# Executive Summary

Our project aims to investigate if ***sympathy play*** is a valid strategy in the stock market under some assumptions.

Sympathy play is the following: for some *highly correlated* companies within the *same sector*, if the stock prices of *all but a few* companies in this sector have substantially risen over a period, then the companies with lower performance will eventually *catch up* and rise later.

## 1. Dataset

We use the *daily closing price* of stocks from several sectors: retail, airline, pharmaceutical, internet, and banks. We add the variable

$$return = \frac{today's \ \ closing\ price}{yesterday's \ \ closing\ price} - 1$$

to the data set.

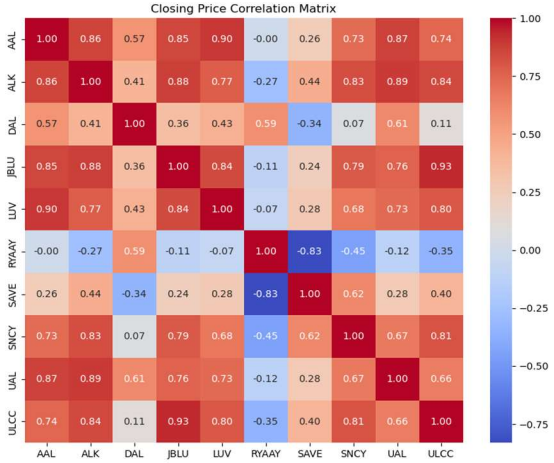
We introduce a new variable, *SP,* which stands for "sympathy play indicator for stock A." as follows. Let

$$SP = \begin{cases} 1 & when\ condition\ 1\ is\ satisfied \\ 0 & otherwise \end{cases}$$

where condition 1 is defined as follows: for the sector containing stock A, at least one stock in the sector increased significantly over the period, and stock A remained flat or fell. Specifically, in this project, "increase significantly" is defined as the daily return is higher than the 75% quantile of all historical daily returns, and "did not move up" is defined as the stock daily return being less than the median of all historical daily returns.

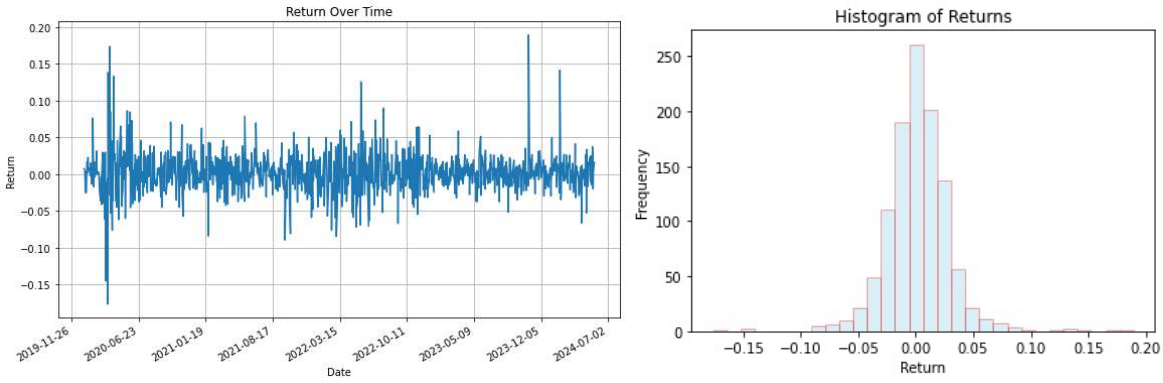
### 2. EDA

We investigate daily closing prices to assess if the companies within the same sector are genuinely correlated. For instance, in the graph below, we give the correlation matrix for some stocks in the airline section.

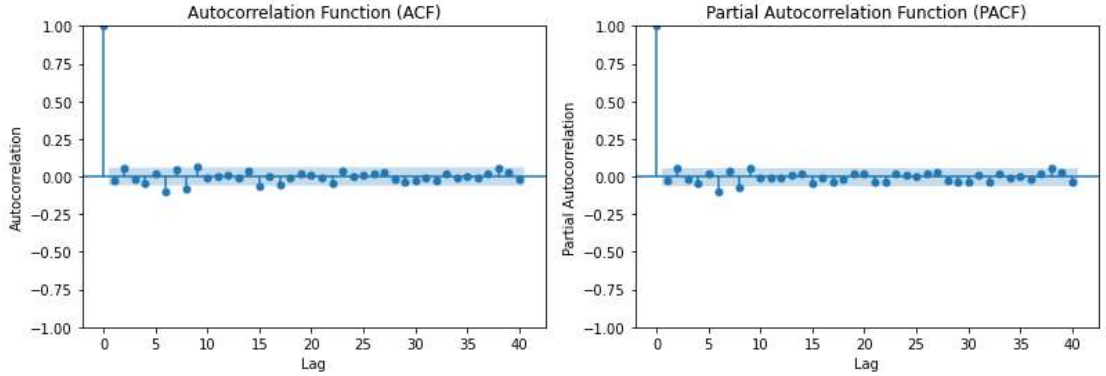


With a few exceptions, the closing prices of most companies are highly correlated. We observe similar results in other sectors.

We use *time series plots* and *histograms* for general investigations. For example, below are the daily return plots for the company DECK from the retail sector




We made a further study time series using the techniques of ACF and PACF:




From the four pictures above, one can see that the daily stock return is roughly a *white noise* with a mean of approximately 0. Based on the graph of ACF and PACF, one may claim there is a very weak seasonality.

### 3. Model selections and results

In this section, we showcase the model selections and results using DECK from the retail sector. Results similar to DECK in retail were observed for other companies and sectors.

Here is an outline of the steps we took:

1. Naïve model
2. Rolling Average
3. Exponential Smoothing
4. ARIMA
5. SRIMAX with sympathy play as a regressor
6. Long Short-Term Memory (LSTM)
    a. LSTM, with yesterday's return
    b. LSTM, with yesterday's return and sympathy play indicator

The parameter choice for ARIMA and SRIMAX are the results of a direct search for a set of parameter values that minimize the model's AIC. The LSTM used 100 epochs with a batch size of 25.

The table below gives the MSE (mean square error) and MAE (mean absolute error) for the above-selected models:

| DECK | | | | |
|---|---|---|---|---|
| | Training | | Test | |
| | MAE | MSE | MAE | MSE |
| Naïve | 0.01974 | 0.00081 | 0.01494 | 0.00041 |
| Rolling average | 0.01849 | 0.00064 | 0.01562 | 0.00045 |
| simple smooth | 0.01983 | 0.00082 | 0.01534 | 0.00043 |
| double smooth | 0.02054 | 0.00086 | 0.01611 | 0.00047 |
| triple smooth | 0.01974 | 0.00081 | 0.01521 | 0.00042 |
| ARIMA(0,0,2) | 0.01979 | 0.00081 | 0.01495 | 0.00042 |
| SRIMAX((0,0,0),(1,0,0,6)) | 0.01988 | 0.0008 | 0.01487 | 0.0004 |
| LSTM | 0.02002 | 0.00082 | 0.01541 | 0.00042 |
| LSTM_SP | 0.01982 | 0.00081 | 0.01511 | 0.00041 |

All models provided similar performance in terms of MAE and MSE, and in fact, the naïve model using training set average as prediction beats most other models.

It is worth noting that despite SRIMAX not providing non-negligible improvements compared to other models, SP as the regressor is significant. This phenomenon can be observed in the following summary:

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| SP | 0.0030 | 0.001 | 2.299 | 0.021 | 0.000 | 0.006 |
| ar.S.L6 | -0.1090 | 0.021 | -5.166 | 0.000 | -0.150 | -0.068 |
| sigma2 | 0.0008 | 1.76e-05 | 45.513 | 0.000 | 0.001 | 0.001 |

## 4. Discussion

In this section, we discuss the above findings. We observed a white noise pattern from the EDA for daily stock increase/decrease. If this is true, different models cannot significantly improve estimation and prediction compared to the naïve model.

In the SRIMAX model, the SP is a significant factor, with a p-value of 0.021. However, the corresponding coefficient being small, it does not capture sufficient variation of the data. This has two interpretations: one possibility is that the sympathy play indicator does explain some variation of the data but at a very small scale, or we observed this correlation by randomness, given our very large dataset (over 1000 observations in the training set). Either way more is needed in terms of prediction.

Surprisingly, the LSTM methods performed slightly worse than the other traditional approaches. However, if the daily return is merely some white noise, then this result is not surprising. After all, neural networks cannot perform a miracle to estimate and forecast something that purely follows a Brownian motion.

## 5. Conclusion and Future Study

This project concluded that the stock price daily return is essentially a Brownian motion, and all models implemented by the team did not provide a significantly better result than the naïve model. The sympathy play indicator appears to be a significant factor in the SRIMAX model, but this provides no improvement in predictions. Including the sympathy play indicator as a factor in the nonlinear model did not provide superior results. *We conclude sympathy play using daily price may not be a valid strategy.*

Two things can be further studied. One can use data per hour or even per second to investigate the phenomenon again. According to the efficient market hypothesis, daily data is unlikely to provide any information that has yet to be priced. Another aspect is that one can discover a better definition of the sympathy play indicator. This might lead to a better estimation and prediction of the given data.