# Clustering & PCA Assignment 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)

**Note**: You don't have to include any images, equations or graphs for this question. Just text should be enough.

**Answer: -**

My job was to identify top countries that are direst need of aid with the provided dataset. I first reviewed the details of the dataset with help of .head, .info & .describe commands and removed the outliers for HIGH gdpp ONLY. Because, when I reviewed the correlation between different columns it showed a huge difference between the max and 75 percentile and other columns were not so high and was having positive correlation with income, life\_expec and others. On the other hand, gdpp was having a negative correlation with child\_mort.

I preserved the country column as to be used in future for INDEX reference and dropped for standardisation with help of StandardScaler and PCA. I then reviewed explained\_variance\_ratio\_ and moved with Scree plot to identify how many PCA components are required. I reviewed all the PCAs with boxplot to review the outliers and couldn’t conclude any pattern hence proceed for HOPKINs analysis which came as 86.9%. This indicates I can proceed with these PCAs for Silhouette score and Elbow curve analysis.

I reviewed both Silhouette & Elbow curve clustering with respect to gdpp, income & child\_mort and observed similar patterns for both of them though and the list of counties are same apart from Botswana was identified by Elbow curve and not by Silhouette score analysis.

Post this I moved with Hierarchical Clustering by Complete linkage as simple linkage I can’t derive any conclusion. I observed, I can go by 4 or 5 clustering and analysed both options and verified the final list for direst need of aid was same on both.

Post this I compared the final list generated by Silhouette score/Elbow curve analysis and Hierarchical Clustering by Complete linkage and concluded both lists are identical. Hence, went with this to be the FINAL list of countries which are direst need of aid.

**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/standardisation before performing Clustering.  
      e) Explain the different linkages used in Hierarchical Clustering.

**Answer: -**

1. K-means clustering can handle big size of data whereas Hierarchical Clustering can’t.

K-means provides concentrated clusters than Hierarchical Clustering

Difficult to predict K-value, but for Hierarchical Clustering its easy

Hierarchical Clustering results are always reproducible, however in K-means its not as we go by random choice of clusters

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

1. Algorithm of K-means clustering –

Let’s say we have N data points X = {x1,x2,x3,….,xn} and its centres V = {v1,v2,……vc}.

We will select c cluster centres randomly.

Add distance between each data point and cluster centres.

Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.

Recalculate the same again and continue in this route. If we don’t get any new centres then we can stop and conclude the centres.

1. K value is chosen when 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.

In statistics terminology this is a task of grouping set of objects in such a way that similar or more similar objects comes closer and create a group and its centroid. This groups are called clusters.

In business aspect we are grouping similar items together to take proper actions on similar items in a quick go. We don’t need to pick and chose what needs to be done, instead we can do the same thing for a group in one go and expedite the work.

1. There are 3 types of Hierarchical Clustering –

Single Linkage – The shortest distance between points in the two clusters

Complete Linkage – The maximum distance between any 2 points in the clusters

Average Linkage – The average distance between every point of one cluster to every other point of the 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: -**

1. 3 application of using PCA –

Facial recognition

Image compression

Data mining

1. Building block of PCA –

It lowers noise sensitivity, the decreased requirements for capacity and memory, and increased efficiency given the processes taking place in a smaller dimension.

We need to first standardise the dataset and fit with a defined random state. Depending upon the Scree plot observation, we can reduce the dimensionality and then fit and transform the original dataset to a smaller dimension.

1. 3 shortcomings of PCA –

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

We must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

Principal Components are not as readable and interpretable as original features. Hence, independent variables become less interpretable.