Clustering Assignment

Assignment: Part II

Question 1: Assignment Summary

Problem Statement:

- As part of distribution of funds raised, by HELP international, deciding on the countries most in need of aid.

Solution Methodology:

- K-Means method and Hierarchical Clustering method were employed to perform clustering.

Steps performed for analysis:

- **Data Analysis**:
  - No missing data was observed.
  - Measures derived as percentages of GDPP were converted to absolute values.

- **Data Correlation:**
  - Pair plots and heat map determined patterns in data.

- **Outlier Analysis and treatment:**
  - Of variables with outliers, high child mortality, high inflation, lower life expectancy and high total fertility typically impact underdeveloped countries, so no outliers were removed here.
  - Higher outliers for exports, imports, health investment, income and GDPP were capped to 99% values.

- **Rescaling**:
  - Data was scaled using StandardScaler

- **Hopkins Test**:
  - With a score of 0.92, data was determined to have a good cluster tendency.

Model Building:

- K-Means algorithm and Hierarchical clustering were used to determine clusters.
    1. **K-Means algorithm**:
        - Using both, Sum of Square Distances (SSD) or Elbow Curve Method and Silhouette Score the cluster value of 3 was determined as most applicable.
        - Cluster members with k=3 showed a good distribution with some difference between clusters with most and least data.
    2. **Hierarchical clustering (Agglomerative technique)**:
        - Single Linkage and Complete Linkage were used to determine clusters. While complete linkage method was better at cluster distribution as compared to Single Linkage method, data was highly skewed towards the cluster with most data.
- Accordingly, K-Means clustering with 3 clusters was determined as final modelling technique.

Model Evaluation

- Box plots determined a reliable distinction between clusters.
- Scatter plots revealed linear correlation between income and GDPP while income and GDPP were lower for countries with higher child mortality.

Since **underdeveloped nations are characterized by higher child mortality and lower income and GDPP** relevant cluster was selected and list of countries in need determined.

Question 2: Clustering

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

Ans:

	K-Means Clustering	Hierarchical clustering
Number of clusters to be selected initially	Yes	No
Centroids to be defined by user (manual or using algorithm)	Yes	No
Impact of re-running algorithm	Clusters can change with change in initial centroid selection	Results are reproducible in Hierarchical clustering since clusters are formed by calculating internal distance between points

		so long as same method of distance calculation is selected.
Complexity	Time complexity of K Means is linear i.e. O(n)	Time complexity of hierarchical clustering is quadratic i.e. $O(n^2)$.
Impact of outliers	Outliers can cause centroids to move closer to them since this method uses Euclidean distance between centroid and data points.	Outliers can cause less-than-optimal merging. Additionally, since merging can't be reversed, this can cause problems for data with outliers.
Cluster selection	Once initial number of clusters is selected, the algorithm needs to be re-run each time we need to change cluster number which can also impact clusters formed.	Hierarchical clustering forms a dendrogram structure which can be cut at any point to decide on number of clusters to be used.

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

Ans: K-Means clustering requires initial selection of K i.e. number of clusters.

Once number of clusters is selected, the following steps are followed:

Step 1: Choose k initial centroids.
Step2 : Assignment: Find the euclidean distance of each point from the each of the centroids. Depending on which centroid is closer/least distance, assign each point to either of the centroids.
Step 3: Optimization: Find the mean of all points assigned to each centroid and move centroid to that location.
Repeat steps 2 and 3 until the centroids no longer update i.e. points assigned remain same post iteration.

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

Ans: K-Means algorithm to distribute data points to clusters requires that the K value or cluster number be selected initially. The initial selection can be done using statistical methods or considering business needs.

Statistical Methods:

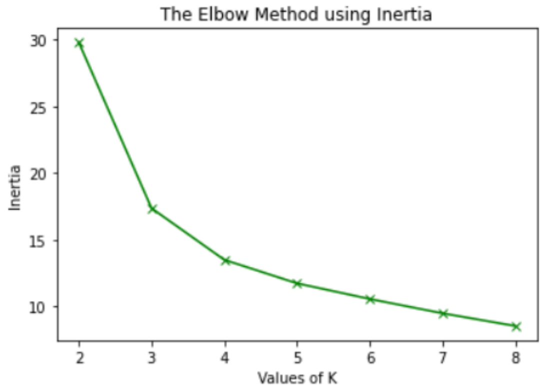
The methods used to find the value of k are:

1. Elbow curve:

Elbow curve method uses one of two variations:

1. Distortion: It is calculated as the average of the squared distances from the cluster centers (Within of the respective clusters. Typically, the Euclidean distance metric is used.
2. Inertia: It is the sum of squared distances of samples to their closest cluster center.

   The concept utilized in these methods is that distortion/inertia is marked for multiple values of k and then the point at which an elbow-like curve(angle) appears can be considered as the most appropriate value of k.



Here the most appropriate value appears at 3.

2. Silhouette score

The silhouette of a data point is a measure of how closely it matches to data within its cluster(cohesion) and how loosely it matches to data of the nearest neighbouring cluster(separation).

So to compute silhouette metric, we need to compute two measures

silhouette score=$(p-q)/max(p,q)$

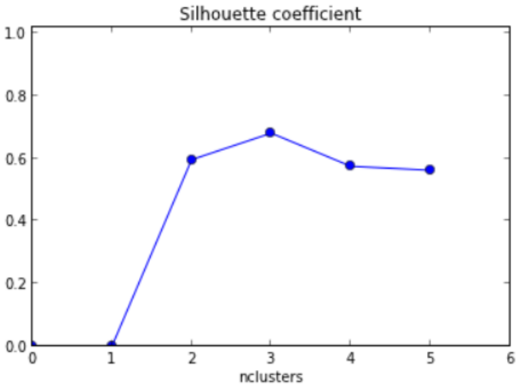
where:

p is the mean distance to the points in the nearest cluster that the data point is not a part of

q is the mean intra-cluster distance to all the points in its own cluster.

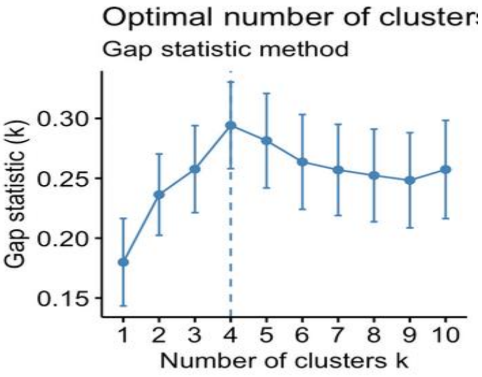
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

For every k, plotting **Average_sil i.e. mean {S(I)}** against k will give a peak value which will indicate optimal k.



The Gap Statistic:

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.



Business aspect:

Sometimes, there are business aspects that need to be satisfied when determining number of clusters.

For example: If we are looking to distribute data to find who would suit the newly created Platinum Members, Gold Members, Silver Members and standard members for a club, we need to distribute data in 4 clusters without really considering the statistical aspects of it.

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

Ans: In any given data set, the numerical data can be distributed across various ranges.

For example, Age may vary from 0 to may be 122 (oldest person recorded). Weight may vary from a few grams to 100+ Kg. Again, weight may also be measured in pounds which will give different values. Height may also be measured in feet and inches or in cm. In each case, the range of values can be highly varied. An additional column of income can cause the range to vary even further.

When creating clusters, each variable will have equal importance in determining clusters. However, in cases like above, the variable with the highest range, income in above case will dominate the other variables.

Hence, we need to convert all variables to a similar scale before performing clustering operations so as to ensure that one variable does not dominate the others.

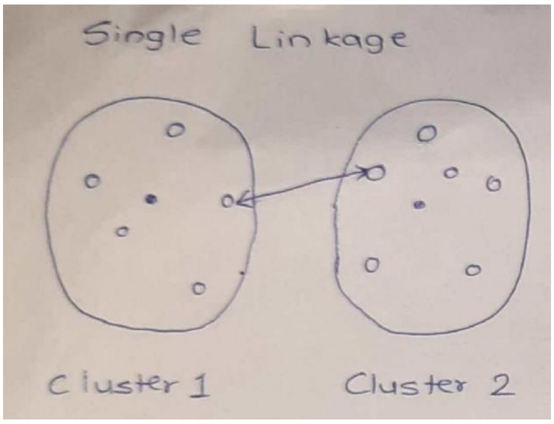
e) Explain the different linkages used in Hierarchical Clustering.

Ans: Linkages are the measure of distance between various data points when configuring Hierarchical Clustering.

Different linkages used for hierarchical clustering are:

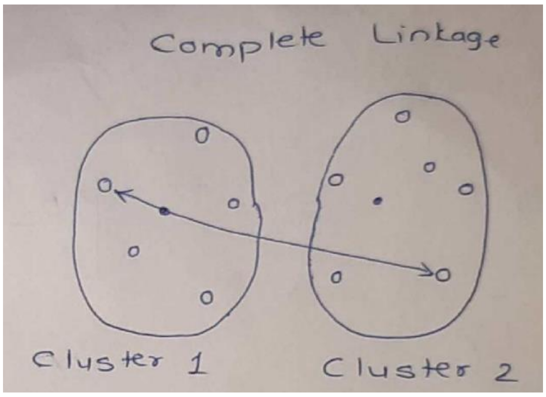
1. Single Linkage:

This is a measure of least distance between points in 2 clusters. Using this method can cause very loose clusters.



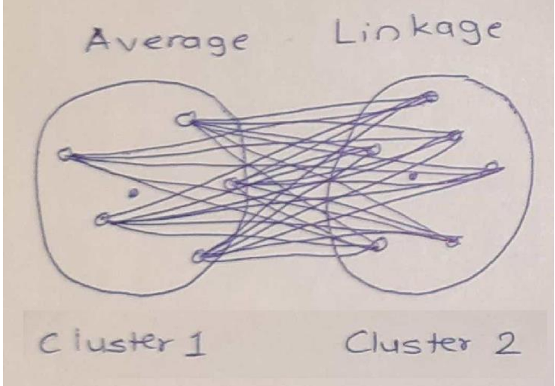
2. Complete Linkage:

This is a measure of highest distance between points in 2 clusters.



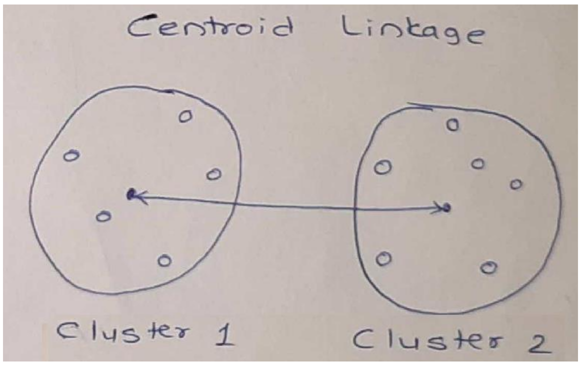
3. Average Linkage:

This is a measure of average distance between all points in one cluster and all points in the 2nd cluster.



4. Centroid Linkage:

This is a measure of distance between centroids of 2 clusters.


Centroid Linkage
Cluster 1
Cluster 2