Udacity_Project_EDA

Edward Carter

7/21/2020

==========

```
## Warning in California$SITE_ID == x: longer object length is not a multiple of
## shorter object length
```

```
## 'data.frame':    979 obs. of  27 variables:
##  $ SITE_ID      : chr  "YOS404" "JOT403" "SEK430" "YOS404" ...
##  $ TYPE         : chr  "C" "C" "C" "C" ...
##  $ DATEON       : Date, format: "2020-04-14" "2020-04-14" ...
##  $ DATEOFF      : Date, format: "2020-04-21" "2020-04-21" ...
##  $ TSO4         : num  0.587 0.663 0.827 0.385 0.531 ...
##  $ TNO3         : num  0.254 0.637 0.387 0.166 0.256 ...
##  $ TNH4         : num  0.133 0.268 0.367 0.137 0.145 ...
##  $ CA           : num  0.0674 0.1408 0.0432 0.0958 0.0486 ...
##  $ MG           : num  0.0281 0.0335 0.017 0.0156 0.0198 0.0196 NA 0.0203 0.0138 0.0328 ...
##  $ NA.          : num  0.114 0.1568 0.0756 0.0307 0.1081 ...
##  $ K            : num  0.0351 0.0334 0.051 0.021 0.0232 0.0496 NA 0.0328 0.0242 0.0454 ...
##  $ CL           : num  0.0146 0.0209 0.0163 0.0145 0.0271 0.0158 NA NA NA NA ...
##  $ NSO4         : num  0.0444 0.0304 0.0484 0.0582 0.0534 0.0959 NA 0.0813 0.0602 0.0605 ...
##  $ NHNO3        : num  0.293 0.865 0.472 0.507 0.179 ...
##  $ WSO2         : num  0.0866 0.1507 0.1875 0.0723 0.0921 ...
##  $ WNO3         : num  NA NA NA NA NA ...
##  $ TOTAL_SO2    : num  0.116 0.171 0.22 0.111 0.128 ...
##  $ TOTAL_NO3    : num  0.543 1.488 0.852 0.665 0.432 ...
##  $ FLOW_VOLUME  : num  34.2 32.9 30.7 34.4 34 ...
##  $ VALID_HOURS  : int  166 159 167 169 166 167 NA 167 146 170 ...
##  $ COMMENT_CODES: chr  "T01 T04" "T04" "T04" "" ...
##  $ STD2LOCAL_CF : num  1.14 1.11 1.02 1.14 1.12 ...
##  $ TEMP_SOURCE  : chr  "sa" "sa" "sa" "sa" ...
##  $ QA_CODE      : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ UPDATE_DATE  : Date, format: "2020-06-30" "2020-06-30" ...
##  $ YEAR         : chr  "2020" "2020" "2020" "2004" ...
##  $ MONTH        : chr  "04" "04" "04" "03" ...
```

The United States Environmental Protection Agency maintains a program known as the Clean Air Status and Network (CASTNET). This program is a long-term environmental monitoring network. This network is maintained in the United States and Canada and consists of 97 sites. Established by the 1990 Clean Air Act Amendments to provide accountability for emission reduction programs by reporting trends in pollutant concentrations and acidic deposition. The dataset utilized in this project is a measurement of the weekly ambient concentrations of Suflate (SO4), Nitrate (NO4), Ammonium (NH4), Calcium (Ca), Magnesium (Mg), Sodium (Na), Potassium (K), Chloride (Cl), Nitric Acid (NO3), and Sulfur Dioxide (NO3).

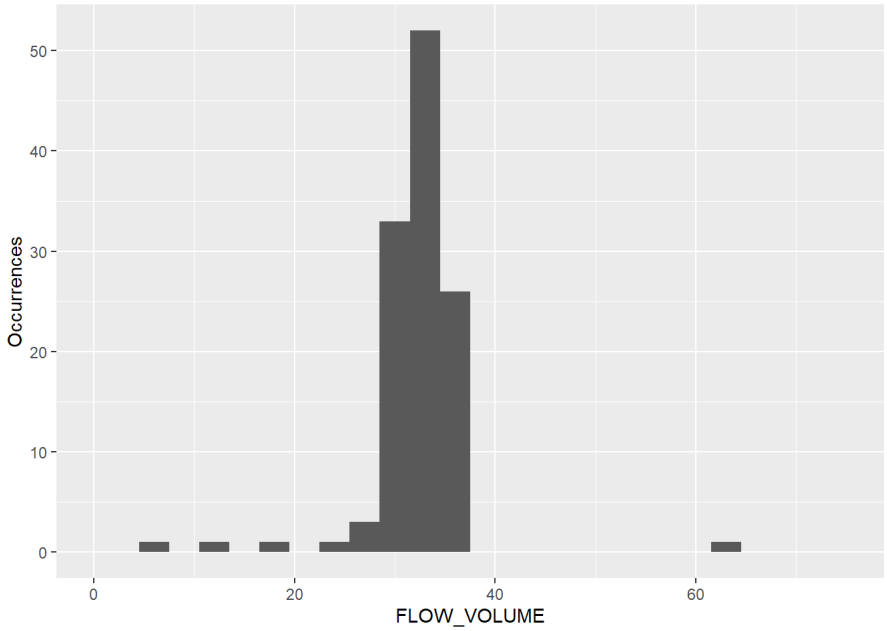
Particle Pollution (PM) inlcudes: Sulfate, Nitrate, Ammonium, Magnesium, Calcium, Potassium, Sodium and Chloride. Based on the 1990 Clean Air Act these fine inhalable particulate matter (PM2.5) thresholds are (per cubic meter of air): 12.0 micrograms (Annual Mean over 3 years) 15.0 micrograms (Annual Mean over 3 years) 35.0 micrograms (98th Percentile averaged over 3 years)

The gaseous pollution of Sulfur Dioxide threshold is 75ppb (196504.05 ug/m3) per hour / .5 ppm (1310.027 ug/m3) per 3 hours.

This dataset is an average of each week from January 2000 to June 2020.

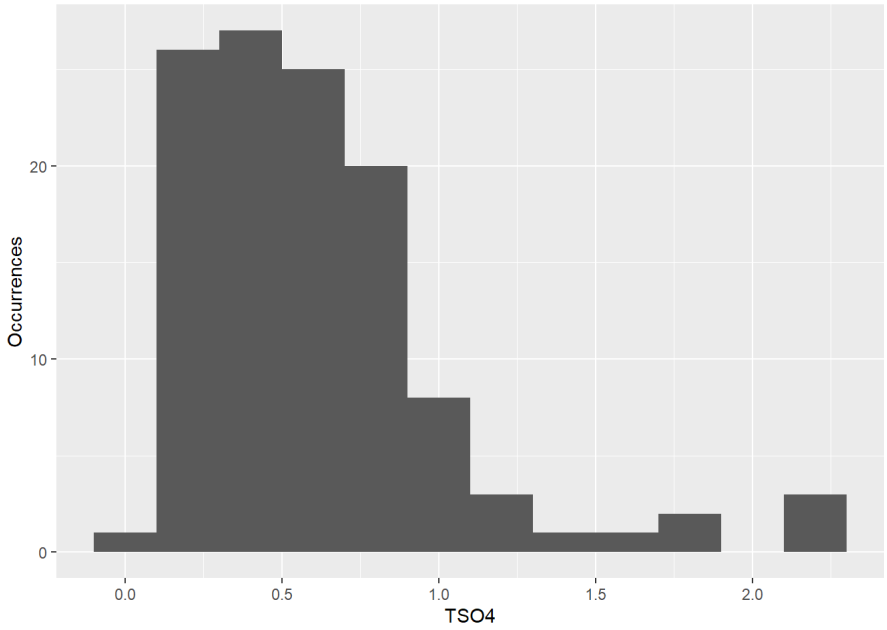
Univariate Plots Section

The following graph shows the Air Flow Volume. This shows a centralized grouping around the 30-35 m3 amount. The graph provides a baseline understanding that the identified sites should all have the same Air Flow Volume.



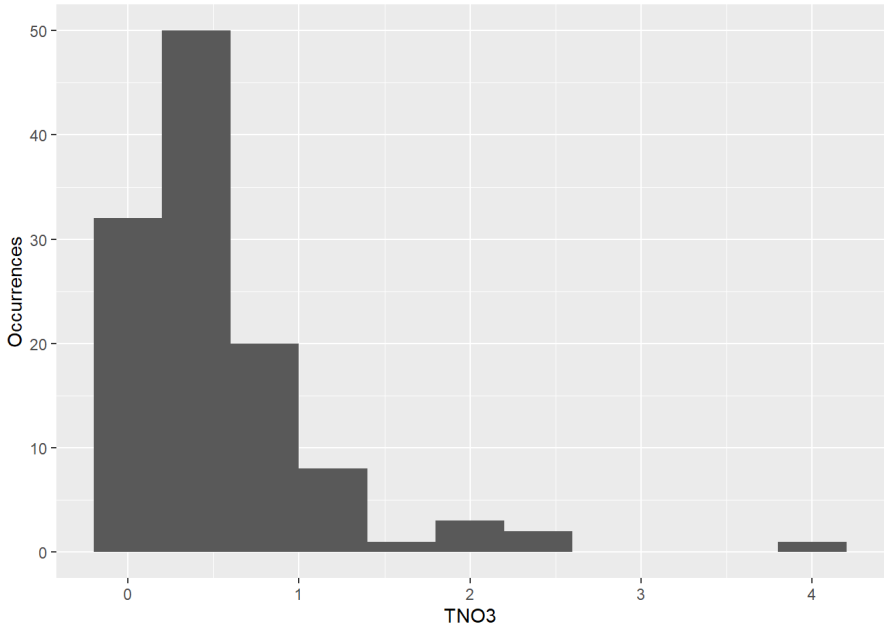

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   5.249  30.923  32.503  35.092  34.394 204.912       1
```

The next graph shows that Sulfate (SO4) is normally low with the majority of occurrences happening around the .5 micro-gram level. The outlier of above 2.0 is a likely contributor to a site's low rating on the Air Quality Rating.



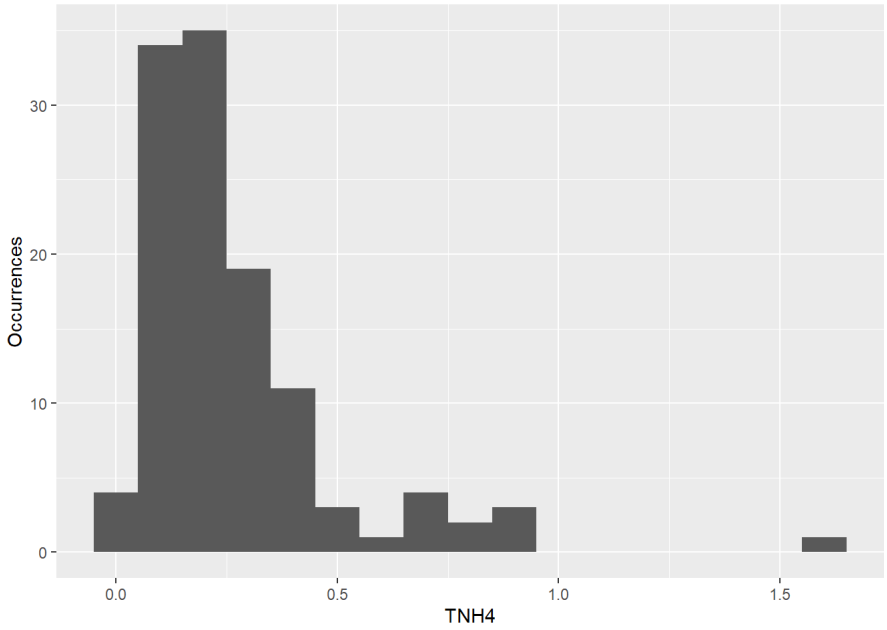

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0995  0.3160  0.5238  0.6115  0.7591  2.2513       5
```

This graph shows that nitrates are usually below the 1.5 micro-gram. Again an outlier is present at 4 micro-gram.



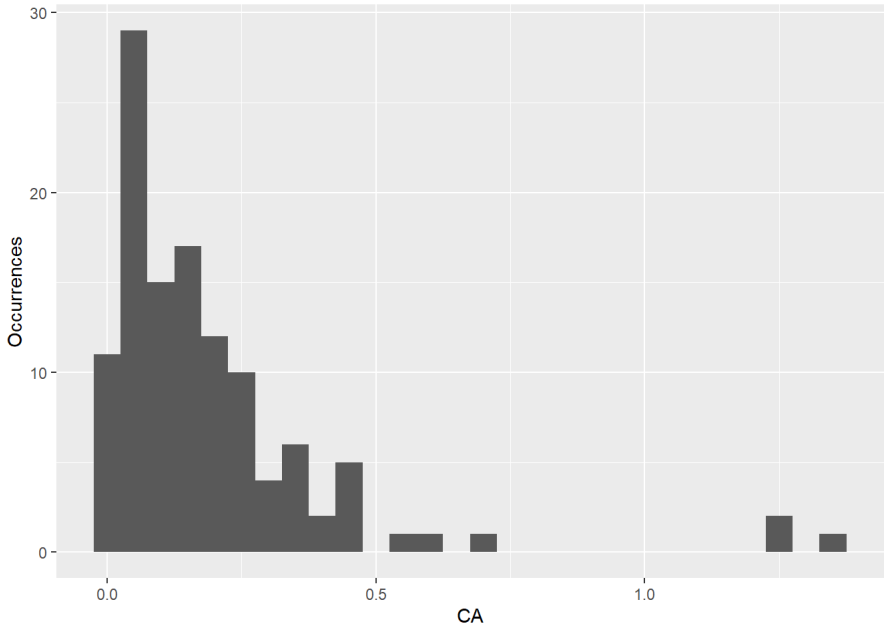

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0316  0.1856  0.3957  0.5359  0.6680  3.8634       5
```

The ammonium graph following continues with the meme the previous graphs have shown. A concentration of most occurrences happening on the low side of the chart with a rare occurrence on the high side.



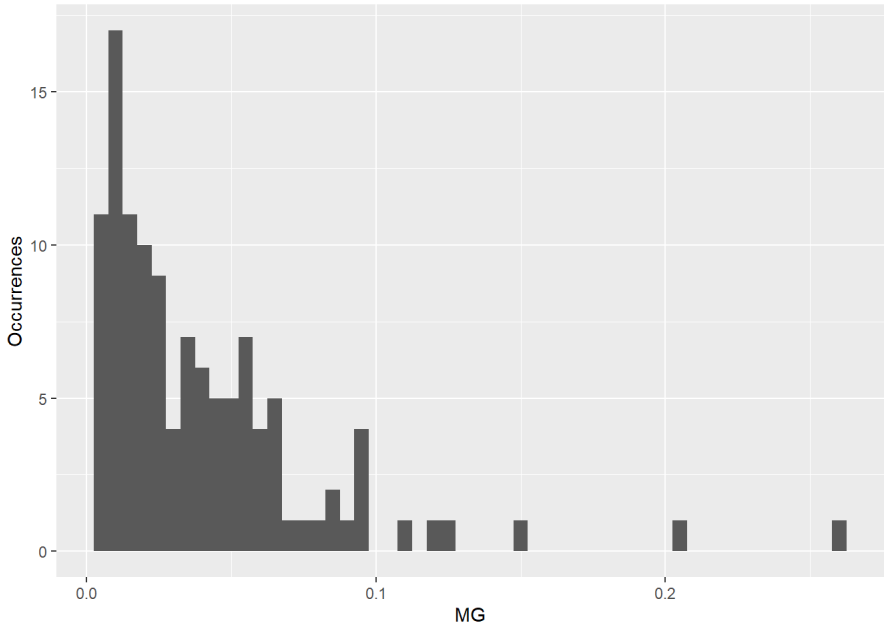

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0305  0.1291  0.2019  0.2641  0.3215  1.5678       5
```

Calcium's chart again shows that calcium is usually found in small amounts. However, there is 3 occurrences that can be seen on the high side.



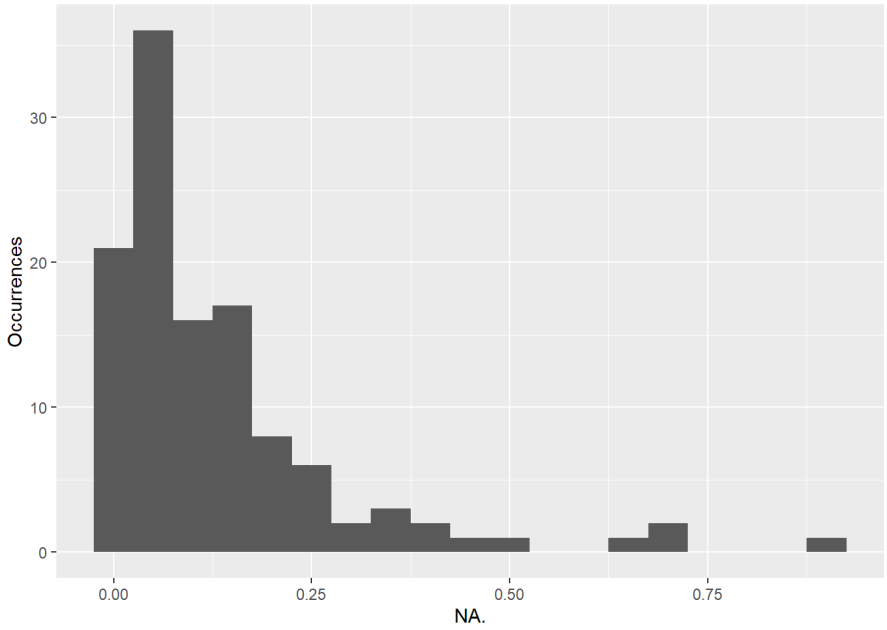

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0125  0.0489  0.1294  0.1940  0.2543  1.3536       5
```

Magnesium exhibits the same characteristics of low level concentrations.



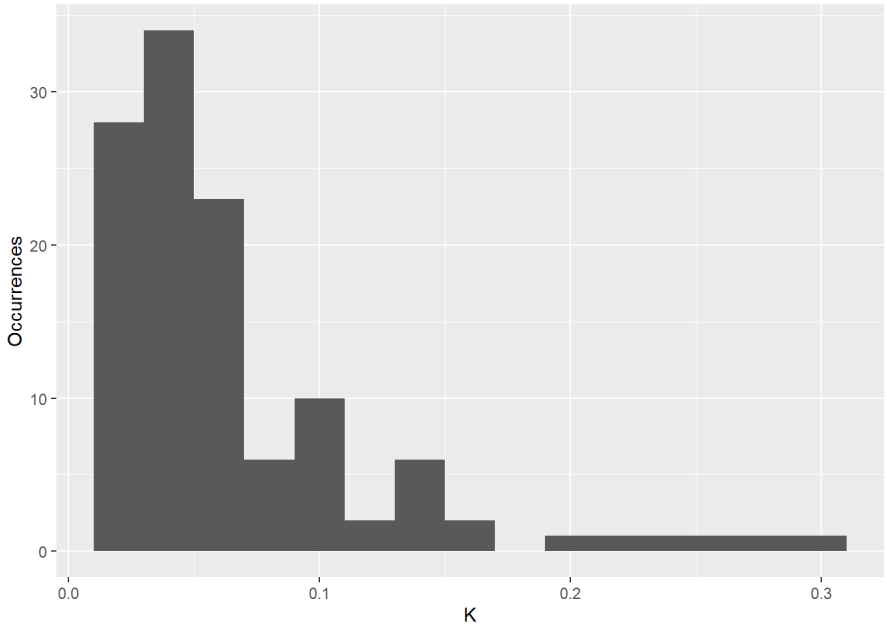

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00370 0.01350 0.02810 0.04012 0.05590 0.25980       5
```

Sodium is completely below the 1 micro-gram level with most closer to the non-existent levels.



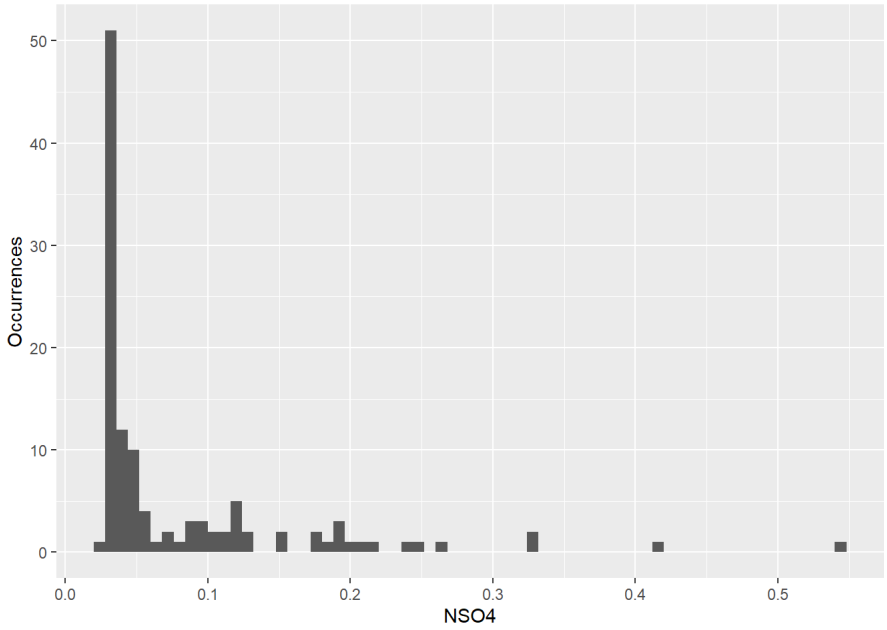

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0076  0.0315  0.0764  0.1315  0.1670  0.9180       5
```

Potassium's occurrences are below .5 micro-grams for all occurrences.



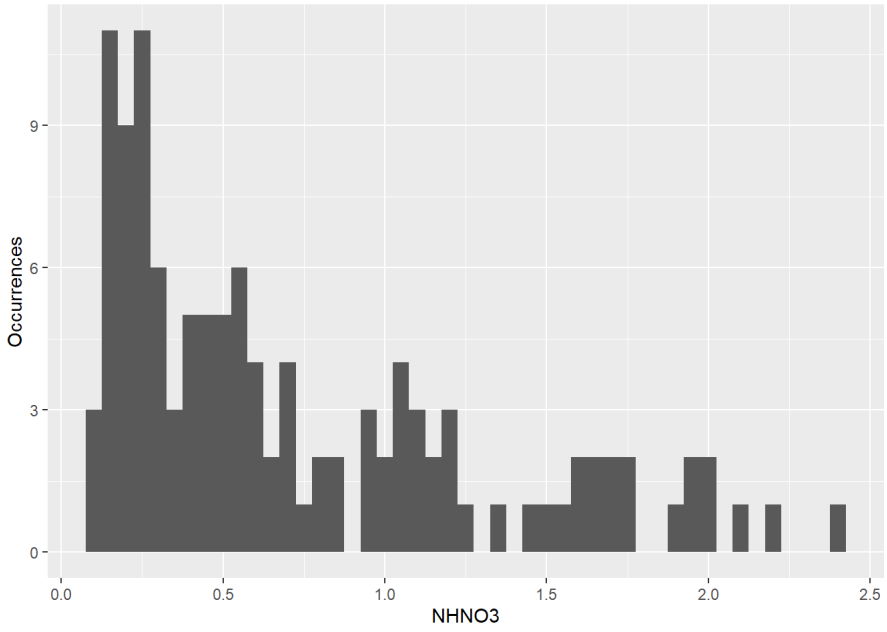

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01020 0.03070 0.04790 0.06513 0.08140 0.29150       5
```

This sulfate graph shows the results from a nylon filter versus the teflon filter of the previous graph. Nylon provides for more refined capture of material. Becuase of this smaller pore size the sample is smaller. The graph shows that Sulfate is concentrated below the .3 micro-gram level with few outliers.



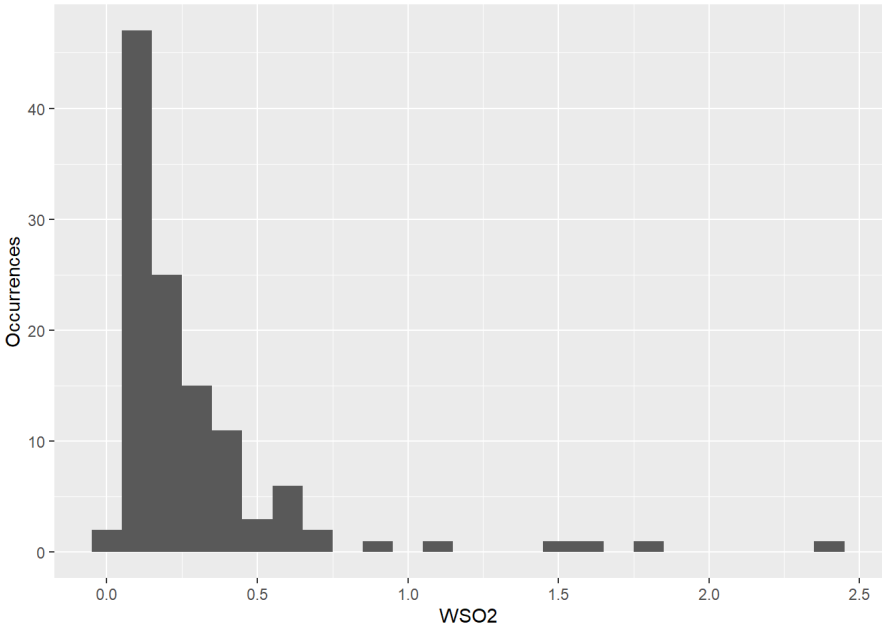

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.02430 0.03140 0.04070 0.07969 0.09430 0.54640       5
```

Nitric Acid is a strong acid that causes acid rain when present in the air. The levels of Nitric Acid in the samples are concentrated below 2 micro-grams.



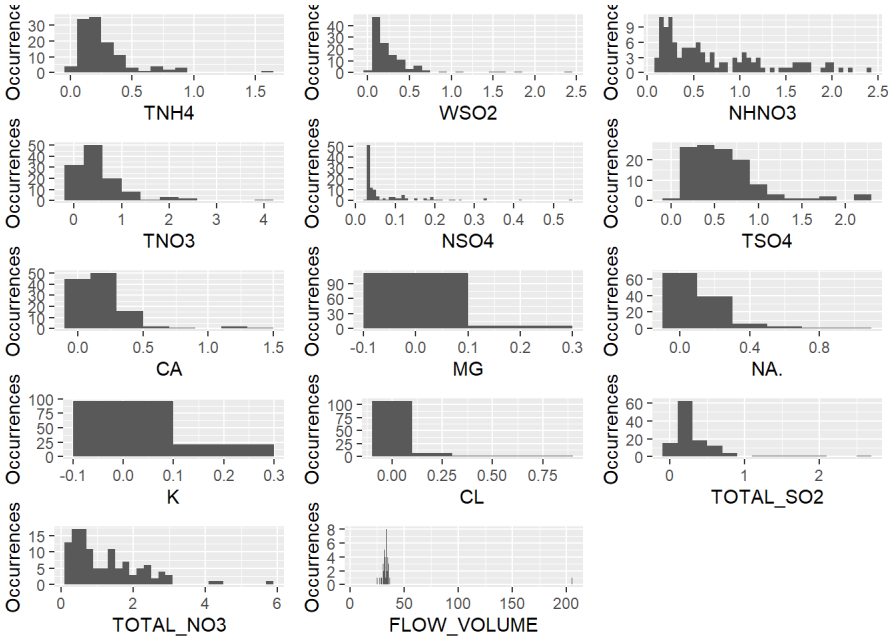

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.1016  0.2450  0.5378  0.7281  1.0649  2.3965       5
```

Another contributor to acid rain is Sulfur Dioxide. The below graph shows that Sulfur Dioxide is mainly concentrated below the 1 micro-gram level.

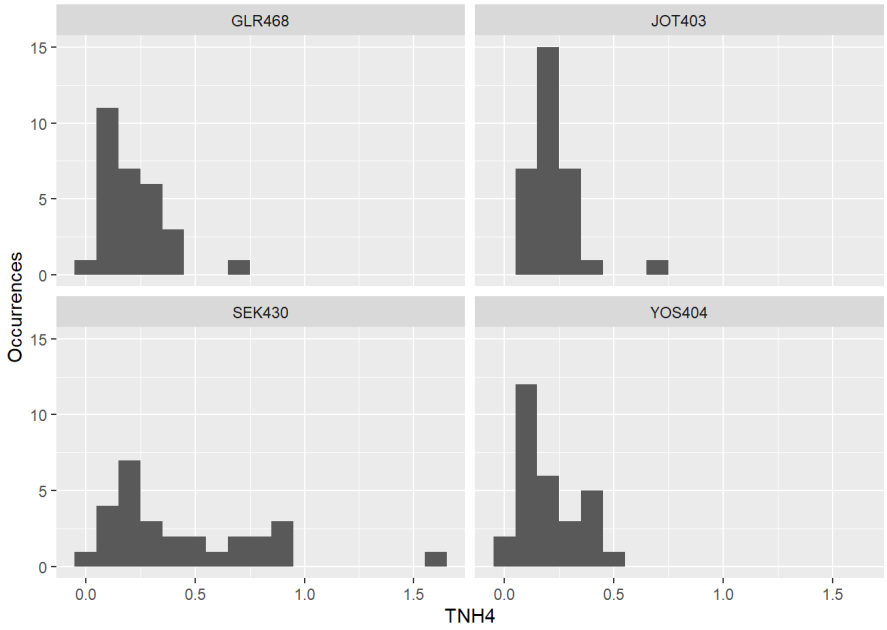



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0463  0.1099  0.1697  0.2912  0.3440  2.3669       5
```

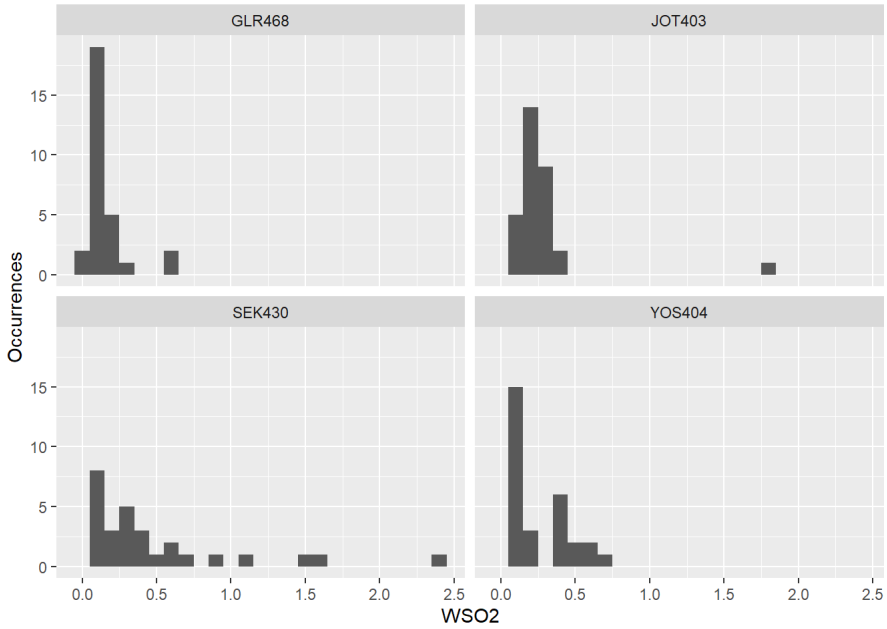
Below is a snap shot of all the above graphs in one.



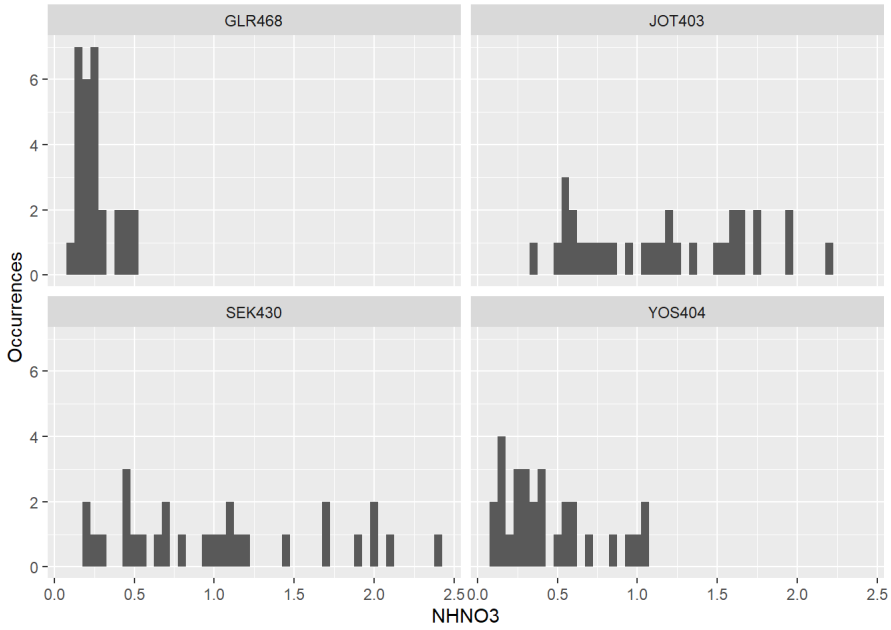
Glacier National Park is used as our baseline for our comparison as it experiences the best Air Quality rating in the US.



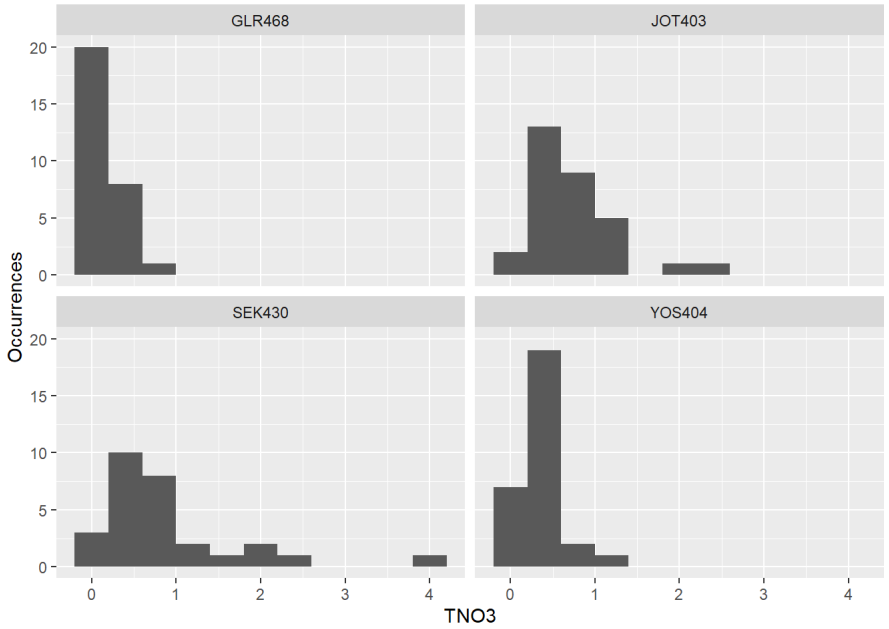
This graph shows that YOS404 is the only site that experiences less Ammonium amongst the four locations.



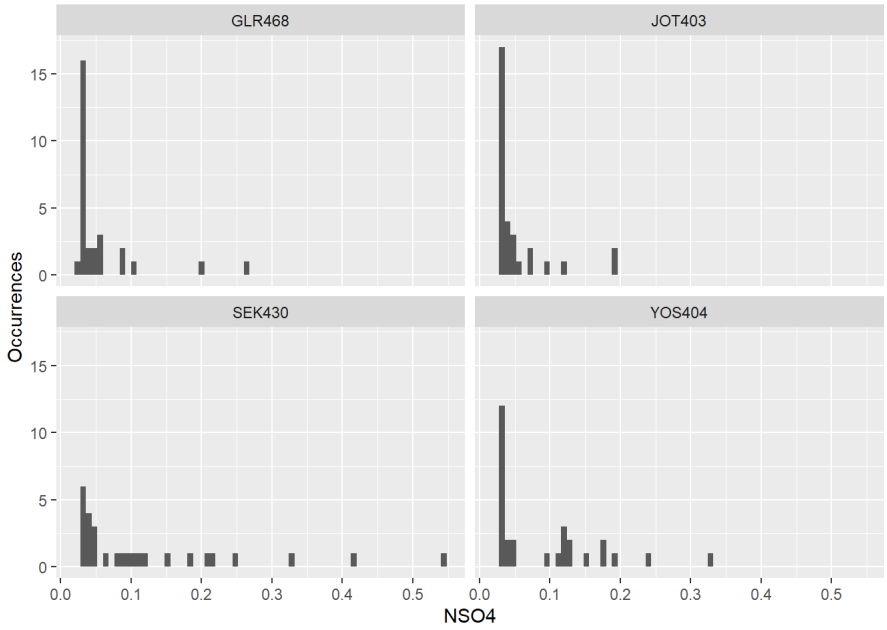
GLR468's Sulfur Dioxide is much lower than SEK430 and YOS 404. However, the concentration for JOT403 and GLR468 are both between 0 and .5 micro-grams.



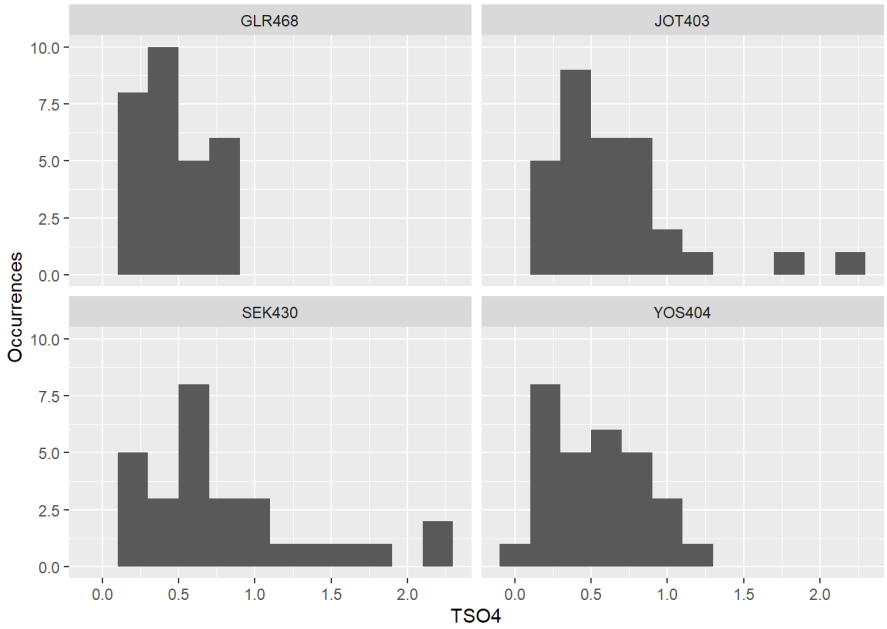
Nitric Acid is extremelow in GLR468. However, JOT403 and SEK430 has nitric acid measurements from .5 to 2.25 micro-grams.



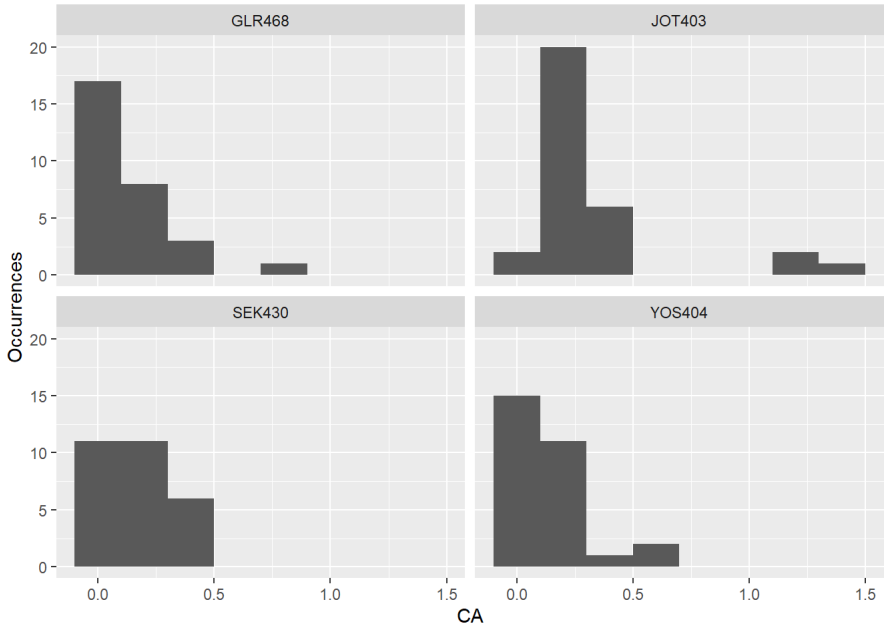
Nitrate is significantly lower in GLR 468 and only comparable to YOS404.



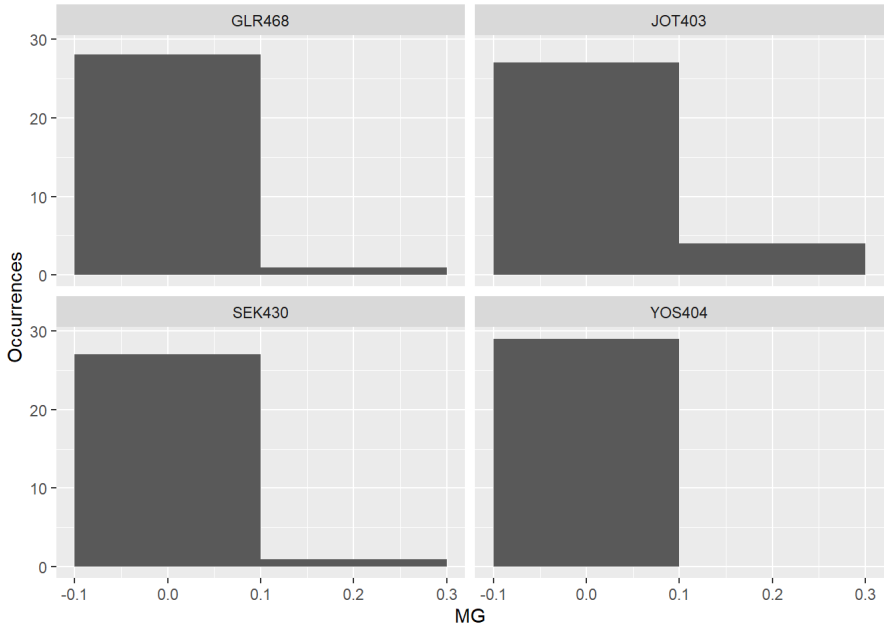
SEK 430 sees Sulfate measurements from 0 to .5 micro-grams. While GLR468 and JOT403 maintain levels at .1.



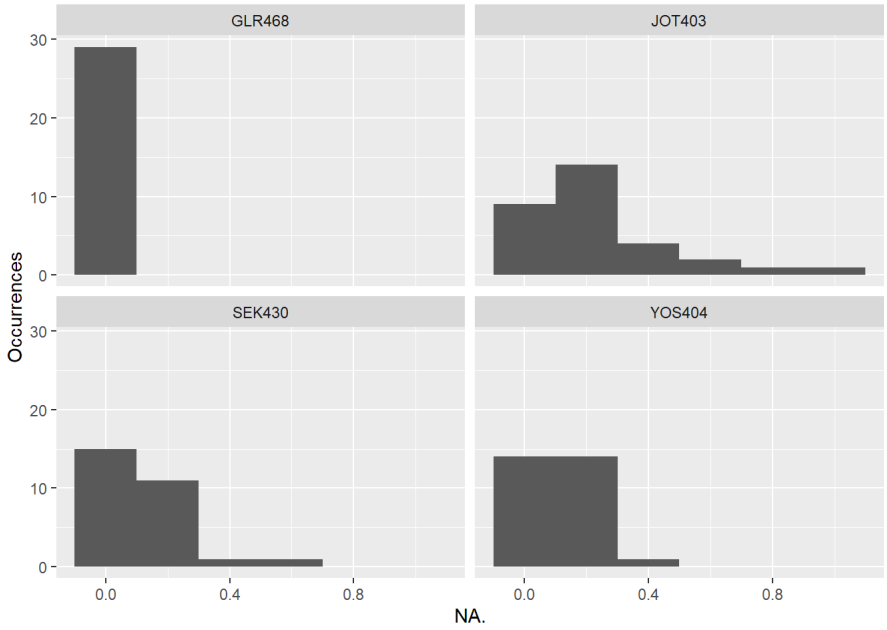
Sulfate through the Teflon filter shows the same as the Nylon filter.



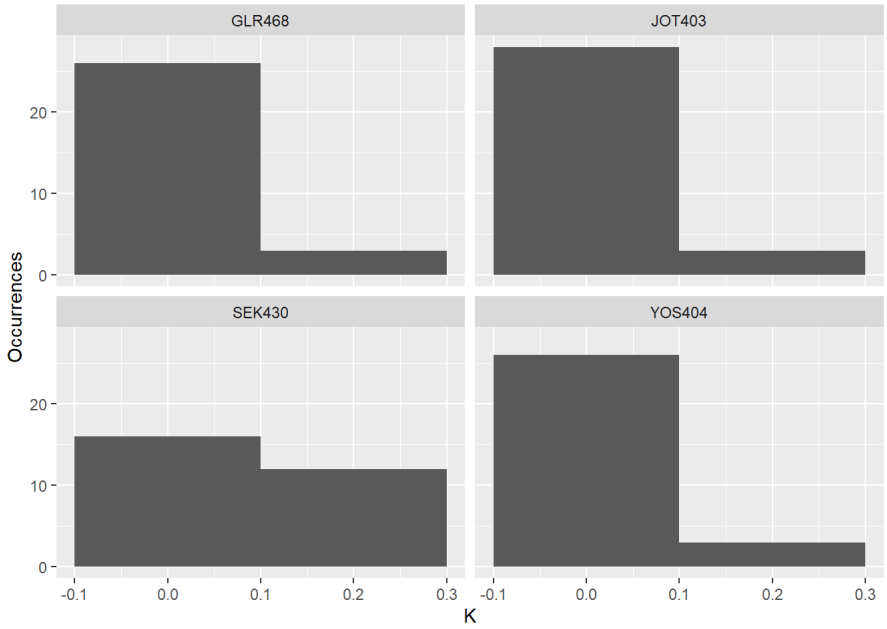
Calcium measured in GLR468, SEK430, and YOS40 are equal and mainly below the .5 micro-gram level.



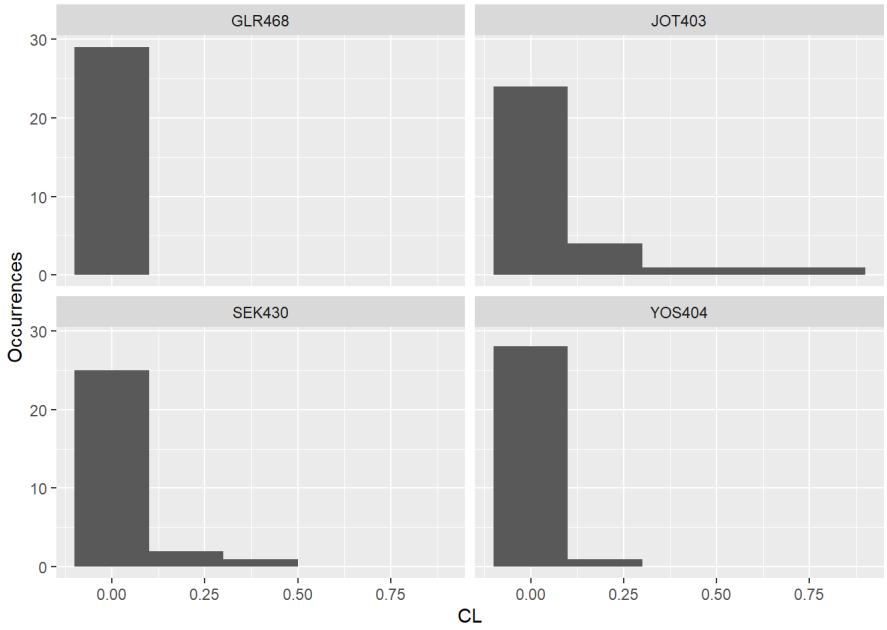
Magnesium in GLR468 adn SEK430 are the exact same while YOS404 has none.



Sodium is not present in GLR468 but is prevalent in JOT403.



Chloride is mostly not present in all location. However, JOT403 and SEK430 does have multiple occurrences of chloride.



Univariate Analysis

What is the structure of your dataset?

The data-set contains 20 years of data collected from 9 different observation posts within the United States of America. While the majority (8) of the observations posts are in California the last one is located in Montana (GLR468). The number of week collected range from the newest at 251 week to the oldest at 1048 weeks. GLR468 has 1040 weeks while YOS404 (Yosemite National Park's Turtleback Dome) has 1048 weeks.

Other Observations:

Based on "https://www.epa.gov/castnet/castnet-ozone-monitoring (https://www.epa.gov/castnet/castnet-ozone-monitoring)" website Glacier National Park's GLR468 had the lowest 8-hour daily maximum ozone concentration at 53. This 53 is tied for the lowest concentration for the United States.

Contrarily, YOS404 (Turtleback Dome - Yosemite National Park), SEK430 (Ash Mountain - Sequoia National Park), JOT403 (Joshua Tree National Park) had ratings of 77, 85, 88 respectively. These rating are the highest in the nation and are all in California.

What is/are the main feature(s) of interest in your dataset?

The main feature of interest is the comparison between GLR458, YOS404, SEK430, and JOT403. The initial (univariate data plots) are counter intuitive because GLR468 has the highest amount of the identified pollutants. LAV410 had the second highest amount of polluntants and also has the second lowest rating of 66.

What other features in the dataset do you think will help support your

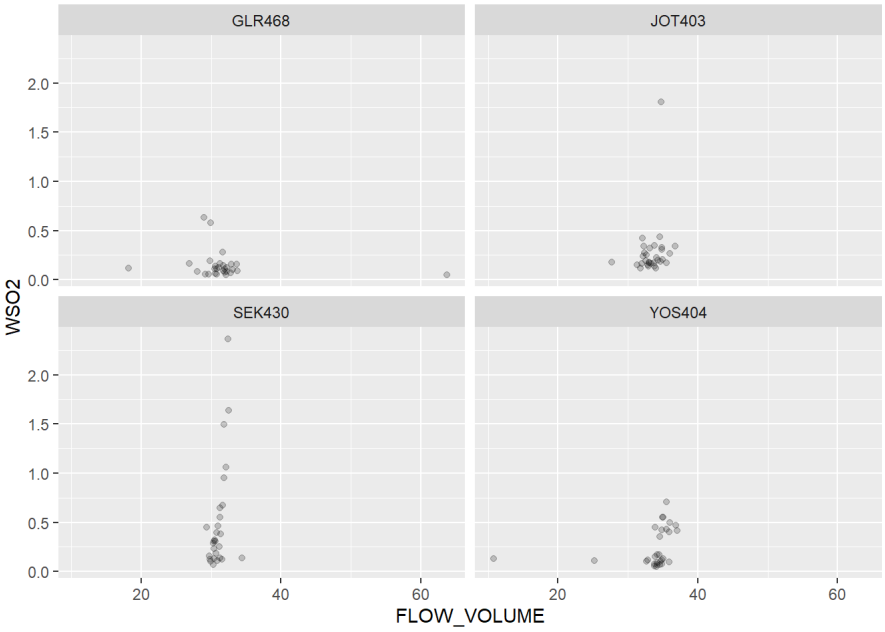
investigation into your feature(s) of interest? A restriction of time frame will likely show a reduction in air pollutants

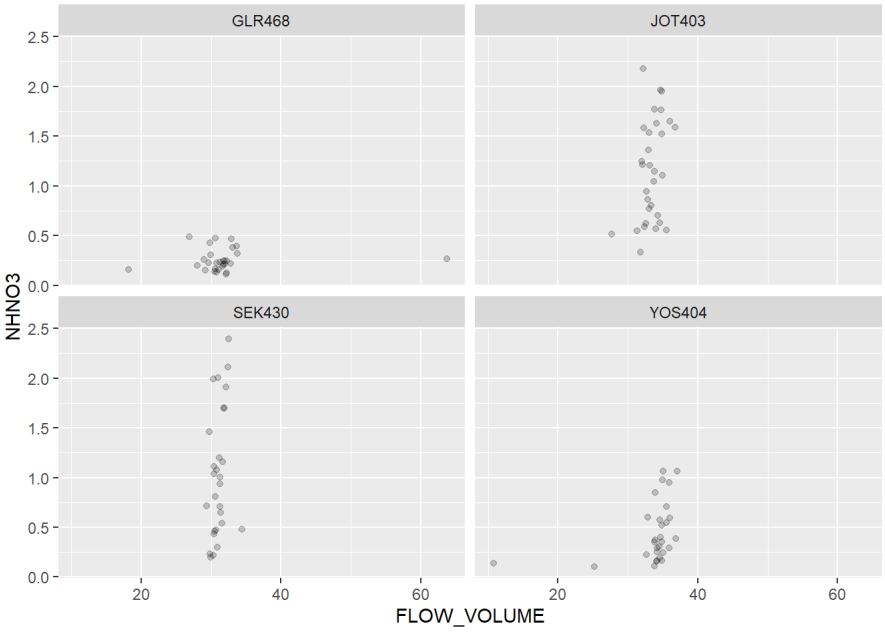
Did you create any new variables from existing variables in the dataset?

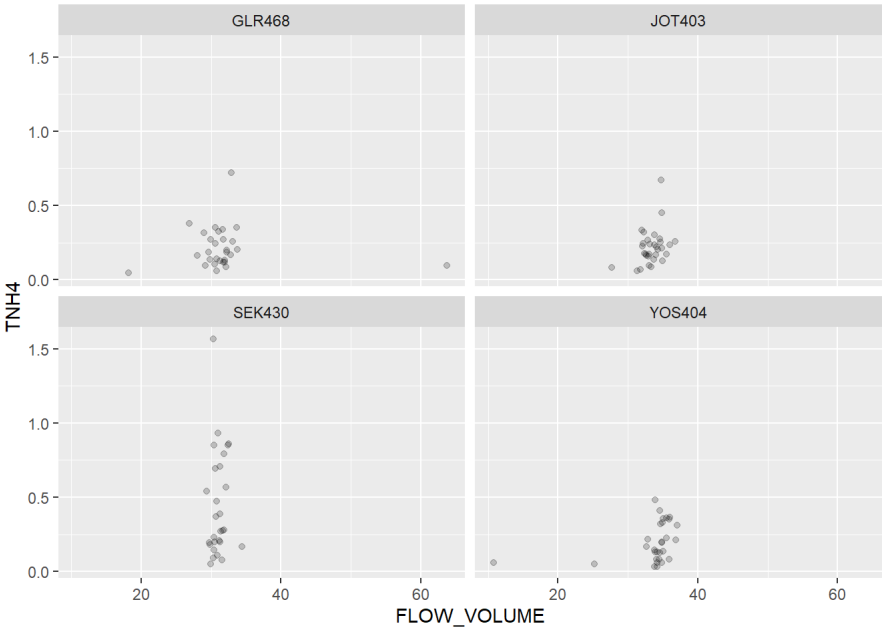
Of the features you investigated, were there any unusual distributions?

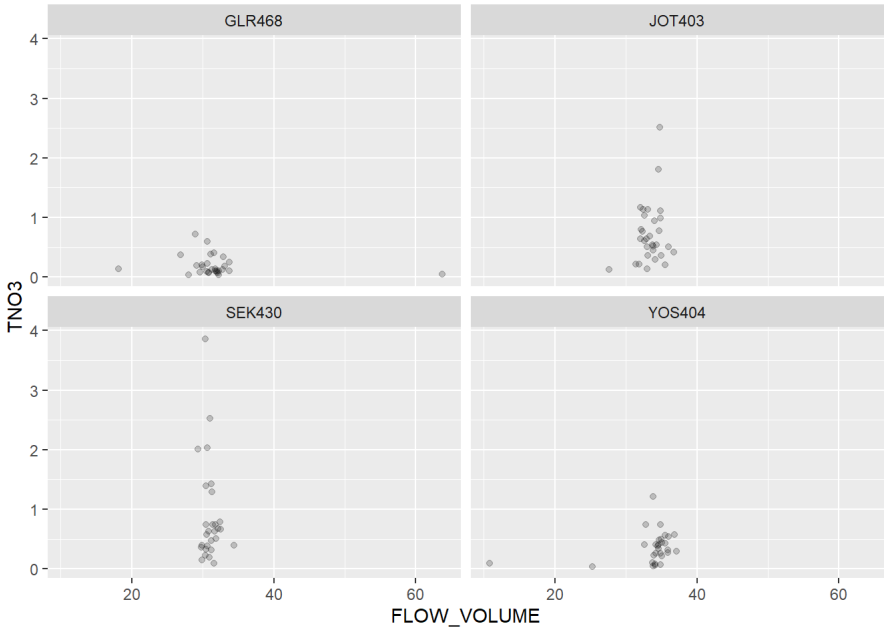
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this?

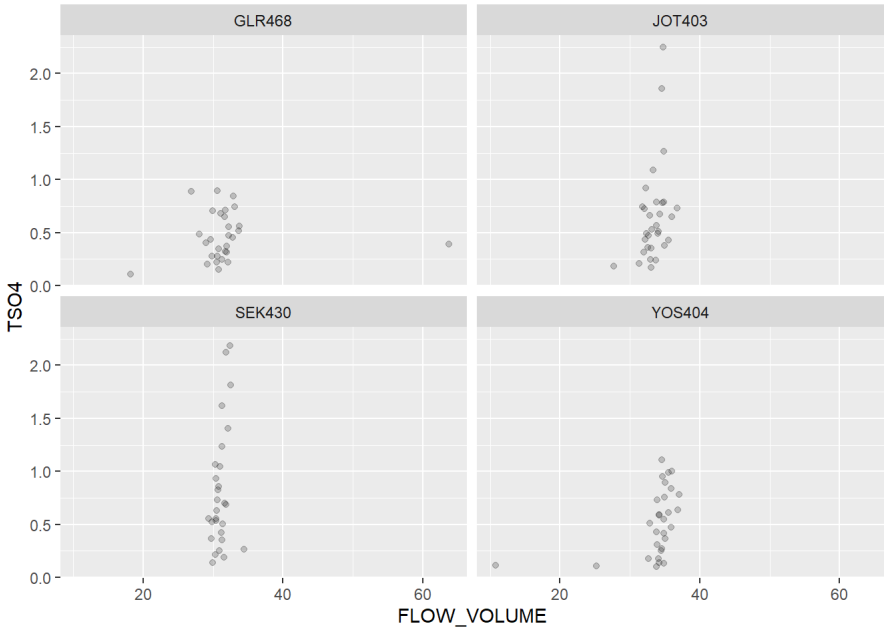
Bivariate Plots Section


GLR468
JOT403
SEK430
YOS404
WSO2
FLOW_VOLUME

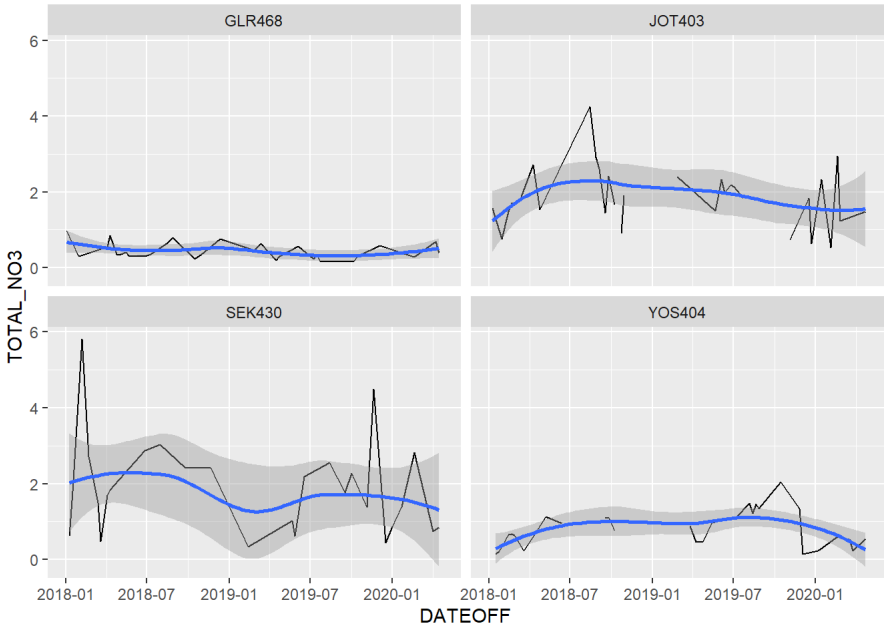

GLR468
JOT403
SEK430
YOS404
NHNO3
FLOW_VOLUME


GLR468
JOT403
SEK430
YOS404
TNH4
FLOW_VOLUME

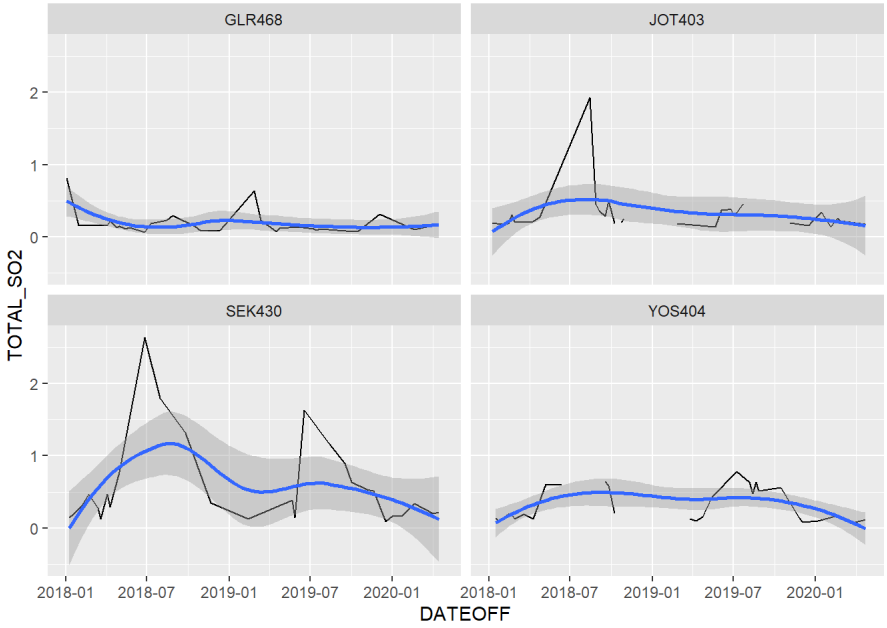




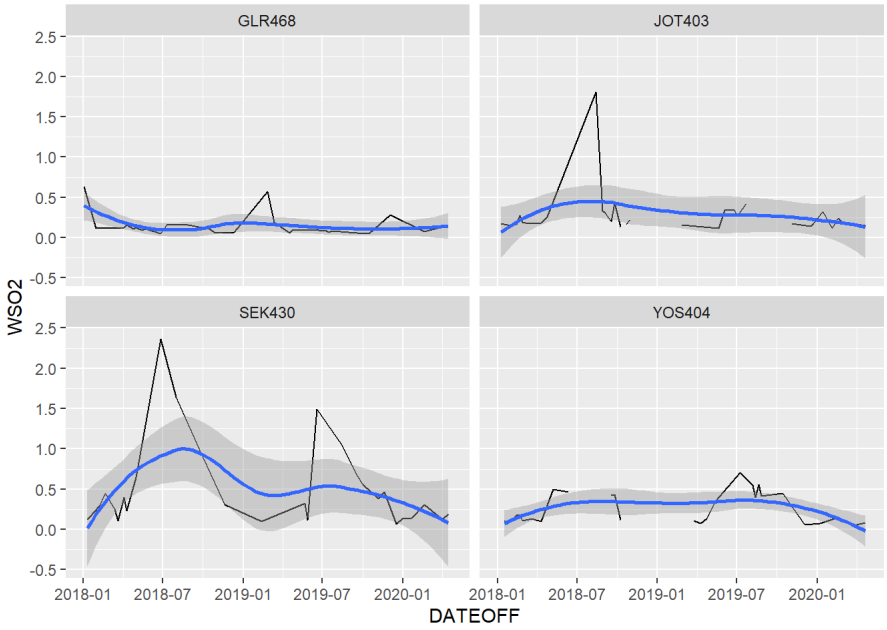

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

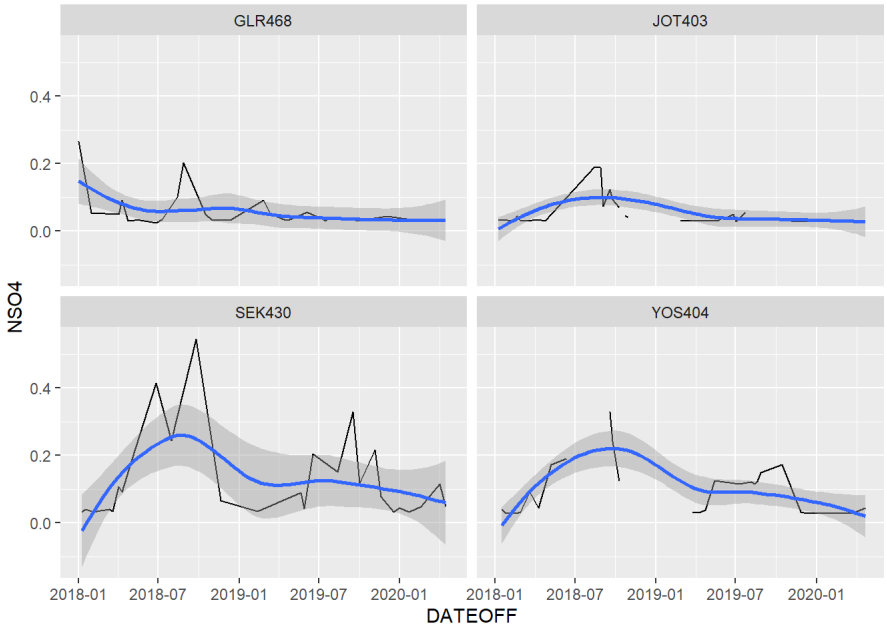



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

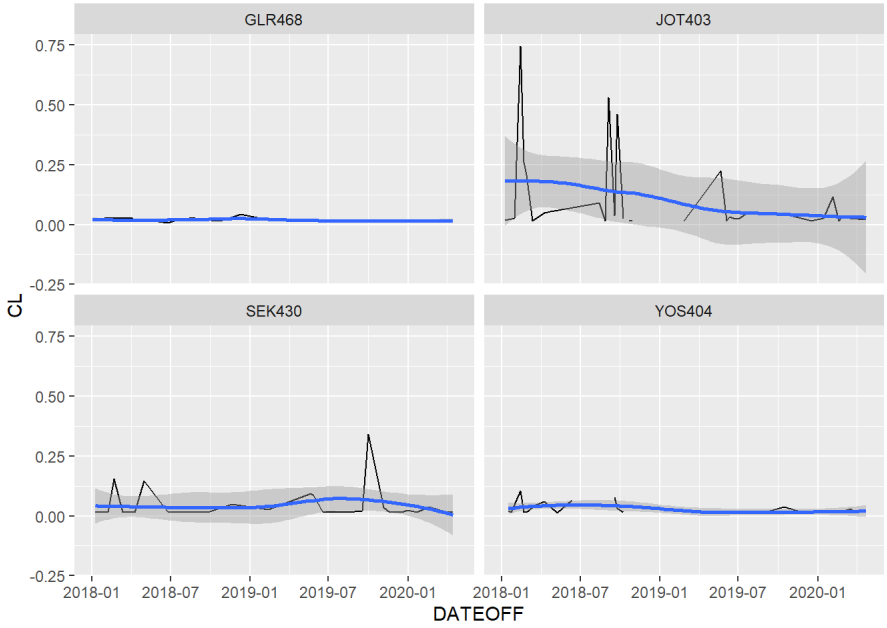



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

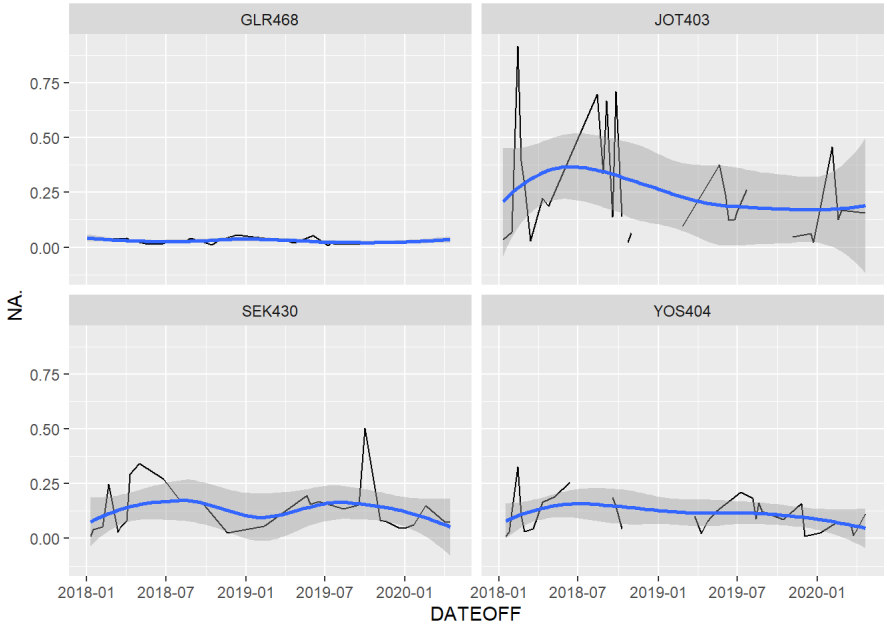



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

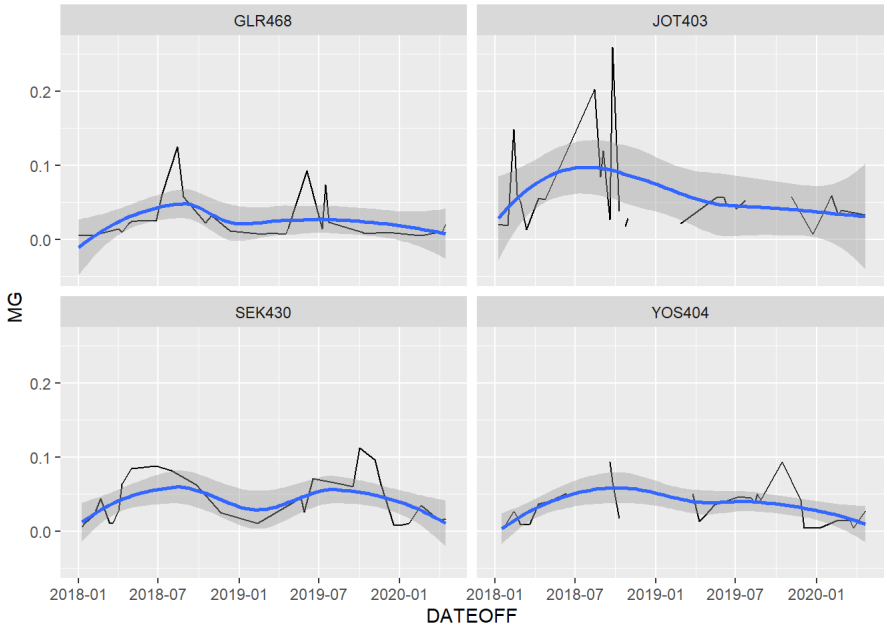



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

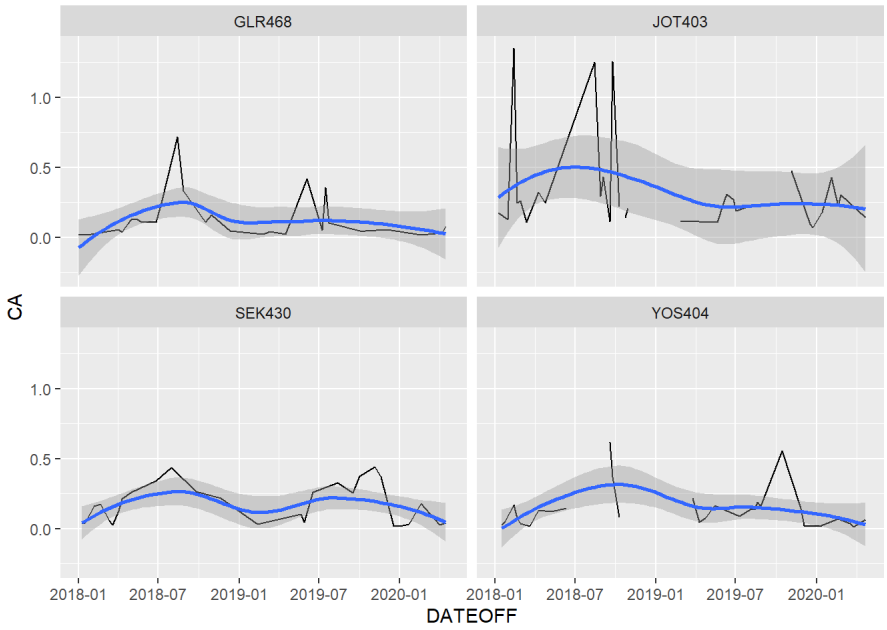

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



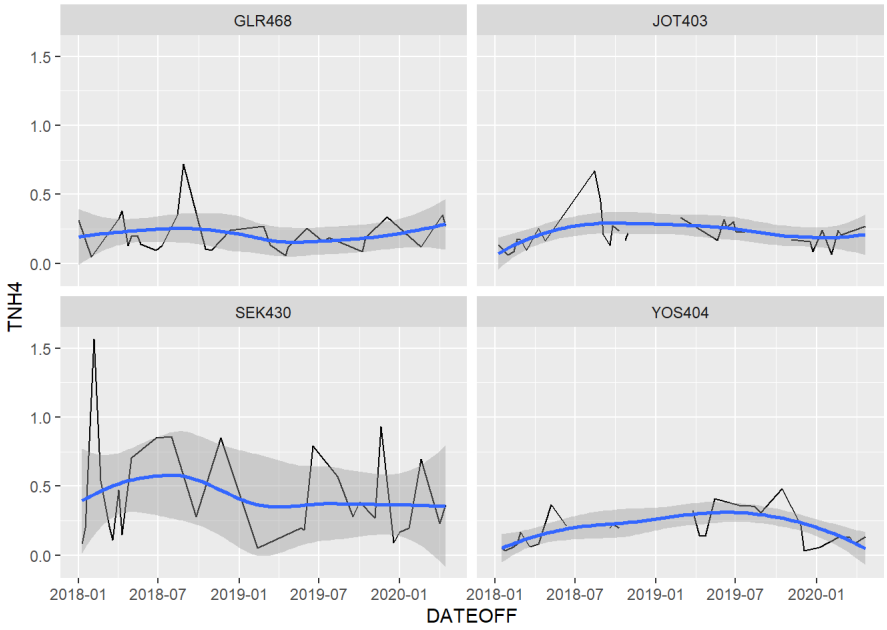

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



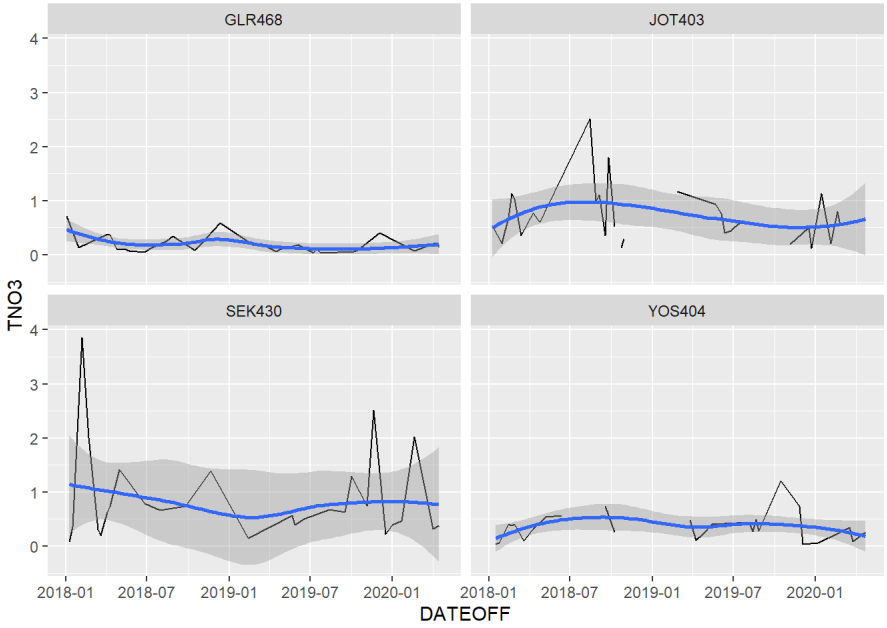

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



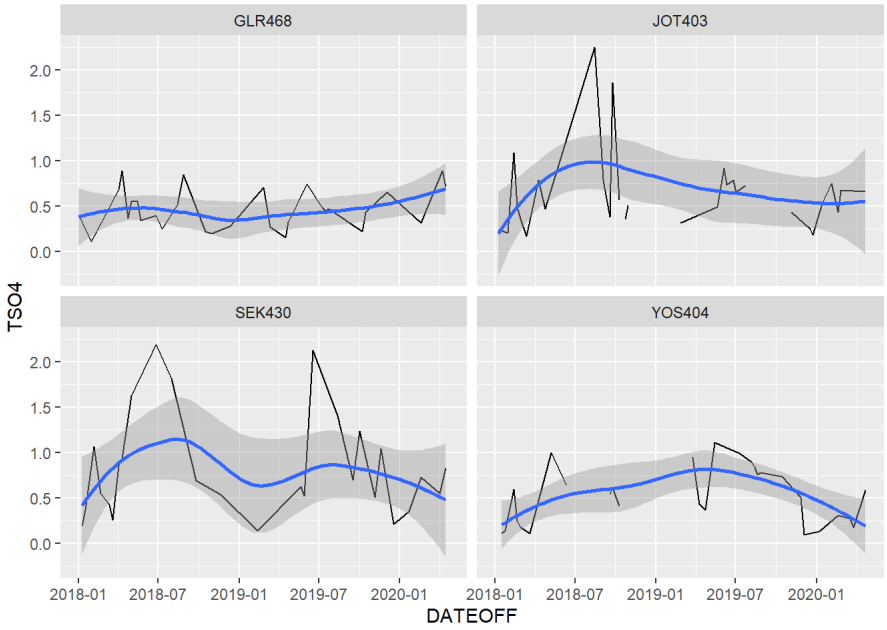

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



A correlation between the Flow Volume and Gaseous measurements does not exist.

```
##
##  Pearson's product-moment correlation
##
## data:  Target_data$FLOW_VOLUME and Target_data$Gaseous
## t = -0.29972, df = 115, p-value = 0.7649
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2084142  0.1543777
## sample estimates:
##         cor
## -0.02793822
```

A correlation between the Flow Volume and Particulate measurement does not exist.

```
##
##  Pearson's product-moment correlation
##
## data:  Target_data$FLOW_VOLUME and Target_data$Particulate
## t = 2.3884, df = 115, p-value = 0.01855
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03733425 0.38378123
## sample estimates:
##       cor
## 0.2173939
```

There appears to be a correlation between Particulate and Gaseous measurements.

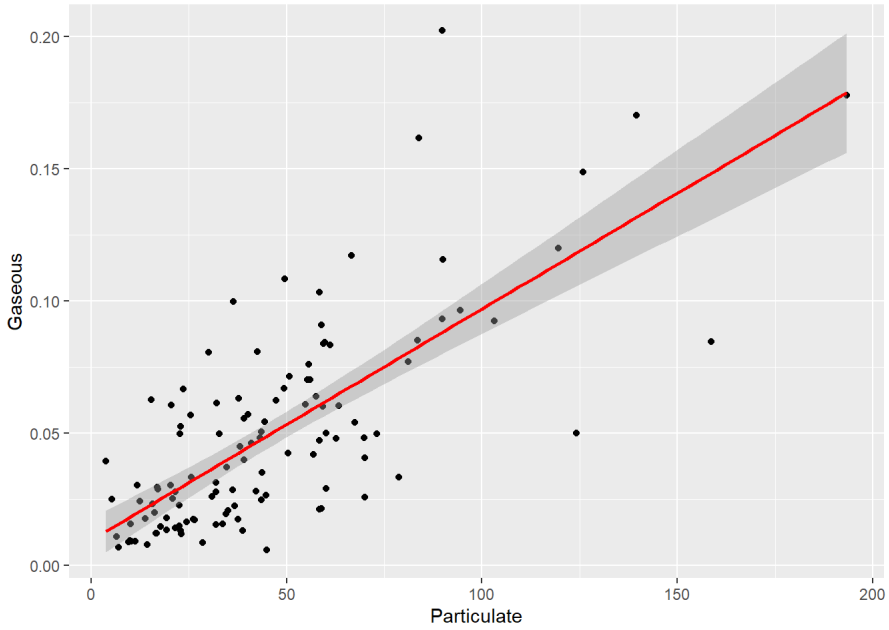
```
##
##  Pearson's product-moment correlation
##
## data:  Target_data$Particulate and Target_data$Gaseous
## t = 11.615, df = 115, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6383242 0.8084282
## sample estimates:
##       cor
## 0.7347201
```

```
ggplot(aes(x = Particulate, y= Gaseous), data = Target_data) +
  geom_point()+
  geom_smooth(method = 'lm', color = 'red')
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



Bivariate Analysis

Talk about some of the relationships you observed in this part of the

investigation. How did the feature(s) of interest vary with other features in
the dataset?

The majority of the data plots have most of their observations around the 30 m^3 air flow volume. Gaseous pollution (S02) is most prevalent in JOT403 and SEK430. SEK430 has significant concentration of pollutants in all of the graphs. JOT403 and SEK430 have the highest and second highest PPM measurements (worst and second worst air quality) within the United States. The graphs confirm that these locations receive a higher concentration of air pollutant.

Did you observe any interesting relationships between the other features

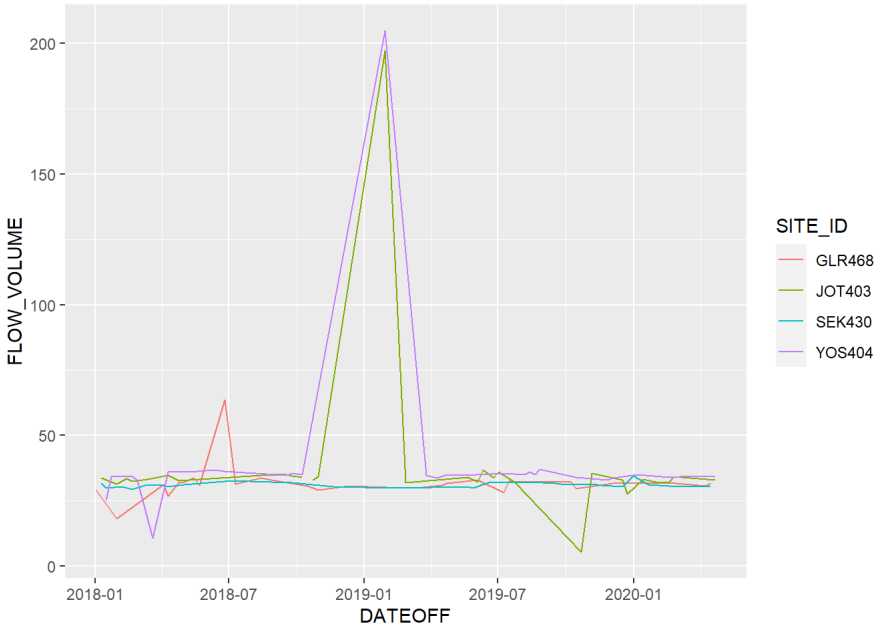
(not the main feature(s) of interest)?

SEK402 which averages 20 m^3 air flow volume consistently shows the highest concentration of PM2.2 and gaseous pollution. However, This inconsistency is not explored because the data for SEK402 ends in the 2005.

What was the strongest relationship you found?

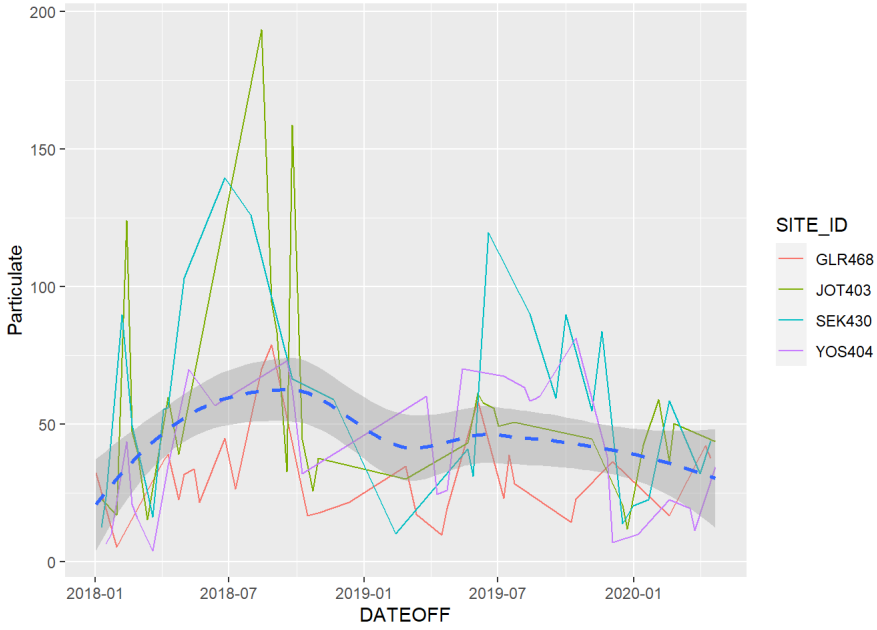
Since the 1990 Clean Air Act states have taken significant steps to reduce emissions of pollution. The graphs explored in this section confirm that states are continuing that effort.

Multivariate Plots Section



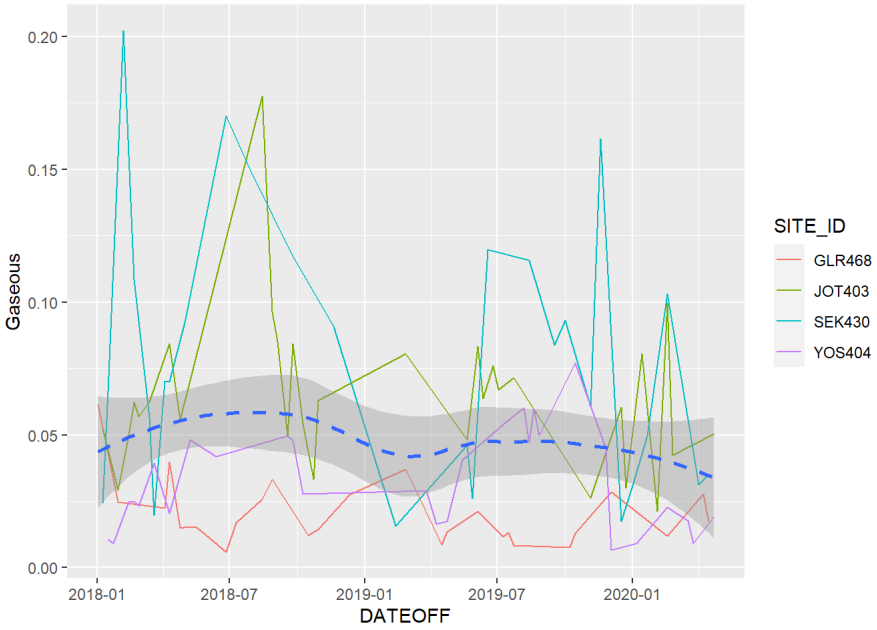

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   5.249  30.923  32.503  35.092  34.394 204.912       1
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   3.815  22.650  38.751  45.559  59.002 193.390       5
```

```
## No summary function supplied, defaulting to `mean_se()`
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## 0.005782 0.020479 0.040650 0.049277 0.063696 0.202223        5
```

Multivariate Analysis

Talk about some of the relationships you observed in this part of the

investigation. Were there features that strengthened each other in terms of
looking at your feature(s) of interest?

When looking at the data Glacier National Park, MT has significantly less gaseous and fine particulate matter than any of the sites located in California. The closest population center to Glacier Park is Columbia falls (population 6000) at 17.5 miles to the South West.

Meanwhile YOS404 168 miles from Sacramento, 193 miles from San Francisco. Joshua Tree (JOT403) is 131 miles from Los Angeles. SEK 430, Ash Mountain, is 60 miles from Fresno and 250 miles from San Francisco. These population centers all have over 500,000 individuals.

Were there any interesting or surprising interactions between features?

An interesting interaction of the features was the extreme increase in air flow volume, around January 2019, correlates with the dates of missing data for SEK430 and YOS404.

Additionally, the location where higher population centers are close by the level of pollutants are significantly higher.

Final Plots and Summary

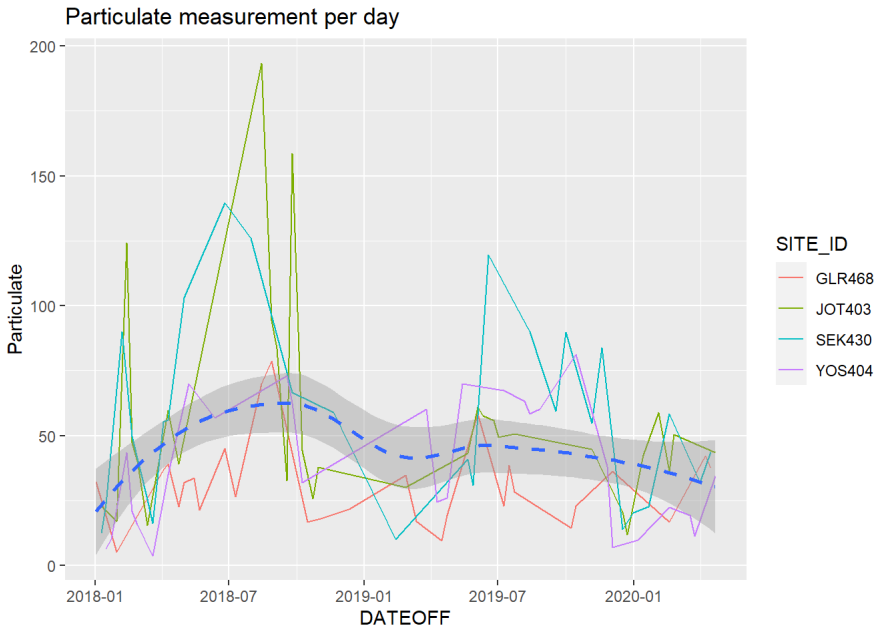
> **Tip**: You've done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings. Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

Plot One

```
## Warning: Removed 5 rows containing non-finite values (stat_summary).
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```



Description One

This Graph shows the mean of all four sites Particulate matter as a dashed blue line. Observing this graph you can see that GLR468's data points (orange) are consistently below this dashed line. Conversely the two locations with the worse Air Quality rating, JOT403 and SEK430, are significantly above the dashed blue line.

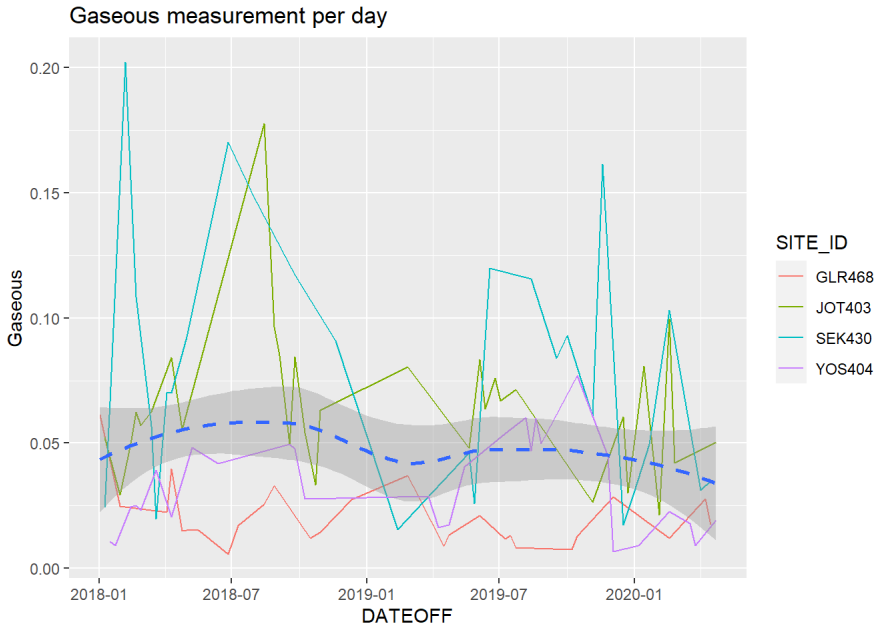
Plot Two

```
## Warning: Removed 5 rows containing non-finite values (stat_summary).
```

```
## No summary function supplied, defaulting to `mean_se()`
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

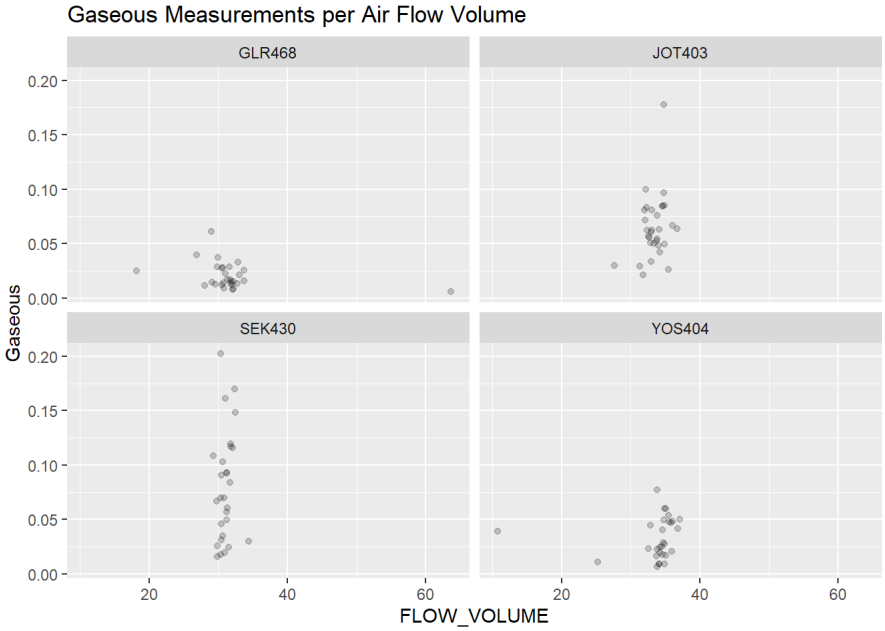
```
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```



Description Two

This Graph shows again the mean of all four sites Gaseous pollution as a dashed blue line. Observing this graph you can see that GLR468's data points (orange) are consistently below this dashed line. Conversely the two locations with the worse Air Quality rating, JOT403 and SEK430, are significantly above the dashed blue line. YOS404 the best rated Air Quality in California also has a significant amount of data points below the dashed blue line.

Plot Three



Description Three

Finally, this graph shows that no four locations averaged about the same ar Flow Volume (between 30-35 m^3). This eliminates the possibility that maybe the air flow volume increase could throw off the data. GLR468 maintains a low weight of gaseous material, a weight lower than JOT403 (site with worse Air Quality rating). SEK430's graph shows no consistency as the data points range from the top to the bottom of the graph.

Reflection

The contemporary thought is that anthropogenic climate warming is one of, if the not the greatest threat to the Earth. Although, this exploration of data does not discuss or delve into this question it is clear through this data that human kind greatly contributes to the Air Quality rating. Looking at GLR468, located in Glacier National Park in Montana we see what the potential Air Quality could be for the US. Located outside of the city of Los Angeles with a population of 4 million people, JOT403, has the highest concentration of all the PM25 and Gaseous pollution and the worst Air Quality rating.

The effects of these pollutants on the environment is not explored in this project either. Nor are the potential causes for these high concentrations. New York City, NY has a population of 8.4 million and the closest monitoring location has a rating of 71, significantly lower than JOT403's 88. Factors that could affect the ratings were not discovered through data. HOwever, Trade Winds, Climatology, volcanic eruptions, geysers, and other environmental factors could explain the discrepencies seen in the Air Quality ratings.

It is highly likely that human-Kind is the cause for the degredation of Air Quality ratings. Likely fixes run the spectrum but states are clearly adhering to reducing the total emission of pollutants as each site in this study shows a decrease from the original year 2000 data points.

Sources: https://www.epa.gov/cas (https://www.epa.gov/cas) https://www.epa.gov/castnet/castnet-ozone-monitoring (https://www.epa.gov/castnet/castnet-ozone-monitoring) https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics# (https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#):~:text=Ozone%20can%20be%20good%20or,on%20where%20it%20is%20found.&text=Learn%20more%20about%20stratospheric%2C%20or,more%20at tnet/castnet-ozone-monitoring https://stackoverflow.com/ (https://stackoverflow.com/) https://catalog.data.gov/dataset (https://catalog.data.gov/dataset) https://www.epa.gov/aqs (https://www.epa.gov/aqs)