Volume - Return Correlation in Finance Assets

May 15, 2023

Andrea Bernini
Donato Francesco Pio Stanco

Abstract

The aim of this work is to investigate the potential correlation between financial stock *Volume* and *Return*. To establish a connection, various techniques such as the DTW algorithm, the DCCA method, and OLS regression models have been employed. To perform the analysis, we looked at daily data from Lobsters and Yahoo Finance. Our findings reveal a correlation between the *Volume* and *Return* prices of various financial indicators. Specifically, we observed that the correlation was mostly positive until the 1990s, after which it turned negative.

1. Introduction

The rise of technology has revolutionized the financial industry, with online trading becoming an increasingly popular way to invest in stocks, commodities, and other financial assets. Online trading has democratized access to financial markets, enabling retail investors to buy and sell securities from the comfort of their own homes. The availability of real-time market data and trading platforms has made it easier for individuals to make informed investment decisions and actively manage their portfolios.

In online trading, the volume of trades and returns are two important indicators that investors use to make investment decisions. Volume refers to the number of shares or contracts traded in a particular asset over a given period of time. High trading volume can indicate increased investor interest in a particular asset, which can lead to higher liquidity and price movements. Return, on the other hand, refers to the profit or loss an investor earns on their investment. High returns are the ultimate goal of any investment, but they also come with higher risks.

An important aspect is that investors need to consider the relationship between volume and return in order to effectively manage their investment portfolios.

Email: Andrea Bernini <bernini.2021867@studenti.uniroma1.it>, Donato Francesco Pio Stanco <stanco.2027523@studenti.uniroma1.it>.

Subsidiary Formative Activity in AI4Trading, Sapienza University of Rome, a.y. 2022/2023.

2. Related works

DCCA A related work is described in this paper (Rodriguez & Alvarez-Ramirez, 2021) where the authors used cross-correlation analysis (DCCA) to study contemporaneous co-movements between trading volume and returns in US stock markets (Dow Jones, NASDAQ, and Standard & Poor's 500).

OLS Another related work is the one described in this paper (Llorente et al., 2002) where the authors examined the dynamic relations between return and volume of individual stock by analyzing the relation between daily volume and first-order return auto-correlation for individual stocks listed on the *NYSE* and *AMEX* by performing OLS regression.

3. Methods

In this section, we discuss all the study cases that we approached in order to verify this kind of relationship.

3.1. Code

The code of the project is available in the following GitHub repository AI4Trading-Project.

3.2. DTW Algorithm

We decided to use the **Dynamic Time Warping (DTW)** Algorithm, widely used for the analysis of time series, to establish a possible correlation between the *Close Price* and the *Volume*. Specifically, we analyzed the stocks of leading Tech companies such as Google, Apple, Microsoft, using the datasets provided by Lobester, which have the format *Message/Orderbook* dated June 21, 2012, so the first step was to transform the datasets of each stock action in **OHLC** format (Open, High, Low, Close), using a unit time of one minute (to simulate days), so as to be able to obtain the *Close Price* and the *Volume* for each unit time.

The DTW algorithm returns a matrix where each element of the two series is represented on the `x`-axis and `y`-axis respectively, and each cell represents the distance between the two values. The objective is to determine the **minimum path** from the bottom left cell to the top right cell, and the more this line looks like a *straight line*, the more there is a correlation between the two time series.

To use the DTW algorithm, we made use of the `dtw-python` Python library, which returns both the **warping curve**, i.e., the alignment curve, the **distance matrix** , and the related shortest path.

However, the *Volume* values are approximately two orders of magnitude higher than the *Close Price* values, which causes the algorithm to not function correctly (we can see an example with Apple shares in Figure 1b) since the *Close Price* is seen as a flat line compared to the *Volume* (Figure 1a).

To overcome this issue, we decided to **normalize** the values of the two time series, so that they are both on the same scale (Figure 2a).

3.3. DCCA Algorithm

Initially, as an alternative method to DTW, we had thought of using **Pearson's correlation coefficient**, but after doing some research, we found that it is not robust and can be misleading in the presence of outliers, which is common in real-world data that are highly non-stationary (Wilcox, 2005). Therefore, following the work done by Rodriguez & Alvarez-Ramirez (2021), we decided to analyze the correlation between *Volume* and *Return*, using the DCCA method.

The **Detrended Cross Correlation Analysis (DCCA)** (Zebende, 2011) is a generalization of the Detrended Fluctuation Analysis (DFA) method. It was proposed to detect long-range auto-correlations embedded in a patch landscape and also avoid spurious detection of apparent long-range auto-correlations. The DCCA returns an index ρ, which varies in the $[-1, +1]$ domain, with

- $\rho = 0$ for the uncorrelated time series;
- $\rho > 0$ for time series that are positively correlated, i.e., an increase in one time series is statistically followed by an increase in the other time series, and vice versa;
- $\rho < 0$ for time series that are negatively correlated, i.e., an increase in one time series is statistically followed by a decrease in the other time series, and vice versa.

We tried to replicate the results of the paper by analyzing the same three U.S. indices using data sets provided by Yahoo Finance, considering the period from 1955 to 2021 for the *Standard & Poor 500* index, from 1990 to 2021 for the *Nasdaq*, and from 1995 to 2021 for the *Dow Jones*.

However, the downloaded datasets appear to have different dimensions from those expected, specifically, for the *Dow Jones* and *Nasdaq* datasets (Table 1), we had less than half the number of samples. For the *Dow Jones*, this is explained by the fact that the *Volume* appears to have been recorded only starting from 1995 (Figure 7b), while for *Nasdaq* the starting dates for recording the *Volume* and the OHLC coincide (Figure 8b), but as mentioned, we have fewer samples, this may be due to an index name change on Yahoo Finance. Finally, for *Standard & Poor*'s, the datasets coincide both with the *Volume* registration start dates and with the number of samples.

	Dow Jones	Nasdaq	S&P 500
# Samples Rodriguez & Alvarez-Ramirez Dataset	23 943	17 588	17 959
# Samples our Dataset	7 368	9 193	17 928

Table 1. Comparison of the number of samples between the datasets obtained by us and those of Rodriguez & Alvarez-Ramirez (2021).

The DCCA analysis was implemented on a moving window of 1056 observations (about 4 calendar years) and a sliding window of 25 observations.

3.4. OLS Regression

Based on the experiment described in this article (Llorente et al., 2002), we decided to find a dynamic relation between *Volume* and *Return*. More specifically we performed the **Ordinary Least Squares (OLS) regression** using *Volume* and the *Daily Return*, which is computed by using the close price and indicates the gain or loss per day for a given stock. In particular, OLS regression is a method that allows us to find a line that best describes the relationship between one or more predictor variables and a response variable by following this equation:

$$y = b_0 + b_1 x$$

where:

- y: the estimated response value
- b_0: the intercept of the regression line
- $b_1 x$: the slope of the regression line

This equation can help us understand the relationship between the predictor and response variables, and it can be used to predict the value of a response variable given the values of the predictor variable.

To test this method, we downloaded from Yahoo Finance the datasets of **AXP** (American Express), **NYA** (NYSE), **RUT** (Russell 2000), with a time range of 5 years (2018-2023).

We tested a different range of years in our experiment than authors Llorente et al., who analyzed the phenomenon between 1982 and 1987. In fact, Yahoo Finance does not provide volume data for this period, so we decided to use modern data to test the existence of the relationship. For the results obtained with this approach, see Section 4.3.

4. Experimental Results

4.1. DTW Algorithm

After normalizing the *Volume* and *Close Price* values, as described in Section 3.2, we achieved a better result (Figure 2b), but there does not seem to be a strong correlation since, as we can see from Figure 2b, the DTW returns a matrix in which the shortest path takes the form of an **increasing exponential functions curve** which approaches the ideal straight line towards the last values. So the algorithm suggests that there is no correlation in the first two-thirds, while in the last part we have a higher but still weaker correlation.

To explore the problem further, we decided to use the Yahoo Finance datasets, which among other features directly provide the daily *Volume* and *Close Price*, and thus more samples than the Lobster datasets that refer to a single day.

However, we did not find a drastic change in the results compared to those obtained previously, so we decided to analyze the correlation between *Volume* and *Return* (instead of using the simple *Return*, whose return values can range from minus infinity to plus infinity, we used the **log Return**, since its return values are symmetric about 0), again using the DTW algorithm, and then normalizing the two time series. The output of the algorithm turns out to be a matrix in which the minimum path turns out to be very close to the ideal path, i.e., a straight line, so the DTW would suggest a slight correlation between *Volume* and *Log Return* of several stocks including Apple, CocaCola, Volkswagen (Figures 3, 4, 5).

One problem related to DTW is that it does not return an **actual metric** to accurately assess whether a correlation exists and whether it is positive or negative, so we decided to explore other methodologies to answer our question of whether a correlation exists between *Volume* and *Return*.

4.2. DCCA Algorithm

To replicate the work done by Rodriguez & Alvarez-Ramirez (2021) and use the DCCA algorithm for the correlation analysis, we utilized the Python library `fathon`, which allows us to calculate the ρ value of DCCA using a rolling window. However, as introduced earlier, due to the different shapes of the datasets, our results are slightly different from those obtained by Rodriguez & Alvarez-Ramirez (2021), due to the time windows examined or the number of data points.

Dow Jones For the Dow Jones index, we have a very narrow time window (1995 to 2021), compared to that of Rodriguez & Alvarez-Ramirez's work (1935 to 2021), which shows a mostly slightly negative correlation over the last twenty years, then becoming slightly positive in 2020 (Figure 7c), it is most likely a COVID-19 related event.

Nasdaq For the Nasdaq index, we have the same time window, however, the number of samples turns out to be much lower, but the algorithm returns the same overview, i.e., a slightly positive correlation until about 2008, to then become slightly negative (Figure 8c), this change could be related to the financial crisis of 2008.

Standard & Poor 500 While with regard to the *Standard & Poor 500* index the results are comparable, i.e., the correlation appears to be mostly **positive** before the beginning of the 1990s (much stronger between about 1975 and 1985), to then become weakly **negative** and decrease more and more until 2020 (Figure 6c).

The authors Rodriguez & Alvarez-Ramirez explain that when traders who seek to buy and sell stocks in a hurry collide with those who prefer to buy and sell more thoughtfully, an abnormal increase in the volume of trading in the market can occur. In this context, risk-averse investors dominate market dynamics and may be prompted to sell their shares due to external or internal risk factors, such as large changes in market conditions (for example, the dot-com bubble burst and the 2008 Great Recession).

4.3. OLS Regression

In this section, we'll discuss the results obtained with this model, as mentioned in Section 3.4 we tested three different datasets in order to find a possible correlation. We're going to analyze the outputs obtained with the *AXP*, *NYA*, *RUT* datasets. To perform OLS in Python, we used the `statsmodel api` provided by the language. First, we compute the *daily return* by using the `pct_change` method provided by `Pandas` to calculate the **percentage** of changes in the daily returns. The output produced by the OLS regression is of two types:

- summary tables of all the coefficients produced by the model,
- graph that shows how the line fits the points.

In the following paragraphs, we describe the results obtained with each dataset.

Results with AXP and RUT datasets As we can observe from Figures (9a, 10a), we have a great oscillation around 2020, probably due to some *spike* in the US stock market, for instance the COVID pandemic. In tables (2, 3) we have the results produced by the model, where we can analyze the possible correlation.

	coef	std err	t	$P > \|t\|$	R-squared
const	0.1022	0.147	0.695	0.487	0.000
volume	$-9.414^{e^{-09}}$	$3.43^{e^{-08}}$	-0.275	0.784	0.000

Table 2. OLS Regression Results with AXP

	coef	std err	t	$P > \|t\|$	R-squared
const	0.3702	0.196	1.889	0.059	0.003
volume	$-8.13^{e^{-11}}$	$4.45^{e^{-11}}$	-1.827	0.068	0.003

Table 3. OLS Regression Results with RUT

From the **coef** columns in tables (2, 3) we can see the regression coefficients and can write the following fitting regression equations:

$$Score = 0.1022 + (-9.414^{e^{-09}}) * (Volume) \quad (1)$$

$$Score = 0.3702 + (-8.13^{e^{-11}}) * (Volume) \quad (2)$$

Now let's analyze this equations:

- the equation (1) describes the fitting equation of AXP and tells us that each additional volume is associated with an average decrease of -9.414 points and the intercept value of 0.102 tells us the average expected daily return for volume
- the equation (2) describes the fitting equation of RUT tells us that each additional volume is associated with an average decrease of -8.13 points and the intercept value of 0.3702 tells us the average expected daily return for the volume

Here is how to interpret the rest of the model summary:

- $P(> |t|)$: this is the p-value associated with the model coefficients. In the case of **AXP** this is (0.784) and for **RUT** is equal to (0.068). In both cases this is greater than .05, so we can say that there is **no statistically** significant association between *Volume* and *Daily Return*.
- $R - squared$: this tells us the percentage of the variation in the daily return that can be explained by the volume. In the case of **AXP** we have 0% and for **RUT** is 0.3% so we don't have **significant variation**.

As we can observe from the Figures (9b, 10b), we can notice that the line doesn't fit all the points but only a part of them. Indeed, this leads to the fact that the model does not "struggle" to catch the relation, and also because the **p-value** is just a probability, we cannot assert the relation.

Results with NYA dataset As we can observe from Figure 11a, we have a great oscillation around 2020, probably due to some *spike* in the US stock market, for instance the COVID pandemic. In table 4 we have the results produced the model where we can analyze the possible correlation.

	coef	std err	t	$P > \|t\|$	R-squared
const	0.4394	0.148	2.973	0.003	0.07
volume	$-9.707^{e^{-11}}$	$3.36^{e^{-11}}$	-2.891	0.004	0.07

Table 4. OLS Regression Results with NYA

From the **coef** column in Table 4 we can see the *regression coefficients* and can write the following fitting regression equation:

$$Score = 0.4394 + (-9.707^{e^{-11}}) * Volume$$

This means that each additional volume is associated with an average decrease of -9.707 points.

The intercept value of 0.4394 tell us the average expected daily return for the volume.

Here is how to interpret the rest of the model summary:

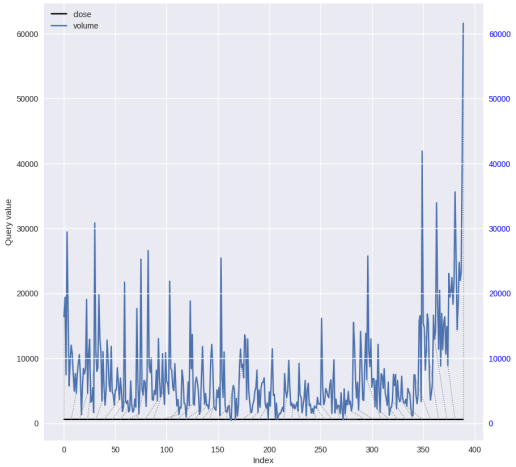
- $P(> |t|)$: this is the $p - value$ associated with the model coefficients. Since the p-value for volume is (0.004) less than .05, we can say that there **is a statistically** significant association between *volume* and *daily return*
- $R - squared$: this tells us the *percentage of the variation* in the *daily Return* can be explained by the *Volume*. In this case we have 7% so we have **significant variation**.

As we can observe from Figure 11b, we can notice that the line doesn't perfectly fit all the points but only a part of them. Indeed, this leads to the fact that the model does not "struggle" to catch the relation, but by the **p-value** which is just a probability, we can assert the relation.

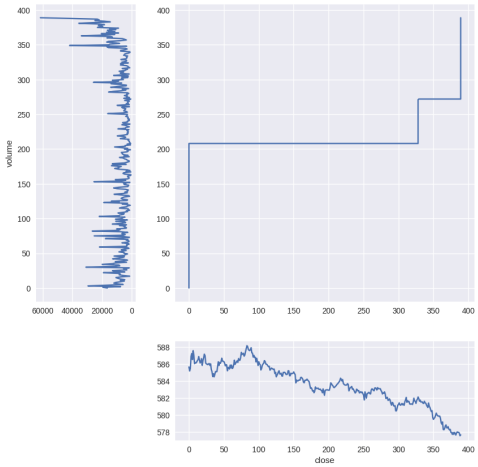
5. Discussion and Conclusion

Based on the analysis conducted using different methods, including DTW, DCCA and OLS Regression, and combining several works already done on the subject, we have come to the conclusion that there is a correlation between *Volume* and the *Return* price of various financial ratios. In particular, we observed that the correlation was largely positive until the 1990s, but then turned negative.

However, it is important to note that the correlation between these two variables **remains relatively weak**, so it does not seem to have such a weight as to be able to influence any investment decisions by investors and financial operators.



(a) Close Price and Volume



(b) DTW Algorithm

Figure 1. Apple shares not normalized from Lobster.

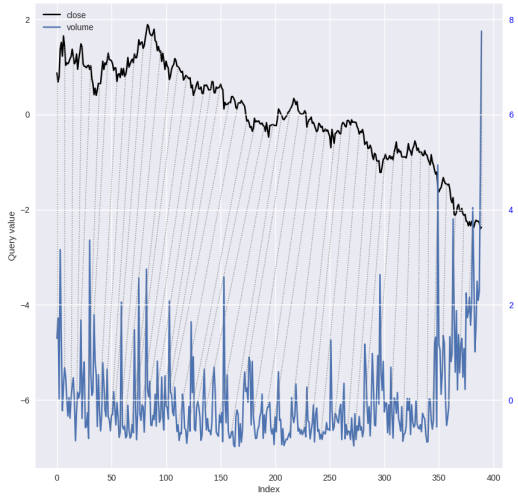
References

Llorente, G., Michaely, R., Saar, G., and Wang, J. Dynamic volume-return relation of individual stocks. *The Review of Financial Studies*, 15(4):1005–1047, 2002. ISSN 08939454, 14657368. URL http://www.jstor.org/stable/1262690.

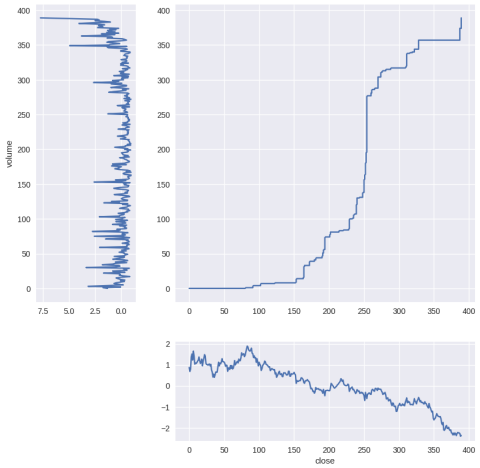
Rodriguez, E. and Alvarez-Ramirez, J. Time-varying cross-correlation between trading volume and returns in us stock markets. *Physica A: Statistical Mechanics and its Applications*, 581:126211, 2021. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2021.126211. URL https://www.sciencedirect.com/science/article/pii/S0378437121004842.

Wilcox, R. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego, CA, 2 edition, 2005.

Zebende, G. Dcca cross-correlation coefficient: Quantifying level of cross-correlation. *Physica A: Statistical Mechanics and its Applications*, 390(4):614–618, 2011. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2010.10.022. URL https://www.sciencedirect.com/science/article/pii/S0378437110008800.

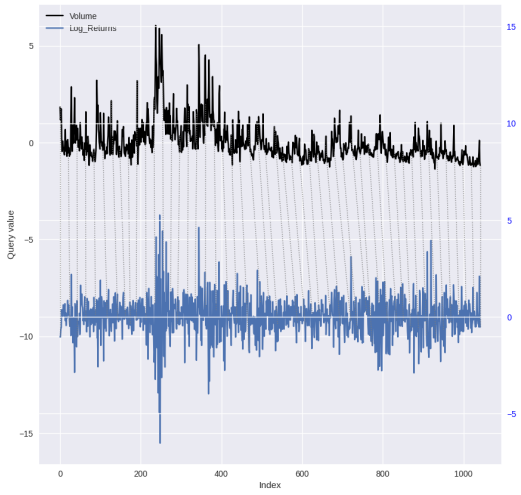


(a) Close Price and Volume Normalized

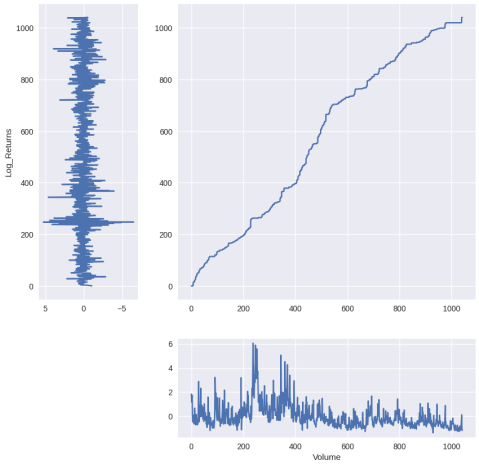


(b) DTW Algorithm

Figure 2. Apple shares normalized from Lobster.

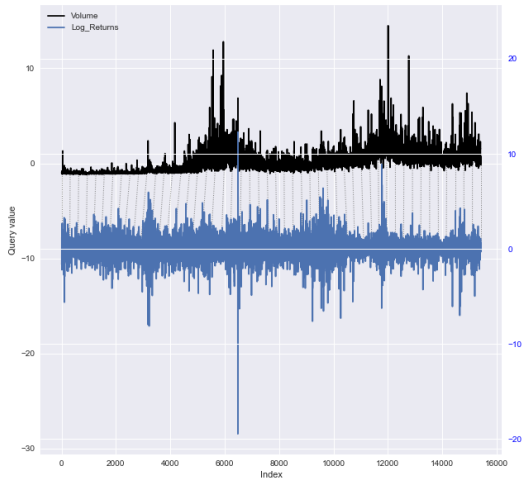


(a) Volume and Log Return Normalized

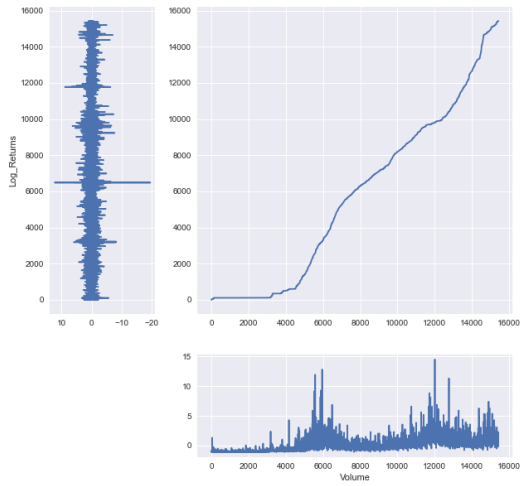


(b) DTW Algorithm

Figure 3. Apple shares from Yahoo Finance.

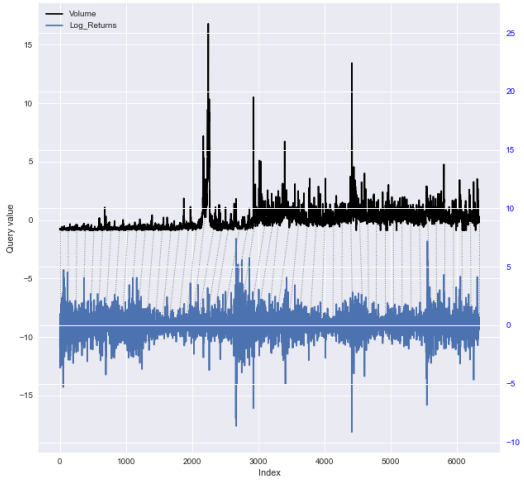


(a) Volume and Log Return Normalized

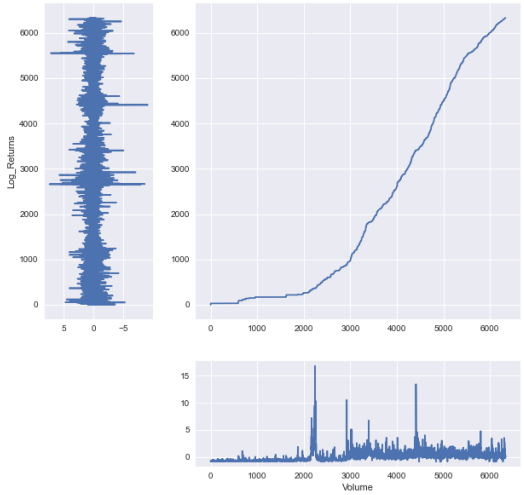


(b) DTW Algorithm

Figure 4. Coca Cola shares from Yahoo Finance.

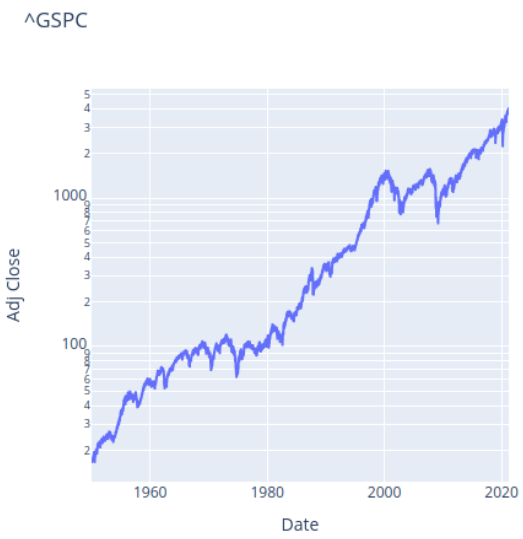


(a) Volume and Log Return Normalized

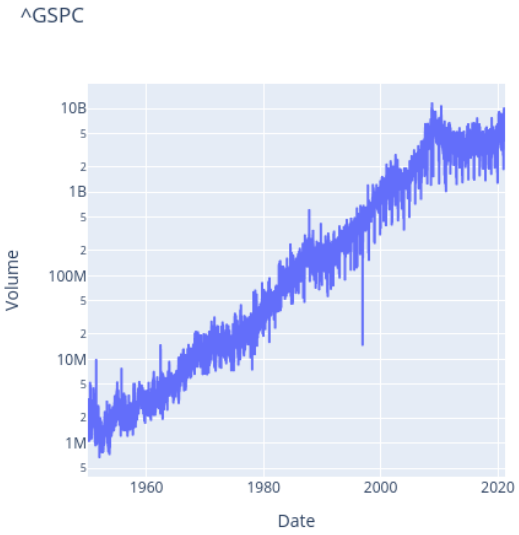


(b) DTW Algorithm

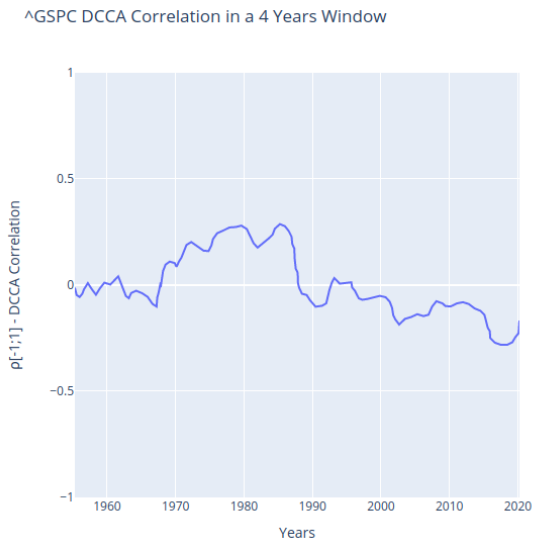
Figure 5. Volkswagen shares from Yahoo Finance.



(a) Close Price



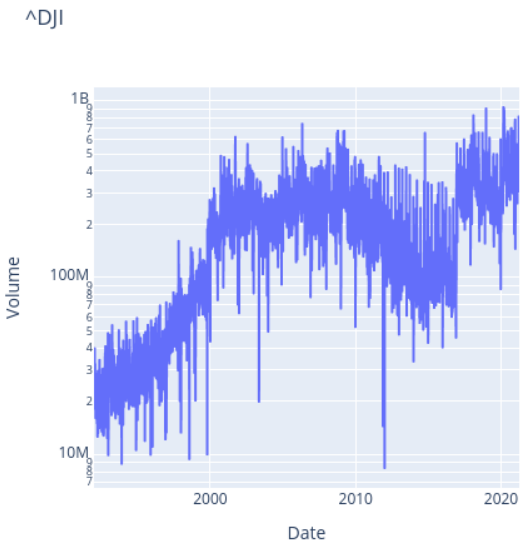
(b) Volume



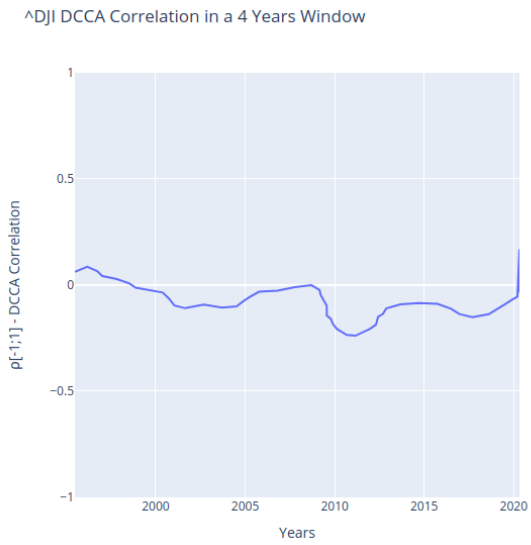
(c) DCCA

Figure 6. Standard & Poor 500 index.

(a) Close Price

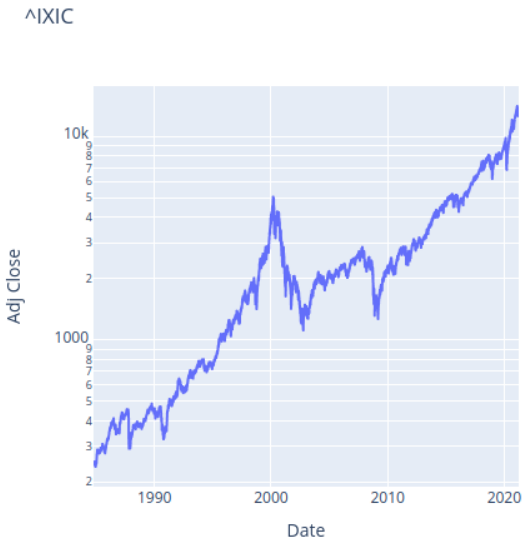


(b) Volume

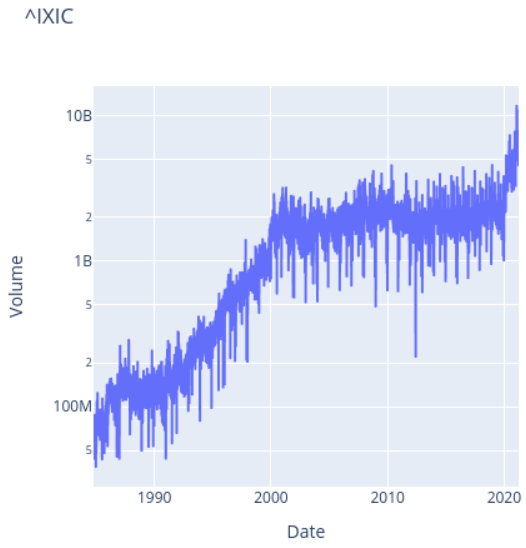


(c) DCCA

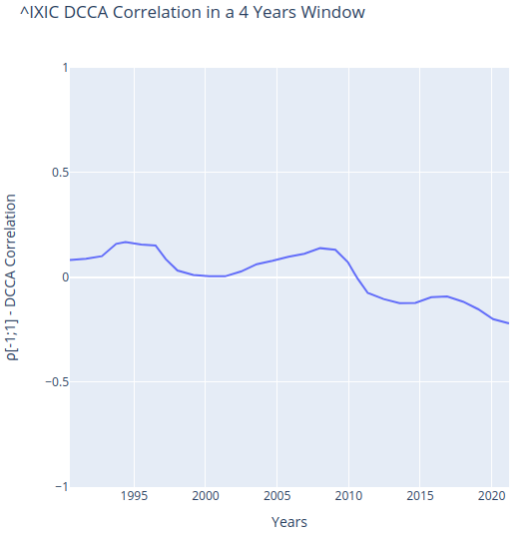
Figure 7. Dow Jones index.



(a) Close Price

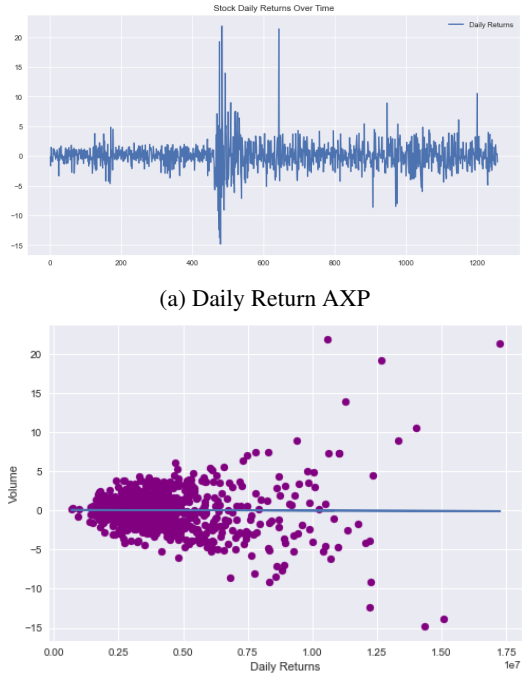


(b) Volume



(c) DCCA

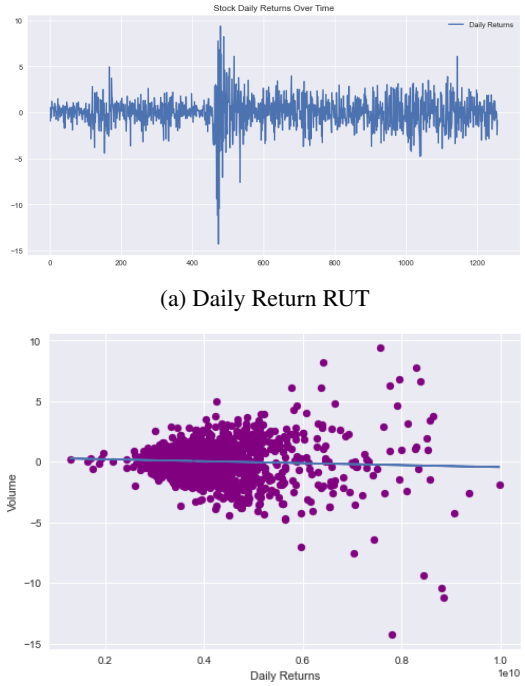
Figure 8. Nasdaq index.



(a) Daily Return AXP

(b) Points fitted with OLS line

Figure 9. OLS model results AXP (American Express)

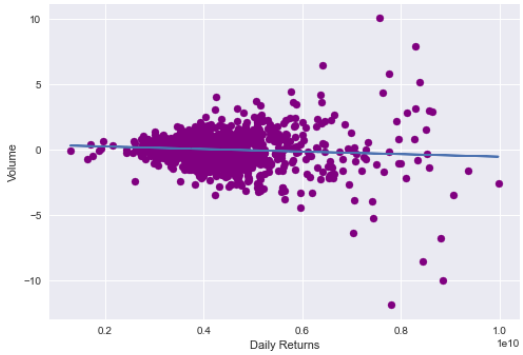


(a) Daily Return RUT

(b) Points fitted with OLS line

Figure 10. OLS model results RUT (Russell 2000)

(a) Daily Return NYA



(b) Points fitted with OLS line

Figure 11. OLS model results NYA