Clustering and PCA Assignment

PGDDS – Cohort 14

Vidhu Jain

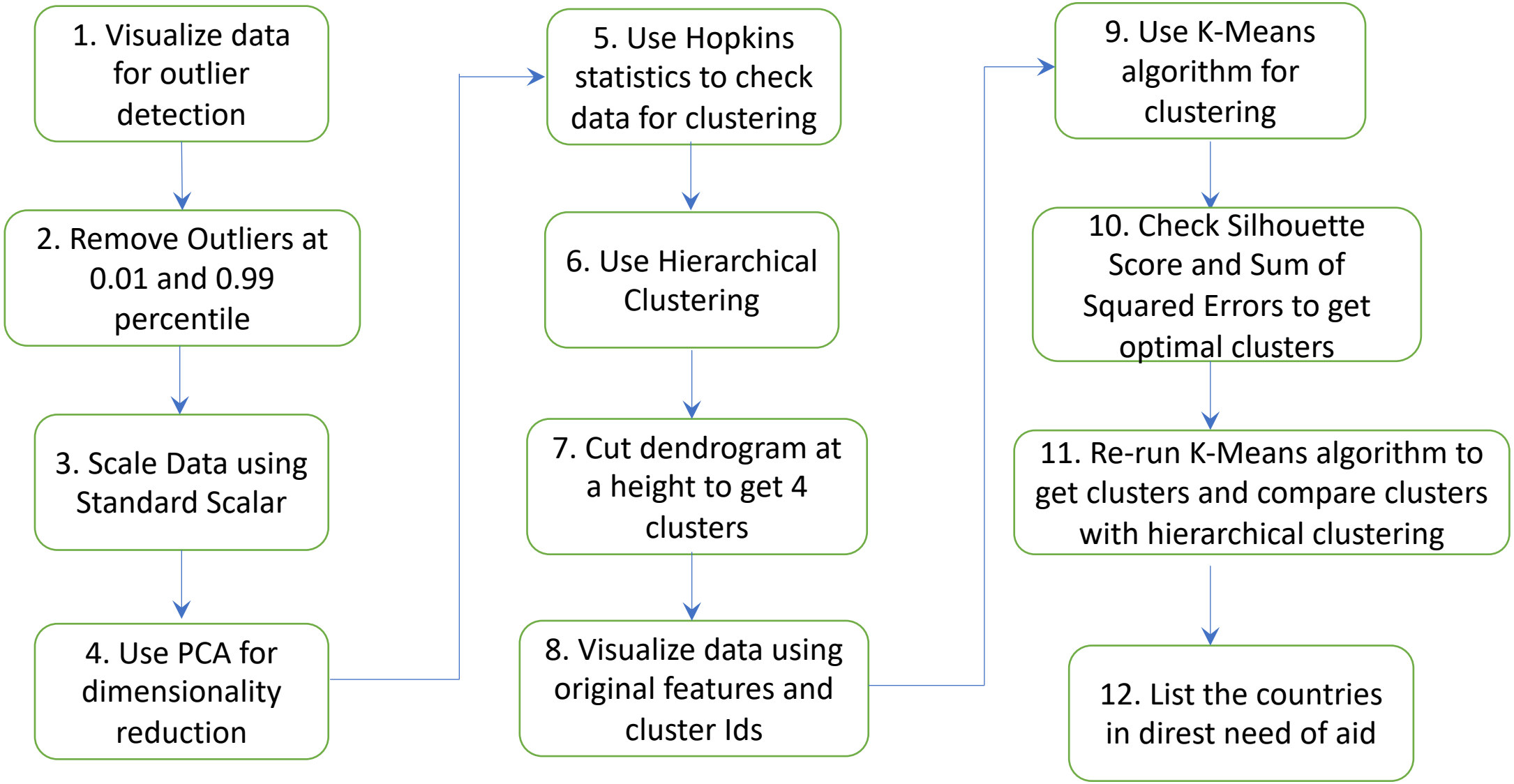
Problem Statement

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

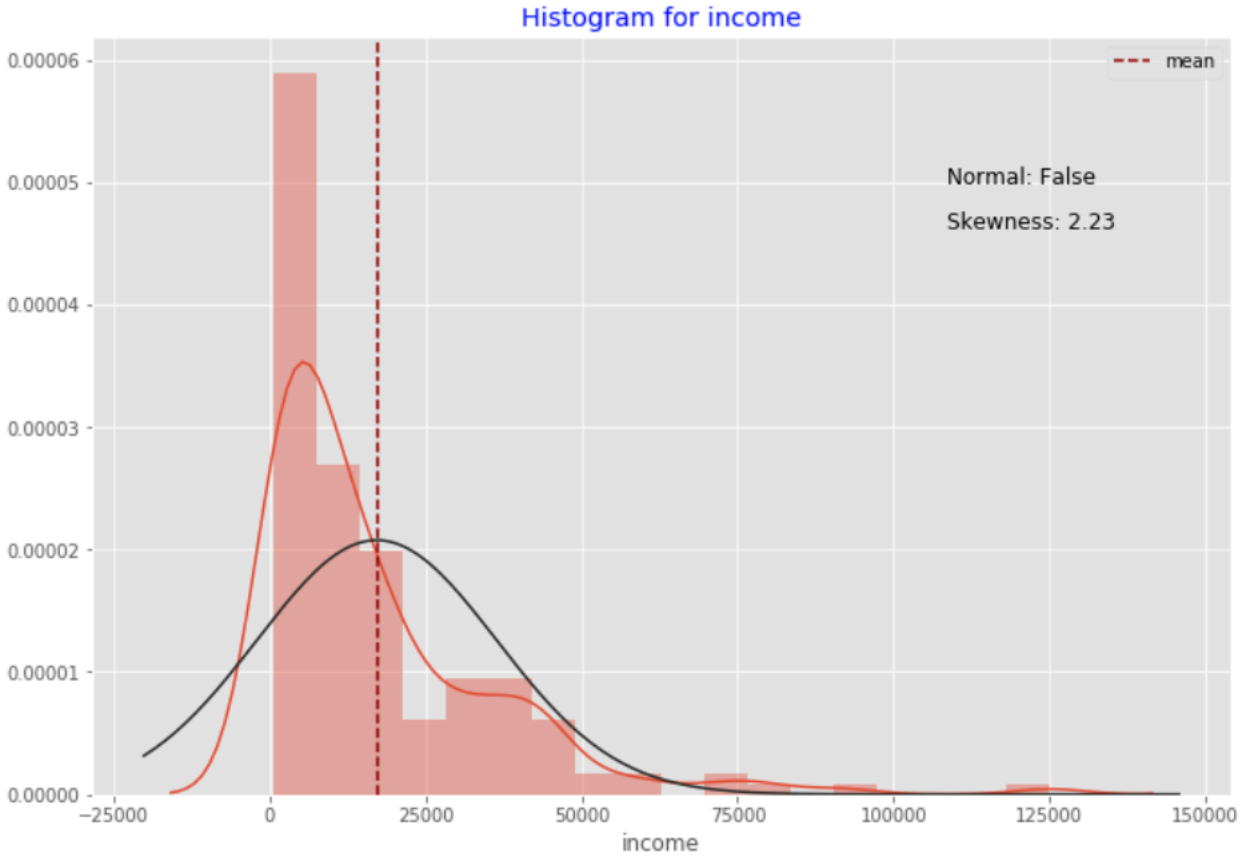
After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Categorise the countries using some socio-economic and health factors that determine the overall development of the country. Suggest the countries which the CEO needs to focus on the most

Approach to categorize countries



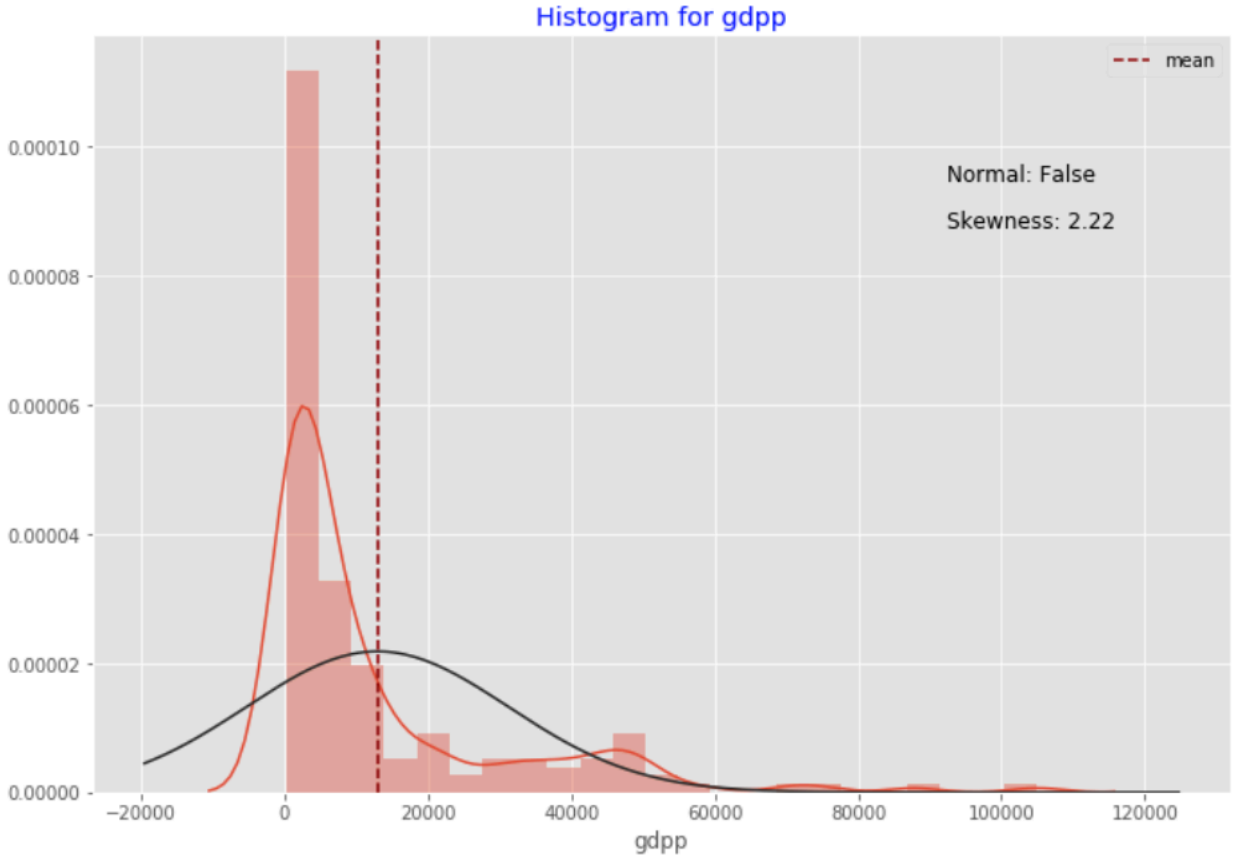
Analyze the given distribution



Net Income per person distribution

- The data is right skewed with skewness of 2.23. Most of the observations have net income below average
- Maximum observations have net income within the range of 600-10000.
- There are few observations with Net income per person within the range of 20000-60000
- There are few outliers between 70000 - 125000

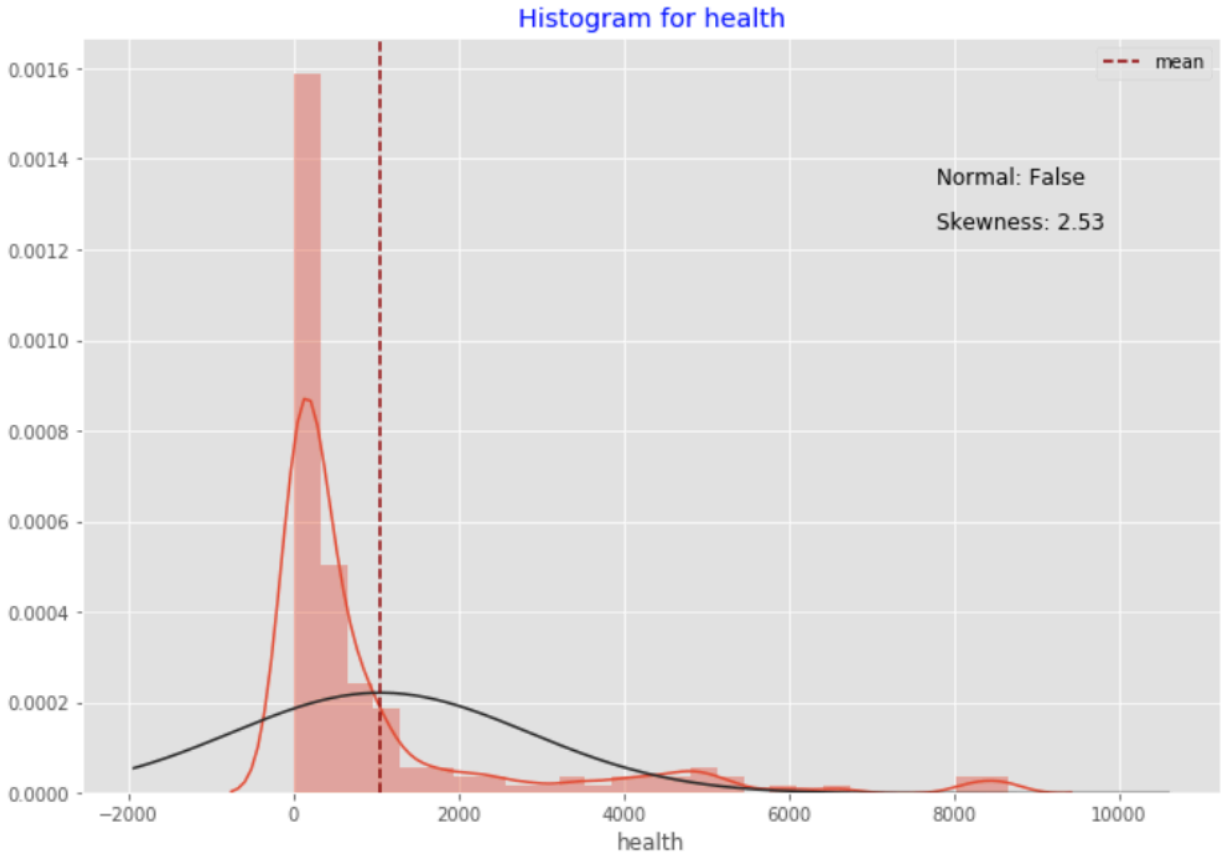
Analyze the given distribution



GDP per capita distribution

- The distribution is right skewed with maximum observations within the range of 0-5000 which is below average.
- There are couple of observations with GDP per capita within the range of 15000-60000.
- There are few outliers as well where GDP per capita is within the range of 79000-105000

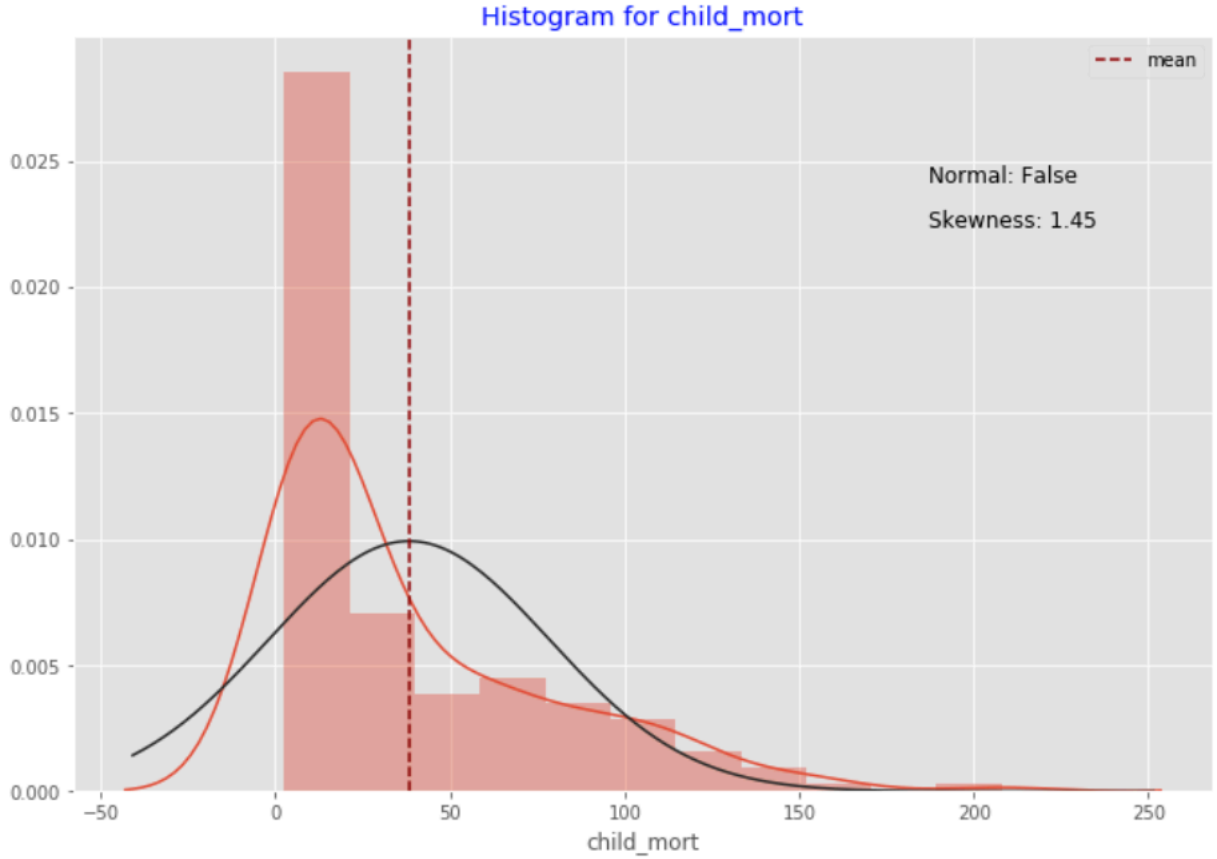
Analyze the given distribution



Health Spending distribution

- This distribution is right skewed meaning there are many observations that have %age health rate spending below average and outliers having health rate spending more than average
- There are many observations with health rate spending within the range of 12-320.
- There are couple of observations within the range of 320-1000 of health rate spending.
- Some observations are within the range of 1000-4000 with few outlier in the range of 8000-9000

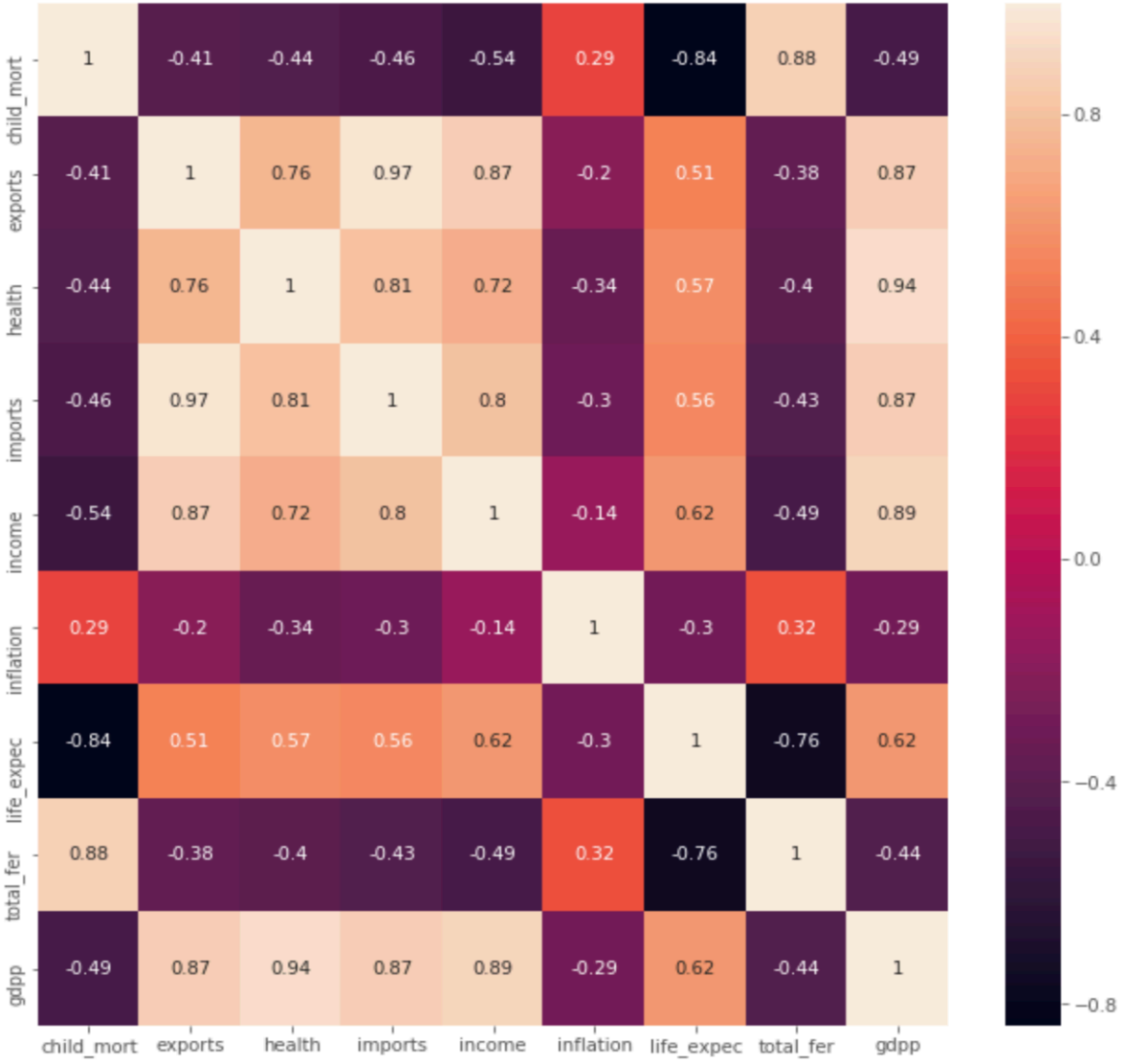
Analyze the given distribution



Child mortality distribution

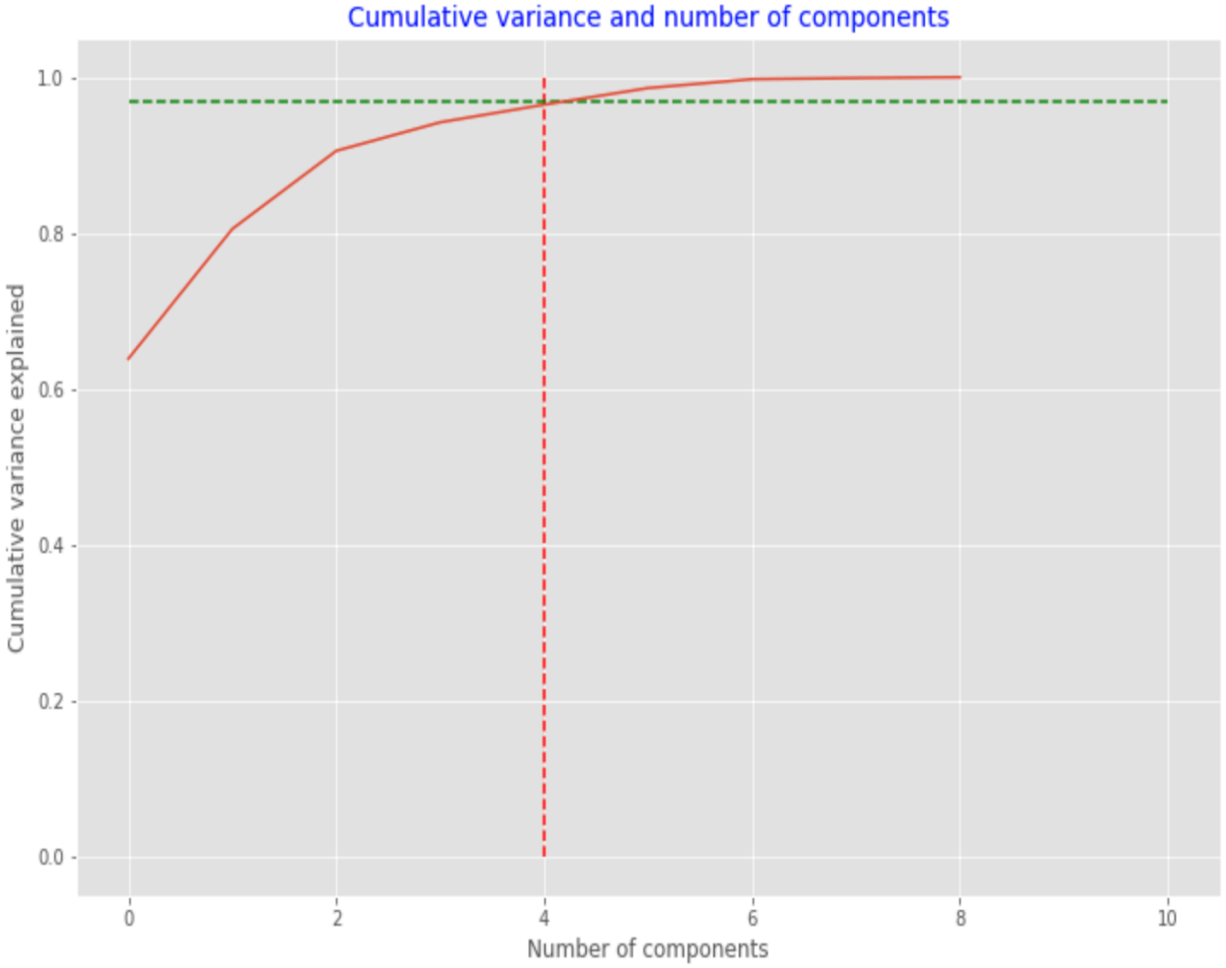
- The data is slightly right-skewed.
- Most of the observations have child mortality rate within the range of 0-20
- There are many observations within the range of 30-150

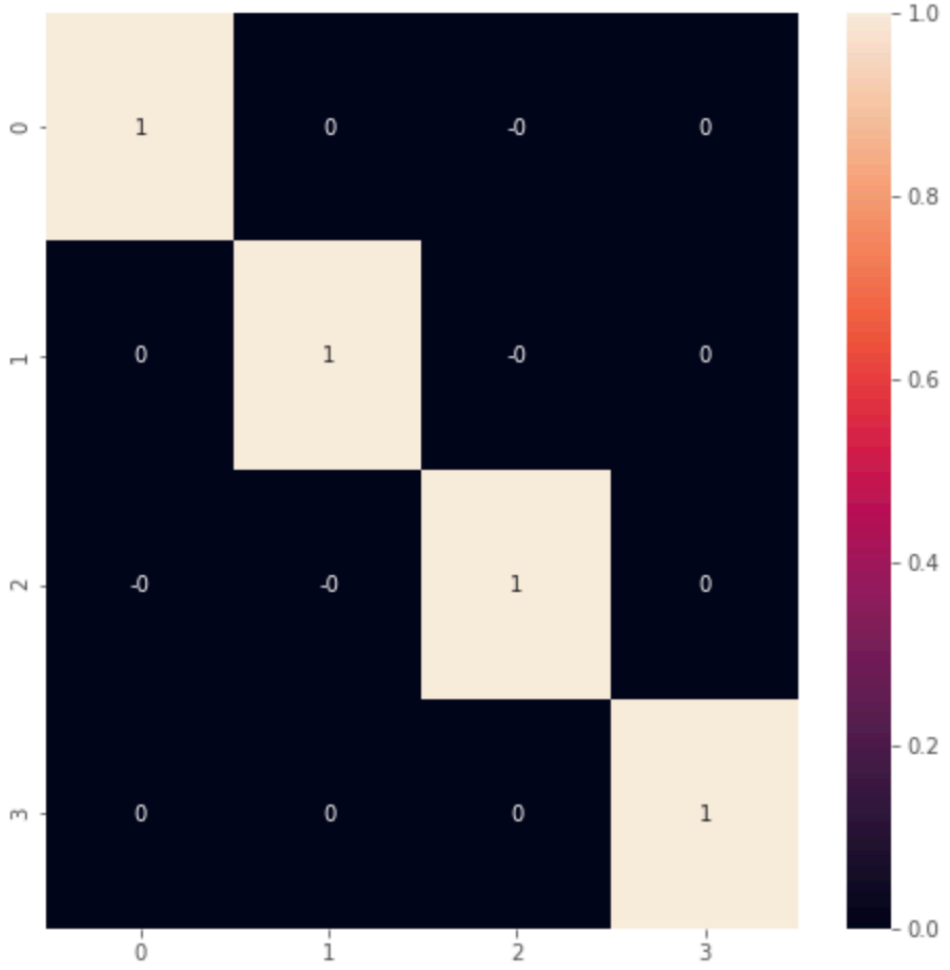
Heatmap to show relationship/correlation between features



- The data present is in different scale for all features so data has been scaled using StandardScalar
- Correlation visualized after scaling data
- There is a very high negative correlation between child_mort & income and child_mort & life_expec. Also there is strong correlation between child_mort & total_fer
- Features exports and imports are highly positively correlated. There is strong positive correlation between exports and income as well
- Income is negatively correlated with child_mort and total_fer. This means as income decreases child mortality goes up. Also there is positive correlation between income and life_expec, this means as the income increases the life expectancy increases. Also as the income increases the gdpp increases because of the positive correlation between the two
- Life_expec has strong negative correlation with child_mort and total_fer and high positive correlation with income and gdpp

Apply PCA and choose optimal number of principal components

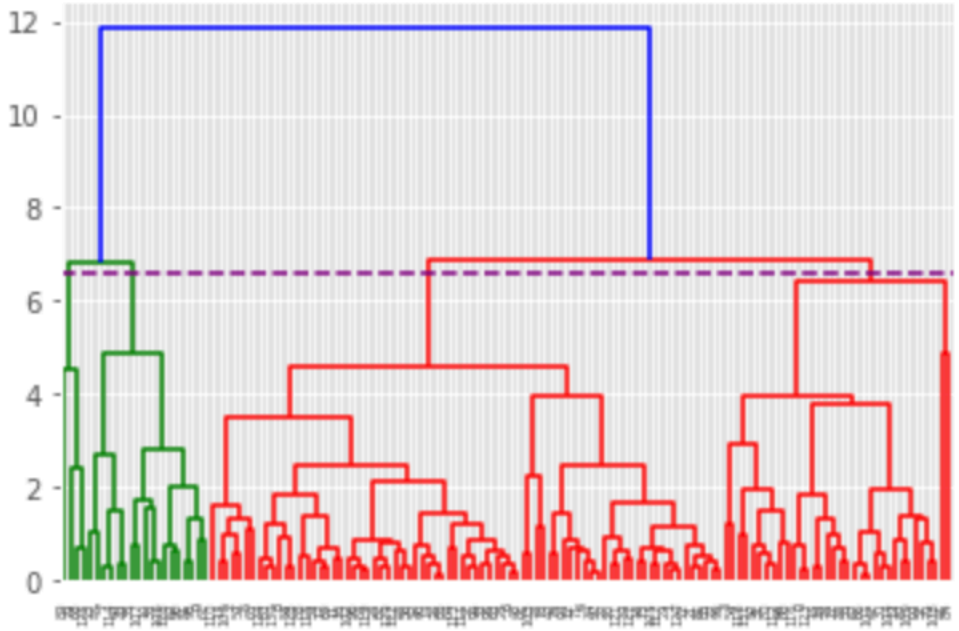




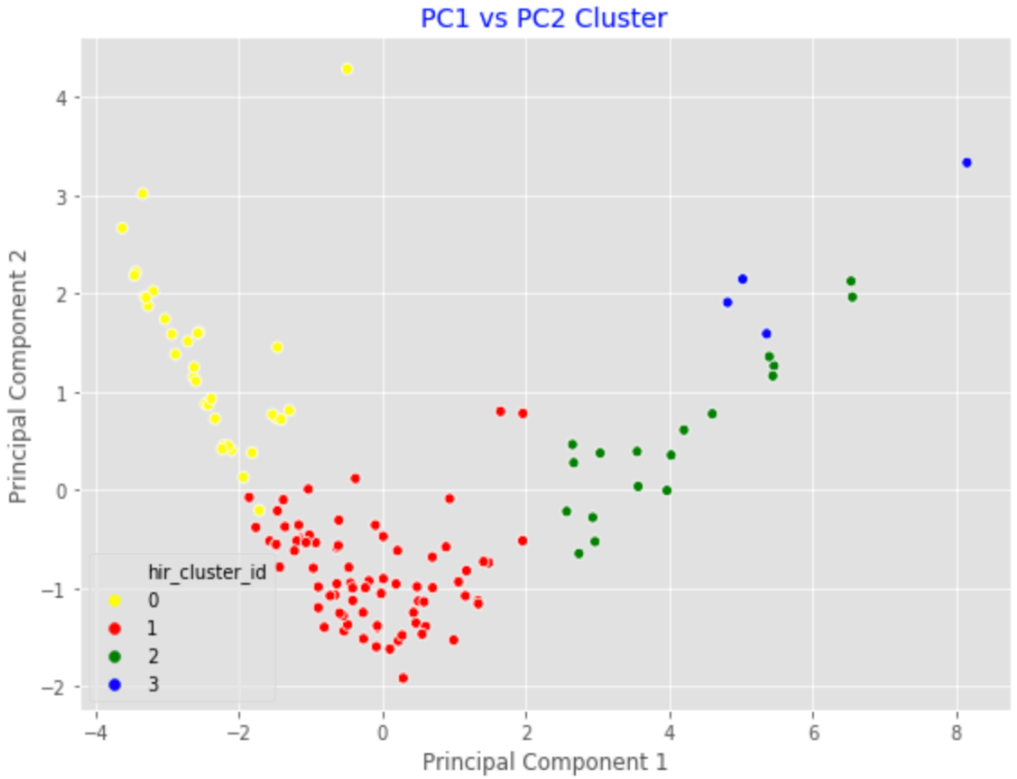
There are 4 Principal components that together explain 97% of variance in data

There is no correlation between these principal components

Apply Hierarchical Clustering to choose optimal number of clusters

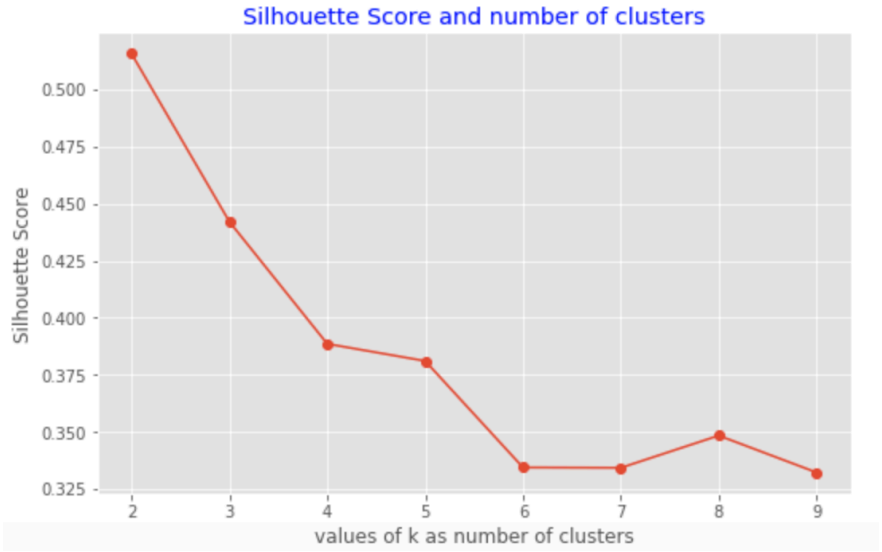


Cut dendrogram at height of 6.6 to choose 4 clusters

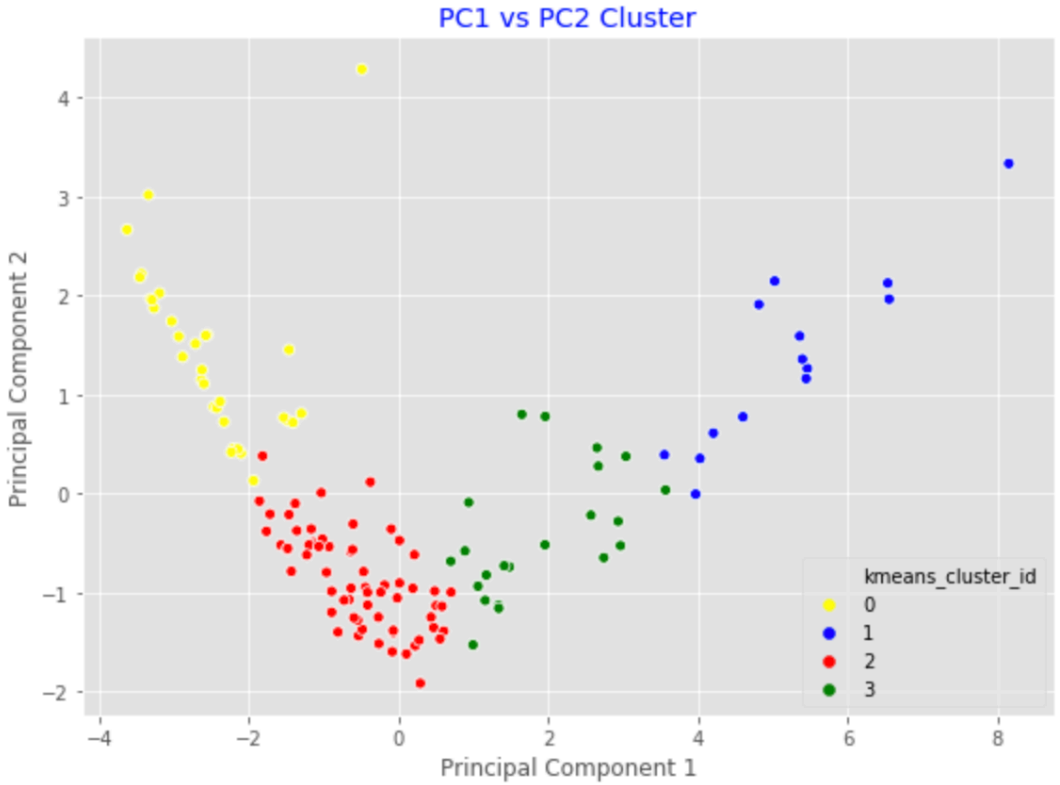


Use K-Means algorithm with various number of clusters as input and check for

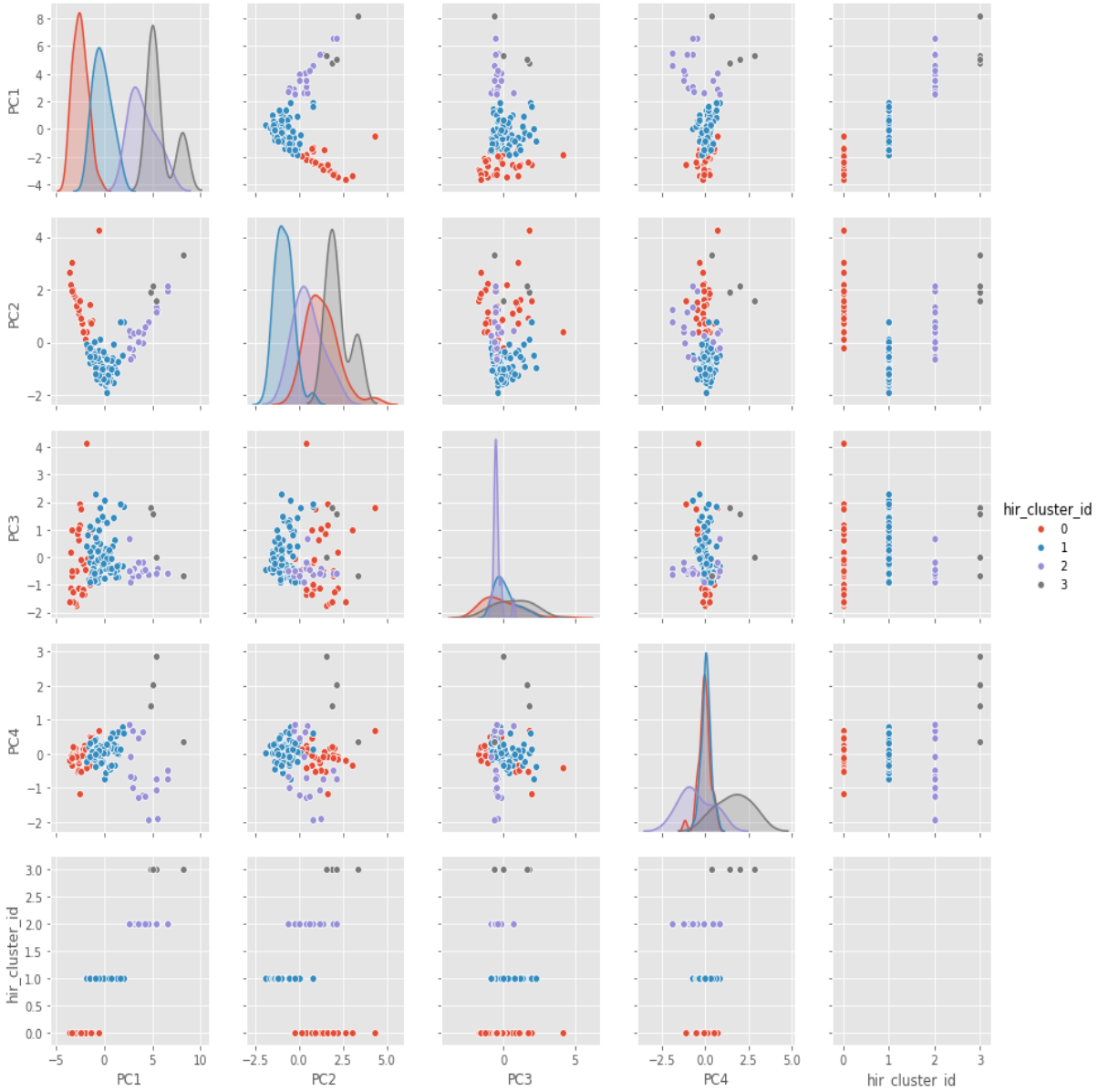
- Sum of Squared Errors
- Silhouette Score



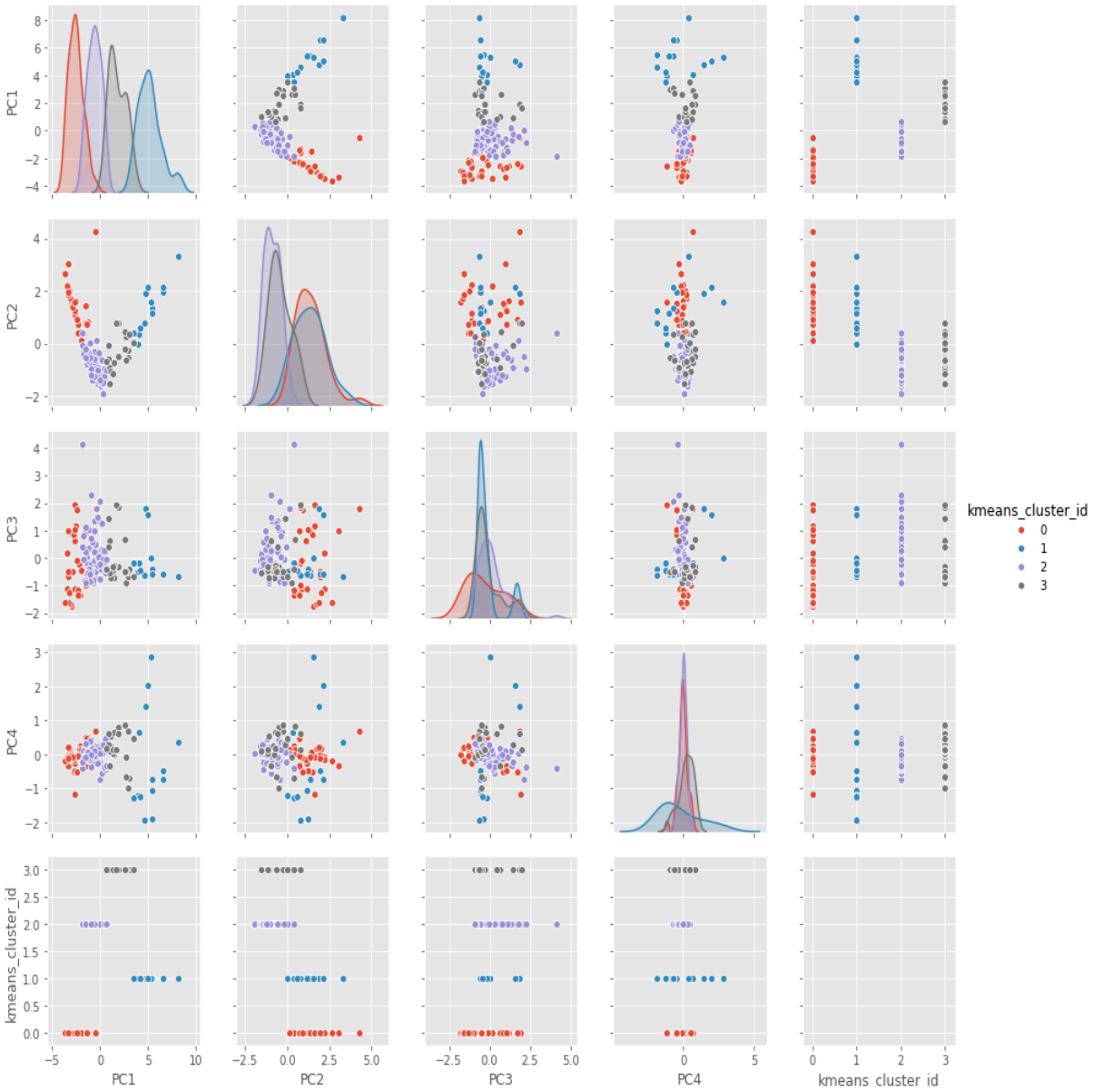
Choose 4 as optimal number of clusters and re-run K-Means algorithm



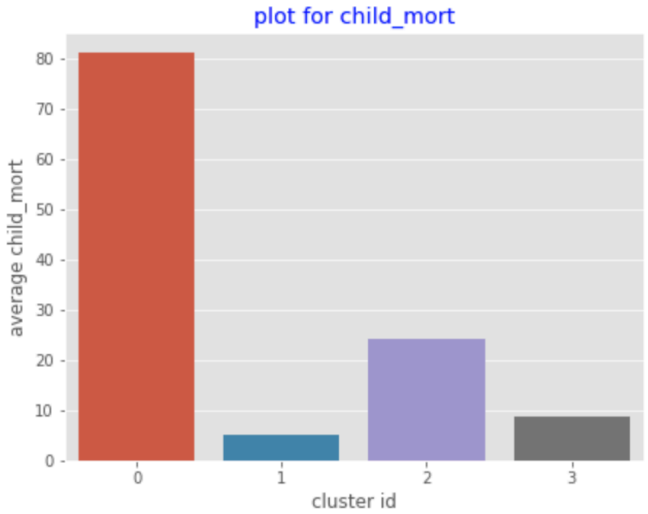
Visualize clusters of PCA using Hierarchical Clustering

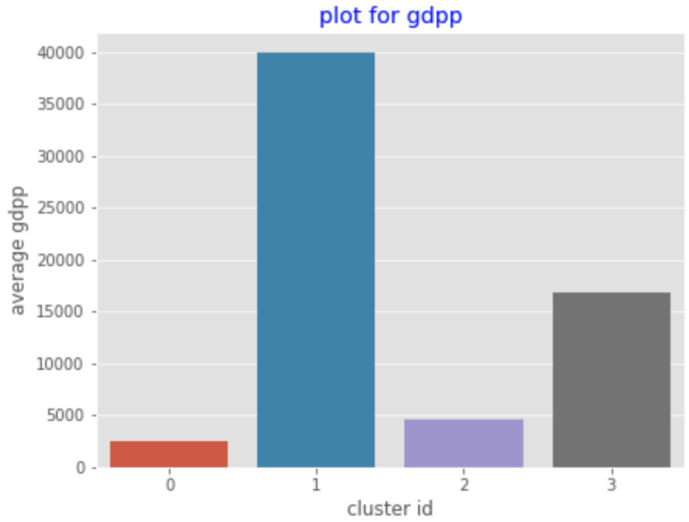


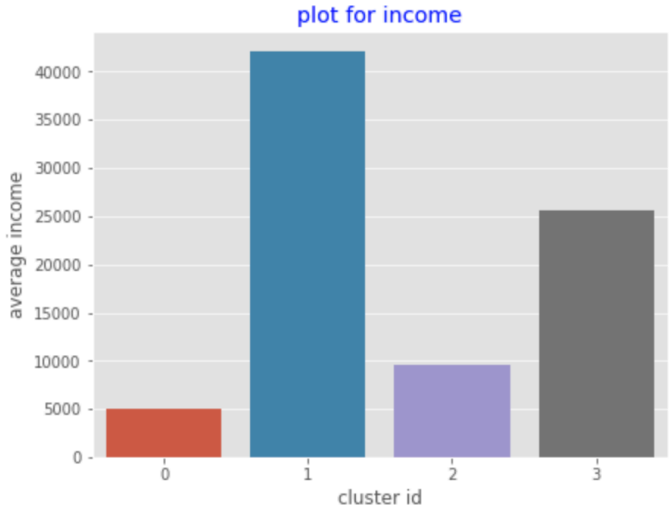
Visualize clusters of PCA using K-Means Clustering



- Countries with cluster label 1 are the developed countries
- Those with label 0 are under developed countries
- Countries with label 2 and 3 are developing countries



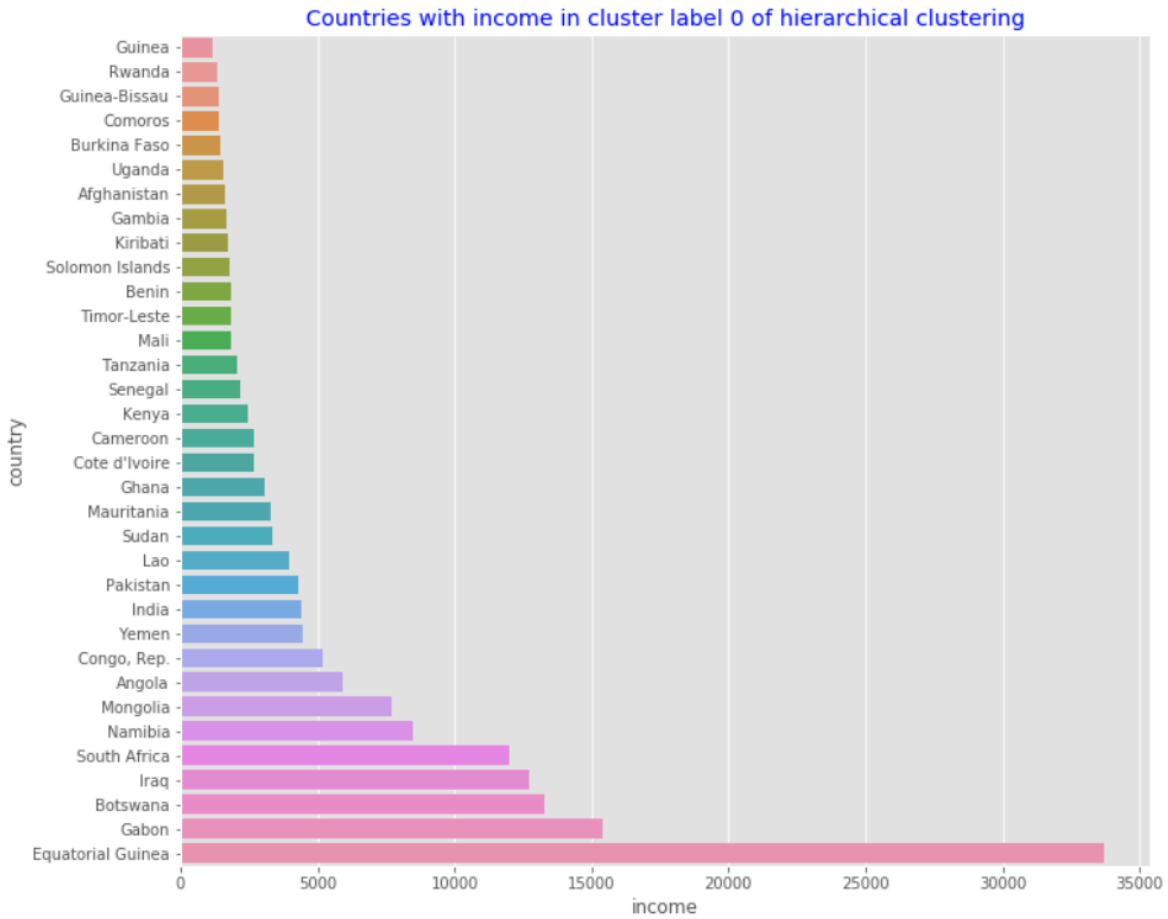


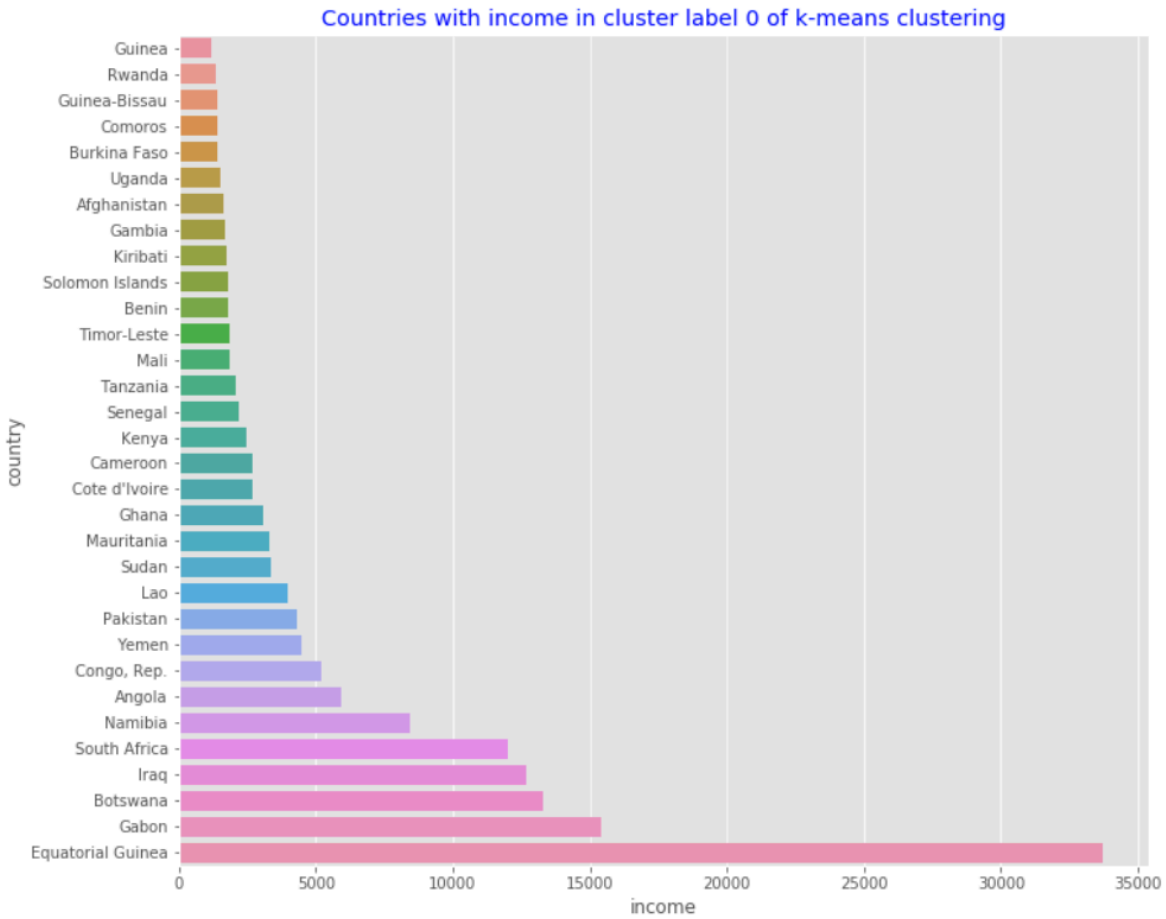


Inferences

- Average GDP per capita is highest for countries with cluster label 1 (39935.71) followed by labelled 3 countries (16832.27), labelled 2 countries (4615.22) and labelled 0 countries (2424.28)
- Average income is highest for countries with cluster label 1 (42078.57) followed by labelled 3 countries (25572.73), labelled 2 countries (9612.03) and labelled 0 countries (5058.44)
- Average inflation is highest for countries with cluster label 0 (10.50) followed by labelled 2 countries (7.93), labelled 3 countries (3.72), labelled 1 countries (3.06)
- Average child mortality rate is highest for countries with cluster label 0 (81.14) followed by labelled 2 countries (24.27), labelled 3 countries (8.79) , labelled 1 countries (5.29)
- Countries with cluster label 0 (under developed countries) need more focus
- Cluster Label 0 countries on average with lowest GDP per capita, lowest income, lowest life expectancy, highest child mortality rate need more focus for improvement

Hierarchical and K-Means Clustering yielded similar list of countries that need focus





Recommendation - List of countries that are in direst need of aid

Afghanistan, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote dIvoire, Equatorial Guinea, Eritrea, Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Kenya, Kiribati, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Pakistan, Rwanda, Senegal, Sierra Leone, South Africa, Sudan, Tajikistan', 'Tanzania', Timor-Leste, Togo, Uganda, Yemen, Zambia