S&P 500 Price Change Predictors

Andy Barnes
Ryan Leveille
Tong Sun
Jared Turner

19 November, 2018

Project Scope

S&P 500

The Standard & Poor's 500, often abbreviated as the S&P 500, or just the S&P, is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ. The NYSE and NASDAQ are the American stock exchanges where the stocks that make up the S&P 500 are traded. Market capitalization (market cap) is the market value of a publicly traded company's outstanding shares. Market capitalization is equal to the share price multiplied by the number of shares outstanding. As outstanding stock is bought and sold in public markets, capitalization could be used as an indicator of public opinion of a company's net worth and is a determining factor in some forms of stock valuation. The reason that we chose to use the S&P 500 for our companies is that it is a very broad and diversified index that includes companies from many different industries. The S&P 500 is one of the most commonly followed equity indices, and many consider it one of the best representations of the U.S. stock market, and a bellwether for the U.S. economy. The share prices for each company within the S&P 500 change daily (during trading days) based on trading interest. A share price is the price of a single share of a company, derivative or other financial asset. The share price is the highest amount someone is willing to pay for the stock, or the lowest amount that it can be bought for.

Time Frame

Our project is going to analyze the companies within the S&P 500 over a 7 year period. This period is going to be from 1/01/2011-12/31/2017.

Approach:

Due to the volatility of the stock market and how the data can sometimes be very misleading depending upon the time frame which you choose to analyze the data from, we are going to approach this project with the intention of minimizing any major outliers for our response variable (share price).

For example, if we were to randomly select some date in 2011 and observe the share price for Microsoft it could be $50 per share. Likewise if we took another random date in 2011 to observe the price of Microsoft it could be $25 per share, perhaps even just a week before or after the price was at $50 per share. Furthermore, when we analyze the response variable in 2017 it could be $50 or $25, depending on the time frame you select for the observation.

Therefore, the way that we will make sure that we do not misrepresent our response variable and by keeping the project within the scope of our class we will be taking the average of the share price for each company within the S&P 500 for the years 2011 and 2017. The 5 years in between (2012, 2013, 2014, 2015, 2016) will allow plenty of time for our response variable to change based on our predictor variables.

Another reason why we want to calculate the average share price for each company by year is because our predictor variables will be yearly as well.

The S&P 500 is based on market cap and some companies will grow in size or fall in size so we will be using companies that were in the S&P 500 starting on January 1, 2011, and are still in the S&P 500 on December 31, 2017. As of now, we have 476 companies out of 500 that are still within the S&P 500 to analyze.

Response Variable

`Price.ch`: Our response variable is going to be the share price of each company within the S&P 500. We are going to take the average share price for 2011 and the average share price for 2017 and then calculate the percent difference for each company.

Predictor Variables

We have multiple variables that we are going to use as our predictor variables. These variables have been chosen based on what what we think investors use to analyze a company before making an investment. Also, some other potential variables' data is incomplete or missing so we choose variables that were best for the scope of the project.

Our goal with this project is to find the predictor variables that have the biggest impact on our response variable (share price).

All of our variables are going to be using their percent change in yearly unit during our time period of 2011-2017. We will start with the predictor variables' 2011 observations and end with the predictor variables' 2017 observations. We chose this time frame because it gave us our yearly data of 2011, a 5 year time frame, then our yearly data of 2017.

1. `EPS`: Earnings per share (EPS) is the portion of a company's profit allocated to each share of common stock. Earnings per share serves as an indicator of a company's profitability.
2. `Div`: A dividend is a payment made by a corporation to its shareholders, usually as a distribution of profits. When a corporation earns a profit or surplus, the corporation is able to re-invest the profit in the business and pay a proportion of the profit as a dividend to shareholders.
3. `BV`: Book value per share indicates the book value (or accounting value) of each share of stock. Book value is a company's net asset value, which is calculated by total assets minus intangible assets and liabilities. An easy way to think of book value per share is what is the expected value of the company.
4. `RoA`: The return on assets shows the percentage of how profitable a company's assets are in generating revenue.
5. `RoE`: Return on equity is a measure of the profitability of a business in relation to the book value of shareholder equity, also known as net assets or assets minus liabilities. ROE is a measure of how well a company uses investments to generate earnings growth.
6. `RoIC`: Return on capital (ROC), or return on invested capital(ROIC), is a ratio used in finance, valuation and accounting, as a measure of the profitability and value-creating potential of companies after taking into account the amount of initial capital invested.
7. `DE`: Debt/Equity (D/E) Ratio, calculated by dividing a company's total liabilities by its stockholders' equity, is a debt ratio used to measure a company's financial leverage. The D/E ratio indicates how much debt a company is using to finance its assets relative to the value of shareholders' equity.
8. `Rev`: Revenue is the income that a business has from its normal business activities, usually from the sale of goods and services to customers.

The Data

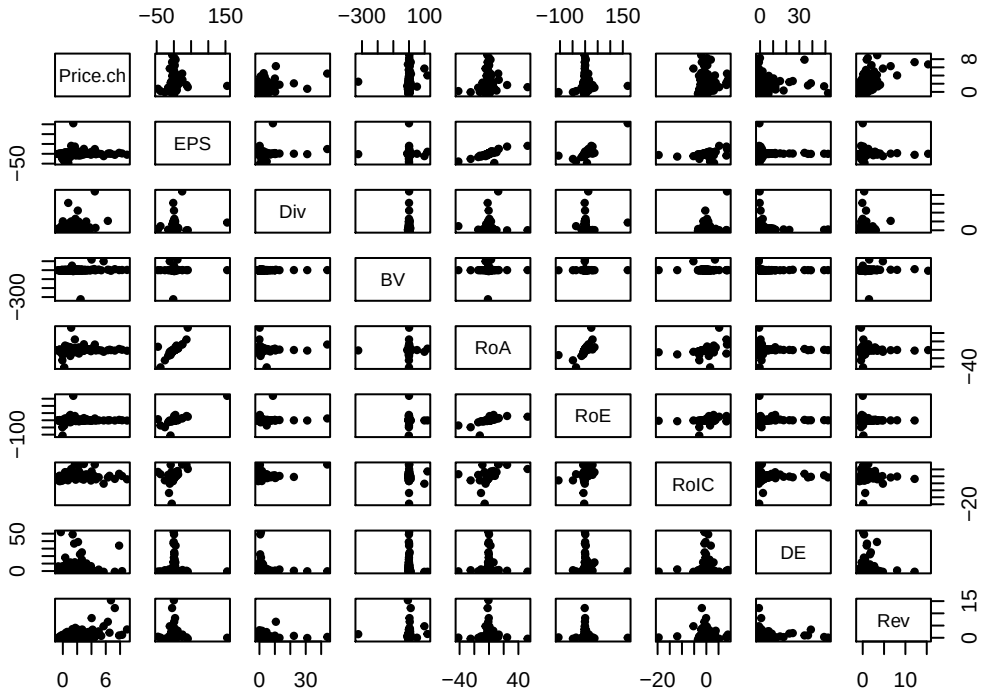
Our data looks at the 476 stocks in in the S&P 500 today with data dating back to 2011. Each variable is given as the percent change in yearly value or average value from 2011 to 2017. The first 6 rows of data are as follows

Co.	Price.ch	EPS	Div	BV	RoA	RoE	RoIC	DE	Rev
A	1.0041	-0.2632	NA	0.1531	-0.2187	-0.4386	-0.3129	-0.1778	-0.3240
AAL	5.3662	-1.2464	NA	-1.2150	-1.4623	NA	NA	NA	0.7602
AAP	0.9557	0.2564	0.0000	3.0918	-0.4973	-0.6409	-0.6045	-0.3673	0.5193
AAPL	1.8952	1.3316	NA	1.1927	-0.4876	-0.1152	-0.5161	NA	1.1177
ABC	1.2260	-0.3543	2.3953	0.0665	-0.7792	-0.2842	-0.5969	3.8824	0.9091
ABMD	8.9755	-4.6562	NA	2.4388	-2.1889	-2.1493	-2.1050	NA	3.4059

from this we can obtain a correlation matrix

	Price.ch	EPS	Div	BV	RoA	RoE	RoIC	DE	Rev
Price.ch	1.0000	0.2264	0.3533	0.2650	0.1495	0.1613	0.3862	0.0228	0.3079
EPS	0.2264	1.0000	0.1746	0.1341	0.7674	0.6148	0.3577	0.0005	0.1163
Div	0.3533	0.1746	1.0000	0.0954	0.0734	0.0707	0.3140	-0.0344	0.0239
BV	0.2650	0.1341	0.0954	1.0000	0.0716	-0.0973	0.0659	-0.1470	0.4739
RoA	0.1495	0.7674	0.0734	0.0716	1.0000	0.7680	0.4130	-0.0146	0.0475
RoE	0.1613	0.6148	0.0707	-0.0973	0.7680	1.0000	0.4423	0.1272	0.0109
RoIC	0.3862	0.3577	0.3140	0.0659	0.4130	0.4423	1.0000	-0.0242	0.1823
DE	0.0228	0.0005	-0.0344	-0.1470	-0.0146	0.1272	-0.0242	1.0000	0.0008
Rev	0.3079	0.1163	0.0239	0.4739	0.0475	0.0109	0.1823	0.0008	1.0000

and scatterplot.



Why we chose this data

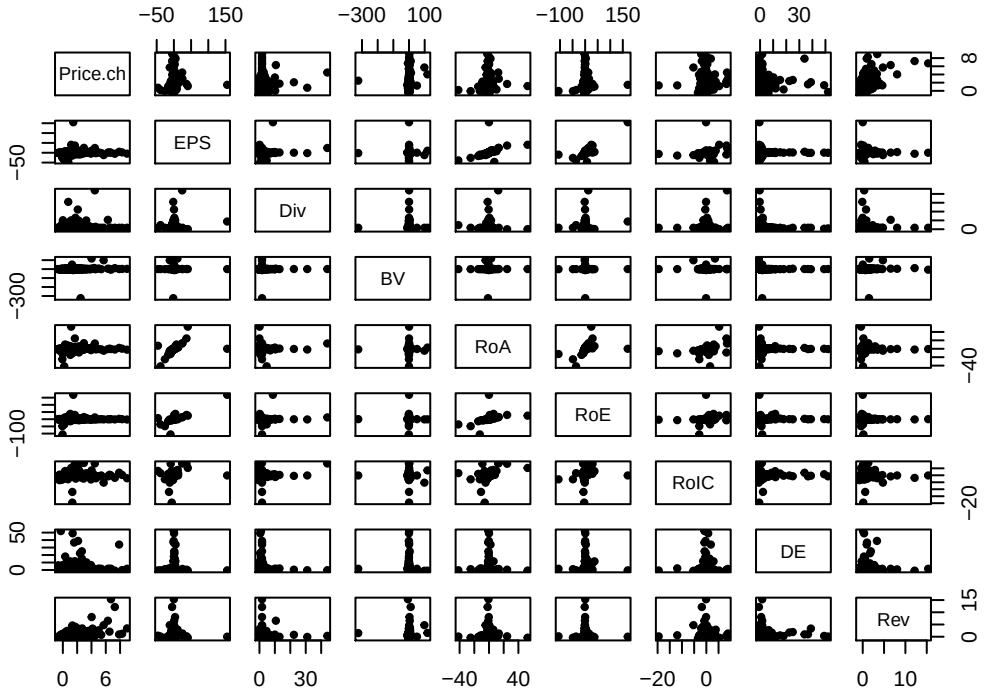
We chose this project due to the vast amounts of data that the stock market encompasses. The stock market is a very broad topic and this project can be applied to many different situations.

With our project we hope to find which predictor variables affect the response variable (share price) the most. We are excited to see what our models come up with. The findings will be interesting as there are many different theories as to what is the most important predictor variables for a companies share price.

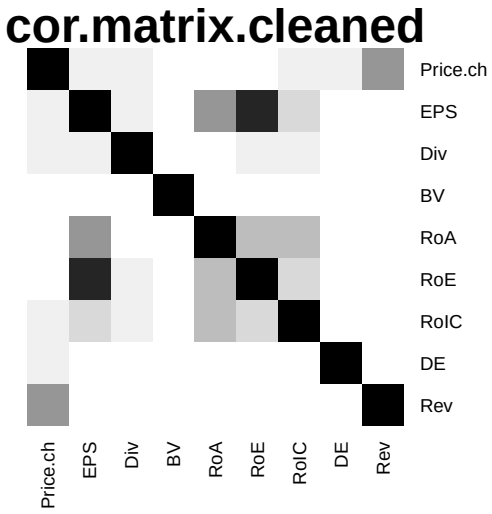
Data preprocessing

Before formulating our model, we cleaned our data by filling in missing values. For the predictor variable `Div`, missing values were assumed to be 0, this is because dividends other than 0 are usually reported and recorded. For all other predictors, missing values were filled in with the mean among all observations for that data. After our data was cleaned we have the following correlation matrix and plot.

	Price.ch	EPS	Div	BV	RoA	RoE	RoIC	DE	Rev
Price.ch	1.0000	0.0819	0.1866	0.0631	0.0791	0.0643	0.0990	0.0889	0.5314
EPS	0.0819	1.0000	0.1558	0.0134	0.4554	0.8153	0.2446	-0.0128	0.0047
Div	0.1866	0.1558	1.0000	0.0051	0.0447	0.0973	0.1529	-0.0348	0.0334
BV	0.0631	0.0134	0.0051	1.0000	0.0131	-0.0046	-0.0092	-0.0201	0.0463
RoA	0.0791	0.4554	0.0447	0.0131	1.0000	0.4084	0.4246	-0.0116	-0.0004
RoE	0.0643	0.8153	0.0973	-0.0046	0.4084	1.0000	0.2076	0.0245	-0.0096
RoIC	0.0990	0.2446	0.1529	-0.0092	0.4246	0.2076	1.0000	0.0082	-0.0141
DE	0.0889	-0.0128	-0.0348	-0.0201	-0.0116	0.0245	0.0082	1.0000	0.0404
Rev	0.5314	0.0047	0.0334	0.0463	-0.0004	-0.0096	-0.0141	0.0404	1.0000



We were also able to generate a heat map of this correlation matrix.



Our Model

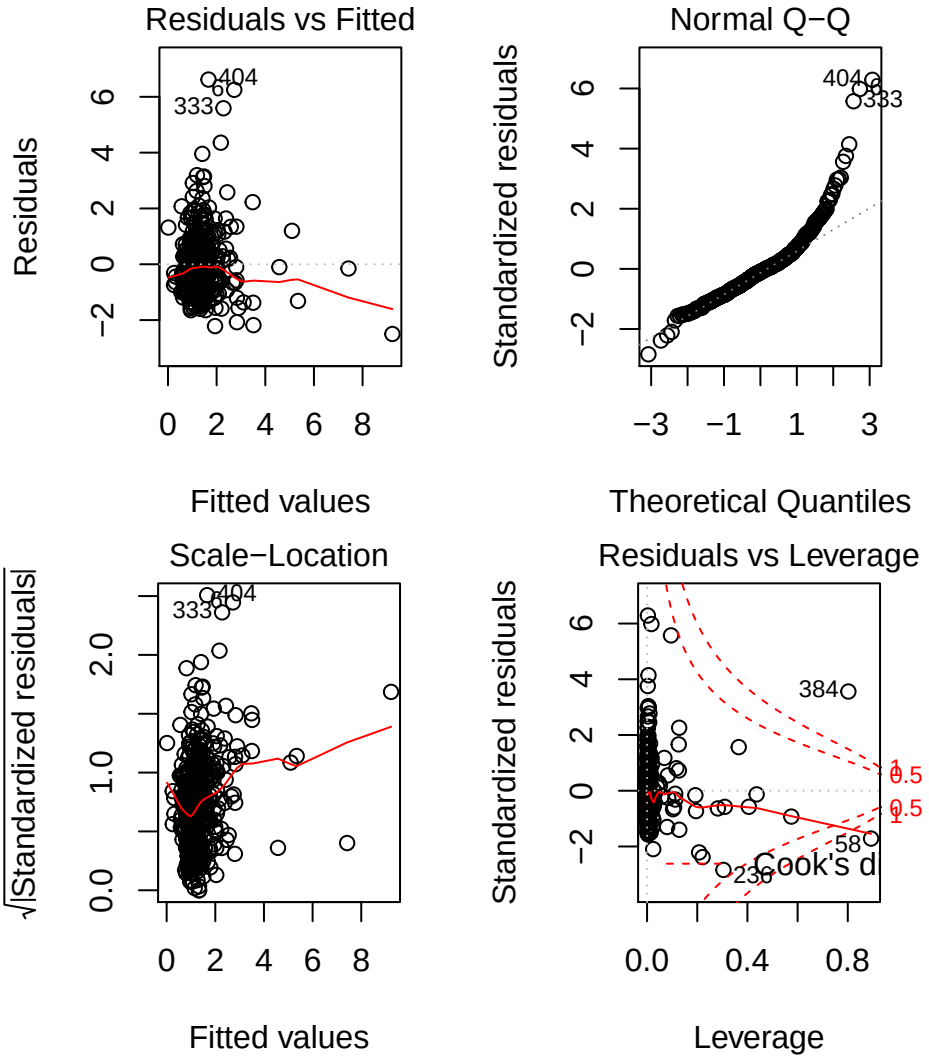
We will formulate a simple linear model, where our response, `Price.ch`, is a linear combination of our predictors

$$\widehat{\text{Price.ch}}_i = \beta_0 + \beta_1 \text{EPS}_i + \beta_2 \text{Div}_i + \beta_3 \text{BV}_i + \beta_4 \text{RoA}_i + \beta_5 \text{RoE}_i + \beta_6 \text{RoIC}_i + \beta_7 \text{DE}_i + \beta_8 \text{Rev}_i + \epsilon_i$$

where $\epsilon_i \sim^{\text{iid}} N(0, \sigma^2)$. Where σ^2 is the population variance. After this, our initial attempt was to take our data and run a multiple linear regression on it. The summary is as follows.

```
##
## Call:
## lm(formula = Price.ch ~ ., data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4984 -0.6388 -0.1045  0.3713  6.6183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.888564   0.061799  14.378  < 2e-16 ***
## EPS         0.001641   0.010075   0.163   0.8707
## Div         0.065480   0.016208   4.040 6.25e-05 ***
## BV          0.003022   0.002900   1.042   0.2978
## RoA         0.011051   0.014615   0.756   0.4500
## RoE         0.002052   0.008244   0.249   0.8035
## RoIC        0.049108   0.034103   1.440   0.1505
## DE          0.018887   0.009769   1.933   0.0538 .
## Rev         0.536356   0.039114  13.713  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.054 on 467 degrees of freedom
## Multiple R-squared:  0.3267, Adjusted R-squared:  0.3151
## F-statistic: 28.32 on 8 and 467 DF,  p-value: < 2.2e-16
```

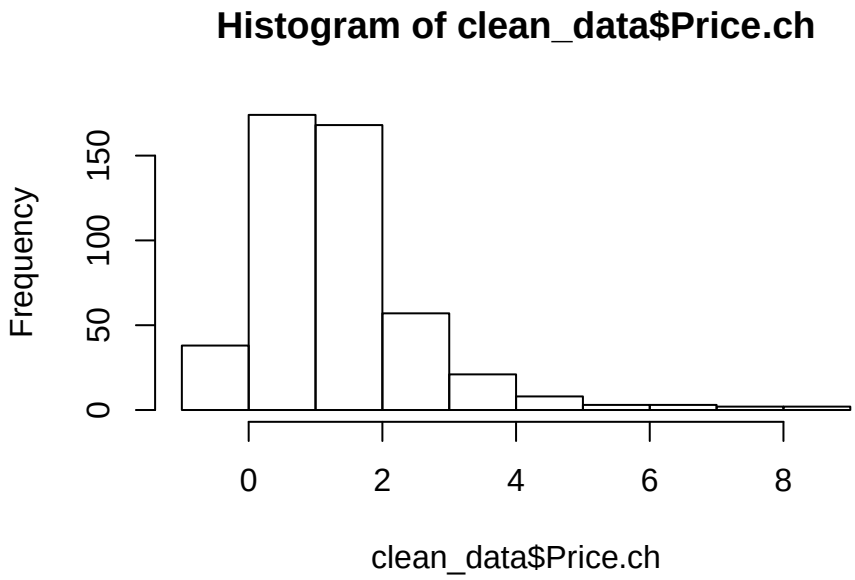
We were able to generate the following diagnostic plots.



In addition to our model having an R-squared of 0.3266805, indicating that only about 30% of variation in our response is explained by our predictors, these plots are of concern. The plot of our Residuals vs. Fitted values indicates that our

Checking for outliers

In hopes of improving our model, we will set out to find a small subset of outliers we can throw out to drastically improve our model. First, let's take a look at the frequency histogram for our response variable `Price.ch`,



which shows that the values of `Price.ch` are skewed right and do not approximate a normal distribution. We will now attempt to drop as few outliers as possible.

STOP HERE..... SCRATCH WORK FOLLOWS

```
## Start:  AIC=59.08
## Price.ch ~ EPS + Div + BV + RoA + RoE + RoIC + DE + Rev
##
##        Df Sum of Sq    RSS     AIC
## - EPS   1     0.029 518.93  57.107
## - RoE   1     0.069 518.97  57.143
## - RoA   1     0.635 519.54  57.662
## - BV    1     1.207 520.11  58.186
## <none>              518.90  59.080
## - RoIC  1     2.304 521.21  59.189
## - DE    1     4.153 523.06  60.875
## - Div   1    18.134 537.04  73.431
## - Rev   1   208.935 727.84 218.140
##
## Step:  AIC=57.11
## Price.ch ~ Div + BV + RoA + RoE + RoIC + DE + Rev
##
##        Df Sum of Sq    RSS     AIC
## - RoE   1     0.389 519.32  55.464
## - RoA   1     0.722 519.66  55.769
## - BV    1     1.217 520.15  56.222
## <none>              518.93  57.107
## - RoIC  1     2.324 521.26  57.234
## - DE    1     4.129 523.06  58.879
## - Div   1    18.618 537.55  71.886
## - Rev   1   209.098 728.03 216.266
##
## Step:  AIC=55.46
## Price.ch ~ Div + BV + RoA + RoIC + DE + Rev
```

```
##
##        Df Sum of Sq    RSS     AIC
## - BV    1     1.204 520.53  54.566
## - RoA   1     1.339 520.66  54.690
## <none>              519.32  55.464
## - RoIC  1     2.379 521.70  55.639
## - DE    1     4.223 523.55  57.319
## - Div   1    19.203 538.53  70.747
## - Rev   1   208.889 728.21 214.384
##
## Step:  AIC=54.57
## Price.ch ~ Div + RoA + RoIC + DE + Rev
##
##        Df Sum of Sq    RSS     AIC
## - RoA   1     1.386 521.91  53.832
## <none>              520.53  54.566
## - RoIC  1     2.326 522.85  54.689
## - DE    1     4.129 524.66  56.327
## - Div   1    19.246 539.77  69.849
## - Rev   1   210.837 731.36 214.439
##
## Step:  AIC=53.83
## Price.ch ~ Div + RoIC + DE + Rev
##
##        Df Sum of Sq    RSS     AIC
## <none>              521.91  53.832
## - DE    1     4.045 525.96  55.508
## - RoIC  1     4.990 526.90  56.362
## - Div   1    19.015 540.93  68.866
## - Rev   1   211.115 733.03 213.521
##
## Call:
## lm(formula = Price.ch ~ Div + RoIC + DE + Rev, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5586 -0.6462 -0.0981  0.3765  6.6185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.891452   0.061592  14.473  < 2e-16 ***
## Div         0.066261   0.015996   4.142 4.07e-05 ***
## RoIC        0.065402   0.030820   2.122   0.0344 *
## DE          0.018599   0.009734   1.911   0.0567 .
## Rev         0.538397   0.039006  13.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 471 degrees of freedom
## Multiple R-squared:  0.3228, Adjusted R-squared:  0.317
## F-statistic: 56.12 on 4 and 471 DF,  p-value: < 2.2e-16
```

	Div	RoIC	DE	Rev
Div	1.0000	0.1529	-0.0348	0.0334
RoIC	0.1529	1.0000	0.0082	-0.0141
DE	-0.0348	0.0082	1.0000	0.0404
Rev	0.0334	-0.0141	0.0404	1.0000