**DONE BY:**

**V.V.S.K. KARTHIK**

CLUSTERING

|  |
| --- |
| UNIVARIATE ANALYSIS on the Data   * Categorical Unordered UNIVARIATE analysis * Categorical ordered UNIVARIATE analysis   UNIVARIATE Analysis helps in fetching insights from single variable which helps in Overall Analysis to identify the distribution of each feature in the data |
|  |



**Distribution on**

**Child\_Mortality**



Child mortality tells us

about the count of deaths

occurred of

children under

5

years of age per 1000 live

births

in a country



We have few observations

where child mortality rate

is greater than 150 which is

a serious consideration in

our outcome.



High

childmortality

indicates countries which

may indicate poor

countries

or less advanced countries

to save the children death

due to various health

problems.



This

feature is strong

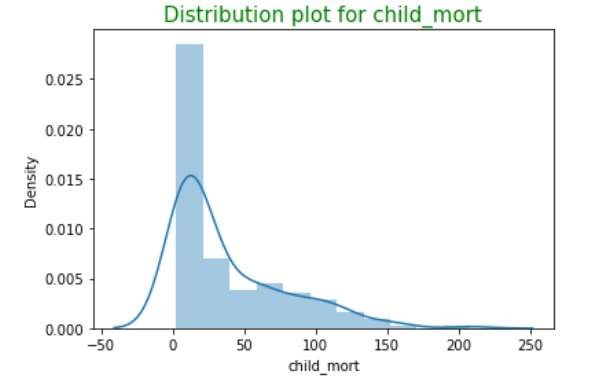
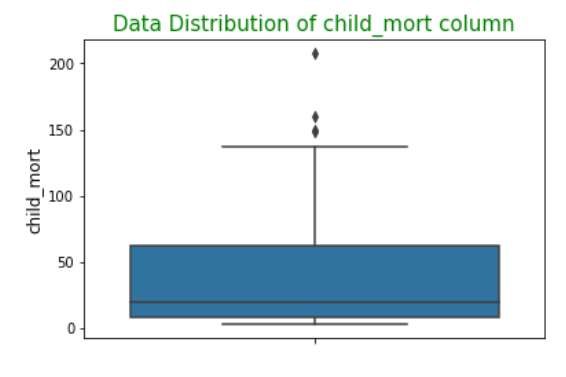
indicator

who

are required

in

aid on priority



**Distribution of**

**Imports Feature**



Import feature says about

the

Imports of goods and

services per capita.



As

shown in the

boxplot

and distribution of Imports

feature we could see data is

right skewed. Since some

of the countries have higher

imports lets treat the right

skewed data since it may

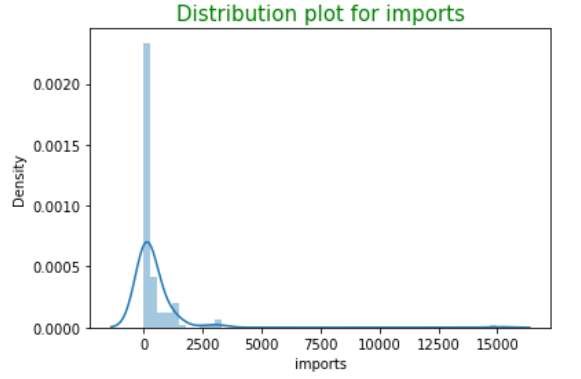
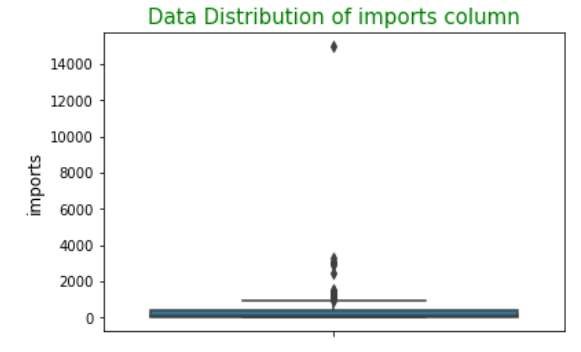
affect our analysis



We will cap the outliers

since we have very less

amount of data.



**Distribution of**

**Exports Feature**



Export feature says about

the e

xports

of goods and

services per capita.



As

shown in the

boxplot

and distribution of exports

feature we could see data is

right skewed. Since some

of the countries have higher

exports lets treat the right

skewed data since it may

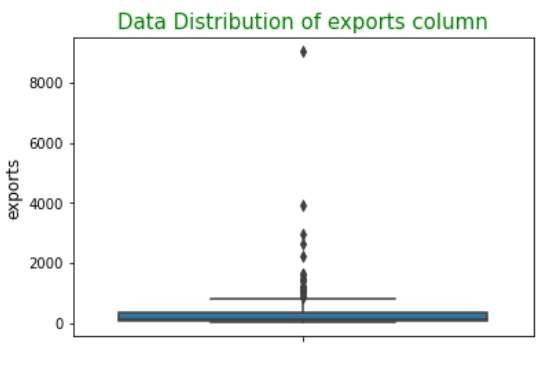
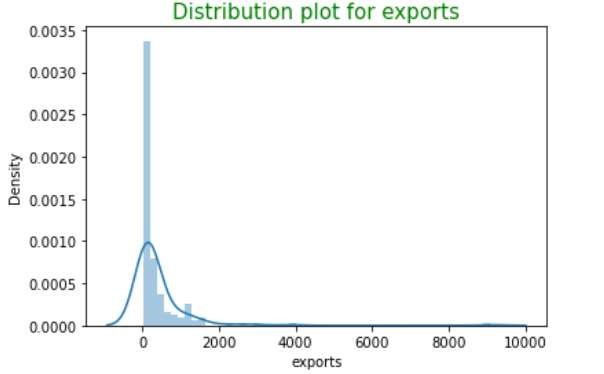
affect our analysis



We will cap the outliers

since we have very less

amount of data.



**Distribution of**

**income Feature**



As

observed

in the box

plot shown income

feature has

huge set of

outliers. Considering this

may affect our analysis

Lets cap them

at

0.95

. So

the model will not be

affected by this

skewness



**Income**

has an outlier

which looks suspicious at

greater than 120000

.We

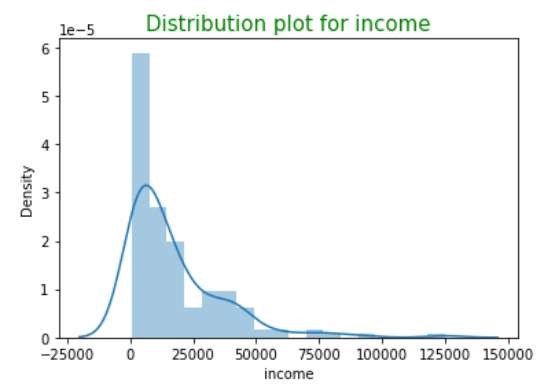
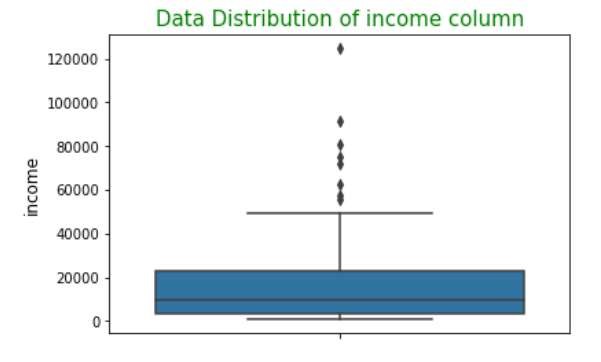
have capped them to

avoid

skewness

of the

model.



**Distribution of**

**GDP Feature**



Gdpp

tells about the

GDP

per capita. Calculated as

the Total GDP divided by

the total population.



GDP of the country is one

of the criteria in

understanding the

requirement or identifying

the needy people where we

can concentrate in funding

or helping from

**HELP**

**International**



gdpp

feature is

having

some good amount of

outliers

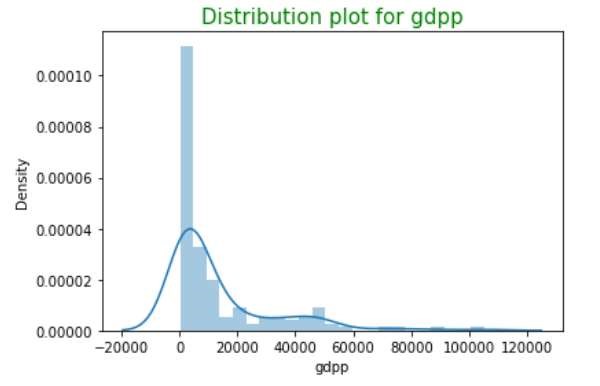
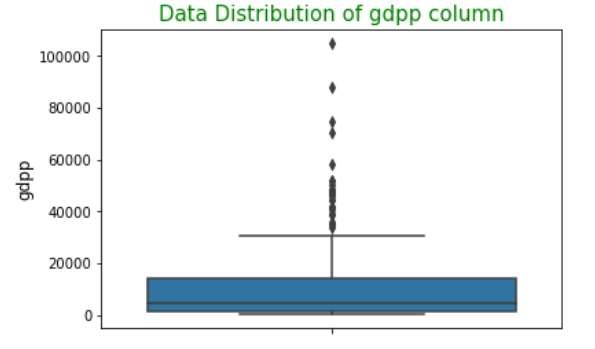
. We have capped

them since the model will

not be affected with the

skewness

of the feature



**Distribution of**

**life\_expec**

**Feature**



Life

expectancey

feature

says about the average

number of years a new born

child would live if the

current mortality patterns

are to remain the same



As

shown

in the plot

**life\_expec**

feature is left

skewed

. Average

human

life expectancy for some

counries

looks weird we

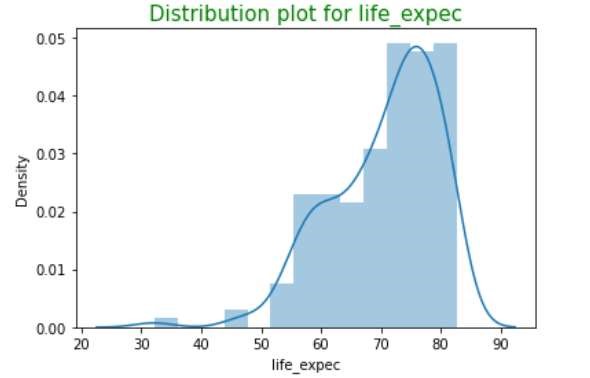
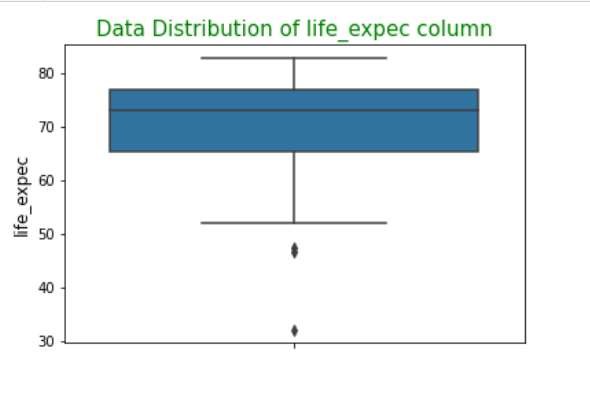
need to treat the left

skewed data where outliers

are

present

.



**Distribution of**

**total\_fertility**

**Feature**



**total\_fer**

column

tells

about

t

he

number of

children that would be

born to each woman if

the current age

-

fertility

rates remain the same.



**total\_fert**

column has

some good spread of

data and nothing

suspicious other than

value near 7

.



Possibility

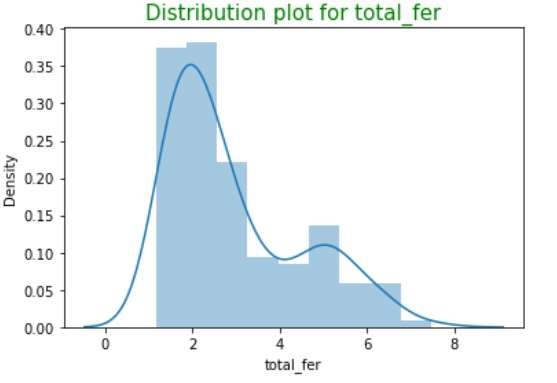
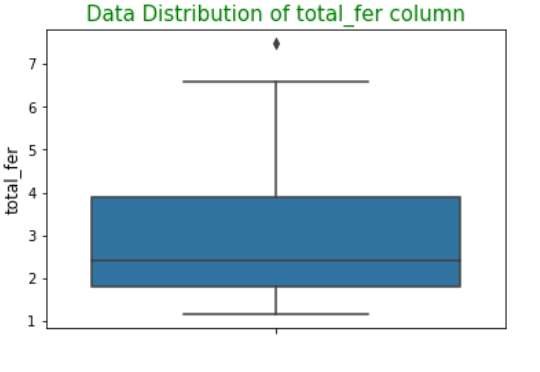
in fertility.

so lets understand in

further plots to treat if

required

.



**Distribution of**

**health Feature**



**Health**

feature is all

about

the total

health

spending per capita

.



As

shown in the

boxplot

health feature

has

outlier at the value near

40000

which look

suspicious

.



We could see that

distribution of health

feature is right skewed

Lets

cap them at 0.95

percentile

in further

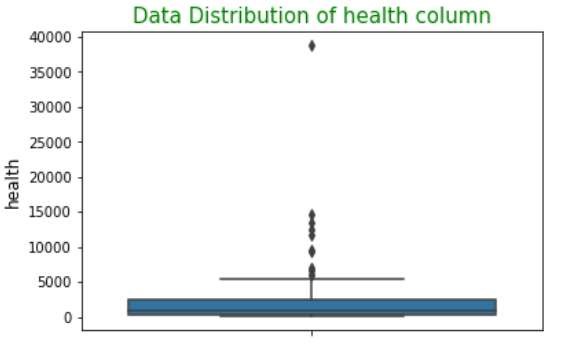
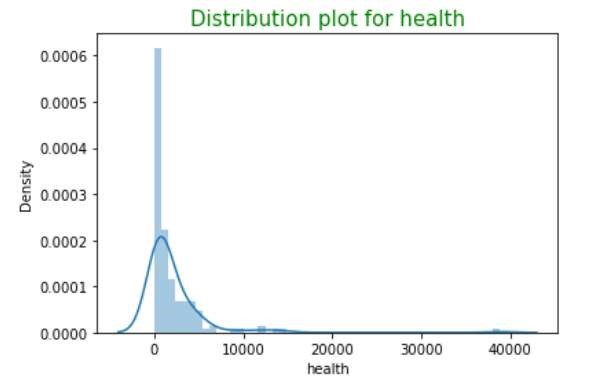
steps

to avoid the model

affected by

skewness

.



|  |
| --- |
| Skewed Data Treatment and Analysing Finding  Correlated variables  • Skewness of the data will affect the model’s Perfomance. Lets treat them and analyse the high correlated variables |
|  |



**Skewed Data**

**Treatment for**

**feature analysis**



As we observed in

previous steps some of

the features are skewed

towards either

positively or negatively

skewed. Which will

affect the analysis so we

treated them and

applied clustering on

that.



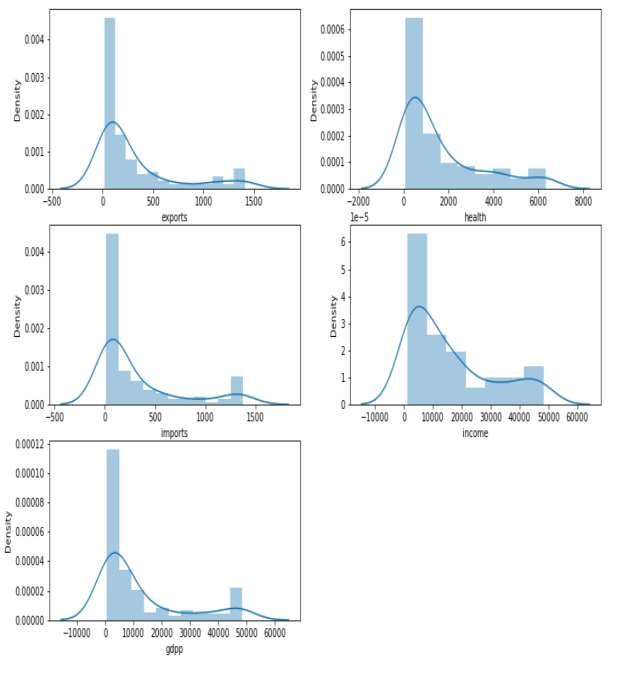
Image show the data

after the

skewness

which we applied

**.**



**Correlation**

**Matrix analysis**



As we can see that health

and income are highly

correlated variables with

0.96



Income and

gdpp

has the

correlation value of 0.94



Imports and exports has

the correlation value of

0.93



Health and

gdpp

has the

correlation value of 0.92



Child mortality and life

expectancy has the

correlation value of

-

0.89



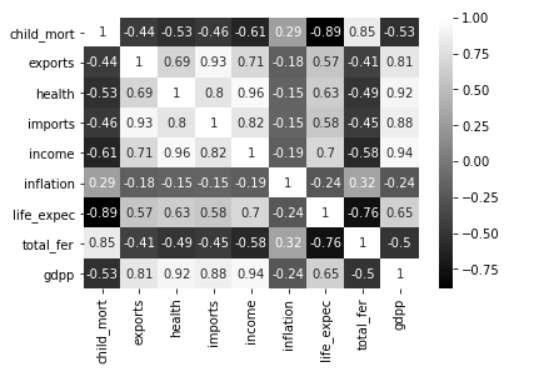
Using this correlation

values business can take

decisions further which

helps in understanding the

correlated variables.



**Correlation**

**Matrix analysis**



As we observed in the image we

are displaying the highest

correlated variables which helps

in understanding the variables.



As we can see that health and

income are highly correlated

variables with 0.96



Income and

gdpp

has the

correlation value of 0.94



Imports and exports has the

correlation value of 0.93



Health and

gdpp

has the

correlation value of 0.92



Child mortality and life

expectancy has the correlation

value of

-

0.89



Using this correlation values

business can take decisions

further which helps in

understanding the correlated

variables.



Top Correlated Variables

|  |
| --- |
| BI-VARIATE ANALYSIS on the country Data   * Numeric – Numeric analysis * Numerical – Categorical analysis * Categorical – Categorical analysis   BIVARIATE Analysis helps in fetching insights by looking into multiple variables which helps in gaining Overall Insights from the hidden data |
|  |



**Income**

***versus***

**GDP**



Image show the relation

between the GDP and

net income of the

person



As

shown in

the

plot

**GDP**

and

**i**

**ncome**

has

highest correlation as

the income increases

gdpp

also increases

linearly.



We can consider GDP

in our clustering to

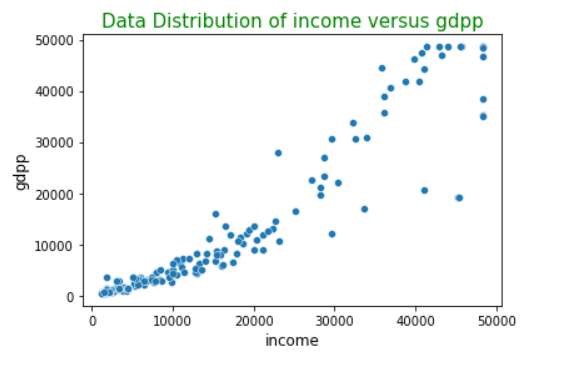
understand how the

countries

gdp

helps in

identifying our outcome



**Income**

***versus***

**Health**



Image show the relation

between the Health

nd

net

income of the person



As

shown in the plot

Health and income has the

high correlation.



As

shown in

the

plot

GDP

and `income` has highest

correlation as the income

increases

GDP also

increases

linearly.



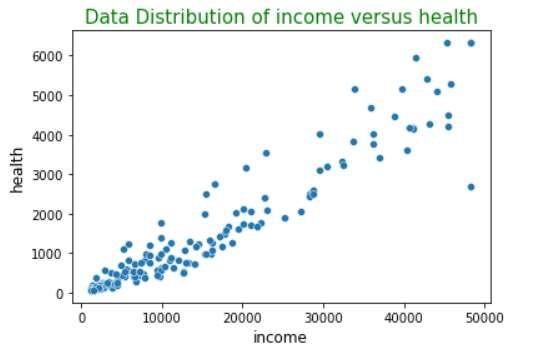
We can consider GDP in

our clustering to understand

how the countries GDP

helps in identifying our

outcome



**Imports**

***versus***

**Exports**



Image show the relation

between the Imports and

Exports of different countries



As shown in the plot exports

and imports have high

correlation.



As the imports increases

Exports also

increases

but

further analysis is required on

this



There is a equal chances for

countries to have higher

import and export values

since all the countries might

not be good in all the

goods and due to lack of

resources there is a goods

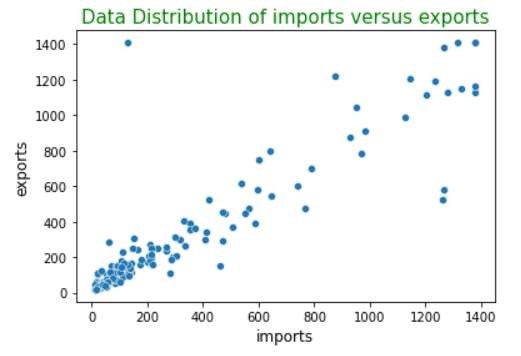
exchange

takes

place between

countries

.



**Total Fertility**

***versus***

**ChildMortality**



Image show the relation

between the total fertility and

child

mortality



As shown in the plot total

fertility and child mortality

have positive correlation.



**total\_fer**

tells about the

number of children that

would be born to each

woman if the current age

-

fertility rates remain the

same.



As

the

**total\_fer**

increases

**child\_mort**

increases this is

one of the reason where

parents are unable to take

care of

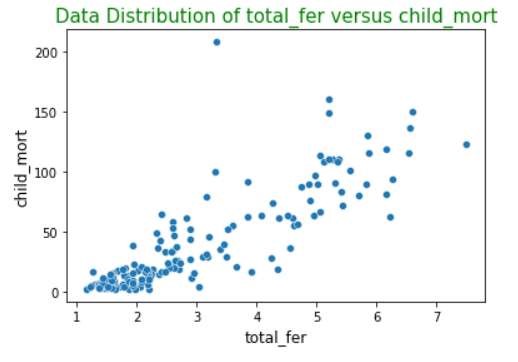
childrens

health

condition due to which child

mortality increases in some

countries.



**Health**

***versus***

**ChildMortality**



Image show the relation

between the Health and

child mortality



As shown in the

plot

we

could understand that when

the `health`(Total health

spending per capita)

increase

child\_mort

decreases.



Less

spending of `health`

Increase in `

child\_mort

`.



This is quite obvious when

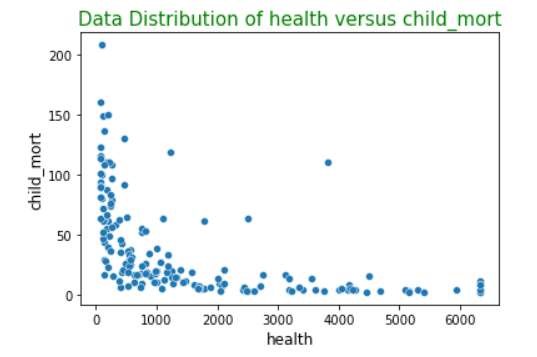
people who are financially

week they may don’t have

health care expenditure due

to which child mortality is

increasing



|  |
| --- |
| HOPKINS Test   * Hopkins test is done to check how well our data is suitable for Clustering * Since we want to conduct our clustering on gdpp, child\_mort and income features we will do hopkins test on the same * As shown in the image below we have conducted hopkins test 10 times to make sure that our data is well suited for clustering purpose. * Every time value is greater than 0.87 which is a good indication for clustering algorithm. |
|  |

|  |
| --- |
| Optimal K value using within cluster sum of squares Method   * Finding the optimal k value using within cluster sum of squares method. * One among the k value i.e.., k=2 and k=3 can become the optimal values. * we analysed the data with both the k values and when k=3 data is well explained by the model and we could see some good interpretations or spread of the data. |
|  |

|  |
| --- |
| Optimal K value using Silhouette Method     * Finding the optimal k value using Silhouette Method. * As observed in the above analysis we have better silhouette score when k=2: 0.65 and k=3: 0.52 * As we already observed the k value using ssd tells the same analysis but we saw better distribution of clusters when k=3 in the above step. Lets consider k=3 as optimal value and visualise the spread of clusters. |
|  |

|  |
| --- |
| Distribution of Variables when 3 clusters are formed  Plot 1   * As shown in the plot 1 we have created three clusters displaying different income and child\_mort values.Now after creation of clusters data is well explained   Plot 2   * As shown in the plot 2 we have created three clusters displaying different income and gdpp values. Now after creation of clusters data is well explained |
|  |

|  |
| --- |
| Distribution of Variables when 3 clusters are formed  Plot 3   * As shown in the plot 3 we have created three clusters displaying different gdpp and child\_mort values.Now after creation of clusters data is well explained. * Cluster 0(zero) is where child\_mortality is high where we need to concentrate |
|  |

|  |
| --- |
| Distribution of gdpp,child\_mort and income across 3 clusters  • As shown in the plot cluster 0 has high child\_mortality and less GDP and income. We can concentrate on this countries where we have this properties. |
|  |

|  |
| --- |
| Kmeans: Distribution of gdpp,child\_mort and income across 3 clusters and Top 5 countries requires help  PLOT 1   * As shown in the plot 1 cluster 0(zero) has high   child\_mortality and less GDP and income. We can concentrate on this countries where we have this properties.   * Table 1 gives the top 5 countries which requires help   on priority due to high  TABLE 1 child\_mort,less gdpp and  income.   * HELP International can concentrate on this countries on priority. |
|  |

|  |
| --- |
| Agglomerative clustering  Here we consider each data point as one cluster and we find the distance from each point to the other data points.Then the data points with minimum distance is formed as a cluster resulting in n-1 cluster. This process iterates until all the data points are formed as a single cluster.  Single Linkage |
|  |

|  |
| --- |
| Agglomerative clustering Complete Linkage    -> When compared with Single and complete Linkage method. Complete linkage explains the data better and some good clusters can be formed which we can understand from the dendograms obtained. |
|  |

|  |
| --- |
| Hierarchical clustering optimal cluster value which is obtained from dendograms  PLOT 1  -> As observed in plot 2 we have 96 countries which fall under cluster 0, 31 countries fall under cluster1 and 40 countries fall under cluster 2.  PLOT 2  -> As observed in plot 2 we have 96 countries which fall under cluster 0, 31 countries fall under cluster1 and 40 countries fall under cluster 2. Data is well explained in cluster 3 considering GDP,income and child mortality |
|  |

|  |
| --- |
| Hierarchical clustering: Distribution of gdpp,child\_mort and income across 3 clusters and Top 5 countries where HELP International can concentrate into   * As shown in the plot 1 cluster   PLOT 1  0(zero) has high child\_mortality and less GDP and income. We can concentrate on this countries where we have this properties.   * Table 1 gives the top 5 countries which requires help on priority due to high child\_mort,less gdpp and income. * HELP International can concentrate on this countries on priority.   TABLE 1 • Country list : Haiti,Sierra  Leone,Chad,Central African Republic and Mali are the top 5 countries which Help internationcla can concentrate based on the study of data. |
|  |

CONCLUSION

•

HELP International should concentrate on following countries

based on the categories we studied:

•

Help

international

can

concentrate

on

the

countries

like

Haiti,Sierra

Leone,Chad,Central

African

Republic

and

Mali

.

Which

are

top

5

countries

which

require

NGO

help

to

save

them

during

difficult

times

and

Natural

calamities

.

•

As

per

the

analysis

we

did

we

observed

that

countries

like

Nigeria,Niger,Angola,Conga

and

Burkina

Faso

are

the

countries

which

are

in

top

10

required

NGO

help

.

