Final Project: Air Pollution and Respiratory Health

Data Science II

Ahlam Abuawad, Lizzy Gibson, & Yanelli Nunez

5/2/2018

Contents

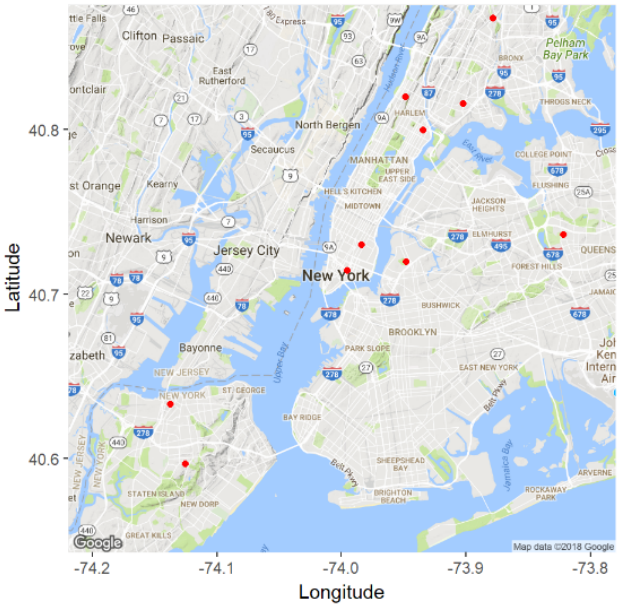
Introduction

The World Health Organization (WHO) has placed air pollution as the world's largest environmental health risk factor. Air pollution is a leading environmental threat to the health of urban populations overall. Although clean air laws and regulations have improved the air quality in New York and most other large cities, several pollutants in the city's air are at levels that are harmful. In the present study, we assessed the potential association between exposure to high levels of air pollution and risk for hospitalizations due to respiratory diseases in New York City (NYC) for the year 2015. We leveraged data from the New York Statewide Planning and Research Cooperative System (SPARCS) and the Environmental Protection Agency's (EPA's) Air Quality System (AQS) database.

Air Pollution

We obtained air pollution data from the US EPA AQS database, which has been extensively used in previous health studies. The AQS includes data summarized on a daily basis for criteria gases, federal reference method particulates (PM 10 and PM 2.5), meteorological variables, toxicants, ozone precursors, and lead. In this analysis, specifically, we obtained daily data for total PM 2.5 mass concentrations, criteria gases (ozone, nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide(CO)) along with the following PM species (individual component particles of PM 2.5): aluminum, selenium, calcium, iron, silicon, lead, manganese, zinc, bromine, copper, nickel, sulfur, titanium, sodium, barium, chlorine, and vanadium. PM 2.5 species were chosen based on a previous study, Kioumourtzoglou et al. 2014. The ten AQS monitor stations in NYC are depicted in Figure 1 below.

Figure 1. AQS Monitor Stations in NYC Map



Health Outcome

The health outcome of interest is the number of hospitalizations due to respiratory disease. This count data follows a Poisson or Quasi-Poisson distribution. The New York Department of Health Statewide Planning and Research Cooperative System (SPARCS) is a comprehensive data reporting system that collects information on hospital admissions and emergency department (ED) visits within New York State. The SPARCS dataset contains information on approximately 98% of all hospitalizations in non-federal acute care facilities regardless of insurance status. Information on patient characteristics, diagnoses, treatments, services, and charges was also collected for each hospitalization or ED visit. Additionally, the dataset includes demographic information such as age, sex, race, and residential address. A unique patient ID is assigned to each person in order to allow for tracking hospitalizations or ED visits over time. We will use SPARCS data on hospital admissions and ED visits from the year 2015 to determine the number of respiratory disease hospitalizations in NYC. Our data is de-identified and consists of counts of inpatient hospitalizations or ED visits per day in the five NYC boroughs: Brooklyn, Manhattan, Staten Island, Queens, and the Bronx.

Confounders

The unit of analysis in this example is day, thus confounders can only be variables that vary from day to day and covary with both exposure and outcome. Weather conditions influence air pollution levels by concentrating, diluting, or chemically processing pollutants; therefore, we will include temperature as a covariate in our analysis. Data for the temperature variable was also obtained from the AQS database (see description above).

Research Question

Through this study we aim to answer the question of whether exposure to high levels of air pollution increases the number of respiratory disease hospitalizations in NYC. For this, we will use a combination of supervised and unsupervised data analyses including lasso, boosted random forest, and principal component analysis.

Data Preparation

Exposure data were downloaded from the EPA AQS website. We downloaded separate datasets for each criteria gas, PM 2.5, PM species, and temperature. Next, the data was cleaned using the Tidy philosophy to obtain a dataset with days as rows and corresponding pollutants as columns. As the PM speciation filters are expensive to analyze, they are only measured every three days. Thus, we restricted our analyses to every third day of 2015 when we had complete data (n = 108). We obtained the outcome data from SPARCS from a colleague. It includes counts of inpatient hospital or ED visits per day in NYC. As this data is de-identified, this analysis was exempt from IRB oversight. Exposure and outcome datasets were merged for final analyses. We created separate training and testing datasets and set the same seed for all steps involving randomness to ensure rigor and reproducibility in our analyses.

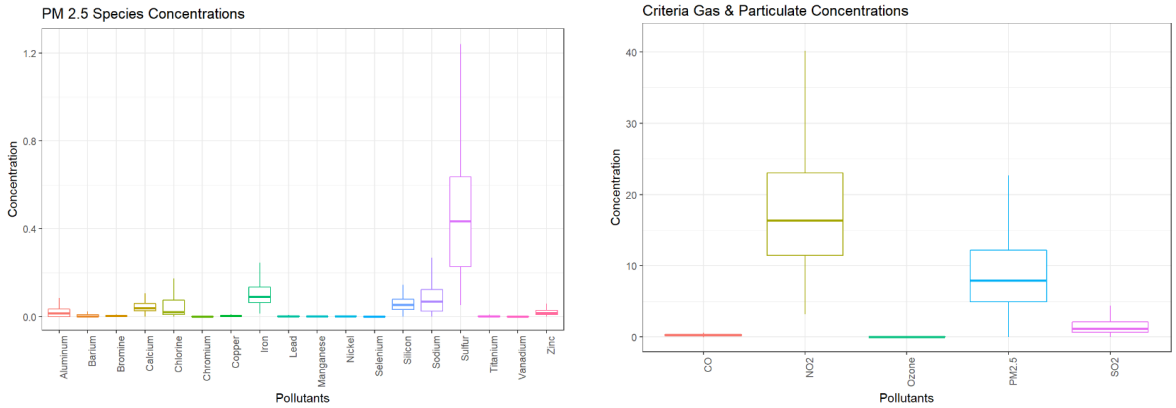
The code for data preparation, cleaning, and analyses can be found in our GitHub repository (https://github.com/lizzyagibson/Air.Pollution.Health).

Unsupervised Analysis

Exploratory Data Analysis

In the exploratory data analysis (EDA), the first boxplot (Figure 2a) of the PM 2.5 species shows that by a large difference, sulfur is the pollutant with the highest concentration (and standard deviation). The second boxplot (Figure 2b) of criteria gas and particulate concentrations shows that NO2 has a larger concentration than all of the PM 2.5 species combined.

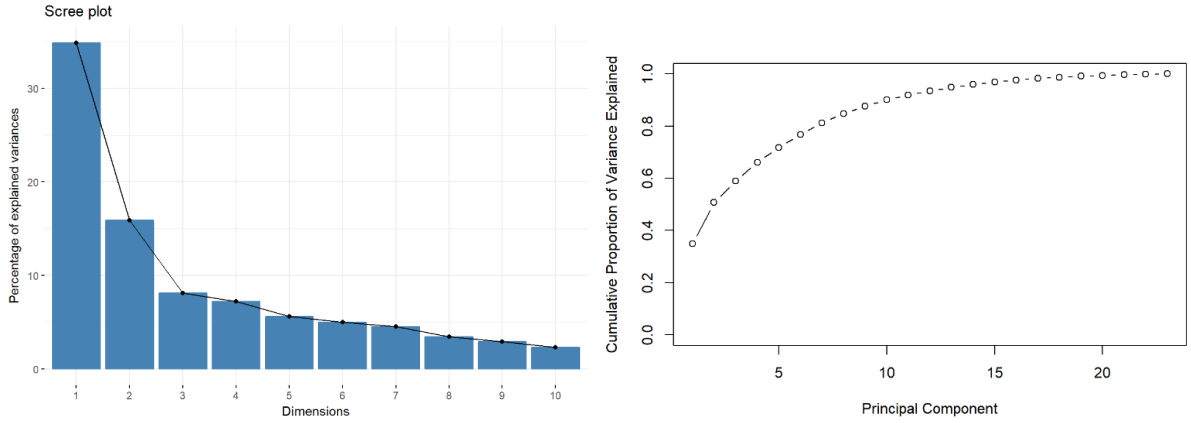
Figure 2. (a) PM 2.5 Species and (b) Criteria Gases



Principal Component Analysis

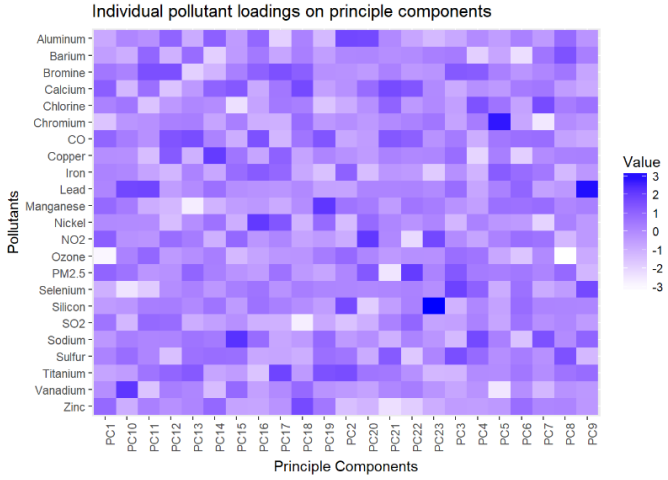
For the unsupervised analysis, we also performed a Principal Component Analysis (PCA). This is a low-dimensional representation of the data that captures as much of the information as possible. PCA seeks a small number of dimensions that are all as "interesting" as possible (not all dimensions are equally interesting), which is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the p original features. The first two principal components (PC's) of a dataset span the plane that is closest to the n observations in terms of average squared Euclidean distance. PCA requires continuous variables and some strong assumptions, including a linear relationship between all variables, a large enough sample size, and no significant outliers (as PCA is very susceptible to outliers). We chose PCA because the dataset contains a large amount of correlated variables, and PCA will help to understand which components comprise a majority of the variability in the data. The model fit was first assessed using Figures 3a and b shown below.

Figure 3. Assessing Model Fit Using (a) Scree Plot (b) Variance Plot



The scree plot reveals that a majority of the variance in the data (~35%) was captured in the first principal component (PC1). The remaining 65% of the variance was captured by nine other PCs. The "elbow" of the scree plot tells us that the first three components do a sufficient job of capturing the variability in the data (~60%).

Figure 4. Heat Map of PCA Data



As shown in Figure 4, PC1 contains three out of four of the criteria air pollutants (NO2, CO, and SO2) and PM 2.5. The fourth criteria air pollutant, ozone, was not in PC1 but did appear in the other PC's to varying extents. In fact, the plot of the eigenvectors (Figure 5a) shows that ozone is in a different dimensional space compared to all of the other variables in the dataset. NO2, CO, and SO2 are in a similar dimensional space, and PM 2.5 is more similar to those three criteria air pollutants, but is also in a different dimensional space.

In comparing week days and weekends, there is a bit of overlap in the PCA. In comparing days in different seasons, winter has a strongly different distribution of pollutants compared with fall, spring, and summer.

Figure 5. Variables plots of (a) Eigenvectors and (b) Loadings

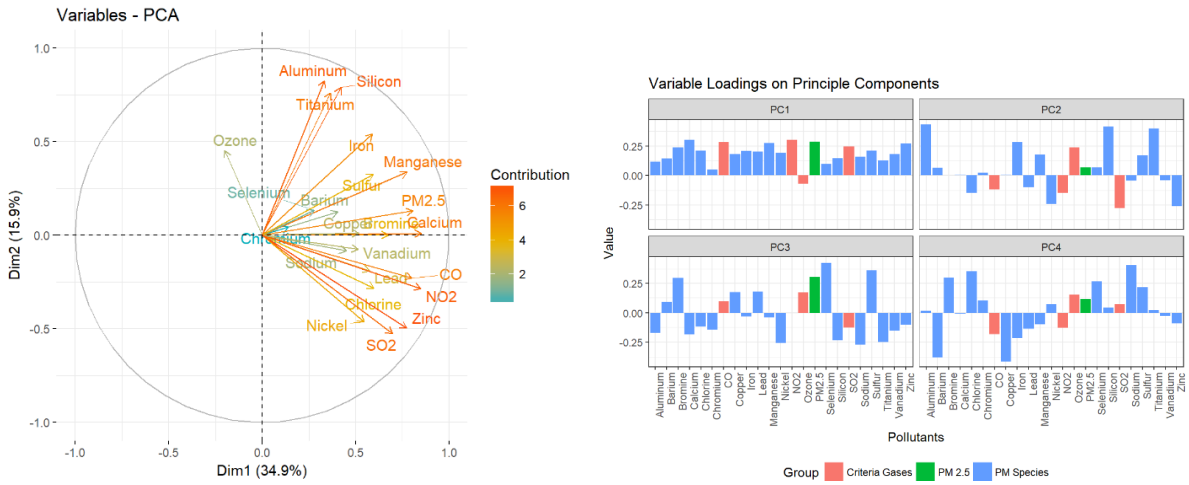
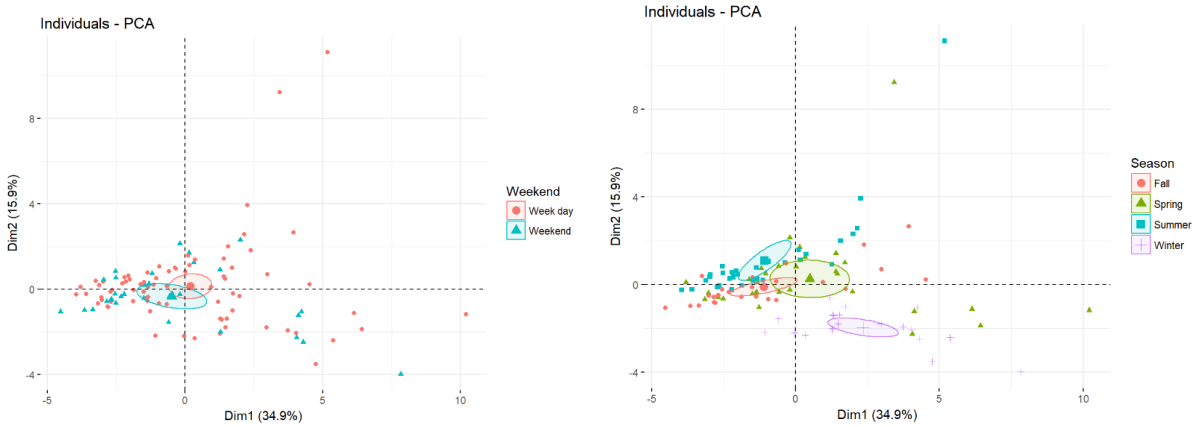


Figure 6. PCA (a) Weekly and (b) Seasonal Plots



Supervised Analysis

What predictor variables did you include? What technique did you use, and why did you choose it? What assumptions, if any, are being made by using this technique?

All air pollution predictor variables from the cleaned dataset were used in the supervised analyses (total PM 2.5 mass concentrations, ozone, sulfate, nitrate, carbon monoxide, aluminum, selenium, calcium, iron, silicon, lead, manganese, zinc, bromine, copper, nickel, sulfur, titanium, sodium, barium, chlorine, and vanadium). Temperature was included as a covariate. The variables excluded were month, day of the week, humidity, and date. Month and day of the week were used for grouping in the PCA, but were not hypothesized as predictors of respiratory health. Humidity (a potential covariate) was excluded due to 50% missing data. Date was excluded as it was the unique identifier and not an air pollution predictor. For our supervised analyses we used boosted random forests and lasso models. The dataset was split into training and test sets in order to perform cross-validation for tuning parameters, fitting the models, and testing their performances. We chose lasso as a variable selection method because many of our exposure variables are highly correlated. Lasso assumes a linear relationship between predictors and outcome (which is quite restictive), and when predictors are highly correlated, lasso will chose the one that is more strongly correlated with the outcome and push the coefficient for the other to zero. We chose boosted random forests as a more flexible method to compare prediction error. The gbm package includes a Poisson option (the RandomForest package does not). Boosted random forests make no assumptions about the shape of the association between predictor and outcome but do require decisions on the appropriateness of tuning parameters.

Boosted Random Forests

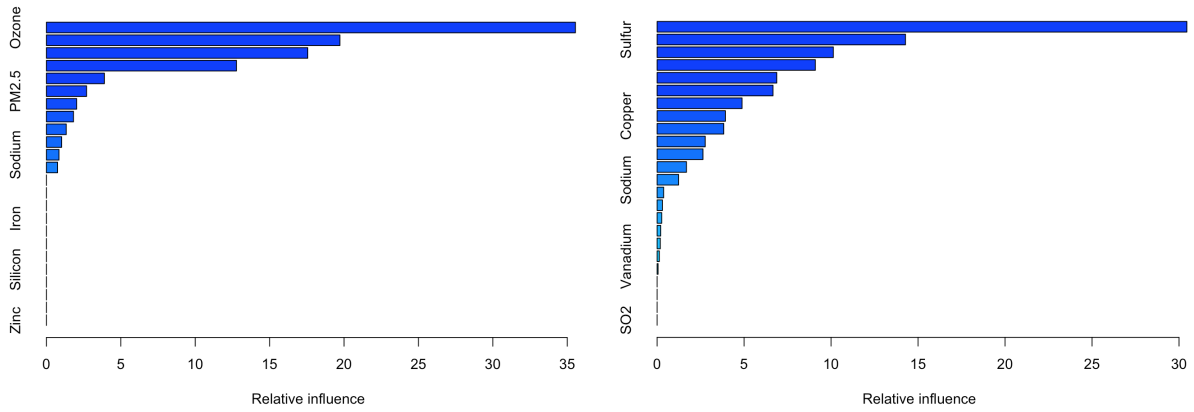
Boosted random forests can help to classify variables, similar to that of regression trees but in a more stable manner, and without assumptions on any information regarding the distribution or collinearity of variables in the model. In boosted random forests, there are three tuning parameters: the number of trees, amount of shrinkage, and number of splits in each tree (interaction depth). The number of trees was selected by comparing various values. With 100 iterations (Figure 7a), both the training and test set mean square errors (MSEs) were large (3201.43 and 3086.71, respectively), but 12 variables were removed. On the other hand, 5000 iterations resulted in smaller MSEs but none of the variables were excluded. Thus, 1000 iterations (Figure 7b) were chosen because this amount of trees resulted in smaller training and test set MSEs (2552.90 and 2674.02, respectively), as well as removed four variables. The shrinkage value is a measure of the learning rate and 0.001 is a recommended value for growing a lot of trees using a small dataset. Lastly, as this is an additive model, the depth was set to 1.

In assessing the model fit, a comparison of the true test values with the predicted test values and the corresponding deviance scores was completed. Table 1 below shows the first six value comparisons:

Table 1. Boosted Random Forest True and Predicted Values from 100 vs 1000 Iterations with Corresponding Deviance Scores

True	Predicted for 100	Deviance Score for 100	Predicted for 1000	Deviance Score for 1000
55	61	3.24	60	3.18
47	60	4.31	60	4.35
55	61	3.22	61	3.19
29	60	12.75	60	12.35
52	59	3.31	58	3.24
51	61	3.77	61	3.69

Figure 7. Boosted Random Forest Relative Influence of Variables (a) 100 vs (b) 1000 Iterations



The plots above reveal that ozone has the highest relative influence on respiratory hospital admissions and ED visits, regardless of the number of iterations. The table below shows other variables of interest that may also play a role:

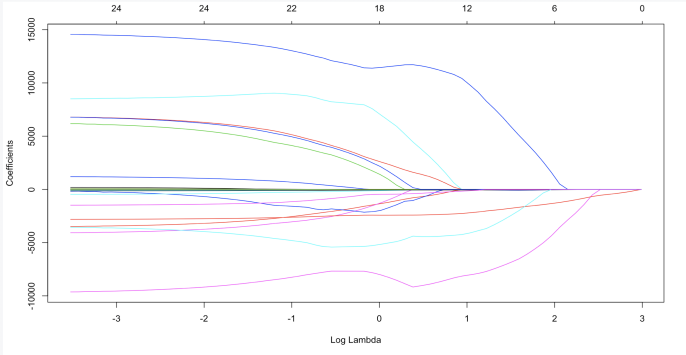
Table 2. Boosted Random Forest Relative Influence Values for 1000 Iterations

	var	rel.inf
Ozone	Ozone	30.4527408
Sulfur	Sulfur	14.2772359
Aluminum	Aluminum	10.1310589
Nickel	Nickel	9.0971719
Temp	Temp	6.8787764
PM2.5	PM2.5	6.6631122
Lead	Lead	4.8838923
Copper	Copper	3.9275707
Silicon	Silicon	3.8311974
Chlorine	Chlorine	2.7654382
Calcium	Calcium	2.6369787
Iron	Iron	1.6881118
Sodium	Sodium	1.2357212
Chromium	Chromium	0.3776950
NO2	NO2	0.3055970
Zinc	Zinc	0.2675830
Manganese	Manganese	0.2082209
Bromine	Bromine	0.1847169
Vanadium	Vanadium	0.1290026
Titanium	Titanium	0.0581783
Barium	Barium	0.0000000
CO	CO	0.0000000
Selenium	Selenium	0.0000000
SO2	SO2	0.0000000

Lasso

For one of our supervised analyses we used lasso. Lasso is a variable selection method that yields sparse models, that is, models that are easier to interpret in comparison to ridge regression. This is due to the fact that depending on the choice of the lambda parameter (Figure 8), some of the coefficients will be exactly equal to zero which helps produce a more simplified model. It is also important to mention that lasso relies on a linear model and assumes noncorrelation between the variables. Lasso is also easier to interpret than boosted random forests (though lasso is less flexible) because lasso provides explicit beta coefficients, or effect estimates.

Figure 8. Lambda Values



We used the training dataset to fit the lasso model and find the best lambda value through cross-validation. Then we used the test set to check the performance of our model.

Results

The best lambda value that we found through cross-validation was 4.479. In Figure 9, we can also see that when testing multiple values in the training dataset, the lambda value that produces the lowest MSE is around 4. The lasso model with the lambda chosen by cross-validation contained only 10 variables which are presented in Table 3.

The MSE from the prediction using the training dataset is 2177.17 and the MSE using the test set is 6587.115.

Figure 9. Best Lambda Value

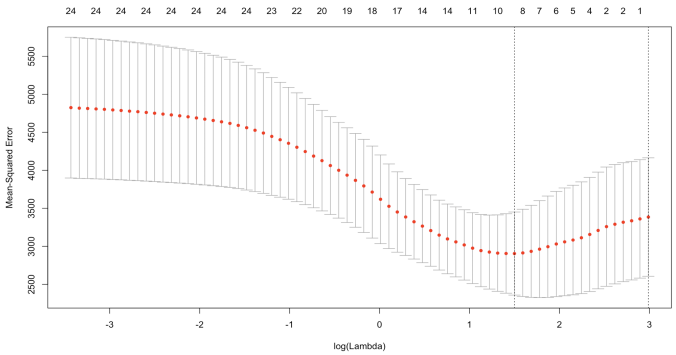


Table 3. Lasso Variable Coefficients

Variable	Coefficient value
Calcium	-52.448
Chlorine	-17.64
Lead	7588.49
Nickle	-7444.53
Ozone	1948.54
SO2	-1.49
Sodium	-51.67
Sulfur	-14.36
Vanadium	-3325
PM2.5	-0.4

Conclusion

As shown in Table 4 below, a comparison of the boosted and lasso test MSE's resulted with the lasso producing a higher test set MSE. This is not surprising as lasso is a less flexible method than boosted random forests.

Table 4. Boosted Random Forest and Lasso Test MSEs

Boosted Test MSE	Lasso Test MSE
2674.017	6835.888

Although the boosted random forests resulted in a smaller test set MSE, this model is much less interpretable than the lasso model. Because the lasso model is more interpretable, and because in this analysis we are interested in interpreting effects over predicting outcomes, we used the lasso model to make predictions. All of the models reveal that ozone is an important factor in predicting the number of respiratory hospital admissions and ED visits per day. This corresponds with current literature that shows ozone's ability to damage the lungs once inhaled. Sulfur and PM 2.5 are also factors of interest. The PCA showed grouping of most PM 2.5 species together, and a strong separation from ozone, which is what we would expect. Furthermore, it is important to note for the supervised analyses that weekly and seasonal variability in exposures also impact respiratory hospital admissions and ED visits. In terms of policy interventions, these analyses are still preliminary and are the first steps toward assessing the association between exposure to high levels of air pollution and the amount of respiratory disease hospitalizations in NYC. This analysis shows how data science methods – including dimension reduction, variable selection, tree-based methods, and cross-validation – can be used for rigorous and reproducible hypothesis generation and testing in the environmental health sciences.