Homework #4

Xinyuan Cao (xc2461)

Nov 14 2018

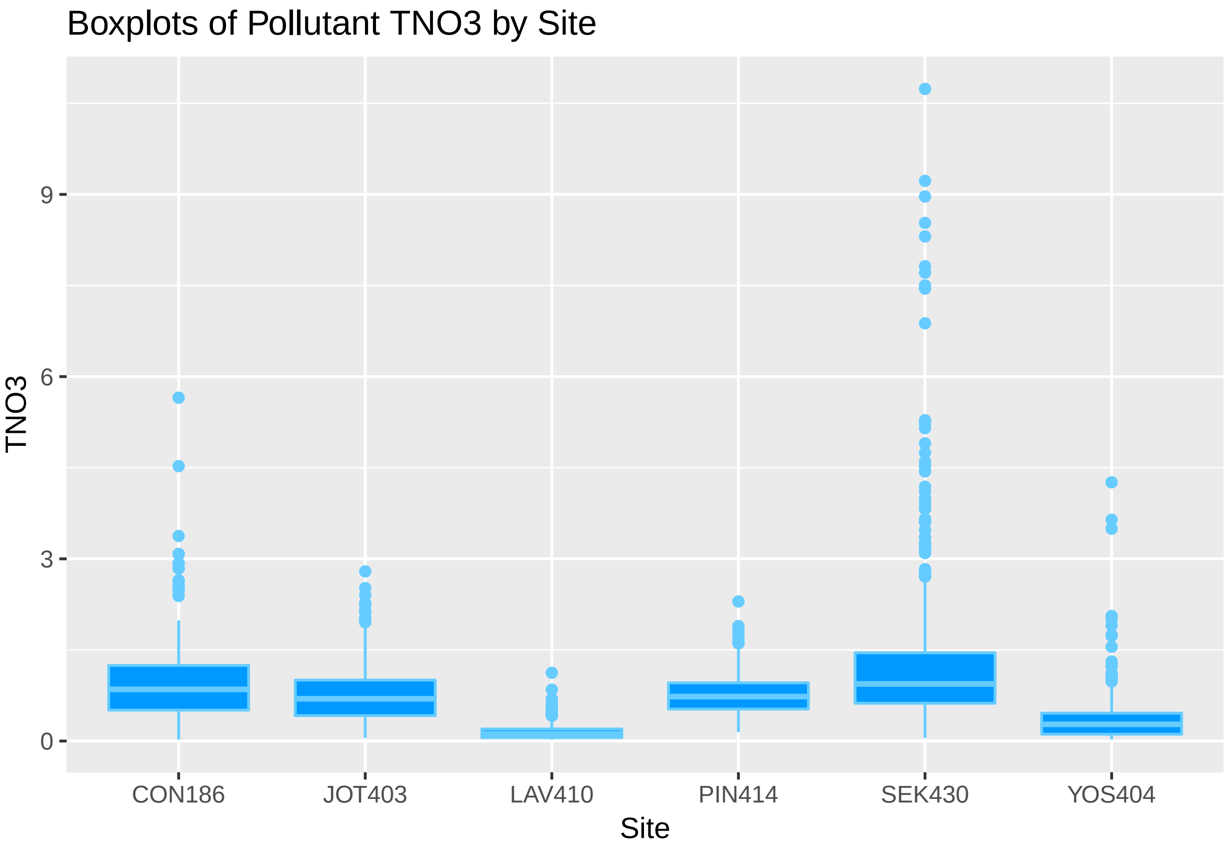
3. Plot in the final pj

Begin the analysis of one variable in the dataset you are using the final project. As this is an individual homework assignment, each group member should choose a different variable. Choose three visualizations as appropriate to show the distribution of the variable, conditioned on another variable if desired (for example, the distribution of income by region). Write a few sentences describing what you found and what new questions your visualizations have generated. (Faceted graphs count as one graph; graphs put together with grid.arrange() or similar count as multiple graphs.)

- The data we use is the air pollution data in California. The variable I choose is the pollutant TNO3. I'll draw a boxplot, a histogram and a time series line chart.

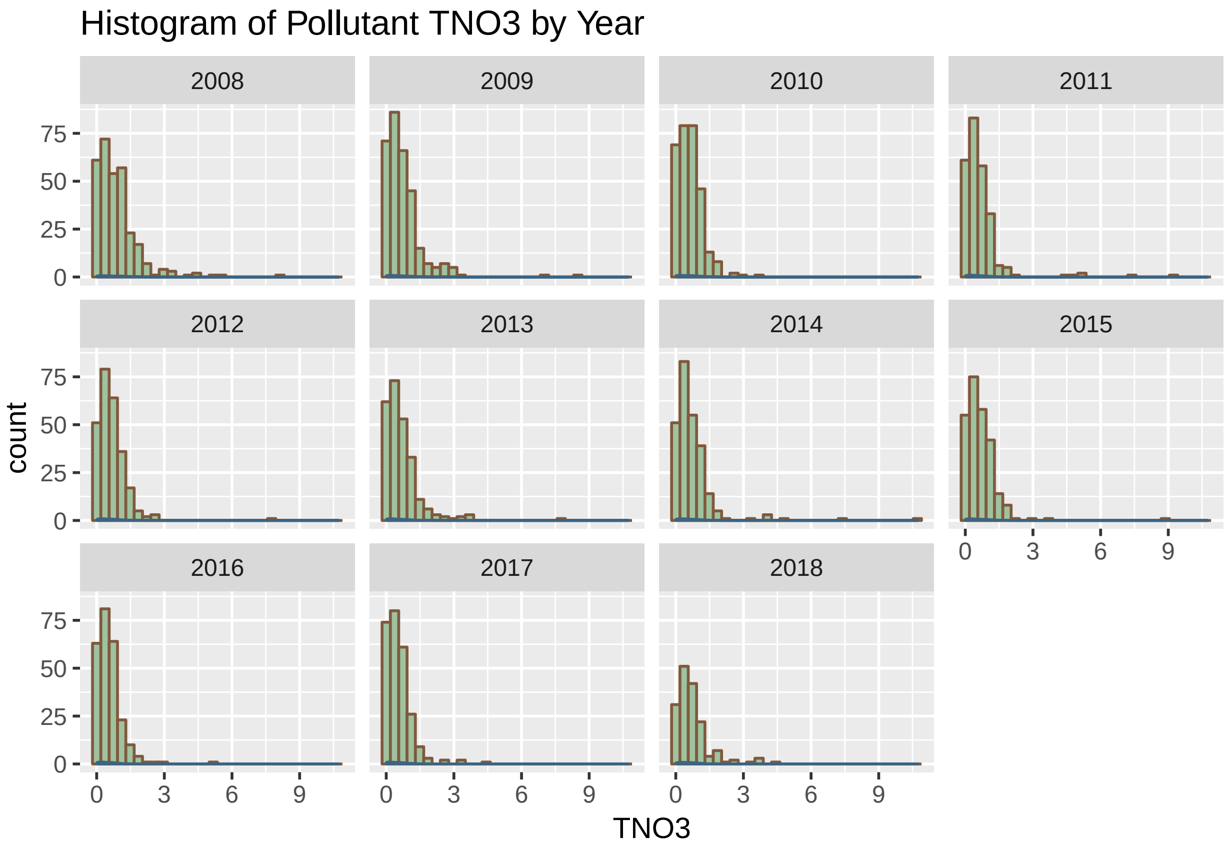
```r
library(tidyverse)
library(ggplot2)
library(lubridate)
pollution <- read.csv("california_pollution.csv")
site <- factor(pollution$SITE_ID)
date.on <- mdy_hms(pollution$DATEON)
year <- factor(year(date.on))

# boxplot
p1 <- ggplot(pollution, aes(x = site, y = pollution$TNO3)) +
  geom_boxplot(fill = "#0099FF", color = "#66CCFF") +
  ggtitle("Boxplots of Pollutant TNO3 by Site") +
  labs(x = "Site", y = "TNO3")
p1
```



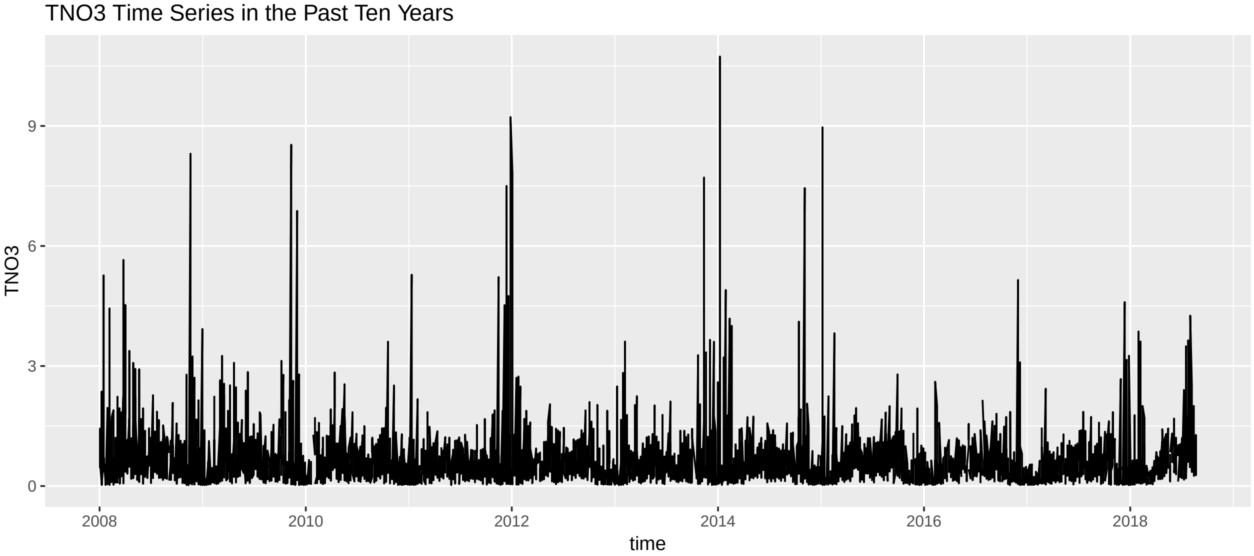
- I draw the boxplot by site. And it can be seen that the pollutant TNO3 in LAV410 is the least while that in SEK430 is the most.
- The pollutant TNO3 in LAV410 is the densest, YOS404 ranks second, and PIN414 ranks third. SEK430 has the largest range, and most outliers.
- In most cases TNO3 is less than 1.5, but on some days it exceeds 1.5.
- So we can make investigation further concentrating on SEK430, and find why TNO3 ranges so big and the pollutant amount is so large. We also should find out in the next step what factors (maybe climate, terrain, factory) affect the distribution of TNO3.

```r
# draw the histogram
p2 <- ggplot(pollution, aes(x = pollution$TNO3)) +
  geom_histogram(colour = "#80593D", fill = "#9FC29F") +
  geom_density(color = "#3D6480") +
  facet_wrap(year) +
  ggtitle("Histogram of Pollutant TNO3 by Year ") +
  labs(x = "TNO3")
p2
```



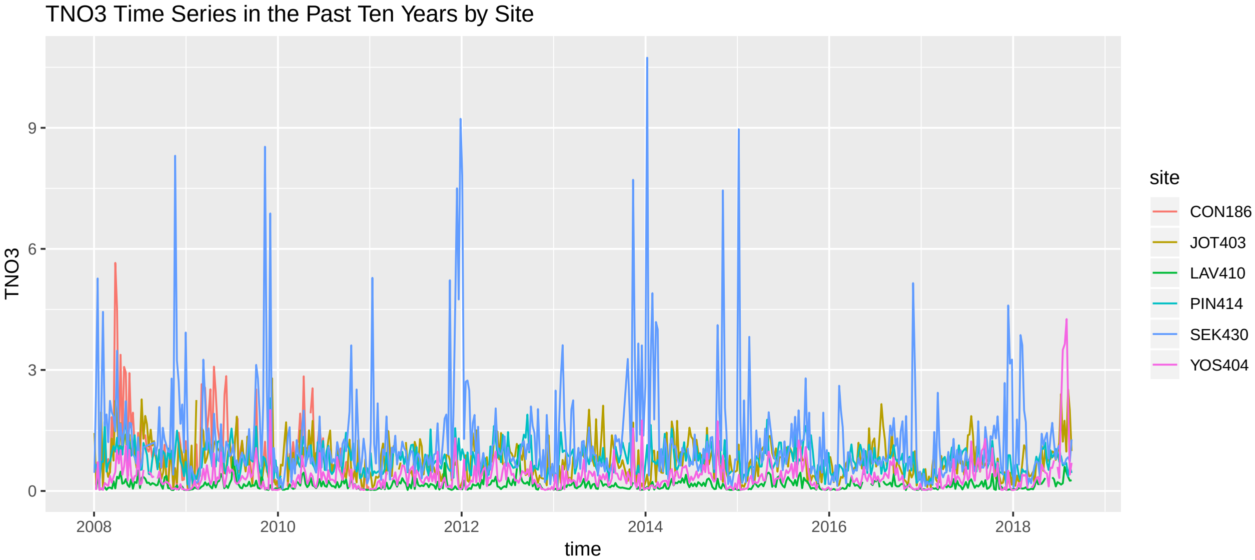
- I draw the histogram on the amount of pollutant TNO3 year by year. It is shown that in every year, the data is right skewed; most of the TNO3 is less than 1.5; and the peak is from 0.3 to 0.6.
- Also I can find that in 2018, the data amount is obviously smaller than others. That is because our data is updated on August, so we only get eight months of data in 2018. This makes sense.
- In the past ten years the distribution of TNO3 is almost the same, which shows that the source of this pollutant almost have not changed in the past. We can further explored the source and make a reasonable interpretation.

```
# draw the time series line chart
p3a <- ggplot(pollution, aes(x = date.on, y = pollution$TNO3)) +
  geom_line() +
  labs(x = "time", y = "TNO3") +
  ggtitle("TNO3 Time Series in the Past Ten Years")
p3a
```



- This is the time series of TNO3 over the past ten years. It can be seen that fluctuation is really big, and there are several days that has very high TNO3. We can investigate to find out the reasons behind it.

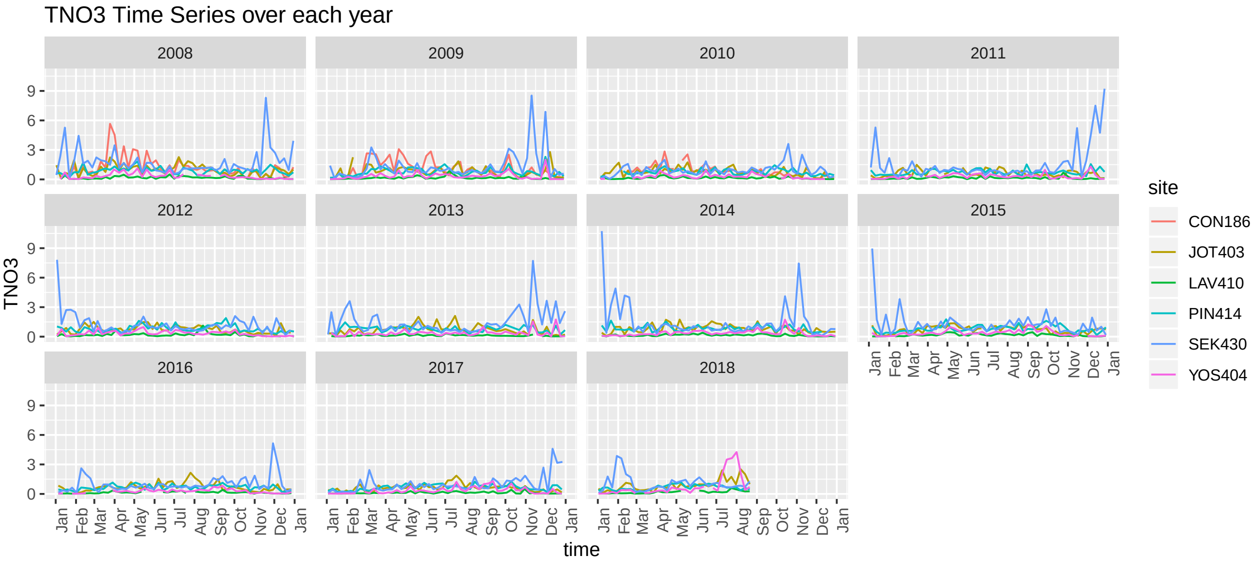
```r
# draw the time series line chart by site
p3b <- ggplot(pollution, aes(x = date.on, y = pollution$TNO3)) +
  geom_line(aes(color = site)) +
  labs(x = "time", y = "TNO3") +
  ggtitle("TNO3 Time Series in the Past Ten Years by Site")
p3b
```



- I also draw the time series by site, and we get the conclusion similar to the boxplot that SEK430 is highest; LAV410 is the lowest; and YOS404 is the second lowest.
- The high outliers are mostly from SEK430. But in 2008-2011 there are some high outliers in CON186, and in recent days there high ourliers in YOS404. We can look into these particular period of time in the particular sites further.

```r
# draw the time series line chart in the year
month.day.str <- format(as.Date(date.on), "%m%d")
month.day <- as.Date(month.day.str, tryFormats = "%m%d")
```

```
p3c <- ggplot(pollution, aes(x = month.day, y = pollution$TNO3)) +
  geom_line(aes(color=site)) +
  labs(x = "time", y = "TNO3") +
  facet_wrap(year) +
  scale_x_date(date_breaks = "months" , date_labels = "%b") +
  ggtitle("TNO3 Time Series over each year") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
p3c
```



- From this chart, I try to see the change of TNO3 in the whole year. I find that the outliers tend to be in the beginnig or end of the year from SEK430. In the middle of the year, the pollutant TNO3 is always very low. We may relate it to the seasons, maybe TNO3 is higher in winter. But it needs more evidence to proove it or to make other assumptions.
- We can also easily capture an abnormal high TNO3 in YOS404 this summer, we can dig into it further.
- Besides, we can also relate TNO3 to other pollutants to find whether there is relation, and is there any cause and effect among different pollutants in the next steps.