Assignment 3: Clustering and fitting

By: Ali Hamza Yasin (21090741) School of physics, Engineering and Computer Science
Supervisor: Ralf Napiwotzki University of Hertfordshire
GitHub link: https://github.com/AliHamzaYasin/Applied-Data-Science-Assignment-3-.git

Introduction

Developing countries are expected to bear the brunt of climate change. The effects of climate change, such as elevated temperatures, changes in rainfall patterns, rising sea levels, and more frequent weather-related disasters, pose significant risks to agriculture, food, and water supplies in these countries. Such risks jeopardize the progress made in reducing poverty, hunger, and disease, and could significantly impact the lives and livelihoods of billions of people in developing countries. To mitigate these risks and contribute to a global solution, the World Bank Group is providing support to developing countries. The Climate Change Dataset covers various indicators related to climate systems, greenhouse gas emissions, resilience, energy use, and exposure to climate impacts. Other relevant indicators can be found under different data pages, including Agriculture & Rural Development, Environment, Energy & Mining, Health, Infrastructure, Poverty, and Urban Development.

The Dataset

The climate change dataset comprises 76 unique indicators and covers data from 1960 to 2021, including information for 266 different countries. However, certain indicators have missing data throughout the entire period. The dataset includes columns such as Country Name, Country Code, Indicator Name, Indicator Code, and year columns spanning from 1960 to 2021. During the pre-processing stage, the dataset will be carefully cleaned to remove null values. Additionally, some data molding and scaling steps may be performed to facilitate analysis of a specific indicator across all countries.

Methodology

The following steps will be taken to achieve meaningful clusters:

1. Collect and preprocess the data.
2. Select relevant indicators and years to compare the data and gain insights.
3. Scale/normalize the data using the Standard Scaler function in sklearn.
4. Utilize the KMEANS unsupervised learning method to create clusters for the selected indicators and years.

To generate the curve_fit line and predict for the next 10-20 years, the following steps will be taken:

1. Select the required data for the desired Country Name.
2. Preprocess the data and apply exponential growth.
3. Use the predicted array to obtain the best-fit line and generate the curve_fit as shown in the figure.

References

https://sparkbyexamples.com/pandas/pandas-count-unique-values-in-column/https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.htm lhttps://data.worldbank.org/topic/climate-changehttps://www.geeksforgeeks.org/python-scipy-curve_fit-with-multiple-independent-variables/

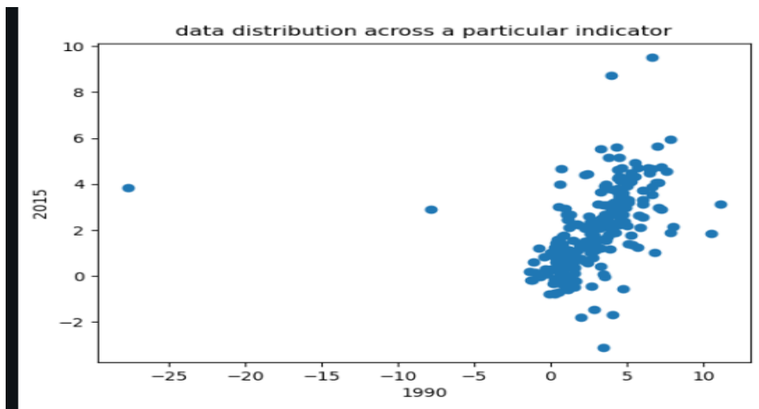
https://sparkbyexamples.com/pandas/pandas-reshape-series/#:~:text=We%20can%20reshape%20the%20pandas,you%20wanted%20to%20reshape%20to

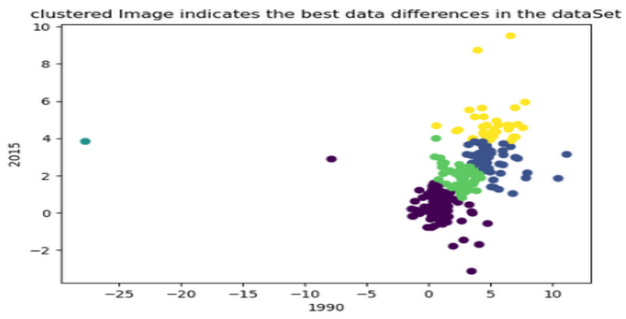
https://datagy.io/numpy-exp-exponential/#:~:text=Understanding%20the%20np.,-exp()%20Function&text=The%20NumPy%20exp()%20function,element%2C%20passed%20into%20the%20function

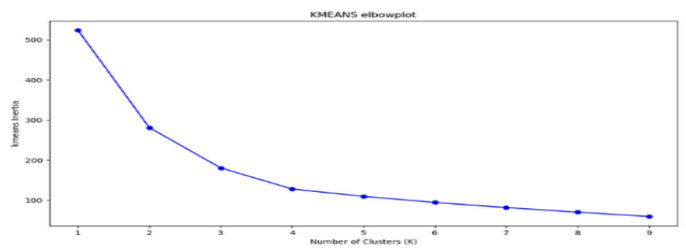
The Project

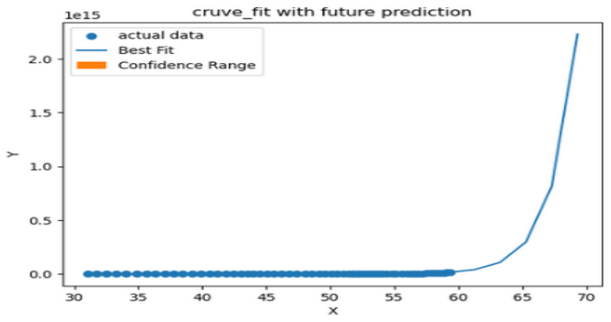
The primary objective of this project is to utilize clustering and curve-fitting techniques on the Climate Change dataset, utilizing the KMEANS clustering algorithm and curve_fit respectively. Obtaining meaningful clusters is crucial to gaining insights into the distribution of data for various indicators in the Climate Change dataset, such as Urban population (% of total population), Urban population, Urban population growth (annual %), Population growth (annual %), among others. The curve_fit approach employs non-linear least squares to fit the data and derive the optimal parameters from the dataset.

Preliminary Result and Analysis









The abore mentioned figures depict the overall structure and data distribution of a specific indicator in the Climate Change dataset. Some key points related to these figures include:

A) The left figure displays the complete data distribution of the selected indicator, i.e., Urban population growth (annual %).

B) The right figure shows the performance of the KMEANS algorithm for 12 different k values. Based on the figure, the 4th k_value seems to be optimal since it does not show any significant improvement for a lower k value.

C) The left figure in the bottom row presents the 4 different clusters obtained from our dataset using the KMEANS algorithm. This is in line with the k_value selected earlier.

D) The right figure in the bottom row illustrates the best-fitted line for a specific country (Arab World) and the selected indicator (Urban population (% of total population)). It also depicts the future prediction for the same using a simple exponential growth technique.

The Next Steps

Once the data have been properly clustered after the prepossessing stage, classification of the data to identify/get the desired insights by comparing the different years using any indicator. Next step is to get the different comparing insights using the different generic functions loops to get the more insights that a human eye just can't see for all the years at a single time.

University of Hertfordshire UH