Capstone Project Report

A COINTEGRATION & CLUSTERING-BASED APPROACH TO PAIR TRADING

Presented By:

Aayush Sahu

Anirudh Pratap Singh

Ashutosh Pradhan

Gopad Shukla

Saif Rizvi

Umesh Kumar

Praxis Business School

Capstone Project Report

A COINTEGRATION & CLUSTERING-BASED APPROACH TO PAIR TRADING OF STOCKS FROM THE SECTORS OF THE INDIAN STOCK MARKET

Project Submitted By:

Aayush Sahu

Anirudh Pratap Singh

Ashutosh Pradhan

Gopar Kumar Shukla

Saif Rizvi

Umesh Kumar

Supervisor:

Jaydip Sen

A report submitted in fulfilment of the requirements for the Certificate of PGPDS

in the

July 2022 Batch
Data Science

26th April 2023

A COINTEGRATION AND CLUSTERING BASED APPROACH TO PAIR TRADING OF STOCKS FROM SELECTED SECTORS OF THE INDIAN STOCK MARKET

Aayush Sahu-A22001
Anirudh Pratap Singh-A22007
Ashutosh Pradhan-A22011
Gopad Kumar Shukla-A22016
Saif Rizvi-A22028
Umesh Kumar Bugata-A22039

April 29, 2023

AKNOWLEDGEMENT

We feel great pleasure in expressing deep sense of gratitude and sincere thanks to the venerable supervisor Prof. Jaydip Sen, Professor, Praxis Business School, Kolkata for his illuminating and precious guidance, invaluable suggestions, constant encouragement and support throughout the course of the project work. His patience, vision and understanding during the numerous crisis in this study were privilege to gain experience, learn and acquire the values of research, time management and independent thinking, under his esteemed supervision. Working with him, has been very exciting and enriching.

We would also take this opportunity to thank all those that helped make this project a success- Aji Thomas-A22004, Nishant Kr. Todi-A22023, Olive Ollemmyan-A22024, Swapnil Tripathi-A22037, Khushi Agarwal-BM22028, Priya Agarwal-BM22048 and Sadhvi Kumari-BM22056 for their invaluable contribution.

Group 7

ABSTRACT

The project makes three distinct contributions. First, the work provides a cointegration-based pair trading strategy for stock portfolio creation, that can be used to earn profit by the investors in the stock market. Second, the pair-trading models are trained and tested on real-world stock market data, with the results displayed to illustrate the models' efficacy. Finally, since the stocks utilized in the pair trading portfolio designs are drawn from various NSE sectors, the outcomes of the pairings are an excellent signal of the possible profit that investors could make if they invest in those sectors using the recommended pair-trading technique.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Stock Market

A stock market, stock market, or equity market is an aggregation of buyers and sellers of shares (also called shares) that represent ownership claims to businesses; they can include securities listed on a public stock exchange as well as stocks that are only traded privately, such as shares of private companies that are sold to investors through equity crowd-funding platforms. Investment is usually made with an investment strategy in mind.

Both "stock market" and "stock exchange" are often used interchangeably. Traders in the stock market buy or sell shares on one or more of the stock exchanges that are part of the overall stock market.

Stock markets provide a safe and regulated environment where market participants can confidently trade stocks and other eligible financial instruments, with zero to low operational risk.

Once new securities have been sold in the primary market, they are traded in the secondary market where one investor buys shares from another investor at the prevailing market price or at whatever price both the buyer and seller agree upon. The secondary market or the stock exchanges are regulated by the regulatory authority. In India, the secondary and primary markets are governed by the Security and Exchange Board of India (SEBI).

A stock exchange facilitates stockbrokers to trade company stocks and other securities. A stock may be bought or sold only if it is listed on an exchange. Thus, it is the meeting place of the stock buyers and sellers. India's premier

stock exchanges are the Bombay Stock Exchange and the National Stock Exchange.

As a primary market, the stock exchange allows companies to issue and sell their shares to the public for the first time through the initial public offering (IPO) process. This activity helps companies to raise the necessary capital from investors.

The company splits into several shares and sells some of those shares to the public at a price per share.

To facilitate this process, a company needs a marketplace where these shares can be sold, and this is achieved in the stock market. A listed company may also offer new, additional shares through other offers at a later stage, such as rights issues or follow-on offers. They can even buy back or remove their shares.

Investors will own a company's stock in the expectation that the stock will increase in value or receive a dividend payment, or both. The exchange acts as an intermediary in this capital-raising process and receives a fee from the company and its financial partners for its services.

Using exchanges, investors can also buy and sell securities they already own in the so-called secondary market.

1.2 Pairs Trading

Pairs trading was first introduced in the mid-1980s by a group of technical analyst researchers in Morgan Stanley, a multinational investment bank and financial services company. The pairs trading strategy uses statistical and technical analysis to find potential market-neutral profits.

A market-neutral strategy is a type of investment strategy undertaken by an investor or an investment manager that seeks to profit from both increasing and decreasing prices in one or more markets while also attempting to

completely avoid some specific form of market risk.

Market-neutral strategies are often attained by taking matching long and short positions in different stocks to increase the return from making good stock selections and decreasing the return from broad market movements.

Pair trading is a market-neutral strategy to identify and pair two stocks, usually from the same sector or industry, that show identical positive movements. When the move breaks down, the two stocks deviate from the path, causing a spread that becomes a trading opportunity.

In pair trading, you buy the better-performing stock and sell the other one when that trading opportunity comes up to balance the profit and loss. The direction of the broad market does not affect the loss or profit of the trade's strategy. Traders tend to lose less when trading pairs when they thoroughly understand and use pairs trading strategies.

1.3 Correlation and Co-integration

A pair trading strategy is based on the historical correlation of two securities. The securities in the pair trade must have a high positive correlation, which is the main driver of the strategy's profits. A pair trading strategy is best deployed when a trader identifies a correlation discrepancy. Relying on the historical notion that two securities will maintain a specified correlation, a pair trade can be placed when that correlation fluctuates.

Example: imagine that the historical price movements of HDFC Bank (HDFCBANK) and SBI Bank (SBIN) show a high degree of positive correlation. If a price divergence were to occur, such that the correlation temporarily burst, traders using the pairs approach could see this as an opportunity to deploy a spread.

When trading stocks and futures, pairs are most used. For the second group, the most famous pairs include oil versus natural gas, gold versus silver, wheat versus corn, and many others.

The basis of pair trades is that a strong positive or negative correlation has been established between them over a prolonged period. The ability to profit from a pair position depends on the belief that the pair will revert to its historical average after a temporary breakdown in correlation.

In principle, cointegration analysis can give traders more confidence that a temporary breakdown in correlation will indeed return to "normal" trading behaviour.

Because of the strong correlation/cointegration that exists between the optimal pairs, there is theoretically less risk in these trade structures as compared to placing bare directional risk in a single underlying.

2. METHODOLOGY

In this section, particulars of the data used and the methods employed in this work are described. Using training data from five NSE sectors, co integrated stock pairs are first identified, and then portfolios are formed on those pairs. The pair-trading portfolios' yearly returns are calculated. The returns of the portfolios over the test period are used to compare the pairs and the sectors. As the part of research methodology, following steps were involved

2.1 Sectors for analysis

From all the sectors listed in NSE and which are listed below, five have been chosen for analysis along with NIFTY 50 stocks. The sectors include – (i) Auto, (ii) IT, (iii) PSU, (iv) Bank, (v) FMCG, (vi) Media, (vii) Metal, (viii) Oil and Gas, (ix) Pharma. Top ten stocks from each sector have been taken for analysis based on their market capitalization.

2.2 Data Collection

Historical stock prices have been downloaded using Yfinance library in python which allows access to the Yahoo Finance API. As the training dataset, close values of the stock prices from '2018-1-1' to '2021-12-31' have been used. For evaluation, close values of the stock prices from '2022-1-1' to '2022-12-31' have been used as the test dataset.

2.3 Correlation Matrix

In order to obtain a comprehensive visualization of the correlation coefficients between the pairs of stocks in each sector, a correlation matrix is constructed for each sector at this stage using the heatmap function of the seaborn library. Given that a sector has ten stocks, the Pearson's product-moment correlation coefficients will be available for 45 different pairs since each sector contains 10 stocks. The correlation coefficients are primarily calculated using the close prices' return values. Using the Python function pct-change function, the return series for each stock is calculated. The correlation coefficients can be used to estimate the degree of linear relationship between two stocks, but they cannot detect whether there is cointegration between the pairs.

2.4 Identification of Co integrated stock pairs

A pair of stocks are said to be cointegrated if they exhibit significant correlation in the long term. If two stocks are cointegrated, there exists some linear combination between the stocks that will vary around a mean(Quantopian Inc.). The coint function, which is part of the Python statsmodels module's stattools submodule, is used to determine whether pairings are cointegrated. The coint test's null hypothesis presupposes that a given pair, which is supplied as the function's two parameters, is not cointegrated (MacKinnon, 2012). Therefore, pairings are presumed to be cointegrated if their p-values fall below the cutoff of 0.05. For the purpose of constructing pair-trading portfolios, pairs for which p-values are less than 0.05 are taken into consideration.

2.5 OLS regression models with the pairs

An ordinary least square (OLS) regression model is built for the pairs for which the cointegration test null hypothesis has been rejected (i.e., the pairs that display cointegration) in the preceding stage. Conventionally, the pre-

dictor (i.e., independent variable) in the OLS model is chosen to be the stock with the greater mean value of its close price. The model is constructed using the OLS function specified in the linear-model submodule of the Python statsmodels module. There are various parts to the OLS function's output. The following are a few of the most significant ones. (i) The hedge ratio and the p-value of its t-statistics, (ii) The p-value of the F-statistics, (iii) The prob. of the Omnibus test statistics, (iv) The value of the Durbin-Watson test statistics, (v) The prob. of the Jarque-Bera test statistics. The hedging ratio is the coefficient parameter, and the p-value of the predictor's t-statistics must be significant for the predictor to have a meaningful level of explanatory power for the target. For the regression model to make accurate predictions, the p-value of the F-statistic needs to be significant (lower than 0.05). In order to ensure that the predictor has a substantial impact on the target, the Omnibus test statistic should likewise have a significant p-value. In order to prevent autocorrelation among the residuals, the Durbin-Watson test statistic's value should be close to 2. There should not be any significant autocorrelation among the model residuals and the cointegrated pair of stocks should have significant p-values for the F-statistics and the Omnibus test. There is one more prerequisite for cointegration, though: the residuals must be stationary which would be checked in the next step.

2.6 Trading Signals for the pairs

This stage involves creating the trading signals for the pair after the cointegration for the pairs has been verified. This requires numerous further steps. First, the daily value of the ratio between the close value of the predictor time series (also known as asset1) and the close value of the target (also known as asset2) is designed as a variable called ratio. Second, by deducting the mean value from each ratio value and dividing the result by the ratio values' standard deviation, the standardized values of the variable ratio are calculated. The standard deviation is added to and subtracted from the mean value, respectively, to determine the upper and lower bounds of the Z-scores for the ratio values. The maximum and lower bounds of the Z-score for the ratio are 1 and -1, respectively, because the mean value of the Z-scores is 0 and their standard deviation is 1. Third, the following six columns are

now added to a pandas data frame called signals: Date, asset1 (time series of the close values of the first stock), asset2 (time series of the close values of the second stock), Z score (standardized values of the ratio), upper limit (all entries are "1" under this column), lower limit (storing the lower limit of the Z scores), and date. The trading signals for asset1 are now developed using the Z-scores and their upper and lower bounds. Asset1 should be used with the short strategy if the Z-score at a point exceeds the upper limit; otherwise, the long strategy should be employed. Selling asset 1 is part of the short strategy, and increasing asset 1 purchases is part of the long plan. The signals data frame now includes a new column called signals1, with the values "-1" for short cases and "1" for long cases for asset1. The signals data frame for asset2 also gains a new column, signals2. Signals2 entries are different from Signals1 entries in terms of sign. At this stage the data frame consists of the following columns - date, asset1, asset2, Z-score, upper limit, lower limit, signals1, and signals2.

2.7 Identifying the points of investment opportunities

Two further columns are added to the signals data frame to identify the portfolio positions, or the areas of undervalued investing opportunities, for the two companies. Column positions1 is added for the initial stock, or asset 1. The first-order difference of the singals1 column is used to calculate the items under this column. Like the first stock, column positions2 for the second stock are added based on the signals2 column's first-order difference values. If "0" is entered under positions1, it means that signals1's prior and current values have not changed. A signal1 entry with a value of "1" denotes a change from "0" to "1," while one with a value of "-1" denotes a transition from "1" to "0." The entries in column positions 2 are chosen in a similar manner. Records with a positions1 value of "1" indicate a long trigger for asset 1, while records with a positions1 value of "0" suggest a short trigger. Signal1's value has changed from "0" to "1" when a record has a position1 entry of "1". So, asset1 has been given a long trigger. Similar to this, a positions1 entry of "-1" indicates a short trigger. An input of "0" for position1 denotes that asset1 does not require action. Like position 2, entries of "-1", "1," and "0" denote a short trigger, a long trigger, or no

action required for asset 2, respectively. During the portfolio testing phase, a graph is drawn to show the short and long trigger points for both stocks.

2.8 Computation of returns

The portfolio returns for the year 2021 are calculated in the final phase. On January 1, 2022, the pair-trading portfolio is started with a capital investment of 100000 units for each asset (i.e., a total investment of 200000 units). The entire holding values and cash amount for both stocks are added up for each day of the test period and kept in a variable called total portfolio value. To determine the profit (or loss) as of December 31, 2022, the overall portfolio value's excess (or deficit) over the total original investment of 200000 units is determined. Transaction costs are not considered in this calculation. The return on the portfolio is generated from the calculated profit.

3. RESULTS

3.1 Auto-Sector

3.1.1 Correlation

Correlation is the easiest approach to look for stock pairs that exhibit long term similar behaviour, which can be exploited to get pairs that give good profit for a given testing period. The Correlation matrix given in 3.1. The stock pairs exhibiting pearson correlation coefficient of more than 0.5 are selected as the correlated stock pairs. The correlated stock pairs obtained with this selection criteria are:

- MARUTI, TVSMOTOR
- M& M, MARUTI
- HEROMOTOCOP, TVSMOTOR
- HEROMOTOCOP, MARUTI
- EICHERMOTORS, MARUTI
- BHARATFORGE, TATMOTORS
- BAJAJAUTO, TVSMOTORS
- BAJAJAUTO, MARUTI
- BAJAJAUTO, HEROMOTOCOP
- AHSHOKLEY, BHARATFORGE

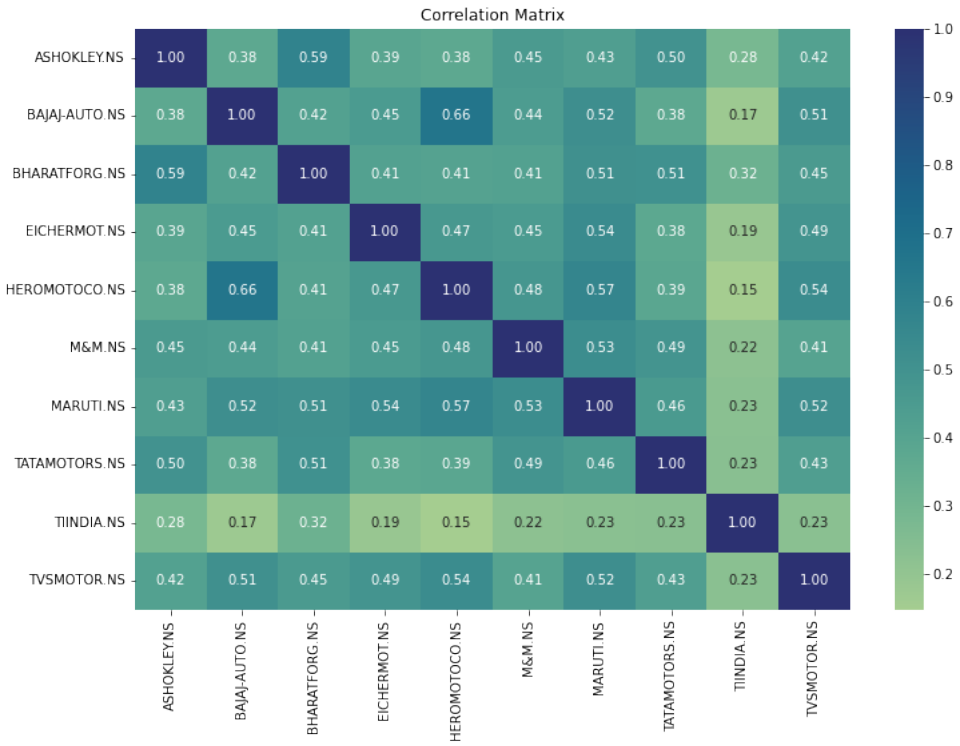

Fig. 3.1: Correlation heat map with respective p-values

The 3.1 below gives the final value of portfolio with the returns

Tab. 3.1: THE ANNUAL RETURNS OF THE PAIR-TRADING PORTFOLIOS OF THE AUTO SECTOR STOCKS (PERIOD: JANUARY 1, 2022 - DECEMBER 31, 2022).

Stock pairs	Initial Investment	Final value	**Profit**	Return
MARUTI TVSMOTORS	200000	406406.45	6406.45	1.60
M& M MARUTI	200000	416955.32	16955.32	4.24
HEROMOTOCO TVSMOTOR	200000	441278.19	41278.19	10.32
HEROMOTOCOP EICHERMOTOR	200000	469844.07	69844.07	17.46
'BHARATFORG' MARUTI	200000	445110.83	45110.83	11.28
BHARATFORGE TATMOTORS	200000	348449.80	-51550.20	-12.89
BHARATFORG MARUTI	200000	433626.17	33626.17	8.41
BAJAJ AUTO TVSMOTORS	200000	393704.01	-6295.99	-1.57
BAJAJ AUTO MARUTI	200000	464142.30	64142.30	16.04
BAJAJ AUTO HEROMOTOCO	200000	431294.64	31294.64	7.82
ASHOKLEY BHARATFORGE	200000	444855.03	44855.03	11.21

3.1.2 Co-integration

As per the report published by the NSE on December 31, 2022, the ten most important stocks of the auto sector are as follows: Mahindra and Mahindra, Maruti Suzuki India, Tata Motors, Eicher Motors, Bajaj Auto, Hero MotoCorp, Tube Investments of India, TVS Motor Company, Bharat Forge, and Ashok Leyland (NSE Website). Six pairs in this sector exhibited significant p-values for their cointegration, as depicted in 3.2 These six pairs are as follows:

- 'ASHOKLEY.NS', 'BHARATFORG.NS'
- 'ASHOKLEY.NS', 'EICHERMOT.NS'
- 'ASHOKLEY.NS', 'TVSMOTOR.NS'
- 'BHARATFORG.NS', 'EICHERMOT.NS'
- 'BHARATFORG.NS', 'TVSMOTOR.NS'
- 'EICHERMOT.NS', 'MandM.NS'

Due to space constraints, the results of the pair Ashok Leyland and Bharat Forge are presented in detail. For the other pairs, only the annual return figures for the year 2022 are presented. The daily close prices of Ashok Leyland and Bharat Forge are depicted for the training period (i.e., January 1, 2018, to December 31, 2021) in 3.3 to visually check the long-term association between them.

The OLS regression results with Bharat Forge as the target and Ashok Leyland as the predictor variable are presented in 3.4. The hedge ratio (i.e., the coefficient parameter for Ashok Leyland) is found to be 5.6851 with a standard error of 0.017. The highly significant p-value of the F-statistic indicates that there is a strong linear relationship between the two series. The Jarque-Bera test yielded a weakly significant p-value for the test statistic indicating the skewness and the kurtosis of the residuals of the model are not like those of a normal distribution. The Durbin-Watson test statistic lies between 0

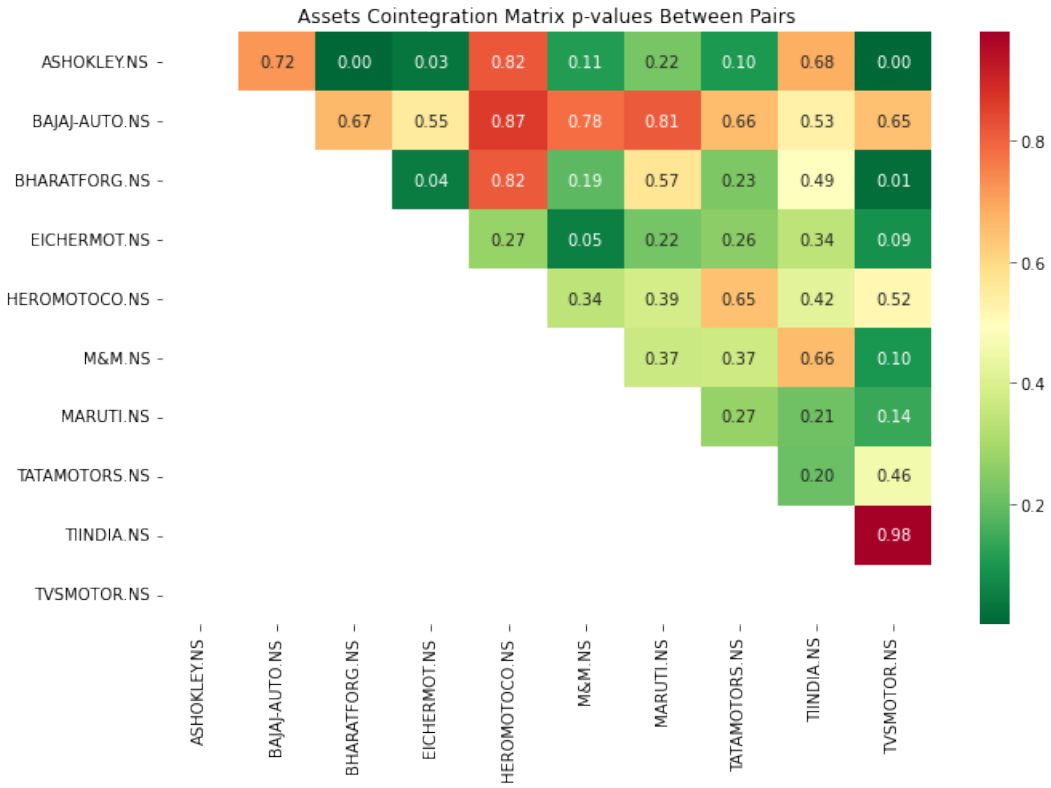

Fig. 3.2: The matrix of the p-values of the cointegration tests for the pairs of stocks of the auto sector.

and 2 indicating a positive autocorrelation among the residual. Finally, the Omnibus test results support the p-value of the F statistic indicating the presence of a strong linear relationship between the target and the predictor. In the next step, the residuals of the model are computed and plotted. 3.5 depicts the residual plot which looks like a stationary series.

The Augmented Dickey Fuller (ADF) test on the residual series yielded a value of -4.3084 while the critical value at the 1% level of significance is -3.4392 (Dickey & Fuller, 1979). The results of the ADF test indicated that the residual series is not stationary. However, as the Bharat Forge and Ashok Leyland series are cointegrated, we proceed to execute their pair-trading strategies. 3.6 depicts the plot of the Z-values of the ratio series and their upper and lower limits. The ratio series represents the ratio of the daily close

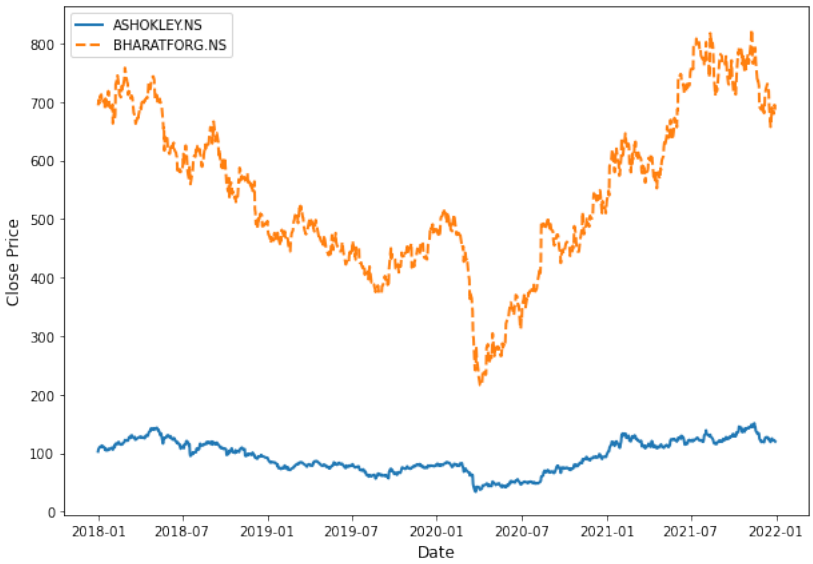

Fig. 3.3: The plot depicting the daily close prices of the stocks Ashok Leyland (ASHOKLEY.NS) and Bharat Forge (BHARATFORG.NS), from January 1, 2018, to December 31, 2021.

prices of Bharat Forge and Ashok Leyland for the test data

The pair-trading signal points are plotted in 3.7, in which four types of signal points are identified, long strategy for Bharat Forge, short strategy for Bharat Forge, long strategy for Ashok Leyland, and short strategy for Ashok Leyland. The four strategies are denoted by different symbols in 3.7. Finally, with an initial investment of 200000 units of capital for each of the two stocks, the trading strategy is initiated on January 1, 2022. 3.8 depicts the daily values of the portfolio for the year 2022. The value of the portfolio at the end of the year is found to be 444845 which corresponds to an annual return of 11.21%. 3.2 presents the annual returns for six pairs of stocks from the auto sector which exhibited significant p-values for their cointegration. Except for two, all the pairs are found to produce positive returns with the Ashok Leyland – Eicher Motors pair producing the highest return.

OLS Regression Results

Dep. Variable:	asset2	R-squared (uncentered):	0.992
Model:	OLS	Adj. R-squared (uncentered):	0.992
Method:	Least Squares	F-statistic:	1.175e+05
Date:	Mon, 13 Mar 2023	Prob (F-statistic):	0.00
Time:	16:00:12	Log-Likelihood:	-5288.6
No. Observations:	987	AIC:	1.058e+04
Df Residuals:	986	BIC:	1.058e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
asset1	5.6851	0.017	342.776	0.000	5.653	5.718

Omnibus:	4.227	Durbin-Watson:	0.064
Prob(Omnibus):	0.121	Jarque-Bera (JB):	4.295
Skew:	-0.149	Prob(JB):	0.117
Kurtosis:	2.876	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 3.4: The results of the OLS regression model with Bharat Forge as the target variable (denoted by asset2) and Ashok Leyland as the predictor variable (denoted by asset1).

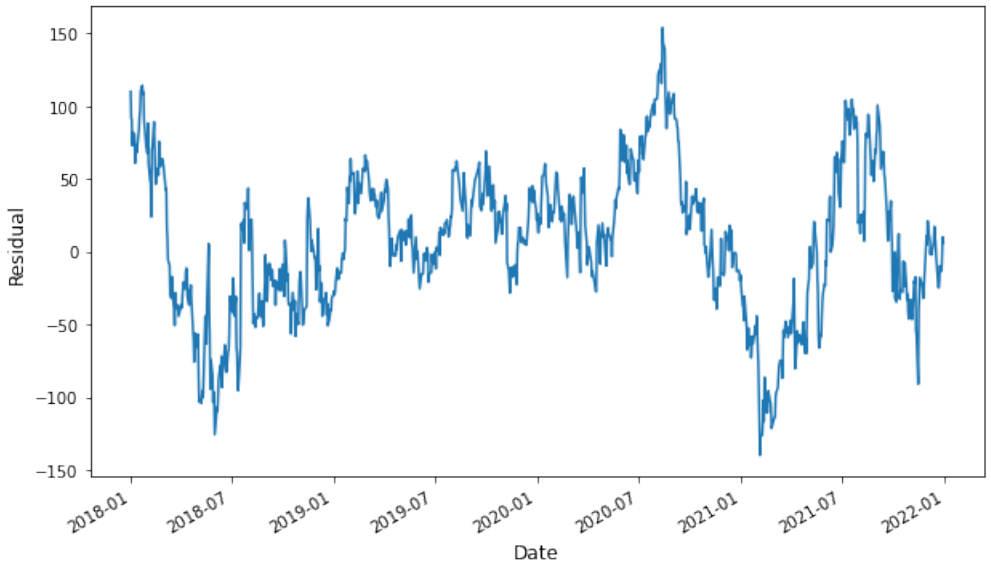

Fig. 3.5: The residual plot for the OLS regression model with Bharat Forge as the predictor and Ashok Leyland as the target variable.

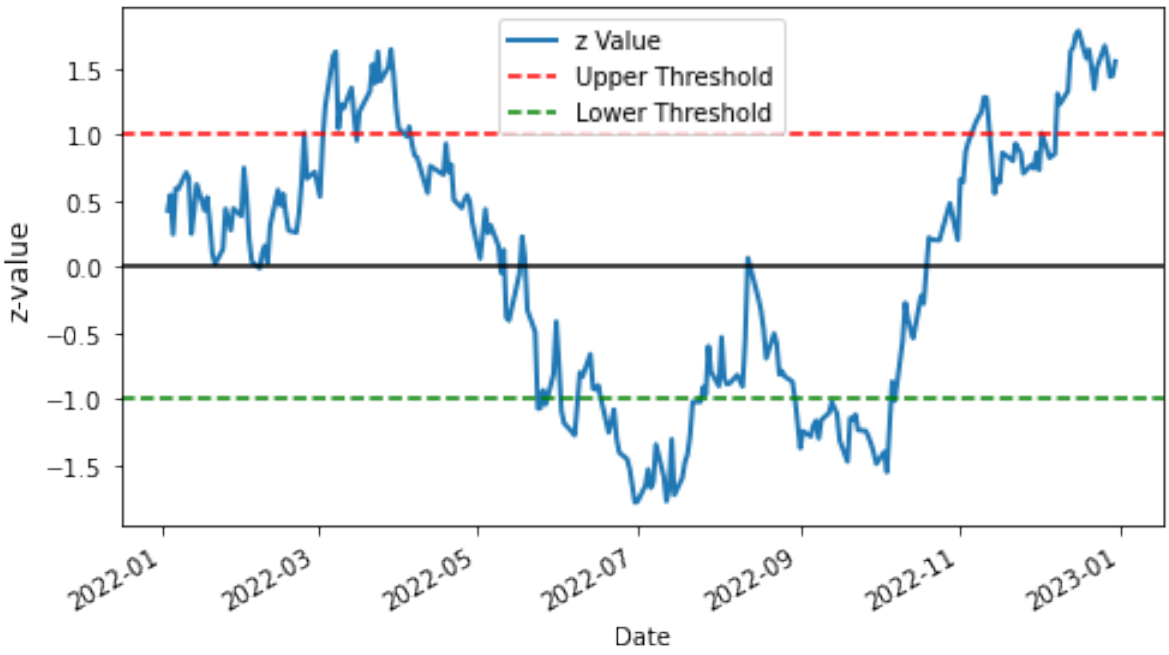

Fig. 3.6: The plot of the Z-values of the ratio series for the Bharat Forge – Ashok Leyland pair and their upper and lower limits.

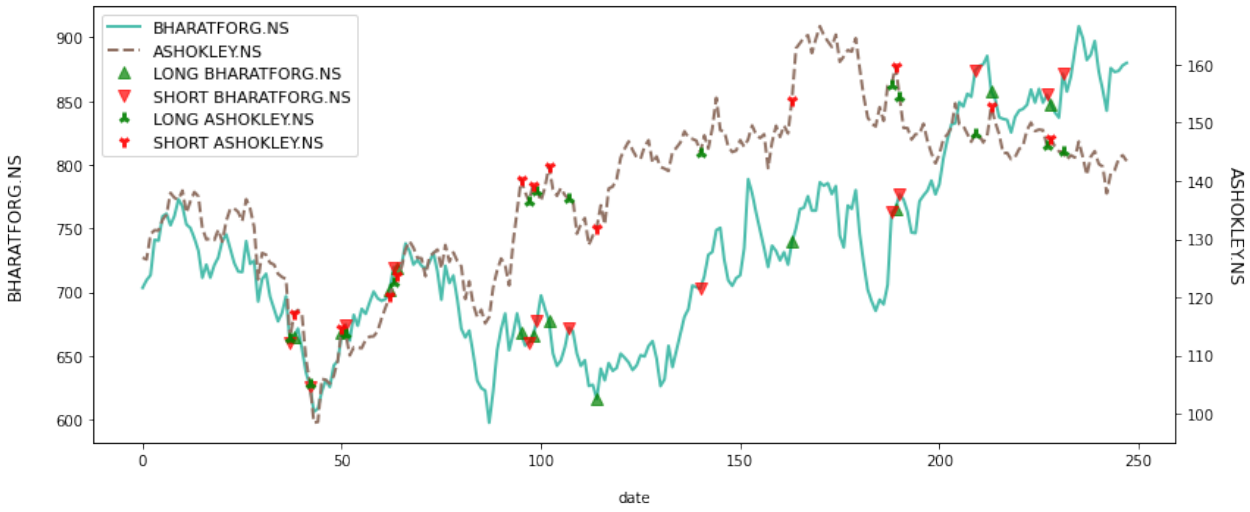

Fig. 3.7: The pair-trading scenarios for the stocks Bharat Forge and Ashok Leyland – the trading signals and their positions identified.

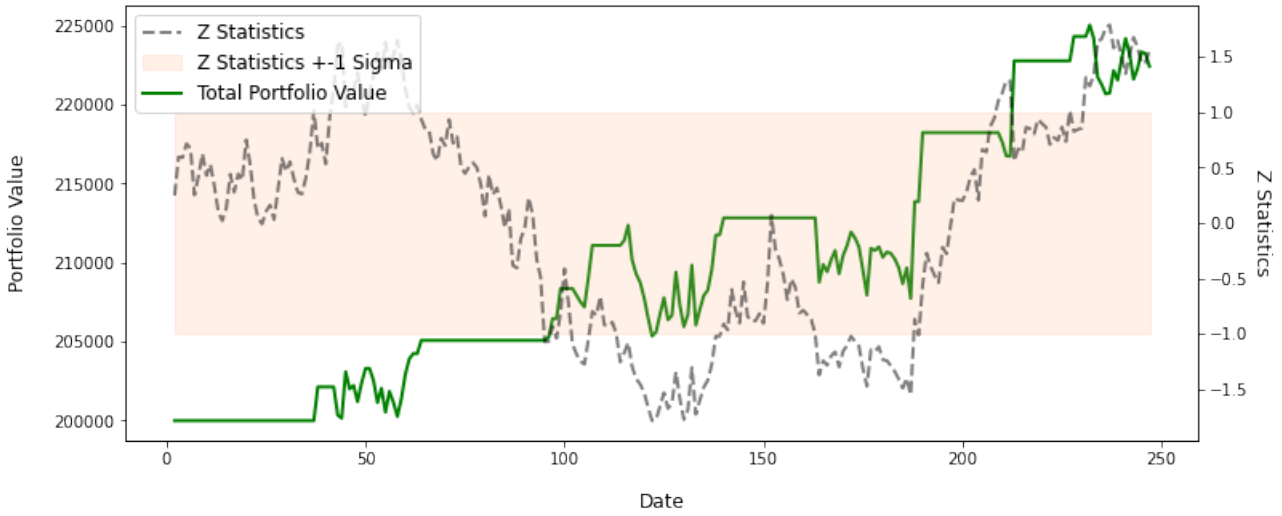

Fig. 3.8: The plot of the daily value of the pair-trading portfolio of Bharat Forge and Ashok Leyland from January 1, 2022, to December 31, 2022.

Tab. 3.2: THE ANNUAL RETURNS OF THE PAIR-TRADING PORTFOLIOS OF THE AUTO SECTOR STOCKS (PERIOD: JANUARY 1, 2022 - DECEMBER 31, 2022).

Stock pairs	Initial Investment	Final value	**Profit**	Return
ASHOKLEY 'BHARATFORG'	200000	444855.02	44855.02	11.21
'ASHOKLEY' 'EICHERMOT'	200000	482467.33	82467.33	20.62
'ASHOKLEY' 'TVSMOTOR'	200000	358160.46	-41839.54	-10.46
'BHARATFORG' 'EICHERMOT'	200000	440510.21	40510.21	10.13
'BHARATFORG' 'TVSMOTOR'	200000	397865.59	-2134.41	-0.53
'EICHERMOT' 'MandM'	200000	425700.04	25700.04	6.425

3.1.3 Clustering

K-means clustering KMeans module has been imported from sklearn.cluster package in python in order to perform the clustering. To find the optimum number of clusters in k-means technique we look at the silhouette score for different values of K (no. of clusters). The y-axis denotes silhouette scores and x-axis denotes the number of cluster(k values)

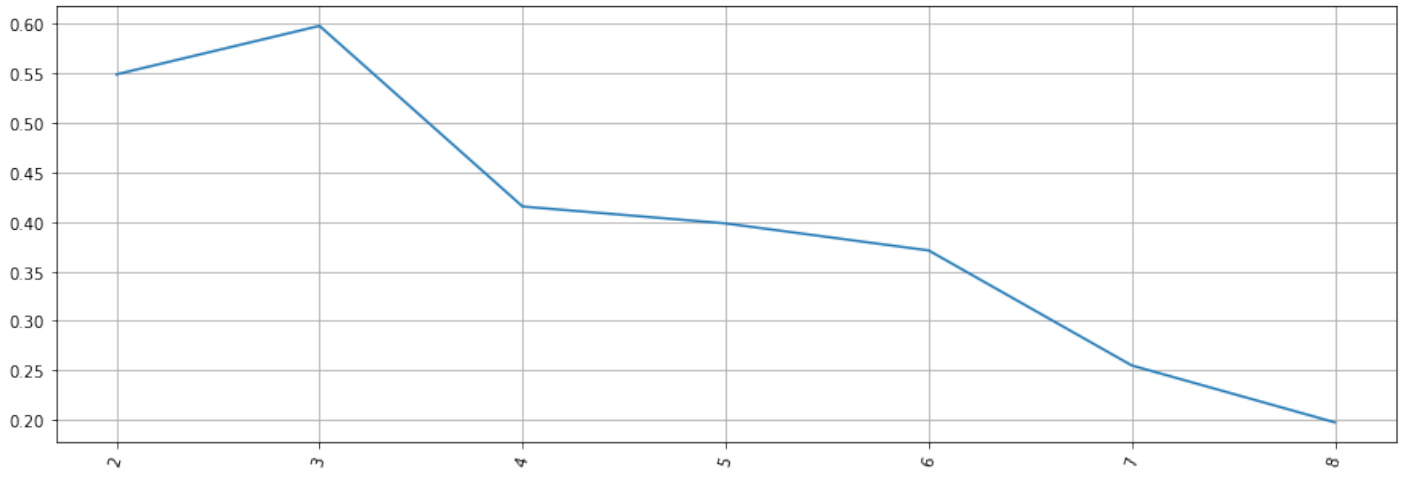

Fig. 3.9: Silhouette score vs. number of clusters

The silhouette score maximizes at k=3. Hence, number of clusters chosen are 3.

Clusters were be visualized using "k-means cluster centers" and scatter plot. The cluster centroids could be seen as green squares, and the number of stocks associated with each cluster could be seen as circles around squares with distinct colors. We can clearly see three clusters.

Number of stocks in each cluster can be further visualized using bar chart -

As per the above chart, there are 7,2 and 1 stocks in clusters 0,1 and 2 respectively.

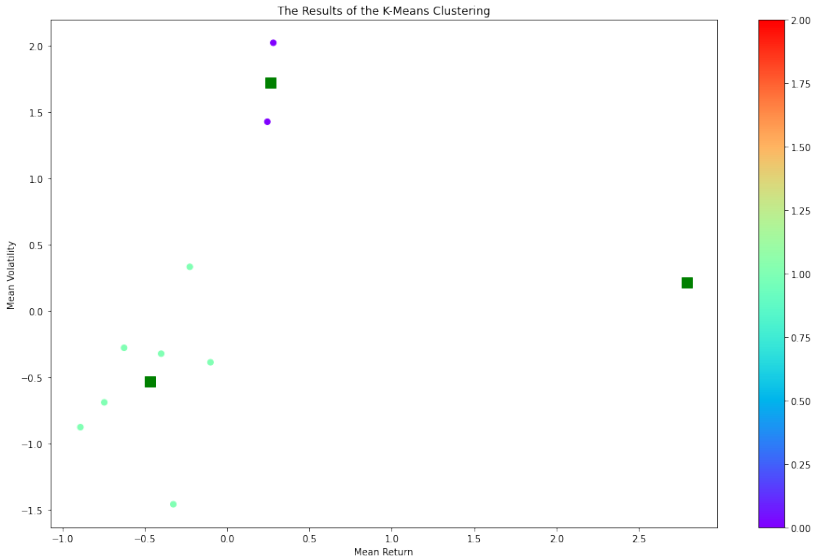

Fig. 3.10: The results of K-means clustering

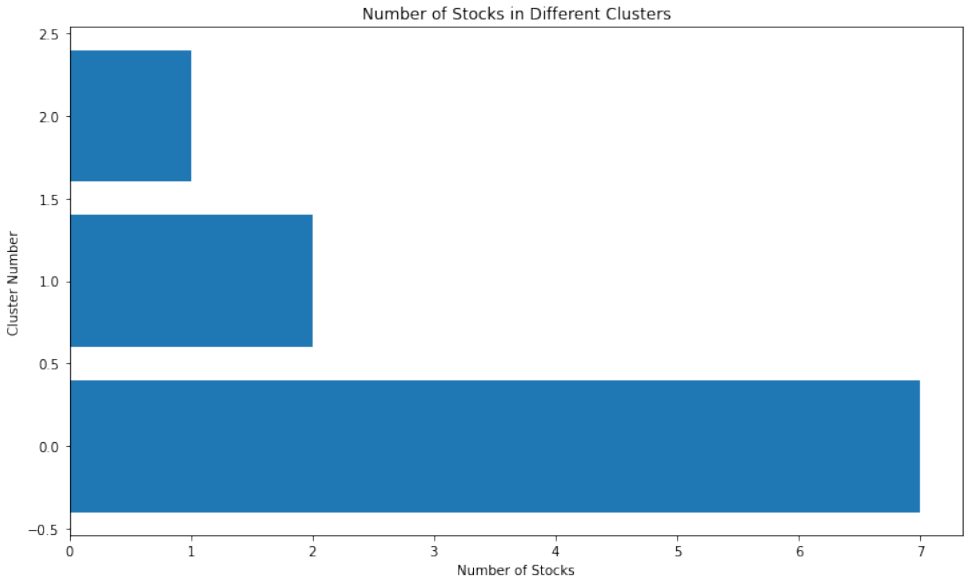

Fig. 3.11: Number of stocks in different clusters

Hierarchical clustering – The best way to visualize an agglomerate clustering algorithm is through a dendrogram, which displays a cluster tree, the leaves being the individual stocks and the root being the final single cluster. The "distance" between each cluster is shown on the y-axis, and thus the longer the branches are, the less correlated two clusters are. In order to perform this clustering and visualize dendrogram, linkage, ward, and AgglomerativeClustering modules were imported.

For Auto Sector the dendrogram plot looked like –

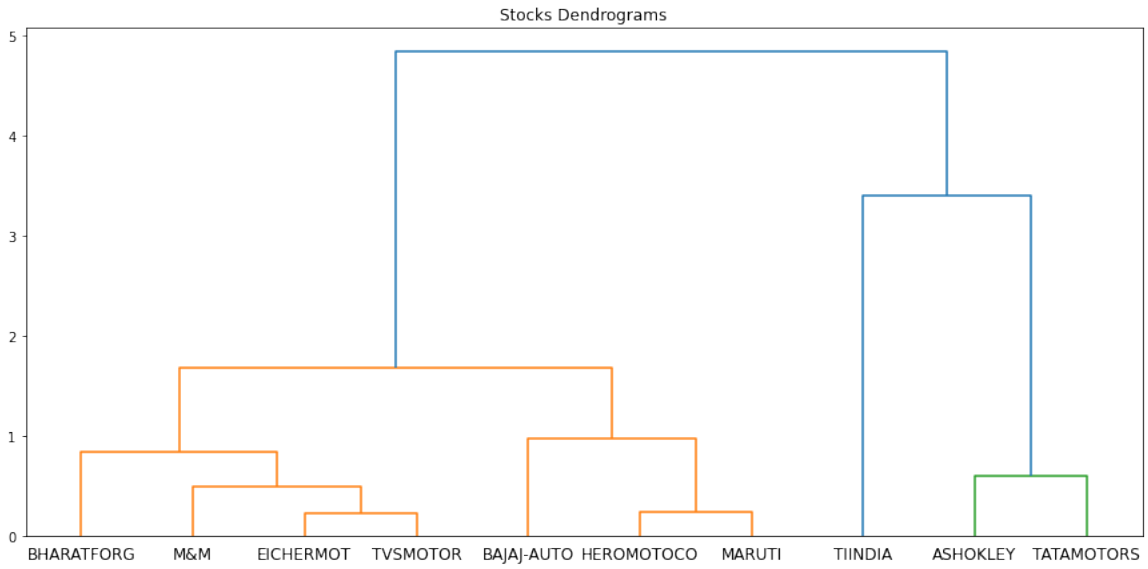

Fig. 3.12: Stocks Dendrogram

Three clusters are formed as can be distinguished by the distinct colors. Further to visualize the clusters, Agglomerative Clustering() and scatter-plot was used specifying the number of clusters as 3. The plot looked like

Similar to the plot of k-means clustering, we see that there are some distinct clusters separated by different colors.

Affinity Propagation – Affinity Propagation() module has been imported from sklearn.cluster package in python in order to perform the clustering.

Similar to the plot of k-means clustering, we see that there are some distinct

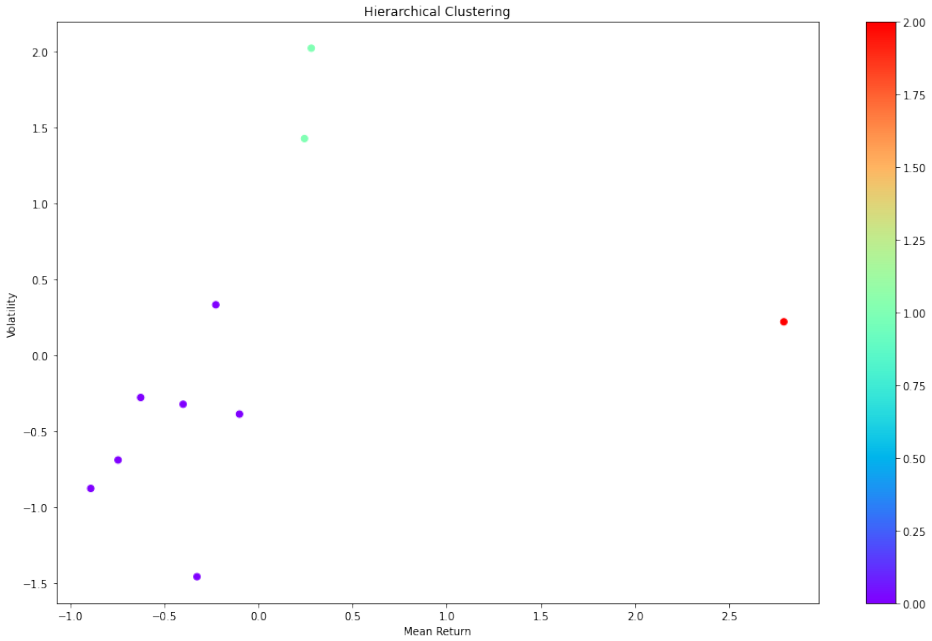

Fig. 3.13: Hierarchical Clustering

clusters separated by different colors.

Bar chart has been used to visualize the number of stocks in each cluster. There are 7,2 and 1 stocks in cluster 0,1, and 2 respectively.

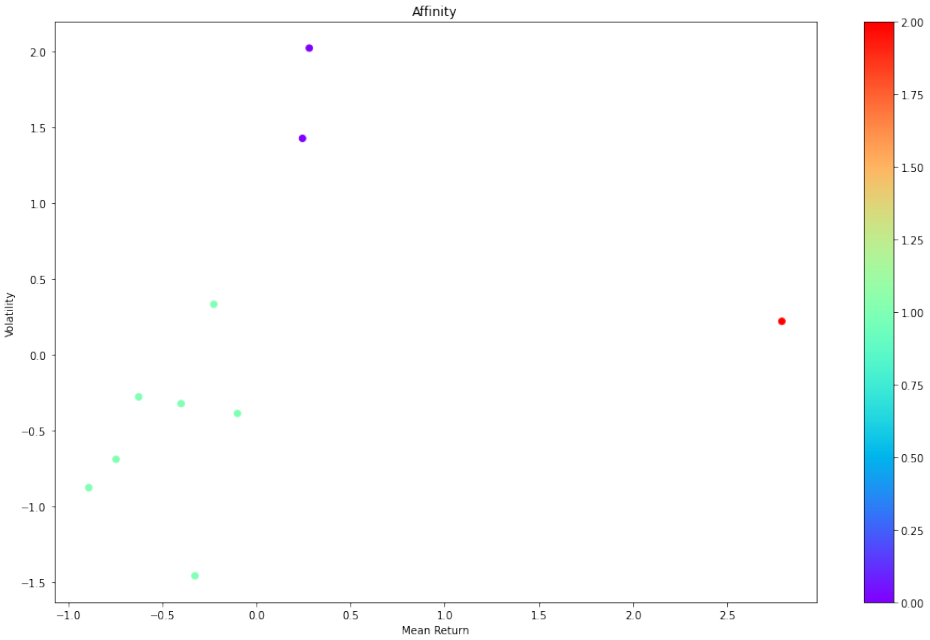

Fig. 3.14: Clusters as per Affinity Propagation

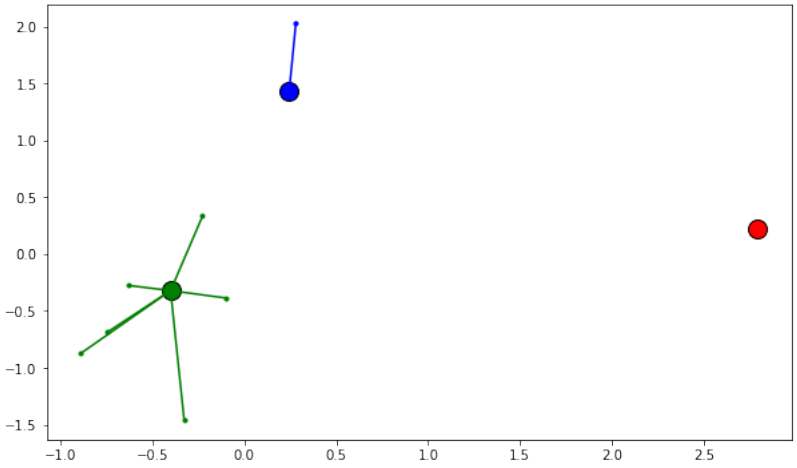

Fig. 3.15: Estimated number of clusters using AP

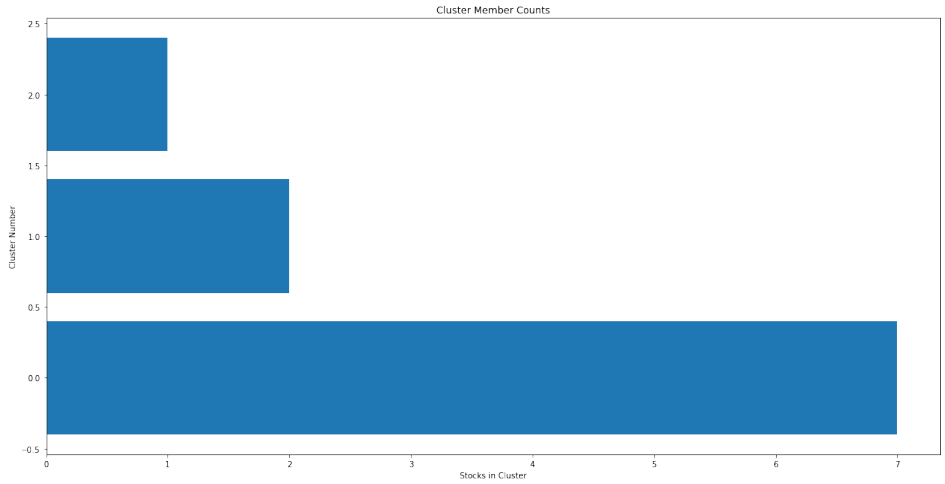

Fig. 3.16: Cluster member counts

Evaluation of Clusters for AUTO Stocks- The Silhouette Coefficient score for the three techniques are presented in the table below –

Tab. 3.3: Cluster output for auto sector

KMeans	0.5978
Hierarchical	0.5978
Affinity Propagation	0.5978

From the silhouette scores above, we proceed with Affinity Propagation clustering as the score is highest for Affinity Propagation. As the next step, we proceed with forming the stock pairs from different clusters as shown below. All the stock pairs from all the clusters are depicted below-

For cluster 0 stock pairs are-

- ASHOKLEY,TATAMOTORS

For Cluster 1 stock pairs are-

- MARUTI, TVSMOTOR

- M&M, TVSMOTOR
- Ma&M, MARUTI
- HEROMOTOCO, TVSMOTOR
- HEROMOTOCO, MARUTI
- HEROMOTOCO, M&M
- EICHERMOT, TVSMOTOR
- EICHERMOT, MARUTI
- EICHERMOT, M&M
- EICHERMOT, HEROMOTOCO
- BHARATFORG, TVSMOTOR
- BHARATFORG, MARUTI
- BHARATFORG, M&M
- BHARATFORG, HEROMOTOCO
- BHARATFORG, EICHERMOT
- BAJAJ-AUTO, TVSMOTOR
- BAJAJ-AUTO, MARUTI
- BAJAJ-AUTO, M&M
- BAJAJ-AUTO, HEROMOTOCO
- BAJAJ-AUTO, EICHERMOT
- BAJAJ-AUTO, BHARATFORG

For Cluster 2 stock pairs are-There are no pairs formed in cluster 2 as there is only one stock present in cluster 2.

Cluster Evaluation using K-Means

Here we observe how different stocks in a cluster vary with time. It is noticeable that the stocks belonging to one cluster show similar behavior.

For Cluster 0

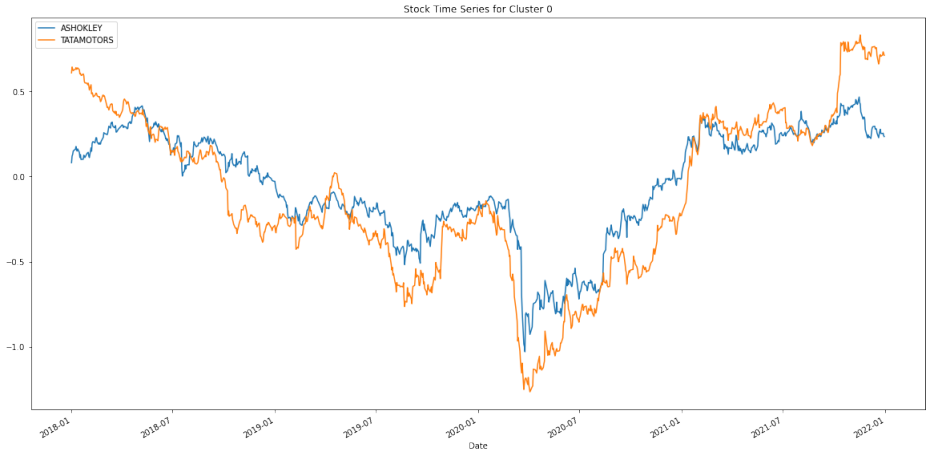

Fig. 3.17: Stock time series for cluster 0

For Cluster 1

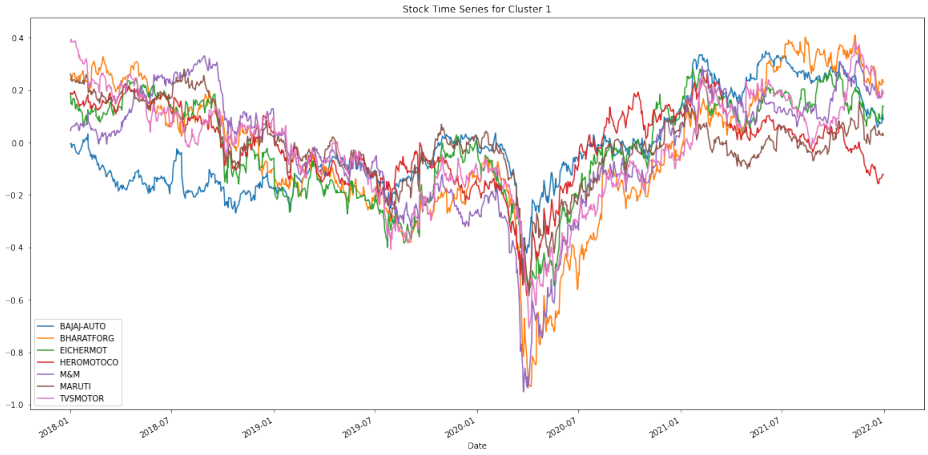

Fig. 3.18: Stock time series for cluster 1

For Cluster 2

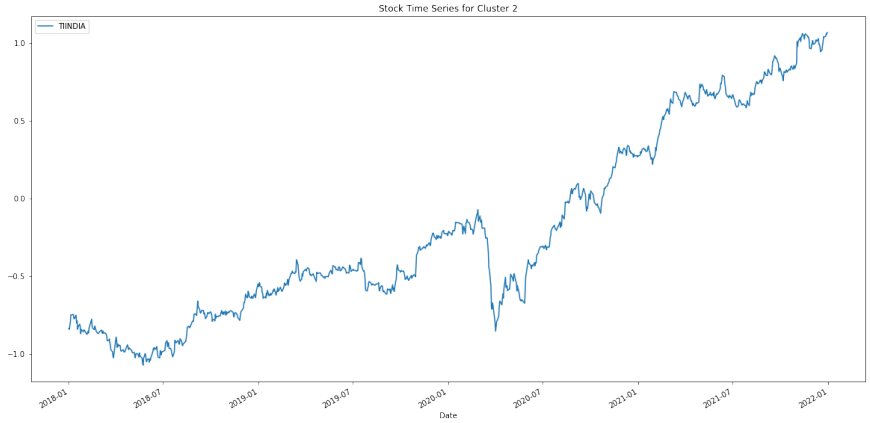

Fig. 3.19: Stock time series for cluster 2

The purpose of clustering was to verify the results of co-integration which means we must look for the pairs of stocks formed common in co-integration and clustering. Following we the common pairs of stock for NIFTY-50 stocks are-

[('BHARATFORG', 'EICHERMOT'), ('BHARATFORGE', 'TVSMOTOR'), ('EICHERMOT', 'MandM')]

Common Pair Visualization using TSNE -

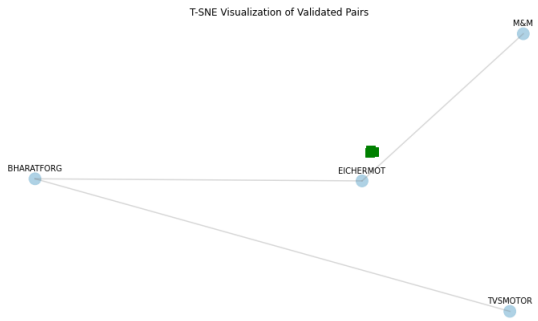

Fig. 3.20: TSNE visualization of validated pairs

3.2 IT-Sector

3.2.1 Co-integration

Information Technology (IT) sector:

As per the report published by the NSE on December 31, 2022, the ten most important stocks of the IT are as follows: Tata Consultancy Services, Infosys, HCL Technologies, Wipro, Tech Mahindra, LTIMindTree, Persistent Systems, MphasiS, Coforge and L&T Technology Services (NSE Website). 3.21 depicts the matrix of the p-values of the co-integration tests for each pair of IT sector stocks. As depicted in 3.21, three pairs exhibited significant p-values for their co-integration. These three pairs are as follows:

- HCLTECH.NS,INFY.NS
- INFY.NS, TCS.NS
- LTTS.NS,TECHM.NS

The results for the pair HCL Technologies – Infosys are presented in detail. Figure 1-9 exhibits the close values of the HCL and the Infosys series from January 1, 2018, to December 31, 2021. The results of the OLS regression model built with HCL Technologies as the predictor and Infosys as the target are presented in 3.23. The hedge ratio is 1.4341. The highly significant p-value of the F-statistics indicates that there is a strong linear relationship between the two series.

The OLS regression results with Bharat Forge as the target and Ashok Leyland as the predictor variable are presented in Figure 1-3. The hedge ratio (i.e., the coefficient parameter for Ashok Leyland) is found to be 5.6851 with a standard error of 0.017. The highly significant p-value of the F-statistic indicates that there is a strong linear relationship between the two series. The Jarque-Bera test yielded a weakly significant p-value for the test statistic indicating the skewness and the kurtosis of the residuals of the model are not like those of a normal distribution. The Durbin-Watson test statistic lies

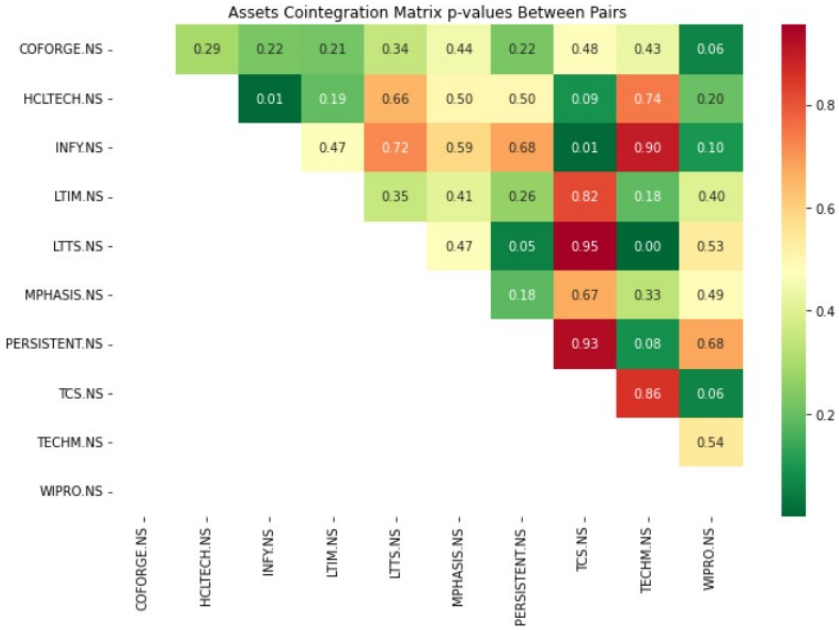

Fig. 3.21: The matrix of the p-values of the cointegration tests for the pairs of stocks of the IT sector.

between 0 and 2 indicating a positive autocorrelation among the residual. Finally, the Omnibus test results support the p-value of the F statistic indicating the presence of a strong linear relationship between the target and the predictor. In the next step, the residuals of the model are computed and plotted. Figure 1-4 depicts the residual plot which looks like a stationary series.

The residuals of the OLS model are plotted in Figure 1-11. The Z-values of the ratio with their upper and lower bounds are depicted in 3.25

The ADF test on the ratio series yielded a test statistic value of -3.0944, while the critical value at the 1% level of significance is -3.4392. The ADF test results indicate that the residual series is not stationary. However, the pair trading strategy is executed since the two series were found to be cointegrated. The signal points of the pair trading during 2022 are depicted in 3.26. 3.27 exhibits the daily portfolio value for the pair for the year 2022. The total portfolio value for the HCL Technologies – Infosys pair at the end

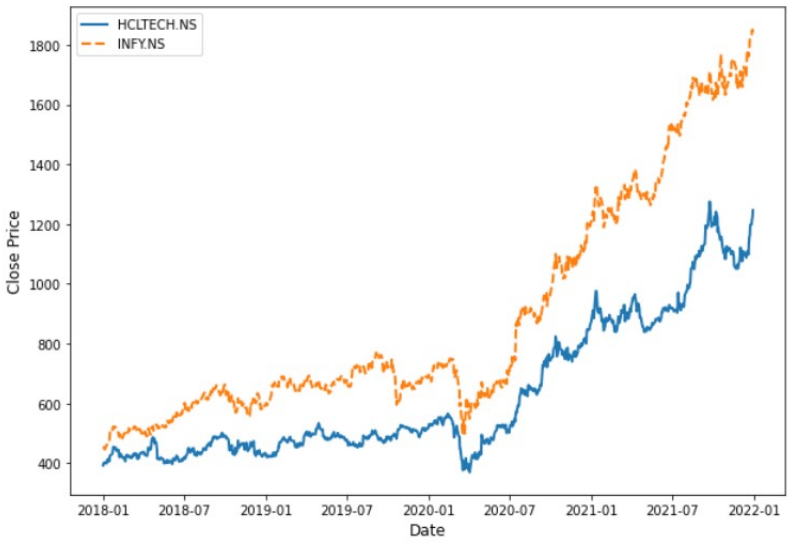

Fig. 3.22: The plot depicting the daily close prices of the stocks HCL Technologies (HCLTECH.NS) and Infosys (INFY.NS), from January 1, 2018, to December 31, 2021.

of 2022 is 440109.19. For an initial investment of 200000, the pair yields a return of 10.03%. Table 1-2 presents the portfolio return for the three pairs from the IT sector. All the pairs have yielded positive returns with ('LTTS.NS', 'TECHM.NS') pair producing the highest return.

OLS Regression Results

Dep. Variable:	asset2	R-squared (uncentered):	0.994
Model:	OLS	Adj. R-squared (uncentered):	0.994
Method:	Least Squares	F-statistic:	1.546e+05
Date:	Mon, 13 Mar 2023	Prob (F-statistic):	0.00
Time:	09:35:24	Log-Likelihood:	-5688.2
No. Observations:	987	AIC:	1.138e+04
Df Residuals:	986	BIC:	1.138e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
asset1	1.4341	0.004	393.136	0.000	1.427	1.441

Omnibus:	160.360	Durbin-Watson:	0.036
Prob(Omnibus):	0.000	Jarque-Bera (JB):	252.373
Skew:	1.079	Prob(JB):	1.58e-55
Kurtosis:	4.218	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 3.23: The results of the OLS regression model with Infosys as the target variable (denoted by asset2) and HCL Technologies as the predictor variable (denoted by asset1).

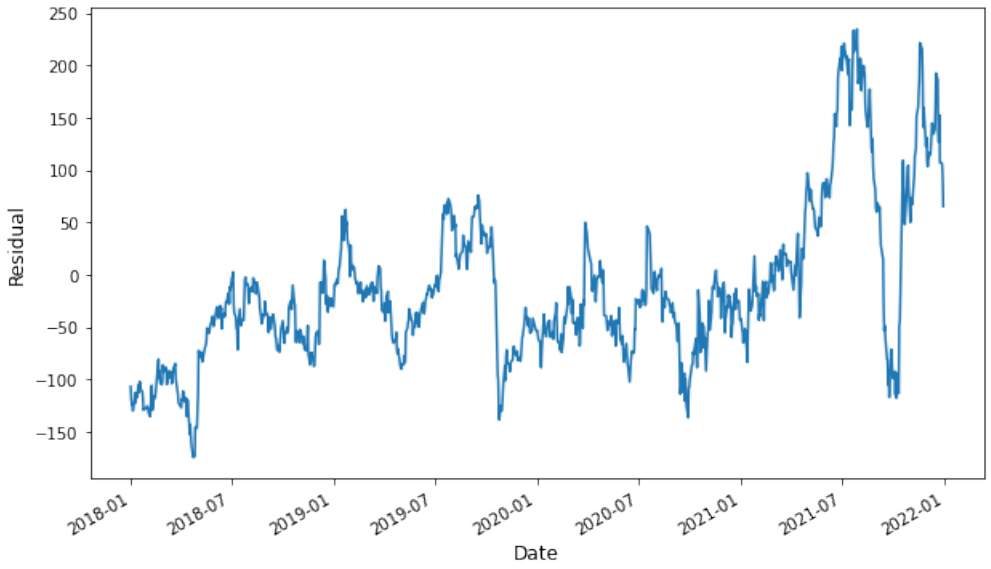

Fig. 3.24: The residual plot for the OLS regression model with HCL Technologies as the predictor and Infosys as the target variable.

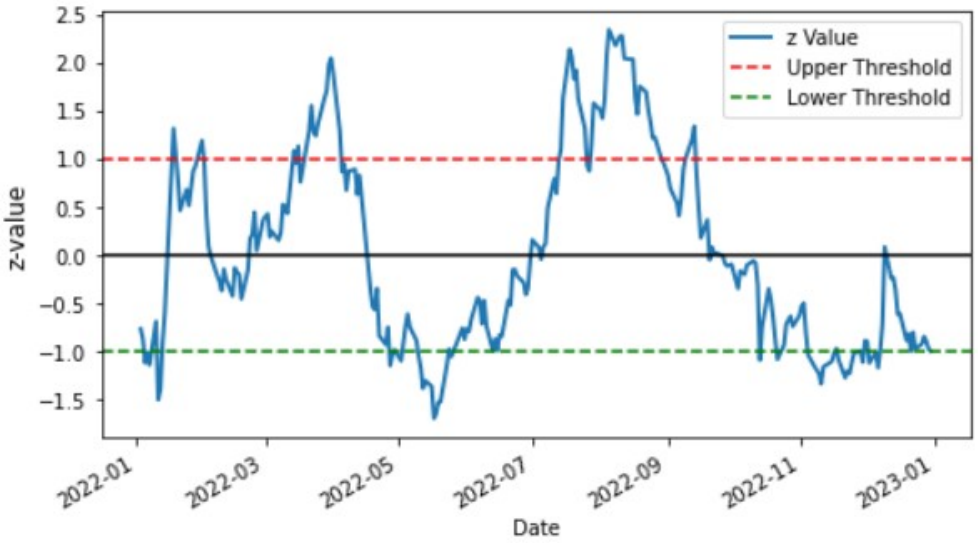

Fig. 3.25: The plot of the Z-values of the ratio series for the HCL Technologies – Infosys pair and their upper and lower limits.

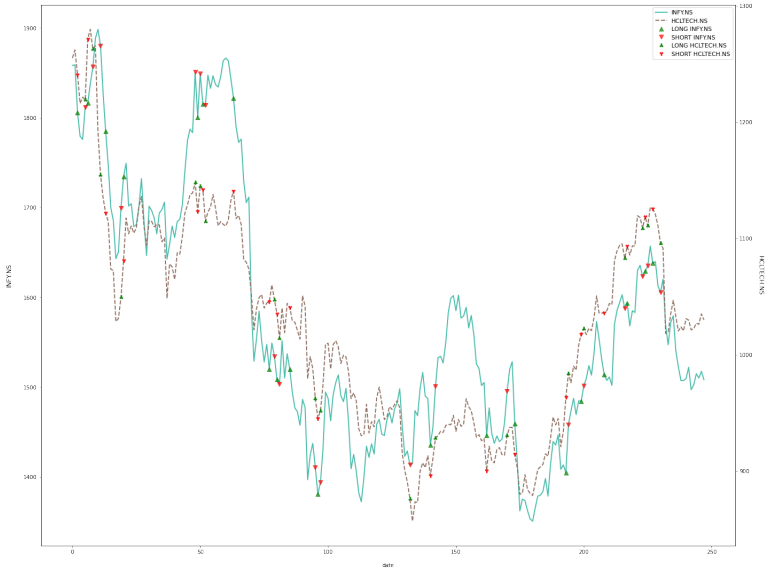

Fig. 3.26: The pair-trading scenarios for the stocks HCL Technologies and Infosys – the trading signals and their positions identified.

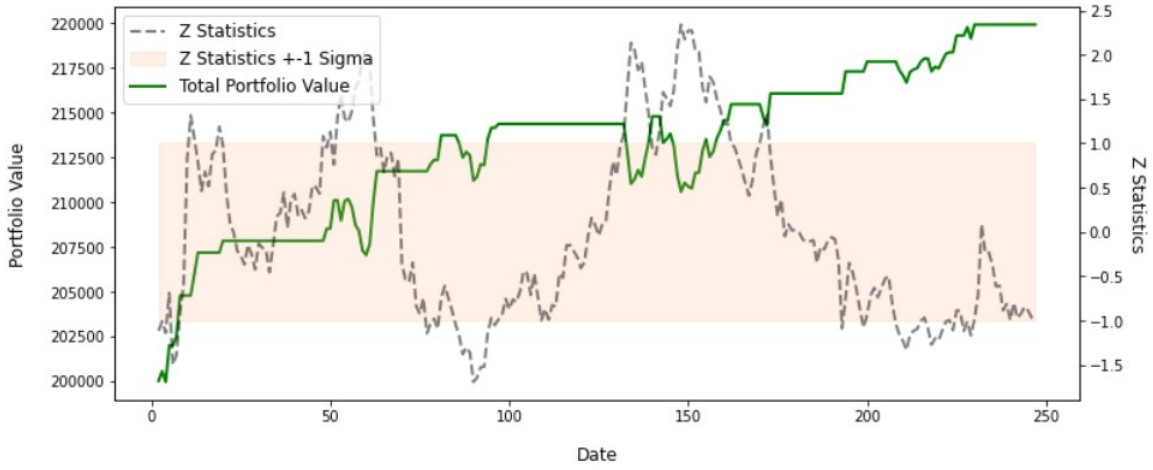

Fig. 3.27: The plot of the daily value of the pair-trading portfolio of HCL Technologies and Infosys from January 1, 2022, to December 31, 2022.

Tab. 3.4: THE ANNUAL RETURNS OF THE PAIR-TRADING PORTFOLIOS OF THE IT SECTOR STOCKS (PERIOD: JANUARY 1, 2022 - DECEMBER 31, 2022).

Stock pairs	Initial Investment	Final value	**Profit**	Return
HCLTECH INFY	200000	440109.19.02	40109.19	10.03
INFY TCS	200000	425684.03	25684.03	6.42
LTTS TECHMAHINDRA	200000	470378.45	70378.45	17.59

3.2.2 CLUSTERING

IT Sector stocks- Based on the NSE's report published on December 31, 2022, the fifty stocks that have the most significant contributions to the overall index of this sector are as follows: Tata Consultancy Services Ltd., Infosys Ltd., HCL Technologies Ltd., Wipro Ltd., Tech Mahindra Ltd., LTIMindtree Ltd., Persistent Systems Ltd., MphasiS Ltd., Coforge Ltd., L&T Technology Services Ltd. from (NSE Website).

K-means clustering – KMeans module has been imported from sklearn.cluster package in python in order to perform the clustering. To find the optimum number of clusters in k-means technique we look at the silhouette score for different values of K (no. of clusters). The y-axis denotes silhouette scores and x-axis denotes the number of cluster(k values)

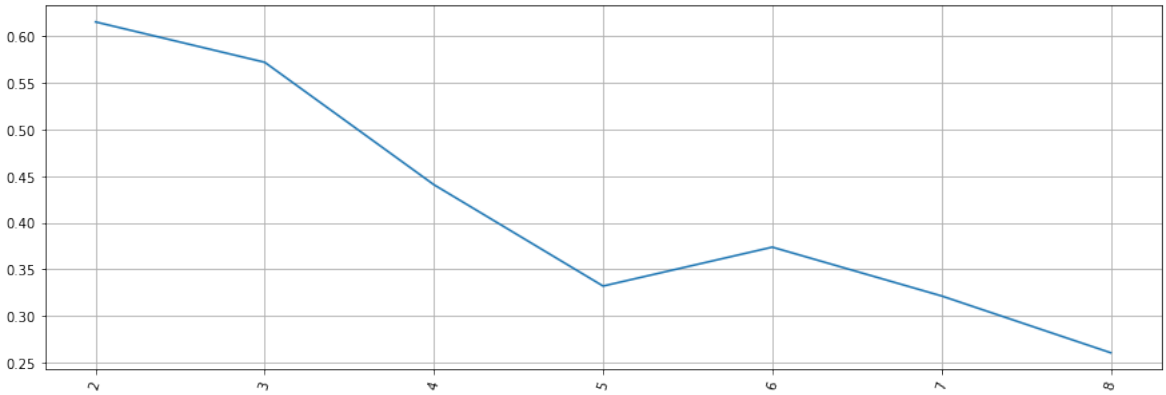

Fig. 3.28: Silhouette score vs number of clusters

The silhouette score maximizes at k=2. Hence, number of clusters chosen are 2.

Clusters were be visualized using 'k-means cluster centers' and scatter plot. The cluster centroids could be seen as green squares, and the number of stocks associated with each cluster could be seen as circles around squares with distinct colors. We can clearly see two clusters.

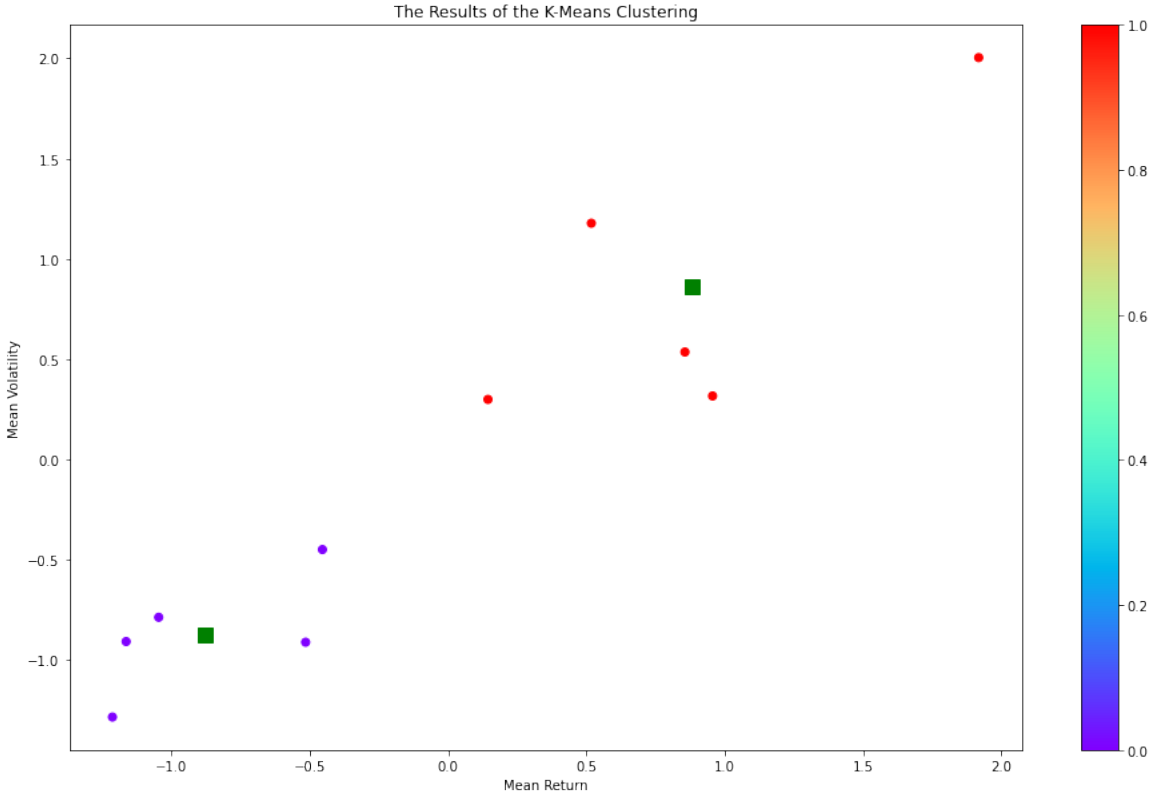

Fig. 3.29: The results of K-means clustering

Number of stocks in each cluster can be further visualized using bar chart -

As per the above chart, there are 5, 5 stocks in clusters 0 and 1 respectively.

Hierarchical clustering – The best way to visualize an agglomerate clustering algorithm is through a dendrogram, which displays a cluster tree, the leaves being the individual stocks and the root being the final single cluster. The "distance" between each cluster is shown on the y-axis, and thus the longer the branches are, the less correlated two clusters are. In order to perform this

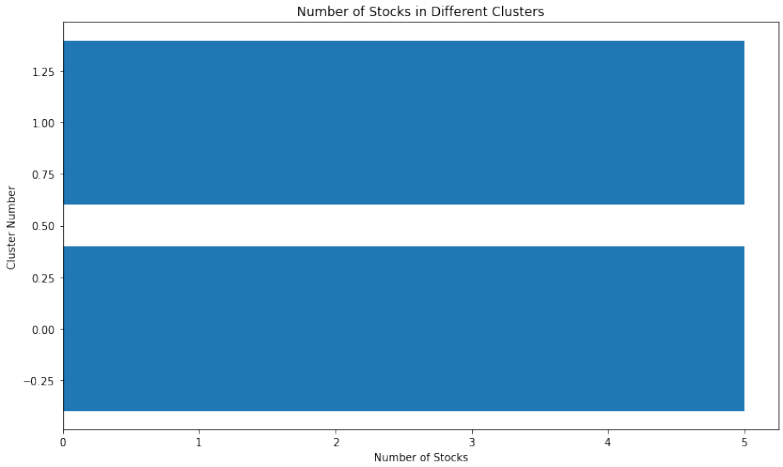

Fig. 3.30: Number of stocks in different clusters

clustering and visualize dendrogram, linkage, ward, and AgglomerativeClustering modules were imported.

For IT Sector the dendrogram plot looked like –

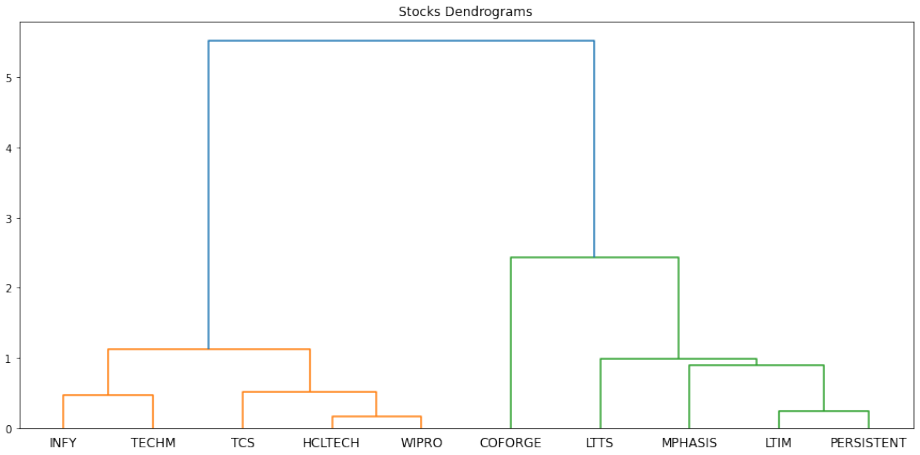

Fig. 3.31: Stocks dendogram

Three clusters are formed as can be distinguished by the distinct colors. Further to visualize the clusters, Agglomerative Clustering() and scatter-plot was used specifying the number of clusters as 3. The plot is given in 3.32

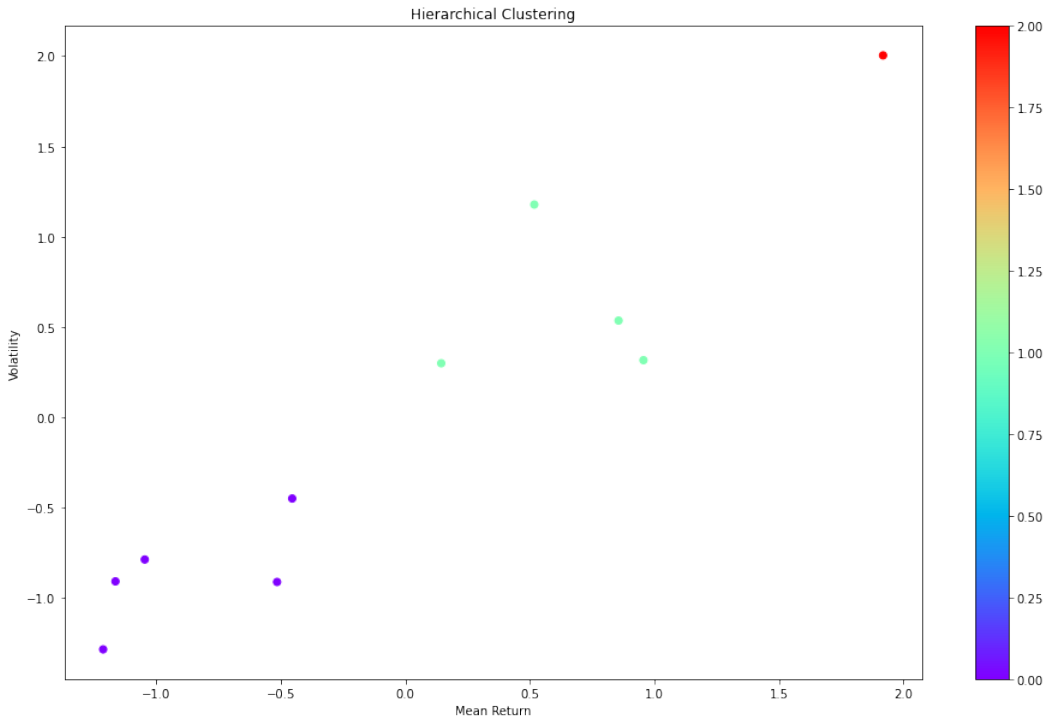

Fig. 3.32: Hierarchical Clustering

Similar to the plot of k-means clustering, we see that there are some distinct clusters separated by different colors.

Affinity Propagation – Affinity Propagation() module has been imported from sklearn.cluster package in python in order to perform the clustering.

Similar to the plot of k-means clustering, we see that there are some distinct clusters separated by different colors.

Bar chart has been used to visualize the number of stocks in each cluster. There are 5,4 and 1 stocks in cluster 0,1, and 2 respectively.

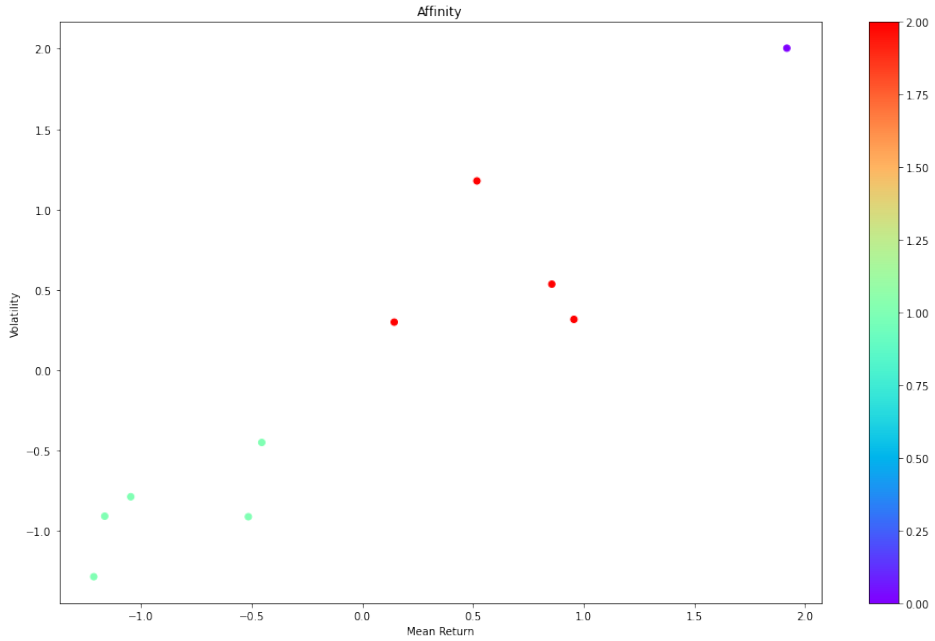

Fig. 3.33: Affinity Propagation

Evaluation of Clusters for NIFTY-50 Stocks- The Silhouette Coefficient score for the three techniques are presented in the table below –

Tab. 3.5: Cluster output for auto sector

K-Means	0.6151
Hierarchical	0.5718
Affinity Propagation	0.5718

From the silhouette scores above, we proceed with Affinity Propagation clustering as the score is highest for K-Means Clustering. As the next step, we proceed with forming the stock pairs from different clusters as shown below. All the stock pairs from all the clusters are depicted below-

For Cluster 0 stock pairs are-

- TECHM, WIPRO
- TCS, WIPRO

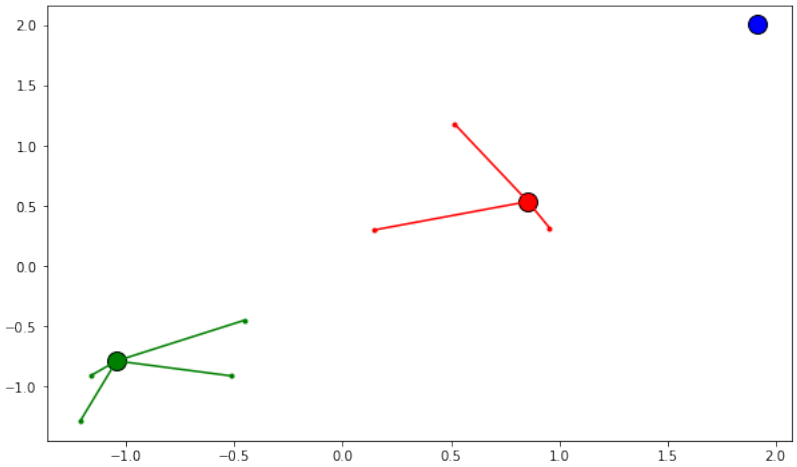

Fig. 3.34: Estimated number of clusters using AP are 3

- TCS, TECHM
- INFY, WIPRO
- INFY, TECHM
- INFY, TCS
- HCLTECH, WIPRO
- HCLTECH, TECHM
- HCLTECH, TCS
- HCLTECH, INFY

For Cluster 1 stock pairs are-

- MPHASIS, PERSISTENT
- LTTS, PERSISTENT
- LTTS, MPHASIS

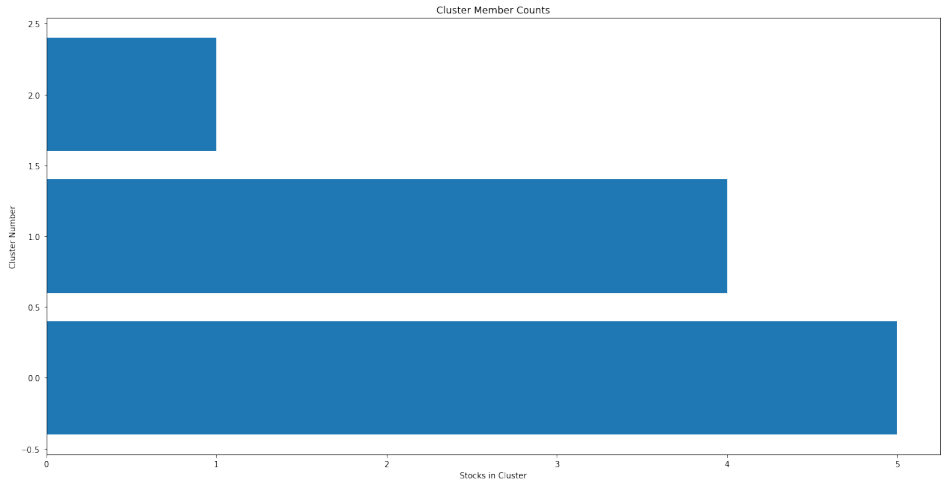

Fig. 3.35: There are 5,4 and 1 stocks in cluster 0,1, and 2 respectively

- LTIM, PERSISTENT
- LTIM, MPHASIS
- LTIM, LTTS
- COFORGE, PERSISTENT
- COFORGE, MPHASIS
- COFORGE, LTTS
- COFORGE, LTIM

Cluster Evaluation using K-Means

Here we observe how different stocks in a cluster vary with time. It is noticeable that the stocks belonging to one cluster show similar behavior.

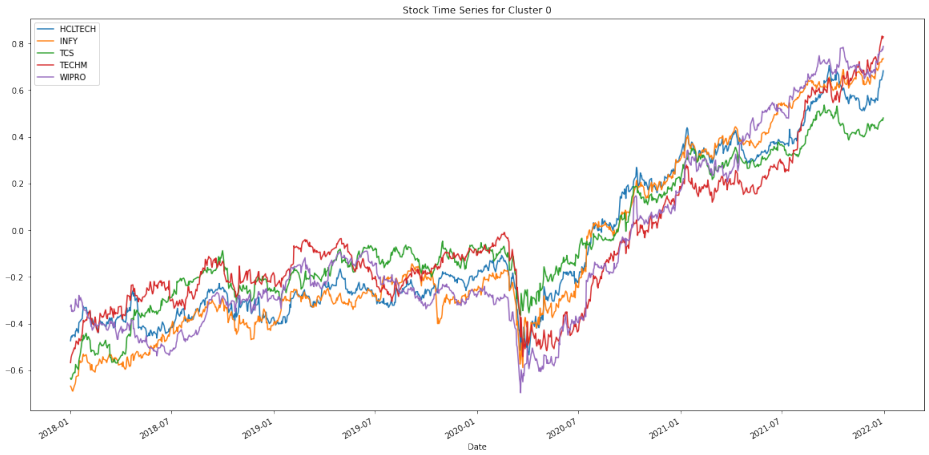

Fig. 3.36: Stock Time-series for cluter 0

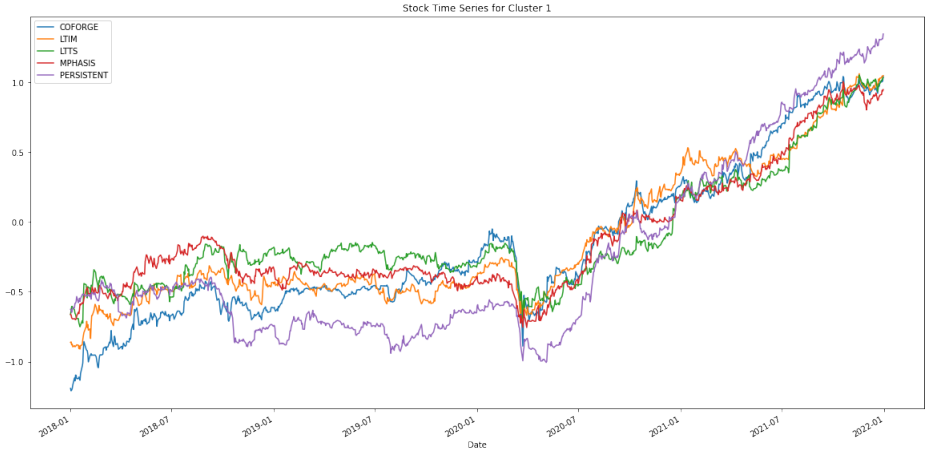

Fig. 3.37: Stock Time-series for cluter 1

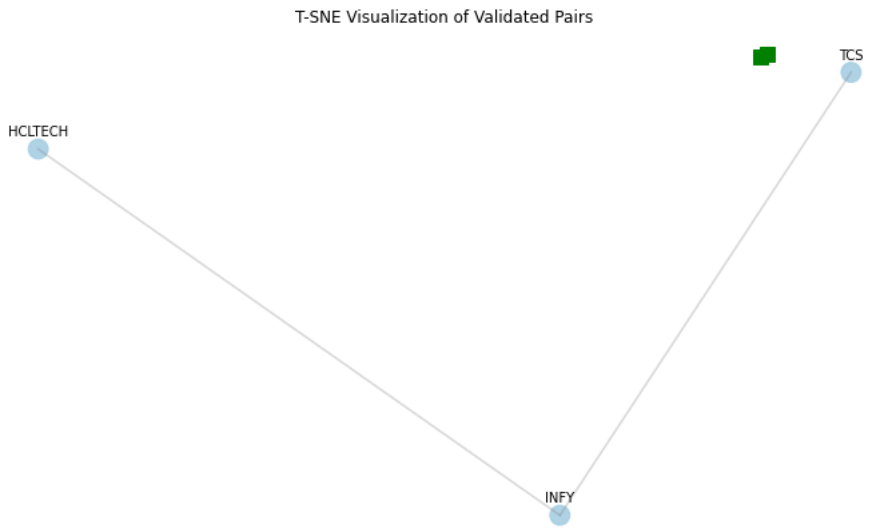

Fig. 3.38: Common Pair Visualization using TSNE

The purpose of clustering was to verify the results of co-integration which means we must look for the pairs of stocks formed common in co-integration and clustering. Following are the common pairs of stock for NIFTY-50 stocks are-

[('HCLTECH', 'INFY'), ('INFY', 'TCS')]

4. CONCLUSION AND FUTURE SCOPE OF WORK

The results are concluded and summarized below

1. The analysis done here was presented for 4 year time period which was chosen on elaborate analysis. The testing period encompassed whole of year 2022. 4 years prior to this was taken as the training period.x
2. The correlated pairs with pearson correlation greater than 0.5 were taken as correlated stock pairs. For these pairs for 11 different sectors the analysis was done on profit/loss calculation. For example as given above, there are 11 correlated stock pairs, of which 2 exhibited negative ROI. Others gave a decent return. But, this may not always happen. For reference IT sector showcased 6 correlated stock pairs of which 2 gave negative stock pairs
3. In next step the Co-integrated stock pairs were identified. Co-integration looks into long term correlation but very less or no correlation in the short period of time. With the Co-integrated stock pairs identified, the same profit loss calculation was done. For two sectors - Auto and IT the final portfolio value and ROI were presented. It was thus observed that even though Co-integrated stock pairs exhibited negative ROIs. This can be attributed to the stochastic nature of stock market. The training period and testing may exhibit completely differing nature and as such stock pairs that would have been Co-integrated in the training may not exhibit same behaviour in the testing period
4. Analysis of 11 sectors revealed the following things.
    - 4 Sectors - Oil & Gas, Pharma, Media and IT exhibited all positive returns. Oil & gas gave 2 stocks both of which were giving

ROI of greater than 20. Media and IT gave 4 and 3 stock pairs respectively in the given training and testing period.

- All the remaining sectors gave some negative ROI stock pairs too
- Moreover in Metal sector Hindalco and Tata Steel showcased a negative ROI of 24 which can be considered for future analysis.
- Similarly, in FMCG Dabur and VBL exhibited the same behaviour.
- Apart from all these sectors Nifty50 was also considered. And there were 57 Co-integrated pairs of which 11 gave negative ROI, rest all gave positive ROI.

5. Another observation to shed light on is that the number of signals generated in a given year were 10-15. Which means that the window of investment opportunities were limited to about 1 in a month.

6. Finally, a clustering based Co-integration approach was also used to arrive at more reliable pair of stocks. This way the stock pairs were backed by the clustering.

7. The stock pairs identified as exhibiting erroneous behaviour can be taken for further in-depth analysis which can shed more light on the stock market as well.

8. The study can be extended to inter-sectoral analysis as well global stock markets.

9. With the knowledge of stock pairs a reliable investment strategy can be devised by integrating predictive algorithms like LSTM based predictive models.

5. BIBLIOGRAPHY

1. Brim, A. (2020) "Deep reinforcement learning pairs trading with a double deep Q-network", Proceedings of the 10th Annual Computing and Communication Workshop and Conference (CCWC'20), pp. 222-227, January 6-8, 2020, Las Vegas, Nevada, USA. DOI: 10.1109/CCWC47524.2020.9031159

2. Chatterjee, A., Bhowmick, H. and Sen, J. (2021) "Stock price prediction using time series, econometric, machine learning, and deep learning models", Proceedings of the IEEE Mysore Sub Section International Conference (MysuruCon'21), pp. 289-296, October 24-25, Hassan, Karnataka, India. DOI: 10.1109/MysuruCon52639.2021.9641610.

3. Cheng, D., Liu, Y., Niu, Z. and Zhang, L. (2018) "Modeling similarities among multi-dimensional financial time series", IEEE Access, Vol. 6, pp. 43404-43414. DOI: 19.1109/ACCESS.2018.2862908.

4. Chong, E., Han, C. and Park, F. C. (2017) "Deep learning networks for stock market analysis and prediction: Methodology, data", Experts Systems with Applications, Vol. 85, pp. 187-205, October 2017. DOI: 10.1016/j.eswa.2017.04.030.

5. Corazza, M., di Tollo, G., Fasano, G. and Pesenti, R. (2021) "A novel hybrid PSO-based metaheuristic for costly portfolio selection problem", Annals of Operations Research, Vol. 304, pp. 109-137. DOI: 10.1007/s10479-021-04075-3.

6. Engle, R.F. and Granger, C. W. J. (1987) "Co-integration and error correction: Representation, estimation, and testing", Econometrica, Vol. 55, No. 2, pp. 251–276. DOI: 10.2307/1913236.

7. Fengqian, D. and Chao, L. (2020) "An adaptive financial trading system using deep reinforcement learning with candlestick decomposing features", IEEE Access, Vol. 8, pp. 63666–63678. DOI: 10.1109/ACCESS.2020.2982662.

8. Gupta, K. and Chatterjee, N. (2020) "Selecting stock pairs for pairs trading while incorporating lead-lag relationship", Physica A: Statistical Mechanics and its Applications, Vol. 551, Art ID. 124103. DOI: 10.1016/j.physa.2019.124103.

9. Hung, C-C. and Chen, Y-J. (2021) "DPP: Deep predictor for price movement from candlestick charts", PloS ONE, Vol. 16, No. 6, Art ID. e0252404, June 2021. DOI: 10.1371/journal.pone.0252404.

10. Karimi, M., Tahayori, H., Tirdad, K. and Sadeghian, A. (2022) "A perceptual computer for hierarchical portfolio selection based on interval type-2 fuzzy sets", Granular Computing, 2022. DOI: 10.1007/s41066-021-00311-0.

11. Kim, T. and Kim, H. Y. (2019) "Optimizing the pair-trading strategy using deep reinforcement learning with trading and stop-loss boundaries", Complexity in Financial Markets, Vol. 2019, Art ID. 3582516. DOI: 10.1155/2019/3582516.

12. Lei, K., Zhang, B., Li, Y., Yang, M. and Shen, Y. (2020) "Time-driven feature-aware jointly deep reinforcement learning for financial signal representation and algorithmic trading", Expert Systems with Applications, Vol. 140, Art ID. 112872. DOI: 10.1016/j.eswa.2019.112872.

13. Mehtab, S. and Sen, J. (2021) "A time series analysis-based stock price prediction using machine learning and deep learning models", International Journal of Business Forecasting and Marketing Intelligence, Vol. 6, No. 4, pp. 272-335. DOI: 10.1504/IJBFMI.2020.115691.

14. Rad, H., Kwong, R., Low, Y. and Faff, R. (2016) "The profitability of pairs trading strategies: Distance, cointegration and copula methods", Quantitative Finance, Vol. 16, No. 10, pp. 1541-1558. DOI: 10.1080/14697688.2016.1164337.