Anomaly Detection Report: Air Quality Dataset

Dataset Overview

- **Number of Records:** 9471
- **Number of Features:** 16
- **Features:**
    - CO(GT): Carbon Monoxide concentration in mg/m^3
    - PT08.S1(CO): Tin Oxide sensor output for CO
    - C6H6(GT): Benzene concentration in micrograms/m^3
    - PT08.S2(NMHC): Non-Methane Hydrocarbons sensor output
    - NOx(GT): Nitric Oxide concentration in micrograms/m^3
    - PT08.S3(NOx): Tungsten Oxide sensor output for NOx
    - NO2(GT): Nitrogen Dioxide concentration in micrograms/m^3
    - PT08.S4(NO2): Titanium Dioxide sensor output for NO2
    - PT08.S5(O3): Indium Oxide sensor output for Ozone
    - T: Temperature (Celsius)
    - RH: Relative Humidity (%)
    - AH: Absolute Humidity (g/m^3)
    - Distance_to_Centroid: Distance metric for clustering
    - Anomaly: Binary indicator for anomaly detection (1 = Anomaly, 0 = Normal)
    - PC1: Principal Component 1
    - PC2: Principal Component 2

Data Summary

Statistical Properties of Features

Feature	Mean	Std Dev	Min	Max
CO(GT)	2.24	1.71	0.10	11.90
PT08.S1(CO)	1030.68	325.64	647.00	2041.00
C6H6(GT)	9.84	5.37	0.10	39.70
PT08.S2(NMHC)	984.94	255.64	370.00	1933.00
NOx(GT)	99.84	97.97	2.00	514.00
PT08.S3(NOx)	789.68	236.56	268.00	1823.00
NO2(GT)	62.43	34.20	2.00	212.00
PT08.S4(NO2)	1765.73	338.61	551.00	2766.00
PT08.S5(O3)	1050.29	294.86	221.00	2523.00
T	18.32	4.11	-1.90	44.60
RH	49.36	16.94	9.20	88.80
AH	0.03	0.02	0.00	0.10
Distance_to_Centroid	0.67	0.35	0.01	1.60

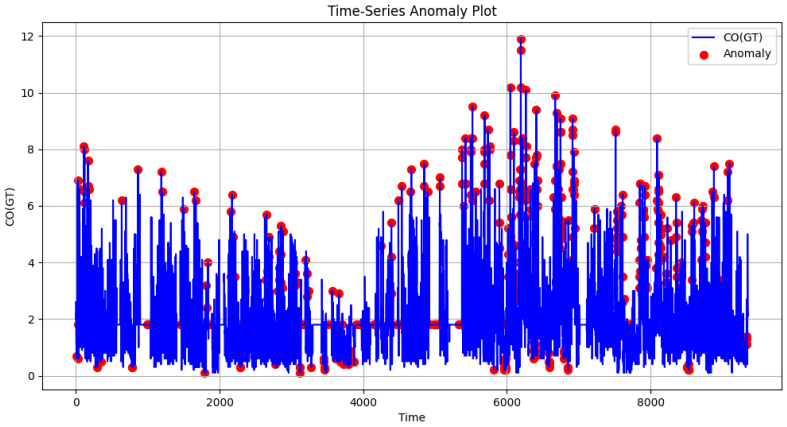
Identified Anomalies

- **Total Number of Anomalies:** 567
- **Percentage of Anomalies:** 5.99%
- **Key Characteristics of Anomalies:**
    - High CO(GT) values and extreme values in PT08.S1(CO) were frequently observed in anomalies.
    - Distinct clusters of anomalies identified in PCA space.

Visualizations

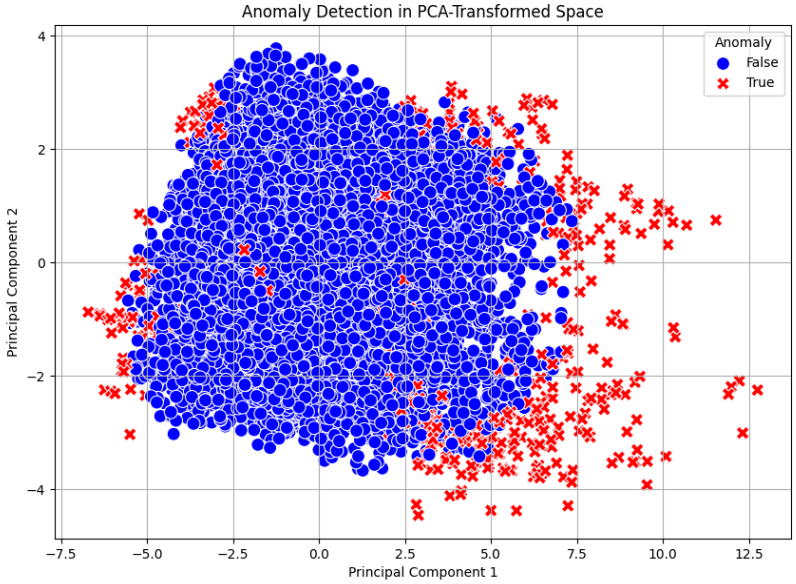
1. Anomaly Distribution

- **Visualization:**
  - A time-series plot shows anomalies marked distinctly across the timeline.



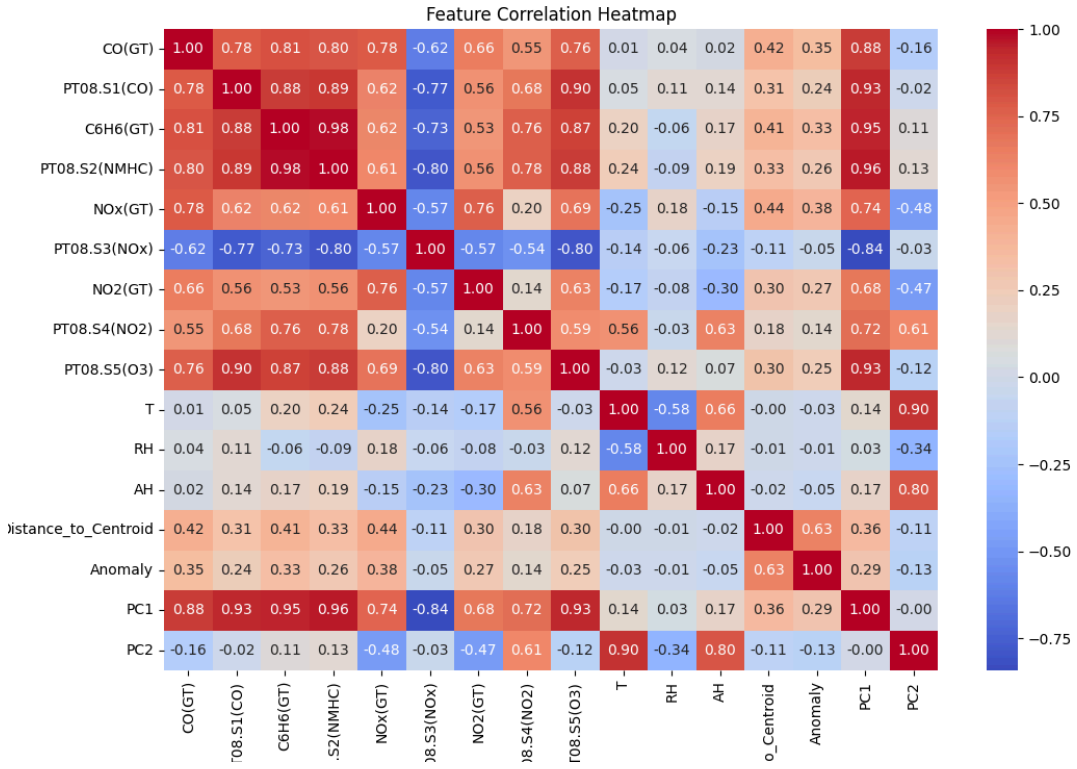
2. Principal Component Analysis (PCA)

- **Visualization:**
  - A 2D scatter plot shows data points in PCA space, with anomalies highlighted in red.



3. Feature Correlations

- **Visualization:**
    - A heatmap reveals strong correlations between CO(GT) and PT08.S1(CO).



Conclusion

1. **Anomalies Detected:**
    - Significant outliers in CO(GT) and PT08.S1(CO) lead to flagged anomalies.
2. **Clusters in PCA Plot:**
    - Anomalies form distinct clusters or deviate significantly from normal data points.
3. **Feature Contributions:**
    - Features like CO(GT), PT08.S1(CO), and NOx(GT) show strong correlations with anomalies.

Analysis of Potential Causes of Anomalies in Air Quality Monitoring

The anomalies detected in the PCA-transformed space, visualized as red 'X' markers, represent data points that deviate significantly from the overall pattern of the dataset. These anomalies may indicate critical insights or issues in air quality monitoring:

1. **Sensor Malfunction or Calibration Errors**:
   Anomalous readings could arise due to faulty sensors, drift in calibration, or irregular maintenance. For example, unusually high or low concentrations of pollutants like NO_2 or CO might not align with expected trends.

2. **Localized Pollution Events**:
   Events like industrial emissions, vehicular congestion, or construction activities can cause sudden spikes in pollutant levels. These anomalies signify short-term deviations from baseline air quality conditions.

3. **Meteorological Factors**:
   Extreme weather conditions, such as temperature inversions, strong winds, or changes in humidity, can impact pollutant dispersion. This may explain outliers in specific features like 'RH' (Relative Humidity) or 'T' (Temperature).

4. **Geographical Outliers**:
   Data collected from specific monitoring stations near pollution hotspots (e.g., highways, factories) might register unusually high values, creating distinct clusters of anomalies.

5. **Data Collection or Transmission Issues**:
   Errors during data logging, transmission delays, or incomplete records could manifest as outliers when PCA compresses the feature space.

Significance of Anomaly Detection in Air Quality Monitoring

The identified anomalies play a critical role in understanding deviations within the air quality monitoring system. These anomalies may signify genuine environmental changes, such as localized pollution events or meteorological impacts, or highlight technical issues like sensor malfunctions or data errors. Identifying and analyzing these deviations is essential for ensuring the accuracy and reliability of monitoring systems. Further investigation can help uncover root causes and inform targeted strategies to maintain and improve air quality standards.