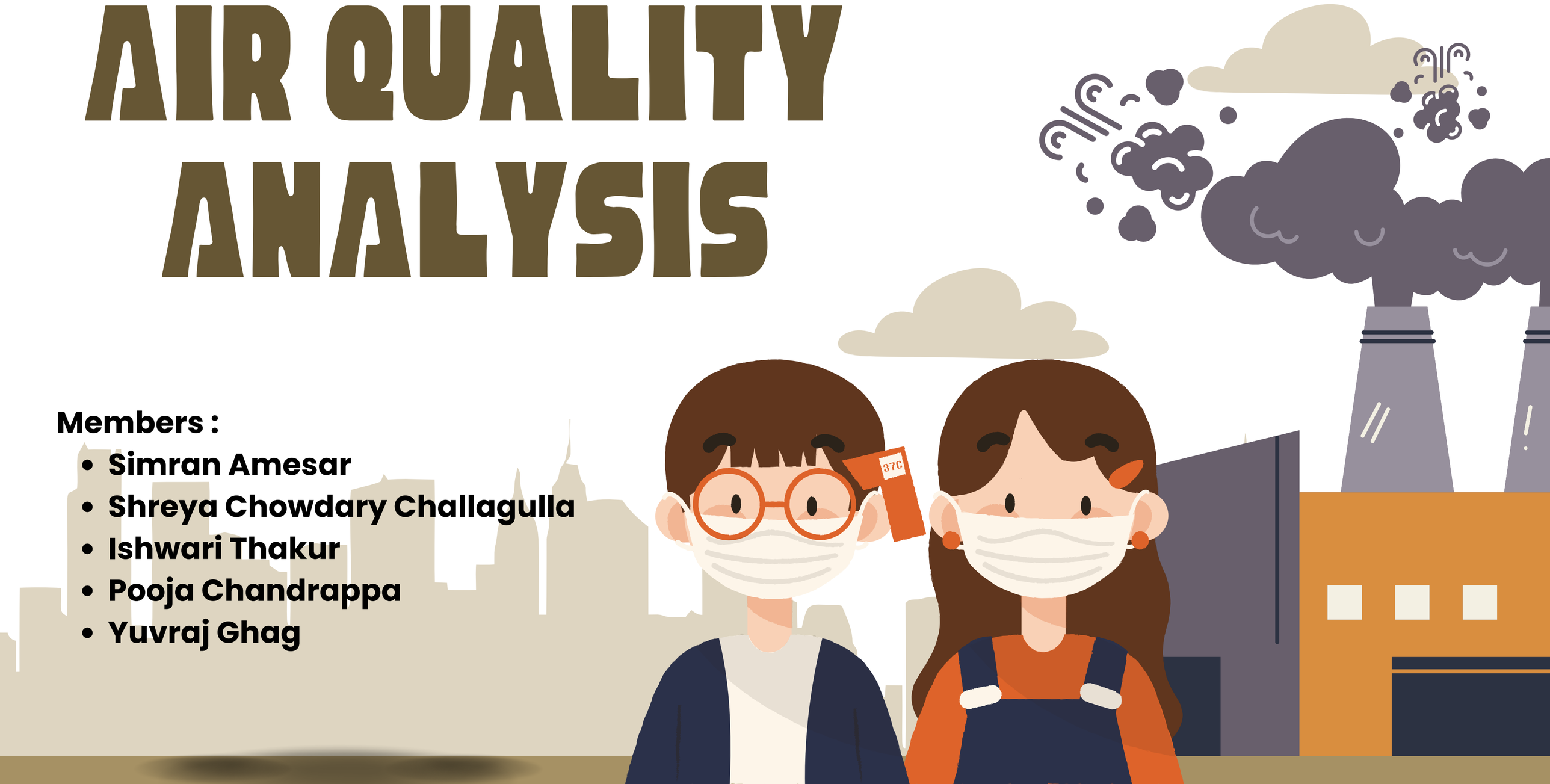
AIR QUALITY ANALYSIS



Members :

- **Simran Amesar**
- **Shreya Chowdary Challagulla**
- **Ishwari Thakur**
- **Pooja Chandrappa**
- **Yuvraj Ghag**

DATASET

The UCI Air Quality Dataset contains 9,357 hourly observations collected over 13 months (Mar 2004 - April 2005) from a single urban monitoring station.

- Target: CO(GT) - true CO concentration
- Features: 13 variables including CO sensor (PT08.S1(CO)), benzene, NOx, NO2 sensors, temperature, humidity
- Challenges: 18% missing values, high multicollinearity between gas sensors, temporal dependencies

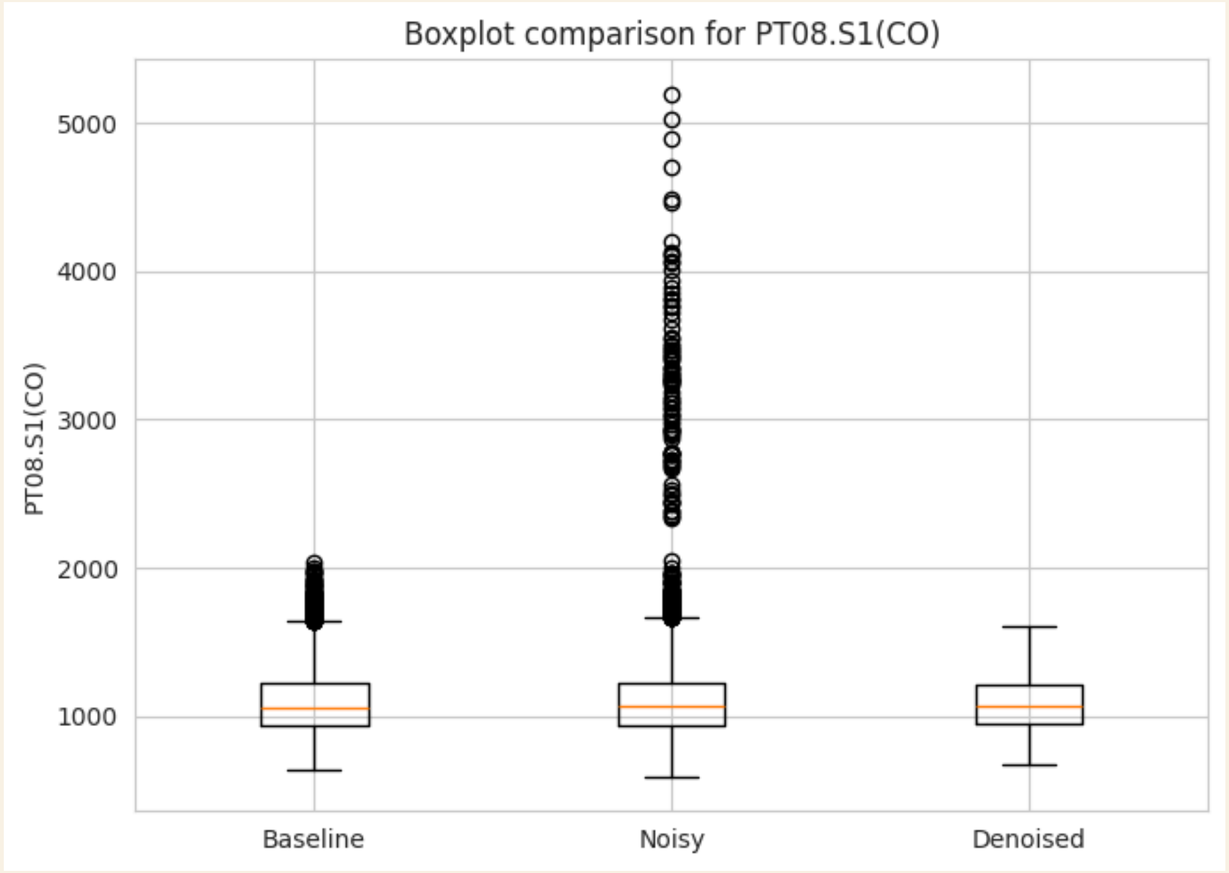
PREPROCESSING

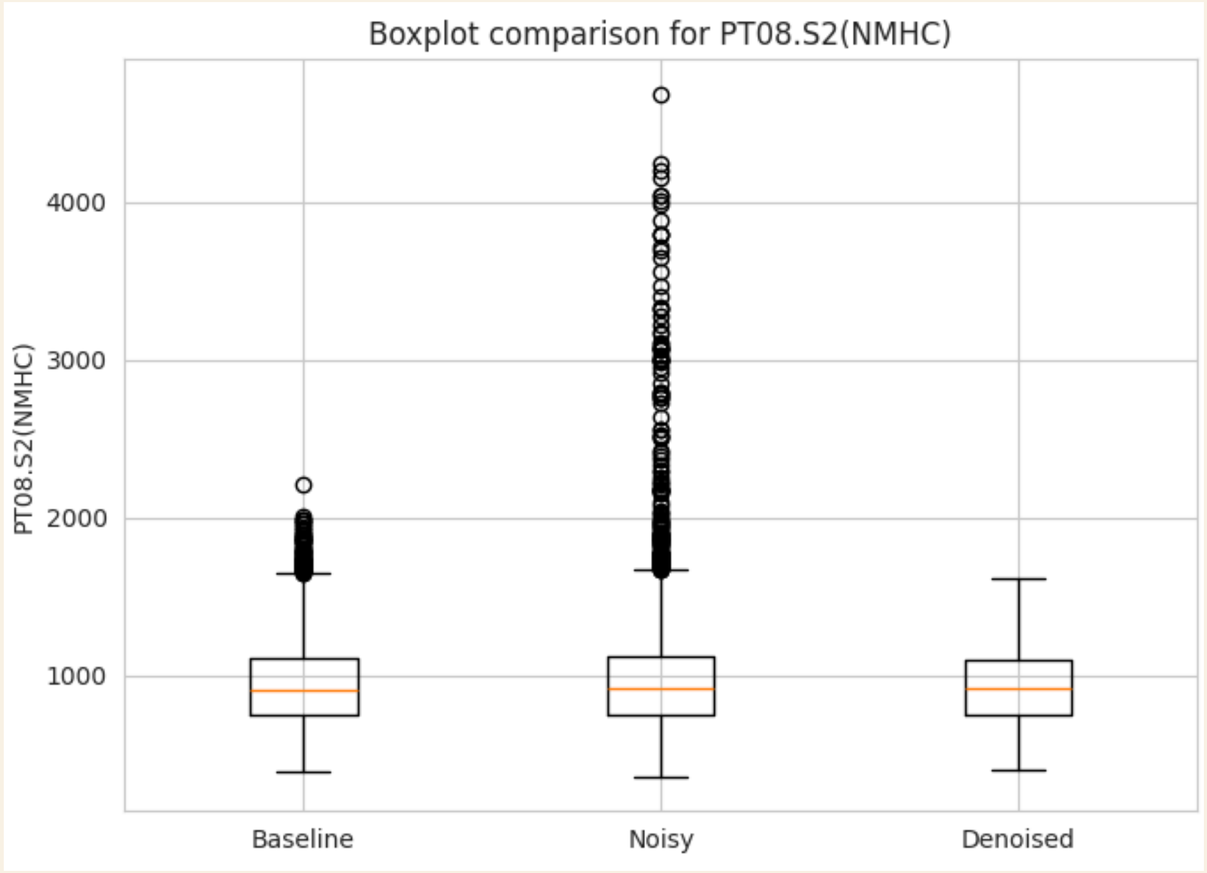
- Missing Values: Values coded as -200 represent sensor failures
- Total missing: 16,701 instances across 13 numeric features (18% of data)
- Treatment: Median imputation (robust to skewness)
- Temporal Structure: Hourly measurements with daily/seasonal patterns
- Date/Time columns combined into DateTime index
- Target-Feature Separation : 'CO(GT)'

NOISE INJECTION

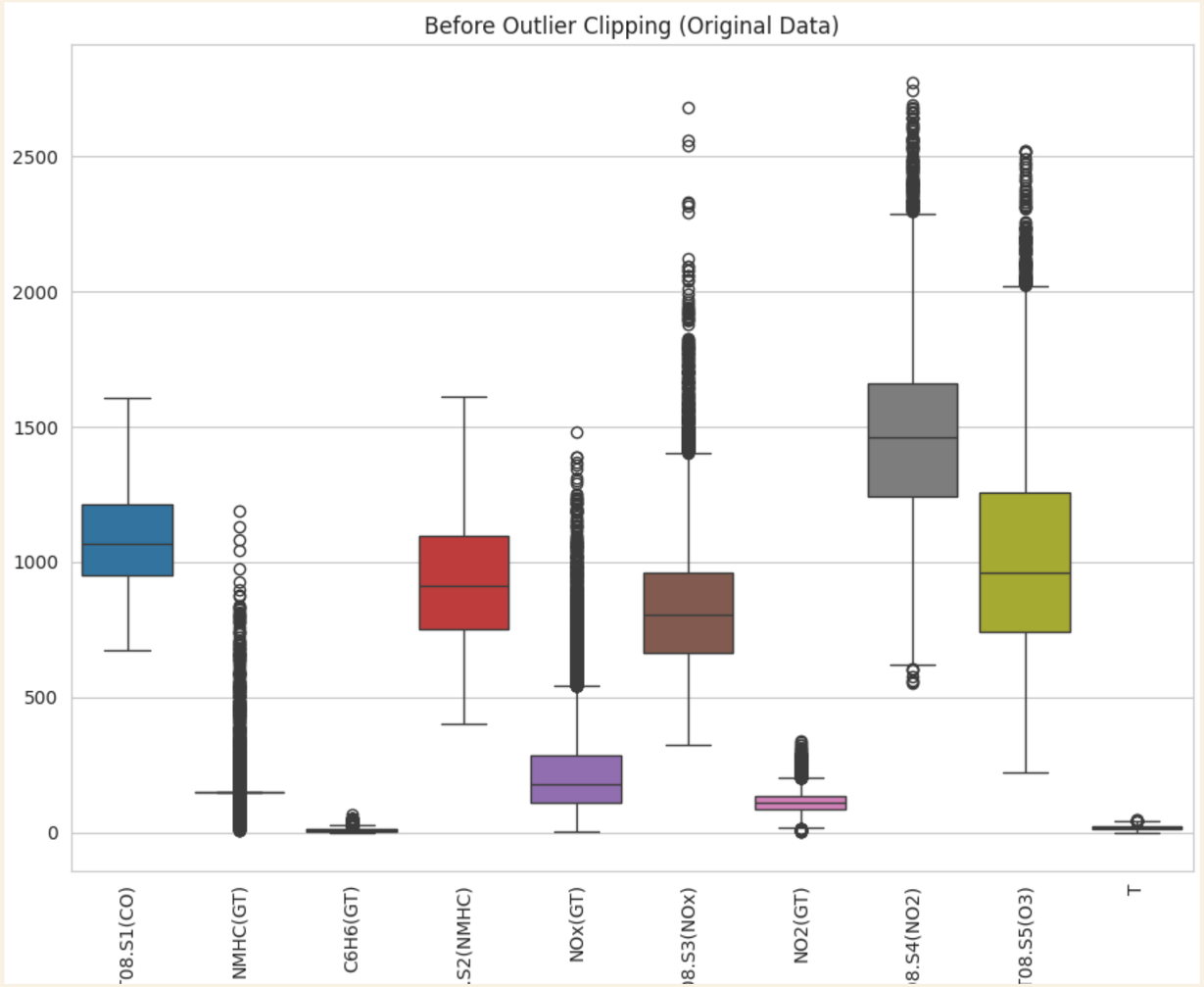
- Add artificial noise to 10% of samples to test data cleaning methods
- Randomly inject outlier values (0-1 range) into selected features : "PT08.S1(CO)" and "PT08.S2(NMHC)"
- Tests outlier detection performance and model robustness against noisy data

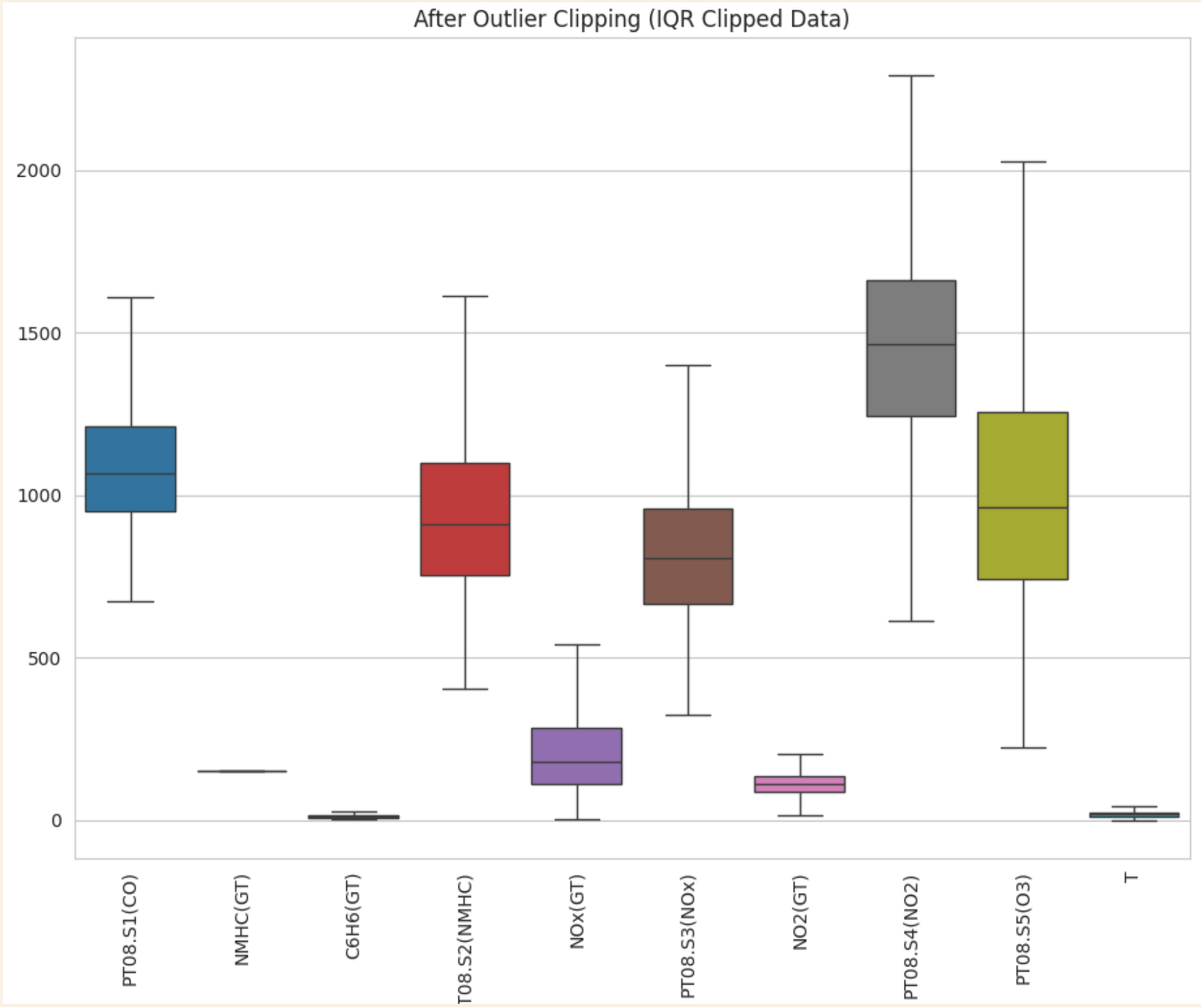
NOISE INJECTION



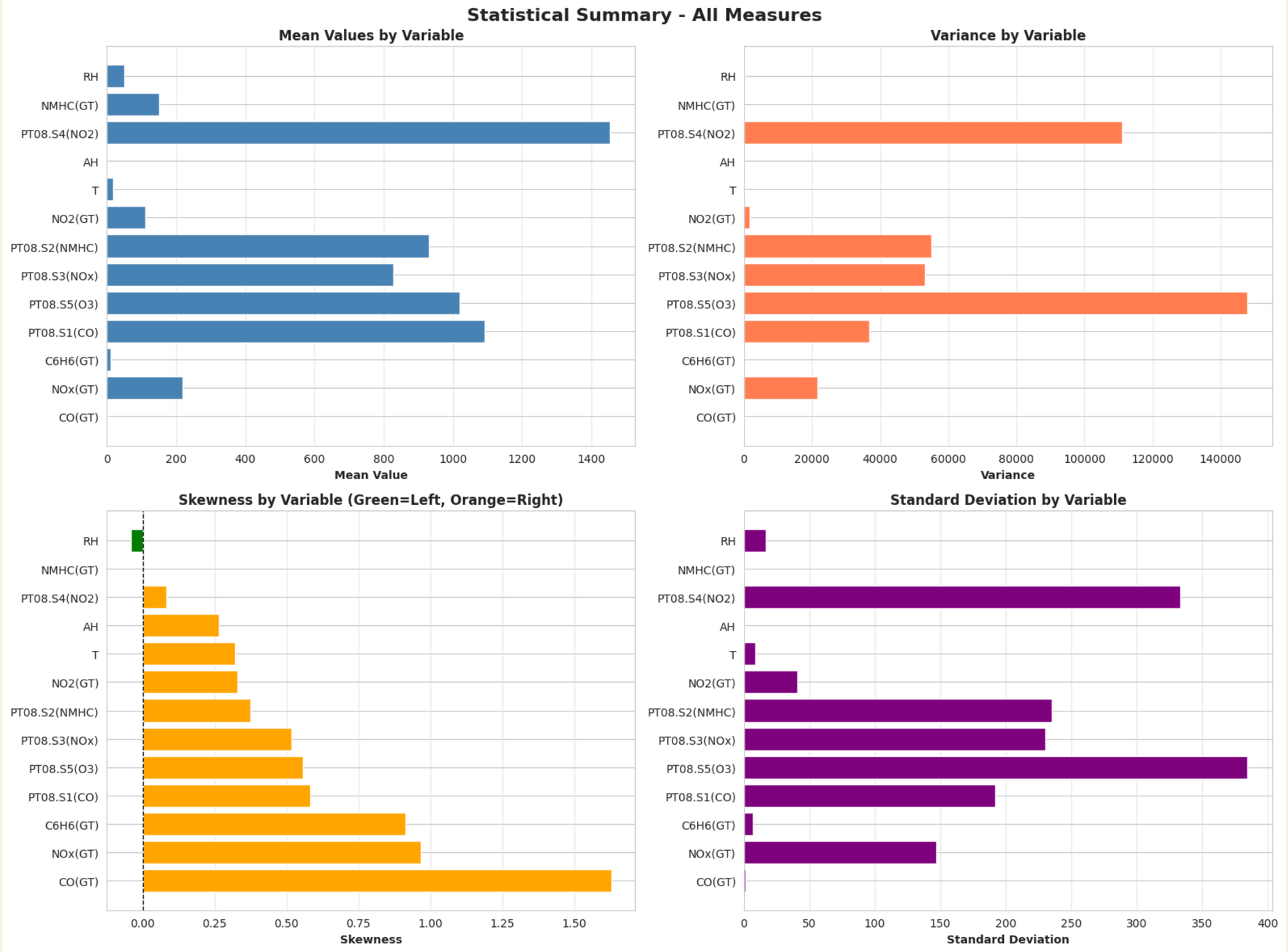


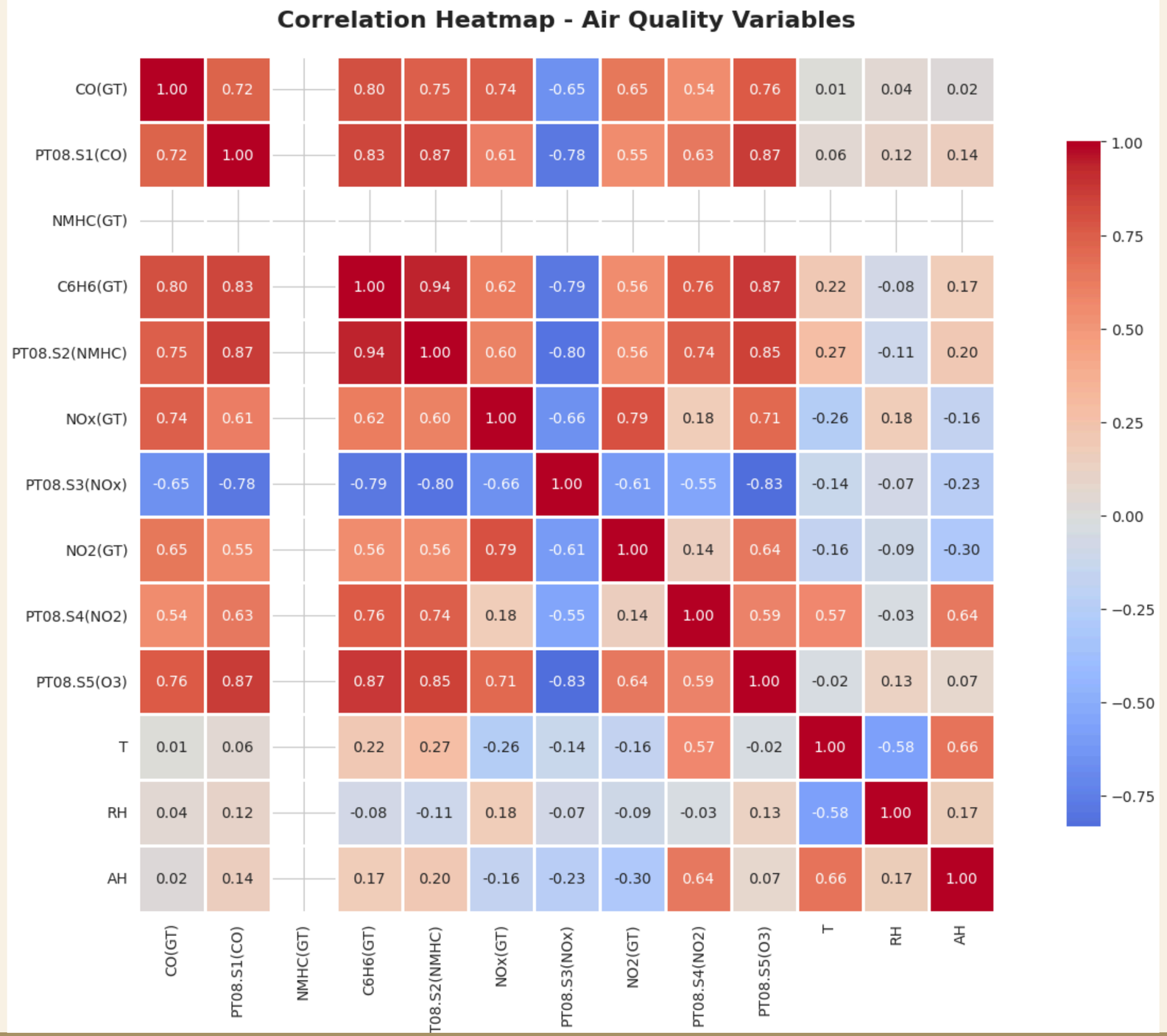
OUTLIER CLIPPING





STATISTICAL ANALYSIS


Statistical Summary - All Measures
Mean Values by Variable
RH
NMHC(GT)
PT08.S4(NO2)
AH
T
NO2(GT)
PT08.S2(NMHC)
PT08.S3(NOx)
PT08.S5(O3)
PT08.S1(CO)
C6H6(GT)
NOx(GT)
CO(GT)
0
200
400
600
800
1000
1200
1400
Mean Value
Variance by Variable
0
20000
40000
60000
80000
100000
120000
140000
Variance
Skewness by Variable (Green=Left, Orange=Right)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
Skewness
Standard Deviation by Variable
0
50
100
150
200
250
300
350
400
Standard Deviation


Correlation Heatmap - Air Quality Variables
CO(GT) 1.00 0.72 0.80 0.75 0.74 -0.65 0.65 0.54 0.76 0.01 0.04 0.02
PT08.S1(CO) 0.72 1.00 0.83 0.87 0.61 -0.78 0.55 0.63 0.87 0.06 0.12 0.14
NMHC(GT)
C6H6(GT) 0.80 0.83 1.00 0.94 0.62 -0.79 0.56 0.76 0.87 0.22 -0.08 0.17
PT08.S2(NMHC) 0.75 0.87 0.94 1.00 0.60 -0.80 0.56 0.74 0.85 0.27 -0.11 0.20
NOx(GT) 0.74 0.61 0.62 0.60 1.00 -0.66 0.79 0.18 0.71 -0.26 0.18 -0.16
PT08.S3(NOx) -0.65 -0.78 -0.79 -0.80 -0.66 1.00 -0.61 -0.55 -0.83 -0.14 -0.07 -0.23
NO2(GT) 0.65 0.55 0.56 0.56 0.79 -0.61 1.00 0.14 0.64 -0.16 -0.09 -0.30
PT08.S4(NO2) 0.54 0.63 0.76 0.74 0.18 -0.55 0.14 1.00 0.59 0.57 -0.03 0.64
PT08.S5(O3) 0.76 0.87 0.87 0.85 0.71 -0.83 0.64 0.59 1.00 -0.02 0.13 0.07
T 0.01 0.06 0.22 0.27 -0.26 -0.14 -0.16 0.57 -0.02 1.00 -0.58 0.66
RH 0.04 0.12 -0.08 -0.11 0.18 -0.07 -0.09 -0.03 0.13 -0.58 1.00 0.17
AH 0.02 0.14 0.17 0.20 -0.16 -0.23 -0.30 0.64 0.07 0.66 0.17 1.00
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

SELECTED FEATURES

PT08.S1(CO)

C6H6(GT)

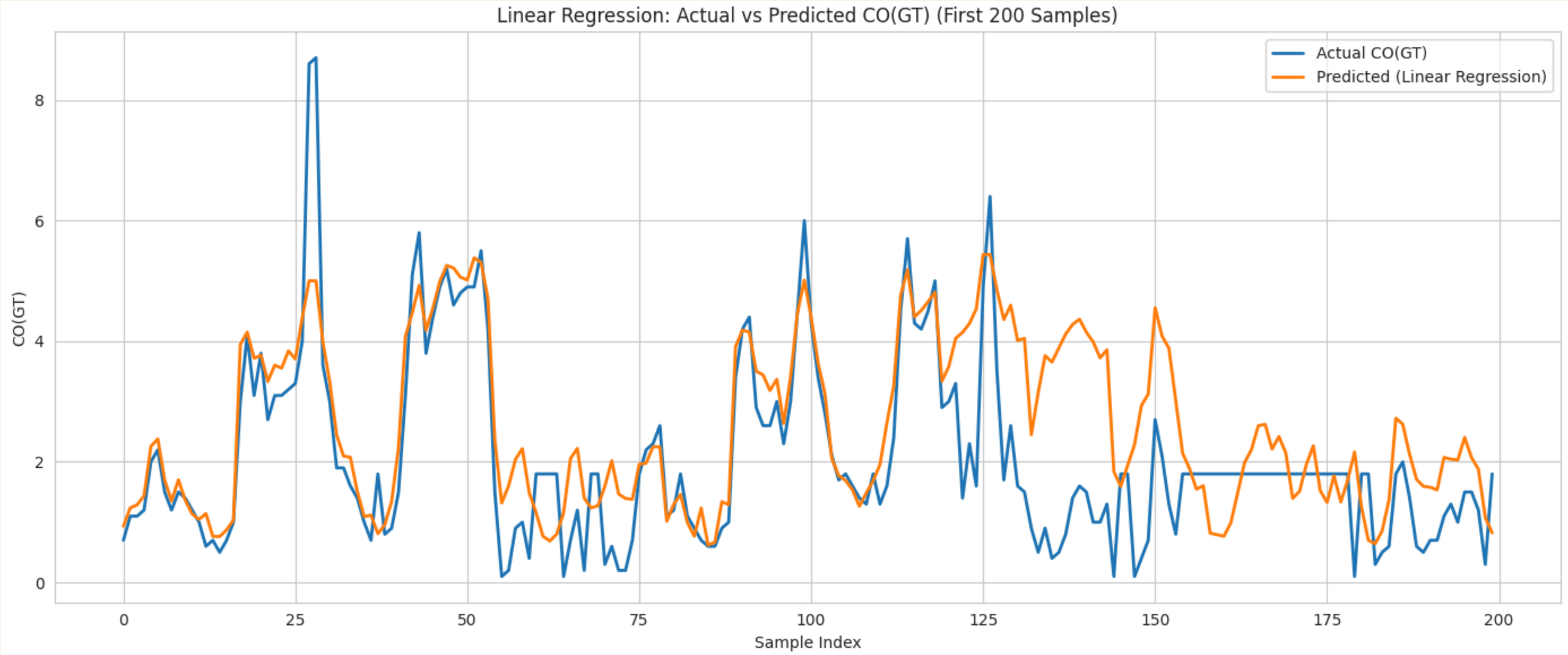
PT08.S2(NMHC)

NOx(GT)

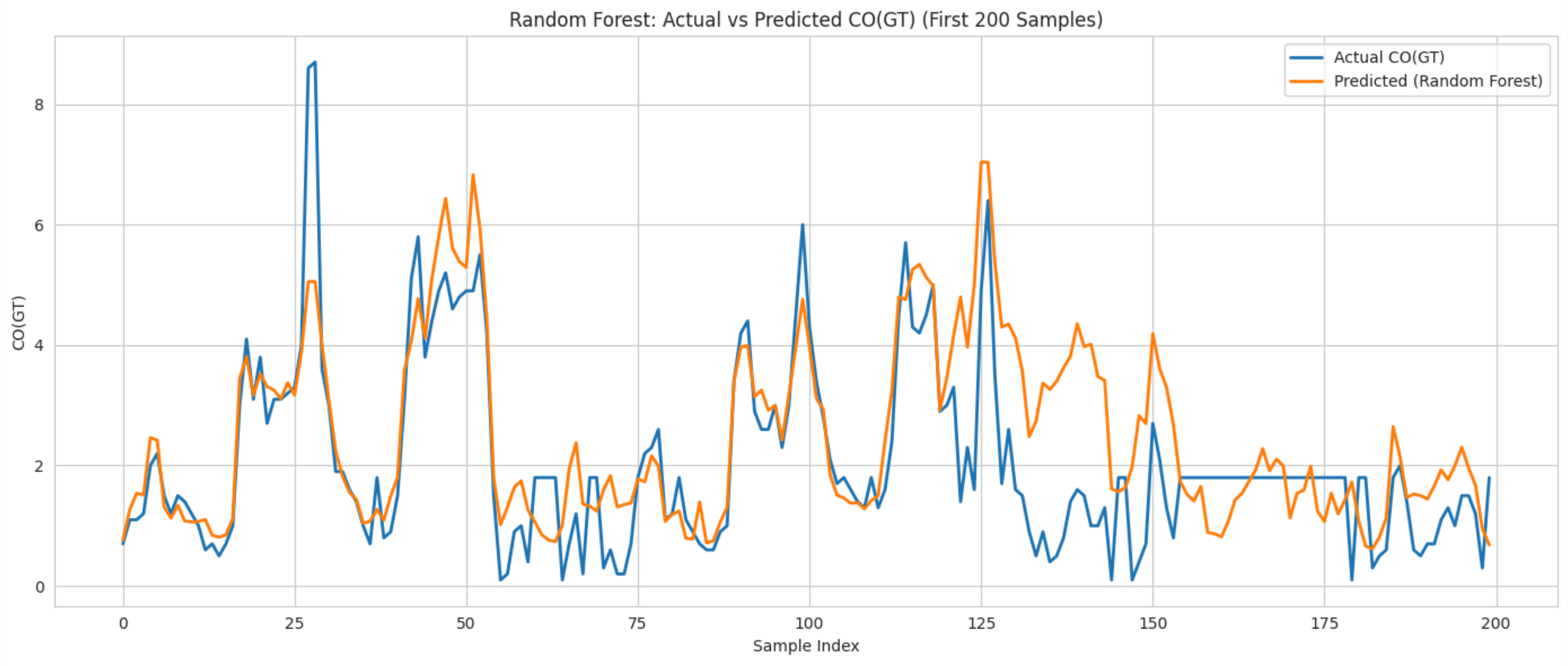
PT08.S5(O3)

SUPERVISED LEARNING

LINEAR REGRESSION



RANDOM FOREST

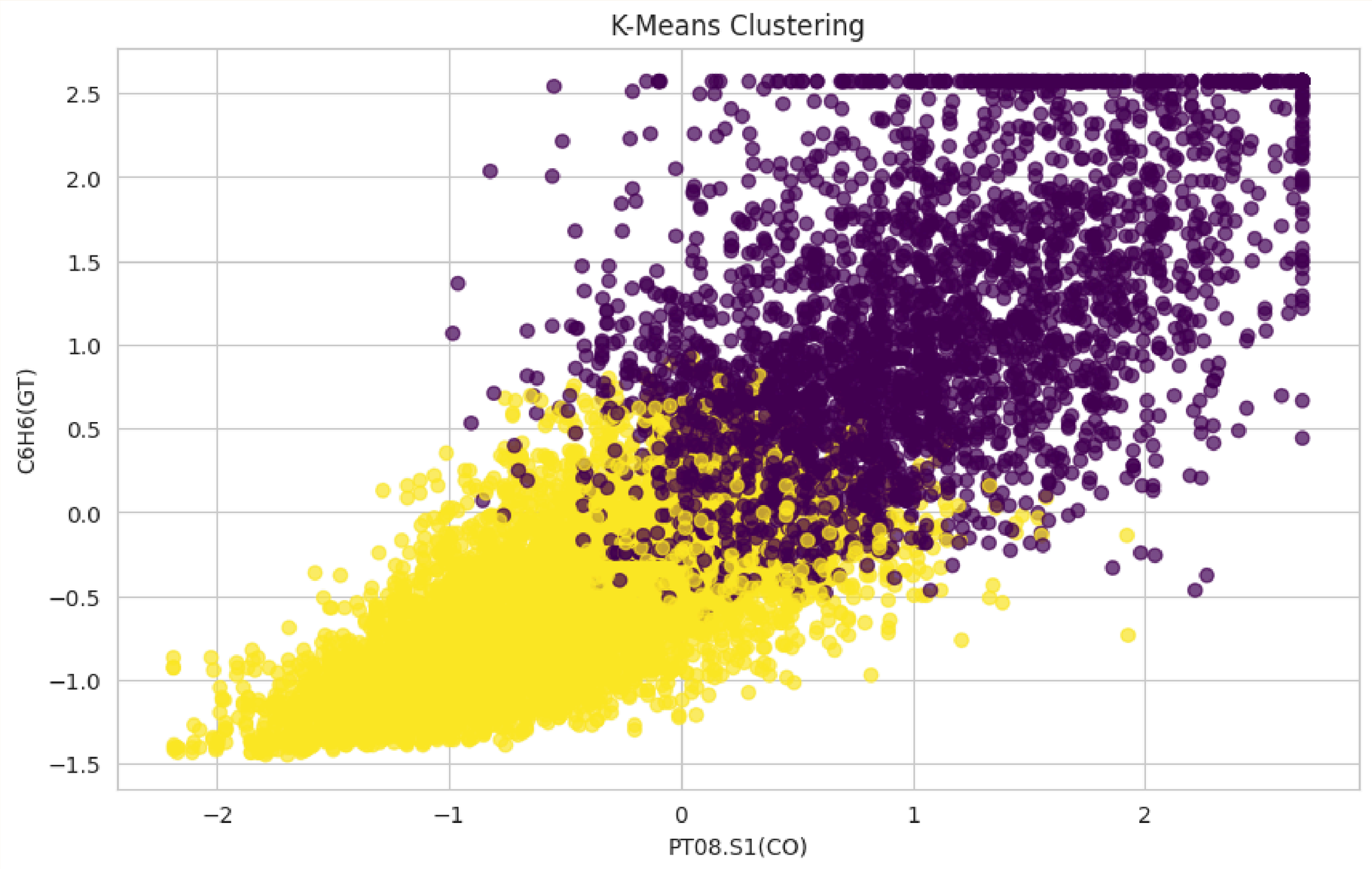


Model Comparison

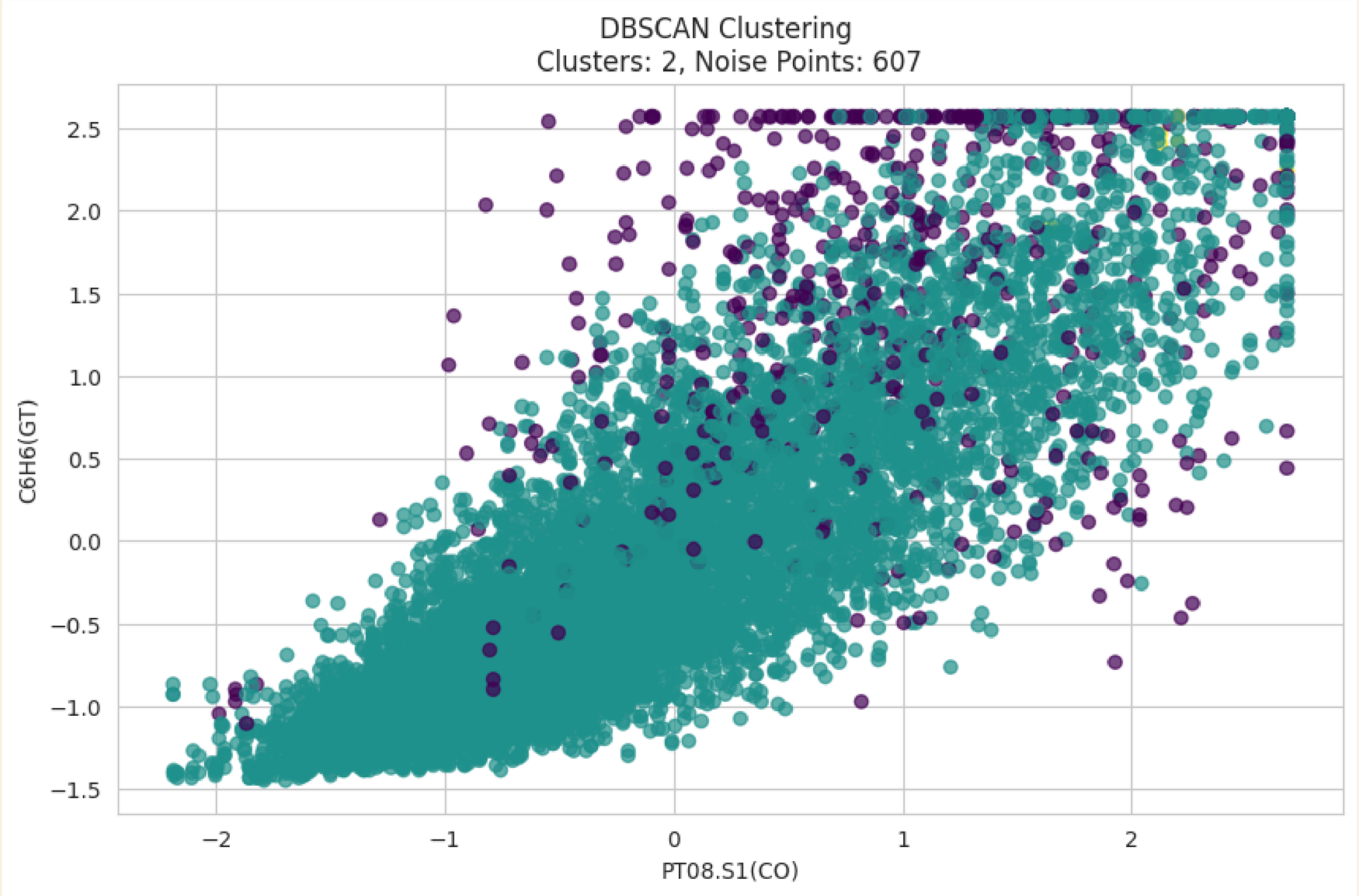
Model	RMSE	MAE	R2
Random Forest	0.655281	0.435162	0.760967
Linear Regression	0.726689	0.520099	0.706032

UNSUPERVISED LEARNING

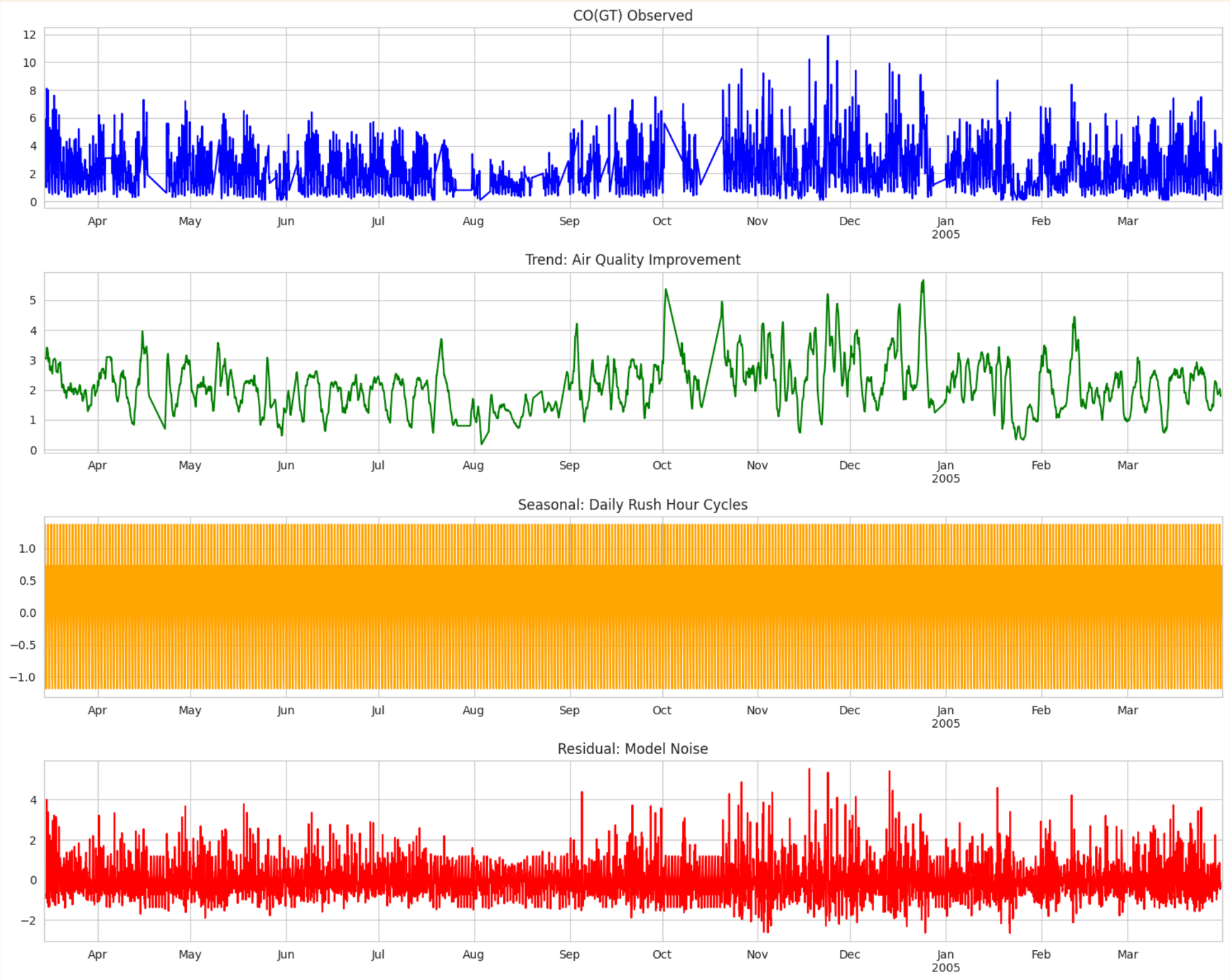
K-MEANS CLUSTERING

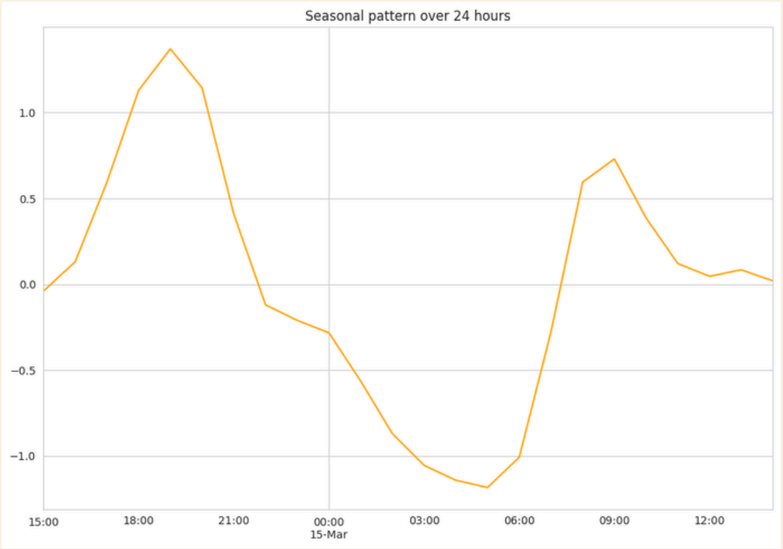


DBSCAN CLUSTERING



TIME-SERIES DECOMPOSITION


CO(GT) Observed
Trend: Air Quality Improvement
Seasonal: Daily Rush Hour Cycles
Residual: Model Noise
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
2005
Feb
Mar


Seasonal pattern over 24 hours
15:00
18:00
21:00
00:00
15-Mar
03:00
06:00
09:00
12:00

THANK YOU