# EXPERIMENT-3

```
# eda_data_cleaning.py


import pandas as pd
from sklearn.preprocessing import StandardScaler, MinMaxScaler


# Load your dataset (replace with your actual file)
df = pd.read_csv(r"C:\Users\REC\Downloads\final_dataset.csv")  # <-- Change this


print("\nMissing values:\n", df.isnull().sum())


# ---------------------------------------
# 1. Handling Missing Values
# ---------------------------------------


# Fill missing values (customize based on your data)
df.fillna(method='ffill', inplace=True)       # Forward fill
df.fillna(method='bfill', inplace=True)       # Backward fill


# Optional: Drop remaining NaNs if needed
df.dropna(inplace=True)


# ---------------------------------------
# 2. Remove Duplicates & Unnecessary Columns
# ---------------------------------------


# Remove duplicate rows
duplicates = df.duplicated().sum()
if duplicates > 0:
```

```
    df.drop_duplicates(inplace=True)

# Drop unwanted columns (edit these column names)
columns_to_drop = ['Unnamed: 0', 'ID', 'Notes']  # Example columns
df.drop(columns=[col for col in columns_to_drop if col in df.columns], inplace=True)

# ---------------------------------------
# 3. Data Type Conversion & Consistency
# ---------------------------------------

# Convert to datetime
if 'Date' in df.columns:
    df['Date'] = pd.to_datetime(df['Date'], errors='coerce')

# Convert numerical columns to correct types (customize)
for col in df.select_dtypes(include='object').columns:
    try:
        df[col] = pd.to_numeric(df[col])
    except:
        pass  # Skip if not convertible

# Strip spaces and lowercase for categorical text columns
df.columns = df.columns.str.strip()
for col in df.select_dtypes(include='object').columns:
    df[col] = df[col].str.strip().str.lower()

# ---------------------------------------
# 4. Normalization (Standardization & Min-Max)
# ---------------------------------------
```

```
# Choose numerical columns to normalize
numeric_cols = df.select_dtypes(include='number').columns.tolist()

# Standardization
scaler_std = StandardScaler()
df[numeric_cols] = scaler_std.fit_transform(df[numeric_cols])

# OR Min-Max Scaling
# scaler_minmax = MinMaxScaler()
# df[numeric_cols] = scaler_minmax.fit_transform(df[numeric_cols])

# ----------------------------------------
# Final Check
# ----------------------------------------

print("\n✅ Cleaned Data Info:")
print(df.info())

print("\n📊 Statistical Summary:")
print(df.describe())

print("\n🔍 Preview:")
print(df.head())

# Save cleaned data if needed
df.to_csv('cleaned_dataset.csv', index=False)
print("\n💾 Cleaned data saved to 'cleaned_dataset.csv'")
```

```
Missing values:
 Date                0
Month               0
Year                0
Holidays_Count      0
Days                0
PM2.5               0
PM10                0
NO2                 0
SO2                 0
CO                  0
Ozone               0
AQI                 0
dtype: int64

✅ Cleaned Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1461 entries, 0 to 1460
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Date             1461 non-null   datetime64[ns]
 1   Month            1461 non-null   float64
 2   Year             1461 non-null   float64
 3   Holidays_Count   1461 non-null   float64
 4   Days             1461 non-null   float64
 5   PM2.5            1461 non-null   float64
 6   PM10             1461 non-null   float64
 7   NO2              1461 non-null   float64
 8   SO2              1461 non-null   float64
 9   CO               1461 non-null   float64
 10  Ozone            1461 non-null   float64
 11  AQI              1461 non-null   float64
dtypes: datetime64[ns](1), float64(11)
memory usage: 137.1 KB
None
```

📊 Statistical Summary:

| | Date | Month | Year | \ |
|---|---|---|---|---|
| count | 1461 | 1461.000000 | 1.461000e+03 | |
| mean | 1970-01-01 00:00:00.000000015 | 0.000000 | 7.143362e-14 | |
| min | 1970-01-01 00:00:00.000000001 | -1.601451 | -1.342192e+00 | |
| 25% | 1970-01-01 00:00:00.000000008 | -0.731559 | -4.480094e-01 | |
| 50% | 1970-01-01 00:00:00.000000016 | 0.138333 | 4.461733e-01 | |
| 75% | 1970-01-01 00:00:00.000000023 | 1.008226 | 1.340356e+00 | |
| max | 1970-01-01 00:00:00.000000031 | 1.588154 | 1.340356e+00 | |
| std | NaN | 1.000342 | 1.000342e+00 | |

| | Holidays_Count | Days | PM2.5 | PM10 | NO2 | \ |
|---|---|---|---|---|---|---|
| count | 1.461000e+03 | 1.461000e+03 | 1.461000e+03 | 1.461000e+03 | 1.461000e+03 | |
| mean | 3.404380e-17 | 8.936497e-17 | 2.723504e-16 | 2.456017e-16 | -1.021314e-16 | |
| min | -4.836866e-01 | -1.499445e+00 | -1.266642e+00 | -1.613336e+00 | -9.946512e-01 | |
| 25% | -4.836866e-01 | -9.997437e-01 | -6.910130e-01 | -7.977291e-01 | -5.652676e-01 | |
| 50% | -4.836866e-01 | -3.420266e-04 | -2.612812e-01 | -1.425050e-01 | -1.901250e-01 | |
| 75% | -4.836866e-01 | 9.990597e-01 | 3.870863e-01 | 6.153083e-01 | 2.222196e-01 | |
| max | 2.067455e+00 | 1.498761e+00 | 1.269406e+01 | 6.048431e+00 | 1.126834e+01 | |
| std | 1.000342e+00 | 1.000342e+00 | 1.000342e+00 | 1.000342e+00 | 1.000342e+00 | |

| | SO2 | CO | Ozone | AQI |
|---|---|---|---|---|
| count | 1.461000e+03 | 1.461000e+03 | 1.461000e+03 | 1.461000e+03 |
| mean | 1.556288e-16 | -2.431700e-16 | 2.334432e-16 | 2.918040e-17 |
| min | -1.142516e+00 | -1.242946e+00 | -1.775633e+00 | -1.700109e+00 |
| 25% | -7.494815e-01 | -6.838246e-01 | -6.460308e-01 | -8.742313e-01 |
| 50% | -2.826776e-01 | -2.891507e-01 | -2.042190e-01 | -1.225900e-01 |
| 75% | 3.939461e-01 | 3.521945e-01 | 4.957123e-01 | 7.589645e-01 |
| max | 5.641257e+00 | 6.042077e+00 | 4.198064e+00 | 2.763341e+00 |
| std | 1.000342e+00 | 1.000342e+00 | 1.000342e+00 | 1.000342e+00 |

🔍 Preview:

| | Date | Month | Year | Holidays_Count | Days | \ |
|---|---|---|---|---|---|---|
| 0 | 1970-01-01 00:00:00.000000001 | -1.601451 | -1.342192 | -0.483687 | 0.499359 | |
| 1 | 1970-01-01 00:00:00.000000002 | -1.601451 | -1.342192 | -0.483687 | 0.999060 | |
| 2 | 1970-01-01 00:00:00.000000003 | -1.601451 | -1.342192 | 2.067455 | 1.498761 | |
| 3 | 1970-01-01 00:00:00.000000004 | -1.601451 | -1.342192 | -0.483687 | -1.499445 | |
| 4 | 1970-01-01 00:00:00.000000005 | -1.601451 | -1.342192 | -0.483687 | -0.999744 | |

| | PM2.5 | PM10 | NO2 | SO2 | CO | Ozone | AQI |
|---|---|---|---|---|---|---|---|
| 0 | 4.440081 | 1.734582 | 3.505073 | -0.432635 | 2.868241 | 0.361638 | 2.410719 |
| 1 | 4.373624 | 2.659354 | 0.444863 | -0.902463 | 2.588680 | -1.050893 | 2.596310 |
| 2 | 1.874953 | 0.161085 | 3.798712 | -0.554778 | 0.615310 | 0.419702 | 0.564095 |
| 3 | -0.017096 | -0.666437 | 3.316792 | -0.585616 | -0.026035 | 0.678349 | 0.044441 |
| 4 | -0.512586 | -1.258606 | 2.427354 | -0.629152 | -0.634490 | 0.661986 | -0.493771 |