Clustering & PCA Assignment - Subjective Question

Q 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: In this assignment our primary objective was to divide 167 countries into various clusters based on certain socio-economic and health factors that determine the overall development of the country, so that the financial aids can be provided to the countries in dire need for Aids.

For this assignment, the dataset was provided with 10 features representing socio-economic & health related indicators.
In this assignment, we performed the null/nan value treatment (no null value) and the scaled the dataset with Standard scaler. We also did Outlier treatment which lead to dropping of 4 rows.
After Performing data preparation & preprocessing steps, we created Principal components i.e. converted set of correlated variables into a set of values of linearly uncorrelated variables called **principal components**. After doing Scree Plot we decided to create 4 PCs as it incorporated 96% of the information in data.
After doing PCA, we have performed K-Mean as well as Hierarchical Clustering (Single & Complete Linkage) to divide the countries in 3 clusters. I preferred Hierarchical Clustering as clusters from Hierarchical were with more well defined boundaries and calculation and has less impact due to outliers.
After clustering we evaluated socio-economic & health factors (child_mort, income, gdpp) for countries across clusters to categorise countries as developed and under-developed. We also did binning of countries based on median values to extract 5 countries urgently looking for financial Aids.

Q 2: 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.

Ans:

a) K - Mean Clustering is basically based on dividing n datapoints into k group of clusters also called centroids. here choice of inter cluster center has great effect on final clustering. It also gets seriously impacted due to presence of outliers.
Hierarchical clustering are of two types - top down and bottom-up.
In top-down case, we divide the data into 2 clusters and for each cluster, we can repeat this process, until all the clusters are too similar for further clustering.
In bottom-up hierarchical clustering, we start with each data item as a cluster and look for other datapoint most similar, so that it can be combined to larger cluster. After repeating this step we will get clusters with very high dissimilarity.

c) There are three type of linkages:

1. Simple Linkage : The distance between 2 cluster is defined as the shortest distance between points in the two clusters.
2. Complete Linkage: The distance between 2 cluster is defined as the Maximum distance between points in the two clusters.
3. The distance between 2 cluster 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.

Ans:

a) It is primarily used in dimension reduction in these three fields:

i) Facial Recognition

ii) Computer Vision

iii) Image Compression

c) These are some limitations of PCA:

i) The PCs have to be linear combinations of the original columns.

ii) PCA requires the PCs to be uncorrelated/orthogonal/perpendicular.

iii)PCA assumes low variance components are not very useful