1)

First, use the IQR method to remove outliers and generate a processed CSV file.

IQR Method for Outlier Removal

Steps:

1. Compute the first quartile (Q1) and the third quartile (Q3).
2. Calculate the interquartile range (IQR): IQR=Q3−Q1IQR = Q3 - Q1
3. Set the lower and upper bounds:

   Lower bound = Q1−1.5×IQRQ1 - 1.5 \times IQR

   Upper bound = Q3+1.5×IQRQ3 + 1.5 \times IQR
4. Identify outliers: Any value below the lower bound or above the upper bound is considered an outlier.
5. Replace outliers with NaN to facilitate further missing value processing.

Generate air_quality_no_outliers.csv

Filling Missing Values Using KNN

After removing outliers, we use KNN to fill missing values. KNN estimates missing values based on similar data points, providing better results than simple imputation methods.

KNN Imputation Steps:

1. Load the data after outlier removal (air_quality_no_outliers.csv).
2. Apply KNNImputer for imputation:

   Choose an appropriate n_neighbors (e.g., 5).

   Use Euclidean distance to find the 5 nearest neighbors and compute the mean to fill in missing values.
3. Save the imputed data.

Generate air_quality_filled.csv

Z-Score Standardization

After imputation, we apply Z-score normalization:

$$X' = (X - \mu)/\sigma$$

where:

- X is the original value,
- μ is the mean,
- σ is the standard deviation.

Effects of Standardization:

The processed data will have a mean of 0 and a standard deviation of 1.

Applicable Scenarios:

Suitable for data with different units, such as PM2.5 (μg/m³), temperature (°C), and humidity (%).

Works well for normally distributed or approximately normal data.

Beneficial for most machine learning algorithms (e.g., PCA, KNN, linear models).

Generate air_quality_standardized.csv

2)

PCA Component Contribution Rate and Mahalanobis Distance

When the principal component contribution rate is high in PCA, the Mahalanobis Distance becomes more reasonable.

In data analysis, Mahalanobis Distance is considered more reasonable than Euclidean Distance when variables are correlated, and PCA (Principal Component Analysis) provides a way to quantify the main directions (principal components) of the data. The connection between these two concepts is as follows:

1. **PCA Reflects the Main Directions of the Data** PCA is a dimensionality reduction method, and its main functions are:
    - Analyzing the main directions of data variation through the covariance matrix of features.
    - Finding the Principal Components (PCs) that maximize the variance of the data in these directions.
    - Principal components with high contribution rates indicate that most of the

information in the data is concentrated in a few directions.

If PCA reveals that:

- The first principal component has a high contribution rate (e.g., >70%), it indicates that the data is mainly distributed in one direction.
- This means the variation in the data is uneven across different dimensions, i.e., certain directions have much more information than others.

2. **The Problem with Euclidean Distance: Ignoring the Correlation Between Variables** The calculation of Euclidean Distance is as follows:

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- This assumes that all variables are independent, i.e., there is no correlation between different air quality metrics.
- In cases where PCA has a high contribution rate, this assumption may not hold.
- For example, PM2.5 and PM10 are likely highly correlated, and their changes are often linked. Using Euclidean Distance in this case may overestimate the difference.

3. **Advantages of Mahalanobis Distance: Considering the Covariance of the Data** Mahalanobis Distance is defined as:

$$d_M(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$$

where:

- S is the covariance matrix of the data.
- S^{-1} reflects the correlation between variables.

Why is Mahalanobis Distance more reasonable when PCA's contribution rate is high?

- Mahalanobis Distance adjusts the distance calculation based on the data's covariance, meaning that if two variables are correlated, their changes are not counted twice.
- When PCA reveals a high contribution rate from principal components, the

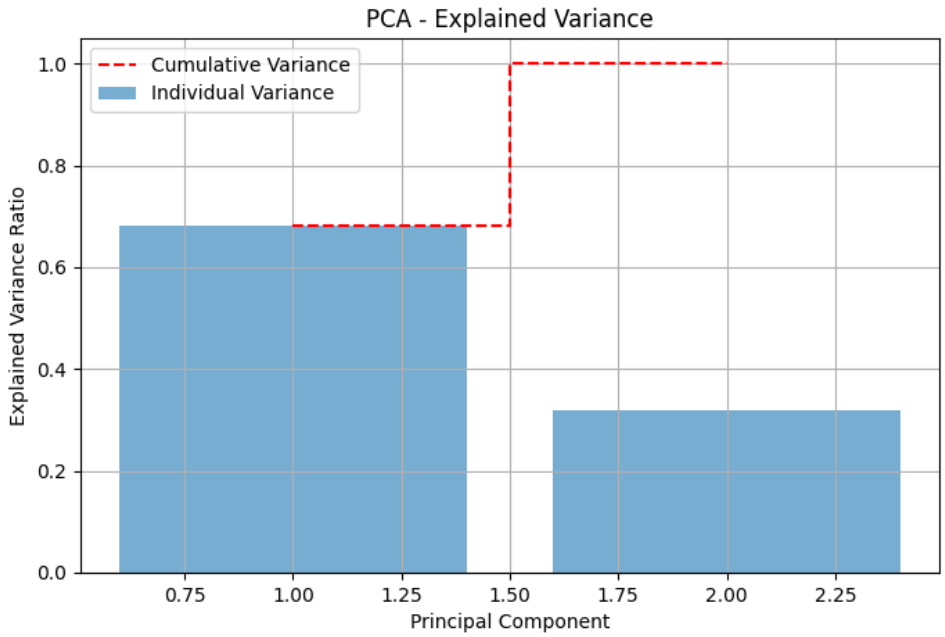
data mainly varies along a few directions, and Mahalanobis Distance automatically adjusts the measurement scale, ignoring less important directions.

- For example:
  - If PCA finds that PM2.5 and PM10 have a 90% correlation, their changes essentially affect one principal component.
  - Euclidean Distance would mistakenly treat PM2.5 and PM10 as two independent dimensions, leading to an overestimation of the difference.
  - Mahalanobis Distance would correctly reflect their joint changes, providing a more reasonable similarity measure.

4. **Conclusion**
   - When PCA's principal component contribution rate is high, it means that data variation mainly occurs along specific directions rather than being evenly distributed across all variables.
   - Euclidean Distance assumes variable independence and may overestimate similarity differences.
   - Mahalanobis Distance adjusts for the influence of different directions using covariance, making it more reasonable.

Therefore, in the case of air quality datasets, if PCA identifies high correlation between certain pollutants (e.g., PM2.5 and PM10), Mahalanobis Distance will provide a more accurate similarity measure.

In **2.py**, we perform PCA analysis on the dataset after z-score normalization with **component=2**, and the analysis results are as follows:




```
主成分 1: 贡献率 = 0.6811
主成分 2: 贡献率 = 0.3189

主成分载荷矩阵（每个变量的贡献）：
        PM2.5      PM10       SO2       NO2        CO        O3      Temp  \
PC1 -0.045952 -0.213120  0.283643 -0.496789 -0.202837 -0.408426 -0.259925
PC2  0.125925 -0.226734  0.276055  0.306212  0.044616  0.466361 -0.393270

     Humidity  WindSpeed  Pressure
PC1  0.566805  -0.167393 -0.020296
PC2  0.372516   0.498160 -0.041069
```

The analysis shows that the contribution rate of Principal Component 1 (68%) significantly exceeds that of Principal Component 2 (31%). This indicates that the direction of data variation is mainly along a specific direction, rather than being evenly distributed across all variables. In this case, using Mahalanobis distance will provide a more accurate similarity measure.

3)

Before applying PCA for dimensionality reduction, data preprocessing (such as outlier removal, missing value imputation, and standardization) directly impacts clustering quality. We can compare PCA results under different preprocessing strategies and analyze their effects on KMeans clustering.

Impact of Data Preprocessing on PCA and Clustering

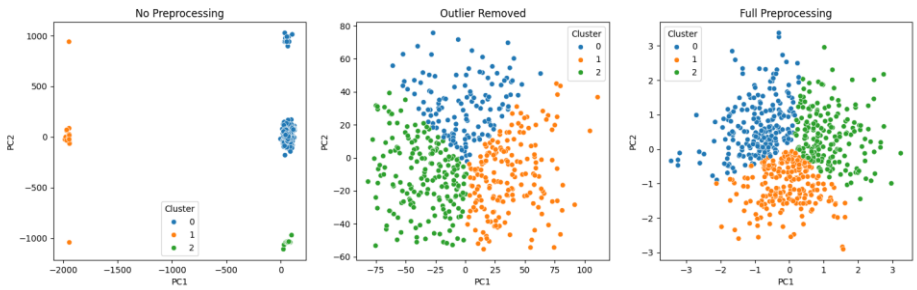
PCA fundamentally finds the directions of maximum variance through a linear transformation, meaning data quality and scale directly affect dimensionality reduction outcomes. Below are key preprocessing steps and their impacts:

Preprocessing Step	Impact on PCA	Impact on Clustering
Outlier Removal	Prevents extreme values from distorting principal component directions	Improves clustering accuracy by avoiding misclassification due to outliers
Missing Value Imputation (Mean vs. KNN)	Affects covariance matrix computation	Influences cluster center positions
Standardization (Z-score or Min-Max Normalization)	Ensures all variables have the same scale, preventing certain features from dominating PCA	Makes Euclidean distance calculations more reasonable, improving KMeans stability

Comparison of Three Preprocessing Strategies

1. **No Preprocessing** (Raw data directly processed with PCA + KMeans)
2. **Outlier Removal Only** (IQR method for outlier removal before PCA + KMeans)
3. **Complete Preprocessing** (Outlier removal + KNN missing value imputation + Standardization)

We then visualize the clustering results after PCA dimensionality reduction and analyze the impact of different preprocessing strategies.



Result Analysis

- **No Preprocessing (Left Image):**
    - Since the data is not standardized, some features dominate the PCA calculation, leading to unstable clustering results.
    - Improper handling of missing values may result in data loss, affecting dimensionality reduction.
- **Outlier Removal Only (Middle Image):**
    - After removing outliers, the clustering results are clearer than the raw data, avoiding interference from extreme values.
    - However, the feature scale still affects clustering, and some clusters remain unclear.
- **Complete Preprocessing (Right Image):**
    - After **outlier removal + KNN imputation + standardization**, the data is more evenly distributed, and PCA performs better.
    - The clustering results are more stable, with clearer boundaries between the three clusters.