**Title**

Beyond GDP: A Multi-Dimensional Clustering of Nations

# Abstract

Traditional methods of evaluating global development rely on rankings based on just one metric, which produce an incomplete picture of a nation's progress. By applying a clustering method to a set of World Bank indicators from 2018–2020 for a group of 15 countries, our study challenges such simplicity. The study uses a RobustScaler to preprocess the data to control outliers and standardize different scales. The K-means clustering algorithm is based on four basic indicators: GDP, life expectancy, primary school enrollment, and safe internet servers.The analysis found four different and understandable clusters: one economic superpower (USA), a group of advanced high-tech economies, a group of major emerging economies lead by China, and a developing nation category. The results show that this method gives a more genuine and detailed classification of countries compared to standard indicators, highlighting the importance of unsupervised machine learning for uncovering patterns in global socio-economic data.

# 1. Introduction

For a long time , the method for assessing a country’s standing has been through economic metrics. Indicators like GDP (Gross Domestic Product) have been common benchmarks for progress and to let us know the economic standing of the country. But , in such an interconnected world taking only one economic indicator is not enough , we need to take in account the country’s health facilities , education and digital infrastructure to provide a more comprehensive understanding of global development.

The core problem is that One-dimensional rankings produce an inaccurate and incomplete picture of global development, which is the main study issue. This study addresses this by analyzing multiple factors simultaneously using K-means clustering, which gives a more realistic and authentic classification of countries. It tries to understand the natural grouping of the countries through similar development levels of their respective indicators. The objectives are as follows :

1. To fetch a relevant and reliable dataset from the World Bank which covers key economic , social and technological indicators for various countries with different levels of economies.
2. To use unsupervised learning and apply K-means clustering for separating these nations into different , data driven groups.
3. To analyze and interpret the results of the clusters that have been obtained through clustering , identify similar levels of countries and profile them together to create a meaningful tier of global development.

The paper will explain how the analysis was done , show which countries were grouped together after clustering and discuss what did the findings mean.

# 2. Methodology

The methodology taken place in this study , ensures a robust and reproducible analysis. It has 3 essential phases: Data fetching and selection , Data preparation and preprocessing, the final clustering model and its analysis.

## 2.1 Data fetching and selection

The data source for this research was the World Bank's official World Development Indicators (WDI) database , which is a trusted and reliable source for global socio-economic data . To ensure a systematic and reproducible process , the data was acquired using a Python script and the wbgapi library , which is linked directly to the World Bank’s API. The selection of indicators were done keeping in mind to have a rich multi-dimensional view of national development. Four indicators were chosen to show distinct but complementary aspects of a country’s profile :

* GDP (current US$): To measure the economic scale.
* Life Expectancy at birth (years): Represents social well-being and public health.
* School Enrollment, primary (% gross): Serves as a proxy for investment in human capital and basic education.
* Secure Internet Servers: To represent technological/digital infrastructure.

These sets of indicators were collected for 20 countries for a three-year period from 2018 to 2020.

## 2.2. Data Preprocessing and Preparation

Once the raw data was collected , it underwent various important preprocessing steps in order to be prepared for the clustering algorithm. The raw data was first checked for its completeness ( to check if there were any missing values ) , this visually showed us the presence of sparse data for our literacy rate indicator . Thus , this indicator was replaced by a school enrollment indicator to ensure data integrity. The process of removing entries with any missing data resulted in a final, clean dataset of 15 countries with complete records for all indicators from 2018-2020.

Then the data was pivoted into a “wide” format where each row represented a single country and each column a specific indicator and its corresponding year. Then an outlier analysis took place using box plots and Z - score which confirmed the presence of certain extreme values particularly in GDP and technological indicators , however these were intentionally not removed as they showed us the realistic economic situation for our research problem. Lastly , the scales of the indicators were vastly different which is why it was necessary to standardize them . This was done using the RobustScaler in Python as it standardizes the data based on its median and interquartile range

## 2.3 Clustering Model

We have used the K-means clustering model for our study mainly because of its effectiveness and simplicity for splitting the data into non overlapping clusters and storing them into separate groups. Before starting out with the K means algorithm we had to figure out what would be the appropriate number of clusters or groups to be created. In order to do this analysis , the Elbow method technique was used which indicated that k = 4 would be the appropriate amount of clusters that would be required for our dataset. We then did our final clustering using the Orange Data Mining software which provided us with interactive visualization tools. Here the K means tool was used from the “Unsupervised” widget set with a fixed random seed , k = 4 and was given 10 re-runs to be performed and the best output to be selected.

# 3.Results

The utilization of the K-means algorithm helped us in getting stable and clear results which showed a clear segmentation/separation of 15 countries into a group of 4 clusters . The results that were obtained are presented first with the specific country assigned for each cluster and then through various visualization tools and techniques that help us understand the nature of these groupings and why they were grouped together.

## 3.1 Clusters Assigned

The final clustering model is set with the number of clusters to be taken as 4 ( k=4 ) , assigned each of the 15 countries into 4 clusters . The components of each cluster are given in **Table 1 Final Clusters Assigned.**

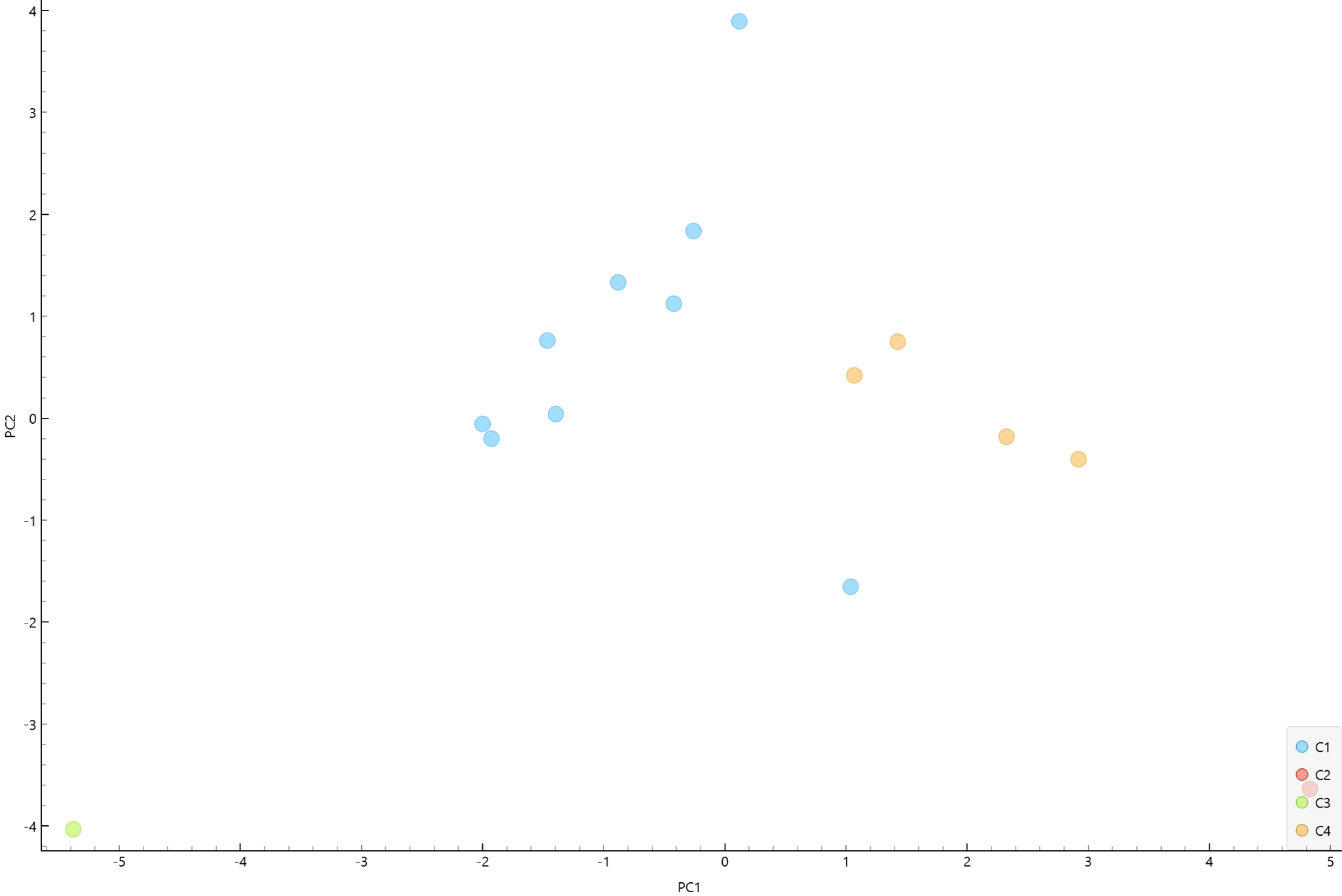
### **Table 1: Final Clusters Assigned**

| Cluster | Profile | Member Countries |
| --- | --- | --- |
| C1 | Major Emerging Economies | Brazil, China, Egypt, Indonesia, India, Mexico, Turkey, Vietnam, South Africa. |
| C2 | The Economic Superpower | United States |
| C3 | Unique Developing Profile | Nigeria |
| C4 | Advanced High-Tech Economies | Australia, Germany, Japan, Singapore |

## 3.2. Visualizations

To dive deeper and explore the relationships between these clusters, several visualizations were used namely the PCA plot , Geo Map , and the GDP vs Life expectancy scatter plot.

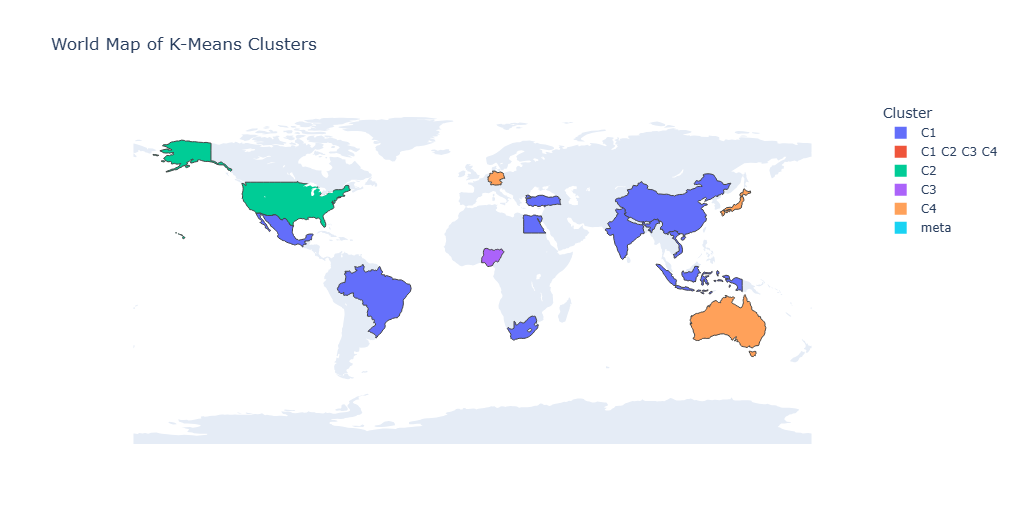
**Figure 1** shows a Principal Component Analysis Scatter plot , PCA in general reduces the dimensionality in the dataset while keeping the most important information . In our case , the 12 indicator variables are now reduced to two principal components that show the maximum variance present in our data. As we can see from **Figure 1** the plot shows four distinct clusters that are well separated , which also confirms and validates our model’s output. We can see how C2 ( Cluster assigned for United States ) is quite isolated which shows its outlier economic scale. While C4 ( Cluster assigned for Advanced High Tech Economies ) are grouped together.



### **Figure 1: PCA Visualization Scatter Plot**

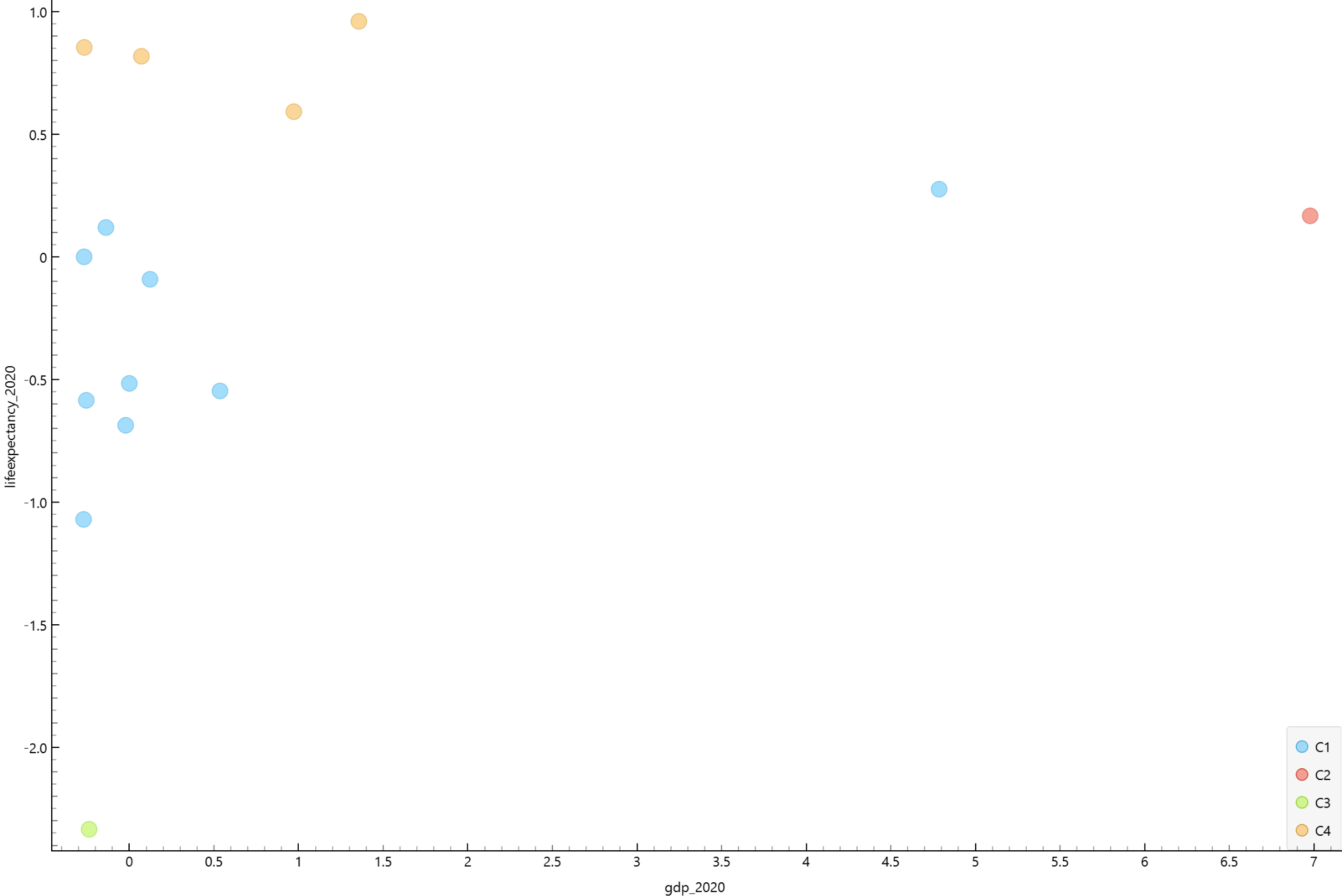
**Figure 2** shows a choropleth world map , which illustrates the geographic distribution of the 4 clusters present in our data. The visualization we see through the choropleth map reveals that our clusters and the countries present in those clusters are not confined to the regions but are genuinely based on the various developmental profiles they are assigned to. For example , our C4 cluster which has the advanced economies comprising Australia , Germany , Japan and Singapore are spread across different locations like Asia , Europe and Oceania .

The largest group that we see is the Cluster 1 has Major emerging economies and also the BRICS member countries : Brazil , India , China , South Africa grouped together in the same cluster but wide spread across the map. Likewise , C2 the economic superpower (USA) is also clearly visible and distinct on the global map highlighting its dominance compared to other clusters , while C3 highlights Nigeria in a different and unique light in Africa. This map ultimately validates that our clustering results are not just restricted to the geographical boundaries but are indeed segmented based on their developmental status.



### Figure 2: Geographic distribution of the four clusters on a world map.

Figure 3 gives us a comprehensive view of the clusters as we plot two indicators against each other . The two variables are: GDP(2020) and Life Expectancy(2020). The scatter plot shown in Figure 3 gives a crucial insight by showing a clear pattern where the advanced economies C4 and C2 are present in the top right quadrant indicating better economic output and higher social well being. While C3 (Nigeria) is low in both the indicators gdp as well as social well being indicating lower economic and developmental output.



### **Figure 3:** Scatter plot of countries by GDP and Life Expectancy, colored by cluster.

# 4. Discussion

The K-means analysis successfully segmented all the 15 nations into four distinct and explainable tiers, moving beyond a simple one metric ranking to tell a more detailed story of global development. The results, visualized in the sections above, show that countries can be grouped based on their combined economic, social, and technological profiles. The most noticeable cluster C2 ( Economic Superpower) only has the USA . We even see this in the Principal Component Analysis plot where , US is a heavy outlier on the primary component and is driven entirely by its high GDP. This suggests that the scale and magnitude of the US economy is a category on its own and no other countries in our sample economically match this superpower , which also makes it so distinct from the other countries.

In contrast C4 , the Advanced high tech economies show a different kind of top tier . The cluster comprising Australia , Germany , Japan and Singapore are not grouped together by their GDP but because of their consistent , high level performance across multiple indicators. The visualization that we have carried out confirms that these nations lead both in life expectancy and secure internet servers , which tells us that they are a profile of stable , high quality of life economies with good technological infrastructure.

The largest group, Cluster C1, the Major Emerging Economies, forms the central global development landscape. This cluster includes nations like China, Brazil, and India, which occupy the middle ground across most indicators. China in this group is a key insight as although its GDP is quite significant, its overall development aligns it more closely with other major emerging economies than with the established, high-income economies of Cluster C4.

Finally, C3 has only Nigeria. Its isolation from the other emerging economies shows that the specific combination of development indicators creates a profile that is unique. This highlights that there are significant complexities even within the broader category of developing nations.

In conclusion, this multi-dimensional approach provides an insightful framework than a simple rich-poor list. It discloses different types of development , the economic scale of the US vs the high-tech stability of Germany , Singapore and Japan and better represents the various complexities of the world's emerging economies. While this analysis is based on a limited sample of 15 countries and four indicators, it effectively shows the capability of clustering to uncover meaningful patterns in global scale data.

# 5.Conclusion

This research started by addressing the problem that one-dimensional rankings create an inaccurate and incomplete picture of global development. By applying K-means clustering algorithm to a multi-dimensional dataset of economic, social, and technological indicators, this study successfully segmented a diverse sample of 15 nations into four distinct and meaningful tiers: an economic superpower, a group of advanced high-tech economies, a large group of major emerging economies, and a unique developing nation.

The findings show that this data driven approach gives a more authentic and useful classification of nations than traditional methods of ranking. It discloses that global development cannot be measured by a single indicator or metric but rather should use a multi dimensional approach for understanding the complexities of economies and the various reasons behind their development. This study shows the capability of using clustering methods to uncover patterns in global data.

Further research could expand by using a higher number of countries and a wider range of indicators like environmental sustainability or political stability or by using different clustering algorithms to compare the results.

# 6.References

### Data Sources

World Bank. (2025). **World Development Indicators**. [Data set]. The World Bank. Retrieved from<https://databank.worldbank.org/source/world-development-indicators>

### Software and API

Wbdata. (2025). **wbgapi: World Bank Data API** (Version 1.0.3). [Computer software]. Retrieved from<https://pypi.org/project/wbgapi/>