Analysis of Air Quality in the United States

DS 5110 Spring 2020
Term Project
14 April, 2020

Carson Cook
Erin Keough
Paola Reyes
Aditya Samaroo
Priyank Shelat

Where does the data come from?

- Air Quality System (AQS) DataMart from the EPA
- Measurements of criteria pollutants
  - Ambient air pollution measurements from thousands of state and local monitoring locations
  - Particulate matter, sulfur dioxide, carbon monoxide, nitrogen dioxide, ozone, air toxics, lead
  - Hourly, 8-hour, daily, and annual concentrations
- Used by research, regulatory, and healthcare research communities
  - Clean Air Act compliance

https://www.epa.gov

Where does the data come from?

CDC

CENTERS FOR DISEASE CONTROL AND PREVENTION

https://www.cdc.gov

- National Environmental Public Health Data tracking program from the CDC
- Data from a network of partners
  - National organizations, federal agencies, and fellowship participants
- Health and environmental data from city, state, and national sources
  - Air pollutant concentrations at county- and state-level resolutions
  - Prevalence of health conditions at county- and state-level resolutions
- Data available from 1999 - 2016

A look at the EPA data

```
  state_code         county_code          site_num         parameter_code        poc            latitude        longitude
 Length:1624705     Length:1624705     Length:1624705     Min.   :11101   Min.   : 1.000   Min.   :17.70   Min.   :-162.97
 Class :character   Class :character   Class :character   1st Qu.:43811   1st Qu.: 1.000   1st Qu.:34.00   1st Qu.:-109.10
 Mode  :character   Mode  :character   Mode  :character   Median :62106   Median : 1.000   Median :38.86   Median : -90.60
                                                          Mean   :62846   Mean   : 2.713   Mean   :38.07   Mean   : -94.72
                                                          3rd Qu.:88102   3rd Qu.: 5.000   3rd Qu.:41.82   3rd Qu.: -81.49
                                                          Max.   :92142   Max.   :99.000   Max.   :68.64   Max.   : -64.75

    datum           parameter_name     sample_duration    pollutant_standard metric_used        method_name             year
 Length:1624705     Length:1624705     Length:1624705     Length:1624705     Length:1624705     Length:1624705     Min.   :2000
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:2004
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2009
                                                                                                                   Mean   :2009
                                                                                                                   3rd Qu.:2014
                                                                                                                   Max.   :2019

 units_of_measure    event_type        observation_count observation_percent completeness_indicator valid_day_count required_day_count
 Length:1624705     Length:1624705     Min.   :     1    Min.   :  0.00      Length:1624705         Min.   :  0.0   Min.   :  0.0
 Class :character   Class :character   1st Qu.:    54    1st Qu.: 80.00      Class :character       1st Qu.: 50.0   1st Qu.: 61.0
 Mode  :character   Mode  :character   Median :    97    Median : 95.00      Mode  :character       Median : 60.0   Median : 61.0
                                       Mean   :  1677    Mean   : 82.74                             Mean   :118.9   Mean   :145.9
                                       3rd Qu.:   352    3rd Qu.: 99.00                             3rd Qu.:122.0   3rd Qu.:214.0
                                       Max.   :105025    Max.   :100.00                             Max.   :366.0   Max.   :366.0

 exceptional_data_count null_data_count     primary_exceedance_count secondary_exceedance_count certification_indicator num_obs_below_mdl
 Min.   :     0.00      Min.   :     0.00   Min.   :  0.0            Min.   :  0.0              Length:1624705          Min.   :0
 1st Qu.:     0.00      1st Qu.:     0.00   1st Qu.:  0.0            1st Qu.:  0.0              Class :character        1st Qu.:0
 Median :     0.00      Median :     2.00   Median :  0.0            Median :  0.0              Mode  :character        Median :0
 Mean   :    30.89      Mean   :    69.48   Mean   :  2.6            Mean   :  2.6                                      Mean   :0
 3rd Qu.:     0.00      3rd Qu.:     9.00   3rd Qu.:  1.0            3rd Qu.:  1.0                                      3rd Qu.:0
 Max.   :102471.00      Max.   :105096.00   Max.   :327.0            Max.   :302.0                                      Max.   :0
                                            NA's   :1375186          NA's   :1394666
 arithmetic_mean    arithmetic_standard_dev first_max_value    first_max_datetime             second_max_value    second_max_datetime
 Min.   : -58.00    Min.   :    0.00        Min.   :   -58.0   Min.   :2000-01-01 00:00:00   Min.   :   -38.30   Min.   :2000-01-01 00:00:00
 1st Qu.:   0.03    1st Qu.:    0.01        1st Qu.:     0.1   1st Qu.:2004-10-15 00:00:00   1st Qu.:     0.07   1st Qu.:2004-10-06 00:00:00
 Median :   0.41    Median :    0.30        Median :     1.9   Median :2009-02-05 09:00:00   Median :     1.45   Median :2009-02-06 00:00:00
 Mean   :  54.09    Mean   :    5.99        Mean   :    88.6   Mean   :2009-06-12 10:38:06   Mean   :    81.14   Mean   :2009-06-12 12:07:59
 3rd Qu.:   8.64    3rd Qu.:    4.52        3rd Qu.:    30.1   3rd Qu.:2014-01-11 00:00:00   3rd Qu.:    26.40   3rd Qu.:2014-01-11 00:00:00
 Max.   :56337.82   Max.   :45501.15        Max.   :336583.0   Max.   :2019-10-29 18:00:00   Max.   :304500.00   Max.   :2019-10-31 07:00:00
 NA's   :5          NA's   :5               NA's   :3          NA's   :3                     NA's   :8047        NA's   :8047
 third_max_value     third_max_datetime            fourth_max_value    fourth_max_datetime           first_max_non_overlapping_value
 Min.   :   -24.20   Min.   :2000-01-01 00:00:00   Min.   :   -28.70   Min.   :2000-01-01 00:00:00   Min.   : 0.0
 1st Qu.:     0.07   1st Qu.:2004-10-06 00:00:00   1st Qu.:     0.06   1st Qu.:2004-10-06 00:00:00   1st Qu.: 1.2
 Median :     1.26   Median :2009-02-12 00:00:00   Median :     1.13   Median :2009-02-15 00:00:00   Median : 1.9
 Mean   :    78.13   Mean   :2009-06-15 07:35:16   Mean   :    76.36   Mean   :2009-06-16 16:54:25   Mean   : 2.4
 3rd Qu.:    24.40   3rd Qu.:2014-01-17 00:00:00   3rd Qu.:    24.00   3rd Qu.:2014-01-23 00:00:00   3rd Qu.: 2.9
 Max.   :299667.00   Max.   :2019-10-31 06:00:00   Max.   :294333.00   Max.   :2019-10-29 03:00:00   Max.   :45.7
 NA's   :14096       NA's   :14096                 NA's   :20880       NA's   :20880                 NA's   :1616863
 first_no_max_datetime         second_max_non_overlapping_value second_no_max_datetime        ninety_nine_percentile ninety_eight_percentile
 Min.   :2000-01-01 05:00:00   Min.   : 0.0                     Min.   :2000-01-01 05:00:00   Min.   :   -58.00      Min.   :   -58.00
 1st Qu.:2003-11-30 08:45:00   1st Qu.: 1.1                     1st Qu.:2003-11-24 17:00:00   1st Qu.:     0.08      1st Qu.:     0.07
 Median :2008-01-12 00:30:00   Median : 1.7                     Median :2008-01-12 00:00:00   Median :     1.60      Median :     1.31
 Mean   :2008-10-16 04:57:49   Mean   : 2.1                     Mean   :2008-10-16 11:28:30   Mean   :    71.49      Mean   :    68.04
 3rd Qu.:2013-11-29 06:00:00   3rd Qu.: 2.6                     3rd Qu.:2013-11-28 00:00:00   3rd Qu.:    24.90      3rd Qu.:    23.40
 Max.   :2019-10-26 01:00:00   Max.   :45.1                     Max.   :2019-10-24 23:00:00   Max.   :227583.00      Max.   :198250.00
 NA's   :1616863               NA's   :1616864                  NA's   :1616864
 ninety_five_percentile ninety_percentile   seventy_five_percentile fifty_percentile   ten_percentile     local_site_name
 Min.   :   -58.00      Min.   :   -58.00   Min.   :  -58.00        Min.   :  -58.00   Min.   :  -99.00   Length:1624705
 1st Qu.:     0.06      1st Qu.:     0.05   1st Qu.:    0.03        1st Qu.:    0.00   1st Qu.:    0.00   Class :character
 Median :     1.02      Median :     0.80   Median :    0.50        Median :    0.30   Median :    0.05   Mode  :character
 Mean   :    63.94      Mean   :    60.89   Mean   :   56.89        Mean   :   53.33   Mean   :   48.19
 3rd Qu.:    19.60      3rd Qu.:    16.00   3rd Qu.:   11.10        3rd Qu.:    7.30   3rd Qu.:    2.61
 Max.   :149092.00      Max.   :110500.00   Max.   :75950.00        Max.   :44583.00   Max.   :34236.50

   address           state_name         county_name         city_name          cbsa_name         date_of_last_change
 Length:1624705     Length:1624705     Length:1624705     Length:1624705     Length:1624705     Min.   :2010-03-05
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:2013-01-02
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2016-03-04
                                                                                                Mean   :2015-05-21
                                                                                                3rd Qu.:2017-06-06
                                                                                                Max.   :2019-11-13
```

- 1.6 million rows of annual summary data
- Data can be roughly broken into:
  - Station information
  - Parameter information
  - Observation information

How is the EPA data collected?

- 5624 air monitoring locations in the contiguous US
- 1052 distinct parameters measured and recorded
- Air monitoring locations can measure one or more parameters

https://www.mdaqmd.ca.gov/permitting/compliance

Observations from raw EPA data

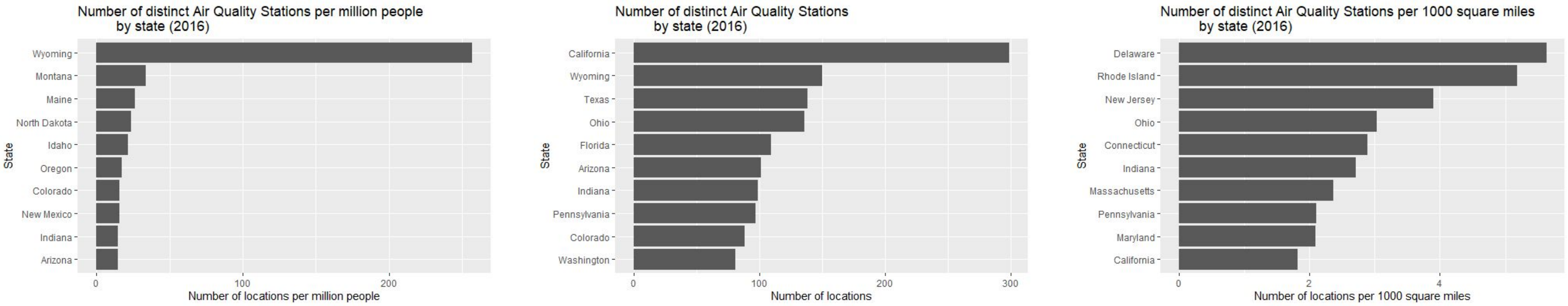


Figure 1. Number of Air Quality Stations per state does not seem to be dependent on population nor square mileage

Observations from raw EPA data

Any Air Quality Station

PM2.5 Air Quality Stations

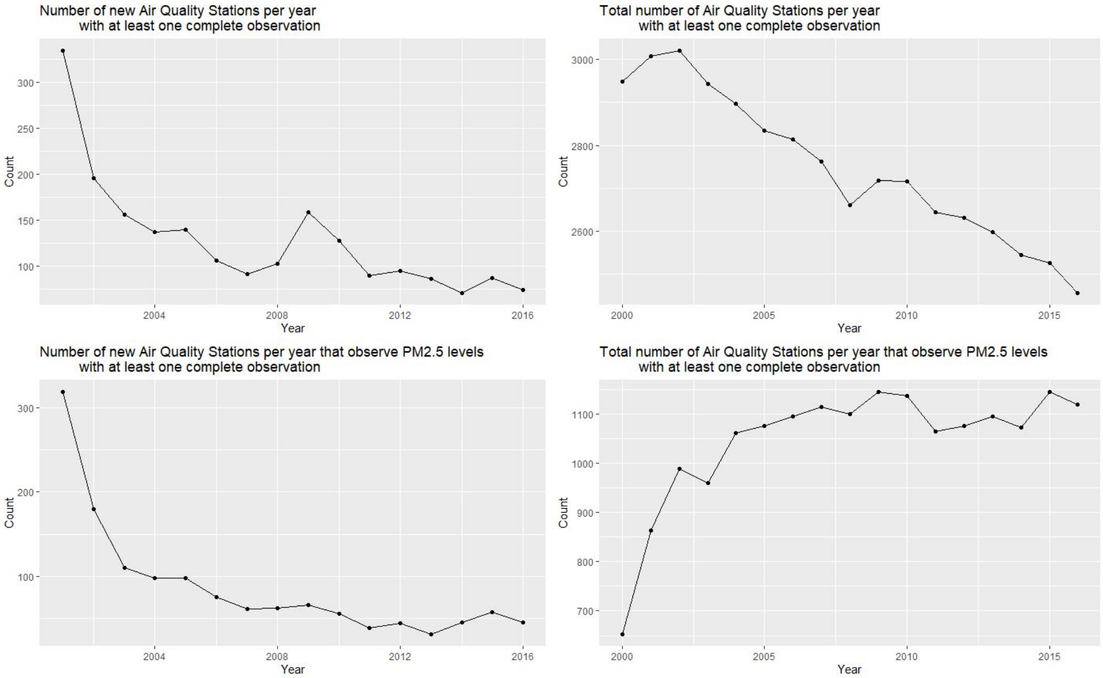


Figure 2. Initial analysis of the AQS dataset from the EPA.

Data Pre-Processing

Selection	Preprocessing	Transformation

Selection

- Filtered out variables with more than 50% missing values.
- Removed duplicates.

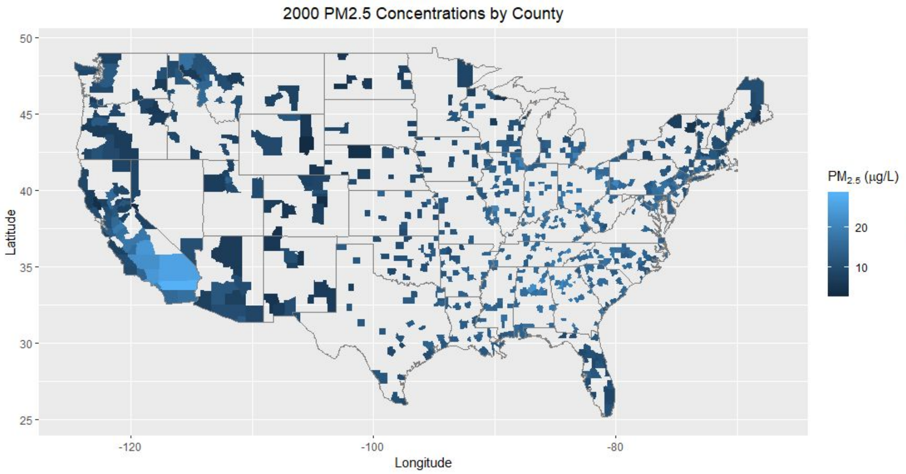
Preprocessing

- Tidy dataset: pivoted variables.
- Merged datasets: EPA + CDC
- Partition Dataset: 80% training 20% testing

Transformation

- Imputed missing values.
- Log-transformed both target and explanatory variables.

Exploratory Data Analysis - $PM_{2.5}$



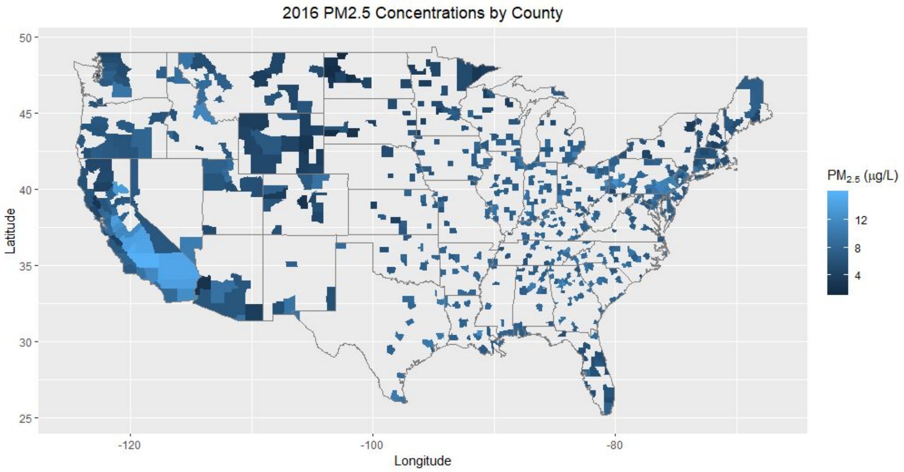


Figure 3. Comparison of particulate matter concentrations by county in 2000 and 2016.

Exploratory Data Analysis - $PM_{2.5}$

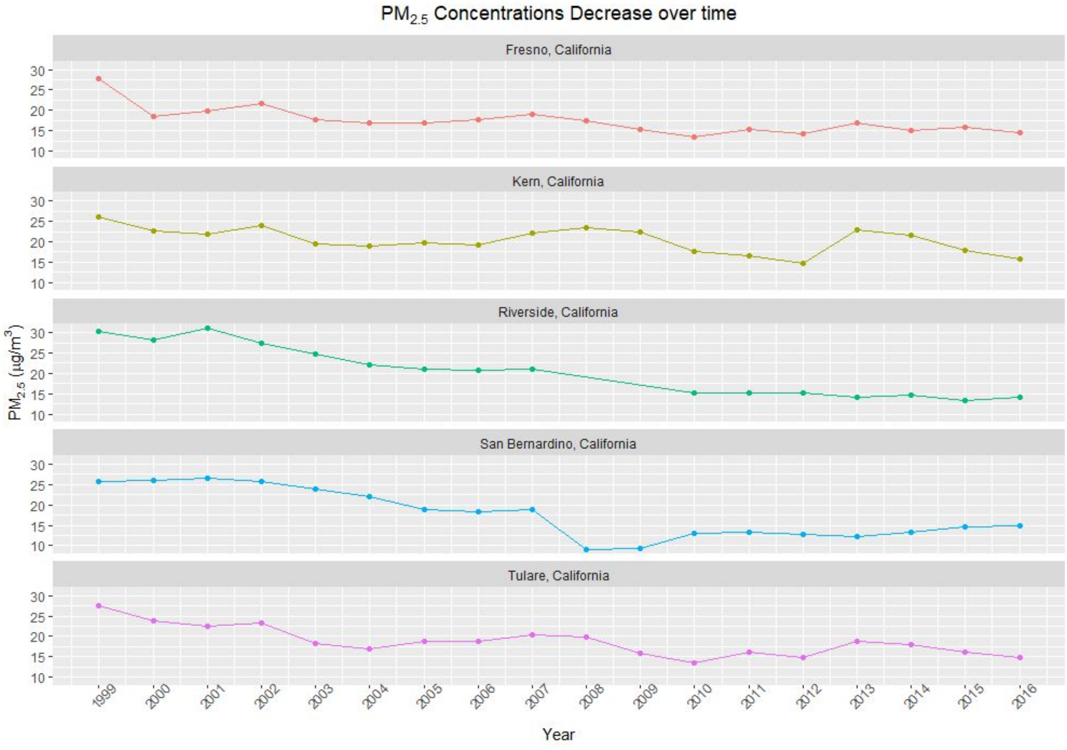


Figure 4. $PM_{2.5}$ concentrations in the areas with the highest overall concentrations from 1999 to 2016.

Table 1. Average, minimum, and maximum concentrations for each of the top 5 counties.

CountyID	Average	Maximum	Minimum
Riverside, California	20.53750	31.0	13.4
Kern, California	20.41667	26.2	14.7
Tulare, California	18.79444	27.6	13.6
San Bernardino, California	17.67222	26.5	9.1
Fresno, California	17.43333	27.7	13.4

Exploratory Data Analysis - $PM_{2.5}$ and Ozone

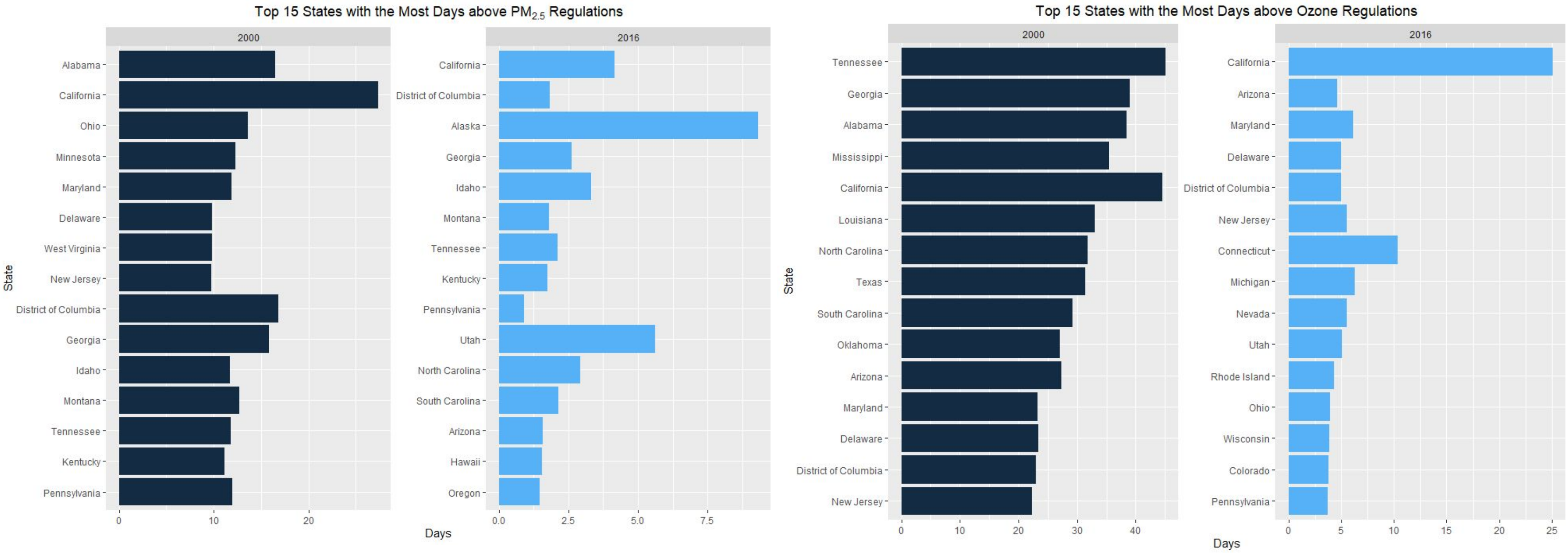
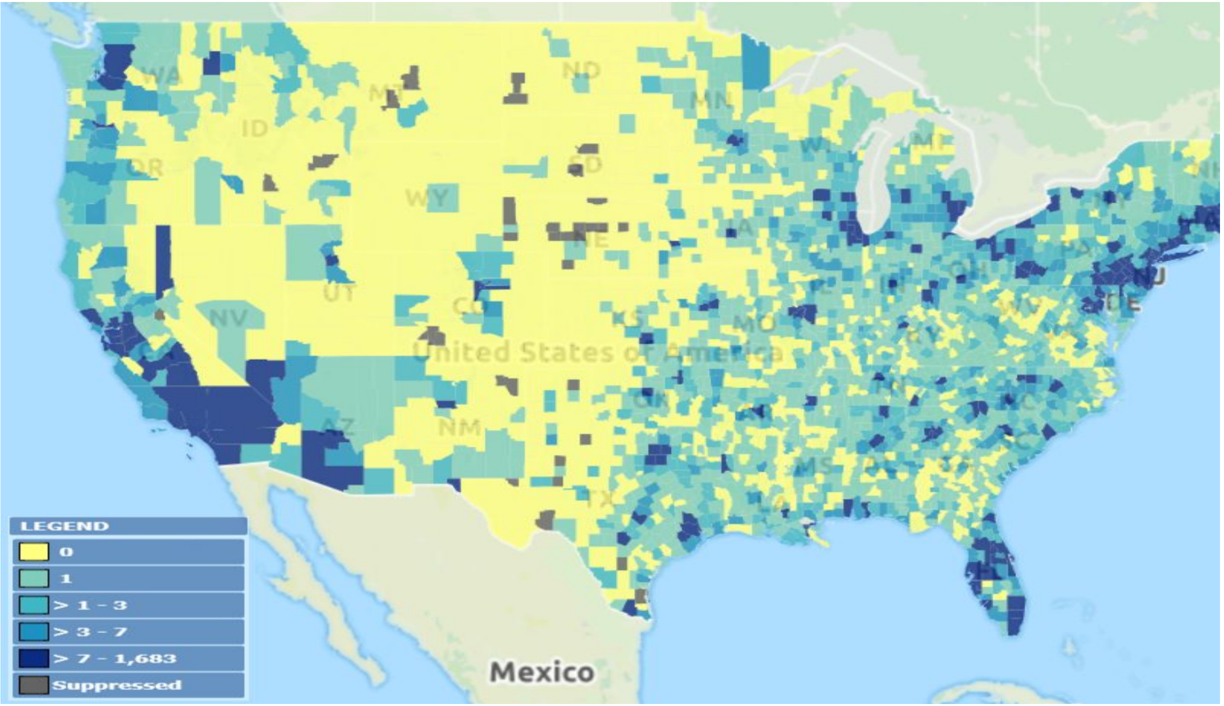


Figure 5. States with the most days above $PM_{2.5}$ and Ozone regulation in 2000 and 2016.

Exploratory Data Analysis - $PM_{2.5}$

Reduction Percent of PM 2.5 <chr>	Minimum Mortality Benefit <dbl>	Maximum Mortality Benefit <dbl>	Average Mortality Benefit <dbl>	Median Mortality Benefit <dbl>
Reduction Percent: 10% Reduction	0	679	4.785960	1
Reduction Percent: 15% Reduction	0	958	7.155908	2
Reduction Percent: 20% Reduction	0	972	9.435453	3
Reduction Percent: 25% Reduction	0	942	11.599772	4
Reduction Percent: 5% Reduction	0	341	2.366039	1

Table 2. $PM_{2.5}$ Mortality Benefit Table.



Deaths Avoided by 5% Reduction in PM 2.5

Figure 6. $PM_{2.5}$ Mortality Benefit Map.

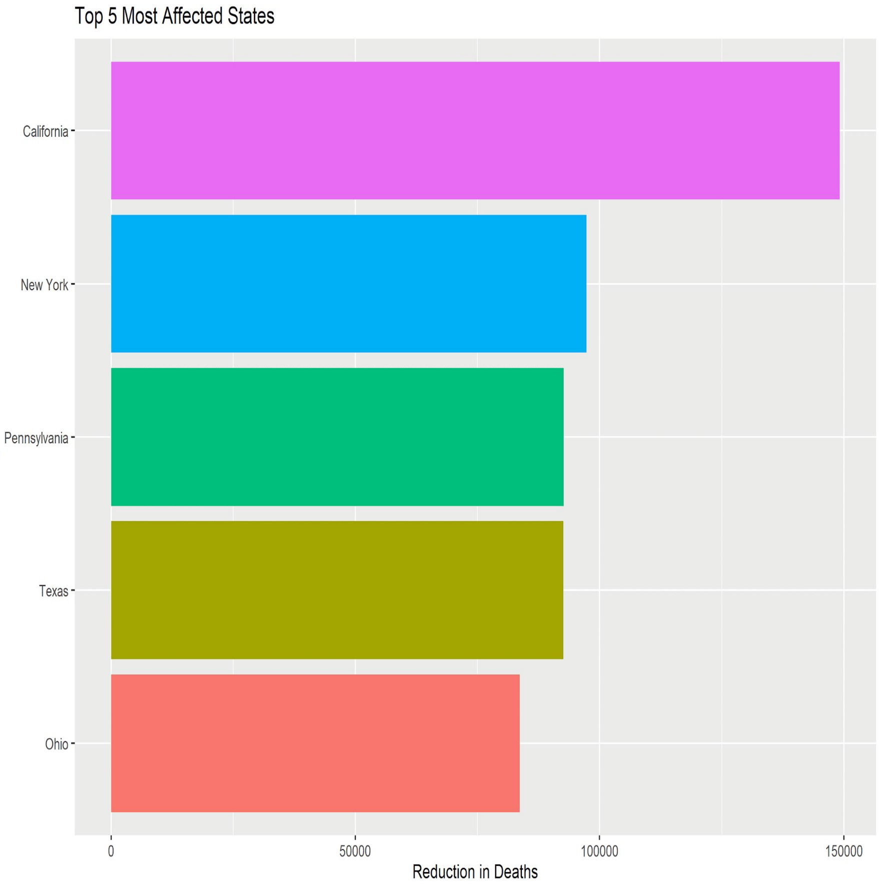


Figure 7. Top states impacted by reducing $PM_{2.5}$ concentrations.

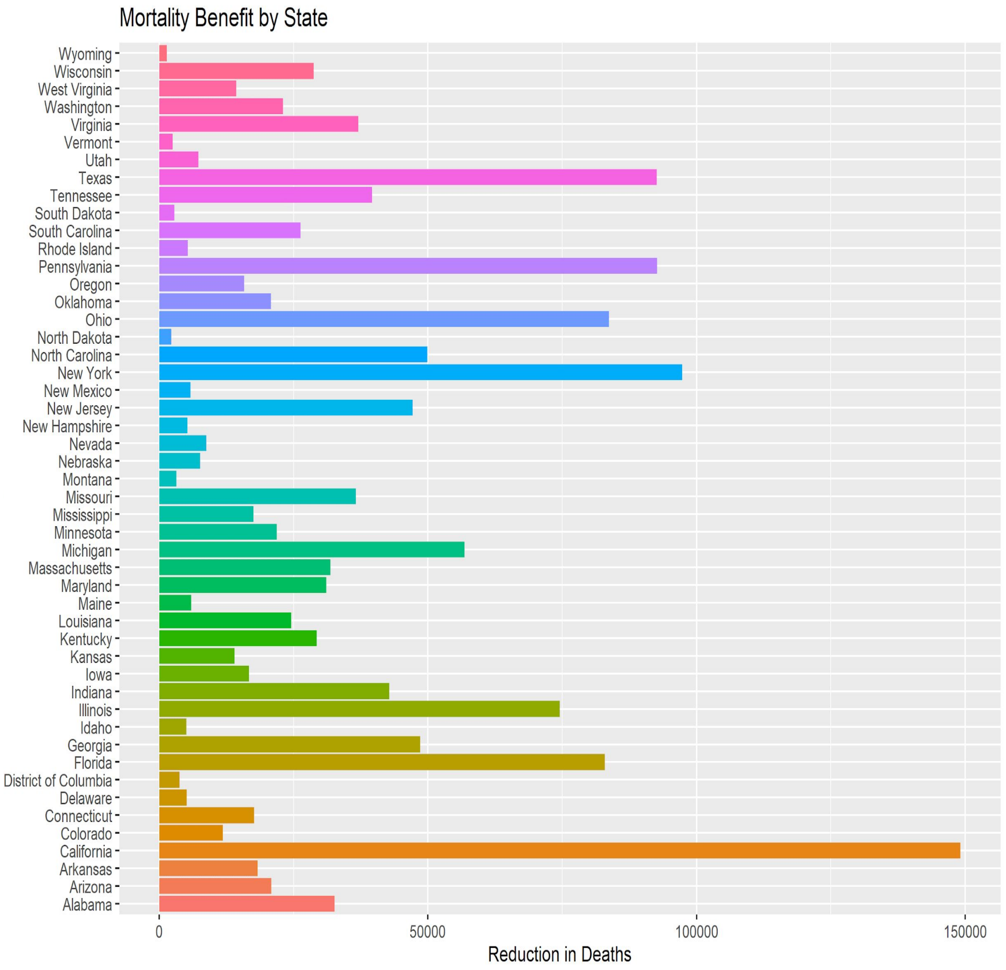


Figure 8. Benefits of reducing $PM_{2.5}$ concentrations by state.

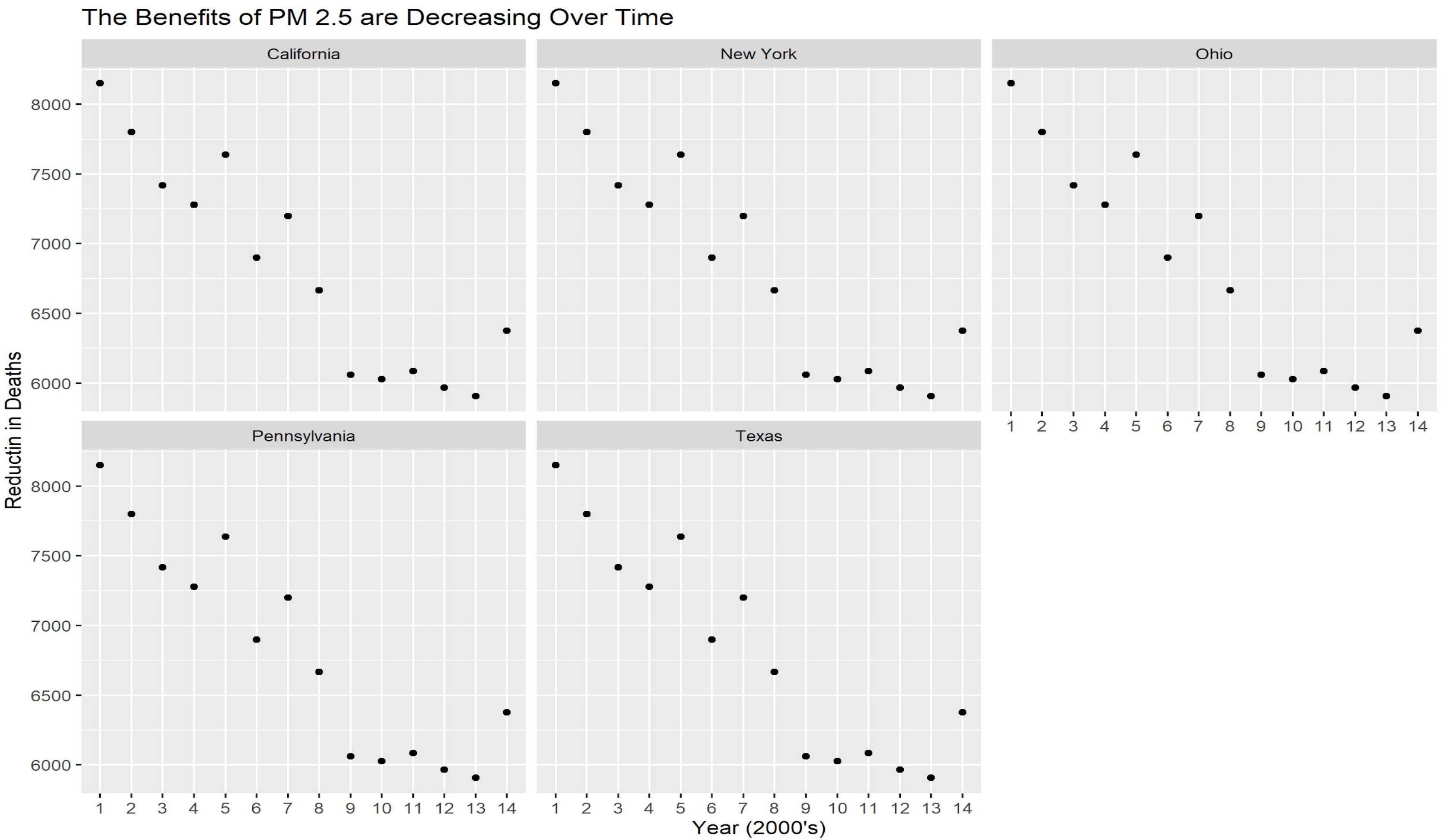


Figure 9. Mortality benefits of decreasing $PM_{2.5}$ over time.

Exploratory Data Analysis - Asthma

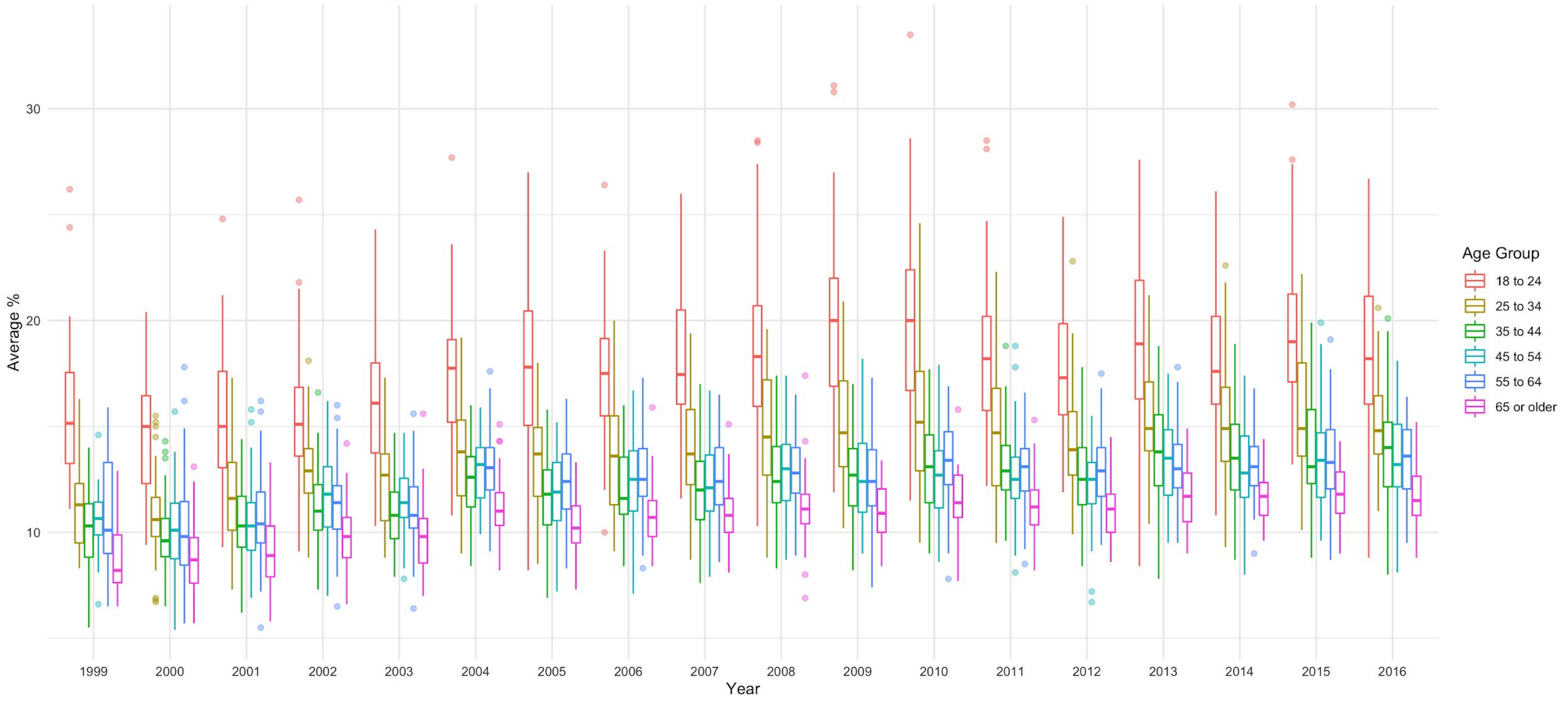


Figure 10. Asthma Prevalence by Age Group and Year.

Exploratory Data Analysis - Asthma

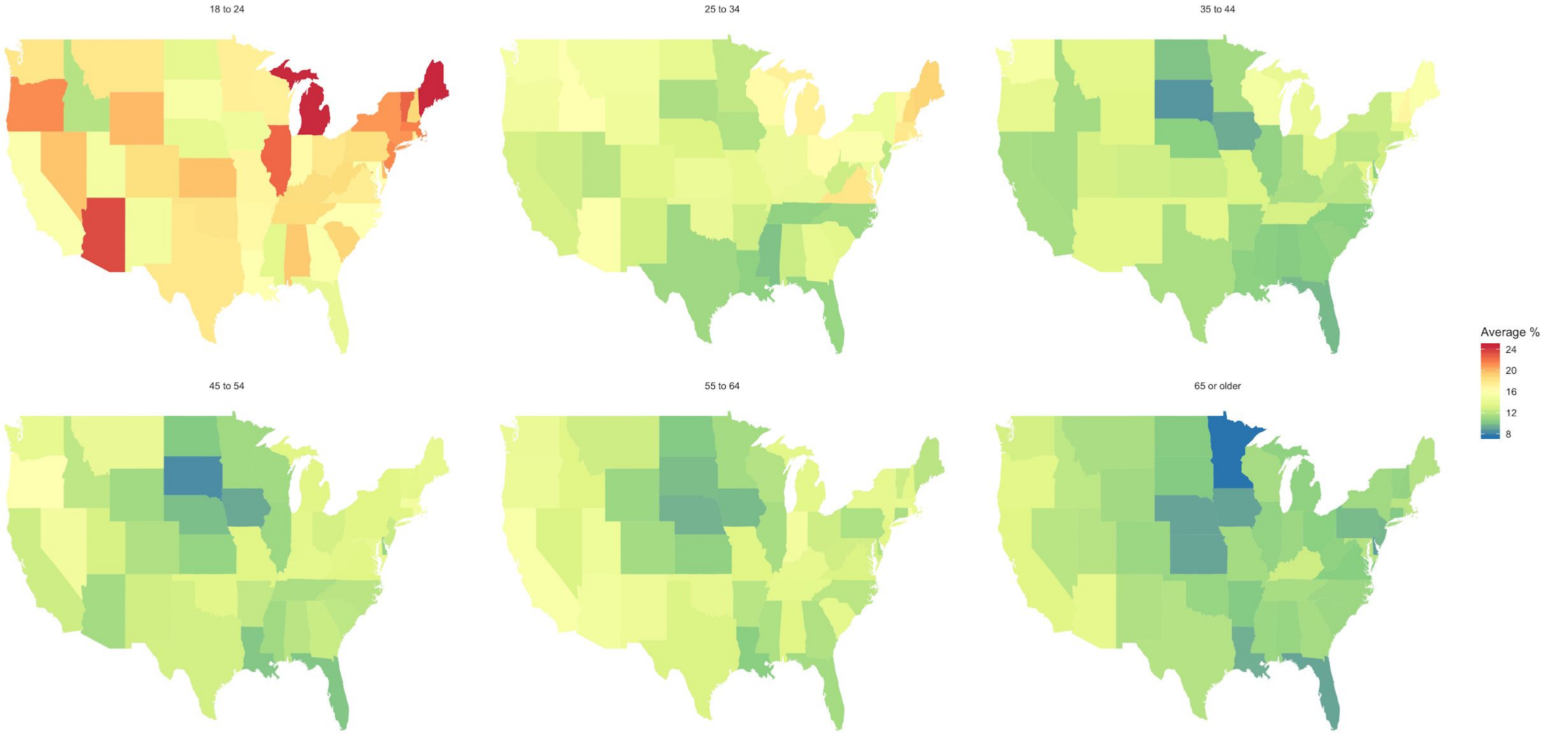


Figure 11. Asthma Prevalence by State and Age Group from 2007 - 2008.

Modeling

Objective

Model Asthma and COPD health outcomes based on air quality and weather data and identify which of these variables are most important for predicting:

- Asthma Emergency Department Visits
- COPD Hospitalizations
- COPD Mortality

Modeling - Missing Value Imputation

Density Plot of Observed (blue) vs Imputed (red) values.

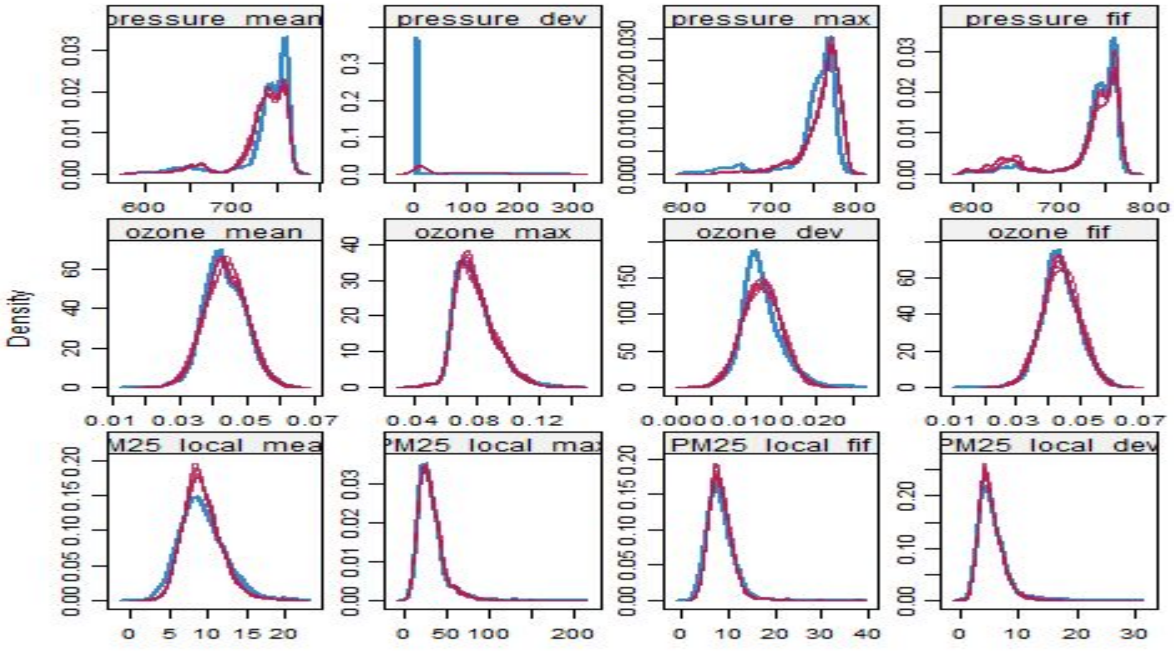


Figure 12. Observed versus imputed values for predictor variables.

Missing values were imputed using Conditional Multiple Imputation with the MICE package.

Here blue represents the observed data and red shows the imputed data.

Modeling - Variable Importance

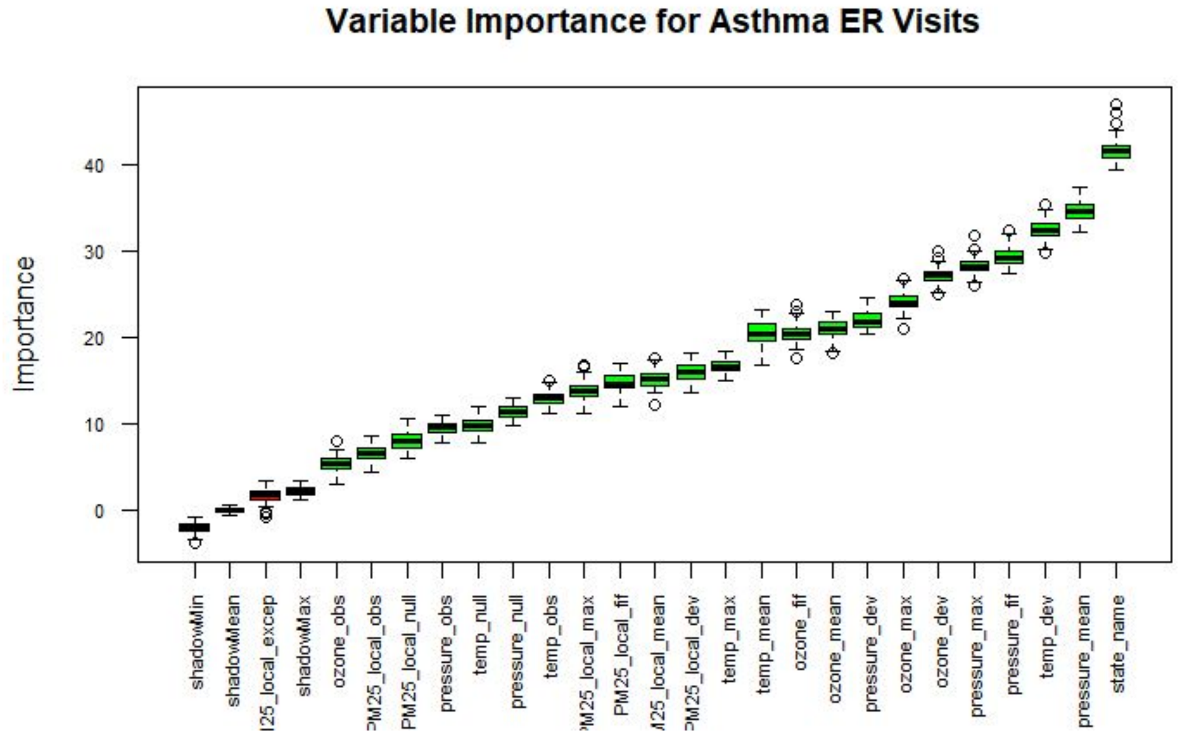


Figure 13. Importance of each predictor variable.

We used Boruta package to rank all variables by importance.

Boruta is a feature ranking and selection algorithm based on a random forests algorithm.

Modeling - Asthma ER Visits

Variable Importance Using Stepwise Backward Selection

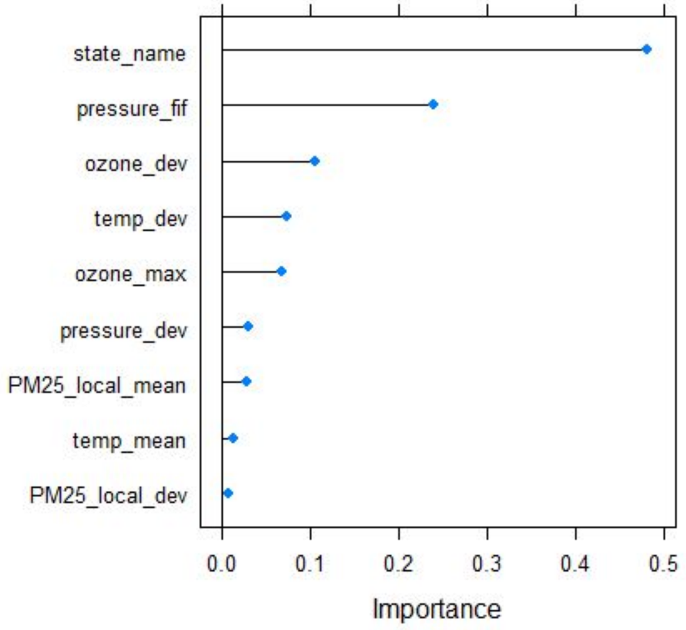


Figure 14. Importance of variables using stepwise backward selection.

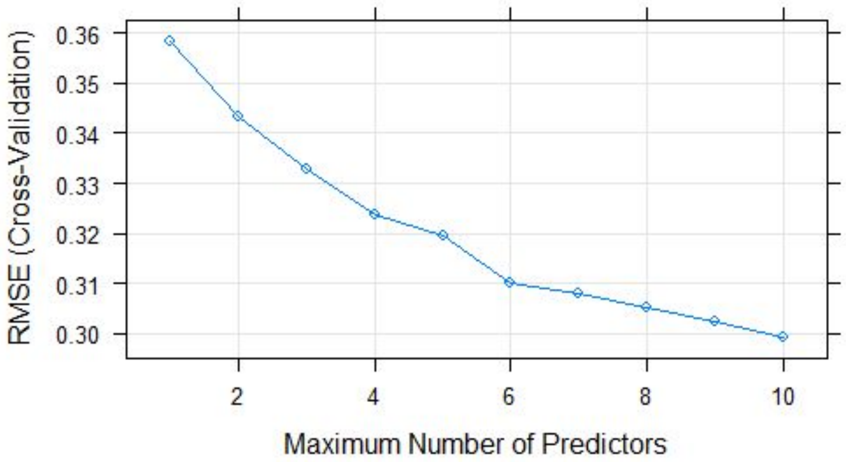


Figure 15. Cross-validated RMSE as number of predictors increases.

RMSE	R-squared	MAE
1.3387169	0.4079990	1.237774

Modeling - COPD Hospitalizations

Variable Importance Using Stepwise Backward Selection

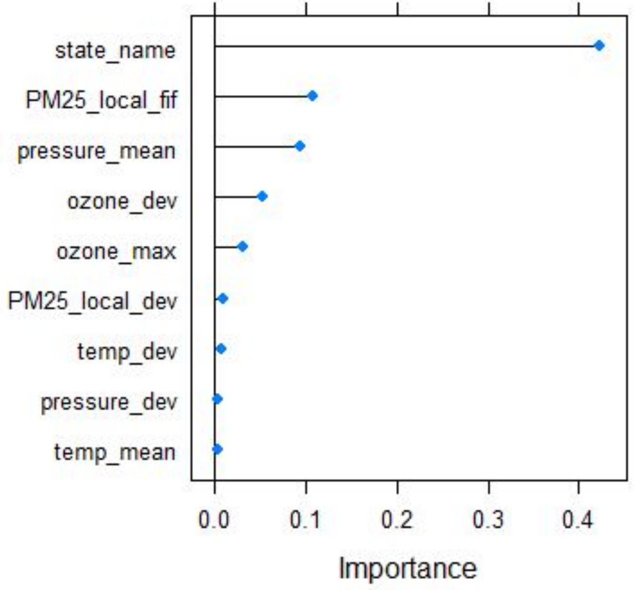


Figure 16. Importance of variables using stepwise backward selection.

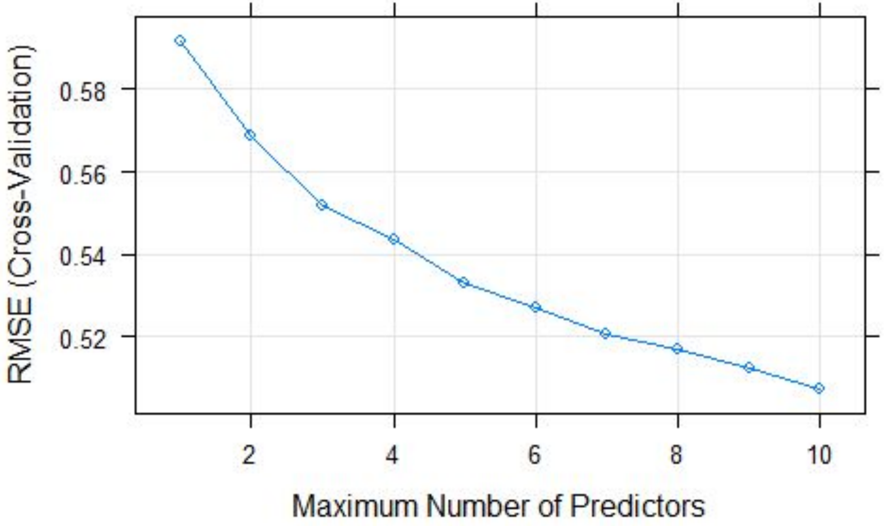


Figure 17. Cross-validated RMSE as number of predictors increases.

RMSE	R-squared	MAE
1.6968100	0.3275164	1.488146

Modeling - COPD Mortality

Variable Importance Using Stepwise Backward Selection

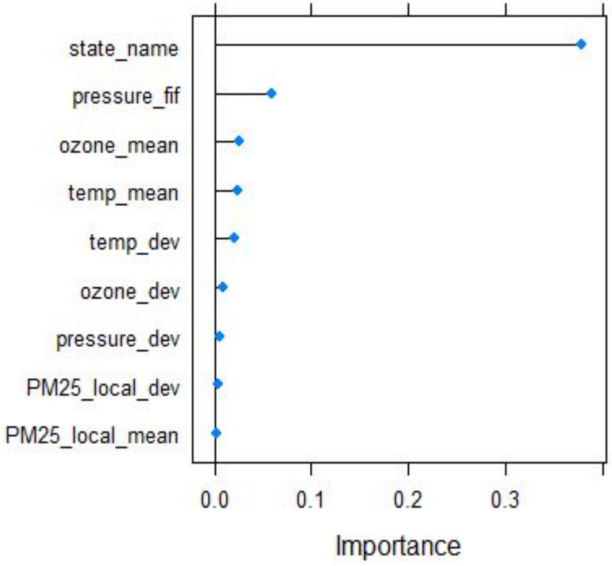


Figure 16. Importance of variables using stepwise backward selection.

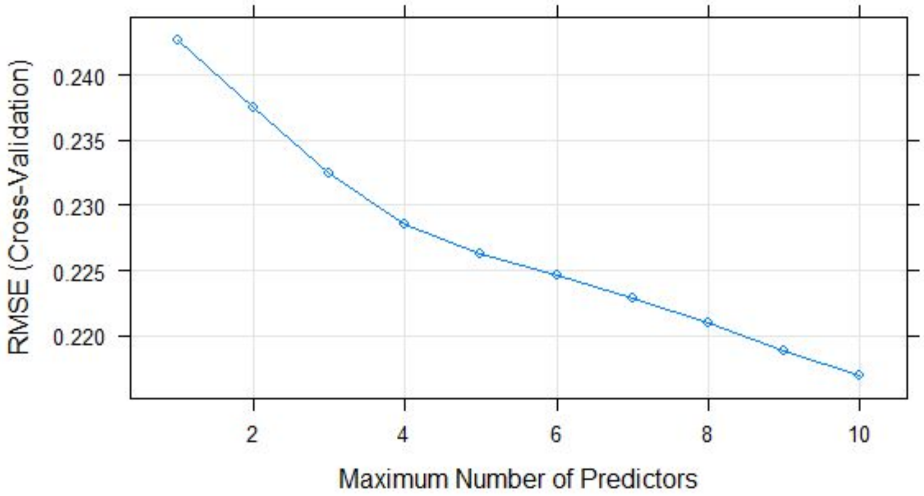


Figure 17. Cross-validated RMSE as number of predictors increases.

RMSE	R-squared	MAE
1.2450260	0.2021450	1.185401

Results & Conclusions

Top 3 Most Important Variables After Adjusting for State and Normalized RMSE

Outcome / Top 3 Variables	RMSE	Rsquared
Asthma ER Visits		
1. Ambient Pressure 50th **2.** Ozone SD **3.** Ambient Temperature SD	1.3387	0.4079
COPD Hospitalizations		
1. PM2.5 Local Conditions 50th **2.** Mean Ambient Pressure **3.** Ozone SD	1.6968	0.3275
COPD Mortality		
1. Ambient Pressure 50th **2.** Mean Ozone **3.** Mean Temperature	1.2450	0.2021

Questions?

References

Environmental Public Health Data Tracking Network, by the Center for Disease Control:

Link to data: https://ephtracking.cdc.gov/DataExplorer/#/

Link to background information: https://www.cdc.gov/nceh/tracking/about.htm

Historical Air Quality Data from the Air Quality System, by the Environmental Protection Agency:

Link to data: https://www.kaggle.com/epa/epa-historical-air-quality

Link to background information: https://aqs.epa.gov/aqsweb/documents/About_aqs_data.html