Data 102 Final Project

Fall 2022

John Bae, Sophia Chadsey, Ronit Gupta, Risheek Somu

Contents

1. Data Overview

Throughout the course of the project, our group utilized and created multiple merged datasets to answer our research questions. The U.S. Centers for Disease Control and Prevention (CDC) dataset on annual state-level U.S. chronic disease indicators was used to provide data on asthma cases. Furthermore, the U.S. CDC Daily Census-Tract datasets were used for particulate matter 2.5 (PM2.5) concentration and ozone concentration numbers. The Census and American Community Survey was also incorporated to include demographic data on minority status and income.

The asthma dataset contains a row for each new case of asthma recorded, with location and year information included. The PM2.5 and ozone concentration datasets contain a row for each day with the levels of each indicator recorded, respectively, along with location information. The demographic datasets contained yearly state data on factors including minority status, household income, and more.

The asthma dataset did not include county-level data on new cases, instead only providing at the state-level. This was inconsistent with what we had from the other PM2.5/ozone datasets and from the demographic datasets, and so our group pivoted to analyzing five states rather than conducting a county-level analysis of California. Furthermore, the PM2.5/ozone data had daily granularity, whereas the asthma and demographic datasets had yearly granularity, and so our group's investigation was by year. Census data is often reported to undercount minority populations, and so it is important to keep this in mind when drawing conclusions.

We chose to subset data from the timespan of 2010 to 2019, inclusive. This avoided any confounding effect of the COVID-19 pandemic on our data, research questions, and results. Our data consists of a sample as well as supplemental census data. The sample is anyone that the

CDC reported as developing asthma. This does not include cases that went unreported. This could ignore poorer patients that could not afford getting checked or locations where identification is not widely available. Patients are required under state and national law to have their cases reported, however reporting race and gender is voluntary. The entire data set of the population of chronic illness cases in the U.S. is subdivided into each individual patient that has the disease. Each row represents an individual patient with a chronic illness. Since every row corresponds to a specific person, we can interpret the results of our findings to what additional risk someone on average of chronic illness occurs due to the treatment effect.

There could be selection bias due to the need for the patient to have been diagnosed to be reported in the census. The bias occurs if there is a large amount of people who are undiagnosed for chronic illness. It would be helpful to know more specific demographic characteristics besides race such as age, notable previous health conditions, sex, and a more specific geographic location than US state. These features would help our model in accounting for the possible confounding factor that could contribute to the advent of chronic illness besides the particulate pollution we are measuring.

2. Research Questions

Causal Inference

Our first research question evaluated the causal relationship between PM2.5 concentration levels on the onset of asthma cases amongst five states in the U.S. – California, Texas, Pennsylvania, Ohio, and New York – incorporating race and income data from demographics. Causal inference was chosen because we are aiming to assess the degree of

impact of the treatment on the outcome. Answering this question has the potential to help progress analyses focused on understanding how PM2.5 concentrations are connected, or not connected, to asthma cases, and guide policy decisions as to how those in disadvantaged strata of income and non-white populations in these five states can best be supported to reduce their risk of asthma and boost their overall health.

The treatment is exposure to PM2.5 pollution and the outcome we chose to associate with the treatment was the onset of asthma. The units include PM2.5 pollution levels in parts per thousand particulates on average every year measured against the number of new asthma cases measured in every year. Possible confounders include access to healthcare, education level, environment, pre-existing conditions, nutrition, and age. The instrumental variables we used were median income and race. Since this is a natural experiment with many confounding variables we will be using matching techniques as well as identifying sources of randomization to get around the confounding variables.

Prediction with GLMs and Nonparametric Methods

For our second research question we wanted to predict the risk of onset of asthma from personal demographics and exposure to climate factors for people in our five states, using negative binomial regression compared to logistic regression. Answering this question can help guide policy decisions as to how the risk of asthma can be reduced through environmental programs, health initiatives, etc. Predicting with GLMs and nonparametric methods was a good fit because we were not positive that the data is necessarily linear but we wanted a linear result. Ultimately we wanted to compare our two models to see which had a more accurate prediction result.

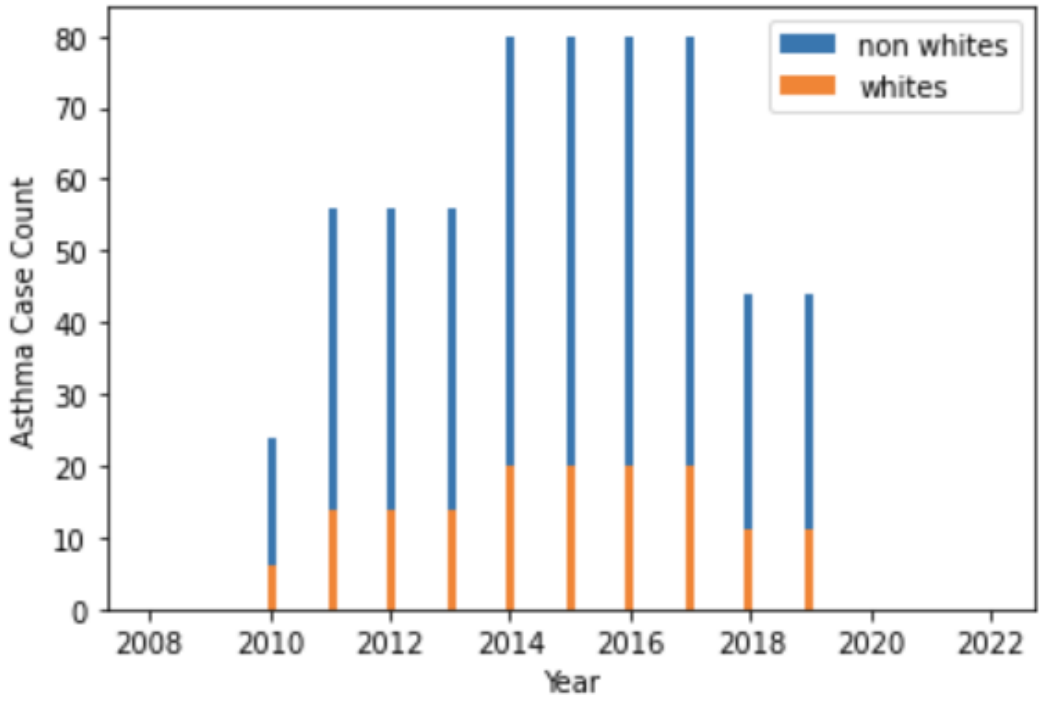
For our non-parametric method we choose logistic regression. Logistic regression was chosen since it is simple and we only have a few parameters. We are assuming linearity between the various variables we are measuring

For our GLM method we chose Negative Binomial regression because we were trying to measure the distribution of counts for the risk of asthma through using the log link function and the exponential inverse link function. The main goal was to determine the number of asthma cases based on personal demographics and exposure to climate factors. In this case choosing negative binomial regression was very useful since this regression technique can be used for overdispersed count data.

The demographic variables we are measuring are median earnings and non-white percentage in the given population. And the climate factors being PM2.5 concentrations and ozone concentrations. Data was collected per state level for California, New York, Pennsylvania, Ohio, and Texas based on yearly averages from 2010 to 2019.

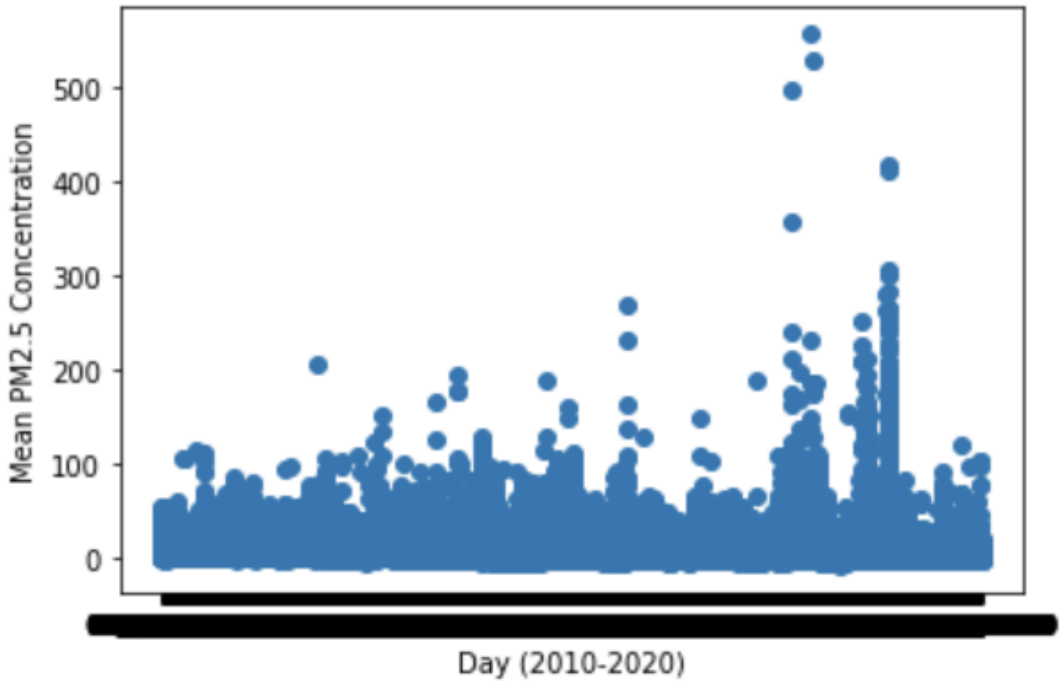
3. EDA

For research question 1, this figure compares asthma cases between White/non-White people from 2010-2020 in California.



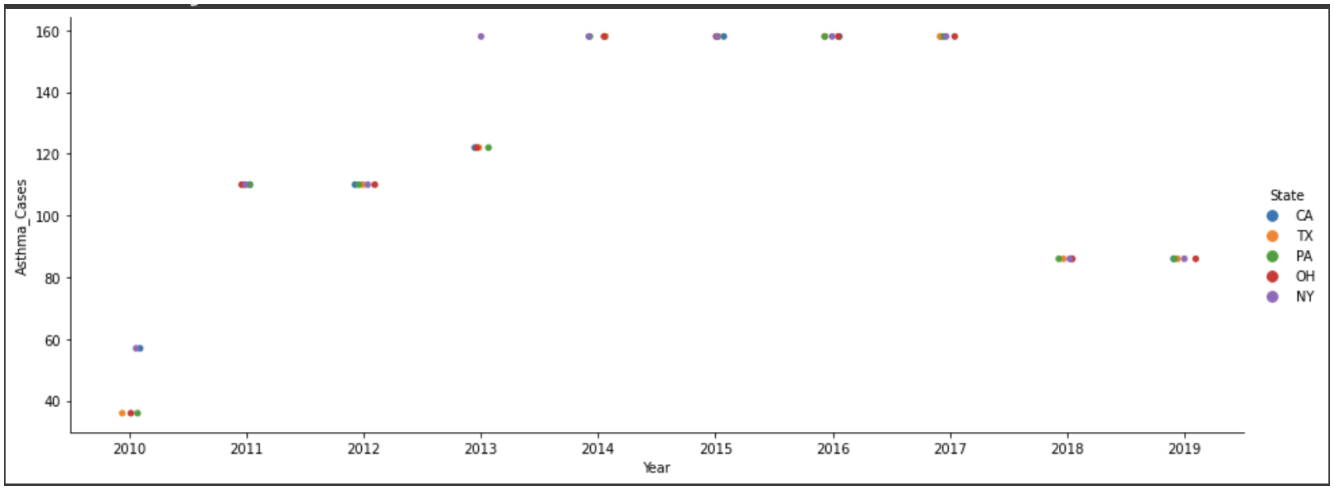
From this, we can observe the trend that over time, both White and non-White populations follow a similar trajectory of asthma cases year over year in California. Cases rose from 2010 to 2014, plateaued until 2018, and then decreased through 2020. However, there are also clearly more cases of asthma amongst the latter group, which we aim to investigate further in our research questions.

This figure showcases the mean daily PM2.5 concentration from 2010-2020 in California.



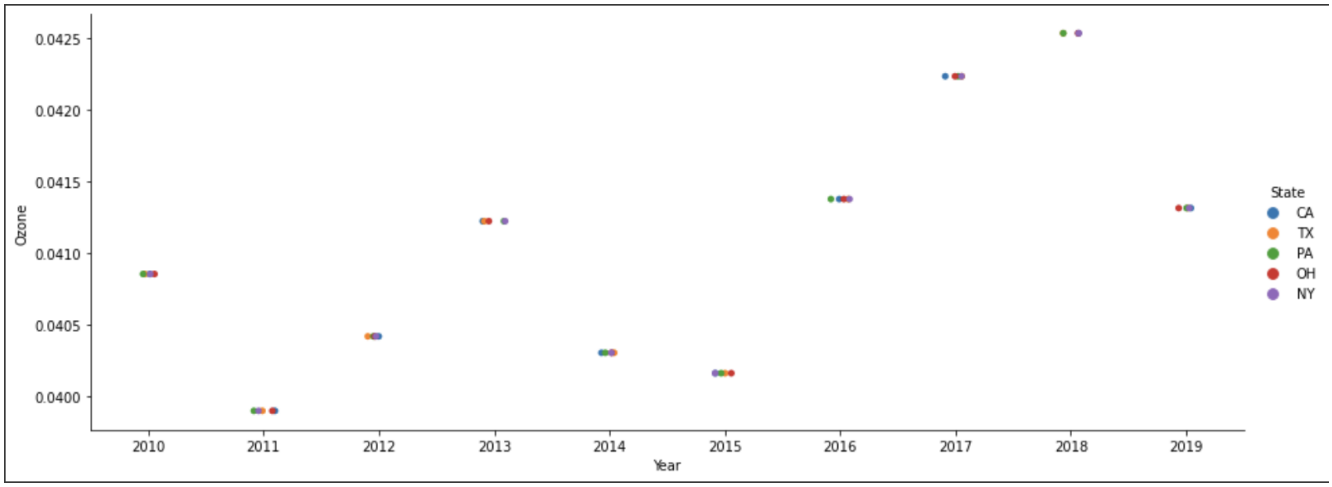
From this, we can observe the trend that while PM2.5 concentration levels generally maintained a set range of values from 2010 to 2020, they certainly increased towards the end of the time spectrum. It will be intriguing to scrutinize how this trend affects the onset of asthma cases.

The first figure was generated by stratifying the asthma dataset by race (White and non-White) and by location (California), and then creating a histogram by year. The latter figure was generated by combining the PM2.5 datasets for each year for California and then creating a scatterplot by year.



X-Axis = Date (by year from 2010 to 2019)

Y-Axis = Asthma_Counts recorded per year



X-Axis: (By Year From 2010 to 2019)

Y-Axis: Ozone Concentration (g/m^3)

The general trends of Asthma Counts for the top 5 most populated states follows a similar pattern over the course of the 10 years but there are some differences for specific states early on. For example, California and New York had higher counts of Asthma Cases during 2010 and California reached a peak of almost 160 cases in 2013 which was much higher than the rest of the states. There were two big spikes in the data showing drastic increases in asthma cases. One was from 2010 to 2011 where Asthma cases increased from around 35-55 cases to all 5 states

having 110 cases and another was from 2013 to 2014 where asthma cases for the 5 states increased from about 120 cases to 160 cases. These rises could have been due to spikes in global temperature spikes from those time periods resulting in dangerous climate factors that are related to increased chronic diseases.

In the data cleaning for the first question, the asthma cases from the CDC case report for the five states we selected for every year from 2010 to 2019 were summed and categorized for each respective state and year. The PM2.5 data was also categorized accordingly. For the confounding variables measured in the demographic data of median income and race, the variables were also summed by the five states and their respective year. The OLS model was able to use this data to calculate the linear relationship between any combination of variables.

For the second question, we made sure to join in Ozone concentration under the same time frame as asthma counts. We were able to find that there was Ozone data for each year we began to populate these ozone level values by a similar method to finding the pm 2.5 values. First we started by gathering data through EPA.gov and uploaded data from 2010 to 2019 into github in order to utilize it in our code. Since the data gathered was from daily counts of the ozone levels for each year we averaged out these levels to per year to fit our other data. After averaging out the ozone levels for each year (2010 to 2019) we then added this to our main dataframe to use in our future models and analysis.

Since this research question looks at how ozone level concentration and asthma counts are related and if one could predict the other, it was important to see how asthma counts changed over time. Additionally, we were able to see how the Ozone concentration levels changed over time. Ozone concentrations fluctuated over the course of the ten year time period for our data with the concentrations eventually reaching a steady increase from 2015 to 2018 which

corresponds to the peak asthma counts of 160 from the range of 2015 to 2017. The rise in ozone levels during that time likely created problems for people with chronic conditions such as asthma resulting in them needing to adjust during that three year time period.

4. Inference and Decisions

Causal Inference

Our research question aims to estimate the causal effect of PM2.5 on the development of asthma, utilizing race and income demographic data as supplements. The data does not represent a randomized trial on any person's development of asthma, so there may be confounding variables between the treatment and outcome.

METHODS

In constructing the causal inference question, the treatment group for our question is the reported average PM2.5 pollution levels for the entire year which was collected every day and averaged. This set includes daily averages from 2010 to 2019 across five states: California, Texas, Pennsylvania, Ohio, and New York. The corresponding outcome is the recorded new asthma cases for that year in that state reported to the CDC.

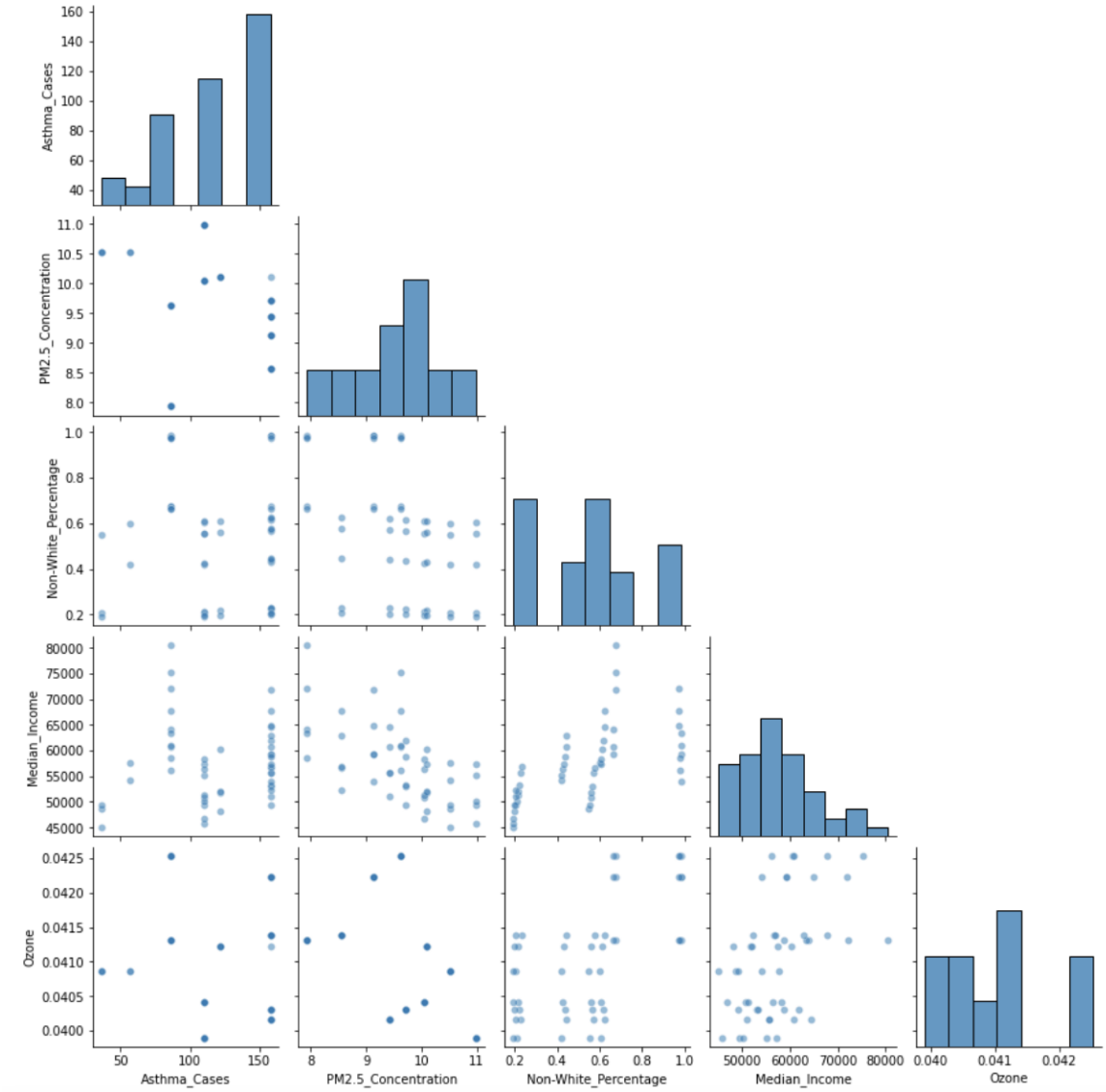
The variables our study identified and used as confounding variables were race and median income. These variables were measured in the demographic data and would help to factor out any increases or decreases in asthma cases that could just be accounted for by these variables instead of PM2.5 pollution level increases. We first used a naive OLS model to measure asthma cases against PM2.5 pollution then compared the results to a model using multiple combinations of individual state data, median income, and race to determine the effect

of confounding on the model. There are no colliders in this dataset we were aware of since this is an intersectional research question there is no clear way to tell.

RESULTS

All States

This pairplot provides pairwise correlations between our different variables of interest for all five states combined.



Here we have the regression results for comparing PM2.5 concentration to asthma cases for all five states.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.888
Model:	OLS	Adj. R-squared (uncentered):	0.886
Method:	Least Squares	F-statistic:	388.8
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	6.00e-25
Time:	05:33:04	Log-Likelihood:	-257.87
No. Observations:	50	AIC:	517.7
Df Residuals:	49	BIC:	519.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	12.2757	0.623	19.717	0.000	11.025	13.527

Omnibus:	3.079	Durbin-Watson:	0.971
Prob(Omnibus):	0.214	Jarque-Bera (JB):	2.922
Skew:	-0.539	Prob(JB):	0.232
Kurtosis:	2.511	Cond. No.	1.00

The coefficient tells us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~12 on the number of asthma cases when examined alone.

Here we have the regression results for comparing PM2.5 concentration, non-white percentage, and median income to asthma cases for all five states.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.908
Model:	OLS	Adj. R-squared (uncentered):	0.902
Method:	Least Squares	F-statistic:	154.2
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	2.50e-24
Time:	05:33:04	Log-Likelihood:	-253.02
No. Observations:	50	AIC:	512.0
Df Residuals:	47	BIC:	517.8
Df Model:	3		
Covariance Type:	nonrobust		

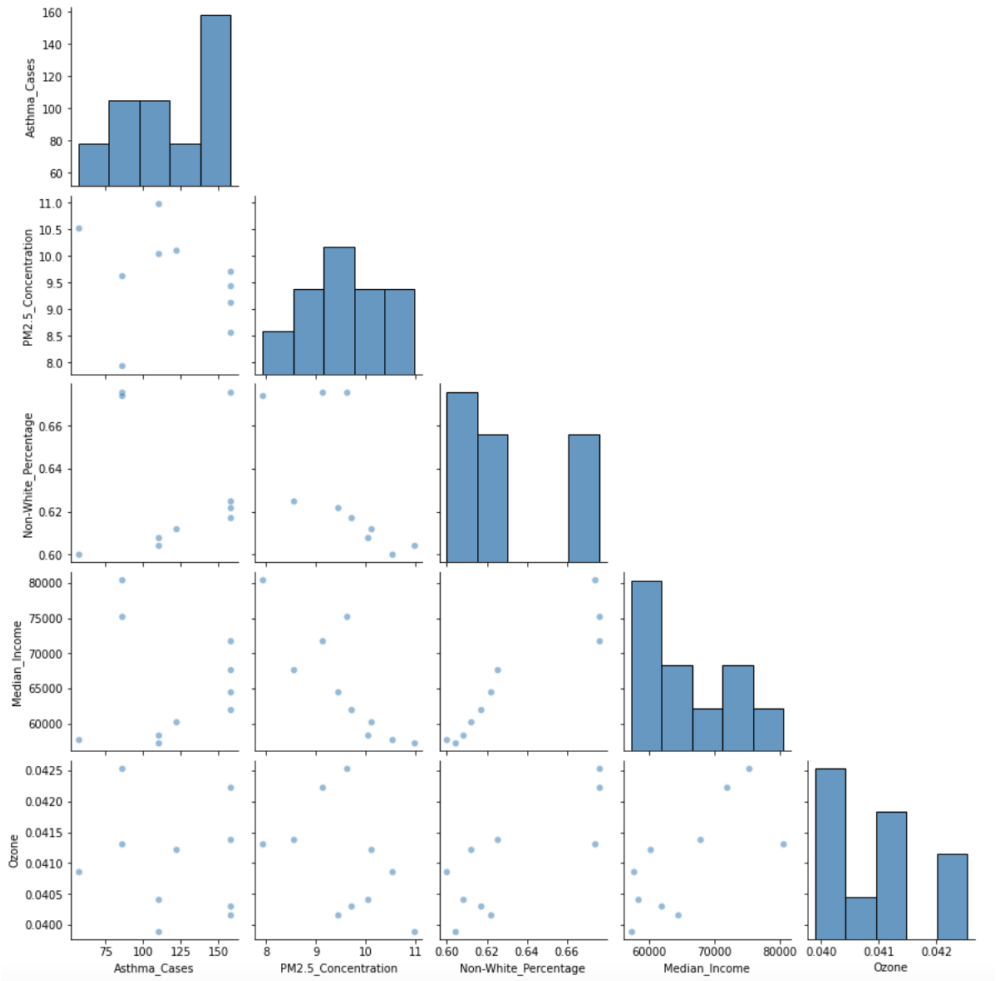
	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	2.3612	3.196	0.739	0.464	-4.069	8.791
Non-White_Percentage	-41.2801	26.387	-1.564	0.124	-94.364	11.804
Median_Income	0.0021	0.001	3.079	0.003	0.001	0.003

Omnibus:	3.761	Durbin-Watson:	0.922
Prob(Omnibus):	0.153	Jarque-Bera (JB):	3.129
Skew:	-0.503	Prob(JB):	0.209
Kurtosis:	2.300	Cond. No.	2.75e+05

The coefficients tell us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~2 on the number of asthma cases when examined in conjunction with the other variables, while the same increase in non-white percentage has an estimated effect of approximately -41 and median income has ~0.

California

This pairplot provides pairwise correlations between our different variables of interest for California.



Here we have the regression results for comparing PM2.5 concentration to asthma cases for California.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Asthma_Cases   R-squared (uncentered):                   0.902
Model:                            OLS   Adj. R-squared (uncentered):              0.891
Method:                 Least Squares   F-statistic:                              83.05
Date:                Tue, 13 Dec 2022   Prob (F-statistic):                    7.71e-06
Time:                        05:02:58   Log-Likelihood:                         -50.870
No. Observations:                  10   AIC:                                      103.7
Df Residuals:                       9   BIC:                                      104.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
PM2.5_Concentration    12.3400      1.354      9.113      0.000       9.277      15.403
==============================================================================
Omnibus:                        0.472   Durbin-Watson:                   0.707
Prob(Omnibus):                  0.790   Jarque-Bera (JB):                0.517
Skew:                          -0.267   Prob(JB):                        0.772
Kurtosis:                       2.022   Cond. No.                         1.00
==============================================================================
```

The coefficient tells us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~12 on the number of asthma cases when examined alone.

Here we have the regression results for comparing PM2.5 concentration, non-white percentage, and median income to asthma cases for California.

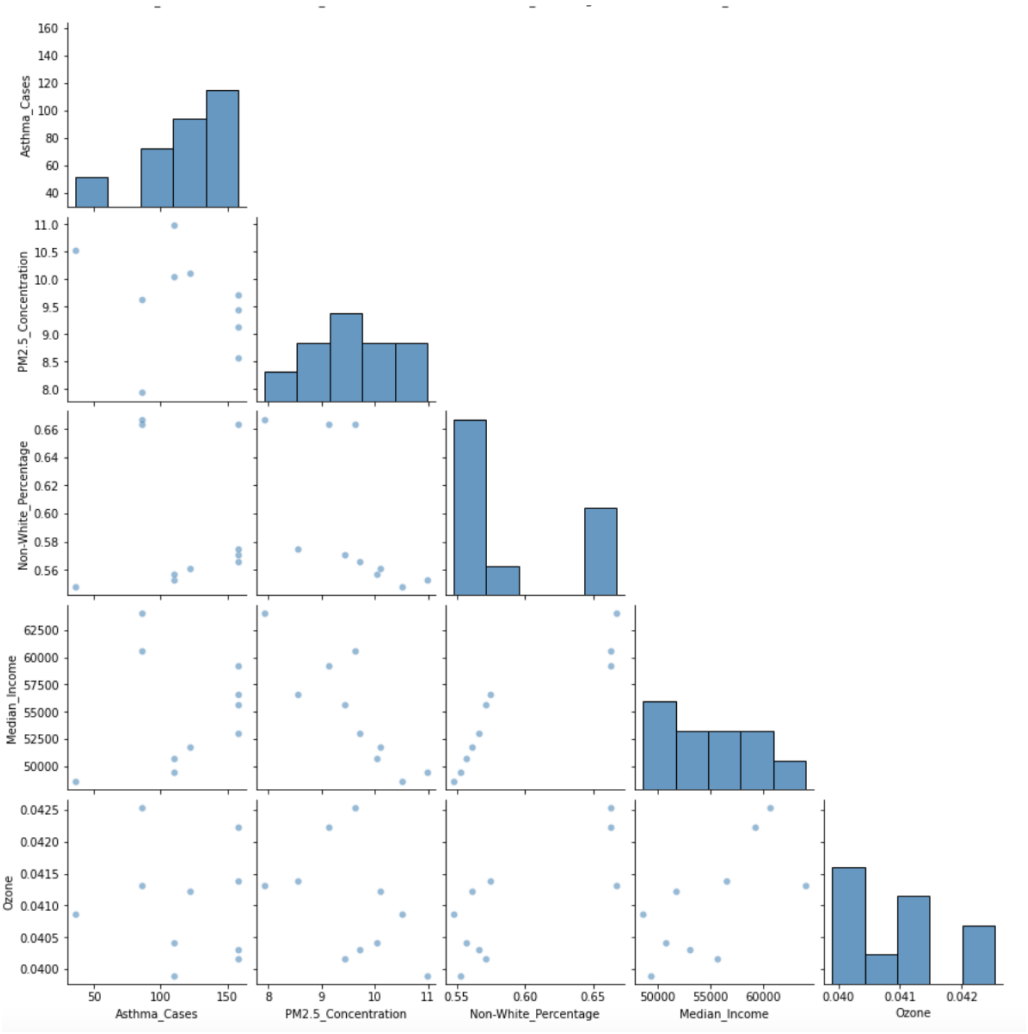
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Asthma_Cases   R-squared (uncentered):                   0.957
Model:                            OLS   Adj. R-squared (uncentered):              0.938
Method:                 Least Squares   F-statistic:                              51.82
Date:                Tue, 13 Dec 2022   Prob (F-statistic):                    3.80e-05
Time:                        05:02:58   Log-Likelihood:                         -46.773
No. Observations:                  10   AIC:                                      99.55
Df Residuals:                       7   BIC:                                      100.5
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
PM2.5_Concentration    -60.5602     25.521     -2.373      0.049    -120.908      -0.212
Non-White_Percentage  2723.6980   1035.098      2.631      0.034     276.081    5171.315
Median_Income           -0.0155      0.006     -2.444      0.044      -0.031      -0.001
==============================================================================
Omnibus:                        1.226   Durbin-Watson:                   1.596
Prob(Omnibus):                  0.542   Jarque-Bera (JB):                0.705
Skew:                          -0.160   Prob(JB):                        0.703
Kurtosis:                       1.739   Cond. No.                     6.95e+06
==============================================================================
```

The coefficients tell us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of approximately -60 on the number of asthma cases when examined in conjunction with

the other variables, while the same increase in non-white percentage has an estimated effect of ~2724 and median income has ~0.

Texas

This pairplot provides pairwise correlations between our different variables of interest for Texas.



Here we have the regression results for comparing PM2.5 concentration to asthma cases for Texas.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.879
Model:	OLS	Adj. R-squared (uncentered):	0.865
Method:	Least Squares	F-statistic:	65.23
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	2.05e-05
Time:	05:13:26	Log-Likelihood:	-51.883
No. Observations:	10	AIC:	105.8
Df Residuals:	9	BIC:	106.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	12.1024	1.498	8.077	0.000	8.713	15.492

Omnibus:	1.413	Durbin-Watson:	0.705
Prob(Omnibus):	0.493	Jarque-Bera (JB):	0.664
Skew:	-0.611	Prob(JB):	0.718
Kurtosis:	2.682	Cond. No.	1.00

The coefficient tells us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~12 on the number of asthma cases when examined alone.

Here we have the regression results for comparing PM2.5 concentration, non-white percentage, and median income to asthma cases for California.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.919
Model:	OLS	Adj. R-squared (uncentered):	0.884
Method:	Least Squares	F-statistic:	26.36
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	0.000345
Time:	05:13:26	Log-Likelihood:	-49.886
No. Observations:	10	AIC:	105.8
Df Residuals:	7	BIC:	106.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	7.4726	10.002	0.747	0.479	-16.177	31.123
Non-White_Percentage	-925.0455	782.296	-1.182	0.276	-2774.882	924.791
Median_Income	0.0108	0.008	1.431	0.196	-0.007	0.029

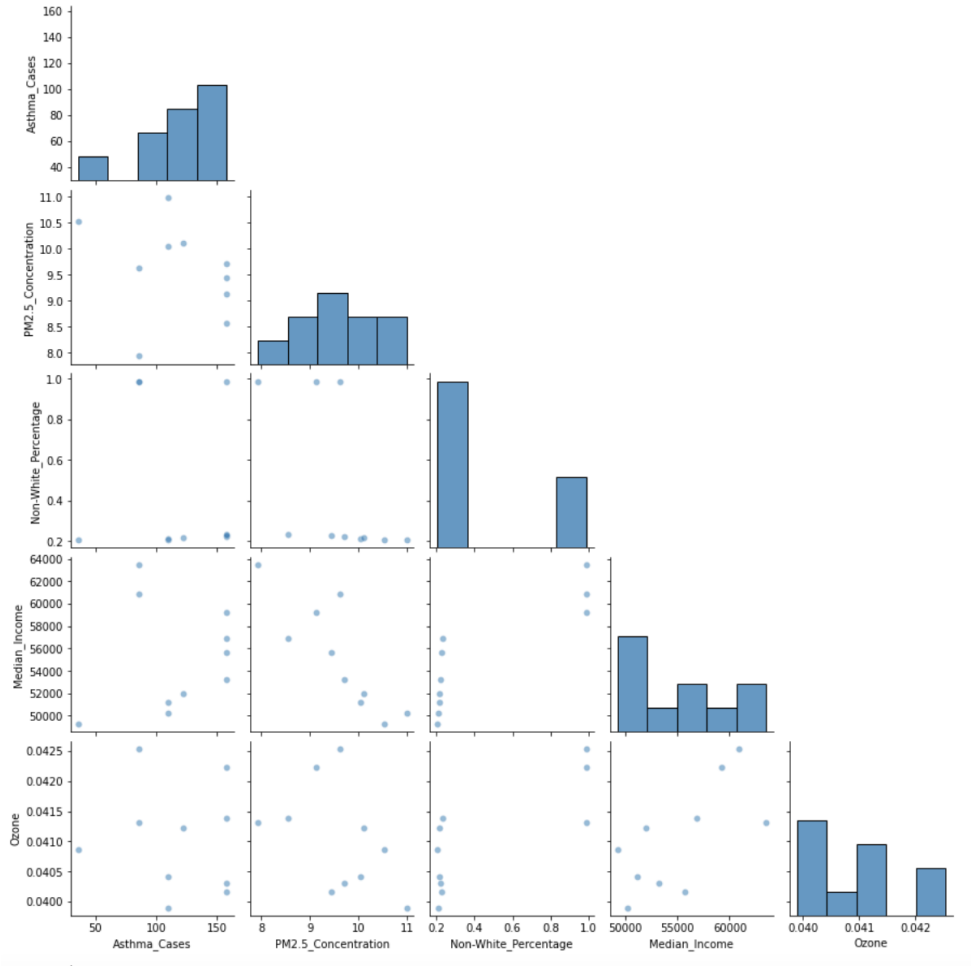
Omnibus:	0.099	Durbin-Watson:	1.338
Prob(Omnibus):	0.952	Jarque-Bera (JB):	0.212
Skew:	-0.170	Prob(JB):	0.899
Kurtosis:	2.373	Cond. No.	3.22e+06

The coefficients tell us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~7 on the number of asthma cases when examined in conjunction with the other

variables, while the same increase in non-white percentage has an estimated effect of -925 and median income has ~2.

Pennsylvania

This pairplot provides pairwise correlations between our different variables of interest for Pennsylvania.



Here we have the regression results for comparing PM2.5 concentration to asthma cases for Pennsylvania.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.879
Model:	OLS	Adj. R-squared (uncentered):	0.865
Method:	Least Squares	F-statistic:	65.23
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	2.05e-05
Time:	05:14:08	Log-Likelihood:	-51.883
No. Observations:	10	AIC:	105.8
Df Residuals:	9	BIC:	106.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	12.1024	1.498	8.077	0.000	8.713	15.492

Omnibus:	1.413	Durbin-Watson:	0.705
Prob(Omnibus):	0.493	Jarque-Bera (JB):	0.664
Skew:	-0.611	Prob(JB):	0.718
Kurtosis:	2.682	Cond. No.	1.00

The coefficient tells us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~12 on the number of asthma cases when examined alone.

Here we have the regression results for comparing PM2.5 concentration, non-white percentage, and median income to asthma cases for Pennsylvania.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.926
Model:	OLS	Adj. R-squared (uncentered):	0.894
Method:	Least Squares	F-statistic:	29.18
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	0.000250
Time:	05:14:08	Log-Likelihood:	-49.417
No. Observations:	10	AIC:	104.8
Df Residuals:	7	BIC:	105.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	-11.3052	11.160	-1.013	0.345	-37.695	15.084
Non-White_Percentage	-82.4938	55.237	-1.493	0.179	-213.110	48.122
Median_Income	0.0048	0.002	2.097	0.074	-0.001	0.010

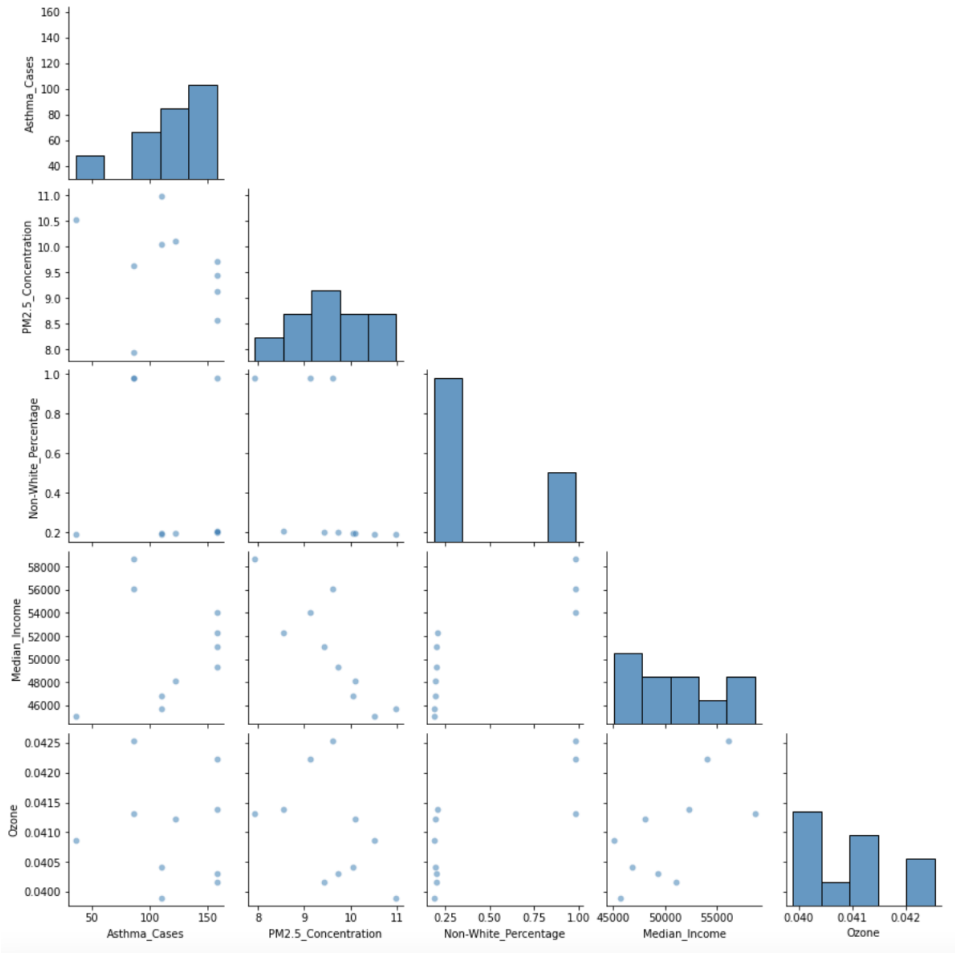
Omnibus:	0.448	Durbin-Watson:	1.482
Prob(Omnibus):	0.799	Jarque-Bera (JB):	0.169
Skew:	-0.268	Prob(JB):	0.919
Kurtosis:	2.656	Cond. No.	2.41e+05

The coefficients tell us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of -11 on the number of asthma cases when examined in conjunction with the other

variables, while the same increase in non-white percentage has an estimated effect of -82 and median income has ~0.

Ohio

This pairplot provides pairwise correlations between our different variables of interest for Ohio.



Here we have the regression results for comparing PM2.5 concentration to asthma cases for Ohio.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.879
Model:	OLS	Adj. R-squared (uncentered):	0.865
Method:	Least Squares	F-statistic:	65.23
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	2.05e-05
Time:	05:15:35	Log-Likelihood:	-51.883
No. Observations:	10	AIC:	105.8
Df Residuals:	9	BIC:	106.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	12.1024	1.498	8.077	0.000	8.713	15.492

Omnibus:	1.413	Durbin-Watson:	0.705
Prob(Omnibus):	0.493	Jarque-Bera (JB):	0.664
Skew:	-0.611	Prob(JB):	0.718
Kurtosis:	2.682	Cond. No.	1.00

The coefficient tells us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~12 on the number of asthma cases when examined alone.

Here we have the regression results for comparing PM2.5 concentration, non-white percentage, and median income to asthma cases for Ohio.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.925
Model:	OLS	Adj. R-squared (uncentered):	0.893
Method:	Least Squares	F-statistic:	28.82
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	0.000260
Time:	05:15:35	Log-Likelihood:	-49.475
No. Observations:	10	AIC:	105.0
Df Residuals:	7	BIC:	105.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	-10.4010	10.899	-0.954	0.372	-36.173	15.371
Non-White_Percentage	-77.9608	53.354	-1.461	0.187	-204.123	48.202
Median_Income	0.0050	0.002	2.069	0.077	-0.001	0.011

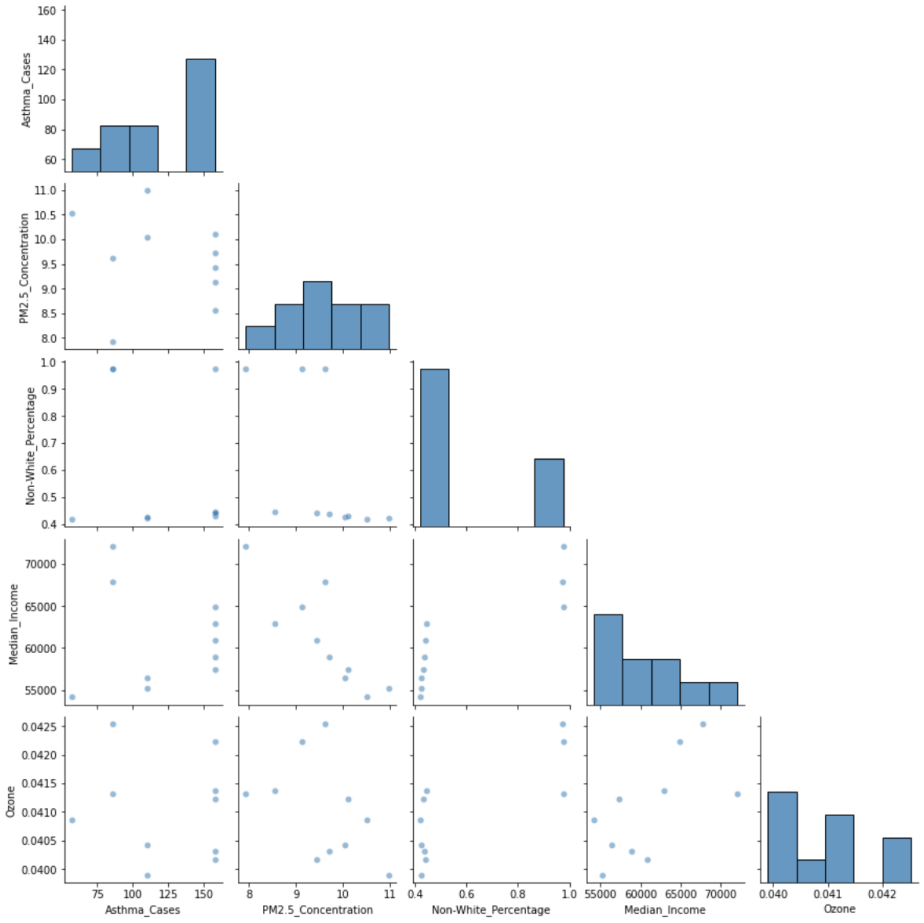
Omnibus:	0.430	Durbin-Watson:	1.492
Prob(Omnibus):	0.806	Jarque-Bera (JB):	0.135
Skew:	-0.236	Prob(JB):	0.935
Kurtosis:	2.683	Cond. No.	2.13e+05

The coefficients tell us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of -10 on the number of asthma cases when examined in conjunction with the other

variables, while the same increase in non-white percentage has an estimated effect of -77 and median income has ~0.

New York

This pairplot provides pairwise correlations between our different variables of interest for New York.



Here we have the regression results for comparing PM2.5 concentration to asthma cases for New York.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.902
Model:	OLS	Adj. R-squared (uncentered):	0.892
Method:	Least Squares	F-statistic:	83.19
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	7.65e-06
Time:	05:16:09	Log-Likelihood:	-51.174
No. Observations:	10	AIC:	104.3
Df Residuals:	9	BIC:	104.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	12.7310	1.396	9.121	0.000	9.573	15.889

Omnibus:	1.021	Durbin-Watson:	0.698
Prob(Omnibus):	0.600	Jarque-Bera (JB):	0.800
Skew:	-0.469	Prob(JB):	0.670
Kurtosis:	1.980	Cond. No.	1.00

The coefficient tells us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of ~13 on the number of asthma cases when examined alone.

Here we have the regression results for comparing PM2.5 concentration, non-white percentage, and median income to asthma cases for New York.

OLS Regression Results

Dep. Variable:	Asthma_Cases	R-squared (uncentered):	0.932
Model:	OLS	Adj. R-squared (uncentered):	0.902
Method:	Least Squares	F-statistic:	31.82
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	0.000189
Time:	05:16:09	Log-Likelihood:	-49.389
No. Observations:	10	AIC:	104.8
Df Residuals:	7	BIC:	105.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PM2.5_Concentration	-4.2778	10.114	-0.423	0.685	-28.194	19.639
Non-White_Percentage	-113.3265	82.220	-1.378	0.211	-307.745	81.092
Median_Income	0.0038	0.002	1.729	0.127	-0.001	0.009

Omnibus:	0.010	Durbin-Watson:	1.329
Prob(Omnibus):	0.995	Jarque-Bera (JB):	0.232
Skew:	-0.003	Prob(JB):	0.890
Kurtosis:	2.254	Cond. No.	3.96e+05

The coefficients tell us that a 1 unit increase in PM2.5 concentration has an estimated causal effect of -4 on the number of asthma cases when examined in conjunction with the other

variables, while the same increase in non-white percentage has an estimated effect of -113 and median income has ~0.

Taking all of our results into consideration, overall, there were not any clear signs of causality. Since this is an intersectional research question that relies on observational data we were unable to pinpoint if specific demographic variables such as income and minority status directly caused the onset of asthma. With so many confounding variables it is quite difficult to accurately provide evidence of causality. While we were overall able to see that a 1 unit increase in PM2.5 concentration had an estimated causal effect of ~12 on the number of asthma cases when investigated in isolation across all five states, the estimated effect dropped to ~2 when non-white percentage and median income were included. Additionally, there were many disparities present between states. These disparities are one of the reasons we were unable to prove causality.

This study would benefit from the isolation of instrumental variables that will facilitate the use of a two-stage least-squares regression model. This will be beneficial to more clearly assess the effect of PM2.5 on asthma cases with less influence by confounding variables.

DISCUSSION

Our model tried to compare PM2.5 and asthma cases; however, our results reject the causal relationship between the two. Despite these results, this does not necessarily mean that PM2.5 pollution has no adverse effects on people. Our data was limited by various factors mainly in the reliance on observational data to find connections to the treatment rather than administration of a treatment to find a connection. Another limitation is the rarity of new asthma cases. In addition, it is also questionable to determine whether new asthma cases are a result of pollution or due to another unseen factor, or normal discovery from unrelated natural causes.

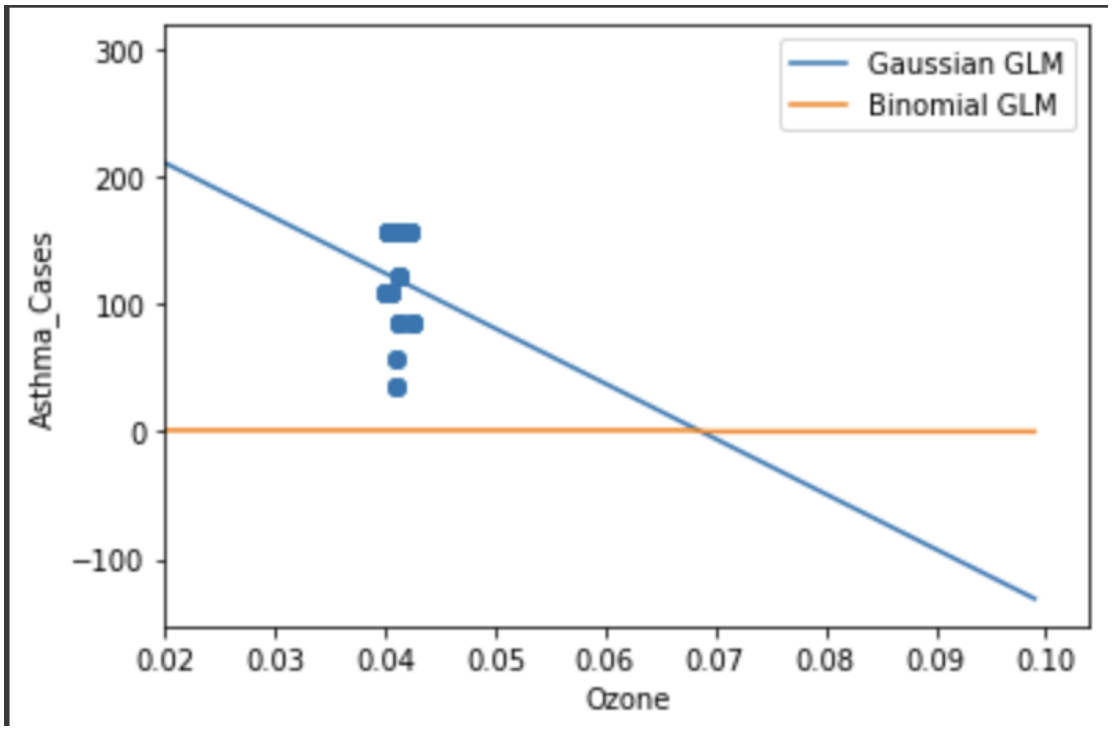
This stems from the reliance on observational data. Another consideration is that PM2.5 may have other adverse effects other than asthma cases, but we did not have the data to measure these as they are not individually reported and publicly available. The length of the study is also a consideration where development and identification of asthma from PM2.5 exposure could take many decades.

Many confounding factors are not taken into account and are most likely unknown. We only took median income and race into consideration and found that this accounted for most of the variation in asthma cases. Ideally, we could control for these factors to find the true effect of PM2.5, but that is impossible with only observational data. In addition, the exact amount of PM2.5 exposure of any patient case is unknown, and the only data we can use is geographic and yearly data to try to predict how much exposure patients probably had in that given time and location. Developing asthma is based on a complex combination of factors such as genetics, subject health conditions, etc. other than pollutants, so knowing how much these other factors contribute combined with PM2.5 concentration is most likely skewed to random variations in these other effects. Since this is an observational study, we cannot control for these factors but simply try to factor them out by estimating the confounding effect with the data we have. Furthermore, the standard errors of our causal inference models remain something to improve upon in future model iterations.

Prediction with GLMs and Nonparametric Methods

METHODS/RESULTS

We are trying to predict the change in Asthma counts and we are using Ozone concentration as the predictive feature.



We decided to use a Gaussian model because Poisson, gamma, and inverse Gaussian did not work. If we look into the scatter plot, the linear line for Gaussian seems to be a better fit for the data than the Binomial GLM.

Therefore, we decided to use GLM as a model. When using Gaussian and the identity link function, we are essentially reducing our research question into a regular linear regression problem. We are trying to predict “Asthma_Counts” by using the explanatory variable of “Ozone”.

For our nonparametric method we choose to use logistic regression. Logistic regression was used since it is a simple form of regression and we only have a few parameters.

One assumption we made was that there would be linearity between the 4 variables we are measuring and the onset of asthma. Another assumption we had to make was that the onset of asthma was independent of every state and each asthma metric did not affect each other. While

there may be some dependence due to geography reasons of air quality by picking only the top five most populous states to take data from which were also geographically distinct. And the last major assumption we made was that there were not any extreme outliers in our data, this was a fairly easy assumption to make since during our exploratory data analysis process we were able to make visualizations to verify the lack of extreme outliers.

For logistic regression using various libraries, especially *sklearn*, we were able to look at the confusion matrix, training matrix, and lastly an accuracy score. Starting off the confusion matrix this matrix was used to evaluate the accuracy of a classification. Since the confusion matrix is used to see how well the prediction of the model is, it allows for a more visual method of seeing accuracy. The training matrix allowed us to do similar evaluations as to the confusion matrix but used scaled data. And lastly the most important evaluation was an accuracy score for each combination of features in logistic regression.

For logistic regression there were three types of results from the model. First the results of the confusion matrix for each combination of variables as follows. These confusion matrices show that none of the features were that accurate of a predictor of asthma.

```
Confusion Matrix for ozone and pm 2.5:
 [[0 0 1 0 0]
 [0 1 0 0 2]
 [0 0 0 0 2]
 [0 0 0 0 1]
 [0 2 0 0 4]]
```

```
Confusion Matrix for all variables:
 [[0 0 1 0 0]
 [0 2 0 0 1]
 [0 0 1 0 1]
 [0 0 0 0 1]
 [0 0 0 0 6]]
```

```
Confusion Matrix for only ozone:
 [[0 0 0 0 1]
 [0 1 0 0 2]
 [0 0 0 0 2]
 [0 0 0 0 1]
 [0 2 2 0 2]]
```

```
Confusion Matrix for median earnings and non-white percentage:
 [[0 0 1 0 0]
 [0 2 0 0 1]
 [0 0 2 0 0]
 [0 0 0 0 1]
 [0 1 0 0 5]]
```

The next set of matrices are similar in that they show the scaled prediction. There was also low performance shown in these matrices.

```
Standard Scalar Matric for all variables:
[[-0.08103518  0.55459526  2.19730045  1.82015462 -1.29062235  1.29419243]
 [ 0.47299548  0.30661913  0.29232022  0.26301032 -1.29062235 -0.44077568]
 [-1.31168143  0.35698928  1.2455048   0.44577066 -1.29062235  0.60020518]
 [ 1.49199415 -0.42956    -0.33194055 -1.30859207 -0.58837195 -1.13476292]
 [-1.31168143 -0.34044358  0.63563839  0.44577066 -0.58837195  0.60020518]
 [-0.64217525  1.71698337  0.88627707  1.46325915 -0.58837195  0.9471988 ]
 [-0.30183769  0.14775942 -0.28055015 -0.99687901  1.51837924  0.25321156]
 [-2.03412815  0.54684601  2.84921355  0.3712039  -1.29062235  1.64118605]
 [-1.31168143  0.16325793 -0.16539525  0.44577066  1.51837924  0.60020518]
 [ 0.02824631 -1.20061078 -0.58598842 -0.82755524  0.81612884 -0.09378206]]

Standard Scalar Matric for ozone:
[[ 1.82015462]
 [ 0.26301032]
 [ 0.44577066]
 [-1.30859207]
 [ 0.44577066]
 [ 1.46325915]
 [-0.99687901]
 [ 0.3712039 ]
 [ 0.44577066]
 [-0.82755524]]

Standard Scalar Matric for environmental factors:
[[ 1.82015462 -0.08103518]
 [ 0.26301032  0.47299548]
 [ 0.44577066 -1.31168143]
 [-1.30859207  1.49199415]
 [ 0.44577066 -1.31168143]
 [ 1.46325915 -0.64217525]
 [-0.99687901 -0.30183769]
 [ 0.3712039  -2.03412815]
 [ 0.44577066 -1.31168143]
 [-0.82755524  0.02824631]]

Standard Scalar Matric for social factors:
[[ 0.55459526  2.19730045]
 [ 0.30661913  0.29232022]
 [ 0.35698928  1.2455048 ]
 [-0.42956    -0.33194055]
 [-0.34044358  0.63563839]
 [ 1.71698337  0.88627707]
 [ 0.14775942 -0.28055015]
 [ 0.54684601  2.84921355]
 [ 0.16325793 -0.16539525]
 [-1.20061078 -0.58598842]]
```

And for the last result of logistic regression we were able to predict accuracy.

```
Accuracy for all variables: 0.6923076923076923
Accuracy for only ozone: 0.23076923076923078
Accuracy for ozone and pm 2.5: 0.38461538461538464
Accuracy for median earnings and non-white percentage: 0.6923076923076923
```

```
Accuracy for only ozone: 0.23076923076923078
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:           Asthma_Cases   No. Observations:                   50
Model:                            GLM   Df Residuals:                       48
Model Family:                Gaussian   Df Model:                            1
Link Function:               identity   Scale:                          1486.6
Method:                          IRLS   Log-Likelihood:                -252.53
Date:                Tue, 13 Dec 2022   Deviance:                       71355.
Time:                        05:02:19   Pearson chi2:                 7.14e+04
No. Iterations:                     3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        297.3986    269.153      1.105      0.269    -230.131     824.928
Ozone      -4329.2686   6558.228     -0.660      0.509   -1.72e+04    8524.623
==============================================================================
```

From our GLM model, we can see that the constant base is 297.40 and the coefficient for Ozone is -4329.27. This indicates that for every 1 unit increase in Ozone, the dependent variable decreases by -4329.27. Our 95% confidence bound is very large and is between -1.72e+04 and between 8524.93. This indicates that the model is not really sure whether or not the Ozone

concentration has a negative or positive effect on the dependent variable. This really shows the limitations of our model since we are not really sure whether the Ozone concentration has a true negative or positive effect. In the future, it would be smart to introduce more covariates so that the relationship between the two variables can be examined better. This analysis can also apply to the constant value as well. Overall, the model could be flawed as a whole or it might need some more complexity in order to strengthen its importance.

DISCUSSION

From our results, for GLM, it seems that our Gaussian GLM frequentist model fits the data the best because the Guassian distribution itself is linear regression. Therefore, it is better than our Bayesian perspective. The proper distributions or families were not looked at from the Bayesian's perspective, therefore, it is not as conclusive.

The logistic regression fit the data reasonably but did not produce any high accuracy predictions. Due to a lack of data since we were unable to find asthma data per county and had to switch to state level, there were not enough data points to have the model be able to fit the data. Especially with logistic regression, it is easier when there are many data points in order to reasonably fit the data.

Generalized Linear Model Regression Results

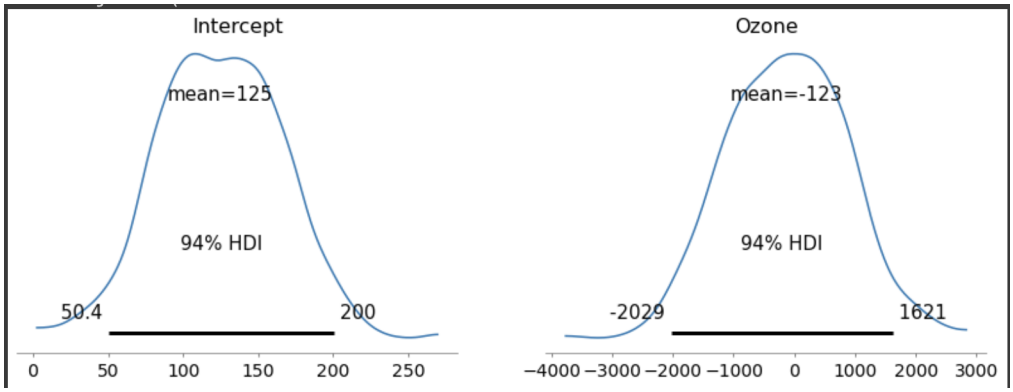
Dep. Variable:	Asthma_Cases	No. Observations:	50
Model:	GLM	Df Residuals:	48
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	1486.6
Method:	IRLS	Log-Likelihood:	-252.53
Date:	Tue, 13 Dec 2022	Deviance:	71355.
Time:	05:02:19	Pearson chi2:	7.14e+04
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	297.3986	269.153	1.105	0.269	-230.131	824.928
Ozone	-4329.2686	6558.228	-0.660	0.509	-1.72e+04	8524.623

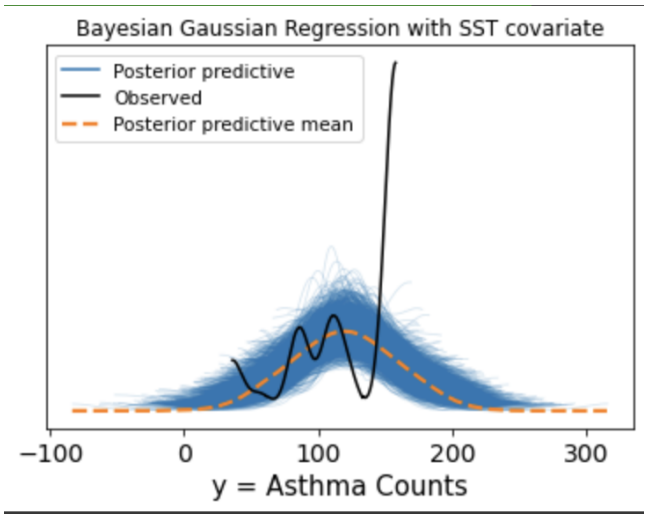
When comparing the two GLM methods from the frequentist perspective, we realized that both models did not fit the data too well. The frequentist Binomial Model does not fit the data very well as the chi squared is a very large number and the p-value is greater than the alpha value that is set at 0.05, showing that the findings are not statistically significant.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:           Asthma_Cases   No. Observations:                   50
Model:                            GLM   Df Residuals:                       48
Model Family:                Binomial   Df Model:                            1
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                   -inf
Date:                Tue, 13 Dec 2022   Deviance:                   4.8604e+05
Time:                        05:02:19   Pearson chi2:                 3.50e+21
No. Iterations:                     2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.003e+18   4.68e+08   4.28e+09      0.000       2e+18       2e+18
Ozone      -2.925e+19   1.14e+10  -2.56e+09      0.000   -2.92e+19   -2.92e+19
==============================================================================
```

Additionally, the frequentist Binomial model does not fit the data well either and is significantly worse than the Gaussian one because the chi-squared value is much greater and the p-value is zero which means that we fail to reject the null and the results are statistically significant. Limitations of this model include the fact that it may not have been accurately run because the log-likelihood shows up as negative infinity which essentially means that our likelihood is 0 and the coefficients are very small values. All these characteristics gathered from this model point to the fact that there may have been an error in setting up the model or that it is simply not a good fit.



From these credible interval plots gathered through Bayesian Regression, we can see the posterior distribution values for each set of x values for our intercept and our predictor which was the Ozone concentration. When comparing the estimates of Frequentist Regression and the Bayesian Regression, we can see that they are not that similar and the ranges for the 95% confidence interval are not very close to each other.



From this Bayesian Gaussian Regression with covariates, we can see that the Bayesian Gaussian model does not fit the model too well as the observed data does not align with the posterior predictive values. A better covariate might result in a better finding that allows Bayesian Gaussian Regression to better fit the data.

Interpreting the results from the logistic regression did not produce any significant results. The highest prediction was from using all variables with a 69.23% accuracy rate. Although this is quite high it is more than likely high due to over-fitting of the data and therefore was discarded. The more accurate prediction accuracy result was from using the environmental factors of PM2.5 and ozone. These produced an accuracy of 38.46% and when using just ozone alone a 23.08% accuracy. Overall the results did not provide any clarity to our research question.

There are many limitations for logistic regression. As was discussed earlier logistic regression is the best model on many points of data; this was a limitation since we did not have an exorbitant amount of data. Without a clear linear relationship between our features and the onset of asthma logistic regression does not provide accurate predictions and was likely overfitted since the relationship between the variables was not exactly linear.

For logistic regression either finding data per asthma or including more years to have more data points would improve the model since not having enough data points was a failing in the logistic regression model. Overall, however, adding different features would not necessarily improve our model.

5. Conclusion

Our first research question investigated the degree of the causal relationship between PM2.5 concentration and number of asthma cases by state, taking into account minority status and median income. Our second research question aimed to predict, using negative binomial regression and logistic regression, how personal demographics and exposure to climate factors affect the risk of onset of asthma. Our results did not display a causal relationship between the treatment and the outcome for any of these models. These findings are likely quite narrow

because they include multiple confounding variables, look at a 10-year range (pre-COVID-19), include only 5 out of 50 U.S. states, and should be refined to account for other potential factors like education, family history, medical history, and more.

There was not a clear link shown through both our causal inference and prediction models. With further research and more domain expertise there could be more evidence to support a call to action to be taken in the real-world. Currently though our call to action is more research to be able to statistically demonstrate a link between both social-demographic factors (median earnings and non-white percentage) and environmental factors (PM2.5 and ozone concentrations). As well as statistical findings we urge more sociological and qualitative research to be conducted to be able to uncover any findings statistical analysis might be unable to look at.

For our research question on causal inference, the demographic data was used against the PM2.5 concentration data and the CDC asthma cases data. The benefit of using multiple data sources was not relying on a single source that could have biases in its collection, representation or measurement. Using multiple data sources was used to check against confounding and help answer the causal inference question. The main consequence of using multiple data sets was lack of synchronization where most of the association had to be either assumed or generalized. Any deeper connection could not be used as any variable to be measured had to be present in every dataset to have some point of reference against the others.

Based on our findings in the second question, we were not able to find a clear relationship between Asthma Cases and Ozone concentration levels. This is surprising considering Ozone is one of the leading causes in asthma development and long-term exposure to Ozone can lead to aggravation of Asthma.

With most real-world data there were quite a few inconsistencies with our data that limited our analysis. For example instead of choosing the top 5 most populous states to gather data from we instead had to gather data from the top 4 most populous states and Ohio due to missing asthma data in the top 5th most populous state. Missing data values like these from asthma as well as in other data sources could not be accounted for. Accounting for missing values in our analysis without thorough domain expertise as well as more knowledge on how data was collected could be researched more in the future.

There are a myriad of directions that future studies could take to build upon our work. They can incorporate other confounding variables such as education status, medical history, and expand upon our study by finding and utilizing county-level data in order to more narrowly examine the connections between environmental factors, asthma, and demographic factors. Furthermore, adding data from post-2019 that includes the time during the COVID-19 pandemic may provide intriguing new avenues to explore. These possibilities have potentially large policy-related implications because they can affect how government funding, outside resources, public health institutions and hospitals, and more focus on supporting the populations most affected by asthma and other diseases.

6. References

Kjellstrom, T., Butler, A.J., Lucas, R.M. *et al.* Public health impact of global heating due to climate change: potential effects on chronic non-communicable diseases. *Int J Public Health* 55, 97–103 (2010). https://doi.org/10.1007/s00038-009-0090-2.