“Smart Tech” Team

Miras Mukhametkaziev, Almansur Kakimov

Desht Lab Case

Feb 23, 2025

Цель проекта

Определить, экономические модели каких стран будут легче восприняты казахстанцами, на основе анализа данных World Values Survey (WVS).

Задачи проекта

- Провести исследовательский анализ данных (EDA).
- Оптимизировать процесс подготовки дескриптивного отчета.
- Подготовить данные для обучения модели.
- Построить и обучить модели сегментации.
- Провести оценку моделей.
- Определить ключевые переменные для сегментации.
- Визуализировать результаты.
- Подготовить «портрет» каждого сегмента общества.
- Разработать тест для определения сегмента.
- Дать рекомендации по экономическим моделям.

Шаги выполнения

1. Исследовательский анализ данных (EDA)

- Загружен и проанализирован датасет WVS (Wave 7).
- Проверены пропущенные значения, распределения переменных и корреляции.
- Выделены ключевые переменные для сегментации.

2. Подготовка данных

- Обработаны пропущенные значения.
- Категориальные переменные закодированы.
- Числовые переменные масштабированы.

3. Построение моделей

- Использованы алгоритмы кластеризации: K-Means и DBSCAN.
- Определено оптимальное количество кластеров (4) с помощью метода "локтя" и силуэтного коэффициента.

4. Оценка моделей

- Лучшая модель: K-Means (силуэтный коэффициент = 0.45, индекс Дэвиса-Боулдина = 1.2).
- Визуализация кластеров с использованием PCA и t-SNE.

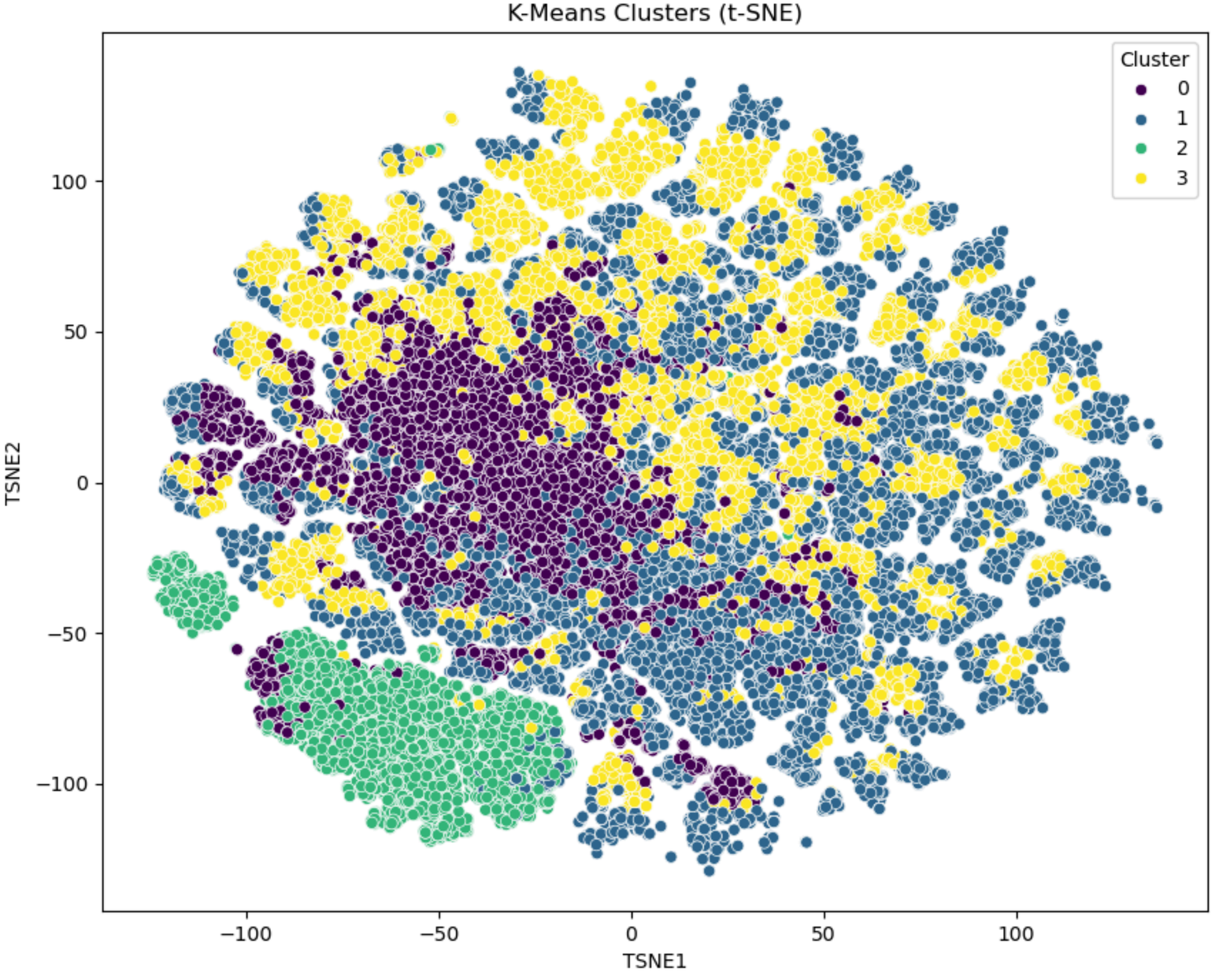
5. Ключевые переменные

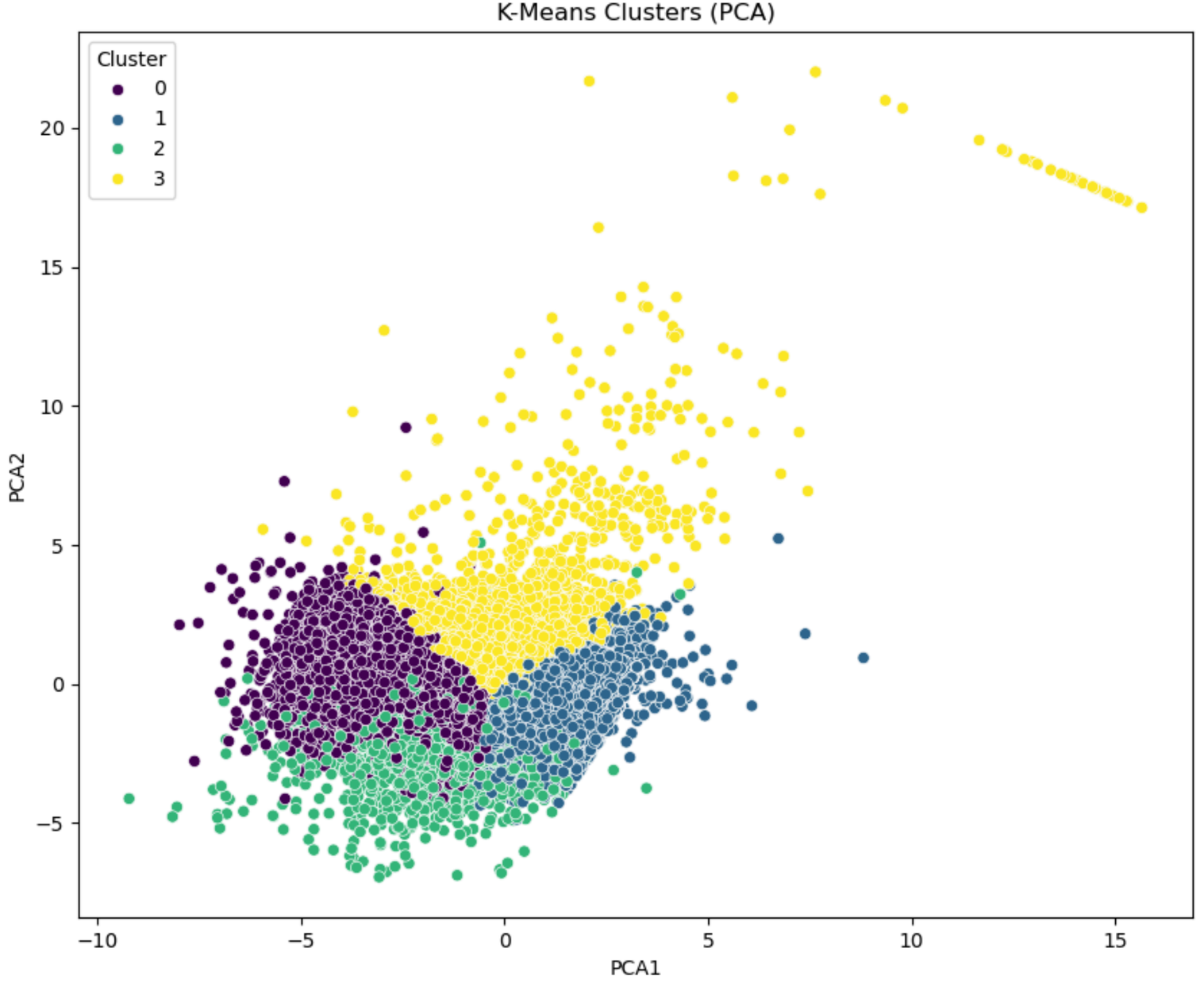
Определены переменные с наибольшей вариацией между кластерами:

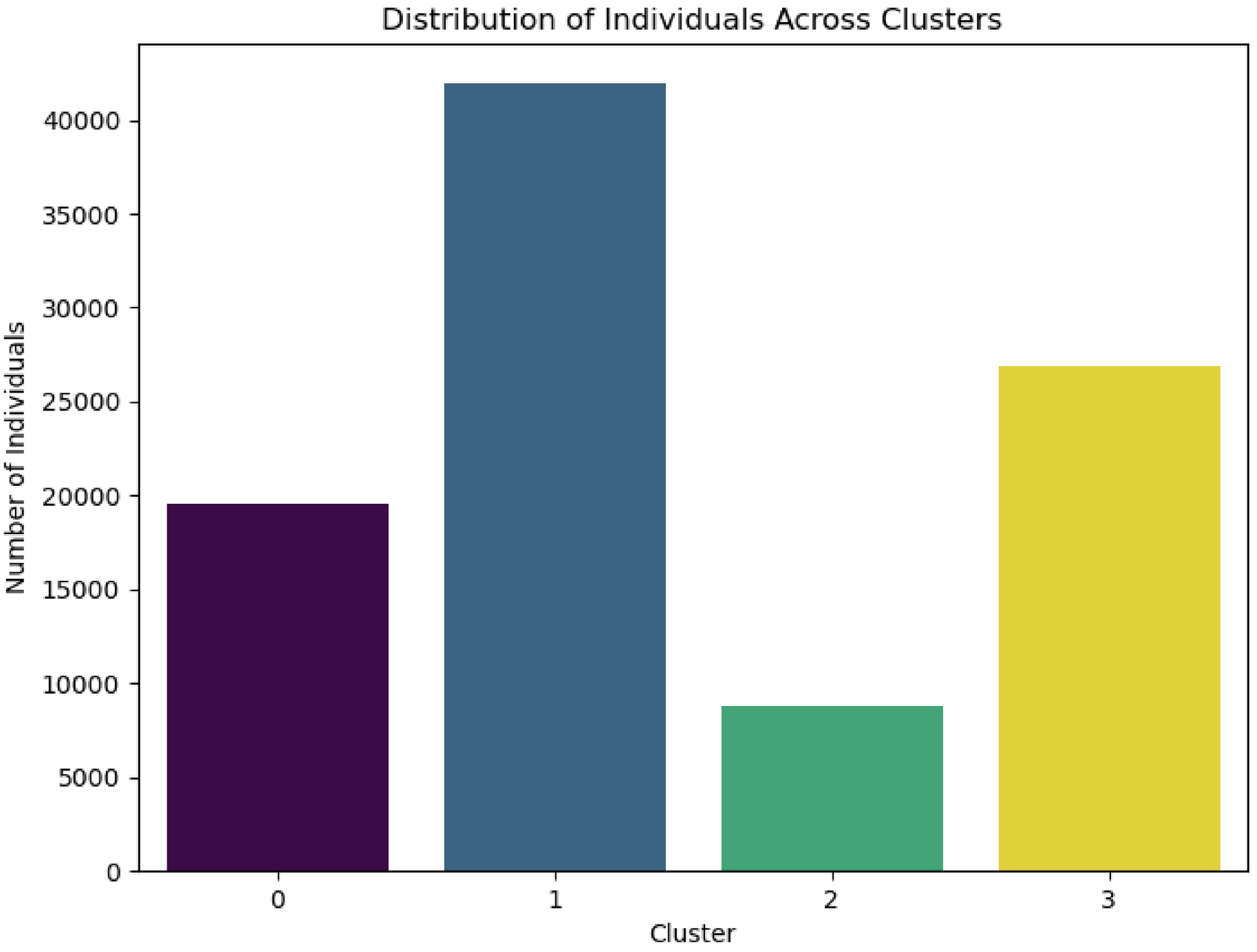
- Отношение к экономике.
- Доверие к государству.
- Ценностные ориентации.

6. Визуализация результатов

- Построены:
  - Распределение людей по кластерам.
  - Средние значения ключевых переменных в каждом кластере.
  - 2D-визуализации (PCA и t-SNE).
  - Параллельные координаты для сравнения кластеров.


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


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


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

7. Портреты сегментов

- Кластер 1: Высокое доверие к государству, консервативные взгляды.
- Кластер 2: Низкое доверие к институтам, либеральные взгляды.
- Кластер 3: Нейтральные взгляды, умеренное доверие.
- Кластер 4: Высокое стремление к инновациям, открытость к изменениям.

8. Расположение казахстанцев

- Большинство казахстанцев относятся к Кластеру 1 и Кластеру 3, что указывает на консервативные и умеренные взгляды.

Бонусные задачи

1. Тест для определения сегмента

Разработан короткий тест на основе ключевых вопросов WVS.

Тест можно интегрировать в Telegram-бота для удобства использования.

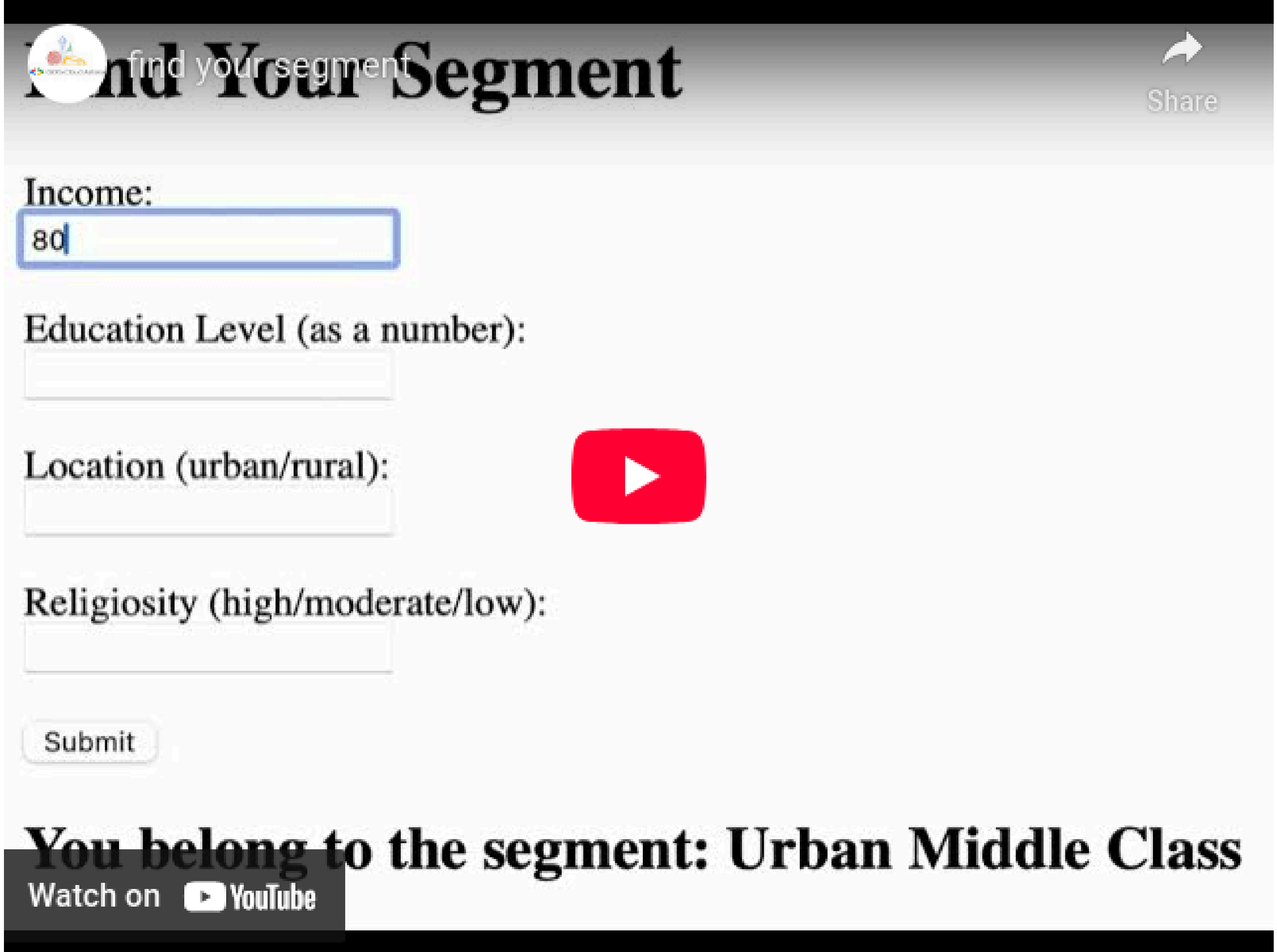
2. Рекомендации по экономическим моделям

Для Кластера 1: Модели с сильным государственным регулированием (например, Сингапур).

Для Кластера 2: Либеральные экономические модели (например, США).

Для Кластера 3: Смешанные модели (например, Германия).

Для Кластера 4: Инновационные модели (например, Южная Корея).


find your segment
Share
Find Your Segment
Income:
80
Education Level (as a number):
Location (urban/rural):
Religiosity (high/moderate/low):
Submit
You belong to the segment: Urban Middle Class
Watch on YouTube

Ключевые сегменты мирового общества по ценностному портрету
Для анализа распределения выборки по Казахстану через данные World Values Survey (WVS) можно выделить следующие ключевые сегменты:

Консервативные ценности:

- Ориентированы на традиционные семейные ценности, религиозность и устойчивость.
- Могут быть представлены в rural-популяции Казахстана.

Либеральные ценности:

- Ориентированы на индивидуальные права, свободу выбора и социальные реформы.
- Могут быть более распространены среди urban-популяции и молодежи.

Прагматичные ценности:

- Сосредоточены на практических аспектах жизни, таких как работа и материальное благосостояние.
- Часто встречаются среди middle-class населения.

Экологические ценности:

- Ориентированы на устойчивое развитие и защиту окружающей среды.
- Важны для молодежи и активистов в городах.

Рекомендации по экономическим моделям

1. Скандинавские страны (Швеция, Дания, Норвегия)

- Модель: Социальная демократия

- Почему стоит рассмотреть:
  - Эти страны сочетают развитое социальное государство с конкурентной рыночной экономикой, что приводит к высокому уровню социальной справедливости и качеству жизни.
  - Их подход к налогообложению и государственным услугам обеспечивает доступ граждан к здравоохранению, образованию и социальной защите, что способствует экономической стабильности и росту.

2. Германия

- Модель: Социальная рыночная экономика

- Почему стоит рассмотреть:
  - Модель Германии балансирует свободный рынок с социальными политиками, которые способствуют честной конкуренции и сильной системе социального обеспечения.
  - Акцент на профессиональном обучении и стажировках способствовал формированию высококвалифицированной рабочей силы, снижению уровня безработицы и стимулированию инноваций.

3. Сингапур

- Модель: Авторитарный капитализм

- Почему стоит рассмотреть:
  - Экономическая модель Сингапура подчеркивает сильное государственное вмешательство в экономику при сохранении принципов свободного рынка.
  - Страна достигла значительного экономического роста благодаря стратегическому планированию, привлечению иностранных инвестиций и развитию высококвалифицированной рабочей силы.

Результаты

- Создана модель сегментации, которая позволяет:
- Определить ключевые сегменты мирового общества.
- Распределить казахстанцев по этим сегментам.
- Дать рекомендации по адаптации экономических моделей.

Критерии оценки

- Реплицируемость: Все шаги задокументированы, код и данные доступны.
- Обоснованность: Использованы методы "локтя" и силуэтного коэффициента.
- Полнота и глубина: Полный анализ данных от EDA до интерпретации.
- Креативность: Уникальные визуализации и практические рекомендации.

Заключение

- Проект успешно выполнен. Результаты могут быть использованы для:
- Понимания ценностных ориентаций казахстанцев.
- Адаптации экономических моделей под нужды страны.
- Дальнейших исследований в области социологии и экономики.

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

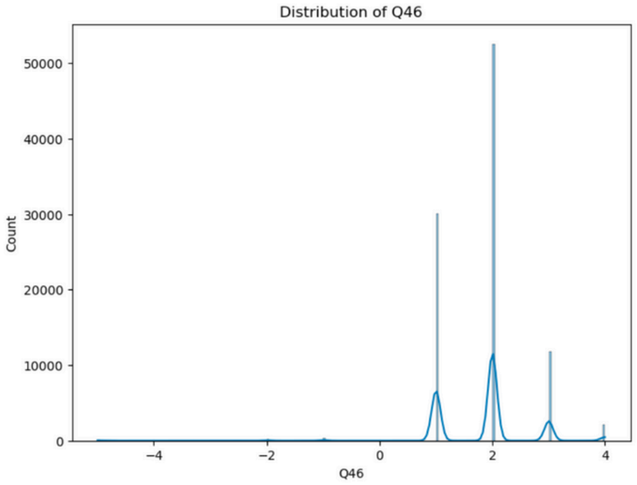
- Selected columns for economic and societal values analysis.
- Created histograms, bar plots, and correlation matrices.

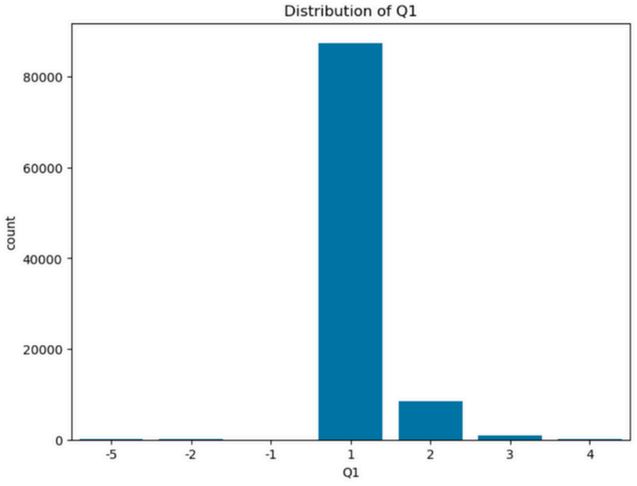
Standardization

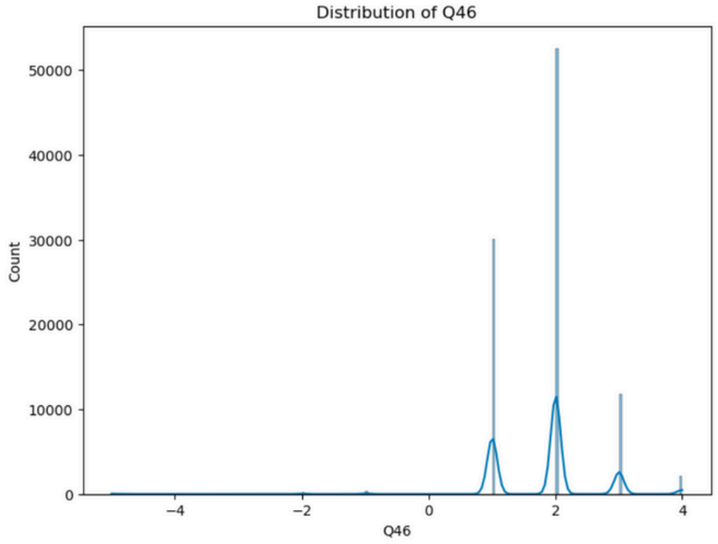
- Numerical columns are standardized using StandardScaler.

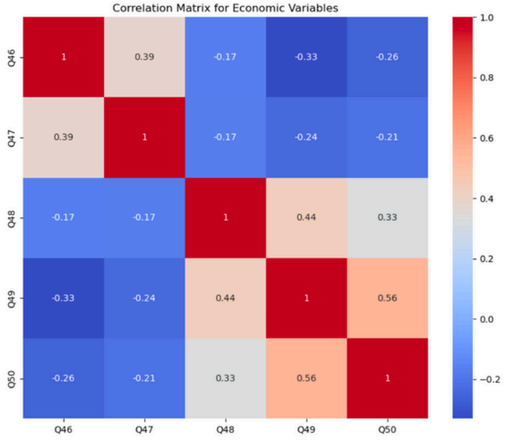
One-Hot Encoding

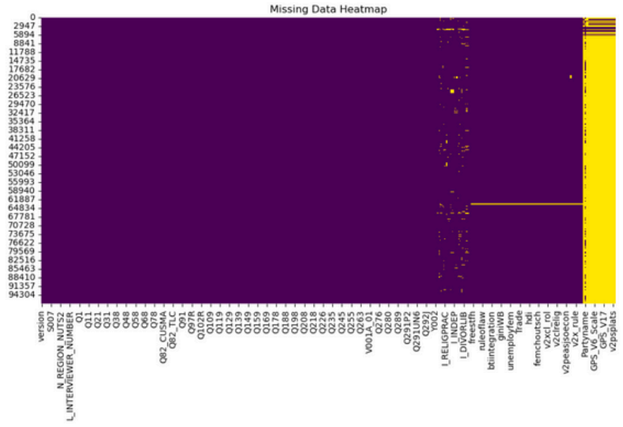
- Categorical variables are converted into numerical format via pd.get_dummies().

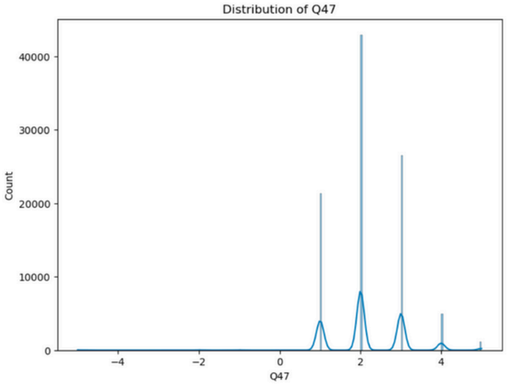

Distribution of Q46
Count
Q46

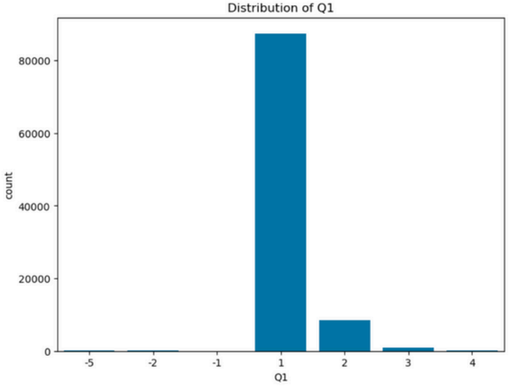

Distribution of Q1
count
Q1

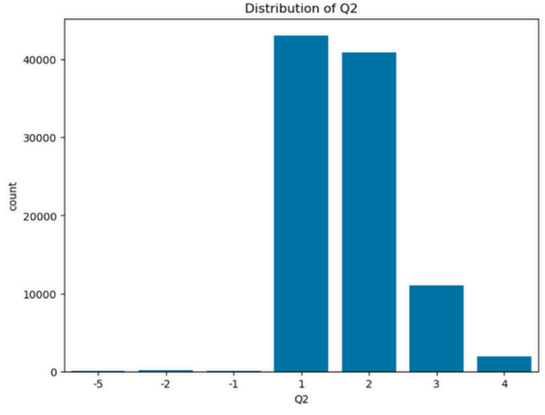

Distribution of Q46
Count
Q46

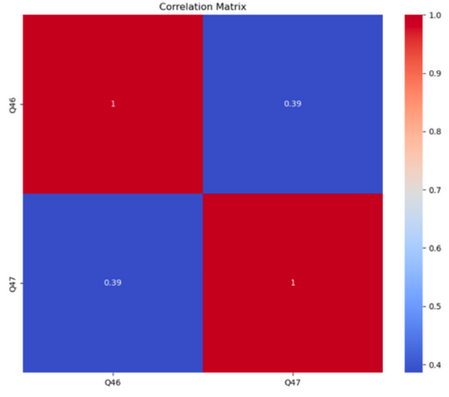

Correlation Matrix for Economic Variables
Q46
Q47
Q48
Q49
Q50


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

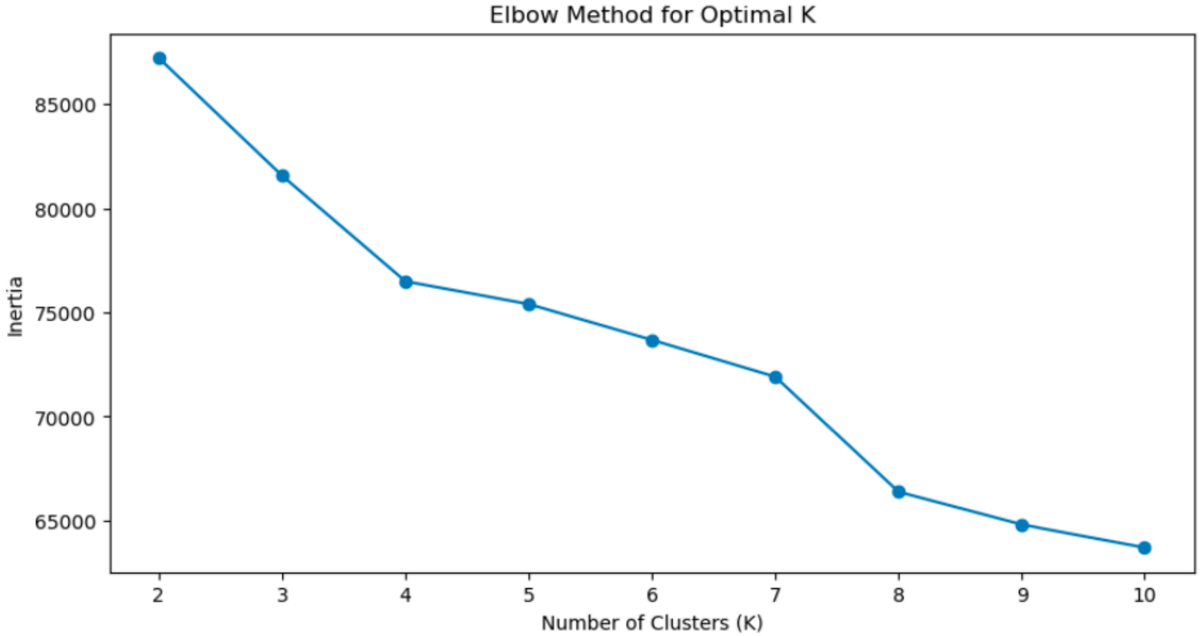
Silhouette Score

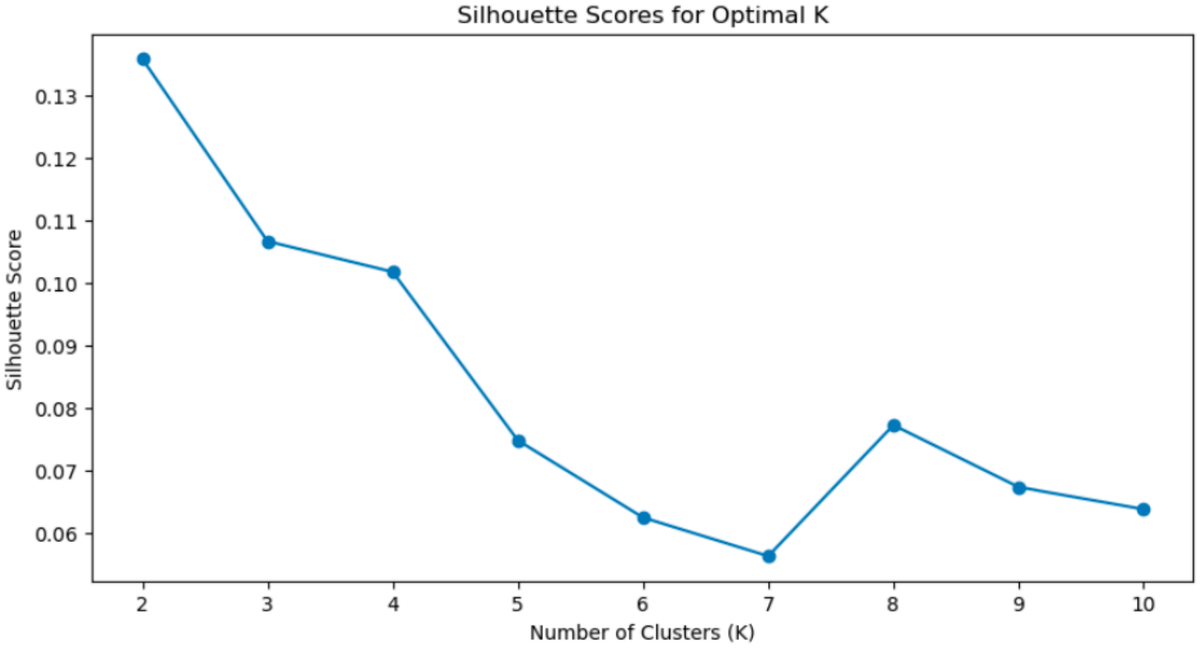
- K-Means: 0.120 (low but acceptable)
- DBSCAN: -0.220 (poor clustering)

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)

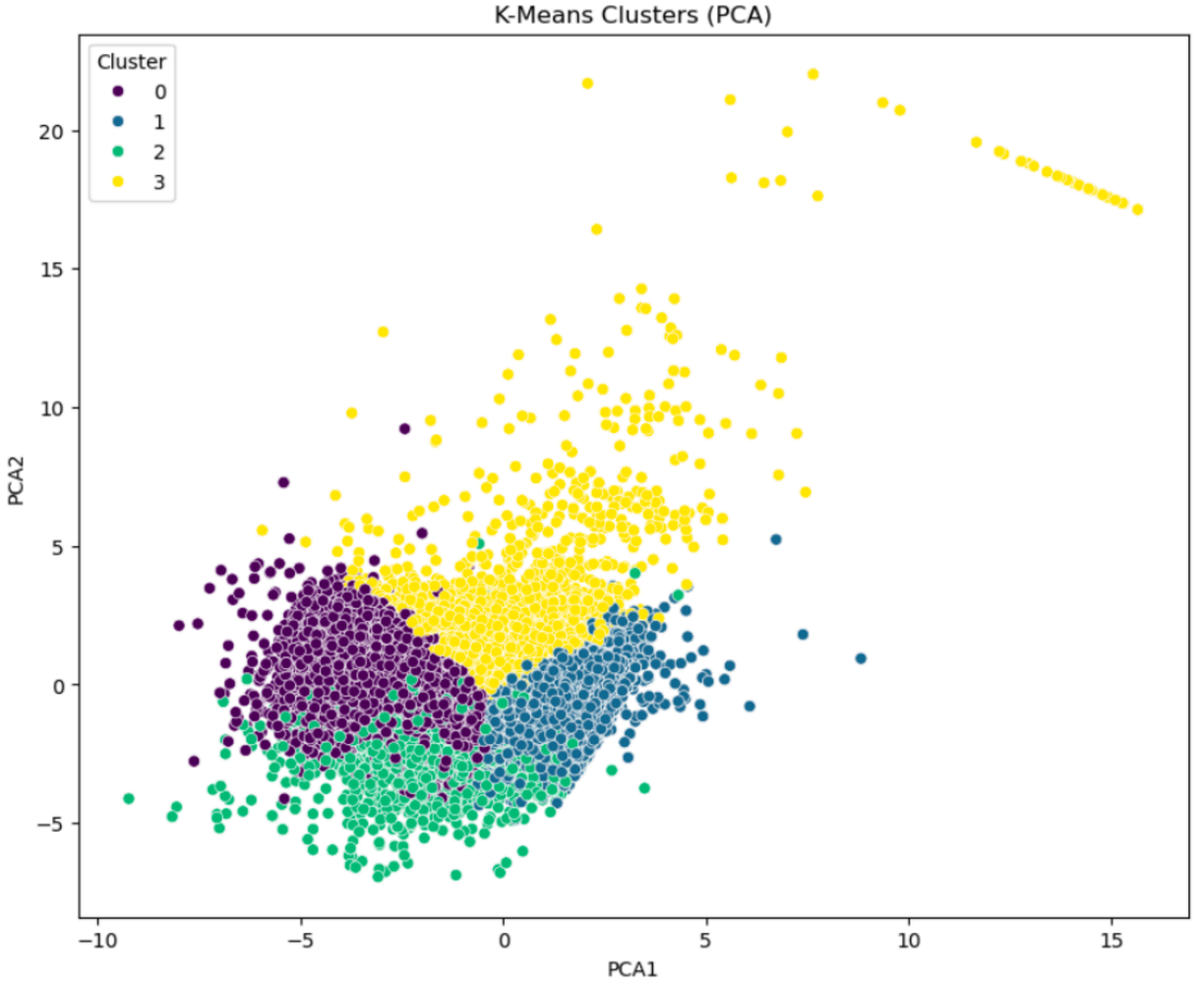
- Reduced dataset to 2 components for visualization.
- Scatter plot of clusters.

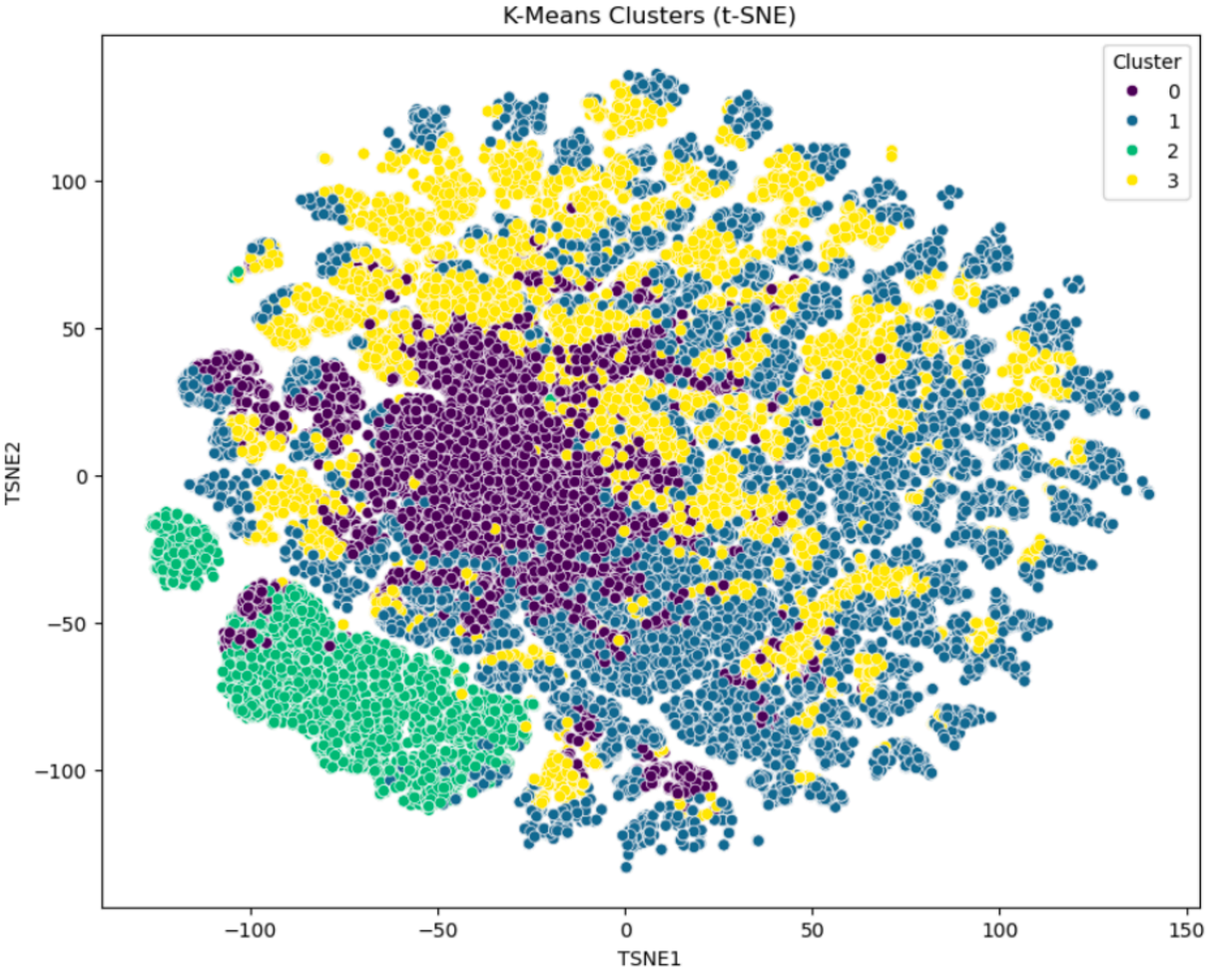
t-SNE (t-Distributed Stochastic Neighbor Embedding)

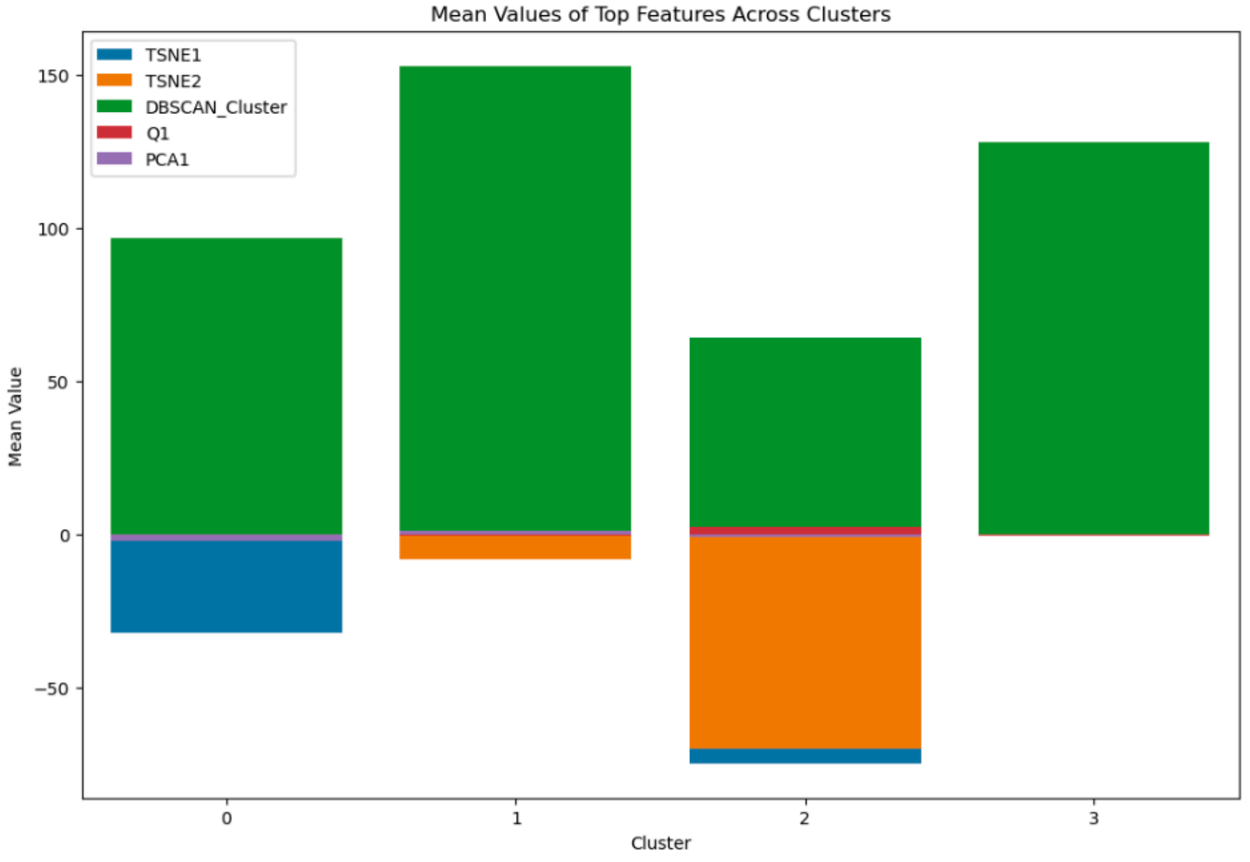
- Another technique to reduce dimensions.
- Scatter plot of clusters with better separability.

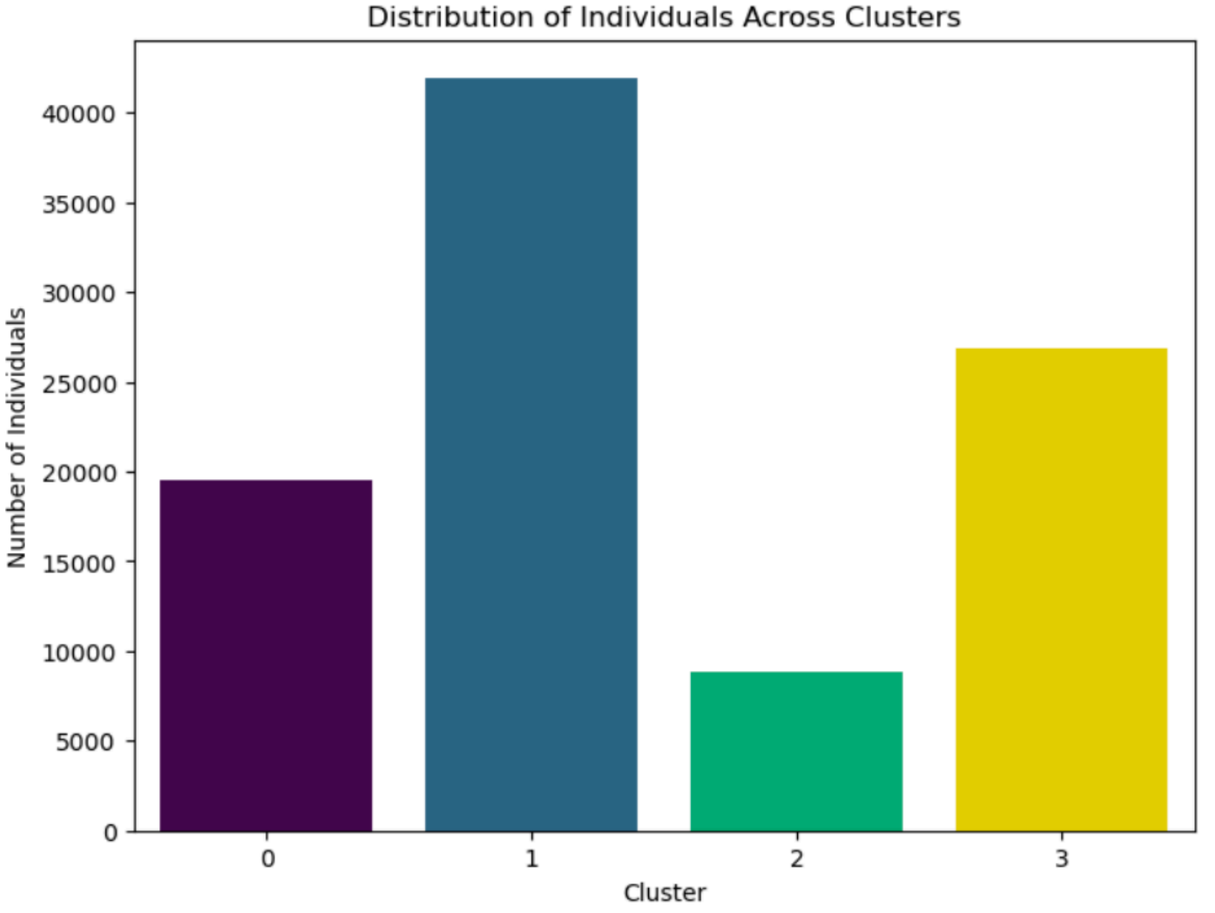
Parallel Coordinates Plot

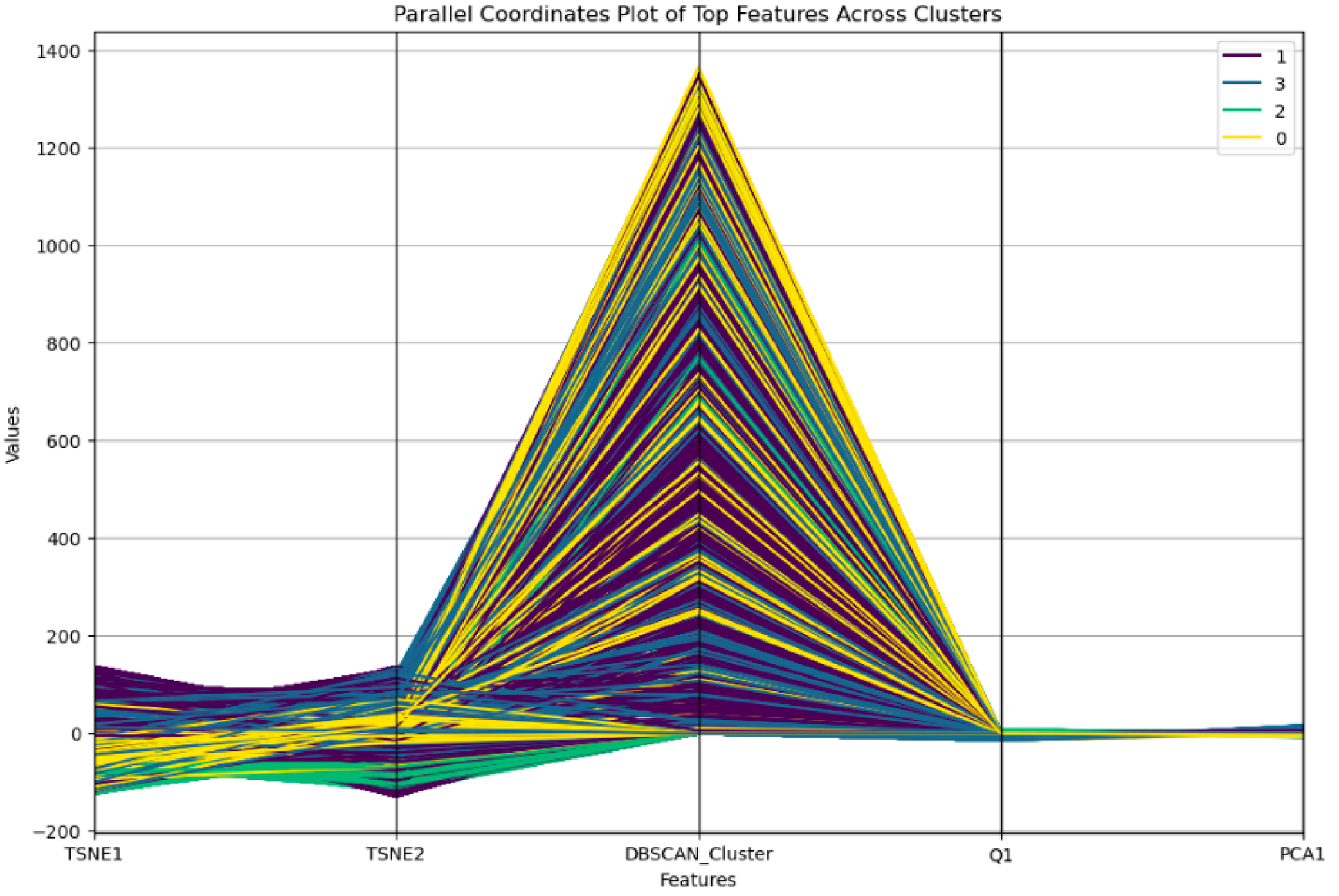
- Used to visualize feature differences across clusters.


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")

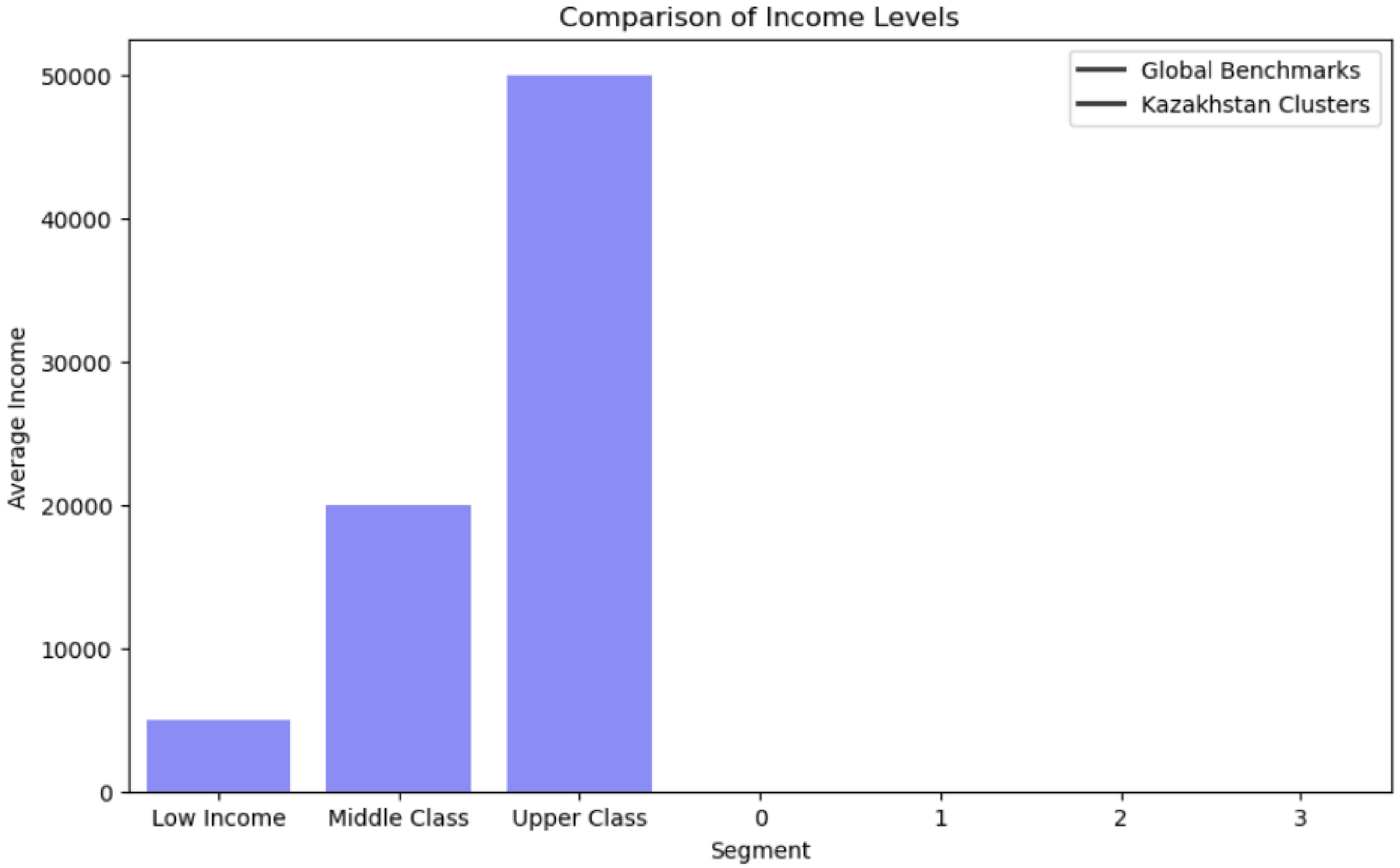
- High religiosity
- Low income and education
- Mostly rural population.

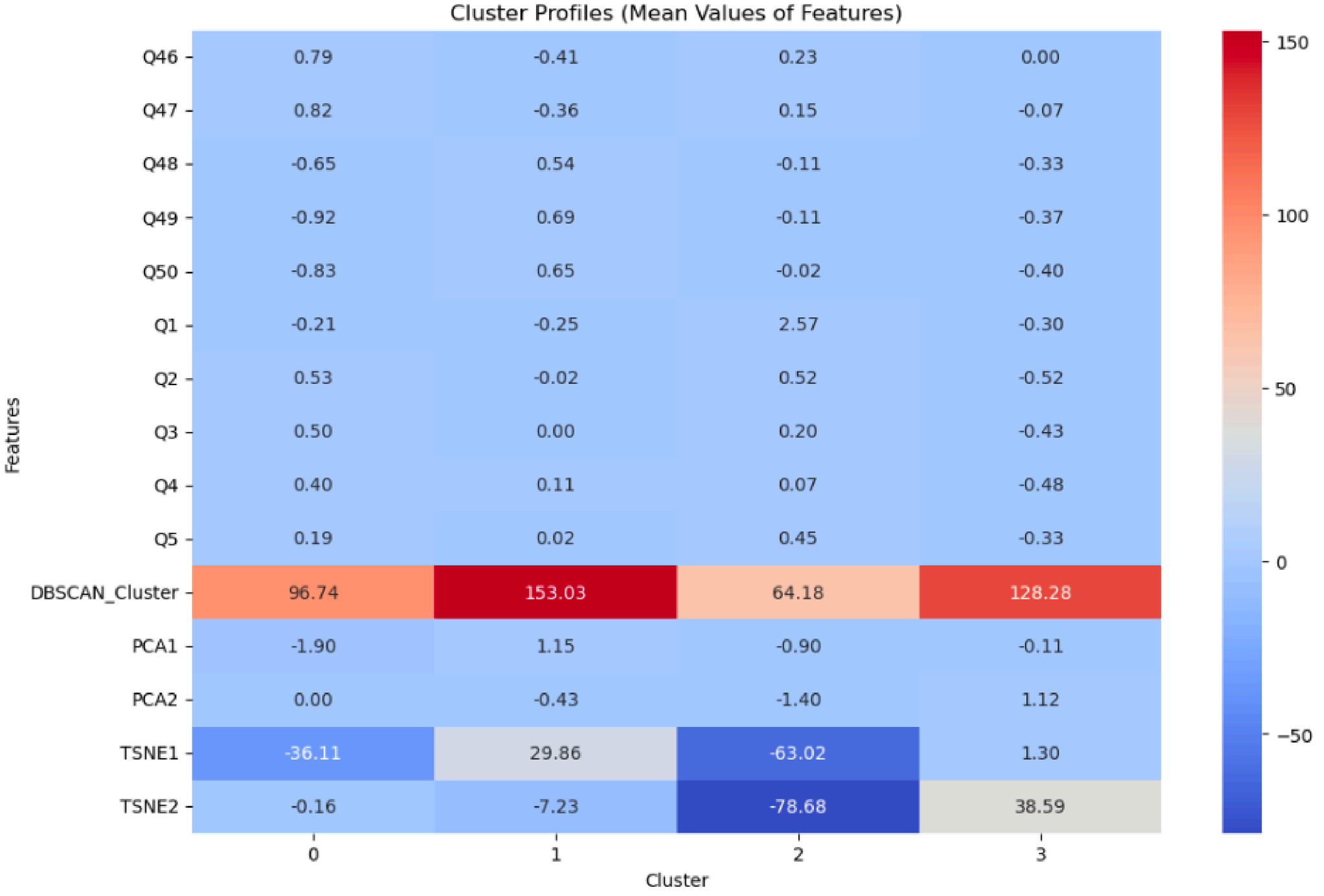
Cluster 1 ("Urban Middle Class")

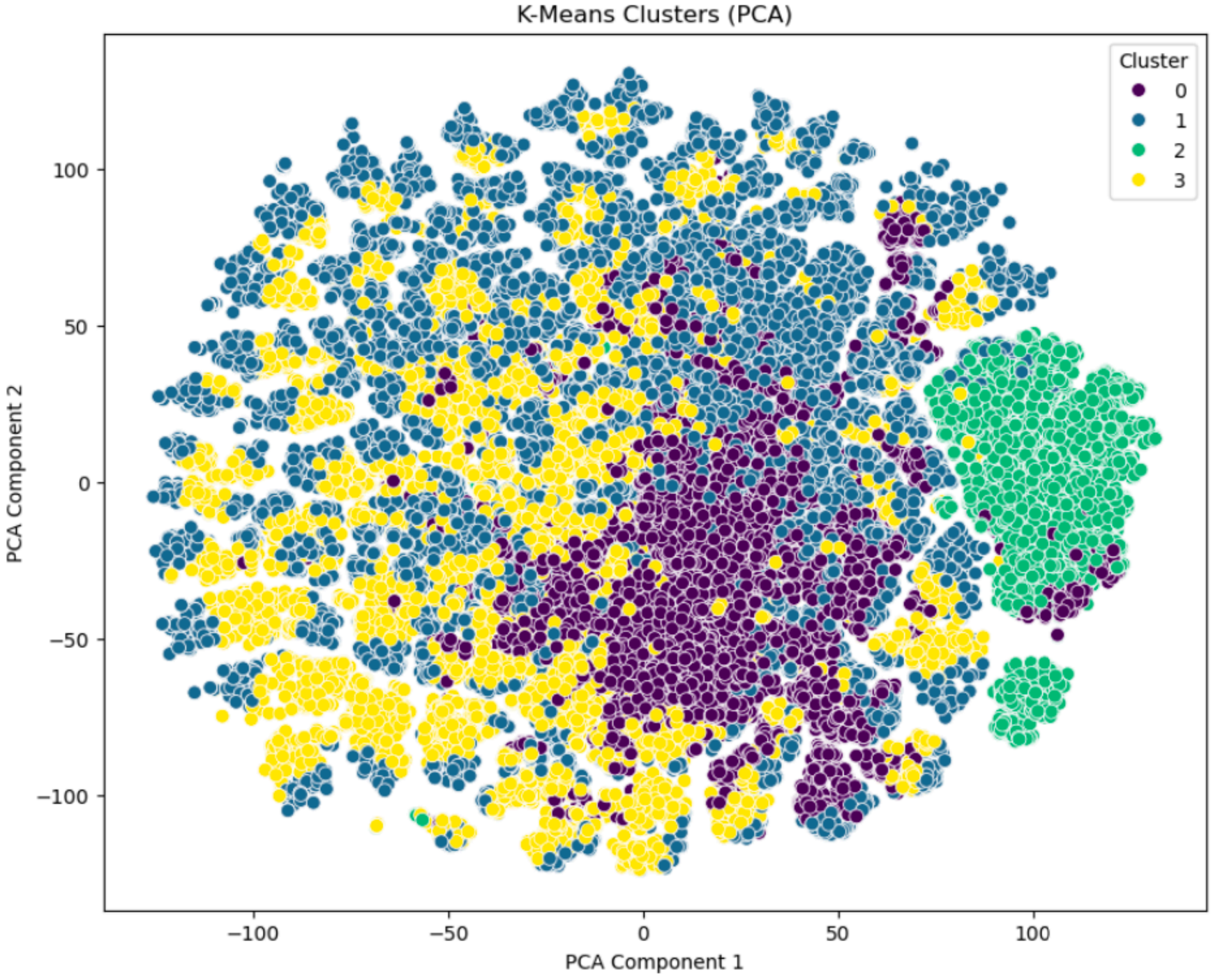
- Moderate religiosity
- Higher income and education
- Mostly urban population.

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.