**Question 1: Assignment Summary:**

**Business Problem:** Cluster the countries based on child mortality, GDP per capita , and net income per person. So, that HELP organization can provide financial aid during times of disaster, natural calamities to countries which are direst need of financial aid(help).

**Answer:**

* Loaded the data, visualized and checked for missing values
* Removed statistical outliers in exports, imports, income, gdpp. So, that none of the under developed countries get removed.
* Scaled the data using standard scaler
* Applied PCA
  + Found optimal components using scree plot
  + Reduce dimensionality of the data set
* Performed Hopkins test on the data to check for cluster tendency
* Applied K-Means clustering on transformed data
  + Found number of clusters
    - Elbow curve
    - Silhouette score
  + Visualized cluster plot using 2 principal components
  + Added output labels to original dataframe
  + Visualized box plots of GDPP, Income, Child\_mort
  + Identified the cluster Id of countries, which are in need of aid and formed a list.
* Applied Hierarchical clustering on transformed data
  + Performed Single linkage
    - Plotted dendrogram
    - Cut the tree to form clusters
    - Visualized cluster plot using 2 principal components
    - Added output labels to original dataframe
    - Visualized box plots of GDPP, Income, Child\_mort
    - Identified the cluster Id of countries, which are in need of aid and formed a list.
    - Clusters formed are loose and haven’t produced desired results
  + Performed Complete linkage
    - Plotted dendrogram
    - Cut the tree to form clusters
    - Visualized cluster plot using 2 principal components
    - Added clustered output labels to original dataframe
    - Visualized box plots of GDPP, Income, Child\_mort
    - Identified the cluster Id of countries, which are in need of aid and formed a list.
* Formed a final list of countries from K-Means as it produced results without leaving any under developed countries compare to hierarchical.

**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/standardization before performing Clustering.  
      e) Explain the different linkages used in Hierarchical Clustering.

**Answer:**

a)

|  |  |
| --- | --- |
| **K Means** | **Hierarchical** |
| K value should be choosed before modeling | There is no such choosing in hierarchical before modeling |
| Initial centroid values determine the final clustering | There is no such dependency |
| K means is group having similar objects among them and dissimilar to other clusters | It forms tree like structure (parent child relation). We use dendrogram to plot it. |
| Based on k value choosed initially before modeling, it creates clusters based on it.  If we want to change the number of clusters. we must run the entire model again. | We can decide on number of clusters based on the height where to cut the dendrogram. We can change number of number clusters without running entire model again. |
| Low time complexity compared to Hierarchical | High time complexity compared to K means |
| We have cost function to optimize | There is no such cost function |
| It can be used only for numerical variables | It can be used for numerical and categorical |

b) Steps of K means Clustering:

1. Randomly pick k points in space or use k-means++ or other initializers to pick the initial centroids.
2. Assign each observation in given data set to the cluster that has minimum the euclidean distance between the data point and its centroid. This step is call assignment step
3. Recompute the centroids of each cluster by taking the means of each of the observations in each cluster. This step is call optimization step
4. Repeat assignment and optimization steps until either the centroids don't change, or maximum number of iterations has been reached.

c) Statistically aspect:

k value in k-means is chosen using elbow curve or silhouette score. Curved parts in elbow curve are choosed as, k value or choose the number of clusters which has high silhouette score. Silhouette coefficient is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation)

Business aspect:

We do hierarchal clustering and cut the tree to get 3-15 clusters. Based on the cohesion and separation and number of observations in each cluster. Use the number of clusters which are having good number of observations on which business can monetize, with good cohesion and heterogenous for other clusters. Use that number as k value in K means

d) K means and hierarchical clustering uses euclidean distance to determine the distance between the centroid and data point to form clusters. As such distance metrics are sensitive to variations within magnitude and scales of variables. This results in overpowering one variable over the other. To prevent this, we do scale the data before clustering.

e) In hierarchical clustering we have 3 types of linkages based on the measure of distance between two clusters used in proximity matrix

1. Single linkage: Distance between two clusters is the shortest distance between any two data point in each cluster
2. Complete linkage: Distance between two clusters is the longest distance between any two data points in each cluster.
3. Average linkage: Distance between two clusters is the average distance between each point in the cluster to every point in other cluster.

**Question 3:**

    a) Give at least three applications of using PCA.  
    b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.  
    c) State at least three shortcomings of using Principal Component Analysis.

**Answer:**

1. **Applications of PCA:**

* Dimensionality reduction, reduces number of columns in a data set without losing much information
* Data visualization, it helps in visualizing the larger data set where n dimensional plots are required, after applying PCA, we can visualize the transformed data (to less dimensions) using principal components
* Improving the performance of model, as number of features decrease models converge quickly with less computation
* Noise reduction
* Reduce multicollinearity in the data set, after applying PCA, formed principal components have no multicollinearity.
* Helps finding latent themes in data.

1. **Basis transformation:**

Basis of vector space is a set of vectors that can be used as coordinates. It is a unit in which vector of matrix can be expressed Ex: i, j are standard basis vectors.

Any vector spaces as multiple basis, there exists a relation between vectors of one space to the other space. The ways of translating vectors in terms of basis (one vector space) to another basis (another vector space) is known as basis transformation.

Ex: {ai-bj, ci-dj} are basis vectors of a vector space V1. {i, j} are standard basis V2

We can rewrite ai-bj and ci-di basis vectors as

Standard basis as

Equate them to get the change matrix

is the change matrix.

Multiplying any vector in V1 vectors space with change matrix transform it to V2.

**Variance as information:**

Variance is equivalent to information that can be extracted from data. How important a column is explained by how varied the data that is present in the column. If the data in the columns is has no variance it means it doesn’t give any information and can be removed.

Column with more variance explains(provides) more information

µ is mean of the column, N is number of points in the column

To choose the basis vectors, we use the variance as information. The direction in which maximum variance is explained is taken as basis vectors to find PC in that direction.

1. **Shortcomings of PCA**
   1. PCA assumes linear combination, which doesn’t produce proper result if the data is non linearly related.
   2. In PCA components are orthogonal which might not provide better results in some cases
   3. PCA assumes columns with low variances are not useful which is not true in case of imbalanced data