HELP International Case Study

CLUSTERING ASSIGNMENT

ANALYSIS BY: ANKIT GOREGAONKAR

Problem Statement:

- 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. With the given dataset, find top 10 countries that are in dire need of aid.
- Furthermore, find the top 5 countries with the direst need of aid.

Approach:

- The dataset was inspected and **no null values** were found to perform missing value treatment.
- Few variables, **Income**, **Health** and **Exports** were in percentages and they were converted back to actual numbers.
- Few variables **Exports**, **Imports**, **Health** and **Income** had outliers and they were treated by computing **Mean of the Outliers**.
- The dataset was then scaled using **Standard Scaler** and **Hopkins Statistics** was used, which generated a score of **0.92**.

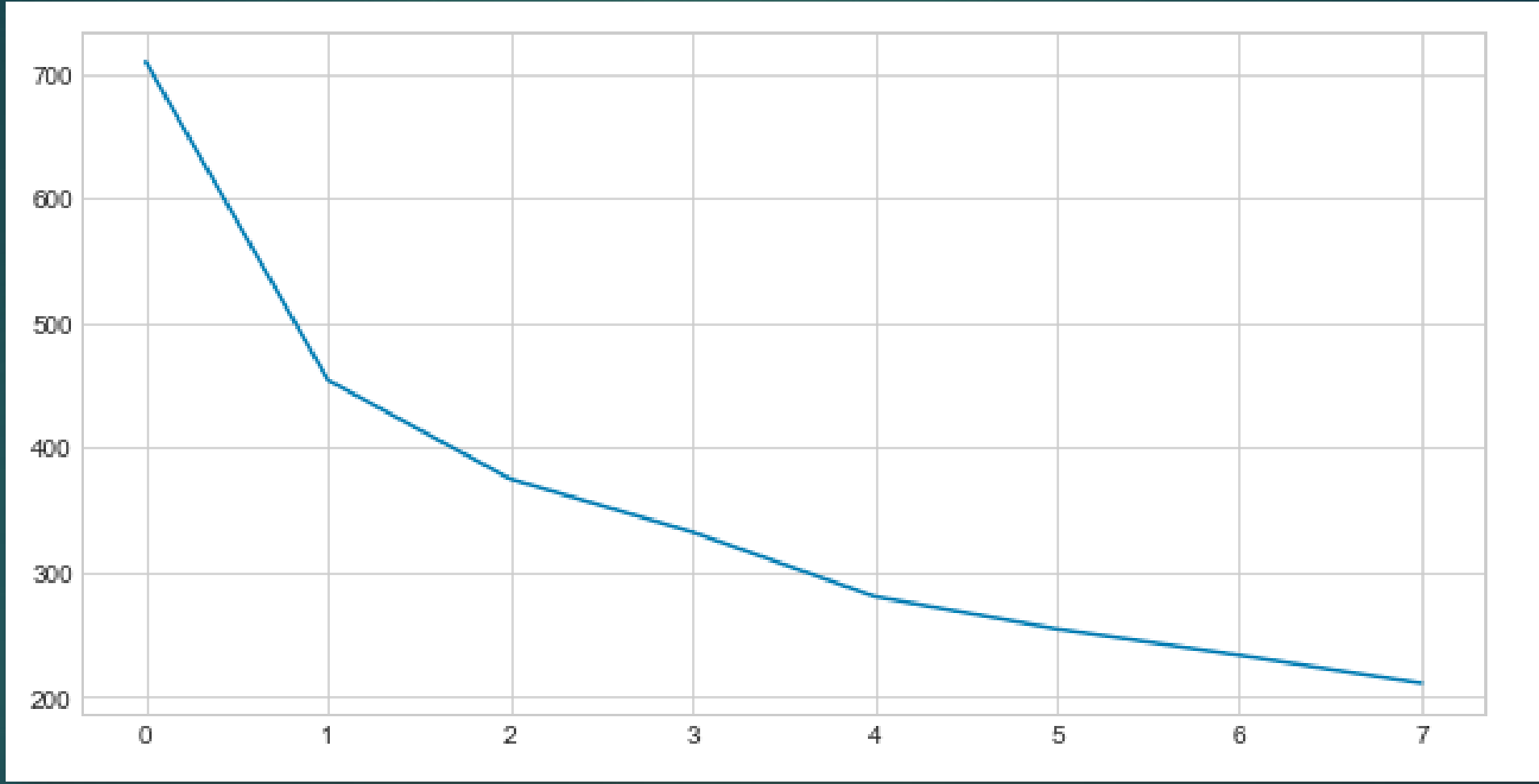
Top 10 countries that need aid: Clustering through K-Means

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali
- Nigeria
- Niger
- Angola
- Congo, Dem. Rep.
- Burkina Faso

Elaboration on K-Means Clustering approach

1. **Elbow Curve**:
   - **Sum of squared distances (SSD)** was found for each data-points based on the number of clusters
   - With elbow curve method, we could infer that the **SSD** is extremely high **(700+)** for **2** clusters. It **decreases significantly** with an **additional cluster**.
   - When we have **3** clusters, the **SSD** did not drop that significantly and therefore 3 clusters were chosen.
   - To further confirm if **3** clusters are optimum to work on, we calculated the Silhouette score (*Slide No. 6*).

Elbow Curve

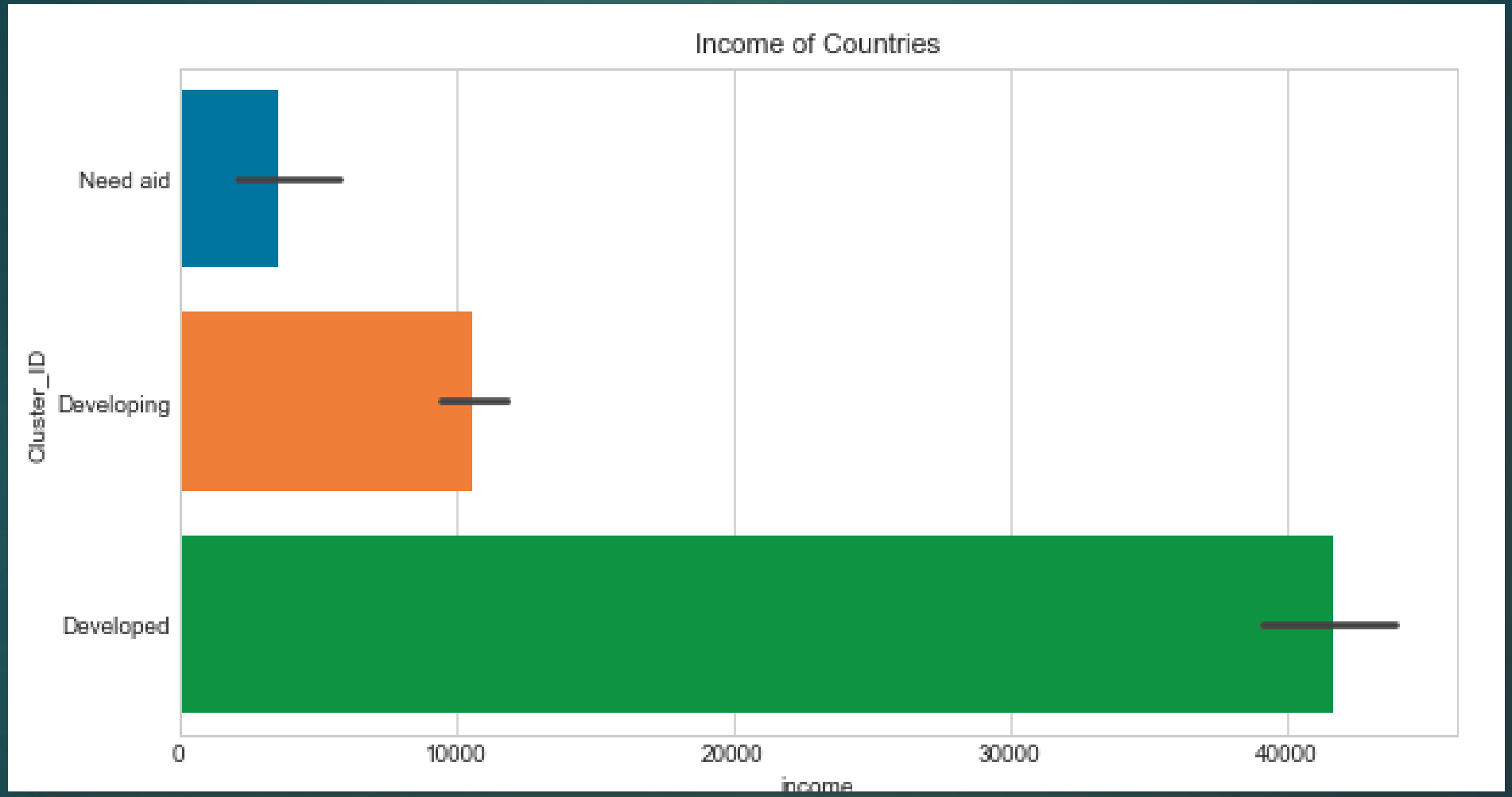


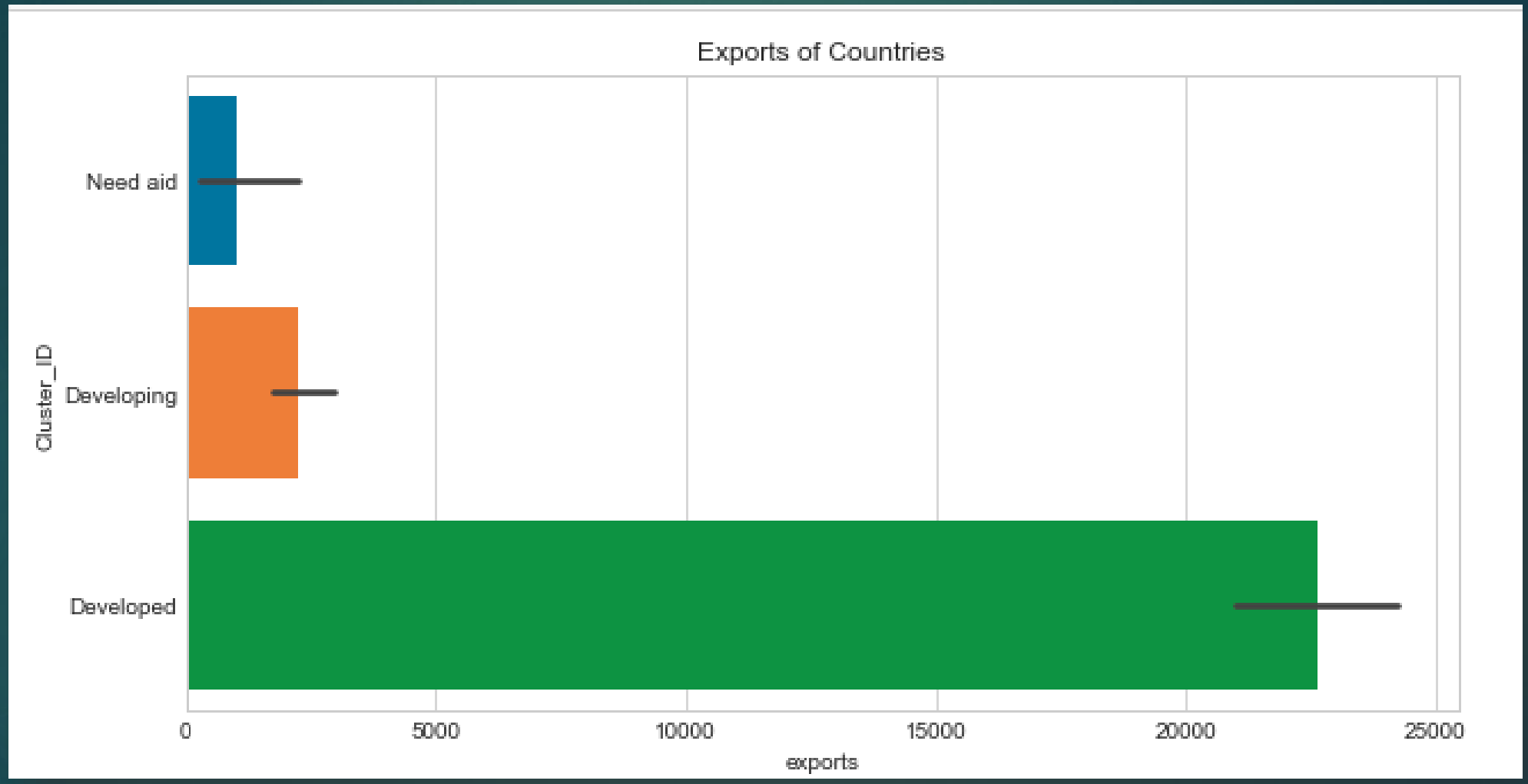
Silhouette Score

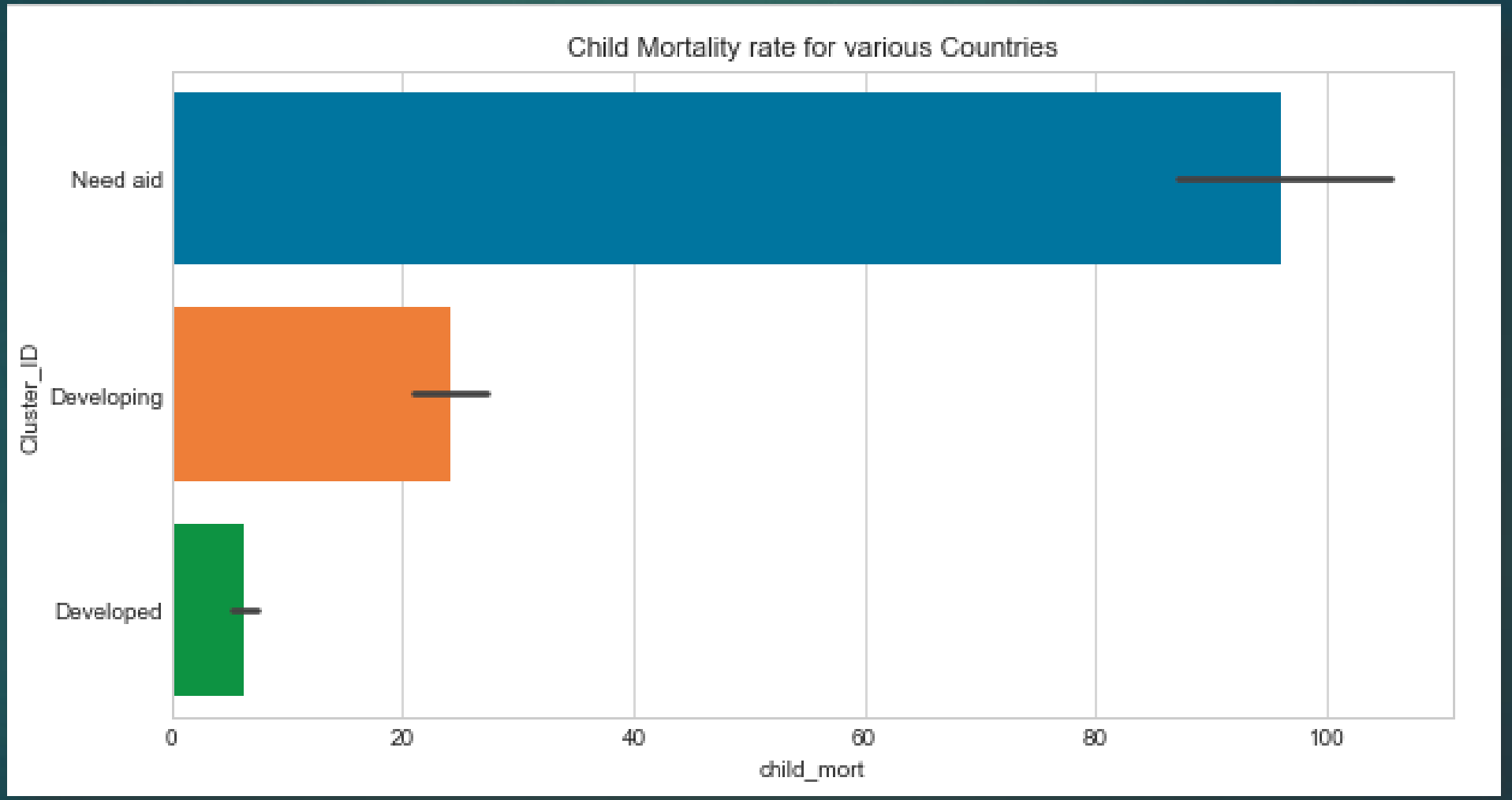
- For n cluster=2, the silhouette score is 0.5271051754280778
- For n cluster=3, the silhouette score is 0.4659769167124918
- For n cluster=4, the silhouette score is 0.46495851174674113
- For n cluster=5, the silhouette score is 0.3435271209548506 6
- For n cluster=6, the silhouette score is 0.3544710682059553 6
- For n cluster=7, the silhouette score is 0.341368444640427 7
- For n cluster=8, the silhouette score is 0.28306330286564957
- For n cluster=9, the silhouette score is 0.298927563242296 6

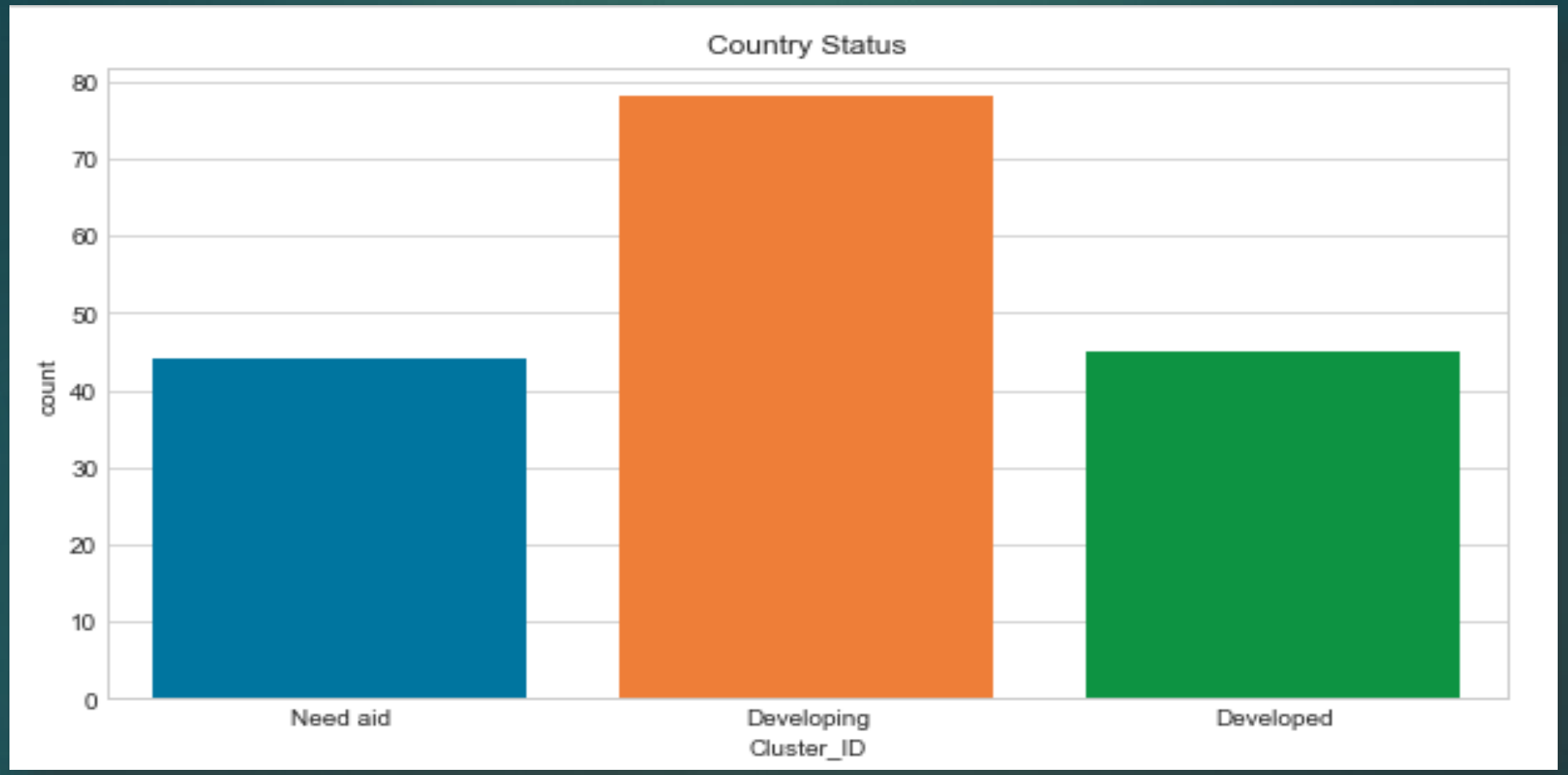
Thus we could infer that the **Silhouette score** for 3 clusters is decent and they are good to work with.

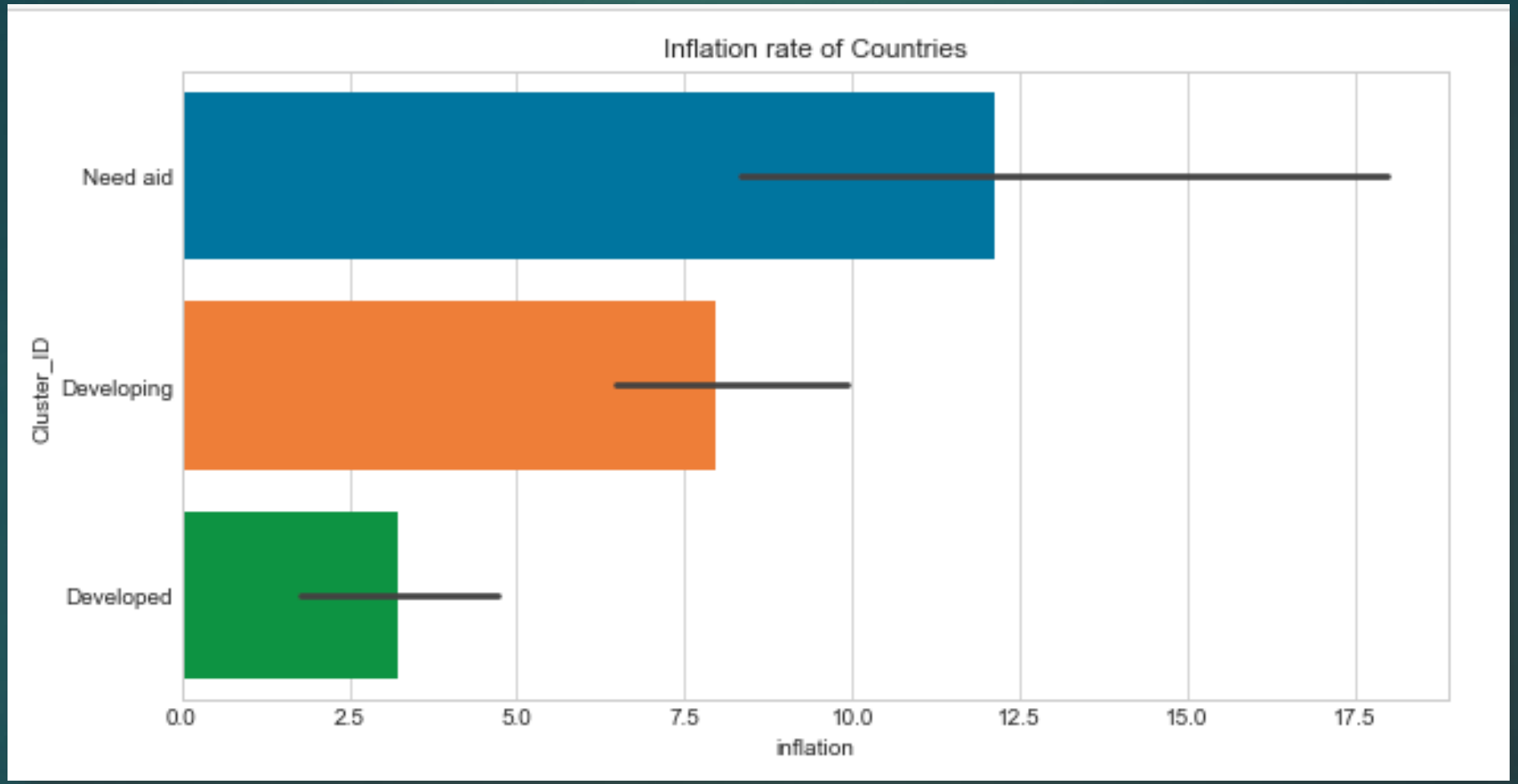
K-Means Cluster Visualisations

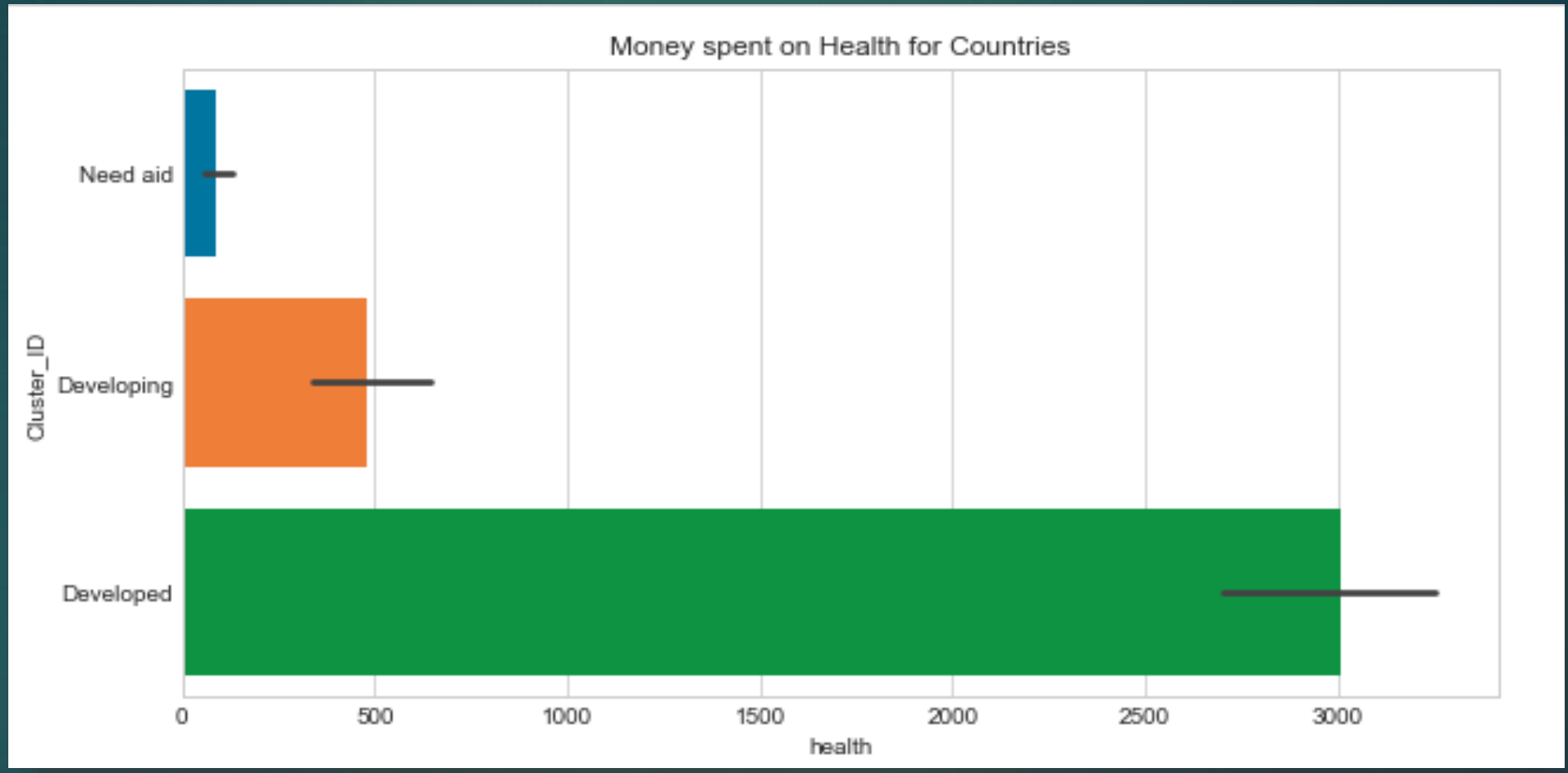



Exports of Countries
Need aid
Developing
Developed
Cluster_ID
0
5000
10000
15000
20000
25000
exports


Child Mortality rate for various Countries
Need aid
Developing
Developed
Cluster_ID
0
20
40
60
80
100
child_mort


Country Status
80
70
60
50
40
30
20
10
0
count
Need aid
Developing
Developed
Cluster_ID


Inflation rate of Countries
Need aid
Developing
Developed
Cluster_ID
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
inflation


Money spent on Health for Countries
Need aid
Developing
Developed
Cluster_ID
0
500
1000
1500
2000
2500
3000
health

Top 10 countries that need aid: Hierarchical Clustering

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali
- Niger
- Angola
- Congo, Dem. Rep.
- Burkina Faso
- Guinea-Bissau

Elaboration on Hierarchical Clustering approach

- While implementing Hierarchical clustering, '**Complete**' linkage was used to generate a well visualised dendrogram.
- After plotting the dendrogram, **3** clusters were used and the dendrogram bifurcated the data in to **3** clusters.
- '**Single**' linkage would make the linkage appear cluttered and was therefore not chosen to decipher the optimum number of clusters.

Top 5 countries with direst need of Aid

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

Thoughts behind the Country Selection:

- **Sierra Leone** has the **highest Child mortality rate** coupled with **poor exports** and **lower money spent on health**.
- **Central African Republic** has the **lowest money spent on Health**, along with **poor GDP per Capita**, **lower Imports and Exports**.
- To be continued..

- **Niger comparatively** has a **lower child mortality rate** than **Central African Republic**, but it has **poor exports**, coupled with **lower money spent on Health** and **poor Income**.
- **Guinea Bissau** has **higher imports** but it is **poor** in **exports** and also has **lower income rate** coupled with **higher inflation rate**.
- **Congo, Dem. Rep.** has the **lowest income** amongst the countries mentioned. It suffers due to **lower imports** as well as **exports** and a **higher child mortality rate**.

Thank You