Project_DataVisulization

Yanxi Peng (Krista)

2022-11-14

Set up R session

Obtain basic summaries of the data

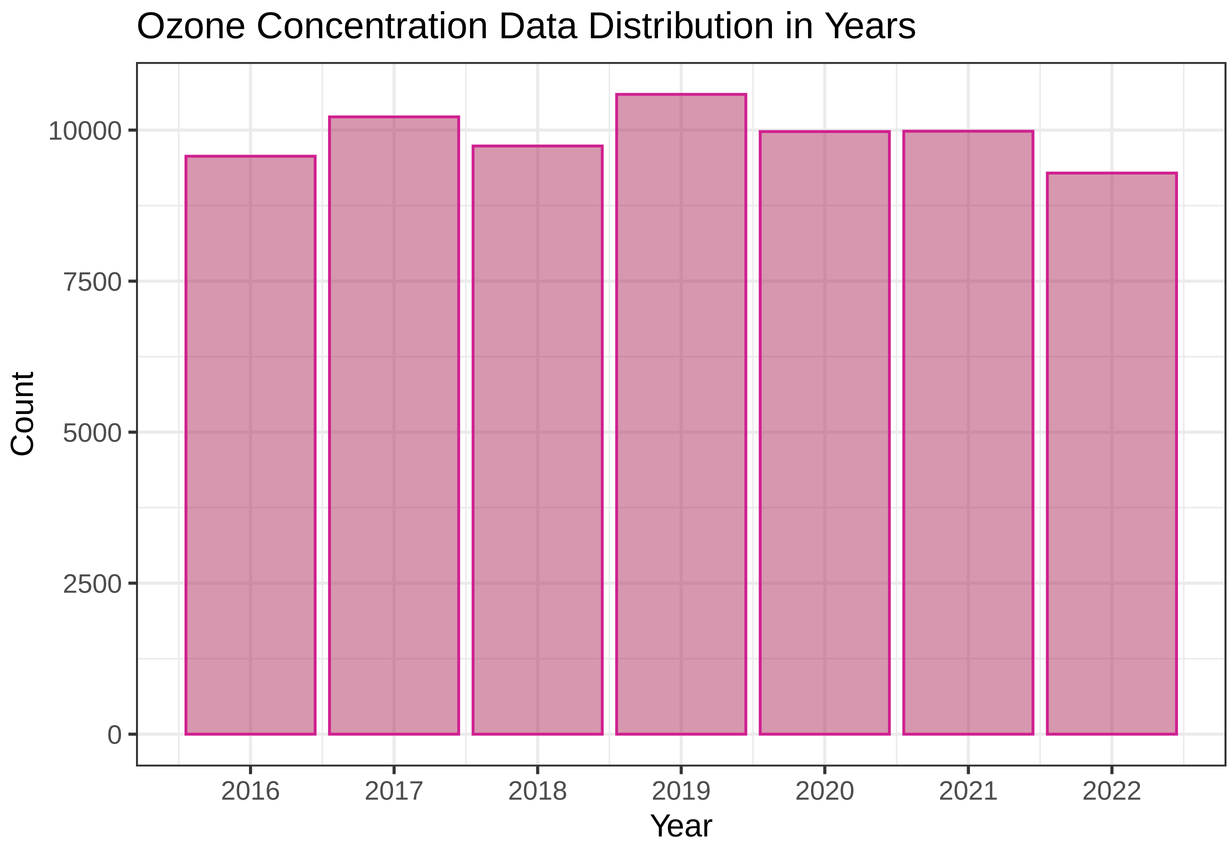
```
## [1] 69363    10
```

```
##        X                 Date        Concentration        DailyAQI     
##  Min.   :    1   10/26/2016:   40   Min.   :0.00000   Min.   :  0.00  
##  1st Qu.:17342   4/1/2018  :   40   1st Qu.:0.03400   1st Qu.: 31.00  
##  Median :34682   4/11/2016 :   40   Median :0.04200   Median : 39.00  
##  Mean   :34682   4/12/2018 :   40   Mean   :0.04163   Mean   : 39.51  
##  3rd Qu.:52023   4/13/2017 :   40   3rd Qu.:0.04900   3rd Qu.: 45.00  
##  Max.   :69363   4/13/2018 :   40   Max.   :0.08800   Max.   :156.00  
##                   (Other)   :69123                                      
##    SiteName                 County         Latitude       Longitude     
##  Length:69363       Haywood    : 5029   Min.   :34.36   Min.   :-83.80  
##  Class :character   Forsyth    : 5024   1st Qu.:35.26   1st Qu.:-82.05  
##  Mode  :character   Mecklenburg: 4234   Median :35.55   Median :-80.34  
##                     Avery      : 3793   Mean   :35.61   Mean   :-80.35  
##                     Cumberland : 3263   3rd Qu.:35.99   3rd Qu.:-78.77  
##                     Swain      : 3088   Max.   :36.31   Max.   :-76.62  
##                     (Other)    :44932                                   
##      Month             Year     
##  Min.   : 1.000   Min.   :2016  
##  1st Qu.: 4.000   1st Qu.:2017  
##  Median : 6.000   Median :2019  
##  Mean   : 6.448   Mean   :2019  
##  3rd Qu.: 9.000   3rd Qu.:2021  
##  Max.   :12.000   Max.   :2022  
## 
```

> This dataset has an overall 69363 observations and 9 variables, including collecting date, Ozone concentration, Daily AQI value, site name, county name, latitude, longitude, month, and year. From the summary, we can conclude that more data are collected in the Garinger high school, Rockwell, and Millbrook schools than in other sites, and we can also use ggplot to study the data distribution in different sites and years and determine the best site name and year that can be used for analysis.

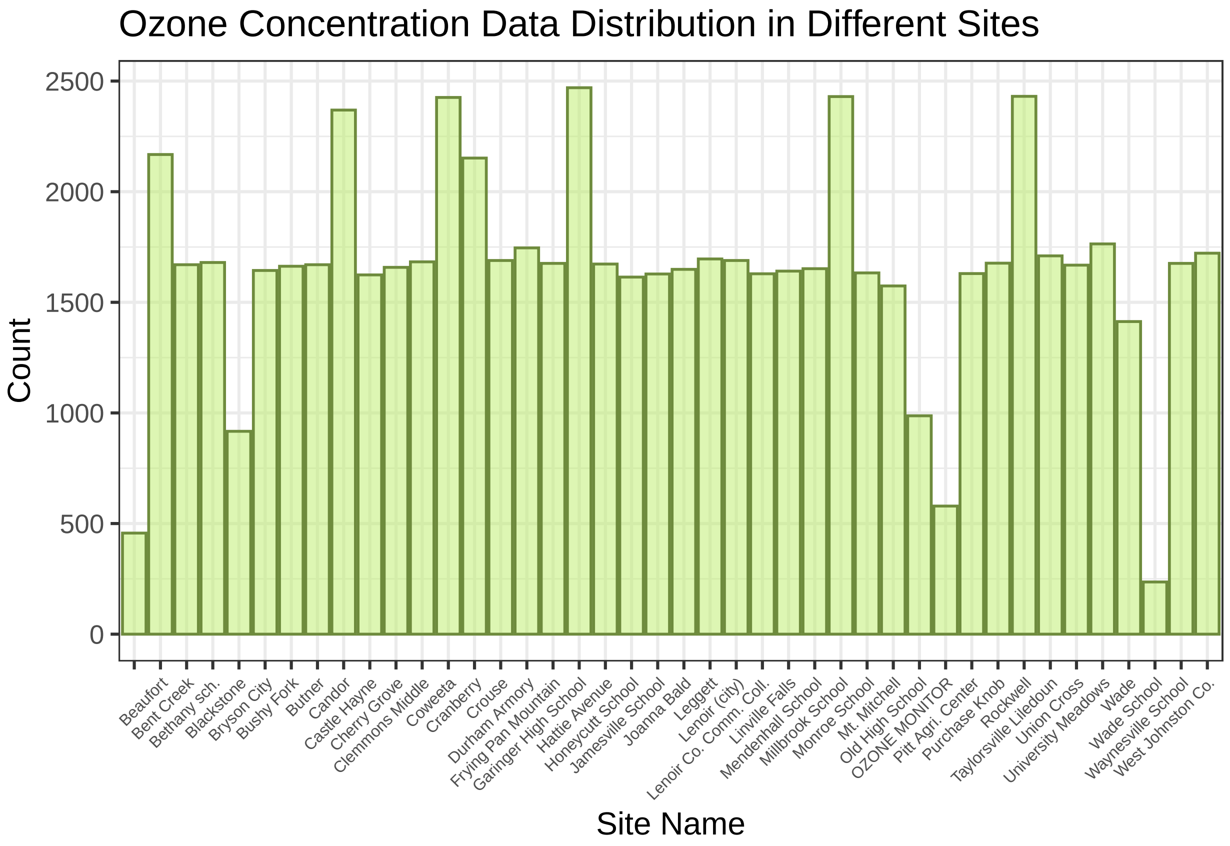
Explore data graphically

Generate a plot of the number of ozone concentrations collected conducted by year



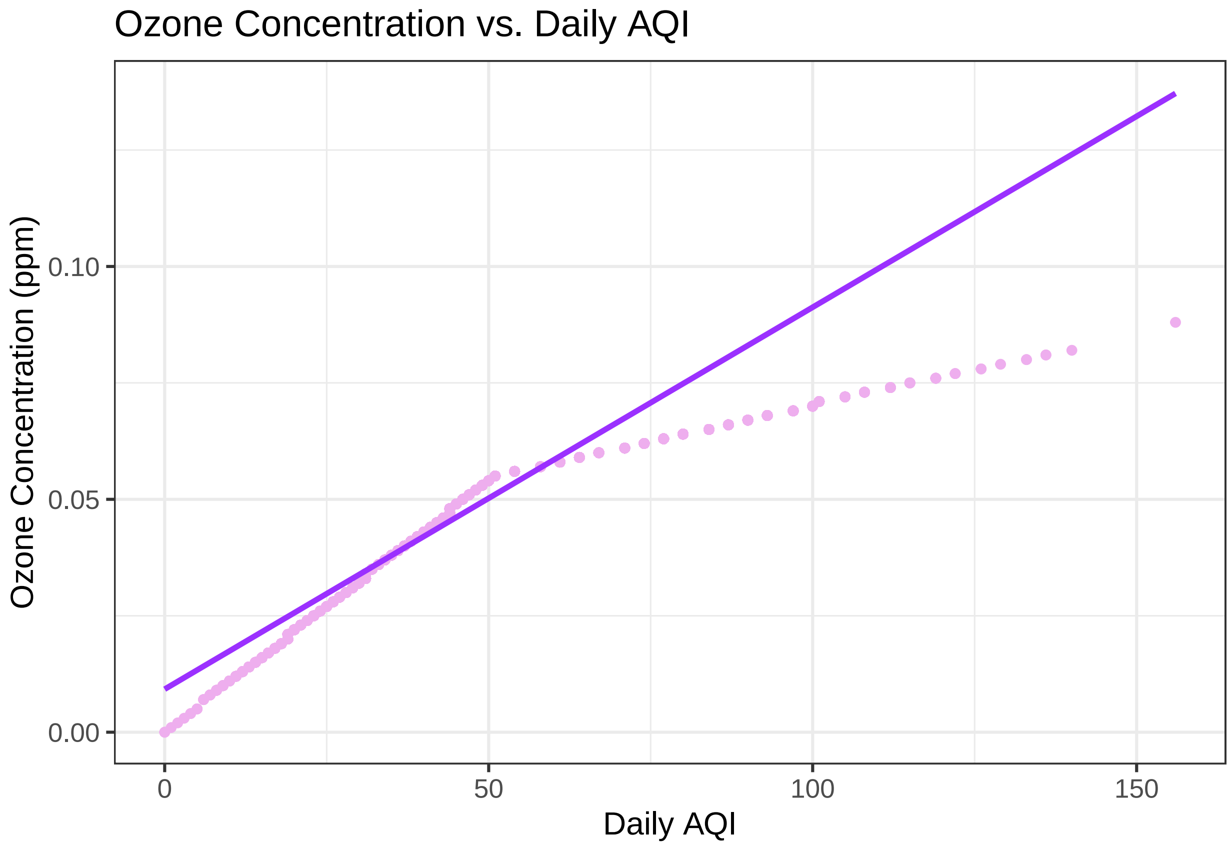
From the bar plot, the number of ozone concentration data collected in the different years is different. For further research, we can either use the newest from the year 2022 or the year that has the most count, which is the year 2019 from this plot.

Generate a plot of the number of ozone concentrations collected conducted by site



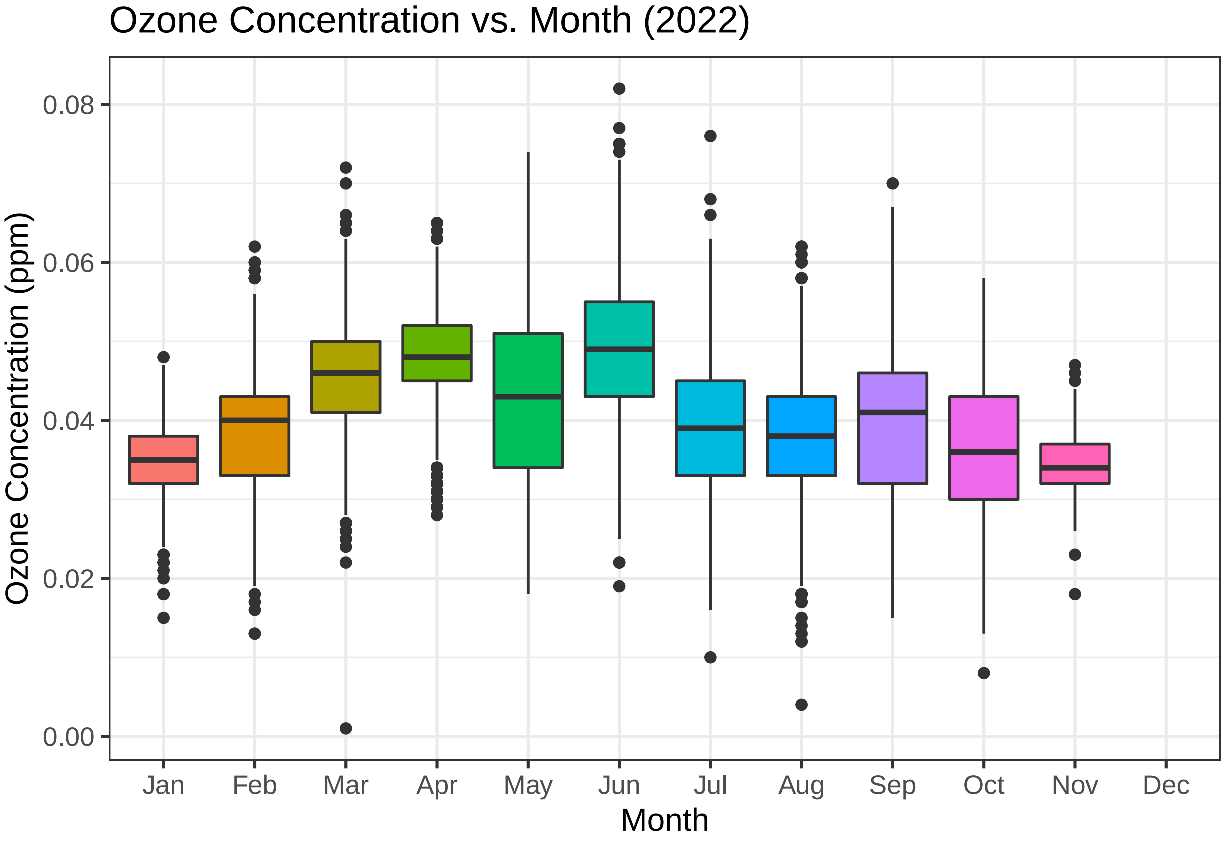
From the bar plot, the number of data samples collected at different sites varied considerably. For further analysis in this project, to ensure the accuracy of the results, we should choose the site with more data extraction as the analysis object. In this case, Garinger high school, Rockwell, Millbrook School, and Coweeta can be considered to use in the data analysis part.

Is there any relation between Daily AQI and O3 Concentration?



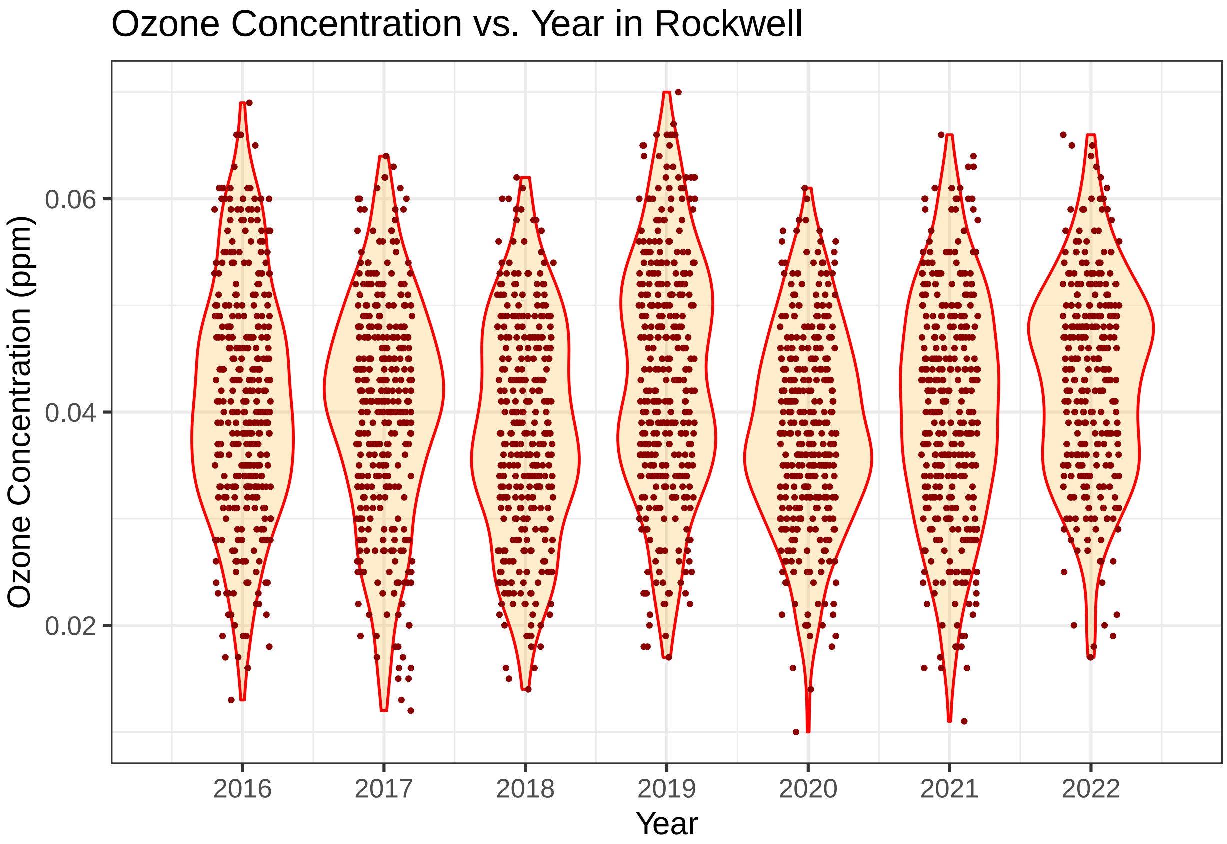
From the plot, we do observe a linear relation between Ozone concentration and daily AQI. Hence, in the data analysis part, the relationship can be treated as a simple linear regression model to answer the research question. For further analysis, we can generate the linear model between these two parameters in the data analysis part.

Boxplot: Relationship between O3 Concentration and month (2022)



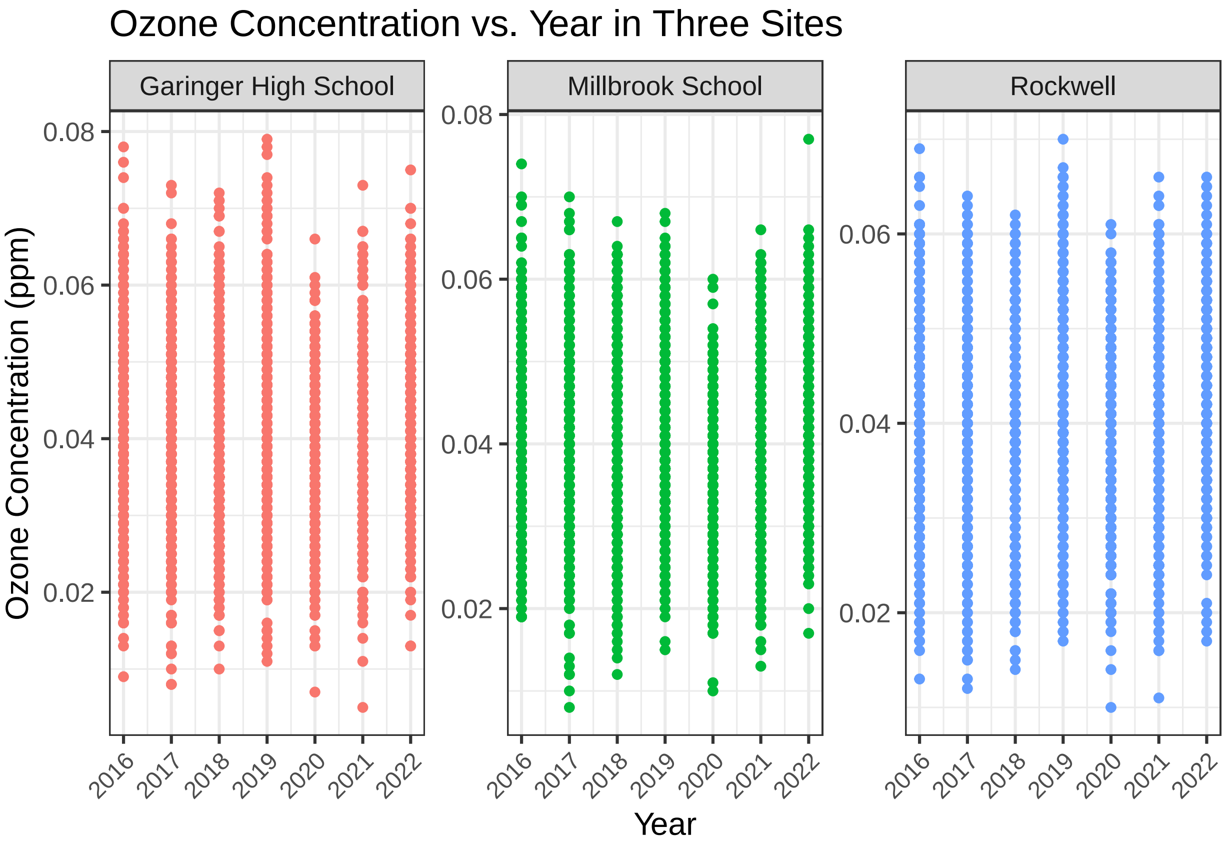
Note that the data for December is missing because 2022 is not over yet. It can be seen from the plot that the ozone concentration varies according to the month, and the average ozone concentration of each month is different. We can apply the Shapiro-Wilk test, Bartlett's test, and ANOVA to further explore the relationship between the average ozone concentration and the month. Overall, June has the highest ozone concentration values, and November, as well as January, has the lowest ozone concentration values. In addition, the relationship between ozone concentration and the season is not obvious. For example, the ozone concentration in April (spring) is higher than that in August (summer), but the highest concentration value also happened in summer (June).

Violin plot: Relationship 03 Concentration and Year



Since the distribution of Ozone concentration differs from the site, so we decided to choose Rockwell as our focus. From the violin plot, the ozone concentration values from the year 2016 to 2022 are most gathered around 0.04, and each year has different max and min. To further analyze the change in ozone concentrations from 2016 to 2022, we can apply time series to this research question.

Plot a subset of sitename



We selected the three locations with the maximum counts of concentration value and plot the distribution of concentration by years. And from that, we did not observe much difference between different locations. Thus, we could select Rockwell as our main focus.