Distance Based and Machine Learning Methods for Pairs Trading

A thesis submitted in fulfilment of the requirements for the degree of

Master of Science in Global Finance and Banking

by

Partha Pratim Sharma

Department of Banking and Finance

King's College London

October 16, 2024

Abstract

This research presents a comprehensive analysis of pairs trading strategies using four distinct methods: Distance Method, K-Means Clustering, DBSCAN Clustering, and OPTICS Clustering. Each method's performance was evaluated based on its ability to identify profitable pairs, optimize risk-adjusted returns, and manage transaction costs effectively. The study's findings reveal that the K-Means method consistently outperformed the other methods in terms of cumulative returns and risk-adjusted metrics. It achieved the highest mean cumulative return and superior Sharpe and Sortino ratios, establishing it as the most effective strategy for maximizing returns while managing risk. The identified pairs were tested using a mean-reversion strategy, with their performance assessed based on key metrics such as cumulative return, Sharpe ratio, Sortino ratio, and maximum drawdown. The analysis also considers the impact of transaction costs on strategy profitability, offering insights into the robustness of each method for pairs trading.

Acknowledgements

I would like to thank my supervisor, Dr. Fotis Papailias, for his invaluable guidance and support throughout this research. Special thanks to my wife, my son, my parents, and my friends for their encouragement and unwavering support.

I would also like to express my gratitude to Hudson & Thames for providing the ArbitrageLab Python library, which played a significant role in helping me develop my independent software used in this study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

1.1.1 Background on Pairs Trading

Pairs trading is a market-neutral trading strategy that involves identifying two securities with a historically correlated price relationship and betting on their mean spread reverting. The core idea is to take a long position on the undervalued asset while simultaneously shorting the overvalued one, expecting their prices to converge. This technique leverages the concept of mean reversion, where temporary price deviations between the two securities are exploited for profit as they return to their historical equilibrium.

Initially developed by quantitative analysts in the 1980s, pairs trading has evolved to incorporate more sophisticated approaches, including statistical measures, cointegration tests, and machine learning techniques. These advancements have made pairs trading a cornerstone strategy in statistical arbitrage, allowing traders to build market-neutral positions that are less sensitive to overall market movements.

1.1.2 Research Objective

The primary objective of this research is to evaluate and compare different methodologies for selecting trading pairs from a universe of stocks within the S&P 500 index. Specifically, the study aims to analyze the effectiveness of various approaches, including the distance method, K-Means clustering, DBSCAN, and OPTICS, in identifying pairs suitable for a mean-reversion strategy. By applying these techniques, the research seeks to determine which method provides the most robust and reliable pairs for trading, thereby enhancing the overall profitability and risk management of pairs trading strategies.

Furthermore, this study aims to contribute to the understanding of how machine learning techniques can be integrated with traditional statistical methods to improve pairs selection in financial markets. The analysis will provide insights into the potential benefits and limitations of each method, helping traders and researchers to refine their strategies for pairs trading.

Chapter 2

Literature Review

2.1 Literature Review

The field of statistical arbitrage and pairs trading has seen significant advancements in the past few decades. The literature can be broadly classified into several approaches, including the distance approach, cointegration approach, time series approach, stochastic control approach, and other emerging techniques. In this study, we focus on the distance approach, the cointegration approach, and machine learning methods, as they represent a diverse set of strategies for implementing pairs trading with varying levels of complexity and robustness. As discussed in the report on pairs trading by Hudson & Thames (2023) [12], the strategies and methodologies covered provide a comprehensive understanding of the domain.

2.1.1 Distance Approach

The distance approach, popularized by the seminal work of Gatev, Goetzmann, and Rouwenhorst (2006) [9], is one of the earliest and most straightforward methodologies in pairs trading. This approach uses nonparametric distance metrics, specifically Euclidean distance, to identify pairs of securities whose prices historically move together. The core idea is to find pairs with the smallest distance between their normalized price series over a historical period, implying a close relationship that might suggest potential trading opportunities.

During the trading phase, positions are opened when the price spread between the two selected securities diverges beyond a predefined threshold, under the assumption that the spread will eventually revert to its historical mean. Once the spread reverts, the position is closed to capture the profit. The simplicity of this approach, alongside its model-free nature, makes it easy to implement and relatively robust against various market conditions.

Gatev et al. (2006) conducted a comprehensive empirical analysis of pairs trading strategies using this method, finding that it could generate annualized returns of up to 11%, even after accounting for transaction costs. Their study demonstrated that this strategy had relatively low exposure to market risk and highlighted its potential as a form of statistical arbitrage that could consistently deliver positive returns over different market cycles. The distance approach thus became a foundational strategy, widely adopted by both academic researchers and financial practitioners.

Building on this foundational work, Do and Faff (2010, 2012) [7, 6] expanded the distance approach by addressing several of its limitations. They emphasized the critical role of transaction costs in pairs trading profitability, noting that ignoring these costs could lead to an overestimation of expected returns. Do and Faff proposed modifications to the pairs selection criteria, such as incorporating industry classification and accounting for the frequency of zero-crossings in the spread to enhance the robustness of the strategy. Their analysis indicated that these improvements could lead to more stable performance, even in markets where mean-reversion signals are less pronounced.

Despite its simplicity and intuitive appeal, the distance approach has faced some criticism. As subsequent studies by Chen et al. (2012) [5] and Huck (2015) [10] have pointed out, distance-based methods can sometimes result in the selection of pairs that exhibit low spread variance, limiting the profitability potential. These studies suggest that while the distance approach is a useful starting point, it may be beneficial to incorporate more sophisticated techniques, such as cointegration tests, to explicitly model the long-term equilibrium relationships between securities and enhance the reliability of pairs trading signals.

Overall, the distance approach remains one of the most extensively researched frameworks for pairs trading due to its simplicity, ease of implementation, and proven effectiveness. It serves as a benchmark against which more advanced and computationally intensive methodologies, such as cointegration and machine learning-based techniques, are compared.

2.1.2 Cointegration Approach

The cointegration approach relies on formal econometric tests to identify pairs of securities whose price spread forms a stationary time series, indicating that while the individual prices of the securities may drift apart, their spread will eventually revert to a mean. This mean-reverting property is the core concept behind pairs trading using cointegration, as it allows traders to profit from temporary deviations from this equilibrium relationship.

Vidyamurthy (2004) [13] laid the groundwork for the cointegration-based approach by presenting a systematic method for identifying mean-reverting pairs through econometric tests. His work emphasized the significance of using cointegration to ensure that pairs of assets have a statistically validated relationship, rather than merely relying on historical price correlations. Vidyamurthy's approach suggested that by focusing on cointegrated pairs, traders could significantly reduce the risk of entering trades based on spurious correlations, which might lead to unprofitable positions.

Building on Vidyamurthy's foundational work, Caldeira and Moura (2013) [3] further refined the cointegration-based strategy by employing both the Engle-Granger two-step method and the Johansen test to select pairs in the Brazilian market. The Engle-Granger method involves testing the residuals of a linear regression between two time series for stationarity, which is a critical indicator of mean-reversion in the spread. The Johansen test, a more sophisticated approach, is capable of detecting multiple cointegrating relationships among several assets, making it suitable for more complex portfolios.

Caldeira and Moura's application of these tests in the context of the Brazilian market demonstrated robust performance, even in emerging markets characterized by higher volatility and lower market efficiency. Their study employed a ranking mechanism based on the in-sample Sharpe ratio to select the most promising pairs for trading. This enhancement not only improved the robustness of the strategy but also made it adaptable to diverse market conditions, highlighting the flexibility and scalability of the cointegration approach in pairs trading.

The advantage of the cointegration approach lies in its ability to explicitly model the long-term equilibrium relationship between securities, as opposed to merely exploiting historical price correlations. This econometric rigor allows traders to design strategies that capitalize on temporary price dislocations with a higher level of confidence in their expected mean-reversion. Additionally, cointegration provides a theoretical framework that underpins the statistical arbitrage strategy, making it more appealing for institutional investors who require a solid basis for their trading models.

Despite its strengths, the cointegration approach does come with certain challenges. It requires more sophisticated statistical analysis compared to simpler approaches like distance metrics, and the stability of cointegrated relationships may change over time, necessitating continuous monitoring and recalibration of the pairs. Additionally, while cointegration-based strategies are generally more robust, their effectiveness can still be influenced by market conditions and the presence of structural breaks in the price series.

Overall, the cointegration approach represents a significant advancement in pairs trading strategies, providing a more reliable framework for identifying trading opportunities with mean-reverting characteristics. It continues to be a popular choice among researchers and practitioners who seek to leverage the predictive power of econometric techniques to design more resilient pairs trading strategies.

2.1.3 Machine Learning Approach

The evolution of pairs trading has seen the emergence of several innovative approaches that leverage machine learning, copula models, and Principal Component Analysis (PCA) to enhance strategy performance. These methods have shown promise in addressing some of the limitations inherent in traditional approaches like distance and cointegration methods.

One significant development in this area was the PCA-based approach developed by Avellaneda and Lee (2010) [2]. They utilized PCA to identify pairs of assets that exhibit mean-reversion properties. PCA helps reduce the dimensionality of the dataset, highlighting the most critical components that influence price movements. This method enables traders to filter out noise from irrelevant data, thereby identifying more robust trading signals.

Building upon this concept, Sarmento and Horta (2020) [11] proposed a sophisticated framework that integrates PCA with the OPTICS clustering algorithm for pairs selection. By applying PCA to reduce the dimensionality of the dataset, they created a more compact representation of each asset, which was then used in the OPTICS clustering algorithm to identify potential pairs. This clustering approach helps manage the challenge of finding profitable pairs by automatically determining the number of clusters and handling varying cluster densities. Their results demonstrated that this machine learning-based approach could outperform traditional methods, achieving a higher Sharpe ratio and more stable performance even in volatile market conditions.

2.1.4 Unsupervised Clustering Methods

Unsupervised clustering methods are widely used in machine learning to identify patterns or groups in datasets without predefined labels. These methods play a crucial role in financial markets, where identifying clusters of similar assets can aid in portfolio construction, risk management, and trading strategy development. In this study, we focus on three commonly used unsupervised clustering techniques: K-Means, DBSCAN, and OPTICS.

K-Means Clustering

K-Means is a centroid-based clustering algorithm that partitions data into a predetermined number of clusters. The algorithm iteratively assigns data points to clusters by minimizing the sum of squared distances from each point to its assigned cluster's centroid. K-Means is known for its simplicity and computational efficiency, making it one of the most widely used clustering methods. However, its primary limitation is the need to specify the number of clusters beforehand.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that groups data points based on their density in the feature space. Unlike K-Means, DBSCAN does not require the number of clusters to be specified in advance. It identifies clusters of arbitrary shape by finding core points (points with a specified number of neighbors) and expanding the clusters from these core points. DBSCAN is robust to noise and can detect outliers effectively, making it suitable for applications where data might have irregular cluster structures [8].

OPTICS (Ordering Points To Identify the Clustering Structure)

OPTICS is an extension of DBSCAN that addresses its limitations by providing a more comprehensive view of the cluster structure. While DBSCAN relies on a fixed density threshold, OPTICS generates an ordering of data points that captures the clustering structure at varying density levels. This makes OPTICS more versatile in identifying clusters of different densities without the need to manually set parameters for each density level. OPTICS is particularly useful for detecting hierarchical structures in the data [1].

These clustering methods, with their distinct strengths and limitations, offer different approaches for uncovering hidden structures in financial data. K-Means excels in computational speed, DBSCAN effectively handles noise and irregular shapes, and OPTICS provides insights into multi-density clustering structures, making them all valuable tools in financial analysis and machine learning.

2.1.5 Mean Reversion and Half-Life

Mean reversion is a core concept in pairs trading, where the underlying assumption is that asset prices will revert to their historical mean over time. Ernie Chan (2013) [4] has made significant contributions to the understanding and practical application of mean reversion strategies, particularly in the context of pairs trading. His work on half-life calculation provides a quantitative approach to determining the speed at which a price series returns to its mean, which is essential for timing trades effectively.

Chan's method to calculate the half-life of mean reversion is based on the Ornstein-Uhlenbeck process, which he adapts from the Augmented Dickey-Fuller (ADF) test. Chan also highlights the importance of using the calculated half-life to optimize the lookback period for trades. By aligning the lookback period with the half-life, traders can improve the timing of their entry and exit points, maximizing the profitability of their mean-reverting strategies.

2.1.6 Transaction Costs

Transaction costs are a critical factor in the profitability of pairs trading and other high-frequency trading strategies. These costs include brokerage fees, bid-ask spreads, market impact, and slippage, which can significantly erode potential profits, especially in strategies that involve frequent trading. Do and Faff (2012) [6] emphasized the importance of accounting for transaction costs in evaluating the performance of pairs trading strategies. They demonstrated that even strategies that appear profitable on paper can yield negative returns when realistic transaction costs are considered, highlighting the need for robust cost management techniques to sustain profitability in trading.

Effective management of transaction costs requires optimizing trade execution and reducing slippage through advanced order routing algorithms and liquidity analysis. Understanding the market microstructure is also essential for minimizing the impact of transaction costs on trading returns, ensuring that trading strategies remain both efficient and cost-effective.

Chapter 3

Methodology

3.1 Research Design

The research design for this study involved a systematic approach to identifying trading pairs from the universe of 503 stocks in the S&P 500 index. Four different methods were employed to screen for pairs, each providing a unique perspective on potential relationships between stocks. These methods included:

- **Distance Method:** A statistical approach that identifies pairs based on the similarity of their price movements, minimizing the sum of squared deviations between their normalized price series.
- **K-Means Clustering:** An unsupervised machine learning technique used to group stocks into clusters, within which pairs with similar characteristics were selected.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** A clustering algorithm that identifies pairs based on the density of data points, allowing for the detection of pairs even in non-linear distributions.
- **OPTICS (Ordering Points To Identify the Clustering Structure):** An extension of DBSCAN that orders data points to identify clusters of varying densities, enhancing the selection of pairs with more nuanced relationships.

These methods were applied to ensure a comprehensive exploration of stock relationships, with each approach contributing to a diverse pool of candidate pairs. This multi-method approach increases the robustness of the pair selection process, providing a strong foundation for developing and testing pairs trading strategies.

The pairs identified using each of the four methods were then tested using a unified mean-reversion trading strategy. This strategy is based on the assumption that the prices of the selected pairs will revert to their historical mean over time. By continuously monitoring the spread between the pairs, trades were executed when the spread deviated significantly from its mean, with the expectation that it would eventually revert. This approach allowed for a consistent evaluation of the pairs across all methods, ensuring that the performance of each pair was assessed under the same trading conditions. The use of a mean-reversion strategy is well-suited to pairs trading, as it leverages the tendency of correlated assets to return to their equilibrium, thereby maximizing the potential for profit while minimizing risk.

3.2 Data

3.2.1 Data Collection

To perform a comprehensive analysis, stock data for all 503 companies in the S&P 500 index was collected for the period from August 21, 2019, to August 20, 2024. The data was sourced from Yahoo Finance, which provided daily stock prices and other relevant financial metrics needed for this study. The companies in the S&P 500 index are distributed across 11 different sectors, as classified by the Global Industry Classification Standard (GICS). Table 3.1 provides a breakdown of the number of companies in each sector. This sectoral distribution ensures a diverse representation of industries, allowing for a more balanced analysis and a better understanding of sector-specific dynamics in the stock market.

Sector	Companies
Technology	79
Industrials	71
Financial	66
Healthcare	64
Consumer Cyclical	57
Consumer Defensive	37
Utilities	32
Real Estate	31
Basic Materials	22
Communication Services	22
Energy	22

Table 3.1: Sector-wise Companies

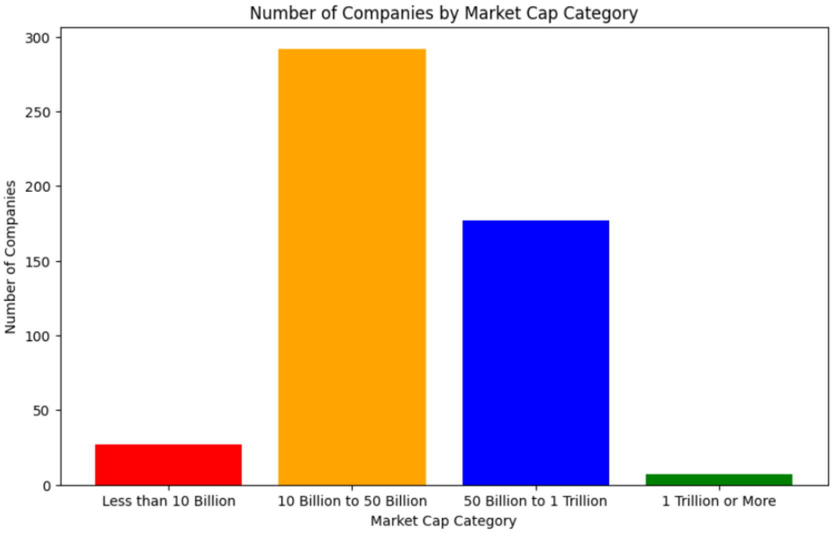


Figure 3.1: Number of Companies by Market Cap

Figure 3.1 illustrates the distribution of companies by market capitalization within the S&P 500 index. The majority of the companies fall within the market cap range of 10B to 50B dollars, indicating a significant representation of mid-cap firms in the dataset. This five-year dataset spans a broad range of market conditions, capturing the influence of various economic events on stock performance, from periods of growth to market volatility.

This comprehensive dataset forms the basis for the subsequent stages of the analysis. It enables the application of advanced statistical techniques and machine learning models aimed at identifying potential trading pairs and constructing robust investment strategies. By leveraging the diverse market cap distribution, the analysis can better understand how companies of different sizes respond to market dynamics, ultimately enhancing the decision-making process for portfolio management.

3.2.2 Data Processing

To evaluate the performance of the pairs trading strategy, the dataset was divided into training and test sets using a time-based split. The methodology involved using four years of historical data for training and the remaining one year for testing. This approach ensures that the model is trained on a substantial amount of data while being tested on a separate, unseen period to validate its performance.

The train-test split was performed using the following steps:

1. **Training Period**: The first four years of the dataset were designated as the training period. This portion of the data was used to estimate the parameters of the trading strategy, such as the cointegration relationship, Z-score thresholds, and other relevant metrics.

2. **Test Period**: The remaining one year of data was reserved for testing. This period was used to evaluate the out-of-sample performance of the strategy, ensuring that the model's predictions and trading signals were validated on data that was not used during the training phase.

The split date was determined based on the starting date of the dataset, with the training set ending exactly four years from that point. This method guarantees a sequential split, maintaining the temporal order of the data, which is crucial for time series analysis and financial modeling.

3.3 Pair Formation Methods

3.3.1 Distance Method

The Distance Method was employed to identify potential trading pairs from a universe of 503 stocks. For each pair, the time series of stock prices was min-max normalized to ensure comparability, transforming the prices into a consistent range. The normalized series for each stock was then used to compute the squared distance between the pairs over the training period. The squared distance metric served as a measure of similarity, capturing how closely the price movements of the two stocks were aligned.

The Sum of Squared Distance (SSD) for each pair was calculated using the following formula:

$$\text{SSD} = \sum_{t=1}^{T} (P_{1,t} - P_{2,t})^2$$

where $P_{1,t}$ and $P_{2,t}$ represent the min-max normalized prices of the two stocks in the pair at time t, and T denotes the length of the training period. The calculated SSD values were then used to rank all pairs in ascending order, with lower distances indicating greater similarity in price movements. This ranking process facilitated the selection of the most closely related pairs for further analysis using mean-reversion strategies.

3.3.2 Machine Learning Methods

The first step in preparing the data for clustering involved computing the weekly log returns for each stock ticker. Log returns are calculated using the following formula:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

where:

- r_t is the log return at time t.
- P_t is the closing price of the stock at time t.
- P_{t-1} is the closing price of the stock at the previous time period.
- $\ln$ denotes the natural logarithm.

Log returns are preferred in finance due to their several advantages:

- **Additivity over time**: Log returns can be summed over multiple periods, which is useful for multi-period return analysis. For example, the log return over two periods can be expressed as the sum of individual log returns:

  $$r_{t,t-2} = r_t + r_{t-1}$$

  This property simplifies the calculation of cumulative returns over longer time horizons.

- **Normalization**: Log returns tend to follow a more normal distribution compared to simple returns, which is beneficial for many statistical models that assume normally distributed data.

These log returns were then used as features in the KMeans, DBSCAN, and OPTICS machine learning models. The use of log returns ensures that the clustering algorithms operate on data that are more suitable for statistical analysis, enhancing the accuracy and interpretability of the resulting clusters. The full vector of features was utilized for clustering to preserve the maximum amount of information in the dataset, as reducing dimensions with techniques like PCA can sometimes lead to the loss of important variability in the data, which might be crucial for distinguishing subtle differences between clusters. However, to facilitate visualization, PCA with three components was employed to project the high-dimensional data into a 3D space, making it easier to interpret the clustering results graphically.

3.4 Spread Computation

The spread is generally defined as the difference between the prices of two assets, representing the deviations from their long-term equilibrium relationship. The Engle-Granger two-step method, this concept is used to test for cointegration between two time series. The first step involves estimating the cointegrating relationship through linear regression to compute the cointegrating coefficient, which defines the equilibrium relationship between the series. The residuals from this regression, known as the "spread," capture the deviations from this equilibrium. In the second step, the stationarity of the spread is tested using the Augmented Dickey-Fuller (ADF) test. If the spread is found to be stationary, it suggests that the two series are cointegrated, indicating a mean-reverting relationship that is well-suited for pairs trading strategies.

3.4.1 Beta Computation

The first step in the Engle-Granger two-step method involves computing the cointegration coefficient, known as β, which quantifies the relationship between the price series of two stocks. To estimate β, a linear regression is performed using the price series of the two stocks, denoted as $P_{1,t}$ and $P_{2,t}$, as follows:

$$P_{1,t} = \alpha + \beta P_{2,t} + \epsilon_t$$

where:

- $P_{1,t}$ is the price of the first stock at time t.
- $P_{2,t}$ is the price of the second stock at time t.
- α is the intercept term.
- β is the cointegration coefficient that indicates the relationship between the two stocks.
- ϵ_t represents the residuals from the regression.

The residuals ϵ_t from this regression capture the deviations from the equilibrium relationship defined by the linear combination of the two stock prices. This step sets the foundation for determining whether the two stocks are cointegrated by analyzing the stationarity of these residuals.

3.4.2 Cointegration Test

In the second step of the Engle-Granger method, the residuals ϵ_t obtained from the linear regression are subjected to a stationarity test using the Augmented Dickey-Fuller (ADF) test. The purpose of this test is to determine whether the residuals have a unit root, which would indicate non-stationarity. The null hypothesis of the ADF test is that the residuals are non-stationary, while the alternative hypothesis is that they are stationary:

$$\Delta\epsilon_t = \gamma\epsilon_{t-1} + \sum_{i=1}^{n} \phi_i \Delta\epsilon_{t-i} + \nu_t$$

where:

- $\Delta\epsilon_t = \epsilon_t - \epsilon_{t-1}$ is the first difference of the residuals.
- γ is the coefficient that indicates stationarity.
- ϕ_i are the coefficients of the lagged differences of the residuals.
- ν_t is the error term.

If the ADF test rejects the null hypothesis, indicating that the residuals are stationary, then the two stocks are considered to be cointegrated. This implies that the price series of the stocks maintain a mean-reverting relationship, making them suitable candidates for pairs trading strategies.

Pairs that pass the cointegration test are ranked based on the strength of their statistical relationship, with the most significant pairs selected for further analysis and trading implementation.

3.4.3 Normalized Spread

The normalized spread is calculated to measure the deviation between two stock prices, adjusted for their relative importance as defined by the cointegrating coefficient (β). This approach ensures that the spread accurately reflects the relationship between the two stocks, accounting for differences in their movements.

The process of computing the normalized spread involves the following steps:

1. **Log Transformation of Prices**: To stabilize the variance and make the price series additive over time, the stock prices are transformed into their natural logarithms. Given two stocks in a pair, denoted as S_1 and S_2, the log-transformed prices are calculated as:

   $$\text{Log\_Price}_{S_1} = \ln(P_{S_1})$$

   $$\text{Log\_Price}_{S_2} = \ln(P_{S_2})$$

   where:

   - P_{S_1} and P_{S_2} represent the prices of stocks S_1 and S_2, respectively.
   - $\ln$ denotes the natural logarithm.

2. **Computation of Normalized Weights**: The weights for the two stocks are derived based on the cointegration coefficient, β. The weights, denoted as w_1 and w_2, are calculated using a normalization function that ensures their sum is equal to 1, thereby balancing the influence of

each stock in the spread. If the β value is provided, the weights are computed as:

$$w_1 = \frac{1}{1 + |\beta|}$$

$$w_2 = \frac{|\beta|}{1 + |\beta|}$$

3. **Calculation of Normalized Spread**: The normalized spread is then calculated as a weighted difference between the log-transformed price series of the two stocks. The formula for the normalized spread is given by:

$$\text{Normalized Spread} = w_1 \cdot \text{Log\_Price}_{S_1} - w_2 \cdot \text{Log\_Price}_{S_2}$$

   where w_1 and w_2 are the weights corresponding to stocks S_1 and S_2, respectively. The resulting normalized spread represents the relative deviation between the two stock prices. This spread serves as the basis for identifying trading signals when implementing mean-reversion strategies.

This methodology ensures that the spread is normalized and accurately captures the relationship between the two stocks, allowing for a more robust analysis when applying statistical trading models.

3.4.4 Z-score

The Z-score normalization of the spread is performed to standardize the spread values, allowing for consistent comparison across different pairs. This process transforms the spread into units of standard deviations from its mean, which helps to identify extreme values that may signal potential trading opportunities.

The computation of the Z-score normalized spread involves the following steps:

1. **Calculation of Mean and Standard Deviation**: For each pair's normalized spread, the mean (μ) and standard deviation (σ) are calculated. Given a normalized spread series, X, these values are computed as:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \mu)^2}$$

   where:

   - X_i represents the value of the normalized spread at time i.
   - N is the number of observations in the spread series.

- μ is the mean of the spread series.
- σ is the standard deviation of the spread series.

2. **Z-score Normalization**: The Z-score for each value of the normalized spread is then computed using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

where:

- X is the value of the normalized spread.
- μ is the mean of the spread series.
- σ is the standard deviation of the spread series.

This Z-score transformation standardizes the spread such that its mean becomes zero and its standard deviation becomes one. This process is performed in a vectorized manner to optimize computation speed and efficiency.

3. **Interpretation of Z-score**: The resulting Z-score normalized spread indicates how many standard deviations each spread value is from its mean. Extreme Z-score values (both positive and negative) suggest that the spread has deviated significantly from its mean, which can be used as a signal for potential trading opportunities in a mean-reversion strategy.

This methodology ensures that the spreads are normalized and can be compared consistently across different pairs, enhancing the ability to detect statistically significant deviations that may indicate trade signals.

3.4.5 Half-life Filter

The half-life of mean reversion measures the time it takes for a spread to revert halfway back to its mean. This metric is essential in pairs trading as it helps determine the expected duration for a mean-reverting process to decay to half of its deviation from the mean. The calculation of half-life involves the following steps:

1. **Lagging the Spread Series**: The spread series is lagged by one period to create a comparison between its current and previous values. Denote the spread at time t as S_t and the lagged spread as S_{t-1}.

2. **Calculation of Spread Change**: Compute the change in the spread (ΔS_t) as the difference between the current spread and its lagged value:

$$\Delta S_t = S_t - S_{t-1}$$

where:

- S_t is the spread at time t.
- S_{t-1} is the lagged spread at time $t-1$.

3. **Linear Regression on Lagged Spread**: Perform a linear regression of the spread change (ΔS_t) on the lagged spread (S_{t-1}) to model the relationship:

$$\Delta S_t = \alpha + \theta S_{t-1} + \epsilon_t$$

where:

- α is the intercept term.
- θ is the slope coefficient that indicates the speed of mean reversion.
- ϵ_t is the error term.

4. **Calculation of Half-Life**: Using the estimated slope coefficient (θ) from the regression, calculate the half-life of mean reversion using the formula:

$$\text{Half-Life} = \frac{-\ln(2)}{\theta}$$

where:

- $\ln(2)$ is the natural logarithm of 2.
- θ is the slope coefficient from the regression.

The half-life represents the number of time periods required for the spread to revert halfway to its mean value.

This methodology provides a quantitative measure of the mean-reversion speed, enabling traders to assess how quickly a spread is expected to decay back to its mean. A shorter half-life indicates a faster mean-reversion process, which is desirable for executing pairs trading strategies.

3.5 Mean Reverting Strategy

The pairs trading strategy aims to exploit the mean-reverting behavior of the spread between two stocks in a pair. This strategy involves entering long or short positions based on the Z-score of the spread, using predefined quantile thresholds to identify trading signals. The detailed steps of the strategy are outlined below:

1. **Z-score Spread Calculation**: The strategy utilizes the previously computed Z-score of the spread to identify optimal entry and exit points. Trading signals are generated when the Z-score crosses predefined thresholds (Q_{high}, Q_{low}), indicating overbought or oversold con-

ditions. This approach helps in quantifying the divergence from the mean, allowing for a systematic and disciplined trading process.

2. **Trading Signals**: The strategy generates trading signals based on the Z-score spread:
   - **Long Entry**: Enter a long position when $Z_t < Q_{\text{low}}$.
   - **Short Entry**: Enter a short position when $Z_t > Q_{\text{high}}$.
   - **Exit Signal**: Close the position when the spread crosses zero or reaches the opposite quantile threshold.
3. **Position Management**: The position at each time step is updated based on the trading signals. Positions are maintained as follows:
   - Maintain a long position if the Z-score remains below the exit signal level.
   - Switch to a short position or exit the position entirely when the conditions are met.
4. **Portfolio Returns Calculation**: The daily returns of the stocks in the pair are calculated, and the weighted portfolio returns are computed as:

   $$R_{\text{portfolio}} = w_1 \times R_{\text{stock1}} - w_2 \times R_{\text{stock2}}$$

   where:

   - $R_{\text{portfolio}}$ is the return of the portfolio on a given day.
   - R_{stock1} and R_{stock2} are the daily returns of stock 1 and stock 2, respectively.
   - w_1 and w_2 are the normalized weights for the pair.

   The equity curve is constructed by compounding the portfolio returns over time:

   $$\text{Equity Curve} = 100 \times (1 + R_{\text{portfolio}}).\text{cumprod()}$$

5. **Plotting and Visualization (Optional)**: The strategy optionally generates a visual representation of the trading process, highlighting entry and exit points, Z-score levels, and the equity curve.

This methodology outlines a structured approach to executing a pairs trading strategy using Z-score normalization, ensuring consistent and robust entry and exit signals for mean-reversion trades.

3.5.1 Performance Metrics

The performance metrics for the pairs trading strategy were calculated to evaluate the risk-adjusted returns of each equity curve generated from the trading signals. These metrics provide insights into the profitability, risk, and consistency of the strategy. The following steps and formulas outline the calculations performed:

1. **Cumulative Return**: The cumulative return R_t measures the total return generated by the strategy over the entire trading period.

2. **Excess Returns**: The excess returns are calculated by subtracting the risk-free rate from the daily returns:
$$\text{Excess Returns} = R_t - \frac{\text{Risk-Free Rate}}{252}$$
where the risk-free rate is annualized and adjusted for daily frequency.

3. **Sharpe Ratio**: The Sharpe ratio evaluates the risk-adjusted return of the strategy. It is calculated using the excess returns as:
$$\text{Sharpe Ratio} = \frac{\sqrt{252} \times \text{Mean(Excess Returns)}}{\text{Standard Deviation(Excess Returns)}}$$

4. **Sortino Ratio**: The Sortino ratio is similar to the Sharpe ratio but focuses on downside risk by only considering negative returns. It is defined as:
$$\text{Sortino Ratio} = \frac{\sqrt{252} \times \text{Mean(Excess Returns)}}{\text{Standard Deviation(Downside Returns)}}$$
where downside returns are those values of excess returns that are less than zero.

5. **Treynor Ratio**: The Treynor ratio measures the returns earned in excess of the risk-free rate per unit of market risk (beta). Assuming a beta of 1 for simplicity, it is calculated as:
$$\text{Treynor Ratio} = \frac{\text{Mean(Returns)} \times 252}{\beta}$$

6. **Maximum Drawdown**: Maximum drawdown represents the maximum observed loss from a peak to a trough of the equity curve before a new peak is achieved.

These performance metrics were calculated using vectorized operations to ensure computational efficiency, with the help of parallel processing for faster execution. The metrics provide a comprehensive view of the strategy's performance, balancing returns against risk and drawdown levels.

Chapter 4

Results

4.1 Distance Method

From the 503 stocks, the following 11 symbols were excluded as they lacked complete data for the full 5-year period: ABNB, CARR, CEG, GEHC, GEV, KVUE, NCLH, OTIS, SOLV, SW, and VLTO. After filtering out these incomplete datasets, the analysis proceeded with the remaining 492 stocks. Using these, combinatorics were applied to generate all possible pairs, resulting in 120,786 unique ticker combinations. The squared distance between each pair was calculated, serving as a critical metric for evaluating pair similarities.

The resulting squared distances are visualized in Figure 4.1, providing insights into the spread of distances between stock pairs. A corresponding histogram (Figure 4.2) presents the frequency distribution of these squared distances, highlighting the concentration of values and identifying potential outliers. This histogram allows for a better understanding of the overall data distribution and offers a clearer picture of how many pairs exhibit lower versus higher distances, which may guide further pair selection or filtering criteria in subsequent analyses.

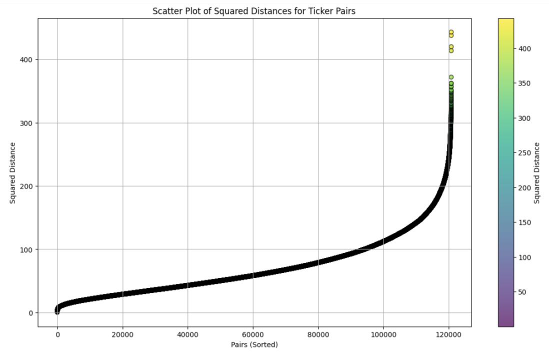


Figure 4.1: Scatter Plot of Squared Distances for Ticker Pairs

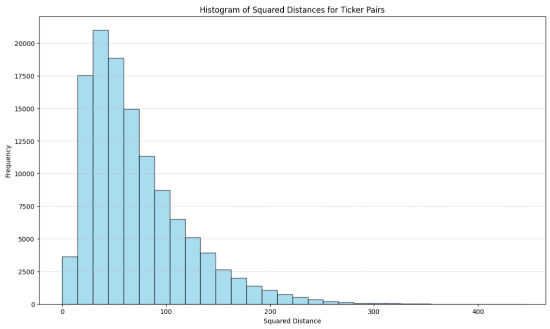


Figure 4.2: Histogram of Squared Distances for Ticker Pairs

To optimize performance and ensure the computations were completed efficiently, vectorized operations in Pandas were utilized, enabling the simultaneous processing of large arrays of data. This approach significantly reduced computational time compared to traditional loop-based methods, making the analysis feasible even with a high volume of pairs. The efficiency gained through vectorization allowed for quicker iterations and faster identification of relevant stock pairs.

After calculating the squared distances for all pairs, the results were sorted in ascending order to facilitate the selection of pairs with the highest potential for trading. The lowest 10th percentile of these distances was chosen as a threshold for pairs trading, ensuring the selection of stock pairs with minimal deviation. This selection process resulted in a total of **12,079** pairs being shortlisted for further analysis. The maximum squared distance within this percentile was 22.86, providing a clear benchmark for filtering out pairs with higher divergence. By focusing on this subset, the strategy aims to capture stocks with tighter correlations, enhancing the likelihood of mean-reverting behavior crucial for pairs trading strategies.

4.1.1 Spread Computations

The spreads for the 12,079 pairs were generated using the methodology outlined in Chapter 3. The cointegration status and half-life of the pairs are presented in the abridged Table 4.1 (sorted by half-life). The complete table contains 12,079 rows. For brevity, only the top five and bottom five rows are displayed.

Pair	Cointegrated	Half-life
AZO_MCK	no	-7457.51
DVN_FANG	no	-4421.90
IQV_LH	yes	9.55
FI_ROST	yes	9.85
AVY_NXPI	yes	9.99
… … … … … …		
CVX_DVN	no	1458.78
AZO_DVN	yes	1476.77
NUE_URI	no	1512.05
DVN_EOG	no	3401.84
DVN_HES	no	4059.22

Table 4.1: Cointegration and Half-life Statistics (Abridged)

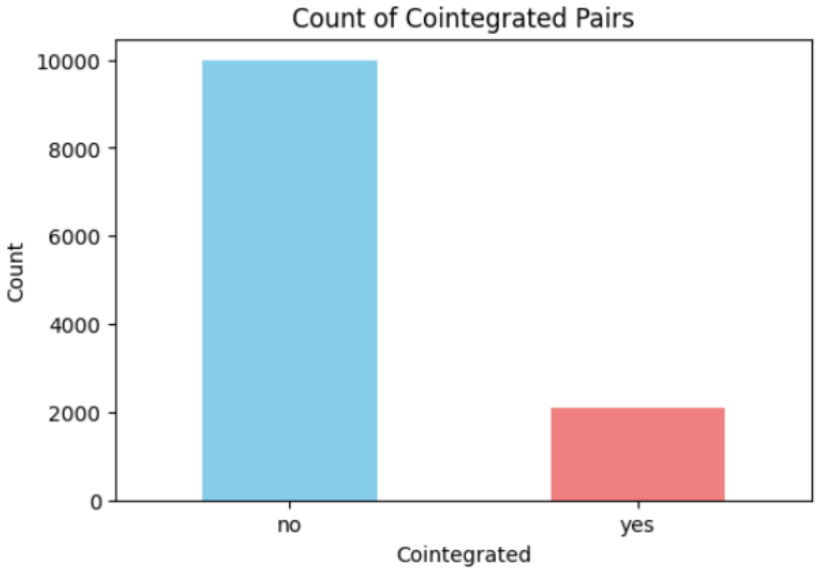
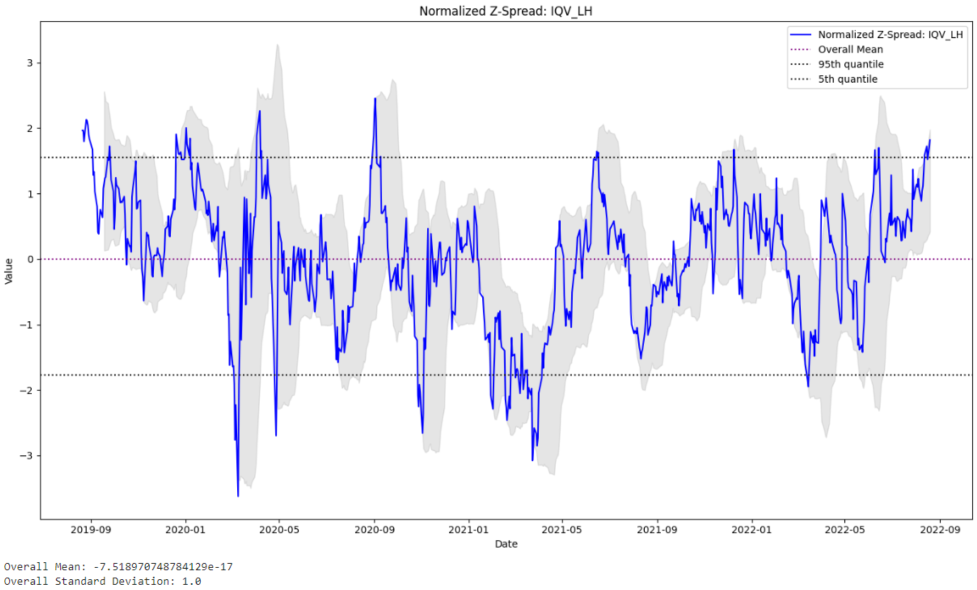


Figure 4.3: Count of Cointegrated Pairs

Two of these pairs exhibited negative half-lives, while several others had exceptionally long half-lives. This suggests a potential divergence in the behavior of certain stock pairs, which could indicate underlying structural differences or external market influences. Of the pairs analyzed, 2,106 were identified as cointegrated (Figure 4.3), while the remaining 9,973 pairs were not.

4.1.2 Spread Analysis: IQV_LH

From Table 4.1, the pair IQV_LH is selected for further analysis. The ticker IQV represents Iqvia Holdings Inc, while LH stands for Labcorp Holdings Inc. Both companies operate within the healthcare sector, sharing similar market dynamics and external influences, which may contribute to their strong cointegration. The normalized Z-spread for this pair is plotted in Figure 4.4, illustrating the mean-reverting behavior of the spread over time. In addition, Figure 4.5 presents the histogram and Q-Q plot of the Z-spread, providing insights into the distribution of the spread and its alignment with a normal distribution.



```
Overall Mean: -7.518970748784129e-17
Overall Standard Deviation: 1.0
```

Figure 4.4: Normalized Z-Spread for Pair: IQV_LH

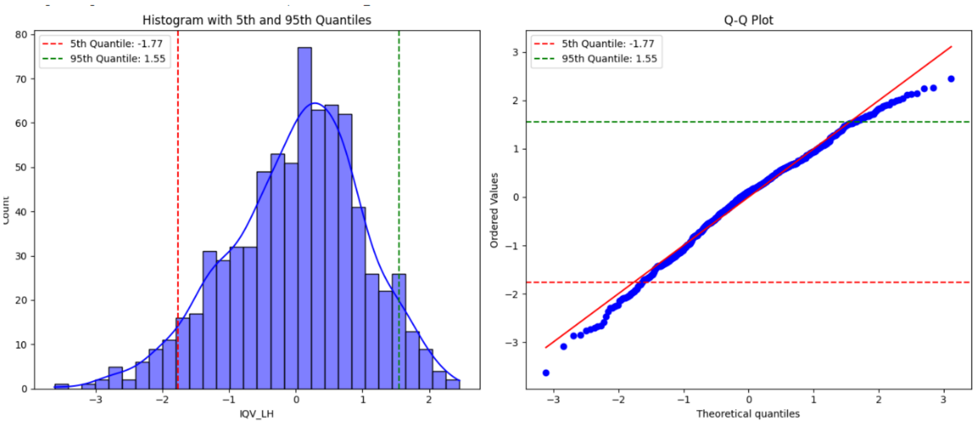
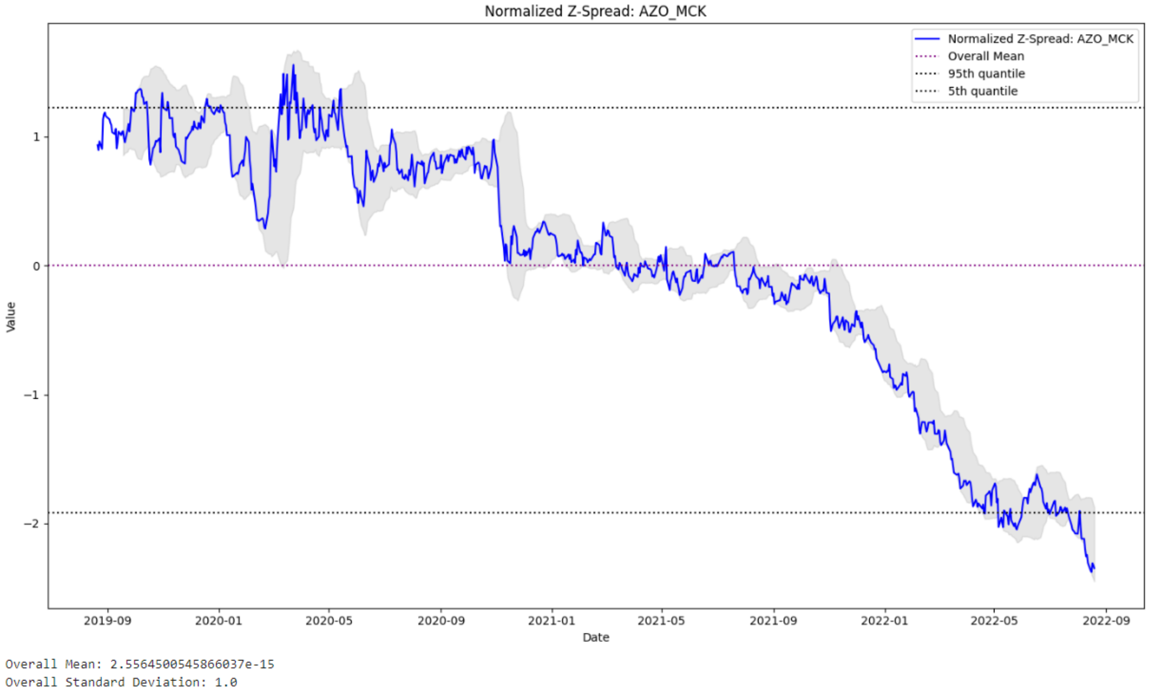


Figure 4.5: Histogram and QQ plot of Normalized Z-Spread for Pair: IQV_LH

The 95th and 5th quantiles are also depicted in these figures, serving as key decision points for the trading strategy. The 95th quantile marks the upper bound for short positions and the 5th quantile serves as the lower bound for long positions. These thresholds are critical in defining the optimal entry and exit points for pairs trading, ensuring that positions are opened when the spread deviates significantly from its mean and closed as it returns to equilibrium. The half-life of the pair IQV_LH is 9.55, indicating that any deviation from the equilibrium tends to correct itself relatively quickly. Given its cointegration status, this pair is well-suited for pairs trading, as it mean-reverts frequently, providing numerous trading opportunities. Since both companies are large, well-established firms, their stock prices are less likely to be influenced by speculative volatility, further reinforcing the reliability of this trading pair. This type of pairs are ideal for spread trading.

4.1.3 Spread Analysis: AZO_MCK

We also examine the normalized Z-spread of the AZO_MCK pair. AZO represents Autozone Inc, a prominent player in the industrial sector, while MCK stands for McKesson Corp, a leading firm in the healthcare sector. These two companies operate in distinct industries, which is reflected in their lack of cointegration. From Table 4.1, the AZO_MCK pair has a negative half-life, suggesting that the spread between their prices tends to widen rather than revert to a mean over time. In Figure 4.6, the normalized Z-spread of AZO_MCK further demonstrates the divergence, with no clear pattern of mean reversion.



```
Overall Mean: 2.5564500545866037e-15
Overall Standard Deviation: 1.0
```

Figure 4.6: Normalized Z-Spread for Pair: AZO_MCK

A negative half-life indicates that the relationship between these two stocks is diverging rather than mean-reverting, which implies that any short-term convergence is likely temporary. This makes the AZO_MCK pair unsuitable for pairs trading, as successful pairs trading relies on the spread reverting to its mean after divergence. In this case, attempting to trade on such a pair would introduce significant risk, as the positions could continue moving apart indefinitely without returning to a predictable equilibrium. The lack of cointegration and the negative half-life suggest that external market forces, sector-specific trends, or fundamental differences between the companies are driving the divergence. As a result, rather than relying on statistical arbitrage, traders would need to explore alternative strategies, such as directional trading, if they wish to take advantage of movements in either stock.

4.1.4 Spread Trading

To identify the most suitable pairs for trading, we filtered out pairs with a negative half-life as well as those with a half-life longer than 250 days. The threshold of 250 days was chosen because it approximates the number of trading days in a year (252), ensuring that only pairs with mean-reversion within a reasonable and practical time frame are considered for trading. This criterion ensures that positions in a pair can be closed within a reasonable period, avoiding prolonged exposure to market risks. After applying this filter, 11,759 pairs remained from the original 12,079 pairs, as shown in the abridged Table 4.2.

Pair	**Cointegrated**	**Beta**	**5th Quantile**	**95th Quantile**	**Half-life**
IQV_LH	yes	1.07	-1.77	1.55	9.55
FI_ROST	yes	0.49	-1.55	1.61	9.85
AVY_NXPI	yes	0.93	-1.46	1.63	9.99
FI_ZBH	yes	0.41	-1.51	1.63	10.51
AWK_RVTY	yes	0.50	-1.64	1.44	10.54
… … … … … …					
EFX_JCI	no	3.10	-1.33	1.94	248.75
HSY_IRM	no	2.75	-1.44	1.48	249.16
COP_NOC	yes	0.36	-1.62	1.60	249.21
ADP_CTVA	no	2.57	-1.47	1.50	249.68
DE_JCI	no	5.75	-1.56	1.86	249.71

Table 4.2: Pair Statistics - Train Data (Abridged)

The five years of stock data were divided into training and test datasets, with the training set covering three years (756 trading days) and the test set covering two years (502 trading days). This split allows us to model and validate the trading strategy in a way that mimics real-world conditions, where historical data is used to inform future trades. In the training set, we calculated the upper (95th quantile) and lower (5th quantile) thresholds for short and long entries, respectively. These quantile-based thresholds serve as key decision points in the trading strategy, marking points at which the spread has deviated significantly from the mean, signaling potential opportunities to enter trades.

By applying the thresholds and Beta (hedge ratio) calculated from the training set to the test set, we aim to prevent overfitting and data snooping, ensuring that the model is evaluated on unseen data. This process leads to a more robust assessment of the strategy's performance. The application of historical data for both threshold and Beta calculation allows us to simulate the trading strategy in a forward-looking manner, reflecting how it would perform in real trading scenarios.

The selection process, by filtering pairs with appropriate half-lives and using quantile-based thresholds, helps to refine the strategy and focus only on pairs that exhibit the most reliable mean-reversion behavior. This approach minimizes unnecessary trades on pairs that may not revert quickly or at all, ultimately improving the efficiency and profitability of the pairs trading strategy.

4.1.5 Performance Metrics

After determining the pair trading parameters (thresholds and Beta) from the training period, we applied the pairs trading strategy to the identified 11,759 pairs over a 5-year period (1,258 trading days). Of these pairs, 10,982 generated positive cumulative returns, indicating that the majority of pairs exhibited profitable trading potential throughout the evaluation period. Table 4.3 summarizes the key performance metrics of the strategy, evaluated using the distance method. These metrics include Cumulative Return (%), Sharpe Ratio, Sortino Ratio, Maximum Drawdown, and Total Trades, offering a comprehensive view of the strategy's overall effectiveness and risk-adjusted performance.

Metric	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown	Total Trades
Mean	56.00	0.63	0.68	-0.23	6.95
Median	49.61	0.63	0.64	-0.20	6.00
Min	-58.62	-0.52	-0.63	-0.80	2.00
Max	518.71	2.10	2.90	-0.01	37.00
Standard Deviation	46.08	0.35	0.40	0.12	3.87

Table 4.3: Performance Metrics Summary: Distance Method

The trading strategy demonstrates positive performance, with a mean cumulative return of 56%, indicating its potential for profitability. However, the standard deviation of 46.08% shows significant variability in returns, suggesting that while some trades may generate substantial gains, others may result in losses. Risk-adjusted metrics, such as the Sharpe Ratio of 0.63 and the Sortino Ratio of 0.68, indicate moderately favorable performance relative to the risk involved. While these metrics suggest that the strategy is effective in balancing returns with risk, they also highlight room for improvement, particularly in managing downside risk.

The Maximum Drawdown, with a mean of -23%, reflects the largest observed loss, a key metric for assessing the strategy's performance during adverse market conditions. The strategy's execution appears selective, with a median of 6 trades, pointing to a focused approach. However, extreme outcomes are possible, as seen in the maximum return of 518.71% and the minimum return of -58.62%. These figures underscore the high-risk, high-reward nature of the strategy, where significant gains may be realized in favorable markets, but substantial losses may occur during unfavorable conditions.

Figure 4.7 presents the distribution of cumulative returns for the trading strategy evaluated using the distance method. This histogram provides insight into the variability of returns, indicating the range and frequency of outcomes.

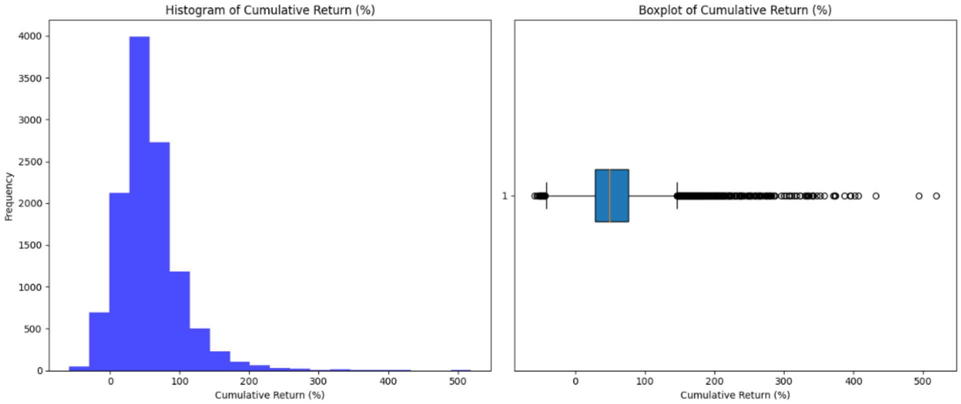


Figure 4.7: Histogram of Cumulative Returns: Distance Method

Similarly, Figure 4.8 illustrates the distribution of Sharpe Ratios for the same strategy. This chart highlights how the strategy's risk-adjusted returns are spread out, helping to assess its overall performance consistency.

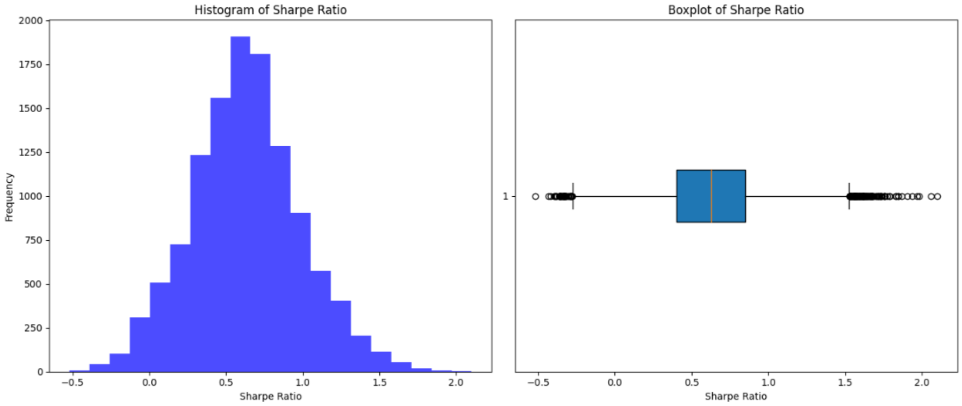


Figure 4.8: Histogram of Sharpe Ratio: Distance Method

4.1.6 Feature Importance

We analyze the results of the pairs trading strategy using a Random Forest Regressor to identify the features that contribute most significantly to the Sharpe ratio. In particular, we focus on the pairs with positive Sharpe ratio. The analysis highlights that the half-life of mean reversion is the most influential feature in predicting the Sharpe ratio, suggesting that pairs with a quicker mean-reverting behavior tend to perform better. This is consistent with the theoretical expectation that pairs trading strategies rely on mean reversion to generate profits.

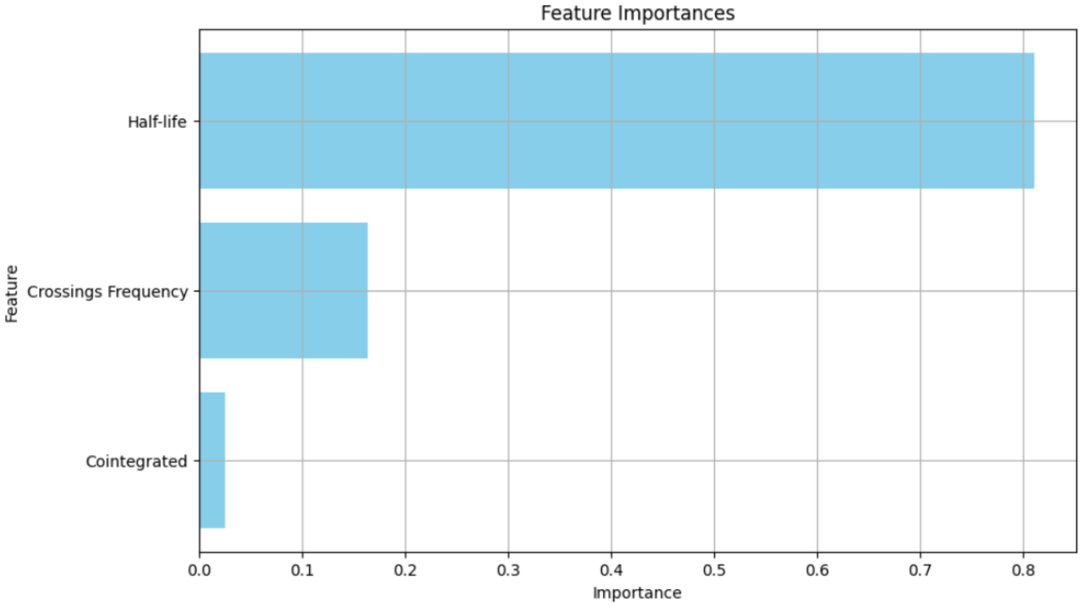


Figure 4.9: Feature Importance

In addition to the half-life, the total number of threshold crossings emerges as the second most important feature. This feature measures the frequency with which the spread between the two stocks crosses a predefined threshold, indicating potential trading opportunities. Pairs with a higher number of threshold crossings tend to have more trading signals, which may lead to better performance, as long as the trades are executed efficiently.

Surprisingly, cointegration, which is often considered a cornerstone of pairs trading strategies, appears to be less important in comparison to half-life and threshold crossings. This may suggest that while cointegration ensures a long-term statistical relationship between the pair, short-term dynamics like mean reversion speed and trading signals play a more critical role in driving performance. The feature importance rankings are visualized in Figure 4.9, which clearly illustrates the dominance of half-life and threshold crossings over other features. The relatively lower importance of cointegration highlights the need for a more nuanced approach when selecting pairs for trading, where factors beyond just statistical relationships should be considered.

4.1.7 Cluster Analysis

Cluster analysis was performed using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to categorize the pairs based on two key features: the Sharpe ratio and half-life of mean reversion. The DBSCAN clustering results are visualized in Figure 4.10.

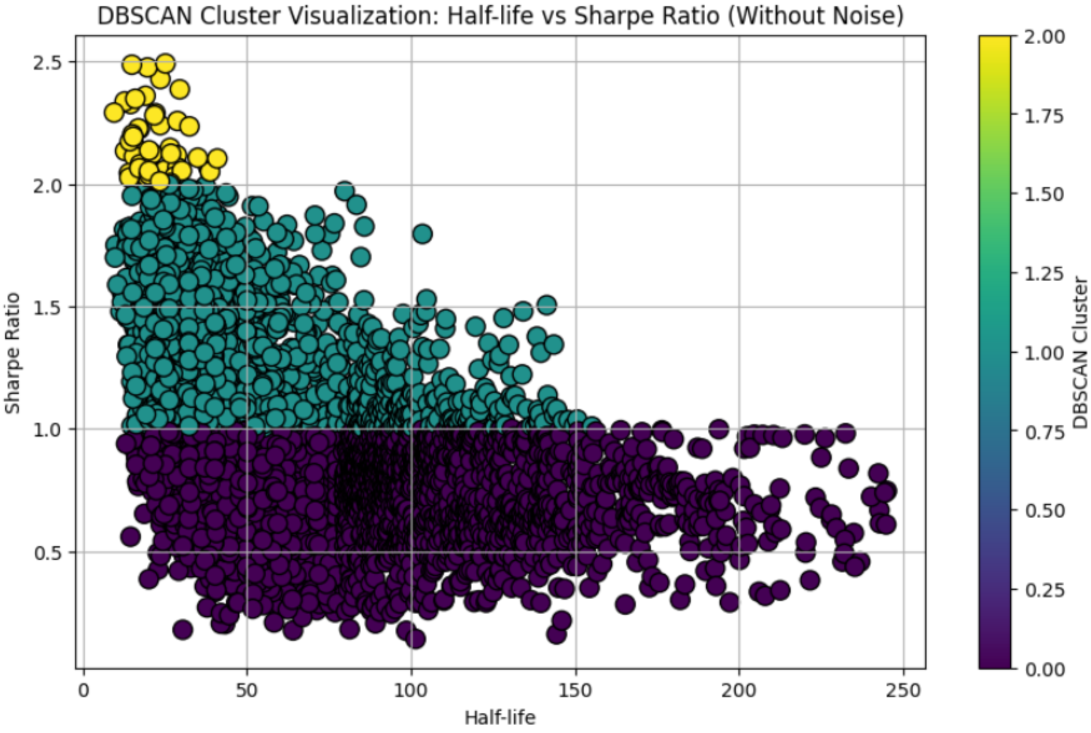


Figure 4.10: DBSCAN Cluster Analysis of Sharpe Ratio and Half-life

DBSCAN is particularly useful in this context because it can identify clusters of pairs that share similar characteristics without requiring a predefined number of clusters, and it also effectively handles noise in the data. The figure illustrates the separation of pairs into distinct clusters based on their Sharpe ratio and half-life, with the higher-performing pairs (i.e., those with shorter half-lives and higher Sharpe ratios) clearly distinguishable from those with suboptimal performance. Additionally, the noise points identified by DBSCAN represent pairs that do not fit well into any cluster, further emphasizing the variability in performance across different pairs.

The results of the clustering revealed distinct patterns in the performance of pairs. One key finding is that pairs with a half-life longer than 150 days tend to produce suboptimal results according to the trading algorithm, as these pairs often exhibit slower mean reversion and generate fewer profitable trading signals. These pairs were typically grouped into a separate cluster of lower Sharpe ratios, indicating that a longer half-life negatively impacts the performance of a pairs trading strategy.

In contrast, pairs with a half-life shorter than 50 days significantly improve the Sharpe ratio. These pairs tend to revert to their mean much more quickly, providing more frequent and timely trading opportunities. This observation was reflected in a cluster with higher Sharpe ratios, confirming the importance of short-term mean reversion dynamics in achieving optimal performance.

4.1.8 Top 5 Pairs

The Table 4.4 highlights the top five pairs (sorted by Sharpe ratio) selected using the distance method, which demonstrate superior performance as seen by their higher cumulative returns, Sortino ratios, and lower maximum drawdowns. The average cumulative return of these top five pairs is 152.01%, with an average Sharpe ratio of 2.01 and an average Sortino ratio of 2.33. The maximum drawdown is also minimal, averaging at -0.08, with an average of 26 trades per pair. These results underscore the effectiveness of the pairs in delivering strong risk-adjusted returns with limited downside exposure. For instance, the pair IQV_LH exhibits a cumulative return of 156.05% and a Sharpe ratio of 2.10, while maintaining a low maximum drawdown of -0.06. Similarly, the pair AEE_PAYX leads with the highest cumulative return of 183.17%, further showcasing the robustness of these pairs in a distance-based pairs trading strategy.

Pair	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown	Total Trades
IQV_LH	156.05	2.10	2.34	-0.06	37
GM_IP	144.38	2.06	2.03	-0.06	17
CL_MAS	127.48	1.98	2.22	-0.07	25
AEE_PAYX	183.17	1.97	2.16	-0.13	27
PM_SO	148.94	1.94	2.90	-0.07	25

Table 4.4: Performance of Top 5 Pairs: Distance Method

Table 4.5 provides a summary of the key characteristics of the top 5 pairs identified using the distance method for optimal portfolio selection in a pairs trading strategy. The table outlines several important features that are essential for understanding the dynamics of each pair and their potential for mean-reversion-based trading.

Pair	Squared Distance	Sector 1	Sector 2	Half-life	Cointegrated
IQV_LH	2.59	Healthcare	Healthcare	9.56	yes
GM_IP	13.30	Consumer Cyclical	Consumer Cyclical	29.35	no
CL_MAS	22.86	Consumer Defensive	Industrials	22.17	no
AEE_PAYX	11.68	Utilities	Industrials	14.99	yes
PM_SO	16.82	Consumer Defensive	Utilities	23.68	no

Table 4.5: Features of Top 5 Pairs - Distance Method

Performance Analysis: IQV_LH

The pair IQV_LH is the best performing pair identified by the distance method. Table 4.6 provides a comparison between the pairs trading strategy and the Buy & Hold approaches for the individual stocks IQV and LH. The results show that the pairs trading strategy significantly outperforms the Buy & Hold strategies in terms of both return and risk-adjusted performance. The cumulative return for the pairs trading strategy is 156.05%, far exceeding the returns of Buy & Hold for IQV (57.10%) and LH (60.35%).

Column	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown
Strategy	156.05	2.10	2.34	-0.06
Buy & Hold IQV	57.10	0.44	0.61	-0.49
Buy & Hold LH	60.35	0.46	0.57	-0.47

Table 4.6: Performance Comparison: Strategy vs. Buy & Hold Approaches

In terms of risk-adjusted performance, the Sharpe ratio for the strategy (2.10) is substantially higher than that for the Buy & Hold approaches (0.44 for IQV and 0.46 for LH). A higher Sharpe ratio indicates that the pairs trading strategy generates more return per unit of risk, making it a more efficient strategy. Similarly, the Sortino ratio, which emphasizes downside risk by penalizing negative volatility, is much higher for the pairs trading strategy (2.34) compared to the Buy & Hold approaches (0.61 for IQV and 0.57 for LH). This suggests that the strategy effectively mitigates harmful fluctuations in price, leading to a smoother equity curve.

Additionally, the maximum drawdown for the pairs trading strategy is much smaller (-0.06), highlighting its better ability to manage risk. In contrast, the Buy & Hold strategies for IQV and LH experience significantly larger drawdowns (-0.49 for IQV and -0.47 for LH), which indicates greater exposure to market downturns. The lower drawdown for the pairs strategy underscores its robustness during volatile market conditions, as it limits potential losses and recovers more quickly from price shocks.

Figure 4.11 compares the equity curves for the pairs trading strategy on the IQV_LH pair against the Buy & Hold strategies for the individual stocks IQV and LH.

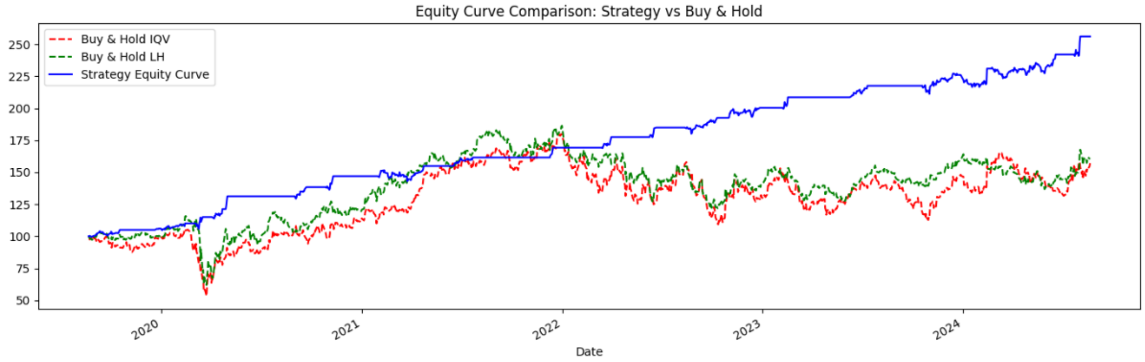


Figure 4.11: Equity Curves: IQV_LH

Figure 4.12 illustrates the Z-Spread between IQV and LH, as well as the corresponding trading positions.

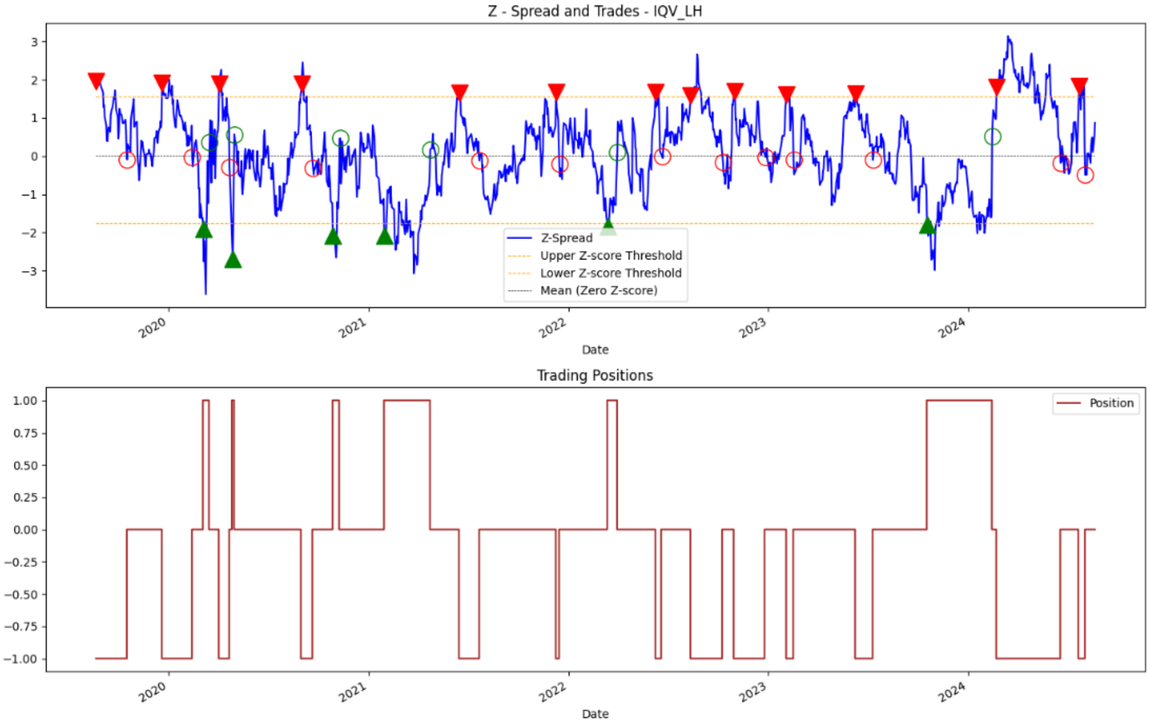


Figure 4.12: Z-Spread and Trading Positions for Pair: IQV_LH

The Z-Spread helps identify trading opportunities as it indicates deviations from the mean, with positions being entered when the spread exceeds a certain threshold, as shown in the plot.

4.2 K-Means Method

Weekly log returns were computed and used as input features for the K-Means clustering algorithm. K-Means is an unsupervised machine learning algorithm that partitions data into clusters, with each stock ticker assigned to the cluster whose centroid (mean log return) is closest to it.

The clustering process proceeded as follows:

1. **Feature Matrix**: A matrix was created where each row represented a stock ticker, and the columns represented the weekly log returns over the selected time period.
2. **Cluster Assignment**: K-Means assigned each stock to one of the clusters by minimizing the sum of squared distances between the stock's log returns and the centroid of the assigned cluster.

4.2.1 Optimal Cluster Size Selection

Selecting the optimal number of clusters for K-Means clustering is essential. Too few clusters can overlook important differences, while too many clusters may introduce unnecessary complexity.

The elbow plot is used to determine the optimal number of clusters. It plots the *within-cluster sum of squares (WCSS)* against the number of clusters:

- **WCSS**: This metric measures the total distance between points in each cluster and their respective centroids. A lower WCSS indicates that points are closer to their centroids, and the clusters are more compact.
- As the number of clusters increases, WCSS decreases since additional clusters reduce the distance between points and their centroids.

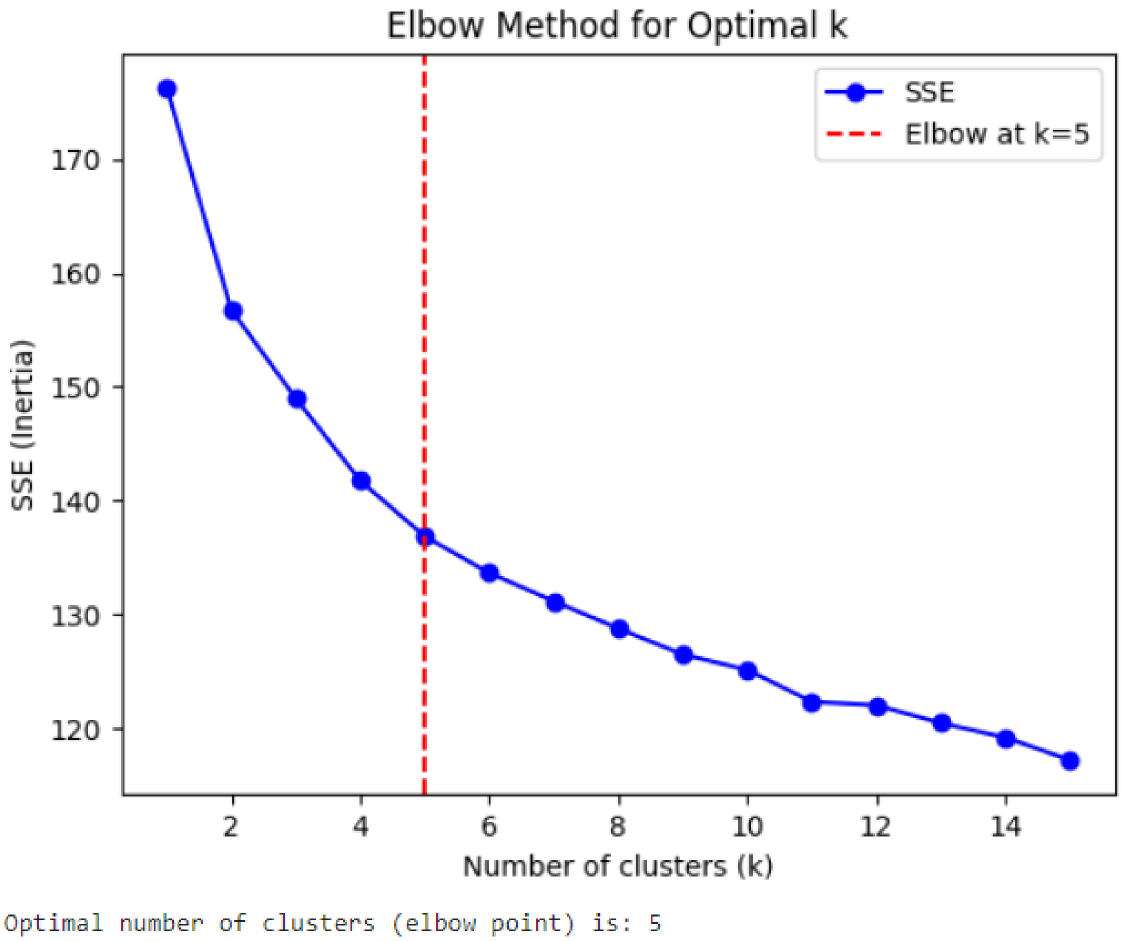
The elbow plot (Figure 4.13)shows the WCSS for a range of different cluster sizes. Initially, as the number of clusters increases, the WCSS decreases sharply. However, after a certain point, the rate of decrease slows down, forming an "elbow." This elbow indicates the point where adding more clusters does not significantly improve the model, thus identifying the optimal number of clusters.

Based on the elbow plot, a cluster size of 5 was selected. By selecting 5 clusters, the following benefits are achieved:

- The stocks are grouped into a manageable number of clusters, making the analysis easier to interpret.
- Each cluster contains stocks with similar patterns in their weekly log returns, indicating similar behaviors in terms of volatility and performance.

Overall, the selection of 5 clusters allows for deeper insights into different groups of stocks, potentially highlighting sectors or specific stock characteristics that drive similar return profiles. The resulting clusters in 3D space is shown in Figure 4.14

This decision corresponds to the point where the reduction in WCSS begins to diminish, indicating that 5 clusters offer a good balance between capturing meaningful differences among the stock tickers and avoiding overfitting.



```
Optimal number of clusters (elbow point) is: 5
```

Figure 4.13: K-Means Elbow Plot

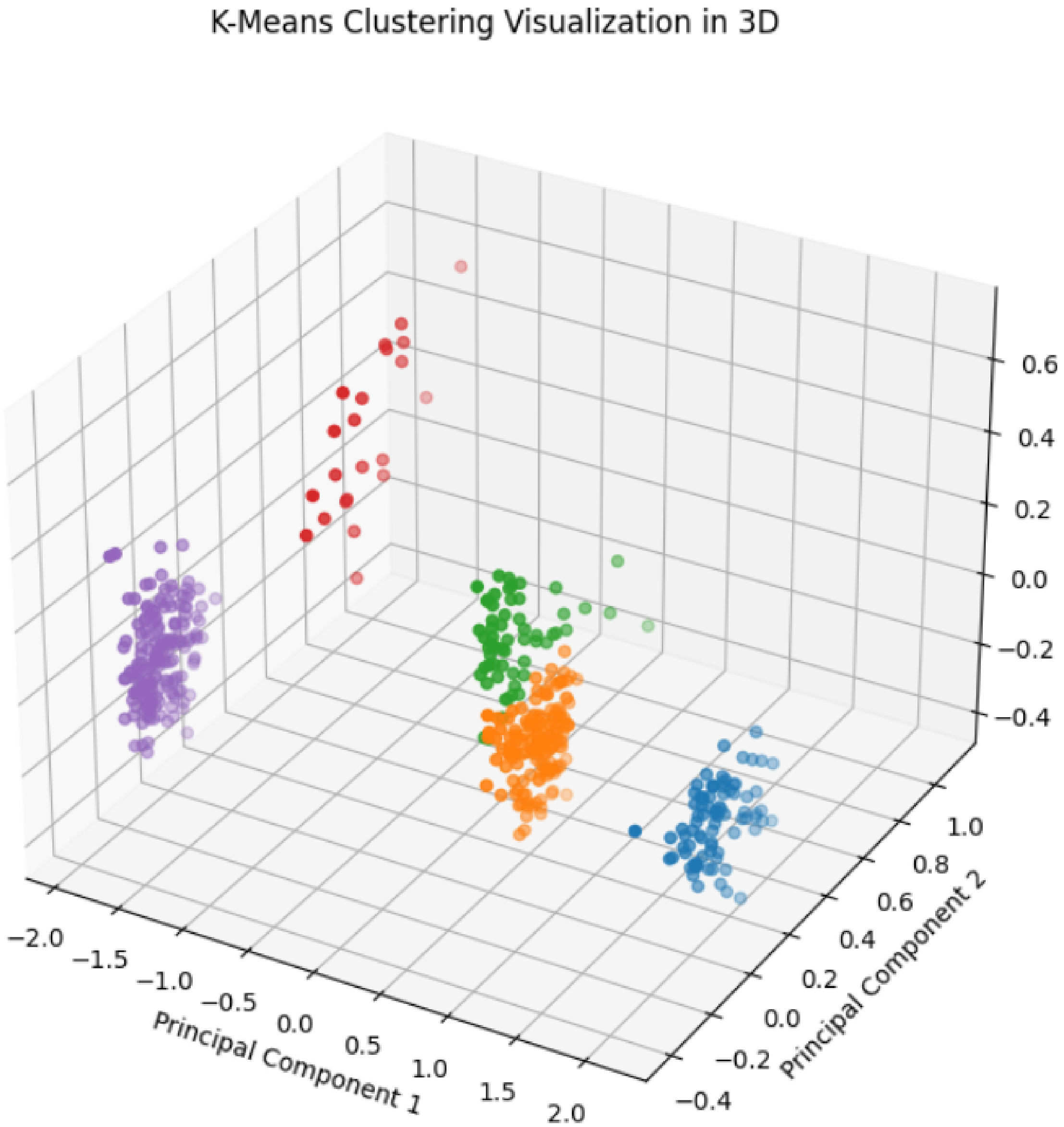


Figure 4.14: K-Means Clusters

The overall K-Means Silhouette Score for the clustering is 0.472, which provides insight into the quality of the clusters formed by the K-Means algorithm. The *Silhouette Score* ranges from -1 to 1 and is used to measure how similar an object is to its own cluster compared to other clusters:

- A **score close to 1** indicates that the stock tickers are well-matched to their assigned cluster and poorly matched to other neighboring clusters. This suggests that the clusters are well-separated and distinct.
- A **score close to 0** indicates that the stock tickers lie near the boundary of their assigned cluster, making them difficult to distinguish from tickers in neighboring clusters.
- A **negative score** indicates that some stock tickers may have been incorrectly assigned to a cluster.

With a **Silhouette Score of 0.472**, the K-Means model demonstrates moderate clustering performance. This score suggests that the clusters are reasonably well-separated, but there may still be some overlap between clusters, particularly for stock tickers that exhibit less distinct return patterns. Overall, this score indicates that the selected cluster size of 5 is appropriate for identifying meaningful groupings of stock tickers based on their weekly log returns, although there could be areas for further refinement to enhance the cluster separation.

After the stocks were grouped into 5 clusters using the K-Means algorithm, combinatorics was applied within each cluster to explore potential stock pairs for further analysis. The combinatorics process involves generating all possible unique pairs of stocks from within each cluster.

The total number of unique pairs generated across all clusters was **32,087**. This extensive set of pairs provides a broad range of stock combinations that can be analyzed for potential trading strategies, such as pairs trading or correlation analysis. By limiting the combinatorics to stocks within the same cluster, we ensure that the stocks being paired together share similar return patterns or characteristics, as identified by the K-Means clustering algorithm. This approach reduces the noise and increases the likelihood of identifying meaningful relationships between the paired stocks.

4.2.2 Performance Metrics

The table 4.7 summarizes the performance metrics of the trading strategy based on pairs selected by the K-Means method. The metrics include cumulative return, Sharpe ratio, Sortino ratio, maximum drawdown, and total trades. Of the 30,224 pairs analyzed, 27,284 recorded positive cumulative returns, indicating that the majority of these pairs exhibited profitable trading potential over the five-year period. The mean cumulative return is 59.53%, indicating a positive overall performance, while the median return is slightly lower at 51.45%, showing some variability in outcomes. The maximum cumulative return is 800.27%, highlighting the potential for significant profit, but the minimum value of -96.52% demonstrates the risk of substantial losses. The strategy's risk-adjusted performance is captured by the Sharpe and Sortino ratios, with mean values of 0.60 and 0.65, respectively. The maximum drawdown, with a mean of -0.26, signifies the average peak-to-trough decline, with a minimum drawdown as severe as -0.99. The total trades range from 2 to 33, with an average of 7.01 trades, indicating the strategy's trading frequency.

Metric	**Cumulative Return (%)**	**Sharpe Ratio**	**Sortino Ratio**	**Maximum Drawdown**	**Total Trades**
Mean	59.53	0.60	0.65	-0.26	7.01
Median	51.45	0.61	0.63	-0.23	6.00
Min	-96.52	-0.98	-0.85	-0.99	2.00
Max	800.27	2.43	3.13	-0.01	33.00
Standard Deviation	59.21	0.38	0.43	0.13	3.77

Table 4.7: Performance Metrics Summary: K-Means Method

The standard deviation values in table 4.7 provide further insights into the variability of the strategy's performance. The standard deviation of cumulative returns is 59.21%, indicating significant variation between the best and worst-performing periods. Similarly, the Sharpe and Sortino ratios have standard deviations of 0.38 and 0.43, respectively, showing moderate volatility in risk-adjusted returns. The maximum drawdown has a standard deviation of 0.13, suggesting that while the average drawdown is moderate, some extreme scenarios of capital decline exist. The total trades have a standard deviation of 3.77, implying that the strategy's trading frequency can vary considerably across different instances.

Figure 4.15 illustrates the distribution of cumulative returns for the trading strategy based on the K-Means method. The histogram reveals the spread and frequency of the returns, offering a clear view of the strategy's variability in performance.

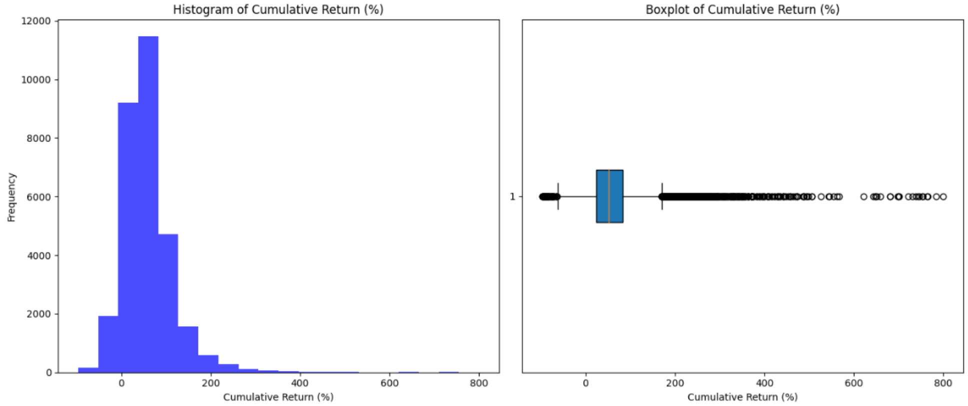


Figure 4.15: Histogram of Cumulative Returns: K-Means Method

Likewise, Figure 4.16 depicts the distribution of Sharpe ratios for the K-Means-based strategy. This chart emphasizes the dispersion of risk-adjusted returns, offering insights into the consistency and reliability of the strategy's performance.

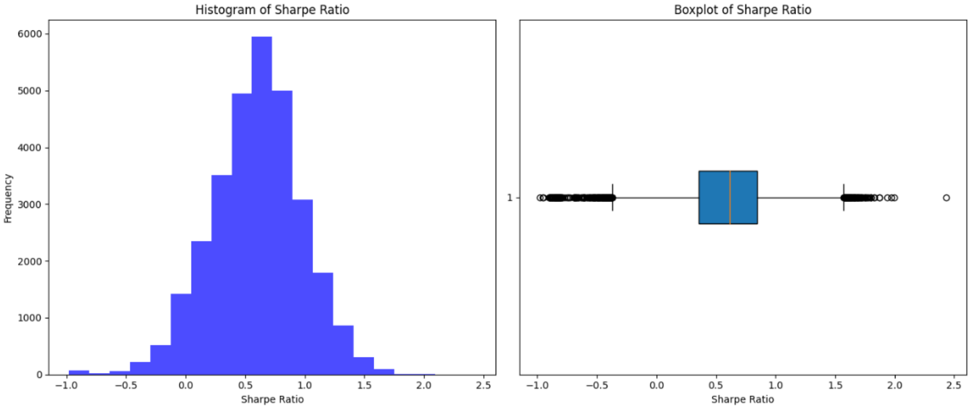


Figure 4.16: Histogram of Sharpe Ratio: K-Means Method

4.2.3 Top 5 Pairs

Table 4.8 presents the performance metrics of the top 5 pairs (sorted by Sharpe ratio), including Cumulative Return (%), Sharpe Ratio, Sortino Ratio, Maximum Drawdown, and the Total Trades executed. For instance, the pair AWK_TGT achieved a cumulative return of 338.69% with a Sharpe ratio of 2.43 and a Sortino ratio of 3.13, reflecting strong risk-adjusted performance. Additionally, this pair executed 33 trades over the observed period with a relatively controlled maximum drawdown of -0.09. Other pairs, such as DVA_JPM and AEE_DLTR, also demonstrate favorable performance with cumulative returns exceeding 200% and strong Sharpe and Sortino ratios, highlighting their profitability and risk management.

Pair	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown	Total Trades
AWK_TGT	338.69	2.43	3.13	-0.09	33
DVA_JPM	230.18	1.99	2.81	-0.08	12
AEE_DLTR	241.07	1.97	2.54	-0.08	23
PM_SO	148.94	1.94	2.90	-0.07	25
AMT_TSN	210.41	1.87	2.59	-0.10	16

Table 4.8: Performance of Top 5 Pairs: K-Means Method

Pair	Squared Distance	Sector 1	Sector 2	Half-life	Cointegrated
AWK_TGT	31.67	Utilities	Consumer Defensive	16.58	yes
DVA_JPM	49.20	Healthcare	Financial	58.82	no
AEE_DLTR	28.78	Utilities	Consumer Defensive	24.42	no
PM_SO	16.82	Consumer Defensive	Utilities	23.68	no
AMT_TSN	50.60	Real Estate	Consumer Defensive	39.45	no

Table 4.9: Features of Top 5 Pairs - K-Means Method

Table 4.9 complements this by detailing the characteristics of the same top 5 pairs. It includes information about the Squared Distance, the sectors for each pair (Sector 1 and Sector 2), the Half-life of the pairs' mean reversion process, and whether the pairs are Cointegrated. For example, AWK_TGT has a squared distance of 31.67, both stocks belong to the Utilities and Consumer Defensive sectors, and the pair exhibits a relatively short half-life of 16.58 weeks, indicating quicker mean reversion. Additionally, this pair is cointegrated, suggesting a stable long-term relationship. On the other hand, pairs like DVA_JPM and AMT_TSN show larger squared distances and longer half-lives, with no cointegration, which may impact their long-term reliability despite their high cumulative returns.

Together, these tables provide both performance and sector-based characteristics, offering a holistic view of the top 5 pairs selected through K-Means clustering. The performance metrics help assess the profitability and risk profile of each pair, while the features provide insights into the structural relationships between the pairs.

Performance Analysis: AWK_TGT

The pair AWK_TGT has been identified by the K-Means method as the top performing pair. Table 4.10 provides a comparison between the performance of the trading strategy and the Buy & Hold approaches for the stocks AWK and TGT.

The strategy significantly outperforms the Buy & Hold approaches in terms of cumulative return and risk-adjusted metrics. The strategy achieves a cumulative return of 338.69%, with a Sharpe ratio of 2.43 and a Sortino ratio of 3.13, indicating strong risk-adjusted performance. The maximum drawdown for the strategy is relatively small at -0.09, showcasing its ability to manage downside risk effectively.

Column	**Cumulative Return (%)**	**Sharpe Ratio**	**Sortino Ratio**	**Maximum Drawdown**
Strategy	338.69	2.43	3.13	-0.09
Buy & Hold AWK	24.33	0.30	0.41	-0.37
Buy & Hold TGT	55.89	0.43	0.56	-0.59

Table 4.10: Performance Comparison: Strategy vs. Buy & Hold Approaches

In contrast, the Buy & Hold approach for AWK yields a cumulative return of 24.33%, with a Sharpe ratio of 0.30 and a Sortino ratio of 0.41, alongside a maximum drawdown of -0.37. Similarly, the Buy & Hold approach for TGT results in a cumulative return of 55.89%, with a Sharpe ratio of 0.43 and a Sortino ratio of 0.56, but experiences a more significant drawdown of -0.59.

Overall, the table highlights the superior performance of the strategy compared to the passive Buy & Hold approaches for both stocks, particularly in terms of managing risk and generating higher returns.

Figure 4.17 illustrates the equity curves of the pairs trading strategy applied to the AWK_TGT pair, compared to the Buy & Hold strategies for the individual stocks AWK and TGT.

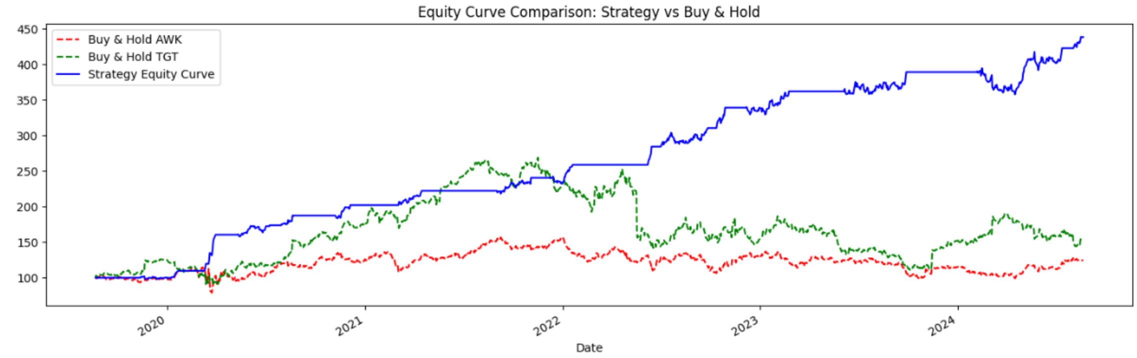


Figure 4.17: Equity Curves: AWK_TGT

Figure 4.18 depicts the Z-Spread between AWK and TGT along with the associated trading positions.

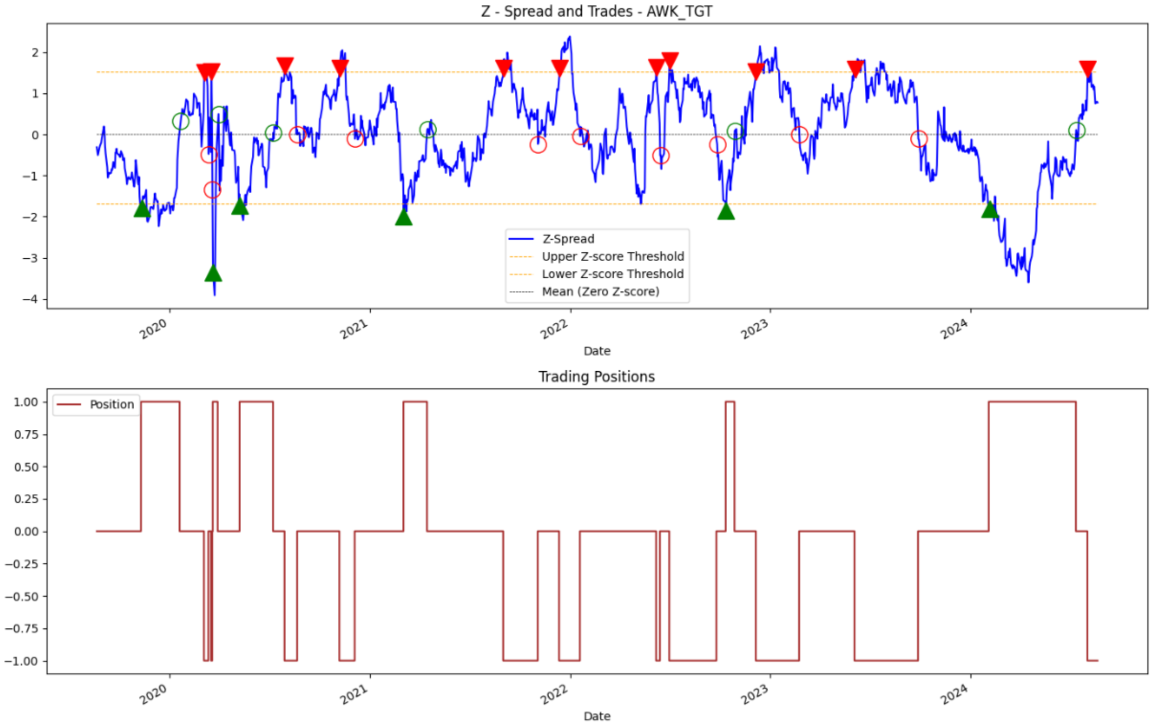


Figure 4.18: Z-Spread and Trading Positions for Pair: AWK_TGT

4.3 DBSCAN Method

4.3.1 Selecting the eps Parameter

Choosing an appropriate value for the eps parameter is crucial for the performance of the DBSCAN algorithm. The eps value defines the maximum distance between two points for them to be considered as neighbors. Selecting a proper value for eps can significantly affect the resulting clusters, as too small a value may lead to many points being classified as noise, while too large a value can cause different clusters to merge, reducing the ability to identify distinct groups in the data.

There are several strategies to determine the optimal eps value:

- **K-Nearest Neighbors Plot (KNN Plot)**: One common approach is to plot the distances to the k-nearest neighbor for each point in the dataset, typically with $k = min\_samples$. The "elbow" in the plot, where the distance begins to increase significantly, is often a good indication of the appropriate eps value.
- **Grid Search**: A grid search can be performed over different values of eps to identify the value that yields the best clustering results based on evaluation metrics like silhouette score.

A grid search was performed to identify the optimal eps value for the DBSCAN algorithm. The results, as shown in Table 4.11, include the silhouette score, number of clusters, and noise count for each eps value tested.

eps	Silhouette Score	Number of Clusters	Noise Count
0.40	0.8931	2	429
0.45	0.5244	3	337
0.50	0.7099	4	233
0.55	0.6954	4	163
0.60	0.6827	4	117
0.65	0.6063	5	73
0.70	0.6020	5	54
0.75	0.5960	5	36
0.80	0.5939	5	29
0.85	0.5899	5	20

Table 4.11: Grid Search Results for eps Selection

After assessing the trade-offs between clustering performance and noise, an eps value of 0.65 was selected. This value strikes a balance between the number of clusters and the silhouette score, resulting in 5 clusters with a moderate silhouette score of 0.6063 and a relatively low noise count of 73. The minimum sample size per cluster was set to 10. Figure 4.19 presents a 3D visualization of the clusters identified by the DBSCAN algorithm. After applying combinatorics to the clusters, a total of **26,588** unique pairs were generated.

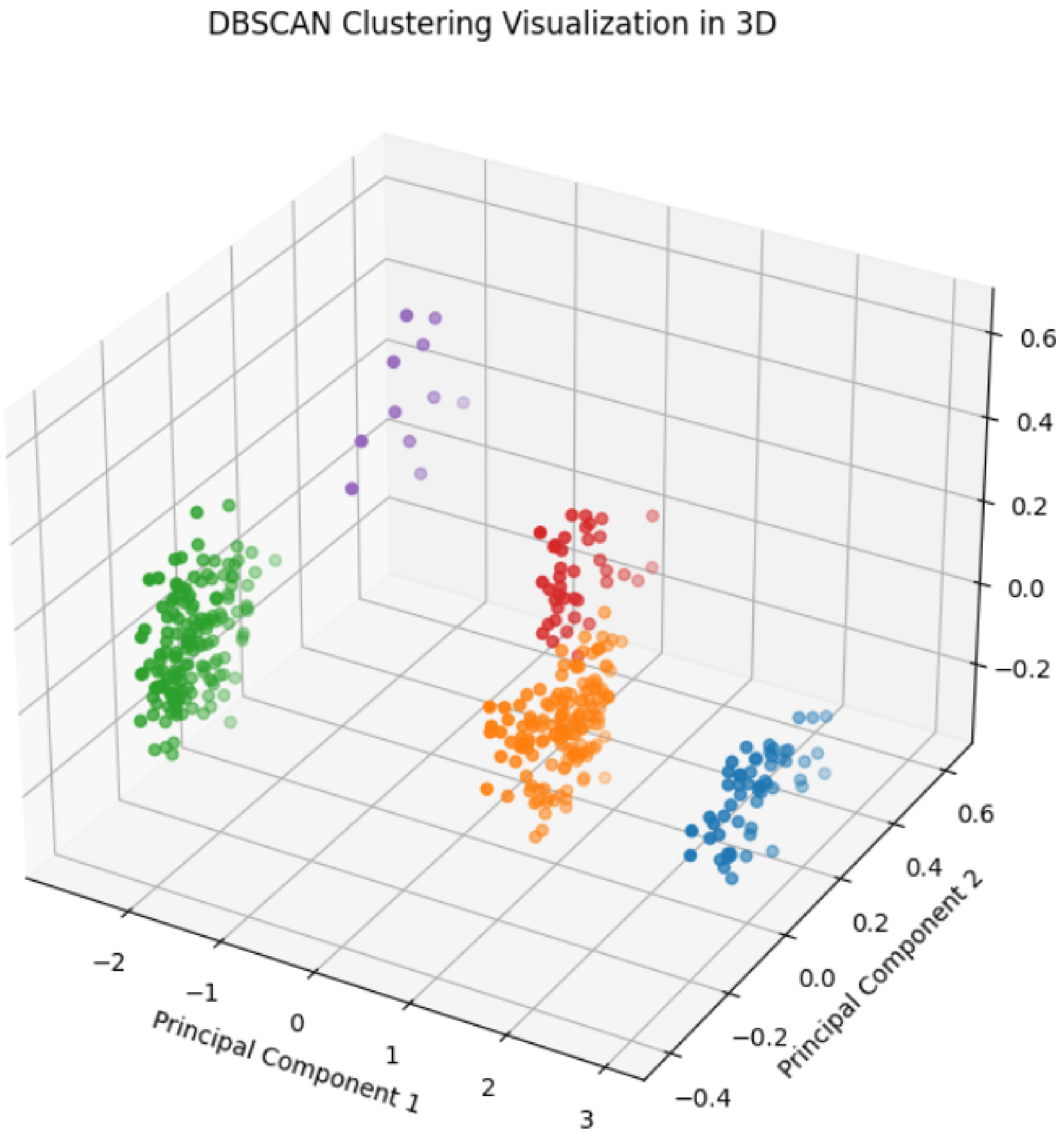


Figure 4.19: DBSCAN Clusters

4.3.2 Performance Metrics

The table 4.12 summarizes the performance metrics of a trading strategy based on the DBSCAN method. The metrics provided include Cumulative Return (%), Sharpe Ratio, Sortino Ratio, Maximum Drawdown, and Total Trades. On average, the strategy delivered a cumulative return of 56.24%, with a Sharpe ratio of 0.61, reflecting moderate risk-adjusted performance. The median values indicate that most trades achieved a cumulative return of 49.87%, a Sharpe ratio of 0.62, and a Sortino ratio of 0.64, suggesting consistent performance across the trades. Of the 26,588 pairs analyzed, 23,161 pairs recorded positive cumulative returns, highlighting the strategy's general profitability. The maximum return observed was 544.55%, demonstrating the potential for significant gains, while the minimum return of -60.42% highlights the risk of substantial losses.

Metric	**Cumulative Return (%)**	**Sharpe Ratio**	**Sortino Ratio**	**Maximum Drawdown**	**Total Trades**
Mean	56.24	0.61	0.66	-0.24	7.18
Median	49.87	0.62	0.64	-0.22	6.00
Min	-60.42	-0.79	-0.78	-0.84	2.00
Max	544.55	2.43	3.13	-0.01	33.00
Standard Deviation	49.96	0.37	0.43	0.12	3.89

Table 4.12: Performance Metrics Summary: DBSCAN Method

Additionally, the strategy's risk exposure is quantified by the maximum drawdown, with a mean value of -0.24, indicating moderate average capital loss from peak to trough. The standard deviation values for all metrics provide insights into the variability of the strategy's outcomes, with a standard deviation of 49.96% for cumulative returns and 0.37 for the Sharpe ratio. The number of trades, with a mean of 7.18 trades and a standard deviation of 3.89, suggests that the strategy was moderately active. Overall, the DBSCAN method demonstrates balanced performance with moderate risk-adjusted returns and relatively controlled drawdowns, though there are instances of high variability and potential for both large gains and losses.

Figure 4.20 shows the distribution of cumulative returns for the trading pairs identified using the DBSCAN method, highlighting the range and frequency of returns.

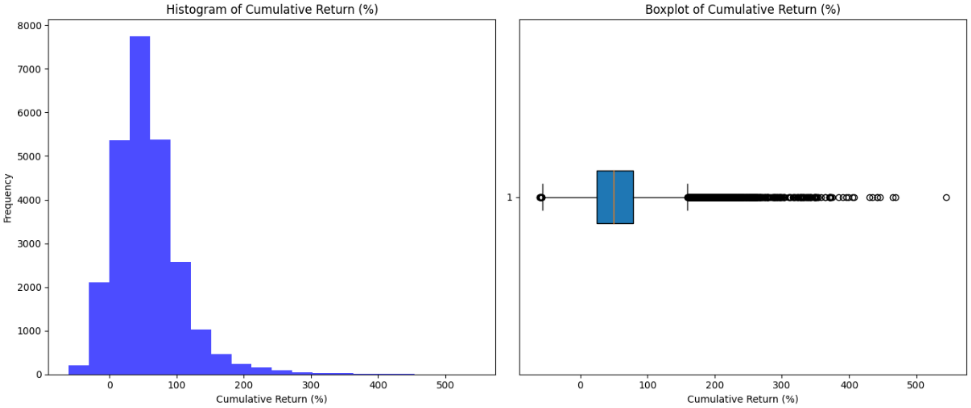


Figure 4.20: Histogram of Cumulative Returns: DBSCAN Method

Figure 4.21 displays the distribution of Sharpe ratios for the trading pairs identified using the DBSCAN method, providing insights into the risk-adjusted performance across pairs.

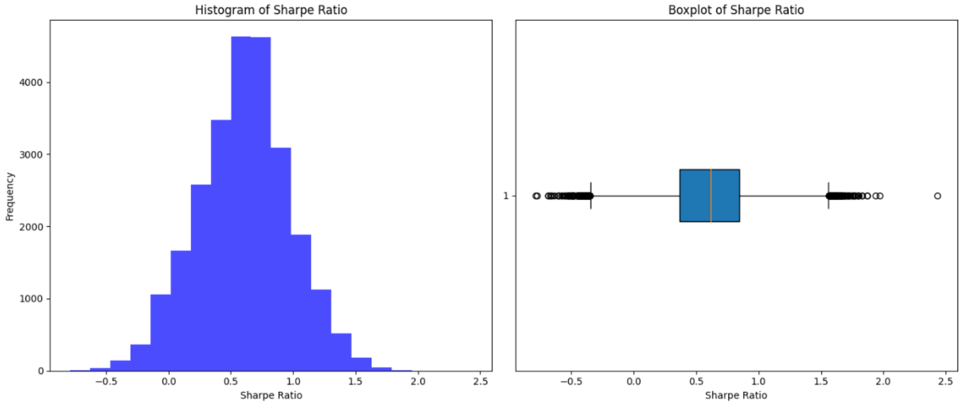


Figure 4.21: Histogram of Sharpe Ratio: DBSCAN Method

4.3.3 Top 5 Pairs

The table 4.13 presents the performance metrics of the top 5 stock pairs (sorted by Sharpe ratio) identified using the DBSCAN clustering method. The key metrics include cumulative return (%), Sharpe ratio, Sortino ratio, maximum drawdown, and total trades. The pair AWK_TGT recorded the highest cumulative return of 338.69%, with a Sharpe ratio of 2.43 and a Sortino ratio of 3.13, indicating strong risk-adjusted performance. This pair executed 33 trades with a maximum drawdown of -0.09, suggesting that the strategy was relatively stable during adverse market conditions.

Other pairs, such as AEE_DLTR and AMT_TSN, also showed solid performance, with cumulative returns of 241.07% and 210.41%, respectively. Their Sharpe and Sortino ratios suggest consistent risk-adjusted returns. The pair ABBV_PM, while having a lower cumulative return of 116.73%, still maintained a Sharpe ratio of 1.87 and a Sortino ratio of 2.14, indicating decent performance with a controlled maximum drawdown of -0.09.

Pair	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown	Total Trades
AWK_TGT	338.69	2.43	3.13	-0.09	33
AEE_DLTR	241.07	1.97	2.54	-0.08	23
PM_SO	148.94	1.94	2.90	-0.07	25
AMT_TSN	210.41	1.87	2.59	-0.10	16
ABBV_PM	116.73	1.87	2.14	-0.09	15

Table 4.13: Performance of Top 5 Pairs: DBSCAN Method

Pair	Squared Distance	Sector 1	Sector 2	Half-life	Cointegrated
AWK_TGT	31.67	Utilities	Consumer Defensive	16.58	yes
AEE_DLTR	28.78	Utilities	Consumer Defensive	24.42	no
PM_SO	16.82	Consumer Defensive	Utilities	23.68	no
AMT_TSN	50.60	Real Estate	Consumer Defensive	39.45	no
ABBV_PM	22.45	Healthcare	Consumer Defensive	29.63	no

Table 4.14: Features of Top 5 Pairs - DBSCAN Method

The table 4.14 presents the key characteristics of the top 5 pairs identified using the DBSCAN method. It includes the squared distance between pairs, the sectors of the two stocks, the half-life of their mean reversion, and whether the pairs are cointegrated. For example, the pair AWK_TGT has a squared distance of 31.67, belongs to the Utilities and Consumer Defensive sectors, has a half-life of 16.58 weeks, and is cointegrated. Most pairs exhibit relatively short half-lives, but only AWK_TGT is cointegrated.

Performance Analysis: AEE_DLTR

The pair AWK_TGT has been identified by the K-Means as well as DBSCAN methods as the top performing pair (based on Sharpe ratio). Since we have analysed this pair in the previous section, we shall analyze the second best pair AEE_DLTR here.

The table 4.15 compares the performance of the trading strategy applied to the AEE_DLTR pair against a simple buy-and-hold approach for each stock individually. The strategy outperforms both buy-and-hold approaches significantly, achieving a cumulative return of 241.07%, with a Sharpe ratio of 1.97 and a Sortino ratio of 2.54, indicating strong risk-adjusted returns. The maximum drawdown of -0.08 shows that the strategy was able to minimize losses, making it more stable compared to the buy-and-hold methods.

Column	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown
Strategy	241.07	1.97	2.54	-0.08
Buy & Hold AEE	23.05	0.29	0.37	-0.29
Buy & Hold DLTR	3.24	0.21	0.26	-0.47

Table 4.15: Performance Comparison: Strategy vs. Buy & Hold Approaches for AEE_DLTR

In contrast, the buy-and-hold approach for AEE yielded a cumulative return of 23.05%, with a much lower Sharpe ratio of 0.29 and a Sortino ratio of 0.37, while DLTR returned only 3.24% with a Sharpe ratio of 0.21 and a Sortino ratio of 0.26. Additionally, both buy-and-hold strategies experienced much higher maximum drawdowns, -0.29 for AEE and -0.47 for DLTR, indicating greater exposure to downside risk. These comparisons highlight the effectiveness of the trading strategy in generating higher returns while managing risk more effectively.

Figure 4.22 illustrates the equity curves of the trading strategy compared to the buy-and-hold approaches for the AEE_DLTR pair.

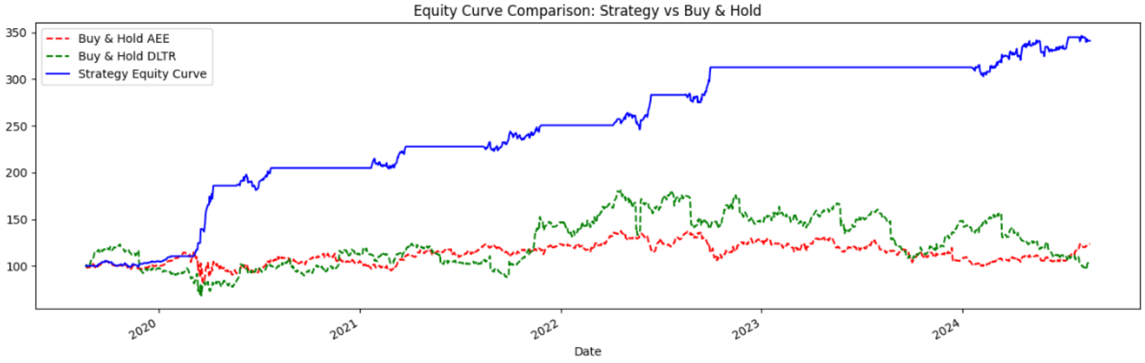


Figure 4.22: Equity Curves: AEE_DLTR

Figure 4.23 shows the Z-Spread and the corresponding trading positions for the AEE_DLTR pair, providing insights into the timing of trades based on the spread.

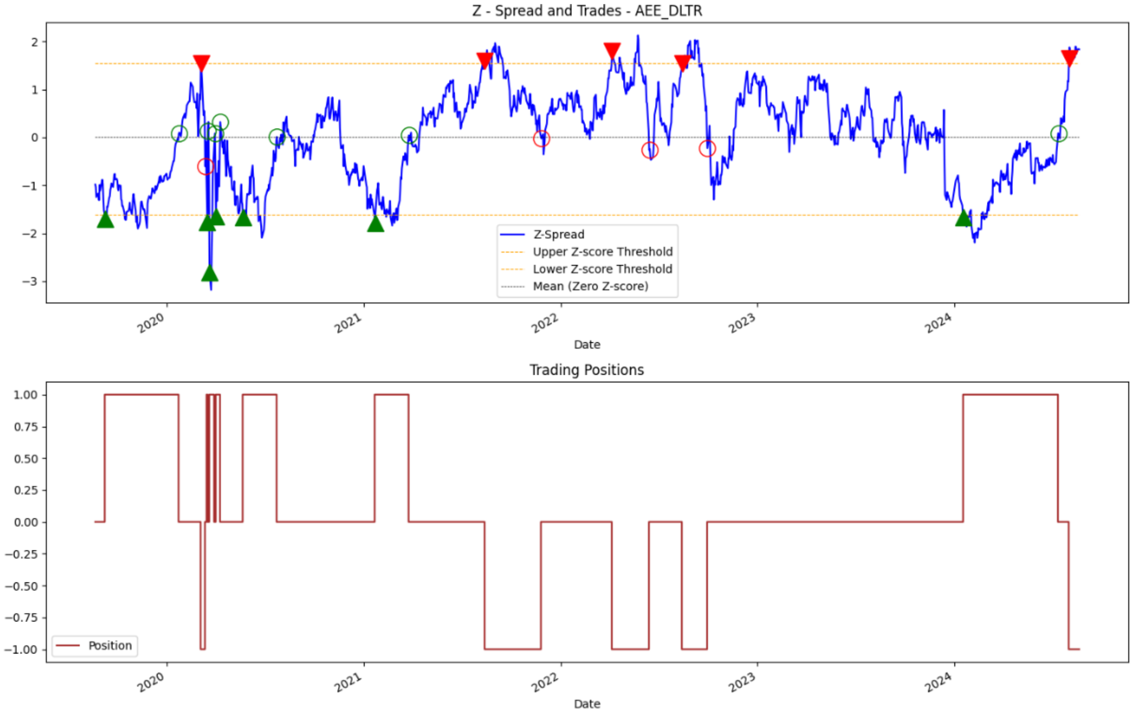


Figure 4.23: Z-Spread and Trading Positions for Pair: AEE_DLTR

4.4 OPTICS Method

4.4.1 Parameter Selection

In the OPTICS model, the min_samples parameter was set to 10, consistent with the settings used in the DBSCAN model, ensuring a minimum density requirement for core points. All other parameters were left at their default values, allowing OPTICS to automatically adjust to the data's density structure.

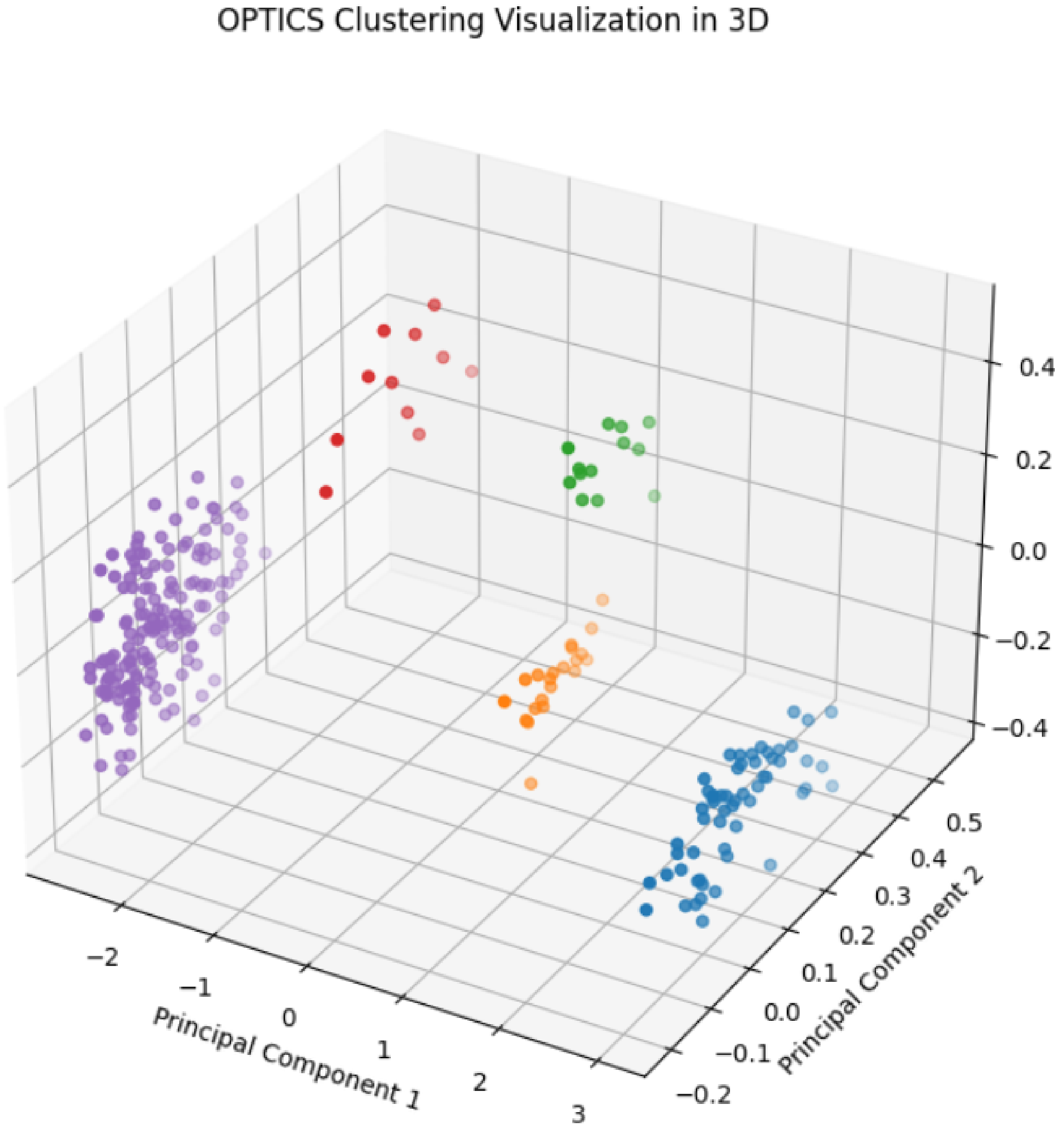


Figure 4.24: OPTICS Clusters

By relying on the default settings for parameters such as max_eps and xi, the model leveraged its inherent flexibility to detect clusters of varying densities while still maintaining robustness against noise. The OPTICS model achieved a Silhouette Score of 0.623 (excluding noise), indicating a good level of cluster cohesion and separation. Five clusters were formed and a total of **14,659** unique pairs were identified, highlighting the model's effectiveness in discovering meaningful relationships within the data.

4.4.2 Performance Metrics

The table 4.16 summarizes the performance metrics of the trading strategy based on the OPTICS method. The metrics include Cumulative Return (%), Sharpe Ratio, Sortino Ratio, Maximum Drawdown, and Total Trades. The strategy's mean cumulative return of 55.76% indicates a solid overall performance, although it is slightly lower than the maximum return of 544.55%, reflecting variability in returns across different trades. The mean Sharpe ratio of 0.59 and the Sortino ratio of 0.64 suggest that the strategy provides moderate risk-adjusted returns. The median values for these metrics are consistent with the means, indicating that the majority of trades exhibit similar performance levels.

Metric	**Cumulative Return (%)**	**Sharpe Ratio**	**Sortino Ratio**	**Maximum Drawdown**	**Total Trades**
Mean	55.76	0.59	0.64	-0.25	6.43
Median	49.61	0.61	0.62	-0.22	6.00
Min	-60.42	-0.66	-0.73	-0.80	2.00
Max	544.55	1.76	2.77	-0.01	29.00
Standard Deviation	51.21	0.35	0.40	0.12	3.17

Table 4.16: Performance Metrics Summary: OPTICS Method

In terms of risk management, the maximum drawdown has a mean of -0.25, suggesting that, on average, the strategy managed to limit losses during adverse market conditions. However, the minimum drawdown reached -0.80, highlighting potential risks in extreme scenarios. The standard deviation of the cumulative returns is 51.21%, pointing to a relatively high variability in performance. The total number of trades, with a mean of 6.43 and a standard deviation of 3.17, indicates a moderately active strategy that balances trading frequency with a focus on risk and return. Overall, the OPTICS method demonstrates a balanced approach to generating returns while managing downside risk.

Figure 4.25 displays the distribution of cumulative returns for the trading strategy using the OPTICS method, highlighting the range and frequency of returns.

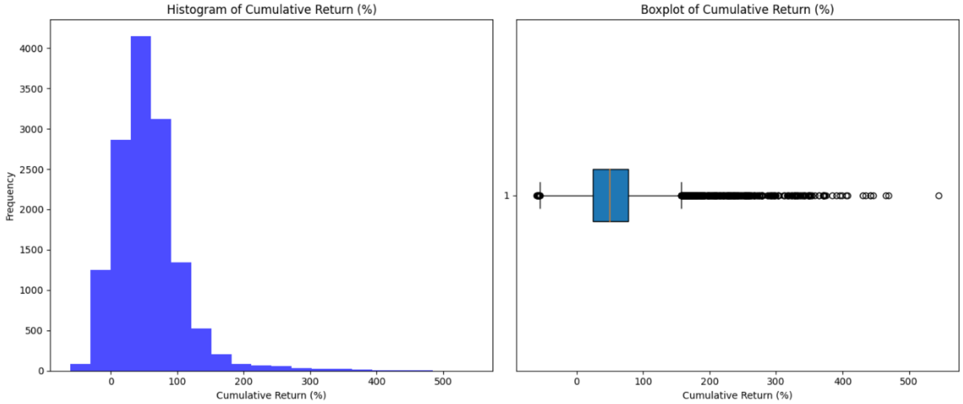


Figure 4.25: Histogram of Cumulative Returns: OPTICS Method

Figure 4.26 illustrates the distribution of Sharpe ratios for the trading strategy using the OPTICS method, providing insights into the risk-adjusted performance of the strategy.

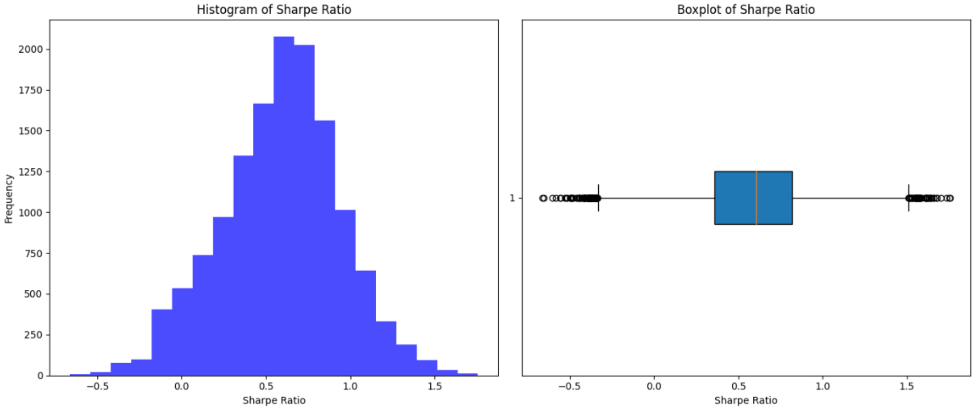


Figure 4.26: Histogram of Sharpe Ratio: OPTICS Method

4.4.3 Top 5 Pairs

The table 4.17 provides a summary of the performance metrics for the top 5 pairs identified using the OPTICS method. It includes key metrics such as Cumulative Return (%), Sharpe Ratio, Sortino Ratio, Maximum Drawdown, and Total Trades for each pair. The pair FOX_GPC achieved the highest cumulative return of 371.16%, with a Sharpe ratio of 1.68 and a Sortino ratio of 1.65, indicating a favorable balance between risk and return despite a relatively higher maximum drawdown of -0.13.

Other pairs, such as AMP_HCA and GL_RJF, also performed well, with cumulative returns of 197.84% and 297.16% respectively, along with high Sharpe and Sortino ratios, suggesting strong risk-adjusted returns. The pairs J_XYL and CMS_DUK exhibited moderate cumulative returns but maintained relatively low maximum drawdowns of -0.05, highlighting their stability during adverse market conditions. Overall, the table demonstrates the effectiveness of the OPTICS method in identifying pairs that offer substantial returns while managing risk effectively.

Pair	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown	Total Trades
AMP_HCA	197.84	1.76	2.42	-0.10	20
GL_RJF	297.16	1.75	1.71	-0.09	13
J_XYL	88.22	1.74	2.06	-0.05	13
CMS_DUK	81.79	1.70	1.65	-0.05	16
FOX_GPC	371.16	1.68	1.65	-0.13	16

Table 4.17: Performance of Top 5 Pairs: OPTICS Method

Pair	Squared Distance	Sector 1	Sector 2	Half-life	Cointegrated
AMP_HCA	6.35	Financial	Healthcare	17.29	yes
GL_RJF	36.95	Financial	Financial	32.07	no
J_XYL	36.35	Industrials	Industrials	86.05	no
CMS_DUK	15.94	Utilities	Utilities	20.66	no
FOX_GPC	64.93	Comm. Services	Consumer Cyclical	45.80	no

Table 4.18: Features of Top 5 Pairs - OPTICS Method

The table 4.18 outlines the characteristics of the top 5 pairs identified using the OPTICS method. It includes the squared distance between pairs, the sectors of each stock in the pair, the half-life of their mean reversion, and whether the pairs are cointegrated. The pair AMP_HCA shows a squared distance of 6.35, suggesting a close relationship between the two stocks, both of which belong to the Financial and Healthcare sectors. This pair also has a relatively short half-life of 17.29, indicating faster mean reversion, and is cointegrated, pointing to a stable long-term relationship.

In contrast, pairs like GL_RJF and J_XYL, while having larger squared distances of 36.95 and 36.35, respectively, are not cointegrated, indicating less stable long-term relationships. The pair

FOX_GPC, with the highest squared distance of 64.93, represents stocks from the Communication Services and Consumer Cyclical sectors, with a longer half-life of 45.80, implying slower mean reversion. The varied characteristics across these pairs demonstrate the OPTICS method's capability to identify both closely related pairs with quick mean reversion and those with more divergent behaviors.

Performance Analysis:AMP_HCA

The table 4.19 presents a comparison of the performance metrics for the trading strategy applied to the AMP_HCA pair against the buy-and-hold approaches for AMP and HCA individually. The trading strategy achieved a cumulative return of 197.84% with a Sharpe ratio of 1.76 and a Sortino ratio of 2.42, indicating strong risk-adjusted performance. The strategy's maximum drawdown of -0.10 highlights its ability to effectively manage downside risk while still generating substantial returns, making it a robust approach compared to the buy-and-hold strategies.

Column	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown
Strategy	197.84	1.76	2.42	-0.10
Buy & Hold AMP	272.71	0.87	1.11	-0.54
Buy & Hold HCA	217.28	0.80	1.03	-0.55

Table 4.19: Performance Comparison: Strategy vs. Buy & Hold Approaches for AMP_HCA

In contrast, the buy-and-hold approach for AMP resulted in a higher cumulative return of 272.71% but exhibited a significantly lower Sharpe ratio of 0.87 and Sortino ratio of 1.11, indicating less efficient risk-adjusted returns. Similarly, the buy-and-hold approach for HCA produced a cumulative return of 217.28% with a Sharpe ratio of 0.80 and a Sortino ratio of 1.03. Both buy-and-hold strategies experienced larger maximum drawdowns of -0.54 for AMP and -0.55 for HCA, reflecting greater vulnerability to market declines. These comparisons underscore the advantage of the trading strategy in achieving balanced returns with improved risk management.

Figure 4.27 displays the equity curves of the trading strategy compared to the buy-and-hold approaches for the AMP_HCA pair.

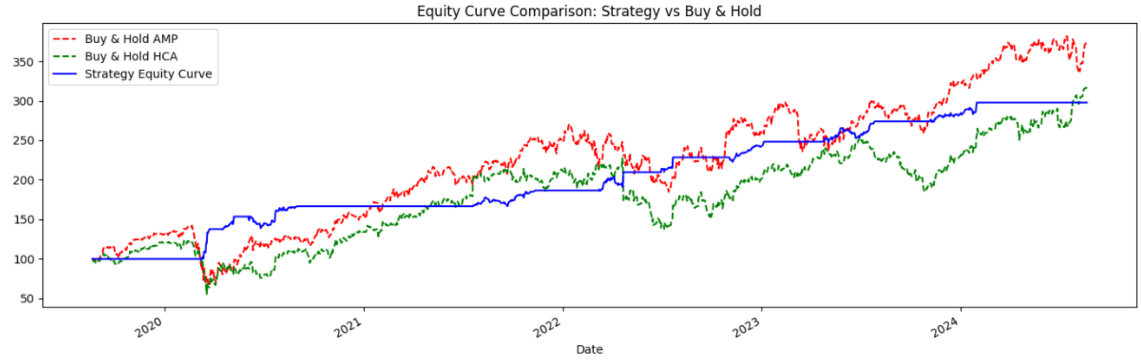


Figure 4.27: Equity Curves: AMP_HCA

Figure 4.28 illustrates the Z-Spread and the corresponding trading positions for the AMP_HCA pair, highlighting the timing of trades based on the spread.

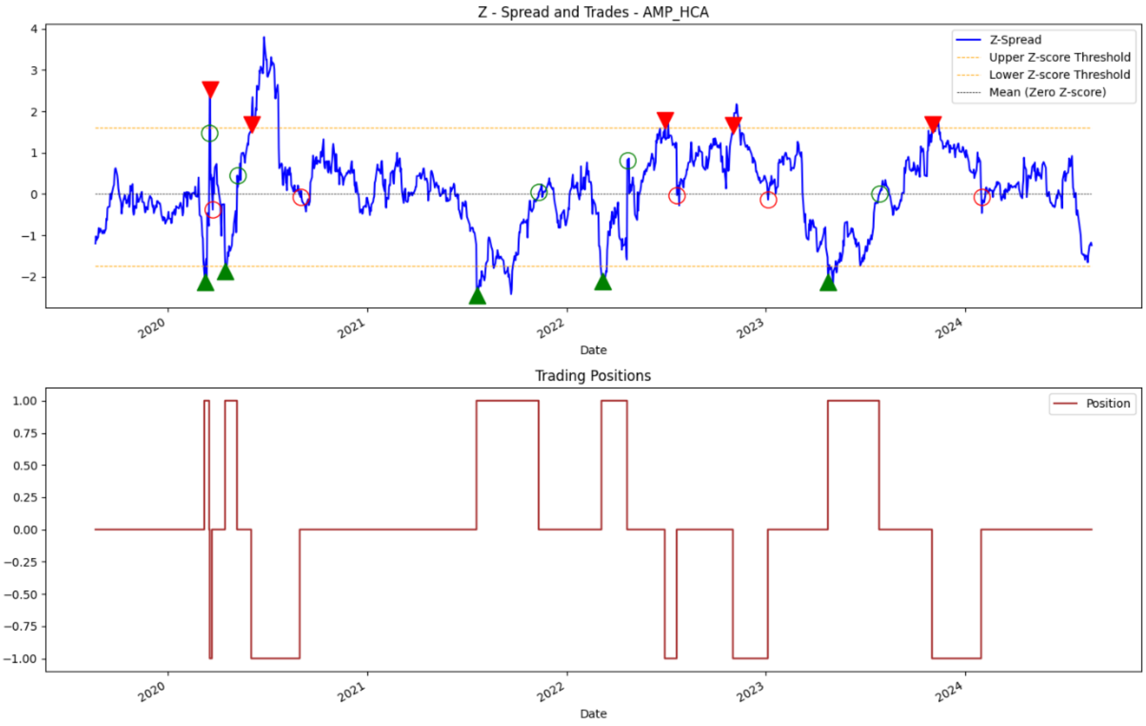


Figure 4.28: Z-Spread and Trading Positions for Pair: AMP_HCA

Chapter 5

Discussion

5.1 Comparison of Methods

5.1.1 Portfolio Analysis

This section examines the performance of the portfolio (Table A.1 to Table A.4) of the top 25 pairs identified through the distance method and various clustering techniques (K-Means, DBSCAN, and OPTICS). The focus of the analysis is to compare each method based on key performance metrics, including cumulative return, Sharpe ratio, and other relevant indicators. Furthermore, the discussion provides a critical assessment of each method's overall effectiveness, highlighting their strengths and limitations in identifying profitable pairs and optimizing risk-adjusted returns.

The comparison table (Table 5.1) provides a comprehensive analysis of the performance metrics of portfolios constructed using the top 25 pairs identified by four different methods: Distance Method, K-Means Method, DBSCAN Method, and OPTICS Method. The metrics evaluated include mean, median, and standard deviation values for cumulative return, Sharpe ratio, Sortino ratio, maximum drawdown, and total trades.

Metric	Distance Portfolio	K-Means Portfolio	DBSCAN Portfolio	OPTICS Portfolio
Mean Values				
Cumul. Return (%)	147.69	188.19	186.36	175.38
Sharpe Ratio	1.83	1.82	1.81	1.64
Sortino Ratio	2.07	2.26	2.24	1.82
Max Drawdown	-0.09	-0.10	-0.10	-0.09
Total Trades	19.72	18.80	19.40	15.80
Median Values				
Cumul. Return (%)	144.38	166.82	167.76	129.05
Sharpe Ratio	1.79	1.78	1.77	1.63
Sortino Ratio	2.04	2.18	2.18	1.76
Max Drawdown	-0.09	-0.09	-0.09	-0.08
Total Trades	18.00	19.00	19.00	14.00
Standard Deviation Values				
Cumul. Return (%)	35.43	64.73	64.06	111.44
Sharpe Ratio	0.11	0.15	0.15	0.05
Sortino Ratio	0.33	0.38	0.37	0.28
Max Drawdown	0.02	0.02	0.02	0.03
Total Trades	6.69	5.61	5.09	4.44

Table 5.1: Performance Comparison Across Methods

5.1.2 Mean Values Analysis

- **Cumulative Return (%)**: The K-Means method has the highest mean cumulative return at 188.19%, followed closely by DBSCAN with 186.36%, and OPTICS at 175.38%. The

Distance method has the lowest mean cumulative return at 147.69%, indicating that the K-Means method was the most effective in generating overall portfolio returns.

- **Sharpe Ratio**: The mean Sharpe ratio values are close for the Distance (1.83), K-Means (1.82), and DBSCAN (1.81) methods, suggesting similar risk-adjusted returns. The OPTICS method has a lower Sharpe ratio of 1.64, indicating lower efficiency in risk-adjusted performance.
- **Sortino Ratio**: The K-Means method leads with a Sortino ratio of 2.26, focusing on downside risk, followed closely by DBSCAN with 2.24. The Distance method (2.07) and OPTICS (1.82) lag behind, indicating less effective downside protection.
- **Maximum Drawdown**: The maximum drawdown is similar across methods, with values ranging from -0.09 to -0.10, indicating comparable worst-case performance across the portfolios.
- **Total Trades**: The Distance method has the highest average number of trades (19.72), indicating a more active trading approach, while OPTICS has the lowest average with 15.80 trades, suggesting a more conservative strategy.

5.1.3 Median Values Analysis

- **Cumulative Return (%)**: The median values confirm that K-Means (166.82%) slightly outperforms DBSCAN (167.76%) in cumulative returns, followed by OPTICS (129.05%), with the Distance method remaining the lowest at 144.38%.
- **Sharpe Ratio**: The Distance (1.79), K-Means (1.78), and DBSCAN (1.77) methods are closely aligned, with OPTICS slightly lower at 1.63, indicating relative underperformance in risk-adjusted returns.
- **Sortino Ratio**: K-Means maintains its lead with a median Sortino ratio of 2.18, closely followed by DBSCAN, while Distance and OPTICS show lower performance in downside risk adjustment.
- **Maximum Drawdown**: Median drawdown values are similar across the methods, with DBSCAN and OPTICS showing slightly better resistance to drawdowns at -0.09.
- **Total Trades**: The Distance method's higher median number of trades (18.00) suggests a consistently active trading strategy, while OPTICS has a lower trading frequency, with a median of 14.00 trades.

5.1.4 Standard Deviation Values Analysis

- **Cumulative Return (%)**: The Distance method exhibits the lowest standard deviation in cumulative returns (35.43), indicating more stable performance, while OPTICS has the highest variability (111.44), reflecting more significant fluctuations.
- **Sharpe Ratio**: The OPTICS method shows the lowest standard deviation in the Sharpe ratio (0.05), indicating more consistent risk-adjusted returns. In contrast, K-Means and DBSCAN display higher variability (0.15).
- **Sortino Ratio**: K-Means and DBSCAN exhibit relatively high standard deviation in Sortino ratios (0.38 and 0.37), indicating variability in downside risk management, while Distance has more stable risk-adjusted returns.
- **Maximum Drawdown**: The maximum drawdown values are relatively consistent across methods, with OPTICS showing slightly higher variability (0.03).
- **Total Trades**: The Distance method has a lower deviation in trading activity (6.69), suggesting a more uniform trading approach, while OPTICS has the least variability (4.44), indicating consistent selectivity.

5.1.5 Summary of Findings

- **Best Performing Method**: The K-Means method is the most effective approach, achieving the highest cumulative returns and superior risk-adjusted performance, making it the most profitable and efficient method.
- **Stability**: The Distance method demonstrates the most stable returns, indicated by the lowest standard deviation in cumulative returns, making it a reliable choice for stable performance.
- **Risk Management**: Both K-Means and DBSCAN excel in managing downside risk, as reflected in their higher Sortino ratios, while OPTICS shows lower risk-adjusted returns with higher variability.

5.2 Impact of Transaction Cost

To assess the impact of transaction costs on the cumulative return of different trading strategies, we adjusted the cumulative return by incorporating a fixed transaction cost (0.5%) per trade. This adjustment helps to better understand the actual profitability of each strategy after accounting for trading expenses. The methodology involves the following steps:

- **Calculate the Average Number of Trades:** We first determine the average number of trades executed for each method. This is crucial because strategies with a higher frequency of trades incur more transaction costs, which can significantly impact overall profitability.

- **Determine the Transaction Cost:** A constant transaction cost is applied to each trade. This cost represents the expenses associated with executing a trade, such as brokerage fees, slippage, and other charges.
- **Adjust the Cumulative Return:** The adjusted cumulative return is calculated by subtracting the total transaction costs (average number of trades multiplied by the transaction cost) from the initial cumulative return. This adjusted value provides a clearer picture of the strategy's performance after accounting for trading expenses.

Table 5.2 compares the impact of transaction costs on the average cumulative return across different clustering methods by factoring in the average number of trades executed. The adjusted cumulative return reflects a reduction from the initial average cumulative return after accounting for transaction costs. The K-Means method maintains the highest adjusted cumulative return at 170.50%, despite a moderate trade frequency, indicating its relative efficiency in generating returns.

Metric	**Distance Portfolio**	**K-Means Portfolio**	**DBSCAN Portfolio**	**OPTICS Portfolio**
Avg Cumul. Return (%)	147.69	188.19	186.36	175.38
Avg Trades	19.72	18.80	19.40	15.80
Adj Cumul. Return (%)	133.13	170.50	168.28	161.52

Table 5.2: Adjusted Cumulative Return for Each Method Considering Transaction Costs

The Distance and DBSCAN methods show slightly lower adjusted returns at 133.13% and 168.28%, respectively, largely due to a higher number of trades increasing transaction costs. The OPTICS method, with the fewest trades, demonstrates a competitive adjusted cumulative return of 161.52%, suggesting that a lower trading frequency helps mitigate the impact of transaction costs, even if the initial returns are not the highest. This analysis highlights the importance of balancing trading frequency and returns to optimize the net performance of trading strategies.

5.3 Combined Portfolio

The Combined Portfolio A.5 was constructed by integrating pairs from four different methods. During this process, duplicate pairs were removed to ensure that each pair in the portfolio was unique. The remaining pairs were then ranked by their Sharpe ratios, and the top 25 pairs were selected to form the portfolio. This approach guarantees that the Combined Portfolio consists of the highest-performing pairs based on their risk-adjusted returns.

The results of the Combined Portfolio as shown in Table 5.3, demonstrate a well-constructed portfolio with strong performance metrics. The mean cumulative return of 173.93% indicates a significant overall gain, while the Sharpe ratio of 1.90 suggests the portfolio has an excellent risk-adjusted return compared to typical market benchmarks.

Statistic	Cumulative Return (%)	Sharpe Ratio	Sortino Ratio	Maximum Drawdown	Total Trades
Mean	173.93	1.90	2.29	-0.09	20.04
Median	153.78	1.84	2.22	-0.09	20.00
Standard Deviation	66.18	0.15	0.39	0.02	6.66

Table 5.3: Summary Statistics for Combined Portfolio

The median values of the cumulative return (153.78%) and Sharpe ratio (1.84) are close to the mean, highlighting that the performance distribution of the pairs is relatively consistent. This consistency is further supported by the low standard deviation of the Sharpe ratio (0.15), indicating that most pairs in the portfolio have similar risk-adjusted returns.

The Sortino ratio, with a mean of 2.29, reflects strong downside protection, suggesting that the portfolio has a good reward-to-risk ratio when considering negative volatility. The maximum drawdown of -0.09 indicates that the portfolio has relatively low exposure to significant losses, enhancing its stability.

By eliminating duplicate pairs and selecting the 25 best pairs based on their Sharpe ratios, the Combined 25 Portfolio aims to maximize returns while managing risk effectively. This careful selection process ensures that the portfolio consists of unique and high-performing pairs, making it a robust choice for achieving optimal performance.

5.4 Limitations of the Study

Despite the comprehensive analysis conducted in this research, several limitations must be acknowledged. First, the study's reliance on historical stock price data for backtesting does not fully account for the potential impact of changing market conditions or macroeconomic factors that could influence the future performance of pairs trading strategies. The assumption that past price patterns will continue to hold in the future may lead to overfitting and limit the strategies' effectiveness during periods of heightened market volatility or structural shifts in the economy.

Second, the sensitivity of the results to the specific period in which the analysis begins poses a significant challenge. Strategies that perform well during one historical period may not necessarily exhibit the same level of success in different market environments. This time-dependence can introduce biases and overstate the robustness of the trading strategy. Without methods to account for this sensitivity, such as cross-validation or rolling window analysis, the true reliability of the strategy remains uncertain.

Additionally, the fixed transaction cost model used in the analysis oversimplifies the complexities of real-world trading. In practice, transaction costs can vary significantly due to factors like slippage, liquidity constraints, and dynamic brokerage fees, which can have a substantial impact on the net profitability of trading strategies. Ignoring these variable costs could lead to an overestimation of the actual returns and fail to reflect the true costs of implementing the strategies.

While clustering algorithms like K-Means, DBSCAN, and OPTICS were employed to identify trading pairs, these methods have inherent limitations in capturing complex non-linear relationships between stock pairs. Such limitations may result in suboptimal pair selection, particularly when market dynamics deviate from traditional linear correlations. Advanced techniques that account for non-linear dependencies could provide a more accurate assessment of stock relationships.

Finally, the evaluation of pairs trading strategies primarily relied on standard risk-adjusted performance metrics like the Sharpe and Sortino ratios. While these metrics are useful, they may not fully capture the strategies' behavior under extreme market conditions or provide a complete picture of downside risks. The lack of alternative risk measures means that important aspects of the risk dynamics in pairs trading might have been overlooked, limiting the depth of the performance analysis.

Chapter 6

Conclusion

6.1 Summary of Research

This research presents a comprehensive analysis of pairs trading strategies using four distinct methods: Distance Method, K-Means Clustering, DBSCAN Clustering, and OPTICS Clustering. Each method's performance was evaluated based on its ability to identify profitable pairs, optimize risk-adjusted returns, and manage transaction costs effectively. The study's findings reveal that the K-Means method consistently outperformed the other methods in terms of cumulative returns and risk-adjusted metrics. It achieved the highest mean cumulative return and superior Sharpe and Sortino ratios, establishing it as the most effective strategy for maximizing returns while managing risk.

In contrast, the Distance method, although not producing the highest returns, demonstrated the most stable performance with the lowest standard deviation in cumulative returns, indicating more consistent results over time. The OPTICS method, despite its competitive risk management and the fewest trades, exhibited lower mean cumulative returns and risk-adjusted ratios compared to the other techniques, highlighting areas for potential improvement.

When transaction costs were incorporated into the analysis, the importance of balancing trading frequency with return optimization became apparent. The K-Means method retained the highest adjusted cumulative return even after accounting for transaction costs, underscoring its relative efficiency in generating returns. Strategies like the Distance and DBSCAN methods experienced a reduction in profitability due to their higher trade frequency, which increased transaction costs, emphasizing the need to consider these costs in strategy evaluation.

The Combined 25 Portfolio, constructed by integrating the best pairs from each method, showed robust performance with a mean cumulative return of 173.93% and a Sharpe ratio of 1.90. The portfolio's careful selection process ensured that it consisted of the highest-performing pairs, leading to consistent risk-adjusted returns and effective downside risk management. Its low maximum drawdown further underscored its stability and resilience in volatile market conditions.

In conclusion, this research highlights that while K-Means clustering is the most effective method for identifying profitable pairs with strong risk-adjusted returns, incorporating stability through approaches like the Distance method can yield more consistent performance. Moreover, the study underscores the significance of accounting for transaction costs to assess the true profitability of trading strategies accurately. Future research could explore advanced machine learning techniques to enhance clustering algorithms, develop dynamic transaction cost models, and investigate how different market conditions impact strategy performance. These areas present opportunities to refine pairs trading strategies further and optimize returns.

6.2 Recommendations

Several recommendations for future improvements are suggested to address the limitations identified in this study. First, incorporating machine learning and artificial intelligence techniques into the pairs selection process could significantly enhance the identification of non-linear relationships and patterns within the stock data. Techniques such as neural networks or reinforcement learning can be explored to develop adaptive trading strategies that can dynamically adjust to changing market conditions in real-time, leading to more robust and responsive models.

To account for the variability in trading strategy performance due to changes in market conditions, a dynamic transaction cost model should be implemented. This model would consider factors such as market liquidity, slippage, and variable brokerage fees, providing a more realistic and accurate evaluation of the strategy's profitability. Additionally, incorporating sensitivity analysis could help understand how different transaction cost levels impact the net returns of trading strategies, thereby improving cost efficiency.

Furthermore, to address the sensitivity of results to the specific period where the backtesting starts, incorporating cross-validation techniques is highly recommended. Cross-validation can be used to divide the data into multiple training and test periods, allowing the model's performance to be evaluated across different timeframes. This approach would provide a more reliable estimate of the expected mean returns by reducing overfitting to a specific period and highlighting the robustness of the strategy under diverse market conditions.

Another important recommendation is to incorporate backtesting using synthetic data to further enhance the strategy's robustness. Synthetic data can be generated to simulate different market conditions, including extreme scenarios that may not have occurred in the historical dataset. This approach allows for a more extensive evaluation of the trading strategies by testing them under a variety of conditions, thus providing insights into how well the strategies would perform in unpredictable or volatile markets. By training models on synthetic data, it is possible to uncover hidden risks and ensure that the trading strategies remain effective even in the face of sudden market changes.

Expanding the evaluation metrics to include alternative risk measures such as the Calmar ratio, Ulcer index, and tail risk metrics is also advised. These measures would offer a more comprehensive understanding of the strategies' risk-return profiles by focusing on downside risks and performance during extreme market conditions.

These recommendations aim to address the current study's limitations and guide future research towards developing more adaptive, robust, and profitable pairs trading strategies. By integrating machine learning, dynamic transaction cost models, cross-validation techniques, backtesting with synthetic data, and enhanced risk measures, future research can significantly improve the accuracy and stability of pairs trading strategies.

Bibliography

[1] Mihael Ankerst et al. "OPTICS: Ordering Points To Identify the Clustering Structure". In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data* (1999), pp. 49–60.

[2] Marco Avellaneda and Jeong-Hyun Lee. "Statistical Arbitrage in the U.S. Equities Market". In: *Quantitative Finance* 10.7 (2010), pp. 761–782.

[3] Joao Filipe Caldeira and Guilherme Victor Moura. "Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy". In: *Brazilian Review of Finance* 11.1 (2013), pp. 49–80.

[4] Ernie Chan. *Algorithmic Trading: Winning Strategies and Their Rationale*. Hoboken, NJ: Wiley, 2013.

[5] H. Chen, S. J. Chen, and F. Li. "Empirical Investigation of an Equity Pairs Trading Strategy". In: *SSRN Electronic Journal* (2012). DOI: `10.2139/ssrn.2042653`.

[6] Baozhong Do and Robert Faff. "Are Pairs Trading Profits Robust to Trading Costs?" In: *Journal of Financial Research* 35.2 (2012), pp. 261–287.

[7] Baozhong Do and Robert Faff. "Does Simple Pairs Trading Still Work?" In: *Financial Analysts Journal* 66.4 (2010), pp. 83–95.

[8] Martin Ester et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (1996), pp. 226–231.

[9] Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. *Pairs Trading: Performance of a Relative Value Arbitrage Rule*. Tech. rep. Yale ICF Working Paper No. 08-03, 2006. DOI: `10.2139/ssrn.141615`. URL: `https://ssrn.com/abstract=141615`.

[10] Nicolas Huck and Koffi Afawubo. "Pairs Trading and Selection Methods: Is Cointegration Superior?" In: *Applied Economics* 47.6 (2015), pp. 599–613.

[11] Simão Moraes Sarmento and Nuno Horta. "Enhancing a Pairs Trading Strategy with the Application of Machine Learning". In: *Expert Systems with Applications* 158 (2020), p. 113490. DOI: `10.1016/j.eswa.2020.113490`.

[12] Hudson & Thames. *The Definitive Guide to Pairs Trading*. Hudson & Thames Quantitative Research. 2023. URL: `https://hudsonthames.org`.

[13] Ganapathy Vidyamurthy. *Pairs Trading: Quantitative Methods and Analysis*. Hoboken, NJ: John Wiley & Sons, 2004.

Appendix A: Portfolio Returns

Table A.1: Distance 25 Portfolio

Pair	Cumul. Return (%)	Sharpe Ratio	Sortino Ratio	Max. Drawdown	Total Trades
IQV-LH	156.05	2.10	2.34	-0.06	37.00
GM-IP	144.38	2.06	2.03	-0.06	17.00
CL-MAS	127.48	1.98	2.22	-0.07	25.00
AEE-PAYX	183.17	1.97	2.16	-0.13	27.00
PM-SO	148.94	1.94	2.90	-0.07	25.00
AMP-MCD	130.22	1.90	2.58	-0.09	18.00
ABBV-PM	116.73	1.87	2.14	-0.09	15.00
CB-WM	108.29	1.85	2.17	-0.07	19.00
AXP-J	133.77	1.84	2.30	-0.08	20.00
CB-DUK	132.04	1.83	2.16	-0.10	18.00
AAPL-VMC	154.20	1.83	2.04	-0.07	9.00
DTE-NEE	99.57	1.79	2.01	-0.09	11.00
DUK-RTX	61.38	1.79	1.16	-0.06	12.00
DLTR-HSY	210.76	1.77	1.93	-0.14	19.00
NDSN-NEE	216.06	1.76	1.81	-0.11	16.00
AMP-HCA	197.84	1.76	2.42	-0.10	20.00
ABT-STT	160.19	1.76	2.01	-0.07	16.00
CMS-HSIC	135.20	1.75	1.94	-0.09	25.00
MDLZ-ON	174.58	1.74	1.52	-0.09	22.00
DFS-MS	168.99	1.74	2.02	-0.08	13.00
DGX-VMC	157.71	1.73	2.01	-0.11	17.00
CINF-MAR	185.43	1.73	2.12	-0.12	14.00
DUK-PAYX	131.55	1.73	2.12	-0.09	33.00
HSIC-LNT	120.04	1.72	1.74	-0.08	27.00
SYK-VRSK	137.56	1.72	1.82	-0.08	18.00
Mean	147.69	1.83	2.07	-0.09	19.72
Median	144.38	1.79	2.04	-0.09	18.00
Standard Deviation	35.43	0.11	0.33	0.02	6.69

Table A.2: K-Means 25 Portfolio

Pair	Cumul. Return (%)	Sharpe Ratio	Sortino Ratio	Max. Drawdown	Total Trades
AWK-TGT	338.69	2.43	3.13	-0.09	33.00
DVA-JPM	230.18	1.99	2.81	-0.08	12.00
AEE-DLTR	241.07	1.97	2.54	-0.08	23.00
PM-SO	148.94	1.94	2.90	-0.07	25.00
AMT-TSN	210.41	1.87	2.59	-0.10	16.00
ABBV-PM	116.73	1.87	2.14	-0.09	15.00
COST-WEC	310.49	1.83	2.31	-0.12	25.00
SHW-WTW	150.03	1.83	2.49	-0.08	23.00
EXC-KHC	137.46	1.80	2.05	-0.13	12.00
DPZ-WEC	250.41	1.80	2.40	-0.13	22.00
CMS-PM	153.78	1.80	1.93	-0.09	20.00
DTE-NEE	99.57	1.79	2.01	-0.09	11.00
AKAM-TTWO	246.66	1.78	2.60	-0.09	21.00
DLTR-HSY	210.76	1.77	1.93	-0.14	19.00
LHX-WEC	172.12	1.77	2.18	-0.09	22.00
RMD-XEL	166.82	1.76	1.59	-0.08	19.00
AMP-HCA	197.84	1.76	2.42	-0.10	20.00
GL-RJF	297.16	1.75	1.71	-0.09	13.00
CMS-EIX	124.48	1.75	2.58	-0.09	19.00
J-XYL	88.22	1.74	2.06	-0.05	13.00
TGT-XEL	177.80	1.73	1.81	-0.12	18.00
CMI-DVA	160.78	1.72	2.04	-0.09	10.00
CMS-CVS	155.32	1.72	1.94	-0.11	23.00
AEP-IBM	164.64	1.72	2.04	-0.11	24.00
CI-MSI	154.31	1.72	2.27	-0.13	12.00
Mean	188.19	1.82	2.26	-0.10	18.80
Median	166.82	1.78	2.18	-0.09	19.00
Standard Deviation	64.73	0.15	0.38	0.02	5.61

Table A.3: DBSCAN 25 Portfolio

Pair	Cumul. Return (%)	Sharpe Ratio	Sortino Ratio	Max. Drawdown	Total Trades
AWK-TGT	338.69	2.43	3.13	-0.09	33.00
AEE-DLTR	241.07	1.97	2.54	-0.08	23.00
PM-SO	148.94	1.94	2.90	-0.07	25.00
AMT-TSN	210.41	1.87	2.59	-0.10	16.00
ABBV-PM	116.73	1.87	2.14	-0.09	15.00
COST-WEC	310.49	1.83	2.31	-0.12	25.00
SHW-WTW	150.03	1.83	2.49	-0.08	23.00
EXC-KHC	137.46	1.80	2.05	-0.13	12.00
DPZ-WEC	250.41	1.80	2.40	-0.13	22.00
CMS-PM	153.78	1.80	1.93	-0.09	20.00
DTE-NEE	99.57	1.79	2.01	-0.09	11.00
AKAM-TTWO	246.66	1.78	2.60	-0.09	21.00
DLTR-HSY	210.76	1.77	1.93	-0.14	19.00
LHX-WEC	172.12	1.77	2.18	-0.09	22.00
RMD-XEL	166.82	1.76	1.59	-0.08	19.00
AMP-HCA	197.84	1.76	2.42	-0.10	20.00
GL-RJF	297.16	1.75	1.71	-0.09	13.00
CMS-EIX	124.48	1.75	2.58	-0.09	19.00
J-XYL	88.22	1.74	2.06	-0.05	13.00
TGT-XEL	177.80	1.73	1.81	-0.12	18.00
CMS-CVS	155.32	1.72	1.94	-0.11	23.00
AEP-IBM	164.64	1.72	2.04	-0.11	24.00
CI-MSI	154.31	1.72	2.27	-0.13	12.00
AWK-FFIV	177.43	1.72	2.50	-0.13	18.00
CMS-DLTR	167.76	1.71	1.95	-0.13	19.00
Mean	186.36	1.81	2.24	-0.10	19.40
Median	167.76	1.77	2.18	-0.09	19.00
Standard Deviation	64.06	0.15	0.37	0.02	5.09

Table A.4: OPTICS 25 Portfolio

Pair	Cumul. Return (%)	Sharpe Ratio	Sortino Ratio	Max. Drawdown	Total Trades
AMP-HCA	197.84	1.76	2.42	-0.10	20.00
GL-RJF	297.16	1.75	1.71	-0.09	13.00
J-XYL	88.22	1.74	2.06	-0.05	13.00
CMS-DUK	81.79	1.70	1.65	-0.05	16.00
FOX-GPC	371.16	1.68	1.65	-0.13	16.00
ADP-SNA	92.56	1.68	1.68	-0.05	14.00
GPC-TAP	138.84	1.66	2.16	-0.12	12.00
ADP-FOX	349.57	1.66	1.67	-0.14	16.00
EW-TECH	165.09	1.65	2.19	-0.07	17.00
GL-PWR	430.49	1.64	1.67	-0.13	13.00
FOX-PPL	129.05	1.64	1.95	-0.09	21.00
BBY-NKE	239.08	1.64	1.84	-0.12	17.00
INTU-LRCX	120.33	1.63	1.28	-0.10	12.00
AEP-CMS	51.02	1.63	1.96	-0.03	22.00
JBHT-LOW	140.42	1.62	1.56	-0.08	29.00
ALLE-BRK-B	294.80	1.62	1.95	-0.14	15.00
DOV-J	72.65	1.62	1.76	-0.04	13.00
GL-KKR	324.05	1.61	2.17	-0.12	13.00
QRVO-SWKS	82.15	1.61	1.55	-0.08	18.00
J-SNA	107.73	1.61	1.88	-0.06	14.00
EFX-GS	128.97	1.61	1.65	-0.07	12.00
FOX-MAS	251.60	1.59	1.56	-0.11	14.00
CDW-TRV	92.34	1.58	1.77	-0.08	14.00
PH-XYL	85.92	1.58	1.54	-0.08	8.00
AEE-CMS	51.61	1.58	2.32	-0.05	23.00
Mean	175.38	1.64	1.82	-0.09	15.80
Median	129.05	1.63	1.76	-0.08	14.00
Standard Deviation	111.44	0.05	0.28	0.03	4.44

Table A.5: Combined 25 Portfolio

Pair	Cumul. Return (%)	Sharpe Ratio	Sortino Ratio	Max. Drawdown	Total Trades
AWK-TGT	338.69	2.43	3.13	-0.09	33.00
IQV-LH	156.05	2.10	2.34	-0.06	37.00
GM-IP	144.38	2.06	2.03	-0.06	17.00
DVA-JPM	230.18	1.99	2.81	-0.08	12.00
CL-MAS	127.48	1.98	2.22	-0.07	25.00
AEE-DLTR	241.07	1.97	2.54	-0.08	23.00
AEE-PAYX	183.17	1.97	2.16	-0.13	27.00
PM-SO	148.94	1.94	2.90	-0.07	25.00
AMP-MCD	130.22	1.90	2.58	-0.09	18.00
AMT-TSN	210.41	1.87	2.59	-0.10	16.00
ABBV-PM	116.73	1.87	2.14	-0.09	15.00
CB-WM	108.29	1.85	2.17	-0.07	19.00
AXP-J	133.77	1.84	2.30	-0.08	20.00
CB-DUK	132.04	1.83	2.16	-0.10	18.00
COST-WEC	310.49	1.83	2.31	-0.12	25.00
AAPL-VMC	154.20	1.83	2.04	-0.07	9.00
SHW-WTW	150.03	1.83	2.49	-0.08	23.00
EXC-KHC	137.46	1.80	2.05	-0.13	12.00
DPZ-WEC	250.41	1.80	2.40	-0.13	22.00
CMS-PM	153.78	1.80	1.93	-0.09	20.00
DTE-NEE	99.57	1.79	2.01	-0.09	11.00
DUK-RTX	61.38	1.79	1.16	-0.06	12.00
AKAM-TTWO	246.66	1.78	2.60	-0.09	21.00
DLTR-HSY	210.76	1.77	1.93	-0.14	19.00
LHX-WEC	172.12	1.77	2.18	-0.09	22.00
Mean	173.93	1.90	2.29	-0.09	20.04
Median	153.78	1.84	2.22	-0.09	20.00
Standard Deviation	66.18	0.15	0.39	0.02	6.66