TASK 1

Data Cleaning and Preprocessing

Code for the project

```
import pandas as pd
import numpy as np

# Create a sample DataFrame with dirty data
data = {
    'TransactionID': [1, 2, 3, 4, 5, 6, 7, 8],
    'CustomerID': ['CUST100', 'CUST101', 'CUST102', 'CUST103', 'CUST100', 'CUST104', np.nan, 'cust105'],
    'Product': ['Laptop', 'Mouse', 'Keyboard', 'Monitor', 'Laptop', 'Webcam', 'Mouse', 'LAPTOP'],
    'Price': [1200.50, 25.0, 75.0, np.nan, 1200.50, 45.0, 25.0, '1500'],
```

```
    'TransactionDate': ['2023-01-01', '2023-01-02', '2023-01-03', '2023-01-04', '2023-01-01', '2023-01-05', '2023-01-06', '2023/01/07']
}
df = pd.DataFrame(data)

print("Original DataFrame:")
print(df)
print("-" * 30)

# --- Step 1: Handling Missing Values ---
# For 'Price' column, fill missing NaN with the median
median_price = df['Price'].median()
df['Price'].fillna(median_price, inplace=True)
print("DataFrame after filling missing 'Price' values:")
print(df)
```

```
print("-" * 30)

# --- Step 2: Removing Duplicates ---
# The row with TransactionID 1 and 5 are duplicates
df.drop_duplicates(subset=['CustomerID', 'Product', 'Price'], keep='first', inplace=True)
print("DataFrame after removing duplicates:")
print(df)
print("-" * 30)

# --- Step 3: Addressing Inconsistent Formats ---

# Standardize text in 'Product' column to lowercase
df['Product'] = df['Product'].str.lower()
```

```
print("DataFrame after standardizing 'Product' text:")
print(df)
print("-" * 30)

# Convert 'Price' column to numeric type
df['Price'] = pd.to_numeric(df['Price'])
print("DataFrame after converting 'Price' to numeric:")
print(df)
print("-" * 30)

# Standardize 'TransactionDate' column to datetime objects
df['TransactionDate'] = pd.to_datetime(df['TransactionDate'])
print("DataFrame after converting 'TransactionDate' to datetime:")
```

```
print(df)
print("-" * 30)

# --- Step 4: Final Cleaned DataFrame ---
# Reset the index after dropping rows
df.reset_index(drop=True, inplace=True)
print("Final Cleaned and Prepared DataFrame:")
print(df)
```

End of code

I have used python pandas. This script demonstrates how to handle missing values, duplicates, and inconsistent data types.

THANK YOU