Case Study: Data Cleaning and Preprocessing

Background

A mid-sized company has collected employee records for analysis. However, before performing any meaningful data-driven insights, the HR analytics team needs to clean and preprocess the data. You, as a data analyst, are assigned to handle missing values, remove duplicate entries, transform data for consistency, detect and remove outliers, and format the dataset properly.

Dataset Overview

The dataset consists of **10,000+ records** of employees, with the following fields:

1. **ID** – Unique identifier for each employee.
2. **Name** – Employee full name.
3. **Age** – Employee age (some values are missing).
4. **Salary** – Annual salary of the employee (contains missing values and outliers).
5. **Joining_Date** – The date the employee joined the company.
6. **Department** – The department in which the employee works (some values have inconsistent casing).
7. **Email** – Employee email address (some are uppercase, some lowercase).
8. **Performance_Score** – Employee performance rating (1-10 scale).

Key Issues in the Dataset

1. **Missing Values**
   - **Age** and **Salary** columns contain missing values.
   - Some employees do not have salary data, requiring imputation techniques.
2. **Duplicate Data**
   - There are 200 duplicate records in the dataset that need to be identified and removed.
3. **Data Transformation**
   - Email addresses have inconsistent formatting (some in uppercase, others in lowercase).
   - Department names have inconsistent casing.
4. **Outlier Detection and Removal**
   - Some salary values are unrealistically high due to data entry errors.
5. **Data Formatting Issues**
   - Some date formats might be inconsistent.
   - Employee names might have variations in spacing or casing.

Tasks for Participants

1. **Handle Missing Values**
   - Identify missing values in **Age** and **Salary** columns.
   - Use appropriate imputation techniques (mean, median, or mode) to fill missing values.
2. **Detect and Remove Duplicate Records**
   - Identify and drop duplicate rows while keeping only unique employee records.
3. **Transform and Standardize Data**
   - Convert all email addresses to lowercase.
   - Ensure all department names have proper capitalization.
4. **Detect and Handle Outliers**
   - Use statistical techniques (e.g., IQR, Z-score) to detect and remove salary outliers.
5. **Format the Data Properly**
   - Ensure date values are formatted consistently.
   - Standardize employee names for uniformity.

Expected Outcome

After completing the above tasks, the cleaned dataset should be:

- **Free of missing values** (handled using imputation techniques).
- **De-duplicated** (no redundant records remain).
- **Standardized** (consistent email and department formats).
- **Free of extreme outliers** (realistic salary values maintained).
- **Well-formatted** for further analysis.

Discussion Questions

1. What are the best practices for handling missing data in business analytics?
2. How can you efficiently detect and remove duplicate records in large datasets?
3. Why is it important to standardize categorical data like department names and email formats?
4. What are some common techniques to detect and handle outliers in numerical data?
5. How does data preprocessing impact the accuracy of machine learning models?