Q1. Assignment summary

I. Importing Dataset:

- Import necessary Libraries and load the data from .csv file to a data frame.

II. Data Understanding:

- Verify various properties of data like shape, info, summary statistics and check for null values.
- "Exports, Health and imports" variables were given as percentage of "gdpp". Convert them to actual numbers.

III. Visualization of Data (EDA):

- Drop the "country" variable and perform EDA to draw valuable insights from the data.

IV. Data Preparation:

- From EDA, we have seen that "exports, health, imports, income, inflation and gdpp" have extreme values on higher side, where as "life_expec" had an extreme value on the lower side. "child_mort and total_fer" have some outliers but they are not extreme values. Hence, they are acceptable.
- To handle the outliers, we have used soft cut off and capped the extreme values so that we are not losing any data.
- After handling outliers, Hopkins check was performed. Hopkins score for our data set was high which means our data has tendency to cluster.
- Scale the data before proceeding with modelling.

V. K-Means clustering:

- Find the optimal K value using Elbow curve and Silhouette analysis.
- Build the model and assign the labels to data frame with "country" variable.
- Plot the clusters with all 3 combinations of "gdpp, child_mort and income".
- Calculate the mean of "gdpp, child_mort and income" variables for all clusters.
- Choose the cluster with high "child_mort" and low "income and gdpp".
- Extract the top 5 countries with our target label and sort the data frame using "gdpp, child_mort and income" variables in ascending, descending and ascending order respectively.
-

VI. Hierarchical clustering:

- Plot the dendrograms using Single and Complete linkages.
- Cut the dendrogram of Complete linkage and assign the labels to data frame with “country” variable.
- Plot the clusters with all 3 combinations of “gdpp, child_mort and income”.
- Calculate the mean of “gdpp, child_mort and income” variables for all clusters.
- Choose the cluster with high “child_mort” and low “income and gdpp”.
- Extract the top 5 countries with our target label and sort the data frame using “gdpp, child_mort and income” variables in ascending, descending and ascending order respectively.

Q2. Clustering

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

I. Hierarchical Clustering

- In Hierarchical clustering, we start by considering each data point as one cluster and we group them until we get 1 cluster.
- Hierarchical clustering is a linear method. Once an element is added to a branch then it cannot be added to another branch.
- Hierarchical clustering needs high computational power.
- Prior knowledge of K is not required as we can interpret the same from Dendrogram.
- Hierarchical clustering is used on small sets of data.
- Results are reproduceable.

II. K-Means Clustering

- In K-Means, we start with K random cluster centres and clusters are formed with the nearest points to centres. In 2nd step, cluster centres will move to their new position based on the clusters formed.
- K-Means clustering is a non-linear method. Once an element is added to a cluster then the same element may be added to another cluster in next iteration.
- K-Means clustering needs low computational power.
- Prior knowledge of K is required.
- K-Means clustering is used on large sets of data.
- In K-Means, we start with random choice of clusters, so we may get different results when we run the algorithm for multiple times.

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

I. **Step-1:**

- We start by choosing K initial centres. This initial choice of cluster centres is completely random.

I. **Step-2:**

- After choosing the centres, we assign each point in the data set to nearest cluster centre.
- To do this we will calculate the distance from the data point to cluster centres and allocate the data point to cluster centre with least distance.
- Euclidean distance is the common measure in calculating the distance.
- This step is called as “Assignment Step”.

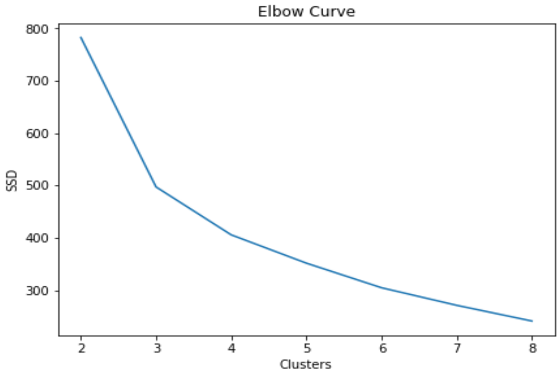
II. **Step-3:**

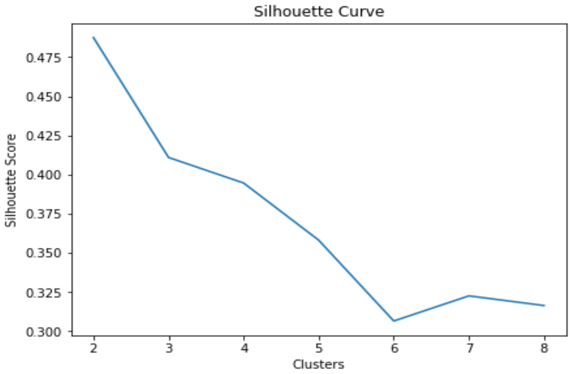
- After assigning the data points to cluster centres, the cluster centres will move to their new centres.
- New centres are calculated by taking the mean of individual points in each of the clusters.
- This step is called as “Optimization Step”.
- Now the Assignment Step is performed again to re-assign the data points to nearest cluster and Optimization Step is also performed again to calculate new cluster centres.
- We keep iterating through this process of Assignment and Optimization till the centroids no longer update.
- At this point, algorithm reached optimal grouping and we have got our K clusters.

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

I. Elbow Curve:

- First, we have to compute K-Means clustering algorithm for different values of K.
- For each K value, calculate total within Sum of squared distances (SSD).
- Plot the curve with K on x-axis and SSD on y-axis.
- In general, the bend of the curve will be considered as appropriate K value for clusters.





II. Silhouette Method:

- First, we have to compute K-Means clustering algorithm for different values of K.
- For each K value, calculate the average silhouette score using the cluster labels and scaled data.
- Plot the curve with K on x-axis and average silhouette score on y-axis.
- In general, K at peak of the curve will be considered as appropriate K value for clusters.

III. Business Aspect:

- Sometimes elbow curve is not reliable because at times the curve will tell us that there are 100 segments which is impractical as no business action can be taken based on 100 segments.
- From the above elbow curve, we can go ahead with K=3 for modelling.
- From the above figure, Silhouette curve is suggesting us to proceed with 2 clusters but 2 is not advisable as it will not give any significant insights. So, lets consider K as 3 and proceed with modelling.
- We can also consider certain Business level points, let’s say Business can only create 4 different actions that they can monitor, then we can go ahead with K=4 for modelling.

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

Scaling is a method used to normalize the range of numeric variables or features of data. In simple words, scaling means converting a value of numeric variables from 1 scale to another (common) scale.

Standardisation helps in bringing the variables which have broad and narrow scales to a common scale with mean 0 and standard deviation as 1.

In below example, we are performing cluster analysis on “Total Population” and “Mean Commute Time” variables to split all the countries in to 2 groups where Total Population is Sum and Mean commute time is Average.

When clustering was performed using raw data (fig 1), it is clear that Total Populations is the primary driver in clustering because there is a population threshold which was used to divide the data into 2 clusters.

But after standardization (fig 2), both variables have their influence in clustering the data.

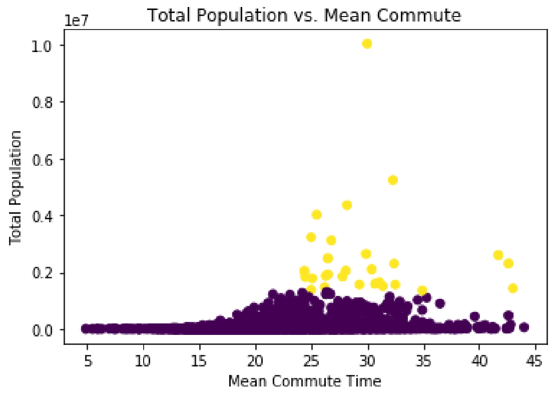


fig 1

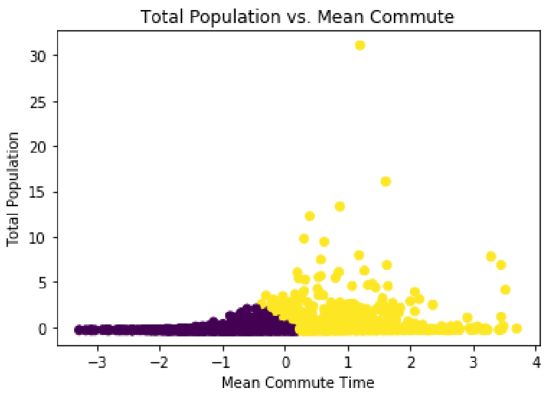


fig 2

So, standardization, prohibits the variables with broad scale from defining the clusters.

e. Explain the different linkages used in Hierarchical Clustering.

In Hierarchical clustering, we either group sub clusters which are data points in first iteration into a large cluster in bottom-up manner or dividing the large cluster into smaller sub clusters in top-down manner. In order to group them we need to measure distance between two clusters. This measure of dissimilarity between clusters is called **Linkage.**

Below are different types of Linkages:

I. **Single Linkage**:

- For 2 clusters A and B, Single Linkage considers the minimum distance between 2 points m and n where m belongs to A and N belongs to B.
- Leads to loosen clusters i.e. intra cluster variance is too high.

II. **Complete Linkage**:

- For 2 clusters A and B, Complete Linkage considers the maximum distance between 2 points m and n where m belongs to A and N belongs to B.
- Leads to stable and tight clusters.

III. **Average Linkage**:

- For 2 clusters A and B, Average Linkage considers the distance between 2 clusters defined by average distance between every point of cluster A to every point of cluster B.

We have to decide on the type of linkage which we have use. One convenient way to decide is to look at how the dendrogram looks. Usually Single Linkage type will produce dendrograms which are not structured properly, where as Complete or Average will produce clusters which have a proper tree-like structure.