Assignment: Part I

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.  The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

Objectives

Your main task is to cluster the countries by the factors mentioned above and then present your solution and recommendations to the CEO using a PPT.  The following approach is suggested :

Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of components k. You need to perform your clustering activity on this new dataset, i.e. the PCA modified dataset with the k components.

Outlier Analysis: You must perform the Outlier Analysis on the dataset, before or after performing PCA, as per your choice. However, you do have the flexibility of not removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, all you need to do is find the outliers in the dataset, and then choose whether to keep them or remove them depending on the results you get.

Try both K-means and Hierarchical clustering(both single and complete linkage) on this dataset to create the clusters. [Note that both the methods may not produce identical results and you might have to choose one of them for the final list of countries.]

Analyse the clusters and identify the ones which are in dire need of aid. You can analyse the clusters by comparing how these three variables - [gdpp, child\_mort and income] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries. Note that you perform clustering on the PCA modified dataset and the clusters that are formed are being analysed now using the original variables to identify the countries which you finally want to select.

Also, you need to perform visualisations on the clusters that have been formed.  You can do this by choosing the first two Principal Components (on the X-Y axes) and plotting a scatter plot of all the countries and differentiating the clusters. You should also do the same visualisation using any two of the original variables (like gdpp, child\_mort, etc.) on the X-Y axes as well. You can also choose other types of plots like boxplots, etc.

The final list of countries depends on the number of components that you choose and the number of clusters that you finally form. Also, both K-means and Hierarchical may give different results. Hence, there might be some subjectivity in the final number of countries that you think should be reported back to the CEO. Here, make sure that you report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

Assignment: Part II

This part of the assignment is subjective and hence, you are required to write the answers and submit them in a PDF file. For writing normal text, you can use MS Word — or any other similar software which can convert word files to the .pdf format. For writing equations, drawing figures, etc. you can do so on a blank sheet of paper, photograph the images and upload them in the same word document.

Please limit your answers to 200-300 words per question. While calculating values, ensure you write all the necessary steps and formulae. Also, use the correct terminology to present the solution.

**Question 1: Assignment Summary**

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly( why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

**Note**: You don't have to include any images, equations or graphs for this question. Just text should be enough.

**Question 2: Clustering**

      a) Compare and contrast K-means Clustering and Hierarchical Clustering.  
      b) Briefly explain the steps of the K-means clustering algorithm.   
      c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well            as the business aspect of it.  
      d) Explain the necessity for scaling/standardisation before performing Clustering.  
      e) Explain the different linkages used in Hierarchical Clustering.

**Question 3: Principal Component Analysis**

      a) Give at least three applications of using PCA.  
      b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.  
      c) State at least three shortcomings of using Principal Component Analysis.