PCA AND CLUSTERING ASSIGNMENT PART2

Question 1:

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly.

Solution 1:

Problem--

HELP International is an international humanitarian NGO that 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this 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.

Objective—

Our main task is to cluster the countries by the factors mentioned and then present our solution and recommendations to the CEO

Methodology—

1. Import— Import all the libraries and data given

2. Data Understanding—Check with the length of data, null values of data, duplicates and converted percentage values to absolute values to get a clear scenario.

3. Standardizing of data-- Created the correlation matrix and checked the correlation between the variables that is checking the multicollinearity.

4. PCA (Principal Component Analysis) or Incremental PCA—For Dimensionality reduction and before that we standardized the scaler because of different scale of values for each variables.

5. We explained the variance ratio using the scree plot which gave us the result as 4 since 94% of the data was explained by this. So we got number of principal components as 4.

6. Basic Transformation—PCA components are calculated and a data frame is created for the same having original variables also.

7. Visualization and Outlier analysis is done.

8. Clustering—Hopkins value is calculated which was 0.80, hence good. Then silhouette and elbow curve is formed to decide the number of clusters (k) to be performed to perform k means clustering. Hence we got the value ok k as 3 and cluster ids are performed which are further merged with data frame having the original values.

9. A new data frame created having original variable's mean values and PCA's mean values too, then further visualizations are done using scatter plot and box plot which resulted that cluster 0 is having the countries which are in dire need as they are having high child_mort and high tot_fer with low income,gdpp,health etc.

10. Hierarchical Clustering— We performed single and complete clustering here and decided the number of clusters 3 as this is the point where maximum dendograms can be cut. Performed further analysis as did in k means clustering. Here also cluster 0 is the suitable for need of aid but proper clusters are not formed in this case.

11. Binning-- On the basis of mean values of original variables we binned the data and obtained the final country dataset.

In this case **k means clustering was better than hierarchical** clustering as proper clusters were formed which satisfied the condition but the same was not happening with hierarchical clustering.

List of countries which are in dire need of aid are:

Afghanistan
Angola
Bangladesh

```
                    Benin
                   Bhutan
                  Bolivia
             Burkina Faso
                  Burundi
                 Cambodia
                 Cameroon
               Cape Verde
Central African Republic
                    Chad
                 Comoros
          Congo, Dem. Rep.
               Congo, Rep.
            Cote d'Ivoire
                   Egypt
                  Eritrea
                   Gambia
                    Ghana
                   Guinea
             Guinea-Bissau
                   Haiti
                    India
                Indonesia
                    Kenya
           Kyrgyz Republic
                      Lao
                  Lesotho
                  Liberia
               Madagascar
                   Malawi
                     Mali
               Mauritania
              Mozambique
                 Myanmar
                   Nepal
                   Niger
                 Nigeria
                Pakistan
             Philippines
                  Rwanda
                 Senegal
            Sierra Leone
          Solomon Islands
                   Sudan
               Tajikistan
                 Tanzania
                     Togo
            Turkmenistan
                  Uganda
              Uzbekistan
                 Vietnam
                   Yemen
                  Zambia
```

Question 2

State at least three shortcomings of using Principal Component Analysis.

Solution:

1. Principal components are orthogonal to the others it's a restriction to find projections with the highest variance.

   Large variance = low covariance = high importance

2. Assume that, we have two-dimensional data (i.e., two features) and the joint distribution of the data follows multivariate normal distribution. One of the important properties of multivariate normal distribution is that, if the correlation between the features is zero, it means that features are orthogonal. The main job of PCA is to represent the data in lower dimensional by removing the redundant features. It achieves that through finding **orthogonal** principal components. The above property is not applicable if the joint distribution of data (not individual distribution of feature) follows other distribution instead of multivariate normal distribution. We also use a covariance matrix (covariance matrix is a function of correlation matrix) to find the principal components. The only one distribution (zero-mean probability distribution) which allows us to represent the whole data in a compact form is Gaussian distribution. It shows that PCA make an implicit assumption that data should follows Gaussian distribution. If data didn't follow Gaussian distribution, it would be difficult to extract independent statistical components by PCA.

3. The standard PCA always finds linear principal components to represent the data in lower dimension. Sometime, we need non-linear principal components.

4. PCA always considered the low variance components in the data as noise and recommend us to throw away those components. But, sometimes those components play a major role in supervised learning task.

5. Also Incremental PCA is more efficient than PCA.

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

Solution:

1. In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.

   K- means is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.

2. Hierarchical clustering, instead, builds clusters incrementally, producing a dendogram.

   The k-means clustering is parameterized by the value *k*, which is the number of clusters that you want to create

3. Hierarchical clustering has fewer assumptions about the distribution of your data - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points.

   In contrast k-means has a lot many assumptions.

4. Hierarchical clustering can be more computationally expensive but usually produces more intuitive results.

   k-means will often give unintuitive results

5. In Hierarchical methods, we create hierarchical decomposition of the given set of data. We create hierarchical decomposition in two ways such as from bottom to the top or top to down.

In k-means method, we find the mutually exclusive cluster of spherical shape based on distance. In this case, we can use mean or median as a cluster centre to represent each cluster. It is helpful in the small and medium size of data.