Table of Contents

Data Overview

Dataset 1 CDC: Daily Census-Tract Ozone Concentrations

The dataset has information on predicted ozone levels at the census tract level across the United States from the years of 2011 to 2014. According to the metadata, the dataset contains estimates for each of the 2010 US Census Tracts within the contiguous United States. Therefore, this data seems to be from a census.

We included data on confounding variables from the American Community Survey (ACS) to include in our models. In particular, data on state unemployment rates and poverty rates in 2014 were collected from the Selected Economic Characteristics dataset to be included in our analysis. We considered other rates found in the ACS but ultimately decided not to include them since we could not see how they were confounding. It was important to include the data on the confounders because we would have a severely misspecified model if we failed to include them and any conclusions drawn from such a model will not likely be accurate.

Since there is only mention of census tracts within the contiguous United States, this suggests the exclusion of states like Alaska and Hawaii, as well as overseas US territories. Since the data was generated from the US Environmental Protection Agency Air Quality System (AQS) and a deterministic prediction model, the "participants" were probably not aware of the collection or use of this data.
What is the granularity of your data? What does each row represent? How will that impact the interpretation of your findings?

Each row represents a census tract in the US. This means that we can only really analyze the aggregate effect of ozone concentration on asthma hospitalizations rather than analyzing the effect on individuals.

Selection bias and measurement error are relevant concerts in context of this data. Selection bias is a concern because certain areas of the US were not included in the dataset. Measurement error is a concern because the data in the dataset are predictions rather than actual measurements.There does not appear to be convenience sampling occurring here. This dataset does not appear to be modified for differential privacy.

One possible feature that might be helpful is population size and area of each census tract. This might provide more information on just how many people are exposed to each concentration of ozone found in the dataset. There does not appear to be missing data in this dataset.

Since the dataset was pretty large, we had to filter by year to and then concatenate each year's dataset to reproduce the whole dataset. We then dropped all the columns from the dataset except "statefips" and "ds_o3_pred" where "statefips" is the states FIPS code and "ds_o3_pred" is the estimated 24-hour average ozone concentration for each census tract. These are the only columns we needed to merge to the other datasets. Including these columns, as well as the columns from the other datasets, ensures that we have specified the model as best we can in order to be able to draw the most accurate conclusions.

Dataset 2 CDC: Daily Census-Tract PM2.5 Concentrations, 2011-2014

The dataset contains information about daily concentrations of PM2.5, a type of air pollution, at the census tract level across the United States from 2011 to 2014. This data represents a sample of census tracts rather than a complete census, so I would consider it to be sample data.

We incorporated data from the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program and the U.S. Census Bureau's American Community Survey (ACS) 5-Year Estimates in order to obtain more detailed information about poverty and unemployment rates at the State level. For our research purposes, we needed to obtain information on unemployment rates and poverty rates. Therefore, we incorporated this data into our dataset.

The 'ds_pm_pred' provides the predicted PM2.5 concentration of the state. Some states that have historically experienced high levels of air pollution and correspondingly high PM2.5 levels include California, Texas, Ohio, Pennsylvania, and Illinois. Therefore, it would be expected that these states have a high 'ds_pm_pred'. We found that this was true for the majority of the data from these states. If these states are overrepresented in the dataset, it could lead to a bias in the analysis and affect the generalizability of the results. This is because the results may not accurately reflect the population as a whole, and the findings may not be applicable to other states or regions with different pollution levels.

This data was generated through the monitoring of PM2.5 concentrations by the United States Environmental Protection Agency (EPA). In terms of granularity is at the census tract level so each row represents the daily PM2.5 concentration for a specific census tract.This can impact our interpretation of findings because it is at the census tract level and does not provide much information about individual exposure to PM2.5 and direct relation to asthma hospitalizations. This level of detail is important as people move from state to state.

Selection bias and measurement error are relevant concerns in the context of this dataset. Selection bias may occur if the data was only collected in urban areas with high levels of pollution. If this were the case, the results may not be generalizable to rural areas or areas with lower levels of pollution. Measurement error may occur if there are inaccuracies in the measurement of PM2.5 concentrations. This may occur through human error when recording data or this may occur if the instruments used to measure PM2.5 concentrations were used improperly.

This dataset was not modified for differential privacy

It would have been extremely useful to have individual-based data. Because the dataset does not provide much information about individual exposure to PM2.5, it is difficult to truly associate individual asthma cases to an entire state's PM2.5 concentration. Such broad generalizations may lead to inaccuracies. Individual-level factors such as age, gender, socio-economic status, and smoking history of those living in places with high PM2.5 concentrations would have been useful data to compare with Asthma hospitalizations.

There are three columns with missing data on the same row. The columns are "longitude", "ds_pm_pred", and "ds_pm_stdd". The missing data is only on one row, so there are only 3 data entries that are missing from the dataset. The missing dataset is probably due to lack of information. In order to accommodate for the missing data, we will remove the row with the three missing data entries. This will not drastically affect our results because there are 2568594 rows in total.

For our analysis, we found that only a few rows were helpful for us. Therefore we dropped all of the columns except for "statefips" and "ds_pm_pred".

- statefips - State FIPS code
- ds_pm_pred - Mean estimated 24-hour average PM2.5 concentration

We merged this dataset with the asthma dataset by using the state FIPS code. Therefore, the statefips was necessary in order to merge with the Asthma dataset and the PM2.5 concentration was necessary for us to model our research question. By adding two columns (unemployment_rate, poverty_rate), we had finally prepared our dataset for our research question. Because of our preprocessing, our model will probably have a high accuracy rate.

Dataset 3 CDC: Annual State-Level Chronic Disease Indicators: Asthma

Since this dataset does not appear to cover all ages and populations across the country, I would consider it to be sample data.

This data has several filters relating to the prevalence of asthma, other diseases, and race and gender, but I would not say it is representative of the populations. It seems that this dataset was constructed with samples known to relate to asthma instead of samples collected from the general population since the categories are “Asthma prevalence among women aged 18-44”, “Emergency department visit rate for asthma” , etc. This tells us that our results will be built on a very specific subset of the population and we have to remember in our results that we cannot generalize to the greater population.

This dataset excludes individuals that are less than 18 years old. This could be due to the fact that asthma is a more prevalent issue in adults, or that asthma developed later in life is more related to environmental factors than asthma in early life.

In terms of the granularity, each row has a YearStart and a LocationAbbr column representing the Year from which the data was collected and the State from which it was collected. From there, there are many filters for the data such as what the raw number represents (ex. Emergency hospital rate for asthma versus number of asthma cases per 10,000 people) and the qualities of the individuals in each row (ex. Male, Female, Race, Overall). Each row represents all the collected data within the specified year start and year end within that state. The data value provides a numerical answer to the “Question” column with a short data type description in the “DataValueType” column.

Considering that this is sample data most likely collected manually across states, there is likely some presence of selection bias, measurement error, as well as convenience sampling. It is difficult in practice to fully randomly collect data. Asthma cases are likely to have been collected with some association to hospitals. Furthermore, there are likely undiagnosed asthma cases that aren’t taken into consideration resulting in selection bias as well as measurement error.

The dataset itself is not modified for differential privacy. No personal information is revealed however.

Important features that are missing include confounding variables for asthma such as pre-existing health conditions. These include but are not limited to whether they smoke, have obesity, previous long term exposure to harmful chemicals, etc. These variables could help us answer whether the difference in asthma cases is only attributed to geographical location or whether it's due to other lifestyle factors.

A significant number of columns as well as rows have no values. The missing column and row data don't seem to hold much meaning aside from perhaps not having enough information as well as resources to compile a more comprehensive data set. Working around the NaN values, we simply filtered them out when we were data cleaning. Moreover while manually examining through the data, we found that the majority of the relevant information is captured in the following columns:

- YearStart - what year the data started being collected from
- LocationAbbr - State
- Question - What question does the DataValue column answer
- DataValue - Actual captured data
- StratificationCategory1 - stratification type separated into (Gender, Race, Overall)
- Stratification1 - actual identified stratification category

As a result, we eliminated the majority of the other columns in order to have a well cleaned dataset that can be modeled on. Because of our filtering, our model likely will have greater prediction accuracy than if we didn't filter the dataset. Consequently, our inferences could be oversimplified and won't necessarily apply to the greater general population.

Research Questions

Research Question 1

Question: How do levels of particulate matter and ozone affect the onset of asthma?

There are many potential real-world decisions that involve pollution and public health that could be made after learning if pollutant concentrations affect the onset of asthma. For example, if particulate matter is found to have a direct causal effect on the onset of asthma, policymakers might be more motivated to tackle the problem of pollution in order to mitigate the negative effects on public health. Explain why the method you will use is a good fit for the question (for example, if you choose causal inference, you should explain why causal inference is a good fit for answering your research question).

We have opted to answer this question with causal inference techniques. This is a good fit because if causality can be demonstrated, it would make the argument for addressing pollution in the interest of public health that much stronger.

One limitation of this method is the difficulty of demonstrating causality. Causality is notoriously difficult to demonstrate in many cases. If the assumptions that are required to make causal inference work are not valid, then it may not be possible to demonstrate causality, at least with the techniques covered in this course.

Research Question 2

Question: Are there any geographical trends relating to the prevalence of asthma?

We want to pursue this research question so that real-world decisions can be made about asthma research across the United States. If a certain region shows high correlation with asthma, policy-makers should ensure that the proper healthcare is available and that funding is delegated to health and environmental researchers to improve the conditions.

We are choosing to use prediction with GLMs and nonparametric methods. This is a good fit for answering the research question because this type of modeling allows us to feed in general geographic regions and examine how the model correlates each region with our target variable. One limitation of this method is that feeding in variables other than the geographic regions can diminish the differences we see between states / regions.

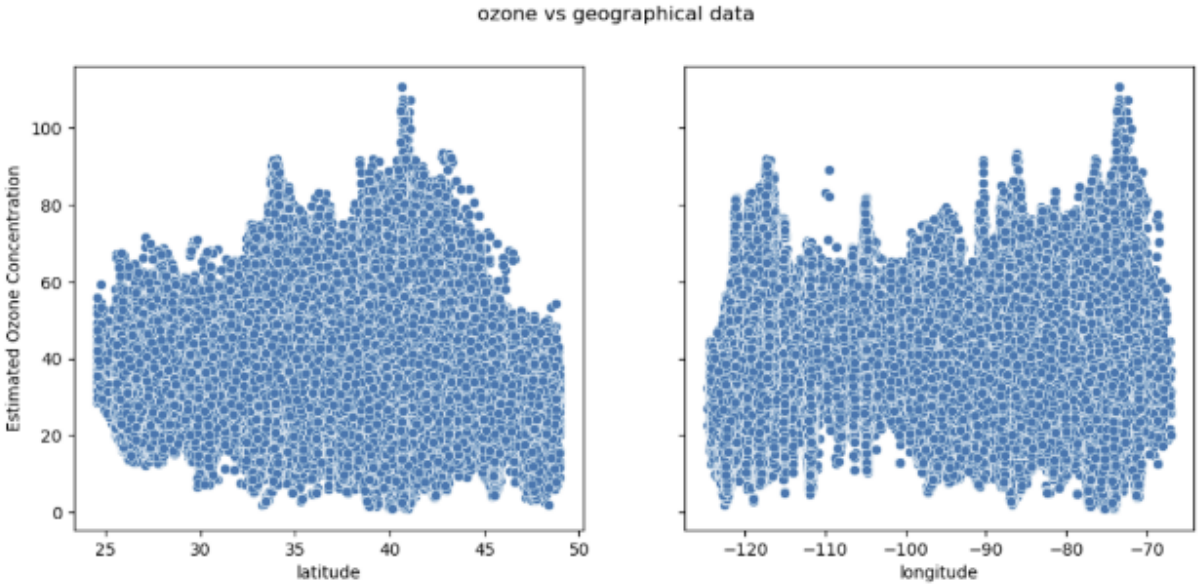
EDA

Research Question 1

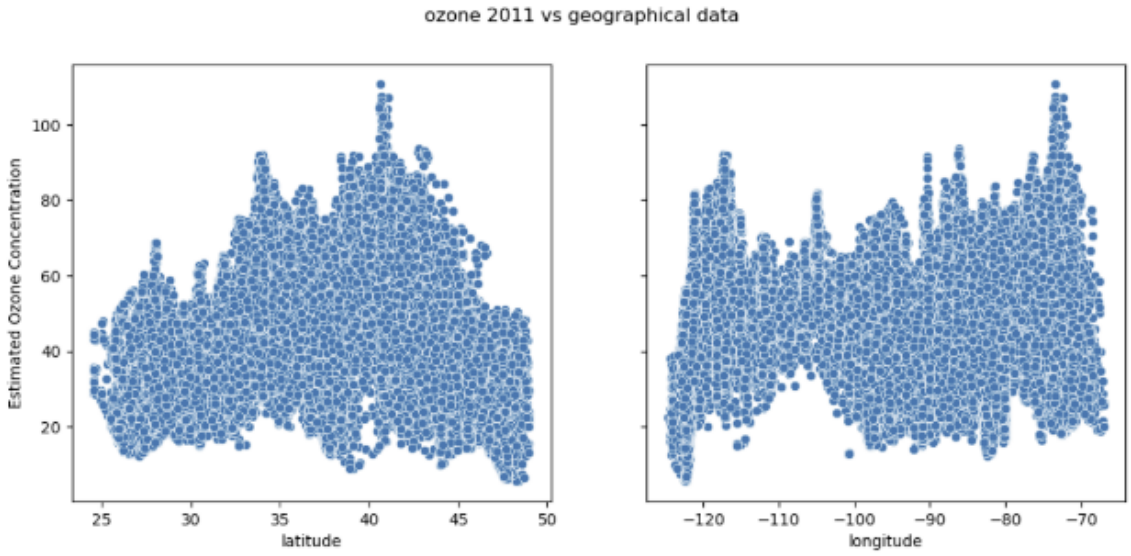
Data: CDC: Daily Census-Tract Ozone Concentrations

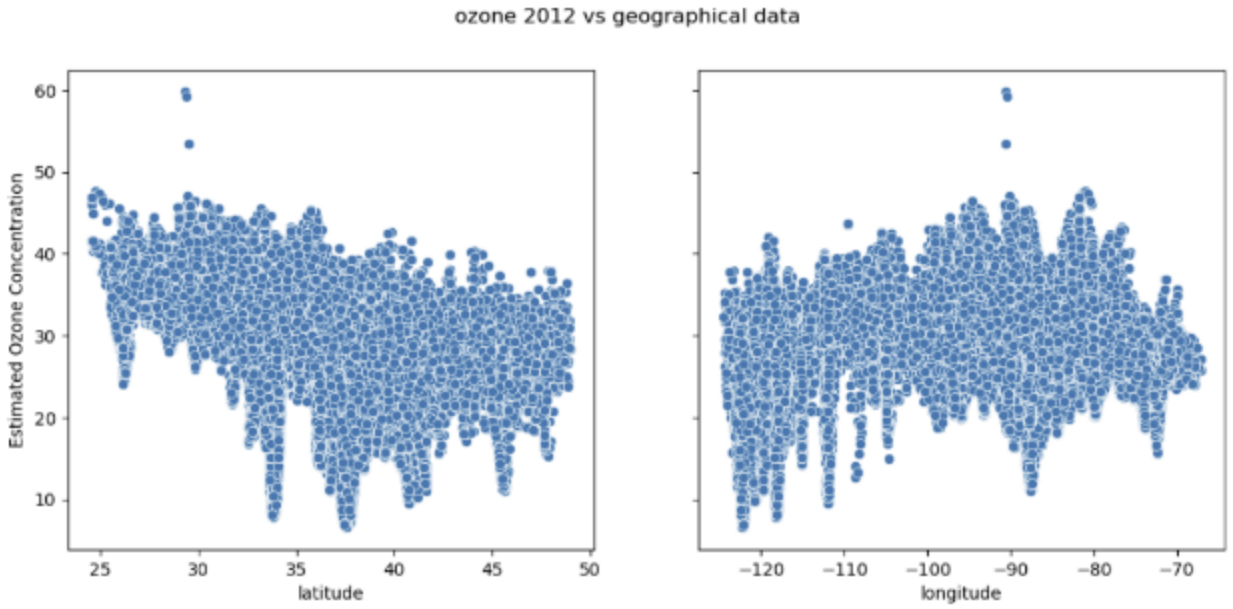
Quantitative Variable

We randomly sampled 100,000 data points from each year 2011-2014. There are 400,000 data points in total.

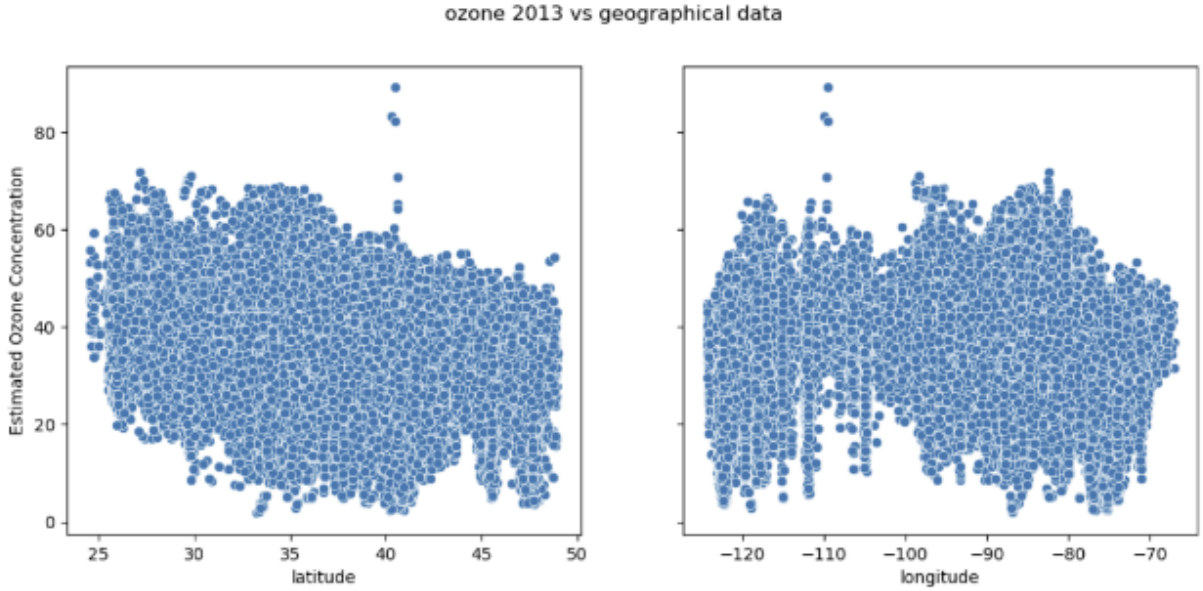


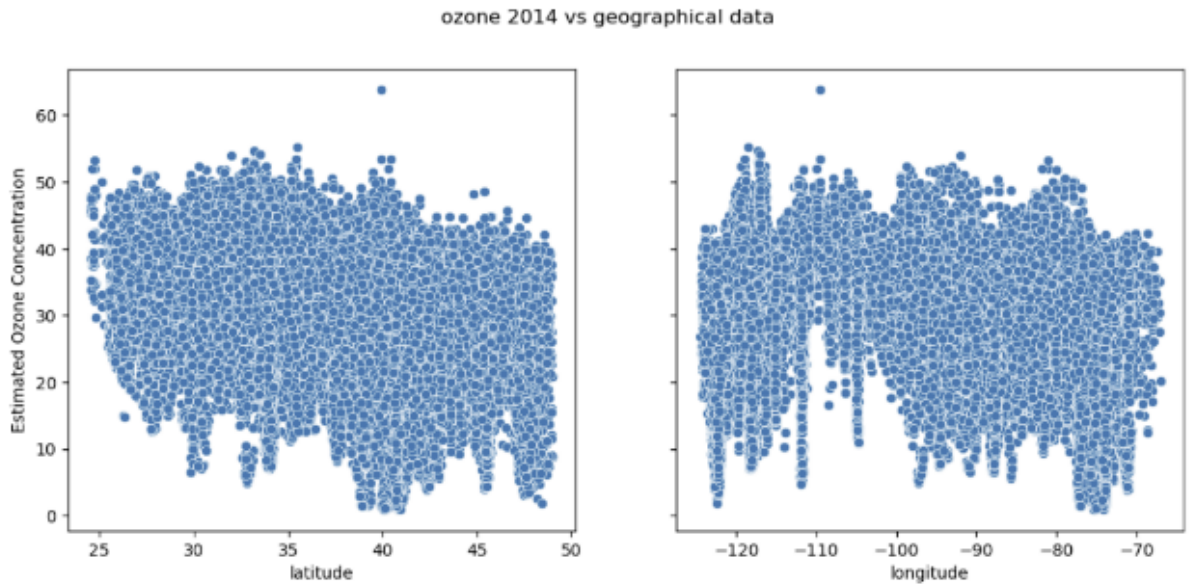
We estimated daily ozone concentration plotted against latitude and longitude data coordinates
Filtered by year:





The 2011 vs 2012 data seem to be a little different with the 2012 data fitting more closely to some sort of a linear correlation

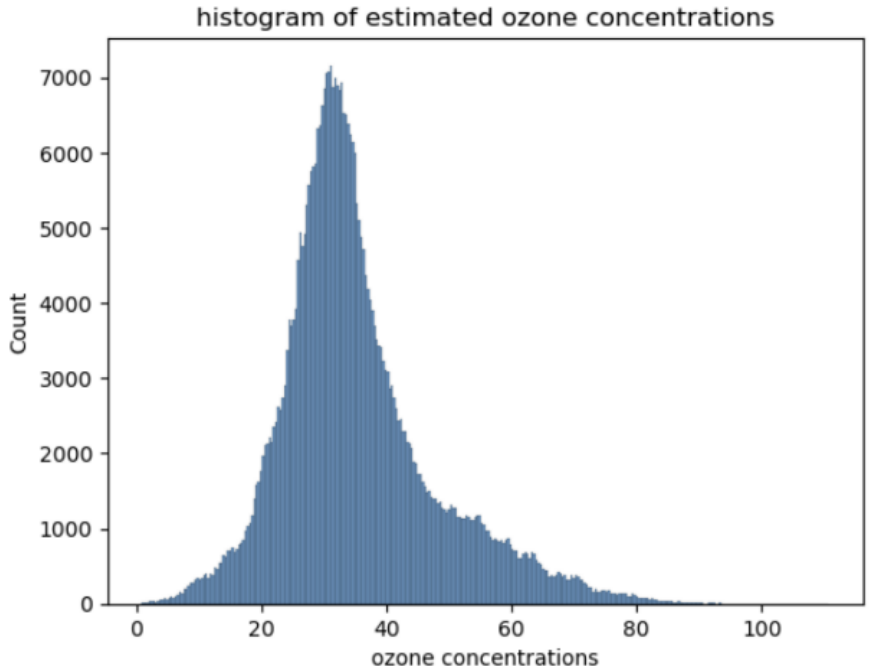




The 2013 vs 2014 seem pretty similar in comparison.

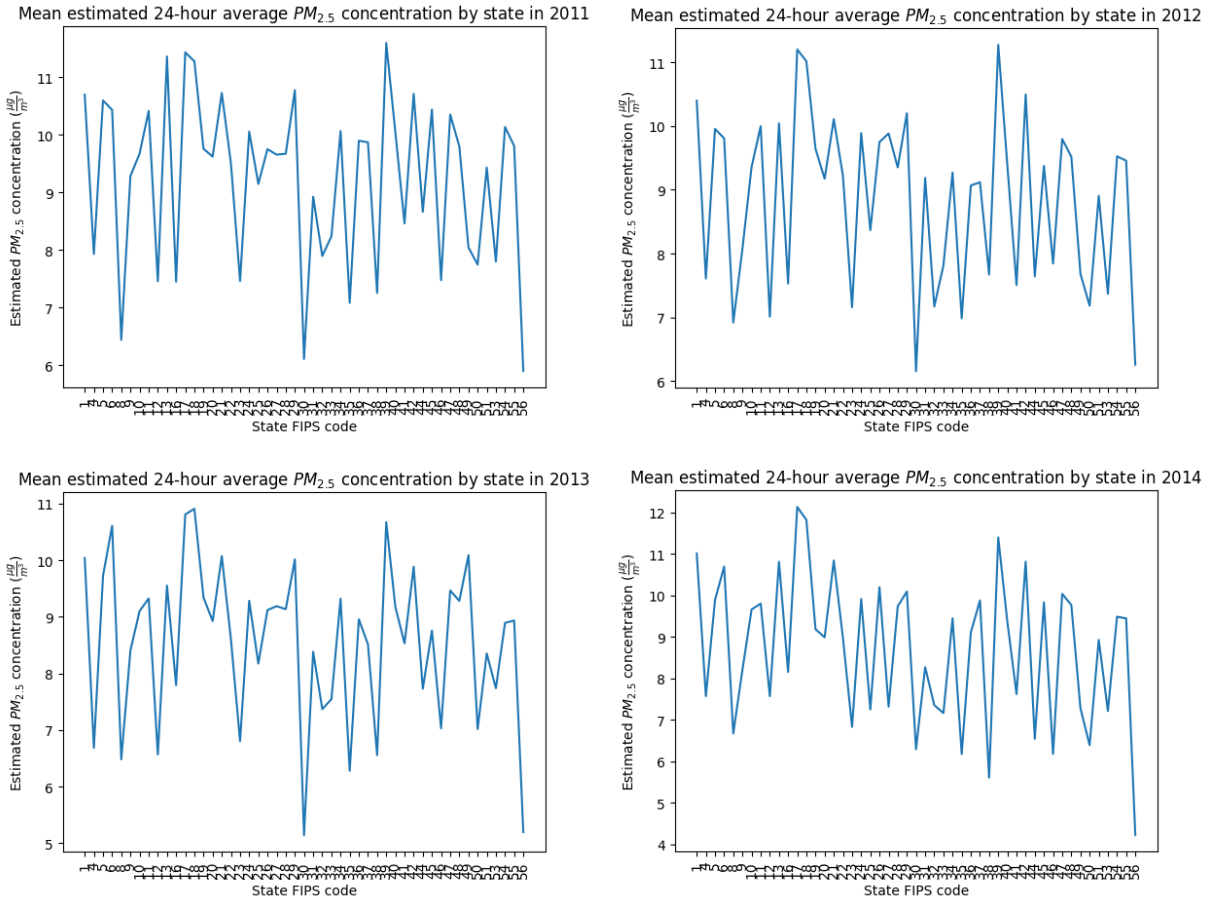
Overall, there seems to be some sort of trend between ozone data and latitude/longitude data. There also seems to be some correlation between the year the ozone data was collected in and latitude/longitude data. This will be helpful in our report as it shows us a probable trend between geographic variables and ozone concentration variables.

Histogram of estimated daily ozone concentrations distribution

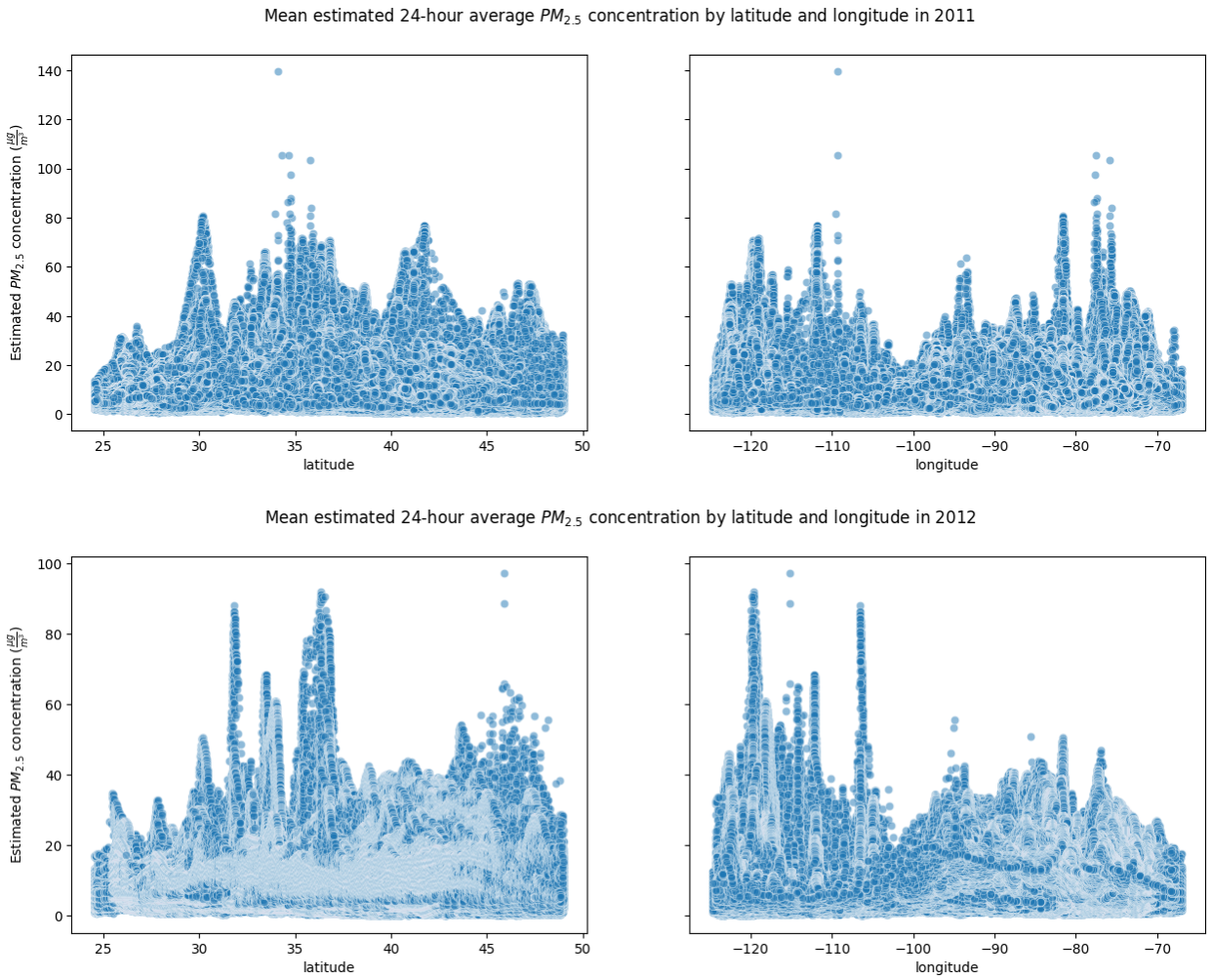


The average ozone concentration seems to be 35 with a skewed right tail

Qualitative Variable

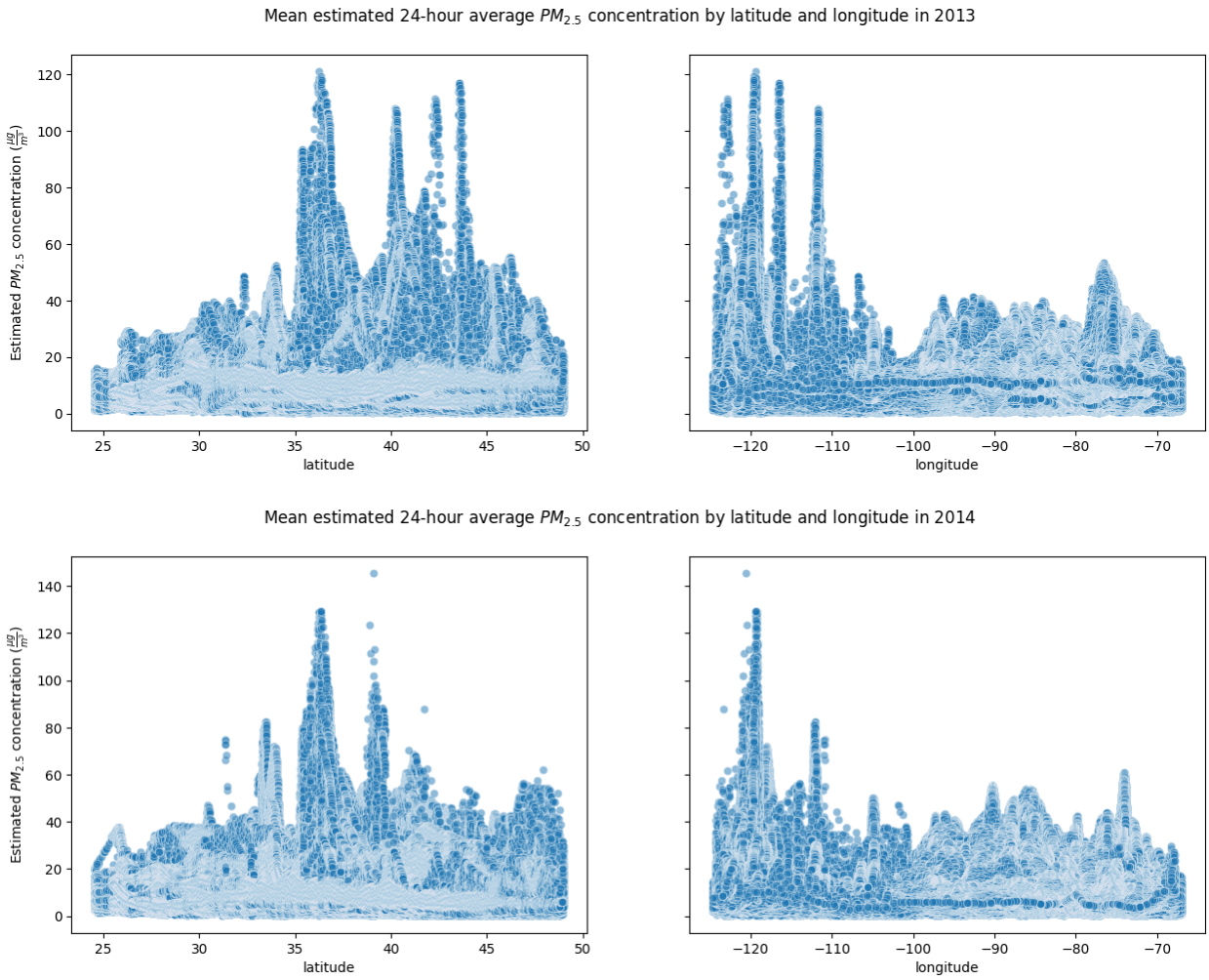


It is very difficult to see a trend in estimated PM2.5 concentration by using FIPS codes. Since FIPS codes are assigned alphabetically, they might not be the best for visualizing geographical trends. The only obvious trend is the increase in the range of possible values that estimated PM2.5 concentration takes across the states with a range of about slightly over 6 to slightly under 12 µg/m³ in 2011 to a range of slightly above 4 to slightly above 12 µg/m³ in 2014.



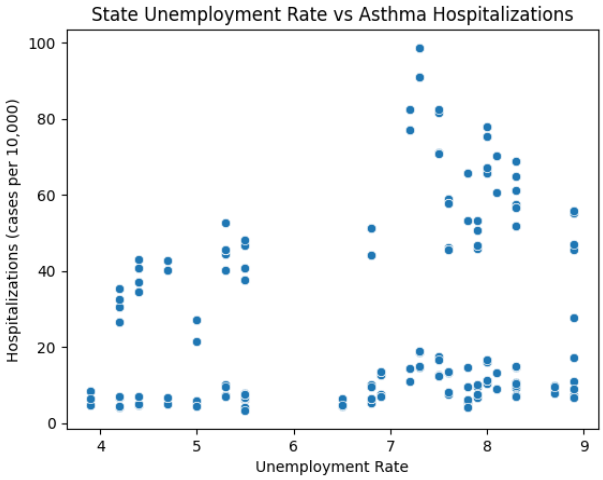
There appears to be an overall increase in PM2.5 concentration between 2011 and 2012 in both latitude and longitude. There appears to be a high concentration at about 33° to 37° latitude and -122° to -118° longitude. This area roughly corresponds to the state of California. There are a few other peaks in PM2.5 concentrations at other latitudes and longitudes that are also muted from 2011 to 2012.

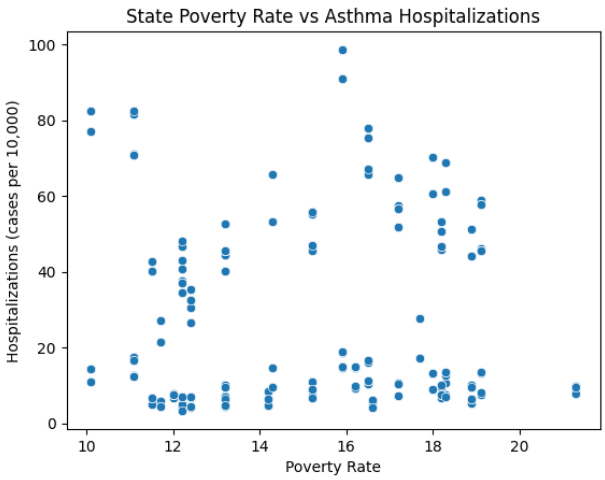
There is a dramatic increase in PM2.5 concentrations in the more northern latitudes and western longitudes in the United States from 2012 to 2013. There is also a slight increase in the concentrations in the eastern longitudes but the highest peaks in concentration at those longitudes have gotten shorter.

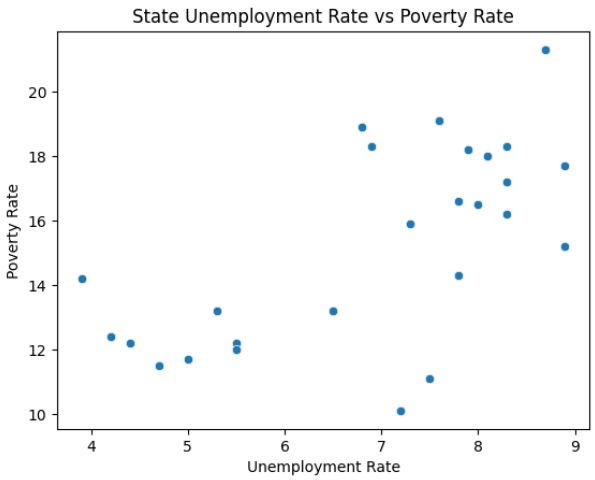


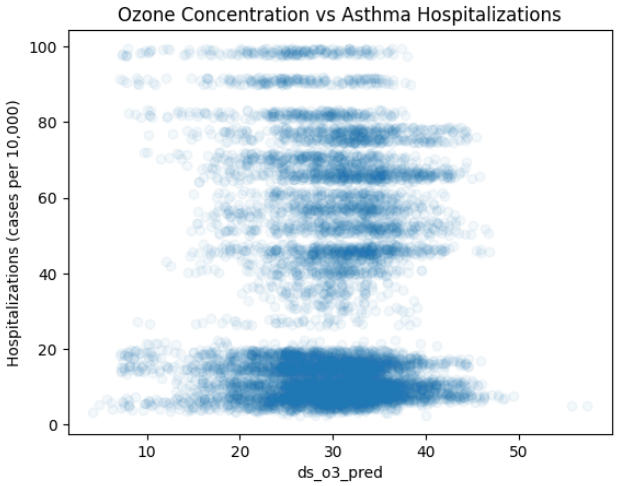
Many of the peaks in PM2.5 concentrations in 2013 seemingly disappeared in 2014. However, the remaining peaks that roughly correspond to Central California have gotten higher.

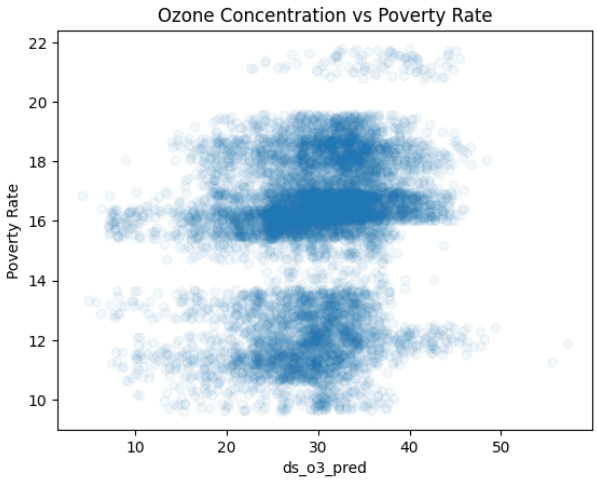
Throughout the years, there tends to be a much greater PM2.5 concentration along the West Coast of the United States. It is also surprising that there is not a single direction that the trend in concentration follows. Although the highest peaks have gotten higher, there have been dramatic fluctuations in PM2.5 concentrations throughout the years rather than a general increase over time. It may be worth investigating what is causing such drastic fluctuations and how the concentrations correlate with chronic disease rates.

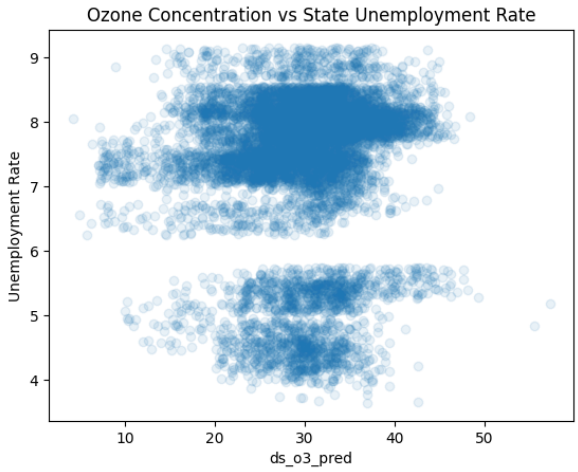












There appears to be at least a slight linear relationship between the confounding variables. However, to simplify the computation of the causal effect, we are going to assume that all variables have a linear relationship between each other. This is required for the unconfoundedness assumption to hold.

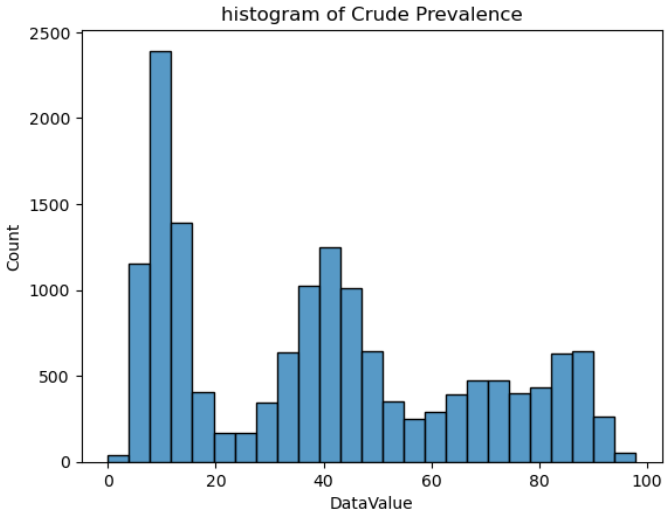
Research Question 2

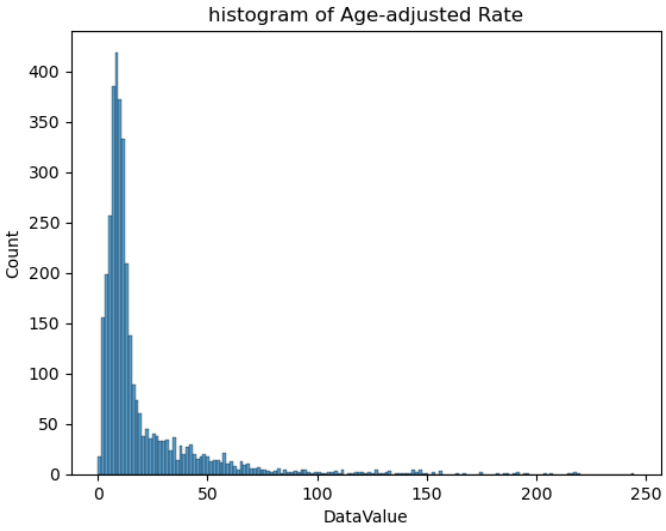
Data: CDC Infectious Diseases Filtered for Asthma

Quantitative Variable

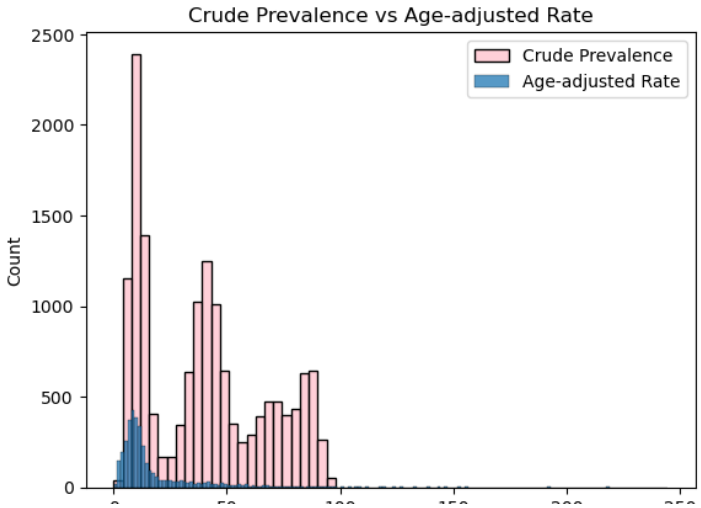
In order to have a better understanding of our Asthma dataset, we performed an EDA on the “DataValue” variable. The “DataValueType” variable is a categorical variable that indicates the type of value that is present in the "DataValue" column for a given row of data. The “DataValueType” column contained two different values: “Crude Prevalence” and “Age-Adjusted Rate”. Each of these values indicates a different way of calculating or interpreting the data in the "DataValue" column. We visualized their corresponding DataValues separately.

By analyzing the distribution of asthma prevalence rates using histograms, we can gain insights into the range of rates observed in the dataset, the typical range of values, and the presence of any outliers. This can help us identify potential patterns or trends in the data that may be explored further using geographic analysis

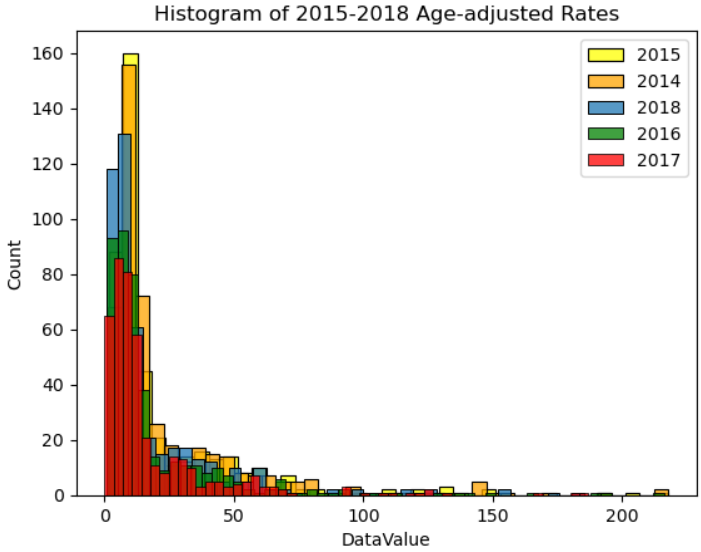




The average DataValue for the crude prevalence DataValueType is around 40. The average DataValue for the age-adjusted rate is around 20 with a skewed right tail.



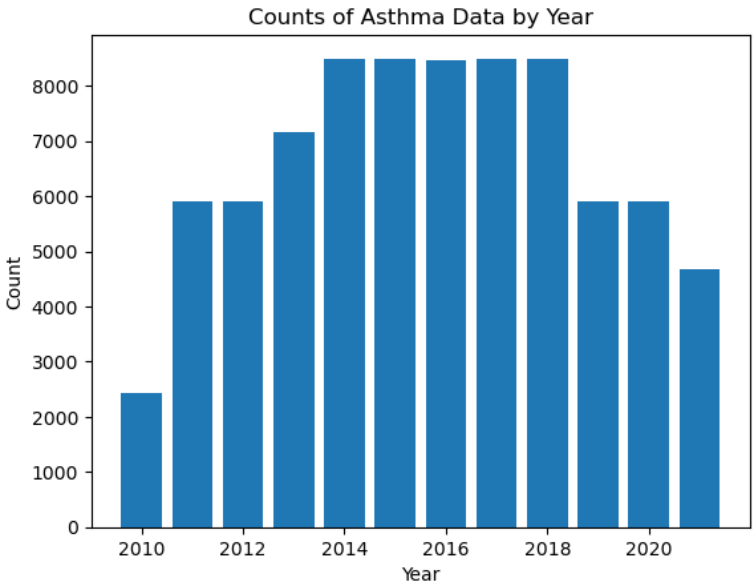
When comparing the crude prevalence DataValues to the age-adjusted rate, it is clear that both distributions are different. It is also noticeable that there are more DataValues for the age-adjusted rate as compared to the crude rate. We found that there were 9,639 DataValues for the age-adjusted rate and 27,500 DataValues for crude prevalence.



Looking closer, it is also clear that there are varying counts of DataValues based on the year we are observing.

Knowing the years with the most data can help us identify which time periods and locations have been the focus of previous research, and whether there have been any changes or trends in asthma prevalence over time.

Therefore, we also visualized the "YearStart" column to see which Years the dataset was most centered around. We found that it would be best to work with a subset of the data. Based on the bar chart, it is clear that years 2014 to 2018 should be in the subset because these years are shown in the dataset the most.

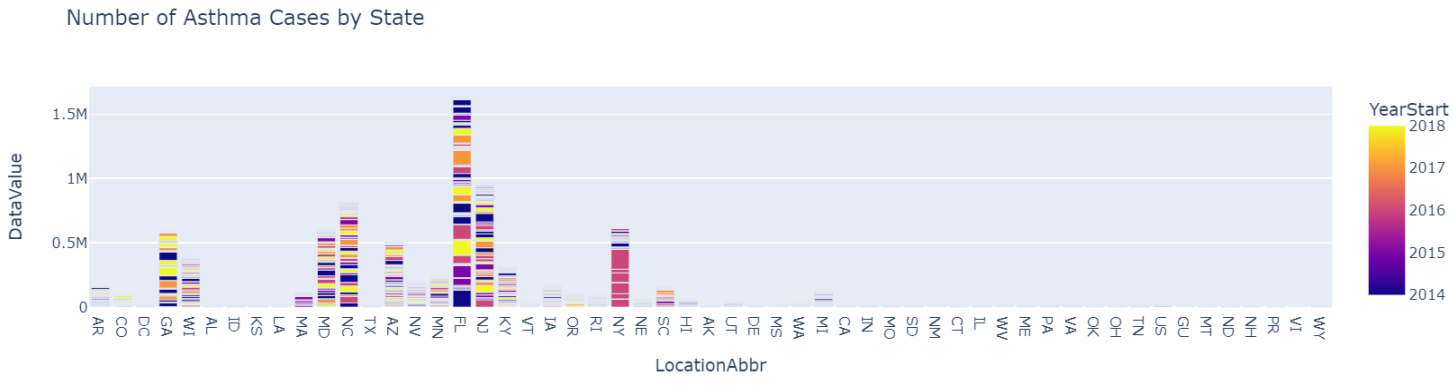


Qualitative Variable

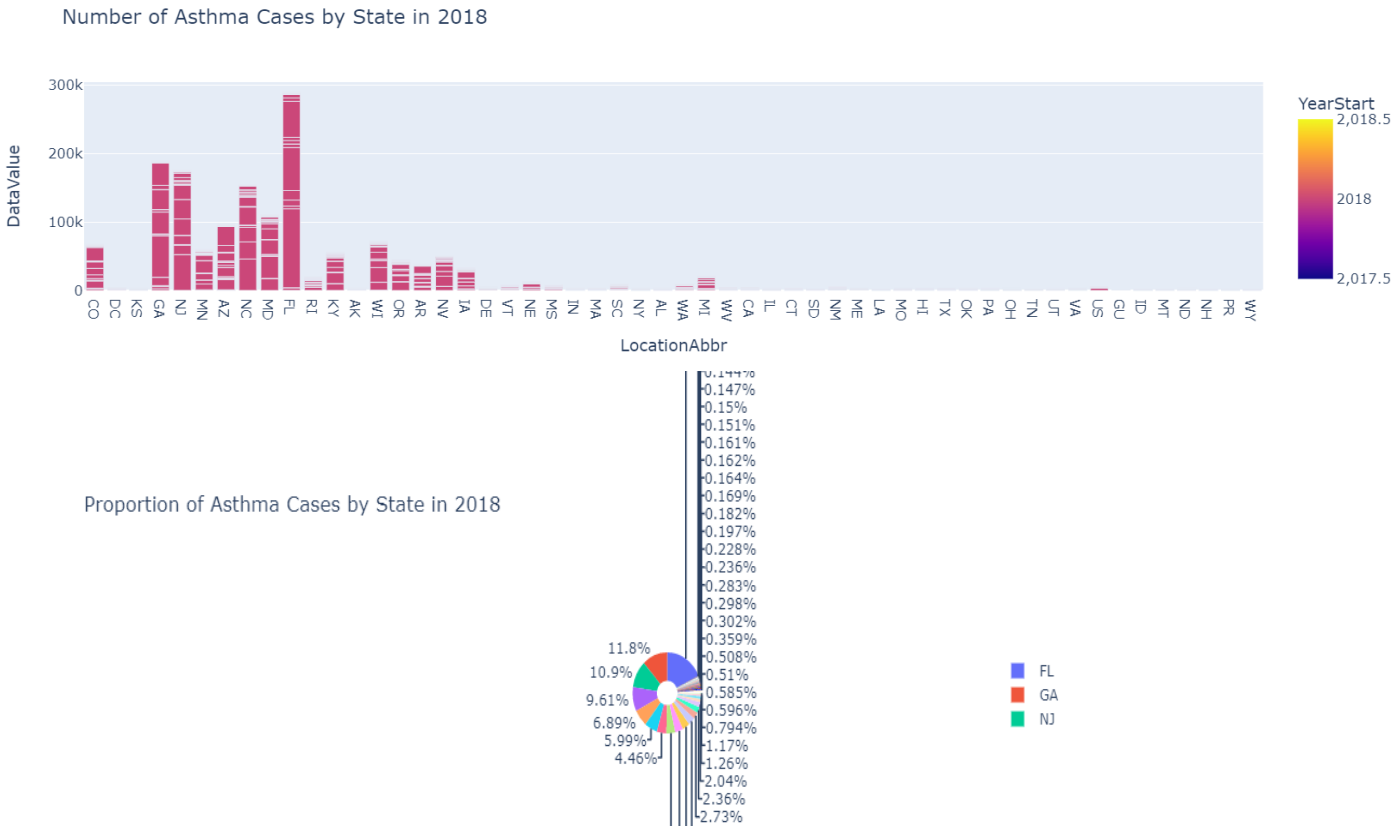
For our second research question, we are exploring whether geographical location can be linked with the prevalence of asthma cases. The following EDA was performed using the YearStart, LocationAbbr, and DataValue (number of asthma cases per 10,000 people) columns in the dataframe to explore our qualitative variable State.

The first step in this EDA is to map out the number of asthma cases by Year and State. We found that this heatmap gives us the best visualization. We can see that for some years there appears to be no data available for any state, and for some states we are missing values as well. This helps us narrow down the range of years we should focus on to have solid data available.

After filtering our dataframe to only include the years 2014 - 2018 (that seems to be where the bulk of the data is) we created a stacked bar chart for an alternative view of our data.



We also created bar charts and proportion donuts filtered by year to get a cleaner, closer look at the relationship between State and our other variables.



Finally, since we are looking at state codes we can map the asthma cases by year onto a map of the United States. As we attempt to model the way geographical location correlates with asthma cases, visualizing our data on the map will be an important tool to confirm our findings.

Research Question 1: Causal Inference

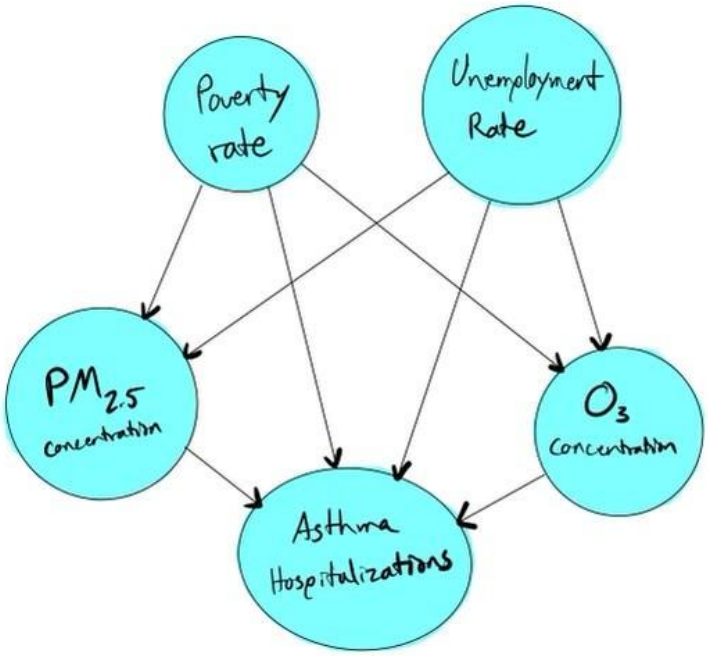
Methods

Our treatment variable is the concentrations of $PM_{2.5}$ and ozone ($\mu g/m^3$ and parts per billion). Our outcome variable is the rate of asthma hospitalizations (cases per 10,000).

Our confounding variables were unemployment rate and poverty rate. Unemployment rate is a confounding variable because people who are unemployed are less likely to have the means to treat their asthma and are possibly more likely to be outside where they are exposed to pollutants Poverty rate is a confounding variable because people who live in poverty are less likely to receive quality medical care and more likely to live in areas that have higher concentrations of pollutants

From our work done in the EDA, we are assuming that the relationship between the confounding variables is linear. Under this assumption, the unconfoundedness property also holds. We are using outcome regression to adjust for the confounding variables. It is not likely that there are colliders in the dataset.

Causal DAG:



Results

PM2.5 OLS results:

OLS Regression Results

Dep. Variable:	DataValue	**R-squared:**	0.027
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	6.771e+04
Date:	Mon, 08 May 2023	**Prob (F-statistic):**	0.00
Time:	13:29:30	**Log-Likelihood:**	-3.4963e+07
No. Observations:	7367532	**AIC:**	6.993e+07
Df Residuals:	7367528	**BIC:**	6.993e+07
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	19.8358	0.085	234.540	0.000	19.670	20.002
ds_pm_pred	0.4595	0.003	-161.230	0.000	-0.465	-0.454
unemployment_rate	-0.8531	0.005	-170.200	0.000	-0.863	-0.843
poverty_rate	3.5543	0.009	415.083	0.000	3.538	3.571

Omnibus:	1054616.877	**Durbin-Watson:**	0.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	900848.433
Skew:	0.772	**Prob(JB):**	0.00
Kurtosis:	2.259	**Cond. No.**	160.

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

Based on the OLS results, it appears that there is a statistically significant relationship between asthma cases and PM2.5 concentrations. The coefficient for PM2.5 (ds_pm_pred) is negative (-0.4595) and statistically significant (p-value < 0.001), which suggests that higher PM2.5 concentrations are associated with lower asthma cases after controlling for the effects of unemployment rate and poverty rate.

Furthermore, the coefficients for unemployment rate and poverty rate are both statistically significant (p-values < 0.001) and negative. This suggests that higher unemployment rate and poverty rate are associated with lower asthma cases after controlling for the effects of PM2.5 concentrations. The coefficient for poverty rate is positive (3.5543), indicating that higher poverty rate is associated with higher asthma cases after controlling for the other variables in the mode

In terms of effect size, the coefficient for PM2.5 indicates that for every one unit increase in PM2.5 concentration, there is a decrease of 0.4595 in asthma cases, after controlling for the effects of unemployment rate and poverty rate. Similarly, the coefficients for unemployment rate and poverty rate indicate that for every one unit increase in these variables, there is a decrease of 0.8531 and an increase of 3.5543 in asthma cases, respectively, after controlling for the other variables in the model.

While the statistical significance of the estimates are strong, it is important to note that the magnitude of the effects are relatively small. Further research is necessary to confirm these findings and to investigate potential mechanisms underlying the observed associations. Also, both poverty rate and unemployment had stronger causal effects on hospitalizations than PM2.5 concentration. Therefore, further research would be needed to confirm our findings.

Ozone OLS Results:

OLS Regression Results

Dep. Variable:	DataValue	**R-squared:**	0.041
Model:	OLS	**Adj. R-squared:**	0.041
Method:	Least Squares	**F-statistic:**	3.982e+04
Date:	Mon, 08 May 2023	**Prob (F-statistic):**	0.00
Time:	15:14:12	**Log-Likelihood:**	-1.3378e+07
No. Observations:	2825736	**AIC:**	2.676e+07
Df Residuals:	2825732	**BIC:**	2.676e+07
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.3335	0.133	100.627	0.000	13.074	13.593
ds_o3_pred	0.0222	0.003	8.017	0.000	0.017	0.028
Poverty Rate	-1.6838	0.008	-205.603	0.000	-1.700	-1.668
Unemployment Rate	6.0730	0.018	345.250	0.000	6.039	6.107

Omnibus:	537286.841	**Durbin-Watson:**	0.000
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	310155.065
Skew:	0.681	**Prob(JB):**	0.00
Kurtosis:	2.118	**Cond. No.**	284.

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

There appears to be at most a very slight causal relationship between ozone concentration and rate of hospitalizations and emergency room visits for asthma. The results show that the regression coefficient of 0.0222 for ozone concentration has a low p-value which indicates that it is a statistically significant result. This means that for every increase in ozone concentration, the hospitalization rate can be expected to increase by about 0.02.

The regression coefficient of -1.6838 for poverty rate is also statistically significant and suggests that for every unit increase in poverty rate, hospitalization rate decreases by about 1.68 when adjusting for the other variables.

Unemployment rate has a statistically significant regression coefficient of 6.0730 that indicates that a unit increase in unemployment rate causes asthma hospitalization rate to increase by about 6.07. This confounding variable appears to have the greatest causal effect on hospitalizations according to our model.

Although the model suggests that there might be a causal relationship between ozone concentration and asthma hospitalizations, both poverty rate and unemployment had stronger causal effects on hospitalizations than ozone concentration.

Discussions

One limitation of this procedure is that it relies on a lot of assumptions. If any of these assumptions do not necessarily hold, then the results will likely not be accurate. We also had to find data that fit our datasets perfectly, or else our OLS results would have been inaccurate. If the additional data does not map perfectly to our ozone dataset, pm2.5 dataset, and asthma dataset, our results would be inaccurate. Hence, this method is prone to include easy errors if one does not work with caution.

Additional data would be useful for answering this causal question. We lack the domain knowledge to think of any other confounding variables. Therefore, if we had the ability to find other confounders and had data on those variables, we would be in a better position to answer this causal question. This is because including that data will make the regression model more representative of reality.

We are not particularly confident that there is a causal relationship between ozone concentration/PM2.5 concentrations and asthma hospitalizations. Although there is a statistically significant positive relationship between PM2.5/ozone concentrations and asthma cases, causality cannot be established solely based on these results. This is because we have evidence to suggest that there are other factors with a greater causal effect on hospitalizations (such as unemployment rates and poverty rates).

Research Question 2: Prediction with GLMs and Nonparametric Methods

Methods

We are trying to predict the “Current asthma prevalence among adults aged >= 18 years” numerical variable. The features we used to predict this variable are “YearStart”, which contains the start year of the data collection by row, and the geographical regions (Midwest, Northeast, South, West) based on the provided State Code. Each individual state of the original data set was mapped into one of those geographical regions and one hot encoded. We thus have an additional 4 columns for each of those regions that were used as features in addition to year start.

Additional filtering on the dataset includes separating the dataset by the “StratificationCategory1” which contains categories such as “Overall”, “Gender”, and “Race / Ethnicity”. We perform parametric and nonparametric modeling on the “Overall”, “Male”, and “Female” (with “Male” and “Female” being subsets of “Gender”) subsets to ensure that we capture more than one view of the data in our research while keeping our samples consistent

We will be using a Frequentist GLM. We chose the Poisson distribution with a log link because our dataset includes the Year feature. Based on our EDA, we want to use the subset of data from 2012 - 2019 which translates well to a Poisson distribution given our time element. For the Bayesian model, we explored our data and determined that the values we want to model follow a rough Normal distribution. Our Region feature is one-hot encoded which is represented by a Beta prior.

We will be using a Random forest as our nonparametric method. Being this is a nonparametric method, we aren’t making that many assumptions and the forest itself isn’t very interpretable. However, the decision to go with a random forest was mainly due to its greater performance accuracy in comparison

to a decision tree. Moreover, KNN was not used as it was unclear what a good X1 and X2 would be in addition to likely overlaps between the regions.

We will evaluate each model's performance in different ways. For the Frequentist GLM, we split the data into training and test sets to determine how well our model can predict the data. We obtained a RMSE so it can be compared with the nonparametric method and also observed the confidence intervals in the model to quantify uncertainty. For the Bayesian model, we are including credible intervals to quantify uncertainty as well as visualizing the alpha values we obtain. For Random Forest, we split the data into training and test sets. We used RMSE to evaluate the model and compare it to the frequentist GLM. We also calculated prediction intervals to quantify uncertainty. We are comparing all three methods to evaluate which model had the best performance per our data and also to determine our conclusion.

Results

Frequentist GLM:

Below are the coefficients we obtained for the Overall, Male, and Female datasets respectively. Based on these results, it would appear that the Northeast region has the strongest correlation with asthma prevalence among the four regions. Across the three subsets of data, the four regions performed roughly the same with the Northeast and the West having the strongest coefficients and the South and the Midwest having the weakest.

Overall:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-15.1347	18.324	-0.826	0.409	-51.049	20.780
YearStart	0.0086	0.009	0.947	0.344	-0.009	0.026
Midwest	-0.0057	0.065	-0.088	0.930	-0.133	0.121
Northeast	0.1552	0.067	2.318	0.020	0.024	0.286
South	-0.0133	0.064	-0.208	0.835	-0.138	0.112
West	0.0056	0.065	0.085	0.932	-0.122	0.134

Male:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-9.0294	21.001	-0.430	0.667	-50.191	32.133
YearStart	0.0054	0.010	0.519	0.604	-0.015	0.026
Midwest	0.0273	0.075	0.367	0.714	-0.119	0.174
Northeast	0.1681	0.075	2.229	0.026	0.020	0.316
South	0.0023	0.073	0.032	0.975	-0.141	0.146
West	0.0620	0.074	0.842	0.400	-0.082	0.206

Female:

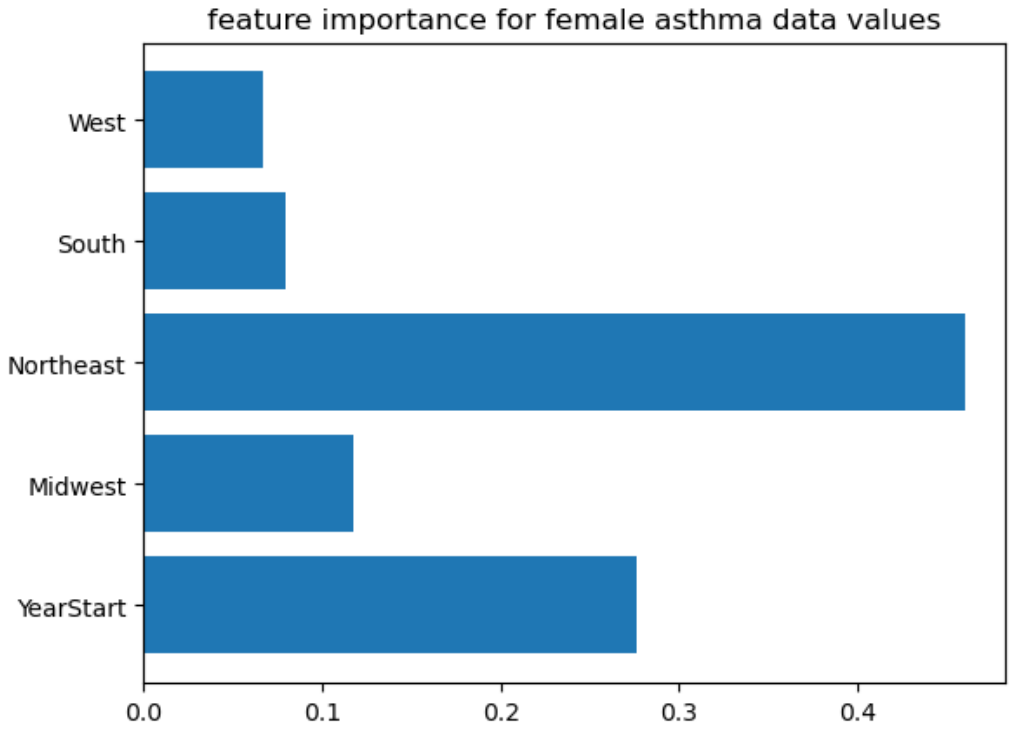
	coef	std err	z	P>|z|	[0.025	0.975]
const	-15.8363	16.195	-0.978	0.328	-47.578	15.906
YearStart	0.0091	0.008	1.129	0.259	-0.007	0.025
Midwest	-0.0152	0.058	-0.261	0.794	-0.129	0.099
Northeast	0.1519	0.060	2.543	0.011	0.035	0.269
South	0.0031	0.057	0.055	0.957	-0.109	0.115
West	0.0091	0.058	0.155	0.877	-0.106	0.124

Bayesian:

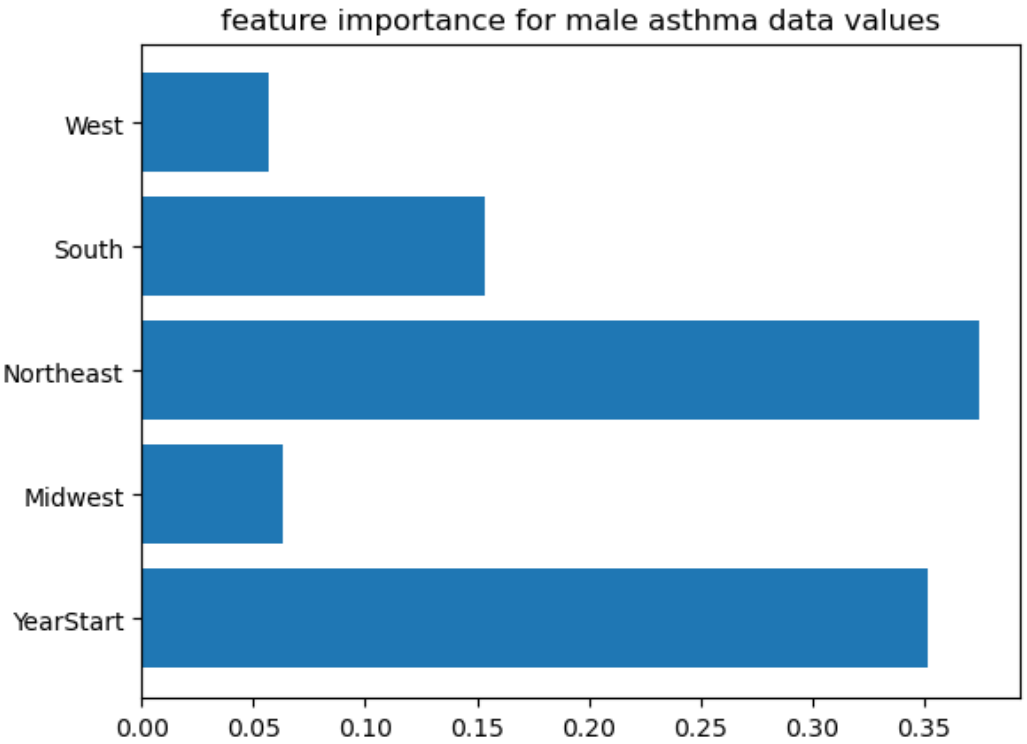
Consistent with the Frequentist model, the Bayesian model summary also shows much stronger correlation in the Northeast and the West. In the model summaries, the means for Northeast are consistently higher than the other regions as it hovers around 0.9 while the other regions are closer to 0.03

Random Forest:

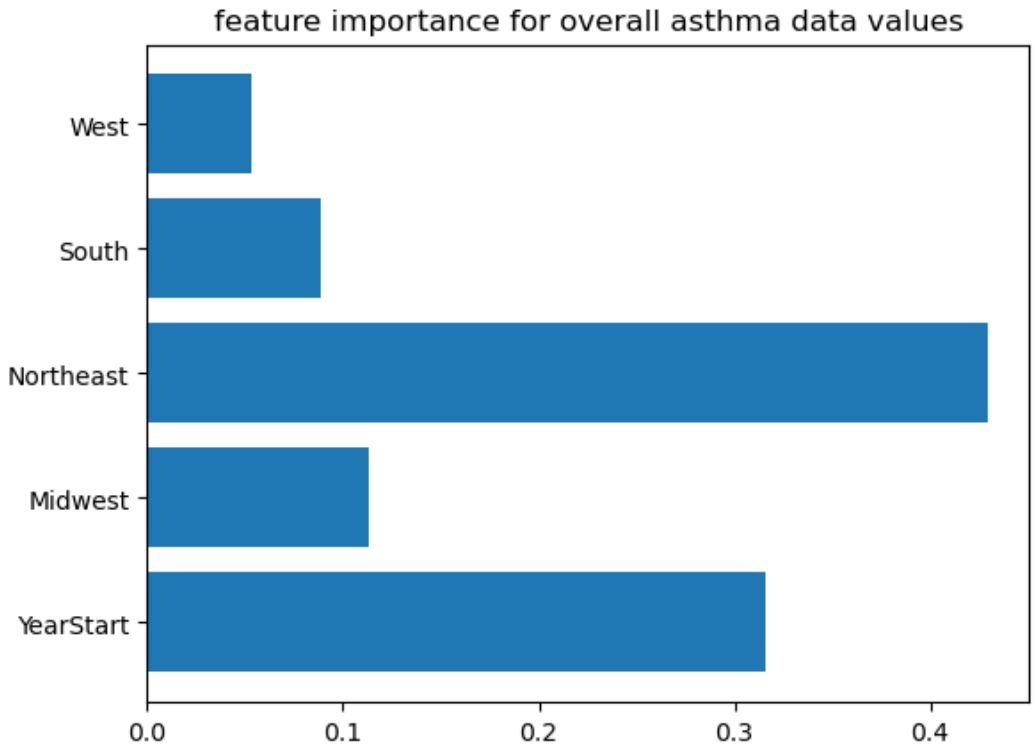
Female:



Male:



Overall:



- RMSE:
    - Overall Test: 1.304
    - Male Test: 1.125
    - Female Test: 1.786

Looking at the feature importance bar graph, the Northeast region and the Yearstart seem to have the most influence on all three stratification categories. Yearstart seems to have a greater effect on data values stratified for males than females and overall. Northeast however appears to be most important in all 3 random forest models.

We can interpret this as the number of asthma cases in the Northeast is higher in comparison to the other regions. This conclusion is consistent with the coefficients obtained from the Frequentist GLM model.

We quantify uncertainty in the Frequentist Model with confidence intervals and RMSE values. The confidence intervals tell us how confident we are in the coefficients we obtain and it helps us compare across the coefficients. For example, for the Overall subset the Northeast confidence interval is the only one with a positive lower and upper bound. This allows us to confirm that the Northeast does indeed have the largest coefficient despite some uncertainty in our model. The RMSE values for the Overall, Male, and Female, models are as follows: 9.39, 6.55, 11.98 These values are to quantify uncertainty in the Bayesian model for comparison with the nonparametric model. This will allow us to give different weights to our models in the conclusion

Similar to the Frequentist model, we are including the HDI 3% and the HDI 97% generated in the Bayesian model summary to confirm that the correlations we see between geographic location and asthma

prevalence are representative of the true relationship. Across the three subsets, regions such as the South and Midwest have a very low lower bound around 0 while the Northeast consistently has a lower bound around 0.9 - this is a very large difference indicating that the Northeast has a much stronger correlation with asthma prevalence than the other regions.

In the Random Forest model, we also quantified uncertainty by creating a bootstrapped prediction interval for every row in the test set of the overall stratification category. It was generated using 100 bootstrapped samples of the training data, each trained to a random forest regressor model and generated predictions of the test set. The 2.5th and the 97.5th percentiles were then calculated to obtain the prediction interval and inserted into the test set table. Other mainstream random forest uncertainty quantifications are, according to professor Ramesh, beyond the scope of this course.

Discussions

The Frequentist model had a relatively high RMSE when compared to the nonparametric methods (9.39, 6.55, 11.98 vs 1.304, 1.125, 1.78)

Across the three subsets that we used, it appears that the Male subset also performed better than the other two subsets in both models. It is possible that this has to do with how well-represented each state is within the subsets, despite the subsets having the same number of rows. This would affect the way that the training and testing datasets got split and how close they compare. As we know, Frequentist modeling relies solely on the observed data. In this situation where we may not have equal representation across the regions, the Frequentist way can fall short in providing a representative and conclusive model.

The Bayesian model revealed lower bounds that were often at or close to 0.0%. This again brings up the question of representation of each region. However, overall we observed the same result as in the Frequentist model where the Northeast had a consistent higher correlation with asthma prevalence.

The random forest model performed better than both the Frequentist and Bayesian GLM. This is primarily because the random forest model can handle messier data and relationships that aren't linear in nature. The random forests capture interactions between features whereas GLMs assume that the effects of features are independent of each other. Moreover, the random forest model could also have performed better because it's less prone to the influence of outliers in comparison to GLM.

It's difficult to fully interpret the results from our random forest models because they're an ensemble of decision trees, which can make it difficult to understand how each individual tree is contributing to the overall prediction. However, the feature importance graphs above demonstrate that the most important features that contribute to the model's predictions are the Northeast region and Yearstart of the data.

One limitation of our research in general is that we had to choose between several filters in the beginning. We ultimately chose to explore "current asthma prevalence among adults aged >= 18", but the current state of things may not reflect the historical prevalence of asthma in the United States. We also had to omit a number of years from the lower and upper bound in our dataset because the EDA revealed that there was not any data available. The range of years we focused on was from 2012 - 2019; this subset may reveal a certain outcome, but we cannot say for sure that our models would adapt well to data from 2019 and on or 2012 and before. Finally, it would have been helpful to have more detailed information on the participants in the dataset. We got general categorizations such as "gender", "age above 18", and "race / ethnicity", but details about medical history, pre-existing conditions, and lifestyle would have helped us be more thorough and confident in our findings.

Conclusions

For research question 1, our results show that there is a statistically significant relationship between asthma hospitalizations and PM2.5 concentration and ozone concentrations, as well as poverty rate and unemployment rate. The effect size of PM2.5 concentration on asthma hospitalizations was relatively small, while poverty rate and unemployment rate had stronger causal effects on hospitalizations. Additionally, while there appeared to be a slight causal relationship between ozone concentration and asthma hospitalizations, the effects were weaker compared to poverty rate and unemployment rate. Assuming the assumptions hold, then we have demonstrated that there is at least a slight causal effect on asthma hospitalizations caused by ozone concentration and PM2.5 concentration.

Our findings for research question 1 are limited to the United States and the years 2011 to 2014 due to the lack of data. However, the relationships between air pollution and health outcomes, as well as the impact of social determinants of health, are widely recognized and may be applicable to other regions. Therefore, our findings are narrow but they can be generalizable.

Based on your results, suggest a call to action. What interventions, policies, real-world decisions, or action should be taken in light of your findings?

Our findings for research question 1 suggest that reducing air pollution and addressing social determinants of health, such as poverty and unemployment, may help reduce asthma hospitalizations. Policy interventions, such as regulations on industrial emissions and investments in public transportation, may help reduce air pollution, while social policies that address poverty and unemployment may help improve access to healthcare and reduce exposure to pollution.Because our results indicated that there is at best a weak causal relationship between ozone concentration/pm2.5 concentration and asthma hospitalization, there might not be much of a demand for policies to keep ozone concentration and pm2.5 concentration under control for the purposes of tackling asthma and other public health issues.

We merged different data sources to obtain information about unemployment rates and poverty rates. The benefit of combining different sources is that it allowed us to access a more comprehensive set of information than we would have had otherwise. By incorporating data from the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program and the U.S. Census Bureau's American Community Survey (ACS) 5-Year Estimates, we were able to obtain more detailed information about poverty and unemployment rates at the State level, which enabled us to control for the effects of these variables more effectively in our analysis. The CDC indicators dataset in particular was very messy. There were several columns with missing data and many different categories that lead to a pretty convoluted filtering process.

One limitation is that we lack the domain knowledge needed to know what other variables could work as confounders. Thus, there might be some confounding variables present in the data that we did not properly take into account.

Future studies could further explore the causal mechanisms underlying the relationships between air pollution, social determinants of health, and asthma hospitalizations. Additionally, studies that use more comprehensive and accurate data on air pollution and health outcomes may help further elucidate these relationships.

Our main key finding from research question 2is that the Northeast region of the United States appears to have a much stronger correlation with asthma prevalence than the other four regions. Not only did the Northeast demonstrate the strongest correlation with asthma across the models, but the order of

correlation also remained consistent. The region with the second-highest coefficients was the West, followed by South and the Midwest.

Our findings are considerably narrow considering that the CDC dataset was not a random sample but a targeted collection of asthma-related individuals. In our filtering, we focused on current asthma cases ages greater than or equal to 18, and we further divided this population into Overall, Male, and Female categories. Despite the three separate models, our results across the populations are consistent with one another, so we can say that for asthma cases occurring from 2012 to 2019, ages 18 and above, our results are applicable.

Real-world decisions include healthcare policies. If people in the Northeast have a higher chance of being diagnosed with asthma, health officials need to prepare the proper resources and support for public health.

For research question 2, we did not merge different sources in this scenario because our focus was on large geographic regions and the general prevalence of asthma. However, our findings do lead to next steps that require additional sources relating to more specific features of the Northeast region.

We were limited by the fact that our data source had multiple different categories. This meant that one row was not necessarily comparable to the other and that in filtering our dataset we omitted a considerable amount of samples. In addition, the target variable in our dataset did not show individuals but rather groups of individuals. This means we could not examine individuals and their specific symptoms or geographic area within the region. If there is a particular state that holds the majority of asthma cases in the region, our model would not reflect that.

Now that we have discovered a strong correlation between the Northeast and asthma prevalence, the next steps would be investigating what qualities of the Northeast may be causing this relationship. Perhaps it is due to the air quality, the industries, or the geography. For researchers hoping to find better treatments for asthma, our work can also be a good reference for where to find the desired subjects.

We learned many lessons during this project. While working on research question 1, we learned the importance of properly preparing data. This project involved significant data preparation, such as cleaning and merging datasets from multiple sources. It was important for us to carefully document all of the steps involved in this process to stay on track of our work. We also learned that it is important to carefully consider which variables to include in the model, especially confounders or colliders that may need to be adjusted for. Without choosing our variables properly, our results looked inaccurate. After properly considering confounding variables, our results drastically improved. While working on research question 2, we learned how important it is to take the granularity of the data into account before feeding it into a model. We also learned that it is important to model data with several different methods and models before making a final conclusion. We also investigated several different ways to quantify the uncertainties and errors of our models in order to determine the best fit for our data.