Final Project Proposal

Group 7 - Logan Talton, JP Pham, Richard Huynh

Submit an approximately 400-word description of what your group plans to do. This includes some commentary specific to each of the 6 components described in the Final Project Report instructions.

The main question that our analysis aims to answer is whether we can predict future stock prices or stock price movement and performance, based on historical data of ten big tech companies. We aim to use a machine learning algorithm and train our data using linear regression or a recurrent neural network such as an LSTM model to obtain results. Linear regression has benefits of being straightforward and easier to implement, as the relationship of the independent and dependent variable will be represented by a linear function. However, there are weaknesses, since linear regression assumes linearity of data, while stock prices are often influenced by non-linear factors. An LSTM, however, is more sophisticated and is particularly good at predicting time series data such as stock prices as the data is sequential. Lastly, we may include some correlation analysis to see how some stocks affect related stocks. For example, a rally in NVIDIA stock may have a similar rally effect in Intel stock since they are both semiconductor companies.

The data we used will consist of stock prices within our datacamp assignments. The module already contains CSV files with various information from specific tech companies. If we need more data, we can also opt to use external data from places like Yahoo Finance.

To clean and process the data, we plan to use Numpy and Pandas. However, the data that is already in our datacamp module is very good and clean. We will most likely separate the data into smaller, more manageable chunks if need be. From the dataset that we already have, there seem to be no NULL values or values that will cause issues either. However, to help mitigate some weaknesses of predictive modeling, we can identify and manage outliers before modeling training, to reduce its impact on a linear regression model.

After the data is cleaned and prepared, we can start using visualization techniques and modeling algorithms to summarize the data. We can use the linear regression as a starting point, to get a basic understanding of machine learning algorithms and of future predictions of the stock prices, but then we can compare the performances of the linear regression modeling to an LSTM to gain a more insightful analysis of the data, which would account and help cover weaknesses of the linear regression model.

To interpret our data, we will visualize the actual stock prices and overlay our stock prices from what our model has predicted. We plan to use Root Mean Squared Error (RMSE) to measure the average magnitude of the errors between the actual stock price and our predicted value. To make our model more robust, we also plan to use cross-validation to ensure the model is consistent. To help us visualize

the results of our predicted prices, we can plot the actual stock prices compared to our predictions, as the last date of the data we use ends in December 2021.

As a group, we will plan to start the project early, to make sure we all gain a high understanding of the modeling techniques we are using, to implement, analyze, and interpret our results properly. We will all contribute to the completion of the different components of Docker, Github, data analysis, and interpretation of our findings.