Foundation to Data Science

SARFARAZ JAMAL-17093

ASSIGNMENT 2

Problem Statement

Identify the countries that are in most need of aid. Categorize countries according socio-economic factors and determine the overall development level of the country. Then recommend countries that need the most attention.

Data Understanding

- The data provided consists a mixture of **socio-economic** and **health** variables.
- It includes data from **167** countries.

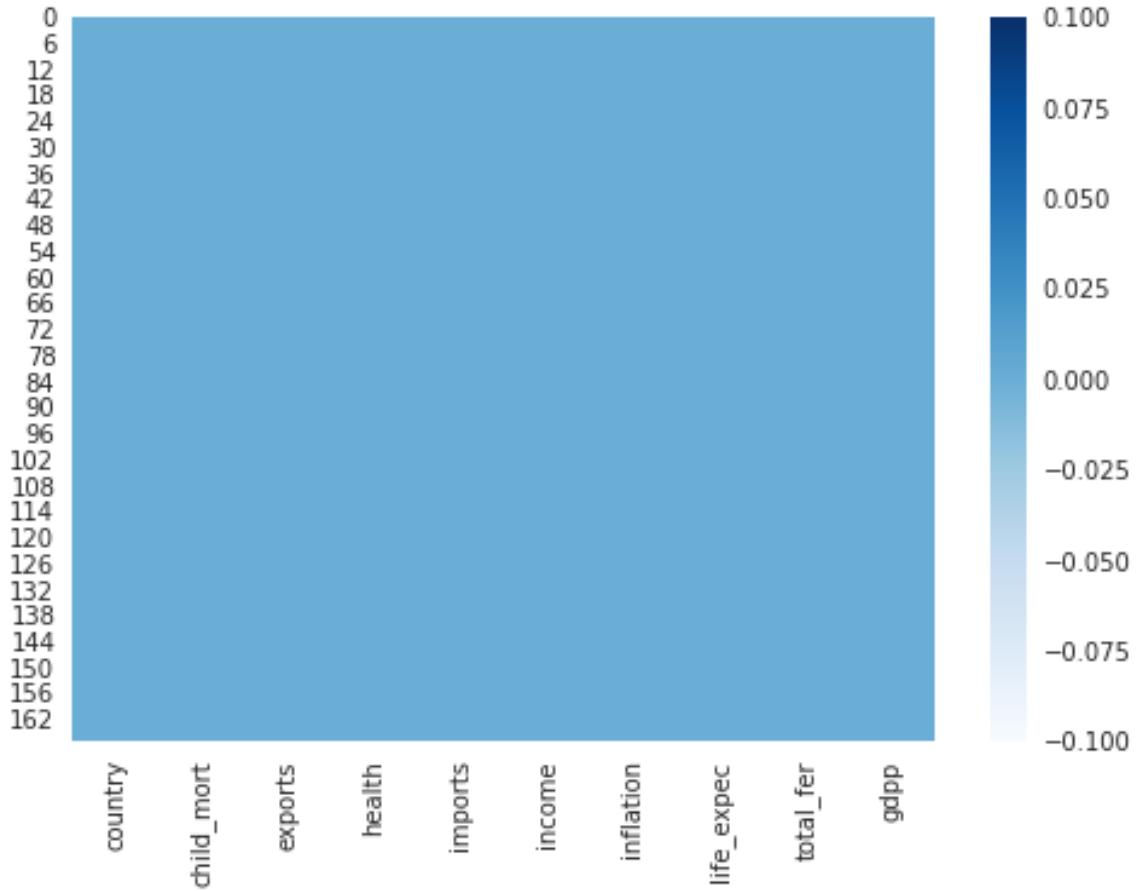
Data Understanding

Data Dictionary

	Column Name	Description
0	country	Name of the country
1	child_mort	Death of children under 5 years of age per 1000 live births
2	exports	Exports of goods and services per capita. Given as %age of the GDP per capita
3	health	Total health spending per capita. Given as %age of GDP per capita
4	imports	Imports of goods and services per capita. Given as %age of the GDP per capita
5	Income	Net income per person
6	Inflation	The measurement of the annual growth rate of the Total GDP
7	life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
8	total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
9	gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

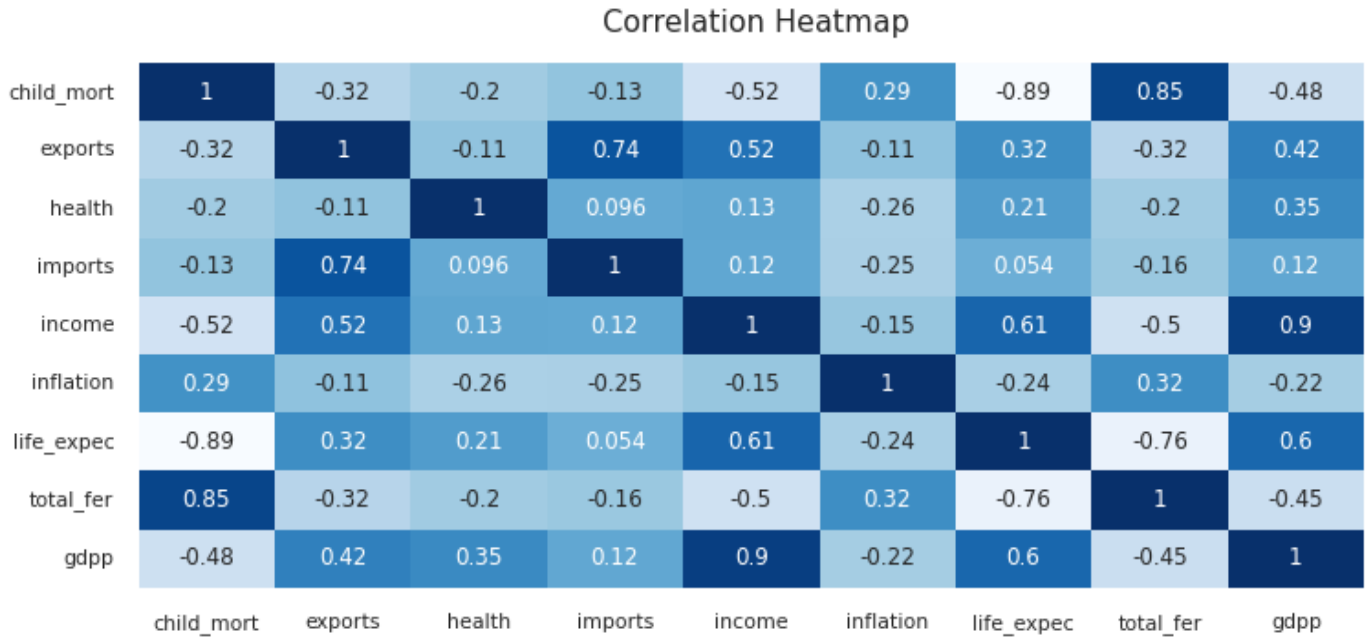
Data Understanding
Missing Values

- No Missing Values were found in the data

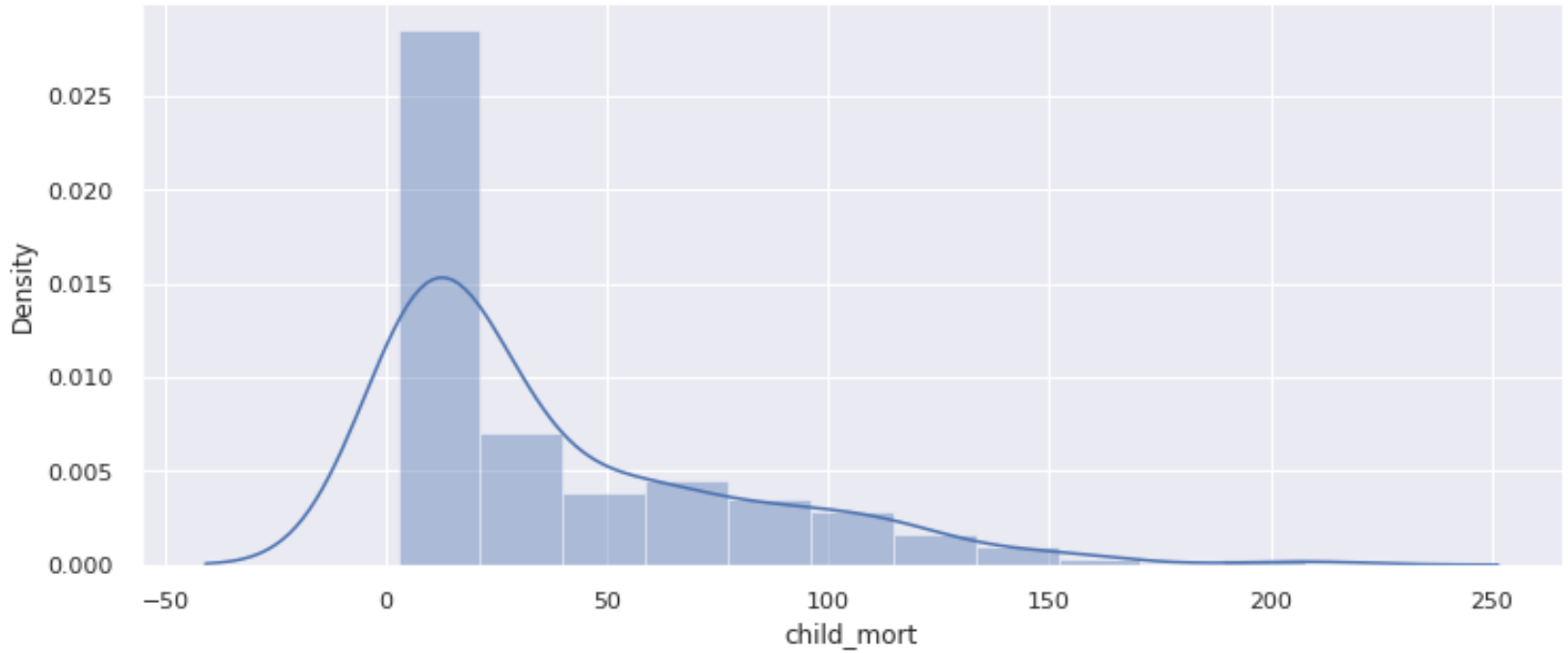


Data Understanding
Correlation Matrix

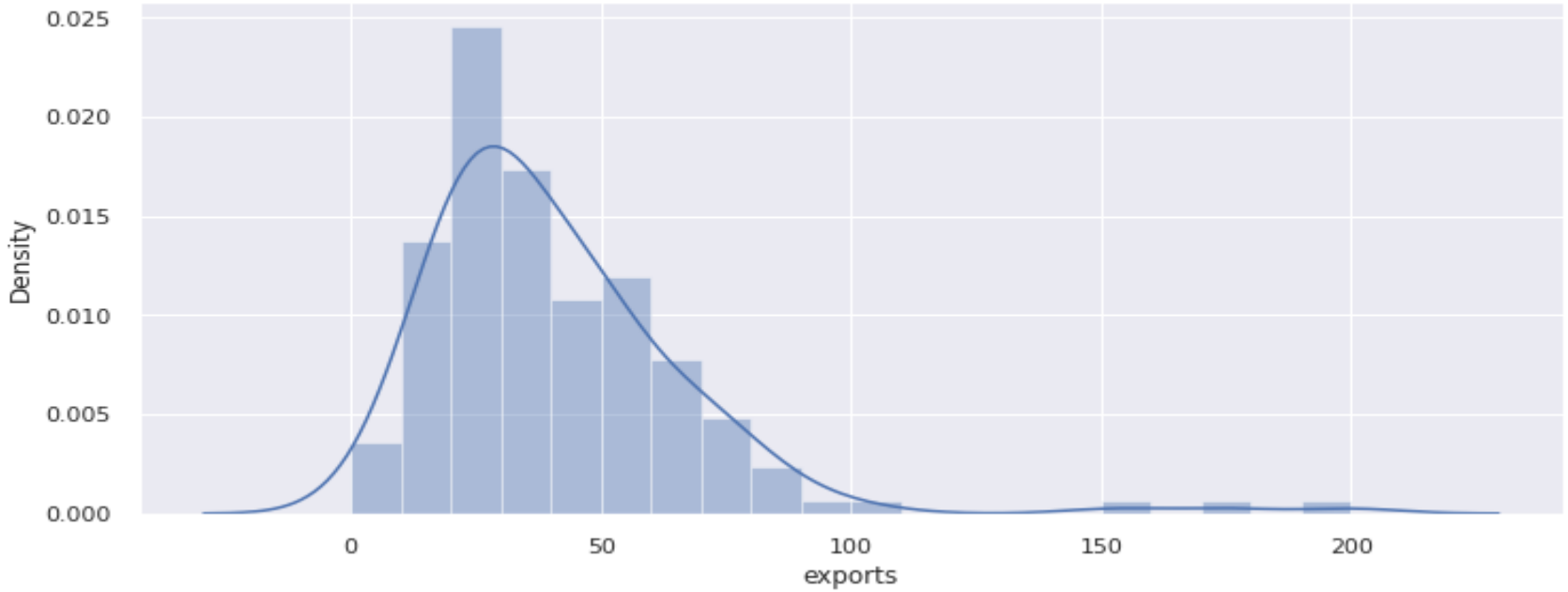
Life Expectancy and Total Fertility are **highly correlated** with Child Mortality. The coefficients are **-0.89 and 0.85**, respectively. Therefore, we have to remove these features.



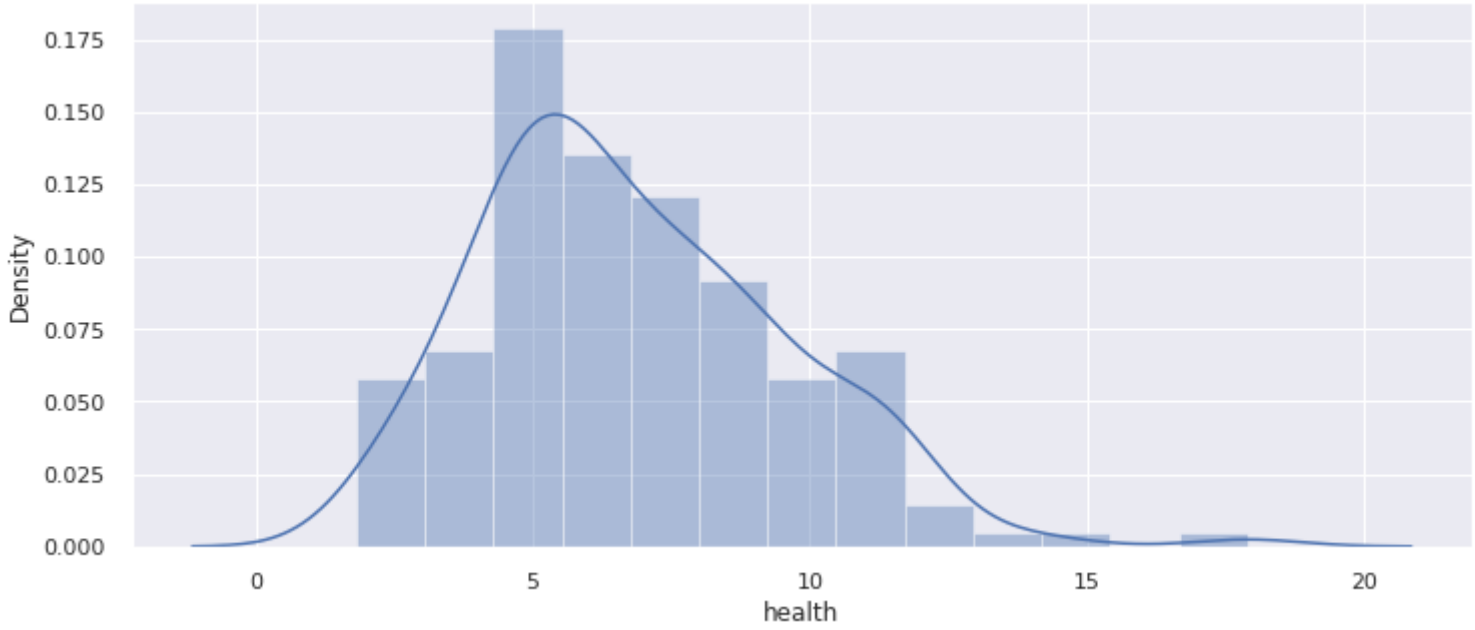
Data Understanding
Distributions
Child Mortality



Data Understanding
Distributions
Exports



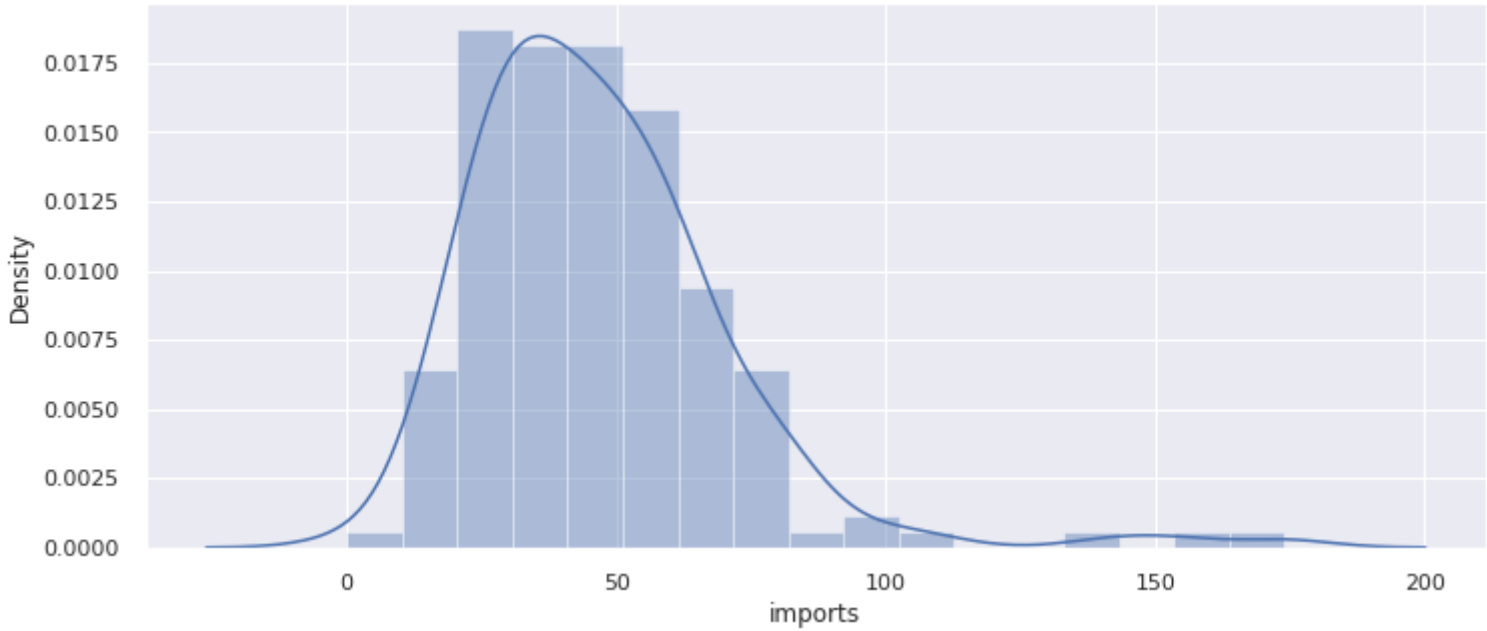
Data Understanding
Distributions
Health Spending



Data Understanding

Distributions

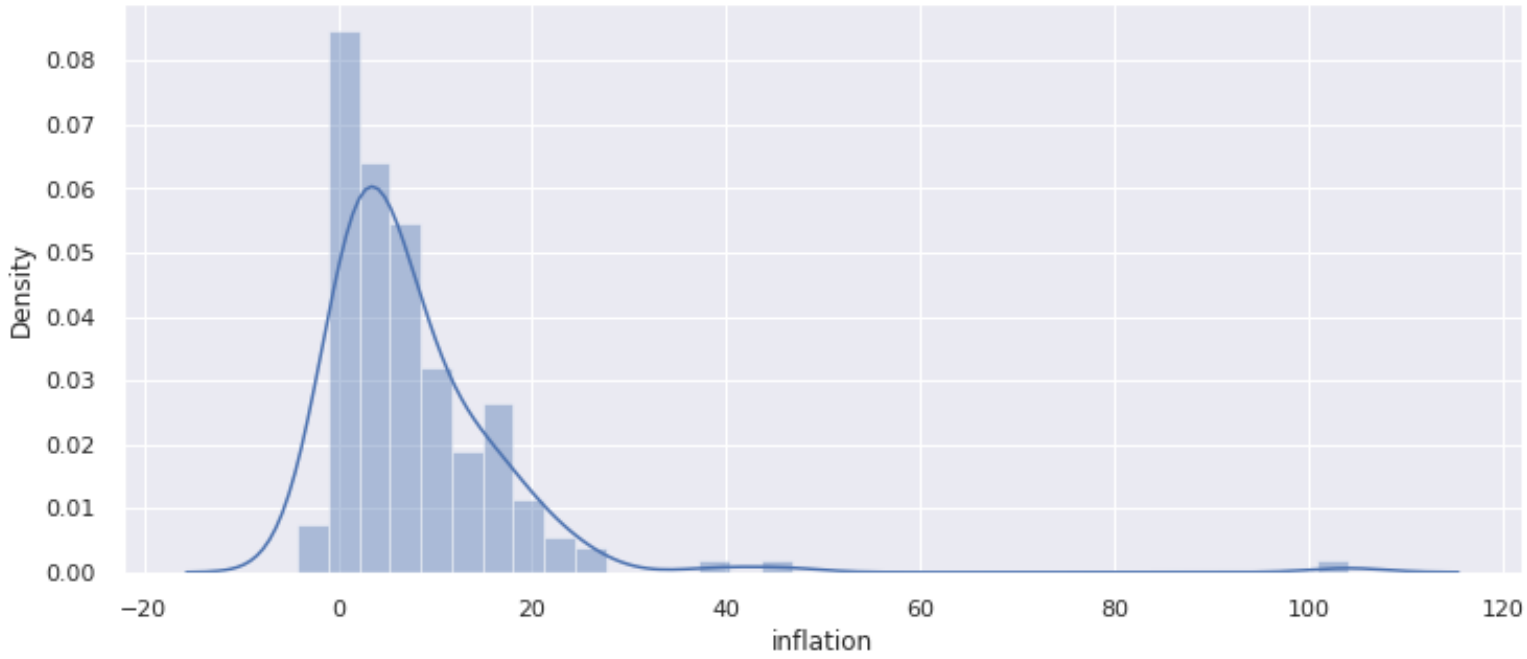
Imports



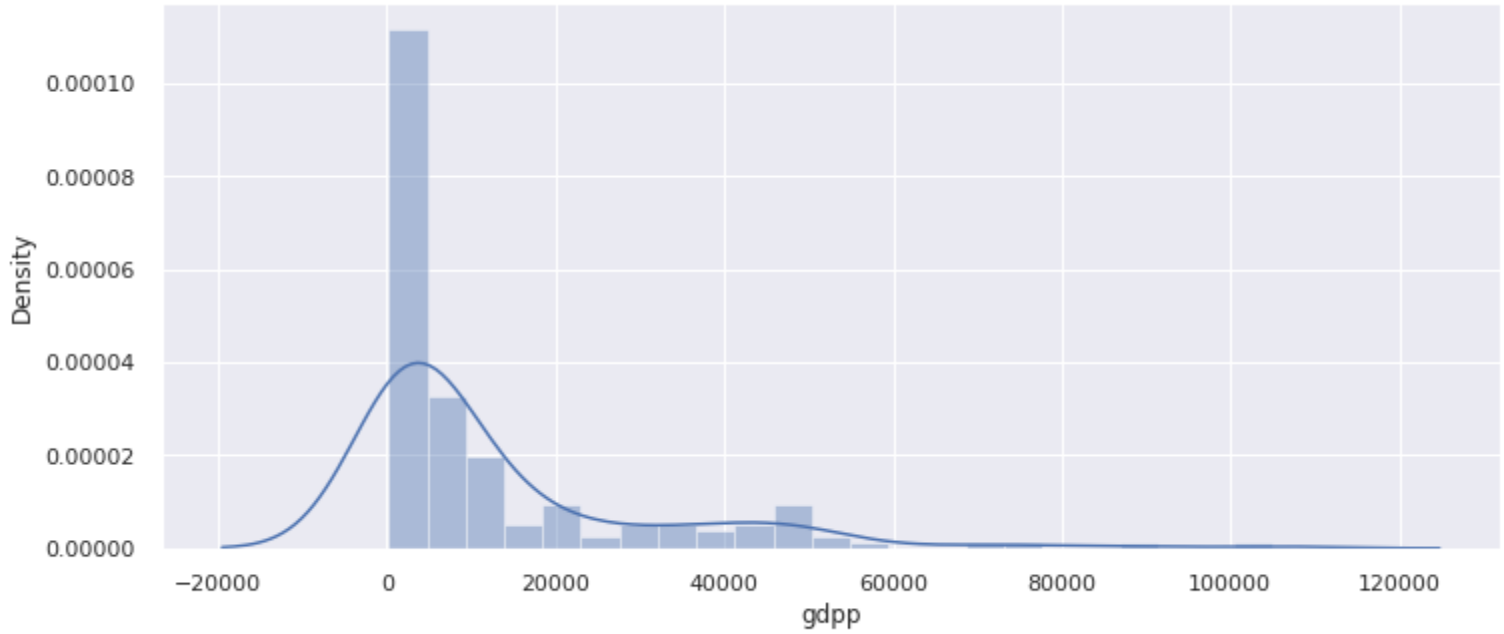
Data Understanding

Distributions

Inflation



Data Understanding
Distributions
GDP per Person



Data Understanding
Distributions

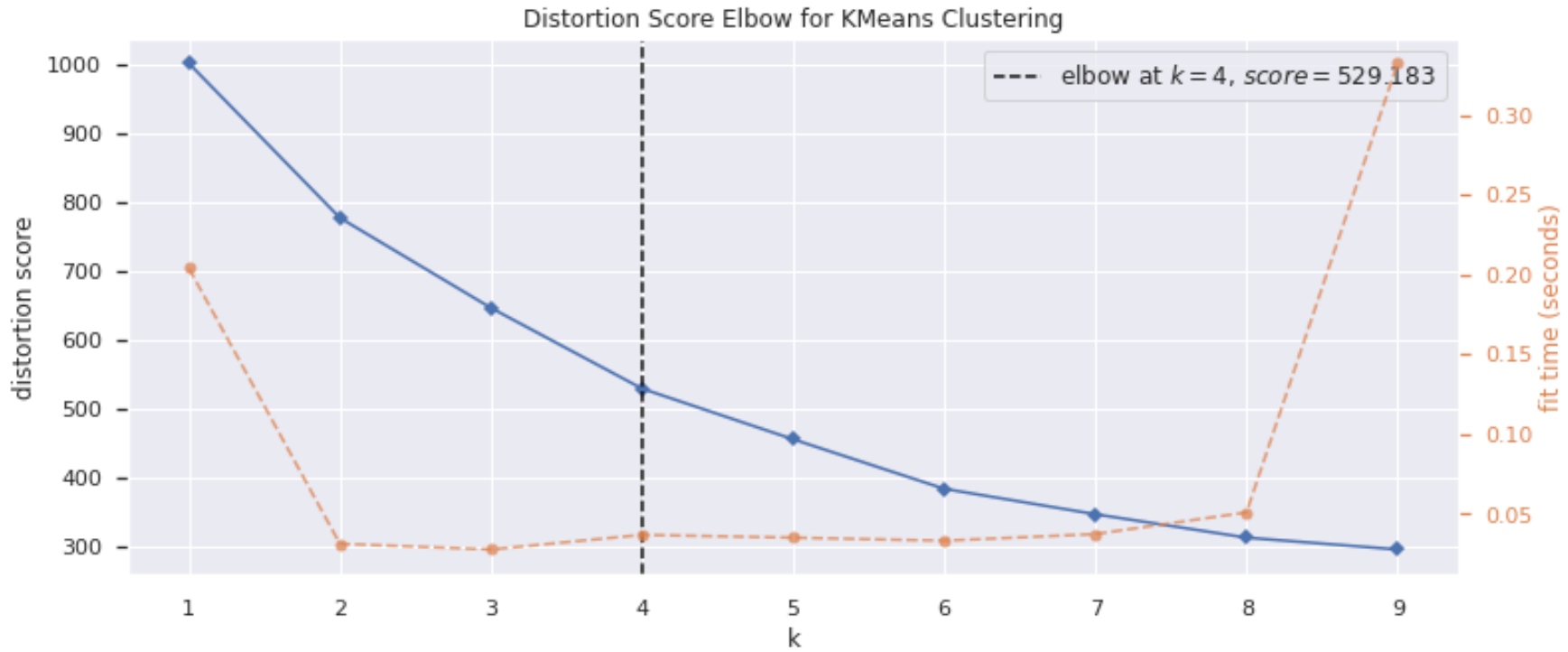
- Some features are close to normal, while others are highly skewed. We have to scale every feature.

Data Preparation

- The columns “country”, “life_expec”, “total_fer”, and “income” were dropped. “country” was non numeric while other features highly correlated with other features.
- The remaining features were standardized using the “StandardScaler()” function from “sklearn.preprossing” library.

Data Modelling

Elbow Curve for Optimal No. of Clusters



Data Modelling

K-means clustering was performed using 4 clusters.

The distribution of countries were as follows:

Cluster 1 => 25 countries

Cluster 2 => 69 countries

Cluster 3 => 3 countries

Cluster 4 => 70 countries

Data Evaluation

Comparing means of clusters with whole data of all features

	Whole Data	Cluster 1	Cluster 2	Cluster 3	Cluster 4
child_mort	38.270060	4.8640	69.904348	4.133333	20.481429
exports	41.108976	41.3640	27.924623	176.000000	48.232857
health	6.815689	10.0604	5.660000	6.793333	6.797000
imports	46.890215	37.8880	33.909651	156.666667	58.195714
income	17144.688623	42960.0000	6869.608696	64033.333333	16043.714286
inflation	7.781832	1.4916	13.260507	2.468000	4.855671
life_expec	70.555689	80.5520	64.362319	81.433333	72.624286
total_fer	2.947964	1.8232	4.116232	1.380000	2.265286
gdpp	12964.155689	45716.0000	3226.391304	57566.666667	8954.185714
clusters	1.706587	0.0000	1.000000	2.000000	3.000000

Recommendation

- According to the averages, countries in cluster 2 have the worst socio-economic and health indicators. Therefore, these countries are in the direst need for help.

Links

Notebook

https://www.kaggle.com/code/sarfarazjamal/sarfarazjamal17093assignment2?scriptVersionId=92823170

Data

https://www.kaggle.com/code/sarfarazjamal/sarfarazjamal17093assignment2/data?scriptVersionId=92823170

Thank You