PCA-Clustering Assignment Part – II

Question 1: Assignment Summary

- **Aim:** We want to suggest the CEO to decide how to use the raised money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
  We will suggest the countries which the CEO needs to focus on the most using some socio-economic and health factors that determine the overall development of the country

- **Methodolody:**
  1. First, check for null values and data quality issues, correct some columns given as % of gdpp.
  2. Then applied normal scaler to scale the data in z-scored
  3. Then applied the PCA and checked using scree plot , how many components can give high variance. It turned out to be 5 PC giving 96% variance.
  4. Then performed outlier analysis for each PC. But outliers were not removed as they either themselves formed a cluster or were special in needs like Nigeria.
  5. Now whether to use the clustering, we checked with Hopkin's method which came above 90% everytime, so good to go.
  6. Found out the k for k-means algo. 4 seemed a good number from elbow and silhouette.
  7. Performed clustering and analysed using box plots for different parameters, group 2 came least in all.
  8. Performed hierarchical clustering for both single and complete linkage.
  9. But data was not conformant to hierarchy and so dropped the idea of using it.
  10. Furthured with k-means clustering
  11. Found out top 10 least performing countries in all the 3 categories and took those countries for recommendation to help (about 18 countries).

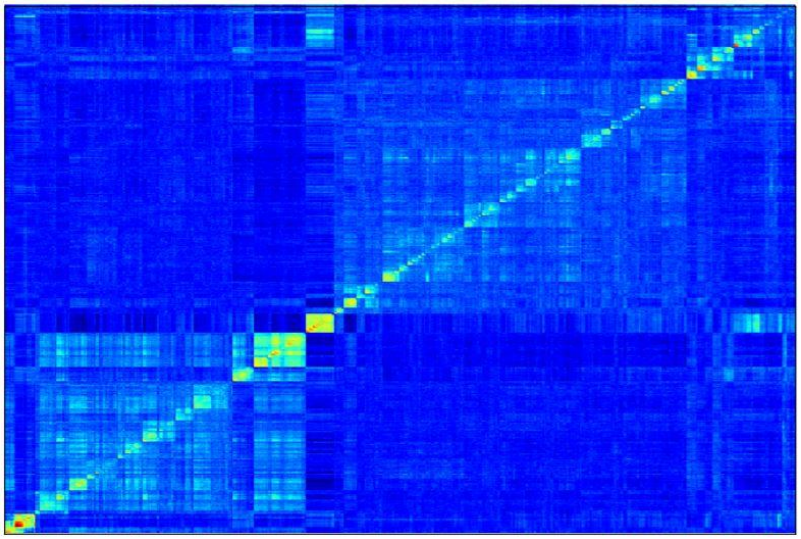
Question 2: Clustering

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

Ans. –

S.No	K-means	Heirarchical
1	K-means requires K(number of clusters) to be defined in advance.	Heirarchical clustering doesn't require this as clusters can be got from breaking the dendogram at suitable positions.
2	It will depend on initial cluster centers and may give different clusters when different cluster centers are taken into account.	There is no such concept of initial clusters here. It is always giving the same result for a given data.
3	Termination point of algorithm is either when cluster centers are stabilized or max. number of iterations are reached.	It has fixed number of iterations of order n^2 , where n is size of dataset

4	It can be used for large amount of data as it will take less time (num_iter * n*k)	It takes more time to perform which is square of order of n.
5	It gives better results when clusters are spherically aligned(like circle in 2D, sphere in 3D, and so on) and may perform poor if data is not like that.	It gives better results when data is itself hierarchical in nature like shown below. Here big rectangles have smaller rectangles which themselves constitute smaller cluster.



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

Ans. There are following steps in the algorithm:

1. Choose a k for the k initial clusters. This can be done using Elbow curve and silhouette analysis.
2. Mark the initial k centroids. These will be the initial centers and act as centers for clusters.
3. Assign each data point to the clusters whose distance is minimum among the cluster centers. The distance function can be Euclidean distance, Manhattan distance, etc. according to need.
4. Now from each cluster, calculate the new centroid(mean) and now these are the new centroids.
5. Repeat steps 3 and 4 until:
    a. Centroids do not change in successive iterations OR
    b. You reach a max number of iterations you selected in the starting.

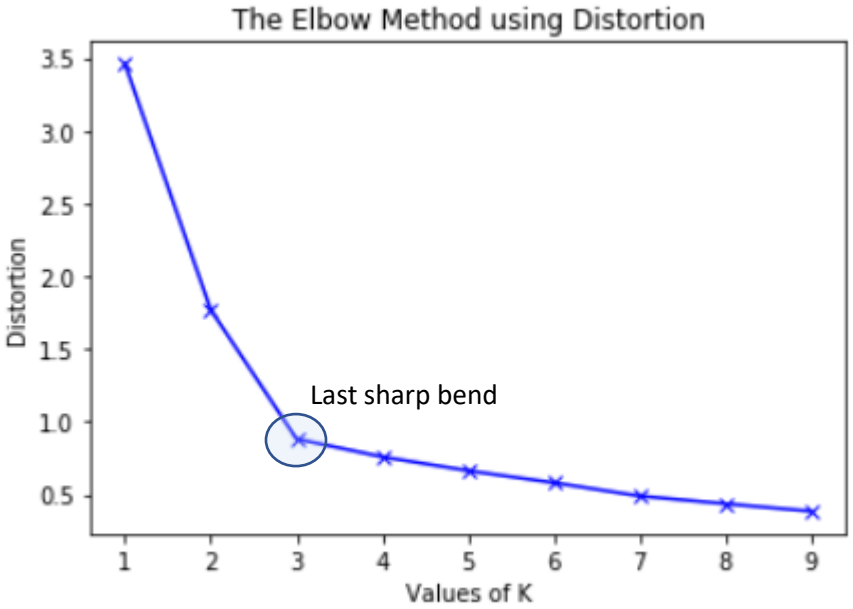
The result is a set of k clusters with each point belonging to some cluster.

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

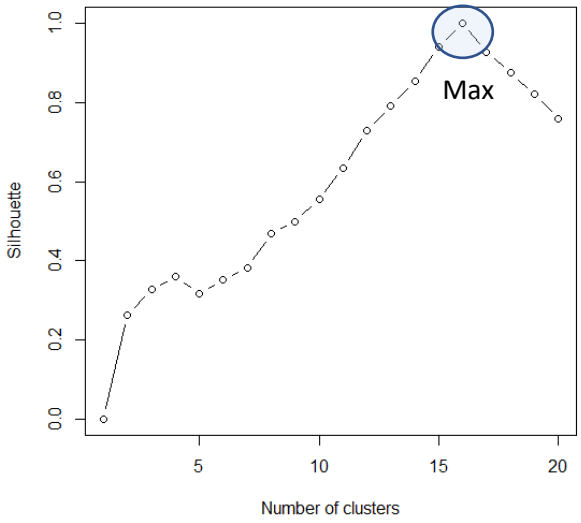
Ans: We can choose k with following statistical methods:

1. Elbow method:
    a. Perform clustering for different values of k.
    b. For each k, find the total within cluster sum of squares(WSS)

c. Draw a graph of k vs. respective WSS
d. We can get the correct number of clusters where the last sharp bend is there



2. Silhouette Method:
   a. Perform clustering for different k values
   b. Find the average silhouette of observations for each k
   c. Plot above vs. k
   d. The max on the curve represents optimal cluster size.



Now these are statistical methods but this may give results like 150 or 50 as k.
This may not be feasible in terms of Business perspective where there are only 2 or 3 segments required as in case of customer segmentation.

Thus it depends on business needs as well.

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

Ans. Scaling is important in situations where numerical values of the attributes are differing in order of magnitudes.
We can see this for example in data like height in millimeters and weight in kg.
If 2 points are 1500mm and 50 kg and 1600mm and 70kg then the distance will have more effect of height difference than weight difference despite of the fact that there isn't much difference in height as there is in weight.

There are a few different options for standardization/scaling, but two of the most frequently used are z-score and unit interval:

1. Z-score: transforms data by subtracting the mean value for each field from the values of the file and then dividing by the standard deviation of the field, resulting in data with a mean of zero and a standard deviation of one.
2. Min-max scaling: is calculated by subtracting the minimum value of the field and then dividing by the range of the field (maximum minus minimum) which results in a field with values ranging from 0 to 1.

e) Explain the different linkages used in Hierarchical Clustering.

Ans. Linkages refer to how the inter cluster differences are calculated in the Hierarchical Clustering.

These can be of mainly 3 types:

Single linkage: In this the distance is taken as the minimum distance between the distances of each point of one cluster to each other point of another cluster (or: the two clusters with the smallest minimum pairwise distance is chosen to merge).

Complete linkage: In this, the distance is taken as the maximum distance between the distances of each point of one cluster to each other point of another cluster (or: the two clusters with the smallest maximum pairwise distance is chosen to merge).

Average linkage: In this, clusters are chosen having minimum average pairwise distances between each points of clusters. It is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

Ans. PCA helps us in following ways:

1. Data Visualisation and EDA : It helps us to visualise data as number of dimensions are reduced and all PC are orthogonal to each other. So first 2 PC can be used for visualisation.

2. Reducing Multicollinearity: With a smaller number of uncorrelated features, the modelling process is faster and more stable as well as PC's are uncorrelated with each other and they are linear combinations of the original variables.
3. They help in capturing maximum information in the data set
4. Latent themes can also be found which may not be visible clearly as of now like genres of music within the ratings on apps like Gaana.
5. Noise reduction: As unnecessary and redundant features are merged in a single component, it also helps in noise reduction.

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

Ans.

1.Basis Transformation:

Basis: A set B of elements (vectors) in a vector space V is called a basis, if every element of V may be written in a unique way as a (finite) linear combination of elements of B.
The coefficients of this linear combination are referred to as components or coordinates on B of the vector. The elements of a basis are called basis vectors.

But the same vector V can also be written in some other basis.

Basis Transformation: Let V be a vector space of dimension n over a field F. Given two (ordered) bases
$B_{old} = (v_1, v_2, \ldots v_n)$ and
$B_{new} = (w_1, w_2, \ldots w_n)$ of V, it is often useful to express the coordinates of a vector x with respect to B_{old} in terms of the coordinates with respect to B_{new} .

This can be done by the *change-of-basis formula*, that is described below:

If $(x_1, x_2, \ldots x_n)$ and $(y_1, y_2, \ldots y_n)$ are the coordinates of a vector x over the old and the new basis respectively, the change-of-basis formula is

$$x_i = \sum_{j=1}^{n} a_{i,j} y_j,$$

for $i = 1, \ldots, n.$ In matrix form,

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$X = MY$

This M is called basis matrix and can be calculated as $M = XY^{-1}$

PCA makes use of this and changes the basis to some other vector in which correlated vectors are aligned.

2.Variance as information: More variance in a data may be attributed to more information(not always true). Thus, more spread means more variation. This is used by PCA to find the direction of

component vectors that contribute to more variance. These directions are called Principal Components.

Thus, Ideal basis vectors that we need to transform our original dataset to have following properties:

1. They explain the directions of maximum variance
2. When used as the new set of basis vectors, the transformed dataset is now suitable for dimensionality reduction.
3. These directions explaining the maximum variance are called the Principal Components of our data.

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

Ans. Following are the shortcomings of the PCA:

1. PCA is scale sensitive. It is a rotation trasnformation of the dataset, which means that doens't affect the scale of your data. That means that if you change the scale of just some of the variables in the data set, we will get different results by applying PCA .
2. PCA will work well only when components are orthogonal to each other. But that may not be the best case always. It's alternative technique is Independent Component Analysis
3. PCA works on linearity , that is, it makes linear transformation of the attributes. But there are other cases where combination may not be linear but other forms as well.