Report: K-Means based Clustering (unsupervised Learning)

Dataset Overview

The World Bank dataset included a range of social indicators for , such as: United States of America, China, India, Brazil, Russia, Canada, France, Japan, Germany and United Kingdom

Social Indicators used:

1. GDP per capita (code: NY.GDP.PCAP.CD)
2. Gini Coefficient (code: SI.POV.GINI)
3. Poverty Headcount Ratio at $1.90/day (code: SI.POV.DDAY)
4. Unemployment Rate (code: SL.UEM.TOTL. ZS)
5. Debt-to-GDP Ratio (code: GC.DOD.TOTL.GD.ZS)

Data Preprocessing

1. Handling missing data through imputation or exclusion.
2. Standardizing the features due to the varied scales of social indicators.

Methodology

Clustering Without PCA

In the first approach, the K-Means clustering algorithm was applied directly to the original dataset without reducing the number of features.

Clustering With PCA (n = 2)

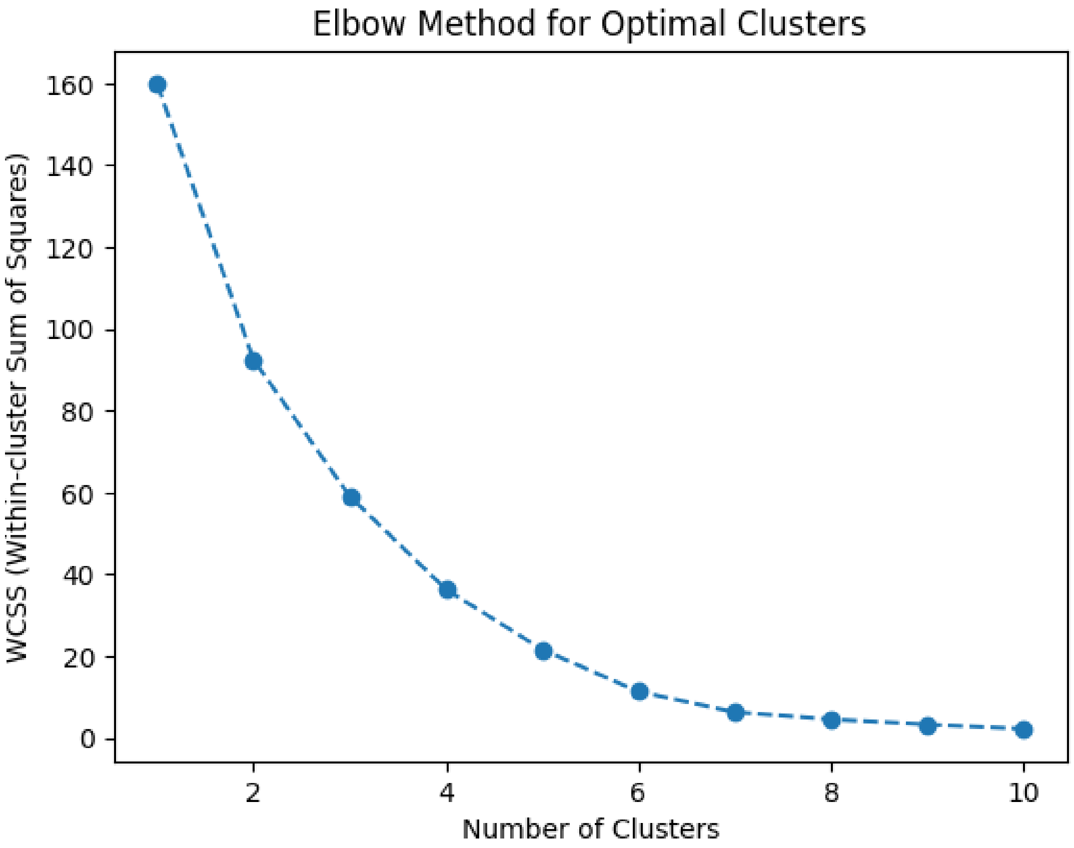
For the second approach, PCA was used to reduce the dataset to 2 principal components before applying K-Means clustering. PCA helps to capture the most significant variance in the data while reducing the dimensionality, which can improve the clustering performance and visualization.

Determining the Optimal Number of Clusters

To determine the optimal number of clusters, Elbow method was applied:

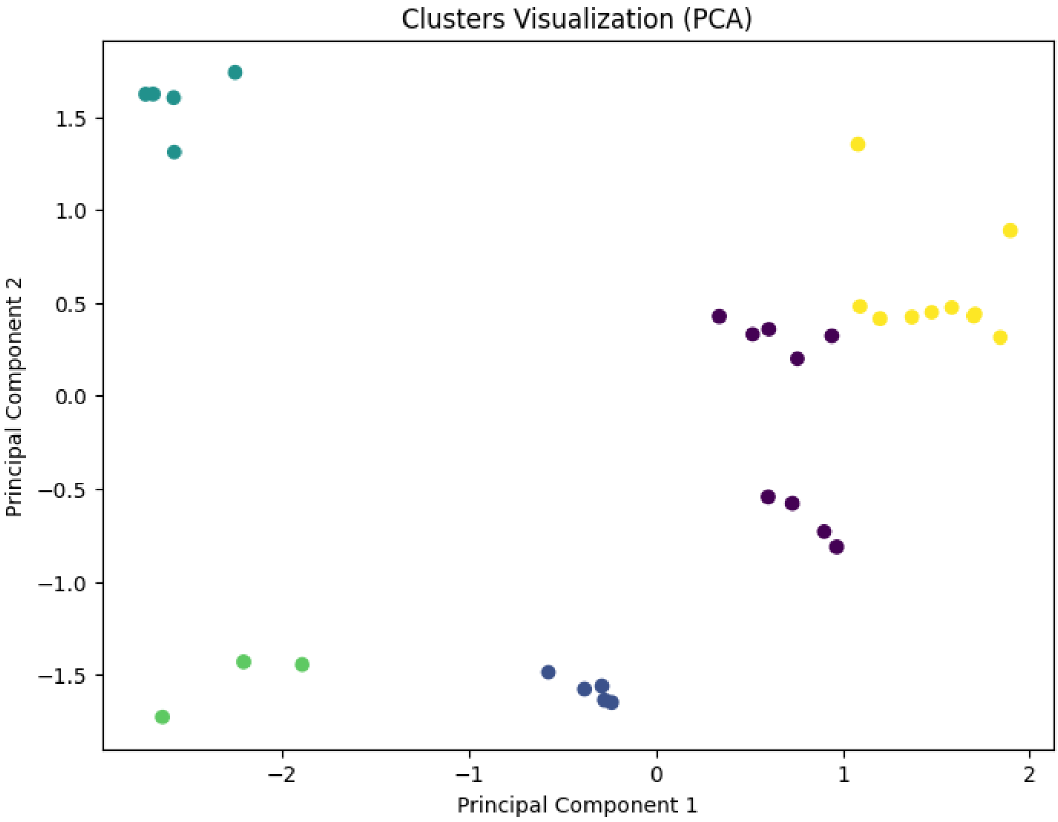
- Elbow Method: This method helps identify the number of clusters were adding more clusters no longer provides a significant reduction in the within-cluster sum of squares

(WCSS). The "elbow" in the plot indicates the point of diminishing returns, which is considered the optimal number of clusters.



Observations

1. **Cluster Composition**: The composition of the clusters was similar between the two methods. Countries like [mention any specific examples] remained in the same cluster regardless of PCA use, while others like [mention examples] shifted between clusters.
2. **Visualization**: Clustering after PCA provides a more interpretable 2D visualization, helping to see the relationships between countries based on the most significant components of their socioeconomic indicators.



Conclusion

The comparison between clustering with and without PCA demonstrates the trade-offs:

1. Without PCA, clustering uses all the original features, which may capture more detailed but potentially redundant information, leading to higher-dimensional and less interpretable clusters.
2. With PCA, dimensionality reduction simplifies the feature space, making the clusters more interpretable and visually distinct while retaining the core structure of the data. In this case, using PCA provided clearer cluster separation and easier visualization without significant loss of information.