Clustering & PCA Assignment

**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)**

Our problem statement was to categorise the countries using some socio-economic and health factors that determine the overall development of the country and then suggest the organisation CEO which segment of countries needs to be targeted first. Since the budget is limited we can concentrate only on a specific cluster of countries,

We tried and applied both KMeans and Hierarchical Clustering algorithm to the data set after performing outliers capping and PCA into 4 cluster.

**Outlier Treatment:**

We choose, capping the outliers instead of removing it as we have only 167 records, by removing outliers we may lose few data. Also we may remove the countries, which are in dire need of aid.

On the first attempt which ever values were below 2 percentile and above the 95 percentile, we replaced all values to the 2% percentile and the 95 percentile for all and 2-85 percentile for “gdpp” column.

**PCA Step:**

After outlier treatment and scaling of the dataset, we found out the explained\_variance values of each vectors. Now, we can notice in order to explain 95% of the data we needed to take up to 5 components indicated in the cumulative sum Scree plot.

**KMeans:**

Based on the Elbow Curve and silhouette score we decided to make the no of clusters (k) to be 4.

**Hierarchical:**

The dendrogram using complete linkage, if cut around y=6.9, we get clearly 3 clusters.

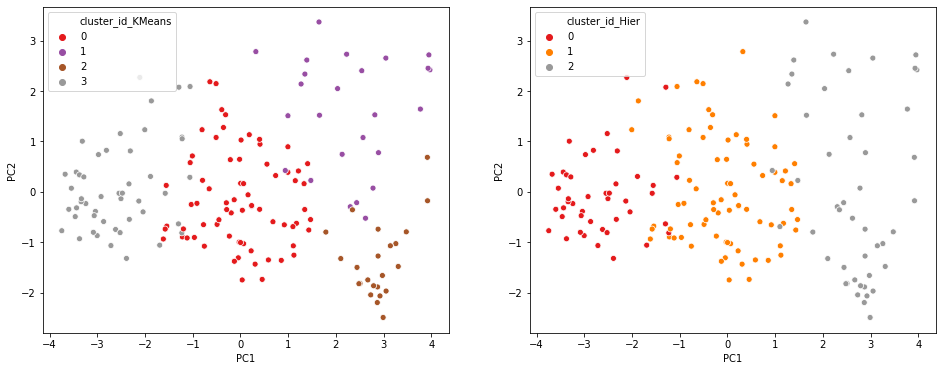
As per the scatter plots, we noticed hierarchical clustering segmented the features clearly than KMeans. Although our result is more or less similar in both clustering methods.

**Question 2: Clustering**

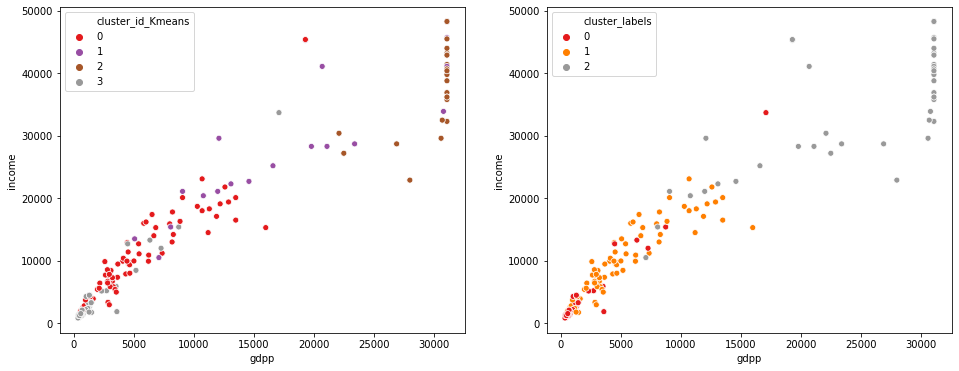
1. **Compare and contrast K-means Clustering and Hierarchical Clustering.**

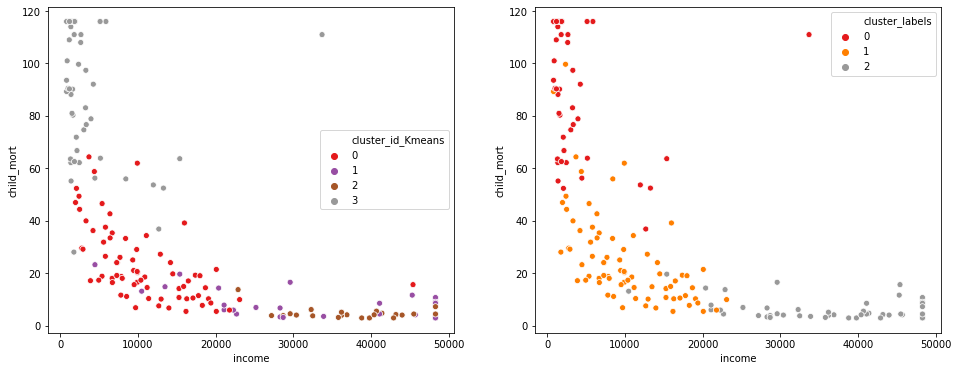
For comparison between K-Means and Hierarchical Clustering, the below graph is plotted which has Principal Components 1 and 2 in x and y axis respectively.

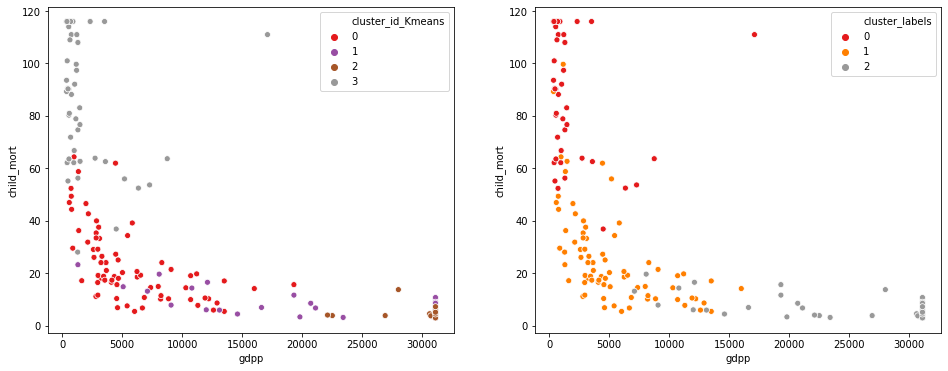
Here PC1 and PC2 explains around 66 % of the data and we notice both algorithm K-Means on the left and Hierarchical on the right could identify a unique are of 4 and 3 clusters respectively.



As per our requirements and data when focussed on these three variables - [**gdpp**, **child\_mort** and **income**] we can notice the following plots.







It’s clear that Hierarchical clusters are more prominent and separable.

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

The algorithm for K-means algorithm is as follows:

* **Select initial centroids**. The input regarding the number of centroids should be given by the user.
* Assign the data points to the closest centroid
* Recalculate the centroid for each cluster and assign the data objects again
* Follow the same procedure until convergence.

Convergence is achieved when there is no more assignment of data objects from one cluster to another, or when there is no change in the centroid of clusters.

Well in Python most of the job is done using

kmeans = KMeans(n\_clusters= 3, max\_iter = 50)

kmeans.fit(df)

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

For K-Means clustering we need to initially have to determine the value of ‘k’. So to find the optimal here are two approaches

**There are two ways to find the optimal value of k.**

* Elbow Curve or Sum of Squared Distance Curve
* Silhouette Analysis

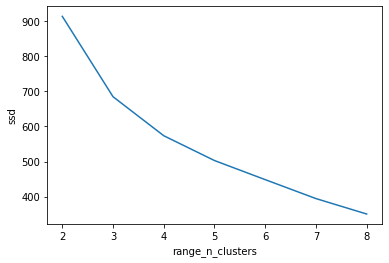
In both of the case we take all possible k –values like from 2 to 8.

**Elbow Curve or Sum of Squared Distance Curve**

Now to plot the elbow curve we try to perform K-Means modelling and try to fit the scaled data frame using all the possible values k-values one by one.

Then we find the sum of squared distances of samples to their closest cluster centres, using inertia\_ attribute for each value of k and plot the same.

In our case, we plotted the below elbow curve we noticed at k = 3 the slope is quite significant, beyond that there is a gradual decrease in the slope forming an edge, which gives us the optimal value of k.



**Silhouette Analysis**

Silhouette Score =

p is the mean distance to the points in the nearest cluster that the data point is not a part of

q is the mean intra-cluster distance to all the points in its own cluster.

* The value of the silhouette score range lies between -1 to 1.
* A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
* A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

In order to calculate the silhouette score we need to import silhouette score as below

from sklearn.metrics import silhouette\_score

Similarly like SSD we can calculate the silhouette score for each value of k and once we get all the values. Now, we consider the maximum of the score and business understanding to get the optimal value of k

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

Scaling /Standardisation is a very import step before we perform clustering or any predictive modelling as we are dependent on the significance of the each features. If the features are in a different scale all together, our output might be aligned to the higher variance side.

The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space. In this situation leaving variances unequal is equivalent to putting more weight on variables with smaller variance

In order to keep all the features in a similar scale we perform scaling or standardisation.

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

Based on the representative of the distance between 2 clusters there are 3 different types of linkages.

* **Single Linkage**: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
* **Complete Linkage**: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
* **Average Linkage**: Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

**Question 3: Principal Component Analysis**

1. **Give at least three applications of using PCA.**

There are a lot of applications of using PCA, here are the top 3 applications

1. **Dimensionality Reduction:** If we can represent the original dataset into fewer columns and that are non-correlated. This also helps in Data Visualization.
2. **EDA & Data Visualisation:** It is not possible to visualise more than two variables at the same time using any 2-D plot. Therefore, finding relationships between the observations in a data set having several variables through visualisation is quite difficult.
3. **Building Predictive Models:** Having a lot of correlated features lead to the multicollinearity problem. Iteratively removing features using VIF and p-value is time-consuming and also leads to some information loss.

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

**Basis transformation:**

Basis is essentially the fundamental units in which you express your data. In vectors and vector spaces, we use basis vectors to represent the points in space. *Basis is as a unit of representation, different basis vectors can be used to represent the same observations, just like you can represent the weight of a person in kilograms or pounds value may be different but actual weight remains same.*

Now, transformation of Basis can occur from 1 basis to another basis, or standard basis to non-standard basis or vice versa.

If*B1 is a set of Basis vectors and B2 is another set of basis vectors and you want to move from B1 to B2 then our equation for****M****would come out to be*

B1=M.B2

**Certain rules to keep in mind**

*1.) If you're moving from a basis space B to the standard basis, then the* ***change of basis*** *matrix M is the same as the basis vectors of B written as its column vectors. Therefore, if there is a vector v represented in B and you want to find its representation in the standard basis, then you'd have to perform*

M.v

*2.) If you want to go the other way around, where you have v represented in the standard basis and want to find its representation in B you multiply it by its inverse*

M−1.v

*3.) Finally, if you want to find the change of basis matrix M where you move from two non-standard basis vectors - say from B1 to B2 then you can get that by calculating this value -*

B2−1. B1

*Note that in all the above cases, the basis vectors should be represented in the same units.*

**Variance as information:**

The importance of a column or a feature is measured by checking its variance values. If a column has more variance, then this column will contain more information. There could be some columns having much less variance than others, it is easier to remove those columns and do dimensionality reduction.

For example: If we have two columns age and height in a dataset such as

|  |  |
| --- | --- |
| Age | Height |
| 18 | 163 |
| 18 | 155 |
| 18 | 147 |
| 18 | 180 |

We can clearly notice that Age column has no variance all rows are same and could be removed for dimensionality reduction.

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

PCA: Principal Component Analysis certainly has some shortcomings and they are

* PCA is limited to linearity, though we can use non-linear techniques such as t-SNE.
* PCA needs the components to be perpendicular, though in some cases, that may not be the best solution. The alternative technique is to use Independent Components Analysis.
* PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problem with a high class imbalance).