Question 1: Assignment Summary

Problem statement: -

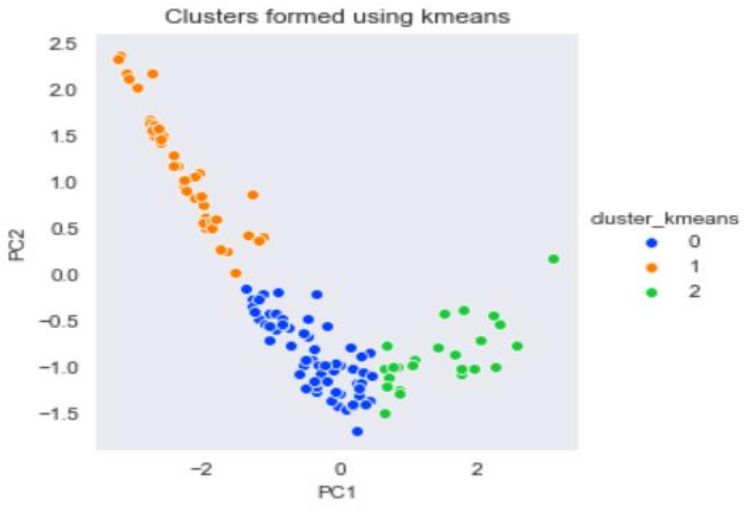
HELP International has raised $10M for fighting poverty and providing aid to backward countries. Based on various socio-economic factors, group the countries into various groups for analysis. Based on the groups formed determine countries in dire need of funds that have been raised.

Solution:-

a. Data Analysis-
    a. Analysis reveals that there are no missing values in the data.
b. Data preparation-
    a. Scale data before stating the modeling process so that there are on same scale and do not impact the centroid selection.
c. PCA on Data-
    a. Use PCA for dimensionality reduction and solve multicollinearity issue.
    b. Looking at the cumulative probability of variance explained, and scree plot it was evident that 5 PCs is the optimal choice.
    c. Convert the data to new basis i.e PCs and use modified dataset (PCs) for model building.
d. Remove outliers-
    a. Clustering is sensitive to outliers and hence we will remove the outliers using statistical approach i.e. using IQR.
e. KMeans clustering-
    a. Use cost function (SSD) and plot SSD to determine optimal cluster for the given data.
    b. Use Silhouette analysis to review/confirm the optimal cluster selection. Based on both elbow curve and silhouette score, 3 is the optimal cluster selection.
    c. Cluster the data into 3 groups and check the clusters formed.
    d. Cluster the original dataframe so that the clusters and visualization are more meaningful.
f. Hierarchical clustering-
    a. One prime advantage of Hierarchical clustering is the need to predefine cluster i.e. select cluster beforehand.

b. Use both single linkage and complete linkage method for clustering. Review the dendrograms and determine linkage to be used for further clustering of data.
c. Single Linkage cluster takes the shortest distance between points in 2 clusters as distance whereas Complete linkage takes the longest distance between 2 points in 2 clusters as distance.
d. Complete linkage clustering methods gives better defined clusters and will be used for further clustering. This is evident from dendrograms.
e. Cut the dendrogram to form clusters.

g. KMeans clustering provided a better clustering in this case. Population is not concentrated in 1 single group. This is evident from value count for each cluster.
h. Below image shows separation of countries when plotted on PC1, PC2



i. Conclusion:
    a. From Cluster which has countries with lowest income, gdpp and highest child_mort, we will use child mortality and life expectancy factors to identify the countries for funds. These additional funds can be utilized for improving the health condition of the country.
    b. Top 5 countries that need aid are
        i. Sierra Leone
        ii. Chad
        iii. Mali
        iv. Niger
        v. Angola

Question 2: Clustering

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

	KMeans Clustering	Hierarchical Clustering
Cluster Selection	Selection of # of cluster is pre-requisite	Predefined selection of clusters not required.
Process	This is a non-linear process	This is a linear process
Dataset Size	Preferable for large dataset	Preferable for small dataset only since this is linear process.
RAM	This is a non-linear process and does not needs more RAM	This is a linear process and is computationally intensive
Centroid	Initial choice of centroid impacts the cluster formation.	No limitation on initial choice of centroid.

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

- In our assignment, we will take pca modified dataset to form clusters.
- Determine number of clusters (k) to be formed from the given data.
- Select random k centroid.
- Follow below 2 steps repetitive till convergence (or max iteration specified)

  Assignment Step
  i. Assign each point to nearest cluster using shortest distance (Euclidean distance).
  ii. At the end each point is assigned to one of the k clusters

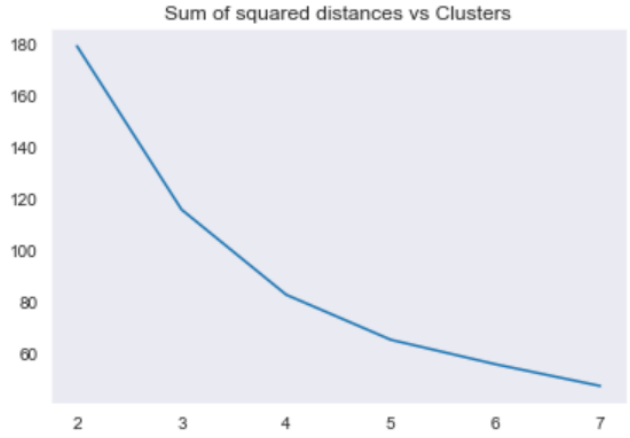
  Optimization Step
  i. Calculate new centroid from the data points assigned to each cluster.
  ii. Go back to assignment step to re-assign the points to clusters
- Follow assignment and optimizations step till centroids no longer change (converge) or max iterations has been achieved.

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

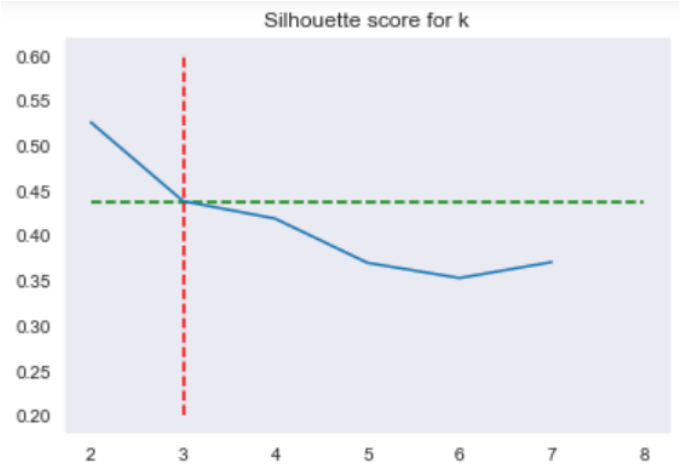
Statistically there are 2 techniques that are predominantly used.

Elbow Curve – In this method k-means clustering is run for series of clusters ranging from 2 to 8 (in our case) and sum of squared distance (SSD) is calculated. KMeans.intertia_ gives the SSD for a given k (cluster). These SSDs are plotted

against the k to determine optimal value of k. Value of k where we have last curve of the elbow is usually a good optimal k.



Silhouette score – Represents how similar a datapoint is to its own cluster compared to other clusters. This is calculated using inter cluster and intra cluster distance.
Ex – a(i) – Average inter cluster distance
b(i) – Average intra cluster distance
Silhouette sore = b(i) – a(i)/max(a(i),b(i))
A positive value indicates google clustering whereas a negative score indicated bad clustering.



From business standpoint, We should not create too less or too many clusters as it would become difficult to interpret with such data. Usually clusters are kept within range of 3 to 15.

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

One of the major components of clustering is Euclidean distance (or distance). If we do not standardize the data then when we calculate Euclidean distance the feature which is on higher scale (like amount, points, balance, weight) will overpower other features which are on lower scale (like rating, interest rate etc). This will result in datapoint being assigned to clusters just based on features having higher scale.

If you take below data which is not standardized, the cluster is insensitive by change in value of "Rating".

2 clusters are

Centroid	X	Y
C1	100	4
C2	80	2

Price	Rating	Price Std	Rating Std	C1 Distance	C2 Distance	Cluster
50	4	-0.37372	0.1072	50	30.066593	2
100	5	1.121153	0.9649	1	20.223748	1
120	5	1.719101	0.9649	20.024984	40.112342	1
50	4	-0.37372	0.1072	50	30.066593	2
20	4	-1.27064	0.1072	80	60.033324	2
80	2	0.523205	-1.6082	20.099751	0	2
60	5	-0.07474	0.9649	40.012498	20.223748	2
20	2	-1.27064	-1.6082	80.024996	60	2

If we scale the data, then we observe the both the features have impact on cluster assignment.

Centroid	X	Y
C1	1.121153	0.886796
C2	0.523205	-0.68973

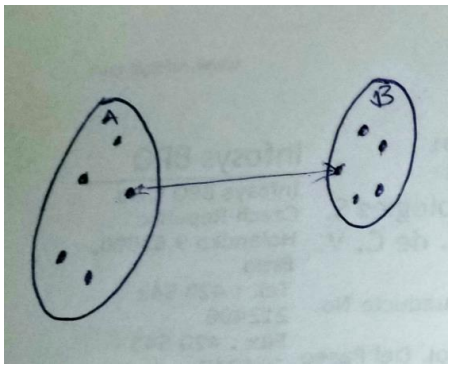
Price	Rating	Price Std	Rating Std	C1std	C2std	Cluster_Std
50	4	-0.37372	0.1072	1.6859	1.1998	2
100	5	1.121153	0.9649	0.0781	1.7594	1
120	5	1.719101	0.9649	0.6030	2.0416	1
50	4	-0.37372	0.1072	1.6859	1.1998	2

20	4	-1.27064	0.1072	2.5156	1.9629	2
80	2	0.523205	-1.6082	2.5656	0.9184	2
60	5	-0.07474	0.9649	1.1984	1.7594	1
20	2	-1.27064	-1.6082	3.4562	2.0153	2

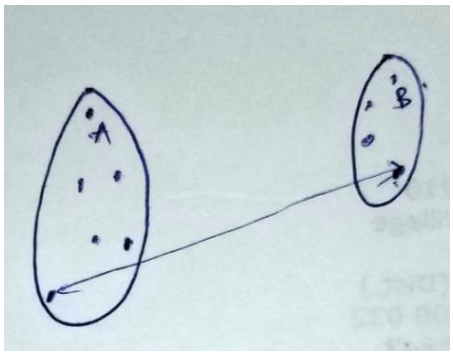
e. Explain the different linkages used in Hierarchical Clustering.

Linkage is measure of dissimilarities between clusters having multiple observations.

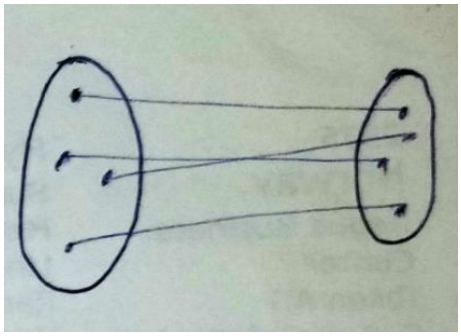
i. Single Linkage cluster takes the shortest distance between points in 2 clusters as distance.



ii. Complete linkage takes the longest distance between 2 points in 2 clusters as distance.



iii. Average Linkage cluster takes the average distance between points in 2 clusters.



iv. Single linkage clusters lead to generation of loose clusters i.e. Intra cluster variance is high.
v. Complete linkage clustering methods gives better defined clusters.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

i. PCA is used in banking, insurance and retail domain where amount of data is huge, and task is to find patterns in the data. Example – Finding good, average and bad customers.
ii. PCA is also widely used in facial recognition, OCR (Optical Character Recognition) predominantly because of dimensionality reduction capability.
iii. PCA is also used in dataset where there is strong multicollinearity in variables.
iv. PCA when used along with other linear models yields better performance.

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

Basis Transformation – Basis is the unit of matrix. If we take standard basis then it is [1,0] and [0,1]. We can represent a vector as linear combination of standard basis. We can represent the same data (vector) in some other basis by transforming it. See the example below.

Car	Mileage (mpg)	Top Speed (m/hr)	Mileage(km/l)	Top Speed(km/hr)
1	40	100	17.2	162
2	30	110	12.9	178.2
3	25	120	10.75	194.4
4	10	140	4.3	226.8
5	15	125	6.45	202.5

Let's take example of C1 represented in Basis vector

C1 in Bstd = $40\begin{bmatrix}1\\0\end{bmatrix} + 100\begin{bmatrix}0\\1\end{bmatrix}$. Now same C1 can be represented in new basis as below

$40\begin{bmatrix}0.43\\0\end{bmatrix} + 100\begin{bmatrix}0\\1.62\end{bmatrix}$.

In PCA, one of the fundamental concepts is to change the basis so that it can be used better for representation of the data.

Variance as information – Variance is nothing but how varied the information is in column. If the variance in column is very less then it does not add any value to the dataset.

Example – Consider a dataset for various mobile models. If this dataset has a column, say category, with all values as Electronic Gadget then it does not add any value to the datapoints.

When such columns are removed from dataset it results in dimensionality reduction. Directions in which maximum variance is explained are ideal basis vectors and are our Principal components.

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

i. PCA mandates that the principle components are perpendicular to each other. This is not always optimal solution. We can use ICA (Independent Component Analysis) to overcome this limitation. ICA is slower than PCA.
ii. One of the assumptions of PCA is that PCA have to linear combination of original component. Sometimes data requires non-linear combinations. T-SNE is alternative/solution to this limitation. However, t-SNE is costlier computational.
iii. PCA assumes that features with less variance are not useful and can be dropped to achieve dimensionality reduction. However, in imbalanced data (Eg – Fraud detection) dropping data with small variance can result in incorrect result.