# STAT 605 Project

## IntroDuction

With the 2020 election, the stock market has fluctuated a lot. Many companies had ups and downs. In this project, the data we are going to use is the StockMarketData- From1996To2020 dataset from Kaggle.The data is from 1st Jan,1996 to 7th Aug,2020. This dataset almost includes every stock in the market and has more than 100k rows. There are some stocks which behave really well in some certain years. Suppose we have some stocks whose behaviors are really similar to those stocks' early stage, we want to see if they are more likely to become a great stock than the others. For now, we use correlation to measure the similarities between stocks and find some stock with high similarities with Tesla. The searching results are great. In the next steps, we will try more advanced measurements and see their performance.

## Contents

The dataset of this project comes from Kaggle, which is a large and famous platform for data science competition and practice. You can access the data through this link: https://www.kaggle.com/aceofit/stockmarketdatafrom1996to2020/discussion/178080 This dataset contains approximately 100k stock time series and their price from Data is from 1st Jan,1996 to 7th Aug,2020. The size of this dataset is more than 10GB and therefore meets our requirement for this project. This dataset contains two parts: One is Tickers.xlsx. This file consists of Tickers' names for more than 1,00,000 companies across the globe. It consists of 5 column

Ticker : Ticker name used in Yahoo Finance

Name : Name of the Company

Exchange : Short Name of Exchange in which Company is registered.

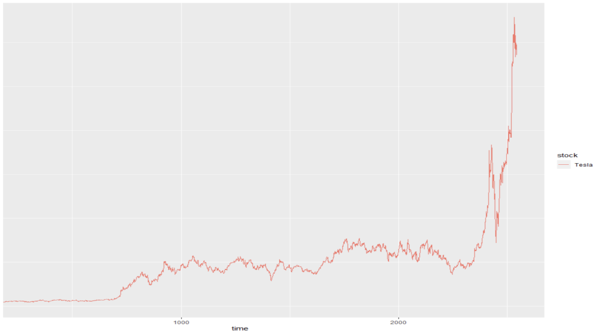
Category : Category of Business, Like Electronic Equipment, Internet Information Providers etc

Country : Country name in which Company is Registered.

The other is Data folder.This Directory Consist of more than 1,00,000 directory, in which company data is Stored.Each Company has separate directory for its data file.The variables in the Data Folder are as follows:

- Date: The data that this data was collected
- Open: The Stock at which opens at the start of market
- High: The particular stock which made high during that particular day - Low: The Particular stock which made Low during that particular day
- Close: The stock closing at the end of the Market hours

We decide to choose Close as a measurement of stock price and below is one visualization of one stock time series:



This dataset contains some missing values, and wrong values. For example, some stock contains a null time series, and some stocks' values are obviously wrong. We spent some time in the data cleaning part. One of our team members writes scripts to get the right stock data corresponding to the stock symbol from Yahoo Finance.

Another thing to note is that there are several folders like AAPL, AAPL34F.SA, AAPL34.SA, AAPL.BA, AAPL.MX, AAPL.SW They are Apple Inc prices in different exchanges around the globe and are calculated in different currencies.Therefore, when we find stocks with high similarities, we need to remove duplicates like this.

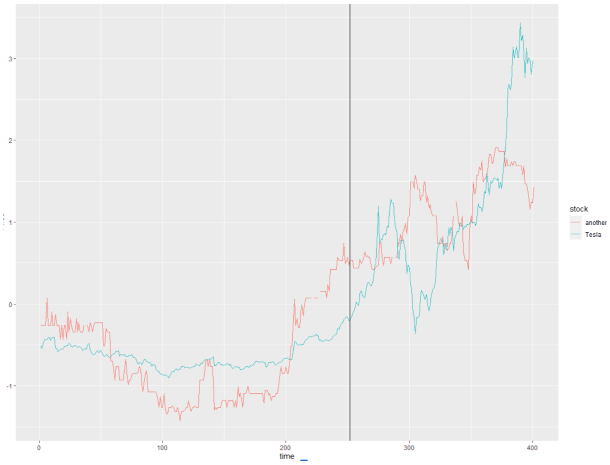
There are some stocks which behave really well in some certain years, like Tesla in 2020 and Amazon in these 5 years. The purpose of this project is: suppose we have some stocks whose behaviors are really similar to those stocks' early stage, we want to see if they are more likely to become a great stock than the others. There are two statistical questions: First, how do we measure the similarity of stocks. Second, once we find the stocks which are similar to some good stock, will they become a great stock in the next year?

Our plans in this proposal are:

1. We are going to load the data to CHTC.To have better computational performance, we plan to write shell scripts to split 100k stocks to 1000 folders and run these 1000 folders parallelly.
2. We will try different R models to measure the similarity of time series, like Dynamic time warping and more.
3. We will extract 100 stocks whose performance is the best among all the stocks in a given year as a target set. And we run the R file in CHTC to find out the one with the highest similarity. We will compare it with the target set and calculate the accuracy.
4. We will change different models and see which model performs best.

For now, our progress is:

1. We wrote bash scripts to split 100k into 1000 folders and run these jobs parallel on CHTC successfully .
2. We have successfully done the experiments on the simplest similarity measurement - correlation. We use this method to verify if our results can work, and it is easier to debug in the beginning.
3. We ranked all the 100k by their similarities with the target stock. For example, the figure below shows a stock and Tesla stock. These two stocks are extremely similar before the black line. After the black line, the red stock goes well. We think that our way of finding good stocks has its practical meaning.



# Future work

In the following weeks, we are going to try different measurements of similarity between time series, like pattern model representation (PMR) of time series. PMR is based on a piecewise linear representation (PLR) and is effective at describing the tendency of time series. Then, the pattern distance can be calculated to measure the similarity of tendency. This method overcomes the problem of time series mismatch based on point distance. According to the numbers of series' segmentations,pattern distance has a multi-scale feature and can reflect different similarities with various bandwidths. Because normalization is unnecessary, the calculation consumption of pattern distance is low. Therefore, this might be a good measurement in our project. We will try and see if this can improve our search.