Clustering Assignment

Problem statement

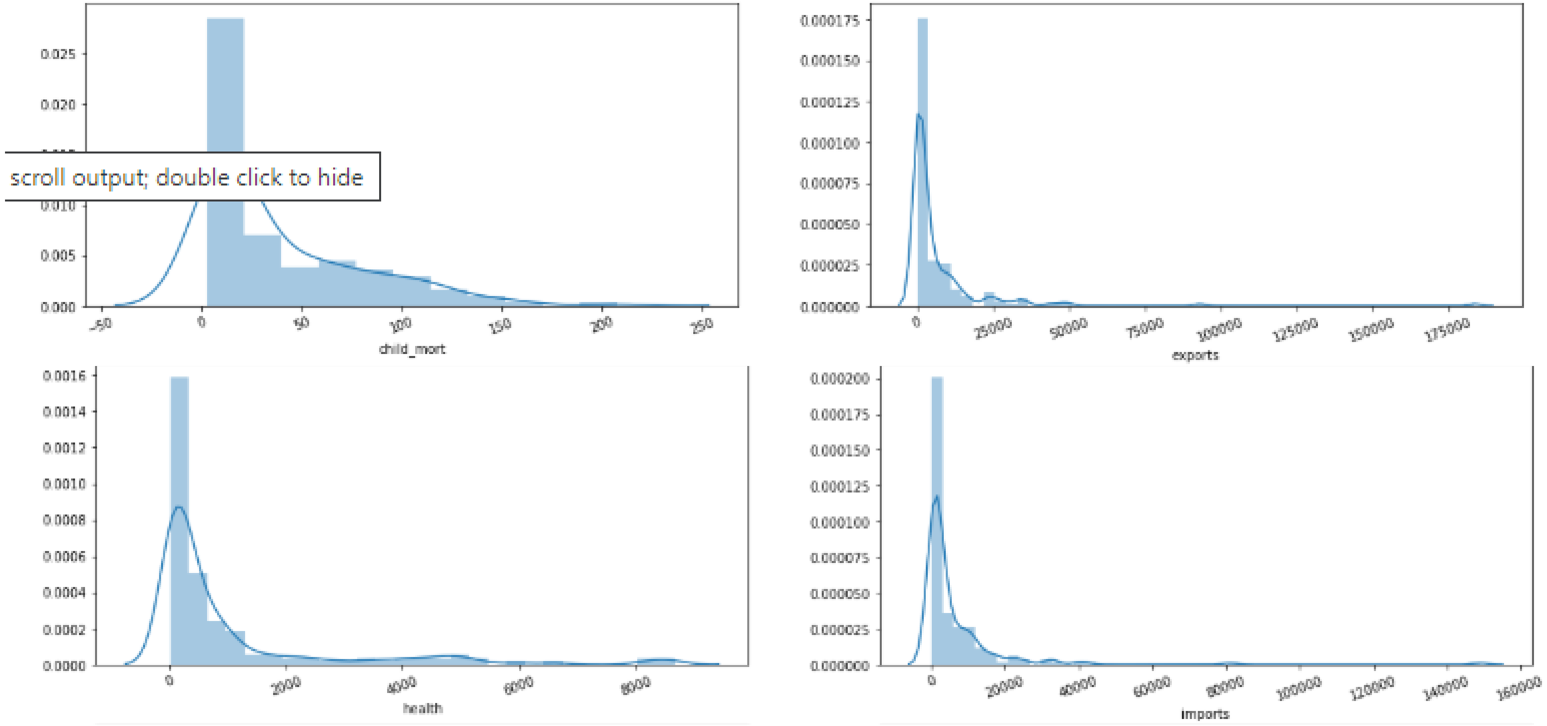
- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.
- After the recent funding programs, they have been able to raise around $10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
- As a Data analyst I'm asked to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then need to suggest the countries which the CEO needs to focus on the most.

Approach used in the analysis

- Reading, understanding and cleaning the data
- Visualizing the data by performing EDA
- Outlier Treatment
- Hopkins test to see if the data is good for clustering
- Preparing the data for modelling by Scaling
- Clustering with K-Means Clustering technique
- Performed Hierarchical Clustering
- Cluster Profiling
- Conclusion

Data Visualization

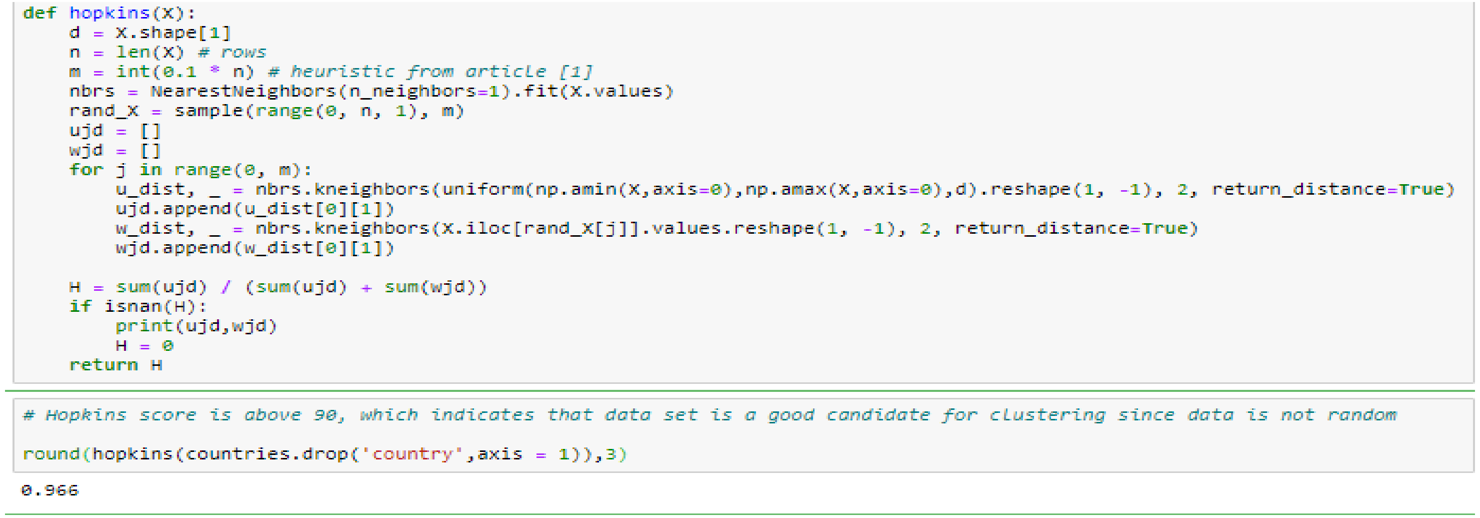
- Did Univariate and Bivariate analysis on the data.



It can be noted that child mortality is normally distributed whereas other attributes are not. This makes the data suitable for clustering.

Outlier treatment & Hopkins Test

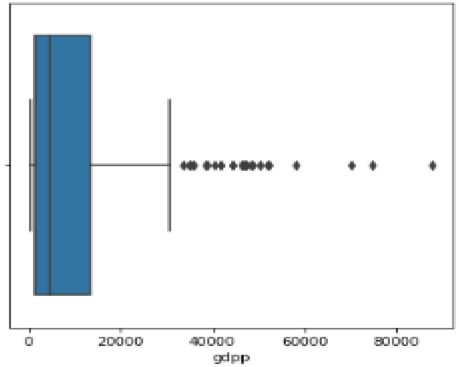
- Outlier treatment is performed by keeping boundaries for quantiles as 1 & 99th Percentile and not in a extreme way as it could remove many countries and that is not right for the analysis. Outliers from Gdpp and income gets removed. IQR method is used here for the same.
- Hopkins Test gave the score of above 95% which ensures that data is good for clustering.

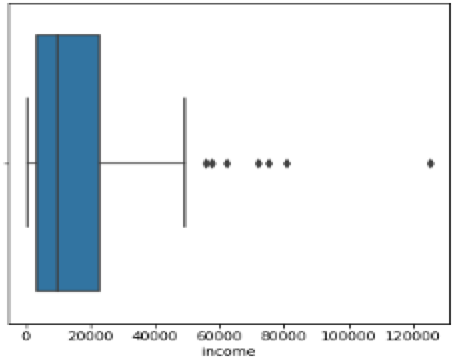


```
def hopkins(X):
    d = X.shape[1]
    n = len(X) # rows
    m = int(0.1 * n) # heuristic from article [1]
    nbrs = NearestNeighbors(n_neighbors=1).fit(X.values)
    rand_X = sample(range(0, n, 1), m)
    ujd = []
    wjd = []
    for j in range(0, m):
        u_dist, _ = nbrs.kneighbors(uniform(np.amin(X,axis=0),np.amax(X,axis=0),d).reshape(1, -1), 2, return_distance=True)
        ujd.append(u_dist[0][1])
        w_dist, _ = nbrs.kneighbors(X.iloc[rand_X[j]].values.reshape(1, -1), 2, return_distance=True)
        wjd.append(w_dist[0][1])

    H = sum(ujd) / (sum(ujd) + sum(wjd))
    if isnan(H):
        print(ujd,wjd)
        H = 0
    return H
```

```
# Hopkins score is above 90, which indicates that data set is a good candidate for clustering since data is not random

round(hopkins(countries.drop('country',axis = 1)),3)
```

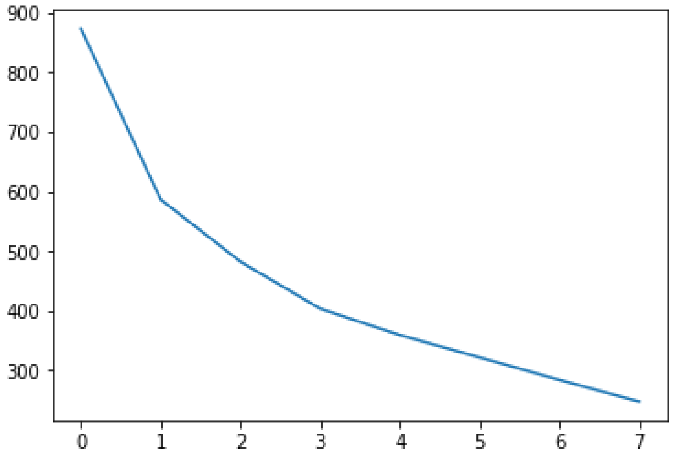
```
0.966
```

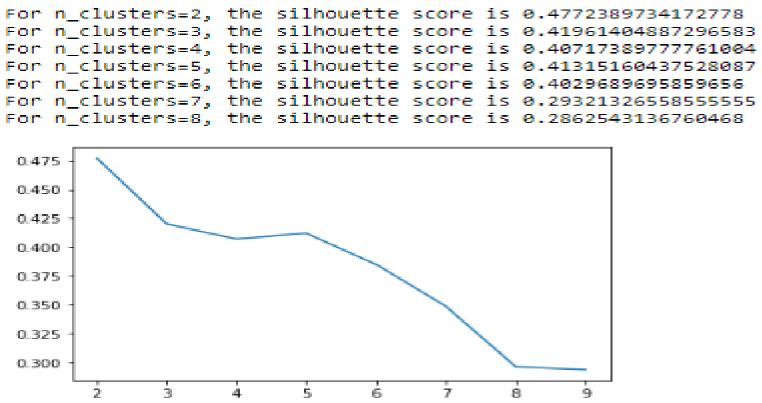




K-Means Clustering

- Elbow Curve and Silhouette Score gives an idea of number of clusters required for the analysis in K-Means Clustering.





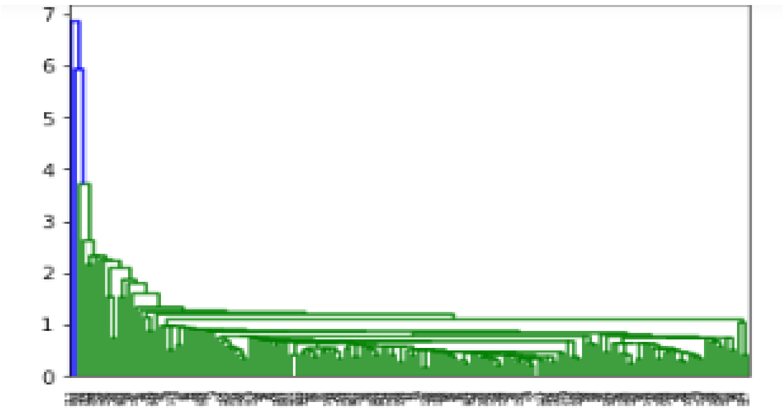
```
For n_clusters=2, the silhouette score is 0.4772389734172778
For n_clusters=3, the silhouette score is 0.41961404887296583
For n_clusters=4, the silhouette score is 0.40717389777761004
For n_clusters=5, the silhouette score is 0.41315160437528087
For n_clusters=6, the silhouette score is 0.4029689695859656
For n_clusters=7, the silhouette score is 0.29321326558555555
For n_clusters=8, the silhouette score is 0.2862543136760468
```

1. Elbow Curve – Cluster with Index 1 i.e 3 seems to be a good choice as the value of SSD doesn’t decrease significantly

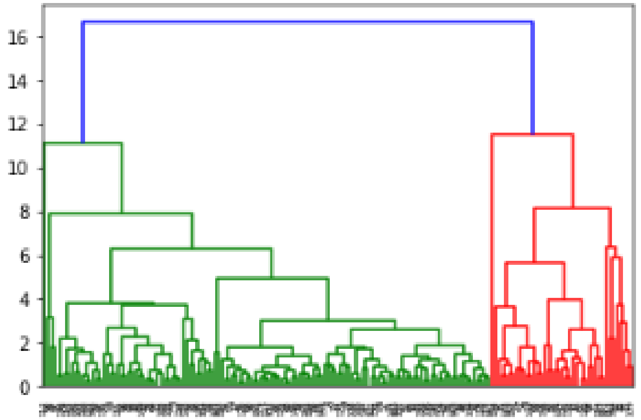
2. Silhouette Score also imply that 3 is the optimal number of cluster as from 2 to 3, there is a certain difference in score.

Hierarchical Clustering

- Hierarchical Clustering is performed using Single and Complete Linkage



Dendrogram created using Single HC is not structured correctly. This happens because clusters are solely based on Neighborhood proximity



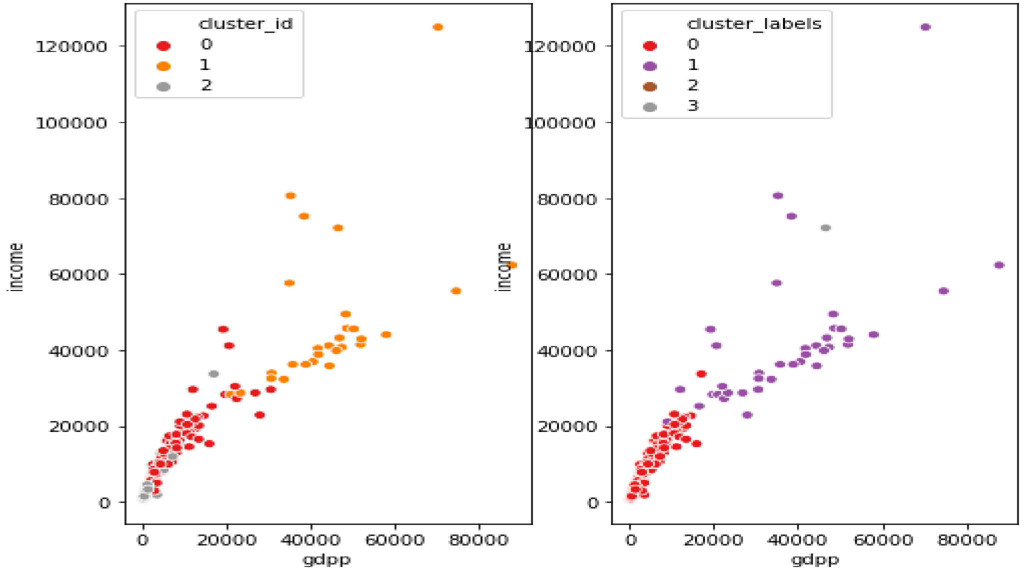
Complete linkage shows 2,4 as the good number of clusters. Business team wants to go ahead with 4 clusters. So choosing 4 from Hierarchical Clustering Method.

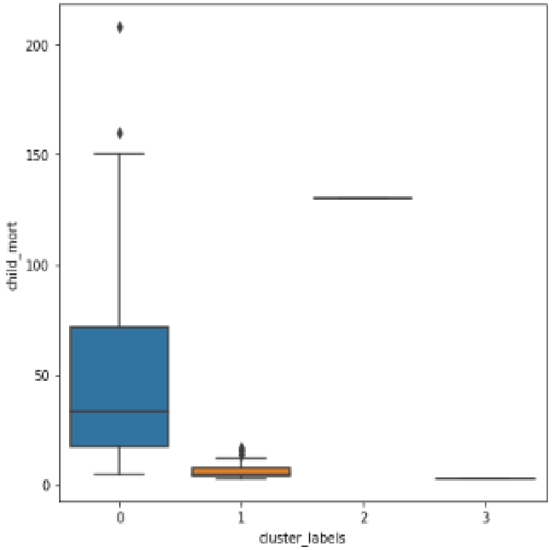
K-Means vs Hierarchical Clustering Visualization

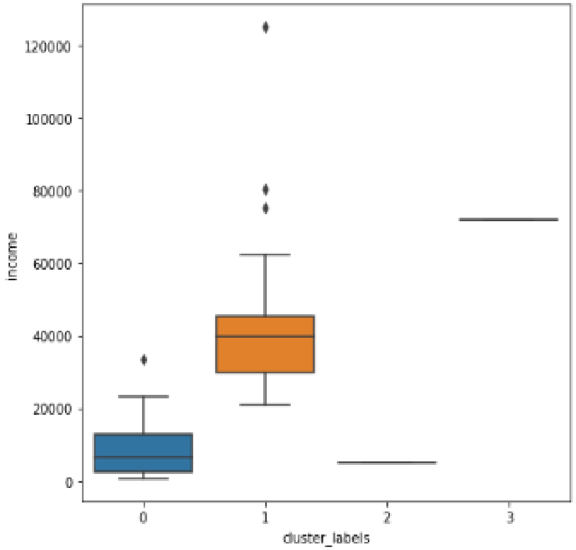
3 clusters with k-means have created a good cluster spread of almost non overlapping clusters

4 Clusters described by Hierarchical clustering are not evenly distributed, clusters 0 and 1 have the dominating effect.

Choosing number of clusters from Hierarchical Clustering - have resulted in badly distributed clusters. Thus choosing optimal number of clusters as 3 given by K-Means.

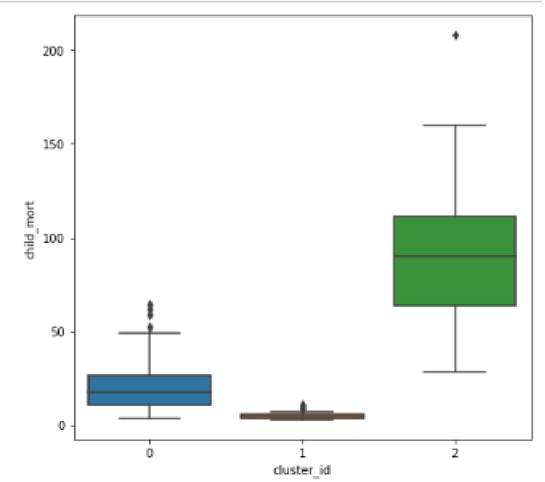


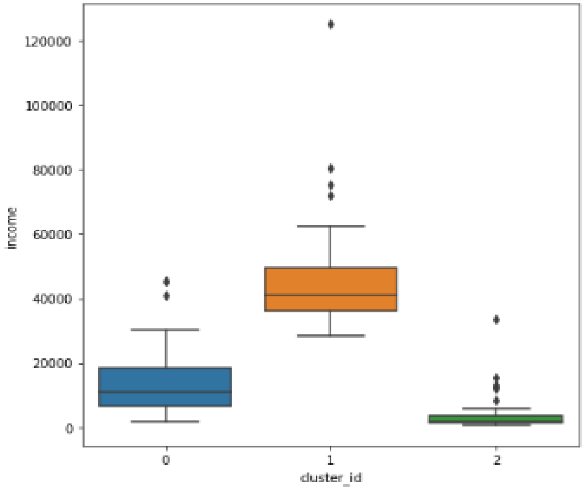


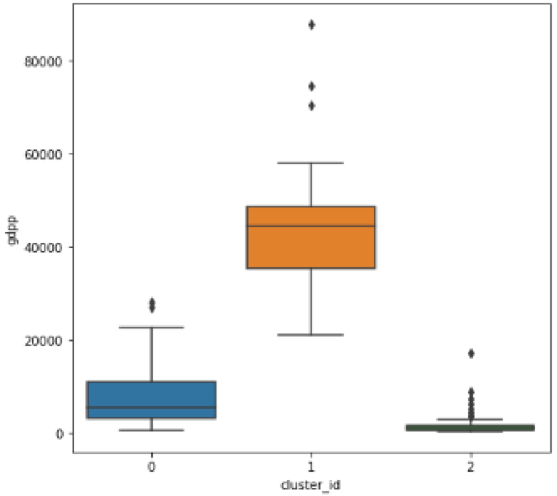


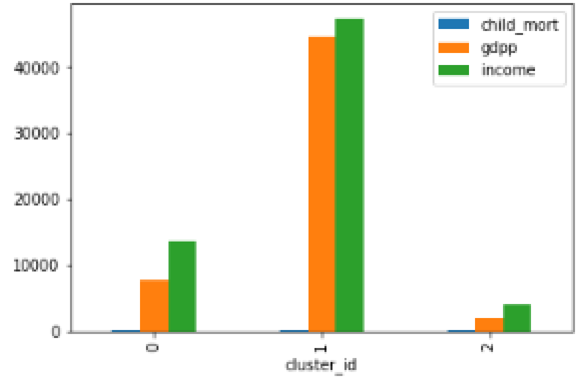
Cluster Profiling

- For the cluster Profiling – 3 variables – Gdpp, Child Mortality, Income are analysed to determine how much they vary for each cluster of countries to recognize and differentiate the clusters of developed countries from the clusters of under-developed countries.
- Plotting boxplots to visualize the same.
- Cluster with low gdpp, high child mortality, low income will consist of the countries that are in dire need of financial help. It is found that cluster 2 is that cluster.









Conclusion

- Cluster with Id 2 is sorted based on low gdpp, high child mortality rates,low income and 5 deserving countries are chosen.
- As per my analysis, countries which are in dire need of finance are as follows:

1. **Burundi**
2. **Liberia**
3. **Congo**
4. **Niger**
5. **Sierra Leone**

Thank You,
Deepa Kamath