AI-Driven ETF Grouping Using Clustering

By: Daniel Mehta

Problem & Goal

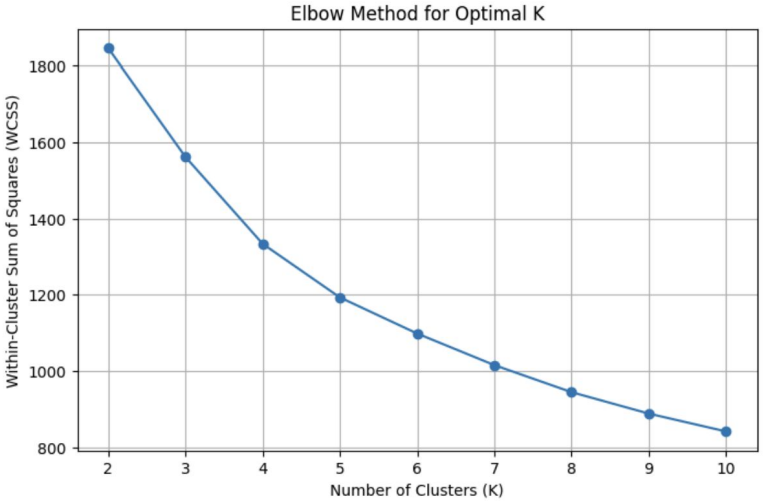
Problem:

- ETF construction is slow, subjective, and can be inconsistent.

Goal:

- Use AI clustering to group stocks based on performance and risk metrics for faster, repeatable ETF idea generation.

Data & Method



- **Dataset**: S&P 500 (5 years daily OHLCV from Kaggle).
- **Features**: Avg daily return, volatility, 30-day momentum, avg volume, max drawdown.
- **Algorithms Tested:**
  - K-Means (chosen)
  - Hierarchical Clustering (compared)
- **Cluster Selection**: Elbow method -> k = 4

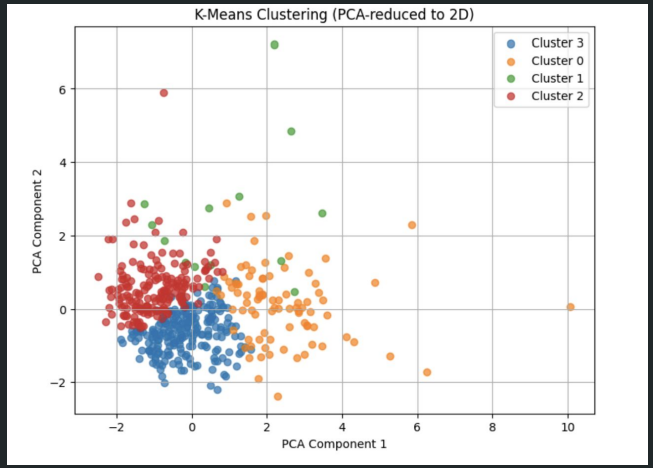
Results

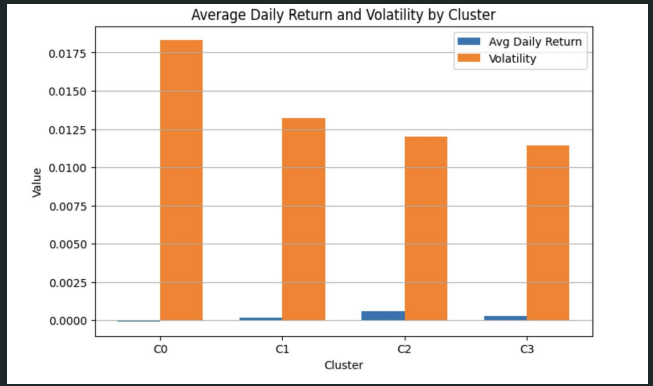
Metrics:

- K-Means had higher Silhouette score and lower Davies-Bouldin Index than Hierarchical.

Why K-Means:

- Tighter, more distinct clusters.
- Scalable for larger datasets.





Limitations & Next Steps

Limitations:

- Only uses historical data; no guarantee of future performance.
- Does not include fundamentals like P/E, dividend yield, earnings growth.
- Low silhouette scores common in financial data.

Ethics: Must label output as “Not Financial Advice.”

Next Steps:

- Add more features and live market data.
- Explore correlation-based grouping.