p01 Discovery and Data Prep

Kaleb Crisci

October 11, 2019

Introduction

Perhaps our most important duty as a species is to protect the Earth. Our very existence depends on it. There are many things that contribute to harming the planet. The amount of pollution released into the atmosphere is one of those things. I chose this topic because I hope to get an idea about how much of the harmful molecules are being released now, as opposed to several years ago, and determine how this has changed over time.

Source Analysis

To perform this analysis, I will be using a dataset from the **United States Environmental Protection Agency on Clean Air Status and Trends Network (CASTNET)**. Because this data was retrieved from a Federal Agency, I consider it to be a valid source for information. In addition, the data has full and complete measurements of harmful particulates from several locations all over California, which can provide an accurate assessment of overall air quality in California.
Some limitations that these data may present is that some locations provide many more observations than others, which could lead to data clustering around one particular station. In addition, there are some observations that are dated and provide a location site, but no data was recorded.

About the Dataset

The dataset provides measurements of several different molecules in the air, at different locations all over California. The measurements were gathered consistently for seven days, and the average of each masurement is used for each observation.

Elements measured:

1. Sulfur Dioxide (SO2)
2. Sulfate (S04)
3. Nitrates (N03)
4. Nitric Acid (HNO3)
5. Magnesium (Mg)
6. Chlorine (Cl)
7. Total Nitrate (TNO3)
8. Ammonium (NH4)
9. Calium (Ca)
10. Sodium (Na)
11. Potassium (K)
    These elements are originally treated as variables in the data, but can easily be represented as factors, or categories of data.

Other Variables

The other variables in the data are the **SITE_ID**, which gives the location that the air was being tested, **YEAR**, which is the year the sample was taken, **WEEK**, which represents the week number in the year (1 to 51), **DATEON**, which is the date the observation started, and **DATEOFF**, which is the date the observation ended (7 day increments). Site Id will be treated as a categorical variable, and the rest of the variables will be considered continuous variables. **It is important to note that all observations started at 9AM and ended at 8AM 7 days later.**

Data Manipulation

To begin analysis we need to read in the data. To do this, I'll load the "**tidyverse**" library and store the database into a variable called "data."

```
library(tidyverse)
data <- read_csv("./Concentration - Weekly.csv")
```

The data contained some entries where the station did not observe anything So we need to filter out this missing data. To use piping (%>%), load the **dplyr** library

```
library(dplyr)
data <- data %>%
  na.omit
```

Next, we want to clean up the names of the columns that represent the harmful molecules and elements.

```
colnames(data)[colnames(data)=="SO2_CONC"] <- "SO2" #Sulfur Dioxide
colnames(data)[colnames(data)=="SO4_CONC"] <- "SO4" #Sulfate
colnames(data)[colnames(data)=="NO3_CONC"] <- "NO3" #Nitrates
colnames(data)[colnames(data)=="HNO3_CONC"] <- "HNO3" #Nitric Acid
colnames(data)[colnames(data)=="TNO3_CONC"] <- "TNO3" #Total Nitrate
colnames(data)[colnames(data)=="NH4_CONC"] <- "NH4" #Ammonium
colnames(data)[colnames(data)=="CA_CONC"] <- "Ca" #Calcium
colnames(data)[colnames(data)=="NA_CONC"] <- "Na" #Soduim
colnames(data)[colnames(data)=="MG_CONC"] <- "Mg" #Magnesium
colnames(data)[colnames(data)=="K_CONC"] <- "K" #Potassioum
colnames(data)[colnames(data)=="CL_CONC"] <- "Cl" #Chlorine
colnames(data)
```

```
##  [1] "SITE_ID" "YEAR"    "WEEK"    "DATEON"  "DATEOFF" "SO2"     "SO4"
##  [8] "NO3"     "HNO3"    "TNO3"    "NH4"     "Ca"      "Na"      "Mg"
## [15] "K"       "Cl"
```

Then, we want to put this data into a format that makes it easier to observe the changes of each molecule over time, individually. This will require creating a new column called Molucule, and making each of the columns that name a molecule an occurrence of the new column “Molecule”. We will also need to add another column called “Concentration” to store the observed data for that molecule.

```
data <- gather(data, key = "Molecule", value = "Concentration", "SO2":"Cl", convert = FA
LSE, factor_key = TRUE)
head(data)
```

```
## # A tibble: 6 x 7
##   SITE_ID  YEAR  WEEK DATEON           DATEOFF        Molecule Concentration
##   <chr>   <dbl> <dbl> <chr>            <chr>          <fct>            <dbl>
## 1 DEV412   2007    12 03/20/2007 09:… 03/27/2007 08… SO2              0.292
## 2 DEV412   2006    13 03/28/2006 09:… 04/04/2006 08… SO2              0.268
## 3 DEV412   2005    13 03/29/2005 09:… 04/05/2005 08… SO2              0.411
## 4 DEV412   2005    11 03/15/2005 09:… 03/22/2005 08… SO2              0.355
## 5 DEV412   2007    11 03/13/2007 09:… 03/20/2007 08… SO2              0.233
## 6 DEV412   2006    12 03/21/2006 09:… 03/28/2006 08… SO2              0.129
```

Next, we change the siteId column to a category (factor), since there are only a few of them. We can then see the different categories that are present using the levels() function

```
data$SITE_ID <- as.factor(data$SITE_ID)
levels(data$SITE_ID)
```

```
## [1] "CON186" "DEV412" "JOT403" "LAV410" "PIN414" "SEK402" "SEK430" "YOS404"
```

We also want to create a new column to give us more information on the location. Using the SITE_ID variable, I was able to find the county information for each station.

```
index <- c("CON186", "DEV412", "JOT403", "LAV410", "PIN414", "SEK402", "SEK430", "YOS40
4")
values <- c("San Bernardino", "Inyo", "San Bernadino", "Shasta", "San Benito", "Tulare",
"Tulare", "Mariposa")

#create new column
data$County <- values[match(data$SITE_ID, index)]
```

The date columns are currently set as characters. We need to parse them into actual dates.

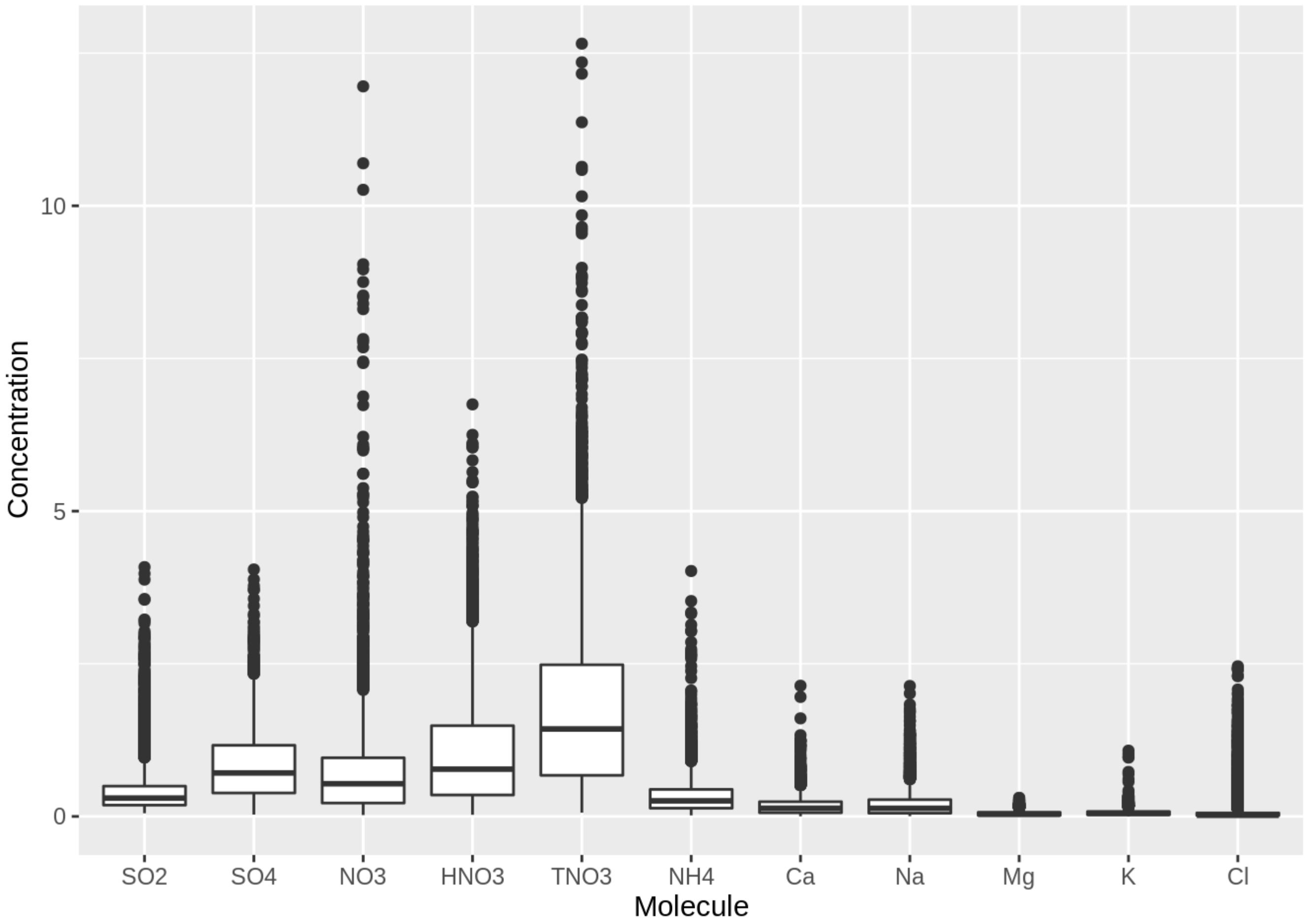
```
data$DATEON <- as.POSIXlt(parse_date(data$DATEON, format="%m/%d/%Y %H:%M:%S"))

data$DATEOFF <- as.POSIXlt(parse_date(data$DATEOFF, format="%m/%d/%Y %H:%M:%S"))
```

Finally, we have a table of tidy data that is easy to analyse. Now, we can get into visualization.

Visualizations

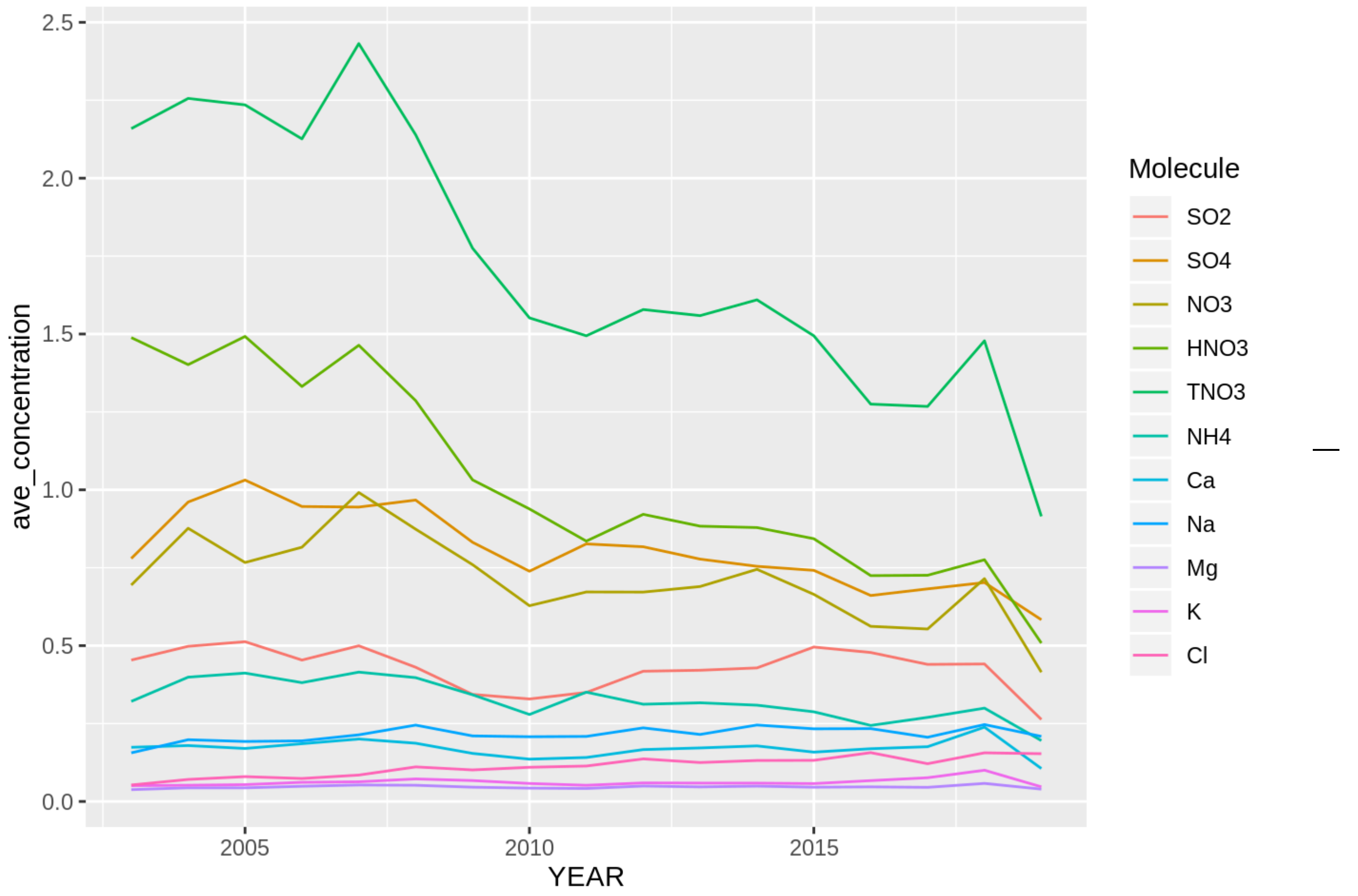
Here, we create a boxplot which shows the concentration of each of the harmful molecules. By observing the data this way, we can see the observed concentration for each molecule, where most of the data is concentrated, as well as any outliers. For example, we can see that TNO3 is generally measured in higher concentration than the others.

```
ggplot() + geom_boxplot(data, mapping=aes(x = Molecule, y = Concentration))
```



Another way to view this data is by observing how each molecule has increased or decreased over time. To do this, we have to set up a group and use an aggregate function to summarize the data so it can be accurately plotted. This graph shows that as time progresses, the average air concentration of these harmful elements actually decreases.

```
years <- group_by(data, YEAR, Molecule)
summary <- summarize(years, ave_concentration = mean(Concentration))
ggplot() + geom_line(summary, mapping=aes(x = YEAR, y = ave_concentration, colour = Mole
cule))
```



Research Questions

This dataset provides a few different options as to where to focus the research:

1. What are the potential causes for the decrease in harmful air particles?
2. Why do some counties have higher concentrations of certain harmful molecules than others?
3. How has the incorporation of clean air acts in California affected the changes in air quality?
4. What is the relationship between human population in a given area and the levels of each of the harmful molecules in the air?

This study can help to identify which areas in California are doing a good job at preventing air pollution, and which areas need to do more to prevent it. It can also help to provide a measure of how effective our current pollution prevention methods have been over the last decade. Also, by comparing the population of people in a particular area to the measured levels of each of the compounds in the air, we can determine which compounds are directly releated to human presence, and which are not.