Amariah Robinson
02/17/2026
DTSC 2301-001
Dr. Blekking and Dr. Benedict

Project 1 Portfolio

The dataset that I will be using to answer this question came from Kaggle. It contains global data such as GDP, life expectancy, education information, and health information, to say the least. Each row represents one of the 204 countries represented in this dataset. Some key features include name, pop_density, co2_emissions, urban_population, urban_population_growth, and forested_area. The dataset is about 204 rows with 38 columns, so it contains over 7700 cells of information total.

My research question is: Are we able to see if there is a possible correlation between the population of people living in urban areas vs the amount of CO2 exposure there is? This question is interesting because it helps to understand the reasoning behind maybe someone in an urban area and newer development wanting to migrate somewhere outside of that area, maybe more rural. In more densely populated areas, vehicle exhaust is a common reason for CO2 levels to be high and rise. Having a densely populated and built area can prevent safe ventilation systems to work and prevent polluting the areas and giving carbon monoxide poisoning. Someone who might care about my question and answer would be urban developers and people looking to relocate into urban areas might want to reconsider their options.

Some things that might be missing in the dataset could be the fact that the amount for the urban_population feature is not labeled at all. It could be millions, thousands, billions, etc. Some assumptions that are being made is that the data is actually accurate. That each country is reporting their GDP and information correctly, and that there has been no data poisoning occurring. Another assumption that I had to make was that the urban population feature is in the millions. This makes the most sense as each region listed has a population over 1 million. But this is also not a safe assumption, and could result in having bad data overall that cannot be used.

Some cleaning steps that I decided to take were to even see if the null values or NaN values were able to be taken out or if they should be replaced by a 0 value. I decided against this decision because it would affect the GDP statistics for the countries (observations) used, resulting in bad data. I also decided to get rid of any countries that have not reported their co2 emissions, as they are not going to be able to tell me if there is a correlation between the amount of CO2 emissions and their urban population. I decided against imputing the data with the median amount of CO2 emissions, as that would range due to regional differences. This could incorrectly report ranges that do not quite fit with their region, especially if outliers are present (which they are) for more developed countries.

For my visualizations, I decided to create a scatterplot of the regions and their CO2 Emissions to showcase any outliers (there were about 5 for the most part that stuck out more than others) and just to see the spread of the amount of CO2 emissions across the different regions. I also created a bar plot of the regions and their to get a different perspective of the outliers that were shown before in the scatterplot, but this shows that there aren't too many outliers as much as the scatterplot did. From here you can see that Eastern Asia has the biggest amount of CO2 emissions from their region. Next is a scatterplot of the regions and their urban population growths. This shows that for each region their countries have different growth rates and some even have a negative growth rate, meaning that people are actually leaving the country. For the next plot, we have the bar chart as well, but for each country. There were about 145 countries found in the dataset to not have any null values for CO2 emissions, and were used to see their values on CO2 emissions and urban population. As you can see, China, which is located in Eastern Asia, has the biggest amount of CO2 emissions. This aligns with the scatterplot from before, showcasing that Eastern Asia has the biggest amount of CO2 emissions from their region. So China is the country that is skewing the rest of the datasets' mean CO2 amounts.

There could be some misleading information here as well such as thinking that just because it is the Eastern Asia region that has the highest amount of CO2 emissions, that it means that all of the countries inside of Eastern Asia contributed to that high amount. Until you dig further into the data, you will realize that different countries have different amounts of CO2 emissions, and that they will not always contribute the same amount. This is shown with the bar graph that represents the 145 countries, vs the bar graph that only shows the regions.

My data does not capture the true values of what the CO2 emissions represent, along with the urban population and urban population growth rates and amounts. There is no clear definition of what these values represent. The user of the dataset is left guessing or assuming that these values are what they need them to be, leading to potential bias and uncertainty. These assumptions of urban population being in the millions and not a different value can lead to misrepresentation of the regions and countries as a whole. This can misrepresent and mislead the public into thinking that their country may be doing better or worse than the others, based on this biased thinking. I think that I was more surprised that there wasn't a clear correlation between CO2 emissions and their urban populations, especially in the outliers. This goes to show that assumptions should be left out when it comes to data analysis and shows how it can affect your analysis as a whole as well.

References

- Emojis used in my code: https://www.w3schools.com/charsets/ref_emoji_smileys.asp
- Dataset: https://www.kaggle.com/datasets/arslaan5/global-data-gdp-life-expectancy-and-more
- Looking at the CO2 emissions per capita: https://ourworldindata.org/grapher/co-emissions-per-capita
- Seaborn Documentation for a scatterplothttps://seaborn.pydata.org/generated/seaborn.scatterplot.html
- isNull function for pythonhttps://pandas.pydata.org/docs/reference/api/pandas.isnull.html
- Figure sizing for matplots: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.figure.html
- Use of Gemini through Google search: https://www.google.com/search?q=how+to+drop+specific+columns+with+rows+that+are+null+in+pandas&sca_esv=7d344b0fa4f660b1&rlz=1C1UEAD_enUS1028US1029&sxsrf=ANbL-n6IShnX9BgRpx1DMiupDiRtPdG24A%3A1771864758118&ei=toKcaargBv3XwN4PktnA0Ak&biw=1536&bih=730&ved=0ahUKEwiqk_3LhvCSAxX9K9AFHZIsEJoQ4dUDCBE&uact=5&oq=how+to+drop+specific+columns+with+rows+that+are+null+in+pandas&gs_lp=Egxnd3Mtd2l6LXNlcnAiPmhvdyB0byBkcm9wIHNwZWNpZmljIGNvbHVtbnMgd2l0aCByb3dzIHRoYXQgYXJlIG51bGwgaW4gcGFuZGFzSPAbUIYDWJgacAJ4AZABAJgBkAGgAbUTqgEENy4xNrgBA8gBAPgBAZgCC6ACnAnCAgoQABiwAxjWBBhHwgIKECEYoAEYwwQYCsICCBAAGKIEGIkFwgIIEAAYgAQYogTCAgUQABjvBcICBBAhGAqYAwCIBgGQBgiSBwMyLjmgB75nsgcDMS45uAeMCcIHBzAuMy43LjHIBzKACAA&sclient=gws-wiz-serp
- Use of Gemini through Google search: https://www.google.com/search?q=how+to+change+the+values+of+the+y+axis+ticks+in+python&sca_esv=7d344b0fa4f660b1&rlz=1C1UEAD_enUS1028US1029&biw=1536&bih=730&sxsrf=ANbL-n6iCvZr1U2ARYZpS5wQHYjB5QDNMA%3A1771865320045&ei=6IScadS3ArTDp84PyJ3P8QY&ved=0ahUKEwjUy_bXiPCSAxW04ckDHcjOM24Q4dUDCBE&uact=5&oq=how+to+change+the+values+of+the+y+axis+ticks+in+python&gs_lp=Egxnd3Mtd2l6LXNlcnAiNmhvdyB0byBjaGFuZ2UgdGhlIHZhbHVlcyBvZiB0aGUgeSBheGlzIHRpY2tzIGluIHB5dGhvbjIIEAAYogQYiQUyCBAAGIAEGKIEMggQABiiBBiJBTIIEAAYogQYiQUyCBAAGKIEGIkFSPEkULkNWJojcAd4AZABAJgBiQGgAa0MqgEEMy4xMbgBA8gBAPgBAZgCEqAC6ArCAgoQABiwAxjWBBhHwgIKECEYoAEYwwQYCsICBBAhGArCAggQIRigARjDBJgDAIgGAZAGCJIHBDEwLjigB8JMsgcDMy44uAelCsIHCDAuNC4xMi4yyAdSgAgA&sclient=gws-wiz-serp
- Trying to get a subset of the original dataframe: https://pandas.pydata.org/docs/dev/getting_started/intro_tutorials/03_subset_data.html#:~:text=When%20specifically%20interested%20in%20certain,of%20the%20selection%20brackets%20%5B%5D%20.
- Example of the matplotlib.pyplot bar chart: https://www.google.com/search?q=plt+bar&rlz=1C1UEAD_enUS1028US1029&oq=plt+b

ar+&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIHCAEQIRiPAjIHCAIQIRiPAtIBCDE5NzdqMGo3qAIAsAIA&sourceid=chrome&ie=UTF-8

- Matplotlib.pyplot documentation for a bar plot: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.bar.html
- More matplotlib documentation for a bar chart: https://www.geeksforgeeks.org/pandas/bar-plot-in-matplotlib/
- How to make the xlabel and ylabels bigger in python (gen AI answer): https://www.google.com/search?sca_esv=48cff7b3afdc9477&rlz=1C1UEAD_enUS1028US1029&sxsrf=ANbL-n7FCIGkT9pHRs7qmktpLPsUBedR5g:1771886964626&q=how+to+make+the+label+bigger+in+python&spell=1&sa=X&ved=2ahUKEwjI5Oqo2fCSAxX-48kDHSg6GWwQBSgAegQIERAB&biw=1536&bih=730&dpr=1.25
- How to increase the size of the datapoints on a scatter plot (Gen AI result from google search): https://www.google.com/search?q=how+to+make+datapoints+larger+in+python&sca_esv=48cff7b3afdc9477&rlz=1C1UEAD_enUS1028US1029&sxsrf=ANbL-n6YF_1wjOYwqlLXdcX26BQiIX6I2A%3A1771887402583&ei=KtucaaWjI7HiwN4PtPDD8Qg&biw=1536&bih=730&ved=0ahUKEwil4Nn52vCSAxUxMdAFHTT4MI4Q4dUDCBE&uact=5&oq=how+to+make+datapoints+larger+in+python&gs_lp=Egxnd3Mtd2l6LXNlcnAiJ2hvdyB0byBtYWtlIGRhdGFwb2ludHMgbGFyZ2VyIGluIHB5dGhvbjIHECEYoAEYCjIHECEYoAEYCjIHECEYoAEYCjIHECEYoAEYCjIHECEYoAEYCjIFECEYnwUyBRAhGJ8FSMoLUD1YlgpwAXgBkAEAmAGCAaAB0weqAQM2LjS4AQPIAQD4AQGYAgugAo8IwgIKEAAYsAMY1gQYR8ICBhAAGBYYHsICCxAAGIAEGIYDGIoFwgIIEAAYgAQYogTCAgUQABjvBcICBRAhGKsCmAMA4gMFEgExIECIBgGQBgiSBwM2LjWgB_VOsgcDNS41uAeGCMIHBzAuNy4zLjHIByaACAA&sclient=gws-wiz-serp
- How to find the statistics of a dataset in python (google search, Gen AI response):https://www.google.com/search?q=how+to+find+the+statistics+of+a+dataset+in+python&rlz=1C1UEAD_enUS1028US1029&oq=how+to+find+the+statistics+of+a+dataset+in+python&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIICAEQABgWGB4yCAgCEAAYFhgeMggIAxAAGBYYHjIICAQQABgWGB4yCAgFEAAYFhgeMggIBhAAGBYYHjIICAcQABgWGB4yDQgIEAAYhgMYgAQYigUyDQgJEAAYhgMYgAQYigXSAQg5NDYzajBqN6gCCLACAfEFzuu0Wvz6zGI&sourceid=chrome&ie=UTF-8
- How to capture certain rows and columns in python (google search, Gen AI result): https://www.google.com/search?q=how+to+see+certain+rows+from+a+python+dataframe&rlz=1C1UEAD_enUS1028US1029&oq=how+to+see+certain+rows+from+a+python+dataframe&gs_lcrp=EgZjaHJvbWUyCQgAEEUYORigATIHCAEQIRigATIHCAIQIRigATIHCAMQIRigATIHCAQQIRigATIHCAUQIRiPAtIBCTEwNzkzajBqN6gCALACAA&sourceid=chrome&ie=UTF-8
- How to change a value in a column (google search, Gen AI result): https://www.google.com/search?q=how+to+see+certain+rows+from+a+python+dataframe&rlz=1C1UEAD_enUS1028US1029&oq=how+to+see+certain+rows+from+a+python+dataframe&gs_lcrp=EgZjaHJvbWUyCQgAEEUYORigATIHCAEQIRigATIHCAIQIRigATIHCAMQIRigATIHCAQQIRigATIHCAUQIRiPAtIBCTEwNzkzajBqN6gCALACAA&sourceid=chrome&ie=UTF-8

- The use of Gemini was used as the resulting agent for most of my google searches! The AI Overview was really helpful and quick in giving me my answers, and it wasn't necessary to "Dive deeper in AI". So I'm honestly not quite sure what AI agent was used with the specific versions, but hopefully the links above work.