Summary

Introduction

The dataset consists of 11000 records and 7 features. The features are ID, Name, Age, Email, Join Date, Salary, and Department. The data cleaning process was performed to prepare the dataset for further analysis.

Initial Data Assessment

The datatypes of the columns and non-null value counts are:

Index	Feature	Datatype	Non-null Count
1	ID	object	11000
2	Name	object	8667
3	Age	float	9253
4	Email	object	9731
5	Join Date	object	8808
6	Salary	float	8761
7	Department	object	8745

The dataset contains duplicate values in the ID column, which should be unique. Additionally, some records only have an ID value with all other fields missing.

Data Cleaning Steps

ID Column

- The ID column is expected to have unique values; hence, duplicate ID entries were removed.
- Additionally, records that contained only an ID with all other fields missing were also dropped to improve data quality.

Name Column:

- Missing values in the Name column were replaced with `"Unknown"` to ensure completeness.
- Names containing unwanted leading or trailing characters were cleaned.
- The column includes both first and last names, which were capitalized for consistency.

Age Column:

- Since Age is a numerical (float) value, missing entries were imputed using the median to prevent skewness caused by outliers.

Email Column:

- Email addresses should follow the standard format (`username@domain.com`).
- Invalid or missing email addresses were replaced with `"unknown@domain.com"` to maintain a consistent format.

Join Date Column:

- Dates were formatted in `'YYYY-MM-DD'`.
- Missing values in the Join Date column were filled with `"1900-01-01"` as a placeholder.
- The column was converted to a date format to facilitate date-based comparisons.

Salary Column:

- Missing values were imputed using the median to minimize the impact of outliers.
- Salaries were rounded to two decimal places for consistency.

Department Column:

- Missing values were replaced with the mode (most frequently occurring value) since it is a categorical field.

Handling Duplicates (ID & Name Combination):

- To remove additional duplicates, a combination of ID and Name was used for duplicate detection.
- In cases where multiple Join Dates existed for the same person, the most recent date was retained to ensure accuracy and prevent inconsistencies between Join Date and Age.

Final Data Summary

The raw dataset initially contained 11,000 rows and 7 features. After the data cleaning process, the number of entries was reduced to 8,908 rows, while the number of features remained the same.

- All missing values in the six non-ID columns were appropriately handled through imputation or replacement.

- Duplicate records were removed by identifying duplicates in the ID column and using a combination of ID and Name for further deduplication.
- The ID column, being a unique identifier, was ensured to have no duplicate values.

This cleaned dataset is now structured and ready for further analysis.

	Raw Data	Cleaned Data
Records	11000	8908
Features	7	7

Challenges and Considerations

The primary challenges in the data cleaning process were handling missing values and duplicate records:

- **Handling Missing Values**:
  - Missing values were imputed using appropriate strategies:
    - Age and Salary were filled with the median to minimize the impact of outliers.
    - Department, being categorical, was filled with the mode (most frequent value).
    - Name column had many unwanted characters, requiring extensive cleaning. Names were capitalized for consistency.
    - Email values that were missing or invalid were replaced with `"unknown@domain.com"`.
- **Duplicate Removal**:
  - Duplicate records were removed based on the ID column and a combination of ID and Name to maintain data integrity.

Limitations

Despite thorough cleaning, some limitations remain:

- **Imputation Quality**:
  - Replacing missing names and emails with placeholder values (`"Unknown"`, `"unknown@domain.com"`) reduces data accuracy.
  - Filling Department with the mode may misclassify an individual's actual department.
  - Age and Salary values, though imputed using the median, may not always reflect the true values.
- **Time Complexity**:

- Cleaning the Name column was particularly time-intensive, as hundreds of unnecessary words needed to be identified and removed.

While these limitations exist, the cleaned dataset is significantly improved in terms of completeness, consistency, and usability for further analysis.

Conclusion

The raw dataset was transformed into a structured and high-quality dataset through systematic data cleaning using various Python modules. The cleaned data is now consistent, complete, and ready for further data analysis and analytics applications.