Find out which countries need our aid using Machine Learning Models

A non profit organization, Help International's mission is to fight poverty and empower people. The organization connects people and companies to communities that need the most of our resources. In many ways, it creates sustainable progress, maintains urban development programs, and connects with local people that are in need. There are a lot of ways to contribute but in this case the organization encourages us to donate directly to projects or partner up with volunteers to participate in some of the many carefully evaluated projects to alleviate poverty. Ultimately, Help International's goal is to create a group of network and social entrepreneurs who can support the difficulties that impoverished communities face daily by being an active participant in an effort to improve the lives of the most vulnerable populations in the world.

The organization successfully raised nearly $10 million and it's going to be distributed strategically for those who are in need. But in order to figure out which countries need the aid, we are going to focus on a data driven approach. So it's important to categorize the countries by utilizing socio - economic and health factors that can help in deciding the improvement of the country. So based on the clusters of the countries which depend on their individual conditions, the charity allocated is based on disasters and natural catastrophes.

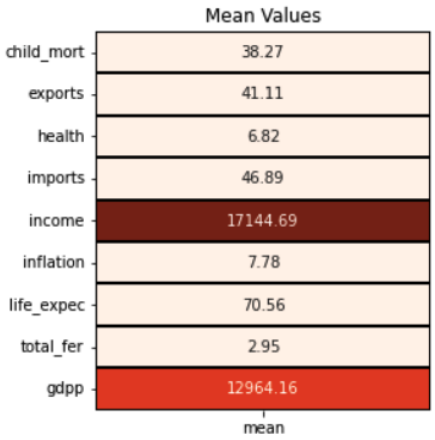
Data Summary

To cluster countries depending on some numerical features. This project aims to show a case study of unsupervised learning wherein we find clusters of each country that are based on selected features.

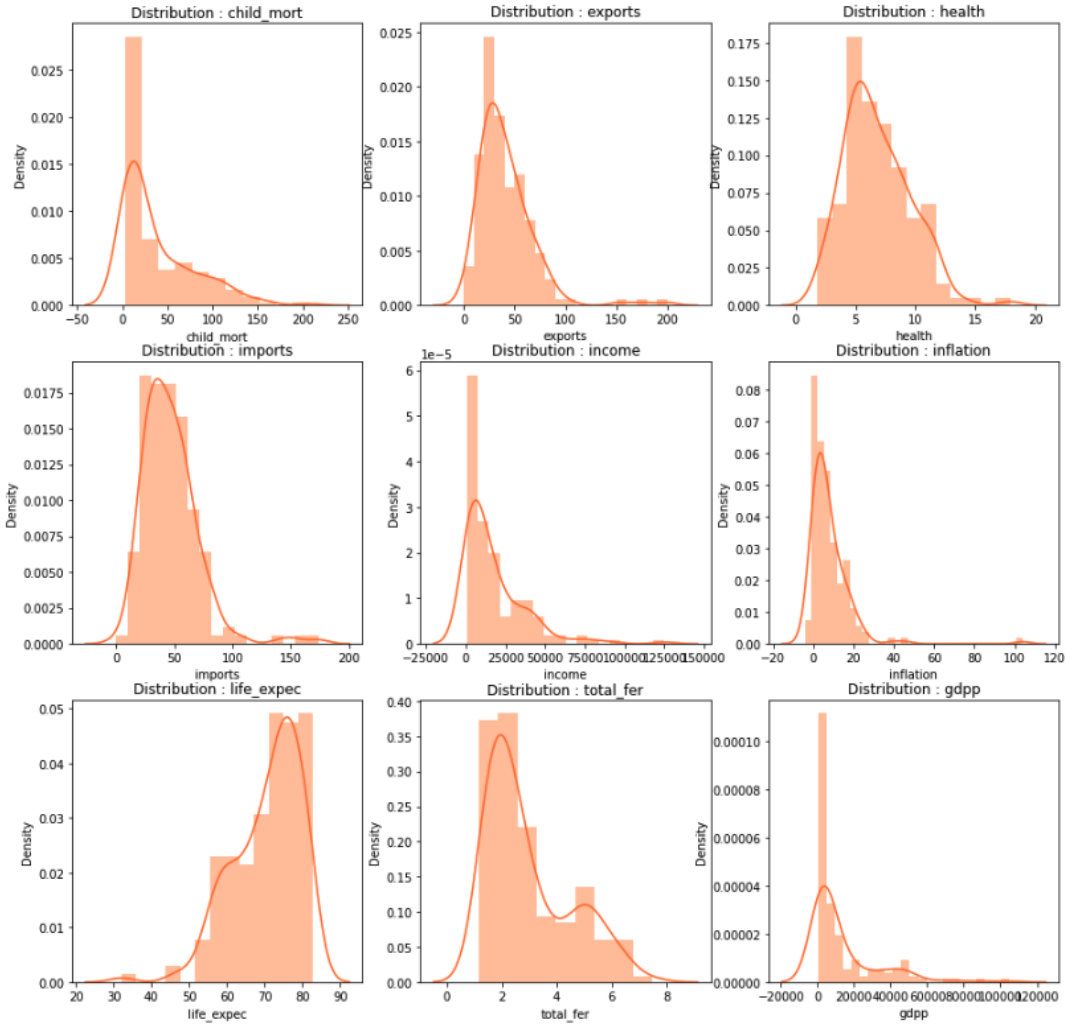
- **country** : Name of the country
- **child_mort** : Death of children under 5 years of age per 1000 live births
- **exports** : Exports of goods and services per capita. Given as %age of the GDP per capita
- **health** : Total health spending per capita. Given as %age of GDP per capita
- **imports** : Imports of goods and services per capita. Given as %age of the GDP per capita
- **Income** : Net income per person
- **Inflation** : The measurement of the annual growth rate of the Total GDP
- **life_expec** : The average number of years a newborn child would live if the current mortality patterns are to remain constant
- **total_fer** : The number of children that would be born to each woman if the current age-fertility rates remain
- **gdpp** : The GDP per capita. Calculated as the Total GDP divided by the total population.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.20	10.00	7.58	44.90	1610	9.44	56.20	5.82	553
1	Albania	16.60	28.00	6.55	48.60	9930	4.49	76.30	1.65	4090
2	Algeria	27.30	38.40	4.17	31.40	12900	16.10	76.50	2.89	4460
3	Angola	119.00	62.30	2.85	42.90	5900	22.40	60.10	6.16	3530
4	Antigua and Barbuda	10.30	45.50	6.03	58.90	19100	1.44	76.80	2.13	12200

The data contains 167 rows and 10 features, cleaned and removed null values. Here's a summary of the data in a heat map.



For this dataset, the number of features are less and we're going to check each dataset. We also know that each variable in the model is quantitative with an element of either a float or integer. Down below, we have the distribution of each feature.



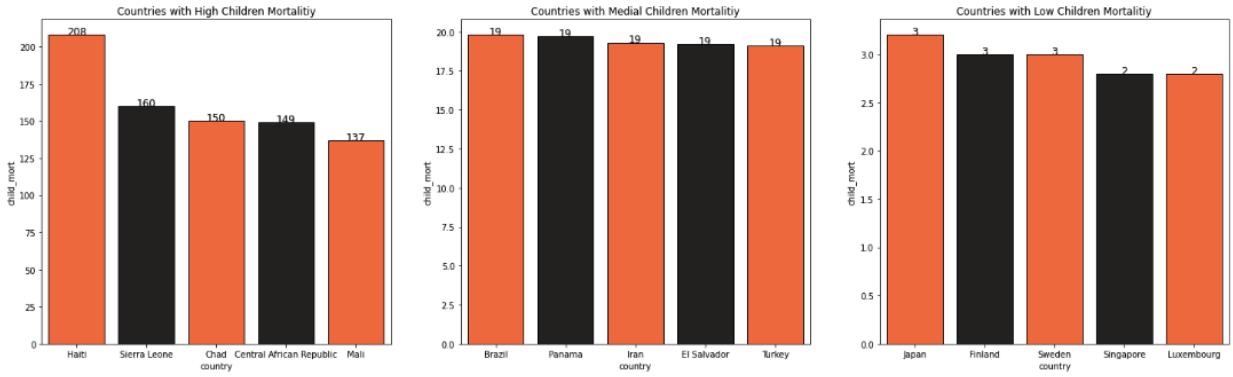
We can see on the third row first column, life expectancy displays a skewed distribution having a left or negative distribution. Health has normally distributed data and while all the rest of the features present a right or positively skewed distribution.

Features of Countries that needs the most aid

- Per capita income is low
- High population leaving low resources
- High unemployment because of low resources
- Low country wealth that leads to low capital
- High disparity of wealth and income
- Absence of educational resources
- Low level of living standards
- No technical advancement.
- Need of health service

The organization is going to target those countries that are showing the above characteristics. Here we will display the data to find out which countries are within the extremes or the center of each feature to find out which countries need us.

Country by Children Mortality

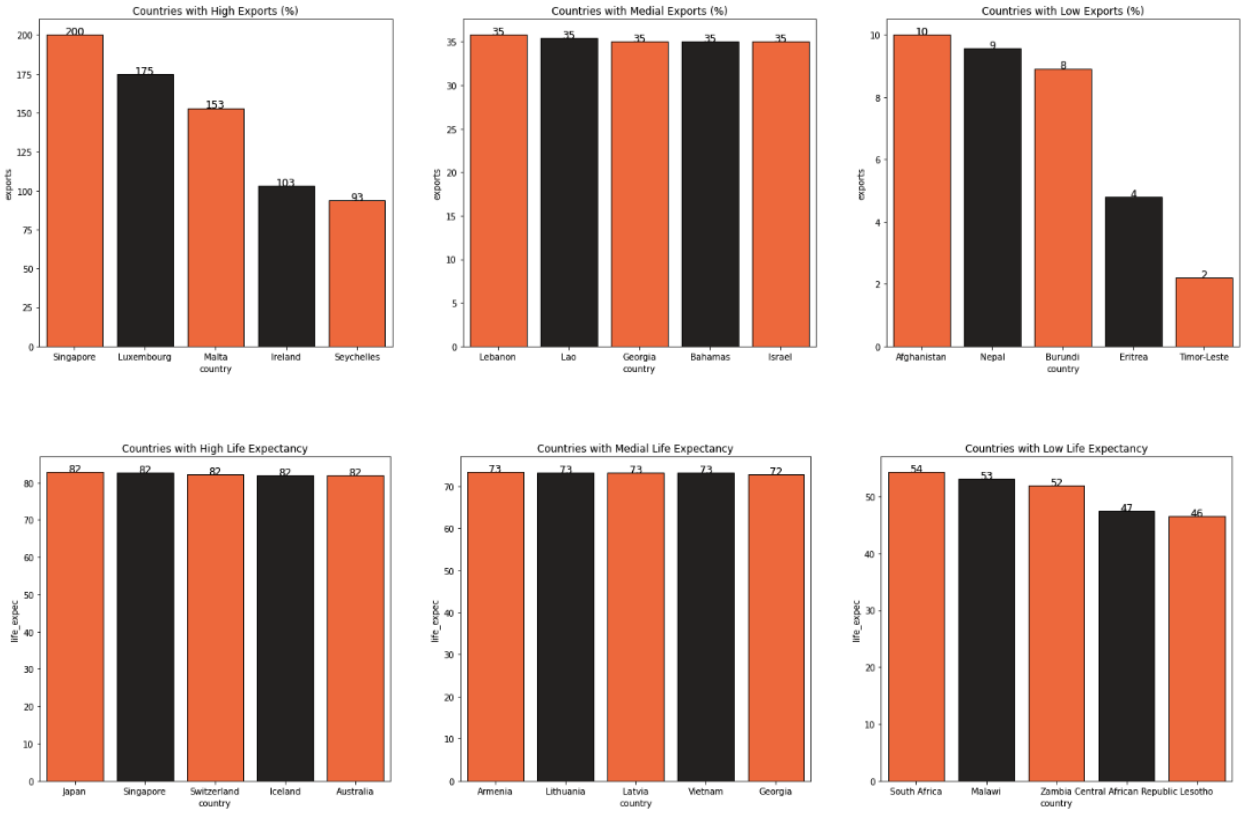


The three graphs display countries with the highest children mortality rate and in this case we clearly see that Haiti has the highest. Note that each graph represents different scales, so the far right begins from 0 - 3. While the left wing is scaled out from 0 - 200. In this case we can see that African countries have a significantly higher statistic. Asia and Europe do hold significance in this presentation too.

Country by Exports

The graph below shows high, medium and low export percentage. We know that exports are a good measurement of the country's health when they sell goods and services. But it also depends on the geographical location, natural resources, and population size. So we can see that Afghanistan and Nepal

are in the lower tail of the exports. I believe it's because of the geographical location which influences its ability to export.

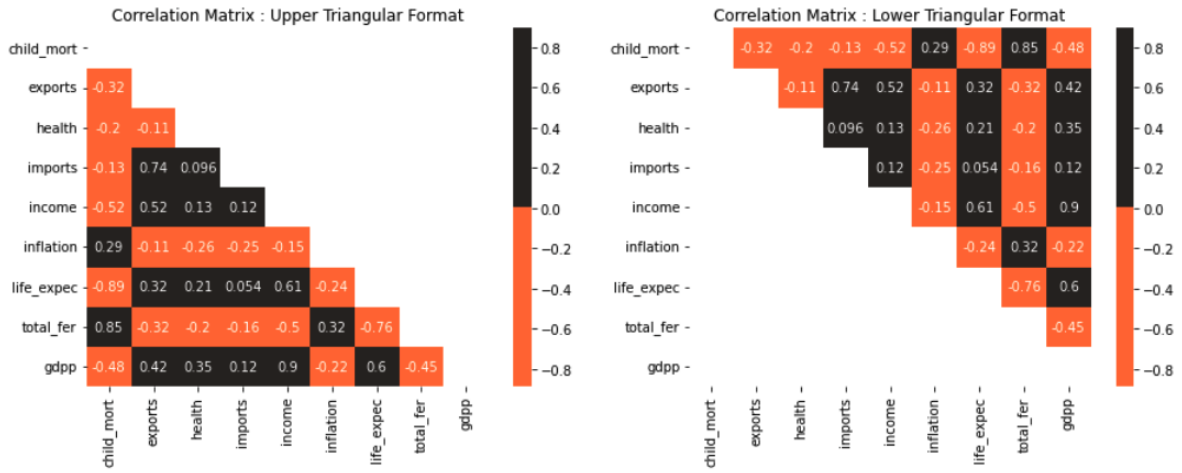


As we continue the visual exploration of the data we have and from it we can summarize the countries that are in need. We found out that African countries are at risk and hold the highest in child mortality and low life expectancy. While all these difficulties present serious challenges like the United States present the most spenders in health but they're not in the top 5 ranks of life expectancy. Singapore, Malta, and Luxembourg are top 5 export countries as well as imports. Population size and geographical locations have an effect when it comes to GDP.

The graph below shows a possibility of multicollinearity as you might find correlations that are significantly strong enough between the features. Note that child mortality increases when GDP, exports, and income goes down. The rise in inflation also causes child mortality, so all these economic factors have an effect on infant mortality. The rise in exports presents an increase in GDP, imports, and income. Observe that income and GDP has a high correlation at 0.9, so high income brings into higher life expectancy.

We can see that some of the variables are almost from the same category and shows the same effect to other features like for example:

Health – child mortality, life expectancy, health
Trade – imports, exports
Finance – GDP, income, inflation



Modeling / Method

In machine learning systems, we often group examples as the first step towards understanding the dataset. Grouping an unlabelled example is called clustering. As the samples are unlabelled, clustering relies on unsupervised machine learning. If the examples are labeled, then it becomes classification.

Clustering is a type of unsupervised learning where the references need to be drawn from unlabelled datasets. Generally, it is used to capture meaningful structure, underlying processes, and grouping inherent in a dataset. In clustering, the task is to divide the population into several groups in such a way that the data points in the same groups are more similar to each other than the data points in other groups. In short, it is a collection of objects based on their similarities and dissimilarities.

Given a set of observations ($x_1, x_2, \ldots x_n$) where each observation is a d-dimensional real vector, k means clustering aims to partition the n observations into k ($< n$) sets S = {$S_1, S_2 \ldots S_k$) so to minimize the within cluster sum of squares or WCSS

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Where μ_i is the mean of points in S_j and this is equivalent to minimizing the pairwise square deviation of points in the same cluster:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{\mathbf{x},\mathbf{y}\in S_i} \|\mathbf{x}-\mathbf{y}\|^2$$

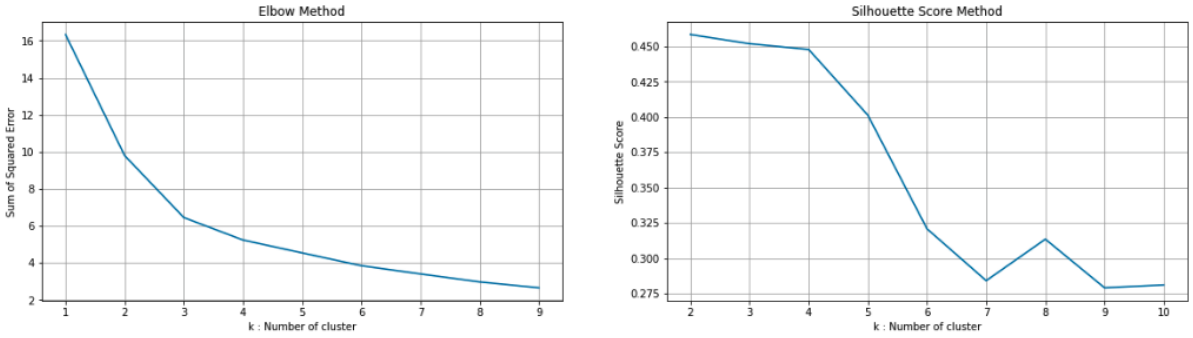
The equivalence can be deduced from identity

$$|S_i| \sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \sum_{\mathbf{x}\neq\mathbf{y}\in S_i} \|\mathbf{x}-\mathbf{y}\|^2$$

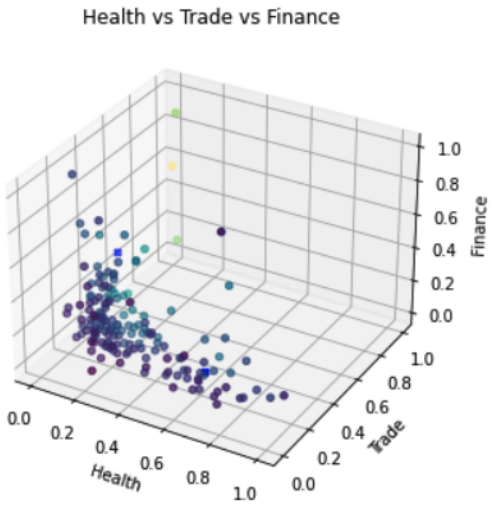
Since the total variance is constant, this is the same as maximizing the sum of squared deviation between points in different clusters

Statistical Test for selection the values of K

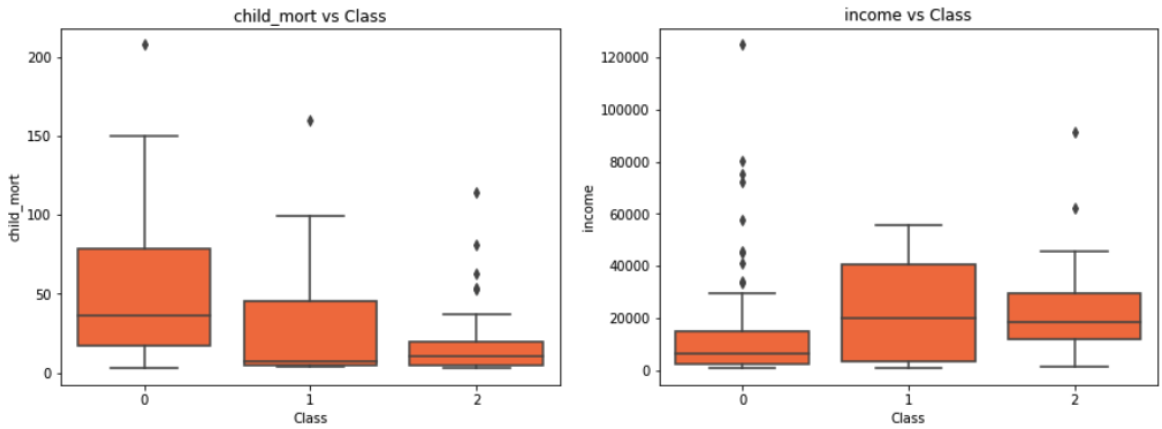
Elbow method and Silhouette Score Method using the feature combination of Health, Trade, and Finance



From here we select k as Clusters = 3



Although the plot doesn't really inform us which value corresponds to what but we do know that low income and high child mortality is not a good indicator of a healthy country



0: Help needed – 1: May need help – 2: No help

Reference

1. https://www.who.int/data/data-collection-tools/who-mortality-database
2. https://help-international.org/
3. https://neptune.ai/blog/k-means-clustering#:~:text=The%20objective%20is%20to%20minimize,a%20mean%20value%20to%20each.
4. https://en.wikipedia.org/wiki/K-means_clustering
5. https://towardsdatascience.com/k-means-explained-10349949bd10