Assignment: Part II

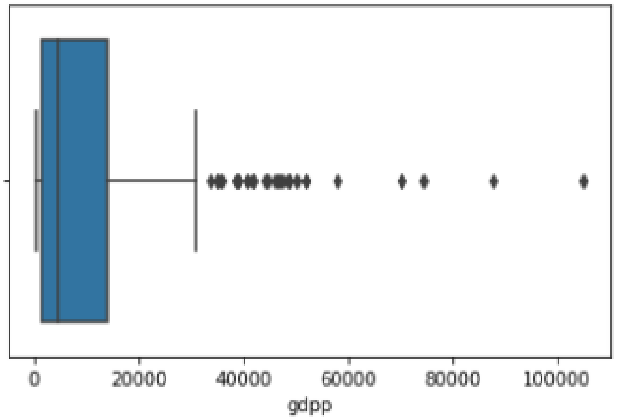
Question 1: Assignment Summary

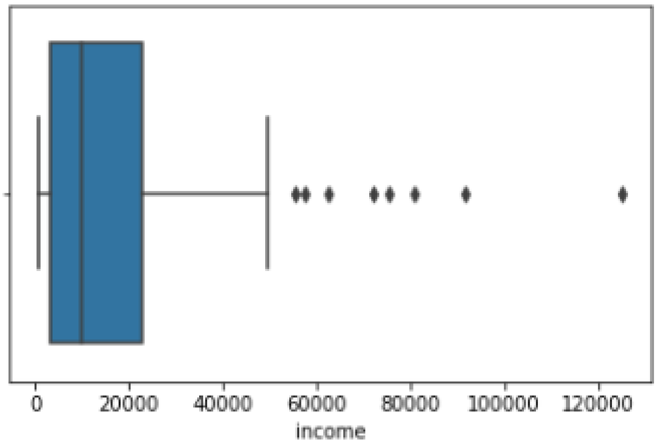
Data reading and EDA

- The dataset is read intially using read_csv function and is seen to have 167 rows and 10 columns.
- The data given has no null columns
- A barplot is done for univariate analysis for all the variables is done to know the trend watch.
- A boxplot is graphed to see the outliers and where most of the values fall. We see that **gdpp** and **income** has a number of outliers (which means that there are some countable number of countries with gdp, as well as income more than the median countries - rich and developed countries on intuition analysis)
- From the data dictionary we figure out that **imports**, **exports** and **health** spending is given in terms of percentage of gdpp. So, to find the actual value, we mutiply it with gdpp. We don't use it in terms of gdpp value because the gdpp might be different for different country and that leads to variation from the original value.

Outlier treatment and analysis

- We are required to take into consideration the (low income, low gdpp, high child_mort) metrics into consideration for this client specifically.
- Outlier treatment of (high income, high gdpp, low child_mort) can be considered as outlier. We see that gdpp, income, exports, etc have a significant number of outliers.





- Since a method called outlier capping can done instead of removing the outlier completely, it is considered in this analysis. Outlier capping does consider the datapoint as within the quantile range so that it is removed in the process. Hence, we are able to preserve as many countries and continue with the analysis.

Hopkins Test Result

- Hopkins' test tells us the tendency of the cluster formation of the dataset. If the value is less than 50% then it is not possible to have the dataset agglomerated into clusters.
- Since, we have around 85-90% (on several trials), our dataset is fit for cluster formation

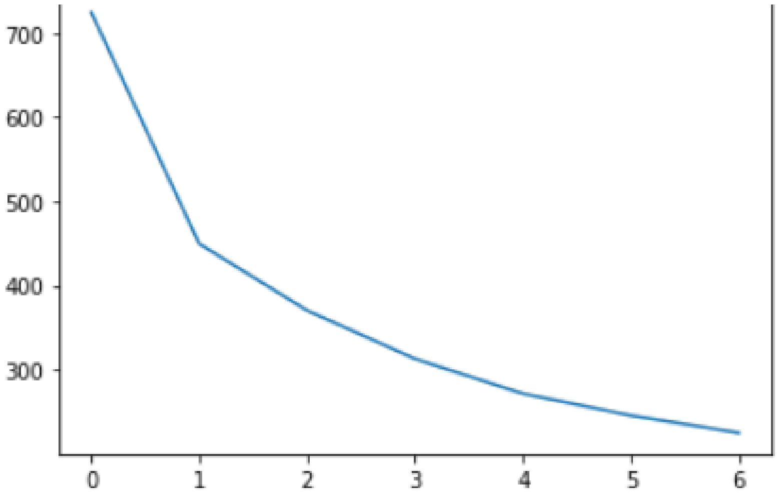
Standardization

- It is required that we do standardization before we do K- means clustering because of fairly two reasons - the intra-cluster assignment may be high and it might not be possible for it to form a cluster, and the tendency of misclassification is too high in the process.

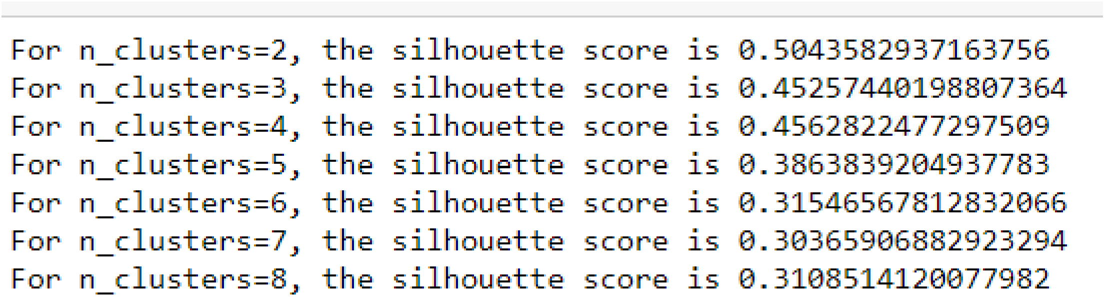
- StandardScaler().fit_transform(data) is used to standardise all the data

Finding the optimal number of clusters

- Finding the optimal number of clusters is done in two ways here - One is SSD or elbow curve method. The other is silhouette analysis. Elbow curve is minimising intra cluster distance to maximising intercluster distance, while silhouette analysis gives a value that measures the effectiveness of a given cluster
- The elbow curve forms an 'elbow' at index = 1 (i.e. number of clusters = 3). The picked value here is where the elbow forms. The elbow is also formed at index = 3(i.e., number of clusters = 5) but there is a slight decreasing trend at 5, but higher decreasing trend at clusters = 3.



- Silhouette analysis tells the effectiveness. Higher the value, higher is the formation of effective clusters. We see that the silhouette score is higher when the number of clusters = 2. Second in the line is 3 and 4, which have almost similar scores. But choosing 2 clusters is generally not preferred by business, and hence we choose the number of clusters as 3



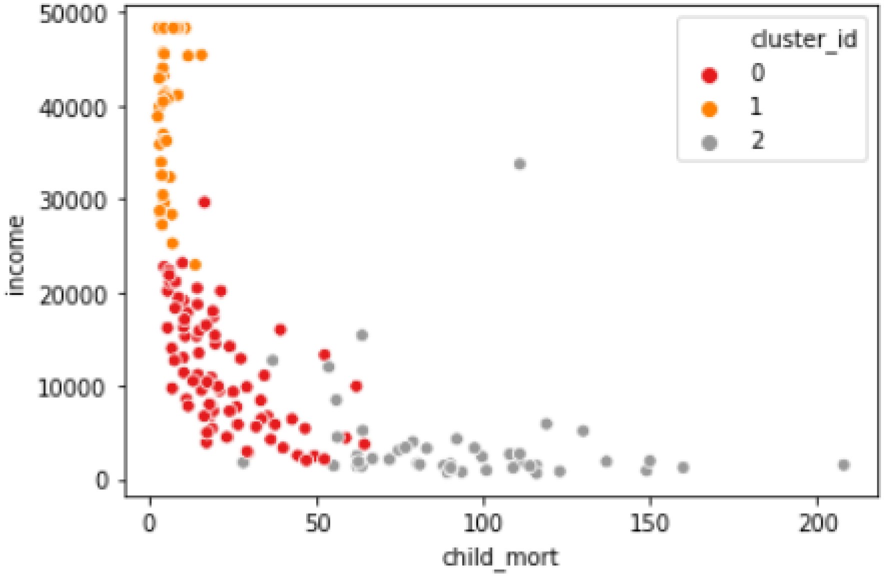
```
For n_clusters=2, the silhouette score is 0.5043582937163756
For n_clusters=3, the silhouette score is 0.45257440198807364
For n_clusters=4, the silhouette score is 0.4562822477297509
For n_clusters=5, the silhouette score is 0.3863839204937783
For n_clusters=6, the silhouette score is 0.31546567812832066
For n_clusters=7, the silhouette score is 0.30365906882923294
For n_clusters=8, the silhouette score is 0.3108514120077982
```

K- means and heirarchical clustering

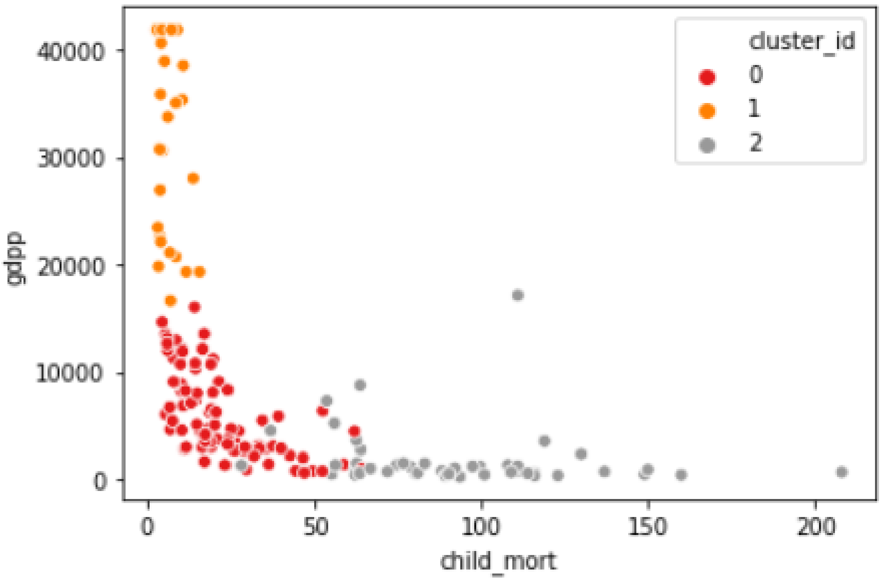
- K- means clustering requires a pre-specific number of clusters for it to decide the number of centroids. 3 is chosen as the number of clusters from above methods, and fed to K-means algorithm. It creates 3 clusters which is effectively seen in the scatter plots between gdpp, income and child_mort
- Heirarchical clustering does not require a prespecified number of clusters. It creates a heirarchy diagram called a dendogram, with linkages to the nearest cluster. Heirarchical clustering can be of two types - Single and complete heirarchical clustering. Complete heirarchical clustering chooses the highest amount of distance between the two clusters and hence, as you can see from the graph, complete clustering has distinct linkages - and hence, is chosen.

Plotting gdpp, income and child_mort to see the clusters

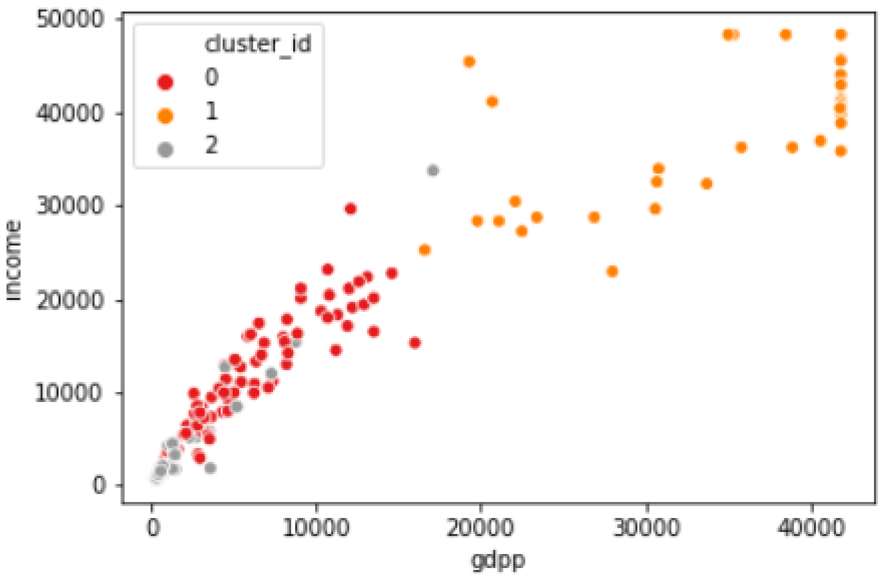
- To see the distinct division of the clusters, a simple scatterplot can be created to show how clusters are formed with a plot of gdpp against income, gdpp against child_mort and income against child_mort
- Here, seaborn scatterplot is used with hue given to the different clusters that are formed out of k-means clustering
- Child_mort vs income



- child_mort vs gdpp

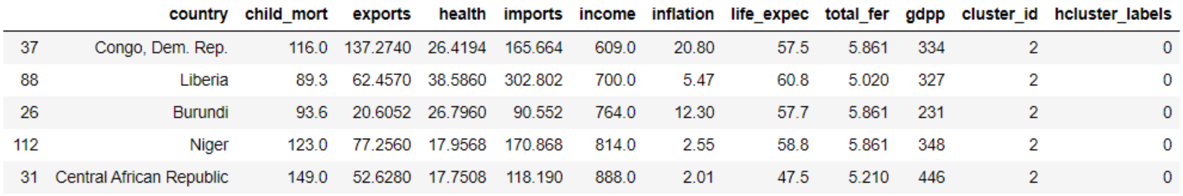


- gdpp vs income



Finding the countries that are in requirement of the aid by the NGO

- Last step is to evaluate which cluster contains the countries that need the aid. For this to be solved, we take income and gdpp in the ascending order, so as to consider the low gdpp and low income countries, as well as the child_mort in the decreasing order so as to have the high child_mort rate first.
- Taking the head() of this will helps us get the 5 countries which are in requirement of the aid



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	hcluster_labels
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	5.861	334	2	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.020	327	2	0
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	5.861	231	2	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	5.861	348	2	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.210	446	2	0

Question 2: Clustering

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

K-means Clustering:

- k-means uses a pre-specified number of clusters and assigns records to each cluster to find the mutually exclusive cluster based on distance.
- K Means clustering needs K pre-specified i.e. no. of clusters one want to divide your data. (either based on the knowledge about the dataset or, the business needs)
- One can use median or mean, or any central tendency as a cluster centre to represent each cluster.
- Methods used are normally less computationally intensive and are suited with very large datasets.
- Since the algorithm starts off with random choice of clusters, the results produced by running it many times may differ.
- K- means clustering a simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset).
- K Means clustering is found to work well when the structure of the clusters is hyper spherical (like circle in 2D, sphere in 3D).

Hierarchical Clustering :

- Hierarchical methods can be either divisive or agglomerative.
- In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.
- Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
- Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy
- In Hierarchical Clustering, results are reproducible in Hierarchical clustering
- A hierarchical clustering is a set of nested clusters that are arranged as a tree.
- Hierarchical clustering don't work as well as, k means when the shape of the clusters is hyper spherical.

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

- Step 1 : Initialization - First thing in K-means clustering algorithm is to assign random centroid to the clusters, choosing K number of centroids as designated by business problem.
- Step 2: Cluster Assignment - After initialization of the centroids, then the data points close to that centroid forms a cluster. If we use the euclidian distance as a measure to calculate the distance between the centoid and the data point, then distance is calculated and the minimum distance is used in assigning the datapoint to that particular cluster
- Step 3: Recalculation of Centoid - After the datapoints converge to form a cluster, the centroid is recaculated for the newly formed cluster. This created new centoid is the mean of all the newly assigned datapoints belonging to a particular cluster
- All the above steps are iterated over and over again till they converge, i.e., till the calculated centroid doesn't change in value or reamins almost the same.

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

- The basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to
  be as small as possible.The Elbow method looks at the total WSS as a function of the number
  of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.
- Average silhouette method briefly measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.
- These above methods are the **statistic methods** of finding out the value of K in accordance to the higher inter-cluster distance and low intra-cluster distance. The **business aspect** usually deals with the requirement of the client. Sometimes, the requirement might be to have a certain number of

clusters to understand the business problem. Then the clustering assignment would involve that number of clusters.

- For example, if a retailer asks to classify their customers into 3 clusters. He has to classify the customers as always buying, mostly buying and very rarely buying from their store, that client would specify to have 3 as the number of customers in a general sense.

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

- Standardizing either input or target variables tends to make the training process better behaved by improving the numerical condition of the optimization problem and ensuring that various default values involved in initialization and termination are appropriate. Standardizing targets can also affect the objective function.
- Here, we are more worried about the objective function of reducing the intra-cluster distance and increasing the inter-cluster distance
- Had it been all the variables in their own values, then the datapoints are bound to be far apart and can't be aggregated as a clusters. There are also possibilities of being misclassified as different cluster because of the high intra-cluster distances. The possible disadvantages of not standardizing is increased computational time to coverge and misclassifications.

e) Explain the different linkages used in Hierarchical Clustering.

- 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.

$$L(R,S) = min(D(i,j)), i\epsilon R, j\epsilon S$$

  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.

$$L(R, S) = max(D(i, j)), i\epsilon R, j\epsilon S$$

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.

$$L(R, S) = \frac{1}{n_R + n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i, j), i\epsilon R, j\epsilon S$$

where

nr – Number of data-points in R

ns – Number of data-points in S