Task 1: Data Cleaning & Preprocessing

Data Cleaning & Preprocessing — House Prices Dataset Code :

```
# Step 1️⃣ — Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
```

```
# Step 2️⃣ — Import the Dataset and Explore Basic Info
df = pd.read_csv("train.csv")   # Replace with your dataset path
print("✅ Dataset Loaded Successfully!")
print("\nShape of Data:", df.shape)

print("\n--- Basic Info ---")
df.info()

print("\n--- Missing Values Count ---")
print(df.isnull().sum().sort_values(ascending=False).head(10))

print("\n--- Data Preview ---")
print(df.head())
```

```
# Step 3️⃣ — Handle Missing Values (Mean/Median/Mode/Imputation)
# Fill numeric columns with median
for col in df.select_dtypes(include=[np.number]).columns:
    df[col].fillna(df[col].median(), inplace=True)

# Fill categorical columns with mode
for col in df.select_dtypes(include=['object']).columns:
    df[col].fillna(df[col].mode()[0], inplace=True)
```

```
print("\n✅ Missing Values Handled!")

# Step 4️⃣ — Convert Categorical Features into Numerical (Encoding)
df = pd.get_dummies(df, drop_first=True)
print("\n✅ Categorical Columns Encoded Successfully!")
print("New Shape after Encoding:", df.shape)

# Step 5️⃣ — Normalize / Standardize Numerical Features
scaler = StandardScaler()
num_cols = df.select_dtypes(include=[np.number]).columns
```

```
df[num_cols] = scaler.fit_transform(df[num_cols])

print("\n✅ Numerical Features Standardized!")

# Step 6️⃣ — Visualize Outliers using Boxplots
plt.figure(figsize=(10,5))
sns.boxplot(x=df['SalePrice'])
plt.title("Boxplot — SalePrice (with Outliers)")
plt.show()

# Step 7️⃣ — Remove Outliers using IQR
Q1 = df['SalePrice'].quantile(0.25)
Q3 = df['SalePrice'].quantile(0.75)
```

```
IQR = Q3 - Q1
df = df[(df['SalePrice'] >= Q1 - 1.5 * IQR) &
(df['SalePrice'] <= Q3 + 1.5 * IQR)]

print("\n✅ Outliers Removed!")
print("Final Data Shape:", df.shape)
print("Remaining Missing Values:",
df.isnull().sum().sum())

print("\n🎯 Data Cleaning & Preprocessing
Completed Successfully!")
```

❖ Output :

```
✅ Dataset Loaded Successfully!
Shape of Data: (1460, 81)


--- Basic Info ---
<class ‘pandas.core.frame.DataFrame’>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
 #   Column          Non-Null Count  Dtype
---  --------------  --------------  -----
 0   Id              1460 non-null   int64
 1   MSSubClass      1460 non-null   int64
 2   MSZoning        1460 non-null   object
 3   LotFrontage     1201 non-null   float64
 ...
 80  SalePrice      1460 non-null   int64
Dtypes: float64(3), int64(35), object(43)
Memory usage: 924.0+ KB

--- Missing Values Count ---
PoolQC        1453
MiscFeature    1406
Alley         1369
Fence         1179
FireplaceQu     690
LotFrontage     259
GarageCond       81
GarageType       81
GarageYrBlt      81
```

```
GarageFinish     81
Dtype: int64

--- Data Preview ---
  Id  MSSubClass MSZoning  LotFrontage
LotArea  Street Alley  LotShape  ... SalePrice
0   1        60      RL       65.0    8450   Pave   NaN
Reg   ...   208500
1   2        20      RL       80.0    9600   Pave   NaN
Reg   ...   181500
2   3        60      RL       68.0   11250   Pave
NaN     IR1   ...   223500
3   4        70      RL       60.0    9550   Pave   NaN
IR1   ...   140000
4   5        60      RL       84.0   14260   Pave
NaN     IR1   ...   250000
```

✅ Missing Values Handled!

✅ Categorical Columns Encoded Successfully!

New Shape after Encoding: (1460, 240)

✅ Numerical Features Standardized!

📊 Boxplot — SalePrice (with Outliers)

(A figure appears showing SalePrice distribution and outliers.)

✅ Outliers Removed!

Final Data Shape: (1379, 240)

```
Remaining Missing Values: 0

🎯 Data Cleaning & Preprocessing
Completed Successfully!
```

❖ **Explanation for code and output**:

This Python program performs complete data cleaning and preprocessing on the House Prices dataset using Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. First, the dataset is imported using pandas.read_csv(), and its structure is explored to view the number of rows, columns, data types, and missing values. Next, missing values are handled by filling numeric columns with their median values and categorical columns with their mode values, ensuring no empty data remains. Then, all categorical (text) columns are converted into numeric form using one-hot encoding so they can be used in machine learning models. After encoding, the numerical columns are standardized using StandardScaler that all features are on a similar scale. Outliers in the SalePrice column are visualized using a boxplot created with Seaborn, and extreme values are removed using the Interquartile Range (IQR) method. Finally, the cleaned dataset is checked again to confirm there are no missing values and that the data is consistent and ready for analysis. The output shows that the dataset initially had 1460 rows and 81 columns, which became 1379 rows and about 240 columns after encoding and cleaning. The process ends with a message confirming that data cleaning and preprocessing have been completed successfully