**EXP 3:**

**Develop a python script for data cleanup on child labour anda marriage data.xlsx**

1. **Check duplicate and missing data**
2. **Cleans line breaks, spaces, and special characters.**

**import pandas as pd**

# Read the CSV file

**df = pd.read\_csv(r"C:\Users\91984\Downloads\Child Labour in India.csv")**

**print("CSV file read successfully.")**

# Display the first few rows

**print("First few rows of the DataFrame:")**

**print(df.head())**

# Get basic info about the DataFrame

**print("Shape of the DataFrame (rows, columns):", df.shape)** # Rows, columns count

**print("List of columns in the DataFrame:", df.columns)**  # List of columns

**print("Data types of each column:")**

**print(df.dtypes)**  # Data types of each column

**print("More info about the DataFrame:")**

**print(df.info())** # More info: non-null counts, types

**print("Statistical summary of numeric columns:")**

**print(df.describe())** # Statistical summary (numeric columns)

# Check and count missing/NULL values

**print("Missing values per column:")**

**print(df.isnull().sum())**  # Missing values per column

**print("Total missing values in the DataFrame:", df.isnull().sum().sum())** # Total missing values

# Check and remove duplicates

**print("Number of duplicate rows in the DataFrame:", df.duplicated().sum())**  # Number of duplicate rows

**df = df.drop\_duplicates()**  # Remove duplicates

**print("Duplicates removed. New shape of the DataFrame:", df.shape)**

# Optional: Clean column names (remove spaces, lowercase, etc.)

**df.columns = df.columns.str.strip().str.lower().str.replace(' ', '\_')**

**print("Column names cleaned. New column names:")**

**print(df.columns)**

# Save cleaned file (optional)

**df.to\_csv("child\_labour\_in\_india\_cleaned.csv", index=False)**

**print("Cleaned DataFrame saved to 'child\_labour\_in\_india\_cleaned.csv'.")**

OUTPUT IN TEXT

CSV file read successfully.

First few rows of the DataFrame:

Category of States States Agriculture Manufacturing \

0 Non Special Category states Andhra Pradesh 69.0 9.7

1 Non Special Category states Assam 69.3 8.4

2 Non Special Category states Bihar 71.8 11.2

3 Non Special Category states Chhattisgarh 87.9 2.4

4 Non Special Category states Delhi NaN 11.1

Construction Trade Hotels & Restaurants \

0 3.2 9.0

1 1.8 7.8

2 0.0 15.5

3 0.9 7.2

4 NaN 57.8

Community, Social and Personal Services Others Total

0 7.1 2.0 100

1 11.0 1.8 100

2 1.1 0.5 100

3 1.7 NaN 100

4 31.1 NaN 100

Shape of the DataFrame (rows, columns): (25, 9)

List of columns in the DataFrame: Index(['Category of States', 'States', 'Agriculture', 'Manufacturing',

'Construction', 'Trade Hotels & Restaurants',

'Community, Social and Personal Services', 'Others', 'Total'],

dtype='object')

Data types of each column:

Category of States object

States object

Agriculture float64

Manufacturing object

Construction float64

Trade Hotels & Restaurants float64

Community, Social and Personal Services float64

Others float64

Total int64

dtype: object

More info about the DataFrame:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 25 entries, 0 to 24

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Category of States 25 non-null object

1 States 25 non-null object

2 Agriculture 24 non-null float64

3 Manufacturing 24 non-null object

4 Construction 22 non-null float64

5 Trade Hotels & Restaurants 25 non-null float64

6 Community, Social and Personal Services 23 non-null float64

7 Others 19 non-null float64

8 Total 25 non-null int64

dtypes: float64(5), int64(1), object(3)

memory usage: 1.9+ KB

None

Statistical summary of numeric columns:

Agriculture Construction Trade Hotels & Restaurants \

count 24.000000 22.000000 25.000000

mean 68.500000 2.540909 11.720000

std 16.613536 2.046753 11.186748

min 19.200000 0.000000 3.300000

25% 65.525000 0.975000 7.200000

50% 70.550000 1.950000 8.500000

75% 80.700000 3.825000 9.700000

max 87.900000 7.000000 57.800000

Community, Social and Personal Services Others Total

count 23.000000 19.000000 25.0

mean 5.795652 1.436842 100.0

std 7.119658 0.916643 0.0

min 0.300000 0.000000 100.0

25% 1.400000 0.900000 100.0

50% 2.800000 1.200000 100.0

75% 7.100000 1.900000 100.0

max 31.100000 4.000000 100.0

Missing values per column:

Category of States 0

States 0

Agriculture 1

Manufacturing 1

Construction 3

Trade Hotels & Restaurants 0

Community, Social and Personal Services 2

Others 6

Total 0

dtype: int64

Total missing values in the DataFrame: 13

Number of duplicate rows in the DataFrame: 2

Duplicates removed. New shape of the DataFrame: (23, 9)

Column names cleaned. New column names:

Index(['category\_of\_states', 'states', 'agriculture', 'manufacturing',

'construction', 'trade\_hotels\_&\_restaurants',

'community,\_social\_and\_personal\_services', 'others', 'total'],

dtype='object')

Cleaned DataFrame saved to 'child\_labour\_in\_india\_cleaned.csv'.