Project Proposal

- Chenyang Jiang / cjiang77
- Haoran Teng / hteng22
- Chenhao Fang / cfang45
- Shikun Liu / sliu673
- Zixiang Xu / zxu475

Description

With the 2020 election, the stock martket 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. In Data Folder, there are some Company wise folders and in each folder there are csv files which consist of seven columns, namely: Date,Open,High,Low,Close,Adj Close,Volume. The data is from 1st Jan,1996 to 7th Aug,2020. This dataset almost includes every stocks in the market and has more than 1 million rows. Also, the size of this dataset is more than 10GB and therefore meets our requirement for this project.

Statistical question

There are some stocks which behaviors really well in some certain years, like Tesla in 2020 and Amazon in these 5 years. Suppose we have some stocks whose behaviour

are really similar to Tesla's early stage, we want to see if they are more likey to become a great stock than the others. The statistical questions in our project are: First, how do we measure the similarity of stocks. Second, once we find the stocks which are similar to the Tesla, will they become a great stock in the next year.

Short code snippet that reads the data onto our laptop

```
u=read.csv("D:/Desktop/archive/Data/Data/-B.TO/-B.TO.csv")
```

(The data set is easy to read since all the data are stored seperately in csv files.)

Description of the variables available

Below are the variables in our dataset:

- 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

Computational tools

We are going to use R as our software and CHTC as our hardware in this project. The schedule is as following:

1. We are going to load the data to CHTC. To have better computational performance, we plan to write shell scripts to split 1 million 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.