CLUSTERING & PCA ASSIGNMENT

SUBMITTED BY:

DARSHNA

PROBLEM STATEMENT

- Categorise the countries using some socio-economic and health factors that determine the overall development of the country.

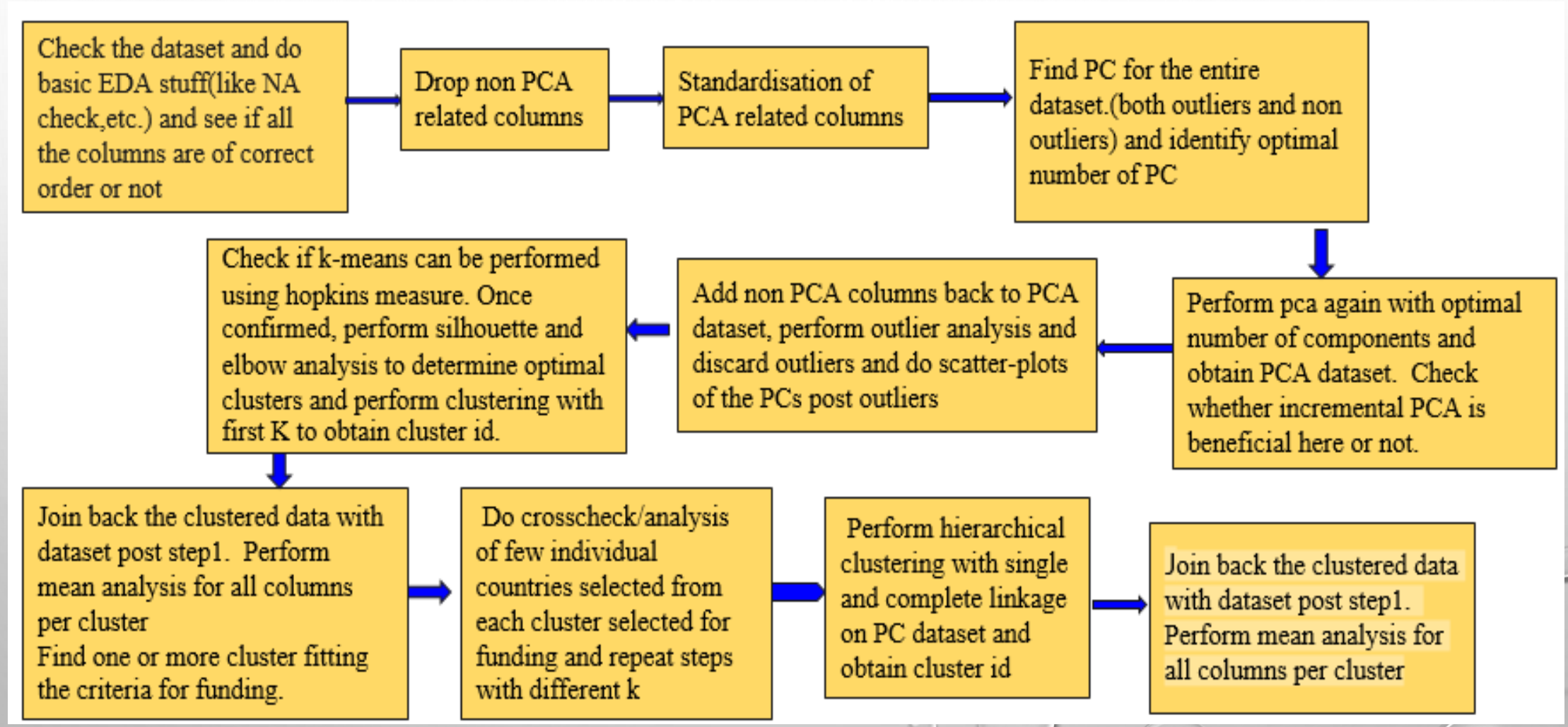
OBJECTIVE

- Aim of the analysis is to cluster the countries by the factors mentioned and then present solution.

DATA PROVIDED

- Country:- name of the country
- Child_mort:- death of children under 5 years of age per 1000 live births
- Exports:- exports of goods and services. Given as %age of the total GDP
- Health:- total health spending as%age of total GDP
- Imports:- imports of goods and services. Given as %age of the total GDP
- Income:- net income per person
- Inflation:- the measurement of the annual growth rate of the total GDP
- Life_expec:-the average number of years a newborn child would live if the current mortality patterns are to remain the same
- Total_fer:- the number of children that would be born to each women if the current age fertility rate remain the same
- Gdpp:- the GDP per capita. Calculated as the total GDP divided by the total population

METHODOLOGY

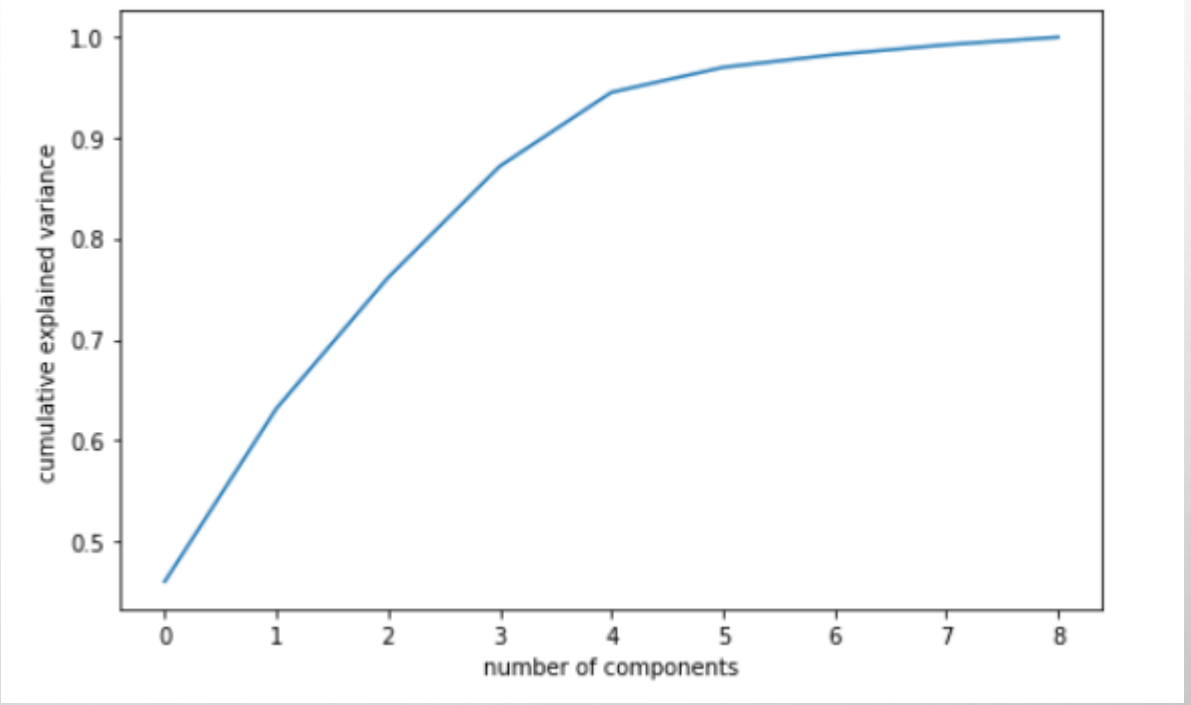


PRINCIPAL COMPONENT ANALYSIS

- Number of component selected

Above the 80% cumulative explained variance.

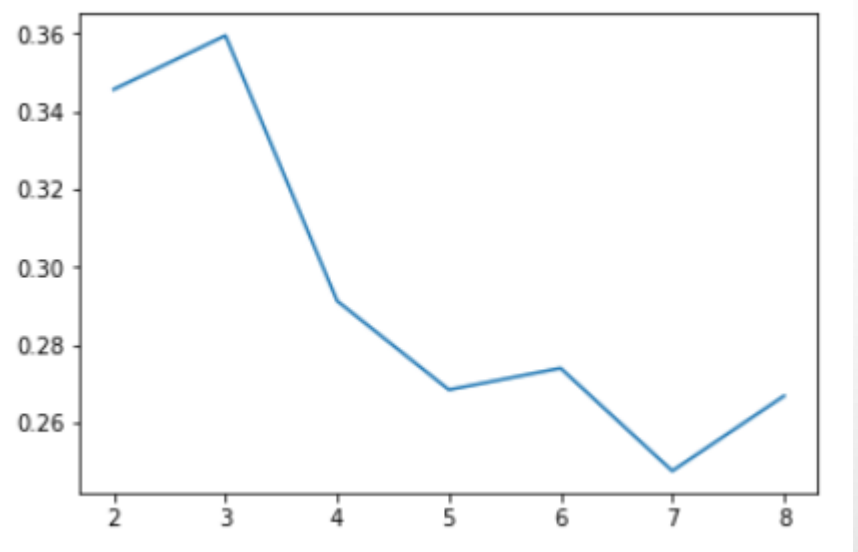
- Here I select 5 component which gives approx. 95% cumulative explained Variance.
- Pc=5



K-MEAN CLUSTERING

Silhouette analysis

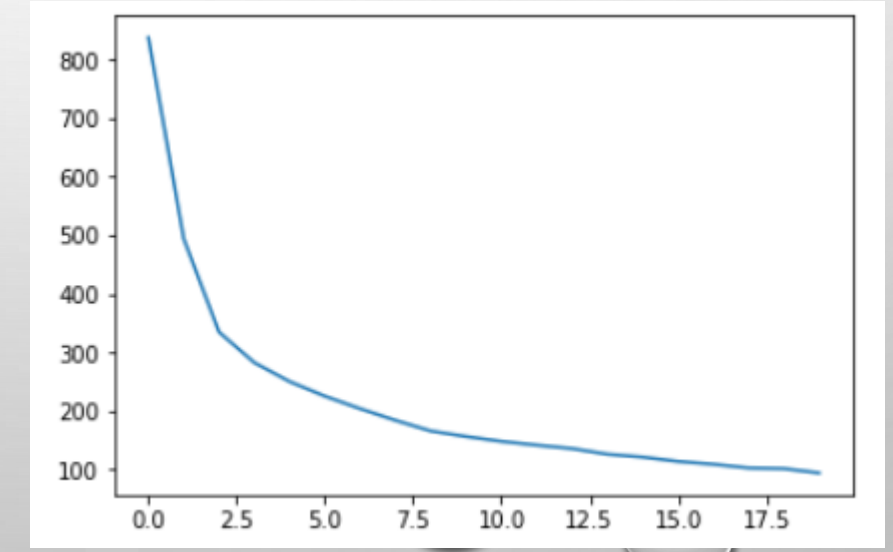
According to this graph we can select value of k between 3 and 4.



ii. Sum of squared distances

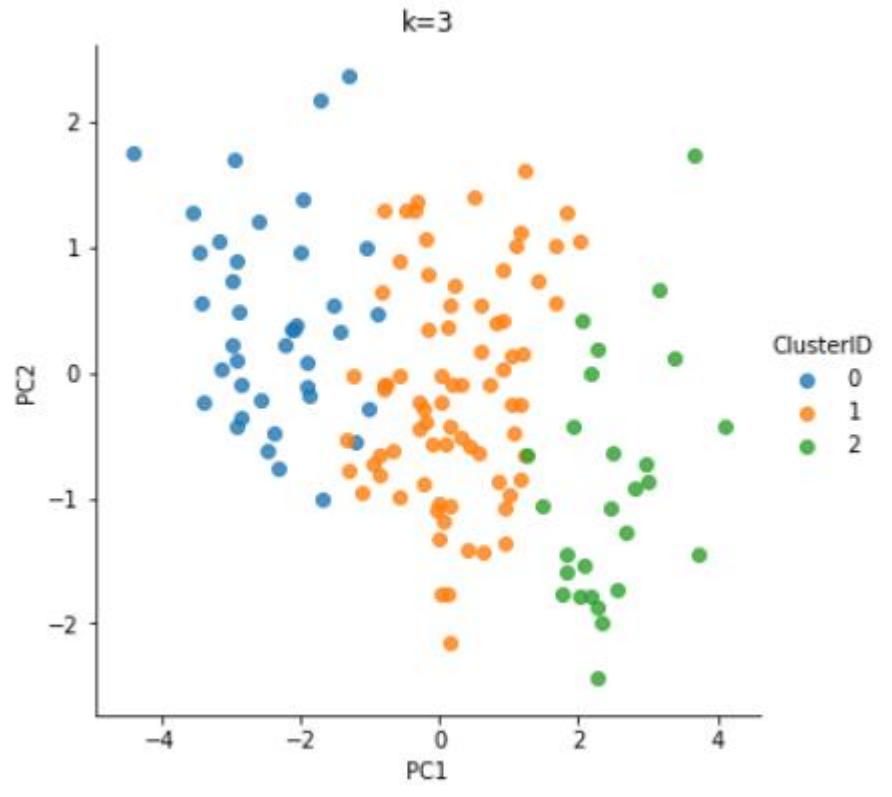
Here elbow is bent between 3 to 6.

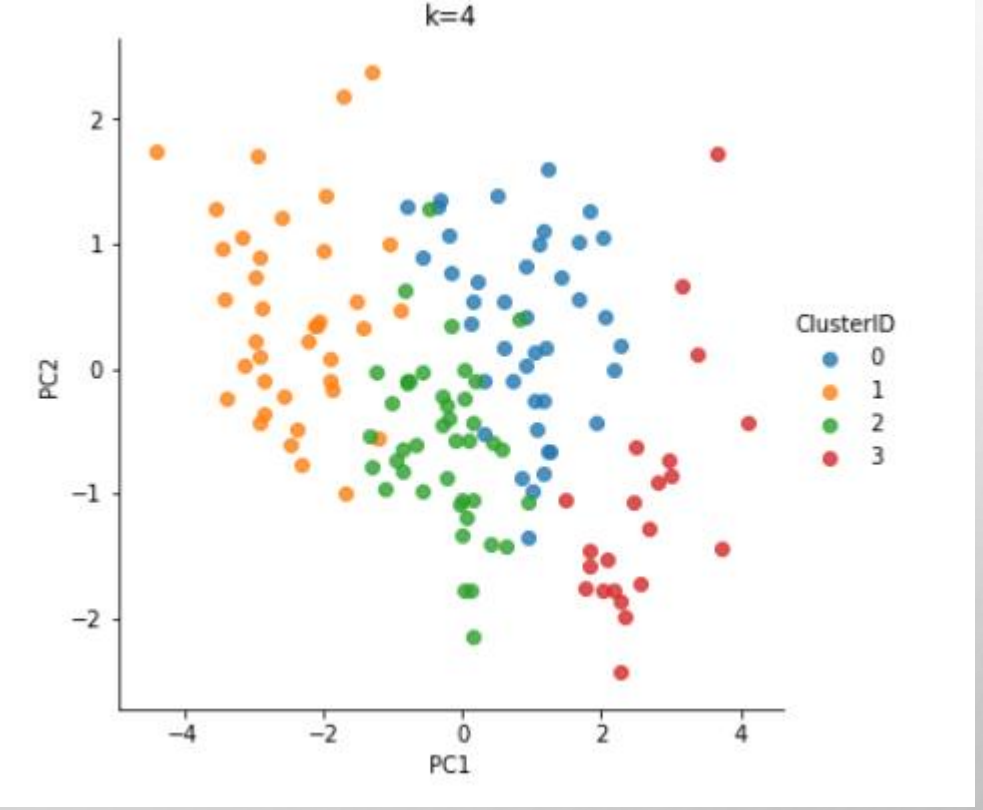
So, according to both the graph I select value of K for clustering is 3 and 4.



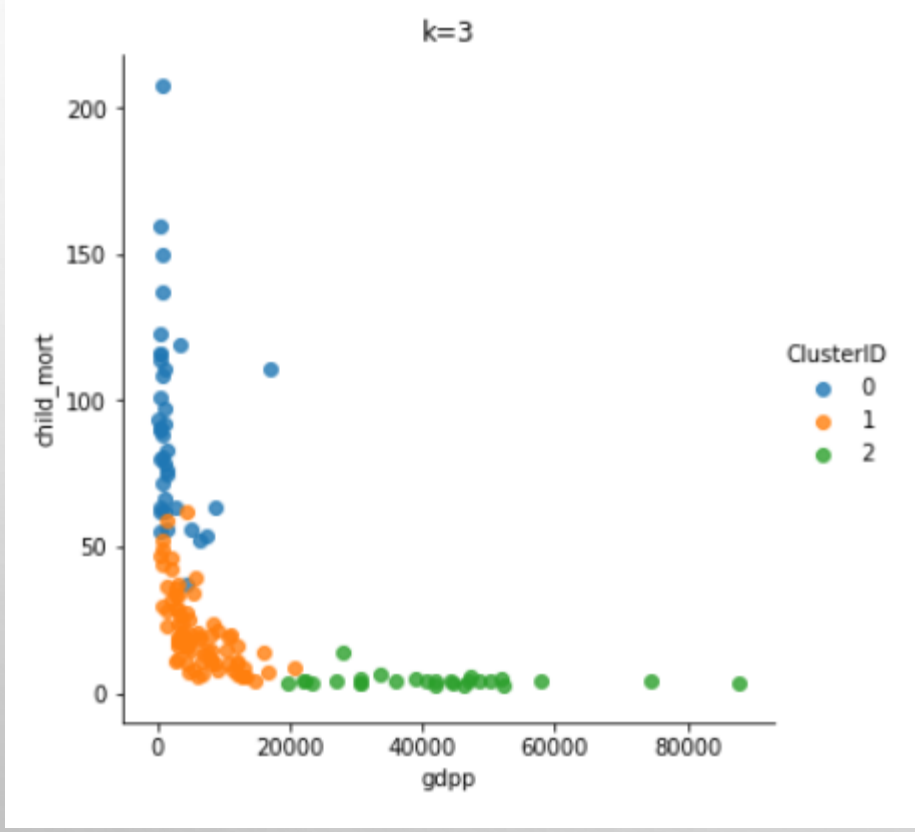
VISUALISATION

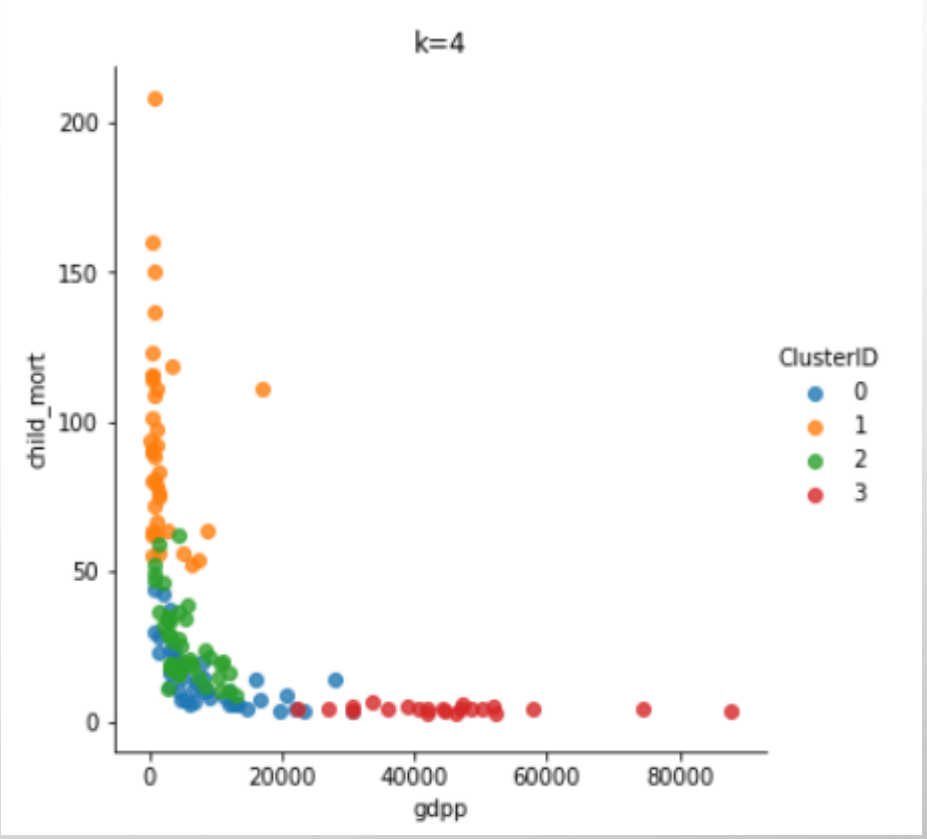
- 2-D plot between PC1 and PC2 using different cluster (k=3 and 4) without outlier



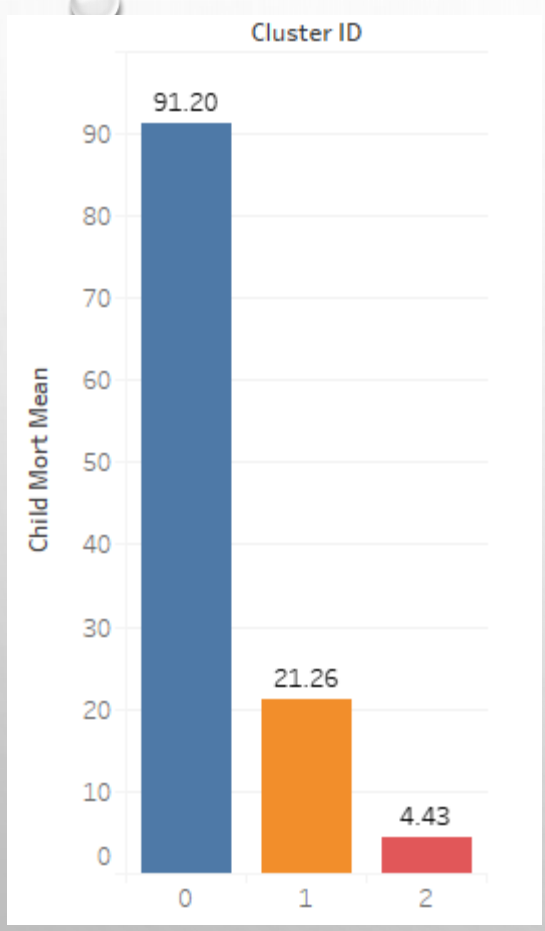


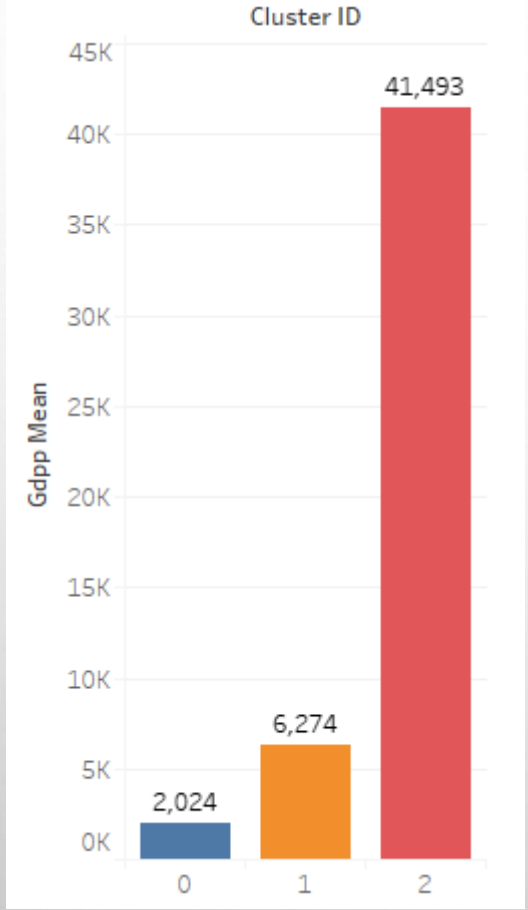
- 2-D plot between child_mort and gdpp using different cluster (k=3 and 4) without outlier

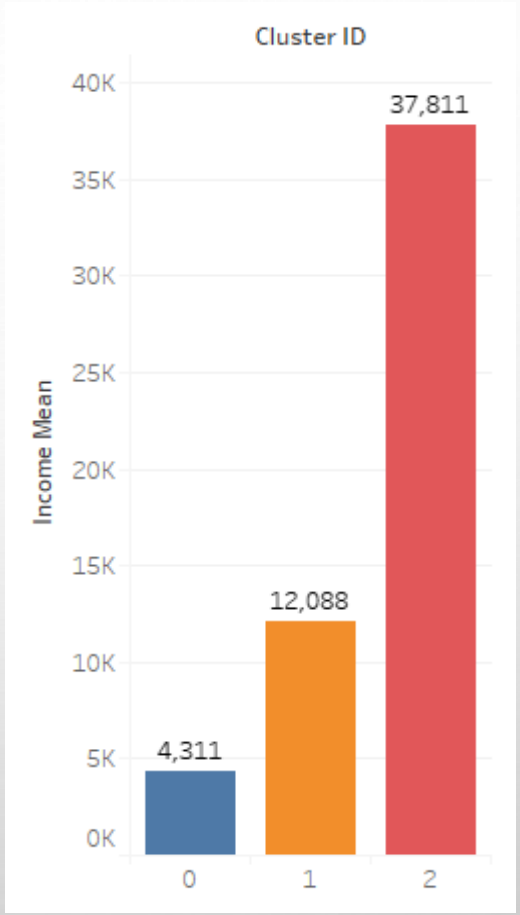


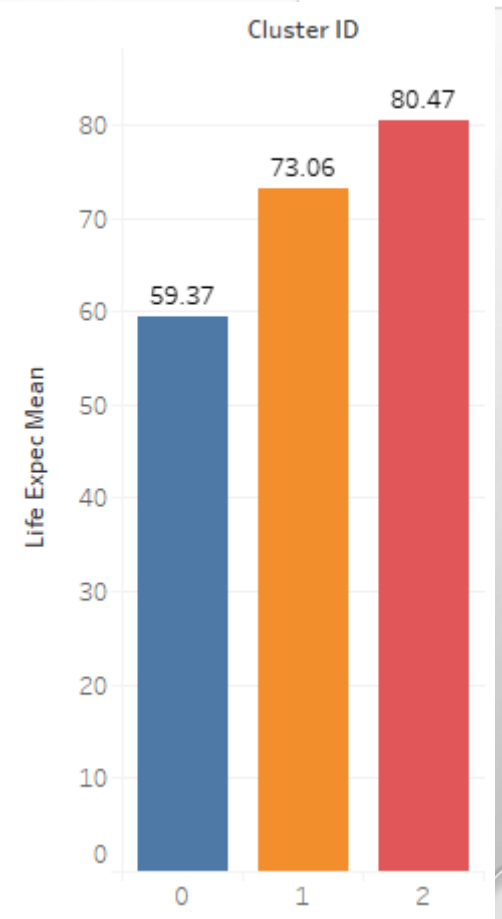


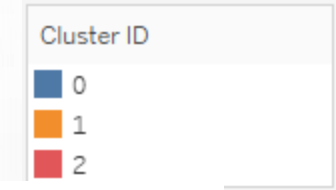
K-MEAN CLUSTERING WITHOUT OUTLIER











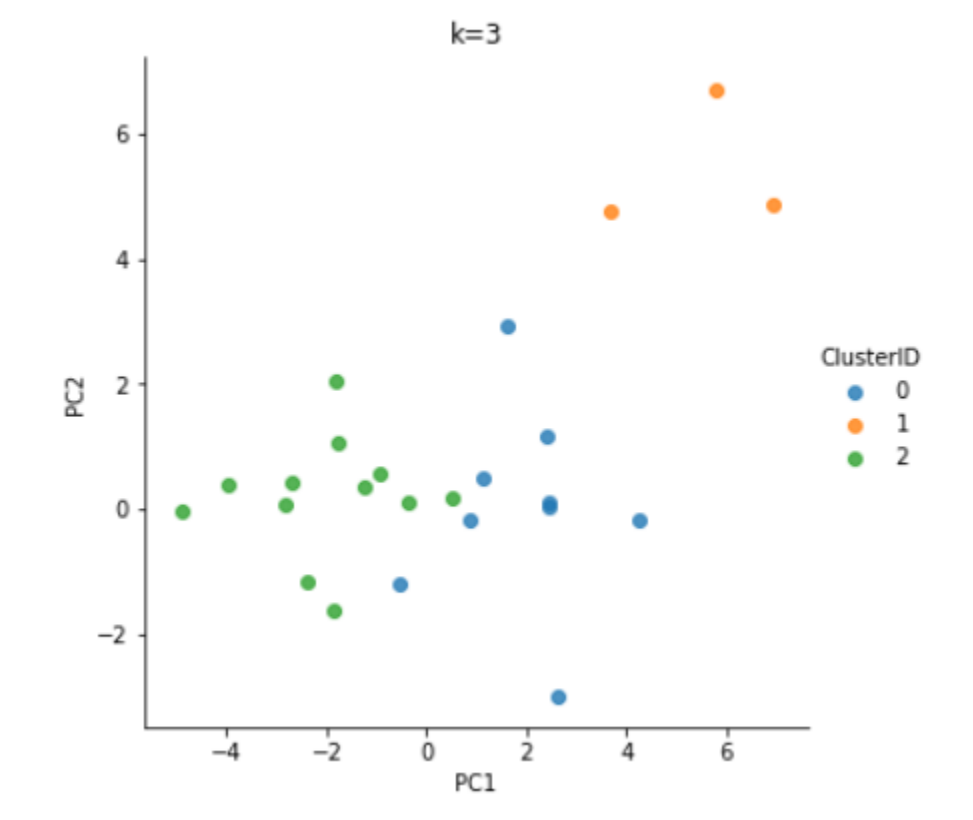
ANALYSING COUNTRIES WHICH NEED AID USING THE ABOVE GRAPH

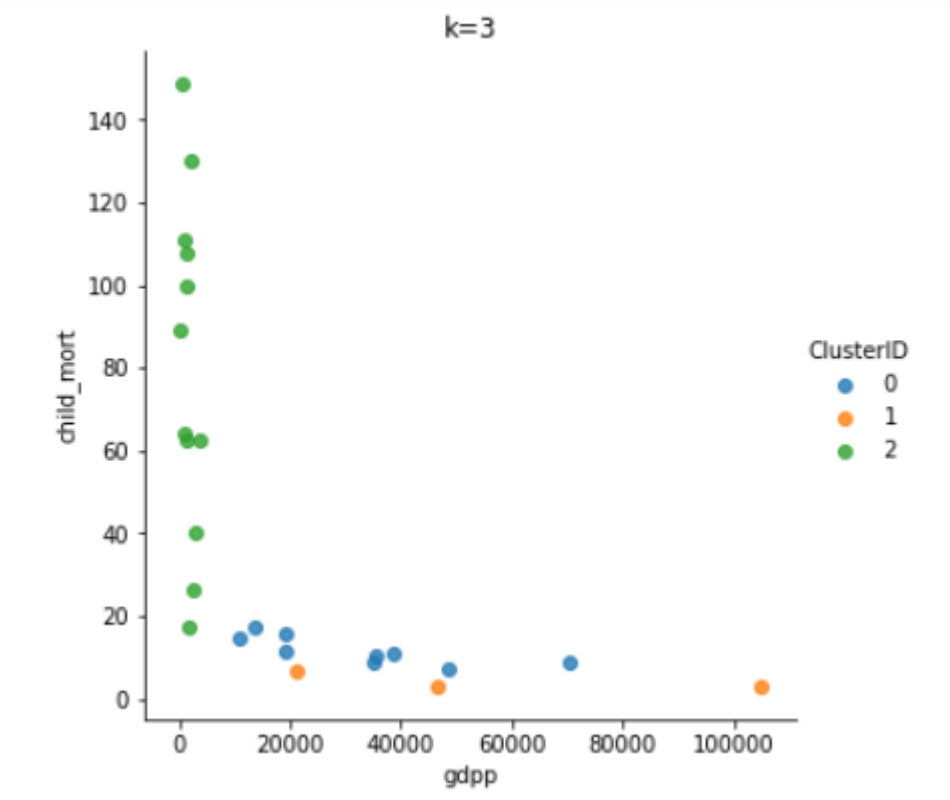
- **Some socio-economic and health factors that determine the overall development of the country of the without outlier data.**

Major factors are child mortality, GDPP, income , health and life expec.

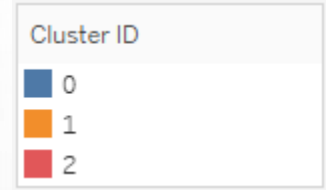
1. Child mortality mean of cluster 0 is high which means death of children under 5 years of age per 1000 live births is high in cluster 0 countries. So, country under cluster 0 needs aid.
2. GDPP mean of cluster 0 is low which means GDP per population of cluster 0 is less. Similarly cluster 0 needs aid according to the GDPP also.
3. Income mean is also low for cluster 0 country so again taking income factor cluster 0 needs aid.
4. Life expec tells about the average number of years a new born child would live if the current mortality patterns are to remain the same. So according to this factor also country cluster 0 needs aid with respect to other country cluster.

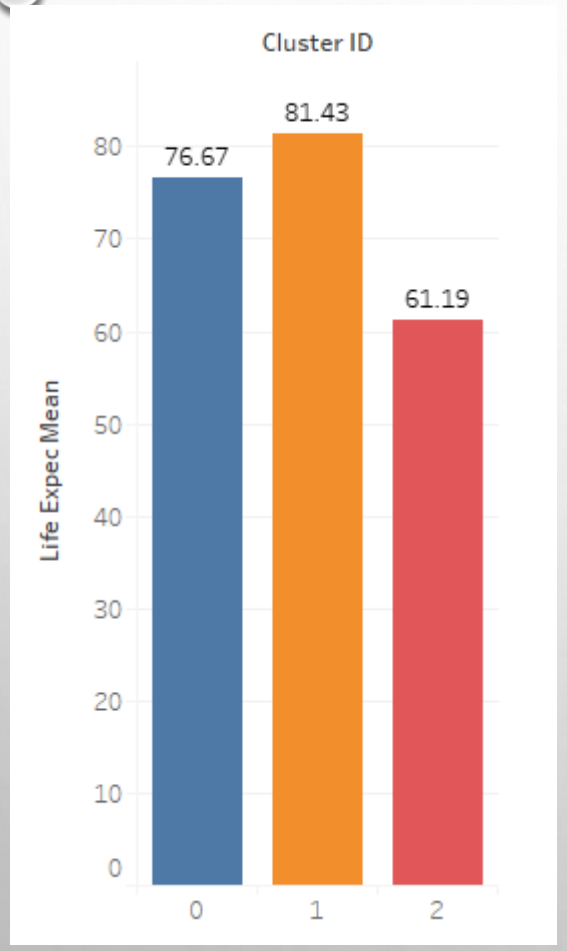
- 2-D plot between PC1 and PC2 using different cluster (k=3) with outlier
- 2-D plot between child_mort and gdpp using different cluster (k=3) with outlier

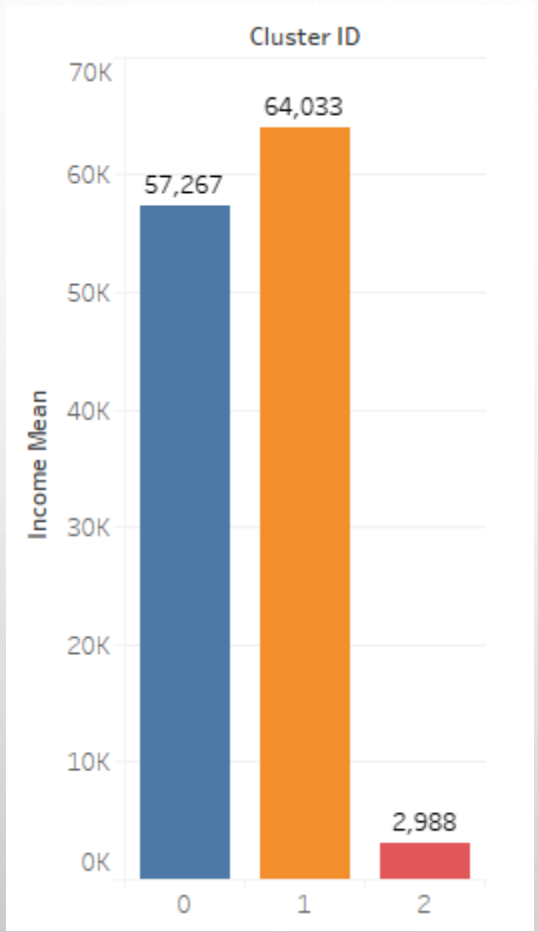


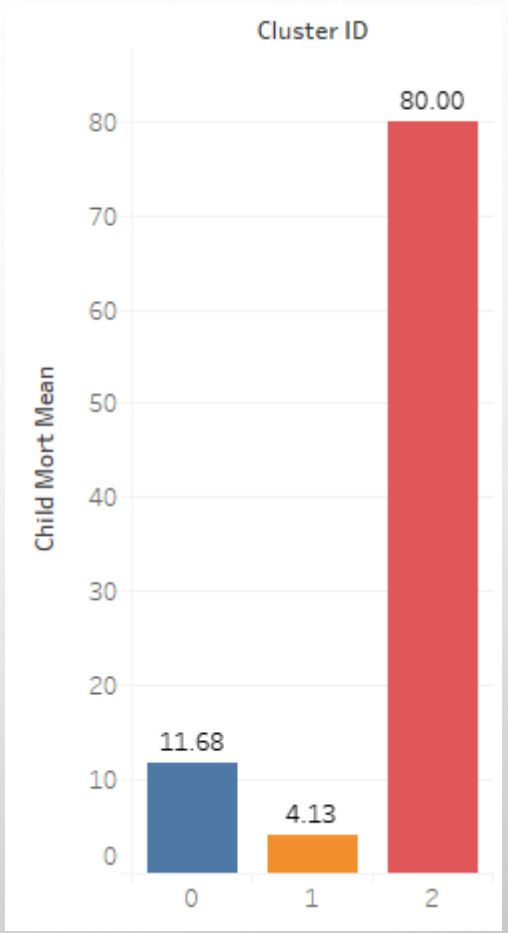


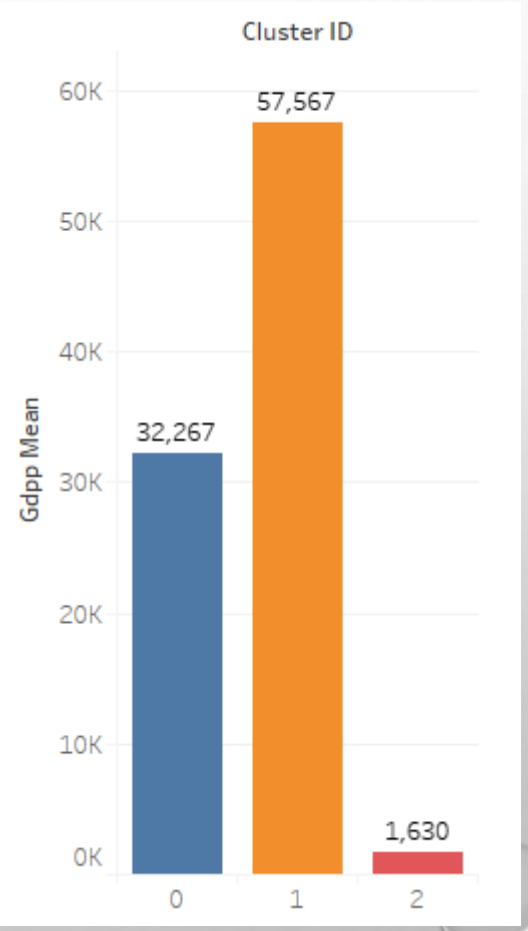
K-MEAN CLUSTERING WITH OUTLIER FOR SOME IMPORTANT FACTORS











ANALYSING COUNTRIES WHICH NEED AID USING THE ABOVE GRAPH

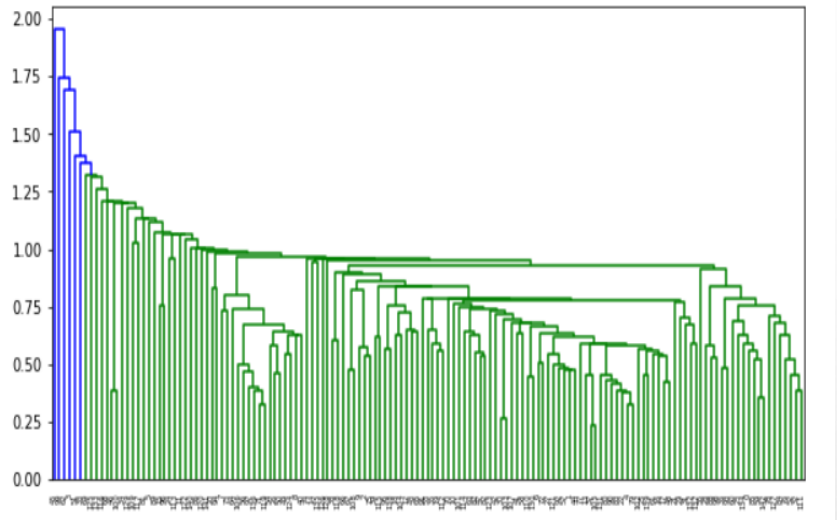
- **Some socio-economic and health factors that determine the overall development of the country**

Major factors are child mortality, GDPP, income , health and life expec.

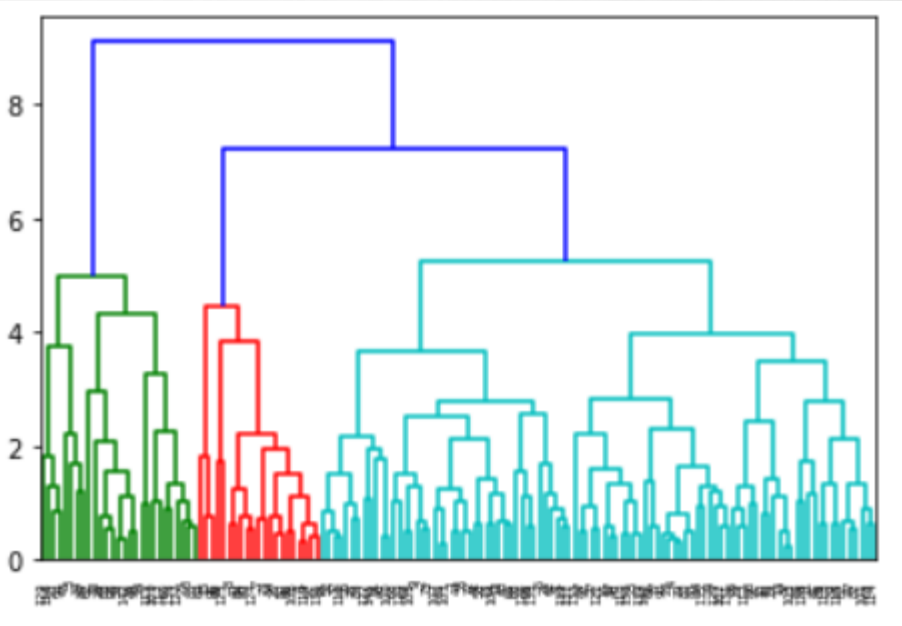
1. Child mortality mean of cluster 2 is high which means death of children under 5 years of age per 1000 live births is high in cluster 2 countries. So, country under cluster 2 needs aid.
2. GDPP mean of cluster 2 is low which means GDP per population of cluster 2 is less. Similarly cluster 2 needs aid according to the GDPP also.
3. Income mean is also low for cluster 2 country so again taking income factor cluster 2 needs aid.
4. Life expec tells about the average number of years a new born child would live if the current mortality patterns are to remain the same. So according to this factor also country cluster 2 needs aid with respect to other country cluster.

HIERARCHAL CLUSTERING

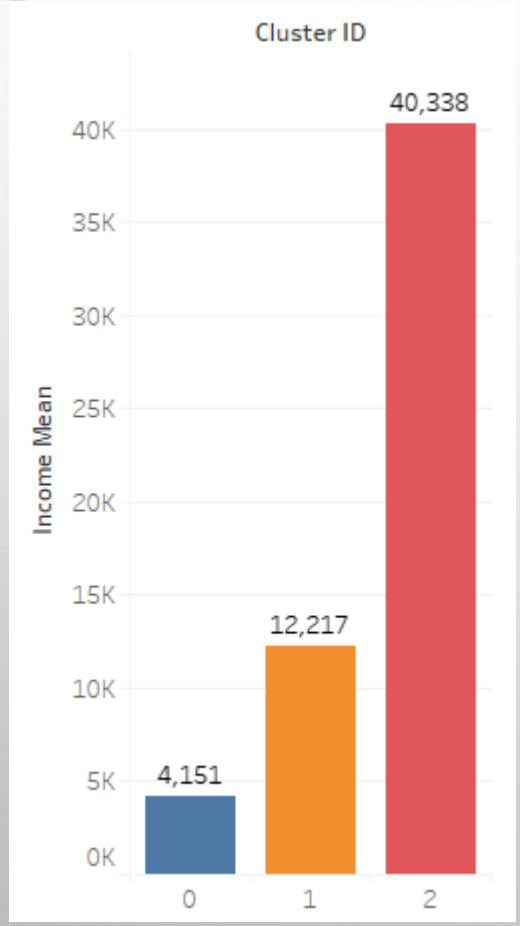
- **Single hierarchal clustering**

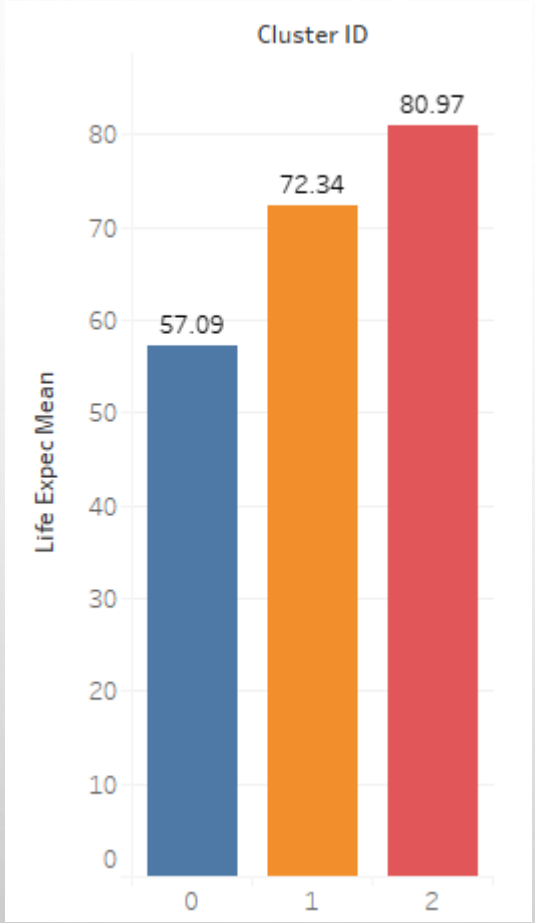


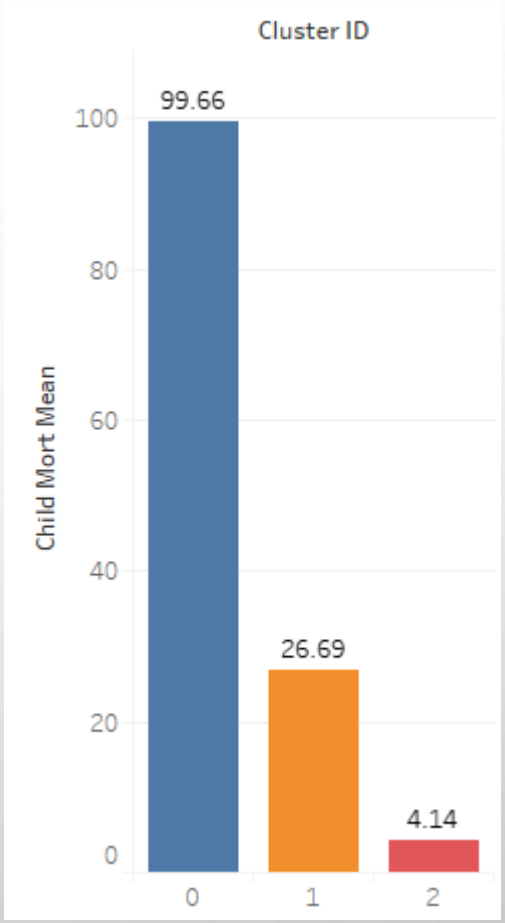
- **Complete hierarchal clustering**

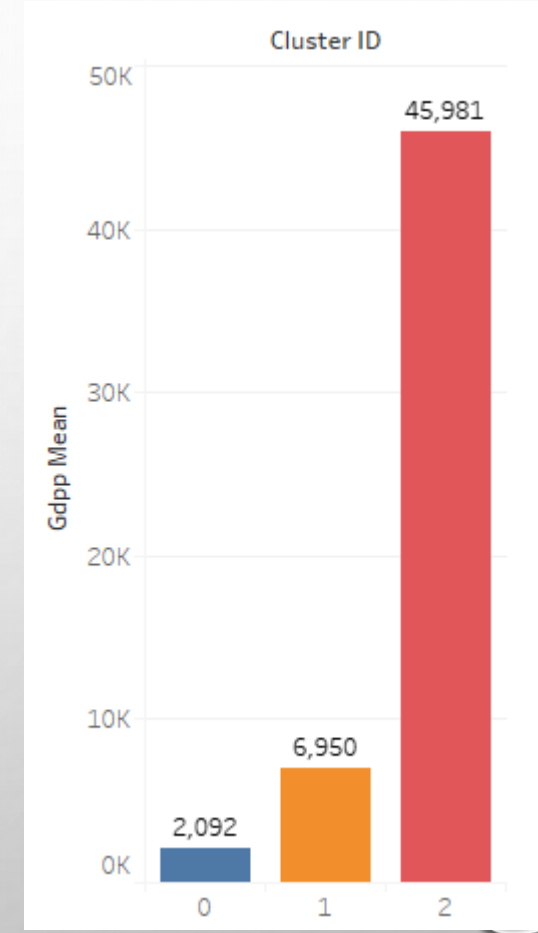


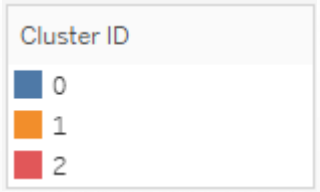
WITHOUT OUTLIER











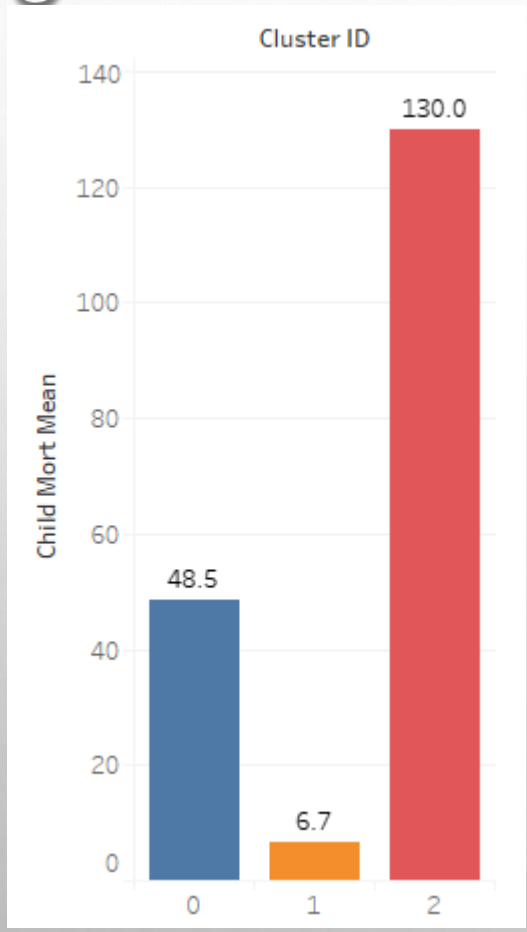
ANALYSING COUNTRIES WHICH NEED AID USING THE ABOVE GRAPH

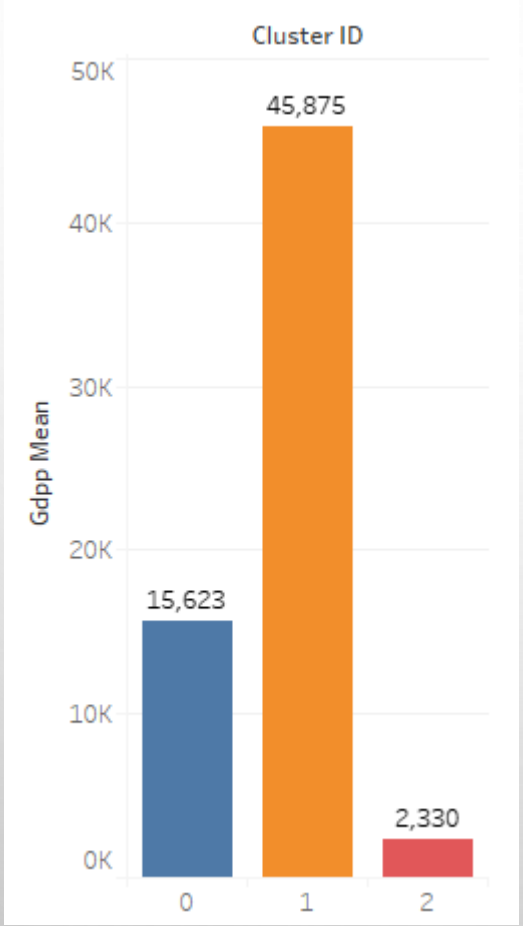
- **Some socio-economic and health factors that determine the overall development of the country of the without outlier data.**

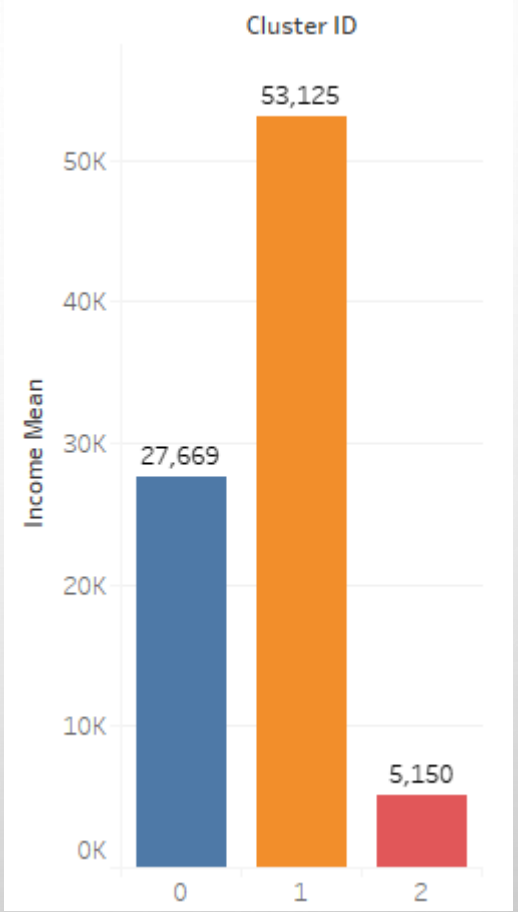
Major factors are child mortality, GDPP, income , health and life expec.

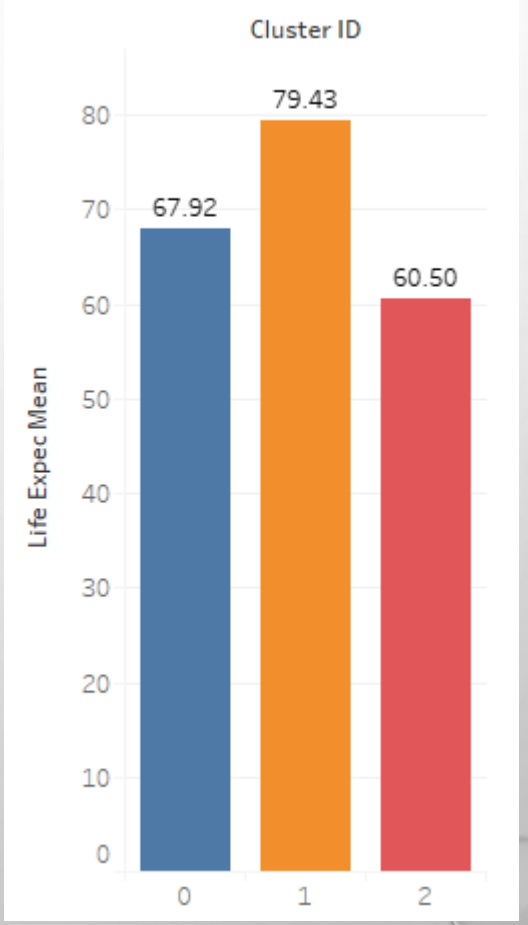
1. Child mortality mean of cluster 0 is high which means death of children under 5 years of age per 1000 live births is high in cluster 0 countries. So, country under cluster 0 needs aid.
2. GDPP mean of cluster 0 is low which means GDP per population of cluster0 is less. Similarly cluster 0 needs aid according to the GDPP also.
3. Income mean is also low for cluster 0 country so again taking income factor cluster 0 needs aid.
4. Life expec tells about the average number of years a new born child would live if the current mortality patterns are to remain the same. So according to this factor also country cluster 0 needs aid with respect to other country cluster.

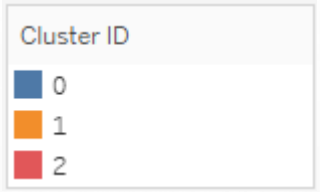
WITH OUTLIER











ANALYSING COUNTRIES WHICH NEED AID USING THE ABOVE GRAPH

- **Some socio-economic and health factors that determine the overall development of the country of the outlier data.**

Major factors are child mortality, GDPP, income , health and life expec.

1. Child mortality mean of cluster 2 is high which means death of children under 5 years of age per 1000 live births is high in cluster 2 countries. So, country under cluster 2 needs aid.
2. GDPP mean of cluster 2 is low which means GDP per population of cluster 2 is less. Similarly cluster 2 needs aid according to the GDPP also.
3. Income mean is also low for cluster 2 country so again taking income factor cluster 2 needs aid.
4. Life expec tells about the average number of years a new born child would live if the current mortality patterns are to remain the same. So according to this factor also country cluster 2 needs aid with respect to other country cluster.

THANK YOU