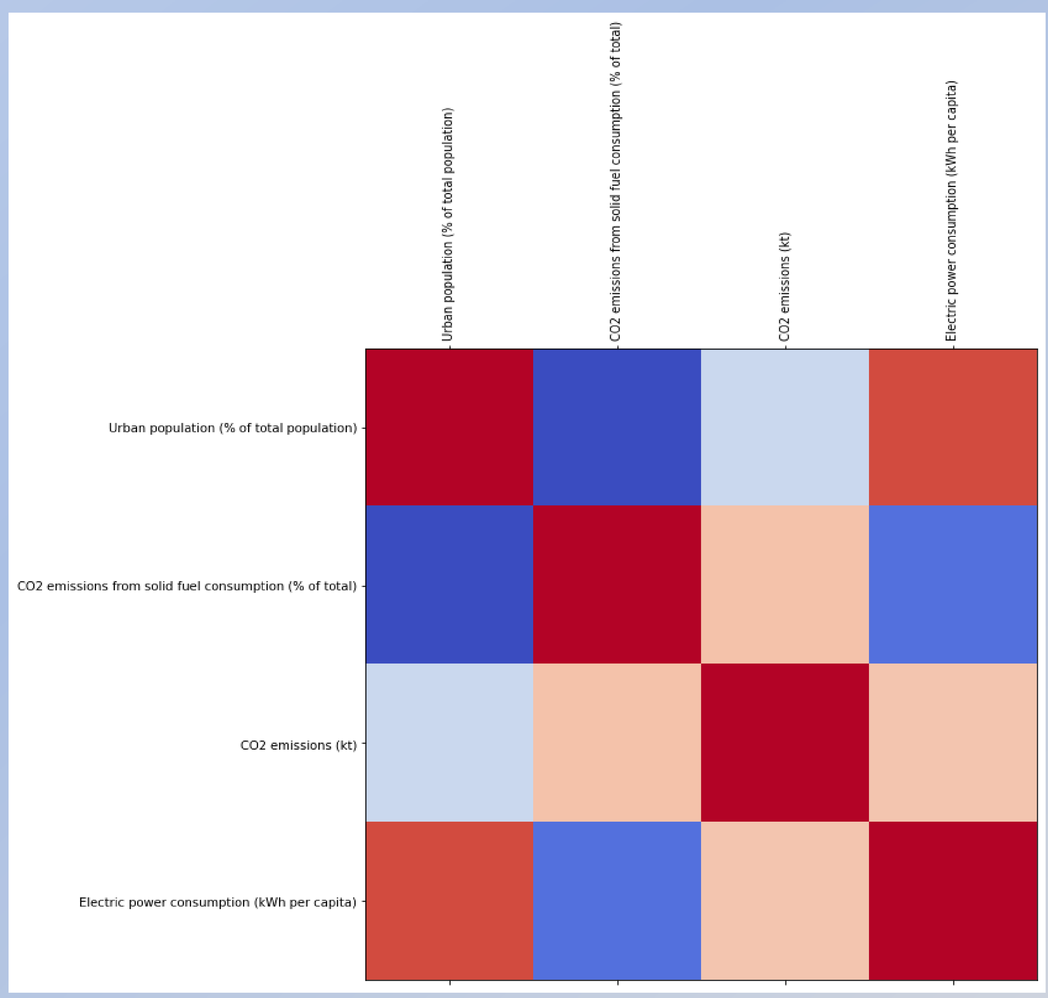
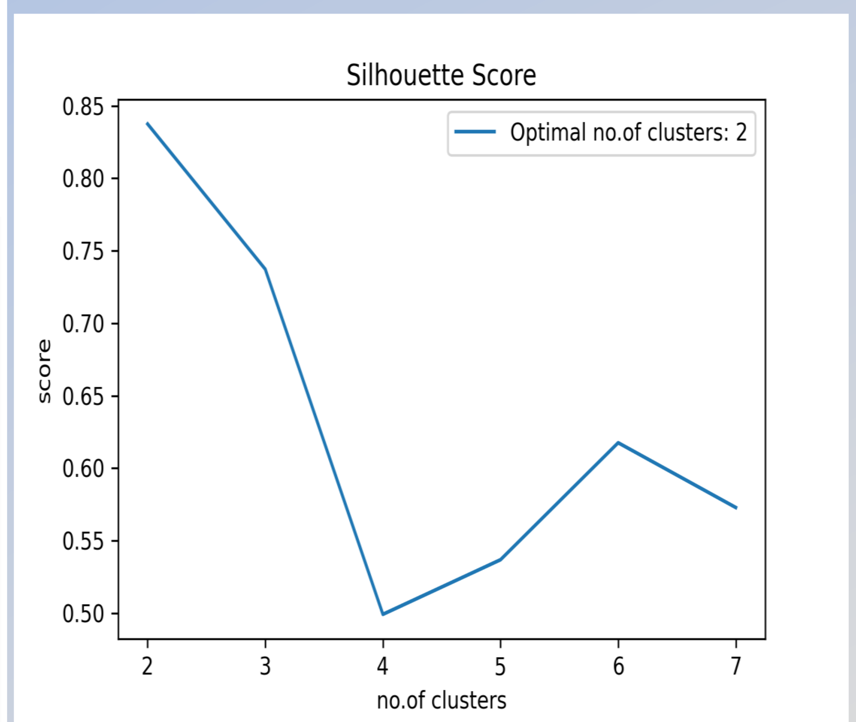
Clustering and Fitting using World Bank - Climate change Data

Jerome Paulraj – 22074395

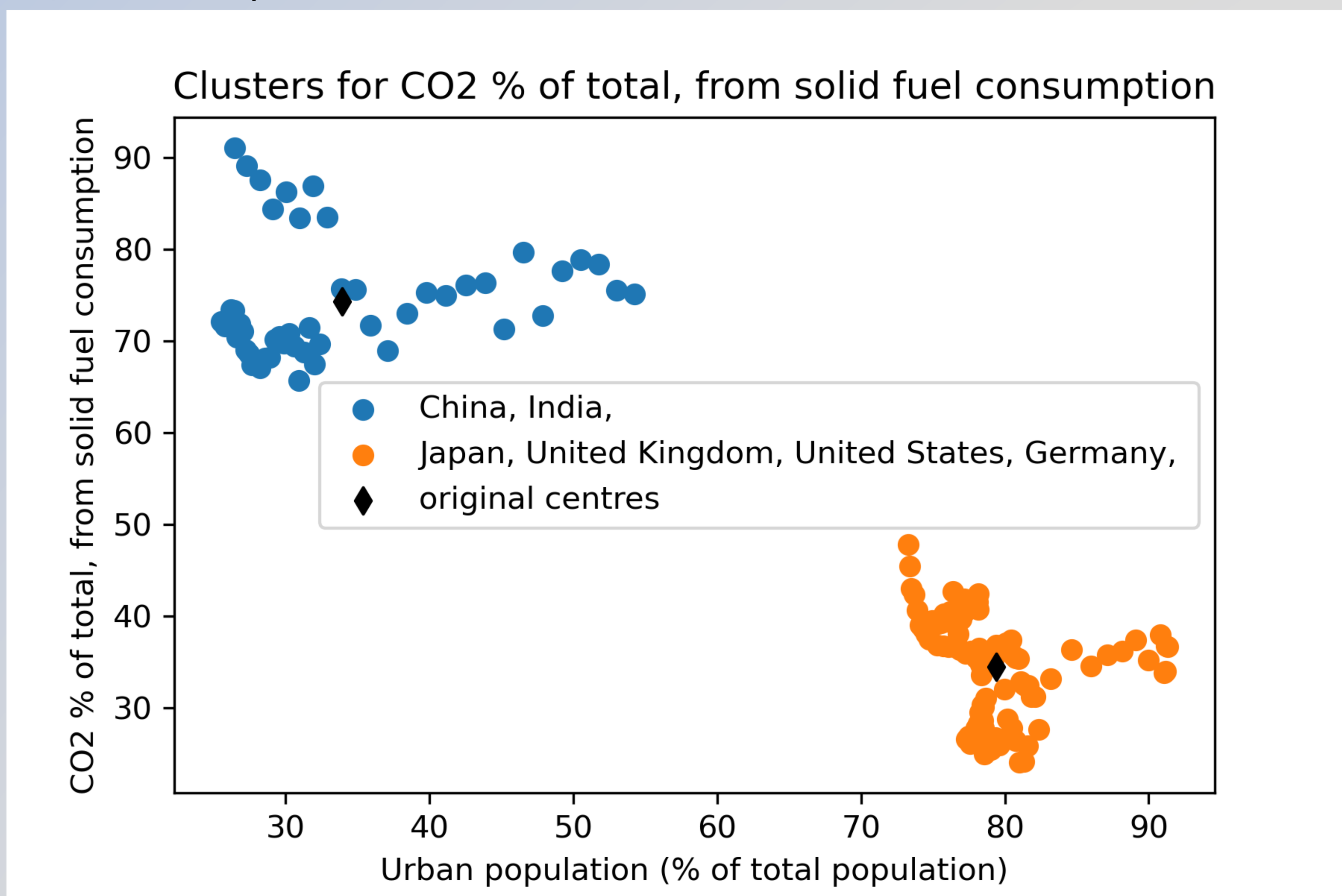
In this analysis, we are going to look at 6 countries. Namely, India, China, Japan, Germany, United States and United Kingdom. The dataset used contains information regarding climate change for countries from the year 1960 to 2022. This project identifies clusters in the dataset and attempts to fit an exponential function to predict the % of urban population for a selected country. Clusters are formed based on the urban population % and CO2 emission from solid fuel and the electricity consumption by each country. K-means along with silhouette score was used to achieve clustering.

The following heatmap tells us about how correlated each of the chosen indicators are. From this we select two indicators, namely CO2 emissions from solid fuel and electricity consumption by each country and compare it along with the urban population % of each country.



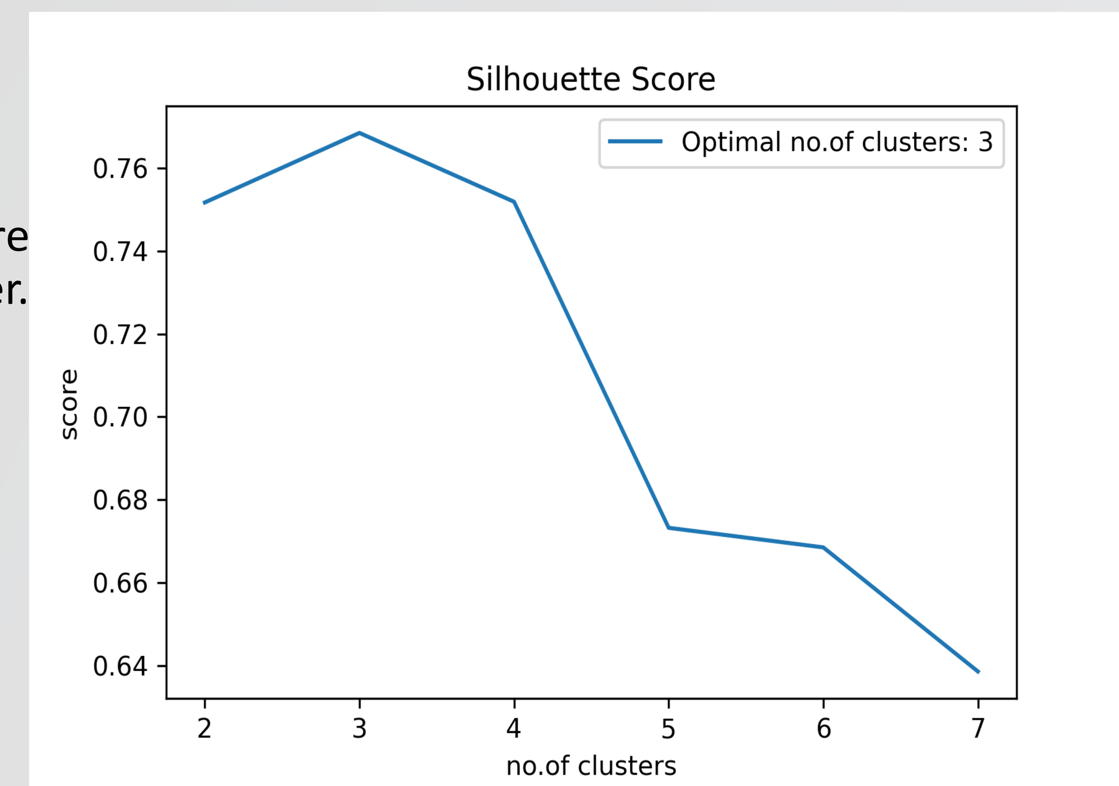


The number of clusters for clustering was decided using the silhouette score.
Let us look into the amount of CO2 emitted from solid fuel consumption first. Using K-means algorithm, we identify two clusters

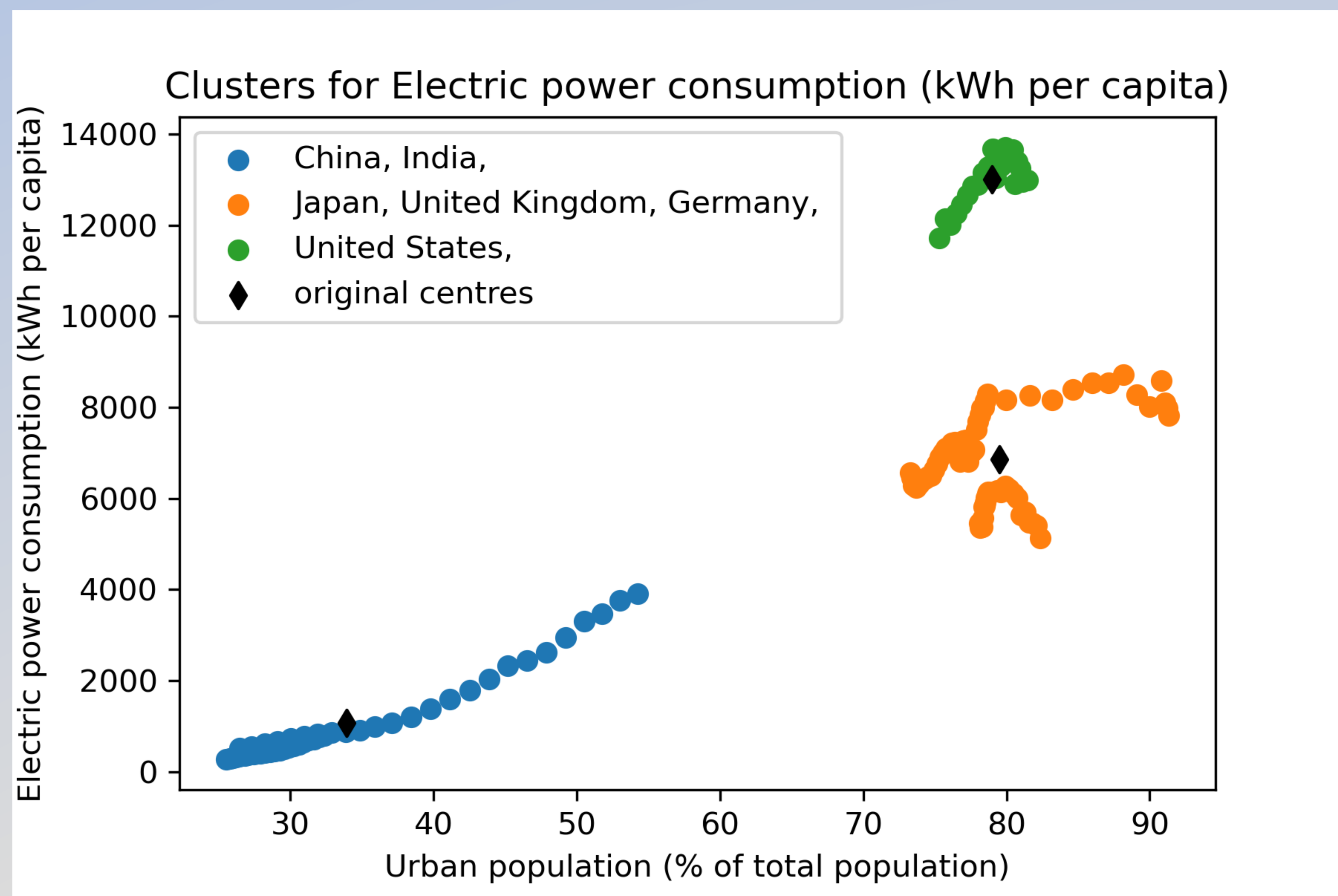


From the 6 countries, India and China have been grouped together and Japan, United Kingdom, United States and Germany have been grouped as one. What we can derive from this is that, the cluster has identified countries that are developing and developed. Which might explain why 4 countries have relatively very high urban population percentage. Contrary to normal belief that more urban cities cause pollution, it can be seen that despite having a significant difference in urban population, carbon output from solid fuels is way lesser in comparison to India and China. This could be due to availability of other power sources and in general a reduced use of solid fuel. For example most of the farming equipment might use solid fuels and electricity for rural areas might still use older technology which requires coal etc. Along with this, it could be assumed that the urban population has more options to reduce their carbon output by trying to use different sources of energy in comparison to rural or non-urban population.

Let us now look into the electricity power consumption of each country in the selected dataset. Which could give us more information about the first cluster. Using the Silhouette score graph, we can say that the data is now clustered into 3 groups.

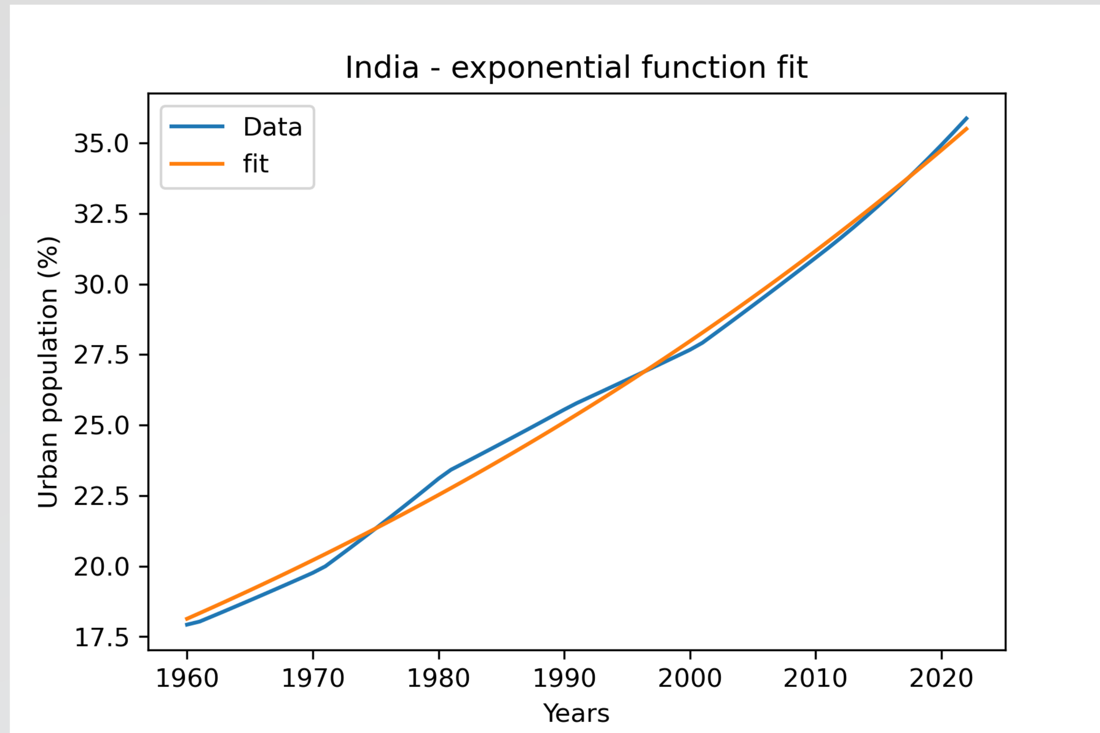


From the second cluster graph below, we can see that United States has the highest electric power consumption per capita, and interestingly even though India and China has the largest populations, both of them have relatively low electric power consumption. This could be explained by both of them having lower urban population. In the previous cluster graph, we assumed that urban population must have more options other than solid fuel for their power/electricity needs. This conclusion seems to be justified in this plot. It can be said that with more urban population, the electricity power consumption increases as well. Despite Germany being relatively bigger than both Japan and United Kingdom, it has been clustered together with them. This could be due Germany's goal of reducing carbon output and shifting 60% of electricity production to renewable sources by 2050. Another reason for Germany, UK and Japan having lesser consumption could be that the USA has way bigger houses when compared to all the other countries.



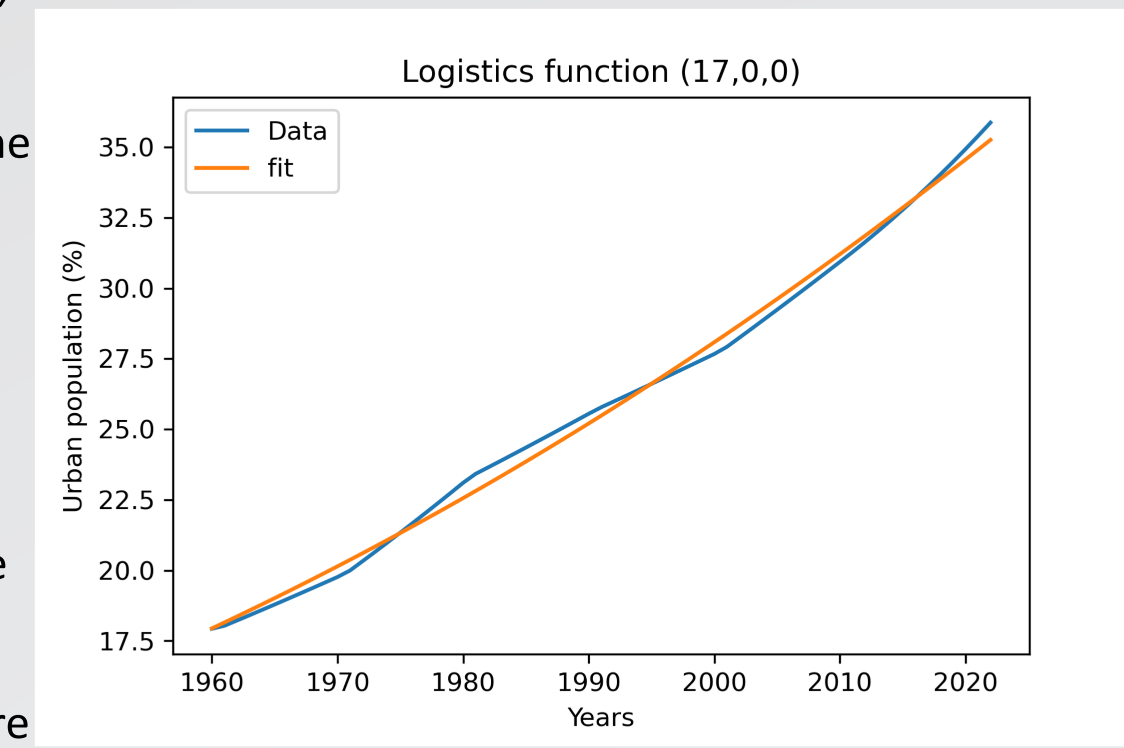
Curve Fitting

The exponential model starts of by slightly overestimating the urban population % and then seems underestimate between 1980 and 1995. Growth rate for 1960 according fit data is 1.08%.



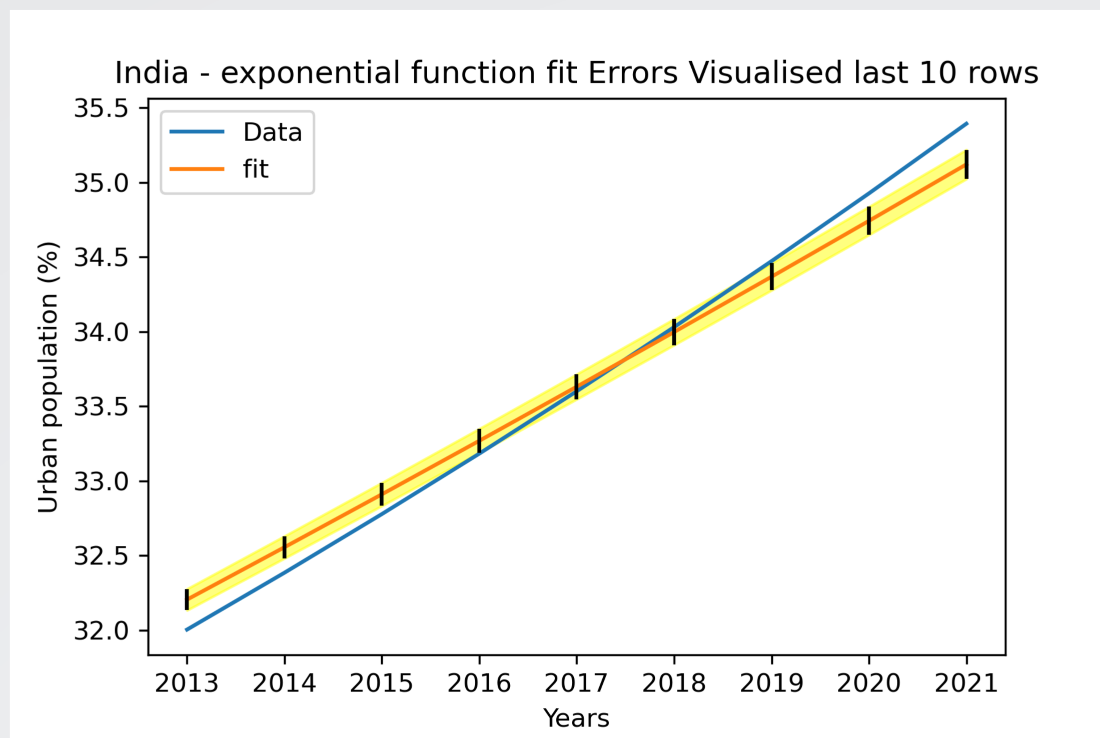
The exponential model was given a guessed start point of (0,0). The fit seems to be good, as it seems to be predicting values close to the original values.
The exponential function used is, f = scale*exp(growth *(t-1960))

The logistical model was given the guessed start points of (17,0,0). The function used for this model is, f = a / (1.0 + exp(-k * (t - t0))).

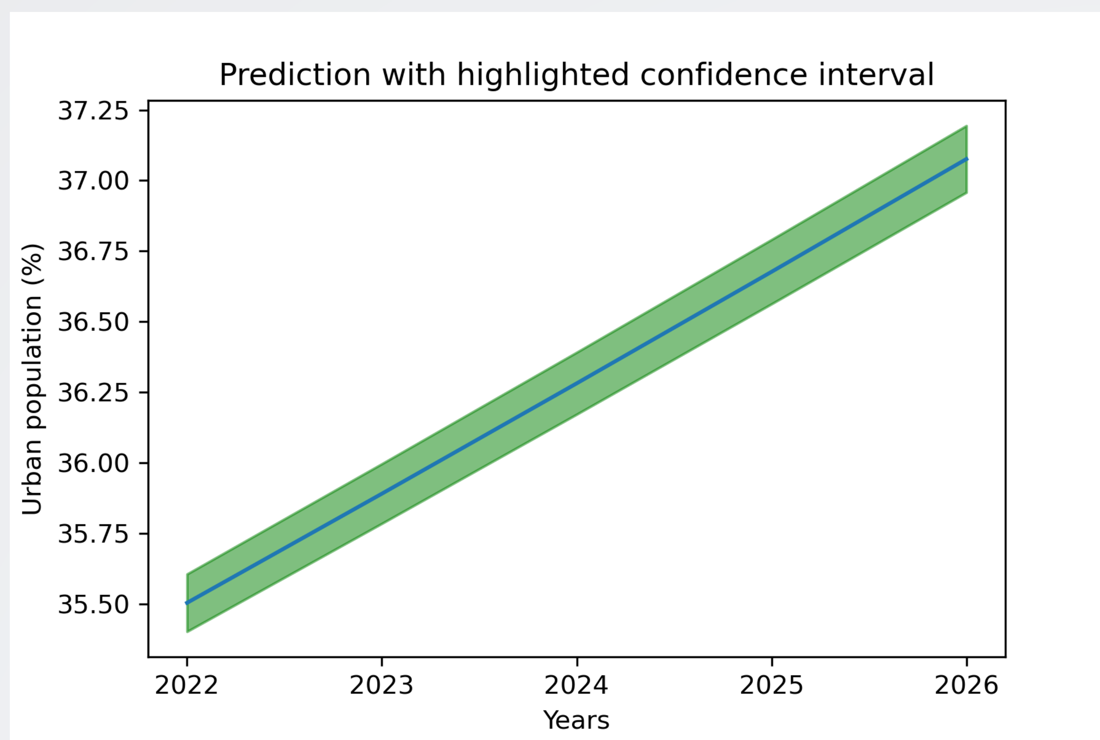


Let us look at the exponential function's predictions.

The quantity of data seems to be too large to be able to see the error range for the fitted data. As a result the last few years were selected in an attempt to be able to 'zoom into' the plot and visualise the confidence range.



For the year 2023 the urban population % according to the exponential fitted data is predicted to be around 35.8901 +/- 0.105. For the year 2026, it is predicted to be 37.0761 +/- 0.118.



Using the Logistical function, the urban population % for 2023 is predicted to be around 35.6119 +/- 0.150383 and for the year 2026, it is predicted to be 36.685 +/- 0.192137. Which is slightly lesser than what the exponential function predicts. But it can be observed that the logistic function seems to be underfitting slightly towards the end,in comparison to the exponential function. Which would explain difference.

Conclusion:

Urban Population could be an indicator for the way a country consumes power and handles CO2 emissions. From the two cluster plots, we can see that developed countries tend to have lower CO2 output (albeit only from solid fuel). But it also indicates that there is an increased demand for electricity per person. Urban population seems to be compensating for this in ways that do not increase CO2 emissions. Maybe by the use of alternative and renewable sources of energy. By predicting the growth of urban population, we could in theory, also guess the demand for non-renewable solid fuels. But more indicators need to be considered before taking a decision.

References used: - https://energytransition.org/2015/03/lower-energy-consumption-in-germany-explained/
Link to code: https://github.com/jerome-p/ads1_assignment3