Clustering & PCA Assignment

Contributors:
- Aditya Chopra

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 project that included a lot of awareness drives and funding programmes, 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.

Objectives

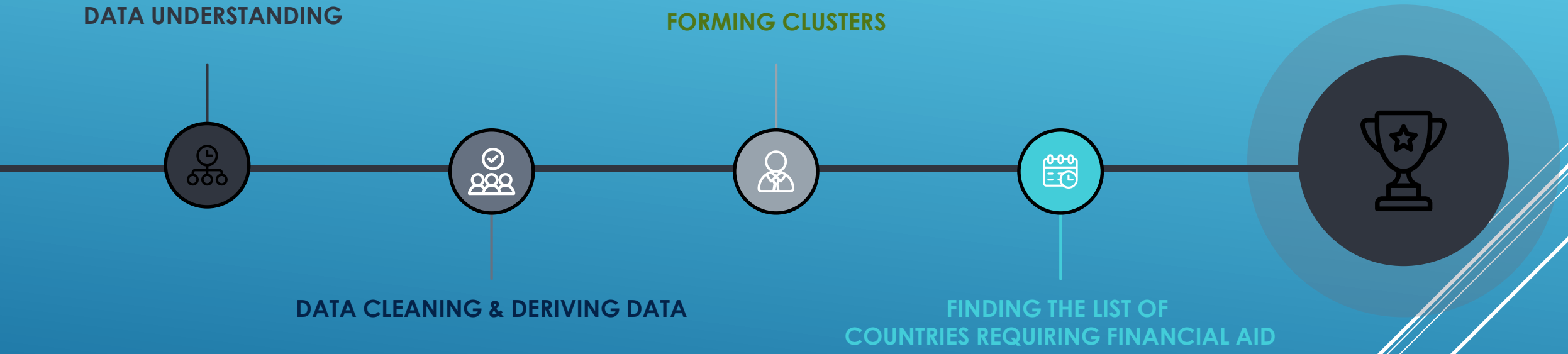
- Main task is to cluster the countries by the factors mentioned.
- Present your solution and recommendations to the CEO using a PPT.
- Use dimensionality reduction using PCA to get the visualizations of the clusters in a 2-D form.

Introduction Contd.

- And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country.

- Then we need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

Problem Solving Methodology



DATA UNDERSTANDING

- Going through the whole dataset
- Understanding the meaning of the columns given
- Looking at the dataset both using Python as well as Excel
- Creating a list of columns that can be derived further into other columns to get useful insights

DATA CLEANING & DERIVING DATA

- Finding out the null percentages of the columns
- Removing rows with null values to get a clean dataset
- Converting the columns to the required datatype
- Creating new columns out of existing columns to get better insights

FORMING CLUSTERS

- Cluster formation using K-means
- Then dimensionality reduction using PCA
- Again performing the K-means clustering on the dataset prepared by PCA

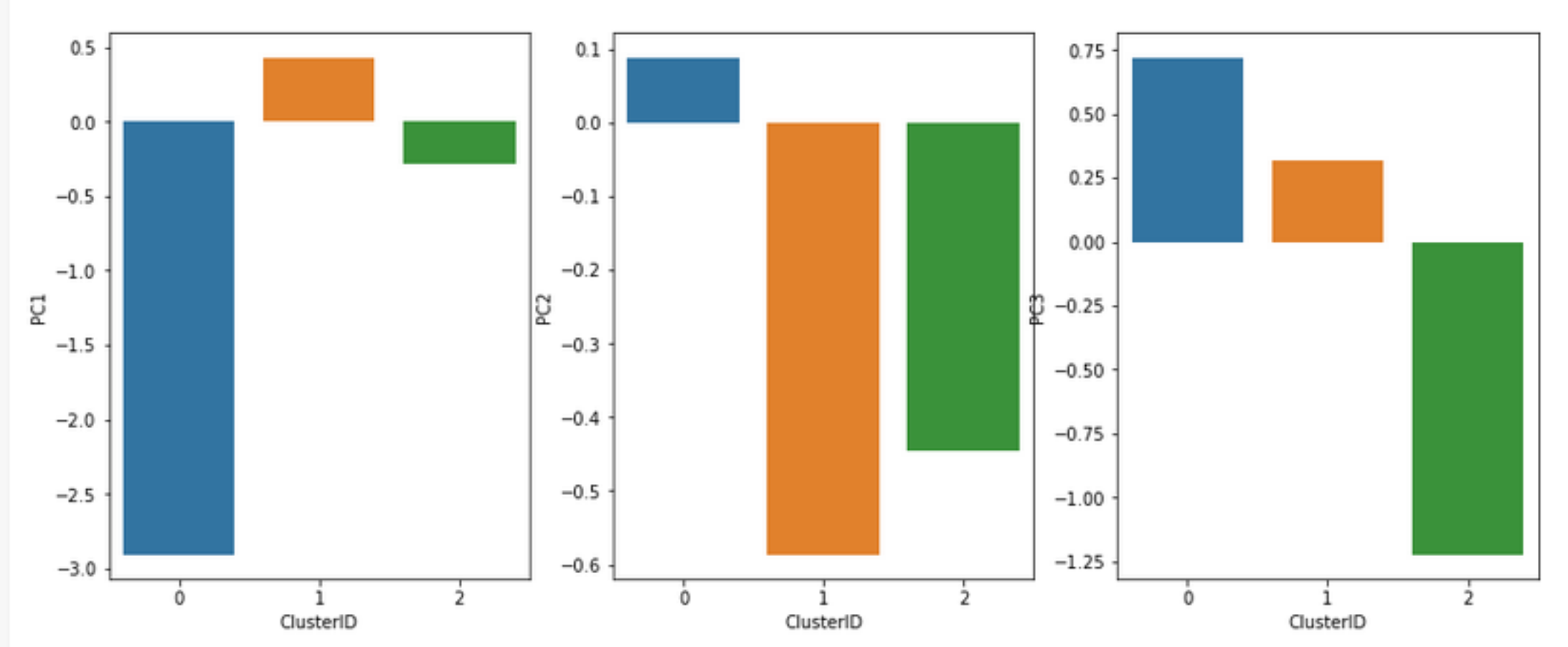
FINDING THE LIST OF COUNTRIES REQUIRING FINANCIAL AID

- Plotting bar graphs on various features to understand the cluster behaviour
- Getting the lists of countries which require financial aid
- Jotting down the insights that we get from the analysis

DISPLAYING THE RESULTS

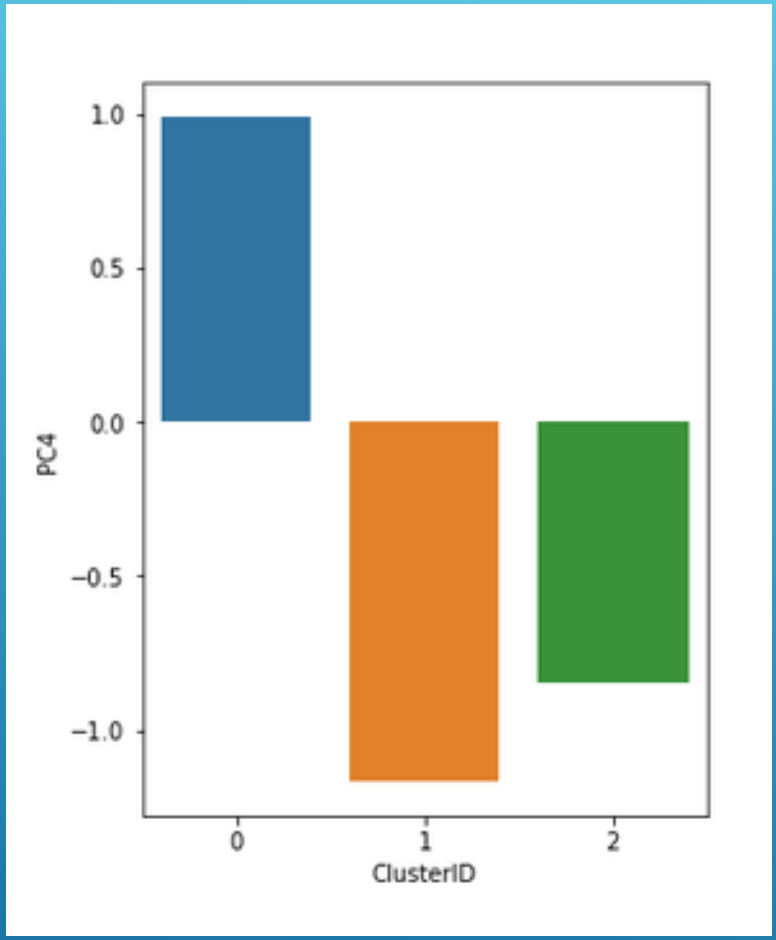
- Creating the final presentation to display the analysis
- Also, describing the factors that may be used to decide which countries the company should donate to

Principal Component Analysis Plots



The above plots are b/w the Principal Components & the Cluster IDs

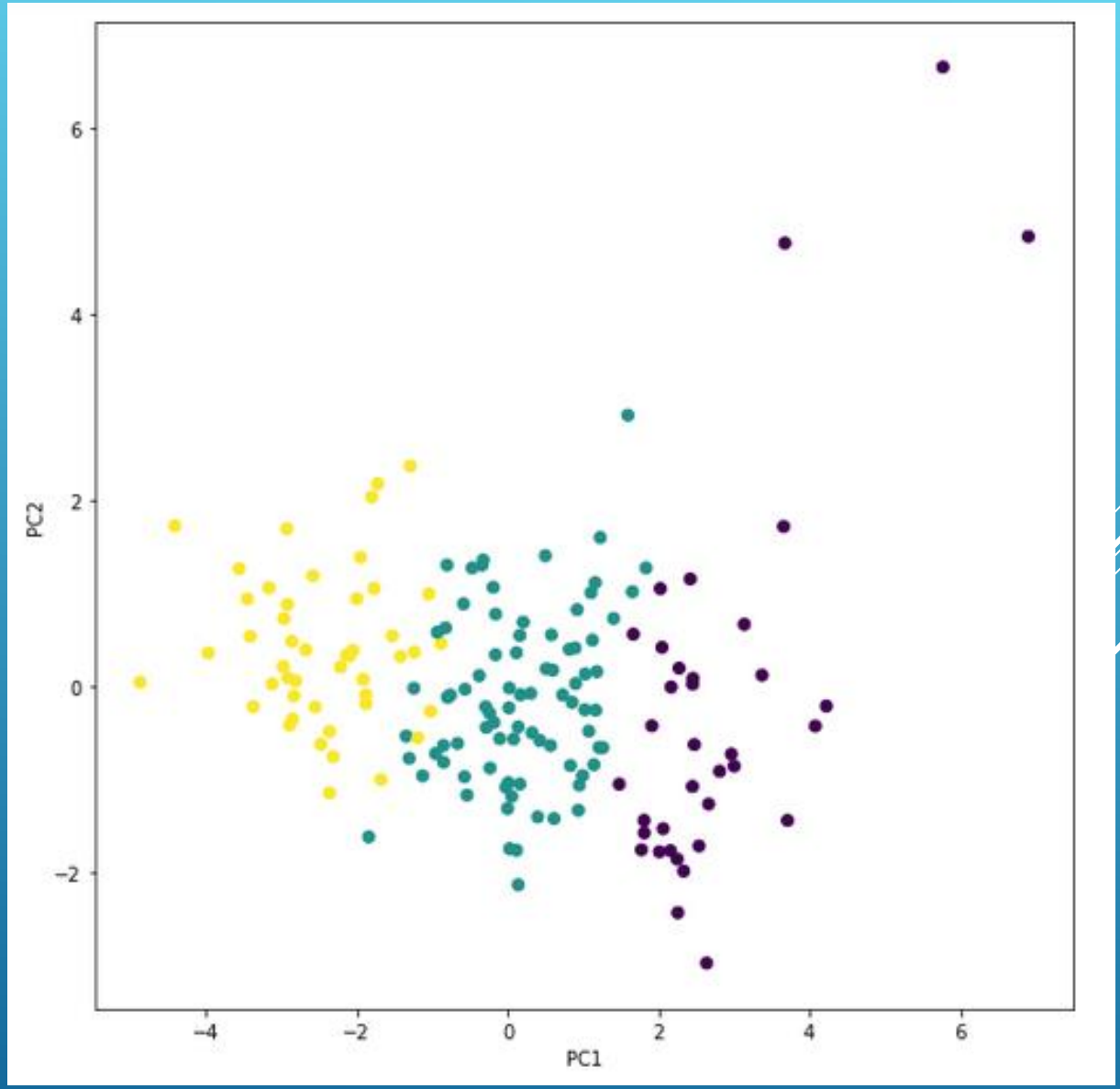
Principal Component Analysis Plots Cont.



The above plots is b/w the Principal Component 4 & the Cluster IDs

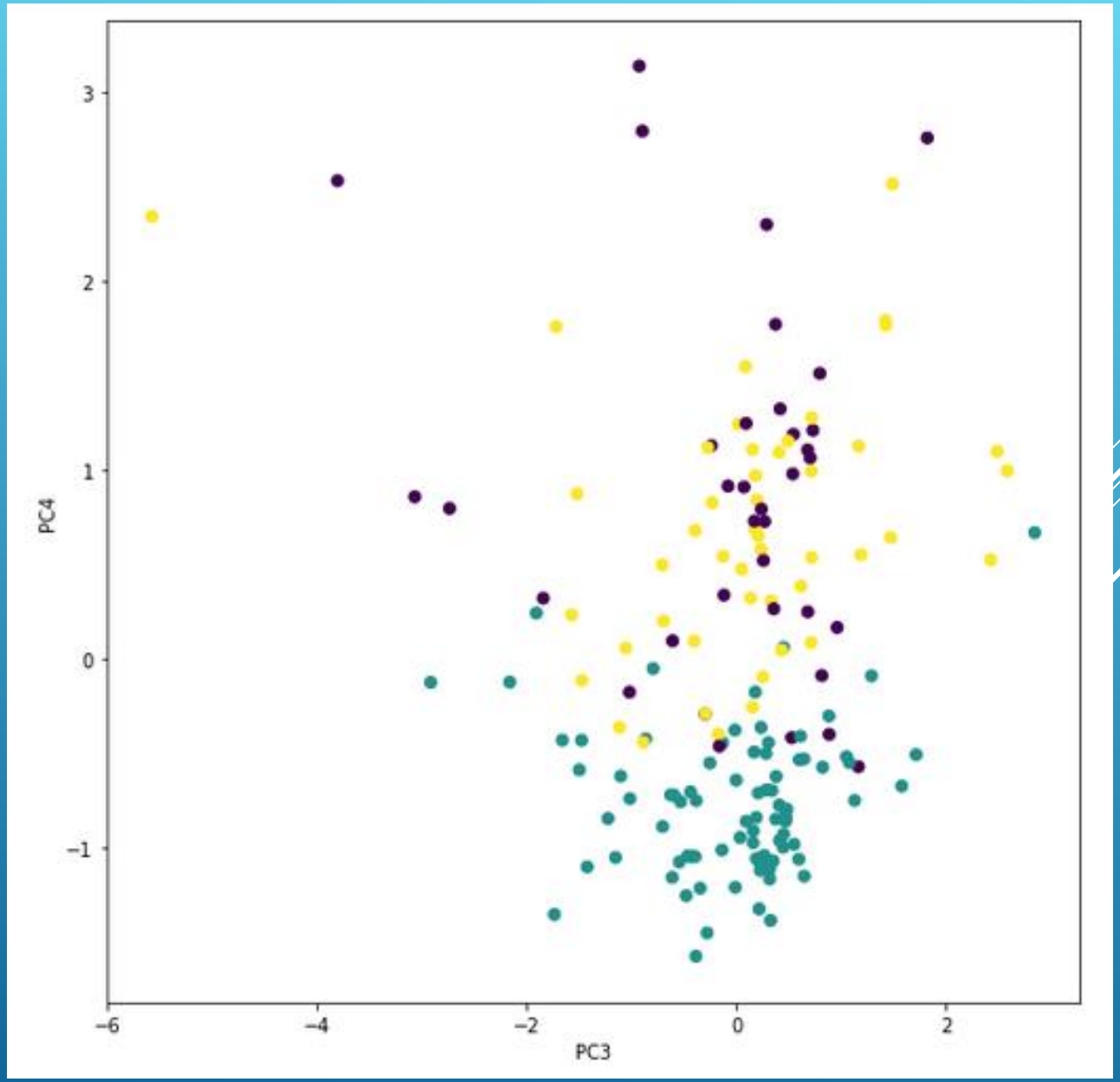
Principal Component Analysis Contd.

- This plot shows the three clusters where each cluster is being represented with a different color.
- The plot has been created after doing the Principal Component Analysis



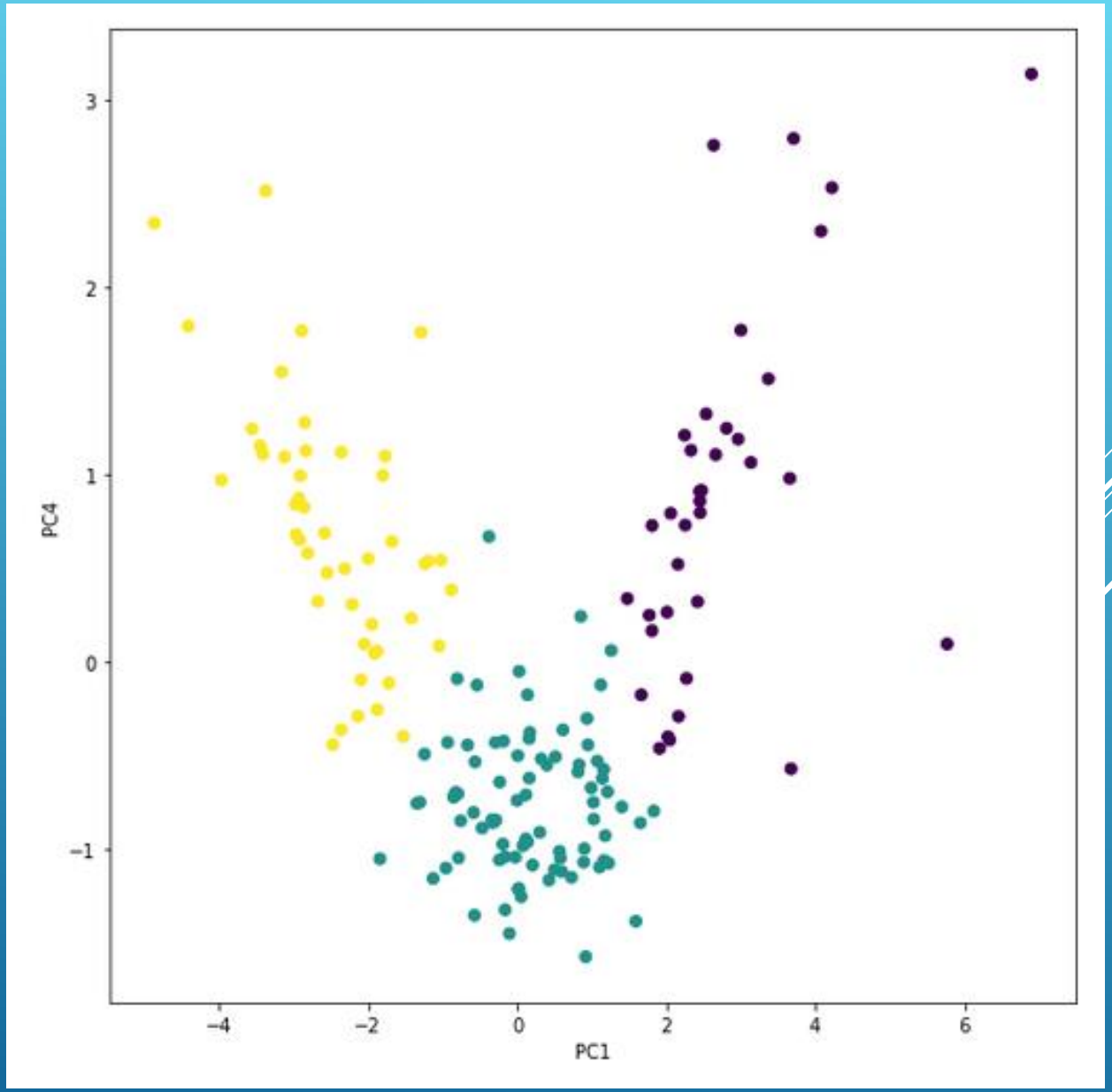
Principal Component Analysis Contd.

- This plot shows the three clusters where each cluster is being represented with a different color.
- The plot has been created after doing the Principal Component Analysis

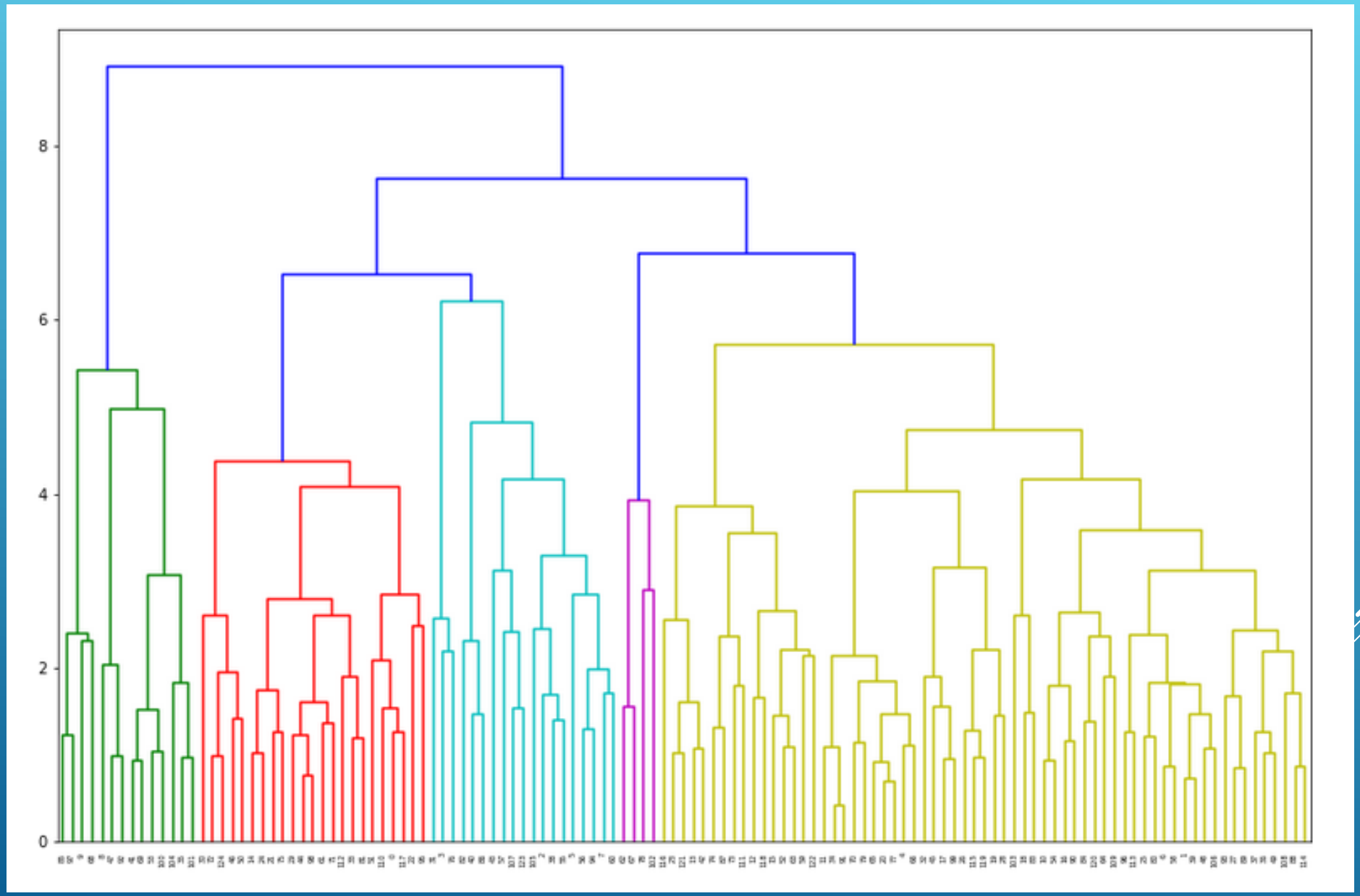


Principal Component Analysis Contd.

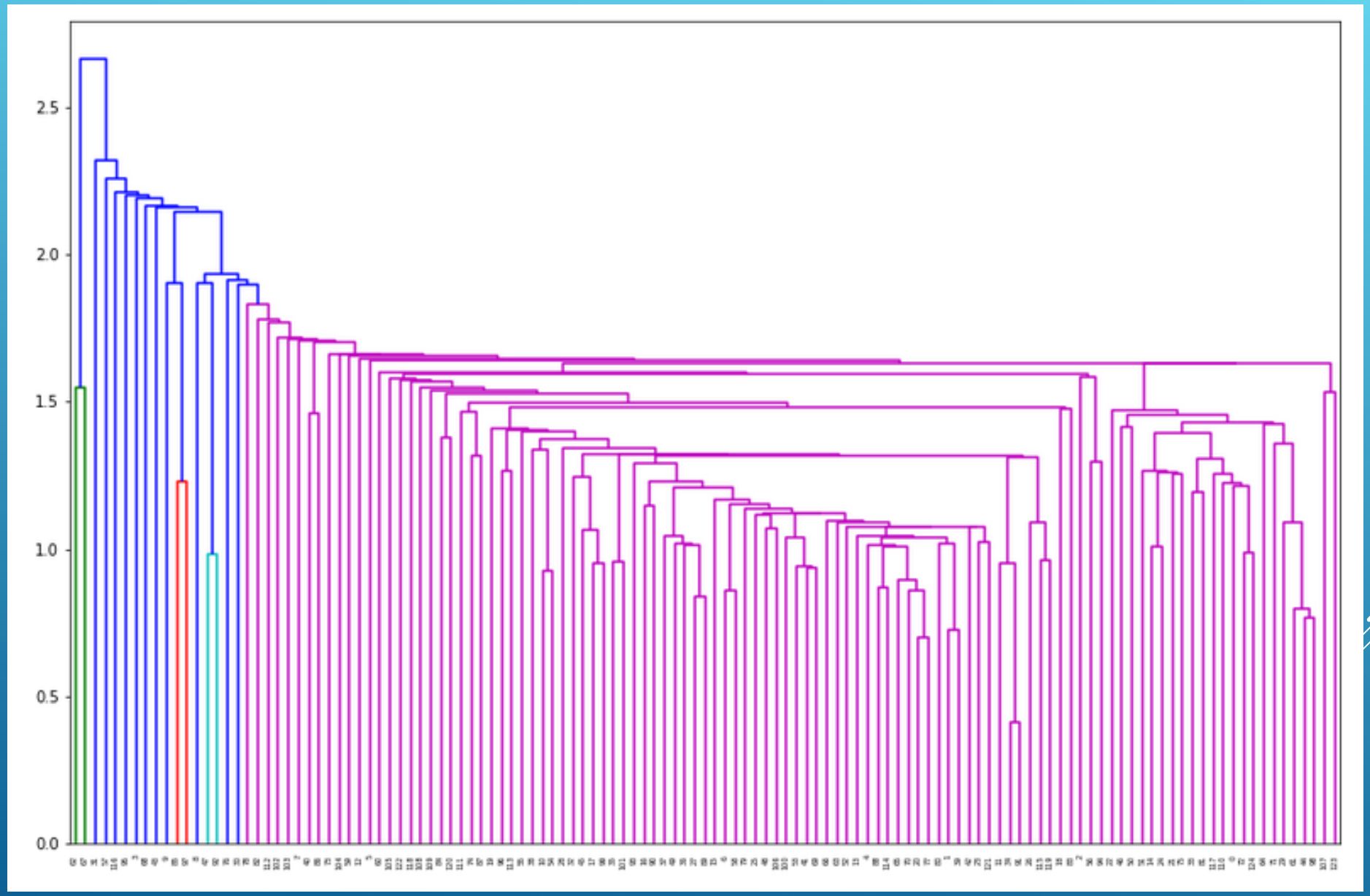
- This plot shows the three clusters where each cluster is being represented with a different color.
- The plot has been created after doing the Principal Component Analysis



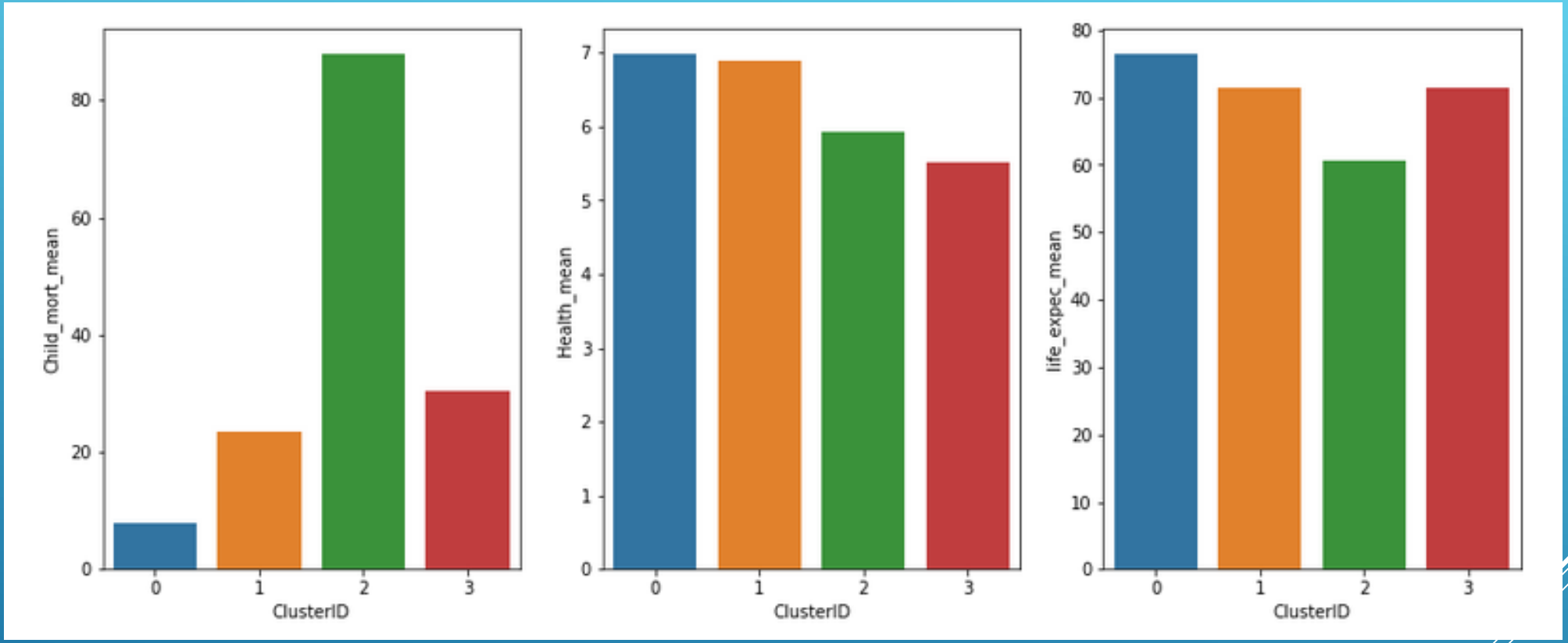
Heirarchical clustering – complete method



Heirarchical clustering – complete method

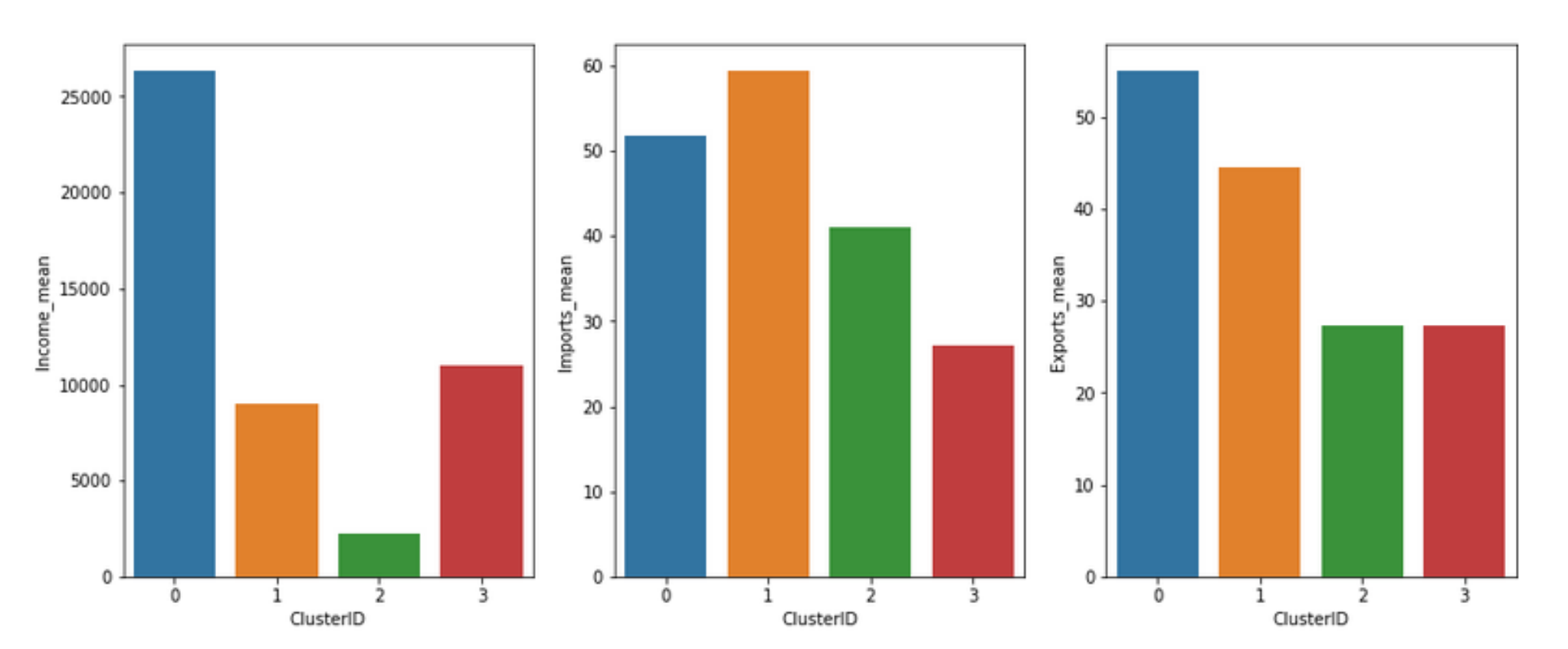


K-Means Clustering (K = 4)



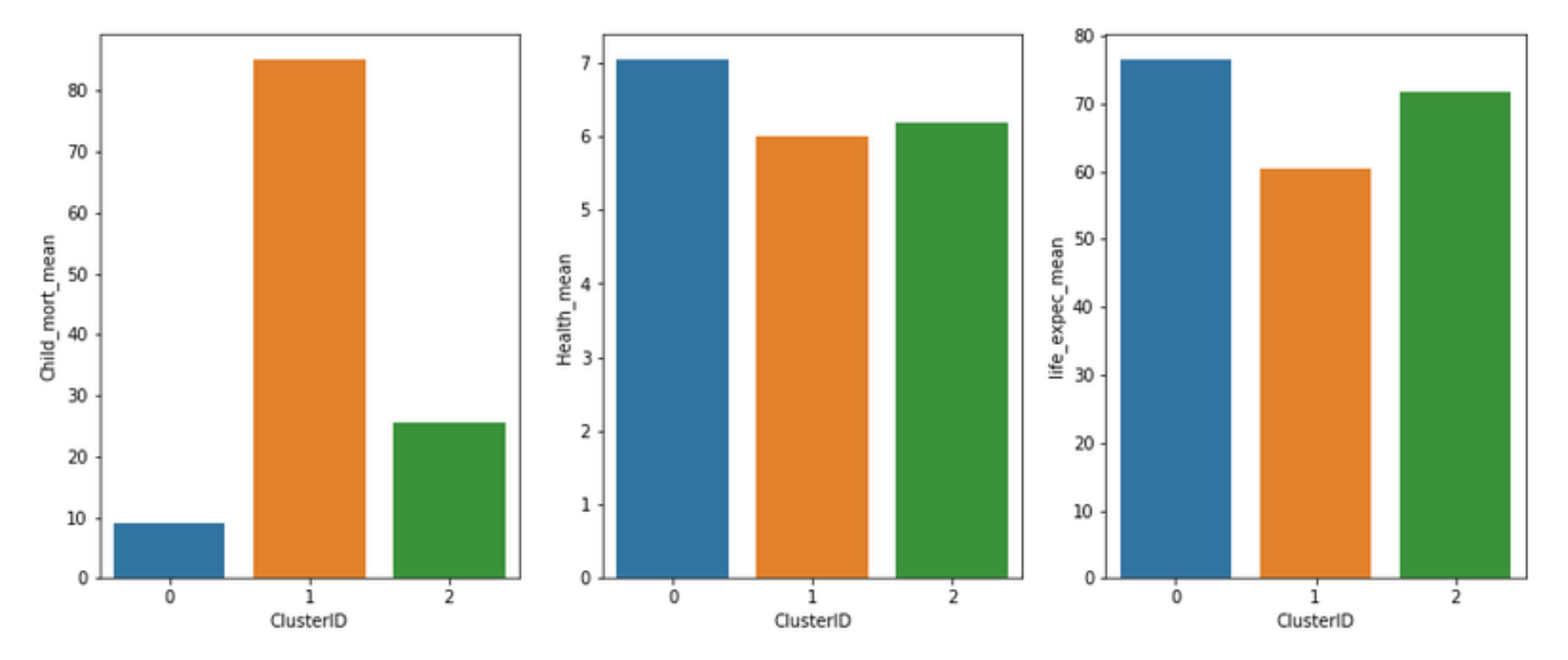
Cluster 2 has the highest child_mortality rate, lowest life_expectancy.

K-Means Clustering (K = 4) Contd.



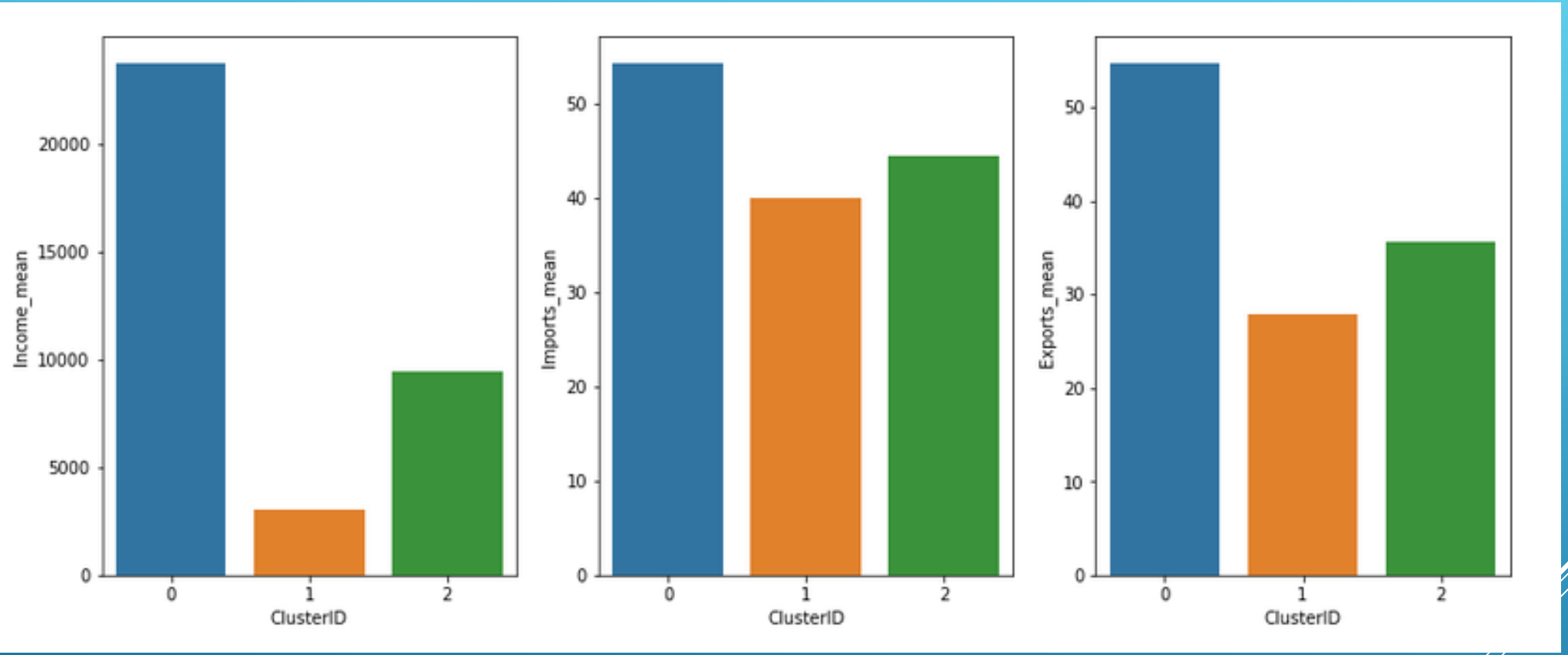
Cluster 2 has the lowest income, but imports and exports are seen at second-last position

K-Means Clustering (K = 3)



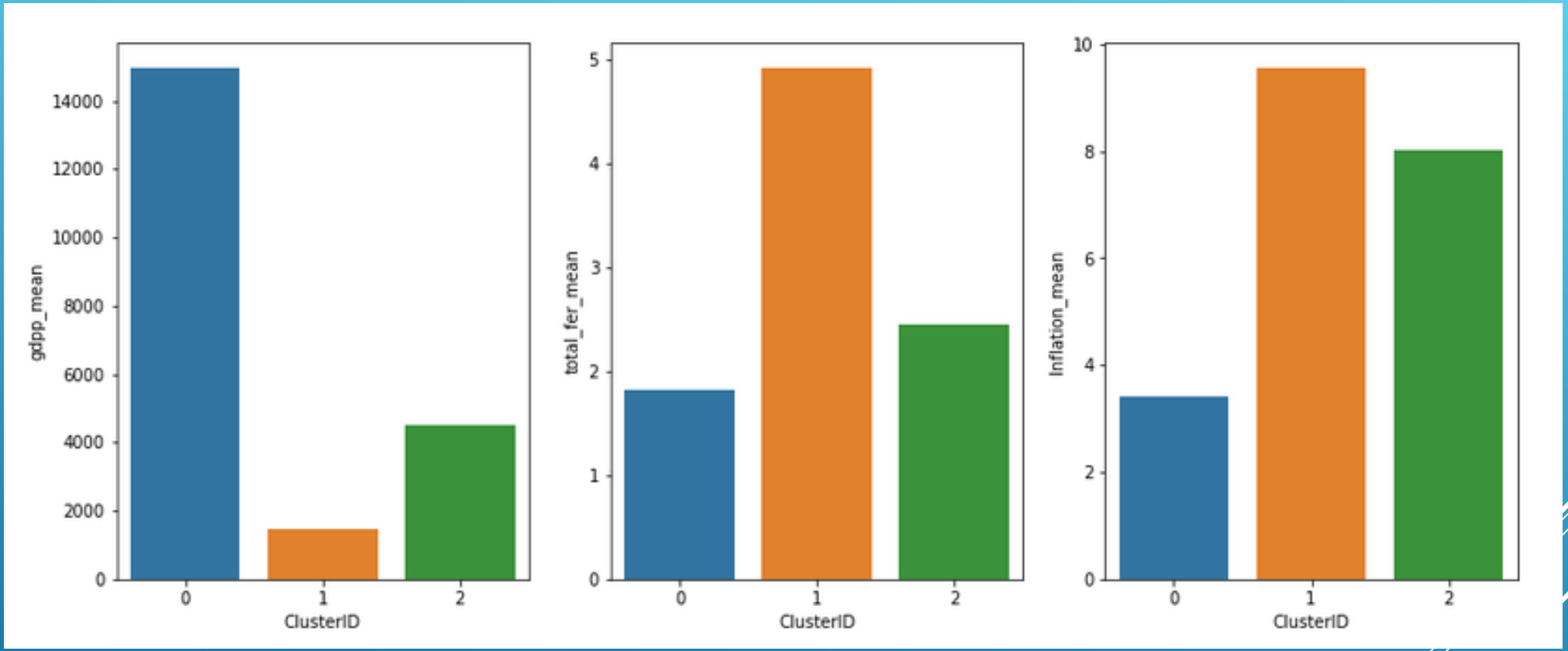
Cluster 1 has the highest child_mortality rate, lowest health and lowest life expectancy. Thus, cluster 1 should be the one to receive aid. Let's see other factors as well.

K-Means Clustering (K = 3)



Cluster 1 has the highest income rate, lowest imports and lowest exports. Thus, again cluster 1 should be the one to receive aid. Let's see other factors as well.

K-Means Clustering (K = 3)



Cluster 1 has the highest gdpp, highest total_fer and highest inflation. Thus, finally considering all the factors cluster 1 should be the one to receive aid.

LIST OF COUNTRIES UNDER CLUSTER 1:

Antigua and Barbuda
Bahamas
Bahrain
Barbados
Bulgaria
Chile
Croatia
Czech Republic
Estonia
Greece
Hungary
Latvia
Lebanon
Libya
Lithuania
Malaysia
Maldives
Mauritius
Montenegro
Oman
Panama
Poland
Portugal
Saudi Arabia
Slovak Republic
Slovenia
South Korea

CONCLUSION:

- Cluster 1 should be the one which should receive financial aid
- As seen, the countries under cluster 1 are having
  - Higheste child_mortality rate
  - Lowest health
  - Loweste life_expectancy
  - Lowest income
  - Lowest GDPP
  - Lowest in imports & exports
  - Highest inflation
- Thus, the companies should give preference to these countries so that these countries can improve in all these factors.


THANK YOU