EIU

TRƯỜNG ĐẠI HỌC
QUỐC TẾ
MIỀN ĐÔNG
EASTERN
INTERNATIONAL
UNIVERSITY

15 Year Anniversary 2010 - 2025

FINAL PROJECT

Clustering Analysis of Vietnamese Banks (Q3 - 2023)

Course: MIS 451 - Machine Learning for Business

Lecturer: Dang Thai Doan

Group: HKT

Student's Name	IRN
Trần Phương Thảo	2132300447
Phạm Tuấn Kiệt	2032300089
Thái Gia Huy	2132300332

Quarter 3, 2024-2025

TABLE OF CONTENT

I. Introduction

In the competitive and evolving landscape of Vietnam's banking industry, understanding the strategic positioning of banks is crucial. Clustering offers a data-driven approach to segment banks into meaningful groups based on their financial performance and operational characteristics.

1. Business Problem

Vietnamese banks vary significantly in scale, profitability, and operational focus. However, there is currently no standardized segmentation that helps regulators, investors, or the banks themselves understand peer groups or benchmark performance effectively.

2. Proposed Solution

To address this gap, we propose a clustering-based approach that groups Vietnamese banks based on key financial indicators. By identifying homogeneous segments, we can uncover patterns and offer tailored strategic recommendations.

3. Objective

The primary objective of this project is to apply clustering techniques—specifically K-Means and Hierarchical Clustering - to segment Vietnamese banks using financial data (KQKD metrics), and derive strategic insights from the resulting clusters.

II. Data Understanding and Preparation

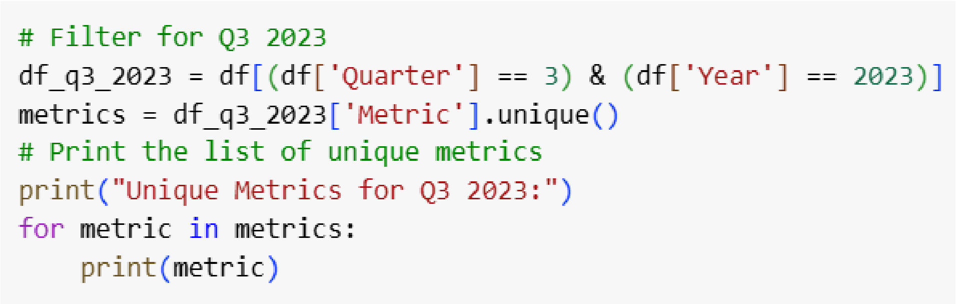
1. Data Overview

- **Columns (9 columns)**: included ID, TICKER, NAME, exchangeCode, dataType, Quarter, Year, Metric, and Value. This indicates a long format where each row represents a specific financial metric for a bank at a particular time.
- **Data Types**: Data types were generally appropriate, with the crucial Value column for financial metrics being a float64.

- **Missing Values**: A significant number of missing values, 27,741 (approximately 10.95%), were identified in the Value column. This needed to be addressed in subsequent preprocessing steps, particularly after data transformation.
- **Unique Values**: The dataset comprised data for 27 unique bank TICKERs across 3 exchange codes, covering 4 quarters over 17 years, with 195 distinct financial metrics. The financial metrics were noted to be in Vietnamese, covering balance sheets (CĐKT), income statements (KQKD), cash flow statements (LCTT), and other financial activities (TM).

2. Data Filtering and Selection

- **Time period filtering:** The dataset was filtered to retain only records where the Quarter was 3 and the Year was 2023. This focused the analysis on the specific period of interest.
- **Metric selection for Q3 2023:** After filtering by time period, the number of unique financial metrics available for Q3 2023 was 192.
- **Focus on income statement (KQKD) metrics:** For the clustering analysis, a decision was made to concentrate on Income Statement metrics (those prefixed with "KQKD."). This was based on the rationale that these metrics directly reflect a bank's profitability, operational efficiency, and overall financial performance, which are key aspects for strategic positioning. This selection resulted in 21 KQKD metrics being used as features for the clustering model for the 27 banks.
- **Missing values in selected features:** Importantly, after filtering for Q3 2023 and selecting the 21 KQKD metrics, an assessment of the kqkd_df DataFrame (containing these selected features for the 27 banks) showed no missing values. This was a positive outcome, as it eliminated the need for data imputation for the core features used in the clustering.

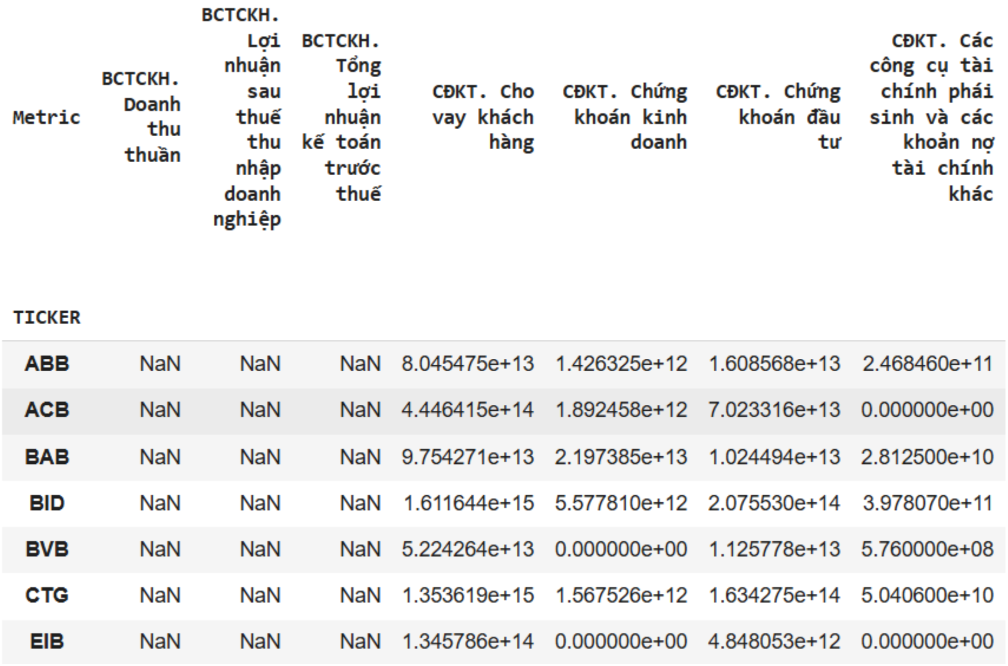


```
# Filter for Q3 2023
df_q3_2023 = df[(df['Quarter'] == 3) & (df['Year'] == 2023)]
metrics = df_q3_2023['Metric'].unique()
# Print the list of unique metrics
print("Unique Metrics for Q3 2023:")
for metric in metrics:
    print(metric)
```

3. Data Transformation (Pivoting)

The original long-format data needed to be restructured for clustering, where each bank is an observation and each financial metric is a feature:

- The filtered Q3 2023 data (specifically the subset containing the KQKD metrics) was pivoted. The bank TICKER was set as the index (rows), the Metric names became the columns (features), and the corresponding Value filled the cells. This transformation resulted in a DataFrame where each row represented a bank and each column represented one of the 21 selected KQKD financial metrics.

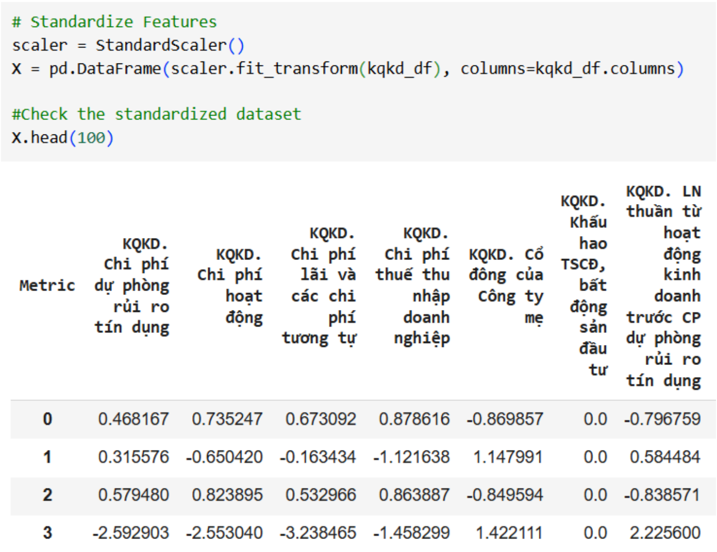


Metric	BCTCKH. Doanh thu thuần	BCTCKH. Lợi nhuận sau thuế thu nhập doanh nghiệp	BCTCKH. Tổng lợi nhuận kế toán trước thuế	CĐKT. Cho vay khách hàng	CĐKT. Chứng khoán kinh doanh	CĐKT. Chứng khoán đầu tư	CĐKT. Các công cụ tài chính phái sinh và các khoản nợ tài chính khác
TICKER							
ABB	NaN	NaN	NaN	8.045475e+13	1.426325e+12	1.608568e+13	2.468460e+11
ACB	NaN	NaN	NaN	4.446415e+14	1.892458e+12	7.023316e+13	0.000000e+00
BAB	NaN	NaN	NaN	9.754271e+13	2.197385e+13	1.024494e+13	2.812500e+10
BID	NaN	NaN	NaN	1.611644e+15	5.577810e+12	2.075530e+14	3.978070e+11
BVB	NaN	NaN	NaN	5.224264e+13	0.000000e+00	1.125778e+13	5.760000e+08
CTG	NaN	NaN	NaN	1.353619e+15	1.567526e+12	1.634275e+14	5.040600e+10
EIB	NaN	NaN	NaN	1.345786e+14	0.000000e+00	4.848053e+12	0.000000e+00

4. Feature Standardization

To ensure that all financial metrics contribute equally to the clustering process, regardless of their original scale or units, feature standardization was applied:

- The StandardScaler from sklearn.preprocessing was used to standardize the 21 KQKD financial metrics. This process transforms the data by subtracting the mean and dividing by the standard deviation for each feature, resulting in features with a mean of 0 and a standard deviation of 1. This is a standard requirement for distance-based algorithms like K-Means to prevent features with larger values from dominating the clustering outcome.



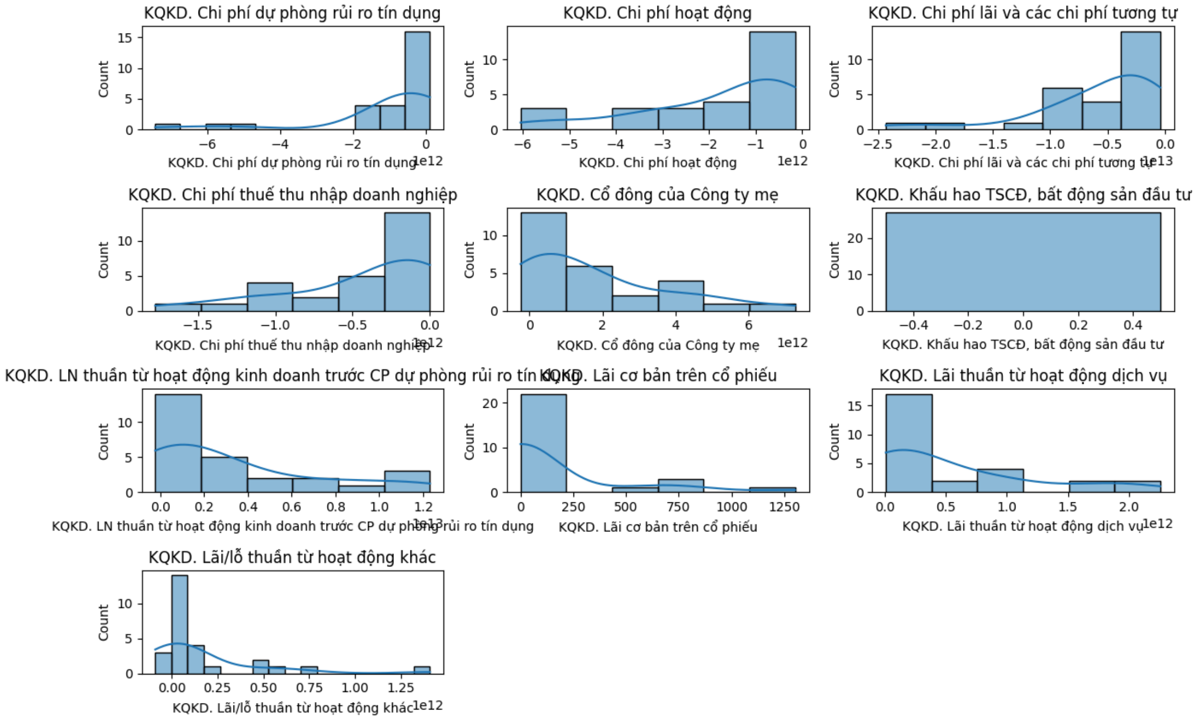
```
# Standardize Features
scaler = StandardScaler()
X = pd.DataFrame(scaler.fit_transform(kqkd_df), columns=kqkd_df.columns)

#Check the standardized dataset
X.head(100)
```

Metric	KQKD. Chi phí dự phòng rủi ro tín dụng	KQKD. Chi phí hoạt động	KQKD. Chi phí lãi và các chi phí tương tự	KQKD. Chi phí thuế thu nhập doanh nghiệp	KQKD. Cổ đông của Công ty mẹ	KQKD. Khấu hao TSCĐ, bất động sản đầu tư	KQKD. LN thuần từ hoạt động kinh doanh trước CP dự phòng rủi ro tín dụng
0	0.468167	0.735247	0.673092	0.878616	-0.869857	0.0	-0.796759
1	0.315576	-0.650420	-0.163434	-1.121638	1.147991	0.0	0.584484
2	0.579480	0.823895	0.532966	0.863887	-0.849594	0.0	-0.838571
3	-2.592903	-2.553040	-3.238465	-1.458299	1.422111	0.0	2.225600

5. Exploratory Data Analysis (EDA) Visualization

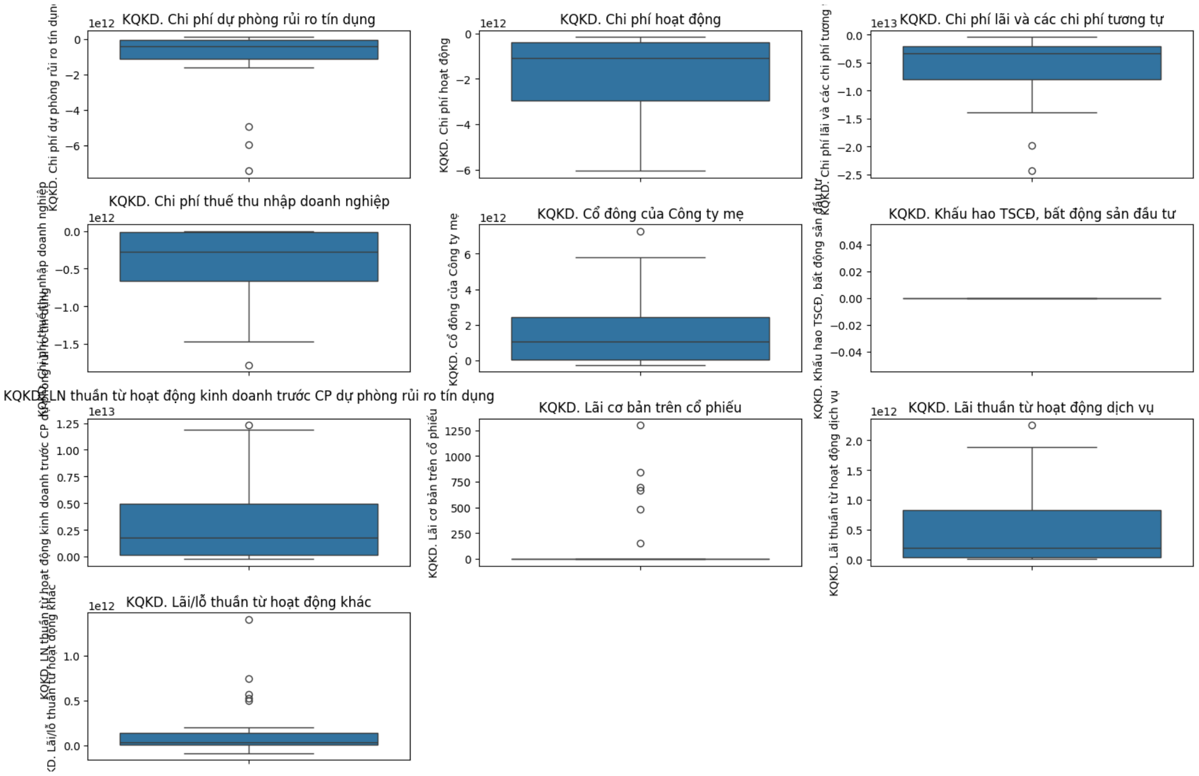
5.1. Feature Distribution Analysis



- Several cost-related metrics, such as KQKD. Chi phí dự phòng rủi ro tín dụng (Cost of Credit Loss Provision), KQKD. Chi phí hoạt động (Operating Costs), and KQKD. Chi phí lãi và các chi phí tương tự (Interest and Similar Expenses), exhibited left-skewed distributions.
- KQKD. Chi phí thuế thu nhập doanh nghiệp (Corporate Income Tax Expense) was also left-skewed, with most values concentrated near zero or slightly negative.

- Profit-related metrics like KQKD. Cổ đông của Công ty mẹ (Profit Attributable to Parent Company Shareholders) and KQKD. LN thuần từ hoạt động kinh doanh trước CP dự phòng rủi ro tín dụng (Net Profit from Business Operations before Credit Loss Provision) generally showed right-skewed distributions.
- KQKD. Lãi cơ bản trên cổ phiếu (Basic Earnings Per Share) was highly right-skewed.
- Notably, the metric KQKD. Khấu hao TSCĐ, bất động sản đầu tư (Depreciation of Fixed Assets and Investment Properties) showed all values at zero for the selected banks in Q3 2023, indicating no variance for this feature in the dataset.

5.2. Outlier Detection Analysis



- Outliers were observed in several metrics. For instance, KQKD. Chi phí dự phòng rủi ro tín dụng, KQKD. Chi phí hoạt động, and KQKD. Chi phí lãi và các chi phí tương tự showed a few outliers on the lower end (more negative, indicating higher expenses or provisions for some banks).
- KQKD. Cổ đông của Công ty mẹ (Profit Attributable to Parent Company Shareholders) and KQKD. Lãi cơ bản trên cổ phiếu exhibited several outliers on the higher end, indicating some banks had exceptionally high profits or EPS.
- KQKD. LN thuần từ hoạt động kinh doanh trước CP dự phòng rủi ro tín dụng and KQKD. Lãi/lỗ thuần từ hoạt động khác (Net Gain/Loss from Other Activities) showed potential outliers on both the lower and higher ends.

- The boxplot for KQKD. Khấu hao TSCĐ, bất động sản đầu tư confirmed the lack of variance, appearing as a single line at zero.

III. Clustering Methodology

To segment the Vietnamese banks based on their Q3 2023 income statement (KQKD) profiles, K-Means clustering was primarily employed. Principal Component Analysis (PCA) was subsequently used to aid in the visualization of these clusters.

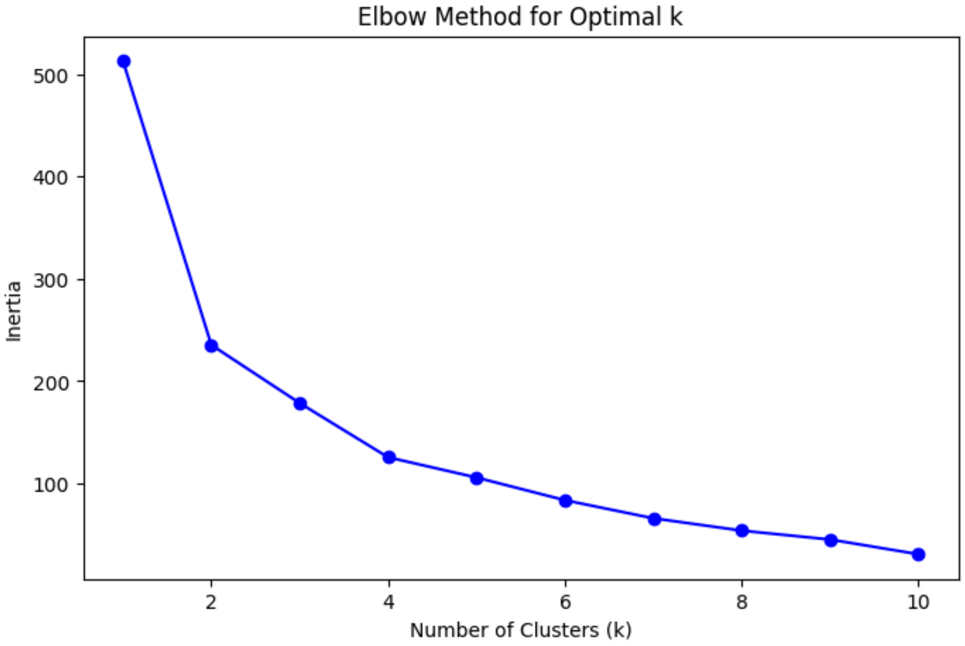
1. K-Means Clustering

The K-Means algorithm is a popular partitioning method that aims to group data points into a pre-defined number of clusters (k) such that the within-cluster sum of squares (inertia) is minimized.

1.1. Determining the Optimal Number of Clusters (k)

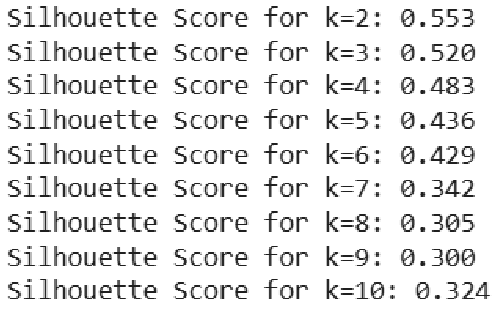
- Elbow Method (Inertia vs. k).

The K-Means algorithm was executed for k values ranging from 1 to 10. The inertia, which measures the within-cluster sum of squares (WCSS), was calculated for each k. The Elbow Curve plots inertia against the number of clusters. A distinct "elbow" or bend in the plot can indicate an optimal k, as adding more clusters beyond this point yields diminishing returns in reducing inertia. Observing the plot, a noticeable bend occurs at k=2, and another, less pronounced one, around k=3 or k=4.



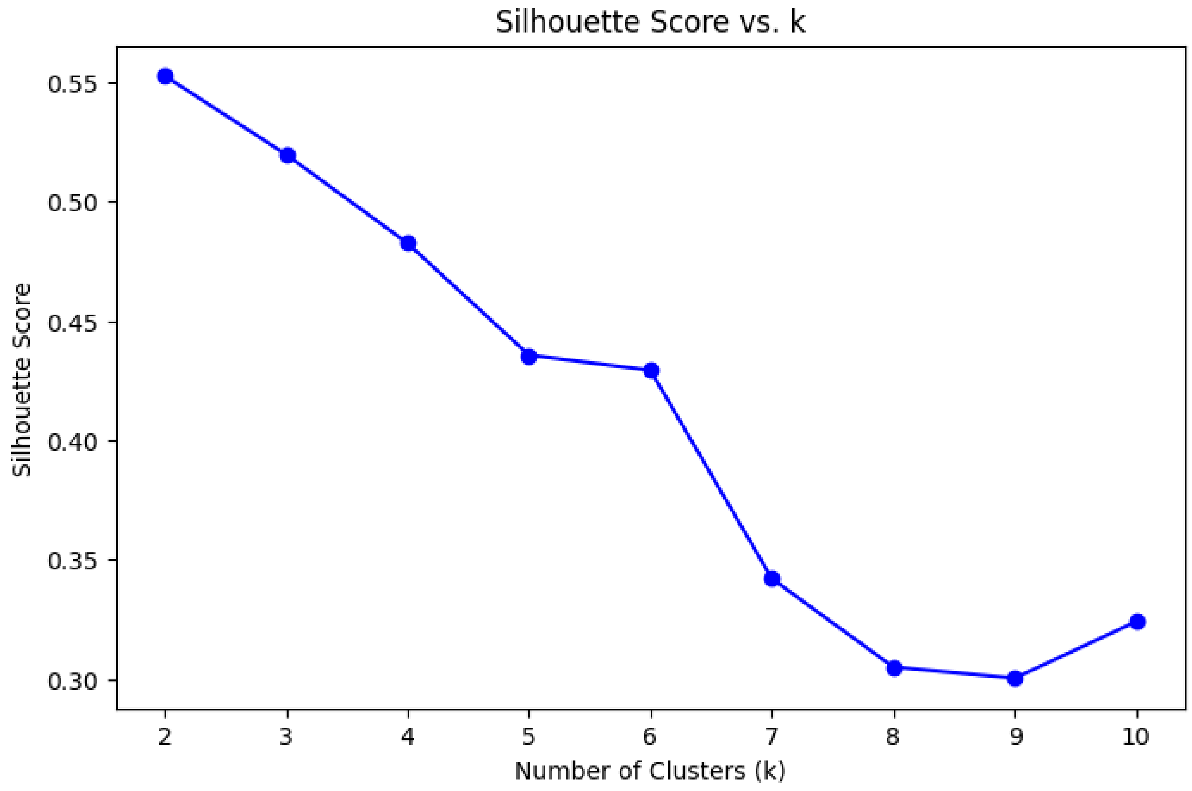
- Silhouette Score Analysis.

The Silhouette Score was calculated for k values from 2 to 10 to measure how well-separated the resulting clusters were. The score ranges from -1 to 1, with values closer to 1 indicating better-defined clusters.



```
Silhouette Score for k=2: 0.553
Silhouette Score for k=3: 0.520
Silhouette Score for k=4: 0.483
Silhouette Score for k=5: 0.436
Silhouette Score for k=6: 0.429
Silhouette Score for k=7: 0.342
Silhouette Score for k=8: 0.305
Silhouette Score for k=9: 0.300
Silhouette Score for k=10: 0.324
```

The highest Silhouette Score was observed at k=2, suggesting it as a strong candidate. The score for k=3 was the second highest.



=> While the Elbow Method showed a prominent elbow at k=2 and the Silhouette Score was also highest for k=2, a value of k=3 was chosen for the final K-Means model.

1.2. K-Means Model Application and Evaluation

The K-Means clustering algorithm was applied to the standardized KQKD data (X) with the chosen k=3 clusters.

Evaluation Metrics for k=3:

- Silhouette Score: 0.520
- Inertia (Within-Cluster Sum of Squares): 178.56

The Silhouette Score of 0.520 indicates that the clusters are reasonably well-defined and distinct.

1.3. Principal Component Analysis (PCA) for K-Means

To aid in understanding the feature space and for later visualization of the clusters, Principal Component Analysis (PCA) was performed on the standardized KQKD feature set (X).

- Application of PCA for Dimensionality Reduction: PCA was initially applied to determine the number of components needed to capture a significant portion of the variance in the data.
- Variance Explained: The analysis showed that the first principal component (PC1) accounted for approximately 68.45% of the variance, PC2 for 9.78%, PC3 for 8.56%, PC4 for 5.26%, and PC5 for 3.15%. Cumulatively, 5 principal components were found to explain approximately 95.19% of the total variance in the 21 KQKD features.

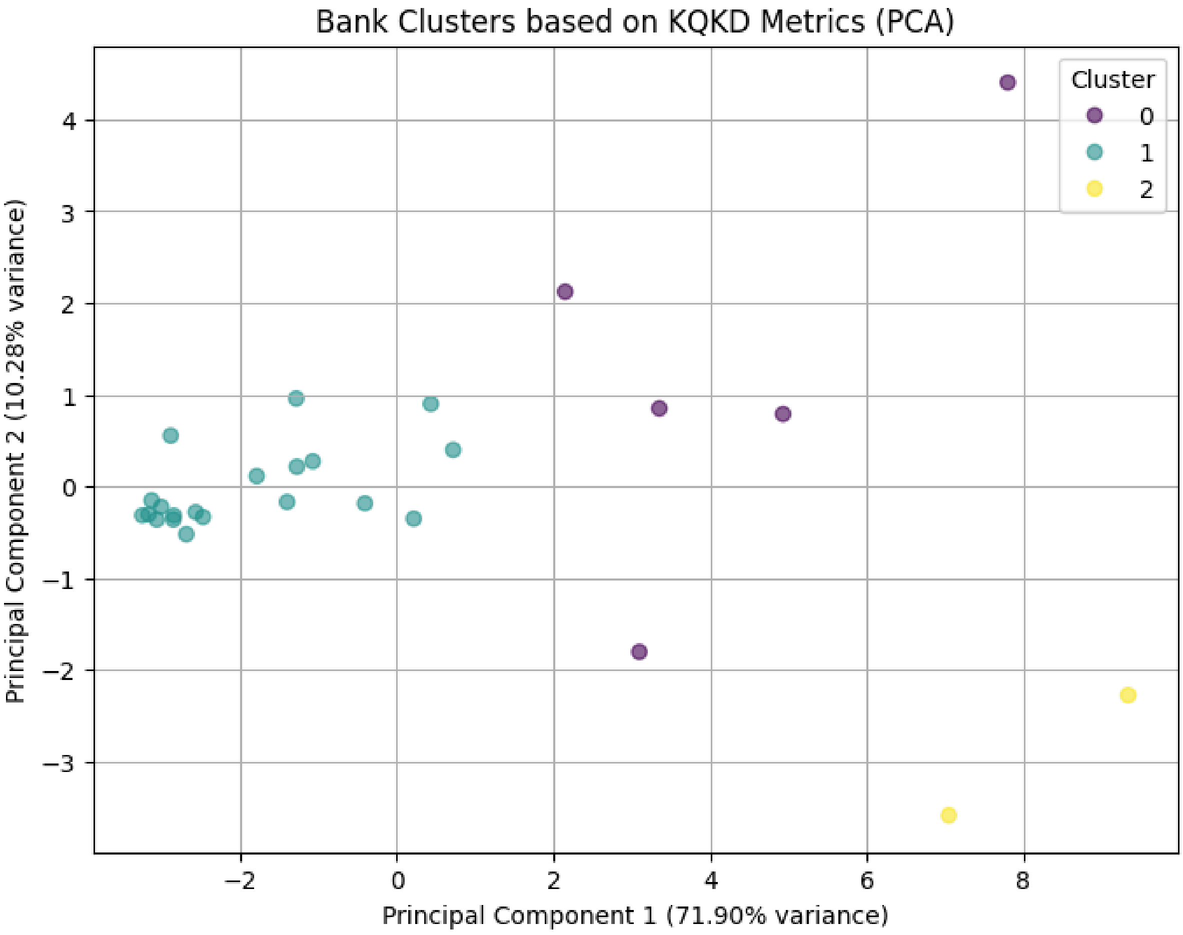
1.4. K-Means Cluster Visualization

Following the determination of k=3 as the optimal number of clusters, the K-Means algorithm was applied to the standardized KQKD (Income Statement) data for the 27 Vietnamese banks.

Distribution of Banks into Clusters:

The 27 banks were segmented into 3 distinct clusters. The distribution of banks within these clusters is as follows:

- Cluster 0: 5 banks
- Cluster 1: 20 banks
- Cluster 2: 2 banks



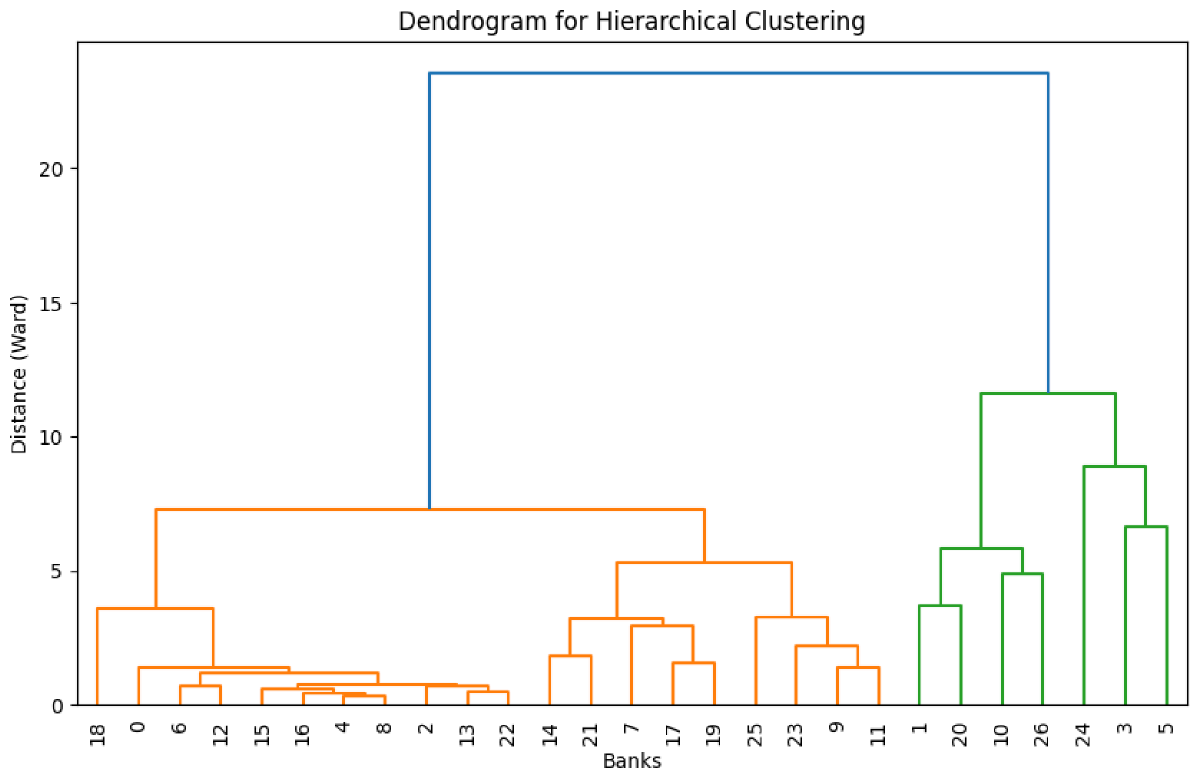
2. Hierarchical Clustering

In addition to K-Means, Hierarchical Clustering was performed as an alternative method to segment the Vietnamese banks based on their standardized Q3 2023 KQKD metrics.

2.1. Dendrogram Analysis

Hierarchical clustering was conducted using the 'ward' linkage method with the 'euclidean' distance metric on the standardized feature set (X). The 'ward' method aims to minimize the variance within each cluster.

A dendrogram was generated to visualize the hierarchical structure of the clusters and to assist in determining an appropriate number of clusters.



- The dendrogram illustrates how individual banks (or small clusters) are successively merged based on similarity.
- Visually inspecting the dendrogram, cutting the tree where there are significant vertical distances between merges can suggest a natural number of clusters. For example, a horizontal cut across the dendrogram at a distance (Ward) of approximately 10-15 would suggest the formation of 2 or 3 major clusters.

2.2. Hierarchical Cluster Formation

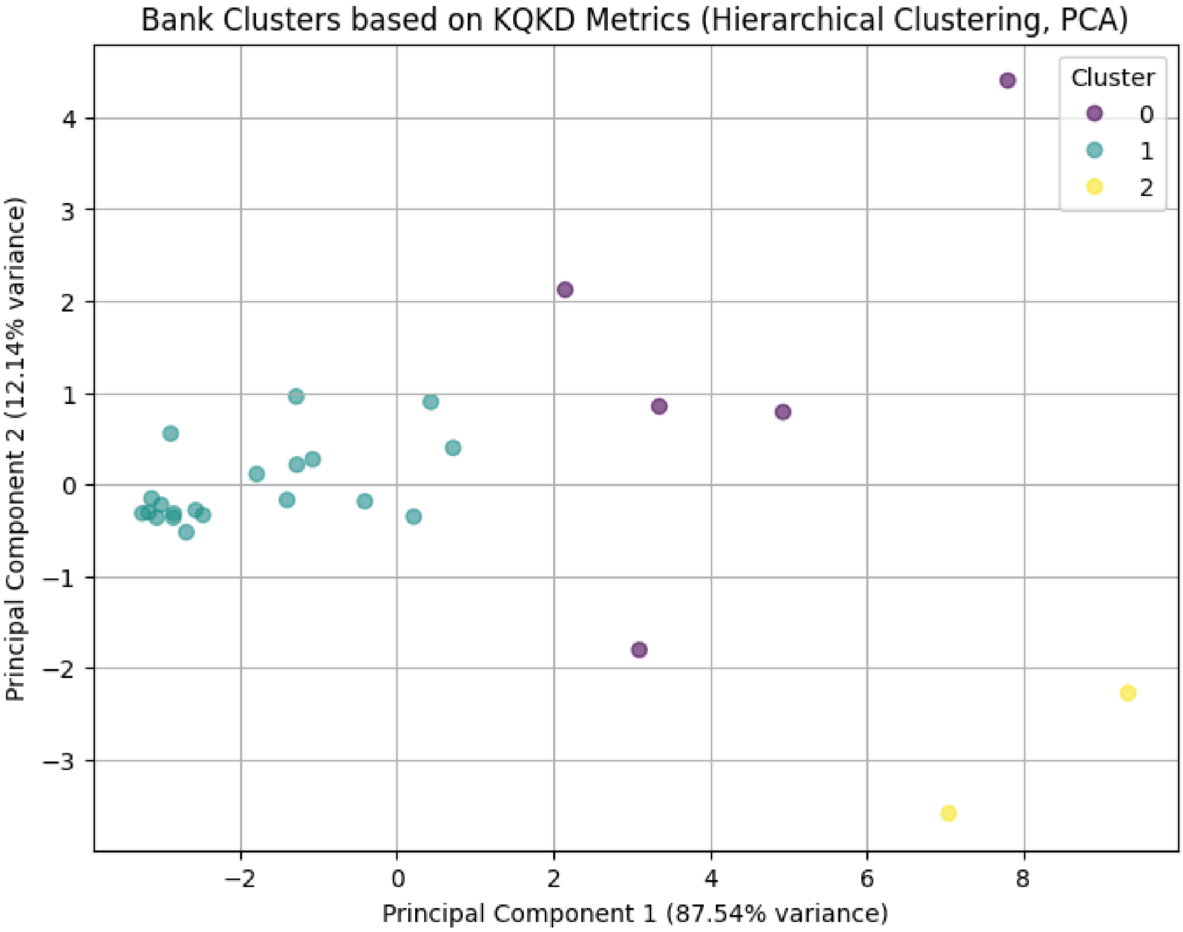
Based on the dendrogram analysis and aiming for a comparable number of segments as with K-Means for consistency, 3 clusters were formally extracted from the linkage matrix.

2.3. Hierarchical Cluster Assignment and Visualization

The 27 banks were assigned to one of the 3 hierarchical clusters.

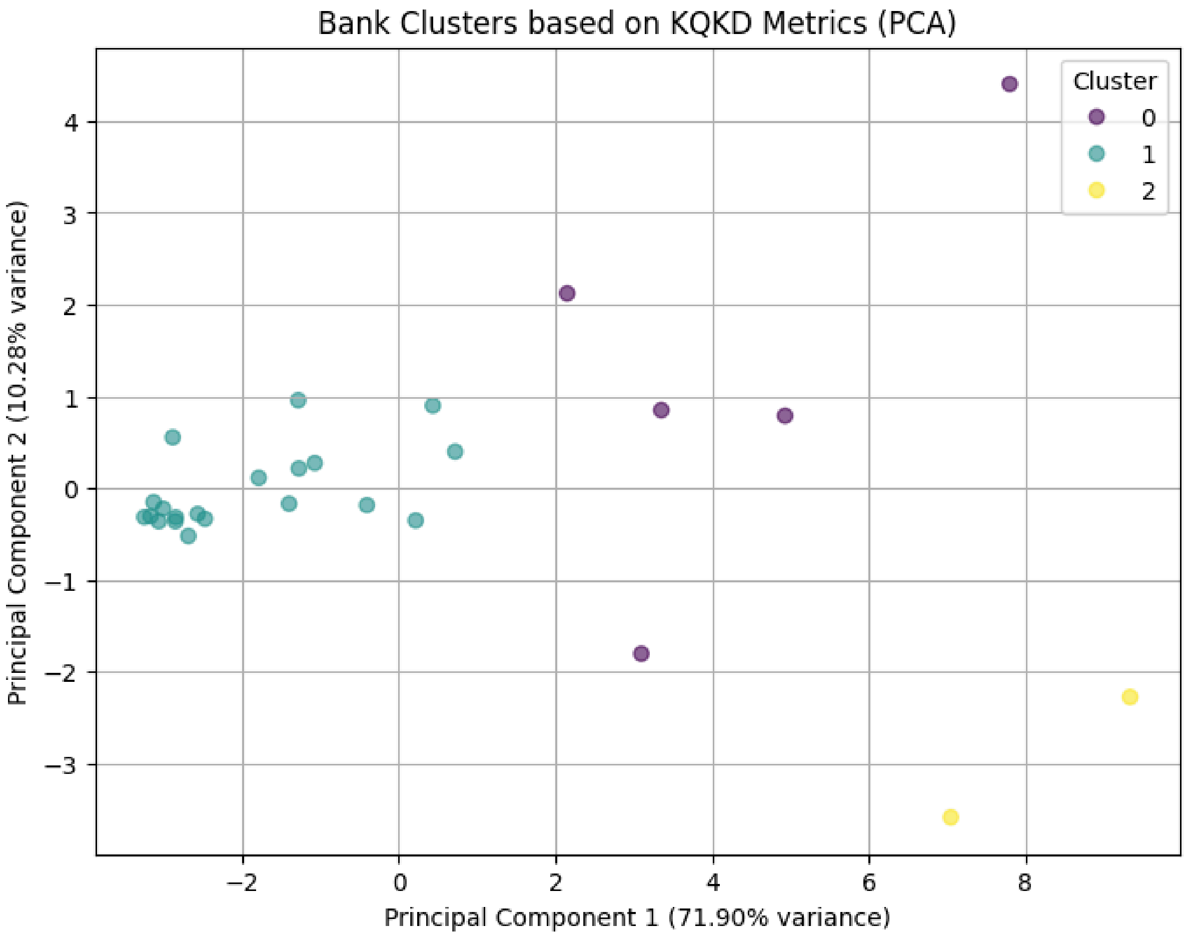
Distribution of Banks:

- Cluster 0: 20 banks
- Cluster 1: 4 banks
- Cluster 2: 3 banks


Bank Clusters based on KQKD Metrics (Hierarchical Clustering, PCA)
Principal Component 2 (12.14% variance)
4
3
2
1
0
−1
−2
−3
Principal Component 1 (87.54% variance)
−2
0
2
4
6
8
Cluster
0
1
2

IV. Cluster Profiles and Interpretation

1. K-Means Cluster Profiles



Cluster 0: High Profitability & Income Leaders

Number of Banks: 5

Banks: ACB, MBB, TCB, VCB, VPB

PCA Visualization Profile: These banks (purple points) are generally located towards the center-right and upper-right of the PCA plot.

Financial Profile (based on mean KQKD metrics):

- This cluster demonstrates the highest average KQKD. Lợi nhuận sau thuế thu nhập doanh nghiệp (Profit After Tax) at approximately 4.82e+12.
- It exhibits very high KQKD. Tổng thu nhập hoạt động (Total Operating Income) (average 1.16e+13) and KQKD. Thu nhập lãi thuần (Net Interest Income) (average 8.95e+12), indicating strong core income generation.

- Their KQKD. Chi phí hoạt động (Operating Costs) are moderate (average -3.65e+12, costs are negative in the data), suggesting relatively good cost management for their income level.
- The KQKD. Chi phí dự phòng rủi ro tín dụng (Credit Loss Provision Cost) is also at a moderate level (average -1.87e+12).
- This group can be characterized as high-performing banks with leading profitability and strong income generation capabilities.

Cluster 1: Moderate Performers / Smaller Operations

Number of Banks: 20

Banks: ABB, BAB, BVB, EIB, HDB, KLB, LPB, MSB, NAB, NVB, OCB, PGB, SGB, SHB, SSB, STB, TPB, VAB, VBB, VIB

PCA Visualization Profile (image_7b8f17.png): This is the largest cluster (teal points) and occupies the left side of the PCA plot, showing a greater spread but generally distinct from the other two.

Financial Profile (based on mean KQKD metrics):

- This cluster shows the lowest average KQKD. Lợi nhuận sau thuế thu nhập doanh nghiệp (Profit After Tax) (average 7.40e+11).
- It also has the lowest average KQKD. Tổng thu nhập hoạt động (Total Operating Income) (2.22e+12) and KQKD. Thu nhập lãi thuần (Net Interest Income) (1.76e+12), suggesting smaller operational scale or lower income efficiency compared to other clusters.
- Their KQKD. Chi phí hoạt động (Operating Costs) are the lowest in magnitude (average -9.37e+11), which is consistent with smaller operations.
- Similarly, their KQKD. Chi phí dự phòng rủi ro tín dụng (Credit Loss Provision Cost) is the lowest in magnitude (average -3.61e+11).
- This group likely represents banks with smaller operational scales or more moderate financial performance across key income and profitability metrics.

Cluster 2: High Income, High Cost & Provision Leaders

Number of Banks: 2

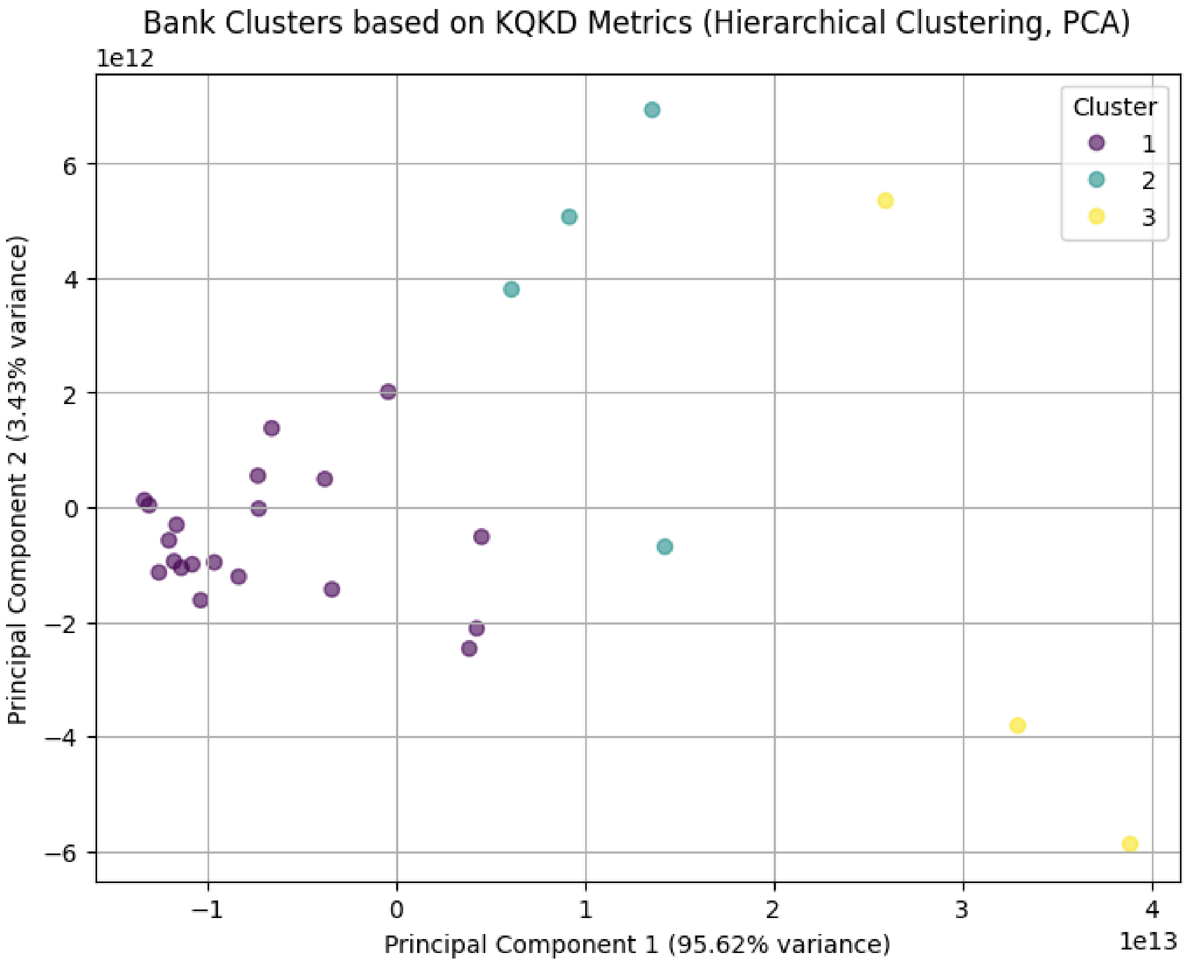
Banks: BID, CTG

PCA Visualization Profile (image_7b8f17.png): These two banks (yellow points) are distinctly separated on the lower right of the PCA plot, showing very high PC1 values but negative PC2 values.

Financial Profile (based on mean KQKD metrics):

- This cluster has a very high KQKD. Lợi nhuận sau thuế thu nhập doanh nghiệp (Profit After Tax) (average 4.31e+12), second only to Cluster 0.
- Notably, it reports the highest average KQKD. Tổng thu nhập hoạt động (Total Operating Income) (1.76e+13) and the highest KQKD. Thu nhập lãi thuần (Net Interest Income) (1.34e+13) among all clusters, indicating very large-scale income generation.
- However, it also incurs the highest KQKD. Chi phí hoạt động (Operating Costs) (average -5.56e+12) and by far the highest KQKD. Chi phí dự phòng rủi ro tín dụng (Credit Loss Provision Cost) (average -6.69e+12).
- This small cluster represents very large banks with exceptionally high income, but also the highest operating costs and credit provisions. They appear to be managing larger risks or are in a phase requiring significant provisioning.

2. Hierarchical Cluster Profiles



Hierarchical Cluster 0 (20 banks): ABB, BAB, BVB, EIB, HDB, KLB, LPB, MSB, NAB, NVB, OCB, PGB, SGB, SHB, SSB, STB, TPB, VAB, VBB, VIB

Hierarchical Cluster 1 (4 banks): ACB, MBB, TCB, VPB

Hierarchical Cluster 2 (3 banks): BID, CTG, VCB

3. Comparison of Clustering Methods

Both K-Means and Hierarchical clustering were applied to segment the 27 Vietnamese banks based on their Q3 2023 KQKD (Income Statement) metrics. While both aimed to group similar banks, their underlying algorithms and the resulting visualizations showed differences.

The reason is: K-Means partitions data based on distance to centroids, aiming for spherical clusters. Hierarchical clustering (Ward's) builds a hierarchy by merging clusters that result in the minimum increase in within-cluster variance.

V. Discussion and Strategic Recommendations

1. Identification of Peer Groups

- **Cluster 0 (High Profitability & Income Leaders):** Comprising 5 banks (ACB, MBB, TCB, VCB, VPB), this group is characterized by leading average Profit After Tax, robust Total Operating Income and Net Interest Income, with moderate operating costs and credit loss provisions.
- **Cluster 1 (Moderate Performers / Smaller Operations):** This largest group includes 20 banks (ABB, BAB, BVB, EIB, HDB, KLB, LPB, MSB, NAB, NVB, OCB, PGB, SGB, SHB, SSB, STB, TPB, VAB, VBB, VIB). They typically show lower average profitability and income figures, matched with the lowest magnitude of operating costs and provisions, indicative of smaller operational scales or more moderate overall performance.
- **Cluster 2 (High Income, High Cost & Provision Leaders):** This distinct, small cluster of 2 banks (BID, CTG) reports the highest Total Operating Income and Net Interest Income. While their Profit After Tax is very high (second highest), they also face the highest Operating Costs and significantly the highest Credit Loss Provisions.

2. Inferred Risk Profiles (based on KQKD performance)

Cluster 0 (High Profitability & Income Leaders):

- Risk Profile: Generally lower immediate financial risk due to strong profitability and substantial income generation capabilities. Their moderate credit loss provisions suggest effective risk management relative to their operations.
- Potential Risks: Risks could include maintaining high growth and profitability in a competitive market, managing customer retention, and potential challenges if expanding into new, less familiar business areas.

Cluster 1 (Moderate Performers / Smaller Operations):

- Risk Profile: Moderate financial risk. Their lower average profitability and income levels might make them more vulnerable to economic fluctuations or aggressive competition.
- Potential Risks: Difficulty in achieving economies of scale, potential capital constraints for significant technological upgrades or expansion, and risk of being outpaced by larger or more specialized competitors. Their low provisions, while aligned with lower income, need careful monitoring to ensure adequacy if asset quality changes.

Cluster 2 (High Income, High Cost & Provision Leaders):

- Risk Profile: Higher and more complex risk profile. While demonstrating the highest income generation capacity, this is counterbalanced by the highest operating costs and exceptionally high credit loss provisions.
- Potential Risks: Significant underlying credit risk within their loan portfolios is evident. High operating costs could indicate operational inefficiencies. Their net profitability, though substantial, is sensitive to income volatility or further deterioration in asset quality, which would necessitate even higher provisions. These banks might be dealing with legacy asset quality issues or operate in segments requiring higher risk provisioning.

3. Strategic Recommendations

Cluster 0 (High Profitability & Income Leaders - ACB, MBB, TCB, VCB, VPB):

- Sustain Excellence & Drive Innovation: Focus on leveraging their strong market position and brand by investing in financial technology (FinTech), enhancing digital customer experiences, and developing innovative, high-margin products and services.
- Strategic Growth & Diversification: Prudently explore opportunities for organic and inorganic growth, possibly diversifying into related financial services like wealth management or specialized corporate advisory, while maintaining rigorous risk assessment.
- Talent Management & Efficiency: Continue to invest in attracting and retaining top talent. Further optimize operational efficiencies to protect and enhance their leading profitability.

For Cluster 1 (Moderate Performers / Smaller Operations - ABB, BAB, BVB, EIB, etc.):

- Niche Market Focus & Specialization: Identify and cultivate specific customer segments or product niches where they can build a strong competitive advantage and achieve better profitability.
- Enhance Operational Efficiency: Invest strategically in technology and process improvements to reduce the cost-to-income ratio and improve scalability. This could include adopting lean methodologies or shared service models.
- Customer-Centric Growth: Develop targeted customer acquisition strategies and focus on deepening relationships with existing customers through cross-selling and improved service quality.

- Prudent Risk & Capital Management: While current provisions are low, ensure robust risk management frameworks are in place to manage potential increases in credit risk as they grow their loan books. Proactive capital planning is essential.

For Cluster 2 (High Income, High Cost & Provision Leaders - BID, CTG):

- Aggressive Asset Quality Management: Implement a focused strategy to manage and resolve non-performing loans (NPLs). This includes strengthening credit underwriting standards, enhancing early warning systems for credit deterioration, and proactive loan restructuring or recovery efforts.
- Comprehensive Cost Rationalization: Undertake a thorough review of their operating cost structure to identify areas for significant efficiency improvements and cost reductions without negatively impacting core services or risk controls.
- Optimize Risk-Return Profile: Re-evaluate their overall risk appetite and the risk-return profiles of different business segments and loan products. Consider strategic shifts towards more stable, lower-risk income streams or segments.
- Strengthen Capital Adequacy: Given the high provisions, it is crucial to ensure robust capital buffers are maintained and potentially enhanced to absorb any further credit shocks and meet regulatory expectations.

VI. Conclusion

This KQKD-based segmentation provides valuable insights into the diverse financial performance and risk landscapes of Vietnamese banks in Q3 2023. The identified peer groups and tailored recommendations can serve as a foundation for financial consulting engagements, enabling more informed strategic planning, risk management, and competitive positioning for individual banks within the sector.