Experiment no: 03 Date:

30-07-2025

EDA – Data Cleaning

**Aim:** To perform data cleaning by handling missing values, removing duplicates, converting data types, and normalizing data.

# Code:

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

# Step 1: Create sample dataset data

= {

"ID": [1, 2, 3, 4, 5, 5],

"Name": ["Alice", "Bob", "Charlie", "David", None, "David"], "Age": [23, 25, np.nan, 24, 22, 22],

"Marks": [85, 78, 90, np.nan, 95, 95],

"Department": ["CSE", "ECE", "ME", "CIVIL", "AI", "AI"]

}

df = pd.DataFrame(data) print("Original Data:") print(df)

# Step 2: Handling missing values print("\nHandling Missing Values:") print("Detect missing:\n", df.isnull().sum())

# Fill missing with mean for Age, fill with mode for Marks df["Age"].fillna(df["Age"].mean(), inplace=True)

df["Marks"].fillna(df["Marks"].mode()[0], inplace=True)

# Fill missing Name with "Unknown" df["Name"].fillna("Unknown", inplace=True)

print("\nAfter Filling Missing Values:") print(df)

# Step 3: Remove duplicates df = df.drop\_duplicates() print("\nAfter Removing Duplicates:") print(df)

# Step 4: Data type conversion df["ID"] = df["ID"].astype(str) # Convert ID to string print("\nAfter Data Type Conversion:") print(df.dtypes)

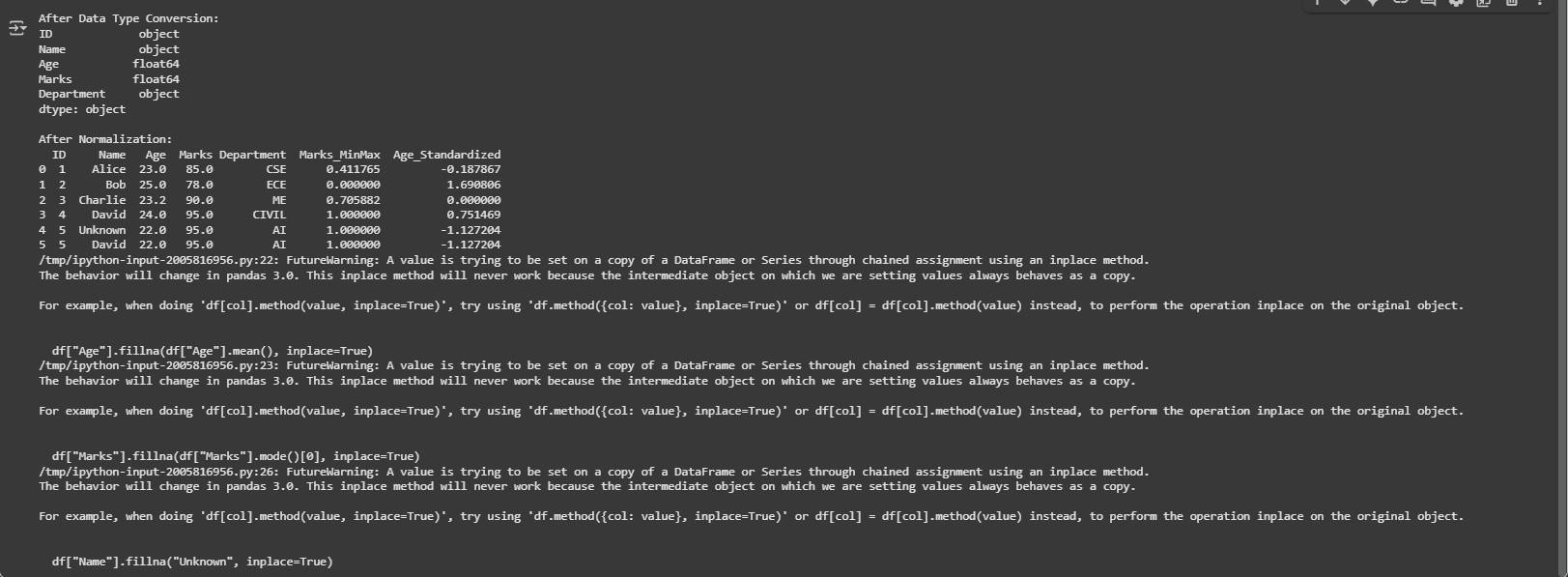
# Step 5: Normalization scaler = MinMaxScaler() df["Marks\_MinMax"] = scaler.fit\_transform(df[["Marks"]])

standard\_scaler = StandardScaler() df["Age\_Standardized"] = standard\_scaler.fit\_transform(df[["Age"]])

print("\nAfter Normalization:") print(df)

# Output:

****



**Result:** Successfully cleaned the dataset by handling missing values, removing duplicates, converting data types, and normalizing data.