NYSE Analysis With MySQL

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.

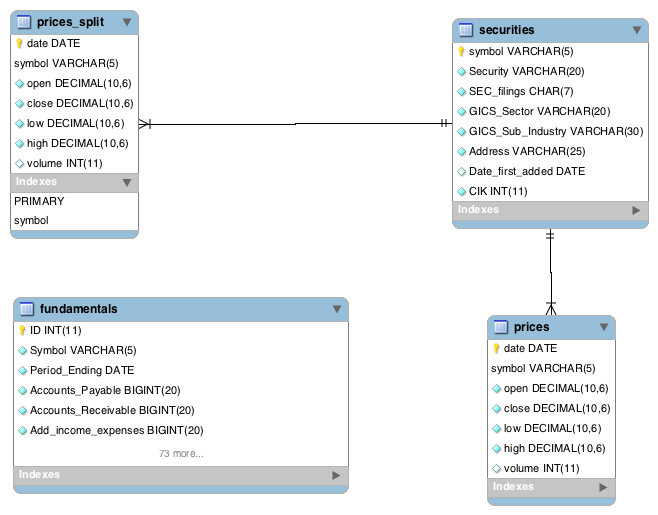
Once the data is loaded into a database, we can perform several queries on the data and report the results.

**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:

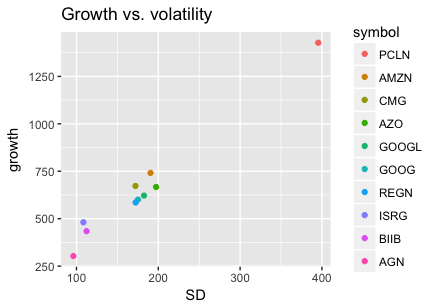
1. Daily price of stocks
2. Daily price of stocks with adjustment for stock splits
3. description of each security and the industry sector it belongs to
4. metrics extracted from annual SEC 10K fillings.

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



**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:



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;

**4.2 Insight 2** Comparing the highest growth stock to the lowest group stock using SQL subqueries provides insight. 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

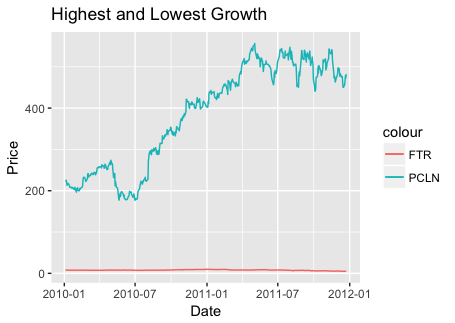
GROUP BY symbol

) AS subTable ORDER BY growth ASC LIMIT 1;

|  |  |
| --- | --- |
| symbol | growth |
| FTR | 6.78 |

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.



**4.3 Insight 3** Looking at the differnet stocks we wanted to know if the different industies 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. Therefor the comparision of the top five and bottom 5 records is a better metrics used for this paper. This will show the highest growing industries verse 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 |

### 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 |

**4.4 Insight 4** 5 By joining split\_prices and fundamentals, there are a huge 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

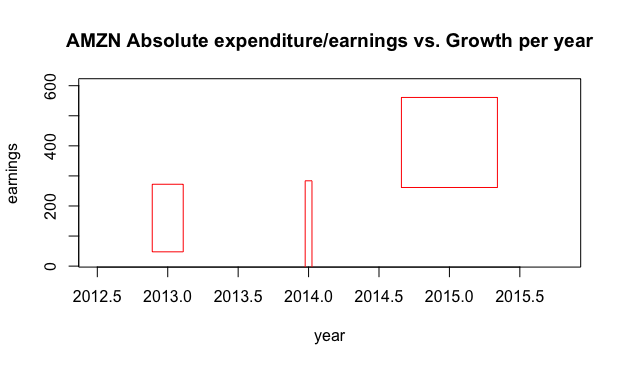
prices\_split.symbol = fundamentals.symbol AND For\_Year = YEAR(date)

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 |

We can chart this using absolute values as



**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. Visualization**

**6. Conclusiton**

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

2. General Instructions

The maximum length of a manu­script is 4 pages, printed single-sided. Print all text, including section titles and figures, in two-column format where each column is 8.5 cm by 24.5 cm (3.35 in by 9.65 in) and there is a 0.6 cm (0.24 in) space between the two columns. Excep­tions to the two-column format include the title at the top of the first page and any full-width figures or tables. Start all pages directly under the top margin. Text should be centered on each page. On A4 paper, this roughly means leaving 1.7 cm (0.67 in) margins on left and right sides of each page as well as a 2.5 cm (1 in) margin on the top and bottom of each page. Type single-spaced. Indent when starting a new paragraph. Use standard fonts such as Times New Roman or Computer Modern Roman, 10 points for text, 11 points (bold) sub­section headings, 12 points (bold) for section headings, 14 points (bold) for title, 11 points for authors’ names, and 10 points for their affiliations.

**2.1 The First Page**

Center the title across both columns. Use the two-column format only when you begin the abstract.

**Title:** Place the title at the top of the first page, followed by the authors’ names and their affiliations. Long title should be typed on two lines without a blank line intervening. Leave approximately 1 cm (0.39 in) between the title and the body of the first page.

**Abstract:** Type the abstract at the beginning of the first column. The abstract should be no longer than 200 words.

**Text:** Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in this example.

**2.2 Sections**

**Headings:** Type and label section and subsection headings in the style shown on these pages. Use numbered sections, in order to facilitate cross references.

**References:** Citations within the text appear in brackets as [ref. number]. Gather the full set of references together under the heading **References**; place the section before any **Appendices**, unless they contain references. Arrange the references in the order that they are cited in the text. Provide as com­plete a citation as possible, using a consistent format.

**Appendixes:** Appendixes, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A Title of Appendix.**

**2.3 Footnotes**

Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[[1]](#footnote-1) Footnotes should be separated from the text by a line.[[2]](#footnote-2)

**2.4 Graphics**

**Illustrations:** Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns.

**Captions:** Provide a caption for every illus­tration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions for figures below the figures. Type the captions for tables above the tables.

**3. Length of Camera-ready Manuscript**

For the length of camera-ready manuscripts, **a paper is limited up to 4 pages**. All illustrations, references, and appendices must be accommodated within these page limits. Any extra page beyond the first four pages will be deleted. **Please DO NOT put a page number in each pag**

**4. Submission Process**

1. Format your paper using this template.

2. Turn the hardcopy by Dec 4th before the lecture starts

### References

[1] T.A. Jones, “Writing a good paper,” *IEEE Trans. on General Writing*, Vol. 1, no. 2, pp.1-10, May 2002.

[2] K. Hwang, *Computer Arithmetic*, John Wiley, 1997.

You have to refer to the following rubric to get the good score as the isntructor will evaluate your paper based on it.

**Team Presentation Rubric: 100%**

1.      Presentation Slide Format: 30%

a.      Font Size (at least 24 font size): 9%

b.      Data Size (How much MB or GB): 7%

c.      Data Source URL: 6%

d.      H/W experimental Specifications (Database server, CPU speed, Memory Size): 8%

2.      Originality: 20%

* 1. How unique your idea different from the existing data analysis (10%)
  2. and your deliverable, that is,  what insights you find out (10%)

3.      Relevance with the topic in the class: 35%

a.      Data Analysis using RDB or NoSQL DB: 10%

b.      Implementation using SQL or other scrpts: 15%

c. Github link and the content that has your presentation file (ppt) and term paper: 10%

4.      Communicate with the instructor about the topic to get approval about the topic: 15%

**NOTE:**

1. **Peer Evaluation (Optional)**: You have to email to the instructor the peer evaluation about your team members for the term project and the presentation. It should be composed of:

If you don’t email me peer evaluation, I assume, all of you contribute the work fairly well.

**For example,** your team score is 95% and your peer evaluation by your team members are 100%, your score is 95 (= 95 x 100%)

|  |  |  |  |
| --- | --- | --- | --- |
| **Group NAME: ?**  **Member** **Name** | **Weight (100%)**  **Ex: 90% (of 100%)** | **1-2 lines of explanation what the member has done and the member’s contribution is** | **NOTE** |
| Yourself |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. **Plagiarism**: If you make a copy of others, it should violate the academic integrity so that you should get 0 in the term project or F in the course in the worst case.

**Term Paper Rubric: 100%**

It should be almost same as the team presentation. But, mostly, I will take a look at if you revise the content per my **comment** at the presentation. Thus, any penalty at the presentation can be recovered. You also need to email the instructor the **peer evaluation** for the term paper. If you don’t email me peer evaluation, I assume, all of you contribute the work fairly well.

**For example,** your team score is 95% and your peer evaluation by your team members are 100%, your score is 95 (= 95 x 100%)

|  |  |  |  |
| --- | --- | --- | --- |
| **Group NAME: ?**  **Member** **Name** | **Weight (100%)**  **Ex: 90% (of 100%)** | **1-2 lines of explanation what the member has done and the member’s contribution is** | **NOTE** |
| Yourself |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. **Plagiarism**: If you make a copy of others, it should violate the academic integrity so that you should get 0 in the term project or F in the course in the worst case.

1. This is how a footnote should appear [↑](#footnote-ref-1)
2. Note the line separating the footnotes from the text [↑](#footnote-ref-2)