Cluster Analysis of Worldwide Countries by the Level of Development

Author: Tsu-Hao Fu

1. Introduction

Overview

For a long time, some international humanitarian aid or food and health-related charity organizations have struggled to allocate limited resources to help the countries most in need. There is no unified standard internationally for the overall condition of a country. Some may be classified as developing countries, but their situation is worse than some undeveloped countries. Therefore, I want to identify the countries most in need of help based on some socio-economic and health welfare variables.

In this midterm report, the dataset I used consists of socio-economic and health welfare variables from 167 countries. I will use what I have learned in class to analyze these data, cluster the countries most in need of help, and identify the appropriate model for data clustering.

Dataset Introduction

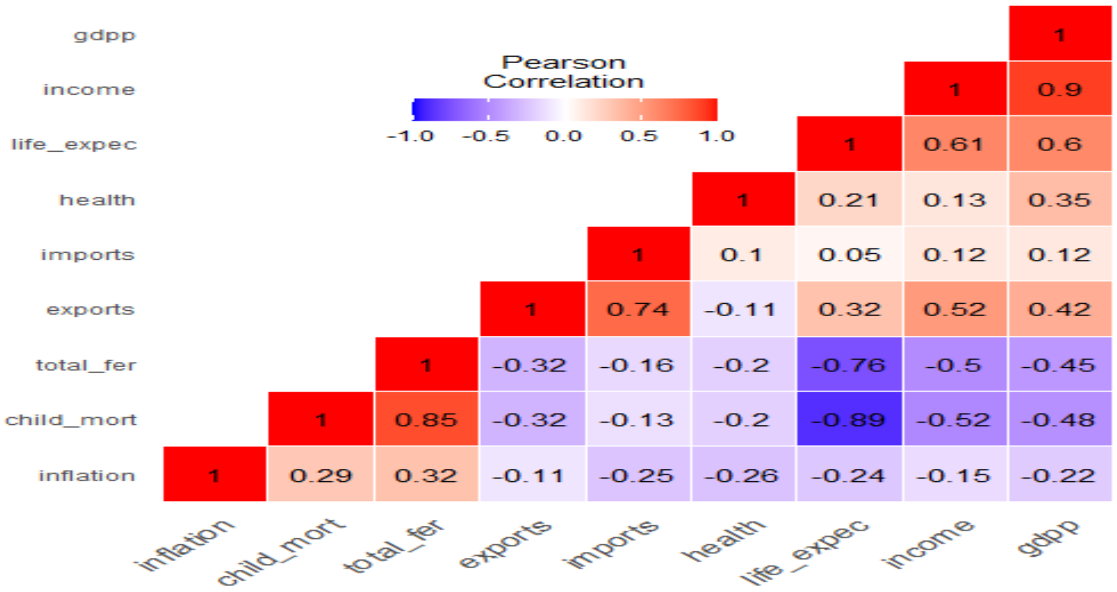
Variable Name	Variable Definition
country	Country Name
child_mort	Number of deaths per thousand children under five
exports	Percentage of goods and services exports in per capita GDP
health	Total medical expenditure as a percentage of per capita GDP
imports	Percentage of goods and services imports in per capita GDP
Income	Per capita net income
Inflation	GDP annual growth rate
life_expec	Life expectancy

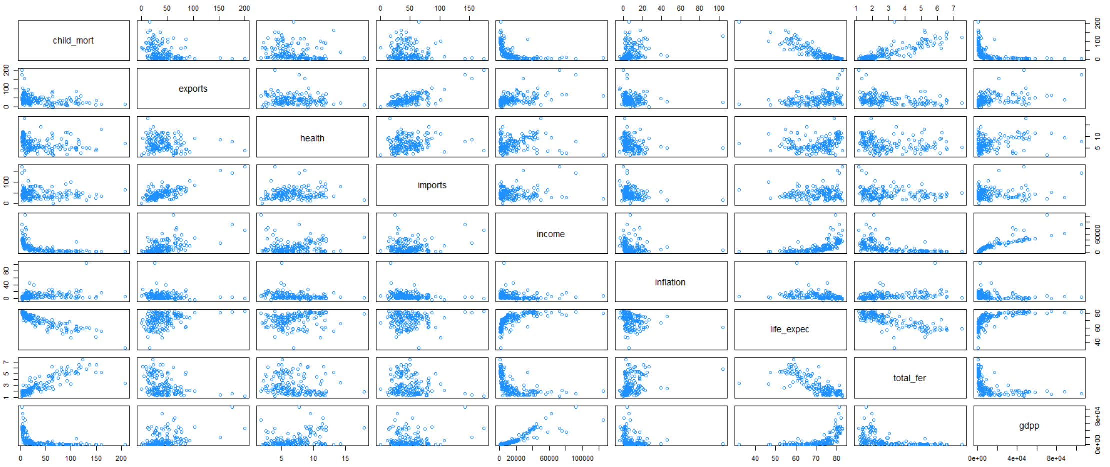
total_fer	Number of children each woman is expected to give birth to
gdpp	GDP per capita

2. Exploratory Data Analysis

Correlation Matrix

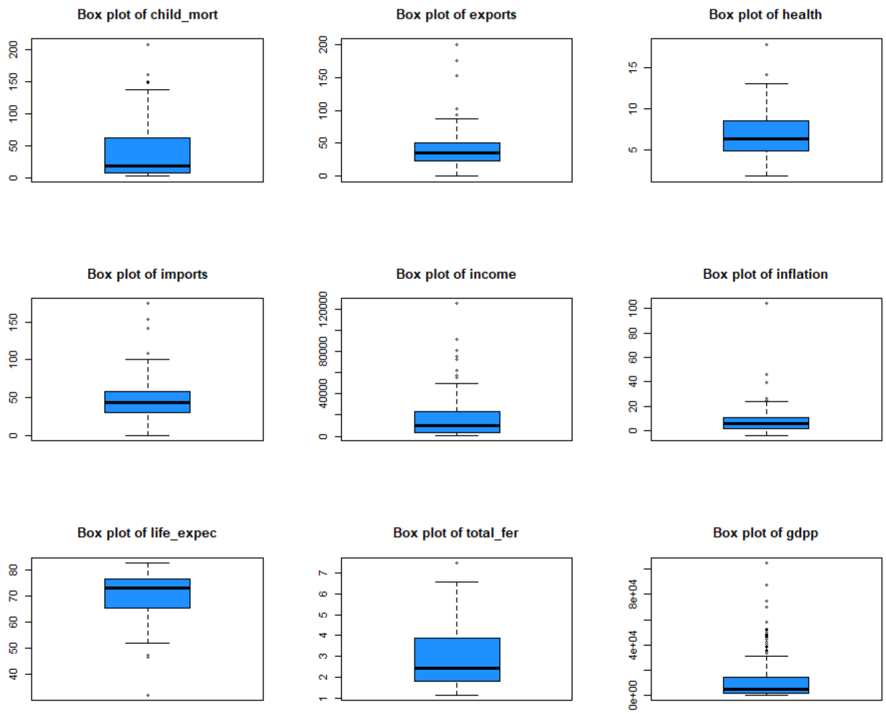
First, using the figure below, observe the correlations and scatter situations among the variables. It can be seen that most countries' GDP annual growth rate, per capita GDP, and per capita net income are concentrated within a narrow range, with only a few countries having higher values. Then, using the heatmap below, we get a more precise correlation of each variable. From the figure, we can find certain correlations between per capita GDP and per capita net income, life expectancy and child mortality, the number of children each woman is expected to give birth to, and between imports and exports.





Box Plot

Next, using the box plots of each variable below, observe if there are any outliers. It is found that every variable has outliers, especially per capita GDP and per capita net income, which have quite a few. However, I do not plan to remove these outliers because the dataset is relatively small, and these outliers may represent one of the clusters.


Box plot of child_mort

Box plot of exports

Box plot of health

Box plot of imports

Box plot of income

Box plot of inflation

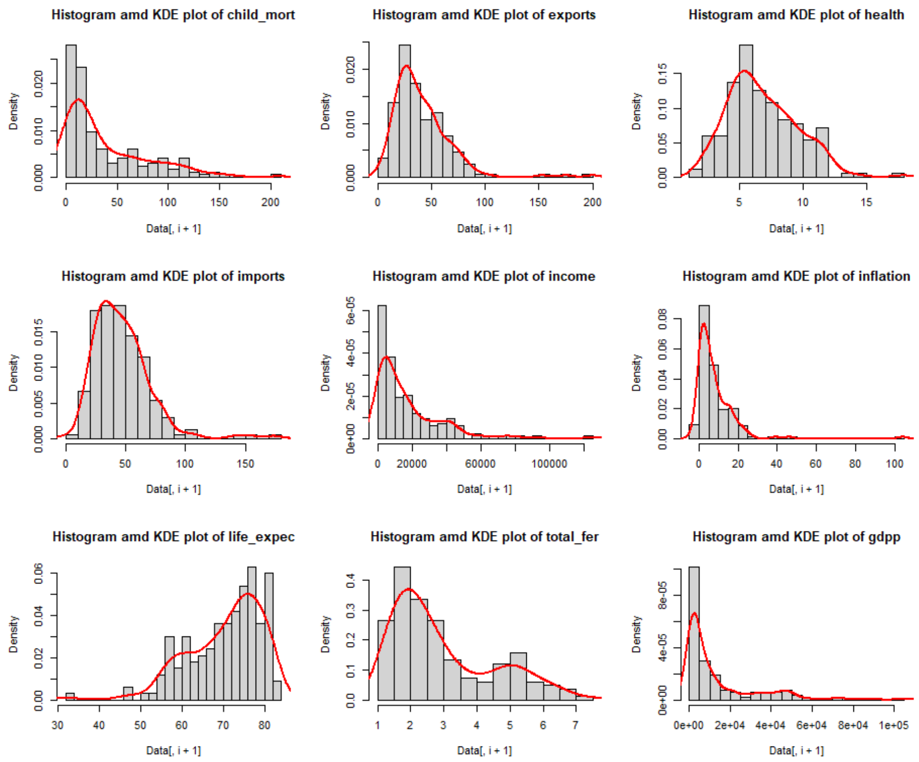
Box plot of life_expec

Box plot of total_fer

Box plot of gdpp

Histogram and Kernel Density Estimate

Finally, using the figure below to observe the distribution of values in each variable, it is found that most variables do not show a normal distribution, which may affect the later calculation of Euclidean distance. Therefore, I will standardize the dataset.



3. Methods

K-means

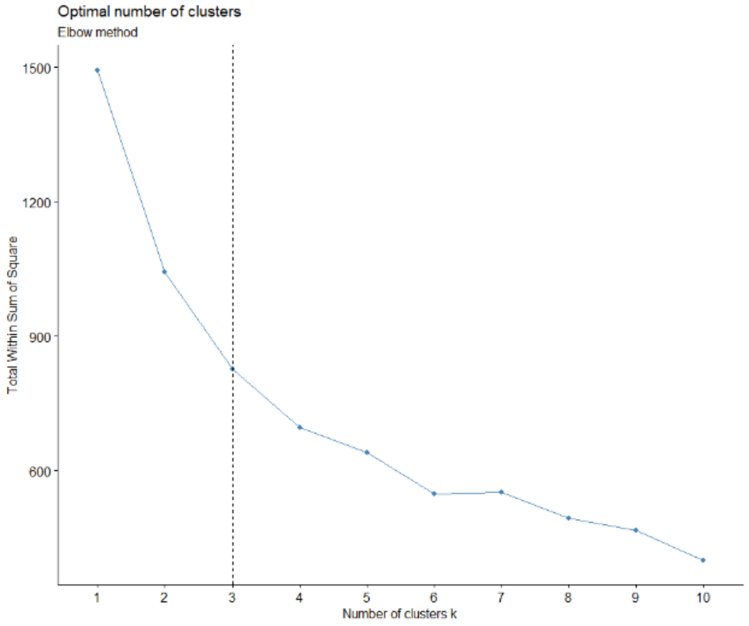
Hierarchical clustering

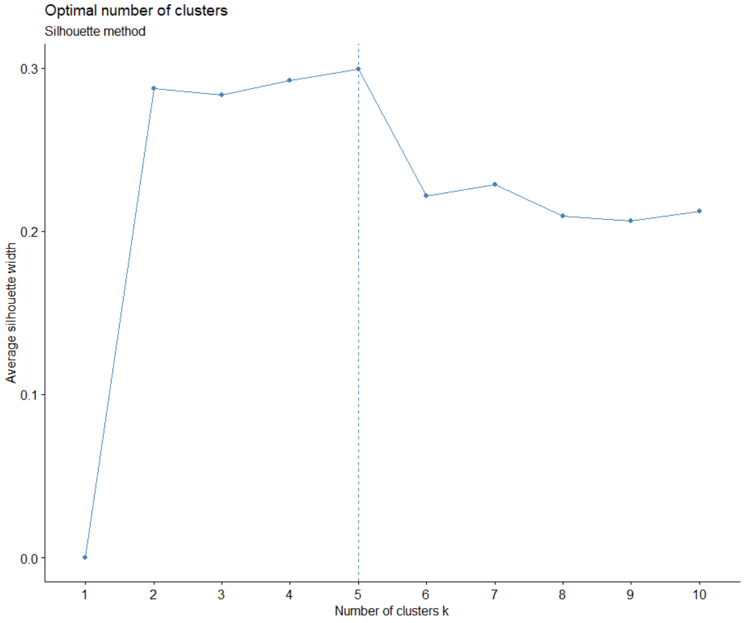
GMM (Gaussian Mixture Model)

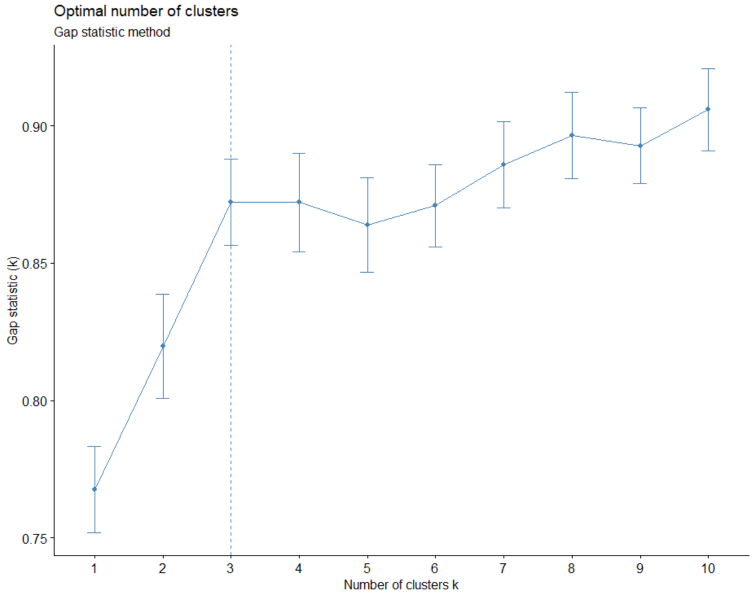
4. Cluster Analysis

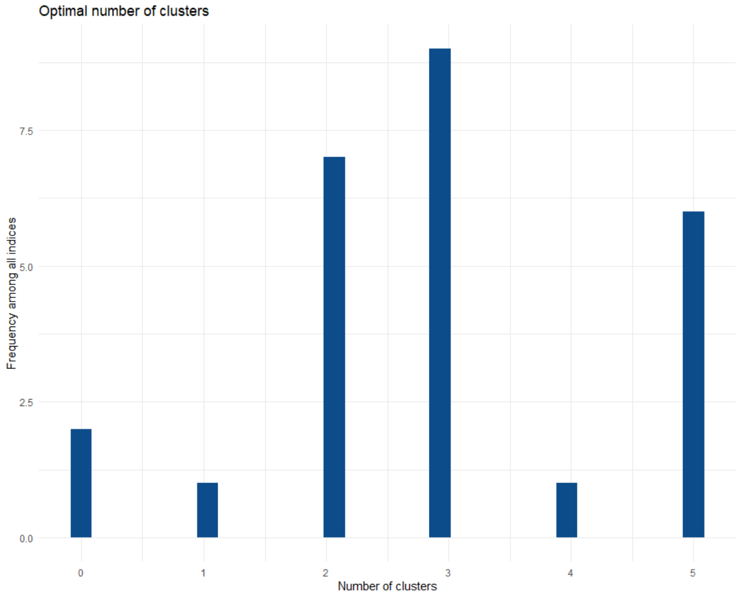
K-means

Based on the methods for selecting the number of clusters below, including the Elbow method and Average Silhouette Gap Statistics, I finally decided to use 3 clusters for the following analysis. The results using K-means (k=3) show that K-means basically divided the data into three groups well.

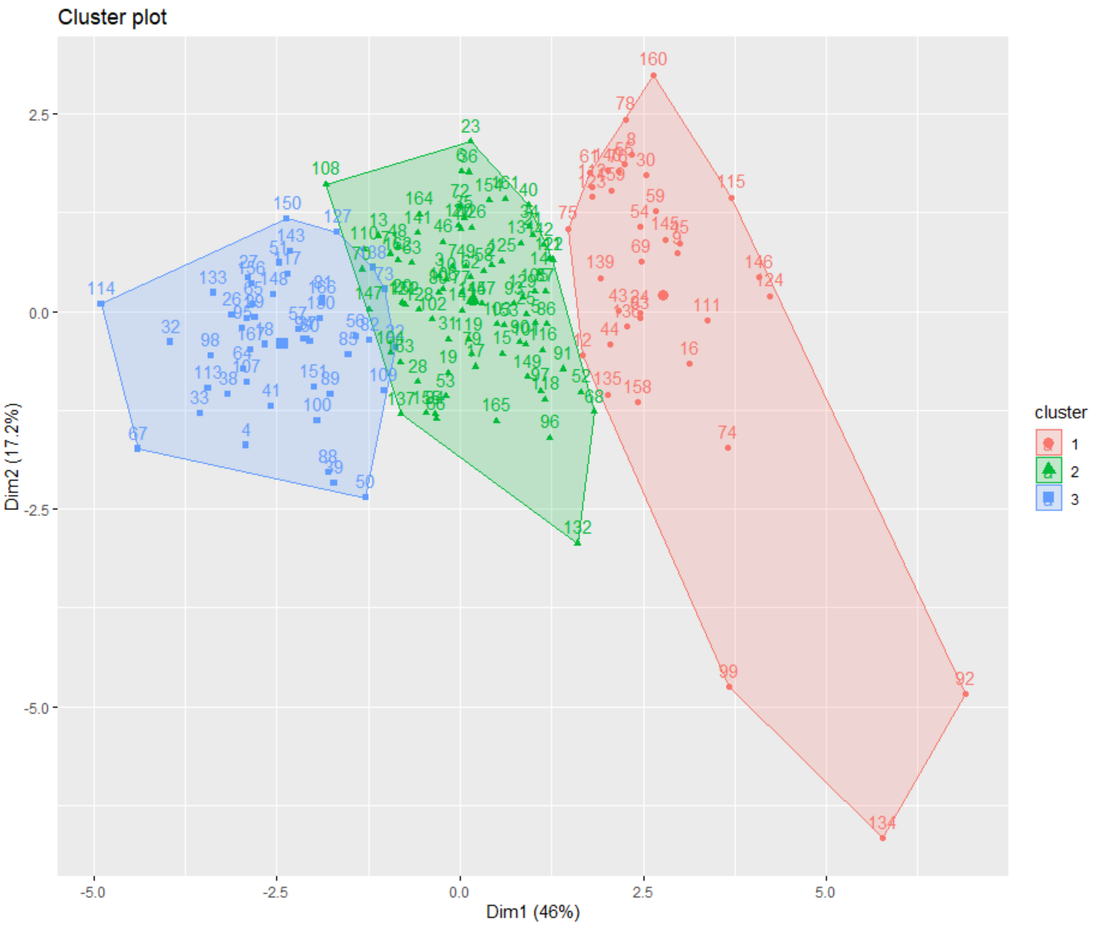


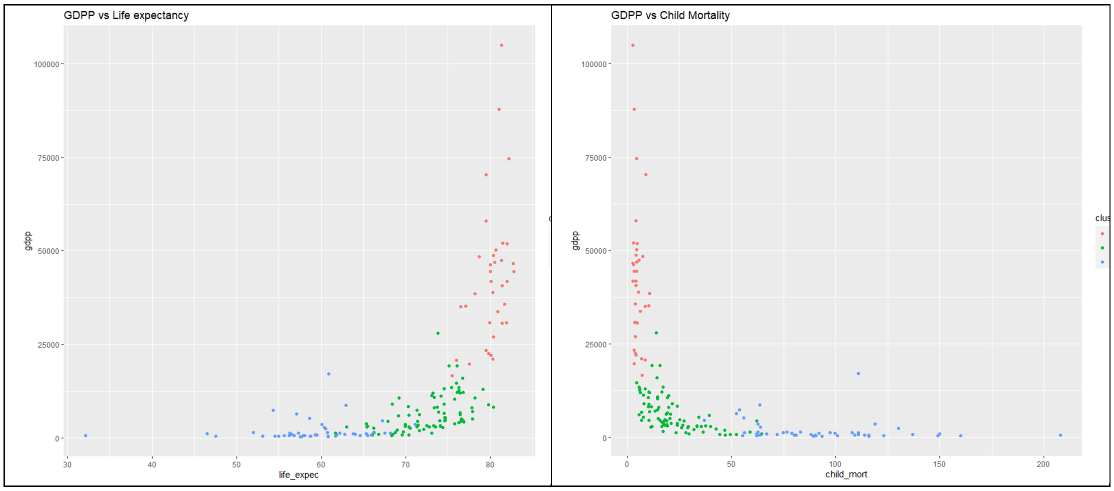


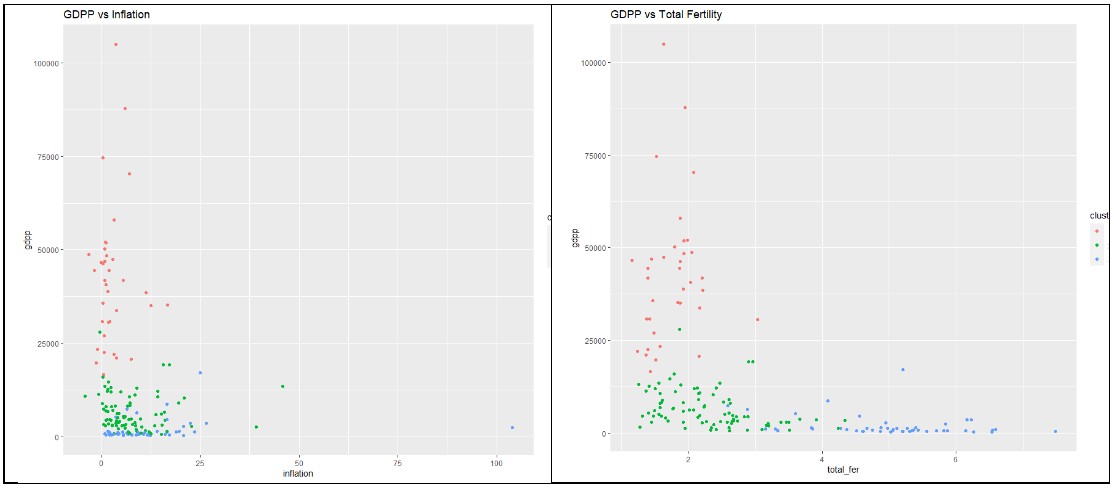




According to the relationship graphs of the four variables below, we can infer the characteristics of the three clusters. Cluster 1 has low child mortality, high per capita GDP, and high life expectancy, indicating these countries are developed. Cluster 2 has moderate child mortality, per capita GDP, and life expectancy, indicating these countries are developing. Cluster 3 has high child mortality, low per capita GDP, and low life expectancy, indicating these countries are undeveloped.

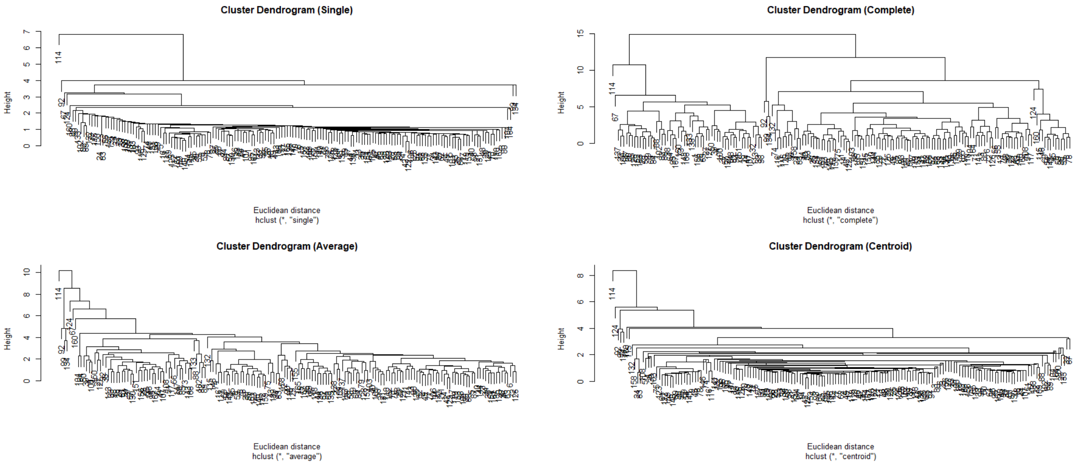

Cluster plot
Dim2 (17.2%)
Dim1 (46%)
2.5
0.0
-2.5
-5.0
-5.0
-2.5
0.0
2.5
5.0
cluster
1
2
3
160
78
115
146
124
111
16
74
99
92
134
23
108
132
96
165
150
114
67
50
4
88


GDPP vs Life expectancy
gdpp
100000
75000
50000
25000
0
30
40
50
60
70
80
life_expec
GDPP vs Child Mortality
gdpp
100000
75000
50000
25000
0
0
50
100
150
200
child_mort
clus

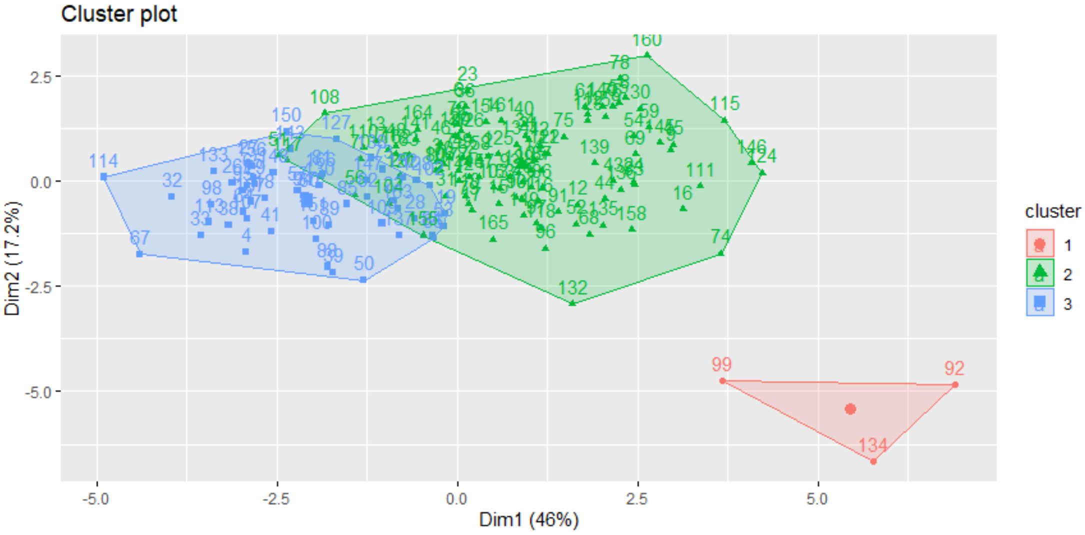


Hierarchical Clustering

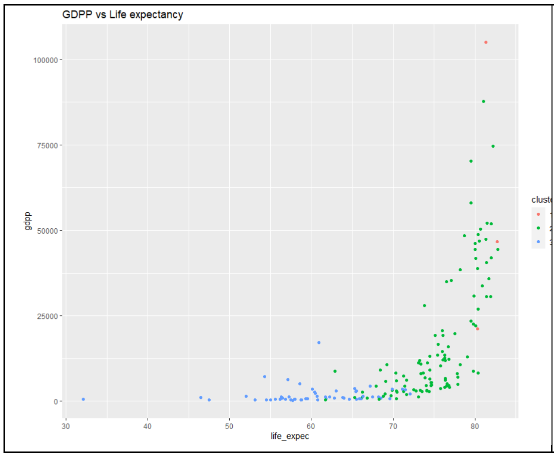
Below are the dendrograms drawn using four different hierarchical clustering methods, including Complete, Average, Single, and Centroid. It is clear that only Complete can maximize the segmentation of the data, so I chose the Complete mode of hierarchical clustering for the following analysis.

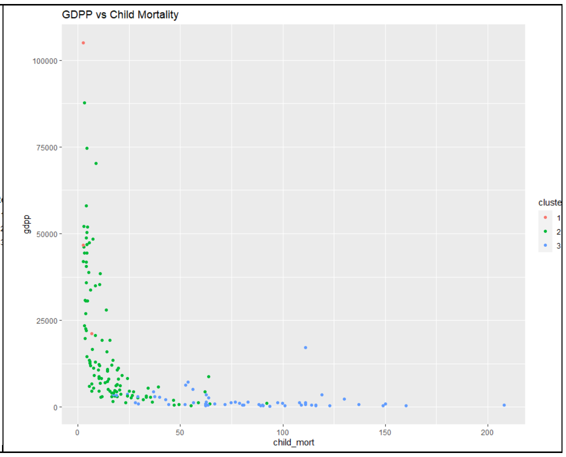


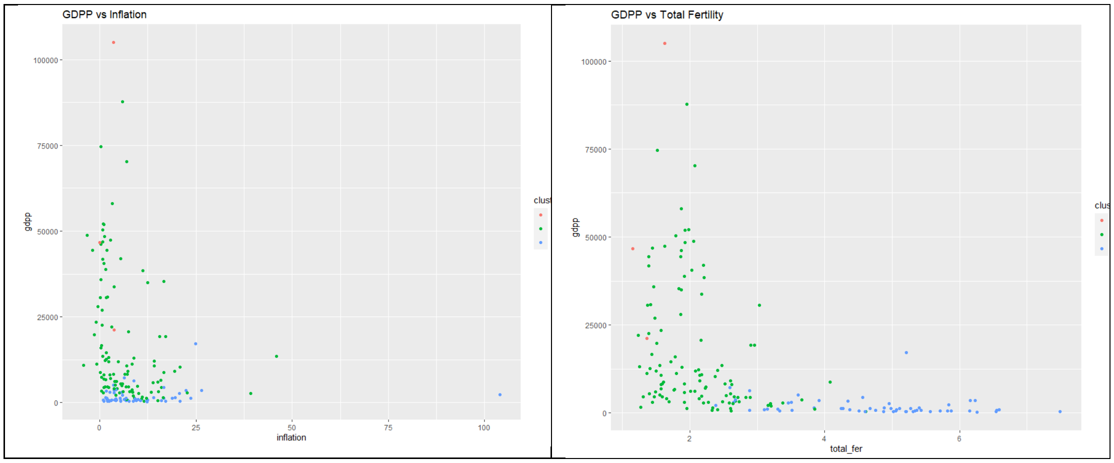
Based on the Complete dendrogram, I chose to divide the data into three groups for analysis. The results show that the clustering effect is not good; clusters 2 and 3 overlap significantly, and cluster 1 has too few data points. Therefore, cluster 1 should be classified into cluster 2, only dividing into two clusters.



Based on the relationship graphs of the four variables below, we can infer the characteristics of the two clusters. Cluster 2 has low child mortality, high per capita GDP, and high life expectancy, indicating these countries are developed or highly developing. Cluster 3 has high child mortality, low per capita GDP, and high life expectancy, indicating these countries are undeveloped to low-developing.

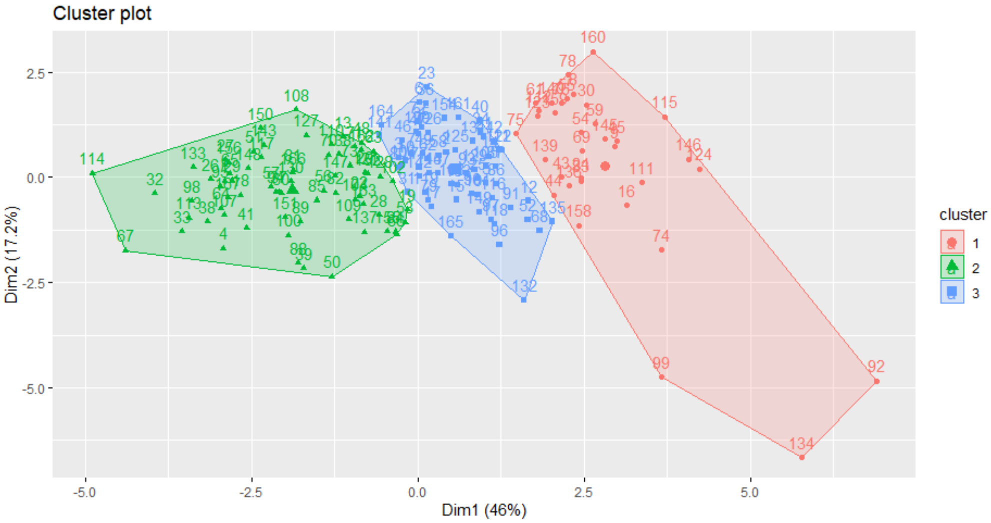


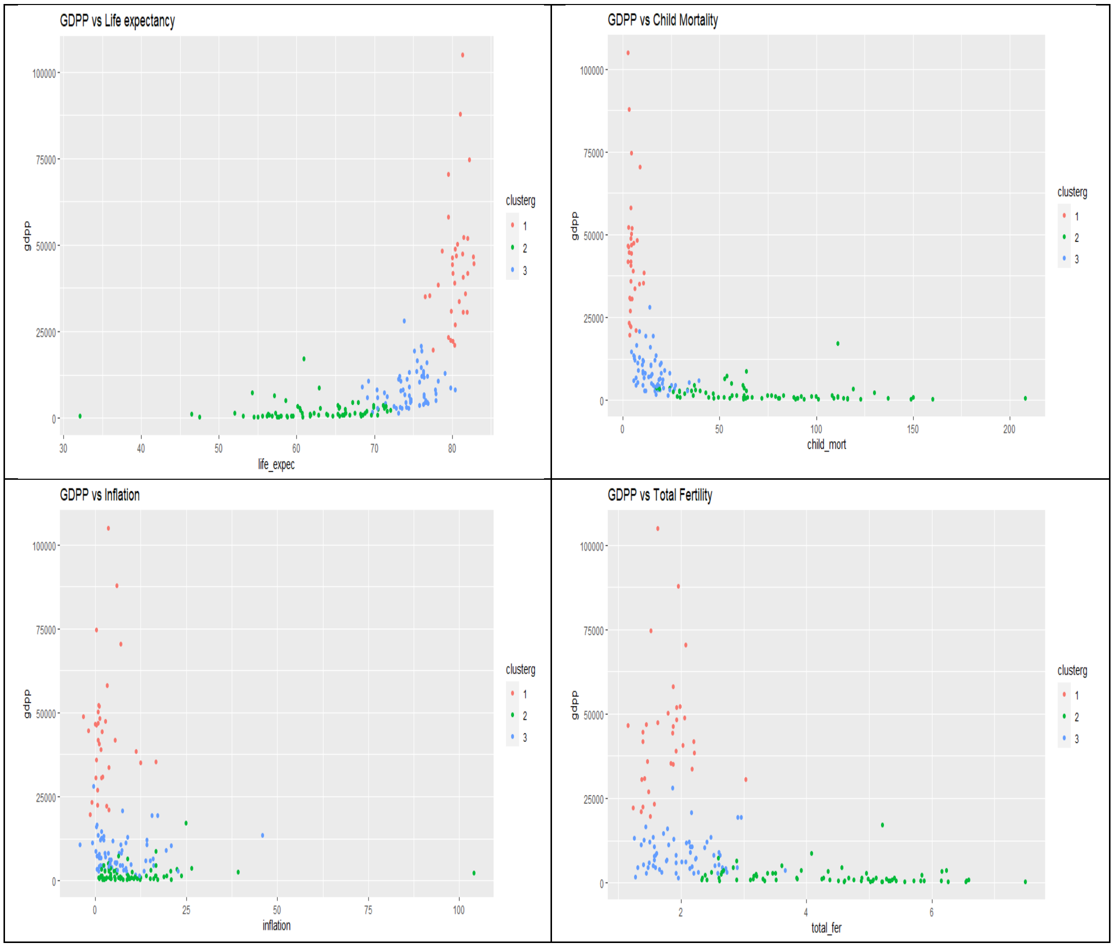




GMM (Gaussian Mixture Model)

To better compare with the other two models, I chose to use 3 Multivariate Normal Distributions to approximate this dataset. By attributing the data points to the cluster with the highest Log-likelihood, it was divided into 3 clusters. The results are very similar to those obtained by K-means, but with no overlap at all, indicating that the GMM clustering results are better.





5. Summary

Based on the results of the three clustering algorithms above, I found that K-means and GMM can better segment the data than Hierarchical clustering. Although the results of GMM and K-means are very close, GMM is more precise in segmentation. I speculate this is because K-means is affected by random initial values, while the initial values of GMM are obtained after K-means, making it better converge and more precise. The results of Hierarchical clustering are not satisfactory, possibly because it is less suitable for handling high-dimensional data.