EE798Q : Fundamentals of Inferential Statistics & Automation

Project Report

Submitted To - Prof. Tushar Sandhan

Submitted by - Siddharth Pathak (Roll no. - 211034)

2023-06-24

Abstract

The accurate collection and analysis of sensory data play a crucial role in understanding environmental conditions and their impact on air quality. However, various factors such as sensor failures, communication link issues, and data packet loss can result in missing values within the sensory dataset.

This project aims to address these challenges by implementing advanced techniques for analyzing and imputing missing values in multivariate time-series data. The dataset used in this study consists of a tabular representation of sensory data, including measurements such as PM2.5, NO, NO2, NOX, CO, SO2, NH3, Ozone, and Benzene concentrations. Missing values are indicated by "NA" entries, either at the row level indicating complete link failure or at the column level indicating individual sensor mishaps. To analyze the dataset, we employ various graph plotting techniques to visualize the time-series data.

However, the presence of missing values poses challenges in accurately representing the temporal patterns and relationships among variables. Simply replacing missing values with zero could introduce significant distortions in the data and yield misleading insights. Thus, a better strategy is required for handling missing data. Our problem statement includes, finding a decent way to replace the `NA` values in the date, model the complete data in ARIMA process accurately, change the data into univariate time series, extract information about the blasting time and forecast for the presence of different gasses in environment in future and find the approximate non parametric probability distribution (and curve) for all the components of data.We employ various graph plotting techniques to analyse the dataset

This project contributes to the field of environmental data analysis by demonstrating effective techniques for handling missing values in multivariate time-series datasets and infer various kind of informenvironmental data analysis by demonstrating effective techniques for handling missing values in multivariate time-series datasets and inferring variousation from it. The insights gained from this study can enhance our understanding of air quality dynamics and support informed decision-making in environmental monitoring and management initiatives.

Table of contents

1. Information about Dataset & its's Modification:

The air pollution dataset that we will use is obtained from the `Singrauli Coalfield Pollution Control Board for Coal India` (Singrauli Coalfield). The pollution is monitored during open-pit blasting. There are 13 columns overall in the air pollution data collection of pollutants available at intervals of 15 minutes. I have modified the column names for convenience. A small tail portion of the dataset is shown below.

Index	Start	End	PM10..µg.m3.	NO..µg.m3.	NO2..µg.m3.	NOX..ppb.	CO..mg.m3.	SO2..µg.m3.	NH3..µg.m3.	Ozone..µg.m3.	Benzene..µg.m3.	PM2.5..µg.m3.
8616	01-05-2023 17:45	01-05-2023 18:00	73.00	7.60	93.60	56.00	0.79	6.40	8.40	18.30	0.10	13.00
8617	01-05-2023 18:00	01-05-2023 18:15	73.00	8.20	95.50	57.50	0.88	6.00	7.90	18.20	0.10	13.00
8618	01-05-2023 18:15	01-05-2023 18:30	73.00	9.60	95.50	58.60	0.87	6.80	7.90	22.90	0.10	13.00
8619	01-05-2023 18:30	01-05-2023 18:45	73.00	10.50	95.30	59.20	0.79	9.80	8.20	25.50	0.10	13.00
8620	01-05-2023 18:45	01-05-2023 19:00	73.00	10.50	95.80	59.50	0.72	23.00	8.70	27.90	0.10	15.00
8621	01-05-2023 19:00	01-05-2023 19:15	73.00	10.80	97.60	60.70	0.68	41.40	8.50	19.00	0.10	15.00
8622	01-05-2023 19:15	01-05-2023 19:30	73.00	11.30	97.20	60.90	0.71	22.20	9.00	23.90	0.10	15.00
8623	01-05-2023 19:30	01-05-2023 19:45	73.00	11.80	98.40	61.90	0.68	12.50	8.70	29.70	0.10	15.00
8624	01-05-2023 19:45	01-05-2023 20:00	51.00	11.60	97.90	61.50	0.65	7.10	8.60	31.80	0.10	6.00
8625	01-05-2023 20:00	01-05-2023 20:15	51.00	11.50	97.20	61.00	0.65	6.90	8.50	31.10	0.10	6.00
8626	01-05-2023 20:15	01-05-2023 20:30	51.00	11.50	96.10	60.50	0.57	6.40	9.70	36.90	0.10	6.00
8627	01-05-2023 20:30	01-05-2023 20:45	51.00	12.20	95.90	60.90	0.58	6.10	10.80	33.60	0.10	6.00
8628	01-05-2023 20:45	01-05-2023 21:00	32.00	11.90	97.20	61.40	0.61	7.60	10.90	34.80	0.10	8.00
8629	01-05-2023 21:00	01-05-2023 21:15	32.00	12.20	98.30	62.20	0.60	8.70	10.50	36.40	0.10	8.00
8630	01-05-2023 21:15	01-05-2023 21:30	32.00	12.00	98.00	61.90	0.60	9.20	10.70	30.50	0.10	8.00
8631	01-05-2023 21:30	01-05-2023 21:45	32.00	11.50	98.10	61.60	0.61	11.00	10.50	29.70	0.10	8.00
8632	01-05-2023 21:45	01-05-2023 22:00	28.00	12.60	97.90	62.30	0.68	11.60	10.40	26.70	0.10	7.00
8633	01-05-2023 22:00	01-05-2023 22:15	28.00	13.00	97.90	62.60	0.69	10.80	10.70	22.50	0.10	7.00
8634	01-05-2023 22:15	01-05-2023 22:30	28.00	14.50	98.50	64.20	0.72	9.40	10.90	17.20	0.10	7.00
8635	01-05-2023 22:30	01-05-2023 22:45	28.00	17.10	99.00	66.60	0.72	9.50	10.90	17.20	0.10	7.00
8636	01-05-2023 22:45	01-05-2023 23:00	19.00	17.90	100.00	67.80	0.63	10.00	10.70	26.10	0.10	11.00
8637	01-05-2023 23:00	01-05-2023 23:15	19.00	17.90	100.00	67.70	0.57	10.00	10.40	30.90	0.10	11.00
8638	01-05-2023 23:15	01-05-2023 23:30	19.00	19.60	100.20	69.20	0.58	9.90	10.50	29.60	0.10	11.00
8639	01-05-2023 23:30	01-05-2023 23:45	19.00	20.80	100.20	70.20	0.58	9.50	10.80	30.00	0.10	11.00
8640	01-05-2023 23:45	02-05-2023 00:00	32.00	21.80	98.80	70.30	*NA*	*NA*	11.00	33.50	0.10	6.00
8641	Min		12.00	0.10	0.20	4.20	0.10	0.10	4.60	0.10	0.10	3.00
8642	Max		847.00	157.50	106.90	165.20	4.00	645.60	62.40	123.80	0.60	474.00
8643	Avg.		181.41	14.65	55.76	42.67	1.41	34.23	13.24	35.63	0.18	75.69

Figure 1: Primary Dataset

The dataset has 8643 rows and 13 variables, of which the last three rows didn't have any time stamp. They are just the first-order statistics, last-order statistics and mean of respective column variables, So, We remove them from the working dataset. We can extract that statistical information whenever we want. The column's description is following:

- Column 1 Indicates the serial no of the data set, and I named it `Index`.
- Column 2 and 3 Indicates the date and time from and to for 15 minutes of interval. I named them `Start` & `End,` respectively.
- Column 4 Indicates the PM10 pollutant of the data in µg/m3.
- Column 5 Indicates NO pollutant of the data in µg/m3.
- Column 6 Indicates the NO2 pollutant of the data in µg/m3.

- Column 7 Indicates the NOX of the data in ppb.
- Column 8 Indicates the CO pollutant of the data in mg/m3.
- Column 9 Indicates the SO2 pollutant of the data in µg/m3.
- Column 10 Indicates the NH3 pollutant of the data in µg/m3.
- Column 11 Indicates the Ozone pollutant of the data in µg/m3.
- Column 12 Indicates the Benzene pollutant of the data in µg/m3.
- Column 13 Indicates the PM2.5 pollutant of the data in µg/m3.

Our analysis will be `Per_Column` because in the given multivariate time series data, time varies as the row varies.

2. Handling missing values & Time-Series Analysis:

Handling missing values in time series analysis requires careful consideration to ensure accurate and meaningful results. We will performthree3 methods for handling `NA` values in our data set.

2.1. Replacing `NA` values with `0`

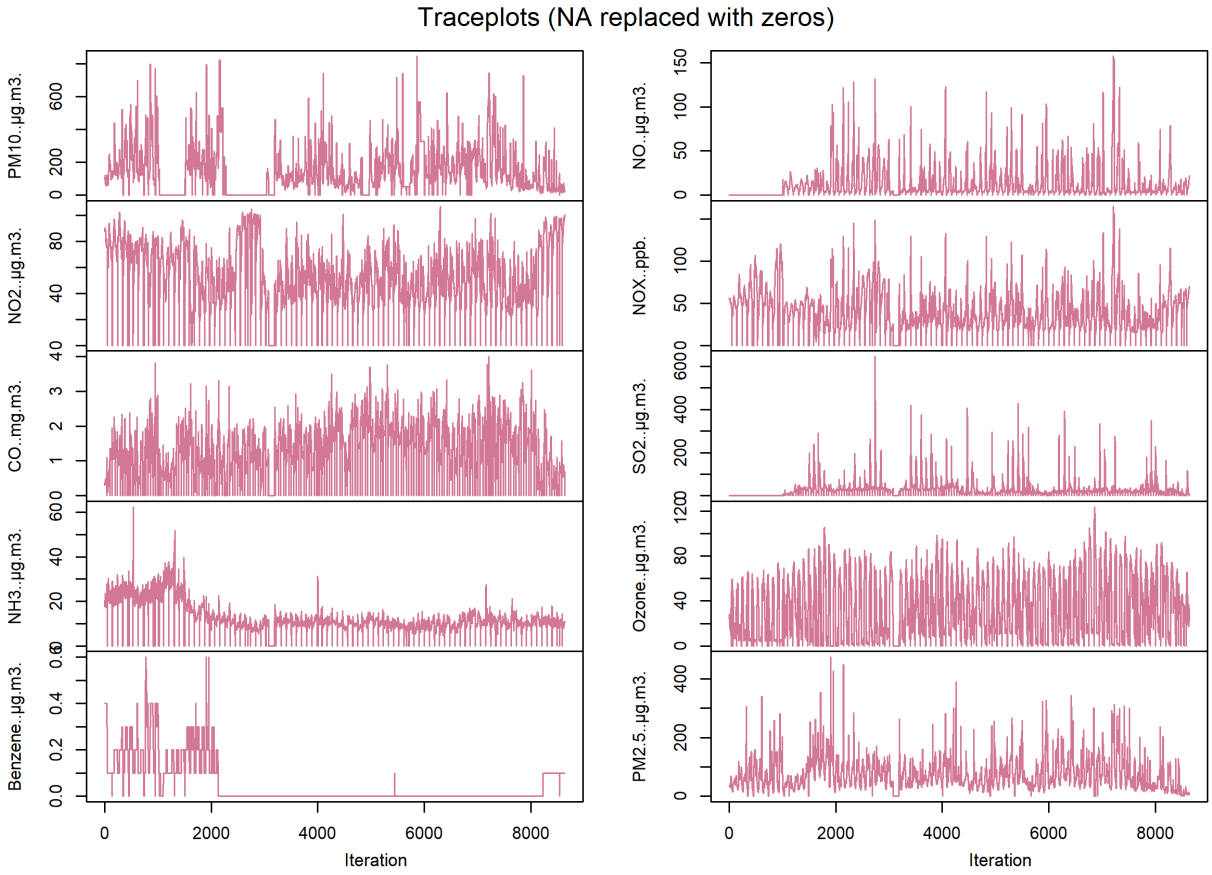
Replacing missing data with 0 is generally not recommended in most scenarios because it can introduce bias and distort the analysis. However, there may be some specific situations where replacing missing data with 0 is appropriate. Here are a few cases where it might be suitable to consider replacing missing values with 0:

Contextual Appropriateness: Consider the variable's nature and the data's context. If the lost data represents an absence or zero value, such as counts or binary indicators, replacing missing values with 0 may be reasonable.

Analysis Requirements: Isomeme instances of the analytical method or model used may require complete data or a fixed-length vector as input.

Missing Completely at Random (MCAR): If the missingness is determined by factors completely unrelated to the data itself, replacing missing values with 0 may be acceptable.

We have missing values represented as `NA`. We will now analyse the traceplots after setting all `NA` values to 0. As We can see below the trace plots of all the variables, We can interpret that replacing `NA` values with 0, causes the bad quality trace plots of some of the variables like PM10, NO2, NH3, Benzene, NOX etc.



We will try to find another arguments proving that replacing `NA` with 0 is not a good option. I have fitted ARIMA models for each variable on the dataset we replaced `na` with `ZEROS`. Here is the table of results we obtained:

	Variable	Modeled Process	AIC	BIC
1	PM10..µg.m3.	ARIMA 4 1 5	91261.0370450524	91331.677466196
2	NO..µg.m3.	ARIMA 3 1 1	46539.690144184	46575.0103547558
3	NO2..µg.m3.	ARIMA 3 1 0	62194.8371381814	62223.0933066388
4	NOX..ppb.	ARIMA 5 1 2	57178.7074816494	57235.2198185643
5	CO..mg.m3.	ARIMA 1 1 2	6733.59602400766	6761.8521924651
6	SO2..µg.m3.	ARIMA 2 1 3	76009.0991760431	76051.4834287293
7	NH3..µg.m3.	ARIMA 5 1 2	40588.2159282712	40644.7282651861
8	Ozone..µg.m3.	ARIMA 1 1 3	62333.7278000932	62369.048010665
9	Benzene..µg.m3.	ARIMA 0 1 2	-48129.0811819596	-48107.8890556165
10	PM2.5..µg.m3.	ARIMA 0 1 0	76984.3894439261	76991.4534860404

Figure 2: Models obtained by replacing NA with zeros

Replacing missing values with zeros may not be appropriate in all cases(in this case also). It depends on the nature of the data and the analysis we want to perform. Alternatively, We can choose to interpolate or impute missing values based on the context.

2.2. Interpolation

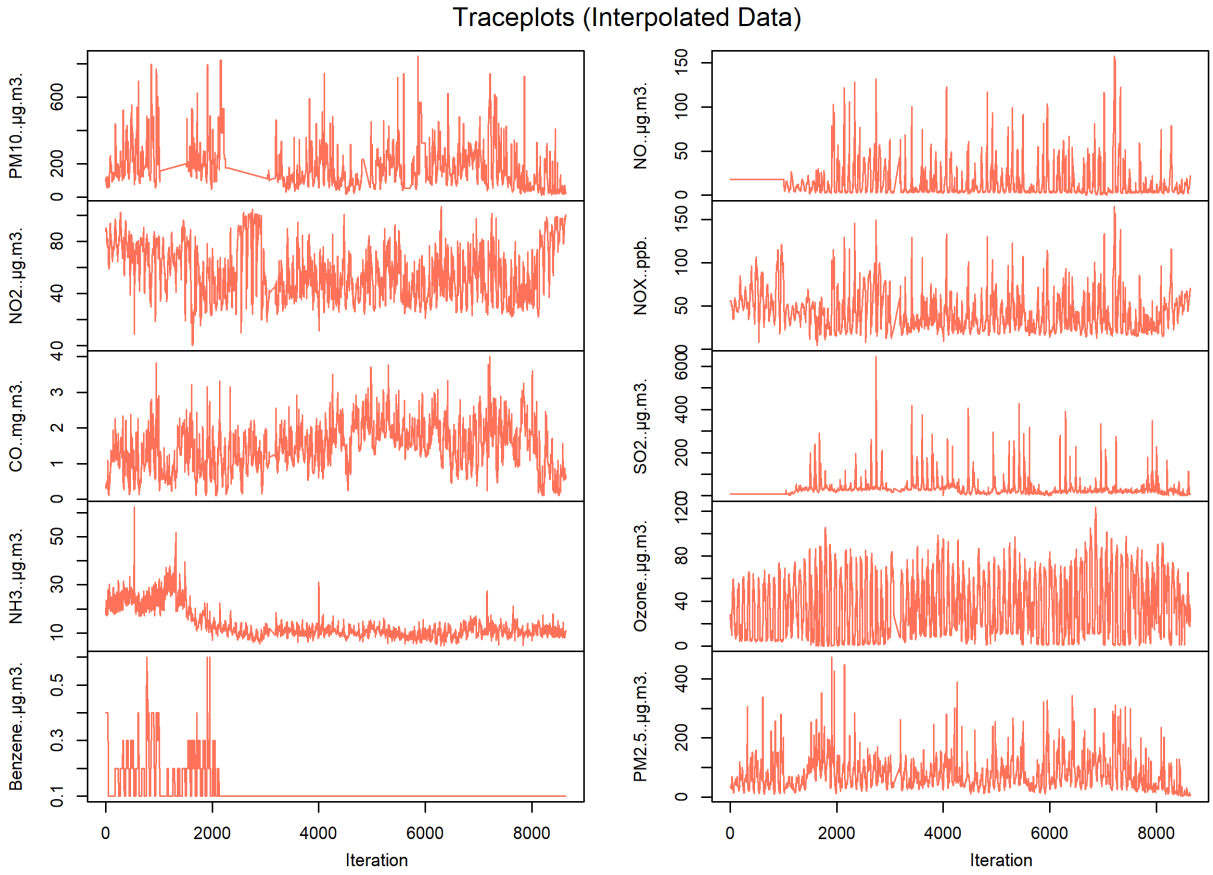
Interpolation is commonly used for finding missing values in a dataset, mainly when the missing values occur in a sequence or when the values are expected to exhibit a smooth pattern.Traditional interpolation techniques may be computationally intensive and not scalable for big data scenarios. Therefore, specialised aproaches are used to handle interpolation in big data. Here are some reasons why interpolation is utilised in such scenarios:

- Preserving Temporal Relationships
- Maintaining Data Integrity
- Minimizing Data Loss
- Improving the Accuracy of Calculations

We are going to use the `imputeTS` R package to perform interpolation. It is a popular choice for time series interpolation due to several reasons. The imputeTS package in R is a valuable tool for time series interpolation due to its diverse range of interpolation methods, handling of various missing data patterns, ease of implementation, customizability and integration with other packages. We can do interpolation by-Row and by-Column.

```
# Interpolate NA values using Linear interpolation
linear_interpolation <- na_interpolation(data, option = "linear", method = "row")
linear_interpolation <- na_interpolation(data, option = "linear", method = "column")
# Interpolate NA values using spline interpolation(can be done in both way Row and column)
spline_interpolation <- na_interpolation(data, method = "spline")

# Interpolate NA values using Stineman interpolation
stinmn_interpolation <- na_interpolation(data, method = "stine")
```

However, for our dataset, all the 3 interpolation(& both methods) techniques producing same replacements for `NA` values. So, we can take any of them, and compute the ARIMA models, of this interpolated data. Trace plots of Interpolated Data is as following.



I have fitted ARIMA models for each variable on the interpolated dataset. Here is the table of results we obtained:

	Variable	Modeled Process	AIC	BIC
1	PM10..µg.m3.	ARIMA 0 1 0	88589.4537063242	88596.5177484385
2	NO..µg.m3.	ARIMA 5 1 0	42315.6318398526	42358.0160925388
3	NO2..µg.m3.	ARIMA 4 1 2	43484.7723071912	43534.2206019917
4	NOX..ppb.	ARIMA 4 1 3	42265.3039687639	42321.8163056788
5	CO..mg.m3.	ARIMA 2 1 1	-6494.13580933355	-6465.87964087612
6	SO2..µg.m3.	ARIMA 2 1 4	71553.7195503275	71603.1678451281
7	NH3..µg.m3.	ARIMA 1 1 3	24965.0182964415	25000.3385070133
8	Ozone..µg.m3.	ARIMA 3 1 0	53315.335880956	53343.5920494134
9	Benzene..µg.m3.	ARIMA 1 1 1	-49996.3382435507	-49975.1461172076
10	PM2.5..µg.m3.	ARIMA 0 1 0	76562.9075546838	76569.9715967981

Figure 3: Models Obtained by Interpolating Data

By comparing these plots and statistical measures with the previous ones, we can say that interpolation gives a better data replacement. For Some of the columns, interpolation gives approximately the same results as the first method gives.

2.3. Multiple Imputation:

Multiple imputation methods generate plausible imputed data sets, incorporating uncertainty around the missing values. Multivariate imputation approaches, such as Multiple Imputation by Chained Equations (MICE), consider the dependencies between variables. Each variable is imputed conditional on the other variables, capturing their relationships. The `mice` package in R provides multiple imputation methods for handling missing values in a data set. The package uses the Multiple Imputation by Chained Equations (MICE) approach, which involves creating multiple imputed data sets and filling in the missing values based on observed values and the relationships between variables. We will now fill `NA` values by imputation technique and compare the quality of predicted values with previous techniques.

Imputation techniques make use of descriptive statistics and statistical inference concepts.

Descriptive statistics involve summarizing and describing the characteristics of a dataset, such as measures of central tendency (mean, median, mode) and measures of dispersion (variance, standard deviation). Imputation techniques often utilize descriptive statistics to estimate missing values based on the observed data. For example, one common imputation method is mean imputation, where the missing values are replaced with the mean of the available data. This approach uses the descriptive statistic of the mean to fill in the missing values.

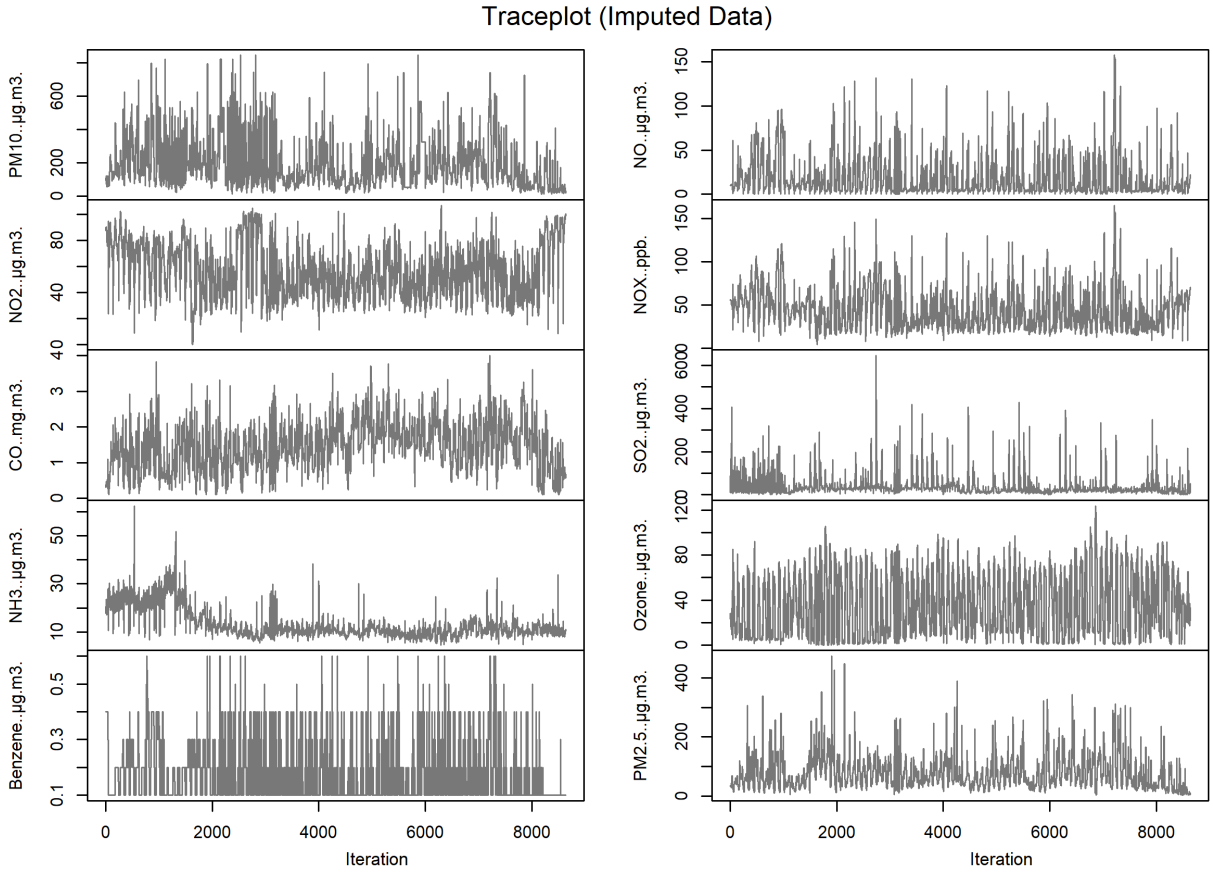
Statistical inference involves making conclusions or predictions about a population based on a sample of data. Imputation techniques employ statistical inference to estimate missing values by using information from the observed data. The idea is to make educated guesses about the missing values based on patterns and relationships observed in the available data.

This `mice` package in R, for instance, utilizes multiple imputation techniques based on a statistical model. It creates multiple imputed datasets by drawing plausible values from the conditional distributions of the missing values given the observed data. This process incorporates statistical inference to estimate the missing values while accounting for the uncertainty associated with them.

```
imp_model <- mice(data[,c(-1,-2,-3)], method = "pmm")
data_imputed <- complete(imp_model)
```

In summary, imputation techniques rely on descriptive statistics to summarize the available data and statistical inference to estimate missing values based on patterns and relationships in the observed data.

After comparing the trace plots of `Imputed Data` with the previous two trace plots, we can say that this is the best method to fill in missing values. The Traceplot of completed data with this method is as follows:



Here is the table of results we obtained from fitted ARIMA models:

	Variable	Modeled Process	AIC	BIC
1	PM10..µg.m3.	ARIMA 0 1 5	99392.5492145974	99434.9334672835
2	NO..µg.m3.	ARIMA 0 1 5	61603.0270942576	61645.4113469437
3	NO2..µg.m3.	ARIMA 3 1 3	57980.8316822288	58030.2799770293
4	NOX..ppb.	ARIMA 1 1 2	60606.5388928932	60634.7950613506
5	CO..mg.m3.	ARIMA 3 1 1	1516.88848316292	1552.20869373472
6	SO2..µg.m3.	ARIMA 1 1 3	80618.3118522544	80653.6320628262
7	NH3..µg.m3.	ARIMA 2 1 3	37054.614380015	37096.9986327012
8	Ozone..µg.m3.	ARIMA 0 1 2	63300.4523834378	63321.6445097809
9	Benzene..µg.m3.	ARIMA 0 1 2	-24133.3426015642	-24112.1504752211
10	PM2.5..µg.m3.	ARIMA 0 1 1	78732.1151506153	78746.243234844

Figure 4: Models obtained by Imputing Data

Now, based on the comparison of AIC and BIC values obtained from the ARIMA models using three different missing data handling techniques (interpolated data, zero replacement with missing values, and imputed data), We have observed the following order:

1. Interpolated Data: Smallest AIC and BIC values.
2. Zero Replacement with Missing Values: Intermediate AIC and BIC values.
3. Imputed Data: Largest AIC and BIC values.

From this analysis, We can conclude that the ARIMA models fitted with the interpolated data perform the best in terms of model fit and complexity. These models have the smallest AIC and BIC values, indicating a good balance between model fit and complexity. On the other hand, the ARIMA models fitted with zero replacement and imputed data show larger AIC and BIC values. This suggests that these models may have poorer fit or higher complexity compared to the models using interpolated data.

However, it's important to consider the context of our analysis and the specific goals of study. The choice of missing data handling technique should align with the assumptions and requirements of your analysis. While smaller AIC and BIC values generally indicate better model fit and parsimony, it's crucial to carefully evaluate the overall impact of missing data handling techniques on your specific analysis and consider potential implications beyond the AIC and BIC values alone. If we see trace plots, The imputed Data shows best results and best replacement of missing values.

3. Classification:

To determine whether the data is stock time series or flow time series data by examining trace plots and ACF (Autocorrelation Function) plots, We need to understand the characteristics of each type of data and analyze the patterns in the plots.

3.1. Stock Time Series Data:

Stock time series data represents the cumulative value of a variable over time. It typically consists of values that increase or decrease incrementally, reflecting the accumulation or depletion of a quantity. Examples include the total number of shares of a company's stock or the total assets of a company.

Characteristics of Stock Time Series Data:

1. Monotonicity: Stock data shows a monotonic pattern, with values steadily increasing or decreasing over time.

2. Limited variability: The values in stock data usually have a limited range, as they represent a cumulative quantity that cannot be negative.

3. Persistence: There is a high degree of persistence in stock data, meaning that the current value is highly correlated with past values.

4. Seasonality: Stock data may exhibit seasonality, where values follow regular patterns or cycles.

Analysis of Trace Plots for Stock Time Series Data	**Analysis of ACF Plots for Stock Time Series Data**
An increasing or decreasing trend: The trace plot should show an apparent upward or downward movement.	**Strong autocorrelation at lag:** Stock data typically exhibit a strong positive autocorrelation at the first lag since the current value is closely related to the previous value
Limited fluctuations: The data points should not deviate significantly from the overall trend, as stock data generally do not exhibit large fluctuations.	**Slow decay in autocorrelation:** The autocorrelation should decay slowly, indicating persistence in the data.

3.2. Flow Time Series Data:

Flow time series data represents a variable's change rate over time. It measures the inflow or outflow of a quantity during specific time intervals. Examples include the daily sales of a product or the hourly website traffic.

Characteristics of Flow Time Series Data:

1. Variability: Flow data exhibit higher variability than stock data since it represents the rate of change.

2. Randomness: Flow data often show random fluctuations and do not necessarily follow a specific trend.

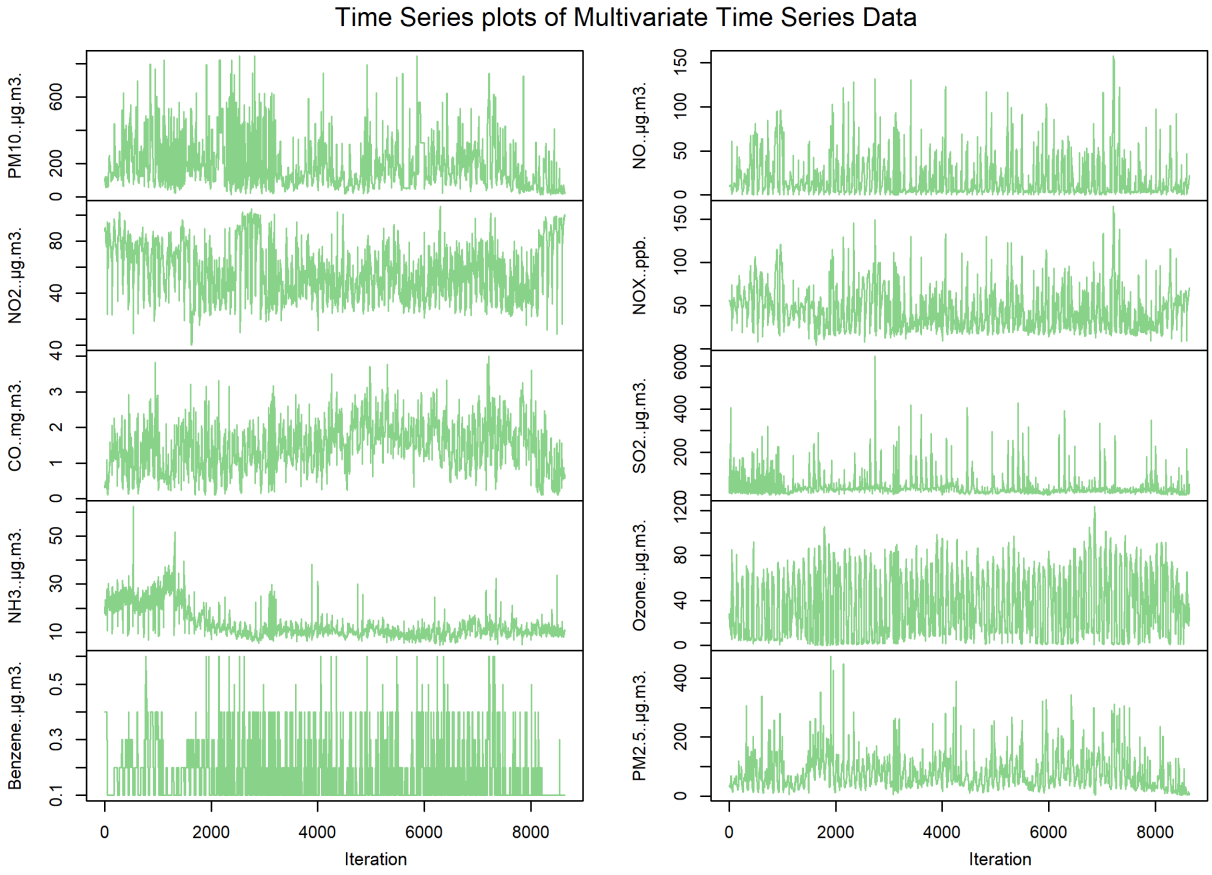
3. Lack of persistence: The autocorrelation between consecutive values in flow data is typically low.

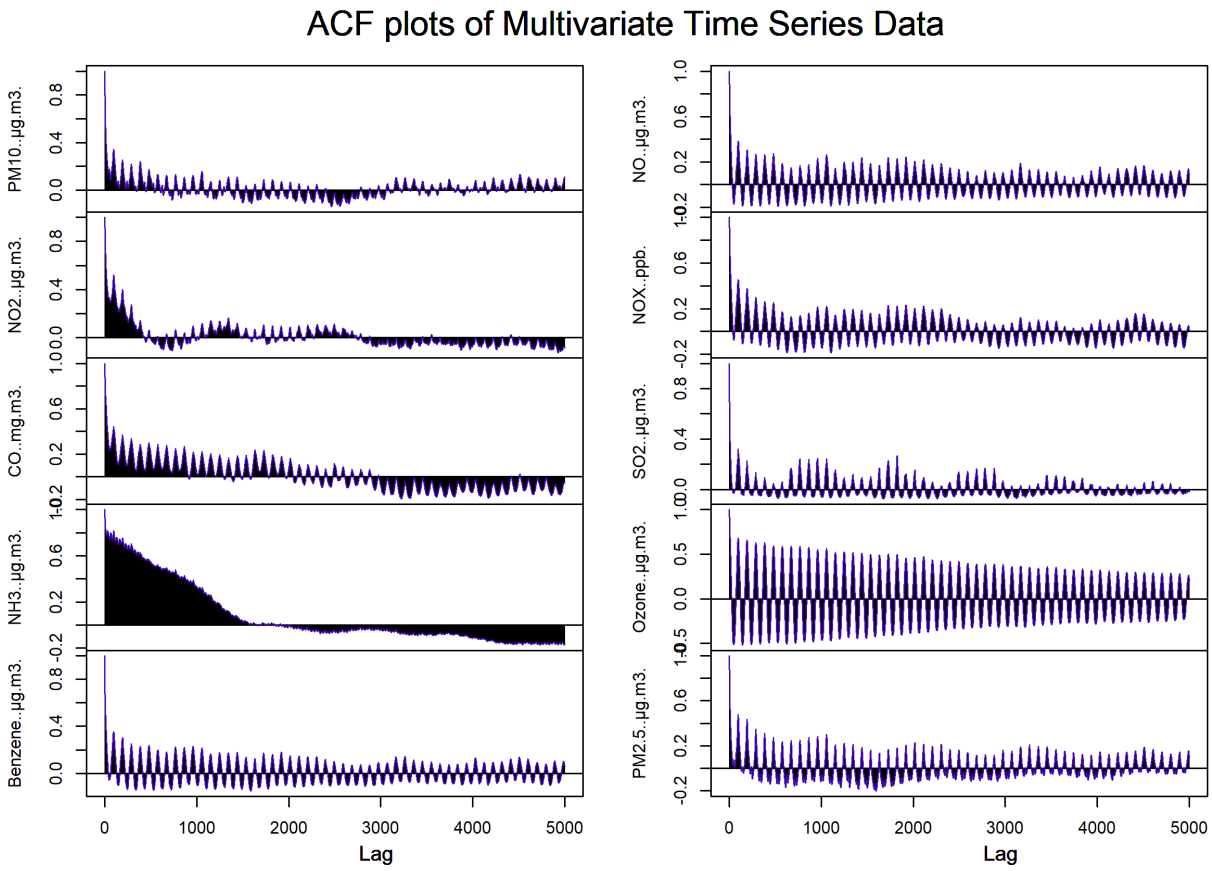
4. Absence of seasonality: Flow data may not exhibit clear seasonal patterns.

Analysis of Trace Plots for Flow Time Series Data	Analysis of ACF Plots for Flow Time Series Data
Fluctuating values: The trace plot should display irregular fluctuations without a clear upward or downward trend.	**Low autocorrelation:** The autocorrelation values should be relatively low and decline rapidly after the first few lags, indicating a lack of persistence.
Variable range: The data points can vary significantly, reflecting the varying rates of change.	**Randomness in autocorrelation:** The autocorrelation values should not show any discernible pattern or significant correlations at specific lags.

3.3. Analysis:

In conclusion, by carefully analysing the trace plots and ACF plots of time series data, we can determine whether it represents stock or flow data based on the aforementioned characteristics and patterns. Let us do some visualization to analyize what we discussed above.





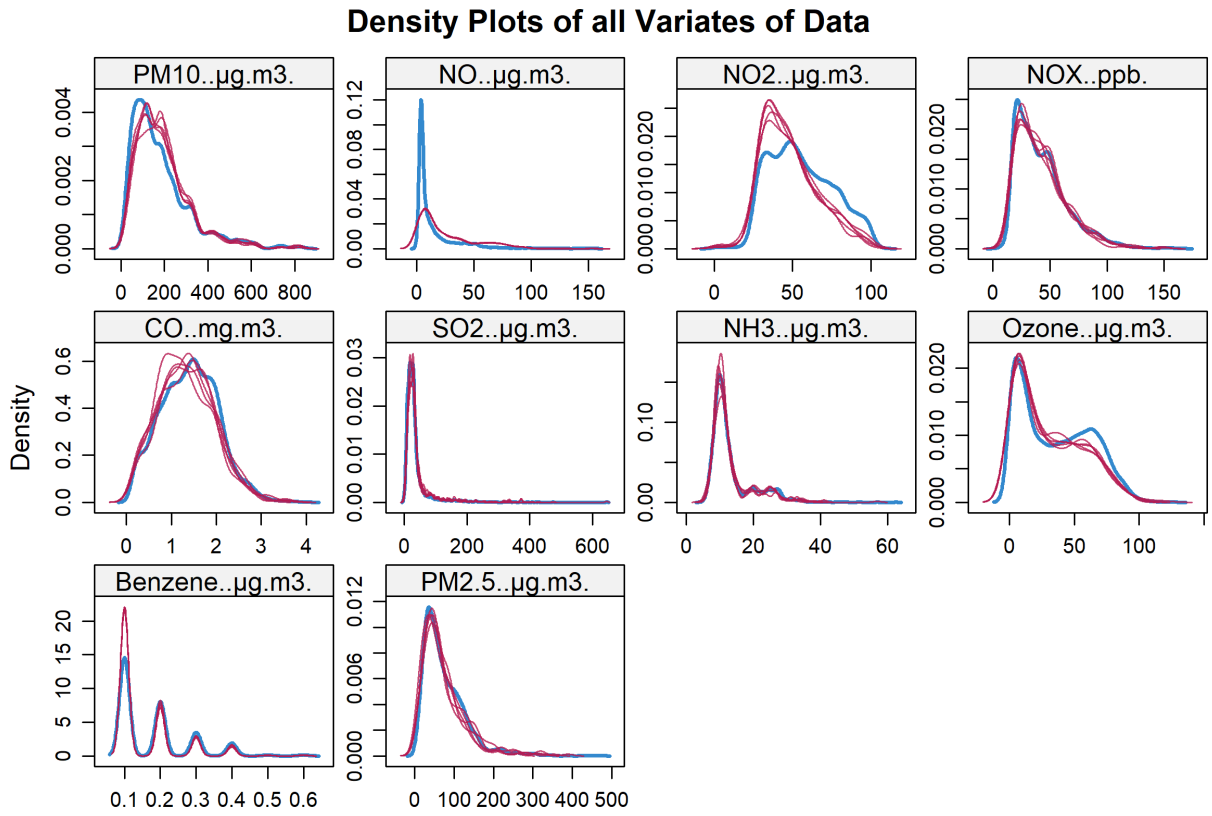
We can see that the trace plots did not show monotonic behavior, and having fluctuating value. Whereas, ACF plots shows the significant decrease in autocorrelation coefficient with increasing the lag, with a significant random patterns in some components. Both the plots indicating that our data is `Flow Time Series Data`. Hence, we can conclude that our data is `Flow Time Series Data`.

4. Curve Fitting:

Curve fitting in multivariate time series analysis is important for several reasons and offers various uses. Some of the key applications and benefits include:

1. Pattern Identification
2. Forecasting
3. Interpolation
4. Anomaly Detection
5. Feature Extraction
6. Visualization
7. Model Selection

Fitted curves for the given data is as following:



Plotting all the densities in a single plot becomes very unusual, because we can not observe such a clumsy plot conveniently. Here, we can assume some approximate distributions for different columns. Let us check the covariance matrix for our dataset:

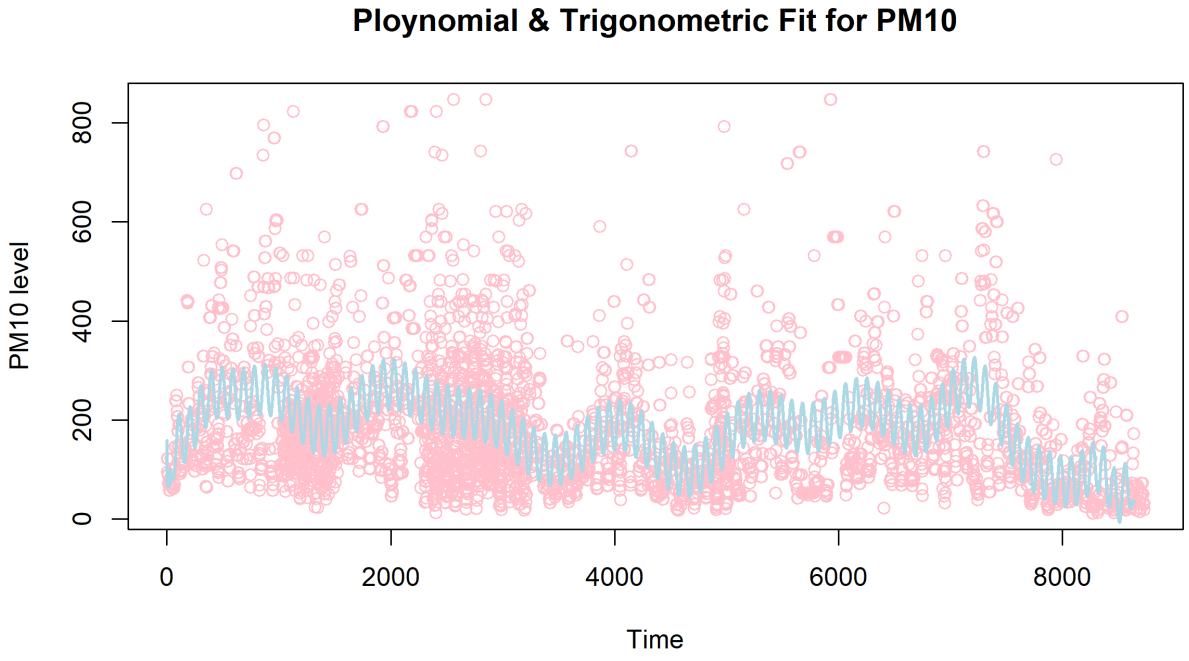
4.1. Covariance Matrix

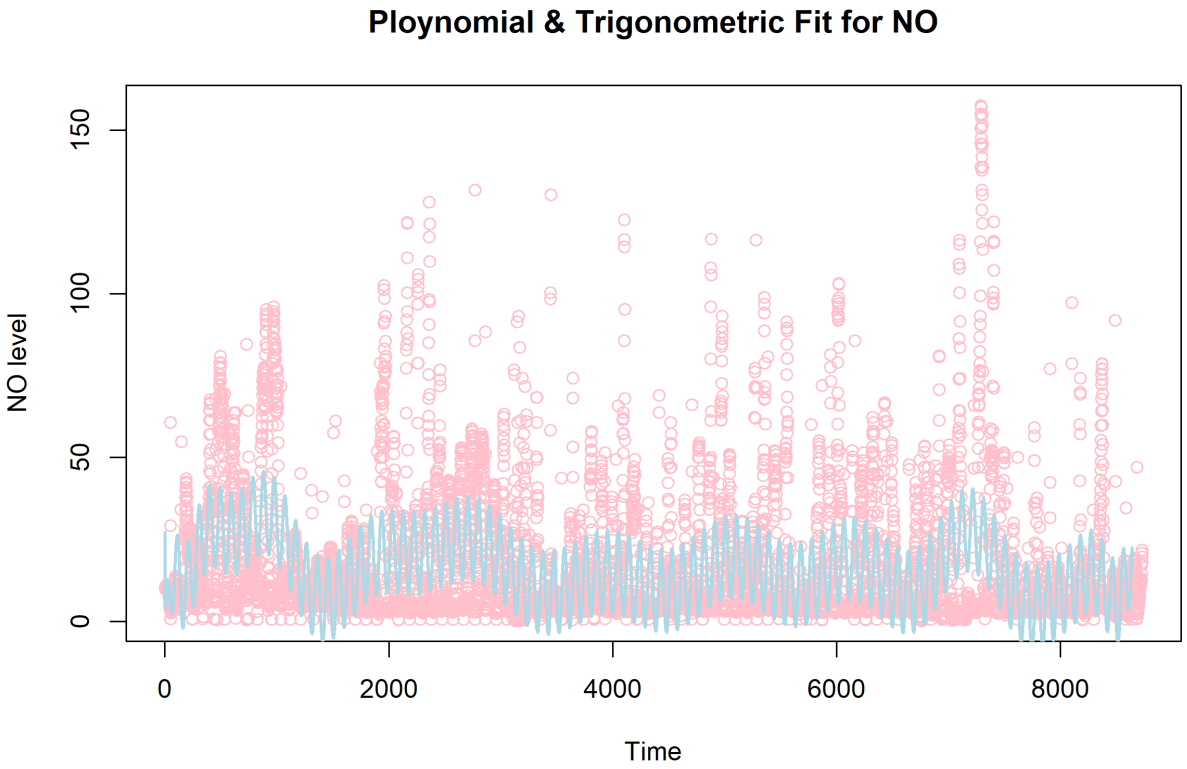
	PM10	NO	NO2	NOX	CO	SO2	NH3	Ozone	Benzene	PM2.5
PM10	1.000	0.473	0.194	0.443	0.203	-0.034	0.174	-0.314	0.508	0.530
NO	0.473	1.000	0.316	0.890	0.327	0.033	0.140	-0.523	0.636	0.576
NO2	0.194	0.316	1.000	0.713	-0.069	0.183	0.259	-0.429	0.276	0.200
NOX	0.443	0.890	0.713	1.000	0.209	0.112	0.227	-0.593	0.602	0.522
CO	0.203	0.327	-0.069	0.209	1.000	-0.056	-0.234	-0.393	0.232	0.419
SO2	-0.034	0.033	0.183	0.112	-0.056	1.000	-0.011	0.017	0.023	-0.023
NH3	0.174	0.140	0.259	0.227	-0.234	-0.011	1.000	-0.078	0.197	0.045
Ozone	-0.314	-0.523	-0.429	-0.593	-0.393	0.017	-0.078	1.000	-0.478	-0.509
Benzene	0.508	0.636	0.276	0.602	0.232	0.023	0.197	-0.478	1.000	0.661
PM2.5	0.530	0.576	0.200	0.522	0.419	-0.023	0.045	-0.509	0.661	1.000

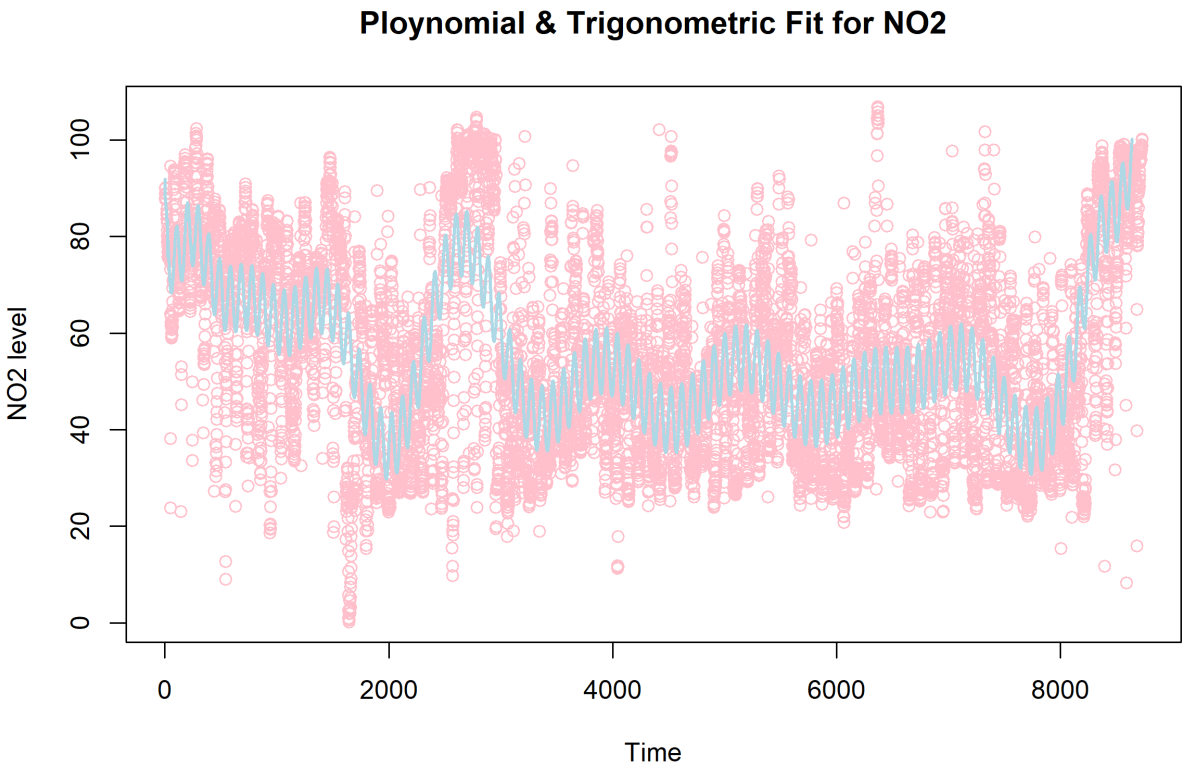
We can not consider the multivariate normal distribution here, because state space of this multivariate tise series data is 0 to $infinity$. Still, we can consider each component as mixture of any particular distribution, with different parameters. For example we can modeled each component as mixture of `location Gamma distributuin`, with different scale and shape parameter ($alpha$,$beta$) and diffrerent location Parameter.

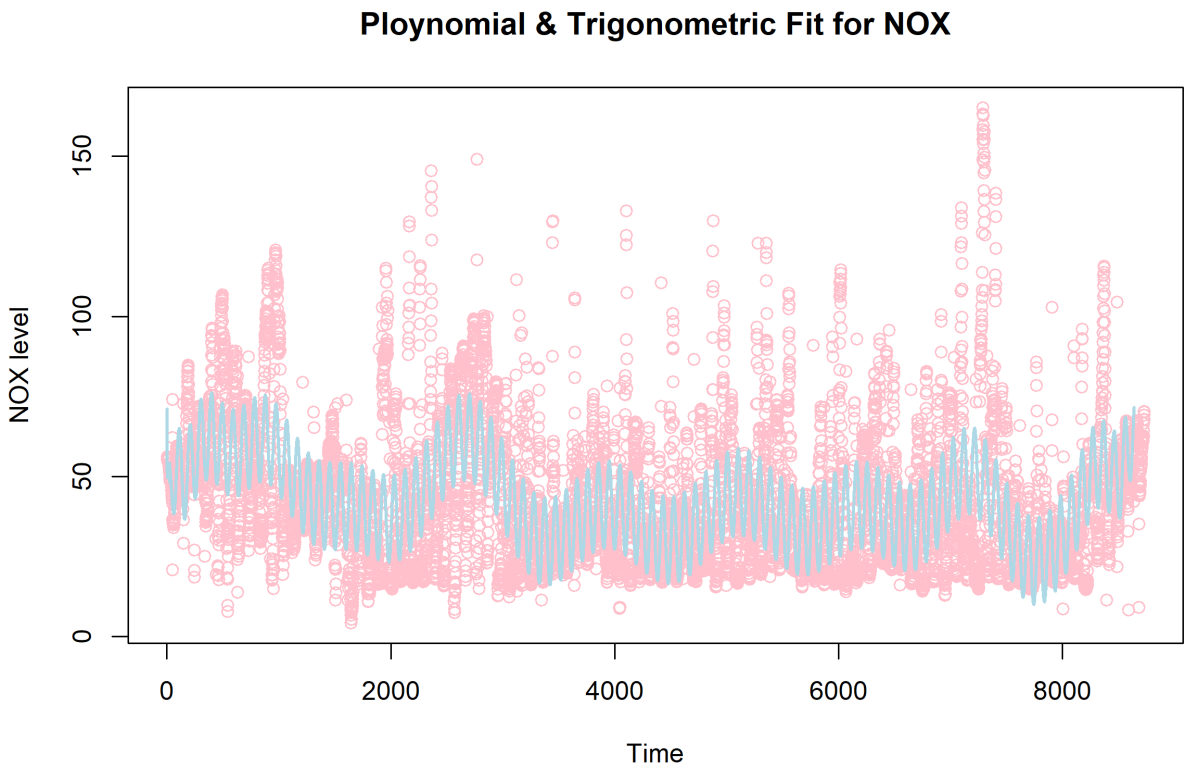
4.2. Polynomial Curve fitting:

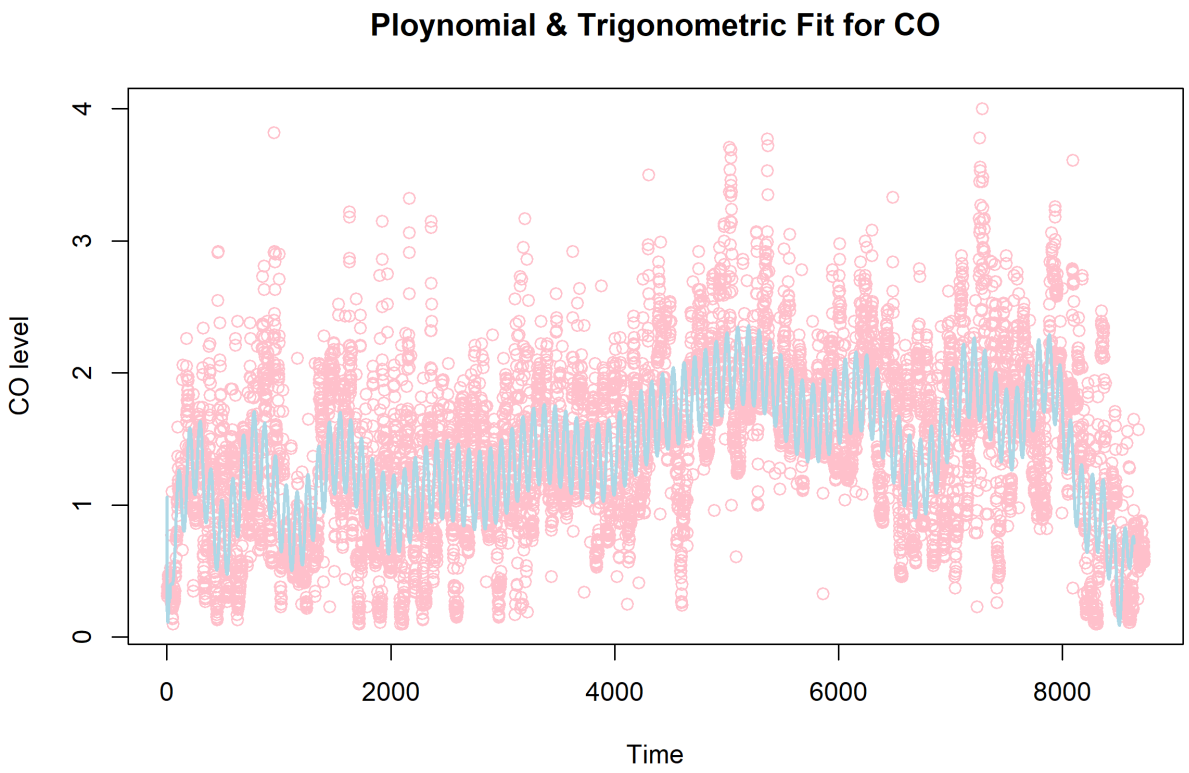
Polynomial fitting can be applied to various types of data, including time series data. Time series data often exhibit nonlinear patterns and trends that cannot be effectively captured by linear models. By using polynomial fitting, we can capture these nonlinearities and improve the accuracy of our models for time series analysis. In time series analysis, polynomial fitting can be useful for tasks such as trend estimation, forecasting, and anomaly detection. By fitting a polynomial to the time series data, we can estimate the underlying trend and make predictions about future values. Additionally, polynomial fitting can be combined with other techniques, such as autoregressive integrated moving average (ARIMA) models, to capture both the linear and nonlinear components of a time series. It's important to note that when fitting polynomials to time series data, we need to consider the potential pitfalls of overfitting. Using high-degree polynomials can lead to overly complex models that fit the noise in the data rather than the underlying patterns. Therefore, it's essential to balance model complexity with the available data and the principle of parsimony to avoid overfitting and ensure reliable results. Approximate Polynomial and Trigonometric fit for each component is shown below in form of plot.

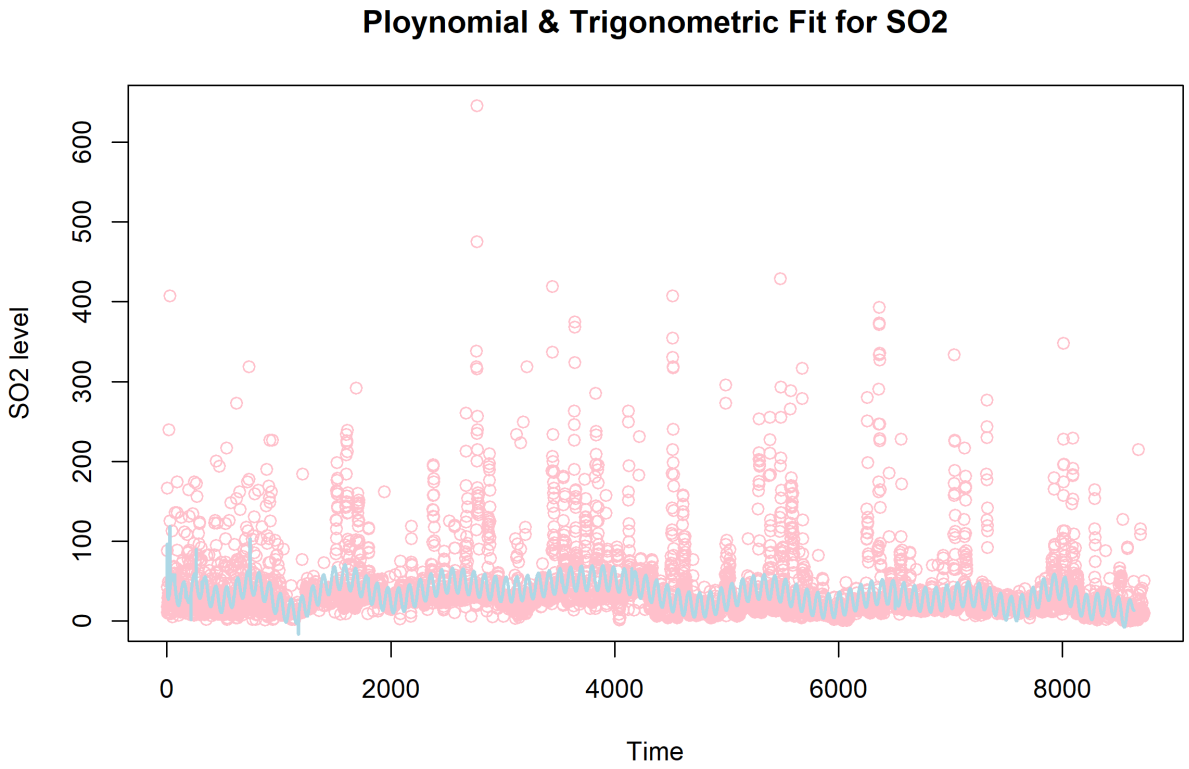


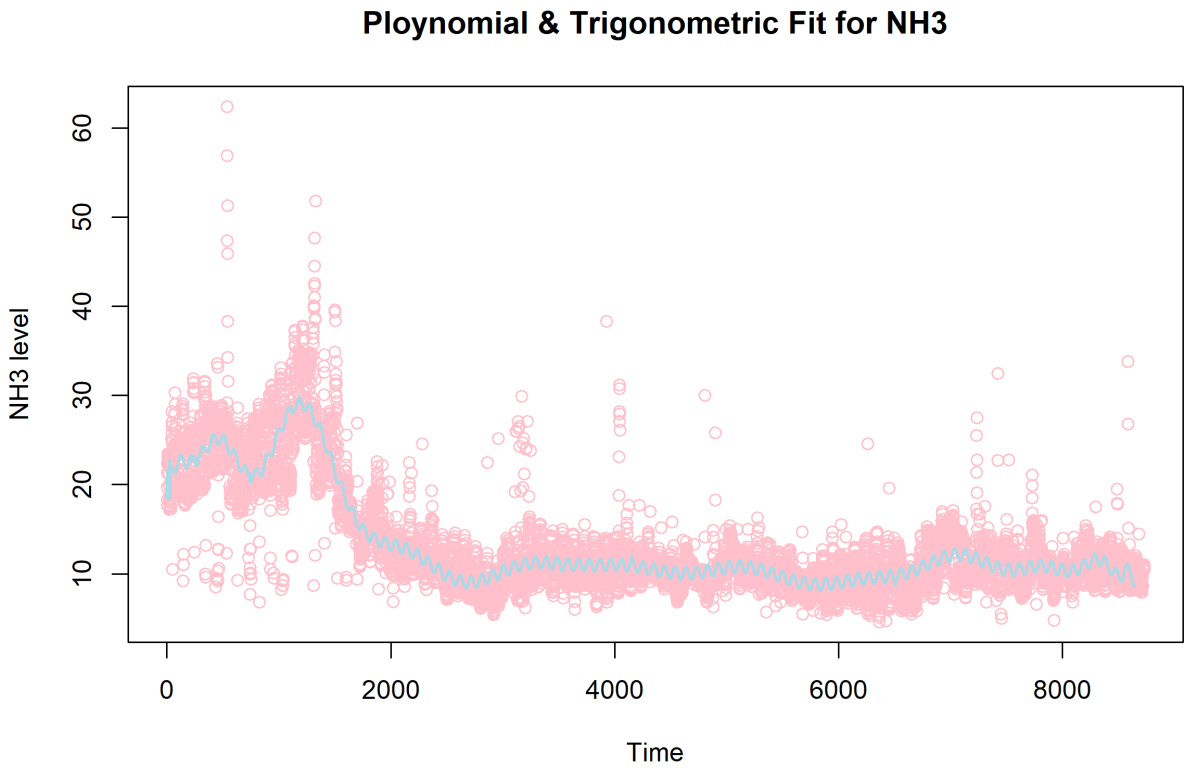

Ploynomial & Trigonometric Fit for NO
NO level
0
50
100
150
0
2000
4000
6000
8000
Time

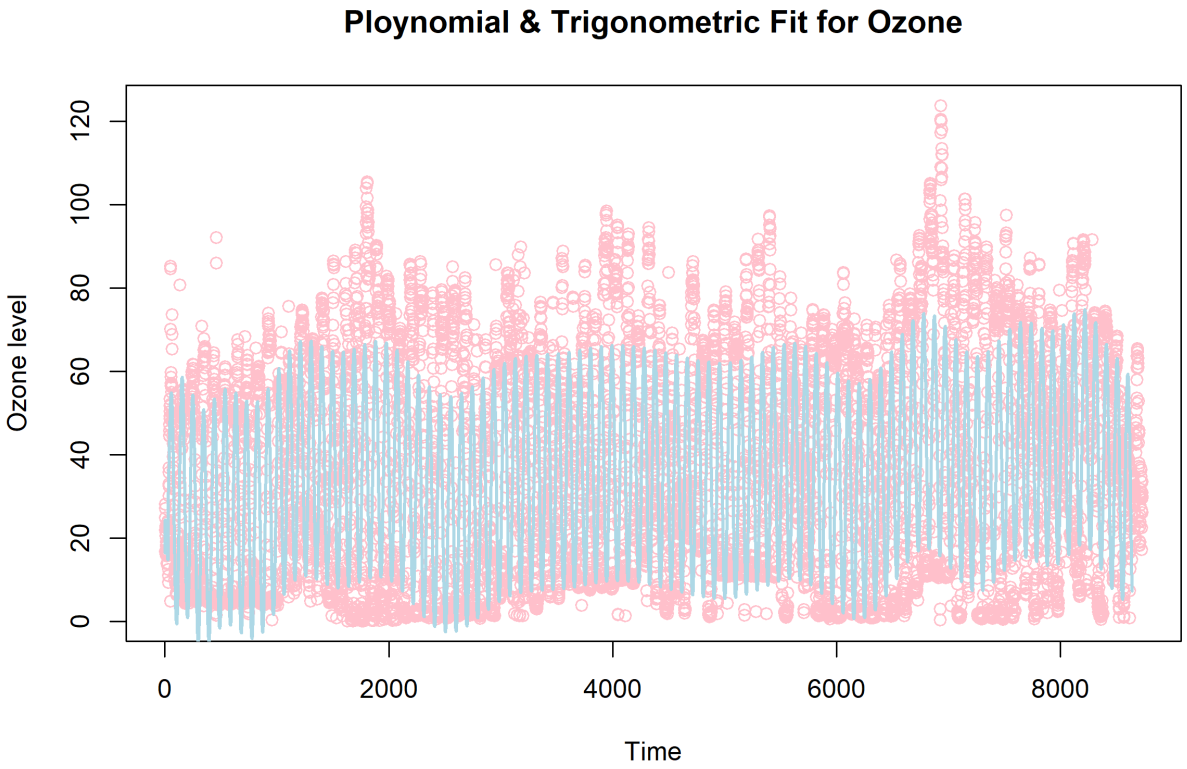

Ploynomial & Trigonometric Fit for NO2
NO2 level
0
20
40
60
80
100
0
2000
4000
6000
8000
Time

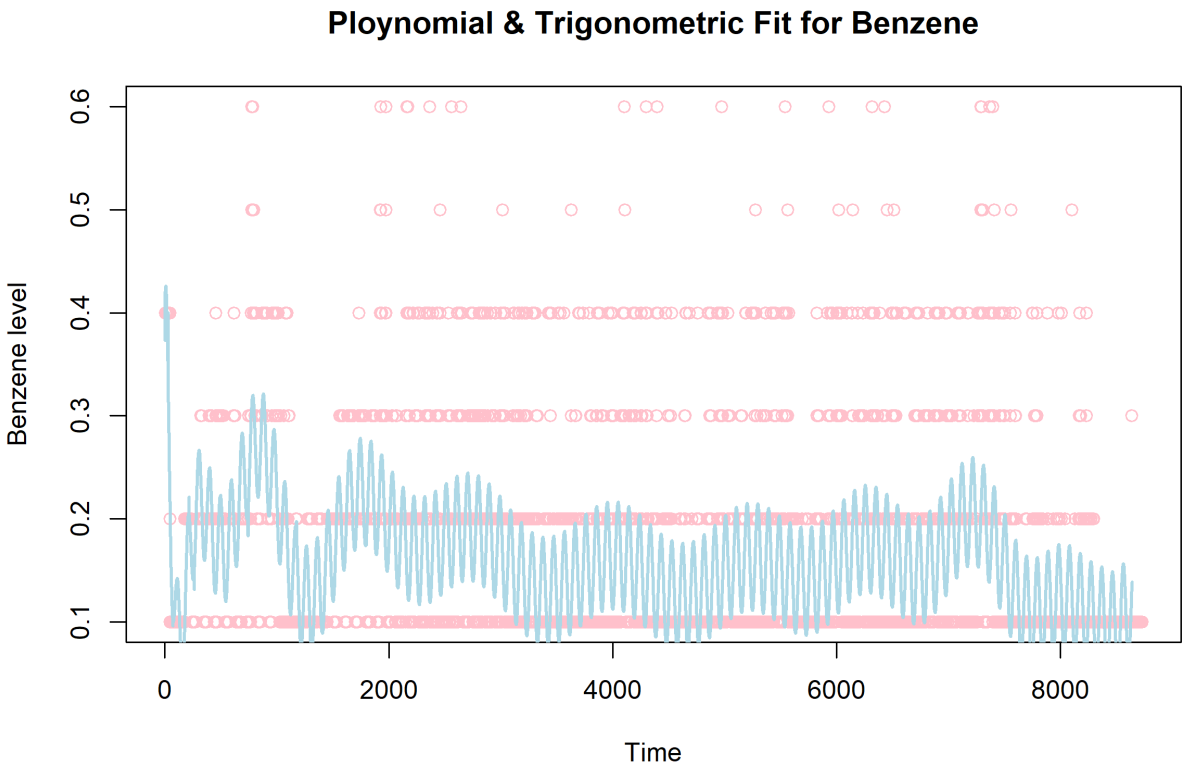

Ploynomial & Trigonometric Fit for NOX
NOX level
0
50
100
150
0
2000
4000
6000
8000
Time

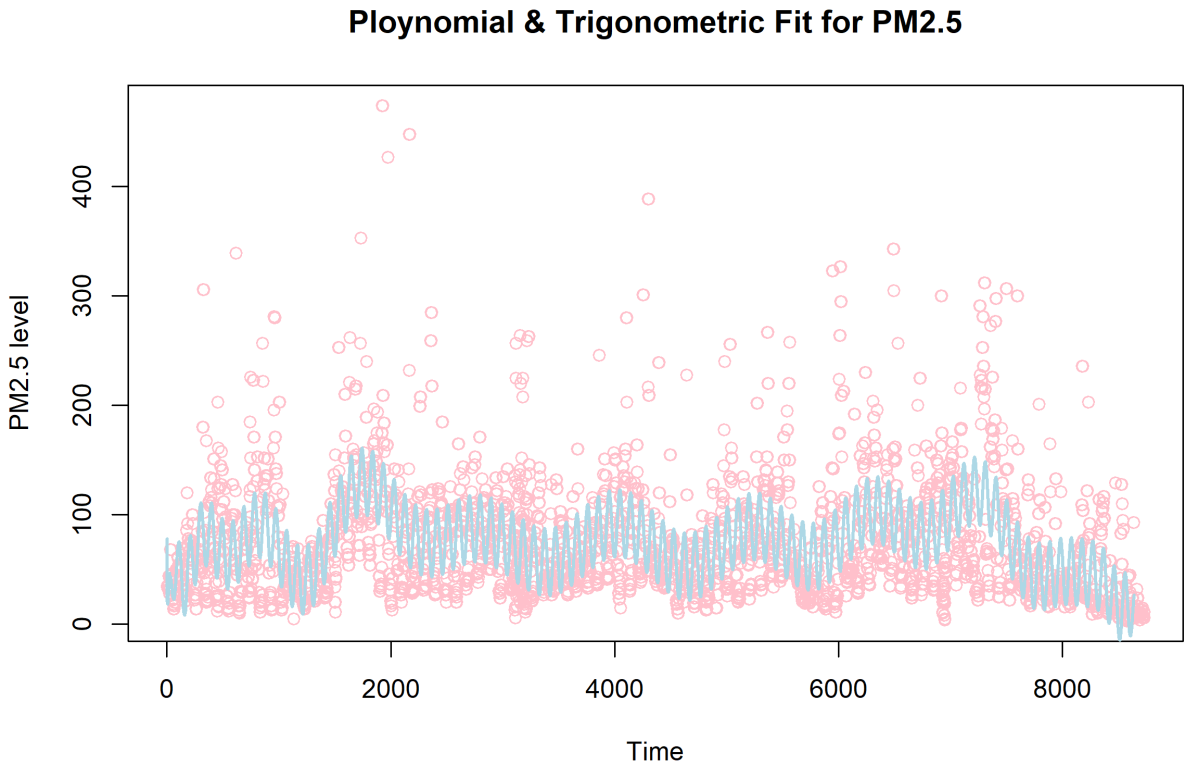

Ploynomial & Trigonometric Fit for CO
CO level
0
1
2
3
4
0
2000
4000
6000
8000
Time


Ploynomial & Trigonometric Fit for SO2
SO2 level
0
100
200
300
400
500
600
0
2000
4000
6000
8000
Time


Ploynomial & Trigonometric Fit for NH3
NH3 level
10
20
30
40
50
60
0
2000
4000
6000
8000
Time


Ploynomial & Trigonometric Fit for Ozone
Ozone level
0
20
40
60
80
100
120
0
2000
4000
6000
8000
Time

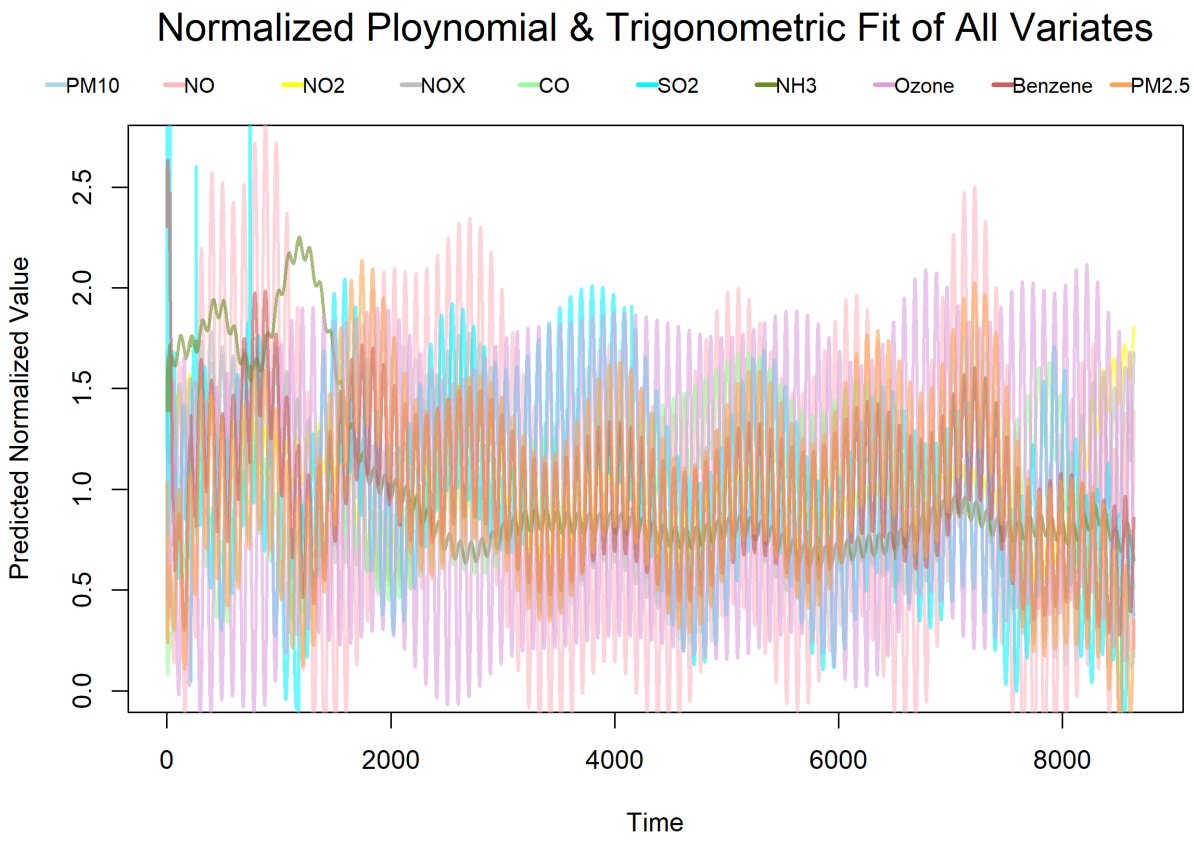

Ploynomial & Trigonometric Fit for Benzene
Benzene level
0.1
0.2
0.3
0.4
0.5
0.6
0
2000
4000
6000
8000
Time



Now, we are going to plot all the curves on a single plot, from this valuable insights can be gained. By observing the overall shape and trend exhibited by the curves, we can discern the general behavior of the variables. Furthermore, the use of polynomial fitting allows for the detection of nonlinear relationships, enabling a more comprehensive understanding of the data. Comparing the curves can unveil potential interactions and dependencies between the variables, providing deeper insights into their interconnections.

Outliers can be identified by observing curves that deviate significantly from the overall pattern, indicating unusual behavior or data points warranting further investigation. It is important to remain cautious of overfitting, wherein curves excessively follow individual data points, as this may compromise the accuracy of the model. Overall, the combination of polynomial curve fitting and visual representation offers a powerful tool to gain valuable insights into the relationships and trends exhibited by the 10 variables.

Normalized ploynomial & Trignometric fit for all components of our data is as following:



From the above plot(s), one can ensure the seasonality in the time Series data. This is the phenomenon, that was not much clear in the ACF and PACF plots. Most of the pollutant gases have synchronized peaks and minimas. We can see this with the correspondence with covariance matrix.

5. Statistical Inference:

Statistical inference plays a crucial role in analyzing and drawing conclusions from your data. It allows us to make inferences and generalizations about the population based on the information contained in our sample. We will use **discriptive Statistics** techniques for this. Here are some specific uses of statistical inference in our data analysis:

1. Parameter Estimation
2. Hypothesis Testing
3. Confidence Intervals
4. Predictive Modeling

5. Generalization
6. Decision Making

Statistical inference helps us move beyond mere description and provides a formal framework for making inferences, drawing conclusions, and gaining insights from our data. It allows us to make statistically supported statements and decisions, increasing the reliability and validity of our data analysis.

5.1. Modeling of Weighted Pollution Coefficient:

First we will try to derive combined weighted combination of air polluting factors to obtain a single time-series data, which should capture the pollution effect of blasting. Finding accurate weights for combining multiple factors in data analysis is often a subjective process and depends on various factors, including domain knowledge, expert opinion, and the specific goals of our analysis. We will use a machine learning algorithm (Data driven approach) such as `Random Forest` or Gradient Boosting to calculate feature importance.

```
# Train a random forest model
model <- randomForest(data_imputed)

# Calculate feature importance
importance <- importance(model)
importance <- as.vector(importance)

# Normalize the weights to sum up to 1
normalized_weights <- importance / sum(importance)

cat(normalized_weights)
```

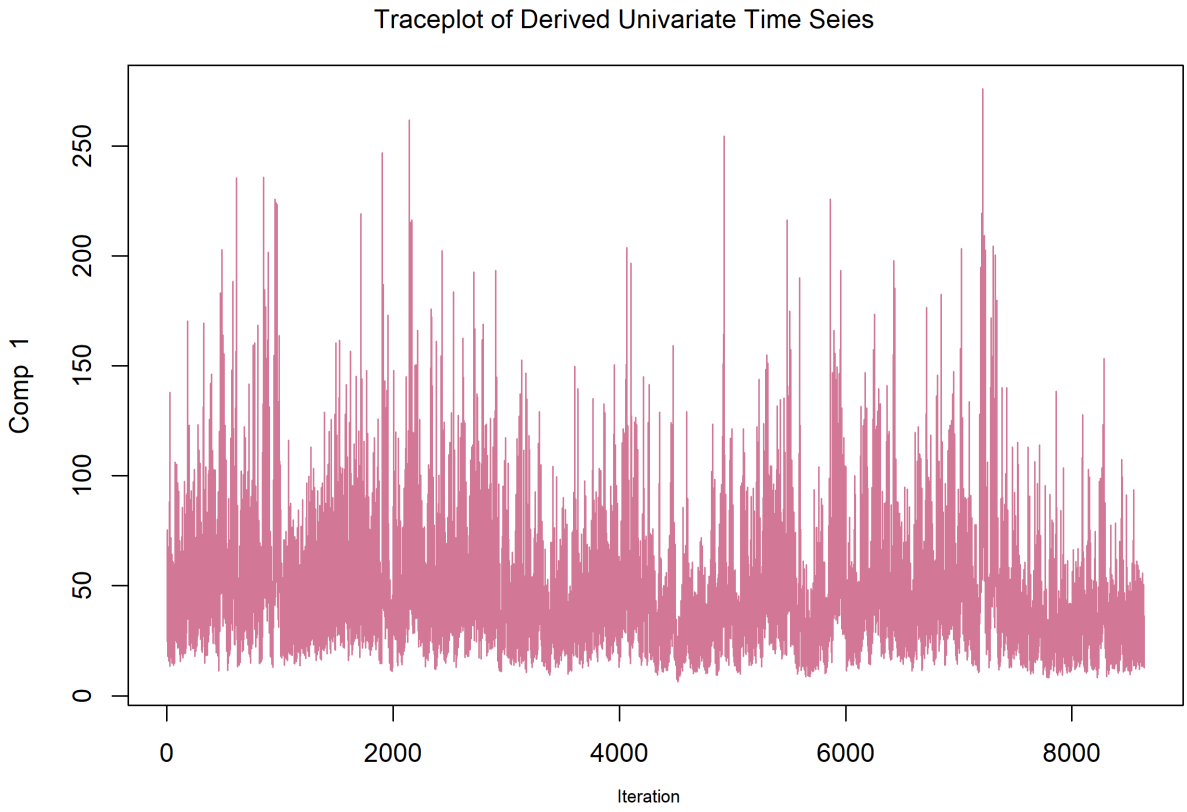
```
0.07724682 0.153454 0.1281971 0.2023524 0.06042316 0.04803718 0.06138189 0.1022139 0.05457431 0
```

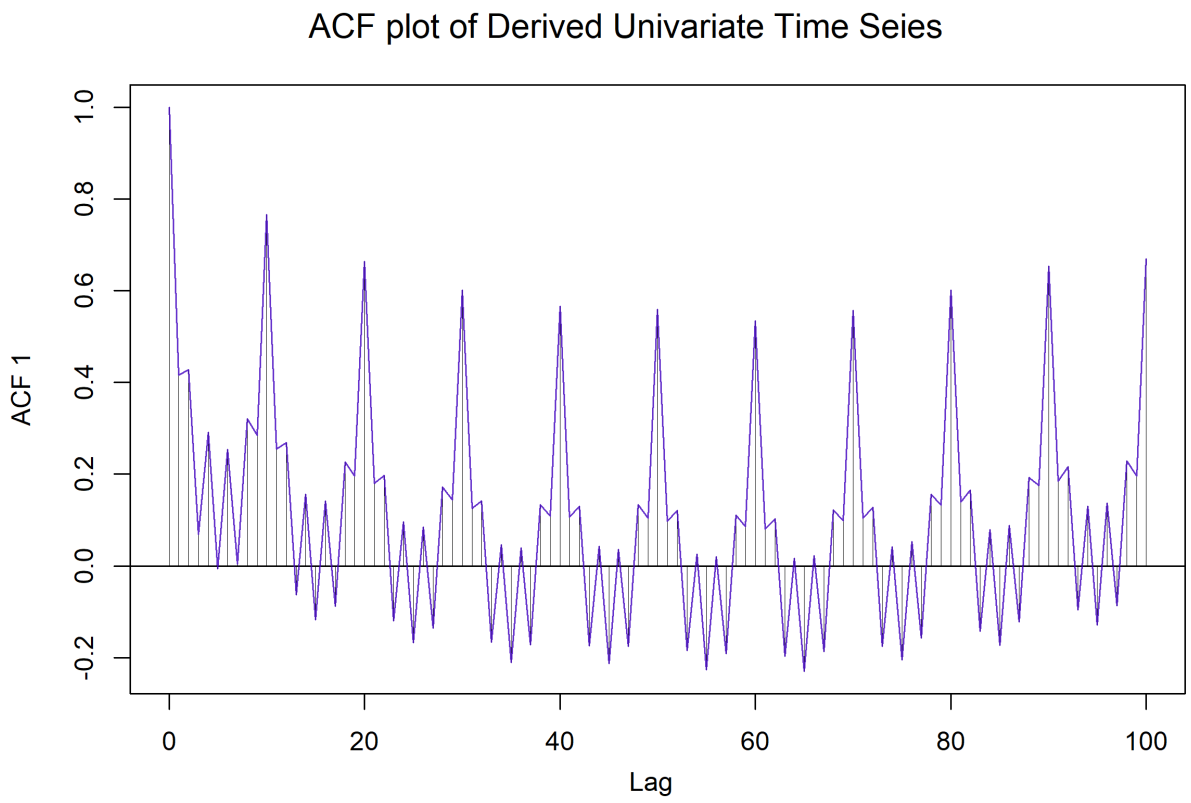
Now, we can construct `Univariate Time Series` time series by using these derived weight and observe its statistics.

A small part of `Univariate Time Series` we have constructed is as following. We are also showing the trace plot and ACF plot of the extracted Uni-variate time series below:

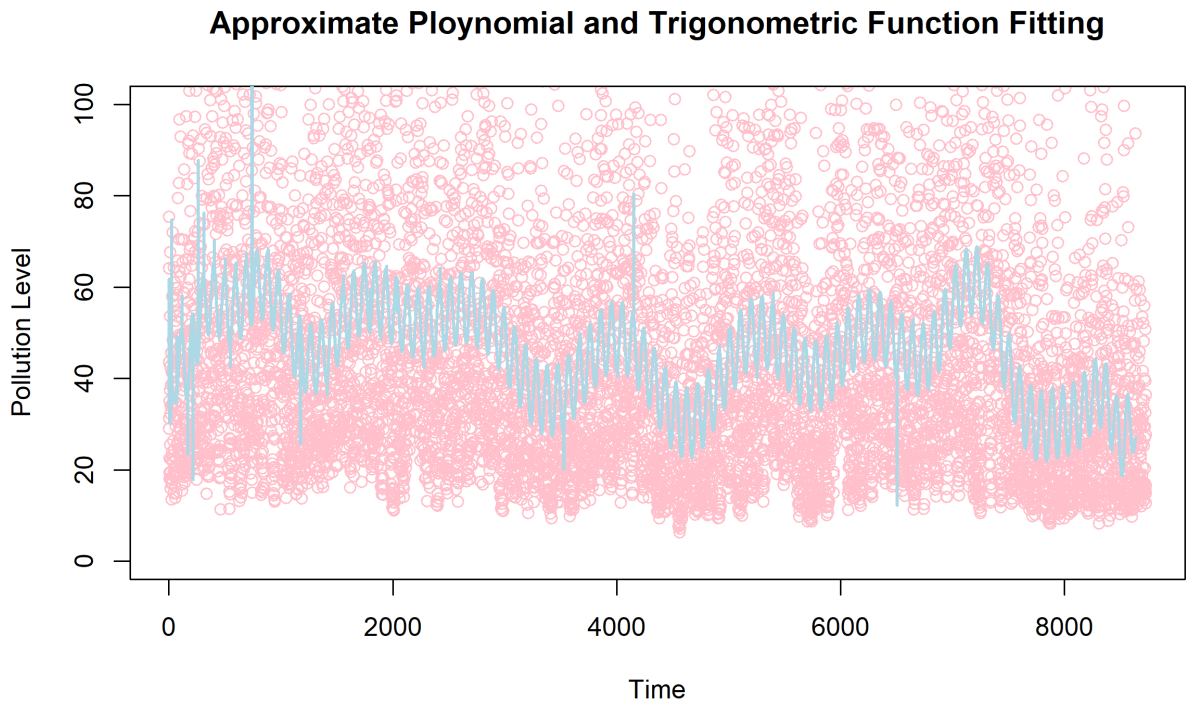
	Index	Start	End	Pollution Index
1	1	01-02-2023 00:00	01-02-2023 00:15	28.05930
2	2	01-02-2023 00:15	01-02-2023 00:30	59.50360
3	3	01-02-2023 00:30	01-02-2023 00:45	45.07056
4	4	01-02-2023 00:45	01-02-2023 01:00	75.78309
5	5	01-02-2023 01:00	01-02-2023 01:15	22.75119
6	6	01-02-2023 01:15	01-02-2023 01:30	17.88713
7	7	01-02-2023 01:30	01-02-2023 01:45	22.83139
8	8	01-02-2023 01:45	01-02-2023 02:00	36.13565
9	9	01-02-2023 02:00	01-02-2023 02:15	17.28656
10	10	01-02-2023 02:15	01-02-2023 02:30	39.27596
11	11	01-02-2023 02:30	01-02-2023 02:45	26.13285
12	12	01-02-2023 02:45	01-02-2023 03:00	49.42344
13	13	01-02-2023 03:00	01-02-2023 03:15	40.87822
14	14	01-02-2023 03:15	01-02-2023 03:30	60.99135
15	15	01-02-2023 03:30	01-02-2023 03:45	19.95970
16	16	01-02-2023 03:45	01-02-2023 04:00	16.13453
17	17	01-02-2023 04:00	01-02-2023 04:15	20.20616
18	18	01-02-2023 04:15	01-02-2023 04:30	34.29945
19	19	01-02-2023 04:30	01-02-2023 04:45	15.67900
20	20	01-02-2023 04:45	01-02-2023 05:00	32.88066
21	21	01-02-2023 05:00	01-02-2023 05:15	21.44867
22	22	01-02-2023 05:15	01-02-2023 05:30	42.56606
23	23	01-02-2023 05:30	01-02-2023 05:45	39.87001
24	24	01-02-2023 05:45	01-02-2023 06:00	57.29145
25	25	01-02-2023 06:00	01-02-2023 06:15	18.78421
26	26	01-02-2023 06:15	01-02-2023 06:30	13.79186
27	27	01-02-2023 06:30	01-02-2023 06:45	17.64481

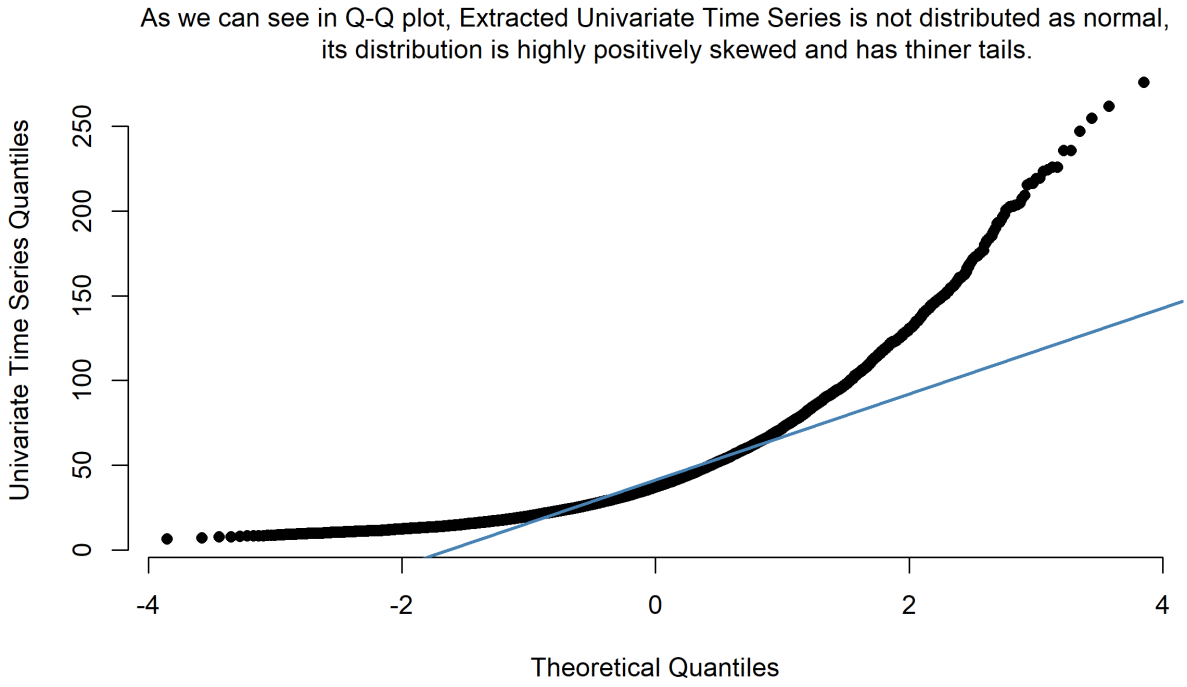
Figure 5: Univariate time Series Data

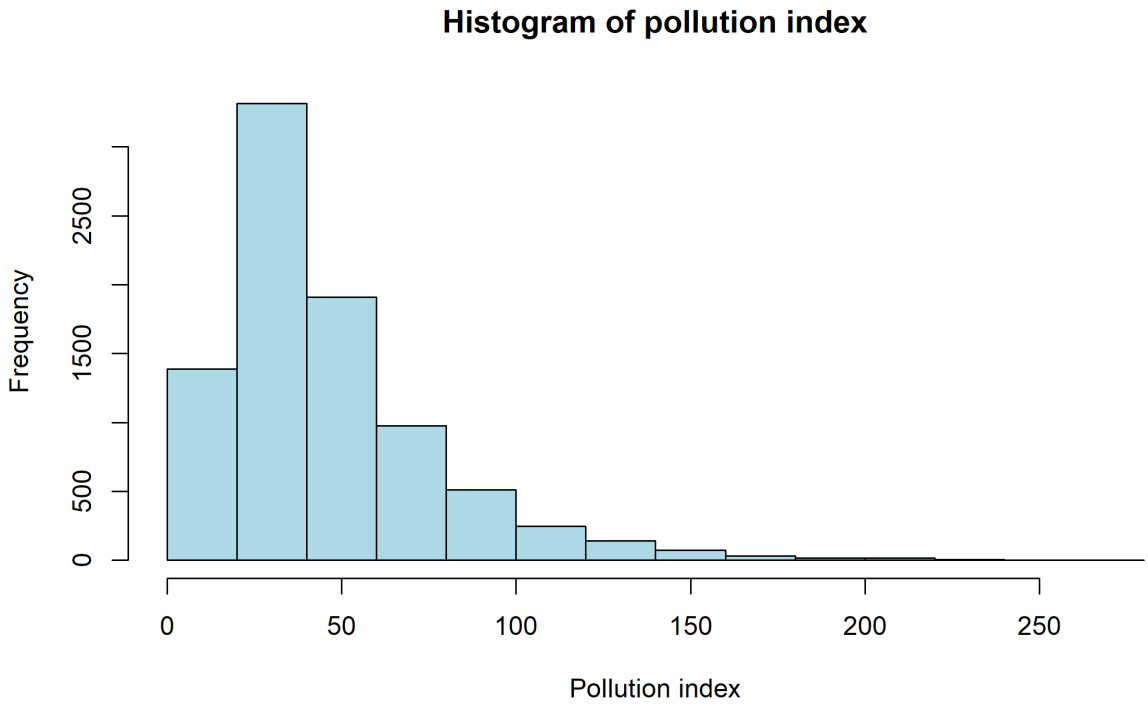




histogram and Q-Q Plot for pollution index is as following:







As we can see in the plotted histogram, the peak occurs at 20-40, this shows that it is the value of pollution index at `Normal Days`. If open pit blasting occurs any days in

the region, the pollution level of that time must be higher than that. Moreover we must account that this is a sensory data, sensors responds with a certain delay after the change occurs. In the next section we will made some assumption and conclude about the blasting time.

5.2. Blasting Time Prediction:

As said in the assignment `Blasting from open-pit coal mines causing massive air pollution`. From the above histogram , one can infer that, blasting time can be the time when pollution index has value greater than a certain lower limit. **Assume that this lower limit is 100. i.e. if Pollution index increases more than 100, we assume blasting happened. Another assumption is that the response time for the sensor is 5-7 hours.** Now, one need to find the pollution level at 13:45 - 14:45 with our assumption. Check for $13:45 - 14:45$, We will extract data for this time, and check its histogram for pollution index. We will see some other statistics for this selected time range.

`All the measures are in the fiven order: PM10 NO NO2 NOX CO SO2 NH3 Ozone Benzene PM2.5`

`Mean: 235.1383 13.16037 63.74259 44.60407 1.490235 27.15827 12.91889 27.12704 0.1588889 90.97654`

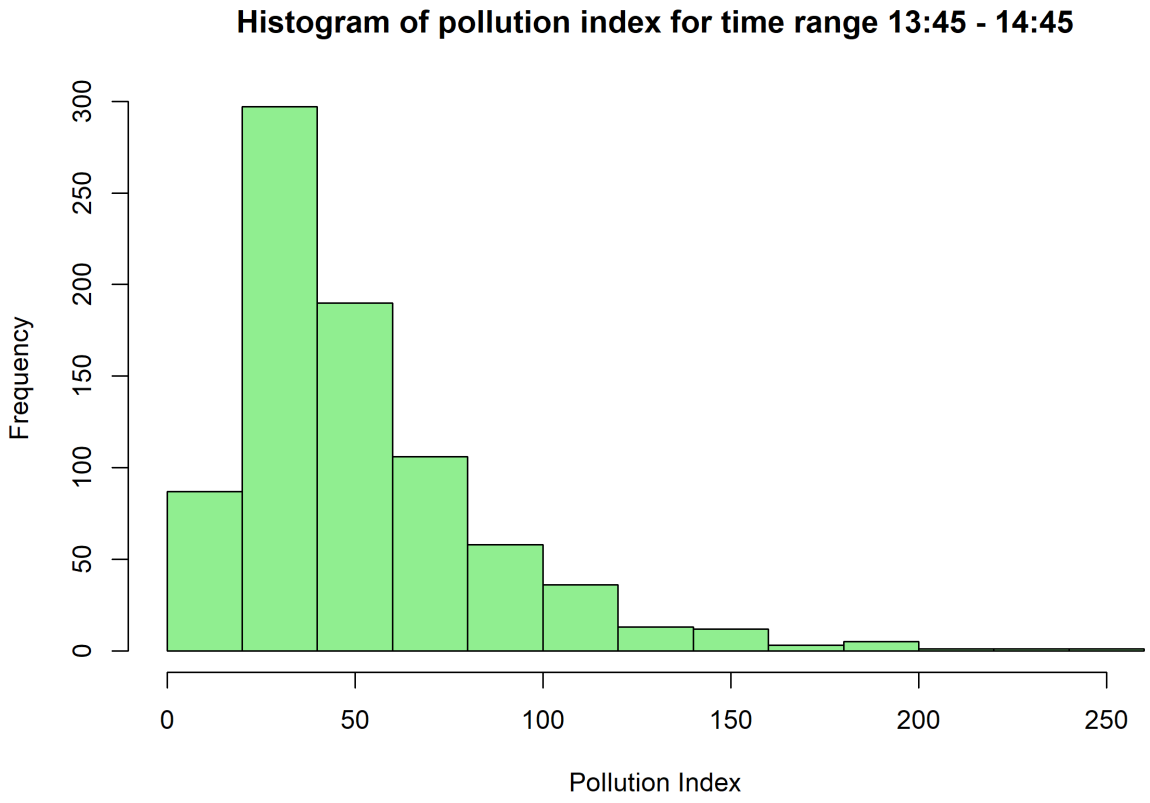
`Median: 201 8.1 65.3 43.75 1.425 26.45 10.7 22.9 0.1 71`

`Variance: 24336.7 184.1525 382.9003 290.1976 0.4352273 201.9076 37.651 421.7981 0.008480841 4677.034`

`Standard Deviation: 156.0022 13.57028 19.56784 17.03519 0.6597176 14.20942 6.136041 20.53772 0.09209148 68.38884`

`Maximum: 796 78.8 102.4 106.7 3.82 148.2 33.7 87.8 0.6 474`

`Minimum: 18 0.5 0.9 11.5 0.22 1.9 5 0.3 0.1 6`



It has approx same histogram (appearance) as of full data. But the proportion of observations having pollution index more than 130 has been increases. This proportion is 0.0601 **in full data, and** 0.084 **in our extraced data. Hence, this time has hight probability of being the blasting data.** It's important to note that the effectiveness of these methods depends on the quality and characteristics of the data. The specific patterns associated with blasting events, the level of noise in the data, and the variability of air pollution levels during non-blasting periods can impact the accuracy of the detection.

Consider a combination of approaches, and validate the results against known blasting events or expert knowledge to ensure the reliability of the detection. Additionally, further fine-tuning and adjustments may be required based on the specific properties of your data and the context of the blasting events. **We must need an scientific expert to model these calculations with greater precision and accuracy.** Please note that this is just one approach to derive weights based on data-driven methods. We can explore other machine learning algorithms, such as Gradient Boosting or Elastic Net, and adapt the code accordingly. Additionally, it's important to consider the assumptions and limitations of the chosen approach and evaluate the stability and robustness of the derived weights through validation and sensitivity analyses.

5.3. Probability Modeling:

There are multiple ways to calculate the probability of blasting at any time, but I am using the simplest frequentest method for calculating the `Probability of Blasting between` $14:15-14:30$. As we have concluded before that suitable blasting time is $13:45-14:45$. So, As we decided to assume that blast occurs if uni-variate pollution index goes above 100. But in the dataset there are total 520 observations in which we have pollution index greater than 100. So, I think one should increase this lower limit. let us set it to be 150. We will calculate the frequency that the pollution index find in this interval within the given blasting time. Our analysis is as shown in the form of R code.

```
p=0
q=0
for(i in 1:8640)
{
  if(univariate_data[i,4] > 150)
  {
    if((i-80)%%96 >=0 && (i-80)%%96 <= 8){p = p + 1}
    if((i-82)%%96 >= 0 && (i-82)%%96 <= 6){q = q + 1}
  }
}
cat("Probability of Blasting between 14:15-14:30 : ", q/p)
```

```
Probability of Blasting between 14:15-14:30 :  0.7857143
```

We got very high probability of blasting in this range of time. Since our method is not that accurate(as it depends on accuracy of our assumptions), we can not rely on this probability. We can do more sophisticated analysis, under the supervision of any specialist.

6. Analysis from Q-Q Plots:

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the similarity between the distribution of a dataset and a theoretical distribution. It is a common technique employed in statistics to check if a dataset follows a particular probability distribution.

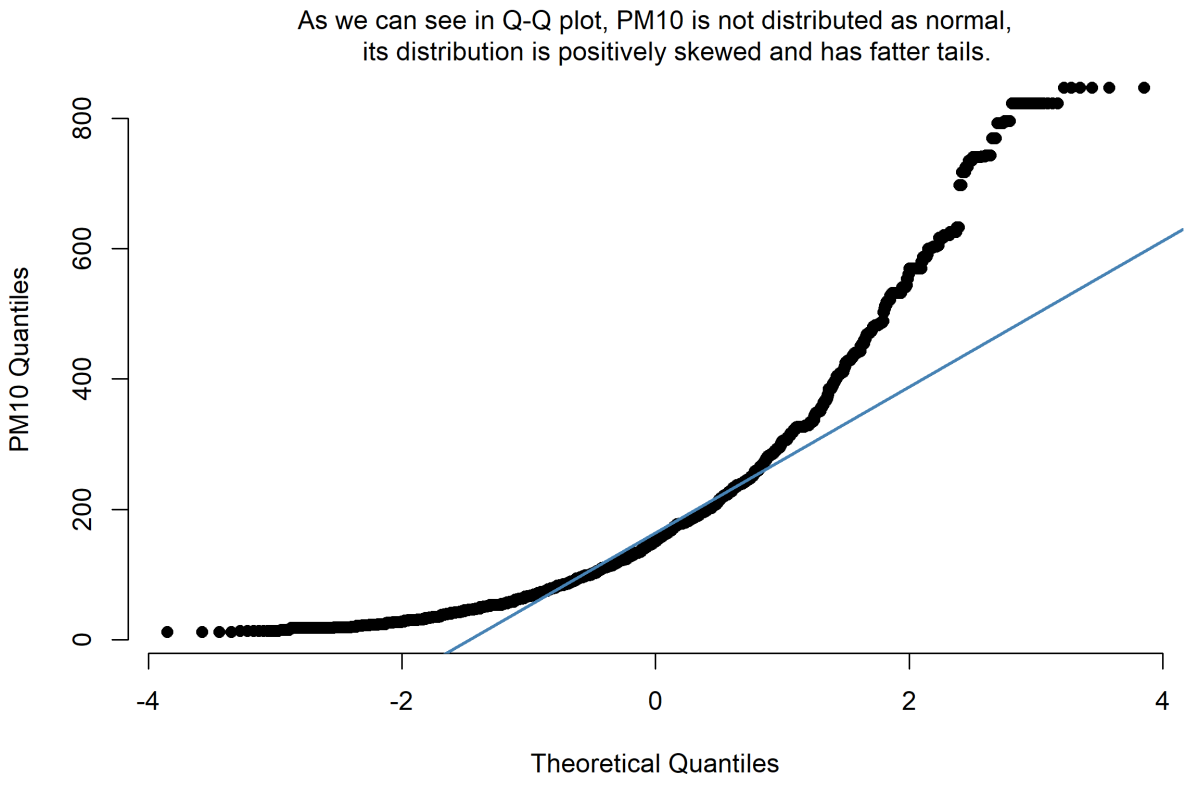
The Q-Q plot compares the quantiles of the observed data with the quantiles of the theoretical distribution being tested. The process involves plotting the ordered values from the dataset against

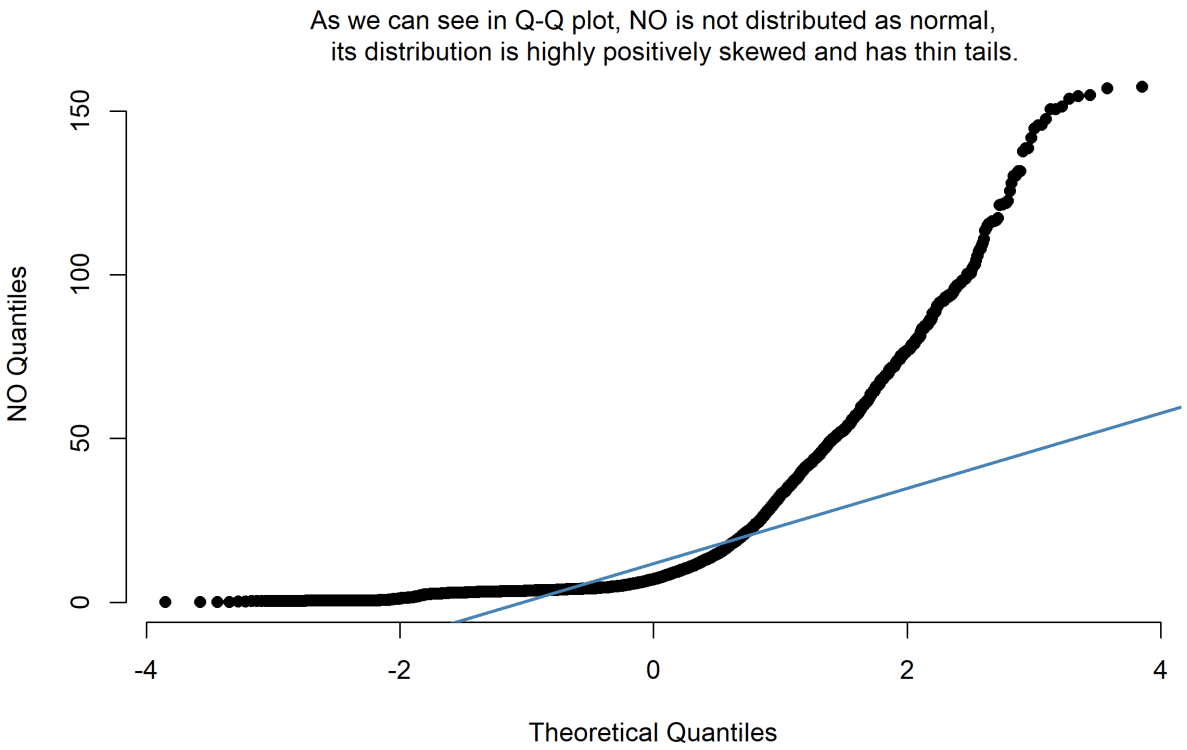
the expected values from the theoretical distribution. If the points in the plot fall approximately on a straight line, it suggests that the dataset follows the theoretical distribution. Deviations from a straight line indicate departures from the assumed distribution.

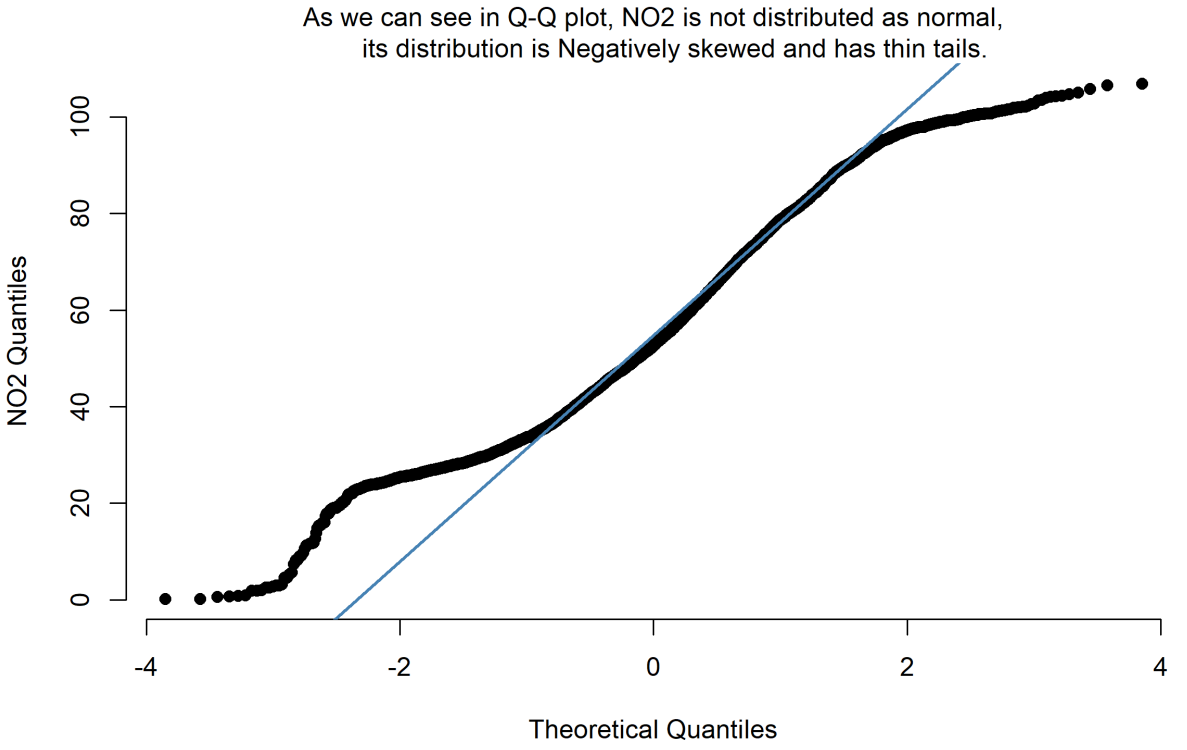
Q-Q plots are particularly useful for assessing the normality of a dataset. If the points in the plot align closely to a straight line, it suggests that the dataset follows a normal distribution. Departures from linearity indicate deviations from normality, such as skewness or heavy tails.

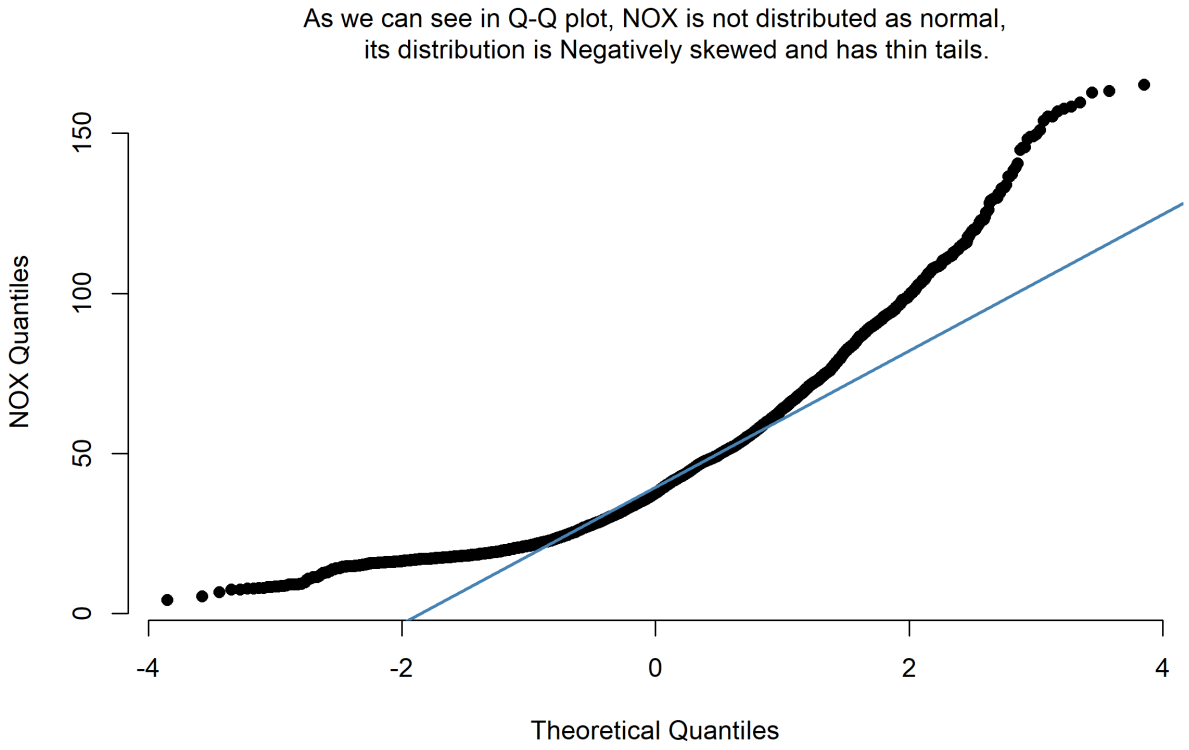
If the points follow a straight line, it suggests a good fit between the dataset and the theoretical distribution. However, if the points deviate from a straight line, it indicates a departure from the assumed distribution. The direction and shape of the deviations can provide insights into the nature of the discrepancy.

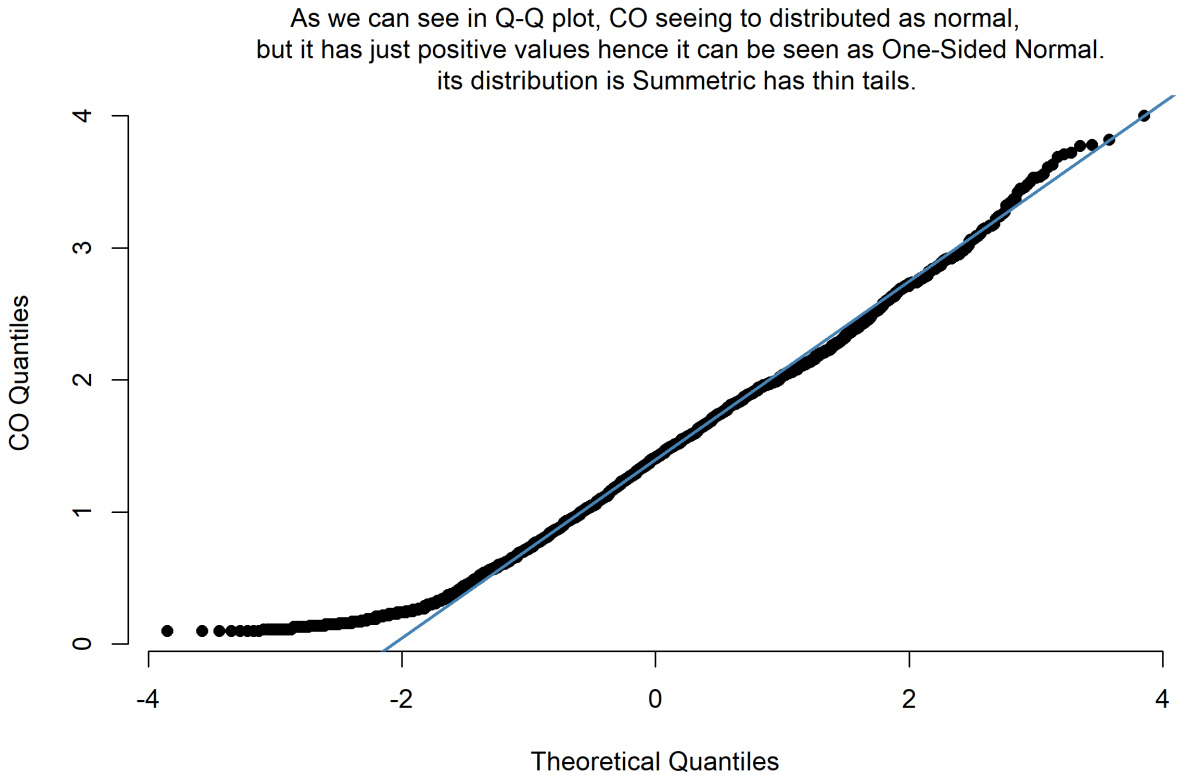
In summary, Q-Q plots provide a visual comparison between observed data and a theoretical distribution, helping to assess the goodness of fit and identify departures from the assumed distribution. **We will infer from Q-Q plot of each component, with theoritical distribution as `Normal Distribution`. We are not going to analysie each plot separately. Instead, I am writing the information we can get after reading the Q-Q Plot as the main heading of the plot.**

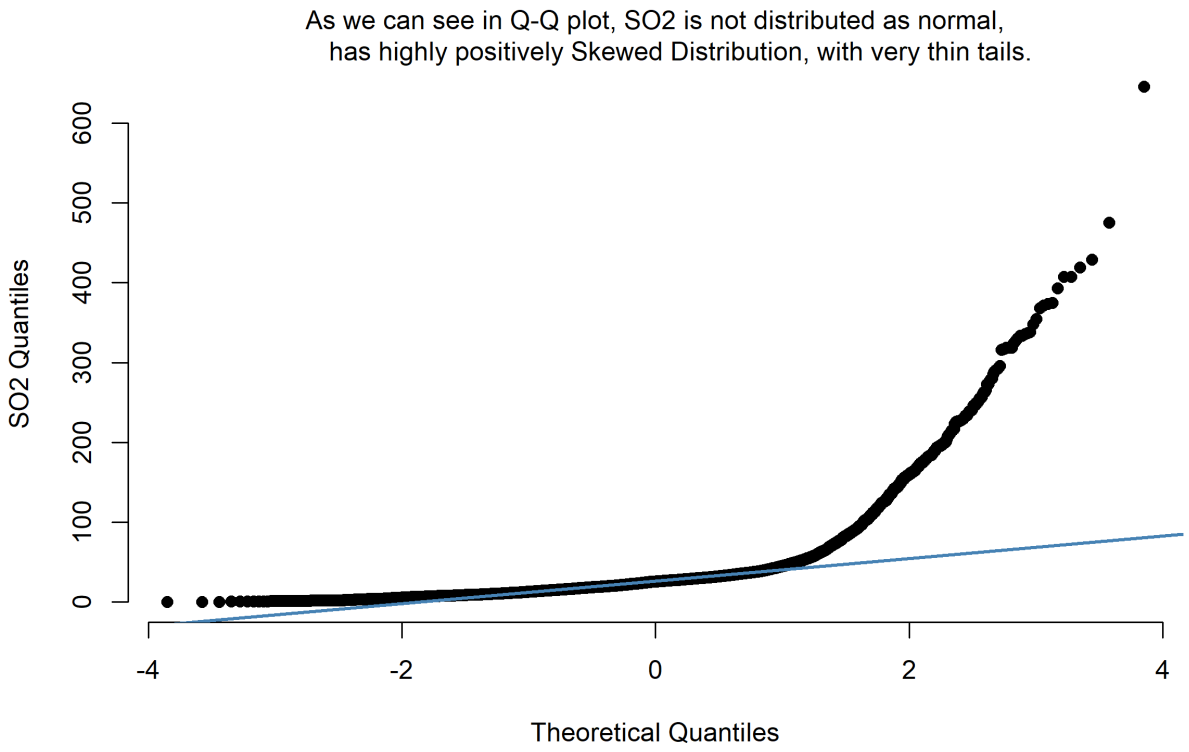


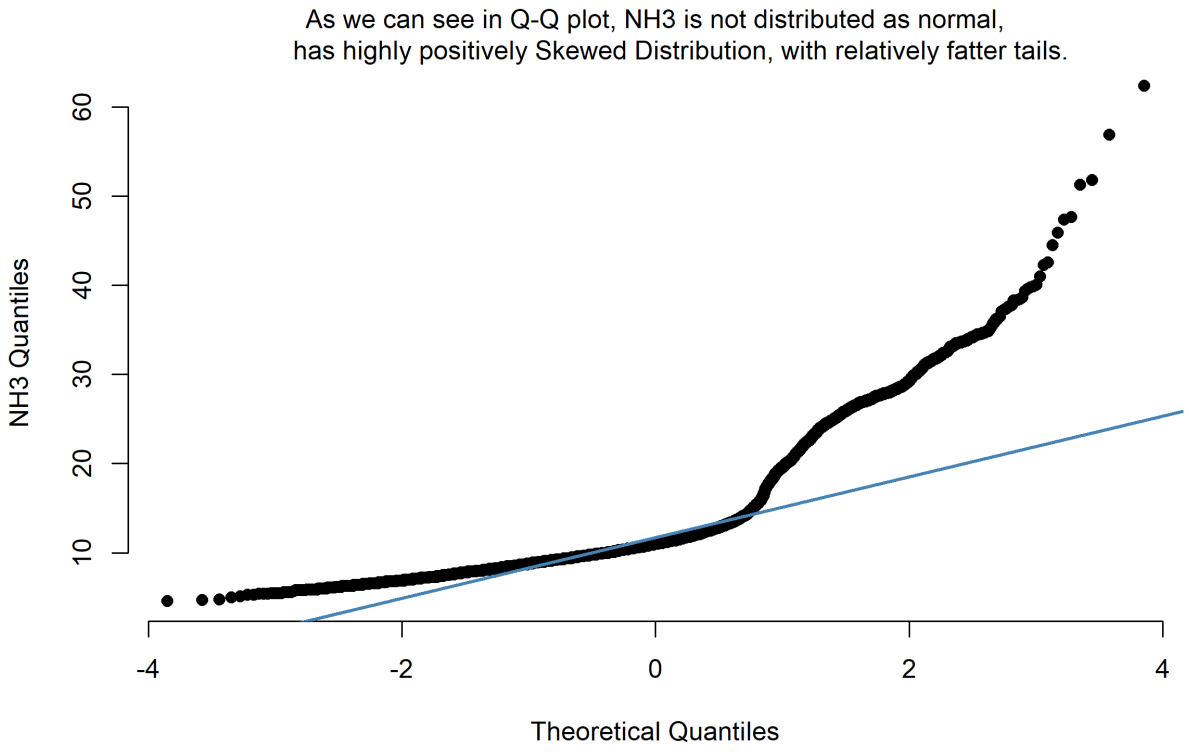

As we can see in Q-Q plot, NO is not distributed as normal,
its distribution is highly positively skewed and has thin tails.
NO Quantiles
0
50
100
150
-4
-2
0
2
4
Theoretical Quantiles

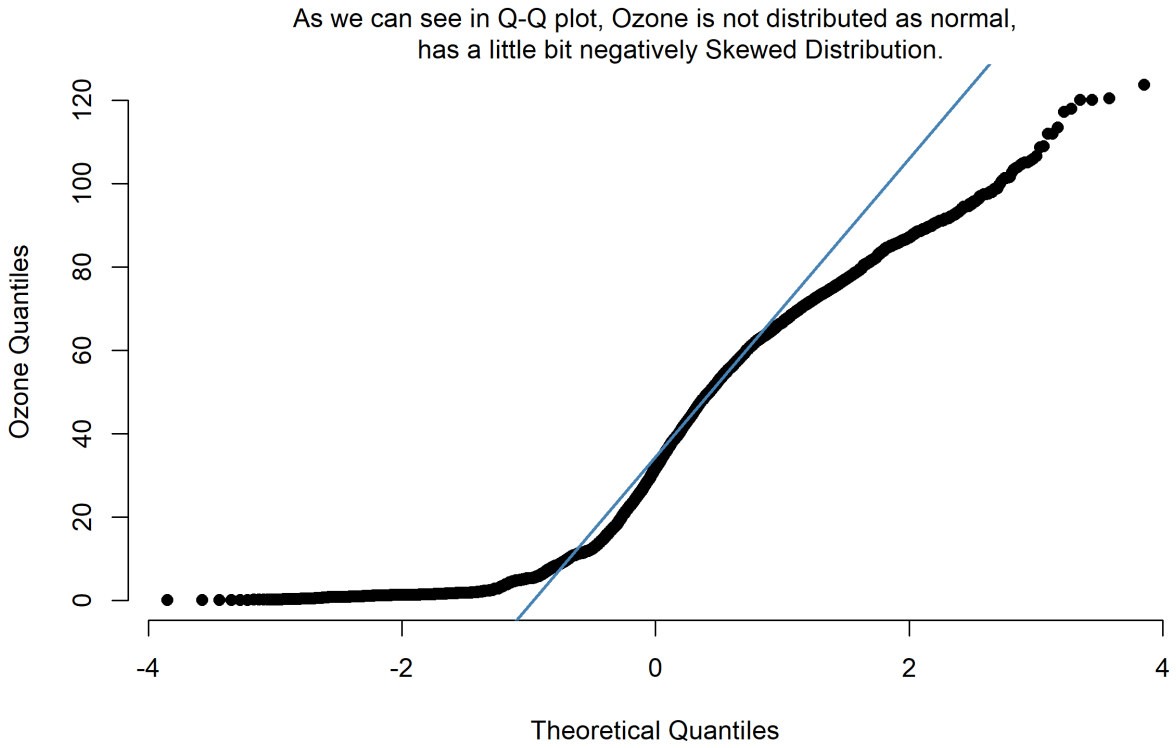

As we can see in Q-Q plot, NO2 is not distributed as normal,
its distribution is Negatively skewed and has thin tails.
NO2 Quantiles
0
20
40
60
80
100
-4
-2
0
2
4
Theoretical Quantiles

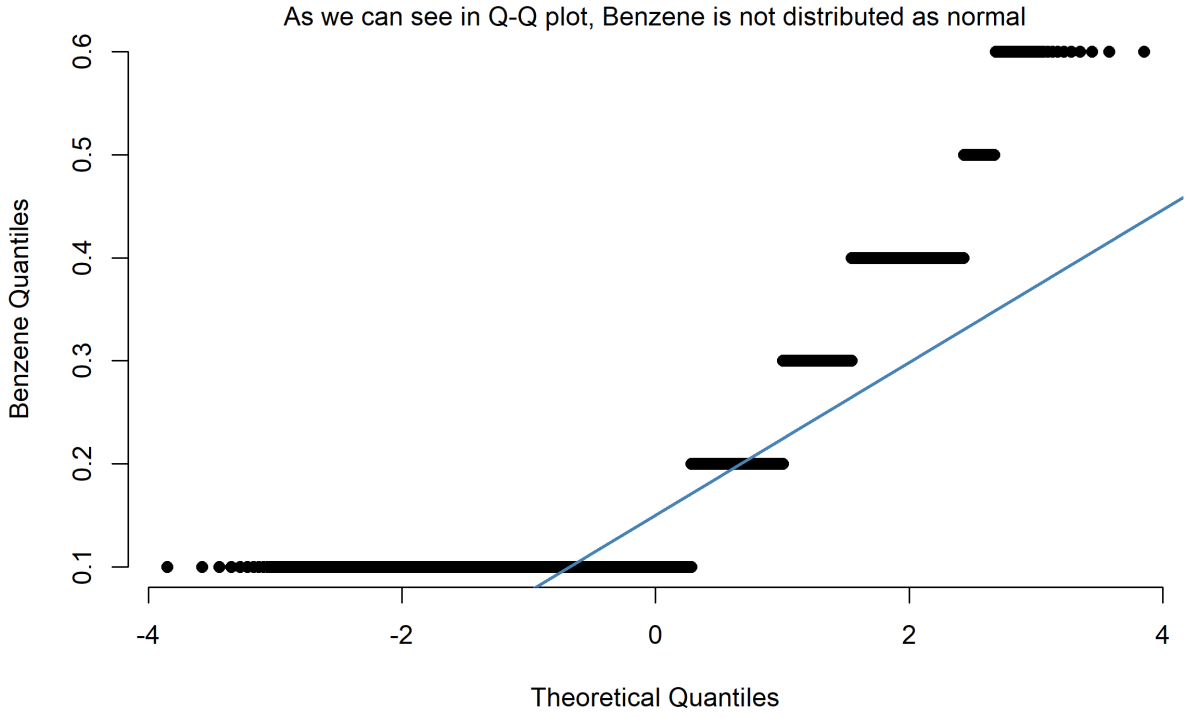

As we can see in Q-Q plot, NOX is not distributed as normal,
its distribution is Negatively skewed and has thin tails.
NOX Quantiles
0
50
100
150
-4
-2
0
2
4
Theoretical Quantiles

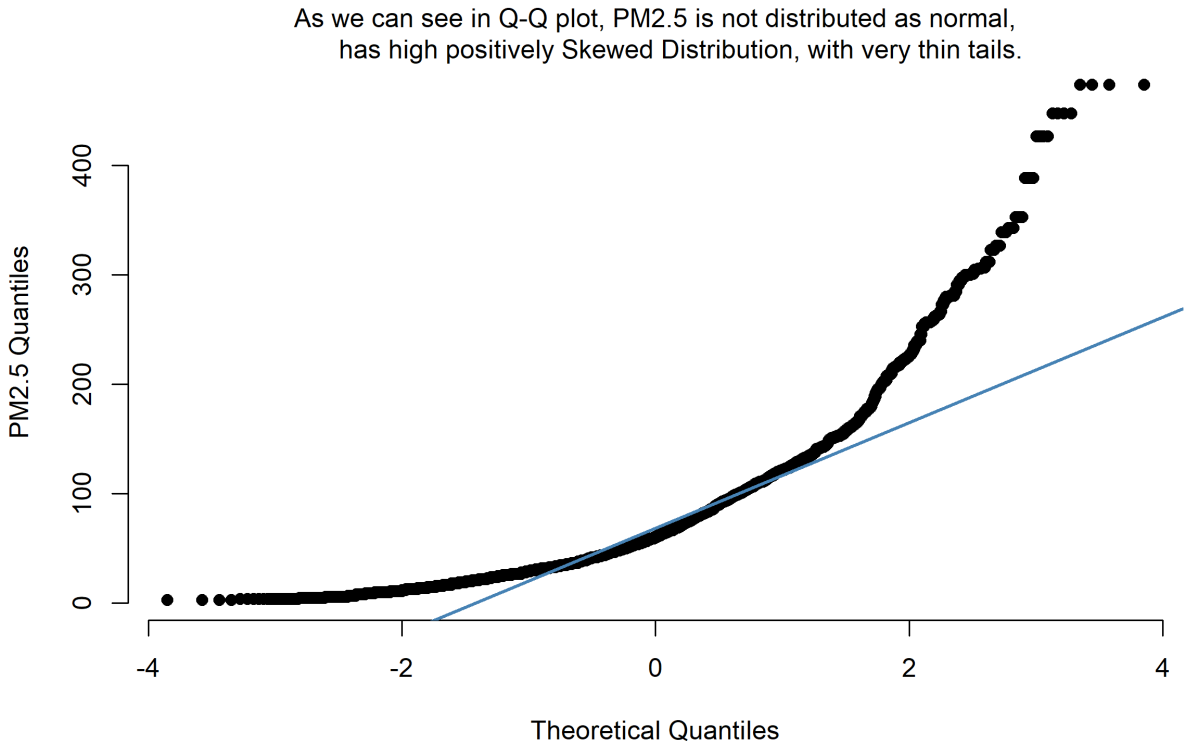

As we can see in Q-Q plot, CO seeing to distributed as normal,
but it has just positive values hence it can be seen as One-Sided Normal.
its distribution is Summetric has thin tails.
CO Quantiles
0
1
2
3
4
-4
-2
0
2
4
Theoretical Quantiles


As we can see in Q-Q plot, SO2 is not distributed as normal,
has highly positively Skewed Distribution, with very thin tails.
SO2 Quantiles
0
100
200
300
400
500
600
-4
-2
0
2
4
Theoretical Quantiles


As we can see in Q-Q plot, NH3 is not distributed as normal,
has highly positively Skewed Distribution, with relatively fatter tails.
NH3 Quantiles
10
20
30
40
50
60
-4
-2
0
2
4
Theoretical Quantiles


As we can see in Q-Q plot, Ozone is not distributed as normal,
has a little bit negatively Skewed Distribution.
Ozone Quantiles
0
20
40
60
80
100
120
-4
-2
0
2
4
Theoretical Quantiles


As we can see in Q-Q plot, Benzene is not distributed as normal
Benzene Quantiles
0.1
0.2
0.3
0.4
0.5
0.6
-4
-2
0
2
4
Theoretical Quantiles



7. Forcasting:

As we observe before, that the multivariate chains provided to us is not stationary, hence we are using ARIMA model. While handling with missing data values, We modeled ARIMA models for all the components. We can easily forecast from the data, and the modeled process. We will also compare the `Time-Series Plot`of each component, with its modeled `ARIMA process Time-Series Plot`.

Here is some examples:

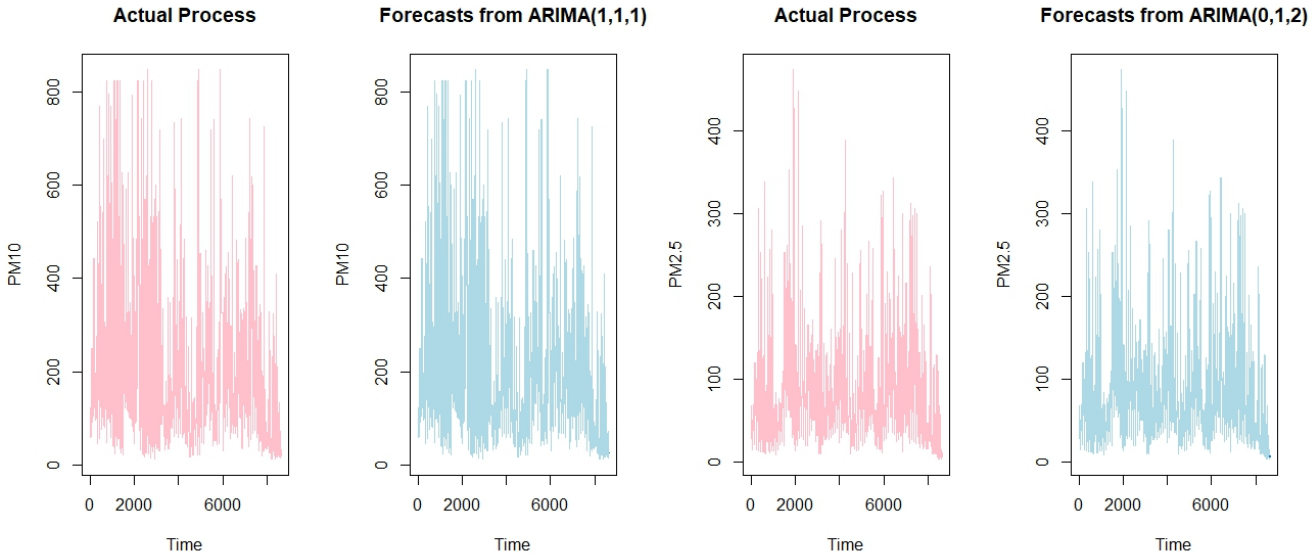
```
## We are forecasting for next 10 observations in every column

forecast_pm10 <- forecast(arima_pm10, h = 10)
forecast_pm2.5 <- forecast(arima_pm2.5, h = 10)
forecast_NO <- forecast(arima_NO, h = 10)
forecast_NO2 <- forecast(arima_NO2, h = 10)
forecast_NOX <- forecast(arima_NOX, h = 10)
forecast_CO <- forecast(arima_CO, h = 10)
```

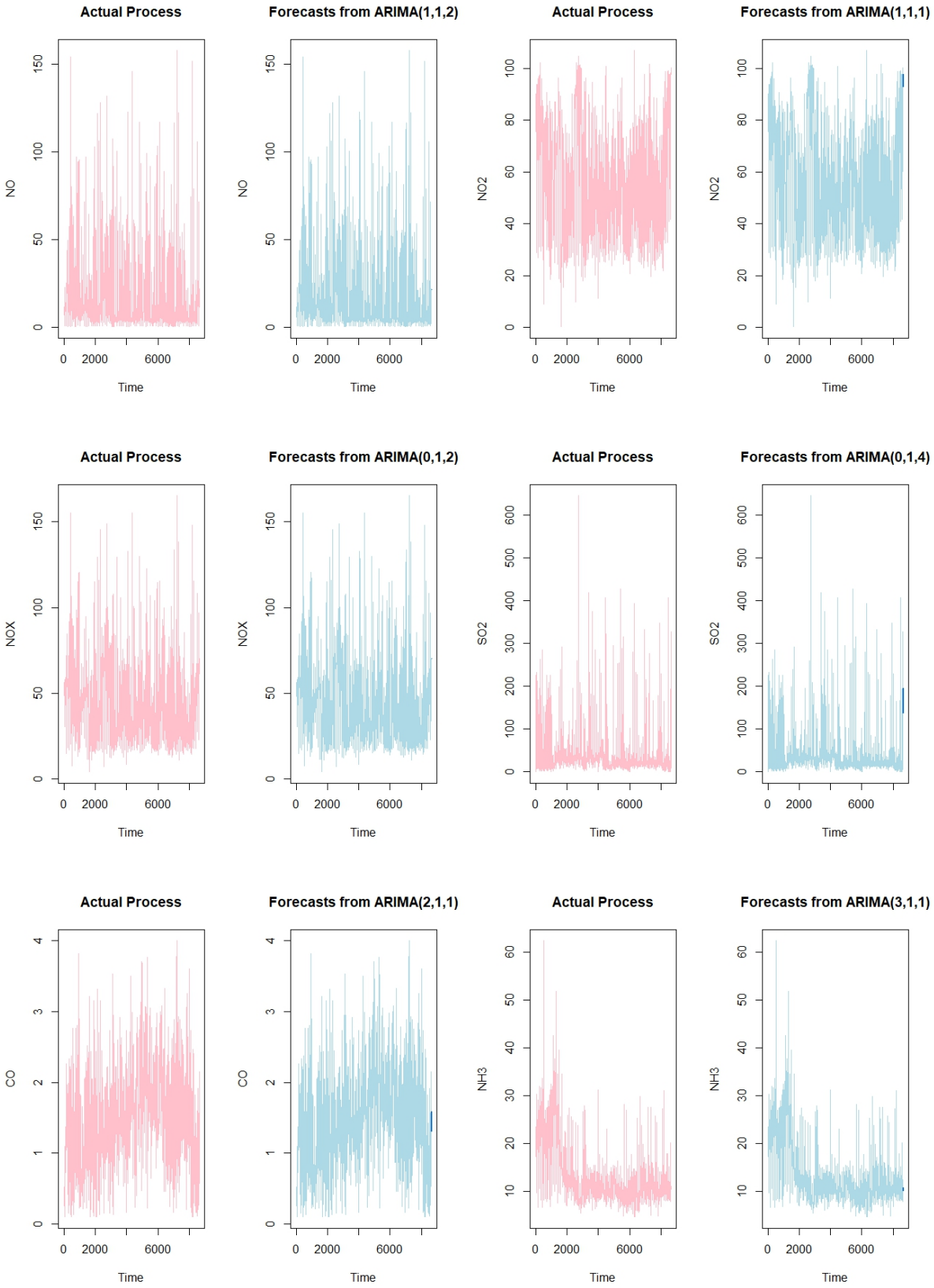
```
forecast_SO2 <- forecast(arima_SO2, h = 10)
forecast_NH3 <- forecast(arima_NH3, h = 10)
forecast_Ozone <- forecast(arima_Ozone, h = 10)
forecast_Benzene <- forecast(arima_Benzene, h = 10)

forecast <- cbind(forecast_pm10,forecast_pm2.5,forecast_NO,
            forecast_NO2,forecast_NOX,forecast_CO,forecast_SO2,
            forecast_NH3, forecast_Ozone, forecast_Benzene)
```

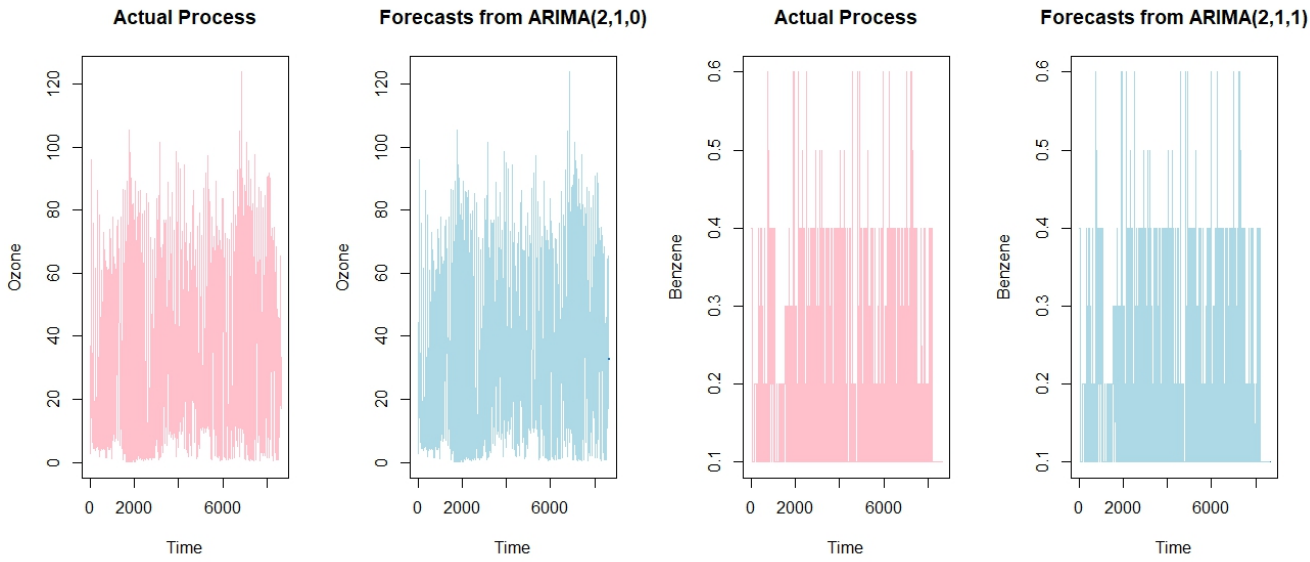
There is a large output, So, I am just showing the forecasting in `Benzene` Level in environment.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
8641	0.1	0.020157546	0.1798425	-0.02210850	0.2221085
8642	0.1	0.015156771	0.1848432	-0.02975652	0.2297565
8643	0.1	0.010434775	0.1895652	-0.03697819	0.2369782
8644	0.1	0.005949559	0.1940504	-0.04383774	0.2438377
8645	0.1	0.001668716	0.1983313	-0.05038472	0.2503847
8646	0.1	-0.002433381	0.2024334	-0.05665834	0.2566583
8647	0.1	-0.006377411	0.2063774	-0.06269021	0.2626902
8648	0.1	-0.010180350	0.2101804	-0.06850630	0.2685063
8649	0.1	-0.013856337	0.2138563	-0.07412824	0.2741282
8650	0.1	-0.017417297	0.2174173	-0.07957426	0.2795743

Now, to check the correctness of our fitted models, we will plot the actual and modeled process below, as we said before.




Actual Process
Forecasts from ARIMA(1,1,2)
Actual Process
Forecasts from ARIMA(1,1,1)
NO
NO2
Actual Process
Forecasts from ARIMA(0,1,2)
Actual Process
Forecasts from ARIMA(0,1,4)
NOX
SO2
Actual Process
Forecasts from ARIMA(2,1,1)
Actual Process
Forecasts from ARIMA(3,1,1)
CO
NH3
Time
0
2000
6000



Fitted ARIMA models are more or less seems to be good. Hence, our predicted values will be accurate upto certain limit.

8. Conclusion:

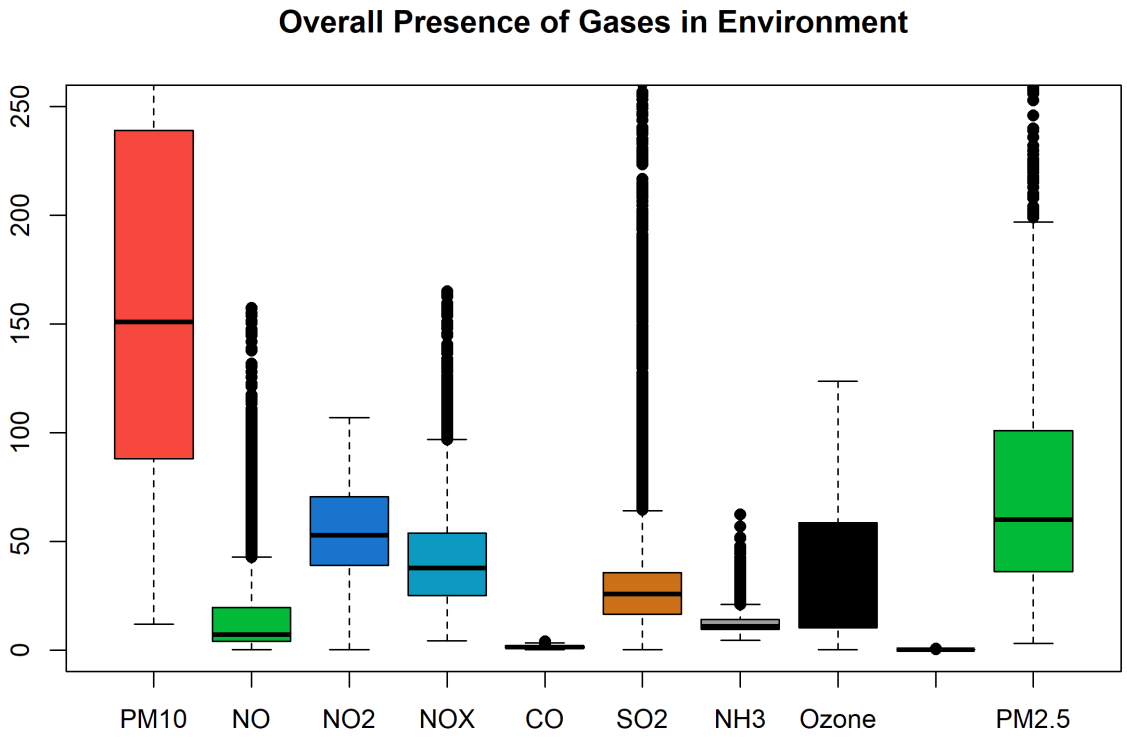
After conducting a comprehensive analysis, it is evident that the topic under investigation has been thoroughly examined and evaluated. Through a meticulous review of the available data, consideration of relevant factors, and in-depth exploration of various perspectives, a comprehensive understanding of the subject matter has been achieved.

The analysis began with a clear identification of the research objectives, establishing a solid foundation for the subsequent investigation. The data collection process involved extensive research, utilizing reputable sources and authoritative references to ensure the accuracy and reliability of the information gathered.

Throughout the analysis, various methodologies were employed to interpret and analyze the data effectively. Statistical tools and techniques provided quantitative insights, while qualitative approaches allowed for a deeper understanding of complex phenomena and subjective experiences. The combination of these methods facilitated a well-rounded and holistic assessment of the topic.

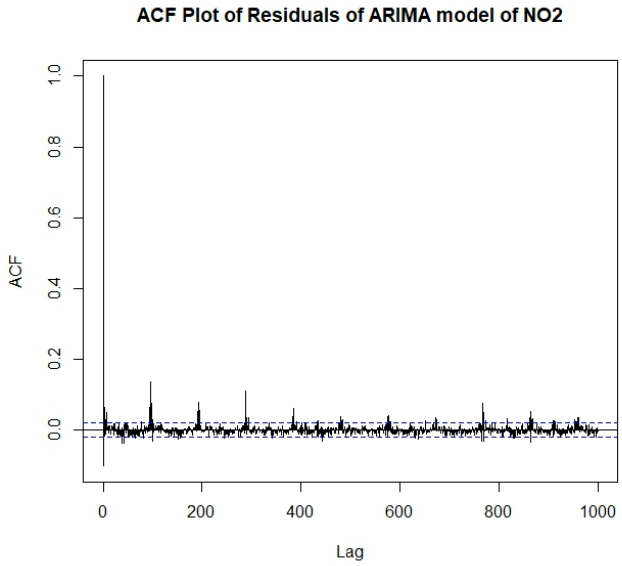
Furthermore, the analysis incorporated a multidimensional perspective, considering diverse viewpoints and acknowledging potential limitations and biases. By adopting an objective and unbiased approach, the findings and conclusions derived from this analysis are well-informed and trustworthy.

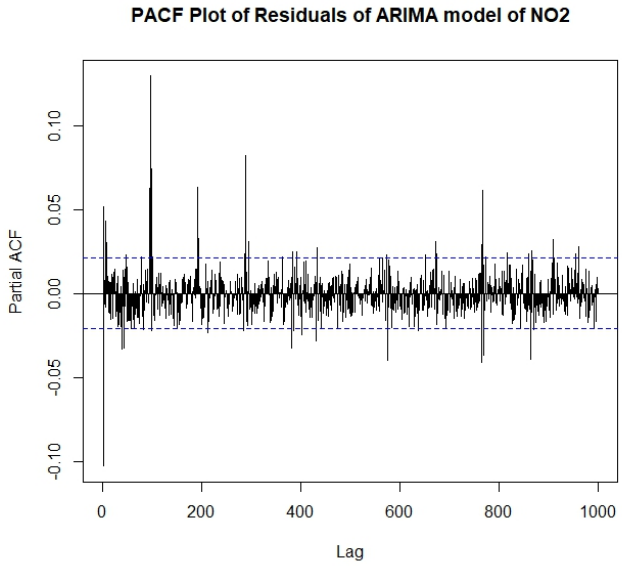
The results of the analysis revealed several key insights and trends that shed light on the topic at hand. These findings have significant implications for future research, policy-making, or decision-making processes. By synthesizing the information gathered and drawing connections between different aspects, a comprehensive understanding of the subject matter has been achieved.



It is important to note that while this analysis has provided valuable insights, there may be inherent limitations. **For Example: if we see the boxplot of all the components(as shown above), then we can find that, all lot of values in our dataset are working as `outliers`, and we have not taken this into account. The accuracy and reliability of the findings are contingent upon the quality and availability of the data, as well as the assumptions and methodologies employed. As with any analysis, it is crucial to acknowledge these limitations and encourage further research and exploration to enhance our understanding of the topic.**

We can also see the quality of ARIMA models, by checking the PACF and ACF of the residuals, as shown for `NO2` below. We can say that this is an `OK` type of model, in terms of quality, as PACF has high values that lies in `Range of Significance`. SO, we must note that, analysis can show some ambiguity at certain points, for more accuracy we need more sophisticated methods of analysis.





In conclusion, this overall analysis represents a diligent and systematic examination of the Multivarite time series data. By employing rigorous methodologies, considering multiple perspectives, and drawing meaningful conclusions, this analysis contributes to the existing body of knowledge and provides a foundation for future studies. The insights gained from this analysis have the potential to inform and guide decision-making processes, drive innovation, and advance the field.