**Assignment: Part II**

**Question 1: Assignment Summary**

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( why you took that many numbers of principal components, which type of Clustering produced a better result and so on).

**Answer:**

The problem is to categorizing the countries using some socio-economic and health factors that determine the overall development of the country. Then suggesting the countries which needs to be focus on the most.

So first we import all the required libraries then we read the dataset then check the given dataset for the defect like null values and duplicate rows. Then we derive any new column in the dataset if needed. Then we separated all the numerical columns into one dataset. We then scale them using standard scaler. Then we applied PCA on the dataset and checked its variance ratio. Then plotted the scree plot, we could see that the around 93% of the variance of the data can be explained by the 4 principle components. Then we went and did the dimensionality reduction using 4 principle components. After that we did the outlier treatment of the components.

Then we go into clustering, but first we did the Hopkins test to see that the dataset will yield a good cluster or not. Then we did the K-Means clustering, to get the optimum cluster number we did the silhouette score and elbow curve and from both we get the value of k =2. So, we perform the k-means clustering for k=2. Then we went for Hierarchical clustering there we first made the dendrogram with single linkage but it wasn’t very clear. Then we made the next dendrogram with the complete linkage, here we cut the dendrogram so that we have 3 clear clusters. Then we did the hierarchical clustering with 3 clusters. Then we add the cluster label to the original dataset and by taking the average value of each of the cluster variable, we saw that the cluster 0 represent the underdeveloped countries and that are our main focus. So we separate them in a separate dataset and by seeing which country has the low gdpp and income and high child\_mort we decided those countries needed most attention.

**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) -** In K-means clustering we have to predefine the number of cluster but where as in Hierarchical clustering we don’t.

**-** K-means clustering is a fast algorithm that it can handle well very large amount of data, where as hierarchical clustering is slow it will take large amount of time when the size data is very big.

**b)** – First we have to specify the number of cluster to be made.

- Randomly assign the centroid for each cluster.

- Assign each of the data point to a cluster by calculating the distance of that point with respect to its nearest centroid.

- Recalculate the centroid of the cluster.

- Repeat last 2 steps iteratively until there is no change.

**c)** K is calculated by using the silhouette score and elbow curve. Elbow curve is the sum of the square of the distance (SSD) of the data points to their closest centroid. As number of cluster increases SSD decreases, but the question is does SSD goes down enough compare to the previous number of cluster, i.e. is adding another cluster is justified. So when we plot the SSD with respect to the number of cluster then plot is decreasing slope having elbow like shape i.e. the negative slope decreased that is the cutoff point of number of cluster.

The silhouette score calculates that the data point in a cluster should be similar and that between other cluster should be dissimilar. The values of Silhouette score lie in between -1 and 1. A score closer to 1 indicates that the data point is very similar to other data point in the cluster. A score closer to -1 indicates that the data point is not similar to the data points in its cluster. So the value of k for which the silhouette score is maximum is the optimum number of k.

But we also have to consider the business aspect of this. E.g if the value of k by statistics comes out to be 3 but from business point of view 4 number cluster makes more sense then we would proceed with k=4.

**d)** Scaling gives all the variable equal weightage. If we don’t scale them then it might create problem, e.g. if the data set contains hair length(in cm) and height of person(in cm) then average hair length is much smaller than height of person, if we don’t scale them then the effect of hair length will be neglected while clustering.

**e)** There are three types of linkages use in the hierarchical clustering.

Single Linkage: Here, distance between two clusters is defined as the shortest distance between points in the two clusters.

Complete Linkage: Here, distance between two clusters is defined as the maximum distance between any 2 points in the clusters.

Average Linkage: Here, the distance between two clusters is defined as the average distance between every point of one cluster to every point of other cluster.

**Question 3: Principal Component Analysis**

      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:**

**a)** PCA is used for dimensionality reduction in the field of fascial recognition, image compression, noise reduction etc.

It is also used to find pattern in the data having high dimension that is very large number of variables, in the field of finance, datamining etc.

It is also used in the field of neuroscience, it is used to discern the identity of a neuron from the shape of its action potential.

**b)** Basis Transformation allows us to represent the same data in multiple basis vectors. The more variance a column has, the more information it has and more important it is for our modelling process. Therefore, the one which explain low variance can be eliminated from our dataset without effecting our result much. But most of the times, dataset doesn’t have column in the order of variance that we want. In some cases, the variance might have been equally distributed amongst all the column. PCA helps in finding the best possible set of basis vectors from given dataset such a way that the variation is non-uniformly distributed amongst them – some columns now explain far more variance than other columns. This makes it easier to choose which column to keep and which to discard.

**c)** Shortcomings of Principle Component Analysis are:

- PCA only performs well on linear model.

- PCA needs the components to be perpendicular to each other, though in some cases, that may not be the best solution.

- PCA assumes that columns with low variance are not useful, which might not be true in prediction setup, especially in classification problem with class imbalance.