**Data Cleaning Process Documentation**

**Step 1:** **Loading the Dataset** - The dataset, `messy\_data.csv`, was loaded into the Jupyter notebook for inspection using the `pandas` library.

**Step 2: Initial Inspection of the Data -** The structure of the data was examined to identify inconsistencies and potential issues such as missing values, duplicates, incorrect data formats, and noisy data.

**Step 3: Documenting QA Issues -** Data quality issues were recorded for each field to plan the cleaning approach. Issues identified included:

•Missing values in columns such as Age, Department, and Salary.

•Duplicate rows in the dataset.

•Invalid email formats in the Email column.

•Inconsistent date formats in the Join Date column.

•Typos and variations in Department names.

•Noise in the Name field due to special characters and numbers.

**Step 4: Handling Missing Values –**

•Missing values in the Age column were replaced with the median.

•Missing values in the Department column were marked as “Unknown”.

• Missing values in the Name column were marked as “Unknown”.

• Missing values in the Join Date column were replaced with the mode.

•Missing values in the salary column were replaced with the median.

•Missing values in the Email column were marked as “Unknown”.

**Step 5: Removing Duplicates**

**Assumption**: Duplicate records were considered errors and were removed to ensure that each record is unique.

**Step 6: Correcting Email Formats**

**Assumption**: Emails that do not follow the standard format (username@domain.com) are considered invalid. Only professional emails were retained.

**Step 7: Cleaning Name Fields**

**Assumption**: Names should consist only of alphabets and spaces. Special characters and numbers were considered noise.

**Step 8: Standardizing Date Formats**

**Assumption**: The Join Date column may contain dates in various formats, so it was standardized to YYYY-MM-DD.

**Step 9: Correcting Department Names**

**Assumption:** Department names were standardized using a mapping dictionary to correct for variations and typos.

**Step 10: Handling Salary Noise**

**Assumption:** Salary values below 20,000 and above 200,000 were considered unreasonable. These values were clipped to remove noise.

**Step 11: Final Inspection and Saving the Cleaned Data**

After completing the cleaning process, the dataset was inspected again to ensure that all issues were addressed. The cleaned dataset was then saved as cleaned\_dataset.csv.

**Conclusion:**

The dataset was successfully cleaned by:

1. Handling missing values in the Age and Department columns.
2. Removing duplicate rows.
3. Correcting invalid email formats and retaining professional email addresses.
4. Cleaning the Name column to remove any noise.
5. Standardizing the Join Date format.
6. Correcting department names for consistency.
7. Handling salary noise by clipping values to a reasonable range.