[ds102\_project\_fa23.pdf - Google Drive](https://drive.google.com/file/d/1cBz7raoqPeYcshWqv89KslZIOTfC91-F/view)

[airqualitytrendsbycity1990-2022.xlsx (live.com)](https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.epa.gov%2Fsystem%2Ffiles%2Fdocuments%2F2023-05%2Fairqualitytrendsbycity1990-2022.xlsx&wdOrigin=BROWSELINK)

[U.S. Pollution Data (kaggle.com)](https://www.kaggle.com/datasets/sogun3/uspollution) \*\*

[Download Files | AirData | US EPA](https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual)

[Global Air Pollution Dataset (kaggle.com)](https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset)

<https://data.cdc.gov/Environmental-Health-Toxicology/Daily-Census-Tract-Level-Ozone-Concentrations-2011/372p-dx3h>

Checkpoint 1:

**COPD and Asthma Datasets:**

• Visualize at least two quantitative variables and two categorical variables. Your visualizations

must be relevant to your research questions!

• Describe any trends you observe, and any relationships you may want to follow up on.

• Explain how your visualizations should be relevant to your research questions: either by moti-

vating the question, or suggesting a potential answer. You must explain why they are relevant.

**Toxic Air Pollution Hotspots Datasets:**

• Visualize at least two quantitative variables and two categorical variables. Your visualizations

must be relevant to your research questions!

• Describe any trends you observe, and any relationships you may want to follow up on.

• Explain how your visualizations should be relevant to your research questions: either by moti-

vating the question, or suggesting a potential answer. You must explain why they are relevant.

Median household income, city population, median age, ethnicity, median educational attainment

Checkpoint 2:

**Prediction**

**Describe what you’re trying to predict, and what features you’re using. Justify your choices.**

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.

**Describe the GLM you’ll be using, justifying your choice. Describe any assumptions being made by your modeling choice.**

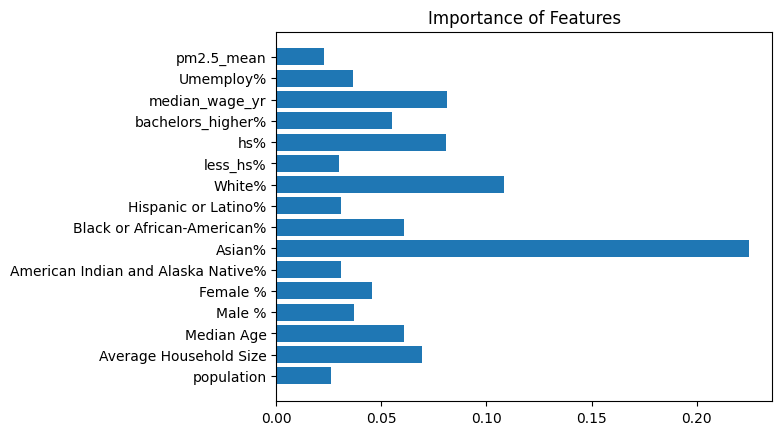
For the GLM requirement, our group is using logistic regression. We are using logistic regression because our outcome is a categorical variable. We are modeling the probability that a city has a high rate of asthma diagnosis w/ a softmax function using the linear combination of the features and weights as the input. Since probabilities are inclusively bounded between 0 and 1, 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 the outcome for all individuals are mutually independent.

**Describe the nonparametric method(s) you’ll be using, justifying your choice. Describe any assumptions being made by your modeling choice.**

**How will you evaluate each model’s performance?**

The nonparametric models we are using for prediction are gradient boosted classification trees and random forests. Both models 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 will be assessed using ROC-AUC because of significant class imbalance in our training dataset; the positive outcome class is overrepresented which could mean that recall is optimistic.

**Summarize and interpret the results from your models.**

****

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.

**Which model performed better, and why? How confident are you in applying this to**

**future datasets?**

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 model performs when selecting random subsets of features during the splitting stage. In effect, the model’s predictions have lower variance which could’ve assisted the model’s generalization to the testing set.

**Elaborate on the limitations of each model.**

Some limitations of our models are that gradient boosted trees have a higher likelihood of having high bias because the trees are grown w/ less depth. 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.

**Explain the uncertainty in your results: is it qualitatively high or low? Explain why this**

**might be (e.g., noisy data, dataset size, variance in the estimation, etc.).**

Sources of uncertainty in our results are that our dataset size is small (< 500 cities in the US) which makes it difficult to generalize our findings to the larger population. Furthermore, the air pollution measurements for PM2.5 was collected from the years 2011-2014 which does not align with the measurements for asthma and COPD rates which is from the year 2017. In addition, data from the ACS (American Community Survey) have a high margin of error/ variance for measurements because of the survey process.

**Causal Inference:**

Formulate a causal question, clearly defining the treatment, control, and units (people, states,

months, etc.). Use one of the techniques you learned in class to answer the question, clearly stating and justifying any and all assumptions you make.

**Question:**

Does pollution cause increased rates of asthma and COPD rates?

* Asthma/COPD are done as a proportion of the population (possibly per 100,000 as a future experiment)
* We are utilizing an IV of minimum distance of a city to the closest air-polluting factory in kilometers
* Our X value is number of days with AQI >= 50, however this is subject to change (likely a more fine grained statistic about pollution will be used)
* We found that only the proportion of Black people in the cities is a cofounder, so after correcting for this we assume unconfoundedness

**Methods:**

**Describe which variables correspond to treatment and outcome.**

Our IV is the distance each city is to the closest factory in miles. Our treatment is the pollution each city undergoes. Currently, it is tracked as days of AQI <=50 because distribution of daily AQI is very right skewed (maybe viable to correct with transformation). The treatment will likely be changed. The outcome is the COPD and Asthma rates each city has.

**Describe which variables (if any) are confounders. If the unconfoundedness assumption holds, make a convincing argument for why.**

After analyzing various demographic and socioeconomic variables in our EDA, the only variable that stood out was the proportion of african americans within each city. This was added as a cofounder, and we assume unconfoundedness of the IV.

**What methods will you use to adjust for confounders?**

We will integrate the cofounder into the preformed regression as well as visually inspect the fit to see if there is a non-linear relationship present.

**Are there any colliders in the dataset? If so, what are they?**

None to our knowledge.

**Draw the causal DAG for your variables.**

Cofounder, IV -> X -> Y <- Cofounder

**Results:**

**Summarize and interpret your results, providing a clear statement about causality (or a**

**lack thereof) including any assumptions necessary. In addition to statistical significance,**

**discuss the magnitude of any effect you find.**

Correlative Causal Effect of X on Y1: 0.0981

Increase of 1% of >=50 AQI days corresponds to an increase in rates of Asthma by ~0.1%

Correlative Causal Effect of X on Y2: 0.1306

Increase of 1% of >=50 AQI days corresponds to an increase in rates of COPD by ~0.13%

Regression of IV on X was not statistically significant, but regression of X\_pred on Y1 and Y2 was. This leads us to think a relationship does exist, but better selection of pollution data (i.e. more fine grained) is needed.

**Where possible, discuss the uncertainty in your estimate and/or the evidence against the**

**hypotheses you are investigating.**

Biggest source of evidence against the hypothesis is the non statistically significant regression coefficient of IV on X, and the possibility for cofounders outside of demographic and socioeconomic data. Difficulty in getting city-wide data for all the 500 most populated cities in the United States is the main barrier to expanding the number of possible confounders we could test. A better dataset for pollution (X) is needed for this analysis.

**Discussion:**

**Elaborate on the limitations of your methods.**

We are limited by the format of the data we are using for X, Y1, and Y2; we can only make conclusions in that specific unit type. Additionally, we assume unfoundedness based on socioeconomic and demographic data, but this may not be the only cofounders to check (in order to ensure our IV is legitimate).

**What additional data would be useful for answering this causal question, and why?**

A more fine-grained datasource on pollution by city and some key data transformations to move it away from a very right skew distribution. More standardized data on the most 500 populous cities in the United States.

**How confident are you that there’s a causal relationship between your chosen treatment and outcome? Why?**

We are confident based on intuitive logical thinking, but the magnitude is under question. Even given a statistically significant result, the magnitude is likely, at best, a rough approximation of the true effect.

**Write up (3000-5000 words)**

You must submit a typed PDF document that contains each of the sections described in “Section Guidelines”. Your report should be between 3000 and 5000 words of text, in addition to tables, figures, and references. All mathematical equations must be rendered properly in LATEX, Equation Editor, or similar. We won’t be strictly enforcing this limit, but reports that are much longer than this are subject to a penalty (reports that are much shorter are probably missing important discussion).

Your report should be a proper written document: you cannot just submit a printed Jupyter notebook (including data sources, code, outputs, etc.) We highly recommend using Overleaf or Google Docs.

If relevant, include a reference page with citations of all outside sources used.

All figures and tables should be included in your written report. Clearly label all figures and include informative captions. These labels should be used to reference figures and tables in your written report. Refer to this guide for instructions on inserting images into a LATEX file.

* Data Overview: Christian, Chinmay
* Research Questions: Christian
* EDA: Christian, Chinmay
* Prediction: Chinmay
* Causal Inference: Sarah
* Conclusions: Sarah, D