AI-Driven ETF Grouping Using Clustering

Daniel Mehta

Faculty Applied Sci & Tech, Humber Polytechnic

AIGC-5503-RNA: AI for Bus. Decision Making

Alex Dela Cruz

August 13, 2025

Introduction

Problem Statement

ETFs package groups of stocks into single, theme-driven products that match investor goals like growth, income, or low volatility. Deciding which stocks belong together takes time and can be inconsistent across analysts. An AI-driven clustering workflow can quickly surface coherent groups based on return, volatility, momentum, liquidity, and drawdown. The result is faster ETF idea generation, more consistent grouping criteria, and a repeatable path from raw market data to candidate ETF lineups.

Dataset

Analysis uses *all_stocks_5yr.csv* from the Kaggle S&P 500 collection [Link]. The file contains five years of daily data for companies in the S&P 500 up to February 2018, with columns: Date, Open, High, Low, Close, Volume, and Name (ticker). This single merged file suits modelling pipelines that compute series-derived features per ticker, such as average daily return, volatility, 30-day momentum, average volume, and maximum drawdown. Coverage across large-cap U.S. equities and multi-year depth make this dataset well matched to clustering experiments that mimic ETF construction based on market behaviour.

AI Algorithms Comparison

K-Means Clustering

K-Means is a partitioning method that assigns each data point to one of a set number of clusters by minimizing the sum of squared distances to the cluster center. It works quickly, handles large datasets well, and is a common approach in market segmentation. The trade-off is that the number of clusters must be set in advance, and the method works best when clusters are compact and similar in size. It can also be influenced by outliers, which may distort the cluster centers.

Hierarchical Clustering

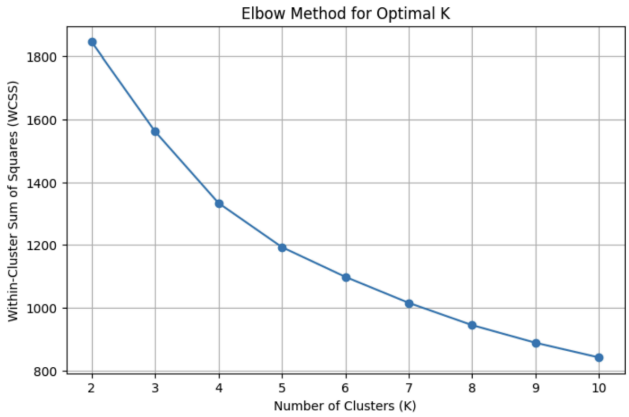
Hierarchical clustering creates a tree-like structure of clusters by progressively merging smaller groups or splitting larger ones. This project uses the agglomerative style with Ward's linkage, which merges clusters to minimize the increase in within-cluster variance. The advantage is flexibility: the number of clusters does not need to be chosen at the start, and the dendrogram makes it easier to explore relationships between groups. The downside is that it can be slower with large datasets and may produce less compact clusters when the data is noisy or not naturally separable.

Evaluation Method

The Elbow Method was used with K-Means to find a suitable number of clusters by checking how the within-cluster sum of squares (WCSS) changed for k values from 2 to 10. The point where improvements began to level off was at k = 4. Both K-Means and Hierarchical clustering were then run with k = 4 to make the comparison fair. Two metrics were used to compare results:

- Silhouette Score measures how well each point fits in its cluster compared to others. Higher values indicate better separation.

- The Davies-Bouldin Index compares the compactness of clusters to their inter-cluster distances. Lower values indicate better-defined clusters.



Advantages and Disadvantages

Algorithm	Advantages	Disadvantages
K-Means	Fast and scalable; simple to apply; effective when clusters are compact and well-separated	Must choose k beforehand; assumes roughly spherical, similar-sized clusters; sensitive to outliers
Hierarchical	Flexible choice of k; dendrogram helps with interpretation; can reveal nested groupings	Slower for large datasets; more affected by noisy data; may produce less compact clusters.

AI System Recommendation

Preprocessing Pipeline

The workflow starts by cleaning the OHLCV (Open, High, Low, Close, and Volume) stock data, removing rows with missing prices, dropping duplicates, and sorting by ticker and date. Feature engineering creates five investment-relevant metrics for each stock: average daily return, volatility, 30-day momentum, average trading volume, and maximum drawdown. These features capture both performance and risk characteristics that matter for ETF construction. All features are standardized with *StandardScaler* so that differences in scale, such as between volume and returns, do not distort the clustering process.

Recommended Algorithm

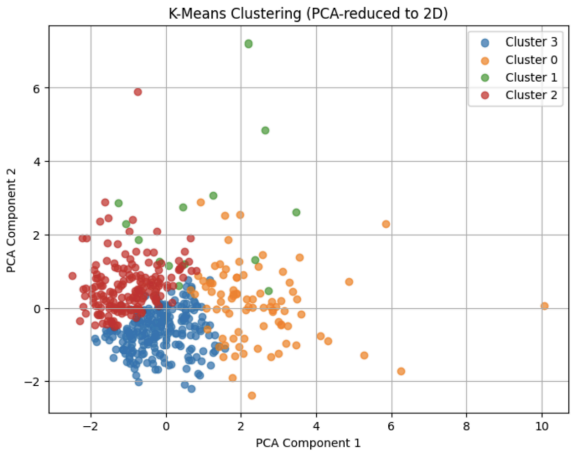
K-Means with four clusters is the preferred choice. It scored higher on the Silhouette metric and lower on the Davies-Bouldin Index compared to hierarchical clustering, indicating more cohesive and distinct groups. The PCA scatterplot shows that K-Means produced tighter, more separable clusters, which makes interpretation easier for portfolio design. The method also runs quickly and can be scaled to larger datasets, making it practical for an ETF grouping tool.

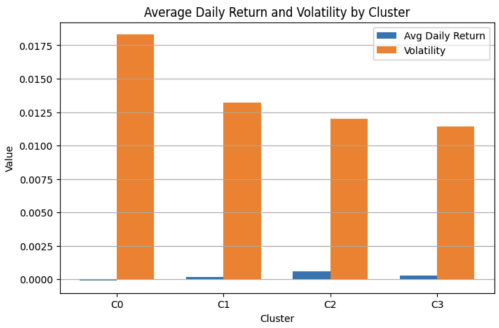
Dashboard

The dashboard presents visual summaries to help decision-makers understand and compare clusters:

- **PCA scatterplot** to show how the clusters separate in reduced dimensions.
- **Cluster summary table** with average values for returns, volatility, momentum, volume, and drawdown in original units.
- **Risk-return bar chart** to compare each cluster's trade-off between average daily return and volatility.
- **Feature heatmap** showing z-scored averages per cluster to highlight distinctive traits such as high momentum or low drawdown.

These views allow an investment professional to quickly spot high-performing, low-risk groups, niche segments, and high-risk, high-volatility sets when considering ETF construction.





Limitations & Ethical Considerations

Limitations

The clusters are based only on historical daily returns, volatility, 30-day momentum, average volume, and maximum drawdown. This means there is no guarantee that the same groupings will perform well in the future. The model does not include fundamental ratios such as P/E, earnings growth, or dividend yield, nor does it consider correlations between stocks, which could improve grouping quality. Silhouette scores in this analysis are relatively low, which is expected in financial data where stock profiles often overlap. The analysis is a static snapshot, and the results may not adapt to shifting market conditions without retraining or using rolling windows.

Ethical Considerations

The output must be clearly labelled as "not financial advice" to avoid misuse. Investors acting solely on these groupings could face significant risk if market conditions change. Transparency is important, so the feature definitions, clustering process, and limitations should be fully documented. Care should also be taken to ensure the process does not systematically favour or exclude specific sectors or company sizes without valid financial reasoning.

References

2.3. clustering. scikit. (n.d.-a). https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index

2.3. clustering. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

2.3. clustering. scikit. (n.d.-c). https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

2.3. clustering. scikit. (n.d.-d). https://scikit-learn.org/stable/modules/clustering.html#k-means

Chen, J. (2025, April 29). *Exchange-traded fund (ETF): What it is and how to invest*. Investopedia. https://www.investopedia.com/terms/e/etf.asp

Hierarchical clustering (scipy.cluster.hierarchy)#. Hierarchical clustering (scipy.cluster.hierarchy) - SciPy v1.16.1 Manual. (n.d.). https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

Matplotlib 3.10.5 documentation#. Matplotlib documentation - Matplotlib 3.10.5 documentation. (n.d.). https://matplotlib.org/stable/index.html

Nugent, C. (2018). all_stocks_5yr.csv. https://www.kaggle.com/datasets/camnugent/sandp500/data.

NumPy documentation#. NumPy documentation - NumPy v2.3 Manual. (n.d.). https://numpy.org/doc/stable/

Pandas documentation#. pandas documentation - pandas 2.3.1 documentation. (2025, July 7). https://pandas.pydata.org/docs/