Question 1: Assignment Summary

In the given Assigment Aim was to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Identify the set of countries that need immediate aid and funding based on the condition

The approach followed is below:

- Understand the data ,Perform EDA.
- Identify the number of clusters to be used .
- Apply clustering – kmeans and Heirarchial , visualize the clusters for better understanding
- Once you get clusters identify the cluster that has high values of child_moratality and low gdpp and income as these the the countries that are in need of aid
- Once the cluster is identify bin the values and sort wrt to ascending (gdpp n income) and descending (child mortality)
- Take different combination and apply to above steps.
- Based on the sorted data decide and pick Countries that need most help.

EDA Performed:

- Converting the exports , imports and health as actual values and not in percent of gdpp
- Checked for missing values (no missing value found)
- Checked for outliers (capped the outliers instead of removing them)
- Scaled the data.

Both clustering and hierarchical clustering yields similar results

Question 2: Clustering

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

K-means Clustering	Hierarchal Clustering
K Means clustering can handle big data.	Hierarchical clustering can't handle big data.
Time complexity of K Means is linear i.e. O(n)	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.
In K-Means Clustering, we need to iterate the model to find out the optimal number of Clusters	In Hierarchical Clustering, it automatically gives result at various number of Clusters.

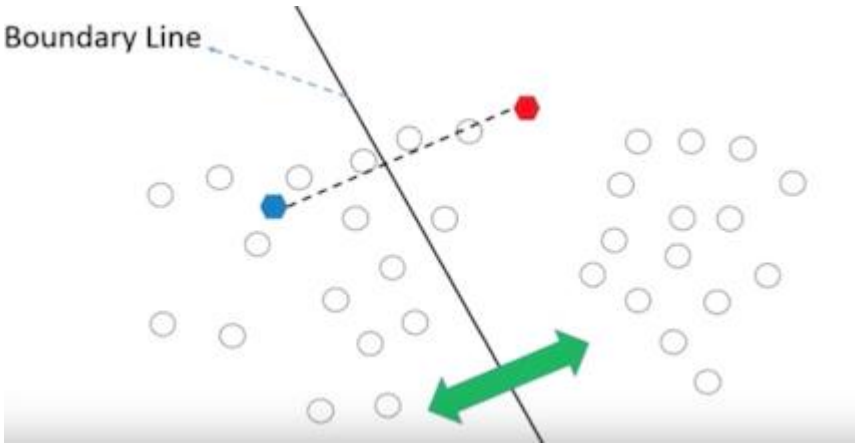
K Means is found to work well when the shape of the clusters is hyper spherical	This approach can separate non-elliptical shapes as long as the gap between the two clusters is not small
Represented by a centroid diagram	The Hierarchical clustering Technique can be visualized using a Dendrogram.

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

k-means algorithm is composed of 3 steps:

Step 1: Initialization
The first thing k-means does, is randomly choose K examples (data points) from the dataset (the 4 green points) as initial centroids and that's simply because it does not know yet where the centre of each cluster is. (a centroid is the centre of a cluster).

Step 2: Cluster Assignment
Then, all the data points that are the closest (similar) to a centroid will create a cluster. If we're using the Euclidean distance between data points and every centroid, a straight line is drawn between two centroids, then a perpendicular bisector (boundary line) divides this line into two clusters.

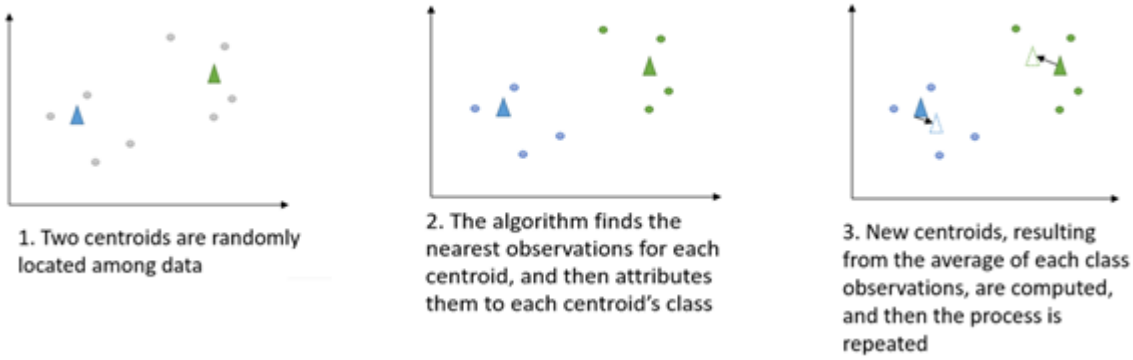


Step 3: Move the centroid
Now, we have new clusters, that need centres. A centroid's new value is going to be the mean of all the examples in a cluster.
We'll keep repeating step 2 and 3 until the centroids stop moving, in other words, K-means algorithm is converged.

Here is the k-means algorithm:

```
randomly chose k examples as initial centroids
while true:
    create k clusters by assigning each
        example to closest centroid
    compute k new centroids by averaging
        examples in each cluster
    if centroids don't change:
        break
```

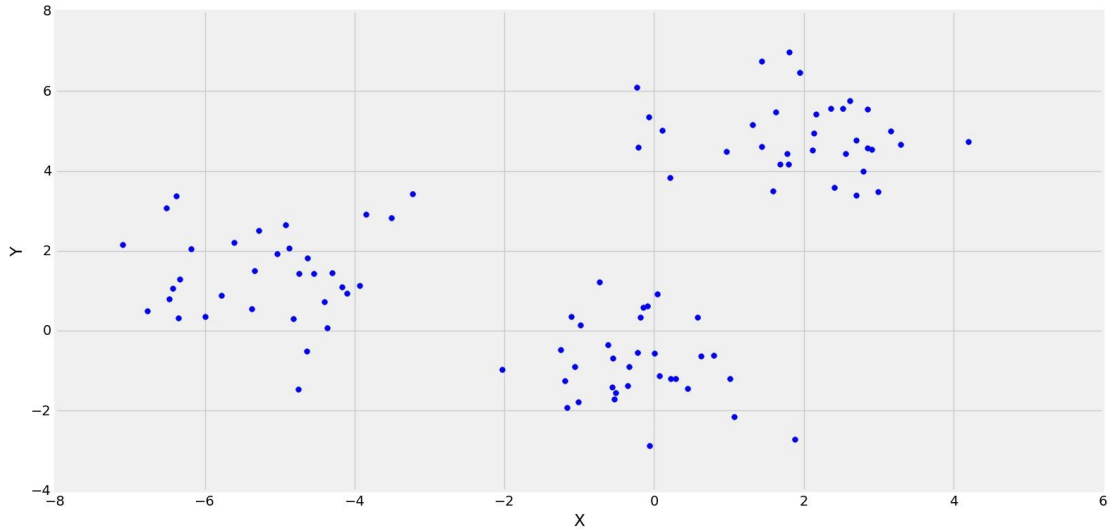
K-means is a fast and efficient method, because the complexity of one iteration is k*n*d where k (number of clusters), n (number of examples), and d (time of computing the Euclidian distance between 2 points).



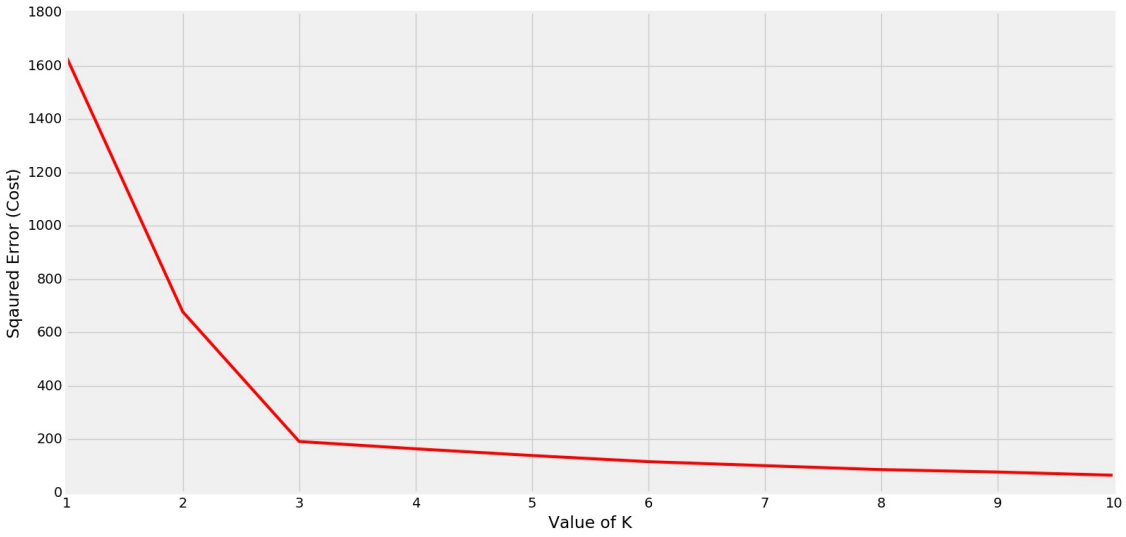
c) How is value of 'K' chosen in K-means clustering? Explain both the statistical as well as business aspect of it.

Statistical Aspect:
1)Elbow Point

There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. 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. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.



In the above figure, its clearly observed that the distribution of points are forming 3 clusters. Now, let's see the plot for the squared error (Cost) for different values of K.



Clearly the elbow is forming at K=3. So, the optimal value will be 3 for performing K-Means.

In case, it is still not clear, we try different values of k, we evaluate them, and we choose the best k value.

```
best = kMeans(points)
for t in range(numTrials):
    C = kMeans(points)
    if dissimilarity(C) < dissimilarity(best):
        best = C
return best
```

Dissimilarity(C) is the sum of all the variabilities of k clusters
Variability is the sum of all Euclidean distances between the centroid and each example in the cluster.
Or you can take a small subset of your data, apply hierarchical clustering on it (it's a slow clustering algorithm) to get an understanding of the data structure before choosing k by hand.

2: Silhouette Analysis

$$\text{silhouette score}=p-qmax(p,q)\text{silhouette score=p−qmax(p,q)}$$

pp is the mean distance to the points in the nearest cluster that the data point is not a part of
qq is the mean intra-cluster distance to all the points in its own cluster.

Central African Republic value of the silhouette score range lies between -1 to 1.

- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

Business Aspect:

Business aspect is how your clusters should be. Similar in the same cluster and different from other clusters.
Basically the clusters should be different enough from each other and lead to business insights , and point inside cluster should be similar

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

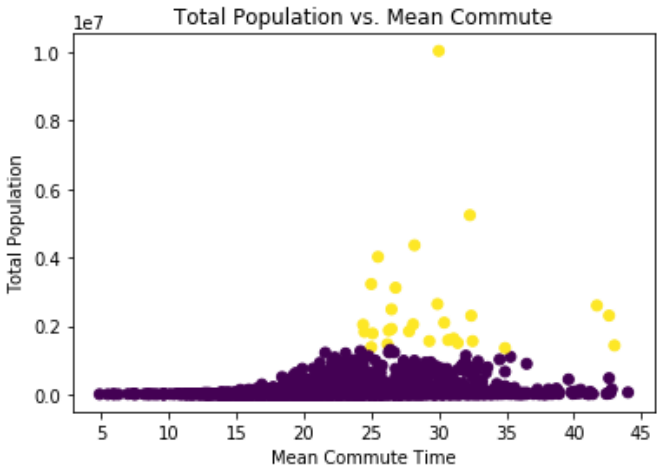
Standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

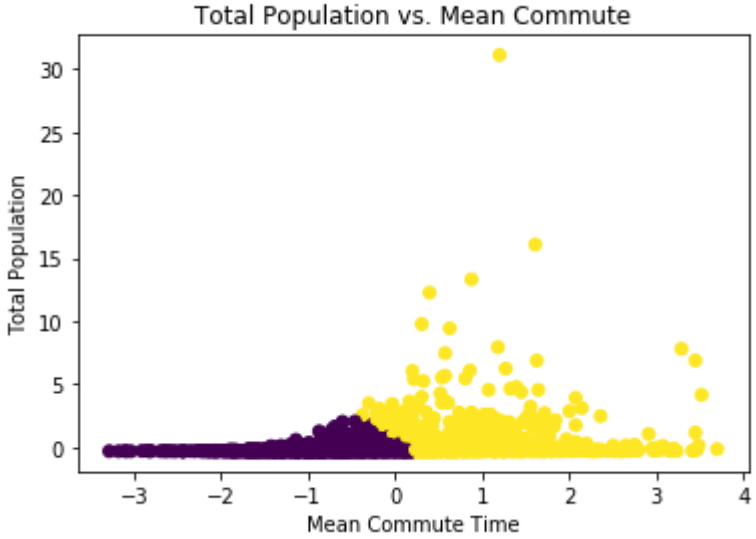
What follows is a couple examples demonstrating how standardization may impact a clustering solution.

In our first example, we are interested in performing cluster analysis on Total Population and Mean Commute Time. We would like to use these two variables to split all the counties into two groups. The units (number of people vs. minutes) and the range of values (85 - 10038388 people vs. 5 - 45 minutes) of these attributes are very different. It is also worth noting that Total Population is a sum, and Mean Commute Time is an average.

When we create clusters with the raw data, we see that Total Population is the primary driver of dividing these two groups. There is an apparent population threshold used to divide the data into two clusters:

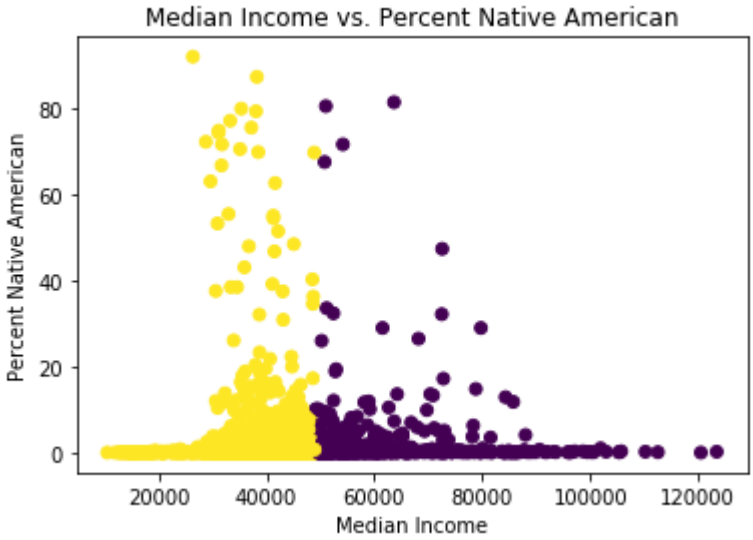


However, after standardization, both Total Population and Mean Commute seem to have an influence on how the clusters are defined.

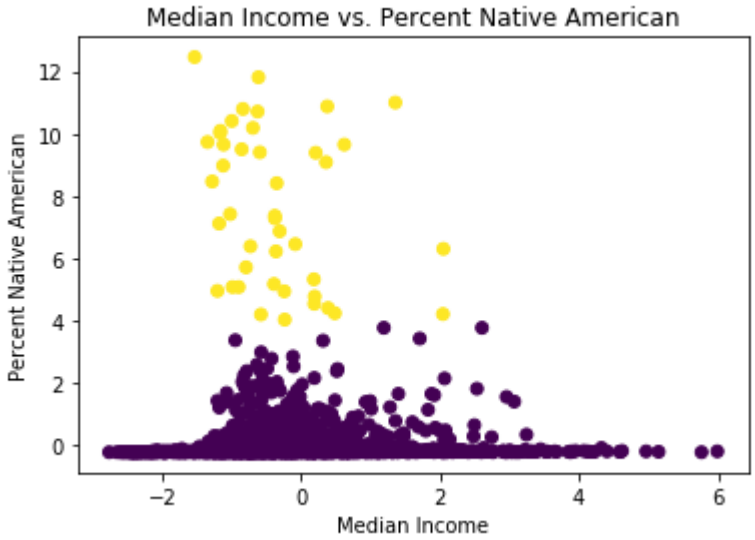


In this next example, we are interested in clustering on Median Income and Percent of the Population that is Native American (by county). Median Income is measured in dollars and represents the "middle" income for a household in a given county, and Native American is a percentage of the total population for that county. Again, the units and ranges of these variables are very different from one another.

When we perform cluster analysis with these two variables without first standardizing, we see that the clusters are primarily split on Income. Income, being measured in dollars, has greater separation in points than percentages, therefore it is the dominant variable.



When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales, the Percent Native American variable more significantly contributes to defining the clusters.



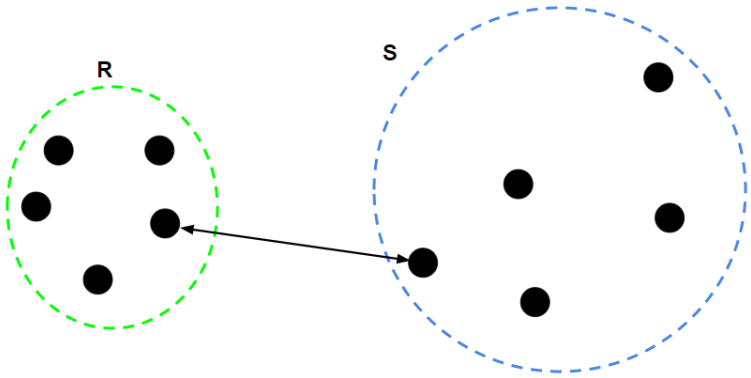
Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

e) Explain the different linkages used in Hierarchical Clustering.

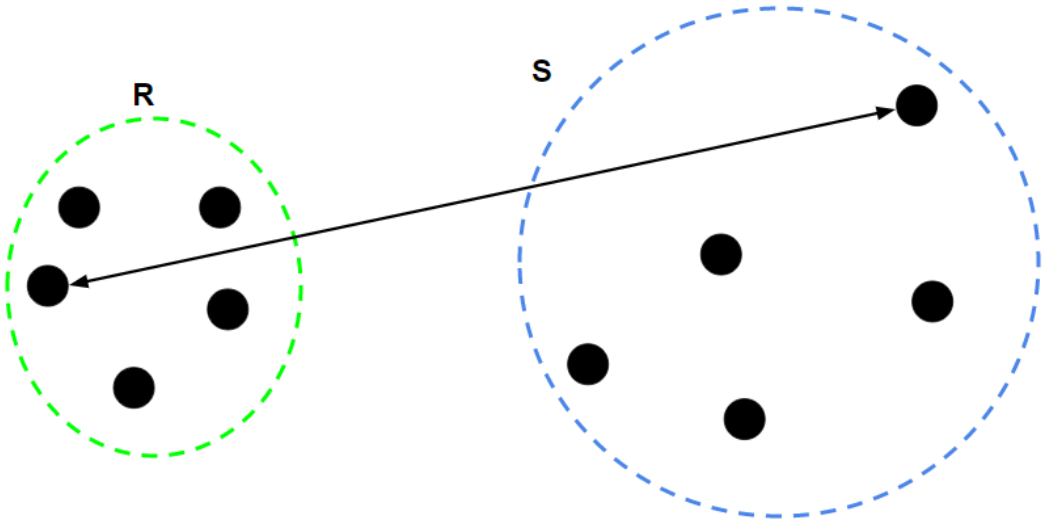
The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages

describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are: -

1. Single Linkage: For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.
   In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest minimum pairwise distance). Single-link clustering can also be described in graph theoretical terms. If dn is the distance of the two clusters merged in step n, and G(n) is the graph that links all data points with a distance of at most dn, then the clusters after step n are the connected components of G(n). A single-link clustering also closely corresponds to a weighted graph's minimum spanning tree.



2. Complete Linkage: For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S. In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest maximum pairwise distance). Complete-link clustering can also be described using the concept of clique. Let dn be the diameter of the cluster created in step n of complete-link clustering. Define graph G(n) as the graph that links all data points with a distance of at most dn. Then the clusters after step n are the cliques of G(n). This motivates the term complete-link clustering.



3. Average Linkage: For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean. Average-link (or group average) clustering (defined below) is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.