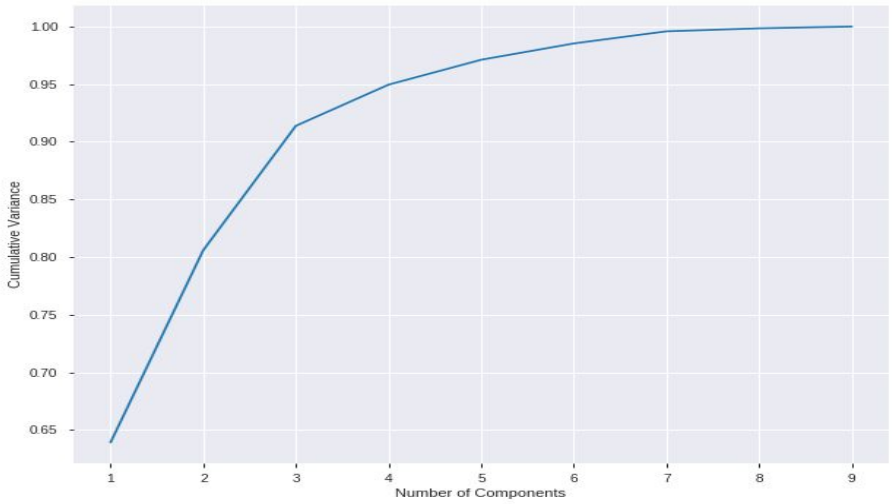
Clustering & PCA Subjective Questions

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)

Answer:
Firstly I have extracted the data and then I have changed all the variables in its desired form especially “Health”, “Imports”, and “Exports” columns as they were in %percentage.
After doing that I fixed the Outliers of each variable and removed as many outliers as possible.
Then I scaled the data and performed Principal Component Analysis.
For calculating the number of components that needs to be considered, I plotted Scree plot on the explained variance of the PCA components obtained.
After plotting the graph I considered the number of components to be 4.
THen comes the Modeling part for which I implemented KMeans, and Hierarchical Clustering.
For choosing the number of clusters for Kmeans, I plotted the Elbow curve and Silhouette score analysis based on which I considered number of clusters to be 3.

Below Graph is the Scree Plot



Then Performed all the necessary cluster profiling techniques and found out the cluster which are in dire need of aid.
After which I extracted the top 5 names of the countries from that cluster.

Similarly performed on Hierarchical clustering: both the type of linkages: Single Linkage and Complete Linkage.

Question 2: Clustering

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

Answer:

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).

- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.

- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).

- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

However, **hierarchical clustering** is usually preferable, as it is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.

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

Answer:

K-Means Clustering works on the unlabeled numerical data and automatically groups the data into clusters of data. Following are the steps:

- **Initialization**: Perform Elbow method and Silhouette score analysis and select an optimal number of clusters that is choosing "K" value, after that randomly select k initial centroids, which will be the center of the K number of clusters.

- **Cluster Assignment**: All those points which are close to chosen point will form the cluster.

- New clusters would be formed and to these clusters centers need to be assigned. Centroid would then be the mean of all the values in that clusters.

Last 2 steps keep getting repeated until all the clusters get stabilized. This is how K-Means Clustering Algorithm works.

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

Answer:
Although the exact value of "K" cannot be determined, but we can find out an optimal value through the following Statistical Method that is "Elbow Method". The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster.

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

Answer: Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms.So it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.
For Example:
scale the variables having heights in meters and weights in KGs **before** calculating the distance.

e) Explain the different linkages used in Hierarchical Clustering

Answer: There are 2 different types of Linkages:

- **Single Linkage**: This type of linkage is based on grouping of clusters in bottom-up way, at each step two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other are combined.
- **Complete Linkage**: In this type of linkage each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster.
- **Average Linkage**: 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.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

Answer:

PCA is predominantly used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression. It is also used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology, etc.

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

Answer:

Basis Transformation: change the basis of all of the vectors from our random distributions, such that they line up mostly along one or two or a small number of those axes. And what this does is to help decorrelate the components of the vector.

Variance as Information: Total variance is the sum of variances of all individual principal components.

The fraction of variance explained by a principal component is the ratio between the variance of that principal component and the total variance.

In Simple terms:
If the **new is bigger than old** then you would **divide** which may lead like your second case:
For instance: converting meter to kilometer: **1meter = 1/1000 kilometer**.
Hence your **M in this case would be 1/1000**.

If the **new is smaller than old** then you can just **simply multiply**:
For instance: converting meter to centimeter: **1meter = 100centimeter**.
Hence your **M in this case would be 100**.

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

Answer: Drawbacks of using Principal Component Analysis

- **Linearity** : PCA assumes that the principle components are a linear combination of the original features. If this is not true, PCA will not give you sensible results.
- **Large variance implies more structure** : PCA uses variance as the measure of how important a particular dimension is. So, high variance axes are treated as principle components, while low variance axes are treated as noise.

- **Orthogonality** : PCA assumes that the principle components are orthogonal.