Trends in US Major Air Pollutants

PROJECT REPORT

AKITHA PINISETTI
AMITH NANDIVADA
KUSHWANTH SAI LALAM
RISSHIKAA YANTRAPATI
SAI MADDULA
VINEETH REDDY BASANI

Contents

1. INTRODUCTION
2. ANALYSIS
3. RESULTS
4. **CONCLUSIONS**

1. INTRODUCTION

What is pollution and its types?

Pollution is defined as the contamination of the natural environment with harmful substances that cause tremendous change. Pollution is typically composed of synthetic or man-made substances (such as plastic), but natural substances such as sediment, nutrients, and carbon dioxide can also become pollutants when levels exceed a certain threshold. Different types of pollution include water, soil, noise, and air. Pollutants can be natural or manmade, such as volcanic ash, trash or residues that are disposed of by industries. These pollutants damage the quality of land, air, and water. A polluted environment can lead to different hazards, such as a burning sensation in the eyes, difficulty breathing and lung cancer. Pollution can degrade ecosystem health by harming or even killing the living things that inhabit those ecosystems.

Air Pollution -

Air pollution is a combination of hazardous substances emitted by both man-made and natural sources. The main causes of man-made air pollution include vehicle emissions, fuel oils and natural gas used to heat homes, waste products from manufacturing and energy production, particularly from coal-fired power plants, and fumes from chemical manufacturing. Nature emits hazardous substances into the atmosphere, such as smoke from wildfires, which are frequently caused by humans; ash and gases from volcanic eruptions; and gases emitted by decomposing organic matter in soils, such as methane. Air pollution and its effects on health are constantly being studied. When exposed to air pollution, it can cause oxidative stress, which can lead to chronic diseases and cancer. Cardiovascular disease, respiratory illnesses, diabetes, obesity, disorders of the nervous and immune systems, and obesity are currently among the public health issues that need to be addressed.

Most Widespread Air Pollutants and Its Origin –

Carbon monoxide, Sulphur dioxide, nitrogen dioxide, and ozone are the four major pollutants. Sulphur dioxide is the most widespread pollutant gas in the atmosphere, and it is commonly encountered in high densities in urban and industrial areas. Smog is the term used to describe air pollution caused by ozone, which significantly contributes to urban air pollution. The majority of the Sulphur dioxide released into the environment comes from power plants, particularly those that burn coal. Other Sulphur dioxide sources include petroleum refineries, cement manufacturing, paper pulp manufacturing, and metal smelting and processing plants. Citizens who commute daily are regularly exposed to air pollution, which has an adverse effect on their health.

To mitigate the impacts of air pollution that are being generated, the US government has decided to incorporate a few remedial measures.

- Keeping your vehicle in good condition.
- Using public transportation or carpooling frequently can help to reduce the amount of gas released.
- Do not burn trash in a fire pit or a burn barrel.

- Plant and maintain more trees to remove toxic gases from the atmosphere and release oxygen.
- Garden with lawn equipment or use electricity to shovel snow. Efficient appliance use can help to reduce pollution.

2. ANALYSIS

Data Collection & Data Cleaning

We gathered data on pollution in the United States and analyzed it from distinct viewpoints. We concentrated on four key pollutants (nitrogen dioxide, Sulphur dioxide, carbon monoxide, and ozone) that significantly contributed to pollution in all 50 states of the United States of America between 2000 and 2016. Furthermore, we examined how these pollutants changed across all states over a 16-year period using data from the Kaggle website, which provides data on all four major pollutants and their measures.

The variables that are used to measure the levels of pollution in the states include

- AQI (Air Quality Index) - It provides an index value for the current pollution level.
- 1st Max Value – It describes the highest value of that day.
- 1st Max Hour – This column specifies the hour of the day that has the highest value.

The data we are working on is extracted from Kaggle, and it consists of 28 columns with 5 distinct columns on all 4 major pollutants that were planned to do analysis on.

Dataset Reference Link: https://www.kaggle.com/datasets/sogun3/uspollution

Raw Data –

```
> head(pollution_us) %>% select(`CO AQI`,`SO2 AQI`,everything())
# A tibble: 6 x 29
  `CO AQI` `SO2 AQI`  ...1 `State Code` `County Code` `Site Num` Address State County City  `Date Local` `NO2 Units`
     <dbl>     <dbl> <dbl>        <dbl>         <dbl>      <dbl> <chr>   <chr> <chr>  <chr> <date>       <chr>
1       NA        13     0            4            13       1002 1645 E… Ariz… Maric… Phoe… 2000-01-01   Parts per …
2       25        13     1            4            13       1002 1645 E… Ariz… Maric… Phoe… 2000-01-01   Parts per …
3       NA        NA     2            4            13       1002 1645 E… Ariz… Maric… Phoe… 2000-01-01   Parts per …
4       25        NA     3            4            13       1002 1645 E… Ariz… Maric… Phoe… 2000-01-01   Parts per …
5       NA         4     4            4            13       1002 1645 E… Ariz… Maric… Phoe… 2000-01-02   Parts per …
6       26         4     5            4            13       1002 1645 E… Ariz… Maric… Phoe… 2000-01-02   Parts per …
# … with 17 more variables: `NO2 Mean` <dbl>, `NO2 1st Max Value` <dbl>, `NO2 1st Max Hour` <dbl>, `NO2 AQI` <dbl>,
#   `O3 Units` <chr>, `O3 Mean` <dbl>, `O3 1st Max Value` <dbl>, `O3 1st Max Hour` <dbl>, `O3 AQI` <dbl>,
#   `SO2 Units` <chr>, `SO2 Mean` <dbl>, `SO2 1st Max Value` <dbl>, `SO2 1st Max Hour` <dbl>, `CO Units` <chr>,
#   `CO Mean` <dbl>, `CO 1st Max Value` <dbl>, `CO 1st Max Hour` <dbl>
>
```

Fig 1: Screenshot of the raw data

Cleaning -

- From the data we have extracted, null values in the AQI columns were removed.
- The data type of the date column has changed from string to date format
- Unnecessary columns were removed from the data set.
- Created a data frame for five different states and named it data_of_five_states.

- Performed feature engineering on the data set and added year and month columns to the data set for further analysis.

```
> head(pol) %>% select(co_aqi,so2_aqi,everything())
# A tibble: 6 x 31
  co_aqi so2_aqi  ...1 state_code county_code site_num address      state county city  date       no2_units no2_mean
   <dbl>   <dbl> <dbl>      <dbl>       <dbl>    <dbl> <chr>        <chr> <chr>  <chr> <date>     <chr>        <dbl>
1     25      13     1          4          13     3002 1645 E ROOS… Ariz… Maric… Phoe… 2000-01-01 Parts pe…     19.0
2     30      21   125          4          13     3002 1645 E ROOS… Ariz… Maric… Phoe… 2000-02-01 Parts pe…     34.6
3     24      14   241          4          13     3002 1645 E ROOS… Ariz… Maric… Phoe… 2000-03-01 Parts pe…     33.9
4     16       6   365          4          13     3002 1645 E ROOS… Ariz… Maric… Phoe… 2000-04-01 Parts pe…     26.8
5     24      13   469          4          13     3002 1645 E ROOS… Ariz… Maric… Phoe… 2000-05-01 Parts pe…      3.79
6     10       7   577          4          13     3002 1645 E ROOS… Ariz… Maric… Phoe… 2000-06-01 Parts pe…     38.5
# … with 18 more variables: no2_1st_max_value <dbl>, no2_1st_max_hour <dbl>, no2_aqi <dbl>, o3_units <chr>,
#   o3_mean <dbl>, o3_1st_max_value <dbl>, o3_1st_max_hour <dbl>, o3_aqi <dbl>, so2_units <chr>, so2_mean <dbl>,
#   so2_1st_max_value <dbl>, so2_1st_max_hour <dbl>, co_units <chr>, co_mean <dbl>, co_1st_max_value <dbl>,
#   co_1st_max_hour <dbl>, year <chr>, month <fct>
```

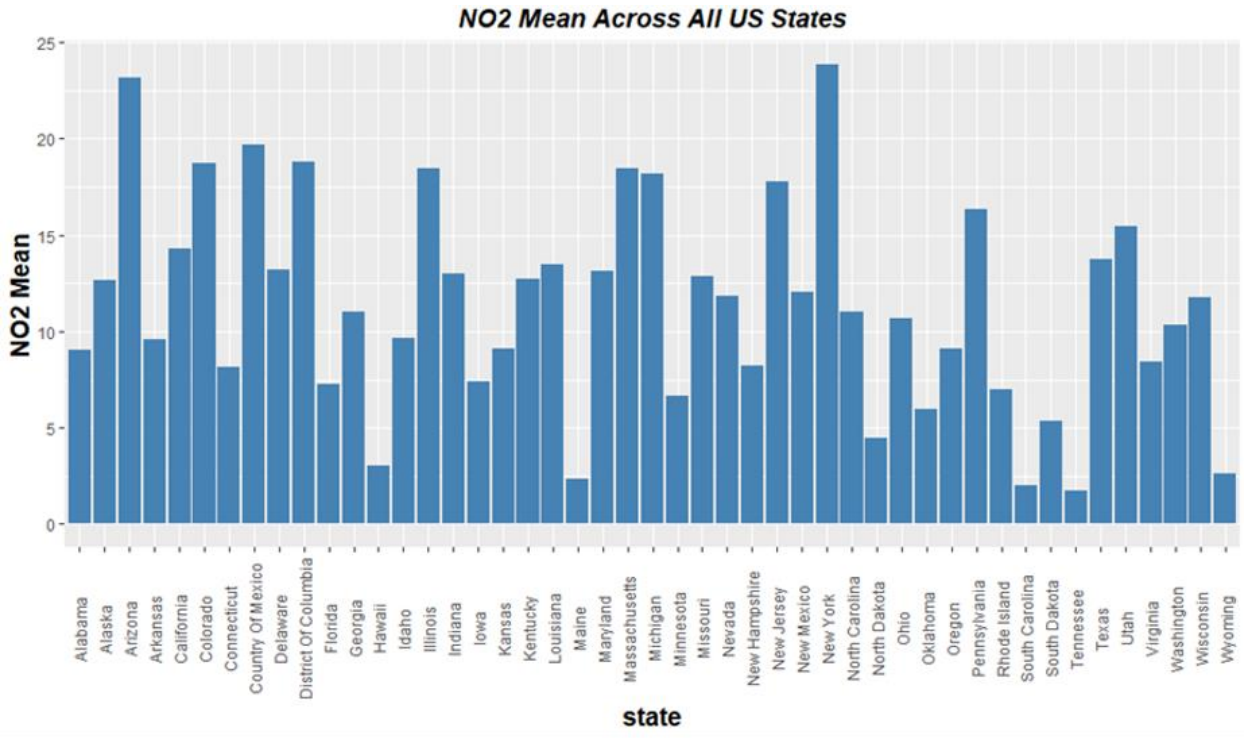
Fig 2: Screenshot of the cleaned data

Data Visualization

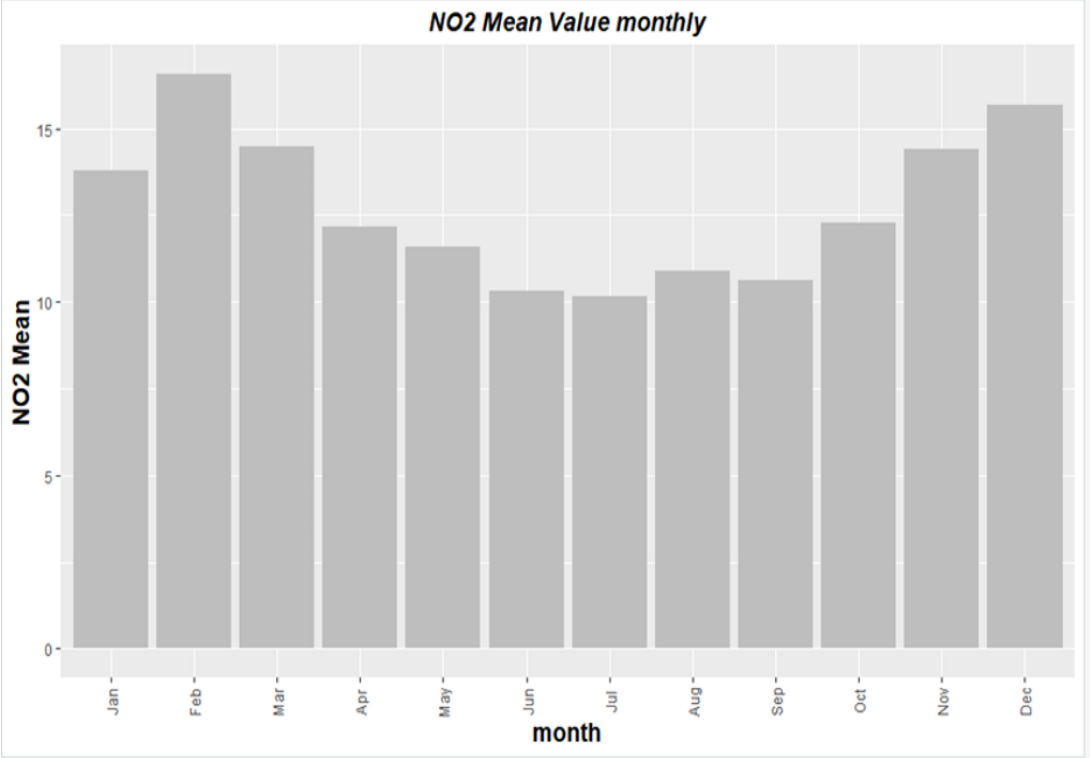
A) NO2:

1. Bar Graph

a. According to the bar graph below, which depicts various NO2 mean trends, New York has the highest rate of NO2 emissions, while Tennessee has the lowest rate.

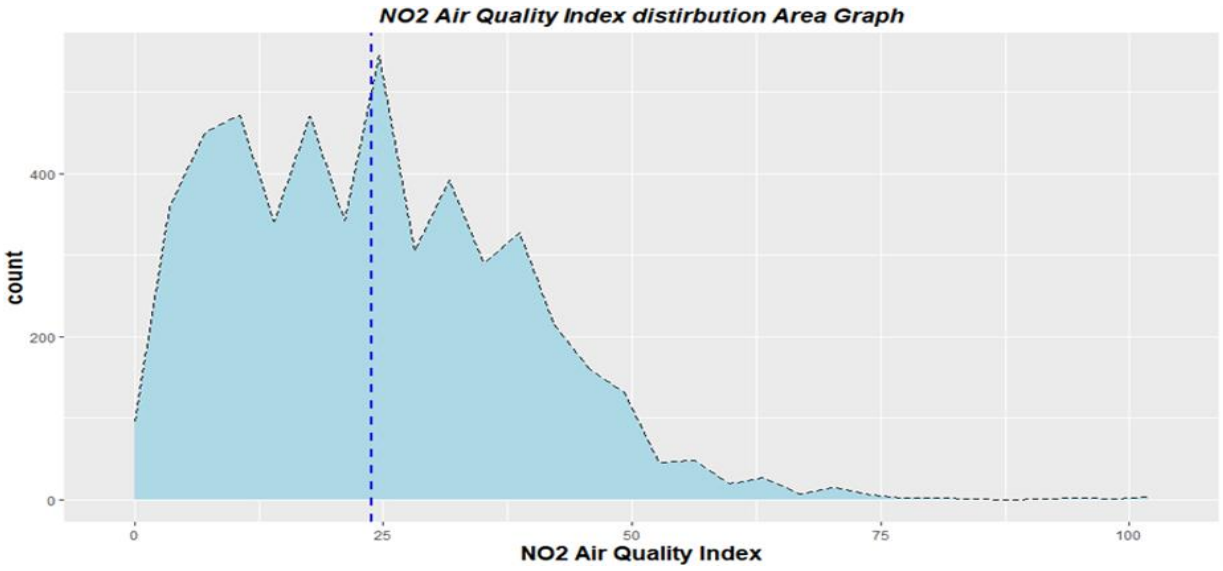


b. Based on the graph pattern analyzed between the months, the NO2 mean value is lower in summer (i.e., June to September) and highest in the months of winter, with more reflection in February.



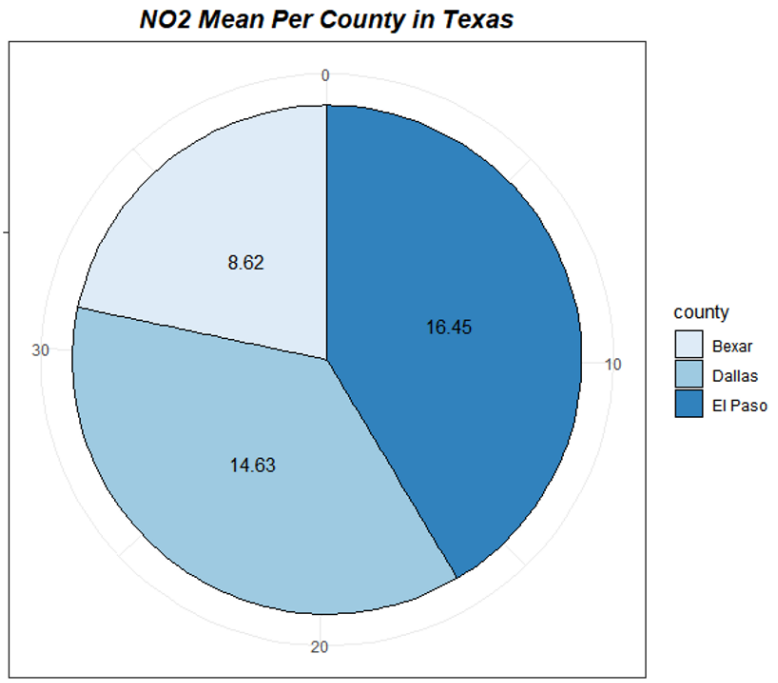
2. Area Graph

a) The mean and the highest values of the air quality index for NO2 gas are near 25, which is good as it lies within the range of 0 to 50, which is considered good as per the AQI forecast used by the government agencies.



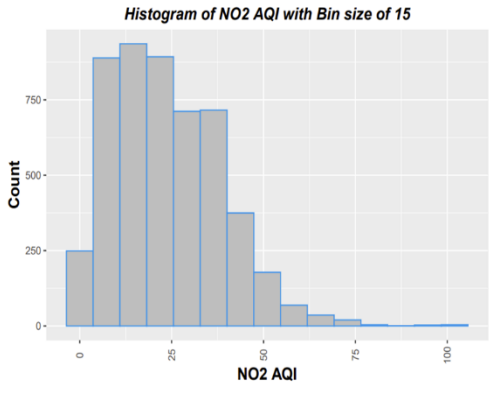
3. Pie Graph

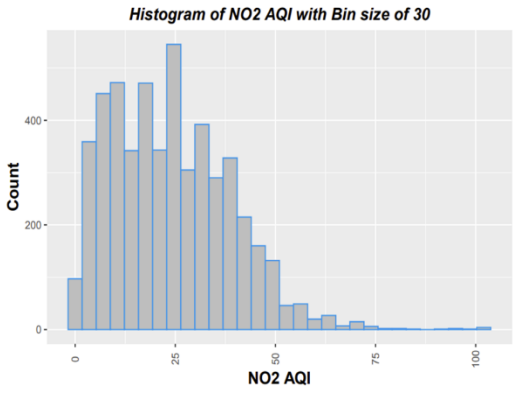
In this graph, we examine the NO2 mean for three counties: Bexar, Dallas, and El Paso. El Paso has the greatest NO2 mean of 16.45, while Bexar has the lowest NO2 mean of 8.62.

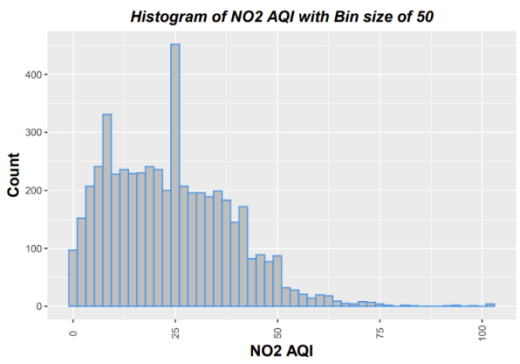


4. Histograms

a. Histogram of different bin sizes on NO2 AQI - For creating a set of histograms, we initially started with NO2 air quality index with different bin sizes of 15, 30 and 50

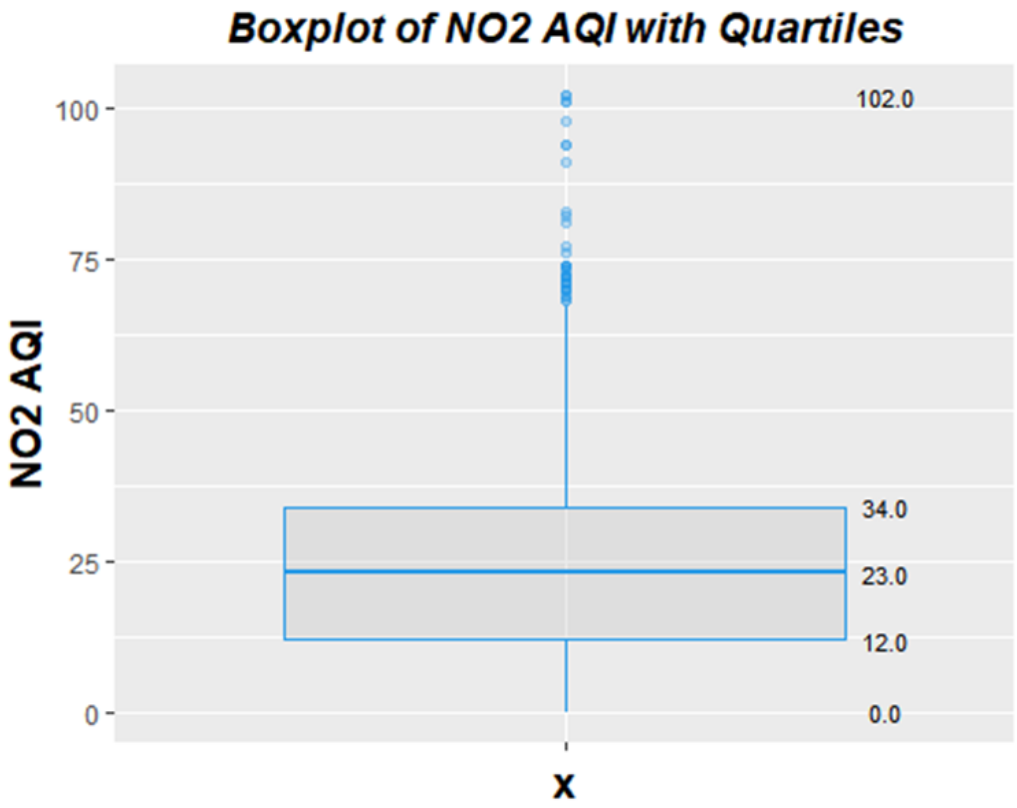






5. Box plot

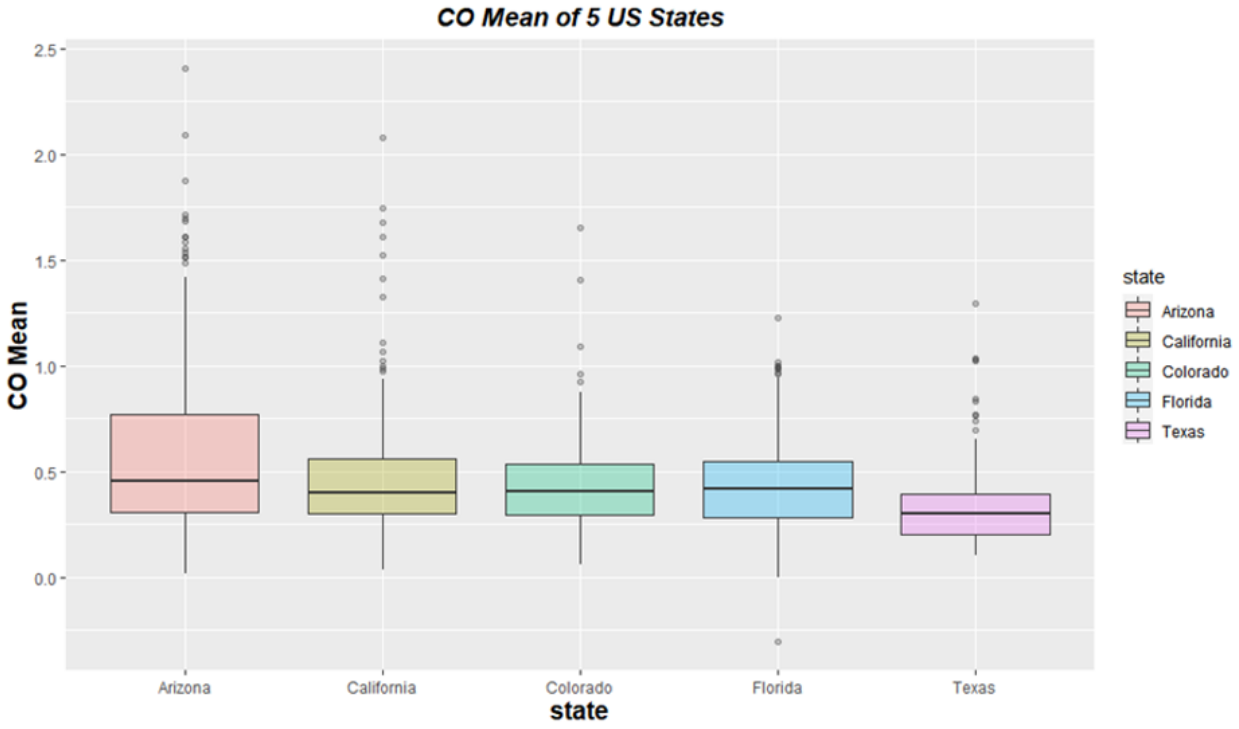
a. The outlier values vary from almost 60 and higher to nearly up to 102, according to the boxplot of NO2 AQI.



B) CO

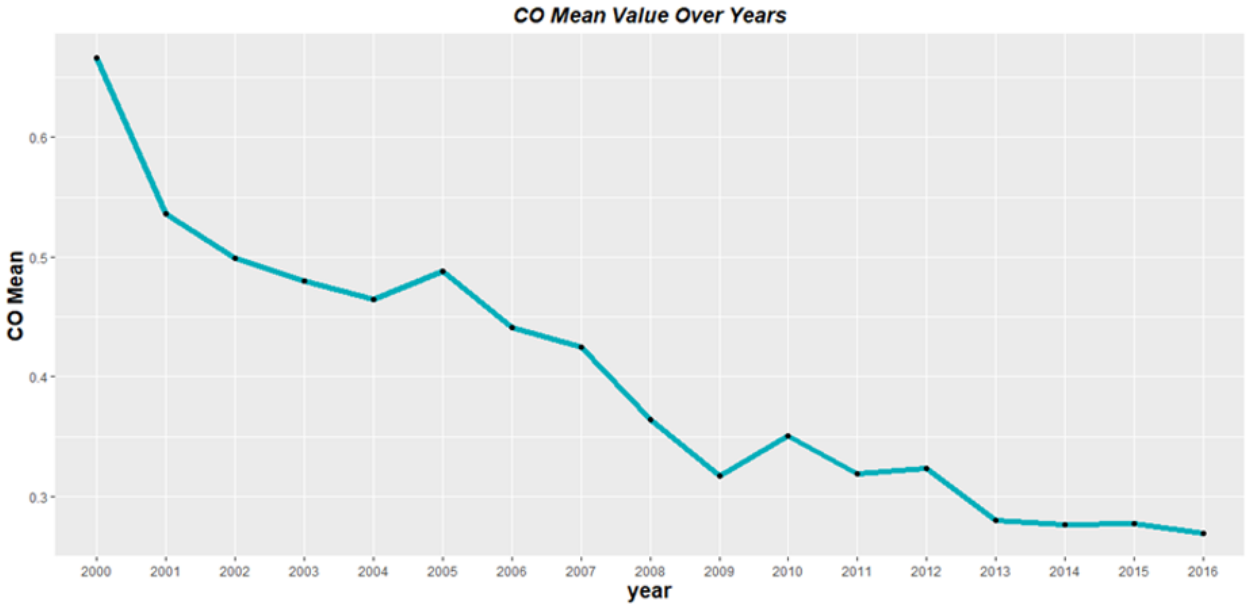
1.Box Plot

a. To observe the highest emission of carbon monoxide gas in five different states, we have used a box plot in which Arizona has the highest CO mean, and Texas stands at the lowest. Furthermore, we have observed from the graph these cities have outlier values.

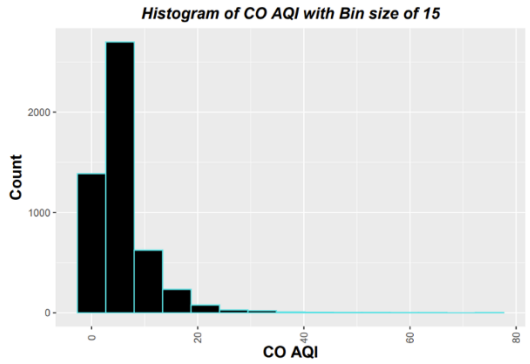


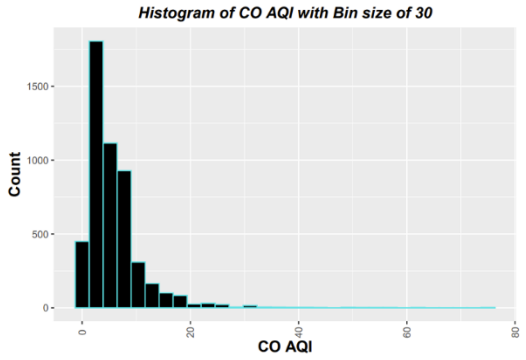
2. Line Chart

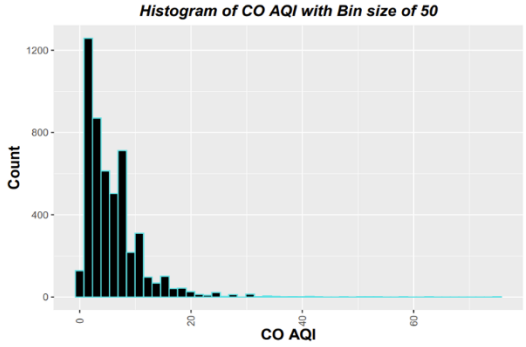
a. From the line chart, we compared from 2000 to 2016 and we have observed that Carbon Monoxide emission has a decreasing trend over the years, with only minor increases.



c. We used the CO2 air quality index with different bins sizes of 15, 30, and 50



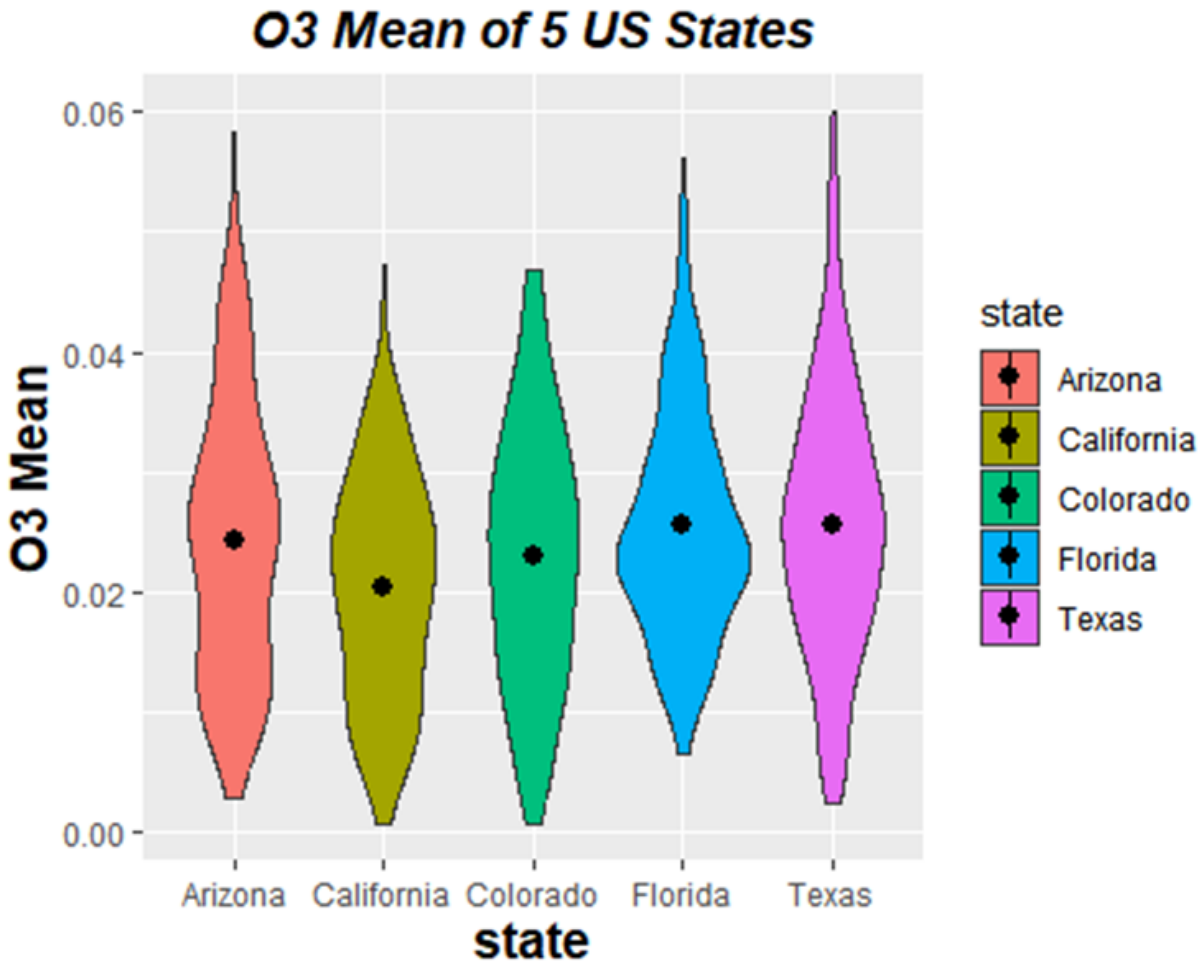




C) O3

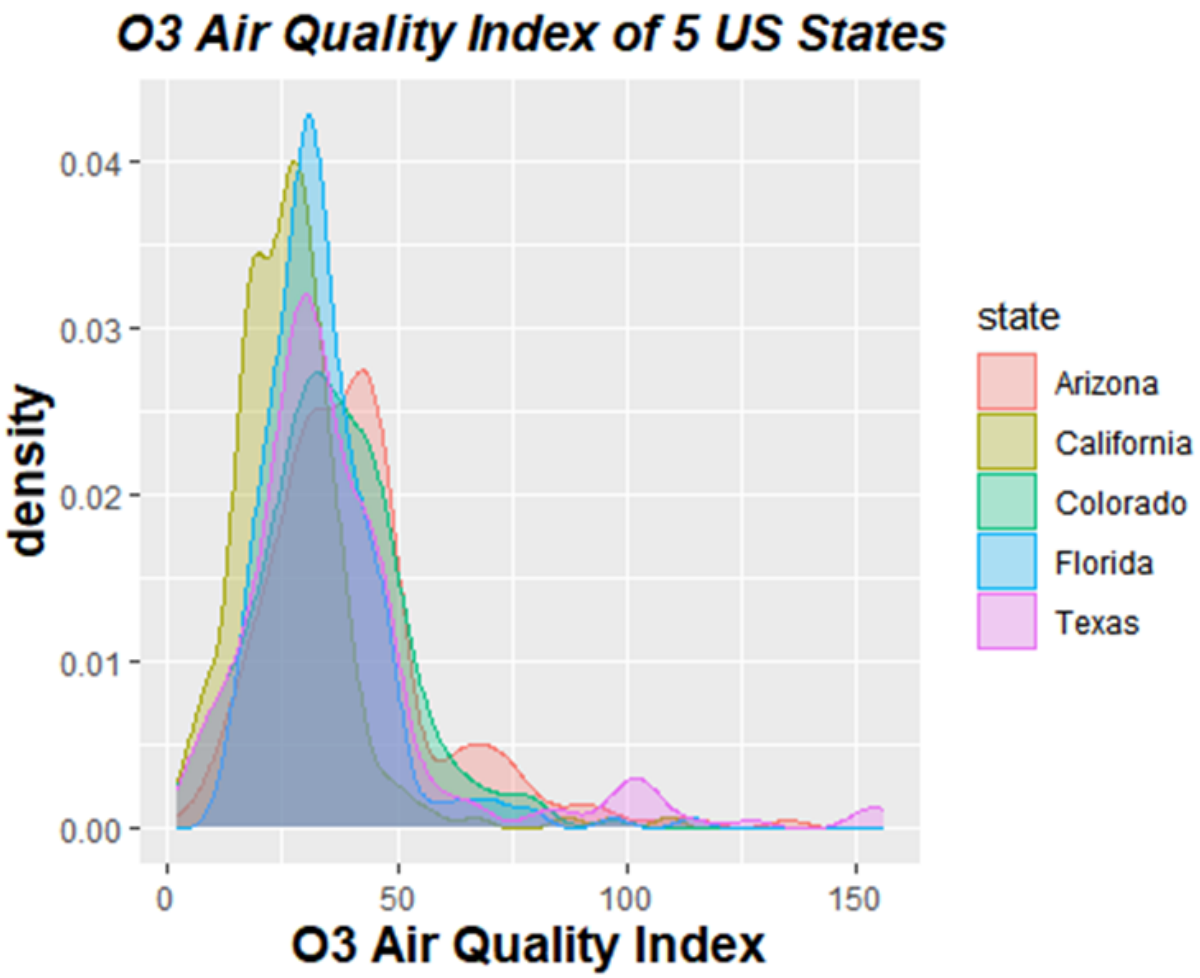
Violin graph

We used a violin plot to identify the states with the highest ozone gas emissions, with Florida having the highest O3 mean and California having the lowest.



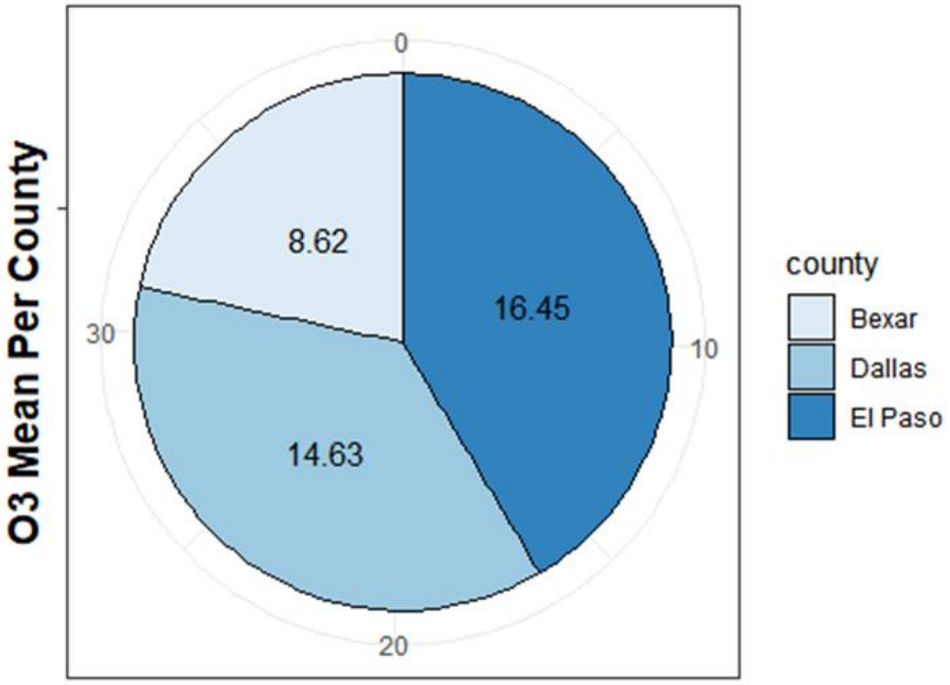
Density Graph

The density graph of O3 AQI values among five states-Arizona, California, Colorado, Florida, and Texas.



Pie Chart:

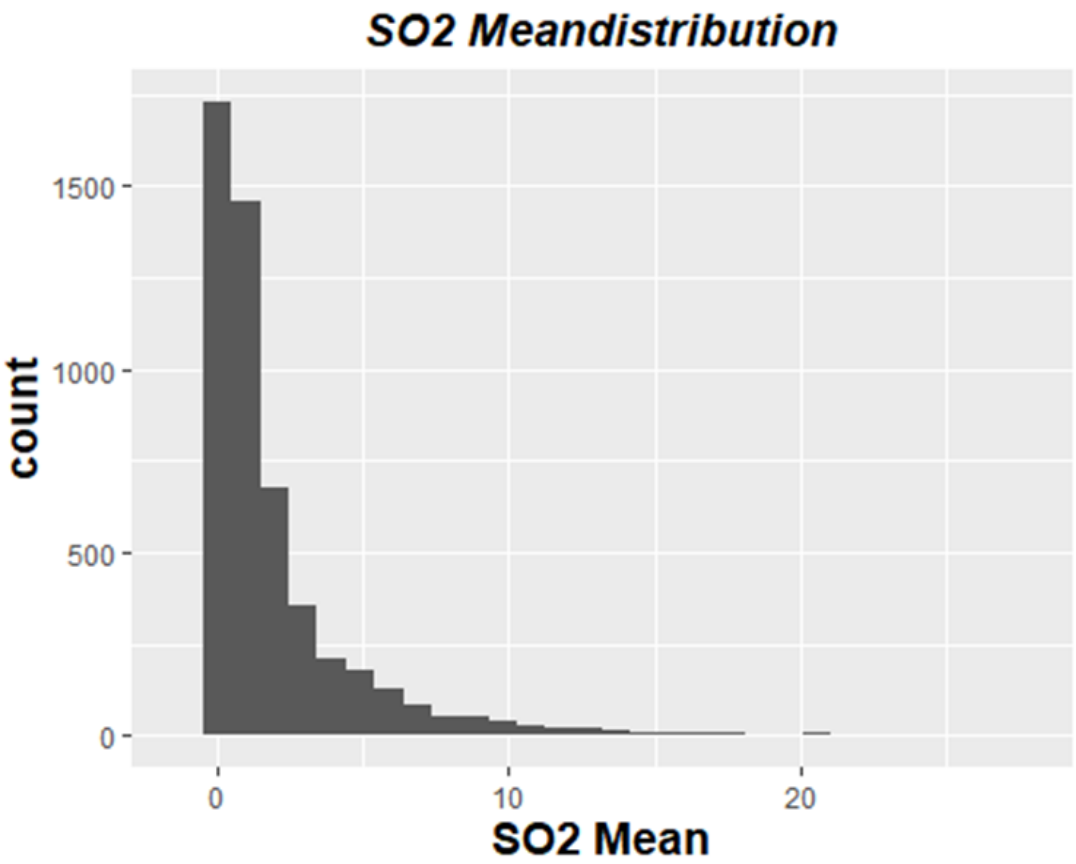
In this graph, we examine the O3 mean for three counties: Bexar, Dallas, and El Paso. El Paso has the greatest O3 mean of 16.45, while Bexar has the lowest O3 mean of 8.62.



D) SO2

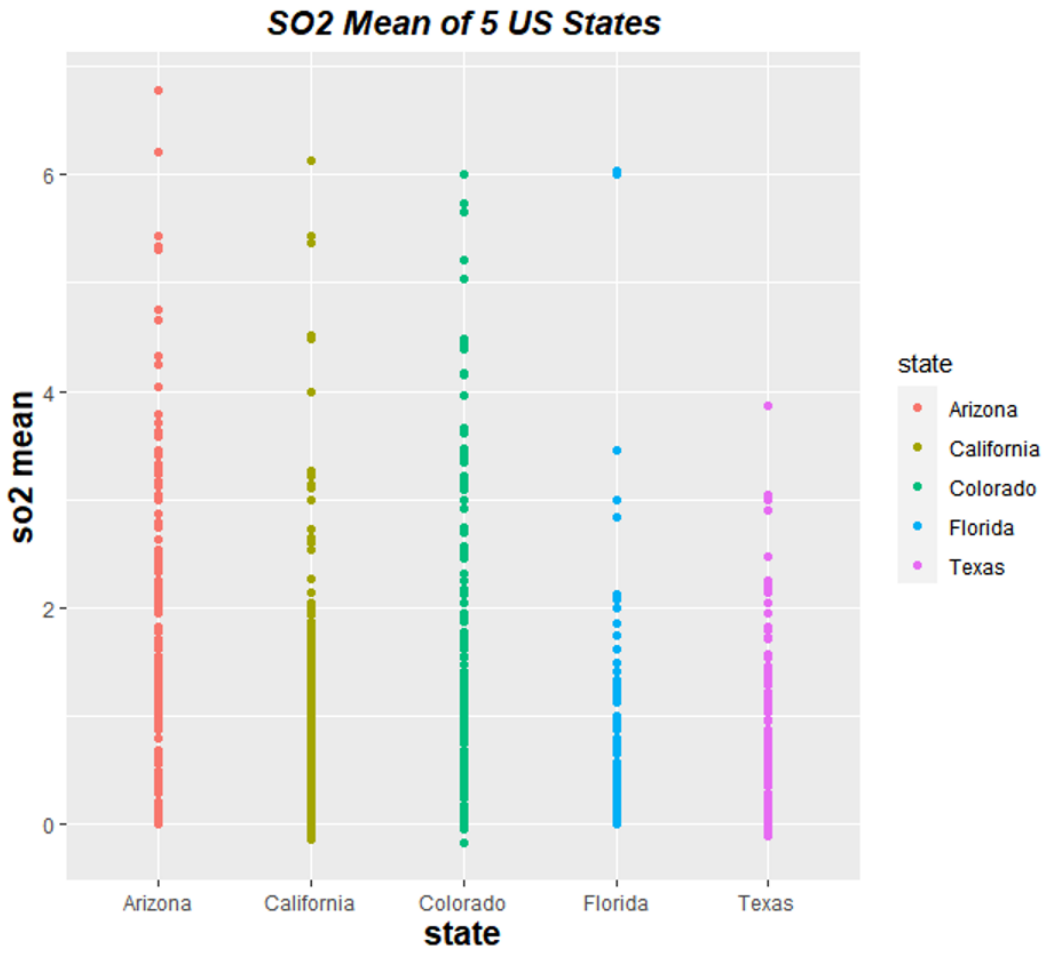
1. Histogram

a. Based on the histogram graph that has been derived, the SO2 mean values are mostly between 0 and 2.



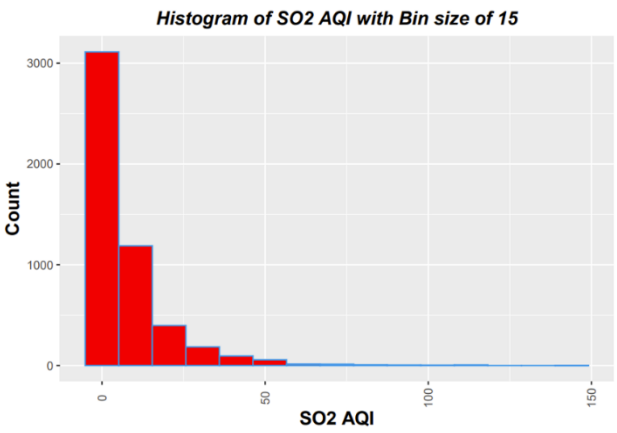
2. Point Graph

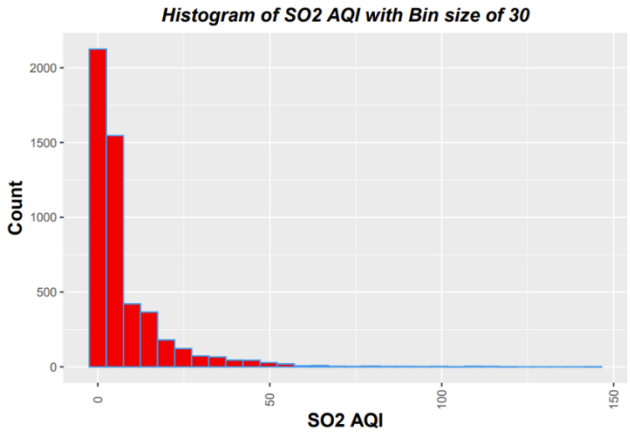
a. Point Graph of SO2 mean values of five states

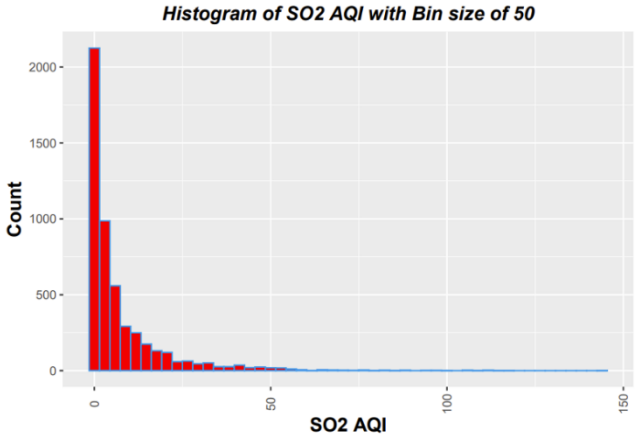


3. Histogram

a. The histograms used in the analysis are shown with different bin sizes and have similar values where the mean, median, and mode are between 0 and 10, range from 0 to 145, and can be represented.





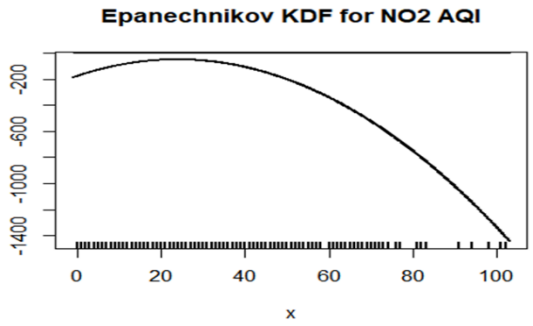


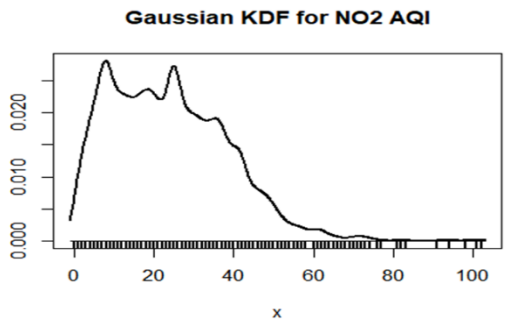
3. RESULTS

I.Kernel Density Distribution -

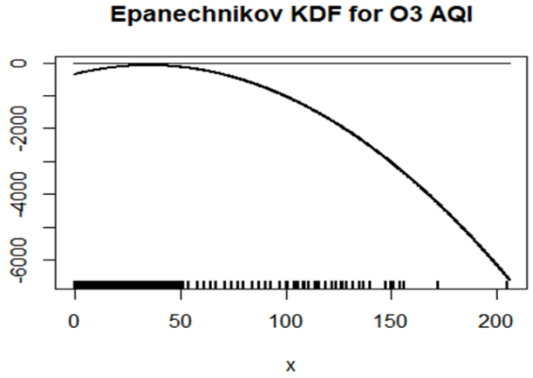
According to the index, any reading between 0-50 indicates good air quality, 51-100 moderate air quality, 101-150 unhealthy for sensitive groups, 151-200 unhealthy air quality, 201-300 very unhealthy air quality, and 301 and above hazardous air quality.

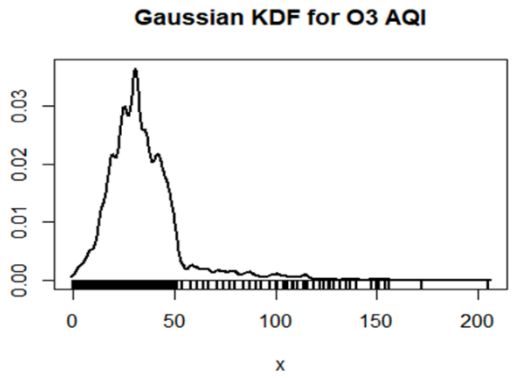
a. An AQI value of 50 or less indicates good air quality, while an AQI value greater than 300 indicates hazardous air quality. For NO2 AQI, the Epanechnikov KDF is smooth, whereas the Gaussian KDF fluctuates. Both KDFs are right skewed, with the majority of NO2 AQI values ranging from 0 to 60, indicating that there are few higher values. The outcomes from the below distributions have a positive impact on the environment.



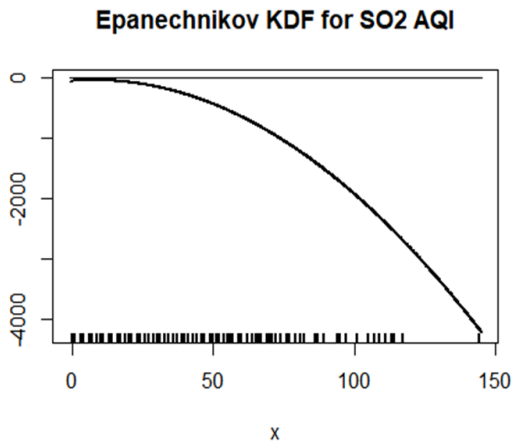


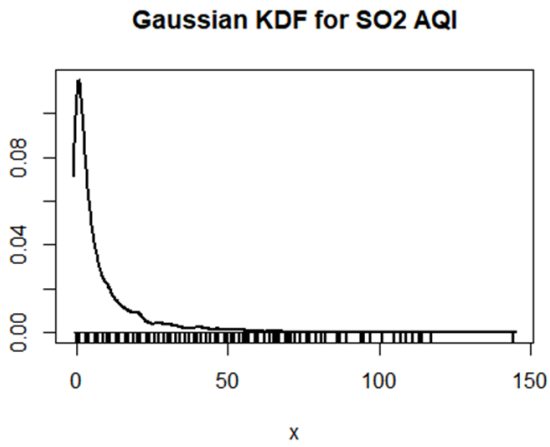
b. Generally, a health alert is issued when the AQI falls between 201 and 300 for O3, indicating that everyone may suffer from more serious health effects. Epanechnikov KDF for O3 AQI is smooth whereas Gaussian KDF has fluctuated. Both the KDFs are right skewed where most of the O3 AQI values are between 0 to 50 which means that the AQI is being reduced over years.



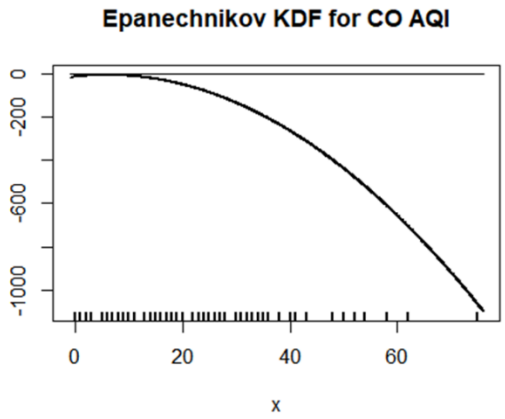


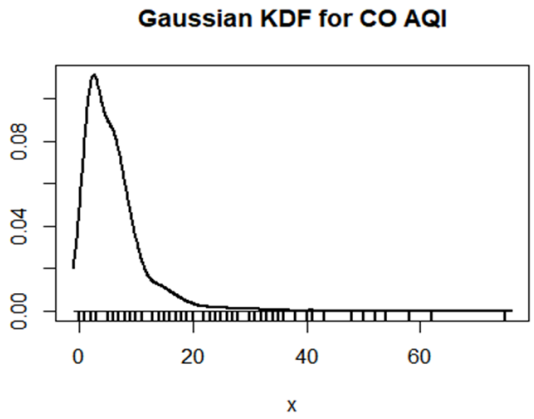
c. SO2 is extremely harmful. A health alert is issued when the AQI falls between 201 and 300, indicating that everyone may suffer from more serious health effects. According to the graphs below, pollution is decreasing on a regular basis. The distribution of the Gaussian KDF for SO2 AQI looks unimodal, and both the KDFs are skewed to the right.



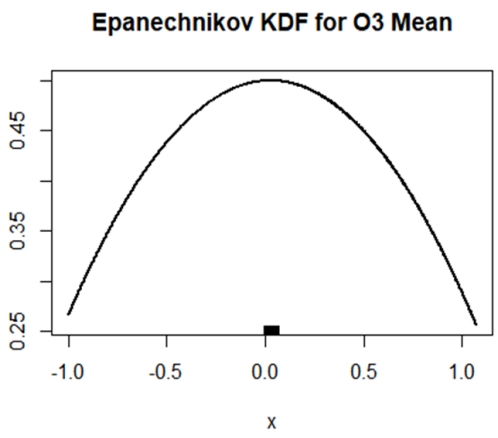


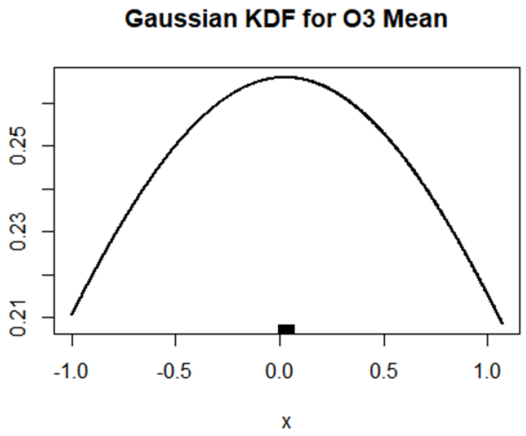
d. Pollution has a positive impact because emissions of air pollutants have gradually decreased. The Gaussian KDF distribution for CO AQI appears unimodal, with both KDFs skewed to the right.



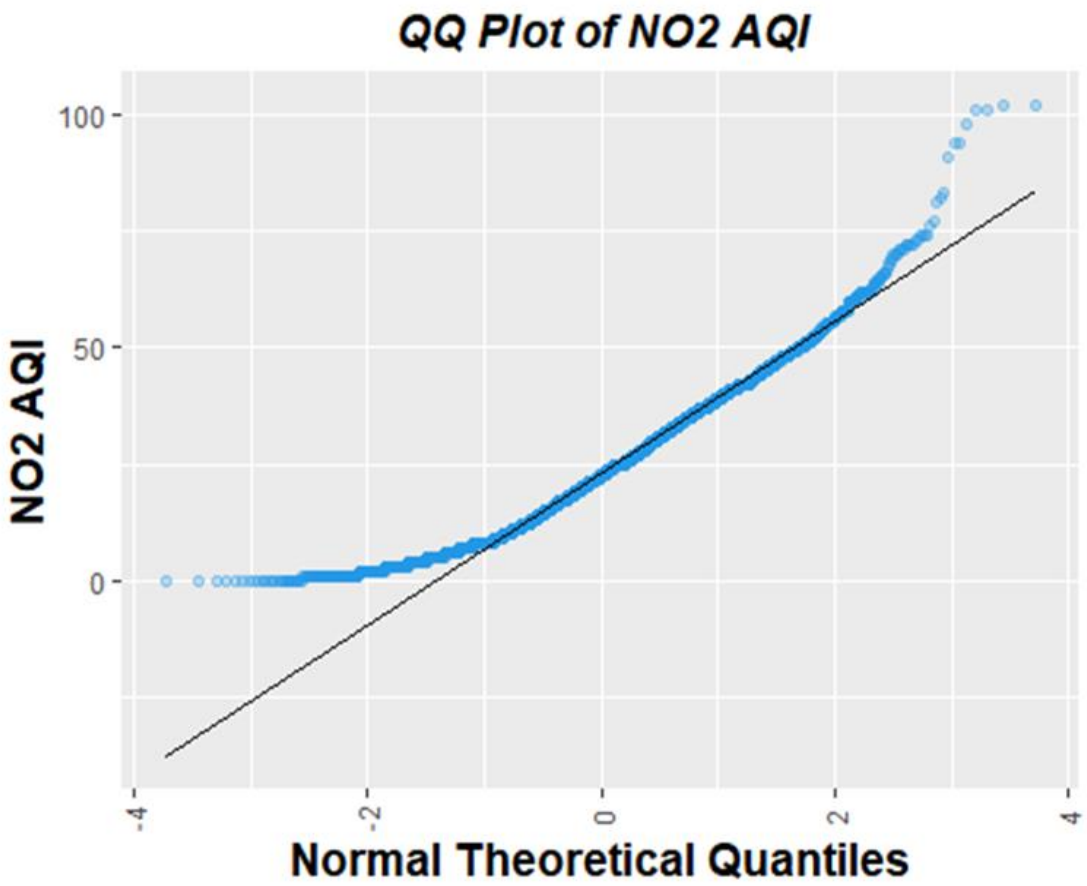


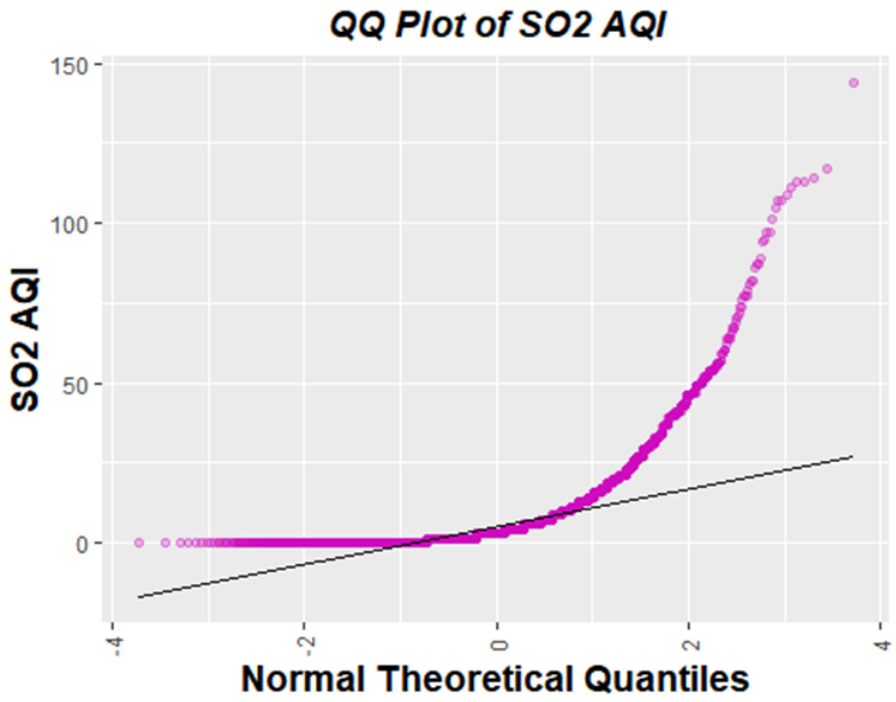
e. The Epanechnikov and Gaussian KDF both have normal distributions with a mean value close to 0. Based on the findings, the values are symmetrical.

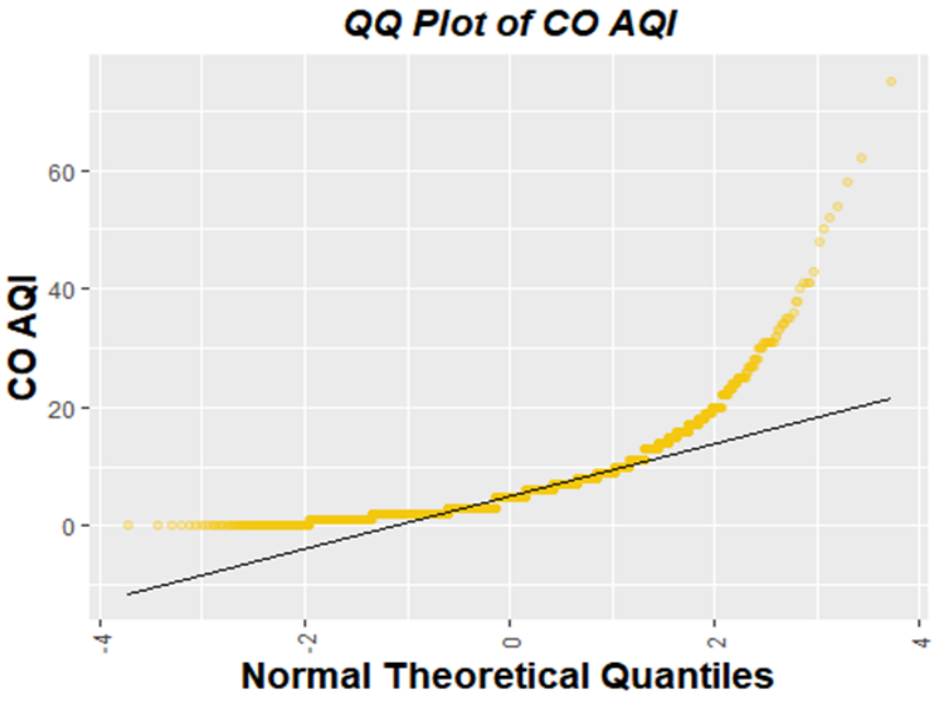

Epanechnikov KDF for O3 Mean
0.45
0.35
0.25
-1.0
-0.5
0.0
0.5
1.0
x

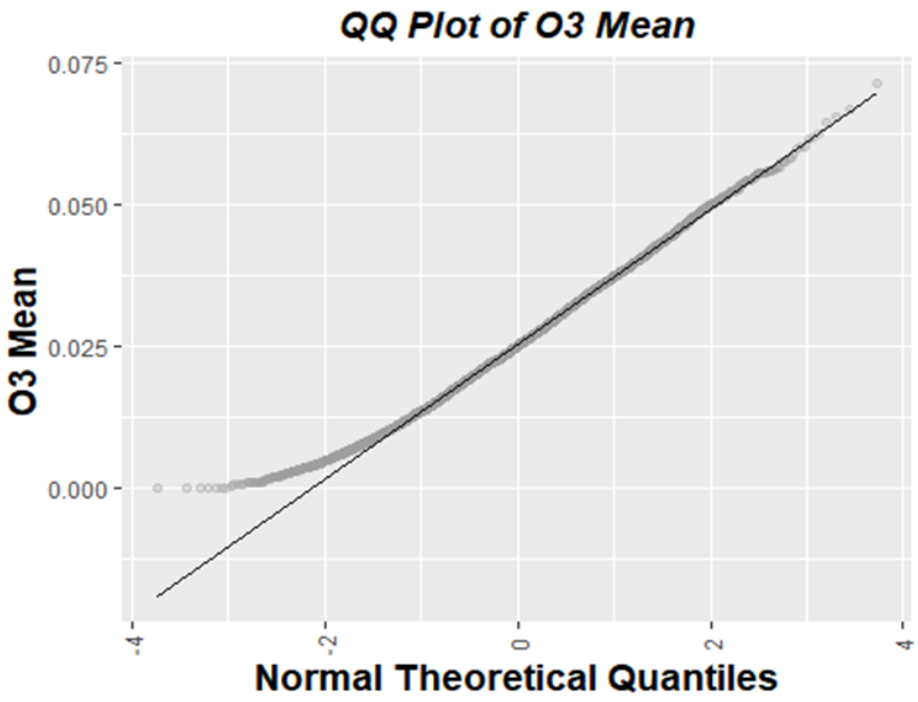

Gaussian KDF for O3 Mean
0.25
0.23
0.21
-1.0
-0.5
0.0
0.5
1.0
x

II. QQ Plot


QQ Plot of NO2 AQI
NO2 AQI
100
50
0
-4
-2
0
2
4
Normal Theoretical Quantiles

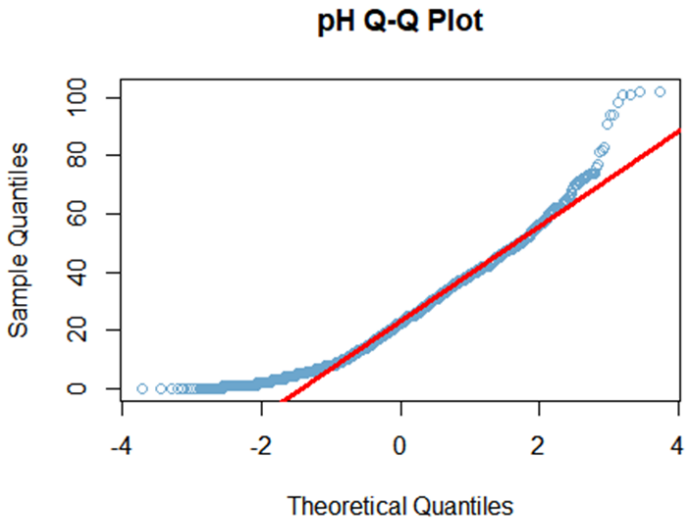

QQ Plot of SO2 AQI
150
100
50
0
SO2 AQI
-4
-2
0
2
4
Normal Theoretical Quantiles


QQ Plot of CO AQI
60
40
20
0
CO AQI
-4
-2
0
2
4
Normal Theoretical Quantiles



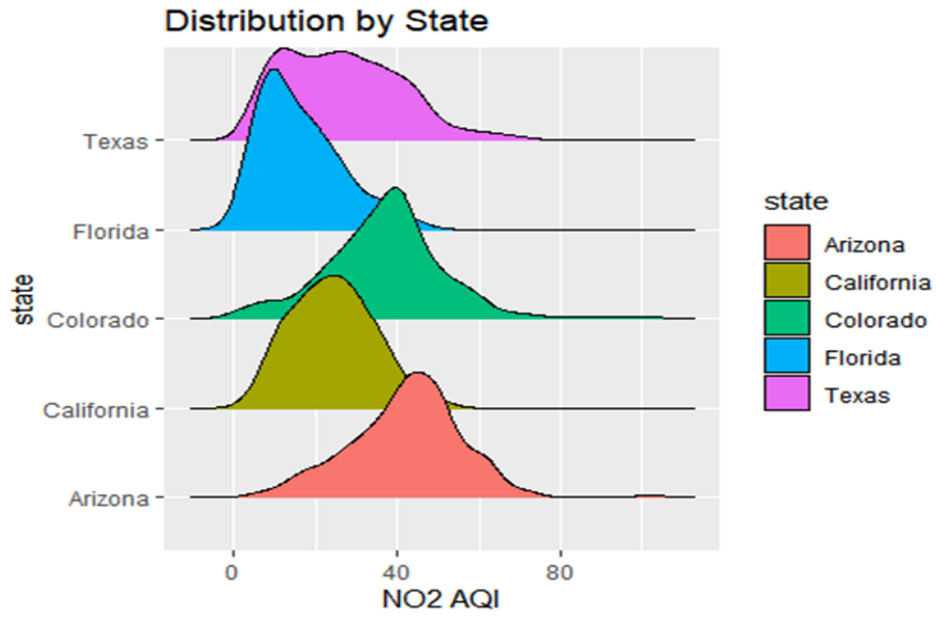
Observations:

- The QQ plot for NO2 AQI, SO2 AQI, and CO AQI is right skewed.
- The QQ plot for O3 is Normally distributed.
- Most of the values are between 2 standard deviations of the mean.



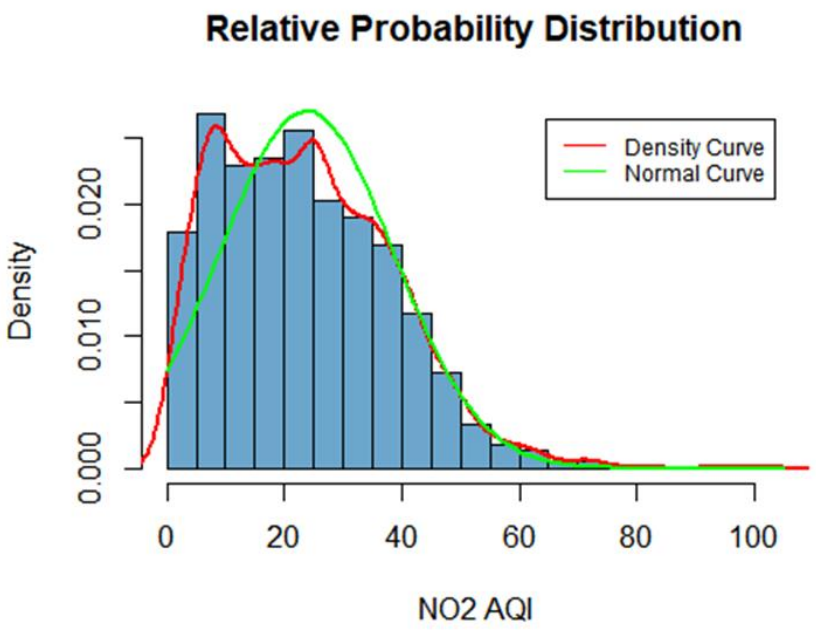
Observations:

- The QQ plot is right skewed.



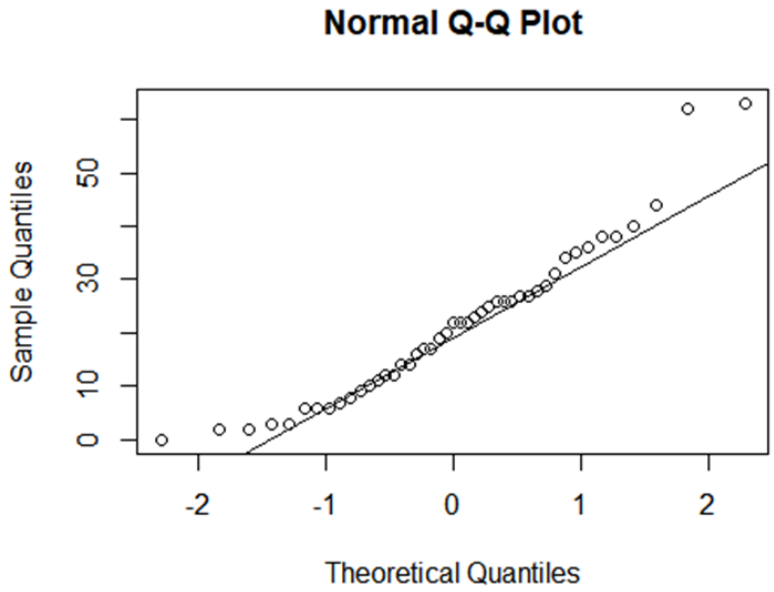
Observations:

- The distribution of NO2 AQI for Texas is widespread. Most of the values are in the range of 0 to 60.
- The distribution of NO2 AQI for Florida is uni-modal. The peak value is around 15.
- The distribution of NO2 AQI for Colorado is Normally distributed. The Mean value lies around 40.
- The distribution of NO2 AQI for California is Normally distributed. The Mean value lies around 25.
- The distribution of NO2 AQI for Arizona is Normally distributed. The Mean value lies around 45.



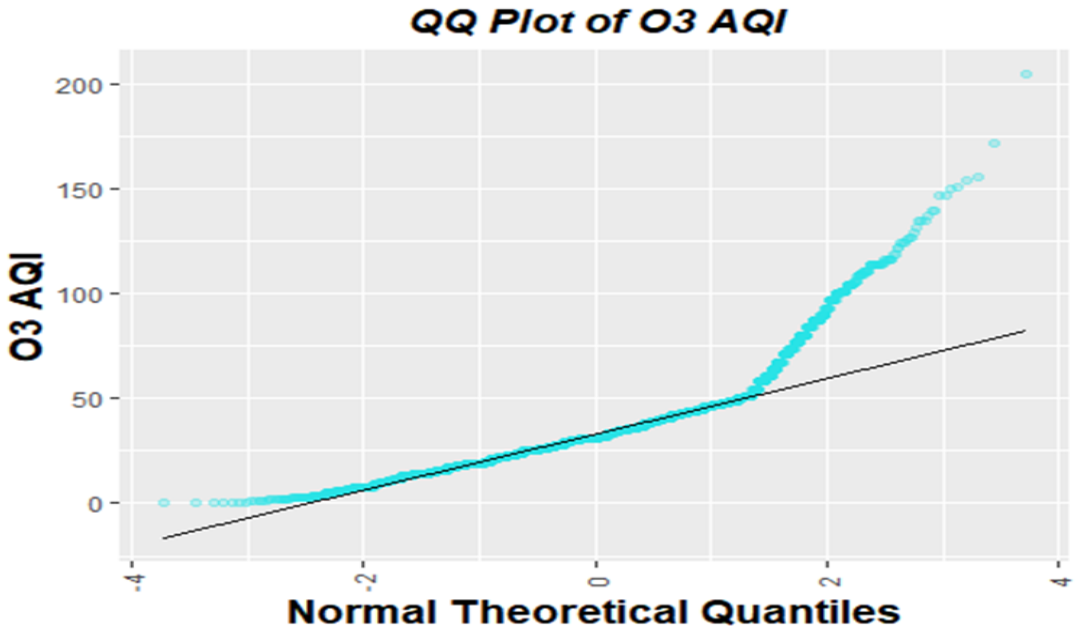
Observations:

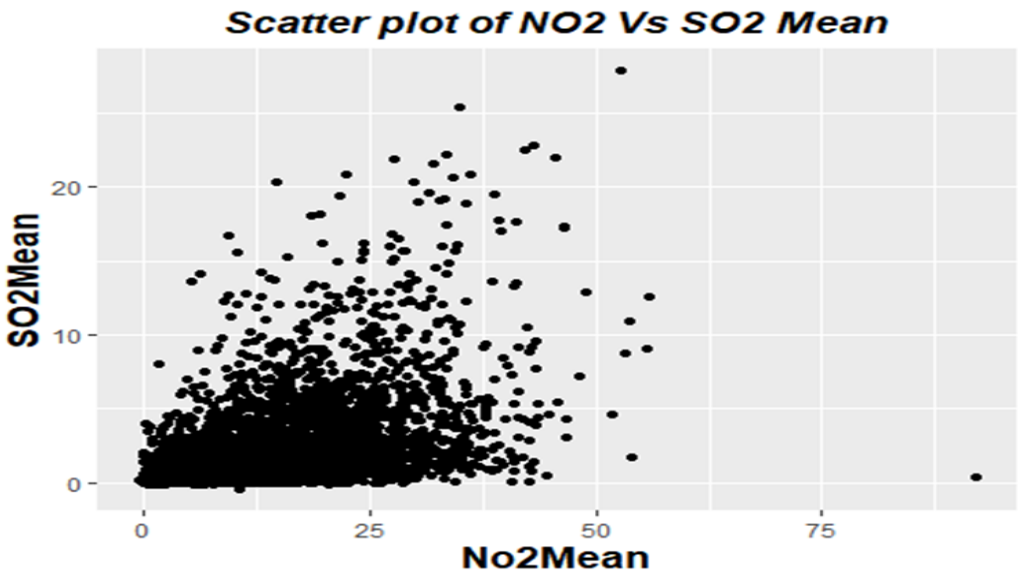
- The probability distribution of NO2 AQI is right skewed.
- Most of the values are in the range of 0 to 60.
- The density curve has a bi-modal distribution and has peak values at 5 and 20.



Observations:

- The Q-Q plot is Normally distributed.
- Most of the values are present in between 1 standard deviation from the mean.





Observations:

- There is a positive correlation between NO2 Mean and SO2 Mean.
- As the value of NO2 Mean increases SO2 value also increases.

III. Hypothesis Testing

The plausibility of a hypothesis is evaluated using sample data in a process known as hypothesis testing. Such information could be derived from a larger population or from a data-generating process. In the following descriptions, "population" will be used to refer to both of these scenarios. A random sample of the population being studied is measured and examined by statistical analysts in order to test a hypothesis. A random population sample is used by all analysts to test two hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis is represented by Ho, and the alternative hypothesis is represented by Ha and is defined as:

Ho: A common statistical theory called the null hypothesis asserts that there is no statistical relationship or significance between two sets of experimental values and measured phenomena for a given single experimental variable.

Ha: The null hypothesis is supplemented by the alternative hypothesis. The exhaustive nature of null and alternative hypotheses ensures that they account for all potential outcomes. As a result, only one of them can be true at once because they are mutually exclusive.

Z-Test:

To proceed with the z-test we would require a sample size that is greater than 30, so we have taken a sample size of 45, with a confidence interval of 0.95. Hence alpha is, $\alpha = 1 - \text{Confidence Interval} = 1 - 0.95 = 0.05$. We get that the value of z is less than the critical value of z (+1.64) from the z table. That means our sample mean value is less than or equal to the population mean. From our one-tailed z-test above, we can infer the same. H - Null is the accepted Hypothesis.

As we are considering the right-tailed z-test (alternative = ‘greater’):

Ho - The sample mean is less than or equal to the population mean (The mean of sample NO2 AQI values is less than or equal to the population mean of NO2 AQI).

Ha – The sample mean is greater than the population mean (This means the sample mean of NO2 AQI values should be greater than the population mean of NO2 AQI), then we can reject the null hypothesis.

$$\text{Z Test} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

T-Test:

For the T-test, we would require a sample size which is less than 30, so we have taken a sample size of 25, and a confidence interval of 0.95. Hence alpha is, $\alpha = 1 -$ Confidence Interval = 1 - 0.95 = 0.05. We get that the value of t is greater than a critical value of t = -1.711 (as df=n-1=24, α=0.05)

That means our sample mean value is greater than or equal to the sample mean taken. From our one-tailed t-test above, we can infer the same. H - Null is the accepted Hypothesis.

As we have considered the left-tailed t-test (alternative = "less"'):

Ho - The sample mean is greater than or equal to the population mean (The mean of sample NO2 AQI values is greater than or equal to the population mean of NO2 AQI).

Ha – The sample mean is less than the population mean (which means the sample mean of NO2 AQI values should be less than the population mean of NO2 AQI), then we can reject the null hypothesis.

t-Test Formula

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Chi Square -Test:

For the Chi Square - test, we are interpreting the relationship between two variables.

We are approximating the results with alpha=0.05

The p-value is less than the alpha value so we reject the Null Hypothesis.

From this it can inferred that, both the samples are significantly different.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

F -Test:

For the F- test, we are interpreting whether the two sample variances are significantly different or not. We are estimating for a confidence interval of 0.95. Hence alpha is, $\alpha = 1 -$ Confidence Interval = 1 - 0.95 = 0.05. We get that the value of F value is greater than the critical value of F.

So we reject the Null Hypothesis and accept the alternate hypothesis.

Ho - The variance is NO2 AQI is equal to the variance of SO2 mean.

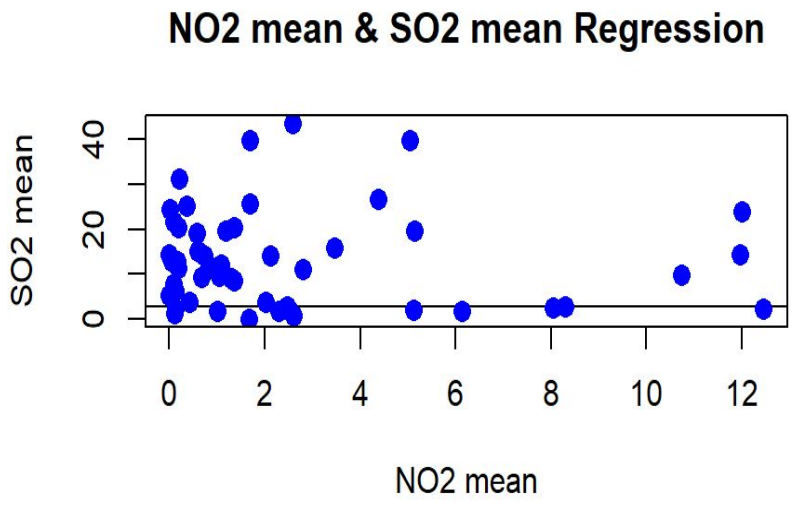
Ha – The variance is NO2 AQI is significantly different to the variance of SO2 mean. .

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

$$where\ \sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

IV. Regression Analysis

Regression analysis is a popular statistical technique for establishing correlations between two variables. One of these variables is known as a predictor variable, and its value is determined through experiments. The other variable is known as the response variable, and its value is determined by the predictor variable.



On comparing the SO2 mean and NO2 mean to find the correlation between the gases, it can be inferred from the graph that there is no specific relationship between the variables.

4. CONCLUSIONS

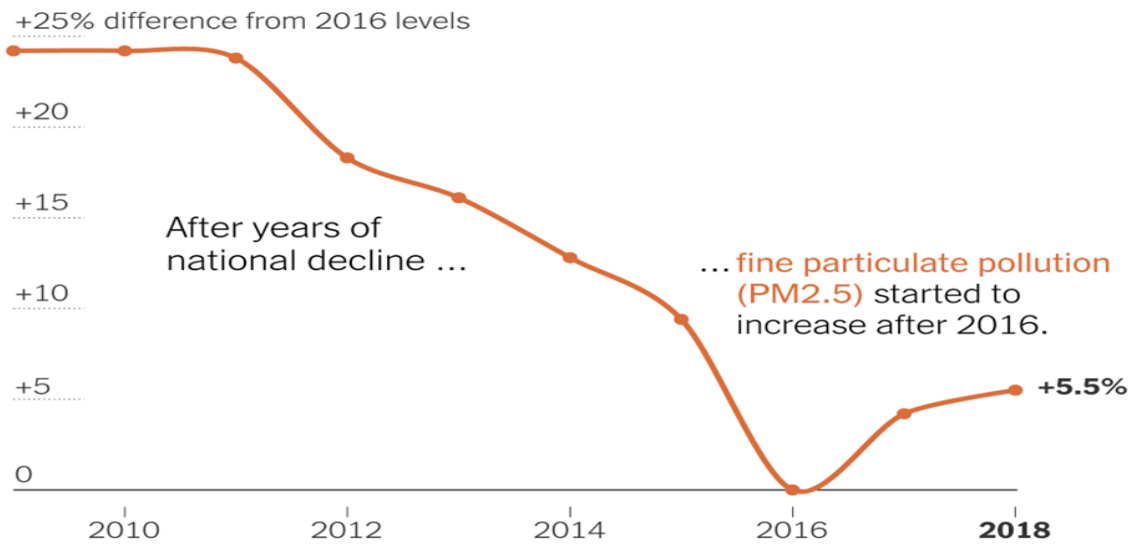


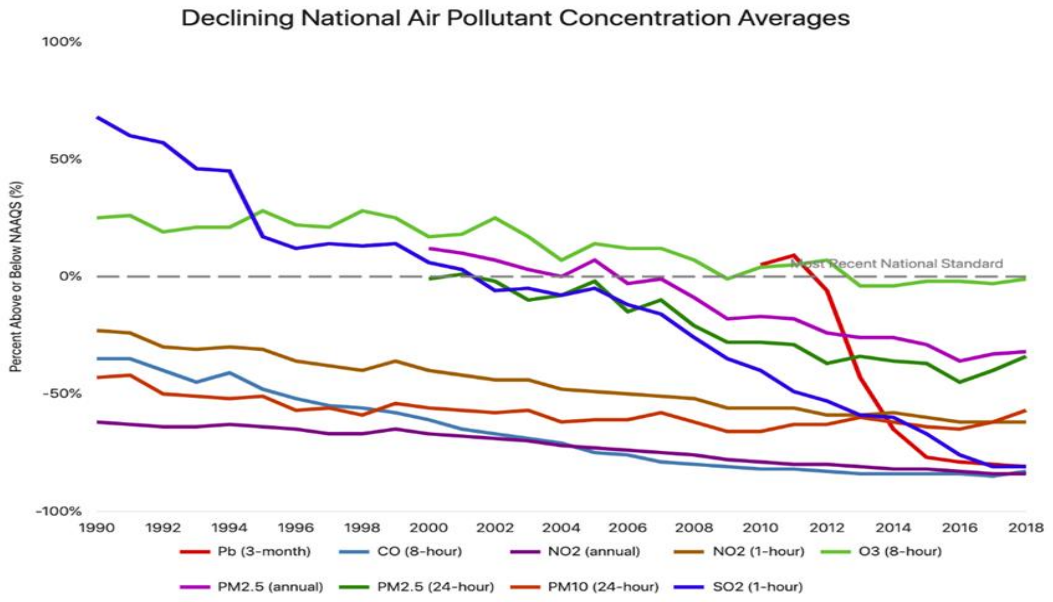
Image Reference Link: https://www.nytimes.com/interactive/2019/10/24/climate/air-pollution-increase.html

The major insight we learned from the analysis is that pollution has been decreasing in the USA over the years. Although New York City's air quality has improved in recent decades, it remains the sixth most polluted city in the United States. It's part of a recent study that found 97% of American communities failed to meet World Health Organization air quality standards. In New York City, nitrogen dioxide is produced by cars, power plants, and industrial pollutants. In contrast, high ozone levels are mostly to blame for New York's poor air quality. When pollutants, nitrogen oxides, and reactive organic compounds from automobile and smokestack combustion react at high temperatures (above 80°F), ozone is generated in the atmosphere. We learned from our previous data and research that Tennessee produces low levels of NO2 gas since air quality refers to the concentration of contaminants in the outside air. Tennessee is located in several prominent geographical features, along with the Appalachian Mountains.

Based on our analysis, we found that in February the air is more polluted than other months. This is because the warm air has a higher moisture content. Since cold air can't hold as much moisture, winter months typically see a decrease in air humidity. One of the more unusual facts about air pollution is that rain can remove a large portion of it. Rain acts like water, washing away the dust in the air. Due to lower precipitation levels in the winter, this is less likely to happen in many locations. Without this purifying effect, the air would continue to be contaminated rather than being able to shed its pollution. By doing so, the natural cycle that removes dust from the air and stops more from accumulating is prevented. Cool, dry air holds more pollution.

Among top 5 major states Arizona, California, Colorado, Texas, Florida we found that Arizona has highest CO mean on the other hand Texas has highest O3 mean but it has lowest CO mean. In major states like Texas, El Paso is the most polluted county. Drought and climate change are exacerbating other variables that contribute to poor air quality in Arizona. The combustion of fuels is a major source of CO emissions. Poor combustion increases CO emissions, but more people are on the road. And those cars produce Greenhouse gases, with road traffic being by far the most significant source, and gasoline-powered vehicles being the most prevalent. The maximum levels of CO often occur during the colder months of the year, when inversion conditions are more common, hence Texas produces less CO gas.

Furthermore, we analysed and found Florida has the highest O3 air quality index whereas Colorado has lowest O3 air quality index. The Spatial Air satisfactory gadget presents the modern-day AQI for all Florida websites equipped with ozone or continuous first-class particle monitors. Even healthy individuals can be harmed by means of ozone pollution, however kids, the elderly, and those tormented by lung conditions like allergies or COPD are at an elevated hazard. Both adults and kids with allergies are prone to attack-inducing ozone exposure. Tampa, St. Petersburg, and Clearwater had extra days with dangerous air and have been the 68th maximum polluted metropolis in the United States for ozone. Whereas, Colorado's ozone air quality index is 88.3 percent, which is lower than the national average of 89.2 percent.



Since 1990, air pollution has decreased, but the decades-long trend may soon be over. Recent environmental rollbacks have resulted in short-term air pollution increases. EPA

Image Reference Link: **https://www.indianaenvironmentalreporter.org/posts/federal-report-indicates-end-of-decades-long-air-quality-improvement**