CO2 Emissions along time
Final Assignment

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).
- Files should be submitted in moodle, by November 6th, 23:59.

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`.

Disclaimer: the `co2_emissions` dataset only contains data on CO2 emissions (rather than all green house gases), and is not adjusted for import and export of goods, off shoring, and carbon sinks. This does not change anything in what you will do subsequently, but most GHG emissions plots you encounter out there might differ from what you produce here.

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

Remember, each dataset is messy in their own way and requires cleaning before you can join them. . .

Question 2: inspecting the dataset

a-) Transform `year`,`co2_emissions`, `total_pop`, `urban_pop_share`, `gdp` to numeric class. Produce a summary of the dataset.

b-) How many countries are there in the dataset?

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 variables in the dataset:

a-) `co2_per_cap`, which is the `co2_emissions` divided by `total_pop`.

b-) `gdp_per_cap`, which is the `gdp` divided by `total_pop`.

Question 5: top emitters

a-) Rank countries by their total `co2_emissions`. What are the top 20 emitters?

b-) Rank countries by their mean `co2_per_cap`. What are top 20 emitters? How does this change from the rank above (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 change in these decades?

Question 6: emissions in time

a-) Create a plot showing the yearly evolution of `co2_per_cap`. Comment on 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. Comment on the yearly evolution of `co2_per_cap` in these three countries.

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

Question 7: emissions, growth, and population

a-) Create a plot showing the relationship between `co2_emissions` and `population` by `income_group`. Comment the relationship between emissions, population, and income group.

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

c-) Re-create the same plot as above (question 7b), now as a histogram (tip: use a `geom_histogram()` for `co2_per_cap` and facet wrap the histograms by `income_group`).

Question 8:urban population and co2 emissions

a-) Are countries with higher share of urban populations, higher emitters on average? Plot the relationship.

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

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

Run a linear model with `co2_emissions` as your dependent variable. The independent variables should be `population` and `gdp`. Print a summary of the model and interpret the output. What other variables you think should be added to the model?