**QUESTION-1: ASSIGNMENT SUMMARY**

PROBLEM STATEMENT:

HELP International is an international humanitarian NGO that is committed to provide financial aid to backward countries. NGO raised funds and now it has to decide on which country to aid. Given data about different countries and their social, economic, health factors, as an analyst, we need to provide countries which are categorised as backward/under developed countries so that NGO can provide financial aid to those countries.

SOLUTION METHODOLOGY:

* Converted columns which have data in percentages to absolute values to capture more information and to compare with other columns.
* Did outlier analysis and decided that outliers should not be removed, as removing them might affect ranking of countries for receiving financial aid.
* Exploratory data analysis:

1. **Univariate analysis**: Plotted distribution plot for each column to get a clear understanding of how each column is distributed
2. **Bivariate analysis:** Plotted pair plot to check how each column is related to other column whether it is positive or negative relationship.
3. **Correlation map:** Plotted correlation map, to quantify relationship between columns.

* Performed scaling and determined Hopkin’s statistic to understand the clustering tendency of the given data and Hopkin’s statistic is greater than 0.9 which tells us that there is a good clustering tendency
* Did Hierarchical clustering using Single linkage but due to it’s unclear

view, switched to complete linkage and after watching dendrogram of complete linkage, decided to go with 3 and 4 clusters.

* After doing hierarchical clustering, the cluster which is categorised as under developed has almost 90 percent of countries. So hierarchical clustering is not giving good solution as 90 percent of data is in one single cluster
* So Did k-means clustering and based on balanced cluster size, silhouette analysis, elbow curve method arrived at a conclusion that optimal number of clusters is 3.
* We arrive to final list of countries by further filtration of the cluster which is categorised as under developed. First filtering is done based on gdpp, next is on income and finally on child mortality. The reason for this is as follows:

1) The reason for first filtering based on gdpp is, gdpp is the primary measure of how country is doing on a whole.

2) The second priority is income as it tells us financial status of people

3) Finally our third priority is child mortality as this child mortality rate depends on income and gdp. if a country has good income, goods, services then it is likely to have less child mortality rate. Therefore, if we first filter countries based on gdpp, income we indirectly are addressing child mortality rate problem too. Therefore, child

mortality rate will be our third priority.

**QUESTION-2: CLUSTERING**

**a) Compare and contrast k-means clustering and hierarchical clustering.**

|  |  |
| --- | --- |
| **K-Means Clustering** | **Hierarchical Clustering** |
| K-means clustering can handle big data due to its time complexity being O(n). | Hierarchical clustering cannot due to its high time complexity being O(n^2) |
| Since we start with random points as cluster centres the results we get after running K-means multiple times differ. | Results are reproducible as there is no random points choosing at the start |
| K-means returns K clusters | Hierarchical clustering returns a tree which is also called dendrogram. |
| K means require advance knowledge about k | In, Hierarchical clustering, we can stop at whatever number of clusters we find appropriate based on dendrogram |

**b) Briefly explain the steps of k means clustering algorithm.**

Step1: Choose k which is the number of clusters

Step2: Select k random points to be cluster centroids

Step3: Assign all the points to closest centroid based on Euclidean distance.

Step4: Now recompute the centroid of newly formed clusters

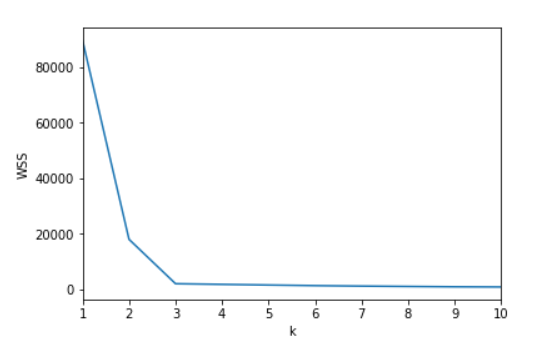
Step5: Repeat step3, step4 until it satisfies one of the following criteria

* Maximum number of iterations is reached
* Cluster centroids do not change and points in the

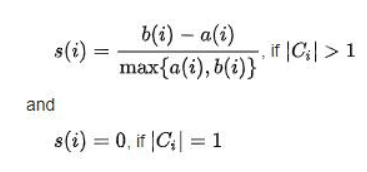
**c) How is the value of k chosen in k-means clustering? Explain both statistical as well as business aspect of it**.

Statistical aspect:

1) Elbow method: For different values of k, calculate within-cluster-sum of squared errors. WSS is calculated as sum of squared distances from samples to their nearest cluster centre. Now after this calculation find k at which WSS drops by a big factor. In the plot of WSS-Vs-k, this is visible as an elbow. The major drop is at k=2 therefore optimal k is 2

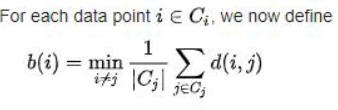
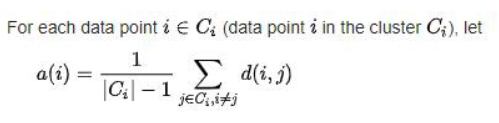
****

**2) Silhouette analysis:** The silhouette value measures how similar a point is to its own cluster compared to other clusters. Silhouette analysis works on the principle that inter cluster distance should be more and intra cluster distance should be less. Silhouette value can take any value 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) 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 intra cluster distance and it is a 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. b(i)is inter cluster distance and it is a measure of dissimilarity of i from points in other clusters. Formulae of both a(i) and b(i) are given below.

****

**Business aspect:**

**Statistical techniques are not completely enough to select optimal clusters. We have to consider business objective of the problem. We have to select k in such a way that after clustering is done, the formed clusters should be interpretable for making decisions in future. We should be able to understand what each cluster represents in a business sense. For example, if a company wants to provide education aid to cities of a state, given data of cities, now going with 3 clusters which represent educated cities, un-educated cities, highly-educated cities make more sense in a business aspect, so that company can focus differently on different cities.**

**d) Explain the necessity of scaling before performing clustering?**

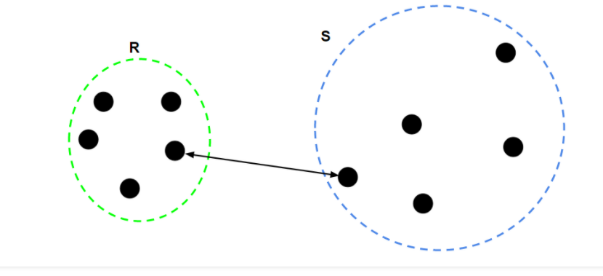
While doing any type of clustering we depend on distance metric between two data points for clustering process. If we do not scale, since this distance can be bigger for bigger scaled data and smaller for smaller scaled data, the algorithm will be biased and will not give correct results. Also, if we do not scale, it takes much higher time for algorithm to execute than when compared to algorithm’s execution time for a scaled data. To prevent delayed execution and biased nature , we scale before performing any type of clustering.

**e) Explain the different linkages used in Hierarchical clustering.**

Linkage means the distance between two clusters. Depending upon the different approaches to find distance between two clusters, different linkages are defined accordingly.

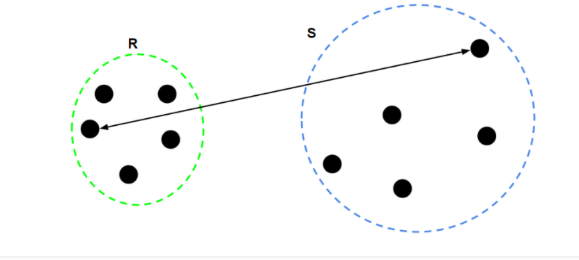
Single Linkage: For two clusters R and S, single linkage returns the minimum distance between two points p and q, such that p belongs to R and q belongs to S.

L (R, S) = min (D (p, q)), p belongs to R and q belongs to S.



Complete Linkage**:** For two clusters R and S, complete linkage returns the maximum distance between two points p and q, such that p belongs to R and q belongs to S.

L (R, S) = max (D (p, q)), p belongs to R and q belongs to S.



Average Linkage**:** For two clusters R and S, first we have to compute distance between points i and j such that i is from cluster R and j is from cluster S. We repeat this process for all pairs of i and j. Then after calculating all distances we average them to get the mean of all distances. Average linkage returns this mean of all distances.

