Project Report
Pairs Trading with Kalman Filter

Mentee : Adarsh Prajapati
Mentor : Nimay Shah & Devvrat Patni

December 30, 2024

Contents

1 About the Project:

Pairs Trading is a market-neutral trading strategy (seeks to profit from both increasing and decreasing prices) where opposite positions are taken in two correlated stocks. The buy and sell signals are initiated when the spread deviates from the mean. Two cointegrated stocks have a property of mean reversion, i.e, if the prices are diverging then they will again come closer eventually. The ratio of the stocks which will be traded is called **Hedge Ratio**.

Kalman Filter is an efficient recursive filter used to estimate the state of a *linear dynamic system* from noisy measurements. A linear dynamic system has states' evolution through a linear relation. Kalman Filter is used to find out the hedge ratio and smoothening the stock price time series. The pairs can be further picked to minimise the risk of portfolio while simultaneously maximising the returns. This helps in getting better indicators like *Sharpe Ratio, Sortino Ratio*

2 Weekly Progress:

2.1 Week 1 - The Basics

I learnt the basics of pairs trading and how can it be utilised for generating profits. The first step includes finding a pair of cointegrated stocks using techniques like *Engle-Granger test* and get validation through P-values. The price ratio is used as an indicator to generate buy and sell signal and not the spread. A suitable window (of days) is chosen to get a moving average(rolling mean) and standard deviation of the prices.

Z statistics of the price ratio were used to compare the movement of stock price relative to its rolling mean. If the Z score dropped a threshold negative value, the pair would have to be bought long (buy numerator and sell denominator). If it increased above a positive threshold, then the pair would have to be sold short (sell numerator and buy denominator). Key point to note is whenever a stock is bought, the complimentary has to be sold. This is evident from the image shown below also.

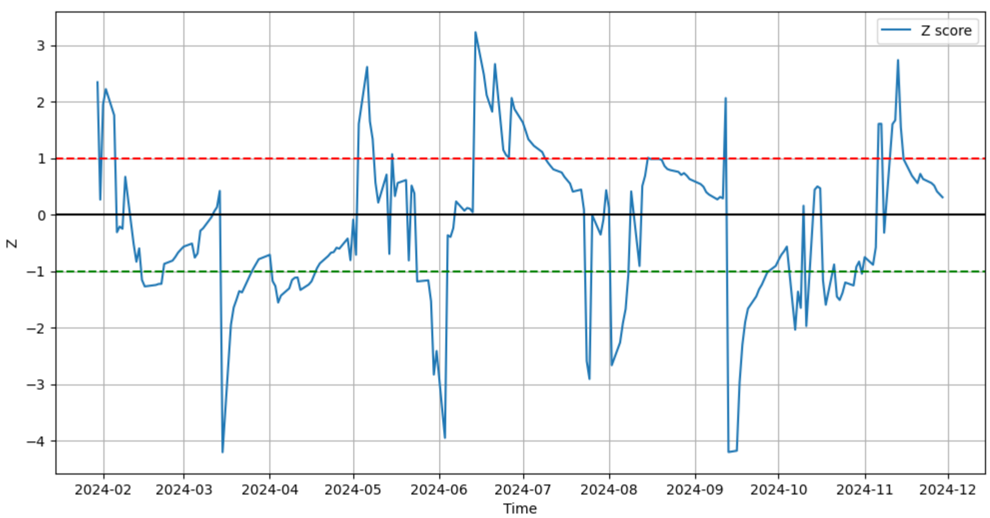

Figure 1: Z score and the threshold value

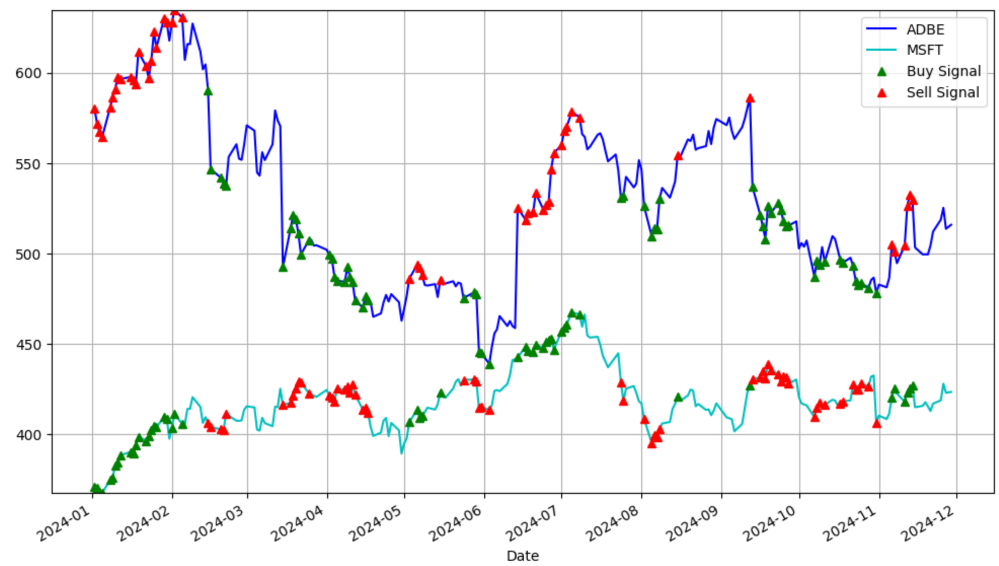

Figure 2: Time series of Adobe and Microsoft containing the Buy and Sell signals

Further optimization in the strategy could be done by finding an optimal window of moving average and the threshold Z score which could be found out by cross-validation.

2.2 Week 2 - Using Kalman Filter

I learnt the usage of Kalman Filter. It is used for predicting the states of a system through noisy measurement. It is a constantly iterating process which will be applied to the application of stock markets to find out an optimal *Hedge Ratio*, which will be used to minimize the spread between the long and short stocks in the pair. Some brief theory about working of Kalman Filter is given below -

1. Prediction Step

The prediction step estimates the current state and its uncertainty based on the previous state.

$$\hat{x}_k^- = F_k \hat{x}_{k-1} + B_k u_k \qquad \text{(Predicted State)} \tag{1}$$
$$P_k^- = F_k P_{k-1} F_k^T + Q_k \qquad \text{(Predicted Covariance)} \tag{2}$$

2. Update Step

The update step corrects the prediction using the new measurement.

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} \qquad \text{(Kalman Gain)} \tag{3}$$
$$\hat{x}_k = \hat{x}_k^- + K_k (z_k - H_k \hat{x}_k^-) \qquad \text{(Updated State)} \tag{4}$$
$$P_k = (I - K_k H_k) P_k^- \qquad \text{(Updated Covariance)} \tag{5}$$

3. Measurement Step

The measurement step calculates the difference between the actual measurement and the predicted measurement, known as the residual or innovation.

$$v_k = z_k - H_k\hat{x}_k^- \qquad \text{(Measurement Residual)} \qquad (6)$$

$$S_k = H_k P_k^- H_k^T + R_k \qquad \text{(Residual Covariance)} \qquad (7)$$

State Representation

The state, denoted as x_k, represents the set of variables required to describe the system at time step k. For example:

- In a tracking system, x_k may represent position and velocity.
- In an economic model, it could represent prices and growth rates.
- In our case, it represents the state of hedge ratio and a margin price

Explanation of Symbols

- $\hat{x}_k$: Estimated state at time k.
- P_k: Covariance matrix representing uncertainty in state variables.
- F_k: State transition matrix.
- B_k: Control input model.
- u_k: Control input like an external force.
- Q_k: Process noise covariance.
- H_k: Observation model.
- z_k: Measurement at time k.
- R_k: Measurement noise covariance.
- K_k: Kalman Gain, determining the weight given to measurements.
- v_k: Measurement residual.
- S_k: Residual covariance.
- I: Identity matrix.

A simple explanation about Kalman Filter can be given through the following image. The states are predicted using the State Transition Matrix, the resulting distribution is multiplied with the noisy state measurements to get the final state mean and covariance.

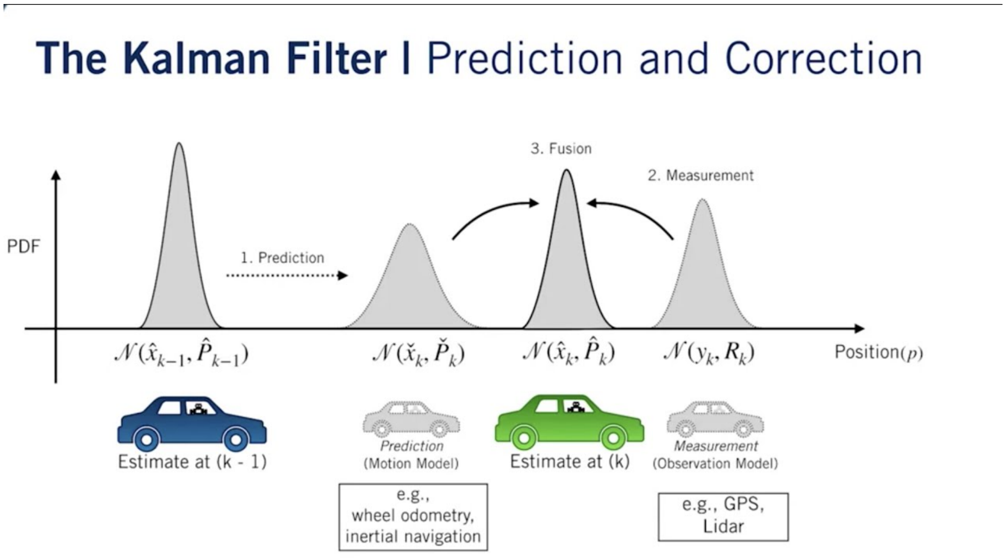

Figure 3: Simple Explanation of Kalman Filter based on car speed example. Reference - Youtube

Kalman Filter can also be used to find a moving average of any time series where the window length is not required giving a better smooth signal. This smooth signal is fed to *KalmanFilterRegression()* model, to get the hedge ratio state.

I learnt that *backtesting* is an essential step to validate the models that we create, testing over historical data. I also learnt various indicators to describe the portfolio like *Max Drawdown, Sharpe Ratio, Sortino Ratio, Kurtosis.*

The assignment required to consider a transaction cost of 0.12% per order and then calculate the cumulative profit.

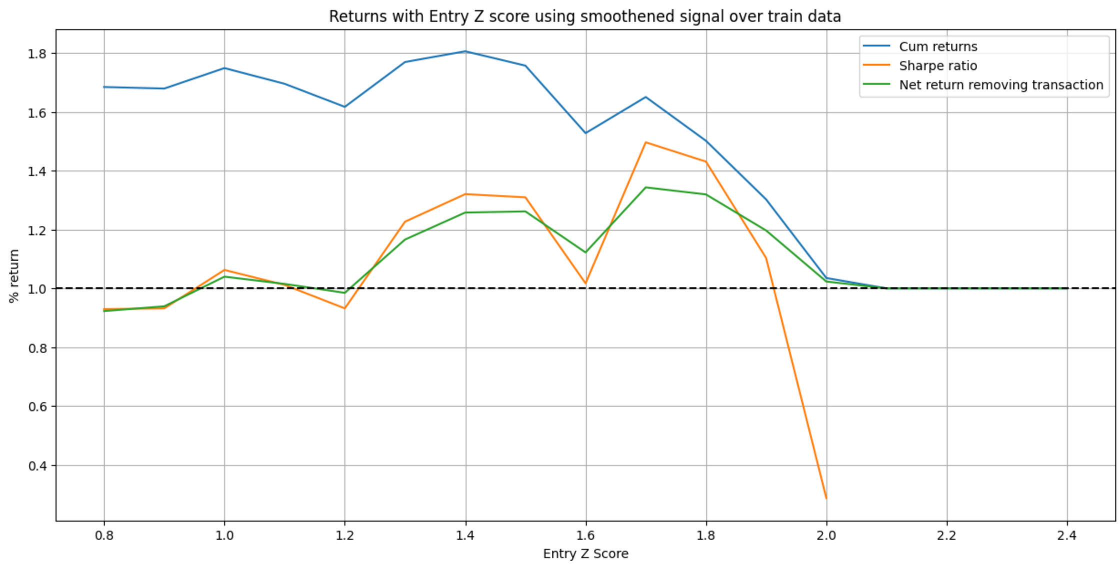

Figure 4: Profit and Sharpe Ratio considering the transaction cost varying with entry Z score

2.2.1 Sidetrack

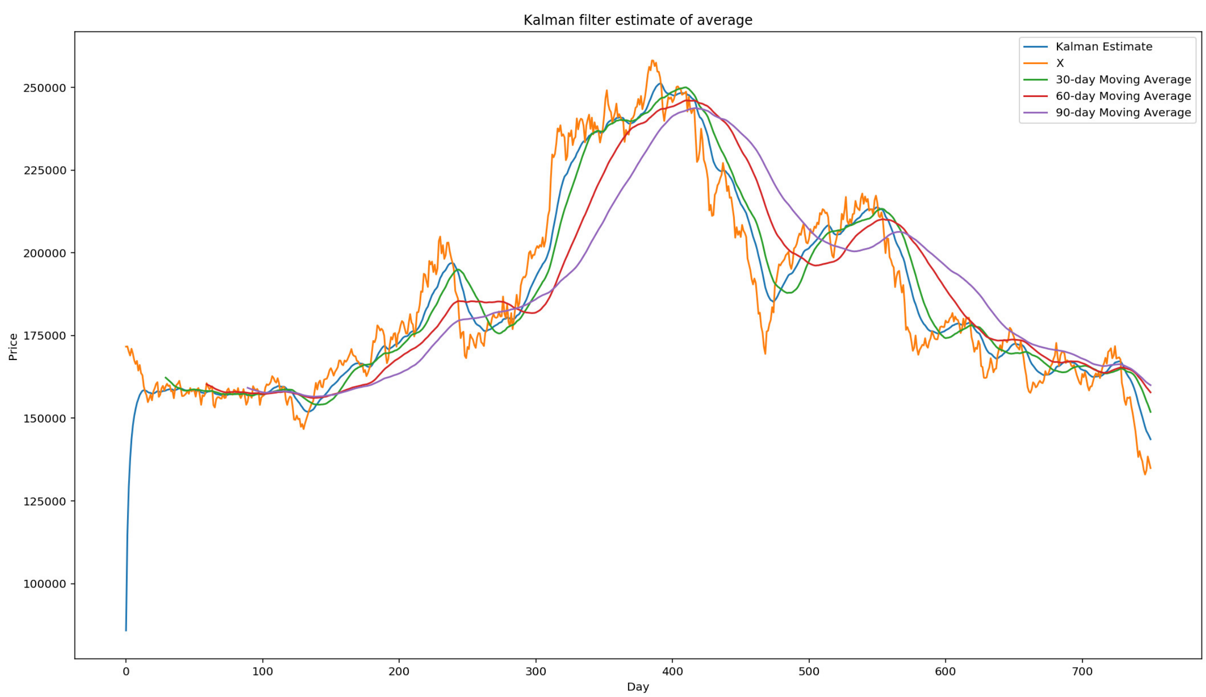

Figure 5: How Kalman Average is different than moving average

I also tried to find out the outcome of passing the unsmoothened signal in the Kalman Regression function, and was getting zero output. This was because the half life of the ADBE-MSFT spread for unsmoothened signal was equal to 1 day. This led to Z score equalling to 0, as the mean equals the current value.

The spread at time t can be related to the lagged spread at time $t-1$ as:

$$\text{Spread}_t = \text{Spread}_0 + \text{Spread}_{t-1}$$

The spread return (change in spread) is:

$$\text{Spread}_{\text{ret},t} = \text{Spread}_t - \text{Spread}_{t-1}$$

We model the relationship between the spread return and the lagged spread using an Ordinary Least Squares (OLS) regression:

$$\text{Spread}_{\text{ret},t} = \alpha + \beta \cdot \text{Spread}_{t-1} + \epsilon_t$$

Where:

- α is the intercept term,
- β is the coefficient of the lagged spread (decay rate),
- ϵ_t is the error term.

The half-life of the spread's decay is:

$$t_{1/2} = \frac{-\ln(2)}{\beta}$$

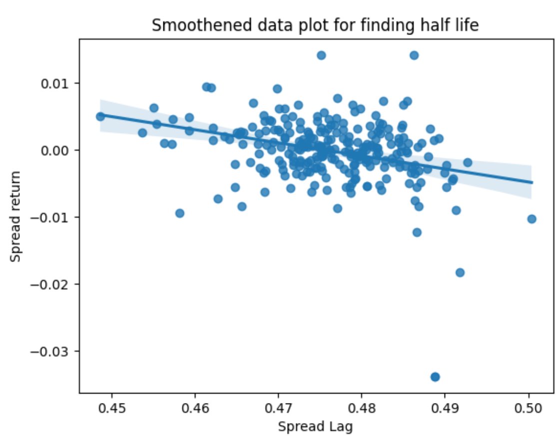

Figure 6: Halflife = 4 days

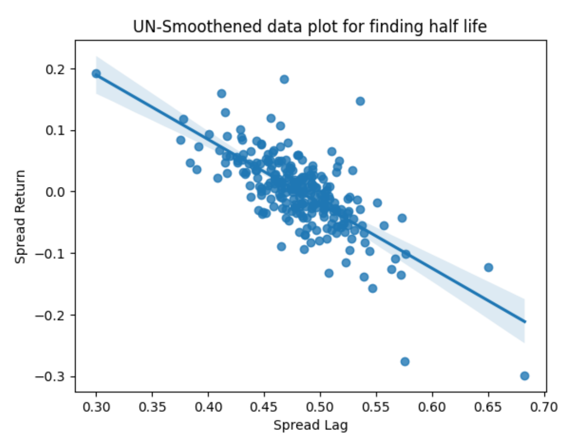

Figure 7: Halflife = 1 day

2.3 Week 3 - Managing Portfolio Risk

Risk management in trading is essential for minimizing losses and ensuring long-term profitability. Key components include **stop-loss orders**, which automatically exit a trade if the spread between two assets exceeds a predefined threshold, protecting capital from adverse movements. **Position sizing** helps determine how much capital to allocate to each trade based on volatility, ensuring no single trade can significantly impact the overall portfolio. **Diversification** across multiple asset pairs reduces the risk of large losses by spreading investments, while portfolio variance measures the risk contribution from each asset. Additionally, risk-adjusted return metrics such as the Sharpe and Sortino ratios assess the effectiveness of trading strategies by factoring in the associated risks.

Volatility is a key measure of risk in trading, often quantified through standard deviation or variance of returns. Higher volatility suggests greater potential for both gains and losses, making it crucial for traders to account for this risk. Historical volatility reflects the past price fluctuations of an asset, while implied volatility gauges the market's expectation of future price movement, often inferred from options prices. Traders use volatility to make informed decisions on position sizing and risk management. A **variance threshold strategy** can also be used, adjusting trading behavior based on changes in volatility to exploit high-volatility periods and avoid low-volatility conditions.

Rolling volatility provides a dynamic, time-sensitive measure of risk by calculating volatility over a moving window, allowing traders to adapt to changing market conditions. This method is useful for determining the optimal times to enter or exit positions. In conclusion, effective risk management in pairs trading involves a combination of strategies such as stop-loss orders, position sizing, diversification, and volatility analysis. By integrating these approaches, traders can mitigate risks, adjust to market fluctuations, and optimize risk-return outcomes for consistent profitability.

3 The Final Task

The final task included the following tasks. I needed to extend the basic pairs trading strategy by managing a portfolio of multiple cointegrated pairs, aiming to diversify risk and improve performance. The Kalman filter is used to dynamically adjust hedge ratios for each pair, optimizing risk exposure. The process involves identifying cointegrated pairs, applying the Kalman filter to estimate hedge ratios, building a diversified portfolio with risk-based weight allocations, and backtesting the multi-pair strategy to evaluate its performance compared to a single-pair approach.

3.1 Choosing Different Sectors

A key player to diversify a portfolio is to pick stocks of different sectors for which I chose FMCG, Tech, Pharma, and Banking.

After choosing the top performing 15 stocks data from 2010 to 2018 in all these sectors, I removed the stocks which were unlisted in the beginning.

There is a huge chance that stocks of similar companies would have cointegration so to be used for Pairs Trading. I found the cointegration matrix for the same and picked the pairs with p value < 0.05.

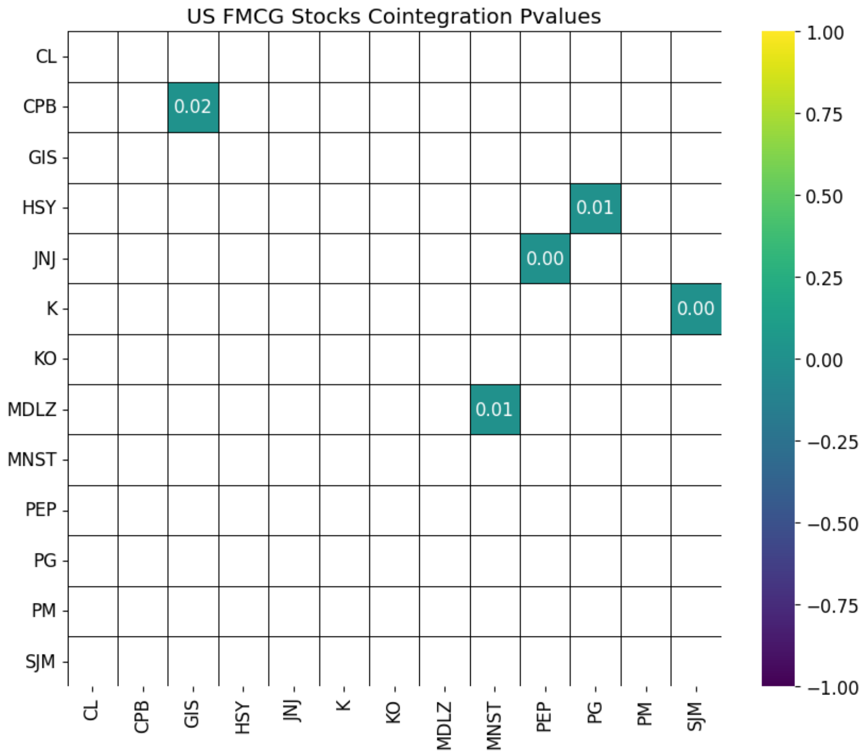

Figure 8: Cointegration Matrix with mask of Pvalue < 0.05

This allowed me to find cointegrated stocks in all the 4 sectors.

3.2 Calculating the Spread and applying Variance Thresholding

For the cointegrated pairs, I found out the hedge ratio and spread using *KalmanFilterRegression()* and *KalmanFilterAverage()*. Variance Thresholding was applied to find two of the most volatile stocks which could generate high returns in each sectors.

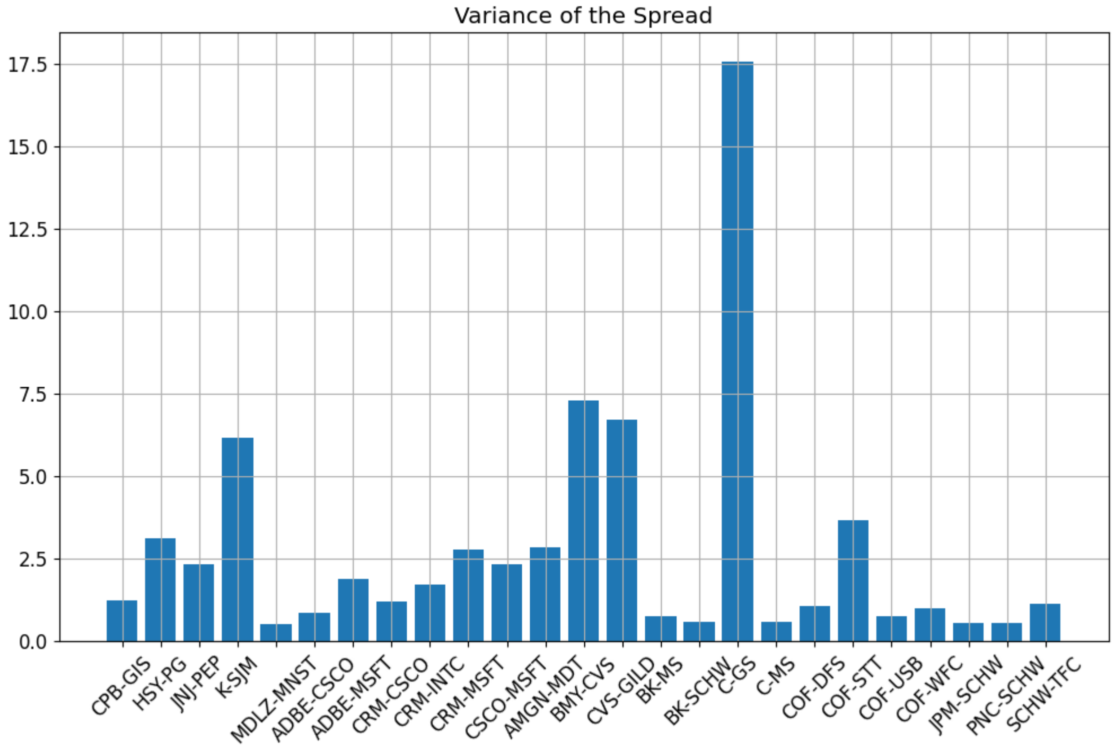

Figure 9: Variance of various pairs of stocks

3.3 Minimising the Covariance of the Portfolio

The final portfolio has to be made using 1 stock of four sectors each. The stocks and their weights have to be decided such that the portfolio covariance is minimum.

For the optimization, the SLSQP (Sequential Least Squares Quadratic Programming) method was used. The optimization algorithm finds the set of asset weights w that minimizes the portfolio variance while satisfying the constraints. The algorithm iteratively adjusts the weights, evaluating the objective function $w^T \sum w$ until it converges to the optimal solution. Here, σ refers to the covariance matrix of all the assets.

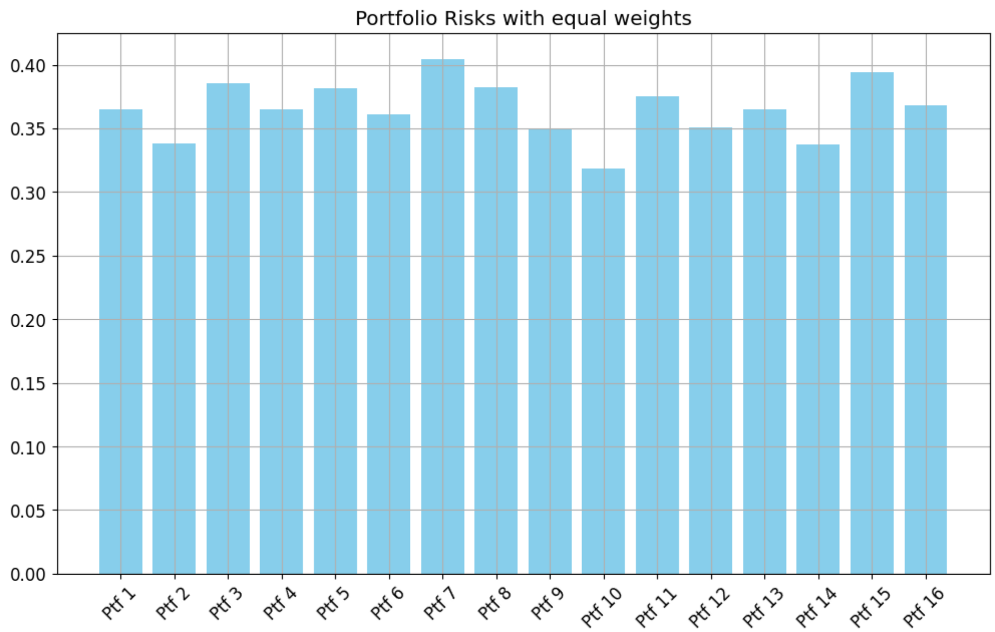

Figure 10: Risk (STD) of various Portfolios without weight optimization

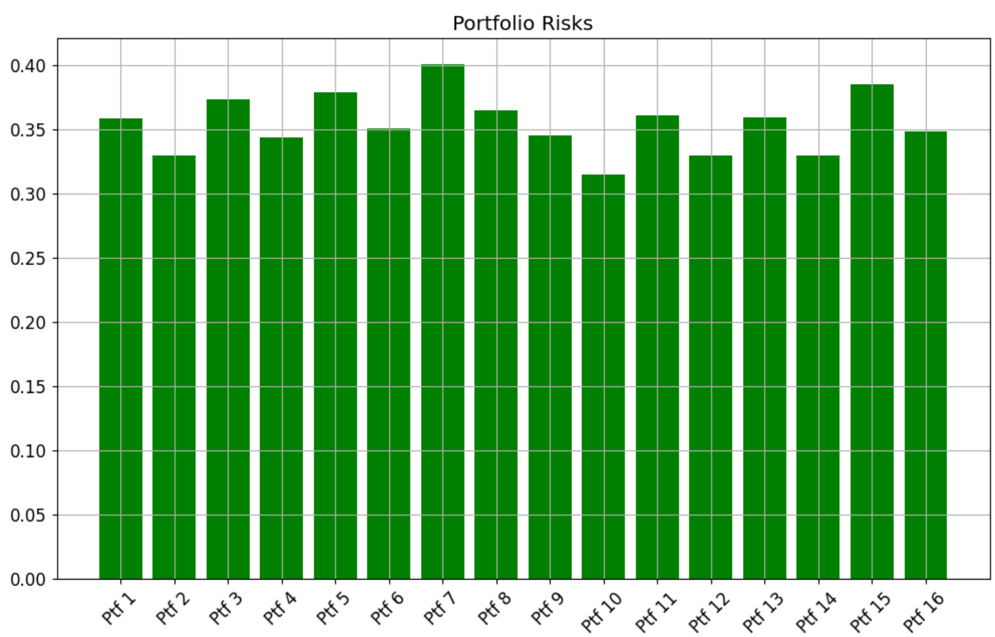

Figure 11: Risk (STD) of various Portfolios after weight optimization

There is not much difference between the risks assuming equal weights or optimizing the weights.

Portfolio 10 has minimum risk. The portfolio contains - [**'HSY-PG', 'C-GS', 'BMY-CVS', 'CSCO-MSFT'**]

3.4 Backtesting

The final step was to backtest the strategy and find the cumulative return.

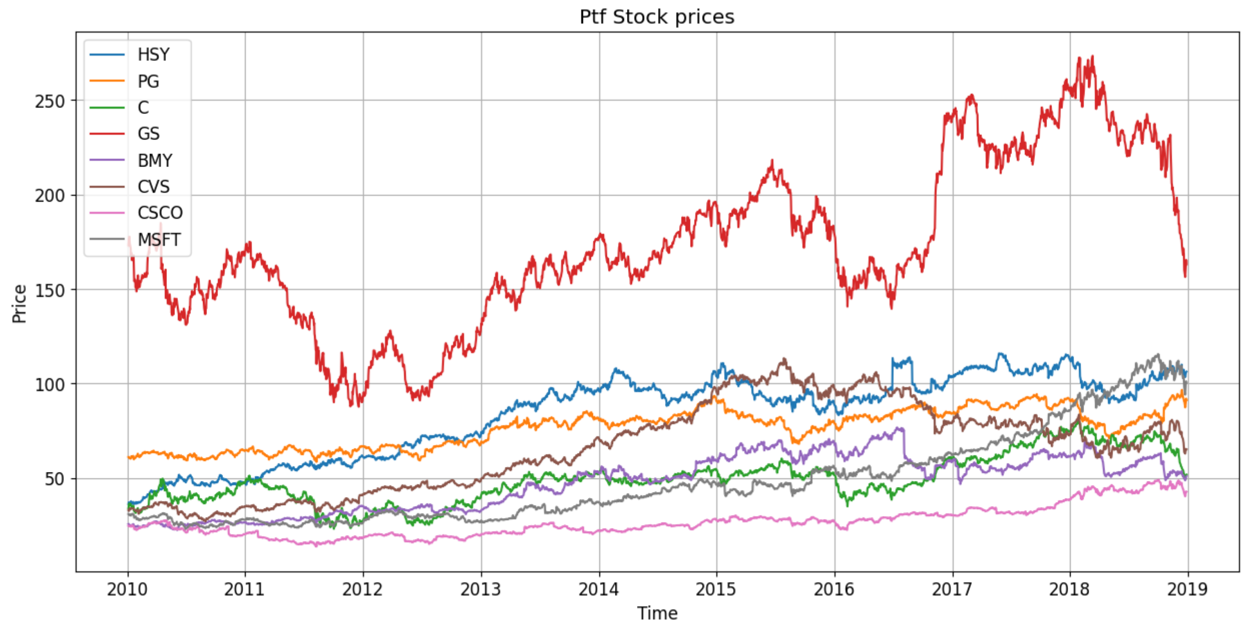

Figure 12: Stock prices varying overtime of the Portfolio

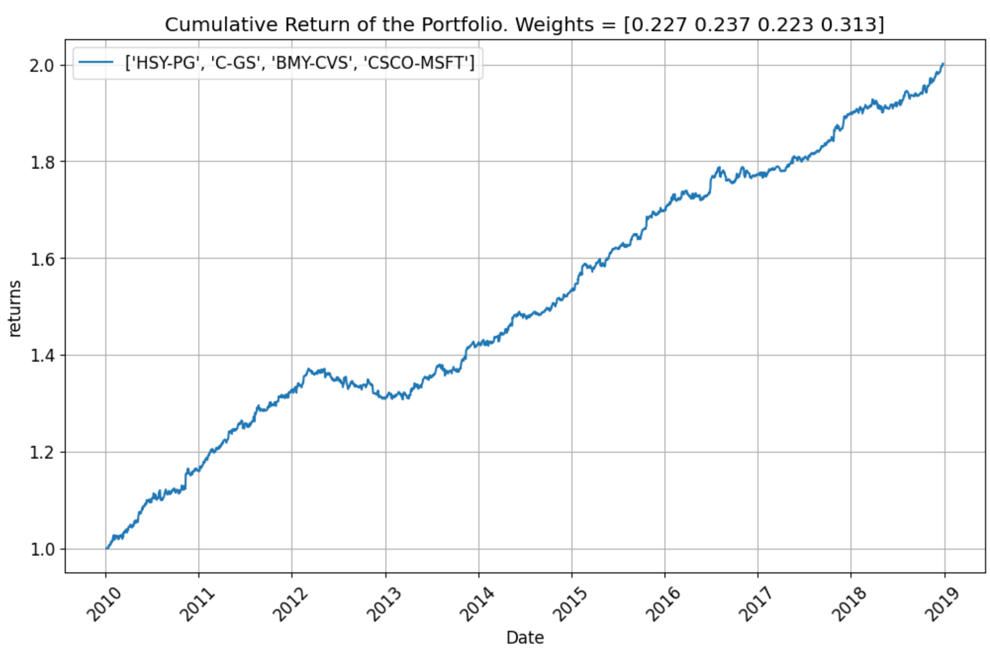

Figure 13: Cumulative Returns of the Portfolio