Document Summery

Steps : Dataset cleaning

1.Load the Data

- We'll assume the data is in a CSV file named `messy_data.csv`.
- Use pandas to load the data into a DataFrame.

```
dataset = pd.read_csv("messy_data.csv")
dataset.head()
```

2. Inspect the Data

- Get a quick overview of the data to understand its structure and identify errors.

```
dataset.shape
dataset.describe()
```

3. Handle Missing Values

Identify missing values and handle them appropriately.

```
dataset.isnull().sum()
```

4. Remove Duplicates

- Find and remove duplicate rows.

```
df_cleaned.duplicated()
df_cleaned = df_cleaned.fillna(0)
```

5. Correct Email Formats

- Validate and correct email formats.
- Filter for professional emails.
- Import re module

```
df_cleaned = df_cleaned[df_cleaned['Email'].apply(is_professional_email)]
```

6. Clean Name Fields

- Standardize and clean name fields.

```
df_cleaned['Name'] = df_cleaned['Name'].apply(clean_name)
```

7. Standardize Date Formats

- Ensure all dates in the 'Join Date' column follow a consistent format (`yyyy-mm-dd`).

- from datetime import datetime

8. Correct Department Names:

- Identify and correct typos in the 'Department' column. Standardise the department names to ensure consistency .

```
  df_cleaned.loc[:, 'Department'] =
df_cleaned['Department'].str.strip().str.title()
```

9.Handle Salary Noise:

- Identify and handle any noise in the 'Salary' column. Ensure salary values are within a reasonable range and free from random fluctuations.

```
df_cleaned = df_cleaned[df_cleaned['Salary'].between(30000,200000)]
```

10.Re name columns

- df_cleaned.rename(columns = {'Unnamed: 0':'Index'}, inplace = True)

11. Sort Index column

- df = df_sorted.sort_index(ascending = True)

12. Arrange the column

- new_column_order = ['Index','Cleaned_ID','Name','Age','Email','Join Date','Salary','Department','ID']
df _sorted = data[new_column_order]

13.Save csv file

- df_sorted.to_csv('my_data_file.csv',index = False)

1. **Document Your Process:**

1 Load the Data
2 Inspect the Data
3 Handle Missing Values
4 Remove Duplicates
5 Correct Email Formats
6 Clean Name Fields
7 Standardize Date Formats
8 Correct Department Names:
9 Handle Salary Noise

#Install python latest version 3.12.4
#install jupyter notebook
#install pip package
# using tools jupyter notebook & python
# used libraries like pandas,numpy,datetime,seaborn
# email validation,name standerdization time error occured message - "copywrite warning" that error detect using loc.accecssor(selecting data row and column)
#date time stadarndization ,email validation type error messege = 'expected string or bytes-like object,got 'int' ,issues solved astype libraries