Clustering of Countries

(Assignment Summary)

Problem Statement – HELP International, a humanitarian NGO need to group countries based on various socio-economic factors and decide which countries need AID. We have data of 167 countries which included parameters like GDP of country, import and export expenditures, child mortality rate and health spending budget of the country.

Solution Approach – Since there is no previous/historical predictive data and we must group the countries based on socio-economic factors; we decide to use clustering.

Data Preparation – We have exports, import, and health spending given as %age of GDP, we convert them to absolute values. This step is required for scaling step before clustering. We have decided to retain the % age columns as well because it can provide useful insights into countries economy.

Exploratory Data Analysis and Outlier Handling- We look at the box plots for all the variables involved. Box plot clearly show the presence of outliers in all variables. Since a lot of things are driven by GDP of the country, we start with that. We inspect the countries with very high GDP. All the factors for these countries are favorable and hence there is not much need of aid here. We can drop these as outliers. We then explore the relation between different variables. We identify variables to explore further (as mentioned in the jupyter notebook after pairplot). We identify the linear relationship income and gdp, inverse relation between GPD and child mortality.

We observe the outliers in inflation. Besides high inflation, Nigeria has low GDPP, child mortality rate is high, health spending budget is also low. We cannot remove this. But we can the outliers to 46. (It still remains the highest value but is now closer to the other elements.)

Hopkins Statistics – We compute the Hopkin's statistics. It comes out to be > 90% in multiple cases. So, we are good to proceed with clustering.

Scaling – Since the clustering algorithms are impacted by outliers and units of the variables involved, it is important to standardize the columns. We use standard scaling to scale all the columns.

Hierarchical clustering- We perform hierarchical clustering as the first step. We use both the single and the complete linkages and metric is selected as Euclidean. Dendrogram from both the linkages indicate that 3 must be a good number for number of clusters to form.

Selective outlier handling also gave a dendrogram indicating 2 clusters. Therefore, we decided to evaluate that also.

Burundi, Liberia, Congo Dem Rep, Niger, sierra leonne came out as top countries in all cases.

K-Means – For K-means we start with trying to determine the value of 'K'. This can be done using

1. Elbow-curve
2. Silhoutte score

We determine that 3 or 4 can be a good number of clusters.

Creating 3 clusters gives us a decent distribution. Profiling clusters through box and scatter plots gives a clear picture of cluster to choose countries from. When we try using 4 clusters, the additional cluster overlaps with existing clusters and does not contribute to decision making. So, k = 3 is a good number.

Final list of countries came out to be same as hierarchical. - Burundi, Liberia, Congo Dem Rep, Niger, sierra leonne

Q1. Compare and Contrast the K-Means Clustering and Hierarchical Clustering.

Clustering is an unsupervised learning technique, which creates different groups or clusters of the given set o f inputs and is also able to put a new input into the appropriate cluster. While doing clustering, the basic objective is to group input points in such a way as to maximize the inter-cluster variance and minimize the intra-cluster variance.

K-Means Clustering -

1. K-Means starts with identified initial value of 'K' and then identifying k-observations as centroids.
2. Then the iterative process starts
    a. All the data points are assigned to the nearest centroid.
    b. The new centroids for each group are calculated.

    The process goes on till the centroids converge i.e. there is no change in centroid values.

Hierarchical Clustering –

As the name suggests, it builds the hierarchy of clusters. This can be done using 2 approaches

1. Agglomerative – Every data point is its own cluster. We then start merging the clusters based on linkage between the clusters. This goes on till we have 1 single cluster.
2. Divisive – In this approach, all the data points are initially in the same cluster. We then start splitting it till each data point is in its own cluster.

    The result of Hierarchical clustering is shown using dendrogram. Number of clusters is decided by interpreting the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

Key Differences between K-Means and Hierarchical clustering

1. Hierarchical clustering is not efficient with large amounts of data. Linkage calculation is a $O(n^2)$ operation. K-means on the other hand is a linear operation in each iteration and hence more efficient on large amount of data.
2. K-means starts with random choice of initial centroids so different iterations can produce different results. On the other hand, the results of Hierarchical clustering can be reproduced across multiple runs of algorithm.
3. K-Means requires prior knowledge of K i.e. the number of clusters you want. Elbow curve method or silhouette score can be used to get a starting value of 'K' but in hierarchical, dendrogram can be used to decide the value of K.

Q2. Explain the steps involved in K-Means Algorithm.

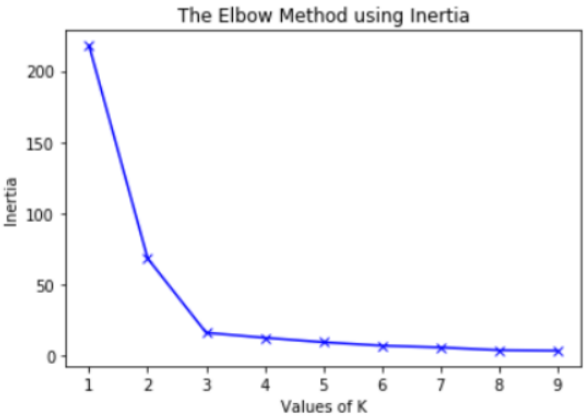
A2. K-Means is a clustering algorithm. Main steps involved in K-Means algorithm are as follows.

1. Specify the number of Clusters – 'K'.
2. Initialize the 'K' centroid points by randomly selecting datapoints.
3. Assign all the data points to 1 centroid based on Euclidean distance or other metric as decided.
4. Re-compute the centroids for each group again.
5. Iterate step 3, 4 till the centroid values do not change.

1. **Specify the number of Clusters- K**

    The first requirement is to determine the number of clusters the data has to split into. 2 most common methods to identify a suitable 'k' are:

a. **Elbow curve**- We plot the **Inertia** - sum of square distances of samples to their closest cluster center, for each value of K in ascending order. The Point where the curve smoothens is likely to be the optimal k. In image below – k=3 looks a decent one to start with.



b. **Silhouette Score -** The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

$$s(i) = \frac{b(i) - a(i)}{Max\{a(i), b(i)\}}$$

a(i) – average distance from points in own cluster. b(i) – average distance from points in another cluster. i– i^{th} data point.

We take the 'K' with highest average silhouette score.

2. **Initialize the 'K' centroid points by randomly selecting datapoints.**
We can choose the initial 'K' centroids randomly. But this choice of centroids can impact the result and there can be different results for different runs of the algorithm.
We can also use K-means++ algorithm to initialize the centroids.

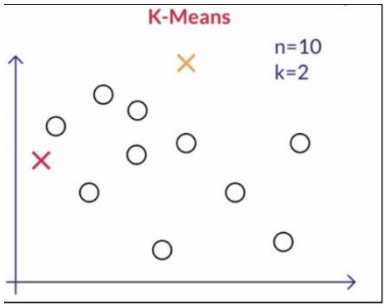
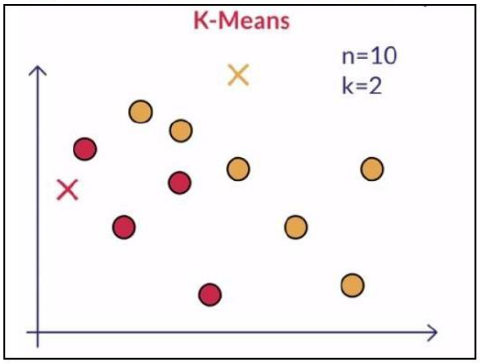


Fig 3: Choosing K random initial cluster centres

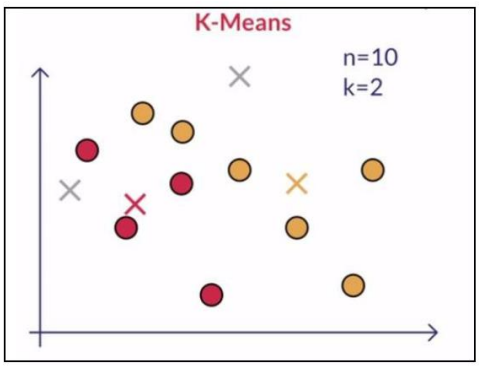
3. **Assign all the data points to 1 centroid based on Euclidean distance or other metric as decided.**

For each given point we calculate Euclidean distance between the point and all the centroids calculated earlier. Euclidean distance between 2 points p, q is given by.

$$d(p,q) = \sqrt{({q_1}^2 - {p_1}^2) + ({q_2}^2 - {p_2}^2) + ({q_3}^2 - {p_3}^2) + \cdots}$$



4. **Re-compute the centroids for each group again.**
Centroid is calculated by taking mean of each dimension of the points in given cluster.



5. **Iterate step 3, 4 till the value of centroids do not change**.

At this point we have our required clusters. We can analyze the various parameters of these clusters and make decisions regarding business significance accordingly.

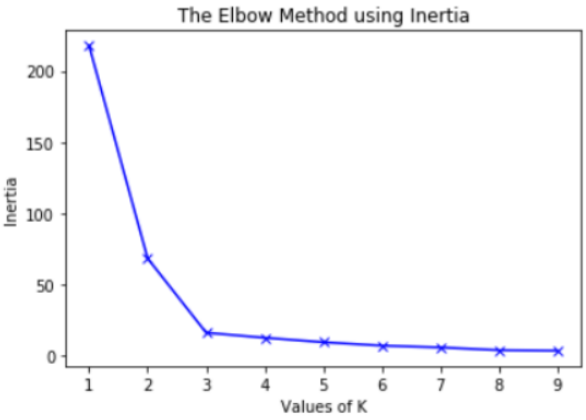
Q3. How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

A3. First Requirement in K-means Algorithm is choosing the value of K i.e. the number of clusters to be formed in the data. Clusters should follow following properties –

1. Cohesion of data points within the cluster – all the data points inside a cluster should be like each other.
2. Coupling of data points across clusters – The data points in 2 clusters should not be very similar to each other.

For example, in a retail store customer segmentation problem, the customers in each cluster should have similar buying habits or shown interest in a category of products, have a similar buying patterns and timeline etc. However, customers in different clusters should be inherently different in their characteristics.

Initial value of 'K' can be chosen using following methods



1. **Elbow curve**- We plot the **Inertia**.
   Inertia calculates the sum of square distances of samples to their closest cluster center for each cluster. Total inertia is the sum of all these values calculated above. So, Inertia is the sum of all intra-cluster distances. Hence, value of inertia should be as much minimum as possible.
   We plot the values of inertia for each value of 'K'. The Point where the curve smoothens is likely to be the optimal k. This is because further granular grouping/clustering does not reduce the inertia value significantly.
   In image below – k=3 looks a decent one to start with.

2. **Silhouette Score -** The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

$$s(i) = \frac{b(i) - a(i)}{Max\{a(i), b(i)\}}$$

   a(i) – average distance from points in own cluster. b(i) – average distance from points in another cluster. i– i^{th} data point. We take the 'K' with highest average silhouette score.

The performance of clustering algorithm is affected by choice of K. So generally, we should choose a range of k values instead of a predefined value. Number of values should be reasonably large to reflect specific characteristics of the data but significantly smaller than number of data points. In K-means algorithm, criterion is the minimization of distortion of clusters. Also, clustering is used to find irregularities in the data and to identify in which objects are concentrated. However, not every region with high concentration of objects can be considered as cluster.

In K means clustering, the distortion of a cluster is a function of the data population and the distance between objects and the cluster centre. Each cluster is represented by its distortion and its impact on entire data set is assessed by its contribution to the sum of all distortions.

It is also important, that k is decided such that on multiple runs of algorithm, the results do not change. Also, the new data points should be assigned to relevant clusters. Any distortion or over kill in cases where critical decisions are to be taken, can have critical business impact. For example, cyber profiling of people based on internet activity, identify potential frauds in insurance claims.

For a simpler situation, like online retail customer behavior problem, if we group a lot customers with high monetary contribution and high frequency to customers with customers not frequent or not contributing high monetary value, we may be running wrong campaigns for wrong target audience. This can lead to ineffective utilization of advertisement/promotional budgets and lead to very less conversion rate. So, k should be such that it is possible to

identify distinctly tell apart the traits of the data points among different clusters and same should not be much different from other points within the cluster.

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

A4. Standardization of the data means converting them into values such that mean of resulting values is 0 and SD = 1.

Formula for standardization is - $X' = \frac{X-\mu}{\sigma}$, μ is the mean and σ is the standard deviation.

Standardization does not impact the outliers in the data. It is necessary to perform the standardization in before clustering because of 2 main reasons.

1. We use Euclidean distance between the data points. $d(p,q) = \sqrt{({q_1}^2 - {p_1}^2) + ({q_2}^2 - {p_2}^2) + ({q_3}^2 - {p_3}^2) + \cdots}$

   As it is evident from the equation for the Euclidean distance, the magnitude and units of the various data points will impact the result.

   The attributes with larger range of values will weigh out the attributes with small range. Therefore, scaling down is required to avoid this.
2. The standardization also helps in making the attributes unit free and uniform.

   Example: Retail store data.

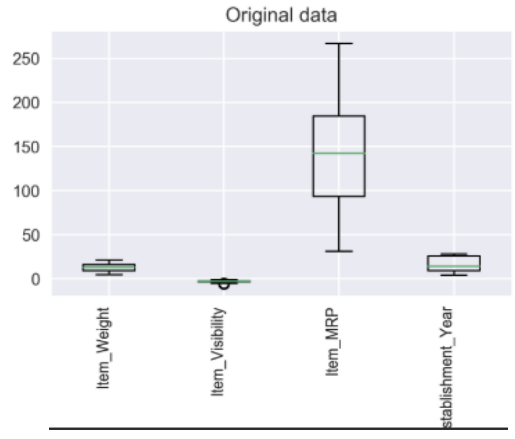
   Original statistics:

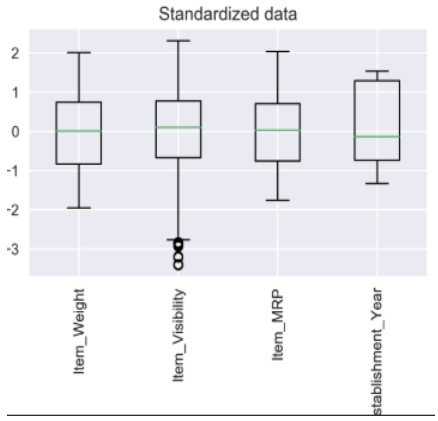
	Item_Weight	Item_Fat_Content	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Outlet_Size	Tier 2	Tier 3	Supermarket Type1	Su
count	6818.000000	6818.000000	6818.000000	6818.000000	6818.000000	6818.000000	6818.000000	6818.000000	6818.000000	6
mean	12.835420	0.355676	-2.940445	140.486413	15.154884	0.830302	0.326049	0.395571	0.651071	
std	4.233450	0.478753	0.791551	62.067053	8.389349	0.598352	0.468800	0.489009	0.476667	
min	4.555000	0.000000	-5.633875	31.290000	4.000000	0.000000	0.000000	0.000000	0.000000	
25%	9.300000	0.000000	-3.467944	93.385700	9.000000	0.000000	0.000000	0.000000	0.000000	
50%	12.857645	0.000000	-2.862535	142.179900	14.000000	1.000000	0.000000	0.000000	1.000000	
75%	16.000000	1.000000	-2.331264	184.495000	26.000000	1.000000	1.000000	1.000000	1.000000	
max	21.350000	1.000000	-1.113550	266.888400	28.000000	2.000000	1.000000	1.000000	1.000000	

   Standardized data statistics:

	Item_Weight	Item_Fat_Content	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Outlet_Size	Tier 2	Tier 3	Supermarket Type1	S
count	6.818000e+03	6818.000000	6.818000e+03	6.818000e+03	6.818000e+03	6818.000000	6818.000000	6818.000000	6818.000000	
mean	1.704754e-16	0.355676	2.342737e-16	1.233002e-16	2.051747e-17	0.830302	0.326049	0.395571	0.651071	
std	1.000073e+00	0.478753	1.000073e+00	1.000073e+00	1.000073e+00	0.598352	0.468800	0.489009	0.476667	
min	-1.956094e+00	0.000000	-3.402972e+00	-1.759459e+00	-1.329746e+00	0.000000	0.000000	0.000000	0.000000	
25%	-8.351767e-01	0.000000	-6.664603e-01	-7.589239e-01	-7.337084e-01	0.000000	0.000000	0.000000	0.000000	
50%	5.250371e-03	0.000000	9.843446e-02	2.728679e-02	-1.376709e-01	1.000000	0.000000	0.000000	1.000000	
75%	7.475728e-01	1.000000	7.696596e-01	7.091011e-01	1.292819e+00	1.000000	1.000000	1.000000	1.000000	
max	2.011410e+00	1.000000	2.308161e+00	2.036689e+00	1.531234e+00	2.000000	1.000000	1.000000	1.000000	

Let us look at some box plots. This helps us visualize how standardization can help in data reading and analyzing the data.





Q5. Explain the different linkages used in Hierarchical Clustering.

A5. Different types of linkages in Hierarchical Clustering.

1. Single Linkage:
   Distance between 2 clusters is defined as shortest distance between any 2 points in the two clusters.
   $$d(u, v) = \min\left(dist(u[i], v[j])\right)$$
   for all points 'i' in cluster u and 'j' in cluster v. This approach is also known as Nearest Point Algorithm.

2. Complete Linkage:
   Distance between 2 clusters is defined as maximum distance between any 2 points of the 2 clusters.
   $$d(u, v) = \max\left(dist(u[i], v[j])\right)$$

   for all points 'i' in cluster u and 'j' in cluster v. This approach is also known as Farthest Point Algorithm or Vor Hees Algorithm.

3. Average Linkage:

   Distance between 2 clusters is defined as average distance between every point of one cluster to every other point of another cluster.

   $$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)}$$

   for all points i, j in both clusters. |u|, |v| are cardinalities of clusters u, v respectively.

4. Weighted Linkage: Also known as Weighted Pair Group Method with Arithmetic Mean is used in Agglomerative hierarchical clustering method.
   At each step, the nearest clusters 'i', 'j' are merged to form $(i \cup j)$. The distance of new cluster to another cluster k is given by following equation:

   $$d_{(i \cup j),k} = \frac{d_{i,k} + d_{j,k}}{2}$$

5. Centroid Linkage:
   Distance between the 2 clusters is defined as the distance between centroids of the 2 clusters.
   $$d(u, v) = |c_u - c_v|$$
   c_u, c_v are the centroids of the clusters u, v.
   When the clusters are merged or a new element is added to the cluster, the new centroids are calculated with all the elements currently in the cluster. In case of merging, distance becomes the distance between centroid of merged cluster and a centroid of a remaining cluster 'v' in the forest. This can some time lead to inversion in dendrogram, as smaller clusters may be more similar to larger cluster than to their individual clusters.

6. Median Linkage: When 2 clusters say – 's', 't' are merged to form a new cluster 'u' average of centroids of 's', 't' give the centroid of new cluster 'u'. Rest calculations are done as in centroid linkage.

7. Ward Linkage: This uses ward variance minimization algorithm. This is used in agglomerative hierarchical clustering. At each step in the algorithm, we merge the clusters such that it leads to minimum increase in the within-cluster variance. Initial distance is set as squared Euclidean distance between the points.