ANALYSIS OF AIR QUALITY AND ASTHMA

Final Report

By

Mary Robinette

Florida Institute of Technology

CIS 5898 – Projects in Computer Information Systems

for

Dr. Bernard Parenteau

December 10, 2021

Air quality affects everyone. Some individuals and segments of the population are more sensitive to poor air quality than others. According to the U.S. Environmental Protection Agency (EPA), air pollution can trigger asthma attacks and can make asthma symptoms worse.

Asthma is a chronic, non-communicable respiratory disease characterized by wheezing, coughing, shortness of breath, and chest tightness. It can be life threating.

Two key pollutants affecting asthma are ozone and particle pollution (EPA). Ozone pollution is found in smog, with ozone levels typically higher on hot summer days. Particle pollution is found in dust, smoke, and haze, with levels independent of season.

Indications of air pollution can sometimes, but not always, be observed. In the U.S., the EPA's tool for communicating air quality to the public is the U.S. Air Quality Index (AQI). The AQI covers five major pollutants—ozone, particle pollution, carbon monoxide, nitrogen dioxide, and sulfur dioxide (Air Now, -b). These pollutants, regulated by the Clean Air Act, require monitoring by state, local, and tribal air pollution control agencies (EPA, 2019).

AQI values run from 0 to 500, with higher values associated with greater levels of air pollution. The AQI is divided into six categories, corresponding to levels of concern: good (0-50), moderate (51-100), unhealthy for sensitive groups (101-150), unhealthy (151-200), very unhealthy (201-300), and hazardous (301 and higher), (Air Now, -a).

The EPA's Air Quality System (AQS) is the repository for air pollution data. The data can be used to assess air quality. I used the data to analyze and visualize air quality data in the U.S. and visualize the relationship between air quality and asthma.

Algorithms/Project Solution

The first part of my project is based on Roger Peng's data analysis: Changes in Fine Particle Air Pollution in the U.S. (Peng, 2020). This study described changes in outdoor air pollution in the U.S. between 1999 and 2012, focused on one pollutant—PM2.5. PM2.5 refers to particulate matter that is 2.5 microns or less in diameter. The author used the R programming language to analyze and visualize the EPA raw text file datasets of PM2.5 for years 1999 and 2012. The study hypothesis was that the levels of PM2.5 in the U.S. would have decreased from 1999 (the first year of required PM2.5 monitoring) to 2012 (the most current dataset at the time of the study publication), ostensibly because of passage of the Clean Air Act. I do not have a copy of the original datasets used by Dr. Peng.

According to the EPA's (2019) About AQS data:

> Historical data can change at any time. Many quality assurance review processes are made on an entire year's worth of data, so it might not be until the middle of this year until the final review and changes have been made to last year's data by a submitter. Also, historical monitoring or calculation methods may be found to be problematic and require that older data be changed. Finally, there is no "versioning" or freezing of data in the Data Mart, so if other people may need the data exactly as it was retrieved to verify or continue an analysis, the user must preserve a copy.

The EPA presently provides pre-generated downloadable CSV files of pollutants at the site https://aqs.epa.gov/aqsweb/airdata/download_files.html. Sample (raw) data is accessed via the AQS API https://aqs.epa.gov/aqsweb/documents/data_api.html#meta and requires

registration. As the data has likely changed, I could not expect my retrievals to correspond exactly with Dr. Peng's. Thus, I used the EPA's downloadable CSV datasets for PM2.5 from 1999 and 2012 rather than raw files. I uploaded the datasets to the project GitHub repository, preserving them in their retrieved form. I used Python to analyze the datasets to see how my results compared to Dr. Peng's.

After the initial comparison review, I repeated the analysis for changes in PM2.5 between 1990 and 2020. Basic questions and visualizations included (1) How does the level of PM2.5 in the U.S. compare between 1999, 2012, and 2020? (2) Which states have the highest and lowest levels of PM 2.5?

After completing the PM2.5 analyses, I downloaded and analyzed the ozone datasets for years 1990, 2012, and 2020. The datasets were uploaded to the project GitHub repository in their retrieved form. Basic questions and visualizations included (1) How does the level of ozone in the U.S. compare between 1999, 2012, and 2020? (2) Which states have the highest and lowest levels of ozone?

Finally, I downloaded the annual AQI by county for years 1990, 2012, and 2020. The datasets were uploaded to the project GitHub repository in their retrieved form. Basic questions and visualizations included (1) Which states have reports of hazardous air quality index (AQI)? (2) Which states have reports of very unhealthy AQI? (3) Which states have reports of unhealthy AQI? (4) Which states have reports of unhealthy for sensitive individuals AQI? (4) What does a map of AQI look like?

In the second part of my project, I analyzed the relationship between air quality and asthma. My primary dataset for asthma came from the Centers for Disease Control and

Prevention. Asthma prevalence data comes from the Behavioral Risk Factor Surveillance System (BRFSS), telephone surveys of U.S. residents concerning their health status. From this U.S. data, states having high reported incidence of asthma were identified and datasets were located for Vermont and Kentucky. Basic questions included (1) What does the reported incidence of asthma look like by state? (2) Which states have the highest and lowest reported incidence of asthma? (3) What are ways to visualize the relationship between air quality and asthma? Finally, I identified California as a state with high AQI reports and repeated analyses for PM2.5, ozone, AQI, and asthma for years 2017-2018.

Implementation

Technologies and resources used to implement the project:

- The project was implemented on an HP OMEN laptop PC with a 64-bit operating system and x64-based processor running Microsoft Windows 10 Home Edition, Version 20H2.
- Anaconda open-source Individual Edition (www.anaconda.com) for Windows was downloaded and installed. The Anaconda installation contained Python 3.8.8, the programming language that was used along with open-source Python libraries NumPy, Pandas, and Matplotlib for data analysis and visualization. NumPy (numpy.org) enables numerical computing with Python. The numpy version used was 1.21.2. Pandas (pandas.pydata.org) is a data analysis and manipulation tool. The pandas version used was 1.3.4. Matplotlib (www.matplotlib.org) is used to create publication quality visualizations. The matplotlib version used was 3.4. Seaborn (seaborn.pydata.org) is a modern-looking python data visualization library based on matplotlib. The version used

was 0.11.2. Geopandas (geopandas.org/en/stable/) is a library used to visualize geospatial data. The version used was 0.10.2 . The Anaconda installation also contained Jupyter Notebooks 6.3.0, which is an interactive computing notebook environment. Jupyter Notebooks was used to run Python code and to create readable documents.

- Microsoft Office 365 was used for creating Microsoft Word and Excel documents used in the project.
- GitHub (www.github.com) was used as a repository for documenting and posting project results.
- The primary data sources for this project were air quality datasets from the United States Environmental Protection Agency (EPA) Air Quality System (AQS) database. Specifically, daily summary data for ozone and particulate matter (PM) 2.5 were analyzed. Separate datasets for each of these components are available by year at https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily. Annual summary data for AQI by County was also analyzed. AQI datasets are available at https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual.
- The second part of the project looked at an asthma dataset from the Centers for Disease Control and Prevention (CDC). This dataset is available at https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Asthma/us8e-ubyj. Additional asthma datasets from Vermont, Kentucky, and California state public health departments were examined. These specific states were found to have high prevalence of asthma and high air quality indices, reported as AQI.

Schedule

TASK	WEEK	START	END
Phase 1 - Research and Proposal Writing			
Select project topic	1	8/23/21	8/29/21
Finalize project/scope	2	8/30/21	9/5/21
DELIVERABLE - project proposal	3	9/6/21	9/12/21
Phase 2 - Design and Implementation			
Set up GitHub Repository for Project	4	9/13/21	9/14/21
Find/download air quality (AQ) datasets	4	9/13/21	9/19/21
Exploratory analysis (EA) of PM2.5 datasets (1999, 2012)	5	9/20/21	9/22/21
Compare results of PM2.5 analyses with case study	5	9/23/21	9/24/21
DELIVERABLE - First Progress Report	5	9/25/21	9/26/21
Update GitHub Repository	5	9/26/21	9/26/21
EA of PM2.5 2020 dataset	6	9/27/21	9/29/21
Create preliminary PM2.5 visualizations	6	9/30/21	10/2/21
EA of Ozone datasets (1999, 2012, 2020)	7	10/4/21	10/8/21
DELIVERABLE - Progress Report 2	7	10/9/21	10/10/21
Update GitHub Repository	7	10/10/21	10/10/21
Clean PM2.5 datasets (2012, 2020)	8	10/11/21	10/14/21
Create preliminary PM2.5 visualizations	8	10/14/21	10/17/21
Clean Ozone datasets and perform EA	9	10/18/21	10/20/21

Create preliminary Ozone visualizations	9	10/20/21	10/21/21
EA of AQI datasets (1999, 2012, 2020)	9	10/21/21	10/22/21
DELIVERABLE - Post-Midterm Progress Report	9	10/23/21	10/24/21
Update GitHub Repository	9	10/24/21	10/24/21
Find/download U.S. asthma dataset	10	10/25/21	10/27/21
Clean U.S. asthma datasets	10	10/27/21	10/29/21
Analysis of U.S. asthma dataset	10	10/30/21	10/31/21
Find/download asthma datasets for Vermont, Kentucky	11	11/1/21	11/6/21
DELIVERABLE - Second Half Progress Report 2	11	11/6/21	11/7/21
Update GitHub Repository	11	11/7/21	11/7/21
Clean state asthma datasets	12	11/8/21	11/11/21
Create preliminary asthma visualizations	13	11/15/21	11/19/21
DELIVERABLE - Last Interim Progress Report	13	11/21/21	11/21/21
Update GitHub Repository	13	11/21/21	11/21/21
Updated PM2.5, Ozone, AQI, and asthma visualizations	14	11/22/21	11/23/21
Plan for/standardize joining of AQI and asthma datasets	14	11/23/21	11/23/21
Join AQ and asthma datasets	14	11/23/21	11/23/21
Analysis of combined datasets	14	11/24/21	11/24/21
Create preliminary AQ and asthma visualizations	14	11/26/21	11/28/21
Download and clean 2011 AQI datasets for states	14	11/26/21	11/28/21

Download and clean California AQ datasets	14	11/26/21	11/28/21
DELIVERABLE - Final Progress Report	14	11/28/21	11/28/21
Update GitHub Repository	14	11/28/21	11/28/21
Download and clean California asthma dataset	15	11/29/21	11/29/21
Join California asthma and AQI datasets	15	11/30/21	11/30/21
Complete analyses for asthma and AQ studies	15	12/1/21	12/3/21
Finalize all visualizations	15	12/3/21	12/5/21
Phase 3 - Final Report/Presentation			
Create Final Project	16	12/6/21	12/6/21
Add Lessons Learned Section	16	12/6/21	12/6/21
Add Appendices	16	12/6/21	12/6/21
Finalize Project Report and Materials	16	12/6/21	12/10/21
Update GitHub Repository	16	12/10/21	12/10/21
Submit DELIVERABLE - Final Project Report & Materials	16	12/10/21	12/10/21

Results

The eight Jupyter notebooks for this project contain code, summaries, and details on cleaning of data. The primary objective of the Python code throughout the study was to answer the target questions in a clear and thorough manner. Code efficiency was not the goal. There may be faster or more complete methods than the code I used. The most relevant visualizations are included here.

PM2.5 Studies

Target questions: How does the level of PM2.5 in the U.S. compare between 1999 and 2012? Can the results of the case study (Peng, 2020) be reproduced? On average, the level of PM2.5 in the U.S. has decreased from 1999 to 2012. There is more variability in the daily levels of PM2.5 in 2012 than there was in 1999. There are many more outliers in the data in 2012 than in 1999. The boxplot in the case study (Fig 1) compares favorably with the boxplot in Fig 2.

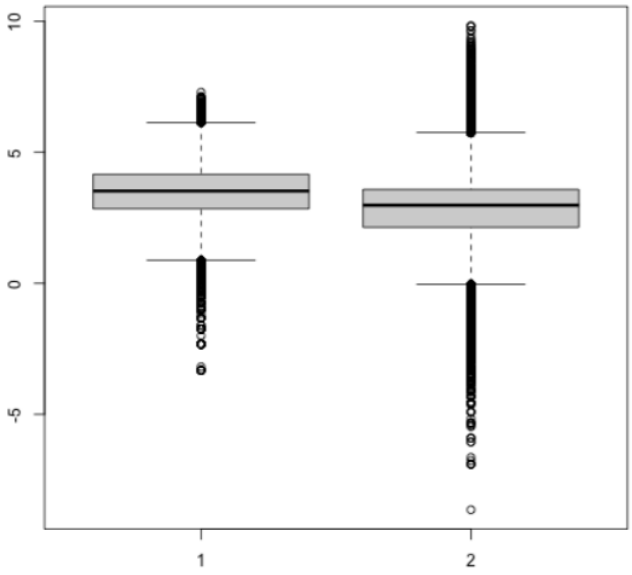

Figure 1 Average PM2.5 levels for years 1999 and 2012. Source: Peng (2020), p.151

Most states had decreased levels of PM2.5 in 2012 compared to levels of PM2.5 in 1999 (Figs 3-4). The slope chart in the case study (Fig 3) compares favorably to the slope chart in Fig 4.

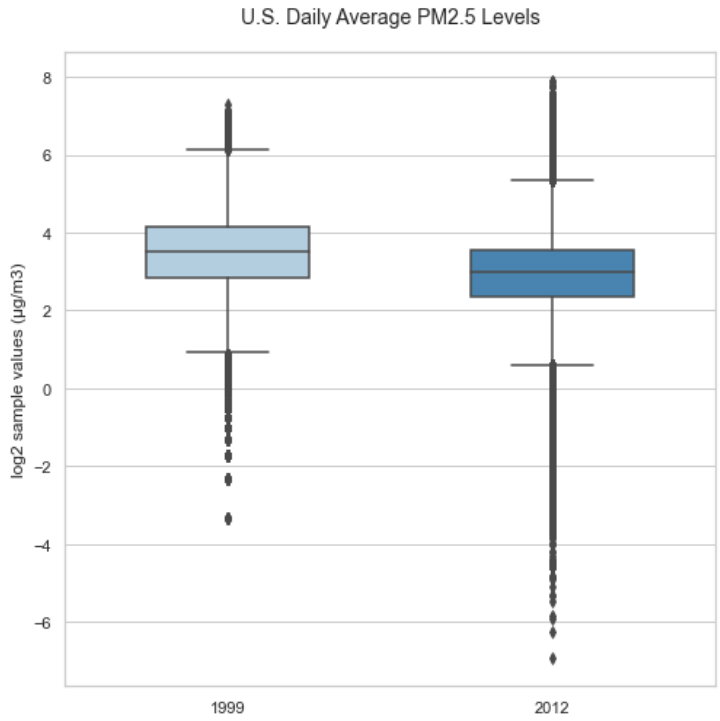

Figure 2 Average PM2.5 levels for years 1999 and 2012 Data Source: Environmental Protection Agency

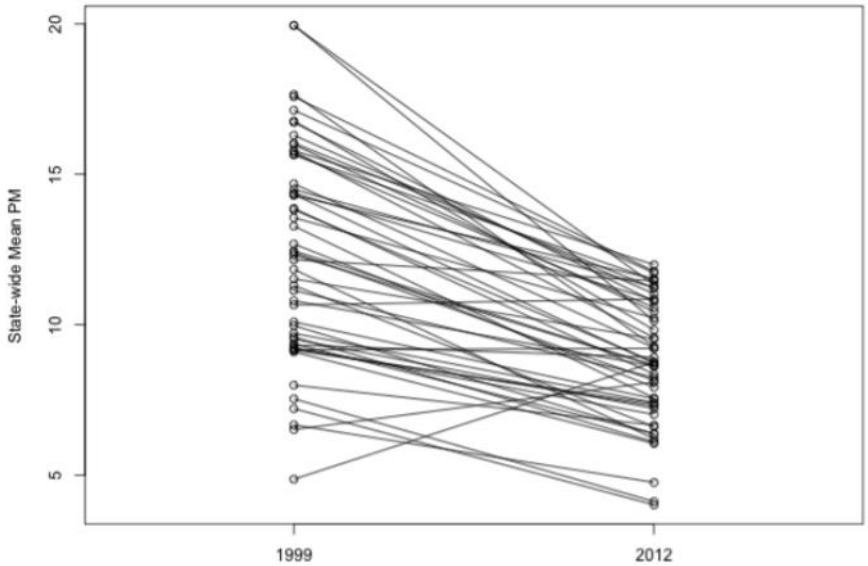

Figure 3 Statewide Average PM2.5 levels for years 1999 and 2012. Source: Peng (2020), p.157

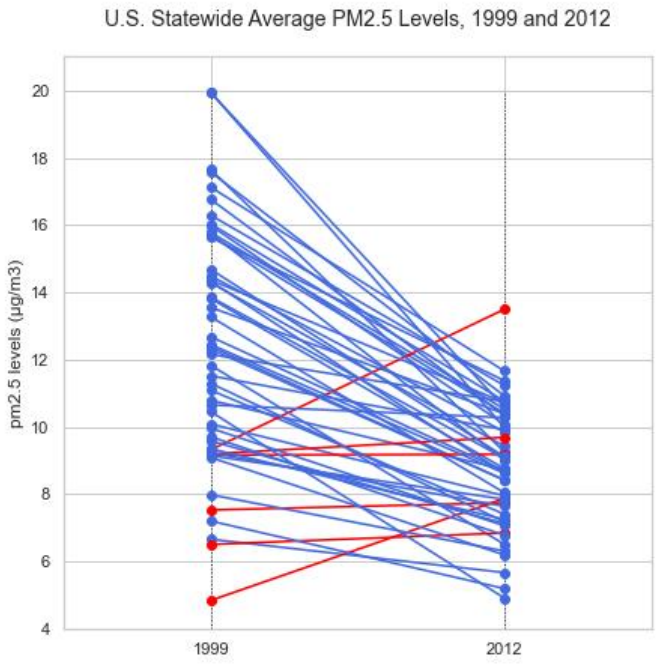

Figure 4 Statewide Average PM2.5 Levels for years 1999 and 2012. States with increases in PM2.5 levels are highlighted in red. Data Source: Environmental Protection Agency

PM2.5 Studies, part 2

Target questions: How does the level of PM 2.5 in the U.S. compare between 1999, 2012, and 2020? Which states have the highest and lowest levels of PM 2.5?

Prior to creating the visualizations in this section, the data was cleaned to remove non-U.S. states, excluded events, 1-hour samples, and 24-hour block average samples when reported on the same day as a 24-hour sample. Fig 5 illustrates the ongoing decline of both the mean PM2.5 levels and the range of values within the 75 percentile. Outliers were removed from the boxplot in Fig 5 because they rendered the plot virtually unreadable except for the outliers.

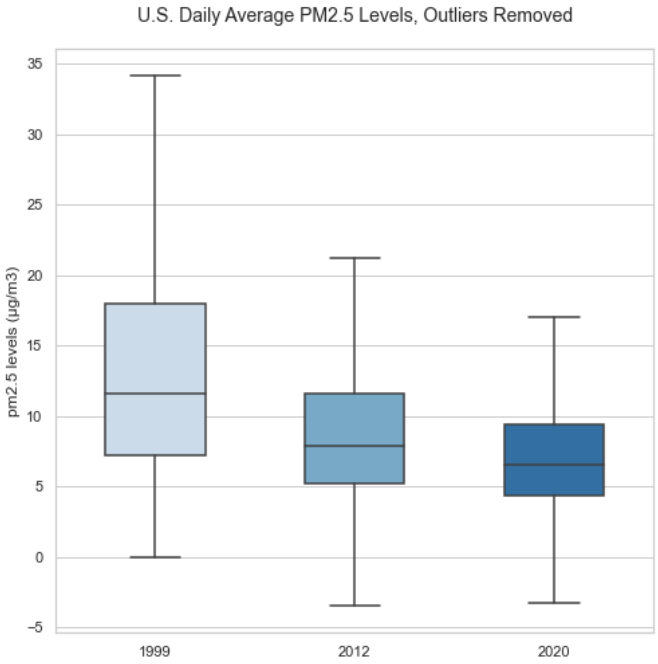

Figure 5 Average PM2.5 levels in the U.S for years 1999, 2012, and 2020. Data Source: Environmental Protection Agency

The slope chart for 1999 and 2012 was re-plotted with cleaned data and the result was indistinguishable from Fig 4. The states with increases from 1999 to 2012 were identified as Idaho, Montana, Colorado, New Mexico, and Hawaii. The slope chart for 2012 and 2020 (Fig 6) illustrates that most states had decreases in the levels of PM2.5. The states with increases from 2012 to 2020 were identified as California, Oregon, Nevada, Colorado, Florida, Washington, and Alaska.

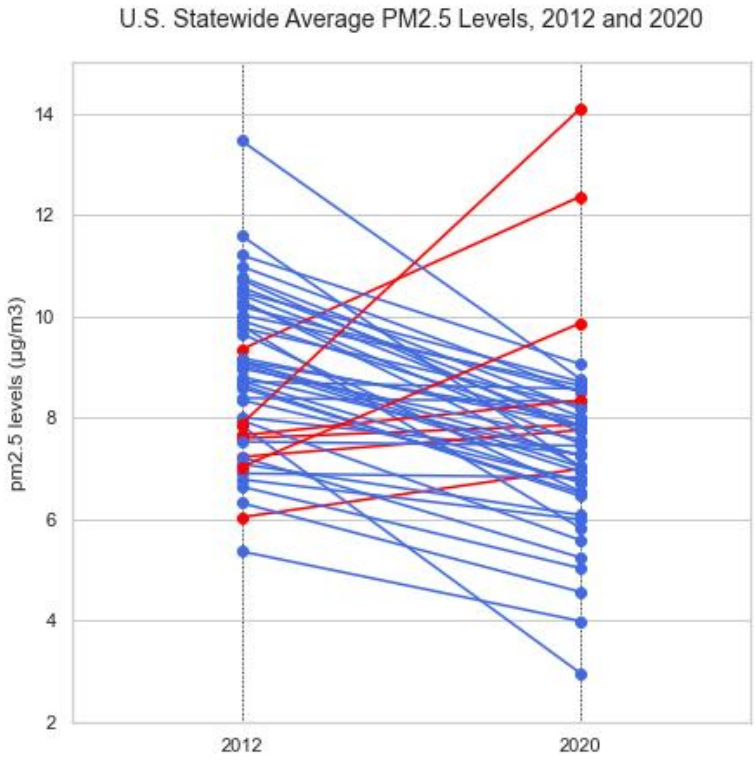

Figure 6 Statewide Average PM2.5 Levels for years 2012 and 2020 Data Source: Environmental Protection Agency

Map visualizations (Fig 7) and bar charts (Figs 8-10) were chosen to illustrate the changing levels of PM2.5 by state for the target years. Due to the difficulty in creating maps showing all 50 states in a reasonably compact manner, I used a map of the United States with boundaries limiting the map to the continental U.S. It is apparent when viewing the PM2.5 maps that the levels are decreasing. The bar charts do include Alaska and Hawaii.

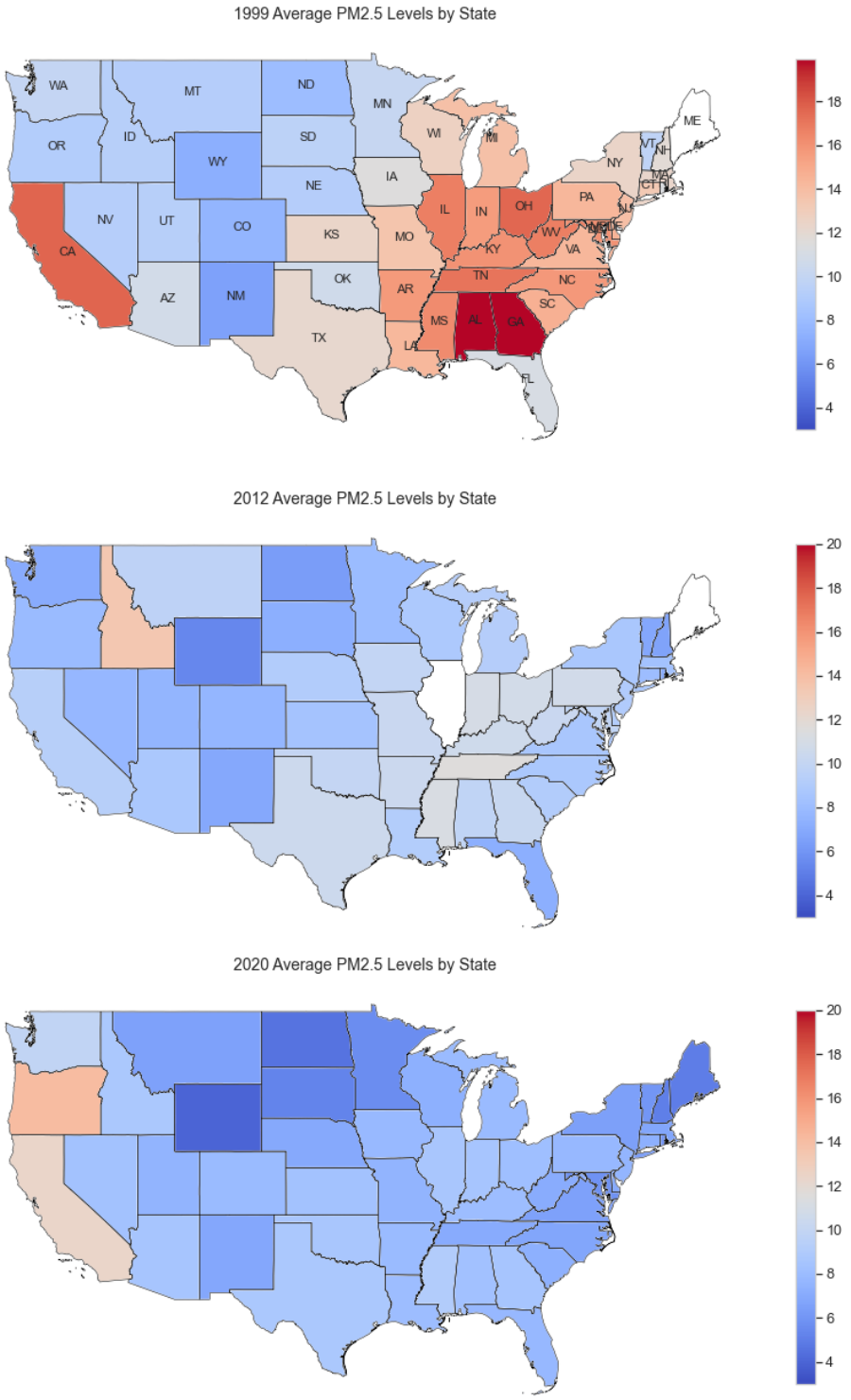

Figure 7 Mean PM2.5 levels by continental U.S. state for years 1999, 2012, and 2020. Data Source: Environmental Protection Agency

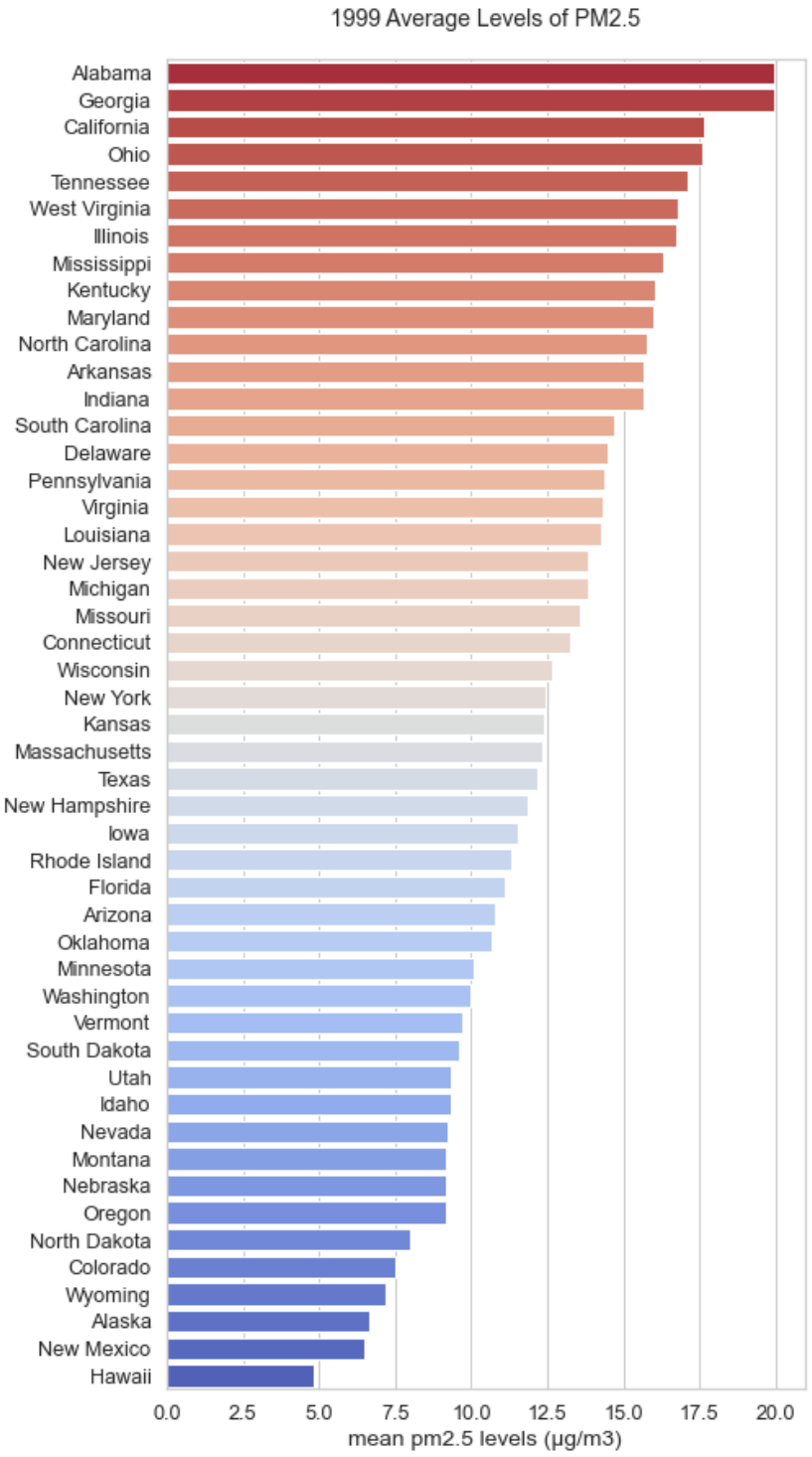

Figure 8 Mean PM2.5 levels for each U.S. state for year 1999 Data Source: Environmental Protection Agency

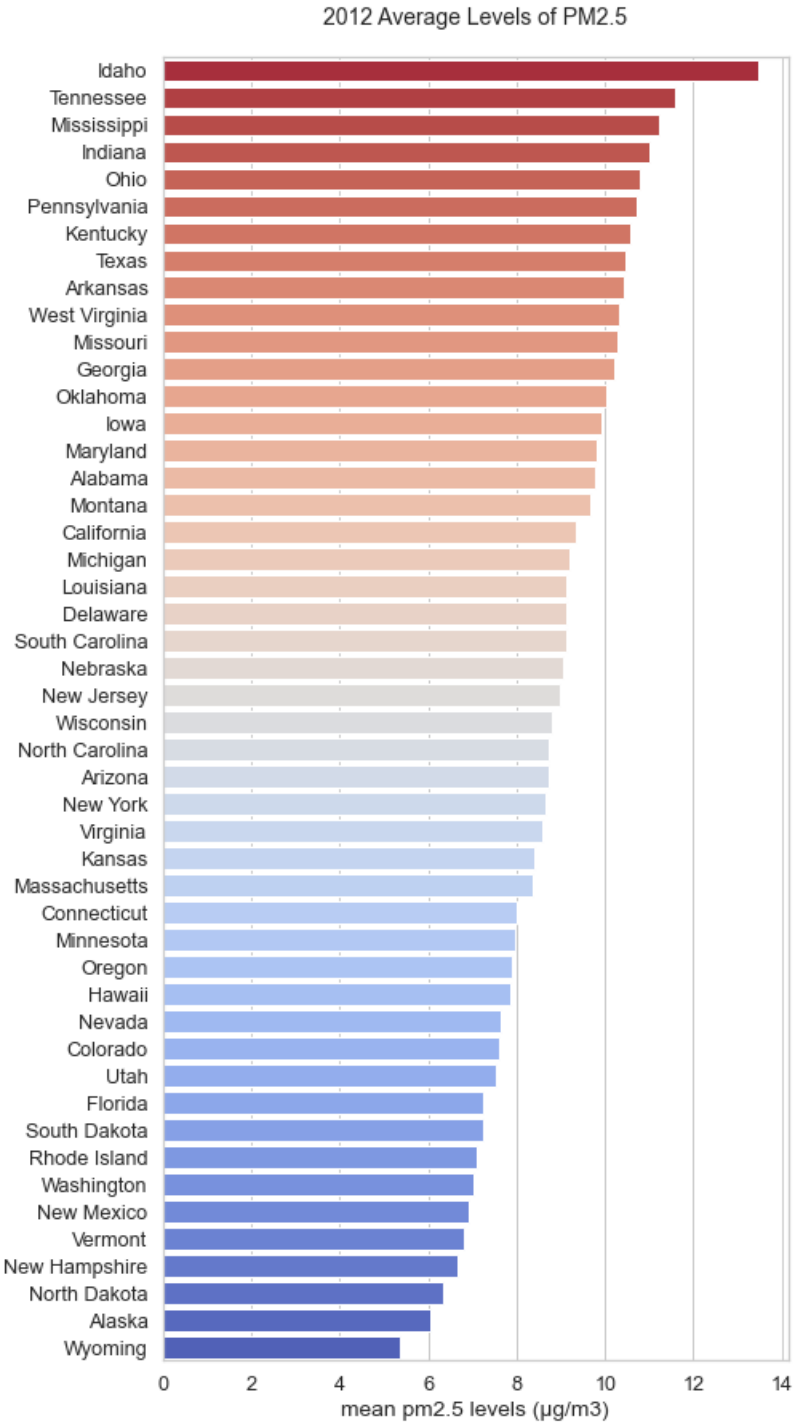

Figure 9 Mean PM2.5 levels for each U.S. state for year 2012 Data Source: Environmental Protection Agency

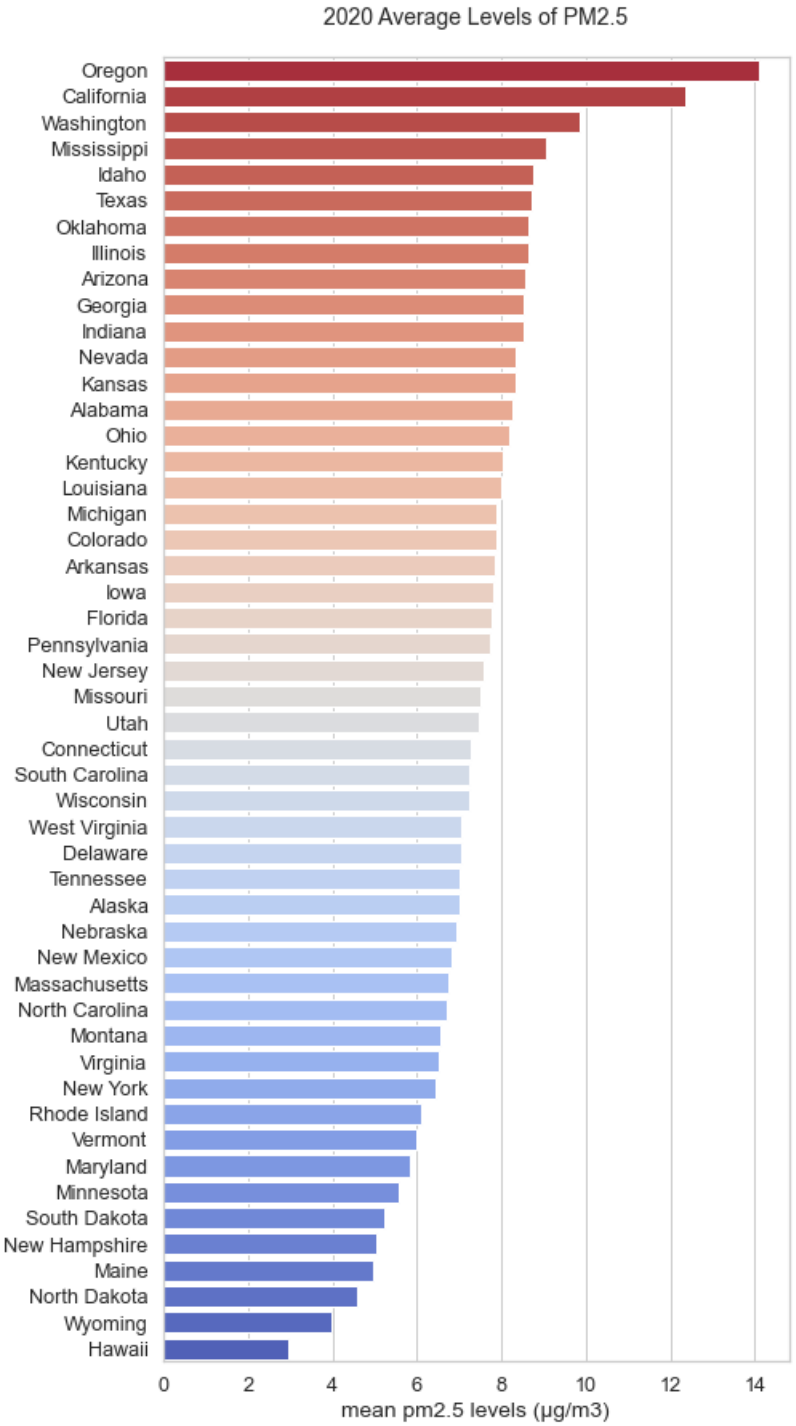

Figure 10 Mean PM2.5 levels for each U.S. state for year 2012 Data Source: Environmental Protection Agency

Ozone Studies

Target questions: How does the level of ozone in the U.S. compare between 1999, 2012, and 2020? Which states have the highest and lowest levels of PM ozone? From the boxplot, Fig 11, it can be seen that the average levels of ozone increased slightly from 1999 to 2012 and then decreased in 2020. The range of values within the 75 percentile decreased steadily from 1999 through 2020. The number of outliers in 2020 have increased dramatically, which means there are extremely high ozone days. A violin plot (Fig 12) was generated to better illustrate the distribution of the data.

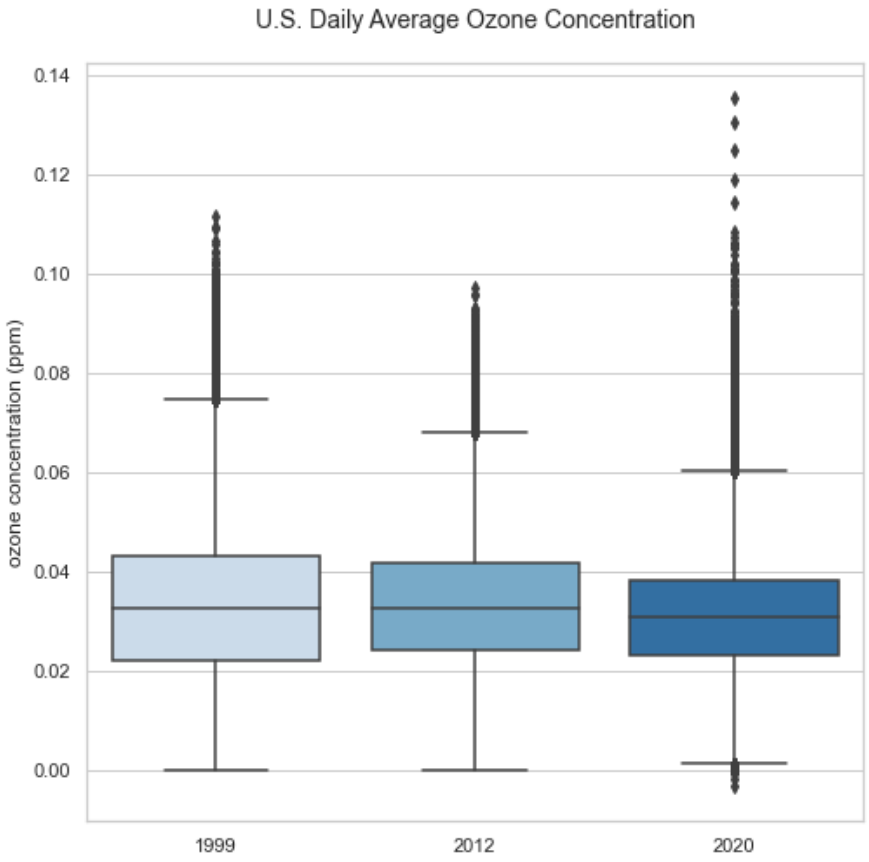

Figure 11 Average ozone levels in the U.S for years 1999, 2012, and 2020 Data Source: Environmental Protection Agency

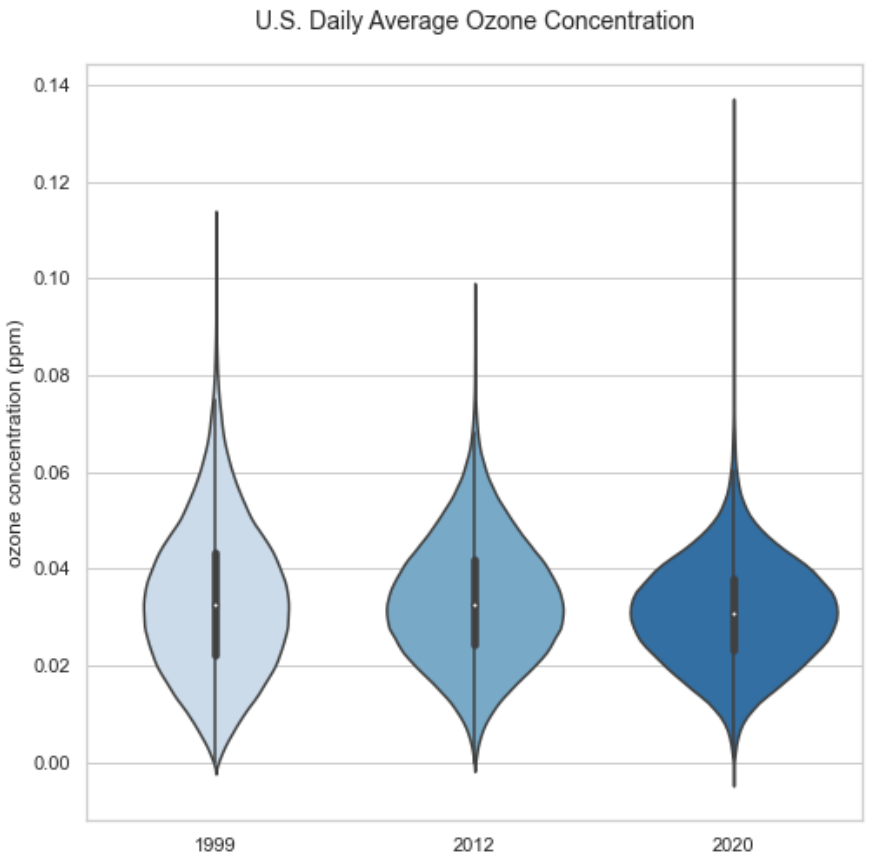

Figure 12 Average ozone levels in the U.S for years 1999, 2012, and 2020 visualized with a violin chart Data Source: Environmental Protection Agency

Slope charts were used again to model the changes in average statewide values of ozone (Figs 13 and 15). Between 1999 and 2012, several states had an increase in their average daily ozone concentrations. Fig 14 depicts those states. Between 2012 and 2020, only two states had an increase in their average daily ozone concentrations. Fig 16 shows those states.

Maps and bar charts show the changes in mean ozone concentration by state. The order of presentation was changed to show the bar charts directly following the map for a given year.

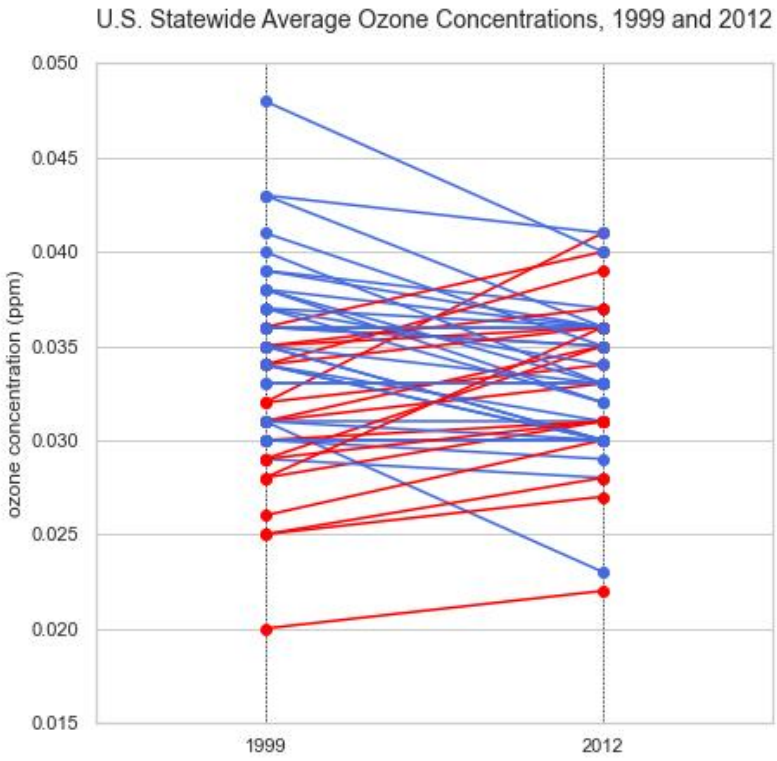

Figure 13 Statewide Average ozone levels for years 1999 and 2012 Data Source: Environmental Protection Agency

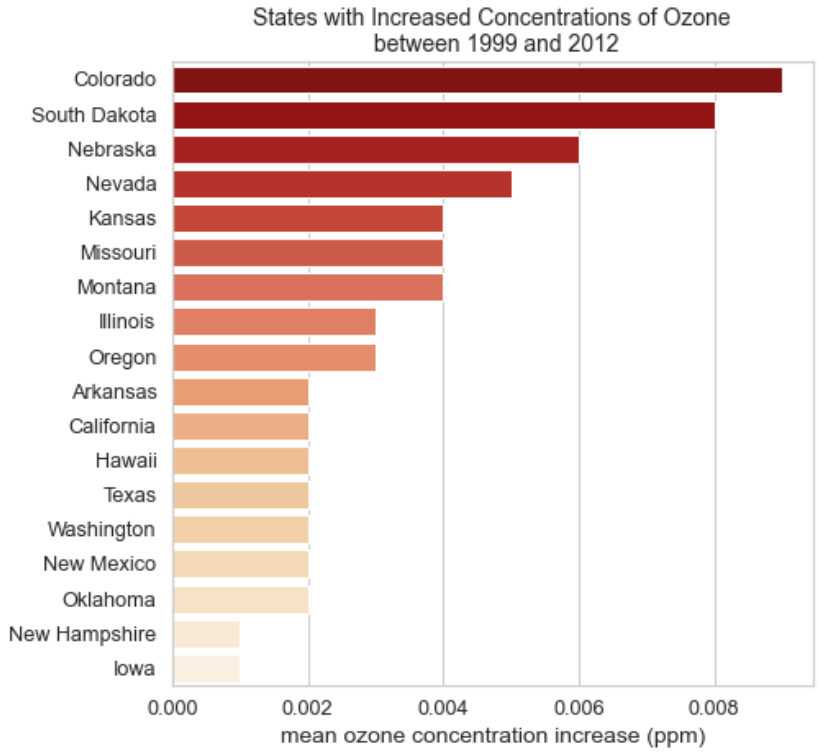

Figure 14 U.S. states with increased concentrations of ozone from years 1999 to 2012 Data Source: EPA

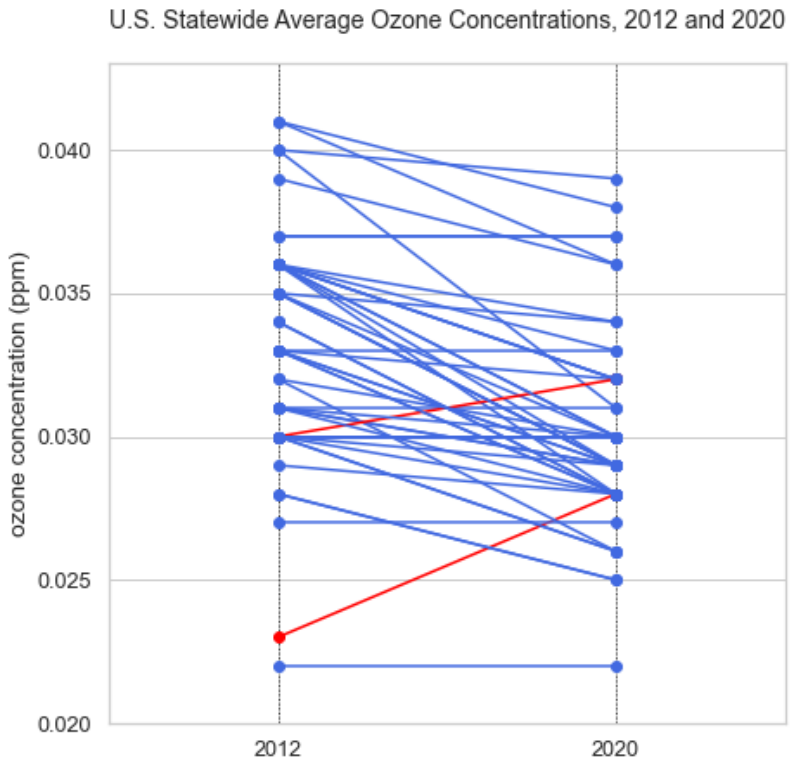

Figure 15 Statewide Average ozone levels for years 2012 and 2020 Data Source: Environmental Protection Agency

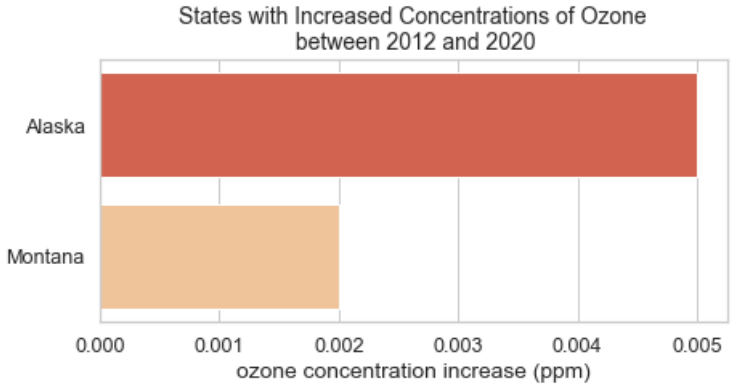

Figure 16 States with increased concentrations of ozone from years 2012 and 2020 Data Source: EPA

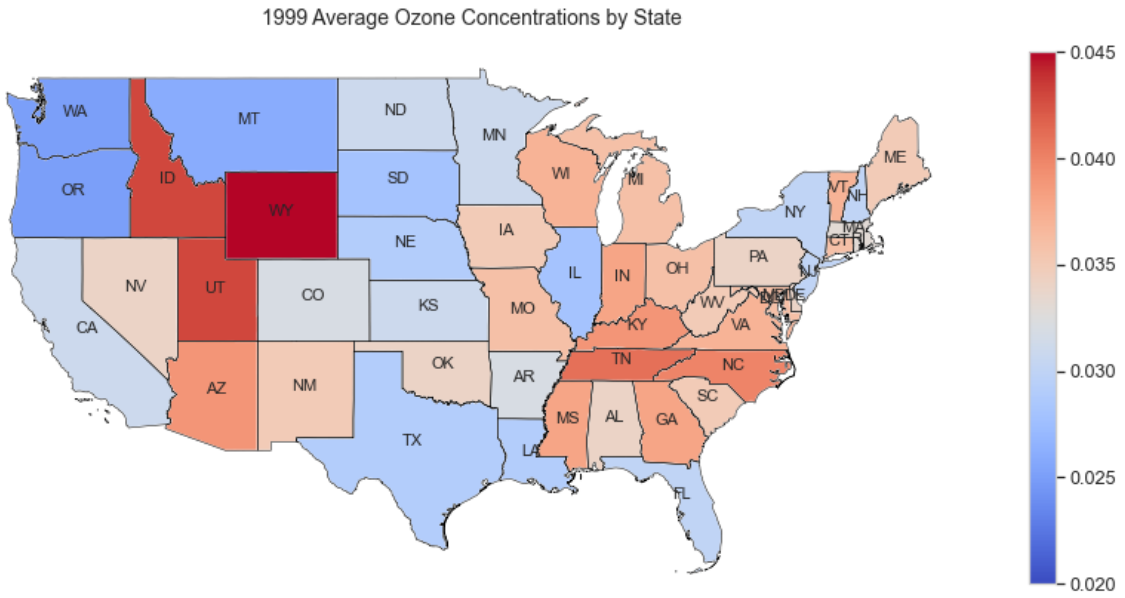

Figure 17 Average ozone concentrations in continental U.S. States in 1999 Data Source: EPA

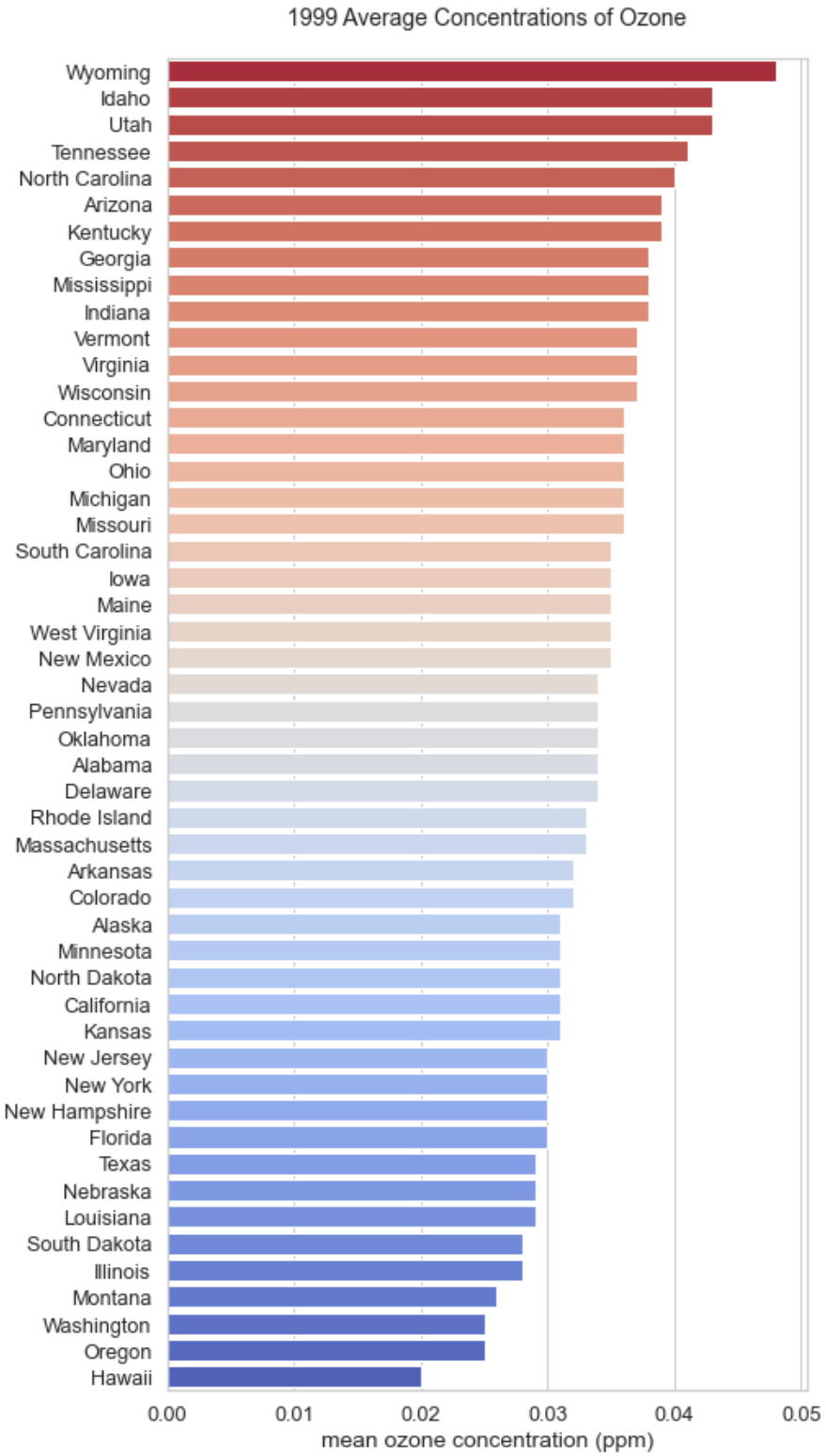

Figure 18 Mean ozone levels for each U.S. state for year 1999 Data Source: Environmental Protection Agency

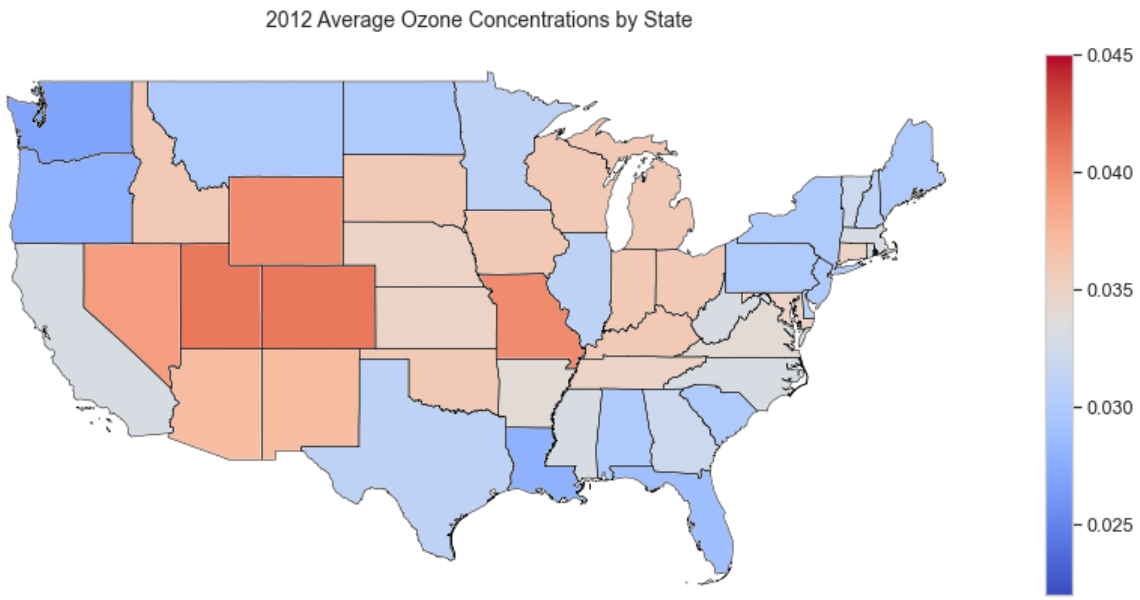

Figure 19 Average ozone concentrations in continental U.S. States in 2012 Data Source: EPA

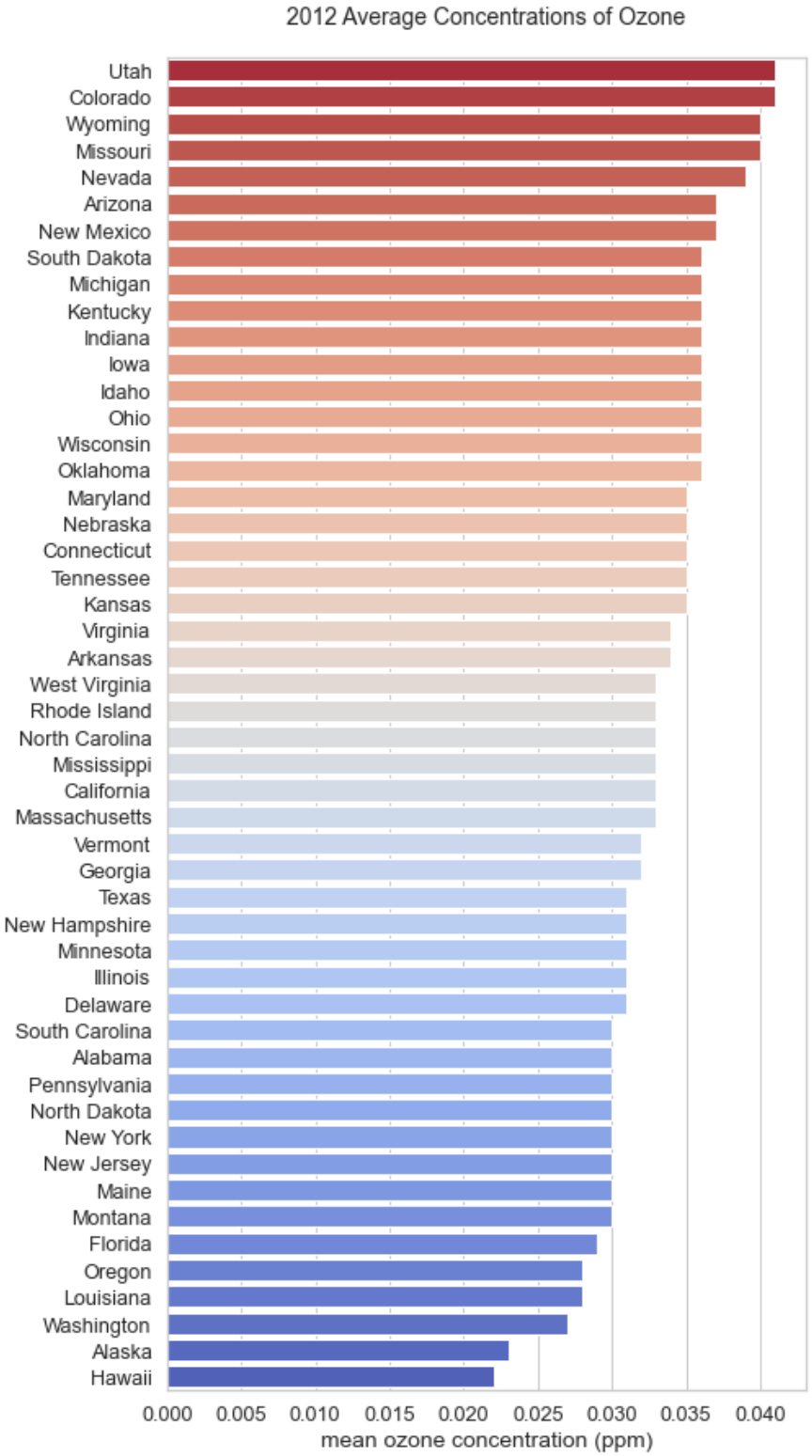

Figure 20 Mean ozone levels for each U.S. state for year 2012 Data Source: Environmental Protection Agency

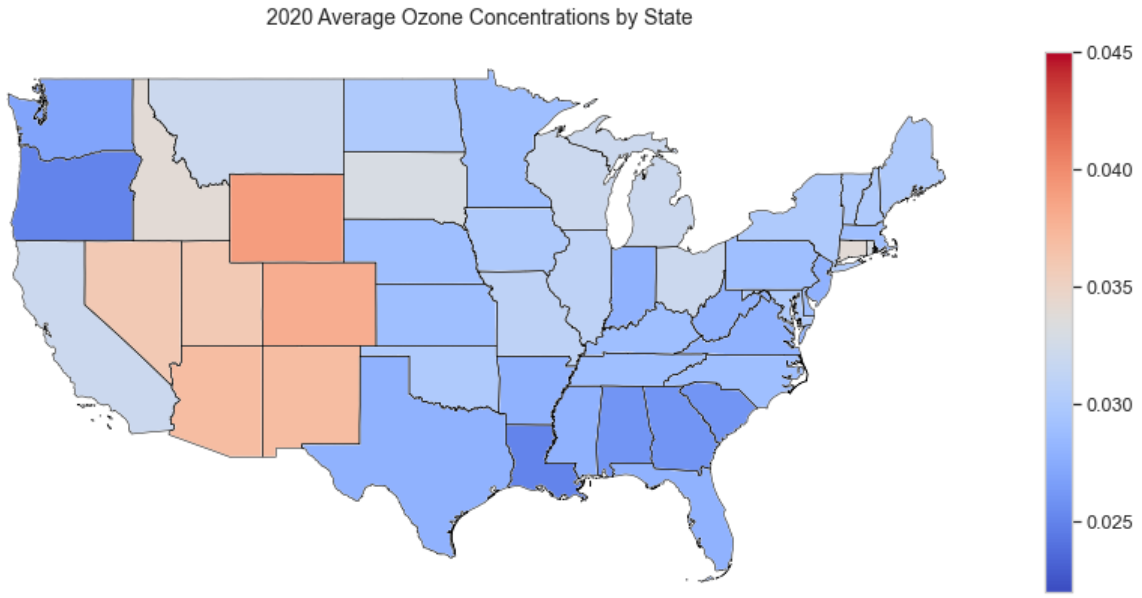

Figure 21 Average ozone concentrations in continental U.S. States in 2020 Data Source: EPA

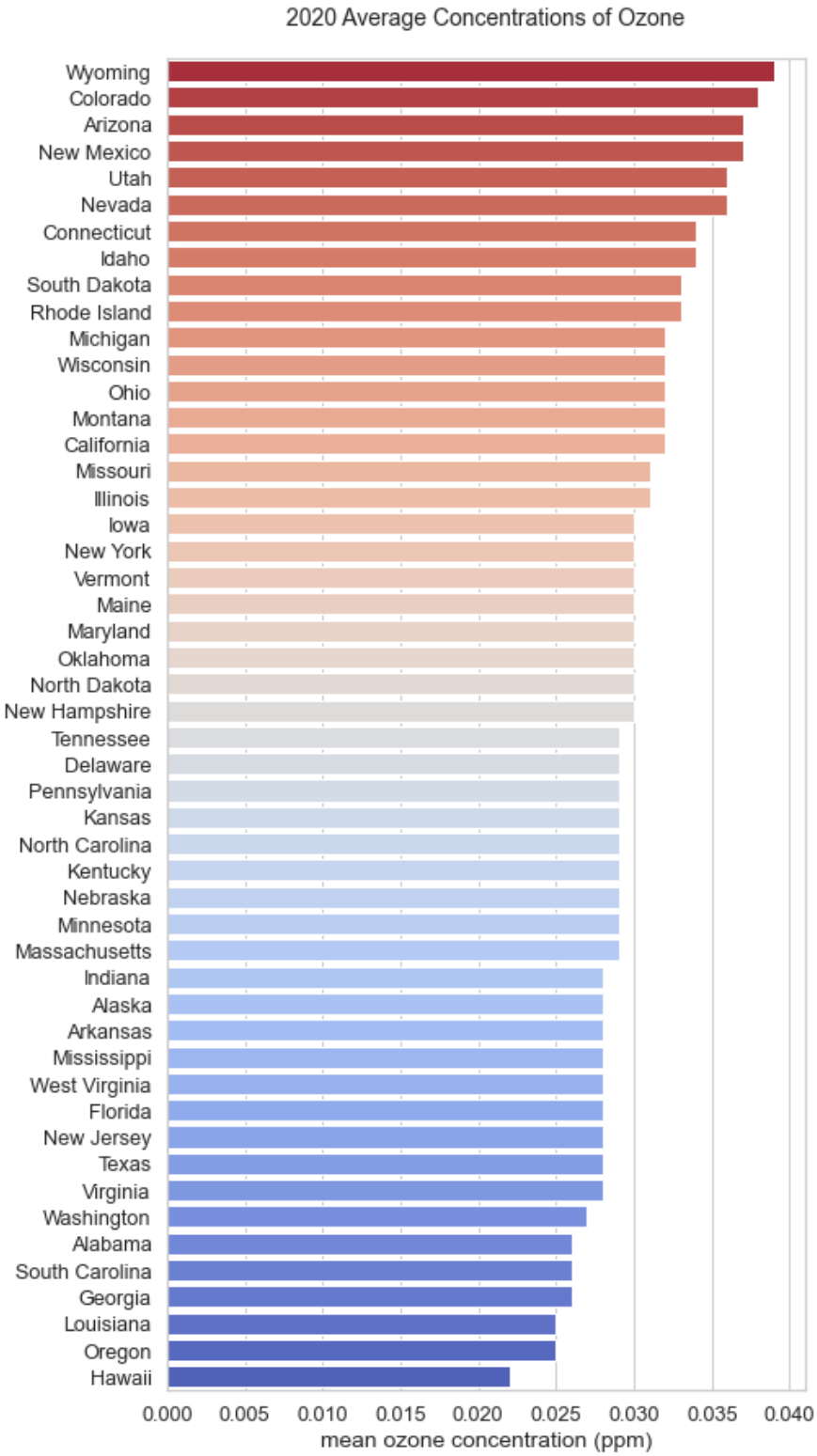

Figure 22 Mean ozone levels for each U.S. state for year 2012 Data Source: Environmental Protection Agency

AQI Studies

Target questions: Which states have reports of hazardous (AQI) in each of the three years of the study? Which states have reports of very unhealthy AQI? Which states have reports of unhealthy AQI? Which states have reports of unhealthy for sensitive individuals AQI? What does a map of AQI look like?

Prior to creating map visualization for the AQI section, I created maps showing the location of counties with AQI monitors. Maps were created for all three target years, but because they were virtually indistinguishable, I am showing the map for 1999 (Fig 23) only. It is evident that not all counties in the U.S. have air quality monitoring sites. Figs 28 and 29 show the difference between mapping data at the state level and at the county level. Clearly, the county level is more accurate than generalizing to a state level.

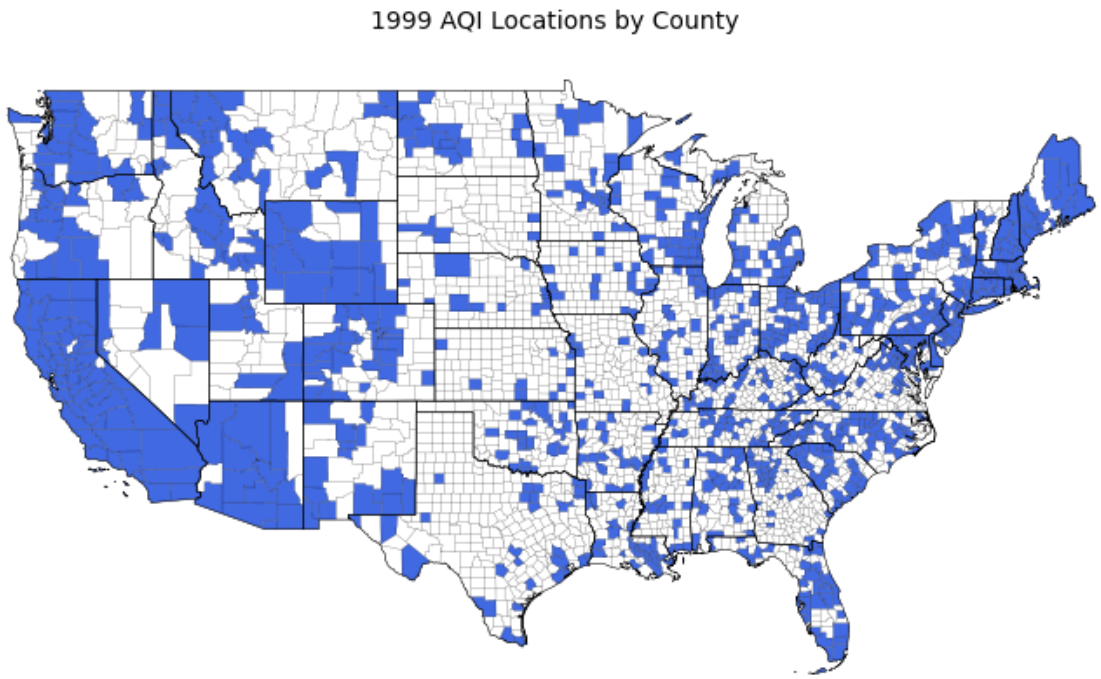

Figure 23 Map of continental U.S. counties that have AQI monitors in 1999. Areas without color are non-reporting counties. Data Source: EPA

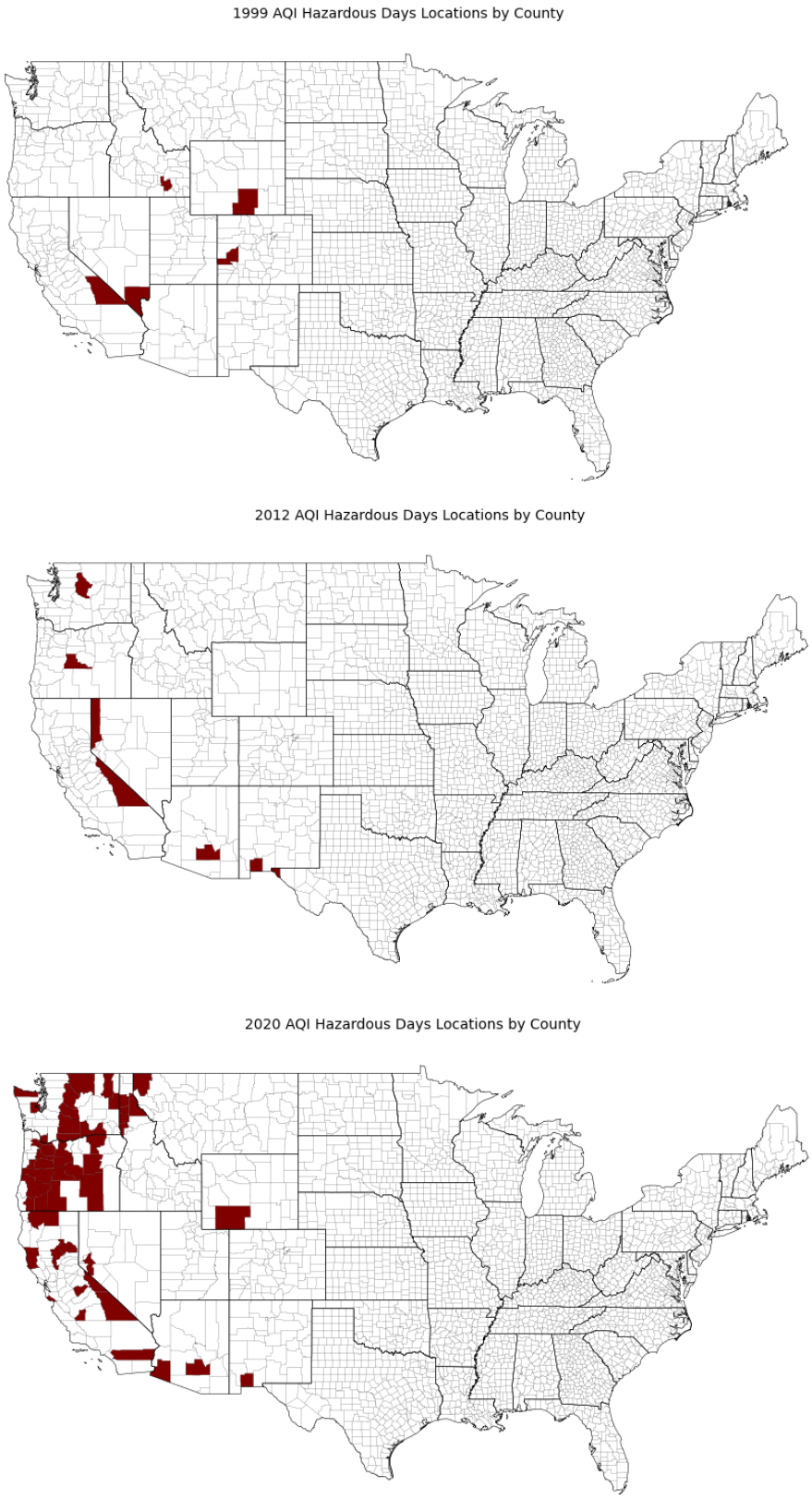

Figure 24 Counties in continental U.S. states reporting hazardous days in years 1999, 2012, and 2020. Data Source: Environmental Protection Agency

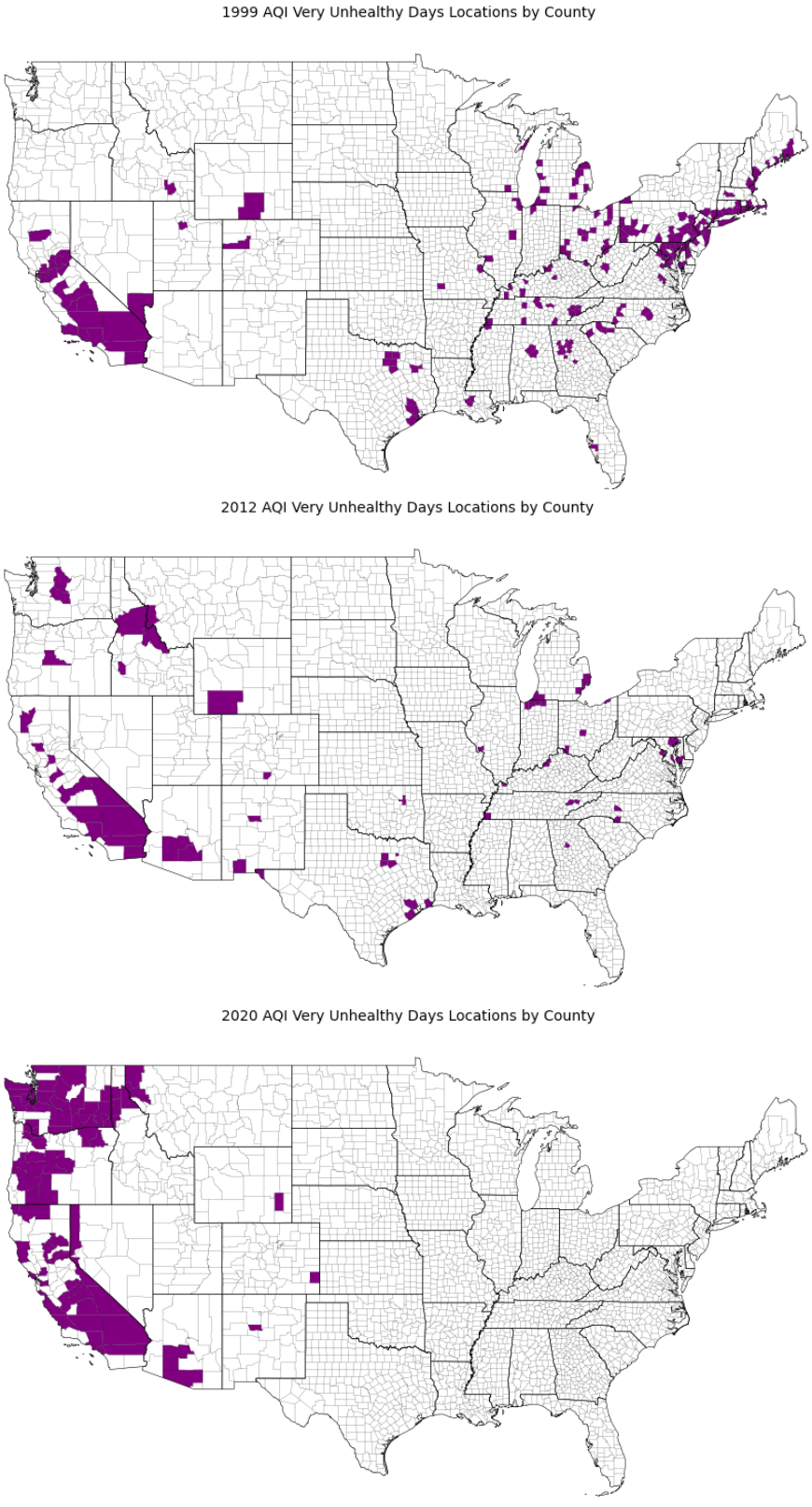

Figure 25 Counties in continental U.S. states reporting very unhealthy days in years 1999, 2012, and 2020. Data Source: Environmental Protection Agency

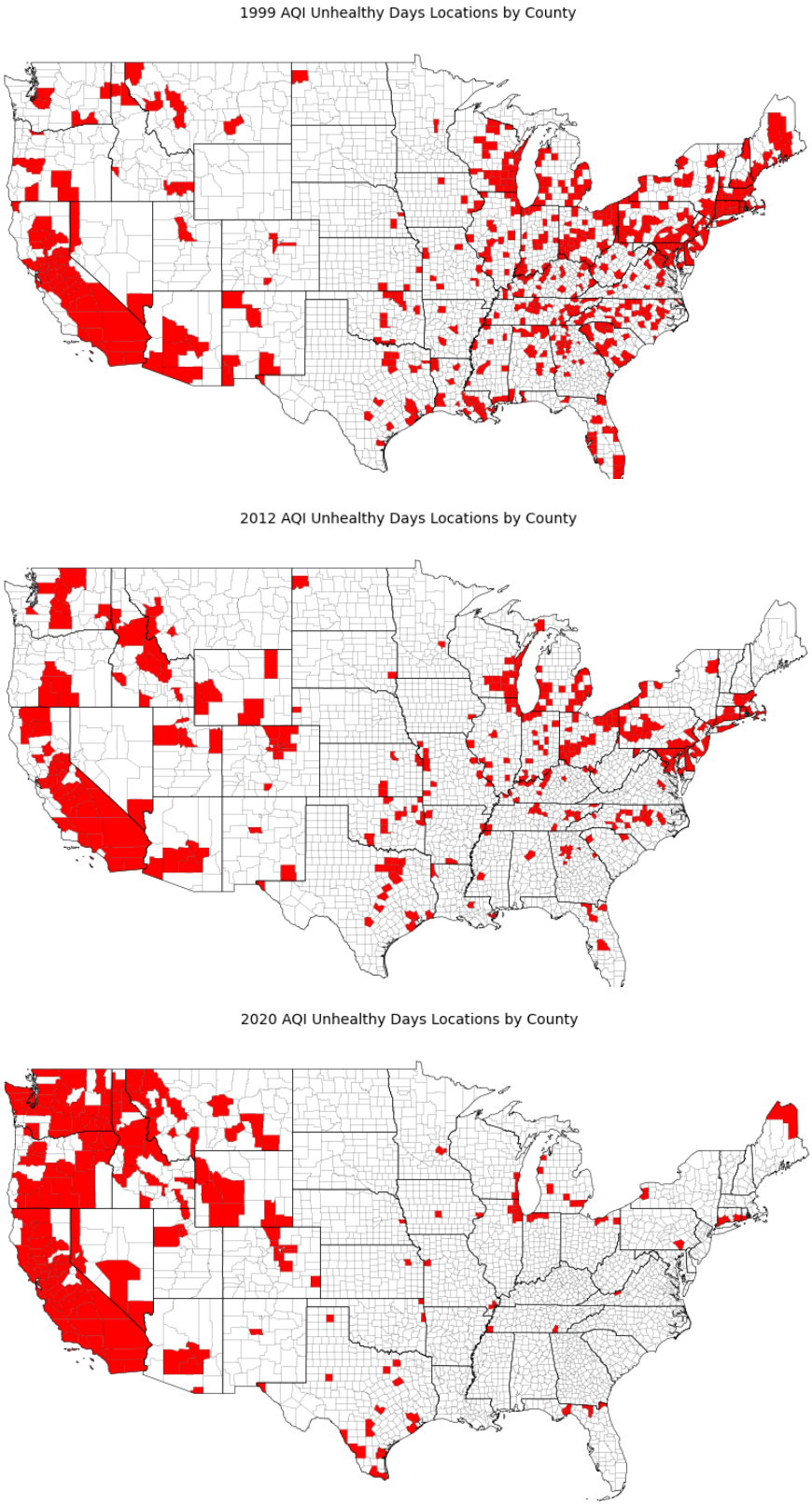

Figure 26 Counties in continental U.S. states reporting unhealthy days in years 1999, 2012, and 2020. Data Source: Environmental Protection Agency

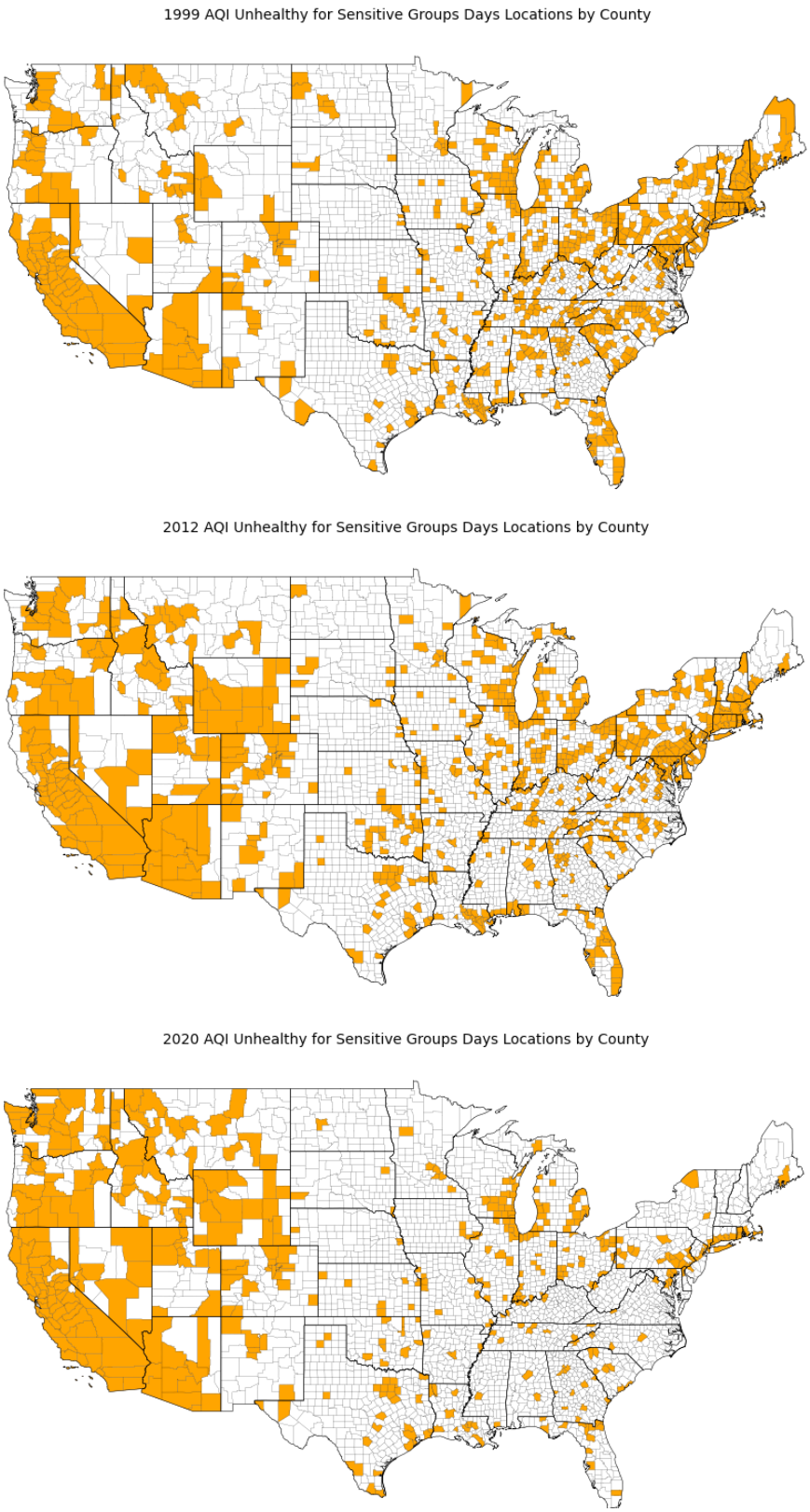

Figure 27 Counties in continental U.S. states reporting unhealthy for sensitive groups days in years 1999, 2012, and 2020. Data Source: EPA

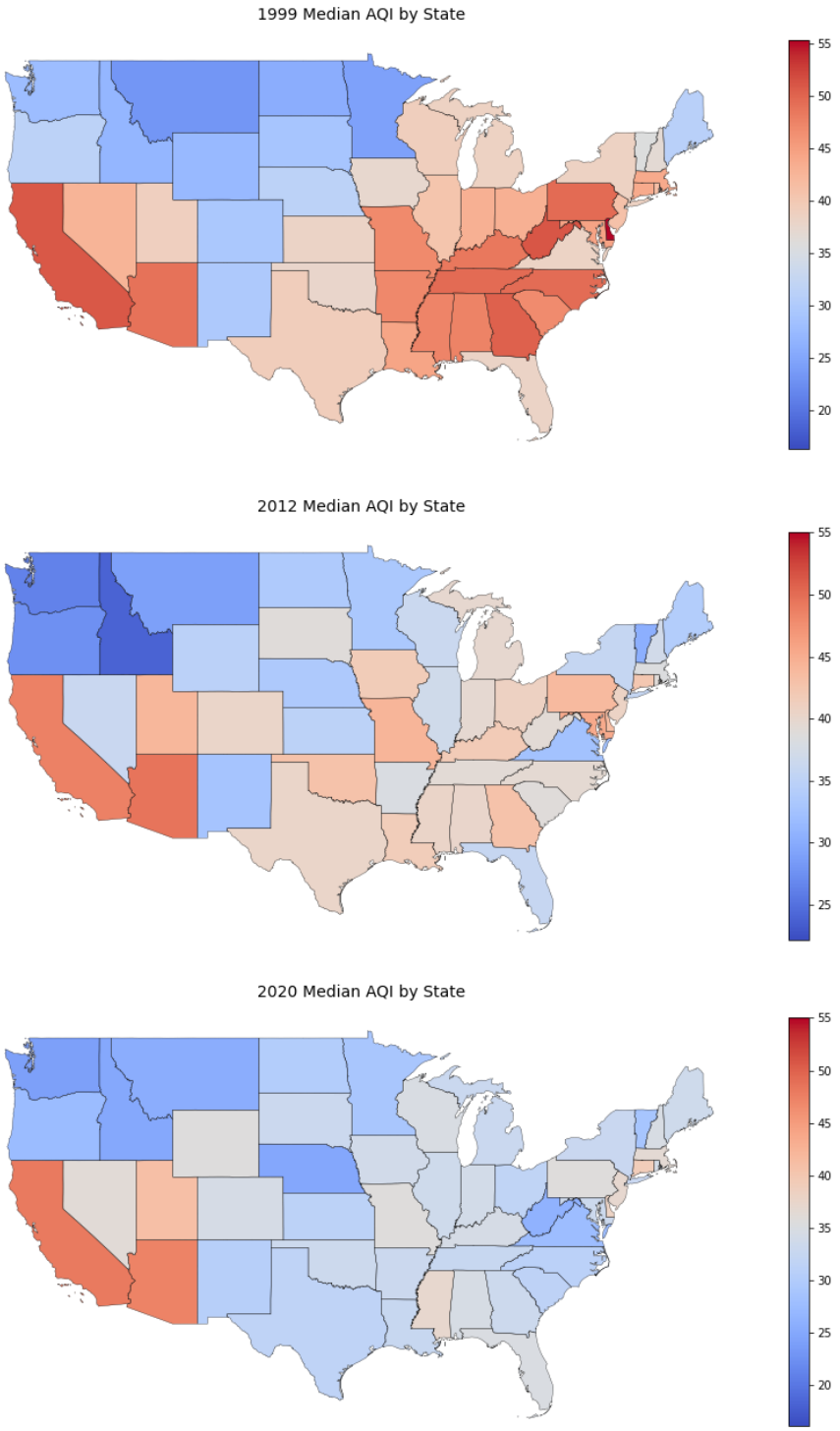

Figure 28 Median AQI in continental U.S. states for years 1999, 2012, and 2020. Data Source: Environmental Protection Agency

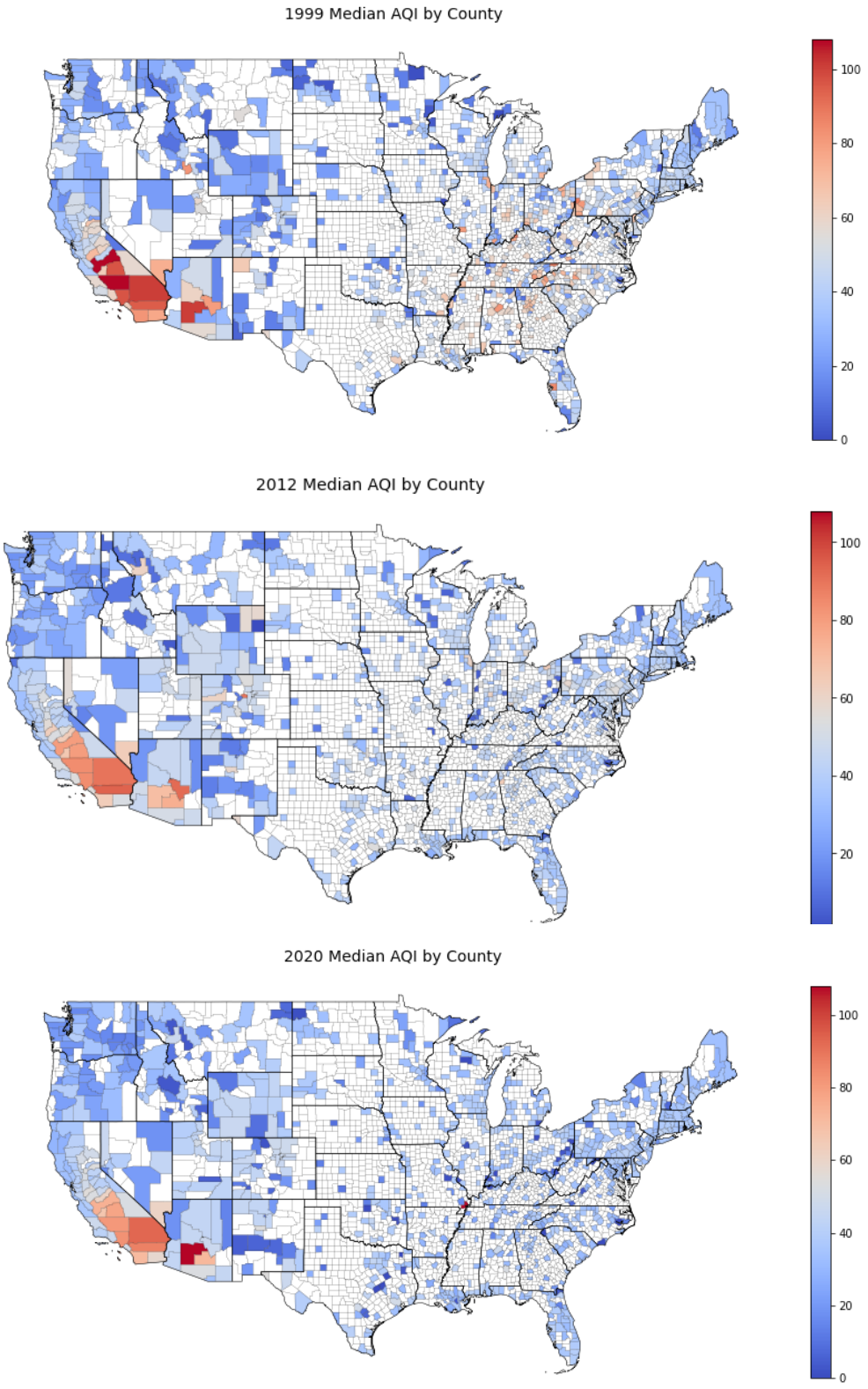

Figure 29 Median AQI by county in continental U.S. for years 1999, 2012, and 2020. Data Source: Environmental Protection Agency

Asthma Studies

Target questions: What does the prevalence of asthma look like by state? Which states having high asthma prevalence also have county level data for asthma prevalence?

2012 is the only year from the project's target years (1999, 2012, and 2020) that had asthma data available at the national level. Asthma prevalence is an estimated percentage of adults with current asthma determined through the use of telephone surveys on health data. For this section, I chose to use a map of the continental U.S. (Fig 30) for geographic / regional insights as well as a bar chart (Fig 31) for state ranking and inclusion of Alaska and Hawaii.

Vermont and Kentucky were identified as states with high levels of asthma with 2012 county asthma data available. Vermont asthma visualizations include a map and a bar chart (Fig 32). The Kentucky asthma visualization (Fig 33) has no bar chart due to many counties.

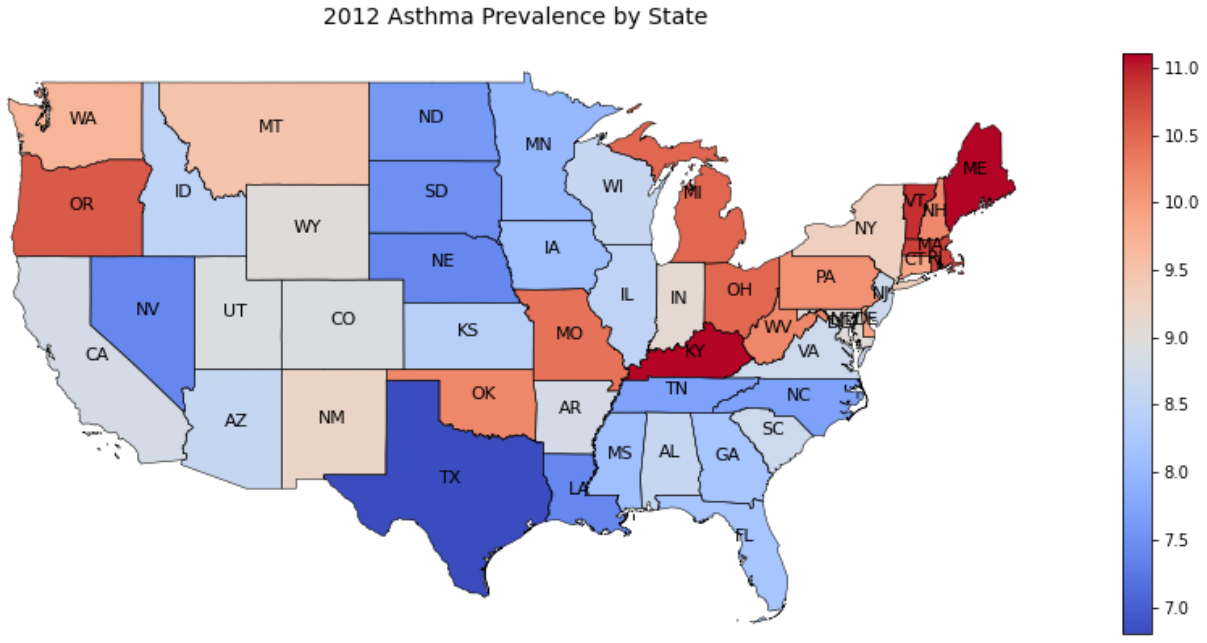

Figure 30 Percentage of adults with history and current asthma in continental U.S. States in 2012 Data Source: Centers for Disease Control and Prevention BRFSS data

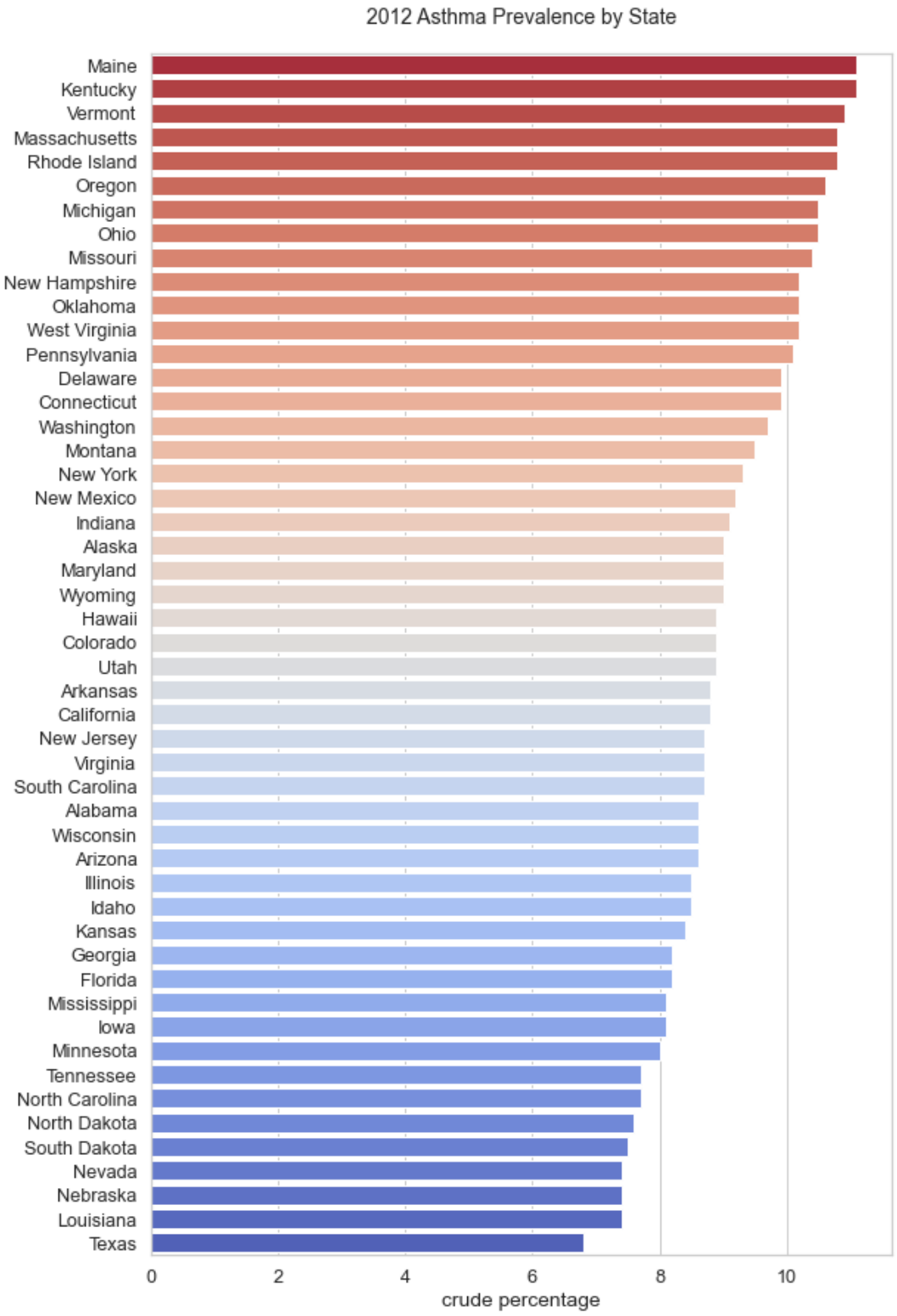

Figure 31 Percentage of U.S. adults aged 18 and older with current asthma in each U.S. State in 2012 Data Source: Centers for Disease Control and Prevention BRFSS data

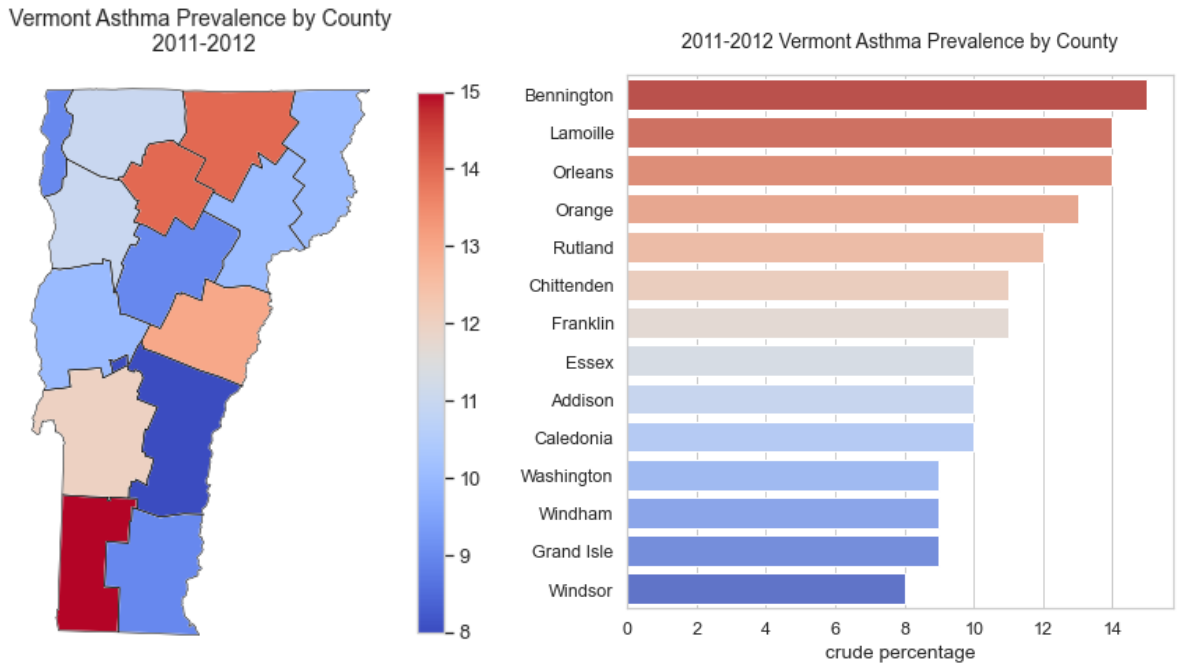

Figure 32 Percentage of Vermont adults aged 18 and older with current asthma by county Data Source: Vermont Department of Health using BRFSS data

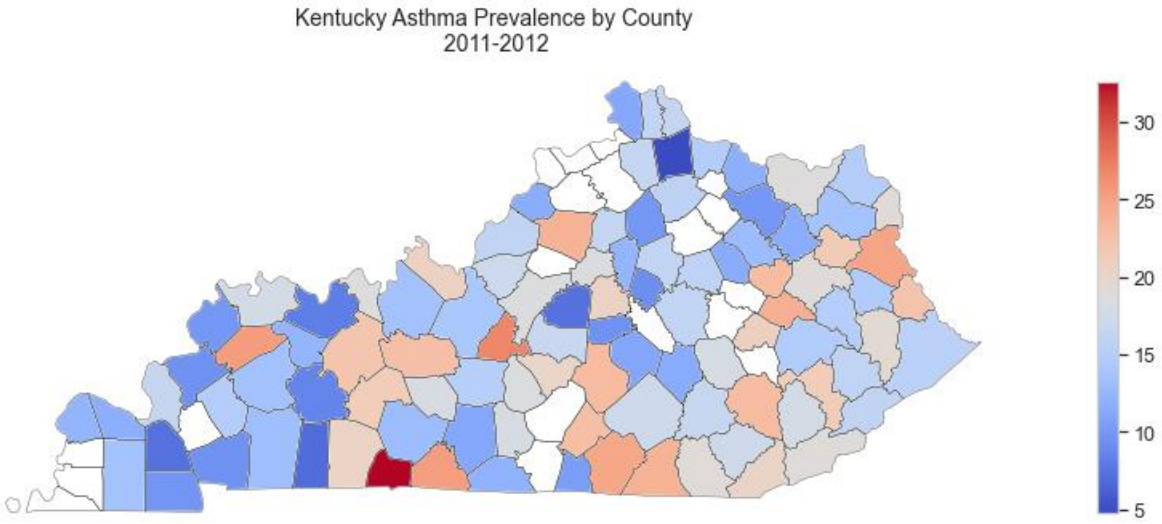

Figure 33 Percentage of Kentucky adults aged 18 and older with current asthma by county Data Sources: Centers for Disease Control and Prevention and Kentucky Department for Public Health using BRFSS data

AQI and Asthma Studies, U.S. State Level

Target question: What are ways to visualize the relationship between air quality and asthma? The question presupposes there is a relationship between air quality and asthma. To examine the relationship, I chose to use a heatmap. At the national level, I found no asthma prevalence data that also included prevalence by county. PM2.5 data, ozone data, and AQI data were rolled up to state levels. A correlation matrix was created, followed by a heatmap (Fig 34). The bottom row of the heatmap shows the correlation between asthma and the air quality measures. All correlations between asthma and the air quality measures were very low.

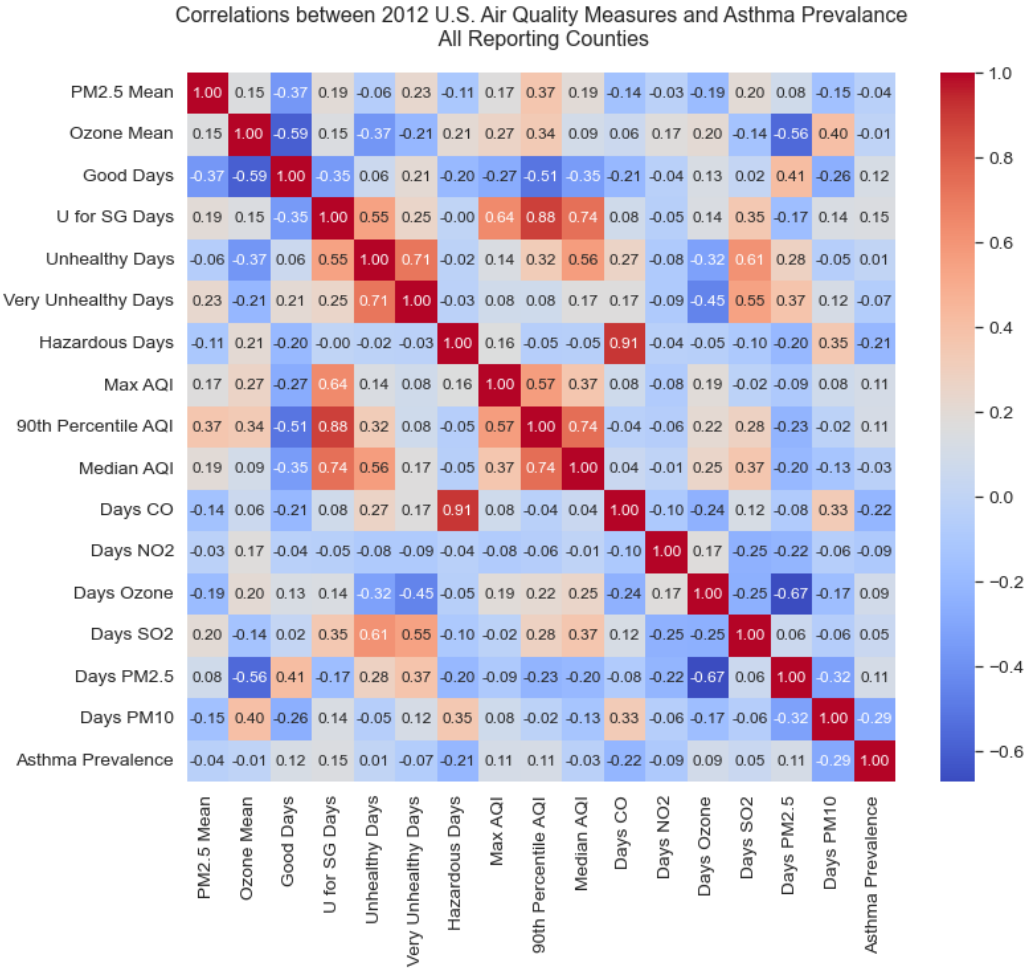

Figure 34 Correlations between air quality measures and asthma prevalence nationwide in 2012 U.S. Data Sources: EPA and CDC

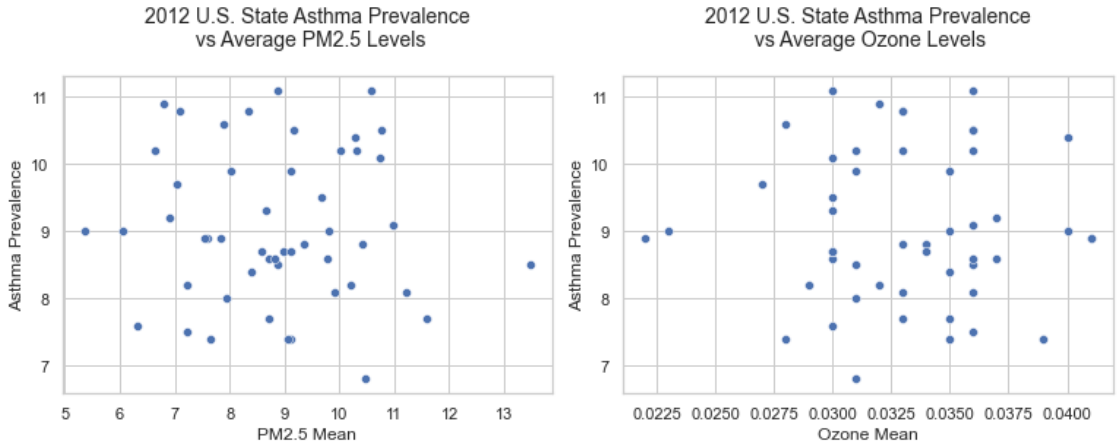

Figure 35 No relationship between average PM2.5 levels and adult asthma prevalence or ozone levels and adult asthma prevalence nationwide Data Sources: EPA and CDC

Scatter plots are frequently used to visualize relationships between variables. I used scatter plots for mean PM2.5 and asthma and ozone and asthma (Fig 35). No relationships can be ascertained from the scatter plots. The correlation coefficients from the heatmap (Fig 34) confirm this.

AQI and Asthma Studies, Vermont and Kentucky

This section extends the AQI and asthma studies at the national level to state levels. The target question remains: What are ways to visualize the relationship between asthma and air quality? In the process of getting the data prepared for analysis, the scant number of counties reporting PM2.5 and ozone data prompted me to include visualizations of PM2.5 and ozone by counties for Vermont (Fig 36) and Kentucky (Figs 39 and 40). Heatmaps and scatter plots were created for these states, as well, although the scatter plots for Vermont had very little data.

The heatmap for Vermont (Fig 38) shows a high correlation (0.8) between max AQI and asthma prevalence. One could expect that correlation. However, there is a high correlation (0.73) between Number of Good Days and asthma prevalence. This would not be expected.

The scatter plots for asthma vs PM2.5 and asthma vs ozone show no relationship. The calculated correlation coefficients were -0.0697 and -0.970, respectively.

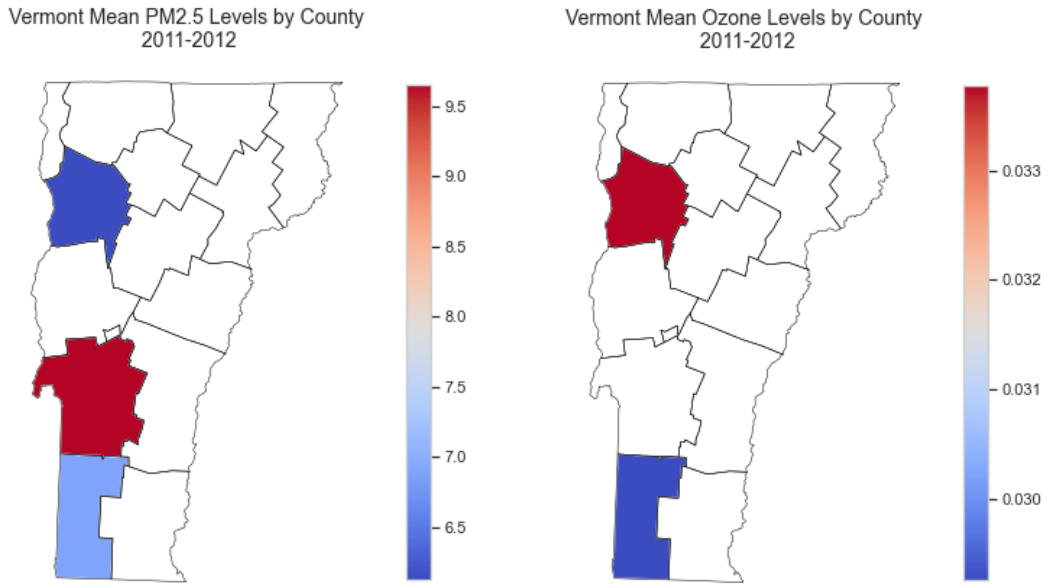

Figure 36 PM2.5 and ozone levels by counties in Vermont. Counties without color are non-reporting counties Data Source: EPA

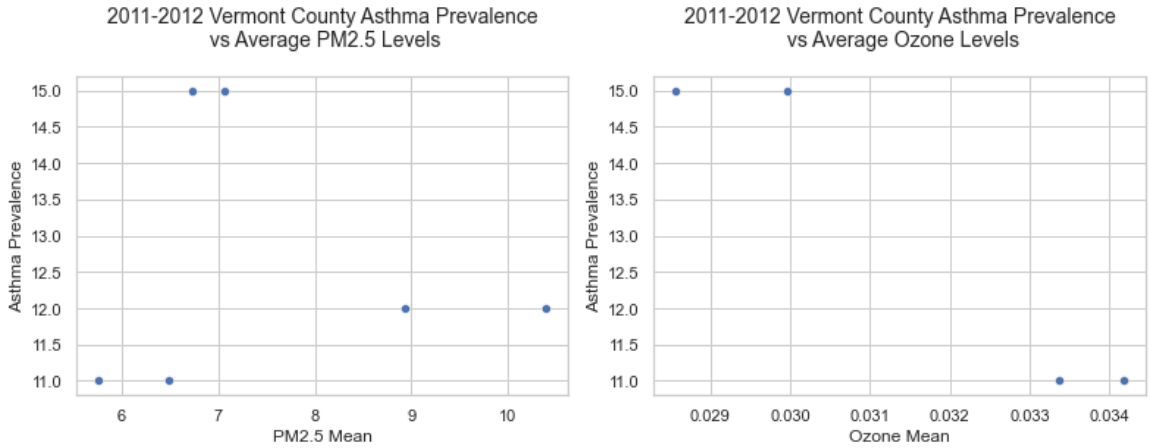

Figure 37 No relationship between average PM2.5 levels and adult asthma prevalence or ozone levels and adult asthma prevalence in Vermont Data Sources: EPA and Vermont Public Health Department using BRFSS

The heatmap for Kentucky (Fig 38) shows very low correlations between AQI measures and asthma prevalence.

The scatter plots for asthma vs PM2.5 and asthma vs ozone (Fig 41) show no relationship. The calculated correlation coefficients were 0.0453 and -2.805, respectively.

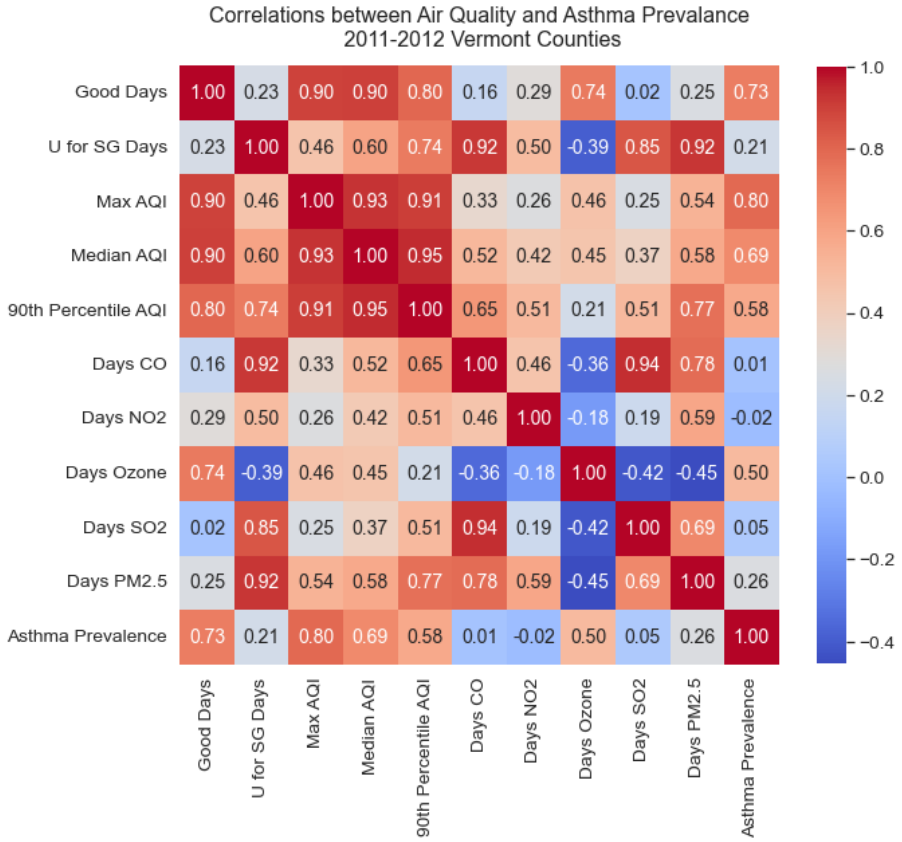

Figure 38 Correlations between air quality measures and adult asthma prevalence in Vermont in 2012 U.S. Data Sources: EPA and Vermont Department of Public Health using BRFSS

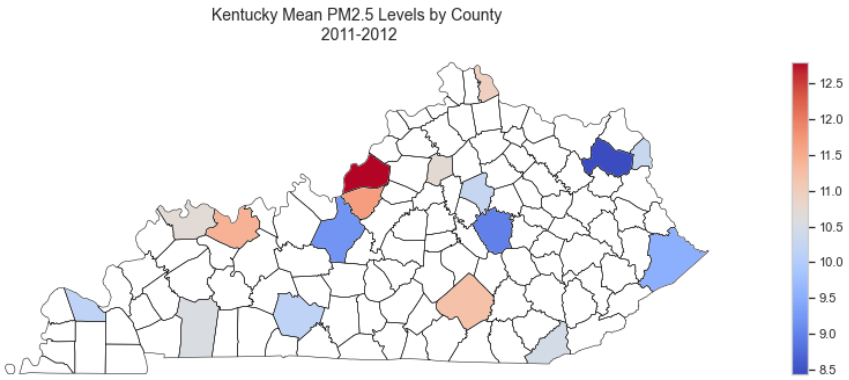

Figure 39 PM2.5 levels by counties in Kentucky. Counties without color are non-reporting counties Data Source: EPA

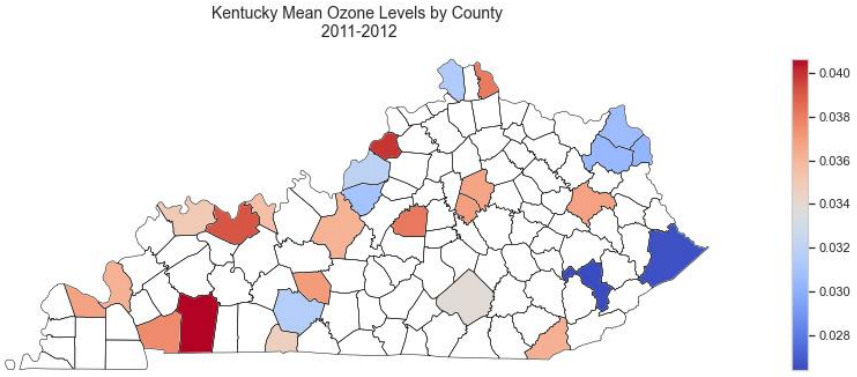

Figure 40 ozone levels by counties in Kentucky. Counties without color are non-reporting counties Data Source: EPA

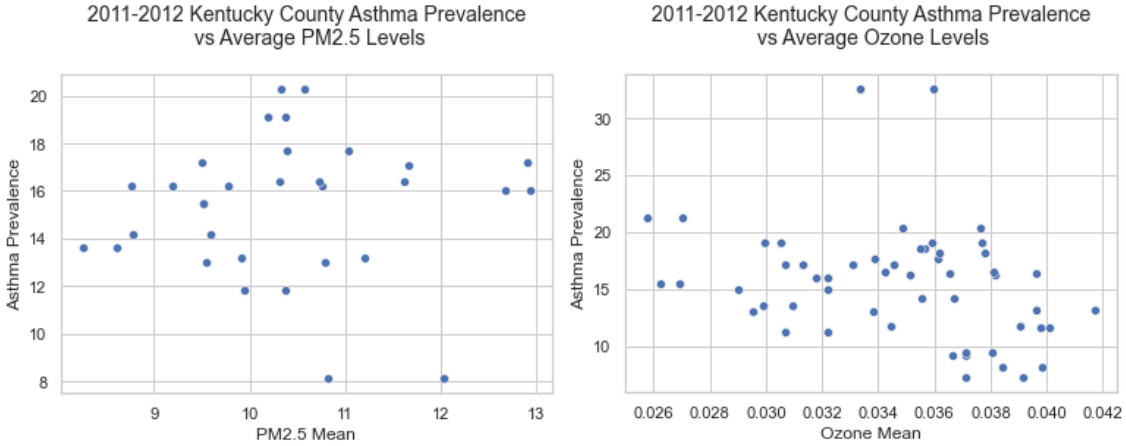

Figure 41 No relationship between average PM2.5 levels and adult asthma prevalence or ozone levels and adult asthma prevalence in Vermont Data Sources: EPA and Kentucky Department for Public Health using BRFSS

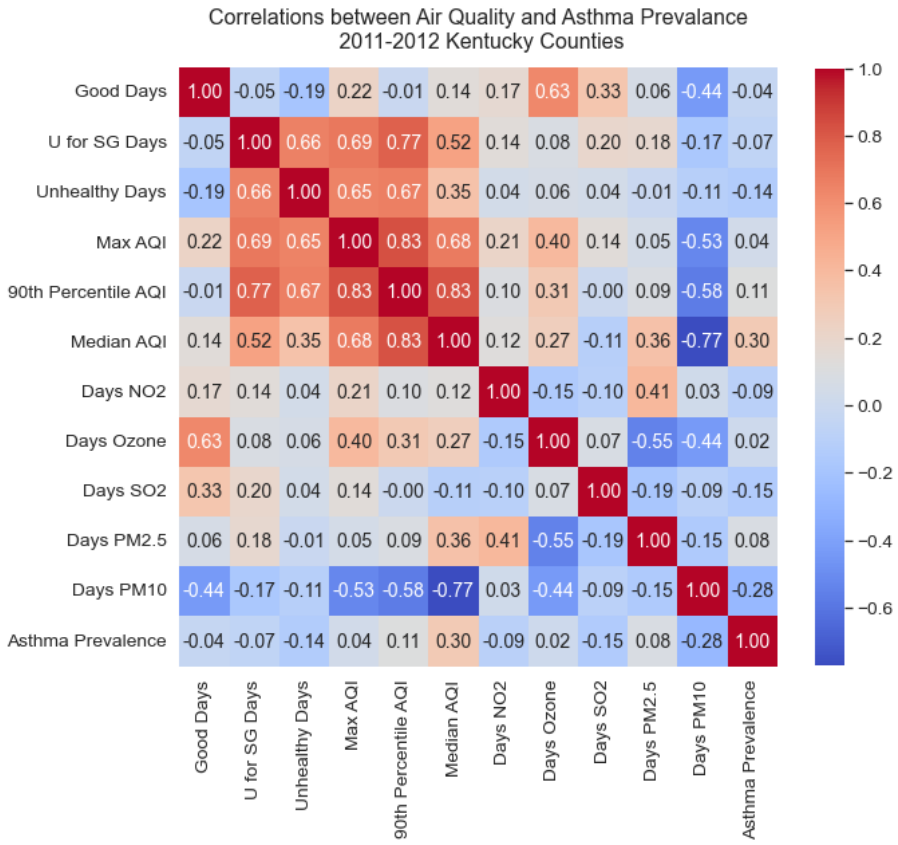

Figure 42 Correlations between air quality measures and adult asthma prevalence in Kentucky in 2012 U.S. Data Sources: EPA and Kentucky Department of Public Health using BRFSS

AQI and Asthma Studies, California

This section expands the AQI and asthma studies at the state level to California. The target question remains: What are ways to visualize the relationship between asthma and air quality?

In previous studies, California counties were found to have high PM2.5, ozone, and AQI indices. In addition, most counties in California have AQI monitoring sites. Unfortunately, there were no California asthma datasets by county for the years 1999, 2012, or 2020. California reports asthma in two-year periods, the most recent asthma data was for 2017-2018. For this study, I read in 2017 and 2018 daily PM2.5 datasets, daily ozone datasets, and annual AQI

datasets. I limited the datasets to California and followed procedures in previous studies for cleaning. Then I joined the datasets to explore relationships between air quality and asthma.

Maps of California counties were used to depict asthma prevalence (Fig 43), PM2.5 levels (Fig 44) and ozone concentrations (Fig 45).

The heatmap for California (Fig 46) shows low correlations between AQI measures and asthma prevalence.

The scatter plots for asthma vs PM2.5 and asthma vs ozone (Fig 47) show no relationship. The calculated correlation coefficients were 0.0997 and 0.1441, respectively.

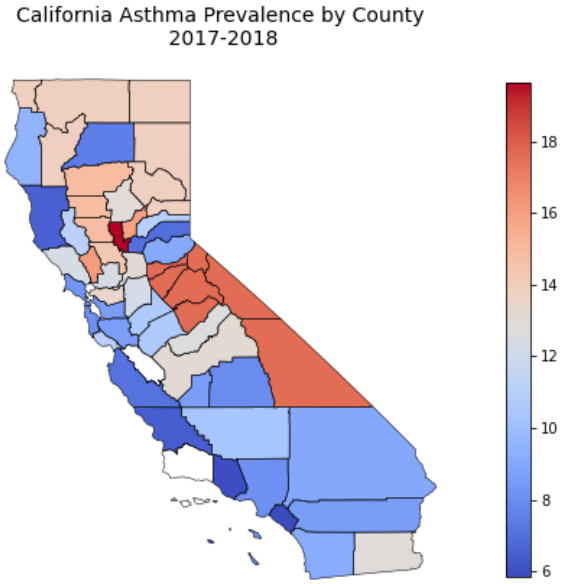

Figure 43 Percentage of California adults aged 18 and older with current asthma by county Data Source: California Department of Health using BRFSS data

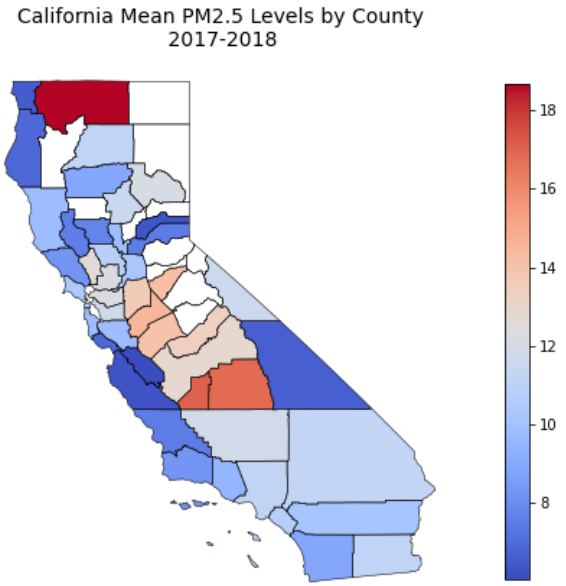

Figure 44 PM2.5 levels by counties in California. Counties without color are non-reporting counties Data Source: EPA

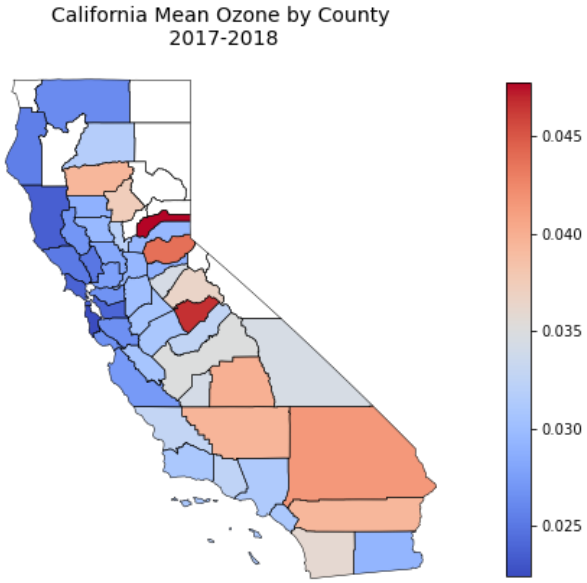

Figure 45 ozone levels by counties in California. Counties without color are non-reporting counties Data Source: EPA

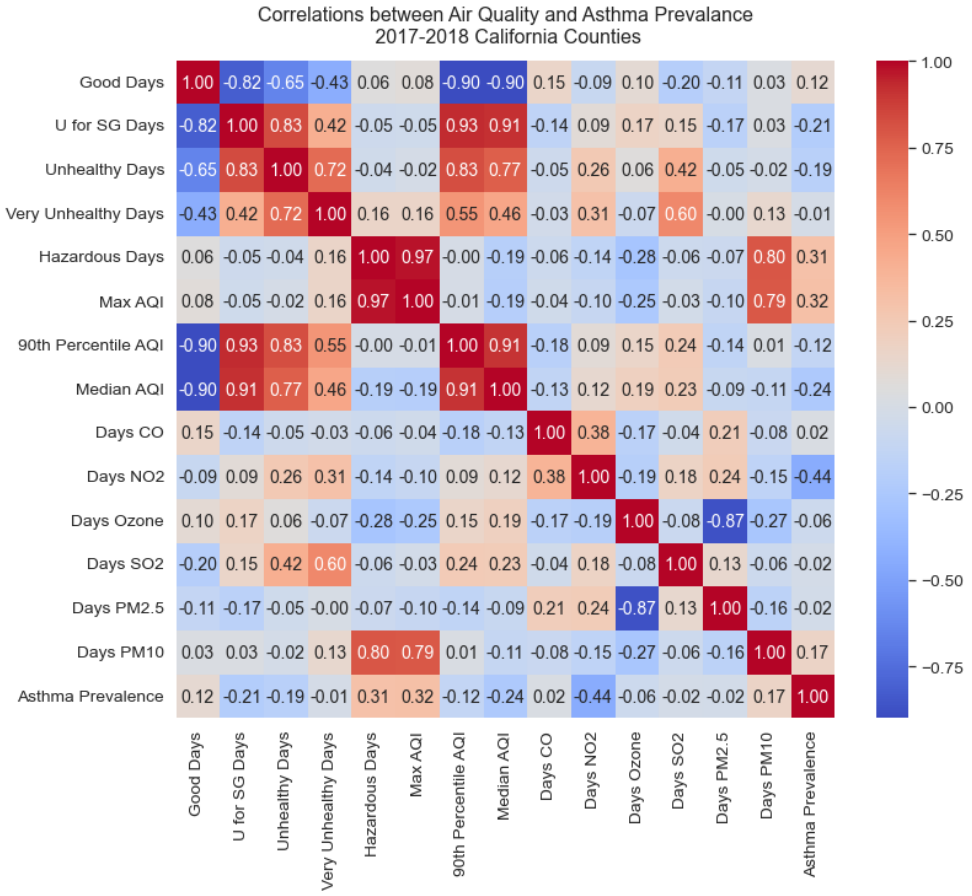

Figure 46 Correlations between air quality measures and adult asthma prevalence in California in 2017-2018 U.S. Data Sources: EPA and California Department of Public Health using BRFSS

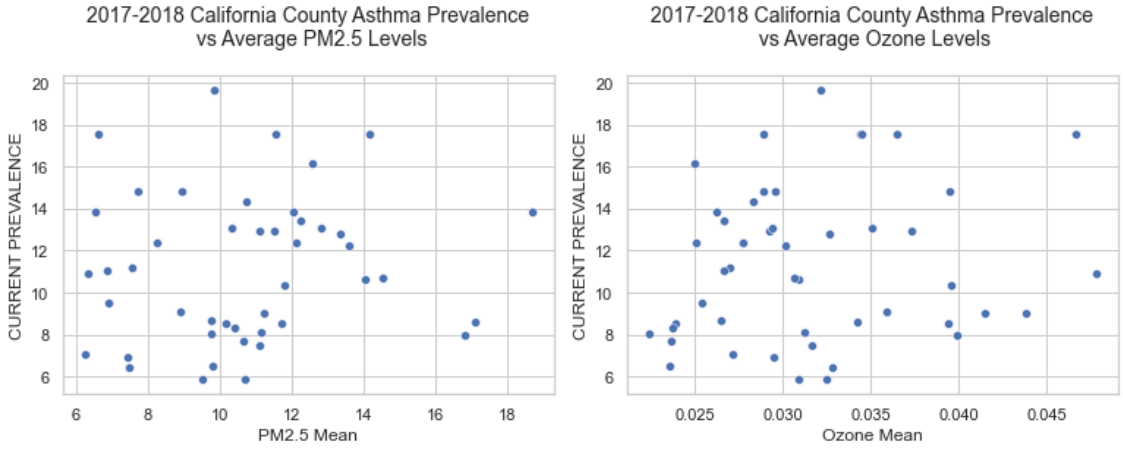

Figure 47 No relationship between average PM2.5 levels and adult asthma prevalence or ozone levels and adult asthma prevalence in California Data Sources: EPA and California Public Health Department using BRFSS

Lessons Learned

Creating maps of the United States using geospatial data is challenging. Many visualizations still show maps of the continental U.S. The Mercator projection is common, but it takes up a great deal of space if Alaska and Hawaii are included. The Albers USA projection is a better choice to visualize map data for the U.S. because it includes Alaska and Hawaii.

State level visualizations may not accurately depict the reality of local data. The absence of data may mask true relationships because if there is no data, no relationship can be examined.

Data that appears to show promise for a study may need to be examined more closely before using. Asthma prevalence comes from the BRFSS, which is a survey. The PM2.5, ozone, and AQI data do not cover all areas of the United States.

Air quality and asthma are both complex entities and the relationship between them may not be simple either.

Appendix A

Status Reports

First Status Report

Date: September 26, 2021

Accomplishments:

- The project GitHub repository was set up at https://github.com/mjlaw/FIT_capstone.
- The project proposal was uploaded to the GitHub repository.
- Air Quality (AQ) datasets were downloaded from the EPA AQS website to include daily summary data for PM2.5 for years 1999, 2012, and 2020; daily summary data for ozone for years 1999, 2012, and 2020; and annual summary for AQI by county for years 1999, 2012, and 2020.
- The datasets were uploaded to a dedicated folder for original datasets on the project's GitHub repository.
- Extracted information on content and format of the Annual Summary Files and the Daily Summary Files from https://aqs.epa.gov/aqsweb/airdata/FileFormats.html, saved the information to PDF documents, and uploaded the documents to the original datasets folder on the GitHub project repository.
- Read in 1999 and 2012 PM2.5 daily summary csv files, extracted arithmetic mean columns, and printed brief summaries in a Jupyter notebook file.
- Compared the summaries to Dr. Peng's summaries of 1999 and 2012 PM2.5 raw text files.
- Uploaded the Jupyter notebook file to the GitHub project repository.
- Uploaded the first status report to the GitHub project repository.

Current Activities: I am currently working on the PM2.5 AQ datasets as well as a synopsis for my GitHub project repository readme file. In addition, I am looking at a better way to incorporate textual content to my Jupyter Notebook files.

Challenges: Interpreting air quality datafiles is the most time-consuming area. Trying to reproduce research results using files with different parameters and numbers of observations is problematic.

Work to be Completed: For the next project milestone, I will complete the PM2.5 studies and visualizations, the ozone studies and visualizations, and the AQI studies and visualizations.

Progress Report 2

Date: October 10, 2021

Accomplishments:

- Updated GitHub project ReadMe File and began tracking local Jupyter notebook files to push to the remote repository for version control.
- Began using markdown for Jupyter notebook textual content in cells rather than the default code format.
- Read in the 2020 PM2.5 daily summary csv file, extracted the arithmetic mean column, and printed a summary in a Jupyter notebook file.
- Delved deeper into negative PM2.5 values by contacting the EPA via email.
- Uploaded the revised PM2.5 Jupyter notebook file to the GitHub project repository.
- Read in the 1999, 2012, and 2020 ozone daily summary csv files, extracted the first maximum value columns, and printed brief summaries in a Jupyter notebook file.
- Uploaded the new ozone Jupyter notebook file to the GitHub project repository
- Uploaded the progress report 2 to the GitHub project repository.

Current Activities: I am currently working on the PM2.5 and ozone AQ datasets and visualizations.

Challenges: Deciding how to move forward when I was unable to re-create similar boxplot visualizations to Dr. Peng's PM2.5 studies has been challenging. I looked more closely at all the datasets because of the negative values in the 2012 and 2020 datasets. The 1999 csv file used 24-hour sample durations only and had no negative values. In the 2012 and 2020 datasets, some sites reported two or three different values on a given day: 1-hour samples, 24-hour block average samples, and 24-hour samples. Re-reading the EPA's description of the data, multiple records for a monitor may be present if the sites report them. All negative values were either 1-hour samples or 24-hour block average samples. Due to correspondence from the EPA and their listed acceptable values at https://aqs.epa.gov/aqsweb/documents/about_aqs_data.html#_acceptable_values, I will not remove the negative values from the datasets as I was initially planning to do. I am planning to remove observations when sites report more than 1 value on a given day. How to prioritize which values to use and how to accomplish removal of multiple records for a monitor using Python has been the most time consuming.

Trying to reproduce visualizations using different programming languages is more complicated than I thought it would be. Deciding how best to visualize the data is also time consuming.

Work to be Completed: For the next project milestone, I will finish what was to be completed in this milestone: the PM2.5 studies and visualizations, the ozone visualizations, and the AQI studies and visualizations. I will also begin the asthma portion of the project by locating and cleaning the dataset.

Post Midterm Progress Report

Date: October 24, 2021

Accomplishments:

- Uploaded an EPA air quality sites csv file to the GitHub project repository. I joined this file to my datasets to obtain state names of state codes.
- Split the PM2.5 studies into two Jupyter notebook parts. The first part looks at the datasets for 1999 and 2012 and compares the results with Dr. Peng's study. The second part expands the study to a 2020 dataset. Cleaned 2012 and 2020 datasets for the second part. Removed excluded values under the events feature, removed all 1-hour samples, removed 24-hour blk average samples if they were present on the same site and on the same day as a 24-hour sample, removed observations from Mexico. (The 1999 dataset did not need cleaning.) Re-wrote and re-organized content within the Jupyter notebooks, to include the use of headings. Completed the PM2.5 study and visualizations for the questions: (1) How does the level of PM 2.5 in the U.S. compare between 1999, 2012, and 2020? (2) Which states have the highest and lowest levels of PM 2.5?
- Cleaned all datasets for the ozone studies. Removed all excluded values under the events feature. Removed all observations from Canada and Mexico. Completed the ozone study and visualizations for the questions: (1) How does the level of ozone in the U.S. compare between 1999, 2012, and 2020? (2) Which states have the highest and lowest levels of ozone?
- Started the AQI studies and visualizations. Completed initial analysis on four of the five target questions at the county level: (1) Which states have reports of hazardous air quality index (AQI)? (2) Which states have reports of very unhealthy AQI? (3) Which states have reports of unhealthy AQI? (4) Which states have reports of unhealthy for sensitive individuals AQI?
- Uploaded the revised PM2.5 Jupyter notebook file to the GitHub project repository.
- Uploaded the revised ozone Jupyter notebook file to the GitHub project repository.
- Uploaded the new PM2.5 part2 Jupyter notebook file to the GitHub project repository.
- Uploaded the new AQI studies Jupyter notebook file to the GitHub project repository.
- Uploaded the post midterm progress report to the GitHub project repository.
- Uploaded a CDC Chronic Disease Indicators: Asthma dataset to the GitHub project repository.

Current Activities: I am currently working on the AQI datasets and visualizations.

Challenges: Although I am combining values from different monitoring sites within a state into state values, I realize it is not as accurate as comparing data from the same site. Furthermore, some states are not represented at all. There are monitoring sites in U.S territories, and in Canada and Mexico. Cleaning data that I initially did not clean and then re-running the analyses on the cleaned data has been the most time consuming.

Work to be Completed: For the next project milestone, I will finish what was to be completed in this milestone: the AQI studies and visualizations and cleaning the asthma dataset. I will also finish the asthma exploratory analysis and visualizations.

Answers to Questions from Last Progress Report: You wrote: "I see the document reference on acceptable values. But I'm puzzled. If it's a concentration or quantity, then a negative value is not possible (unless it's relative to a baseline)."

The email response I received from the EPA was: "Thanks for your message and reaching out. We have a short write up on acceptable values in EPA's Air Quality System (AQS) here: https://aqs.epa.gov/aqsweb/documents/about_aqs_data.html#_acceptable_values, but since that explanation is very short I can elaborate.

Every instrument has an allowable uncertainty, and occasionally as you've noted monitors can yield small negative hourly values. Say it's +/- 10 ppb for whatever substance. If the instrument reads 100 ppb, that means the real concentration will be somewhere between 90 and 110. If the instrument reads -3 that means the real value can be anywhere between 0 and 7 (negative concentrations not being possible). We allow reporting of the negative values to capture valid, quality assured readings that are valid members of the sample set. With PM2.5 monitors, negative hourly concentrations for PM2.5 down to -4.99 ug/m3 (the default QC range check) are used in computing 24-hour averages so as not to bias that average."

I had included the email excerpt in the Jupyter Notebook PM2.5 studies, but I will include it in the final report, as well.

Second Half Progress Report 2

Date: November 7, 2021

Accomplishments:

- Created a new asthma studies Jupyter notebook file and read in the CDC Chronic Disease Indicators Asthma dataset. Although the dataset website https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Asthma/us8e-ubyj reports it was last updated in February, 2021, the most recent year for which there are entries is 2019. 2012 is the only year in common with the air quality studies I've completed. Therefore, I will limit the asthma studies to 2012.
- Completed exploratory analysis of the asthma dataset.
- Cleaned the CDC Chronic Disease Indicators: Asthma dataset. Removed observations from the District of Columbia, Guam, Puerto Rico, the Virgin Islands, and the United States. Narrowed the dataset to 2012 observations and to one question: Current asthma prevalence among adults aged 18 and older. Further narrowed the dataset to overall incidence (not stratified by gender or race) and used the crude prevalence rather than the age adjusted prevalence.
- Completed analysis on the asthma studies target questions: (1) What does the reported incidence of asthma look like by state? (2) Which states have the highest and lowest reported incidence of asthma.

- Searched for 2012 asthma datasets for states with the highest reported prevalence of asthma. Located asthma datasets for Kentucky (tied with Maine for highest) and Vermont (number 3). The state asthma datasets provide granularity at the county level, although the reporting year is not limited to 2012 in either dataset. Kentucky reports asthma in two or three-year periods. Options are 2011-2012 and 2011-2013. Vermont reports asthma for two-year periods. Options are 2011-2012 and 2012-2013.
- Uploaded the asthma Jupyter notebook file to the GitHub project repository.
- Uploaded the Kentucky and Vermont asthma datasets to the GitHub project repository.
- Uploaded the Second Half Progress Report 2 to the GitHub project repository.

Current Activities: I am currently working on air quality and asthma visualizations with U.S. maps.

Challenges: Searching for appropriate asthma datasets has been the most time consuming. Creating visualizations with U.S. maps using Python is very challenging, and although I am confident that I can do those visualizations with Tableau, I want to do them in Python first.

Work to be Completed: For the next project milestone, I will finish the AQI and asthma visualizations, join the AQ and asthma datasets, and complete the exploratory analysis of the combined datasets.

Last Interim Progress Report

Date: November 21, 2021

Accomplishments:

- Reworked PM2.5 part 2 studies, ozone studies, and AQI studies, and uploaded to the GitHub project repository.
- Installed GeoPandas, an open-source project for geospatial data, and began plotting maps. GeoPandas is available from https://geopandas.org
- Downloaded cartographic boundary shapefiles for the United States and for U.S. counties from the United States Census Bureau. These shapefiles are used in creating maps with GeoPandas. The source is https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
- Added state studies for Vermont and Kentucky to the asthma studies. Cleaned the datasets and completed state exploratory analyses and preliminary map visualizations of asthma incidence by county on state maps of Vermont and Kentucky. Completed a preliminary visualization on the reported incidence of asthma by state (U.S. map).
- Uploaded the modified asthma Jupyter notebook file to the GitHub project repository.
- Uploaded the Last Interim Progress Report to the GitHub project repository.

Current Activities: I am currently working on air quality and asthma visualizations with U.S. maps and joining AQ and asthma datasets.

Challenges: Creating visualizations with U.S. maps using GeoPandas is very challenging. I've made progress; however, I will need to spend more time working on them to obtain better results. I was unable to upload several cleaned datasets to the GitHub project repository because the files were too large. I'll try zipping them, first. If that doesn't work, I'll look at the large file storage options in GitHub.

Work to be Completed: For the next project milestone, I will complete all remaining visualizations and complete the asthma and air quality studies.

Final Progress Report

Date: November 28, 2021

Accomplishments:

- Reworked several of the visualizations in PM2.5, PM2.5 part 2, Ozone, AQI, and Asthma studies and uploaded the modified Jupyter notebook files to the GitHub project repository at https://github.com/mjlaw/FIT_capstone.
- Completed the analyses of air quality and asthma at the U.S. state level for 2012. Air quality features were mean PM2.5, mean ozone, median AQI, max AQI, days CO, days NO2, days Ozone, days SO2, days PM2.5, days PM10, unhealthy for sensitive group days, unhealthy days, very unhealthy days, and hazardous days. Created a correlation matrix with a heatmap and scatterplots to visualize the relationship between air quality and asthma. Uploaded the Air Quality and Asthma Jupyter notebook to the GitHub project repository.
- Completed the analyses of air quality and asthma at the county level for two states with a high prevalence of asthma in 2012—Vermont and Kentucky. These states report asthma in two- or three-year periods. The asthma datasets I used were for 2011-2012. Therefore, I cleaned daily PM2.5 datasets, daily ozone datasets, and annual AQI by county datasets for 2011 to add to the previously cleaned and saved 2012 datasets. Air quality features for Kentucky were mean PM2.5, mean ozone, good days, unhealthy for sensitive group days, unhealthy days, max AQI, median AQI, days NO2, days Ozone, and Days SO2. (Other features were not included because the values were all 0 or a max of 1.) Created a correlation matrix with a heatmap and a scatterplot. Air quality features for Vermont were good days, unhealthy for sensitive group days, max AQI, median AQI, days CO, days NO2, days Ozone, and days SO2. Uploaded the Air Quality and Asthma part2 Jupyter notebook to the GitHub project repository.
- Uploaded zip files for annual AQI by county for years 2011, 2017, and 2018 and for daily PM2.5 and daily Ozone for years 2011, 2017, and 2018 to the GitHub project repository Original datasets folder.
- Created a new Jupyter notebook for Air Quality and Asthma part3, which will examine the relationship between air quality and asthma in California. Read in daily PM2.5, daily ozone, and annual AQI files for years 2017 and 2018. Limited the datasets to California. Cleaned the PM2.5 and ozone datasets. Uploaded the Air Quality and Asthma part3 Jupyter notebook to the GitHub project repository.
- Uploaded the Final Progress Report to the GitHub project repository.

Current Activities: I am currently working on a new study on air quality and asthma in California and finalizing visualizations.

Challenges: Working with limited data is challenging. Vermont had three counties with PM2.5 data, two counties with ozone data, and four counties with AQI data. I did not use PM2.5 data or ozone data in the analysis because there weren't enough observations. The correlation matrices and heatmaps created thus far show very little correlation between any of the air quality features and asthma prevalence. Although I had initially intended to limit the scope of the study to specific years—1999, 2012, and 2020 , I added the 2011 datasets because states report asthma in two-year periods. I had also planned to look at asthma prevalence in a state or states having hazardous days. I could not find asthma reports for those states in my target years. After further research into states with the largest numbers of counties reporting air quality data, I decided to examine the relationship between air quality and asthma in California. The most recent reports for asthma in California are for 2017-2018, necessitating the download and cleaning of additional datasets. The map visualizations are taking the largest amount of time and are the most challenging for me. I'm working on adding layers to maps using geopandas and modifying the color/shape/size of color bars.

Work to be Completed: All remaining visualizations, the California air quality and asthma study, and the final project report.

Appendix B

Source Code

The source code for this project consists of eight Jupyter notebook .pynb files. They are zipped and submitted as a separate attachment.

References

AirNow. (n.d. -a). *Air Quality Index (AQI) Basics.* https://www.airnow.gov/aqi/aqi-basics/

AirNow. (n.d. -b). *Using the Air Quality Index.* https://www.airnow.gov/aqi/aqi-basics/using-air-quality-index/

California Health and Human Services Agency. (n.d.). *BRFSS Data-Adult Asthma Prevalence 2012 to Present.* [Data set]. https://data.chhs.ca.gov/dataset/asthma-prevalence/resource/4dc6e239-97b0-475b-94b6-bec1a7f87729

Centers for Disease Control and Prevention. (2021). *U.S. Chronic Disease Indicators: Asthma.* [Data set]. https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Asthma/us8e-ubyj

Cutrer, J. (2020, July 28). *GeoPandas Tutorial: How to plot US Maps in Python.* https://jcutrer.com/python/learn-geopandas-plotting-usmaps

Cutrer, J. (2020, May 6). *GeoPandas Tutorial: Learn geopandas by plotting tornados on a map.* https://jcutrer.com/python/learn-geopandas-plotting-tornados

Dausch, M. (2021, October 28). 211028.ipynb [Jupyter notebook]. In M. Dausch, *Topics in Computer Information Systems*. Florida Institute of Technology. https://fit.instructure.com/courses/593004/files

Environmental Protection Agency. (n.d. -a) *Annual AQI by County*, *1999* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

Environmental Protection Agency. (n.d. -b) *Annual AQI by County*, *2011* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

Environmental Protection Agency. (n.d. -c) *Annual AQI by County*, *2012* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

Environmental Protection Agency. (n.d. -d) *Annual AQI by County*, *2017* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

Environmental Protection Agency. (n.d. -e) *Annual AQI by County*, *2018* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

Environmental Protection Agency. (n.d. -f) *Annual AQI by County*, *2020* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual

Environmental Protection Agency. (n.d. -g). *Asthma and Outdoor Air Pollution, EPA-452-F-04-002.* https://www.airnow.gov/sites/default/files/2018-03/asthma-flyer.pdf

Environmental Protection Agency. (n.d. -h). *Daily Ozone (44201), 1999*. [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -i). *Daily Ozone (44201), 2011*. [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d.-j). *Daily Ozone (44201), 2012*. [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -k). *Daily Ozone (44201), 2017*. [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -l). *Daily Ozone (44201), 2018*. [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -m). *Daily Ozone (44201), 2020*. [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -n). *Daily PM2.5 (88101), 1999*. [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -o). *Daily PM2.5. (88101), 2011* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -p). *Daily PM2.5. (88101), 2012* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -q). *Daily PM2.5. (88101), 2017* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -r). *Daily PM2.5. (88101), 2018* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (n.d. -s). *Daily PM2.5. (88101), 2020* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily

Environmental Protection Agency. (2019). *About AQS Data.* https://aqs.epa.gov/aqsweb/documents/about_aqs_data.html#

Environmental Protection Agency. (2021). *NAAQS Table.* https://www.epa.gov/criteria-air-pollutants/naaqs-table

Foundation for a Healthy Kentucky. (n.d.). *Chronic Disease Indicator, Prevalence of Asthma (percent adults), 2011 -2012.* [Data set]. https://perma.cc/RRV4-84JH

Peng, R. D. (2020). Data Analysis Case Study: Changes in Fine Particle Air Pollution in the U.S. In *R programming for data science*. Leanpub. https://leanpub.com/rprogramming

Prabhakaran, S. (2018, November 28). *Top 50 matplotlib Visualizations – The Master Plots (with full python code).* https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

United States Census Bureau. (2021a). Cartographic Boundary Files – Shapefile. (cb_2014_us_county_20m.zip) [Dataset]. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.2014.html

United States Census Bureau. (2021b). Cartographic Boundary Files – Shapefile. (cb_2014_us_state_20m.zip) [Dataset]. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.2014.html

Vermont Department of Health. (n.d.). Chronic Diseases Community Profile – County [Dataset] https://apps.health.vermont.gov/ias/querytool?topic=HV2020CommunityProfiles&geo=71&date=2012&theme1=HV2020ChronicDiseasesCommunityProfile&tab=DataViewTabular&go=1

World Health Organization. (2021). *Asthma Factsheet*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/asthma.