Assignment Summary

Objective

An international humanitarian NGO 'HELP' 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.

After the recent funding programs, they have been able to raise around $ 10 million and wanted to utilize this money strategically and effectively.

Problem Statement

While the CEO of help wants to ensure the amount raised from funding program is effectively utilized, the significant challenge come while making this decision is mostly related to choosing the countries that are in the direst need of aid.

Hence as a data analyst, we have to is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. We need to suggest the CEO the right countries he should focus most.

Approach to solution

1. **Read and visualize the data:** The first and most important thing to do before starting to analyze the data is to read and visualize it properly. Understand the number of columns and their datatypes.

2. **Data Cleaning:**_ The next important process is to clean the data. This includes:
    - Identifying Missing data
    - Identifying wrong data type
    - Removing duplicates

3. **Data Preparation:**
    - **Derived Metrics**: The variables export, health & imports are percentage values and hence wouldn't give the clear picture of spending by the country. For example two countries (Afghanistan & Albina) have similar import % but not necessarily have the same gdpp which doesn't give accurate of idea of country being develop or under develop. Hence we need to derive the actual value of this variable.

    - **Exploratory Data analysis:** This helps to visualize the top/bottom countries on different socio-economic and health factors.

- **Correlation between different variables:** Plot the correlation matrix and check if the data is indeed highly correlated so that the usage of PCA in this scenario is justified.

- **Scaling the Data :** Most software packages use SVD to compute the principal components and assume that the data is scaled and centered, so it is important to do standardization/normalization

- **PCA (Principal Component Analysis) on the data to remove redundancies:** Principal component analysis (PCA) is one of the most commonly used dimensionality reduction techniques to improve model performance.

  To find out the number of PCA components, which would best describe the Variance, we drew the Scree Plot and found out that first 3 principal components can well explain around 90% variance.

- **Outlier analysis and Treatment:** Post Outlier treatment, the number of rows reduced from 167 to 119.

- **Hopkins statistics test:** This is a way of measuring the cluster tendency of a data set. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0

4. **Building Model:**

K-means: In this method, first we initialize k points, called means, randomly. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far. We repeat the process for a given number of iterations and at the end, we have our clusters.

- **Finding the right number of clusters:** We drew Elbow Curve to get the right number of clusters. Looking at our elbow graph, we inferred its good to proceed with either 4 or 5 clusters
- **Silhouette Analysis:** With this analysis, we observed that there were good number of countries in all the 5 clustered formed.
- **Scatter plot:** From the business understanding we have learnt that **Child Mortality**, **Income**, **Gdpp** are some important factors which decides the development of any country. We have also cross checked with Principal components and found that these variables have good score in PCA. Hence, we will proceed with analyzing these 3 components to build some meaningful *clusters*.
- **Bar plot:** Post the bar plot construction we found that Child Mortality is highest for Cluster 0 and Cluster 3.These clusters need some aid. Income and Gdpp are measures of

development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 0 and 3. Hence, these countries need some help.

- **Final Analysis on K means:** Cluster ID 0 with 23 countries are the ones which are in direst need of aid based on analysis of various Socio-economi factors.

Hierarchical Clustering: It involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy.

- **Single Linkage:** First we tried with the single linkage hierarchical clustering and didn't find good results.
- **Complete Linkage:** Later we tried the complete linkage and chose the number of cluster to be 4.
- **Scatter plot:** We drew Scatter plot on Principal components to visualize the spread of the data.
- **Merged the PCA dataframe with Original dataframe.**
- **Final inference with this clustering:** Cluster ID 0 with 23 countries are the ones which are in direst need of aid based on analysis of various Socio economic factors.

Closing Statement - Both the clustering mechanisms- K means and hierarchical Clustering have identical results and Cluster ID 0 with 23 countries are the ones which are in direst need of aid based on analysis of various Socio economi factors.

Clustering

Compare and contrast K-means Clustering and Hierarchical Clustering.

1. **Time Complexity** - Hierarchical clustering can't handle big data well but K Means clustering can.
   Time complexity of **K Means is linear -> O(n)**
   Time complexity of **Hierarchical clustering is quadratic i.e. O(n2).**

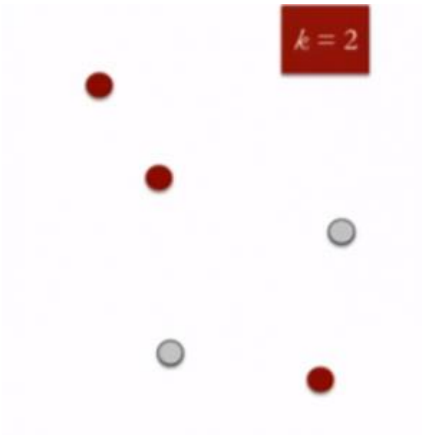
2. **Results –**
   In K Means clustering**,** since we start with random choice of clusters, **the results produced by running the algorithm multiple times might differ**.
   In **Hierarchical** clustering - **Results are reproducible**

3. **K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).**

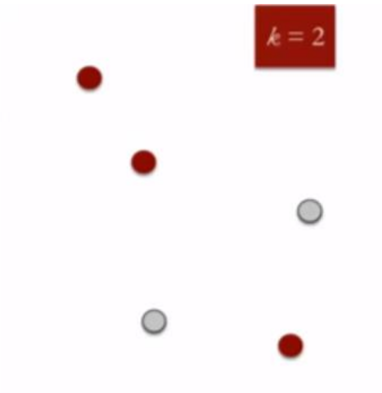
4. **K Means** clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into.
   In **hierarchical** clustering you can stop at whatever number of clusters you find appropriate by interpreting the dendrogram

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

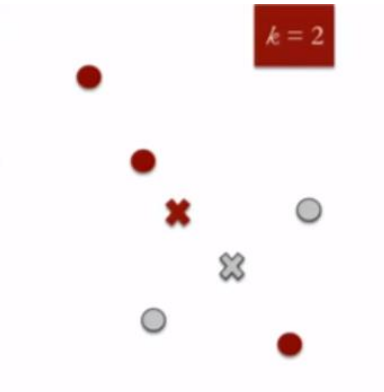
1. Specify the desired number of clusters K: Let us choose k=2 for these 5 data points in 2-D space.



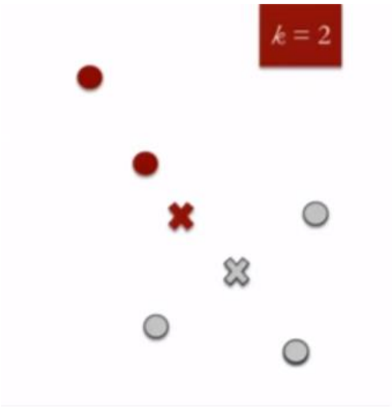
2. Randomly assign each data point to a cluster: Let’s assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.



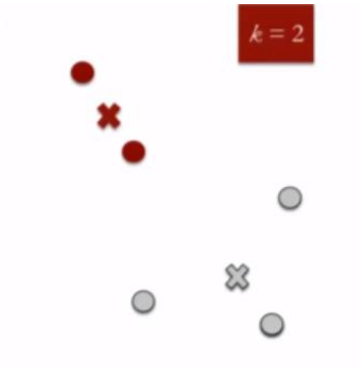
3. Compute cluster centroids: The centroid of data points in the red cluster is shown using Red Cross and those in grey cluster using grey cross.



4. Re-assign each point to the closest cluster centroid: Note that only the data point at the bottom is assigned to the red cluster even though it’s closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster



5. Re-compute cluster centroids: Now, re-computing the centroids for both the clusters.



6. Repeat steps 4 and 5 until no improvements are possible: Similarly, we’ll repeat the 4th and 5th steps until we’ll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

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

Two commonly used methods of determining optimal value of K.

1. Elbow Method
2. Silhouette algorithm

Elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing 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

Silhouette Analysis - Silhouette score analysis to find the ideal number of clusters for K-means clustering. 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. Formulae -

$$\text{silhouette score} = \frac{p-q}{\max(p,q)}$$

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.

Explain the necessity for scaling/standardization before performing Clustering.

Standardization is an important step of Data preprocessing. It controls the variability of the dataset, it convert data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms

It is critical step if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

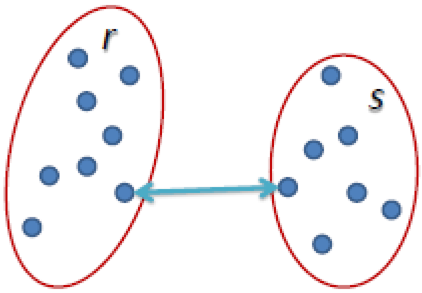
Also, most software packages use SVD to compute the principal components and assume that **the data is scaled and centred,** so it is important to do standardization/normalization

Explain the different linkages used in Hierarchical Clustering

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering,

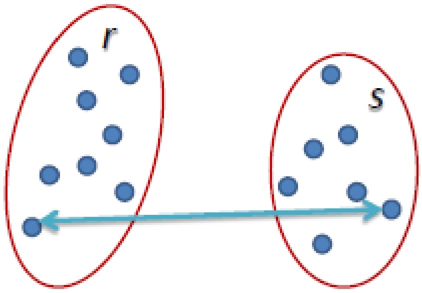
1. Divisive
2. Agglomerative.

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 example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

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 example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Principal Component Analysis

Give at least three applications of using PCA.

PCA is predominantly used as a dimensionality reduction technique in domains like **facial recognition, computer vision and image compression**. It is also used for **finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology**, etc.

1. **PCA for images compression** - an image is compressed by using different principal components, and concepts such as image dimension reduction and image reconstruction quality are explained. Also, using the almost periodicity of the first principal component, a quality comparative analysis of a compressed image using two and eight principal components is carried out. Finally, a novel construction of principal components by periodicity of principal components has been included, in order to reduce the computational cost for their calculation, although decreasing the accuracy.

2. **Quantitative finance** - In quantitative finance, principal component analysis can be directly applied to the risk management of interest rate derivative portfolios. Trading multiple swap instruments which are usually a function of 30-500 other market quotable swap instruments is sought to be reduced to usually 3 or 4 principal components, representing the path of interest rates on a macro basis. Converting risks to be represented as those to factor loadings (or multipliers) provides assessments and understanding beyond that available to simply collectively viewing risks to individual 30-500 buckets

3. **Neuroscience** - PCA is also used to discern the identity of a neuron from the shape of its action potential. Spike sorting is an important procedure because extracellular recording techniques often pick up signals from more than one neuron. In spike sorting, one first uses PCA to reduce the dimensionality of the space of action potential waveforms, and then performs clustering analysis to associate specific action potentials with individual neurons.

Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Basis transformation : PCA takes points expressed in the standard basis and transforms them into points expressed in an eigenvector basis. In this process of transformation, some dimensions with low variance are discarded and hence the resulting dimensional reduction.

Variance as information: In case of PCA, "variance" means summative variance or multivariate variability or overall variability or total variability. ... Their variances are on the diagonal. PCA replaces original variables with new variables, called principal components, which are orthogonal (i.e. they have zero covariations) and have variances (called eigenvalues) in decreasing order.

Below is the covariance matrix of some 3 variables. Their variances are on the diagonal, and the sum of the 3 values (3.448) is the overall variability.

1.343730519	-.160152268	.186470243
-.160152268	.619205620	-.126684273
.186470243	-.126684273	1.485549631

Now, PCA replaces original variables with new variables, called principal components, which are orthogonal (i.e. they have zero covariations) and have variances (called eigenvalues) in decreasing order. So, the covariance matrix between the principal components extracted from the above data is this:

1.651354285	.000000000	.000000000
.000000000	1.220288343	.000000000
.000000000	.000000000	.576843142

Note that the diagonal sum is still 3.448, which says that all 3 components account for all the multivariate variability. The 1st principal component accounts for or "explains" 1.651/3.448 = 47.9% of the overall variability; the 2nd one explains 1.220/3.448 = 35.4% of it; the 3rd one explains .577/3.448 = 16.7% of it.

So, what does it mean to say that "PCA maximizes variance" or "PCA explains maximal variance"? It doesn't mean to find the largest variance among three values 1.343730519 .619205620 1.485549631. PCA finds, in the data space, the dimension (direction) with the largest variance out of the overall variance 1.343730519+.619205620+1.485549631 = 3.448. That largest variance would be 1.651354285. Then it finds the dimension of the second largest variance, orthogonal to the first one, out of the remaining 3.448-1.651354285 overall variance. That 2nd dimension would be 1.220288343 variance. And so on. The last remaining dimension is .576843142 variance.

Mathematically, PCA is performed via linear algebra functions called eigen-decomposition or svd-decomposition.

State at least three shortcomings of using Principal Component Analysis.

While PCA (Principal Component Analysis) helps remove correlated features and improves algorithm performances, it has few shortcomings:

1. Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.

Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be applied.

PCA is affected by scale, so you need to scale the features in your data before applying PCA.

3. Information Loss:

Although Principal Components try to cover maximum variance among the features in a dataset, if we don’t select the number of Principal Components with care, it may miss some information as compared to the original list of features.