Air Quality and COPD

Data Overview

- U.S. Chronic Disease Indicators: Chronic Obstructive Pulmonary Disease
    - Sample dataset of Chronic diseases indicator filtered for COPD. The dataset is collected to allow states, territories, and large metropolitan areas to define, collect, and report chronic disease data uniformly. Dataset seems to have a systematic bias toward devaluing minorities. Compared to the true distribution of the population, the minority groups under the specified limit for the data to be collected will be ignored from the dataset.
    - Each row represents a response from a group of people. We will have to analyze the race, county, or higher or equal level than the group. We have found out that the number of individuals of different races has not been well distributed during the EDA process. Mainly focused on specific races. If the relationship between race and the COPD data, we will have to address race distribution. Although the granularity of the dataset is high, It would have been better if we could go deeper into analysis if the dataset was generated per individual.

- CDC: Daily Census-Tract PM2.5 Concentrations
    - Daily census PM2.5 concentration dataset from CDC to generate air quality measure. Each row represents the PM2.5 concentration of different locations on each day. Hundreds of million rows of data have been collected over a year. Daily data has been collected from the same place over the years. Therefore an area slightly off from the location will be impossible to be calculated accurately. Calculation based on higher-level analysis such as country or state-level will help to address bias.

- CDC: Daily Census-Tract Ozone Concentrations
    - Daily census Ozone concentration dataset from CDC to generate air quality measure. Each row represents the Ozone concentration of different locations on a single day in 2011. About hundreds of million data were collected in 2011. Concerns and bias are identical to the CDC: Daily Census-Tract PM2.5 Concentrations dataset.

- External datasets
    - Common purpose: To address the causal impact of confounding variables while drawing out the causal relationship between smoking and COPD prevalence.

- Greenhouse gas emission
    - Each row represents each facility’s greenhouse gas emission (GHG) information for 2011. The dataset is the census dataset. However, the dataset is mainly collected from large facilities in the States. Therefore, most of the information about smaller businesses has been omitted. The current granularity of the data is yearly based.

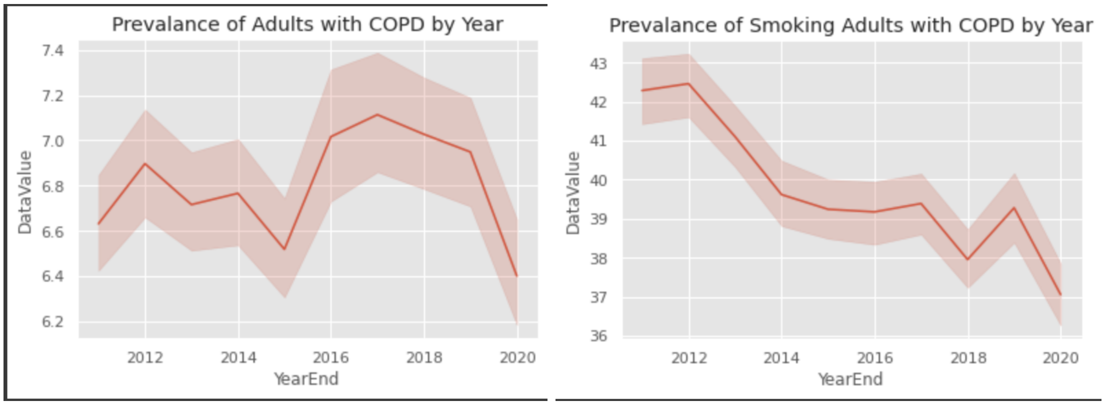
- Wildfire
    - Randomly sampled dataset from census dataset of total incidents of wildfires in the States. Selection bias seems to exist in the sampled dataset since it has been sampled according to its specific longitude and latitude. Each row represents a single wildfire incident in the States. The incident is recorded only if others have found it. Therefore, if the incident happens where people don’t live, it is highly likely that the incident doesn’t get collected.

- Population density
    - State-level population dataset from 2010 to 2019. Census data that Each row represents the number of population for each state or region in the United States. Each column represents estimated or census population for each particular year. Non-response bias is a potential bias in the dataset.

- Smoking prevalence
    - County level smoking prevalence census dataset. Each row represents a single county’s bounded smoking prevalence according to gender. The smallest measurement level of the data set is a county. We should keep other datasets to county level or higher to maintain the consistency of the data and draw out reasonable conclusions. The dataset is collected based on the belief that respondents respond correctly. However, the questionnaire is about bad habits. We cannot always expect truthful responses from the respondents.

EDA

Data Visualization 1: Does Smoking Affect the Prevalence of COPD?



Trends

- We have observed how ‘Prevalence of chronic obstructive pulmonary disease among adults >= 18' and 'Prevalence of current smoking among adults >= 18 with diagnosed chronic obstructive pulmonary disease' differs by year.
- It seems like both of the line plots are showing that the prevalence is gradually decreasing over several years. Suprising fact is that while the lowest average prevalence value is around 6.4% in 2020, the value for the smoking adults was 37% around 2020. This implies that smoking is a significant factor that causes COPD among adults.

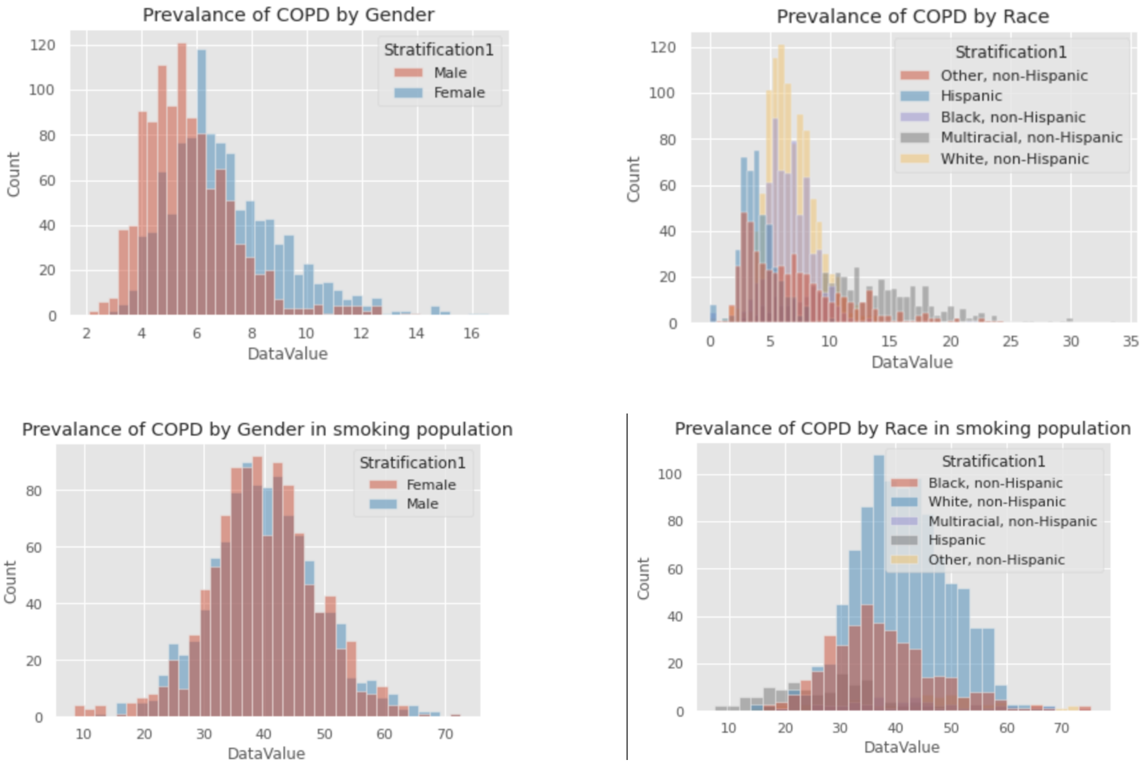
Data Cleaning

- Since our research question is interested in the prevalence of COPD among adults we have filtered out the rows that only contain this information.

Relevance

- I could use this implication to use smoking as a confounding factor for research question 1 since smoking both affects the prevalence of COPD among adults and the air quality. Also,
- It can act as our predictor for research question 2 since the smoking variable significantly affects the prevalence of COPD proven by the visualization.

Data Visualization 2: visualize how prevalence of COPD for combined/smoking adults differs by race and gender



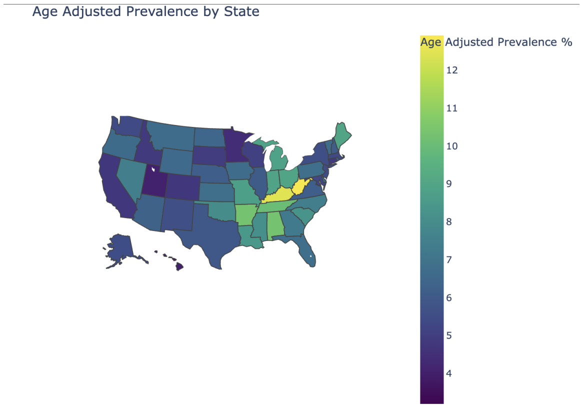
Trends

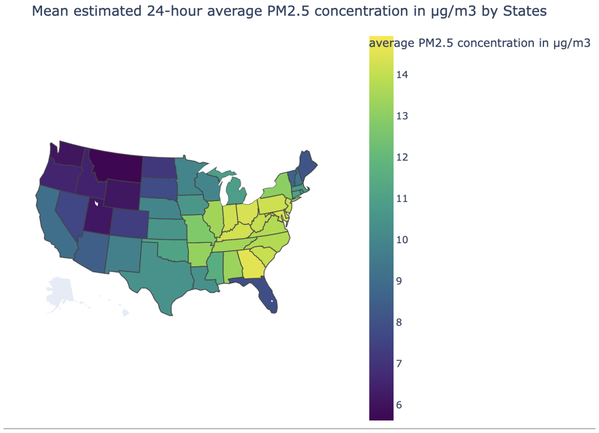
- Prevalence of COPD by Gender: It seems like the distribution of male's prevalence is located on the left of the distribution of female's prevalence which implies that the average of males' prevalence is lower than females'. We could assume that females might have more probability to get diagnosed with COPD.
- Prevalence of COPD by Race: It seems like the distribution of hispanics has relatively lower prevalence than black and white. Also it seems like the distribution of white and black follows a similar distribution. Multiracial, non-hispanic people seem to have relatively higher prevalence than other races.
- Prevalence of COPD by Gender in Smoking Population: Among smoking adults, Gender does not seem to be the critical factor that differentiates in the distribution of prevalence of COPD.
- Prevalence of COPD by Race in Smoking Population: Among smoking adults, it also seems like the distribution of hispanics has relatively lower prevalence than other races.

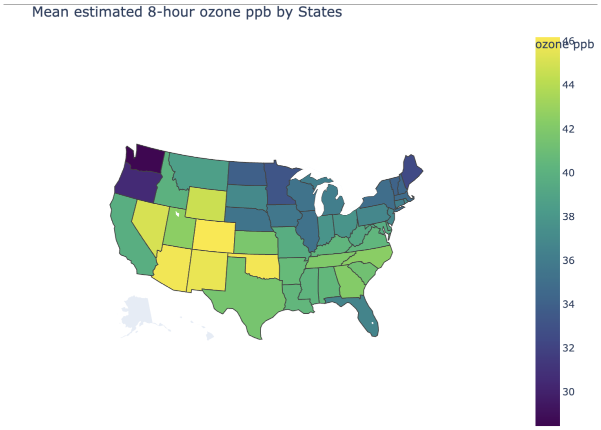
Relevance

- Since it seems like the distribution of prevalence rate differences by both race and gender, we could use these stratification factors as predictors to predict whether people have chronic disease.
- Since it seems like the distribution of prevalence among smoking adults seems to have non-outstanding variables that are helpful to predict the prevalence of COPD, it gives us certainty to stick with the prevalence of COPD among the combined population instead of the smoking population.

Data Visualization 3:


Age Adjusted Prevalence by State


Mean estimated 24-hour average PM2.5 concentration in µg/m3 by States


Mean estimated 8-hour ozone ppb by States

Trend

- These color coded geo maps show how age adjusted prevalence rate of COPD, ozone concentration, and pm 2.5 concentration differs by states. By looking at geomaps of COPD prevalence rate and pm 2.5 concentration, we could observe that regions with high pm 2.5 concentration (Kentucky, West Virginia, etc) also tend to have high COPD prevalence rate. This indicates that there exists some relationship between pm 2.5 and COPD rate; we aim to define this relationship using causal inference technique. It was interesting to see that regions with higher ozone concentration (Nevada, Arizona, etc.) did not necessarily have higher COPD rate.

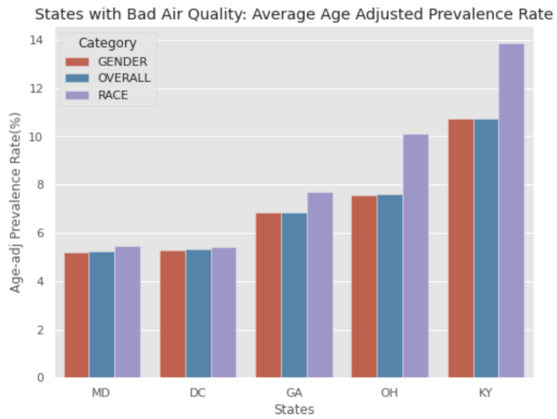
Data cleaning

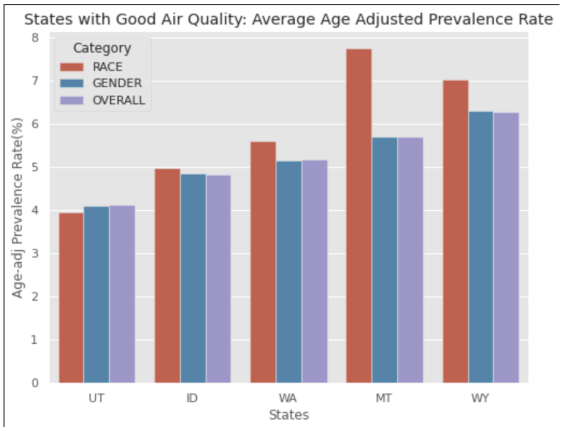
- Since the granularity of COPD data is in each state, we grouped ozone & pm 2.5 tables by states as well, by taking the mean. During the process, ozone & pm 2.5 data could have lost some information. For instance, while taking the mean, we could have accidentally included outliers.

Relevance

- This is relevant to our research question regarding the relationship between air pollutants and age adjusted prevalence rate of COPD. The visualization clearly shows that states with higher pm 2.5 have higher COPD rate.

Data Visualization 4:





Trend

- These two bar plots each compare age adjusted prevalence rate of COPD in states with good and bad air quality. We could observe that states with bad air quality have higher age adjusted prevalence rate of COPD compared to that of states with good air quality. For instance, the COPD rate in states with bad air quality ranges from 6% to 14%, while that of good air quality states ranges from 4% to 7.8%.

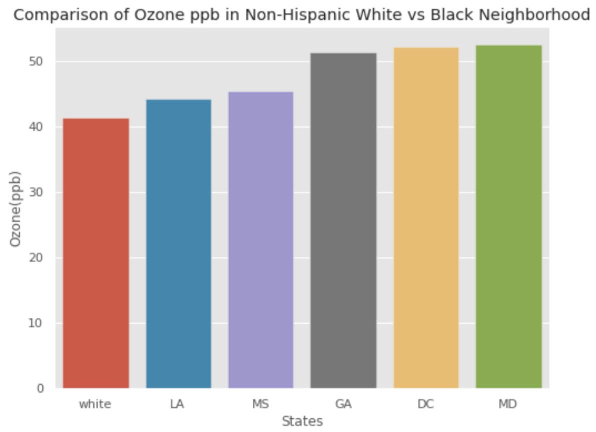
Data cleaning

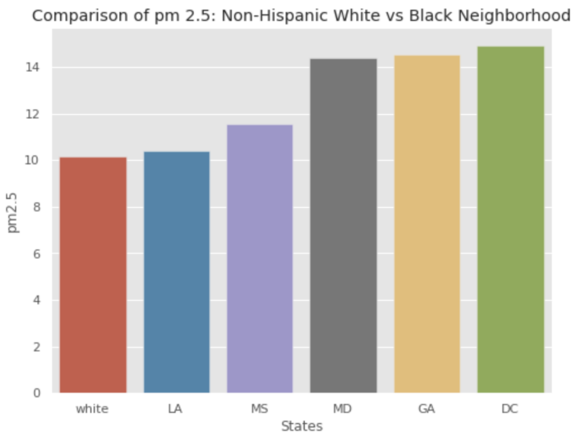
- As stated in the research question, we defined bad air quality as ozone concentration greater than 80 ppb and PM 2.5 concentration greater than 35.4µg/m3. However, since the ozone & pm2.5 table for this visualization only contained 5M rows each, the data was not comprehensive enough to allow strict classification of good and bad air quality. Thus, we took the yearly average ozone/pm2.5 concentration by each state and defined five bad/good air quality states as those with highest/lowest ozone and pm2.5 concentration, respectively.

Relevance

- This is relevant to our research question regarding the relationship between air pollutants and age adjusted prevalence rate of COPD. The visualization clearly shows that states with bad air quality have higher COPD rate.

Data Visualization 5:





Trend

- These two bar plots each compare ozone ppb and pm2.5 concentration in non-hispanic white and Black neighborhoods (top five states with highest non-hispanic/black population, according to census data). Here, we could observe that both ozone ppb and pm2.5 concentration in white neighborhood are relatively lower than that of in Black neighborhood. This trend may suggest that the climate burden exists in minority neighborhoods. This may also imply that since people of color are easily exposed to air pollutants, they are more susceptible to health consequences.

Data cleaning

- Due to memory shortage, I only took 5 million rows for each ozone and pm 2.5 data. This may have affected data coverage thus harming credibility of the data, since it only contained a portion of 2011 data.

Relevance

- This is relevant to our research question regarding the relationship between geographical location, race, and COPD prevalence. Starting from here, we can further examine if most black populated states have a higher chance of getting diagnosed with COPD. From the visualization, we have identified how racial composition differs by states and how these states have different levels of air pollutant.

Inference and Decisions

Research Question 1: Causal Inference

Question: Does low air quality cause the onset of Chronic Obstructive Pulmonary Disease (COPD)?

- By answering the question, we are able to determine whether or not people who residence in states with low air quality are exposed to chronic obstructive pulmonary diseases. It helps us to predict the future and recognize which variables have the capacity to affect the air quality and the prevalence of chronic diseases. We can change the future by intervening to change the variables that affect the air quality negatively, or improve the variables that filter air quality.
- In this case, causal inference would be a good fit for answering our research question because we are trying to make a decision whether the states with bad air quality causes the onset of chronic obstructive pulmonary disease. We will recognize confounding variables that affect both the air quality and the prevalence of COPD, so that our treatment and potential outcomes are conditionally independent given a set of known confounding variables.

Methods

- Methods:
    - Treatment ("Treatment" in merged_df) : treatment =1 contains 16 states with bad air quality and treatment = 0 contains 36 states with good air quality in 2011
        - treatment decisions are not made completely at random
    - Outcome ("Datavalue" in merged_df) : the prevalence of chronic obstructive pulmonary disease among adults greater than or equal to 18.
    - Confounding variables: greenhouse gas emissions, smoking prevalence, number of wildfire occurrences & size of the wildfire, population density of each state
- Unconfoundedness Assumption Holds:
    - States with bad quality might have more intense greenhouse gas emission, higher rate of smoking prevalence, more frequent number of wildfire occurrences, larger magnitude of the wildfire and larger population density .
    - Treatment and the potential outcomes are conditionally independent given these sets of known confounding variables which also means that we observed all the relevant confounding variables.
    - SUTVA (Stable Unit Treatment Value Assumption)
        - We assume that states with bad air quality do not affect the states with good air quality.
    - There are no other additional confounders that have an effect on the treatment and the outcome.
- Methods we use to adjust for confounders
    - Outcome regression
        - We decided to fit a linear model of the following form: (Z is treatment)

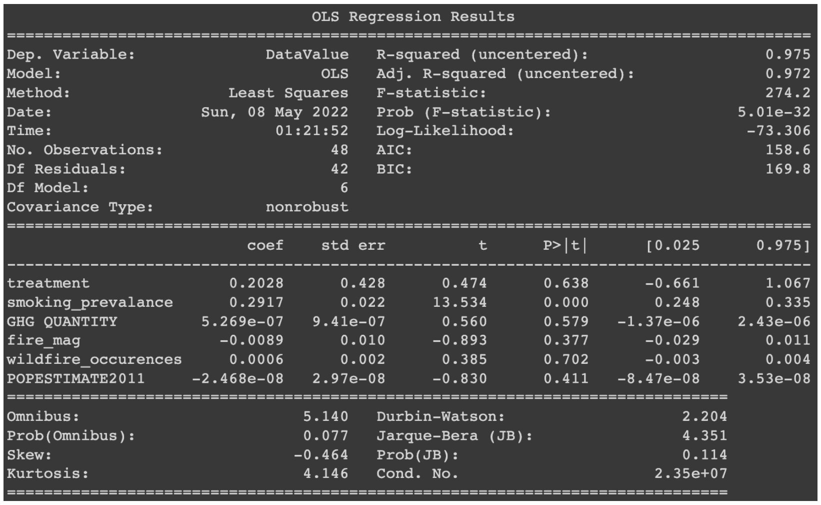
Prevalence = τ * Z + a* Prevalence of Smoking + b * Greenhouse Gas emission + c * Magnitude of Fire + d * Occurrences of wildfire + e * Population density of 2011

        - Under the following assumptions, the estimated coefficient of treatment from OLS, predicted τ, will be an unbiased estimate of the ATE:
            1. Assume unconfoundedness given this set of 5 confounding variables.
            2. Assume this new linear model correctly describes the interaction between the variables.
    - Inverse Propensity Weighting
        - Calculate propensity scores; probability that state has a bad air quality, conditioned on the set of confounding variables using logistic regression
        - Use IPW estimator of the ATE

$$\hat{\tau}_{IPW} = \frac{1}{n} \underbrace{\sum_{i:Z_i=1} \frac{Y_i}{e(X_i)}}_{\text{reweighted treated rows}} - \frac{1}{n} \underbrace{\sum_{i:Z_i=0} \frac{Y_i}{1-e(X_i)}}_{\text{reweighted control rows}}$$

Results

1. Outcome Regression



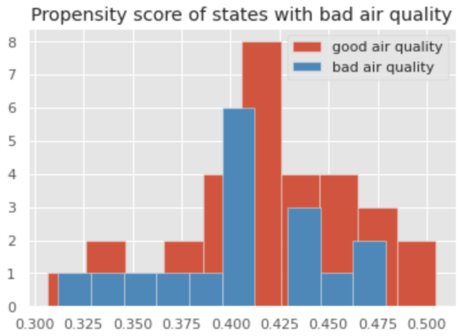
```
                                OLS Regression Results
=======================================================================================
Dep. Variable:              DataValue   R-squared (uncentered):                   0.975
Model:                            OLS   Adj. R-squared (uncentered):              0.972
Method:                 Least Squares   F-statistic:                              274.2
Date:                Sun, 08 May 2022   Prob (F-statistic):                    5.01e-32
Time:                        01:21:52   Log-Likelihood:                         -73.306
No. Observations:                  48   AIC:                                      158.6
Df Residuals:                      42   BIC:                                      169.8
Df Model:                           6
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
treatment               0.2028      0.428      0.474      0.638      -0.661       1.067
smoking_prevalance      0.2917      0.022     13.534      0.000       0.248       0.335
GHG QUANTITY         5.269e-07   9.41e-07      0.560      0.579   -1.37e-06    2.43e-06
fire_mag               -0.0089      0.010     -0.893      0.377      -0.029       0.011
wildfire_occurences     0.0006      0.002      0.385      0.702      -0.003       0.004
POPESTIMATE2011     -2.468e-08   2.97e-08     -0.830      0.411   -8.47e-08    3.53e-08
==============================================================================
Omnibus:                        5.140   Durbin-Watson:                   2.204
Prob(Omnibus):                  0.077   Jarque-Bera (JB):                4.351
Skew:                          -0.464   Prob(JB):                        0.114
Kurtosis:                       4.146   Cond. No.                     2.35e+07
==============================================================================
```

- Summary: Estimated coefficient of treatment from OLS, predicted τ = 0.2028
  Confidence Interval at 5% significance level = [-0.661, 1.067]
- Interpretation: The estimated coefficient is close to zero and the confidence interval of it contains negative value, it implies that the estimated effect of the air quality using outcome regression is that the program does not cause people to have higher prevalence of COPD.
- Uncertainty: We assumed unconfoundedness that there are only 6 confounding variables that cause both treatment and outcome. There might be missing confounders not taken care of by the data and the model which lead us to this decision. Also, it is not clear that confounding variables would all have a linear effect on the prevalence of COPD. This leads us to an uncertainty that our linear model does not model any interactions between the variables.

2. Inverse Propensity Weighting
    - ipw_estimate = 6.82
    - Interpretation: If we assume that the outcome, age-adjusted prevalence of COPD among adults, is unconfounded given the variables and the propensity score is a good model of probability of states that has a bad air quality given the listed confounding variables , then

the estimated effect of bad air quality is that the bad air quality causes people to have 6.83% more prevalence of being diagnosed to COPD than people who residence in good air quality.

  - Uncertainty: Similarly as outcome regression, there might be missing confounders not taken into account by the data. The propensity score is calculated based on logistic regression, its calculated value could be misinterpreted which also leads to uncertainty in our IPW estimate. When we observe the histogram of propensity scores of states with bad/good air quality, it seems like there are no big differences between the distribution of each treatment. It is questionable whether the confounding variables that we have used to model the treatment can differentiate among the treatment groups.


Propensity score of states with bad air quality

Discussion

- Since our PM 2.5 concentration dataset only contains information from 2011 and Ozone dataset only contains information from 2011-2014, we have made our decision based on 2011 to keep consistency in our unit. However, if we had more data for more years we could have relied on more data which gives more certainty on our decision.
- We have identified more confounding variables other than listed above, such as number of factories in a region, number of cars driven on the road, fuel efficiency of those cars, precipitation, presence of mountains, etc, however, it was difficult to find the dataset so we had to assume unconfoundedness with few confounding variables.
- Since we assigned our treatment group as bad and good air quality based on PM 2.5 concentration and Ozone concentration, the SUTVA assumption might not hold because due to complicated nature, units might affect each other which violates the assumption.
- We are not confident on our causal relation between our chosen treatment and outcome because we didn’t include some confounding variables. It is uncertain to make a decision that bad air quality causes the onset of COPD since there are so many determinants of air quality and COPD which makes the model way more complicated. Therefore, our unconfounded assumption does not hold, so our outcome regression and IPW estimator would have been arbitrarily biased.

Conclusion

- To summarize our key findings for causal inference, we have identified the causal effect of air quality on the prevalence of Chronic Obstructive Pulmonary Disease. Our findings are relevant to 2011 since we have used our datasets that only contain the corresponding year. Also, since our treatment units are states that have bad/good air quality we could say our findings are broad compared to findings from individuals who reside in more detailed geographic areas that have bad/air quality are exposed to COPD.

- Based on our results, since we have concluded that bad air quality causes the prevalence of COPD, the states with bad air quality can improve their legislation that regulates air pollutants generated from transportations, manufacturers, or place more air filtering technologies to control air pollutants generated from wildfires. So that individuals could be less exposed to ozone and pm 2.5 concentration which also decreases the prevalence of COPD.
- For causal inference we had to merge different datasets. Datasets that we have combined were COPD, Ozone and PM 2.5 Concentration, Greenhouse gas emission, Wildfire, Smoking prevalence, Population density of 2011. We have grouped each dataset by state and took the mean of each corresponding value that we are interested in. The benefits of combining different sources for our research question, we were able to identify our treatment and outcome by merging air quality and COPD prevalence dataset and by adding confounding variables we were able to improve our model quality and certainty in our causal inference decision.
- For causal inference, we could have found more confounding variables that affect both the prevalence of COPD and the air quality. There are so many possible determinants that affect air quality and prevalence of COPD that we didn't take into consideration due to inaccessibility of data. For future studies, we could have found the relationship between genetic factors and prevalence of COPD or obtain census data that contains information of personal demographics of each state.

Research Question 2: GLM & Random Forest

The second research question is "Can we predict whether people have COPD from geographical location and race/ethnicity?". By answering this question, we will better understand which set of the population is most likely to have COPD compared to the general population. Thus, we can prioritize implementing preventative and treatment measures to those who are more likely to have COPD.

GLM is a good fit for answering our second research question because we are trying to verify if onset of COPD has a relationship with geographical location and race/ethnicity. Thus, if we identify the relationship, we can successfully predict whether people have COPD by using geographical locations and race/ethnicity.

Random forest is a good fit for answering our second question because we aim to predict binary decisions of whether or not people have COPD. It is also least likely to overfit our data. Since our prediction is binary, we thought that the model has plenty of room to overfit our data. Random forest is known to be less likely to overfit to data; thus, we chose random forest.

Methods

GLM:

We chose the individual's race (black, white, hispanic, multiracial, other) and location (longitude and latitude of the states) as the features of our model and the prediction was whether individuals will be likely to have COPD. We were able to find the longitude and latitude of the status from an external dataset (reference). We chose location and race because we thought these were the most relevant features that we could use to predict. After researching, a few articles, including the provided news articles, suggest that racial composition differs by location and that particular areas with dense minority communities tend to have a higher prevalence of respiratory diseases.

Because the model will predict whether individuals are likely to have COPD or not, we chose to build a logistic regression model. We decided to say an individual is likely to have COPD if his COPD likelihood is greater than the median of the sample's COPDs, which was about 6.4. The features related to race are all binary variables: if an individual's race was black, the variables would be 0 for white, hispanic, multiracial, other and 1 for black. Consequently, we will be using logit as our link function and Bernoulli as our likelihood function.

A GLM, logistic regression, is a good fit for our research question because we have to determine whether an individual is likely to have COPD or not. Thus, logistic regression is the best method to clearly show this relationship.

By using GLM to answer this question, we assumed that each Ys (whether an individual is likely to have COPD or not) is independently distributed. Moreover, we assumed that each of the Ys follows a binomial distribution. Finally, we assumed that geographical locations will incorporate different levels of air pollution in each state.

For the Bayesian GLM, we will evaluate the model's performance by conducting a posterior predictive check. If the posterior predictive includes the majority of the data, we can conclude that our model did a fairly good job of fitting to our data.

For the Frequentist GLM, we will evaluate the model's performance by looking at the average log-likelihood of the model and also at the value of pearson chi2. If the value of average log-likelihood is close to 0 and the value of pearson chi2 is close to the number of samples subtracted by the number of parameters, we will be able to verify that our model has a good performance. If the two criteria contradict, we will be able to choose from either of the two.

Random Forest:

We used a random forest for prediction. We chose random forest instead of decision tree because random forest has lower variance than decision tree, and is less prone to overfitting. Furthermore, given that we were working with a specified dataset of COPD (only looking into age-adjusted prevalence rate with specific question type), we believed ~1500 rows of data were not sufficient enough and are prone to outliers. This also meant that data was not sufficient enough for us to use the Neural Network method, since it typically allows plenty of data for accurate prediction. Hence we chose random forest because it is robust to outliers since it handles outliers by binning them. By choosing random forest, we are assuming that there are no formal distributional assumptions. In other words, because random forest does not use any probabilistic model, but rather performs binary split (left & right child node for each decision tree), we are not making any assumptions about underlying data distribution. Moreover, since random forest uses bootstrap aggregation, there exists no model underneath. Thus we are assuming that each sampling is representative of the dataset. We are also assuming that bagging will resolve noise. As seen in class, it is fairly common that at least 10% of the data are corrupted by noise.

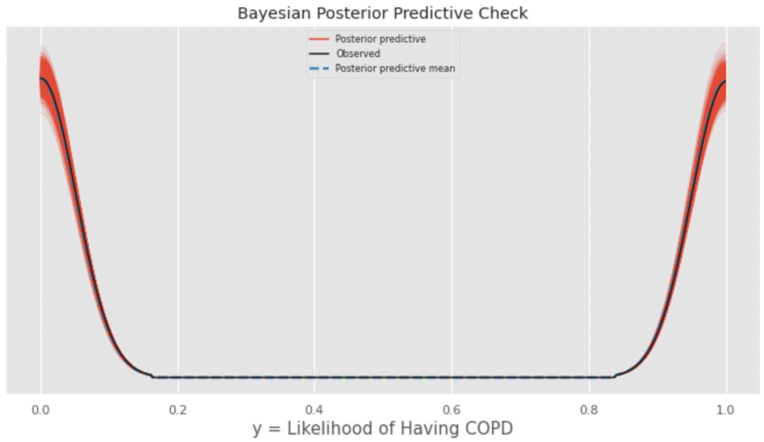
We will evaluate the performance of random forest by computing TPR and TNR. We will define true decision (D=1) as predictions greater than 60%, and define false decision (D=0) as predictions less than 40%. Here we chose 60% and 40% instead of 50% and 50% because the latter defines a random guesser rather than a predictive model. We will also test the performance by computing RMSE of both train and test sets. This will let us understand how off the prediction is from the true observed values.

Results

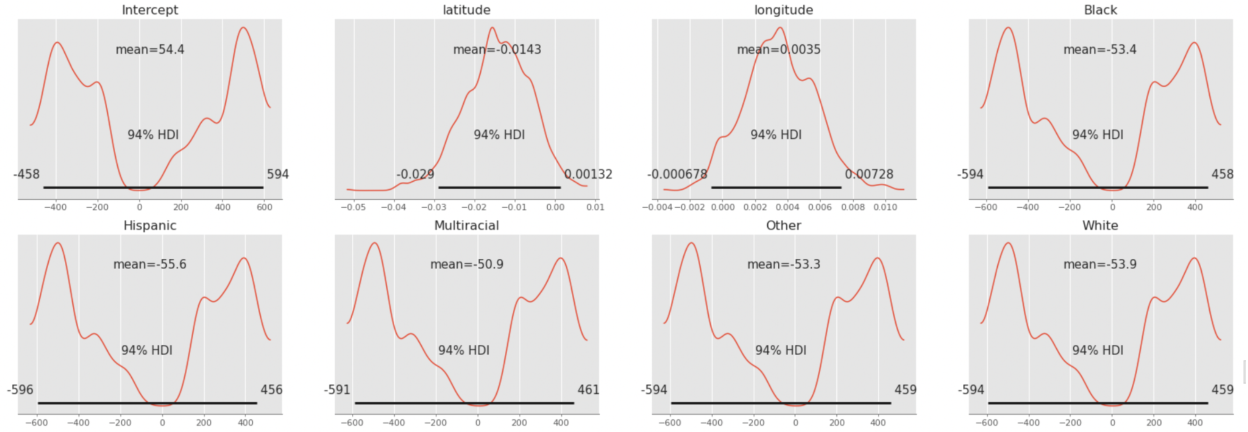
<u>GLM:</u>

1. Bayesian

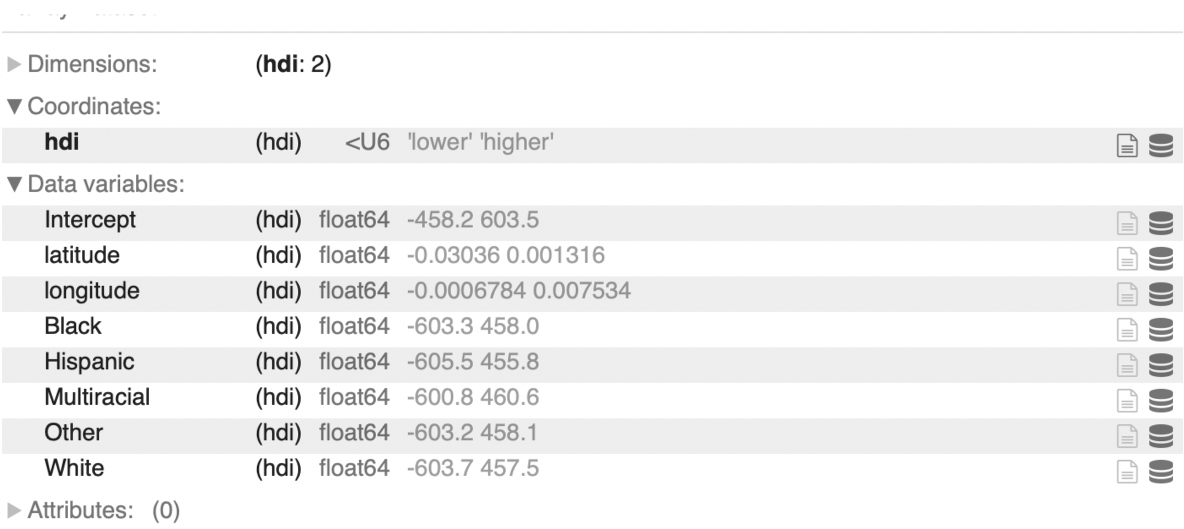
Looking at the PPC diagram below, It seems like our posterior predictive is really close to our actual values of the data; however, it also seems like the model has overfitted to the data as all the values are within the posterior predictive.



Looking at the posterior distributions of features, individuals who are hispanic are most likely to have COPD ceteris paribus. Moreover, it seems like the distributions of latitude and longitude are fairly normal at around 0, with small values of standard deviations. However, the other variables are very skewed and have very large values of standard deviations.



The numbers in the diagram below are 95% credible intervals for each of the features of the model; 95% of the posterior distribution is in the credible intervals for each of the features below. So, for example, in the diagram above, we can see the posterior distributions of the feature “intercept”. Now looking at the diagram below, we can see that the 95% credible interval for “intercept” is [-458.2, 603.5]. What this means is that about 95% of the posterior distribution for “intercept” is between the [-458.2, 603.5].

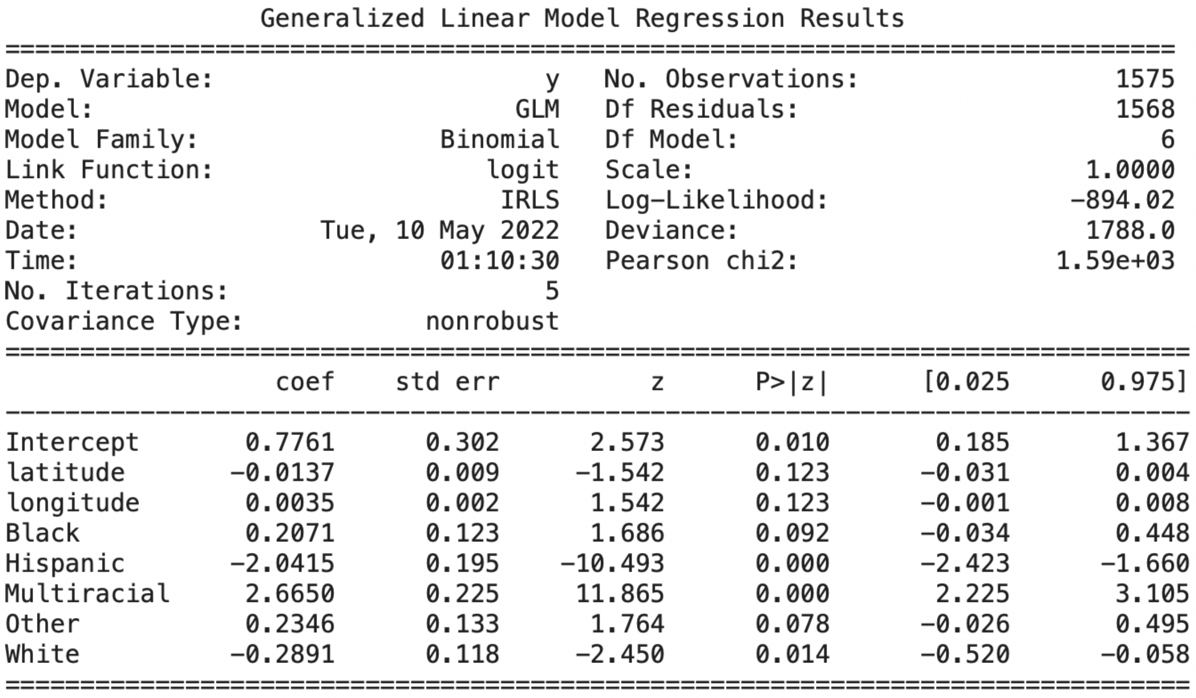


```
▶ Dimensions:        (hdi: 2)
▼ Coordinates:
    hdi              (hdi)    <U6  'lower' 'higher'
▼ Data variables:
    Intercept        (hdi)  float64  -458.2 603.5
    latitude         (hdi)  float64  -0.03036 0.001316
    longitude        (hdi)  float64  -0.0006784 0.007534
    Black            (hdi)  float64  -603.3 458.0
    Hispanic         (hdi)  float64  -605.5 455.8
    Multiracial      (hdi)  float64  -600.8 460.6
    Other            (hdi)  float64  -603.2 458.1
    White            (hdi)  float64  -603.7 457.5
▶ Attributes:   (0)
```

2. Frequentist

From the frequentist model, we were able to see that hispanic and white individuals are less likely to have COPD compared to other races, ceteris paribus (hispanic and white's coefficients are negative). Moreover, as the state's latitude is lower and its longitude is higher, individuals in that state are more likely than individuals in other states. The intercept of the model is 0.7761, which means that the model will provide an output of 0.7761, if all other variables are 0; however, it is not possible for all the variables to be 0 as all of the data in our dataset is one of the five races. Moreover, a longitude of 0 and latitude of 0 would indicate a location in the Atlantic Ocean, which is obviously not included in our data.

Moreover, the p-values of latitude, longitude, Black, and Other are greater than 0.05, meaning that they are not statistically significant at the 5% level. The variables that are statistically significant at the 5% level are Hispanic, Multiracial, and White.

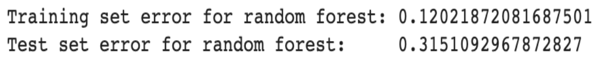
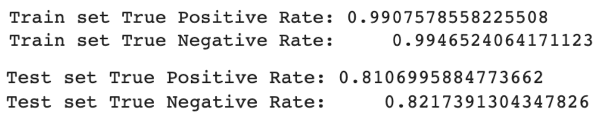


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 1575
Model:                            GLM   Df Residuals:                     1568
Model Family:                Binomial   Df Model:                            6
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -894.02
Date:                Tue, 10 May 2022   Deviance:                       1788.0
Time:                        01:10:30   Pearson chi2:                 1.59e+03
No. Iterations:                     5
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       0.7761      0.302      2.573      0.010       0.185       1.367
latitude       -0.0137      0.009     -1.542      0.123      -0.031       0.004
longitude       0.0035      0.002      1.542      0.123      -0.001       0.008
Black           0.2071      0.123      1.686      0.092      -0.034       0.448
Hispanic       -2.0415      0.195    -10.493      0.000      -2.423      -1.660
Multiracial     2.6650      0.225     11.865      0.000       2.225       3.105
Other           0.2346      0.133      1.764      0.078      -0.026       0.495
White          -0.2891      0.118     -2.450      0.014      -0.520      -0.058
===============================================================================
```

With the average log-likelihood being about -0.57, which is relatively close to 0, our model may be a good fit for the data. However, the value of pearson chi-squared, 1590, is really high meaning that our model may not be a good fit for the data.

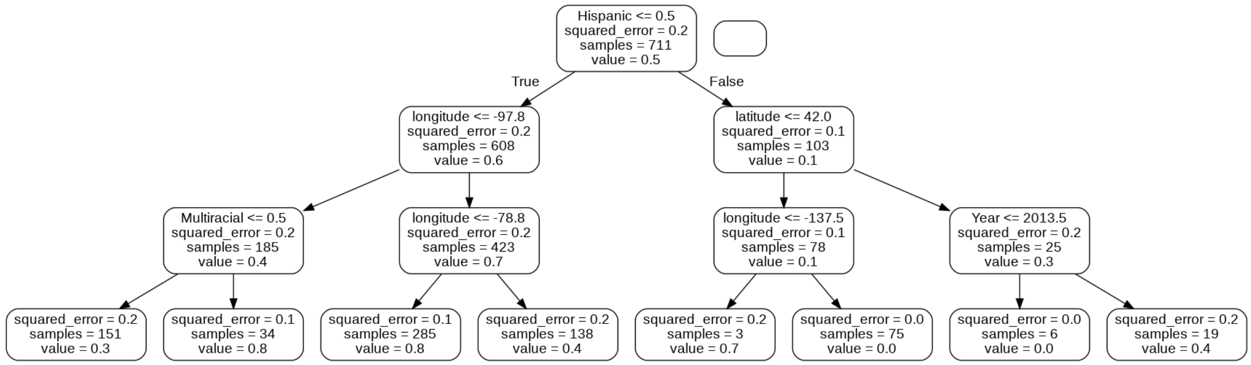
The numbers in the left two columns in the table are 95% confidence intervals for each of the features. The concept of confidence intervals is that the probability of the true value of the features' coefficient falling into the interval is about 95%. For example, the confidence interval of the feature “intercept” is [0.185, 1.367]. Thus, the probability of the true coefficient of “intercept” being in the interval of [0.185, 1.367] is 95%.

Random Forest:

Train/Test Error (RMSE)	Accuracy (TPR/TNR)
 Training set error for random forest: 0.12021872081687501 Test set error for random forest: 0.3151092967872827	 Train set True Positive Rate: 0.9907578558225508 Train set True Negative Rate: 0.9946524064171123 Test set True Positive Rate: 0.8106995884773662 Test set True Negative Rate: 0.8217391304347826

While both training and test set error are relatively low, test set error is slightly greater than training set error. To better understand the prediction, we computed the true positive rate and true negative rate for both the test and the train set. Here, we defined predictions greater than 60% as ‘predicting that one has COPD’. Similarly, we defined predictions less than 40% as ‘predicting that one does not have COPD’. TPR and TNR of our train set are approximately 99%, while that of our test set are approximately 81%. Overall, our model accuracy is high and our model error is also relatively low.

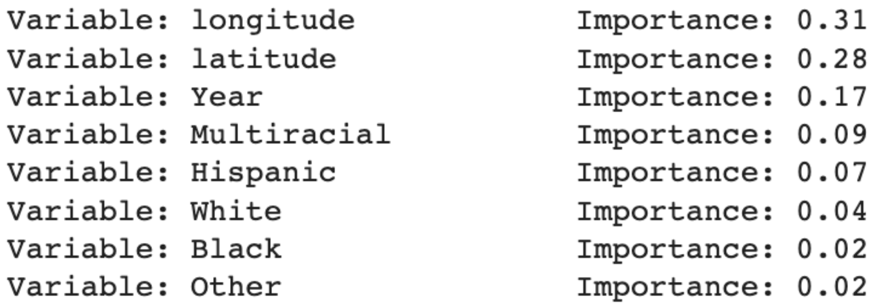
Visualization of a Single Decision Tree



As shown in the visualization of a single decision tree, the root node first looks at multiracial feature to split the data. Then, its children split after looking into longitude and latitude of the data. For instance, the rightmost branch predicts higher value when multiracial feature is greater than 0.5 and latitude is greater than 23.8. Similarly, when the latitude is less than or equal to 23.8, model predicts lower value. The only states that meet this condition (Guam, Hawaii, and Puerto Rico) are the states with lowest age-adjusted prevalence rate of COPD. This indicates that the model is incorporating geographical locations to accurately predict COPD.

Although this is a simple random forest model generated to only show a single decision tree, it seems like most of the layers make decisions based on longitude, latitude, multiracial, and year.

Computing Feature Importance



```
Variable: longitude            Importance: 0.31
Variable: latitude             Importance: 0.28
Variable: Year                 Importance: 0.17
Variable: Multiracial          Importance: 0.09
Variable: Hispanic             Importance: 0.07
Variable: White                Importance: 0.04
Variable: Black                Importance: 0.02
Variable: Other                Importance: 0.02
```

To further examine features and their role in prediction, we computed feature importances. As shown below, longitude is the best predictor of whether or not an individual will be diagnosed with COPD. Then follows latitude, Year, Multiracial, and more. This indicates that geographic location is the best predictor overall.

Discussion

We believe that our random forest model performed better than Frequentist/Bayesian logistic regression GLM. The Bayesian model overfits the provided data and the Frequentist model had several features that were not statistically significant at the 5% level. The random forest model is less prone to overfitting compared to logistic regression GLMs because random forest consists of many classifiers that are independently trained on different subsets of the training data. Random forest, as an ensemble model that uses feature bagging, is less likely to overfit because it has low variance. We are confident that applying the random forest model to future datasets is feasible because most of the current limitations of the random forest come from the limitations of the dataset. If future datasets have a more detailed and specified location and time data (i.e. counties with daily COPD data), we are confident that we could apply the random forest model.

<u>GLM:</u>

As discussed above, it seems like our Bayesian model overfits our data. Additionally, the Frequentist model had an average log-likelihood that was fairly close to 0; however, as its pearson chi^2 was very high, it seems like the model didn't do an excellent job of fitting to our data.

The true value of coefficients for latitude and longitude were similar in both models. For example, the mean of "latitude"'s posterior distribution was around -0.0143 in the Bayesian model and the Frequentist model's coefficient for "latitude" was around -0.0137. However, the remaining features had values that were very different in magnitude for each of the models. For example, the mean of "Black"'s posterior distribution was around -53.4 in the Bayesian model; however, the coefficient in the Frequentist model was around 0.2071.

Results from the Bayesian model suggest that it can be predicted that multiracial individuals are most likely to have COPD, given the same location. It also can be predicted that the hispanic individuals are least likely to have COPD compared to multiracial individuals. Lastly, it can be predicted that Black, White, Other individuals are less likely to have COPD than multiracial but more likely than hispanic individuals, given the same location. In addition, it can be predicted that individuals living in lower latitude and higher longitude are more likely to have COPD.

Results from the Frequentist model suggest that predicting COPD likelihood with longitude and latitude are not statistically significant at the 5% level. In addition, the difference in COPD likelihoods for Black and Other individuals are not statistically significant at the 5% level. However, the difference in COPD likelihoods for Hispanic, Multiracial, and White individuals are statistically significant at the 5% level. In addition, it can be predicted that individuals living in lower latitude and higher longitude are more likely to have COPD.

GLM has several inherent limitations. Firstly, it does not select features to incorporate into its model, but instead uses all the features provided. Taking all the features provided might be the reason why our model overfit to the data. Moreover, it has strict assumptions about the distributions' shapes. Finally, it is sensitive to outliers and have lower predictive power than nonparametric models.

Random Forest:

Results from the random forest model suggest that although we can predict whether people have COPD by using longitude and latitude data, we cannot predict such using racial features. This is because racial features had little to no feature importance (~0.02). This indicates that our model accuracy will sustain even if we exclude racial features and only predict using longitude and latitude data, and possibly year data as well.

One major limitation of random forest is that it cannot extrapolate. In other words, its prediction is only a mean of formerly observed labels. For instance, as shown in our EDA, the prevalence of COPD shows a yearly increasing trend up until 2017, and gradually decreases in the following years. Since the model cannot extrapolate, if the training data is missing year data, the model will likely under-predict or over-predict. For instance, New Jersey's COPD data is missing for 2019, when the overall state's COPD prevalence was decreasing. Likely, our model could not make accurate predictions in cases like this.

GLM/Random Forest:

The granularity of the provided COPD data was by state and years. If we had daily/monthly data by each county, it would have provided more sufficient data. Furthermore, as shown by the computation of feature importance, latitude, longitude and year are the three most important features. In other words, if we had more detailed geographical and time data, we might have observed a better accuracy. We might have been able to provide useful insights based on each county at a specific time frame.

Conclusion

GLM:

The key finding we were able to identify was that the results of Bayesian and Frequentist models produced different results for all the races; however, both Bayesian and Frequentist model produced very similar results for longitude and latitude. It can be predicted that individuals living in lower latitude and higher longitude are more likely to have COPD.

The results were very broad leading it to be very generalizable. In the Bayesian model, the standard deviations of all features related to race had very large standard deviations. Moreover, in the

Frequentist model, the majority of the values were not statistically significant at the 5% level, causing the findings to be very broad.

Random Forest:

Key findings we were able to identify was that we can predict whether people have COPD by using longitude and latitude data. As suggested by feature importance, we could not predict whether people have COPD by using racial features. They had little to no feature importance, which means that geographical data alone suffice to predict COPD. This also means that excluding racial features will have little to no impact to the accuracy of the model.

Using GLM and random forest, we could only determine that longitude and latitude have high feature importance thus relatively higher predictive power than features such as race. Thus, our finding from the random forest is way too broad thus generalizable. Although we couldn't identify specific relationships between a certain state and its COPD prevalence, our finding is broad enough to generalize it to the wider public and situation.

GLM/Random Forest:

Based on random forest results, we have identified that we can predict the onset of COPD from geographical location. Hence, medical institutions and the federal government could further investigate geographical attributes of different locations to predict areas that are most likely to be prone to COPD. Through prediction, they can plan medical interventions in these areas, such as periodic diagnostic assessment, recommendation of clinic visit, symptom checks, and more.

We imported external longitude and latitude data to merge with COPD data. This was because longitude/latitude data for some states were missing in COPD data. The benefit of combining different sources was that we could include all states' COPD and longitude/latitude data instead of dropping N/A rows/columns.

Data used for the models does not include any data points regarding air pollution. This is because we assumed that locations naturally incorporate different levels of air pollution in each state. We also included a year in our dataset and its feature importance (~ 0.17) was higher than that of race features; however, our dataset fails to recognize the yearly trends of air pollution that may differ by state and year. For instance, air pollutants in some states may have shown spikes in certain year due to factors we couldn't account for, such as wildfire, construction, factory pollutants, and more.

We could possibly further examine how location affects the onset of COPD. Using random forests, we've seen that location can predict COPD. Hence, we can use classification techniques (i.e. k-NN, hierarchical clustering) to first cluster locations based on common geographical attributes (i.e. rural/urban, climate, humidity, etc.). Then, we can see if we could identify relationships between a combination of geographical attributes and the prevalence of COPD or other respiratory diseases. For instance, future studies can test if residents in rural areas with dusty winds and/or factory air pollutants are more prone to respiratory diseases than those in urban areas.