart for some individual outliers, such as one for current ratio marked at over twenty-five thousand, the standard deviation seems to be relatively small, generally spanning values between one and two for each variable. Each of the variables also appear to be normally distributed, which aids to the validity of the model. That being said, there is a relatively large portion of the dataset which is considered anomalous. In terms of the variables, the only column with prominent anomalies is the wacc ratio, considering how some of the values are reported in the thousands, or even millions. Based on how the ratio is calculated, this shouldn't necessarily be possible. However, the total amount of points which exist like this is less than .1% of the total dataset, meaning even if they are left unchanged, they shouldn't affect the model in any effective way. In terms of the remaining data, the wacc ratios fall between zero and one, with a mean of .08 and a standard deviation of .02.

One of the main facets of the methods used to create the dataset is using the yfinance package to pull the financial statements for a given ticker symbol. However, for nearly two thousand of the eight thousand companies, yfinance was unable to pull a balance sheet or income statement. Because of this, each of the variables were replaced by null data which, for the purposes of the predictive model, is completely useless. Along with these anomalies, there are another thousand or so datapoints with missing values for certain ratios. For the cases where the data is entirely nullified, a full deletion and purge from the dataset is necessary. To fix the problem of individual

ratios missing, a simple mean imputation is all that will be necessary. Given the large number of variables, each with very little interaction with each other, an imputed case will still yield important data. What is extremely important and separates the anomalous cases between imputation and deletion is whether there is a value for price change. If the predictive variable is there, then it is possible to impute the remaining values and still retain some of the case's value for the model.

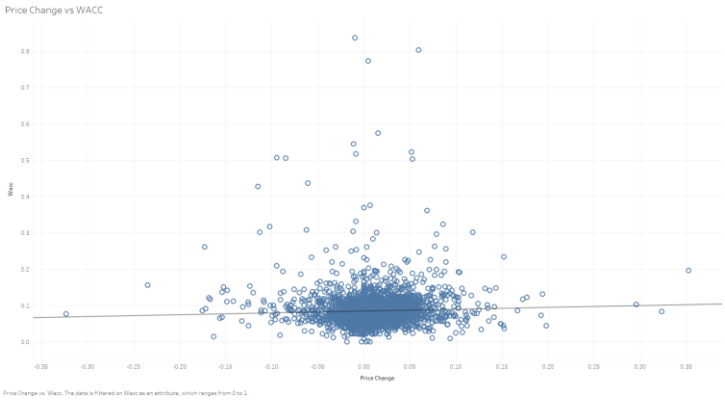
IV: Dataset Graphical Exploration:

Comparison of changes in Revenue and Net income:



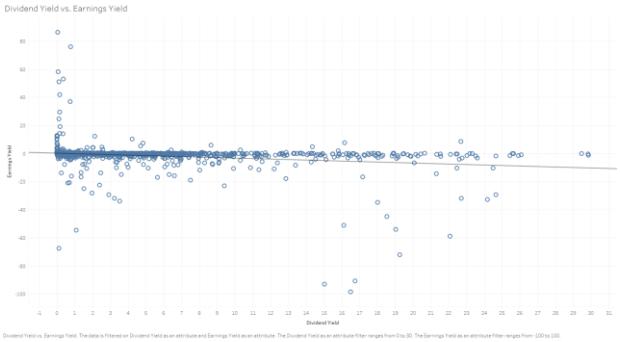
There appears to be a weak, positive, and linear relationship between the revenue and net income changes. This points towards the correlation between these values on the income sheet, but the small slope in the relationship shows that the dilation between revenue and net income tends to be relatively large.

Comparison of WACC percentage vs. Price Change



There appears to be a weak, positive, linear relationship with a near zero slope. Because of this, there doesn't appear to be a large correlation between the variables. Due to outliers, it is possible that the trend line is being pulled away from its true value. However, as there seems to be less than one hundred points in total, the total sway on the slope of the model, while being non-zero, is likely inconsequential.

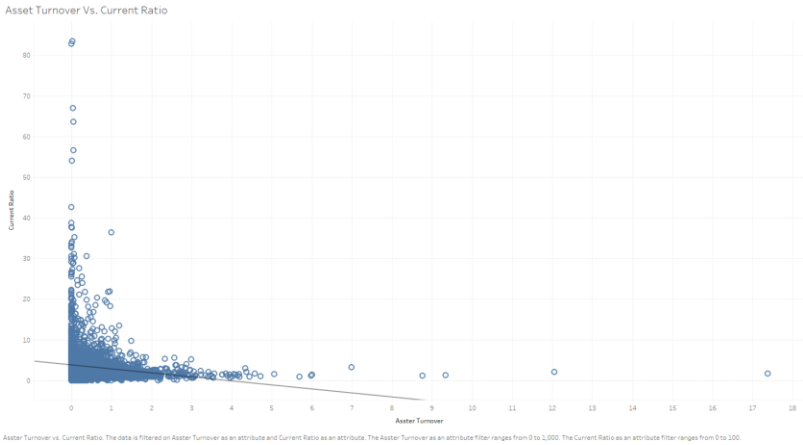
Comparison of Dividend Yield and Earnings Yield



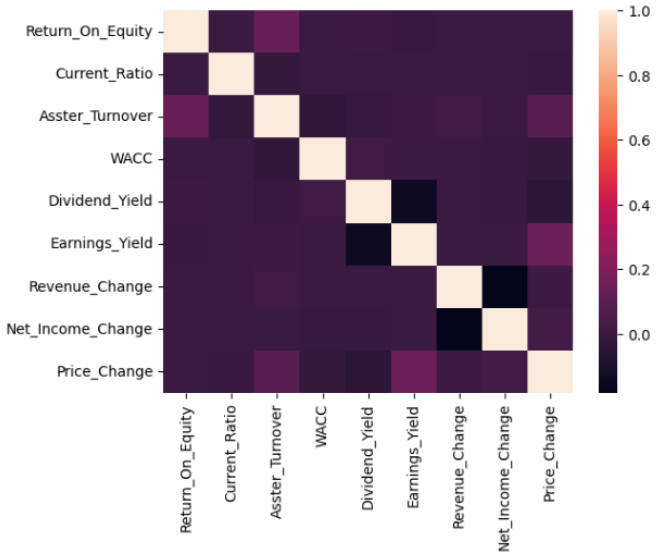
The fact that the slope of the relationship between the two yields shows that there is a very small amount of correlation in the data.

As shown by the three two variable graphs, the slope in the relationships between the variables is relatively low, meaning that there is likely no correlation between the variables, thus they have no relationship with each other. This is important because when there is little correlation between variables, the chance of the existence of confounding variables is extremely low. This means that the p-value for each individual variable in the model will be relatively unchanged based on the elimination of variables.

Comparison of Asset Turnover and Current Ratio:



That being said, there are a few cases where there does seem to be some kind of correlation between variables, such as with the Asset turnover and current ratio. That being said, in all the correlations, this occurrence tends to be small, as seen in the correlation matrix for the entirety of the dataset. While there is a small trend in the data which suggests that correlation, the sheer amount of datapoints shows how there is very little correlation in variables and thus, very little chance that the listed variables will confound or bias other variables in the model.



V: Summary Of findings:

While there are quite a few datapoints which must be removed from the dataset for lacking any information, it is clear that each individual variable will have a strong contribution in the model to predict the percentage change in price. In the correlation matrix, the only spots which show any relationship at all seem to be involving the price change variable. This means that the chance of interaction in variables is relatively low and thus will not bias the model. Because of this, I believe that the set of variables I have chosen are the most productive towards predicting price change over the course of a year. Additionally, the rather large dataset ensures that the variation in trend data will have minimal sway over the predictions, and that point is even more accurate given that the dataset can incorporate a second year's data, resulting in nearly twelve thousand total datapoints. From this, I believe I will be able to create an effective model for predicting price change in a stock based on its ratio analysis.