Clustering and PCA Assignment: Part II

Question 1. 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)

Solution. HELP organization needs decision how to invest money **strategically** and **effectively,** and this can be done if **countries** can be identified those are in **direst need of aid,** using this the suggestion can be given. A dataset was given that includes the Country details of **socio economic** and **health factors**. **Data Exploration** was done in the beginning, that involves the analysis of **Dimension, Different Data Types**, whether any **Null values** or **Duplicates** were being there or not. **Univariate** and **Bivariate Analysis** were done, to understand the data characteristics. While doing the Data Exploration, we found some features has a **huge data spread**, **Outlier** Analysis was done, and it was found, some of the features like **gdpp**, **income** & … had a huge spread, this was calculated after calculating the **IQR**. Then using correlation matrices, it was checked whether **Multicollinearity exists or not,** and it was found there exist some. Prior to applying PCA, the **Data** was scaled using the **Standard Scaler**. Then **PCA** was applied on the whole dataset, **Covariance** and **Principal Components** were determined. Then using the **Scree plot**, which involves the plotting of **Cumulative variance** explained by the **Components**, **95% variance** was considered, and the number of components was **5**, then Data frame was transformed, then the **Correlation Matrice** was plotted, and it was inferred nearly zero correlation exists. Prior to applying clustering, **Hopkins Statistic** score was calculated, and result was **0.87**, so data points have **high tendency to cluster**. Then using **K-Means** and **Hierarchical Clustering** with **Single** and **Complete linkage, cluster labels** were formulated, and it was found for **Single Linkage** the **clusters were not so distinct** by **visualization** the **Top two Principal Components with Clyster IDs**, so we proceeded with rest two. Using the features **health, child_mort, gdpp, income**, cluster ids were identified for both K-Means and Hierarchical Clusters for Countries in need. For both **K-Means and Hierarchical** the results were nearly identical, so we proceed with K-Means. Then using that information, Final List of Countries were determined.

Question 2.a Compare and contrast K-means Clustering and Hierarchical Clustering.

Solution.

a. Hierarchical Clustering can't handle large amount of data, but K-Means works well with any amount of data w.r.t Time Complexity**,** because for K-Means is Linear i.e. O(n) while for Hierarchical Clustering its Quadratic $O(n^2)$.
b. In K-Means clustering, because of random cluster centroids point, the output will differ from time to time. But for Hierarchical Clustering results are always identical for each run.
c. K-Means works well if the data points distribution shape is spherical.

d. Prior to using K-Means, K number of cluster and Cluster Centroid must be decided at the very beginning. But for Hierarchical Clustering, using the Dendrogram, the number of clusters to consider can be determined based on Cutoff Value of Similarity.

Question 2.b Briefly explain the steps of the K-means clustering algorithm.

Solution. K-Means is an **Unsupervised algorithm**, and this useful when we have unlabeled data. Goal is finding the cluster or groups in the data. So, the algorithm is

1. First determine the **number of K** (cluster points or Centroid or mean)
2. The algorithm starts with the initial estimates for the K-Centroids, which can either be randomly generated or randomly some data point is selected from the data set.
3. **Data Assignment Step**: In this step, each cluster centroid determines the Cluster, so now each data point is assigned to its near most centroid, by using the Euclidean distance.

$$\text{argmin}_k \, || \, x_i - \mu_k \, ||^2$$

where $x_i = i^{th}$ Data Point, μ_k = Mean/Centroid of K^{th} cluster

4. **Optimization Step:** Recompute the Centroid of the newly formed cluster. This is done by taking the mean of all the data points assigned to that centroid cluster

$$\boldsymbol{\mu} = \frac{1}{n_k} \sum_{i:Z_i=k} \boldsymbol{x_i} = \frac{\textbf{\textit{Find all the Datapoints belonging to the Cluster K and then sum them up}}}{\textbf{\textit{Number of Datapoint belonging to cluster K}}}$$

5. The above two steps are repeated until there is no change in the cluster centers or the algorithm doesn't converges.

Question 2.c How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Solution: There are mainly two well known method to choose k,

a. **Elbow Method**: This method is also known as SSD (Sum of Squared Distance). The main objective of clustering is to define Cluster in such a manner that total intra cluster distance is minimized. Now the within cluster distance will be computed using **Inertia,** it says how far the points are within the cluster.
   Algorithm to find k via Elbow curve:
   i. Compute the K Means for different values of k, i.e. value of k cluster from 1 to 10
   ii. Now calculate the inertia or the total intra cluster distance
   iii. Now plot for each value of K what is the inertia
   iv. In the end find the optimal value of k or number of clusters, we need the "elbow", i.e. the value after which the inertia starts decreasing in linear fashion.

b. **Silhouette Method:** It mainly captures the cohesion & dissimilarity. It is also knows as the measure of goodness. Here mainly two measures are computed.
   i. **a(i)** Average distance from own cluster (cohesion / intra cluster distance)
   ii. **b(i)** Average distance from the nearest neighbor cluster i.e. separation
   Here ***i*** stands for each data point

Objective here is to keep a(i) as less as possible and b(i) as large as possible. Because objective of clustering is to keep low intra cluster distance and high inter cluster distance

$$\text{Silhouette of } i^{th} \text{ point} = S(i) = \frac{b(i) - a(i)}{max\ (b(i), a(i))}$$

Algorithm to find k via Silhouette Method

i. Compute the K Means for different values of k, i.e. value of k cluster from 1 to 10
ii. For each value of k, calculate the average silhouette of observation is calculated
iii. Plot the curve of average silhouette values of cluster k
iv. The location of maximum is considered as the ideal value of cluster

Silhouette Analysis score usually lies within the range **[-1,1].**

Using both either of the metric in statistical way we can have a say what could be an ideal number of clusters to consider. But sometime from the business perspective we can have some other consideration, like lets says a business strategy can be to group the customer in let's say 3 categories then the business people would want to devise a strategy or some business rules for those customer segment, then the it's needs to check using one such metric it can said in terms of grouping it seems better to choose three as compared to 4/5 …. So, this brings the role of both qualitative consideration/strategy consideration as well as metric consideration to choose what k is ideal. Sometimes also there are cases where we first use the Clustering methods to segment the data and then they are used for prediction. Then we can play with different value k to check how the prediction works, so it can be one of the business needs.

Question 2.d Explain the necessity for scaling/standardisation before performing Clustering.

Solution: In Standardization of datapoint, data is being converted in such a manner that the distribution will have a mean of 0 and standard deviation of 1.

There can be a possibility for different data points the scales are different, or they have different units like (metre and km or in cm or inch). So scaling is very much important in cluster analysis, because groups are defined based on the distance between points in mathematical space. For computation of Euclidean distance, between the data points, it is very important to ensure that the data points with higher value doesn't outweigh the attribute with smaller range, so scaling plays a crucial role here.

Question 2.e Explain the different linkages used in Hierarchical Clustering.

Solution. The problem arises, when a cluster contains more than one-point, Euclidean distance can only be calculated between a pair of scores. But the goal is still difference in scores between pairs of clusters, however in this case the clusters do not contain one single

value per variable. So, **Linkage** helps to find the distance between the pair of clusters. There are mainly three types of linkage

a. **Single Linkage**: It is also referred to as nearest neighbor or minimum method, this measure defines the distance between two clusters as the minimum distance found between one data point from the first cluster and one data point from a different cluster. For example cluster 1 contains data point a,b and cluster 2 contains data point c,d. So now the distance between cluster 1 and cluster 2 is smallest distance found between the pairs (a,c), (a,d), (b,c) and (b,d). Sometime Single linkage produces it produces chaining among the cluster, several cluster may be joined together because one of the data points is in proximity with a data point from a separate cluster.
b. **Complete Linkage:** It is also referred to as furthest neighbor or maximum method. It's pretty like Single Linkage except the part where it was searching the minimum distance between the pairs, this algorithm considers maximum/furthest distance between the cluster. This is the solution of Chaining which was a problem for Single Linkage. But this has one more problem, lets say from the above example a, b and c are within close proximity to one another, but if d differs significantly from the rest, then cluster 1 and cluster 2 won't be joined together, because of the difference in distance between (a,d) and (b,d), so it prevent close cluster to merge together.
c. **Average Linkage:** To overcome the issue of both the above linkage, average of the distance values between the cluster data point will be considered. The distance between each data point in the first cluster and every data point in the second cluster are calculated and averaged. That means for the above example, distance between the cluster 1 and cluster 2 would the average of (a,c), (a,d),(b,c) and (b,d).

Question 3.a Give at least three applications of using PCA

Solution: Application of PCA are

i. **Image Analysis**: PCA can be applied to compress the image, as image have a lots of data points, imagine an RGB image then we will have data point of

(256 * 256) * 3 [for pixels and color channels] many dimensions. This helps in Image Compression. PCA also used to find Patterns used especially in Image Recognition.

ii. Projection of data point in **2D space** to **1D space.**
iii. **Neuroscience:** To identify the specific properties of a stimulus that will increase **neuron's probability of generating an action potential.**
iv. Principal Component Analysis provides a general frame for **systemic approaches in pharmacology**.
v. PCA can be used in the field of analysis with **Gene data**
vi. It's also used in the field **Noise Reduction**

Question 3.b Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Solution Basis is a unit in which express the vector of a matrix. Vectors in any dimensional space or matrix can be represented as a linear combination of basis vector.

Using the analogy of Basis as a unit of representation, different basis vector can be used to represent the same information. So, in below you can see the data frame and their respective transformation and their Basis Vectors.

Patient ID	Height (cm)	Weight (kg)
p1	165	55
p2	155	71
p3	165	88
p4	160	105
p5	160	94

⟷

Patient ID	Height (ft)	Weight (lbs)
p1	5.4	121.3
p2	5.1	156.5
p3	5.4	194.0
p4	5.2	231.5
p5	5.2	207.2

Basis $\left\{ \begin{bmatrix} 1\text{cm} \\ 0\text{ kg} \end{bmatrix}, \begin{bmatrix} 0\text{cm} \\ 1\text{ kg} \end{bmatrix} \right\}$ $\left\{ \begin{bmatrix} 1\text{ft} \\ 0\text{ lbs} \end{bmatrix}, \begin{bmatrix} 0\text{ft} \\ 1\text{ lbs} \end{bmatrix} \right\}$

As it can be seen when changing the height and weight it's not changed so significantly. So, it can be seen $\begin{bmatrix} 165 \\ 55 \end{bmatrix}$ is same as $\begin{bmatrix} 5.4 \\ 121.3 \end{bmatrix}$ with different basis vector is being used.

$$\text{New Basis Representation} = M * \text{Old Basis Representation}$$

$$\text{Old Basis Representation} = M^{-1} * \text{New Basis Representation}$$

So here M is a representation of Old Basis (ft and lbs) in new Basis (cm and kg) and M^{-1} is the representation of new Basis (cm and kg) in old Basis (ft and lbs).

In the above pic, ft and lbs basis vector and their corresponding representation in cm and kg are provided. So, these are the two-basis vector, one in ft and lbs as basis and other for cm and kg as basis. Now if the vectors are represented from ft and lbs space to cm and kg space, all we need to do is , Here everything in green is M and $[5.4\ 121.3]^{-1}$ is the old Representation and $[165\ 55]^{-1}$ is the new Basis representation

$$\begin{bmatrix} 30.48 & 0 \\ 0 & 0.45 \end{bmatrix} \times \begin{bmatrix} 5.4 \\ 121.3 \end{bmatrix} = \begin{bmatrix} 165 \\ 55 \end{bmatrix}$$

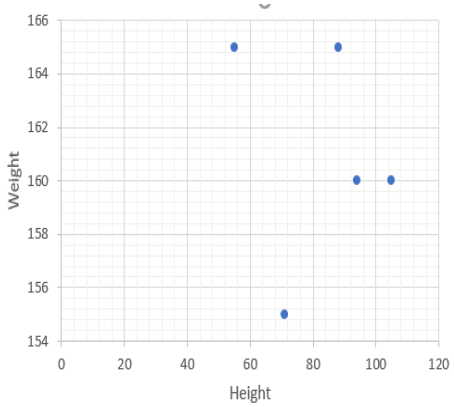
But if we want to go from cm and kg space to ft and lbs, all we need to do is M^{-1} $\begin{bmatrix} 0.0328 & 0.0 \\ 0.0 & 2.22 \end{bmatrix} \begin{bmatrix} 165 \\ 55 \end{bmatrix}$ M^{-1}

In PCA the basic idea is we want to change the basis of the data so that we find some new basis in which representation so that it becomes much more useful, i.e. in context of dimensionality reduction, it means we want to find a new basis vectors where it is very helpful to identify whether it is informative or whether it is not so informative whether some features can be dropped and to keep some features.

Variance as Information: Variance is a numerical measure for the variation that occurs in the field. If a Feature has lot of Variation then the features will have high Variance, if low variation then low Variance. If a Feature has low variance i.e. the information content in that feature is very low, but if a feature has higher variance then the information content is higher. So, based on this explanation higher variance i.e. the feature has high importance, lower variance i.e. the feature has low importance.

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

	Height (cm)	Weight (kg)	Age
	165	55	22
	155	71	22
	165	88	22
	160	105	22
	160	94	22
Variance (σ^2)	**14.0**	**311.44**	**0.0**



As it can be seen here as an example, the variance content by Age is 0, then information content here is very less. And we also see from scatter plot, weight has a higher variance as compared to Height. So, some columns have much less variance than others, it is easier to remove those columns and do dimensionality reduction.

So now if the variance is similar, then there is a need to find the direction in which the maximum variance of data can be found. So, if the variance along the axis is comparable, then it changes the basis vector in such a way, that the new basis vector captures the maximum variance or information.

So, these two fundamental blocks help in determining the ideal basis vector i.e. M.

1. Explains the direction of maximum Variance
2. When it is being used as a new set of vectors, the transformed dataset is now suitable for dimensionality reduction
3. Directions explaining the maximum variance are called the Principal Components

Question 3.c State at least three shortcomings of using Principal Component Analysis.

Solution Some of the limitations of PCA are

a. PCA is limited to **linearity**, it is beneficial to use when the data is linear correlated, but if the data is not linearity correlated then PCA is not an ideal approach.
b. PCA needs the components to be **perpendicular**/ it mainly relies on **orthogonal projection**, but it might not be the case every time.
c. PCA assumes that the variable with **low variance** has very **less importance**, which might not be the case in case of prediction (for a classification problem if there is a class imbalance)