**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.

**Problem Statement:**  To categorise the countries using some socio economic and health factors which are used to determine the overall development of the country, and report the countries which are under developed and require the financial aid.

**Solution Methodology:**

**Data Preparation and Visualisation**: Prepare the dataset (missing values, duplicates and datatypes). Treat the outliers and scale the data.

**Applying PCA:** PCA is applied on the data. And by explained variance ratio the first 4 components are explaining the 95% variance of the data. So, 4 components are chosen.

**KMeans Clustering:** On the PC Data using elbow curve and Silhouette analysis no of clusters K as 3 is chosen and clustering is done on the dataset and then identify the labels. Now the identified labels are merged with the original data and the segmentation with high child mortality, low gdpp and low income is identified as the under developed country segment. Now using filters sort the data to find the lowest under developed countries are found.

**Hierarchical Clustering:** Hierarchical clustering is done both by single linkage and complete linkage. Complete linkage is found to be more effective than single linkage as it is explained by a smaller number of clusters.

**My Choice:** Overall K means clustering and Hierarchical has given the same results for 10 least developed countries but the clustering size is more for Hierarchical (62 countries) than KMeans (47 countries). KMeans has given better results for segmentation of the clusters than Hierarchical.

**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.

**Ans:**

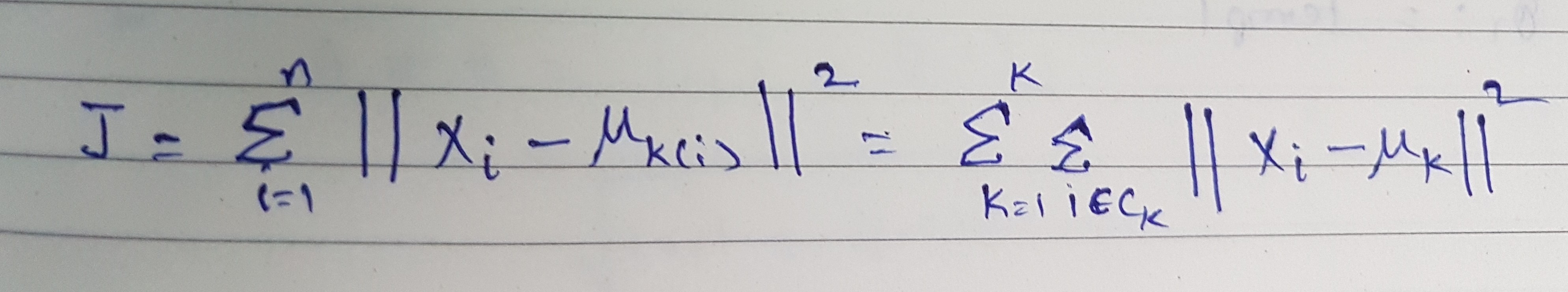
**a) Compare and contrast K-Means and Hierarchical:**

1. In K-Means clustering the no of clusters to be formed K need to be known first. While as in Hierarchical clustering the number of clusters can be found out from the dendrogram.
2. K-Means clusters will not be uniform as we run multiple times because of the random selection of the centroids which is not the case with the Hierarchical clustering.
3. K-Means clustering is found to be better if the shape of the cluster is hyper spherical.
4. With big data K-Means clustering works better as its time complexity is O(n). While as the Hierarchical clustering is not advisable as its time complexity is O(n2)

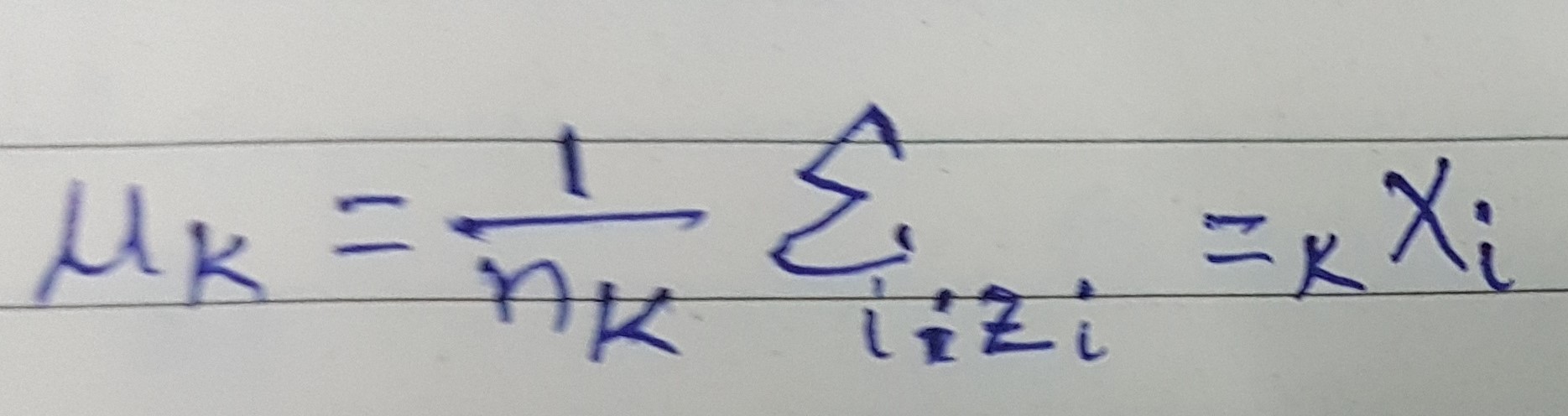
**b) Briefly explain the steps of the K-means clustering algorithm:**

1. The K-Means algorithm uses the concept od centroid to create the clusters.
2. The number of clusters k is defined.
3. K points are taken as centroids. Each observation Xi is associated to the closest cluster µk based on Euclidean distance.
4. Centroid for each cluster is updated and again distance between each observation is calculated with the new centroid.
5. Likewise repeated till the centroid won’t change.

Cost Function of the K Means clustering:



Equation for the optimisation is:



**c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.**

**Statistic Method:**

**Elbow curve:**

1. Compute the algorithm for different values of K i.e. for K from 1 to n clusters.
2. For each value of K, calculate the within distance square of each data point to the centroid of K clusters which is called as WSS
3. Plot the curve of WSS of different K
4. The location of bend in the plot is generally considered as the appropriate number of the clusters.

**Silhouette Method:**

1. Compute the algorithm for different values of K i.e. from 1 to n clusters.
2. For each value of K, calculate the average silhouette of the observations called as avg.sil .
3. Plot the curve for avg.sil for different values of K.
4. The location of the maximum value is considered as the best number of clusters K.

**Business Method:**

A T shirts manufacturing company has to segment its data based on their BMI to manufacture the T-Shirts of different sizes. Here in this scenario with the data they have provided we may get best clusters as 4 or any K by the statistical method elbow or silhouette method. Instead of taking the best K we got from statistical method we still have to segment the data based on the sizes they manufacture like (XS, S, M, L, XL, XXL) 6 segments or 6 clusters.

**d) Explain the necessity for scaling/standardisation before performing Clustering.**

The dataset will be containing features with various measurements like % values, Kg, Millions, cms, lts etc

1. There will be high variance in the data because of the uneven measurement of the features. If PCA is applied on such data set the cluster will tend to get separated with the variables having greater variance. Which means the Principal components will be biased towards features with high variance, leading to false results.
2. Also, PCA can be applied on numerical columns only so the categorical variables to be converted in to the numerical variables and then should be standardised to get the better results.
3. PCA is affected by scale, so we need to scale the data before applying PCA. Standard scaler can be used to standardize the data for optimal performance of PCA.

**e) Explain the different linkages used in Hierarchical Clustering.**

There are three types of linkages:

1.Single Linkage: The distance between 2 clusters is defined as the shortest distance between the two points in the clusters. For two clusters A and B, between i and j, I from cluster A and j from cluster B is calculated as:

L(A,B)=min(Dist(i,j)) , i ϵ A and j ϵ B



2.Complete Linkage: The distance between 2 clusters is defines as the maximum distance between any two points in the clusters.

L(A,B)=max(Dist(i,j)) , i ϵ A and j ϵ B



3.Average Linkage: The distance between the 2 clusters is defined as the average distance between every point of one cluster to every point of other cluster.

L(A,B)= ∑inA=1 ∑jnB=1 D(i,j) i ϵ A and j ϵ B





**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.

a) Give at least three applications of using PCA.

PCA is mainly used for dimensionality reduction and is used for the following:

1. Facial Recognition.
2. Image Compression.
3. Data Mining.
4. Bio Informatics.

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

**Basis Transformation**: Since all features will not be the same and we have various dimensions we can reduce the dimensionality by changing its basis.

If we have one dimension and to represent that into the new representation all we have to do is multiply the old basis with the multiplying factor M to change that into the new basis.

Equation being:

New Basis Representation(B1) = M(multiplying factor) \* Old Basis Representation(B2)

B1=M\*B2

**Variance as Information:**

The more variance a feature has the more information we get from that feature.

The **total variance** is the sum of variances of all individual principal components.

The fraction of **variance explained** by a principal component is the ratio between the variance of that principal component and the total variance.

For several principal components, add up their variances and divide by the total variance.

Generally, the first k principal components (where k can be 1, 2, 3 etc.) explain the most variance any k variables can explain, and the last k variables explain the least variance any k variables can explain, under some general restrictions. (The restrictions ensure, for example, that we cannot adjust a variable’s explained variance simply by scaling it.)

**c) State at least three shortcomings of using Principal Component Analysis.**

1. **Variable Interpretation:** Independent variables becoming less interpretable. After implementing the PCA on the dataset, the original features will turn into combination of principal components. Therefore, there are not interpretable and readable like the original features in the dataset.
2. **Linearity:** PCA assumes that the principle components are a linear combination of the original features. If this is not true, it won’t give sensible results. In other words, PCA works PCA is limited to linearity.
3. **Orthogonality:** PCA assumes that the principal components are orthogonal. PCA needs the components to be perpendicular, though in some case that might not be the best solution.
4. **Variance:** PCA uses variance as the measure of how important a particular dimension is. So, high variance axes are treated as principal components, while low variance axes are treated as noise. (Which might not be true in prediction setups like classification of dataset which has high imbalance.)