Gabriella Weis, Mason Polier

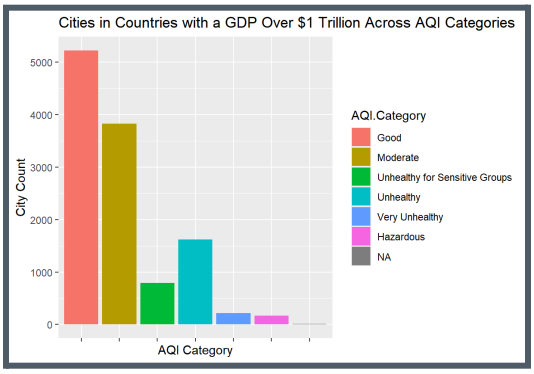
DATA-151

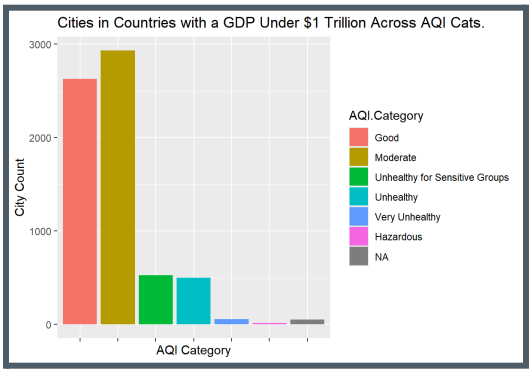
6 December 2024

Pollution and Prosperity: An Essay on Air Quality and Economics

This project was focused on the link between economics and air quality. It involved two datasets, one based around GDP from the world bank with 222 country-based observations, and another on indicators of air quality from Kaggle with 23,463 city-based observations (World Development Indicators | The World Bank, Global Air Pollution Dataset). The variables of interest were AQI Category, or what category a city's air quality rating falls under (e.g. good, moderate, unhealthy, etc.), AQI Value, or a number ranging from 0-500 (0 being the best and 500 being the worst), Country GDP for the year 2023 (in billions USD), and Country Agriculture, Industry, Manufacturing, and Services. These last four variables denote what percentage of its GDP a country allocates toward a specific sector of their economy.

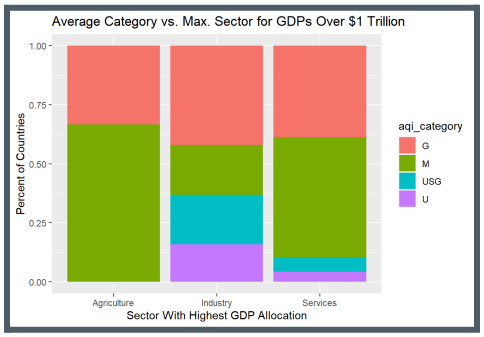
To begin, it was important to know if GDP affected the distribution of cities' AQI categories. In other words, what is the distribution of cities across AQI categories in countries with a GDP over $1 trillion? How does this compare with countries with a GDP under $1 trillion? These distributions are pictured in the bar graphs below.





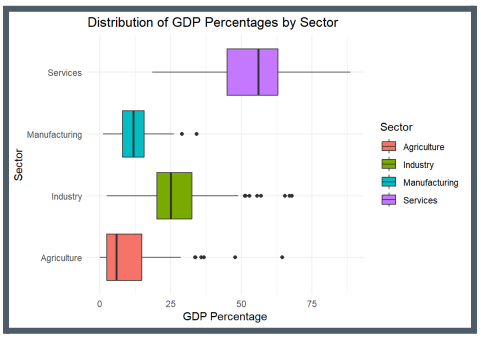
In the bar graph on the left, for cities in countries with a GDP over $1 trillion, the majority of cities earn Good AQI Category ratings. On the contrary, for under $1 trillion, the majority of cities earn Moderate AQI Category ratings. Additionally, the percentages of cities with Hazardous and Unhealthy ratings appear to be lower in countries with a GDP under $1 trillion than those in countries with a GDP over $1 trillion.

The exploration moved on to sector allocation, specifically countries' most funded sectors; how does the average AQI category vary among countries based on their primary GDP sector allocation? Answering this meant creating a stacked bar graph.



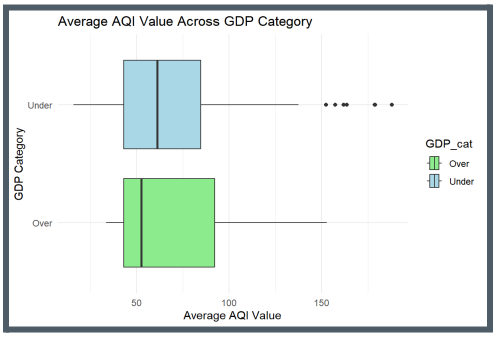
For the above graph, two new categorical variables were created. One was called "Max Sector" and referenced a country's most funded sector, and the other, "Average AQI Category," represented the average AQI category of a country instead of just its cities. The graph shows the layout of average AQI category ratings based on countries' maximum GDP sector allocations (with NAs removed). It appears that over 50% of countries that allocate most of their GDP to agriculture have an average AQI rating of Moderate, and that compared to other countries with other maximum sectors, countries that allocate most of their GDP to Industry have more Unhealthy ratings.

Continuing with sectors and taking a closer look at the percentage of country GDP attributed to each, how does GDP percentage vary across sectors? Below is a graph of comparative boxplots with the sectors plotted against each other.



From the boxplots, the Services sector appears to have the greatest percentage of GDP allocation among countries, with Industry coming in second. Services also has the largest Interquartile Range (the largest spread within the middle 50% of its data).

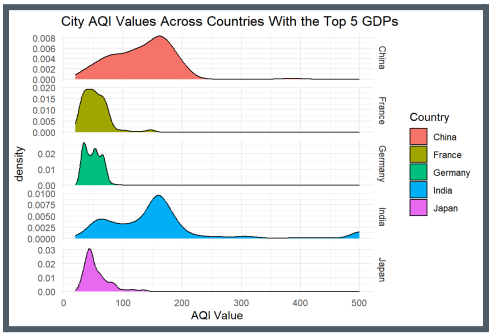
The next question embodies the entire project because it strives to discover a numerical relationship between two of the main variables of interest: is there a correlation between overall GDP and a country's average AQI value? Once again, this involved the construction of comparative boxplots.



To make these horizontal side by side boxplots, a new categorical variable called "GDP_cat" and a new numerical variable called "Average AQI Value" were created. "GDP_cat" has two levels: under (country's GDP is under $1 trillion) and over (country's

GDP is over $1 trillion). “Average AQI Value” holds the average AQI value per country. These plots seem to suggest that countries with higher GDPs tend to have better air quality on average. However, wealthier countries also show more variability, and there are probably additional variables and factors besides GDP affecting these distributions.

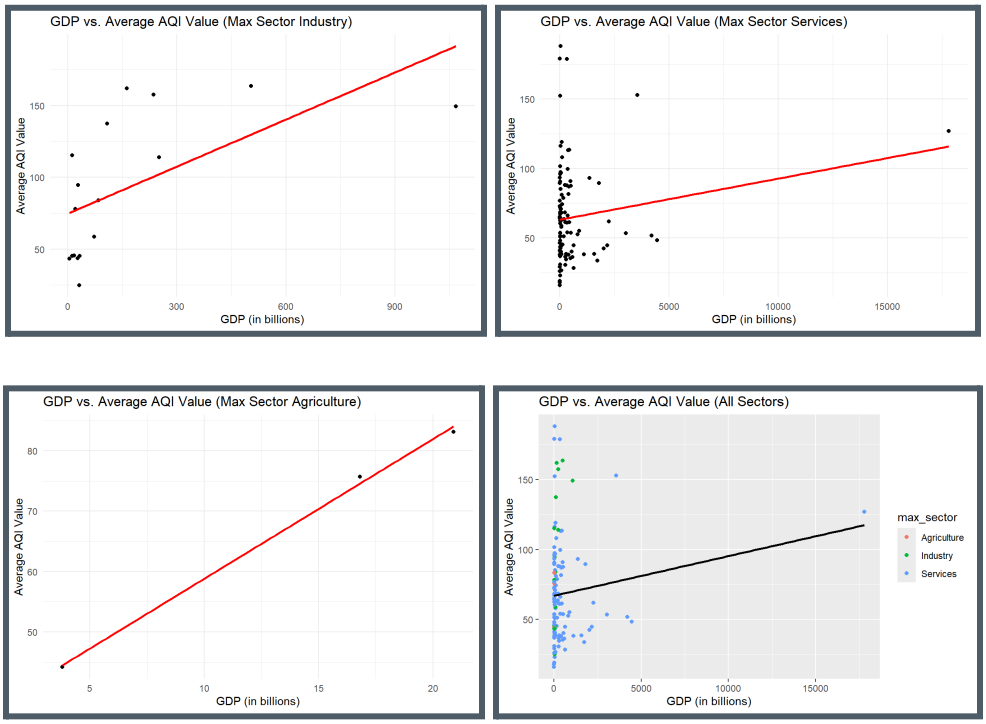
Focusing more on significant countries in terms of GDP, a new question was crafted. How do the countries with the top five GDPs compare in terms of air quality? Below are facet gridded density plots of these countries.



After filtering out the countries with the top five GDPs, it was easy to visualize the distributions of city AQI across countries and compare them (notice the absence of the US here; the data lacked all air quality information for the US). Take Japan and India for example. India has extreme outliers on the higher end of the spectrum, meaning it has a large skew. Its range and IQR are massive. Japan, on the other hand, has a very compact distribution and one distinct peak.

The final question returns to max sector allocation, focusing this time on industry. Industry is important because in the stacked bar graph from earlier, countries that allocate more of their GDP to it than any other sector tend to earn the most Average Country AQI Category ratings of Unhealthy. Therefore, the question was: how do average AQI values differ between countries that allocate the maximum percentage of

their GDP to industry versus those that allocate the maximum to another sector? This led to the creation of four graphs, three filtered by a country's max sector (no countries had a max sector of Manufacturing) and one of every sector plotted together.



Focusing on the Industry sector, the graph and its correlation coefficient of 0.6016 indicate a moderately strong, positive, linear correlation with a possible outlier on the x-axis. The Services scatter plot appears to have very little correlation, and although the Agriculture graph has a correlation coefficient of 0.9987, there are only three data points, so it is hard to make a definitive conclusion. Furthermore, these graphs seem to exhibit an example of Simpson's Paradox; when the sectors are separated, there seems to be some correlation, but when combined, there is very little.

Using these questions, graphs, and summary statistics, it is possible to formulate a series of small conclusions. However, it is difficult to draw sweeping ones that encompass the entire dataset, or to say definitively that a country's GDP has an effect

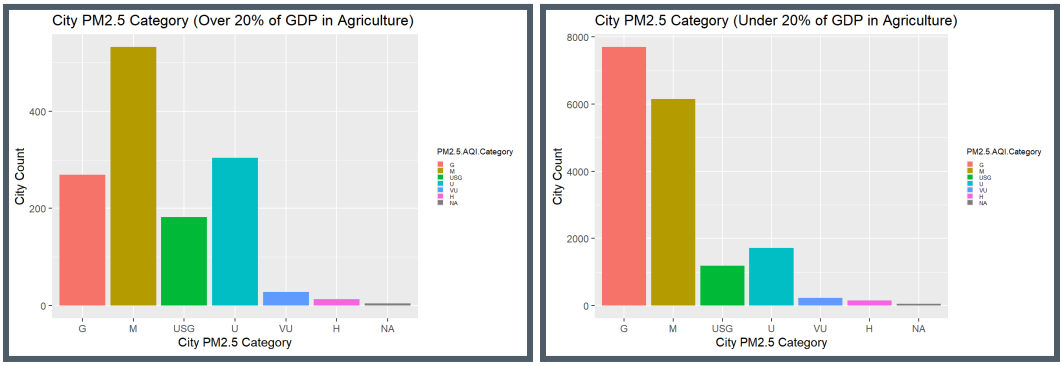
on air quality. In the future, it would be very interesting to explore the rate of increasing GDP with AQI. This would have also been fairly feasible with the data; the GDP dataset has data from both the year 2023 and 2015. Lastly, it would be beneficial to use more easily manipulatable datasets, and to include a wider range of variables in the analyses.

Appendix:

Extra Question:

There existed another question that was answered and analyzed, but it was too much to include in the core of the paper and presentation. It is explained in detail in this appendix, along with an array of other unused graphs and analyses.

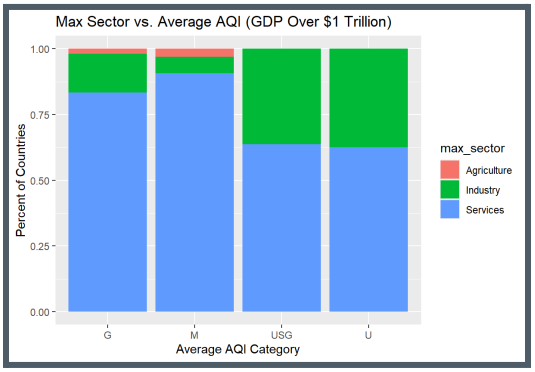
The question was: how does PM2.5 (a size of particulate matter) AQI category differ between cities in countries that allocate a higher percentage of their GDP to agriculture (over 20%) versus those that allocate a lower percentage? This meant creating two more bar graphs.



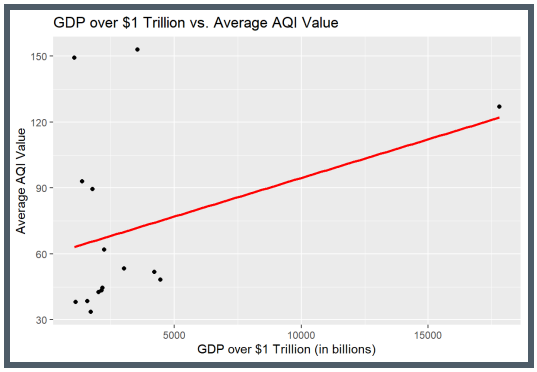
The creation of these graphs involved filtering for countries with an agricultural allocation over and under 20%. The bar graph on the left represents the distribution of cities among PM2.5 Categories for countries with over 20% of their GDP allocated to agriculture while the bar graph on the right represents the category distribution for countries with under 20% allocation. There seems to be (most notably) a higher frequency of Good ratings and a lower frequency of Unhealthy ratings in the right bar graph than its counterpart, but it also contains a far higher number of cities. Percentages were used to get a clearer view because of the count difference between

the graphs. Cities with over 20% GDP allocation receive a Good rating 20.2% of the time and a Moderate rating 39.94% of the time. Cities with under 20% allocation receive a Good rating 44.85% of the time and a Moderate rating 35.82% of the time. Therefore, cities with over 20% of their country's GDP allocated to agriculture tend to have a higher chance of being in the Moderate AQI category and a lower chance of being in the Good AQI category, and cities with under 20% of their country's GDP allocated to agriculture tend to have the opposite chances.

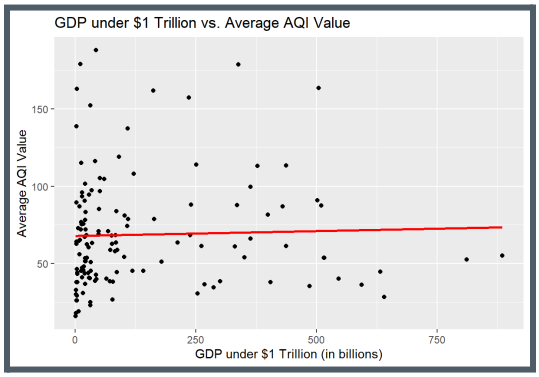
Additional graphs and their descriptions:



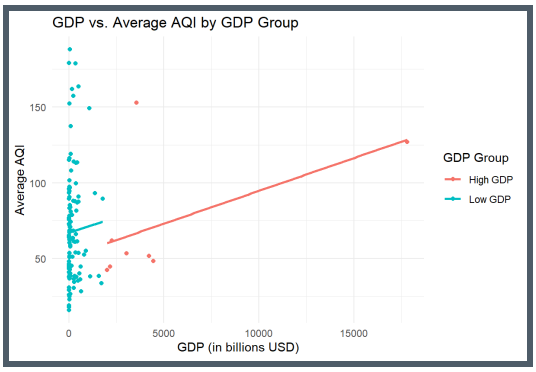
This stacked bar graph shows the layout of maximum GDP sector allocations based on countries' average AQI ratings. It seems to be a less meaningful graph than the stacked bar graph depicting category by sector (since most of the max sectors are Services), but it is still possible to glean some things. For example, it seems that countries with average AQI ratings of Unhealthy and Unhealthy for Sensitive Groups also appear at a higher percentage with max sectors in industry.



The R value of 0.352 and the graph indicate a weak, positive, linear correlation with one major outlier on the x-axis. The slope of the regression line is 0.003522 and the y-intercept is 59.369121. This means that for every one-unit increase in GDP for countries with GDP over $1 trillion, the predicted average AQI value increases by 0.003522 units.



The R value of 0.0326 and the graph indicate a very weak, positive, linear correlation with multiple possible outliers on the x and y-axis. The slope of the regression line is 0.006256 and the y-intercept is 67.911430. This means that for every one-unit increase in GDP for countries with GDP under $1 trillion, the predicted average AQI value increases by 0.006256 units.



For this scatter plot, a new categorical variable was created for GDP grouping called “GDP_Group,” where GDP is sorted into either above 2000 billion (High GDP) or under 2000 billion (Low GDP). Separate models were also fitted for each GDP group.

For the High GDP model:

The R squared value of 0.29 suggests a weak, positive, linear correlation. The slope of the regression line is 0.004332 and the y-intercept is 51.404082. This means that for every one-unit increase in GDP, the predicted average AQI value increases by 0.004332 units.

For the Low GDP model:

The R squared value of 0.0014 suggests a very weak, positive, linear correlation. The slope of the regression line is 0.003966 and the y-intercept is 67.240328. This means that for every one-unit increase in GDP, the predicted average AQI value increases by 0.003966 units.