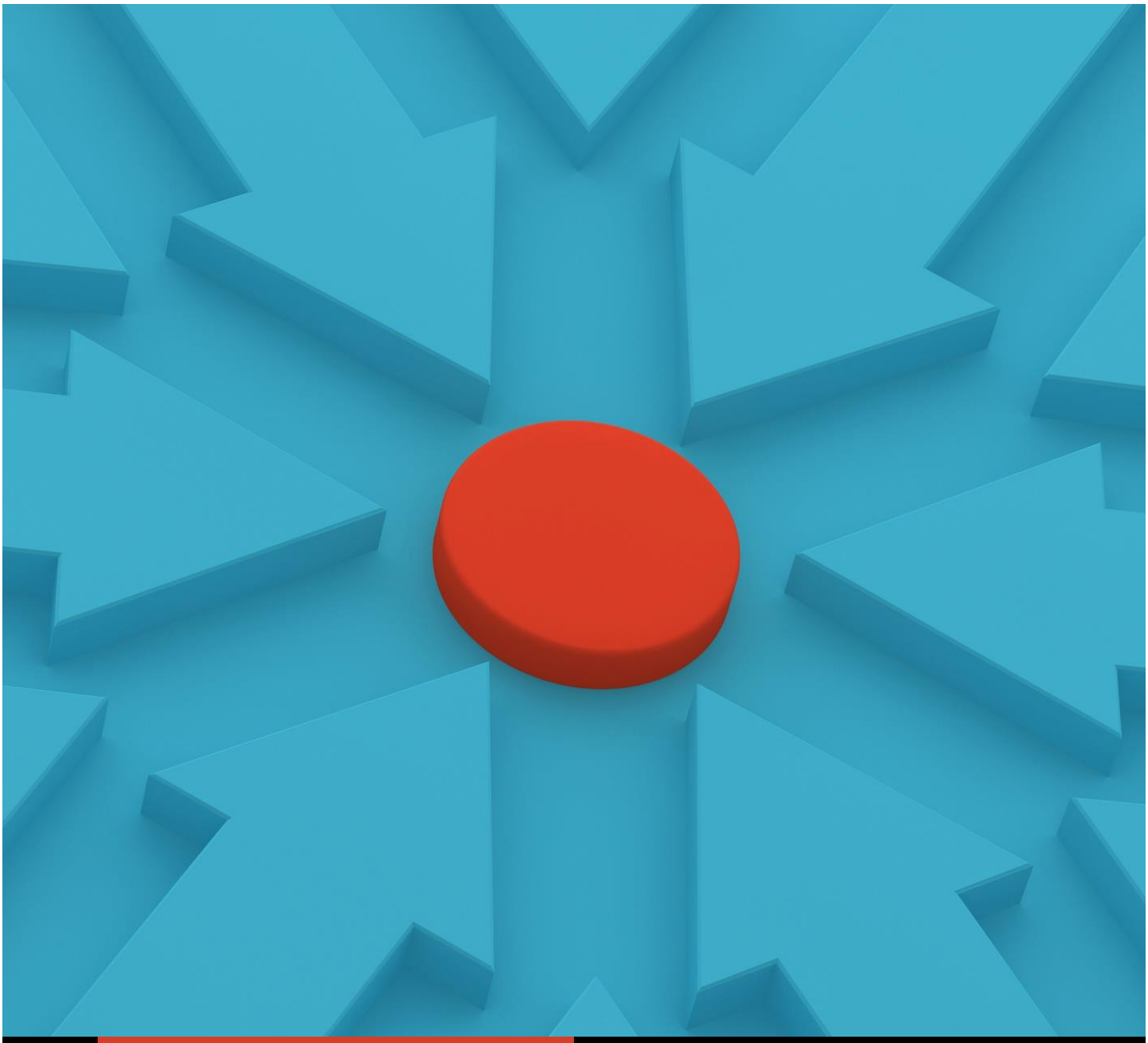
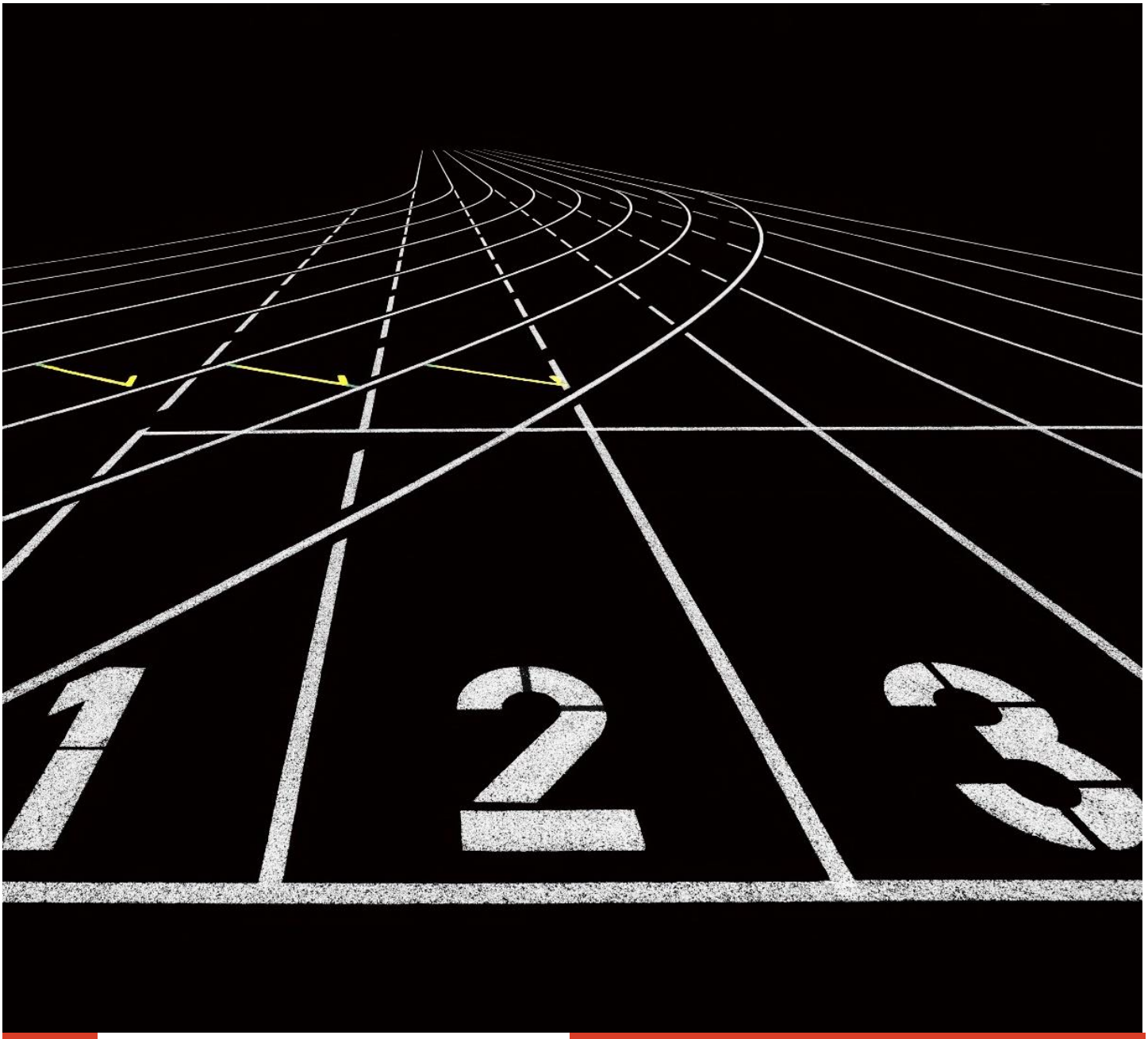


COUNTRIES

CLUSTERING – DIRE NEED OF FINANCIAL AID

DS C21

SRINIVASA VARADHAN V dmvaradhan4u@gmail.com



PROBLEM STATEMENT

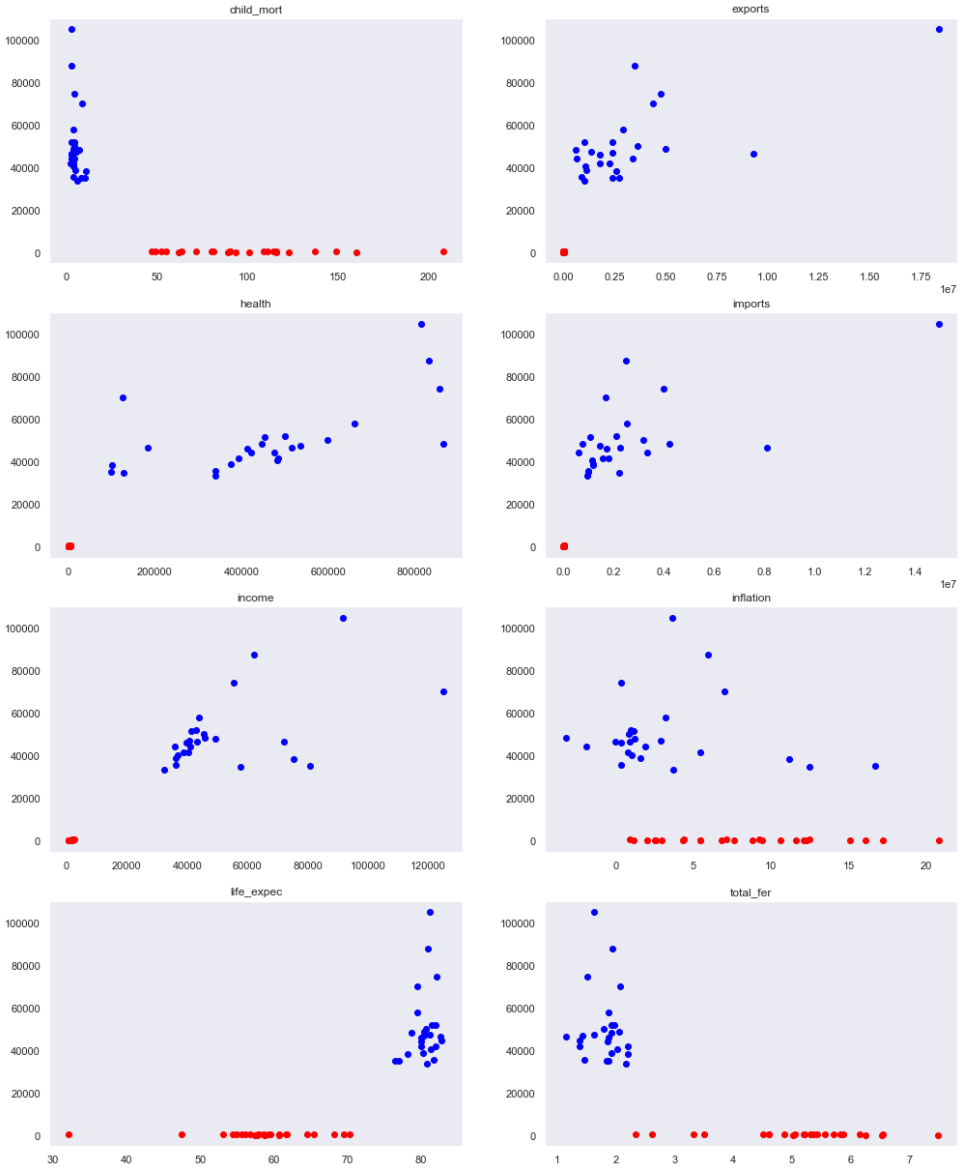
To perform clustering based on the given socio-economic data and arrive at top 5-10 countries which are in dire need of financial aid.

INSIGHTS FROM DATA

country Vs gdpp

The countries having a very low gdpp had

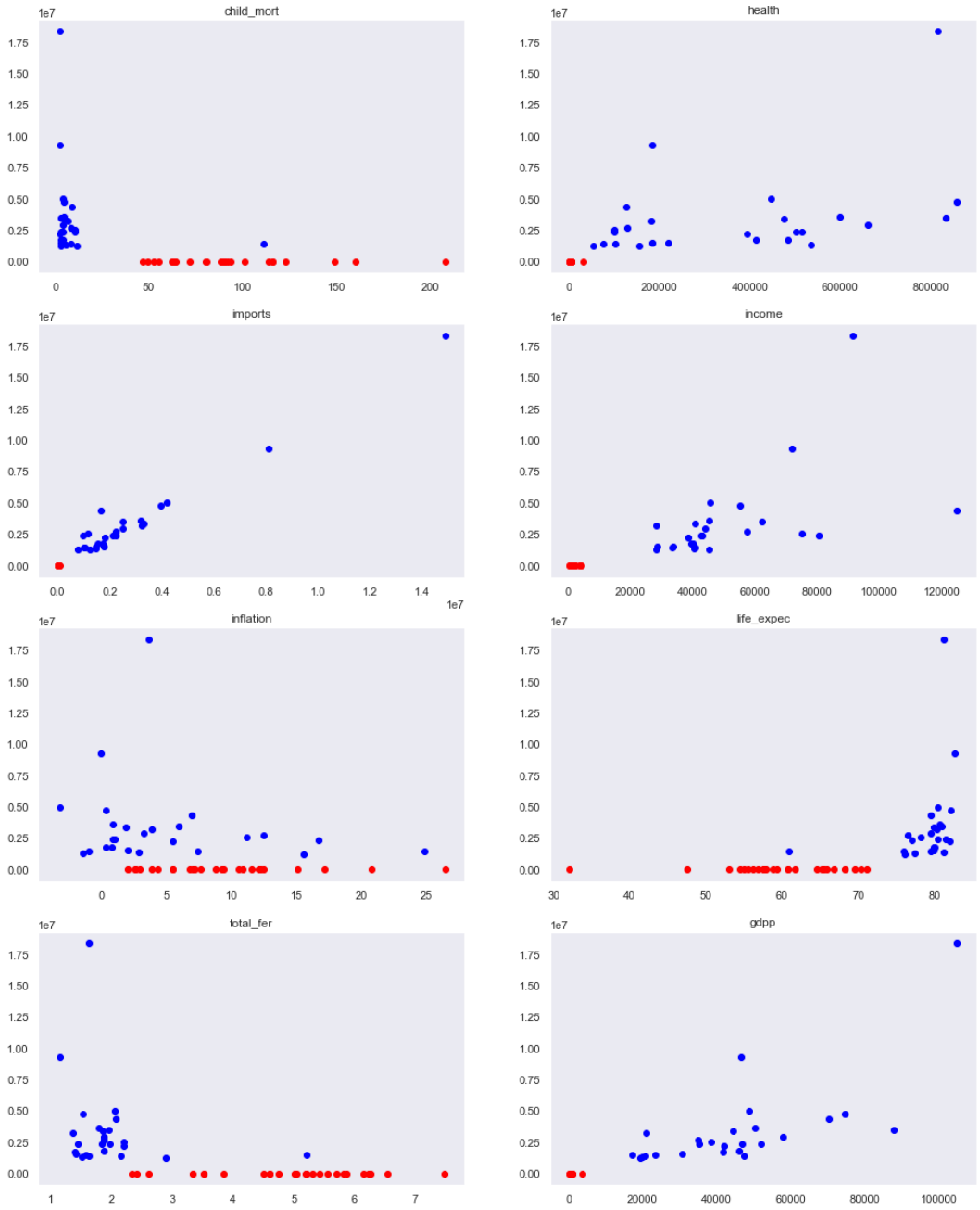
- Very High child_mort
- Very Low Exports
- Very Low expenditure on health
- Very Low imports
- Very Low income
- High inflation
- Life expec not greater than 70
- Very High fert



country Vs exports

The countries having a very low export had

- Very High child_mort
- Very Low expenditure on health
- Very Low imports
- Very Low income
- High inflation
- Life expec not greater than 70
- Very High fert
- Very low gdpp



Results

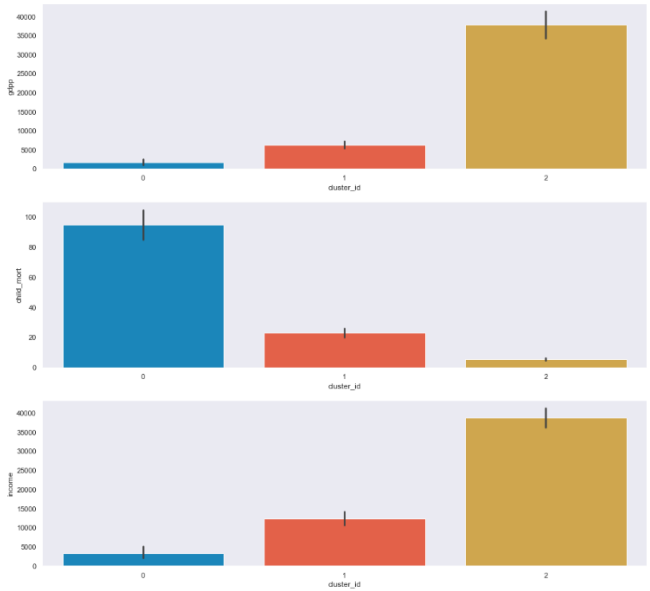


Figure 1 - Result of k-means

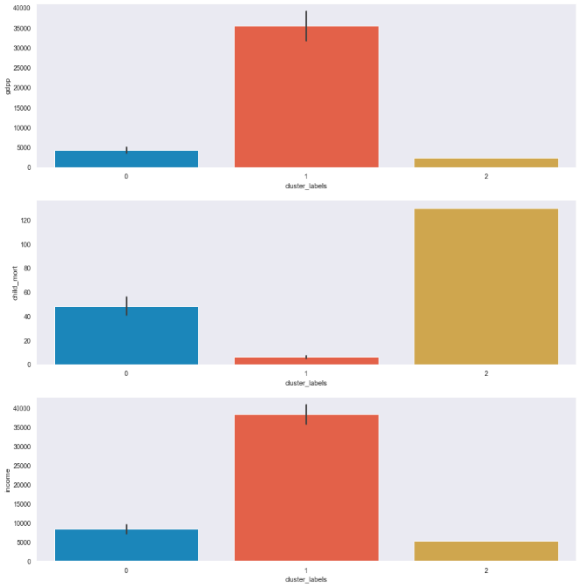


Figure 2 - Result of Hierarchical

Both the k-means and hierarchical clustering provided the same cluster results.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
26	Burundi	93.599998	2060.520018	2679.600088	9055.200176	764.0	12.300000	57.700001	6.26	231	0	0
88	Liberia	89.300003	6245.700125	3858.600062	30280.199501	700.0	5.470000	60.799999	5.02	327	0	0
37	Congo, Dem. Rep.	116.000000	13727.399490	2641.939949	16566.399490	609.0	20.799999	57.500000	6.54	334	0	0
112	Niger	123.000000	7725.600266	1795.679947	17086.799469	814.0	2.550000	58.799999	7.49	348	0	0
132	Sierra Leone	160.000000	6703.199696	5226.900152	13765.500000	1220.0	17.200001	55.000000	5.20	399	0	0
93	Madagascar	62.200001	10325.000000	1557.009992	17759.000000	1390.0	8.790000	60.799999	4.60	413	0	0
106	Mozambique	101.000000	13198.500000	2182.990016	19357.800320	918.0	7.640000	54.500000	5.56	419	0	0
31	Central African Republic	149.000000	5262.800085	1775.080009	11819.000000	888.0	2.010000	47.500000	5.21	446	0	0
94	Malawi	90.500000	10465.199650	3024.810070	16019.100700	1030.0	12.100000	53.099998	5.31	459	0	0
50	Eritrea	55.200001	2308.779982	1282.120041	11230.599632	1420.0	11.600000	61.700001	4.61	482	0	0

Figure 3 - Cluster Results

The above said countries are in dire need of aid in sequential order of priority. The fund shall accommodate the needs of the country in the order of priority.

APPROACH

First data was imported using an optimization technique to save memory. The method employed was to import sample count of data and down casting of datatype and finding the suitable datatype of each column and importing.

Then data was validated for missing values. There were no missing values. While performing EDA, the gdpp was set as the reference and the other socio-economic status against gdpp was visualized. The results were similar while performing with exports vs gdpp.

Outliers were treated using capping. The columns where selected on which upper capping where essential and cutoff was set at 95th percentile based on data. Next the data was checked for the Hopkins score and on multiple iteration received a value always greater than 87. Concluded good for clustering. Next with silhouette score and elbow curve, number of clusters were decided as three.

Performed k-means clustering and mapped those labels with the countries. Requirement is find countries with low gdpp, low income, high child_mort.This was achieved by sorting. In hierarchical effective clustering did not happen with Euclidean metrics, hence employed cityblock. Final results 45 countries in cluster of need for fund.

At top level, following the best case benchmark and worst performing countries for a specific parameter.

	Parameter	Avg Value	Max Country	Max Value	Min Country	Min Value
0	child_mort	38.27	Haiti	208.00	Iceland	2.60
1	exports	742061.89	Luxembourg	18375000.00	Myanmar	107.69
2	health	105673.32	United States	866359.98	Eritrea	1282.12
3	imports	658835.21	Luxembourg	14910000.00	Myanmar	65.11
4	income	17144.69	Qatar	125000.00	Congo, Dem. Rep.	609.00
5	inflation	7.78	Nigeria	104.00	Seychelles	-4.21
6	life_expec	70.56	Japan	82.80	Haiti	32.10
7	total_fer	2.95	Niger	7.49	Singapore	1.15
8	gdpp	12964.16	Luxembourg	105000.00	Burundi	231.00

Compare and contrast K-means Clustering and Hierarchical Clustering.

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e., O(n) while that of 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.
- K Means clustering requires prior knowledge of K i.e., no. of clusters you want to divide your data into. But you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

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

- Deciding the number of clusters(k) using Silhouette score and Elbow curve
- Setting random k cluster centroids
- Finding the distance between the cluster centroid and all points
- The point with low distance is added into the cluster
- Cluster centroid shifts to a new point based on the newly added cluster points
- Step continue to repeat until the cluster observes no further changes
- In case of a K-means ++, all centroids are not randomly initiated once. Only one centroid is initiated.

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

- Statistically, the k value is chosen using two metrics. The Silhouette score & Elbow curve
- In Silhouette score, the highest value k is chosen. In general k=2 is never considered in business, hence any value greater than 2 which has higher value is chosen.
- In Elbow curve, the point is chosen where there is a steep curve point similar to an elbow
- In business aspect, the k value is chosen based on two things – the number of actions that take or act on, the number of KPI they can track continuously on long term.

Explain the necessity for scaling/standardization before performing Clustering.

- Each feature might be in different units or measure system and have completely different scales of value
- When working with different scale, the features are not comparable
- In Clustering analysis groups are classified based on the distance in mathematical space
- Standardization/Scaling helps in converting all into a comparable measure/unit and perform clustering groups.

Explain the different linkages used in Hierarchical Clustering.

- Single Linkage: it returns the minimum inter-cluster distance two cluster points
- Complete Linkage: it returns the maximum inter-cluster distance between two cluster points
- Average Linkage: it computes all the inter-cluster distance between all points against each cluster and then arithmetic mean of the distance is returned.