NYC Weather and Trading Behavior

CSPB 4502

Connor Tree
CSPB
CU Boulder

Tara Panzarino
CSPB
CU Boulder

Sam Steingard
CSPB
CU Boulder

ABSTRACT

Global weather and the US stock market are both highly volatile systems that depend on a variety of independent factors. Throughout this paper, the NOAA API and the Kaggle stocks database will be used in combination with state-of-the-art data mining techniques, to try and answer a few different questions regarding patterns that exist in the weather/stocks data. These questions include:

- Does inclement weather in NYC have an effect on stock trading activity?
- How do seasonal weather patterns affect stock trading behavior?

It is pretty easy to see the potential benefits of doing this kind of analysis on these datasets - humans care about the weather and how it affects their lives. For example, a stock broker might want to use this research to adjust his trading strategies or choose when to buy or sell a stock. The dataset from NOAA also includes all regional atmospheric measurements (temperature, pressure, humidity, etc.) so this should be handy when trying to answer questions about seasonal weather and its effect on certain industries or companies - things that a farmer, government policymaker, or lobbyist might like to know for the future.

1 Literature Survey

A key study that this work will be based on was done by Edward Saunders, stock prices of various New York exchanges were compared to the percentage of cloud cover during the day. Cloud coverage was binned into 3 main groups: 0-30, 40-70, and 80-100 which were then compared to the mean percentage daily change. The study concluded that cloud coverage in the 80-100% range was significantly different from that of the 0-30% cloud coverage days and that there was a surprisingly large economic effect produced by the weather [3]. Building upon this work, a study in 2015 sought to determine whether this trend was observable globally and compared data from 2011-2015 of cloud coverage, humidity, and temperature to data of financial hubs in Asia to their respective exchanges. The results of the study showed that the effect of weather was insignificant and could not be distinguished from market trends [4]. Further work has been done comparing the S&P 500, NYSE, NASDAQ, and various other United States based exchanges in comparison to larger geographic regions which have found conflicting results [1, 2].

2 Proposed Work

Full historical data for the NYSE was obtained with information of over 7000 individual stocks and 1300 exchange traded funds (ETFs). As ETFs are a better indicator of overall market performance those will be selected rather than using individual stock data. To track daily performance, percentage gain had to be calculated by taking the difference between the previous day's closing price and the current day's close then determining the percentage gain. The NOAA data was joined to the stock market data by date so that only days in which the markets were in operation were included in the weather data. To differentiate from the previous studies, daily precipitation/snowfall will be used to compare to the stock market rather than cloud coverage or temperature. The data can then be further analyzed by comparing the trends for each successive decade to see whether computers and high frequency trading (HFT) has lessened the effect of weather on trading over time. Changes in these trends will indicate why previous studies from the 1990s showed observable weather effects while modern studies proved inconclusive.

3 **Datasets**

NOAA Dataset (NOAA) - This is an API from NOAA/NCEI that can be used to obtain weather data for any location or time in the US. The API can be accessed using Python and returns a JSON response that can be transformed into any database type that we choose to use. The API can be used to fetch all kinds of atmospheric data (temp., precip., # of rainy days, etc.) and allows a user to choose which attributes are included in the response. This dataset is very comprehensive and the main challenge will be working efficiently with the large amount of data it contains.

US Stock Market Dataset (Kaggle) - This dataset contains thousands of text documents that contain historical stock data for all companies listed on the US stock market. Each individual text document contains attributes for the company that includes daily open/close prices, trading volume, high/low prices and the date. Luckily, these files are all in one folder and named for the company who's data they contain so searching for the right dataset should be a trivial task.

4 **Evaluation Methods**.

Pattern Evaluation will primarily be done with Python. Almost all data is in decimal form, allowing for straightforward use of these values for data analysis. Data points can consist of Numpy vectors to be used for clustering and classification.

Data will undergo least squares classification to predict stock behavior outcomes. Vectors will consist of weather data for each day. A boolean classifier will be used to predict the outcome, which will have two options such as a positive percent change from the previous day's close or a negative change from the previous day's close. A confusion matrix will be used to evaluate the model and calculate error rate, true positives, false positives, true negatives, and precision.

Data will undergo clustering using the K-Means algorithm, partitioning vectors of data into distinct clusters. In this case, clustering will be used to view associations between weather conditions and stock behavior. There are many possibilities of vectors of data to cluster. Vectors consisting of both weather and stock data can be included into a vector of one day's data, or vectors can consist of either daily weather or stock data to correspond to an overall value for stock behavior or weather behavior, respectively. The number of groups will be chosen based on what insights are to be explored. The K-means algorithm from *Introduction to applied linear algebra: Vectors, matrices, and least squares* will be utilized:

> Given a list of N vectors $x_1,...,x_N$, and an initial list of k group representative
>
> vectors $z_1,.., z_k$ repeat until convergence
>
> 1. *Partition the vectors into k groups*. For each vector $i = 1,...,N$, assign x_i to the group associated with the nearest representative.
>
> 2. *Update representatives*. For each group $j = 1,..., k$, set z_j to be the mean of the vectors in group j. [5]

Vectors of data will be assigned based on the euclidean distance between itself and the representative vector. Clusters will be judged based on the mean square distance from the data vectors to the representative in that cluster, with the goal of getting that value as close to zero as possible [5]. These clusters will be plotted and visually analyzed along with the consideration of a low mean square distance to their respective representatives.

5 Tools

The tools to be utilized for this project can be broken down into the purpose of data visualization, data analysis, and data interchanging. For data visualization purposes, Python will be used, specifically the Matplotlib and Plotly libraries to plot data of interest. Additionally, PowerBI and Tableau will be used in addition for data visualization. For data analysis purposes, Python will be used, specifically the Numpy and Pandas libraries. The functionality to work with numpy arrays as vectors of data is integral to the pattern evaluation that will be done with k-means clustering and classification analysis. The Json package in Python will be used to parse and access data that is formatted in the Json format. The data analysis, data interchanging, and much of the data visualization will be done in Python, allowing for streamlining of tasks.

6 Milestones

Finish Data Preprocessing - *3/22/22*

- By this point the data selection, cleaning, combination should be completed and the next step will be to begin writing code to analyze/display the data

Finish Code - *4/1/22*

- The Python code used to display and analyze the data should be finished by this point in order to allow enough time for analysis

Finish Data Analysis - *4/14/22*

- Analysis of the results of the code should be completed at this time - trends should be identified and questions answered

Finish Presentation/Final Report - *4/25/22*

- At this time the report and presentation that summarize the results will be complete

REFERENCES

[1] Chandra, M., 2021. *Weather Effects on Stock Market Returns in the United States*. [online] University of New Hampshire Scholars' Repository. Available at: <https://scholars.unh.edu/honors/585/> [Accessed 13 March 2022].

[2] Ong, B., 2016. *Weather vs. the Stock Market*. [online] Data Science Blog. Available at: <https://nycdatascience.com/blog/student-works/nyc-weather-affect-stock-market/> [Accessed 13 March 2022].

[3] Saunders, E., 1993. *Stock Prices and Wall Street*. [online] Www-jstor-org.colorado.idm.oclc.org. Available at: <https://www-jstor-org.colorado.idm.oclc.org/stable/2117565?pq-origsite=summon&seq=1#metadata_info_tab_contents> [Accessed 13 March 2022].

[4] Wang, Y., Shih, K. and Jang, J., 2018. *Relationship among Weather Effects, Investors' Moods and Stock Market Risk: An Analysis of Bull and Bear Markets in Taiwan, Japan and Hong Kong*. [online] Pdfs.semanticscholar.org. Available at: <https://pdfs.semanticscholar.org/893d/02411f9bd303b06124eb11fad126553ff43d.pdf> [Accessed 13 March 2022].

[5] S. Boyd and L. Vandenberghe, *Introduction to applied linear algebra: Vectors, matrices, and least squares*. Cambridge: Cambridge University Press, 2019.