Clustering Assignment-based Subjective Questions

Que 1: Assignment Summary.

Answer: Our objective is to find the countries that are in dire need of aid.
In order to get the list of countries that need aid we need to perform following steps:

1. Understanding Data:
    a. Import necessary libraries
    b. Load the dataset
    c. Check the shape of the dataset: Here we have 167 rows and 10 columns
    d. Check for data types: All columns are numeric except country column
    e. Check for null values: No null values found
    f. Check for duplicate values: No duplicate values found
    g. Check for unique values present in each column
    h. Check statistical summary of Dataset: It indicates presence of outliers
2. Data Transformation:
    a. Converting health, imports and exports columns (which were in %) to their actual values to make sense
3. Data Visualisation:
    a. Creating distribution plots for all columns: Decided to use all columns for clustering but we will use only 'gdpp' 'child_mort' and 'income' columns for profiling
    b. Creating pairplot of numerical variables: Some variables display collinearity
    c. Creating heat map to understand the correlation: We can observe high correlation between the following pairs - child_mort and total_fer, imports and exports, gdpp and health, income and gdpp
    d. Univariate Analysis: Checking for top/bottom 10 countries from each column as per column definition
4. Outlier analysis and treatment:
    We can see that all the columns are having Outliers. We cannot exclude these outliers as we have only 167 samples with us and excluding outliers can led to loss of Data. Instead, we can perform soft capping for some columns.
    a. For columns such as child_mort, inflation, total_fer we should not do anything to the upper range outliers but we may deal with the lower range outliers through capping.
    b. But for the rest of the columns, we should not do anything for the lower range outliers but we may deal with the upper range outliers through capping.
    c. Checking results after capping: Capping done with IQR method
5. Hopkins Test: Checking clustering tendency of Dataset
    a. After running the Hopkins test multiple times we get the results greater than 0.80.
    b. Since the value is > 0.8 the given dataset has a good tendency to form clusters.
6. Scaling Dataset for better interpretability
7. Find best value for K (number of clusters to form)
    a. Silhouette score: We look at the maximum value in case of Silhouette Score. From the graph we can infer that the ideal number of cluster values can be 3 or 4.

    b. SSD (Elbow curve): We can observe that the elbow is formed between 3 to 5 and hence value between 3-5 will be suitable for cluster formation.

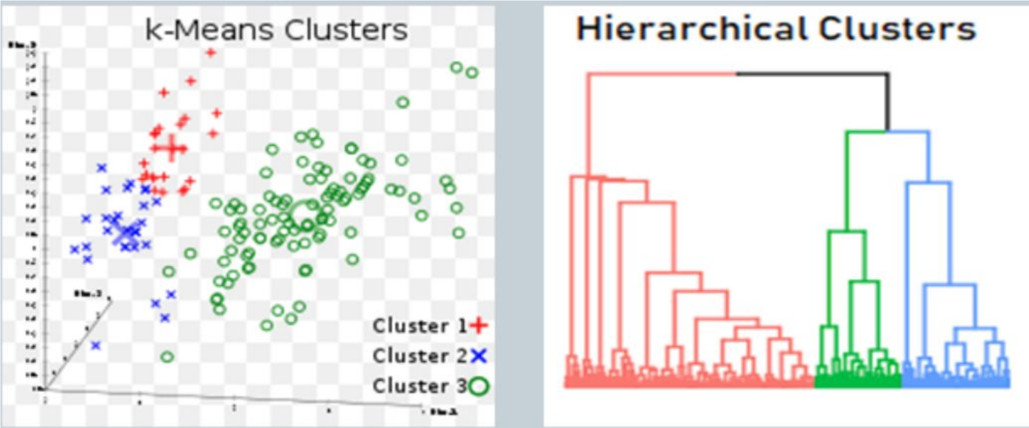
8. Proceed for K means clustering:
    a. Try with k = 4: We get 4 clusters but there aren't enough data points in each cluster
    b. Try with k = 3: We get 3 clusters and there are enough data points in each cluster. So I decided to proceed with the value of K as 3.
9. Plotting the clusters obtained from K means method with K = 3:
    a. We can observe that cluster Id 0 has low GDPP, low Income and high Child mortality rate and hence cluster 0 contains data points of our interest
    b. Countries which belong to the needy cluster (Cluster 0) are:
        i. 'Burundi',
        ii. 'Liberia',
        iii. 'Congo, Dem. Rep.',
        iv. 'Niger',
        v. 'Sierra Leone',
        vi. 'Madagascar',
        vii. 'Mozambique',
        viii. 'Central African Republic',
        ix. 'Malawi',
        x. 'Eritrea'
10. Perform Hierarchical clustering on the scaled dataset with single linkage and complete linkage method.
    a. Results of complete linkage are much better than single linkage method
    b. I decided to use k as 4 and obtained 4 clusters
11. Plotting the clusters obtained from Hierarchical clustering method with K = 4:
    a. We can observe that cluster Id 0 has low GDPP, low Income and second highest Child mortality rate and hence cluster 0 contains data points of our interest
    b. Countries which belong to the needy cluster (Cluster 0) are:
        i. 'Burundi',
        ii. 'Liberia',
        iii. 'Congo, Dem. Rep.',
        iv. 'Niger',
        v. 'Sierra Leone',
        vi. 'Madagascar',
        vii. 'Mozambique',
        viii. 'Central African Republic',
        ix. 'Malawi',
        x. 'Eritrea'
    c. We can observe that there is only 1 data point in cluster 3 which belongs to country 'Nigeria'.
    d. Nigeria has highest mortality rate which might be due to poor state of pregnant women in Nigeria which is affecting maternal health as well as child mortality

12. Conclusion:
    a. We performed analysis on the Dataset using K means clustering and Hierarchical clustering
    b. Both the approaches are giving the same results and hence we can say that above mentioned countries are in urgent need of financial aid.
    c. Nigeria has the highest mortality rate and hence this country should also be provided with good medical care for pregnant women and children.

Que 2. a): Compare and contrast K means clustering and Hierarchical clustering.

Answer:

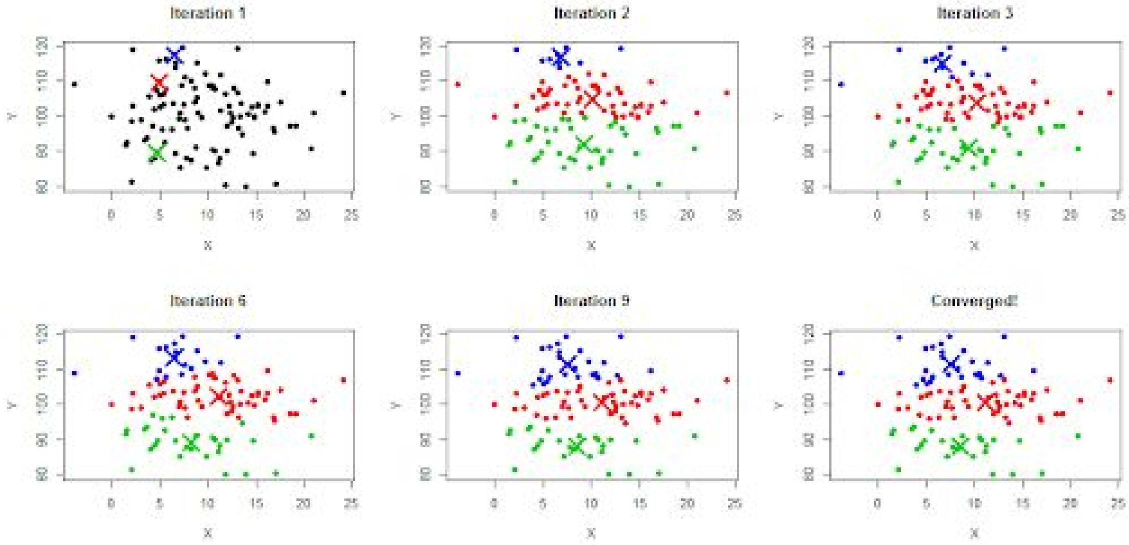
No	**K means clustering**	**Hierarchical clustering**
1.	Number of clusters to be formed (K) has to be given explicitly to the algorithm	Run the hierarchical clustering algorithm, obtain dendrogram and then cut dendrogram at different heights to obtain desired number of clusters
2.	It is good for processing large datasets	Hierarchical clustering generally produces better clusters, but is more computationally intensive and hence performance slows down when used for large datasets
3.	Outliers are not evaluated properly	Outliers are properly explained
4.	K means is used on numeric data only	Hierarchical method can be used for variety of data
5.	Time complexity of K Means is linear: O(n)	Time complexity of Hierarchical clustering is quadratic: O(n2)
6.	Results might differ if we run the algorithm multiple times with different choice of K	Results are reproducible in hierarchical clustering
7.	Similar data points belong to same cluster and dissimilar data points belong to different clusters	Tree like structure is formed in which similar clusters combine first and it continues till we get single cluster



Que 2. b): Briefly explain the steps of K means clustering algorithm.

Answer: Steps to follow for performing K means clustering are:

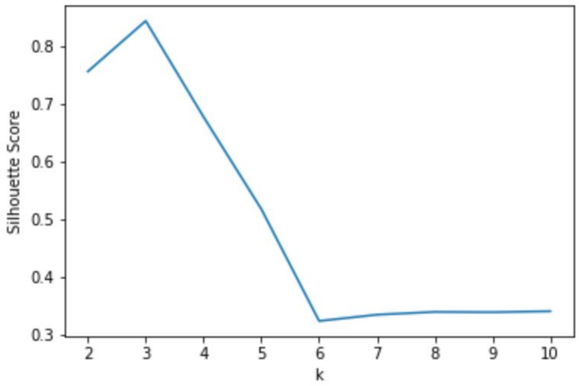
1. Choose the number of clusters K: You can make choice for K by performing silhouette score analysis or SSD (elbow curve) analysis
2. Select K random points from dataset as centroids
3. Assign all the points to their closest cluster centroids one by one
4. Recompute the centroids of the newly formed clusters
5. Repeat the 3rd and 4th step till you get the centroid value which is not changing anymore (known as convergence)
6. You can also specify how many iterations to perform explicitly



Que 2. c): How is the value of ‘k’ chosen in K means clustering? Explain both the statistical and business aspect of it.

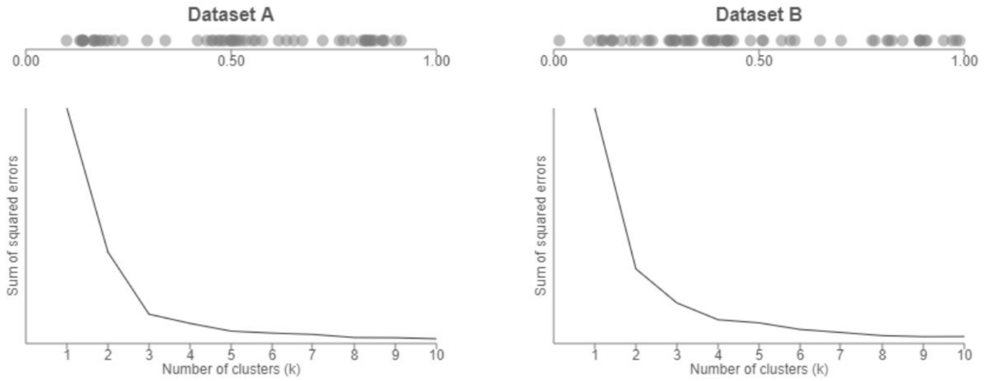
Answer: There are 2 ways to choose ideal value of K:

1. Statistical method:
    a. Silhouette score: This method will measure how similar a point is to its own cluster in comparison to other clusters. It lies between -1 to +1. Higher silhouette score is desirable.



    Eg: Here Peak is at 3 which means it is ideal to choose K as 3

    b. SSD (Elbow curve): It uses a sum of squared distances approach with Euclidean Distance or the Manhattan Distance as distance metric. It is a naive method and will not always give a clear and sharp elbow in graphs if the data is not clustered properly.



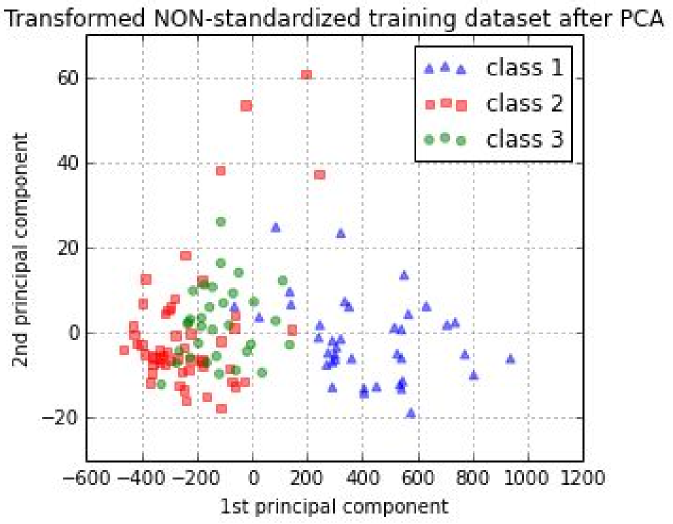
    Eg: In the above figure, Dataset A has clear elbow at 3 and we can choose K = 3
    Whereas in Dataset B choice of K is ambiguous. It can be 3 or 4.

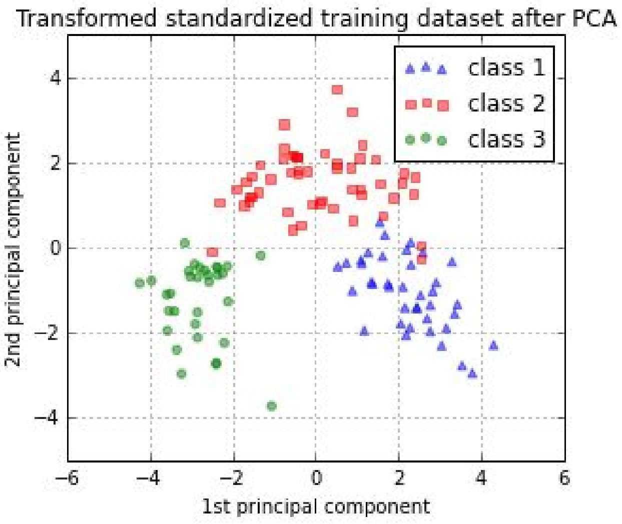
2. Business knowledge: We can use business/ domain knowledge to predict the number of clusters. For example if we have variables like books, canvas, painting brush, iphone, speakers, laptops in our dataset then with our business understanding we can create 2 clusters namely stationary and electronics.

Que 2. d): Explain the necessity of scaling/standardisation before performing clustering.

Answer: Importance of Scaling:

1. In our dataset we have variables which have units at different scales.
   In clustering, groups are defined on the basis of distance between points in mathematical space.
2. If we work with data where each variable means something different (for example if some variables have unit kilograms and some others have meters), it is not possible to compare the variables directly.
3. All the units have different weightage in mathematics. Like 1 pound is not equal to 1 kilogram and so on.
4. If we use a dataset without scaling, the variable with higher range of values will be the driving factor for defining clusters and variables with lower range of values will get lower weightage.
5. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.
6. Scaling increases the overall performance of the model.



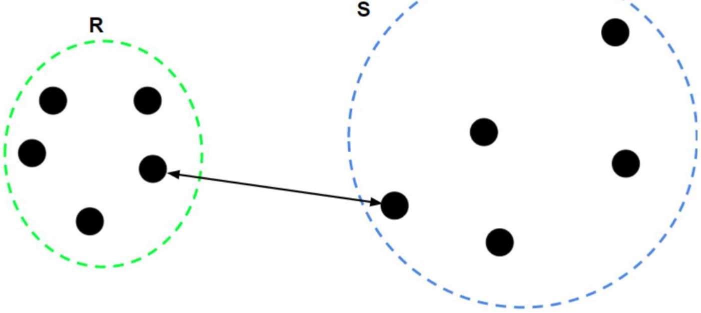


Que 2. e): Explain the different linkages used in Hierarchical clustering.

Answer: Hierarchical clustering follows bottom-up approach (combining sub clusters till we get one single large cluster) or top-down approach (diving large cluster to form smaller sub clusters).
During both the approaches we have to compute the distance between the two sub clusters. In order to compute the distance between two clusters, we use different linkage methods such as single linkage, average linkage, complete linkage, ward linkage and centroids.

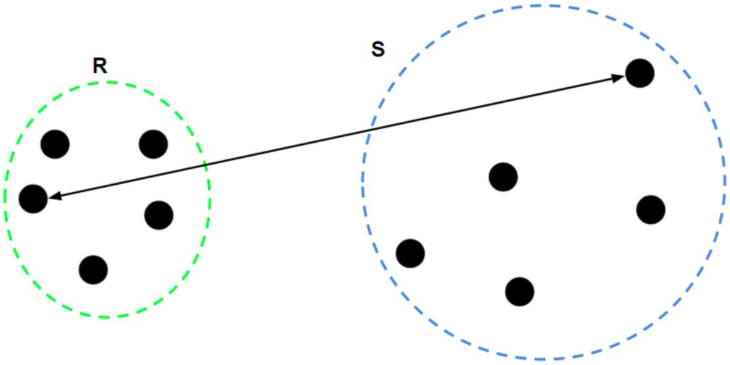
1. Single linkage: Returns minimum distance between two points from each cluster.

$$L(R,S) = min(D(i,j)), i\epsilon R, j\epsilon S$$



2. Complete linkage: Returns maximum distance between two points from each cluster.

$$L(R,S) = max(D(i,j)), i\epsilon R, j\epsilon S$$



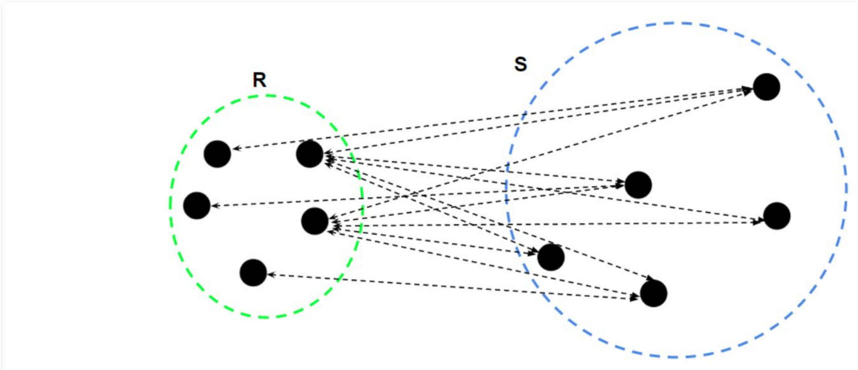
3. Average linkage: Returns average distance between every point of one cluster to every point of another cluster.

$$L(R, S) = \frac{1}{n_R+n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i, j), i\epsilon R, j\epsilon S$$

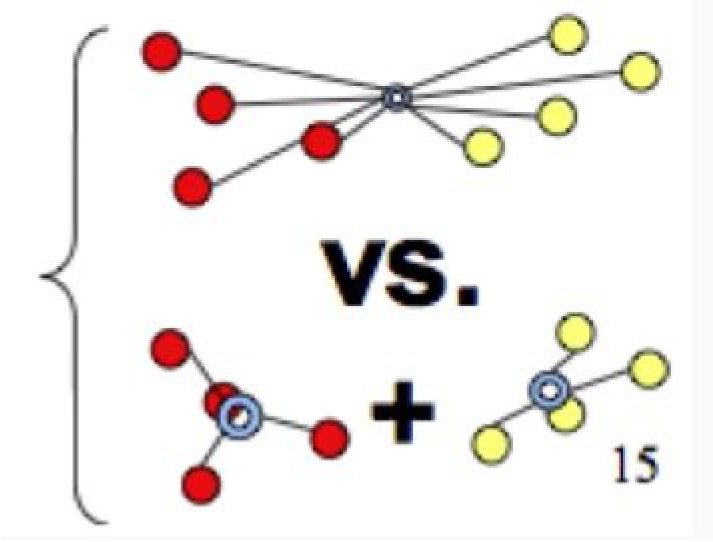
where

n_R - Number of data-points in R

n_S - Number of data-points in S



4. Ward linkage: calculates the distance between clusters by sum of squared differences with all clusters.



5. Centroids: Returns the distance between centroids of two clusters.