DATAHUT QA ASSIGNMENT

Introduction

The dataset provided for this assignment consists of 11,000 records and 8 columns. Upon initial inspection, multiple data quality issues were identified, including missing values, inconsistent formatting, incorrect data entries, duplicate records, and outliers. These issues can significantly impact data accuracy and reliability, making it unsuitable for analysis without proper cleaning and standardization.

The primary objective of this data cleaning process was to enhance the accuracy, consistency, and completeness of the dataset. This was achieved by addressing missing values, standardizing formats, correcting errors, removing duplicates, and ensuring uniformity in data representation. The following document outlines the identified issues, the methodologies used for rectification, and the final outcome of the cleaning process.

Identified Issues and Fixes

1. Missing Values

Several fields in the dataset contained missing values, which were handled as follows:

- Name: 8,667/11,000 records had valid names. Missing values were replaced with "Unknown."
- Age: 9,253/11,000 records had valid age data. Missing values were either interpolated or removed if necessary.
- Email: 9,731/11,000 records had valid email addresses. Missing values were replaced with "missing@example.com."
- Join Date: 8,808/11,000 records had valid dates. Missing dates were set to "Unknown."
- Salary: 8,761/11,000 records had valid salary values. Missing values were replaced with the median salary.
- Department: 8,745/11,000 records had valid department values. Missing values were assigned based on similar employee data.

2. Formatting and Data Standardization

- Join Date: The date format was inconsistent across records. It was standardized to YYYY-MM-DD format.
- Department Names: Typographical errors and inconsistencies were corrected by mapping them to a predefined list of department labels (e.g., "SupportJ" was corrected to "Support").
- Email Addresses: Invalid email formats were identified and replaced with "invalid@example.com" while ensuring professional email addresses were retained.
- Names: Extraneous words and special characters were removed to maintain uniformity.

3. Invalid and Outlier Data

- Salary Values: Extreme outliers were identified and capped within a reasonable range (₹20,000 - ₹200,000) to prevent data distortion.
- Age Values: Unreasonable age values (below 18 or above 65) were either corrected or removed to maintain data credibility.
- Duplicate Records: Identified and removed to maintain data integrity.

Assumptions & Methodologies

To ensure an effective data cleaning process, the following assumptions and methodologies were applied:

- **Missing names** were replaced with "Unknown" to retain the record while acknowledging incomplete information.
- **Invalid email addresses** were substituted with "invalid@example.com" to differentiate them from legitimate records.
- **Extreme salary values** were capped within an acceptable range to prevent data distortion.
- **Unrealistic ages** (below 18 or above 65) were considered erroneous and removed.
- **Name fields** were cleaned by removing special characters and ensuring uniform capitalization.

- **Email validation** was performed using **regular expressions (regex)** to ensure compliance with standard email formats.

Tools Used

The following tools and technologies were utilized to carry out the data cleaning process:

- **Python (Pandas, NumPy, Regex):** For data processing, transformation, and cleaning.
- **Google Colab:** For executing the data cleaning pipeline in a cloud-based environment.
- **WPS Office (Spreadsheet Tool):** For documenting data quality issues and maintaining an organized record of changes.

Conclusion

After implementing the data cleaning process, the dataset has been transformed into a **structured, accurate, and reliable format** suitable for further analysis. The applied cleaning steps have significantly improved data integrity by eliminating inconsistencies, standardizing formats, correcting erroneous values, and handling missing data effectively.

The cleaned dataset ensures:

- **Higher data accuracy**, minimizing errors that could impact decision-making.
- **Consistency across all fields**, preventing discrepancies in analysis.
- **Completeness**, with properly handled missing values to retain as much information as possible.
- **Readiness for analysis**, making the dataset suitable for predictive modeling, reporting, or business intelligence applications.

This structured approach ensures that the dataset is now fully optimized for any downstream analytical processes, making it a valuable asset for data-driven insights and decision-making.