Clustering & PCA Part-II

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)

Answer:
Problem Statement:

HELP International is an international humanitarian NGO has been able to raise around $ 10 million. The objective is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest about the countries which the CEO needs to focus on the most.

As a data analyst, our job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries list that are in the direst need of aid to the CEO who needs to focus on the most.

Solution Methodology:

1. Data Collection and Cleaning:
   - Importing all required libraries.
   - Reading and understanding the given Country data.
   - Sanity checks on the data.
   - Checking for duplicate data and removing the duplicates if found any.
   - Checking for Spelling mistakes in the data.
2. Outlier Analysis and Removals
   - Detecting Outliers in variables like gdpp and income. Capping Those outliers and making the data stable.
3. Visualizing the Data
   - Plotting Correlation matrix and heat map to visualize the data and identify the variables with high correlation for targeting as main factors of study.
4. Scaling the Data
   - Continuous data are scaled by using StandardScaler() object of Standardization technique.
5. Principal Component Analysis on the Data
   The principal component analysis is used in deriving the components from original data. There were total 5 PCs for 5 variables. We derived 3 PCs from the country data set using the variance ratio. Scree plot is used to find the optimal Principal Components. It is found that the first 3 components explain about 95% of the variance in the data. Then PCA is performed using Incremental PCA technique with 3 components. The Correlation is again checked and is found to be approximately 0 which is a good indicator.

6. Hopkins Statistics

   The Hopkins statistics is performed on PCA dataset. The obtained score is greater than 0.5. hence, the given dataset has a good tendency to form clusters.

7. K means Clustering

   - The number of clusters is determined by using Silhouette Analysis and Sum of Squared distances.
   - The Silhouette Score has the highest peak at 3 and the sum of squared distances (Elbow Curve) ranges from 2 to 3.
   - Hence the optimal number of clusters comes out to be 3. The data is grouped into 3 clusters.
   - The variables are analyzed according to cluster id using bar plot and Scatterplots. From this data the countries are sorted according to low gdpp, high Child mortality value and low income.
   - These countries are then concluded as the countries that need direst aid from the government.

8. Hierarchical Clustering

   - This technique uses Dendrogram for determining the clusters. Single and complete linkages are performed.
   - Single linkage does not show any clear view and hence we can go forward with complete linkage.
   - The dendrogram is cut at 3 cluster value which is arbitrary. Hence 3 clusters are created.
   - These 3 clusters are analyzed using bar plots, boxplots and scatterplots.
   - The countries at the final are sorted using the same 3 variables and the final list of countries is generated which needs direct help from the government.

9. Evaluating and Decision Making

   Finally, the countries obtained from both the clustering techniques are reported. These are the countries which are in severe concern and thus needs aid from the government.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

i. K-means Clustering is easy and simple for understanding. In Hierarchical clustering, single linkage may not show clear view and Complete linkage may appear complex sometimes.
ii. K-means clustering generates several flat clusters. Hierarchical Clustering generates a hierarchy of clusters.
iii. The Performance of K-means clustering is better than hierarchical clustering.
iv. K-means is very sensitive to Outliers whereas Hierarchical Clustering is less sensitive to Outliers.
v. K-means always require pre-defined number of clusters. Hierarchical clustering doesn't require any input for number of clusters.
vi. K-Means requires more time to execution, Hierarchical clustering is executed in less time.
vii. K-means is useful for large datasets, Hierarchical clustering is useful for small data sets.
viii. K-means clustering shows less quality, Hierarchical clustering shows high quality.

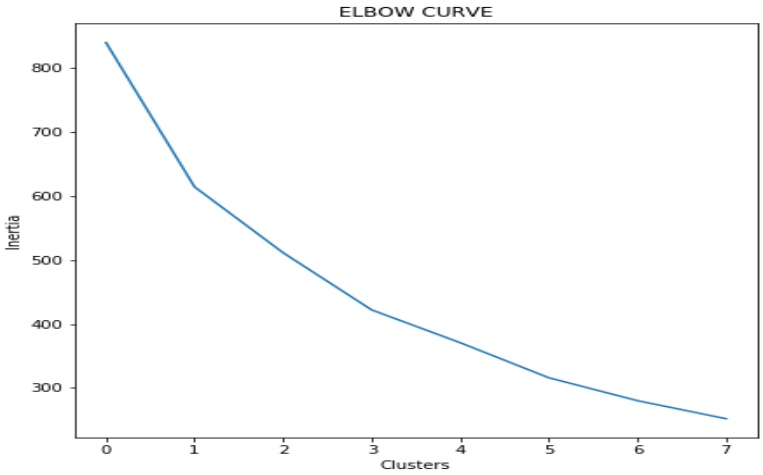
b) Briefly explain the steps of the K-means clustering algorithm.

1) Initialization
K-means randomly choose K (data points) from the dataset as initial centroids of a cluster.
2) Distance Metric Calculation.
Then, all the data points that are the closest to a centroid will create a cluster. Calculate the distance from the data point to each cluster.
This is mostly done using Euclidean Distance. Choose the minimum of those distances from each point to the random centroid.
3) Cluster Assigning
Choose the minimum of those distances from each point to the random centroid. Assign the data point to the respective cluster. We have new clusters that will need new centers. We need to update the centroids.
4) Optimization
A centroid's new value will be the mean of all the data points in a cluster. The centroid is updated with the mean value. This process is called Optimization.
5) Repeat steps 3 and 4 until the cluster centers get converge (same centers).

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

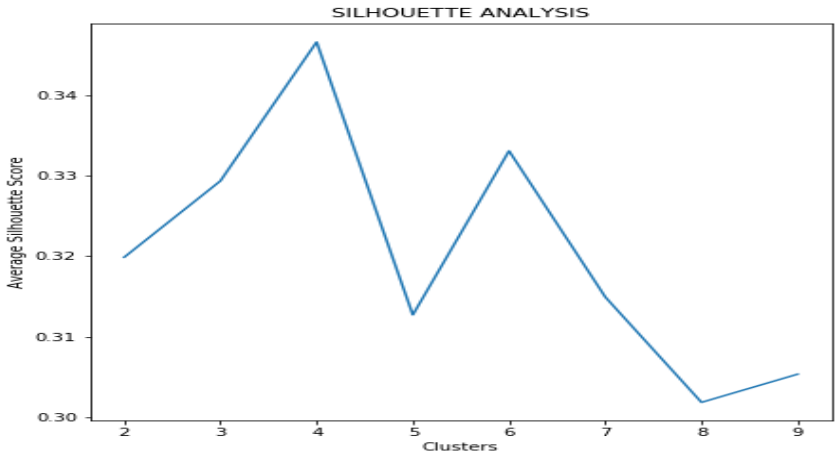
1. ELBOW METHOD:

- The Elbow Method is used to determine optimal value of k. Compute the sum of squared error (SSE) for some values of k.
- If you plot k against the SSE, you will see that the error decreases as, k gets larger.
- 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.
- To determine the optimal number of clusters, we have to select the value of k at the "elbow" i.e. the point after which the inertia starts decreasing in a linear fashion.



2. SILHOUETTE ANALYSIS:

- 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 us that the optimal number of clusters to be used.
- Silhouette values +1 indicates that the sample is far away from the neighboring clusters.
- A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters.
- Negative values indicate that those samples might have been assigned to the wrong cluster.



Business aspect of determining K:

- Monitoring or visualizing the distribution of data points across groups is a useful and easy way to gain insight into how the algorithm is splitting the data for each K.
- K value can be decided using some business domain. Likewise, 2 clusters can hardly explain any data properly; hence we need more than 2 clusters. Also, the value of K depends on the data provided.
- A Movie related data can be grouped into Horror, Comedy, Thriller types of clusters.
- The data related to items in a grocery store can be clustered according to their prices, quantity.
- Hence these factors can be used in addition to determine the value of K.

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

- 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 not desired.
- The idea is that, if different components of data have different scales, then derivatives tend to align along the directions with higher variance, which leads to the poor convergence.
- Groups are defined based on the distances between them.
- Standardization helps to make the relative weight of each variable equal by converting each variable to a unit less measure.
- Scaling eliminates redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering.

e) Explain the different linkages used in Hierarchical Clustering.

1. **SINGLE LINKAGE**: (Nearest Neighbor)

   In Single linkage hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance. Most of the time Single linkage does not give a clear view.

   A concern of using single linkage is that it can sometimes produce several clusters that may be joined together simply because one of their cases is within closest proximity of case from a separate cluster.

2. **COMPLETE LINKAGE**: (Furthest Neighbor)

   In complete linkage hierarchical clustering the two clusters are merged whose two closest members have the maximum distance.

3. **AVERAGE LINKAGE:**

   In Average linkage hierarchical clustering, we merge in each step the two clusters whose two closest members have the average distance.

   In this type linkage every point of first cluster are merged with every other point in the other cluster.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

i. **DIMENSIONALITY REDUCTION:**

PCA allows you to reduce the dimensionality of your data so most of the variation that exists in our data across many high dimensions is captured in fewer dimensions. PCA is used to reduce the number of variables by isolating the directions of maximal variance. We use the covariance matrix and find its eigenvalues and associated eigenvectors.

Also, the transformed variables are fewer than the original ones, so PCA is considered one of the dimensional reduction techniques.

ii. **DATA VISUALISATION AND EDA:**

PCA helps in the visualization of high-dimensional data. The MINST data set has 784 dimensions and hence it is difficult to visualize all the dimensions, PCA reduces the dimensions and hence can be useful to visualize the data easily. The components are visualized separately to see which picture shows clear output.

iii. **NOISE REDUCTION:**

When the dimensions are reduced, indirectly the noise or the erroneous data is removed and hence result is Noise Reduction. Noise may include some different data points, random patterns.

iv. **BUILDING PREDICTIVE MODELS:**

PCA aims to orthogonally transform correlated variables to a smaller set of linearly uncorrelated variables (principal components). Hence, this will reduce the multicollinearity.

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

i. **BASIS TRANSFORMATION:**

Basis is a unit in which we express the vectors of a matrix.

Basis Vectors are certain set of vectors whose linear combination is able to explain any other vector in that space.

Any data observation can be represented in terms of some fundamental units known as basis or basis vectors.

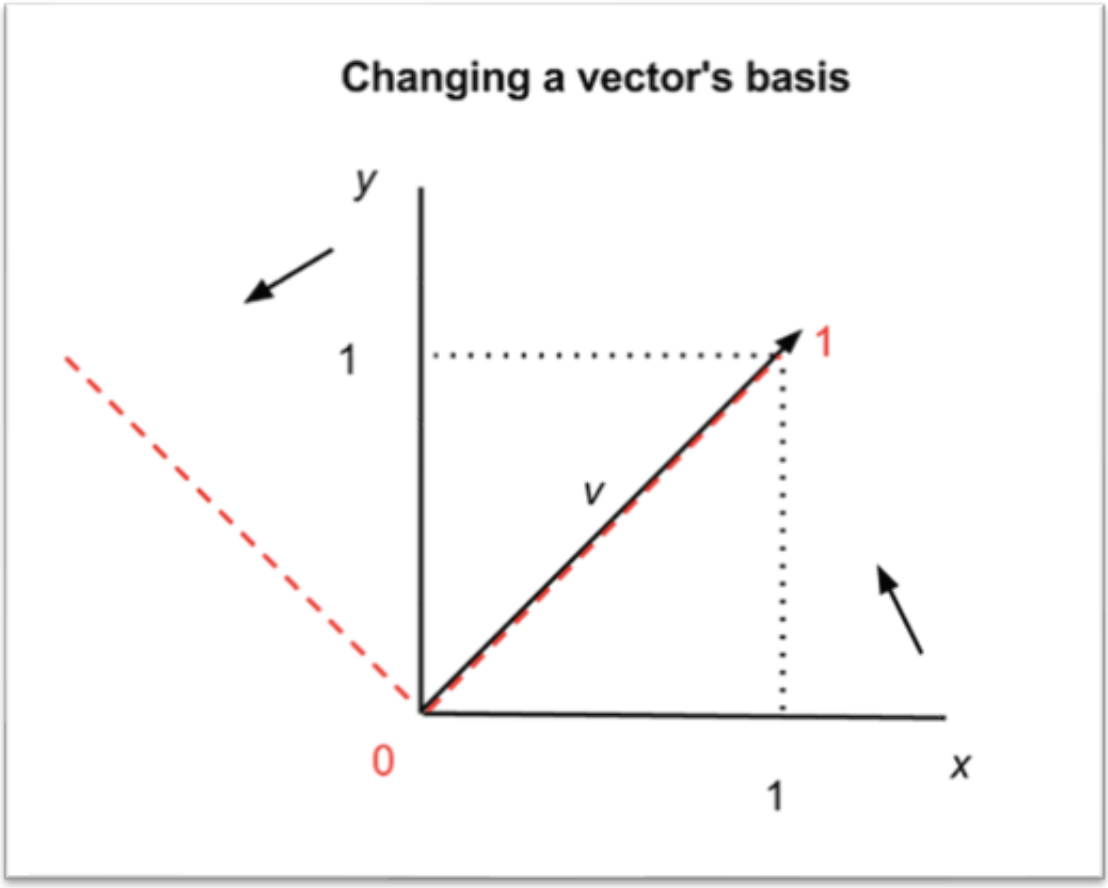
FORMULAS:

$$\text{New basis Represention (ft)} = M \times \text{Old basis Representation (cm)}$$

$$M^{-1} \times \text{New basis Representation (ft)} = \text{Old basis Representation (cm)}$$

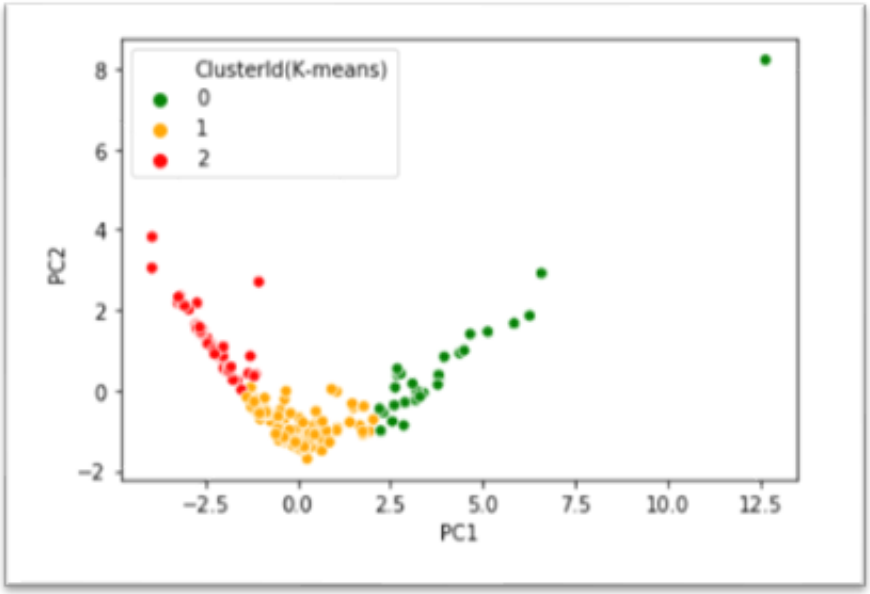
The change of basis matrix M represents the old basis in terms of the new basis.

We can transform the original data set so that the eigenvectors are the basis vectors and find the new coordinates of the data points with respect to this new basis. Changing the basis may achieve dimensionality reduction.



ii. **VARIANCE AS INFORMATION:**

When there is more variance in the data, we can say it produces more information. Hence we can drop the features with less variance, which means dimension reduction. Basis vectors or directions that capture the maximum variance are called Principal Components.



Here, we can see the Variance explained by the two components of PCA from our country data set.

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

i. **Outliers***:* PCA is also affected by outliers, and normalization of the data needs to be an essential component of any workflow.

ii. **Interpretability***:* Each principal component is a combination of original features and does not allow for the individual feature importance to be recognized.

iii. **Performance of the Model**: PCA can lead to a reduction in model performance on datasets with no or low feature correlation or does not meet the assumptions of linearity.

iv. **Linear assumptions:** If we have some of the variables in our dataset that are linearly correlated, PCA can find directions that represent our data. But if the data is not linearly correlated PCA is not enough.