Lesson 29: K-Means Clustering

Data Science with Python – BSc Course

45 Minutes

Learning Objectives

The Problem: We have 500 stocks with dozens of features. How do we group similar stocks together without predefined categories? This is unsupervised learning.

After this lesson, you will be able to:

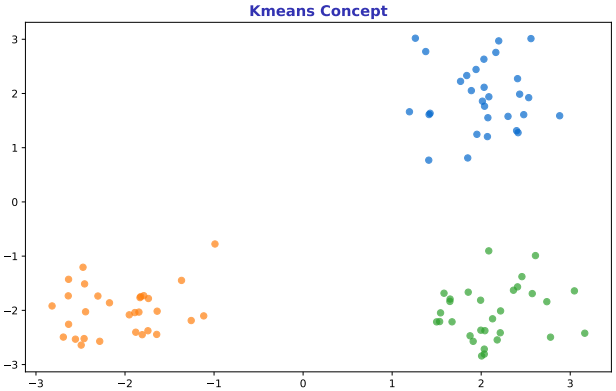
- Apply K-Means clustering algorithm
- Choose optimal K using elbow method and silhouette score
- Interpret cluster centers (centroids)
- Segment financial assets by behavior

Finance Application: Stock segmentation, customer clustering, regime detection

K-Means Concept

Partitioning Data into K Groups

- Goal: minimize within-cluster variance (inertia)
- Each point belongs to cluster with nearest centroid

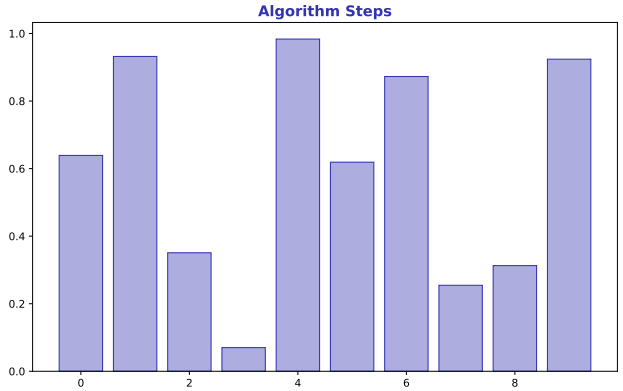


K-Means finds compact, spherical clusters – works well when clusters are well-separated

Algorithm Steps

Iterative Refinement

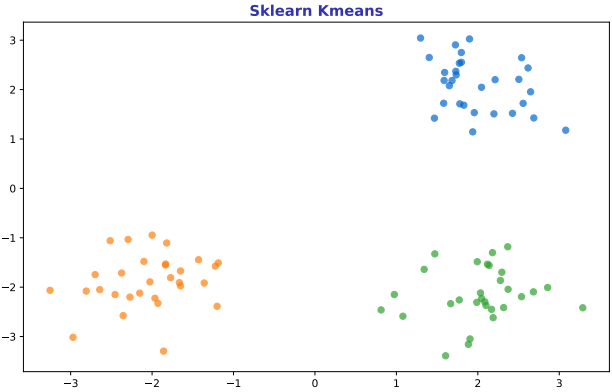
- Initialize: randomly place K centroids
- Repeat: (1) assign points to nearest centroid, (2) update centroids to cluster means



Convergence guaranteed but may find local minimum – use multiple random starts

Implementation in Python

- `from sklearn.cluster import KMeans`
- `kmeans = KMeans(n_clusters=K, n_init=10).fit(X)`

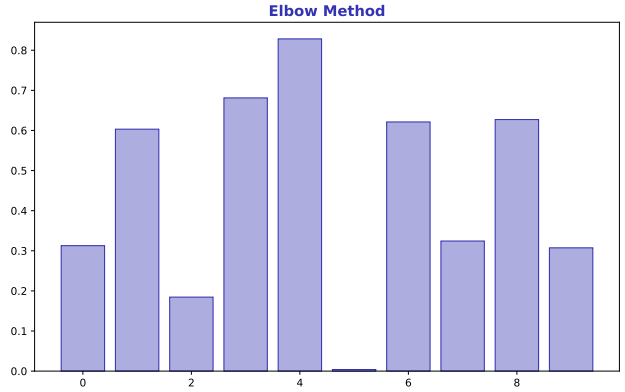


Access labels via `kmeans.labels_` **and centers via** `kmeans.cluster_centers_`

Elbow Method

How Many Clusters?

- Plot inertia vs K – look for "elbow" where improvement slows
- Inertia always decreases with K; find diminishing returns

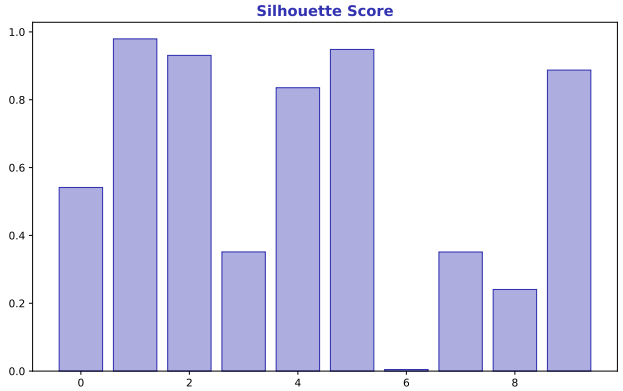


The elbow is subjective – combine with domain knowledge and silhouette score

Silhouette Score

Measuring Cluster Quality

- Range: -1 to 1. Higher = better-defined clusters
- Compares within-cluster distance to nearest-cluster distance

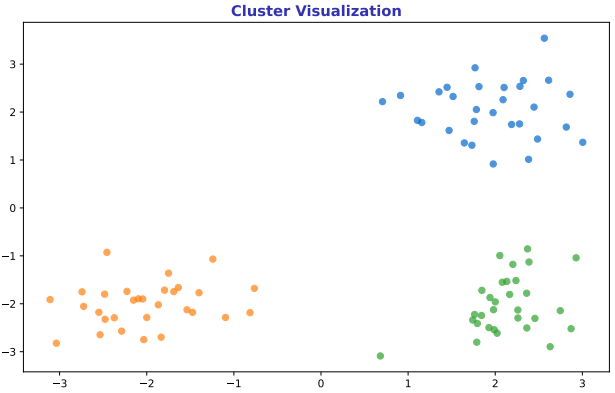


Silhouette > 0.5 indicates reasonable structure. Negative = probably wrong cluster.

Cluster Visualization

Seeing the Results

- 2D scatter plot with cluster colors
- For high-dimensional data: use PCA first, then plot

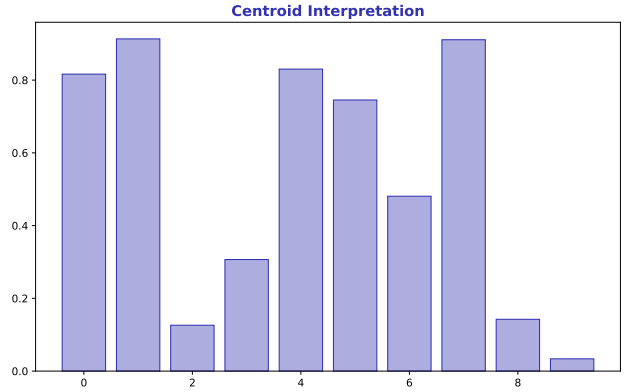


Always visualize clusters to sanity-check results

Centroid Interpretation

What Does Each Cluster Represent?

- Centroid = average feature values for that cluster
- Compare centroids to understand cluster characteristics

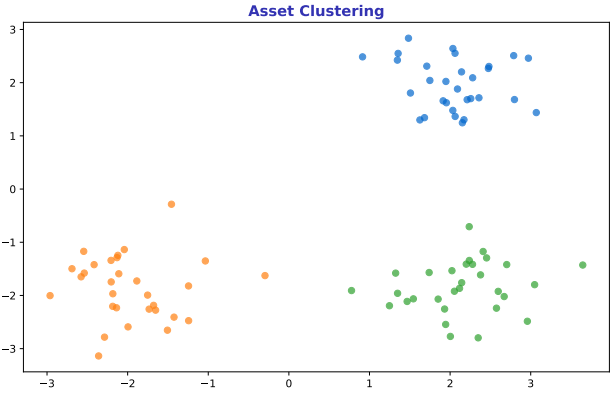


Example: Cluster 1 = high volatility, low volume; Cluster 2 = low volatility, high volume

Asset Clustering

Finance Application: Stock Segmentation

- Features: returns, volatility, beta, market cap, sector
- Clusters reveal natural groupings beyond traditional sectors



Use for diversification: select one stock per cluster for uncorrelated portfolio

Hands-On Exercise (25 min)

Task: Cluster Stocks by Behavior

1. Calculate features: 1-year return, volatility, beta for 50 stocks
2. Standardize features (important for K-Means!)
3. Run K-Means with K=2,3,4,5 – plot elbow curve
4. Choose best K using elbow + silhouette
5. Interpret centroids: what characterizes each cluster?

Deliverable: Elbow plot + scatter plot of clusters with labels.

Extension: Compare clusters to GICS sectors – do they align?

Lesson Summary

Problem Solved: We can now discover natural groupings in data without predefined labels.

Key Takeaways:

- K-Means: iteratively assign points to K cluster centers
- Choose K: elbow method + silhouette score
- Always standardize features before clustering
- Interpret centroids to understand cluster meaning

Next Lesson: Hierarchical Clustering (L30) – clusters within clusters

Memory: K-Means minimizes inertia. Elbow = where curve bends. Silhouette = cluster quality.