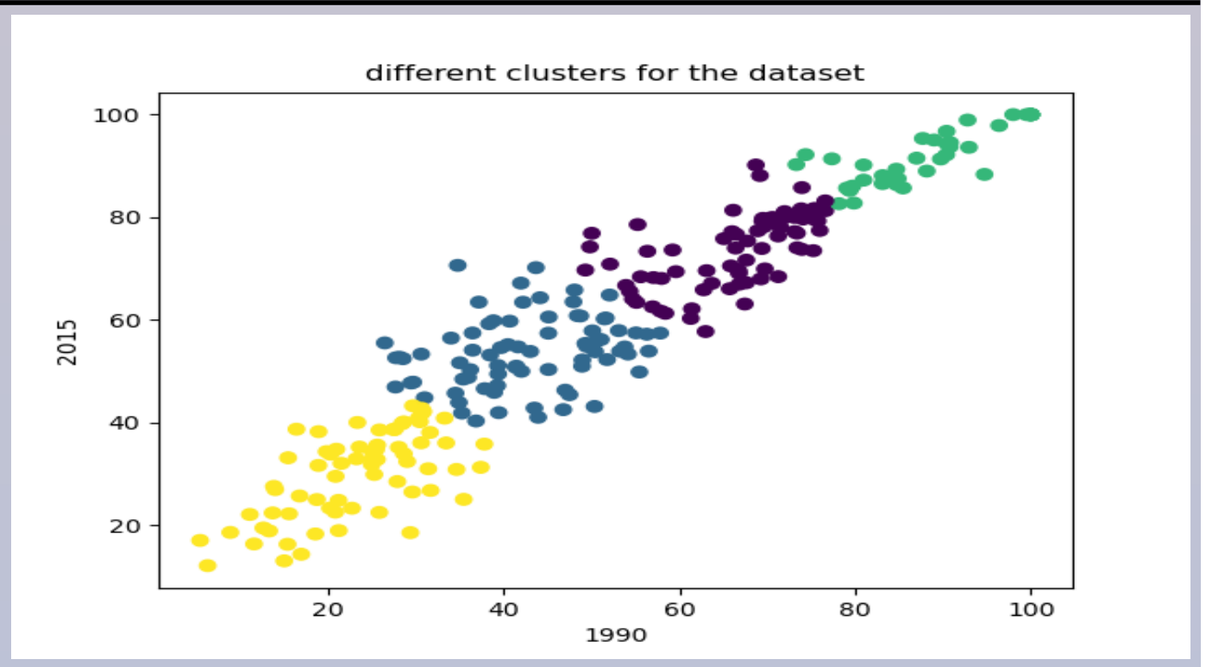
University of Hertfordshire UH

ASSIGNMENT 3: CLUSTERING AND FITTING

BY – Sabir Khan (22020950)
Supervised by – Ralf Napiwotzki

Introduction:

Climate change is expected to hit developing countries the hardest. Its effects—higher temperatures, changes in precipitation patterns, rising sea levels, and more frequent weather-related disasters—pose risks for agriculture, food, and water supplies. At stake are recent gains in the fight against poverty, hunger and disease, and the lives and livelihoods of billions of people in developing countries. The World Bank Group is helping support developing countries and contributing to a global solution. Climate change Dataset cover climate systems, exposure to climate impacts, greenhouse gas emissions, resilience, and energy use. Other indicators relevant to climate change are found under other data pages, Agriculture & Rural Development, particularly Environment, Energy & Mining, Health, Infrastructure, Poverty, and Urban Development.



The Dataset:

Climate change datasets contains the 76 different kind of indicators and have the data from 1960 to 2021. The dataset also hold data for 266 different countries. Some of the indicator's data missing for all years all over throughout the years from 1960 to 2021.
The dataset contains the specific columns like Country Name, Country Code, Indicator Name, Indicator Code, and the year columns from 1960 to 2021.
The dataset will be thoroughly refined at the preprocessed stage like nulls removals.
Some of the Data Molding and scaling steps also will be done to make the for any specific indicator throughout all the countries.

Primarily result and analysis

the figs show the overall structure and data distribution for the climate change dataset.
Some of these points related to that figures are:

A) first left figure shows the overall data distribution for the indicator (Urban population (% of total population))

B) 2nd fig shows the 12 different k value for KMEANS algorithm which we can choose ourselves, as we can see that 4th k value will be more than enough for our case, because it just stopped moving downwards. That means this could be the best value as k_value score.

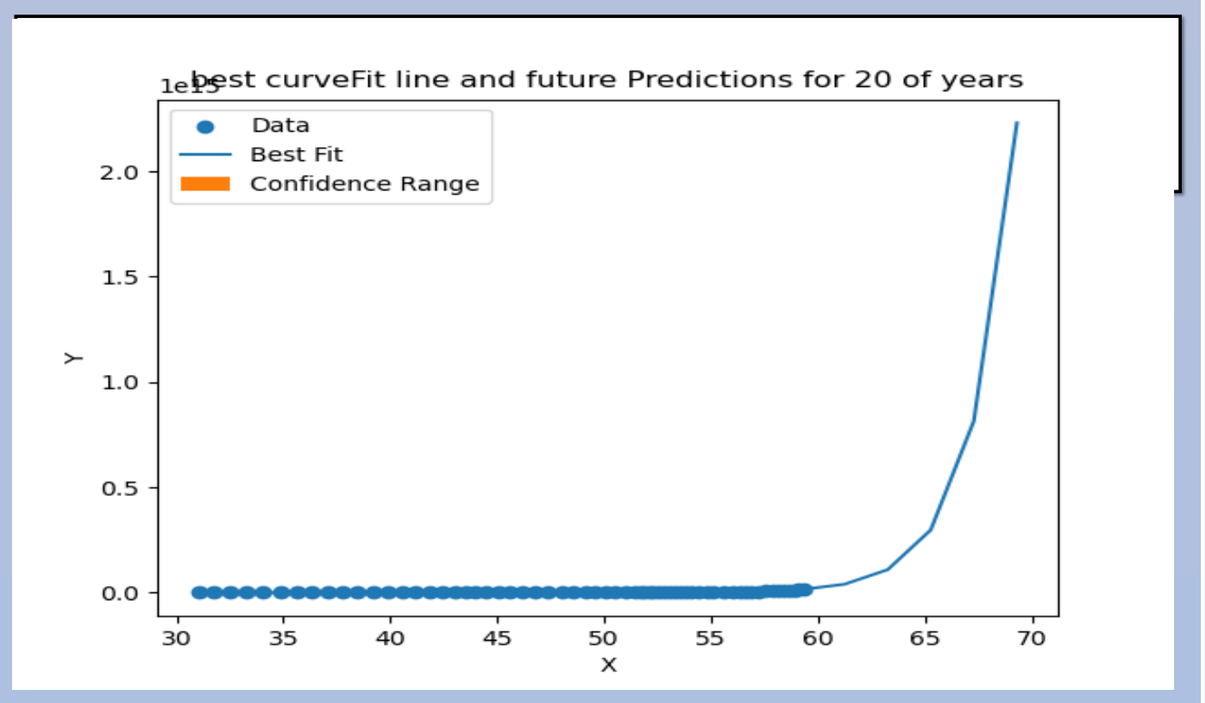
C) Third figure shows the 4 different clusters in our dataset indicator (Urban population (% of total population)) for the country (Arab World), it shows 4 just because of we select the k_value as 4.

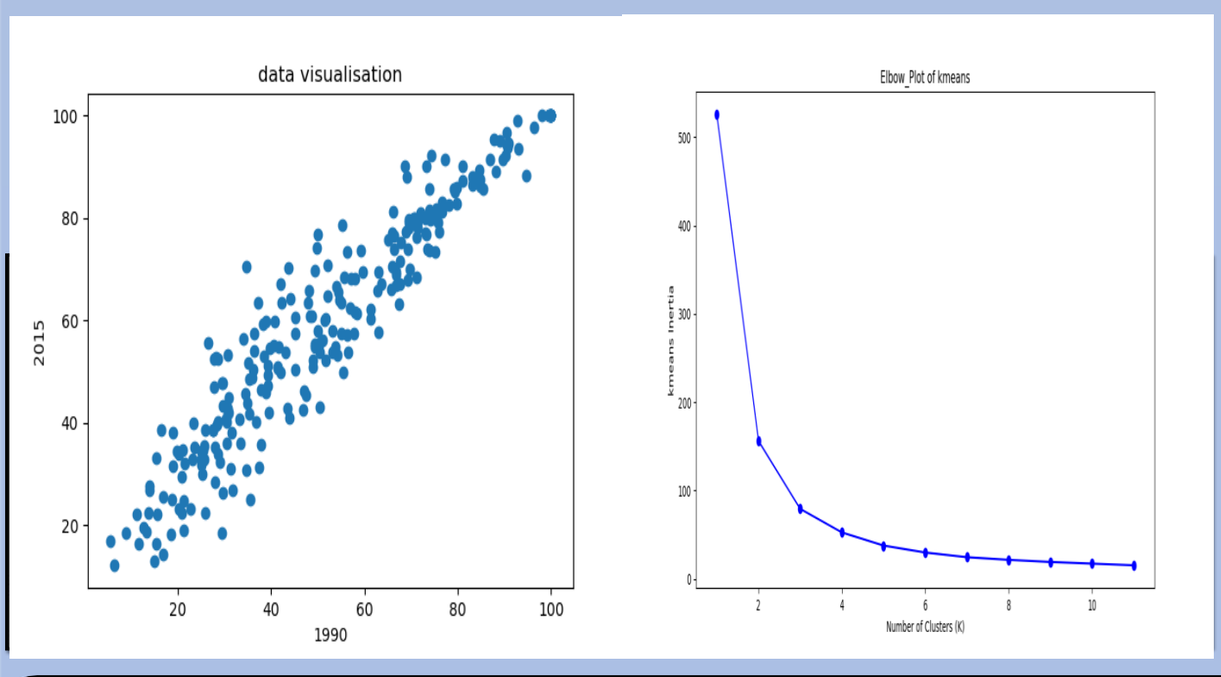
D) last right figure indicates the best fitted_line and future prediction after following the simple exponential growth technique.

Methodology:

The meaningful clusters, will be achieved in these different steps:
1. data collection and data pre-processing
2. indicator selection and year selection to compare the data on different years to get the insights.
3. scale/normalize the data using the sklearn's Standard Scaler function
4. Follows the KMEANS unsupervised learning method to create the cluster for the selected indicator and years.

The second step is to generate the curve_fit line and predict for the next 10- or twenty-years' time including confidence range. They could be achieved by these steps:
1. first step is the data selection as required to get the curve_fit for required Country Name.
2. data pre-processing and exponential growth for the data.
3. predict the array to best fit the line and get the curve_fit as shown in the fig.





The Next Step:

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 particular 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 of the years at a single time.

The Project

This project mainly aims to focus on clustering and fitting techniques on the Climate Change dataset using one of the cluster-based algorithms like KMEANS and the curve_fit respectively.
Meaningful cluster is a good idea to get the insights and data distributions for the Climate change dataset indicators e.g(Urban population (% of total population),Urban population, Urban population growth (annual %), Population growth (annual %) etc.)

References:

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

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

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

GitHub Link: https://github.com/Sabir1stt/Applied-Data-Science-asssignmnt-3.git