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.

We hypothesized that poor weather conditions such as rain or snow could have an effect on general sentiment on the stock market. We also seek to determine whether seasonal weather patterns have any effect on the agricultural sector of the stock market. After thorough data analysis our results were inconclusive, there was no significant correlation between inclement weather and that of the percent change in the stock market that we could find. However, agricultural ETF stock data indicated that there was a very slight positive correlation with precipitation and both stock price and volume.

1 Introduction

The questions we sought to answer were twofold:

- 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 - weather has a pronounced effect on the human psyche and mood. Numerous studies have linked the effects of weather on feelings of depression or anxiety, Seasonal Affective Disorder (SAD) has been shown to cause depression in winter months when there is little sunlight. For example, a stock broker might want to use this research to adjust their 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.

2 Related Works

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].

3 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.

4 Datasets

NOAA Dataset (NOAA) - This is an API from NOAA/NCEI that was used to obtain weather data for any location or time in the US. The API was accessed using Python and returned a JSON response that was transformed into a pandas dataframe. 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 lists 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.

5 Evaluation Methods

Pattern Evaluation and Classification is done with Python. Almost all data is in decimal form, allowing for straightforward use of these values for data analysis. Data points consist of Numpy vectors to be used for clustering and classification. Pandas Dataframes contain corresponding precipitation, maximum temperature, minimum temperature, stock volume, percent change, and percent change binned values for a date. Scikit-learn library was also used for Linear Regression analysis, and Matplotlib and Seaborn was used to plot data.

Firstly, z-score was calculated for all numerical data. This allows for us to see the relation between the data from that day and the mean data. Z-score data was plotted using Matplotlib for visual comparison.

Boxplots were created using the % change values from Agricultural EFT data, S&P 500 data, and Top 10 ETFS data in order to view extreme outliers. Seaborn was used to create these plots.

Linear Regression analysis was done to see relationships between Volume or % change with Precipitation, Minimum Temperature, and Maximum Temperature. The regression line is then superimposed on the plotted observations to determine if there is a relationship between stock and weather data.

Data underwent least squares classification to predict stock behavior outcomes. Vectors consist of weather data for each day. A boolean classifier is 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. This classifier is in the form of $f_{hat}(x) = \mathbf{sign}(x^T \beta + v)$[5]. :Performance metrics from *Introduction to applied linear algebra: Vectors, matrices, and least squares* are calculated:

- The *error rate* is the total number of errors (of both kinds) divided by the total number of examples, *i.e.,* $(N_{fp} + N_{fn})/N$.
- The *true positive rate* (also known as the sensitivity or recall rate) is N_{tp}/N_p. This gives the fraction of the data points with y = +1 for which we correctly guessed ^y = +1.
- The *false positive rate* (also known as the false alarm rate) is N_{fp}/N_n. The false positive rate is the fraction of data points with y = -1 for which we incorrectly guess ^y = +1.
- The *specificity or true negative rate* is one minus the false positive rate, i.e., N_{tn}/N_n. The true negative rate is the fraction of the data points with y = -1 for which we correctly guess ^y = -1.
- The *precision* is $N_{tp}/(N_{tp} + N_{fp})$, the fraction of true predictions that are correct.[5]

For future exploration, data can undergo clustering using the K-Means algorithm, partitioning vectors of data into distinct clusters. In this case, clustering can 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 may 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 would be assigned based on the euclidean distance between itself and the representative vector. Clusters would then 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 could be plotted and visually analyzed along with the consideration of a low mean square distance to their respective representatives.

6 Milestones Completed

Finished Data Preprocessing - *4/16/22*

- The final datasets have been pulled from the internet and cleaned to the level that is useful for further analysis
- Data was cleaned of NaN values, cast to the proper types, in dataframe format with ISO date indices

Finish Code - *4/20/22*

- In order to finish the code, sections need to be added that include further data analysis of the datasets
- A stretch goal for this project would be to add another dataset containing data relating to the second farming/seasonality question

Finish Data Analysis - *4/24/22*

- Final analysis needs to be performed, now that the precursory data search has been completed the final analysis can target questions we obtained from it

Finish Presentation/Final Report - *4/28/22*

- The presentation draft has been started but needs the final analysis done in order to finalize it for submission

7 Data Preprocessing

For the initial feasibility study, a subset of data was selected to perform the data cleaning and analysis. Weather data from a NOAA weather station in Central Park from 1970-2020 was pulled from NOAA using their API. Originally, the data was being obtained dynamically from the NOAA database as a .dly file, but translating this file type proved to be time consuming and confusing for what the data was needed for. Instead, a CSV file that contained just the information needed was downloaded from the NOAA NCEI data catalog/online data ordering system. From this file, a Pandas Dataframe was created with columns for date, maximum temperature, minimum temperature, precipitation, and snowfall. A new column, 'Total Precipitation', was generated by summing the values for rainfall, snowfall and current snow depth to account for days in which both could occur and also account for days with large "pile-ups" of snow. All units in the CSV file were measured in inches so unit conversion was unnecessary. Total precipitation was then binned into 3 separate bins, low, medium, and high with (0-1), (1-4), (4-100) inches of total precipitation respectively. These binned precipitation numbers will be used more as the code is finished and further analysis is completed/

The subset of stock data chosen was the SPDR S&P 500 ETF Trust (SPY ETF) which is composed of the top 500 largest U.S based stocks. This ETF was selected as it represents the general market sentiment and reflects overall behavior of the market. This data pulled from our Kaggle dataset spanned from 2005-2017 which totaled 3201 active trading days. The dataset was also read into a Pandas Dataframe with columns for date, volume, open, and close. A percent change column was then calculated by taking the difference between 'Open' and 'Close', then dividing by 'Open' to normalize the data.

Both datasets were then combined using a left join on *df_stocks* and *df_weather* to generate a combined dataset of the weather on each active trading day with values for date, volume, % change, max temperature, min temperature, total precipitation, and weather category.

After initial proof of concept studies on S&P 500 data proved successful, we moved to expand that to a larger dataset to hopefully see more concrete results. From the same dataset obtained from Kaggle, the top 10 largest ETFs by assets under management (AUM) were selected in order to do our analysis on. These 10 ETFs included both large, medium, and small cap companies, bonds, and other types of assets to give a much more rounded view of the economy. These 10 ETFs were grouped by date and their average value taken as the average percent change over the entire market during that day. This data was then joined again with the NOAA weather dataset to generate a combined dataframe of the weather and subsequent percentage change of the stock values for that day.

After analyzing the overall market picture we wanted to see the effects of weather on the agricultural sector. To do this we selected the largest ETF that represented the agricultural sector, the Invesco DB Agriculture Fund (DBA). This data again was cleaned to replace NaNs with 0 values, had a percent change calculated, and binned into various weather categories. It was again merged with the NOAA weather data to compare weather on each day with the DBA performance.

8 Tools

The tools that are utilized for this project can be broken down into the purpose of data visualization, data analysis, and data interchanging. For data visualization purposes, Python was used, specifically the Matplotlib, Plotly, and Seaborn libraries to plot data of interest. For data analysis purposes, Python will be used, specifically the Numpy, Pandas, and

Scikit-learn libraries. The functionality to work with numpy arrays as vectors of data is integral to the pattern evaluation that was done for the least squares classification analysis and for any future clustering using k-means algorithm. The Json package in Python was used to parse and access data that is formatted in the Json format to create the Pandas Dataframes. The data analysis, data interchanging, and much of the data visualization was done in Python, which allowed for streamlining of tasks.

9 Techniques Applied

Z-Score calculation

- Z-Score calculations from preprocessed stock data and preprocessed weather data were calculated.
- By viewing all Z-Score data, we are able to see if any weather conditions that differ greatly from the norm visually correspond to stock data that differed greatly from the norm.
- From our initial plots we are unable to determine whether there is a correlation between weather conditions and stock behavior.

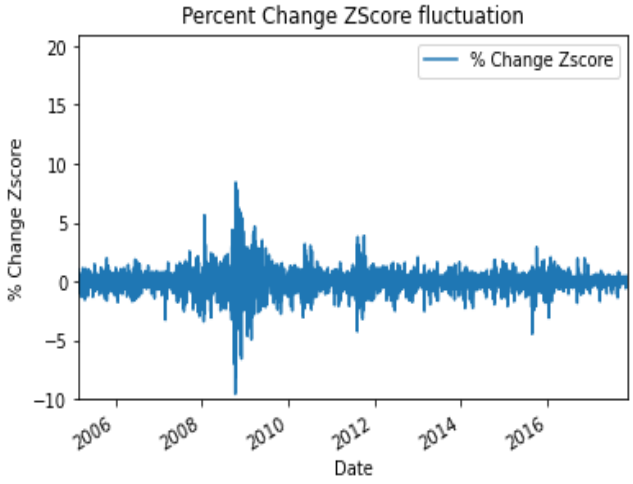


Fig. 1

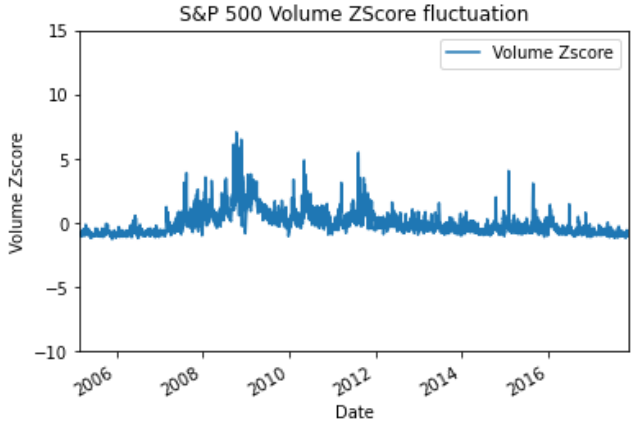


Fig. 2

Following the 2008 financial crisis we see significant increases in fluctuations of both volume and percent change, as seen in Figure 1 and Figure 2 above, indicating that there was significant price volatility during that period of time. This then levels out around 2010 which is representative of economic stabilization and recovery.

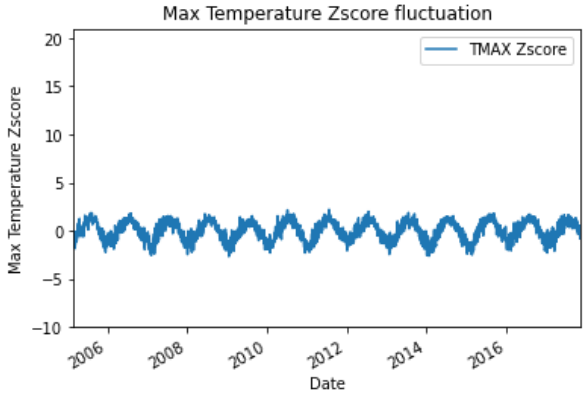


Fig. 3

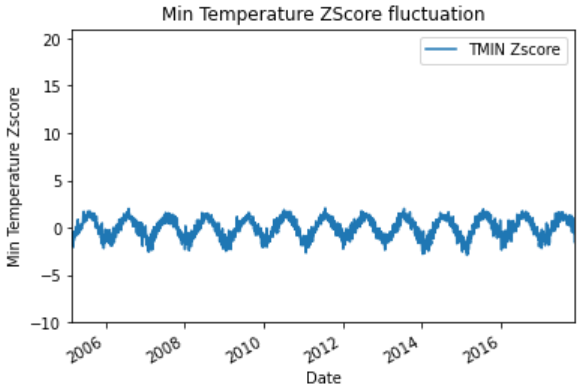


Fig. 4

The plotted z-scores of temperature are also expected, showing the swings in temperature based on the changing of seasons with each year. This can be seen in Figure 3 and Figure 4 above.

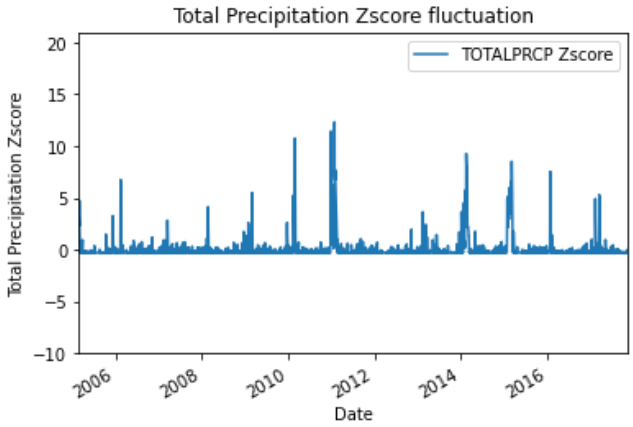


Fig. 5

Total precipitation is also shown to be cyclical with more precipitation in the winter and spring months. This is seen in Figure 5 above.

Outlier Detection

- Boxplots were created using the % change values from Agricultural EFT data, S&P 500 data, and Top 10 ETFS data.
- Extreme outliers can be seen in these visualizations.

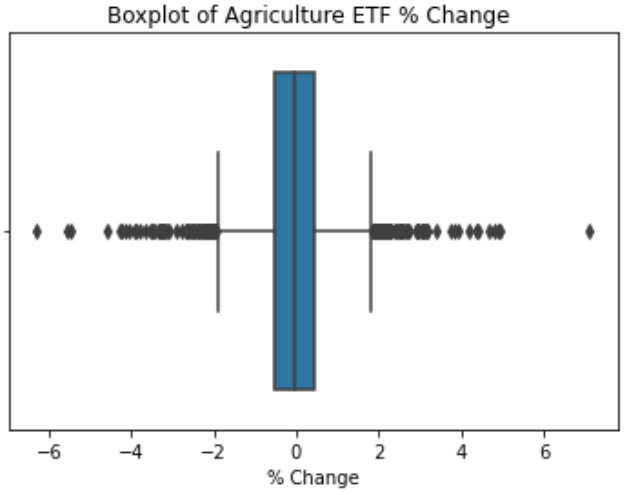


Fig. 6

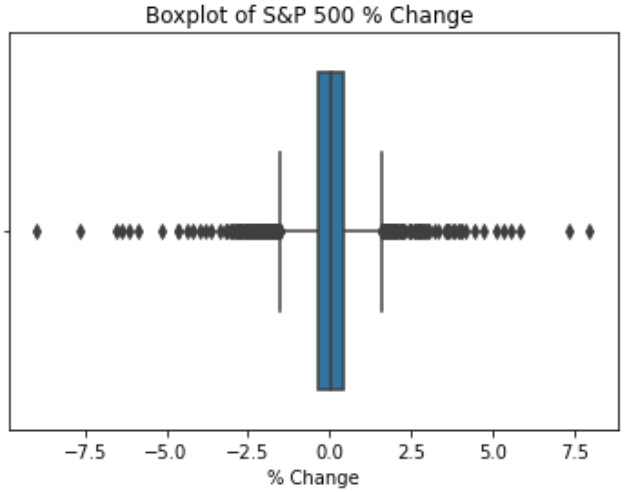
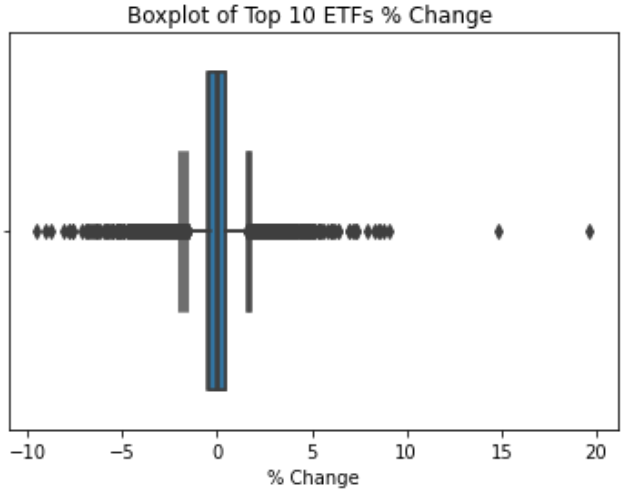


Fig. 7 (Above) & Fig. 8 (Below)



From these visualizations (Figure 6, Figure 7, and Figure 8) we can see that there is a large amount of outlier data for all three datasets, which can be used for further analysis. These outliers may be days where there were significant world events or a day that an industry/company was hit with new regulations and this kind of information could prove useful if further research is conducted. We can also see that there seems to be outliers with large positive % change values, particularly with the top 10 ETFs. This might indicate that ETFs are more volatile but in a positive way for investors - something that would also warrant further research.

Linear Regression

- Linear Regression calculations were done, using total precipitation, minimum temperature, and maximum temperature as independent variables.
- Dependent variables included stock percent change initially and later stock volume.
- So far we have only seen flat regression lines, suggesting that there is no correlation between high/low temps or precipitation values and stock market % change.

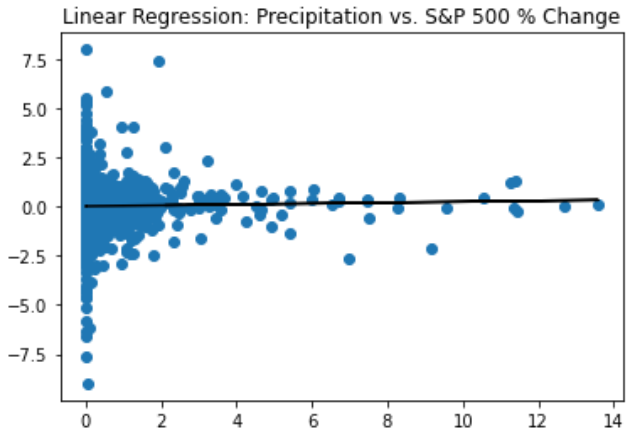


Fig. 9

In Figure 9 we see total precipitation plotted against stock percent change. If they were correlated we would see a linear relationship between them. Here we see a large spread of data points clustered around with no obvious relationships between them.

Interestingly, the data seems to be skewed heavily towards the “average” or low-precipitation days. Initially, it was thought that the days with the highest amount of precipitation in NYC would be the days with the most volatility, but we clearly see that is not the case and the most volatile days occur when there is no precipitation at all. This will definitely require further analysis and consideration - and we may want to consider adding more datasets to the analysis.

Below are the charts obtained from graphing the daily max temperatures against the daily percent change values.

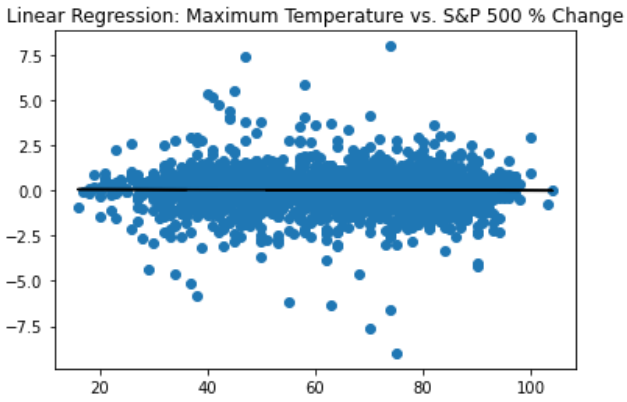
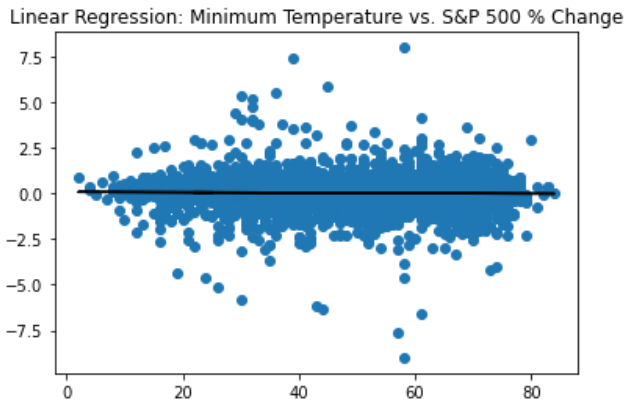


Fig. 10 (Above) & Fig. 11 (Below)



There is also no discernible relationship when comparing temperatures to stock percent changes, as seen in Figure 10 and Figure 11 above. There is a

significant spread of temperatures for each percent change in stock price with no obvious trends.

However, we can also tell that there is a slight bell-shape to these graphs and the values seem to be "highest" around the approximate center of the temp. values. This means to say that the most volatile days on the stock market are also the most temperate or moderately hot or cold. This tracks along with the results obtained from the other analysis on the precipitation values. This definitely warrants further research because if this is true, and the most boring weather days are the worst on the stock market, it would be interesting to try and figure out why that occurs.

Least Squares Classification

- Least Squares classification was performed, using total precipitation, minimum temperature, and maximum temperature as independent variables.
- Percent change was binned into two values: positive and negative and used to create our classifiers.
- Performance metrics were calculated to determine quality of least squares classifier, including error rate, true positive, false positive, true negative and precision based on classification of positive and negative % change values.

	Calculated from Positives	Calculated From Negatives
Error Rate	0.466886	0.466886
True Positive	0.970954	0.041181
False Positive	0.958819	0.0290456
True Negative	0.041181	0.970954
Precision	0.532221	0.557895
Final Error Rate	0.533114	

Table 1 - Agriculture ETF Performance Metrics

	Calculated from Positives	Calculated From Negatives
Error Rate	0.462668	0.462668
True Positive	1	0
False Positive	1	0
True Negative	0	1
Final Error Rate	0.462668	

Table 2 - S&P 500 Performance Metrics

	Calculated from Positives	Calculated From Negatives
Error Rate	0.480961	0.480961
True Positive	0.0304905	0.972518
False Positive	0.0274815	0.96951
True Negative	0.972518	0.0304905
Precision	0.507353	0.519387
Final Error Rate	0.519039	

Table 3 - 10 Largest ETF Performance Metrics

As seen in Table 1, Table 2, and Table 3 above, It is interesting to note that the performance metrics varied when positive or negative values were chosen as the basis of the analysis. Using this method of creating a classification model was inconclusive. The final error rate was close to 50% for classification of all 3 sets of data. It is possible that the relationship between weather conditions and stock data is not linear, so linear analysis may not yield conclusive results.

10 Key Results

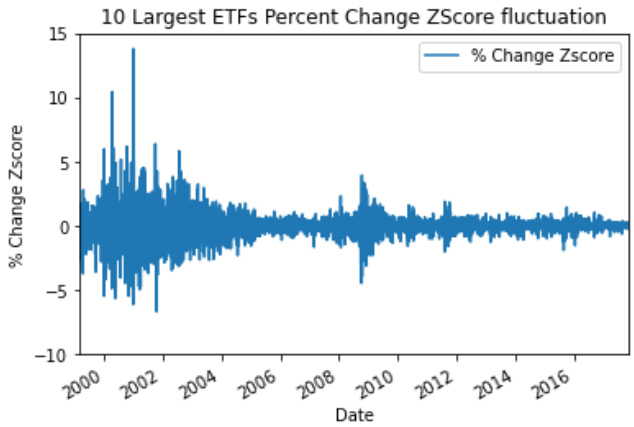


Fig. 12

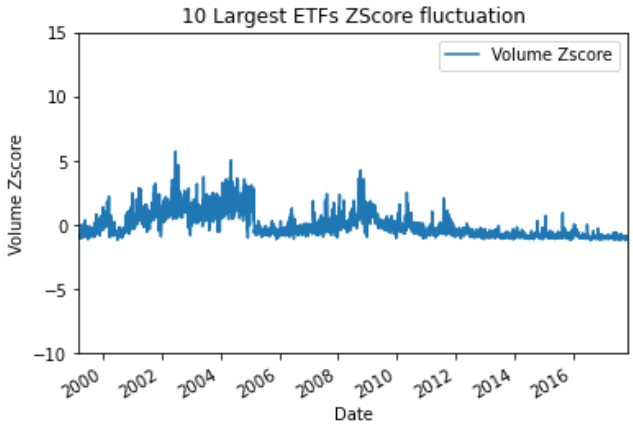


Fig. 13

Results from our aggregate datasets in Figure 12 and Figure 13 above revealed similar trends to that of our initial feasibility study from the S&P 500 data. Z-score and volume showed significant fluctuations in values in times of economic instability such as those that occurred in 2001 and 2008. These again are indicative of volatility in the market due to the uncertain economic outlook of those time periods.

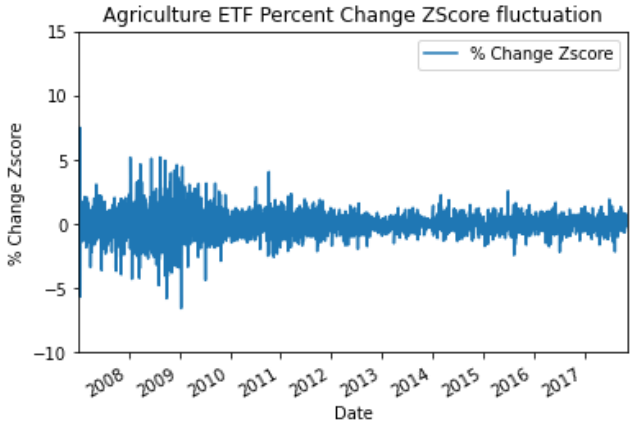


Fig. 14

Contrasting that with the Z-scores of percent change and volume for the agricultural sectors as shown above in Figure 14, there was significantly less volatility and fluctuations during this time. As you can see from the period of 2008 onwards, while there are still noticeable effects on both graphs, they are less drastic and smaller in measure. This can be attributed to the fact that agriculture is seen as a safer investment due to its nature as a consumer staple. Whereas during a recession, the first stocks to be sold are high growth stocks like technologies and other consumer discretionary stocks which make up the majority of high valued companies.

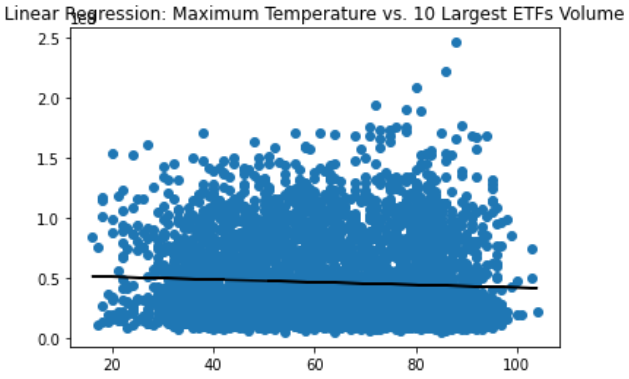


Fig. 15

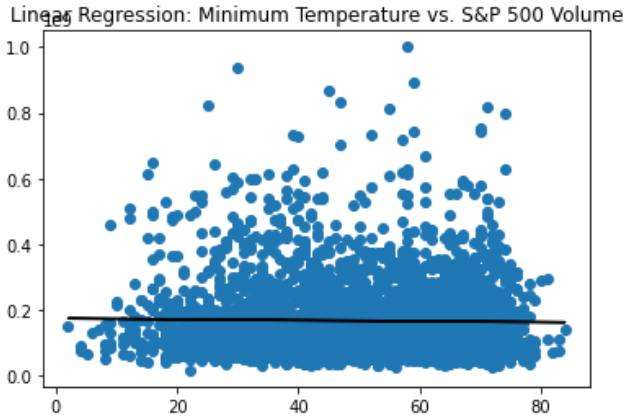


Fig. 16

Looking at the temperature variations in Figure 15 and Figure 16, we can see an interesting pattern, days with higher temperatures are slightly correlated with having lower volumes. This can both be seen in the minimum and maximum temperature comparisons. This provides minor support to our hypothesis that weather can have an effect on the stock market. Higher temperature days could indicate lower volatility as there is less trading activity during those days. While this is a very slight correlation, it is significantly larger than that of the S&P 500, indicating that this is a wider market trend only visible due to the increased amount of data.

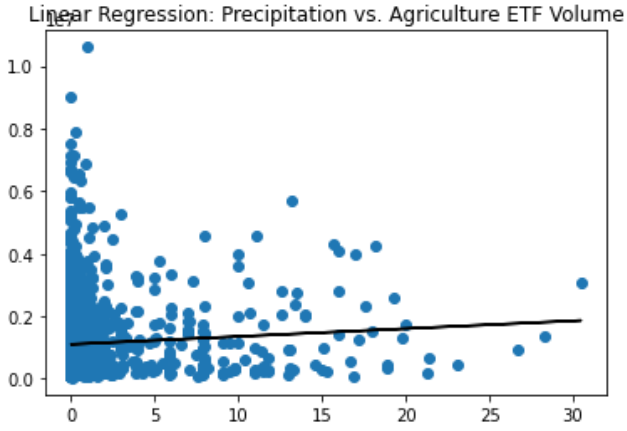


Fig. 17

When analyzing the results from agricultural data shown in Figure 17, we can see a slight positive correlation between the levels of precipitation and volume, but no correlation between precipitation and percent change. This somewhat supports our claim that weather can have effects on the agricultural stock market but more analysis is necessary to understand why there is no pronounced effect on the change in price while there is a change in volume.

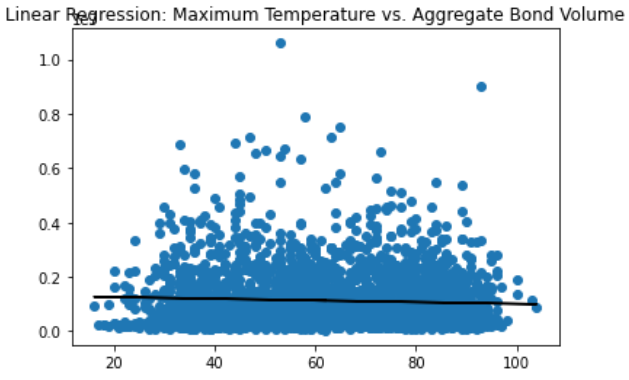


Fig. 18

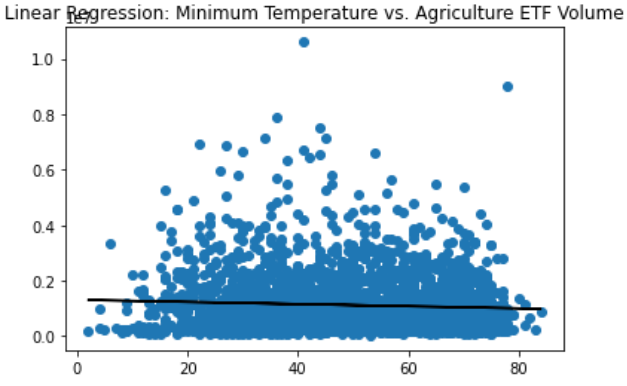


Fig. 19

While the precipitation data showed positive correlation, the temperature data paints a slightly different picture. Both minimum and maximum temperatures, shown in Figure 18 and Figure 19 above, when plotted against agricultural ETF volume showed a slight negative correlation between them. Combining the results showed that while the weather was warmer, the volume of agricultural stocks tended to be lower and when temperatures were cooler there was an increase in volume. This again can be attributed to seasonal fluctuations in the weather, it seems that cooler seasons such as fall and winter led to more trading of agricultural stocks while warmer weather had less trading. This could be that agricultural tools and capital must be purchased prior to the growing seasons and thus have better financial performance during these months. This better financial performance could lead to higher trading activity, but again more detailed analysis would have to be done to draw any significant conclusions.

While previous studies have shown correlation between weather patterns and stock price, we conclude that due to the increase in High Frequency Trading (HFT) and various computerized algorithms for daily trading, the effect of the weather has diminished significantly on that of the stock price. Much of the human element has been removed from stock trading and thus weather patterns have little effect on the modern, data driven technical analysis of the stock market. Our data focused mainly from 2000 - 2017 has shown negligible relationships between

stock price and inclement weather. We can conclude, however, that seasonal weather patterns can have an effect on the volume of trades but this information is not actionable and would not generate significant revenue.

11 Applications

While our analysis proved inconclusive, the applications of this type of work are myriad. It is an undeniable fact that weather events have some effect on the performance of the stock market, and financial institutions are always seeking the next leg up on their competitors. Analyzing possible relationships and correlations further could prove beneficial for the next iteration of quantitative trading algorithms. Our research proved that there are some correlations that exist between these two data sets and with further data collection it would definitely be possible to find stronger correlations.

Being able to predict stock behavior is crucial in building a robust predictive model that can accurately predict stock prices and trends. Although our predictive model was not very accurate and prone to high rates of error, we believe that with further data clarification, classification and time an accurate algorithm could be developed. This model would be helpful for companies/brokerage firms trying to understand the near and long term effects of weather patterns and events.

This algorithm could be potentially useful for farmers as well if it were adapted to track with the price of nutrients, water, etc.. These are all goods that fluctuate in price and change depending on environmental conditions like weather or weather events. A future version of the model we developed could potentially even track global weather patterns to increase the scope of the results and provide users with predictive price information built upon world-wide data. There are lots of potential applications for this kind of research and given more data/time, results would have proven more fruitful.

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.