# Traffic Volume and Air Quality

Hannah Nguyen

2025-02-27

## Introduction

Air quality is a critical factor influencing public health and environmental sustainability, particularly in regions with significant traffic congestion.[1] This project investigates the relationship between air quality and traffic volume across counties in Pennsylvania from 2018 to 2023 using panel regression among other techniques. Panel regression helps control for both temporal and spatial variations,[2] allowing for more precise estimations of traffic's effects on air quality. The findings of this study can inform policymakers and urban planners in designing targeted interventions to mitigate air pollution, improve public health, and enhance transportation policies in Pennsylvania.

## Data Available

### a. Daily Air Quality Index by County

Air quality in the US is measured by the Air Quality Index (AQI), defined by the US Environmental Protection Agency (EPA).[3] The EPA determines an AQI for a geographic location through the ratio of different pollutant factors found in air monitors in the area. The higher the AQI, the more hazardous such air is to public health.[4]

Data for daily AQI per county in the US is collected through the EPA's pre-generated data website.[5] The dataset also contains corresponding quality categorization ("Good", "Moderate", "Unhealthy",... quality of air) and the specific pollutant component that informs the index calculation. See section 1 in the Appendix for more detail.

### b. Traffic Volume Data

The Pennsylvania Department of Transportation (PennDOT) collects and publishes open data on traffic volume, defined as "amounts of vehicle traffic that travel the section of road,"[6] for specific roadways in the state. The data does not include every street nor every county due to availability of traffic counting methods.

The variables include estimate counts of vehicle and truck traffic in specific date (Current AADT - Annual Average Daily Traffic and Current AADTT - Annual Average Daily Truck Traffic) as well as base counts or historical counts for vehicle and truck traffics (Base AADT and Base AADTT). That is, if the current traffic on a street is 2,000 while its base traffic is 1,500, that street is 500 vehicles more crowded on that date than its usual traffic rate. The data also includes categories of Traffic Pattern Groups to specify types of roads for the traffic count (urban interstates, rural local roads, etc.). More information on the metadata of this dataset is available on PennDOT's GIS site.[7] While the data collected goes back to 1990, the final dataset will only include data from 2018 to 2023. An example of how traffic volume was recorded for York county (the full PA road network is too dense for this visualization) was included in section 2 of the Appendix.

[1] National Institute of Environmental Health Sciences, "Air Pollution and Your Health," National Institute of Environmental Health Sciences, February 26, 2025, https://www.niehs.nih.gov/health/topics/agents/air-pollution.

[2] Oscar Torres-Reyna, "Getting Started in Fixed/Random Effects Models using R/RStudio," Princeton University, 2010, accessed Feb. 21, 2025, https://www.princeton.edu/~otorres/Panel101R.pdf, 2.

[3] US EPA, "AQI Basics," AirNow.gov, accessed February 7, 2025, https://www.airnow.gov/aqi/aqi-basics.

[4] Ibid.

[5] US EPA, "Pre-Generated Data Files," Data & Tools, Air Data, November 19, 2024, https://aqs.epa.gov/aqsweb/airdata/download_files.html.

[6] PennDOT, "RMSTRAFFIC (Traffic Volumes)," PennShare, January 21, 2025, https://data-pennshare.opendata.arcgis.com/datasets/a17c20bf71dd40fea24363bb9f0ae0e4_0/about.

[7] PennDOT, "Traffic Volumes," Data & Tools, PennDOT GIS Data Dictionary Hub, accessed February 7, 2025, https://docs-pennshare.hub.arcgis.com/pages/traffic-volumes.

### c. Daily average temperature and precipitation

This analysis takes into account daily average temperature and precipitation to observe occurrences that might have an impact on traffic patterns and overall air quality. Daily climate summaries from 4/1/2018 to 12/31/2023 for all 133 counties in Pennsylvania were collected from National Oceanic and Atmospheric Association.[8] See section 3 in the Appendix for summary statistics.

### d. Final dataset

From 81,000 records for AQI indexes and 43,000 records for traffic volumes, merging by both county code and date results in a dataframe of 7831 records. From base volumes, delta change values were calculated for traffic and truck volumes. That is, "Volume Change" or "Delta.Traffic" is the difference between "Base Traffic" and "Current Traffic."

Because traffic volume data was measured at the street level but on different dates, there were duplicated values when grouping data by county and date. Specifically, there may be multiple streets in a specific county, of which traffic counts were recorded in different date. When merged with weather and air quality data by county codes, there were many rows with different volume counts for the same county on the same date. This is a big problem because redundancy skews the distributions and misrepresents variables of interest.[9] Therefore, the merged dataset was filtered by county code and date, and duplication was handled as follow:

- **Traffic volume variables:** Taking the sum of all vehicle counts in a county on a specific date.
- **AQI, average temperature, and average precipitation:** Taking the mean of these indicators for a county on a specific date.
- **Air Quality Category and Defining Parameter:** Taking the unique value across a county on a specific date. Because these categorical values are tied to county-specific information pre-merge, they should only have 1 unique value.
- **Traffic patterns:** Not including this variable because it is tied to street-specific information and cannot be tallied correctly across a county.

The most problematic variable to handle was traffic pattern due to its inability to be represented fairly across a county. There is no good way to tally the type of roads being counted for traffic for a county on a specific date and compare such information to that of another county. Therefore, these patterns will not be a factor in the regression, though included in visualizations. The filtered dataset includes 1,460 rows and 13 features (see Table 1 for summary statistics of numerical values in the dataset).

Table 1: Descriptive Statistics of Dataset

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Air Quality Index | 1,460 | 47.6 | 22.0 | 3 | 271 |
| Daily Average Precipitation | 1,460 | 2.6 | 6.6 | 0.0 | 107.2 |
| Daily Average Temperature | 1,460 | 16.3 | 7.1 | −7.6 | 28.8 |
| Truck Volume | 1,460 | 1,454.4 | 2,849.8 | 0 | 57,107 |
| Change in Truck Volume | 1,460 | 107.8 | 269.2 | −171 | 4,796 |
| Traffic Volume | 1,460 | 18,502.8 | 29,267.4 | 4 | 447,058 |
| Change in Traffic Volume | 1,460 | 1,493.4 | 3,082.3 | 0 | 38,428 |

[8] NOAA, "Index of /Data/Nclimgrid-Daily/Archive," accessed February 7, 2025, https://www.ncei.noaa.gov/data/nclimgrid-daily/archive/.

[9] Dataddo, "Data Duplication: Understanding and Resolving Common Issues," accessed February 20, 2025, https://docs.dataddo.com/docs/data-duplication.

## Preliminary Analysis

The dataset is quite unbalanced, based on initial observation. Figures 1 provides snapshots of traffic information in the dataset, which is heavily biased towards local roads and main urban streets. This is simply due to data collection complexities as estimating traffic counts is resource-intensive when the street is not equipped with vision technology; thus, traffic counting is often done by private companies as opposed to public, government organization.[10] We also see a long tail to the right of the histogram, even with outliers removed. We can expect the median and mean to be much great than the mode, or that our data contains outliers with very high traffic counts.

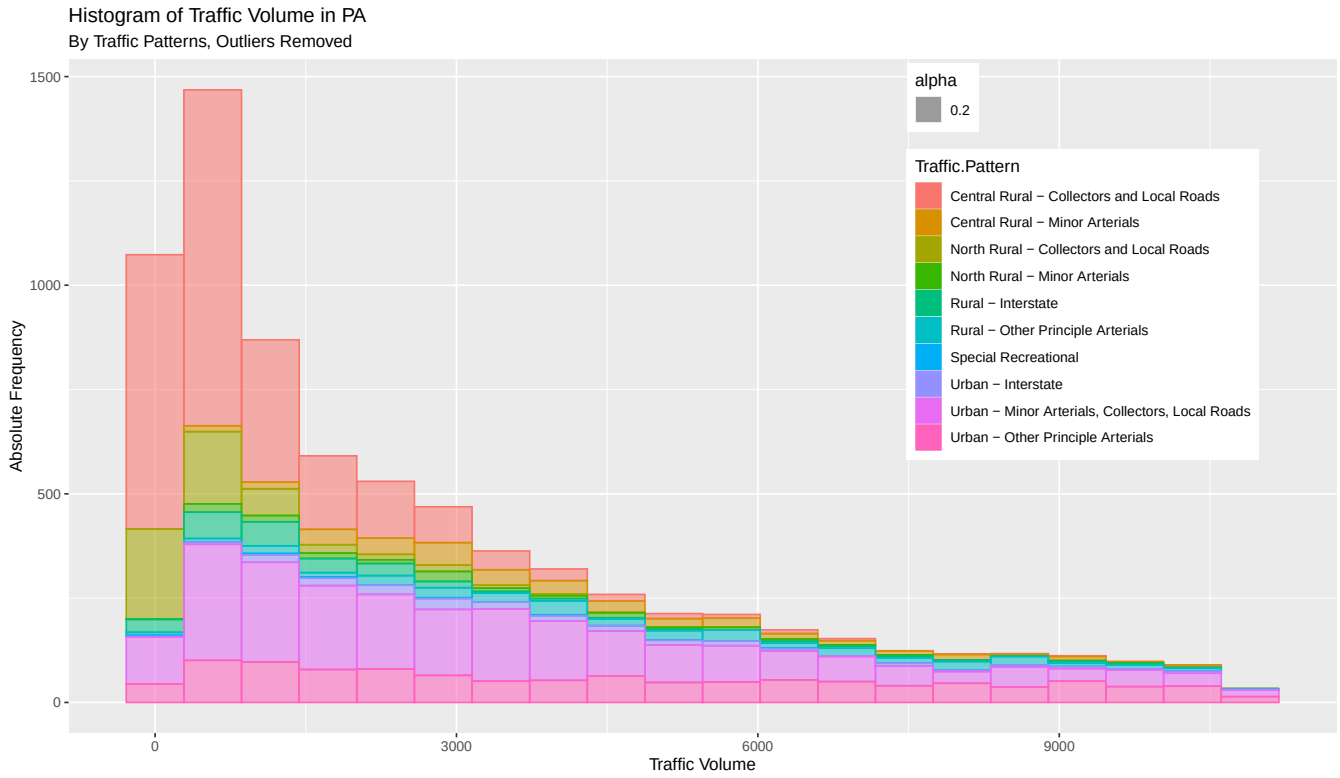
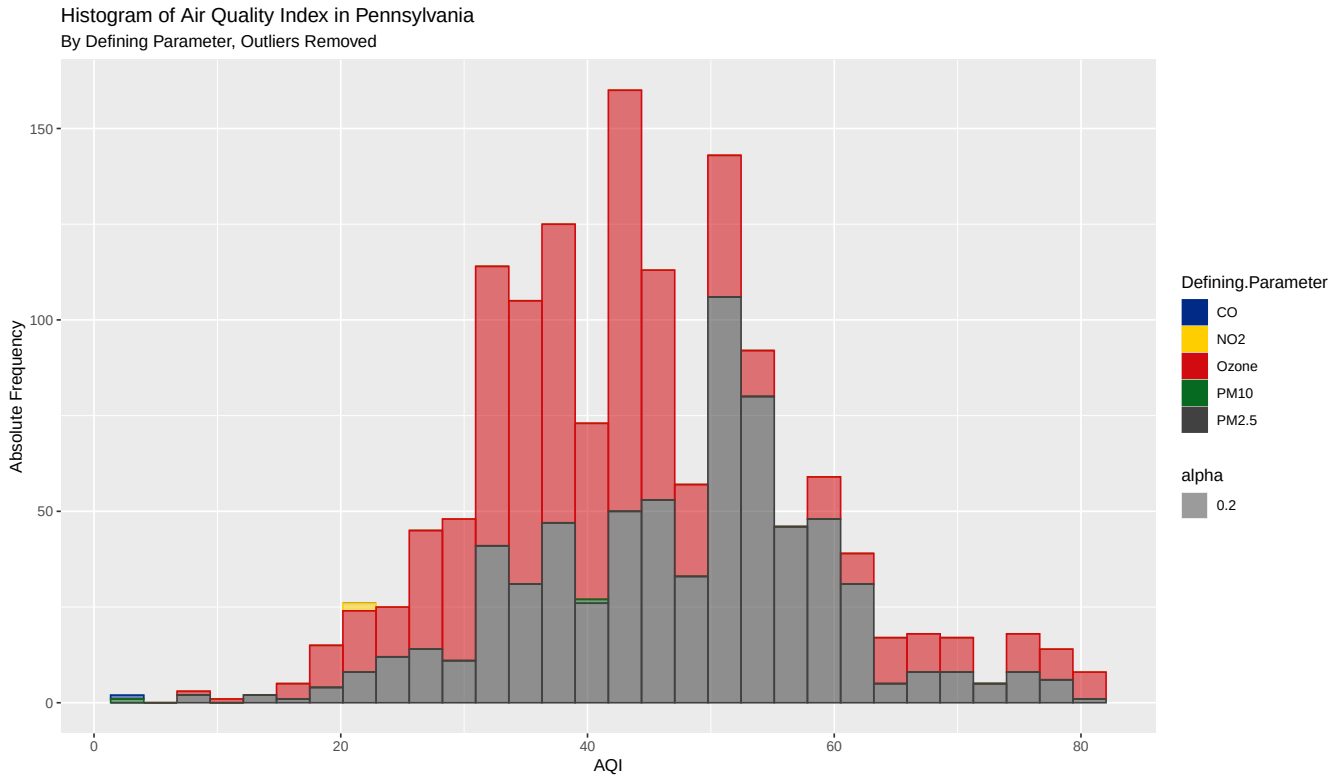



Figure 1: Distribution of Traffic Volumes by Traffic Patterns

We also observe imbalances in air quality attributes. Figure 2 demonstrates that the data contains mostly good and moderate air quality with defining parameters being almost exclusively "Ozone" and "PM2.5".Ozone is the main ingredient in "smog" while PM2.5 are micro-particles that can get into one's bloodstream.[11] When Ozone is a defining parameter for "Good" air quality, it might mean that the level of "smog" is low enough for the area to have healthy air. With PM2.5, however, we see that it more often defines less "Good" air quality. Figure 2 also seems relatively bell-shaped without outliers, centering around 40-50 in air quality.

[10]Sarah Penny, "6 Traffic Counts and Classification Study Methods," *SMATS* (blog), July 21, 2021, https://www.smatstraffic.com/2021/07/21/counts-and-classification-study-methods/; VentureRadar, "Top Traffic Counting Companies," accessed February 20, 2025, https://www.ventureradar.com/keyword/Traffic%20Counting.

[11]OAR US EPA, "Health Effects of Ozone and Particulate Matter," Other Policies and Guidance, June 21, 2022, https://www.epa.gov/advance/health-effects-ozone-and-particulate-matter.




Finally, there are imbalances in the counties represented (Figure 3), which can be due to differences in the number of streets with traffic counts recorded between 2018 and 2023. Counties with more representation such as Allegheny, Cambria, and Dauphin are reasonably justified since they have more streets operating in and pouring *into* big cities (Pittsburgh, Harrisburg, etc.). Bradford being the most represented county is quite surprising as it is a small, rural county.[12]

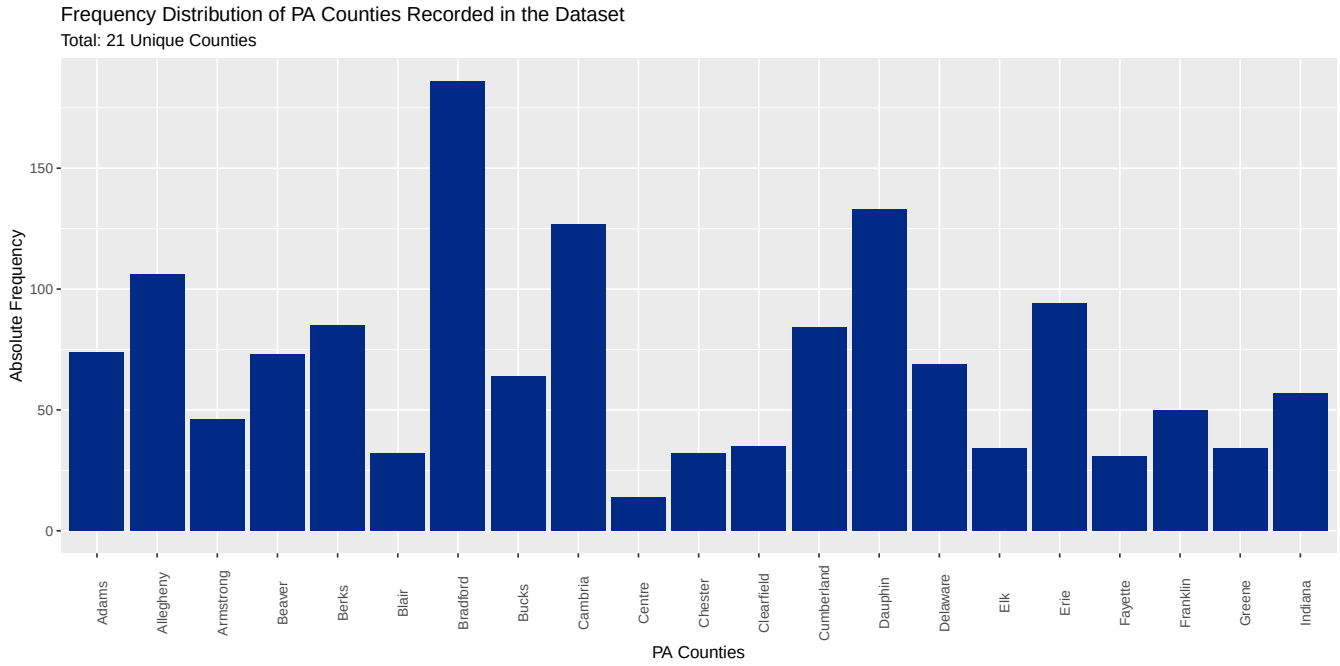



Figure 2: Absolute Frequency of PA Counties

[12]Bradford County, "Visit Bradford County," accessed February 20, 2025, https://bradfordcountypa.org/visitors/.

Before moving to statistical analysis, plotting a correlation heat map of all numerical variables (Figure 4) is helpful. We see a strong relationship between traffic count variables and little relationship elsewhere. It is intuitive to see the traffic count correlations because high traffic volume might also mean a stark deviation from historical counts. Moreover, it is reasonable to expect a high truck volume when there is a high traffic volume for certain streets.

Initial visualizations have not only reveal imbalances in the data but also imply innate differences between counties. Accounting for these differences will allow us to make causal inferences, building a stronger case to justify the relationship, or lack thereof, between variables compared to a regular regression model.

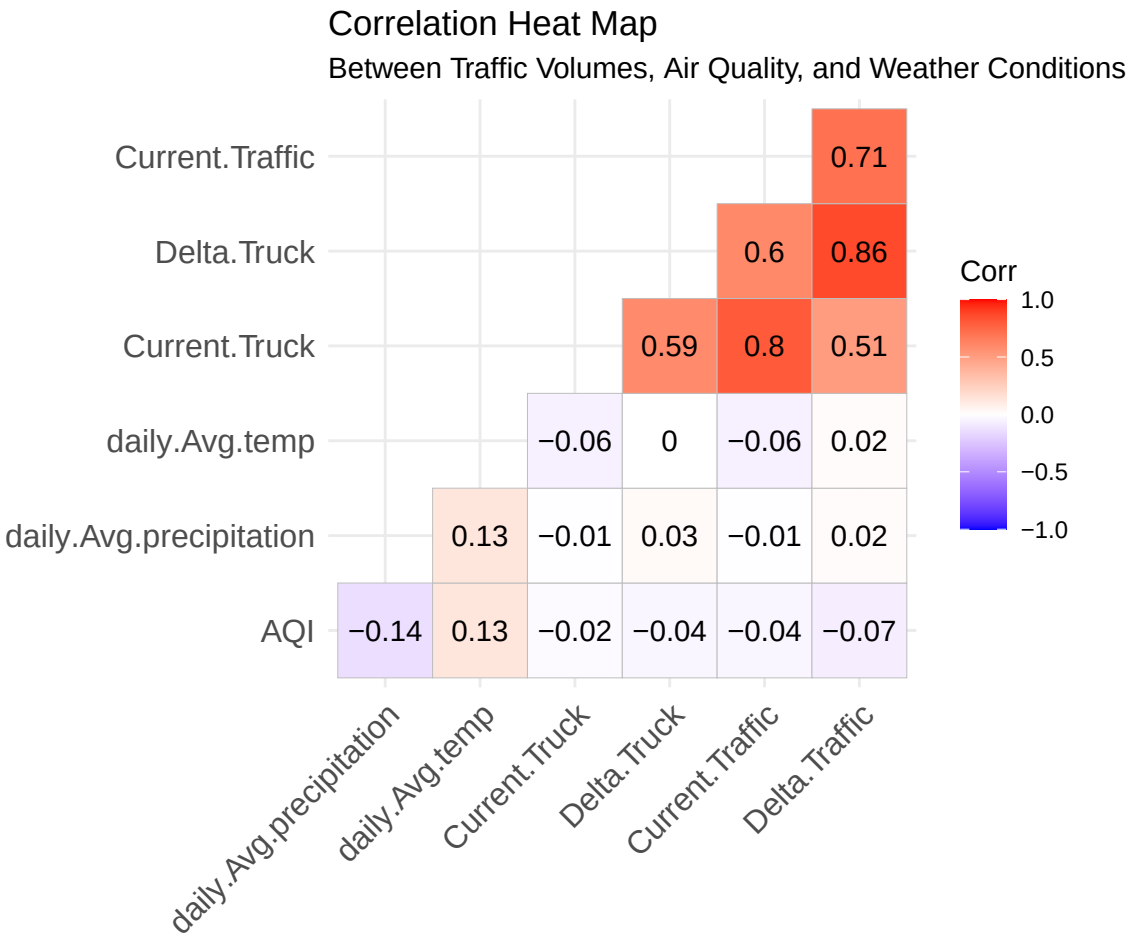



Figure 3: Correlation Heat Map Across Numerical Variables

## Statistical Test

### a. Ordinary Least Square

After conducting variable selection (see section 8 in the Appendix), the recommended regression involves: average daily precipitation, average daily temperature, traffic volume, change in traffic volume, air quality categories, defining parameters, and *county names*. In other words, the "best" ols model suggests accounting for differences between counties.

For this "best" ols model (see Table 2), we see some significant positive relationship between AQI and traffic volume, indicating that more traffic is related to higher AQI (more unhealthy air). However, the negative significant coefficent for *change* in traffic volume seems to contradict the traffic volume - AQI relationship above it: More traffic *than usual* correlates to a decrease in AQI (healthier air). We also established the strong, positive correlation between traffic volume and change in traffic volume in the last section, so this ols result seems unjustifiable.

Table 2:

| | *Dependent variable:* |
|---|---|
| | AQI |
| daily.Avg.precipitation | $-0.218^{***}$ (0.037) |
| daily.Avg.temp | $0.172^{***}$ (0.035) |
| Current.Traffic | $0.00002^{**}$ (0.00001) |
| Delta.Traffic | $-0.0003^{**}$ (0.0001) |
| factor(Category)Moderate | $22.180^{***}$ (0.565) |
| factor(Category)Unhealthy | $143.544^{***}$ (4.083) |
| factor(Category)Unhealthy for Sensitive Groups | $73.041^{***}$ (1.937) |
| factor(Category)Very Unhealthy | $195.569^{***}$ (3.735) |
| factor(Defining.Parameter)NO2 | $22.291^{**}$ (11.080) |
| factor(Defining.Parameter)Ozone | $36.731^{***}$ (9.032) |
| factor(Defining.Parameter)PM10 | $18.415^{*}$ (11.054) |
| factor(Defining.Parameter)PM2.5 | $34.861^{***}$ (9.028) |
| factor(county.Name)Allegheny | $2.596^{*}$ (1.383) |
| factor(county.Name)Armstrong | $-2.295$ (1.704) |
| factor(county.Name)Beaver | $-0.126$ (1.491) |
| factor(county.Name)Berks | $-0.703$ (1.442) |
| factor(county.Name)Blair | $-2.589$ (1.909) |
| factor(county.Name)Bradford | $-3.831^{***}$ (1.256) |
| factor(county.Name)Bucks | $-3.912^{**}$ (1.559) |
| factor(county.Name)Cambria | $-1.499$ (1.337) |
| factor(county.Name)Centre | $-0.257$ (2.640) |
| factor(county.Name)Chester | 1.371 (1.914) |
| factor(county.Name)Clearfield | $-5.851^{***}$ (1.864) |
| factor(county.Name)Cumberland | $-4.777^{***}$ (1.479) |
| factor(county.Name)Dauphin | $-0.393$ (1.318) |
| factor(county.Name)Delaware | $-0.776$ (1.522) |
| factor(county.Name)Elk | $-0.709$ (1.886) |
| factor(county.Name)Erie | $-3.085^{**}$ (1.410) |
| factor(county.Name)Fayette | $-1.791$ (1.925) |
| factor(county.Name)Franklin | $-5.031^{***}$ (1.672) |
| factor(county.Name)Greene | 1.743 (1.866) |
| factor(county.Name)Indiana | $-1.864$ (1.589) |
| Constant | 1.642 (9.105) |
| Observations | 1,460 |
| $R^2$ | 0.837 |
| Adjusted $R^2$ | 0.833 |
| Residual Std. Error | 8.983 (df = 1427) |
| F Statistic | $228.197^{***}$ (df = 32; 1427) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Figure 5 plots the relationship between traffic volume change and AQI. We see that most values are concentrated in the bottom left corner where AQI and volume change are low, showing that Pennsylvania counties have good air quality and relatively slow-changing traffic volume. Moreover, there are no data in the top right corner (more traffic than usual and high AQI index), which does not corroborate common understanding of car exhaust's impact to air

quality.[13] There are also strong outliers that might have affected the regression model. These confusing conclusions are warning signs that an ols might not be the best option for this dataset.

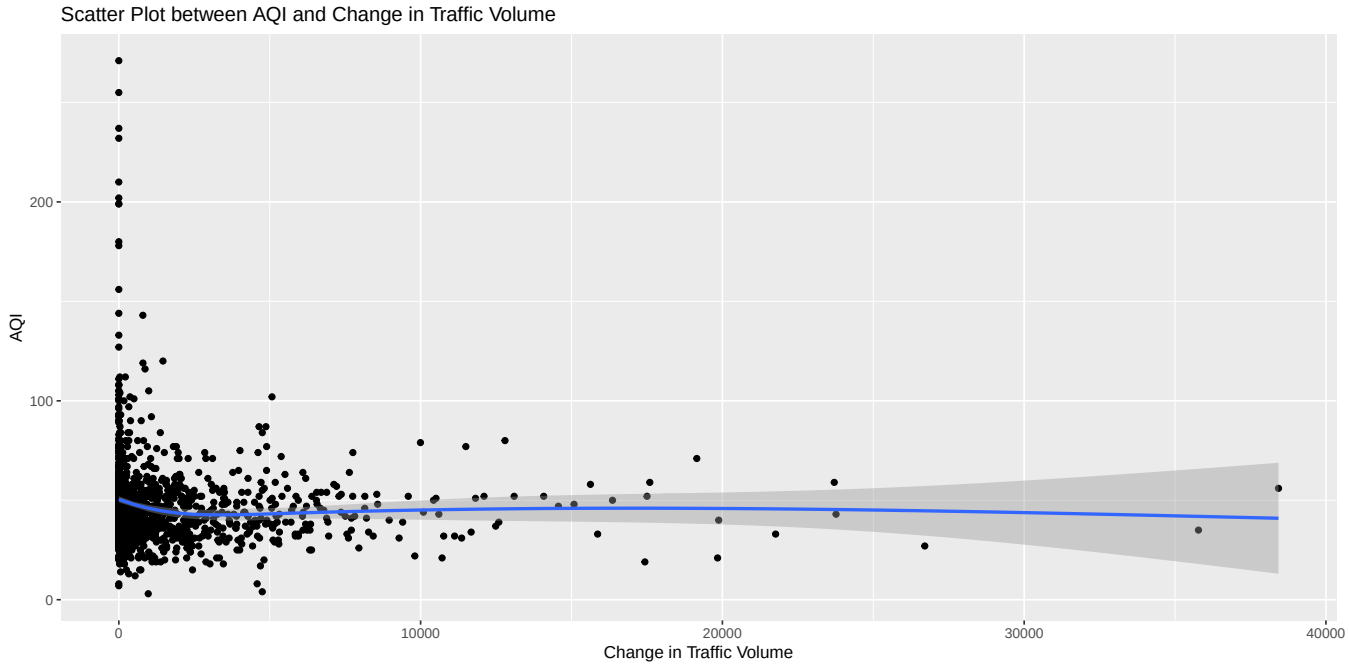



Figure 4: Relationship between Change in Traffic Volume and Air Quality

### b. Panel Regression Models

We move on to panel regression models. Due to the nature of panel regression (controlling for innate differences between counties and date), we only need to pass our variables of interest (traffic volume) to the model and not concern about colinearity. I built a fixed-effects model and a random-effects model with traffic volume, truck volume, change in traffic volume, and change in truck volume as independent variables to AQI. I then employed different statistical tests to choose the best model among my fixed-effects, random-effects, and ols models. See section 8 in the Appendix for more details on my comparison tests.

The best model for this data is a random effects model that concludes no significant relationship between any traffic variables to AQI (see Table 3).

Table 3:

| | *Dependent variable:* |
|---|---|
| | AQI |
| Current.Traffic | −0.00000 (0.00003) |
| Current.Truck | −0.00004 (0.0003) |
| Delta.Traffic | −0.001 (0.0004) |
| Delta.Truck | 0.004 (0.004) |
| Constant | 47.587*** (0.879) |
| Observations | 1,460 |
| $R^2$ | 0.094 |
| Adjusted $R^2$ | 0.092 |
| F Statistic | 4.880 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

[13]David L Buckeridge et al., "Effect of Motor Vehicle Emissions on Respiratory Health in an Urban Area.," *Environmental Health Perspectives* 110, no. 3 (March 2002): 293–300.

I performed the Breusch-Pagan Lagrange multiplier test and found a significant p-value, rejecting the null hypothesis that there exists no significant differences across units. We accept the alternative hypothesis that the data has panel effects.[14]

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  AQI ~ Current.Traffic + Delta.Traffic + Current.Truck + Delta.Truck
## chisq = 403.68, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Through this section, we arrive at the following key conclusions:

- The data has panel effects that are more appropriately modeled through panel regression, specifically a random-effects model.
- The results of the random-effects model shows no relationship between AQI and traffic/truck volume, controlling for differences between counties in PA over time.
- This is not to say that exhaust fumes are not linked to air pollution because this analysis is constrained within specific Pennsylvania counties on specific dates in time. It might also show the value and effectiveness of air quality monitoring efforts and car emission controls in Pennsylvania.

**c. Clustering with DBSCAN:**

I will conduct clustering through the DBSCAN algorithm. DBSCAN stands for density-based spatial clustering of applications with noise, which not only identifies outliers in hyper-dimensional planes but also does not assume that the data clusters by spherical shapes.[15] We first plot a k-NN distance plot to determine the best epsilon hyperparameter (the elbow of the plot) for a chosen k nearest neighbors.[16] k is typically chosen to be twice the dimension of the dataset.[17] Since we have 12 variables in the filtered dataset, I chose k = 24.

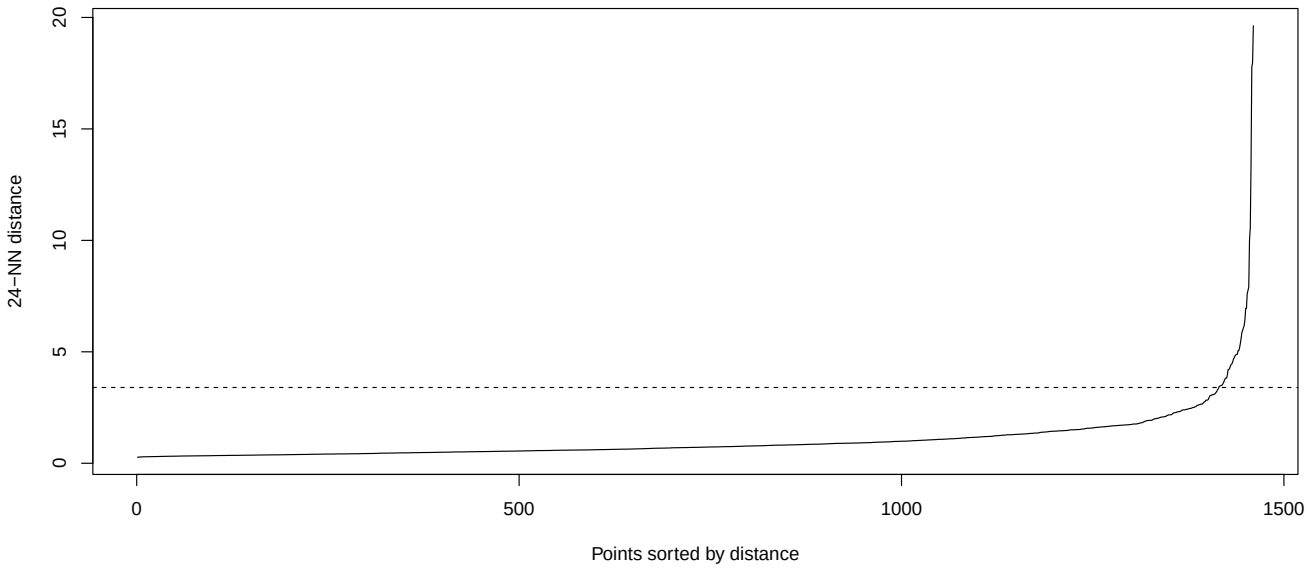



Figure 5: 24-NN Distance Plot

[14] Oscar Torres-Reyna, "Getting Started in Fixed/Random Effects", 19.

[15] Okan Yenigun, "DBSCAN Clustering Algorithm Demystified," Built In, March 11, 2024, https://builtin.com/articles/dbscan.

[16] STHDA, "DBSCAN: Density-Based Clustering for Discovering Clusters in Large Datasets with Noise - Unsupervised Machine Learning - Easy Guides," accessed February 21, 2025, https://www.sthda.com/english/wiki/wiki.php?id_contents=7940.

[17] Tara Mullin, "DBSCAN Parameter Estimation Using Python," *Medium* (blog), July 15, 2020, https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd.

Around a height of 3.4 is where I determined the elbow of the 24-NN distance plot. I passed the scaled, filtered data to a DBSCAN algorithm with eps = 3.4 and MinPts = 24. The model, surprisingly, found only two clusters: 0 and 1 where 0 are the "noises" or outliers in hyperdimensional planes.

Figure 7 plots the data on the AQI - Traffic Volume plane and colors the points by their cluster assignments. We can clearly see the points on the far left and top right are considered "outliers" by DBSCAN. This result seems intuitive and justifiable considering the large distance between these points to the bottom right corner where most points gather.

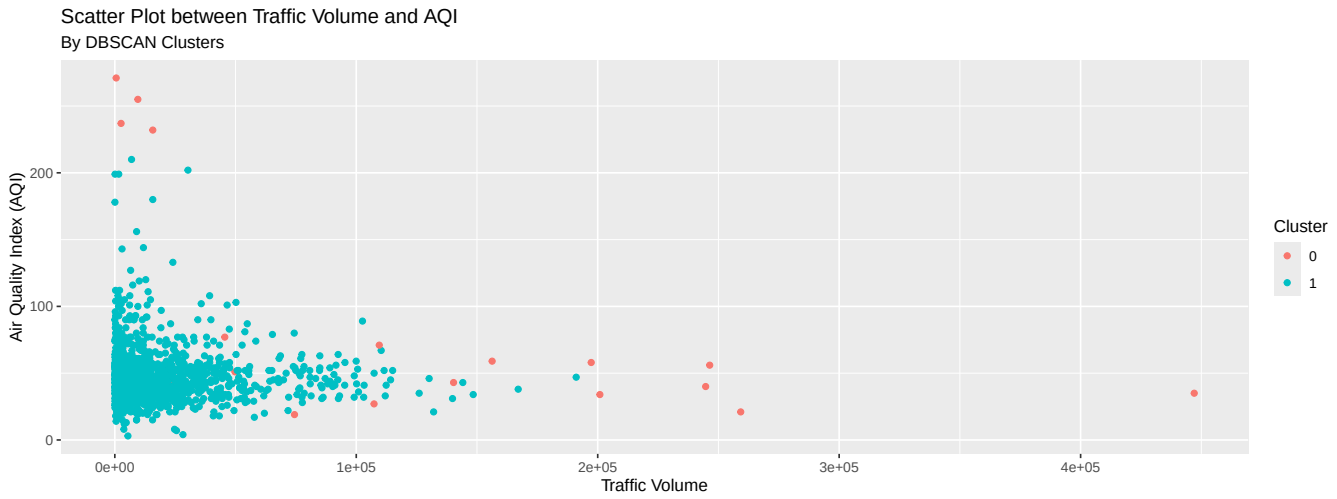



Figure 6: Traffic Volume and AQI by DBSCAN clusters

## Conclusion and Future Work

Through this project, I explored panel regression techniques in R that control for innate differences in place and time, resulting in a more trustworthy regression model for data with panel effects. The overall conclusion is that there is not enough evidence to affirm a relationship between traffic volume and air quality in specific Pennsylvania counties from April 2018 to December 2023. This is not to refuse robust, scientific research that links air pollution to car exhaust fumes.[18] However, it is supportive evidence that Pennsylvania's environmental policies, air quality monitoring efforts, and stringent exhaust controls might have played a role in providing clean air for residents.[19]

This analysis can be improved through better traffic volume data, potentially collected by a private entity with more resources. I recommend county-level data, aggregating traffic counts of all streets in a county. I also believe there are gaps in the data between April 2018 and December 2023. Not all dates were represented, which might have impacted my coefficients.

[18] David L Buckeridge et al., "Effect of Motor Vehicle".

[19] Arthur van Benthem et al., "How Effective Are Vehicle Exhaust Standards?," *Kleinman Center for Energy Policy* (blog), December 7, 2022, https://kleinmanenergy.upenn.edu/research/publications/how-effective-are-vehicle-exhaust-standards/.

## Appendix

**1. Descriptive table for pre-merge AQI dataset:**

Table 4: Summary Statistics of Pennsylvania AQI

| Category | Min AQI | Max AQI | Mean | Standard Deviation |
|---|---|---|---|---|
| Good | 0 | 50 | 35.95 | 8.32 |
| Moderate | 51 | 100 | 60.93 | 10.11 |
| Unhealthy for Sensitive Groups | 101 | 150 | 114.86 | 13.04 |
| Unhealthy | 151 | 200 | 167.05 | 13.58 |
| Very Unhealthy | 201 | 292 | 236.79 | 27.77 |
| Hazardous | 321 | 368 | 339.25 | 20.27 |

**2. Example of pre-merge traffic volume data, mapped onto York County, PA:**

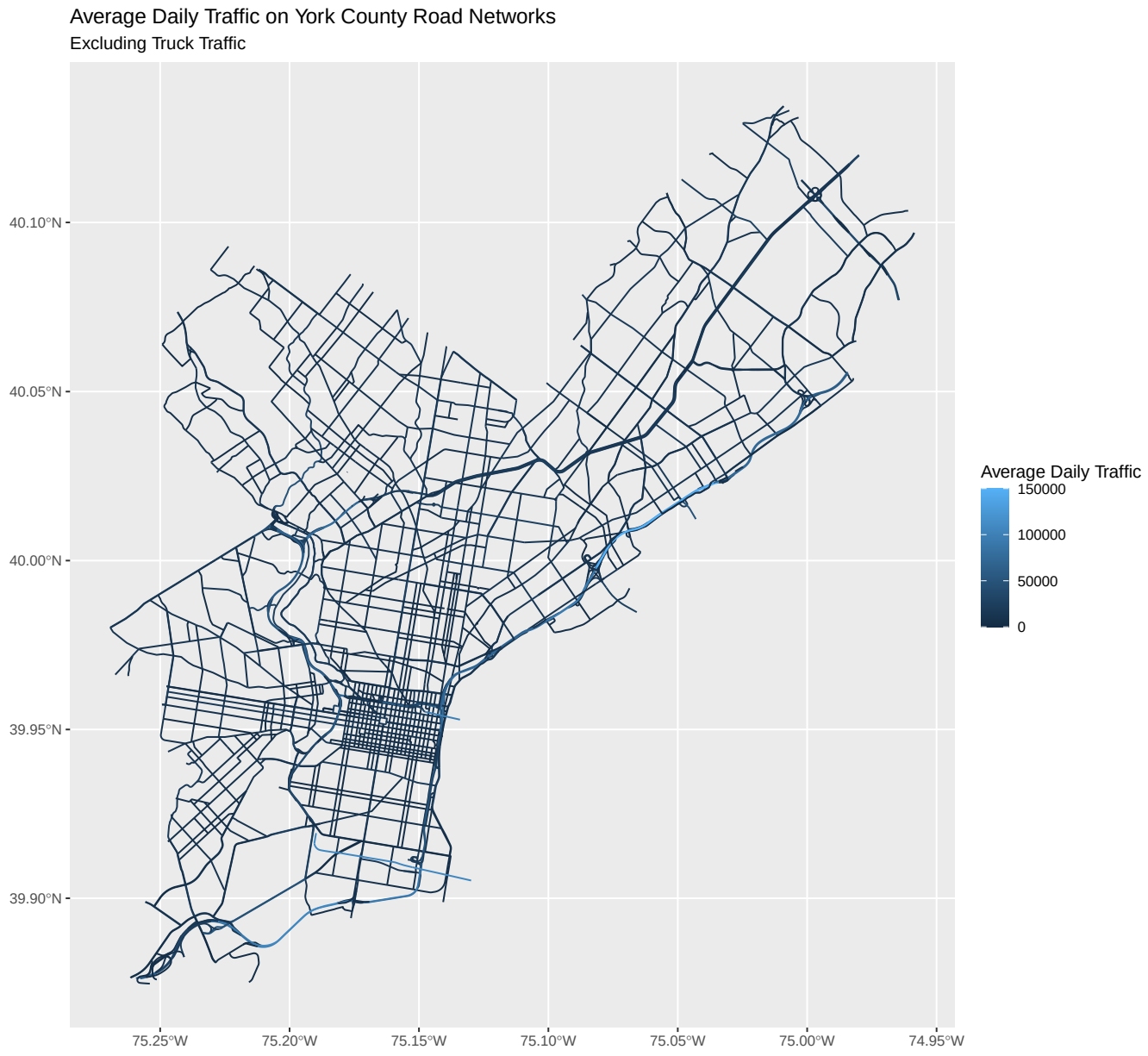



Figure 7: Example of Traffic Volume Data Recorded for York County

**3. Summary statistics for pre-merge temperature and precipitation data:**

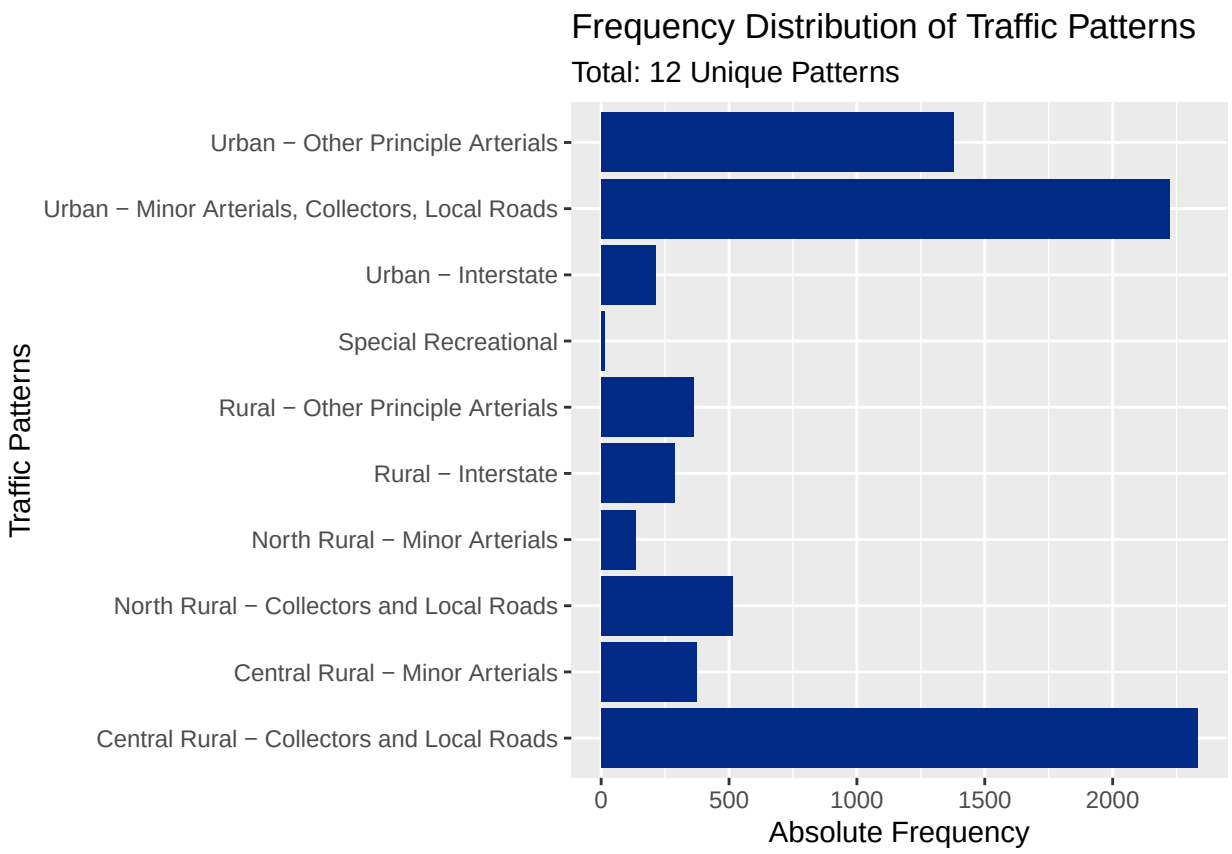
Table 5: Summary Statistics for Daily Average Temperature and Precipitation in Pennsylvania

| Date | County Code | Temperature | Precipitation |
|---|---|---|---|
| Min. :2018-04-01 | Min. : 1 | Min. :-19.57 | Min. : 0.000 |
| 1st Qu.:2019-09-08 | 1st Qu.: 33 | 1st Qu.: 3.14 | 1st Qu.: 0.000 |

| | | | |
|---|---|---|---|
| Median :2021-02-14 | Median : 67 | Median : 11.39 | Median : 0.050 |
| Mean :2021-02-14 | Mean : 67 | Mean : 11.02 | Mean : 3.354 |
| 3rd Qu.:2022-07-24 | 3rd Qu.:101 | 3rd Qu.: 19.49 | 3rd Qu.: 3.170 |
| Max. :2023-12-31 | Max. :133 | Max. : 31.53 | Max. :136.220 |

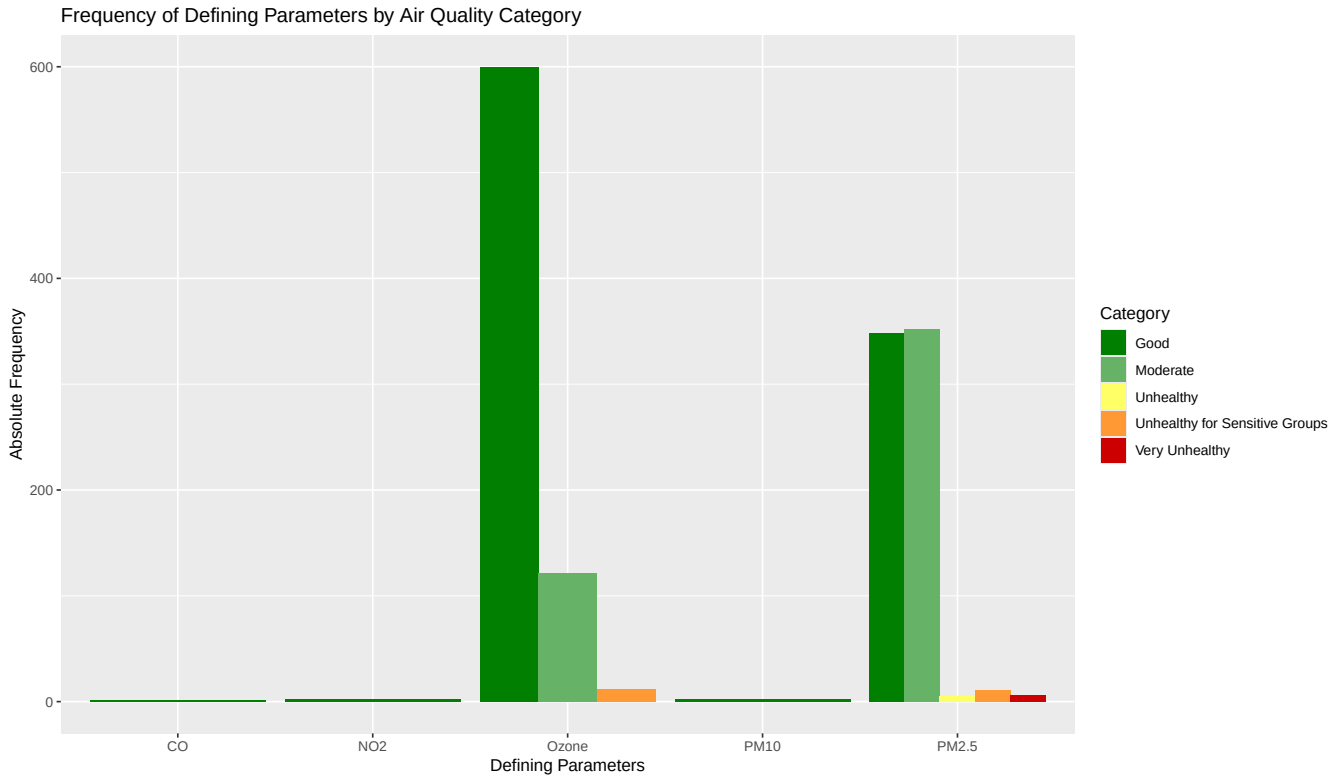
**4. Frequency distribution of Traffic Patterns:**

This section illustrates imbalances in traffic variables.




**5. Frequency of defining parameters by air quality category:**

This section illustrates imbalances in air quality variables.




### 6. Trends in temperature and precipitation over time:

We clearly see seasonality with daily temperature due to natural seasonal changes in Pennsylvania, which might impact the regression analysis in unpredictable ways.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

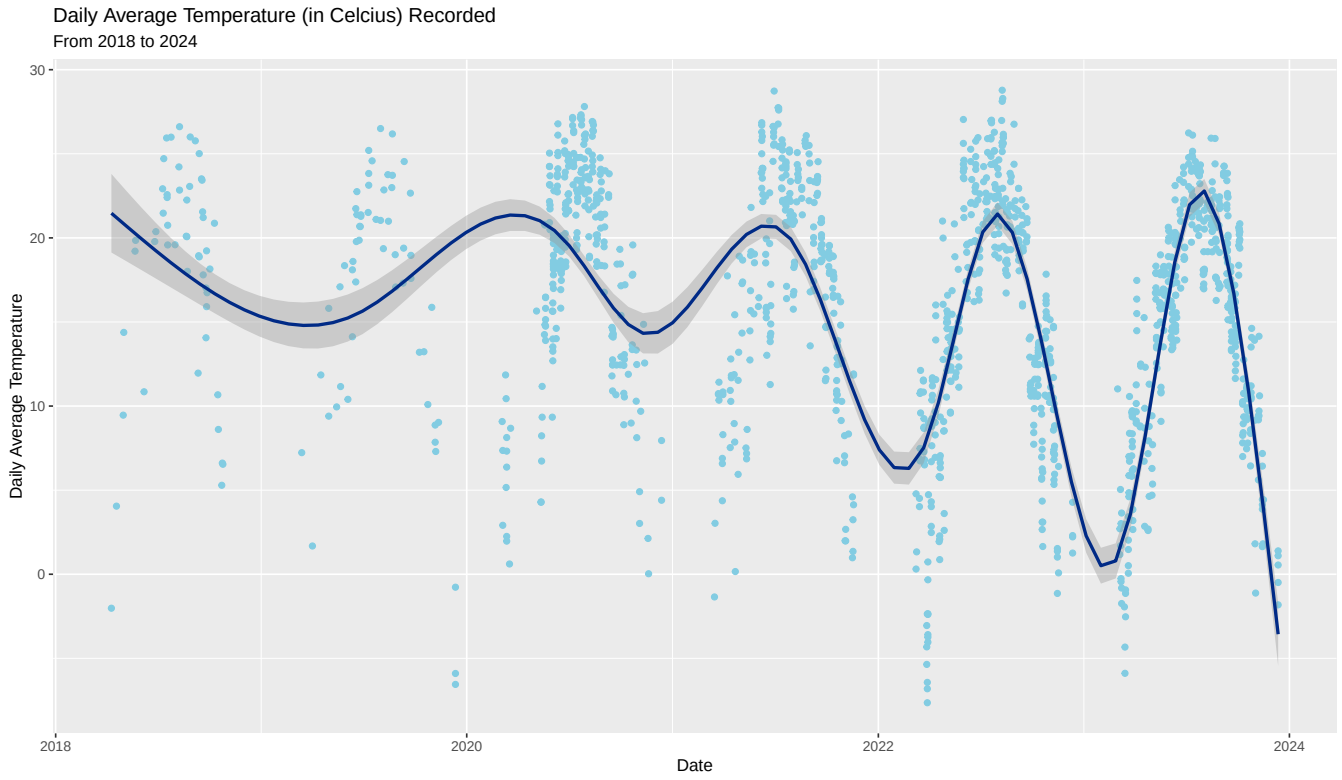



Figure 8: Time Series Visualization of Average Temperature

For precipitation, we do not see a clear linear trend, but there seems to be gaps in the dataset. Pennsylvania seems to be a dry area in general with some scattered rainy days throughout a year.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

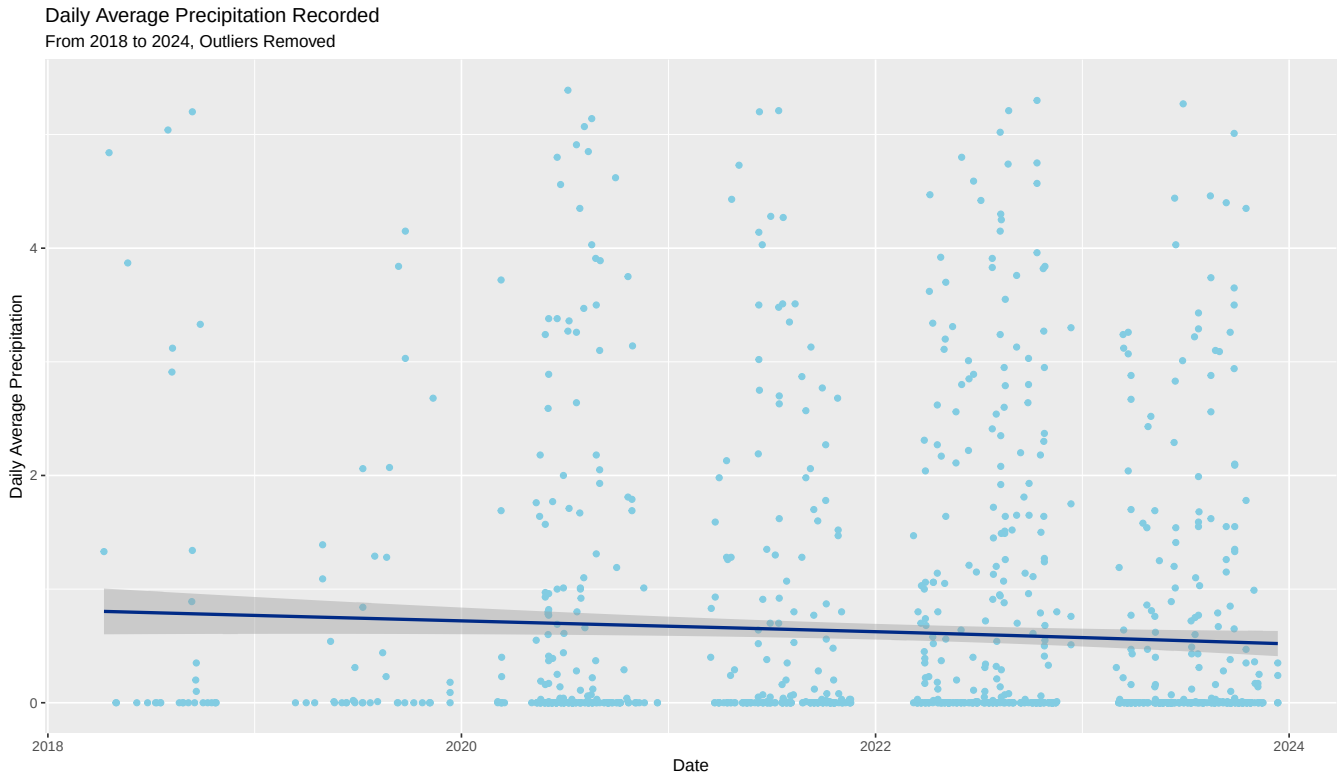



Figure 9: Time Series Visualization of Average Precipitation

### 7. Plots of AQI over time by counties:

While we see some trends in AQI between certain counties spiking near 2024, AQI values seem quite different between counties in general.

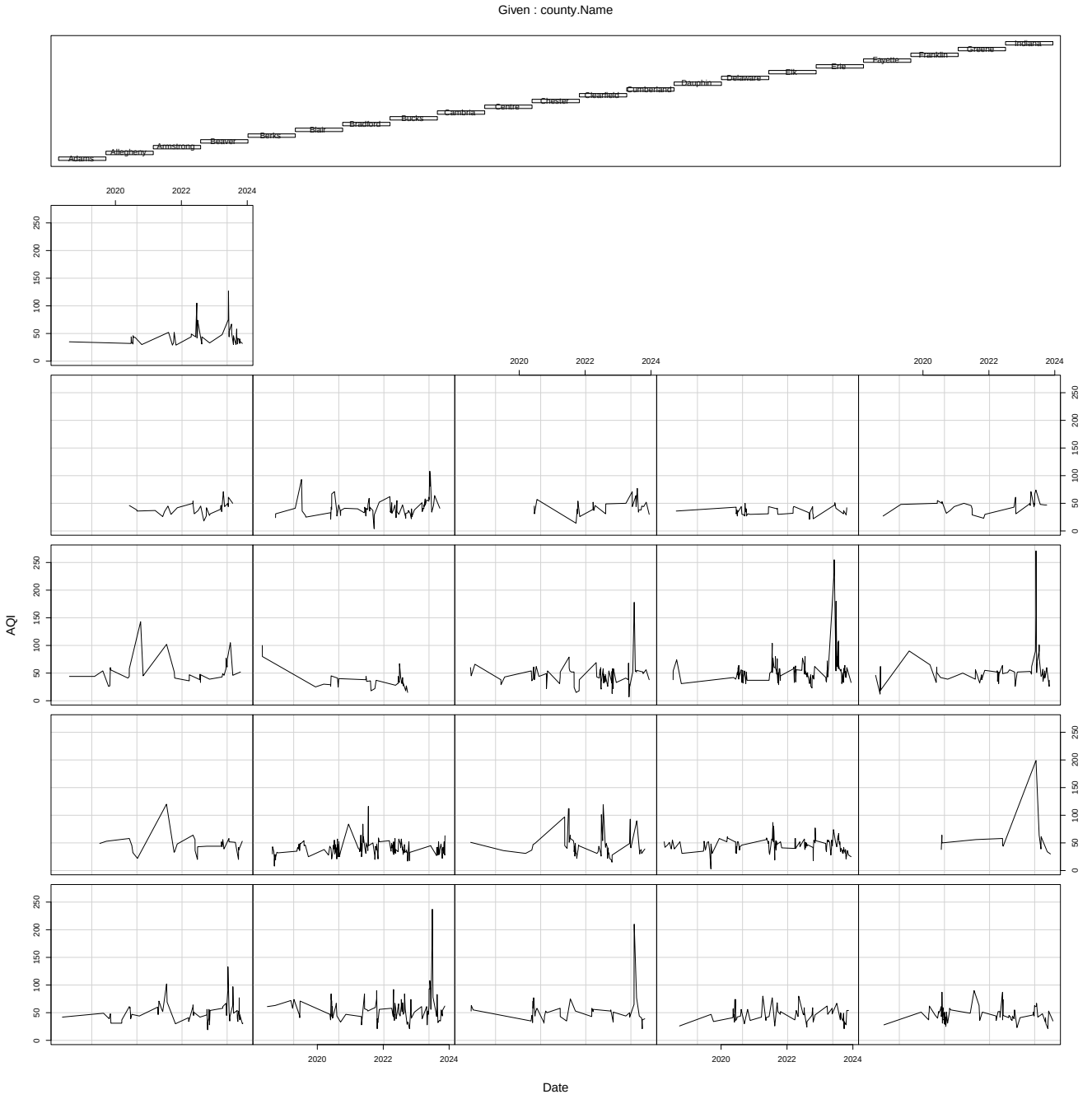



Figure 10: AQI over time, deaggregated by counties

The following plot shows differences in AQI mean between counties, reiterating the idea that an ols model might not account for heterogeneity across groups over time.

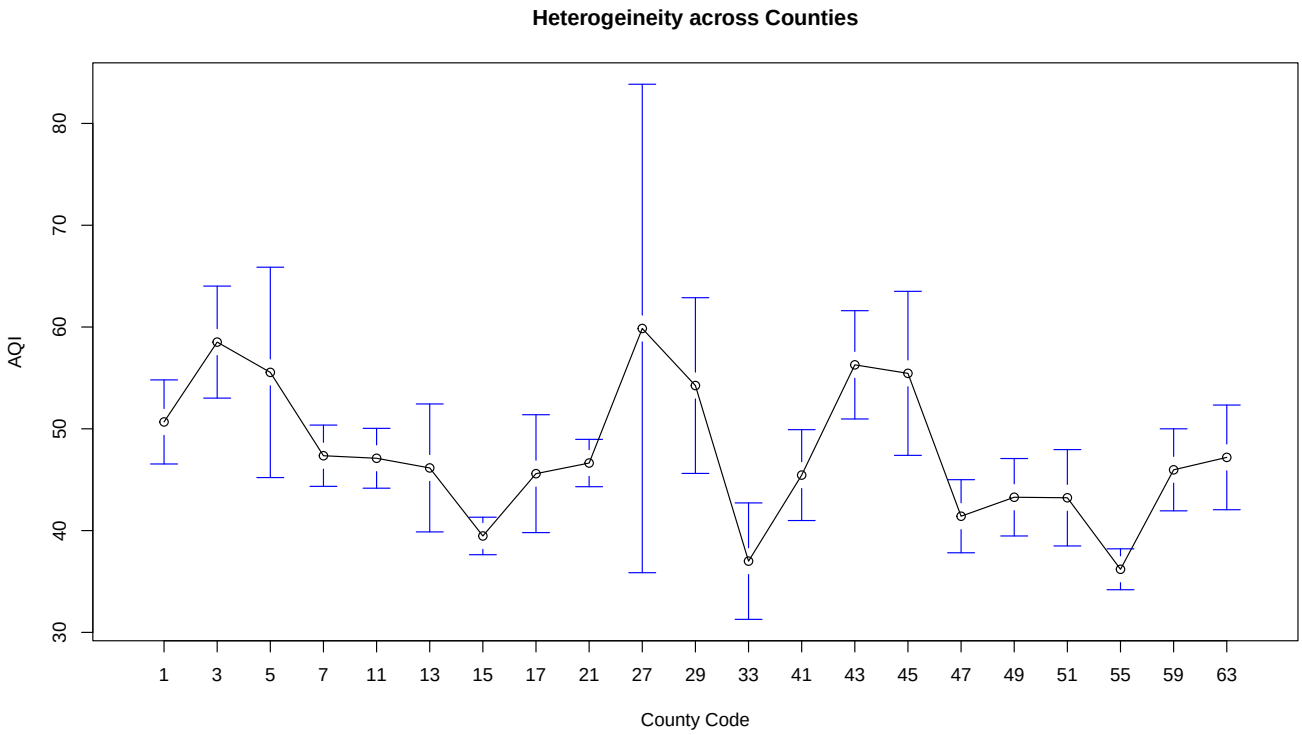



Figure 11: Differences in AQI means across counties

### 8. Variable selection for Ordinary Least Square:

I started with a simple regression model between traffic volume (traffic/truck counts and changes) with AQI values without adding the remaining variables. The regression shows a strong relationship between change in traffic volume and change in truck volume with air quality, but the coefficients are quite contradictory: While traffic volume change and AQI is inversely related (an increase in one leads to a decrease in the other), truck volume change and AQI is not. Moreover, a small AQI means healthier air quality, so an increase in traffic leading to a smaller AQI seems deviant from our expectations. This result cannot be intuitively interpreted and justified.

```
##
## Call:
## lm(formula = AQI ~ Current.Traffic + Delta.Traffic + Current.Truck +
##     Delta.Truck, data = filtered_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.336 -11.707  -3.423   6.043 222.706
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.828e+01  6.795e-01  71.050  < 2e-16 ***
## Current.Traffic  4.787e-05  4.422e-05   1.083  0.27920
## Delta.Traffic   -1.424e-03  4.785e-04  -2.976  0.00297 **
## Current.Truck   -3.480e-04  3.972e-04  -0.876  0.38103
## Delta.Truck      1.000e-02  4.926e-03   2.030  0.04250 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.91 on 1455 degrees of freedom
## Multiple R-squared:  0.008285,   Adjusted R-squared:  0.005559
```

```
## F-statistic: 3.039 on 4 and 1455 DF,  p-value: 0.01653
```

I then pass all the data through a step function to get the "best" ols model.

```
## Start:  AIC=6446.13
## AQI ~ County.Code + Date + daily.Avg.precipitation + daily.Avg.temp +
##     Current.Truck + Delta.Truck + Current.Traffic + Delta.Traffic +
##     Category + Defining.Parameter + county.Name
##
##
## Step:  AIC=6446.13
## AQI ~ Date + daily.Avg.precipitation + daily.Avg.temp + Current.Truck +
##     Delta.Truck + Current.Traffic + Delta.Traffic + Category +
##     Defining.Parameter + county.Name
##
##                            Df Sum of Sq    RSS    AIC
## - Date                      1         2 114927 6444.2
## - Current.Truck             1        64 114989 6444.9
## <none>                                  114925 6446.1
## - Delta.Truck               1       209 115134 6446.8
## - Current.Traffic           1       272 115197 6447.6
## - Delta.Traffic             1       472 115397 6450.1
## - daily.Avg.temp            1      1916 116841 6468.3
## - Defining.Parameter        4      3030 117955 6476.1
## - daily.Avg.precipitation   1      2859 117784 6480.0
## - county.Name              20      6244 121169 6483.4
## - Category                  4    479267 594192 8836.8
##
## Step:  AIC=6444.16
## AQI ~ daily.Avg.precipitation + daily.Avg.temp + Current.Truck +
##     Delta.Truck + Current.Traffic + Delta.Traffic + Category +
##     Defining.Parameter + county.Name
##
##                            Df Sum of Sq    RSS    AIC
## - Current.Truck             1        64 114991 6443.0
## <none>                                  114927 6444.2
## - Delta.Truck               1       214 115141 6444.9
## - Current.Traffic           1       325 115252 6446.3
## - Delta.Traffic             1       564 115491 6449.3
## - daily.Avg.temp            1      1953 116880 6466.8
## - Defining.Parameter        4      3030 117957 6474.1
## - daily.Avg.precipitation   1      2882 117809 6478.3
## - county.Name              20      6304 121231 6482.1
## - Category                  4    483068 597995 8844.1
##
## Step:  AIC=6442.97
## AQI ~ daily.Avg.precipitation + daily.Avg.temp + Delta.Truck +
##     Current.Traffic + Delta.Traffic + Category + Defining.Parameter +
##     county.Name
##
##                            Df Sum of Sq    RSS    AIC
## - Delta.Truck               1       150 115142 6442.9
## <none>                                  114991 6443.0
## - Current.Traffic           1       351 115342 6445.4
## - Delta.Traffic             1       516 115507 6447.5
## - daily.Avg.temp            1      1942 116934 6465.4
## - Defining.Parameter        4      3024 118015 6472.9
## - daily.Avg.precipitation   1      2879 117871 6477.1
```

```
## - county.Name             20      6272 121264 6480.5
## - Category                 4    483475 598466 8843.3
##
## Step:  AIC=6442.88
## AQI ~ daily.Avg.precipitation + daily.Avg.temp + Current.Traffic +
##     Delta.Traffic + Category + Defining.Parameter + county.Name
##
##                           Df Sum of Sq    RSS    AIC
## <none>                                  115142 6442.9
## - Current.Traffic          1       344 115486 6445.2
## - Delta.Traffic            1       441 115582 6446.5
## - daily.Avg.temp           1      1903 117044 6464.8
## - Defining.Parameter       4      3083 118225 6473.5
## - daily.Avg.precipitation  1      2842 117984 6476.5
## - county.Name             20      6344 121486 6481.2
## - Category                 4    484359 599501 8843.8
##
## Call:
## lm(formula = AQI ~ daily.Avg.precipitation + daily.Avg.temp +
##     Current.Traffic + Delta.Traffic + Category + Defining.Parameter +
##     county.Name, data = filtered_df)
##
## Coefficients:
##                             (Intercept)                 daily.Avg.precipitation
##                               1.642e+00                               -2.177e-01
##                           daily.Avg.temp                          Current.Traffic
##                               1.719e-01                                2.409e-05
##                            Delta.Traffic                          CategoryModerate
##                               -2.555e-04                                2.218e+01
##                      CategoryUnhealthy  CategoryUnhealthy for Sensitive Groups
##                               1.435e+02                                7.304e+01
##                  CategoryVery Unhealthy                    Defining.ParameterNO2
##                               1.956e+02                                2.229e+01
##                 Defining.ParameterOzone                   Defining.ParameterPM10
##                               3.673e+01                                1.841e+01
##                 Defining.ParameterPM2.5                    county.NameAllegheny
##                               3.486e+01                                2.596e+00
##                   county.NameArmstrong                       county.NameBeaver
##                               -2.295e+00                               -1.260e-01
##                       county.NameBerks                        county.NameBlair
##                               -7.032e-01                               -2.589e+00
##                    county.NameBradford                        county.NameBucks
##                               -3.831e+00                               -3.912e+00
##                     county.NameCambria                       county.NameCentre
##                               -1.499e+00                               -2.571e-01
##                    county.NameChester                     county.NameClearfield
##                               1.371e+00                                -5.851e+00
##                  county.NameCumberland                       county.NameDauphin
##                               -4.777e+00                               -3.929e-01
##                    county.NameDelaware                          county.NameElk
##                               -7.757e-01                               -7.094e-01
##                        county.NameErie                       county.NameFayette
##                               -3.085e+00                               -1.791e+00
##                    county.NameFranklin                        county.NameGreene
##                               -5.031e+00                                1.743e+00
##                     county.NameIndiana
```

```
##                               -1.864e+00
```

**9. Compare between fixed-effects, random-effects, and ols models:**

The following fixed-effects model shows no relationship between AQI and traffic volume across counties between 2018 and 2023.

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = AQI ~ Current.Traffic + Current.Truck + Delta.Traffic +
##     Delta.Truck, data = filtered_df, model = "within", index = c("Date",
##     "County.Code"))
##
## Unbalanced Panel: n = 513, T = 1-12, N = 1460
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -85.3080  -4.5018   0.0000   3.8410 129.0995
##
## Coefficients:
##                     Estimate  Std. Error t-value Pr(>|t|)
## Current.Traffic -1.9371e-05  3.5798e-05 -0.5411   0.5886
## Current.Truck   -6.4554e-06  2.9417e-04 -0.0219   0.9825
## Delta.Traffic   -1.8140e-04  4.1959e-04 -0.4323   0.6656
## Delta.Truck      2.3882e-03  4.0374e-03  0.5915   0.5543
##
## Total Sum of Squares:    182730
## Residual Sum of Squares: 182320
## R-Squared:      0.0022538
## Adj. R-Squared: -0.5437
## F-statistic: 0.532539 on 4 and 943 DF, p-value: 0.71186
```

We arrive at the same conclusion of no significant relationship with a random-effects model.

To choose the most appropriate model, we can conduct a Hausman test.[20] With p-value $> 0.05$, we fail to reject the null hypothesis that unique errors are not correlated with the regressors.[21] We should use the random-effects model.

```
##
##  Hausman Test
##
## data:  AQI ~ Current.Traffic + Current.Truck + Delta.Traffic + Delta.Truck
## chisq = 3.6215, df = 4, p-value = 0.4597
## alternative hypothesis: one model is inconsistent
```

Using a simple F test for effects, we can also see that the fixed-effects model is still better than ols. At a p-value $>$ 0.05, we fail to reject the null that the "best" ols model is better than than the fixed-effects model.[22]

```
##
##  F test for individual effects
##
## data:  AQI ~ Current.Traffic + Current.Truck + Delta.Traffic + Delta.Truck
## F = -0.71789, df1 = 484, df2 = 943, p-value = 1
## alternative hypothesis: significant effects
```

[20] Oscar Torres-Reyna, "Getting Started in Fixed/Random Effects Models using R/RStudio," 16
[21] Ibid, 16
[22] Ibid, 12.

**10. More detail on cluster 0:**

The following plot shows the hyperdimensional outliers and specifies the county of each point. Among the outliers, we see two groups: high AQI, low traffic volume and low AQI, (mostly) high traffic volume. This inverse relationship contradicts common understand of the relationship between air quality and car exhaust fumes; thus, it is reasonable for them to be "outliers" compared to other points. We also see that Bradford and Cambria appear more often than other counties (4 points each), which is consistent with the imbalances inherent in the dataset.

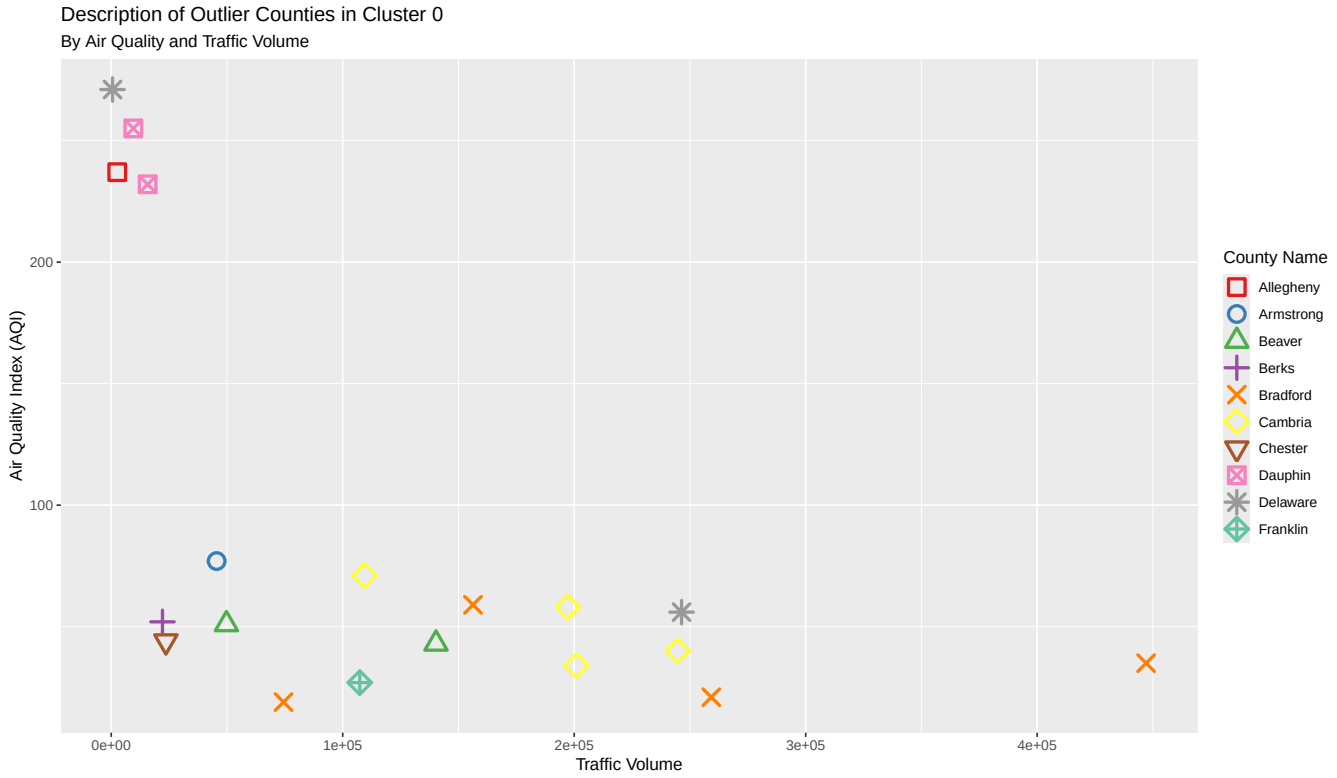



Figure 12: Characteristics of Counties in Cluster 0

I want to provide some insights on these intense deviations from the norm. Extreme AQI values might just be a special weather event happening in the county that has no relevance to traffic for that date. Table 4 summarizes the date, AQI, and traffic volumes of all records in cluster 0. For the first record of Allegheny county on June 29, 2023, there was a wildfire in Canada, of which smoke traveled to the county and tanked air quality.[23]

In terms of traffic volume, due to the differences between the number of streets in a county as well as imbalances in the dataset, an intense traffic volume might just be due to the data cleaning process where traffic counts for each street on a particular date is added together, which has no relationship to AQI. An ols regression might be quite vulnerable to these data points, which is why controlling for differences between counties *and* dates is important.

Table 6: Counties and Dates in Cluster 0

| | County | Date | AQI | Traffic Volume |
|---|---|---|---|---|
| 169 | Allegheny | 2023-06-29 | 237 | 2593 |
| 190 | Armstrong | 2020-06-09 | 77 | 45549 |
| 231 | Beaver | 2020-05-13 | 51 | 49775 |
| 256 | Beaver | 2021-06-16 | 43 | 140296 |
| 319 | Berks | 2020-08-05 | 52 | 22132 |

[23] CBS Pittsburgh, "Canadian Wildfire Smoke Creates Unhealthy Air Quality across Pittsburgh Area," June 30, 2023, https://www.cbsnews.com/pittsburgh/live-updates/live-updates-canadian-wildfire-smoke-clouds-pittsburgh-skies-air-quality-alerts-issued-for-western-pa/.

| | | | | |
|---|---|---|---|---|
| 426 | Bradford | 2018-09-25 | 19 | 74426 |
| 518 | Bradford | 2021-11-09 | 59 | 156275 |
| 557 | Bradford | 2022-08-23 | 35 | 447058 |
| 563 | Bradford | 2022-09-27 | 21 | 259234 |
| 718 | Cambria | 2021-08-09 | 71 | 109532 |
| 727 | Cambria | 2022-03-29 | 40 | 244709 |
| 729 | Cambria | 2022-04-05 | 58 | 197308 |
| 782 | Cambria | 2023-09-13 | 34 | 200896 |
| 809 | Chester | 2019-05-15 | 44 | 23622 |
| 1053 | Dauphin | 2023-06-07 | 232 | 15729 |
| 1054 | Dauphin | 2023-06-08 | 255 | 9547 |
| 1108 | Delaware | 2021-08-05 | 56 | 246385 |
| 1140 | Delaware | 2023-06-07 | 271 | 566 |
| 1335 | Franklin | 2020-09-17 | 27 | 107369 |