

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 (what EDA you performed, which type of Clustering produced a better result and so on)

Answer: Main goal of "clustering of countries" case study was to identify the countries that are in severe need of aid; using health and socio-economic factors that showcases if, on the whole, a country is developed or not. The dataset is called "Country-data.csv" (here after referred to as "df" for ease) Below are the steps followed for analysis:

Data reading and understand

1. Import all libraries like numpy, pandas, matplotlib, seaborn, filter warnings, sklearn and scipy.
2. Then the data set was read using `pd.read_csv("filename")`.
3. Dataset can then be viewed in form of a dataframe with 167 rows and 10 columns. This was found using `df.shape` method.
4. Information on missing values and datatypes was checked using `df.info()`
5. Statistical description was found using `df.describe()`

Data cleaning and preparation

1. Checked for null values using `df.isnull().sum()`. This gives the sum of null values in all columns in a list form. None were found in this dataset.
2. For data preparation, three columns named "exports", "imports" and health were taken to change values from percentage of GDP to actual GDP values.

Performing exploratory data analysis

1. Performed univariate analysis on all numeric columns by defining a function. Statistical description, boxplot and distribution plot was done simultaneously for all numeric columns. Outliers were found in all columns.
2. Bivariate analysis was done between all numeric columns to see if there are any patterns. Such inferences are mentioned in the notebook.
3. Heatmap was plotted. There are features that are highly correlated to one another. For clustering, we do not drop any correlated columns as of now.
4. Outlier treatment was done by capping.

Scaling data and Hopkins score

1. Dataset scaled using `StandardScaler` from sklearn.
2. Hopkins score was checked to see if the dataset can be clustered. 0.90 score was obtained;

K-means approach

1. importing k means library from `sklearn.cluster`
2. kmeans was checked using both elbow curve method at $k = 3$ and silhouette score at $k = 2$

3. for further analysis with k means, $k = 3$ cluster was taken since that gave balanced data points between the clusters
4. cluster profiling and visualisation was done using plots
5. concluded that cluster 0 had datapoints with lowest GDP, lowest income and highest child mortality rate
6. 48 countries were present in this cluster; out of which top ten were listed based on high child_mort and lowest GDP

Hierarchical Approach

1. Importing libraries from scipy
2. simple and complete linkage was performed
3. complete linkage dendrogram was considered for further analysis as it was more readable
4. initially, dendrogram tree was cut at 3; but then there was one data point alone as a cluster.
5. cut the tree again at $k = 2$; and performed further analysis
6. cluster profiling and visualisation was done using plots
7. concluded that cluster 0 had datapoints with lowest GDP, lowest income and highest child mortality rate
8. 126 countries were present in this cluster; out of which top ten were listed based on high child_mort and lowest GDP

*Same top ten list of countries was obtained using both the methods

BASED ON HIGHEST CHILD MORTALITY RATE	BASED ON LOWEST GDPP
1. Haiti 2. Sierra Leone 3. Chad 4. Central African Republic 5. Mali 6. Nigeria 7. Niger 8. Angola 9. Congo, Dem. Rep 10. Burkina Faso	1. Burundi 2. Liberia 3. Congo, Dem. Rep 4. Niger 5. Sierra Leone 6. Madagascar 7. Mozambique 8. Central African Republic 9. Malawi 10. Eritrea

K-Means approach with number of clusters being 3 gave a better result. Not only did it give the same list of countries as the output, but also all three clusters were well balanced in terms of datapoints distribution. Final list will have 48 countries that needs immediate aid; out of which, ten are listed above based on high child mortality rate and low GDP rate.

Question 2: Clustering

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

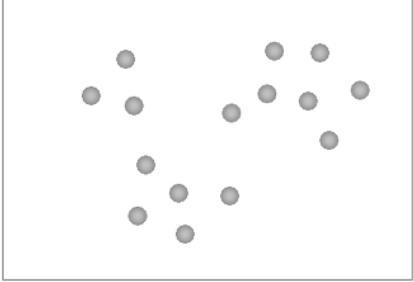
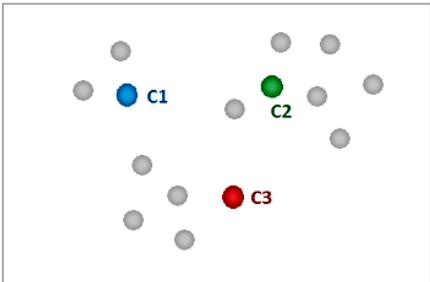
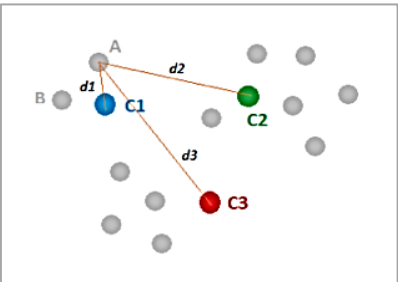
Answer: Clustering, in simple words can be described as the statistical process of dividing the data points into several groups; and each group will have datapoints that possess similar characteristics. For example, consider an airline frequent flyer program. Clustering can be used to identify different groups of customers based on the frequency of their travel; so that airline can grant them with different milage offers. Types of clustering algorithm and their differences are as follows:

K-means algorithm	Hierarchical algorithm
<i>k-means clustering is a clustering algorithm which requires prior specification on the number of clusters (depicted by k).</i>	<i>Hierarchical clustering is another form of clustering algorithm, which builds clusters in a categorized manner. No of 'k' is depicted using dendrogram.</i>
<i>K means works well when the cluster shape represents a 2D circle or 3D sphere.</i>	<i>Hierarchical clustering retain difficulty in handling convex shapes.</i>
<i>First step is to assign each point to the nearest centroid in terms of distance (Euclidean distance); and second step is to calculate average of these points which in turn becomes the new centroid. Process is iterative until there are no more changes in position of centroid.</i>	<i>Initial step of this algorithm is to assign data points to a cluster of their own. Next, nearest clusters get merged into the same cluster. This step is repeated until a single cluster is obtained.</i>
<i>Since the initial number of clusters are a random choice, results obtained by running the algorithm iteratively might differ.</i>	<i>Results obtained using hierarchical clustering are reproducible.</i>
<i>Suits with large data sets, as the computation is usually less intensive; with time complexity being $O(n)$.</i>	<i>Cannot handle larger data sets as it is computationally intensive with a time complexity of $O(n^3)$.</i>
<i>Elbow curve and silhouette scores are the evaluation methods used.</i>	<i>Hierarchical can either follow a divisive method or agglomerative method.</i>
<u>Advantages:</u> <ol style="list-style-type: none"> 1. Convergence is guaranteed. 	<u>Advantages:</u> <ol style="list-style-type: none"> 1. No prior specification of k (number of clusters) is required. 2. Ease of understanding and handling.

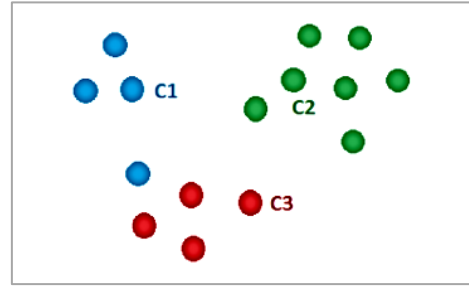
<p>2. Clusters of different size and shapes can be generalized with k means. For e.g. Elliptical clusters</p>	
<p><u>Disadvantages:</u></p> <ol style="list-style-type: none"> 1. Dependent on the initial value of k chosen at random. 2. Affected by the presence of outliers. Outliers must be either removed or capped. Otherwise, instead of being ignored, they might get a cluster of their own. 3. Clusters of different density needs to be generalized as mentioned in advantages section. 	<p><u>Disadvantages:</u></p> <ol style="list-style-type: none"> 1. Previous steps cannot be undone. 2. Sensitive to outliers. Outliers must be either removed or capped. 3. Distance metric and linkage criteria needs to be specified

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

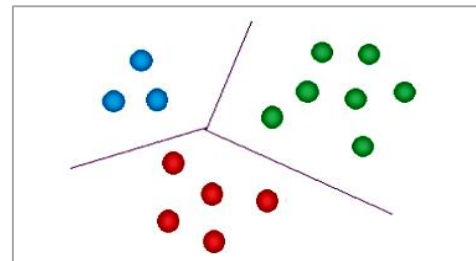
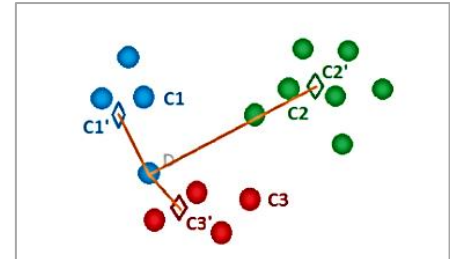
Answer: Following is the workflow of k means algorithm. (image source and credits: healthcare.ai)

<p><u>Step 1:</u> select the number of clusters by assigning a value to 'k'. for example, if the image shown needs to be clustered into 3 groups, then k is initialised with the value $k=3$</p>	
<p><u>Step 2:</u> Select centroids for each k at random. So, we get three points c_1, c_2 and c_3 as initial centroids.</p>	
<p><u>Step 3:</u> Assign each data point to their closest centroid based on the minimum distance to the cluster centre. This will result in k clusters which was predefined.</p>	

Step 4: recompute cluster centroids with the mean of formed clusters centroids. As in, in the image shown, new centroid for blue cluster can be found by dividing cluster points by total points. Similarly, do the same for other clusters and new centroids are obtained.



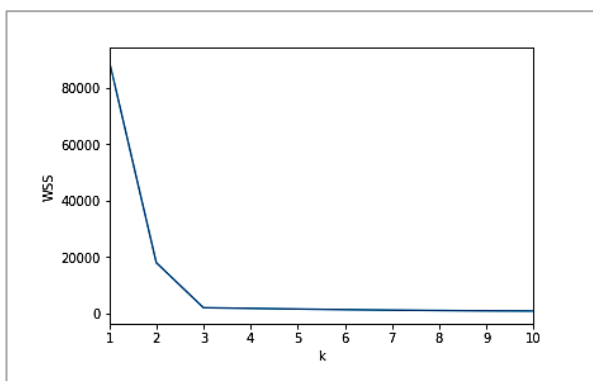
Step 5: Repeat step 3 and 4 until there is no modification to cluster centroids.



Points to note: The results obtained by k means might not be the most optimal because it converges to local optimum. The quality of the result solely depends on the initialization of centres and the number of k chosen prior to performing clustering.

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

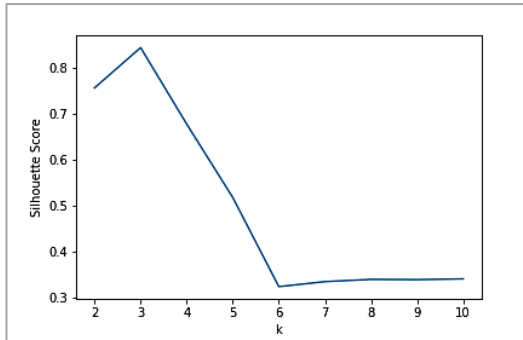
Answer: Initially, value of 'k' is set at random. But, after performing the algorithm, there are two main methods to statistically identify the optimal number of k . They are called as elbow curve and silhouette score. Once elbow curve method is performed, a plot can be obtained for WSS (within cluster sum of squared) and k . An example of such plot is shown below:



Here, it can be clearly seen that the graph looks like an arm with elbow at $k=3$. But, in some cases, there could be a curve at some other point too. Say, $k=4$. In such cases, silhouette score is used.

(image source and credits: analyticsvidhya.com)

A silhouette score measures how one data point is similar to its own cluster (i.e. cohesion) compared to other clusters (i.e. separation). Silhouette score ranges from +1 to -1. A higher score close to +1 is more desirable. More number of negative scores indicates that too many or fewer clusters are created.



Silhouette score reaches its global maximum for the optimal value of k and it is seen as a peak when plotted. Here, it can be clearly seen that the graph has a peak at $k=3$.

(image source and credits: analyticsvidhya.com)

When it comes to business aspect of k , understanding of dataset is a crucial step. This is because, only if the data and business requirement is understood, clustering can be applied correctly. For example, take an airline into consideration; and they want to group their customers based on frequent flyers to provide them with rewards/benefits. Market segmentation can be done using k means to find customer groups which are frequent flyers that are not explicitly labelled in the data. Choosing k value here must be done wisely considering the business requirement along with evaluation methods. There must be a reasonable balance.

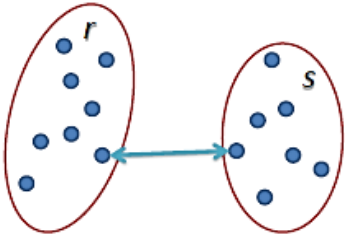
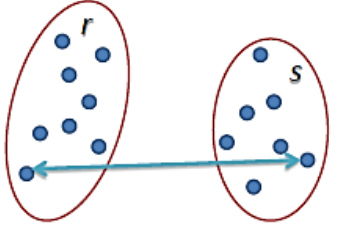
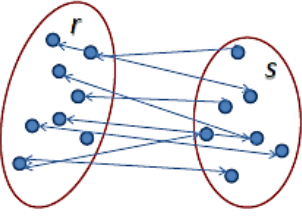
d) Explain the necessity for scaling/standardisation before performing Clustering.

Answer: It is always good to scale the data before clustering. But even more so, if the numerical data present in the dataset possess completely different scales to one another, then scaling or standardisation is necessary. For example, age and income or height and weight. This is because, if they are not scaled or standardised, then distance metric would not perform correctly as some of the columns have higher values, and clusters tend to move towards higher values. Scaling assures a better performing model.

e) Explain the different linkages used in Hierarchical Clustering.

Answer: Hierarchical clustering is another form of clustering algorithm, which builds clusters in a categorized tree structure manner. No of ' k ' is depicted using dendrogram. Initial step of this algorithm is to assign data points to a cluster of their own. Then nearest clusters get merged into the same cluster using linkage methods. Following are the different types of linkage methods.

(image source and credits: saedsayad.com)

 $L(r, s) = \min(D(x_{ri}, x_{sj}))$	<p><u>Single linkage</u>: Minimum distance between two data points of two different clusters is calculated. That is, if cluster r and s is considered, the distance between these two clusters are length of the arrow representing the two nearest data points</p>
 $L(r, s) = \max(D(x_{ri}, x_{sj}))$	<p><u>Complete linkage</u>: Maximum distance between two clusters is calculated. This method produces a tighter cluster compared to single linkage.</p>
 $L(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$	<p><u>Average linkage</u>: sum of each pair of observation in each cluster divided by the total number of pairs is calculated to get an average distance.</p>