**Title:** Regression of fossil fuel and air quality with demographic attributes

**By:** Annemarie Donohue, Maria Ignacia Araya, Regan Alt, Shu Guan and Tova Gottesman

**Class**: BYGB-7967

**Section:** V03

**Date:** November 7th, 2020

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **Abstract:**

High air quality is critical to all life on Earth. The goal for this project is to find the demographics associated with the correlation of fuel and air quality. Using datasets featuring the consumption of biofuel, demographics, annual air quality index and tax returns (looking at income disparities), this project will look at air pollution trends over the past five years in the Northeast region. Variables were chosen from each dataset to understand their correlation with air pollution. Our findings will expose the part of the population in which northeastern states, Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont, are lacking in air quality index. The impacts of air pollution render a public health and environmental concern, affecting people from all walks of life.

# **Introduction:**

Fossil fuel emissions contribute to a decrease in air quality, but how much are they actually contributing? Who does this affect the most? This can be found by finding the correlation between air quality and fossil fuel emission. Using this information,our analysis will show a baseline for our air quality, what is going into our air, and who is affected by high pollution.

There are variables of air quality that can differ in certain areas depending on the time of year, weather conditions, and extraneous circumstances such as a global pandemic or wildfires. According to the Environmental Defense Fund, however, nearly 40 percent of Americans live in areas with unhealthy levels of smog pollutants (EDF). Smog is formed around the ground-level ozone layer of cities when emissions from power plants, factories, cars, and other fossil fuels react with heat and sunlight (EDF). Nitrogen oxide is a byproduct of all fossil fuel combustions and contributes to acid rain and smog. When smog is inhaled, the particles from the emissions irritate airways and burn lung tissue, threatening asthma and bronchitis, and exacerbating the risk of heart and lung diseases later in life (UCSUSA). Furthermore, when acid rain increases the acidity in lakes and streams, it harms fish and other aquatic life, as well as damages trees and weakens forest ecosystems (EPA-acid rain). Due to these underlying health costs almost half of Americans face from fossil fuel emissions, app developers added an air quality index in the weather app for nearly every phone model, highlighting the impact air quality has on our daily lives.

The economic impact of air pollution overall is equally startling. In the US alone, air pollution accounts for nearly 5 percent of the yearly gross domestic product, generally related to early deaths from exposure to particulate matter such as Nitrogen oxide (Stanford). Concerningly, most economic production depends on energy from burning fossil fuels around the world, whereby the combination of carbon dioxide and nitrogen oxide emitted into the air not only creates air pollution but contributes to global warming. A report published in PNAS thus found that the reduced per-capita economic productivity due to these emissions is projected to be the most negative socioeconomic byproduct of burning fossil fuels. Luckily, according to a study in the *Proceedings of the National Academy of Sciences*, from 2008 to 2014, air pollution fell 20 percent over this six-year span (Stanford). Additionally, over the last few decades as the GDP, population, and car usage in the United States has grown, there is evidence that emissions from the most common pollutants have dropped 74 percent (EPA). These are positive signs air quality is improving in the United States, however, our research will uncover where exactly emissions are most prominent and where alternative energy use such as ethanol can salvage the remaining contaminated air.

# 

# **Data Description:**

We used four different datasets, from different sources:

**1) Demographic: U.S. Census Bureau, Population Division.**

Annual Resident Population Estimates and Estimated Components of Resident Population Change for Metropolitan and Micropolitan Statistical Areas and Their Geographic Components: April 1, 2010 to July 1, 2019

Variables: 97.

* Descriptive variables: CBSA, MDIV, STCOU, NAME, LSAD.
* Demographic variables ( each variable has one column by year from 2010 to 2019): population, numeric change in population, births, deaths, natural increase, international migration, domestic migration, net migration and residual.

Records: 2,797

**2)Taxes:Internal Revenue Service SOI Tax Stats - Individual Income Tax Statistics - 2018 ZIP Code Data (SOI)**

Files: five different datasets by year. From 2014 to 2018.

Variables:153

* Descriptive Variables:statefips, state, zipcode, AGI\_STUB.
* Taxes related variables

Records: 845,688 total

**3) Energy Consumption: U. S. Energy Information Administration. State Energy Data System (SEDS) 201**8.

Variables:61

* Data Status, State and MSN. ( MSN includes 211 combinations of energy source and sector as records), years from 1960 to 2018 ( one column for each year)

Records: 10,975

**4) AQI:United States Environmental Protection Agency. Annual aqi by cbsa Year.**

Files: five different datasets by year. From 2014 to 2018.

Variables:28

* Descriptive variables: CBSA, CBSA Code, Year.
* AQI related variables: 16, includes days with AQI, good days, moderate days, unhealthy days, days CO2, days Ozone, days PM10.
* Descriptive statistics: Median AQI, max AQI, 90th percentile AQI
* An AQI value of 100 is generally the national air quality standard for that pollutant, and any AQI greater is considered a health-risk for sensitive groups, and gradually the whole population as the AQI hits 150.

Records: 3,002 total

# **Problem Statement**

How is air pollution correlated with fossil fuels? What other demographics also contribute?

# **Methodology:**

### Data Collection

AQI

Air Quality Index (AQI) is an index for reporting daily air quality in the United States, representing how clean a designated area’s air is or is not. This report is important due to the numerous health effects that are a consequence of poor air quality on humans, animals and the climate. The days are ranked on a scale, 0-500, with 0 being the best quality days. We used data from the United States Environmental Protection Agency (EPA), who calculates the AQI for the five major air pollutants that are regulated by the Clean Air Act (cfpub.gov). According to the EPA, ground-level ozone and airborne particles pose the greatest health threat to citizens of the United States. For our data collection, we utilized the EPA’s “Annual\_AQI\_by\_CSBA” for the years 2014 through 2018. We downloaded one csv file per year.

## Demographic

The demographic data depicts attributes of the population. This let us gain knowledge of the population that was being analyzed. Our project includes two major variables covered, the number of people and the number of deaths during the five-year period 2014-2018. This data was collected from pregenerated files in csv format from the Census Bureau.

## Use of Energy by State in BTU

Not only does this dataset describe types of energy used, but also the quantity being used. The MSN, which is the use of energy by state, is a code of five characters. First and second characters describe the energy source, the third and fourth describe the energy sector or activity and the last one indicates the measure that in this case is BTU. The combination of energy source and sector or activity is out of 270. The values are presented by year and by state. The dataset is a pre generated file as csv.

## Income(Tax)

One of the demographic variables not mentioned in any other dataset was income. Financial flexibility can provide healthier food, safer neighborhoods, and better schools. Due to the fact that income normally impacts quality of life, this dataset was chosen to see if there was also an impact on air quality and income. The variable of aggregate gross income was chosen from the taxes dataset. The purpose is to better understand the role air pollution can have in correlation to income.

**Data Preprocessing:**

AQI

In the AQI DATASET, The column for CBSA code joined some states together. First, we broke up cities by state and reordered the table of columns by state as first key and city as the secondary key. Then we exclude attributes that we are sure will not be used. The attributes we decide to use include CSBA City, State, CSBA Code, Year (eliminated 2019 data), Good Days, Moderate Days, Unhealthy Days, Very Unhealthy Days, Max AQI, Median AQI. These are the essential attributes for the datasets or data with statistical meaning. After filtering the attribute we checked through the data and dropped either missing or duplicated data. In order to better integrate the data later, we copied each state, data and correlated CSBA code into a new Excel sheet. The sheet was separated by space to identify different regions. After discussion, we decided to limit our research to nine northeastern states and to remove the second state from the 2-state records. The state (Ex: RI - MA) for Providence to Warwick to just Rhode Island. The state is most correlated within public opinion, so it makes it easier to analyze based on the city instead of regional zip code, and easier for the general public to interpret as well. New York-Newark-Jersey City also just became NY for ease. Then we deleted spaces by “regions,” sorted our AQI dataset by state and ordered alphabetically.

Demographic

For simplicity, we first filtered the variables that were not significant like state and county code by hand through excel. In the SPSS, we import the demographic dataset by adding the Var.File node. In the node setting, we created a column based on name to extract the State (input State). Then we added the node filler and connected it to the source node. We applied a formula to substring the characters needed, without specific condition. We filter LSAD only by metropolitan statistical area, micropolitan statistical area and metropolitan division that were arranged by CBSA. The main purpose of this step is to ensure that the accuracy and range of data in different data tables are consistent. After filtering LSAD, we added the aggregate node to condense by state. We used the sum option because we needed the quantity of the variable by state. Due to the format difference between different datasets, (sorted by column or row) we transposed each “variable+year”, using fields to records. We wanted the fields to be converted to only one column, separate from the year. Now with two created columns, one with the name of the variable, for example population2013, and the value in another column. We used 7 different transpose nodes, one for each variable needed. Then we used the filter node to change the name of the variable to Year and Value to the specific variable (for example population, deaths, births). We used the filler node with the formula substring to extract the last character of the string to obtain the year, so the column has only years. Finally, we merged all the 7 nodes with a key value “state+year.” We merged it again with a source node with the north east states (from an excel file) that we want to consider. Then we exported it as a flat file.

### Use of Energy by State in BTU

Using SPSS, first we filtered to delete excess in columns to get the data status that indicated the year of the data. We selected the year range needed, 2014 to 2018, along with the MSN ( combination of energy and sector), BFTCB ( Biofuel Total Consumption), WYTCB (Wind energy total consumption), SOTCB (Solar energy total consumption), HYTCB (Hydropower total consumption), GETCB (Geothermal energy total consumption) and FFTCB (Fossil fuels total consumption). Then the transpose node switched the field to records by converting the five years into one column named Year, and the value of the MSN. This was then exported as a flat file.

### Taxes

For the 2014-2018 time interval chosen, there were 5 yearly datasets. In the beginning, our group was choosing to have either cbsa or state to be the location field, so the cbsa was included in the beginning. The cbsa dataset included a format to convert zip code to cbsa. Starting in Python, pandas, used for data structure operation manipulation and analysis, and numpy, used for mathematical functions on large arrays and matrices, were imported due to their relevance. The cbsa values were sorted by ratio, and duplicates were dropped by only keeping the first zip code associated in that cbsa code. An id column was added to create uniqueness. The cbsa file was then individually merged into each yearly tax dataset through an inner join. Going into SPSS, the 5 tax datasets were each loaded into 5 variable file source nodes. Now seeing that state would be the best field for location to incorporate into every dataset (foreign key here), state was chosen over cbsa throughout the next processing steps. Through the filter tab on the variable node, the inputs were changed to STATE and A00100, which stands for the adjusted gross income. Since there were multiple cities reported within each state, each year was aggregated by mean so each state had a yearly mean. This was done through aggregate nodes attached to each of the 5 variable file nodes. A merge node was connected to the 5 aggregate nodes, with a specification to merge the key “STATE,” and only include matching records (inner join). This transformed all the yearly means from all states into one dataset. Next, the derive node was used to specify the certain states within the field “STATE” that were in the northeast region. By deriving as a “flag,” or boolean variable, the syntax (member(STATE, [list of state names])) with the field and state names were written to be true. The boolean variable is binary, describing a variable to either be true or false here. The select node was used to specify the inclusion of only the states labeled as true, which got the final product of northeast region states and their yearly mean incomes. For reference, the ME is referring to Maine, and MS is referring to Massachusetts. This differs in some files in this project. Finally, the dataset was exported using the flat file node.

### 

### Variable Selection:

AQI

*Variables chosen in AQI*= CSBA City, State, Year (eliminated 2019 data), Good Days, Moderate Days, Unhealthy Days, Very Unhealthy Days, Max AQI, Median AQI

Out of originally twenty-one columns, we parsed the data to just nine. The first column ‘State’ represents the region in which the AQI was measured. Therein, the ‘City’ column comprises a metropolitan area in New England where air pollution would have the most detrimental effects, such as the New York, Newark, and Jersey City area, represented as ‘NewYork-Newark-JerseyCity’. For our data, we retrieved annual Air Quality indices from 2014 through 2018, which is represented in the third column ‘Year’. The following columns named ‘Good\_Days’, ‘Moderate\_Days’, ‘Unhealthy\_Days’, and ‘Very\_Unhealthy\_Days’ represent the number of days within each year where the AQI was designated under one of these labels. The Environmental Protection Agency where we retrieved the data used a scale to represent the number of days for which measurements from any monitoring site in the country needed to report Air Quality to the Air Quality System database. For instance, Good Days were measured on a scale between 0 - 50, Moderate Days between 51-100, Unhealthy Days between 151 - 200, and Very Unhealthy days between 201 - 300. Furthermore, Max AQI represents the highest daily AQI value in the year, and Median AQI is measured by half of the AQI values during the year which were less than or equal to the median value, and half equaling or exceeding it. The Max and Median AQI values help exhibit the overall AQI value for the year and state.

### Demographic

*Variables chosen*: Year, State, Deaths, Births, International Migration, Domestic Migration, Population.

We selected CBSA that is the core statistical base area, and the name of it. Wa had many datasets by cbsa, but one that is composed by state with no more information. That is why we chose to work with states and preprocess the datasets to present a state column. We selected only the years of the scope, that is why we deleted the column census2010pop and estimates base 2010. We did not select Numeric change in residents because it could be redundant as we considered population as one of the fields. Then we selected births, domestic migration and international migration to have more information about the change in population and how this could affect the AQI. We selected Deaths to analyse the correlation between AQI indicators and other demographic variables. We Ignored the net migration and residual.

### Use of Energy by State in BTU

*Variables chosen*: State, Year, Biofuels total consumption (BFTCB) , WYTCB (Wind energy total consumption), SOTCB (Solar energy total consumption), HYTCB (Hydropower total consumption), GETCB (Geothermal energy total consumption) and FFTCB (Fossil fuels total consumption)

These variables were chosen to analyze if these clean energy sources have a correlation with air quality and deaths. With clean energy, the air should be cleaner with more consumption considering that consumption would normally be used by a fossil fuel source. The more consumption of clean energy should lead to better air quality.

### Taxes:

### *Variables chosen:* State, Adjusted Gross Income (A00100)

The goal of this set was to measure the income of the states for the northeast region. The two variables considered were either adjusted gross income or total income amount. The first was chosen due to the probability of increased accuracy.

**Backward selection of variables**

First we Choose a significance level (e.g. SL = 0.15 with a 70% confidence).

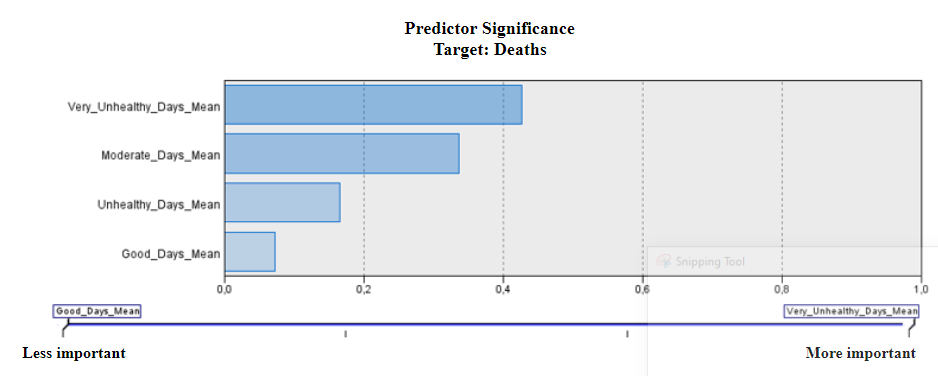
In all possible models, add an additional feature to the previously selected feature. Again, choose the feature with the lowest P value.If p\_value & lt;Then go to step 3, otherwise terminate the process. The code for the backwards selection for each dataset and should be used for reference.

### 

### Model Building: Nonlinear Regression: AQI vs Deaths

**Algorithm**

After understanding the detrimental effects of poor air quality on human health, we wanted to uncover if there is a higher degree of deaths in the presence of more contaminated air. Our goal is to find the variable that best correlates with the target, so we decided to conduct nonlinear regression. First, we imported the combined Demographic and AQI dataset into a variable file node. This was simply taking the demographics and AQI datasets and merging them into one for ease. The Type node was connected to indicate “Deaths” as the target, and “Very Unhealthy Days Mean”, “Moderate Days Mean”, “Unhealthy Days Mean” and “Good Days Mean” as the input variables. Next, a Regression node was connected to analyze the data and indicate which variables were most correlated with predicting “Deaths”.

****

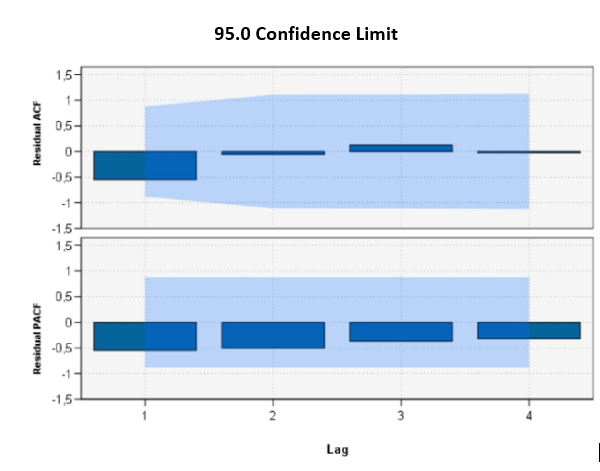
### Evaluation:

For this non-linear regression, the parameter estimates and residual sum of squares were calculated in The Goodness of Fit table, which shows whether the data sample represents the data one would expect to find in the actual population. The adjusted R squared for this dataset is .282, indicating there is some correlation with the variables and actual Deaths in reality. SPSS ran a t-distribution due to the fact that the data was aggregated and therefore a “smaller” size. The standard error of this dataset was estimated to be 73502,374. The beta coefficients for the standardized variables after running the regression analysis were .208 for Very Unhealthy Days, .089 for Good Days, .597 for Moderate Days, and -.294 for Unhealthy days. Based on this Beta, Good Days, or better air quality, is farthest from the target’s mean, indicating it is the least correlated with “Deaths”

**Modeling Building: Time Series for AQI vs Demographic/Biofuel Consumption**

**Algorithm**

Aside from nonlinear regression, we were interested in comparing the mean of the yearly median AQI, against other variables across all years from 2014 to 2018. For this reason, we decided to conduct a time series regression, first looking at the confidence limit. The confidence helps us assess the estimate of median AQI to the observed variable inputs We imported the combined Demographic, Energy Consumption and AQI dataset into a variable file node. Then, within the Time Series node, we ensured ‘Observations’ were in yearly time intervals, ‘Median\_AQI\_Mean’ was the target, and “BFTCB” (Biofuel total Consumption), “Deaths”, “International Increase”, “Births” and “Domestic Increase” were the input variables.



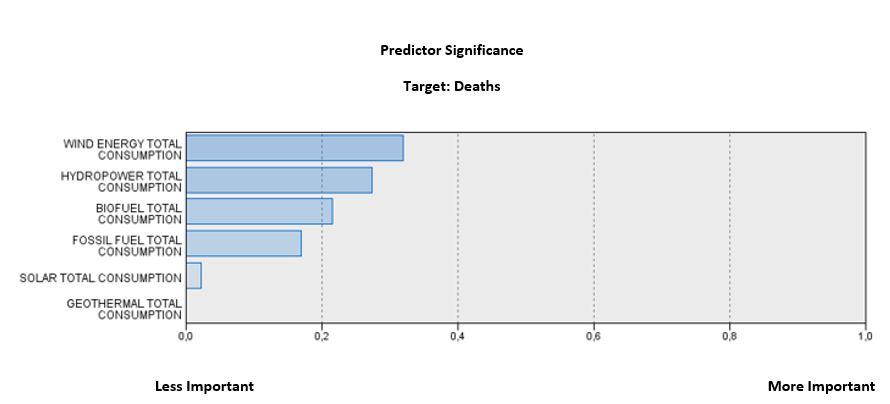
**Evaluation:**

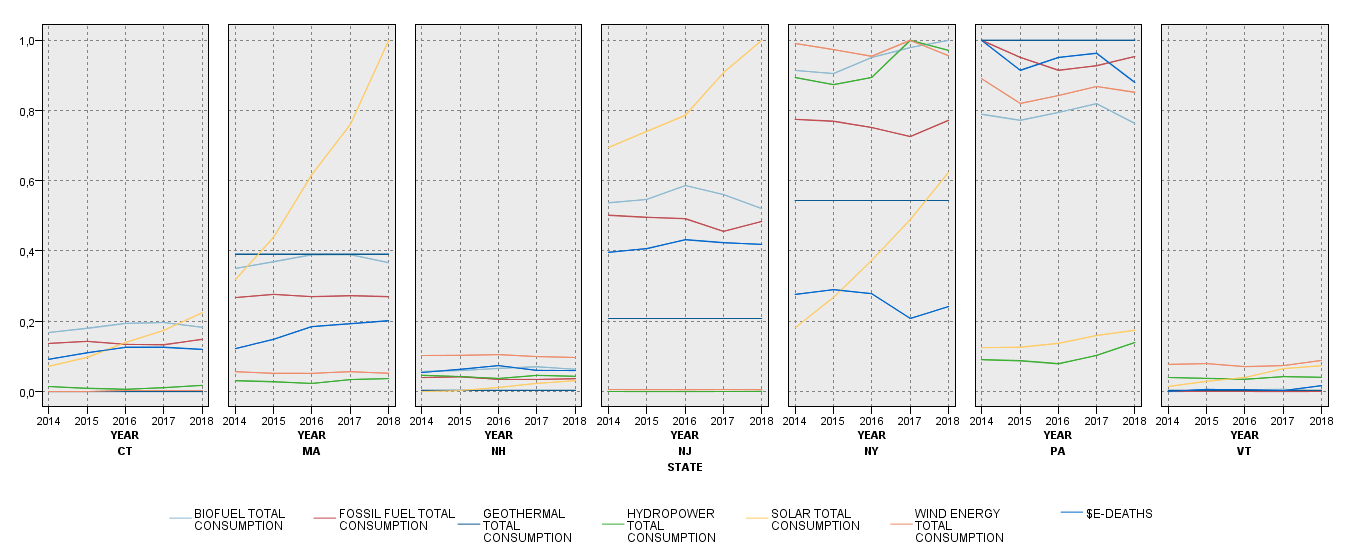
The ACF time series measures the average correlation between data points in a time series measured at different lag lengths. PACF is similar but each partial correlation is controlled for any correlations between lag lengths. We used this method to create a realistic estimate of the deaths by population parameter, represented by the time intervals. Because our aggregated sample is small and we are working with mean values, we decided to use t-distribution tests. For our data, the significance level was 0.713, so the confidence that Deaths are correlated with AQI levels is about 30%.

**Model Building: Demographics for Nonlinear Regression: Deaths vs Energy Sources**

### Algorithms

As governments and businesses seek more renewable energy sources and stray from fossil fuels, we wanted to reveal if this shift has resulted in any correlation with population deaths. For this reason we conducted a nonlinear regression analysis, to determine which energy source predicted deaths the most. First, we imported the combined Demographic, AQI, and Energy Sources dataset into a variable file node. The Type node was connected to indicate “Deaths” as the target, and “Wind Energy Consumption”, “Hydropower Consumption”, “Biofuel Consumption”, “Fossil Fuel Consumption”, “Solar Power Consumption”, and “Geothermal Consumption” as the input variables. Next, a Regression node was connected to analyze the data and indicate which variables were most correlated with predicting “Deaths”.





**Evaluation:**

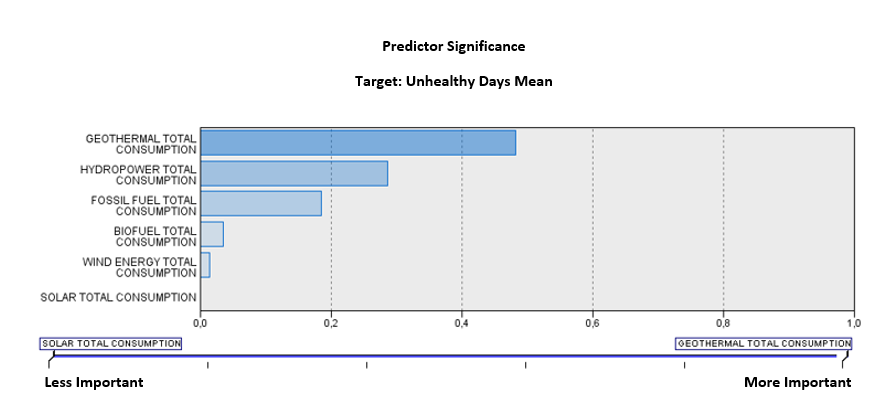
For this non-linear regression, the adjusted R squared for this dataset is .977 in The Goodness of Fit table, indicating there is a high degree of correlation with the variables and actual Deaths in a population. The standard error of this dataset was estimated to be 13017.069. The beta, or risk coefficients for the standardized variables after running the regression analysis were .608 for Biofuel Consumption, .292 for Fossil Fuel Consumption, -.487 for Geothermal Consumption, and -.1.336 for Hydropower Consumption, .124 for Solar Consumption, and 1.272 for Wind Consumption. Additionally, the t-score measures how close a variable is to the mean of Deaths. In this case, Wind Energy was the closest at .015, while Solar Energy was the farthest, at -2,245. This tells us that Wind Energy is the most correlated with predicting Deaths, while Solar is the least.

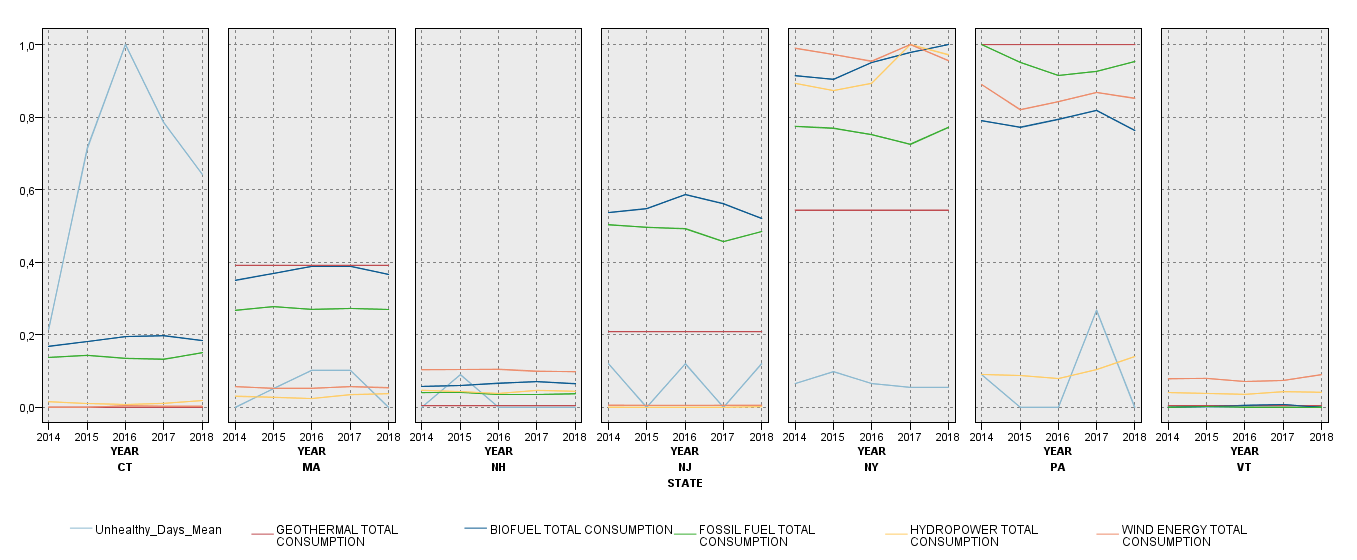
In the graphs generated by the nonlinear algorithm, Death versus Energy Consumption is displayed by State annually from 2014 to 2018. This model visually represents the independent variable “Deaths” as a horizontal blue line, and reveals the trend in each energy consumed over the years. Solar energy is increasing the greatest across most of the states. The graph also shows how states that increased their fossil fuel consumption over the time period at a higher rate, also accounted for more deaths such as in New York, than its counterpart in New Hampshire, for example.

### Model Building: Consumption for Nonlinear Regression: Unhealthy Days (Mean) vs. Energy Consumption

**Algorithm**

To discern whether energy consumption has any impact on air quality, and it’s correlation with the number of days labeled as “Unhealthy” over a year, we constructed a nonlinear regression for Unhealthy Days Mean versus Energy Consumption. First, we imported the combined Demographic, AQI, and Energy Sources dataset into a variable file node. The Type node was connected to indicate “Unhealthy Days Mean” as the target, and “Wind Energy Consumption”, “Hydropower Consumption”, “Biofuel Consumption”, “Fossil Fuel Consumption”, “Solar Power Consumption”, and “Geothermal Consumption” as the input variables. Next, a Regression node was connected to analyze the data and indicate which variables were most correlated with predicting Unhealthy air quality in New England.





### Evaluations: Unhealthy Days (Mean) vs. Energy Consumption

For the non-linear regression table, an ANOVA analysis was conducted which measures the variance of the target variables against the number of Unhealthy Days. The t-distribution test which measures the degree to which a variable is far from the mean of the target, found Geothermal and Solar Energy at opposite sides of Unhealthy Days’ mean. Geothermal’s t-score was -1,182, while Solar’s was -2,245. Interestingly, Fossil Fuel Consumptions was near the mean at -.998. The standard error of this dataset was estimated to be 13017.069. The beta coefficients for the standardized variables after running the regression analysis were .608 for Biofuel Consumption, .292 for Fossil Fuel Consumption, -.487 for Geothermal Consumption, and -.1.336 for Hydropower Consumption, .124 for Solar Consumption, and 1.272 for Wind Consumption. Overall, because our dataset was small, there is a high margin for error that Unhealthy Days is predicted by any of the variables.

## 

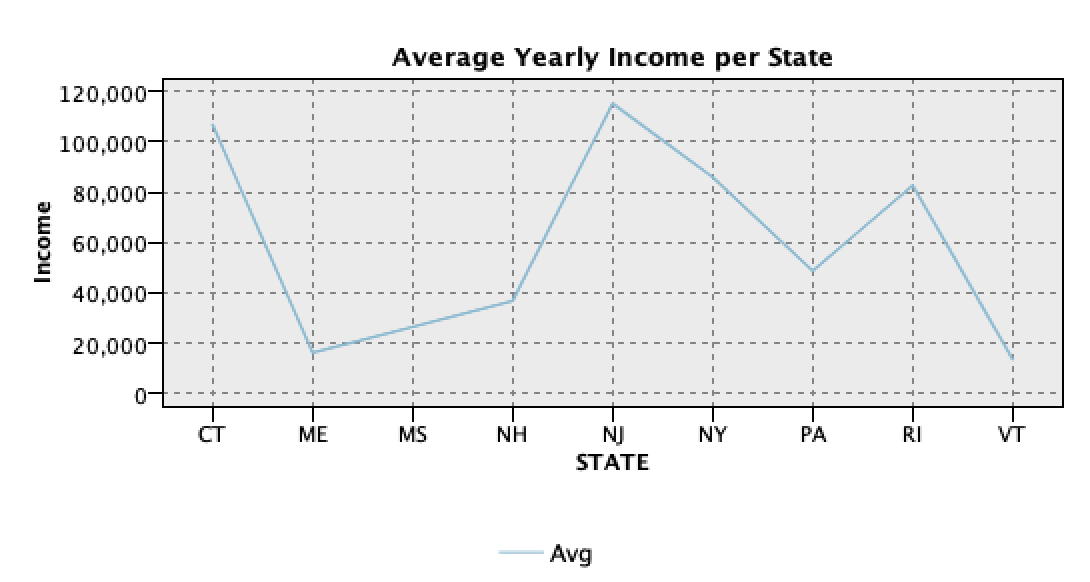
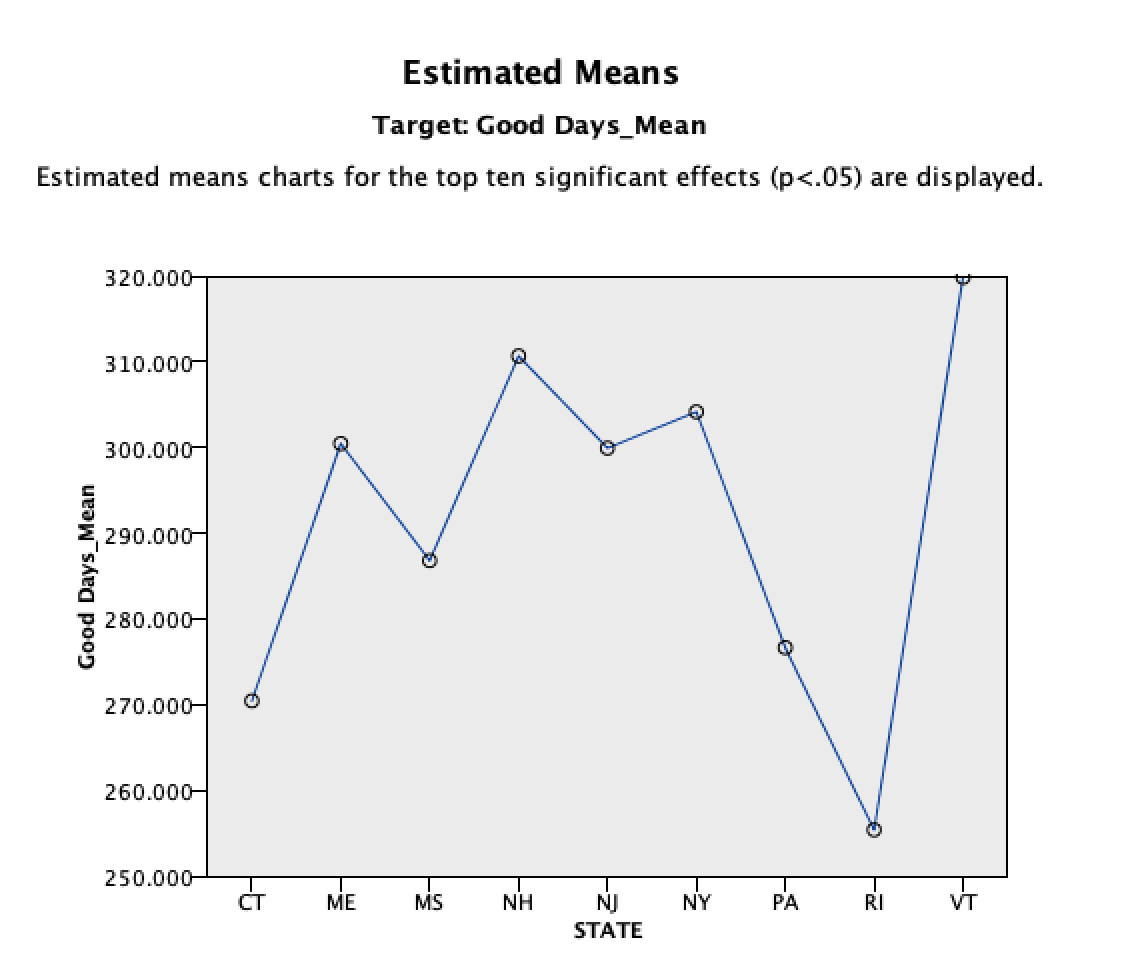
### Model Building: Income vs AQI

### Algorithms

Income normally plays a significant role in the health of a human, so this analysis was constructed to see if there was a correlation between finances and the air quality. For ease, first the taxes dataset was altered in excel to get a cumulated mean for the 5 years, instead of having every year in its own field. Using SPSS, the datasets for the average yearly income of the northeast region and AQI dataset, both imported through a variable file node. The AQI dataset was aggregated from multiple cities per state into just northeastern states by using the key field state, and the aggregated field of good days to mean. The two sets were merged so the state, average yearly amount of good days and average yearly income were condensed. Using the linear node, the target was the mean of good days, with the inputs of state and average yearly income. From there, a multiplot graph was connected to show correlation between state and average yearly income.

### Evaluation:

Since this analysis was to show how income affects air quality, there wasn’t a need for predictive analytics. This analysis was descriptive, rather than predictive or prescriptive, where models like regression would be a better fit. The most useful graph within the linear model was used to show the state to average yearly good days. The mean yearly good days and mean yearly income could not both be on the multiplot graph due to the scale of the numbers, one being in the hundreds and the other being in the tens or hundreds of thousands. A better analysis would be conducted by analyzing the two separate graphs with appropriate scaling for each.



**Results for all Analysis:**

The state of air quality has a significant impact on our general health. More contaminated air can bring about underlying diseases such as asthma, and exacerbate preexisting conditions such as lung disease. It was thus not surprising our results found that Very Unhealthy Days play the highest role in predicting Deaths. On the opposite side of the spectrum, the mean of Good Days played the smallest. This leads us to conclude the greater amount of days where AQI is labeled as “Very Unhealthy” is correlated with a greater amount of deaths in that specific state. Additionally in the time interval analysis, the confidence of Deaths correlated with AQI is about 30%. Further, states with higher AQIs (worse air) are more likely to have larger metropolitan areas than states with lower AQIs. Living in cities already comes with a number of risks itself, such as being exposed to crime, accidents, and diseases. It should therefore be noted, it is difficult to derive whether the exact cause of death is the result of poor air quality, or other extenuating circumstances.

Next, we analyzed if the source of energy had any correlation in predicting deaths within a given population. Our results concluded that over the four years, consumption of wind energy played the greatest role in predicting deaths, while geothermal consumption plays the smallest. This leads us to believe the more a state consumes wind energy, the higher amount of deaths there will be among the population. This evidence goes against our original hypothesis that fossil fuel consumption would result in the most deaths. As previously mentioned, it is difficult to derive whether the exact cause of death is the result of poor air quality or the energy source within a state, especially since our analysis assumes the cleaner energy sources lead to higher deaths, despite fossil fuels posing the greatest risk to human and climate vitality in common discourse regarding the subject.

Predicting the frequency of good or unhealthy air quality days based on a state’s energy consumption was of particular interest to determine whether alternative energy is the best solution to counter air pollution. Interestingly, we found that geothermal energy consumption was most correlated with unhealthy days, while solar energy consumption corresponds with good days. Again, this result goes against our initial hypothesis that fossil fuels create a higher frequency of unhealthy days. Despite the evident contribution of air pollution from burning these fuels, they did not stray far from the mean of the target variable. Our dataset was small, however, so there is a higher probability of error. Additionally, using Geothermal energy may not contribute as much air pollution, but it may be used more in cities that use the energy to power buildings, where there are also thousands of cars burning fossil fuels for power. This parallel could result in a higher number of unhealthy days where geothermal energy in particular is used at a much higher rate than fossil fuels are used in less urban areas.

For the analysis measuring income vs AQI, since income normally provides easier access to a healthier lifestyle, it was surprising to see that an increase in income leads to a decrease in air quality. A state like Vermont does not have any major cities, and the lowest average income. However, this was the highest for air quality. Vermont is known for their outdoor activities, like skiing, which causes less air pollution than a big city full of cars and other emissions. Another surprising state was Connecticut. Although close to New York City and has the medium size city of Stanford, most of Connecticut is suburban space. With the second highest income, Connecticut also has the second lowest amount of good air quality days. This again goes to the belief of an indirect relationship between air quality and income. There are some outliers, like New York. The city holds a majority of the population in a highly polluted area with a high variance of income. The average shows the average yearly income being above average, yet there are an above average amount of good air quality days. Due to the variance in rural to urban areas and levels of income within the state, NY is an outlier. The majority of states show a correlation between a high income and low air quality.

# **Conclusion:**

Initially anticipating that high fossil fuel consumption would be the highest component of low air quality, this was in fact not always the case. Other sources of energy, like wind energy, were correlated with a high death rate. Higher average income normally leads to a higher air quality, but is excluded for states with larger cities like New York. The data also shows a correlation between pollution and an increase in deaths because in lower income states, there is a higher percentage of deaths. This can be due to lower income areas being around coal-powered fire plants or refineries. Both emit extremely hazardous chemicals to the air, and are normally located in less desirable areas. Lower housing costs are in less desirable areas, leading to the conclusion that poorer communities are exposed to more air pollution. The burning of fossil fuels has a wide ranging negative effect on our environment, but it is important that we switch over to environmentally friendly alternatives that are not associated with health hazards. Some biofuels are better than others, like solar power, but most are only considered a temporary fix. There are other variables not shown in this study, like land or water pollution, that can also be effects of energy sources. Although we are moving in the right direction by creating new types of energy, The United States needs to continue to find permanent solutions that can reverse the damage already caused to the environment, and learn how to create a climate neutral solution.

# **I. Appendix 1: Contributions of each group member. Who did what?**

Regan: Data Methodology for AQI dataset; all analysis besides tax vs AQI, additional work with Demographic and Energy Consumption datasets- used for analysis section

Maria: All data collection, Data Methodology for demographic dataset; general assistance for preprocessing

Annemarie: Data Methodology for tax dataset; analytics for tax vs AQI

Shu: Data Methodology for demographic dataset

Tova: Abstract, introduction, conclusion, Data Methodology for consumption (BTU) dataset

**References**

Demographic data

<https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/cities/totals/>

Tax data

<https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi>

Energy consumption

<https://www.eia.gov/totalenergy/data/monthly/index.php>

Daily AQI data

<https://www.epa.gov/outdoor-air-quality-data>

EDF:

<https://www.edf.org/health/why-smog-standards-are-important-our-health?utm_source=google&utm_campaign=edf_clean-air_upd_dmt&utm_medium=cpc&utm_id=1518122557&gclid=CjwKCAjw0On8BRAgEiwAincsHDIeFPMk3tBMuIm81XxfknjYHSdjQ_uUcRHO39otb3mSRo7_u2DXKhoC3u8QAvD_BwE>

UCSUSA

<https://www.ucsusa.org/resources/hidden-costs-fossil-fuels#39>

EPA-acid rain

Environmental Protection Agency (EPA). 2016. Acid rain. Washington, DC. Online at: <http://www.epa.gov/acidrain/>, accessed on May 3, 2016

Stanford: <https://earth.stanford.edu/news/how-much-does-air-pollution-cost-us#:~:text=Air%20pollution%20negatively%20impacts%20the,(%24790%20billion%20in%202014).&text=5%20are%20related%20to%20human,such%20as%20manufacturing%20and%20agriculture>.

Woodard et al : Woodard DL, Davis SJ, Randerson JT (2019) Economic carbon cycle feedbacks may offset additional warming from natural feedbacks. *Proc Natl Acad Sci USA* 116:759–764.[Abstract/FREE Full Text](https://www.pnas.org/lookup/ijlink/YTozOntzOjQ6InBhdGgiO3M6MTQ6Ii9sb29rdXAvaWpsaW5rIjtzOjU6InF1ZXJ5IjthOjQ6e3M6ODoibGlua1R5cGUiO3M6NDoiQUJTVCI7czoxMToiam91cm5hbENvZGUiO3M6NDoicG5hcyI7czo1OiJyZXNpZCI7czo5OiIxMTYvMy83NTkiO3M6NDoiYXRvbSI7czoyMDoiL3BuYXMvMTE2LzMvNzE0LmF0b20iO31zOjg6ImZyYWdtZW50IjtzOjA6IiI7fQ==)[Google Scholar](https://www.pnas.org/lookup/google-scholar?link_type=googlescholar&gs_type=article&author%5B0%5D=DL+Woodard&author%5B1%5D=SJ+Davis&author%5B2%5D=JT+Randerson&title=Economic+carbon+cycle+feedbacks+may+offset+additional+warming+from+natural+feedbacks&publication_year=2019&journal=Proc+Natl+Acad+Sci+USA&volume=116&pages=759-764)

EPA economic impact- <https://gispub.epa.gov/air/trendsreport/2019/#growth_w_cleaner_air>

Forward Selection: <https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/>