Data 102 Final Project

Zcjanin Ollesca, Kloe Yap, Alphonsus Yong, Kasey Zhang

May 6th, 2024

1 Introduction

1.1 Background

The alarming rise of atmospheric temperatures spurred by anthropogenic climate change has been the topic of significant concern and debate in current affairs [2]. With ongoing pressure from the Paris Agreement to limit the temperature increase to 1.5°C above pre-industrial levels, governing bodies must act with urgency to decarbonize. Energy is a promising industry sector with the most potential for reaching carbon reductions goals through the deployment of renewable energy sources, such as solar photovoltaics, wind energy, and hydropower [4]. This project was inspired by trends in current literature that shed light on the impact of socioeconomic disparities on air pollution exposure. Understanding the factors influencing greenhouse gas (GHG) emissions is vital in formulating effective and equitable mitigation strategies [1]. The relationship between a nation's economic status, the wealth distribution among its citizens, and resulting environmental impacts on the quality of life provide vital insights into the complexities of global carbon management [6].

1.2 Research Questions

For the scope of this project, we are interested in exploring the dynamics between GHG emissions and socioeconomic factors to guide impactful decision making in decarbonizing the electric grid while promoting a high quality of life for all democraphic groups. To assess this, we explore the following questions:

1. How well do socioeconomic factors (level of income, unemployment rate, and racial diversity) predict CO2 emissions?
2. What is the true causal estimate of above-national-average plant CO2 emissions on limited life expectancy scores?

1.3 Data Overview

For this study, we decided to use the EPA eGrid data provided by the US Environmental Protection Agency, which consists of sampled data. We imported the xlsx files into our workspace and read them in as excel files to convert them into dataframes.

The granularity of our data differed from dataset to dataset. For example, ST22 has a coarser granularity than DEMO22 and PLNT22 do. Each row of our ST22 dataset represents a state in the

United State, or a U.S. territory, while each row in the DEMO22 and PLNT22 datasets represent a plant (for which there are several for each state). ST22 contains information about the annual emissions of various types of gasses such as CO2 and NO2, while DEMO provides demographic information for each plant across the states and territories, and PLNT22 provides similar data compared to ST22 of different gas emissions, at a finer scale (for each plant rather than for each state). The demographic dataset included data on the unemployment rate, income level, limited lifetime expectancy score, and education levels that are scored from a scale of 1-10 from that are defined by the EPA data document spec. Usually, a higher number indicates a higher extent of the factor, such as a higher number corresponding to a higher unemployment rate. Thus, the primary key for our data is based on each electricity plant that corresponds to a county. We applied data aggregation to generalize our results to various regions across the country.

To our knowledge, our dataset was not modified for differential privacy, because our dataset did not contain individual demographic data or high-privacy personal data. Measurement error is especially common in emissions data. Equipment malfunction during the collection of emissions data may introduce several complications that could lead to inaccuracies in the data or errors.

It would have been beneficial to have a column which represented average age of life expectancy, rather than a limited life expectancy score, which made interpretation of our results in our second question a bit complicated. Having this column would provide us with a more concrete estimate of the effects of carbon emissions on life expectancy, and would provide clearer trends between the treatment and the outcome.

The column "Plant annual CO2 emissions" from the PLNT22 dataset was missing some entries, due to inconsistencies between the sources used by the EPA, as well as ambiguous data entries. To deal with these missing entries, we excluded all rows with any null values in our data analysis.

In terms of data cleaning, we log transformed the data in question 1 to equalize the effects of outliers as the scale is large for counties with very high CO2 emissions. We also filtered our data frame to only include socioeconomic factors. For question 2, we added region columns based on U.S. census data and dropped any rows with null values. To assign our treatment, we one-hot-encoded the CO2 emissions to take on a value of 1 if it was above the national average of annual CO2 emissions, and 0 if not. We also used one-hot-encoding on our region variable to perform regression.

2 Exploratory Data Analysis

To decide which socioeconomic features to use in our models, we creatde a scatter plot to observe trends between our explanatory variables and carbon emissions (response variable) (Figure 1).The features we will potentially use to predict CO2 emissions are: state average of income (numerical feature), state average of people of color (numerical), state average of unemployment rate (numerical), and state average of less than high school education (numerical). We decided to refrain from using the "Demographic Index" featured as provided from the data because this variable is correlated with the aforementioned features so it may not be useful for prediction.

2.1 Examining the Relationship Between Numerical Socioeconomic Factors and Carbon Emissions

Next, we created a visualization that provides us with details on the importance of how different socioeconomic factors such as income levels and unemployment rate affect carbon emission. Based on these scatter plots (Figure 1), we made the following observations with regards to CO2 emission:

1. People of Color Index: Weak negative association
2. Low Income Index: Moderate positive association
3. High School Education: Moderate positive association
4. Unemployment Rate: Weak negative association
5. Demographic Index: No association/defined relationship can be established.

Furthermore, we can also see how many of these points are highly variable at the higher ends of the explanatory variable, and do not provide us more information on how these points affect CO2 emissions. All of these distributions are also fairly scattered, with the exception of low income which shows us that many of these socioeconomic factors may not be strong predictors of our model. Thus, we may observe other categorical variables, such as states as state legislation, that might be involved in the emission of CO2 (e.g. California vs Texas when it comes to CO2 permitted levels). This can also help us decide which features are important, and guide feature engineering techniques on certain explanatory variables to improve our model.

2.2 Assessing Potential Confounding Variables Using a Correlation Matrix

Afterwards, we moved on to assessing potential confounding variables. Figure 2 illustrates a correlation matrix which forms pairwise correlations between the variables which helps show which of them are highly correlated with each other and should not be included together in the model. Additionally, variables with stronger correlations with both the treatment and outcome variables are more likely to be influential confounders that need to be accounted for in causal inference. For instance, the heat map shows us that the low income variable and the less than high school education are highly correlated with a score of 0.79. Therefore, this informs us that when we build our model, we should only be including one of them as a feature. Overall, this matrix helps guide us to control for the potential confounding variables and a further analysis of the income variable is done below.

2.3 Examining Another Potential Confounding Variable

When looking at the true causal estimate of CO2 on limited life expectancy, we believe that income level is a confounding variable. Therefore, we want to look at the relationship between the average state income and CO2 and the relationship between average state income and average life expectancy.

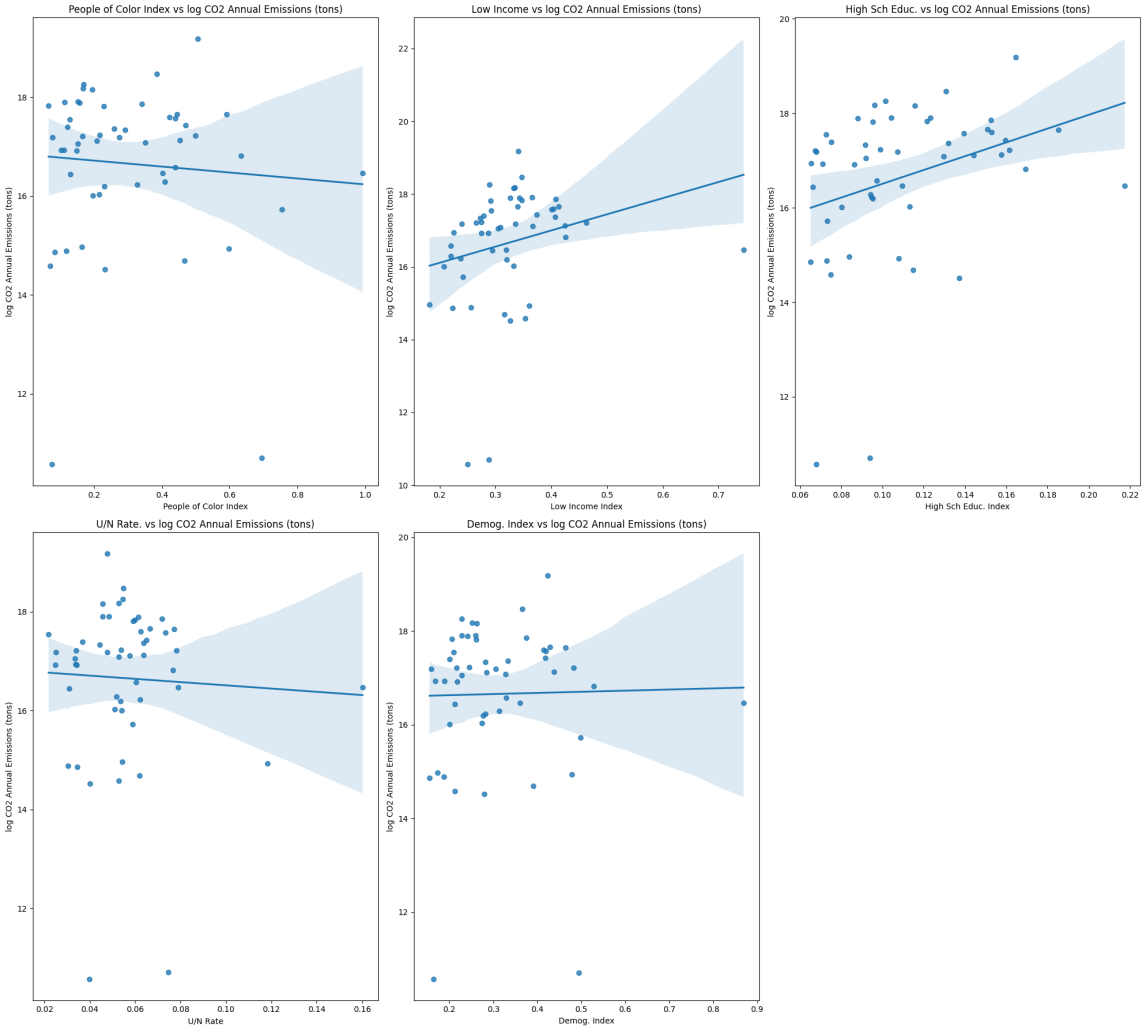


Figure 1: This scatter plot displays the relationship between numerical socioeconomic factors and carbon emissions

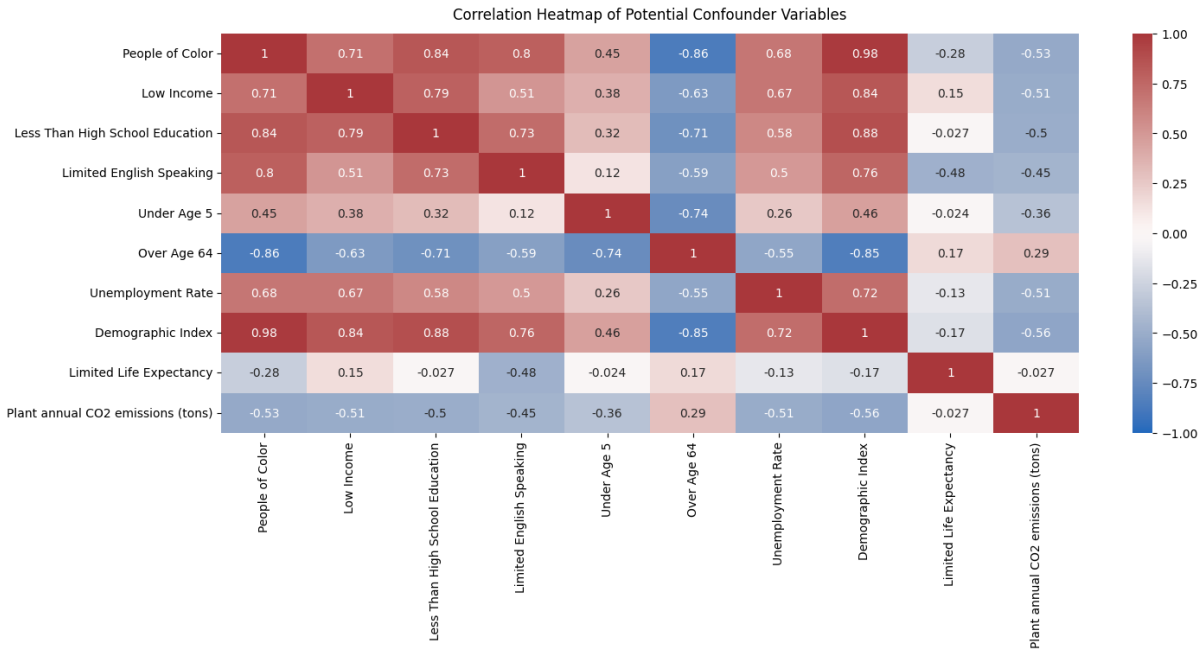


Figure 2: Heatmap of Potential confounder variables using demographic data and carbon emissions data from EGrid.

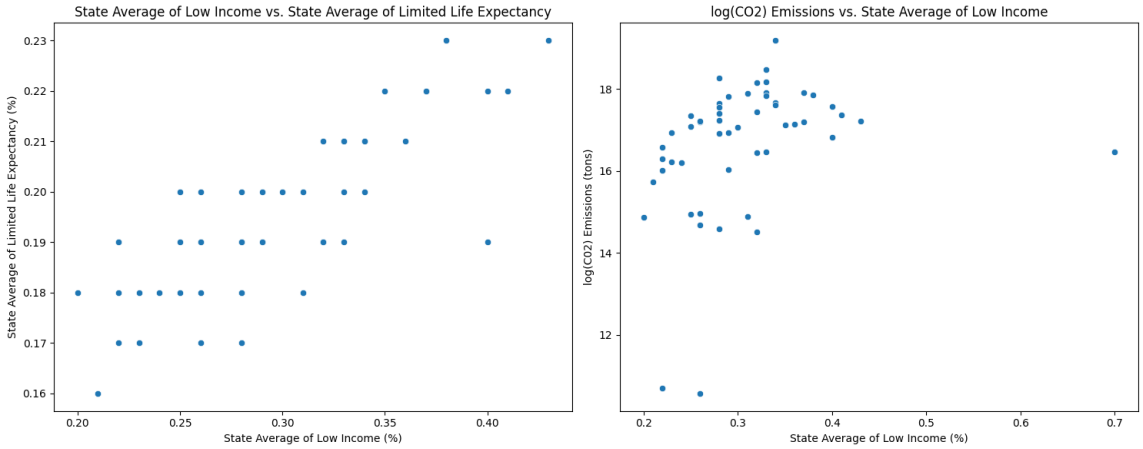


Figure 3: Scatter plots for determining additional confounding variables.

We believe that income level is a confounding variable that would affect the way we measure the causal effect of CO2 greenhouse emissions of life expectancy. Hence, we made 2 scatter plots to visualize whether or not the population in households where the household income is less than or equal to twice the federal "poverty level", measured in percentage, has an effect on the treatment (CO2 emissions) and the outcome (Limited Life Expectancy).

By looking at the scatter plots in Figure 3, we can see that there is an obvious positive correlation between the state average of low income and the state average of limited life expectancy. As the percentage of low income households increase in an area, the percentage of state average limited life expectancy also increases, showing that higher levels of low income is associated with higher percentage of reduced lifespan compared to the average life expectancy. Similarly, the second scatter plot shows that with higher percentage of low income households in an area is associated with higher CO2 emissions (tons). This scatter plot seems to contain a few outliers on the lower and higher ends of percentage of low income households.

These scatter plots enable us to understand the relationship between relevant variables and the potential for confounding in our causal effect question. Since the percentage of low income households in an area is positively correlated with both the treatment and outcome variables, it suggests that income could potentially introduce bias in the estimated causal effect of carbon emissions on limited life expectancy and needs to be controlled for when estimating the effect. In regards to our research question, these visualizations show that to obtain unbiased estimates of the causal effects, income needs to be controlled for and strategies such as matching may need to be utilized to minimize the influence of the confounding variable.

2.4 Examining the Relationship Between Region and Carbon Emissions

Figure 4 represents a bar plot that compares the average annual CO2 emissions between and across U.S. regions in the log metric ton scale. From this bar plot, we observe that the Southern region has the greatest average annual CO2 emissions, while the Western U.S. region has the lowest. The

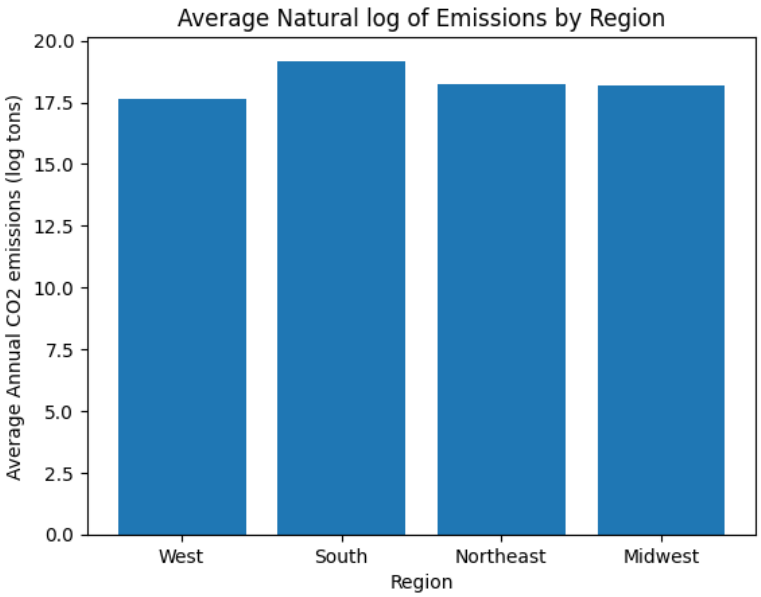


Figure 4: Bar chart of average annual carbon emissions by region.

Northeast and Midwest regions have around the same average carbon emission annually. We may also want to compare other demographic and socioeconomic factors between the South and the West U.S. regions, because these two had the greatest difference in average CO2 emissions.

Based on the trends shown in the bar plot, we can further explore the causal estimate of CO2 on limited life expectancy scores due to differences in crop growth (and therefore diet), lifestyle, and education level across regions. Therefore, our bar plot suggests potential answers to our research question and helps us to decide which features are most important, based on lifestyle factors of each region.

2.5 Examining the Relationship between Region and Type of Primary Fuel for a Plant Generator

Figure 5 allows us to examine the relationship between a geographic region and the types of fuel that power plants are comprised of. Based on the bar chart, we can observe that plants in the Midwest rely primarily on wind power and oil more than any of the other regions. We can also see that solar is the dominating primary fuel category across all regions and the number of plants whose primary fuel category is oil is roughly the same across all regions, with the exception of the Midwest.

This information is relevant to our research question because regions across the United States experience vastly different climate conditions, geography, and resource availability that may drastically impact the choices of most suitable energy sources. The kinds of energy sources most prevalent in these regions may also affect an individual's exposure to carbon emissions and alter their life

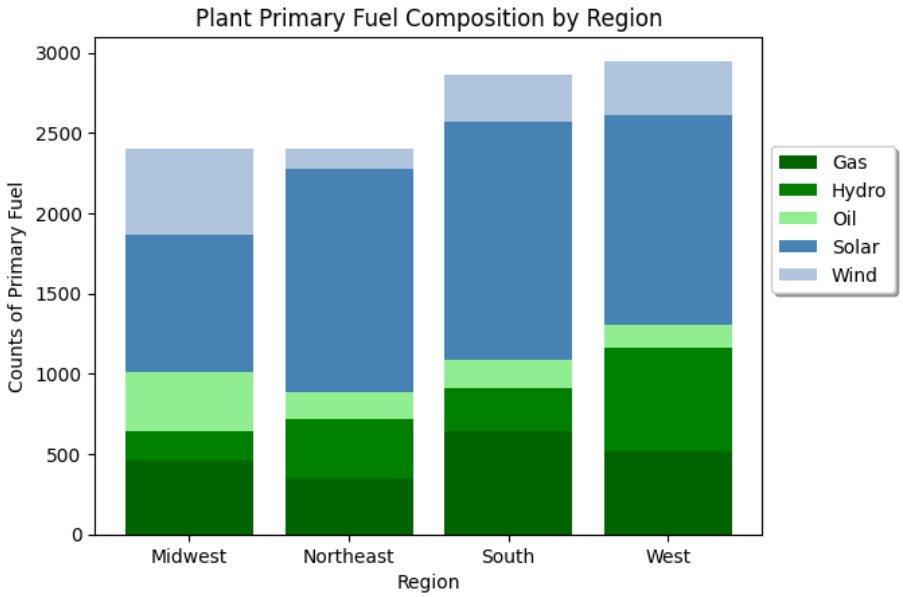


Figure 5: A bar chart illustrating plant primary fuel composition by region.

expectancy differently. This visualization can help to address how the type of primary fuel for a plant generator across regions may be confounding our independent variable, or carbon emissions, and dependent/outcome variable, or life expectancy score.

3 Research Question 1

In our first research question, We are trying to predict carbon emissions based on socioeconomic factors that are collected on a numeric scale from the dataset. These features are collected on a plant basis that are usually based on counties as each county has a main plant generator that can be defined demographically. As such, the distinct features we used were: People of Color, Low Income, and Unemployment Rate. We decided not to include the Demographic Index provided by the data as they were a combination of weighted features of the aforementioned variables (mentioned in EDA). Additionally, we removed High School Education Level based on our Visualization 2 in our EDA as this factor is related to Income which will affect the coefficient due to multi-co-linearity.

We will also use different approaches for evaluating each model's performance. For the Frequentist approach, we will use the Log-Likelihood ratio to assess how well our GLM model fits into our dataset, while also looking at the confidence intervals of each explanatory variable to determine if the association is statistically significant.

In our Bayesian mode, we will be using the Posterior Predictive Check method in the Bambi package to test how the stimulated data fits into our observed data. A good model will show that the observed data falls within the range of the predictive simulations. If the observed data

frequently lies outside the predictive lines, this might suggest the model does not fit the data well.

For our non-parametric methods, as we are dealing with a continuous numerical output, we conducted a RMSE test on the training and test set, and evaluated the models' performance based on having a lower accuracy RMSE score.

3.1 GLM: Multi-Linear Regression

3.1.1 Methods

Since we are trying to predict a continuous quantitative variable (CO2 emissions), we decided to use a multi-linear regression because this allows us to include multiple explanatory variables in our model. We assumed that all of these chosen variables had an association with affecting CO2 emissions linearly as we expect somewhat of a positive or negative linear association between our explanatory and observed variables. First, we decided to explore the Frequentist approach as we are working based on past fixed data, and considering a large sample size of over 3000, we believed the Frequentist method may perform well with this dataset since it relies less on prior information and more on the data itself.

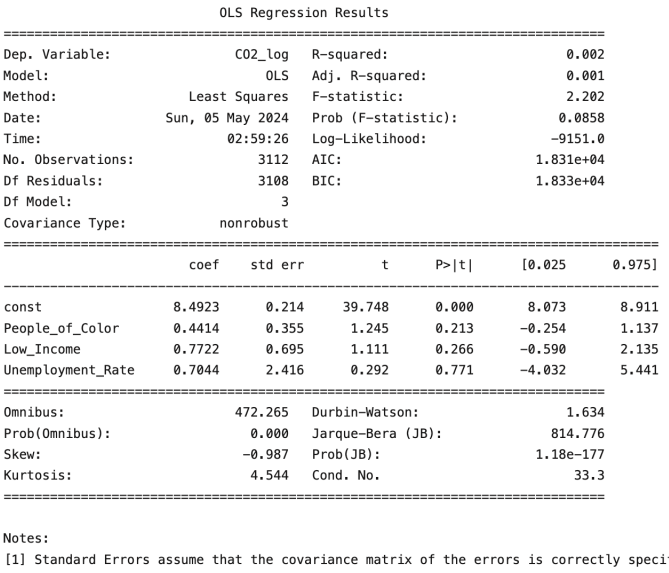
We also conducted a Bayesian GLM as a sanity check, and utilized the prior provided by the Bambi package as we do not have any prior knowledge on the distribution of our explanatory variables. As such, if we were to specify one, it would be using the Normal Distribution to capture variability in the possible values of our coefficients for each explanatory variable. The Normal Distribution is suitable in this instance as we have a large dataset, and by visualization in our Posterior Predictive Check, all of the coefficients follow a normal distribution except for the tail on the left, but this is the best assumption we could use moving forward.

3.1.2 Results

For our Frequentist approach, all three variables "People of Color", "Low Income" and "Unemployment Rate" had positive associations with CO2 emissions on a log scale.

The following literal interpretations can be made using the model summaries (Figure 6, Figure 7):

- A one unit increase in "People of Color" results in an increase by 0.441 units of log-CO2 emissions on average. This implies having a greater fraction of the population with minority races leads to higher CO2 emissions. This makes intuitive sense as areas with a higher fraction of minority populations usually live in areas with higher levels of pollution (disparities on impact of pollution on races).
- A one unit increase in "Low Income" results in an increase by 0.772 units of log-CO2 emissions on average. This implies having a greater fraction of the population with lower-income based on the federal rate leads to higher CO2 emissions. This makes intuitive sense as poorer regions usually face higher pollution levels due to fewer resources to improve air quality and material welfare such as improvements in greener technology.
- A one unit increase in "Unemployment Rate" results in an increase by 0.704 units of log-CO2 emissions on average. This makes intuitive sense with a population with lower income levels attributed to higher unemployment rate as mentioned above.



OLS Regression Results

Dep. Variable:	CO2_log	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	2.202
Date:	Sun, 05 May 2024	Prob (F-statistic):	0.0858
Time:	02:59:26	Log-Likelihood:	-9151.0
No. Observations:	3112	AIC:	1.831e+04
Df Residuals:	3108	BIC:	1.833e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.4923	0.214	39.748	0.000	8.073	8.911
People_of_Color	0.4414	0.355	1.245	0.213	-0.254	1.137
Low_Income	0.7722	0.695	1.111	0.266	-0.590	2.135
Unemployment_Rate	0.7044	2.416	0.292	0.771	-4.032	5.441

Omnibus:	472.265	Durbin-Watson:	1.634
Prob(Omnibus):	0.000	Jarque-Bera (JB):	814.776
Skew:	-0.987	Prob(JB):	1.18e-177
Kurtosis:	4.544	Cond. No.	33.3

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 6: Frequentist GLM model summary

However, it is important to note that all of these variables were not statistically significant, indicated by a high p-value and our 95% confidence intervals containing 0. Thus, there might not be a significant association with our explanatory variables and CO2 emissions.

3.2 Non-parametric Methods: Decision Tree and Random Forest

3.2.1 Methods

For our non-parametric method, we implemented a Decision Tree and Random Forest as these methods do not assume any relationship between CO2 emissions and the demographic factors we chose. Even though we had no categorical variables in our model, these methods are appropriate as they are generally robust to outliers and we can leverage our knowledge on which variables are causing a "tree" to split into their respective "nodes". Random Forest is an additional bootstrap aggregation method to handle the variability of our data and reduce overfitting.

3.2.2 Results

For our Decision Tree and Random Forest, we are unable to ascertain interpretability as we would be having many trees, but we note that our Random Forest model provided a lower RMSE score on our test set as compared to our Decision Tree.

3.2.3 Discussion

Based on the results of our nonparametric models, our Random Forest model performed best as it had a lower RMSE score on our test set as compared to the Decision Tree model. Additionally, the residual plot in our Random Forest model (Figure 9) is variable and randomly scattered, thereby

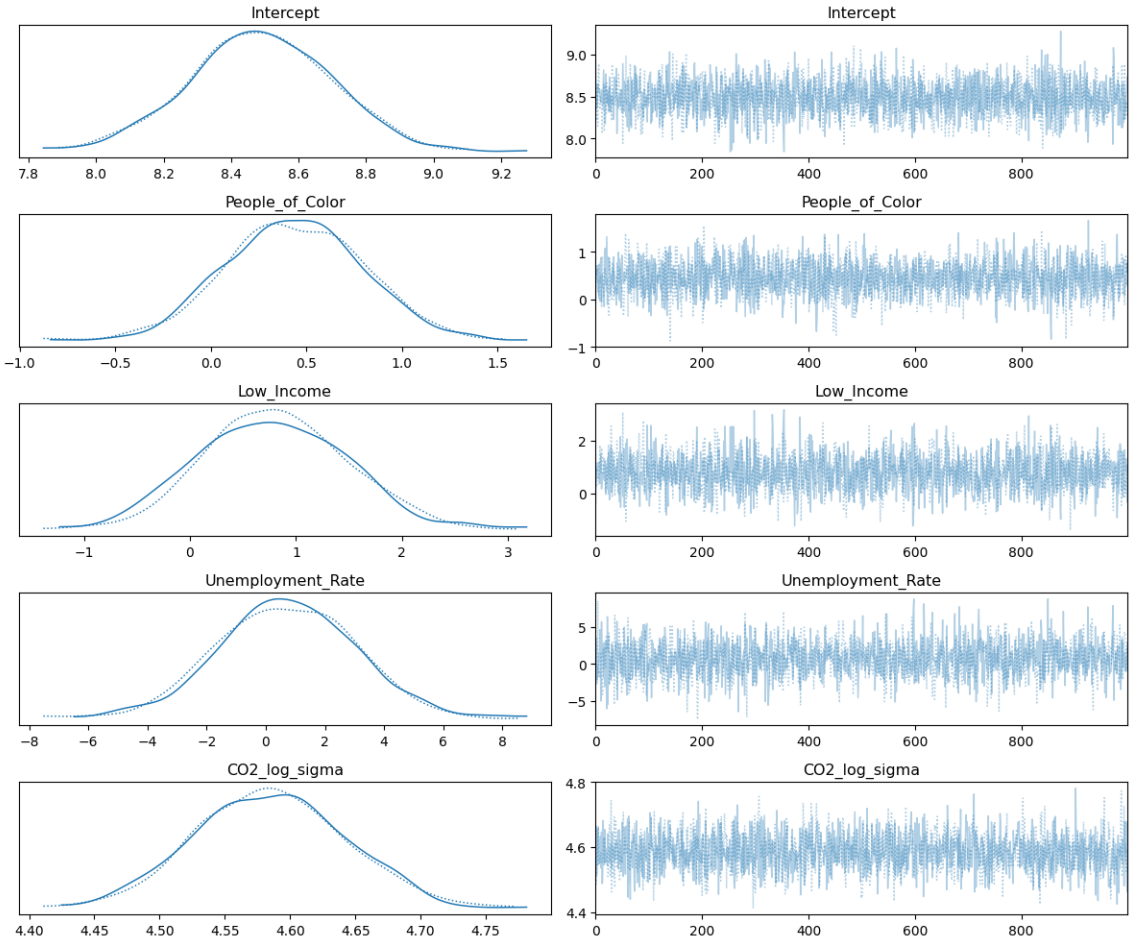


Figure 7: Bayesian GLM model results

being a better fit as compared to our Decision Tree residual plot (Figure 10) which is somewhat clustered into two. Since the Decision Tree residual plot has a clear decreasing trend, we decided that this might not be a good thing to use. Given that the Random Forest performed better, our Decision Tree model led to overfitting due to high variability in our data and is less robust to outliers and noise.

In the case of our Frequentist approach, the Akaike information criterion (AIC) score was very high as our log-likelihood ratio was low as well, thus our data did not fit well into our model. Furthermore, given that all features contained 0 in their confidence intervals, these linear associations were not statistically significant which rendered our model as not credible, thus we would not apply it to future datasets using the Frequentist approach.

Our posterior predictive check was also unable to model the different modes of our observed data, as our posterior predictive mean returned a uni-modal distribution compared to the true multi-modal distribution of our observed data (Figure 8). This implies that there was high variation in our observed data which can be better captured by the use of non-parametric methods. Therefore, we believe that our assumption of linearity placed between the relevant demographic factors and carbon emissions may not hold true. This could mean that there are other factors, besides the demographic ones we have included in our model, that have a larger impact on carbon emissions. Factors such as "People of Color" and "Unemployment Rate" may not have a large impact on behavior modifications such as political opinion or adoption of greener lifestyles on CO2 emissions. High-school level of education was the only statistically significant factor, as seen in our Frequentist and Bayesian approach. This allows us to conclude that there is a strong negative association between level of high school education and CO2 emissions which could likely be due to the increased level of awareness an individual may have on the consequences of carbon emissions.

In terms of the limitations of our model, our Random Forest and Decision tree lacks interpretability since the aggregation of multiple trees, especially in our best model, make it difficult to track the decision-making process. This leads to making the model appear more like a "black box" in terms of understanding how these inputs become outputs. As such, it is harder to determine which features were most important in affecting CO2 emissions unlike our GLM approach which determined education was the most pivotal factor. Additionally, our non-parametric models struggle with high dimensionality as we include additional demographic and socioeconomic variables that can be irrelevant or redundant. As we learned from our GLM model, certain features that had a statistically insignificant relationship with CO2 emissions, we could remove certain features and reduce dimensionality to improve accuracy in our Random Forest or Decision Tree.

For our GLM, this model was limited as it assumes linearity between our explanatory and outcome variables which we have to reconsider based on the insignificance of our results based on our OLS output. As such, this most likely had an impact on our predictive performance as compared to our non-parametric methods since the dataset might feature complex relationships between certain variables that we did not account for. Furthermore, this GLM model is sensitive to outliers and high-leverage points which commonly occur in CO2 emissions as we did not account for the demographic nature of the area (rural vs urban, population density etc.) but only looked at socioeconomic status. These can disproportionately influence the model fit and skew the results, leading to misleading interpretations.

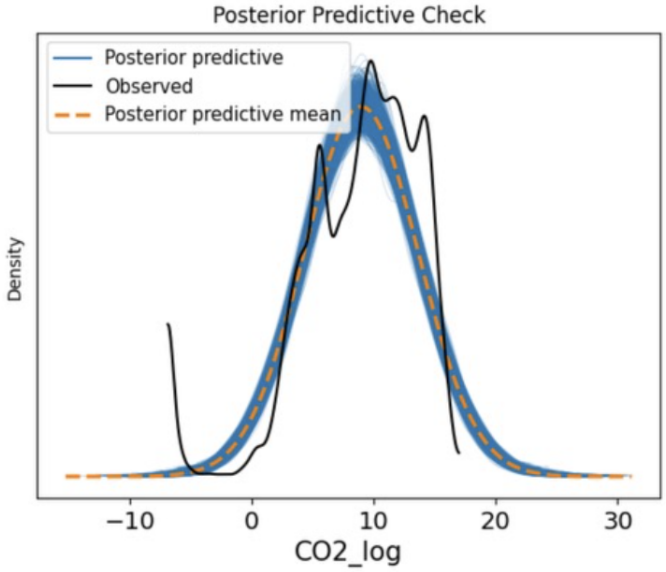


Figure 8: Posterior predictive check using our Bayesian GLM

We believe additional data in our dataset would be helpful to predict future CO2 emissions such as whether the plant existed in a rural or urban area (which can be one-hot encoded), and the population density of the county as these would have been important features to determine CO2 emissions through energy consumption. Though these are factors based less on socioeconomic contexts, demographic features such as population would be crucial in predicting emissions. Additionally, feature engineering could have been done in further studies such as using the Demographic Index which are weighted features on "Income" and "Education" as defined by the data spec to predict these to improve the log-likelihood of our model.

4 Research Question 2

4.1 Outcomes Regression

4.1.1 Methods

To explore the true causal estimate of CO2 emissions on limited life expectancy scores, we defined CO2 emission levels to be our treatment variable, and limited lifetime expectancy score to be our outcome variable in our model (Figure 11). We decided to manipulate the data so that plants that had CO2 emissions greater than the average of all plants' CO2 emissions were considered to have been treated (Z=1), and plants that had CO2 emission levels less than the national average were considered to be untreated (Z=0).

Confounders are variables that affect both the treatment and the outcome variables, in this case,

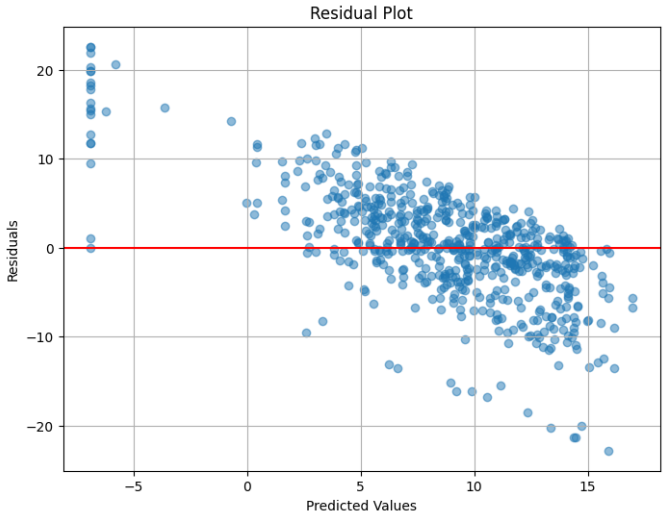


Figure 9: Decision Tree residual plot.

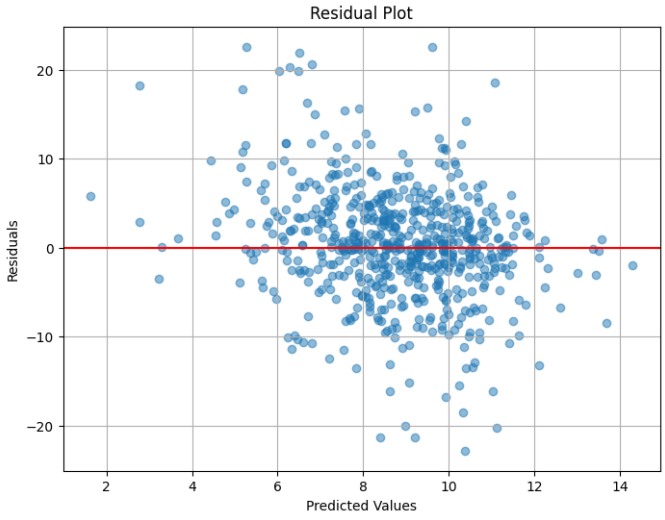


Figure 10: Random forest residual plot.

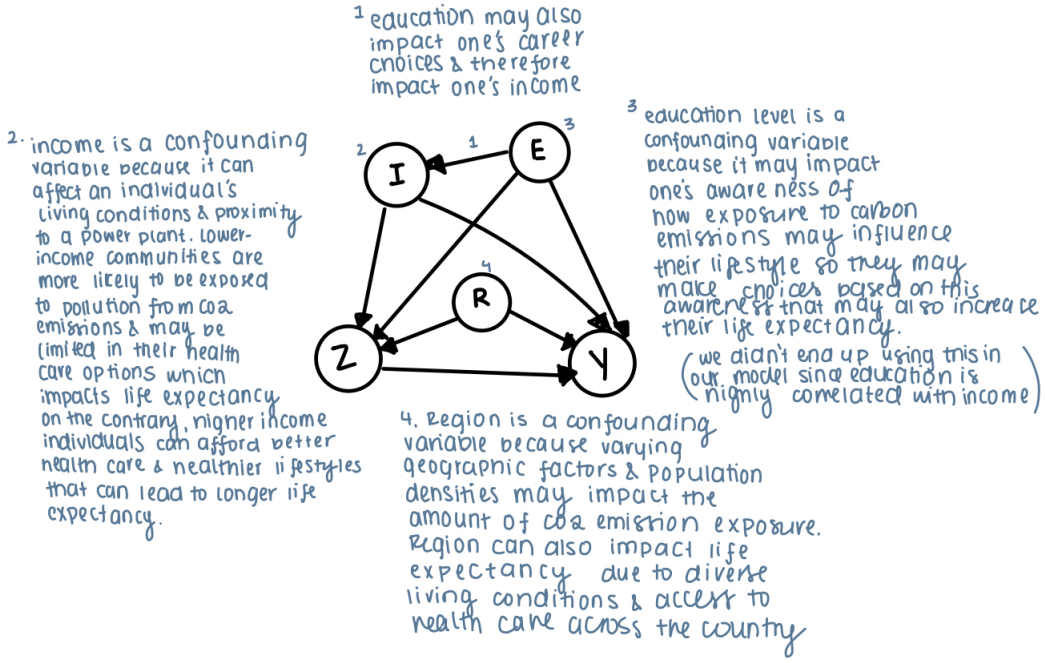


Figure 11: The causal dag for our second research question

examples of confounders would be those that affect both CO2 emissions and limited life expectancy scores. Income, education level, and region are three examples of confounding variables in our set up. Higher income levels may decrease the amount of CO2 emissions because this can mean greater access to cleaner technologies, such as electric vehicles or solar energy, which can oftentimes be expensive. Higher income correlates with a higher life expectancy because those with higher income likely have a higher education level and will likely have access to better healthcare, and might be more knowledgeable in greener lifestyle choices. Region also introduces confounding on our treatment and outcome variables because geographic factors and differing political climates across the country may impact the kinds of primary energy sources that states rely on. Differing energy sources will lead to varying amounts of carbon dioxide emissions emitted, affecting our treatment variable. Region impacts life expectancy because of differing access to healthcare across the country where one region may have greater access due to increased numbers of hospitals and health care professions.

The unconfoundedness assumption holds as we assume that given the confounders of income and region, the treatment variable (CO2 emission levels) and outcome variables (life expectancy scores) are conditionally independent. This implies that these two confounders are the main variables and it encapsulates all confounders influencing both CO2 emissions and life expectancy. We believe this makes sense in the context of our dataset as people of color, limited English speaking, and unemployment rate may not have a material impact since they do not affect material behavioral changes that affect Eco-friendly behavior in an attempt to reduce CO2 emissions. We also did not include variables that are correlated with income, like education level, because this can result in multicollinearity. We discuss these limitations in further detail below.

We employed outcomes regression (Figure 12) which is a method for dealing with confounding variables in the context of an observational study. This method involves assuming unconfoundedness which states that the treatment and the pair of potential outcomes are independent given our confounders. To execute outcome regression, we computed the treatment effect using the inverse propensity estimator and averaged them together. In this case, we are assuming a linear relationship between our variables so we ran logistic regression to fit our features (income, education, and region) and predict life expectancy.

There are no colliders in the dataset based on our treatment and outcome variables. In order for there to be a collider, there must be a variable that is caused by both the presence of higher-than–plant-average CO2 emissions, our treatment, and life expectancy score, our outcome.

4.1.2 Results

Based on our results from running outcomes regression, the average treatment effect of each of our model variables is 0.4114, 0.0337, 0.0455, 0.0583, and 0.0287 with respect to low income, west, south, northeast, and the treatment. Since all of our confidence intervals contain 0, this means that carbon emissions has a significant causal effect on life expectancy. This means that, on average, for each unit increase in CO2 emissions (tons), the limited life expectancy is expected to increase by 0.0287. Although these average treatment effect values are positive numbers, they actually reflect a negative treatment effect because of how the limited lifetime expectancy score is defined. A higher limited lifetime expectancy score represents lower life expectancy, while lower scores represent higher life expectancy (in years). Therefore, our positive ATE values actually represent a negative effect,

OLS Regression Results

Dep. Variable:	outcome	R-squared (uncentered):	0.889
Model:	OLS	Adj. R-squared (uncentered):	0.889
Method:	Least Squares	F-statistic:	5448.
Date:	Fri, 03 May 2024	Prob (F-statistic):	0.00
Time:	20:07:42	Log-Likelihood:	4531.0
No. Observations:	3395	AIC:	-9052.
Df Residuals:	3390	BIC:	-9021.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Low Income	0.4114	0.005	79.610	0.000	0.401	0.422
West	0.0337	0.003	11.514	0.000	0.028	0.039
South	0.0455	0.003	16.677	0.000	0.040	0.051
Northeast	0.0583	0.003	19.320	0.000	0.052	0.064
treatment	0.0287	0.003	9.966	0.000	0.023	0.034

Omnibus:	421.272	Durbin-Watson:	1.345
Prob(Omnibus):	0.000	Jarque-Bera (JB):	732.492
Skew:	-0.828	Prob(JB):	8.74e-160
Kurtosis:	4.562	Cond. No.	3.57

Figure 12: Outcomes Regression model summary

because greater CO2 emissions resulted in lower lifetime expectancy.

There may be uncertainty in our estimate due to limitations of our methods, which is discussed in the discussion section. For example, our estimate of the average treatment effect may not be completely accurate due to omitted variable bias. There may be other variables that could influence both CO2 emissions and life expectancy that were not included in our model to predict outcome. For example, People of Color and Limited English Speaking were omitted from our outcome regression analysis due to our belief that these variables are not good predictors of limited lifetime expectancy scores. Additionally, there may be other variables not defined in our data set that could affect our results, such as regional economic conditions. Additionally uncertainty may stem from measurement errors or from methods of data collection by the EPA, which would affect our ATE calculations.

4.1.3 Discussion

Our methods are limited by the nature of our data. For instance, we were not able to use instrumental variables because this requires meeting the following criteria: (1) the confounding variable is independent of the instrumental variable, (2) life expectancy must be independent of the instru-

mental variable given the treatment, and (3) the IV must have a non-zero effect on the treatment. There is no such instrumental variable in the scope of our data that meets the second criteria because the majority of variables (excluding the treatment and outcome) are correlated with one another. In other words, we could not state that any of the potential candidates for IV affect life expectancy only through our treatment and not also through our confounding variables. This led us to using outcomes regression.

Our methods are also limited by the contents of our data, as there are several other variables that could potentially influence both CO2 emissions and life expectancy that are not included in the model because we did not have access to these datasets. For example, we would expect healthcare access, environmental policies, and regional economic conditions to play significant roles, which may have changed our model's coefficient values and therefore our calculated average treatment effect.

Another limitation of our method is the imbalance of treated and non-treated areas. Recall that we classified the area where the plant is located as treated if its CO2 emissions are greater than the country average and untreated if below the average. From our data, there are more untreated areas (82%) than treated (18%) and having significantly more non-treated samples can potentially bias the prediction towards the majority class. This would affect the estimates of the treatment effect due to less data to learn from the treatment group.

Additional data we could have to answer this causal question is data on healthcare services. Since lifetime expectancy goes hand in hand with health, there are more confounding variables such as healthcare infrastructure, availability of healthcare professionals, and the amount of money some areas might put into healthcare compared to others. Additionally, this data could provide information on cardiovascular or respiratory diseases that may be caused by the CO2 emissions that influence life expectancy. With this additional data, we can control for health outcomes, which can enhance the accuracy of the causal effect of CO2 emissions on limited lifetime expectancy.

We may also benefit from more granular demographic data since life expectancy can be affected by various factors such as race, age, gender, and population density. With this additional information, we can get a better understanding of particular subgroups that may be particularly vulnerable to the health effects of CO2 emissions. This would be useful to answer the causal question since we can control for the additional variables, helping us pinpoint the specific effect of CO2 emissions on limited life expectancy.

To determine our confidence in the model, we consulted the model summary table (Figure 12) generated after running linear regression. Specifically, we inspected whether or not zero was contained in our 95% confidence interval for each variable in our model. A confident model would not contain zero in the 95% confidence intervals for all of our variables. This trend is reflected in our model summary table so we can conclude that there is a statistically significant difference in life expectancy between plants that experience higher-than-national-plant average carbon dioxide emissions and those that did not.

5 Conclusion

In conclusion, we determined that socioeconomic factors such as unemployment rate, income level, and level of diversity do not have a strong negative association with CO2 emissions as the results were statistically insignificant. Additionally, our nonparametric methods were better for prediction as compared to our multi-linear regression GLM as our data was probably highly variable and did not have strong linear associations with CO2 emissions. Our random forest was a better predictor as compared to decision tree as it averages out the variability of our data. For our second research question, we found that there is a significant causal effect of high-than-plant average carbon dioxide emissions on life expectancy, due to the fact that our model summary table's 95% confidence intervals did not contain zero for all of our variables. We can interpret this result to indicate that greater CO2 emissions does in fact cause life expectancy to decrease.

Our models for our first research question were not generalizable when it comes to predicting CO2 emissions since the AIC score for our Frequentist approach was high, while our PPC was not well-fitted in our Bayesian model. In terms of our RMSE for our nonparametric approaches, they were high and could be reduced through further feature engineering or removing outliers from our dataset to improve our accuracy. However, we believe our research question one findings are pretty broad and generable based on our intuitive findings as shown in our primary naive associations from our Frequentist model output. For instance, the factors of unemployment rate, income levels and diversity all generally follow negative trends with increases in these factors as they are areas with lower income which leads to lower investments and priorities to develop these areas into greener spaces as justified by research reports [3].

The results from our model for our second research question are generalizable and applicable across the United States, but most likely not across the world. Some areas of limitation for generalizability would be demographics, as we did not include this in our model due to high correlation to income. We did however account for different regions and the differences in access to and quality of healthcare, as well as lifestyle differences between these regions. However, we would expect healthcare access, quality, and lifestyle to be drastically different in other countries outside of the U.S.

Despite not having a statistically significant conclusion for Research Question 1, we could further assess the relationship between socioeconomic factors and CO2 emissions, and invest in greener technologies in areas that face higher marginalization when it comes to income levels and diversity. Additionally, given that carbon emissions have a significant and negative causal effect on life expectancy, we highlight the need for stringent industrial policies that push for phasing out fossil fuels by substituting in renewables in order to improve the quality of life. The energy sector shows the most potential for driving decarbonization across the nation so it is essential to target this sector and leverage data-driven models to guide investments towards renewables [5].

We did not merge data from another data source, but we did merge two datasets from the eGrid dataset. We merged data from the plants and demographic data together to expand our dataset and with this expansion, we were able to have a more comprehensive understanding of the potential association of limited life expectancy and specific plants. This makes our data more granular to achieve higher levels of specificity in our analysis and uncovers more patterns and trends within the data.

Our data only contains information on certain socioeconomic factors and the electric grid for 2022. Therefore, our analysis is based on the trends captured only in this year and may not be a holistic representation of the long-term trends of the causal effect we are identifying. Additionally, our analysis may be impacted by specific outlier events that occurred in 2022 that may have not occurred in other years. Additionally, the EPA dataset only looks at demographics and CO2 emissions on a county level based on the presence of an energy plant. Thus, counties without an existing energy plant, especially those with a smaller population that rely on other counties for energy, may not be represented in this dataset.

For Research Question 1, our data is limited as our explanatory variables such as people of color, unemployment rate, and income were based on a scale of 1-10 as rated by the US EPA. Despite the data specifications from the EPA, we are unable to ascertain how this numeric feature was calculated as they were based on the definitions by the EPA. If we were able to get an absolute value such as a % of unemployment rate or the median income based on county level, we can implement our model from other data sources and be able to standardize our interpretations from our model.

Our work can be built further upon by performing more research on the long term effects of being exposed to excessive amounts of CO2 emissions. This would involve doing a longitudinal study that can assess the impact of chronic exposure to CO2 emissions on limited life expectancy. There could be potential delayed effects that researchers can gain more information on by following the test subjects. By answering this question, researchers can also build upon it by developing strategies and policy interventions to help reduce exposure to CO2 emissions.

We can expand on this study by including other greenhouse gasses besides CO2 which can provide a comprehensive understanding on some gasses that may have greater influence on people's lives. By comparing the profiles of the different emission gasses, researchers can learn from the regional similarities and differences that may contribute to the development of targeted strategies and policies. To further enhance this study, researchers can aim to broaden the scope beyond CO2 emissions alone. Further exploration into the impact of additional gasses on life expectancy and their predictability could significantly contribute to uncovering the true effects of the treatment.

During the project, we gained valuable insights on the emphasis of considering relationships between various features, especially confounding factors. Utilizing confounding factors that are highly correlated with each other can yield misleading results and heavily influence our coefficient values when running our model. We also discovered the significance of contextualizing our results within the scale of our data. While our ATE appeared to be small, when viewed in the context of the broader dataset scale, where the values were already low, the significance becomes more apparent. This highlights the importance of considering the effects relative to the scope of the study and prevents misleading conclusions from being made.

Overall, our utilization of course material was instrumental in being applied to a real-world dataset. By leveraging the various techniques, we were able to effectively decipher how well the socioeconomic factors predicted CO2 emissions and the effect of CO2 emissions on limited lifetime expectancy. This project provided us practical insights in environmental science research and the significance of employing techniques needed to address complex environmental challenges.

References

[1] Yiting Li, Anikender Kumar, Yin Li, and Michael J. Kleeman. Adoption of low-carbon fuels reduces race/ethnicity disparities in air pollution exposure in california. *Science of The Total Environment*, 834:155230, Aug 2022.

[2] Elisa Papadis and George Tsatsaronis. Challenges in the decarbonization of the energy sector. *Energy*, 205:118025, Aug 2020.

[3] Hiroko Tabuchi and Nadja Popovich. People of color breathe more hazardous air. the sources are everywhere., Apr 2021.

[4] Xuelin Tian, Chunjiang An, and Zhikun Chen. The role of clean energy in achieving decarbonization of electricity generation, transportation, and heating sectors by 2050: A meta-analysis review. *Renewable and Sustainable Energy Reviews*, 182:113404, Aug 2023.

[5] Carnegie Mellon University. An urgent plan to decarbonize electricity by 2035 - engineering and public policy - college of engineering - carnegie mellon university, Sep 2021.

[6] Max Åhman, Lars J. Nilsson, and Bengt Johansson. Global climate policy and deep decarbonization of energy-intensive industries. *Climate Policy*, 17(5):634–649, Jun 2016.