Clustering of Countries

Question 1: Assignment summary

This assignment involved the problem statement where we need to cluster countries based on 3 important variables **gdpp**, **income**, and **child mortality**.
The data we got had no nulls or NAs or in other words, was in no need of cleansing. But when trying to analyze the spread of data for each variable I got to know about the outliers it had. Worked with removing outliers which didn't change the result significantly on the data I was working on. But if you are deleting entire outliers it would lead to losing 25% of the data. So, I left it as it is since clustering would take care of the outliers.
After data was ready PCA was used to reduce dimensionality in the data and scree plot was used to find out the variability that each component would explain based on which I got to choose 5 principal components, which would explain variability in most of the data.
Clustering: Both k-means and Hierarchical clustering were used to see which would offer a better result.
For K-means with the help of the silhouette score and elbow curve, I selected 4 clusters. And Hierarchical clustering looking at Dendrogram helped me to identify clusters and it made sense to choose 4 clusters.
When looking at 2 types of clustering both seem to produce a similar result

Question2:
a) The most important difference is the **hierarchy**. Two different approaches fall under this name: top-down and bottom-up.
In top-down hierarchical clustering, we divide the data into 2 clusters (using k-means with k=2, for example). Then, for each cluster, we can repeat this process, until all the clusters are too small or too similar for further clustering to make sense, or until we reach a preset number of clusters.
In bottom-up hierarchical clustering, we start with each data item having its cluster. We then look for the two most similar items and combine them in a larger cluster. We keep repeating until all the clusters we have left are too dissimilar to be gathered together, or until we reach a preset number of clusters.
In **k-means** clustering, we try to identify the best way to divide the data into kk sets simultaneously. A good approach is to take kk items from the data set as initial cluster representatives, assign all items to the cluster whose representative is closest, and then calculate the cluster mean as the new representative until it converges (all clusters stay the same).

b) The first step of this algorithm is creating, among our unlabeled observations, *c* new observations, randomly located, called 'centroids'. The number of centroids will be representative of the number of output classes (which, remember, we do not know). Now, an iterative process will start, made of two steps:

- First, for each centroid, the algorithm finds the nearest points (in terms of distance that is usually computed as Euclidean distance) to that centroid, and assigns them to its category;

- Second, for each category (represented by one centroid), the algorithm computes the average of all the points which have been attributed to that class. The output of this computation will be the new centroid for that class.

Every time the process is reiterated, some observations, initially classified together with one centroid, might be redirected to another one. Furthermore, after several reiterations, the change in centroids' location should be less and less important since the initial random centroids are converging to the real ones. This process ends when there is no more change in the centroids' position.
The number of centroids, in this case, is chosen by elbow curve and silhouette score

c) Statistical aspect:

Elbow method

The Elbow method looks at the total intra-cluster variation as a function of the number of clusters: One should choose some clusters so that adding another cluster doesn't improve much better the total intra-cluster variation (WSS).
The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Average silhouette method

That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.
The average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k.
The algorithm is similar to the elbow method and can be computed as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (*avg.sil*).
3. Plot the curve of *avg.sil* according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters.

Business aspect:

This comes purely from the business knowledge and experience in a particular domain. In some cases even if the above statistical methods tell us to go with (say 3) clusters we need to select 4 because of the business aspect of that particular domain.

d)

Clustering on the non-normalized data fails. Clustering on the normalized data works very well. A good example is the geolocation data (longitudes and latitudes). If you were seeking to cluster towns, you wouldn't need to scale and center their locations.
For data that are of different physical measurements or units, its probably a good idea to scale and center. For example, when clustering vehicles, the data may contain attributes such as the

number of wheels, number of doors, miles per gallon, horsepower, etc. In this case, it may be a better idea to scale and center since you are unsure of the relationship between each attribute. The intuition behind that is that since many clustering algorithms require some definition of distance if you do not scale and center your data, you may give attributes that have larger magnitudes more importance.

e) **Single Linkage**
In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.

Complete Linkage
In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.

Average Linkage
In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters "r" and "s" to the left is equal to the average length each arrow between connecting the points of one cluster to the other.

4) a) Application of PCA:
1) In quantitative finance, principal component analysis can be directly applied to the risk management of interest rate derivative portfolios.
2) A variant of principal components analysis is used in neuroscience to identify the specific properties of a stimulus that increase a neuron's probability of generating an action potential.
3) Used in reducing dimensionality in image recognition which can be used in facial recognition.

b) **Basis of Transformation**
Essentially, 'basis' is a unit in which we express the vectors of a matrix.
For example, we describe the weight of an object in terms of the kilogram, gram and so on; to describe length, we use a meter, centimeter, etc. So for example, when you say that an object has a length of 23 cm, what you are essentially saying is that the object's length is 23×1 cm. Here, 1 cm is the unit in which you are expressing the length of the object.
So Patient 1's information in the cm/kg space is given by 165·[10]+55·[01] whereas in the ft/lbs space is given by 5.4·[10]+121.3·[01]

Now, 1 ft = 30.48 cm and 1 cm = 0.033 ft
Similarly, 1 kg = 2.205 lbs and 1lbs = 0.454 kg.

Therefore, comparing the basis vectors, we can say
[1ft0lbs]in ft/lbs space = **[30.48cm0kg]** in cm/kg space and
[0ft1lbs] in ft/lbs space = **[0cm0.45kg]**

[16555] = **165[10]** + **55[01]** = **5.4[30.480]** + **121.3[00.45]**in the cm/kg space.

In the above case, we considered the new basis vectors as **[30.480]** and **[00.45]** in the cm/kg space which is equivalent to (1,0) and (0,1) in the ft/lbs space. And using this, we got the representation of **[5.4121.3]** for the patient.

Therefore, we can choose a completely different set of vectors, say **v1** and **v2** as the basis vectors and find the representation of Patient 1 (originally in the standard basis vectors) in the new basis system. They should be satisfying the following linear combination equation
[16555]= a1·v1+a2·v2
where (a1,a2) is the representation of Patient 1 in the v1 and v2 space.

Taking v1=[30.480] and v2=[00.45] we got a1=5.4 and a2=121.3
Similarly, taking v1=[550] and v2=[055] we get a1=3 and a2=1
Again, taking v1=[31] and v2=[20] we get a1=55 and a2=0
and so on...

b) **The variance of Information:**
In the case of PCA, "variance" means *summative variance* or *multivariate variability* or *overall variability* or *total variability*. Below is the covariance matrix of some 3 variables. Their variances are on the diagonal, and the sum of the 3 values (3.448) is the overall variability.

1.343730519 -.160152268 .186470243
-.160152268 .619205620 -.126684273
.186470243 -.126684273 1.485549631

Now, PCA replaces original variables with new variables, called principal components, which are orthogonal (i.e. they have zero covariations) and have variances (called eigenvalues) in decreasing order. So, the covariance matrix between the principal components extracted from the above data is this:

1.651354285 .000000000 .000000000
.000000000 1.220288343 .000000000
.000000000 .000000000 .576843142

Note that the diagonal sum is still 3.448, which says that all 3 components account for all the multivariate variability. The 1st principal component accounts for or "explains" 1.651/3.448 = 47.9% of the overall variability; the 2nd one explains 1.220/3.448 = 35.4% of it; the 3rd one explains .577/3.448 = 16.7% of it.
So, what do they mean when they say that "*PCA maximizes variance*" or "*PCA explains maximal variance*"? That is not, of course, that it finds the largest variance among three values 1.343730519 .619205620 1.485549631, no. PCA finds, in the data space,
the *dimension* (direction) with the largest variance *out of the overall* variance 1.343730519+.619205620+1.485549631 = 3.448. That largest variance would be 1.651354285. Then it finds the dimension of the second largest variance, orthogonal to the first one, out of the remaining 3.448-1.651354285 overall variance. That 2nd dimension would be 1.220288343 variances. And so on. The last remaining dimension is .576843142 variance.

c) Shortcomings of PCA:

- **Linearity**: PCA assumes that the principal components are a linear combination of the original features. If this is not true, PCA will not give you sensible results.
- **Large variance implies more structure**: 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.
- **Orthogonality**: PCA assumes that the principal components are orthogonal.