Assignment: Part II

Question 1: Assignment Summary

Problem Statement:

In the PCA and clustering assignment, the focus of the exercise is on finding countries which are in need of aid. The financially weak countries which are struggling with high child Mortality rate, low life expectancy, low income and low GDPP are unable to spend more money towards health spending's.

EDA/Outlier Analysis:

I started the assignment with EDA to find any missing values, Outliers in the data as well as derived fields. Outliers before PCA listed out very rich as well as poop countries. I have avoided outlier removal, as this will result in removal of counties which actually need the aid.

PCA and selection of principle components & Outlier treatment:

After EDA analysis, I standardized the dataset before applying the PCA algorithm for dimensionality reduction. I have selected 4 principle components as they describe around 94% of variance, so the information loss is minimal.

After PCA I checked for outliers, as important data was being removed in the outlier treatment, I took decision to go ahead without removal of the outliers.

Clustering:

After selection of the 4 principle components, I checked for the clustering ability of the dataset with the help of Hopkins score which is around 0.94 which is excellent.
With the help of Elbow curve analysis and avg. silhouette score, I have selected K =3 for k-means clustering.

K-means clustering: I created K-means clustering model with K=3, after creating 3 clusters. After creation of clusters and merging of original dataset , I did analysis of the cluster data based on the average values of GDPP/child Mortality rate/Income/Health spending etc. I created binning based on the average values. And sorted data in the underdeveloped cluster.

The selection of 5 countries is based on the factors such as Hight Child mortality rate, low income, low GDPP.

Hierarchical clustering: I applied both Single Linkage and complete linkage on the PCA dataset. Single Linkage clustering did not provide very effecting clustering.
The full linkage hierarchical clustering on another hand, provided results for 9 cluster with around 4 cluster with 1 or 2 countries, so this method of clustering was not very useful as it help me to derive 4 countries for recommendation.

Recommendation: Based on K-means clustering, I selected 5 counties which have Hight Child mortality rate, low income, low GDPP.

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali

These are the countries which have very high child mortality rate and low income and GDPP

Question 2: Clustering

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

Clustering is process of dividing the dataset into similar groups called clusters. We have K-means and Hierarchical clustering algorithms clustering.

K-means:-

i. For K means clustering initial choice of K is required
ii. K-means can work on large volume of data as the time complexity of K-means is linear which is an advantage over the Hierarchical clustering
iii. As K-means works with random k-values each run will result in different results, so its not repeatable
iv. For hyper Spherical clusters K-means seems to work better

Hierarchical Clustering:-

i. For Hierarchical clustering no prior selection of no of cluster is required which is an advantage over K-means clustering
ii. Hierarchical clustering is resource extensive as it has quadratic time complexity
iii. Hierarchical clustering can work agglomerative where it starts with sequentially merging similar clusters and it has divisive Hierarchical clustering method where all observations are merged in one group and then successively they are partitioned in clusters with similar observations.
iv. With Hierarchical for each run we will get same results which makes it a repeatable algorithm

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

Following is the brief process followed in the K-means clustering algorithm to create K clusters with N input datapoints.

1. In K-means clustering the initial choice of the K cluster centres is random, so we need to select K random points as centres in the first step.
2. Based on the shortest distance of each datapoint to the Centre, assign all the data points to their nearest centres so that initial clusters are formed.
3. Based on the mean of the distance of datapoints to centre of respective cluster, calculate the new centre
4. Now Again reassign all the datapoints to these newly calculated centres.
5. Repeat the process in step 3 and 4 , till there is no change in the centre values and all datapoints are fixed with their respective centres.
6. This will be our final clustering

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

Depending on the choice of K value, clustering will provide different results. Choice of K in K-means is random but we need to select a optimal value of K which can provide us best results.

For this purpose there are two techniques which are used in selection of K value.

1. Elbow Method:-
   i. In elbow method, clustering algorithm is computed with different values of K. e.g. range of 1 to 10
   ii. For each K clusters, within cluster Sum of Squares(WSS) is calculated This data of WSS is plotted for all the cluster values
   iii. The K value where the plot shows a bend is selected as K value for number of clusters which indicates lower number of clusters with significant lower value of WSS
2. Silhouette Method:-

i. In Silhouette method, clustering algorithm is computed with different values of K. e.g. range of 1 to 10
ii. For each K value, the average Silhouette value is calculated, which is a indicator of how similar is a object to objects in own cluster compared to Objects in other clusters
iii. Plot the avg. sil according to the K values
iv. Location of maximum is the best number of clusters selection.

The best values of K means helps use in following aspects:
a. Its helps us identify the hidden clusters in the dataset
b. Enhances within-cluster similarity and between-cluster dissimilarity.

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

In clustering, we need to find the distance(Euclidian distance) between the datapoint. The measured distance helps to find similarities between observation. So in this scenario, attributes with large scale will dominate the attributes with smaller scale and which will result in incorrect clustering .

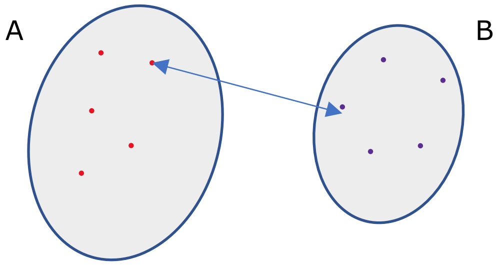
To overcome this drawback, we need to do standardisation of the data so that all data is at same normal scale and which can help in effective Clustering.

The Standardisation process will convert all features to mean of 0 and Standard deviation of 1 and same scale.

e) Explain the different linkages used in Hierarchical Clustering.

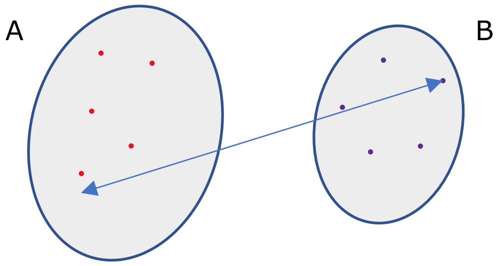
In Hierarchical Clustering to find the representative distance between two clusters we use 3 types of linkages.

1. Single Linkage: In this method the distance between two clusters is defined by the shorted distance between two points in each clusters.



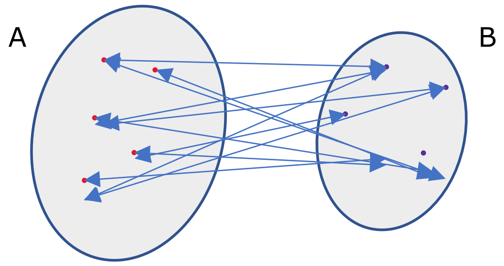
In above 2 clusters with single linkage min(D(Xai,Xbj)) is the distance between two clusters A and B

2. Complete Linkage: In this method the distance between two clusters is defined by the largest distance between two points in each clusters.



In above 2 clusters with single linkage max(D(Xai,Xbj)) is the distance between two clusters A and B

3. Average Linkage: in this method, distance between two clusters is defined by average of the distance between every point in a cluster to each point in other cluster



Question 3: Principal Component Analysis

a) **Give at least three applications of using PCA.**

PCA has multiple application when it comes to dimensionality reduction among data with linear relationship. Below are few applications

1. **Better Visualisation:-** When we have data with high dimensionality it is not feasible to visualize the data. PCA finds components with high variance and helps to show strong pattern in the dataset. Data exploration and analysis is made easy with the help of PCA
2. **Reduce Size of Dataset:-** As PCA works with aim of converting correlated variables into linearly uncorrelated variables with dimensionality reduction. The no of variables are reduced to smaller set of useful variables which describe the dataset better. Which in turns results in the size reductions of the dataset
3. **Different Data Perspective:-** As PCA gives the best linearly independent variables and different combination of variables it can help us uncover hidden trends and give us different perspective of the data
4. **Algorithm Performance:-** As PCA reduces no of features so it helps the algorithms performance as more features means more processing power.

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

1. **Basis transformation:-** Basis is set of B vectors in a vector space V such that every vector in vector space V is represented by unique linear combination of the B vectors.
   So basis transformation is a process where we would be representing our data as it is with new columns which are different that the original columns with minimal loss of the information.
   This is achieved by converting the underlined information from one set of basis to another set such that the information is fully represented by the newly selected basis.

2. **Variance as information:-** This building block works with idea of variance in the data provides information about the data.
   For example: if we have a dataset with women pregnancy related data in hospital with a column "Gender" and the column has only "F" value. This feature does not add any additional information to our dataset.
   On contrary "pregnancy duration" and "weight of the new born" columns capture variance in the data and provide useful information to the dataset.

So ideally it would be useful to drop low or no variance data and keep the data with variance.
Similarly if two columns are highly correlated then both of them do not add additional information, in this case only one can suffice the purpose so second variable can be dropped. This is the basic idea behind variance is information building block of PCA.

c) State at least three shortcomings of using Principal Component Analysis.

1. **Information Loss:-** As PCA tries to identify and cover maximum variance among features into components, if the no of principle components are not selected correct/properly can results in missing out on some of the information in the features which in turn is information loss
2. **Interpretability of the Independent features:-** Principle components are linear combination of original features which is helpful for further processing but it is not readable and interpretable as the original features.
3. **Data Standardisation:-** PCA need data to be always standardized otherwise the discovery of Principle components will not be optimal.
4. **Categorical Variable:-** PCA is not advised to use on the categorical data
5. **Low Variance in Components:-** PCA focus on higher variance in components and assumes that lower variance is not important this can result in loss of information