Data Overview

Dataset: Chronic Disease and Air Quality

1. Daily Census Tract-Level PM2.5 Concentrations, 2011-2014 2. U.S. Chronic Disease Indicators: Cardiovascular Disease

CDC Ozone Levels Data (2011-14):
https://data.cdc.gov/Environmental-Health-Toxicology/Daily-Census-Tract-Level-Ozone-Concentrations-2011/372p-dx3h

Supplementary Data: Educational attainment for adults age 25 and older for the U.S., States, and counties, 1970–2020
https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

Both datasets are a census. The first dataset is a census tract-level dataset that showcases predictions of PM2.5 levels from the EPA's Downscaler model. For every day, there is an estimate of the mean PM2.5 concentration in each county in each state. The second dataset is a consensus that allows states and large metropolitan areas to uniformly define, collect, and report chronic disease data that are important to public health practice. Data from Alaska and Hawaii is excluded from dataset 1.

For dataset one, there were no participants that needed to be asked for permission as the data collection was done county wide. The granularity of the data is the PM2.5 concentration prediction for each county each day from 2011 to 2014. Each row represents each county's predicted PM2.5 concentration for a certain day. Since I was looking at data statewide in order to be comparable with dataset two, I will have to find a way to aggregate or average the predicted PM2.5 concentrations so that they can represent a state in a given year (2014). This dataset might have measurement error since it is predicting PM 2.5 concentration.

For dataset two, the data points are a sum of cases (participants) for each cardiovascular disease indicated, so there is no direct need for the participants to be aware of data collection. The granularity of the data is the crude rate for the purposes of hypothesis 1 which is cases per 1000 or cases per 100,000 for each cardiovascular disease indicator. The data has age-adjusted rates as well, but I will be only using crude rates for hypothesis one. Among the crude rate, the data is segmented into different race/ethnicities as well as gender. I will be averaging among all these crude rates in order to find a representative crude rate that can be used for hypothesis 1.

For the second hypothesis question, I want to discover whether there is an association between ozone concentration and higher education in each state. I will need this dataset in order to join ozone concentration data (each state) with corresponding data on education for that state.

Research Questions

Question 1: Does a high PM2.5 concentration cause an increase in Cardiovascular Disease indicators in the US? If so, which ones?

What decision(s) could be made by answering this question?
If we find that high PM2.5 concentrations cause an increase in Cardiovascular Disease indicators we could try to move the people who are more at risk of cardiovascular disease to areas with less PM2.5 concentration. We could also begin to highly recommend the use of air purifiers for these people to try to limit the concentration in their homes.

Why is the method that I will use is a good fit for the question:

I want to know if high concentrations of PM2.5 cause more cardiovascular disease indicators in the US so in order to determine this causality I should use causal inference. Causal inference will allow me to consider any confounding variables that may affect the results such as age, race, or gender.

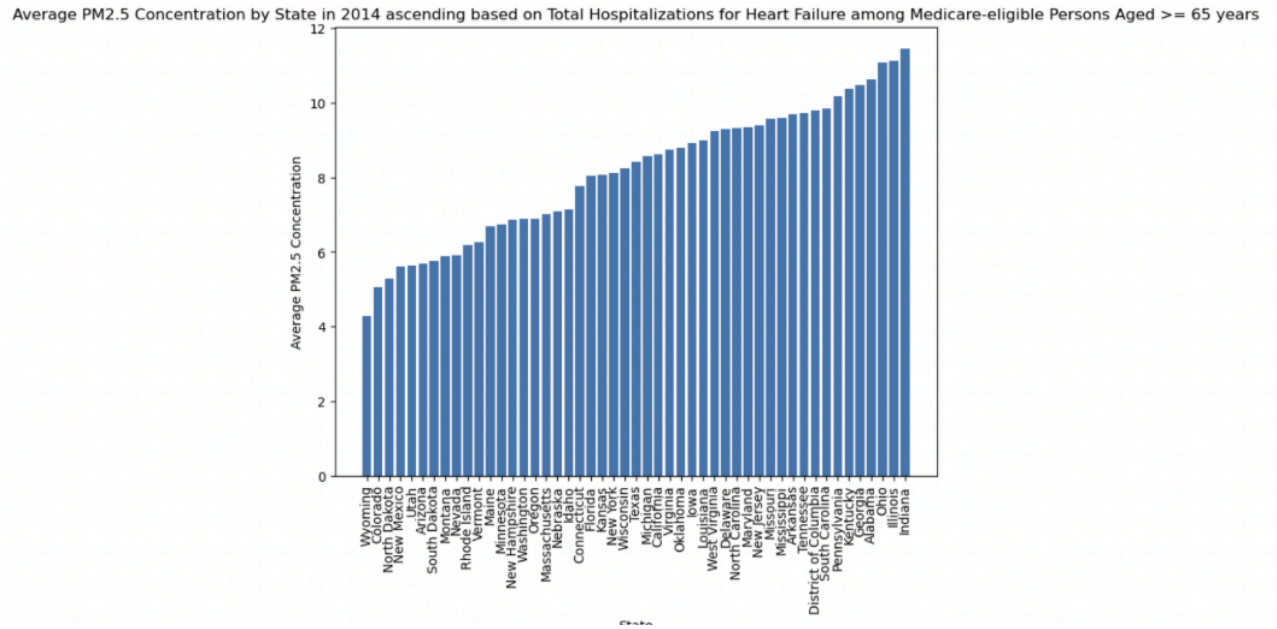
Question 2:For each state in the US, is the educational attainment for adults age 25 and older significantly associated with ozone concentration?

What decision(s) could be made by answering this question?
If we find a relationship between educational attainment for adults age 25 and older and ozone concentration we could use this to predict our future ozone concentration. With this information we can judge how dangerous our ozone concentration is getting and take action before it is too late.

Why is the method that I will use is a good fit for the question:

Since I am testing all the states individually, I have multiple datasets with the same question which leads me to multiple hypothesis testing. There is a risk of false positives since I am doing multiple tests on multiple datasets so in order to prevent these false positives I will use correction methods such as Bonferroni and Benjamini-Hochberg.

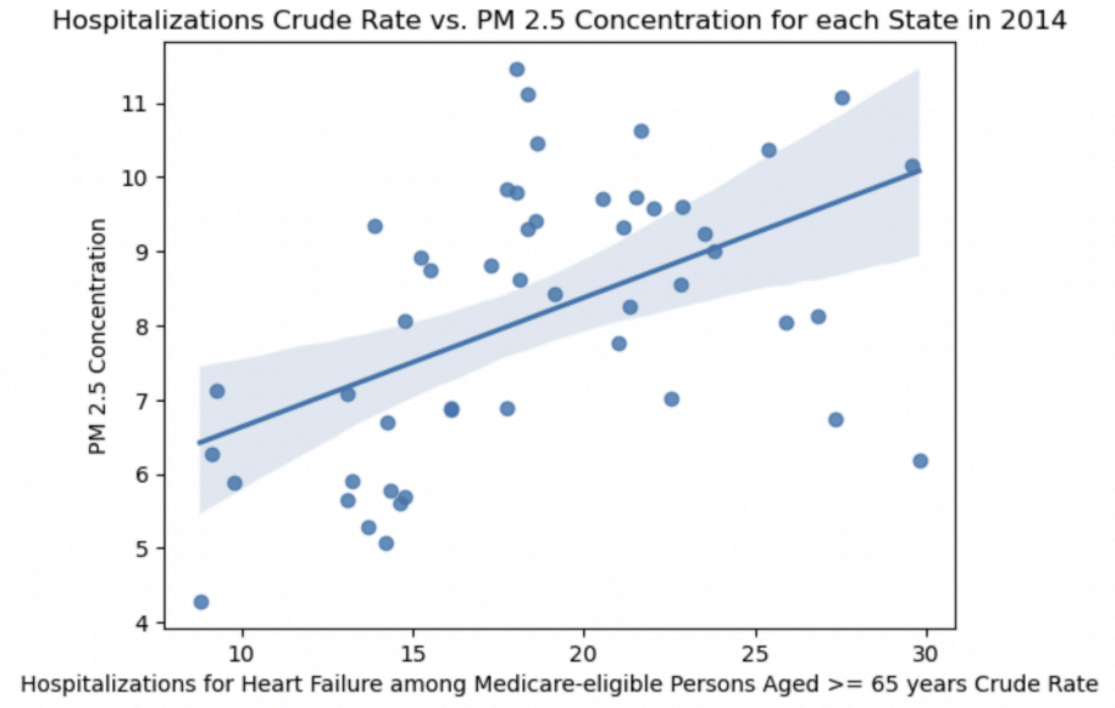
EDA



As the crude rate of hospitalizations for heart failure among Medicare-eligible persons aged above 65 increases, there is a simultaneous increase in average PM 2.5 concentration for the state. The states are ranked from left to right in ascending order based on crude rate of hospitalizations. There is a clear pattern showing that states with a higher rate of a cardiovascular disease indicator like the one mentioned tend to have higher average PM 2.5 concentration levels. Midwestern states have the highest average PM 2.5 concentration in 2014.

I cleaned data so that it only took into account the year 2014 and focused on only one cardiovascular indicator. I also averaged the PM 2.5 concentration for the entire year for each state and averaged the crude rate among all different ethnicities and genders. These decisions narrow the scope of the data and more than one indicator needs to be used in my model to prove causal inference.

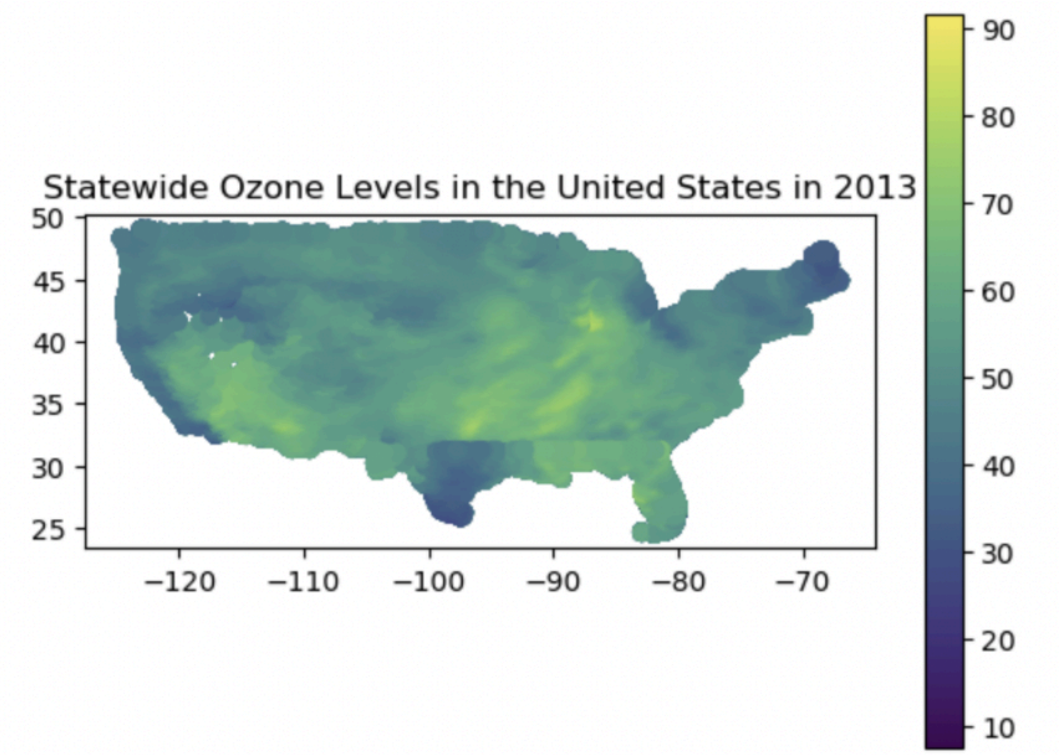
This shows a correlation with high hospitalizations and a high PM concentration which can be used to prove that there is some justification for exploring causal inference between the two.



This shows a somewhat high positive correlation between crude rate of hospitalizations for heart failure and PM 2.5 concentration. There are still certain outliers in the data where the crude rate of hospitalizations is high and the PM2.5 concentration is low and vice versa, so hypothesis 1 should be further explored to see if there is a casual relationship.

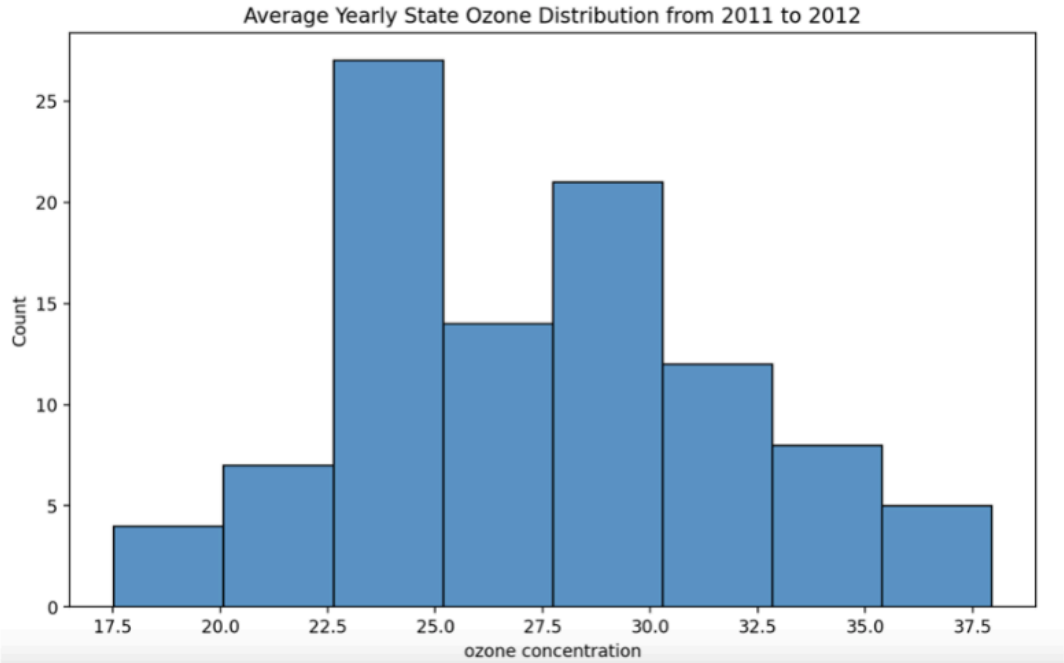
I cleaned data so that it only took into account the year 2014 and focused on only one cardiovascular indicator for now. I also averaged the PM 2.5 concentration for the entire year for each state and averaged the crude rate of the total cases. These decisions narrow the scope of the data and more than one indicator needs to be used in my model to prove causal inference. Additionally, to make the scatter plot more coherent, the x and y axis should be adjusted through functions (ex. Log, exp, etc).

This positive correlation shows that there is a justification for exploring causation between PM 2.5 concentration and cardiovascular disease indicators. This indicator was used as it was the most frequent in the dataset (13.9%).



The data for the Daily census ozone levels only consisted of the year 2013, so I decided to sort the data by date chronologically.

This visualization is relevant to the second research question since it visually shows the ozone levels of each state. The plot visualizes the average ozone levels on the coasts are much lower than Midwestern states, which average around 70-90. This can be further explored by comparing the visualization to the population estimates for associations between the ozone levels and the population estimate.



I wanted to observe how ozone distribution within the United States throughout the years. The data from the cdc consisted of daily ozone concentrations within 8 hour periods from 2011-2014. Unfortunately when downloading the data set I was only able to get the data form 2011-2012. One thing I observed was the fact that ozone concentrations from Hawaii and Alaska were not a part of the dataset, so I took that into account. In addition, I believe it is interesting to explore how the population of each state relates to these distributions.

I cleaned the data by removing any states that were not taken into account and averaging all the ozone concentrations by state and year. I also only focused on the two years that I was able to obtain from the data set.

This visualization is relevant to the second question because the visualization shows the average ozone levels across 48 states. Through the histogram we are able to see that more states have an average ozone concentration of 25. This could also potentially answer whether this distribution has any correlation to the variation of population within each state.

Causal Inference

Hypothesis 1: Does a high PM2.5 concentration cause an increase in Cardiovascular Disease indicators in the US? If so, which ones?

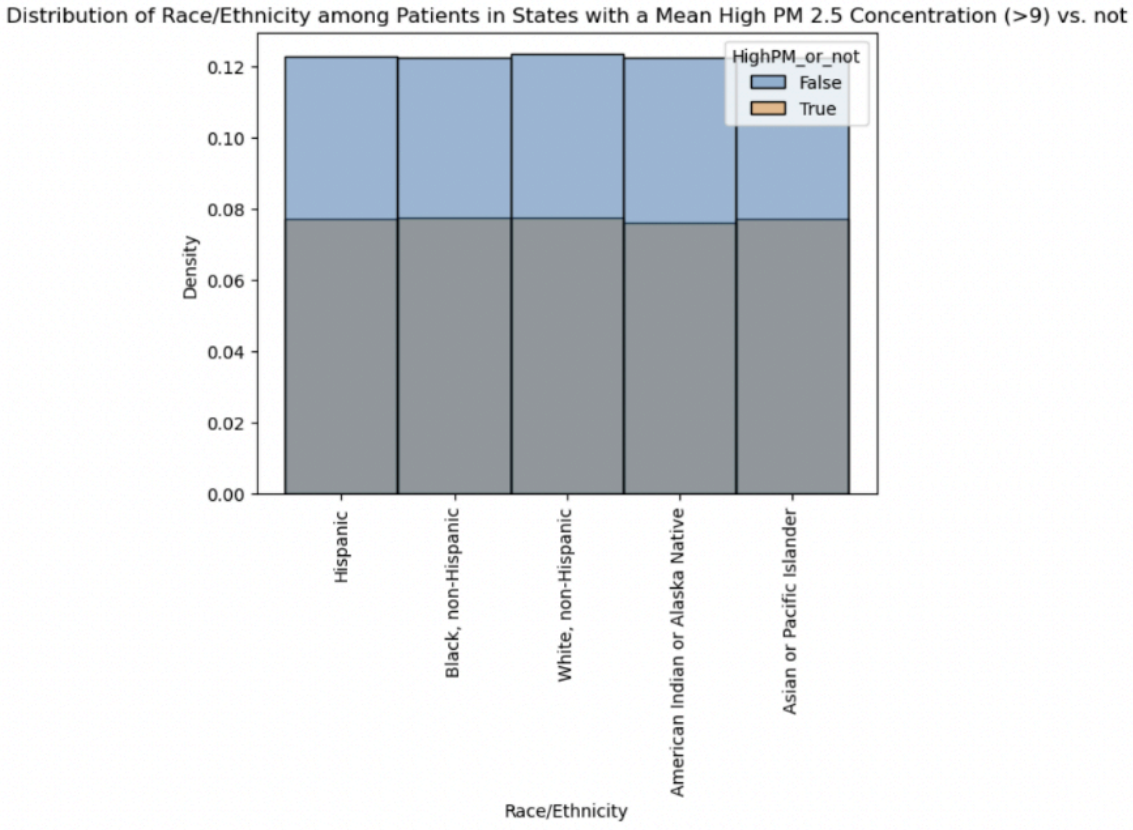
Technique:

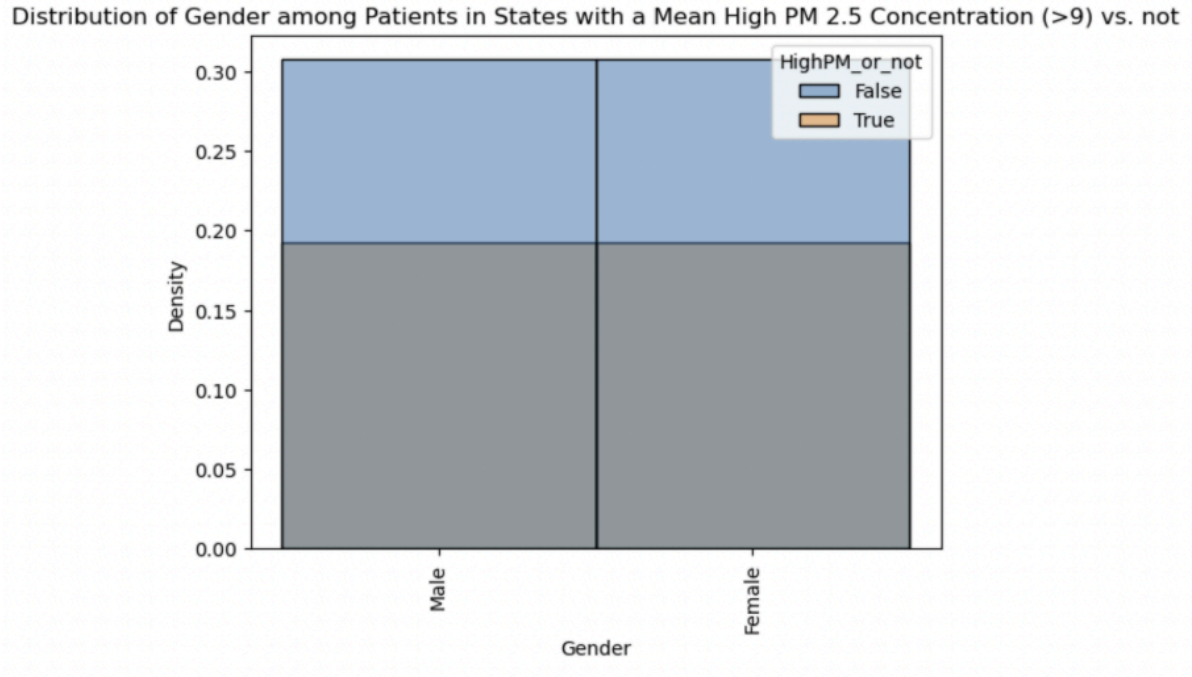
Treatment: High PM2.5 concentration (>= 9 μg/m3)

Outcome: Cardiovascular disease indicators Crude rate for the following indicators:

- *Hospitalizations for heart failure among Medicare-eligible persons aged >= 65 years (cases per 1000)*
- *Mortality from total cardiovascular diseases (cases per 100,000)*
- *Mortality from cerebrovascular disease (stroke) (cases per 100,000)*
- *Mortality from diseases of the heart (cases per 100,000)*
- *Mortality from coronary heart disease (cases per 100,000)*

Units: Each state divided by race and gender
Confounders: Ages, location, health history, race, gender
Instrumental Variables: none
This hypothesis deals with using causal inference for an observational study. To account for confounding variables, I will condition on any found confounding variables.





As seen by the two figures above, both race and gender have a similar distribution on all indicators present in the dataset for states with a high PM2.5 concentration vs. an average or low PM2.5 concentration. Therefore, these two factors are not confounding variables, and we do not see the need to condition on these in the OLS model. Additionally, age was not provided in the dataset rather there were two adjusted rates provided: crude rate and age-adjusted rate. I decided to use the crude rate as it accounts for another vital confounding variable: population. As each state has vastly different populations, accounting for population is important when considering the number of cases for each of the cardiovascular disease indicators. Health history information was not included, so I could not condition on that aspect. Overall, the unconfoundedness assumption holds.

The treatment variable is the PM 2.5 concentration. Treatment group would be a PM 2.5 concentration of above 9, while the baseline group would be for states with a PM 2.5 concentration 9 or below. Above 9 is considered high as the data for PM2.5 concentration for each state showed that there is a somewhat equal amount of data for states below this value and for states above this value. The observed values are the average crude rates for the 5 indicators for cardiovascular disease as listed above in the Technique section.

Results:

Indicator 1: Hospitalizations for heart failure among Medicare-eligible persons aged >= 65 years

OLS Regression Results

Dep. Variable:	DataValue	R-squared (uncentered):	0.940
Model:	OLS	Adj. R-squared (uncentered):	0.938
Method:	Least Squares	F-statistic:	745.7
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	6.76e-31
Time:	16:15:10	Log-Likelihood:	-145.89
No. Observations:	49	AIC:	293.8
Df Residuals:	48	BIC:	295.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pm25_mean_pred	2.2484	0.082	27.308	0.000	2.083	2.414

Omnibus:	10.080	Durbin-Watson:	1.616
Prob(Omnibus):	0.006	Jarque-Bera (JB):	9.857
Skew:	0.880	Prob(JB):	0.00724
Kurtosis:	4.315	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The results show that an increase of 1 in PM 2.5 concentration leads to around a 2.25 increase in the crude rate of hospitalization for heart failure among Medicare-eligible persons aged >= 65 years. The correlation between these two variables is quite strong in the model as shown by the R-squared value of .940. There is shown to be causality assuming factors like prior health history have little to no effect on this indicator of cardiovascular disease.

There is uncertainty with this estimate as calculating the mean difference of the crude rate for this indicator between states with a PM 2.5 concentration above 9 and those with concentration below or equal to 9 was around 4.11. This gap shows that the model slightly underestimates the causal effect of PM2.5 concentration on this indicator's crude rate. This shows that with high confidence that there is a positive causal relationship as predicted, but the extent to which a model can predict this causality is not very strong as the true difference between the means is different from the one predicted using OLS.

The biggest limitation of using OLS is that there is not a strictly linear relationship between PM2.5 concentration and the crude rate of the cardiovascular disease indicator. As seen by the scatterplot in the EDA section, there are several outliers that do not follow the linear pattern proven through this model. These data points can not be represented strongly with OLS. Data on prior health history of each participant would provide a greater understanding of whether there is a true causal relationship between the PM2.5 concentration in the environment one lives in with the chances of them developing risk of cardiovascular disease.

Indicator 2: Mortality from total cardiovascular diseases

OLS Regression Results

Dep. Variable:	DataValue	R-squared (uncentered):	0.941
Model:	OLS	Adj. R-squared (uncentered):	0.940
Method:	Least Squares	F-statistic:	762.8
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	4.05e-31
Time:	16:16:03	Log-Likelihood:	-262.94
No. Observations:	49	AIC:	527.9
Df Residuals:	48	BIC:	529.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pm25_mean_pred	24.7889	0.898	27.620	0.000	22.984	26.593

Omnibus:	3.876	Durbin-Watson:	2.080
Prob(Omnibus):	0.144	Jarque-Bera (JB):	3.712
Skew:	0.652	Prob(JB):	0.156
Kurtosis:	2.659	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The results show that an increase of 1 in PM 2.5 concentration leads to around a 24.78 increase in the crude rate of mortality from total cardiovascular diseases. The correlation between these two variables is quite strong in the model as shown by the R-squared value of .941. There is shown to be causality assuming factors like prior health history have little to no effect on this indicator of cardiovascular disease.

There is uncertainty with this estimate as calculating the mean difference of the crude rate for this indicator between states with a PM 2.5 concentration above 9 and those with concentration below or equal to 9 was around 20.28. This gap shows that the model slightly underestimates the causal effect of PM2.5 concentration on this indicator's crude rate. This shows that with high confidence that there is a high positive causal relationship as predicted, but the extent to which a

model can predict this causality is not very strong as the true difference between the means is different from the one predicted using OLS.

The biggest limitation of using OLS is that there is not a strictly linear relationship between PM2.5 concentration and the crude rate of the cardiovascular disease indicator. As seen by the scatterplots in the Jupyter notebook, there are several outliers that do not follow the linear pattern proven through this model. These data points can not be represented strongly with OLS. Data on prior health history of each participant would provide a greater understanding of whether there is a true causal relationship between the PM2.5 concentration in the environment one lives in with the chances of them developing risk of cardiovascular disease.

Indicator 3: Mortality from cerebrovascular disease (stroke)

OLS Regression Results

Dep. Variable:	DataValue	R-squared (uncentered):	0.941
Model:	OLS	Adj. R-squared (uncentered):	0.940
Method:	Least Squares	F-statistic:	769.5
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	3.33e-31
Time:	16:16:37	Log-Likelihood:	-179.94
No. Observations:	49	AIC:	361.9
Df Residuals:	48	BIC:	363.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pm25_mean_pred	4.5758	0.165	27.739	0.000	4.244	4.907

Omnibus:	1.345	Durbin-Watson:	2.062
Prob(Omnibus):	0.510	Jarque-Bera (JB):	1.234
Skew:	0.240	Prob(JB):	0.539
Kurtosis:	2.388	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The results show that an increase of 1 in PM 2.5 concentration leads to around a 4.58 increase in the crude rate of mortality from cerebrovascular disease (stroke). The correlation between these two variables is quite strong in the model as shown by the R-squared value of .941. There is shown to be causality assuming factors like prior health history have little to no effect on this indicator of cardiovascular disease. 0.2861011640168624

There is uncertainty with this estimate as calculating the mean difference of the crude rate for this indicator between states with a PM 2.5 concentration above 9 and those with concentration below or equal to 9 was around 5.77. This gap shows that the model slightly underestimates the causal effect of PM2.5 concentration on this indicator's crude rate. This shows that with high confidence that there is a positive causal relationship as predicted, but the extent to which a model can predict this causality is not very strong as the true difference between the means is different from the one predicted using OLS.

Indicator 4: Mortality from diseases of the heart

OLS Regression Results

Dep. Variable:	DataValue	R-squared (uncentered):	0.939
Model:	OLS	Adj. R-squared (uncentered):	0.937
Method:	Least Squares	F-statistic:	733.9
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	9.68e-31
Time:	16:16:58	Log-Likelihood:	-250.81
No. Observations:	49	AIC:	503.6
Df Residuals:	48	BIC:	505.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pm25_mean_pred	18.9799	0.701	27.091	0.000	17.571	20.389

Omnibus:	3.005	Durbin-Watson:	2.038
Prob(Omnibus):	0.223	Jarque-Bera (JB):	2.836
Skew:	0.532	Prob(JB):	0.242
Kurtosis:	2.493	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The results show that an increase of 1 in PM 2.5 concentration leads to around a 18.98 increase in the crude rate of Mortality from diseases of the heart. The correlation between these two variables is quite strong in the model as shown by the R-squared value of .939. There is shown to be causality assuming factors like prior health history have little to no effect on this indicator of cardiovascular disease.

There is uncertainty with this estimate as calculating the mean difference of the crude rate for this indicator between states with a PM 2.5 concentration above 9 and those with concentration below or equal to 9 was around 16.95. This gap shows that the model slightly overestimates the

causal effect of PM2.5 concentration on this indicator's crude rate. This shows that with high confidence that there is a positive causal relationship as predicted, but the extent to which a model can predict this causality is not very strong as the true difference between the means is different from the one predicted using OLS.

Indicator 5: Mortality from coronary heart disease

OLS Regression Results

Dep. Variable:	DataValue	R-squared (uncentered):	0.920
Model:	OLS	Adj. R-squared (uncentered):	0.919
Method:	Least Squares	F-statistic:	554.5
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	5.09e-28
Time:	16:17:14	Log-Likelihood:	-233.22
No. Observations:	49	AIC:	468.4
Df Residuals:	48	BIC:	470.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
pm25_mean_pred	11.5238	0.489	23.548	0.000	10.540	12.508

Omnibus:	1.101	Durbin-Watson:	2.158
Prob(Omnibus):	0.577	Jarque-Bera (JB):	1.124
Skew:	0.262	Prob(JB):	0.570
Kurtosis:	2.474	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The results show that an increase of 1 in PM 2.5 concentration leads to around a 11.52 increase in the crude rate of hospitalization for mortality from coronary heart disease. The correlation between these two variables is quite strong in the model as shown by the R-squared value of .920. There is shown to be causality assuming factors like prior health history have little to no effect on this indicator of cardiovascular disease.

There is uncertainty with this estimate as calculating the mean difference of the crude rate for this indicator between states with a PM 2.5 concentration above 9 and those with concentration below or equal to 9 was around 12.11. This gap shows that the model slightly underestimates the causal effect of PM2.5 concentration on this indicator's crude rate. This shows that with high confidence that there is a positive causal relationship as predicted, but the extent to which a model can predict this causality is not very strong as the true difference between the means is different from the one predicted using OLS.

Multiple Hypothesis Testing

Hypothesis 2: For each state in the US, is the educational attainment for adults age 25 and older significantly associated with ozone concentration?

Technique:

H0: the the educational attainment for adults age 25 and older is not significantly associated with ozone concentration
HA: the the educational attainment for adults age 25 and older is significantly associated with ozone concentration

I will test my hypothesis by using a two-tailed test. I will set the p value to 0.05, but I can correct for these tests using Bonferroni and Benjamini-Hochberg.

Methods:

It makes sense to test many hypotheses since I want to test each state. I will be testing each hypothesis with A/B testing using the percentage of adults with a bachelor's degree or higher from 2007-2011 as one of my variables and the ozone concentration as the other. Counties with a percent higher than 20 are considered highly educated. Some states had to be left out because the ozone dataset did not have their information. I will be using Bonferroni correction to control the Family Wise Error Rate(FWER) which is the probability that any test is a false positive. I will also use the Benjamini-Hochberg procedure to control the False Discovery Rate(FDR) which is the expected proportion of discoveries that were wrong.

Results:

Once I perform the hypothesis tests on all the states that I have data for I get a list of p-values [0.835, 0.3866, 0.3006, 0.9578, 1.0, 0.2488, 0.7826, 0.519, 0.9998, 0.9796, 0.9826, 0.014, 0.785, 0.5104, 0.9784, 0.6604, 0.1042, 0.997, 0.7084, 0.9724, 0.9978, 0.1754, 0.0764, 1.0, 0.6852, 0.4852, 0.3286, 0.8296, 0.245, 1.0, 0.4192, 0.9928, 0.9994, 0.5038, 0.9752, 0.335, 0.228, 0.9856, 0.9468, 0.9782, 0.9886, 0.8036, 0.4588, 0.3416].

With these p-values I can first use Bonferroni correction to control the family wise error rate which is the probability that any test is a false positive. If I want this rate to be less than 0.05 and

I have performed 44 tests, the threshold should be 0.05 / 44 = .001. Since the smallest p-value I have is 0.014 none of these tests are significant.

Now I can perform Benjamini-Hochberg to control the false discovery rate which is the expected proportion of discoveries that were wrong. If I sort the p-values and plot the line k * a/m where a = 0.05 and m = 44 I get this graph.

Discussion

Through the graph, we are able to see that there are no p-values that are under the line which means none of the p-values that I have found are significant. From these results I can conclude that there is no significant association between the education level of a county and their ozone concentration levels.

From the individual tests I will fail to reject the null hypothesis.

One of the limitations faced during the analysis was finding categorical variables that were considered confounding variables within the dataset. Originally I wanted to compare gender and race to significant ozone concentrations but there was not enough information within the dataset as these categories were grouped together and there were not enough points recorded.

If given more data, additional tests were discussing whether there is an association between ozone concentrations and climate of each state. Additionally it would be interesting to test whether there is an association between ozone concentrations and rural areas within a state, as many rural areas within California experience ozone being carried downwind from urban communities and truck routes driving through rural areas.

Conclusion

There is a positive causal relationship between PM2.5 concentration and cardiovascular disease indicators. These results are generalizable to statewide data, but it is also important to look at the risk of each individual in the future as things like health history are not accounted for. Additionally, only five indicators were tested, so the findings are narrow in terms of which cardiovascular risk factors have a positive causal relationship with PM2.5 concentration.

My results are generalizable to other similar environmental issues and cardiovascular disease. My second hypothesis did not really have any findings and probably could not be generalized.

Based on my results, a good call to action would be to reduce PM 2.5 by encouraging people to reduce their use of fossil fuels. This could be by encouraging people to drive less and use public transportation more often. It could also be beneficial to use clean energy sources.

For the second hypothesis I had to merge the additional dataset with the provided ozone dataset. This merge allowed us to compare the two dataset but it also caused us to lose some data from the additional dataset since some counties were missing from the ozone data which meant I needed to remove some rows.

Most of the data that I used for my analysis was from around 2011 to 2014. Since the data is quite old there may have been some changes that I cannot account for in my analysis which could make my results less accurate when applied today. There was also some missing data in my second hypothesis which could have affected my results since some states were excluded.

For my first hypothesis there could be future studies on what kind of environmental factors increase the amount of cardiovascular disease indicators. Identifying these factors could result in longer life spans if I am able to manage my environments, especially since cardiovascular disease is the leading cause of death in the United States.