STAT6020 Predictive Analytics Project (20%)

Global Climate Patterns from World Bank Indicators: A PCA & K-Means Exploration

Author: Kazi Shabab Mahfuz
Course: STAT6020 Predictive Analytics
Supervisor: Dr Hongsheng Hu

Abstract

This study examines climate-related indicators from the World Bank to discern global patterns and classifications among nations. Each record in the cross-sectional dataset (wbcc_bc.csv) shows the most recent climate indicator values for a country from 2001 to 2020. The analysis uses Principal Component Analysis (PCA) to lower the number of dimensions and K-Means clustering to find structural groupings. We dealt with missing data by filtering columns and rows and filling in the median. We also made sure that all numeric features were the same.

Results show that the first two principal components explain 35.23 % of total variance, and 23 components capture approximately 90 %. Based on elbow and silhouette analysis, K = 2 clusters offered the best partition. The clusters differ mainly across population scale, land area below 5 m elevation, and greenhouse-gas growth indicators. Australia is present in the dataset (country code: AUS) and is assigned to Cluster 0, which corresponds to larger, industrialised economies.

The PCA–K-Means pipeline provides an interpretable map of climate-indicator variation worldwide, revealing two broad country archetypes separated by size, exposure, and emissions trends. Although purely descriptive, such groupings help benchmark nations for sustainability initiatives and highlight variables driving divergence.

Introduction

Understanding how climate-change indicators vary across countries is vital for evidence-based global policy. The World Bank collates multiple environmental, economic, and demographic variables (e.g., CO_2 emissions, land-use, energy intensity) that jointly describe a nation's climate profile. The aim of this project is to:

1. Reduce the high-dimensional indicator space into interpretable components via PCA.
2. Cluster countries into homogeneous groups using K-Means.
3. Interpret the resulting clusters through indicator means and dominant features.
4. Compare individual countries—particularly Australia if present—to their peers.

This work applies at least two major techniques (dimensionality reduction + unsupervised clustering) with quantitative and visual evaluation.

Data

Source: World Bank Climate Change Indicators. Each row represents a country, and each column represents the latest available value (2001–2020). The dataset contains approximately 217 countries × 79 indicators initially.

Data Preprocessing:
- Dropped columns >30% missing and rows >30% missing.
- Median imputation for remaining gaps.
- Standardized all numeric features.

Final matrix: 188 countries × 60 features.
Limitations: Mixed years and missing values are known limitations in this dataset.

Methods

1. Principal Component Analysis (PCA)

PCA was applied to the standardised data to identify orthogonal directions of maximum variance. A scree plot was used to determine the number of principal components required for 90 % of the total variance (~23 PCs). The first two components (PC1 and PC2) explained 35.23 % of the variance and were visualised in a 2D scatter plot.

2. K-Means Clustering

K-Means clustering was applied to the 2D PCA embedding. The number of clusters k was selected using elbow (inertia) and silhouette analyses across $k \in [2, 8]$, where the highest silhouette score occurred at k = 2.

3. Cluster Profiling

Mean indicator values per cluster were computed from the imputed data. The 15 indicators with the highest variance across clusters were used for interpretation.

Results and Discussion

1. **PCA Findings**
    - The first two principal components explain 35.23 % of the total variance.
    - Approximately 23 components are required to capture 90 % of the total variance.
    - The scree plot demonstrates diminishing returns after 20 components.
    - PC1 represents scale and emissions intensity, while PC2 represents land exposure and agricultural use.

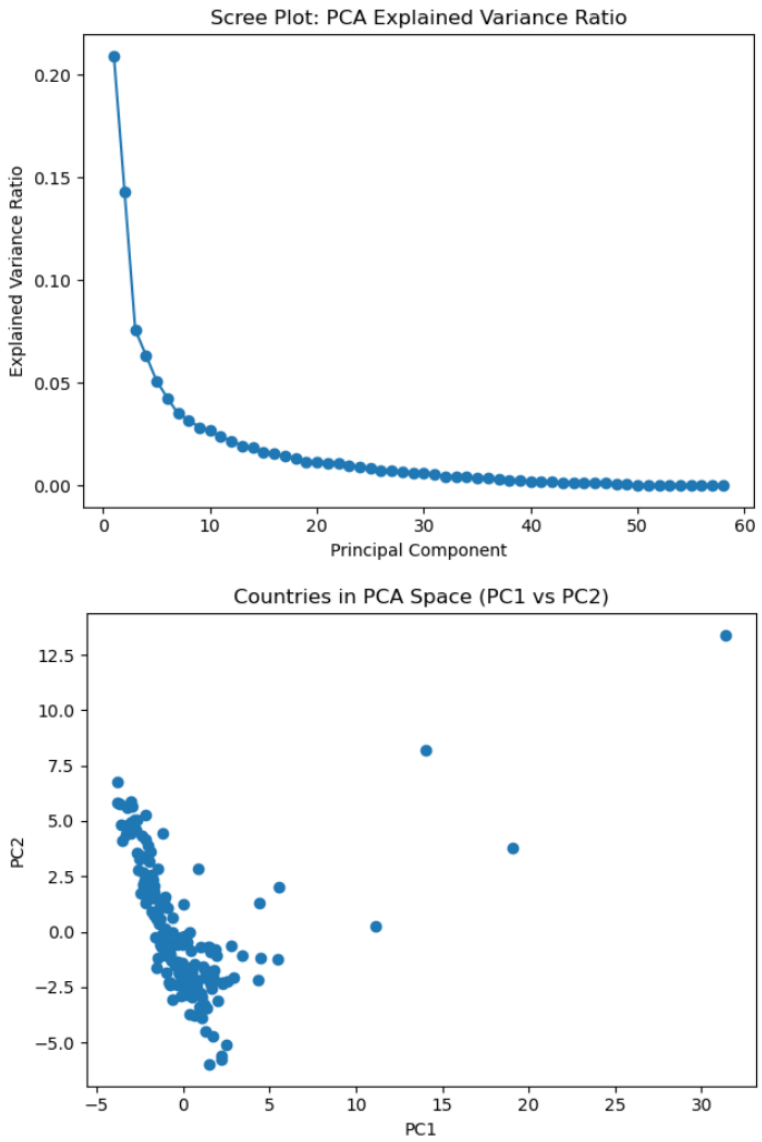
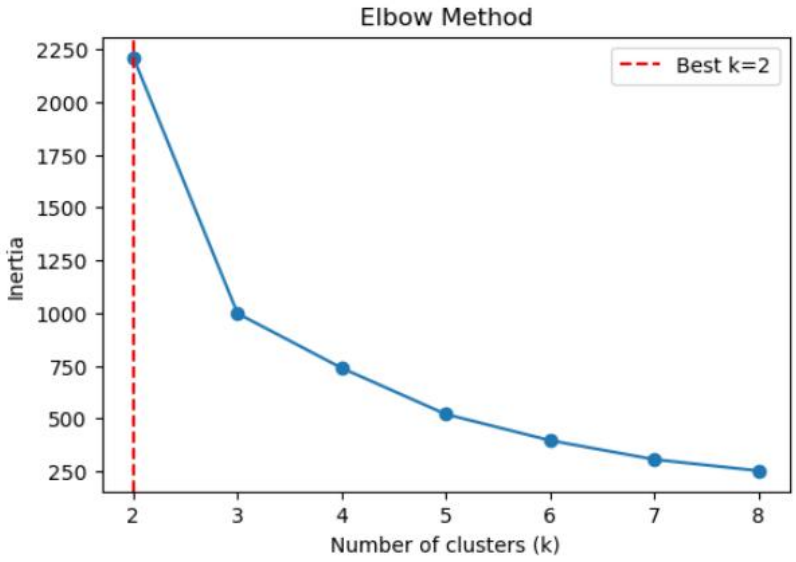
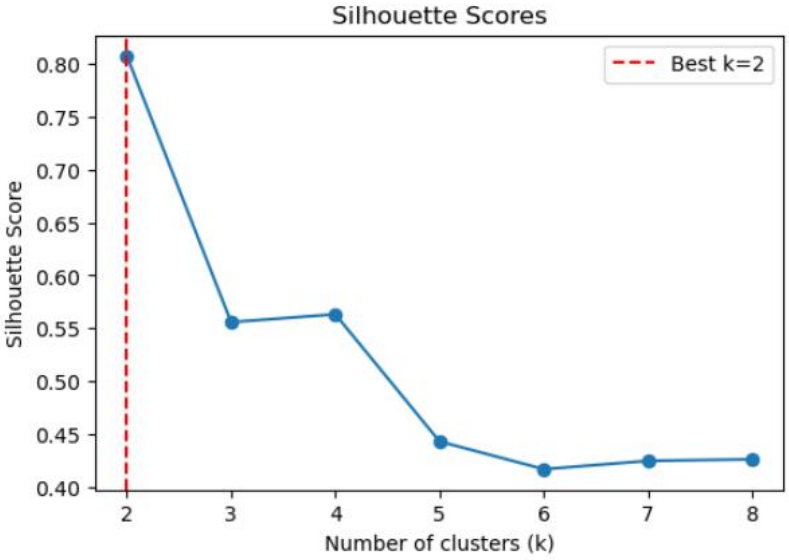


Figure 1 - PCA

2. Optimal Number of Clusters

The elbow curve shows a sharp decrease in inertia at k = 2, followed by a plateau.
The silhouette score also peaks at k = 2, confirming two natural groupings among countries.






```
Best k (by silhouette, fallback=3): 2
```

Figure 2 – K-Means

3. Cluster Characteristics

Two distinct macro-clusters were identified:

- Cluster 0 (Industrialised/Large Economies):
  High population (SP.POP.TOTL), larger urban land areas, high total GHG emissions and CO_2 outputs, stronger infrastructure indicators.
- Cluster 1 (Smaller/Lower Exposure Economies):
  Lower population and emissions totals but higher relative proportions of rural or low-elevation land.

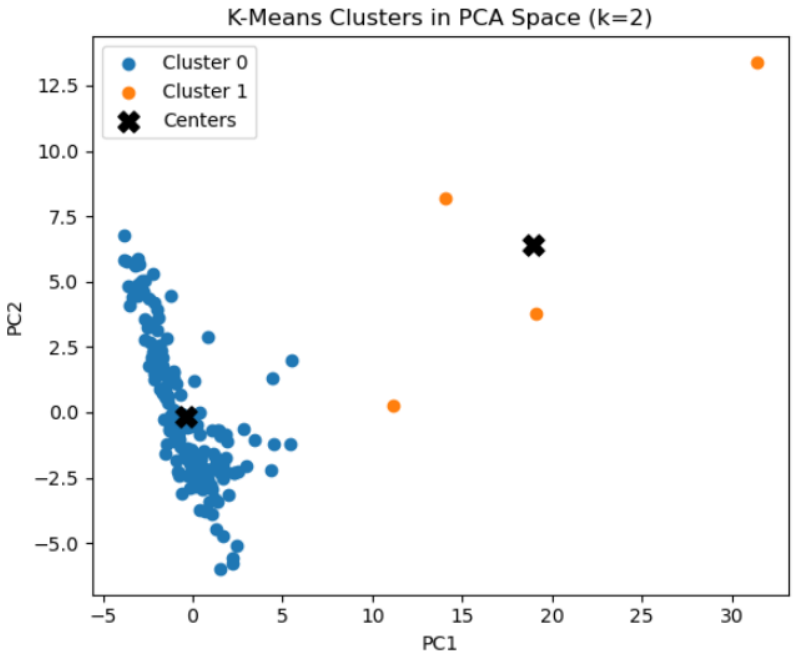


Figure 3 – K-Means Clusters

4. Key Differentiating Indicators

Top 15 indicators that vary most between clusters include:

```
Top-varying indicators: ['SP.POP.TOTL', 'SP.URB.TOTL', 'EN.ATM.GHGT.KT.CE', 'EN.ATM.CO2E.KT', 'AG.LND.FRST.K2', 'AG.LND.AGRI.K2', 'EN.ATM.CO2E.SF.
KT', 'EN.ATM.CO2E.LF.KT', 'EN.ATM.METH.KT.CE', 'EN.ATM.CO2E.GF.KT', 'EN.ATM.NOXE.KT.CE', 'EN.ATM.GHGO.KT.CE', 'AG.LND.EL5M.RU.K2', 'EN.ATM.GHGO.Z
G', 'AG.LND.EL5M.UR.K2']

Cluster profile preview:
          SP.POP.TOTL   SP.URB.TOTL  EN.ATM.GHGT.KT.CE  EN.ATM.CO2E.KT  \
cluster
0        2.414160e+07  1.408363e+07       1.169411e+05    7.739234e+04
1        8.139261e+08  4.308395e+08       6.074312e+06    4.834208e+06

         AG.LND.FRST.K2  AG.LND.AGRI.K2  EN.ATM.CO2E.SF.KT  EN.ATM.CO2E.LF.KT  \
cluster
0          1.426046e+05    1.867252e+05       1.963894e+04       3.185170e+04
1          3.543112e+06    3.323768e+06       2.558348e+06       1.125938e+06

         EN.ATM.METH.KT.CE  EN.ATM.CO2E.GF.KT  EN.ATM.NOXE.KT.CE  \
cluster
0             25885.217391       22121.47644       10109.076087
1            844325.000000       711964.55150      275312.500000

         EN.ATM.GHGO.KT.CE  AG.LND.EL5M.RU.K2  EN.ATM.GHGO.ZG  \
cluster
0               476.496600        4427.710955     12945.60407
1           -190653.863281       86691.761245       167.79881

         AG.LND.EL5M.UR.K2
cluster
0               488.073375
1             11698.361429
```

Figure 4 – Key Indicators

5. Australia in Context

Australia is present in the dataset (country code: AUS) and is assigned to Cluster 0 (k = 2). This cluster represents large, industrialised economies.

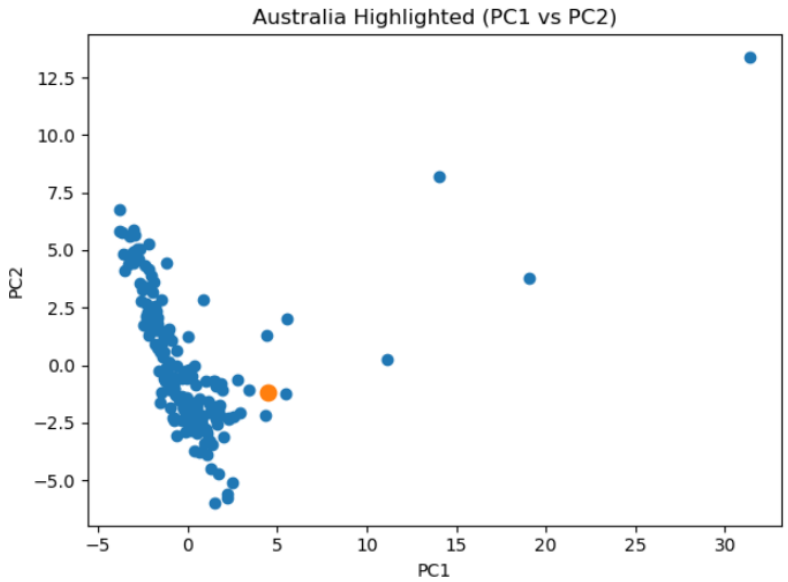


Figure 5 – Australia(Yellow Mark) on PCA scatter

Relative to the mean of its cluster, Australia shows the largest positive deviations in:

- Urban population (SP.URB.TOTL)
- Agricultural land area (AG.LND.AGRI.K2)
- Total population (SP.POP.TOTL)
- Forest area (AG.LND.FRST.K2)
- Total GHG emissions (EN.ATM.GHGT.KT.CE)
- CO_2 from solid fuels (EN.ATM.CO2E.SF.KT)
- Methane emissions (EN.ATM.METH.KT.CE)

Conversely, Australia's largest negative deviations (lower than cluster mean) occur in:

- GHG emissions growth rate (EN.ATM.GHGT.ZG)
- CO_2 from liquid fuels (% of total) (EN.ATM.CO2E.LF.ZS)
- NOx emissions growth (EN.ATM.NOXE.ZG)
- Methane growth (EN.ATM.METH.ZG)
- Freshwater withdrawal (% of resources) (ER.H2O.FWTL.ZS)
- Precipitation (AG.LND.PRCP.MM)
- Cereal yield (AG.YLD.CREL.KG)

Interpretation

Australia's membership in Cluster 0 aligns with industrialised, high-emission economies. It exhibits larger absolute outputs and land-use metrics but lower growth rates across several emission categories—consistent with a mature, stabilising economy. This profile supports Australia's positioning among developed nations with substantial environmental footprints but moderated emission growth.

Conclusion

- Dimensionality reduction: 23 PCs capture approximately 90 % of total variance; PC1 and PC2 explain 35.23 %.
- Clustering: K-Means (k = 2) reveals two clear country archetypes.
- Australia: Belongs to the industrialised cluster with high absolute emissions but slower growth.
- Policy implication: Clusters and PCA components provide an interpretable structure for international comparison of sustainability metrics.
- Limitations: Cross-sectional data, mixed indicator years, limited causal inference.
  Future work should extend to hierarchical clustering, temporal trends, and inclusion of socio-economic predictors.