New York Air Quality Analysis using Regression methods

Daniel Nehemiah Peter Katam
Mentor: Dr. Ali Motamedi | Pace University

PACE UNIVERSITY

Abstract:

The study emphasizes the effects of pollutants including carbon monoxide (CO), nitrogen oxides (NOx), and benzoene (C6H6) on the environment, human health, and the economy in metropolitan areas through the examination of air quality data. In order to investigate these links and create predictive models that can help with air quality management, the investigation combines statistical and machine learning methodologies. The initiative aims to give policymakers and public health professionals practical insights for implementing efficient strategies for pollutant reduction and air quality improvement by offering a thorough analysis of pollutant behaviors and their interactions with meteorological variables.

Introduction:

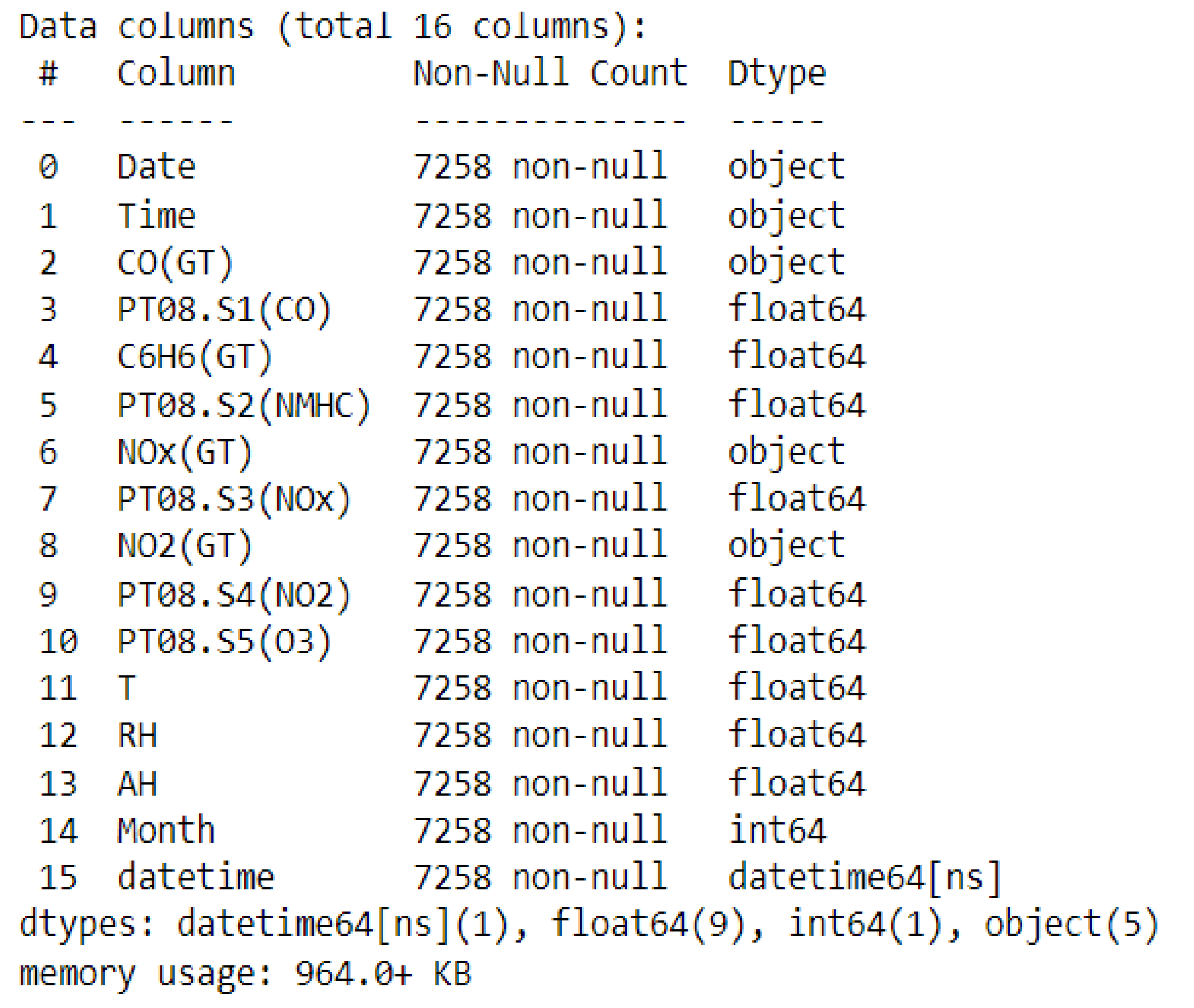
The Air Quality Dataset Analysis Project delves deeply into the intricate relationships between air pollution and its effects on public health, environmental sustainability, and economic growth. This study makes use of an extensive dataset that include hourly amounts of important pollutants, such as air pressure, temperature, humidity, and carbon monoxide (CO), as well as benzene (C6H6). The study intends to determine important patterns and connections and evaluate the impact of these contaminants by utilizing cutting-edge statistical and machine learning approaches.

Goal:

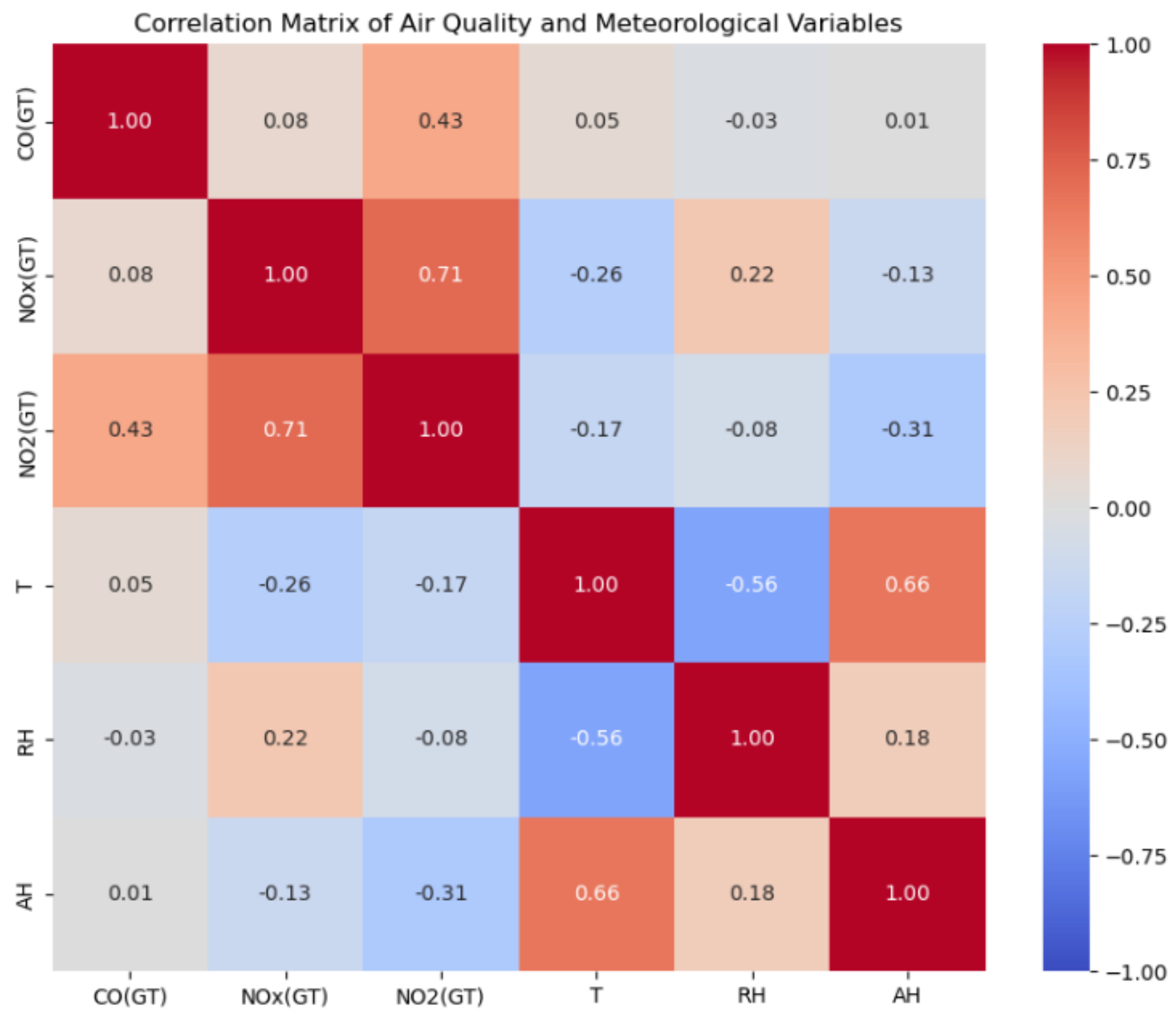
The ultimate goal is to develop robust predictive models capable of forecasting air quality variations. These models are intended to guide the formulation of targeted environmental policies and health advisories, based on a solid foundation of empirical research and data-driven insights. This project not only addresses urgent public health issues but also contributes to the formulation of smarter, more effective environmental policies.

Data:

The dataset consists of several environmental and pollutant measurements recorded on an hourly basisThis data is critical for analysing the impact of air quality on health, environmental, and economic aspects by examining correlations between air pollutants and varying environmental conditions.



```
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Date           7258 non-null   object
 1   Time           7258 non-null   object
 2   CO(GT)         7258 non-null   object
 3   PT08.S1(CO)    7258 non-null   float64
 4   C6H6(GT)       7258 non-null   float64
 5   PT08.S2(NMHC)  7258 non-null   float64
 6   NOx(GT)        7258 non-null   object
 7   PT08.S3(NOx)   7258 non-null   float64
 8   NO2(GT)        7258 non-null   object
 9   PT08.S4(NO2)   7258 non-null   float64
 10  PT08.S5(O3)    7258 non-null   float64
 11  T              7258 non-null   float64
 12  RH             7258 non-null   float64
 13  AH             7258 non-null   float64
 14  Month          7258 non-null   int64
 15  datetime       7258 non-null   datetime64[ns]
dtypes: datetime64[ns](1), float64(9), int64(1), object(5)
memory usage: 964.0+ KB
```



Correlation Matrix of Air Quality and Meteorological Variables

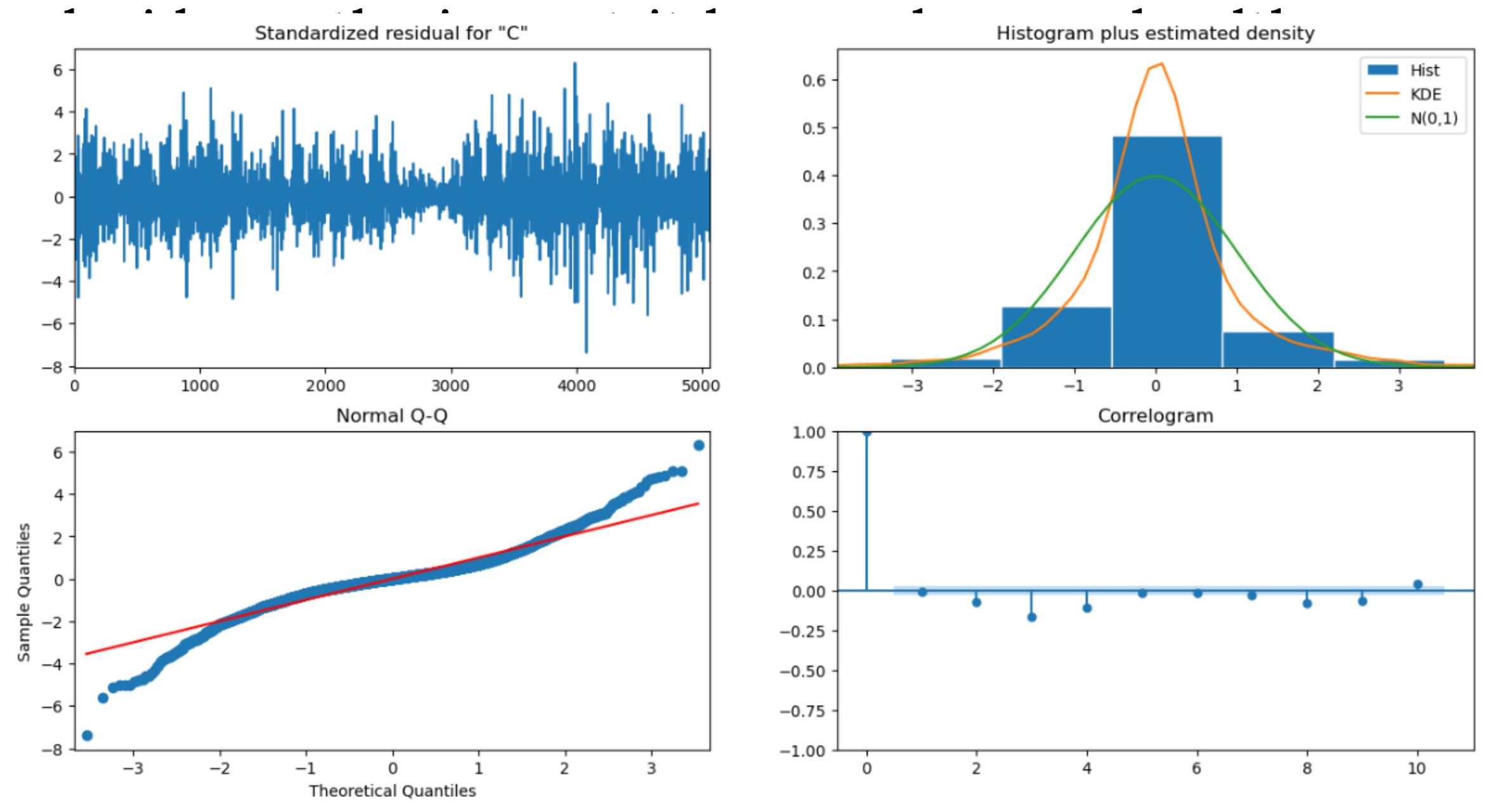
	CO(GT)	NOx(GT)	NO2(GT)	T	RH	AH
CO(GT)	1.00	0.08	0.43	0.05	-0.03	0.01
NOx(GT)	0.08	1.00	0.71	-0.26	0.22	-0.13
NO2(GT)	0.43	0.71	1.00	-0.17	-0.08	-0.31
T	0.05	-0.26	-0.17	1.00	-0.56	0.66
RH	-0.03	0.22	-0.08	-0.56	1.00	0.18
AH	0.01	-0.13	-0.31	0.66	0.18	1.00

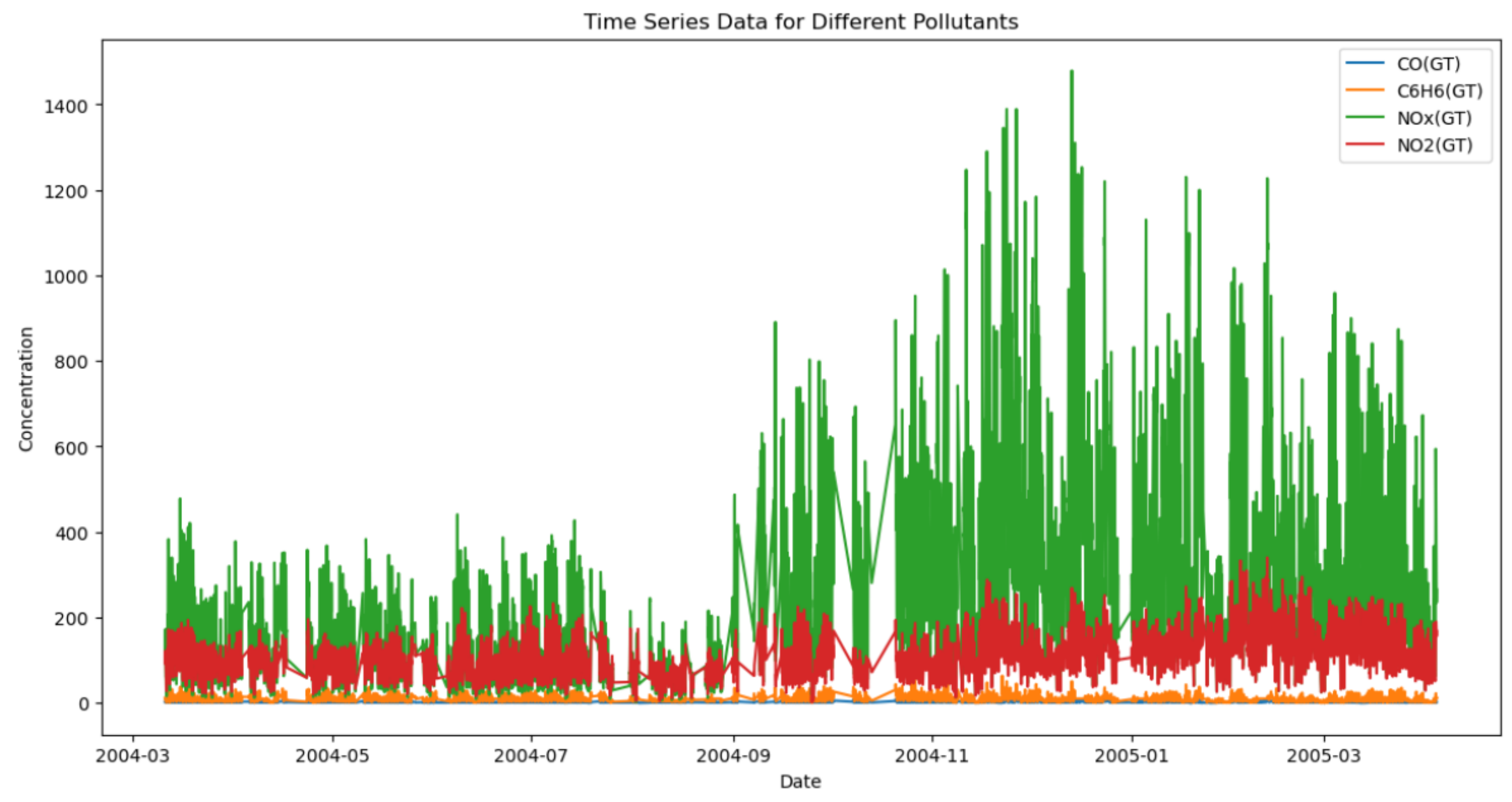
Related work:

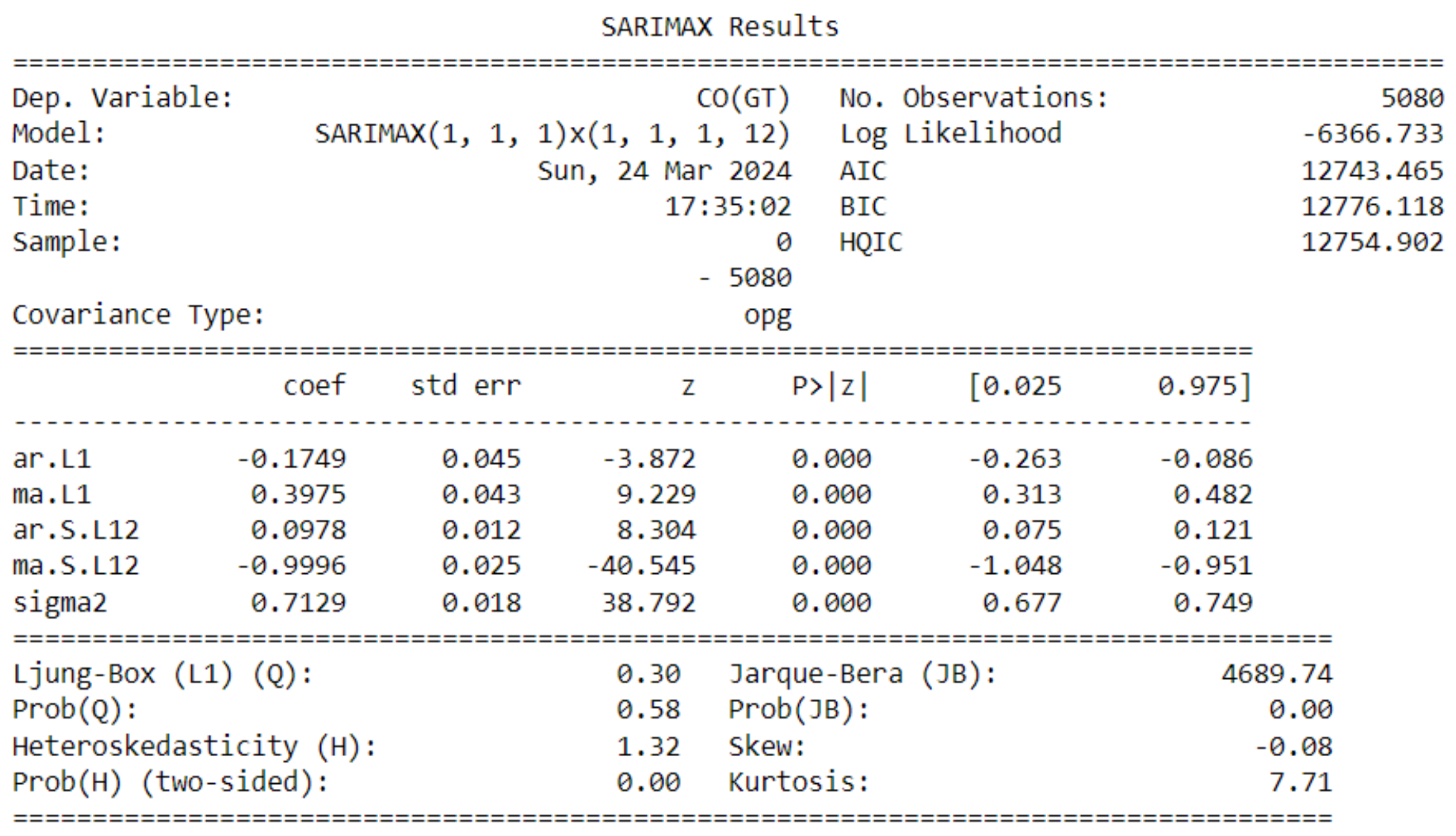
Epidemiologic studies consistently demonstrate associations between long-term air pollution exposure and mortality/morbidity even at low levels (HEI, 2010). Estimating population exposure as accurately as possible is critical for health impact assessment. Simulation studies estimate substantial avoided mortality from reducing air pollution to safe levels (Apte et al., 2015). This informs air quality management strategies and policies to protect public health.

Methodology:

Firstly, we arranged the columns, cleaned the data from missing values and outliers then performed EDA analysis to find the relation between the columns and applied Regression models to check the which pollutant impact on the quality of weather which will help u







```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                             CO(GT)   No. Observations:                 5080
Model:             SARIMAX(1, 1, 1)x(1, 1, 1, 12)   Log Likelihood               -6366.733
Date:                            Sun, 24 Mar 2024   AIC                          12743.465
Time:                                    17:35:02   BIC                          12776.118
Sample:                                         0   HQIC                         12754.902
                                           - 5080
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1749      0.045     -3.872      0.000      -0.263      -0.086
ma.L1          0.3975      0.043      9.229      0.000       0.313       0.482
ar.S.L12       0.0978      0.012      8.304      0.000       0.075       0.121
ma.S.L12      -0.9996      0.025    -40.545      0.000      -1.048      -0.951
sigma2         0.7129      0.018     38.792      0.000       0.677       0.749
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):              4689.74
Prob(Q):                              0.58   Prob(JB):                         0.00
Heteroskedasticity (H):               1.32   Skew:                            -0.08
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.71
===================================================================================
```

Results & Future work:

The time series analysis indicated that there are temporal patterns in the data, with certain pollutants showing seasonal trends or variations over time. The rolling window analysis of temperature (T), relative humidity (RH), and Absolute Humidity (AH) highlighted how these properties change over different time frames, with clear daily and seasonal patterns.

the EDA revealed that the dataset is a rich source of information on air quality, with clear indications of temporal patterns and relationships between pollutants. It also highlighted the need for careful data cleaning and pre-processing to handle missing and erroneous values. The insights gained can be used to inform further research, such as investigating the causes of pollution spikes, understanding the impact of weather on pollutant levels, and developing predictive models for air quality management.

Reference:

- What is the air quality index (AQI)? (iqair.com)
- Air pollution measurement - Wikipedia