Assignment – Part II – Subjective (Submitted by: Vikram Mathur)

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)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Answer 1

Problem Statement:

- Identifying at least 5 countries in direct need of aid for utilization of funding available with HELP International NGO.
- Objective is to determine the overall development of the countries based on socio-economic and health factors to take this decision. Adopting PCA and Clustering technique to solve this problem.

Solution Methodology: Following steps were involved:

1. Reading & Understanding Data
    - Dataset shape, info, describe applied to study and understand it.
2. Cleaning the Data
    - Checked for missing data.
3. Data Preparation and EDA
    - Handled Percentage Formats for exports, imports and health
    - Correlation Analysis (Pair Plots, Heatmap)
    - Outlier Analysis & Treatment – 4 rows removed which were beyond 95th percentile (essentially developed countries records, hence no impact on analysis)
4. Principal Component Analysis – Post scaling, decision taken for three (3) principal components. Reasons:-
    - After computing the cumulative Explained Variance Ratio and plotting the scree plot it was statistically determined that around 90% variance is explained by 3 principal components.
    - Along with this, the review of correlation of the nine features helped conclude the need for 3 PCs.
5. Clustering (Kmeans and Hierarchical) - Decision taken for considering three (3) clusters. Reasons:-
    - Hopkins Statistic of 87% (greater than 50%) suggested good tendency to form clusters.
    - SSD (Elbow Curve) visually depicted a bend of 2 clusters and also a marginal bend on 3 clusters. Also, the Silhouette Score of 51% was coming for 2 clusters and 45% for 3 clusters. Based on this a decision to take 2 clusters was determined and KMeans algorithm was applied.
    - Hierarchical clustering's complete linkage plot indicated 3 clusters giving a fair and clear distribution of all data points. From a business perspective 3 clusters meant – *Developed*, *Developing* and *Under-Developed* countries. Hence, it was concluded to apply 3 clusters.
6. Cluster Analysis
    - Analyzing Features in Clusters – Box plots and scatter plots to visualize data point segregation on clusters.
    - Feature Means in Clusters – To bin the data and filter the countries from the under-developed clusters which are in direct need of aid.

Question 2: 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.

Answer 2 (a)

	K-Means Clustering	**Hierarchical Clustering**
1.	K-Means Clustering is the division of data-points (objects) into clusters so that each data-point is in exactly one cluster. Eventually, the data-points lie is very dissimilar clusters.	In hierarchical clustering, two similar clusters are merged and combined together and this roll up of clusters is repeated so that all clusters containing data points or objects are in a single cluster.
2.	The number of clusters needs to be pre-determined (before the execution of the algorithm)	The number of clusters is determined post visualization in a dendrogram by simply navigating the layers of the tree.
3.	The data is divided into clusters as the first step and in the subsequent steps the clusters are refined to get the most optimal grouping.	The data is not divided into clusters in a single step. A series of divisions / merges happen, which may result in a single cluster containing all objects to k clusters that contains a single object or vice versa.
4.	As the choice of clusters chosen is random in the beginning hence the results by running the same algorithm might give different results.	The algorithm gives similar reproducible results.
5.	Algorithm’s time complexity is linear O(n)	Algorithm’s time complexity is quadratic O(n^2)
6.	Algorithm: 1. Determine random k points (referred as cluster centroids) 2. Calculate the Euclidean distance of n data points with the centroid and determine the minimum distance. 3. Assign the ith data point to a particular kth cluster. 4. Calculate the mean to get the cluster centroid. 5. Optimize to get the correct centroid by re-calcuating the distance of n data points with the kth cluster centroid and determine new value of the centroid. 6. Repeat the process until convergence.	Algorithm: 1. Calculate NxN (similarity) matrix. That is distance of each data point from the other 2. Assign each data point to it’s own cluster. So, N data points means N clusters. 3. Merge closest pair of clusters into a single cluster so as to reduce one cluster. 4. Find distance between new cluster and all old clusters 5. Repeat above two points till we have a single cluster. This can be visually depicted in a dendrogram

Answer 2 (b)

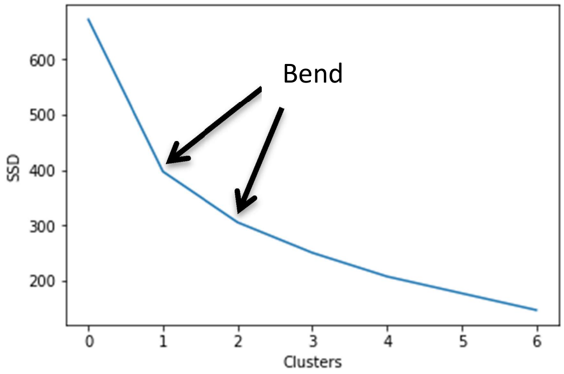
Following are the steps for K-Means Clustering algorithm:-

1. Determine random k points as the center of the clusters (referred as cluster centroids)
2. Calculate the Euclidean distance of n data points with the centroid and determine the minimum distance.
3. Assign the ith data point to a particular kth cluster.
4. Calculate the mean of all data points within a cluster in order to determine the new cluster centroid.
5. Optimize to get the correct centroid by re-calcuating the distance of n data points with the kth cluster centroid and determine new value of the centroid.
6. Repeat the process (above steps 4 and 5) until convergence. That is till we get optimal clusters.

Answer 2 (c)

Value of k (number of clusters) can be statistically determined using the following methods:

1. Elbow Method
    a. Compute the KMeans clustering based on different values of k. Lets say 1 to 8.
    b. For each cluster (k) calculate the sum of squared distances (SSD) of data points to their closest cluster
    c. Plot of sum of squares (y-axis) as per the cluster number (x-axis)
    d. The location of a bend (referred as elbow) is a good indicator depicting appropriate number of clusters.



    e. When SSD tends to decrease to zero with the increase in the value of k. SSD zero suggests that each data point is in its own cluster.
    f. Objective is to choose a small value of k which still has a low SSD. This is represented by the elbow.
2. Average Silhouette
    a. Compute the KMeans clustering based on different values of k. Lets say 1 to 8.
    b. For each cluster (k) we calculate the average Silhouette
    c. The location of maximum is considered as the appropriate number of clusters.

Business Aspect:

- In order to do Segmentation (Business Problem) we need to do Clustering (Analysis Technique)
- We need to finally decide the number of clusters based on what the business need is. The elbow method and Average Silhouette would statistically provide us the number of clusters however, it is important to note that once we get the emerging patterns what action we want to take on it.
- For example for a marketing campaign there may be multiple clusters for emerging patterns of customer behaviours based on statistical inferences. However, based on business need we need to “fine tune” the clusters in order to control the reach of our campaign.

Answer 2 (d)

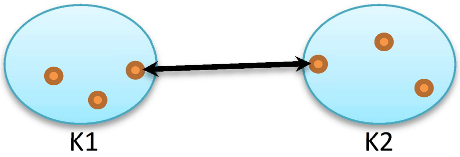
The reason for scaling / standardization is KMeans algorithm is as follows:-

1. Different features / variables may be in different units. By standardizing the features before executing the K-Means algorithm we get them to a normal scale.
2. KMeans algorithm involves computation of Euclidean distance between the data points. The features / variables having larger values should not outweigh the features / variables having smaller values. Hence, it is important to scale all the features to make them free of the associated units.

Answer 2 (e)

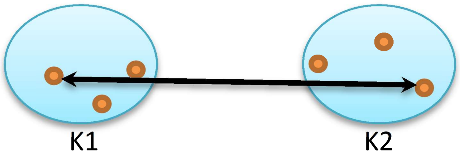
There are three types of linkages in hierarchical clustering. They are:-

1. Single Linkage
    a. Distance between two clusters is the shortest distance between the points in the two clusters. Distance between the data points closest to the other cluster is taken as the distance between the clusters.



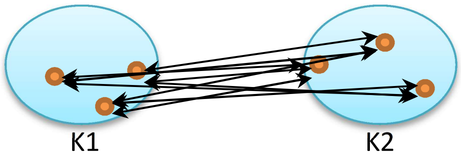
Min(Distance(K1 data point, K2 data point))

    b. Chaining: Single linkage clustering can produce straggling clusters. As the merging is strictly local, a chain of points can be extended for long distances without regard to the overall shape of the emerging cluster.
    c. Outliers: Single linkage is sensitive to noise and outliers.
    d. Cluster Size: Single linkage is good at handling non-elliptical shapes.
2. Complete Linkage
    a. Distance between two clusters is the maximum distance between any two points in the clusters. This distance is the intra cluster distance.



Max(Distance(K1 data point, K2 data point))

    b. Complete Linkage avoids chaining behavior as described in above pt. 1 b. However, it is impacted with crowding.
    c. Outliers: Complete linkage is influenced by outliers in comparison to single linkage. If outliers or extreme values are to be included while modeling then better use this technique as it includes the inliers first and then the outliers. Complete linkage does not necessarily merge groups that are close together due to outlying cases that may be far apart.
    d. Cluster Size: Complete linkage tends to find compact clusters of approximately equal diameters. The clusters are of uniform size.
    e. Computation: Complete linkage is computationally expensive.
3. Average Linkage
    a. Distance between two clusters is the average distance between each point of one cluster to each and every point of the other cluster.



(Sum of pair-wise distances between K1 & K2) / Size of the Cluster

b. Cluster Size: Clusters tend to be relatively compact and relatively far apart.

Question 3: 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.

Answer 3 (a)

Following are the application of using PCA:

1. Image a. Recognition and b. Compression and c. Segmentation
   a. The image can be converted in numerical matrix image of dimensions NxM where each element is the intensity value of the image.
   b. Applying PCA helps extract appropriate features and helps to leave out some components without losing out much information.
   c. PCA reduces the complexity of the problem.
   d. In case of Image Compression by taking out the less significant eigenvectors the size of the image can be reduced considerably for storage.
2. Recommender Systems
   a. PCA helps in reducing the dimensions and leads to faster computation of recommendations.
   b. PCA helps improve accuracy on the Recommender Systems predictions.
3. Technical Trading (Equity Selection)
   a. Data related to stocks have very high number of variables. Principal Component Analysis can help reduce these dimensions significantly to help identify the high performing stocks.
   b. Highly beneficial to reduce investor's time and cost

Answer 3 (b)

Basis Transformation:

- Basis transformation is the process of converting information from one set of basis to another. It allows us to represent the same data in multiple basis vectors.
- Representing data in new columns different from original. That is, original features to new features. We do basis transformation for convenience and efficiency.

- PCA helps to find a new set of basis vector to represent all the points in our dataset. The basis vectors help explain the information of the dataset in the best possible way helping in operations like reducing dimensions, finding latent variables etc.
- These basis vectors form the Principal Components which are nothing but linear combinations of the basis vectors. These basis vectors are all orthogonal and thus uncorrelated with each other.

Variance as Information:

- The variables which capture variance in the data are the variables that capture the information in the data. The more variance a variable or a column has the more informative the variable or the column becomes. That is it is more helpful and more important for modelling.
- If the variable or column has less variance as compared to others, then we can delete or disregard that column and use other columns for modelling.
- The Principal Components (basis vectors) explain the variance in decreasing order.
- This building block helps in dimensionality reduction in PCA.

PCA helps in finding the best possible basis vectors set in such a way that the variation is non-uniformly distributed amongst them. This way we can determine which columns to keep and which ones to discard.

Answer 3 (c)

Following are the shortcomings of using Principal Component Analysis:

1. In Principal Component Analysis has to be with linear components. Limitation of being restricted to linearity. In some business cases, non-linear methods produce better results.
2. Principal Component Analysis requires principal components to be perpendicular, uncorrelated or orthogonal. Sometimes data requires that correlated components should be present and can present better results.
3. Principal Component Analysis assumes that if there is no or low variance, then the components are not very useful. Dropping a component with even 2% variance is not a good idea in reality for all business scenarios. E.g. in fraud cases. We don't want such small variation components to be dropped.