Capstone Project 2: RNN Stock Portfolio Analysis Milestone Report

Problem Statement

This project would implement a RNN that will model, observe, and forecast stock/exchange-traded funds (ETFs) data to determine the best stocks for portfolio growth. This would be relevant for investors, portfolio managers, financial/investment planners, and hedge/endowments to name a few. A few stocks will be selected for implementation and their collective performance will be observed.

My intended client would be any individual or institution concerned with investments. These entities would use my analysis and results for normal operations but also more importantly, for determining critical areas/weaknesses in a given timeframe should an unfortunate economic disaster occur such as the COVID-19 and Great Recessions. My project would serve as a reference and basis for implementing guidance, standard operating procedures, and investment education for the public.

The source time series/stock data I plan to clean and analyze is publicly available on Yahoo Finance as CSV files. In case this data is not enough for my analysis and modeling, I will search for additional datasets from other data repository websites such as Quandl, re3data, and other data marketplaces that offer other historical stock data for free.

After downloading CSV data on some stocks of interest, I will clean, perform EDA, and split the data into a 70/30 training/test set. Then, I predict the performance of buy-hold and buy signal strategies by implementing and training an RNN, more specifically a LSTM, over time. This LSTM is later tested to predict how well a portfolio may perform over a specified timeframe. The results are visualized through Matplotlib and Seaborn using time series plots and of course EDA/statistical testing plots ensuring data/model validity.

Data Wrangling and Cleaning

The following steps were taken to obtain, wrangle, then clean data into a form that is suited for analysis.

1) Obtained source CSV files for the following stocks/ETFs from the Yahoo Finance website:
    - AAPL
    - AMZN
    - DIA
    - FB
    - MSFT
    - NFLX
    - SPY
    - TSLA
    - TWTR
    - XOM

Daily data for each stock's/ETF's full history was downloaded to observe investment potential.

2) Used "describe" and "info" DataFrame methods to determine data validity and integrity for all downloaded CSV files. If a null value were present, the SimpleImputer function that uses the mean of the previous and following data points would be used to impute it.

3) Created time series plots to observe price behavior over time. All five price components (Open, High, Low, Close, and Adj Close) look normal with nothing out of the ordinary. Some of the plots have the Adj Close piercing shifted downward relative to the Close pricing due to dividends and corporate actions. They also appear to have orange dots on top and green dots on the bottom to show High and Low prices, respectively.

4) Looked for outliers as "anomalies" or deviations from the general trend of the data. These were sought by plotting upper and lower yearly rolling bounds that are four standard deviations away from the yearly moving average. If a data point exceeds these bounds, then an outlier is found. However, there is usually a rationale explanation or definite cause. In the stock market, common examples are (but not limited to) stock splits, changes in leadership, poor sales performance, or disasters/sudden phenomena. Thus, these outliers are highlighted and discussed, but not removed or discounted.

Data Storytelling

For this section, stock source data are analyzed to answer the following questions below. Numerous plots were generated and can be viewed in this GitHub repo link to the Jupyter Notebook containing the analysis.

1) How well are selected popular stocks and ETFs currently performing since their Initial Public Offering (IPO)?

Analysis and Outcome: From plotting the source data, nearly all stocks/ETFs (i.e. TWTR being the exception) are performing well since their IPO. In other words, for a specified stock/ETF, if the final Adjusted Closing price is greater than the starting Adjusted Closing price, it should be considered as a potential investment opportunity. Some of them were unfortunately affected by market corrections such as the Great and COVID-19 Recessions (and even recovered over time), but the others continued to perform positively uninfluenced.

2) Are there any correlations or covariances in pricing between the selected stocks and ETFs?

Analysis and Outcome: Yes, most notably between pairings of any stock/ETF listed below.

- AAPL
- AMZN
- FB
- MSFT
- NFLX
- DIA
- SPY

Any unique pairing from this list will result in a high Pearson Adjusted Closing price correlation ranging anywhere from 0.87-0.99 and a high log covariance ranging anywhere from 7.1-13.0 as shown visually in the next section. Since all prices (i.e. High, Low, Open, Close, Adjusted Close), follow nearly identical shape and trend, we can safely assume that these correlations hold for all prices.

Statistical Data Analysis

This section summarizes the steps and corresponding results of performing Statistical Data Analysis on the stock source data. For this analysis, Time Series EDA was performed and heatmaps of correlation and covariance were made. These two steps were performed on Adjusted Closing price data and, as alluded previously, we can safely assume that these results will be very similar for all prices since they follow nearly identical shape and trend.

Steps and Findings

1. Conduct Time Series EDA: Create plots visualizing Density, Seasonality, Trend, and Noise for each stock/ETF to characterize the data as time series. Visualizations can be viewed in the corresponding Jupyter notebook.

   Results and Findings:
   - Density: Here we see that five of them (AAPL, AMZN, MSFT, NFLX, and XOM) were priced closed to zero most of the time possibly due to the market not being familiar with them at first. The other five (TWTR, FB, TSLA, SPY, and DIA) already had popularity prior to their IPOs and caused price movement in the beginning. One minor caveat is for DIA, SPY, MSFT, and XOM, whose Close and Adjusted Close plots are different due to corporate actions and dividends.
   - Seasonality: By plotting the source data as done previously, we can suspect that there are no discernable patterns or cycles in any of the stocks/ETFs. These seasonality plots show straight horizontal lines at y = 0 which confirm this suspicion.
   - Trend and Noise: The source data was plotted alongside its yearly simple moving average to determine the general trend and focus less on the noise/fluctuations.

We see that any of them either trend positive, negatively, or neither. Some even trend linearly or parabolically.

2. Correlation and Covariance Analysis: Calculate Pearson correlation coefficients and covariances between stocks and ETFs (confirm if relationships are present between any stocks/ETFs).

   Results and Findings: As stated in the Data Storytelling section, the Pearson Adjusted Closing price correlation coefficients and covariances of the all stocks/ETFs are shown in Figures 1 and 2, respectively.

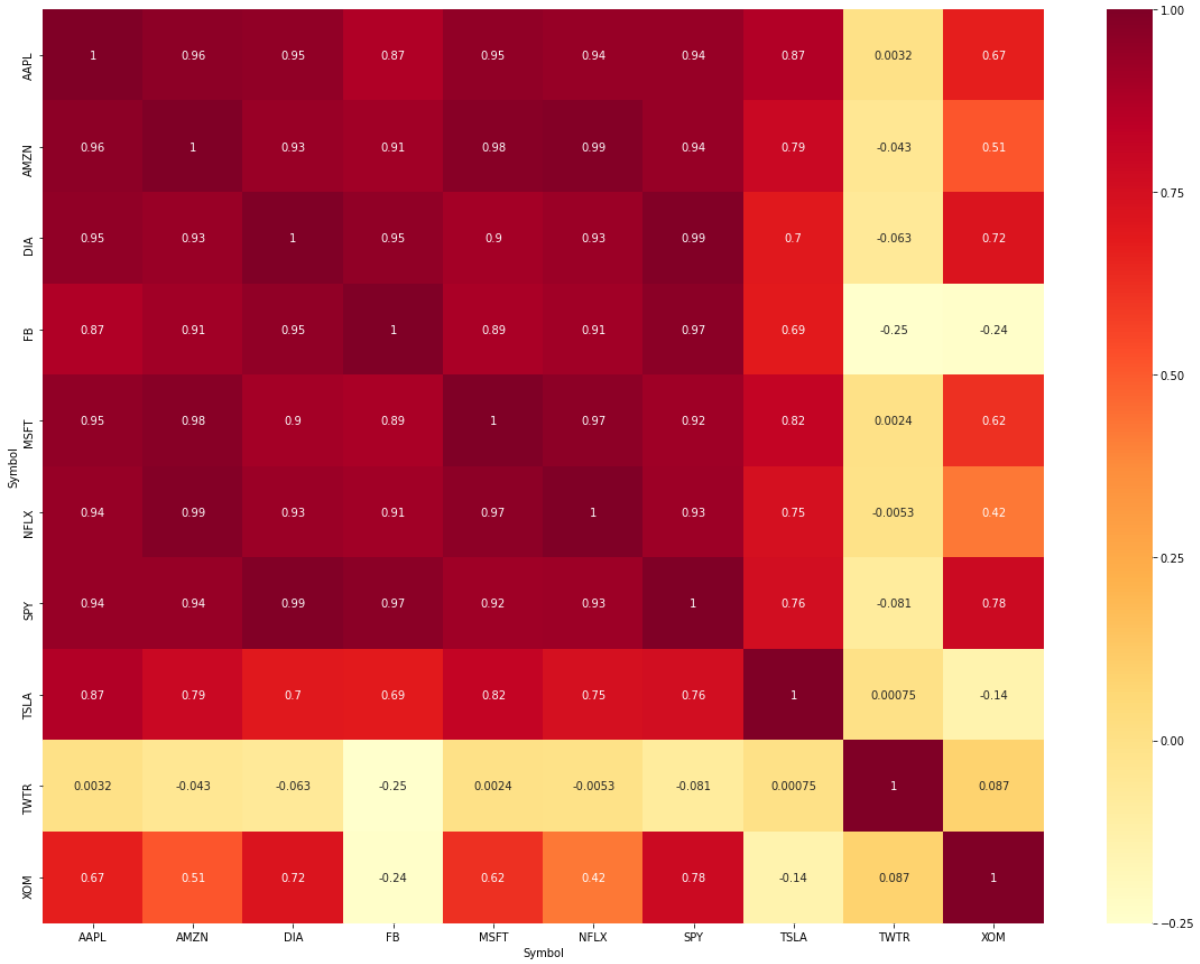

Figure 1. Correlation Coefficient Heatmap

Here in Figure 1, we see the strongest positive correlations shown by the dark red/maroon boxes meaning any pairing made from the following:

- AAPL
- AMZN
- FB
- MSFT
- NFLX
- DIA
- SPY

These make sense since i) these stocks are the top technology companies that have significant influence over the U.S. economy, and ii) these two ETFs are overall indicators of economic health and these technology companies are listed in these ETFs. Thus, for a simple buy-and-hold portfolio, investing in any or all of them should be considered.

On the other hand, we also see that correlations with TSLA, TWTR, and XOM range from moderately positive to essentially non-existent, but these stocks may be considered under different investing strategies.

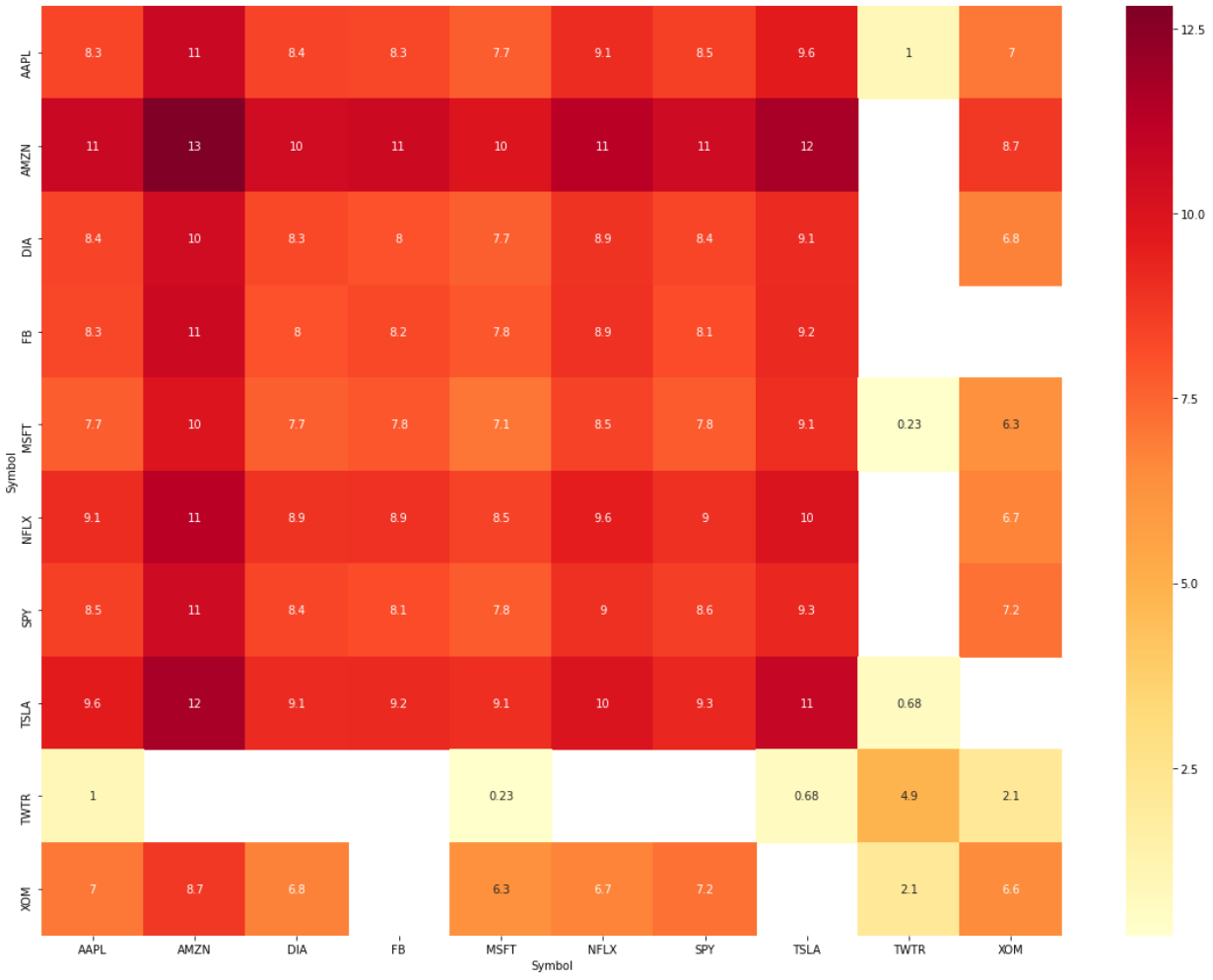

Figure 2. Covariance Heatmap

For Figure 2, the covariance data was logarithmically transformed for clarity since this properly highlights strong positive covariance and reinforces the strong relationships between pairings of the stocks/ETFs discussed previously. Using a linear scale would produce a misleading and confusing visualization.