Final Assignment: C02 emissions along time

October 2022

- All materials for the exercises below are available in the final assignment folder.
- Please submit an R Markdown file and a knitted pdf document containing your answers and plots (no code.)

CO2 Emissions along time

For the final assignment, we will analyze C02 emissions from 1990 to 2019 and whether it correlates with a a few variables. The World Bank Data Repository contains multiple country-year datasets that can help us do so. All the datasets share the same country names and codes, but not all of them are tidy. For the final assignment, we are providing you with four datasets: (1)`co2_emissions`, (2) `country_region`, (3) `gdp`, and (4) `pop_dat`.

Question 1: tidying and joining data

Using the four datasets available, create a single tidy dataset named `co2_dat` containing the following variables:

Name	Description
country_name‘	Name of the country
country_code‘	Code of the country
year‘	Year of the observation
region‘	Region of the country according to world bank
income_group‘	Income group of the country according to the world bank
co2_emissions‘	CO2 emissions in a given year
total_pop‘	Total population in a given year
urban_pop_share‘	Share of urban population in a given year (in percentage)
gdp‘	Gross domestic product in a given year

Print the first 10 rows of the dataset.

Question 2: inspecting the dataset

a-) Transform `year`,`co2_emissions`, `total_pop`, `urban_pop_share`, `gdp` to numeric class.

b-) How many countries are there?

c-) How many observations are there?

Question 3: missing values

Our dataset still contains missing values.

a-) Venezuela is missing values for `income_group`. According to the World Bank, Venezuela , previously classified as a an upper-middle income country, is now unclassified due to a lack of available data in the recent period. Classify Venezuela as an upper-middle income country.

b-) Drop all observations with missing values in any variable.

Question 4: normalizing our measures

We will need a few new variables for the subsequent analysis, create the following:

`co2_per_cap`, which is the `co2_emissions` divided by `total_pop`.

`gdp_per_cap`, which is the `gdp` divided by `total_pop`.

Question 5: top emitters

a-) Rank countries by their total `co2_emissions`, print the top 20 emitters.

b-) Rank countries by their mean `co2_per_cap`, print the top 20 emitters. Write a brief paragraph commenting how this rank is different than the one from question 5a.

c-) Calculate the mean emissions per capita per country in three different decades: 1990s, 2000s, and 2010s.Did the top three emitters per capita changed in these decades?

Question 6: emissions in time

a-) Create a plot showing the yearly evolution of `co2_per_cap`.

b-) Now create a plot showing the yearly evolution of `co2_per_cap` for the United States, China, and India.

c-) Finally, create a plot showing the yearly evolution of `co2_per_cap` by `income_group`.

d-) In light of the three plots, what can you say about co2 emissions along time?

Question 7: emissions, growth, and population

a-) Create a plot showing the relationship between `co2_emissions` and `population` by `income_group`.

b-) Create a box plot showing the distribution of `co2_per_cap` by `income_group`. Comment the relationship between emissions, growth and population.

c-) Re-create the plot as above, now as a histogram (Tip: use a `geom_histogram()`).

Question 8:urban population and co2 emissions

a-) Are countries with higher share of urban populations, higher emitters on average? Create a plot.

b-) Does the relationship between share of urban population and co2 emissions hold across regions? Create a plot.

Question 9: top emitters, revisited

Create a bar plot that ranks countries by their total co2 emissions(`co2_emissions`) overall. Countries should be colored differently if they are within the top 20 emitters in terms of their co2 emissions per capita (`co2_per_cap`) overall.

Bonus Question 10: Modelling

a-) Run a linear model with `co2_emissions` as your dependent variable. The independent variables should be `population` and `gdp`. Print and interpret the output.

b-) Add year as fixed effect. Print and interpret the output.