Nifty 100 Stock Data – Group 1

Marwan Lloyd, Max Schleck, Skylar Shafter, Rachel Studer, James Wan
University of Wisconsin - Madison

Stocks and their movements have always been an interesting statistical question, as there is an immense reward for those who can accurately predict their future price. As this problem is very broad and complex, there are many subfields of quantitative finance with taking their own respective approaches to predicting future stock prices based on historical data.

Another key detail of this complicated question is that the potential size of data is incredibly large. Data on stock prices can be collected in very small increments, down to the second, and so over many years, and with many stocks globally, the potential datasets become exorbitantly large. In turn, this problem is well suited for parallel processing, in that it works with large files that can be subdivided by the stock.

Our dataset for analyzing stocks with clustered computing is a Kaggle dataset - Nifty 100. It is comprised of 100 stocks traded on the Indian stock market, with minute-by-minute observations taken from 2015 to 2021.

Process

First, we use sub/sh files to divide the computing tasks across stock files and run each separately. Each individual job was run individually and simultaneously within the computing cluster, having submitted the necessary arguments and input files by the sub/sh files mentioned earlier. We ran 100 jobs in total, and each job took about 24 minutes, while requiring 3.31 MB memory and 1.14 GB disk space.

The first step, prior to any analysis, was to clean our chosen datasets. The original datasets provided minute-by-minute information on prices, which we reduced to only the day's high, low, average, and volume. From there, we also performed further aggregations to reduce the dataset to a single opening and closing price for each day for each stock and merged the day open dataset and the day close datasets together into a third general day dataset. Next, we created

a new variable that counted the days in which a stock was above the 50-day moving average (MA). Each entry was the number of days in which a stock was above its 50-day MA, with only entries in which the price was above that threshold for more than 2 days consecutively. We used the 50-day MA threshold as it is a finance industry-wide common metric. We then focused on summarizing the durations of these intervals of positive momentum into summary statistics by finding their average and median, which we believe is an important part of extracting insight into how positive momentum impacts stock prices generally across different companies.

Our next major step was to apply the Augmented Dickey-Fuller Test to our day/close information to indicate stationarity. This test verified that stocks typically perform in a random way, so we expect a statistical non-stationary result.

Finally, after this, we returned tables of the stock's name along with the corresponding mean of above-MA duration, median of above-MA duration, and P-value of the ADF, as well as a time-series plot of the Day Close dataset as compared to the MA at each point. This allowed us to easily access the results in an efficient manner once the separate programs for each stock finished running. With the results we could compare distributions between different stocks and generate plots that allow us to compare MA for each stock to the actual stock price.

Results

All the median durations of which a stock performed above the 50-day MA resulted in 5 days. This indicates that the industry standard of 50-day is appropriate for predicting how stocks will behave.

7 company stocks (Bajaj Auto Ltd, Colgate-Palmolive, Dabur Ltd, Hindustan Unilever Ltd, Indus Towers Ltd, Kotak Mahindra Bank Ltd, Reliance Industries Ltd) were statistically shown ($p\text{-value} < 0.05$) to be stationary by the Augmented-Dickey Fuller test.

Stock	N	Dickey-Fuller	p-value
Hindunilvr	1797	-3.70	0.024
Kotakbank	1769	-3.65	0.028
Colpal	1797	-3.65	0.028
Dabur	1797	-3.65	0.028
Industower	1797	-3.55	0.037
Bajajauto	1797	-3.52	0.041
Reliance	1797	-3.48	0.044

Table 1: Significant results from Augmented Dickey-Fuller test.

Upon plotting these time series, we noticed there is slight evidence of trends within each stock. This fact could be due to the peaks and valleys seen over time and the stock performed extremely well or crashed (Figure 1). This leads us to conclude that none of our 100 stocks are truly randomly performing.

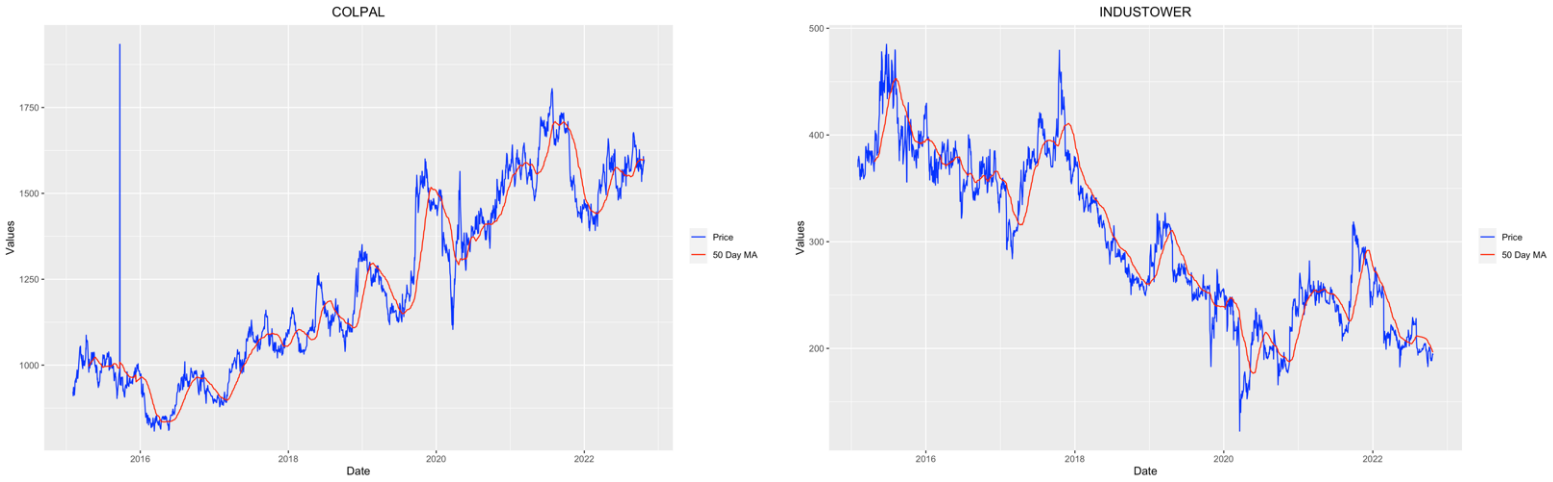

Figure 1: Two stationary stocks via ADF test (Colgate-Palmolive and Indus Towers Ltd)

To summarize, we believe that our approach to this project has brought conclusive results and utilized clustered computing in a way that delivers value, showing why using a major computing cluster can increase efficiency and produce results significantly faster than a linear computational approach. We have used our methodology to derive insights into the impact of positive momentum on stock prices, concluding that there is an effect and that the resultant performance of stocks is not random.

Contributions

All group members contributed to ideas, processes, writings, and presentation preparation.