An Exploration of the Association Between Political Affiliation and AQI Levels

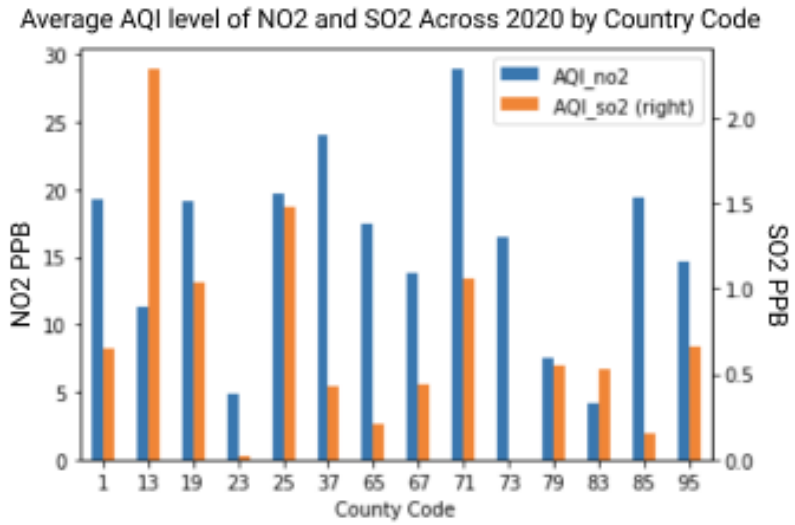
Isabel Zheng, Ayush Misra, AJ Grover, Krish Rambhiya

Open-Ended EDA

[Repeated Guided EDA Analysis on New Data]

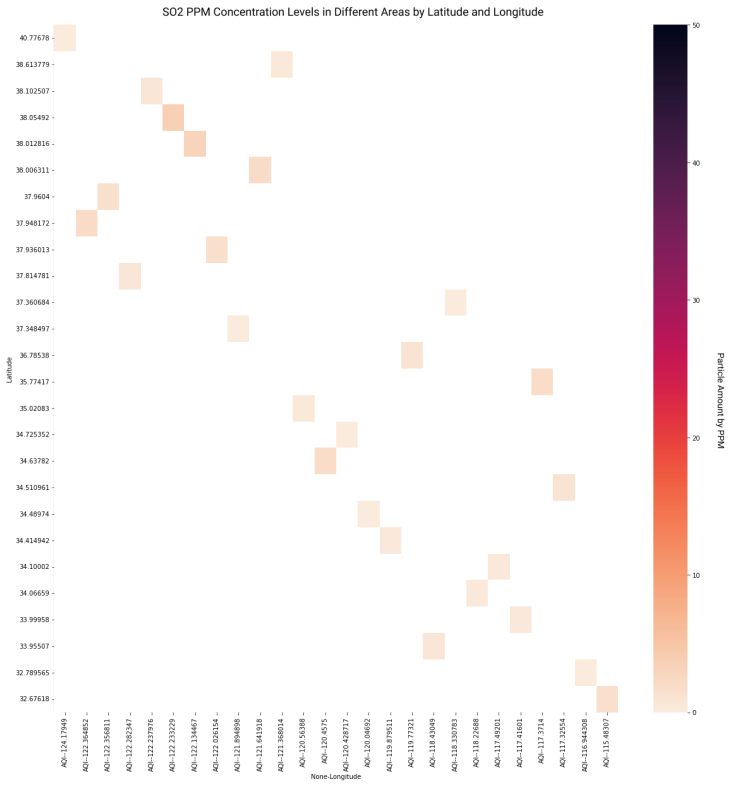
We repeated a guided EDA analysis on our new data focussing on the particle matters of AQI specifically NO2 and SO2. Our first step was cleaning the data, which required converting the datetime column into distinct columns for each day and month respectively. This allowed us to pick out certain years that we wanted to analyze. After that we created a pivot table of latitude vs. longitude and found the average AQI for each combination. This gave us a general sense of the AQI levels depending on geographic location. Since we focused on particle matter and AQI calculation methods, we joined a table of the average AQI over the year for no2 and so2 through the country code. After all of this cleaning and analysis, we were able to start visualizing the data. Our visualizations consisted of a few plots, 2 of which will be addressed in detail.

This graph showcases the average levels of the two pollutants, NO2 and SO2, in 2020 in certain countries. Despite the large variance in bar sizes, taking special note of the axes will reveal a different scale for SO2 and NO2. It ultimately shows that the average concentration of NO2 throughout a year is overall way higher than SO2 in most countries. This begs us to question why certain pollutants may over take the proportion of air pollutants and therefore dominate the calculation of AQI. This leads us to interesting possibilities such as certain powerful and common industries producing a large amount of a certain pollutant. We also wonder if factors such as differences in political affiliation and differences in geographical location may play a role in the varying concentration levels of any



particle in general. This graph greatly influenced the direction of our project as we later dove deeper into the potential causes in AQI concentrations between countries.

To the left, the two plots showcase the relationship between the average NO2 level and SO2 level in 2020 based on longitude and latitude. The heat map effectively visualizes the distribution of this spread as we can clearly see partners between geographical location and pollutant level through colors and map location. As one can see, there is a small clump of higher concentration of NO2 levels in about -117 to -119 degrees of longitude and 33 to 35 degrees of latitude. Meanwhile, SO2 levels are overall a lot lower in all regions. This is quite interesting as we initially expected certain regions to have higher concentrations of NO2 and some with higher concentrations of SO2. This heat map also begs us to question to relevance and impact of SO2 overall on AQI calculations. Looking at our heatmaps currently, it seems as if SO2 is never utilized to calculate AQI as it's concentration levels are consistently being overshadowed by NO2. Evidently, the spreads for NO2 and SO2 are distinctly unique and both play a role in AQI calculation regardless, but NO2 seems much more prevalent overall. The heatmap tells us that based on the darker areas of the spectrum, there exist hotspots within the data. In the few clumps and regions where there is heavy concentration, AQI differs heavily than other regions that may not be part of those hotspots.



After analyzing our new data, we were able to draw multiple interesting conclusions and insights. We were able to conclude with additional insight from the heat map longitude/latitude based EDA graph that -117 to -119 longitude and 33 to 35 were especially prevalent for both particles than in other regions. This means that in this region, both of these particles have a stronger presence than outside of this region. This was an interesting insight as it led us to consider the fact that increasing AQI is likely due to increases in all pollution particles and matter and not simply one. With our insights from the bar graph, we suspected that AQI may increase in all pollution particles and matters in a more proportional sense. Because the variance of SO2 was quite small (between 0-2) in comparison to NO2, we considered that NO2 may increase at a rate faster than SO2. This prompts some open ended questions: is this because different strong pollutant generators may generate a generally proportional level of pollutants? Are these regions hotter, thus increasing the amount of temperature control machinery used and therefore increasing pollution in the form of these particulates? Might political affiliation and environmental policies play a large role in the differing concentrations of AQI?

Problem

AQI generally is measured by the concentration of a particular pollutant in a particular area (namely cities). Based on this definition of AQI, we are curious about the AQI levels in different countries and how it's levels may be influenced by different political standpoints. We hypothesize that there will be a quantifiable relationship between political standpoint and AQI levels and AQI measurement method. We hypothesize that countries leaning left will have an average annual AQI level lower than the average annual AQI levels of those leaning right. Furthermore, we believe we can predict to 75% accuracy what a country's yearly average AQI is given the country's current political standpoint. We define left leaning and right leaning through our personal judgement of answers in a pew research report. For example, those who answered with approval to direct democracy would be more likely to lean left. We quantified left and right leaning through an aggregation of answers to these questions.

Answer

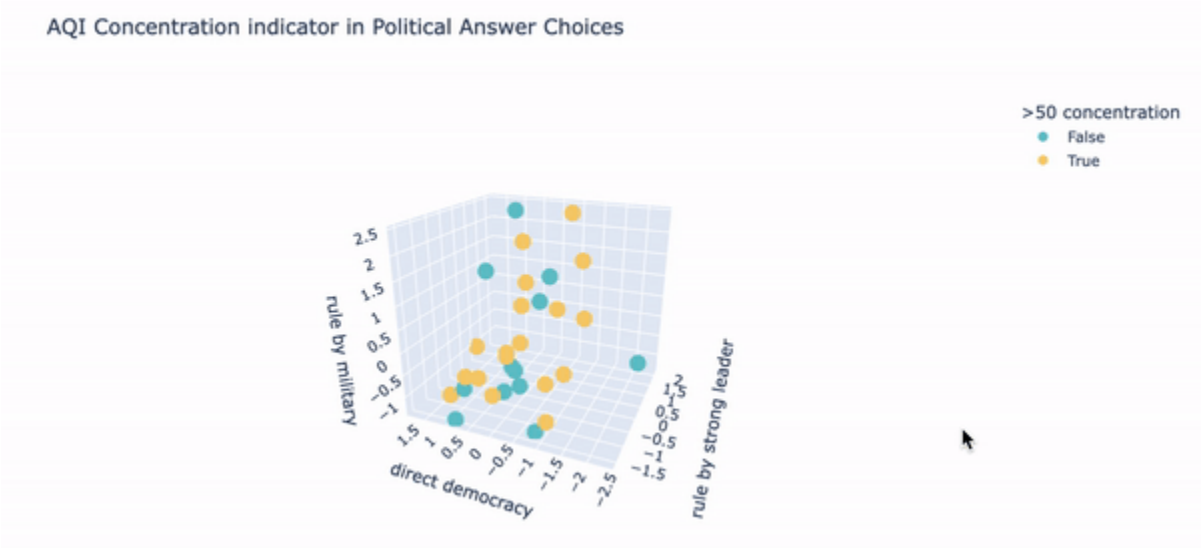
After conducting our research, data cleaning, and data analysis, we concluded that our hypothesis is strongly rejected by the data. Further details in how we went about this process will be touched upon later in this report. The average R^2 coefficient from our data analysis was found to be ~-1.8 which indicates that there is no relationship and using political standpoint to find AQI levels has an even worse accuracy than simply guessing the average.

Modeling

We decided to use multiple linear regression as it works well with multiple independent variables and one dependent variable. Though linear regression is easily influenced by outliers, we suspected there would not be many outliers as we do not have any extremist countries in our dataset. We also decided to use linear regression because the overfitting can be reduced by regularization. The inputs of our model were 3 questions (after we quantified and normalized the responses) in a recent Pew research poll. These three questions specifically looked at a popular favor of "rule by strong leader", "rule by military", and "representative democracy". We decided upon these three questions as they most accurately exemplify and predict the political standpoint of a nation because they are well known preferences of left or right leaning political standpoints. Our model outputted with an average R^2 coefficient of ~-1.8 across 10,000 iterations.

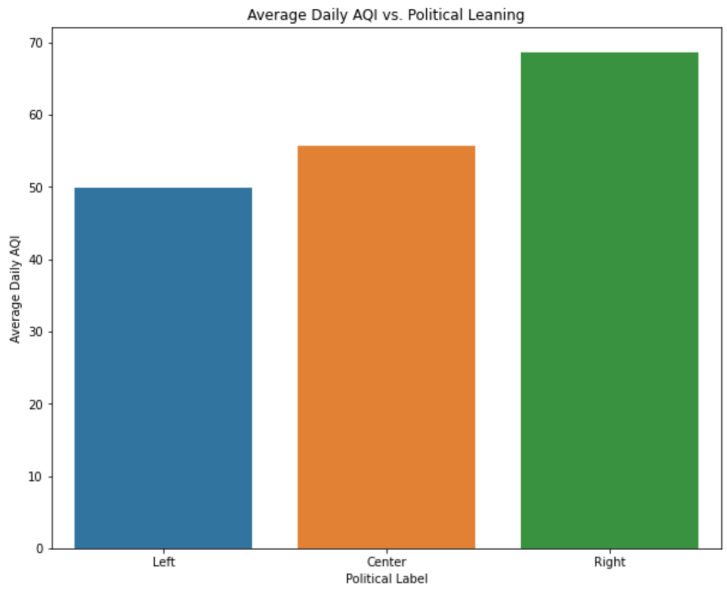
Model Evaluation and Analysis (Visualizations)

After creating our model, we used an R^2 coefficient of determination to determine whether or not our model result rejected or accepted our hypothesis. With the three features, we unfortunately got a model that was considerably worse than just always predicting the average, with an R^2 score of -2.1 across 10,000 iterations. Because the value is below 0, this means that we would have a higher accuracy simply guessing the average AQI than to use political affiliation to find it. This rejects our hypothesis that political affiliation of a country can help predict its AQI.



The plot above showcases the relationship between the answers to our chosen three questions/features and AQI levels. In general, textbook definition of right leaning countries should have high values in rule by military and rule by strong leader and a low value in direct democracy and the opposite for left leaning countries. If our hypothesis was accepted, one would expect to see clumps of similar colors where the orange points would clump by the feature values

that describe right leaning countries and blue for left. However, this plot effectively showcases how our model fantastically rejects our hypothesis as there are absolutely no clumps and no patterns. This prompts us to question interestingly if preferences for direct democracy or rule by strong leader are features that are an accurate portrayal of political standpoint.



This second plot to the left showcases the relationship between political leaning (as we define it) and average AQI. In order to create this plot, we further define left as those countries who approve of direct democracy and disapproval of rule by strong leader and rule by the military and the opposite for right. Those who did not fall in either category were categorized as center. We defined approval as higher than the mean. We previously normalized these values so therefore if a country scored above zero, that means they approve. Based on our hypothesis, we would see a progression of increasing average daily AQI from left leaning as the lowest to center to right leaning as the highest. To our surprise, this bar graph accurately represents our hypothesis. At first glance, it may seem that this visualization strongly supports our hypothesis. However, this visualization is created with a more strict definition of left and right leaning. In order to be left, one must fulfil three requirements, same with the right. Meanwhile our hypothesis defines left and right leaning as an aggregation of quantified answers to the three questions (refer to the first paragraph). This definition of left and right leaning is also simply a personal judgement on our end as data scientists with not much political knowledge. We noticed that Sweden, a primarily left leaning country, was categorized as a center country. This shows that this definition of left and right does not accurately categorize all countries. This graph reminds us of something interesting and hard to address: political standpoint is hard to quantify and is fluid. Some countries may answer in the middle for certain questions while being primarily leaning a certain way. We must be mindful of our definitions while modeling and interpreting data.

Modeling Improvements

After our model proved to have a horrible accuracy rate, we hoped to improve it. The main issue we wanted to address was simply the horrible accuracy rate and we looked for any

potential ways we could improve it. Thus, we decided to add the rest of the questions (normalized and quantified) from the pew research poll as features in our model. We thought this would be an improvement as it would give more clear definitions as to whether a country is left or right leaning and would allow the model to decide for itself. We thought that the horrible accuracy rate may be because of our inaccurate quantification of left and right leaning countries. Adding more features that would tell us whether or not a country is left or right leaning would improve the accuracy of our quantification and sorting. However, unfortunately, our improved model ended up with an even worse prediction than our baseline model. While the baseline model had an average error of ~2.1 across 10,000 trials, our "improvement" of adding the rest of the questions decreased it to ~-2.84. This may be because there is absolutely no relationship between AQI and political standpoint. If this were the case, adding features related to political standpoint would not improve our model at all which was exactly the case.

After seeing that our model predictions got worse after adding more information about political standpoint, we decided to take a different approach. When ideating hypotheses, the topic of a relationship between location and AQI came up. We thought that depending on the geographic location of a country, their AQI can potentially be predicted to be higher or lower. We thought that countries near each other may influence each other environmentally in ways such as policy, practices, and also physically (e.g. smog drifting from one country to another). For this reason, we thought that there may be a relationship between geographic location and AQI. We decided to add location to our model in the form of northwestern/northeastern/southwestern/southeastern dummy variables and to our dismay, our model further worsened to an average R^2 score of ~-3.43. We suspect this may be because geographical location has no relation to environmental standards, practices, and policies. Better indicators for these factors (and potentially AQI in relation) may be economic wealth, GDP, or population numbers.

Future Work

Due to limited time and resources, we barely scratched the surface of what our hypothesis has to offer and the topic of AQI and it's relation to both political leanings and geographic location. With such a vast and interpreted hypothesis, our research is far from the end of it's potential. To further explore this topic, we could look into the features we've mentioned earlier: economic wealth, GDP, or population count. This may help more accurately predict the environmental wellbeing of a country and also it's AQI. This would be quite interesting as it takes in many different features that are way more unrelated to each other than the questions of political standpoint that we used as features.