**Preprocessing Strategy Based on EDA Results**

**1. Missing Values Strategy**

* **Columns with Missing Values:**
  + PM2.5 (1330 missing)
  + PM10 (1136 missing)
  + NO2 (1374 missing)
  + CO (1082 missing)
  + O3 (1071 missing)
* **Should we drop rows/columns?**
  + No, because missing values are present in key features. Dropping rows will lead to significant data loss.
* **Filling missing values strategy:**
  + Since the distributions are right-skewed (based on histograms), using the **median** for imputation is a robust choice.
  + We can also use **KNN Imputer** (K-Nearest Neighbors) to fill missing values based on similar data points.

**2. Outliers Detection**

* **Extreme Outliers Observed:**
  + PM2.5, PM10, NO2, CO, O3 have extreme outliers.
  + Maximum values are significantly higher than the 75th percentile.
* **Should we remove or cap them?**
  + Instead of removing, we should **cap outliers** using the **IQR (Interquartile Range) method**:
    - Upper limit = Q3 + 1.5 \* IQR
    - Lower limit = Q1 - 1.5 \* IQR
  + This will help retain useful data while removing extreme variations.

**3. Feature Transformations**

* **Normalization/Standardization:**
  + Since PM2.5, PM10, NO2, CO, and O3 have skewed distributions, we should use **MinMax Scaling (0-1 scaling)** to normalize them.
  + Alternatively, we can apply **log transformation** to make the distributions more normal.
* **Categorical Data Processing:**
  + StateCode: Only contains "TN" → Can be dropped as it doesn't provide variance.
  + StationId: Already numeric but should be converted to categorical if used in models.

**4. Date Processing**

* Extract meaningful features from the Date column:
  + Year, Month, Day, Day of Week
  + Is\_Weekend (binary feature)

**Final Steps for Preprocessing**

1. **Handle Missing Values**
   * KNN Imputer to fill missing values.
2. **Remove/Capping Outliers**
   * Apply IQR method to cap outliers.
3. **Normalize Data**
   * Use **MinMaxScaler** or **log transformation** on skewed features.
4. **Feature Engineering**
   * Extract Year, Month, Day, Day of Week, Is\_Weekend from Date.
5. **Drop Unnecessary Columns**
   * Remove StateCode as it has only one unique value.

Let me know if you want any modifications before I generate the preprocessing script! 🚀