Indian Institute of Technology Kanpur

Department of Mathematics and Statistics

MTH442A Course Project on

”Pairs Trading: A Strategy Exploration”

Under the Supervision of:

Dr. Amit Mitra

Submitted by:

Aditya Mishra (221256)
Amitesh Singh (221267)
Kaushik Thakkar (221329)
Ranjit Prasad (221387)
Shruti Bindra (221422)

Acknowledgement

We would like to extend our gratitude towards Dr. Amit Mitra for providing us with this opportunity to work on this project for his valuable guidance and feedback which were extremely helpful in the completion of this project.

We would also like to thank everyone who contributed in this project for their collaborative efforts, stimulating discussions, and their assistance, which have significantly contributed to the completion of this project.

Contents

1 Abstract

Pairs Trading is one of the arbitrage techniques that is used for trading in stock market. Here stock pairs are identified which are similar and whose prices tend to move together as suggested by historical data. Any short term mispricing between these stocks is seen as an opportunity for trading and we take on Long and Short positions. We buy the undervalued stock and sell the overvalued stock with the expectation that the mispricing would correct itself and the undervalued stock will become expensive and the overvalued stock will become cheaper. Once the stock prices are back to their long-term relationship we reverse the positions and close the trade. This helps us to make profit from the relative mispricing of the pair.

The stock pairs are identified using the idea of co-integration. Often Stocks belonging to same or related sector have prices that move together in the long run because of similar exposure to risk or some common underlying economic factor influencing them. We identify this relationship using co-integration and hence try to develop a trading strategy to maximize our profits based on the co- integrated stocks.

***Aim of this project** is to identify potentially co-integrated stocks from the major sectors of the economy and to build a trading strategy based on moving averages techniques. Additionally we aim to find an optimized threshold value for generating trading signal that would maximize our profit. We also test our trading strategy and estimate the potential profit.*

2 Introduction

This section contains answers to all the questions that needs to be answered before starting with the implementation of Pairs Trading Strategy.

2.1 What is Stationarity?

If $\{X_t : t \in T\}$ is a Stochastic Process such that **statistical properties** of the process does not change over time i.e. the physical process is in a "steady state of statistical equilibrium". By stationarity here we mean covariance or weak stationarity of a time series.

A time series is stationary in weak sense if: It's mean function is finite and time invariant; Variance function is finite and time invariant; Covariance function depends only on time lag.

2.2 What are Integrated Processes and Order of Integration?

An integrated process refers to a time series that requires differencing to become stationary and is characterized by it's order of integration. The order of integration is the minimum number of differences needed to get an stationary series.

$\{X_t\}$ is integrated of order "d" if d is the smallest integer such that $\Delta^d X_t$ is a stationary process. Some Examples: of $I(1)$ *Series* : Price, yields, Exchange rates and Examples of $I(0)$ *Series*: Returns of assets.

$I(0)$ is what in which we are interested as $I(0)$ is weak-sense stationary. The time invariant mean of $I(0)$ series forms the basis of mean reversion trading strategy.

2.3 What is Co-integration?

A $(n * 1)$ vector time series $\{Y_t\}$ is said to be co-integrated if each of the series taken individually is $I(1)$ i.e. Integrated of order 1 while some Linear Combination of series $\alpha^T Y_t$ is $I(0)$ i.e. Integrated of order 0 or stationary for some non-zero $(n * 1)$ vector α.

It implies there is some long run equilibrium relation tying the individual components of Y_t together, represented as the Linear Combination $\alpha^T Y_t$. In other words we say that two stocks are co-integrated if their linear combination has a constant mean and standard deviation.

2.4 How to identify Pairs for Pairs Trading?

The pairs are identified using the idea of co-integration. Two Stocks which are co-integrated are treated as pairs in our trading strategy. The test applied to identify pairs is explained in detail in Section 5.2.

2.5 What is Mean reversion strategy?

Mean reversion trading strategy as the name suggests is based on idea that asset prices tends to revert to their historical averages. This is the exact idea behind pairs trading. Here we identified a pair of related stocks using co-integration and we would trade with an expectation that the spread of prices between these stocks would revert back to it's historical average. We put on positions when the spread diverges from it's mean (which is determined by historical data) and close the trade when the spread approaches it's mean value.

2.6 What is Pairs Trading?

Pairs Trading is a Statistical Arbitrage technique based on relative pricing between stocks. The basic idea behind this is that securities with similar characteristics should be priced similarly.

Any short term mispricing between these stocks is seen as an opportunity for trading and we take on Long and Short positions. We buy the undervalued stock and sell the overvalued stock with the expectation that the mispricing would correct itself and the undervalued stock will become expensive and the overvalued stock will become cheaper. Once the stock prices are back to their long-term relationship we reverse the positions and close the trade.

2.6.1 Why Pairs Trading?

a) **Cash Neutral** To implement pairs trading strategy we do not have to invest a lot of money as we use money received from short selling one stock to purchase the other stock. We only need some initial capital along with the fixed amount that we need to keep with the exchange for short selling stocks.

b) **Market Neutral** The performance of the overall market does not affect the trading strategy and profits because let's say the overall market goes up by 10% then the 10% loss in the short stock is compensated by 10% profit in the long stock and vice versa if the overall market goes down.

2.7 Why Co-integration and not Correlation?

Co-integration describes a long term relationship between two assets which might not be correlated in the short term. Whereas, Correlation measures the degree of linear association between two assets in a given time period. As for pairs trading we are interested in the co-movement of two stock prices in the long run so, we use co-integration rather than correlation. Mathematically, we have shown below that two co-integrated stocks may have a low correlation. Whereas Figure 1 below shows two stocks which has high positive correlation but are not co-integrated.

Let $\{X_t\}$ and $\{Y_t\}$ be two time series defined by:

$$X_t = W_t + \epsilon_{x,t}$$
$$Y_t = W_t + \epsilon_{y,t}$$
$$W_t = W_{t-1} + \epsilon_t$$

Here:

- W_t is a common component(random walk) shared by both time series,
- $\epsilon_{x,t}$, $\epsilon_{y,t}$ and ϵ_t are independent random noise terms associated with X_t and Y_t respectively with mean 0 and variance $\sigma_x^2 \, and \, \sigma_y^2 \, and \, \sigma^2$,
- X_t and Y_t are I(1) processes in such a way that the linear combination of X_t and Y_t becomes stationary i.e. $Z_t = [X_t, Y_t]$ is co-integrated with co-integration vector (-1, 1) and

  $\mathrm{var}(\Delta X_t) = var(X_t - X_{t-1}) = \sigma^2 + 2\sigma_x^2$
  $\mathrm{var}(\Delta Y_t) = var(Y_t - Y_{t-1}) = \sigma^2 + 2\sigma_y^2$
  $\mathrm{cov}(\Delta X_t, \Delta Y_t) = cov(\epsilon_t) = \sigma^2$

  $\mathrm{corr}(\Delta X_t, \Delta Y_t) = \frac{\mathrm{cov}(\Delta X_t, \Delta Y_t)}{\sqrt{\mathrm{var}(\Delta X_t) \cdot \mathrm{var}(\Delta Y_t)}} = \frac{\sigma^2}{\sqrt{(\sigma^2 + 2\sigma_x^2)(\sigma^2 + 2\sigma_y^2)}}$

  If $\mathrm{var}(\epsilon_{x,t})$ and/or $\mathrm{var}(\epsilon_{y,t})$ are much larger than the variance ϵ_t, the correlation would be low while X_t and Y_t are co-integrated.

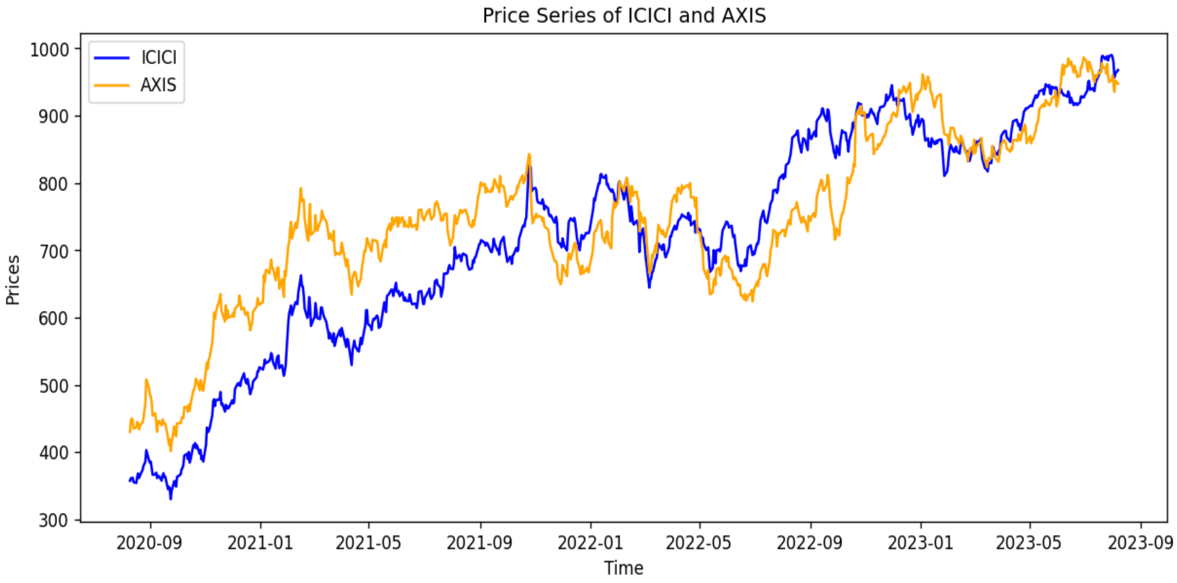

Figure 1

3 An example to illustrate the Pairs Trading approach

We consider two stocks viz. S1 (price = ₹8x/share) and S2 (price = ₹x/share) which are co-integrated and let's say on an average the price of S1 is 8 times the price of S2. This ratio is determined based on this historical data.

Suppose at time t_0 the price of S1 is ₹215.00 and price of S2 is ₹26.00

Now ₹26 $*$ 8 = ₹208 So we see there is a mispricing here. Either S1 is relatively overpriced or S2 is relatively underpriced.

Hence we have a trading opportunity. We expect that in future S1 would become cheaper or S2 would become more expensive based on historical data.

Action at t_0: Sell one unit of S1 and buy 8 units of S2.
Net money we have: ₹215−₹26 $*$ 8 = ₹7

Stocks currently in portfolio: 8 Stocks of S2 and -1 stock of S1
Suppose at time t_1 the price of S1 is ₹215.00 and price of S2 is ₹26.875

Now ₹26.875 $*$ 8 = ₹215 So the mispricing has corrected itself and it's time to close the trade.

Action at t_1: Buy one share of S1 and sell 8 shares of S2 to neutralize our portfolio. (This does not result in extra gain or loss since the amount obtained by selling the stock is used to buy other stock)

Thus the total profit made here is ₹7. Now if this amount is greater than the transaction cost incurred to trade the securities as above we would book the remaining amount as our profit.
Based on this logic a more elaborate explanation of the pairs trading strategy is given in sections 4.3

Conclusion: This strategy does not rely on the true market values of the two stocks (which is very difficult to determine). Rather it is based on the relative pricing of the two co-integrated stocks based on the expectation that the relationship between these two stocks would still continue to hold as suggested by historical data.

4 Description of the Data set

We have collected data from finance.yahoo.com from 10 August 2020 - 07 August 2023 on the following given Sectors :

Banking - 'HDFCBANK.NS', 'ICICIBANK.NS', 'SBIN.NS', 'AXISBANK.NS
Oil - Gas - 'RELIANCE.NS', 'ONGC.NS','IOC.NS', 'GAIL.NS', 'BPCL.NS'
Nifty50 -'NSEI','ADANIPORTS.NS', 'ASIANPAINT.NS', 'BAJAJ-FINANCE.NS'
Automobile - 'MARUTI.NS', 'M&M.NS', 'BAJAJ-AUTO.NS', 'EICHERMOT.NS'
Fmcg - 'HINDUNILVR.NS', 'NESTLEIND.NS', 'GODREJCP.NS', 'MARICO.NS'

Similarly other sectors are as follows- **Telecom, Entertainment, Pharmaceutical, Consumer Goods, Cement, Power, Metail Mining, Healthcare, Infrastructure, Retail, Transportation - Logistic, IT, Sensex 30**

5 Methodology

This section includes the complete workflow of our project starting from scratch up to generating trade signals. It includes the methods used and all the test employed in this project.

5.1 Test for Stationarity

In order to fit any suitable model to our time series data firstly we need to ensure whether our sample data is stationary or not. The stationarity test we use is based on test for a unit root in a time series sample. Presence of unit root indicates that our time data is non-stationary. Generally there are two main test used for this purpose i.e, **Dickey-Fuller test** and **Augmented-Dickey Fuller test** where later is just an augmented version of former including autoregressive terms and lag terms.

Null Hypothesis(H_0): time series data has a unit root.
Alternative Hypothesis(H_a): time series data doesn't have unit root.

The test statistic for ADF is calculated as:

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

where $\hat{\gamma}$ is the estimated value of the unit root, i.e., γ

If the calculated test statistic is less than critical value then null hypothesis is rejected and we can say that our data is stationary.

5.2 Test for Co-integration

In order to test for co-integration between two stock prices, we use **Engle-Granger** test for co-integration developed by **R.F. Engle** and **C.W.F Granger**. This approach involves testing the null hypothesis that there is no co-integration between the two series vs the alternate hypothesis that there is co-integration between the two series.

Consider a vector process Y_t consisting all the stock prices. Before testing for co-integration between the stock prices, we test whether both the series of stock prices are integrated of order 1. This vector process is said to be co-integrated if there exists a linear combination $a^T y_t$ which is stationary.

The Engle Granger test involves two steps:

Step 1: Estimate 'a' using the usual ordinary least square regression.

Step 2: Testing whether the residuals give a stationary series using Dickey-Fuller test. We test for the residuals because if they are stationary, a'y_t would be stationary implying that y_t is co-integrated with co-integrating vector a.
In a pairs trading approach, consider two stock prices u_t and v_t such that the y_t = (u_t,v_t). We want u_t and v_t to be co-integrated such that the spread ϵ_t = v_t - αu_t oscillates around zero. We have the co-integrating vector a = (-α,1). A stationary process has a constant expectation but it is not necessarily zero mean. We can make it zero mean by adding an intercept term α_0 so that the spread becomes ϵ_t = v_t - αu_t - α_0.

To test for co-integration for our data, we used coint() function in python which uses augmented Engle-Granger test to check for co-integration. This function is defined as:
statsmodels.tsa.stattools.coint(y_0, y_1, trend='c', method='aeg', maxlag=None, autolag='aic', return_results=None) where,

- y_0 and y_1 (array-like) are the two stock prices which are to be checked for co-integration
- trend (str) which can take values "c" or "ct"
  - "c" : constant.
  - "ct" : constant and linear trend

  - also available as quadratic trend “ctt” and no constant “n”

- method“aeg” - Only “aeg” (augmented Engle-Granger) is available.

- maxlag(None or int) - Argument for adfuller, largest or given number of lags.

- autolag(str) - Argument for adfuller(function for augmemted Dickey-Fuller test), lag selection criterion:

  - If None, then maxlag lags are used without lag search.
  - If “AIC” (default) or “BIC”, then the number of lags is chosen to minimize the corresponding information criterion.
  - “t-stat” based choice of maxlag. Starts with maxlag and drops a lag until the t-statistic on the last lag length is significant using a 5

- return_results (bool) - For future compatibility, currently only tuple available. If True, then a results instance is returned. Otherwise, a tuple with the test outcome is returned. Set return_results=False to avoid future changes in return.

The coint() function returns the following values:

- coint_t (float) - The t-statistic of unit-root test on residuals.

- pvalue (float) - MacKinnon”s approximate, asymptotic p-value based on MacKinnon (1994).

- crit_value (dict) - Critical values for the test statistic at the 1 %, 5 %, and 10 % levels based on regression curve. This depends on the number of observations.

5.3 Pairs Trading Strategy

This section includes the step by step procedure that we adopted to finally generate the trade signals.

- From here onwards we would work with two stocks “HDFCBANK.NS” and “AXISBANK.NS”. Let’s denote the 2 corresponding time series as : $\{y_t\}_{t=1}^{742}$ and $\{x_t\}_{t=1}^{742}$ respectively. Here number of sample points n = 742 i.e we have closing price of two stocks for 3 years.

- The price series is $I(1)$ and hence is not stationary. The below graph displays the non-stationarity of the price series of the two stocks.

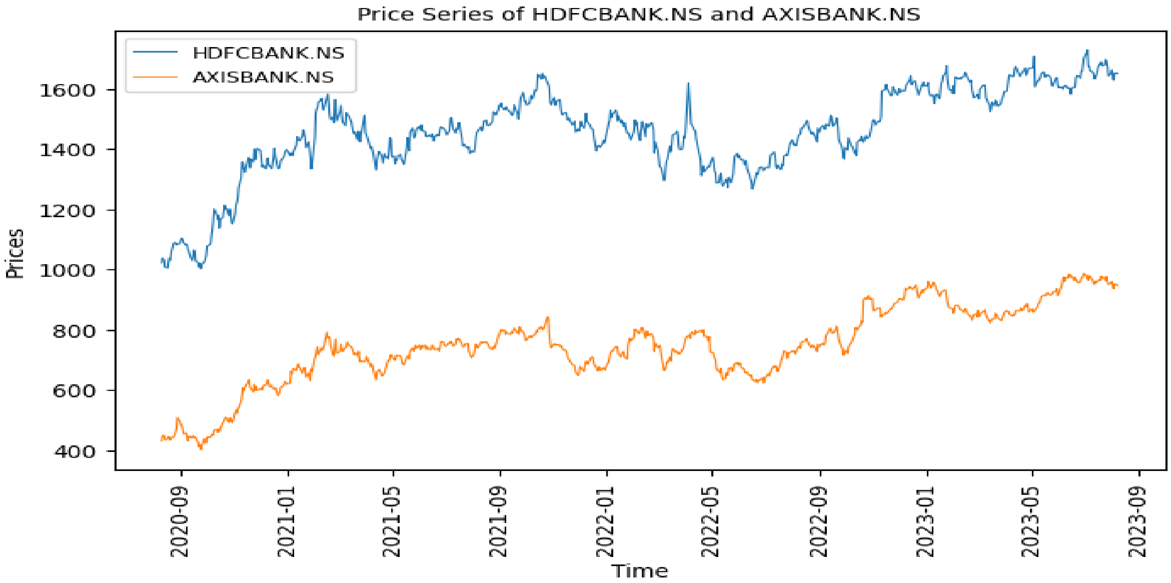

Figure 2

- Where as we expect the return series to be stationary. The returns series is defined as:

$$\frac{x_t - x_{t-1}}{x_{t-1}} \; for\ all\ t = 2, 3, ..., n$$

  We found that the return series is stationary. This result is also evident from figure 3. Hence Return series of ”HDFCBANK.NS” is $I(0)$ and therefore stationary.

- Hence we have obtained that individually the two time series are $I(1)$ i.e. they are non-stationary but becomes stationary after first order differencing.

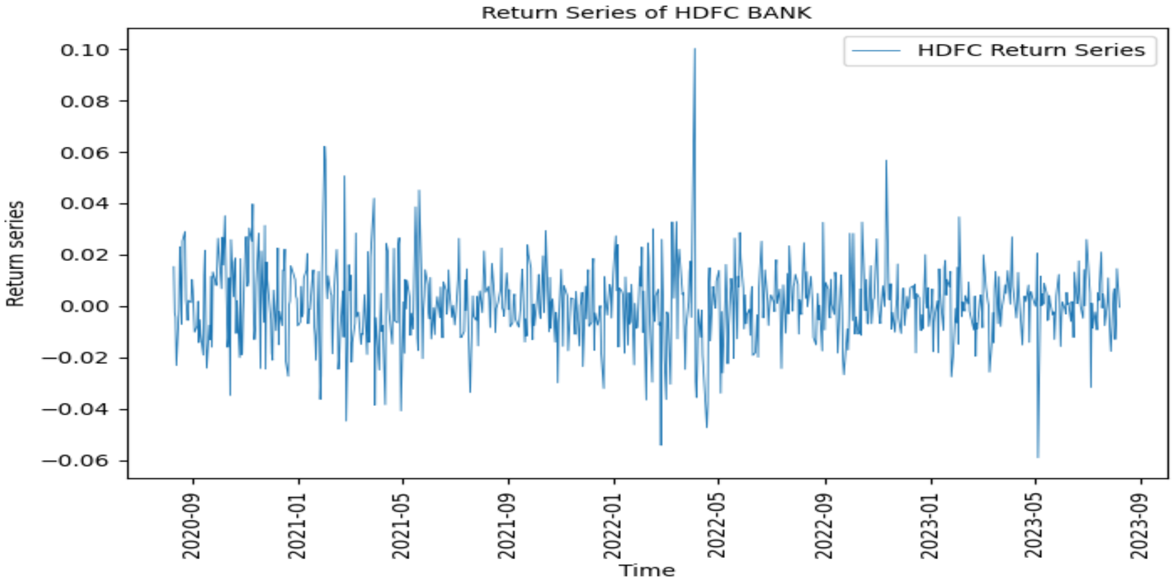

Figure 3

Now we apply the augmented Engle-Granger two-step co-integration test described in section 4.2 for testing co-integration between these two stocks and it turns out that these two stocks are co-integrated with $p - value = 0.001$. Next we are interested in finding the co-integrating vector.

- Having established the co-integration between the price series of HDFC and AXIS Bank, Plotted the Price ratio and Price spread and then tested their stationarity.

1. For calculating the price spread we have used simple linear regression to get the co-efficient estimates and hence calculated their residuals. After that we test the stationarity of the residual series. We have done this exercise taking HDFC as the dependent variable and AXIS as the independent variable first and then vice versa. And finally we selected the model that gave a lower p-value in the test for stationarity of the residual series. It turns out that AXIS bank is the independent variable and HDFC bank is the dependent variable and the OLS coefficients are: $Constant\ term\,(c)$: 665.480237; and AXISBANK.NS (β) : 1.058369

   Figure 4 Displays the spread of prices when we do least squares regression with AXIS bank prices as independent variable and prices of HDFC bank as dependent variable.

2. Figure 5 displays the price ratio series between HDFC and AXIS Bank.

   Now we test the stationarity of spread series and price ratio series using the

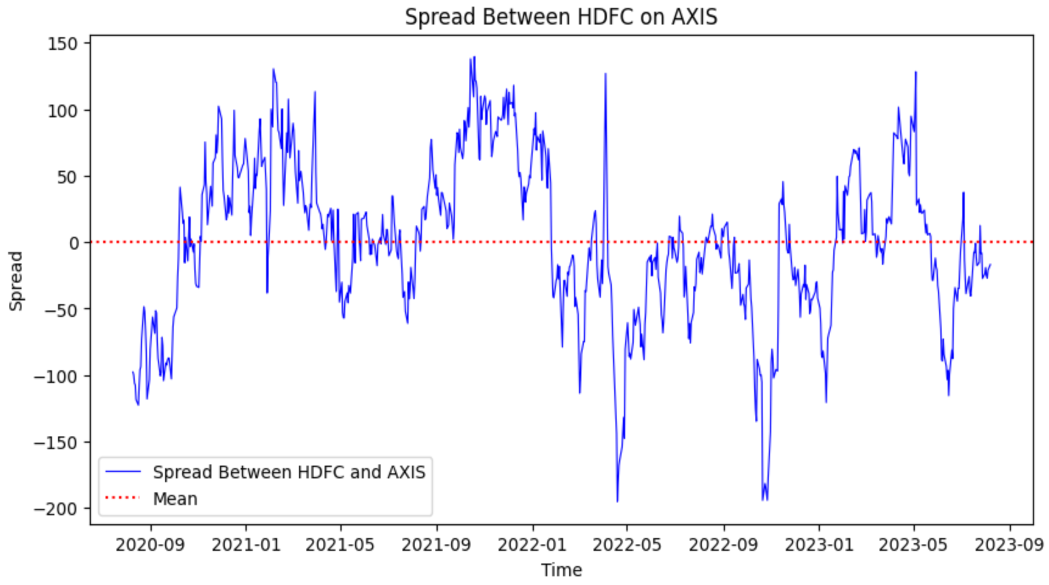

Figure 4

augmented Dickey Fuller test and the p-value turns out to be: $(0.13038, 0.00017)$. The spread p-value is lower so we will be using the spread series as the stationary mean and this would become the base of the mean-reversion trading strategy.

- In the spread series above we are able to see that any divergence from the mean converges back to the mean value and hence we can apply our method.

  First we divide the data into train and test set, we keep 70% data for training and 30% data for testing. From now on wards, we would work with the training data.

- We would be using **Moving Averages** to capture the short term and long term pattern movement in stock prices. Suppose we choose the window length to be 5 and 60. The 5 day rolling moving average(short Term) reacts quickly to some price changes whereas the long term (60 day) moving average doesn't reacts so quickly to the changes in prices due to short fluctuation.

  Figure 6 shows the plot of spread, 5 day Moving Average and 60 day Moving Average.

  The idea here is to see how many standard deviation is the 5 day moving average away from the 60 day moving average. That is if the spread diverges significantly from the long term moving average then we can get an opportunity of pairs trading.

  For this we calculate i^{th} element of Z Scores series as:

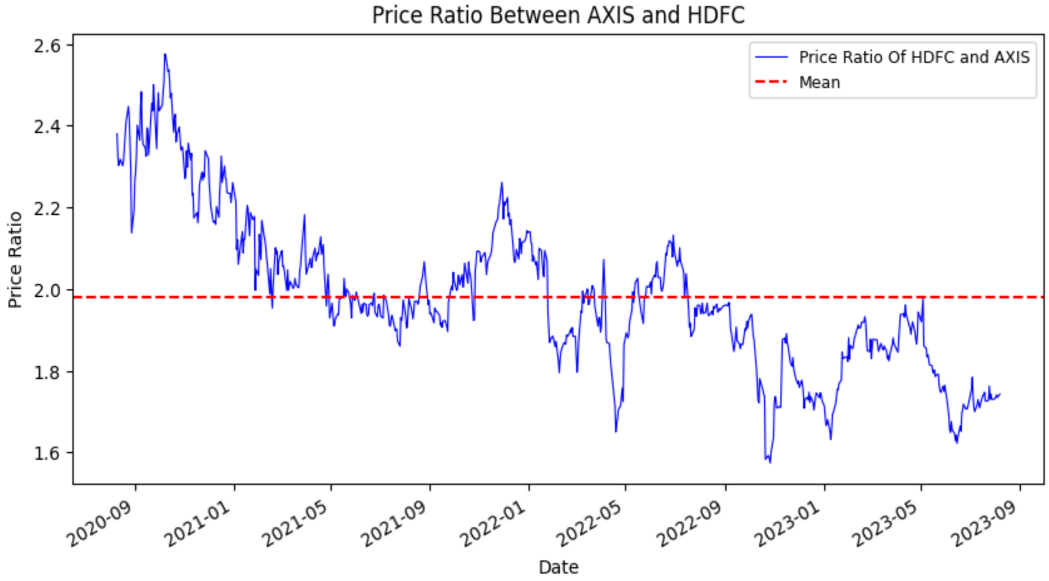

Figure 5

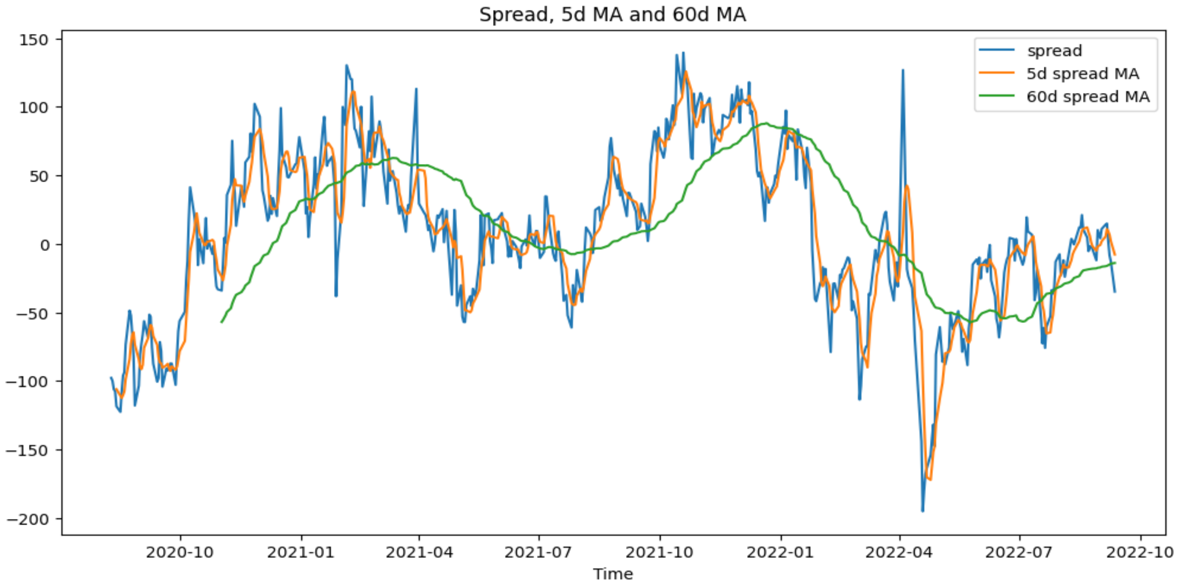

Figure 6

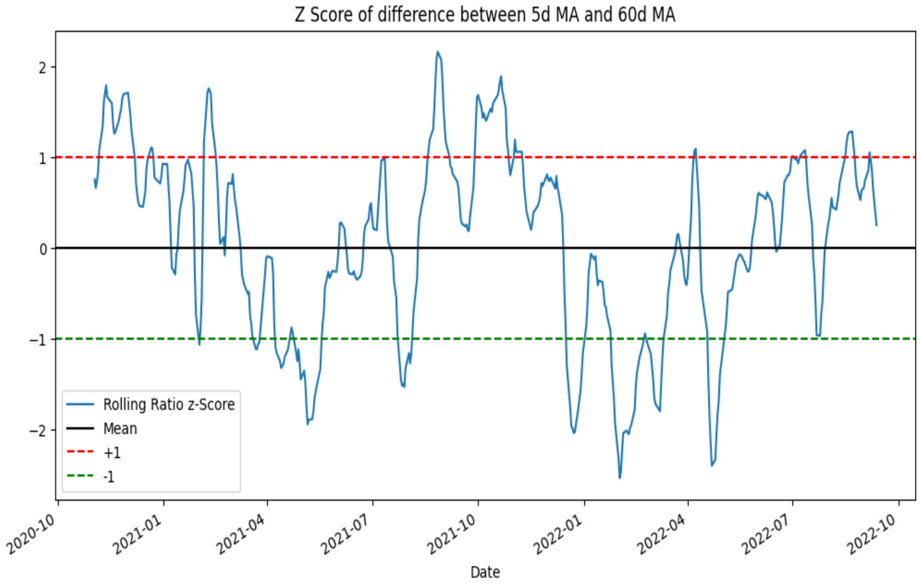

Figure 7

$$Z\,Score_i = \frac{\{(Spread\,5d\,MA)_i - (Spread\,60d\,MA)_i\}}{Std.\,deviation\,of\,60d\,MA\,series}$$

If the Z score is positive and significantly greater than zero, we can say that the short term m.a is more than long term m.a. This means one stock has become relatively overvalued and the trader should take a short position on the overvalued stock and a long position on the undervalued stock.

Similarly , If the Z score is negative and significantly less than zero, we can say that the short term m.a is less than long term m.a. This means one stock has become relatively undervalued and the trader should take a long position on the undervalued stock and a short position on the overvalued stock.

In pairs trading strategy we just exploit short-term deviations from the long-term equilibrium, assuming that these deviations are mean-reverting, i.e., prices will eventually revert to their historical relationship.

Here we are trying to capture these short-term deviations and profit from their correction.

Figure 7 Shows the Z scores of difference between 5 Day MA and 60 Day MA.

- **Trading Signals** are generated based on the following rule: (Here we take the threshold to be 1, but we find the value of threshold that would maximize the profit).

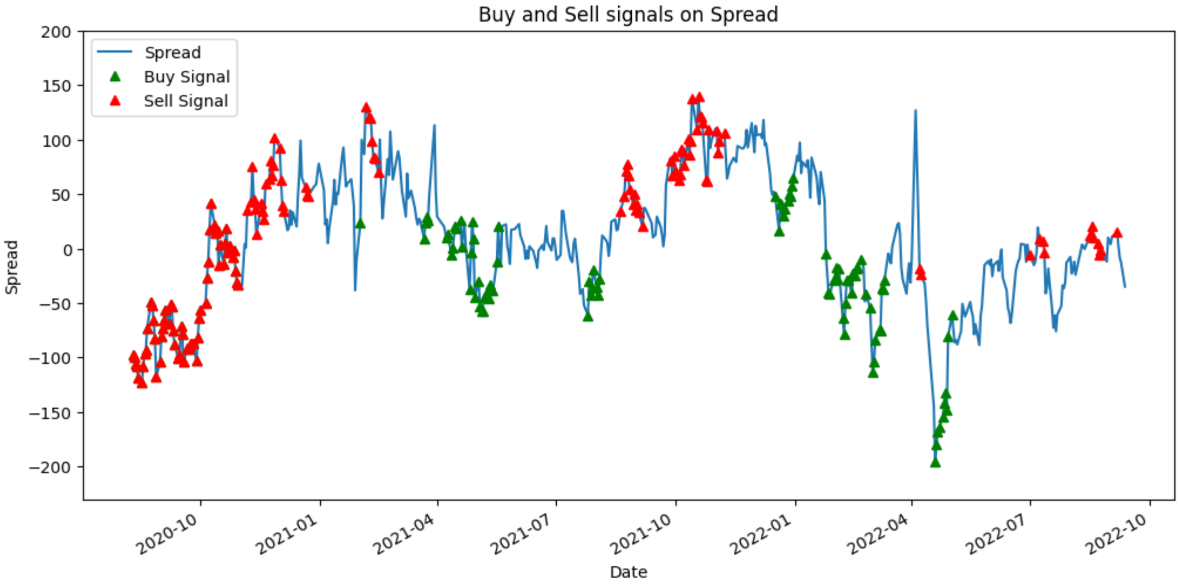

Figure 8

1. Buy(+1) the spread whenever the z_score is below -1, since we expect the spread to increase

2. Sell(-1) the spread whenever the z_score is above +1, since we expect the spread to decrease.

   Figure 7 displays the buy and trade signal on the spread series.

- **Trade signal on Individual Stocks** Consider the spread is $\epsilon_t = S1_t - (\alpha + \beta * S2_t)$. Buying the spread(i.e When ϵ_t is $<$ -1) means Buying Stock1(S1) and Selling Stock2(S2) and selling the spread(i.e When ϵ_t is $>$ 1) means Selling Stock(S1) and Buying Stock2(S2). In our case S1 is HDFCBANK and S2 is AXISBANK. Figure 8 displays the trade and buy signals on individual price series.

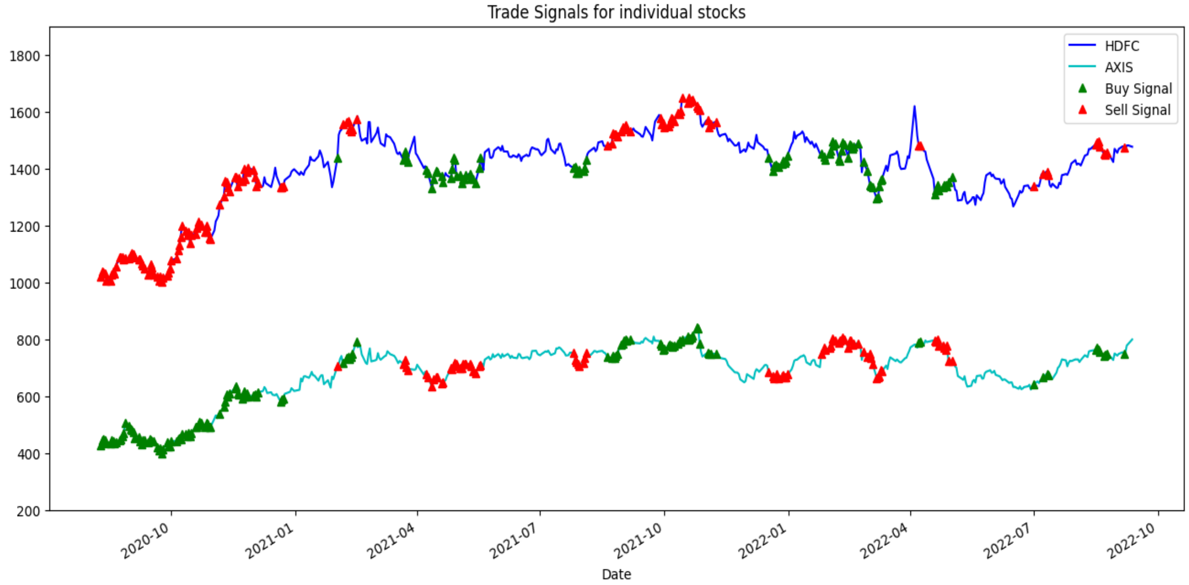

Figure 9

6 Trading Strategy into action and Estimated Profit

After the trade signal have been generated and we know when to buy and when to sell which stock, The question remains How much to buy and How much to sell and what are the expected profits.

For this we created a function called "Trade" which takes the following input: stock prices of both the stocks, window length for calculating moving averages, threshold values for opening and closing a trade and the OLS model that is fit with asset1 as dependent variable and asset2 as independent variable.

The function calculates the spread based on the OLS parameter estimates obtained from historical data. We define the price ratio as: $Price\,ratio = \frac{Stock1}{Stock2}$ and we trade Stock1 and Stock2 at the i^{th} instance according to $1 : Price\,ratio_i$.

We calculate the rolling moving average from spread series and we buy the spread (buy S1 and sell S2) as the Z Score based on the moving averages crosses $-threshold\,for\,opening\,trade$ value and we sell the spread (sell S1 and buy S2) as the Z score based on the moving averages crosses $threshold\,for\,opening\,trade$.

We close all the trades once the absolute value of the spread becomes less than $threshold\,for\,closing\,trade$. Each order is assumed to have an trading cost of 0.5% on the total order value. We subtract the total operational cost from gross profit to get the net profit.

Now we as increase the threshold for opening the trade the number of trade gets reduced, hence reducing the operational cost and profits from each trade increases. So basically we try to find out the best threshold values that would give us the maximum profits.

For input in the respective order as defined above: test data on HDFC and AXIS, 5 (short term moving average window) , 60 (short term moving average window), 0.8 (threshold for opening the trade), 0.4(threshold for closing the trade), model (OLS model fitted as HDFC as dependent variable and AXIS as the independent variable on the train data).
The output is: *Net Profit:* 3718.63
***Gross Profit:* 3838.95**
***Operational Cost:* 120.32**
***No of trades:* 74**

7 Results and Conclusion

Results

The table below shows the most co-integrated pairs from 5 major sectors of the economy and the optimized parameters for the moving average strategy and corresponding profits.

Sectors	**Stocks**	**Window1**	**Window2**	**Threshold open**	**Threshold close**	**Pro -fit**
BANKING	HDFCBANK, AXISBANK	11	30	0.5	0.07	17K
OIL - GAS	RELIANCE, GULFOILLUB	8	30	0.5	0.09	81K
NIFTY50	BAJAJFINANCE, ASIANPAINT	11	40	0.5	0.01	106K
AUTOMOBILE	TVSMOTOR, MARUTI	14	70	0.95	0.01	428K
FMCG	TATACONSUM, MARICO	10	70	0.5	0.15	15k

Table 1: Best co-integrated stocks with their profits

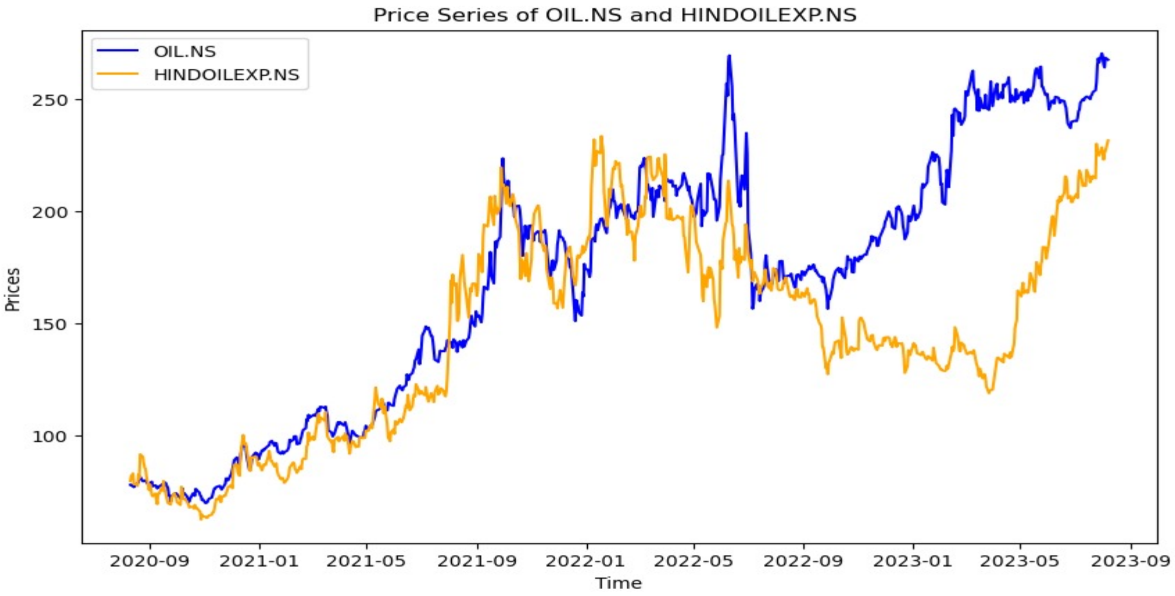

Figure 10

Conclusion Though Pairs Trading seems very fascinating, indeed it is, but it also has a lot of risk involved which needs to be taken into consideration before actually trading in the market.

An very important issue is that the spread should not be away from zero for a long time. There is a chance that the spread will never return to zero and in that case it would cost money to flatten the position. Further, It may happen that the pairs that are initially co-integrated during the training period may not remain co-integrated during the testing time. This may happen because of some market events such mergers, acquisitions or entry or exit of the company from a market. This is displayed in the Figure 10.