Clustering-Based Diversification In Financial Portfolios

Carolyn Soo (csjy@stanford.edu)

Data: • **1225 days of daily closing prices for 1199 ETF tickers** @ Kaggle (Full Historical Daily Price + Volume Data For All U.S. Stocks & ETFs)
• **High-level industry/sector classifications** @ Bloomberg Professional
The ticker names were left-joined with the classification codes, as the two datasets had to be obtained separately.

References: • B. Marjanovic. Huge Stock Market Dataset. • Bloomberg Professional 2017 • F. Cai, N. Le-Khac, M. Kechadi. 2016. "Clustering Approaches for Financial Data Analysis: a Survey."
• G. James, D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2017 • R. Tibshirani. 2013. "Clustering 2: Hierarchical clustering"

Predicting:

Given the volatility of stock markets, **an ideal investment portfolio should be diverse** in both industry type as well as asset class, to avoid steep drops in value when positively-correlated holdings decline together.
Investing in Exchange-Traded Funds (publicly-traded securities which track collections of assets) provides easy diversification, but if we can **identify underlying similarities** within a list of ETFs and classify the tickers into **"similar" (highly-correlated) groups**, then we can **make sure to control the amount of portfolio weight assigned to each such group to achieve PORTFOLIO DIVERSIFICATION.**
This project focused on **clustering-based classification strategies** to group the ETFs, including both a partitional technique (**k-means**) and a hierarchical one (**agglomerative hierarchical**).
Experimented with **Principal Component Analysis** for dimensionality reduction.

Features of interest:

• From raw input: **Trading Date, ETF Ticker Name, Daily Closing Price** (Dividend- and Split-Adjusted) • Derived: **Daily One-Day Log Return**. These features are appropriate: A ticker's single day log return value becomes one dimension of many; this transformation **skirts the issue of time-series modeling**, which removes the need for complex considerations e.g. walk-forward-only validation, seasonality detection, heteroskedasticity. These features are sufficient: The algorithms do not require more informative features (unlike e.g. logistic regression, Naïve Bayes classifier) to operate.

Models (Unsupervised Classification):

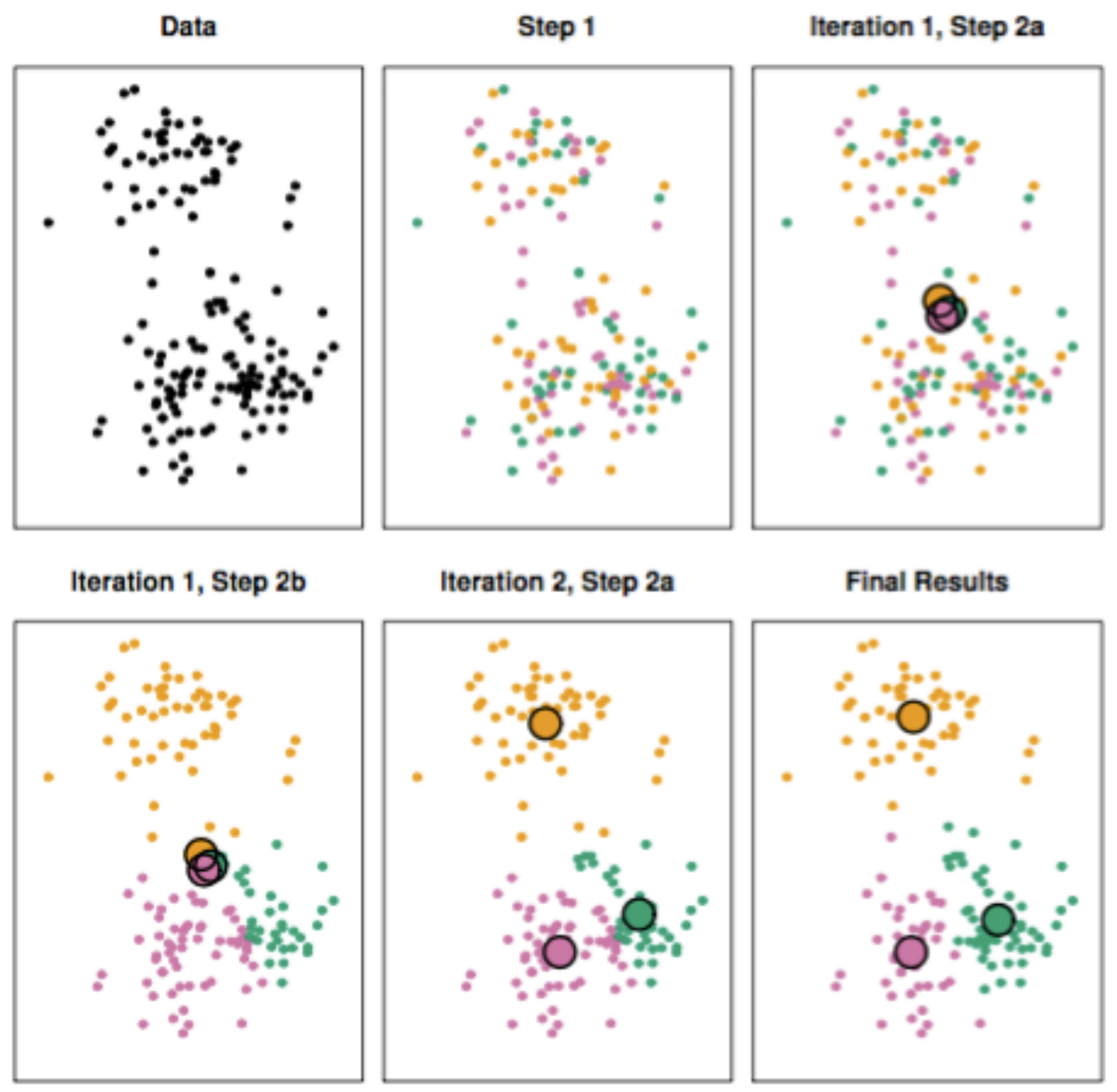
K-Means Clustering:

- The observations are initially and arbitrarily assigned to one of K clusters
- Until the clustering stabilizes, iteratively:
  - ○ Compute the centroid for every cluster $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$
  - ○ Partition each observation x_i to the cluster with the closest centroid of all

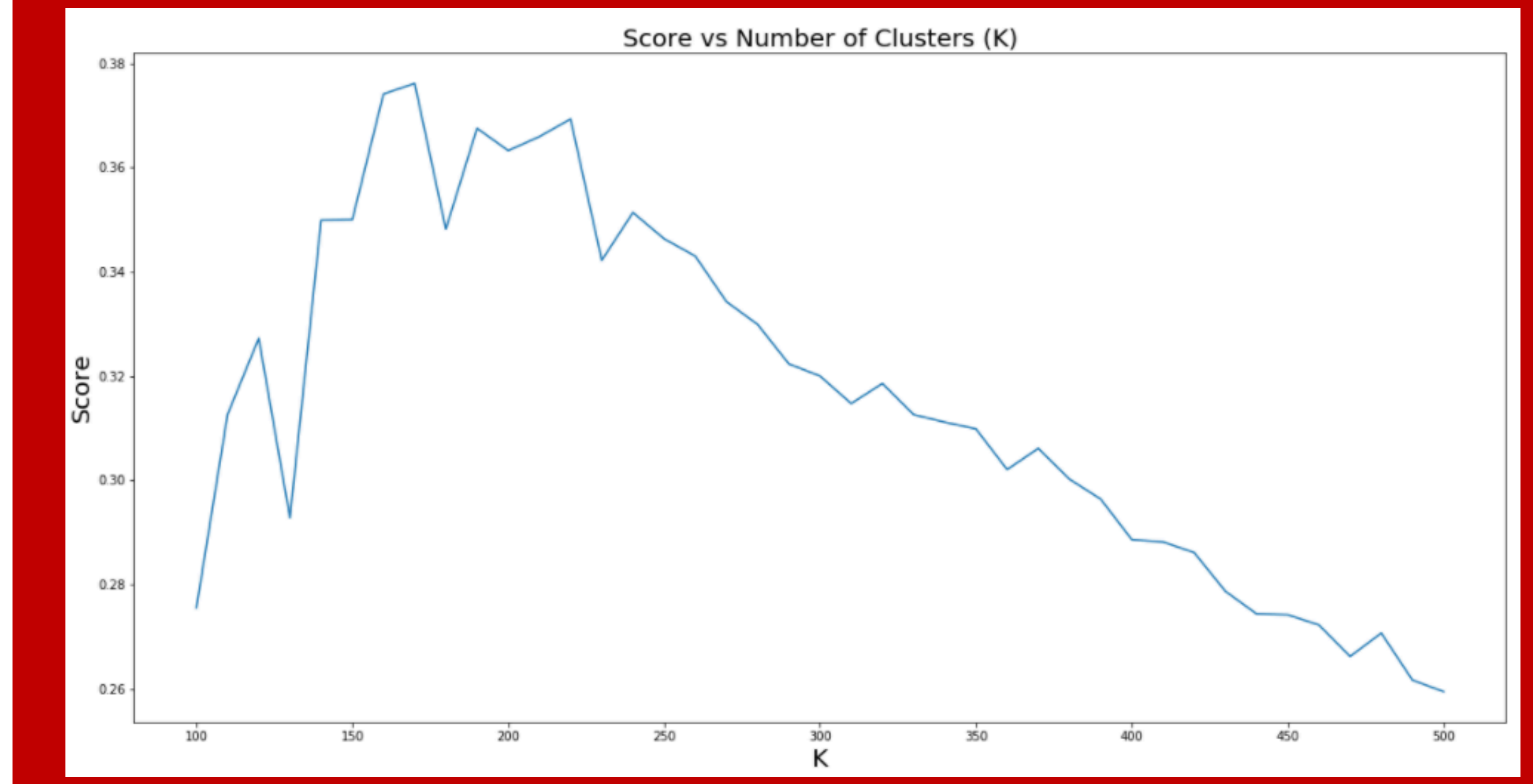
$$\|x_i - \bar{x}_k\|_2^2 \le \|x_i - \bar{x}_{k'}\|_2^2, \qquad \forall k' \neq k.$$

- The final clustering assignment is the solution to the following optimization problem:

$$\underset{C_1,\dots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$



Adjusted Mutual Info Score vs # Clusters

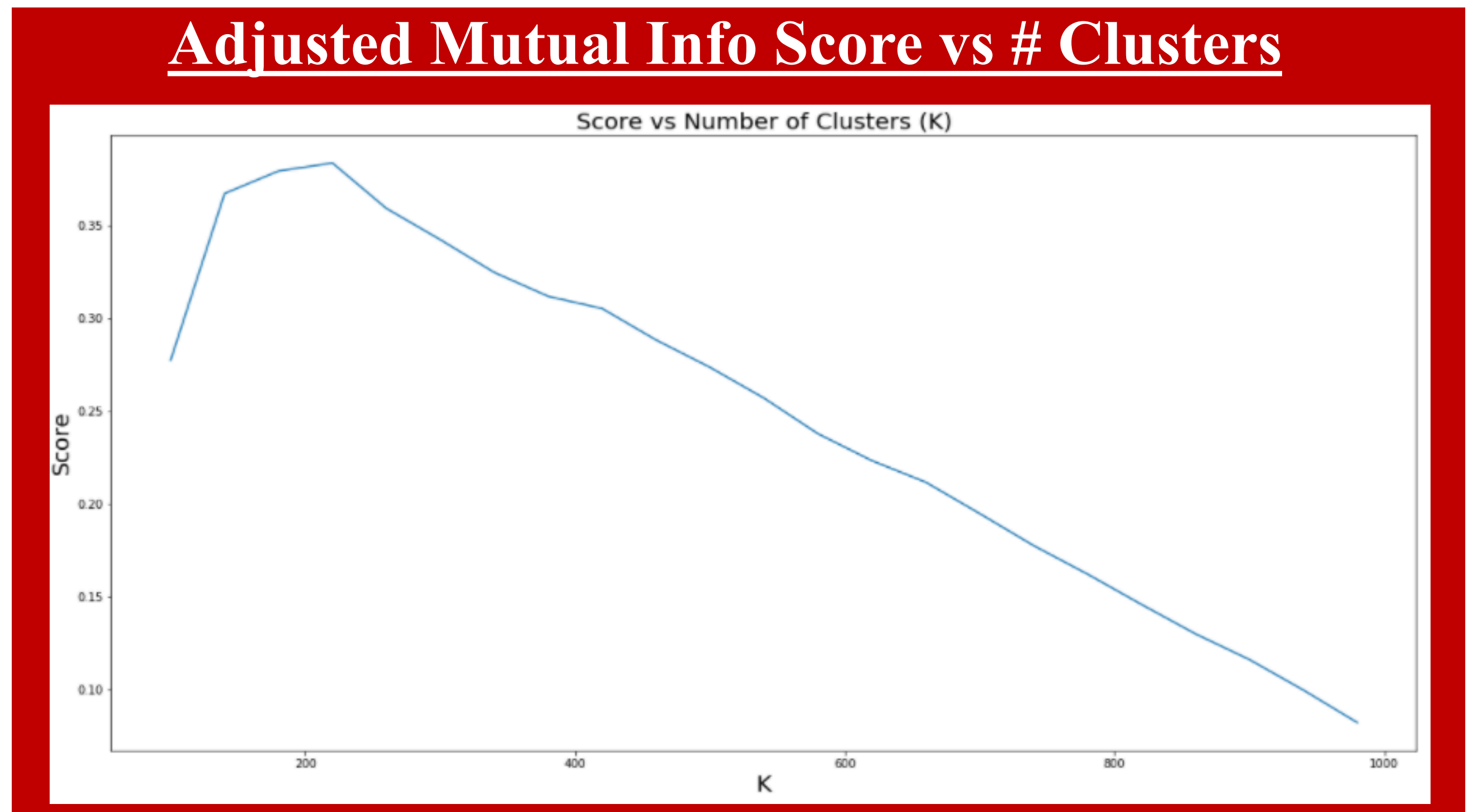


Optimal # Clusters: k = 160
Best AMI Score: 0.375

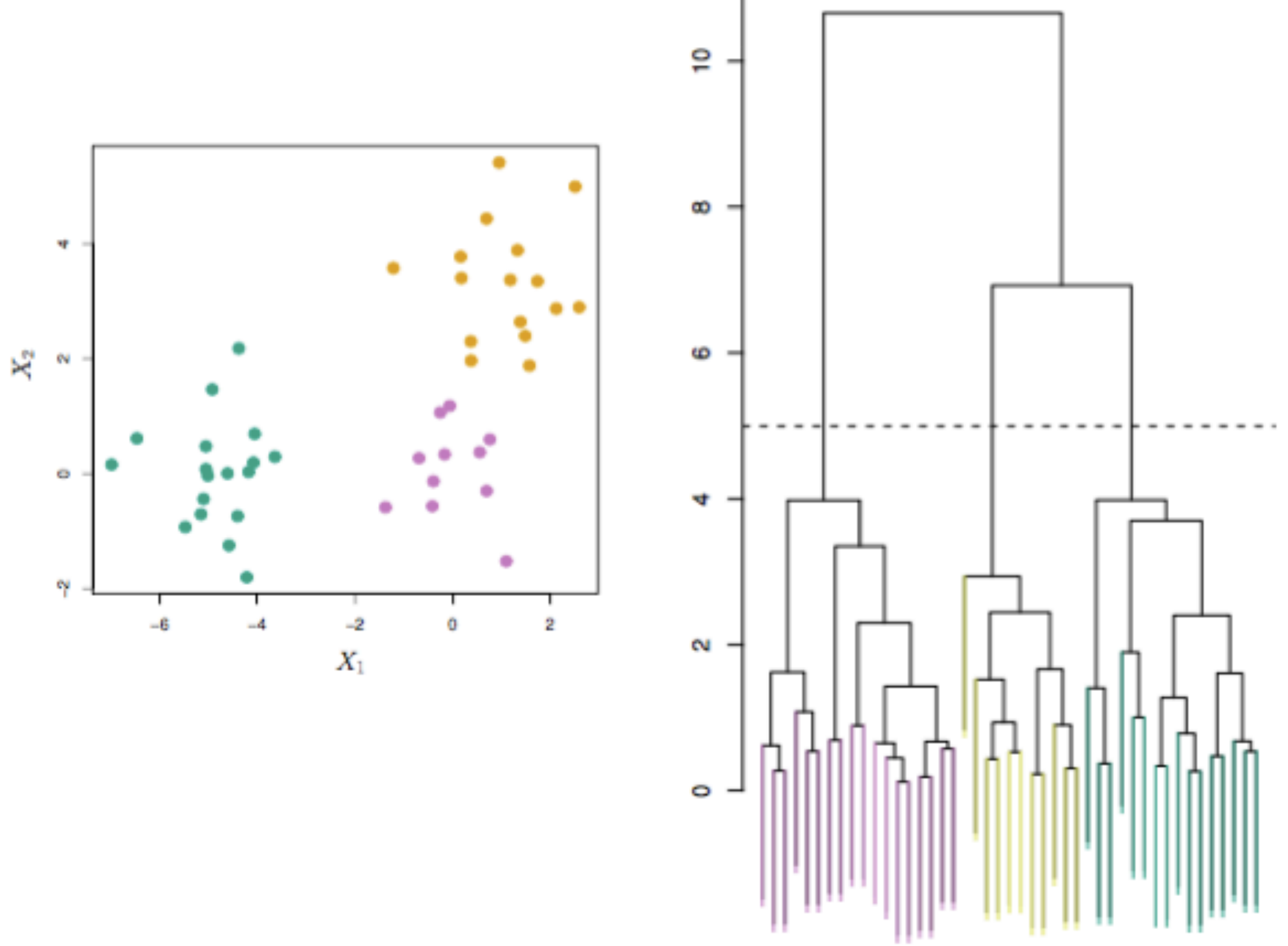
Agglomerative Hierarchical Clustering:

- Iteratively build larger clusters by combining smaller ones in ascending order of inter-cluster "distance"
  - ○ Single linkage: Use the min pairwise distance between two clusters as the distance
  - ○ Complete linkage: Use the max pairwise distance between two clusters as the distance
  - ○ Average linkage: Use the avg pairwise distance between all points in two clusters as the distance
- Represented by a tree whose leaves each represent a single observation; progressively merged as we move upwards

Adjusted Mutual Info Score vs # Clusters



Optimal # Clusters: k = 220
Best AMI Score: 0.384



- The mutual information metric unfairly penalizes partitions which might be valid, but which cannot be verified by the coarse granularity of the Bloomberg truth labels. Yet cluster purity alone cannot prevent # clusters from monotonically decreasing the classification error.
- K-Means assumes that the tickers' series make up globular clusters which span a roughly equal amount of dimensional space. But the ETFs might actually be skewedly distributed amongst the true industries/sectors, so mediocre performance was expected.
- Hierarchical clustering intuitively parallels the way in which sectors are made up of industries made up of subindustries made up of… made up of indiv stocks, yet performs only slightly better! Truth labels not granular enough, else could devise fairer metric that penalizes unnecessary splits less heavily.

Suggested future steps: (1) Get more granular truths (2) Use risk parity to convert clustering to portfolio (3) Evaluate clustered portfolio by comparing returns/risks vs S&P500