Project – 3: Data Cleaning Utility Using Pandas

Objective

The objective of this project is to clean and preprocess raw data using the Pandas library by handling missing values, removing duplicate records, correcting data types, and standardizing column names. This project aims to improve data quality and prepare a clean dataset suitable for further analysis or modelling.

CODES:

```
import pandas as pd
import numpy as np

# Step 1: Create a raw sample dataset
data = {
    "Name ": ["Amit", "Riya", "Sohan", "Riya", None],
    "Age": [23, np.nan, 25, 28, 30],
    "Salary": [35000, 55000, None, 55000, 60000],
    "Department": ["HR", "IT", "Finance", "IT", "IT"]
}

df = pd.DataFrame(data)
print("Raw Dataset:\n", df)

# Step 2: Standardize column names
df.columns = df.columns.str.strip().str.lower()
print("\nStandardized Columns:\n", df.columns)

# Step 3: Handle missing values
df['age'].fillna(df['age'].mean(), inplace=True)
df['salary'].fillna(df['salary'].median(), inplace=True)

# Step 4: Remove duplicate rows
df.drop_duplicates(inplace=True)

# Step 5: Handle missing names
df['name'].fillna("Unknown", inplace=True)

# Step 6: Data type correction
df['age'] = df['age'].astype(int)
df['salary'] = df['salary'].astype(int)

print("\nCleaned Dataset:\n", df)

# Step 7: Save cleaned data
df.to_csv("cleaned_employee_data.csv", index=False)

print("\nCleaned data saved successfully!")
```

OUTPUT:

```
Raw Dataset:
    Name    Age   Salary Department
0   Amit  23.0  35000.0         HR
1   Riya   NaN  55000.0         IT
2  Sohan  25.0      NaN    Finance
3   Riya  28.0  55000.0         IT
4   None  30.0  60000.0         IT

Standardized Columns:
 Index(['name', 'age', 'salary', 'department'], dtype='object')

Cleaned Dataset:
       name  age  salary department
0     Amit   23   35000         HR
1     Riya   26   55000         IT
2    Sohan   25   55000    Finance
3     Riya   28   55000         IT
4  Unknown   30   60000         IT

Cleaned data saved successfully!
```

Output Summary

- *Created a raw dataset containing missing values, duplicate records, and inconsistent column names.*
- *Standardized column names by removing extra spaces and converting them to lowercase.*
- *Handled missing numerical values using mean and median imputation.*
- *Removed duplicate records to ensure data consistency.*
- *Replaced missing categorical values with meaningful placeholders.*
- *Corrected data types for numerical columns.*
- *Exported the cleaned dataset into a CSV file for further analysis**.***

Result: *The dataset was successfully cleaned and structured, making it reliable and ready for analytical or machine learning tasks.*

Conclusion

This project demonstrates essential data cleaning techniques required in real-world data analysis. Proper data preprocessing improves data reliability, accuracy, and usability for decision-making and predictive modelling.