Final Project

Course: Digital Strategies for Social Sciences

Author: Mishu Dhar

Computational Social Science, Linköping University

Email: misdh783@student.liu.se

I. Introduction:

Socialism has been a matter of attraction and a subject of debate in academia and public forums for many decades. In the study of political and economic systems, socialism stands out as a model that emphasizes the fair distribution of resources, social welfare, and public ownership. Countries that adopt socialist principles often spend a lot on education, healthcare, housing, and social protection to ensure that all citizens have access to essential services and opportunities. This analysis aims to identify the best socialist countries in the world based on key attributes such as government expenditures on social support welfare, public health care, free education for everybody, support in housing, etc. By using principal component analysis and clustering techniques, we can categorize countries into groups that reflect their adherence to socialist principles and evaluate which nations excel in implementing these ideals.

II. Literature Review:

Yangkuo et al. (2022) discuss extensively Xi Jinping's innovative methods to achieve the great rejuvenation of the Chinese nation, injecting a new connotation and spirit of the times into the adherence to and development of socialism with Chinese attributes. Lloyd Cox (2007) in his papers, described what socialism is and what the attributes of a socialist country are. The preconditions of socialism are studied by Tudor (1993) in his research.

III. Data Description and Preparation:

The data for the "Measuring Socialism" dataset was collected by Joseph Nathan Cohen, an Associate Professor of Sociology at the City University of New York, Queens College, and Joseph van der Naald, a Doctoral Student at the City University of New York, Graduate Center. This dataset compiles metrics from various databases, primarily published by the OECD, to capture different facets of government taxation, spending, organizational resources, regulations, and programs. The sources include the World Bank's World Development Indicators, OECD's Governments at a Glance, Government Revenue Statistics, Indicators of Product Market Regulation, and Social Expenditure data.

In this dataset, there were 46 observations and 247 variables. However, to proceed with the analysis, I filtered the data for a few variables including population, country, general government consumption expenditures (exp_consumption), state support in housing (exp_housing), and support in education. There are several reasons behind this selection. First, the data contains many missing observations. If I include all the variables, then imputing could be a really challenging task for me. Secondly, I do not have enough domain knowledge to impute the missing observations properly. Additionally, most of the variables are very similar to each other, which I assumed could be omitted. Finally, I used my knowledge and experience of living in Sweden and reading about other welfare states to choose these variables from the total 247 variables. Simply dropping the observations with missing information is not a good solution as there are only 46 observations and I would lose a large part of the data.

To overcome the missing information issues in my filtered data, I used various techniques for different variables. The missing values in population were imputed manually by collecting data from the internet, which was easy to find. For the missing observations in the country variable, I imputed 255 with Tanzania and dropped the other two due to the lack of information. For the other variables, I imputed the missing observations with the value from the previous rows, as the countries are geographically close to each other, and I assumed their socio-economic conditions are similar. Even after this imputation, there were four missing values in the first four rows of the "state tax" variable. Because the missing values were imputed by the previous value, they could not be imputed this way. I filled these values with the mean value of the variable.

IV. Questions

1. What are the top socialist countries based on the attributes we selected?
2. How do different clustering techniques provide different results in the clusters of socialist countries?

V. Data Mining: Analysis

I initiated the analysis by performing Principal Component Analysis (PCA). From the summary of the PCA, we can see that the first 8 principal components explain 96% of the variance in the data, meaning we lose very little information by focusing on these components.

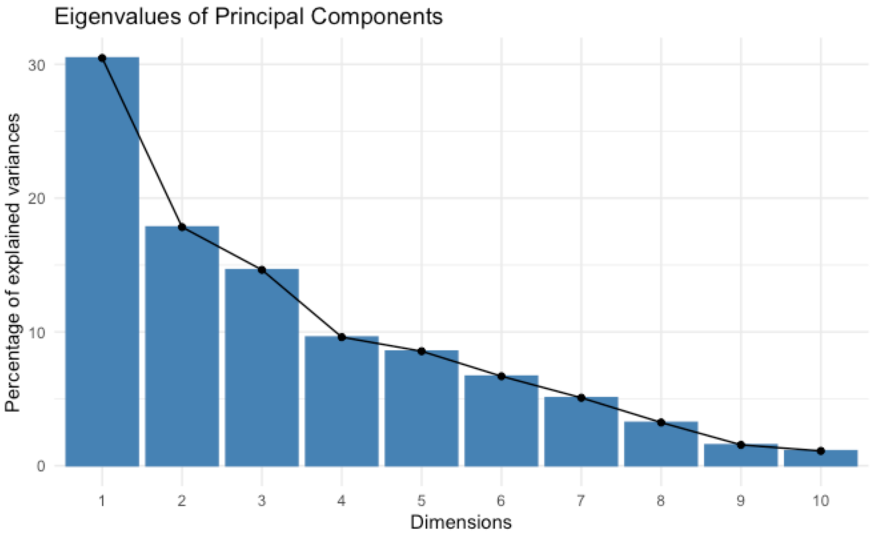


FIGURE 1 EINGEN VALUES OF THE PRINCIPAL COMPONENTS

I further explored the PCA by visualizing the different attributes. Next, I moved on to clustering analysis, starting with K-means clustering using the PCA results.

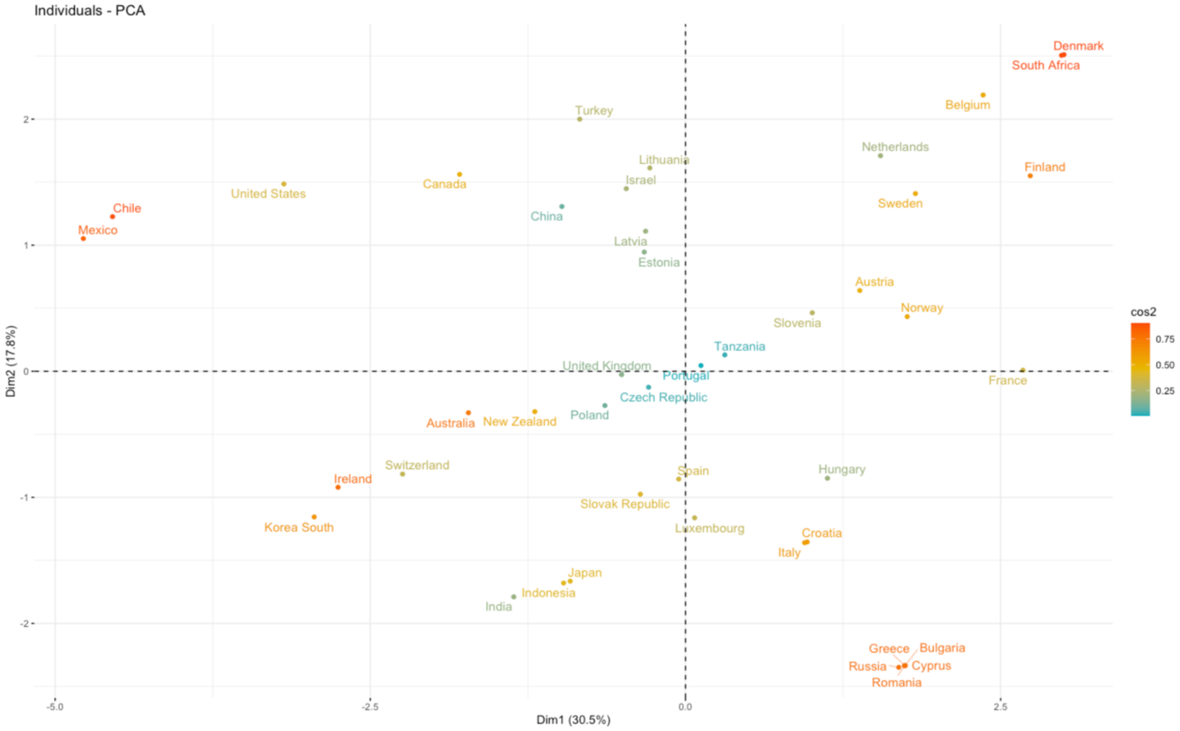


Fig 02: Plot of Countries Based on Selected Socio-Economic Attributes

This plot shows the distribution of countries according to the first two principal components (Dim1 and Dim2), which together explain 48.3% of the total variance. Each point represents a country, with colors indicating the quality of representation (cos2) on the principal components. Countries with similar profiles based on the selected attributes are positioned close to each other. The color gradient ranges from blue (low cos2) to red (high cos2), where red indicates a better representation. The dashed lines represent the origin for each dimension.

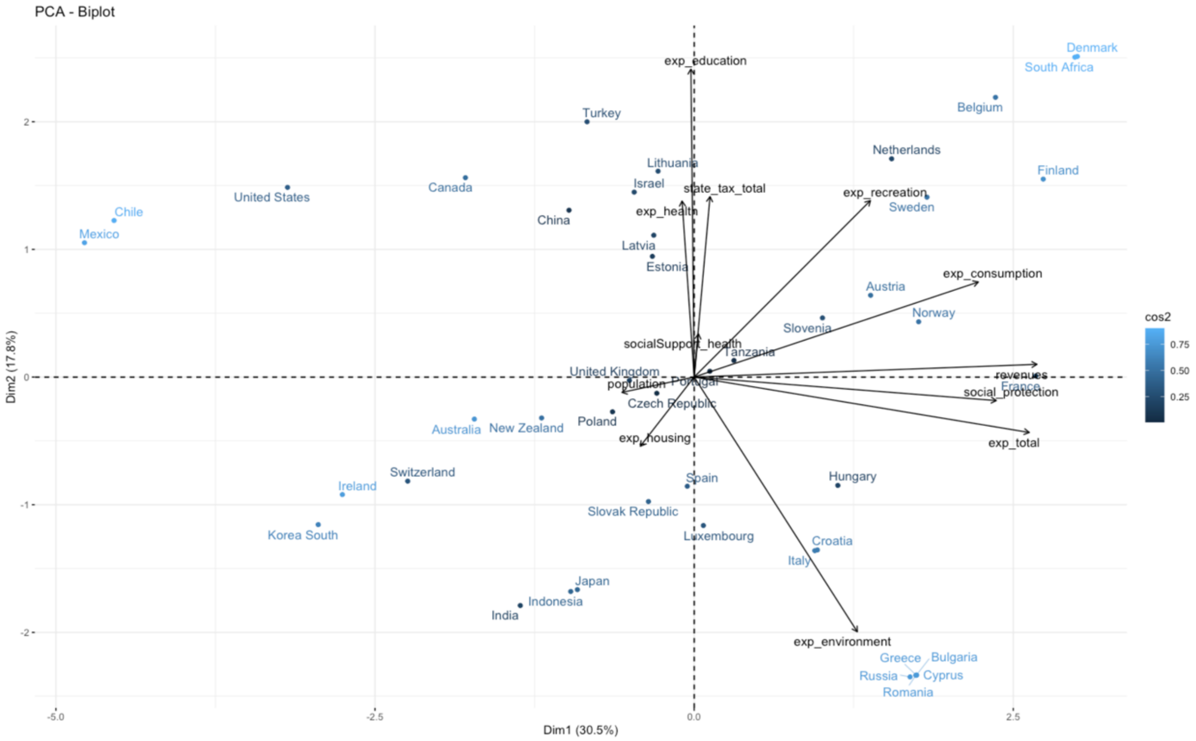


Figure 03: PC1 & PC2 with various loadings (variables)

The optimal number of clusters for K-means clustering was chosen using the elbow method (Fig 04).

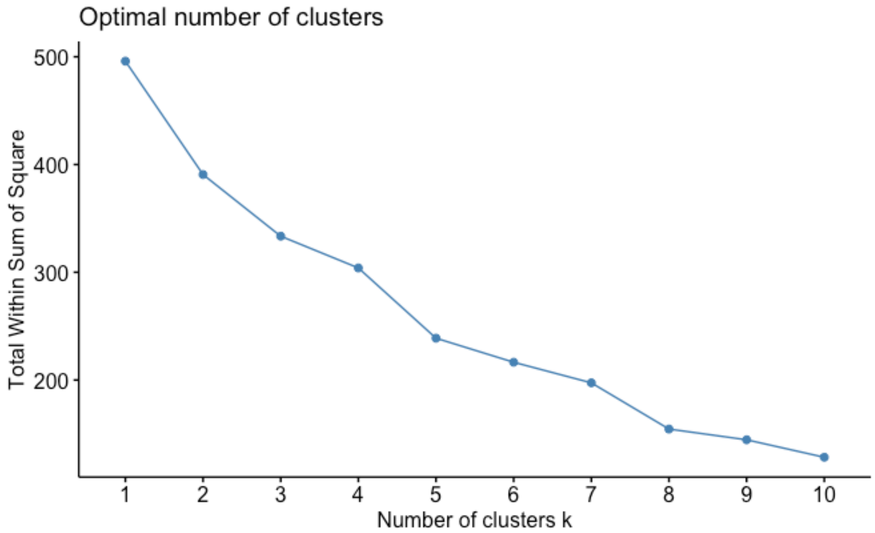


Figure 04: Elbow Method, Optimal Number of Clusters for K-means clustering

For hierarchical clustering, the optimal number of clusters was determined by the dendrogram (fig 05). Additionally, I aimed to identify the top socialist countries and understand how they form groups based on the selected attributes (variables) by using K-means and hierarchical clustering techniques.

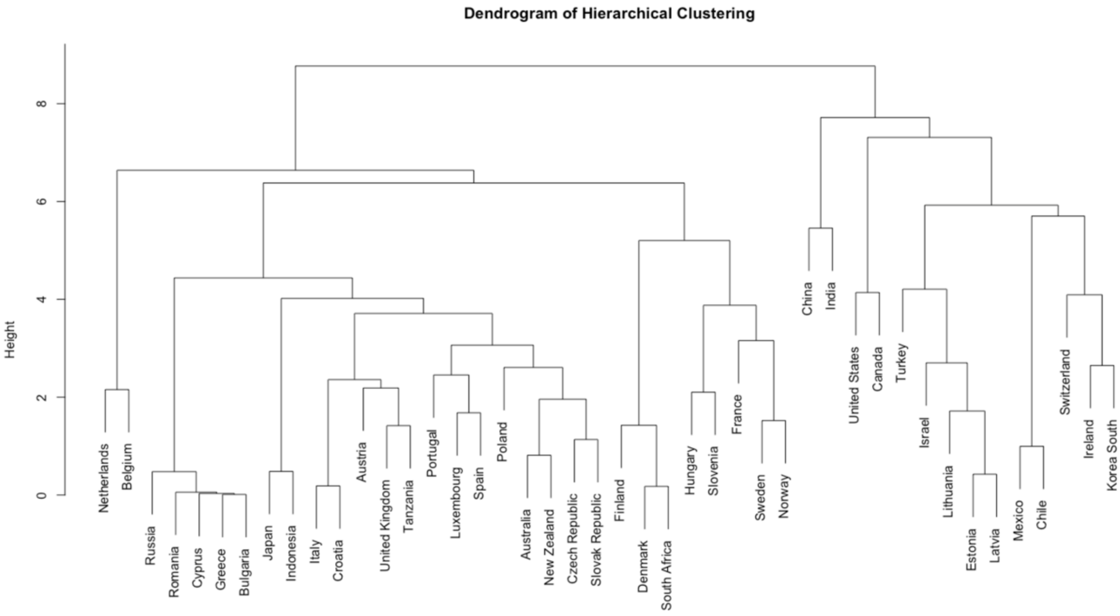


Figure 05: Dendrogram to find the optimal number of clusters for the hierarchical clustering.

Result:

For the K-means and hierarchical clustering with 3 optimal clusters, different results are observable. In K-means clustering, the United Kingdom clusters with countries like New Zealand, Japan, Indonesia, and India (Fig. 06). However, in hierarchical clustering, it clusters with other countries such as Latvia, Austria, and Estonia (Fig. 07). This difference in clustering results is also visible for other countries in both clustering techniques.

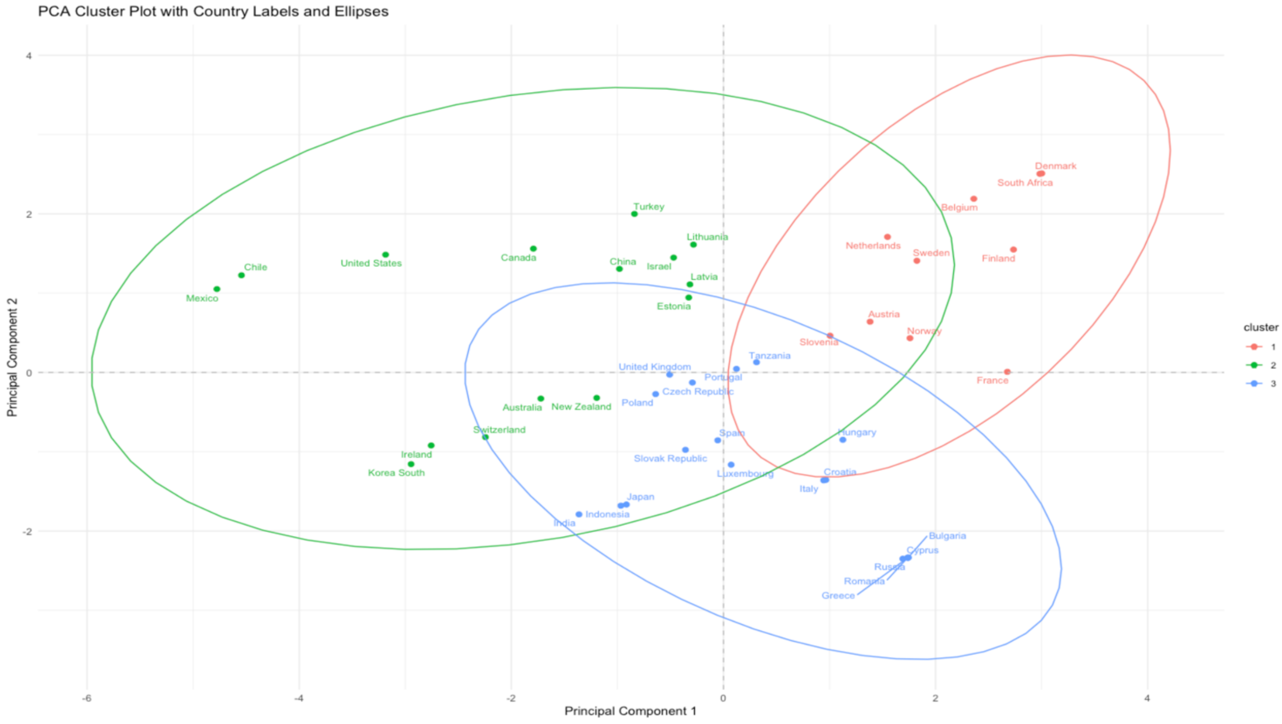


Fig 06: Clusters- K Means Clustering

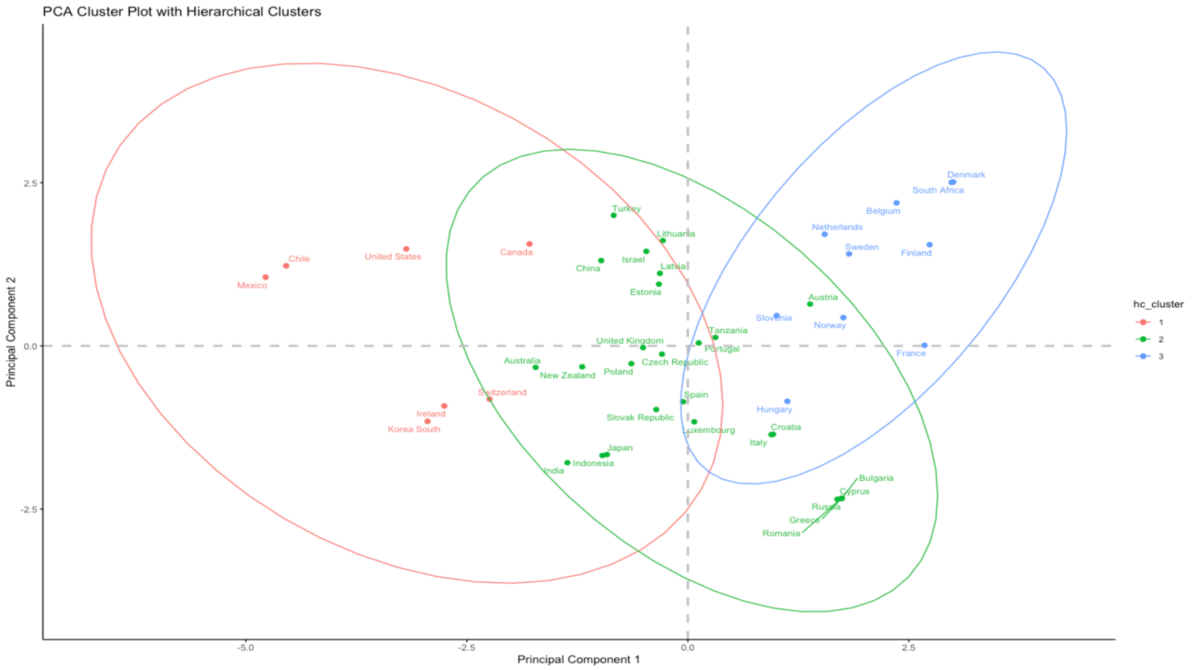


Fig 07: Clusters: Hierarchical Clustering

What Are the Top Socialist Countries Based on the attributes we selected?

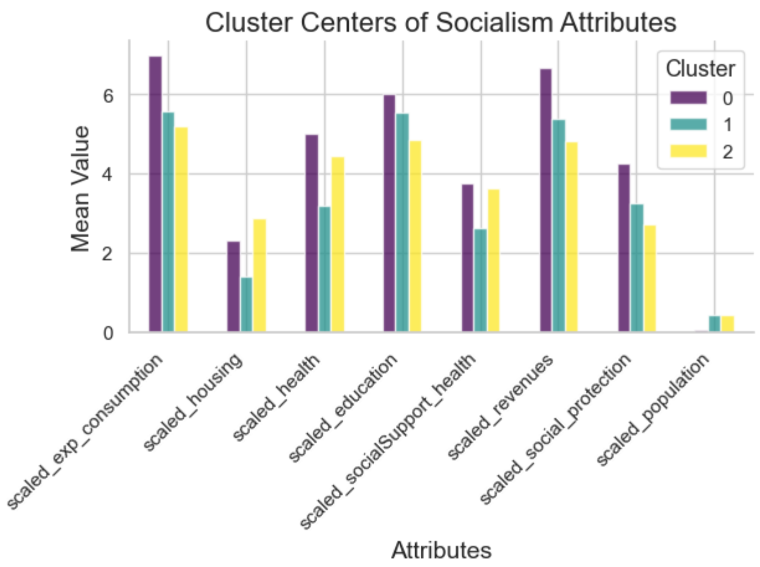


Fig 08: Mean Values of the Selected Attributes in Three Cluster Groups

From the figure above (Fig. 08), we can see that countries in cluster group 0 have the highest values compared to the others. Based on this, welfare states such as France, Finland, and Sweden are the top socialist countries with the highest values of these attributes and belong to cluster group 0. In contrast, the USA, UK, Chile, etc., are fewer socialist states and Luxembourg, Portugal, Greece, etc detected as the moderately socialist country. On the other hand, we can see almost similar results but slightly different outcomes for hierarchical clustering. Welfare countries like France, Sweden, and Denmark remain in the top socialist group, whereas China and India are in the moderate group of clusters. The USA, UK, and Canada still take place in the less socialist groups (fig 09 & 10).

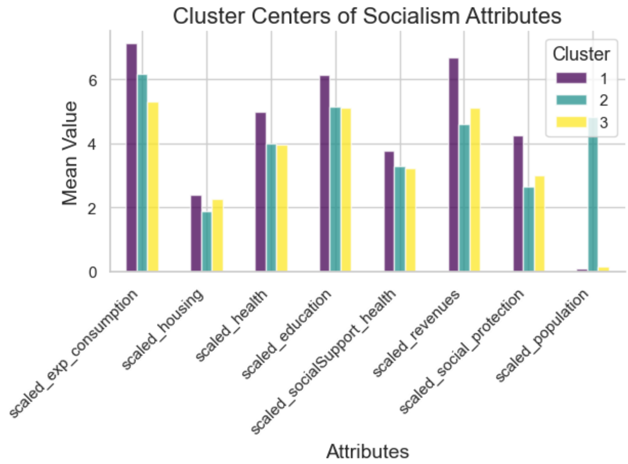
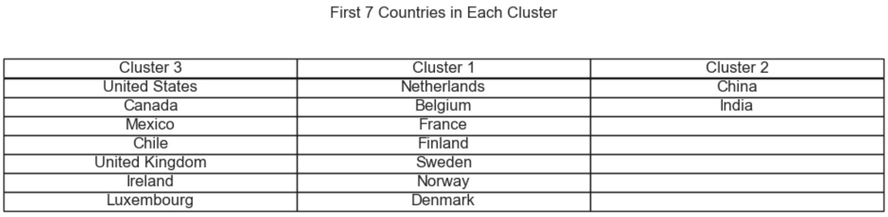


Fig 09: Cluster Centers of Socialism Attributes, Hierarchical Model



First 7 Countries in Each Cluster

Cluster 3	Cluster 1	Cluster 2
United States	Netherlands	China
Canada	Belgium	India
Mexico	France	
Chile	Finland	
United Kingdom	Sweden	
Ireland	Norway	
Luxembourg	Denmark	

VI. Limitations of the Analysis:

The main limitation of this study is the concern about the imputation of the data. The variables presented in this dataset are sensitive and have specific values. For example, the cost of the US government in the environmental sector is listed as 0 in this dataset, which cannot be accurate, and I did not find any specific value to replace it. For countries like India and China, many observations are entirely missing. My imputed information is not an accurate representation of the actual data. These issues have likely impacted the clustering results.

VII. References:

[1] Yangkuo, L., & Guang, T. (2022). Xi Jinping and his thoughts on socialism with Chinese attributes. Journal of Applied Business and Economics., 24(1). Retrieved from https://articlearchives.co/index.php/JABE/article/view/2529

[2] Cox, L. (2007, February 15). Socialism. Encyclopedia of Social Science. https://doi.org/10.1002/9781405165518.wbeoss189

[3] Bernstein, E. (n.d.). The Preconditions of Socialism. In The economic development of modern society (Chapter 3).