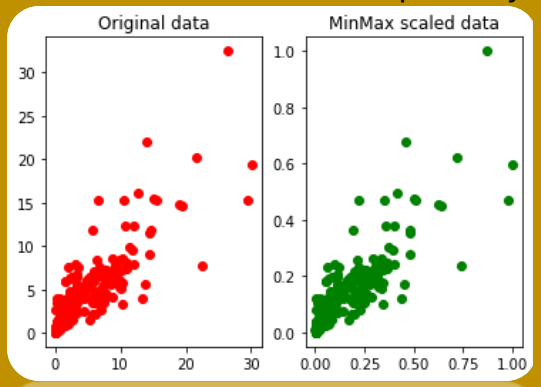
CLUSTERING AND CURVE FITTING OF CO_2 EMISSION OF COUNTRIES

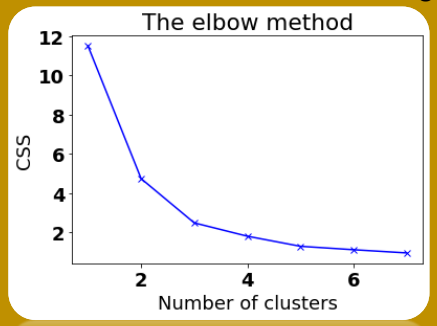
Introduction: World climate is changing drastically due to human activities, so it is quite good thought to analyse factor like CO_2 emission of more than 170 countries. In this poster, CO_2 emission data of year 1990 and 2019 is analysed with K-means clustering. Later on, curve fitting is applied to see the impact and role of any country in climate change in 23 years.

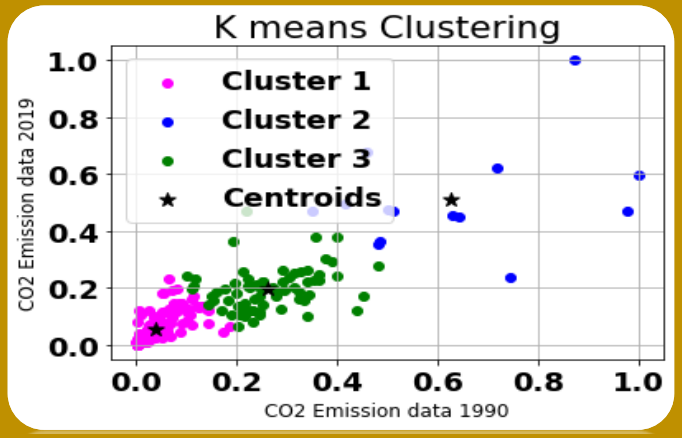
Normalization Results: In this work, Min-max normalization is used distributed the CO_2 emission data set in range [0,1]. The two scatter plots with red and green datapoints are showing original and Min-max scaled data respectively.



K-Means Clustering Results

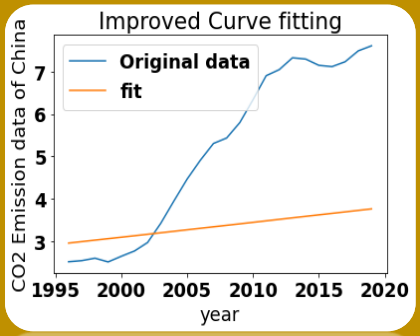
In this model, we limit cluster numbers to 3 after using elbow method. As figure below is showing clusters of three colors whereas centroids are in black color. This graph is clearly depicting that countries in second clusters (blue) are contributing as major part in CO_2 emission in both 1990 and 2019. But countries that are less contributing in CO_2 are in first cluster (pink) whereas countries that are at average position in both years are in third cluster. It can be concluded after interpreting values of second cluster that ***Bahrain*** has highest level of CO_2 emission.

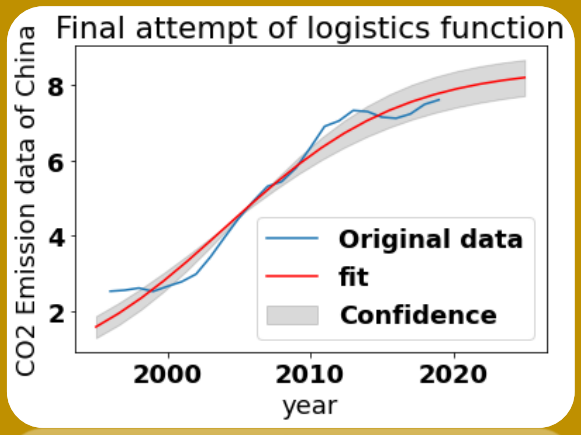




Curve Fitting Results

Logistic curve fitting is used over data of CO_2 emission of China from 1996 to 2019. In first attempt, we fit the data with default initial parameters. But in improved attempt, we got better results by scaling with the 2012-year emission data. In final attempt we got the best fit with confidence interval. Curve fit is also predicted of data after 2022 to 2025 that have upward trend showing increase in CO_2 emission in coming years





Conclusion: K-means clustering and curve fitting shows distribution of entire data into three clusters and fitting of logistic model over data. It distributes the data into three clusters in good manner but due to presence of some outliers, k-means cannot be used as an ideal technique for grouping data. Moreover, logistic curve fitting effectively fitted on data set in some S-shaped with estimate of confidence interval.

Github link: https://github.com/AaMNAHZaAFAR/Clustering-on-World-bank-data