ECON422 Final Paper

Rohit Krishnagopal

December 11, 2020

1 Introduction

This paper will examine the existence of a correlation between trading volumes and weather and the existence of a mood effect caused by weather. Specifically, we will be looking at trading volumes of the SPY index ETF from 2017 to 2020 and corresponding weather data from New York City and Chicago. The reason for choosing the SPY index ETF is since this is historically one of the most popular stocks traded on a day to day basis. New York City and Chicago are chosen as they are financial hubs in the United States. Daily trading volume is a proxy for market liquidity and better informs traders about how easily they can get into or out of a position within a day. Weather can often influence mood and in this case it may cause a difference in daily trading volumes. Analyzing this correlation between trading volume and weather will allow us to see if there is a mood effect caused by changes in weather and what factors contribute the most.

2 Data

To perform the regression, both Weather Data and Stock Volume Data is needed. The Weather data is from the National Oceanic and Atmospheric Association. This data is compromised of daily reports from weather stations in New York City and Chicago from 2017 - 2020. The variables used from these data sets are average wind speed(AWND), precipitation(PRCP),snowfall(SNOW), and average temperature(TAVG). The historical SPY volume data is from the NASDAQ website.

With the original weather data sets being messy, I had to spend time cleaning them before joining them to the volume table using python. For both the Chicago and New York City weather data I dropped metadata columns such as the station name, coordinates, and station ID. Following this I dropped any rows with null values. The last step was to group the data by date. Originally the weather data for each city had multiple stations reporting data. This meant there were multiple rows that had the same date. To solve this I grouped the data by date and took the mean of the data for each date. The final step was to join the weather data to the volume table by date. We are left with a table indexed by date with five columns: AWND, PRCP, SNOW, TAVG, and Volume.

3 Methods

I ran two different sets of regressions. One comparing trading Volume to weather in New York City and the other comparing trading Volume to weather in Chicago. In each one of these sets I ran four regressions each with their own specifications. The full model for both cities as is as follows:

New York: $Volume = 128704534.498 + 330220.038 AWND + 7586109.017 PRCP - 2181069.593 SNOW - 768077.182 TAVG$

Chicago: $Volume = 118057322.606 + 376782.102 AWND + 70825.921 PRCP - 6219621.294 SNOW - 607599.195 TAVG$

There are a few possible problems with this estimation strategy. One issue is the presence of heteroskedasticity. As shown in Figures (1 - 2) the residuals have high variance when Snowfall and Precipitation are near 0. As the magnitude of Snowfall and Precipitation increase, the variance of the residuals falls. This issue could imply that the OLS estimator is not the lowest variance estimator for our model. Another possible issue is the existence of colinearity in certain months. A look at Figures (3 - 4) shows the correlation matrix for all the variables. We see that none of the variables are highly correlated overall but, snowfall and precipitation have high correlation in the winter months. I do not see this as a large issue since we are looking at the entire year.

There are also some issues the prevent us from establishing a true causal relationship. One is the existence of outside shocks that affect trading volume. For example, the stock market took a large hit in early 2020 due to COVID-19. As shown in (Figure 5) the Volume of SPY being traded was at its highest during that time. Fortunately, these external shocks are captured by our error term. There is a chance of a large storm affecting companies represented by the SPY which is also captured by the error term. In this case since weather is a good predictor of storms, the error is correlated with our independent variables which could pose another issue.

4 Results

In New York City, a one mph increase in wind speed meant there was an average 330220.038 increasing in SPY trading volume, holding all other variables constant. A one inch increase in precipitation meant there was an average 7586109.017 increase in SPY trading volume, holding all other variables constant. A one inch increase in snowfall was associated with an average

2181069.593 decrease in volume, holding all other variables constant. Finally, a one inch increase in temperature meant there was an average 768077.182 decrease in volume, holding all other variables constant. The constant term which is the trading volume at 0 degrees, no wind, no snow, and no precipitation was 128704534.498. The highest R^2 value achieved was 0.064.

For Chicago, a one mph increase in wind speed meant there was an average 376782.102 increasing in SPY trading volume, holding all other variables constant. A one inch increase in precipitation meant there was an average 70825.921 increase in SPY trading volume, holding all other variables constant. A one inch increase in snowfall was associated with an average 6219621.294 decrease in volume, holding all other variables constant. Finally, a one inch increase in temperature meant there was an average 607599.195 decrease in volume, holding all other variables constant. The constant term which is the trading volume at 0 degrees, no wind, no snow, and no precipitation was 118057322.606. In the case of Chicago weather data and SPY volume, the highest R^2 was only 0.049.

In terms of statistical significance, the only variable that was statistically significant in both regressions was the TAVG variable which is the average temperature on a given day. At $\alpha = 0.05$ no other variables were significant. Another interesting find related to statistical significance is the Snowfall is more significant in Chicago than New York City. In Chicago, the snowfall variable had a p-value of 0.201 while snowfall in New York City had a p-value of 0.554. Looking at Figures (6 - 7) show that this disparity might be explained by snowfall in Chicago happening more frequently. On the other hand, the precipitation variable had a p-value of 0.211 in New York and 0.991 in Chicago. This can be attributed to the more frequent rain in New York City. in both cases Wind speed was not a good predictor of Volume with p-values of 0.523 and 0.535 for New York City and Chicago respectively.

Along with interpreting the full model for both sets of regressions, there were some interesting outcomes caused by restricting the model. When comparing the New York City model(3) to model(4) we see that the omission of the precipitation variable has a positive bias. This can be seen from the results in (Table 1) and correlation matrix in (Figure 3). For Chicago, when comparing model(3) to model(4) there is also a positive bias caused by omitting the snowfall variable. This can be interpreted from the regression results *Table 2) and correlation matrix(Figure 4). Another interesting outcome of removing a variable occurred in the Chicago regression when comparing model(1) to model(2). The removal to precipitation did not change any of the other coefficients by a large amount. This indicates that precipitation was not at all significant in predicting SPY volume when given Chicago weather data. In the New York City model, when removing Snowfall from model(1) the coefficients for wind speed and precipitation fell. This makes sense given that the coefficient of snowfall was negative and the snowfall was positively correlated with wind speed and precipitation.

As mentioned in the methods section, there is a possible existence of heteroskedasticity. Upon running the White test for both full models, there is indeed heteroskedasticity in both models. The New York City model had a p-value of 0.02107 which is below that 0.05 threshold. The Chicago model had a p-value of 0.00714 which also below our threshold. Another problem found is the that distribution of the residuals is not normal as shown in Figures (8-9). This is further confirmed by the Anderson-Darling test in which the test statistic had a p-value of 0 in both cases. This implies non normal error terms.

5 Conclusion

From these results we can conclude that these wind, snowfall, precipitation, and temperature alone are not good predictors of the SPY trading volume. There are some issues with the choice of variables, heteroskedasticity, and normality that significantly affect our prediction. Despite these poor results, we were able to find that the Average temperature was the only statistically significant variable. This allows us to conclude that there is likely not a mood effect caused by weather. This conclusion makes sense due to the increase in high frequency trading and algorithmic trading. As more trading is done by computers, it makes sense that there is not a mood effect at play.

6 Tables

Table 1: SPY Volume and New York City Weather

	Dependent variable: Volume			
	Full model	Omit Snowfall	Omit Snowfall and Wind	Only Temperature
	(1)	(2)	(3)	(4)
AWND	330220.038	303173.855		
	(516231.787)	(513983.864)		
PRCP	7586109.017	7026208.417	7557839.150	
	(6055953.787)	(5979088.428)	(5908189.719)	
SNOW	-2181069.593			
	(3683166.789)			
TAVG	-768077.182***	-758354.538***	-780129.334***	-768649.125***
	(117668.287)	(116466.536)	(110413.059)	(110094.473)
const	128704534.498***	128391881.614***	132738806.250***	133185846.044***
	(9839012.426)	(9820565.156)	(6488203.350)	(6481536.942)
Observations	751	751	751	751
R^2	0.064	0.064	0.063	0.061
Adjusted R^2	0.059	0.060	0.061	0.060
Residual Std. Error	51335333.764(df = 746)	51313017.221(df = 747)	51290646.046(df = 748)	51312430.992(df = 749)
F Statistic	12.758*** (df = 4.0; 746.0)	16.909*** (df = 3.0; 747.0)	25.211*** (df = 2.0; 748.0)	48.744*** (df = 1.0; 749.0)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: SPY Volume and Chicago Weather

	Dependent variable: Volume			
	Full model	Omit Precipitation	Omit Precipitation and Wind	Only Temperature
	(1)	(2)	(3)	(4)
AWND	376782.102	377483.707		
	(607394.532)	(604006.186)		
PRCP	70825.921			
	(6021955.482)			
SNOW	-6219621.294	-6214867.028	-5919753.752	
	(4854366.123)	(4833890.094)	(4808582.295)	
TAVG	-607599.195***	-607405.003***	-616524.780***	-582909.172***
	(108155.663)	(106808.029)	(105758.373)	(102211.787)
const	118057322.606***	118048912.430***	122032874.805***	119743934.465***
	(8896922.232)	(8861509.893)	(6152669.828)	(5867265.513)
Observations	673	673	673	673
R^2	0.049	0.049	0.048	0.046
Adjusted R^2	0.043	0.045	0.046	0.045
Residual Std. Error	51374546.076(df = 668)	51336140.510(df = 669)	51312788.108(df = 670)	51332497.310(df = 671)
F Statistic	8.593*** (df = 4.0; 668.0)	11.475*** (df = 3.0; 669.0)	17.032*** (df = 2.0; 670.0)	32.524*** (df = 1.0; 671.0)

Note: *p<0.1; **p<0.05; ***p<0.01

7 Figures

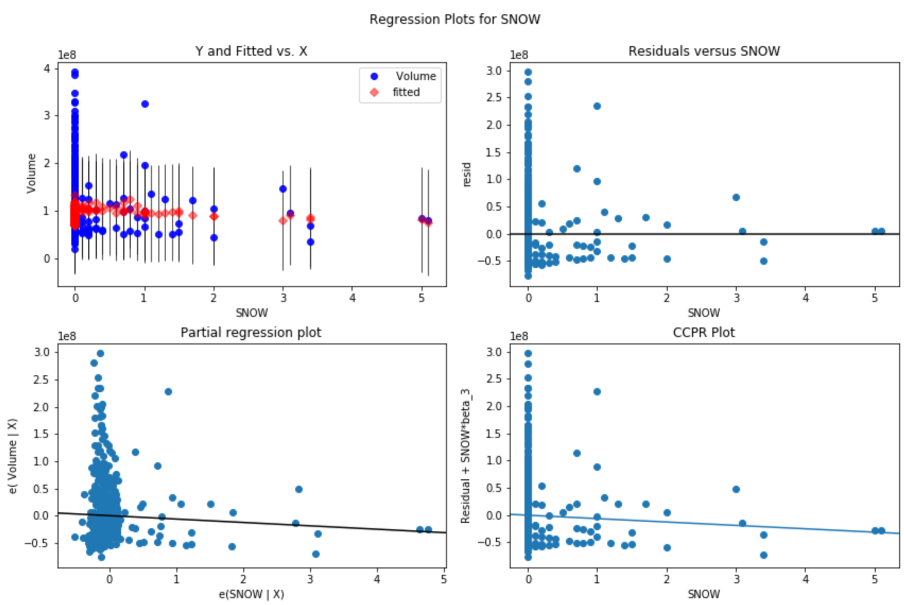

Figure 1: New York City Snow variable diagnostics

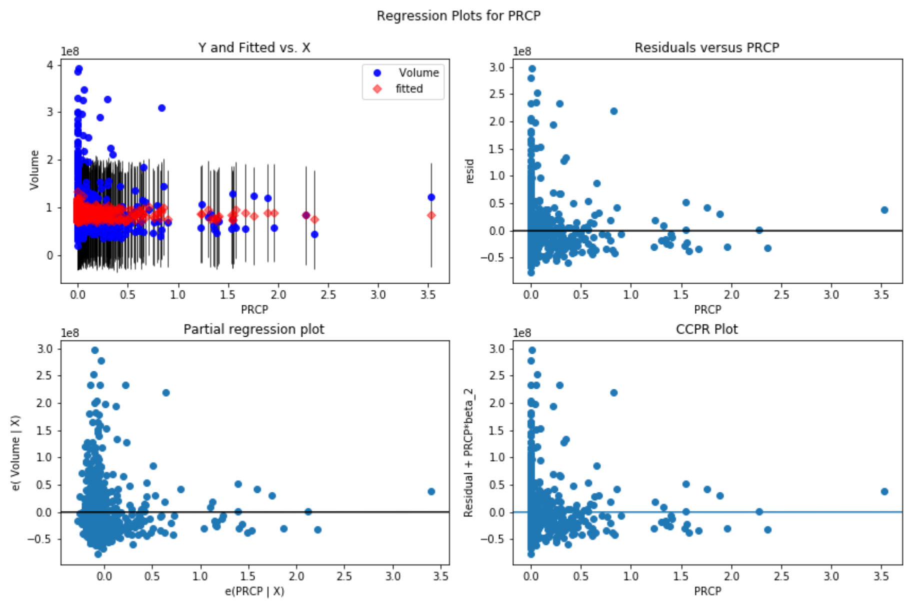

Figure 2: New York City precipitation variable diagnostics

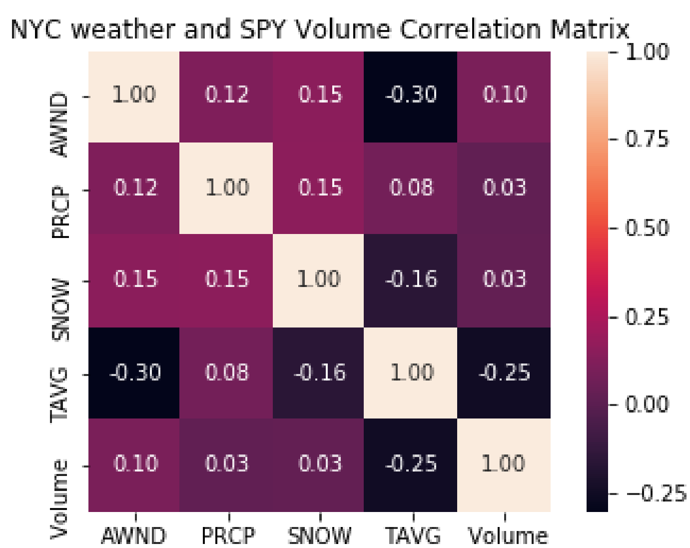

Figure 3: New York City Correlation Matrix

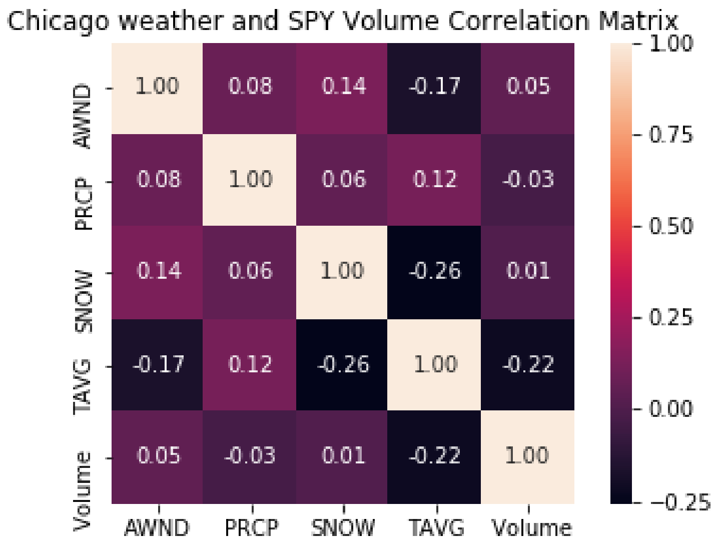

Figure 4: Chicago Correlation Matrix

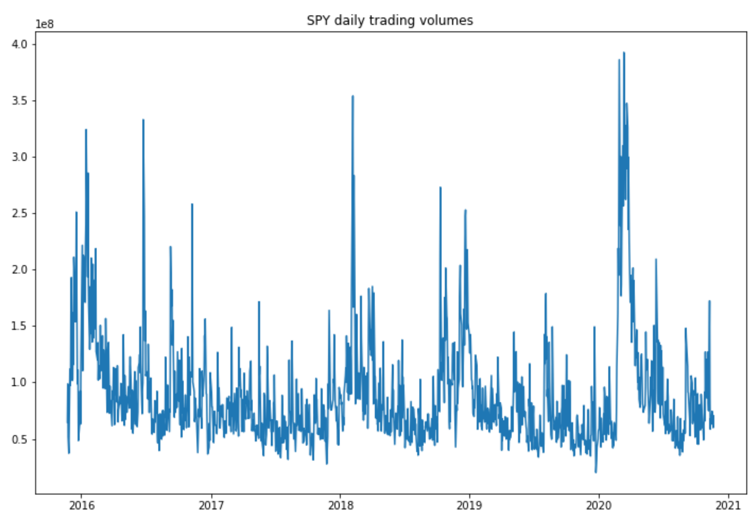

Figure 5: SPY Daily Trading Volumes

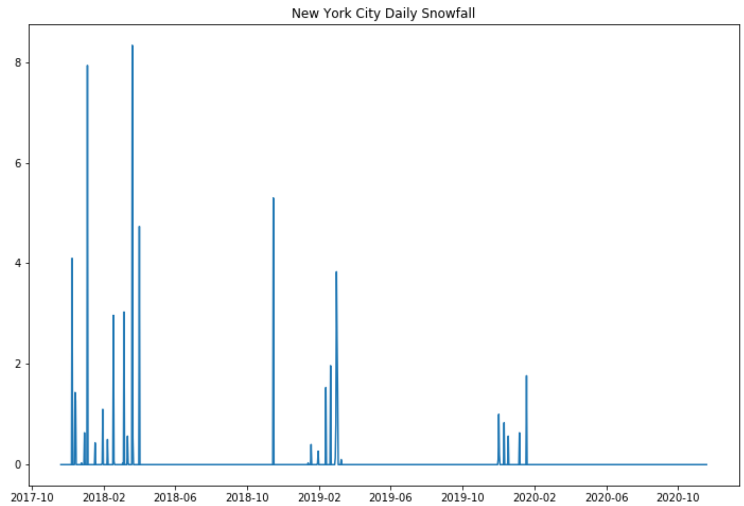

Figure 6: New York City Snowfall

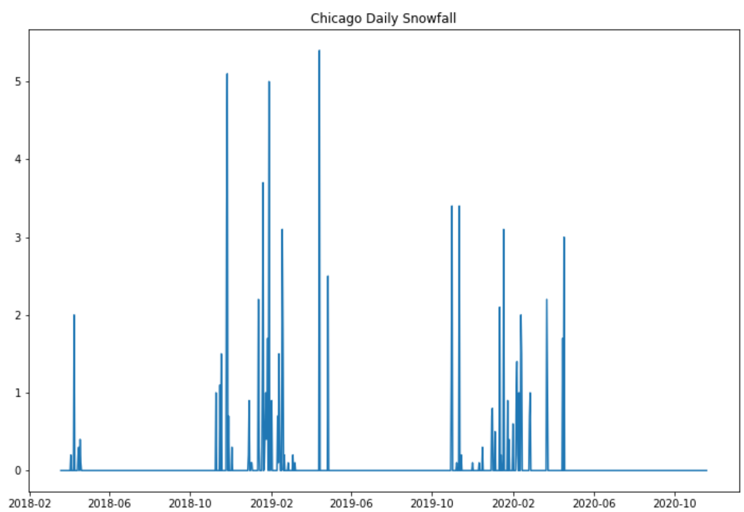

Figure 7: Chicago Snowfall

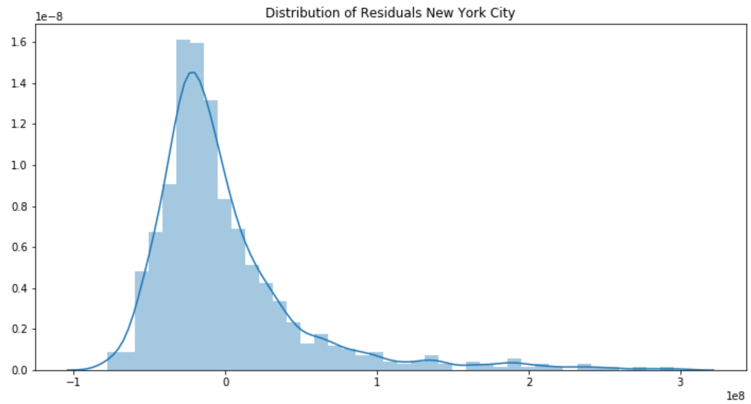

Figure 8: New York City Residual Distribution

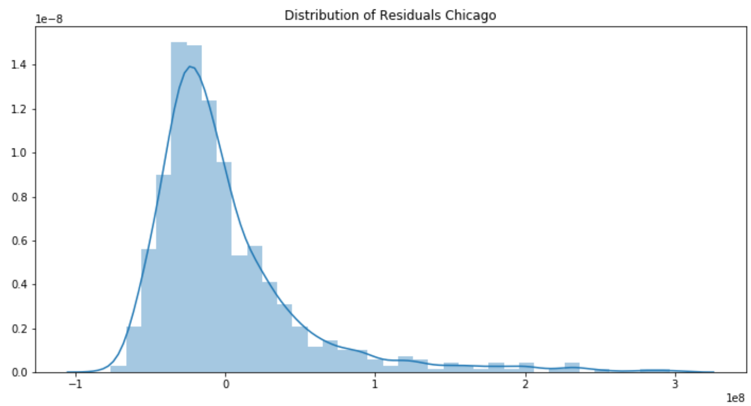

Figure 9: Chicago Residual Distribution

8 References

1. https://www.ncdc.noaa.gov/cdo-web/

2. https://www.nasdaq.com/market-activity/funds-and-etfs/spy/historical