DSCI 510 Final Report:

How is the Concentration of PM2.5 associated with asthmatic attacks in the United States?

Siying Yang

Department of Computer Science, University of Southern California

DSCI 510: Principles of Programming for Data Science

Professor Ulf Hermjakob

May 5, 2023

Abstract

The project involved scraping FIPS code data and collecting PM2.5 concentration data using APIs and Wikipedia. Emergency department visit data for asthma were downloaded as CSV files. The three datasets were combined for analysis, including descriptive results, linear regressions, and visualizations.

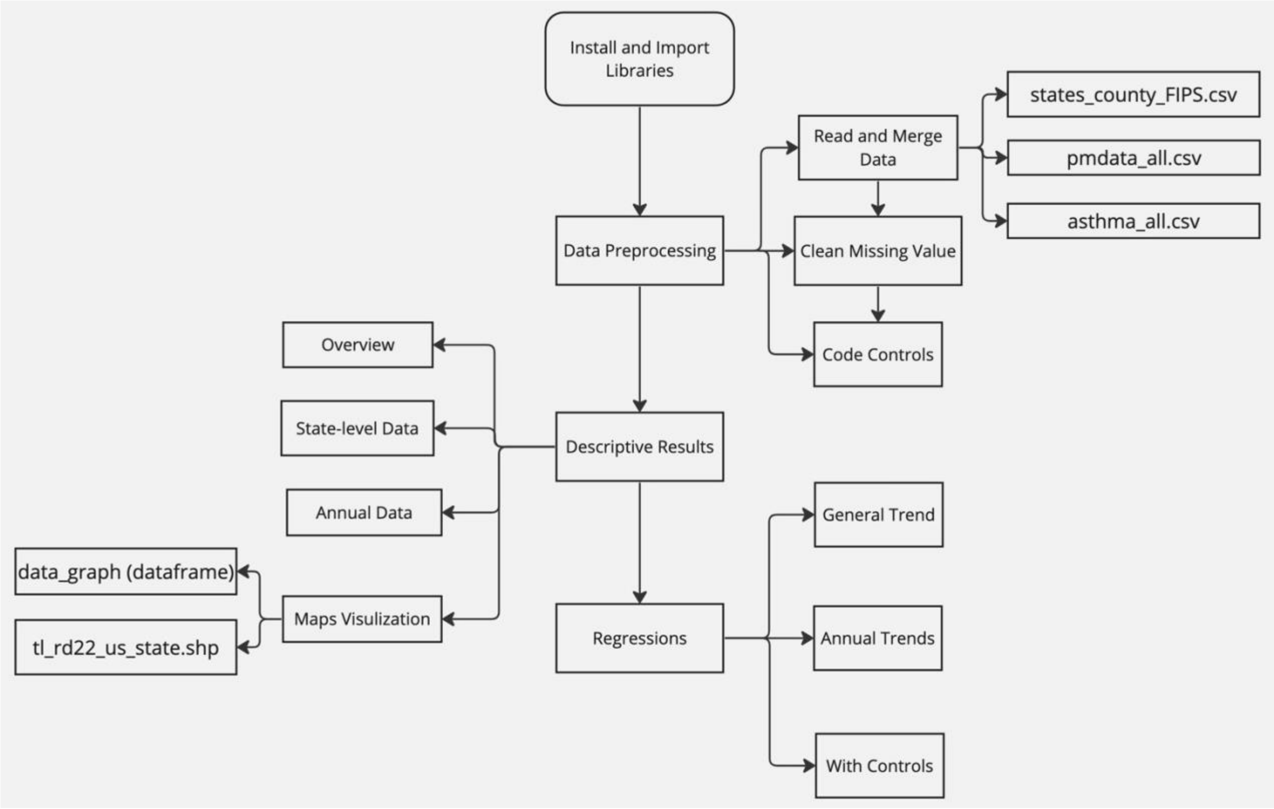
The regression results showed a positive association between PM2.5 concentration and emergency department visits for asthma across the United States. However, when examining annual trends, only the data from 2017 exhibited a statistically significant regression. Control variables were considered, including regions, economic status, and population density. However, the economic status control variable was found to be statistically insignificant in annual trends, while the other two variables were only significant in 2017. This suggests that these variables did not significantly impact the outcome variable once the influence of PM2.5 concentration was accounted for.

Motivation

The project focuses on investigating the correlational relationship between PM2.5 concentration and emergency department visits for asthma, considering PM2.5 as a major threat to public health from anthropogenic sources. Due to the ethical and practical limitations of conducting experimental studies, the project aims to gather data from nationwide air quality monitors and statistical departments to explore the association between PM2.5 concentration and asthma-related emergency department visits. The data collected from legal and ethical sources serves as a substitute for controlled experiments. The project aims to gather data from as many states as possible to provide a comprehensive understanding of the correlational relationship between PM2.5 concentration and asthmatic attacks across the United States.

Research Question: How is the concentration of PM2.5 associated with asthmatic attacks across the United States?

Figure 1. *Overview of Work System*



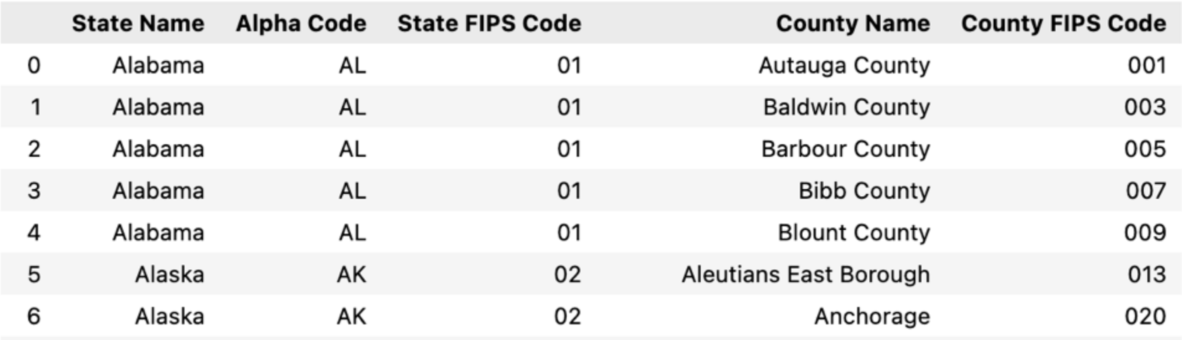
Data Source

FIPS code from Wikipedia

website: https://en.wikipedia.org/wiki/Federal_Information_Processing_Standard_state_code

The FIPS code dataset was collected for accessing PM2.5 data and drawing dynamic geographic maps as they share the same code and county name values from the Wikipedia page "Federal Information Processing Standard state code". The Python program outputs a CSV file with the state and county FIPS codes with a total row number of 3093.

Figure 2. *Sample output for FPIS code dataset*



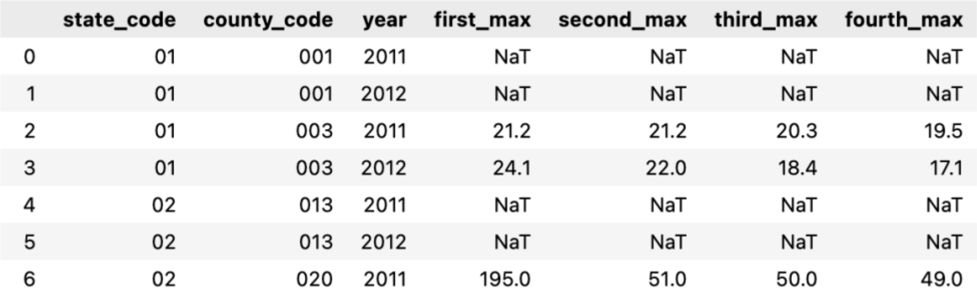
	State Name	Alpha Code	State FIPS Code	County Name	County FIPS Code
0	Alabama	AL	01	Autauga County	001
1	Alabama	AL	01	Baldwin County	003
2	Alabama	AL	01	Barbour County	005
3	Alabama	AL	01	Bibb County	007
4	Alabama	AL	01	Blount County	009
5	Alaska	AK	02	Aleutians East Borough	013
6	Alaska	AK	02	Anchorage	020

PM2.5 concentration across Counties

website: https://aqs.epa.gov/aqsweb/documents/data_api.html#annual

The dataset was collected by querying the United States Environmental Protection Agency (EPA) Air Quality System (AQS) to collect annual PM 2.5 concentration data for counties across multiple states in the United States. Data only included states that were available from 2011 to 2019, which covered the most in dataset 3. The Python program outputs a CSV file with each county's maximum annual PM 2.5 concentration in the given time range with a total row number of 2364.

Figure 3. *Sample output for PM2.5 concentration*



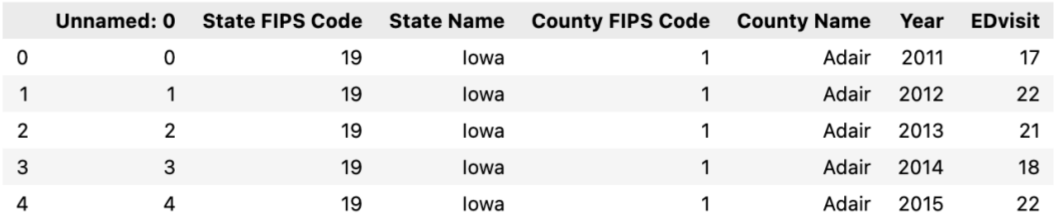
	state_code	county_code	year	first_max	second_max	third_max	fourth_max
0	01	001	2011	NaT	NaT	NaT	NaT
1	01	001	2012	NaT	NaT	NaT	NaT
2	01	003	2011	21.2	21.2	20.3	19.5
3	01	003	2012	24.1	22.0	18.4	17.1
4	02	013	2011	NaT	NaT	NaT	NaT
5	02	013	2012	NaT	NaT	NaT	NaT
6	02	020	2011	195.0	51.0	50.0	49.0

Emergency department visits for asthma

website: https://ephtracking.cdc.gov/DataExplorer/?c=3&i=90&m=-1

The dataset was directly downloaded from the website with a total row number of 10821.

Figure 4. *Sample output for emergency department visits for asthma*



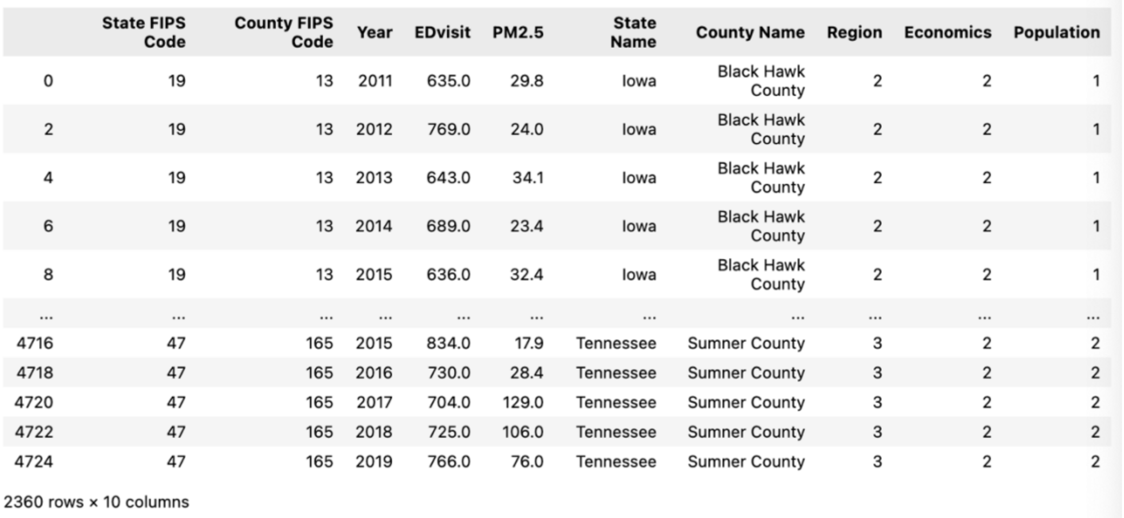
	Unnamed: 0	State FIPS Code	State Name	County FIPS Code	County Name	Year	EDvisit
0	0	19	Iowa	1	Adair	2011	17
1	1	19	Iowa	1	Adair	2012	22
2	2	19	Iowa	1	Adair	2013	21
3	3	19	Iowa	1	Adair	2014	18
4	4	19	Iowa	1	Adair	2015	22

Overview of merged dataset

Read all three datasets and dropped redundant columns. Then, merged the datasets based on FPIS code and year, and dropped duplicate columns. Next, dropped missing value in emergency department visits.

Then, three columns of control variables were categorized by the newest statistical data and available state data for geographic regions, economic status, and population density (See detailed categories in Jupyter Notebook).

Figure 5. *Sample output for cleaned data*



	State FIPS Code	County FIPS Code	Year	EDvisit	PM2.5	State Name	County Name	Region	Economics	Population
0	19	13	2011	635.0	29.8	Iowa	Black Hawk County	2	2	1
2	19	13	2012	769.0	24.0	Iowa	Black Hawk County	2	2	1
4	19	13	2013	643.0	34.1	Iowa	Black Hawk County	2	2	1
6	19	13	2014	689.0	23.4	Iowa	Black Hawk County	2	2	1
8	19	13	2015	636.0	32.4	Iowa	Black Hawk County	2	2	1
...	...	...	...	...	...	...	...	...	...	...
4716	47	165	2015	834.0	17.9	Tennessee	Sumner County	3	2	2
4718	47	165	2016	730.0	28.4	Tennessee	Sumner County	3	2	2
4720	47	165	2017	704.0	129.0	Tennessee	Sumner County	3	2	2
4722	47	165	2018	725.0	106.0	Tennessee	Sumner County	3	2	2
4724	47	165	2019	766.0	76.0	Tennessee	Sumner County	3	2	2

2360 rows × 10 columns

Visualizations in Descriptive Results

Notes:

All visualizations are available in Jupyter Notebook, including tables, charts, and regression results. To save pages here, I only include necessary tables and maps.

General Descriptive Results

The average PM2.5 concentration is relatively high at approximately 72.91 ug/m^3. However, according to a quite large standard deviation of 95.31, we could tell that the concentration varied significantly greatly across the states, with a minimum of 4.3 ug/m^3 and a maximum of 1167 ug/m^3. Meanwhile, the median concentration (50%) is 40.00 ug/m^3, suggesting that a significant portion of the observations fall below this level. Similarly, the average number of ED visits for asthma is around 2252.46, with a standard deviation of 5181.73, indicating a wide range of variability in different states. In detail, the minimum number of ED visits recorded is 6, while the maximum is as high as 51895. The median number of ED visits (50%) is 799.50, suggesting that half of the observations have a lower number of ED visits.

Table 1. *Overview Descriptive Results*

	Count	Mean	SD	Min	25%	50%	75%	Max
PM2.5 Concentration (ug/m^3)	2360	72.91	95.31	4.3	23.78	40.00	84.18	1167.00
ED Visits for Asthma	2360	2552.46	5181.73	6	265.25	799.50	2835.25	51895

Among those with high population density states, the range of PM2.5 concentration is around 61.11 to 65.96 ug/m^3. And the ED visits range is between 1793 o 3798 visits. Among those states with a medium population density, the range of PM2.5 concentration is around 33.23 to 156.45 ug/m^3, as well as the range of ED visits of 709 to 5474 visits. Among those with low population density states, the PM2.5 concentration range is around 37.73 to 131.94 ug/m^3. And the range of ED visits is around 274 o 3779 visits.

Overall, while high population density states exhibit relatively consistent PM2.5 concentrations, there is still some variation in ED visits. Besides, medium and low-population-density states show wider ranges of both PM2.5 concentration and ED visits with a relatively lower boundary in both PM2.5 concentration and ED visits, indicating more significant differences in air quality and healthcare utilization for asthma.

Table 2. *Overview Descriptive Results Across States*

State	M (PM)	SD (PM)	Min (PM)	Max (PM)	M (ED)	SD (ED)	Min (ED)	Max (ED)	Population Density
Iowa	37.2	46.2	16.3	461.0	440	526	8	2353	1
Vermont	37.7	22.5	14.1	92.9	274	134	99	477	1
Connecticut	61.1	49.0	12.3	235.0	3779	2353	447	7461	1
New Mexico	71.2	83.3	7.3	402.0	761	659	70	2365	1
Utah	105.9	126.8	9.8	900.5	603	839	17	3408	1
Maine	131.9	187.1	12.7	870.0	489	276	96	1167	1
Arizona	156.5	198.7	4.3	1167.0	2796	5406	44	18957	1
Louisiana	28.0	16.8	9.9	135.0	1204	1029	64	3863	2
New York	33.2	32.1	8.7	329.5	9427	12032	122	42454	2
New Jersey	38.4	17.9	15.6	104.9	3140	2170	156	11081	2
Florida	42.6	49.5	13.1	448.0	5474	4397	439	18624	2
Tennessee	45.7	47.5	14.3	297.0	1409	2403	92	10422	2
Kentucky	47.9	67.2	12.2	593.0	709	1390	49	6822	2
Wisconsin	52.7	83.6	8.9	901.0	922	2038	14	9757	2
North Carolina	53.0	52.8	12.3	356.5	1155	1421	10	8429	2
Kansas	65.8	77.1	18.2	357.6	1151	1173	6	3596	2
Minnesota	95.3	66.8	11.0	353.0	918	1510	7	6784	2
California	122.4	127.9	5.5	985.0	3889	7587	43	51895	2
Missouri	125.0	68.3	37.2	326.2	2377	2724	29	8945	2
Massachusetts	65.3	57.1	12.2	253.7	3798	2192	371	7433	3
Rhode Island	66.0	59.6	24.0	287.0	1793	1806	259	4631	3

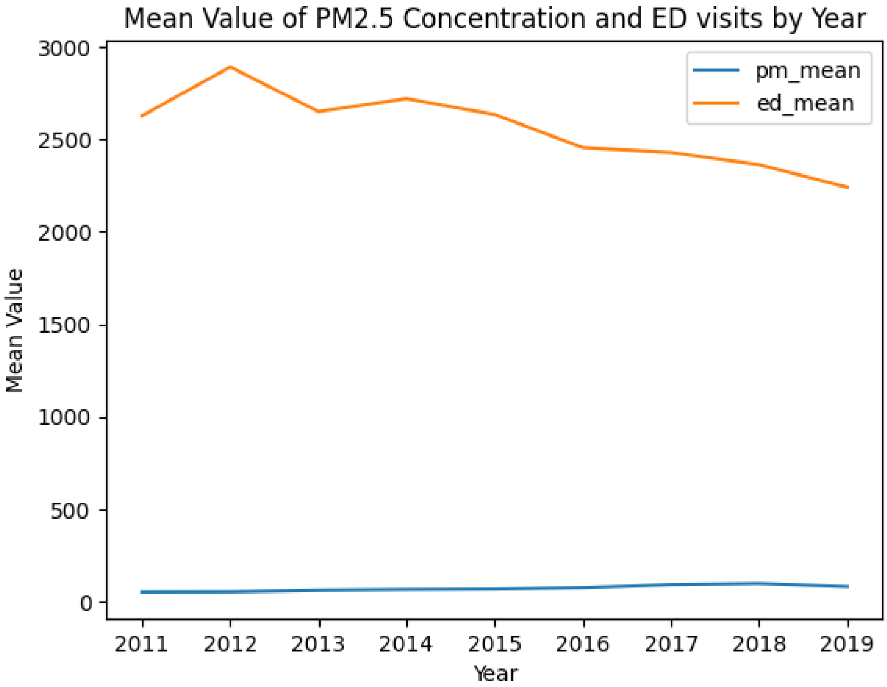
Annual Descriptive Results

Including analysis of outliners in both PM2.5 concentration and ED visits, we could further confirm that the data variability is quite large. The average PM2.5 concentration ranged from 52.56 ug/m^3 in 2011 to 99.02 ug/m^3 in 2018, indicating climbing variations in air pollution levels. However, considering the average number of ED visits for asthma, it dropped in general with a bit of fluctuation in 2012 and 2014. This suggests that despite the rising pollution levels, there might have been improvements in asthma management or other factors contributing to the reduction in asthma-related emergency visits.

Table 3. *Annual descriptive results of PM2.5 concentration and ED visits for asthma*

Year	M (PM)	SD (PM)	Min (PM)	Max (PM)	M (ED)	SD (ED)	Min (ED)	Max (ED)
2011	52.6	68.7	4.6	768.4	2626	5312	11	47510
2012	53.4	81.3	7.3	801.0	2890	5674	6	49391
2013	62.0	96.3	9.7	914.0	2649	5457	8	49732
2014	66.2	116.9	7.3	1167.0	2718	5479	7	50771
2015	68.5	88.1	6.5	985.0	2633	5389	19	51895
2016	75.9	93.1	8.7	804.0	2454	5141	8	50479
2017	92.9	111.4	5.5	918.0	2428	4909	7	49376
2018	99.0	100.9	4.3	870.0	2362	4711	8	46612
2019	82.4	77.3	4.3	727.0	2240	4539	9	46725

Figure 6. *Mean value line chart for PM2.5 concentration and ED visits for asthma by year*



Annual Data by States

Combined with the above analysis, I drew two maps to show the dynamic trends from 2011 to 2019 across the United States.

Figure 7. *Annual trends for PM2.5 concentration on a geographic map (click to view if any error in figure)*

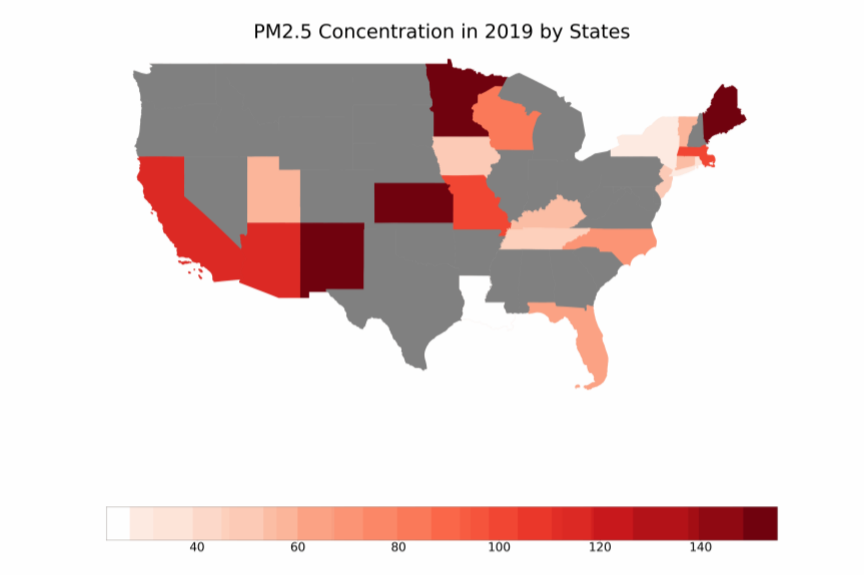
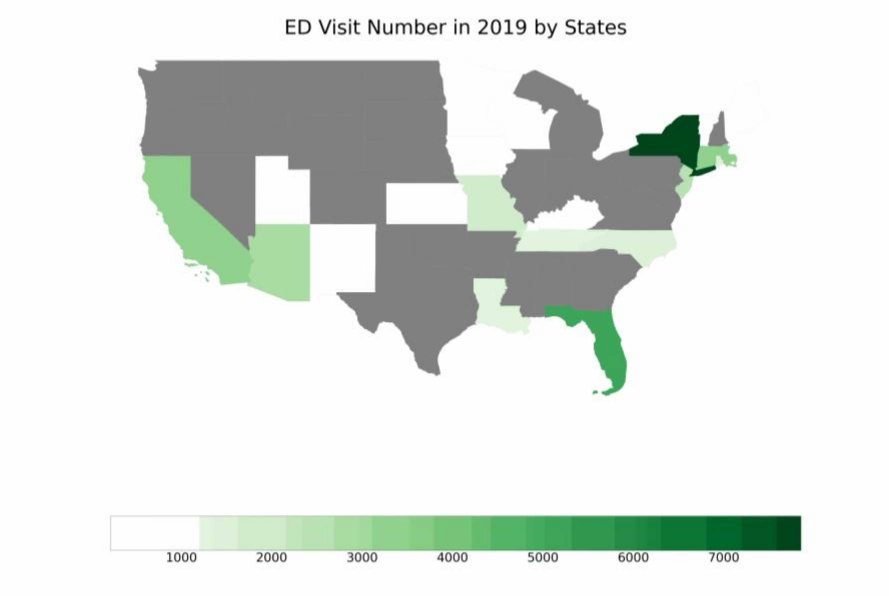


Figure 8. *Annual trends for ED visits on a geographic map (click to view if any error in figure)*



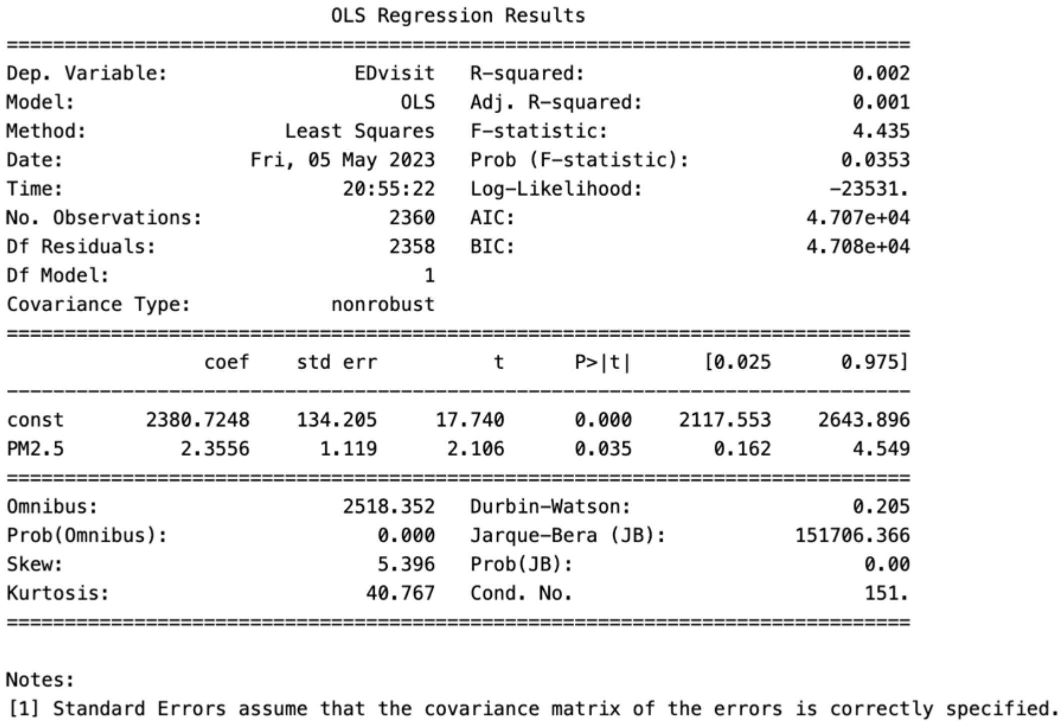
Analysis, Insights and Relative Limitations

General Trend

Analysis and insights. The regression analysis results indicate a weak relationship between the PM2.5 concentration and the number of ED visits for asthma. The R-squared value of 0.002 suggests that only 0.2% of the variation in ED visits can be explained by changes in PM2.5 concentration. The coefficient for PM2.5 is 2.3556, indicating that for every unit increase in PM2.5, there is an expected increase of approximately 2.36 ED visits. Though we could tell it is significant in statistical results, it is important to note that the significance level is relatively low, indicating that the relationship may not be practically significant or strong. These findings have important implications for addressing asthma-related healthcare utilization, as the previous studies valued how to reduce the discharge of PM2.5. While efforts to reduce air pollution, including PM2.5, remain crucial for overall public health, focusing solely on improving air quality may not significantly reduce asthma-related ED visits.

Limitations. Additionally, outliners could potentially influence the relationship between PM2.5 concentration and ED visits. Other factors not included in the model may have a more significant impact on asthma-related ED visits.

Figure 9. *OLS regression results for the general trend*



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                EDvisit   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     4.435
Date:                Fri, 05 May 2023   Prob (F-statistic):             0.0353
Time:                        20:55:22   Log-Likelihood:                -23531.
No. Observations:                2360   AIC:                         4.707e+04
Df Residuals:                    2358   BIC:                         4.708e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2380.7248    134.205     17.740      0.000    2117.553    2643.896
PM2.5          2.3556      1.119      2.106      0.035       0.162       4.549
==============================================================================
Omnibus:                     2518.352   Durbin-Watson:                   0.205
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           151706.366
Skew:                           5.396   Prob(JB):                         0.00
Kurtosis:                      40.767   Cond. No.                         151.
==============================================================================

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

Annual Trend

Analysis and insights. The analysis shows that the slope is not statistically significant for most years, except for 2017, where there is a significant positive relationship between the year and ED visits. This suggests that, on average, there was a significant increase in ED visits for asthma in 2017 compared to the other years. However, the effect size of the slope is relatively small, indicating a limited impact of the year on ED visits, which is in line with the previous findings in general trends. While there may have been a slight increase in ED visits in 2017, the overall impact of the year on ED visits is not substantial. It suggests that other factors, such as individual asthma management, healthcare access, and environmental conditions, may significantly influence ED visits for asthma.

Table 4. *Regression summary for annual trends*

Year	Intercept	Slope	p-value	Significance
2011	2324.98	5.72	0.26	FALSE
2012	2773.89	2.17	0.63	FALSE
2013	2741.64	-1.49	0.67	FALSE
2014	2481.80	3.56	0.21	FALSE
2015	2602.85	0.44	0.91	FALSE
2016	2319.57	1.77	0.60	FALSE
2017	1931.69	5.34	0.05	TRUE
2018	2033.27	3.32	0.25	FALSE
2019	1971.39	3.27	0.37	FALSE

Annual Trends with Control Variables

Analysis and insights in regions as control. The analysis shows that the slope coefficient for regions is not statistically significant for any of the years except for 2017. This suggests that the regional classification does not significantly impact asthma-related ED visits. Therefore, the variation in ED visits cannot be solely attributed to the different regions in the dataset.

Analysis and insights into economic status as control. Similarly, the slope coefficient for economic status is not statistically significant for any of the years. This indicates that economic status does not significantly influence the number of ED visits for asthma. Other factors, such as healthcare access, individual asthma management, or environmental conditions, may have a more prominent role in explaining variations in ED visits.

Analysis and insights into population density as control. For population density, the analysis reveals a significant positive relationship between population density and ED visits in 2017. This suggests that higher population density is associated with a higher number of ED visits for asthma in that particular year. However, the relationship is not statistically significant for the other years. This implies that while population density may play a role in asthma-related ED visits in certain circumstances, its impact is inconsistent across all years.

Limitations. Firstly, the analysis only establishes associations between variables and does not imply causality. Other unmeasured factors could influence the relationship between PM2.5 concentration and ED visits for asthma. Additionally, the quality of the data used in the analysis can impact the accuracy and reliability of the results. Moreover, the model includes a limited number of control variables, and there may be other important variables that were not considered, such as individual health behaviors or environmental factors, which could influence the outcome.

Notes:

Line charts for slopes in regressions with control variables are similar to the one above. Therefore the report would not include the charts. Please see the Jupyter Notebook.

Table 5. *Regression with Control variables*

	Year	Intercept	Slope	p-value	Significant
Regions					
	2011	4343.707	8.485	0.097	FALSE
	2012	4369.460	4.269	0.350	FALSE
	2013	3694.983	-0.152	0.967	FALSE
	2014	3538.551	4.411	0.132	FALSE
	2015	3177.307	0.888	0.815	FALSE
	2016	2744.171	1.962	0.565	FALSE
	2017	2344.361	5.649	0.041	TRUE
	2018	2627.751	4.217	0.163	FALSE
	2019	2242.766	3.417	0.347	FALSE
Economic Status					
	2011	-4301.553	2.904	0.545	FALSE
	2012	-3327.845	-0.025	0.995	FALSE
	2013	-3117.258	-1.907	0.568	FALSE
	2014	-3317.275	3.341	0.225	FALSE
	2015	-3054.150	-0.904	0.802	FALSE
	2016	-2862.967	2.963	0.367	FALSE
	2017	-2644.157	4.401	0.094	FALSE
	2018	-2380.777	3.208	0.248	FALSE
	2019	-2392.739	3.444	0.323	FALSE
Population Density					
	2011	-453.832	5.051	0.315	FALSE
	2012	-114.602	2.059	0.641	FALSE
	2013	128.107	-0.762	0.827	FALSE
	2014	-443.119	4.482	0.119	FALSE
	2015	224.507	0.510	0.891	FALSE
	2016	-146.947	2.748	0.420	FALSE
	2017	-554.057	5.692	0.035	TRUE
	2018	-469.451	4.004	0.162	FALSE
	2019	-148.087	3.336	0.353	FALSE

Conclusion

In conclusion, the regression analysis and the one with control variables indicate that neither alone PM2.5 concentration nor controls as regions, economic status, and population density have a limited impact on asthma-related ED visits. The findings emphasize the need for a comprehensive approach to asthma management that goes beyond these factors. By focusing on preventive measures, healthcare access, and addressing individual and environmental factors, healthcare systems can significantly reduce asthma-related ED visits and improve overall asthma outcomes.

Technical Solution

Overview of Algorithm:

The algorithm consists of data collection, preprocessing, integration, regression analysis, and data visualization.

Data collection and preprocessing phase. Relevant datasets are collected from the internet using web scraping, downloading, and API access. The datasets include PM2.5 concentration, emergency department visit data, FIPS codes for states and counties, and control variables obtained from online sources. The collected datasets are structured into a CSV format compatible with analysis.

Data integration. Involves combining the collected datasets into a single dataset and matching the data based on the 'Year', 'State Name', and 'County Name' variables. Data cleaning procedures are performed, including dropping redundant columns, handling missing values, and addressing outliers. Control variables are coded using Python functions.

Regression analysis. They were conducted using the statsmodels library in Python. Initially, a simple linear regression model is fitted to the overall dataset to identify general trends, with PM2.5 concentration as the independent variable and ED visits as the dependent variable. Subsequently, each year, linear regression is performed separately to analyze annual trends. To incorporate control variables, separate regression analyses are conducted by adding control variables (such as region, economics, and population) to the model. The results of the regression, including intercepts, slopes, and p-values, are stored in separate dataframes.

Data visualization. Types included line plots, tables, and maps. Descriptive statistics are calculated for the 'EDvisit' and 'PM2.5' variables, summarizing the general dataset, annual trends, and state-level perspective. Line charts are used to visualize the annual trends in the mean value of PM2.5 concentration and ED visits. The trends of the slopes over the years are also displayed for annual trends and regression analyses with control variables. Maps are generated by integrating geospatial data with the matplotlib library to create GIFs, showcasing the data for each year and state.

Challenges and Solutions

Data Collection. It is hard to find available datasets for PM2.5 concentration from the past years as most APIs were available for the past 48 hours or in recent years. Also, the data across the states are large and hard to estimate. To gather as much data as I could, I first collected FIPS code and directly accessed through API for 9-year data for 6 hours in total. Then, I cleaned the missing value to determine which state would be included in the analysis and revised the project plan for feasibility.

Statistical Analysis and Visualization. Conducting regression analysis is hard, especially when it comes to a state-level perspective, as there are 21 states in total for 9-year data. To make it easier to analyze and clear in a readable study, I didn’t consider the state-level analysis and only applied general and annual trend analyses instead. Also, I provided two maps to show the trends across the states as a supplement. However, I also came across difficulty in generalizing GIFs, as the map is too wide in coordinates. So, I abandoned some of the regions in the United States and only maintained the mainland part. Besides, the legend size is also hard for me because I used the Geopandas library instead of a pure Matplotlib library, whose parameters were different.