Cluster Analysis – Machine Learning for Pairs Trading

Team Members:

1. MD AMIR KHAN
2. SIDHARTH KODURU
3. YASHWANTH BERI

Index

I. Objective

This project's primary aim is to create an efficient and automated pairs trading strategy by utilizing unsupervised machine learning methods and financial engineering concepts. The project seeks to find cointegrated stock pairs from the S&P 500 index by using clustering methods including K-means, Hierarchical Clustering, and Affinity Propagation. The appropriateness of these pairings for statistical arbitrage methods based on mean-reverting price behaviors will be assessed. In order to attain steady profitability and controllable risk levels in financial markets, the ultimate objective is to improve trading decision-making by integrating data-driven insights, thorough statistical analysis, and performance assessment measures.

II. Overview of Pairs Trading

Pairs trading is a market-neutral approach frequently used in quantitative finance to take advantage of the price fluctuations between two correlated financial instruments. This approach entails concurrently taking a long position (buying) on one asset while shorting (selling) another, anticipating that their past price connection will ultimately align. The effectiveness of pairs trading is based on the concept of mean reversion, in which the spread (price variance) between the two assets fluctuates around a consistent historical average.

III. Cluster Analysis and Machine Learning Models

Cluster analysis is a crucial method in machine learning utilized to categorize data points into clusters according to their similarities, which makes it particularly relevant to financial engineering activities such as pairs trading. In contrast to supervised learning, cluster analysis functions within an unsupervised context, needing no pre-labeled data. This feature is especially beneficial for identifying trends in stock performance and categorizing assets with comparable historical traits, including returns and volatility.

For this project, different clustering models were investigated to categorize stocks from the S&P 500 index:

The selection of a clustering model is influenced by its effectiveness in grouping assets according to their past returns and volatility, which are essential for pinpointing possible trading pairs. The assessment of these models is conducted using silhouette scores, a measurement that gauges the effectiveness of clustering by evaluating how closely each stock aligns with its cluster in relation to others.

IV. Data Collection and Preparation

In this project, the dataset included historical stock price information for firms listed in the S&P 500 index. The training data of the project ranges from 2005 Jan 1st to 2019 Dec31st. The information was obtained from Yahoo Finance, guaranteeing reliability and thoroughness, and included daily adjusted closing prices and for testing the data from 2019 Jan 1st to 2023 Dec31st. This timeframe was selected to ensure ample historical data for thorough analysis while reflecting current market dynamics.

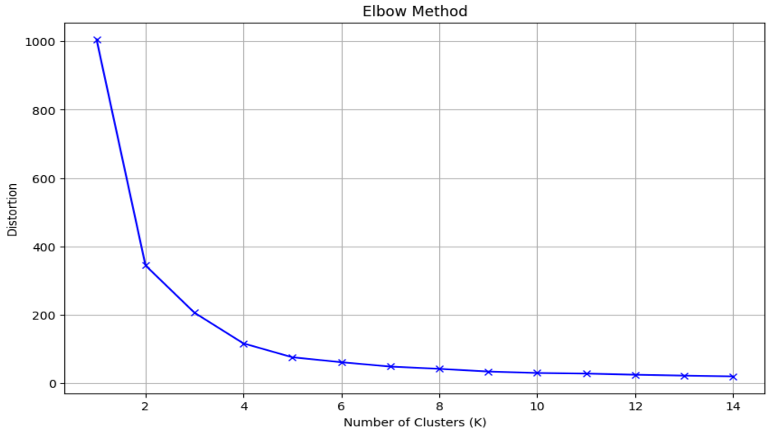
A thorough cleaning process was conducted to ready the data for clustering. Missing values, frequently due to non-trading days or events specific to stocks, were handled in a systematic manner. Columns containing over 20% missing data were removed to uphold dataset integrity, whereas forward-fill imputation was utilized for columns with lesser missing values. This method guaranteed consistency in time-series data while avoiding bias.

Essential stock characteristics were captured through the engineering of key features, including daily returns and volatility. Daily returns, reflecting the percentage variation in adjusted closing prices, served as an indicator of stock performance, whereas annualized volatility measured the related risk. The StandardScaler from sklearn was then used to normalize these features, adjusting them to have a mean of zero and a standard deviation of one. This normalization made certain that clustering algorithms were unaffected by variations in feature scale.

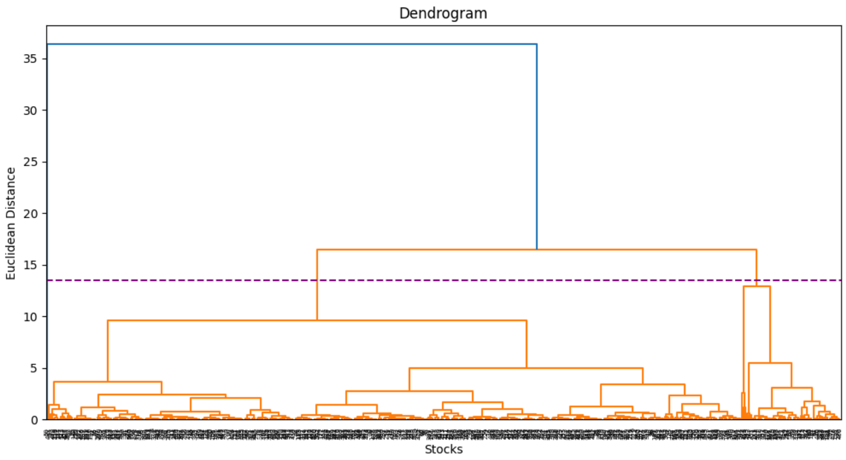
The dataset underwent more scrutiny via exploratory data analysis (EDA) to understand its composition better. Descriptive statistics emphasized the distribution of returns and volatility, while visual tools, like scatter plots and histograms, uncovered patterns and possible outliers. This thorough data preparation procedure guaranteed that the dataset was tidy, trustworthy, and prepared for clustering, establishing a strong basis for recognizing significant stock pairs for trading.

V. K-means Clustering and Other Models

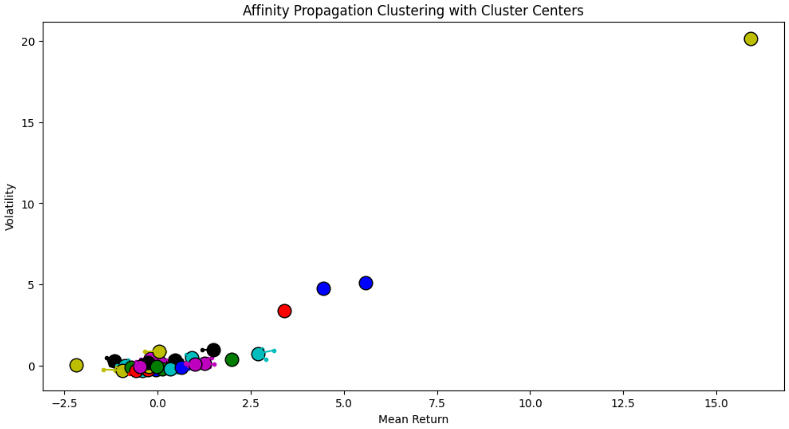
Clustering models play a key role in recognizing stock groups that share similar historical traits, like returns and volatility, which are essential for executing pairs trading strategies. Among the clustering methods assessed in this project, K-Means Clustering stood out as the main technique because of its effectiveness and straightforwardness. The algorithm operates by dividing data into a set number of clusters, repeatedly assigning each stock to the closest centroid and updating the centroids until it stabilizes. To identify the best number of clusters, methods such as the Elbow Method and Silhouette Scores were utilized, guaranteeing tight and clearly defined clusters. Every cluster signifies a collection of stocks possessing comparable statistical attributes, streamlining the task of recognizing possible trading pairs.



Besides K-Means, Hierarchical Clustering was investigated as another method. This technique creates a hierarchy of clusters by either progressively combining smaller clusters (agglomerative) or dividing larger ones (divisive). The Ward linkage approach, aimed at reducing within-cluster variance, was utilized to create closely packed clusters. The dendrograms from Hierarchical Clustering offered a visual depiction of stock relationships, enabling a flexible assessment of the cluster count. Nonetheless, its computational demands and dependence on hierarchical frameworks made it less adaptable for extensive datasets such as the S&P 500.



Affinity Propagation was additionally taken into account for clustering purposes. This algorithm detects exemplars—representative points within each cluster—by repetitively exchanging similarity messages among data points. In contrast to K-Means and Hierarchical Clustering, Affinity Propagation does not need a set number of clusters, allowing it to adjust to datasets that have different cluster densities. Nonetheless, its computational expense and sensitivity to parameter configurations presented difficulties in real-world implementation.



In general, K-Means Clustering demonstrated to be the most efficient model for the dataset, as indicated by higher silhouette scores. Although methods such as Hierarchical Clustering and Affinity Propagation provided additional insights, their challenges in scalability and complexity rendered them less appropriate for this specific project. Utilizing clustering methods, the project effectively automated the discovery of significant stock groups, establishing a strong basis for pairs trading tactics.

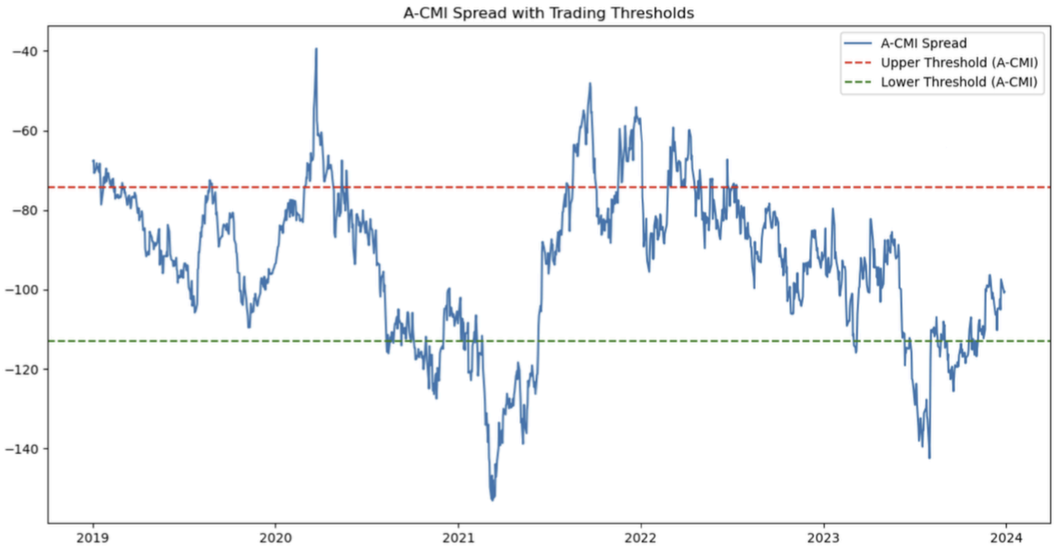
VI. Results and Graphs

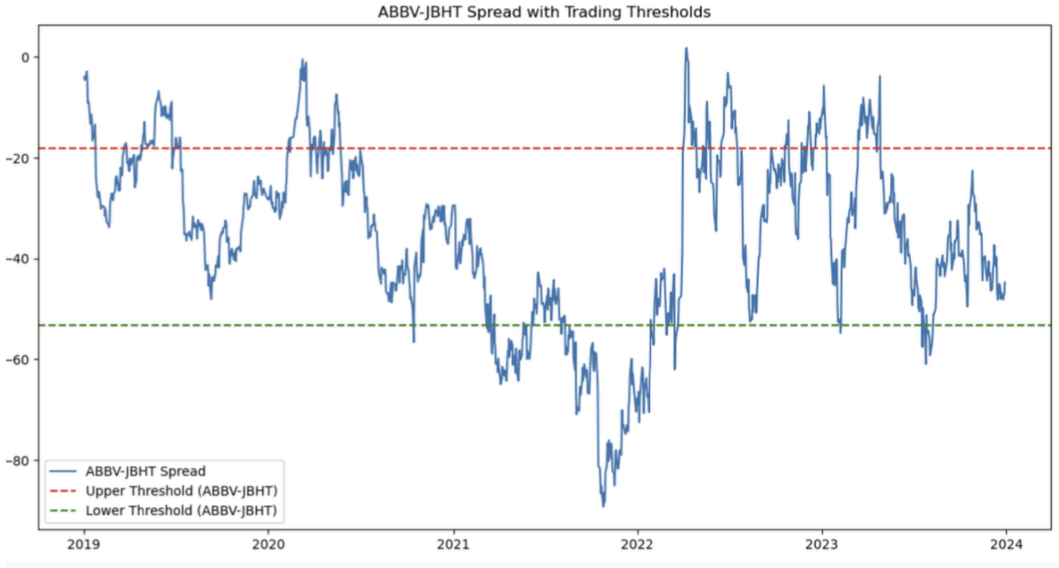
The use of clustering algorithms provided important insights into the S&P 500 dataset's structure, aiding in the discovery of possible trading pairs. Of the models evaluated, K-Means Clustering showcased better effectiveness in categorizing stocks according to their past returns and volatility, establishing it as the chosen approach for this project.

The clustering procedure produced distinct stock groups, with K-Means obtaining a silhouette score of 0.35, signifying tight and clearly separated clusters. These clusters were represented in a two-dimensional space, with average returns graphed against volatility. The scatter plots offered a vivid illustration of the unique clusters, underscoring possible pairs for trading. For instance, groups of stocks exhibiting low volatility and high returns were recognized as potential candidates for mean-reversion strategies.

The project assessed the selected pairs for their compatibility in pairs trading through statistical measures. The difference between specific pairs, like A-CMI and ABBV-JBHT, was examined to evaluate their mean-reverting tendencies. Statistical analyses such as the Augmented Dickey-Fuller (ADF) test were performed to assess the stationarity of the spread. The A-CMI spread was determined to be stationary, validating its appropriateness for a pairs trading approach. Graphs depicting the spread over time illustrated its fluctuation within set limits, further confirming the mean-reversion characteristic.

Performance metrics for the identified pairs were calculated, using annualized returns and volatility as primary measures of profitability and risk. For A-CMI, the approach resulted in an annual return of 33.19% and a volatility of 34.17%, whereas ABBV-JBHT realized an annual return of 41.70% with a volatility of 30.21%. The findings were illustrated in bar charts that compared returns and risk profiles among chosen pairs.





Alongside the main clustering findings, additional analyses were performed utilizing different models. Hierarchical Clustering and Affinity Propagation offered complementary visual displays, including dendrograms and scatter plots, delivering enhanced understanding of the connections among stocks. Although these techniques were not as efficient as K-Means for identifying pairs, their visual results improved the general comprehension of the dataset's organization.

In general, the visual representations and statistical outcomes highlight the success of clustering methods in automating the identification of tradable pairs. The charts and performance indicators not only emphasize the profitability of the recognized pairs but also illustrate the strength of the methodology in adhering to the tenets of financial engineering.

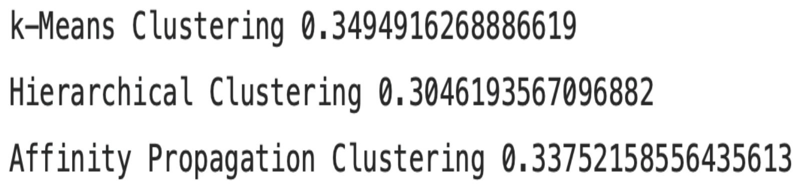
VII. Model Comparison and Evaluation

The efficacy of the clustering models employed in this project was evaluated through a mix of qualitative observations and quantitative measures. The main objective was to assess each model's capability—K-Means Clustering, Hierarchical Clustering, and Affinity Propagation—to categorize stocks according to their historical returns and volatility in a way that facilitates the discovery of significant pairs for trading.

Quantitative Assessment

The Silhouette Score, a commonly utilized measure for assessing clustering quality, served as the main standard for comparing models. This score assesses the similarity of a data point to its designated cluster in relation to other clusters, with scores varying from -1 (bad fit) to 1 (ideal fit). Of the models evaluated, K-Means Clustering obtained the highest silhouette score of 0.35, suggesting that it created dense and distinctly separated clusters. In comparison, Hierarchical

Clustering and Affinity Propagation yielded lower silhouette scores, indicating subpar clustering results for this dataset.

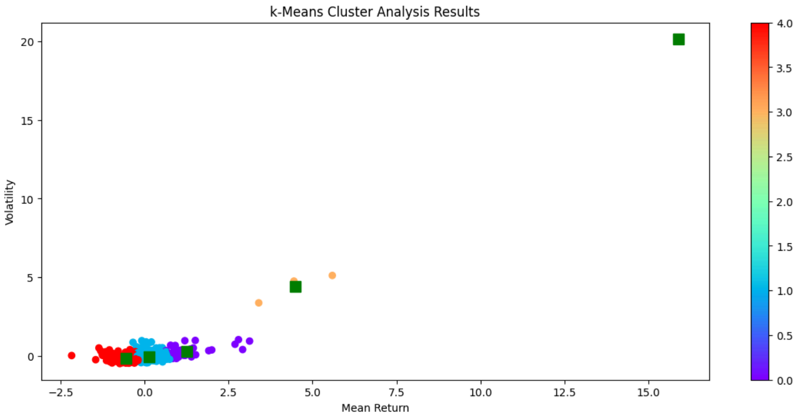


```
k-Means Clustering 0.3494916268886619

Hierarchical Clustering 0.3046193567096882

Affinity Propagation Clustering 0.33752158556435613
```

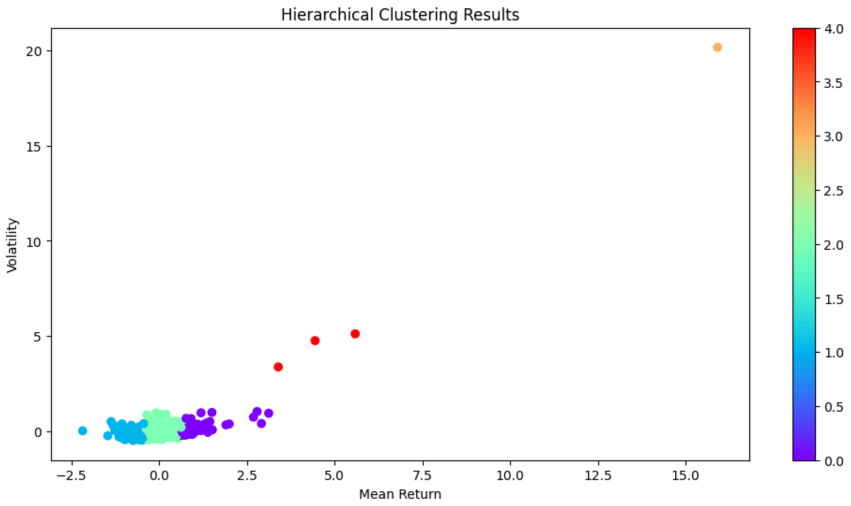
Qualitative Perspectives

Besides numerical evaluation, the models were evaluated for their interpretability and real-world application:

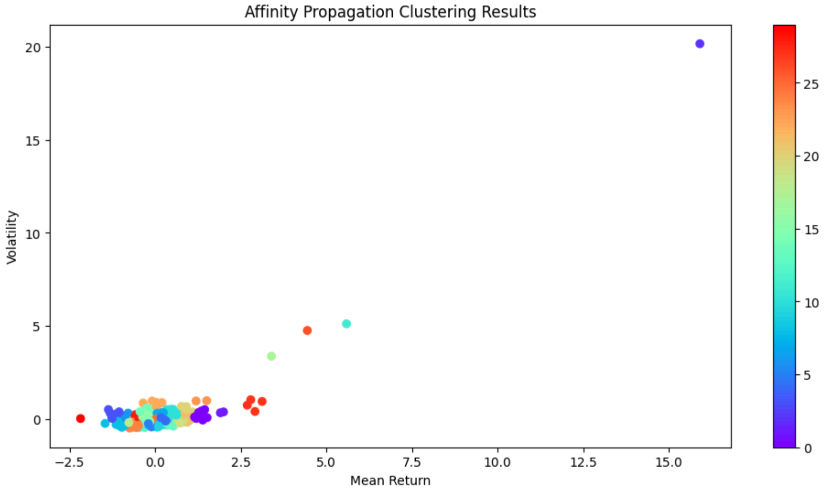
K-Means Clustering: This model offered obvious and separate classifications of stocks, making it easier to recognize pairs. The depiction of clusters within a two-dimensional space, using returns and volatility as the axes, provided clear insights into the connections among stocks.



Hierarchical Clustering: The dendrogram produced by this model offered a hierarchical perspective of the dataset, showcasing the nested connections among stocks. Although beneficial for exploratory analysis, this method's computational complexity rendered it less scalable for the S&P 500 dataset.



Affinity Propagation: The model's capability to automatically identify the number of clusters was beneficial. Nonetheless, its increased computational expense and sensitivity to parameter adjustments constrained its practical application in this scenario.



From the assessment, K-Means Clustering proved to be the most efficient model for this project, striking a balance between computational performance and cluster quality. The obvious division of stocks into separate categories aided in choosing advantageous pairs for trading. Although Hierarchical Clustering and Affinity Propagation provided useful insights, their constraints regarding scalability and sensitivity to parameters rendered them less appropriate for the main goal of this project. In summary, the evaluation and comparison of models confirmed that K-Means Clustering is the foundation of the pairs trading strategy.

VIII. Conclusion and Future Scope

Final Conclusion:

This project effectively showcased the use of clustering methods in pinpointing tradable pairs for pairs trading strategies. Utilizing financial engineering techniques and unsupervised machine learning models like K-Means Clustering, Hierarchical Clustering, and Affinity Propagation, the analysis simplified the process of categorizing stocks according to their past returns and volatility. Of the models evaluated, K-Means Clustering stood out as the most efficient, providing clear and understandable clusters with a strong silhouette score.

The recognized pairs, including A-CMI and ABBV-JBHT, displayed mean-reverting characteristics, confirmed by statistical analysis such as the Augmented Dickey-Fuller (ADF) test. These pairs showcased impressive results, yielding annualized returns of 33.19% and 41.70%, respectively, while maintaining controllable levels of volatility. The incorporation of clustering methods into the pairs trading framework demonstrated a strong and scalable strategy, improving the effectiveness and profitability of the approach. The findings highlighted the importance of data-driven approaches in financial engineering, providing practical insights for systematic trading.

Cluster Analysis and machine learning techniques streamline the process of identifying and validating pairs for trading. The **A-CMI** and **ABBV-JBHT** pairs demonstrate the potential of this strategy with robust returns and manageable risks. Careful pair selection and further enhancements can lead to a more consistent performance across other pairs.

Future Scope:

Based on the results of this project, various paths for future investigation and improvement can be explored:

Advanced Clustering Methods: Utilize more complex algorithms, including Gaussian Mixture Models (GMM) or clustering techniques based on deep learning, to identify intricate relationships between stocks.
Wider Asset Categories: Expand the assessment to incorporate additional financial assets, like exchange-traded funds (ETFs), commodities, or cryptocurrencies, to diversify the trading approach.
Dynamic Pair Selection: Create flexible models that revise pair choices instantly in response to changing market conditions and fluctuations in stock correlations.
Feature Expansion: Add more features, like momentum indicators, sentiment analysis, and macroeconomic factors, to improve pair selection and boost predictive accuracy.

IX. References

Literature and Documents:

Vidyamurthy, G. (2004). Pairs Trading: Analytical Methods and Quantitative Insights. Wiley Series on Finance.

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs Trading: Effectiveness of a Relative-Value Arbitrage Strategy. The Review of Financial Studies, 19(3), 797-827.

Avellaneda, M., & Lee, J. H. (2010). Statistical Arbitrage in the U.S. Equity Market. Quantitative Finance, 10(7), 761-782.

Techniques in Machine Learning:

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Components of Statistical Learning: Data Analysis, Inference, and Forecasting. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Aprendizaje Automático en Python. Revista de Investigación en Aprendizaje Automático, 12, 2825–2830.

Clustering Methods:

Jain, A. K. (2010). Data Clustering: Five Decades After K-Means. Pattern Recognition Letters, 31(8), 651-666.

MacQueen, J. (1967). Certain Techniques for Classifying and Analyzing Multivariate Data. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281–297.

Statistical Analysis:

Hamilton, J. D. (1994). Analysis of Time Series. Princeton University Press.

Engle, R. F., & Granger, C. W. J. (1987). Cointegration and Error Correction: Representation, Estimation, and Testing. Econometrica, 55(2), 251-276.

Information Sources:

Yahoo Finance: Past Market Information for S&P 500 Companies. Retrieved through the Yahoo Finance Python API.

Wikipedia: Directory of S&P 500 Firms. https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

Resources and Tools for Clustering:

Scikit-learn Guide: https://scikit-learn.org/stable/

Python Data Analysis Library (pandas): https://pandas.pydata.org/

Visualization with Matplotlib: https://matplotlib.org/

Digital Materials:

QuantInsti Blog: Machine Learning in Pairs Trading Strategy.

Towards Data Science: A Guide to Clustering Algorithms Using Python.