efrei

PARIS PANTHÉON-ASSAS UNIVERSITÉ

Data Analysis Project

Group Members:
- Amira BOUDAOUD
- Ghofrane BEN RHAIEM

Supervisor:
MOHSEN Zahraa

Date of Submission:
December 16, 2024

Academic year: 2024/2025

Abstract

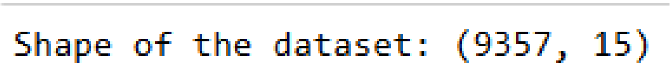
This project applies Principal Component Analysis (PCA) to analyze air quality data collected in an Italian city over one year. The dataset contains hourly-averaged pollutant concentrations, sensor responses, and environmental measurements. PCA was chosen to reduce dimensionality, identify key patterns, and simplify complex relationships in the dataset while retaining critical information. The study aims to visualize principal components, interpret the results, and understand the contributions of individual features to air quality trends.

Contents

1 Dataset Description

The dataset was sourced from the UCI Machine Learning Repository and contains measurements of air quality recorded from March 2004 to February 2005. The data were collected at road level in a highly polluted area within an Italian city using a multisensor air quality monitoring device. Key details of the dataset are summarized below:

1.1 Dataset Overview



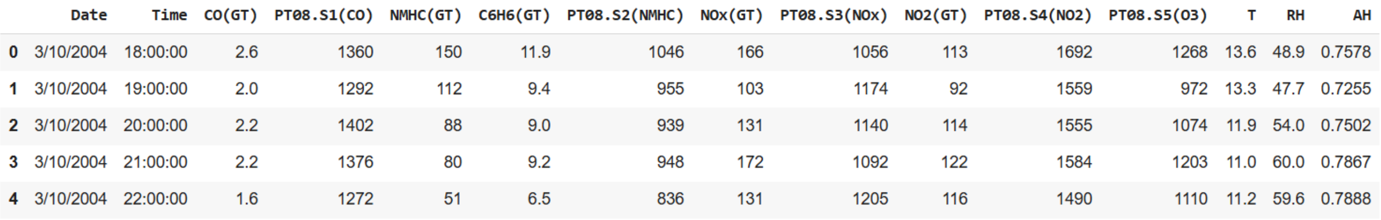
```
Shape of the dataset: (9357, 15)
```

The output ”**Shape of the dataset: (9357, 15)**” provides the following information about the dataset:

- ***9357 rows:***
  This indicates the dataset contains 9357 observations. Each row represents a single record, most likely an hourly measurement of air quality data over a period of time.

- ***15 columns:***
  The dataset has 15 features (variables). These features include environmental parameters like CO (carbon monoxide), NOx, NO2, sensor responses, temperature, relative humidity, and absolute humidity, as described previously.

- Preview of the first 5 rows of the dataset, showing air quality measurements, sensor responses, and environmental conditions.



	Date	Time	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
0	3/10/2004	18:00:00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
1	3/10/2004	19:00:00	2.0	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
2	3/10/2004	20:00:00	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0	0.7502
3	3/10/2004	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0	0.7867
4	3/10/2004	22:00:00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888

1.2 Features

The dataset includes the following features:

Variable Name	Role	Type	Description	Units
Date	Feature	Date	Recording date (DD/MM/YYYY)	-
Time	Feature	Categorical	Recording time (HH.MM.SS)	-
CO(GT)	Feature	Integer	True hourly-averaged concentration of Carbon Monoxide (CO)	mg/m^3
PT08.S1(CO)	Feature	Categorical	Sensor response nominally targeting CO	-
NMHC(GT)	Feature	Integer	True hourly-averaged concentration of Non-Methane Hydrocarbons (NMHC)	µg/m^3
C6H6(GT)	Feature	Continuous	True hourly-averaged concentration of Benzene (C6H6)	µg/m^3
PT08.S2(NMHC)	Feature	Categorical	Sensor response nominally targeting NMHC	-
NOx(GT)	Feature	Integer	True hourly-averaged concentration of Nitrogen Oxides (NOx)	ppb
PT08.S3(NOx)	Feature	Categorical	Sensor response nominally targeting NOx	-
NO2(GT)	Feature	Integer	True hourly-averaged concentration of Nitrogen Dioxide (NO2)	µg/m^3
PT08.S4(NO2)	Feature	Categorical	Sensor response nominally targeting NO2	-
PT08.S5(O3)	Feature	Categorical	Sensor response nominally targeting Ozone (O3)	-
T	Feature	Continuous	Temperature recorded in degrees Celsius	°C
RH	Feature	Continuous	Relative Humidity recorded as a percentage	%
AH	Feature	Continuous	Absolute Humidity (moisture content in the air)	-

Table 1: Description of features in the air quality dataset.

1.3 Additional information about the features

Pollutant Concentrations (Ground Truth):

These variables represent the true hourly-averaged concentrations of pollutants measured using certified reference analyzers.

- **CO(GT)**:
  - Represents the concentration of Carbon Monoxide (CO) in mg/m^3.

- *Analysis:* CO is a colorless, odorless gas emitted from incomplete combustion processes, such as vehicle engines and industrial activities. High levels indicate areas with heavy traffic or industrial emissions.

- **NMHC(GT)**:
  - Represents the concentration of Non-Methane Hydrocarbons (NMHC) in µg/m^3.
  - *Analysis:* NMHC includes volatile organic compounds (VOCs) except methane. These compounds are precursors to smog formation, particularly in urban and industrial regions.

- **C6H6(GT)**:
  - Represents the concentration of Benzene (C6H6) in µg/m^3.
  - *Analysis:* Benzene is a toxic hydrocarbon mainly emitted by fuel combustion and industrial processes. Monitoring Benzene levels is critical due to its carcinogenic effects.

- **NOx(GT)**:
  - Represents the concentration of Nitrogen Oxides (NOx) in parts per billion (ppb).
  - *Analysis:* NOx gases are significant air pollutants emitted by vehicles and power plants. They contribute to acid rain and photochemical smog, affecting both the environment and human health.

- **NO2(GT)**:
  - Represents the concentration of Nitrogen Dioxide (NO2) in µg/m^3.
  - *Analysis:* NO2, a subset of NOx, is a harmful pollutant that aggravates respiratory conditions. Its levels are closely related to vehicular and industrial emissions.

Sensor Responses:
These variables capture sensor responses aimed at monitoring specific pollutants.

- **PT08.S1(CO)**:
  - Sensor response targeting CO.
  - *Analysis:* Metal-oxide-based sensors measure CO indirectly. While responsive, these sensors are prone to cross-sensitivity with other gases.

- **PT08.S2(NMHC)**:
  - Sensor response targeting NMHC.
  - *Analysis:* Useful for estimating hydrocarbon levels. Variations in sensor response may indicate sensor drift or environmental interference.

- **PT08.S3(NOx)**:
  - Sensor response targeting NOx.

  - *Analysis:* Correlates with NOx levels but may suffer from inaccuracies in areas with mixed gas sources.

- **PT08.S4(NO2)**:
  - Sensor response targeting NO2.
  - *Analysis:* Indicates NO2 levels through chemical interactions within the sensor. Environmental factors like humidity may influence sensor readings.

- **PT08.S5(O3)**:
  - Sensor response targeting Ozone (O3).
  - *Analysis:* Monitors ozone levels indirectly. Ozone is formed by sunlight-driven reactions between NOx and VOCs.

Environmental Conditions:
These variables capture meteorological data influencing air quality.

- **T (Temperature)**:
  - Measured in degrees Celsius.
  - *Analysis:* Higher temperatures accelerate chemical reactions in the atmosphere, influencing ozone formation and pollutant behavior.

- **RH (Relative Humidity)**:
  - Measured as a percentage.
  - *Analysis:* High humidity can suppress airborne particulate matter but may also distort sensor responses.

- **AH (Absolute Humidity)**:
  - Represents the actual concentration of water vapor in the air.
  - *Analysis:* Complements RH for understanding atmospheric moisture, which can interact with pollutants and influence sensor performance.

1.4 Visualization of Original Data Features

To understand the distribution and behavior of each feature, we plotted the original dataset. Each feature is visualized individually to highlight its range, variability, and trends across all observations.

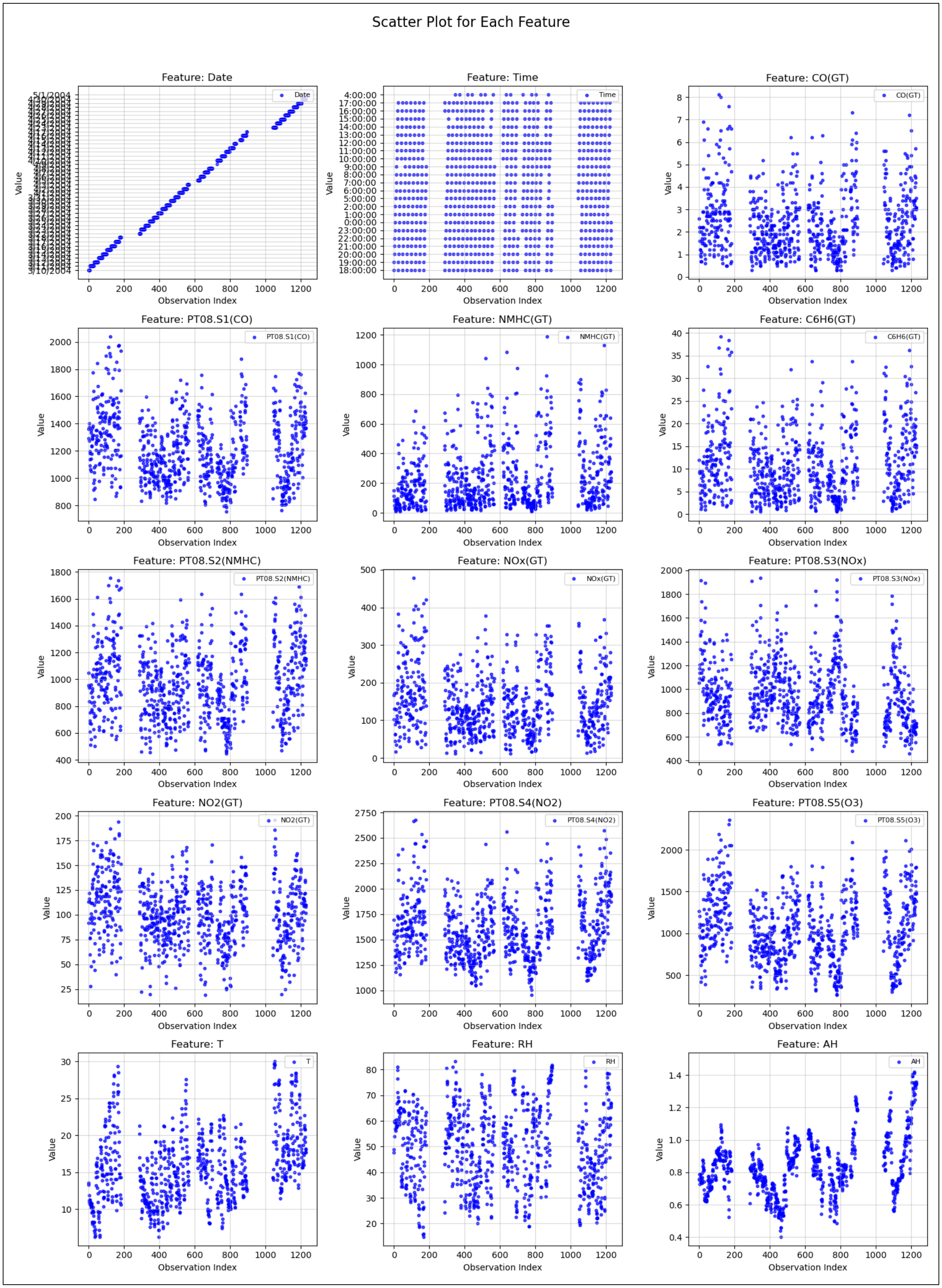
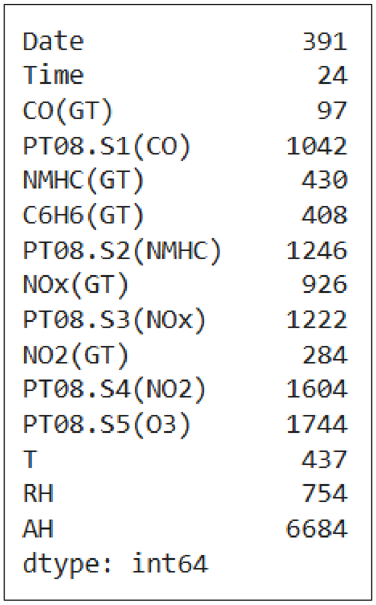


Figure 1: Quality of Representation for Individuals.

2 Data Preprocessing and cleaning

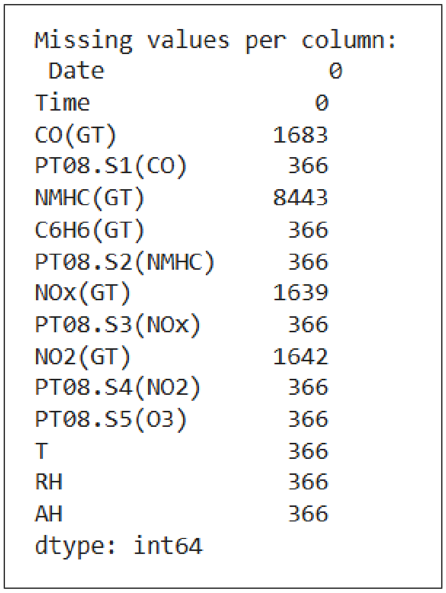
2.1 Missing Values

- **Identification of Missing Values:** The dataset documentation indicates that missing values are represented by the value `-200`. To address this, a summary of unique values and missing data was analyzed.



```
Date              391
Time               24
CO(GT)             97
PT08.S1(CO)      1042
NMHC(GT)          430
C6H6(GT)          408
PT08.S2(NMHC)    1246
NOx(GT)           926
PT08.S3(NOx)     1222
NO2(GT)           284
PT08.S4(NO2)     1604
PT08.S5(O3)      1744
T                 437
RH                754
AH               6684
dtype: int64
```

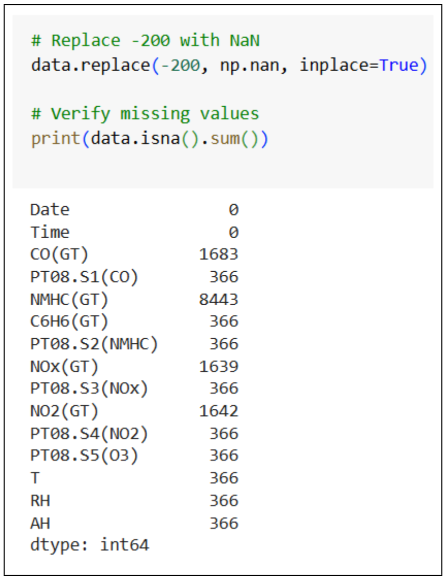
(a) Unique values summary.



```
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
```

(b) Missing values identified.

- **Handling Missing Values:** Missing values (indicated as `-200`) were replaced with `NaN` for easier handling. The count of missing values per column was then calculated to evaluate their distribution.



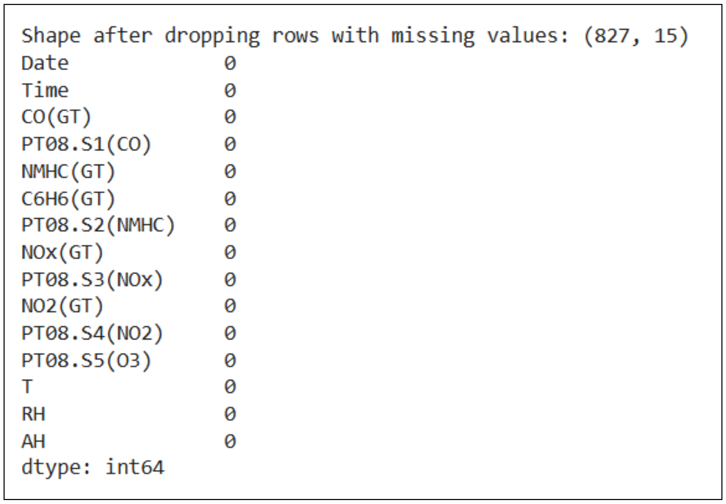
```
# Replace -200 with NaN
data.replace(-200, np.nan, inplace=True)

# Verify missing values
print(data.isna().sum())
```

```
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
```

Figure 3: Count of missing values per column.

- **Cleaning the Dataset:** Rows containing missing values were removed, resulting in a cleaned dataset with no missing entries.

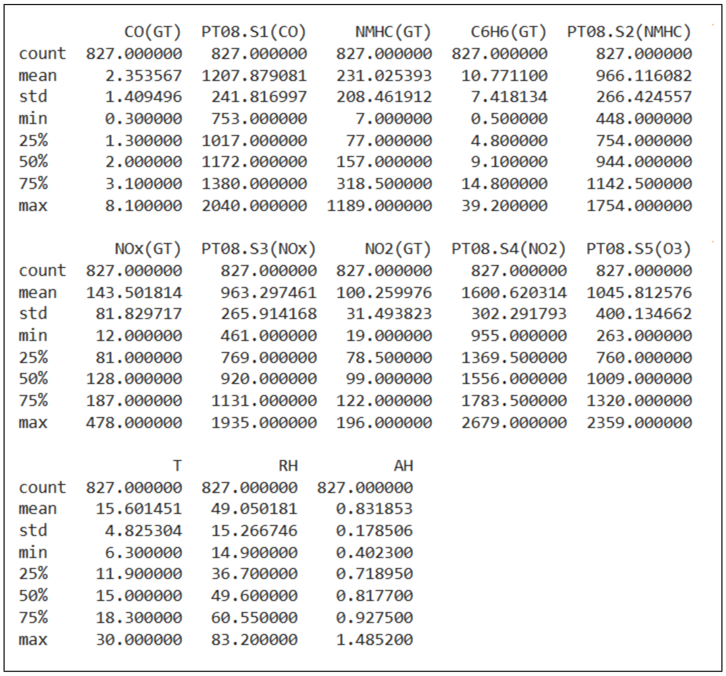


```
Shape after dropping rows with missing values: (827, 15)
Date             0
Time             0
CO(GT)           0
PT08.S1(CO)      0
NMHC(GT)         0
C6H6(GT)         0
PT08.S2(NMHC)    0
NOx(GT)          0
PT08.S3(NOx)     0
NO2(GT)          0
PT08.S4(NO2)     0
PT08.S5(O3)      0
T                0
RH               0
AH               0
dtype: int64
```

Figure 4: Cleaned dataset with no missing values.

2.2 Summary Statistics and Standard Deviation

The summary statistics for the cleaned dataset (`data_cleaned`) provide critical insights into the distribution of variables. These include metrics such as the mean, standard deviation, minimum, 25th percentile (Q1), median, 75th percentile (Q3), and maximum.



```
           CO(GT)  PT08.S1(CO)     NMHC(GT)    C6H6(GT)  PT08.S2(NMHC)
count  827.000000   827.000000   827.000000  827.000000     827.000000
mean     2.353567  1207.879081   231.025393   10.771100     966.116082
std      1.409496   241.816997   208.461912    7.418134     266.424557
min      0.300000   753.000000     7.000000    0.500000     448.000000
25%      1.300000  1017.000000    77.000000    4.800000     754.000000
50%      2.000000  1172.000000   157.000000    9.100000     944.000000
75%      3.100000  1380.000000   318.500000   14.800000    1142.500000
max      8.100000  2040.000000  1189.000000   39.200000    1754.000000

           NOx(GT)  PT08.S3(NOx)     NO2(GT)  PT08.S4(NO2)  PT08.S5(O3)
count  827.000000    827.000000  827.000000    827.000000   827.000000
mean   143.501814    963.297461  100.259976   1600.620314  1045.812576
std     81.829717    265.914168   31.493823    302.291793   400.134662
min     12.000000    461.000000   19.000000    955.000000   263.000000
25%     81.000000    769.000000   78.500000   1369.500000   760.000000
50%    128.000000    920.000000   99.000000   1556.000000  1009.000000
75%    187.000000   1131.000000  122.000000   1783.500000  1320.000000
max    478.000000   1935.000000  196.000000   2679.000000  2359.000000

                T          RH          AH
count  827.000000  827.000000  827.000000
mean    15.601451   49.050181    0.831853
std      4.825304   15.266746    0.178506
min      6.300000   14.900000    0.402300
25%     11.900000   36.700000    0.718950
50%     15.000000   49.600000    0.817700
75%     18.300000   60.550000    0.927500
max     30.000000   83.200000    1.485200
```

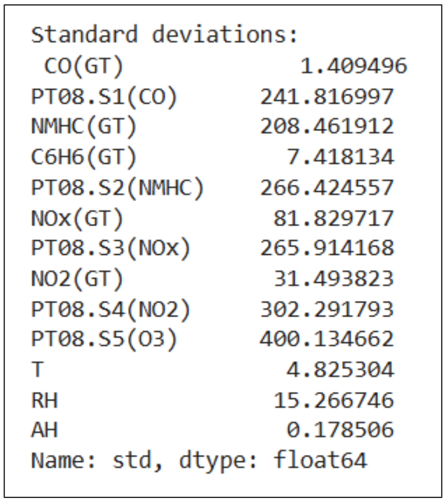
Figure 5: Summary statistics of the cleaned dataset.

Key Insights:

- **Standard Deviations:** The standard deviation values provide a measure of data variability. Higher values indicate greater dispersion, while lower values signify consistent observations.

- **High Variability Variables:** Variables such as `PT08.S1(CO)`, `NMHC(GT)`, `PT08.S2(NMHC)`, `PT08.S3(NOx)`, `PT08.S4(NO2)`, and `PT08.S5(O3)` exhibit high standard deviations, reflecting significant fluctuations. These variables may be more sensitive to environmental changes or potential data inconsistencies.

Standard Deviation Visualization:

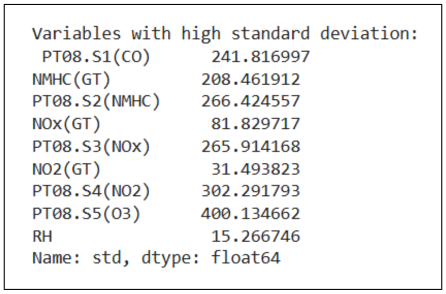


```
Standard deviations:
 CO(GT)             1.409496
PT08.S1(CO)       241.816997
NMHC(GT)          208.461912
C6H6(GT)            7.418134
PT08.S2(NMHC)     266.424557
NOx(GT)            81.829717
PT08.S3(NOx)      265.914168
NO2(GT)            31.493823
PT08.S4(NO2)      302.291793
PT08.S5(O3)       400.134662
T                   4.825304
RH                 15.266746
AH                  0.178506
Name: std, dtype: float64
```

Figure 6: Standard deviation values of key variables.

- Variables like `CO(GT)` and `T` exhibit relatively smaller standard deviations, indicating consistent measurements.
- Conversely, variables such as `PT08.S5(O3)` and `PT08.S4(NO2)` show larger deviations, suggesting higher variability across observations.

Variables with High Standard Deviations (greater than a threshold of 10):



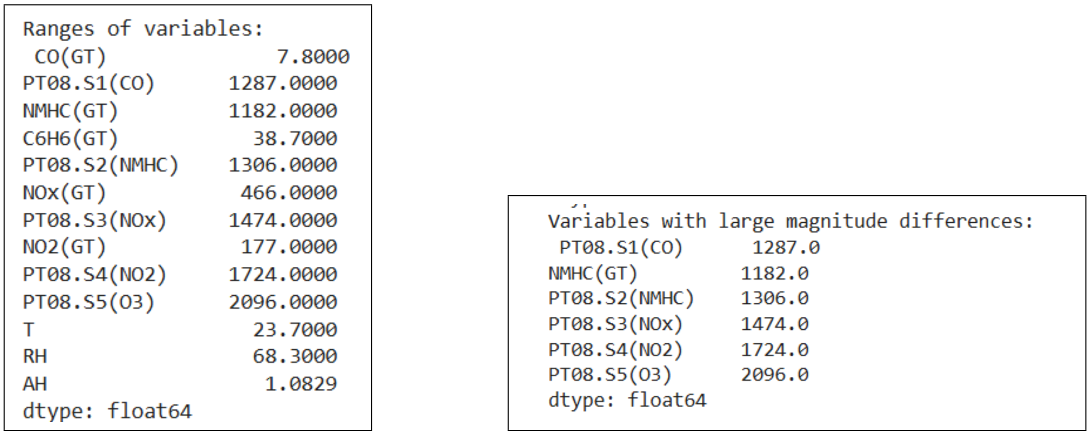
```
Variables with high standard deviation:
 PT08.S1(CO)      241.816997
NMHC(GT)         208.461912
PT08.S2(NMHC)    266.424557
NOx(GT)           81.829717
PT08.S3(NOx)     265.914168
NO2(GT)           31.493823
PT08.S4(NO2)     302.291793
PT08.S5(O3)      400.134662
RH                15.266746
Name: std, dtype: float64
```

Figure 7: Variables with high standard deviations.

These variables demonstrate substantial variability, warranting further exploration or potential data transformations to address fluctuations during analysis.

2.3 Analysis of Magnitude and Range

The **range** of a variable, calculated as the difference between its maximum and minimum values, provides insights into its magnitude and variability. Examining this range helps identify variables with significant or minimal fluctuations.



```
Ranges of variables:
 CO(GT)              7.8000
PT08.S1(CO)       1287.0000
NMHC(GT)          1182.0000
C6H6(GT)            38.7000
PT08.S2(NMHC)     1306.0000
NOx(GT)            466.0000
PT08.S3(NOx)      1474.0000
NO2(GT)            177.0000
PT08.S4(NO2)      1724.0000
PT08.S5(O3)       2096.0000
T                   23.7000
RH                  68.3000
AH                   1.0829
dtype: float64
```

(a) Variable magnitudes.

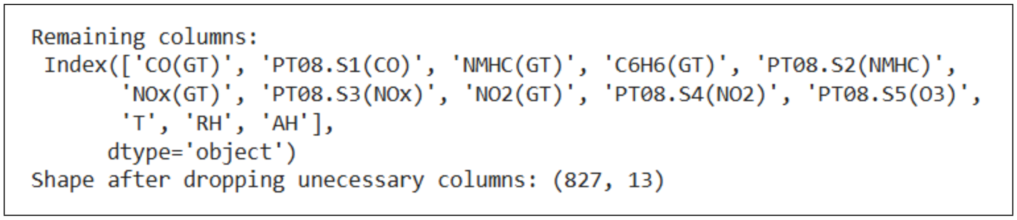
```
Variables with large magnitude differences:
 PT08.S1(CO)      1287.0
NMHC(GT)         1182.0
PT08.S2(NMHC)    1306.0
PT08.S3(NOx)     1474.0
PT08.S4(NO2)     1724.0
PT08.S5(O3)      2096.0
dtype: float64
```

(b) Range of variables.

Figure 8: Magnitude and range analysis of key variables.

Data Cleaning and Reducing the Dataset

After analyzing the dataset for missing values and standard deviations, **columns like `Date` and `Time` were dropped**, as they were not relevant to the analysis of air quality and environmental conditions and they are not interesting for our study. This reduced the dataset to the following 13 columns:



```
Remaining columns:
 Index(['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'],
      dtype='object')
Shape after dropping unecessary columns: (827, 13)
```

By removing these columns, the dataset now focuses on the environmental and sensor-related variables, which are central to air quality and atmospheric studies.

3 Principal Component Analysis (PCA): A Data-Driven Exploration

3.1 Why We Choose PCA

Principal Component Analysis (PCA) was selected due to the nature of the dataset:

1. **Quantitative Data:** The dataset consists of numerical variables, which align well with the mathematical foundations of PCA.

2. **Highly Correlated Variables:** Several features exhibit strong intercorrelations, making PCA an ideal technique to reduce dimensionality while retaining the most critical information.

3.2 Type of PCA

The decision to use the **centered and reduced** version of Principal Component Analysis (PCA) was made for several important reasons:

1. **High Standard Deviations:** Some features in the dataset exhibit significantly high standard deviations, indicating that these variables have large variations in their values.

2. **Different Magnitudes:** The variables in the dataset span a wide range of values.

3. **Different Units:** The dataset includes features with different units of measurement, such as concentration values (e.g., `CO(GT)` in ppm) and temperature (e.g., `T` in degrees Celsius).

By centering (subtracting the mean) and reducing (dividing by the standard deviation) each feature, we ensure that PCA treats all variables equally, leading to a more accurate and meaningful decomposition of the data.

3.3 Standardizing the Data for PCA

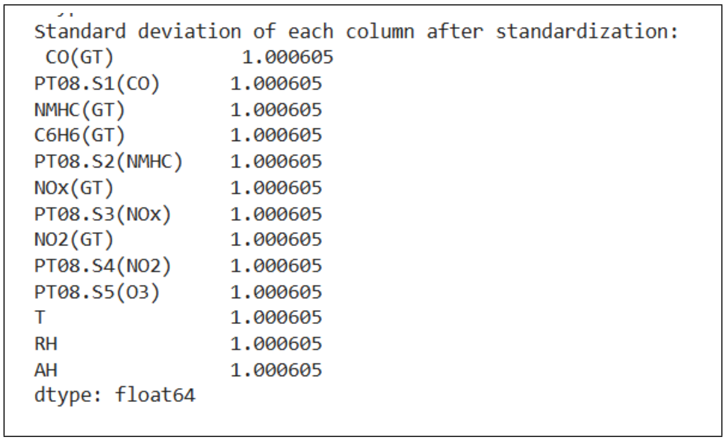
To ensure the effectiveness of PCA, the dataset was standardized, as the technique is sensitive to the scale of the variables. Standardization adjusts the features such that each has:

- A mean of **0** (centered).
- A standard deviation of **1** (reduced).

3.3.1 Steps to Standardize the Data

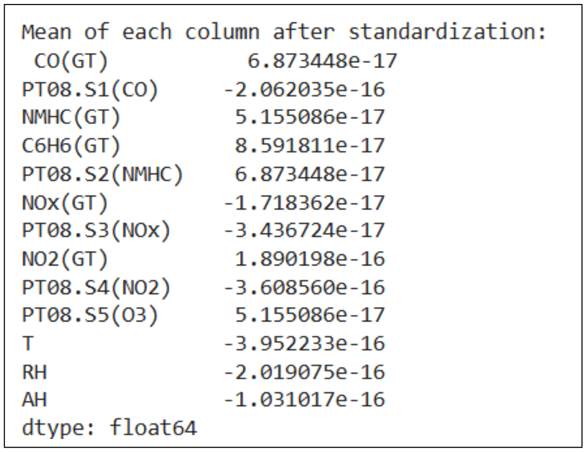
1. **Centering:** Subtract the mean of each feature to ensure the data is centered around zero.
2. **Scaling:** Divide by the standard deviation to ensure all features contribute equally, regardless of their original scale.

This process is essential when variables have different units, scales, or magnitudes, as is the case with this dataset.



```
Standard deviation of each column after standardization:
 CO(GT)           1.000605
PT08.S1(CO)       1.000605
NMHC(GT)          1.000605
C6H6(GT)          1.000605
PT08.S2(NMHC)     1.000605
NOx(GT)           1.000605
PT08.S3(NOx)      1.000605
NO2(GT)           1.000605
PT08.S4(NO2)      1.000605
PT08.S5(O3)       1.000605
T                 1.000605
RH                1.000605
AH                1.000605
dtype: float64
```

(a) Ensuring that std is 1



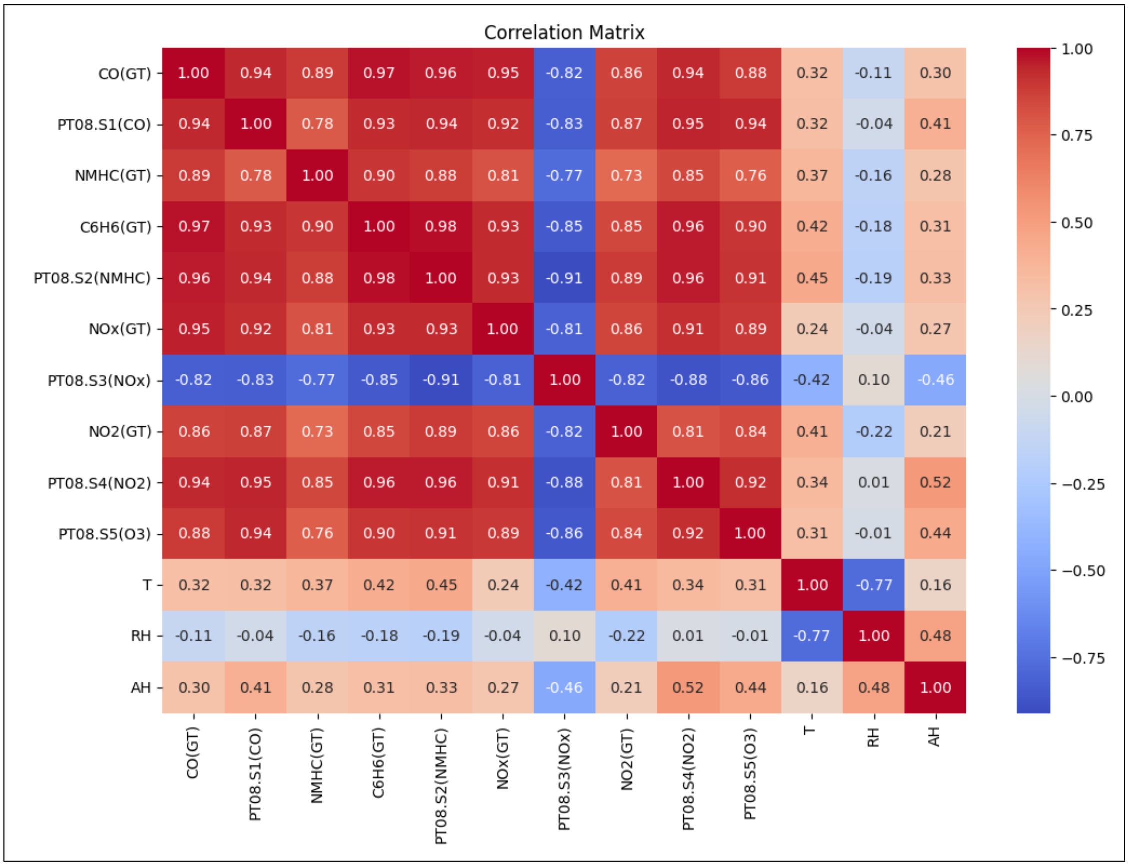
```
Mean of each column after standardization:
 CO(GT)           6.873448e-17
PT08.S1(CO)      -2.062035e-16
NMHC(GT)          5.155086e-17
C6H6(GT)          8.591811e-17
PT08.S2(NMHC)     6.873448e-17
NOx(GT)          -1.718362e-17
PT08.S3(NOx)     -3.436724e-17
NO2(GT)           1.890198e-16
PT08.S4(NO2)     -3.608560e-16
PT08.S5(O3)       5.155086e-17
T                -3.952233e-16
RH               -2.019075e-16
AH               -1.031017e-16
dtype: float64
```

(b) Ensuring the mean of data is 0

Mean and Standard Deviation after Standardization: All columns now have a mean close to zero and a standard deviation close to one, ensuring fair contributions of features in PCA.

3.4 Analyzing the Correlation Matrix

The correlation matrix reveals relationships between variables. A threshold of **0.8** was chosen to classify relationships as **strongly correlated**, **strongly anti-correlated**, or **weakly correlated**. This threshold provides clarity for identifying relevant patterns in the dataset.



3.4.1 Strong Correlations ($|r| \geq 0.8$)

Variables with correlations above the threshold are highly related, as shown below:

Variable 1	Variable 2	Correlation Coefficient
CO(GT)	C6H6(GT)	0.97
CO(GT)	PT08.S1(CO)	0.94
C6H6(GT)	PT08.S2(NMHC)	0.98
PT08.S1(CO)	PT08.S2(NMHC)	0.94
CO(GT)	NOx(GT)	0.95
C6H6(GT)	PT08.S1(CO)	0.93
PT08.S4(NO2)	PT08.S5(O3)	0.92

3.4.2 Strong Anti-Correlations ($r \leq -0.8$)

Negative correlations exceeding the threshold suggest inverse relationships:

Variable 1	Variable 2	Correlation Coefficient
PT08.S3(NOx)	CO(GT)	-0.82
PT08.S3(NOx)	PT08.S1(CO)	-0.83
PT08.S3(NOx)	C6H6(GT)	-0.85
PT08.S3(NOx)	PT08.S2(NMHC)	-0.91
PT08.S3(NOx)	PT08.S4(NO2)	-0.88
PT08.S3(NOx)	PT08.S5(O3)	-0.86

3.4.3 Weak or No Correlations ($|r| < 0.8$)

Variables below the threshold show little to no relationship, as seen here:

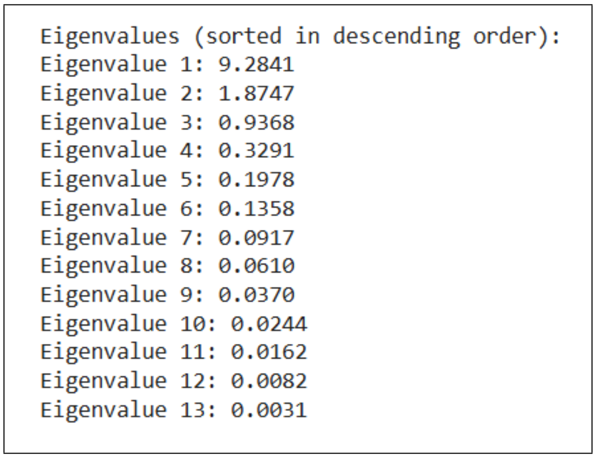
Variable 1	Variable 2	Correlation Coefficient
RH	CO(GT)	-0.11
RH	PT08.S1(CO)	-0.04
RH	C6H6(GT)	-0.18
AH	T	0.16

4 Eigenvectors and Eigenvalues

In Principal Component Analysis (PCA), **eigenvalues** represent the amount of variance captured by each principal component (PC). A larger eigenvalue indicates that the corresponding eigenvector captures a greater portion of the total variance in the dataset. Consequently, principal components with larger eigenvalues are more significant in explaining the structure of the data.

4.1 Eigenvalues Analysis

The eigenvalues for the correlation matrix, sorted in descending order, are as follows:



```
Eigenvalues (sorted in descending order):
Eigenvalue 1: 9.2841
Eigenvalue 2: 1.8747
Eigenvalue 3: 0.9368
Eigenvalue 4: 0.3291
Eigenvalue 5: 0.1978
Eigenvalue 6: 0.1358
Eigenvalue 7: 0.0917
Eigenvalue 8: 0.0610
Eigenvalue 9: 0.0370
Eigenvalue 10: 0.0244
Eigenvalue 11: 0.0162
Eigenvalue 12: 0.0082
Eigenvalue 13: 0.0031
```

Figure 10: Sorted eigenvalues of the correlation matrix.

Significance:

- Most of the variance is captured by the first few eigenvalues (typically the first two or three), while the remaining eigenvalues contribute relatively little.

- This pattern indicates that a significant portion of the dataset's variability can be explained using a smaller number of principal components.

- By focusing on the principal components corresponding to the largest eigenvalues, **dimensionality reduction** can be effectively achieved without losing much information.

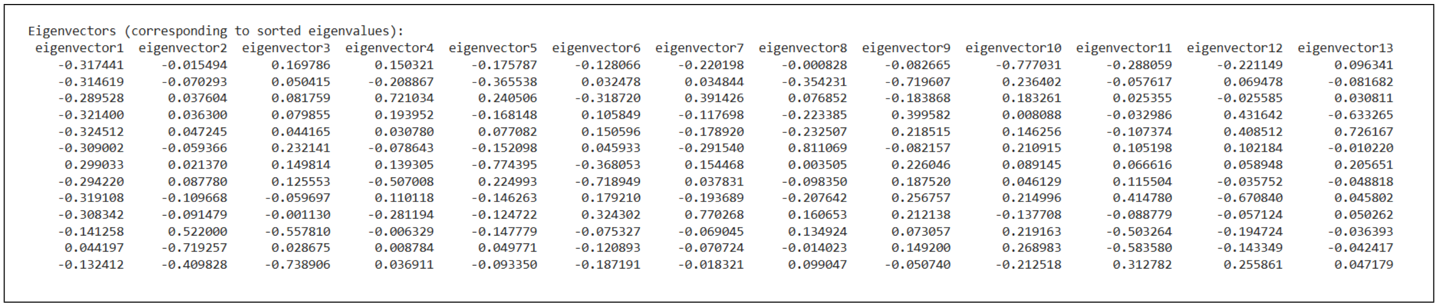
4.2 Eigenvectors and Their Interpretation

The **eigenvectors** are the directions in the data space that correspond to the eigenvalues. Each eigenvector represents a principal component and can be interpreted as a **linear combination of the original features (variables)**.

- The *weights* of the features within each eigenvector indicate their relative importance in defining the corresponding principal component.

- Features with larger absolute weights contribute more significantly to the direction of the principal component.

The following figure summarizes the eigenvectors for the sorted eigenvalues:



```
Eigenvectors (corresponding to sorted eigenvalues):
 eigenvector1  eigenvector2  eigenvector3  eigenvector4  eigenvector5  eigenvector6  eigenvector7  eigenvector8  eigenvector9  eigenvector10  eigenvector11  eigenvector12  eigenvector13
    -0.317441     -0.015494      0.169786      0.150321     -0.175787     -0.128066     -0.220198     -0.000828     -0.082665      -0.777031      -0.288059      -0.221149       0.096341
    -0.314619     -0.070293      0.050415     -0.208867     -0.365538      0.032478      0.034844     -0.354231     -0.719607       0.236402      -0.057617       0.069478      -0.081682
    -0.289528      0.037604      0.081759      0.721034      0.240506     -0.318720      0.391426      0.076852     -0.183868       0.183261       0.025355      -0.025585       0.030811
    -0.321400      0.036300      0.079855      0.193952     -0.168148      0.105849     -0.117698     -0.223385      0.399582       0.008088      -0.032986       0.431642      -0.633265
    -0.324512      0.047245      0.044165      0.030780      0.077082      0.150596     -0.178920     -0.232507      0.218515       0.146256      -0.107374       0.408512       0.726167
    -0.309002     -0.059366      0.232141     -0.078643     -0.152098      0.045933     -0.291540      0.811069     -0.082157       0.210915       0.105198       0.102184      -0.010220
     0.299033      0.021370      0.149814      0.139305     -0.774395     -0.368053      0.154468      0.003505      0.226046       0.089145       0.066616       0.058948       0.205651
    -0.294220      0.087780      0.125553     -0.507008      0.224993     -0.718949      0.037831     -0.098350      0.187520       0.046129       0.115504      -0.035752      -0.048818
    -0.319108     -0.109668     -0.059697      0.110118     -0.146263      0.179210     -0.193689     -0.207642      0.256757       0.214996       0.414780      -0.670840       0.045802
    -0.308342     -0.091479     -0.001130     -0.281194     -0.124722      0.324302      0.770268      0.160653      0.212138      -0.137708      -0.088779      -0.057124       0.050262
    -0.141258      0.522000     -0.557810     -0.006329     -0.147779     -0.075327     -0.069045      0.134924      0.073057       0.219163      -0.503264      -0.194724      -0.036393
     0.044197     -0.719257      0.028675      0.008784      0.049771     -0.120893     -0.070724     -0.014023      0.149200       0.268983      -0.583580      -0.143349      -0.042417
    -0.132412     -0.409828     -0.738906      0.036911     -0.093350     -0.187191     -0.018321      0.099047     -0.050740      -0.212518       0.312782       0.255861       0.047179
```

Figure 11: Eigenvectors corresponding to the sorted eigenvalues.

By analyzing the eigenvalues and their corresponding eigenvectors:

1. We can identify the most important principal components based on the magnitude of their eigenvalues.

2. The eigenvectors help explain how the original variables contribute to each principal component.

This allows us to reduce the dimensionality of the dataset while retaining the features that contribute the most to the variability in the data.

5 Matrix of New Principal Components F

The matrix of new principal components F represents the transformed dataset in the principal component space. This transformation is performed using the formula:

$$F = X_s \cdot U$$

Where:

- F: The new matrix in the principal component space.
- X_s: The standardized data matrix.
- U: The eigenvector matrix, where each column corresponds to a principal component.

5.1 Structure of F

The matrix F has the following structure:

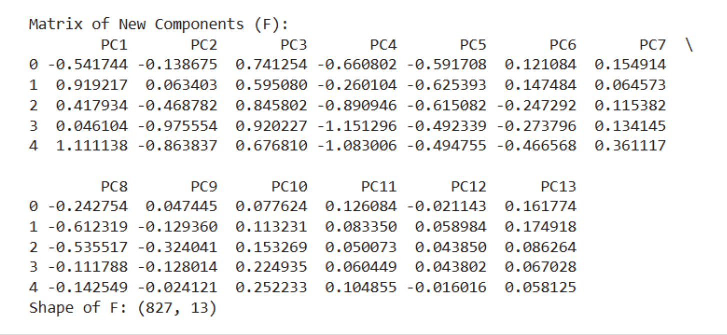
- Each **row** in F represents a data point (observation) in the new principal component space.
- Each **column** in F represents one of the principal components.

The dimensions of F are $(827, 13)$, where:

- 827: Number of observations (data points) in the dataset.
- 13: Number of principal components, which equals the number of original variables.

5.2 Principal Component Scores

Each cell in F contains the **score** of a particular observation on a specific principal component. These scores quantify the projection of each data point onto the principal components.



```
Matrix of New Components (F):
        PC1       PC2       PC3       PC4       PC5       PC6       PC7  \
0 -0.541744 -0.138675  0.741254 -0.660802 -0.591708  0.121084  0.154914
1  0.919217  0.063403  0.595080 -0.260104 -0.625393  0.147484  0.064573
2  0.417934 -0.468782  0.845802 -0.890946 -0.615082 -0.247292  0.115382
3  0.046104 -0.975554  0.920227 -1.151296 -0.492339 -0.273796  0.134145
4  1.111138 -0.863837  0.676810 -1.083006 -0.494755 -0.466568  0.361117

        PC8       PC9      PC10      PC11      PC12      PC13
0 -0.242754  0.047445  0.077624  0.126084 -0.021143  0.161774
1 -0.612319 -0.129360  0.113231  0.083350  0.058984  0.174918
2 -0.535517 -0.324041  0.153269  0.050073  0.043850  0.086264
3 -0.111788 -0.128014  0.224935  0.060449  0.043802  0.067028
4 -0.142549 -0.024121  0.252233  0.104855 -0.016016  0.058125
Shape of F: (827, 13)
```

Figure 12: the matrix of principal components for the first 5 rows

For instance:

- The first observation has a score of -0.5417 for PC1 and -0.1387 for PC2.

6 The Saturation Matrix S

The saturation matrix S describes the relationship between the original variables and the principal components. It is calculated using the following formula:

$$S = t_X \times F \times D^{\frac{1}{2}}$$

Where:

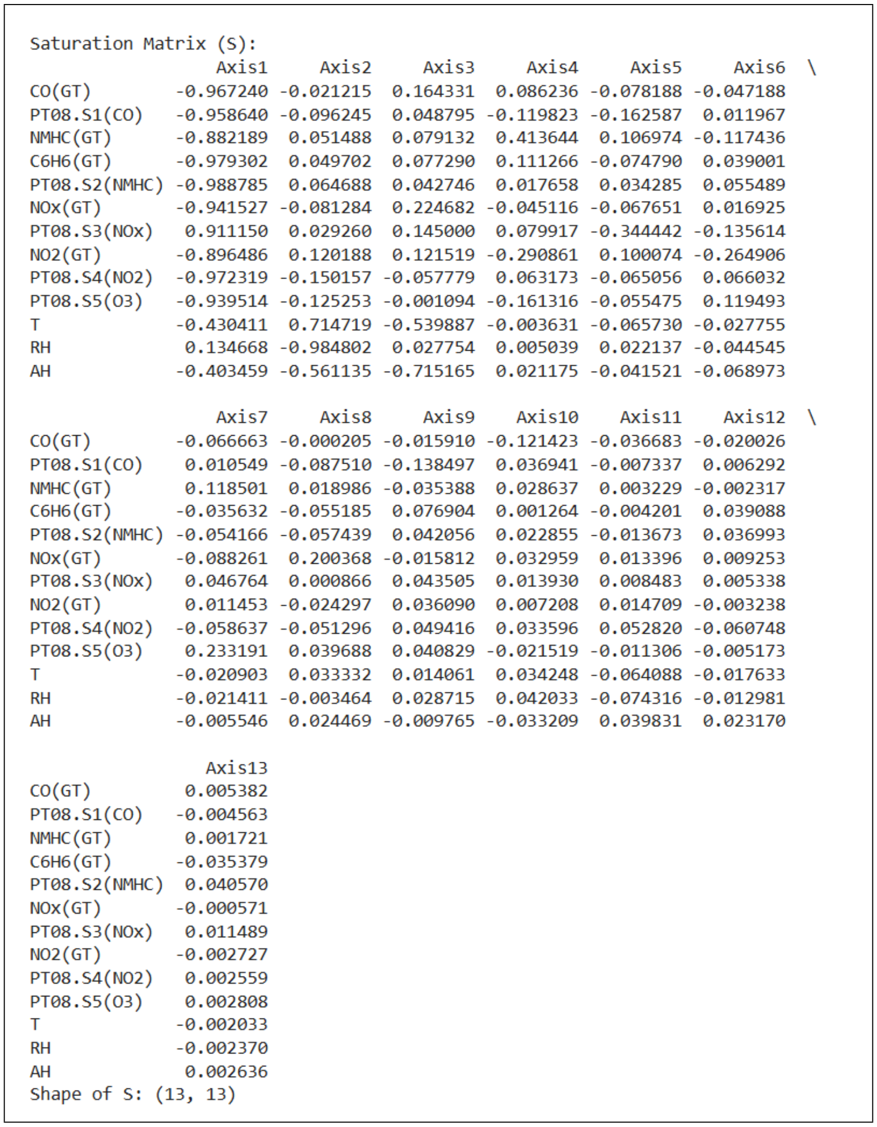
- t_X: The transpose of the standardized data matrix.
- F: The matrix of new principal components.
- $D^{\frac{1}{2}}$: The square root of the diagonal matrix of eigenvalues D.

6.1 Structure of S

The saturation matrix S has dimensions $(13, 13)$, where:

- Each **row** corresponds to an original variable.
- Each **column** corresponds to a principal component.

The values in S represent the **relationship** between each original variable and the principal components. This latter indicates both the strength and the direction of the relationships (the correlation between the new factors and original features).



```
Saturation Matrix (S):
                 Axis1     Axis2     Axis3     Axis4     Axis5     Axis6  \
CO(GT)       -0.967240 -0.021215  0.164331  0.086236 -0.078188 -0.047188
PT08.S1(CO)  -0.958640 -0.096245  0.048795 -0.119823 -0.162587  0.011967
NMHC(GT)     -0.882189  0.051488  0.079132  0.413644  0.106974 -0.117436
C6H6(GT)     -0.979302  0.049702  0.077290  0.111266 -0.074790  0.039001
PT08.S2(NMHC) -0.988785  0.064688  0.042746  0.017658  0.034285  0.055489
NOx(GT)      -0.941527 -0.081284  0.224682 -0.045116 -0.067651  0.016925
PT08.S3(NOx)  0.911150  0.029260  0.145000  0.079917 -0.344442 -0.135614
NO2(GT)      -0.896486  0.120188  0.121519 -0.290861  0.100074 -0.264906
PT08.S4(NO2) -0.972319 -0.150157 -0.057779  0.063173 -0.065056  0.066032
PT08.S5(O3)  -0.939514 -0.125253 -0.001094 -0.161316 -0.055475  0.119493
T            -0.430411  0.714719 -0.539887 -0.003631 -0.065730 -0.027755
RH            0.134668 -0.984802  0.027754  0.005039  0.022137 -0.044545
AH           -0.403459 -0.561135 -0.715165  0.021175 -0.041521 -0.068973

                 Axis7     Axis8     Axis9    Axis10    Axis11    Axis12  \
CO(GT)       -0.066663 -0.000205 -0.015910 -0.121423 -0.036683 -0.020026
PT08.S1(CO)   0.010549 -0.087510 -0.138497  0.036941 -0.007337  0.006292
NMHC(GT)      0.118501  0.018986 -0.035388  0.028637  0.003229 -0.002317
C6H6(GT)     -0.035632 -0.055185  0.076904  0.001264 -0.004201  0.039088
PT08.S2(NMHC) -0.054166 -0.057439  0.042056  0.022855 -0.013673  0.036993
NOx(GT)      -0.088261  0.200368 -0.015812  0.032959  0.013396  0.009253
PT08.S3(NOx)  0.046764  0.000866  0.043505  0.013930  0.008483  0.005338
NO2(GT)       0.011453 -0.024297  0.036090  0.007208  0.014709 -0.003238
PT08.S4(NO2) -0.058637 -0.051296  0.049416  0.033596  0.052820 -0.060748
PT08.S5(O3)   0.233191  0.039688  0.040829 -0.021519 -0.011306 -0.005173
T            -0.020903  0.033332  0.014061  0.034248 -0.064088 -0.017633
RH           -0.021411 -0.003464  0.028715  0.042033 -0.074316 -0.012981
AH           -0.005546  0.024469 -0.009765 -0.033209  0.039831  0.023170

                Axis13
CO(GT)        0.005382
PT08.S1(CO)  -0.004563
NMHC(GT)      0.001721
C6H6(GT)     -0.035379
PT08.S2(NMHC)  0.040570
NOx(GT)      -0.000571
PT08.S3(NOx)  0.011489
NO2(GT)      -0.002727
PT08.S4(NO2)  0.002559
PT08.S5(O3)   0.002808
T            -0.002033
RH           -0.002370
AH            0.002636
Shape of S: (13, 13)
```

Figure 13: the saturation matrix

6.2 Significance of the Saturation Matrix

The saturation matrix provides critical insights into the principal component analysis:

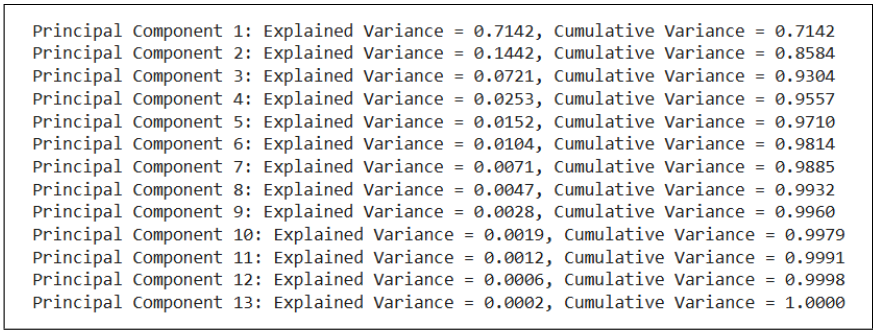
- It reveals how much each principal component contributes to explaining the variation in each original variable.
- It identifies which variables **dominate** or have the strongest influence on each principal component.
- It confirms that earlier principal components (e.g., PC1, PC2) explain the most variance, while later components primarily capture minor trends or noise.

By examining the saturation matrix:

- We can determine the variables most strongly associated with each principal component.
- This helps in identifying the key contributors to each axis, supporting dimensionality reduction and further analysis.

Overall, the matrix of new principal components F and the saturation matrix S together enable a comprehensive understanding of the transformed dataset and the relationships between the original variables and the principal components.

7 Choice of Factorial Axes and Computation of Overall Quality of Explanation



```
Principal Component 1: Explained Variance = 0.7142, Cumulative Variance = 0.7142
Principal Component 2: Explained Variance = 0.1442, Cumulative Variance = 0.8584
Principal Component 3: Explained Variance = 0.0721, Cumulative Variance = 0.9304
Principal Component 4: Explained Variance = 0.0253, Cumulative Variance = 0.9557
Principal Component 5: Explained Variance = 0.0152, Cumulative Variance = 0.9710
Principal Component 6: Explained Variance = 0.0104, Cumulative Variance = 0.9814
Principal Component 7: Explained Variance = 0.0071, Cumulative Variance = 0.9885
Principal Component 8: Explained Variance = 0.0047, Cumulative Variance = 0.9932
Principal Component 9: Explained Variance = 0.0028, Cumulative Variance = 0.9960
Principal Component 10: Explained Variance = 0.0019, Cumulative Variance = 0.9979
Principal Component 11: Explained Variance = 0.0012, Cumulative Variance = 0.9991
Principal Component 12: Explained Variance = 0.0006, Cumulative Variance = 0.9998
Principal Component 13: Explained Variance = 0.0002, Cumulative Variance = 1.0000
```

Figure 14: Explained variance for each principal component.

The selection of factorial axes is based on the proportion of variance explained by each principal component (PC). The key results are as follows:

- **PC1 (Principal Component 1)** explains 71.42% of the total variance, capturing the dominant pattern in the data.
- **PC2** explains 14.42%, bringing the cumulative explained variance for the first two components to 85.84%.
- **PC3** explains 7.21%, leading to a cumulative total of 93.04%.
- **PC4** explains 2.53%, and **PC5** contributes 2.90%, resulting in a cumulative explained variance of 97.10%.

and so on.

7.1 Cumulative Explained Variance Plot

The plot of Cumulative Explained Variance versus the Number of Principal Components illustrates how variance is distributed across the components. Key levels of cumulative variance are highlighted in the figure:

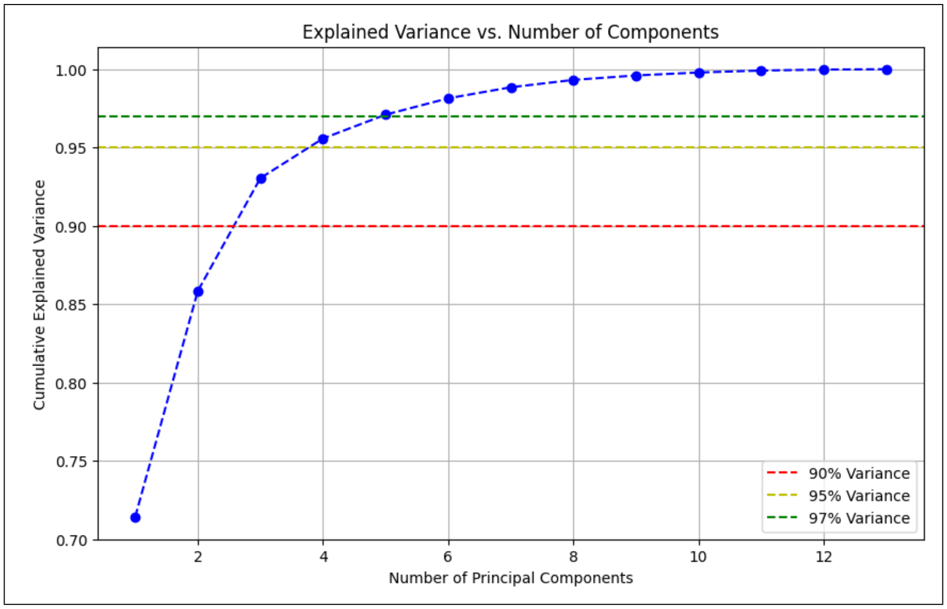


Figure 15: Cumulative explained variance as a function of the number of components.

Key Observations:

- **90% Variance (Red Line)**: Achieved with the first **3 components**, explaining 93.04% of the variance.
- **95% Variance (Yellow Line)**: Reached with **4 components**, explaining 95.57% of the variance.
- **97% Variance (Green Line)**: Achieved with the first **5 components**, explaining 97.10% of the total variance.

7.2 Choosing the Number of Principal Components

Based on the cumulative explained variance and the trade-off between information retention and dimensionality reduction:

- To explain **97%** of the total variance, the optimal number of components is **5**.
- Retaining these 5 components captures an overall quality of explanation of **97.10%** is achieved.

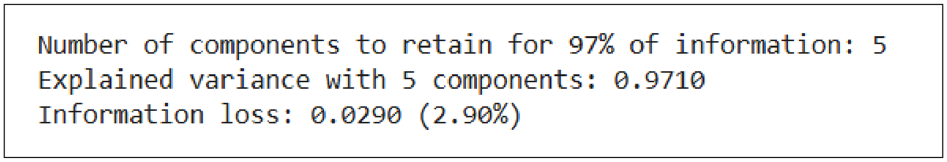
7.3 Explained Variance and Information Loss

Explained Variance with 5 Components: The first 5 principal components together explain **97.10%** of the total variance, providing a strong representation of the original dataset.
Information Loss: The remaining **2.90%** of the variance is not captured by the first 5 components. This small loss indicates that the reduction in dimensionality has minimal impact on the overall structure of the data.

7.4 Conclusion

By retaining the first **5 principal components**:

- The majority of the variability in the data (**97.10%**) is preserved.
- The dimensionality of the dataset is significantly reduced while maintaining its core structure.
- The minimal information loss (**2.90%**) ensures that the data remains well-represented for further analysis.

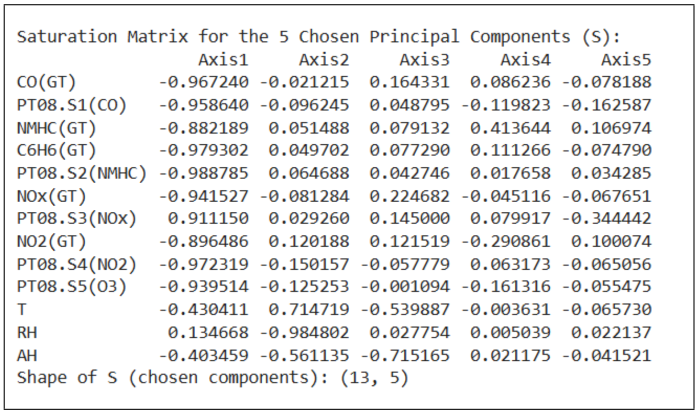


```
Number of components to retain for 97% of information: 5
Explained variance with 5 components: 0.9710
Information loss: 0.0290 (2.90%)
```

Figure 16: Retaining 5 components to explain 97.10% of the variance.

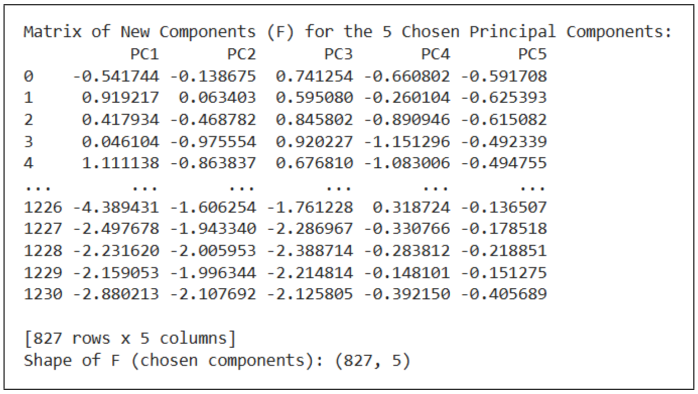
8 Analysis of Correlation Circles and Saturation Matrix

8.1 Visualization of the Saturation Matrix and Principal Component Matrix for the 5 components



```
Saturation Matrix for the 5 Chosen Principal Components (S):
                 Axis1     Axis2     Axis3     Axis4     Axis5
CO(GT)       -0.967240 -0.021215  0.164331  0.086236 -0.078188
PT08.S1(CO)  -0.958640 -0.096245  0.048795 -0.119823 -0.162587
NMHC(GT)     -0.882189  0.051488  0.079132  0.413644  0.106974
C6H6(GT)     -0.979302  0.049702  0.077290  0.111266 -0.074790
PT08.S2(NMHC) -0.988785  0.064688  0.042746  0.017658  0.034285
NOx(GT)      -0.941527 -0.081284  0.224682 -0.045116 -0.067651
PT08.S3(NOx)  0.911150  0.029260  0.145000  0.079917 -0.344442
NO2(GT)      -0.896486  0.120188  0.121519 -0.290861  0.100074
PT08.S4(NO2) -0.972319 -0.150157 -0.057779  0.063173 -0.065056
PT08.S5(O3)  -0.939514 -0.125253 -0.001094 -0.161316 -0.055475
T            -0.430411  0.714719 -0.539887 -0.003631 -0.065730
RH            0.134668 -0.984802  0.027754  0.005039  0.022137
AH           -0.403459 -0.561135 -0.715165  0.021175 -0.041521
Shape of S (chosen components): (13, 5)
```

Figure 17: Saturation Matrix S.



```
Matrix of New Components (F) for the 5 Chosen Principal Components:
           PC1       PC2       PC3       PC4       PC5
0    -0.541744 -0.138675  0.741254 -0.660802 -0.591708
1     0.919217  0.063403  0.595080 -0.260104 -0.625393
2     0.417934 -0.468782  0.845802 -0.890946 -0.615082
3     0.046104 -0.975554  0.920227 -1.151296 -0.492339
4     1.111138 -0.863837  0.676810 -1.083006 -0.494755
...        ...       ...       ...       ...       ...
1226 -4.389431 -1.606254 -1.761228  0.318724 -0.136507
1227 -2.497678 -1.943340 -2.286967 -0.330766 -0.178518
1228 -2.231620 -2.005953 -2.388714 -0.283812 -0.218851
1229 -2.159053 -1.996344 -2.214814 -0.148101 -0.151275
1230 -2.880213 -2.107692 -2.125805 -0.392150 -0.405689

[827 rows x 5 columns]
Shape of F (chosen components): (827, 5)
```

Figure 18: Matrix of New Principal Components F.

8.2 Correlation Circles: Methodology explanation

The correlation circles provide a visual representation of how the original variables contribute to the principal components. To plot them, we applied a dynamic threshold based on the 70^{th} percentile of the absolute loadings for each axis. This ensures that only the most significant variables are highlighted in the plots.

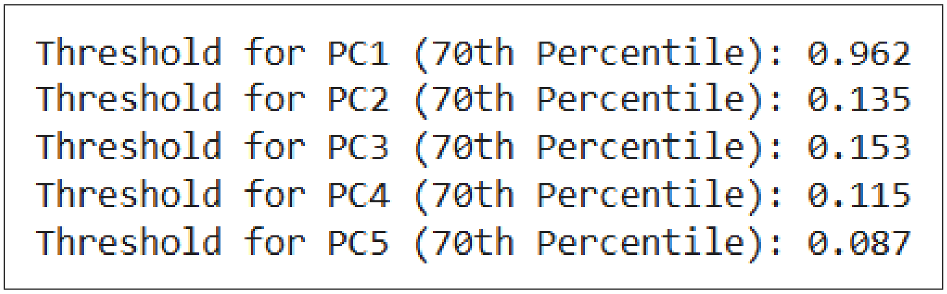
The following steps were used to generate the correlation circles:

1. Compute the **dynamic threshold** for each principal component using the 70^{th} percentile of the absolute values of the saturation matrix S.

2. For each pair of principal components (e.g., PC1 vs. PC2), plot the variables meeting the threshold on both axes.

3. Add a unit circle to highlight the boundaries of the correlation space.

4. Use labeled points instead of arrows for better clarity.



```
Threshold for PC1 (70th Percentile): 0.962
Threshold for PC2 (70th Percentile): 0.135
Threshold for PC3 (70th Percentile): 0.153
Threshold for PC4 (70th Percentile): 0.115
Threshold for PC5 (70th Percentile): 0.087
```

Figure 19: the definied threshold for each axis

8.3 Correlation Circle Analysis and Interpretation

8.4 Visualizing the Correlation Circles

The correlation circles for the first five principal components are displayed below. Each plot shows the relationships between variables for a pair of components:

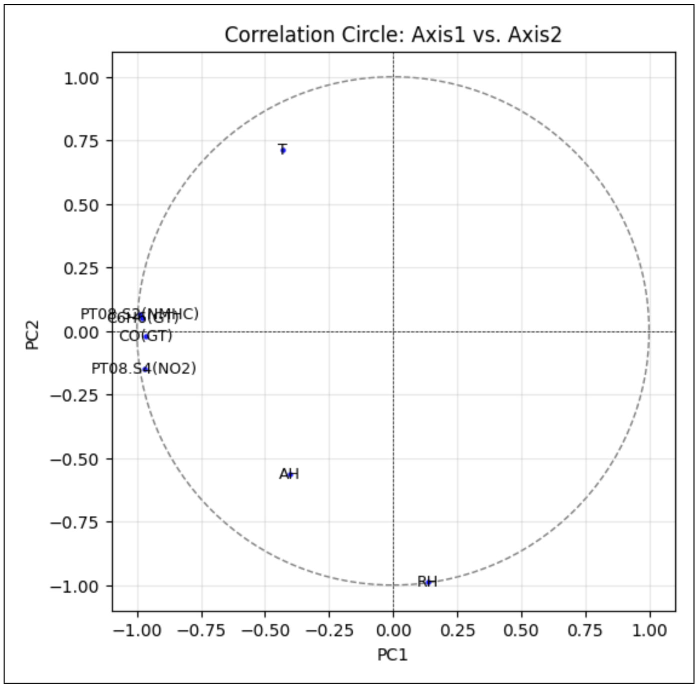


Figure 20: Correlation Circle: PC1 vs. PC2.

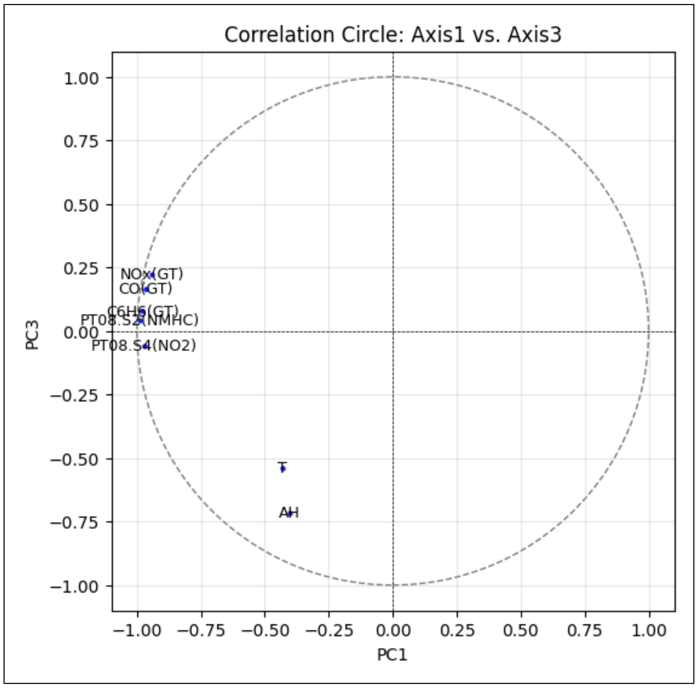


Figure 21: Correlation Circle: PC1 vs. PC3.

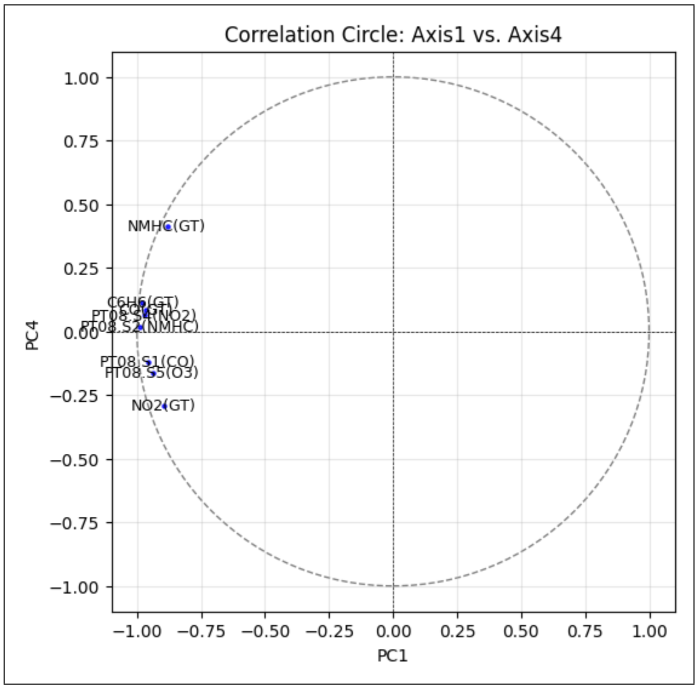


Figure 22: Correlation Circle: PC1 vs. PC4.

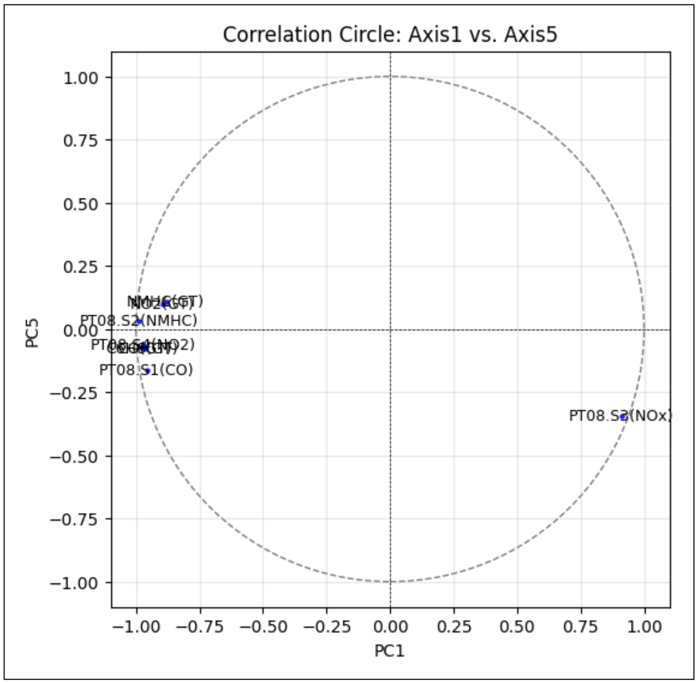


Figure 23: Correlation Circle: PC1 vs. PC5.

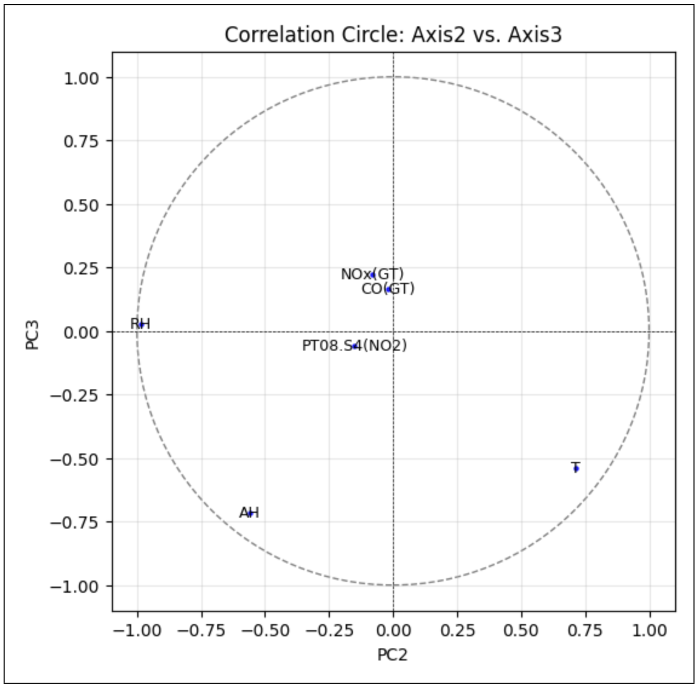


Figure 24: Correlation Circle: PC2 vs. PC3.

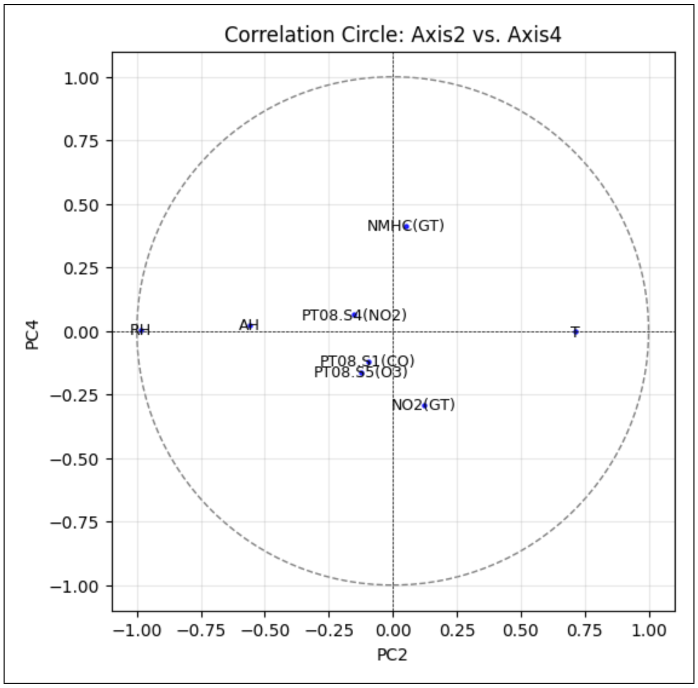


Figure 25: Correlation Circle: PC2 vs. PC4.

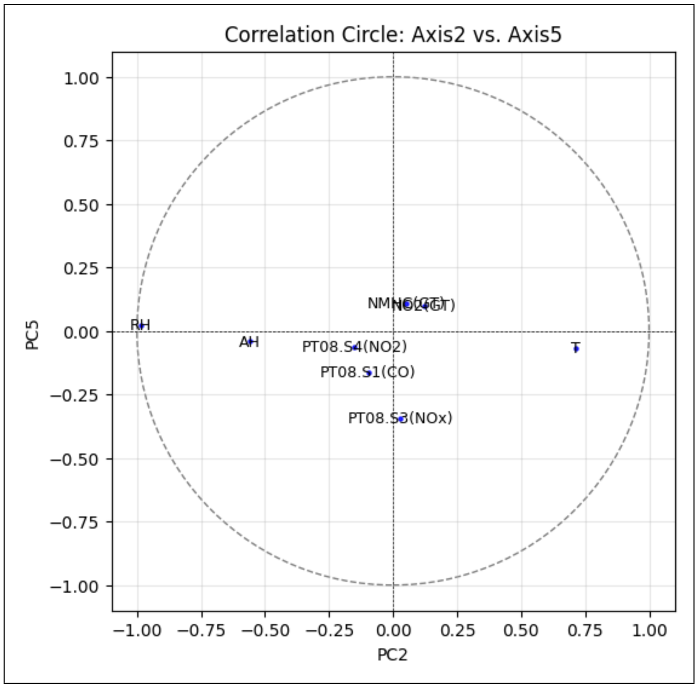


Figure 26: Correlation Circle: PC2 vs. PC5.

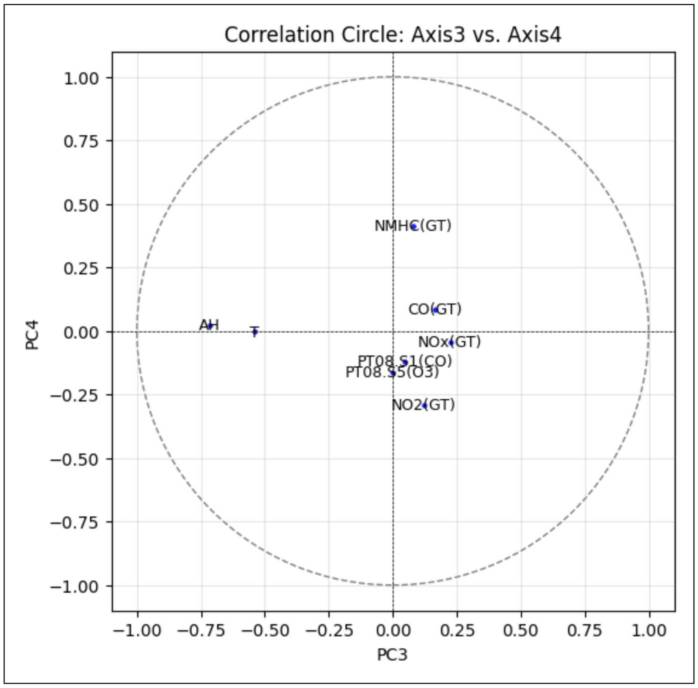


Figure 27: Correlation Circle: PC3 vs. PC4.

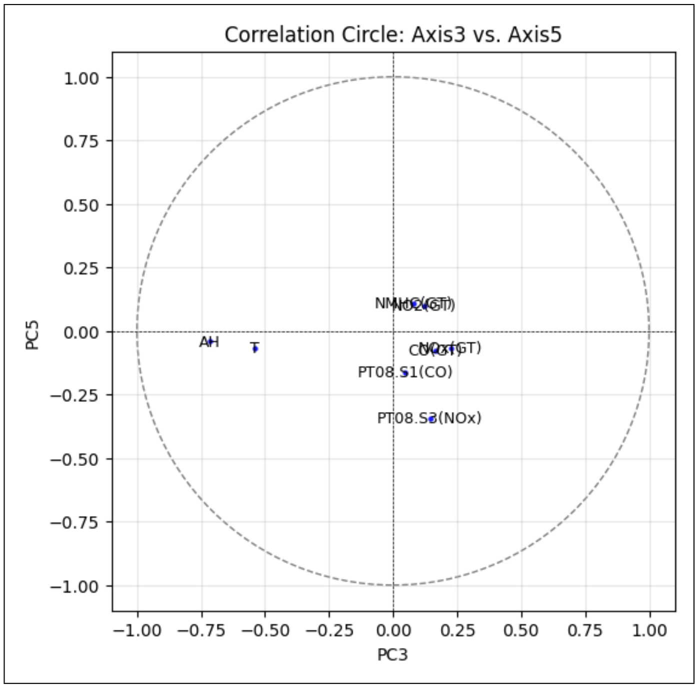


Figure 28: Correlation Circle: PC3 vs. PC5.

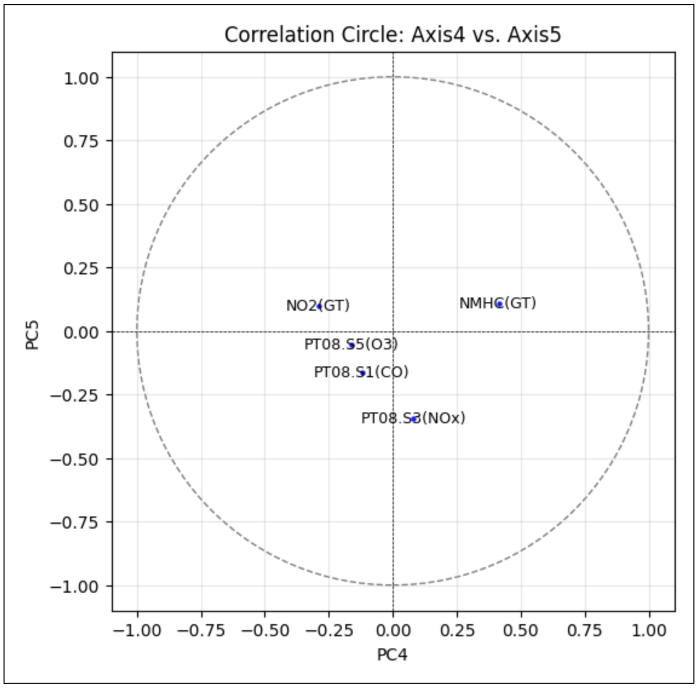
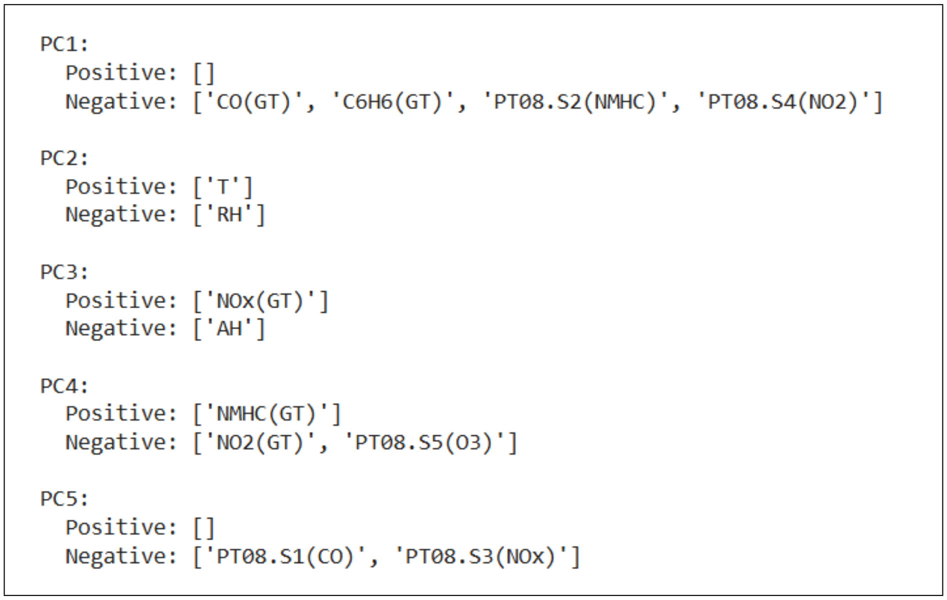


Figure 29: Correlation Circle: PC4 vs. PC5.



```
PC1:
  Positive: []
  Negative: ['CO(GT)', 'C6H6(GT)', 'PT08.S2(NMHC)', 'PT08.S4(NO2)']

PC2:
  Positive: ['T']
  Negative: ['RH']

PC3:
  Positive: ['NOx(GT)']
  Negative: ['AH']

PC4:
  Positive: ['NMHC(GT)']
  Negative: ['NO2(GT)', 'PT08.S5(O3)']

PC5:
  Positive: []
  Negative: ['PT08.S1(CO)', 'PT08.S3(NOx)']
```

Figure 30: analysis of correlation cercles

The correlation circle results reveal the variables contributing most to each principal component. Below is a detailed interpretation:

8.4.1 PC1: Pollutant Levels and Gas Emissions

- **Defining Features:** Strong negative contributions from ‘CO(GT)‘, ‘C6H6(GT)‘, ‘PT08.S2(NMHC)‘, and ‘PT08.S4(NO2)‘.
- **Interpretation:** Represents true pollutant concentrations, particularly carbon monoxide and benzene. It captures the influence of primary pollutants.

8.4.2 PC2: Environmental Temperature and Humidity Effects

- **Defining Features:** Positive contribution from ‘T‘ (Temperature) and negative contribution from ‘RH‘ (Relative Humidity).
- **Interpretation:** Reflects the environmental factors affecting air quality, where temperature positively influences pollutant levels and humidity has an opposing effect.

8.4.3 PC3: Nitrogen Oxides and Atmospheric Dryness.

- **Defining Features:** Positive contribution from ‘NOx(GT)‘ and negative contribution from ‘AH‘ (Absolute Humidity).
- **Interpretation:** Highlights the relationship between nitrogen oxides and atmospheric moisture, where drier conditions coincide with elevated NOx levels.

8.4.4 PC4: Hydrocarbon Emissions and Secondary Pollution Balance.

- **Defining Features:** Positive contribution from ‘NMHC(GT)‘ and negative contributions from ‘NO2(GT)‘ and ‘PT08.S5(O3)‘.
- **Interpretation:** Emphasizes the balance between hydrocarbons (precursor pollutants) and secondary pollutants like ozone.

8.4.5 PC5: NOx Sensor Sensitivity

- **Defining Features:** Negative contribution from ‘PT08.S1(CO)‘ and ‘PT08.S3(NOx)‘.
- **Interpretation:** Reflects sensor-specific responses, particularly for NOx, and identifies potential sensor drift or sensitivity issues.

Summary

The analysis of the saturation matrix and correlation circles reveals the key variables contributing to each principal component. Based on the dataset description and the correlation circle analysis, the principal components can be interpreted as follows:

- **PC1: Pollutant Levels and Gas Emissions.**
  Captures the primary pollutants such as CO, Benzene, and sensor responses for NMHC and NO2.

- **PC2: Environmental Temperature and Humidity Effects.**
  Reflects the influence of temperature (positive) and relative humidity (negative) on air quality.
- **PC3: Nitrogen Oxides and Atmospheric Dryness.**
  Highlights the relationship between nitrogen oxides (NOx) and low humidity (dry conditions).
- **PC4: Hydrocarbon Emissions and Secondary Pollution Balance.**
  Balances hydrocarbons (NMHC) with secondary pollutants like ozone (O_3) and NO2.
- **PC5: NOx Sensor Sensitivity.**
  Represents sensor-specific responses, particularly those targeting NOx.

The principal components provide a clear understanding of the key factors affecting air quality, supporting dimensionality reduction and targeted analysis of the dataset.

9 Quality of Representation for Individuals

The quality of representation measures how well each individual (data point) is represented by the chosen principal components. This is computed using the matrix of new principal components F, where each row represents an individual and each column represents a principal component. The process is described below:

1. **Matrix of New Components**: The matrix F contains the values of the individuals in the new principal component space, focusing on the first 5 components.
2. **Total Sum of Squares**: For each individual, the total sum of squares is calculated as the sum of the squares of their coordinates across the 5 principal components:

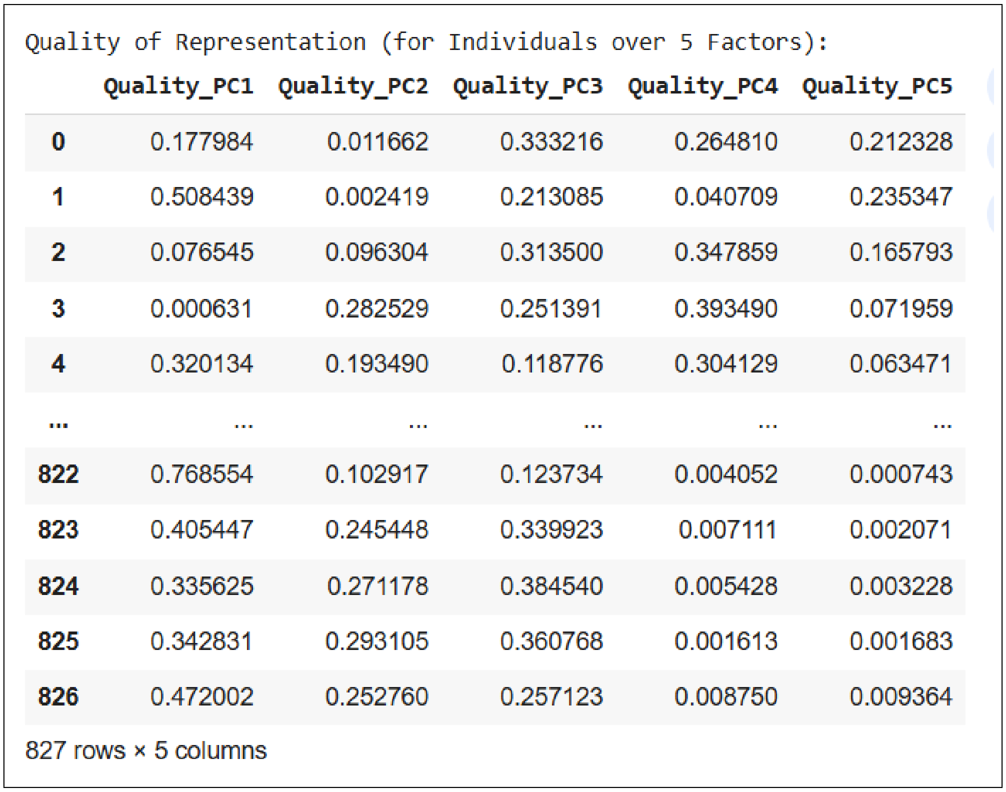
$$\text{Total Sum of Squares} = \sum_{j=1}^{5} F_{ij}^2$$

3. **Quality of Representation**: The quality of representation for an individual on a particular principal component is obtained by dividing the squared coordinate of that component by the total sum of squares:

$$\text{Quality}_{ij} = \frac{F_{ij}^2}{\text{Total Sum of Squares}_i}$$

4. **Output**: The resulting quality matrix contains values between 0 and 1, where higher values indicate better representation of the individual on the corresponding principal component.

The following table provides an example of the quality of representation for individuals over the first 5 principal components:



Quality of Representation (for Individuals over 5 Factors):

	Quality_PC1	Quality_PC2	Quality_PC3	Quality_PC4	Quality_PC5
0	0.177984	0.011662	0.333216	0.264810	0.212328
1	0.508439	0.002419	0.213085	0.040709	0.235347
2	0.076545	0.096304	0.313500	0.347859	0.165793
3	0.000631	0.282529	0.251391	0.393490	0.071959
4	0.320134	0.193490	0.118776	0.304129	0.063471
...	...	...	...	...	...
822	0.768554	0.102917	0.123734	0.004052	0.000743
823	0.405447	0.245448	0.339923	0.007111	0.002071
824	0.335625	0.271178	0.384540	0.005428	0.003228
825	0.342831	0.293105	0.360768	0.001613	0.001683
826	0.472002	0.252760	0.257123	0.008750	0.009364

827 rows × 5 columns

Figure 31: Quality of Representation for Individuals.

10 Graphical Representation of Individuals in the Reduced Space

The principal components not only simplify the dataset but also enable a meaningful visualization of individuals in the reduced space. By projecting the data onto the principal components, we can analyze patterns, clusters, and relationships between individuals.

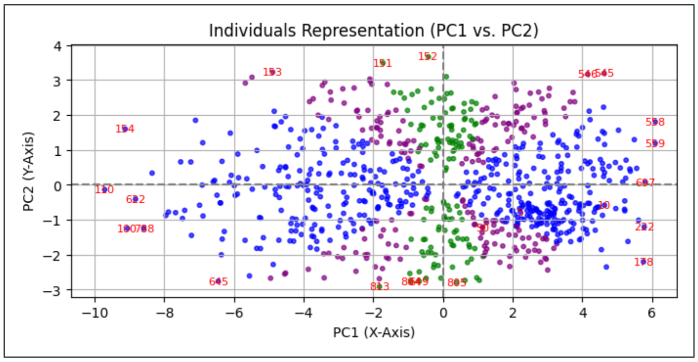


Figure 32: PC1 vs. PC2.

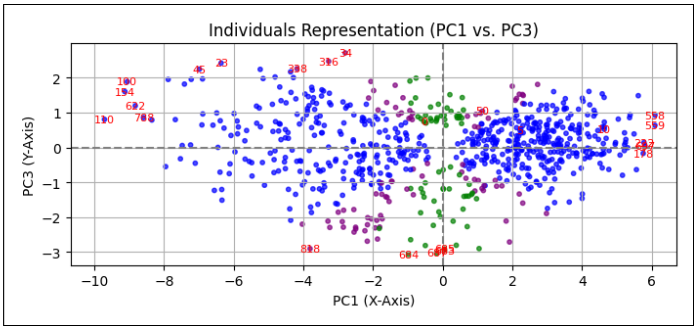


Figure 33: PC1 vs. PC3.

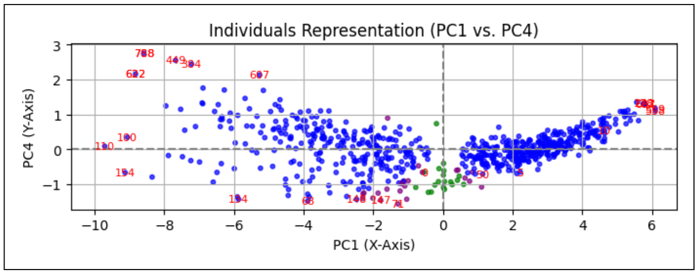


Figure 34: PC1 vs. PC4.

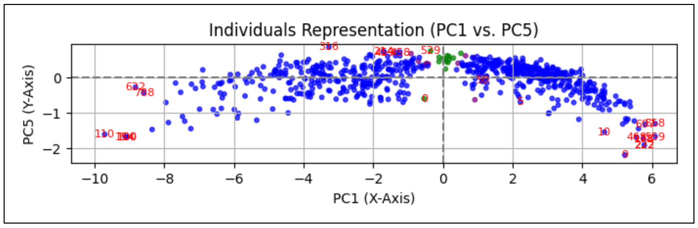


Figure 35: PC1 vs. PC5.

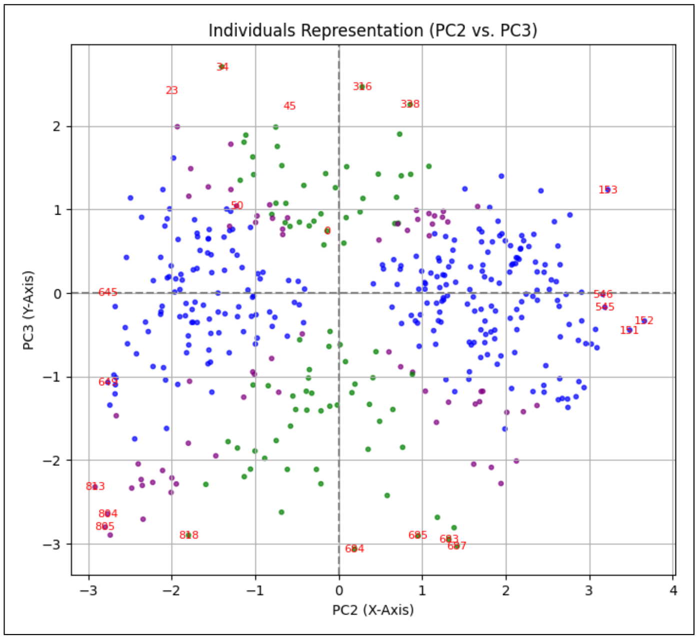


Figure 36: PC2 vs. PC3.

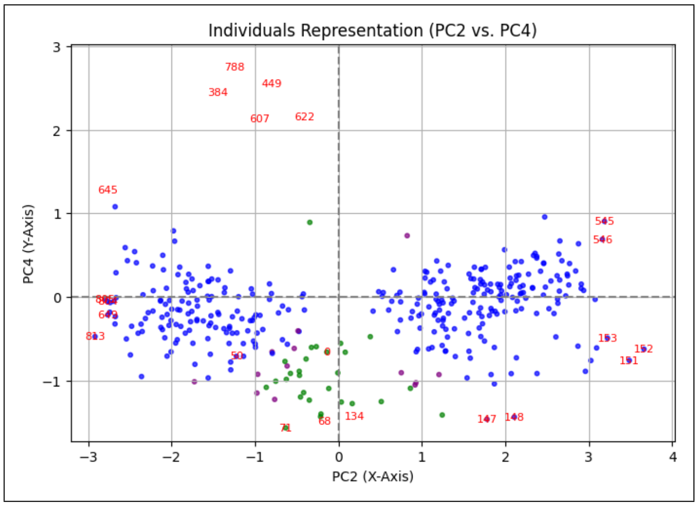


Figure 37: PC2 vs. PC4.

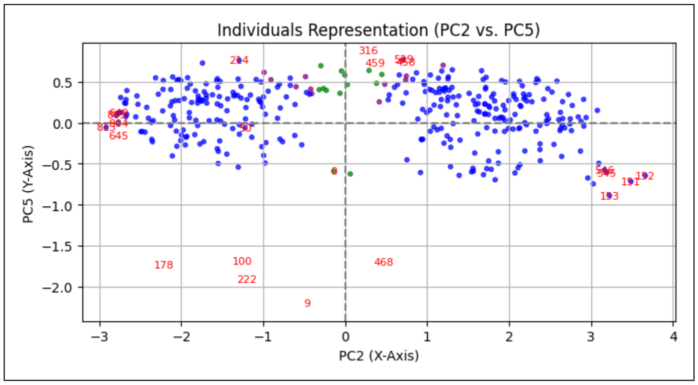


Figure 38: PC2 vs. PC5.

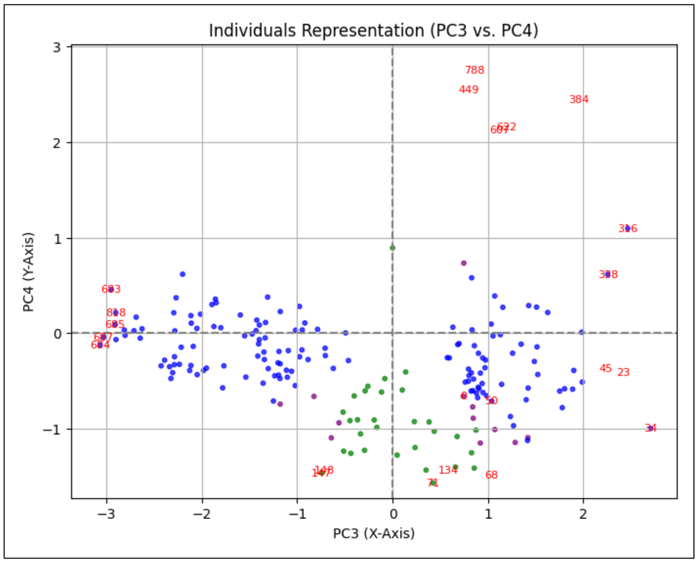


Figure 39: PC3 vs. PC4.

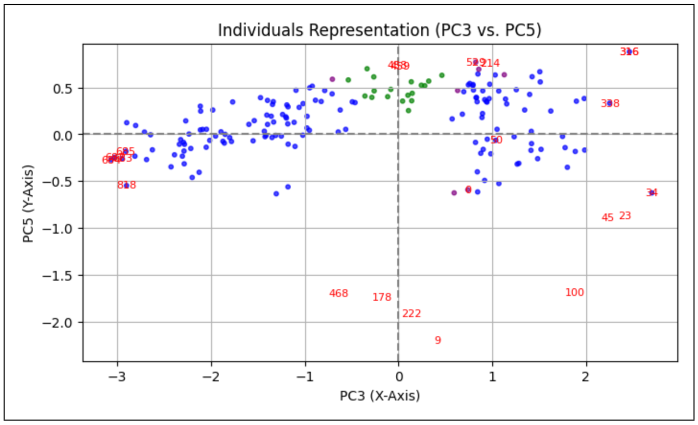


Figure 40: PC3 vs. PC5.

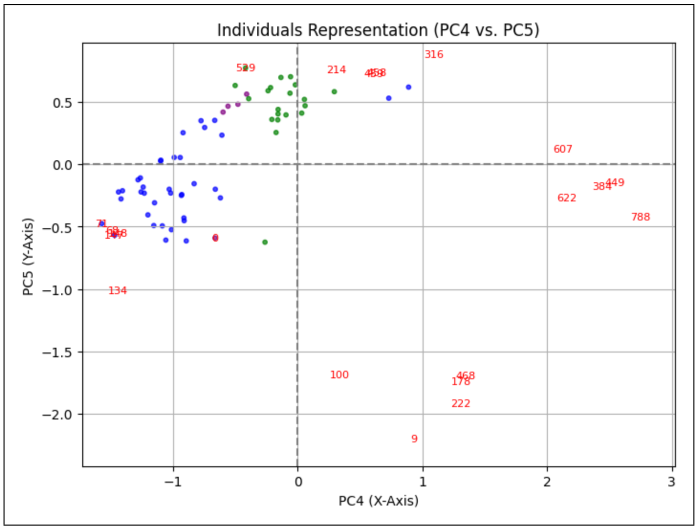


Figure 41: PC4 vs. PC5.

The correlation circle results reveal the variables contributing most to each principal component. Below is a detailed interpretation:

Below are the interpretations of the principal components with examples:

10.1 PC1: Pollutant Levels and Gas Emissions

- **Defining Variables**: Strong negative contributions from ‘CO(GT)‘, ‘C6H6(GT)‘, ‘PT08.S2(NMHC)‘, and ‘PT08.S4(NO2)‘.
- **Axis Interpretation**: This axis represents **air pollutant concentrations**. Individuals with high negative values correspond to periods of elevated pollution, particularly for carbon monoxide and benzene.
- **Examples**:
  - Individual **558**: Strong negative value on PC1, indicating high pollution levels dominated by CO and Benzene.
  - Individual **110**: Positive value on PC1, reflecting a cleaner environment.

10.2 PC2: Environmental Temperature and Humidity Effects

- **Defining Variables**: Positive contribution from ‘T‘ (Temperature) and negative contribution from ‘RH‘ (Relative Humidity).
- **Axis Interpretation**: This axis captures **environmental conditions**, with positive values representing hot and dry conditions and negative values reflecting humid and cooler periods.
- **Examples**:
  - Individual **152**: High positive value on PC2, corresponding to hot and dry conditions.
  - Individual **813**: Strong negative value, indicating high humidity and lower temperatures.

10.3 PC3: Nitrogen Oxides and Atmospheric Dryness

- **Defining Variables**: Positive contribution from ‘NOx(GT)‘ and negative contribution from ‘AH‘ (Absolute Humidity).
- **Axis Interpretation**: This axis highlights the relationship between **nitrogen oxides and humidity levels**. Positive values indicate high NOx levels in dry conditions, while negative values are linked to higher humidity.
- **Examples**:
  - Individual **34**: High positive value on PC3, representing elevated NOx concentrations.
  - Individual **684**: Strong negative value, showing high humidity conditions reducing NOx.

10.4 PC4: Hydrocarbon Emissions and Secondary Pollution Balance

- **Defining Variables**: Positive contribution from ‘NMHC(GT)‘ and negative contributions from ‘NO2(GT)‘ and ‘PT08.S5(O3)‘.
- **Axis Interpretation**: This axis contrasts **hydrocarbon emissions** with secondary pollutants like ozone and NO2, reflecting their interactions in photochemical pollution.
- **Examples**:
  - Individual **788**: Strong positive value on PC4, indicating high hydrocarbon emissions.
  - Individual **71**: Negative value on PC4, reflecting elevated levels of ozone and NO2.

10.5 PC5: NOx Sensor Sensitivity

- **Defining Variables**: Negative contribution from ‘PT08.S3(NOx)‘.
- **Axis Interpretation**: This axis reflects the **NOx sensor responses**, where negative values indicate strong sensor readings for NOx, possibly linked to peak pollution events or sensor drift.
- **Examples**:
  - Individual **316**: Strong negative value, highlighting high NOx sensor activity.
  - Individual **9**: Lower magnitude, representing minimal NOx sensor response.

By analyzing the quality of representation and visualizing individuals in the new reduced space, we can identify key patterns and trends in air quality data while reducing dimensionality and preserving significant information.

11 Applications in Air Quality Management

Targeted Interventions

- Focus on pollutants identified in PC1 and PC4 (e.g., CO, Benzene, NMHC) to reduce emissions.
- Implement policies for industries and vehicles to limit these pollutants.

Sensor Deployment and Calibration

- Use PC5 to assess and improve sensor accuracy, especially for NOx.
- Prioritize sensors sensitive to key pollutants in high-risk areas.

Seasonal and Environmental Adjustments

- Adjust monitoring strategies based on temperature and humidity trends highlighted by PC2.

Predictive Modeling

- Use principal components (PC1, PC2, PC3) as inputs for machine learning models to forecast pollution levels and implement **proactive interventions**.

Public Health Implications

- Link pollution trends (PC1 and PC4) with health outcomes like respiratory diseases.
- Use results to:
  - Implement regulations on fuel standards.
  - Promote urban greenery to counteract pollution.

12 Conclusion

This study highlights the effective application of Principal Component Analysis (PCA) for analyzing a complex air quality dataset, collected over a year in a heavily polluted Italian city. By transforming the high-dimensional dataset into a reduced subspace of five principal components, the analysis preserved 97% of the data's variability while simplifying its interpretation. Key findings include the identification of pollutant contributions (e.g., CO, Benzene, NOx), the impact of environmental factors such as temperature and humidity, and sensor-specific patterns. These results not only enhance our understanding of air quality dynamics but also provide actionable recommendations for pollution management, sensor optimization, and public health policy. This methodology demonstrates PCA's potential as a powerful tool for environmental data analysis and decision-making.