Application of Unsupervised Machine Learning for Quantitative Trading

ACADEMIC YEAR:

2023/2024, Semester 1

SECTION:

QF209 – Machine Learning in Quantitative Finance

G1 – Team 7

NAME	STUDENT ID
Jerome Wong Jen Hoe	01416326
Lim Wei Jie	01371965
Htoo Myat Naing	01395605
Liu Sirui	01394015
Nguyen Ngoc Quynh Tram	01495579

1. Contents

2. Executive Summary *2*

3. Introduction *3*

a. **Problem Statement** 3

b. **Overview of Unsupervised Learning** 3

c. **Overview of Clustering in Trading** 3

d. **Overview of Trading Strategy** 3

e. **Step-by-Step Implementation of Trading Strategy** 4

f. **Success Metrics of Trading Strategies** 4

4. Dataset Collection and Preprocessing *4*

5. Machine Learning Models: K-Means and Agglomerative Clustering *6*

a. **Overview of K-Means** 6

Why is K-Means Clustering used? 7

b. **Overview of Agglomerative Clustering** 7

How the algorithm works 7

Why is it used and why is it different than K-Means Clustering? 8

6. Results: *9*

a. **K-Means Clustering Result** 9

b. **Agglomerative Clustering Result** 11

c. **Comparing Results of 2 Clustering Methods** 15

7. Recommendation for improvements *16*

Further Investigation of Agglomerative Clustering Results 16

Alternative Clustering Algorithms 17

8. Conclusion *18*

9. References *20*

2. Executive Summary

The financial industry continually seeks innovative strategies to enhance portfolio management and investment returns. This report explores the integration of unsupervised machine learning techniques—specifically K-Means and Agglomerative Clustering—into the development of trading strategies. By identifying coherent clusters of financial assets, these techniques have shown promise in improving risk-adjusted returns, outperforming traditional benchmarks such as the SPY Buy & Hold strategy.

Our analysis demonstrates that clustering-based strategies, particularly employing K-Means Clustering with k = 6 and Agglomerative Clustering with k=3, offered a superior balance of risk and return, as evidenced by their Sharpe ratios. Both methods yielded impressive cumulative returns over the last 5 years (2018 – 2023) compared to the S&P500 benchmark. However, these strategies also present a higher maximum drawdown during the same period, indicating a potential increase in risk exposure that came with higher returns. This trade-off is a critical consideration for portfolio managers aiming to optimize returns while managing risk levels.

The report underscores the dynamic nature of financial markets, where asset behaviour is influenced by market conditions and external events, such as economic disruptions and geopolitical tensions. It suggests that an adaptive strategy, which includes regular re-evaluation of cluster configurations and reallocation of assets, can better respond to such fluctuations.

In conclusion, the application of Unsupervised Learning, particularly K-Means and Agglomerative Clustering presents a significant advancement in financial analytics, offering a data-driven approach to portfolio management. However, investors may need to consider the risk-return trade-offs when choosing these strategies. Further experiments in other Clustering techniques can lead to better understanding of the underlying factors driving the markets, more informed decision-making and improved investment outcomes.

3. Introduction

a. Problem Statement

The realm of finance and trading has witnessed an increasing integration of machine learning techniques to gain insights, manage risk, and optimize investment strategies. In this context, a fundamental challenge exists - the identification of coherent groups or clusters of financial assets that exhibit similar characteristics. These clusters can guide portfolio diversification and risk management. However, conventional approaches often fall short when dealing with the complex and high-dimensional data found in the financial markets. This academic report addresses the problem of cluster-based trading strategy formulation in financial markets by utilizing unsupervised machine learning techniques, specifically K-Means and Agglomerative Clustering, to create clusters of assets, followed by the construction of a trading strategy based on these clusters. The primary objective is to enhance the performance and risk-adjusted returns of the trading strategy when compared to traditional investment approaches.

b. Overview of Unsupervised Learning

Unsupervised machine learning represents a branch of artificial intelligence dedicated to the extraction of latent patterns and structures from data, operating without the reliance on pre-assigned labels or categorical annotations. Diverging from the conventional paradigm of supervised learning, where algorithms are trained on labelled datasets to make predictions or classifications, unsupervised learning engages in an exploratory process, ideally suited for circumstances where the inherent relationships within data remain ambiguously defined. It emerges as a pivotal instrument for uncovering concealed insights and underlying information across a spectrum of domains, including the intricate landscape of finance and trading.

c. Overview of Clustering in Trading

Within the domain of trading, unsupervised learning, notably clustering, assumes a prominent role. Clustering methodologies serve as a method to categorize financial assets, such as stocks, bonds, or commodities, into groups or clusters, contingent on shared underlying similarities. These resemblances may encompass historical price dynamics, trading volumes, correlations with other assets, or other pertinent attributes. Clustering furnishes traders with the ability to discern coherent clusters of assets exhibiting akin behaviours, a facet instrumental for portfolio diversification, risk mitigation, and the development of astute trading strategies. By incorporating clustering techniques in trading, market participants can elevate their decision-making acumen, enhance investment strategies, and navigate the intricate contours of financial markets with a heightened level of precision and sophistication.

d. Overview of Trading Strategy

This study introduces a novel trading strategy that leverages unsupervised machine learning, specifically K-Means and Agglomerative Clustering, to group financial assets based on similarities in their historical price dynamics, trading volumes, and other relevant features. The clustering process identifies clusters of assets that exhibit akin behaviours, providing a foundation for the trading strategy. By categorizing assets into clusters, traders can make more informed decisions regarding portfolio diversification and risk mitigation. Moreover, the strategy capitalizes on the insights gained from unsupervised learning to allocate capital to assets that have higher expected returns while managing volatility. The approach aims to enhance trading strategy performance by integrating state-of-the-art machine learning techniques.

e. Step-by-Step Implementation of Trading Strategy

1. **Initialization**: The process begins with the initialization of K-Means and Agglomerative Clustering algorithms. K-Means randomly or strategically selects K initial cluster centroids, whereas Agglomerative Clustering starts with individual data points, each in its own cluster.
2. **Similarity Computation (Agglomerative Clustering)**: In Agglomerative Clustering, a proximity matrix is computed to measure the distance between each pair of clusters, using various distance metrics such as Euclidean, Manhattan, or Cosine distance.
3. **Assignment (K-Means)**: For K-Means, the algorithm assigns each data point to the cluster whose centroid is closest to it based on a chosen distance metric, such as Euclidean distance.
4. **Cluster Merging (Agglomerative Clustering)**: In Agglomerative Clustering, the closest pair of clusters are iteratively merged into larger clusters.
5. **Update**: For both K-Means and Agglomerative Clustering, centroids are recalculated to reflect the centre of their respective clusters.
6. **Iteration**: Steps 3-5 are iteratively repeated until convergence is achieved, ensuring that data points are assigned to the best-fit clusters.
7. **Final Clustering**: Once the algorithms converge, data points are assigned to their final clusters.
8. **Optimizing Portfolio**: The selected cluster(s) determine which stocks to trade each month, with stock weights adjusted to optimize the Sharpe ratio, reflecting risk-adjusted returns.

f. Success Metrics of Trading Strategies

To assess the success of the trading strategy, a set of performance metrics are employed. These metrics include ***annualized volatility, annualized return, annualized Sharpe ratio, Sortino ratio, Calmar ratio and maximum drawdown***. Annualized volatility measures the variation in the portfolio's returns over time, annualized return calculates the expected returns of the portfolio, and the Sharpe ratio quantifies the risk-adjusted returns. Sortino ratio, like the Sharpe ratio, measures the risk-adjusted returns, but penalises only those returns falling below a user-specified target. Calmar ratio is a function of the average compounded annual rate of return versus its maximum drawdown. The maximum drawdown indicates the maximum loss experienced by the portfolio during its trading history. The ratios are visualised through a heatmap. By analysing these metrics, the efficacy of the trading strategy is evaluated in terms of risk, return, and overall performance. This comprehensive assessment ensures that the strategy's benefits are rigorously analysed and benchmarked against traditional investment benchmarks, such as the S&P 500, enabling a sound judgment of its success.

4. Dataset Collection and Preprocessing

We collect a list of S&P 500 companies from Wikipedia and download the historical stock data for these companies using Yahoo Finance. Then we used the open, high, low, close, adjusted close, and volume data for each stock to calculate various features and technical indicators for each stock, such as Garman-Klass Volatility, Relative Strength Index (RSI), Bollinger Bands, Average True Range (ATR), Moving Average Convergence Divergence (MACD), and Dollar Volume.

Garman Klass Volatility is calculated using the following formula:

$$\text{Garman-Klass Volatility} = \frac{(\ln(\text{High}) - \ln(\text{Low}))^2}{2} - (2\ln(2) - 1)(\ln(\text{Adj Close}) - \ln(\text{Open}))^2$$

The remaining technical indicators are calculated using pandas-ta, a Python 3 pandas extension with 130+ technical analysis indicators. We then converted the data from daily to monthly data and filtered the top 150 most liquid stocks. Then we calculated monthly returns for different time periods (1, 2, 3, 6, 9, and 12 months) for each stock. We then introduced the Fama-French 5 factors (2x3) model (Kenneth R. French - Description of Fama/French Factors, 2023) to explain asset pricing and assess the risk and expected return on equity, and included these factors as financial features for our clustering models. The 5 factors include market risk, size, value, operating profitability, and investment (S.M. Ikhtiar Alam, 2022). We accessed the factor returns using the pandas-datareader and computed rolling factor betas for each stock using a rolling Ordinary Least Squares (OLS) approach with a maximum window of 24 periods. Finally, we join the betas with the original data, shifting them by one period. We now have a total of 18 features in the final dataset, contributing to the clustering models. The 18 features are as mentioned below along with their definition.

1. **Garman-Klass Volatility:** A measure of volatility that uses the high, low, and closing prices of an asset. It's considered more accurate than other models because it accounts for the intraday price range. It's included in the dataset as volatility is a key measure of risk and can influence investment decisions. This indicator is read as a percentage and represents the annualized volatility based on intraday price ranges.
2. **RSI (Relative Strength Index):** A momentum oscillator that measures the speed and change of price movements. RSI values range from 0 to 100 and can indicate overbought or oversold conditions. It's a popular technical indicator used to predict potential reversals in price. A value above 70 typically indicates that an asset may be overbought, while a value below 30 may indicate it is oversold.
3. **Bollinger Band Low:** The lower band of the Bollinger Bands, which is typically two standard deviations below the simple moving average (set default as 20-period or 20-day in our case). It's used to measure the low-price level relative to recent volatility and can signal oversold conditions. Prices touching or falling below this band may suggest an oversold condition. It's included in the dataset to signal potential buying opportunities or to warn of a possible increase in volatility.
4. **Bollinger Band Mid:** The middle band of the Bollinger Bands, which is the simple moving average (set default as 20-period or 20-day in our case). It serves as a reference point for the low and high bands and is used to gauge the intermediate-term trend based on the asset's price in relation to it.
5. **Bollinger Band High:** The upper band of the Bollinger Bands, which is typically two standard deviations above the simple moving average (set default as 20-period or 20-day in our case).. It's used to measure the high price level relative to recent volatility and can signal overbought conditions. Prices touching or exceeding this band may suggest an overbought condition. It's included to signal potential selling opportunities or to indicate a heightened possibility of volatility or price retracement.
6. **Average True Range Standardised:** A normalized measure of the true range of price movement over a given period. It's normalized by dividing by the price to allow comparison across different price levels. It's included as it provides insight into the volatility and potential price movement range of an asset.
7. **MACD (Moving Average Convergence Divergence) Standardised:** When the MACD line crosses above the signal line, it's a bullish signal, and when it crosses below, it's a bearish signal. The normalization allows for comparison across different trading instruments. A trend-following momentum indicator that shows the relationship between two moving averages of an asset's price. The normalization helps to compare it across different assets or time frames.
8. **1 Month Return:** The percentage change in price over the past month. It's a short-term performance metric that can indicate recent trends or momentum.

9. **2 Month Return:** The percentage change in price over the past two months. It provides a slightly longer view than the 1-month return to assess performance consistency over time.
10. **3 Month Return:** The percentage change in price over the past three months. This quarterly performance metric is commonly used to assess short-term investment performance.
11. **6 Month Return:** The percentage change in price over the past six months. It reflects medium-term performance and can be indicative of sustained trends.
12. **9 Month Return:** The percentage change in price over the past nine months. This feature can capture the performance across three quarters, providing insight into longer-term trends.
13. **12 Month Return:** The percentage change in price over the past year. Annual return is a standard measure for comparing performance year-over-year.
14. **Mkt_RF (Market Risk Premium):** The excess return of investing in the stock market over a risk-free rate. It's a measure of the market's risk premium and is crucial for understanding the reward for bearing market risk.
15. **SMB (Small Minus Big):** A factor that measures the average return on the portfolios of small stocks minus the average return on the portfolios of big stocks. This factor's value indicates whether small-cap stocks are expected to perform better than large-cap stocks. It's included to exploit size-based return anomalies in the market.
16. **HML (High Minus Low):** Another factor from the Fama-French model, HML represents the difference in returns between stocks with high book-to-market ratios and those with low ratios, capturing the "value" premium. A positive value suggests that value stocks (high book-to-market ratio) are in favour compared to growth stocks (low book-to-market ratio). It's included to capture the value premium in asset pricing.
17. **RMW (Robust Minus Weak):** A factor that captures the performance difference between companies with strong profitability and those with weak profitability. It's part of the Fama-French five-factor model and is included to account for profitability's impact on returns. A positive value indicates that stocks of companies with robust profitability are performing better than those with weaker profitability. It's included to take advantage of profitability-based return differentials.
18. **CMA (Conservative Minus Aggressive):** The final factor in the Fama-French five-factor model, CMA represents the difference in returns between conservative investment firms and aggressive investment firms, capturing the "investment" style premium. A positive value suggests that conservative firms are outperforming aggressive firms. It's included to capture the effects of investment style on returns.

5. Machine Learning Models: K-Means and Agglomerative Clustering

a. Overview of K-Means

K-Means Clustering is a popular unsupervised machine learning algorithm widely used in quantitative finance and various other fields for data analysis and pattern recognition. It is particularly useful for segmenting data into distinct groups or clusters based on similarity patterns, with the primary goal of minimizing the within-cluster variance, which is the sum of squared distances between data points and their respective cluster centroids. This optimization objective is achieved by iteratively reassigning data points to clusters and updating centroids until convergence. The result is a partitioning of the data into K distinct clusters, where K is a user-defined parameter representing the number of clusters created. Each cluster contains its unique set of data points, which exhibit greater similarity to each other compared to data points in different clusters.

The algorithm can be summarized in several key steps:

1. **Initialization:** K initial cluster centroids are randomly or strategically chosen from the data points. These centroids serve as the starting points for cluster formation.
2. **Assignment:** For each data point in the dataset, the algorithm calculates its similarity or distance to each of the K centroids. The data point is then assigned to the cluster whose centroid is closest in terms of distance, typically using a distance metric like Euclidean distance.
3. **Update:** After all data points have been assigned to clusters, the algorithm recalculates the centroids for each cluster by taking the mean of all the data points within that cluster. These new centroids represent the "center" of each cluster.
4. **Iteration:** Steps 2 and 3 are repeated iteratively until convergence, typically defined by either a maximum number of iterations or until the centroids no longer change significantly.
5. **Final Clustering:** Once the algorithm converges, the data points are assigned to their final clusters based on the last set of centroids.

Why is K-Means Clustering used?

K-Means Clustering is used in quantitative finance because of its simplicity and ease of implementation. Its intuitive approach of partitioning data into clusters based on similarity appeals to financial analysts and researchers, as it requires minimal prior knowledge and can quickly yield meaningful insights from large and complex datasets. This simplicity makes it an ideal baseline model for understanding financial data patterns as it provides a solid foundation for more advanced clustering algorithms. Once the data has been segmented using K-Means, more sophisticated methods can be applied to refine the clusters or capture non-linear relationships. K-Means, therefore, acts as a stepping stone for developing more tailored and accurate unsupervised machine learning models suited to specific financial applications.

Also, K-Means Clustering is computationally efficient and scales well to large datasets, which is crucial in quantitative finance, where data can be vast and high-dimensional. Its computational efficiency allows for rapid experimentation and prototyping of models, enabling analysts to quickly assess the feasibility of a particular approach before delving into more complex and computationally intensive techniques.

b. Overview of Agglomerative Clustering

Agglomerative Clustering is a hierarchical clustering method that builds a multilevel hierarchy of clusters by starting with individual data points and iteratively merging them into larger clusters. It is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

How the algorithm works

The algorithm works as follows:

1. **Initialisation:** Begin with n clusters, where n is the number of data points.
2. **Similarity Computation:** Compute a proximity matrix that measures the distance between each pair of clusters. Various distance metrics can be used, such as Euclidean, Manhattan, or Cosine distance.
3. **Cluster Merging:** Find the closest (most similar) pair of clusters and merge them into a single cluster.
4. **Proximity Matrix Update:** Update the proximity matrix to reflect the distance between the new cluster and the original clusters.
5. **Iteration:** Repeat steps 3 and 4 until all data points are clustered into a single cluster of size n.

The result of agglomerative clustering can be visualised using a dendrogram, which is a tree-like diagram that records the sequences of merges and shows the arrangement of the clusters produced by the algorithm.

Why is it used and why is it different than K-Means Clustering?

In the context of stock market data, where the relationships between different stocks can be complex and not well-defined, agglomerative clustering can be advantageous for several reasons:

1. Flexibility in Cluster Shapes: Unlike K-means, which assumes spherical clusters, agglomerative clustering can find clusters of various shapes and sizes.
2. Interpretability: The hierarchical structure produced by agglomerative clustering can be easily interpreted through a dendrogram, allowing for a nuanced understanding of data groupings.
3. Intuitive and Interpretable: The hierarchical nature of agglomerative clustering provides a natural way of understanding the data structure at different levels of granularity.
4. Robustness to Noise and Outliers: It can be more robust to noise and outliers compared to centroid-based clustering methods. More on this below

Centroid-based clustering methods, such as K-means, work by defining clusters around central points. These central points, or centroids, are calculated as the mean of all points within a cluster. Because the mean is sensitive to extreme values, a few outliers can significantly shift the position of the centroid. This can lead to misrepresentative clustering, where the structure of the clusters is skewed by the presence of noise and outliers.

Agglomerative clustering, on the other hand, does not compute a central point that represents a cluster. Instead, it builds clusters based on the proximity of individual data points to one another. This method merges clusters based on the chosen linkage criterion, which determines how the distance between clusters is measured. The common linkage criteria are:

1. **Single Linkage:** The distance between two clusters is defined as the shortest distance from any member of one cluster to any member of the other cluster. This method can lead to "chain-like" clusters that can capture the continuity of the data but may be sensitive to noise.
2. **Complete Linkage:** The distance between two clusters is the longest distance from any member of one cluster to any member of the other cluster. This method is less sensitive to outliers than single linkage because it considers the worst-case scenario in terms of distance, thus not allowing a single outlier to have a large influence on the clustering process.
3. **Average Linkage:** The distance between two clusters is the average distance between all pairs of points in the two clusters. This balances the sensitivity between the single and complete linkage methods and is less affected by outliers than the mean calculation used in centroid-based methods.
4. **Ward's Method:** This linkage criterion minimises the total within-cluster variance. At each step, the pair of clusters with the minimum between-cluster distance are merged. This method tends to create more compact and equally sized clusters, which can be more robust to outliers compared to methods that are influenced by extreme values. In our project, we used Ward's Method to minimize impact of noises in the data.

When applying these clustering methods to financial dataset, the clusters formed by agglomerative clustering might provide a more nuanced view of the stock market, revealing intricate structures and relationships between stocks. This could be particularly useful for identifying investment opportunities that are not immediately obvious.

On the other hand, K-Means might be more useful for segmenting the market into a predefined number of distinct groups based on the chosen features. However, the assumption of spherical clusters and sensitivity to outliers could lead to less meaningful groupings in the context of financial data, which is often noisy and non-linear.

6. Results:

a. K-Means Clustering Result

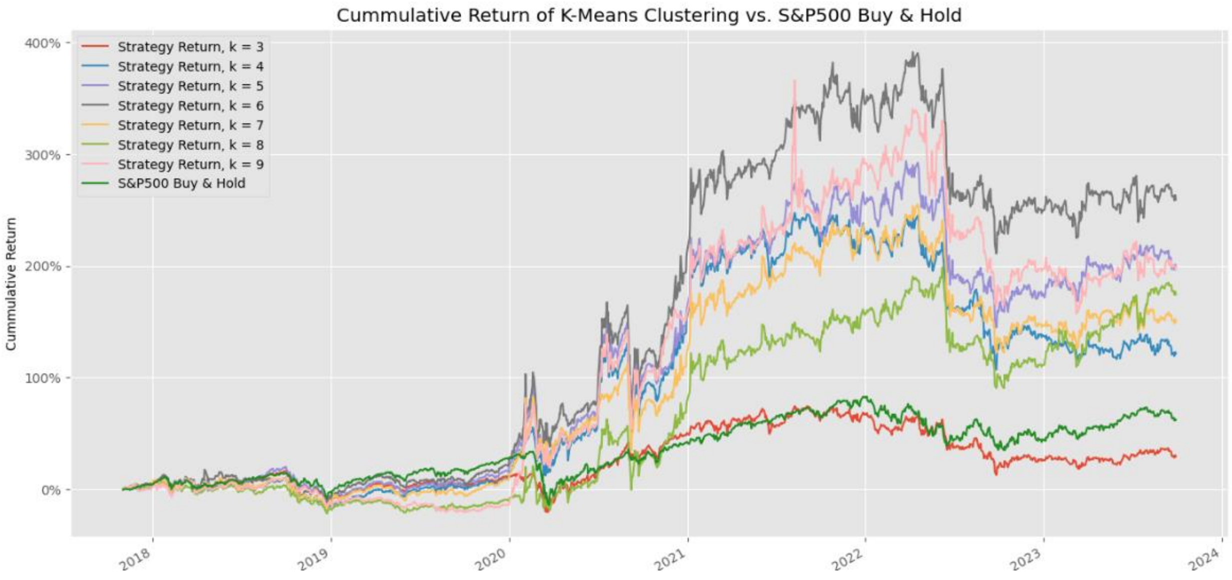


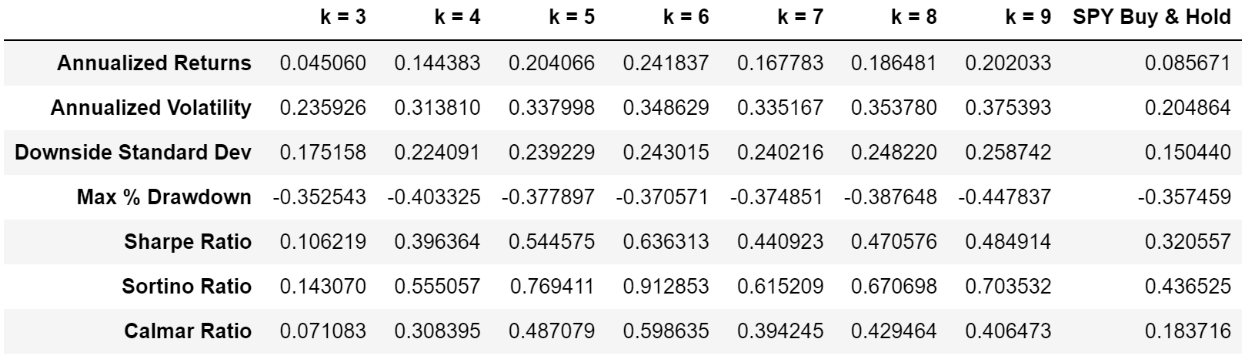
Figure 1: Portfolio Returns Over Time using K-Means Clustering

Figure 1 illustrates the time-based returns for the K Means clustering-based trading strategy, with the selection criteria being the highest mean Relative Strength Index (RSI) values for clusters of stocks to be traded each month. Stock weights within each cluster are adjusted to optimize the Sharpe ratio, reflecting risk-adjusted returns.

The depicted strategy's performance mirrors SPY's overall market trends closely, indicating a positive correlation. As the SPY index rises or falls, the portfolio's returns tend to follow similarly. Notably, the graph suggests an increase in the portfolio's sensitivity to SPY fluctuations as the value of k rises, particularly to k = 6. The heightened responsiveness to market movements with larger k values could stem from a more selective stock-picking process within each cluster as the k values increases from 3 to 6, leading to larger than proportionate movements in relation to SPY. However, as the value of k gets larger from 6 to 9, the portfolio's sensitivity to SPY fluctuations decreases. This is likely due to the selective stock-picking process causing the chosen cluster to have a highly elevated mean RSI value, resulting in the stocks chosen to likely be largely overvalued, leading to smaller returns compare to the k = 6 portfolio.

Market occurrences also seem to exert a considerable influence on the strategy's outcomes. In early 2020, the synchronized downturn of the portfolios and the SPY, followed by a recovery, likely reflects the market's initial reaction to the COVID-19 pandemic and its subsequent rally as global restrictions eased. In the third quarter of 2020, a notable rally in portfolio returns can be ascribed to the performance of specific stocks, like Amazon and Microsoft, which significantly exceeded industry forecasts, thereby significantly boosting its share price. Concurrently, escalating tensions between the US and China contributed to market volatility and widespread selling during this period, leading to a fall in portfolio returns thereafter.

Moving to 2022, the portfolio experienced a downturn and heightened volatility, which can be traced back to the geopolitical unrest initiated by Russia's invasion of Ukraine, leading to substantial market instability. The outset of 2023 marked a considerable recovery in returns, propelled by exceptional gains in stocks such as Tesla and Warner Bros Discovery, each ascending more than 60%, with companies like Catalent, Align Technology, and Royal Caribbean seeing approximately 50% increases. Despite their diverse industry representation, these stocks shared a commonality in which they were among the previous year's underperformers, each with losses surpassing 35%. In portfolios where k is equal to or greater than 5, the influence of individual stocks is magnified due to the smaller cluster size, which intensifies their impact on the portfolio's overall performance, culminating in the significant returns observed.



	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	SPY Buy & Hold
Annualized Returns	0.045060	0.144383	0.204066	0.241837	0.167783	0.186481	0.202033	0.085671
Annualized Volatility	0.235926	0.313810	0.337998	0.348629	0.335167	0.353780	0.375393	0.204864
Downside Standard Dev	0.175158	0.224091	0.239229	0.243015	0.240216	0.248220	0.258742	0.150440
Max % Drawdown	-0.352543	-0.403325	-0.377897	-0.370571	-0.374851	-0.387648	-0.447837	-0.357459
Sharpe Ratio	0.106219	0.396364	0.544575	0.636313	0.440923	0.470576	0.484914	0.320557
Sortino Ratio	0.143070	0.555057	0.769411	0.912853	0.615209	0.670698	0.703532	0.436525
Calmar Ratio	0.071083	0.308395	0.487079	0.598635	0.394245	0.429464	0.406473	0.183716

Figure 2: Key metrics of trading strategy using K-Means Clustering

Based on Figure 2, we observe that the K Means clustering trading strategy consistently outpaces the SPY, as evidenced by the higher annualized returns for the portfolio for each value of k. Notably, though, there is also an uptick in annualized volatility for the portfolio in comparison to SPY, which could be tied to the portfolio's amplified response to SPY's movements. Moreover, as the k value increases, the annualized volatility of the portfolio also rises, indicating greater variability in the returns as the cluster size decreases.

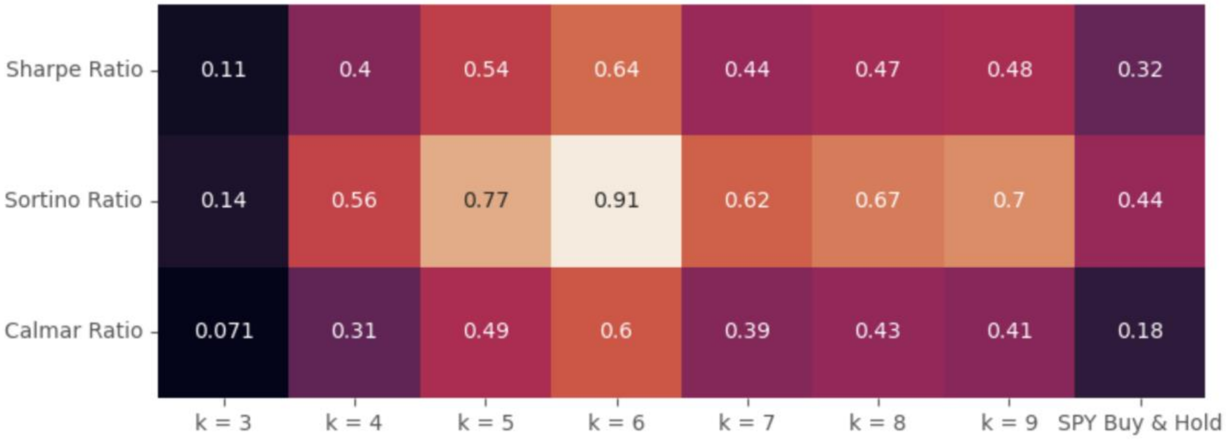


Figure 3: Key ratios using K-Means Clustering

Figure 3 translates the performance metrics from Figure 2 into a visual heatmap. This representation reveals that the performance of the portfolio peaks at k = 6, suggesting an optimal clustering scenario for the trading strategy. The heatmap also highlights that the Sortino Ratios are consistently higher than the Sharpe Ratios for all instances where k is 4 or greater, signalling that the strategy's returns are resilient to downside risks. However, this may also be influenced by the prevailing bull market conditions depicted in

Figure 4, which would naturally elevate the upside potential over the observed period, resulting in lower downside volatility.

Further, a comparison of the Sharpe and Calmar Ratios across the portfolios indicates a uniformity in risk-adjusted returns and drawdown risks, underscoring a consistent performance relative to the risks undertaken. Such a congruence is often a favourable indicator for investors, denoting that the trading strategy is delivering steady results after accounting for both volatility and potential declines.

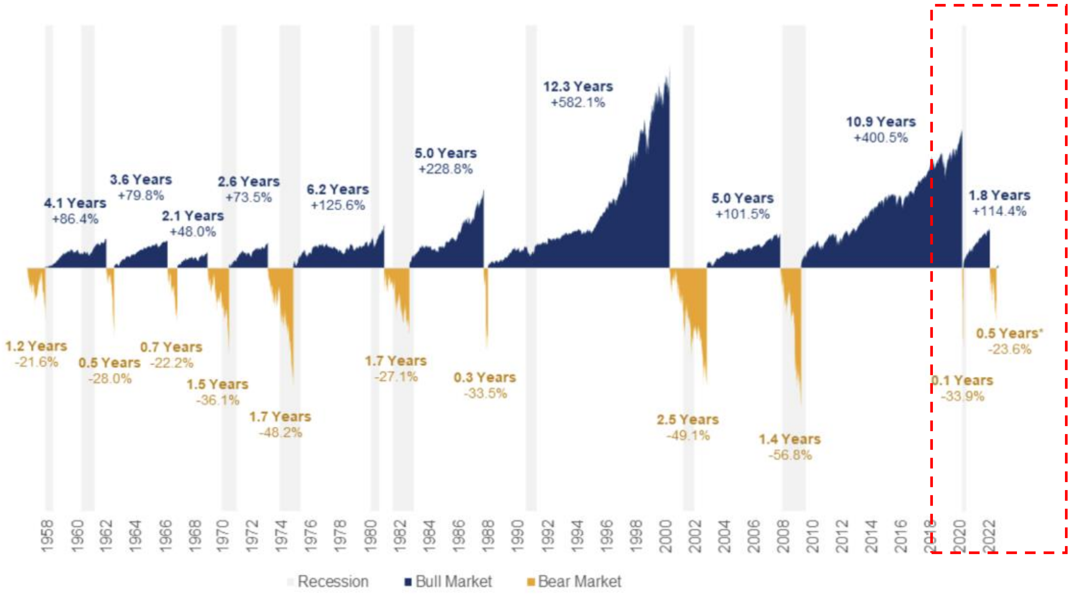


Figure 4: History of U.S equity of bull & bear markets by Royal Bank of Canada

b. Agglomerative Clustering Result

Figure 5 illustrates the time-based returns for the Agglomerative clustering-based trading strategy, with the selection criteria being the highest mean Relative Strength Index (RSI) values for clusters of stocks to be traded each month. Stock weights within each cluster are adjusted to optimize the Sharpe ratio, reflecting risk-adjusted returns. We have reiterated the project 7 times over, each with a differing number of clusters set from 3 to 9.

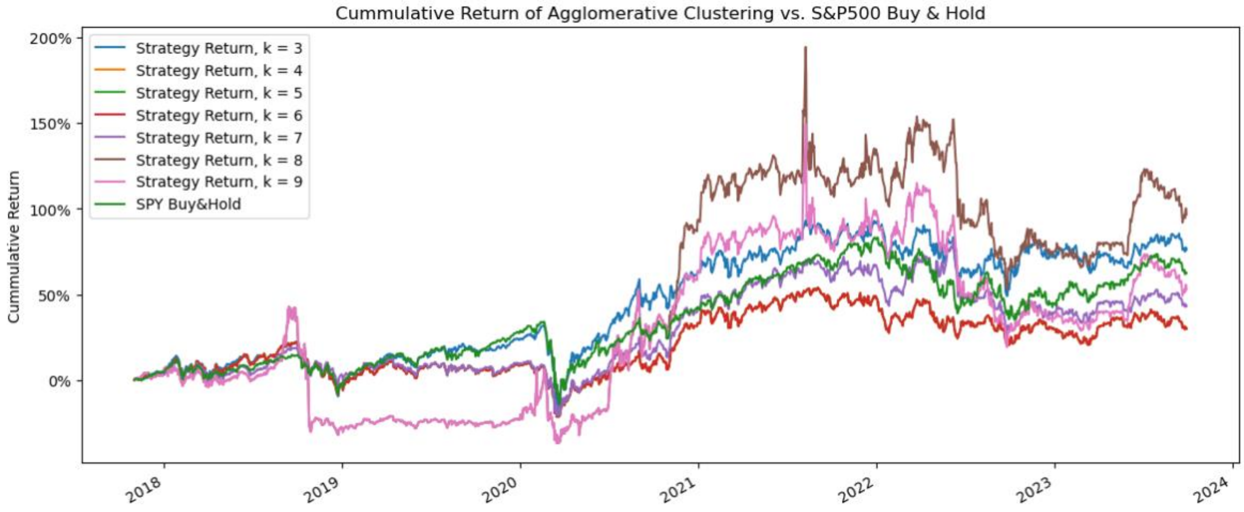
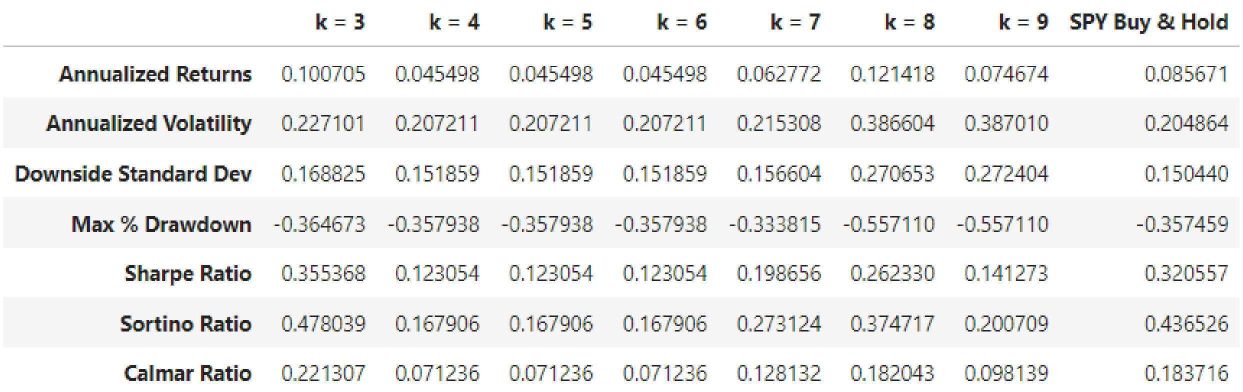


Figure 5: Returns of Strategy with varying number of clusters and S&P500 Buy and Hold strategy

The depicted strategy's performance mirrors SPY's overall market trends closely, indicating a positive correlation. As the SPY index rises or falls, the portfolio's returns tend to follow similarly. Notably, the graph suggests an increase in the portfolio's sensitivity to SPY fluctuations as the value of k rises. This heightened responsiveness to market movements with larger k values could stem from a more selective stock-picking process within each cluster, which may more closely resonate with the adopted trading approach.



	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	SPY Buy & Hold
Annualized Returns	0.100705	0.045498	0.045498	0.045498	0.062772	0.121418	0.074674	0.085671
Annualized Volatility	0.227101	0.207211	0.207211	0.207211	0.215308	0.386604	0.387010	0.204864
Downside Standard Dev	0.168825	0.151859	0.151859	0.151859	0.156604	0.270653	0.272404	0.150440
Max % Drawdown	-0.364673	-0.357938	-0.357938	-0.357938	-0.333815	-0.557110	-0.557110	-0.357459
Sharpe Ratio	0.355368	0.123054	0.123054	0.123054	0.198656	0.262330	0.141273	0.320557
Sortino Ratio	0.478039	0.167906	0.167906	0.167906	0.273124	0.374717	0.200709	0.436526
Calmar Ratio	0.221307	0.071236	0.071236	0.071236	0.128132	0.182043	0.098139	0.183716

Figure 6: Key metrics of trading strategy using Agglomerative Clustering

Starting with the k=3 strategy, it stands out with an annualized return of approximately 10.07%, which is a robust performance, particularly when compared to the SPY Buy & Hold's return of 8.57%. Despite having a higher annualized volatility (22.71% vs. 20.49% for SPY), the k=3 strategy's Sharpe ratio is 0.355, which is higher than that of the SPY strategy at 0.321. This suggests that the k=3 strategy, while riskier, has been more efficient in converting risk into return over the period measured. However, it's important to note that the k=3 strategy also experienced the highest maximum drawdown at -36.47%, indicating that investors in this strategy would have seen a significant temporary decline in portfolio value at some point, which is a concern for risk-averse investors.

For strategies k=4 through k=6, the results are identical, which is unusual and might imply that these strategies perform similarly across different market conditions or that there's an error in the data. They all have an annualized volatility of 20.72%, an annualized return of 4.55%, and a Sharpe ratio of 0.123. The maximum drawdown for these strategies is also the same at -35.79%. These metrics indicate a lower performance compared to both the k=3 strategy and the SPY Buy & Hold, with a significantly lower return and a similar level of risk as the SPY, which is not an attractive combination.

The k=7 strategy shows a slight improvement over k=4 to k=6, with a higher annualized return of 6.28% and a Sharpe ratio of 0.199. Its maximum drawdown is the lowest among all strategies at -33.38%, which may make it more appealing to those who prioritize capital preservation. However, its Sharpe ratio is still below that of the SPY Buy & Hold, indicating that it may not be as efficient in terms of risk-adjusted returns.

The introduction of k = 8 and k = 9 strategies presents a more comprehensive view of the clustering results. The k = 8 strategy stands out with its significantly higher annualized return at 12.14% but comes with elevated volatility, with maximum drawdown reaching -55.7%. However, compared to k=3 strategy with , the additional risk does not justify a small increase in annualized return. As such, its Sharpe Ratio, which is risk-adjusted returns still falls short of outperforming the SPY Buy & Hold strategy or k=3 strategy.

Conversely, the k = 9 strategy, while offering a higher return compared to k = 4 to k = 7, exhibits lower risk-adjusted returns and shares the same -55.7% maximum drawdown as the k = 8 strategy. This strategy is even less appealing compared to k = 8 given the insignificant improvement in annualized return at ~7.5%

In comparison, the k = 3 strategy continues to demonstrate strong risk-adjusted returns, making it appealing to investors willing to accept higher volatility for the potential of superior long-term returns. The SPY Buy & Hold strategy remains a relatively safer choice for risk-averse investors, offering lower but respectable risk-adjusted returns.

The relationship between the number of clusters (k) in and the performance of the strategy is not straightforward and can be influenced by various factors. The nature of the data is critical; if the dataset lacks a natural cluster structure or the clusters are not pertinent to the predictive task, increasing k may not enhance performance. The stability of clusters is another factor; agglomerative clustering's results can vary depending on how 'k' is chosen when cutting the dendrogram. This variability can lead to inconsistent performance if the data doesn't exhibit well-defined groups. Moreover, a higher k risks overfitting to noise, which can degrade out-of-sample performance. The dimensionality of the data and the choice of distance metric are also crucial; high-dimensional data can dilute the effectiveness of distance measures, and an inappropriate metric may yield unhelpful clusters.

In comparing all the strategies, the k=3 strategy seems to offer the best risk-adjusted returns, as evidenced by its higher Sharpe ratio. This would be attractive to investors who are willing to accept higher volatility and the potential for significant temporary losses in exchange for higher long-term returns. The SPY Buy & Hold strategy, while less volatile, offers lower returns but still maintains a respectable Sharpe ratio, making it a potentially safer choice for the risk averse. The identical performance of the k=4 to k=6 strategies raise questions and may require further investigation to understand the underlying reasons for their performance parity. The k=7 strategy, with its lower drawdown, could be considered by those looking for a more conservative approach among the k strategies but still falls short of the SPY benchmark in terms of efficiency.

The superior performance of the k=3 strategy over the S&P buy-and-hold approach suggests that a lower number of clusters may offer an optimal balance for the dataset and period analysed. Three clusters might align with significant market segments or asset behaviours, providing a robust strategy less prone to overfitting and better suited to the economic conditions during the back testing period. This strategy may also benefit from a more stable and less data-sensitive clustering, leading to a more consistent and reliable performance.

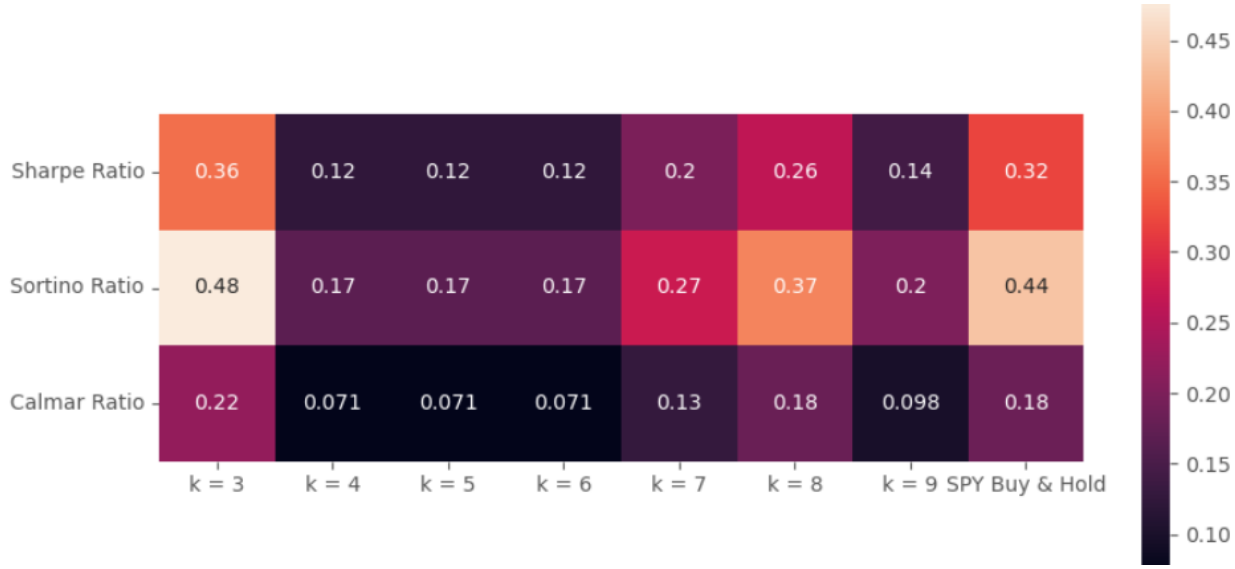


Figure 7: Key ratios using Agglomerative Clustering

Figure 7 translates the performance metrics of *Agglomerative Clustering* from Figure 6 into a visual heatmap.

The k=3 strategy demonstrates strong risk-adjusted performance with a relatively high Sharpe ratio (0.355) and an even higher Sortino ratio (0.478). This suggests that the strategy has been efficient in converting risk into return over the measured period. However, the strategy comes with a relatively high annualized volatility of 22.71%, and it experienced a significant maximum drawdown of -36.47%.

For k=4 to k=6, the strategies exhibit lower risk-adjusted performance compared to k=3. The Sharpe ratios for these strategies are notably lower, ranging from 0.123 to 0.168. These strategies have a lower annualized return (around 4.55%) and lower Sortino ratios, indicating lower efficiency in managing downside risk. The maximum drawdown is still relatively high at around -35.79%.

The k=7 strategy shows improvement over k=4 to k=6 in terms of risk-adjusted performance. It has a higher annualized return of 6.28% and a higher Sharpe ratio of 0.199. The Sortino ratio also indicates better efficiency in managing downside risk (0.273). The maximum drawdown is the lowest among these strategies at -33.38%, which may make it more appealing to those who prioritize capital preservation.

The k=8 strategy exhibits significantly higher annualized volatility (38.66%) and downside volatility (27.07%) compared to previous strategies. However, it also has the highest annualized return of 12.14% among the strategies considered. The Sharpe ratio (0.262) and Sortino ratio (0.375) indicate better risk-adjusted performance, although the strategy comes with a substantial maximum drawdown of -55.71%.

Similar to k=8, the k=9 strategy has high volatility measures (38.70% annualized volatility and 27.24% downside volatility) but offers a lower annualized return of 7.47%. The Sharpe ratio (0.141) and Sortino ratio (0.201) are lower compared to previous strategies, indicating less efficient risk-adjusted performance. The maximum drawdown remains at -55.71%.

In summary, the choice of k in agglomerative clustering significantly impacts the risk-return profile of the trading strategy. However, there isn’t a linear relationship between increase in the number of clusters and increase in risk-adjusted return. As such, we will need to look into the fundamental aspects, i.e. investigating characteristics of the companies in each cluster at different k values.

c. Comparing Results of 2 Clustering Methods

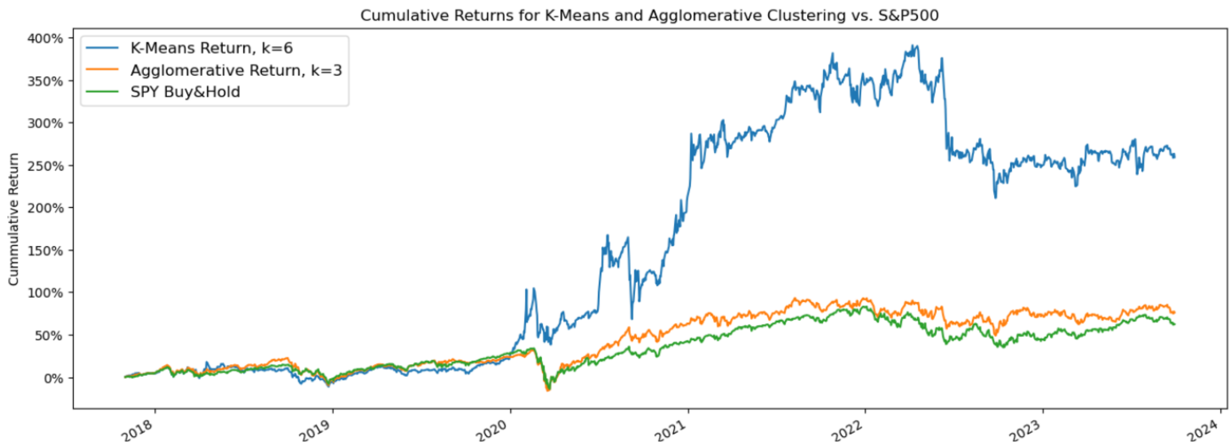
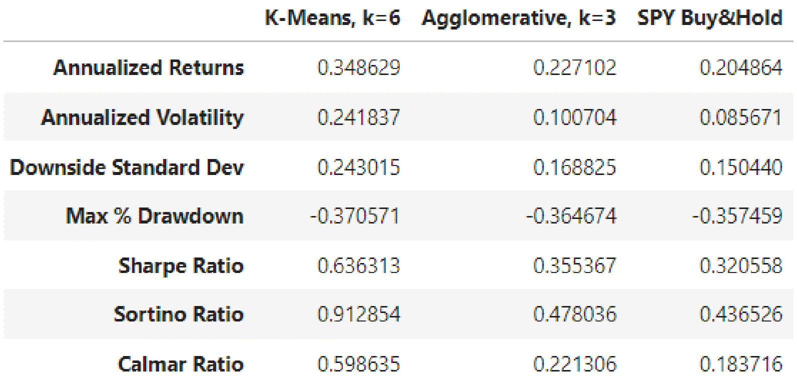


Figure 8: Cumulative Returns for K-Means and Agglomerative Clustering vs. S&P500

Comparing both models, we evaluate their performance based on the strategy with the highest Sharpe Ratio. For K-Means Clustering, Strategy with 6 clusters is selected, while for Agglomerative Clustering, Strategy with 3 clusters is chosen. Analysing the Cumulative Return visualization in Figure 8, it becomes evident that K-Means Clustering outperforms both Agglomerative Clustering and the S&P 500 Buy & Hold Strategy.



	K-Means, k=6	Agglomerative, k=3	SPY Buy&Hold
Annualized Returns	0.348629	0.227102	0.204864
Annualized Volatility	0.241837	0.100704	0.085671
Downside Standard Dev	0.243015	0.168825	0.150440
Max % Drawdown	-0.370571	-0.364674	-0.357459
Sharpe Ratio	0.636313	0.355367	0.320558
Sortino Ratio	0.912854	0.478036	0.436526
Calmar Ratio	0.598635	0.221306	0.183716

Figure 9: Key Metrics for performance in K-Means and Agglomerative Clustering vs. S&P500

Annualized Returns: The K-Means strategy (k=6) has the highest annualized return at 34.86%, indicating it performed the best in terms of returns among the three. Agglomerative (k=3) also performed well with a 22.71% annualized return, outperforming the SPY Buy & Hold benchmark (20.49%).

Volatility: K-Means (k=6) has the highest volatility (24.18%), suggesting it carries higher risk compared to the other strategies. Agglomerative (k=3) has lower volatility (10.07%) and is less risky than both K-Means and the benchmark.

Downside Risk: Downside Standard Deviation measures the risk of negative returns. K-Means (k=6) has the highest downside risk (24.30%), indicating a higher chance of loss in the strategy. Agglomerative (k=3) has a lower downside risk (16.88%), making it more resilient to downturns than K-Means.

Max % Drawdown: K-Means (k=6) and Agglomerative (k=3) have similar drawdowns, both exceeding -36%, while the SPY Buy & Hold benchmark has a slightly lower drawdown (-35.75%). Lower drawdowns are generally preferable as they represent lower losses during market downturns.

Risk-Adjusted Metrics: The Sharpe Ratio, Sortino Ratio, and Calmar Ratio are measures of risk-adjusted performance. K-Means (k=6) has the highest Sharpe and Sortino ratios, indicating better risk-adjusted returns. However, Agglomerative (k=3) has a higher Calmar ratio compared to K-Means, suggesting that it performed better relative to drawdown risk.

Furthermore, the observation that Agglomerative Clustering closely follows the S&P 500 in almost identical patterns while K-Means exhibits more significant and diverse movements can have several implications:

Market Dependence: K-Means, with its sensitivity to changes in market dynamics, can help identify assets that perform well in specific market conditions. Whereas Agglomerative Clustering's alignment with the S&P 500 suggests that it tends to mirror overall market trends. This can be advantageous during stable or bullish market conditions when the broader market is performing well. However, it may also mean that Agglomerative Clustering is less effective at identifying assets that outperform the market during downturns compared to K-Means Clustering.

Diversification: K-Means' diverse movements may indicate that it has the potential to identify assets with varying performance characteristics across different market conditions. This can be valuable for diversifying a portfolio. Investors can use K-Means to select assets that behave differently from the market, potentially reducing overall portfolio risk.

Risk Management: The significant and diverse movements of K-Means may imply higher risk during certain periods, even though the Maximum Drawdown of K-Means Clustering in our example is similar to that of Agglomerative Clustering or the S&P 500 Buy & Hold strategies. It is uncertain to conclude definitively if K-Means is a superior option. However, investors using K-Means should remain vigilant regarding its sensitivity to market dynamics and consider taking appropriate risk management measures. These measures might include adjusting portfolio allocations or implementing stop-loss strategies.

In conclusion, the difference in movement patterns between Agglomerative Clustering and K-Means suggests that each method has its strengths and weaknesses. Investors can use this information to tailor their strategies to the prevailing market conditions and achieve a more balanced and adaptive investment approach. By combining both strategies or adjusting the allocation of assets based on market conditions, investors can potentially achieve a more balanced and resilient portfolio.

7. Recommendation for improvements

Further Investigation of Agglomerative Clustering Results

The observed performance parity among strategies with k=4 to k=9 warrants an in-depth examination of the entire clustering process, from feature selection to the choice of distance metric and linkage method. This ensures that the clustering parameters are optimized for each value of k. It's possible that a suboptimal choice of parameters led to similar cluster assignments.

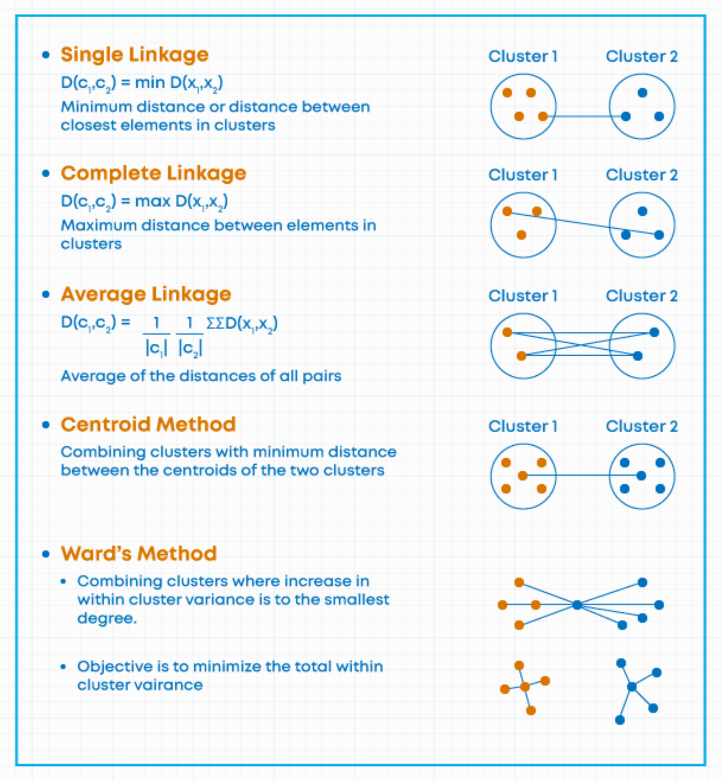

Single Linkage
D(c1,c2) = min D(x1,x2)
Minimum distance or distance between closest elements in clusters
Cluster 1
Cluster 2
Complete Linkage
D(c1,c2) = max D(x1,x2)
Maximum distance between elements in clusters
Cluster 1
Cluster 2
Average Linkage
D(c1,c2) = 1/|c1| 1/|c2| ΣΣD(x1,x2)
Average of the distances of all pairs
Cluster 1
Cluster 2
Centroid Method
Combining clusters with minimum distance between the centroids of the two clusters
Cluster 1
Cluster 2
Ward's Method
Combining clusters where increase in within cluster variance is to the smallest degree.
Objective is to minimize the total within cluster vairance

Figure 10: Different linkage methods in Agglomerative Clustering

Regarding the choice of the agglomerative approach, we selected Ward's Method, which is commonly used to perform Agglomerative Clustering with the goal of noise reduction (Figure 10). However, it's important to note that Ward's method is biased towards globular clusters, which have spherical or nearly spherical shapes. This bias may not accurately represent the underlying structure of the data when clusters have different shapes, such as elongated or irregular forms.

As a result, it may be worthwhile to explore alternative hierarchical clustering algorithms, such as Complete or Average Linkage in Figure 10. This exploration can help determine if the observed pattern of performance parity persists across different linkage methods and whether if other linkage methods lead to a better performance.

Alternative Clustering Algorithms

During our experiments, we exclusively explored *K-Means, Agglomerative Clustering and DBSCAN*. While *K-Means, Agglomerative Clustering* generated satisfactory results and were our focus in this report, the DBSCAN Method didn't yield satisfactory results. The issue was that the clusters generated by DBSCAN were heavily skewed, with almost all data points concentrated in a single cluster. Additionally, some clusters had almost no data points at all, indicating a challenge in identifying meaningful clusters (Figure 11). Despite these challenges, DBSCAN remains valuable for its ability to handle complex, non-linear data structures and noisy data, and it's worth revisiting with adjusted parameters.

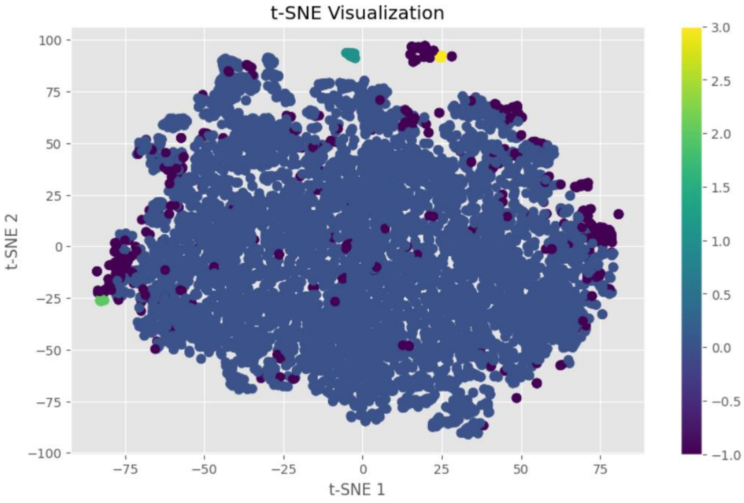


Figure 11: DSBCAN failed to identify meaningful clusters

The quantitative trading domain offers a diverse array of clustering methods (Figure 12), such as *Birch*, and *Gaussian Mixture*. Expanding our experimentation may yield improved results and offer deeper insights into the relationships among factors within our dataset.

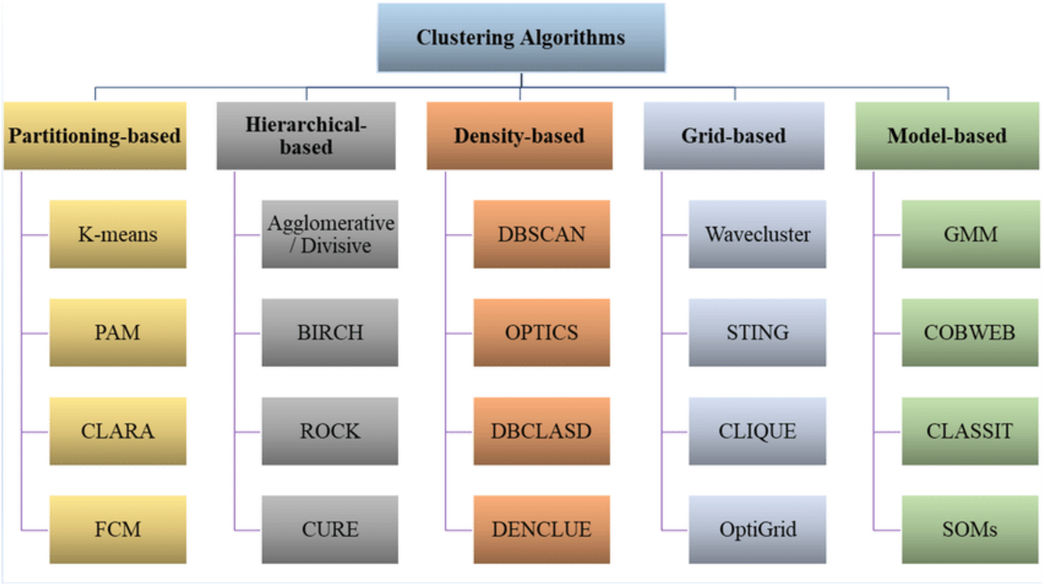


Figure 12: Potential Clustering Methods to experiment

8. Conclusion

The consolidation of findings from the application of K-Means and Agglomerative Clustering to trading strategy formulation presents a compelling case for the integration of unsupervised machine learning techniques in financial markets. The business problem at hand is the identification of coherent groups or clusters of financial assets for enhanced portfolio management. The application of these techniques has demonstrated potential in addressing this challenge, as evidenced by the performance metrics of the developed trading strategies.

The results indicate that the use of clustering to form investment strategies can lead to improved risk-adjusted returns compared to traditional market strategies such as the SPY Buy & Hold approach. For example, the three-cluster strategy derived from Agglomerative Clustering outperformed other strategies in terms of the Sharpe ratio, suggesting a superior balance of risk and return. This finding underscores the value of determining the optimal number of clusters to maximize the efficacy of the trading strategy.

However, the increased maximum drawdown associated with the three-cluster strategy also highlights the trade-off between risk and return. While the strategy may offer higher returns, it also exposes investors to potential higher losses. This trade-off is a critical consideration for portfolio managers who must balance the desire for higher returns against the tolerance for risk.

The use of K-Means and Agglomerative Clustering also offers insights into the dynamic nature of financial markets. The variation in portfolio performance with different numbers of clusters (k values) suggests that market conditions and external events significantly influence asset behaviour. For instance, the impact of the COVID-19 pandemic and geopolitical tensions on market volatility underscores the need for adaptive strategies that can respond to sudden market shifts.

To solve the business problem effectively, portfolio managers should consider the following steps:

1. **Dynamic Cluster Adjustment:** Regularly re-evaluate and adjust the number of clusters to reflect current market conditions and data. This adaptive approach can help maintain an optimal balance between risk and return.
2. **Diversification Across Clusters:** Allocate investments across multiple clusters to spread risk. This strategy can mitigate the impact of a poor performance in any single cluster.
3. **Tailored Risk Management:** Implement risk management strategies tailored to the characteristics of each cluster. For example, clusters with higher volatility may require more conservative investment approaches.
4. **Continuous Monitoring and Rebalancing:** Monitor cluster performance continuously and rebalance portfolios as needed. This practice can capitalize on the strengths of machine learning to detect and respond to changes in asset behaviour.
5. **Integration with Other Analytical Tools:** Combine clustering with other analytical tools and financial models to create a more robust trading strategy. Machine learning can complement traditional financial analysis by providing additional layers of insight.
6. **Back-testing and Forward Testing:** Extensively back-test and forward test trading strategies based on clustering to validate their effectiveness over different time periods and market conditions.

In conclusion, the integration of K-Means and Agglomerative Clustering into trading strategies represents a promising advancement in financial analytics. By leveraging the strengths of unsupervised machine learning, financial professionals can enhance decision-making processes, achieve better risk-adjusted returns, and navigate the complexities of the market with greater confidence. The key to success lies in the careful application, continuous evaluation, and strategic combination of these techniques with conventional financial wisdom.

9. References

Kenneth R. French - Description of Fama/French Factors. (2023). Dartmouth.edu. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_5_factors_2x3.html

S.M. Ikhtiar Alam. (2022, March 28). Fama-French 5-Factor Model and Its Applications. ResearchGate; unknown. https://www.researchgate.net/publication/359505728_Fama-French_5-Factor_Model_and_Its_Applications.

Broad classification of clustering algorithms - researchgate. (n.d.). https://www.researchgate.net/figure/Broad-classification-of-clustering-algorithms_fig2_341111980

Great Learning Team. (2023, July 19). What is hierarchical clustering? an introduction to hierarchical clustering. Great Learning Blog: Free Resources what Matters to shape your Career! https://www.mygreatlearning.com/blog/hierarchical-clustering/