Assignment (Clustering & PCA)

Part 2:

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)

Solution:

Clustering of Countries :

Problem statement: HELP International humanitarian NGO wants to utilize their recently raised funds of around $10 million to help fight poverty and provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities in the most strategic and effective way.

They want the help of data analyst to solve the significant issue of choosing the countries which are in the direst need for their aid.

Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most.

Solution Methodology: In the original dataset, some columns such as imports, exports and health were given as percentage of total GDP, so we converted them to absolute values using GDPP.

Then we performed PCA and found 93% of the information is explained by 4 principle components. We did outlier analysis using all principle components.

Then we went ahead and performed K-means clustering on the principle components dataset. Our Hopkins's score was a good value and therefore our dataset has a good tendency to form clusters. Using Silhouette's score analysis and elbow curve method we found out that k = 3 is a good value to proceed. So we performed the clustering and divided the dataset in 3 clusters and plotted the countries against PC1 and PC2, which gave us distinct clusters.

Then we performed analysis on these clusters by looking into the average values per cluster. So we found out that out focus cluster here is cluster 0 where life_expec was very low, income was very low, where as child_mort, inflation were very high. So we determined the countries which require assistance by filtering all those countries which have all the matrices below average in their cluster.

The Countries which require aid from HELP international are :

1. Burundi
2. Congo, dem.rep.
3. Guinea
4. Malawi

5. Sierra Leone

Now we confirm our analysis using Hierarchical Clustering. We plotted dendrograms using single linkage as well as complete linkage. We divided the data into 3 clusters. After the cluster analysis the list of countries which came out was similar to the one after K-means clustering.

Hence performing K-means and Hierarchical clustering we found out, below are the countries which need aid urgently:

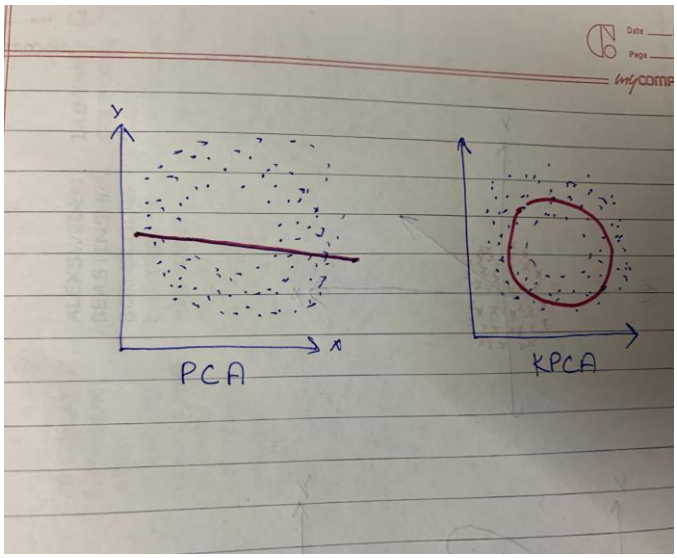
1. Burundi
2. Congo, dem.rep
3. Guinea
4. Malawi
5. Sierra Leone

Question 2

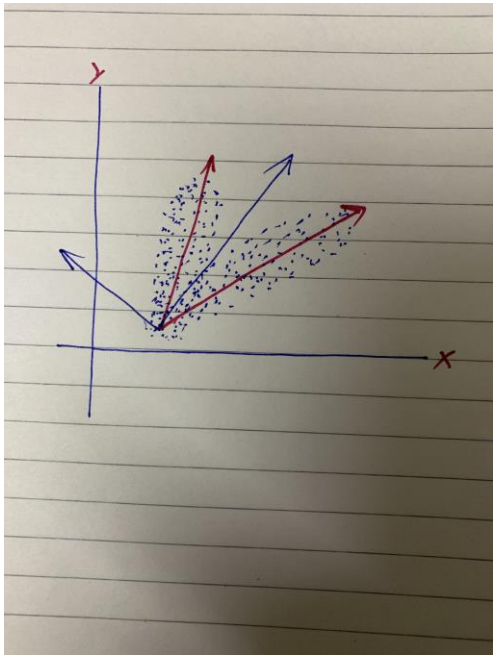
State at least three shortcomings of using Principal Component Analysis.

Answer:

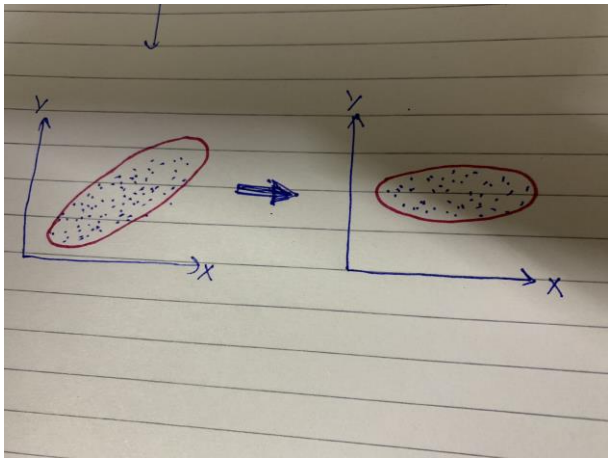
1. PCA finds the hidden linear correlation between the data, and focussed on the projections of data which can be defined linearly. It essentially means that if you have some variables in your dataset which can be represented using the mathematical correlations, the PCA will be able to find direction for the projections. But if a case comes when the data is correlated (other than in linear fashion) like y = x.cos x PCA is not enough alone.



2. PCA can only find orthogonal principal components. Sometimes, our data demands non-orthogonal principal components to represent the data.



3. 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.
4. Before selecting the K dimensions using PCA, one has to apply an affine transformation to uncorrelate the axis.



In the above example you have to pick one dimension out of the two, in the left image, picking either X or Y axis leads to a loss of significant amount of information (Project data on either X and Y, and you should see that it alters the shape and pairwise distance in distributions). One the contrary, in the right image, we have aligned our axis to be uncorrelated.

Assignment (Clustering & PCA)

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

Solution:

In k-means clustering, we try to identify the best way to divide the data into k clusters simultaneously. With k-Means clustering, you need to have an idea ahead about how many desired clusters 'k' value is. IT might often give unintuitive results if your data is not well-separated in the clusters. or if you pick a not suitable value of 'k' for your data. Means picking a very high value for k or picking a value too low will affect your analysis. In the beginning of analysis you have totally random values as your centroids, therefore we will be required to run the algorithm in loop.

In contrast, 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. Hierarchical clustering typically 'joins' nearby points into a cluster, and then successively adds nearby points to the nearest group. You end up with a 'dendrogram', or a sort of connectivity plot. You can use that plot to decide after the fact of how many clusters your data has, by cutting the dendrogram at different heights. Of course, if you need to pre-decide how many clusters you want (based on some sort of business need) you can do that too. Hierarchical clustering can be more computationally expensive but usually produces more intuitive results.