Assessing the effects of particulate matter (PM2.5) and ozone concentrations on asthma prevalence: Data science evidence from the United States

Assessing the effects of particulate matter (PM2.5) and ozone concentrations on asthma prevalence: Data science evidence from the United States..... 1
1 Data Overview..... 4
1.1 Data Sourcing..... 4
1.2 Representation of Population..... 4
1.3 Data Collection & Privacy..... 4
1.4 Understanding the Data..... 5
Granularity..... 5
Limitations & Solutions..... 5
2 Research Questions..... 7
2.1 Question 1: Does exposure to air pollution, particularly fine particulate matter (PM2.5) and Ozone, increase the risk of asthma in the US population?..... 7
Implications of this question..... 7
Method..... 7
Limitations..... 7
2.2 Question 2: What is the causal effect of PM2.5 concentrations on the prevalence of asthma in the United States?..... 7
Implications of this question..... 7
Method..... 8
Limitations..... 8
3 Exploratory Data Analysis (EDA)..... 9
3.1 Question 1 EDA..... 9
Visualizations..... 9
4 Results..... 13
4.1 Question 1 Results..... 13
Our Question..... 13
Methods..... 13
Calculating P- Values..... 13
Results..... 15
P-Value Results..... 15
Bonferroni Correction Results..... 16
Benjamini-Hochberg Results..... 17
Discussion..... 17
4.2 Question 2 Results..... 18
Methods..... 18
Results..... 19
Discussion..... 19
4 Conclusion..... 21

1 Data Overview

1.1 Data Sourcing

For our study we used the following datasets:

- Daily Census Tract-Level PM2.5 Concentrations 2011-2014
- Daily Census Tract-Level Ozone Concentrations 2011-2014
- US Chronic Disease Indicators: Asthma

All of our dataset's are census datasets provided by the Centers for Disease Control and Prevention (CDC). Census datasets were available for this study. We are grateful for this because a census data set is typically better than a sample dataset because it provides a complete representation of the target population, eliminates sampling error, reduces the risk of bias, offers greater statistical power, and better handles rare events and subgroups.

1.2 Representation of Population

The Centers for Disease Control and Prevention (CDC) strives to create representative census datasets to ensure accurate and reliable information about the population. However, despite their best efforts, they may not always achieve a perfect representation due to factors such as non-response, undercounting, or data quality issues.

In our datasets specifically we ran into issues where several American territories and non-contiguous states were not represented in the individual datasets or were missing from others. To ensure the accuracy of our study we decided to drop all states and territories that were not representative in all of the datasets.

1.3 Data Collection & Privacy

All of the individual chronic disease indicator (CDI) datasets within the CDI dataset are a derived from the following data sources: the Behavioral Risk Factor Surveillance System (BRFSS), state cancer registries, the American Community Survey (ACS), birth and death certificates data in the National Vital Statistics System (NVSS), the State Tobacco Activities Tracking and Evaluation System, the United States Renal Data System, and the Youth Risk Behavior Surveillance System, Pregnancy Risk Assessment Monitoring System, the Alcohol Epidemiologic Data System, the Alcohol Policy Information System, alcohol policy legal research, the National Survey of Children's Health, State Emergency Department Databases, State Inpatient Databases, the Centers for Medicare and Medicaid Services Chronic Condition Warehouse and the Medicare Current Beneficiary Survey, the U.S. Department of Agriculture, the CDC School Health Profiles, Achieving a State of Healthy Weight, Maternal Practices in Infant Nutrition and Care, the Breastfeeding Report Card, the Health Resources and Services Administration Uniform Data System, the National Immunization Survey, and the Water Fluoridation Reporting System.

According to the CDC, “if the data providers suppress data for quality or confidentiality reasons, CDC does not report those particular data elements on the CDI website. CDC follows all data use policies of the data providers”.

Additionally the datasets that comprise much of CDI are required to follow the model state vital statistics act and regulations. These regulations have been revised numerous times to address privacy concerns, confidentiality, and fraudulent use of vital records, and strengthened penalty provisions of the Model Act as a deterrent to illegal use of vital records.

1.4 Understanding the Data

Granularity

Both the PM2.5 and ozone concentration datasets are rather granular. The files contain estimates of the mean prediction and associated standard error for each of the 2010 U.S. Census Tracts within the contiguous U.S. for each day of the modeling year from January 1, 2001 to December 31, 2014.

The asthma disease indicator dataset is less granualr that the PM2.5 and ozone concentration datasets with regard to their locationary data and frequency of observation. This dataset has numerous questions that separate the disease prevalence into numerous different age and gender demographics. Provided with each question are several different answer rate types of the prevalence of the disease that take into account different types of rates.

Limitations & Solutions

PM2.5 and ozone concentrations offer a daily average estimate while the asthma dataset offers a yearly estimate. Additionally the PM2.5 and ozone concentrations datasets are very large (~106 M data points) so we decided to preprocess these datasets by aggregating the estimates into monthly estimate averages for each of the 2010 U.S. Census Tracts within the contiguous U.S. with a simple python script reducing the the dataset to 2352 data points. Now with the dataset in a more usable format we were then able to reduce the PM2.5 and ozone concentrations estimates into yearly averages that would allow for their values to agree with the asthma dataset values.

There were also numerous inconsistencies within each of the datasets of the areas that they are representative of. To solve this problem we ensured that for each year there was a consistent data point for each location (i.e. fips code) for every datapoint that was present in all datasets.

The asthma disease dataset is stratified between three different categories: gender, race, and overall. In order to generalize our model we decided to go with the overall category.

Additionally the answers to the prevalence questions vary into several different groups. Since the questions themselves already stratify the data into different demographic categories, we decided to go with crude prevalence in order to generalize our model as much as possible.

2 Research Questions

2.1 Question 1: Does exposure to air pollution, particularly fine particulate matter (PM2.5) and Ozone, increase the risk of asthma in the US population?

Implications of this question

Answering the research question about the relationship between air pollution and asthma risk could have significant real-world implications. Key decisions that could be made based on the findings include strengthening air quality regulations, informing urban planning and zoning policies, developing targeted public health interventions, and promoting environmental justice initiatives. The results could also inspire public awareness campaigns, guide research funding allocation, and encourage personal lifestyle changes to reduce exposure to air pollution.

Method

We decided to use multiple hypothesis testing. We found that multiple hypothesis would be a good choice for this kind of study because it allows us to simultaneously investigate the effects of different air pollutants on asthma prevalence within the U.S. population. This approach helps control the overall Type I error rate (false positive findings) while exploring multiple relationships. This increases the study's rigor but also provides more comprehensive insights into the complex interplay between air pollution and respiratory health.

Limitations

With multiple hypothesis testing comes its limitations. Some of the limitations include increased type II errors (false negative findings) due to stringent adjustments for multiple comparisons which can result in potentially significant associations being overlooked. Additionally, there are many complex interactions that may not be covered or considered in the scope of this type of analysis such as the cumulative impact of multiple airborne pollutants and respiratory health. Ultimately it requires careful and well understood interpretation of results and analytical methods to understand all of the relationships at play in such an analysis.

2.2 Question 2: What is the causal effect of PM2.5 concentrations on the prevalence of asthma in the United States?

Implications of this question

By answering the research question regarding the causal relationship between PM2.5 concentrations and asthma prevalence, important real-world decisions can be made in public health policies, environmental protection measures, and resource allocation. This knowledge can guide policymakers and healthcare professionals in developing targeted strategies to mitigate the impact of air pollution on public health.

Method

The method we chose was causal inference, and specifically propensity score matching. We decided this method was best suited for addressing our research question because propensity score matching allows us to eliminate confounding variables through matching those who had similar scores of PM2.5 concentrations and creating groupings for the different treatment assignments. With this method, we can reduce the impact of confounding variables and concentrate on the causal effect of PM2.5 concentrations on the prevalence of asthma.

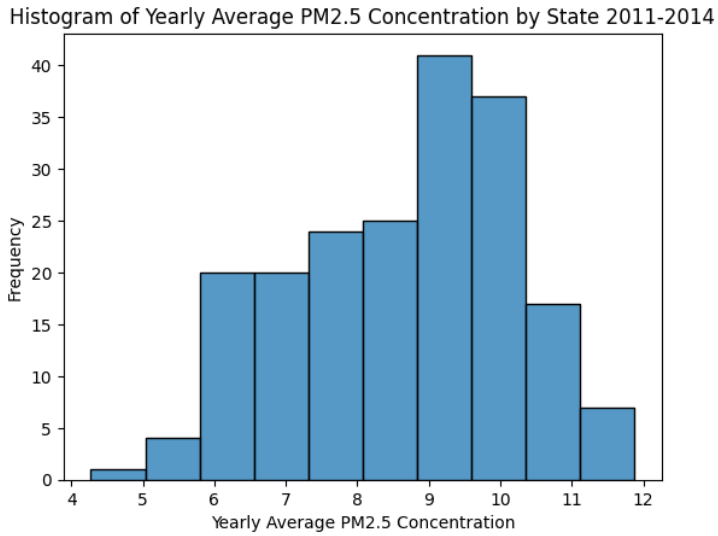
Limitations

One of the limitations we found in our causal inference model was that despite the fact that propensity score matching accounted for the confounding variables that we found, there may have been confounding variables that we did not account for, or that were not observed, that could create bias in the estimation of our ATE and thus our causal effect may have also been biased. We also found that since propensity score matching relied heavily on the unconfoundedness assumption, that there could have also been some violations of this assumption that also introduced bias within our project.

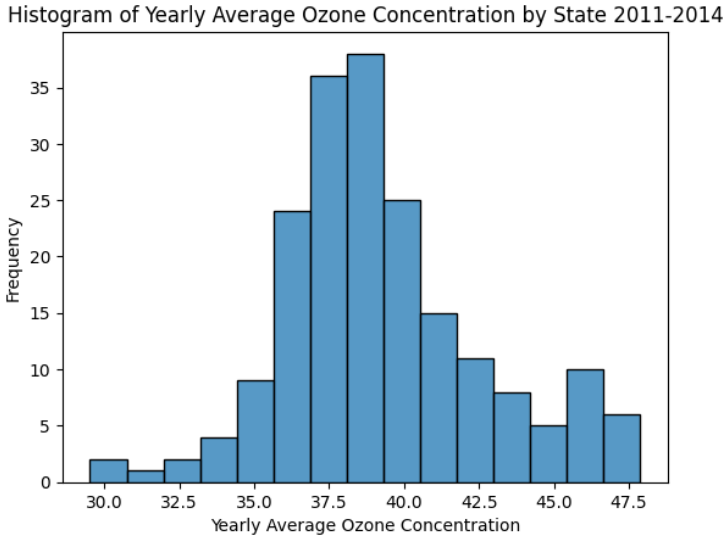
3 Exploratory Data Analysis (EDA)

3.1 Question 1 EDA

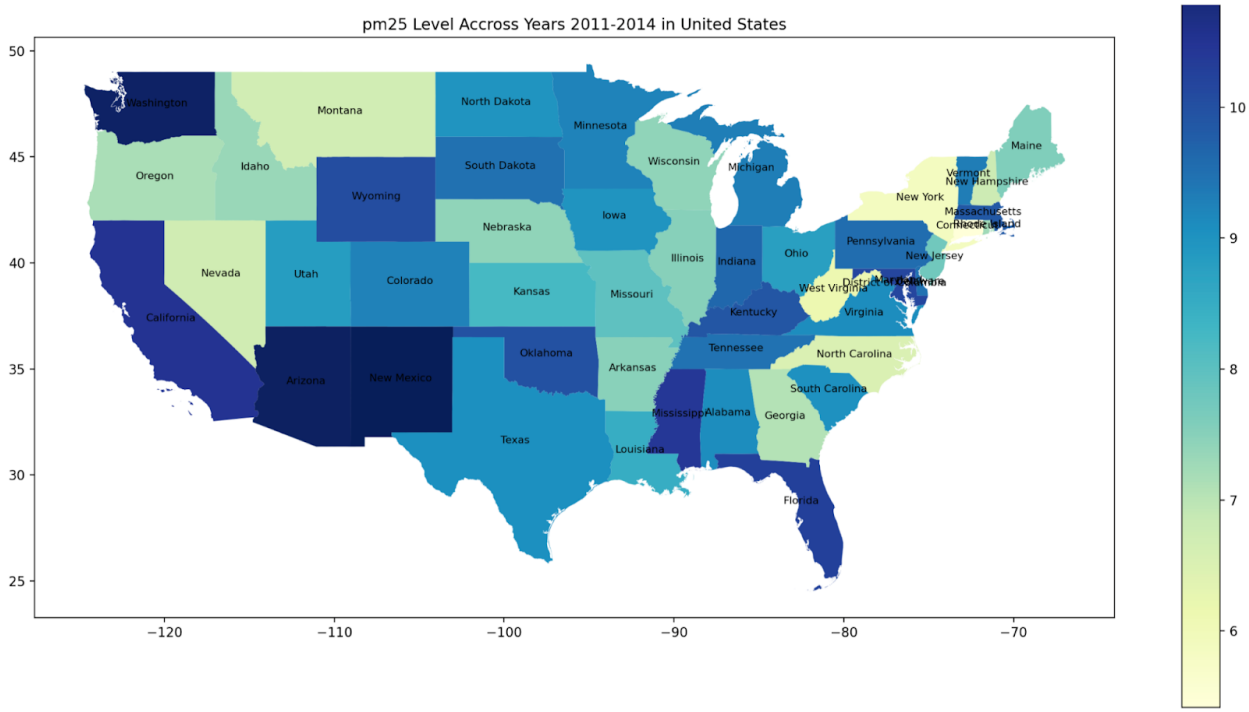
Visualizations



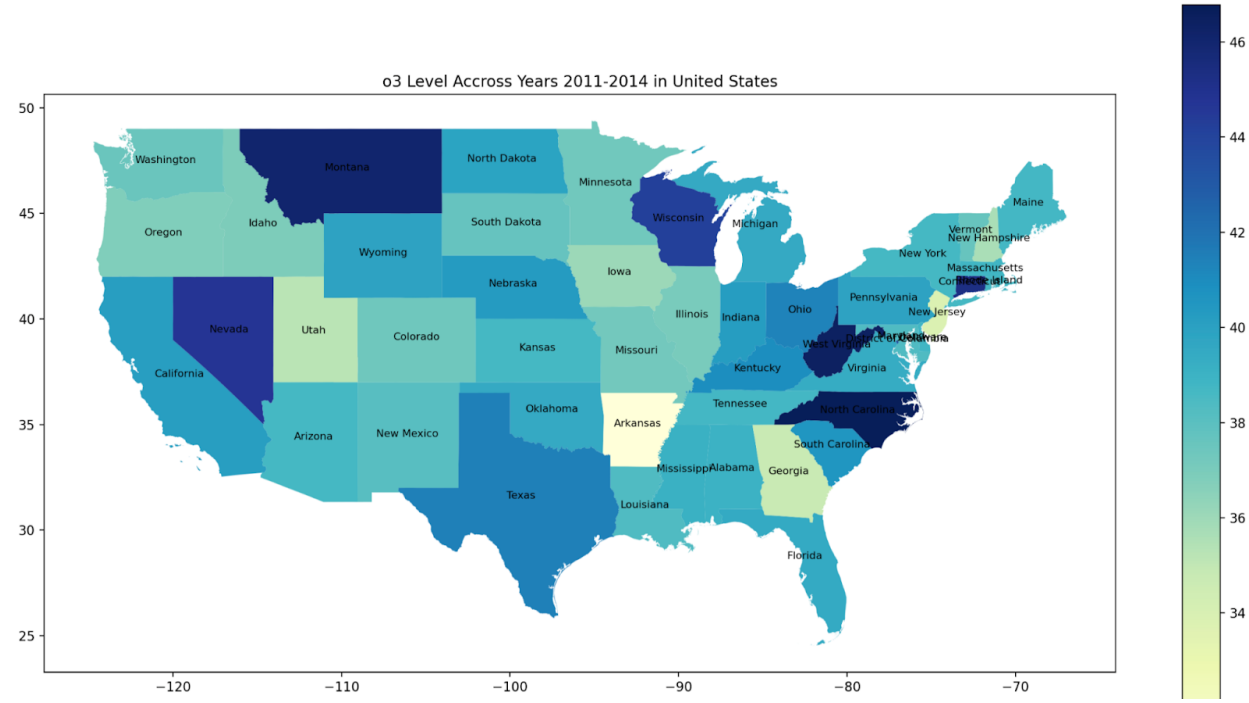
This histogram represents yearly PM2.5 Concentrations from 2011-2014 in each of the 49 states that are represented in the ozone concentration and asthma datasets. From this histogram we find that the average state PM2.5 average is approximately 8.6405 with a standard deviation of 1.5281. The distribution is unimodal and slightly left skewed.



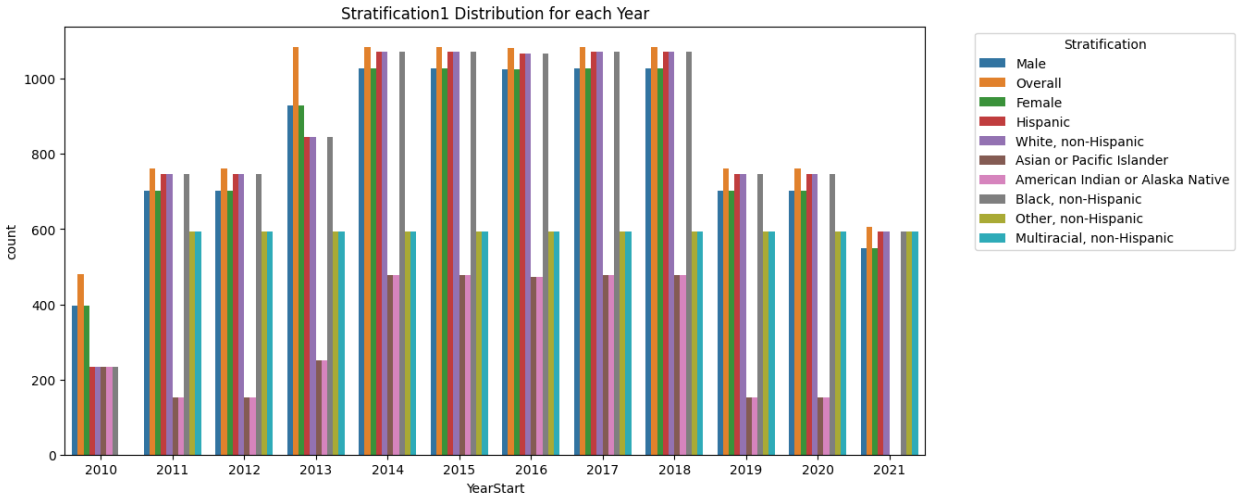
This histogram represents the ozone concentrations from 2011-2014 in 49 states that are represented in the PM2.5 concentrations and asthma datasets. From this histogram we can fine that the average ozone concentration average is approximately 39.2671 with a standard deviation of 3.3990. The distribution is fairly unimodal and normally distributed.



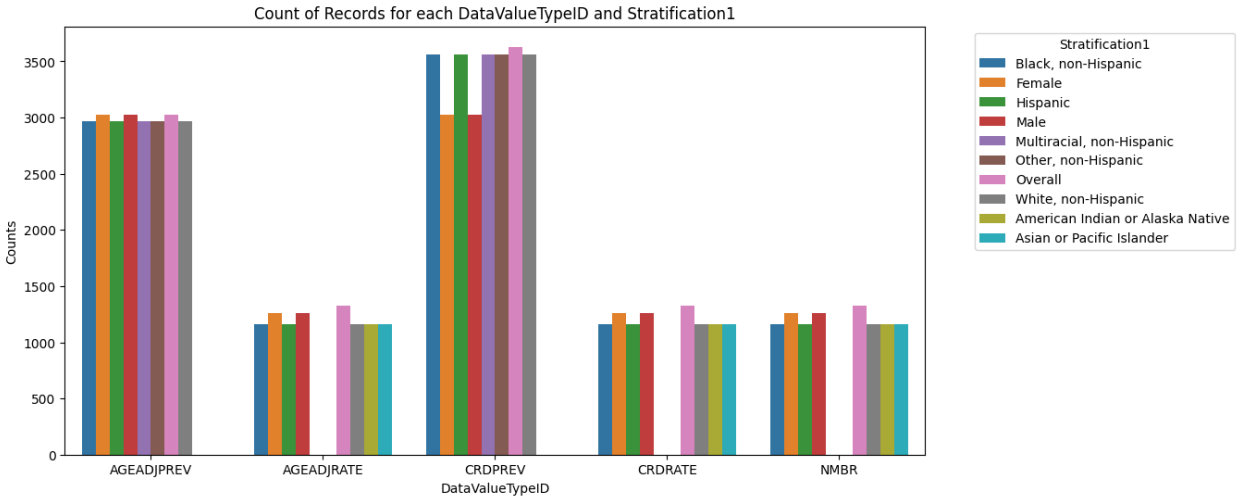
The above map of the United States visually represents the levels of PM2.5 particulate matter pollution across 49 states. The map is color-coded, with darker shades indicating higher levels of pollution, and displays the average PM2.5 concentration for each state over the years 2011-2014. The title of the map is "PM2.5 Level Across Years 2011-2014 in the United States" and it provides a clear and concise overview of the pollution levels across the country.



The above map of the United States visually represents the levels of ozone pollution across 49 states. The map is color-coded, with darker shades indicating higher levels of pollution, and displays the average ozone concentration for each state over the years 2011-2014. The title of the map is "o3 Level Across Years 2011-2014 in the United States" and it provides a clear and concise overview of the pollution levels across the country.



This assemblage of bar plots represents the count stratification categories available for all of the different asthma research questions for each year. This visualization is useful in determining which of the stratification categories is most useful in generalizing the model. When closely analyzing the barplots we realize that the overall category encapsulates all of the values stratification values well and this led us to decide this as our stratification value.



This assemblage of bar plots represents the count of data value types for each asthma stratification category. This visualization is useful because it helps us determine which of the data values are most prevalent in the different stratification categories. When closely analyzing

the bar plots we realize that the crude prevalence data type is the most prevalent among the overall category and all other categories as well.

4 Results

4.1 Question 1 Results

Our Question

Does the yearly state average exposure to air pollution, particularly fine particulate matter (PM2.5) and Ozone, increase the risk of asthma in the US population?

Methods

We decided to use multiple hypothesis testing for our analysis of this question. In multiple hypothesis testing, it makes sense to conduct many hypothesis tests instead of just one because we often want to explore multiple relationships, effects, or comparisons within the same dataset. While multiple hypothesis testing is useful in these situations, it comes with a caveat: the risk of false discoveries (Type I errors) increases as we conduct more tests. To address this issue, we need to apply techniques, such as the Bonferroni correction or the False Discovery Rate (FDR), to adjust the significance level and maintain the accuracy of our conclusions.

Calculating P- Values

To calculate p-values we need to make both pm2.5 and ozone averages into binary values. We have decided to encode pm2.5 values with 0 if they are less than 9.751921804482512 (ie. within the 0-25th percentile). We have decided to encode ozone values with 0 if they are less than 41.058259064972034 (ie. within the 0-25th percentile).

To calculate our p-values we decided to write a function called `p_values` that calculates the p-values for a multiple linear regression model with two independent variables and one dependent variable. This is how our function works:

1. Extracts the independent treatment variables (X) (i.e. ozone and pm2.5 binary concentration averages) and dependent variable (y) (i.e. asthma crude prevalence) from the input data.

2. Fits a multiple linear regression model using the independent and dependent variables.

3. Calculates the residuals, which are the differences between the observed values of the dependent variable and the predicted values from the model.

4. Calculates the residual sum of squares (RSS) and the mean squared error (MSE) by dividing the RSS by the degrees of freedom (number of observations minus the number of independent variables minus 1), given by:

$$\text{MSE} = \frac{\text{RSS}}{n - p - 1}$$

   where n is the number of observations and p is the number of independent variables.

5. Calculates the standard errors for the regression coefficients by finding the square root of the diagonal elements of the product of the MSE and the inverse of the matrix product of the transposed independent variables matrix (X) and the original independent variables matrix (X), given by:

$$\text{SE} = \sqrt{\text{diag}\left(\text{MSE} \times (X^T X)^{-1}\right)}$$

6. Calculates the t-statistics for each regression coefficient by dividing the coefficients by their respective standard errors, given by:

$$t_i = \frac{\beta_i}{\text{SE}_i}$$

   where β_i is the coefficient for the i-th independent variable and SE_i is the standard error for the i-th independent variable.

7. Calculates the p-values for each t-statistic using the cumulative distribution function (CDF) of the t-distribution with the given degrees of freedom, given by:

$$p_i = 2 \times \left(1 - F_{t_{n-p-1}}(|t_i|)\right)$$

   where $F_{t_{n-p-1}}$ is the CDF of the t-distribution with $n - p - 1$ degrees of freedom and t_i is the t-statistic for the i-th independent variable.

8. Returns the calculated p-values.

We calculate the p-values for all of the of our possible research questions:

1. Asthma prevalence among women aged 18-44 years
2. Influenza vaccination among noninstitutionalized adults aged 18-64 years with asthma
3. Pneumococcal vaccination among noninstitutionalized adults aged >= 65 years with asthma
4. Current asthma prevalence among adults aged >= 18 years
5. Influenza vaccination among noninstitutionalized adults aged >= 65 years with asthma
6. Pneumococcal vaccination among noninstitutionalized adults aged 18-64 years with asthma

Results

P-Value Results

Our p-value results and their respective questions:

1. Asthma prevalence among women aged 18-44 years
2. Influenza vaccination among noninstitutionalized adults aged 18-64 years with asthma
3. Pneumococcal vaccination among noninstitutionalized adults aged >= 65 years with asthma
4. Current asthma prevalence among adults aged >= 18 years
5. Influenza vaccination among noninstitutionalized adults aged >= 65 years with asthma
6. Pneumococcal vaccination among noninstitutionalized adults aged 18-64 years with asthma

	pm25_pval float...	o3_pval float64
0	0.38476408647637816	3.3368441078485134e-05
1	0.018303867199827240	0.002801006544647766
2	0.07110581337055555	0.8534483442573317
3	0.28516571622328524	0.0030567353288870613
4	0.2065060871844273	0.5502754399408638
5	0.45317599258330565	0.15718172804173203

Bonferroni Correction Results

```
Adjusted significance level (Bonferroni correction): 0.004166666666666667

Rejected null hypotheses for o3 (True = rejected, False = not rejected):
0    False
1    False
2    False
3    False
4    False
5    False
Name: pm25_pval, dtype: bool

Rejected null hypotheses for pm25 (True = rejected, False = not rejected):
0     True
1     True
2    False
3     True
4    False
5    False
Name: o3_pval, dtype: bool
```

The Bonferroni correction method helps us correct the family-wise error rate (FWER). This helps us minimize the amount of times we could possibly incorrectly reject the null hypothesis.

For our analysis we fail to reject the null for all of our different stratification questions with regards to PM2.5 concentrations.Additionally we do reject the null for questions 1,2, and 4 of our different stratification questions with regards to ozone concentrations meaning there is not enough evidence to conclude a significant relationship between ozone and asthma crude prevalence. We do not reject the null however for questions 3,5, and 6.

Benjamini-Hochberg Results

```
Benjamini-Hochberg critical values:
   pm25_pval   o3_pval
0   0.041667  0.008333
1   0.008333  0.016667
2   0.016667  0.050000
3   0.033333  0.025000
4   0.025000  0.041667
5   0.050000  0.033333

Rejected null hypotheses for o3 (True = rejected, False = not rejected):
0    False
1    False
2    False
3    False
4    False
5    False
Name: pm25_pval, dtype: bool

Rejected null hypotheses for pm25 (True = rejected, False = not rejected):
0     True
1     True
2    False
3     True
4    False
5    False
Name: o3_pval, dtype: bool
```

The Benjamini-Hochberg correction controls the false discovery rate (FDR). The FDR is the expected proportion of false positive findings among all reject null hypotheses. The B-H correction is less conservative than methods that control FWER. In being less conservative the B-H correction allows for a higher rate of false positives in exchange for increased power to detect true effects.

For our analysis we fail to reject the null for all of our different stratification questions with regards to ozone concentrations. Additionally we do reject the null for questions 1,2, and 4 of our different stratification questions with regards to PM2.5 concentrations meaning there is not enough evidence to conclude a significant relationship between PM2.5 and asthma crude prevalence. We do not reject the null however for questions 3,5, and 6.

Discussion

After applying correction procedures for multiple testing, it appears that none of the discoveries remain significant at a standard alpha level of 0.05. The Bonferroni correction, for example, would adjust the alpha level to 0.05/6 = 0.0083. None of the p-values for the individual tests fall below this threshold, indicating that the null hypothesis cannot be rejected at this level of significance.

From the individual tests, we can conclude that there is no evidence to support the hypotheses that were tested. This means that we cannot conclude that there is a significant association between yearly state average exposure to air pollution, particularly PM2.5 and Ozone, and asthma prevalence.

Taken together, the results suggest that there may not be a significant relationship between exposure to air pollution and asthma in the US population. However, it is important to

note that this does not necessarily mean that there is no relationship. There may be other factors that were not accounted for in the analysis that could be influencing the results.

A limitation of this analysis is that it only considers two specific pollutants, PM2.5 and Ozone, and does not account for other potential environmental factors that may contribute to asthma prevalence. Additionally, the analysis is limited to observational data, which means that it cannot establish a causal relationship between exposure to air pollution and asthma.

To avoid p-hacking, the analysis could have been pre-registered before conducting the tests, specifying the hypotheses to be tested and the statistical methods to be used. This would help to reduce the risk of selectively reporting only significant results.

If more data were available, additional tests could be conducted to explore other potential relationships between air pollution and asthma prevalence. For example, other pollutants could be considered, and more sophisticated statistical methods could be used to account for potential confounding variables. Additionally, the analysis could be expanded to consider other outcomes related to respiratory health, such as hospitalizations or mortality.

4.2 Question 2 Results

Methods

The causal question that we decided to investigate for this report was: What is the causal effect of PM2.5 concentrations on the prevalence of asthma among individuals in the United States over a specified time period?"

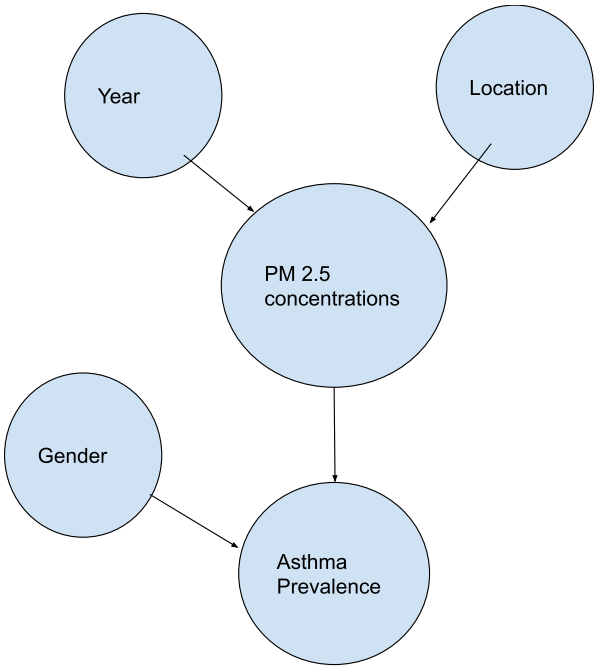
The PM2.5 concentrations represent the treatment, the prevalence of asthma represents the outcome, and the unit of analysis is the individuals in the US who are affected by asthma. Our confounding variables were the following from the Asthma data set: DataValue, Location, Year, DataSources, Female/Male, Pm25_binary. DataValue is a confounding variable because it represents the number of asthma hospitalizations per state and thus could affect the prevalence of asthma (outcome) and could also be impacted by the PM2.5 concentrations. The DataSources different sources have different metrics and methodologies for determining if someone has asthma or not but this data could also affect PM2.5 concentrations because of the same reason - independently, which is what makes it a confounding variable. The Female/Male encodings are also confounding variables because gender at birth may have different biological functions and could result in varying outcomes of the prevalence of asthma and independently could be impacted by PM2.5 concentrations. Another potential confounding variable is the year that is aggregated. We know from the lecture that this can be a potential confounder when we try to aggregate data; the outcome can change due to Simpson's paradox and can violate the Stable Unit Treatment Value Assumption. The last confounding variable we found was the Pm25_binary - the treatment variable will be either 1 or 0, depending on the severity of the air quality, the control and the treatment, and thus could impact our treatment and outcome variables differently.

Results

In order to adjust for confounders, we used the Propensity Score Matching technique. This will allow us to estimate the probability of the PM2.5 concentrations and then we can match

the individuals who have similar propensity scores and different treatment assignments in order to create balanced treatment and control groups. This method will help with eliminating the influence of confounding variables and in estimating the effect of the treatment.
The colliding variables in the datasets were the following: Year and Location, as shown in our causal DAG. Our collider variables are affected by both the treatment (PM2.5 concentrations) and the outcome (prevalence of asthma). The Location variable is a collider variable because different places in the US could be affected by PM2.5 concentrations and could also be affected by the prevalence of asthma in that state. The Year is a collider variable because it could be affected by changes in PM2.5 concentrations which could have potential increases/decreases and it could also be impacted by the prevalence of asthma because of genetics/births of children with asthma in that given year. Since the lower and upper bounds of the confidence interval were -12.7088 and 14.1773, this may also indicate that the ATE may be slightly different from an actual point estimate, and we would need to further estimate to confirm the causal relationship with more certainty.

Causal DAG:



Discussion

One of the limitations we found within this research question was the controlling of confounding variables. We realized that it was a challenging process to decide which variables were confounders and which were colliders and that if we did not include any confounding variables in our analysis, it could affect the results in different ways that we could not control. Another limitation of our methods was the data itself. Since we could not choose the quality or choose data with specific metrics in terms of where and how it was collected, there could be measurement errors, errors in values, or selection bias that could create negative effects on the confidence of our causal question.

Additional data that could have been useful for answering this causal question would have been more confounding variables that create more of an intricate and complex understanding of how PM2.5 concentrations affect the prevalence of asthma. For example, variables such as socioeconomic status, smoking history, or medical/genetic history could help with drafting a much deeper picture of how these variables may affect the causal relationship we were researching.

In terms of our confidence in the causal relationship between PM2.5 concentrations and asthma prevalence, with the acknowledgement that a deeper study and more complex methodology of experimentation and research would most definitely result in a higher confidence level, we are fairly confident that there is a causal relationship between our chosen treatment and outcome because our resulting ATE points towards the presence of a causal relationship.

4 Conclusion

In our multiple hypothesis testing model analysis we fail to reject the null for all of our different stratification questions with regards to PM2.5. Additionally we do reject the null for questions 1, 2, and 4 of our different stratification questions with regards to ozone concentrations meaning there is not enough evidence to conclude a significant relationship between ozone and asthma crude prevalence. We do not reject the null however for questions 3, 5, and 6.

From our analysis there is some evidence to conclude a relationship between PM2.5 concentrations and asthma crude prevalence, however there are several confounding factors that we must consider when directly associating PM2.5 concentrations and asthma crude prevalence. Confounding factors that directly affect PM2.5 concentrations could be linked to certain types of industry in the areas where the PM2.5 data is collected that have additional chemicals and pollutants that this study does not take into consideration. Additionally there could be several other confounding factors that directly affect asthma prevalence such as whether a person smokes or is exposed to second-hand smoke on a regular basis. This study does not take into account numerous other causal factors thus we must take these results with a grain of salt when determining the exact cause of asthma prevalence in the population.

Because we do find that there is evidence suggesting that PM2.5 concentrations have an adverse effect on asthma prevalence in the population, it could prove to be useful for some government intervention. The government should conduct studies on airborne pollutants to at least classify which pollutants have the strongest quantifiable effect on the population. While this would only be a start, it could help reduce the amount of those in the population that suffer from this respiratory disease as well as numerous others that directly affect the health of people not only domestically but around the globe.

Our analysis was slightly limited because of the way our data is structured. Ideally we would have been able to directly compare the crude prevalence of asthma with the entire population rather than comparing several different sub-populations within the population. While the dataset does encapsulate the majority of the population our model is not as generalizable because of this facet of our datasets.

In our causal inference study the average treatment effect ATE of the PM2.5 concentrations on asthma prevalence was 8.5011, with a lower bound of -12.7088 and an upper bound of 14.1773. This points us towards a causal relationship between PM2.5 concentrations and asthma prevalence. Since the magnitude of the effect is on the smaller side we can conclude that PM2.5 concentrations could also have a limited effect on asthma prevalence due to uncertainty. Although merging datasets created a more diverse understanding of the research we were trying to accomplish, we faced challenges with integrating some of the data and figuring out how to create compatibility between different sources.

One of the limitations we found was understanding how to choose confounding and collider variables within our datasets. We also could not directly control the quality of our data, which could result in selection bias or general bias within our data. In terms of future studies, I believe we could address some of the limitations that were mentioned by utilizing more comprehensive datasets, in addition to exploring any other confounding variables that could affect our results. Perhaps, a different method such as Bayesian Hierarchical Modeling could

also be used by determining relationships between how PM2.5 concentrations may affect asthma prevalence with different implications, such as geographical, socioeconomic, and age related models. Based on our research and causal inference model, we can consider the implications of our project to be focused on making change with the impact of air pollution and the impact that weak environmental policies have for corporations in the US. I think that it is important for us to think about how to implement stricter policies within pollution control measures, in addition to allocating more resources to protecting those who suffer from asthma and other cardiovascular issues.

In light of this project, we believe healthcare professionals and government officials should come together to address how we can work together to reduce carbon emissions and prioritize the care of those who are affected by these ailments. By combining different data sources, we were able to draft a more comprehensive dataset that could capture the different aspects of PM2.5 concentrations and asthma prevalence. Overall, our group learned several different things throughout this project. We discovered the importance and the challenge of determining confounding variables and understanding which methods will best serve us when it comes to addressing those variables. We also further solidified our understanding of how to use causal inference methods in order to create a call to action.