ML101 Assignment : PCA and K-Means Clustering on Indian District Data

Name: ***KRITIKA SINGH***

Introduction

This report applies unsupervised machine learning techniques to demographic and socio-economic data from Indian districts. The objective is to identify natural groupings among districts using Principal Component Analysis (PCA) for dimensionality reduction and K-Means clustering to detect underlying clusters. All methods were implemented from scratch without the use of machine learning libraries.

Data Preprocessing

The dataset contained 610 rows and 7 columns, including district names, states, population, growth rates, sex ratio, and literacy levels.

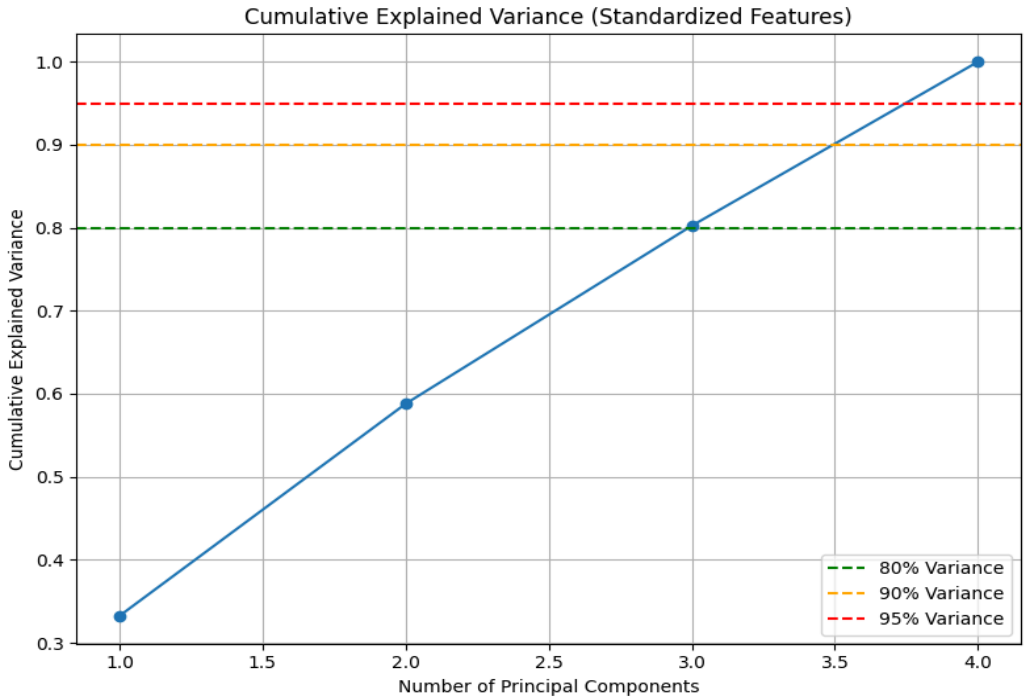
- ***Columns like `District`, `State`, and `Ranking` were excluded from analysis as they are identifiers.***
- ***The `Population` column was cleaned by removing commas, and `Growth` values were stripped of percentage signs.***
- ***Missing values were filled using column-wise means to preserve all rows.***
- ***All features were standardized using z-score normalization to ensure equal contribution to PCA and clustering.***

Principal Component Analysis (PCA)

PCA was applied to the standardized data matrix to reduce its dimensionality while retaining as much variance as possible.

- ***The cumulative explained variance showed that:***
    - ***3 principal components retained 80% of the variance***
    - ***All 4 features retained over 90% of the variance***
    - ***Thus, all 4 features were kept for clustering, while only the first 2 principal components were used for visualization in 2D.***

Cumulative Explained Variance Plot



K-Means Clustering

K-Means clustering was implemented from scratch with support for:

- ***Distance metrics: Euclidean and Manhattan***
- ***Values of K: 5, 7, and 9***
- ***Random initialization of centroids***

The number of iterations to convergence was recorded for each configuration:

K	*Distance Metric*	*Iterations*
5	*Euclidean*	*20*

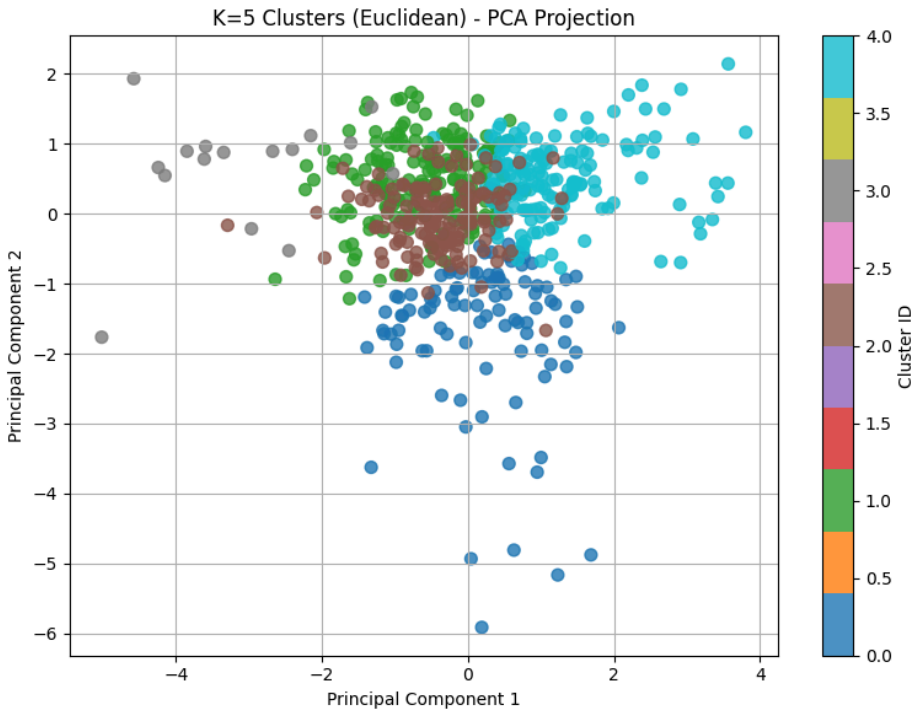
7	***Euclidean***	***23***
9	***Euclidean***	***26***
5	***Manhattan***	***13***
7	***Manhattan***	***34***
9	***Manhattan***	***41***

Visualization and Interpretation

Districts were projected into 2D PCA space using the first two principal components, and scatter plots were generated to visualize cluster assignments.

- ***Euclidean (K = 5, 7, 9)***
- ***Manhattan (K = 5, 7, 9)***

Example: K = 5, Euclidean Distance

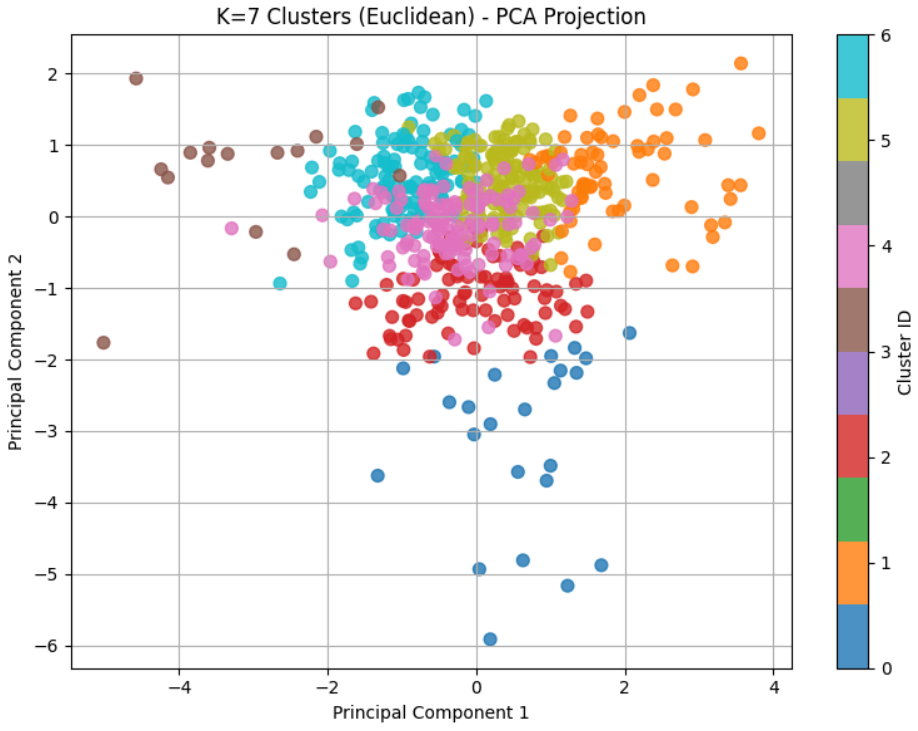


Cluster	*Population*	*Growth*	*Sex-Ratio*	*Literacy*
0	*4.6M*	*19.2%*	*939*	*72.8%*
1	*1.4M*	*22.2%*	*955*	*62.1%*
2	*1.3M*	*17.0%*	*885*	*74.4%*
3	*0.29M*	*88.0%*	*908*	*76.7%*
4	*1.3M*	*11.8%*	*996*	*81.6%*

Cluster Interpretations:

- ***Cluster 0: Large districts with moderate growth and literacy; likely urban centers.***
- ***Cluster 1: Rural, high-growth districts with low literacy.***
- ***Cluster 2: Low sex ratio and moderate literacy — possibly patriarchal northern regions.***
- ***Cluster 3: Very small districts with extreme growth — likely newly formed or administrative splits.***
- ***Cluster 4: High literacy and balanced gender ratio — well-developed districts (e.g., southern states).***

K = 7, Euclidean Distance



Cluster Summary (K = 7, Euclidean)

Cluster	*Population*	*Growth*	*Sex-Ratio*	*Literacy*
0	*6.7M*	*21.52%*	*920*	*81.37%*
1	*1.3M*	*8.87%*	*1029*	*87.36%*
2	*3.75M*	*18.91%*	*945*	*68.96%*
3	*0.29M*	*88.03%*	*908*	*76.71%*
4	*1.35M*	*16.50%*	*882*	*76.05%*
5	*1.29M*	*15.62%*	*971*	*73.37%*
6	*1.25M*	*24.34%*	*943*	*58.91%*

Cluster Interpretations:

Cluster 0: Large, well-developed metro areas with high literacy.

Cluster 1: Highly educated districts with excellent sex ratios and low growth — possibly Kerala or Northeast states.

Cluster 2: Large districts with moderate literacy — possibly industrial or transitioning zones.

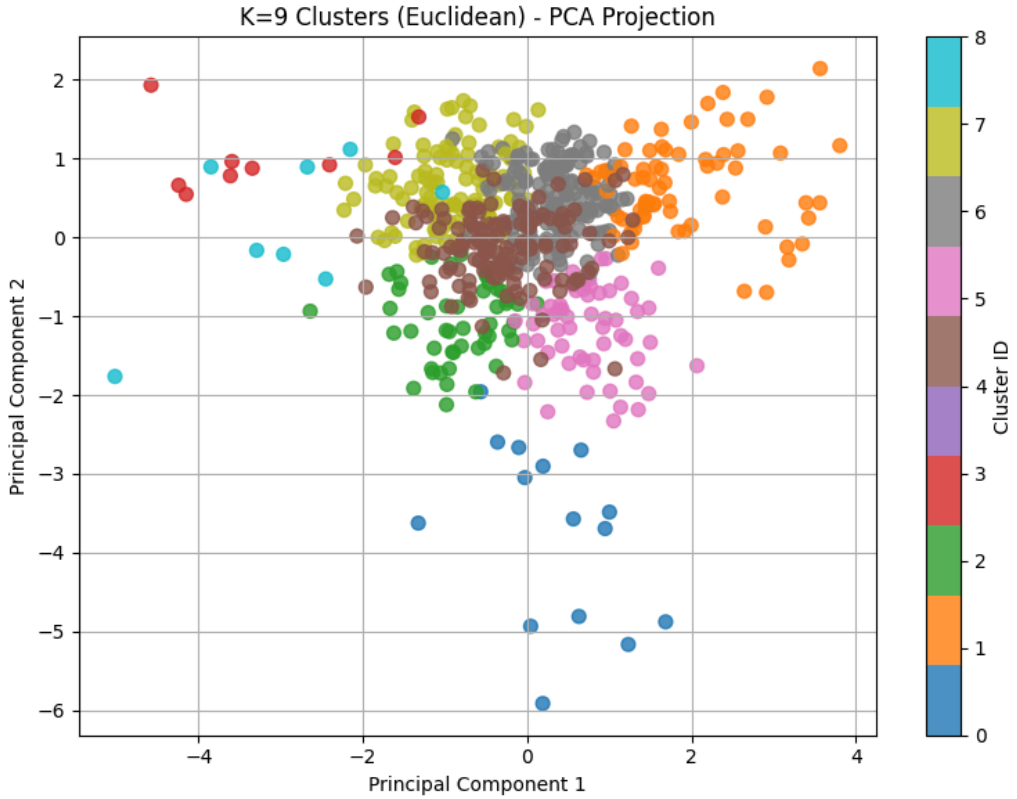
Cluster 3: Very small, explosive-growth districts — likely new administrative regions.

Cluster 4: Low sex ratio, moderate development — patriarchal regions.

Cluster 5: Balanced and moderate literacy — semi-urban growth areas.

Cluster 6: High growth, very low literacy — underdeveloped rural districts.

K = 9, Euclidean Distance



Cluster Summary (K = 9, Euclidean)

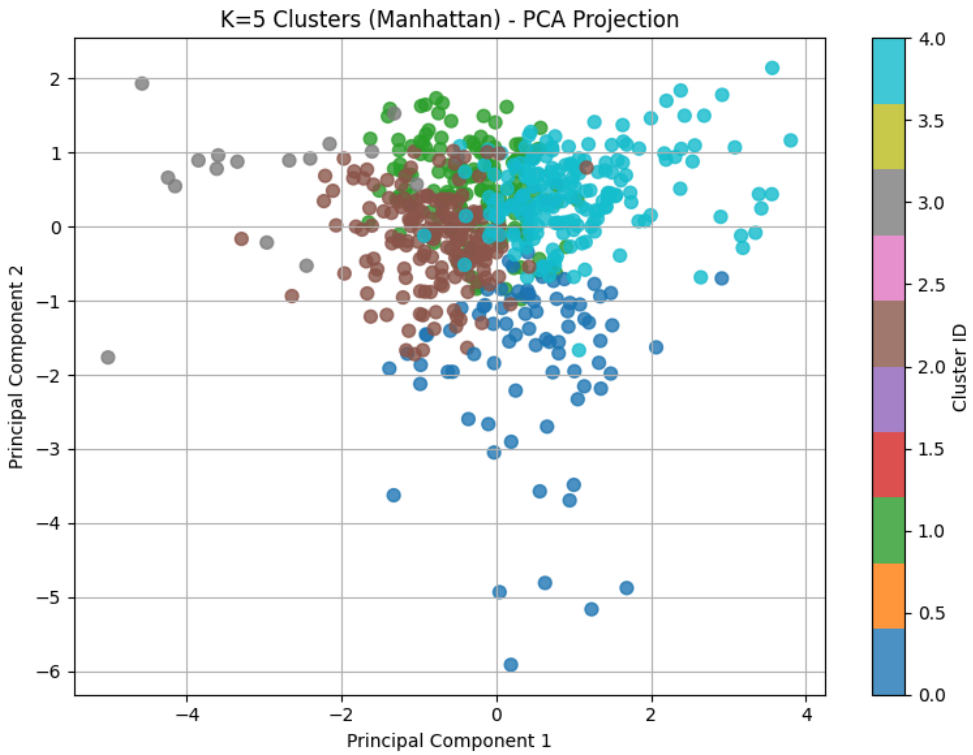
Cluster	*Population*	*Growth*	*Sex-Ratio*	*Literacy*
0	*7.7M*	*26.16%*	*912*	*80.01%*
1	*1.2M*	*8.65%*	*1027*	*87.38%*

2	*3.6M*	*24.34%*	*912*	*62.89%*
3	*0.17M*	*104.5%*	*962*	*77.81%*
4	*1.34M*	*16.30%*	*883*	*76.04%*
5	*3.9M*	*15.14%*	*967*	*76.06%*
6	*1.3M*	*16.10%*	*969*	*72.71%*
7	*1.1M*	*23.89%*	*947*	*58.74%*
8	*0.39M*	*62.04%*	*823*	*73.26%*

Cluster Interpretations:

- ***Cluster 0: Very large, growing cities with high literacy.***
- ***Cluster 1: Stable, educated, and gender-balanced districts.***
- ***Cluster 2: High-growth zones with lower education — emerging metros or industrial belts.***
- ***Cluster 3: Very small and extremely fast-growing — possible administrative splits.***
- ***Cluster 4: Low sex ratio and modest literacy — patriarchal rural areas.***
- ***Cluster 5: Large, balanced regions — well-developed tier-2 cities.***
- ***Cluster 6: Mid-sized districts with above-average sex ratio and literacy.***
- ***Cluster 7: High-growth, undereducated regions — developing rural zones.***
- ***Cluster 8: Low literacy and poor gender ratio — likely isolated or tribal regions.***

K = 5, Manhattan Distance



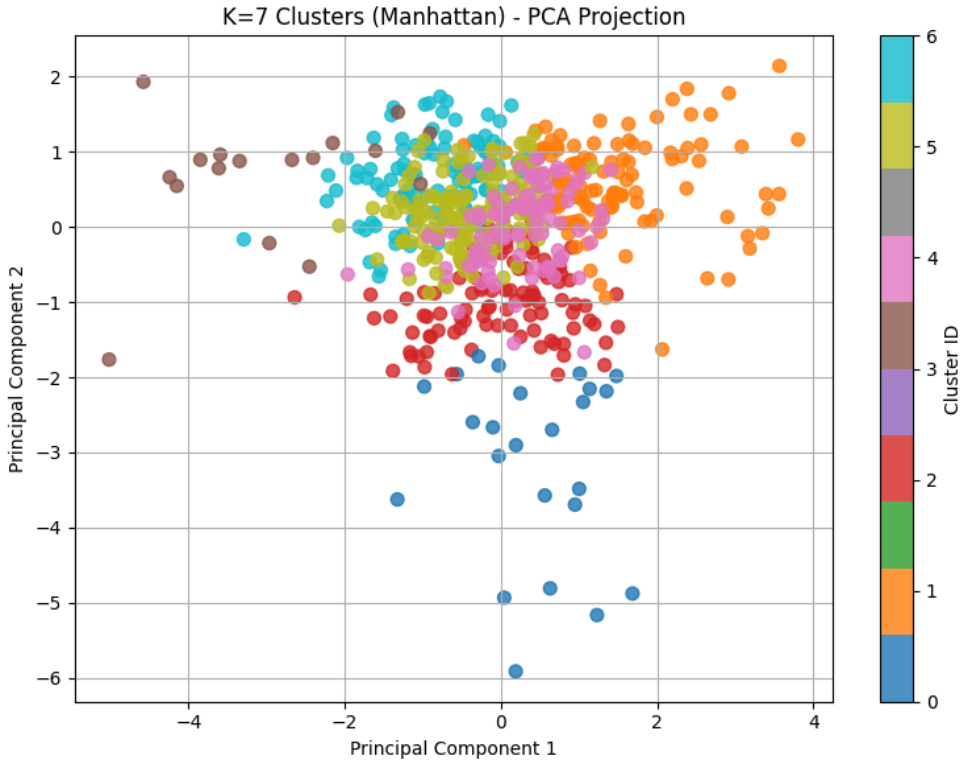
Cluster Summary (K = 5, Manhattan)

Cluster	*Population*	*Growth*	*Sex-Ratio*	*Literacy*
0	*4.8M*	*19.23%*	*944*	*75.89%*
1	*1.5M*	*20.41%*	*977*	*62.70%*
2	*1.7M*	*21.22%*	*892*	*67.88%*
3	*0.29M*	*88.03%*	*907*	*76.71%*
4	*1.18M*	*11.67%*	*972*	*81.50%*

Cluster Interpretations:

- ***Cluster 0: Large and moderately developed districts — older metros or capitals.***
- ***Cluster 1: High sex ratio and low literacy — possibly rural female-balanced states.***
- ***Cluster 2: Low sex ratio and moderate literacy — northern patriarchal regions.***
- ***Cluster 3: Small, high-growth districts — newly formed or reorganized.***
- ***Cluster 4: Highly literate and balanced — possibly southern, developed districts.***

K = 7, Manhattan Distance



Cluster Summary (K = 7, Manhattan)

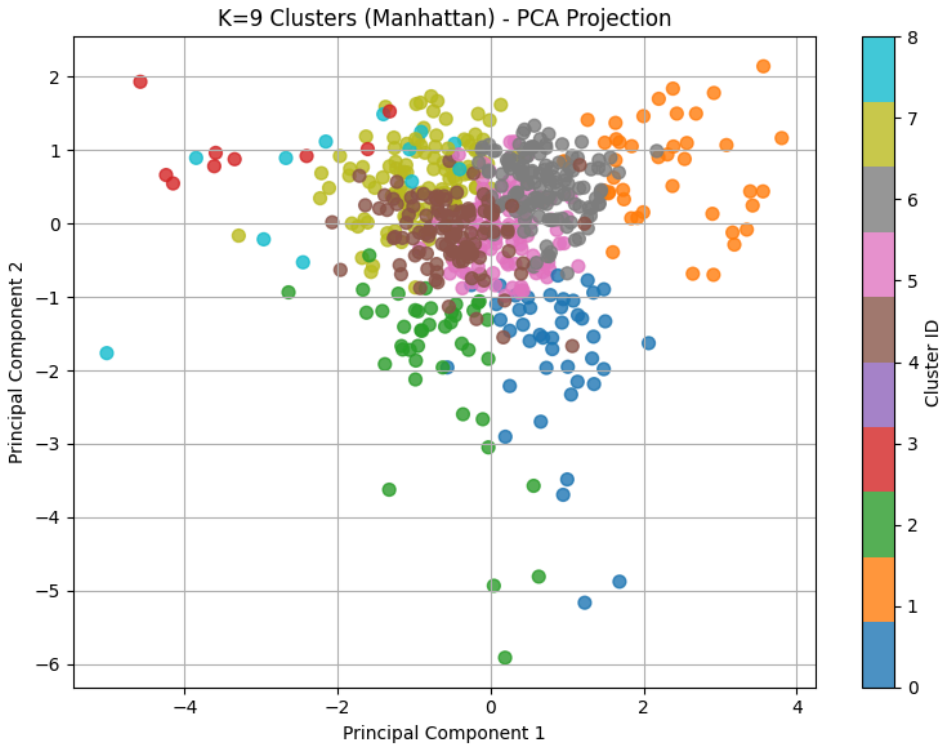
Cluster	*Population*	*Growth*	*Sex-Ratio*	*Literacy*
0	*6.6M*	*22.86%*	*913*	*80.67%*

1	*1.4M*	*10.75%*	*1016*	*82.34%*
2	*3.7M*	*19.11%*	*948*	*68.83%*
3	*0.28M*	*85.73%*	*912*	*76.57%*
4	*1.2M*	*14.61%*	*904*	*81.37%*
5	*1.3M*	*18.61%*	*919*	*69.43%*
6	*1.2M*	*23.31%*	*955*	*58.06%*

Cluster Interpretations:

- ***Cluster 0: Large, educated metros with balanced demographics.***
- ***Cluster 1: Top-performing districts in literacy and gender equality.***
- ***Cluster 2: High population but literacy lagging — industrial or semi-urban zones.***
- ***Cluster 3: Very small, fast-growing districts — administrative expansions.***
- ***Cluster 4: Educated and demographically stable mid-sized districts.***
- ***Cluster 5: Average literacy and development — suburban or tier-2 regions.***
- ***Cluster 6: Underdeveloped, high-growth districts — likely backward rural states.***

K = 9, Manhattan Distance



Cluster Summary (K = 9, Manhattan)

Cluster	*Population*	*Growth*	*Sex-Ratio*	*Literacy*
0	*4.9M*	*15.19%*	*962*	*76.97%*
1	*1.4M*	*7.14%*	*1060*	*87.83%*
2	*4.7M*	*26.96%*	*899*	*67.96%*
3	*0.17M*	*104.5%*	*962*	*77.81%*

4	*1.3M*	*16.94%*	*873*	*73.58%*
5	*2.3M*	*17.31%*	*962*	*70.47%*
6	*1.0M*	*12.76%*	*963*	*78.74%*
7	*1.2M*	*24.12%*	*949*	*59.01%*
8	*0.3M*	*57.58%*	*892*	*75.15%*

Cluster Interpretations:

- ***Cluster 0: Large, stable, literate districts with balanced demographics.***
- ***Cluster 1: Excellent gender balance and top-tier literacy — leading performers.***
- ***Cluster 2: Rapid-growth, lower-literacy districts — urban pressure zones.***
- ***Cluster 3: Very small, new districts growing rapidly — special attention zones.***
- ***Cluster 4: Poor gender ratio, mid-level development — needs social intervention.***
- ***Cluster 5: Mid-population, average literacy — stable semi-urban areas.***
- ***Cluster 6: Small and fairly literate districts — well-governed smaller towns.***
- ***Cluster 7: High-growth, low-literacy — classic backward zones.***
- ***Cluster 8: Small districts with fast growth and low sex ratio — neglected rural areas.***

Conclusion

The application of PCA and K-Means clustering revealed meaningful patterns in the socio-economic structure of Indian districts:

- ***PCA allowed us to reduce feature space while retaining essential variance.***
- ***K-Means clustering exposed distinct groupings based on growth, literacy, population, and sex ratio.***
- ***Cluster interpretations helped identify progressive regions, rural lagging districts, urban centers, and administrative outliers.***
- ***The results can assist in targeted development planning and resource allocation.***