NYSE Analysis With SQL

Ethan Graham, Scott Gozdzialski, Huynh Vu

Graduate Students Master of Data Science

Southern Methodist University

Dallas, Texas

**1. Abstract:** We endeavor to pull in several years of financial securities (stocks) information and save them into a data management system for easy querying and analysis. We can use the stocks package for Kaggle, found at <https://www.kaggle.com/benjibb/lstm-stock-prediction-20170507/data>. We can export the data into a data warehousing and querying system, such as MySQL.

Once the data is loaded into a database, we can perform several queries on the data and report the results.

**2. Introduction:** Database management systems (DBMS) have a proven track record handling arbitrary datasets. For example, stock trading: between 50% and 84% of all trades are completed by high-frequency computers, and it is predicted as time goes on computer trading will expand. Individual day traders need insight applications to drive investment decisions, and databases are a core technology to enable that. MySQL can help let the data speak for itself and drive informed investment.

**3. Data and Database construction:** The data is available at kaggle.com. It contains New York Stock Exchange (NYSE) securities pricing data from 2010 to 2016. The data is broken up into four different files:

1. Daily price of stocks
2. Daily price of stocks with adjustment for stock splits
3. Description of each security and the industry sector it belongs to
4. Metrics extracted from annual SEC 10K fillings.

The Data on all four files totals 103.6 MB and was run on the Bluemix server and a Windows 10 machine with a dual core 2.5GHz I7 processor and 6.00 GB of RAM with no issues.

These files can be consumed by MySQL to create a database for efficient analysis. Insights are drawn from the MySQL database named “stocks”.

The Database was put into a relational database with the “Symbol” variable used as the primary key for each table.

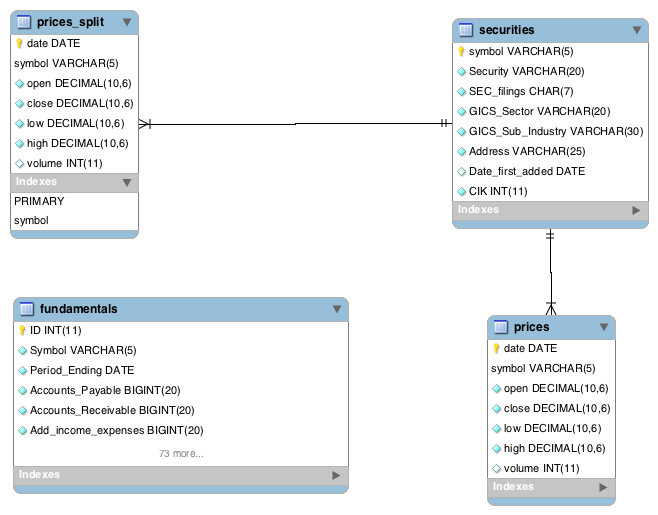


Fig 1

**4. Data Insights**

**4.1 Insight 1:** What is the relationship of high volatility stocks to their growth? Using ggplot2 in R shows a striking positive linear relationship:

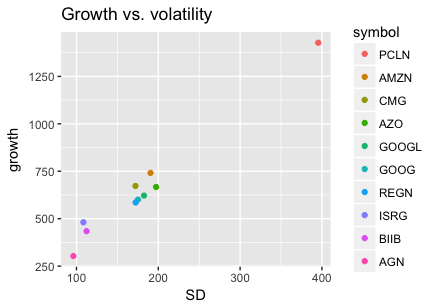


Fig 2

Data provided from SQL query:

SELECT symbol, stddev(high) as SD,

max(high) - min(low) as growth

FROM prices\_split

GROUP BY symbol

ORDER BY SD DESC LIMIT 10;

We can see in fig 2 there is a linear correlation between the voliatility of a stock and the growth.

**4.2 Insight 2:** Using the knowledge gained in insight 1, we are looking for a way to determine the actual growth and volatility. We do this by comparing the highest growth stock to the lowest group stock using SQL subqueries which provide insight. We plot the difference instead of using numbers to highlight the actual difference in a more relatable format. From insight 1, the highest growth stock has the symbol “PCLN”. The lowest growth stock is found with an SQL query with an embedded subquery:

SELECT \* FROM

(

SELECT symbol, max(high)-min(low) AS growth FROM prices\_split

GROUP BY symbol

) AS subTable ORDER BY growth ASC LIMIT 1;

The high growth data is provided by, SELECT \* FROM prices\_split WHERE symbol = 'PCLN';

Charting the lowest and highest growth stocks shows large differentials in stock performance.

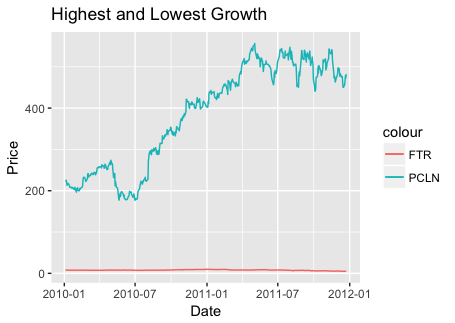


Fig 3

**4.3 Insight 3:** Looking at the different stocks, we wanted to know if the different industries had diffeerent growth during this period.

SELECT GICS\_Sub\_Industry , AVG(newhigh)-AVG(newlow) AS avg\_growth

FROM(

SELECT AVG(high) AS newhigh, AVG(low) AS newlow, GICS\_Sub\_Industry

FROM (

SELECT \*

FROM securities natural join prices\_split) AS combined

GROUP BY symbol

) AS temp

GROUP BY GICS\_Sub\_Industry

ORDER BY avg\_growth DESC;

The SQL query above, created 124 different results. Therefore the comparision of the top five and bottom five records is a better metrics used for this paper,making the information more managable. This will show the highest growing industries verses those that are tapering off. The results are listed in the tables below.

### Top 5

|  |  |
| --- | --- |
| Internet & Direct Marketing Retail | 7.53716222651872 |
| Life Sciences Tools & Services | 3.93256237883045 |
| Casinos & Gaming | 3.72577750510828 |
| Industrial Materials | 3.69075463507389 |
| Biotechnology | 3.48485485142692 |

### Table 1

### Bottom 5

|  |  |
| --- | --- |
| Integrated Telecommunications Services | 0.451242797284351 |
| Publishing | 0.443185394110685 |
| Independent Power Producers & Energy Traders | 0.425368906923989 |
| Technology Hardware Software and Supplies | 0.424676514188445 |
| Thrifts & Mortgage Finance | 0.244812714528964 |

Table 2

**4.4 Insight 4:** We are looking into different metrics of stocks. By joining split\_prices and fundamentals, there are a large number of ways we can compare growth to other company metrics. This insight focuses on Amazon and how growth compares year over to capital expenditures and earnings before tax.

SELECT MAX(high) - min(low) AS growth, capital\_expenditures, YEAR(date), Earnings\_Before\_Tax

FROM prices\_split JOIN fundamentals ON

prices\_split.symbol = fundamentals.symbol AND For\_Year = YEAR(date)

WHERE prices\_split.symbol = "AMZN"

GROUP BY symbol, YEAR(date) LIMIT 10;

|  |  |  |  |
| --- | --- | --- | --- |
| growth | capital\_expenditures | YEAR(date) | Earnings\_Before\_Tax |
| 159.880005 | -3444000000 | 2013 | 506000000 |
| 124.059998 | -4893000000 | 2014 | -111000000 |
| 411.190002 | -4589000000 | 2015 | 1568000000 |

Table3

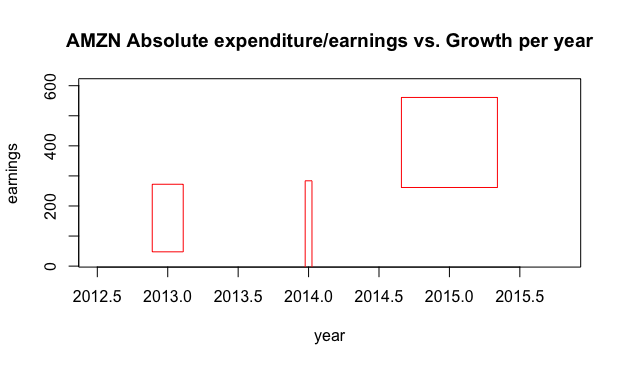


Fig 4

Figure 4 shows the differences of Amazon stock for 2013 through 2015. The width of the boxes highlights the Earnings before stock and the height shows the Capital Expendatures.

**4.5 Insight 5:** It is important to know which industries and stocks provide the highest earnings per share. Using the Global Industry Classification Standard (GICS) Sector, the 2016 data can be found:

SELECT DISTINCT GICS\_Sector, `security`, MAX(Earnings\_Per\_Share)AS `Max EPS`

FROM fundamentals

INNER JOIN securities

ON fundamentals.symbol = securities.symbol

WHERE For\_Year = 2016

GROUP BY GICS\_Sector

ORDER BY MAX(Earnings\_Per\_Share) DESC;

|  |  |  |
| --- | --- | --- |
| GICS\_Sector | security | Max EPS |
| Consumer Discretiona | AutoZone Inc | 42 |
| Health Care | AmerisourceBergen Co | 10 |
| Information Technolo | Apple Inc. | 8 |
| Industrials | Acuity Brands Inc | 7 |
| Consumer Staples | ConAgra Foods Inc. | 6 |
| Materials | Air Products & Chemi | 3 |
| Financials | Block H&R | 2 |
| Energy | Helmerich & Payne | -1 |

**5. Comparable studies** When we look into other studies to see if there are any comparable studies we find a 1999 study by Ruth Judson and Athanasios Orphanides comparing inflation volatility and growth, with a focus on inflation and not stock prices in the Volume 2, Issue 1 April 1999 “International Finance” journal. The study focused on the growth in different international markets and their markets as a whole not individual stocks. The study can be found at <http://onlinelibrary.wiley.com/doi/10.1111/1468-2362.00021/full>

There are many studies with focus on the different growth of stocks in different sectors, that is the study of economics. But, when looking for any related to the data we used, none were found. I cannot prove there are no other studies out there which cover the same subject, but I found none in the time I spent researching.

**6. Conclusion** We have looked at the Kaggle market data from the NYSE from 2010 through 2016. Where we researched the correlation between the volatility of stock vs the growth rate of stocks, the research showed a distinct linear trend. We showed the difference visually between the highest growth stock vs the lowest growth stock, displaying the vast difference in volatility. We looked to see if the different markets had a similar or different growth rate during the time period. We did this by looking at the top 5 and bottom 5 GICS sectors average growth rate. We focused on one stock, “Amazon”, and explored relationships in different SEC fundamentals. Finally, we wanted to highlight the top performers of the different GICS sectors with a focus on earning per share.

Finally, we would like to highlight that this study was conducted with a very limited dataset and any correlations or revelations cannot be applied to the large stock population. All our data and code can be found at github in the website listed below.

<https://github.com/DSSMU/MSDS7330_Project>

7. References

<http://www.zerohedge.com/contributed/2012-17-26/84-all-stock-trades-are-high-frequency-computers-%E2%80%A6-only-16-are-done-human-tra>

https://www.kaggle.com/benjibb/lstm-stock-prediction-20170507/data