upGrad

International Institute of Information Technology Bangalore

Clustering and PCA Assignment – Subjective Questions

Author: Anish Mahapatra
E-mail id: anishmahapatra01@gmail.com

Q1: 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).

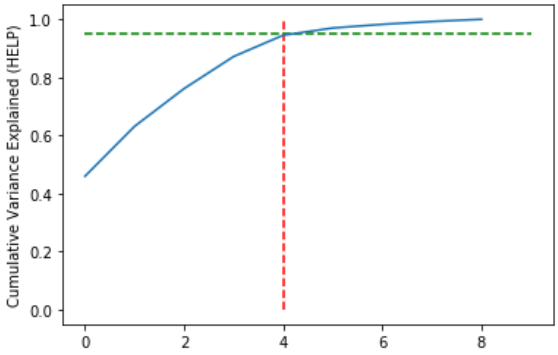
A.

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

The factors on which the analysis was carried out to cluster the countries include the following features:
Child Mortality Rate, Exports, Health, Income, Inflation, Life Expectancy, Total Fertility, GDP

The data used was clean and there were no missing values. The data has considerable number of outliers that were found during outlier analysis, but, they were not removed as it was expected that clustering would put them in another cluster.

The data was normalized to get all the features on the same scale, post which a method called PCA was used to capture 95% of the variance. For this, the **scree plot** as shown below was used.



Heatmap of the Principal Components

This plot indicated that 95% of the variance can be captured with the help of **5** Principal components (0 to 4).

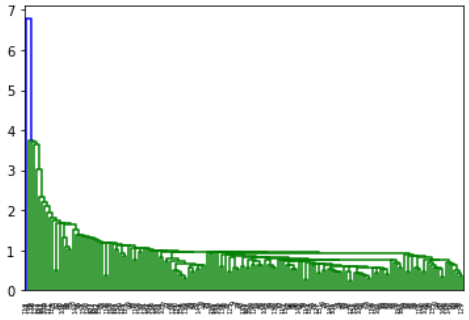
So, we have chosen **5 principal components.** To assess if the new dataset with 5 principal components tends to cluster, **Hopkin's Statistics** was leveraged and a score between 0.75 to 0.9 was achieved. The closer the score to 1, the higher is the tendency to cluster.

For clustering, there are two main methods that were leveraged to find the number of clusters.

1. Hierarchical Clustering
2. K-Means Clustering

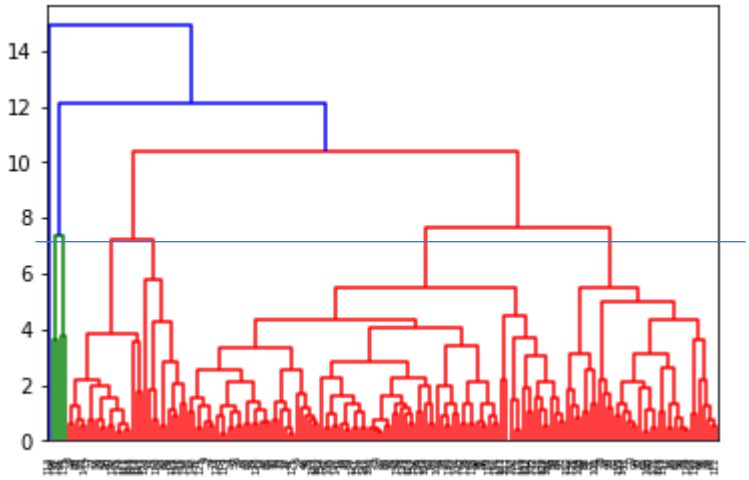
1**. Hierarchical Clustering**

Hierarchical Clustering using *Single Linkage*:



Nothing much can be interpreted from this.

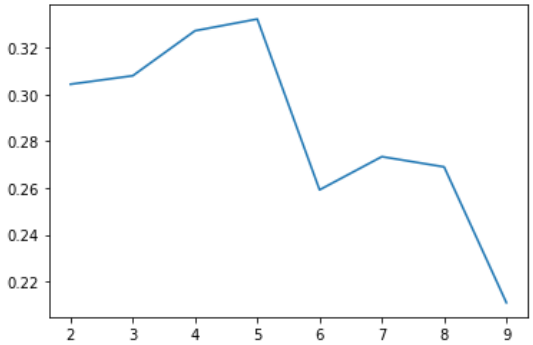
Hierarchical Clustering using *Complete Linkage*



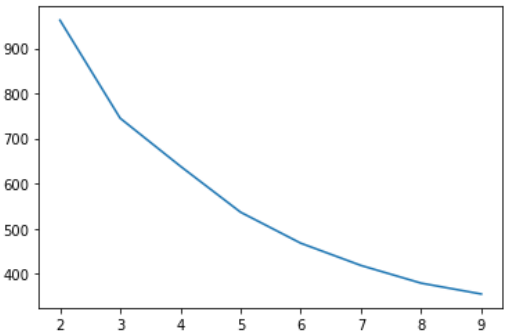
From this, we can interpret that the number of cluster can be between 4 and 6. As the green-coloured ones are very close, they can be clustered as a single cluster.

2. **K-Means Clustering:**

Silhouette Score Plot:

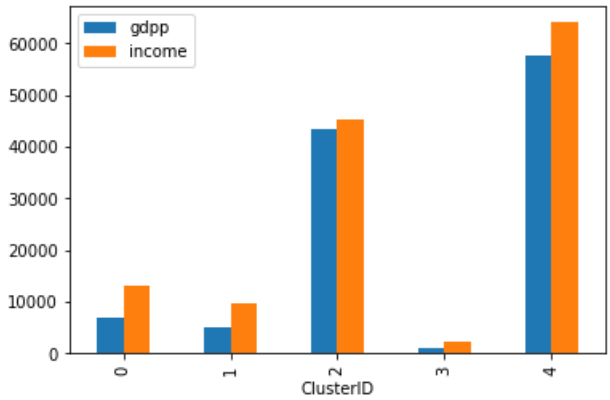


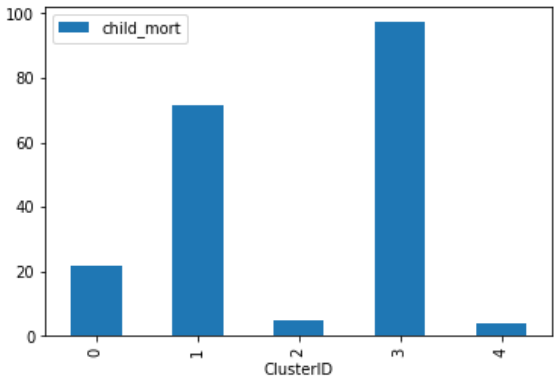
Elbow Curve:



Based on the analysis, we have chosen 5 clusters to be made.

Let us now analyse the GDP, income and mortality rates of the chosen clusters from 0 to 4 (5 clusters)





The GDP of Cluster 3 is the lowest and the mortality rate (average) of Cluster 3 is the highest, hence, indicating that it houses the countries that are in the direst need for monetary relief.

Let us now observe the distribution of the GDP, income and mortality rates across the clusters.

Cluster 3 has the highest mortality rate, lowest income and GDP.

Find below the observations cluster-wise:

1. Cluster 0: The child mortality rate is low, the income and GDP are barely above Cluster 1
2. Cluster 1: The child mortality rate is alarmingly high; The GDP and income are the second-lowest
3. Cluster 2: The child mortality rate is quite low, the income and GDP are the second highest
4. Cluster 3: The child mortality rate is the highest (very bad), the income and GDP are the lowest of all the clusters.
5. Cluster 4: The child mortality rate is the lowest, the income and the GDP are the highest of all the clusters

Cluster-3 > Cluster-1 > Cluster-0 > Cluster-2 > Cluster-4

Q2: 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.

A.

a) **K-Means versus Hierarchical Clustering**

- K-means is a top to bottom approach to clustering, where as Hierarchical is a bottom- up approach to clustering
- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is much lesser than that of Hierarchical Clustering
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical.
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

b) **Steps of K Means Algorithm:**

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centres.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centres.
4) Recalculate the new cluster center using:

$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$ where, 'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centres.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

c) **How is k chosen?**

In the K-means algorithm, there are the **Elbow Method** and **Silhouette method**:

Elbow Method:

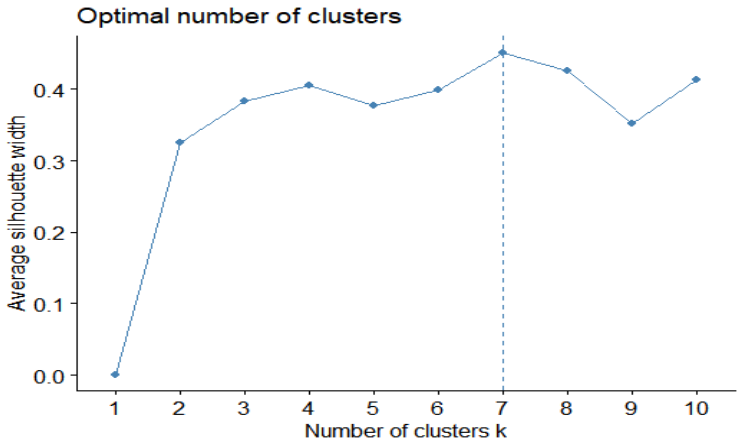
First, compute the sum of squared error (SSE) for some values of k. The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid. Mathematically:

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

If you plot k against the SSE, you will see that the error decreases ask gets larger; this is because when the number of clusters increases, they should be smaller, so distortion is also smaller. The idea of the elbow method is to choose the k at which the SSE decreases abruptly. This produces an "elbow effect" in the graph.

Silhouette method:

This method helps in measuring the quality of the clustering i.e. how well an object lies within its cluster. A good silhouette width indicates good clustering. Also, it shows you the optimal number of clusters to be used.



This was in terms of statistics. For the business approach to K-means, it is critical that the clusters chosen make business sense. There cannot be too many clusters or less number of clusters as the clusters must be leveraged by the business to make decisions on that cluster group.

If there is sufficient computing power, **Hierarchical Clustering** can be used as well, even on larger data sets.

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

Yes, it is necessary to scale/ standardize before performing clustering.

- If there are features that are on different scales, then the variation of the smaller-scaled features does not get captured and they may get ignored when PCA is done on the dataset.
- Also, K-means clustering is "isotropic" in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. In this situation leaving variances unequal is equivalent to putting more weight on variables with smaller variance.
- For larger datasets, standardizing the data will make the algorithm more efficient.
- To avoid this dependence on the choice of measurement units, one has the option of standardizing the data. This converts the original measurements to unitless variables.
- Standardizing data is recommended because otherwise the range of values in each feature will act as a weight when determining how to cluster data, which is typically undesired.

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

- In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest maximum pairwise distance).
- In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest minimum pairwise distance).

- Complete-link clustering can also be described using the concept of clique. Let dn be the diameter of the cluster created in step n of complete-link clustering. Define graph G(n) as the graph that links all data points with a distance of at most dn. Then the clusters after step n are the cliques of G(n). This motivates the term complete-link clustering.
- Single-link clustering can also be described in graph theoretical terms. If dn is the distance of the two clusters merged in step n, and G(n) is the graph that links all data points with a distance of at most dn, then the clusters after step n are the connected components of G(n). A single-link clustering also closely corresponds to a weighted graph's

Q3. 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.

a**) 3 applications of using PCA**

- PCA is used in image-processing to perform dimensionality reduction on extremely large sparse matrices to capture relevant information
- Data Compression: PCA is used to reduce multicollinearity in the data. PCA aims to orthogonally transform correlated variables to a smaller set of linearly uncorrelated variables (principal components). Hence, this will reduce the multicollinearity.
- Used to speed up algorithms as it captures the essence of the data in a way that makes the algorithm more efficient when running models on the Principal Components. This, in return produces better results while saving time

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

The basis is the set of direction or the relative reference point to explain the weight of the data. Whilst performing PCA, essentially, the basis of the data is changed to capture maximum variance in the data.

The way PCA works is mentioned below:

- Represent all the information in the dataset as a covariance matrix
- Eigen Decomposition in covariance matrix
- **New basis** is the Eigen Vectors of covariance based on Step I
- Represent the data in new basis. This new basis is essentially the Principal Component.

In case of PCA, "variance" means *summative variance* or *multivariate variability* or *overall variability* or *total variability*. 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. "*PCA maximizes variance.*

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

Following are the limitations of PCA:

- PCA relies on linear assumptions. If the data is not linearly correlated, then, PCA is insufficient
- Mean and covariance doesn't describe some distributions. There are many statistics distributions in which mean and covariance do not give relevant informatio.
- 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. Standardizing data sometimes leads to information loss
- 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.