North Carolina Ozone Concentration Monitoring from 2016 to 2022

https://github.com/KristaPeng/PengYangYu_ENV872_EDA_FinalProject.git

Yanxi Peng, Kaichun Yang, Changxin Yu

Rationale and Research Questions: The concentration of tropospheric ozone plays an important role in the atmosphere to global climate change, plant growth and human health. Ozone absorbs solar UVA, UVB irradiation and terrestrial IR irradiation. Also, it's a strong greenhouse gas whose conctentration trend is highly related to global climate change. The elevation of ozone has been linked to adverse health outcomes in many studies for both of human, animal and plants. Therefore, the long-term trend of ozone concentration is crucial. In class, we used ozone data of North Carolina in 2018 and 2019 and we are interested in how the ozone concentration develops in a longer period at different sites of North Carolina. To study the long-term ozone concentration trend, we used data from 2016 to 2022 and conduct linear regression, time series analysis and spatial analysis with the research questions listed below.

Research Question 1: Is there any correlation between ozone concentration and daily AQI in 2022? The null hypothesis: ozone concentration has no correlation with daily AQI. The alternative hypothesis: ozone concentration has correlation with daily AQI.

Research Question 2: Do different sites have equal mean of ozone concentrations in 2022? The null hypothesis: they have equal mean of concentrations. The alternative hypothesis: they do not have equal mean of concentrations.

Research Question 3: Is the mean of ozone concentrations in 2021 and 2022 equivalent? The null hypothesis: the mean between 2021 and 2022 is equivalent. The alternative hypothesis: the mean between 2021 and 2022 is not equivalent.

Research Question 4: Is there any trend of ozone concentrations in space?

Research Question 5: Have ozone concentrations changed from 2016 to 2022 at Rockwell? The null hypothesis: the ozone concentration is stationary over time. The alternative hypothesis: the ozone concentration change over time.

Dataset Information:

Here we studied the ozone concentration trend at North Carolina from 2016 to 2022. The datasets were downloaded directly from EPA Outdoor Air Quality Data section (https://www.epa.gov/outdoor-air-quality-data/download-daily-data). The raw data contains detailed information including Date, Source, Site ID, POC, Daily Max, Units etc. Since not all variables are required in our project, we selected the necessary data and form a new dataframe which is comprised of variables including Date, Daily.Max.8.hour.Ozone.Concentration, DAILY_AQI_VALUE, Site.Name, COUNTY, SITE_LATITUDE, SITE_LONGITUDE. After the data selection, we add year and month column for further analysis and fill out NA values in the blank date with linear interpolation.

Analysis: Part 1: Data wrangling

Part 2: Data exploring and visualization # Set up R session

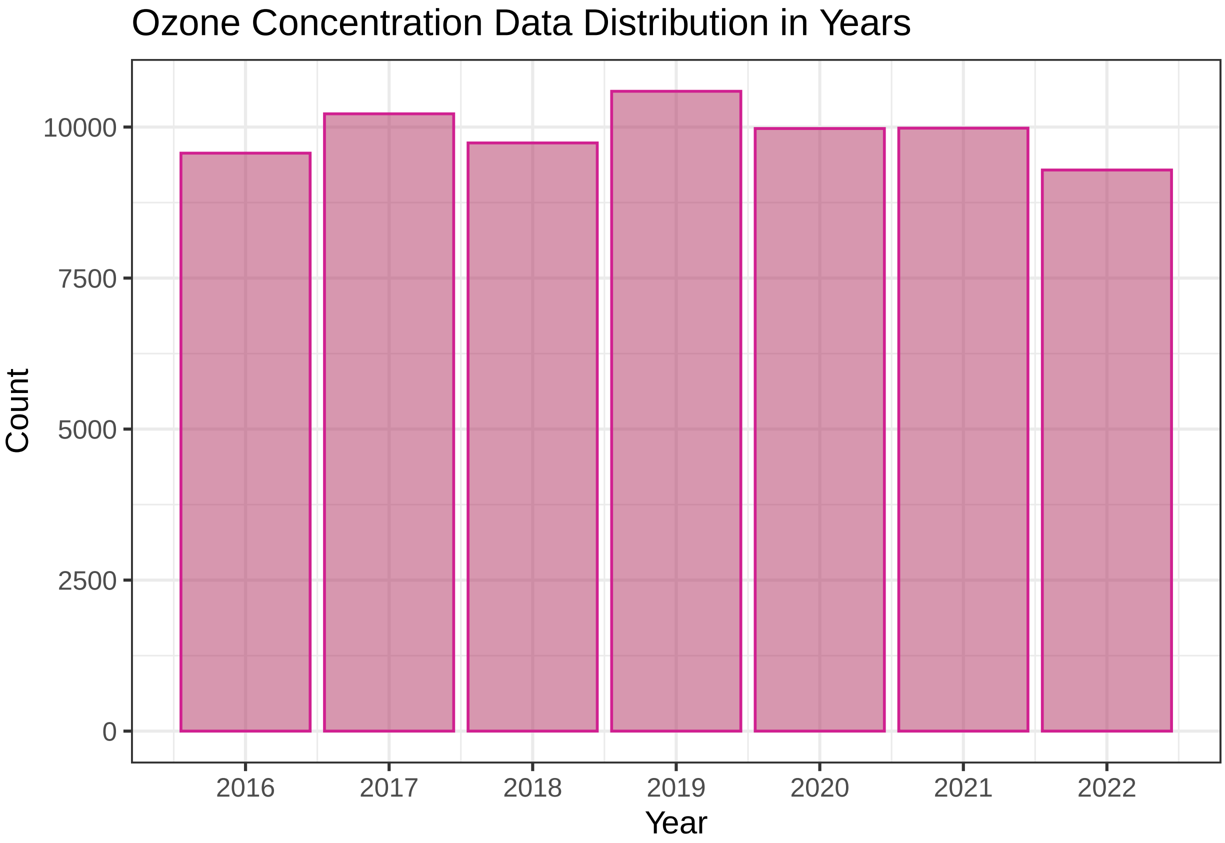
1 Obtain basic summaries of the data

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

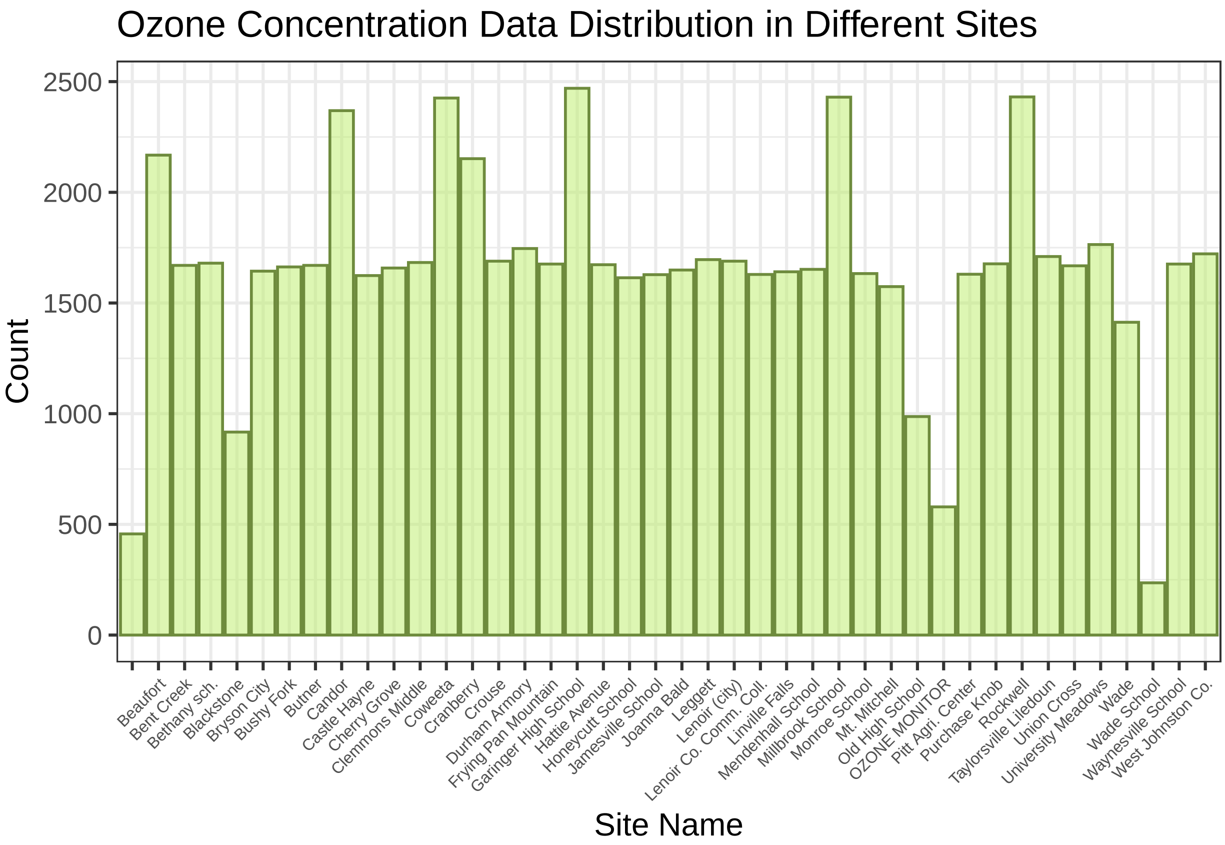
```
##        X                  Date       Concentration       DailyAQI
##  Min.   :    1   2016-04-11:   40   Min.   :0.00000   Min.   :  0.00
##  1st Qu.:17342   2016-04-14:   40   1st Qu.:0.03400   1st Qu.: 31.00
##  Median :34682   2016-04-15:   40   Median :0.04200   Median : 39.00
##  Mean   :34682   2016-04-16:   40   Mean   :0.04163   Mean   : 39.51
##  3rd Qu.:52023   2016-04-17:   40   3rd Qu.:0.04900   3rd Qu.: 45.00
##  Max.   :69363   2016-04-20:   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

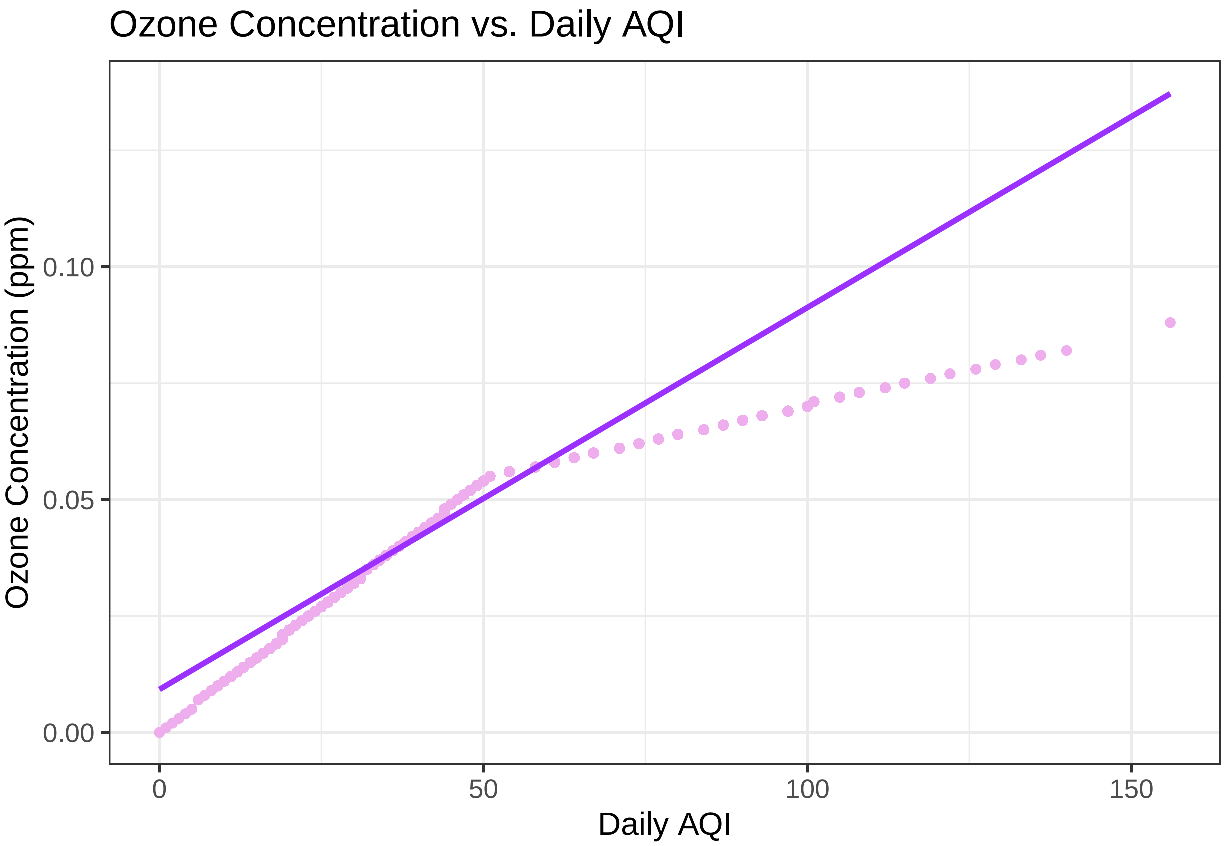
1.0.1 Generate a plot of the number of ozone concentrations collected conducted by year



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

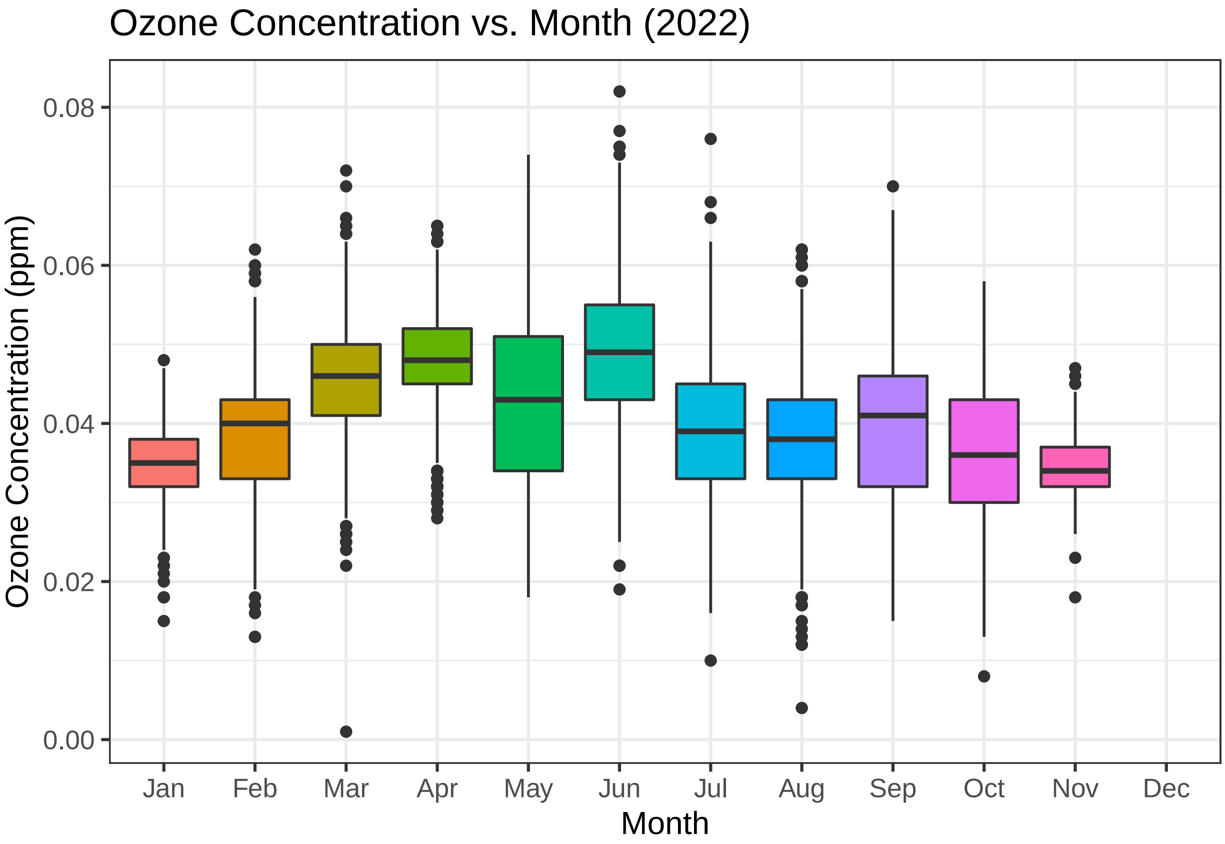


From the bar plot, the number of data samples collected at different sites varied considerably. For further analysis in this project, to ensure the accuracy of the results, we should choose the site with more data extraction as the analysis object. In this case, Garinger high school, Rockwell, Millbrook School, and Coweeta can be considered to use in the data analysis part. ### Is there any relation between Daily AQI and O3 Concentration?

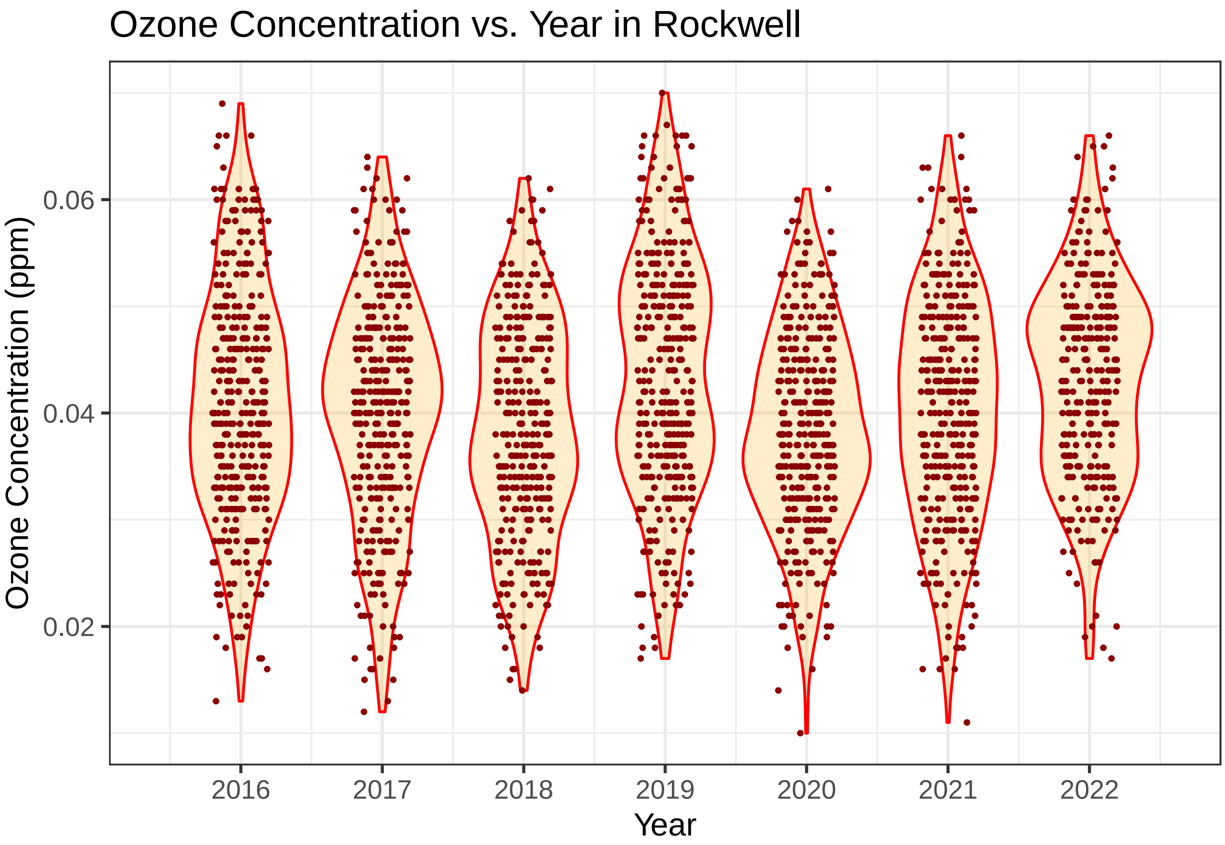


From the plot, we do observe a linear relation between Ozone concentration and daily AQI. Hence, in the data analysis part, the relationship can be treated as a simple linear regression model to answer the research question. For further analysis, we can generate the linear model between these two parameters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

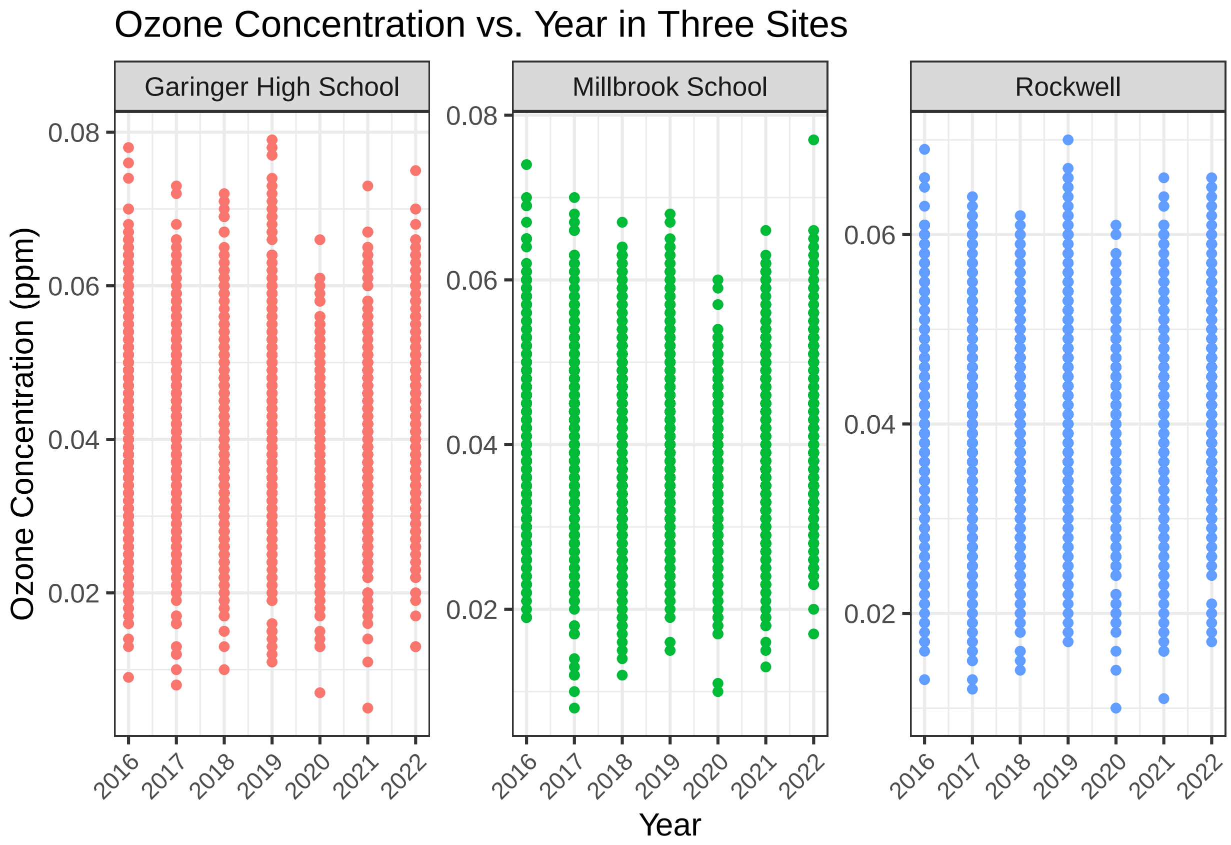
1.0.2 Boxplot: Relationship between O3 Concentration and month (2022)



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.

Part 3: Data Analysis

1.0.3 Research question 1:

Is there any correlation between ozone concentration and daily AQI in 2022?

The null hypothesis: ozone concentration has no correlation with daily AQI

The alternative hypothesis: ozone concentration has correlation with daily AQI

```
## Reading layer ‘cb_2018_us_county_20m’ from data source
##   ‘E:\things\Duke University\study\2022 Fall\ENVIRON 872\PengYangYu_ENV872_EDA_FinalP
##   using driver ‘ESRI Shapefile’
## Simple feature collection with 3220 features and 9 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -179.1743 ymin: 17.91377 xmax: 179.7739 ymax: 71.35256
## Geodetic CRS:  NAD83
```

```
##
## Call:
## lm(formula = Concentration ~ DailyAQI, data = EPAair_O3_22)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.042585 -0.000866  0.000372  0.001434  0.003672
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.350e-03  1.110e-04   84.25   <2e-16 ***
## DailyAQI    8.231e-04  2.688e-06  306.23   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.002882 on 9287 degrees of freedom
## Multiple R-squared:  0.9099, Adjusted R-squared:  0.9099
## F-statistic: 9.378e+04 on 1 and 9287 DF,  p-value: < 2.2e-16
```

The R-squared of this linear model is 0.9099, which means that 90.99% variability in daily air quality index (AQI) is explained by changes in ozone concentration. The degrees of freedom of the model is 9287 = 9289-2 since the number of observations is 9289 and the number of parameters is 2. According to the p-value of slope, which is smaller than 0.05, we reject the null hypothesis, so the correlation between ozone concentration and daily AQI in 2022 is significant.

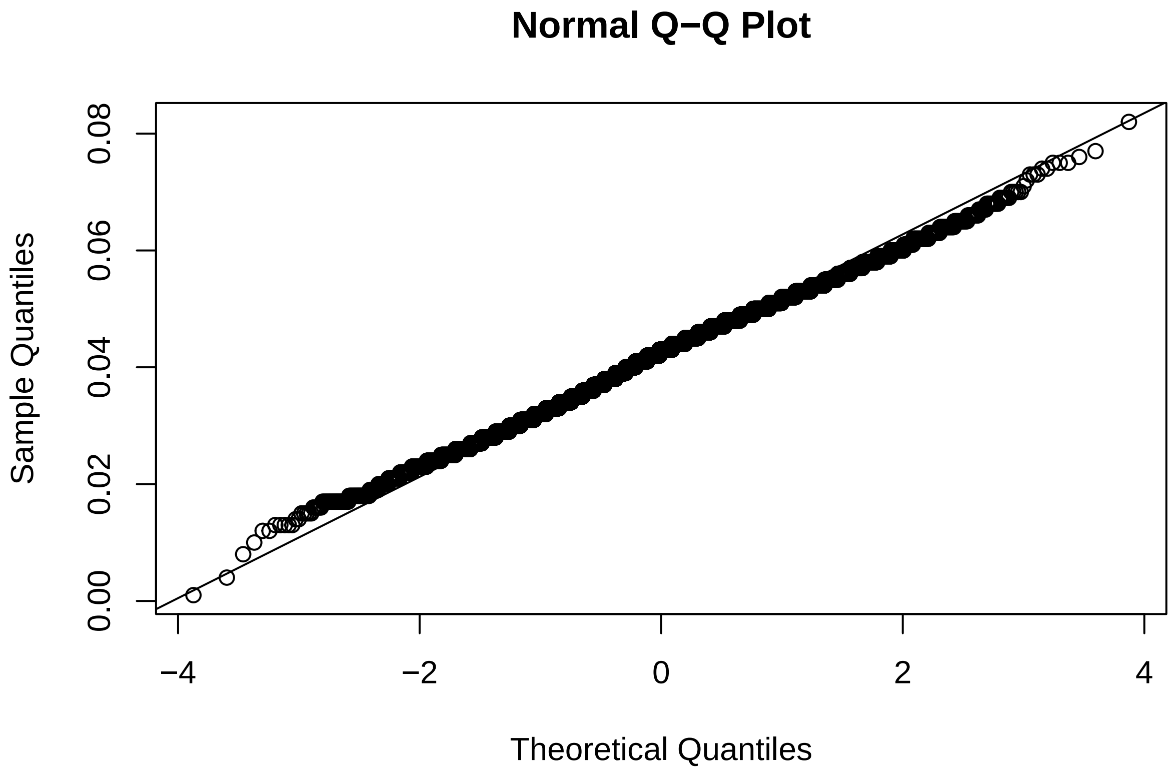
1.0.4 Research question 2:

Do different sites have equal mean of ozone concentrations in 2022?

The null hypothesis: they have equal mean of concentrations

The alternative hypothesis: they do not have equal mean of concentrations

```
## [1] "There are 38 sites in total and 28 of them follow a normal distribution."
```



```
##
##  Bartlett test of homogeneity of variances
##
## data:  EPAair_O3_22$Concentration and EPAair_O3_22$SiteName
## Bartlett's K-squared = 114.06, df = 37, p-value = 8.769e-10
```

```
##               Df Sum Sq   Mean Sq F value Pr(>F)
## SiteName      37 0.0546 0.0014746   17.02 <2e-16 ***
## Residuals   9251 0.8016 0.0000867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Shapiro-Wilk test result and normal Q-Q plot, we find that most of the sites are conform to normal population distribution assumption. However, the Bartlett's test result shows that the null hypothesis that the variances in each sites are the same is rejected. Since ANOVA is robust against departures from equal variance, we can still apply one-way ANOVA on our dataset.

The p-value of the ANOVA is smaller than 0.05, so we reject the null hypothesis. Therefore, the mean of ozone concentrations in 2022 significantly differ among sites.

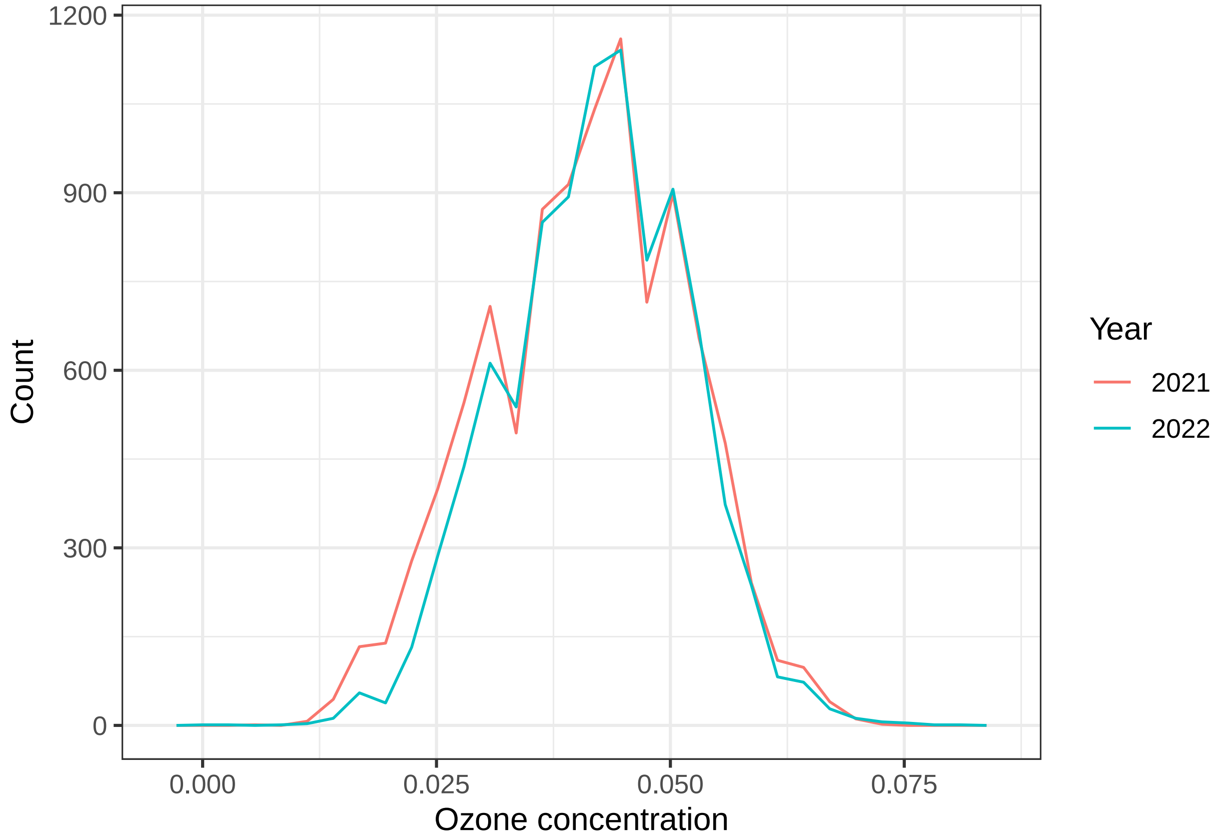
1.0.5 Research question 3:

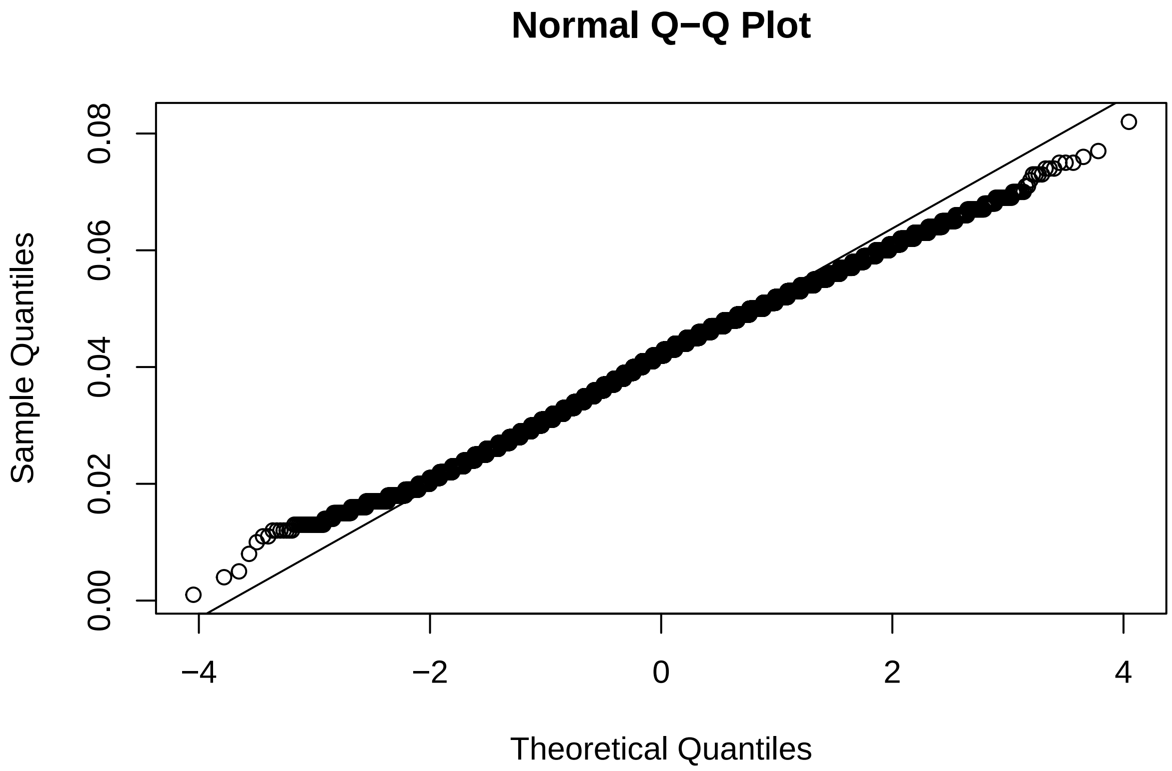
Is the mean of ozone concentrations in 2021 and 2022 equivalent?

The null hypothesis: the mean between 2021 and 2022 is equivalent

The alternative hypothesis: the mean between 2021 and 2022 is not equivalent

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





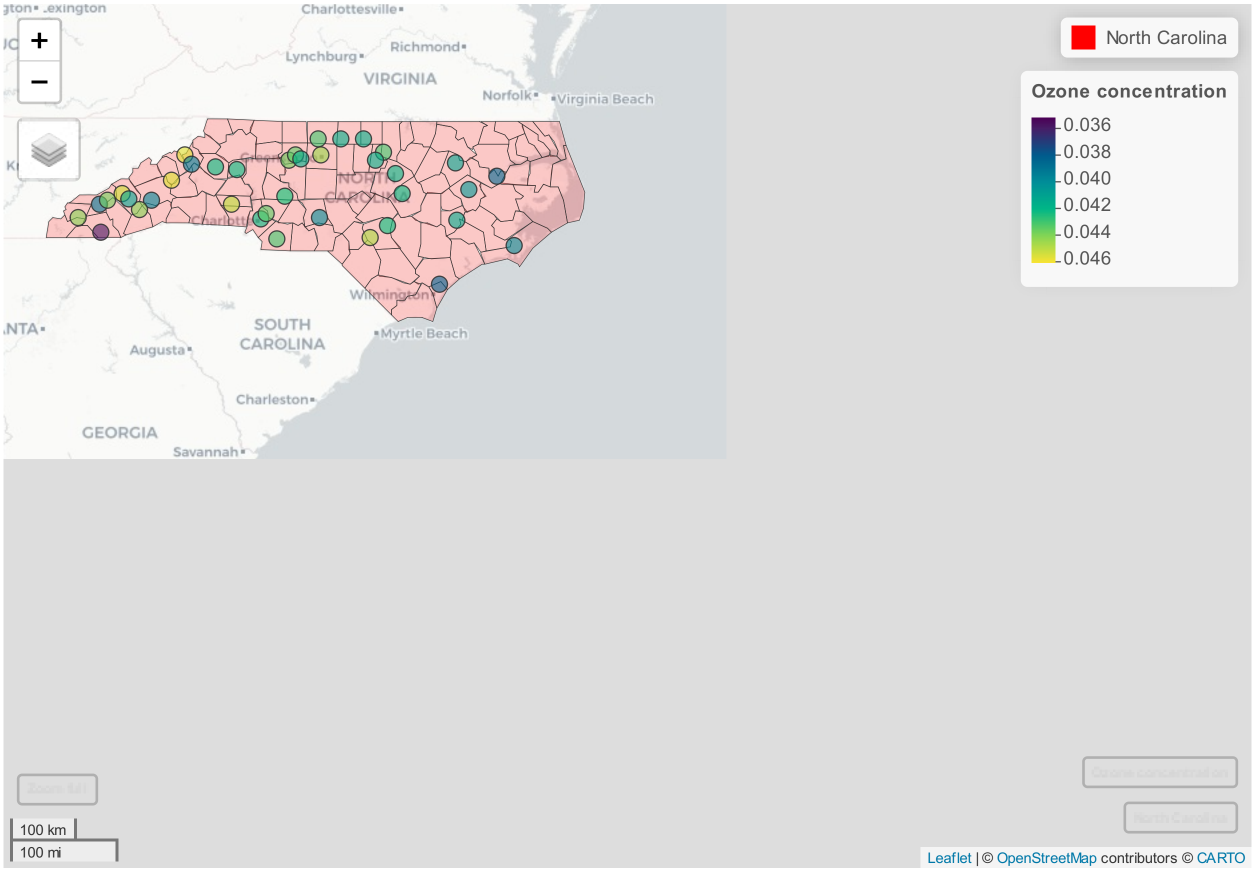
```
##
##  Bartlett test of homogeneity of variances
##
## data:  EPAair_O3_2122$Concentration and EPAair_O3_2122$Year
## Bartlett's K-squared = 103.52, df = 1, p-value < 2.2e-16
```

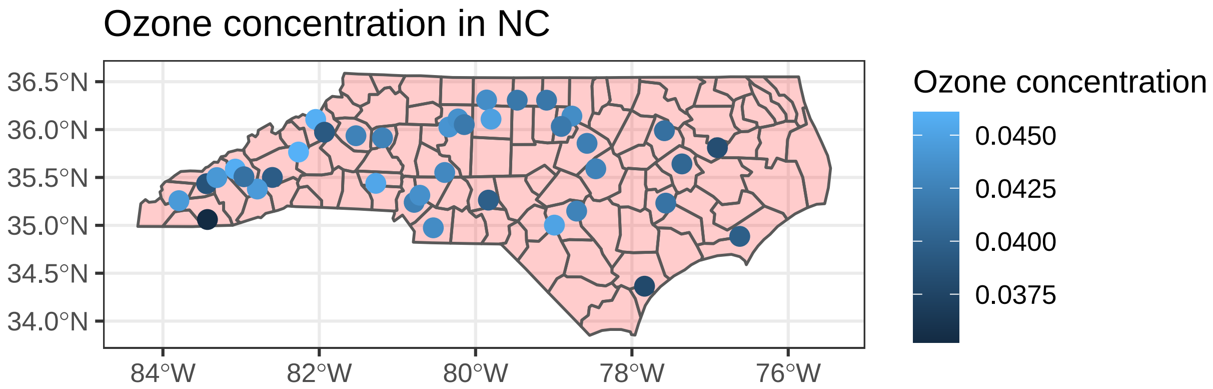
```
##
##  Welch Two Sample t-test
##
## data:  Concentration by Year
## t = -7.0983, df = 19249, p-value = 1.307e-12
## alternative hypothesis: true difference in means between group 2021 and group 2022 is
## 95 percent confidence interval:
##  -0.0013217999 -0.0007497675
## sample estimates:
## mean in group 2021 mean in group 2022
##         0.04104097         0.04207676
```

The normal Q-Q plot shows that the data has small deviations from normal distribution. The Bartlett's test result shows that the variances between 2021 and 2022 are different. Again, t-test is robust to these. The T-test suggests that the mean of ozone concentrations in 2021 and 2022 is not equivalent with p-value smaller than 0.05.

1.0.6 Research question 4:

Is there any trend of ozone concentrations in space?





It seems there is no obvious trend between the location of sites and the mean ozone concentration collected from the sites.

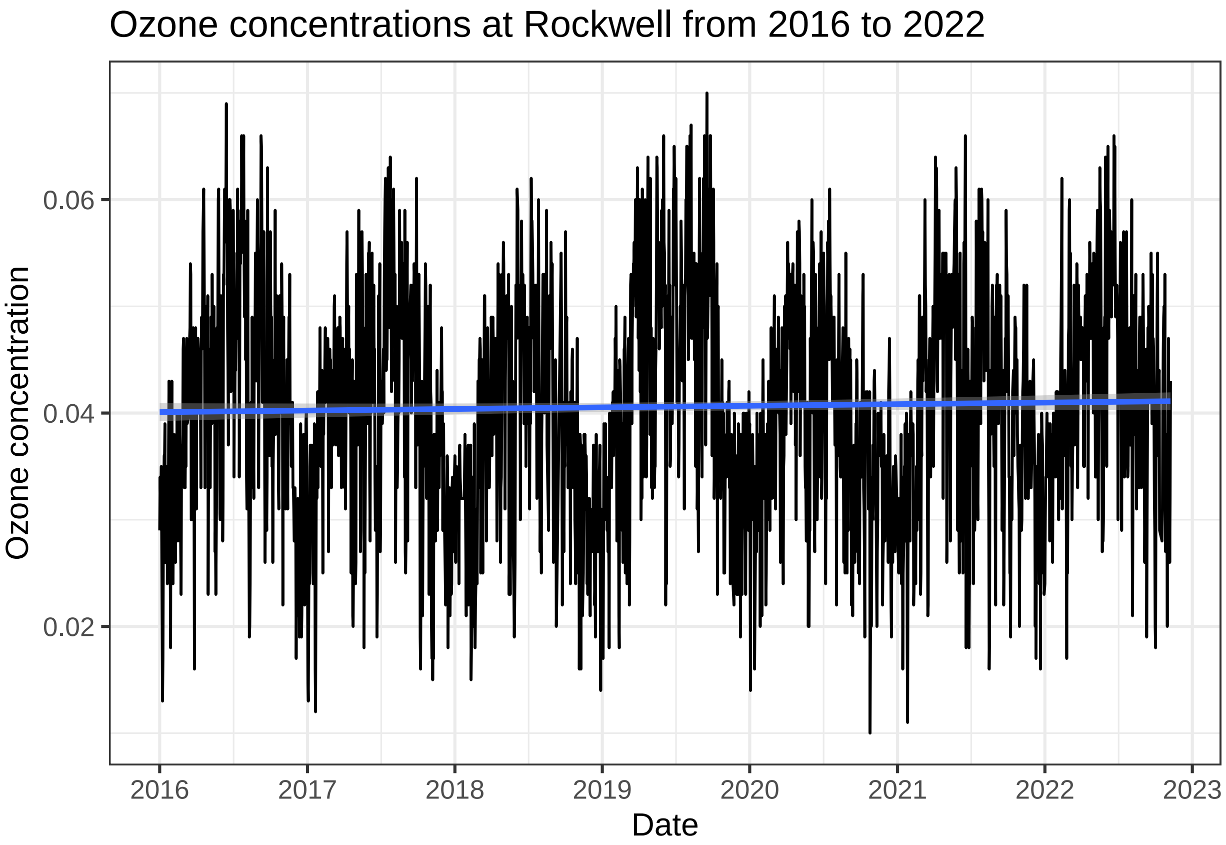
1.0.7 Research question 5:

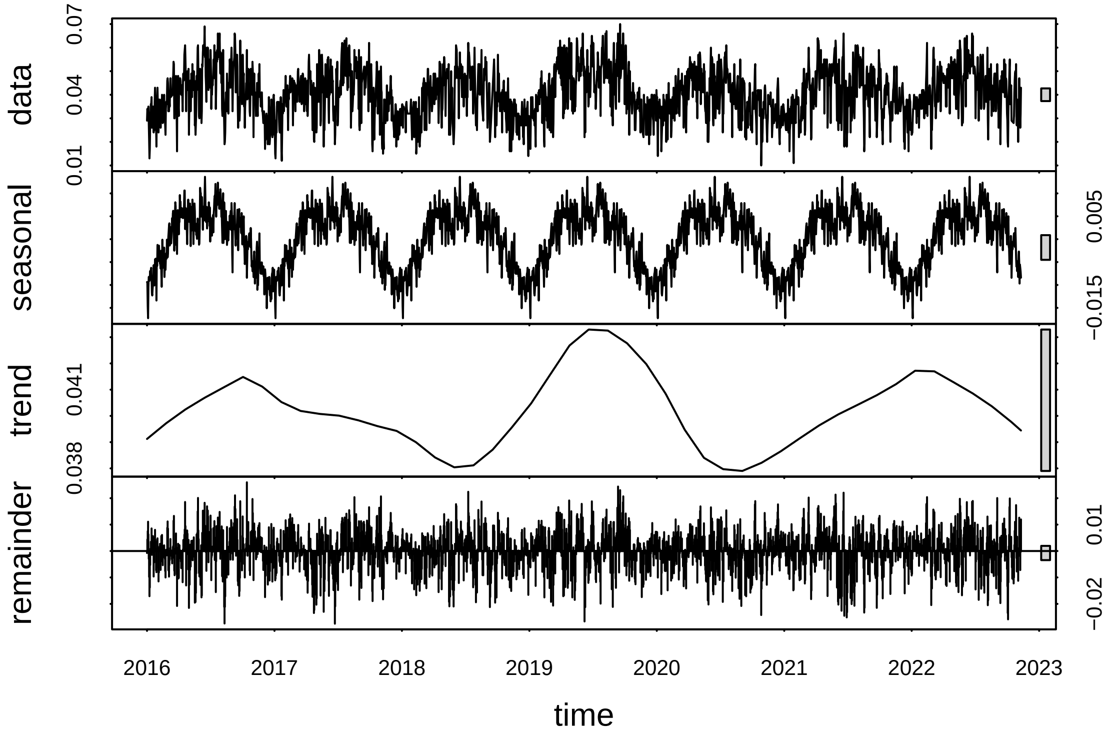
Have ozone concentrations changed from 2016 to 2022 at Rockwell?

The null hypothesis: the ozone concentration is stationary over time

The alternative hypothesis: the ozone concentration change over time

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```


Ozone concentrations at Rockwell from 2016 to 2022
Ozone concentration
0.06
0.04
0.02
Date
2016
2017
2018
2019
2020
2021
2022
2023



```
## Score =  87 , Var(Score) = 15119
## denominator =  7239.089
## tau = 0.012, 2-sided pvalue =0.47922
```

Summary and Conclusions:

The Seasonal Mann-Kendall test is chosen to test monotonic trend because the decomposed figure shows that the time series object has a strong seasonal component. From the result, we accept the null hypothesis since the p-value is greater than 0.05. Therefore, the ozone concentration at Rockwell is stationary from 2016 to 2022.