Clustering Assignment Submission

Problem Statement

- Aim is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Identify the set of countries that need immediate aid and funding based on the condition
- Suggest the countries which the CEO needs to focus on the most.

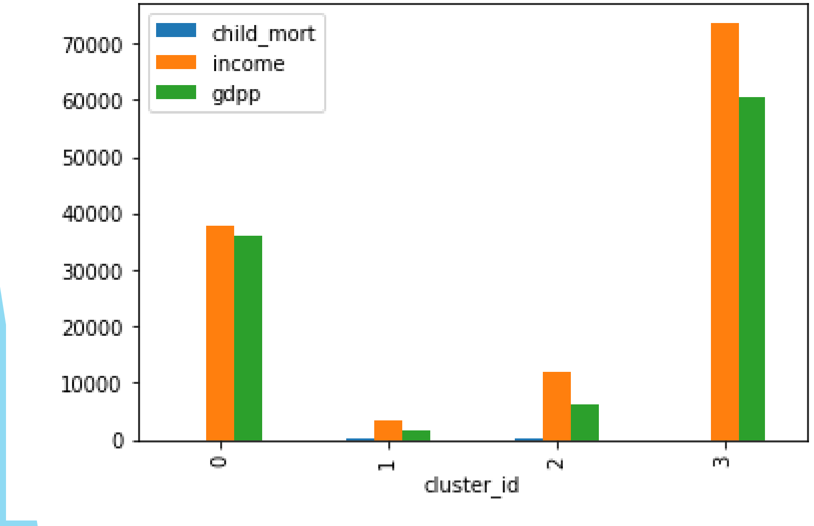
Analysis approach

The approach followed is below:

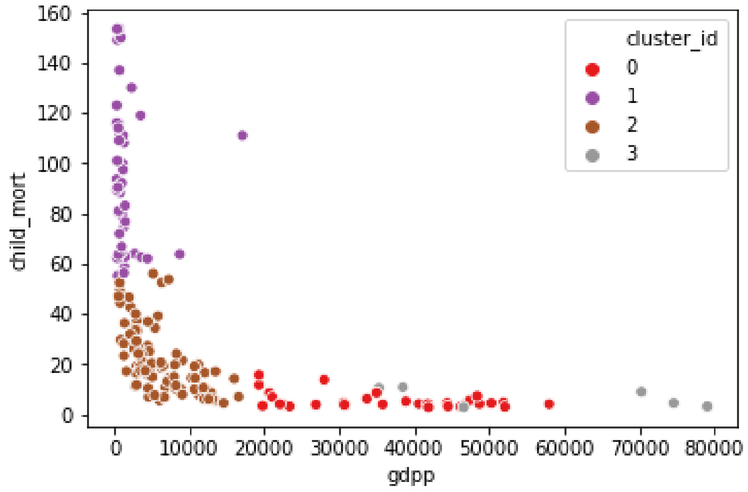
- Understand the data , cap the outliers.
- Identify the number of clusters to be used .
- Apply clustering – kmeans and Heirarchial , visualize the clusters for better understanding
- Once you get clusters identify the cluster that has high values of child_moratality and low gdpp and income as these the the countries that are in need of aid
- Once the cluster is identify bin the values and sort wrt to ascending (gdpp n income) and descending (child mortality)
- Take different combination and apply to above steps.
- Based on the sorted data decide and pick Countries that need most help.

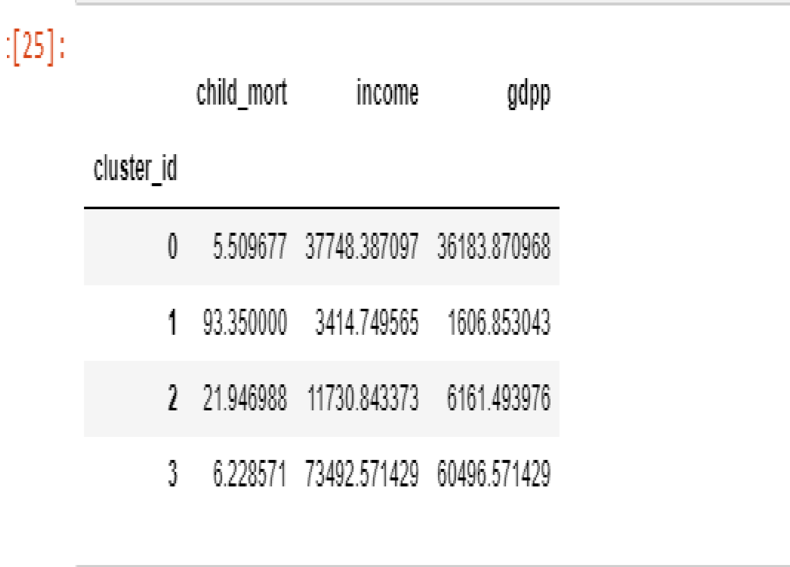
Results of Clustering Model

- Result 1: Identifying the correct cluster – bases on below visualization I identified cluster 1 is best for further analysis as it has high child mortality rate and very low gdpp , income as compared to other clusters



```
: <matplotlib.axes._subplots.AxesSubplot at 0x1c012f101d0>
```

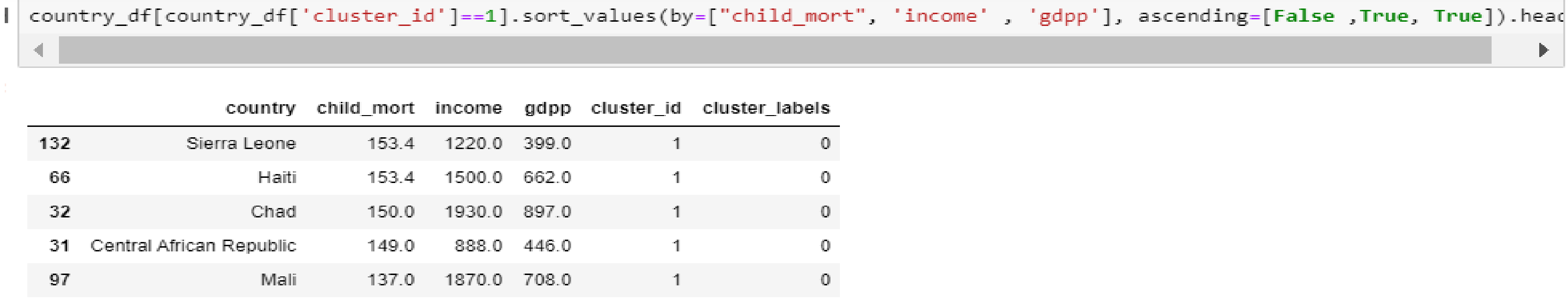




:[25]:

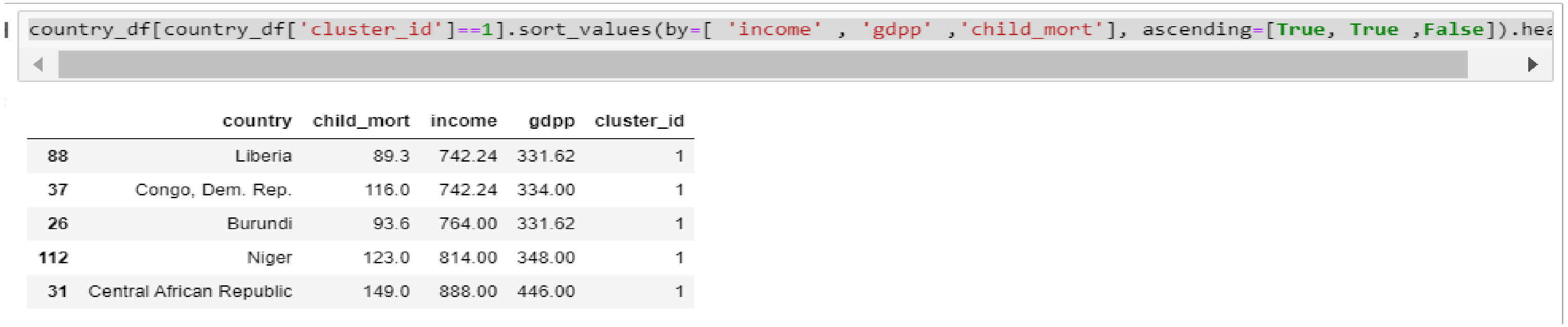
cluster_id	child_mort	income	gdpp
0	5.509677	37748.387097	36183.870968
1	93.350000	3414.749565	1606.853043
2	21.946988	11730.843373	6161.493976
3	6.228571	73492.571429	60496.571429

- Result 2: In cluster identify the countries:
- Sorted data in three combinations and identified countries that need most help



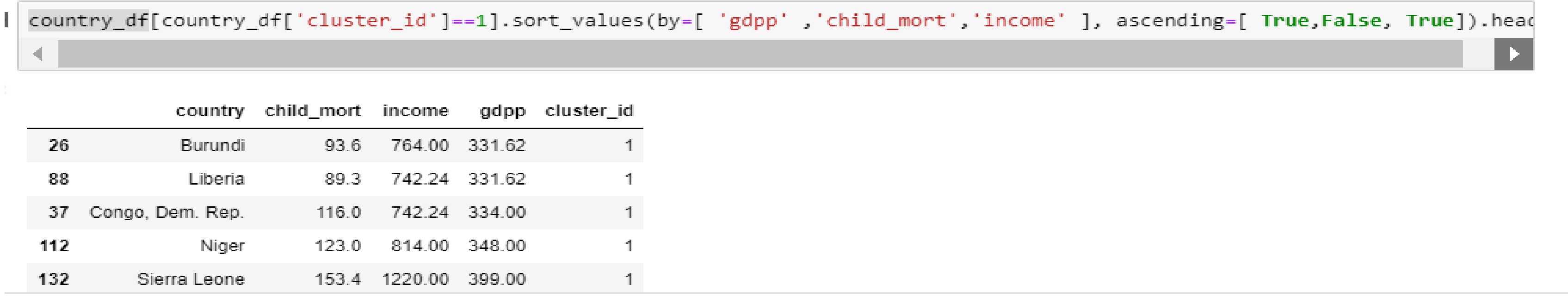
```
country_df[country_df['cluster_id']==1].sort_values(by=["child_mort", 'income' , 'gdpp'], ascending=[False ,True, True]).heac
```

	country	child_mort	income	gdpp	cluster_id	cluster_labels
132	Sierra Leone	153.4	1220.0	399.0	1	0
66	Haiti	153.4	1500.0	662.0	1	0
32	Chad	150.0	1930.0	897.0	1	0
31	Central African Republic	149.0	888.0	446.0	1	0
97	Mali	137.0	1870.0	708.0	1	0



```
country_df[country_df['cluster_id']==1].sort_values(by=[ 'income' , 'gdpp' ,'child_mort'], ascending=[True, True ,False]).hea
```

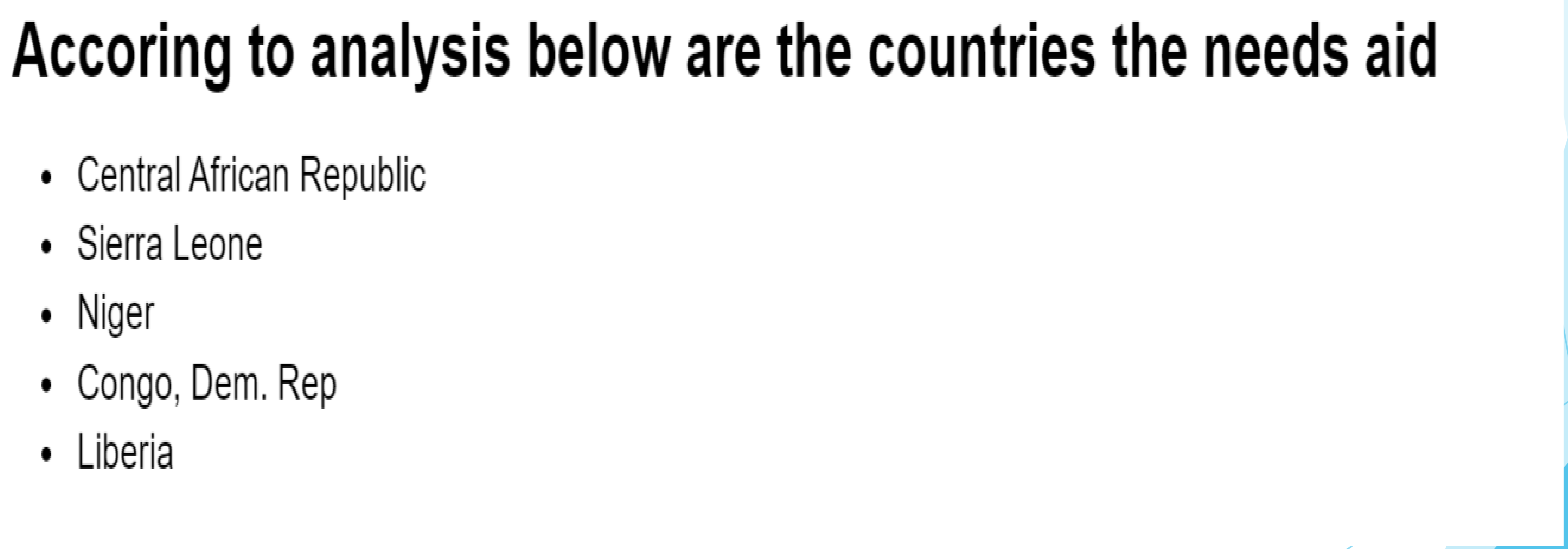
	country	child_mort	income	gdpp	cluster_id
88	Liberia	89.3	742.24	331.62	1
37	Congo, Dem. Rep.	116.0	742.24	334.00	1
26	Burundi	93.6	764.00	331.62	1
112	Niger	123.0	814.00	348.00	1
31	Central African Republic	149.0	888.00	446.00	1



```
country_df[country_df['cluster_id']==1].sort_values(by=[ 'gdpp' ,'child_mort','income' ], ascending=[ True,False, True]).heac
```

	country	child_mort	income	gdpp	cluster_id
26	Burundi	93.6	764.00	331.62	1
88	Liberia	89.3	742.24	331.62	1
37	Congo, Dem. Rep.	116.0	742.24	334.00	1
112	Niger	123.0	814.00	348.00	1
132	Sierra Leone	153.4	1220.00	399.00	1

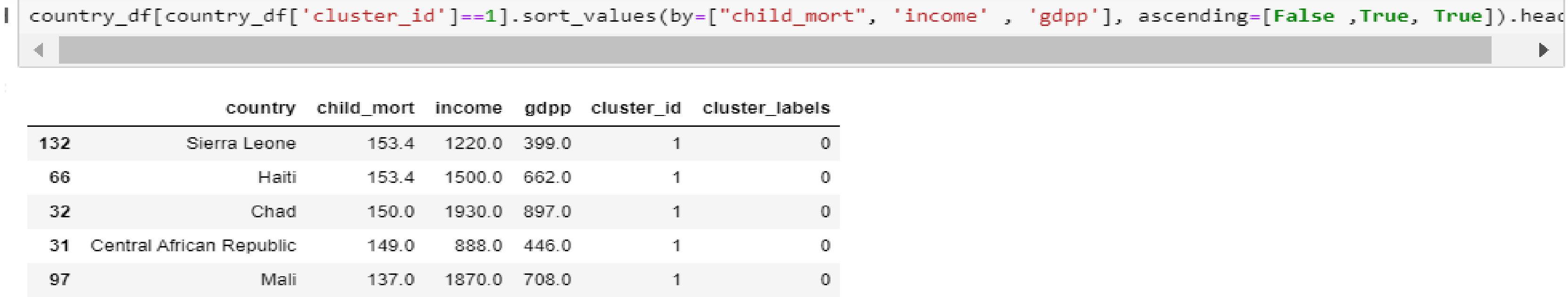
- Step 3 :
- We need to consider both health and business aspect , for example if country A has slightly high child mortality that country B , but very high gdpp that B , we need to first consider B , bases on this approach following the are suggested countries



Accoring to analysis below are the countries the needs aid

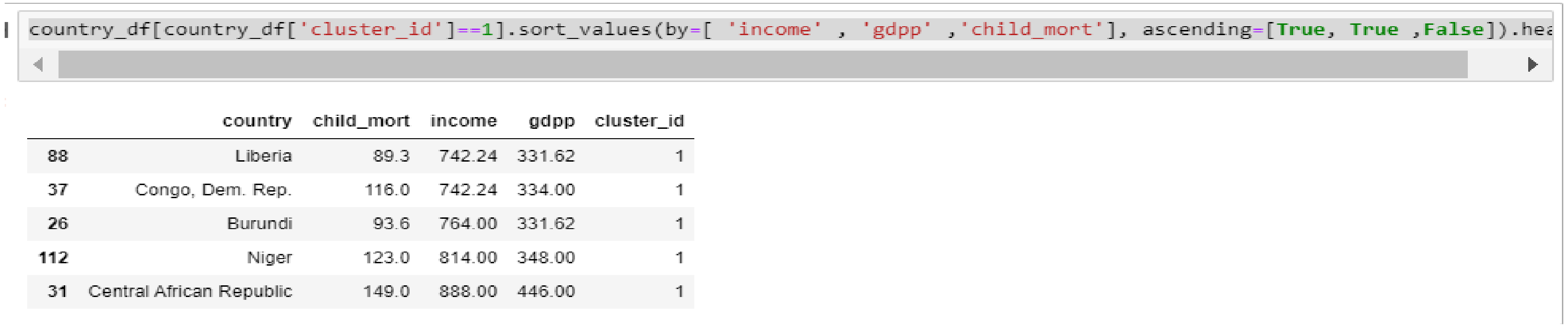
- Central African Republic
- Sierra Leone
- Niger
- Congo, Dem. Rep
- Liberia

- Result 2: In cluster identify the countries:
- Sorted data in three combinations and identified countries that need most help



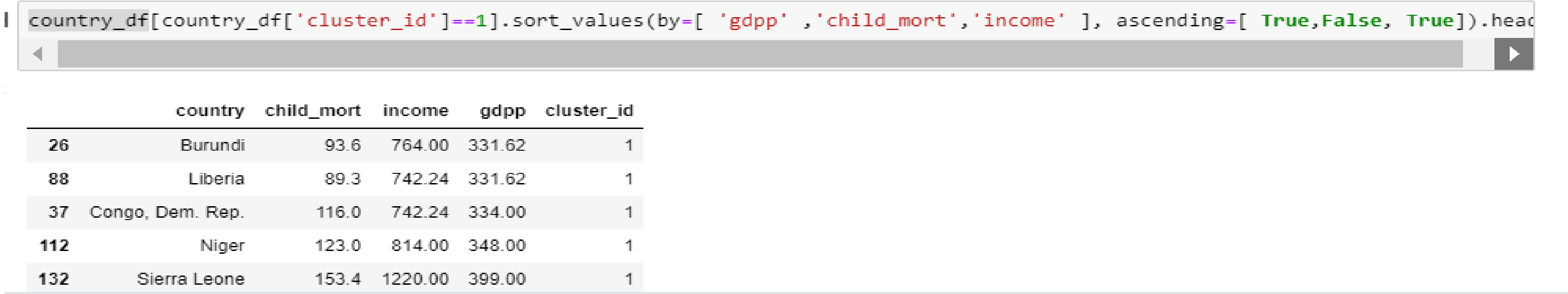
```
country_df[country_df['cluster_id']==1].sort_values(by=["child_mort", 'income' , 'gdpp'], ascending=[False ,True, True]).heac
```

	country	child_mort	income	gdpp	cluster_id	cluster_labels
132	Sierra Leone	153.4	1220.0	399.0	1	0
66	Haiti	153.4	1500.0	662.0	1	0
32	Chad	150.0	1930.0	897.0	1	0
31	Central African Republic	149.0	888.0	446.0	1	0
97	Mali	137.0	1870.0	708.0	1	0



```
country_df[country_df['cluster_id']==1].sort_values(by=[ 'income' , 'gdpp' ,'child_mort'], ascending=[True, True ,False]).hea
```

	country	child_mort	income	gdpp	cluster_id
88	Liberia	89.3	742.24	331.62	1
37	Congo, Dem. Rep.	116.0	742.24	334.00	1
26	Burundi	93.6	764.00	331.62	1
112	Niger	123.0	814.00	348.00	1
31	Central African Republic	149.0	888.00	446.00	1



```
country_df[country_df['cluster_id']==1].sort_values(by=[ 'gdpp' ,'child_mort','income' ], ascending=[ True,False, True]).heac
```

	country	child_mort	income	gdpp	cluster_id
26	Burundi	93.6	764.00	331.62	1
88	Liberia	89.3	742.24	331.62	1
37	Congo, Dem. Rep.	116.0	742.24	334.00	1
112	Niger	123.0	814.00	348.00	1
132	Sierra Leone	153.4	1220.00	399.00	1