upGrad

PG Diploma in Data Science Sep 2020

Clustering Assignemnt

By

Srihari R

Problem Statement

Introduction

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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

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.

Dataset

The data given contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

•**The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,

•**All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1.**Approved:** The Company has approved loan Application

2.**Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

3.**Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).

4.**Unused offer:** Loan has been cancelled by the client but on different stages of the process

Data Cleaning

Application Data

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

Shape before cleaning

Rows: 167
Columns: 10

Shape after cleaning

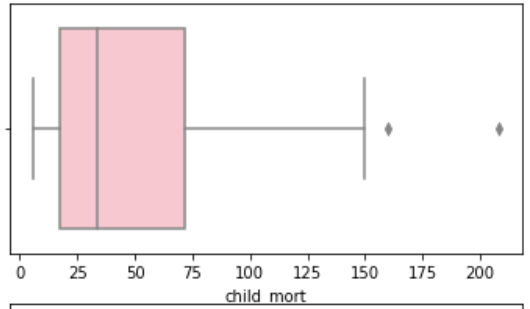
Rows: 125
Columns: 10

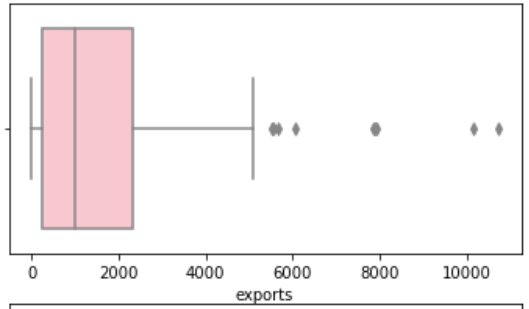
Some basic cleaning like removing the outliers are done the data and there are some outliers still which does not affect the data in any way

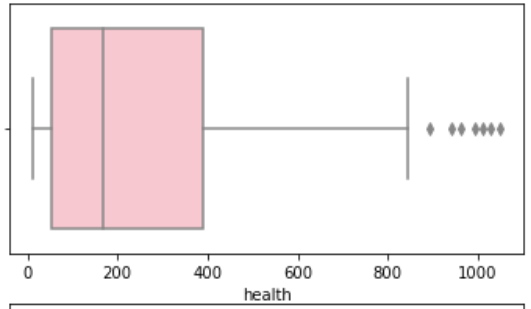
Univariate Analysis

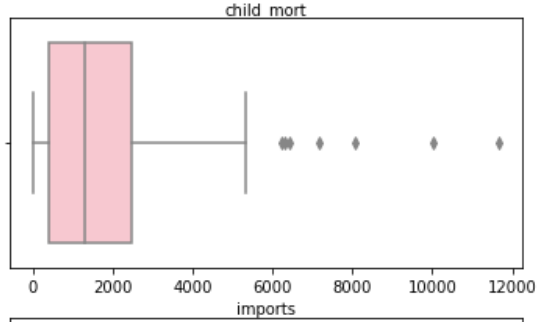
01 Boxplot for all numeric features

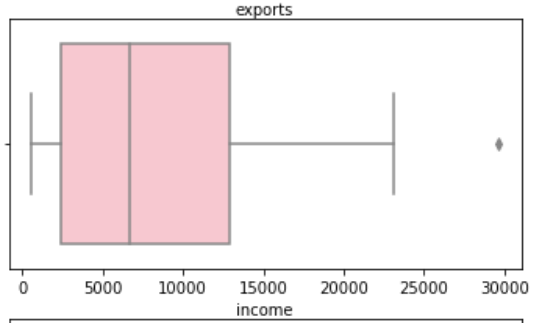
- There are outliers in every numerical column, but not every outlier can be removed which is domain specific outliers
- We can also see higher inflation values above 50, which is not good for a country, so it is advised to keep that data point
- The GDPP variable has too many outliers at the higher range, since our main moto is to find countries that need immediate help, higher GDPP countries obviously doesn't need any immediate help

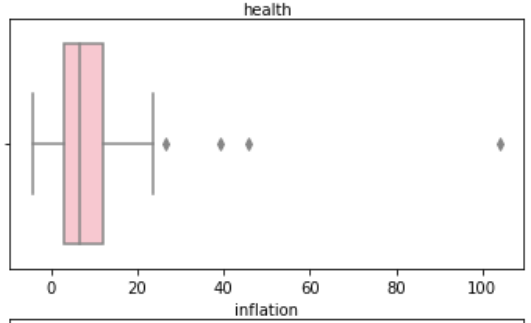


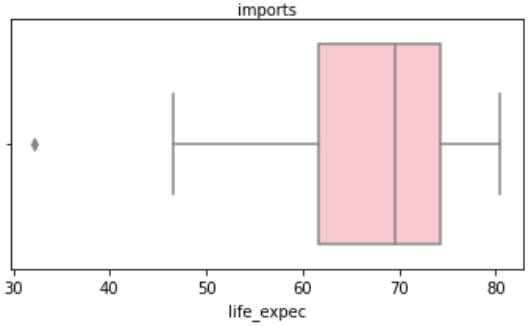


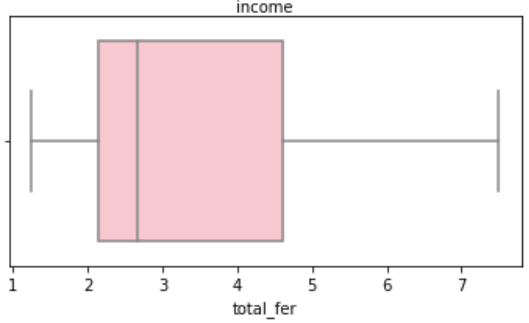


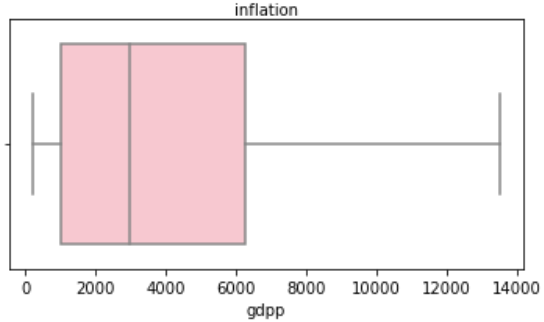






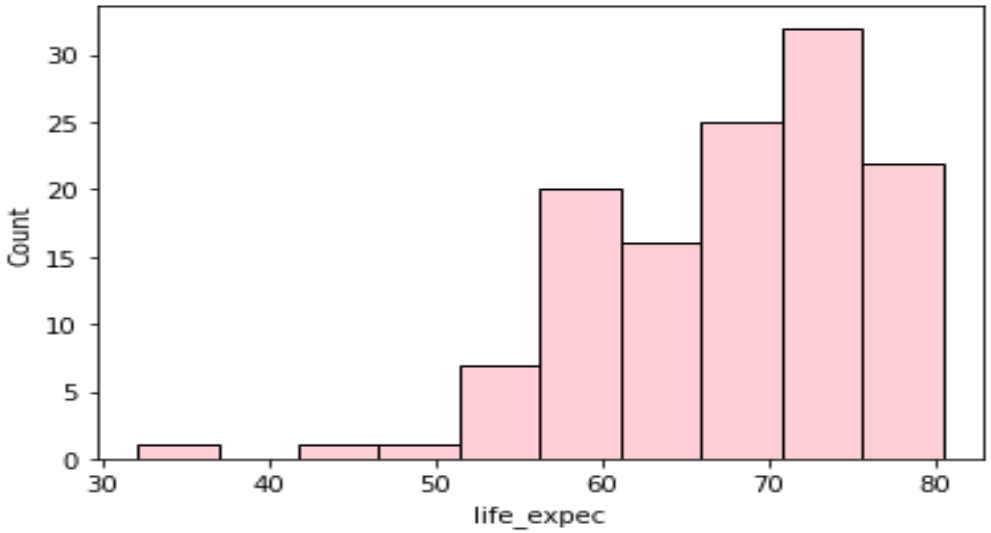






Univariate Analysis

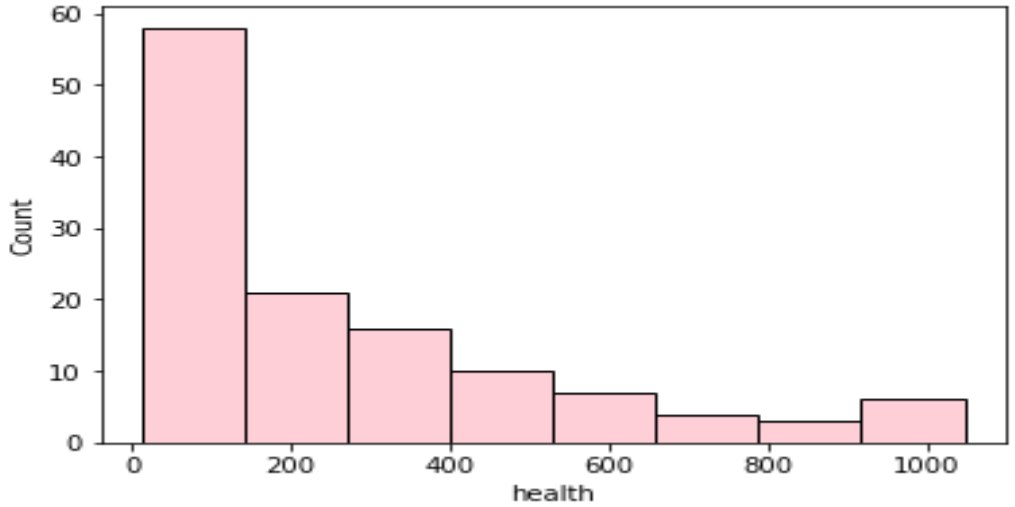
01 Hist plot for life expentency



Inferences

- There are more than 50 countries whose average life expentency is more than 60
- There are less than 5 countries with avergae life expentency of less then 50

02 Hist plot for health

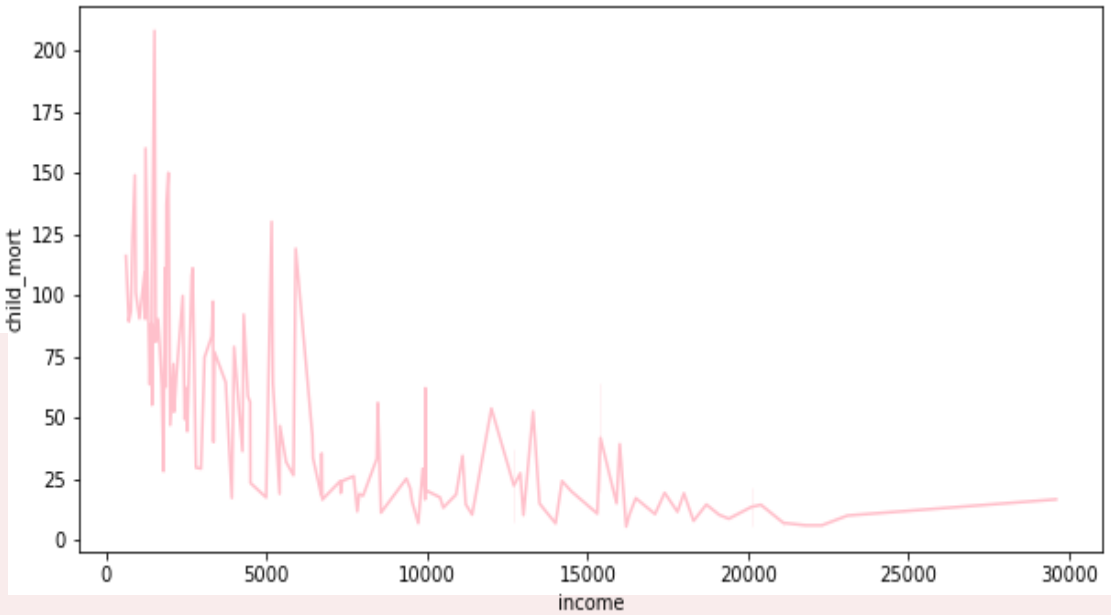


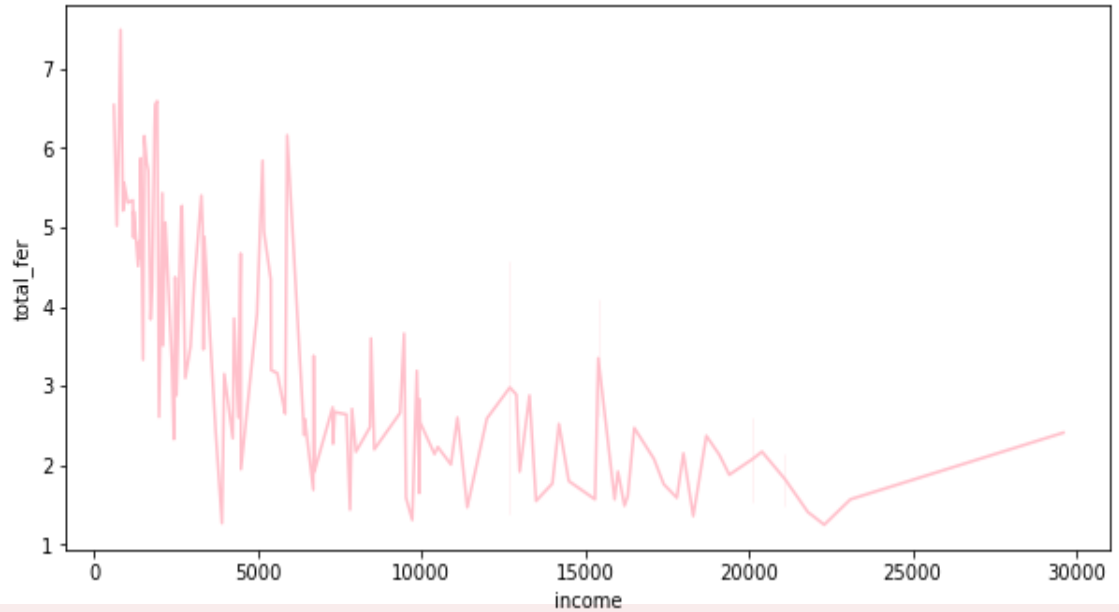
Inferences

- Most countries (nearly 60) spends on average 100 million from the gdpp
- Less than 10 countries spends about 1000million from the gdpp

Bivariate Analysis

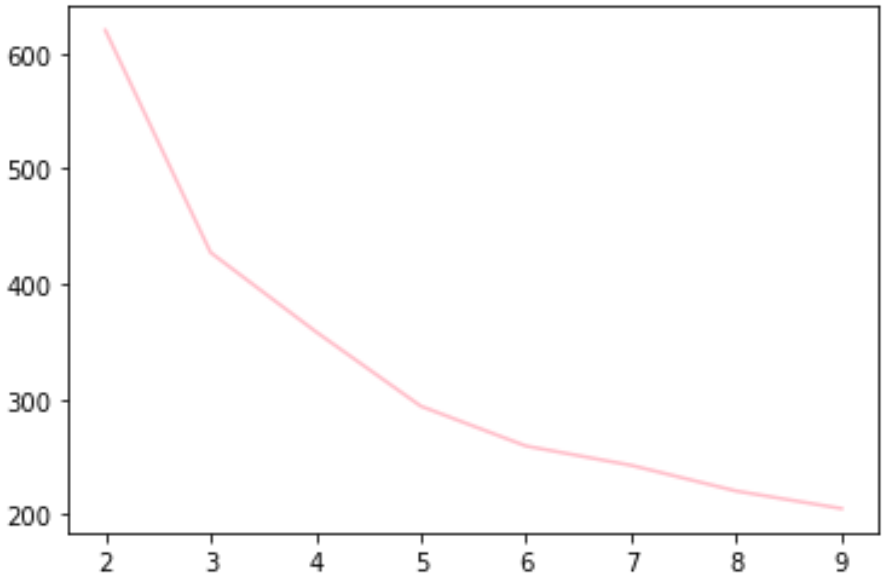
Line plot for Income vs. Child mortality and income vs. Total fertitlity





- Child mortality declines as the income of the country increases
- Countries with net income of less than 5000, 100 children die on average for every 1000 live births
- Countries with net income 22000, child mortality rate almost has a plateau with less than 25 deaths for every 1000 live births
- Families of Countries with net income higher than 20000 tends to have less than 3 children
- Families of Countries with less income of less than 5000 mostly have more than 5 children

Choosing the K number- Elbow test and Silhoutte Score

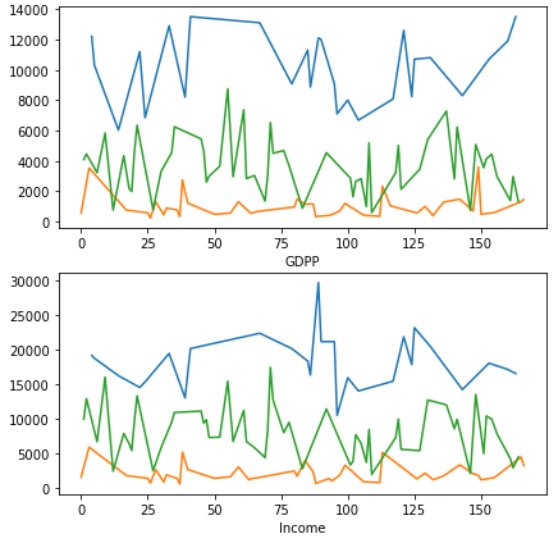


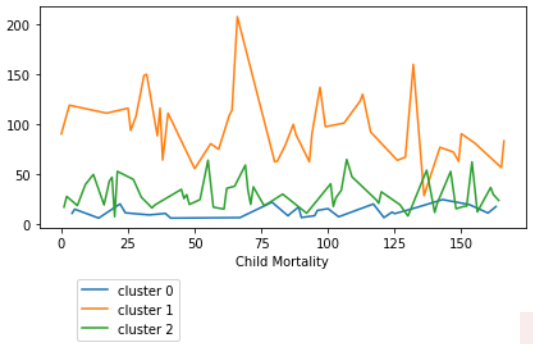
Here, 3 seems to be a good number by comparing both elbow method and silhouette score but we can also verify by using 4 clusters

```
for n_clusters= 2, the silhoutte score is: 0.3882853947975552
for n_clusters= 3, the silhoutte score is: 0.3683534308559713
for n_clusters= 4, the silhoutte score is: 0.375116742473736 44
for n_clusters= 5, the silhoutte score is: 0.3363484500395424
for n_clusters= 6, the silhoutte score is: 0.2689742417250885
for n_clusters= 7, the silhoutte score is: 0.2630603016026452
for n_clusters= 8, the silhoutte score is: 0.2153514904432701
for n_clusters= 9, the silhoutte score is: 0.218297820962218 36
```

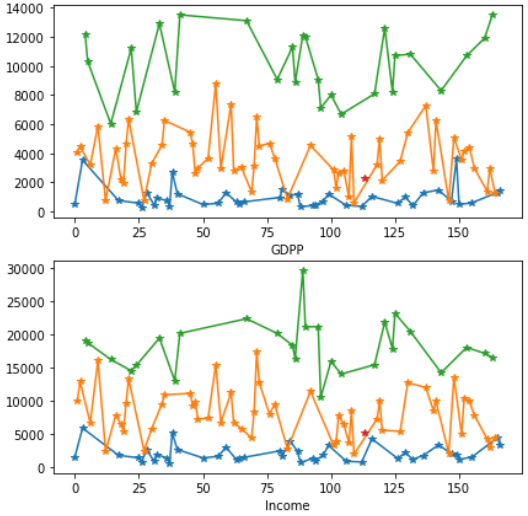
K means

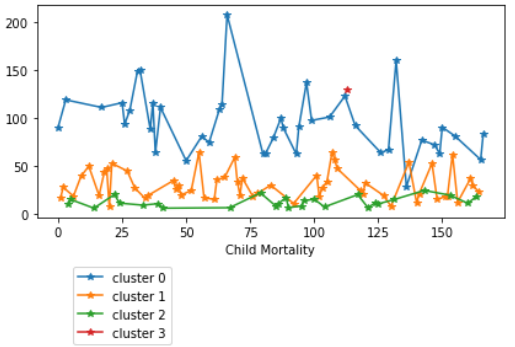
K = 3





K = 4

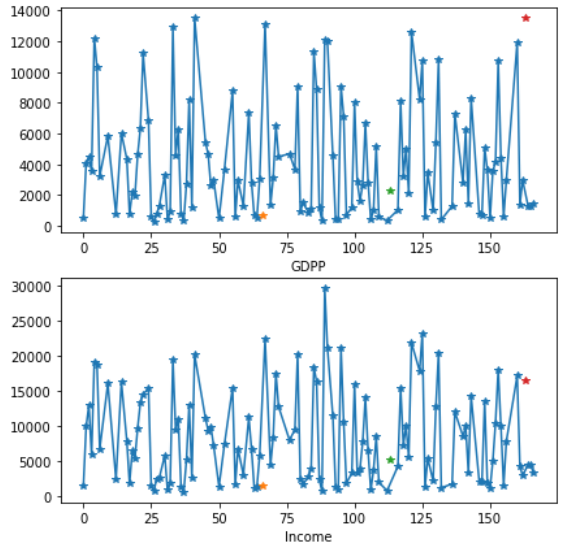


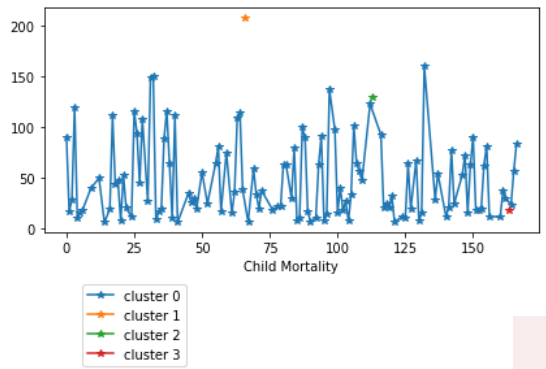


when modelling is done for 4 clusters, the 4th cluster has only one single data point, and that data point seems to be a part of a cluster (cluster that is formed in the middle) and does not variation from that cluster as well. Therefoew, 3 seems to be perfect number

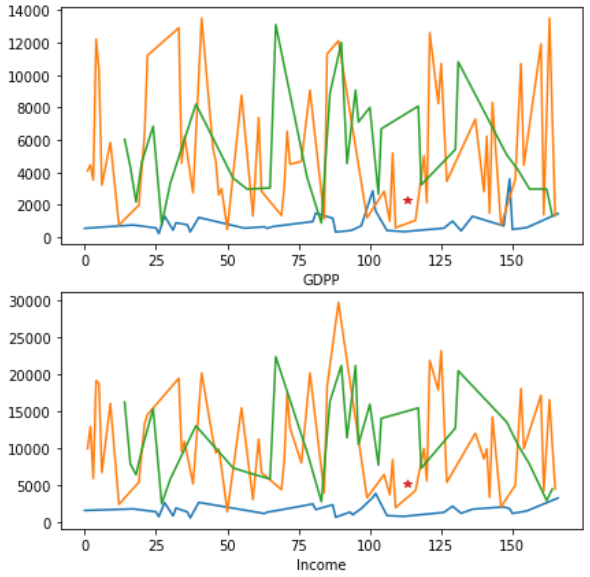
Hierarchical Clustering

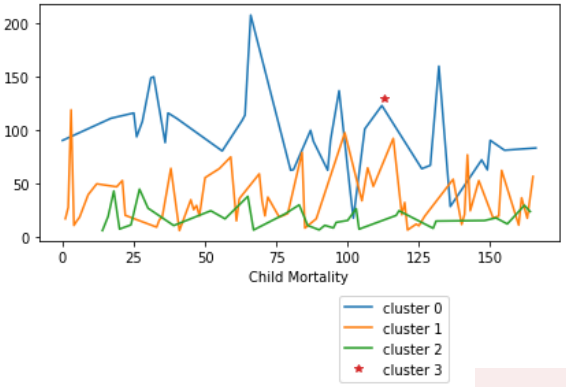
Single Linkage





Complete Linkage



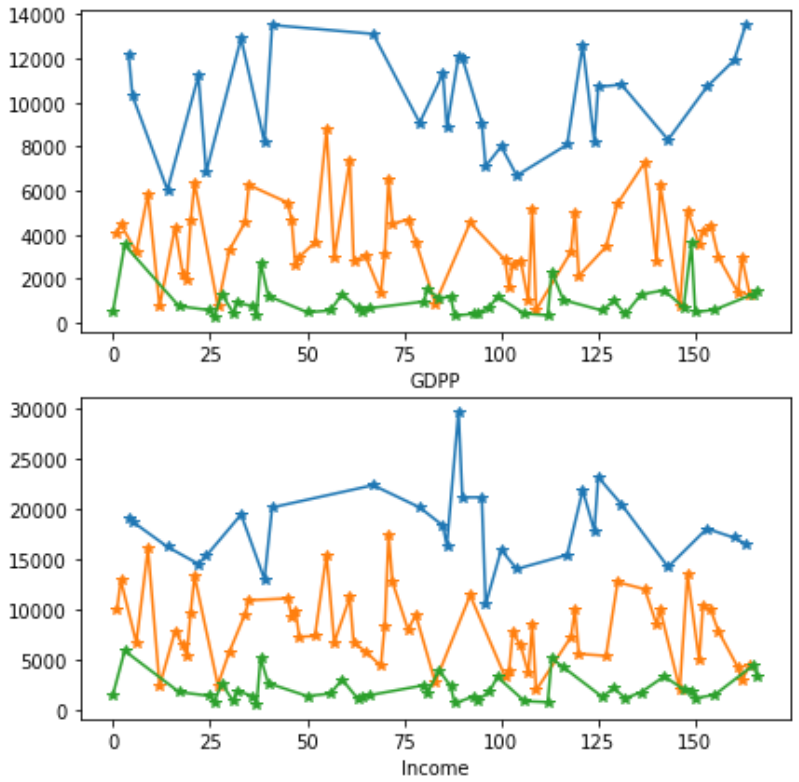


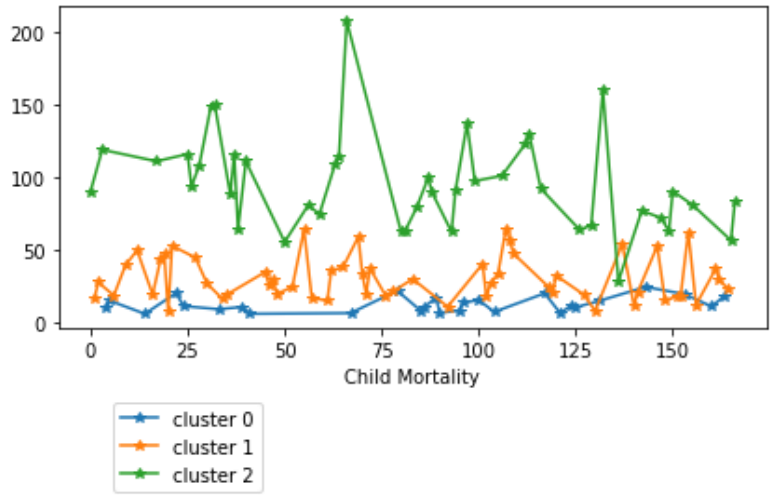
Even with complete linkage hierarchical clustering did not perform well in this data as we can see in the above graphs, hierarchical clustering failed to form unique clusters as clusters 2 and 3 from above graphs, show they overlap each other.

On the other hand we see K-means clustering can perfectly form 3 clusters, that are different from each other

Therefore the final model chosen is K means clustering with 3 clusters

Final Model - K means (K = 3)





- The 3 clusters have been formed, where each cluster can be easily differentiated from one another
- Cluster 0 are countries which has great GDPP, income and least child mortality
- Cluster 1 are countries which has GDPP and income not as great as cluster 0 and not worse as cluster 2
- Cluster 2 are countries which are in dire need of help since their GDPP and income is bad and child mortality rate is worse than others

country	gdpp	income	child_mort
Burundi	231.0	764.0	93.6
Liberia	327.0	700.0	89.3
Congo, Dem. Rep.	334.0	609.0	116.0
Niger	348.0	814.0	123.0
Sierra Leone	399.0	1220.0	160.0
Madagascar	413.0	1390.0	62.2
Mozambique	419.0	918.0	101.0
Central African Republic	446.0	888.0	149.0
Malawi	459.0	1030.0	90.5
Eritrea	482.0	1420.0	55.2

Top Countries which need help

The Final list of countries is chosen based on sorting of countries from cluster 2 based on sorting the countries where gdpp and income are sorted in ascending and child mortality is sorted in descending

The Final list of countries are:

'Burundi',
'Liberia',
'Congo, Dem. Rep.',
'Niger',
'Sierra Leone',
'Madagascar',
'Mozambique',
'Central African Republic',
'Malawi',
'Eritrea'

Thank You

- Srihari R (sriharikrishna06@gmail.com)