Neil Opitz
DSC 630 Final Project

Impact of Energy Production on Air Quality in the United States: 1990-2018

Impact of Energy Production on Air Quality in the United States: 1990-2018

Background

The goal of this project is to measure the impact that energy production has on air quality. The project will include megawatt hours production data from conventional energy sources such as coal and natural gas as well as alternative (green) energy sources such as wind and hydroelectric. The goal is to determine if transitioning to green energy sources will impact airborne pollutant levels, and if so, how much of an impact.

It is understood that there are factors beyond energy production that contribute to changes in air quality. While inclusion of all potential contributors is beyond the scope of this project, annual wildfire data from the state of California is going to be included due to the significant impact these events have upon air quality.

Impact of Energy Production on Air Quality in the United States: 1990-2018

AQI AIR QUALITY INDEX

Data: Energy Production

For this project I am using data collected over the past 30 years (starting in 1990):

The energy production data includes annual production rates for all energy types (both conventional and green) across all 50 states in the United States.

The values in the raw data are comprised of total megawatt hours of energy produced for each energy type for each state during calendar years beginning in 1990 and ending in year-end 2018.

The data was obtained from the United States Energy Information Administration website. The data is contained in a single excel file.

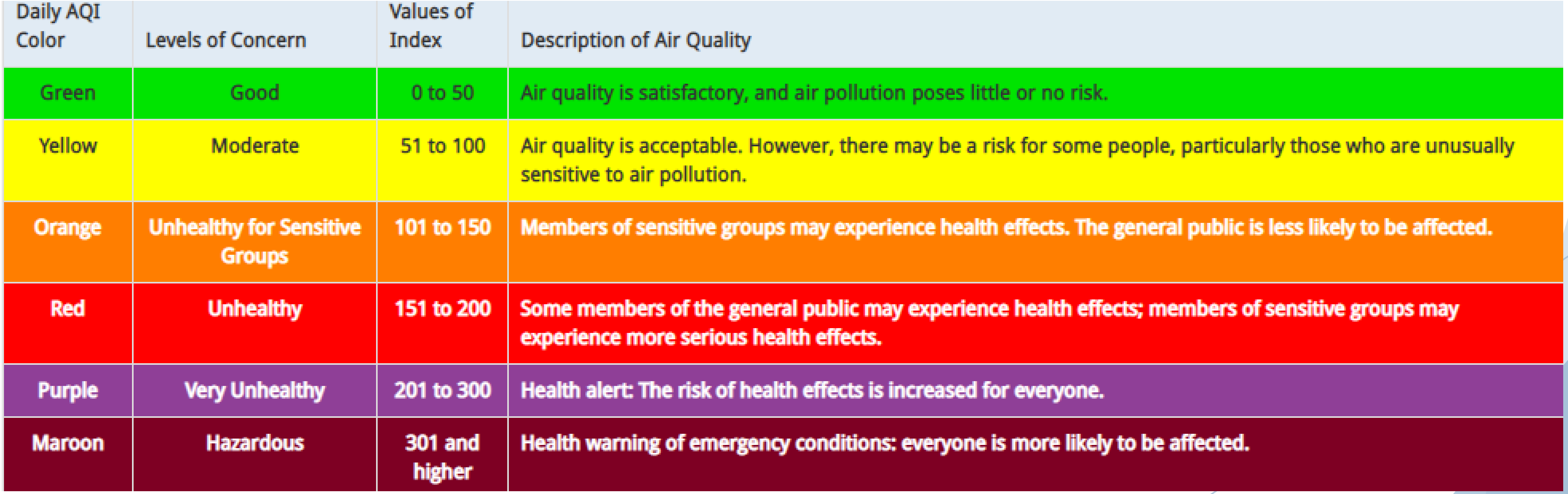
Impact of Energy Production on Air Quality in the United States: 1990-2018

Data: Air Quality

Air Quality Index data is comprised of measurements from air quality monitoring stations from hundreds of locations across the United States. The AQI is a score that can range from 0 to 500 based on the amounts of pollutants captured at a monitoring station, with 0 indicating no pollutants and 500 representing extremely hazardous levels of pollution.

The AQI data provides median values for each monitoring station as well as the number of days that exceed defined thresholds: Moderate, Unhealthy, Hazardous. AQI data for this project includes the years 1990 through 2018.

The air quality data was obtained from AirNow.gov, which contains data from the United States Air Quality Index. The data is contained in individual annual files.



Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

DO NOT STRAP DOWN ACROSS THE ROOF
STRAP OVER THE PALLET
LSA
Lear Siegler Australasia
AIR QUALITY MONITORING STATION

Impact of Energy Production on Air Quality in the United States: 1990-2018

Data: California Wildfires

The California wildfire data represents the total number of acres of land that were impacted by wildfires in the state of California on an annual basis. The California wildfire data includes total acres burned for each year starting in 1990 through 2018.

The wildfire data was collected from the California Department of Forestry and Fire Protection. The data was extracted from a pdf table which contained total acres of land burned in California wildfires dating back to 1987.

Impact of Energy Production on Air Quality in the United States: 1990-2018

AIR QUALITY INDEX

Data: Green New Deal

The Green New Deal is a resolution that proposes that human activity is the dominant cause of observed climate changes over the past 100 years and would require that 100% of power produced in the United States be from 'green' energy sources such as solar, wind, and geothermal.

The Green New Deal data is a quantification of the verbiage proposed in the language for the Green New Deal.

This data was obtained from the United States Congressional website.

Impact of Energy Production on Air Quality in the United States: 1990-2018

Data

Explanatory Variables

I will be using the energy production data and the California wildfire data as explanatory variables.

I created feature variables from the energy production data, including a summation of all green energy sources into a single variable.

Other feature variables include converting the raw megawatt hour data into percentages which will detail the rate of relative change in the production of green energy and other energy sources at the state level across the timeframe.

Response Variables

Response variables for this project come from the Air Quality Index data.

I used the annual Median AQI values as a response variable by creating a binary outcome. The response variable is coded 0 if the median AQI value is ≤38 and 1 if it is greater than 38. Using this value provided a large sample of values in both buckets.

I created a feature response variable which calculates the percentage of monitored days that were classified as “unhealthy” for each year. The AirNow website indicates that a day is considered ‘unhealthy’ if the monitored result was 100 or greater at any time during the day.

Data Preparation

Briefly, data preparation consisted of:

Energy Production Data

- Removing redundant data rows
- Convert from a vertical setup to a horizontal setup with 1 row for each state for each year

Air Quality Index Data

- Merging 30 years of Air Quality Index data
- Aggregating county-level data at the state level

Wildfire Data

- Scrape data from a pdf table into a dataframe
- Remove unnecessary columns and rows

Each dataset includes a ‘YEAR’ and ‘STATE’ variable which was used to merge the three datasets into a single dataframe. Each row contained aggregated annual data for each individual state.

Exploratory Data Analysis

Check for missing data

1. As the Air Quality, Energy Production, and Wildfire data are all a census of the data (and not a sample) I have all of the data for each of these data sources.
2. A check for missing data was conducted for each variable in each data set (Air Quality Index, Energy Production, Wildfire)
   - No missing data was found
   - In some instances the value is zero and not missing (for example the state of Alaska generated zero megawatts of wind power in the year 1990).

Data frequencies

1. Frequency counts of ‘State’ data indicated several locations that are in exception to the 50 states of the United States for which I am conducting the analysis.
   - I removed data for Guam, Canada, District of Columbia, Puerto Rico, Mexico, and the Virgin Islands.
2. I also created a dictionary to convert state names into abbreviations for purposes of merging the data as one data source used state abbreviation while another used state name.

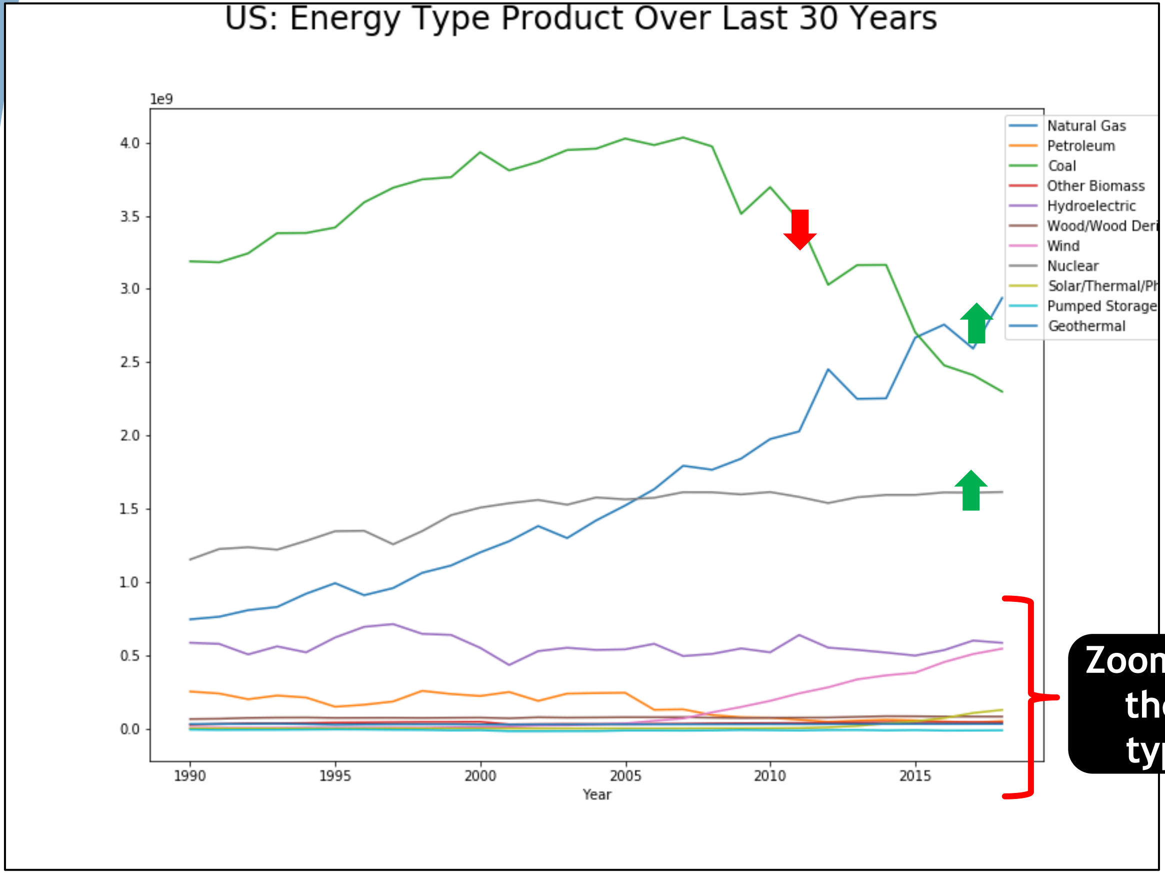
Impact of Energy Production on Air Quality in the United States: 1990-2018

Exploratory Data Analysis

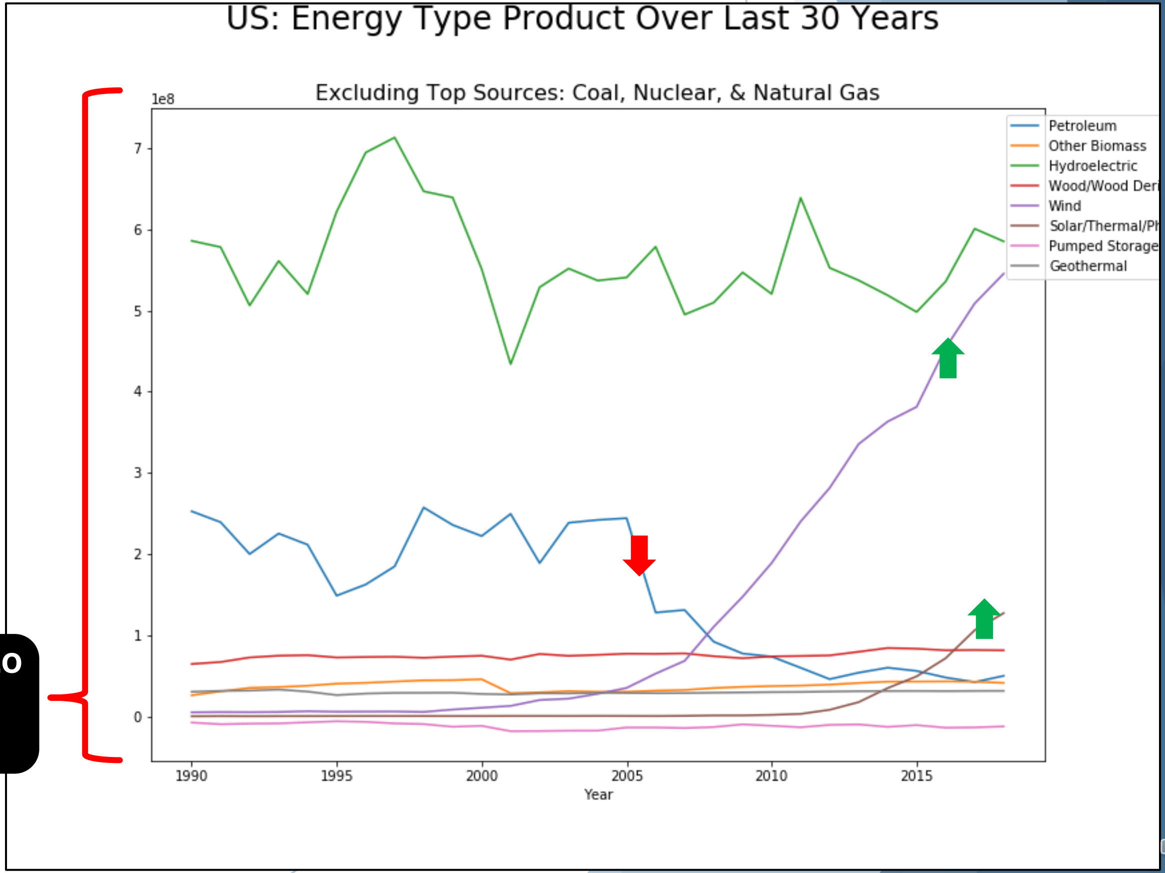
1. To view production trends, I created charts for the annual sum of megawatt hour production for each energy type across all 50 states
    1. The left chart is all energy types while the chart on the right highlights the lower value types shown at the bottom of the left chart.

⬇ Production for coal and petroleum are in decline

⬆ Production for natural gas, nuclear, wind, and solar/thermal/photovoltaic are increasing



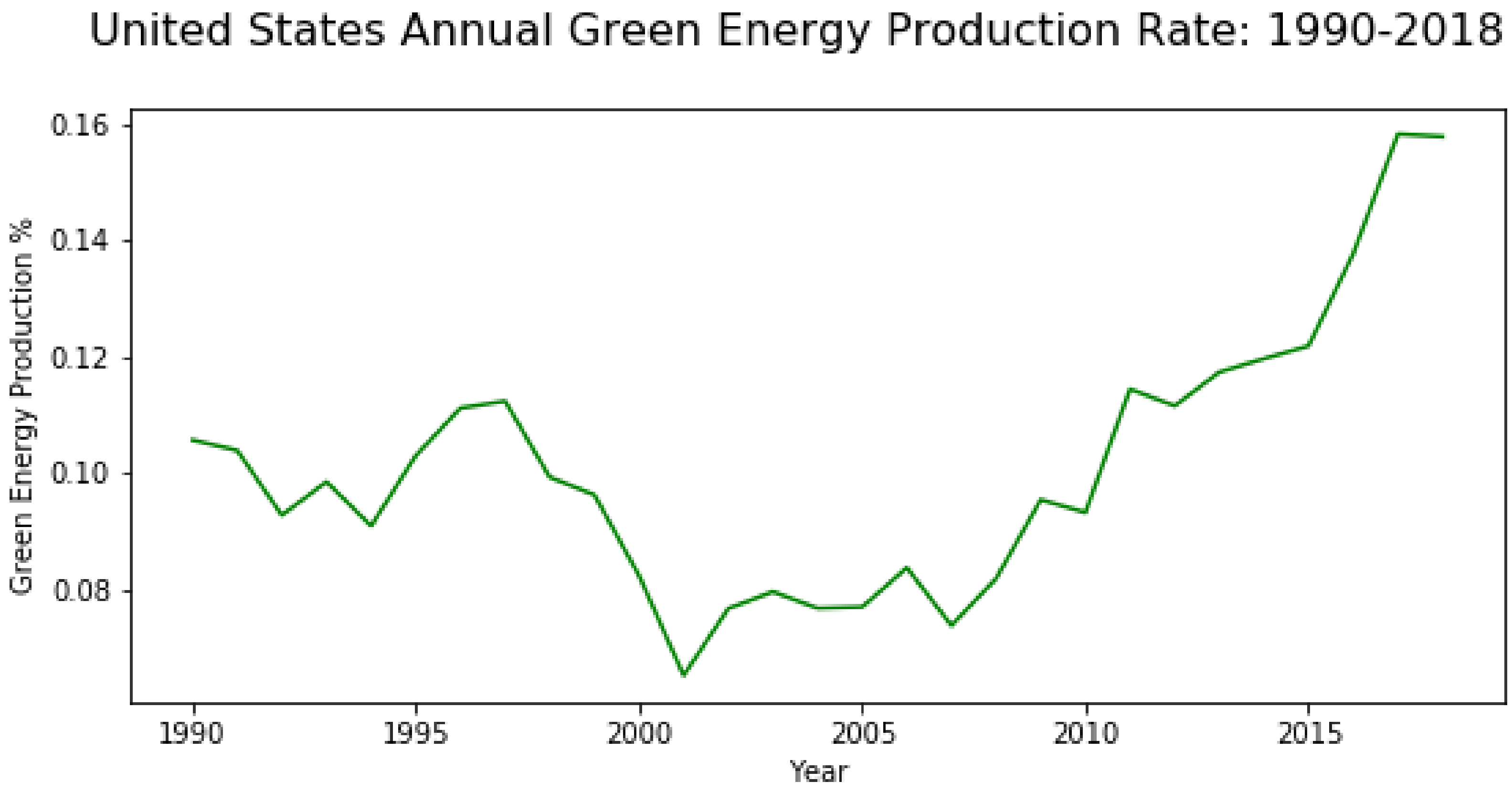
Zoom into these types



Impact of Energy Production on Air Quality in the United States: 1990-2018

Exploratory Data Analysis

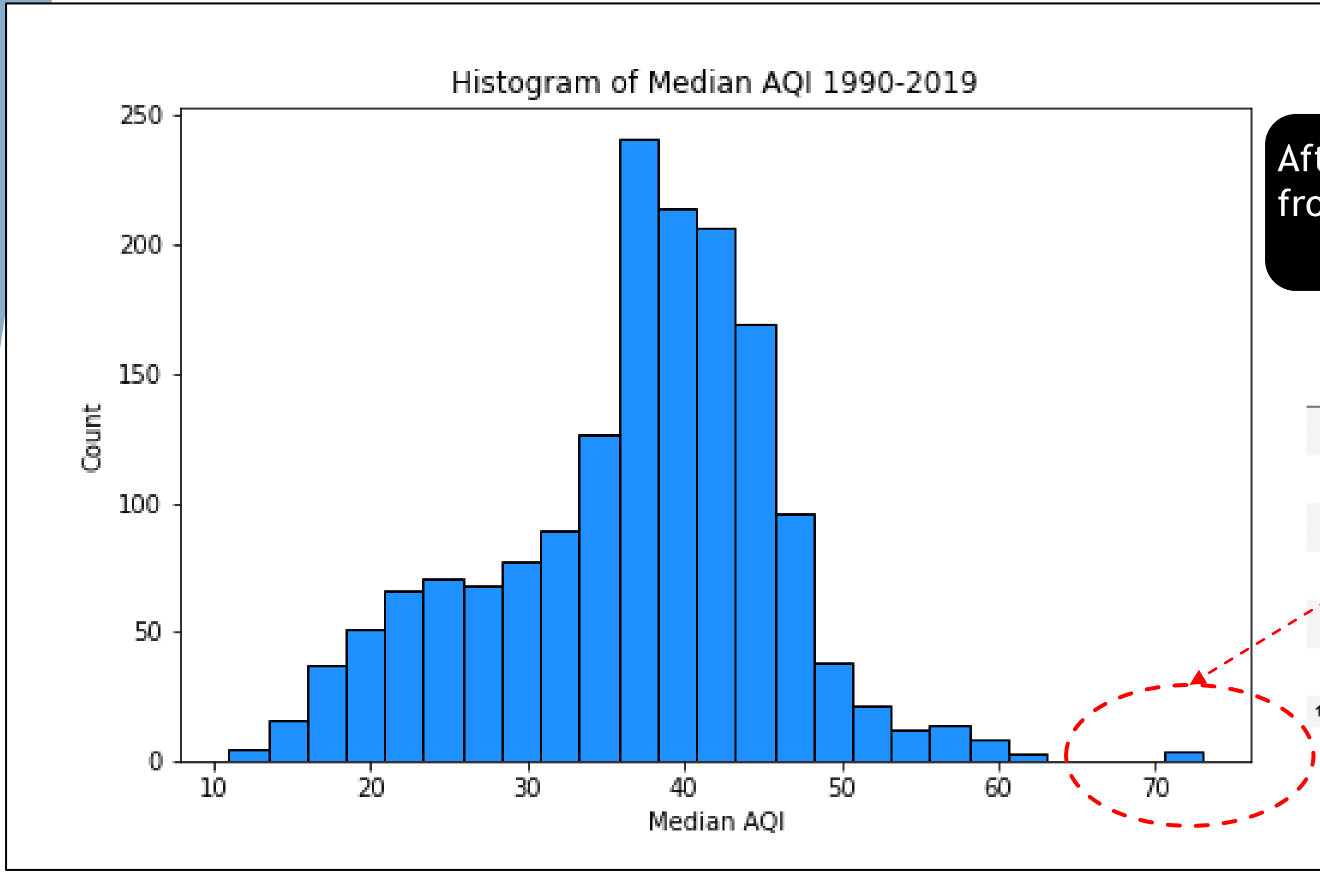
While the relative production of ‘green’ energy declined in the late 1990s and early 2000s, it has been increasing since 2008.



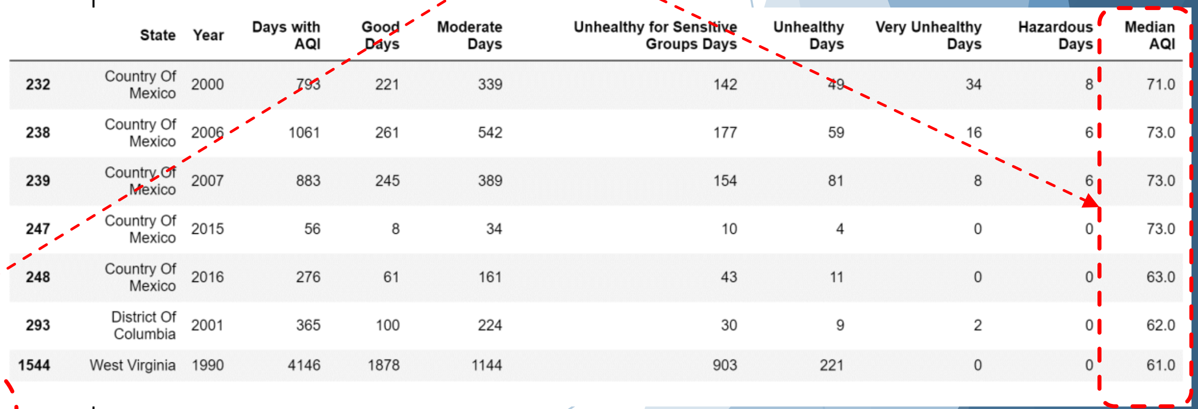
Impact of Energy Production on Air Quality in the United States: 1990-2018

Exploratory Data Analysis

1. To gain a better understanding of the Air Quality Index values, I created a histogram to see the distribution of values.
    1. The histogram indicates that median AQI values have a slight negative skew and the presence of an outlier in the >70 range.



After a deeper look, it turns out that outlier Median AQI values are from Mexico and are going to be removed from the dataset as I am only using data from the 50 United States.



	State	Year	Days with AQI	Good Days	Moderate Days	Unhealthy for Sensitive Groups Days	Unhealthy Days	Very Unhealthy Days	Hazardous Days	Median AQI
232	Country Of Mexico	2000	793	221	339	142	49	34	8	71.0
238	Country Of Mexico	2006	1061	261	542	177	59	16	6	73.0
239	Country Of Mexico	2007	883	245	389	154	81	8	6	73.0
247	Country Of Mexico	2015	56	8	34	10	4	0	0	73.0
248	Country Of Mexico	2016	276	61	161	43	11	0	0	63.0
293	District Of Columbia	2001	365	100	224	30	9	2	0	62.0
1544	West Virginia	1990	4146	1878	1144	903	221	0	0	61.0

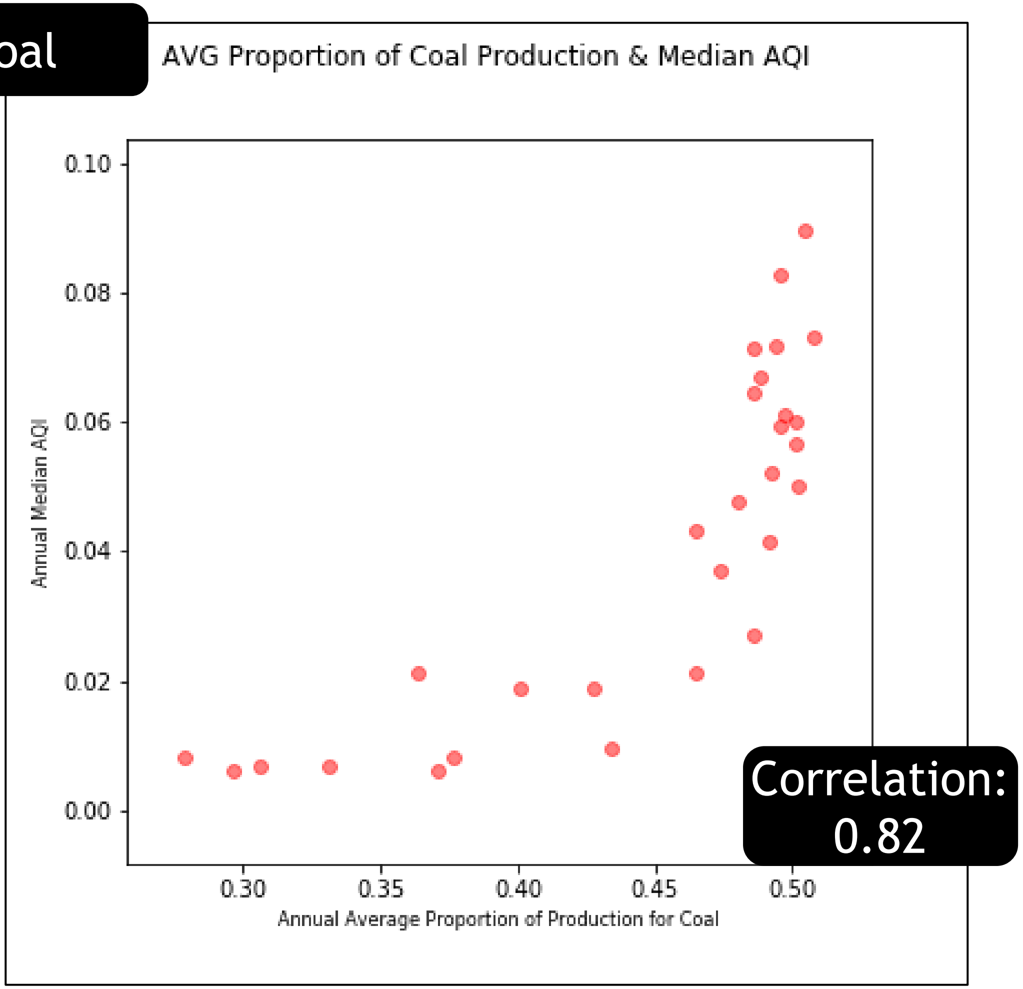
Impact of Energy Production on Air Quality in the United States: 1990-2018

AIR QUALITY INDEX

Exploratory Data Analysis

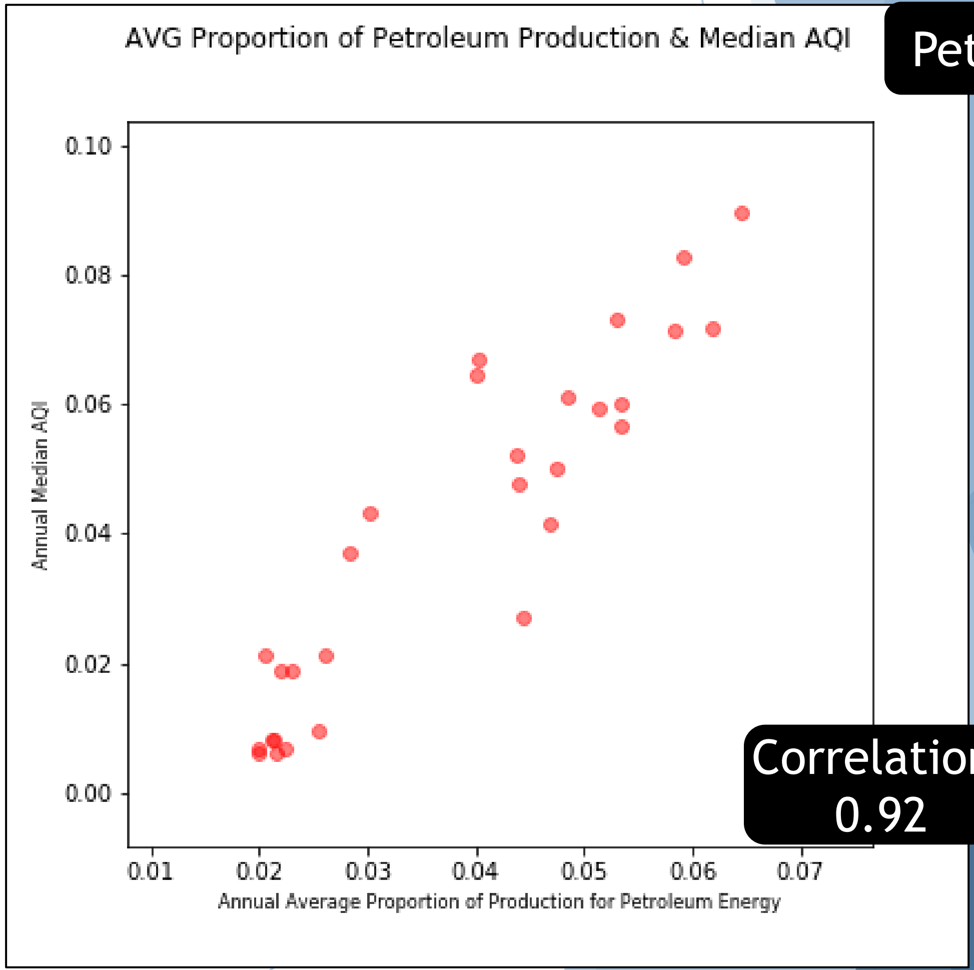
Data indicates strong positive relationships between production of coal and petroleum energy sources and higher AQI levels (higher pollution levels).

Coal



Correlation: 0.82

Petroleum



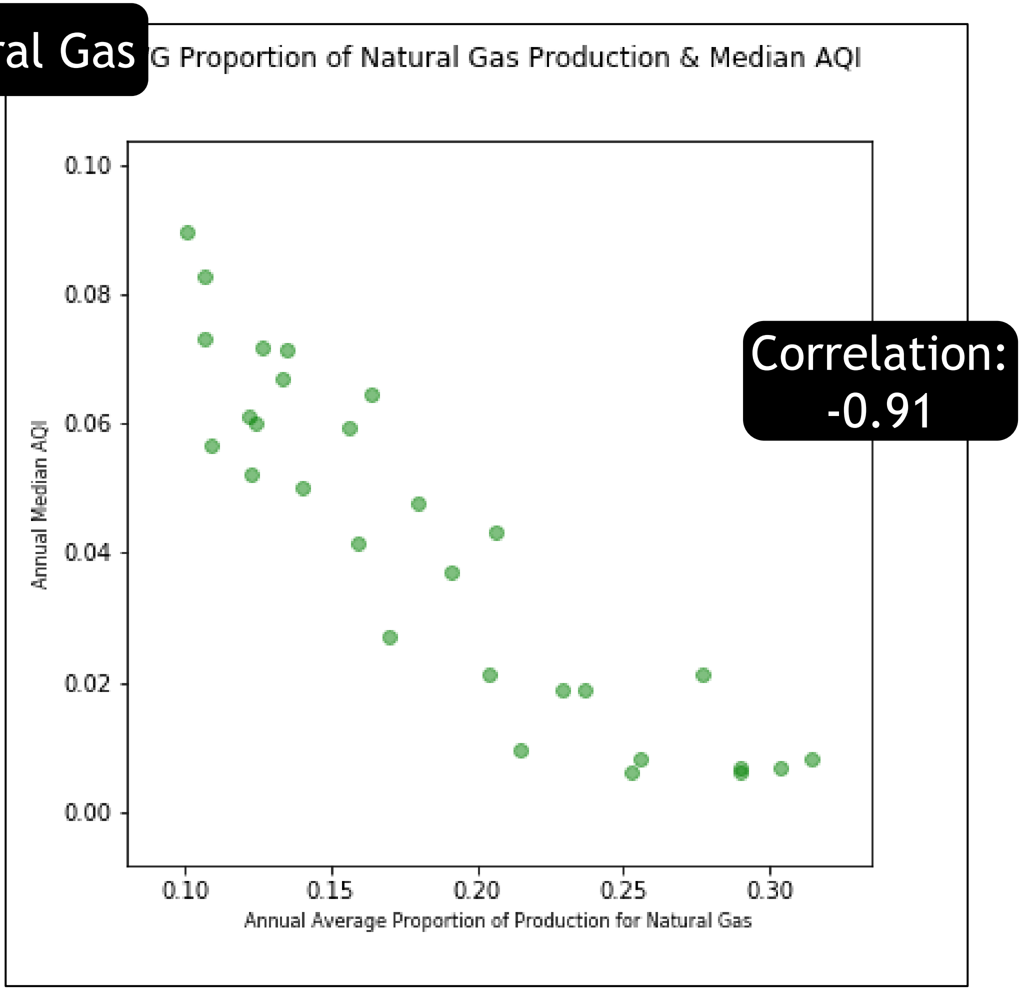
Correlation: 0.92

Impact of Energy Production on Air Quality in the United States: 1990-2018

Exploratory Data Analysis

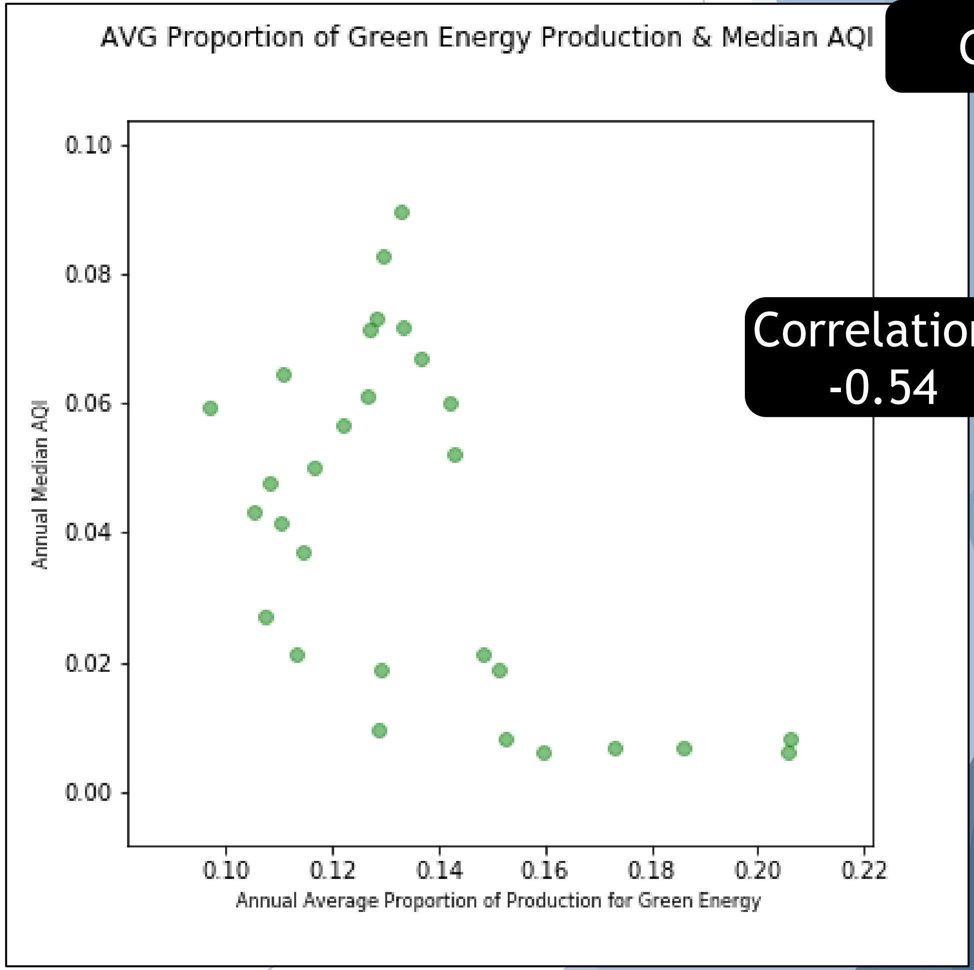
Conversely, there are strong negative relationships between production of natural gas and green energy sources and lower AQI levels (lower pollution levels).

Natural Gas



Correlation: -0.91

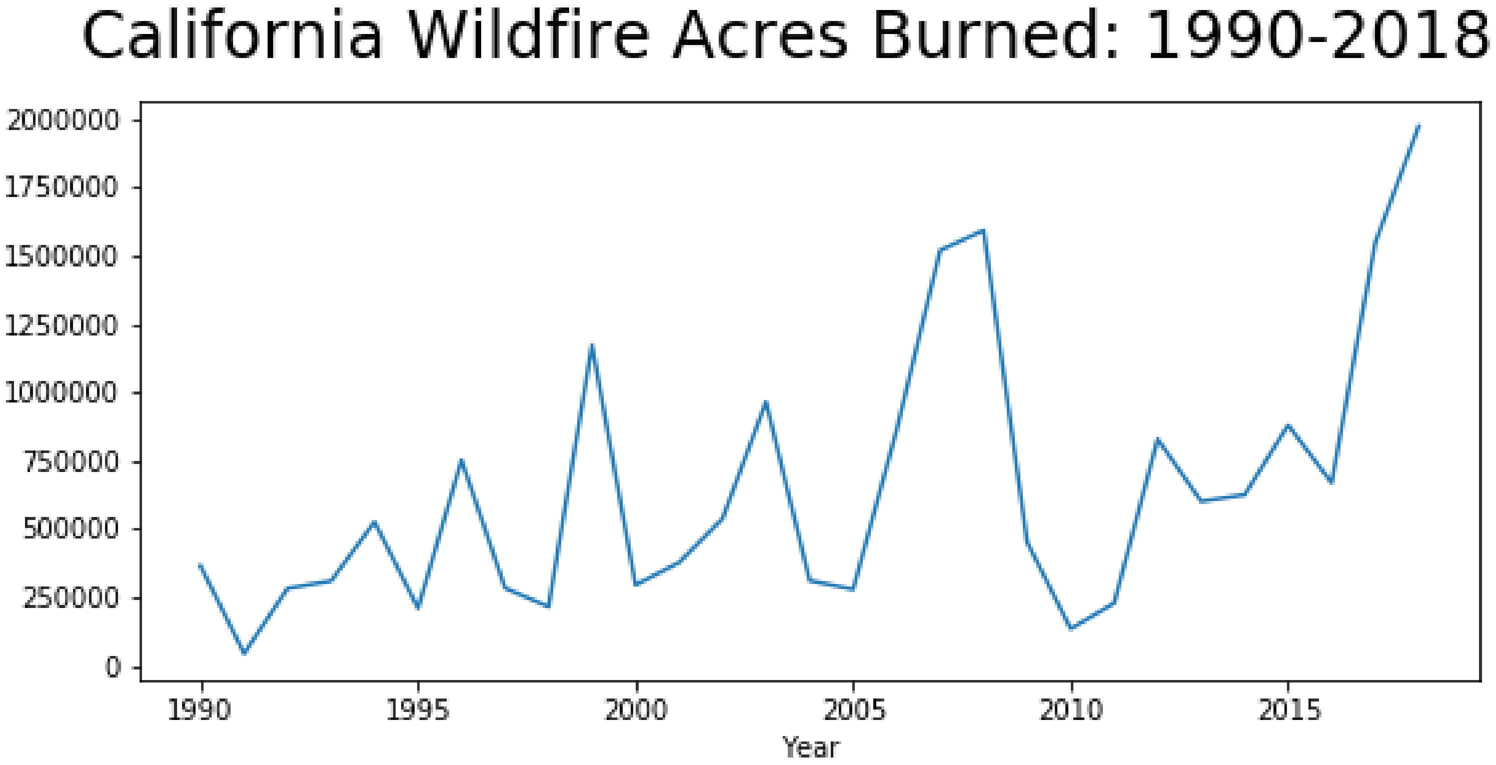
Green



Correlation: -0.54

Exploratory Data Analysis

California wildfire data shows a fluctuating, but clearly upward trend, in the number of acres burned due to wildfires since 1990. The data indicates a spike in the number of acres burned about very 3 to 4 years.

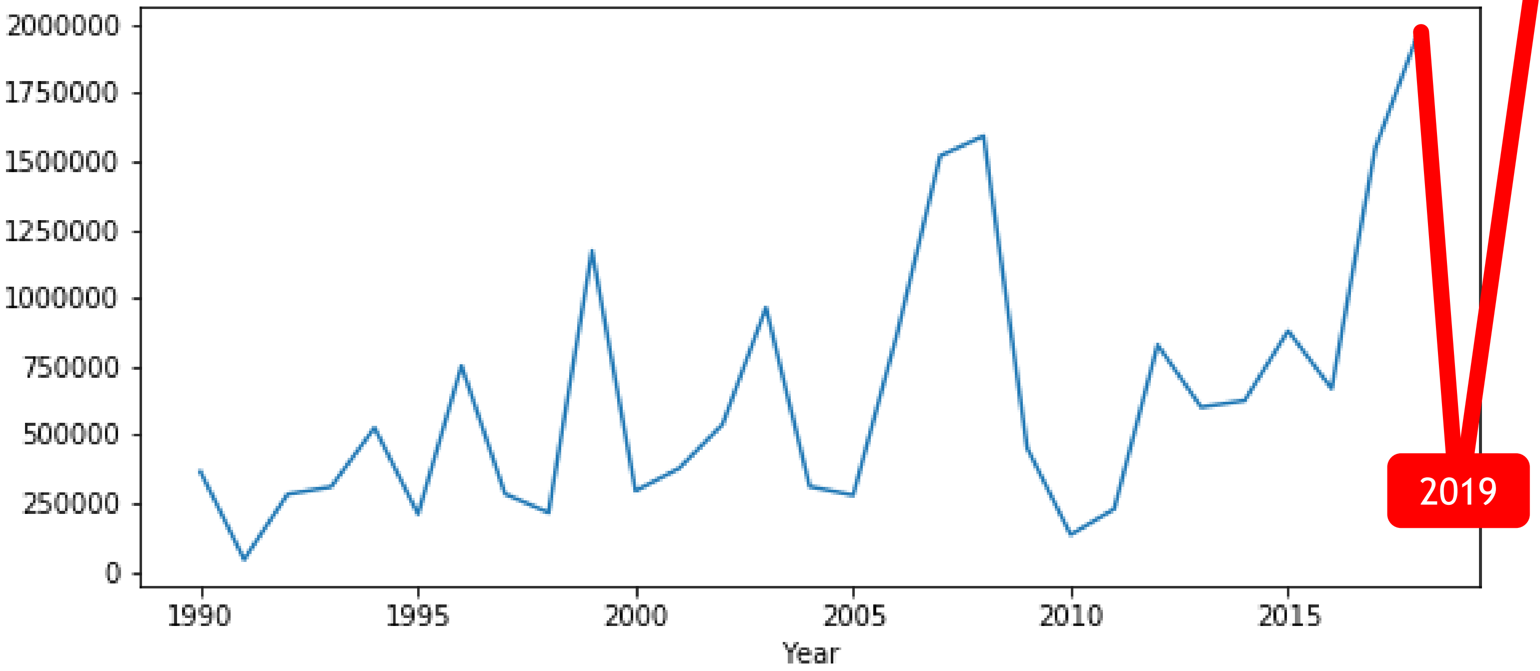


Impact of Energy Production on Air Quality in the United States: 1990-2018

Exploratory Data Analysis

If 2019 and 2020 acres burned were added to this chart

California Wildfire Acres Burned: 1990-2018



2020

4,400,000

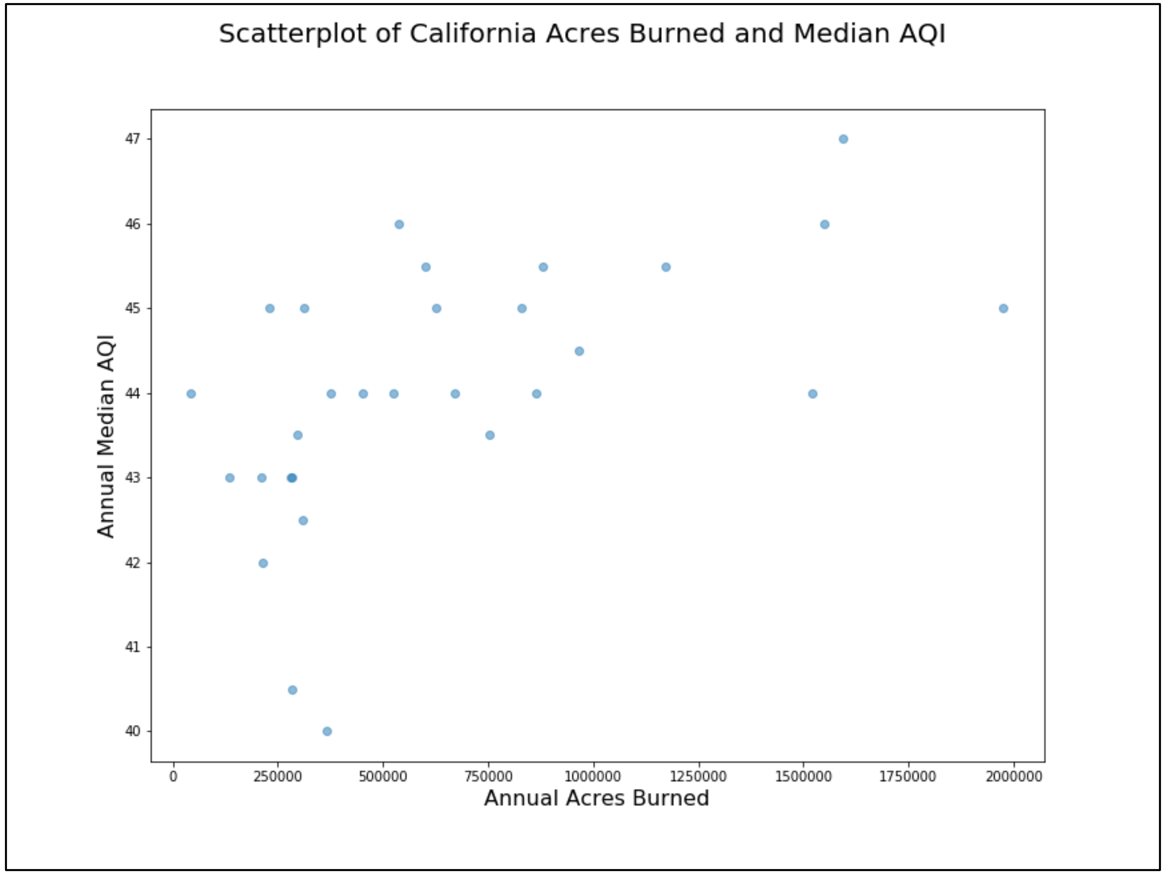
2019

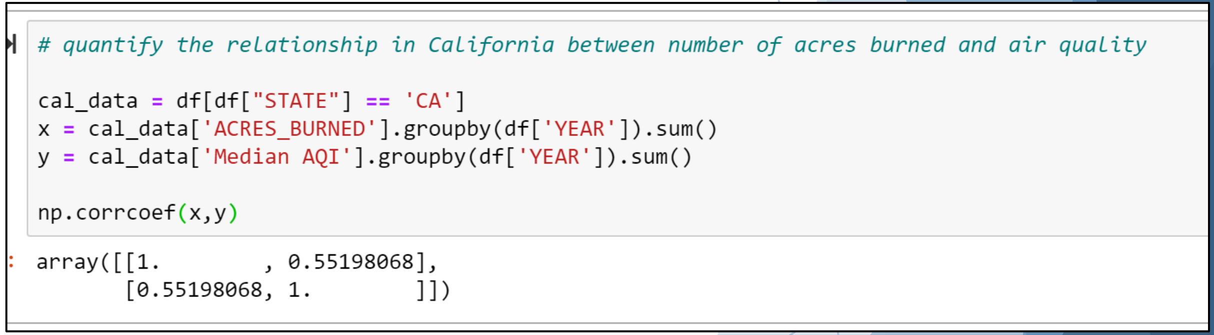
260,000

Impact of Energy Production on Air Quality in the United States: 1990-2018

Exploratory Data Analysis

Isolating the data to the state of California reveals a relationship between the number of acres burned in wildfires and the median Air Quality Index values. Analysis of the correlation coefficient indicates a moderate relationship (0.55) between the number of acres burned and air quality.


Scatterplot of California Acres Burned and Median AQI



```
# quantify the relationship in California between number of acres burned and air quality

cal_data = df[df["STATE"] == 'CA']
x = cal_data['ACRES_BURNED'].groupby(df['YEAR']).sum()
y = cal_data['Median AQI'].groupby(df['YEAR']).sum()

np.corrcoef(x,y)
```

```
array([[1.        , 0.55198068],
       [0.55198068, 1.        ]])
```

Impact of Energy Production on Air Quality in the United States: 1990-2018

Modeling

I used two different predictive modeling algorithms for this project:

- Linear Regression
- Binomial Logistic Regression

I ran the predictive models across the entire dataset and then using data for the state of California as this state has relatively high levels of green energy production across the timeframe.

The outcome variable for the linear regression model is the percent of unhealthy days monitored for one model and median AQI for a second model.

The outcome for the logistic regression dataset is the median AQI rate higher or lower than 38.

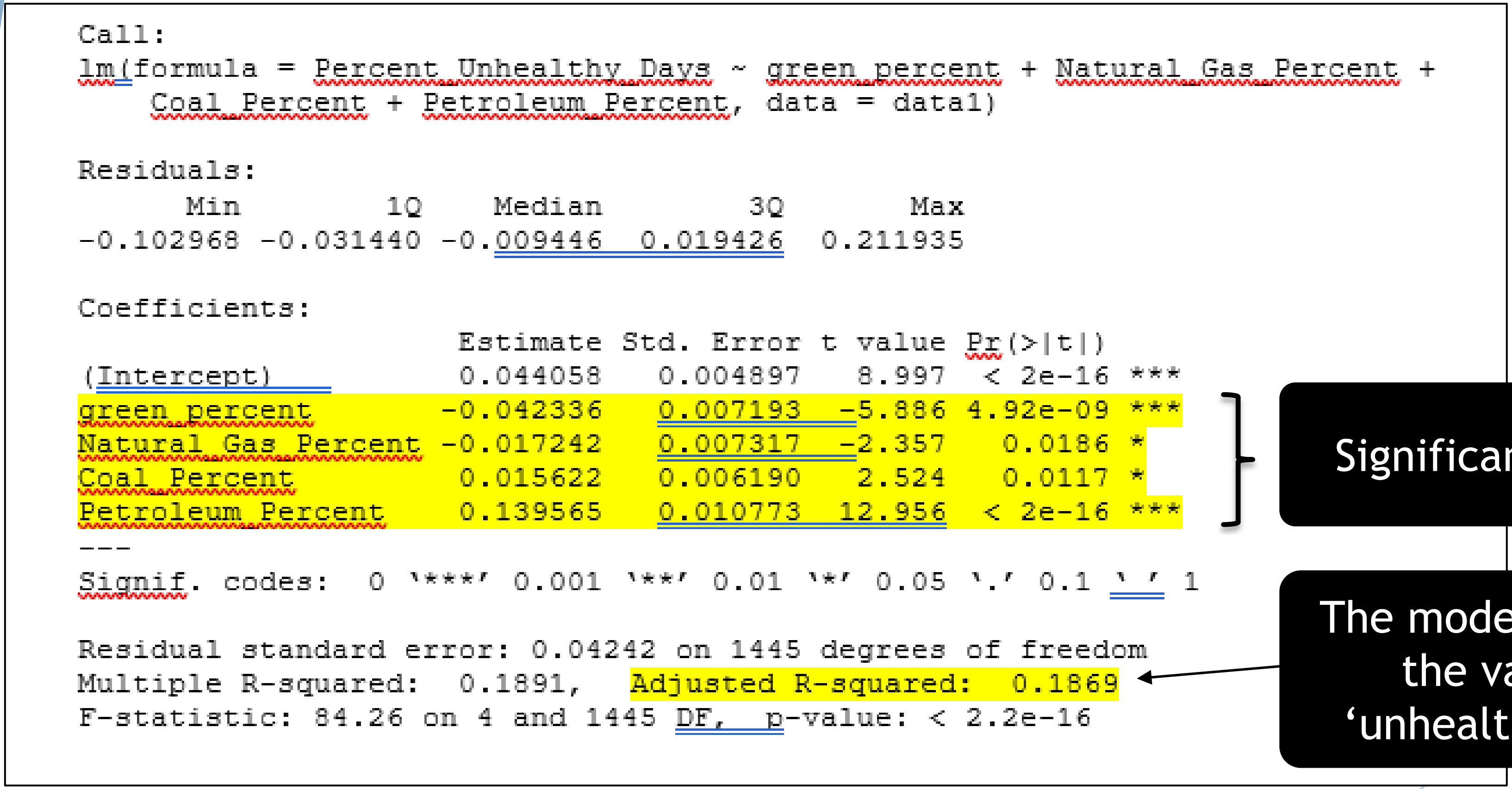
I created training and test sets for the logistic regression model and created a confusion matrix to measure the accuracy of the classification model on the test dataset.

Impact of Energy Production on Air Quality in the United States: 1990-2018

Results: All States

Linear regression results indicate production of Coal and Petroleum have a statistical relationship to a higher rate of ‘unhealthy’ days while production of ‘Green’ energy and Natural Gas have a statistical relationship to a lower rate of ‘unhealthy’ days.

These results are a validation of the relationships that are evident in the scatterplots for these 4 energy production types and the percent of unhealthy days.



```
Call:
lm(formula = Percent_Unhealthy_Days ~ green_percent + Natural_Gas_Percent +
    Coal_Percent + Petroleum_Percent, data = data1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.102968 -0.031440 -0.009446  0.019426  0.211935

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.044058   0.004897   8.997  < 2e-16 ***
green_percent       -0.042336   0.007193  -5.886 4.92e-09 ***
Natural_Gas_Percent -0.017242   0.007317  -2.357   0.0186 *
Coal_Percent         0.015622   0.006190   2.524   0.0117 *
Petroleum_Percent    0.139565   0.010773  12.956  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04242 on 1445 degrees of freedom
Multiple R-squared:  0.1891,    Adjusted R-squared:  0.1869
F-statistic: 84.26 on 4 and 1445 DF,  p-value: < 2.2e-16
```

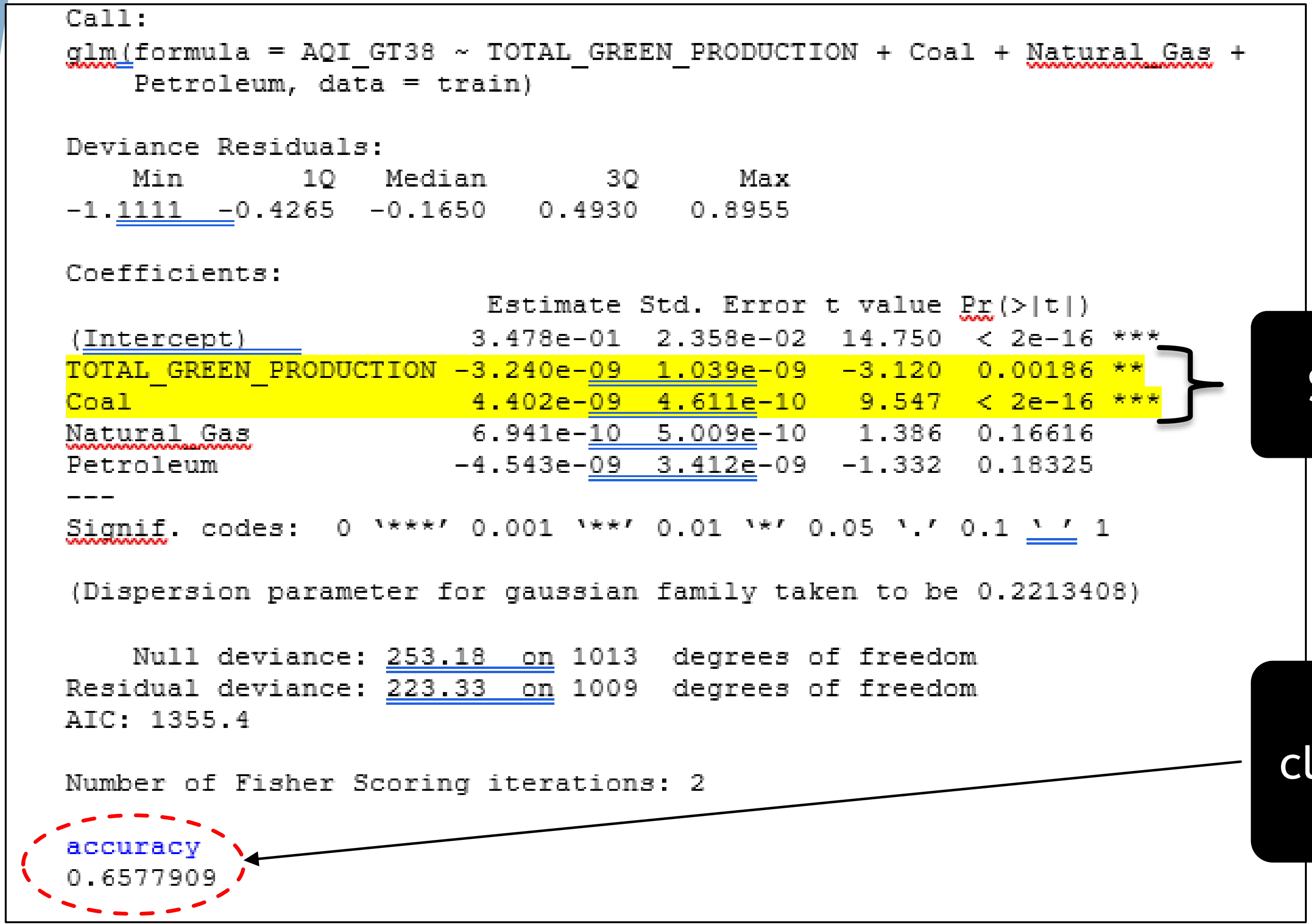
Significant at .05 or lower

The model explains ~20% of the variation in the ‘unhealthy days’ variable

Impact of Energy Production on Air Quality in the United States: 1990-2018

Results: All States

Binomial Logistic regression shows that Coal production has a significant relationship to higher median AQI values while production of Green energy sources has a significant relationship to lower median AQI values.



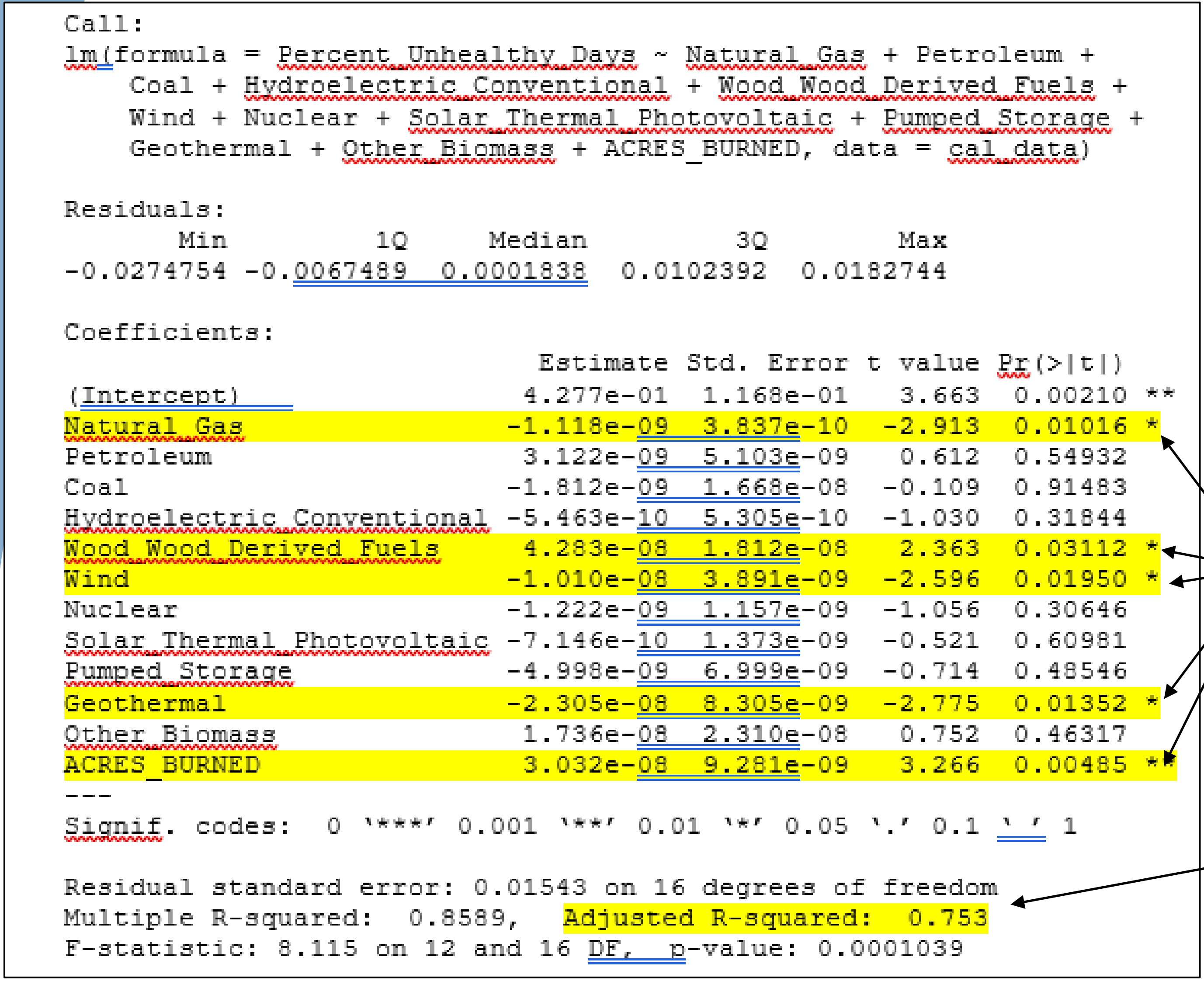
```
Call:
glm(formula = AQI_GT38 ~ TOTAL_GREEN_PRODUCTION + Coal + Natural_Gas +
    Petroleum, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1111  -0.4265  -0.1650   0.4930   0.8955

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.478e-01  2.358e-02  14.750  < 2e-16 ***
TOTAL_GREEN_PRODUCTION -3.240e-09  1.039e-09  -3.120  0.00186 **
Coal                    4.402e-09  4.611e-10   9.547  < 2e-16 ***
Natural_Gas             6.941e-10  5.009e-10   1.386  0.16616
Petroleum              -4.543e-09  3.412e-09  -1.332  0.18325
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2213408)

    Null deviance: 253.18  on 1013  degrees of freedom
Residual deviance: 223.33  on 1009  degrees of freedom
AIC: 1355.4

Number of Fisher Scoring iterations: 2

accuracy
0.6577909
```

Significant at .05 or lower

The model correctly classified 66% of cases in the test dataset

Impact of Energy Production on Air Quality in the United States: 1990-2018

Results: California



```
Call:
lm(formula = Percent_Unhealthy_Days ~ Natural_Gas + Petroleum + 
    Coal + Hydroelectric_Conventional + Wood_Wood_Derived_Fuels + 
    Wind + Nuclear + Solar_Thermal_Photovoltaic + Pumped_Storage + 
    Geothermal + Other_Biomass + ACRES_BURNED, data = cal_data)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0274754 -0.0067489  0.0001838  0.0102392  0.0182744 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)   
(Intercept)                 4.277e-01  1.168e-01   3.663  0.00210 **
Natural_Gas                -1.118e-09  3.837e-10  -2.913  0.01016 * 
Petroleum                   3.122e-09  5.103e-09   0.612  0.54932   
Coal                       -1.812e-09  1.668e-08  -0.109  0.91483   
Hydroelectric_Conventional -5.463e-10  5.305e-10  -1.030  0.31844   
Wood_Wood_Derived_Fuels     4.283e-08  1.812e-08   2.363  0.03112 * 
Wind                       -1.010e-08  3.891e-09  -2.596  0.01950 * 
Nuclear                    -1.222e-09  1.157e-09  -1.056  0.30646   
Solar_Thermal_Photovoltaic -7.146e-10  1.373e-09  -0.521  0.60981   
Pumped_Storage             -4.998e-09  6.999e-09  -0.714  0.48546   
Geothermal                 -2.305e-08  8.305e-09  -2.775  0.01352 * 
Other_Biomass               1.736e-08  2.310e-08   0.752  0.46317   
ACRES_BURNED                3.032e-08  9.281e-09   3.266  0.00485 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01543 on 16 degrees of freedom
Multiple R-squared:  0.8589,	Adjusted R-squared:  0.753 
F-statistic: 8.115 on 12 and 16 DF,  p-value: 0.0001039
```

Linear regression shows that production of Wood Derived Fuels is related to a higher rate of unhealthy days while Natural Gas, Wind, and Geothermal production are related to a lower rate.

The model also indicates that wildfire pollution has a significant statistical relationship to higher pollution levels.

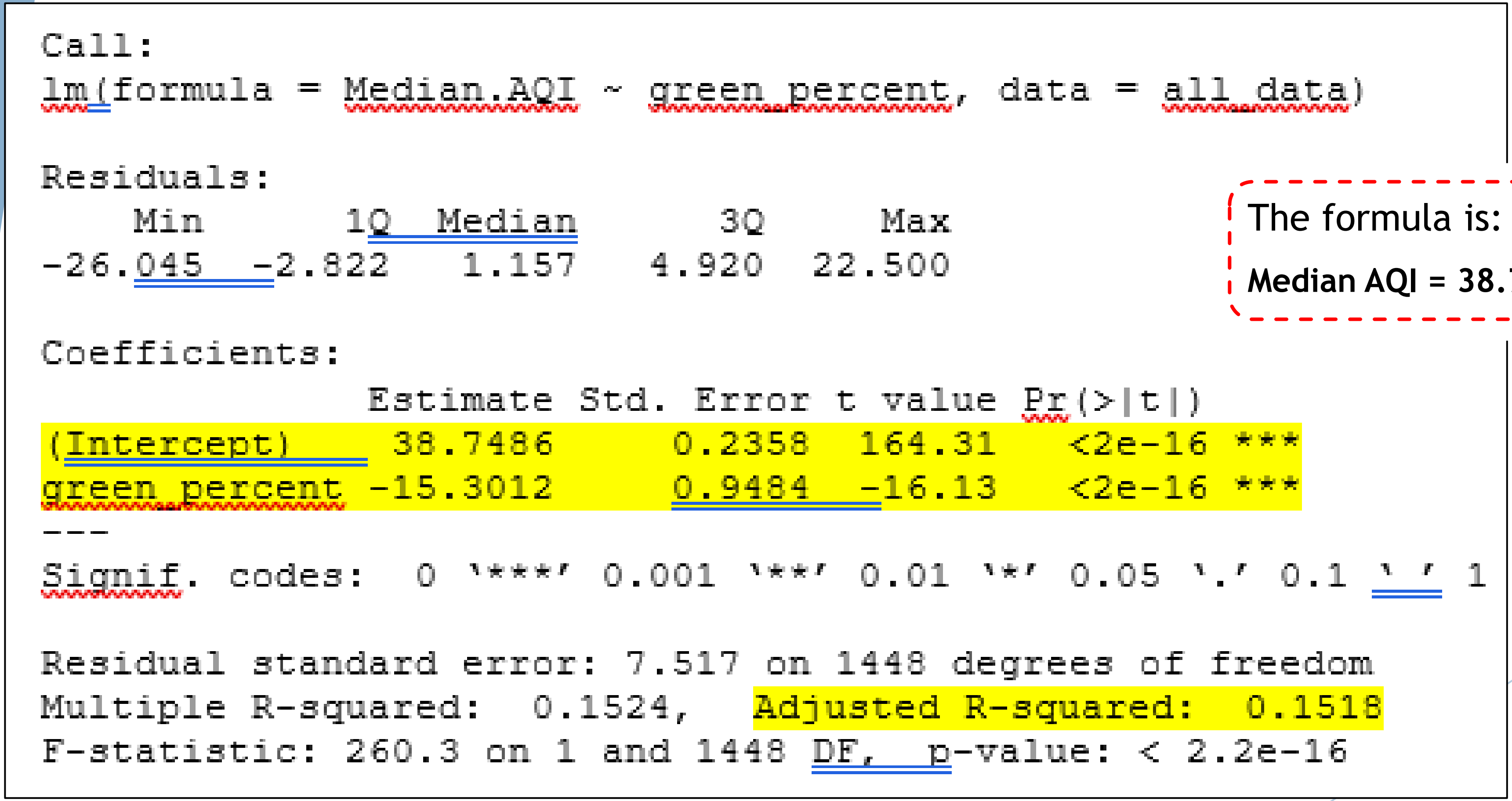
Significant at .05 or lower

The model explains 75% of the variation in the ‘unhealthy days’ variable

Impact of Energy Production on Air Quality in the United States: 1990-2018

Results: All States

A final linear model across all states quantifies the relationship between Green energy production rates and median AQI values. The model indicates that a 1% increase in overall green energy rates (and thus a 1% decrease in conventional rates) is related to a -.153 change in median AQI.



```
Call:
lm(formula = Median.AQI ~ green_percent, data = all_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.045  -2.822   1.157   4.920  22.500 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    38.7486     0.2358  164.31   <2e-16 ***
green_percent -15.3012     0.9484  -16.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.517 on 1448 degrees of freedom
Multiple R-squared:  0.1524,	Adjusted R-squared:  0.1518 
F-statistic: 260.3 on 1 and 1448 DF,  p-value: < 2.2e-16
```

The formula is:

Median AQI = 38.75 + (-.153)*(green energy production %)

Impact of Energy Production on Air Quality in the United States: 1990-2018

Green New Deal

- In 2018, 16% of all energy produced in the United States was green energy
- The median AQI for 2018 was 36.0
- The Green New Deal requires 100% of energy to be produced by green energy sources
- Based on the linear regression equation, transitioning the remaining 84% of energy to green energy would equate to a 12.9 point decline in median AQI, for a median AQI of 23.1
- 2018 estimates suggest is would cost approximately $5.1 trillion to transition to 100% green energy
- This equates to approximately $400 billion for a 6.5% increase in green energy and a 1-point drop in median AQI

Impact of Energy Production on Air Quality in the United States: 1990-2018

Conclusions

- The data indicates a significant relationship between different energy production types and air quality
- Production of coal and petroleum have a significant relationship to higher levels of monitored air pollutants – production rates for these 2 energy sources are in decline
- Production of natural gas and green energy sources have a significant relationship to lower levels of pollutants – production for these energy sources are increasing
- In California, wildfires are related to higher levels of pollutants
- Increasing the proportion of green energy among all types of energy produced would impact median Air Quality Index levels at a rate of minus 1 point for every 6.5% increase in green energy
- The estimated cost for this 1-point decline is $400 billion