Assignment Summary

Q1: Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words

Ans:

Problem Statement:

Provide insights on which countries are in dire need of the help from NGO

- Perform PCA on the dataset and obtain the new dataset with the Principal Components
- Perform Outlier Analysis
- Analyze the clusters and identify the ones which are in dire need of aid.

Methodology Followed:

1. Basic Analysis of the dataset provided
    - Performed the basic analysis on the provided dataset such as checking for the null values, quantile values, min and max values
    - There were no null values found in the dataset

2. Outliers Treatment on the dataset
    - Performed the outlier analysis on the dataset to check if data contains any outliers as if present outliers can significantly influence the clusters and PCA.
    - Removed the outliers from “GDPP” and “Income” columns.
    - Reduced the number of records from 167 to 128.

3. Scaling the data:
    - Scaled the dataset so as to speed up the algorithm pace and avoid any fluctuation in the coefficient values.

4. Applying PCA:
    - Applied the PCA on scaled data.
    - From PCA analysis found that 5 Principal Components describe up to 95 % variance in the data.
    - Hence decided that number of Principal Components used will be 5.
    - Using the defined PC, ran the PCA again so as to get the component values.
    - Plotted the Scatter Plot between obtained PCs to decide if the data can be clustered or not.
    - Calculated the hopkins score to back up the decision obtained using scatter plots that data can be clustered.

5. Clustering :
    - Applied the KMeans Clustering algorithm on the dataset.
    - Using the Silhouette Score Plot and Elbow curve, decided that number of clusters will be 3
    - In silhouette score plot, average max value is at 3 and in Elbow Curve, we have an elbow at 3.
    - Reran the KMeans algorithm using predefined cluster number i.e. k = 3
    - Applied Hierarchical Clustering (Both Single and Complete Linkage) to the dataset.

- When clusters were plotted using the results obtained in Hierarchical Clustering, were not able to distinguish the clusters easily.
- Hence decided to go ahead with the KMeans clustering algorithm.

6. Cluster Analysis:
    - Found that Cluster 0 is made up of “Under Developing” countries.
    - Analyzed the cluster on “GDPP”, “Child Mortality” and “Income” features.
    - The Countries those have low GDPP, low income, and high child mortality are the ones those need help from the NGO.
    - Below are the 5 countries those will need help from NGO
        i. Niger
        ii. Sierra Leone
        iii. Madagascar
        iv. Mozambique
        v. Central African Republic

Clustering

Q1: Compare and contrast K-means Clustering and Hierarchical Clustering.

Ans:

In below table, we can find some of the points those distinguish K-Means and Hierarchical Clustering.

KMeans	Hierarchical
Need to specify desired number of clusters at the start of clustering process.	No need to specify the clusters at the start of clustering process.
Randomly assigning data points to a cluster	Data points are not randomly assigned.
Need to execute steps again and again to find the perfect clusters.	Clusters are formed in one run only.
No dendrograms plotted	Dendrogram needs to be plotted so as to compute the number of clusters
Can handle large data easily.	Cannot handle large data sets easily.

Q2: Briefly explain the steps of the K-means clustering algorithm.

Ans:

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

Steps involved in KMeans clustering:

1. Specify the number of clusters k.
2. Initialize the centroids by shuffling the dataset and then randomly selecting k data points as centroids.
3. Compute the sum of squared distance between data points and the centroid.
4. Assign each data point to the nearest centroid.
5. Repeat step 3 and 4 until there is no change between the centroid.

Q3: How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Ans:

Defining number of clusters is the most important decision to be taken in clustering. Hence whenever data is to be clustered, the analyst must have the knowledge of the data he/she is working on which will help in deciding the number of clusters.

There are two methods via which number of clusters can be decided.

1. Elbow Method:
    a. Elbow method works on the principal of Sum of squared distances between data points and their assigned cluster' centroid.
    b. One should pick the k value where the SSE is flattening out and is forming an elbow curve.

2. Silhouette Score Analysis:
    a. Silhouette analysis can be used to determine the degree of separation between clusters.
    b. Compute the average distance between data points in same cluster
    c. Then find average distance between data points from closest cluster.
    d. Compute the score
    e. The average value where silhouette average score is maximum, that value can be taken as optimal value of clusters.

In the Assignment, we have chosen the value of Clusters as 3. This value is verified using the Elbow curves as well as Silhouette Score.

Also we can see that there are three types of countries,

1. Developed
2. Developing
3. Under Developed

Hence, our k fits true.

In this way, if one has business knowledge of data, then finding optimal number of clusters becomes easy.

Q4: Explain the necessity for scaling/standardization before performing Clustering.

Ans:

Due to below reasons scaling is necessary when performing clustering:

1. Whenever business data is provided it can have any of the units (e.g., inches, meters, tons and kilograms)
2. Scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000).
3. The importance of this is high in clustering is high as groups are defined based on the distance between points in mathematical space.
4. These units are not directly comparable and can cause formation of wrong clusters

5. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unit less measure or relative distance.
6. To avoid forming wrong clusters, data needs to be normalized.
7. After normalizing the data, range of data is proportional no matter what feature is.

Due to these reasons, data must be normalized before clustering.

Q5: Explain the different linkages used in Hierarchical Clustering.

Ans:

Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. A hierarchical clustering is often represented as a dendrogram.

Types of linkages:

1. Single Linkage:
    a. It is the shortest distance between pair of data points in two clusters.
    b. This linkage outputs spread out clusters.

2. Complete Linkage:
    a. In this linkage we measure the distance between two farthest points in two clusters.
    b. It produces tighter clusters than the single linkage.

3. Average Linkage:
    a. Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance.

Principal Component Analysis:

Q1: Give at least three applications of using PCA.

Ans:

The aim of PCA is dimensionality reductions. Using PCA, 100 different variables can be converted to 10 relevant Principal Components.

Applications of PCA:

1. PCA on Images:
    a. An image of N*N pixels can be represented by N*N matrix.
    b. When image detection is to be done, this N*N matrix can be reduced to a lower dimension matrix without losing any credible data by using PCA.

2. Pattern Recognition:
    a. In Pattern Recognition, numbers can be converted to the lower dimensions using PCA and can be plotted so that they can be matched.
3. Data Compression
4. Time Series Prediction
5. Finance

Q2: Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information

Ans:

Basis Transformation:

Any vector can be represent as a linear combination of basis vectors. It is desirable to work with more than one basis for a vector space, it is of fundamental importance in linear algebra to be able to easily transform coordinate-wise representations of vectors and operators taken with respect to one basis to their equivalent representations with respect to another basis. Such a transformation is called a change of basis.

Principal components are simply the eigenvectors of the covariance matrix used as basis vectors. Each of the original data points is expressed as a linear combination of the principal components, giving rise to a new set of coordinates.

Using PCA, we capture the Components those give maximum variance. These components are can be defined using alternative co-ordinates. Due to which we need change in basis concepts.

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

Variance as Information:

In case of PCA, "variance" means summative variance or multivariate variability or overall variability or total variability. Below is the covariance matrix of some 3 variables. Their variances are on the diagonal, and the sum of the 3 values (3.448) is the overall variability.

```
1.343730519   -.160152268    .186470243
-.160152268    .619205620   -.126684273
 .186470243   -.126684273   1.485549631
```

Now, 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. So, the covariance matrix between the principal components extracted from the above data is this:

```
1.651354285    .000000000    .000000000
 .000000000   1.220288343    .000000000
 .000000000    .000000000    .576843142
```

The 1st principal component accounts for or "explains" 1.651/3.448 = 47.9% of the overall variability; the 2nd one explains 1.220/3.448 = 35.4% of it; the 3rd one explains .577/3.448 = 16.7% of it.

Here the variance of first component is 48% which says that approx. half of the information can be derived from the first component only.

Also second component provides 35% variance, hence we can use these two components only to get 83% information of the data. This reduces the issue of multi collinearity.

Q3: State at least three shortcomings of using Principal Component Analysis.

Ans:

Below are the short comings of PCA:

1. PCA is limited to linearity, it works well only on linear data
2. PCA needs the components to be perpendicular, though in some cases, that may not be the best solution.
3. PCA assumes that columns with low variance are not useful, which might not be true in prediction setups.
4. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.
5. 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