Clustering and Fitting in Data Sciences

Introduction

Cluster is combination of data points which have some similarities and different from other clusters data points where fitting is a measurement of model to generalized similar data for which it was trained. Here our aim for clustering of data with over fitting model to produce more accuracy. Here we are going to use clustering and fitting on country's data set to make country's analysis according to DGP, income , CO_2 emission , and fuel consumption.

Data Set

Used two data sets

- Country to country change of climate regarding emission of CO_2 from world bank
- Dataset about countries related to their health, income, , import, exports, gdpp, inflation, child-mort, life expenditures etc.

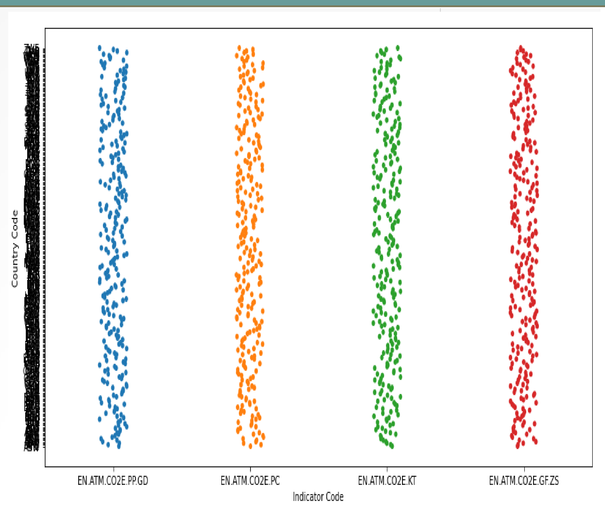

Fig 1: Overall Clustering of indicator code

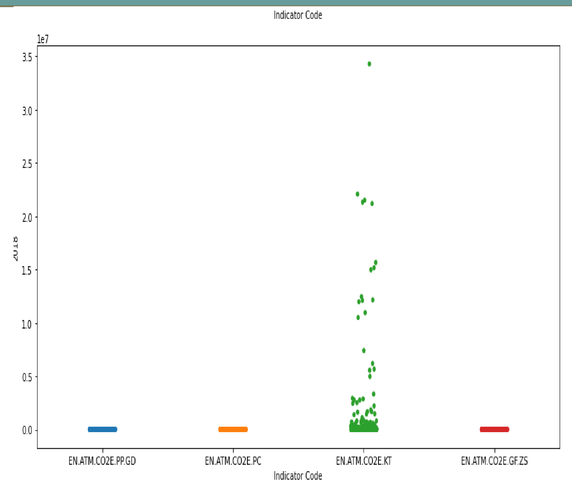

Fig 2: Clustering of indicator code in 2018

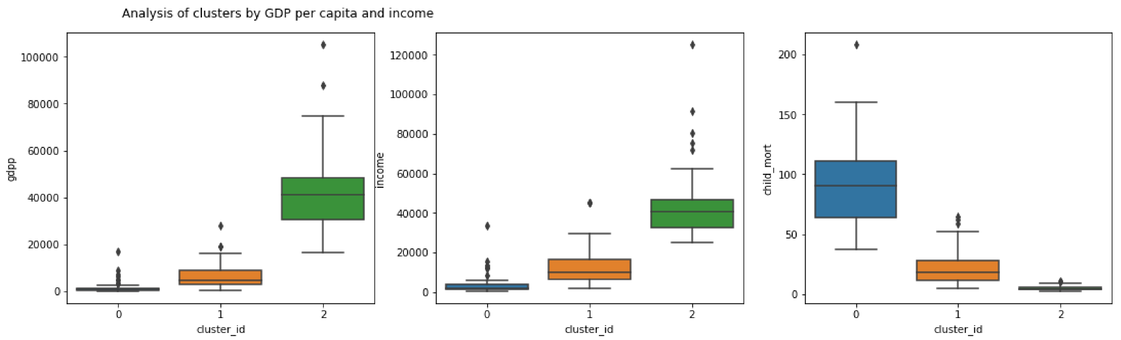

Fig 3: Analysis of clusters by GDP per capita and income of countries

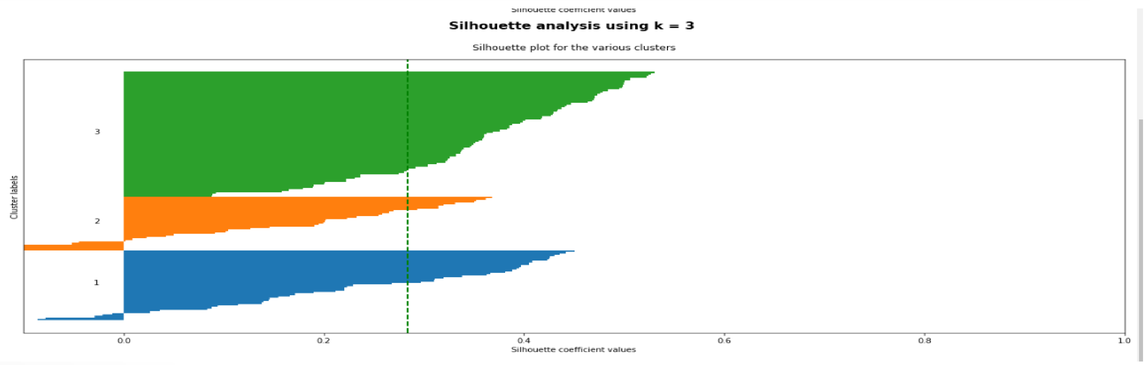

Fig 4: Best optimum cluster for 2nd dataset with k=3

Methodology

- After loading data indicators will be selected data through which we will got data related to indicators with countries
- Take median of null values
- Plot clustering with indicators

1. For 2nd dataset take country as targeted feature and other columns as training features
2. Scaling data and feasibility of cluster
3. Identification of right number of cluster using silhouette and elbow effect analysis and apply K-mean clustering.
4. Making comparison of cluster with plotting on bases of GDP per capita and income.

Conclusion

So overall as a conclusion related to our dataset clustering and fitting become more helpful with considering 1st dataset to make a comparison of courtiers with more CO_2 emission and their GDP per capita from last 40 years we have found the in early years 1960, 1970 and 1980 indicator of more fuel consumptions by countries and in 1990 to 2018 we have found CO_2 per $ of GDP for countries.

With concerning 2nd dataset by using elbow affect and silhouette method w have found best optimum clusters 3. With analyzing clusters according to GDPP, income and child-morality from fig3 we have found cluster 2 contain developed countries like USA, Australia, UK with more GDP, income and low in child morality, cluster 1 contain under development countries with moderate GDP, income and child-morality and cluster 0 contain low develop countries like Afghanistan with low GDP, income and higher in child morality.

References

1. Mishra, S. (n.d.). *Unsupervised Learning and Data Clustering*. Retrieved from towardsdatascience: https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a
2. *Model Fitting*. (n.d.). Retrieved from datarobot: https://www.datarobot.com/wiki/fitting/