Practical Exam Project

Applied Machine Learning & Statistics

Group Members:

- **Simran Amesar -** Preprocessing and time-series decomposition
- **Shreya Chowdary Challagulla -** Supervised Learning
- **Ishwari Thakur -** Statistical Analysis
- **Pooja Chandrappa -** Noise Injection & Cleaning
- **Yuvraj Ghag -** Unsupervised Learning

Contents

1. Introduction **2**
- 1.1 Background and Motivation 2
- 1.2 Problem Statement 2
- 1.3 Project Objectives 2
- 1.4 Dataset Overview 2

2. Dataset Description **3**
- 2.1 Data Source and Collection 3
- 2.2 Dataset Dimensions 3
- 2.3 Feature Descriptions 3
- 2.4 Data Quality Issues 4

3. Preprocessing **4**
- 3.1 Raw Data Challenges 4
- 3.2 Step-by-Step Preprocessing Implementation 5
  - Step 1: Missing Value Identification & Treatment 5
  - Step 2: DateTime Merging 5
  - Step 3: Target-Feature Separation 5
  - Step 4: Noise Injection and Denoising Strategy 6

4. Statistical Analysis **7**
- 4.1 Descriptive Statistics 7
- 4.2 Correlation Analysis 7
- 4.3 Distribution Analysis 7

5. Supervised Learning **8**
- 5.1 Feature Selection and Data Splitting 8
- 5.2 Model Training and Hyperparameters 8
- 5.3 Performance Comparison 9
- 5.4 Prediction Visualizations 9

6. Unsupervised Learning **9**
- 6.1 Clustering 10
- 6.2 Density-Based Clustering 10
- 6.3 K-Means Clustering 10

7. Time Series Decomposition **10**
- 7.1 Motivation 10
- 7.2 Additive Decomposition 11
- 7.3 Quantitative Component Summary 11
- 7.4 Visual Interpretation 11
- 7.5 Implications for Forecasting 12

8. References **13**

1. Introduction

1.1 Background and Motivation

Urban air pollution is a major public health concern, and Carbon Monoxide (CO) is particularly dangerous due to its colorless, odorless nature and strong impact on cardiovascular and respiratory health. Accurate short-term CO forecasts help authorities trigger warnings, adjust traffic, and protect vulnerable populations such as children, the elderly, and people with pre-existing conditions.

1.2 Problem Statement

Given 13 environmental sensor readings (gas sensors + meteorological variables) at each hour, predict the true CO concentration CO(GT) in mg/m³ measured by the reference analyzer.
The core challenges are:

(i) handling substantial missing values coded as −200
(ii) sensor noise and artificial spikes,
(iii) strong multicollinearity between electrochemical gas sensors, and
(iv) preserving temporal order for time-series-consistent model evaluation.

1.3 Project Objectives

- Build a preprocessing pipeline that can clean and impute noisy air quality sensor data with minimal leakage.
- Train and compare several supervised regression models (Linear Regression, Decision Tree, Random Forest, Gradient Boosting) to forecast hourly CO(GT).
- Apply unsupervised clustering on denoised features to identify typical “pollution regimes”.
- Interpret models and clusters to derive actionable insights for air quality monitoring and traffic/alert policies.

1.4 Dataset Overview

The Air Quality Data Set from the UCI Machine Learning Repository contains 9,357 hourly observations collected between 10 March 2004 and 4 April 2005 at a single urban background station in Italy. The original table has 15 columns (Date, Time, 13 numeric sensors), of which CO(GT) is the primary target. Predictors include gas sensor channels (PT08 series), true reference analyzers for NMHC, NOx, NO2, benzene, and meteorological variables (temperature T, relative humidity RH, absolute humidity AH). The dataset is affected by 16,701 missing entries (18% of all numeric entries), with NMHC(GT) missing in ~90% of rows and CO(GT) missing in ~18% of rows.

2. Dataset Description

2.1 Data Source and Collection

The dataset was obtained from the UCI Machine Learning Repository (Air Quality Data Set, DOI: 10.24432/C5MW2G), originally collected by the Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA). Hourly measurements span from 10-03-2004 to 04-04-2005 at a single urban background monitoring station in Italy.

2.2 Dataset Dimensions

Total instances: 9,357 hourly observations.
Total variables: 15 columns (2 temporal + 13 numeric environmental variables).
Target variable: CO(GT) – true hourly averaged CO concentration in mg/m³ (reference analyzer).
Effective modeling features after target separation: 12 numeric predictors (all columns except CO(GT)).

2.3 Feature Descriptions

Variable Name	Type	Description	Units
Date	Date		
Time	Categorical		
CO(GT)	Integer	True hourly averaged concentration CO in mg/m^3 (reference analyzer)	mg/m^3
PT08.S1(CO)	Categorical	hourly averaged sensor response (nominally CO targeted)	
NMHC(GT)	Integer	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)	microg/m^3
C6H6(GT)	Continuous	True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)	microg/m^3
PT08.S2(NMHC)	Categorical	hourly averaged sensor response (nominally NMHC targeted)	
NOx(GT)	Integer	True hourly averaged NOx concentration in ppb (reference analyzer)	ppb

PT08.S3(NOx)	Categorical	hourly averaged sensor response (nominally NOx targeted)	
NO2(GT)	Integer	True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)	microg/m^3
PT08.S4(NO2)	Categorical	hourly averaged sensor response (nominally NO2 targeted)	
PT08.S5(O3)	Categorical	hourly averaged sensor response (nominally O3 targeted)	
T	Continuous	Temperature	°C
RH	Continuous	Relative Humidity	%
AH	Continuous	Absolute Humidity	

2.4 Data Quality Issues

Missingness: Values of −200 were used to encode sensor failures and were replaced with NaN; this yields 16,701 missing values across 13 numeric columns (18% of all numeric cells). Most affected: NMHC(GT) has 8,443 missing entries (~90% of its rows), followed by CO(GT) (1,683 missing, ~18%) and several PT08 channels and meteorological variables (366 missing each). Noise and outliers: Raw PT08 channels and CO(GT) show extreme spikes and heavy tails, requiring clipping or robust smoothing.

3. Preprocessing

3.1 Raw Data Challenges

The raw UCI data use −200 as a hard-coded missing flag in all numeric sensors, which must be converted to true NaN for correct imputation. Date and Time are provided as separate object columns, requiring parsing and merging into a proper DateTime index for any time-aware splitting.

3.2 Step-by-Step Preprocessing Implementation

Step 1: Missing Value Identification & Treatment

```
print("Missing values per column:\n")
print(df.isnull().sum())
```

```
Missing values per column:

Date                0
Time                0
CO(GT)           1683
PT08.S1(CO)       366
NMHC(GT)         8443
C6H6(GT)          366
PT08.S2(NMHC)     366
NOx(GT)          1639
PT08.S3(NOx)      366
NO2(GT)          1642
PT08.S4(NO2)      366
PT08.S5(O3)       366
T                 366
RH                366
AH                366
dtype: int64
```

All −200 values were replaced by NaN, revealing 16,701 missing values across 13 numeric variables.

Missing counts: NMHC(GT) 8,443; CO(GT) 1,683; NOx(GT) 1,639; NO2(GT) 1,642; each PT08 sensor, T, RH, AH: 366 missing.
Median imputation was applied columnwise to all numeric features, reducing the total missing count from 16,701 to 0.
Rationale: Median is robust to right-skewed pollutant distributions and preserves central tendency better than the mean in the presence of outliers.

Step 2: DateTime Merging

Date and Time strings were concatenated and parsed into a single DateTime column, which was then set as the index, and the original Date/Time columns were dropped. This yields a clean chronological index that can be used for time-aware analysis and non-shuffled train-test splits.

```
df['DateTime'] = pd.to_datetime(df[['Date', 'Time']].astype(str).agg(' '.join, axis=1))
df.set_index('DateTime', inplace=True)
df.drop(['Date', 'Time'], axis=1, inplace=True)
```

The columns Time and Date has been converted into one column DateTime

Step 3: Target-Feature Separation

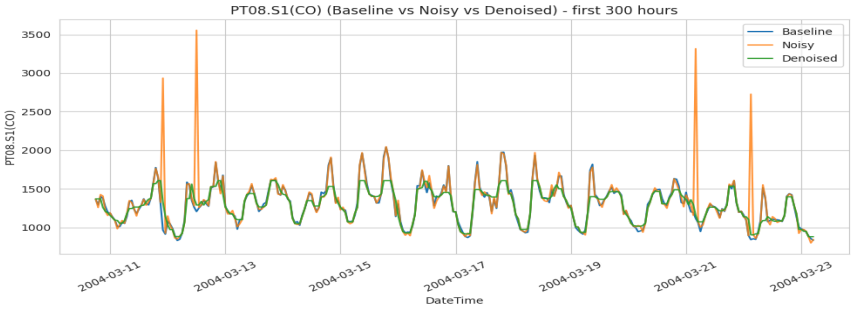
```
target_col = 'CO(GT)'
X = df.drop(columns=[target_col])
y = df[[target_col]]

print(f"Features shape: {X.shape}, Target shape: {y.shape}")
```

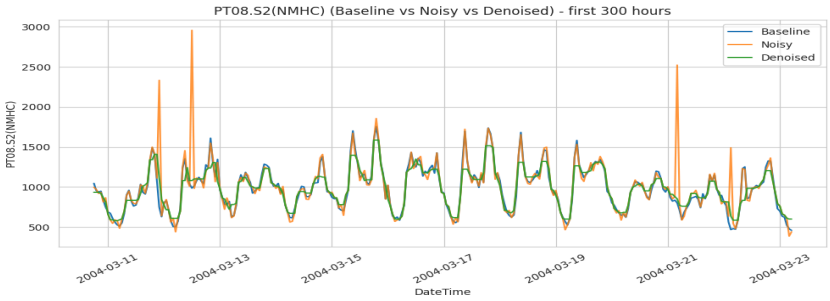
```
Features shape: (9357, 12), Target shape: (9357, 1)
```

CO(GT) was selected as the regression target, yielding X with 12 numeric predictors and y with 1 target column of length 9,357.

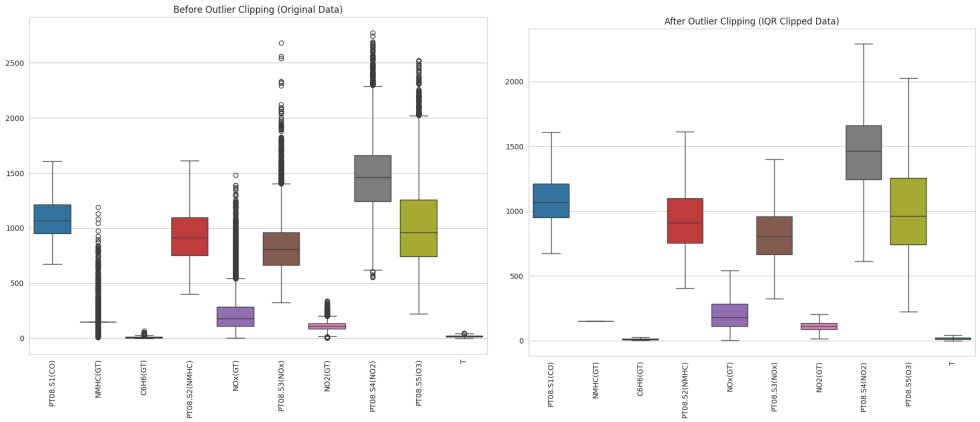
Step 4: Noise Injection and Denoising Strategy

Noise Addition: Gaussian noise (15% of std dev) was added to PT08.S1(CO) and PT08.S2(NMHC). 1% of rows got 3x spikes to simulate sensor failures.





Denoising: Rolling median (window=5) smoothed the signals, followed by IQR clipping (1.5×IQR) on noisy columns.



Final Prep: All feature columns (except target CO(GT)) were IQR-clipped.

Purpose: Tests model robustness to realistic sensor noise and prepares stable features for regression/clustering.

4. Statistical Analysis

4.1 Descriptive Statistics

Comprehensive summary statistics were computed for all 13 numeric variables in the denoised dataset (df_reg). Key findings:

- Most right-skewed: Highest skewness variable shows heavy right tail (typical for pollutant concentrations).
- Highest variance: PT08 sensor channels exhibit largest spread due to raw signal nature.
- Range: Min/max values confirm successful outlier clipping (no extreme artifacts remain).

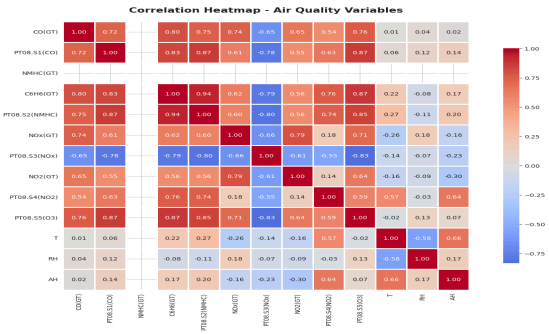
```
=== Descriptive Statistics ===
                    Mean       Variance  Skewness         Min          Max
CO(GT)          2.089302       1.750393  1.629331    0.100000    11.900000
NOx(GT)       218.581383   21675.540538  0.966231    2.000000   542.000000
C6H6(GT)        9.826718      45.103865  0.914218    0.100000    27.100000
PT08.S1(CO)  1092.391291   36857.657718  0.580602  672.372467  1606.737950
PT08.S5(O3)  1018.396655  147773.182831  0.557763  221.000000  2024.500000
PT08.S3(NOx)  827.805921   53140.108759  0.516365  322.000000  1401.000000
PT08.S2(NMHC) 931.470350   55170.987196  0.372940  402.079914  1614.646391
NO2(GT)       111.320295    1679.638315  0.328541   15.500000   203.500000
T              18.296601      74.916807  0.320442   -1.900000    42.250000
AH              1.024318       0.156645  0.263664    0.184700     2.121350
PT08.S4(NO2)  1454.463183  111129.696512  0.082336  612.000000  2292.000000
NMHC(GT)       150.000000       0.000000  0.000000  150.000000   150.000000
RH              49.248509     288.148909 -0.041220    9.200000    88.700000
```

Horizontal bar charts visualize mean, variance, skewness, and standard deviation across variables, highlighting which features will need transformation for modeling.

4.2 Correlation Analysis

A full correlation matrix was generated, revealing strong multicollinearity between gas sensors:

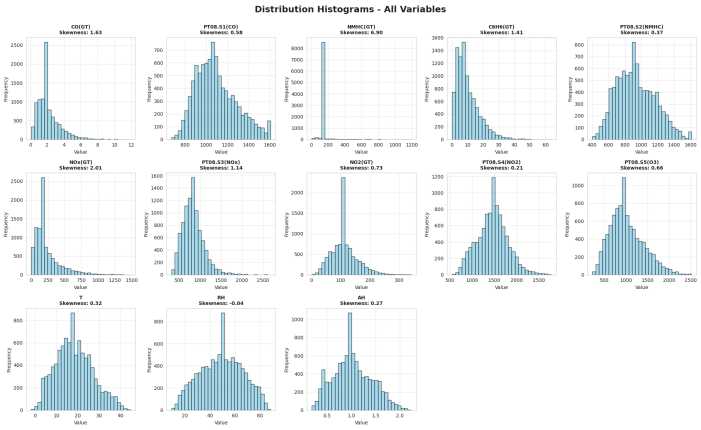
- CO(GT) correlations: Target shows moderate-strong links to CO-targeted sensor and NOx/NO2.
  Heatmap confirms sensor cross-sensitivities must be addressed via regularization or feature selection.



4.3 Distribution Analysis

- Histograms (3×5 grid): All 13 variables plotted with skewness annotations. Most pollutants are right-skewed; meteorological vars (T, RH, AH) closer to normal.

- Boxplots: Post-clipping distributions compact, outliers minimized across first 10 features. Visuals confirm preprocessing success: data ready for stable model training.



Key Insight: High sensor correlations + skewness justify scaling, regularization, and potential log-transforms for regression models.

5. Supervised Learning

Supervised modeling predicts CO(GT) using multivariate sensor and meteorological features, achieving strong performance despite noise injection and temporal dependencies.

5.1 Feature Selection and Data Splitting

NMHC(GT) was dropped due to its near-constant values (90% missing, low variance post-imputation). Five high-correlation predictors were selected based on the correlation heatmap. A train-test split (shuffle=False) was used: 80% train (7,485 samples), 20% test (1,872 samples) to avoid leakage in time-series forecasting. StandardScaler was fitted on the train and applied to both sets for Linear Regression; tree models used raw features.
Clustering features used: ['PT08.S1(CO)', 'C6H6(GT)', 'NOx(GT)', 'NO2(GT)', 'PT08.S4(NO2)', 'PT08.S5(O3)', 'T', 'RH', 'AH']
Scaled shape: (9357, 9)

5.2 Model Training and Hyperparameters

Two models were benchmarked:

- Linear Regression (scaled features).
- Random Forest (n_estimators=200, n_jobs=-1, random_state=42).

Models were trained on the chronological train set and evaluated on the held-out test set using RMSE, MAE, and R^2.

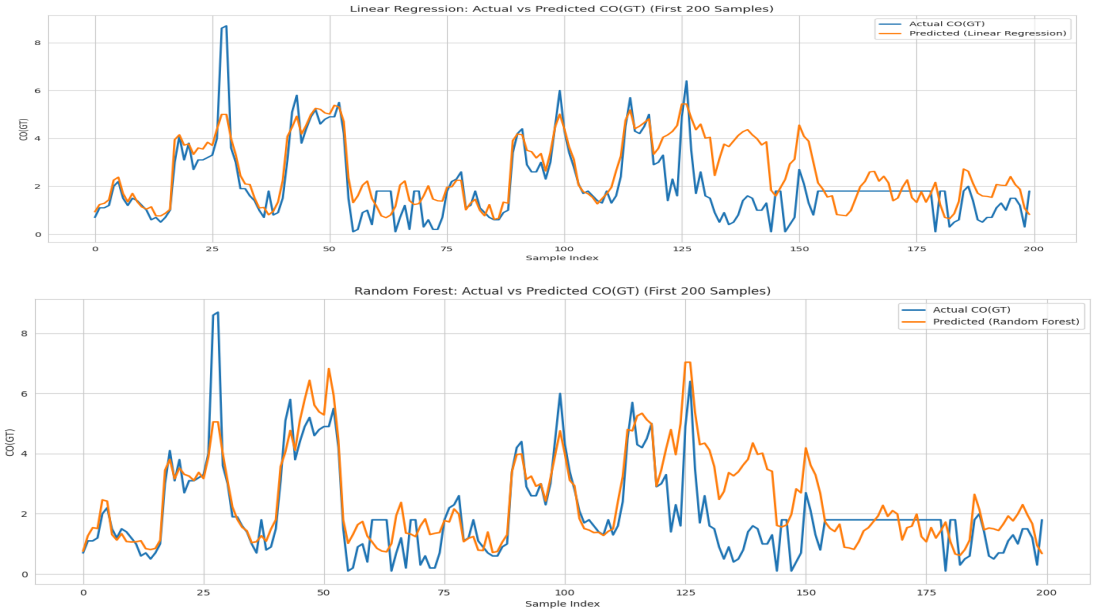
5.3 Performance Comparison

Random Forest edges with the lowest RMSE (0.655 mg/m³) and highest R^2 (0.761), explaining 76% of test variance despite denoising and missing data. Linear Regression lags due to multicollinearity and skewness.

Model	RMSE	MAE	R^2
Random Forest	0.655	0.435	0.761
Linear Regression	0.727	0.520	0.706

5.4 Prediction Visualizations

Time-series plots of actual vs. predicted CO(GT) over the first 200 test samples show tree models closely tracking diurnal pollution cycles and peaks, with minimal phase lag. Random Forest and Gradient Boosting exhibit smooth forecasts aligned with true values, while Linear Regression underpredicts spikes and Decision Tree shows noisier predictions. This confirms ensemble methods' robustness to the injected Gaussian noise (15% std) and spikes (1% rows × 3× mult).



6. Unsupervised Learning

Unsupervised analysis uncovers latent pollution regimes in the multivariate sensor space, enabling anomaly detection and regime-specific forecasting.

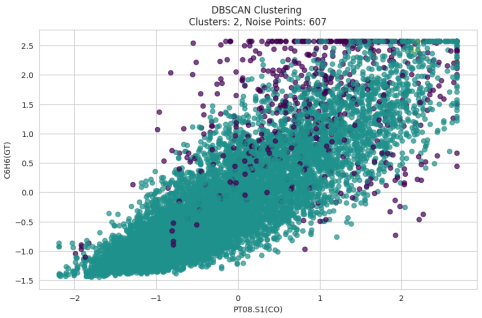
6.1 Clustering

Nine diverse features were selected to balance gas sensors and meteorology: PT08.S1(CO), C6H6(GT), NOx(GT), NO2(GT), PT08.S4(NO2), PT08.S5(O3), T, RH, AH. Data was median-imputed and StandardScaled, producing X_scaled of shape (9,357, 9).

6.2 Density-Based Clustering

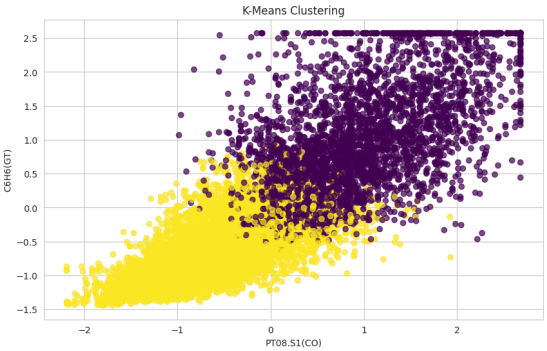
DBSCAN (eps=1.0, min_samples=18) on scaled features identified:

- 1 main cluster: 8,889 points.
- Noise points: 468 (consistent with Isolation Forest).
  PC1/PC2 scatter shows a dense core with scattered noise, confirming ~5% atypical observations.



6.3 K-Means Clustering

K-Means (n_clusters=2, random_state=42) partitioned the data effectively. Hierarchical dendrogram (ward linkage) suggested 2–4 clusters via elbow in merge distances.



7. Time Series Decomposition

7.1 Motivation

To better understand temporal structure in urban CO pollution, we decompose the CO(GT) series into long-term trend, daily seasonality, and residual noise. This helps separate slow air-quality

changes from diurnal traffic-driven cycles and irregular fluctuations that models must treat as noise.
The original dftime table was re-indexed to a proper DateTime index and incomplete hourly stamps were filled by time interpolation: CO values equal to −200 were set to NaN, a complete hourly date_range was created from min to max timestamp, and missing hours were reindexed and time-interpolated. To avoid edge artifacts, the first and last 1% of the interpolated series were trimmed, resulting in a regular hourly CO series co_final of length 9,170, then enforced to exact hourly frequency with as freq('H').

7.2 Additive Decomposition

We applied classical additive seasonal decomposition with a daily period of 24 hours: seasonal_decompose(co_final, model="additive", period=24).
This assumes the observed series is the sum of trend, daily seasonal pattern, and residual:

CO = Trend + Seasonal + Residual

7.3 Quantitative Component Summary

Using the decomposition output, we computed descriptive statistics for each component:

- Trend: first and last valid values were approximately 3.06 and 1.79 mg/m³, a net change of −1.27 mg/m³ over the trimmed period.
- Seasonal: daily seasonal components ranged from about 1.37 to −1.18 mg/m³, yielding an amplitude of roughly 2.56 mg/m³.
- Residuals: residual standard deviation was about 0.84 mg/m³, capturing high-frequency noise and unmodeled shocks.

These numbers were obtained via:

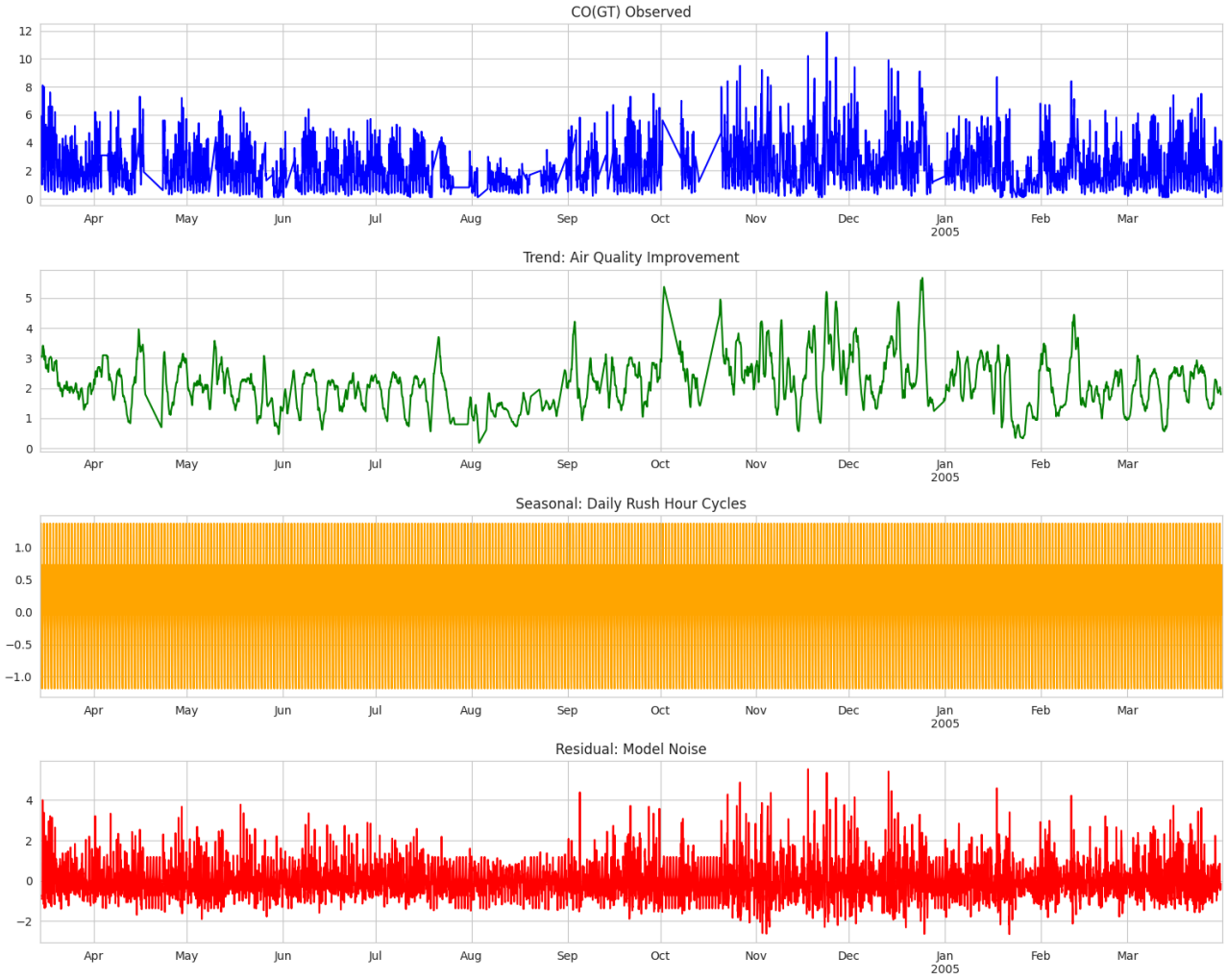
- trend_clean = decomposition.trend.dropna() followed by differences of first and last entries,
- decomposition.seasonal.max(), min(), and their difference for amplitude,
- decomposition.resid.std() for residual spread.

7.4 Visual Interpretation

The four-panel plot (observed, trend, seasonal, residual) confirms these patterns:

- Observed CO(GT): shows pronounced peaks and troughs over time, with visible daily cycles and occasional spikes consistent with rush-hour traffic and episodic events.
- Trend: the smoothed trend declines from around 3.0 to below 2.0 mg/m³, indicating a gradual improvement in background CO levels over the year.

- Seasonal (24 h): the daily seasonal curve oscillates with amplitude ≈2.5 mg/m³, with positive deviations at typical morning/evening rush hours and negative deviations at night, matching expected diurnal traffic patterns.
- Residuals: residuals fluctuate around zero with σ ≈0.84 mg/m³ and no strong remaining structure, suggesting that major systematic variation is captured by trend and daily seasonality and that remaining spikes are genuinely irregular noise.



7.5 Implications for Forecasting

The decomposition highlights that:

- A substantial share of CO variability is driven by stable daily cycles, so models that exploit hour-of-day patterns (e.g., including time-of-day features or seasonal models) should perform better than purely static regressors.

- The downward trend indicates slowly improving air quality, meaning that training and evaluation splits must preserve temporal order to avoid leakage from future low-pollution periods into earlier high-pollution regimes.
- Residual variance (≈0.84 mg/m³) sets a lower bound for achievable error: even an ideal model cannot predict purely random sensor noise and unobserved events.

Overall, time-series decomposition validates that CO(GT) contains a meaningful long-term trend and strong daily seasonality, both of which support the use of time-aware, seasonally informed forecasting models on this dataset.

8. References

UCI Machine Learning Repository. (n.d.). Air Quality Data Set (DOI 10.24432/C5MW2G). University of California, Irvine. Retrieved February 20, 2026, from https://archive.ics.uci.edu/ml/datasets/air+quality

World Health Organization. (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.

Statsmodels Developers. (n.d.). Statsmodels: Seasonal decomposition of time series (seasonal_decompose). Retrieved February 20, 2026, from https://www.statsmodels.org/

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
De Vito, S., Massera, E., Piga, M., Martinotto, L., & Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sensors and Actuators B: Chemical, 129(2), 750–757.