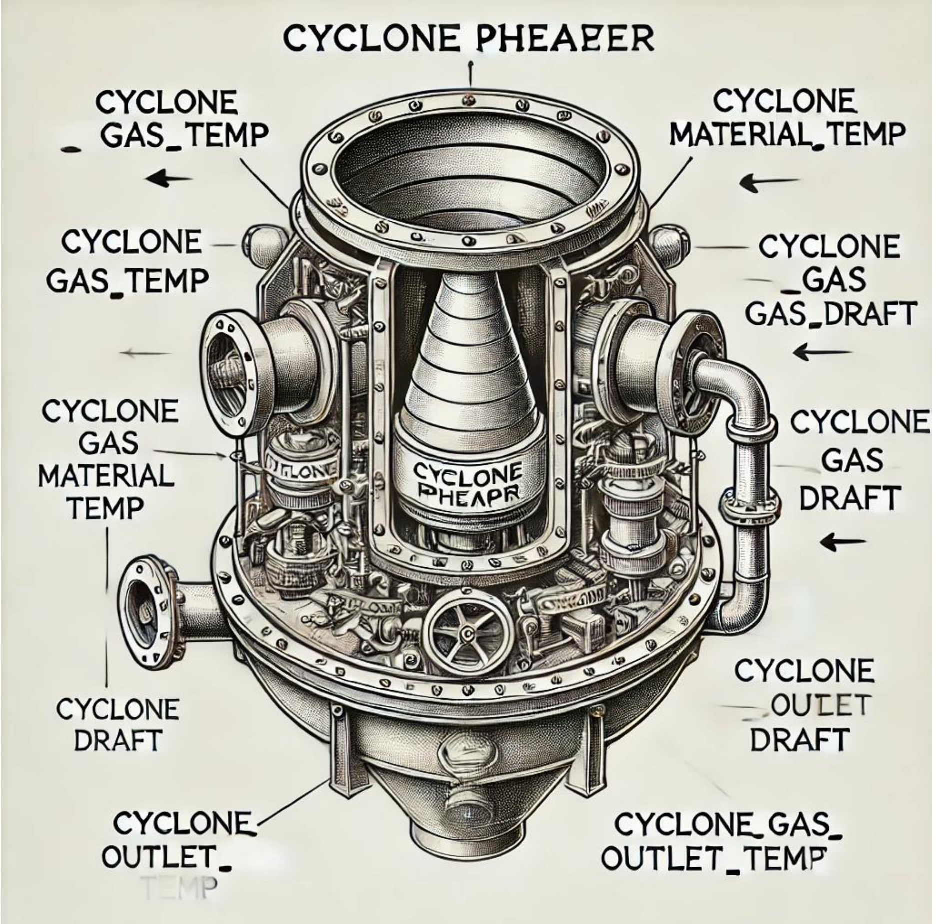
Cyclone Preheater Anomaly Detection

Flow Chart

- **Importing Dataset**
- **Cleaning the Data**
  - Convert string into float
  - Handle Missing Values → Replace missing Values by median
- **Exploratory Data Analysis**
  - Find Correlation Matrix
- **Dimensionality Reduction**
  - Reduce 6 dimensions to 2
- **Model Selection**
  - Isolation Forest → Anomaly Detection



Data Preparation

Data Type Transformation:

•Converted object-type columns to numerical float values where applicable.

•**Reason**: Ensures compatibility with numerical analysis techniques and machine learning algorithms.

```python
for col in columns_to_convert:
    data[col] = pd.to_numeric(data[col], errors='coerce')
print(data.dtypes)
```

```
time                          datetime64[ns]
Cyclone_Inlet_Gas_Temp               float64
Cyclone_Material_Temp                float64
Cyclone_Outlet_Gas_draft             float64
Cyclone_cone_draft                   float64
Cyclone_Gas_Outlet_Temp              float64
Cyclone_Inlet_Draft                  float64
dtype: object
```

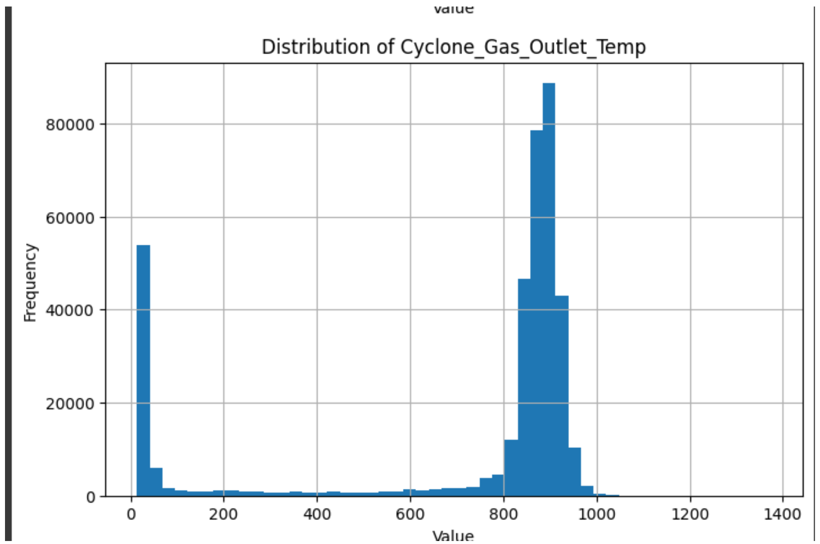
Handling Missing Values:

•**Visualization**:

- Plotted frequency value distributions for each column to understand the spread and nature of missing data.

•**Imputation**:

- Replaced missing values with the **median** of the respective column.
- **Reason**: Median is robust to outliers and preserves the central tendency of the data better than mean imputation, especially in skewed distributions.




```python
for col in columns_to_check:
    data[col] = data[col].fillna(data[col].median())
```

Exploratory Data Analysis (EDA)

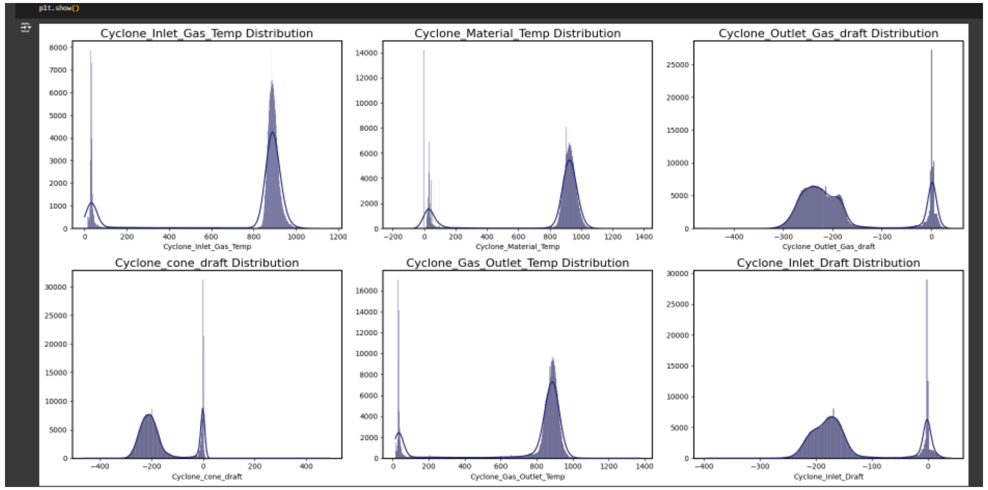
Distribution Analysis:
•Used distplot to visualize the distribution of each numerical variable.
•**Objective**: Identify skewness, modality, and potential outliers in the data.

Subplots for Multi-Variable Insights:
•Created subplots to compare distributions and trends across variables simultaneously.
•**Reason**: Simplifies comparison and highlights inter-variable differences.

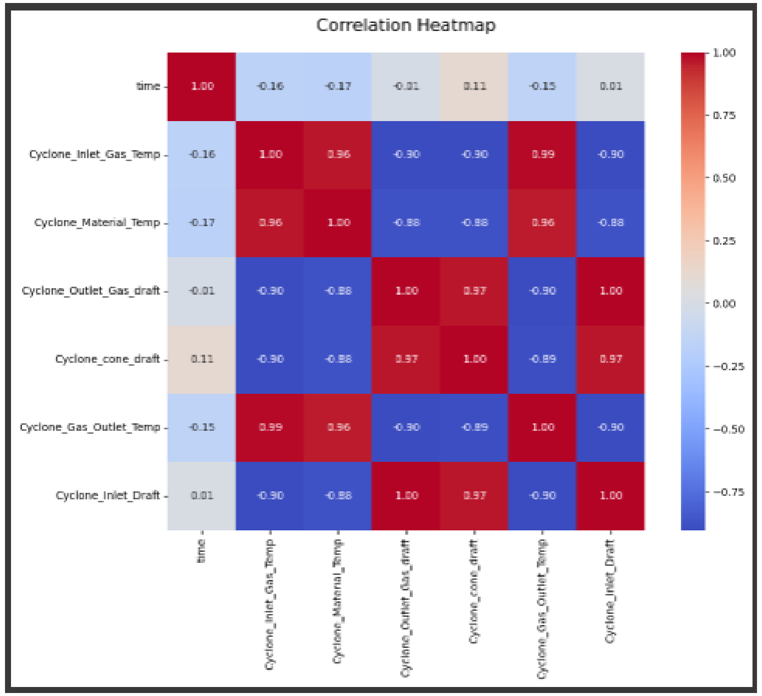
Correlation Matrix:
•Visualized pairwise correlations using a heatmap.
•**Goal**: Identify strongly correlated variables to understand relationships and potential multicollinearity.

Pair Plot:
•Used pair plots to examine scatter plots between variables.
•**Benefit**: Highlights trends, clusters, and linear relationships between variables.

Box Plots:
•Created box plots to identify outliers and examine value ranges for each variable.
•**Utility**: Helps in detecting and visualizing anomalies in the data.



subplot



Correlation Matrix

Dimension Reduction

Purpose:

•Simplify the dataset by reducing it to its most informative components while preserving variance.

Techniques Used:

•**Principal Component Analysis (PCA)**: Reduced the data to 2 principal components for better visualization and analysis.

Process:

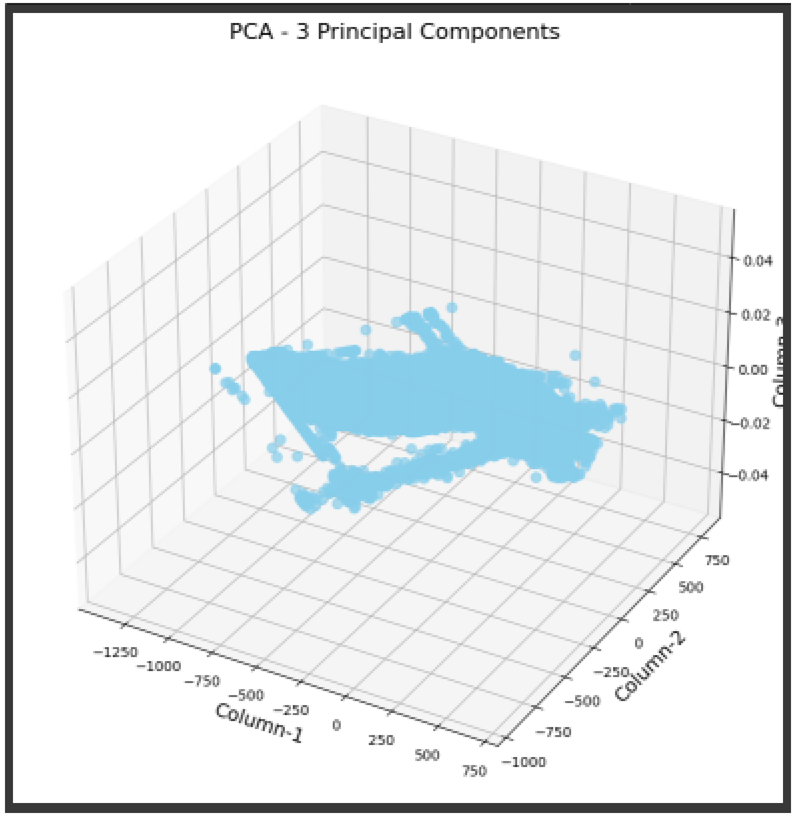
•Extracted key variables from the dataset.
•Applied **PCA** to capture the dominant patterns in the data.
•Standardized the PCA components using **StandardScaler** to ensure consistency and comparability.

Visualization:

•Created a 3D scatter plot to represent relationships and clustering between the PCA components.

Outcome:

•Dimensionality reduction helped uncover hidden patterns and prepare data for anomaly detection.



Model Selection and Visualization

Algorithm: **Isolation Forest**

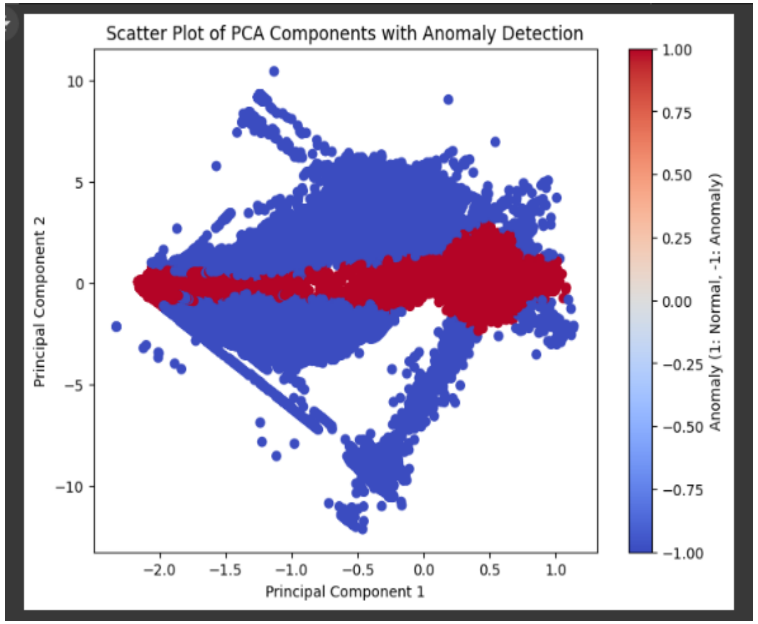
Reason for Selection:

1. **Efficiency**: Handles large datasets like ours (370,000 records) effectively.
2. **Robustness**: Does not assume any specific data distribution.
3. **Interpretability**: Flags anomalies based on the isolation principle, making results easier to understand.
4. **Scalability**: Well-suited for high-dimensional data and capable of detecting global and local anomalies

Out of a total of **358,833** data points, **18,886** anomalies were detected. This highlights the significance of anomaly detection within the dataset.

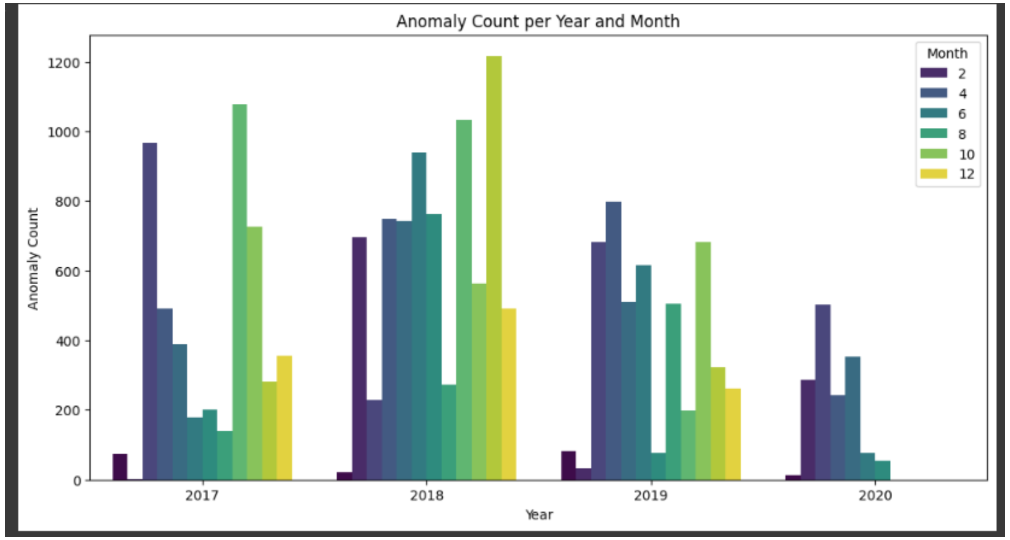
```
Dates of Anomalies:
5297      2017-01-19 09:25:00
5298      2017-01-19 09:30:00
5299      2017-01-19 09:35:00
5300      2017-01-19 09:40:00
5301      2017-01-19 09:45:00
                  ...
375407    2020-07-30 11:40:00
375408    2020-07-30 11:45:00
375409    2020-07-30 11:50:00
375410    2020-07-30 11:55:00
375411    2020-07-30 12:00:00
```

Dates of Anomaly



Insights: Anomaly Trends Analysis

•**Peak Periods**: Highest anomalies occurred in **2017, 2018, and 2019**, with notable peaks in **August (Month 8)** and **December (Month 12)**. **December 2018** had the highest count, exceeding **1200**.
•**Identification**: Abnormal periods were identified by significant peaks in the bar graph, particularly in late-year months.
•**Trends**: Anomaly counts rose from **2017 to 2018**, declined in **2019**, and sharply dropped by **2020**. Recurring high counts in **December** suggest seasonal patterns



Anomaly Count Vs Year and Month

Thank you