Unsupervised Learning: Dimensionality Reduction(PCA) and Clustering Analysis of Water-Economic Indicators

1. Introduction

This section presents an unsupervised learning approach to grouping countries based on selected water and economic indicators. Clustering was performed using K-means, preceded by feature reduction via Principal Component Analysis (PCA). The analysis is based on a curated subset of features, derived and filtered through careful preprocessing steps. In order to reduce short-term fluctuations and place the focus on the recent years, all selected features were averaged over 2012–2021 before analysis.

2. R Packages and Tools Used

Package	Purpose
cluster	Computing silhouette scores and clustering utilities
tidyverse	Data wrangling and visualization
factoextra	PCA and cluster visualization
clusterSim	Calculating the Davies-Bouldin index for cluster evaluation
rnaturalearth & rnaturalearthdata	Mapping countries onto geographic clusters

3. Feature Selection Strategy

3.1. Initial Exclusion

The following indicators were excluded due to their redundancy, as they are directly involved in or derived from other features through the following relationships:

Key relationships:

- Water Productivity = GDP (USD) / Freshwater Withdrawals
- Water Stress = Freshwater Withdrawals / Available Resources × 100
- GDP per capita(PPP) = GDP (PPP) / Population
- Industrial + Agricultural + Domestic = 100

Excluded indicators based on the above:

- Freshwater Withdrawals (total)
- Water Stress

- GDP (PPP)
- Industrial Withdrawals

4. Principal Component Analysis (PCA)

PCA was conducted to reduce the dimensionality of the feature space and understand the contribution of each feature to overall variance.

4.1. Input Features

The following seven features have been averaged from 2012 to 2021 and were initially included in PCA:

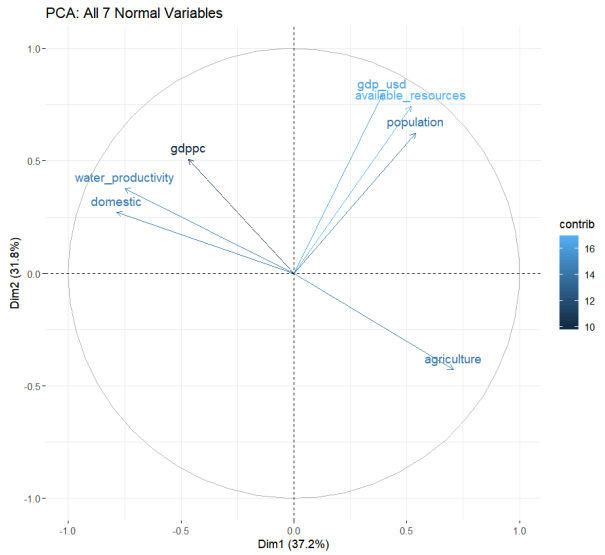
- GDP (USD)
- Available Water Resources
- Water Productivity
- Agricultural Withdrawals
- Domestic Withdrawals
- GDP per Capita
- Population

4.2. PCA Result and Interpretation

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.6143 1.4918 0.9052 0.75182 0.59191 0.51748 0.40698
Proportion of Variance 0.3723 0.3179 0.1171 0.08075 0.05005 0.03826 0.02366
Cumulative Proportion  0.3723 0.6902 0.8073 0.88803 0.93808 0.97634 1.00000
```

The PCA revealed that the first two principal components explain a substantial portion of the variance in the data, with PC1 accounting for 37.2% and PC2 for 31.8%, resulting in a cumulative explained variance of 69.0%. This indicates that a two-dimensional representation retains most of the original information and therefore justifies the use of a 2D PCA contribution plot for interpretation. The variance explained drops sharply after PC2, with PC3 contributing only 11.7% and later components each explaining less than 10%, confirming that higher dimensions add limited new information.

The 2D PCA contribution plot and the corresponding PCA output are shown below.

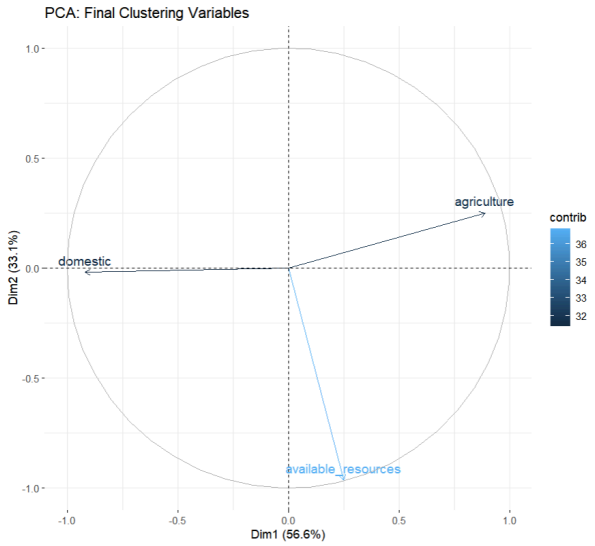


	PC1	PC2	PC3	PC4	PC5
gdp_usd	0.2434258	0.5318134	-0.09015836	-0.2027767758	0.5564709
available_resources	0.3216155	0.4973317	0.08416974	-0.1166244284	0.1235007
water_productivity	-0.4634469	0.2524688	-0.04520644	0.6371018221	0.2481157
agriculture	0.4377296	-0.2868088	-0.20117415	0.6079249516	0.3530565
domestic	-0.4861675	0.1815890	0.52604908	0.0423873636	0.1808006
gdppc	-0.2885806	0.3401951	-0.76090148	0.0003522581	-0.2740718
population	0.3335860	0.4184837	0.29423386	0.4098927063	-0.6173113

	PC6	PC7
gdp_usd	-0.52556540	-0.15097227
available_resources	0.76357640	0.17388943
water_productivity	0.19797229	-0.46128055
agriculture	-0.03888929	0.43580784
domestic	-0.13834188	0.63270443
gdppc	-0.05804661	0.37885691
population	-0.27842199	-0.02035124

It might be unclear which axis represents PC1 and PC2 in the plot. However, one can verify from the PCA output that the horizontal axis corresponds to PC1 (Dim1: 37.2%) and the vertical axis to PC2 (Dim2: 31.8%). Strong drivers of PC1 are Agricultural Withdrawals, Water Productivity, and Domestic Withdrawals. On the other hand, strong drivers of PC2 are GDP (USD), Available Resources, and Population.

According to the 2D PCA contribution plot, Water Productivity, Domestic Withdrawals, and GDP per Capita show overlapping patterns, suggesting some redundancy. A similar overlap is observed among GDP (USD), Available Resources, and Population.
Based on this observation, the features GDP (USD), GDP per capita, Water Productivity, and Population were excluded from clustering to reduce redundancy. The 2D PCA contribution plot after excluding those features is shown below.



The three features are well-separated in the plot, indicating that they contribute distinct information and are suitable for clustering.

5. Clustering Methodology

5.1. Final Feature Set

The final clustering was based on the following features:

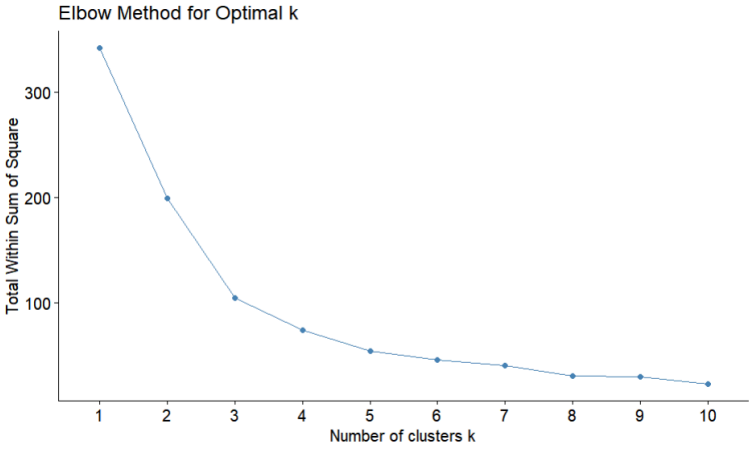
- Available Water Resources
- Agricultural Withdrawals
- Domestic Withdrawals

```
# Standardize feature data
clustering_scaled <- scale(clustering_data)
```

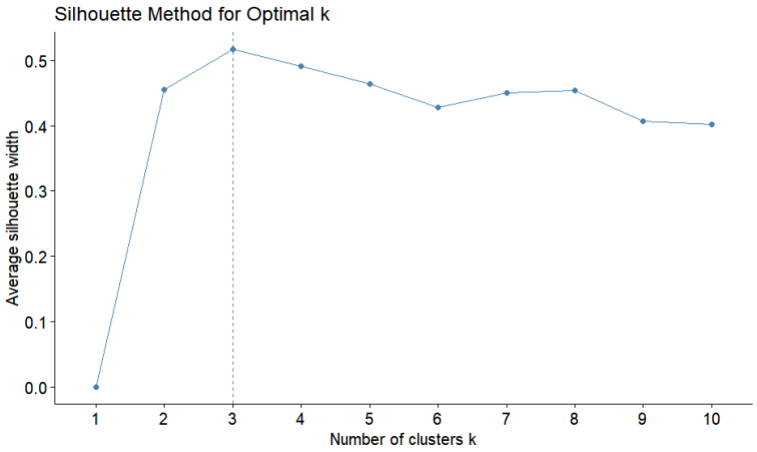
Features were standardized before clustering to eliminate scale differences and prevent any single variable from dominating the clustering outcome.

5.2. K-means Clustering

Initially, **Elbow Method** was used to determine the optimal number of clusters.



In theory, the optimal k appears where the elbow graph shows a sharp drop in slope followed by a plateau. However, no clear elbow point was found in the graph. Therefore, the silhouette method was used instead to determine the number of clusters.



Final number of clusters: **k = 3**

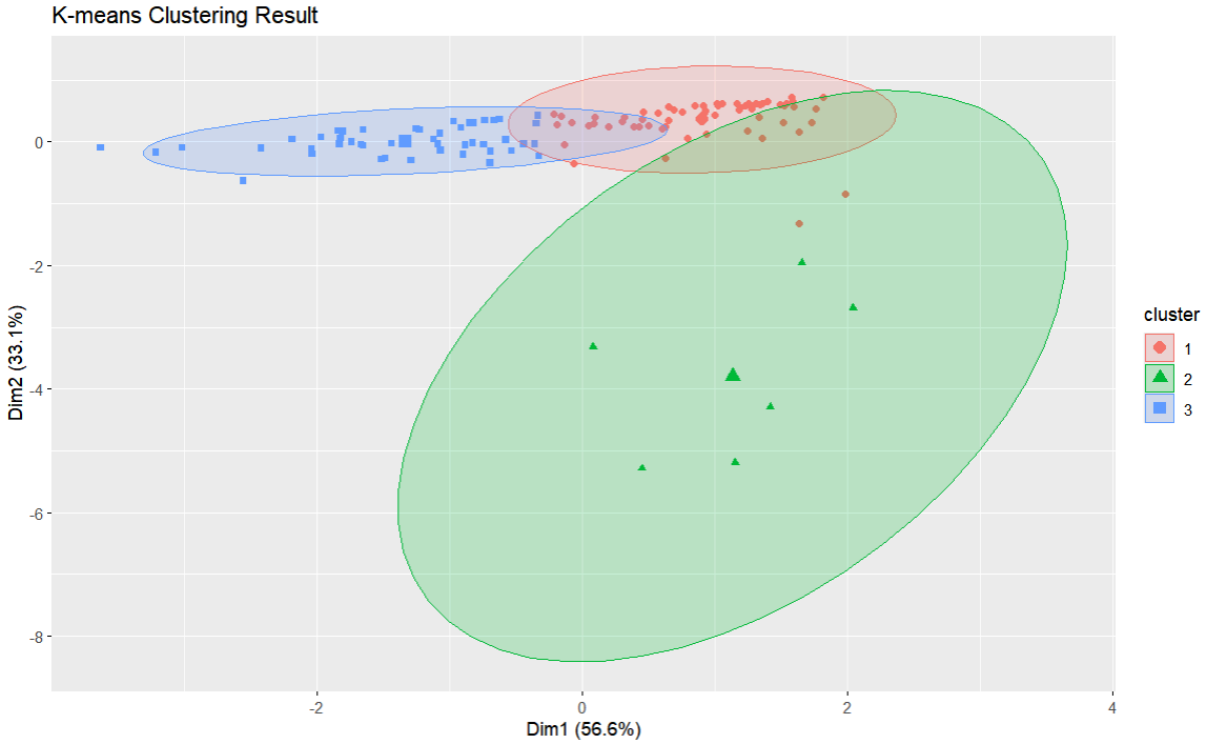
```
# Save the final dataset with cluster assignments
setwd("C:/Users/SEC/Desktop/2024WS/projectR/dataset/clustering")
write.csv(clustering_dataset, "clustering_result.csv", row.names = FALSE)

# Save the kmeans model result (for reproducibility or later analysis)
saveRDS(k_result, "kmeans_result_k3.rds")
```

To ensure reproducibility and allow for future use of the clustering results, the final dataset and the K-means model object were saved.

6. Results and Interpretation

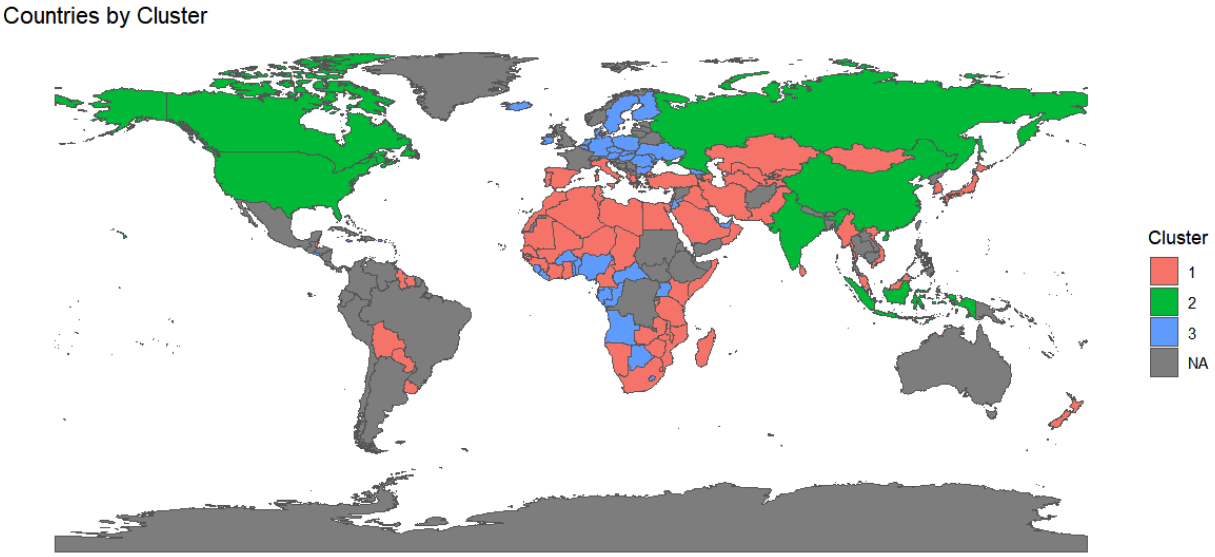
6.1 Visualization



The figure shows the result of K-means clustering projected onto the first two principal components (PC1 and PC2), which together explain 89.7% of the variance. Each point represents a country, colored and shaped by its assigned cluster.

- Cluster 1 (red) is tightly grouped, indicating relatively low internal variance.
- Cluster 2 (green) has more spread, suggesting greater internal diversity.
- Cluster 3 (blue) is tightly grouped and mostly separated along the horizontal (PC1) axis.

The separation of clusters in PCA space suggests that the clustering captures underlying structural differences in the selected features.



In addition to the PCA projection, a geographic map of cluster assignments provides further insight into regional patterns.

- Cluster 1 (red) appears predominantly across Africa, the Middle East, South Asia, and parts of Latin America. These regions may share common characteristics in terms of water use or availability, potentially shaped by agricultural activity or climatic factors.

- Cluster 2 (green) includes countries such as Canada, Russia, China, the USA, Indonesia, and India, which could indicate the presence of abundant natural water resources or distinct usage behaviors.

- Cluster 3 (blue) is more scattered, with countries appearing in Europe, Africa, and some parts of the Middle East. These may represent countries facing water constraints or following unique domestic usage patterns.

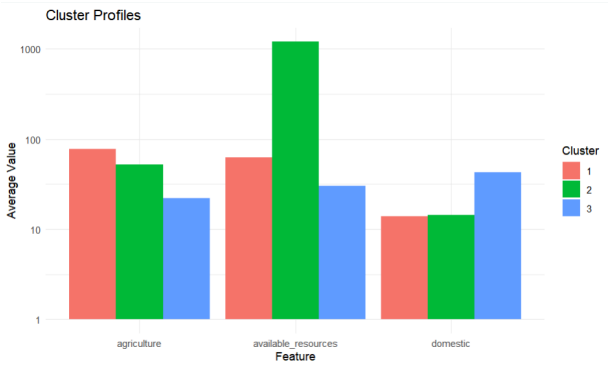
- Countries shown in gray were previously excluded from the clustering due to missing data. The missingness is especially concentrated in parts of Africa, Central & Latin America, and several island nations, reflecting regional disparities in data availability and reporting capacity.

While no assumptions about the exact drivers can be made from geography alone, this spatial distribution suggests that the clustering reflects more than random separation — and warrants further exploration through feature-based profiling in the following section.

6.2. Cluster Profiles

	Cluster	Feature	Mean
1	1	available_resources	62.50049
2	1	agriculture	77.24471
3	1	domestic	13.84198
4	2	available_resources	1202.36083
5	2	agriculture	52.32259
6	2	domestic	14.33468
7	3	available_resources	30.32449
8	3	agriculture	22.09220
9	3	domestic	43.05617

Average feature values by cluster(Table)



Cluster Profiles (Log-Scaled, Barplot)

The table and bar plot above shows the average values of the three selected features — Available Water Resources, Agricultural Withdrawals, and Domestic Withdrawals — for each cluster. These profiles provide insight into the distinct characteristics that define each group:

Cluster	Characteristics

Cluster1	Characterized by moderate available water resources (62.5) and the highest agricultural withdrawals (77.2) among all clusters, this group represents an agriculture-intensive profile with limited but usable water supply. Geographically, these countries are concentrated in Africa, the Middle East, South Asia, and parts of Latin America, suggesting regions where agriculture remains central to the economy and water use is quite skewed toward that sector. Climatic factors and irrigation needs may also contribute to this pattern.
Cluster2	This cluster exhibits extremely high available resources (1202.4) alongside relatively low agricultural (52.3) and domestic (14.3) water use, pointing to resource-rich but low-usage countries. It includes large nations such as Canada, Russia, China, the USA, Indonesia, and India. While some of these are highly populated or agriculturally active, their inclusion here may reflect national-scale water abundance, possibly buffered by regional variation or strong infrastructure.
Cluster3	With low available resources (30.3) and low agricultural withdrawals (22.1) but the highest domestic use (43.1), this cluster likely reflects water-stressed countries with high urban or household demand. Geographically, these nations are scattered across Europe, Africa, and parts of the Middle East, suggesting a mix of urbanized societies, developed infrastructure, and limited natural water availability, especially for agriculture.

This interpretation provides context for the clustering outcome and forms the basis for later discussion, including potential policy recommendations informed by the distinct regional water-use patterns and resource profiles of each cluster.

6.3. Discussion on Policy Implications

Based on the distinct profiles and regional distribution of the clusters, several policy-relevant insights emerge. While specific strategies must be tailored to each country's context, the following implications offer a general direction for water resource management within each cluster.

6.3.1. Cluster 1: Agriculture-Dependent, Moderately Resourced

Countries in this cluster rely quite heavily on agriculture and have moderate water availability. High agricultural withdrawals make them particularly vulnerable to water inefficiencies and climate variability. Therefore, relevant policy directions include:

- Investing in irrigation efficiency, such as drip systems or precision agriculture, to reduce water loss and improve productivity.

- Supporting farmer education and incentives to adopt water-saving practices.

- Improving monitoring of agricultural water use, especially in drought-prone or groundwater-dependent regions.

These measures aim to balance agricultural demand with sustainable water use under increasing environmental stress.

6.3.2. Cluster 2: Water-Abundant, Low Usage

Countries in this cluster have abundant water resources and relatively low current usage. However, this should not lead to complacency. To ensure long-term sustainability, the following policy directions are relevant:

- Sustainable resource management should be maintained to protect existing reserves and prevent overexploitation as demand evolves.

- Proactive development of water infrastructure is needed to accommodate potential increases in water demand from urban expansion, industrial growth, or regional population shifts.

- Environmental protection policies are important to safeguard water quality in large river systems, lakes, and underground aquifers.

With careful planning, countries in this cluster can serve as models for balanced, forward-looking water governance.

6.3.3. Cluster 3: Water-Stressed, Urbanized Demand

Countries in this cluster face low water availability but high domestic demand, often driven by urbanization and population pressure. Managing consumption and improving infrastructure are critical priorities. Policy directions may include:

- Improving urban water supply resilience, through methods such as rainwater harvesting, wastewater reuse, or smart distribution systems.

- Implementing demand-side measures, such as tiered pricing and public conservation campaigns.

- Upgrading infrastructure to reduce leakage and improve delivery efficiency, particularly in densely populated areas.

These measures aim to address the mismatch between limited resources and growing household water needs.

Overall, this clustering framework highlights how data-driven groupings can guide region-specific policy interventions, ensuring that water governance aligns with actual usage patterns and resource constraints.

6.4. Validation

To assess the quality of the clustering result, two internal validation metrics were used: the silhouette score and the Davies-Bouldin (DB) index.

6.4.1. Silhouette score

```
> # Calculate average silhouette score
> sil <- silhouette(k_result$cluster, dist(clustering_scaled))
> mean(sil[, 3])  # Print silhouette score
[1] 0.5169572
```

The average silhouette score was 0.5170, indicating a moderate to good clustering structure. This suggests that, on average, countries are well matched to their assigned clusters and reasonably well separated from others, with low overlap between cluster boundaries.

6.4.2. Davies-Bouldin (DB) index

```
> # Calculate Davies-Bouldin Index
> db_index <- index.DB(clustering_scaled, k_result$cluster)$DB
> print(db_index)
[1] 0.7319856
```

The Davies-Bouldin index was 0.7320, where lower values indicate better-defined and more compact clusters. A DB index below 1 is generally considered acceptable, and the result here confirms that the clusters are distinct yet internally cohesive.

7. Conclusion

This study applied an unsupervised learning approach to group countries based on selected water and economic indicators, using Principal Component Analysis (PCA) for dimensionality reduction, followed by K-means clustering. Three key features — available water resources, agricultural withdrawals, and domestic withdrawals — were selected as the basis for clustering.

The analysis identified three distinct clusters, each reflecting unique patterns of water availability and usage. These groups also showed meaningful geographic and structural differences, further explored through global mapping.

Cluster profiles showed how different combinations of water availability and usage are linked to common country characteristics — such as a reliance on agriculture, large natural water reserves, or high urban water demand. Based on these patterns, policy implications were proposed to help align water governance with the practical needs and constraints of each group.

Overall, the results demonstrate that unsupervised learning methods can support more informed, data-driven decisions in global water management and policy development.