# Assignment 2

## Weight: 60%

- Code + Report + Interview: 40%
- Prediction Result: 20%

## Introduction:

The research of the stock market is currently one of the hottest topics in the field of data mining and machine learning. The purpose of this assignment is to discover how to use data mining and machine learning to solve the real stock market problem. The evaluation method for stock market forecasts will use **stock returns**[1]. Each group only needs to submit one solution.

In this assignment, there are **no restrictions** on data sources and styles. You can use numerical data, textual data, pictures, audios, etc. The following are some examples of data acquisition methods:

- Numerical data - e.g. Stock opening prices, Stock closing price, Volumes, etc.
    1. Yahoo Finance[2]: Web download or using Yahoo finance API
    2. Professional financial platforms:
        1) Bloomberg Terminal
        2) Thomas Reuters Eikon
        3) Wind Financial Terminal
    3. Third-party Platforms
- Textual - e.g. 10K report, 10Q report, Social Media posts, Expert posts, etc.
    1. EDGAR database from US (can find 10k report)
    2. Professional finance platforms (See above)
    3. StockTwits: a social media platform designed for sharing ideas between investors
    4. Twitter (Notice: You may need to register a developer account)
- Etc.

## Evaluation:

For the stock market forecast, we will evaluate the forecast stock's return within a week (day & week). The predicted target time is **from May 24th to May 30th**, 5 days. The target number of stocks is **50**, please refer to **Stock_Ticker_List.xlsx** (on Kaggle). The stock market platform can search for stock data based on tickers (symbols).

Due to the high volatility of the stock market, it is very challenging to actually predict the trend of stock returns. Your prediction will be **ranked** according to the prediction accuracy and you will be scored based on the final group ranking.

---

[1] Appendix 1: “How to Calculate Stock Returns”

[2] Appendix 2: “Sample: Crawl Data from Yahoo Finance”

## Submission:

1. Prediction Result

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code. In this assignment, we will use Kaggle In-class Competition to record your prediction results. You only need to upload the result (**.csv file**) to the Kaggle competition (please refer to Table 1 for the result format). The final prediction result ranking will be displayed on the Kaggle competition leader board.

- Competition Name: **COMP3027J Assignment 2 - BDIC2021**
- Participate URL: https://www.kaggle.com/t/61be73cd9f2d4c8bacf6598aa6fdef9f
- Group name: Group_XX (e.g. Group_01)
- 50 Stocks List: Please refer to **Stock_Ticker_List.xlsx** in Kaggle competition dataset.
  - ✻ Note: The order of the stocks in the final result should be the same as the order of the stocks in the file provided.
- Result Format:
  - Return: Do number prediction. Fill in the number part before the percent sign, and keep two decimal places for the number. e.g. for 5.326% enter 5.33

Table 1: Result Format (.csv)

| tickers | return_day_1 | return_day_2 | return_day_3 | return_day_4 | return_day_5 | return_week |
|---|---|---|---|---|---|---|
| stock_1 | -6.46 | -6.21 | -2.98 | 4.49 | -6.24 | 5.11 |
| stock_2 | -1.86 | -1.68 | 6.52 | -7.82 | -4.28 | 9.08 |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| stock_49 | -8.46 | -8.13 | 3.87 | 4.4 | -7.78 | -8.08 |
| stock_50 | -5.73 | 7.15 | 4.45 | 2.62 | -1.54 | 6.83 |

- Pre-test Period:

In order to avoid using the wrong format in the final submission, we will provide a pre-test experiment. The order of 50 stocks is the same as the sample data. The pre-test training data set will provide the daily opening price, closing price, highest price, lowest price, adjust closing price and trading volume of each stock. The prediction results can be uploaded to Kaggle for evaluation and ranking to check whether the model predicts correctly and whether the format of the model result is correct.

The pre-test phase will start on May 3rd (Week 10, Monday) and end on May 16th (Week 11, Sunday). Training data can be obtained from Kaggle Dataset. After the pre-test phase is over, we will clear the leaderboard and no longer accept test data.

The pre-test data set comes from Yahoo Finance. The training data is derived from 50 stocks for 5 years (~ Apr. 23rd 2021). The target time of the prediction test is the 9th week, April 26th ~ April 30th, 5 days. Please refer to the Sample_Answer.csv format and upload the .csv solution to Kaggle.

Note that in the pre-training phase, please upload your predictions rather than upload real-time stock data to avoid interference with the leaderboard. (Teams whose answers are exactly the same as the test set will be removed and will not be allowed to participate in the pre-test).

- **Deadline: Sun, May 23rd, 23:55 - Beijing Time (Submit on Kaggle)**

✻ Note: Due to the effectiveness of the job review, we do not accept any overdue submissions.

2. Code + Report

Your pdf report should clearly detail how you carried out the experiment to address this challenge.

1. Your report should be written in Overleaf or Word, and use the provided template: https://www.acm.org/publications/proceedings-template.
2. It should be a human-readable document (e.g. do not include code)
3. The final report is expected to be 4-8 pages including references.
4. You should provide student numbers instead of address in the provided template.
5. Use clear headings for each section.
6. Include tables and figures if needed appropriately, such as giving captions, describing your figures or analysing the results provided in your tables in your text etc.
7. The final report filename should be "Comp3027J_GroupXX" (e.g. Comp3027J_Group01)

In your report, it is recommended to discuss the following essential topics, but not limited to these topics:

1. Literatures Review: Before starting the project, you should do literature review to see how the predecessors solved this type of problem.
2. Dataset selection (collection) and Data pre-processing.
    a. Where you find your data (or how do you collect the data and create your dataset)?
    b. How do you analyze your data?
    c. how to pre-process your data to fit your solution?
    d. Any challenges with your dataset? etc.
3. Methodology
    a. Any machine learning algorithm can be used (not limited to the algorithm we have learned).
    b. Creativity is encouraged.
    c. Be careful, a sophisticated approach with little description and explanation will receive little credit.
4. Evaluation
   Except submit your results to Kaggle, you should evaluate your model yourself, for example,
    a. Compare your solution with benchmarks in literature;
    b. Evaluation metrics for your task;
    c. Analysing your results etc.,
    d. Which model finally is chosen to generate predictions submitted to Kaggle.

- **Deadline: Tuesday, May 25th, 23:55 - Beijing Time (Submit on BrightSpace)**

# Appendix 1

## How to Calculate Stock Returns

A **return**, also known as a **financial return**, in its simplest terms, is the money made or lost on an investment over some period of time.

The traditional price return calculation method is as follows:

$$R = \frac{V_f - V_i}{V_i}$$

$R$: Return

$V_f$ : Final Value, including dividends and interest

$V_i$ : Initial Values

In this assignment, we do not consider such as dividends and interest. Therefore, the daily return and weekly return will be calculated according to the following formula:

$$R_t = \frac{V_{t_{close}} - V_{t-1_{close}}}{V_{t-1_{close}}} \text{ Daily Return}$$

$$R_t = \frac{V_{t_{close}} - V_{t-7_{close}}}{V_{t-7_{close}}} \text{ Weekly Return}$$

$R_t$: Return at time t

$V_{t_{close}}$ : Closing price at time t.

$V_{t-1_{close}}$ : Close price at a day before time t.

$V_{t-7_{close}}$ : Close price at a week before time t.

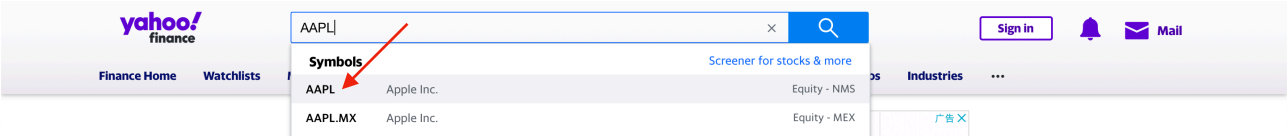
**Note**: The weekly return of the final prediction result will only calculate the result within the prediction period (calculated using the closing price at the end of Week 12 and the closing price at the end of the previous week). So there should be only one answer for each stock.

# Appendix 2

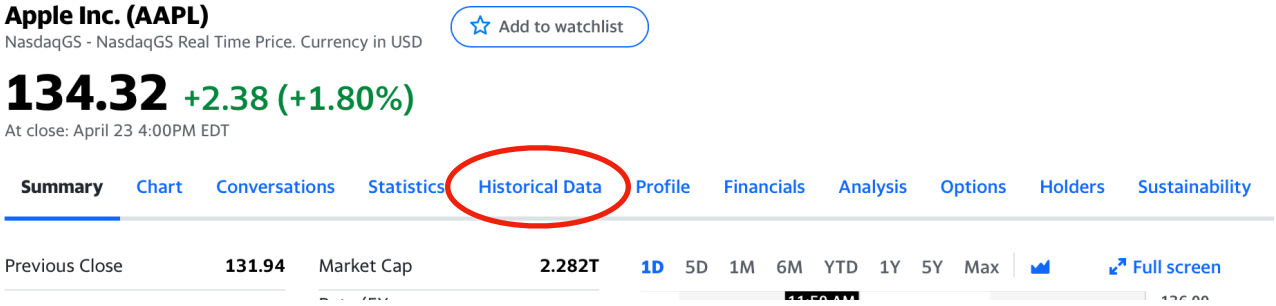
## Sample: Crawl Data from Yahoo Finance[3]

- **Method 1**: Download through the website

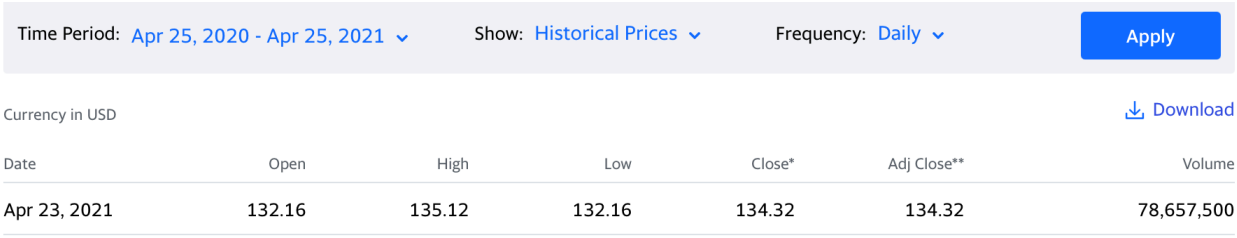
For example, if you want to download the Apple (tickers: AAPL) stock price data from Yahoo finance. You can search Apple's tickers directly through the search box to find Apple's information.




After opening the webpage, click Historical Data to get Apple's historical stock data.




By selecting the start time and the end time, the historical data of the stock can be obtained. The frequency can be selected as required. After the selection is complete, click Apply, and the corresponding historical data will be displayed below. You can download the selected data by clicking the Download button.




- **Method 2:** Download through the Yahoo Finance API

**yFinance**[4] is a third-party API that can download historical stock data from Yahoo finance. Please refer to the yFinance document for specific installation methods and instructions for use.

  - Install: pip install yfinance

[3] Yahoo Finance: https://finance.yahoo.com

[4] yFinance: https://pypi.org/project/yfinance/