K-Means Clustering in Finance

Discovering Natural Groupings in Assets' Characteristics

MSc Banking & Finance

A.U.E.B.

October 22, 2025

The Fundamental Question

Observation

DeFi protocols vary across multiple dimensions:

- Risk profiles (volatility, drawdowns)
- Return characteristics (mean, Sharpe ratio)
- Liquidity (TVL, trading volume)
- Market cap and maturity

The Question

Can we **group similar tokens together** based on their characteristics?

Goal: Discover natural clusters for portfolio construction, risk management, and strategy design

K-Means Clustering: What & Why

What is K-Means?

Unsupervised learning algorithm that:

- Partitions data into k clusters
- Minimizes within-cluster variance
- Assigns each token to nearest cluster center

Objective Function:

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Where μ_i is cluster i centroid

Why Use K-Means?

- **Simple** and fast
- **Interpretable** results
- **Scalable** to many tokens
- Reveals **natural groupings**
- Useful for:
  - Portfolio diversification
  - Risk bucketing
  - Peer comparison
  - Strategy segmentation

The Dataset: 12 DeFi Tokens

Token	Category	Mean Ret (%)	Volatility (%)	TVL ($M)	Mkt Cap ($B)
UNI	DEX	3.2	65	4200	5.8
SUSHI	DEX	2.8	72	850	0.6
AAVE	Lending	2.5	58	6800	1.4
COMP	Lending	2.1	62	3100	0.8
MKR	Stablecoin	2.7	55	5200	1.2
CRV	Stableswap	1.8	48	3900	0.9
SNX	Derivatives	4.5	85	1200	0.7
LDO	Staking	5.2	78	9500	2.1
RPL	Staking	6.8	92	1800	0.5
GMX	Derivatives	7.5	98	650	0.4
DYDX	DEX	3.5	68	380	0.5
FXS	Stablecoin	2.9	61	780	0.6

Key Features for Clustering

We'll use: **Mean Return, Volatility, TVL, Market Cap**

K-Means Algorithm: Step-by-Step

The Process

1. **Choose** k (number of clusters) - we'll use $k = 3$
2. **Initialize:** Randomly select k tokens as initial centroids
3. **Assignment Step:** Assign each token to nearest centroid

$$C_i = \{x : ||x - \mu_i|| \leq ||x - \mu_j|| \text{ for all } j\}$$

4. **Update Step:** Recalculate centroids as mean of assigned tokens

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

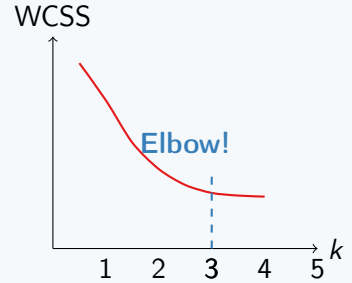
5. **Repeat** steps 3-4 until convergence (centroids don't change)

Typical convergence: 5-10 iterations

Choosing the Number of Clusters

Elbow Method

Plot within-cluster sum of squares (WCSS) vs k:



Elbow at $k = 3$ suggests 3 clusters

Silhouette Score: Measures cluster quality

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]$$

- $a(i)$: avg distance within cluster
- $b(i)$: avg distance to nearest cluster
- $s > 0.5$: good clustering

k	Silhouette
2	0.48
3	**0.62**
4	0.51
5	0.45

Best: $k = 3$

Results: Three Distinct Clusters

Token	Mean Ret	Volatility	TVL	Cluster
GMX	7.5%	98%	$650M	1: High Risk
RPL	6.8%	92%	$1,800M	1: High Risk
SNX	4.5%	85%	$1,200M	1: High Risk
LDO	5.2%	78%	$9,500M	1: High Risk
UNI	3.2%	65%	$4,200M	2: Medium Risk
SUSHI	2.8%	72%	$850M	2: Medium Risk
DYDX	3.5%	68%	$380M	2: Medium Risk
FXS	2.9%	61%	$780M	2: Medium Risk
AAVE	2.5%	58%	$6,800M	3: Low Risk
COMP	2.1%	62%	$3,100M	3: Low Risk
MKR	2.7%	55%	$5,200M	3: Low Risk
CRV	1.8%	48%	$3,900M	3: Low Risk

Cluster Profiles & Characteristics

Cluster 1: High Risk

Tokens: GMX, RPL, SNX, LDO

Profile:

- High return (5-7.5%)
- High volatility (78-98%)
- Smaller/emerging
- Derivatives & Staking

Strategy:

- Growth/aggressive
- Higher allocation in bull
- Tight stop-losses

Cluster 2: Medium Risk

Tokens: UNI, SUSHI, DYDX, FXS

Profile:

- Mod. return (2.8-3.5%)
- Moderate vol (61-72%)
- Established DEX
- Balanced risk/return

Strategy:

- Core holdings
- Market exposure
- Sector rotation plays

Cluster 3: Low Risk

Tokens: AAVE, COMP, MKR, CRV

Profile:

- Lower return (1.8-2.7%)
- Lower vol (48-62%)
- Large TVL
- Lending & Stables

Strategy:

- Defensive/stable
- Bear market hedge
- Yield generation

Practical Applications

1. Portfolio Construction: Cluster Diversification

Balanced Portfolio:

- 25% Cluster 1 (High Risk) - Growth exposure
- 50% Cluster 2 (Medium Risk) - Core holdings
- 25% Cluster 3 (Low Risk) - Stability

Result: True diversification across risk profiles, not just token count

2. Risk Management: Cluster Limits

Set exposure limits by cluster:

- Cluster 1 (High Risk): Max 30% of portfolio
- Cluster 2 (Medium Risk): 30-70%
- Cluster 3 (Low Risk): Min 15%

Prevents: Over-concentration in high-risk tokens

Key Takeaways & Conclusion

What We Learned

1. **K-Means reveals natural groupings** in DeFi tokens
   - 3 clusters: High, Medium, Low Risk
   - Clear differentiation by volatility and return
2. **Clusters align with economic intuition**
   - Derivatives/Staking = High Risk
   - DEX = Medium Risk
   - Lending/Stables = Low Risk
3. **Practical applications**
   - Cluster-based portfolio construction
   - Risk limits by cluster
   - Peer comparison within clusters
   - Dynamic rebalancing across clusters

Backup: Distance Metrics

Euclidean Distance (Default)

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Most common; assumes equal importance of features

Manhattan Distance

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Less sensitive to outliers

Always standardize before clustering, itnsures features on same scale:

$$z = (x - \mu)/\sigma$$

Backup: Limitations & Extensions

K-Means Limitations

- Assumes spherical clusters (equal variance)
- Sensitive to initialization
- Requires specifying k in advance
- Sensitive to outliers

Alternative Clustering Methods

- **Hierarchical:** Creates dendrogram, no need to specify k
- **DBSCAN:** Density-based, finds arbitrary shapes
- **Gaussian Mixture Models:** Probabilistic, allows soft assignments
- **Spectral Clustering:** Uses graph structure