CSIS 4260 – Assignment 1

The purpose of this assignment is to combine research, benchmarking and coding using the provided time-series dataset of stock prices for S&P 500 companies. The project is divided into three parts, each with its own research and coding components.

All your benchmarking will be conducted, assuming the current data size (1x) and future expansion of 10x and 100x. e.g. If you are researching between option 1 and option 2 and both are giving comparable performance at scale 1x and 10x but option 1 is easy to use then you can recommend option 1 in that scenario but if at scale 100x the option 2 is much faster, you can recommend option 2 at that scale.

Dataset: I will provide reasonably clean data for daily stock prices for 505 companies from 2013-02-08 to 2018-02-07 (619,040 rows). You will work with the entire dataset.

Part 1: Storing and retrieving data

The given dataset is in csv format (about 29MB), you will examine whether to keep the data in csv format (as it is) or to store it in parquet format including any compression scheme used (https://arrow.apache.org/docs/python/parquet.html). You will need to do both research and benchmarking to justify your choice at scales 1x, 10x and 100x.

Part 2: Manipulating, analyzing data and building models

In this stage we will primarily compare performance of two dataframe libraries Pandas vs Polars (https://pola.rs/) while doing data analysis and building prediction models. You will start with enhancing dataset with at least 4 technical indicators (https://www.investopedia.com/terms/t/technicalindicator.asp) which can be calculated based on the data given. Once calculated, these values will be added to the dataframe.

In the next part you will choose 2 algorithms to predict the closing price for the next day for all the companies. Please split your training and testing data using 80-20 split for back testing your algorithms.

Part 3: Creating a visual dashboard for the results

In the final part we will research two dashboarding libraries, some popular ones are Streamlit, Dash, and Reflex. You can also use another library after consulting with me. For benchmarking you can use only section a of the dashboard and then do the section b of the dashboard in the chosen framework only.

In dashboard section a, you will build a dashboard to display your benchmark results at all scales. In section b of the dashboard, we will display your price prediction models. The price predictions should be available for all the companies, meaning I should be able to select / search for a company ticker and the charts should update for that.