Data 102 Final Project

Aaron Cabeza, Angelique Nguyen, Sara Hanmonty, Sean Yang

I. Data Overview

Our data was on chronic disease and air quality concerns from multiple datasets provided by the Centers for Disease Control and Prevention (CDC). The primary dataset, "U.S. Chronic Disease Indicators," includes annual state-level data on chronic illnesses and their prevalence. Additionally, we are using daily census-tract level data for PM2.5, CO2, and ozone concentrations, which are published through the CDC's National Environmental Public Health Tracking Network. The chronic disease dataset is a census of state-level observations and the air quality and smoking dataset has samples that are aggregated to state-year levels. These datasets were strong enough for us to utilize because they aim to monitor environmental exposures and health outcomes across the U.S. We also used a smoking dataset which highlighted the smoking disparities by race and ethnicity. This dataset was also published by the CDC and is useful to us because we believe that ethnicity may play a part in whether or not someone smokes as well as whether or not they develop COPD. While these datasets are not a perfect representation of how chronic illnesses and air pollution levels affect one another, they represent large collected data samples that we can utilize to make inferences about their connection.

Each row of the chronic disease dataset represents a state-year observation from 2011 to 2014 that captures different chronic illness prevalence as well as the stratification of ethnicity for the recorded individual. There is also the feature of DataValue, which is the crude rate or age-adjusted rate for each case. Crude rate in this case represents the occurrence of an event in a population while ade-adjusted rate also accounts for differences in age in the population. Similarly, each row in the air quality datasets represents an observation of PM2.5 or ozone levels for each specific state. We've also decided to include the percentage of smoking per state as another feature in our dataset, because it is another factor that can contribute to chronic illness alongside air quality. While the datasets aim to cover all geographic regions, weather patterns and other environmental factors may lead to systematic differences in pollutant measurements across regions. There can be a risk of selection bias due to how air quality is more monitored in urban and industrial areas rather than in rural regions. The PM2.5 data is also based on predicted estimates rather than real measured numbers and can possibly introduce measurement error. Our additional datasets, CO2 levels and smoking rates, provided us more information on a state-to-state basis. We believed that the levels of CO2, PM2.5, ethnicity, state location, and ozone concentrations alongside consideration of smoking levels were sufficient together for understanding air quality implications on chronic illness.

The "U.S. Chronic Disease Indicators," left us with many problems regarding multiple empty columns and many NaN values. The dataset contains a column labeled DataValueType, which includes varying data types such as crude rates and age-adjusted rates. To ensure consistency, we filtered out rows with irrelevant or inconsistent DataValueType values and retained the most relevant ones for our analysis. We identified several columns that were either redundant, irrelevant to our analysis, or entirely populated with missing values (NaN). These columns included:

- YearEnd: It was identical to YearStart, so we retained only one column.
- LocationDesc and LocationID: These were redundant since we already had the state abbreviation.
- DataSource, Response, DataValueAlt, DataValueFootnoteSymbol: These columns were either unclear or duplicative of other columns.
- StratificationCategory2, Stratification2, StratificationCategory3, Stratification3: These columns were filled with NaN values and provided no additional information.
- TopicID, ResponseID, StratificationCategoryID2, StratificationID2, StratificationCategoryID3, StratificationID3: These identifiers were not meaningful for our analysis, as they were only filled with abbreviations for the real columns. For example, our column labelled 'Topic' with values such as 'Chronic obstructive pulmonary disease' would have a second column 'TopicID' with values that just shorten it to'COPD'.

By removing these columns, we streamlined our dataset by making it more manageable and digestible. These issues required a lot of preprocessing and filtering, which may have led to the loss of some data. However, the cleaned dataset became more reliable for us to use and matched the information we needed to answer our research questions. Overall, our preprocessing steps allow us to focus on the most relevant data. We wanted to work with as much data as possible, so we decided to focus our table on Chronic Obstructive Pulmonary Disease (COPD) hospitalizations because it is a common lung disease and had the most rows due to it being the most prevalent chronic disease. Rather than working with the pure number of hospitalizations, we decided to work with the number of cases per 1,000 for each state because it provided more rows to work with compared to the exact number of hospitalizations per state.

For our research question, we used Generalized Linear Models (GLMs) to investigate the relationship between particulate matter (PM2.5), ozone levels, state location, smoking prevalence, ethnicity, and C02 emissions in the effect on the number of COPD hospitalizations. We decided to impute missing values in the DataValue column using group-based mean imputation because removing all rows with missing values in 'DataValue' would have significantly reduced the size of our dataset, potentially introducing bias if the missingness was not random.

II. Research Questions

Our two research questions are:

1. How do environmental factors (ozone and particulate matter levels), smoking prevalence, CO2 emissions per capita, demographic stratifications, and health event rates, and location of state influence hospitalization rates for COPD?
2. What is the causal effect of PM2.5 levels on COPD hospitalization rates, controlling for relevant confounders?

Our first research question aims to understand how a range of environmental and demographic factors contributes to hospitalization rates for chronic obstructive pulmonary disease (COPD). By examining these relationships, we hope to identify the most significant predictors and their relative contributions. We thought GLMs and Random Forest regression were relevant because we were interested in using a variety of continuous and categorical variables to predict COPD hospitalization rates, including environmental factors like ozone and particulate matter levels, smoking prevalence, and CO2 emissions per capita. Given the high dimensionality and diverse nature of our predictors, GLMs provided a robust framework for modeling their individual effects, while Random Forest regression captured complex, nonlinear relationships that GLMs might miss. Additionally, Random Forests can highlight the most important predictors in the model, making them particularly useful for identifying key drivers of COPD hospitalizations. However, a limitation of Random Forests is their reduced interpretability compared to GLMs, which provide clear coefficients for each variable. Despite this, the combination of these methods allows us to assess both linear and nonlinear patterns, offering valuable insights for public health interventions.

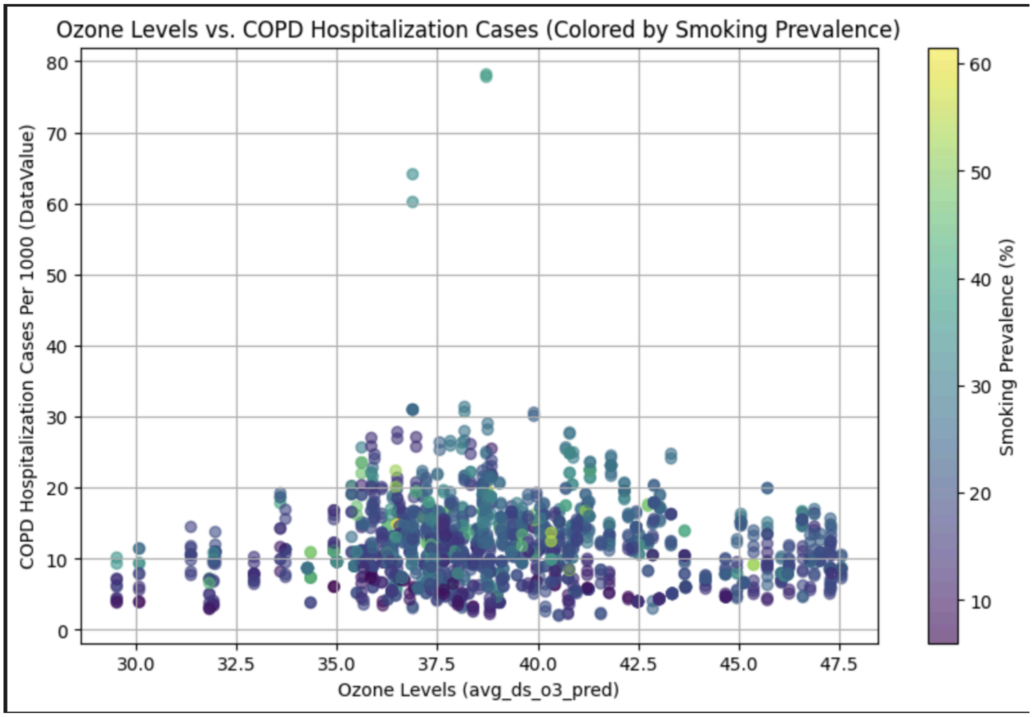
Our second research question investigates the causal effect of PM2.5 levels on COPD hospitalization rates, controlling for relevant confounders. This analysis can help inform public health policies by determining whether reducing PM2.5 levels could significantly lower hospitalization rates. To answer this question we used causal inference methods, specifically outcome regression and inverse propensity weighting. Outcome regression, implemented using Ordinary Least Squares, allowed us to estimate the Average Treatment Effect of high PM2.5 exposure on COPD hospitalization rates while controlling for confounders such as smoking prevalence, ozone levels, CO2 emission per capita, and median household income. IPW adjusted for confounders by reweighting observations based on propensity scores to simulate a randomized experiment to better isolate the causal effect. However, these outcomes still have limitations. For instance, outcome regression assumes that all relevant confounders are included and accurately measured, which may not hold if there are unobserved confounders. IPW is sensitive to misspecification of the propensity score model, which could lead to biased estimates if the model fails to capture the true relationships between the treatment and confounders. Despite these challenges, the combination of these methods provides robust estimates of the causal impact of PM2.5 on COPD hospitalization rates, providing valuable insights for future research and policy development.

III. Prior Work:

We were interested in this dataset from the start because we knew that there would be previous studies and reports on how environmental factors like exposure to PM2.5 rates and ozone levels have effects on developing respiratory issues. Our first source written by Mbabazi Kariisa and other authors concludes that even "low ambient levels of fine particulate matter and ozone can significantly affect respiratory function in COPD subjects" (Kariisa et al. 2014). Since low levels of particulate matter and ozone can make a difference in respiratory function of COPD subjects, we wanted to investigate the actual impact that these environmental factors have in actually developing COPD. We were also interested in how ethnicity factored in when it comes to the number of hospitalizations. The research and study done by Fragoso and other authors displayed a disparity in respiratory impairment among different ethnicities and concluded that more research should be done regarding the subject. Thus, we wanted to see how a combination of ethnicity and other factors would influence the number of COPD hospitalizations. The second visualization in the EDA section discusses this more as we took a look at the distribution of COPD cases in different states by ethnicity.

IV. EDA

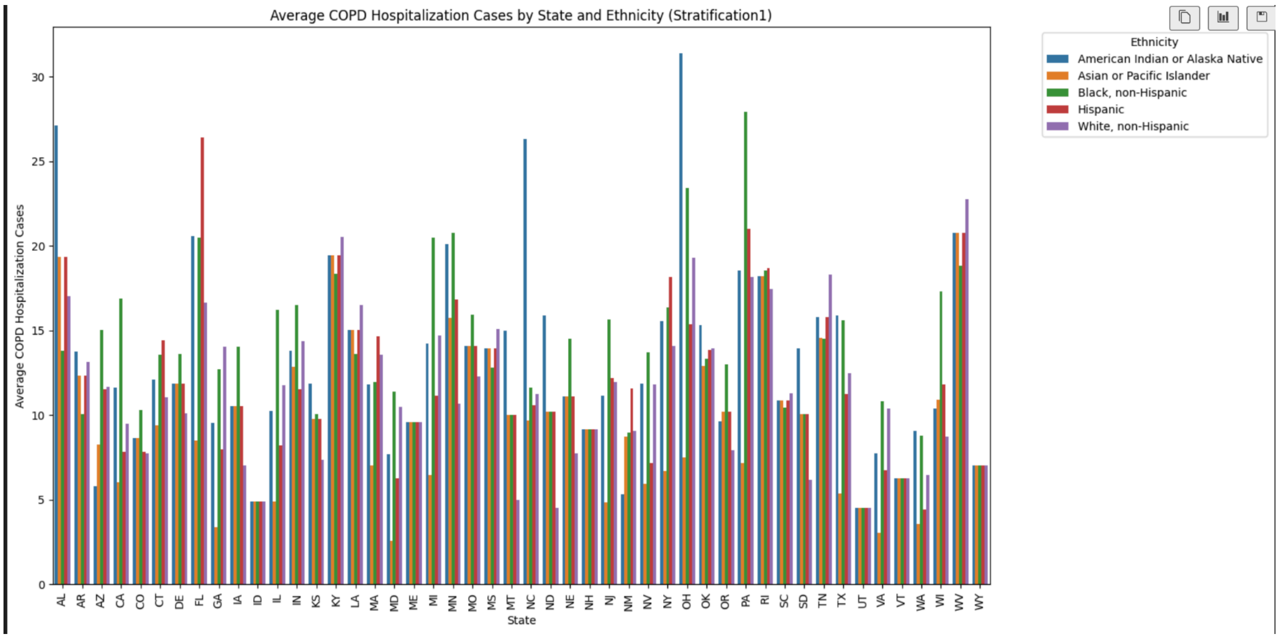
a. **Visualization 1:** Ozone Levels vs. COPD Hospitalization Rates (Quantitative)

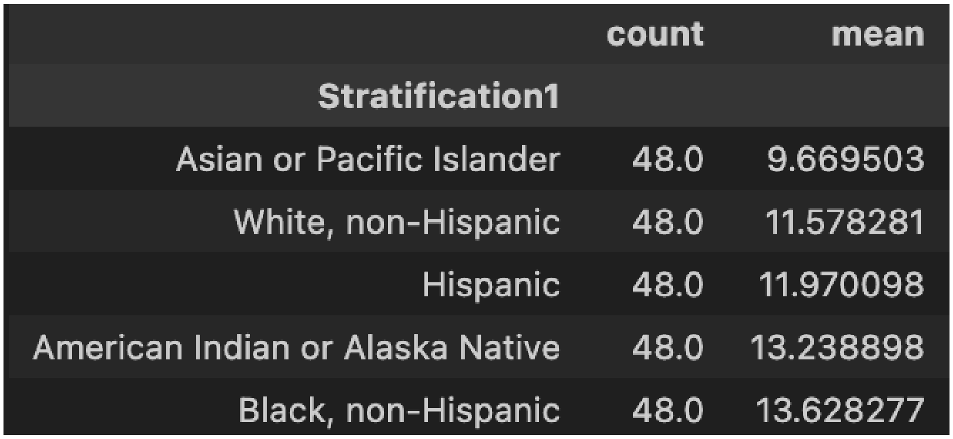


The scatter plot here is pretty scattered and there is not really a clear linear pattern, but it is densely clustered. We can see that most of the data is clustered around the center where ozone levels are between 35.0 and 42.5 and the COPD rate is around 3 to 30 cases per 1000. We can see that points with higher smoking prevalence do have a higher number of hospitalizations compared to the points with a lower number of cases. Since the data is pretty clustered and there is not a clear linear trend, we want to use other predictors like PM2.5 level and other factors that we are curious about listed in our research question. It seems like ozone by itself is not a great indicator of the number of cases, but along with smoking prevalence, we can see more of a trend. Thus, by combining the two with more factors such as particulate matter, we may be able to see a more linear correlation if we do believe that these factors have an influence on the number of COPD hospitalizations.

In our research question, we are interested in the influence of various environmental factors, smoking prevalence, and more on the number of COPD hospitalizations. From our research in prior work done related to this topic, we know that ozone has some sort of effect on the prevalence of respiratory diseases, so visualizing this relationship allows us to get a clear picture on what that kind of effect looks like. Additionally, smoking is a common factor amongst people who develop respiratory diseases so seeing how both smoking rates and ozone levels can influence the number of hospitalizations is relevant to see how rates are influenced.

b. **Visualization 2:** Average COPD Hospitalization Cases by State and Ethnicity (Categorical)



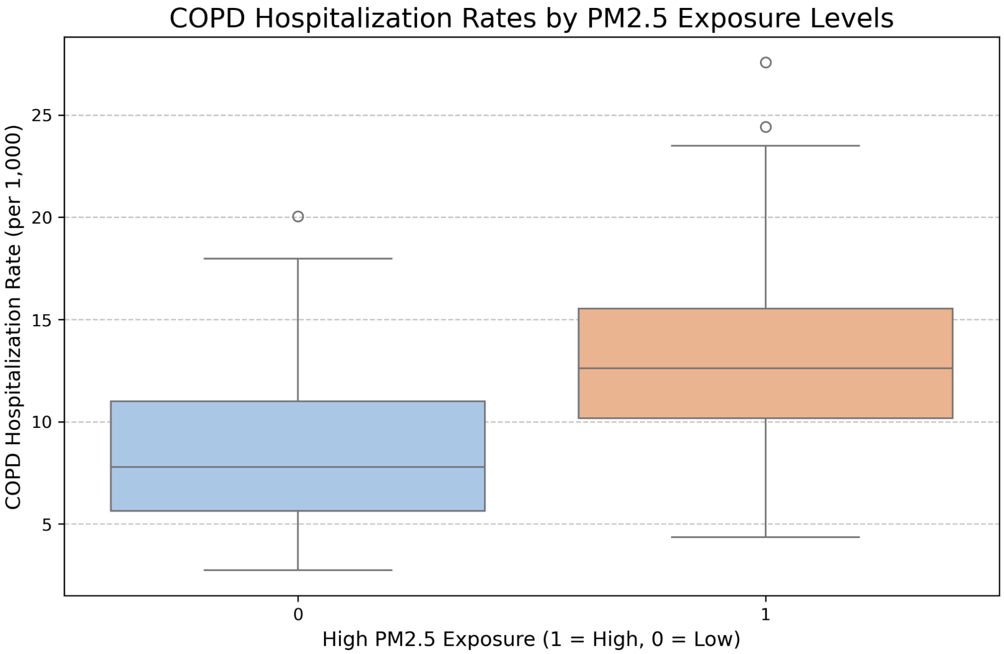


	count	mean
Stratification1		
Asian or Pacific Islander	48.0	9.669503
White, non-Hispanic	48.0	11.578281
Hispanic	48.0	11.970098
American Indian or Alaska Native	48.0	13.238898
Black, non-Hispanic	48.0	13.628277

From the visualization, we can see that the number of COPD cases are very spread out amongst the different states in our dataset depending on the population of the state and the ethnicity breakdown. After taking the average, we can see that Asian or Pacific Islander have the lowest number of cases while Black, non-Hispanic have the highest number of cases. This is interesting because it is different from a previous case that we studied that showed that White-Americans had a higher prevalence rate for chronic bronchitis and emphysema compared to African-Americans and Hispanic-Americans. However, the difference may come from the fact that we are directly studying COPD rather than the other respiratory disease that the study was interested in. We are interested in further following up on how smoking prevalence ties into this relationship among other factors. Just ethnicity and the location of the state unfortunately doesn't tell us much about the number of COPD cases. Thus, by including environmental factors like PM2.5 levels and ozone levels, we may be able to see how the environmental conditions in each state influence the number of cases.

Different ethnic groups in different states will have a varying number of COPD cases which we are interested in. This may come from the fact that each state has different environmental conditions and that states are composed of varying ethnicity groups. This visualization opens up different angles into how we can train our models and the different features we can add into our models in order to get a better understanding on the number of cases of COPD.

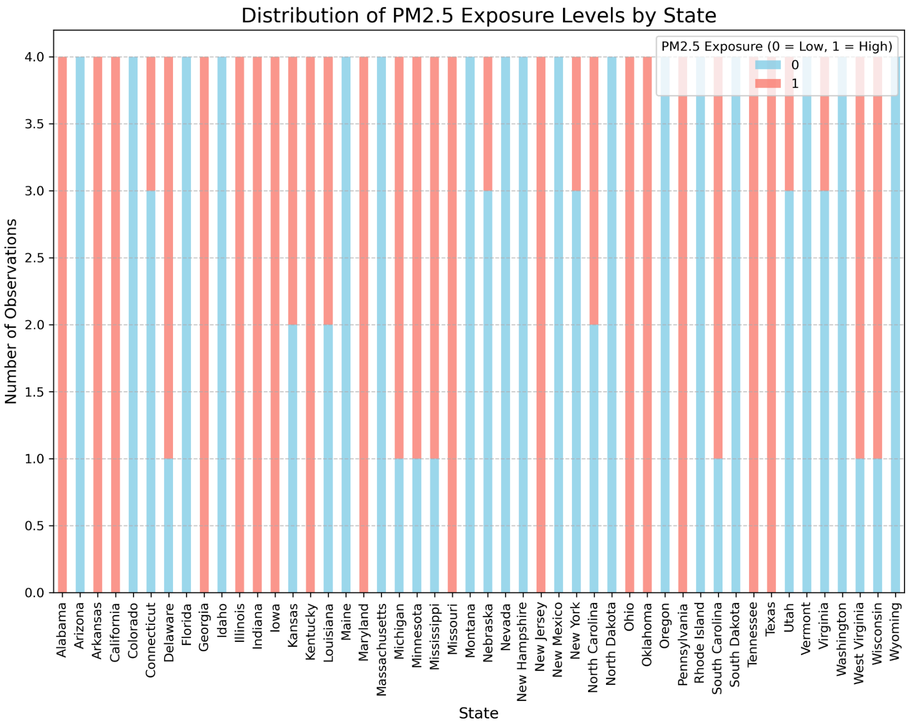
c. **Visualization 3:** COPD Hospitalization Rates by PM2.5 Exposure Levels (Quantitative)



From this visualization, we can observe that COPD hospitalization rates are noticeably different between the individuals exposed to high versus the individuals exposed to low levels of PM2.5. The median COPD hospitalization rate for individuals with high PM2.5 exposure is a lot higher compared to those with low exposure. There is an upward shift of the boxplot for the high-exposure group as well. This suggests that there is a potential relationship between PM2.5 levels and respiratory health outcomes. The interquartile range for the high PM2.5 group is also broader, showing greater variability in hospitalization rates that are in this category. There are also outliers in both groups, but they seem to be more extreme in the high-exposure group. It can reflect how specific regions or populations are disproportionately affected.

This visualization highlights the potential impact of environmental factors like air pollution on public health outcomes. While PM2.5 exposure seems to play a role, there can be other contributing factors such as socioeconomic status, healthcare access, and smoking prevalence that may also influence COPD hospitalization rates. We realized it would be important to explore how these variables interact with PM2.5 exposure levels as it can improve our ability to predict COPD-related outcomes.

d. **Visualization 4:** Distribution of PM2.5 Exposure Levels by State (Categorical)



From the chart, we observe that the distribution of PM2.5 exposure levels is uneven across the states. California and Arizona have a higher representation of observations with high PM2.5 exposure levels (1) while Montana and Wyoming are represented by low PM2.5 exposure levels (0). This indicates that exposure to PM2.5 is not evenly distributed across states, with certain regions consistently experiencing higher or lower pollution levels.

This disparity likely means that there can be geographic, industrial, or urbanization patterns that influence PM2.5 exposure. States with larger city areas or industrial activity like California are more prone to higher PM2.5 exposure, whereas rural or less industrialized states like Montana may have cleaner air.

V. Predictions with GLMs and Nonparametric Methods

a. Methods

Generalized Linear Model (GLM) with a Gamma Family:

We modeled COPD hospitalization rates using a GLM with a Gamma family and a log link function. The response variable (DataValue) is the rate of hospitalizations per 1,000 individuals. We chose the Gamma family because the response variable is continuous, positive, and also right-skewed. Additionally, we used the log link function to make sure that our predictions stayed positive and our predictors were easier to interpret.

The features used for the GLM were ozone levels (avg_ds_o3_pred), particulate matter levels (avg_ds_pm_pred), ethnic group (Stratification1_encoded), state location

(LocationAbbr_encoded), smoking prevalence (smoke_percent), and CO2 emissions per capita (co2_per_capita).

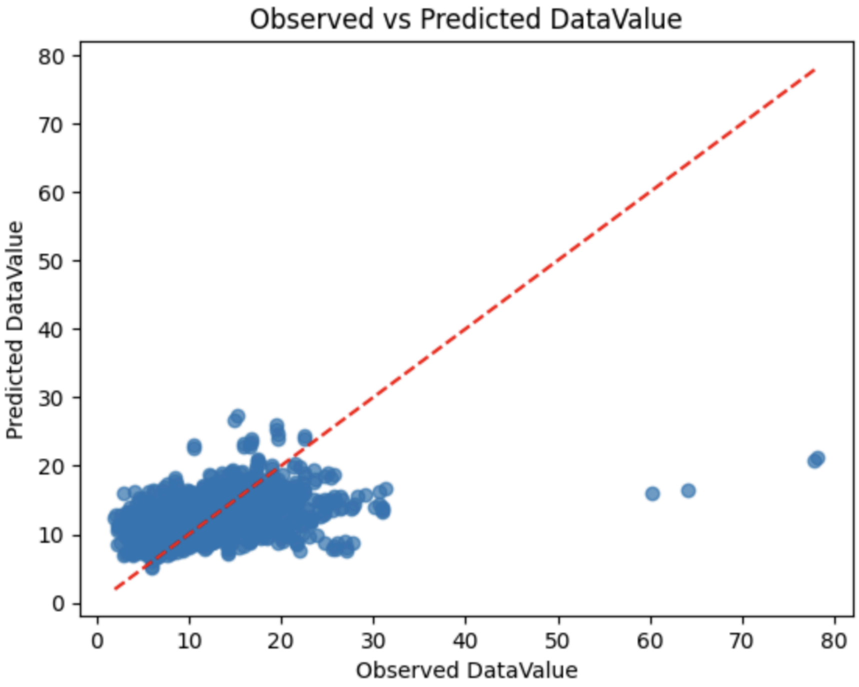
Nonparametric Method: Random Forest

As a nonparametric approach, we used a Random Forest regression model to predict the same response variable. Random forests were chosen because they are able to capture complex, nonlinear relationships and are robust in situations where some features are highly correlated or not very useful. To keep things consistent, we used the same predictors as in the GLM. The Random Forest model was trained and tested using an 80-20 train-test split using 100 trees and we measured model performance using MSE, R-squared, and feature importance scores.

b. Results

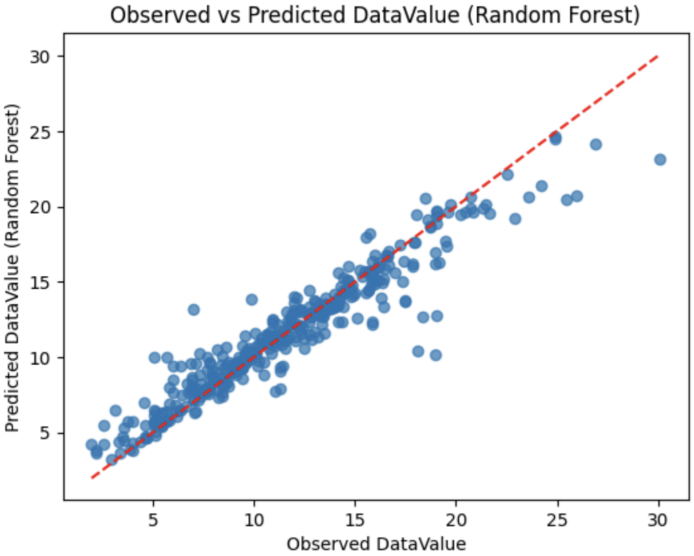
GLM Results:

The GLM achieved an MSE of 19 and R-squared of 0.19 on the test set indicating a modest level of explanatory power. Among the predictors, particulate matter levels (avg_ds_pm_pred) and smoking prevalence (smoke_percent) were the most significant, with particulate matter levels showing a strong positive association with COPD hospitalizations. While the GLM performed reasonably well for lower hospitalization rates, it struggled to capture extreme values, as seen in the residual plots.

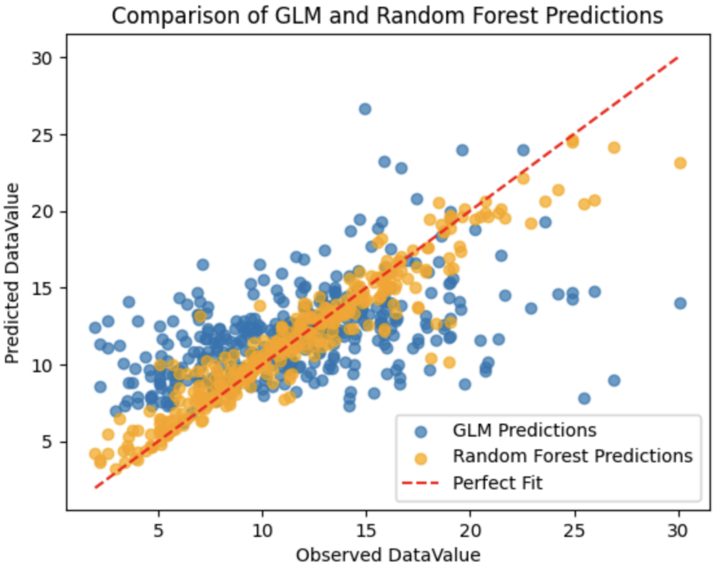


Random Forest Results:

The Random Forest outperformed the GLM, achieving an MSE of 2.69 and an R-squared of 0.90 on the test set. The observed vs. predicted plot for the Random Forest model showed predictions closely aligned with the line of perfect fit. After analyzing what features were the most important, particulate matter levels, smoking prevalence, and C02 emissions per capita emerged as the most influential predictors.



The comparison plot highlights the differences between the GLM and Random Forest predictions. The Random forest provided significantly better predictions, especially for higher hospitalization rates, whereas the GLM's predictions deviated more from the observed values. These differences are also reflected in the metrics, with the Random Forest demonstrating far lower MSE and higher R-squared values.



C. Discussion

The Random Forest model substantially outperformed the GLM in predictive accuracy, likely due to its ability to model more complex, nonlinear relationships. However, the GLM provided results that were easier to interpret, allowing for a better understanding of individual predictor effects. For instance, the GLM indicated that a 1-unit increase in smoke_percent was associated with a 1.72% increase in COPD hospitalization rates, if you hold the other factors constant.

The GLM showed greater uncertainty, as reflected in its lower R-squared and higher MSE. This uncertainty may stem from the assumption of linear relationships and the inability to capture interactions between predictors. On the other hand, the Random Forest's high accuracy suggests it captured these interactions, but its complexity limits interpretability.

VI. Causal Inference

a. Methods

As we progressed with this research question, we performed additional cleaning and aggregations on the dataset to better align it with our analysis goals. These changes, which will be discussed in detail later, resulted in a slightly modified dataset compared to the one initially used to train the GLM and non-parametric models mentioned earlier.

In this class, we mainly focused on binary treatments. However, for this research question, our treatment of interest is PM2.5 levels, which is a continuous variable. To align with the coursework, we created the binary feature high_pm25 by binarizing PM2.5 levels based on the median value (~9.15 µg/m³). The median value is used to ensure equal-sized treatment groups, with half of the observations classified as "high" (1) and the other half as "low" (0) PM 2.5 levels in a state-year. This balanced grouping helps satisfy the overlap assumption required for the propensity weighting method. By ensuring that, for every level of the confounders, there is a non-zero probability of being in either the treatment or control group, the overlap assumption allows for more robust and unbiased estimates of the Average Treatment Effect (ATE).

The outcome variable of this analysis is the age-adjusted COPD hospitalization rate per 1,000 individuals for each state-year. Initially, the dataset provided hospitalization rates for specific ethnic groups, with each row representing a state-year observation for a single group. To analyze the overall effect of PM2.5 exposure on COPD hospitalization rates for the entire population, we aggregated the DataValue column by summing the rates across ethnic groups for each state-year. Since the rates are already age-adjusted, we did not need to include age as a confounder in our analysis. This approach allowed us to combine the data into a single measure representing all racial groups while accounting for age differences within the population.

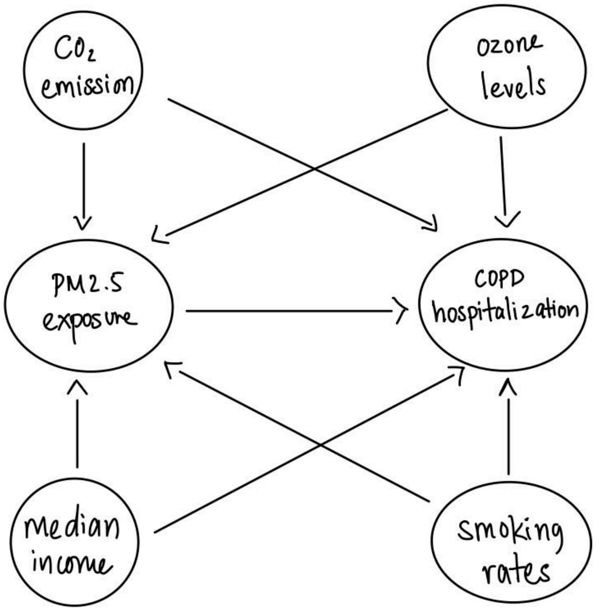
The primary confounders include smoking prevalence, average ozone levels, CO2 emissions per capita and median income. Smoking is a well-documented risk factor for COPD and correlates with PM2.5 levels, as smoke contains fine particles. To ensure consistency, smoking prevalence was also aggregated across all racial groups for each state-year, which further shrinked our dataset. Ozone levels are another air pollutant that independently impacts respiratory health and often coincides with high PM2.5 levels in polluted regions. Similarly, CO2 emissions per capita reflect industrial activity, which contributes to PM2.5 levels and can elevate respiratory health risks through increased exposure to pollutants. Median household income serves as an indicator of socioeconomic status, influencing both environmental exposure and health outcomes. This was done by merging with the median income dataset from the National Center for Education Statistics on the YearStart and LocationDesc columns.

While we made efforts to include relevant confounders in our analysis, some key datasets, such as those on wind patterns or wildfire occurrences, were not available. These variables are critical confounders because they can influence both the treatment (PM 2.5 levels) and the outcome (COPD hospitalization rates). We also explored the use of an instrumental variable (IV) for achieving unconfoundedness, and upon further analysis, we identified that the most ideal option would be policies that impact PM2.5 levels, such as regional restrictions on industrial emissions. However, obtaining detailed and reliable data on such policies proved challenging, limiting our ability to pursue this approach.

Due to the many aggregations performed, our final dataset contains only ~200 observations. We think this smaller sample is acceptable for this kind of question as we are not splitting the dataset into training and test sets, as is common in predictive modeling. Moreover, we opted against using matching because they often result in the loss of data, which would be problematic in our already small dataset. Importantly, there are no colliders in the dataset, as none of the variables are simultaneously caused by both the treatment and the outcome. This ensures that our analysis is not biased by collider conditioning, allowing us to focus on addressing confounding variables instead.

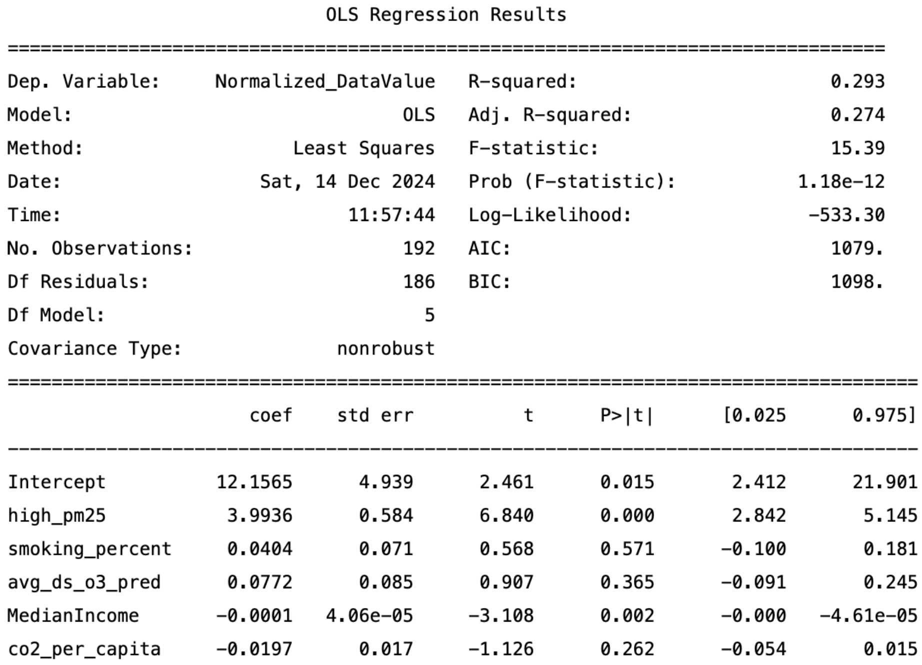
Therefore, to estimate the causal effect, we explored two methods: outcome regression and inverse propensity weighting (IPW). Outcome regression, implemented using Ordinary Least Squares (OLS), estimates the ATE of high_pm25 on DataValue at the state-year level while accounting for confounders. This method offers an interpretable framework to directly quantify the impact of PM2.5 exposure on COPD rates. To account for the uncertainty in our OLS estimates, we used bootstrapping, resampling the data multiple times to generate confidence intervals for the ATE, providing a robust measure of statistical significance and variability.

To adjust for confounders, we utilized IPW, which reweights observations based on their propensity scores. Propensity scores were estimated using a logistic regression model that included relevant covariates. By applying IPW, we aim to simulate a pseudo-randomized experiment where the treatment assignment is independent of the confounders, improving our ability to isolate the causal effect of high PM2.5 exposure on COPD hospitalization rates.



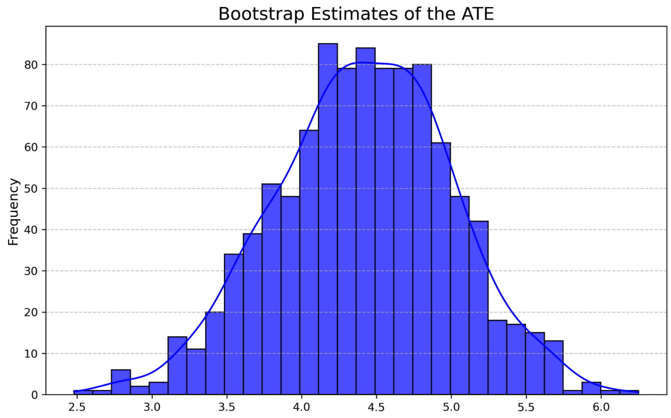
Causal directed acyclic graph (DAG)

b. Results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Normalized_DataValue   R-squared:                       0.293
Model:                              OLS   Adj. R-squared:                  0.274
Method:                   Least Squares   F-statistic:                     15.39
Date:                  Sat, 14 Dec 2024   Prob (F-statistic):           1.18e-12
Time:                          11:57:44   Log-Likelihood:                -533.30
No. Observations:                   192   AIC:                             1079.
Df Residuals:                       186   BIC:                             1098.
Df Model:                             5
Covariance Type:              nonrobust
=================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept         12.1565      4.939      2.461      0.015       2.412      21.901
high_pm25          3.9936      0.584      6.840      0.000       2.842       5.145
smoking_percent    0.0404      0.071      0.568      0.571      -0.100       0.181
avg_ds_o3_pred     0.0772      0.085      0.907      0.365      -0.091       0.245
MedianIncome      -0.0001   4.06e-05     -3.108      0.002      -0.000   -4.61e-05
co2_per_capita    -0.0197      0.017     -1.126      0.262      -0.054       0.015
```

Outcome Regression with Ordinary Least Squares (OLS): The coefficient for the binary treatment variable, high_pm25, is ~3.99, with a p-value of <0.001. This indicates a statistically significant effect of high PM2.5 exposure on COPD hospitalization rates at the state-year level. The 95% confidence interval for this coefficient ranges from 2.842 to 5.145, excluding zero and thereby providing strong evidence of a causal relationship under the assumption of no unmeasured confounding. Interpreting the coefficient in the context of the regression model and state-year aggregation, this result suggests that states with high PM2.5 levels experience ~4 additional COPD hospitalizations per 1,000 individuals annually, compared to states with low PM2.5 levels.



Using 500 bootstrap samples, we calculated the ATE of high PM2.5 exposure on COPD hospitalization rates. The bootstrap ATE estimate is 4.423, with a 95% confidence interval ranging from 3.324 to 5.640. Similar to the OLS results, the confidence interval does not contain zero, reinforcing the conclusion that high PM2.5 levels have a statistically significant causal effect on COPD hospitalization rates at the 95% confidence level.

Inverse Propensity Weighting (IPW): The estimated ATE using IPW is ~3.44. This indicates that, on average, states exposed to high PM2.5 levels experience an increase of 3.44 age-adjusted COPD hospitalization rates per 1,000 individuals compared to states with low PM2.5 levels, after adjusting for confounders. This positive value further supports the hypothesis that higher exposure to PM2.5 levels is associated with an increased risk of COPD-related hospitalizations. However, the slightly lower effect size in the IPW estimate compared to OLS and bootstrap results may reflect differences in how the methods adjust for confounders and weight the data.

c. Discussion

The unconfoundedness assumption in this analysis presents a key challenge, as it is difficult to ensure that all confounders are accounted for. Our dataset lacks key confounders such as wildfires occurrence or wind patterns, which are likely correlated with both PM2.5 exposure and COPD hospitalization rates. These omissions could introduce unobserved variable bias, potentially skewing our estimates of the causal effect. While the included covariates, such as smoking rates, income, and ozone levels, attempt to mitigate this issue, there is no guarantee that our model fully captures all confounding relationships, leaving room for uncertainty. Therefore, access to such data could have brought us closer to achieving unconfoundedness.

Additionally, the PM2.5 levels used in this study are predicted estimates provided by the CDC rather than directly measured values, which introduces a degree of uncertainty and potential bias due to measurement error.

Our dataset is also limited by its small sample size and the temporal scope of the data. With ~ 200 observations spanning from 2011 to 2014, the results may not be representative of broader temporal trends or applicable to more recent years. Air quality and healthcare systems have evolved significantly in the last decade, so the observed associations between PM2.5 levels and COPD hospitalization rates may not generalize to other time periods. Moreover, the state-year-level aggregation of our data prevents us from identifying individual-level variation, making our findings more reflective of statewide trends rather than personal exposure and health outcomes.

Despite these limitations, we are moderately confident that a causal relationship exists between high PM2.5 levels and increased COPD hospitalization rates. Both regression and IPW analyses show statistically significant results, and the confidence intervals do not include zero. However, this confidence is tempered by the limitations of our data and methods. A plausible alternative explanation is that high PM2.5 regions may share other environmental or socioeconomic factors that independently contribute to higher COPD rates. While our results suggest a meaningful relationship, gaps in our data prevent us from being fully confident in this conclusion.

VII. Conclusion

Outcomes Summary: Our analysis using GLMs and a Random Forest revealed that PM2.5 levels, ozone levels, and smoking prevalence were big impacting factors in predicting COPD hospitalization rates. With our GLMs, we were able to catch the effects of these predictors, as PM2.5 showed strong association with COPD hospitalizations. The Random Forest, however, had a much higher accuracy. But, it was hard to interpret compared to using GLMs. Though it outperformed GLMs in accuracy, we learned the trade-offs between using two different models and how they handle dataset complexity. Using causal inference, we looked into the effect of high PM2.5 exposure on COPD hospitalization rates by using outcome regression and inverse propensity weighting (IPW). For outcome regression, it estimated an ATE of about 4 additional hospitalizations per 1,000 cases within states with high PM2.5 levels. The IPW estimated a lower ATE of about 3.44 hospitalizations. Utilizing confounders such as smoking prevalence, ozone levels, CO2 emissions per capita, and median income were important as they are influential to PM2.5 exposure and COPD outcomes.

Critical Evaluation: Our dataset was limited to only using information during 2011-2014 and also omitted important factors such as weather patterns. These factors could’ve been important as they can heavily impact air quality as well. Our PM2.5 levels were also only estimates and not real, direct numbers; therefore, we had potential measurement error.

We also missed out on studying regional healthcare access and seeing how that can impact COPD outcomes. If given the opportunity to ask a domain expert, we'd ask, "How can differences in healthcare access across the country affect COPD hospitalization rates?" Knowing if there was a correlation between the two could've potentially allowed us to make more informed decisions about our model, one being possibly using the healthcare metrics as confounders.

In the end, our findings can be somewhat robust due to the choices we made when picking models. Because GLMs assume linear relationships, it could've oversimplified our factors and how they interact. We could have used a generalized additive model (GAM) to catch nonlinear connections, but this could've caused complications and changed our estimates. Overall, our results act as a baseline in understanding the relationship between PM2.5 levels, ozone levels, and smoking prevalence in predicting COPD hospitalization rates.

VII. Sources

1. Kariisa, M., Foraker, R., Pennell, M., Buckley, T., Diaz, P., Criner, G. J., & Wilkins, J. R. (2014). Short- and Long-Term Effects of Ambient Ozone and Fine Particulate Matter on the Respiratory Health of Chronic Obstructive Pulmonary Disease Subjects. Archives of Environmental & Occupational Health, 70(1), 56–62. https://doi.org/10.1080/19338244.2014.932753
2. Pisoni, E., Volta, M., & Carnevale, C. (2011). A decision support system for air quality planning: The case study of Lombardy Region (Italy). *Progress in Health Sciences*, *1*(2), 144–155. Retrieved from https://www.umb.edu.pl/photo/pliki/progress-file/phs/phs_2011_2/144-155.pdf
3. Liu, Y., Lee, M., & Wang, X. (2014). Ethnic differences in exposure to air pollution and associated respiratory health outcomes: A review. *Environmental Health Perspectives*, *122*(1), 4–10. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC3925402/

Word Count: 4970