SUBJECTIVE QUESTIONS

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 (what EDA you performed, 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.
Answer:
Below is the problem statement given for this assignment:
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.
We have been provided with a data set containing list of countries having child mortality rate, exports, health, imports, income, inflation, life expectancy, total fertility and GDPP values defined.
Methodology used:
We have followed below steps.

1. **Data Understanding:**
   a) **Data Description:**
   In this step, we started by loading the dataset, reading it, inspecting the data set by finding out its shape, columns, information of every columns, describing it by finding its mean, standard deviation, percentile at 25%, 50%, 75% and its minimum and maximum values. Next, we checked for missing values in each column and found that there were no missing values. Also, we looked for each columns information and found that there was no inconsistency in the data . No duplicate values were found.
   We converted imports, exports and health spending from percentage values to actual values of their GDP per capita. The correlation of variables was found by plotting a heatmap.
   b) **EDA**
   Univariate and Bivariate analysis were performed here.
   We are required to choose the countries that are in the direst need of aid. Therefore, we need to identify those countries which are using some socio-economic and health factors, that determine the overall development of the country.
      a) In the **Univariate Analysis**, barplot for all variables were made and below inferences can be made.
      Child mortality rate is highest in Haiti
      Fertility rate is highest in Niger and lowest in Nigeria
      Life expectancy is highest in Afghanistan and lowest in Haiti
      Health situation is worst in Eritrea and highest in Burundi
      GDP per capita is highest in Eritrea
      Sierra Leone has highest per capita income
      Inflation is highest in Nigeria
      Myanmar is lowest in exports and imports

      b) In the **Bivariate Analysis**, we made a pairplot of all numerical variables in the dataset.
   We are required to choose the countries that are in the direst need of aid. Therefore, we need to identify those countries which are using some socio-economic and health factors, that determine the overall development of the country.

2. **Performed Clustering**
   a) **Data Preparation for Clustering**

i) **Outlier Treatment**: Boxplots of all variables were made which are just helpful in detecting the outliers. Later, we caped the outliers to values accordingly for analysis. The capping boundary taken is 0.1 and 0.99. We haven't removed the outliers as the dataset is not so large and removing outliers can lead the dataset to become smaller.

ii) **Hopkins Check:** The Hopkins statistic is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered. Hopkins value came out to be 0.92 which means that dataset has a tendency to form cluster as its value is near to 1.

iii) **Scaling:** For scaling, we first prepared the Numerical columns in a separate dataframe. As we can scale numerical values only . After scaling I describe the data set and now all values are in between -1 to +1 and my data is now ready for clustering.

b) Clustering

i) K-means Clustering:

First I started with K means clustering and below steps has been formed for K Means Clustering:

1. Choose K using both Elbow and silhouette score: I used Elbow and silhouette score methods to determine the optimal number of clusters into which the data may be clustered. From the Elbow Curve graph I decided to take 3 as my cluster value (K).Secondly from silhouette score method since the silhouette score for cluster 3 is the second maximum value so I opted cluster value as 3.
2. Run K means with chosen K: Then I used Kmeans library and performed clustering with 3 clusters.
3. Visualize clusters: I have create new column named cluster_id in my original data frame and assigned value of Kmeans.labels_ to that column.
4. Cluster Profiling using gdpp, child_mort, income:
   From the problem statement I have learnt that Child_Mortality, Income, Gdpp are some important factors which decides the development of any country. Hence, I proceeded with cluster profiling by using these 3 variables. Then I created boxplots of these 3 variables with country and found that cluster 0 has lowest gdpp, lowest income and highest child mortality rate so I can say cluster 0 countries need some aid.

ii) Hierarchical Clustering:

Now after Kmeans I choosed Hierarchical Clustering method to perform clustering:

1. Using both single and complete linkages: Here I created Dendogram for both Single and Complete Linkage. From the complete linkage Dendogram I found that I can cut my cluster at 3 which is most optimum so I will go with 3 clusters. As in Single linkage I didn't got much inference as the dendogram got overlapped.
2. Visualize clusters: I have create new column named cluster_labels1 in my original data frame and assigned value of cluster_labels_ to that column.
3. Cluster profiling using gdpp, child_mort, income:
   From the problem statement I have learnt that Child_Mortality, Income, Gdpp are some important factors which decides the development of any country. Hence, I proceeded with cluster profiling by using these 3 variables.Then I created boxplots of these 3 variables with country and found that cluster 0 has lowest gdpp, lowest income and highest child mortality rate so I can say cluster 0 countries need some aid.

c) Country Identification and Conclusion:

Based on analysis, choose countries in need of aid:

I have analyzed both K-means and Hierarchical clustering and found clusters formed are also identical. The clusters formed in both the cases are great and I can choose anyone of the method. So, I will proceed with the clusters formed by hierarchical clustering as we know whenever we have less data set we should go with hierarchical clustering and based on the information provided by the final clusters I will deduce the final list of countries which are in need of aid.

Choosing countries based on some socio-economic and health factors

Top 10 countries that are in the direst need of aid and on which the CEO needs to focus on the most and provide Aid to these countries:

1. Liberia
2. Congo, Dem. Rep.
3. Burundi
4. Niger
5. Central African Republic
6. Mozambique
7. Malawi
8. Guinea
9. Togo
10. Sierra Leone

Question 2: Clustering

a) **Compare and contrast K-means Clustering and Hierarchical Clustering.**

Answer:

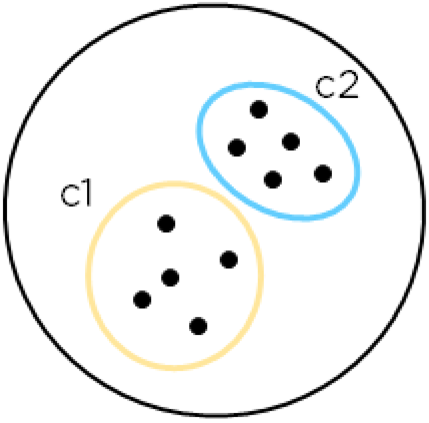
Clustering:

Clustering is a process of keeping similar data into groups. Clustering is an unsupervised learning technique as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data.

In machine learning, clustering is used for analyzing and grouping data which does not include pre-labelled class or even a class attribute at all.

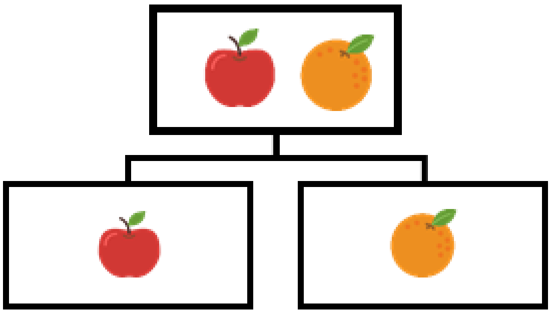
K-means Clustering:

Cluster analysis or simply k means clustering is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.



Hierarchical clustering:

In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.



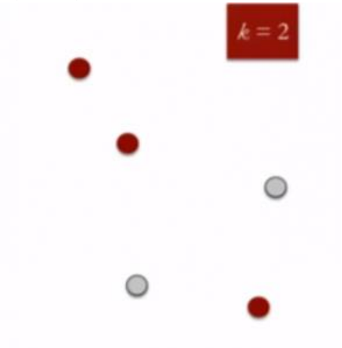
Comparison between K Means and Hierarchical Clustering:

- In K Means clustering we use the Elbow method using WCSS **to find the optimal number of clusters .In Hierarchical Clustering** we use Dendrogram **to find the optimal number of clusters.**
- Directional approach to form clusters in K Means is the only and only Centroid. **But in Hierarchical Clustering** Directional approach to form clusters is Dendograms.
- In K Means clustering 'sklearn – Kmeans ' python library is used and **in Hierarchical Clustering '**sklearn-AgglomerativeClustering' python library is used.
- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the Dendogram.
- Hierarchical Clustering is useful for small dataset, and K-Means Clustering is useful .for large dataset.
- K-Means Clustering, we need to iterate the model to find out the optimal number of Clusters, but in Hierarchical Clustering, it automatically gives result at various number of Clusters.

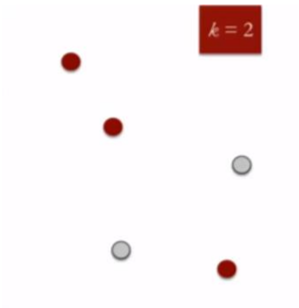
b) Briefly explain the steps of the K-means clustering algorithm.
Answer:

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps:

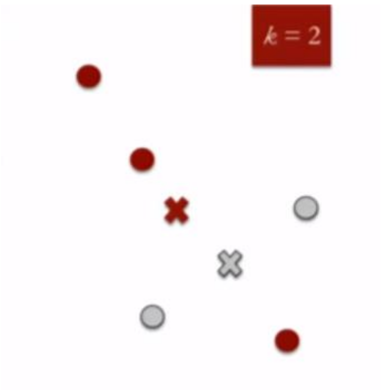
1. **Assignment Step:**
Specify the desired number of clusters K: Let us choose k=2 for these 5 data points in 2-D space.



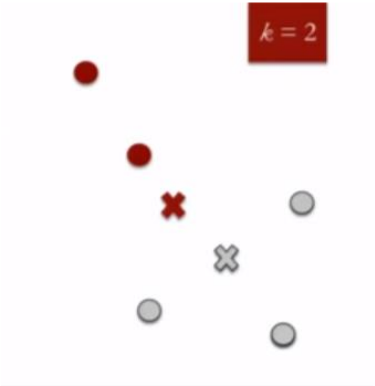
2. **Optimisation Step:**
Randomly assign each data point to a cluster: Let's assign three points in cluster 1 shown using red colour and two points in cluster 2 shown using grey colour.



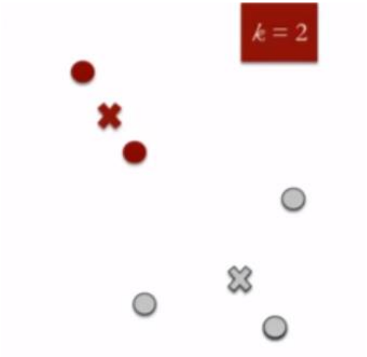
3. Compute cluster centroid: The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



4. Re-assign each point to the closest cluster centroid: Note that only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster



5. Re-compute cluster centroids: Now, re-computing the centroids for both the clusters.

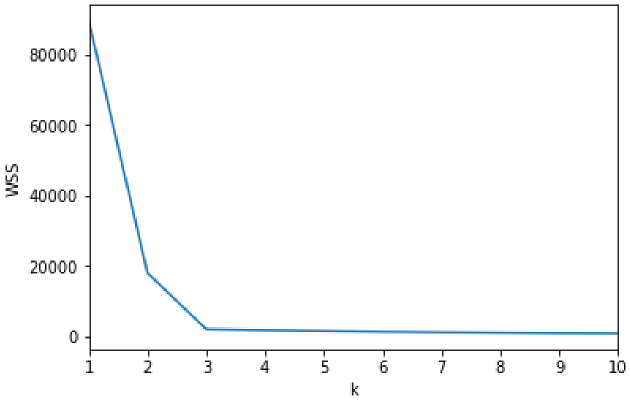


6. Repeat steps 4 and 5 until no improvements are possible: Similarly, we'll repeat the 4th and 5th steps until we'll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
Answer:

Value of k is chosen by following methods in K-means clustering.

1. **The Elbow Method**
   This is probably the most well-known method for determining the optimal number of clusters. It is also a bit naive in its approach.
   Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.



Above plot looks like an arm with a clear elbow at k = 3.

Within-Cluster-Sum of Squared Errors sounds a bit complex. Let's break it down:

- The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster centre.
- The WSS score is the sum of these Squared Errors for all the points.
- Any distance metric like the Euclidean Distance can be used.

Unfortunately, we do not always have such clearly clustered data. This means that the elbow may not be clear and sharp.

2. The Silhouette Method

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

The Silhouette Value s(i) for each data point i is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

s(i) is defined to be equal to zero if i is the only point in the cluster. This is to prevent the number of clusters from increasing significantly with many single-point clusters.

Here, a(i) is the measure of similarity of the point i to its own cluster. It is measured as the average distance of i from other points in the cluster.

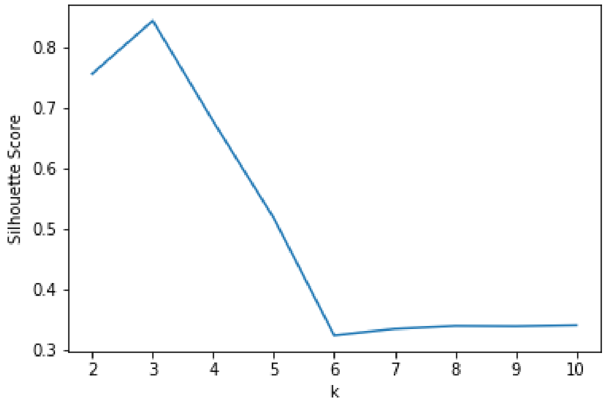
Similarly, b(i) is the measure of dissimilarity of i from points in other clusters.

d(i, j) is the distance between points i and j. Generally, Euclidean Distance is used as the distance metric.

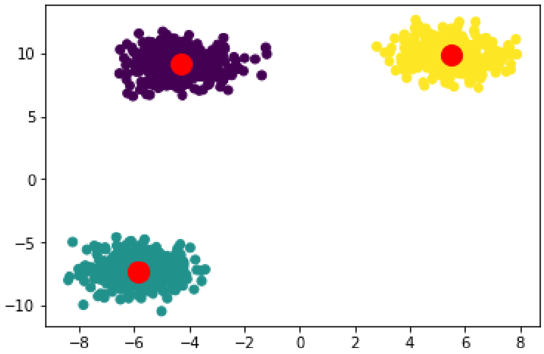
The Silhouette score can be easily calculated in Python using the metrics module of the sklearn library.

As mentioned before that a high Silhouette Score is desirable. The Silhouette Score reaches its global maximum at the optimal k. This should ideally appear as a peak in the Silhouette Value-versus-k plot.

Here is the plot for our own dataset:



There is a clear peak at k = 3. Hence, it is optimal.
Finally, the data can be optimally clustered into 3 clusters as shown below.



d)Explain the necessity for scaling/standardisation before performing Clustering.

Answer:
Standardizing means to rescale your data to have a mean of zero and a standard deviation of one. A standardized variable is sometimes called a z-score or a standard score.
Normalizing is another rescaling method with many meanings in statistics and statistical applications. Most commonly, normalizing rescales numeric data between zero and 1

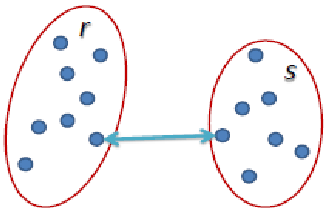
Standardizing your data prior to cluster analysis is extremely critical. Clustering is an unsupervised learning technique that classifies observations into similar groups or clusters. A commonly used measure of similarity is Euclidean distance. The Euclidean distance is calculated by taking the square root of the sum of the squared differences between observations. This distance can be greatly affected by differences in scale among the variables. Generally, variables with large variances have a larger effect on this measure than variables with small variances. In other words, if one of the variables is measured on a much larger scale than the other variables, then whatever measure we use will be overly influenced by other variable. Also, in the case of variables that contain outliers (observations that are much bigger or smaller than the vast majority of the data), this sort of standardization may be too severe, scaling down the outlying observations so that they appear to be closer to the others. For this reason, standardizing multi-scaled variables is advised prior to performing clustering.

e) Explain the different linkages used in Hierarchical Clustering.
Answer:

Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. The following three methods differ in how the distance between each cluster is measured.

Single Linkage

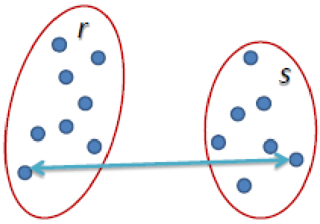
In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length o the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete Linkage

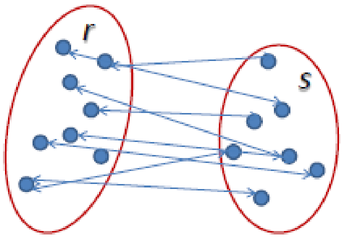
In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average Linkage

:In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between eac point in one cluster to every point in the other cluster. For example, the distance between clusters “r” and “s” to the left is equal to the average length each arrow between connecting the points of one cluster to the other. This is also called the UPGMA algorithm.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Other linkages are:

- **Simple average** linkage:

Method of **equilibrious between-group average** linkage (WPGMA) is the modified previous. Proximity between two clusters is the arithmetic mean of all the proximities between the objects of one, on one side, and the objects of the other, on the other side; while the sub clusters of which each of these two clusters were merged recently have equalized influence on that proximity – even if the sub clusters differed in the number of objects.

$d(u,v)=(dist(s,v)+dist(t,v))/2$

where cluster u was formed with cluster s and t and v is a remaining cluster in the forest.

- **Centroid** linkage:
  Proximity between two clusters is the proximity between their geometric centroids: [squared] Euclidean distance between those. When two clusters s and t are combined into a new cluster u, the new centroid is computed over all the original objects in clusters s and t. The distance then becomes the Euclidean distance between the centroid of u and the centroid of a remaining cluster v in the forest. This is also known as the **UPGMC algorithm.**

$$dist\ (s, t)=||cs-ct||2$$

where cs and ct are the centroids of clusters s and t, respectively.

- **Median**, or **equilibrious centroid** linkage:
  . When two clusters s and t are combined into a new cluster u, the average of centroids s and t give the new centroid u. This is also known as the WPGMC algorithm.

- **Ward's** linkage:
  Ward variance minimization algorithm. This is also known as minimal **increase of sum-of-squares** (MISSQ). This is also known as the incremental algorithm.
  The new entry d(u,v) is computed as follows,

$$d(u,v)=|v|+|s|Td(v,s)2+|v|+|t|Td(v,t)2-|v|Td(s,t)2$$

where u is the newly joined cluster consisting of clusters s and t, v is an unused cluster in the forest, T=|v|+|s|+|t|, and |∗| is the cardinality of its argument