Nifty Stock Data – Group 1

Marwan Lloyd, Max Schleck, Skylar Shafter, Rachel Studer, James Wan

University of Wisconsin - Madison

**Introduction**

Stocks and their movements have always been an interesting statistical question, as there is an immense reward for those who can predict their future price, as compared to their current price. As this problem is so complicated and so general in how one can approach it, there are many subfields of quantitative finance with take their own respective approaches to predicting future stock prices based off 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, 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 from India, looking at stocks from an index entitled Nifty 50, with 100 stocks included, with data taken from 2000 to 2021, respectively. This dataset is sourced from Kaggle.

**Process**

First, we use sub/sh files to divide the computing tasks by each stock file and then run them separately. The below analysis is on each individual job running individually and simultaneously within the computing cluster, having been submitted the necessary arguments and input files by the sub/sh files mentioned earlier.

To begin, we sought to clean our datasets before immediately conducting rigorous analysis. The original datasets provided minute-by-minute information on prices, which we wanted first reduced to only the day’s high, low, average, and volume. From there, we also reduced 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 set that was the count of days in which a stock was above the 50-day moving average, with each entry a separate instance being the number of days in which a stock was above the 50-day MA in a row, with only entries in which the price was above the 50-day MA for more than 2 days. We then focus on summarizing the durations of these intervals of positive momentum by finding their average and median, to get an idea of the general distribution for each, 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 is to apply the Augmented Dickey-Fuller Test to our Day Close information, focusing on the P-Value to determine if we can conclude the stocks perform randomly or not, by looking at if they are statistically shown to be stationary.

Finally, after this, we return 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. Then, we can easily access the results in an efficient manner once the separate programs for each stock have been completed, comparing distributions between different stock, and the plots 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 moving average resulted in 5 days. This indicates that the industry standard 50-days is appropriate at predicting how stocks will behave.

Graphical user interface, chart, histogram

Description automatically generatedGraphical user interface, chart, line chart, histogram

Description automatically generated 7 company stocks (Bajaj Auto Ltd, Colgate-Palmolive, Dabur Ltd, Hindustan Unilever Ltd, Indus Towers Ltd, Kotak Mahindra Bank Ltd, Reliance Industries Lt) were statistically shown (p-value < 0.05) to be stationary by the Augmented-Dickey Fuller test of stationarity. Upon plotting these time series, we noticed there is evidence of trends within each stock.

This leads us to conclude that none of our 100 stocks are truly randomly performing.

**Discussion**

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 having access to a major computing cluster can decisively 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.