MM916 Project 1 (2023): CO2 emissions by European cities

Background

Carbon dioxide (CO2) and other greenhouse gas emissions are a global problem that affects everyone, but some people and places are responsible for far more carbon emissions than others. Both the amount of energy we use, and the source of that energy (for example, coal-burning power plants vs. wind turbines and solar panels), affect our per-person share of the problem ("emissions per capita").

The Global Covenant of Mayors for Climate and Energy (GCoM) is a major international initiative that encourages climate action at city level. A team of affiliated reearchers have produced a dataset containing the greenhouse gas emissions of several thousand cities in the EU and neighbouring regions, along with a breakdown of which sector of the economy the emissions come from (residential buildings, transport, etc.) and some additional variables that might explain why some cities have higher emissions per capita than others: for example, one might predict that cities in wealthier countries produce more emissions than others, or that total emissions are driven by how much energy is required for heating and cooling buildings, or that dense cities have more energy-efficient transportation—among many possible hypotheses.

The purpose of an exploratory analysis is to gain an overview of a dataset, including insight into its limitations, and to develop a good set of hypotheses to test carefully with statistical modelling. This is what you will do for the GCoM emissions data in this project.

To get you started, we downloaded the dataset from the published source (https://essd.copernicus.org/articles/13/3551/2021/) and converted two sheets of the original Excel file into .csv files. We also provided a short script (`read_GCoM_data.R`) that reads them into R, keeps only the variables you will need for this project, and renames them for convenience. Use the published article or general web searches for further background on definitions and units.

Instructions

Write the Results section of a technical report that describes an exploratory analysis of the GCoM emissions dataset above. Make a single Word or PDF document containing figures, tables, and the brief text that accompanies then, with the R code that produced them included at the end as an Appendix.

Include the 8 elements listed on the next page (You can include other analysis if you really want to, but additional material will not normally affect your grade.)

- A brief summary of results from each element should be included in the main text.
- Figures should have captions.
- Tables should have descriptive titles, and be formatted professionally: do not simply copy and paste raw output from R.

It is not necessary to run any statistical tests of your interpretations of the data. This project is about *exploratory* data analysis and so guesses based on what you see by eye are sufficient (this is how statistical hypotheses are generated, not tested).

You can make use of everything on MyPlace and also general web searches, but you may not work in groups or copy from a classmate: these are individual projects and everything below needs to be your own work.

Marking

For each element, * 3 points for correct calculations and basic figure production in R * 1 point for professionally correct presentation (axis labels, NA columns removed from categorical variables, caption, etc.) * 2 points for clear and sensible written interpretation. A final 2 points for clarity and organisation of the document as a whole. 50 points total.

We will give partial credit for figures and calculations that are useful but not exactly what we asked for. However, we will not run your code for you to determine if it produces the expected output: figures and other results must be included in the document you submit.

What to include

Element 1: A dataset overview. Filter out rows in the `GCoM_emissions.csv` dataset where either the emissions per capita or population is missing. Give the number of cities and countries represented in the data, and the names of the countries with the most and fewest cities. (This information can be presented in a table or simply in sentences, as you prefer.) Do there seem to be any imbalances: are certain countries are overrepresented?

Element 2: Range and distribution of city populations. Make a histogram of population. What are the maximum and minimum city populations in the dataset (and the corresponding city names), and the median?

Element 3: Emissions by country. Make a boxplot or similar plot that shows emissions per capita for each country. For full marks, display the countries in a sensible order. Report the top 3 and bottom 3 countries by median emissions per capita (get R to identify these countries for you).

Element 4: Emissions by sector. Using the data in `GCoM_emissions_by_sector.csv`, make a plot of the *total* emissions for each of the six sectors (residential buildings, etc.). Which of the sectors are responsible for the most emissions?

Element 5: Emissions by sector and country. Now join the two datasets (using city ID) so that you can associate city names and countries with the by-sector data. Make a plot showing how the relative importance of the six sectors varies by country. (One nice way to do this is a stacked bar plot giving the fraction of each country's emissions in each sector, but there are many ways to approach it.)

Element 6: Connecting emissions to heating demand. One might hypothesize that emissions are higher in colder cities where more energy is required for heating. Make a scatter plot that helps evaluate the possible link between Heating Degree Days and emissions per capita. Use either total emissions by city, or emissions in one of the six sectors that seems appropriate. Highlight the Scandanavian cities (Sweden, Norway, Finland, Denmark) in a second colour: one might expect these countries to have especially high heating needs and therefore especially high emissions. Based on visual inspection of the plot, does this hypothesis seem likely to be true?

Element 7: Connecting emissions to wealth. An alternate hypothesis is that wealthier countries use more energy and therefore produce more CO2 emissions. Make a scatterplot that helps you examine the relationship between emissions per capita and GDP per capita, removing one outlier city with very high GDP per capita (name which city it is). Use colour and symbol type to include other variables if it helps you evaluate further hypotheses. What does this plot tell you?

Element 8: Summary and recommendations. Conclude with a short paragraph summarising what you have learned from this exploratory analysis, and what hypotheses appear to be most promising for future analysis. If you feel other variables would need to be included in the analysis to find an explanation of why some cities or countries have especially high or low emissions, speculate on what data would be useful. (Note that in a real published report, this summary paragraph would probably go in your Discussion section, but for this project you don't need to break your writing into sections.)