STA 5650: Group Project

Ronald Lencevičius, Kaitlyn McGloin, Kevin Tsoi

December 11, 2020

Abstract

This paper aims to categorize countries based off of socio-economic and health factors that reflect the development of each country. To achieve this, we implement three different algorithms: k-means clustering, agglomerate hierarchical clustering, and divisive hierarchical clustering. We used three different metrics to find an optimal number of clusters to apply the algorithms to the full data set.

1 Source and Description of Data

The data used in this paper comes from Kaggle and the data was collected by HELP International. HELP International is a non-profit organization that helps provide emergency aid relief after natural disasters and works to develop programs that improve public health, education, entrepreneurship/business, and infrastructure development in countries around the world. The data collected is intended to help the organization pinpoint which specific countries could use the money raised to help improve the socio-economic and health conditions of those countries. We determine which countries are in need of funding by analyzing the socio-economic and health factors of each country.

A series of different socio-economic and health factors were recorded for each country. The data is provided for 167 countries. We will be categorizing the countries by considering the following nine different socio-economic and health factors reflective of the development of each country:

child mortality	death of children under the age of five per 1000 live births
exports	goods and services per capita; measured as a percentage of the GDP per capita
health	total health spending per capita; measured as a percentage of GDP per capita
imports	goods and services per capita; measured as a percentage of the GDP per capita
income	net income per person
inflation	annual growth rate of the total GDP
life expectancy	average lifespan of a new-born child in years
total fertility	number of children a woman gives birth
GDP	GDP per capita

Table 1: List of dataset features

2 Preliminary Analysis

Since we will perform k-means clustering in two dimensions and our dataset consists of nine features, we need to reduce the number of features to two features. To do this, we will use Principle Component Analysis. With R, we compute the principle components and then plot the cumulative proportion of variance explained by each principal component.

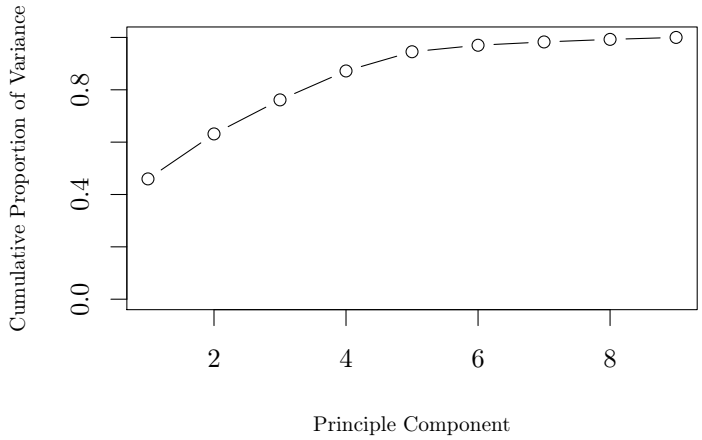

Figure 1: The cumulative proportion of variance explained by each principal component.

Based off the results, the first two principle components account for 63.13 percent of the proportion of the variance. These results are missing about 37 percent of the proportion of the variance. Thus, by including the third principle component, then 76.14 percent of the proportion of the variance is explained. To determine whether we should use the first two or first three principle components, we will first look at the scree plot, constructed using the loadings of the principle components.

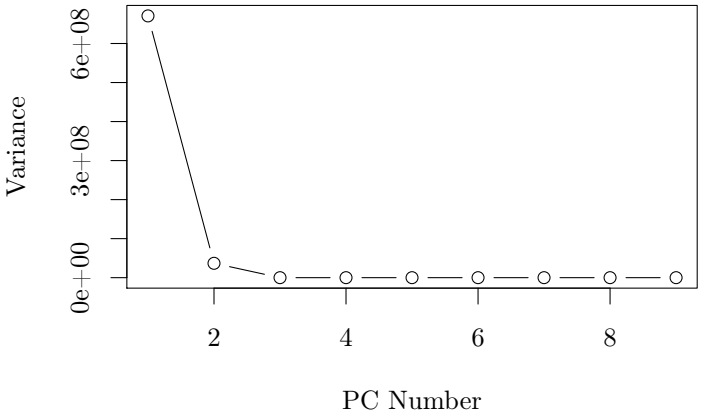

Figure 2: Scree plot to help us determine how many principle components to use.

From the above scree plot, we should use the first two principle components since the "elbow" appears to be at 2. However, since it is sometimes hard to distinguish where the "elbow" is located on a scree plot, we will also look at the biplots, PC1 vs. PC2 and PC2 vs. PC3, to determine which features greatly influence the principle components to help us choose whether to include the third principle component.

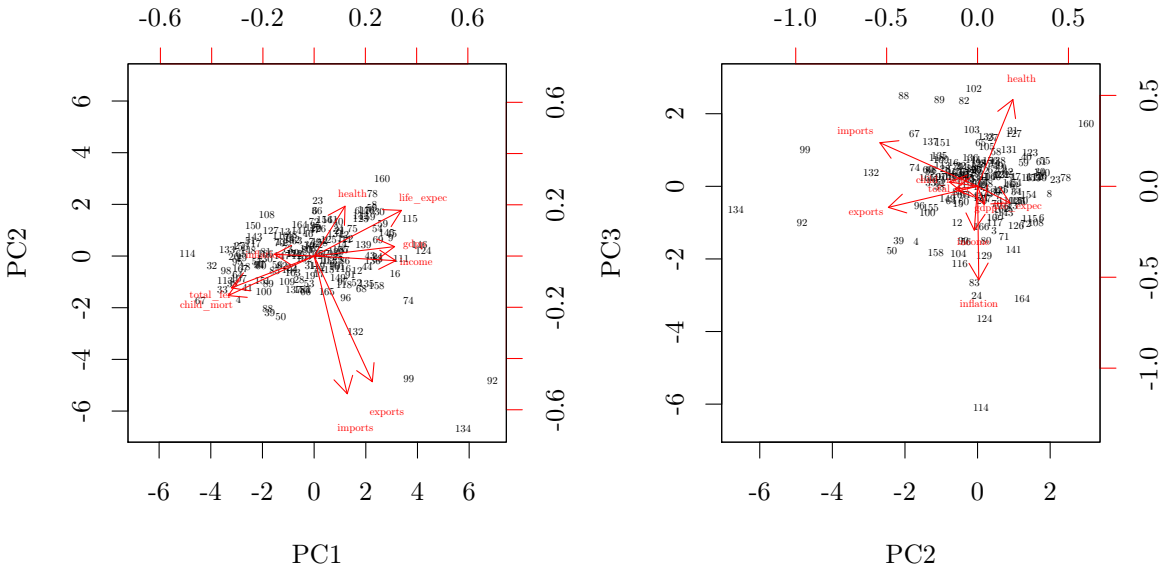

Figure 3: Biplots of PC1 vs PC2 and PC2 vs. PC3, left to right.

3 Main Analysis

3.1 K-Means Clustering

We conduct k-means clustering for all nine features of the data set.

Before applying k-means clustering, we must decide whether to scale the data. For this data set, we choose to scale the data since there are great differences in the variance of the nine features. We begin by performing k-means clustering for where $k = 2, 3, 4, 5$ and visualize the results by plotting the clusters.

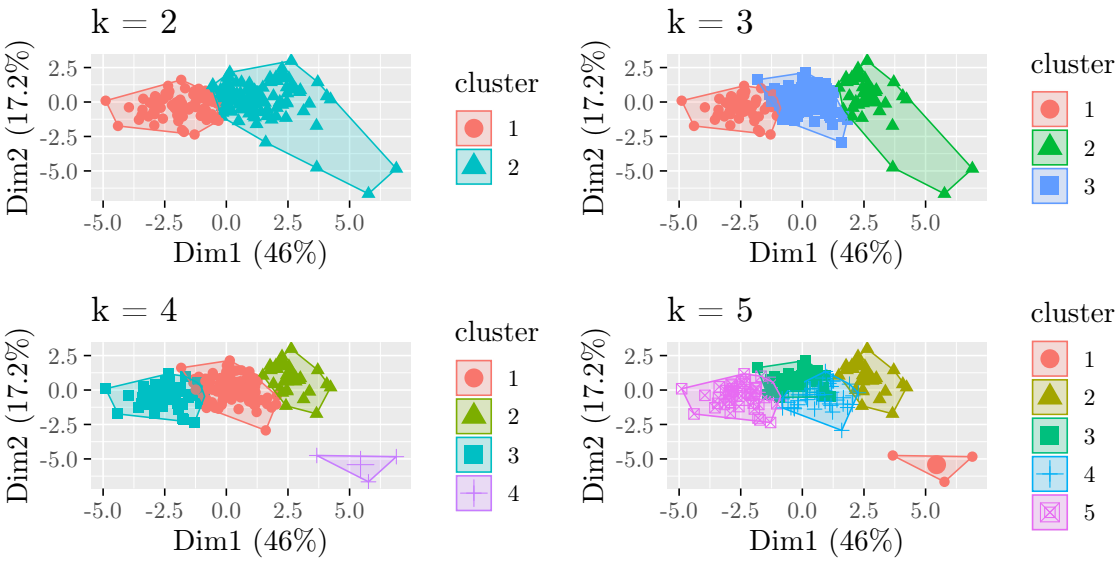

Figure 4: K-means Clustering with all nine features for $k = 2, 3, 4, 5$.

However to determine which number of clusters we should use, we will use three different methods to help us determine the optimal number of clusters. The three methods are called the Elbow Method, Average Silhouette Method, and the Gap Statistic Method. The Elbow Method measures the compactness of the clustering by utilizing the metric total within-cluster sum of squares. The Average Silhouette Method measures the quality of a clustering. In other words, it measures how well each country fits in a cluster. A high average silhouette width is desirable. The Gap Statistic Method compares the total variation within a cluster for k different values with their expected values under a null distribution.

Using R to apply the three methods, we get the following plots:

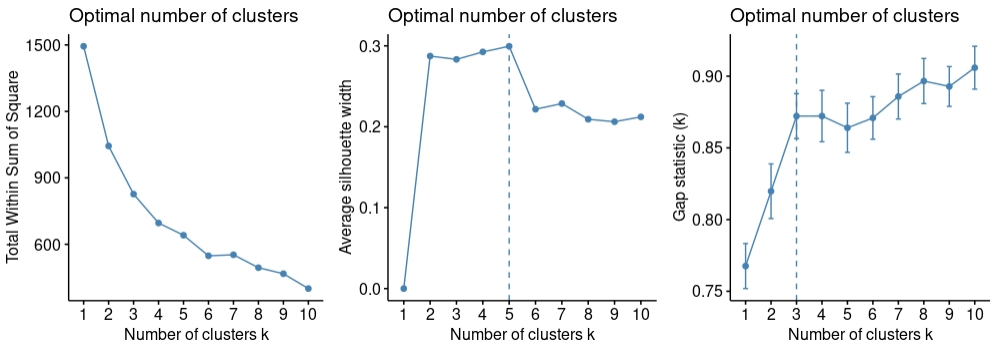

Figure 5: The results from the Elbow Method, Average Silhouette Method, and Gap Statistic Method, left to right.

Although there are conflicting results of the optimal number of clusters by the three methods, we will use $k = 4$. We visualize the four clusters of the countries by plotting a scatter plot matrix.

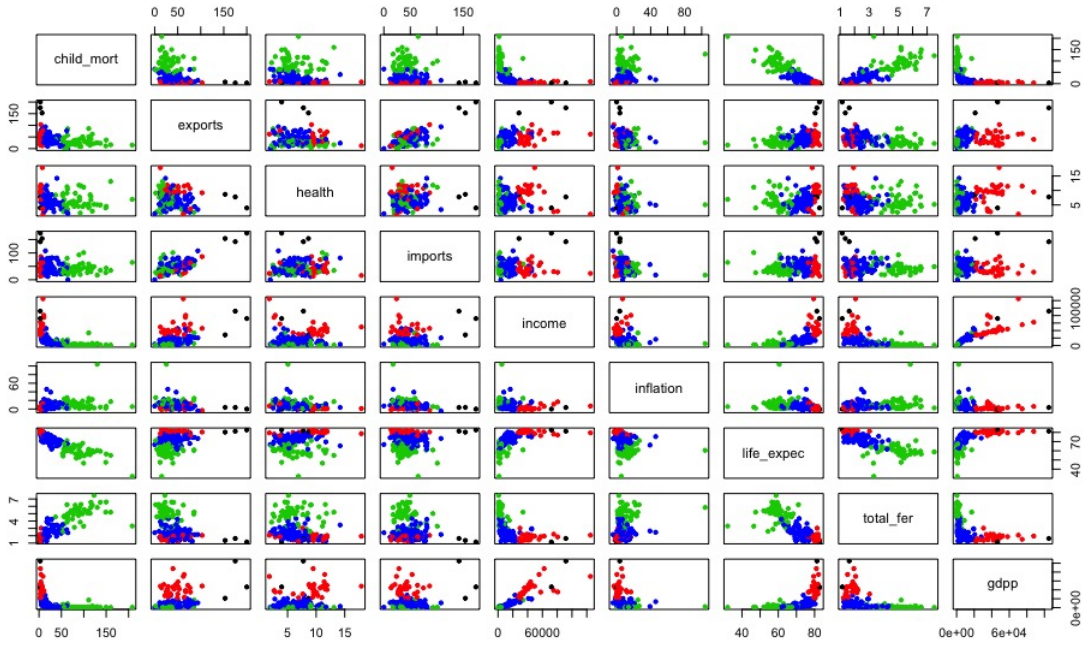

Figure 6: Scatter plot matrix showing final K-Means clustering algorithm of the full dataset.

Using the above scatter plot, we determine the extent of development of each country based off of the socio-economic and health factors.

The countries in the 1st cluster, represented in black, contains only three countries: Luxembourg, Malta, and Singapore. These countries are characterized as having significantly higher rates of exports and imports and the highest net income per person. In addition, the countries of this cluster have high life expectancy rates and low rates of child mortality. The 2nd cluster consists of countries like Australia, Canada, France, the United Kingdom, and the United States. These countries are represented in green and tend to have a higher rate of total health spending per capita compared to the other clusters of countries. These countries also have low rates of child mortality and high life expectancy rates. The 3rd cluster is represented in blue and contains countries such as China, Hungary, Poland, and Russia. These countries tend to have a high rate of child mortality and a low life expectancy rate. Lastly, the 4th cluster, represented in red, consists of countries like Afghanistan, Haiti, Iraq, Kenya, and Nigeria. These countries have the highest rates of child mortality, inflation, and total fertility in comparison to the countries in the other clusters. In addition, these countries also have the lowest life expectancy rate. The table below shows which country is in which cluster.

C	#	Countries	Type
1	3	Luxembourg Malta Singapore	Excellent
2	30	Australia Austria Belgium Brunei Canada Cyprus Denmark Finland France Germany Greece Iceland Ireland Israel Italy Japan Kuwait Netherlands New Zealand Norway Portugal Qatar Slovenia South Korea Spain Sweden Switzerland United Arab Emirates U.K. U.S.	Good
3	87	Albania Algeria Antigua/Barbuda Argentina Armenia Azerbaijan Bahamas Bahrain Bangladesh Barbados Belarus Belize Bhutan Bolivia Bosnia/Herzegovina . . . . Tonga Tunisia Turkey Turkmenistan Ukraine Uruguay Uzbekistan Vanuatu Venezuela Vietnam	Adequate
4	47	Afghanistan Angola Benin Botswana Burkina Faso . . . . Senegal Sierra Leone So. Africa Sudan Tanzania Timor-Leste Togo Uganda Yemen Zambia	Poor

Table 2: Table identifying countries in each cluster(K-means)

3.2 Hierarchical Clustering: Agglomerate

The agglomerate clustering or AGNES is the most common type of hierarchical clustering which starts by treating every observation as a cluster. Then, the pairs of clusters are grouped based on its similarity until all clusters are merged into one large cluster containing all observations. The results can be represented in a tree-like graph called a dendrogram.

Before performing agglomerate clustering, we need to decide what linkage method to use to calculate dissimilarities between observations. A metric we used is called the agglomerative coefficient which measures the strength of the clustering structure. Coefficients approaching one implies a more balanced clustering structure while coefficients approaching zero suggest less well-formed structures. There are several different linkage methods out there, but we decided to look into these four linkage methods: Average, Single, Complete, and Ward. We found that the Ward method had performed the best compared to the other three methods and will be chosen for the clustering process.

In Figure 7, the dendrogram is constructed by using agglomerate clustering with Ward's method. To compare to the results of k-means, we chose four clusters which are distinctly represented by the colored boxes.

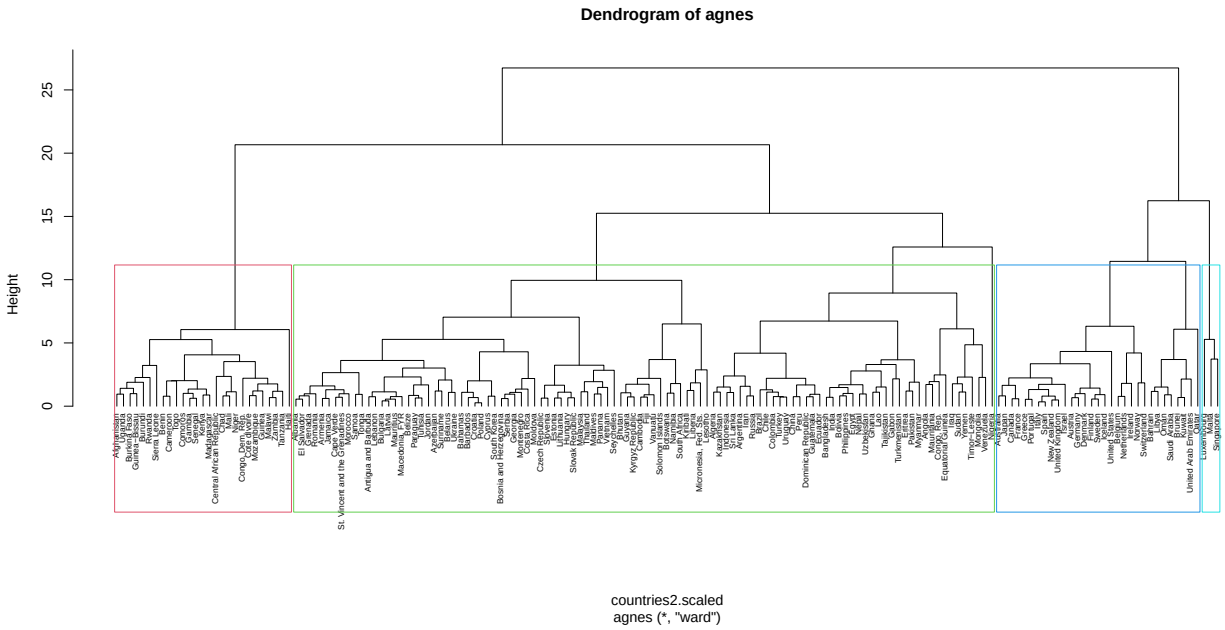

Figure 7: Dendrogram from agglomerate clustering using Ward's method

To understand which country is in what cluster, Table 2 shows a table classifying countries based on their socio-economic and health factors.

C	#	Countries					Type
1	3	Luxembourg	Malta	Singapore			Excellent
2	31	Australia	Japan	Canada	France	Greece	Good
		Portugal	Italy	Spain	New Zealand	U.K.	
		Israel	Austria	Germany	Denmark	Finland	
		.	.	.	.	.	
		.	.	.	.	.	
		.	.	.	.	.	
		Norway	Switzerland	Bahrain	Libya	Bahrain	
		Saudi Arabia	Brunei	Kuwait	United Arab Emirates	Qatar	
3	106	Albania	El Salvador	Grenada	Romania	Armenia	Adequate
		Jamaica	Cape Verde	St. Vincent	Morocco	Samoa	
		Tonga	Antigua/Barbuda	Lebanon	Bulgaria	Latvia	
		.	.	.	.	.	
		.	.	.	.	.	
		.	.	.	.	.	
		South Korea	Equatorial Guinea	Turkey	Iraq	Sudan	
		Yemen	Timor-Leste	Mongolia	Venezuela	Nigeria	
4	27	Afghanistan	Uganda	Burkina Faso	Guinea-Bissau	Burundi	Poor
		Rwanda	Sierra Leone	Benin	Cameroon	Togo	
		.	.	.	.	.	
		.	.	.	.	.	
		.	.	.	.	.	
		Guinea	Malawi	Zambia	Tanzania	Haiti	

Table 3: Table identifying countries in each cluster (Agglomerative

The cluster in light blue, dark blue, green, and red represents the first, second, third and fourth clusters respectively. Table 2 has very similar results to that of k-means. Comparing to k-means, the notable differences are that Libya and Bahrain are moved into cluster two and South Korea is found in cluster 3. Considering how developed South Korea is, this can be considered as a misclassification from this clustering method.

3.3 Hierarchical Clustering: Divisive

Unlike agglomerative clustering, divisive clustering or DIANA starts by having a single cluster. Then, through each iteration, the most heterogeneous cluster is divided into until all observations have its own cluster. Similarly, the results can be represented through a dendrogram. In addition, divisive clustering does not need to specify a linkage method between observations. Similar to the agglomerate coefficient, we can calculate the divisive coefficient which measures the strength of the clustering structure. We found that the divisive coefficient was lower than the agglomerate coefficient with Ward's method. Although the divisive clustering was worse with this metric, we decided to explore what clusters this algorithm provides.

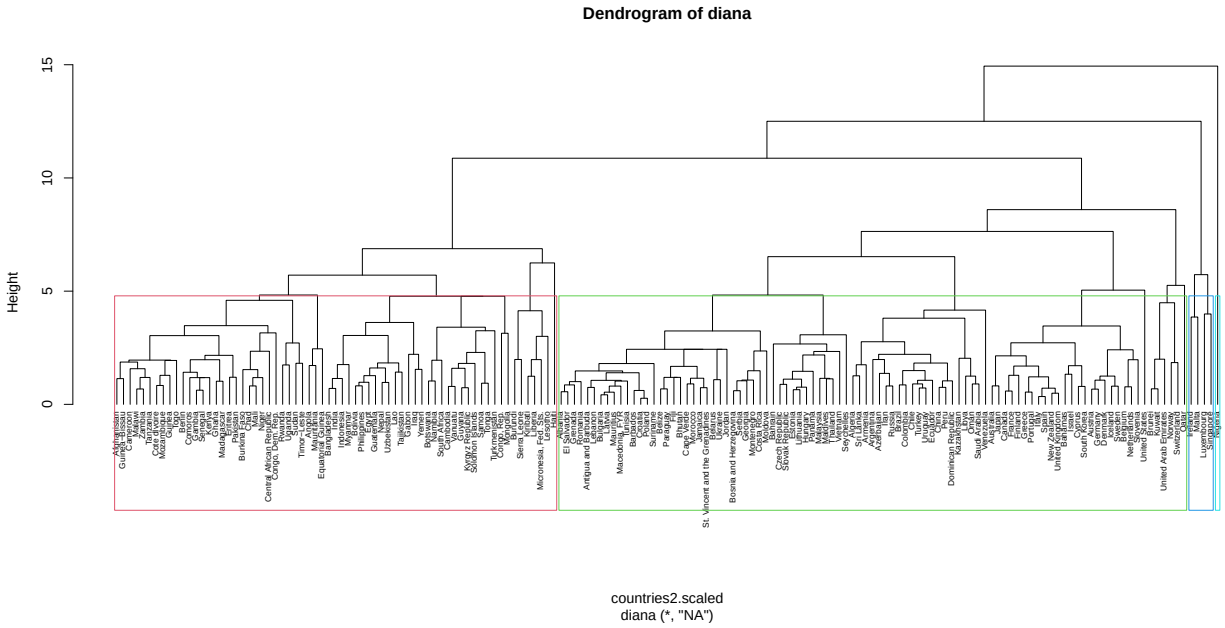

Figure 8: Dendrogram from divisive clustering

In Figure 8, the dendrogram is constructed using divisive clustering. We can see that it is more skewed to the left compared to the other the agglomerate dendrogram. Again, to compare with the other clustering algorithms, we chose four clusters as indicated by the colored boxes.

Table 3 shows a table classifying countries based on their socio-economic and health factors.

C	#	Countries					Type
1	4	Ireland	Malta	Luxembourg	Singapore		Excellent
2	95	Albania	El Salvador	Grenada	Romania	Antigua/Barbuda	Good
		Lebenon	Bulgaria	Latvia	Mauritius	Macedonia, FYR	
		⋮	⋮	⋮	⋮	⋮	
		Congo, Rep.	Equatorial Guinea	Turkey	Iraq	Sudan	
		Yemen	Timor-Leste	Mongolia	Venezuela	United States	
3	67	Afghanistan	Guinea-Bissau	Cameroon	Malawi	Zambia	Adequate
		Tanzania	Cote d'Ivoire	Mozambique	Guinea	Togo	
		⋮	⋮	⋮	⋮	⋮	
		Sierra Leone	Burundi	Mongolia	Congo, Rep.	Turkmenistan	
		Kiribati	Liberia	Micronesia	Lesotho	Haiti	
4	1	Nigeria					Poor

Table 4: Table identifying countries in each cluster (Divisive

The cluster in dark blue, green, red and light blue represents the first, second, third and fourth clusters respectively. Table 3 has drastically different results from that of k-means and the agglomerative algorithm clusters. From cluster one, we found that Ireland has been clustered into countries that tend to have the best socio-economic and health status. Cluster two contained countries that have a very large variation of socio-economic and health factors. For instance, United States and Venezuela are in this cluster, yet have extremely different economies and healthcare. Cluster three contained countries that are third-world countries which do not have fairly good socio-economic and health conditions. Finally, in cluster four, we only have Nigeria. We weren't really too sure why or how this algorithm separated Nigeria from cluster 3 but from looking at the data, we found that Nigeria has the highest inflation rate out of all the countries in this data set. We believe this extreme value has made a big influence on how Nigeria was clustered.

4 Conclusion

The goal of this report is to analyze and categorize countries based on their socio-economic and health factors to determine which country is in need of funding. To categorize these countries, we implemented three different clustering algorithms and had varying results. From looking at the countries in each cluster for all the algorithms, we found that k-means and agglomerate clustering performed similarly well in clustering these countries. On the other hand, divisive clustering performed poorly due to the variation of socio-economic and health factors particularly in its second cluster.