ASSIGNMENT Part-2

Ans 1:
Problem Statement:
Categorising the countries using specified socio-economic and health factors that determine the overall development of the country. Hereby, suggesting the countries which the CEO needs to focus on the most.
Methodology:
1. Data analysis and data cleaning is done.(For this data set there no much null values) . EDA is done and is found that outliers are present (as it removes more number of countries and it isn't necessary as data loss is seen) . Scaling is done due to great difference in values.
2. We can see there is high correlation between some variables, we will use PCA to solve this issue.
3. 4 components are enough to describe 95% of the variance in the dataset.(Observed in the Plot)
4.The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.
a.If the value is between {0.01, ...,0.3}, the data is regularly spaced.
b.If the value is around 0.5, it is random.
c.If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.
As it has high tendency to cluster , Hierarchical clustering is the best in this case.
4.Output: The countries that have low gdpp , income and high child_mort.

Ans 2:
1. Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.

Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be applied.

PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use StandardScaler from Scikit Learn to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning algorithms.

3. Information Loss: 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.

Ans 3:

K-Means Clustering	**Hierarchical**
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.	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.
This is a well known partitioning method. In this objects are classified as belonging to one of K-groups.	This produces a sequence of clusterings in which each clustering is nested into the next clustering in the sequence.
The results of partitioning method is a set of K clusters, each object of data set belonging to one cluster.	Since hierarchical clustering is a greedy search algorithm based on a local search, the merging decision made early in the agglomerative process are not necessarily the right ones.
Euclidean Distance is used for calculating the distance of data point from the particular centroid.	For n samples, agglomerative algorithms begin with n clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes 1 or as specified by the user.
It will often give unintuitive results if (a) your data is not well-separated into sphere-like clusters, (b) you pick a 'k' not well-suited to the shape of your data, i.e. you pick a value too high or too low, or (c) you have weird initial values for your cluster centroids (one strategy is to run a bunch of k-means	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 can be more

algorithms with random starting centroids and take some common clustering result as the final result).	computationally expensive but usually produces more intuitive results.