DATA QA/QC INTERNSHIP TASK

SUMMARY DOCUMENT

Data Overview

Data cleaning is vital for accurate and meaningful analysis. This project focuses on transforming a raw employee database into a reliable and analyzable dataset by addressing errors, inconsistencies, and missing values.

Dataset Description

The employee database contains the following columns:

1. **ID**: Unique identifier for each employee.
2. **Name**: Employee's full name.
3. **Age**: Employee's age.
4. **Email**: Employee's official email address.
5. **Join Date**: Date of employment.
6. **Salary**: Annual salary.
7. **Department**: Department where the employee works.

Data Processing

1. Removing Duplicates

The 'ID' column, intended as a unique identifier, contained 1000 duplicate entries. These duplicates were removed to maintain the integrity of the dataset.

2. Handling Null Values

Dataset contained 6 columns with null values. Proceeded with handling them after conducting Exploratory Data Analysis (EDA) to examine potential relationships between variables. EDA revealed no significant correlations among the variables.

Method of handling **:**

Email

- Removed records with null values in the 'Email' column. The email address is essential as it is the primary means of communication and uniquely identifies each employee. Therefore, retaining only records with valid email addresses ensures data integrity and utility.

Name

- Replaced null values with the placeholder 'Unknown'. Since the 'Name' field is a categorical variable and most people do not share the same first and last name, using a placeholder ensures completeness while avoiding confusion with existing names.

Age

- Imputed null values with random values drawn from the existing age data. The distribution graph of the 'Age' column indicated a roughly uniform distribution. By using random values from the existing data, this method preserves the uniform nature of the age distribution, ensuring the imputed values blend seamlessly into the dataset.

Join Date

- Filled null values with random dates between 2020 and the last recorded join date in 2024. The dataset showed a significant increase in recruitments during these years. This approach maintains the temporal relevance of the data and aligns with the observed hiring trends, providing realistic imputed values that reflect the actual hiring patterns.

Department

- Replaced null entries with random departments. Analysis of the dataset indicated a relatively even distribution of employees across departments, making random assignment suitable for maintaining data uniformity without introducing bias.

Salary

- Imputed null values with the median salary of the respective department. The histogram of salary data showed positive skewness, making the median a more robust measure than the mean since its less affected by skewed data. This method ensures that imputed salary values are representative of typical earnings within each department, maintaining data integrity and consistency for analysis purposes.

3. Correcting Email formats

- Standardized email formats by creating a function that added '@' before domain names that were missing it, ensuring all email addresses conform to the standard format (e.g., username@domain.com).

- Removed addresses that did not follow a standard email format using regex within a function, ensuring that all retained emails are valid and usable for communication and analysis purposes.

4. Standardising Date Formats

- Ensured all dates in the 'Join Date' column followed a consistent format (e.g., YYYY-MM-DD) by applying a function that standardized the date format across the column.

5. Correcting Department Names:

- Standardized department names by correcting typos and variations to ensure consistency. Departments were standardized to common formats such as HR, Engineering, Marketing, Sales, and Support.

6. Handling noise in Salary and Age

- Converted the data type of both Salary and Age columns to integer to remove noise and ensure consistency in numerical representation.

- Ensured Salary values do not contain outliers by visualizing the distribution using a box plot. This graphical representation helps identify any extreme values that could distort analysis, ensuring the dataset accurately represents the salary distribution within reasonable bounds.

7. Cleaning Name Fields

- Removed honorifics such as 'Mr.', 'Mrs.', 'Dr.', etc., from the beginning of names using regex to standardize name entries.

- Eliminated trailing capital letters indicating professional titles like 'DDS', 'MD', etc., using regex to ensure consistency in name formats.

- Corrected trailing capital letters with typos, such as 'DDSraise', by removing them using regex.

- Removed extraneous words or typos attached to last names, such as 'Tonya Philipsnetwork', 'Richard Bryantwhile', using NLTK's WordNet Corpus Reader to enhance the accuracy and clarity of names in the dataset.