EPIC-India Data test solutions

All the codes and files have been saved in the following link on Github as instructed https://github.com/hcgautam/AQLI_assignment/tree/main/Output

1.2. **Data Exploration**

1. Abrupt increase in value of average annual GDP was observed before the year 1980 and in the years leading to 2008 economic crash. The increase can be attributed to increased consumption and better living standards before the economic crash of early 1980s and 2008 respectively.

2. Constant as well as current GDP per capita were plotted for the 6 countries in the Middle East and north African region. Libya was found to have the highest GDP. Most countries showed slightly higher constant GDP as compared to current GDP which signifies deflation in economy and reduced commodity prices while in case of Egypt inflation was observed.

3.

i) GDP per capita (constant) on Poverty rate(1day)

GDP per capita (constant) is showing an inverse relationship with Poverty rate(1day) as increase in GDP (if properly distributed among population) will mean better living standard and resources leading to a lower poverty rate

ii) GDP per capita (constant) on Employment ratio

Initially GDP per capita (constant) is showing a positive relationship with Employment ratio as increase in Employment ratio means more people are employed leading to a higher GDP but if we look at the very high Employment ratio data points it represents the population low-income countries where more and more people have to work for long hours just for their own survival. Therefore, high employment ratio but still low GDP (set aside a few outliers and anomalies)

1.3 Estimation and Causal Inference

1. With and adjusted R^2 of 0.3 we can only say that the variables have a low level of correlation for deciding the causal nature we have to take into account all the confounding factors as well as theoretical framework i.e older data trend

2. Coefficient of 0.1 indicates the expected change in the dependent variable for a unit change in the independent variable, holding all other variables constant. In practical terms, if the independent variable increases by one unit, the dependent variable is expected to increase by 0.1 units. Observing the t-statistic, which is calculated as the coefficient divided by standard error (t=0.2/0.01=5). A large absolute value of t-statistic suggests that the coefficient is significantly different from zero, implying a meaningful relationship between the independent and dependent variable.

3. I will design and experiment to introduce nonlinearities in the linear regression model and consider the following points or their combination in the model:
   (i) Use combination of different Polynomial functions on both the parameters like square, cube etc.
   (ii) Exploratory Data Aanalyisis to find anomalies/outliers
   (iii) Introducing functions like log/exponential/square root will help in handling non linearities occurring due to skewed data
   (iv) Use regularization parameters to penalize the magnitude of coefficients
4. The regression code was run and the output saved in respective csv files. String to integer mapping was done to create numerical classes from the income group variable.
5. Fixed-effects regression of constant GDP per capita on employment ratio with country-specific fixed effects was run and the results were saved as a csv file named Fixed effect regression.

2.1. Basic wrangling tasks and questions

1. Number of GADM2 regions in India :685

2. The most polluted countries in the world in 2021 were:
   Bangladesh, India, Nepal, Pakistan, Myanmar, Democratic Republic of Congo, Cameroon, Republic of Congo, Rwanda, and Burundi

3. The most polluted GADM2 level region in 1998, 2005, and 2021 were:
   Unnao, NCT of Delhi and NCT of Delhi

4. Particulate Matter Conc. for Uttar Pradesh was plotted for the years 1998 to 2021 and save d as png. Drastic rise in PM concentration was observed after the year 1999 with a peak in the year 2011and the trend remained stable in the range of 80-95 $\mu g/m^3$.

2.2 Geospatial tasks and questions

1. Top 10 most polluted countries were plotted in dark red. All the countries were either in the Central African Congo Basin Region or Indian Subcontinent which signifies importance of mitigation measures required in Low-income countries of Africa as well as Indo Gangetic plain.

2. Potential gain in life expectancy (relative to the WHO guideline) for western and eastern Europe in AQLI map colour scheme. As part of Chukot region in easternmost Russia was plotted in the left corner of the plot near Alaska and shifting the whole plot towards right, I dropped Chukot level1 region from eastern Europe map so that it can be visualised properly.

3. Whenever I was trying to plotting the PM concentration across the world my laptop was hanging and I am getting errors like insufficient memory. Hence, I was unable to generate the plot but I have written the code and put it in comment block. I even tried to dissolve the AQLI shape file to name0 (national level) nut it didn't work.