Quantitative Research Coding Project

Stock Volatility and Liquidity Analysis

Kelvin Brinham

GitHub Repository Link

Date: February 19, 2023

Contents

1 Objective

In summary, this project investigates the relationship between liquidity and volatility of four fictional stocks using intraday (volume and price) trading data.

2 Data Cleaning

2.1 Data Cleaning

The data provided for each of the four stocks is a list of trades which detail the time, price and volume traded. Given the frequency of data provided, it is unsurprising that errors are present. We remove these errors according to the following procedure:

1. Remove duplicate time entries (keeping the first occurrence)
2. Remove invalid/missing data (e.g. non-numeric data or negative prices/volumes)
3. Remove data outside market hours (08:00 - 16:30 for stocks A and B; 08:00 - 16:00 for stocks C and D)
4. Remove outlier prices

The statistics relating to how much data was removed are outlined in table 1. We chose to remove invalid/missing data rather than replace it because the data set is large. Figure 1 displays the difference after data cleaning.

Stock	Unclean Size	Repeated Entries	Rows with Missing/Invalid Data	Outliers
A	182288	0	51	121
B	92238	0	91	139
C	115998	0	104	129
D	69588	0	188	177

Table 1: Summary of data removed during the cleaning process.

2.2 Removing Outliers

Outliers were removed by considering a single price which was significantly different to its two neighbouring points which are similar. For example, a price series of $10, 200, 11$ would consider 200 an outlier because it is significantly different from 10 and 11 while 10 and 11 are similar. Further, $10, 200, 198$ would not be edited and assumed a sudden change due to market news or some other factor.

In practice, this was achieved by taking the second difference of the price series. For example, a price series of $9, 10, 200, 8, 10$ has a second difference of $NaN, NaN, 189, -382, 194$. We see this leaves one large value indicating a single point far from its two neighbours. We remove the prices associated with these points using a defined cutoff value (382 above) for each stock by

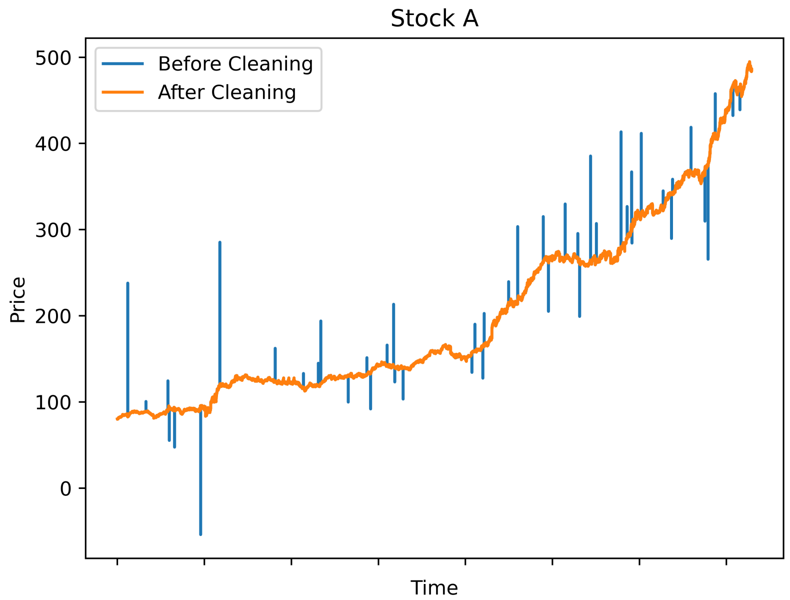


Figure 1: Stock A price before and after data cleaning.

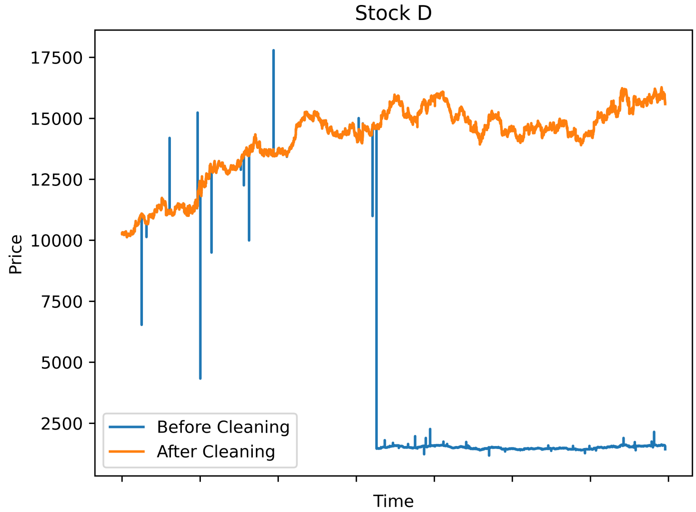
Stock	Cutoff Value
A	20
B	100
C	0.9
D	100

Table 2: Summary of cutoff values for removing outliers

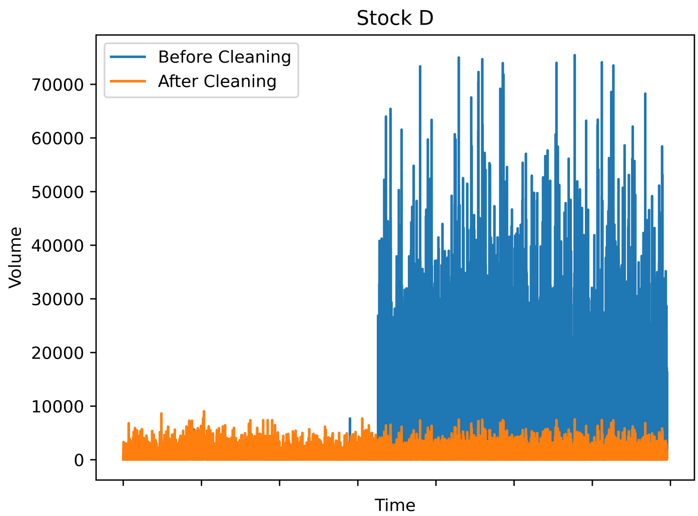
inspection; these are outlined in Table 2. We note that this twice difference method is unable to account for the first and last price points. Further, the first and last points were removed if they were half the cutoff value different to the second and second last points respectively.

2.3 Stock D stock split

Inspecting the raw price and volume data (see Figure 2) for stock D we can see a large drop in price and a similar rise in volume indicative of a stock split. We choose to remove this by multiplying/deviding the price/volume data after the split by the 'split factor'. The 'split factor' is the factor by which the stock price drops from the last/first trade before/after the split. We find the split factor to be 10.015.



(a) Price



(b) Volume

Figure 2: Stock D price (a) and volume (b) time series before and after cleaning and adjusting for the stock split.

3 Data Processing and Calculating Variables of Interest

The raw data is high frequency and thus exhibits significant microstructure noise. In addition, the time series' are asynchronous. Therefore, we transform the data into 5-minute intervals. Price is taken to be the price immediately before the timestamp, except for the 8:00 value which takes the first price value for each trading day. This method is known as the previous tick method which is preferable to linear interpolation in high-frequency economic time series because linear interpolation can introduce significant bias [1][3][4]. Volume is taken to be the summed volume for each 5-minute period.

3.1 Realised Volatility

Volatility can be described as a measure of stock price variation. Volatility is commonly encoded in the Realised volatility (RV). It is calculated from (log) price returns,

$$r(t) = ln\left(\frac{P(t)}{P(t-1)}\right) \tag{1}$$

where $P(t)$ is the price at the timestamp t and $P(t-1)$ is the price at the previous timestamp (5 minutes earlier).

RV is defined as

$$RV = \sum_{t}^{T} r(t)^2 \tag{2}$$

where we choose T to be 1 trading day. This equates to 101 5-minute intervals for stocks A and B and 95 intervals for stocks C and D for which RV is calculated[1]. Further, *our definition of RV is daily*. Volume is taken to be the sum of all volume data for each trading day. To ensure that 5-minute return calculations did not spread overnight we chunked the data into daily data frames for processing.

We note that there was one instance in which there was no data within a 5-minute interval. In this case, we replaced the return data with the mean from the rest of the day.

3.2 Summary of Data

The final step in processing the data was normalised is using a z-score. Table 3 outlines the cleaned dataset which now contains two daily variables, RV and volume.

[1]There are 95 and 101 5-minute intervals respectively rather than 96 and 102 because we have no 5-minute return for 08:00 (see equation 1).

Stock	Time Period	Size (Days)	5-minute Periods Per Day
A	01/03/2017 - 18/08/2017	121	95
B	01/03/2017 - 18/08/2017	121	95
C	01/03/2017 - 18/08/2017	121	101
D	01/03/2017 - 18/08/2017	121	101

Table 3: Summary of data after cleaning and processing.

4 Relationship between Volatility and Liquidity

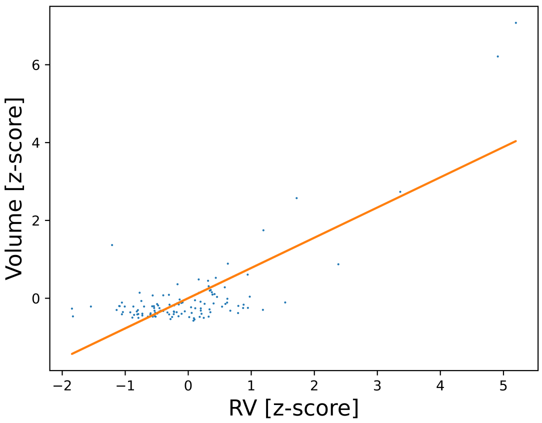
Volume is commonly used as a measure of liquidity in our case daily volume (just volume hereafter) is considered [2][5]. As mentioned above, our chosen measure of volatility is daily RV (just RV hereafter) (see §3.1 for definition).

4.1 Pearson's Correlation

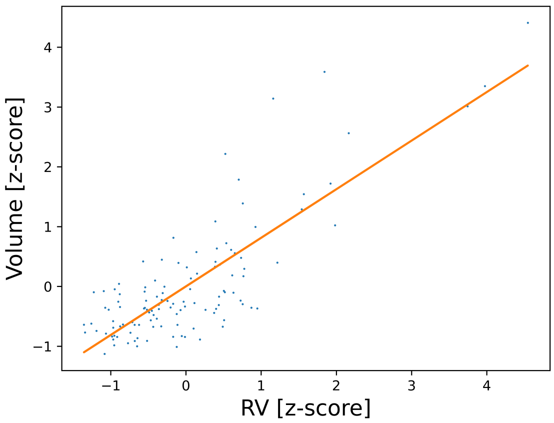
We generate a correlation matrix to examine the basic relationship between RV and volume. We see from Table 4, there is a significant positive correlation between volume and RV for all the stocks except stock C for which there is a weak positive relation. This agrees with the majority of studies which have conclusions summarised in [6]. Figure 3 displays the final data.

Stock	volume
A: RV	0.77639
B: RV	0.81249
C: RV	0.15618
D: RV	0.57806

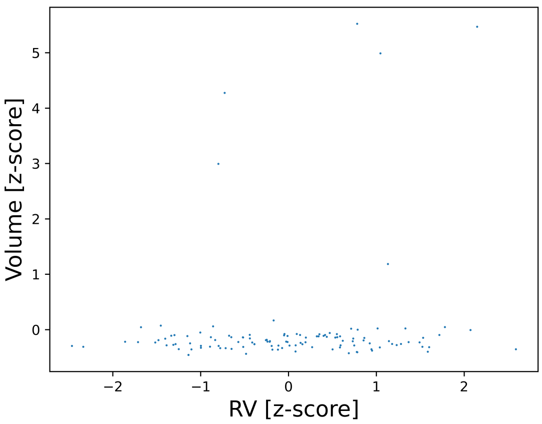
Table 4: Pearson's correlation matrix for each stock using 5-minute returns to calculate RV.



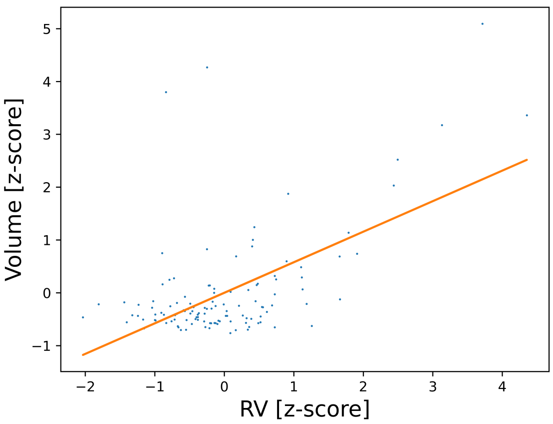
(a) Stock A



(b) Stock B



(c) Stock C



(d) Stock D

Figure 3: Final relationship between volume and RV for each stock. The line of best fit is plotted for each stock except C to display the linear relationship between volume and RV.

References

[1] Emilio Barucci and Roberto Renò. On measuring volatility of diffusion processes with high frequency data. *Economics Letters*, 74(3):371–378, 2002. ISSN 0165-1765. doi: https://doi.org/10.1016/S0165-1765(01)00572-9. URL https://www.sciencedirect.com/science/article/pii/S0165176501005729.

[2] Namitha K Cheriyan and Daniel Lazar. Relationship between liquidity, volatility and trading activity: An intraday analysis of indian stock market. *International Journal of Economics and Financial Issues*, 9(1):17, 2019.

[3] Ramazan Gençay, Michel Dacorogna, Ulrich A Muller, Olivier Pictet, and Richard Olsen. *An introduction to high-frequency finance*. Elsevier, 2001.

[4] Taro Kanatani et al. Integrated volatility measuring from unevenly sampled observations. *Economics Bulletin*, 3(36):1–8, 2004.

[5] Huong Le and Andros Gregoriou. How do you capture liquidity? a review of the literature on low-frequency stock liquidity. *Journal of Economic Surveys*, 34(5):1170–1186, 2020.

[6] Shivaram Shrestha. Contemporaneous relationship between trading volume and stock returns volatility: Evidence from nepalese stock market. *PYC Nepal Journal of Management*, 10(1):40–63, 2017.