Course Project Milestone 3- Preliminary Analysis- Air Quality Prediction

Guruprasad Velikadu Krishnamoorthy and Hemalatha Subbiah

College of Science and Technology, Bellevue University

DSC630-T302: Predictive Analysis

Professor. Fadi Alsaleem

January 17, 2024

Air Quality Prediction

Introduction

Air pollution is a significant concern in the modern world. With the rise of the Industrial Revolution in the 19th century and the invention of new machinery and technologies, economies worldwide saw explosive growth in the following decades, accompanied by the migration of farmers from the countryside to fast-growing, crowded cities in search of factory jobs. (Kiger, 2021) While the industrialists were amazed by the profits being made, the governments were astonished by the progress made in the economy; little did they all know the side effects that changed the landscape of the earth forever, for the worse. The emission of gases from the factories and the pollutants from the automobiles had an everlasting impact on the face of the earth. In this project, the effects of air pollution and the ways to combat it by being able to predict it using machine learning algorithms will be discussed in detail.

Problem statement:

On average, it is estimated that a person inhales about 2000 gallons of air daily. Hence, it is essential that the air we breathe is of good quality. The polluted air, when inhaled, gets straight into our lungs, then enters the bloodstream, and can cause more damage to internal organs such as the brain, heart, etc., and young children are the most affected ones.

The aim of the project is to develop a machine-learning model capable of predicting the Air Quality Index(AQI) based on various environmental parameters and pollutant concentrations. The model should be capable of providing accurate AQI predictions for future time points, allowing for early detection of potential air quality issues.

Why the Problem is important:

According to a recent study, air pollution-related ailments are the fourth largest contributing factor to premature deaths, and about 4.5 million deaths around the world are reported to be related to

air pollution. Some air pollutants such as mercury, Lead, and benzene can cause several health issues and, in some cases, even death. (10 Things You Never Knew Could Cause Lung Cancer, n.d.) In New Delhi, India, where the air quality is ranked among the worst, people reported that it felt like breathing poison during extreme smog conditions. The level of PM2.5 pollutant, which is small enough to enter the bloodstream, was reported to be 25 times more than the limits recommended by the World Health Organization in the city. ("Like Breathing Poison": Children in India's Delhi Hit Hard by Smog, n.d.)

Though some measures were taken in the last decade to bring awareness to people about the impacts of air pollution, it was too little, and it will be too late if all the countries around the globe don't work together to solve this looming Problem. If the current trend continues, in the future, people may be forced to pay for clean, breathable air, making it an absolute priority to take measures to clean up the air and provide a sustainable and healthy environment for our future generations.

Who would be interested in solving the Problem?

Pretty much everyone is affected by the rising air pollution-related problems around the world. Children, older adults, and people living with asthma and other breathing-related disorders are the most affected. People who have lower incomes and those who live near the sources of pollution, such as factory workers, may also face greater harm. Though it is believed that people in bigger cities are the most affected due to the vehicle and industrial emissions, small cities and towns are not spared either. Agricultural activities, Fertilizer manufacturing, and livestock production that release Methane can all contribute to air pollution in rural areas. Hence, this is a universal problem that affects most of the population around the globe.

The availability of comprehensive air quality datasets can help develop predictive models that can forecast the Air Quality Index (AQI) with high accuracy. Being able to predict the air quality based on the pollutant concentration can not only alert the public to take precautions but can also help take measures in advance to be prepared or to avoid the situation altogether.

Source of the Data:

The data is extracted from the Environmental Protection Agency website. (Download Files | AirData | US EPA, 2015). A total of 13 datasets will be used to predict the air quality in this project, such as Ozone concentration datasets, SO2, NO2, and Carbon Monoxide concentration datasets; datasets of pollutants such as PM2.5 and PM10; datasets containing meteorological data such as Temperature, Pressure, Humidity, etc. All these datasets are for the year 2022.

Some of the commonly used fields in the dataset are listed below:

- Columns to identify the site location where the readings were measured:
    - State code, County Code, Site Number, Latitude, Longitude, Local Site name, Address, State Name, County Name, City Name.
- Columns to identify the Pollutant/ Gases/ the Metrological quantity:
    - Parameter Name: Represents the Parameter of the Pollutant /Particulate/Toxin/Meteorological measure.
    - Parameter Code: Unique code assigned to the parameter describing the Pollutant /Particulate / Toxin /Meteorological measure.
    - Units of Measure: The unit in which the parameter is measured.
    - Observation Count: Number of observations captured on the specified date in the site location.
    - Arithmetic Mean: The Mean value of the quantity of the parameter captured on the given date at the given site.
- Column to identify the Air Quality:
    - AQI: Represents Air Quality Index value measured on the specified date at the specified site location.

- Date fields:
    - Date Local: Date when the parameter was measured and recorded.
- Columns for Pollutants:
    - Ozone (O3): Concentration of ozone in the air
    - Carbon Monoxide (CO): The Concentration of carbon monoxide in the air.
    - Nitrogen Dioxide (NO2): Concentration of nitrogen dioxide in the air.
    - Sulfur Dioxide (SO2): Concentration of sulfur dioxide in the air.

Why is the data useful to the Problem:

The datasets contain information about the concentration of Gases such as Ozone, Sulfur Dioxide (SO2), and Nitrogen Dioxide (NO2) measured at different sites around the US for the year 2022. This, along with the concentration of particulates such as PM2.5 and PM10.0, and other parameters such as Temperature, pressure, and humidity, can provide valuable information to predict the Air Quality index. As all these metrics are available in individual datasets, they must be combined to extract useful insights from them. There are about 250,000 rows on average in these datasets, thus providing sufficient data required for the research.

Type of Models to be used:

The target variable that will be predicted in this machine learning project is the Air Quality Index (AQI). As it is a continuous numeric quantity, Regression algorithms will be used to predict the air quality index. Various regression algorithms, such as Linear Regression and Regularization techniques, such as Lasso and Ridge algorithms, will be used in this project.

Also, the project can be expanded to include classification models like Random Forest, Gradient Boosting, K-Neighbor, and Decision Tree to predict the Levels of Health concern based on the range of AQI values.

Plan to evaluate the results:

1. The 13 datasets used in this project will be loaded into their respective Data Frames and then combined based on the common keys.
2. Exploratory Data Analysis will be performed on the combined data frame to handle null values and treat the outliers. In this step, correlation between the variables in the dataset will also be measured.
3. Visualize the dataset to gain insights, identify patterns, and make informed decisions about preprocessing and modeling. More Visualization types like Box plots, Scatter plots, Correlation heat maps, and Confusion Matrix will be incorporated.
4. Features and Target variables will be declared, and the dataset will then be split into train and test sets. During this step, feature selection and extraction techniques will be deployed to reduce the number of features and extract useful information from the features.
5. The regression model algorithms will then be used to build the prediction model. The accuracy of the model and the confusion matrix to demonstrate the performance of the model with predicted vs actual values will be computed.
6. Root Mean Square Error (RMSE), which measures the difference between the actual and predicted AQI values, will be used as an evaluation metric, while R-squared, which explains the proportion of the variance in the AQI, will be used to evaluate the results.
7. Precision, recall, and F1 scores can offer insights into the model's behavior.
8. Also, cross-validation techniques such as K-fold validation will be deployed to reduce overfitting and to improve the performance of the model on unseen data.

What do I hope to learn?

As the project involves combining the data from various datasets with approximately 250,000 records in each, handling huge volumes of data and combining them without causing performance issues is something that we hope to learn from this project. As the dataset resembles a real-life dataset, this project can help us learn to overcome the challenges of handling large volumes of data at different stages of the project, such as Data Preparation, Model Building, Cross-validation, etc. Moreover, from the dataset, I want to understand how pollutant concentrations detected by the sensors are highly correlated with one another and want to analyze the real effect of it.

Risks:

One potential risk of this project is the quality and availability of the data available across the 13 datasets. During the EDA and Data Preparation phase, the data must be carefully studied and joined using common fields such as Site Location and Dates. There is a risk that the data may not be available for all site locations across the different datasets, thus resulting in a much smaller dataset than intended after all joins are performed. Also, the presence of nulls in the essential columns in these datasets can pose an additional risk, resulting in data reduction.

Ethical Implications:

While studying the impacts of Air pollution and its relationship to global warming and the climate change, the ethical implications should be carefully assessed. Some of them are listed below:

- Though gas-powered vehicle emissions and industrial smoke have played a significant role in air pollution, they cannot be entirely replaced by sustainable solutions, as they can lead to many job losses affecting many families employed by the manufacturing industries. Care must be taken while publishing the results, keeping in mind the impact it can have on the families.

- While determining the acceptable levels of greenhouse gases and pollutants for humans, careful assessment should be made while determining the values, as the acceptable levels for humans may cause significant damage to other ecosystems and species. (Brown, 2001)
- The acceptable levels should also be carefully assessed with international considerations in mind, as the gas emissions and pollutants from the developed countries are no longer a local issue. These impacts are already seen on the other side of the world, with extreme floods and drought conditions that were not seen in the past.

Contingency plan:

If we run into problems while combining the datasets or if the performance of the model lags due to a large volume of data, the contingency plan is to reduce the number of datasets. However, the downside of this approach is that some useful information presented by the excluded datasets will be lost and can affect the model performance. If we still face issues, a simplified version of the Air quality dataset with about 9000 rows from Kaggle will be used to continue with the project. (Air Quality Dataset, n.d.) or from the UCI Machine Learning Library. (UCI Machine Learning Repository: Air Quality Data Set, n.d.).

Milestone 3- Preliminary Analysis

Can the questions be answered with the existing data?

The different datasets used in this project contain details about the concentration of harmful gases, Particulates, and other air pollutants. Also, datasets about meteorological parameters such as wind speed, Temperature, and Pressure will be used in the project. In total, 13 datasets from the year 2022 are in the project, and they contain plenty of useful data to build a model. If needed, datasets about other toxic materials and Lead can also be included in the analysis.

What Visualizations can be useful in explaining the data?

As most of the features used in the project are numerical, scatter plots to show the relationship between the pollutants or harmful gases versus the AQI (Air Quality Index) can be a useful visualization to understand the relationship between the two. Also, different scatter plots can be plotted for each category to understand which pollutant has a greater effect on the air quality. For instance, a scatter plot with NO2 (Nitrogen Dioxide), SO2 (Sulphur dioxide), and CO (Carbon Monoxide) concentrations in the same plot can help us understand which gases have more impact on the air quality. Also, Box plots can be used to identify the outliers, and based on the nature of the outlier, a decision can be made to either eliminate them or impute the data. The dataset also contains the date when the measurement was taken, which is another important feature on which many visuals can be built. A bar plot showing the average Air Quality Index for each month of the year can help us understand the months when most pollution occurs. Heat maps can also be used to show similar data for each state or region. Also, the dataset contains information about the state and city, which can be used to plot on the US Map to show the changes in the air quality index for each month.

Changes to the Data or driving questions?

Datasets about the lead concentration and other toxins are not included in the project. If required, they can be included as an expansion as the project progresses. As the dataset contains details of the state and county, it will be useful to include the region information by joining it with the state-to-region dataset. (USA States to Region, n.d.)

The main goal of the project is to be able to predict the Air Quality index using various pollutants and meteorological factors, and there are no changes to it. However, some enhancements can be made to it by providing more details about the pattern in the AQI for a region or a state.

Changes to the model or evaluation choices?

The Air Quality Index, being a continuous numeric quantity, will be used to predict the AQI values using a regression algorithm. Hence, no changes are expected to the original model choices. However, the project can be extended to a classification model by creating categories of AQI per the values in Table 1 (US EPA, 2016). Furthermore, the project can also be extended, and an unsupervised model, such as the Clustering model, can be built to identify different clusters in the data. The clusters can also be plotted on the US Map for visualization, which will provide a better idea about the regions where the Air quality is bad.

Air Quality Index (AQI) measured	Health concern Level	Colors
0-50	Good	Green
51-100	Moderate	Yellow
101-150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

(Table 1)

Are the Original expectations still reasonable?

The original expectation to build a model to predict the Air Quality Index values is still reasonable. However, based on the time available, the scope of further expansion of the project to include a classification and a clustering model must be determined. Another potential challenge is the volume of the data, which can slow down the performance of the model. As 13 datasets have been identified so far, caution must be taken while joining the datasets without causing significant performance issues.

EDA using R

```
# Calling the Libraries used
library(readxl)
library(dplyr)
library(lubridate)
library(readr)
library(ggplot2)
library(ggthemes)
library(tidyr)
library(DT)
library(scales)
library(stringr)
library(knitr)
library(FactoMineR)
library(ggpubr)
library(kableExtra)
library(magrittr)
library(ggfortify)
```

```
# Reading the data from CSV file and loading into a
# Dataframe
ozone_df <- read.csv("daily_ozone_2022.csv")
# Reading few datasamples from the CSV file
ozone_df <- sample_n(ozone_df, 5000)
# Printing the Dimensions of the Dataframe
dim(ozone_df)
```

```
## [1] 5000   29
```

```
# Examining the structure of the Dataframe
str(ozone_df)
```

```
## 'data.frame':    5000 obs. of  29 variables:
##  $ State.Code         : int  4 1 36 48 29 8 48 13 8 1 ...
##  $ County.Code        : int  17 73 81 29 137 123 201 97 41 73 ...
##  $ Site.Num           : int  119 6002 124 59 1 13 417 4 16 1010 ...
##  $ Parameter.Code     : int  44201 44201 44201 44201 44201 44201 44201 44201 44201 44201 ...
##  $ POC                : int  1 1 1 1 2 1 1 1 1 1 ...
##  $ Latitude           : num  34.8 33.6 40.7 29.3 39.5 ...
##  $ Longitude          : num  -109.9 -86.8 -73.8 -98.3 -91.8 ...
##  $ Datum              : chr  "WGS84" "WGS84" "WGS84" "WGS84" ...
##  $ Parameter.Name     : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
##  $ Sample.Duration    : chr  "8-HR RUN AVG BEGIN HOUR" "8-HR RUN AVG BEGIN HOUR" "8-HR RUN AVG BEGIN
##  $ Pollutant.Standard : chr  "Ozone 8-hour 2015" "Ozone 8-hour 2015" "Ozone 8-hour 2015" "Ozone 8-hou
##  $ Date.Local         : chr  "2022-07-15" "2022-08-21" "2022-10-23" "2022-04-18" ...
```

```
##  $ Units.of.Measure   : chr  "Parts per million" "Parts per million" "Parts per million" "Parts per m
##  $ Event.Type         : chr  "None" "None" "None" "None" ...
##  $ Observation.Count  : int  17 17 17 17 17 17 17 17 17 17 ...
##  $ Observation.Percent: num  100 100 100 100 100 100 100 100 100 100 ...
##  $ Arithmetic.Mean    : num  0.0519 0.0159 0.0331 0.0368 0.0217 ...
##  $ X1st.Max.Value     : num  0.061 0.02 0.035 0.046 0.022 0.051 0.028 0.036 0.062 0.05 ...
##  $ X1st.Max.Hour      : int  9 8 7 11 7 10 9 17 10 10 ...
##  $ AQI                : int  71 19 32 43 20 47 26 33 74 46 ...
##  $ Method.Code        : int  47 NA 87 87 47 87 NA 47 87 87 ...
##  $ Method.Name        : chr  "INSTRUMENTAL - ULTRA VIOLET" " - " "INSTRUMENTAL - ULTRA VIOLET ABSORP
##  $ Local.Site.Name    : chr  "Petrified Forest NP - South Entrance" "Tarrant Elementary School" "QUE
##  $ Address            : chr  "PET FOR NAT PARK, NEAR OLD SW ENTERANCE ON OLD ROUTE 180" "TARRANT, EL
##  $ State.Name         : chr  "Arizona" "Alabama" "New York" "Texas" ...
##  $ County.Name        : chr  "Navajo" "Jefferson" "Queens" "Bexar" ...
##  $ City.Name          : chr  "Not in a city" "Tarrant (corporate name for Tarrant City)" "New York"
##  $ CBSA.Name          : chr  "Show Low, AZ" "Birmingham-Hoover, AL" "New York-Newark-Jersey City, NY
##  $ Date.of.Last.Change: chr  "2023-03-16" "2023-04-13" "2023-02-05" "2023-03-20" ...
```

```r
# Examining the Dataframe, it appears come columns should
# be converted to Factors and dates
ozone_df$Parameter.Code <- as.factor(ozone_df$Parameter.Code)
ozone_df$Units.of.Measure <- as.factor(ozone_df$Units.of.Measure)
# Handling Date columns
ozone_df$Date.Local <- as.Date(ozone_df$Date.Local)
ozone_df$Date.of.Last.Change <- as.Date(ozone_df$Date.of.Last.Change)
# Printing sample rows from Dataframe
kbl(head(ozone_df[1:6, c(1:8)]), caption = "Ozone Pollutant Data",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

Table 1: Ozone Pollutant Data

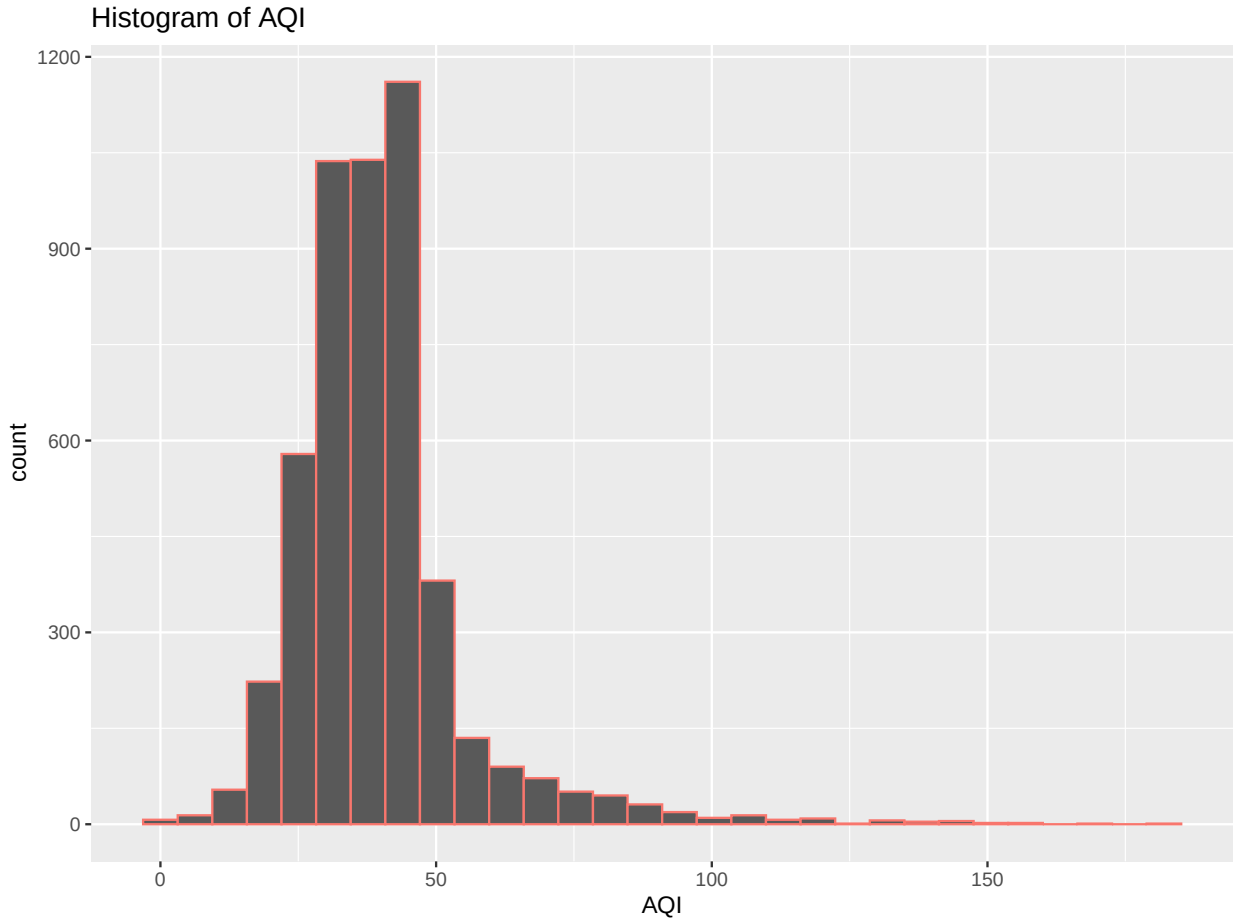
State.Code	County.Code	Site.Num	Parameter.Code	POC	Latitude	Longitude	Datum
4	17	119	44201	1	34.82251	-109.89249	WGS84
1	73	6002	44201	1	33.57833	-86.77389	WGS84
36	81	124	44201	1	40.73614	-73.82153	WGS84
48	29	59	44201	1	29.27538	-98.31169	WGS84
29	137	1	44201	2	39.47498	-91.78899	WGS84
8	123	13	44201	1	40.18162	-104.72613	NAD83

```r
# Checking for Nulls in the Ozone Dataframe
colSums(is.na(ozone_df))
```

```
##          State.Code         County.Code            Site.Num       Parameter.Code
##                   0                   0                   0                   0
##                 POC            Latitude           Longitude               Datum
##                   0                   0                   0                   0
##      Parameter.Name     Sample.Duration  Pollutant.Standard          Date.Local
##                   0                   0                   0                   0
##    Units.of.Measure          Event.Type   Observation.Count Observation.Percent
```

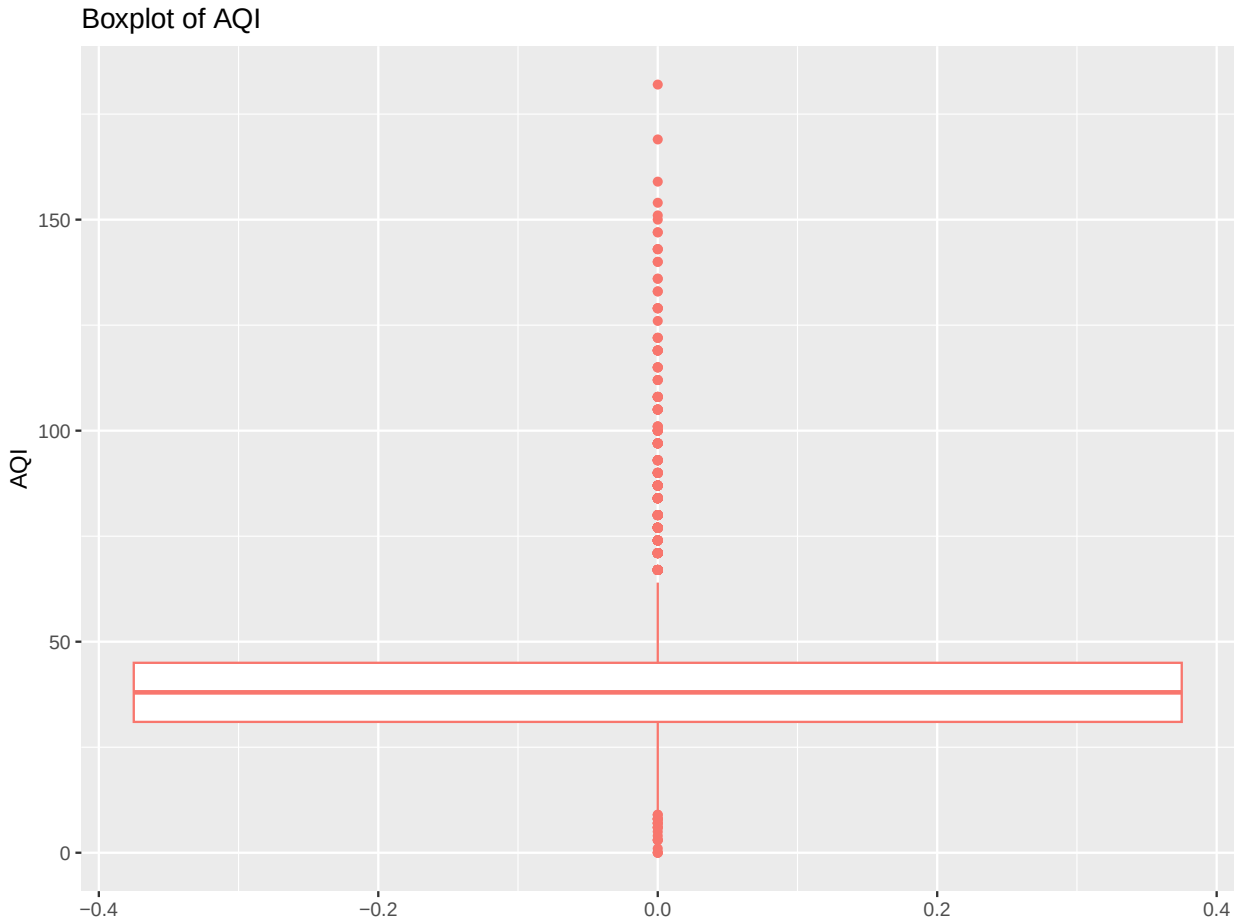
```
##                   0                   0                   0                   0
##      Arithmetic.Mean      X1st.Max.Value       X1st.Max.Hour                 AQI
##                   0                   0                   0                   0
##         Method.Code         Method.Name     Local.Site.Name             Address
##                1672                   0                   0                   0
##          State.Name         County.Name           City.Name         CBSA.Name
##                   0                   0                   0                   0
## Date.of.Last.Change
##                   0
```

```r
# Plotting Histogram to examine the distribution of AQI
ggplot(ozone_df, aes(x = AQI)) + geom_histogram(aes(color = "red")) +
    ggtitle(label = "Histogram of AQI") + theme(legend.position = "none")
```



Histogram Results: The Histogram results indicate that the AQI is heavily distributed between 20 and 65 and has a long tail on the right indicating outliers.

```r
# Plotting Boxplot to examine the distribution of AQI
ggplot(ozone_df, aes(y = AQI)) + geom_boxplot(aes(color = "red")) +
    ggtitle(label = "Boxplot of AQI") + theme(legend.position = "none")
```



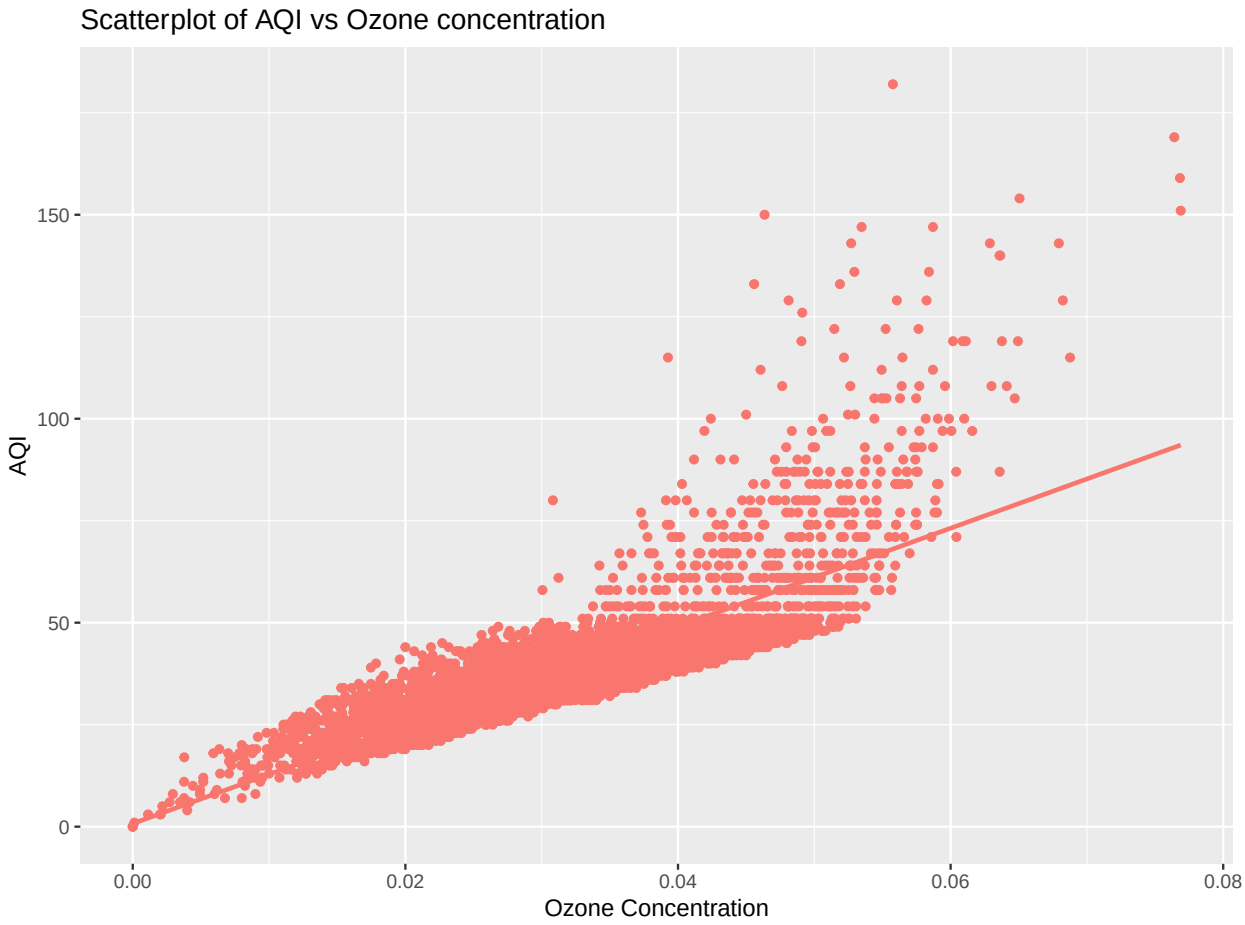
Boxplot Results: The Boxplot results indicate that the presence of Ouliers above an AQI of 65 and also below 20. Most of the Data is between 20 and 65.

```
# Computing the first and third quartiles
q1 <- quantile(ozone_df$AQI, 0.25)
q3 <- quantile(ozone_df$AQI, 0.75)
iqr <- q3 - q1
# Calculate the lower and upper cutoffs for outliers
lower <- q1 - 1.5 * iqr
upper <- q3 + 1.5 * iqr
# Filter AQI to find outliers
AQI_outliers <- ozone_df %>%
    filter(AQI > upper | AQI < lower)
# Printing the top few rows of Outliers
kbl(head(AQI_outliers[1:6, c(1:8)]), caption = "Outliers in the AQI",
    booktabs = T) %>%
    kable_styling(latex_options = c("striped", "hold_position"))
```

```
ggplot(ozone_df, aes(x = Arithmetic.Mean, y = AQI, color = "red")) +
    geom_point() + geom_smooth(method = "lm", se = FALSE) +
    ggtitle(label = "Scatterplot of AQI vs Ozone concentration") +
    xlab("Ozone Concentration") + ylab("AQI") + theme(legend.position = "none")
```

Table 2: Outliers in the AQI

State.Code	County.Code	Site.Num	Parameter.Code	POC	Latitude	Longitude	Datum
4	17	119	44201	1	34.82251	-109.89249	WGS84
8	41	16	44201	1	38.85310	-104.90129	WGS84
40	43	860	44201	1	36.15841	-98.93197	WGS84
6	71	306	44201	1	34.51096	-117.32554	NAD83
6	37	9033	44201	1	34.66974	-118.13051	NAD83
6	65	12	44201	1	33.92086	-116.85841	WGS84



Scatterplot Results: The Scatterplot indicates a strong positive relation between the Ozone Concentration and AQI.

```
# Calculating the correlation coefficient
cor(ozone_df$Arithmetic.Mean, ozone_df$AQI)
```

```
## [1] 0.805749
```

Correlation Results: The output of the correlation coefficient indicates a Strong positive correlation (0.8) between the Ozone concentration and the Air Quality Index values.

References:

Kiger, P. (2021, November 9). Seven negative effects of the Industrial Revolution. History. https://www.history.com/news/industrial-revolution-negative-effects

Air quality. (n.d.). Transport & Environment. https://www.transportenvironment.org/challenges/air-quality/

10 Things You Never Knew Could Cause Lung Cancer. (n.d.). WebMD. https://www.webmd.com/cancer/ss/slideshow-lung-cancer-surprising-causes

"Like breathing poison": Children in India's Delhi hit hard by smog. (n.d.). Al Jazeera. Retrieved December 8, 2023, from https://www.aljazeera.com/gallery/2023/11/10/photos-like-breathing-poison-delhi-children-hardest-hit-by-smog#:~:text=Many%20have%20asthma%20or%20pneumonia

Download Files | AirData | US EPA. (2015). Epa.gov. https://aqs.epa.gov/aqsweb/airdata/download_files.html

Air Quality Dataset. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set

UCI Machine Learning Repository: Air Quality Data Set. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/ml/datasets/Air+Quality

Brown, D. (2001). The Ethical Dimensions of Global Environmental Issues. American Academy of Arts & Sciences. https://www.amacad.org/publication/ethical-dimensions-global-environmental-issues

US EPA, O. (2016, August 30). Air Data Basic Information. US EPA. https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information

USA States to region. (n.d.). Www.kaggle.com. Retrieved January 13, 2024, from https://www.kaggle.com/datasets/omer2040/usa-states-to-region