Data 102, Spring 2024
Final Written Report

Group 16: Karen Ting, Janelle Correa, Ricardo Reyes, Danielle Ma

1 Research Questions

Q1: How do levels of PM2.5 and ozone affect the onset of chronic illnesses - specifically asthma and COPD?

Answering this question has significant implications for real-world decisions and policies related to air pollution and public health. The findings could inform air quality standards, guiding the refinement of ozone and PM2.5 limits to better protect against the onset of chronic illnesses like asthma and COPD. Moreover, the results could guide public health campaigns, educating high-risk groups about air quality, health risks, and preventative measures, empowering them to take action. Additionally, the research could optimize healthcare resource allocation, distributing resources more effectively to areas with higher asthma and COPD prevalence, informed by air pollution data. Finally, the study's conclusions could shape policy initiatives, informing legislation and programs aimed at reducing air pollution and promoting public health, such as the Clean Air Act, to create a healthier environment for all. By addressing this question, we can develop evidence-based solutions to mitigate the impact of air pollution on public health, ultimately improving the well-being of communities and individuals alike.

Q2: To what extent can ozone and PM 2.5 levels predict asthma prevalence rates, and what are the key factors that modify this relationship? Specifically, are lower income counties more susceptible to higher prevalence?

The findings of this question could be used as evidence for change in policy, that would prove that specific communities need more resources to treat asthma related diseases. Furthermore these findings could also mean for a change in systems that contribute towards the ozone and PM 2.5 levels that negatively affect communities. This could lead to less pollution and environmental harm, overall increasing the air quality of affected areas. In turn reducing the air pollution would have a positive impact on those who have asthma related diseases. The method of using a GLM, specially a Poisson GLM, was an optimal choice because of the proximity it gave us for our predictions which we found were about 35% off. Poisson was a better fit for our question because most of our prevalence data was in terms of rates and rates work very well with Poisson GLMs. Nevertheless, it's important to note that having a good model for predicting the probability of asthma prevalence is difficult and has notable limitations. One limitation is that it is difficult to keep account of the permanent residence of the entire population as some might face housing insecurity and might not have a permanent address and thus could be partially excluded from our datasets. Another notable limitation is that it is limited to the state of California, which might have different confounders that other states might not have including

2 Data Overview

We worked with Dataset 1: Chronic Disease and Air Quality. The ozone and PM2.5 Concentration datasets were generated through modeled predictions of the levels from the EPA's Downscaler model, utilizing data at the census tract level from 2011-2014. The asthma and Chronic Obstructive Pulmonary Disease (COPD) datasets were filtered from the 2023 release of the U.S. Chronic Disease Indicators (CDI) dataset, a comprehensive nationwide census dataset based on 124 indicators for each state.

Research Q1 External Datasets

We supplemented our causal inference analysis with external datasets to gain a more comprehensive understanding of ozone and PM2.5 trends. We utilized datasets with aggregated annual values spanning a longer period (1980-2022 for ozone and 2000-2022 for PM2.5), sourced directly from the U.S. Environmental Protection Agency (EPA) website. These datasets were generated through a nationwide census of air quality monitoring sites, providing annual values and a longer time frame that offered a more robust basis for our analysis.

To address a potential instrumental variable in our casual relationship analysis, we utilized the Contiguous U.S. Average Temperature dataset from the National Centers for Environmental Information (NCEI). This dataset, generated through a suite of monitoring services that track key climate indicators, covers the entire contiguous U.S. from 1895 to the present. For our purposes, we filtered and downloaded the data directly on the website, selecting only the relevant years (2010, 2013-2018 to match the years available for asthma/COPD rates) and the corresponding average annual U.S. temperature. As we included all available U.S. temperature data from the specified time period, this dataset represents a census rather than sample.

When interpreting our models and inferences, several key notes should be considered. The EPA identified data quality issues in PM2.5 data from Florida, Illinois, Tennessee, and Kentucky, which were excluded from aggregated yearly values. The average temperature dataset only includes contiguous U.S. territories, excluding territories like Alaska, Hawaii, and Puerto Rico. The CDC's ozone and PM2.5 concentration datasets have a high level of detail, with data collected at specific monitoring stations; each row represents a data point for a specific monitoring station. In contrast, the annual air quality and temperature datasets have a national-level granularity, representing the average values across states for each year. Meanwhile, the asthma and COPD datasets have a state-level granularity, with each row representing a data point for the chronic disease and question at the state level. These notes are crucial to consider, as they may impact the accuracy and generalizability of our findings. For instance, national-level granularity could mask local variations in ozone and PM2.5 levels, while state-level granularity might not capture variations within states. This could result in our models being more accurate for certain regions within the U.S.

Research Q2 External Datasets

Other externally sourced datasets we used included: Income county, County names and Asthma rates. These datasets allowed us to determine predictors for asthma rate predictions because they provide predictors. For the purpose of our datasets, we have filtered each to only display data for

CA. This was primarily done to draw more specific conclusions while still having access to a large amount of data. The income county dataset is an official dataset acquired from the US Department of Health and Human Services which gathers data from hospitals, laboratories and other reliable sources. Nevertheless, it is important to acknowledge that this dataset could be systematically excluding individuals that are homeless and do not have a fixed address, as well as those without any form of insurance. The granularity of the dataset is at the county level in California. In summary, the rows in our dataset set are represented as:

County: The name of the county in California.

Value: This is the median income per county in California

Rank: The rank of the county based on the "Value" column.

FIPS: The Federal Information Processing Standards code, which is a unique identifier for each county.

One objective that the income county dataset will solve is to provide us insight for which counties have higher incomes. We have also brought to light concerns such as selection bias, measurement error and convenience sampling, which are not present given that the data was acquired from a gubermental resource. Our eventual goal was to merge this data with the provided Ozone and PM 2.5 datasets. Given however that these two datasets were quite large, and we only had income data from California, we had lots of cleaning and preprocessing of data to deal with. Like previously said, since our original file sizes were too large to run consistently in our jupyterhub notebook we narrowed down our Ozone and PM datasets to counties just in California while keeping the remaining columns. As a result, we were then able to craft a question with more specificity and were prepared to answer our question with more detail. Given this context, the income county dataset's interpretation goes hand in hand with the county name dataset. The county name data set was external and downloaded from the National Weather Service. The county name dataset allows us to determine which unique identification codes belong to a specific county, which provides us a standardized unit that different sets use compared to just the name of the county .Since the county name dataset is filtered by counties only in the state of California, it ultimately forces us to only draw conclusions of asthma rate predictors in CA going hand-in-hand with the county income, PM 2.5 and ozone level datasets. The county names were assigned a code by the government of California and therefore, there is no significant need to consider groups that were systematically excluded. The main components of this dataset are the FIPS code, which is a unique identification substituting the name for the counties in California. To the right of the FIPS code is the column for the name of the county. Moving back to the county income dataset, a couple of columns we wish the dataset would have is the average cost of living in the county as well as the minimum wage because these could be variables that may suggest there is a stronger correlation between income, different cost of living and minimum wage. This would give us more freedom to use a more traditional GLM model that would account for more variables to draw a conclusion to a potential correlation.

3 EDA

Research Q1 EDA

1.1: Hospitalization Rates for Asthma and COPD per Year (Categorical Variable)

For both COPD and asthma datasets, relevant columns include "YearStart", "Question", "DataValueType", and "DataValue". "YearStart" represents the year when the value was recorded. Then, the data is grouped by ”YearStart”, ”DataValueType”, and ”Question”, with the sum of values calculated within each group. Subsequently, the grouped data is sorted in ascending order of ”YearStart” and the index is reset to ensure proper alignment of data. For the asthma dataset, a subset of the grouped data is then filtered to include only records where ”DataValueType” is ’Number’ and ”Question” is ’Hospitalizations for asthma’. This step ensures that the analysis focuses specifically on the number of hospitalizations for asthma, disregarding other types of data values or questions. For the COPD dataset, a subset of the grouped data is then filtered to include only records where ”DataValueType” is ’Number’ and ”Question” is ’Hospitalization for chronic obstructive pulmonary disease as first-listed diagnosis’. The final tables we got include data values in 2011 and 2013 to 2018.

The filtered subset of data is then prepared for visualization. Using Matplotlib, two bar plots are generated, with year on the x-axis and ”DataValue” (number of hospitalizations for asthma/COPD) on the y-axis. The plots provide a visual representation of the temporal trend in asthma/COPD hospitalizations, allowing for observation of any patterns, fluctuations, or trends over time. Key features are the number of asthma/COPD hospitalizations.

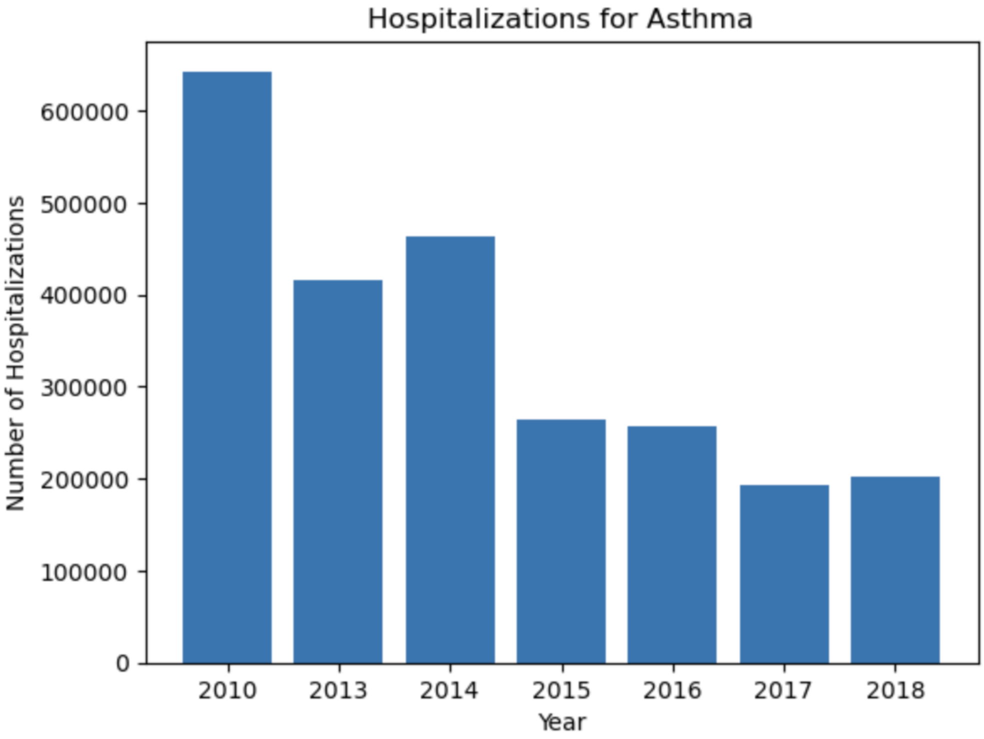


Figure 1: Number of asthma hospitalizations by year

The first graph shows the number of hospitalizations on asthma in year 2010, 2013, 2014, 2015, 2016, 2017, and 2018. Other years that are not shown in this graph are missing data value. 2010 has the highest number of hospitalization, with year 2014 follows. Then from 2015, there appears a decreasing trend. Based on this plot, we would like to see whether levels of ozone and PM2.5 in each year affect number of hospitalizations(yearly) for asthma. More specific, does there appear a causal relationship between levels of ozone and PM2.5 and chronic disease(asthma). One potential answer would be higher levels of ozone and PM2.5 will cause higher number of hospitalizations for asthma.

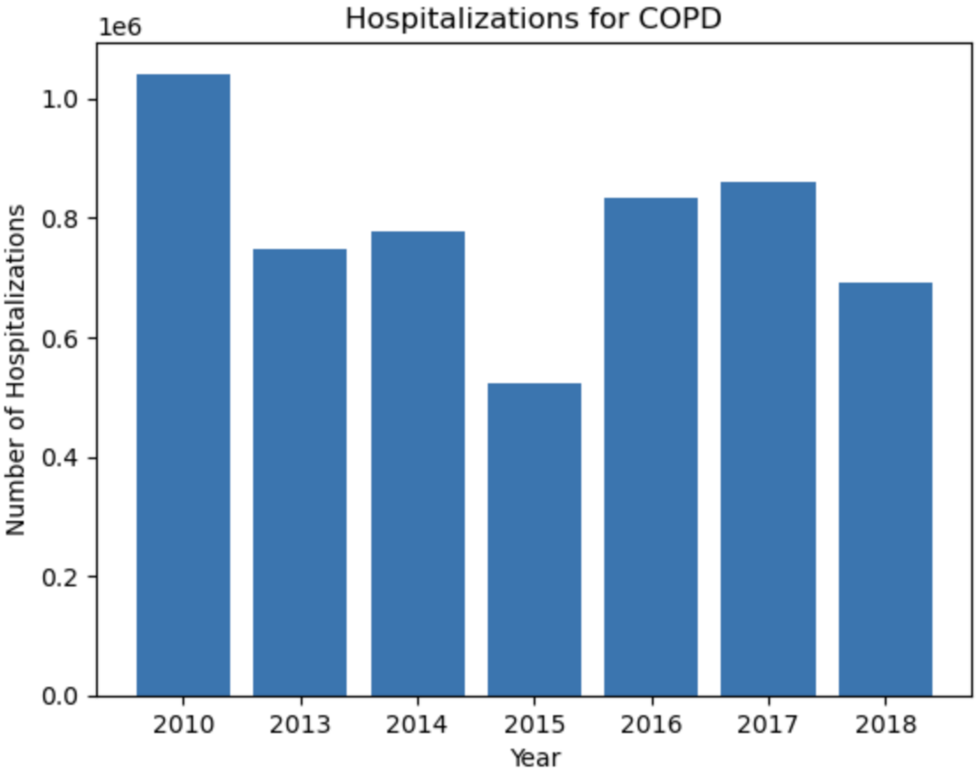


Figure 2: Number of Hospitalization for COPD

The second graph shows the number of hospitalizations on COPD in year 2010, 2013, 2014, 2015, 2016, 2017, and 2018. Other years that are not shown in this graph are missing data value. 2010 has the highest number of hospitalization, with year 2015 being the lowest. An obvious trend does not shown in the graph. Based on this plot, we would like to see whether levels of ozone and PM2.5 in each year affect number of hospitalizations(yearly) for COPD. More specific, does there appear a causal relationship between levels of ozone and PM2.5 and chronic disease(COPD). One potential answer would be higher levels of ozone and PM2.5 will cause higher number of hospitalizations for COPD.

1.2: Ozone and PM2.5 Levels over Time (Quantitative Variable)

The calculations for the ozone and PM2.5 graphs below were created by first averaging the different predictions given for specific dates. Then an average monthly was made from the average date predictions. It's important to note that Simpson's Paradox could be in play from aggregating the data several times, potentially hiding further trends or creating false trends. However, for this EDA portion, we're only searching for potential areas for further examination.

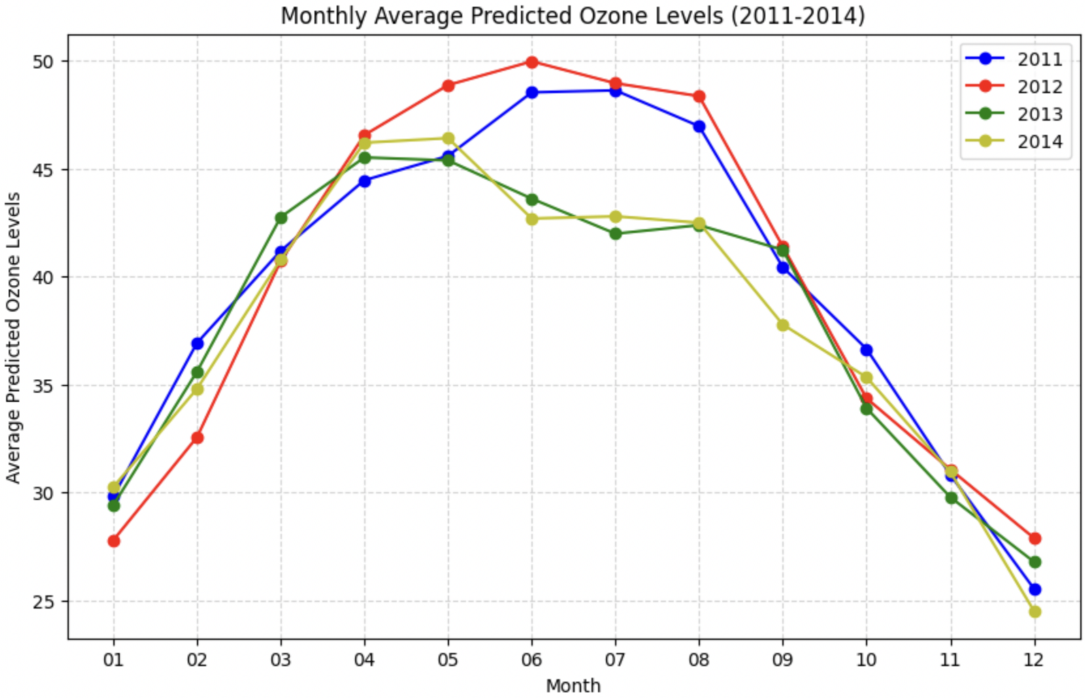


Figure 3: Predicted Ozone Levels over Time

From the graph above, we can observe and compare trends in the monthly average predicted ozone levels over 2011-2014. A clear pattern is a seasonal fluctuation seen in the general ozone increase from January-April and decrease from August-December. A more striking difference can seen between between 2011/2012 and 2013/2014 ozone levels during the summer months of June-August. The higher levels in 2011/2012 compared to 2013/2014 suggest a potential shift in ozone concentration patterns over these two-year intervals. By comparing asthma and COPD cases during these specific intervals, we can further investigate whether there's an change in prevalence/severity that correspond to the ozone shift that we've observed.

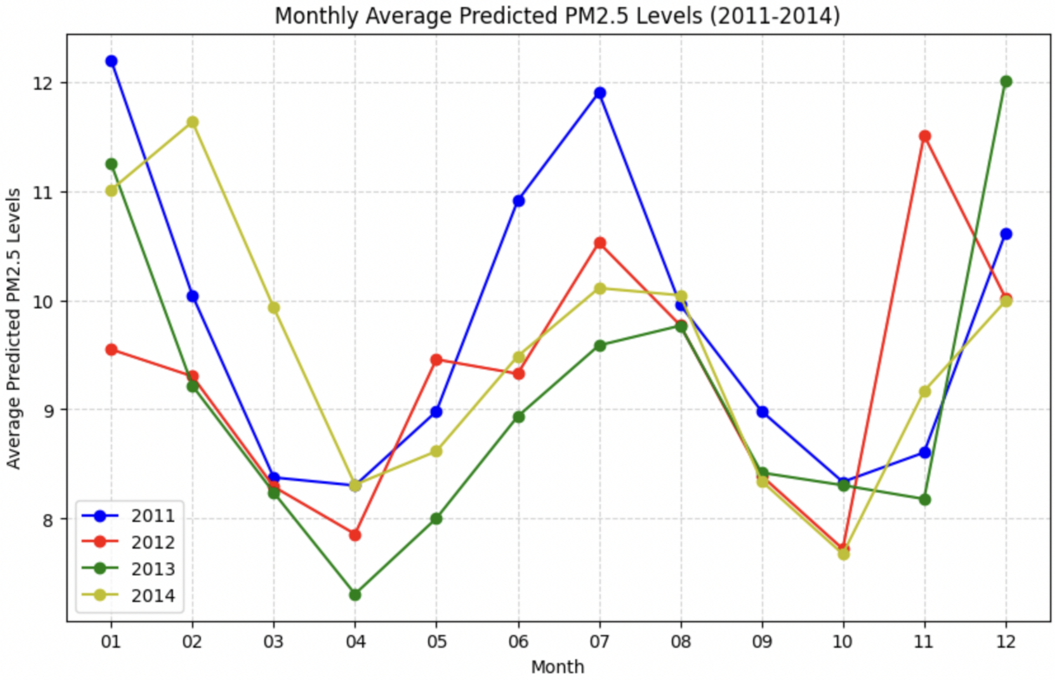


Figure 4: Predicted PM2.5 Levels over Time

From the graph above, we can observe and compare trends in the monthly average predicted PM2.5 levels over 2011-2014 to answer the same questions as we did for ozone levels.

The PM2.5 graph shows a more variable pattern with several peaks and troughs throughout the year. There is, however, a general decrease from February-April, increase from April-July, decrease from August-October, and increase from October-December across 2011-2014. There is also a significant spike in levels between October to November in 2012 and November to December in 2013.

By comparing asthma and COPD cases during these specific intervals, we can investigate whether there's an change in prevalence/severity that correspond to the PM2.5 shift that we've observed. Investigating and comparing during the specific monthly spikes may also be very helpful.

Research Q2 EDA

2.1: Geographical Loation and Exposure Level (Categorical Variable)

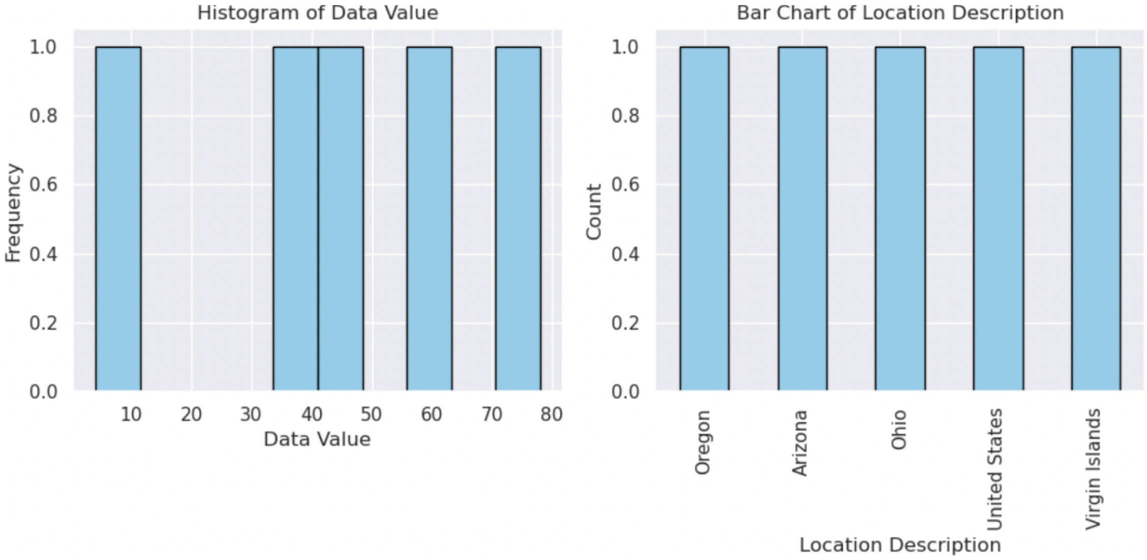


Figure 5: Enter Caption

For research question 2, we observed the relationship of categorical variables of geographical location, specifically states, vs PM 2.5 level which is a quantitative variable. A relationship I observed in this visualization was the relationship between counties and their exposure level. It is also worth noting that the frequency of specific diseases could vary from state to state. This motivates the question of finding factors that influence higher rates of disease in specific regions. This also sparked my curiosity to know if there is a similar relationship that varies from different counties in the same state. In light of this information we are prepared to find out if PM2.5 exposure levels have a correlation with asthma prevalence rates.

2.2: Median Income and Ozone Levels (Quantitative Variable)

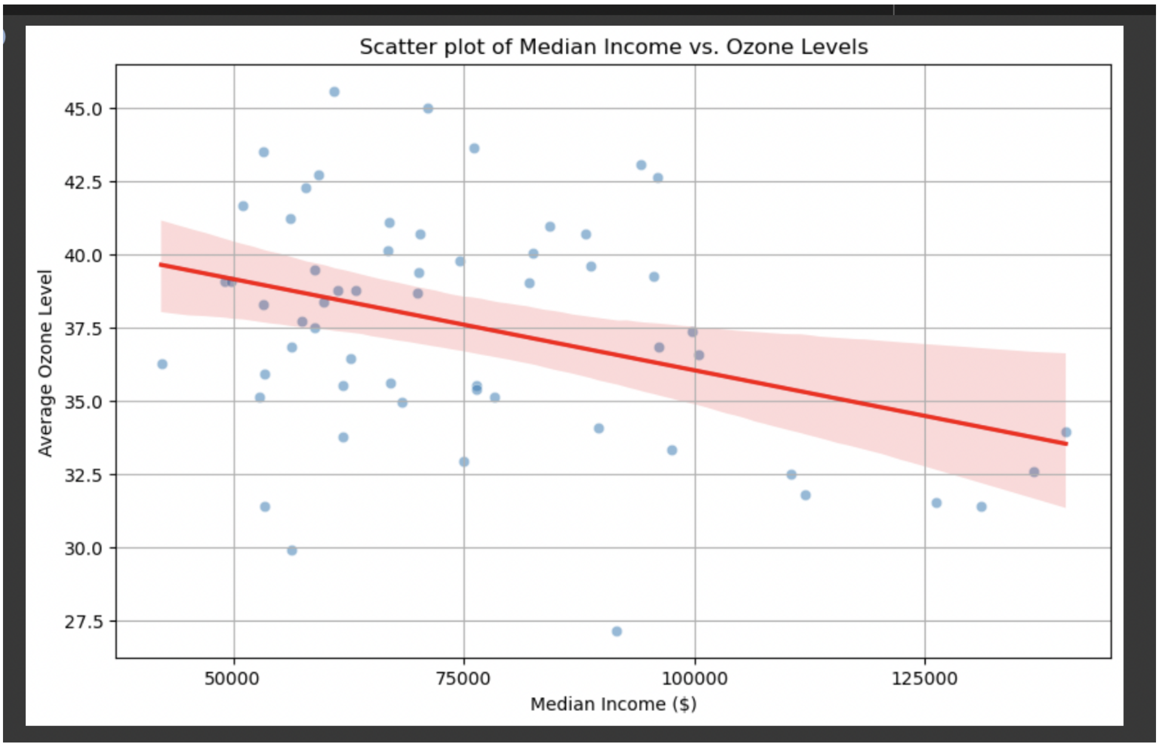


Figure 6:

For our second research question, we observed the quantitative variables of median income and ozone levels. From our scatter plot (Fig 1) we were able to see that a county's median income is correlated with a lower average ozone level. In fact, it appears that the correlation is more significant the closer higher incomes are. This motivates the question of whether or not counties with lower incomes are more likely to have higher rates of asthma due to higher ozone levels. In addition, if we find that such correlation exists and is strong we can then determine if these factors among others are good predictors to determine the likelihood of people in certain counties being hospitalized for an asthma related issue.

4 Q1: Causal Inference

Method:

We hypothesized a causal relationship between PM2.5 and ozone levels (treatment) and the onset of chronic respiratory illnesses like asthma and COPD (outcome). To test this hypothesis, we employed the causal inference technique, leveraging temperature (annual average temperature in the U.S.) as an instrumental variable to address potential biases and confounding factors. Since temperature affects hospitalizations through its influence on PM2.5 and ozone levels, we assumed no confounders in this research question. Additionally, colliders were not included in the dataset. By using temperature as an instrumental variable, we can effectively address potential biases and confounding factors that might obscure the true causal effect.

Our analysis focused on the time frame: 2011, and 2013 to 2018. Due to missing hospitalization data for other years in the asthma and COPD datasets, we limited it to these years. The chronic illness datasets were then grouped by year, question, and data value type. There were several different questions asked in the dataset, such as “emergency department visit rate for asthma” and “asthma mortality rate”. We chose to focus solely on “hospitalizations” because we believed it to be a more comprehensive and reliable metric for assessing air pollution impact. By focusing on hospitalizations, we aimed to capture the most severe cases of asthma and COPD exacerbations that resulted in inpatient care. After grouping, we summed the values of the groups to find an overall number for the different stratification groups and different location groups.

Additionally, we extracted the annual average U.S. temperature data, PM2.5 levels data, and ozone levels data for the same time frame (2011 and 2013-2018). We then merged the temperature data, the chronic illness datasets, and the PM2.5 and ozone level datasets to create a comprehensive dataset for analysis.

To test the causal relationship between PM2.5 and ozone levels and asthma/COPD hospitalizations, we applied a two-stage least squares (2SLS) regression analysis for each relationship. Using asthma as an example, we first constructed separate ordinary least squares (OLS) regression models to estimate the associations between temperature and PM2.5/ozone levels. These models, once fitted to the data, generated predicted values of PM2.5 and ozone, which were then incorporated into a new dataset for the second stage of the analysis. In the second stage, we constructed another OLS regression model to examine the relationship between asthma hospitalizations and the predicted PM2.5 and ozone values. From this second model, we then extracted a summary of the regression results. Overall, this approach enabled us to assess the impact of air quality (PM2.5 and ozone) on asthma hospitalizations while mitigating potential biases that are brought by temperature. We then followed the same process to test COPD hospitalizations and PM2.5/ozone levels.

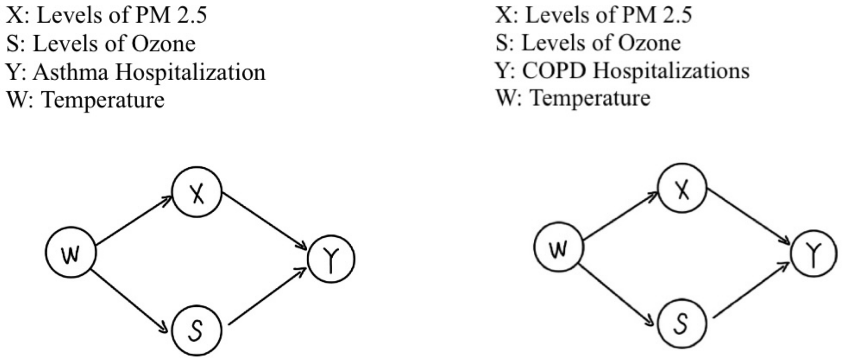
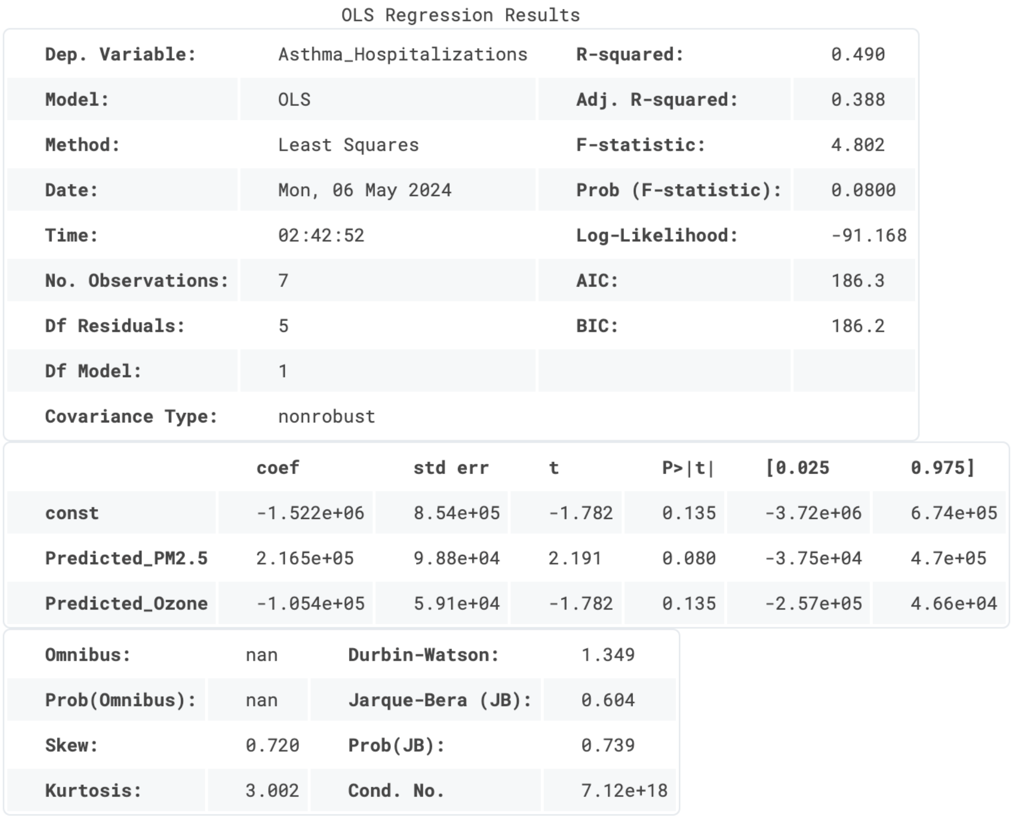


Figure 7: DAG

Results:



OLS Regression Results

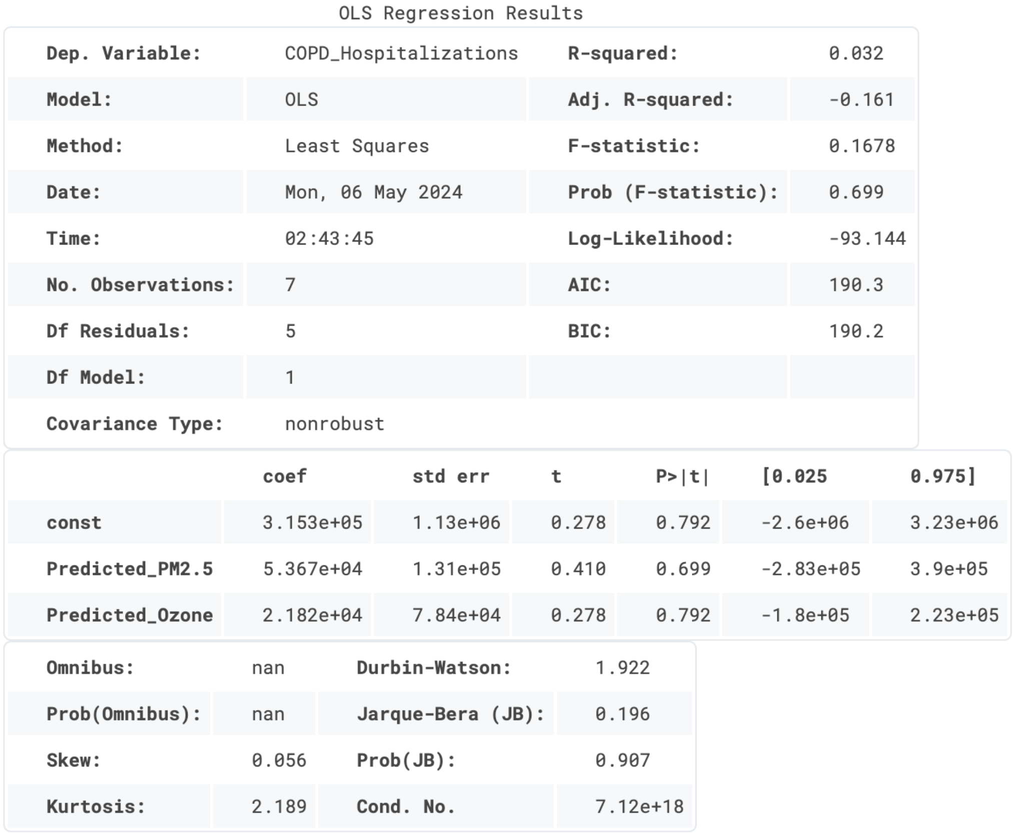
Dep. Variable:	Asthma_Hospitalizations	**R-squared:**	0.490
Model:	OLS	**Adj. R-squared:**	0.388
Method:	Least Squares	**F-statistic:**	4.802
Date:	Mon, 06 May 2024	**Prob (F-statistic):**	0.0800
Time:	02:42:52	**Log-Likelihood:**	-91.168
No. Observations:	7	**AIC:**	186.3
Df Residuals:	5	**BIC:**	186.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-1.522e+06	8.54e+05	-1.782	0.135	-3.72e+06	6.74e+05
Predicted_PM2.5	2.165e+05	9.88e+04	2.191	0.080	-3.75e+04	4.7e+05
Predicted_Ozone	-1.054e+05	5.91e+04	-1.782	0.135	-2.57e+05	4.66e+04

Omnibus:	nan	**Durbin-Watson:**	1.349
Prob(Omnibus):	nan	**Jarque-Bera (JB):**	0.604
Skew:	0.720	**Prob(JB):**	0.739
Kurtosis:	3.002	**Cond. No.**	7.12e+18

Figure 8: Asthma

First, the OLS summary of asthma. The F-statistic of 4.802 with a corresponding p-value of 0.080 suggests that the overall model is statistically significant at the 0.1 significance level, indicating that at least one of the predictors has a non-zero coefficient. The coefficient for predicted PM2.5 stands at 2.165e+05 with a standard error of 9.88e+04. This implies that for every unit increase in predicted PM2.5 levels, there's an estimated increase of approximately 2.165e+05 units in asthma hospitalizations, with a marginal significance level of 0.080 (p-value). The coefficient for predicted ozone, at -1.054e+05 with a standard error of 5.91e+04, suggests a decrease of around 1.054e+05 units in asthma hospitalizations per unit increase in predicted ozone levels, though again, the significance level is marginal at 0.135 (p-value). Both the p-values are greater than 0.05, which means they are not statistically significant. As a result, we can not conclude that levels of PM2.5 and ozone have a causal effect on the onset of asthma.



OLS Regression Results

Dep. Variable:	COPD_Hospitalizations	**R-squared:**	0.032
Model:	OLS	**Adj. R-squared:**	-0.161
Method:	Least Squares	**F-statistic:**	0.1678
Date:	Mon, 06 May 2024	**Prob (F-statistic):**	0.699
Time:	02:43:45	**Log-Likelihood:**	-93.144
No. Observations:	7	**AIC:**	190.3
Df Residuals:	5	**BIC:**	190.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3.153e+05	1.13e+06	0.278	0.792	-2.6e+06	3.23e+06
Predicted_PM2.5	5.367e+04	1.31e+05	0.410	0.699	-2.83e+05	3.9e+05
Predicted_Ozone	2.182e+04	7.84e+04	0.278	0.792	-1.8e+05	2.23e+05

Omnibus:	nan	**Durbin-Watson:**	1.922
Prob(Omnibus):	nan	**Jarque-Bera (JB):**	0.196
Skew:	0.056	**Prob(JB):**	0.907
Kurtosis:	2.189	**Cond. No.**	7.12e+18

Figure 9: COPD

Similarly, for the summary of COPD, the F-statistic of 0.1678, along with its associated p-value of 0.699, indicates that the overall model is not statistically significant at conventional levels, suggesting that neither predicted PM2.5 nor ozone levels significantly predict COPD hospitalizations. Examining the coefficients, neither predicted PM2.5 nor ozone levels exhibit statistically significant relationships with COPD hospitalizations, as indicated by their respective p-values of 0.699 and 0.792. We can not conclude that levels of PM2.5 and ozone have a causal effect on the onset of COPD.

Overall, while the model for asthma hospitalizations demonstrates a moderate explanatory power with an R-squared value of 0.490, the adjusted R-squared value of 0.388 suggests that the model may not fully capture all relevant factors influencing asthma hospitalizations. This discrepancy indicates potential uncertainty in the estimate and suggests that there may be unobserved variables or omitted factors that could affect the relationship between air pollutants and asthma hospitalizations. In the analysis of COPD hospitalizations, the OLS results reveal limited explanatory power of the model, with an R-squared value of only 0.032. Furthermore, the adjusted R-squared value of -0.161 indicates potential overfitting or inadequate inclusion of relevant predictors in the model.

Discussion:

We acknowledge several limitations in our methods. First, our dataset was limited to a specific time frame (2011, 2013-2018) and lacked individual-level data, which may have oversimplified the complexity of respiratory health outcomes. Second, we relied on aggregated national data, which may mask regional variations in the examined relationships. Third, we focused on hospitalization rates, omitting emergency department rates and mortality rates, which may have oversimplified the complexity of the relationships and limited the generalizability of our findings. By omitting these outcomes, we may have missed important nuances in the relationship between air pollution and chronic illness health outcomes as there could be different causal pathways based on the outcome. Most notably, our method relied on a simplifying assumption of no confounders, and the use of temperature as an instrumental variable may not have fully captured the complexities of the relationship.

Based on our results, we are unable to conclude a causal relationship between PM2.5/ozone levels and asthma/COPD hospitalizations due to the non-significant p-values and moderate explanatory power of the models. While our results suggest a potential association, other factors may be at play and our method's limitations may have influenced the results. Therefore, we recommend further investigation with more robust methods and additional data to better account for confounding factors and provide a more definitive answer to this research question. One such method is propensity score matching, which would allow us to create more comparable groups and better control for confounding variables. We would then ideally have data spanning a longer time period, along with more detailed hospitalization information (e.g. readmission rates) and individual-level characteristics (e.g. smoking status) to increase the accuracy of the analysis.

5 Q2: GLM and Nonparametric methods

Method:

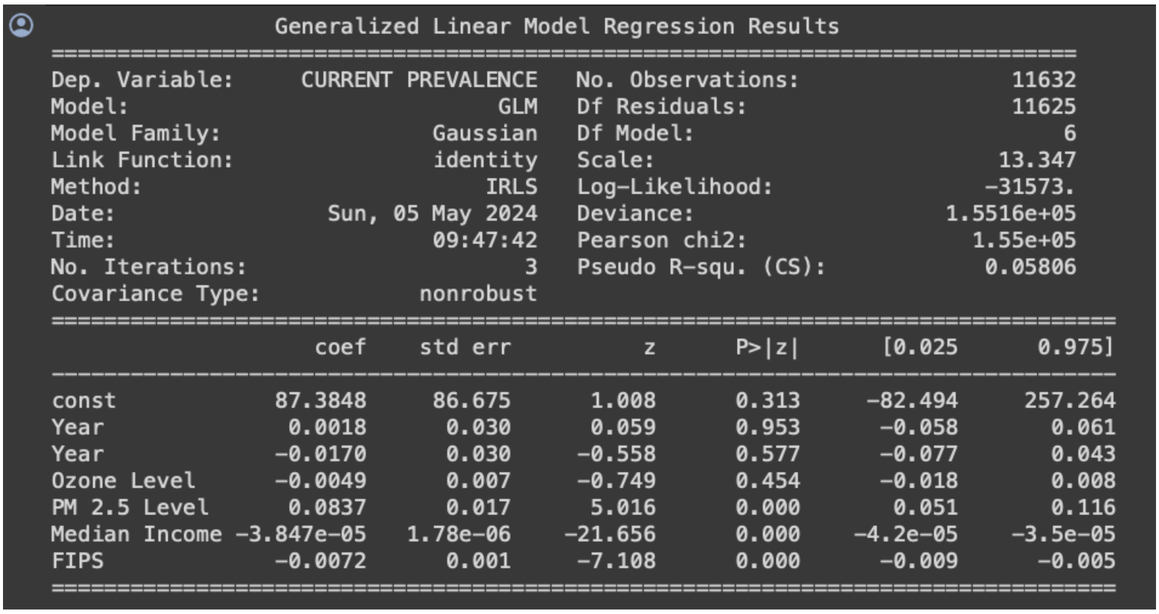
We are trying to predict asthma prevalence rates, by analyzing the relationship of ozone and PM 2.5 exposure levels , specifically in lower income communities. Some features we used to assist us to execute our method were ozone levels, PM 2.5 levels, median income, and FIPS codes of counties. We chose these features because income level is a key socioeconomic factor that determines access to healthcare. Income level can also determine quality housing and therefore could lead to poor air pollution which has a relationship to asthma rates.

We first began by executing a GLM with a Poisson distribution in a frequentist setting. Since Poisson is utilized for count predictions we chose this because of the count nature of asthma cases. The assumptions include the mean-variance relationship inherent in Poisson models. We also decided that the best link function for a Poisson regression was the identity link function. Our next step was to execute nonparametric methods which were Random Forest because of its ability to capture nonlinear relationships among variables and Neural Networks because of the flexibility as they do not make strong assumptions for underlying data distribution.As we include ozone, PM2.5 levels, county income and asthma prevalence rates which are going to be included in model interactions, neural networks are able to learn from these factors. We plan to evaluate the model performance using Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) to quantify prediction accuracy and error magnitude.

Results: As we analyzed a nonparametric method for estimating Asthma prevalence, we found that, in our Random Forest model, we can see that there is a Mean Squared Error (MSE) of 12.09 and a Mean Absolute Percentage Error of 34.45%. Given these outputs, we are able to draw that

FIPS codes had a high importance in the model.

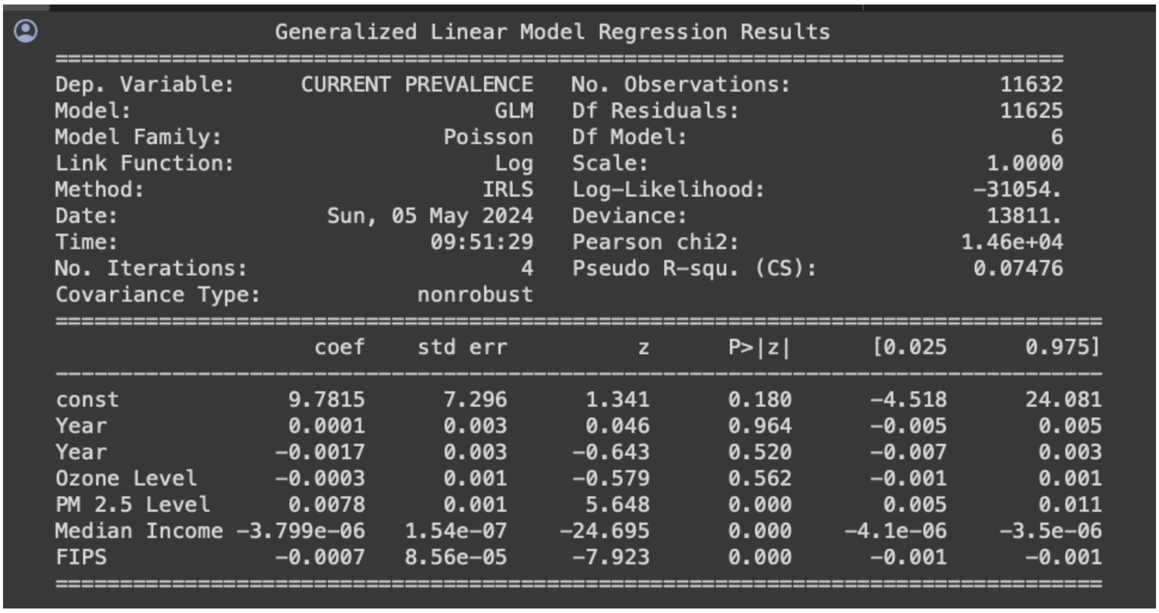
In our first GLM, the results were a MSE of 13.34. The coefficients of the GLM model indicated that median income and FIPS codes were significant predictors of asthma prevalence. When we implemented a Nerual Network model it resulted in a MSE of 16.79, which is insightful but has a slightly higher MSE.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:     CURRENT PREVALENCE   No. Observations:                11632
Model:                            GLM   Df Residuals:                    11625
Model Family:                Gaussian   Df Model:                            6
Link Function:               identity   Scale:                          13.347
Method:                          IRLS   Log-Likelihood:                -31573.
Date:                Sun, 05 May 2024   Deviance:                   1.5516e+05
Time:                        09:47:42   Pearson chi2:                 1.55e+05
No. Iterations:                     3   Pseudo R-squ. (CS):            0.05806
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const            87.3848     86.675      1.008      0.313     -82.494     257.264
Year              0.0018      0.030      0.059      0.953      -0.058       0.061
Year             -0.0170      0.030     -0.558      0.577      -0.077       0.043
Ozone Level      -0.0049      0.007     -0.749      0.454      -0.018       0.008
PM 2.5 Level      0.0837      0.017      5.016      0.000       0.051       0.116
Median Income -3.847e-05   1.78e-06    -21.656      0.000     -4.2e-05    -3.5e-05
FIPS             -0.0072      0.001     -7.108      0.000      -0.009      -0.005
=================================================================================
```

Figure 10: First GLM

Since we found that rates would be used at a significant amount, we implemented a second GLM model that used a Poisson Regression. In terms of uncertainty in the GLM prediction, the standard errors provided in the GLM summary can be used to construct confidence intervals for the coefficients, providing a quantitative measure of uncertainty. For example, the 95% confidence interval for the coefficient of Median Income ranges from -4.2e-05 to -3.5e-05, indicating a level of uncertainty in this estimate.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:     CURRENT PREVALENCE   No. Observations:                11632
Model:                            GLM   Df Residuals:                    11625
Model Family:                 Poisson   Df Model:                            6
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -31054.
Date:                Sun, 05 May 2024   Deviance:                       13811.
Time:                        09:51:29   Pearson chi2:                 1.46e+04
No. Iterations:                     4   Pseudo R-squ. (CS):            0.07476
Covariance Type:            nonrobust
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            9.7815      7.296      1.341      0.180      -4.518      24.081
Year             0.0001      0.003      0.046      0.964      -0.005       0.005
Year            -0.0017      0.003     -0.643      0.520      -0.007       0.003
Ozone Level     -0.0003      0.001     -0.579      0.562      -0.001       0.001
PM 2.5 Level     0.0078      0.001      5.648      0.000       0.005       0.011
Median Income -3.799e-06   1.54e-07    -24.695      0.000    -4.1e-06    -3.5e-06
FIPS            -0.0007   8.56e-05     -7.923      0.000      -0.001      -0.001
================================================================================
```

Figure 11: Second GLM

In summary the results suggest that the FIPS, PM 2.5 Level, and Median Income are the most important features in predicting asthma prevalence according to the Random Forest model. The GLM and Neural Network models also provide valuable insights, but with slightly higher MSEs. The relatively high MAPE for the Random Forest model suggests that there may be room for improving the model or that the relationship between the features and the target variable may not be fully captured by the model.

Discussion:

The Random Forest model performed the best based on the Mean Squared Error (MSE), with a lower MSE (12.09) compared to the GLM (13.34) and Neural Network (16.59) models. This suggests that the Random Forest model may be more accurate in predicting asthma prevalence rates. However, the performance of these models on future datasets would depend on the consistency of the data patterns and may require retraining.

The Random Forest model fits the data best as indicated by the lowest MSE. However, all models have relatively high Mean Absolute Percentage Errors (MAPE), which would mean that the models would be retrained for a potential better prediction.

The feature importance from the Random Forest model suggests that FIPS codes (which could be a proxy for geographical location) are the most important predictor of asthma prevalence. The GLM model indicates that median income and FIPS codes are significant predictors. This suggests that socioeconomic and geographical factors play a significant role in asthma prevalence.

The GLM model assumes a linear relationship between the log of the outcome and the predictors, which may not hold true. The Random Forest model, while able to capture complex relationships, may overfit the data and does not provide as clear interpretability as GLM. We found that the

Neural Network model requires careful selection of architecture and hyperparameters, and also lacks interpretability. We also found that additional data that could be useful might include other environmental factors such as access to healthcare or socioeconomic data such as education levels. Given our MAPE values, the uncertainty could be considered relativity high. One reason why this could be could be due to various factors such as noisy data, the complexity of the relationships being modeled, and the inherent variability in health outcomes like asthma prevalence.

6 Conclusion

In conclusion, our study explored the relationship between air pollution and chronic illnesses, with a focus on asthma prevalence rates. Although we did not establish a causal link between air pollution and chronic illnesses, our findings suggest a potential association that warrants further investigation. Additionally, our GLM model successfully predicted asthma incidents, but we faced challenges in accurately predicting asthma prevalence rates due to limitations in data size, scope, and complexity. To draw stronger conclusions, we need to incorporate more variables and look into more potential confounders.

Our study's limitations included the small data size and geographical scope of our data, as well as the constraints of our coding environment. Nevertheless, we gained valuable insights into the impact of income on ozone and PM2.5 levels, and the accuracy of GLM models compared to non-parametric methods. We also developed practical skills in working with large datasets, optimizing memory and efficiency, and identifying appropriate external datasets. By merging different data sources, we were able create a more comprehensive dataset that enabled better in-depth analysis; however, a limitation of this approach is that a lack of data on certain topics (i.e. asthma and COPD) constrained the time frame we could investigate. More personally, we learned that asking a good question is key to exercising our skills as data scientists. We also discovered that building a simple model can be valuable, as it allows for easier debugging and refinement, whereas more complex models can be challenging to interpret and troubleshoot.

Our results highlight the need for further research into the causal relationship between air pollution and chronic illnesses, as well as the importance of refining our model to be able to determine if lower-income counties are more susceptible to higher ozone/PM2.5 levels. By addressing these limitations and exploring alternative approaches (such as propensity score matching), future studies can build on our findings and the results can potentially inform policy decisions aimed at reducing air pollution and improving public health. For instance, our results could be used to support policies that target air pollution reduction in lower-income areas, ultimately contributing to better health outcomes and quality of life.

7 References

California Health and Human Services Open Data Portal. (n.d.). *Asthma hospitalization rates by county*. Retrieved May 3, 2024, from https://data.chhs.ca.gov/dataset/asthma-hospitalization-rates-by-county

National Institute on Minority Health and Health Disparities. (n.d.). *Data Portal: Social Determinants - Income Level, California*. Retrieved May 4, 2024, from https://hdpulse.nimhd.nih.gov/data-portal/social/table?socialtopic=030&socialtopic_options=social_6&demo=00011&demo_options=income_3&race=00&race_options=race_7&sex=0&sex_options=sexboth_1&age=001&age_options=ageall_1&statefips

=06&statefips_options=area_states

National Weather Service. (n.d.). *California Fire Weather*. Retrieved May 1, 2024, from

https://www.weather.gov/hnx/cafips

NOAA National Centers for Environmental Information. (n.d.). Climate at a Glance. Retrieved from May 2, 2024, from https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/national/time-series/110/tavg/ytd/12/2000-2023?base$_prd = truebegbaseyear = 1901endbaseyear = 2000$

U.S. Environmental Protection Agency. (n.d.). Ozone Trends. Retrieved May 1, 2024, from https://www.epa.gov/air-trends/ozone-trends

U.S. Environmental Protection Agency. (n.d.). Particulate Matter (PM2.5) Trends. Retrieved May 1, 2024, from https://www.epa.gov/air-trends/particulate-matter-pm25-trends