Christian G, Sarah T, Chinmay C, David C
Data 102: Data, Inference, and Decisions
Final Project Report

Data Overview:

Name	Abbrv.	Description + Purpose
CDC Daily PM2.5 Concentrations	PM2.5	Tracks daily PM2.5 air concentrations from 2011-2014 for almost all cities within the USA. Utilized as a covariate in prediction and X variable in causal inference.
ProPublica Cancer-Causing Industrial Air Pollution	AIRP	Tracks the locations of pollution factories around the USA; pollution is measured on a grid system via incremental cancer risk. Utilized in creating the IV for the causal inference.
CDC 500 Cities - Adult COPD Rates	COPD	Tracks the COPD rate, as proportion of the population, for the 500 most populous cities – as of 2019 – in the USA. Utilized as one of the target variables Y in both causal inference and regression.
CDC 500 Cities - Adult Asthma Rates	ASTH	Tracks the asthma rate, as proportion of the population, for the 500 most populous cities – as of 2019 – in the USA. Utilized as one of the target variables Y in both causal inference and regression.
US Census - Educational Attainment	EDUC	Tracks the educational attainment and earnings by city – for almost all cities in the USA. Utilized as a covariate in prediction, and as a check for possible correlation with the selected IV.
US Census - Employment Status	EMPL	Tracks the employment and unemployment figures, by city, for almost all cities in the USA. Utilized as a covariate in prediction, and as a check for possible correlation with the selected IV.
US Census - Demographic Data	DEMO	Tracks demographic data such as race, age, household size, and gender for almost all cities within the USA. Utilized as a covariate in prediction, and as a check for possible correlation with the selected IV.

The goal of this project is to (1) test if there is a causal relationship between city-wide pollution and chronic respiratory diseases – namely Asthma and COPD – and (2) predict these chronic respiratory disease rates from informative covariates. Throughout our project, we accessed multiple datasets regarding pollution and diseases that are recorded throughout the United States. More specifically we focused on and analyzed relationships found between air quality and respiratory diseases. To understand pollution and air quality in the various regions across the US we utilized the Daily PM2.5 Concentrations dataset included in the project guidelines. We used 2 datasets, specifically the CDC's 500 cities community survey, that contained information about COPD and asthma. These come from the U.S. Chronic Disease Indicators dataset which has been getting updated every year since 2016. The target variables – contained in the COPD and ASTH datasets – show city-wide average rates of the diseases as a percent of the total population. This dataset is especially valuable because it does not just find

Disease data on the city level it also goes further to collect data on a smaller scale according to census tracts. This in particular is valuable because many cities may be large and due to regional geographic patterns certain areas may be affected more by pollution or other factors. This allows for the potential further analysis of cities on a smaller scale. Also, when looking towards causal inference, this team tested "distance to closest air-polluting factory" as an Instrumental Variable (IV). The AIRP dataset, which details locations of air-polluting factories – among other data – was valuable in calculating these distances. Along with these datasets to add another angle of analysis to our project we used the ACS 5-Year Estimates from the US Census Bureau to understand demographic information.To ensure the selected IV was independent of any confounders related to COPD and Asthma rates, linear and nonlinear relationships were tested between the IV and various demographic and socioeconomic factors. The demographic and socioeconomic factors – contained in the EDUC, EMPL, and DEMO datasets – are the most recent 2022 US Census Bureau ACS community survey datasets.

Research Questions:

The first research question tackled the prediction of Asthma and COPD rates from city-wide demographic, environmental, socioeconomic, and pollution data. Random forests and Gradient-boosted trees are a non-parametric method used to classify cities into high or low-incidence groups; an advantage of this method is the interpretability of feature importance. Logistic regression was used as a method to predict the probability of a city belonging to either low or high-incidence groups.

Our next research question is does the proximity to power plants – a source of pollution – have a significant effect on the rates of diseases such as COPD and asthma. We modeled this by testing various relationships between the proximity to power plants and the disease rates of COPD/asthma. This second research question addressed casual quantification of the effect of pollution on Asthma and COPD rates; this was done on a city granularity. Initially, the IV – a city's distance to the closest pollution-causing factory – was tested against many socioeconomic and demographic factors to establish any correlation. The only covariate that saw any trend was the proportion of the population that was African American; this was corrected for in the 2SLS regression.

EDA:

There were two primary sections of EDA: the first isolated possible confounders for the selected IV, and the second explored the ASTH and COPD datasets more generally. Both sections required extensive cleaning and manipulation to format the data in a way that could be analyzed, and used for modeling.

The ASTH, COPD, EDUC, EMPL, and DEMO were all joined on their city and state combinations. Several issues and their solutions are detailed below:

Total population between the tables differed significantly. For example, the EDUC table had a total population by city so it could quantify educational achievement via proportions and the EMPL table has total population by city to quantify employment as a proportion. These populations routinely different between the same cities. As a result, demographic and socioeconomic data points were all recorded as raw proportions – not per 100,000 – related to populations in their corresponding tables.

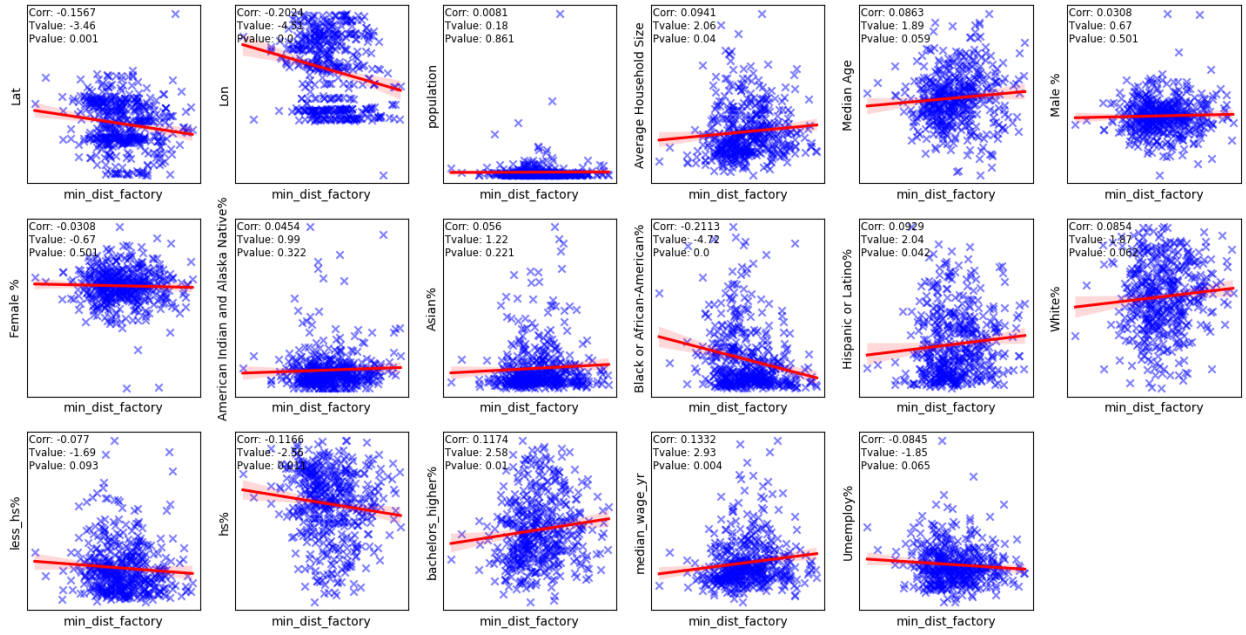
For many of the city and state combinations, there were duplicates that contained different population numbers, latitude, and longitude coordinates. Some of these duplicates were due to multiple different census datasets being combined in the raw data; as a solution, we filtered exclusively for the ACS yearly data. For duplicates that occurred from the same census sampling, an average population, latitude, and longitude were used.

To calculate the values for the IV, Haversine distance was used to rank each city against all factories, and all but the closest factory was dropped. Our IV exclusively looked at a factory's existence and not the magnitude of the pollution by said factory; accounting for factory pollution amount could improve accuracy in a future analysis.

Unfortunately, the PM2.5 pollution dataset was evaluated on a county level versus the city-level granularity features in the majority of the other datasets. As a result, cities were joined with the corresponding county using an outside dataset and then joined to the PM2.5 dataset. This may cause some cities to have the same PM2.5 data.

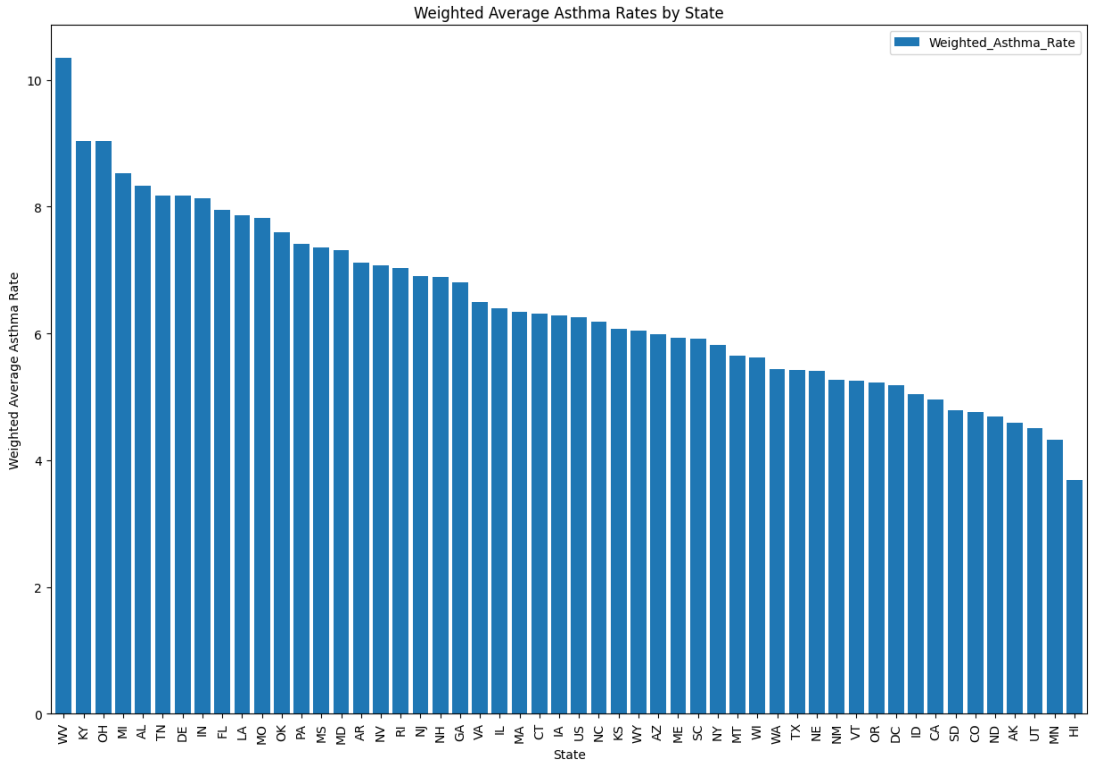
As a result of all the joining and merging of many different datasets, the resulting number of cities decreased from 500 to 487. These cities do not constitute a random sample but are rather the 500 most populous cities in the United States. This non-random sampling is justified as not all cities will have the infrastructure or population to support collection of the robust data needed for this analysis.

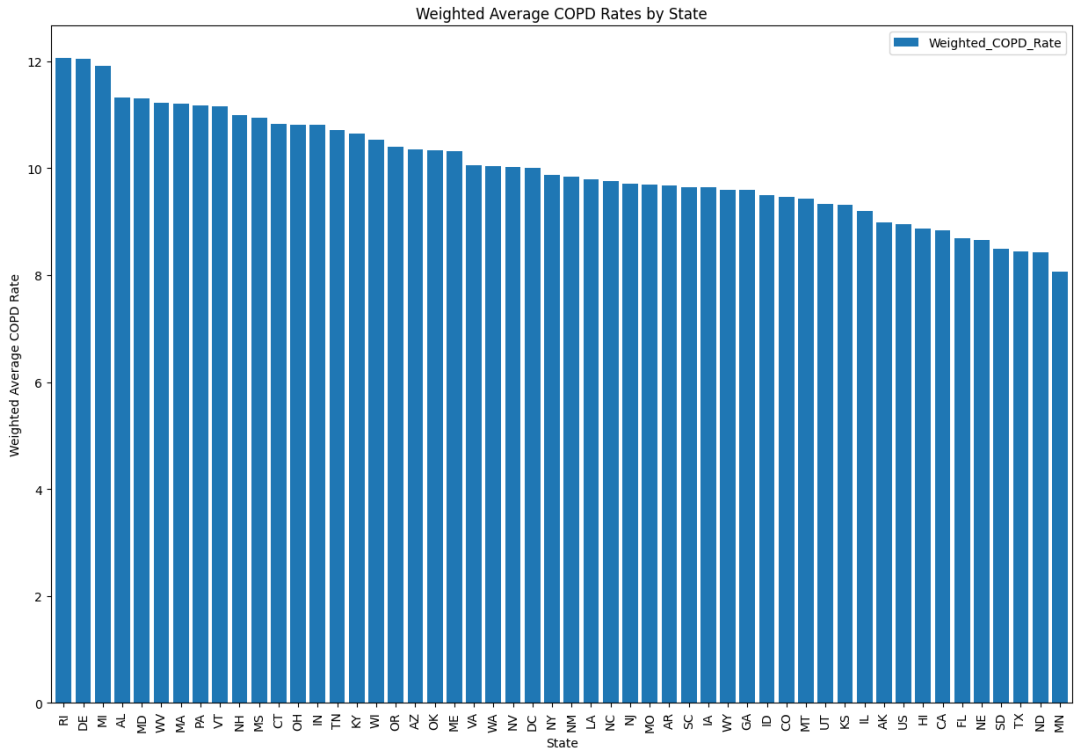
Concerning the relationship between the IV and possible confounding variables, through testing multiple linear and non-linear relationships, the only reasonable cofounder, at a regression coefficient of -0.21 is the proportion of African Americans within a city. The independent variable, the distance of a city to the closest factory in kilometers, was log-transformed due to its right-skewed distribution.



As a result, proportion of African Americans was added during the 2-Stage Least Squares Regression (2SLS) during the causal inference modeling stage.

Also as for exploring the ASTH and COPD data sets, we decided to create graphics to help us understand the data.

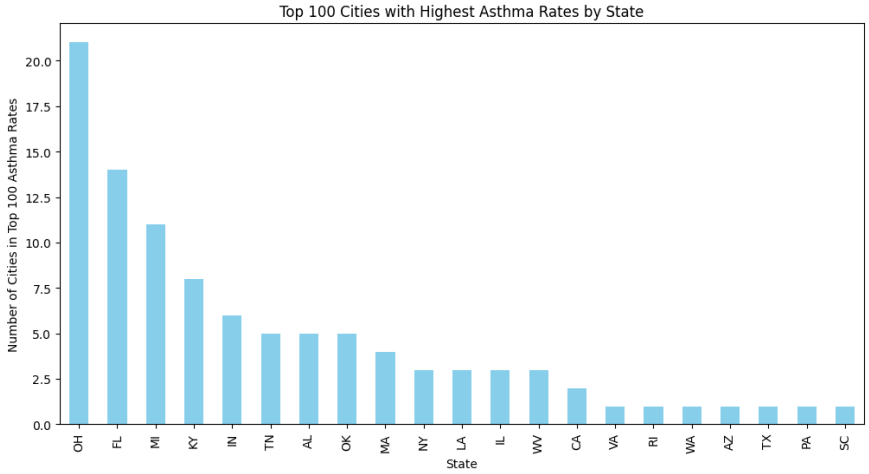


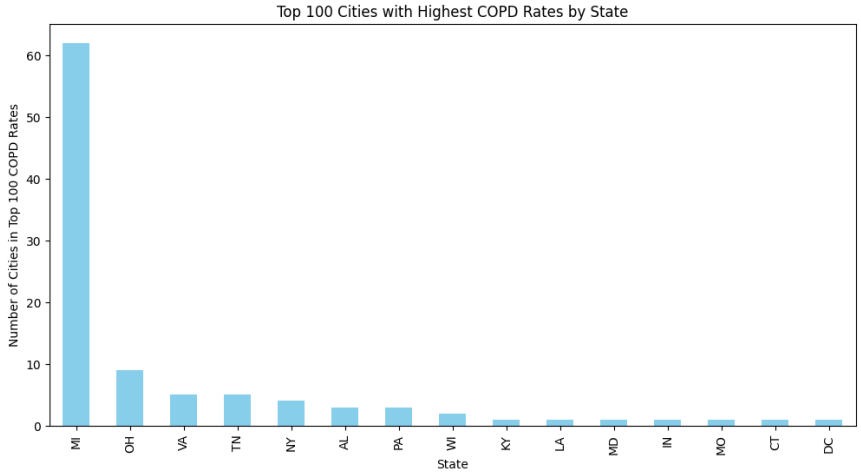


The first graph above shows the Asthma rates per State. When we calculate these rates there are huge differences in the variety of rates produced by this sample of cities. Now that we have established that there is variation in the asthma rates we can start looking into what is different between that states to try to identify factors that may be causing these results. It is clear that there is a difference in either living conditions or choices(or some sort of factor causing this variation) between the state at the top of the list such as West Virginia and the state at the bottom of the list Hawaii. Now that we have established what states have the worst Asthma rates we can target these states and see what general patterns they have that may be causing asthma in the population. We can analyze various factors including distance from power plants to see if pollution may be the cause of these asthma rates.

The second graph above shows the COPD rates per State calculated using the city sample found in our dataset. When we calculate these rates out we see that there are huge differences in the variety of rates produced by this sample of cities. Now that we have established that there is variation in the COPD rates we can start looking into what is different between that states to try to identify factors that may be causing these results. It is clear that there is a difference in either living conditions or choices(or some sort of factor causing this variation) between the state at the top of the list such as Rhode Island and the state at the bottom of the list Minnesota. Now that we have established what states have the worst COPD rates we can target these states and see what general patterns they have that may be causing COPD in the population. We can analyze various

factors including distance from power plants to see if pollution may be the cause of these COPD rates. On top of all of this, it seems that there does not seem to be a pattern between COPD and asthma rates. This is just from a visual assessment of these graphics generated above but it is very much clear that there are different states that have high rates of either condition. This is something we will investigate further because that may imply that there are different factors that are causing these respiratory conditions. And we will of course be tying it back and analyzing the relationship to proximity to power plants.





The first graph above depicts the amount of the top 100 cities with the highest Asthma rates that are in each state. The table below shows the 100 worst cities in order along with their asthma rates. When it comes to asthma it is clear that there is some factor that is causing cities in Ohio and Florida to be overly represented in the top 100 cities with the highest asthma rates. It is interesting that considering this we can see that New York City is the single city with the highest asthma rate but relatively few other cities in New York have made it into the top 100. It is clear that certain city related factors are causing asthma in people of the region. This seems like an interesting point that we will be getting into because we are analyzing the effects of proximity of power plants on the asthma rates but at the same time you would think that power plant would

not be close enough to a major city like New york to cause major health effects. Also as can be seen in the top 100 cities list there are repeating cities. This is because they are data measurements from subsets of the city which means that if a city is repeated a lot then many more census regions are affected by elevated Asthma rates.

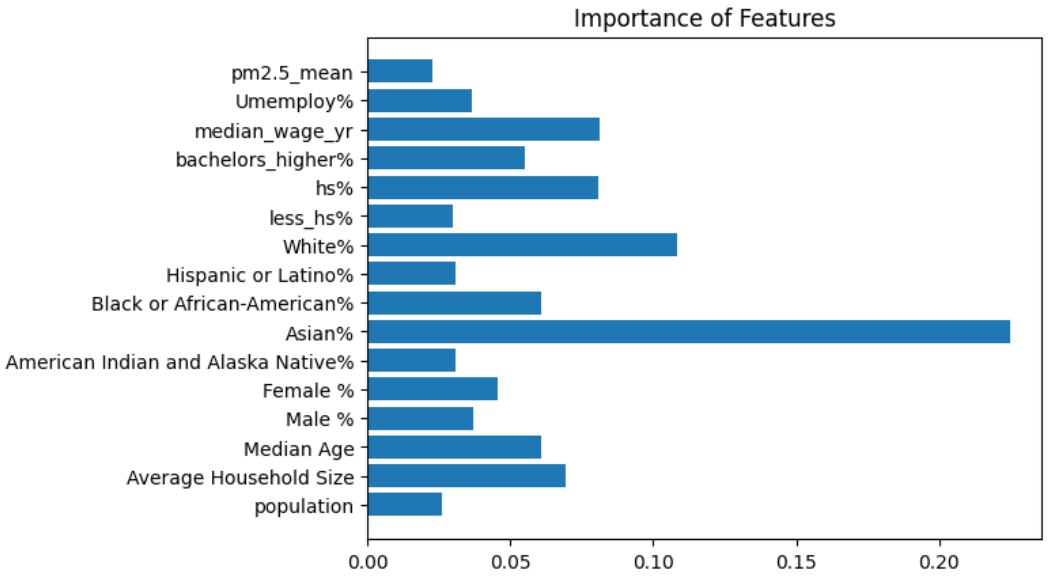
The second graph above depicts the amount of the top 100 cities with the highest COPD rates that are in each state. The table below shows the 100 worst cities in order along with their COPD rates. It seems that it is clear that there is an outlier in this graph with the state of Michigan. The reason for the city of Detroit specifically showing up so many times has to do with the regions that were measured. It seems that the city has a widespread issue or issues that have caused elevated rates of COPD in many different areas of the city. This is an interesting outlier to focus on because Detroit is a city that is relatively spread out and so you would think that over a large area of land it should not all be heavily affected by factors causing health issues in the residents. So we will be analyzing the potential causes of this behavior whether it has to do with the geography, infrastructure, or whatever else may cause this phenomenon.

Prediction with GLMs & Non-Parametric Methods:

Our group is predicting whether or not cities in the US have high rates of asthma and COPD (Chronic Obstructive Pulmonary Disease). The features we are using for prediction include demographic information (e.g. racial/ gender/ age composition, educational attainment, average household size, population), socioeconomic information (e.g. median wage, unemployment rate, etc), and information about pollution levels, specifically average PM2.5 measurements. We use socioeconomic information because cities might differ in available resources/ infrastructure to prevent negative respiratory health; We use PM2.5 measurements because there is an association between poor air quality and negative respiratory health; We use demographic information because different groups of people might not have the same levels of risk with respect to developing asthma or COPD.

For the GLM requirement, our group is using logistic regression. We are using logistic regression because our outcome is a categorical variable. We are modeling the probability that a city has a high rate of asthma diagnosis w/ a softmax function using the linear combination of the features and weights as the input. Since probabilities are bounded between 0 and 1 (inclusive), linear regression is not an appropriate model in the context of our research because the range is unbounded and therefore the outcome would not be interpretable. The assumptions of logistic regression is that the log-odds of the outcome variable has a linear relationship with the features and that all observations in the dataset are independent of one another.

The nonparametric models we are using for prediction are gradient boosted classification trees and random forests. Our group decided on both of the models because they are capable of predicting categorical outcomes. We did not use neural networks because the size of our dataset was too small which could've caused overfitting issues or minimal learning. Each model's performance was assessed using ROC-AUC because of the presence of significant class imbalance in our training dataset; the positive outcome class is overrepresented which likely could cause recall to be optimistic.



According to the summary of the random forest model, the most important features in the prediction of high COPD diagnosis rates were unemployment rate, median wage, educational attainment, and percentage of Asian population. This could be interpreted to be that COPD is largely preventable w/ economic and medical resources/ knowledge regardless of the pollution level of the city. This finding is in contrast to our group's hypothesis that cities with higher levels of pollution would necessarily correspond to higher rates of asthma and COPD rate.

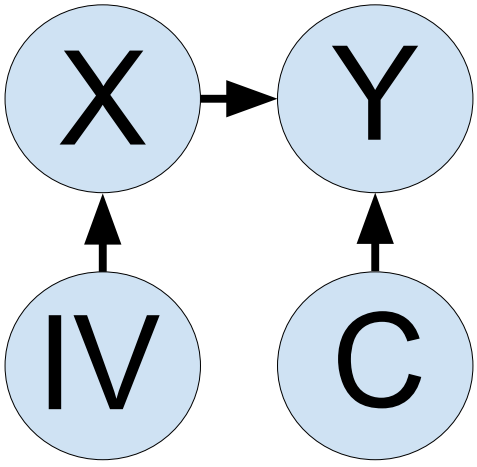
The random forest model had the highest ROC-AUC score on the validation set in comparison to the other models. This outcome could be attributed to the regularization the random forest model performs when selecting random subsets of features during the splitting stage. In effect, the correlation between the predictions of individual trees was lowered, resulting in lower variance of the aggregation of the predictions which could've assisted with the model's generalization to unseen data.

Some limitations of our models are that gradient-boosted trees have a higher likelihood of high bias because the trees are grown with less depth; decision trees with less depth are not

able to learn more complex relationships between the features and outcome. In addition, the random forest model has the same problem because of the noise it introduces into the splitting process. The models are also not interpretable compared to other models such as the linear regression model. It is not possible to read the decision tree for the random forest model or the gradient-boosted classification tree model because the predictions are an aggregation of multiple different trees.

Causal Inference:

To test the question of whether there is a causal relationship between pollutants and pulmonary diseases such as asthma and COPD, we decided to take into consideration possible instrumental variables we could use. In this case, we thought about how the distance from factories could have a big impact on the rates of these diseases. To use this instrumental variable, we did a two-stage least squares regression. We matched up the rates of chronic obstructive pulmonary disease and asthma in the 500 cities datasets and merged them to allow us to use them in regression. For the pollution measures, we used another dataset that had city-state pairings, along with the percentage of good days when it came to air pollution. The definition of a good day is considered to be when the 50 or less AQI. The dataset had 337 cities in the set, 260 of which were the same as the 500 cities in the COPD and asthma datasets. This was not ideal, but we continued on to see how well it would work with the two-stage least squares regression. When doing the original EDA, we found the percentage of people in the city who identified as black or African American turned out the be a confounding factor.



Where Y is the rates of asthma and COPD, X is pollution, IV is out instrumental variable which is the minimum distance to a factory and C is the confounder which is the percentage of people who identify and black or African American.

So, to take this into consideration, we added it in to mitigate its effect on the regression. We also found that a logarithmic transformation on the instrumental variable was more favorable, so we took the log of the minimum distance a city was from a factory.

OLS Regression Results

Dep. Variable:	GD%	**R-squared:**	0.029
Model:	OLS	**Adj. R-squared:**	0.021
Method:	Least Squares	**F-statistic:**	3.833
Date:	Tue, 12 Dec 2023	**Prob (F-statistic):**	0.0229
Time:	03:52:03	**Log-Likelihood:**	-1140.6
No. Observations:	260	**AIC:**	2287.
Df Residuals:	257	**BIC:**	2298.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	58.7081	3.400	17.265	0.000	52.012	65.404
min_dist_factory	1.4852	1.097	1.353	0.177	-0.676	3.646
Black or African-American%	0.2048	0.076	2.677	0.008	0.054	0.355

Omnibus:	39.786	**Durbin-Watson:**	1.298
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	53.609
Skew:	-1.014	**Prob(JB):**	2.29e-12
Kurtosis:	3.916	**Cond. No.**	70.2

First Stage of OLS Regression Results

With the results of the regression, we took them and fitted out predicted values based on the percentage of people with asthma and then those with COPD to check the correlation. We did this by taking the "0 coef" of each illness and the "min_dist_factory coef" to find $\frac{0.1457}{1.4852}$ for Asthma and $\frac{0.1940}{1.4852}$ for COPD. This resulted in an increase of 1% of 50 or more good days (in terms of AQI) corresponding to an increase in rates of Asthma by about 0.1% and an increase of 1% of 50 or more good days (in terms of AQI) corresponding to an increase in rates of COPD by about 0.13%.

OLS Regression Results

Dep. Variable:	Asthma%	R-squared:	0.176
Model:	OLS	Adj. R-squared:	0.172
Method:	Least Squares	F-statistic:	54.98
Date:	Tue, 12 Dec 2023	Prob (F-statistic):	1.75e-12
Time:	03:52:04	Log-Likelihood:	-384.25
No. Observations:	260	AIC:	772.5
Df Residuals:	258	BIC:	779.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2835	1.294	0.219	0.827	-2.265	2.832
0	0.1457	0.020	7.415	0.000	0.107	0.184

Omnibus:	4.725	Durbin-Watson:	0.924
Prob(Omnibus):	0.094	Jarque-Bera (JB):	4.392
Skew:	0.258	Prob(JB):	0.111
Kurtosis:	2.627	Cond. No.	1.29e+03

Second Stage of OLS Regression Results for Asthma

OLS Regression Results

Dep. Variable:	COPD%	R-squared:	0.144
Model:	OLS	Adj. R-squared:	0.140
Method:	Least Squares	F-statistic:	43.27
Date:	Tue, 12 Dec 2023	Prob (F-statistic):	2.64e-10
Time:	03:52:04	Log-Likelihood:	-489.76
No. Observations:	260	AIC:	983.5
Df Residuals:	258	BIC:	990.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.1853	1.942	-3.185	0.002	-10.010	-2.361
0	0.1940	0.029	6.578	0.000	0.136	0.252

Omnibus:	10.160	Durbin-Watson:	1.277
Prob(Omnibus):	0.006	Jarque-Bera (JB):	10.742
Skew:	0.486	Prob(JB):	0.00465
Kurtosis:	2.784	Cond. No.	1.29e+03

Second Stage of OLS Regression Results for COPD

We used the beta coefficient over the minimum distance coefficient, which is inverted, to get our final casual effect. Since the result was very low and not significantly significant, we decided to look for other datasets that would allow for more cities to be accounted for when taking the regression. This took a while since there were many datasets that would have some cities, but not all 500, and others without good indicators of pollution. We finally landed on a dataset that had 487 of the same city-state combinations and had the mean, median, and percentiles of pollution indicators of that year, which was 2014. This was not ideal since the year that the information about the rates of asthma and COPD was taken late on in 2017, which might lead to discrepancies in how much pollution actually was there that year. We then had to combine this dataset with the 500 cities datasets to be able to take the regression. We then re-preform two-stage least squares regression on the newly combined dataset while still accounting for the confounding variable we had previously found. To allow us to use this newly edited dataset, we had to go back and redo some of the EDA to allow for easier access to the full state name since that is what the new data frame had. We then got to the point where we had all the information we needed in one data frame, which included the full state name and abbreviation, city name, percentage of people with asthma, percentage of people with COPD, percentage of people who identify as black or African American, the mean of the pollution, and minimum distance from a factory. We then redid the regression and found that there was a larger negative effect when looking at the mean of the pollution.

OLS Regression Results

Dep. Variable:	mean	**R-squared:**	0.138
Model:	OLS	**Adj. R-squared:**	0.135
Method:	Least Squares	**F-statistic:**	37.72
Date:	Tue, 12 Dec 2023	**Prob (F-statistic):**	6.44e-16
Time:	03:52:20	**Log-Likelihood:**	-834.67
No. Observations:	473	**AIC:**	1675.
Df Residuals:	470	**BIC:**	1688.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.6577	0.188	56.808	0.000	10.289	11.026
min_dist_factory	-0.5071	0.063	-8.074	0.000	-0.631	-0.384
Black or African-American%	0.0053	0.004	1.238	0.216	-0.003	0.014

Omnibus:	4.597	**Durbin-Watson:**	0.869
Prob(Omnibus):	0.100	**Jarque-Bera (JB):**	4.675
Skew:	-0.226	**Prob(JB):**	0.0966
Kurtosis:	2.820	**Cond. No.**	65.2

Second Try: First Stage of Regression Results

We again took this model to fit our prediction and then found the coefficients that correlate to the rates of Asthma and rates of COPD. We did this again by taking the “0 coef” of each illness and the “min_dist_factory coef” to find $\frac{0.6133}{-0.5071}$ for Asthma and $\frac{0.8797}{-0.5071}$ for COPD.

Unfortunately, the team ran into a sign flip when adding the IV into the regression; this is a common phenomenon, but is not a result of faulty assumptions or a misspecification of the model. Logically, the correlation signs make sense. As a factory gets farther away, less pollution exists within a city. As pollution increases, Asthma and COPD rates increase. The IV Regression Coefficient measures the change in Y given we change the IV positively to elicit a one unit change in X. When applying a positive increase to the IV, we get decreasing values of X, and thus decreasing values of Y. This does not mean X and Y are negatively correlated. Thus, the team is justified in taking the absolute value of the IV Regression coefficients to isolate the true, positive causal effect of pollution on COPD and Asthma rates.

OLS Regression Results

Dep. Variable:	Asthma%	R-squared:	0.079
Model:	OLS	Adj. R-squared:	0.078
Method:	Least Squares	F-statistic:	40.67
Date:	Tue, 12 Dec 2023	Prob (F-statistic):	4.30e-10
Time:	05:12:35	Log-Likelihood:	-749.97
No. Observations:	473	AIC:	1504.
Df Residuals:	471	BIC:	1512.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.7891	0.917	4.131	0.000	1.987	5.592
0	0.6133	0.096	6.378	0.000	0.424	0.802

Omnibus:	3.923	Durbin-Watson:	0.999
Prob(Omnibus):	0.141	Jarque-Bera (JB):	3.930
Skew:	0.222	Prob(JB):	0.140
Kurtosis:	2.956	Cond. No.	163.

Second Try: Second Stage of Regression Results for Asthma

OLS Regression Results

Dep. Variable:	COPD%	R-squared:	0.078
Model:	OLS	Adj. R-squared:	0.077
Method:	Least Squares	F-statistic:	40.10
Date:	Tue, 12 Dec 2023	Prob (F-statistic):	5.64e-10
Time:	05:12:35	Log-Likelihood:	-923.97
No. Observations:	473	AIC:	1852.
Df Residuals:	471	BIC:	1860.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.1496	1.325	-1.622	0.105	-4.754	0.454
0	0.8797	0.139	6.332	0.000	0.607	1.153

Omnibus:	17.799	Durbin-Watson:	1.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18.186
Skew:	0.450	Prob(JB):	0.000112
Kurtosis:	2.667	Cond. No.	163.

Second Try: Second Stage of Regression Results For COPD

Conclusion:

With these calculations, we ended up finding that each increase of 1% of the pollution in a city corresponded to an increase in rates of asthma by about 1.2%. This same chance in pollution also corresponded with an increase in rates of chronic obstructive pulmonary disease by about 1.7%. This is consistent with our thoughts at the beginning and we are confident in these results as we have statistical significance and it makes sense that there would be higher rates of pulmonary conditions in cities caused by these higher rates of pollution. This is also a better outcome than what we had with the first dataset and the fact that it is more significant with more data shows that we can be confident in the outcome. To apply this, we should rebuild factories away from cities to allow for less pollution and thus lower rates of these pulmonary conditions. We can also find better ways to run factories that allow less pollutants in the air since it does negatively affect the people living close to the factories.

Based on our findings in the prediction section, we propose that governments upscale other preventative measures for respiratory diseases in addition to reducing greenhouse emissions, whether by building infrastructure or increasing the population’s medical knowledge.

One limitation of our research is that our dataset size is small (< 500 cities in the US) which makes it difficult to generalize our findings to the larger population in the US. Furthermore, the air pollution measurements for PM2.5 were collected from the years 2011-2014 which does not align with the measurements for asthma and COPD rates which were collected from the year 2016; the ACS data was collected from the year 2022. In addition, ACS data have a high margin of error/ variance for measurements because of the survey process.

A possibility for future research on a similar question could be what cities and their socioeconomic status are affected the most. This could show how redlining affects people's health and even their quality if it turns out that factories are built closer to cities with a generally lower socio-economic status.