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(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 1 :

Problem Statement : To suggest HELP International an international humanitarian NGO, to decide how to use recent funding money strategically and effectively to fight poverty.

Solution methodology :

1. Cluster the countries based on its factors provided.
2. Perform PCA on the data and continue both K-means and Hierarchical clustering and create clusters.
3. Analyze the clusters and identify the countries which are in dire need of aid.
4. Create visualizations by plotting Principal Components and the clusters.
5. Summary : to provide top 5 countries which are in direst need of aid.

Observations :

1. After loading the original country data set, I removed the outliers (gdpp more than 95 percentile)
2. I decided to have 4 principal components as we see that it explains around 95% of the total variances in the data set.
3. In K -Means elbow curve suggest having 4 clusters for the data set. The Cluster 2 countries need more funding as they are having low - income, gdpp, import, exports, life expectancy and high in child mortality.
4. In Hierarchical clustering also, I decided to have 4 clusters and I got almost similar data in the clusters. The cluster 1 having low - income, gdpp, import, exports, life expectancy and more health.
5. So the countries under cluster 2 in K means were similar and present in cluster 1 in hierarchical clustering.
6. Burundi, Liberia, Congo, Dem. Rep, Niger and Sierra Leone are the countries in dire need of aid.

Question 2

State at least three shortcomings of using Principal Component Analysis.

1. PCA is a better in large number of data.
2. PCA is an unsupervised technique, meaning that the model does not take into account the label of each data point. PCA looks at the data set as a whole and determines the direction of highest variance. Then it determines the next direction of highest variance which is orthogonal to the previous ones, and so on.
3. **Principal component analysis (PCA)** is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components** in such a way that every principal component tries capturing the maximum of remaining variance. The positions of columns play no role in any calculation.

Question 3

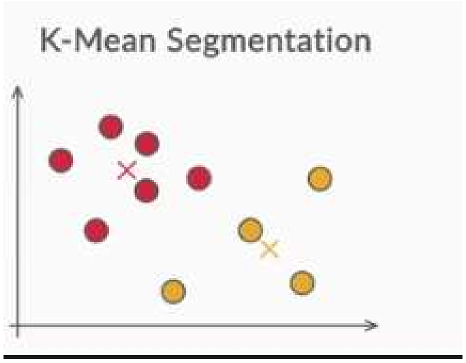
Compare and contrast K-means Clustering and Hierarchical Clustering.

Compare : Both are unsupervised learning methods.

Contrast :

K – Means :

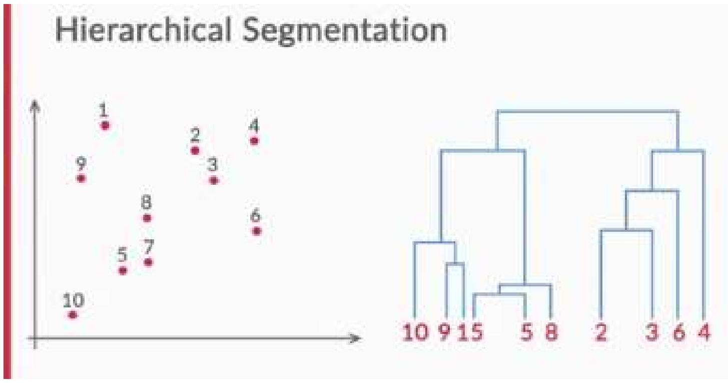
1. For big size of data.
2. Non linear process. We need to mention value of K.
3. It is performed in iterations.
4. Does not need much of RAM as its doing in iterations and not like hierarchical building hierarchy on top of each one.



Hierarchical :

1. For small size of data.
2. Linear method.
3. Computational it needs more RAM. For big size of data we can use cloud system.

4. Hierarchical clustering generally produces better clusters, but is more computationally intensive.



Hierarchical Segmentation