TASK2

Clustering is an unsupervised learning technique that involves grouping together a set of data. In a machine learning system, this kind of grouping helps with understanding a dataset. Kmeans clustering is the method that would be used for this analysis, and this was be programmed on the python file Task2.py attached.

Keynotes

- Elbow method: This is used to figure out the number of clusters in a data set. The process entails determining the number of clusters to adopt by plotting the explained variance as a function of the number of clusters and choosing clusters at the elbow of the curve.
- Centroid: the actual or theoretical spot that serves as the cluster's centre.

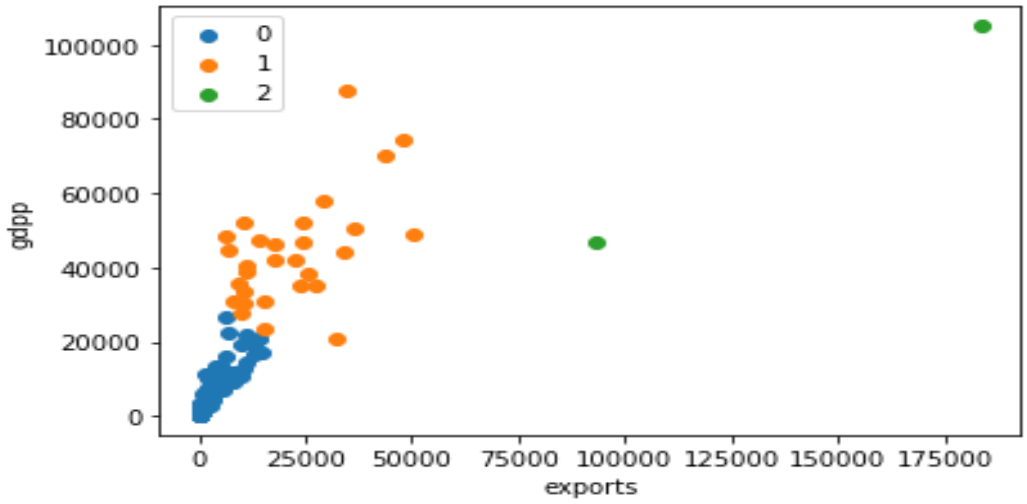
STEP 1

All the necessary libraries and machine-learning tools were imported. There are 167 entries (representing the countries) and 10 features. Export, import, and health were in percentages of GDPP. This was changed in order to achieve similar data for all features. There were no missing values in our data set.

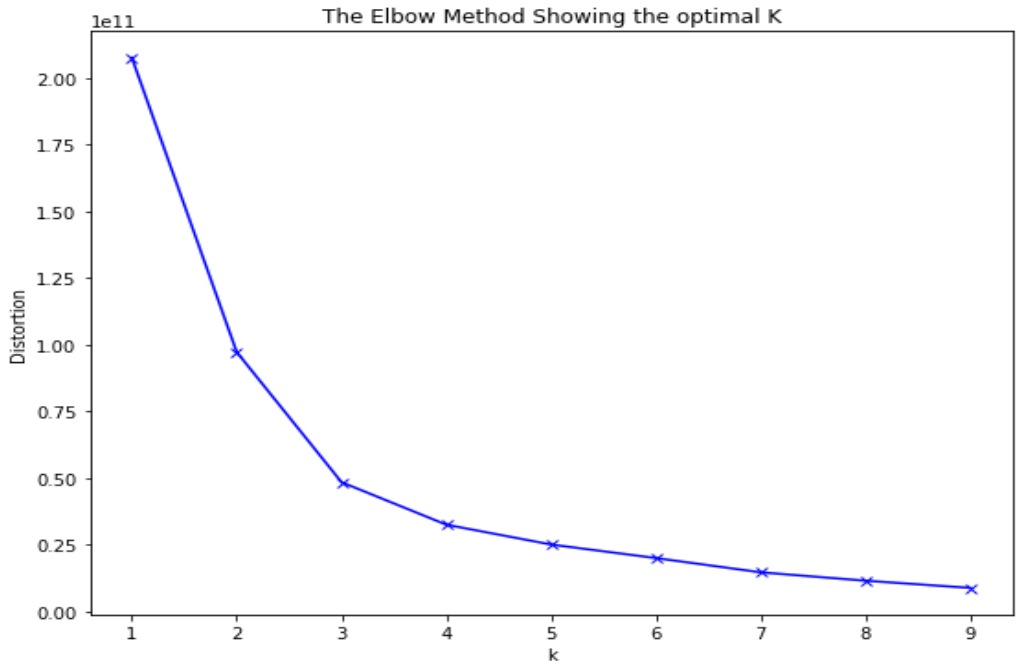
A correlation matrix was then used on the data sets to understand the relationship between the features. GDPP is significantly correlated with income, health, exports, and imports. Also, total fertility is highly correlated with both child mortality and inflation.

STEP2

In this work, 2 features were first visualized. GDPP to export.



Using the Elbow method as shown in Task2.py, the best number of clusters to be used is 3. The countries are then clustered into 3 using three features GDPP, Imports and export (determined based on the correlation matrix).



The total imports and total exports are important in estimating a country's GDPP (National income). These 167 countries were then clustered into 3. Clustering 0 has 136 countries, Clustering 1 has 29 countries and clustering 2 has 2 countries (Countries in details shown in Task2.py). After checking the average of the total export, import and GDPP it was observed that countries in Clustering 2(Luxembourg and Singapore) have better GDPP, imports, and export as compared to clusters 0 and 1.

STEP3

3 other features were added (Health, Income, and life expectancy). The number of countries in the clusters changed as cluster 0 has 129 countries, Cluster 1 has 36 countries and cluster 2 has 2 countries. After getting the average of the features using the mean, it was gotten that cluster 1 is doing better, showing a better average of all 6 features used.

All features (GDPP, imports, exports, life expectancy, income, health, inflation, total fertility, and child mortality) was used and showed clustering 0 to 48countries, clustering 1 to 28countries, and clustering 2 to 91countries respectively. This also showed that child mortality is the highest in cluster 0 and least in cluster 1, Inflation is highest in cluster 0 and least in cluster 1, and total fertility is high in cluster zero and least in cluster 1.

Conclusion

According to the clustering done on this data set, it can be said that this data set helps to cluster countries according to different categories and more informed decisions can be made. Lesser features give deeper insight into the clustering. In a case where an investor wants to invest, start a business or open branches in different countries using the GDPP, export, import, and inflation on this data set can give insight into the best countries to start a business in. Also, if an NGO wants to help people that are in need factors such as health, child mortality, and total fertility (which could lead to increased population) can help determine countries that are best in need of help.