

# # Report for the final-project :

- Course info : I&C SCI\_X426.85 (SUMMER 2022/REG 00267/SEC 1)
- Project theme : Thinking about Purchasing Stock Case Study
- Group info :
  - Group 2 : Qi Zhang & Xinmeng Liu

Notice:

- Please access to the ipynb file appendix to check the codes.
- This PDF is only used to explain the operation.
- Here we have a dictionary

## # Report for the final-project :

Task 0: Comprehend the features the csv file

### Task 1: The choose of data set

### Task 2: Data Cleansing

### Task 3: Analysis

1. Check the trend by close price
2. Check the trend by MA of adjusted price
3. Calculate the Rate of Return by Adj Close
4. Check the correlation between stocks
5. Check the Expected Returns and Risk
6. Train the models to predict the trend of adjusted price
  - Step 1: Feature Engineering
  - Step 2: Data Cleansing & Train\_test\_split
  - Step 3: Choose the regression model and train them
7. Try to predict the adj price by date
8. Compare two stock based on the prediction
9. Summary





## Task 0: Comprehend the features the csv file

| Feature        | Explanation  |
|----------------|--|
| Date           | The recorded date                                      |
| Open           | The price of the stock when the market open            |
| High           | The highest price of the stock today                   |
| Low            | The lowest price of the stock today                    |
| Close          | The price of the stock when the market closed          |
| Adjusted close | The adjusted price of the stock when the market closed |
| Volume         | The number of transactions today                       |

Notice: When it comes to the history stock analysis, adjusted close price is more useful.

## Task 1: The choose of data set

Our group has downloaded 4 stock records from the [Yahoo Finance](#).

|   |                |                      |        |
|---|----------------|----------------------|--------|
|  Apple.csv     | 2022/8/6 16:13 | Microsoft Excel 逗... | 693 KB |
|  Meta.csv      | 2022/8/6 16:13 | Microsoft Excel 逗... | 185 KB |
|  Microsoft.csv | 2022/8/6 16:13 | Microsoft Excel 逗... | 620 KB |
|  Tesla.csv     | 2022/8/6 16:13 | Microsoft Excel 逗... | 208 KB |

And we choose to use the stock records of Apple corporation and Microsoft corporation because their amount of data is big enough.

## Task 2: Data Cleasing

1. Import the csv file through `pandas`

Notice: The 2020 data is not needed at all due to COVID-19 impacts.

2. Drop the null values of these data sets with null values

Analysis: **No value is missing in these 2 data sets**

3. Apply `PCA` analysis or `standarization` if necessary.

Analysis: Unnecessary to do these operation at present.

4. Use the date of the stock record to replace the index.

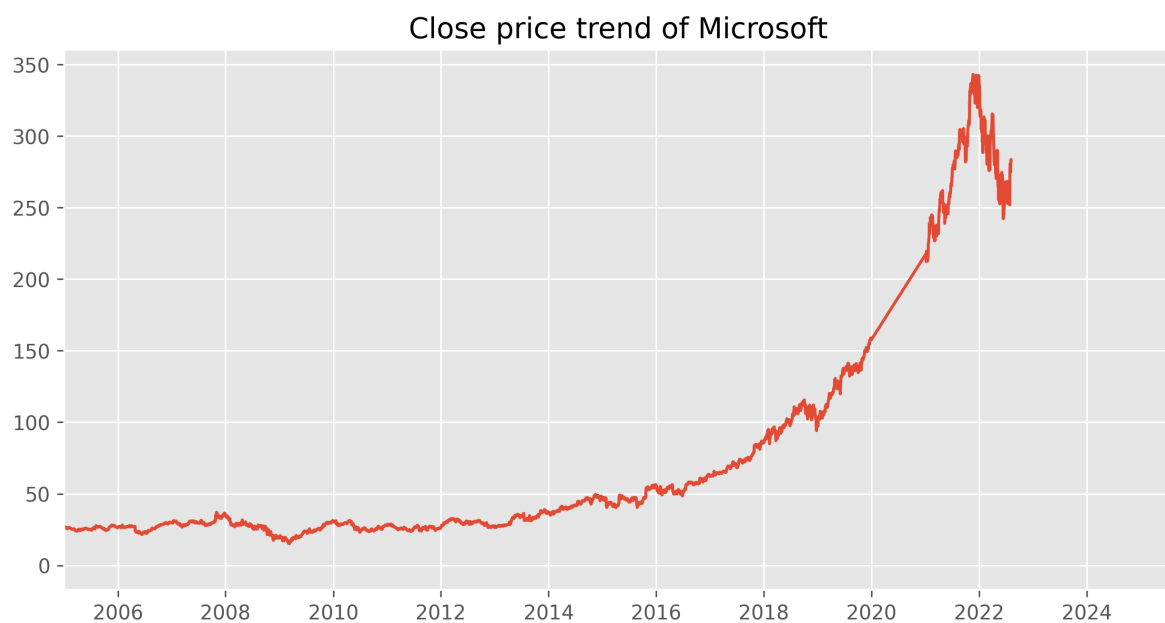
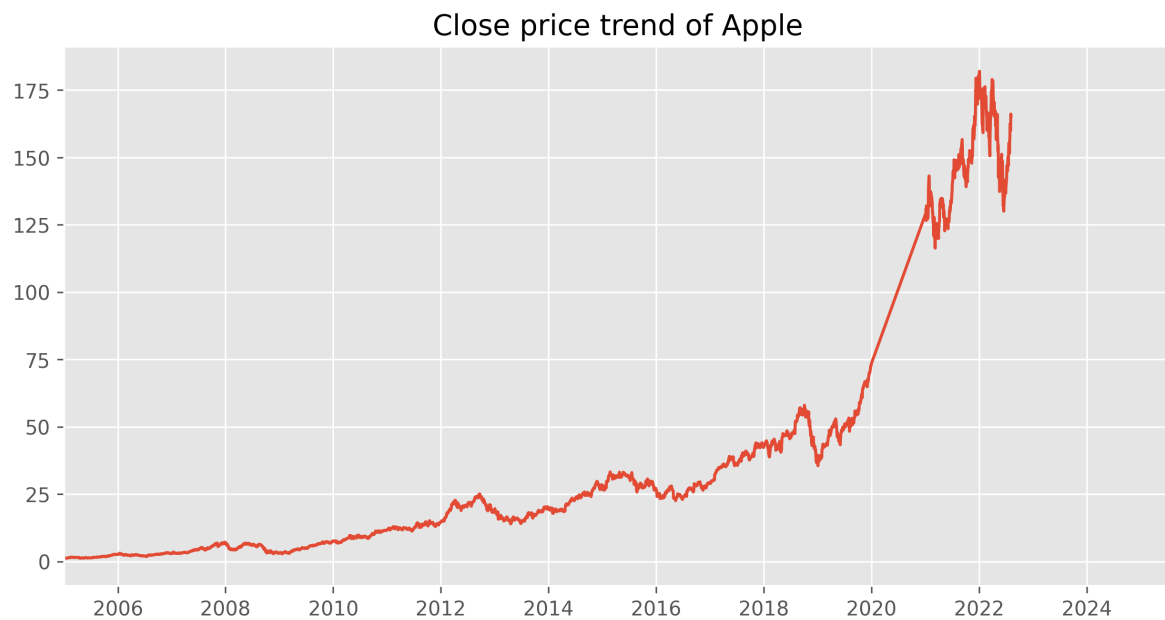
## Task 3: Analysis

### 1. Check the trend by close price

Here we draw the plots according to the daily stock close price of Apple and Microsoft between 2005-1-1 and 2025-8-1 excluding 2020.

We can see that the close price of these 2 company is increasing in general.

The price decrease at the end of 2021 but begin to ascend recently.



## 2. Check the trend by MA of adjusted price

### Moving Average (MA):

- a moving average is a calculation used to analyze data points by creating a series of averages of different subsets of the full data set
- By calculating the moving average, the impacts of random, short-term fluctuations on the price of a stock over a specified time frame are mitigated.

Here we use the SMA:

$$SMA = \frac{A_1 + A_2 + A_3 + \dots + A_n}{n}$$

*A = Average in period*

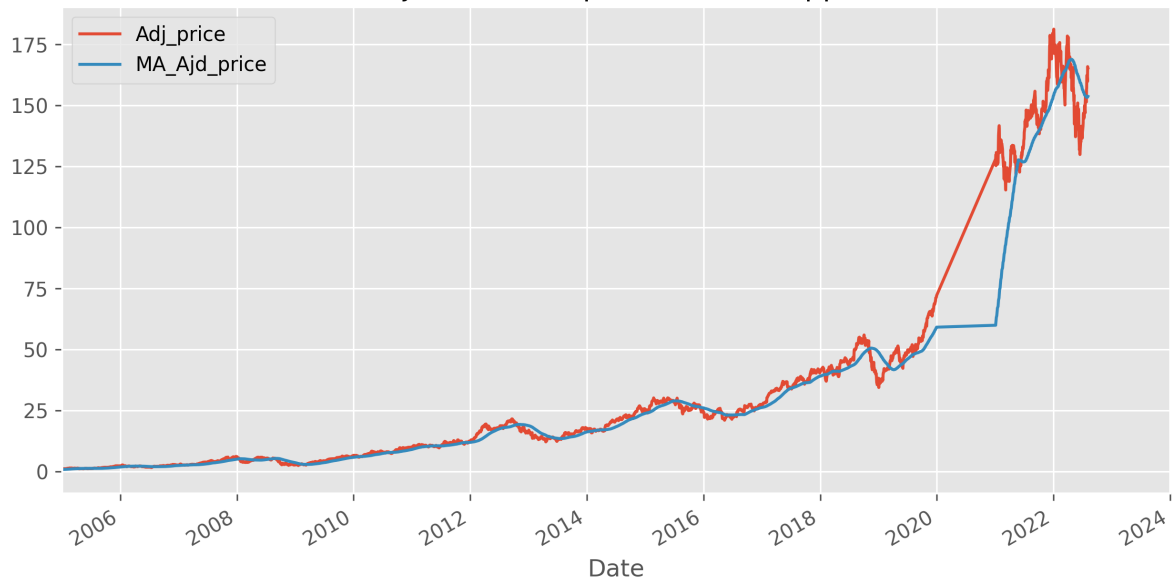
*n = Number of time periods*

According to the plots, we can see that :

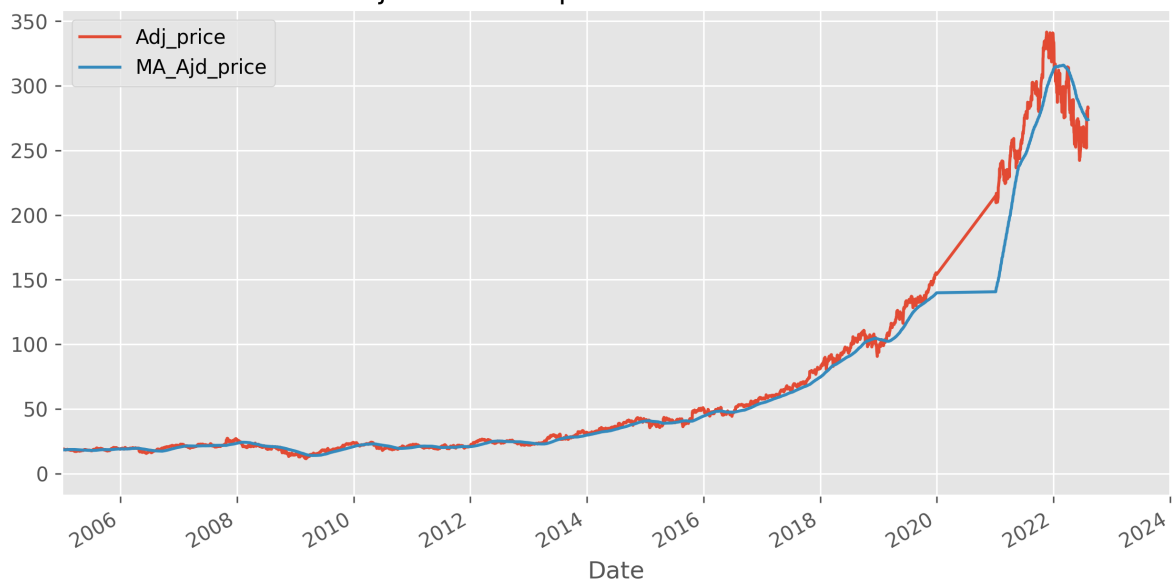
1. The adj price of Apple and Microsoft **keep increasing in general** in this time period.
2. **However, the price ascending of Microsoft seems smoother and the price of it is more stable than Apple.**

- I guess it's because the profit of Apple mainly depends on the release of its new product. So the price of its stock can be up and down or you can say seasonal.  
  
But Microsoft's profit mainly comes from the bonus of softwares. Although the software keeps updating, it can not suddenly attract a large amount of people to buy its product or abandon its product. That's why its stock seems more stable.

Adjusted close price trend of Apple



Adjusted close price trend of Microsoft



### 3. Calculate the Rate of Return by Adj Close

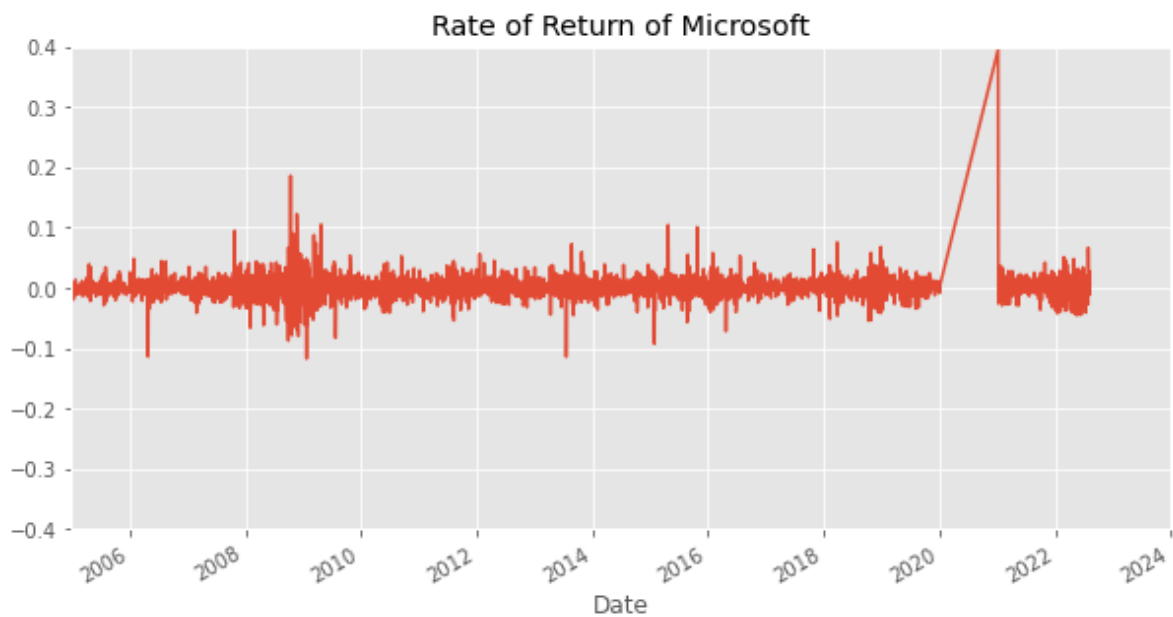
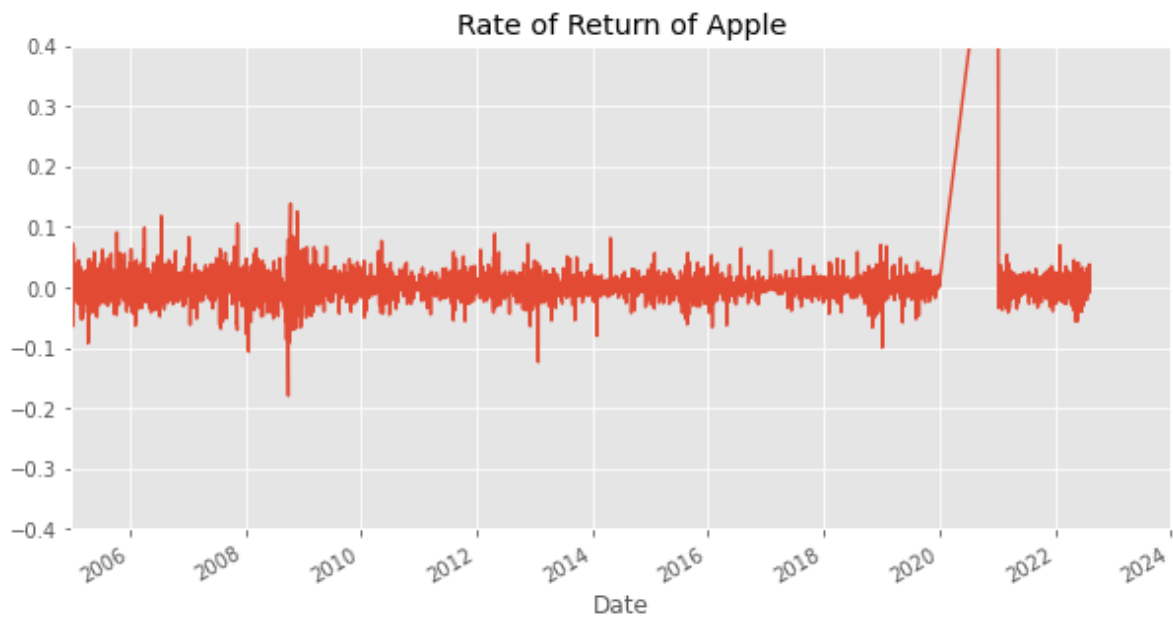
#### Rate of Return (RoR):

- The rate of return is the net gain or loss of an investment over a specified time period, expressed as a percentage of the investment's initial cost.
- The dividends and effects of inflation are not taken into consideration in the simple rate of return calculation.

The formula to calculate the rate of return (RoR) is:

$$\text{Rate of return} = \frac{\text{Current value} - \text{Initial value}}{\text{Initial value}} \times 100$$

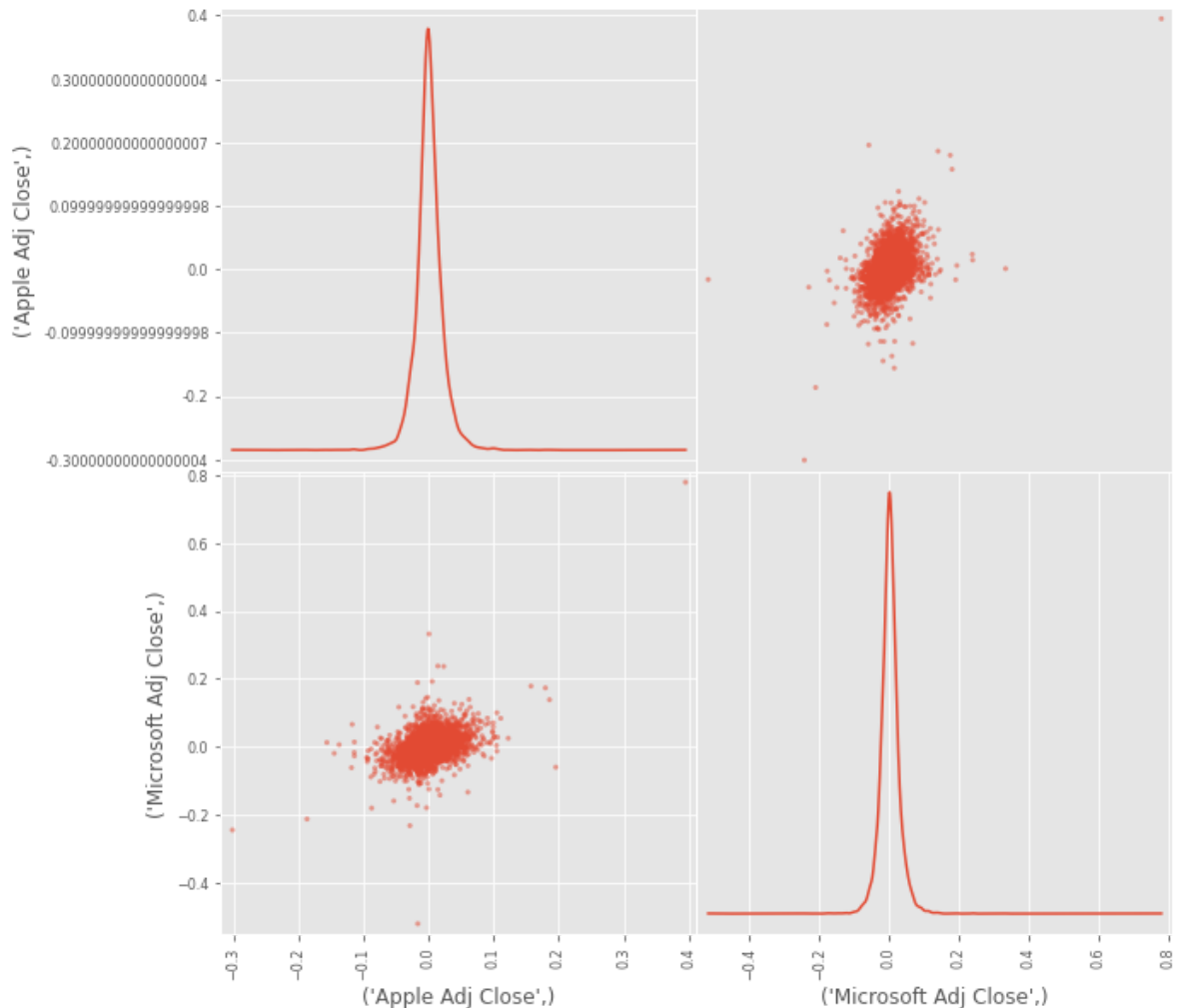
We can see that Apple stock dropped almost 20% in 2008, and Microsoft's return ratio is more stable than Apple.



## 4. Check the correlation between stocks

The scatter matrix shows that the stock prices of Apple and Microsoft are approximately positively correlated.

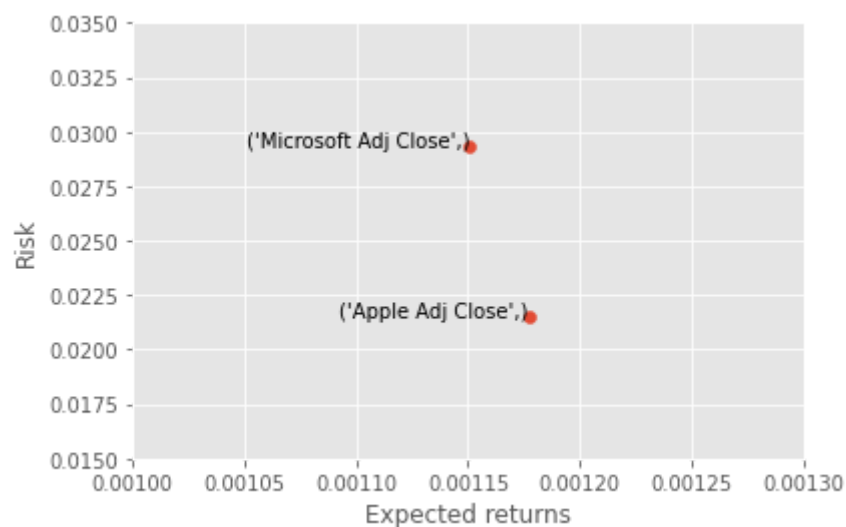
This is because they are both in technology industry and keep growing.



## 5. Check the Expected Returns and Risk

We take the mean of returns as their expected returns and the standard deviation of returns as their risk.

We can see that Apple has more expected return and lower risk than Microsoft.



## 6. Train the models to predict the trend of adjusted price

Adjusted price is useful in analyzing the price of history stock and the trend of it. To gain the profit from stocks, we should buy the stock when it has a low price and sell it out when the price of it is high.

That is to say, if the future trend of a stock is ascending, it's time to buy in and make some money!

So predict the trend of adjusted price helps us to decide which stock is profitable.

Steps:

- Adjust the dataset. Adjusted price is set to be target while others are features.
- Data cleansing and split the data into train data set and test data set.
- Use the dataset to train different kinds of models and find the best out of them.
- Use the best model to predict the future trend of this 2 company relatively.
- Find out the better stock.

## Step 1: Feature Engineering

According to the heat map, some of the features are too related to be used in predict model.

-> As a result, we should develop some new features to feed the model.

Some of the features are listed below.

$$HL\_PCT = \frac{High - Low}{Close} \times 100$$

$$PCT\_change = \frac{Close - Open}{Open} \times 100$$

$$High\_Low = \frac{Hight - Low}{Low} \times 100$$

## Step 2: Data Cleansing && Train\_test\_split

1. drop the null values
2. Set the adj price to be y and other features to be x.
3. X values varies so different that standrization is necessary.
4. Split the data

## Step 3: Choose the regression model and train them

Apple :

| Model                 | Score              |
|-----------------------|--------------------|
| SVR                   | 0.1347301809273994 |
| RandomForestRegressor | 0.9004766402249255 |
| AdaBoostRegressor     | 0.7477720545636599 |

| Model                     | Score               |
|---------------------------|---------------------|
| GradientBoostingRegressor | 0.8956349972426464  |
| LinearRegression          | 0.46750795307588877 |
| DecisionTreeRegressor     | 0.3297451092942806  |
| KNeighborsRegressor       | 0.7410970518931721  |

Microsoft:

| Model                     | Score                |
|---------------------------|----------------------|
| SVR                       | -0.11970083235958495 |
| RandomForestRegressor     | 0.6456887918544316   |
| AdaBoostRegressor         | 0.1932097008238315   |
| GradientBoostingRegressor | 0.6774150367473095   |
| LinearRegression          | 0.1651828307485914   |
| DecisionTreeRegressor     | -0.2360732479219958  |
| KNeighborsRegressor       | 0.318847926671497    |

It seems that if we find the suitable inputs for these models to predict the trend of adj price, the stock price of `Apple` is more predictable than `Microsoft`.

## 7. Try to predict the adj price by date

Because we can't find appropriate inputs to feed the models above to predict the future price, we try to modify this model.

Now the model takes date as input only, so that we can use the model to predict the future price.

Apple Training Model:

| Model                     | Score              |
|---------------------------|--------------------|
| SVR                       | 0.9144209675551294 |
| RandomForestRegressor     | 0.9997331050660618 |
| AdaBoostRegressor         | 0.9998237811047961 |
| GradientBoostingRegressor | 0.9991219486655357 |
| LinearRegression          | 0.6381134088091925 |
| DecisionTreeRegressor     | 0.999521941588286  |
| KNeighborsRegressor       | 0.9998237811047961 |



The best model is : AdaBoostRegressor

Microsoft Training Model :

| Model                     | Score              |
|---------------------------|--------------------|
| SVR                       | 0.886046846266471  |
| RandomForestRegressor     | 0.9998639986573902 |
| AdaBoostRegressor         | 0.9998848066914476 |
| GradientBoostingRegressor | 0.9997137079984763 |
| LinearRegression          | 0.6224787693482164 |
| DecisionTreeRegressor     | 0.9997133475396855 |
| KNeighborsRegressor       | 0.9998848066914476 |

The best model is also : AdaBoostRegressor

Let's use the model to predict new data. Than use the new data to train the model.

-> Keeping iterating this operation until we get the results we want !

## 8. Compare two stock based on the prediction

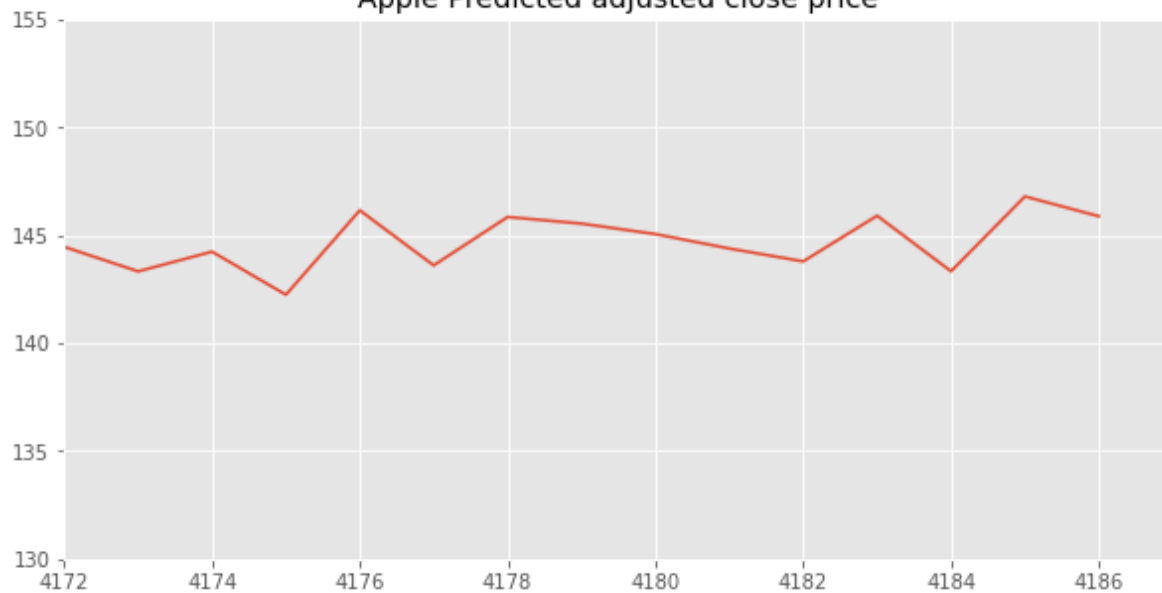
---

We use the model to predict the next 15 days adjusted close, then plot them and their return ratio.

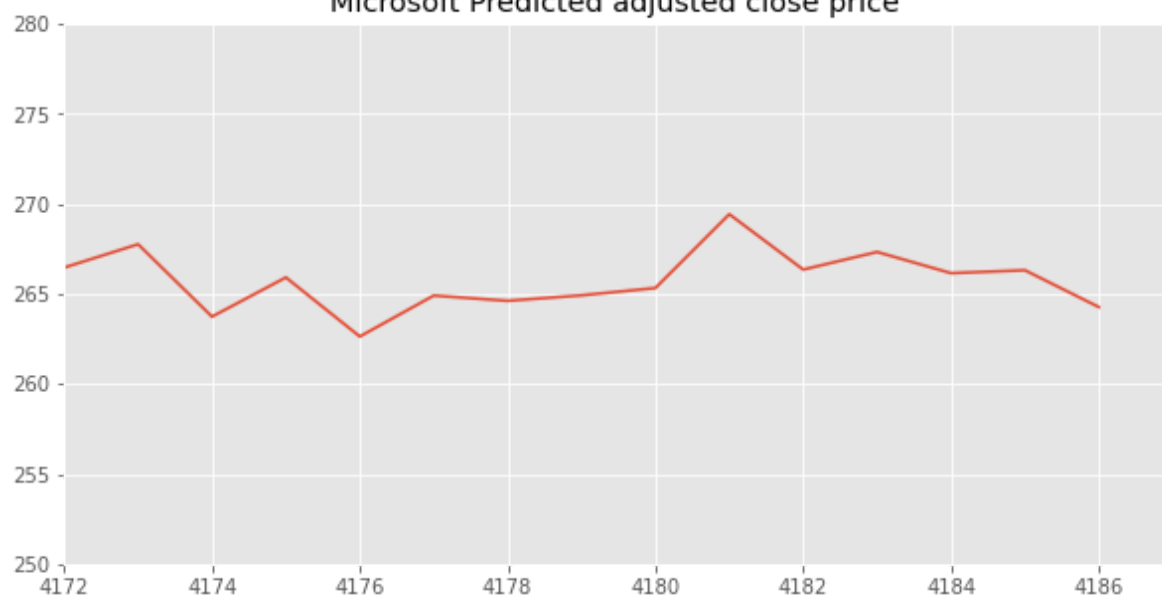
Because in the original data, Microsoft has higher price than Apple, the price of Microsoft is still higher than Apple in our prediction.

On the one hand, the chart shows that sometimes Apple will have a higher rate of return than Microsoft. On the other hand, Apple has more fluctuation than Microsoft.

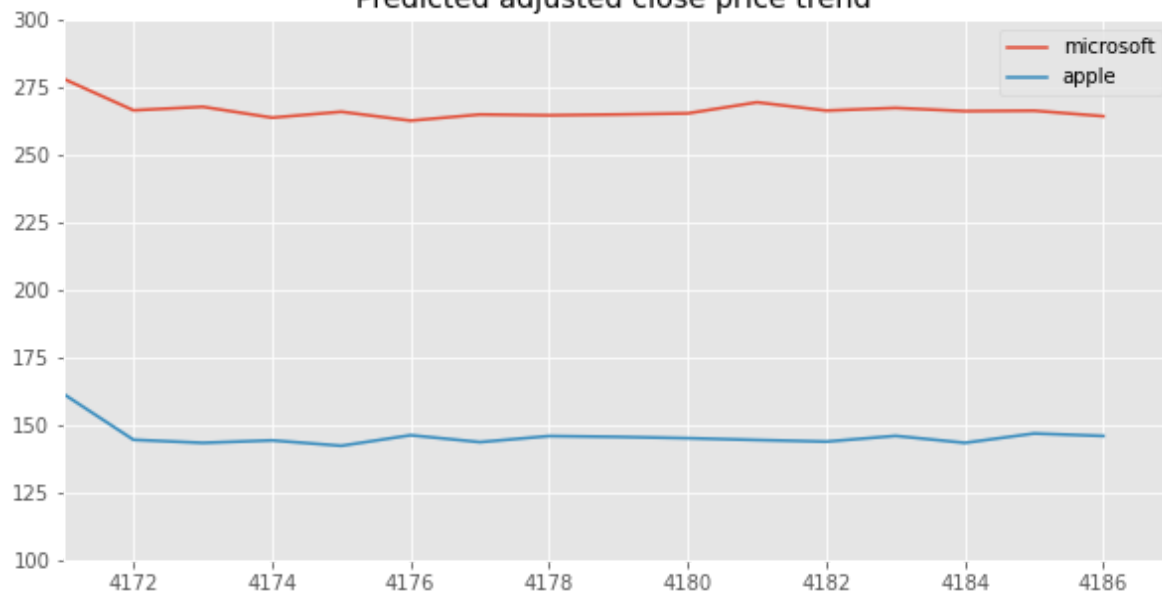
Apple Predicted adjusted close price

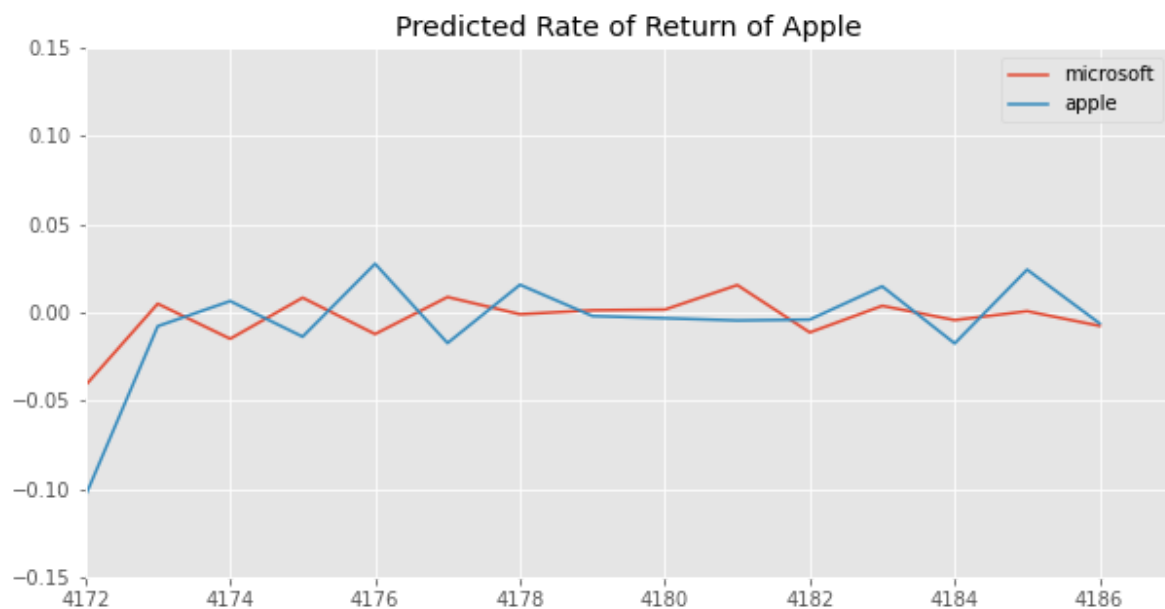


Microsoft Predicted adjusted close price



Predicted adjusted close price trend





## 9. Summary

First, we preprocess the dataset by removing 2020 data. Then we visualize basic information of the two stocks. Finally we use AdaBoostRegressor model to predict the adjusted close price for next 15 days.

-> Overall, we choose Microsoft because it has higher price and more stability.

Since we only use date as the variable to build this simple model and the stock market is complex, the prediction is inaccurate and rough. This model can be improved if we involved other data such as stock dividend or market capitalization.

Thank you for your reading !