Exploring Search Relevance as an Indicator for Corporate Quarterly Financial Reports

Jason Coffman (jascoffm@) and Nicolas Calo (ncalo@)

24 May 2021

Table of Contents

Motivation

Quarterly earnings reports are documents provided by publicly traded companies to report their quarterly fiscal performance (Tuovila, 2020). Within these reports are a number of key metrics - including revenues, profits, and cash flow - which all convey the recent financial

status of a business. These reports traditionally have profound implications on stock prices and stock trading volume as traders use these reports to decide whether or not to buy or sell shares of a company (Terzo, 2016). Having the ability to predict the results of an earnings report, or some elements of the report, would allow traders to purchase shares before stock prices increase - in the case of positive earnings - and to offload shares before a price decrease - in the case of negative earnings.
For this project, we will examine the relationship between a single element of the quarterly earnings report for a company, earnings per share (defined as a company's quarterly profit divided by the number of outstanding shares), and proxies for public interest in the same company. The primary proxy for public interest within this report is the relative Google search volume for the company's stock symbol. The ultimate intent of this research is to examine if search volume can be used to explain, and perhaps predict, the earnings per share for a company.

Data Sources

Google Trends

Google Trends is a publicly available dataset that provides analyses of terms and subjects searched within the Google search engine. While there are myriad data points provided by Google Trends, this investigation was solely concerned with *Interest over time*. The interest over time data was retrieved via the Python 3.7 library pytrends, which provides seamless Python integration with the Google Trends API. The pytrends library ingests a search term, this project used a company's stock symbol concatenated with the word *stock* (i.e. MMM stock), and returns a pandas dataframe of interest over time for the previous 5 years; the earliest data available was May 2016. Data for the stock symbols listed in the S&P 500 Index were collected via the pytrends library. In total 504 independent queries were made, for each symbol tracked in the S&P 500 index, with each query containing 260 7-day interest over time values for a total of 131,040 unique values.

Table 1: Summary of utilized variables within the Google Trends dataset

Variable Name	**Description**
Interest over time (column name as stock symbol)	Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.
date	Date associated with the interest over time variable. Date represents the start of a 7-day period; the method used to consolidate interest

	over time into a 7-day period is performed by the data provided and is unknown.
isPartial	Boolean value used to indicate whether the interest over time value represents a complete 7-day period of data, or whether the 7-day collection is still in process.

It should be noted that the Google Trends API is rate limited; to prevent IP blocking, each stock symbol was retrieved individually with a one second timeout between queries.

US Historical Stock Prices with Earnings Data (Kaggle) [x]

The secondary dataset used in this analysis was the US historical stock prices with earnings data provided within the online data science community, Kaggle. The dataset, accessible via the link above, provides three key datasets as independent files of comma-separated values (csv). This analysis was primarily concerned with the *earnings_latest.csv* file.

Table 2: Summary of utilized variables within the US Historical Stock Prices with Earnings Data

Variable Name	**Description**
symbol	Company stock symbol as a string
qtr	Fiscal quarter that the data is provided for
eps_est	Analyst estimate for the quarterly earnings per share
eps	Reported earnings per share for the given quarter and company

The *earnings_latest* document provided 20-years of historical earnings, for applicable companies, and contained 5,354 unique company stock symbols. Given that the Google trend data only dated back 5-years, and only current S&P 500 stocks were considered, a subset of the total earnings data was utilized.

Data Manipulation Methods

Google Trends

A key manipulation of the Google Trends dataset was compressing the data in a way that was conducive to merging the data with the secondary dataset. The secondary dataset contains

dates that correspond to the final month in a fiscal quarter with the relevant year in MM/YYYY format, whilst the Google Trends dataset contains dates in YYYY-MM-DD format which represent aggregated interest over time during 7-day periods. To provide the most granular data possible that was still consistent across quarters, the 7-day periods were grouped into one-month periods with the values being averaged. While this is not a perfect method - there are likely 7-day periods that span multiple months - it was utilized as the interest over time data is not publicly available on a per-day basis. Formatting the Google Trends data in this manner allowed for each earnings value to have three interest over time values, corresponding to the three months within the fiscal quarter.

Table 3: Summary of Google Trends manipulation to condense 7-day aggregate data into data points averaged over one month

date	MMM	AOS	...
2016-06-05	16	19	...
2016-06-12	6	19	...
2016-06-19	7	0	...
2016-06-26	11	19	...

→

date	MMM	AOS
06/2016	10	14.25
07/2016	...	...
08/2016	...	...
09/2016	...	...

After aggregating the data into values representing one-month of relative search interest, the data was then converted into a format that was advantageous for merging the datasets. In particular, a format that was *date*, *symbol*, *value* was desired to allow for merging on common date and symbol pairs. To achieve this manipulation the Pandas *melt* function was employed; a nuance to utilizing this function was to provide the date field under the *id_vars* parameter to ensure the correct dates remained tied to the symbol and value.

Table 4: Summary of Google Trends melt manipulation to convert data into single row entries

date	MMM	AOS	...
06/2016	10	...	...
07/2016	8.4	...	...
08/2016	6.25	...	...
09/2016	11.75	...	...

→

date	symbol	value
06/2016	MMM	10
07/2016	MMM	8.4
08/2016	MMM	6.25
09/2016	MMM	11.75

The queries made to Google Trends returned no missing data, however the *isPartial* field did denote that a data point was incomplete. Since these data points provided no additional context to the analysis - they contained data of the current week and financial results are not available at the same granularity - they were dropped from the dataset.

US Historical Stock Prices with Earnings Data (earnings)

The US historical stock prices with earnings data set contained a significant amount of data that was unnecessary due to the fact that the Google Trends dataset limited the time frame this analysis could examine. Thus, the first manipulations performed were to prune the earnings dataset to improve computational efficiency, as it is wasteful to perform operations on data that will be lost when the datasets are merged. This was done by first removing all data from before May 2016, as this was the first date available in the Google Trends dataset. This reduced the dataset by 53.1% from 160,660 rows to 75,293 rows. Next, the symbol column was filtered to only include stock symbols that were present in the S&P 500 index. This step further reduced the dataset by 88.2% from 75,293 rows to 8,872 rows. The resulting dataset was much more concise than the initial dataset and contained only data points necessary for the analysis.

This particular dataset had a number of missing values, particularly in the *eps*, and *eps_est* columns. Upon initial inspection of the data, 35.6% of *eps_est* values and 20.2% of *eps* values were missing. Once data points with dates before May 2016 were removed the percentage of missing *eps_est* and *eps* values fell to 14.6% and 0.6% respectively. Next, the earnings data was further refined to only include companies that are tracked in the S&P 500 Index. After companies not within the S&P 500 were removed, the percentage of missing *eps_est* and *eps* values fell to 0.6% and 0.2% respectively. Since the value this analysis is concerned with is the relationship between estimated earnings and actual earnings, rows without entries for *eps_est* and *eps*.

After the dataset was pared down to the essential data points, more data cleaning was necessary before merging was possible. The expected format of the *qtr* column was MM/YYYY date format, but a number of companies had both MM/YYYY date formats and strings explicitly representing the quarter (i.e. Q1, Q2, Q3, or Q4). To ensure that the datasets could be merged successfully the string values were converted to the correct date format. This presented a challenge as not all companies follow the same fiscal quarters. For example, Agilent Technologies has a Q1 ending on January 31st annually, whilst American Airlines Group has a Q1 ending March 31st.

Merging Datasets

With the two dataset cleaned and mutated to a state close to permitting a simple merge, a few adjustments were made to preserve the greatest amount of data. With the data in the state described by the right-hand schema shown in *table 4*, the quarter to MM/YYYY mapping created during the formatting of the earnings dataset to aggregate all months within a shared fiscal quarter into a single row.

Table 5: Summary of final Google Trends DataFrame, where asv is an acronym for Average Search Volume, and the supercedeing integer represents the ordered month of the fiscal quarter

date	symbol	value
06/2016	MMM	10
07/2016	MMM	8.4
08/2016	MMM	6.25
09/2016	MMM	11.75

→

date	symbol	asv1	asv2	asv3
06/2016	MMM	...	...	10
09/2016	MMM	8.4	6.25	11.75
12/2016	MMM	...	...	...
01/2017	MMM	...	...	...

Formatting and manipulating each of the two datasets with the end goal of merging them allowed for a simple final merge operation. Once the tasks outlined above were completed, the two datasets were prepared to merge on a *date, symbol* joint key. Performing an inner merge between the two datasets generated a DataFrame with a schema of *date, symbol, asv1, asv2, asv3, eps_est,* and *eps*.

Analysis and Visualization

Relationship Analysis via SPLOM Visualization

The goal of this analysis was to determine whether search volume was a valid indicator for a company's quarterly earnings per share. The first step in exploring the relationship between search volume and eps was visual inspection of the data. To aid in this inspection a number of graphical visualizations were generated.

To perform simple visual analysis the *eps_est* and *eps* were condensed to a single value *result*. This value was +1 if *eps* was greater than *eps_est*, 0 if they were equal, and -1 if *eps* was less than *eps_est*. To investigate how average search volume for a month (ASV**N or average search volume for the N-th month of the quarter*) relates to the *result* field a scatter plot

matrix (SPLOM) was generated with ASV*N as the X- and Y-coordinates and result encoded via color. The SPLOM suggested that while there may be a linear relationship between the pairs of average search volume (i.e. ASV1 with ASV2, ASV1 with ASV3, etc.), given the distribution of color (encoding the result) the search volume doesn't seem to inform the earnings results.

The results of the SPLOM analysis seemed to initially suggest that the search volume did not have a notable relationship with the earnings result. To further investigate this a subset of the 504 stock symbols were selected to generate an additional SPLOM. The subset was the top 10 companies within the S&P 500 by index weight. These top 10 companies represent 27.2% of the S&P 500 index value which motivated their selection (Alpert, 21). The resulting plot yielded similar results to the first SPLOM of all 504 stock symbols; however, there were some regions, toward the upper right of the three in *Figure 2* plots, that were solely composed of positive earnings reports.

Figure 1: Isolated view of scatter plot matrix (SPLOM) of all S&P 500 companies reduced from 9 figures to 3 unique views for brevity. Full SPLOM available in EarningsEDA.ipynb

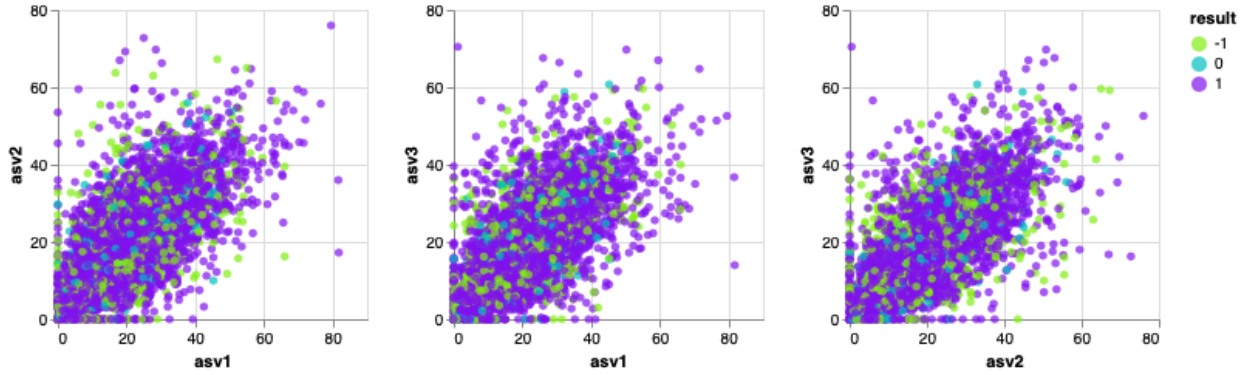
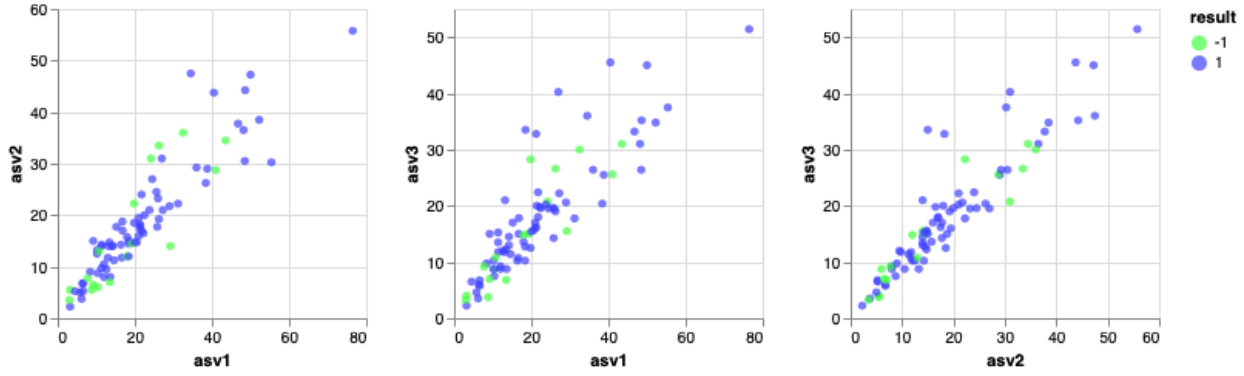


Figure 2: Isolated view of scatter plot matrix (SPLOM) of top 10 S&P 500 companies by index weight reduced from 9 figures to 3 unique views for brevity. Full SPLOM available in EarningsEDA.ipynb



Investigation of Predictive Models

With the intent to generate a predictive model, a number of slices of the data were examined independently. The first slice of the dataset examined was over different fiscal quarters to examine whether the relationship between ASV and *eps* was strong for certain quarters. This analysis was performed once for each quarter, and for each of the three individual months defining the quarter; an additional analysis was performed on the summed ASV over the three months.

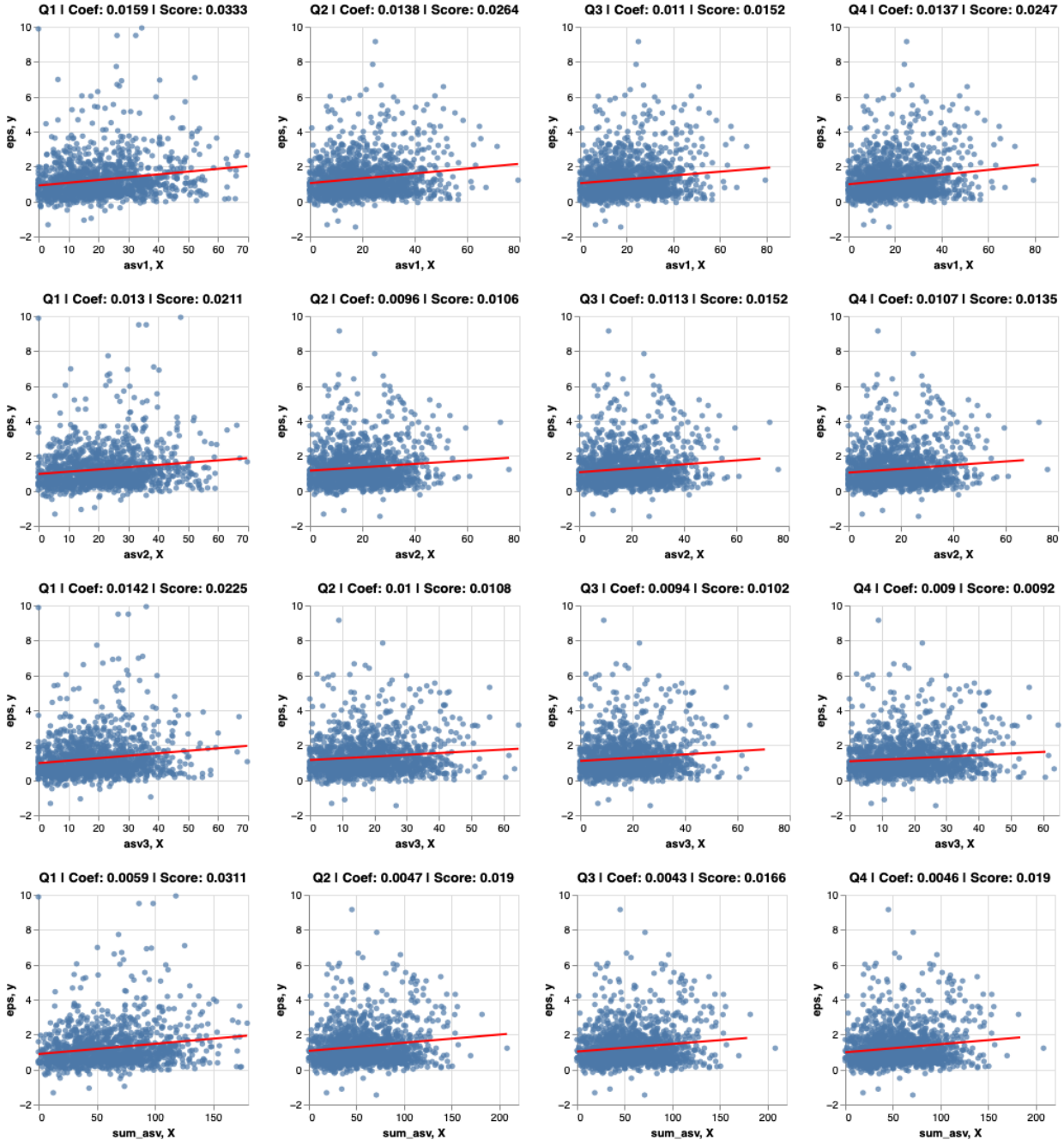
Examining each of the 16 linear regressions that were performed, the R^2 for all of the fits were notably small, with a mean of 0.0101 and a standard deviation of 0.0035. Furthermore, the slope of the best fit lines had a mean value of 0.0187 and a standard deviation of 0.0072. Finally, visually there does not appear to be significant differences between any of the 16 scatter plots. It does not appear that asv explained the variance in eps within any particular slice that was examined.

Analysis Conclusion

While there may be some intuitive suggestions that raw public interest could serve as an indicator for a company's quarterly performance - intuition supported by fact that performance is driven by revenues derived from consumer action - there was no indication in the data collected to support this.

In future extension of this work including a general sentiment analysis for a stock symbol, or company, could elucidate the goals proposed in this analysis. This sentiment could be derived from public news and opinion journalism, or social media streams. Perhaps linking sentiment to search volume would provide stronger relationships with quarterly financial performance, and be more constructive for generating a predictive model.

Figure 3: Visualization of asv and eps sliced on fiscal quarter. Linear regression was performed for each, with the best fit line plotted in red; the title of each chart includes the relevant quarter, the regression coefficient, and the R^2 score of the fit.



Statement of Work

Both Nicolas and Jason worked collaboratively to identify key data sources. Once identified, Jason collected the Google Trends data for the necessary company stocks and created a central document of the data. Nicolas led the task of merging the two data sources into a single source.

Once the datasets were merged, Nicolas perused the data with the aim of creating a predictive model. Jason visually examined the data and generated the visualizations presented in the report.

For the final report, Jason wrote the first draft and Nicolas provided edits, corrections, and code artifacts used in his analysis.

References

Alpert, G. (21, May 8). *Top 10 S&P 500 Stocks by Index Weight*. Investopedia.

https://www.investopedia.com/top-10-s-and-p-500-stocks-by-index-weight-4843111

Marr, B. (2018). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Forbes.

https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=3d4b1a5360ba

Terzo, G. (2016). *What Do Quarterly Earnings Mean for Stocks?* Zacks Investment Research.

https://finance.zacks.com/quarterly-earnings-mean-stocks-3875.html

Tuovila, A. (2020). *Guide to Company Earnings*. Dotdash.

https://www.investopedia.com/terms/e/earningsreport.asp