TASK REPORT

Introduction:

This report outlines the data cleaning process undertaken on the dataset messy_data.csv to prepare it for analysis. The steps include handling missing values, removing duplicates, correcting email formats, cleaning name fields, standardizing date formats, correcting department names, and handling salary noise.

Data Cleaning Steps:

1. Load the Data:

The dataset messy_Data.csv was loaded into a Jupyter notebook using the Pandas library. The initial inspection checked the structure, data types, and basic statistics.

2. Inspect the Data:

An initial inspection was conducted to understand its structure and identify errors and inconsistencies. This included:

- Viewing the first few rows of the dataset.
- Checking data types and summary statistics.
- Identifying missing values and duplicate records.

3. Handle Missing Values:

I assumed missing values could impact data analysis and interpretation. I decided that rows with missing values in all columns should be removed to maintain data integrity. I removed rows with missing values using the dropna() method.

4. Remove Duplicates:

Duplicate rows were identified and removed to ensure that each record in the dataset was unique. Used Pandas drop_duplicates() method to identify and remove duplicate rows based on all columns. This step is crucial for accurate analysis and reporting.

5. Correct Email Formats:

I Assumed emails might contain formatting errors such as '@' symbol or incorrect domain. I also assumed professional emails typically have domains like 'gmail.com', 'yahoo.com', or company-specific domains. I defined a function to correct email formats using string manipulation and regular expressions. This function removed spaces and added missing '@' symbols with appropriate domains. I filtered out non-professional emails based on predefined domains.

6. Clean Name Fields:

I Assumed names might contain titles (e.g., Mr., Mrs., Dr.) at the beginning that should be removed. I also assumed last names might have appended extraneous words that should be removed to clean the name properly. I used regex (re.sub()) to remove titles from the beginning of names. Then splits name into parts and cleans the last name by removing extraneous words using regex patterns.

7. Standardize Date Formats:

I Assumed dates might be stored in different formats (e.g., 'DD/MM/YYYY' vs 'YYYY-MM-DD')and also assumed a need to handle various date formats and convert them into a consistent format (YYYY-MM-DD). I used Pandas' to_datetime() function to parse and convert dates into a consistent format. Then I handled potential errors in date parsing by trying multiple formats and returning None for invalid dates.

8. Correct Department Names:

I Assumed department names might contain typos or variations that need standardization. I also assumed departments start with specific keywords (e.g., 'Engineering', 'Support', 'HR', 'Sales', 'Marketing'). I Defined a list of known department keywords, then matches department names to keywords and returns the closest match.

9. Handle Salary Noise:

I Assumed salaries should be displayed as whole numbers for clarity and consistency. So I used Python's string formatting ('{:.0f}'.format) to convert salary values to whole numbers.

Methodologies:

Pandas Library: Used for data manipulation and cleaning tasks.

Regular Expressions: Employed for validating and filtering email formats.

Conclusion:

The dataset messy_data.csv was successfully cleaned and prepared for analysis. The cleaned dataset was saved as cleaned_dataset.csv, and this summary document outlines the steps, assumptions, and methodologies used in the data cleaning process.