Clustering of Countries Assignment

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)

Answer 1:

Overview/Background:

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, the NGO have been able to raise an amount of $ 10 million.

Now, the CEO of the NGO needs to decide how to use this money strategically and effectively to help the countries which are in dire need of aid.

Problem Statement:

As a data analyst, our job is to categorise the countries using socio-economic and health factors that determine the overall development of the country. We need give top 5 countries which are in dire need of funds.

Solution Methodology:

We performed the following steps to arrive our solution:

1. **Data Collection and Cleaning**

    - Import the data – Data is given as a csv file.

- Identify and eradicate the data quality issues.
    - Some columns in the dataset, viz. health, imports and income were converted from %-age value to absolute value.
    - Null value treatment was not required.

2. Outlier Analysis and Removal

- Deletion of rows/columns is not appropriate since the dataset size is very small.
- So, capping of outliers is done wherever required.

3. EDA Analysis on Data

- EDA is performed on the data to get the insights of data and analyse patterns in the data.
- Following EDA is performed –
    - Distribution of all feature variables by plotting distplots to get insight on their behaviour.
    - Checking outliers in all the feature variables – This is very helpful to identify whether capping is required at lower end or upper end for each variable.
    - Scatter plots for -
        - Income v/s Life expectancy
        - Income v/s Health
        - Income v/s Child Mortality
        - GDPP v/s Child Mortality
        - GDPP v/s Life expectancy
        - GDPP v/s Health
    - Top 10 countries analysis –
        - Highest Child Mortality
        - Lowest Health index
        - Highest Inflation rate
        - Lowest Life Expectancy
        - Lowest GDPP
        - Lowest Income

4. Scaling the data

- Data is scaled by using Standard Scalar.
- Scaling is required so that the clustering algorithms works correctly.
- K-means does not performs well when the data is unscaled.

5. Hopkins Statistics

- Hopkins statistics is performed to check if the data is appropriate for clustering.
- The data has Hopkins statistics value of greater than 85%, which states that the data is good for clustering.

6. K-Means Clustering

- The value of "K" is identified as **3** by performing Silhouette analysis & plotting Elbow curve.
- The clustering is performed with K=3 and plotting is done to visualize the clusters.
- Profiling is done to get inference about the clusters made after clustering.
- Countries with dire need of aid identified by K-Means clusters are - **Burundi, Liberia, Congo, Dem. Rep., Niger, Sierra Leone**

7. Hierarchical Clustering

- Both single linkage and complete linkage are performed. Complete linkage gives more interpretable results.
- We choose N=3, to cut the dendrogram to form the clusters.
- Plotting is done to visualize the clusters.
- Profile is done to get the inferences about the clusters made after hierarchical clustering.
- Countries with dire need of aid identified by complete linkage Hierarchical clustering are - **Burundi, Liberia, Congo, Dem. Rep., Niger, Sierra Leone**

8. Conclusion

- The list of countries for dire need of aid identified from both K-Means clustering and complete linkage Hierarchical clustering are same.
- Final list of top 5 countries in dire need of aid are –
    - Burundi
    - Liberia
    - Congo
    - Dem. Rep., Niger
    - Sierra Leone

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.

Answer - (a):

Comparison and contrast between K-Means and Hierarchical clustering

K means is an iterative clustering algorithm that aims to find local maxima in each iteration.

Hierarchical clustering starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.

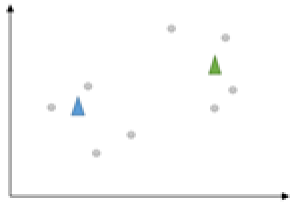
k-means Clustering	Hierarchical Clustering
It requires advance knowledge of K i.e. no. of clusters one want to divide your data.	We can stop at any number of clusters by interpreting the dendrogram.
Median or Mean as a cluster centre is used to represent each cluster.	It uses agglomerative method. It begin with 'n' clusters and sequentially combine similar clusters till only one cluster is obtained.
It is less computationally intensive and is suited for very large datasets.	It is more computationally expensive. It is not suitable for very large datasets.
The results produced by the algorithm may differ on consecutive runs, since we start with a random choice of clusters.	The results are reproducible in Hierarchical clustering

Answer - (b):

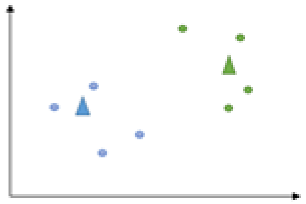
Briefly explain the steps of the K-means clustering algorithm.

The k-means algorithm is an iterative algorithm in which has 3 main steps as follows –

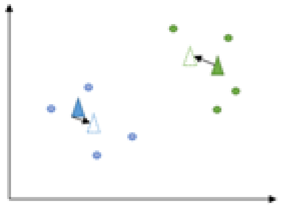
1. Choosing “k” points also referred to as centroids in the data (randomly located).
2. For each centroid the algorithm finds the nearest points, in terms of distance that is usually computed as Euclidean distance, to that centroid, and assigns them to its category.
3. For each category represented by a centroid, the algorithm computes the average of all the points in that cluster and that average point is the new cluster.
4. Repeat step 2 and 3 until there is no change in the centroid point.



1. Two centroids are randomly located among data



2. The algorithm finds the nearest observations for each centroid, and then attributes them to each centroid’s class



3. New centroids, resulting from the average of each class observations, are computed, and then the process is repeated

Answer - (c):

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

Statistical Aspect:

There are 2 methods used to find the optimal value of K.

1. Elbow Method
2. Silhouette Method

We use the **elbow method** to determine the optimal value of K.

The basic idea behind this method is that it plots the various values of cost with changing values of *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease.

The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the **elbow point**.

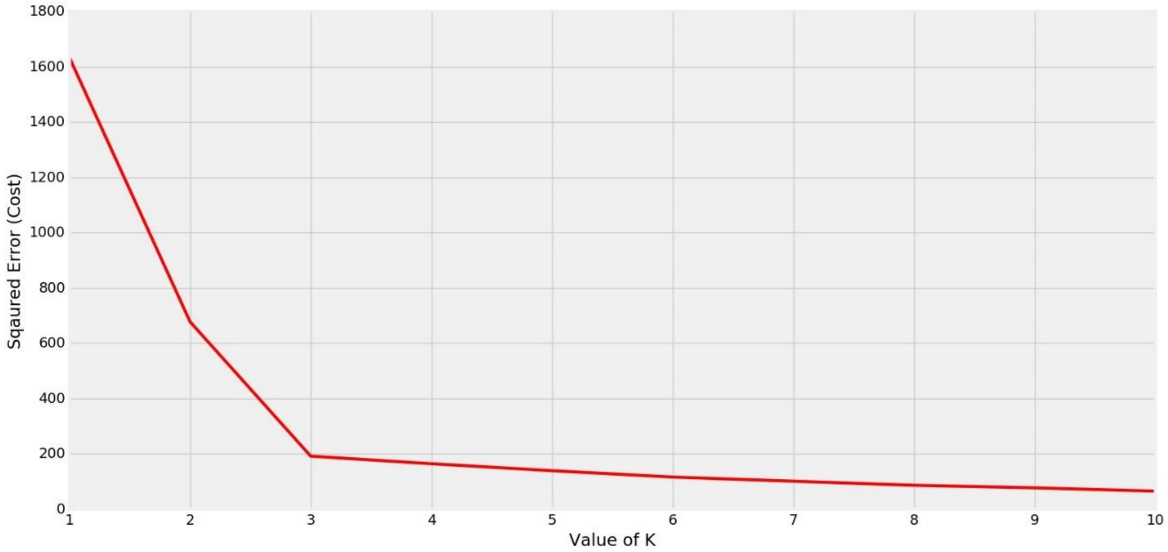
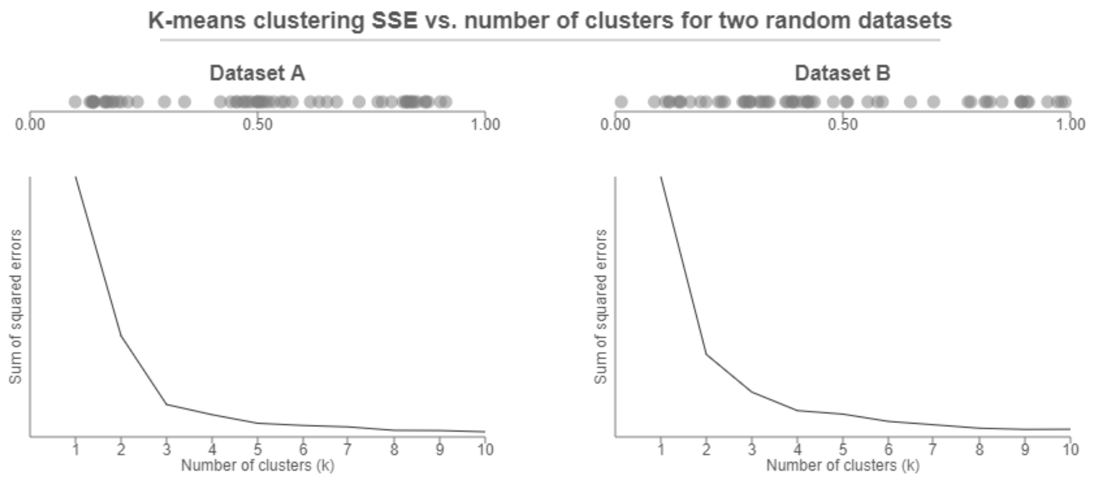


Figure 1 Elbow curve

- From above elbow curve, the optimal value of K is 3.

In cases where the elbow is not clear (right diagram), the choice of k becomes ambiguous, we use the Silhouette method.

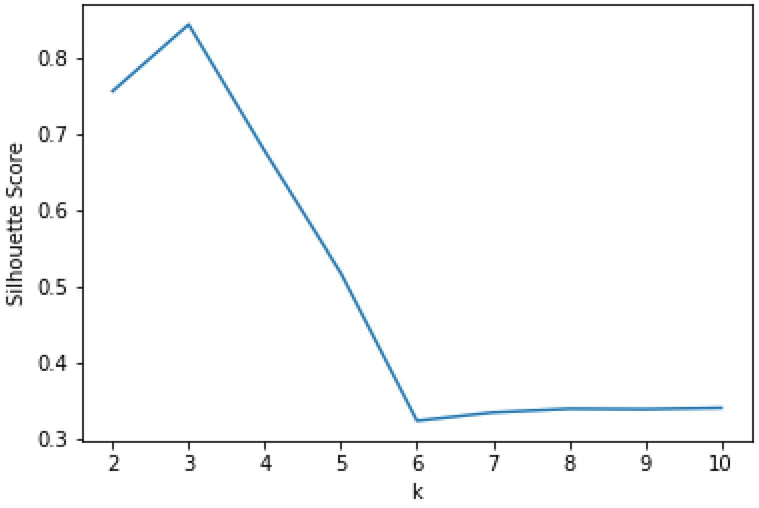


The **Silhouette value** measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation)

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 Score reaches its *global maximum at the optimal k*. This should ideally appear as a peak in the Silhouette Value-versus-k plot.



- From above silhouette curve, the optimal value of K is 3.

Business Aspect:

Business aspect of choosing the values of “K” is equally important.

Say, in a dataset, if a business already knows the number of categories available in the data, we can simply choose the value of K, same as business.

The implications comes when we chose different number of “K” than the business need.

Even though our model is able to classify the data into “K” categories, but if it is not interpretable by the business, we can’t work on that data further.

This situation/scenario restricts the clustering problem to a technical problem only and not the complete business case solution.

Hence, business aspect is equally important in choosing the value of “K”.

Answer - (d):

Explain the necessity for scaling/standardization before performing Clustering.

Clustering techniques like K-Means use **Euclidean Distance** to form the clusters.

That is why it is important to scale the variables before calculating the distances.

All such distance-based algorithms are affected by the scale of the variables.

For e.g.:

Say, you are working with data where each variable means something different. Like, age, weight, year, etc. Then, these fields are not directly comparable.

1 year is not equivalent to 1 kg and may or may not have the same level of importance in sorting a group of records.

In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters.

Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

Answer - (e):

Explain the different linkages used in Hierarchical Clustering.

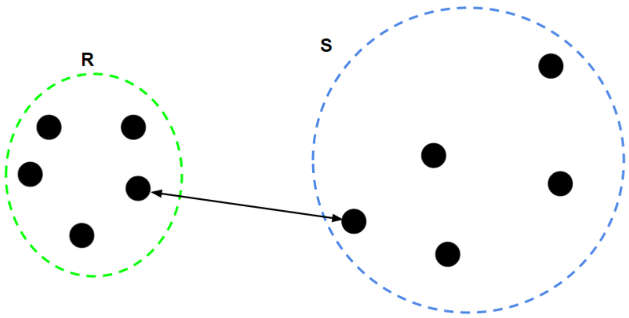
The Hierarchical Clustering involves either clustering sub-clusters into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner.

During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed.

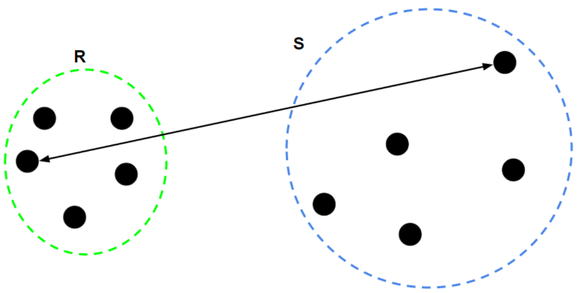
The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points.

The different types of linkages are –

1. **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 two clusters “R” and “S”, the single linkage returns the minimum distance between two points “I” and “j” such that “I” belongs to “R” and “j” belongs to “S”.



2. **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 two clusters “R” and “S”, the single linkage returns the maximum distance between two points “I” and “j” such that “I” belongs to “R” and “j” belongs to “S”.



3. **Average Linkage:**

   In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.