

# Project Phase 1 Report

## Group 1

### 1. INTRODUCTION

In our project we will be exploring patterns within stock and general economic data with the intent of successfully classifying high and low risk stock options. For our sources we are pulling data from yahoo finance via yfinance and Quandl. We are also pulling from the World Bank and the Federal Reserve Bank of St Louis to find a number of yearly economic indicators for the US. Each row represents key data for a singular day of trading while dealing with a given year's economic factors.

### 2. DATA DESCRIPTION

This data uses the following elements.

1. **Date:** It's important to be able to connect each day of trading to outside events and date is how we will do so. This represents the day the trading took place.
2. **Open:** Prices can change while the stock market is closed so the price at the start of each day is important for understanding the baseline other traders will be working with.
3. **High:** Great indicator for stock volatility especially when used with Low. Tells us the highest price the stock reached on that particular day.
4. **Low:** Great indicator for stock volatility especially when used with High. Tells us the lowest price the stock reached on that particular day.
5. **Close:** Final price of the stock before trading stopped for the day, tells us results of traders and external economic impact on a given day, separating the two will be an ongoing challenge during data mining.
6. **Volume:** Total number of stocks traded over the course of a day.
7. **Dividends:** The dividend is the reward for the stockholders given by the corporation. This is usually done when the company is booming with profits. The rewards are offered either in form of cash or some extra stocks(also known as splits). The dividend is usually rewarded quarterly or twice a year or monthly. Some companies offer a one time dividend across the year. Having said that the dividends are not offered by every company and the dividend policies of a company are completely dependent on the corporation. These are important because they have a significant impact on the relevant buying and selling strategies associated with this stock as they incentives more conservative trading strategies.
8. **Company:** It's important not to forget what stocks are meant to represent. Specifically they are a fraction of the total value of a public company that any member of the public can invest in with the possible reward of that company's value growing, therefore growing the value of the investment. The company column is very simply the name of the company in question that is being bought and sold on the open market with all other columns representing it's value.
9. **Adjusted closing price:** The closing price of a stock is nothing but the value of the stock at the end of the day. However this is not the accurate representation of the stock price. As there are multiple factors that affect the stock value. For instance the stock appreciates the corporation decides to reward the stockholders with a reward usually in the form of dividend. This dividend can be rewarded in the form of cash or it can be in the form of some extra stock. This changes the true value of the stock. Eg: The closing price of a stock is 100\$ , the dividend rewarded to the stockholders is 5% i.e. 5\$ then the adjusted closing price is 95\$.
10. **Splits:** In order to combat over inflated stock prices a company will sometimes perform a split. This means that it will (usually) decrease the price of the stock in exchange for adding more of it into circulation. The number here tells you how many stocks you now have after the split by multiplying whatever number you had originally by the split variable. A split is a rare occurrence and a 0 does not mean that you have 0 stocks after a given day, simply that a split did not happen. This falls into the same field as dividends, companies that issue splits are ones that are looking to be a part of a more conservative investment strategy or are looking to combat inflation. Either way it's an informative statistic about the stock at large.
11. **GDP:** Gross domestic product: the total value of goods produced and services provided in a country during one year in US Dollars. Simple way of evaluating if this particular trade occurred during a net positive or net negative year, can be used alongside aggregated stock data to show how reliant stock price is on the economy at large.

12. **Industrial Output Index:** A percent production of Industrial production is a measure of output of manufacturing based industries, including those producing goods for consumers and businesses based on the output of the same for 2012. Correlation (or lack thereof) can show how reliant a given company is on demand for US manufactured goods. Because the US has changed to being a service industry over time this is an important statistic as it can indicate the viability of this business into the going future.

### 3. PLANNED DELIVERABLES

- Create proposal and choose starting datasets. (6/8/2020)
- Add additional data from yahoo finance and Quandl, evaluate if additional economic indicators are needed and add them as appropriate to the dataset. Ultimately aim to finalize data that will be used. (6/15/2020)
- Begin classification analysis. Aim to use general indicators of stock volatility and risk and make programs

to classify based on certain borderline values. Develop and improve methods in use and evaluate results. (6/15/2020)

- Integrate Association. Aim to look at what stocks that were deemed the most volatile and look at what fields with what range of values was causing the greatest trouble.
- Work and submit a revised version of Phase 1 report to reflect our progress, and any changes anticipated in the final submission for Phase 2. As well as all project files including source code, dataset description and sample subsets of the data, and documentation for installation and usage. (6/23/2020)
- Keep working on the analysis on the datasets, while working on the research report, presentation, document all the source files, and README file. Submit all of them as Phase 3 as required at the end of week 10. As well as the peer evaluation. (7/27/2020)