Date : 17-02-2026

Data Cleaning Task Report using Regular Expressions

In this task, data cleaning was done using Python basic functions and regular expressions using Jupyter Notebook.

Dataset Name : Employee dataset

- **Size: 100 KB**
- **Columns: 12**
- **Rows: 1020**

Libraries like **Pandas and NumPy** were imported for data processing, and the **Re** library was used for applying regular expressions.

The shape of the dataset was checked using **df.shape** to see how many rows and columns were present. The **info()** method was used to understand the structure of the dataset. Also, **df.dtypes()** was used to check data types, and **describe()** was used to get a basic statistical summary of the data.

Then, null values and duplicate records were checked. No duplicate rows were there. Missing values were handled using the **fillna()** method.

Next, **regular expressions** were used for cleaning specific columns.

• **Numbers** were extracted from **employee ID** values that contained letters and numbers.
• **First name** and **last name** columns were cleaned by removing extra spaces and then **combined** into a single name column (**Full Name**).
• The **department** and **region** column was split into two separate columns using a hyphen.
• **Email** addresses were checked to see whether they have a **valid format**.
• **Phone numbers** were cleaned by removing special characters and keeping only **numbers**.
• **Username** and **domain** columns were created from **email** addresses to **classify** domains.
• This dataset has only one domain, but these columns were added so the same method can be used for real-time datasets with different domains.
• The **username** column helps to **identify users**, and the **domain** column helps to **group and classify email domains**.

• After completing all cleaning steps, the shape of the dataset was checked again. The final dataset contains **1020 rows and 15 columns**.

• Column changes after cleaning:
– Old column names: First Name, Last Name, Department Region, Email
– New column names: Full Name, Department, Region, Username, Domain

The final cleaned data was saved.