IEEE Machine Learning Hackathon at Nazarbayev University

“Smart Tech” team

Miras Mukhametkaziev, Almansur Kakimov

Desht Lab Case

Feb 23, 2025

1. Data Loading and Exploration

The dataset (WVS_Cross-National_Wave_7_csv_v6_0.csv) is loaded into a Pandas DataFrame.

The initial structure of the dataset is examined:

- Shape: (97220, 613) (97,220 rows and 613 columns)
- Data types: float64, int64, object
- Missing values: Checked per column, with some columns having significant missing data.
- Summary statistics (mean, std, min, max, quartiles) are computed.

2. Data Preprocessing

Missing Values Handling

- Columns with more than 50% missing values are dropped.
- Numerical missing values are filled with the median.
- Categorical missing values are replaced with the mode.

Feature Engineering

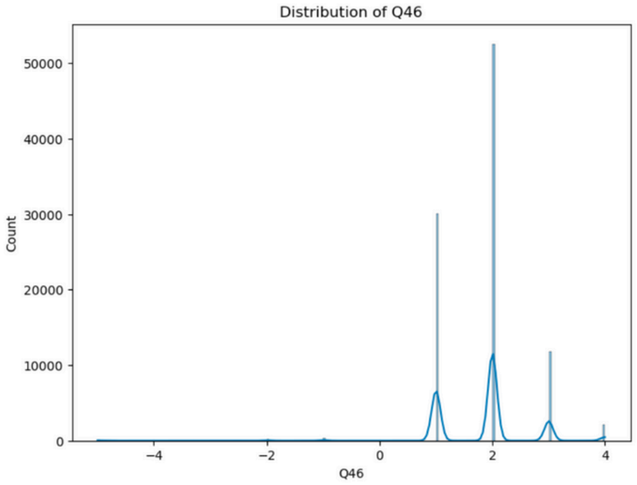
- Selected columns for economic and societal values analysis.
- Created histograms, bar plots, and correlation matrices.

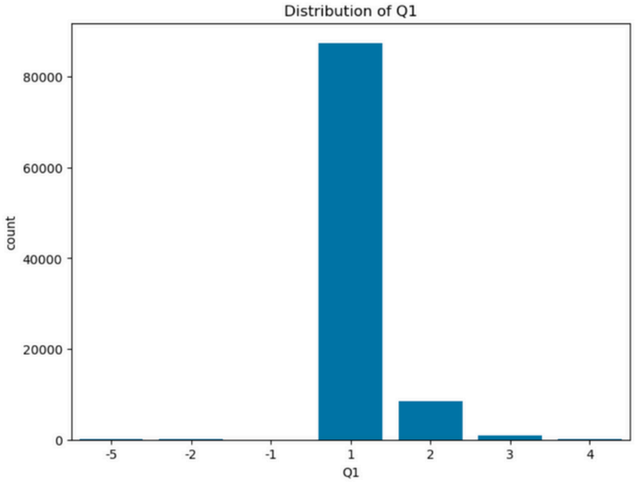
Standardization

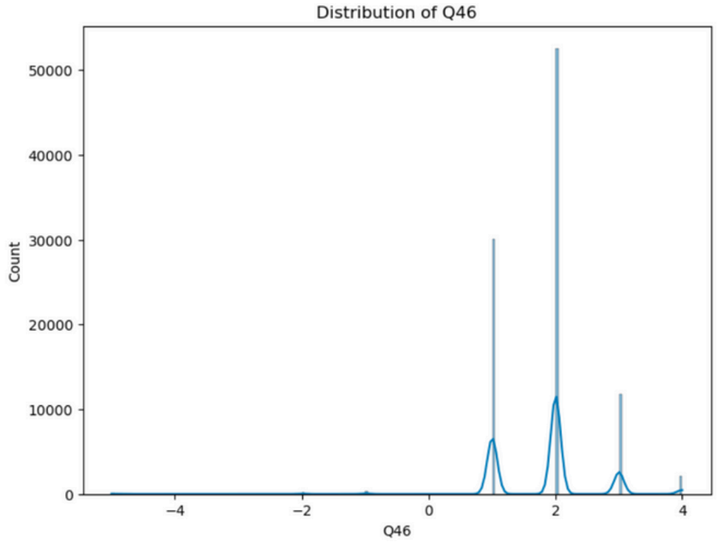
- Numerical columns are standardized using StandardScaler.

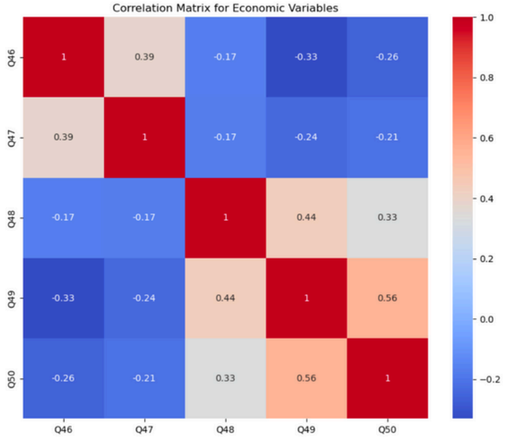
One-Hot Encoding

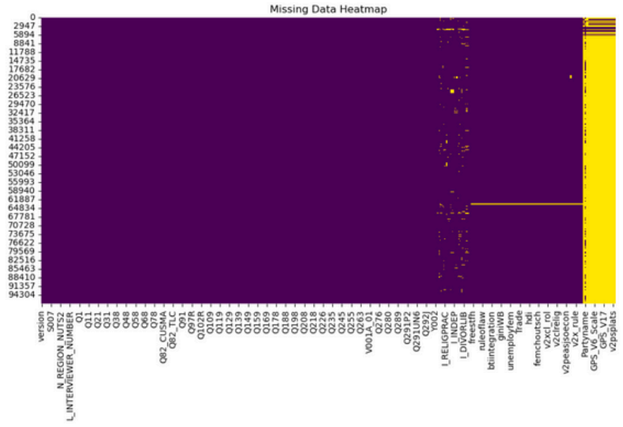
- Categorical variables are converted into numerical format via pd.get_dummies().

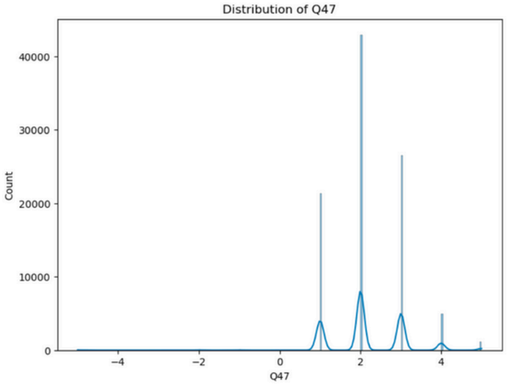

Distribution of Q46
Count
Q46

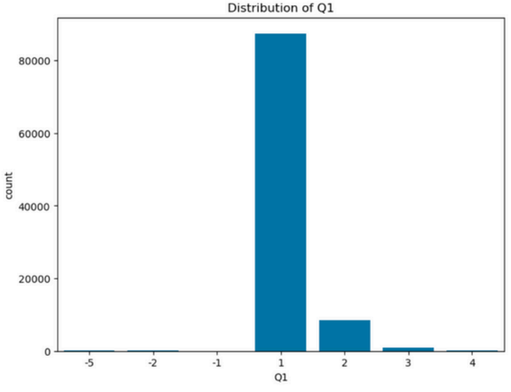

Distribution of Q1
count
Q1

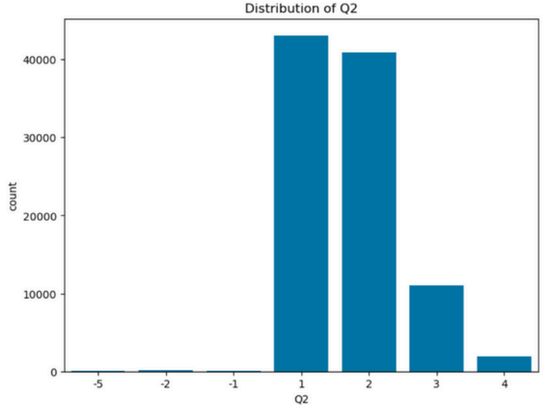

Distribution of Q46
Count
Q46

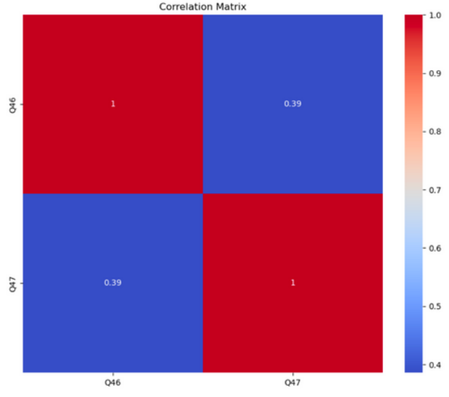

Correlation Matrix for Economic Variables
Q46
Q47
Q48
Q49
Q50


Missing Data Heatmap


Distribution of Q47
Count
Q47


Distribution of Q1
count
Q1


Distribution of Q2
count
Q2


Correlation Matrix
Q46
Q47

3. Clustering Methods

K-Means Clustering

MiniBatchKMeans is used for clustering:

- A sample of 10,000 observations is taken for efficiency.

The Elbow Method (inertia) and Silhouette Score are used to determine the optimal number of clusters (K).
K=4 is selected based on the Elbow and Silhouette analysis.

Final K-Means Model: the dataset is clustered into 4 groups.

Distribution:

- Cluster 1: 41,970 samples
- Cluster 3: 26,916 samples
- Cluster 0: 19,522 samples
- Cluster 2: 8,812 samples

Cluster labels are added to the dataset.

DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering) is also applied:
Parameters: eps=0.5, min_samples=5
Many samples (-1) are classified as noise, meaning DBSCAN is not very effective on this dataset.

4. Clustering Evaluation

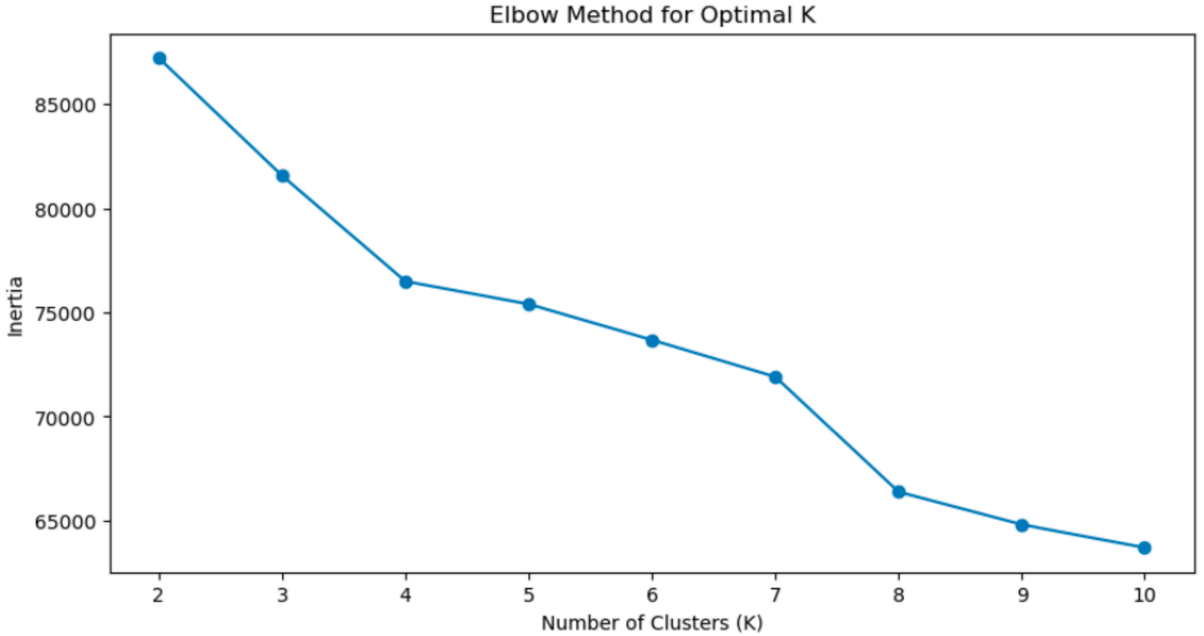
Silhouette Score

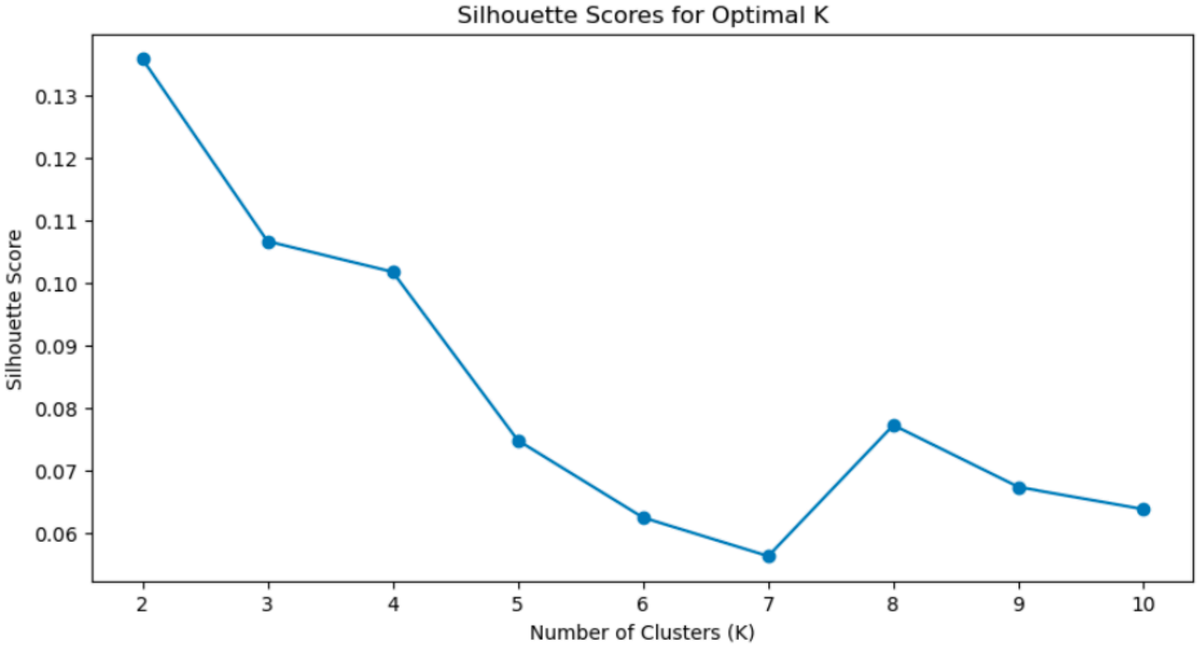
- K-Means: 0.120 (low but acceptable)
- DBSCAN: -0.220 (poor clustering)

Davies-Bouldin Index

- K-Means: 2.195 (lower is better)
- DBSCAN: 1.539

DBSCAN performs poorly compared to K-Means, likely due to the dataset's high dimensionality and sparse density regions.


Elbow Method for Optimal K
Inertia
85000
80000
75000
70000
65000
2
3
4
5
6
7
8
9
10
Number of Clusters (K)


Silhouette Scores for Optimal K
Silhouette Score
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
2
3
4
5
6
7
8
9
10
Number of Clusters (K)

5. Dimensionality Reduction and Visualization

PCA (Principal Component Analysis)

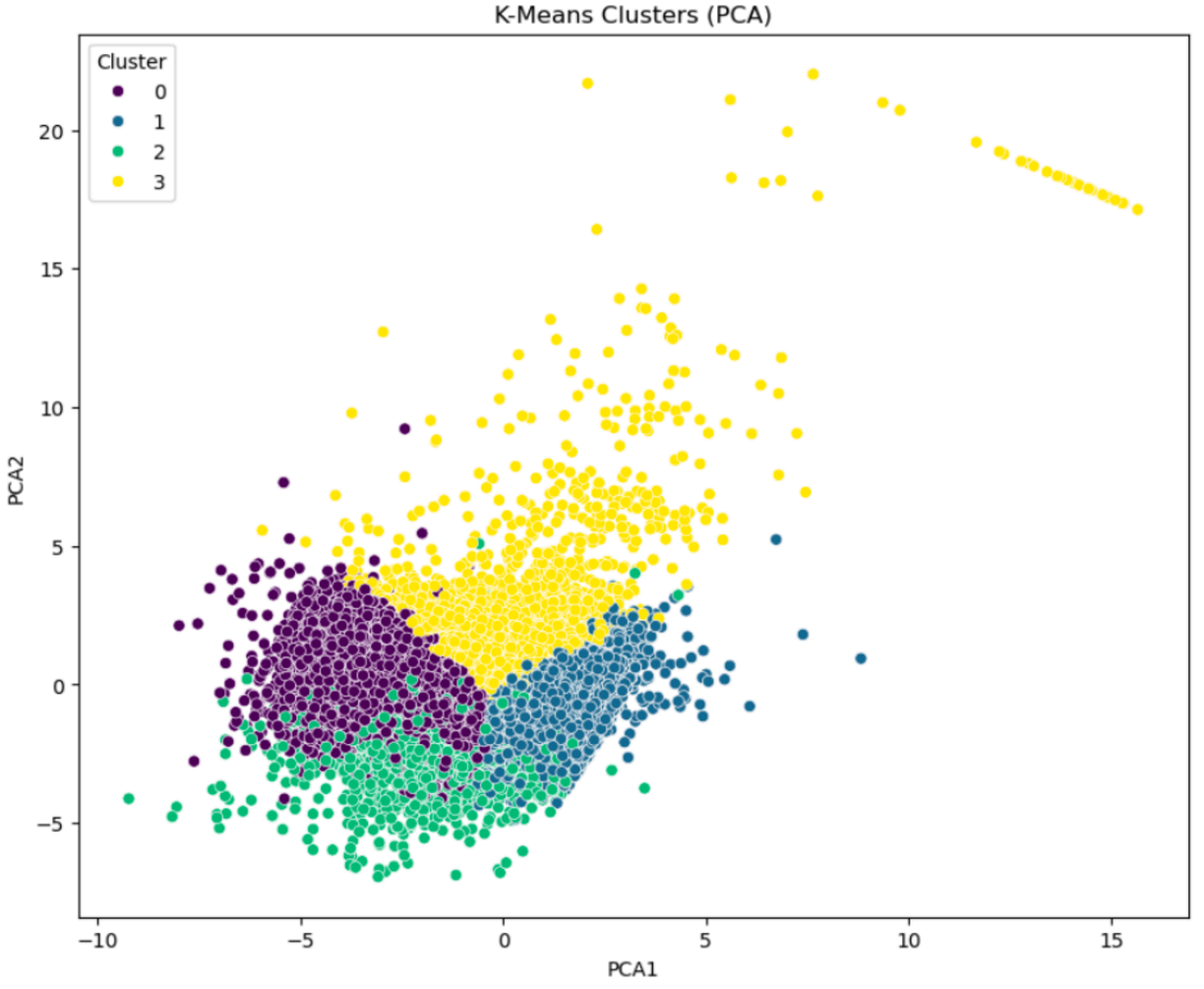
- Reduced dataset to 2 components for visualization.
- Scatter plot of clusters.

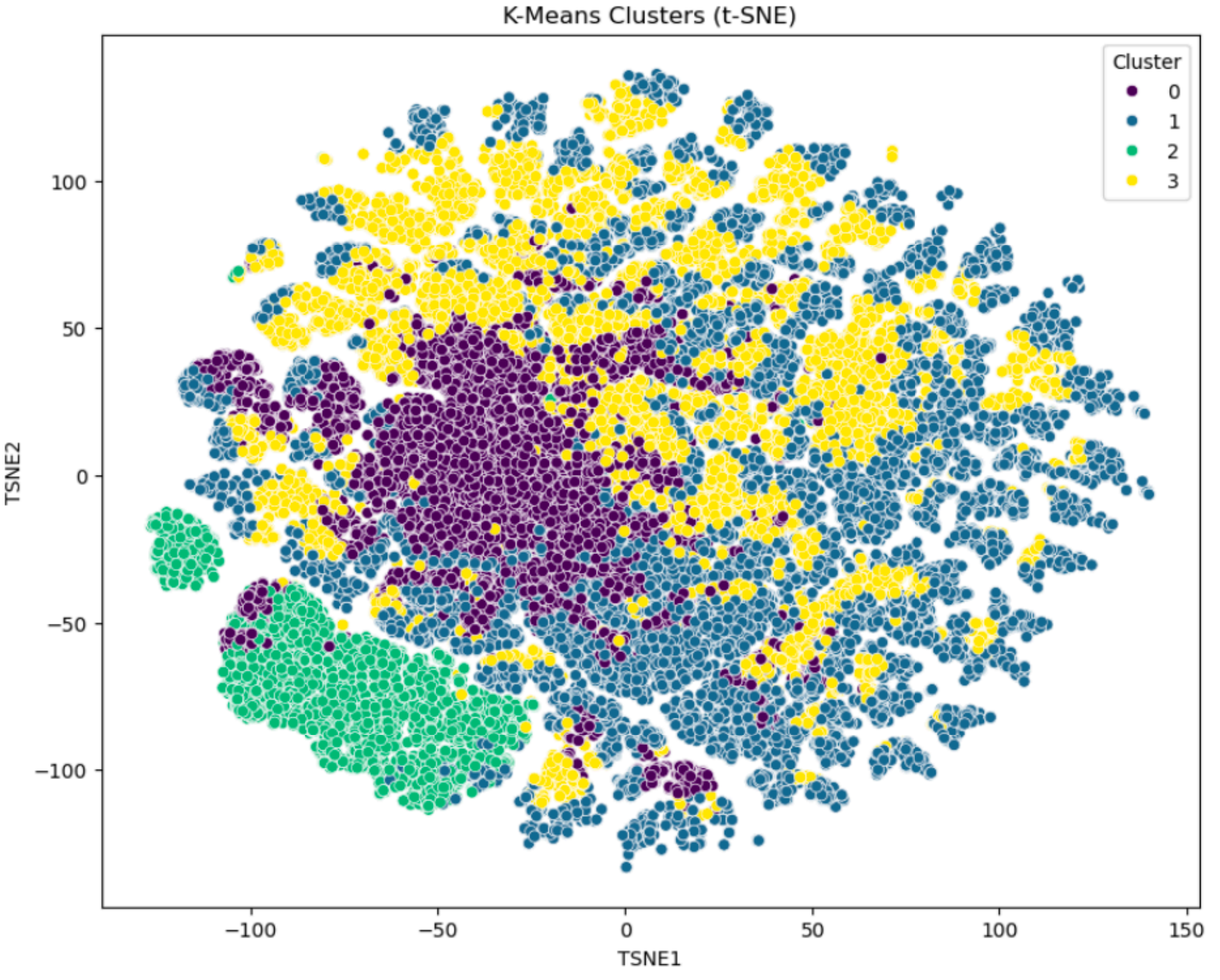
t-SNE (t-Distributed Stochastic Neighbor Embedding)

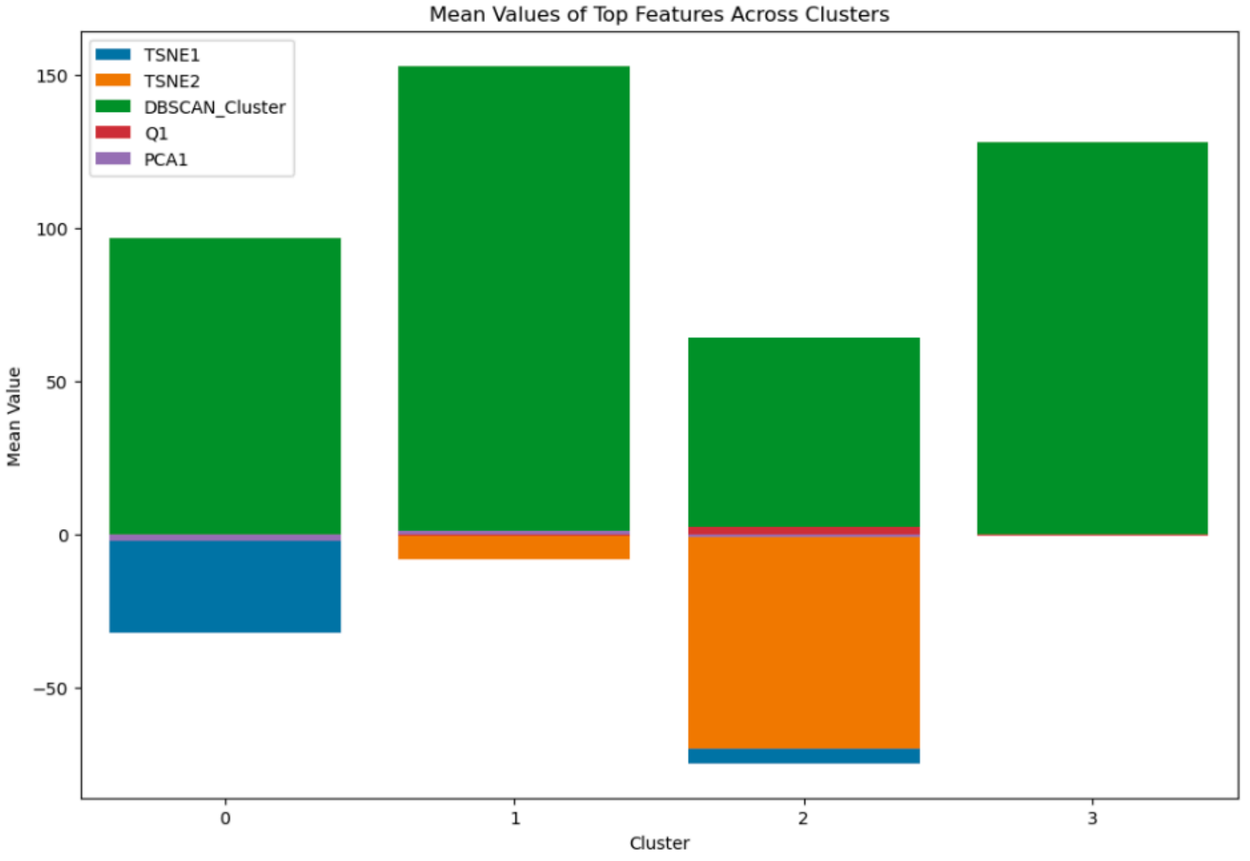
- Another technique to reduce dimensions.
- Scatter plot of clusters with better separability.

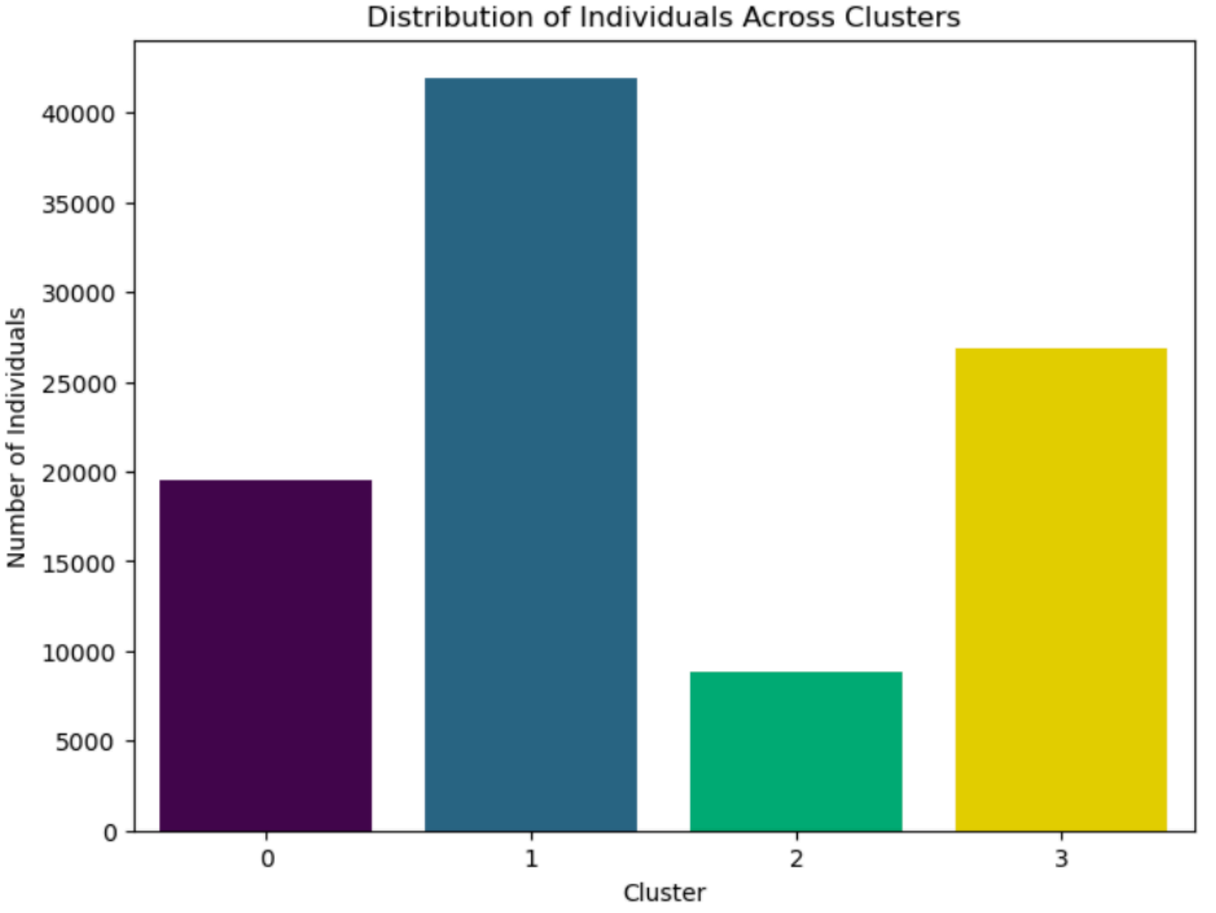
Parallel Coordinates Plot

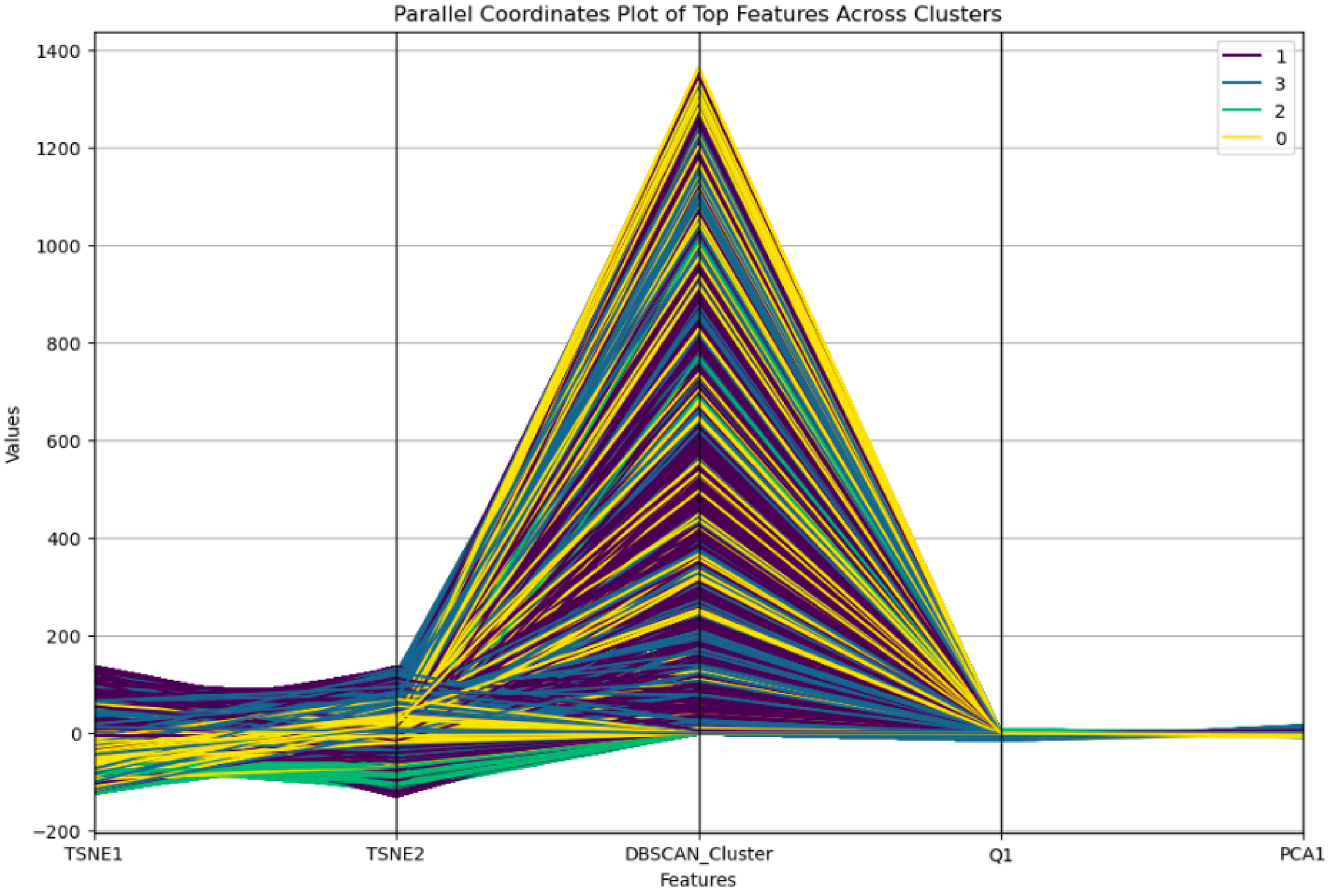
- Used to visualize feature differences across clusters.


K-Means Clusters (PCA)
Cluster
0
1
2
3
PCA1
PCA2
−10
−5
0
5
10
15
20
−5
0
5
10
15


K-Means Clusters (t-SNE)
Cluster
0
1
2
3
TSNE1
TSNE2
−100
−50
0
50
100
150
−100
−50
0
50
100


Mean Values of Top Features Across Clusters
TSNE1
TSNE2
DBSCAN_Cluster
Q1
PCA1
150
100
50
0
−50
Mean Value
0
1
2
3
Cluster


Distribution of Individuals Across Clusters
40000
35000
30000
25000
20000
15000
10000
5000
0
Number of Individuals
0
1
2
3
Cluster


Parallel Coordinates Plot of Top Features Across Clusters
1400
1200
1000
800
600
400
200
0
−200
Values
1
3
2
0
TSNE1
TSNE2
DBSCAN_Cluster
Q1
PCA1
Features

6. Interpretation of Clusters

Cluster 0 ("Rural Poor")

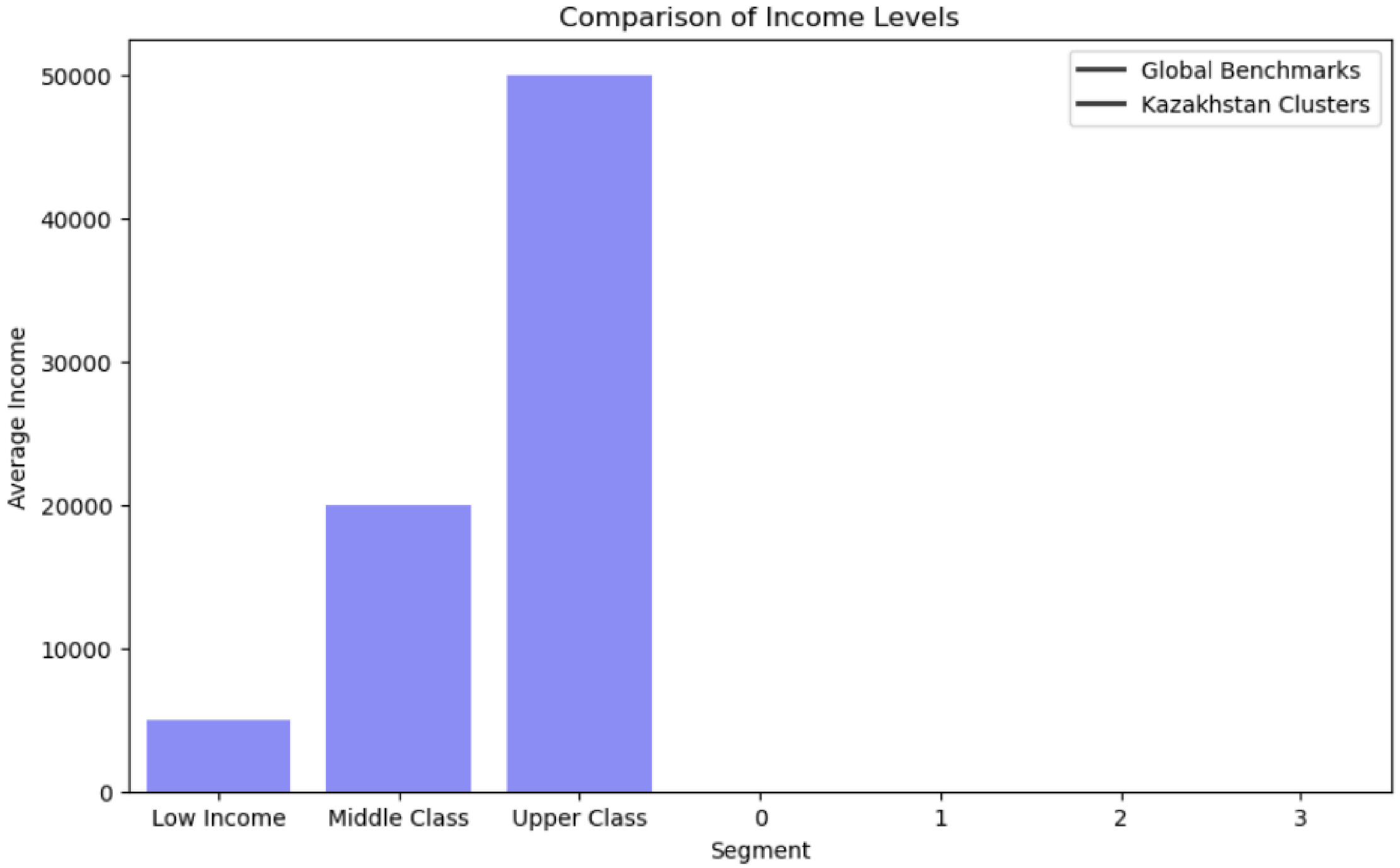
- High religiosity
- Low income and education
- Mostly rural population.

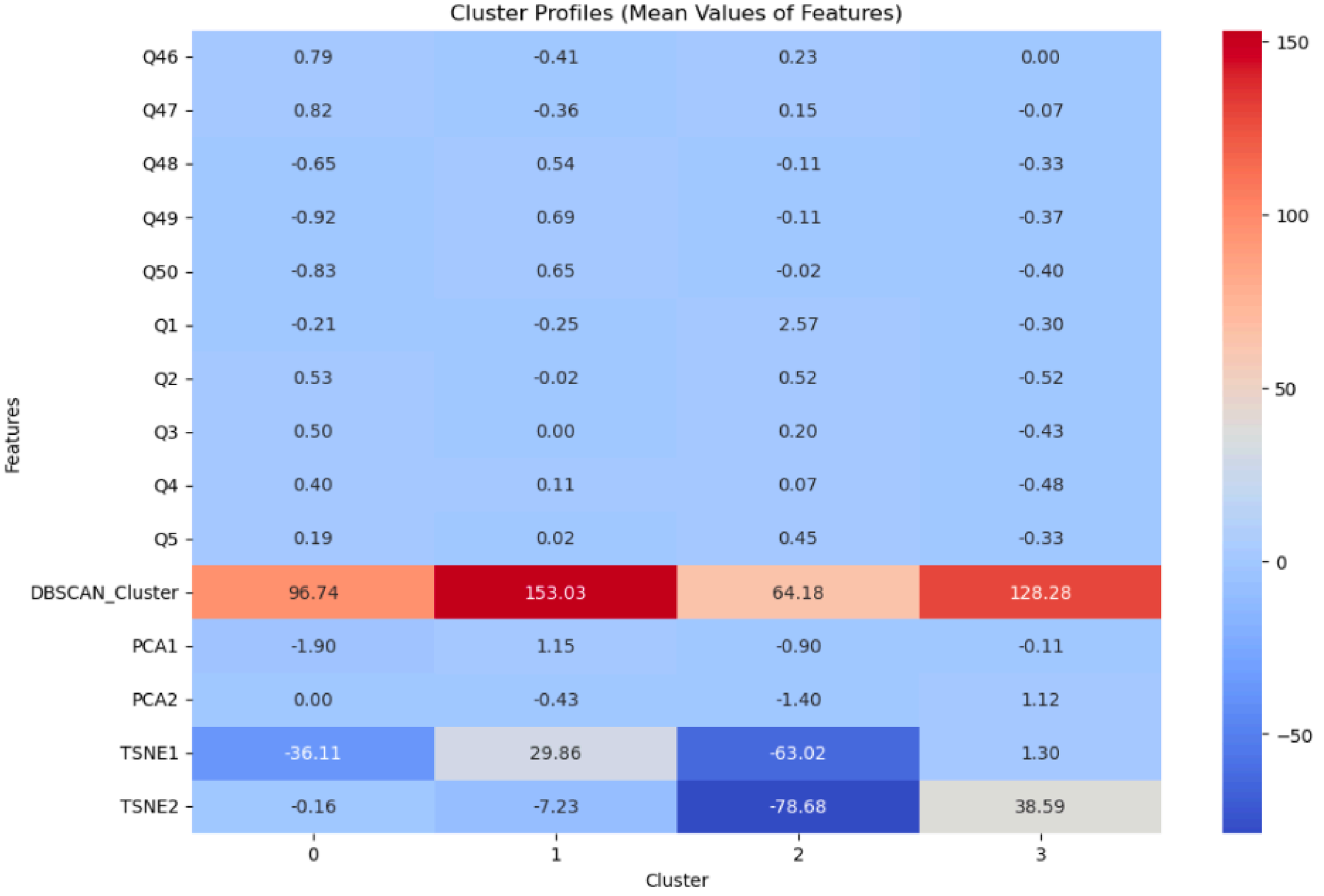
Cluster 1 ("Urban Middle Class")

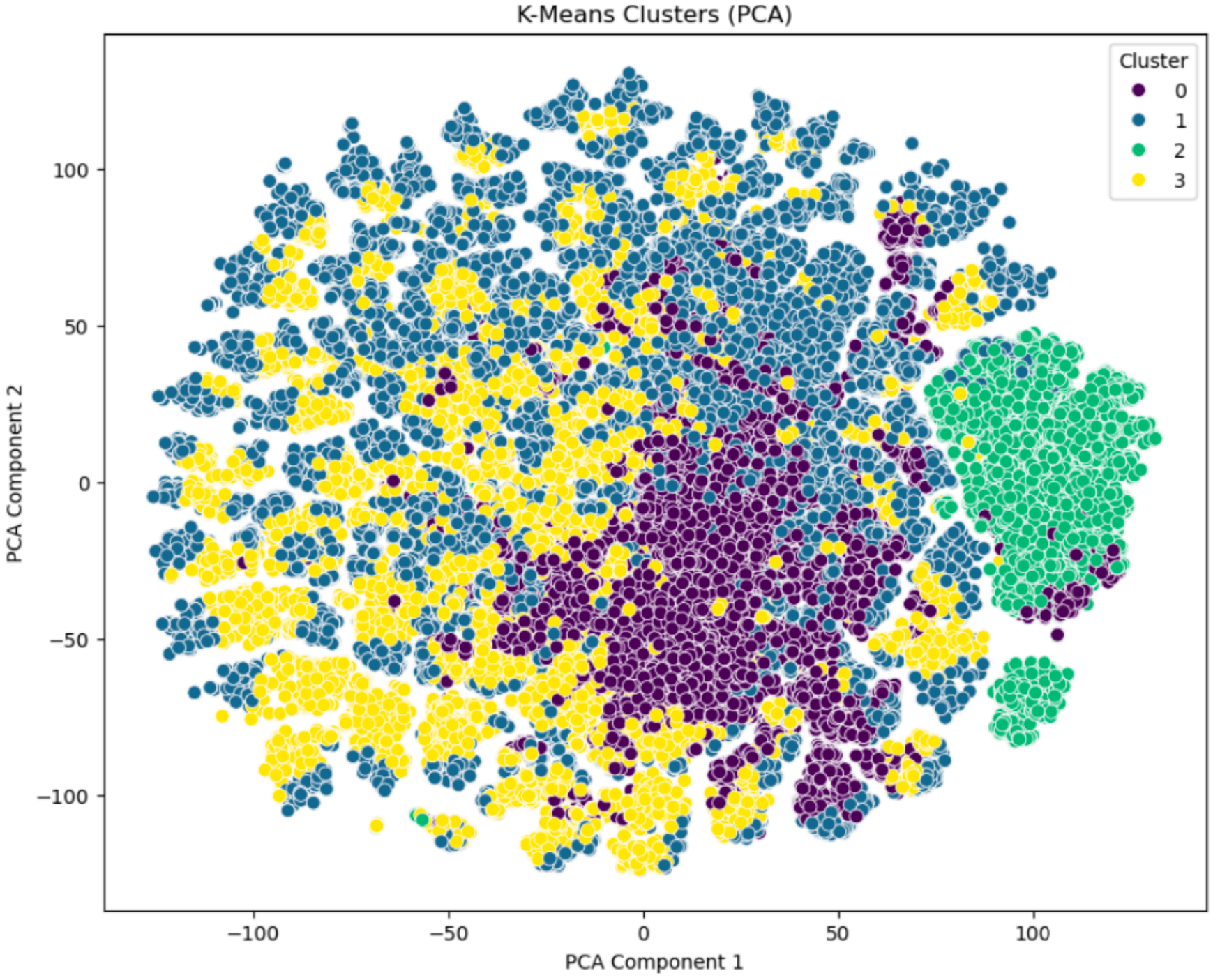
- Moderate religiosity
- Higher income and education
- Mostly urban population.

Cluster 2 ("Elite")

- High income and education
- Low religiosity
- Urban-based.


Comparison of Income Levels
Global Benchmarks
Kazakhstan Clusters
Average Income
50000
40000
30000
20000
10000
0
Low Income
Middle Class
Upper Class
0
1
2
3
Segment


Cluster Profiles (Mean Values of Features)
Features
Q46 0.79 -0.41 0.23 0.00
Q47 0.82 -0.36 0.15 -0.07
Q48 -0.65 0.54 -0.11 -0.33
Q49 -0.92 0.69 -0.11 -0.37
Q50 -0.83 0.65 -0.02 -0.40
Q1 -0.21 -0.25 2.57 -0.30
Q2 0.53 -0.02 0.52 -0.52
Q3 0.50 0.00 0.20 -0.43
Q4 0.40 0.11 0.07 -0.48
Q5 0.19 0.02 0.45 -0.33
DBSCAN_Cluster 96.74 153.03 64.18 128.28
PCA1 -1.90 1.15 -0.90 -0.11
PCA2 0.00 -0.43 -1.40 1.12
TSNE1 -36.11 29.86 -63.02 1.30
TSNE2 -0.16 -7.23 -78.68 38.59
0 1 2 3
Cluster
150
100
50
0
−50


K-Means Clusters (PCA)
Cluster
0
1
2
3
PCA Component 2
100
50
0
−50
−100
PCA Component 1
−100
−50
0
50
100

7. Saving Results

Cleaned dataset (cleaned_wvs_data.csv)
Clustered dataset (clustered_wvs_data.csv)
Cluster summary (cluster_summary.csv)
Important feature list (important_features.csv)

Conclusion

- K-Means performed better than DBSCAN, as the latter produced too many noise points.
- 4 distinct clusters were identified, with clear socio-economic differences.
- PCA and t-SNE helped visualize cluster separability.
- The results offer insights into different socio-economic groups based on values, income, and education.