**Assignment for Data QA & QC Internship @ Datahut**

**Data Cleaning Task Instructions**

I am provided with a dataset named messy\_Data.csv. My task is to clean this dataset and ensure it is ready for analysis.

Following Steps were taken for the data cleaning process.

1. **Load the Data:**

* Load the dataset into a jupyter notebook.
* pd: read\_csv used for this method

1. **Inspect the Data:**

* Understood the dimension
* understood the repetition in some columns to study the data patterns
* I feel there is some commonality between Unnamed 0 and ID
* Renamed Unnamed 0 to Serial No.

Handling the date format and Department Correction before handling the NULL values

1. **Standardise Date Formats:**

**A screenshot of a computer

Description automatically generated**

* After checking the df info() converted the data format datetime format

(YYYY-MM-DD).

1. **Correct Department Names:**

* Corrected the Department name before handling the missing values

by using check the unique department names that are mentioned.

1. **Handle Missing Values:**

* Almost all column have null values

**A screenshot of a computer

Description automatically generated**

* Handled: Nameand age together being NULL in many places.
* Age and Salary are normally distributed. Filled the NULL with **Mean**
* Department NULL was filled with **Most frequent** value
* Join date was filled with **Median** Value

1. **Remove Duplicates:**

* There are some rows with same Serial No and ID. I have kept only 1 row of such data
* Same email id was repeated in more than 1 entry. Removed the duplicate and kept the first occurrence only

1. **Correct Email Formats:**

* Regular expression to make all email addresses follow a standard format (e.g., username@domain.com).
* Regular expression used: r'^[\w\.-]+@[a-zA-Z\d\.-]+\.(com|info|net|org|biz)$'

1. **Clean Name Fields:**

* NULL names are dropped
* Found some name is having 3rd name. Trimmed out the 3rd name
* Found some name with title. Dropped the title and just kept the first name and Last name
* Found some extraneous words added towards the end of the Surname code added to remove those. Keeping all the remaining as same.
* Logic Followed to trim out the extraneous word is as follows:
  + List1 : Found all the Last name from df[‘Name’]
  + List2: Found repeating Last names (more than 2) from df[‘Name’]
  + List3 : keeping the List1 as it is but just if the last name matched the string in surname then replace the last name with the **valid surname**

This will remove the extraneous word from the **List1.**

* Some code to verify the correction:

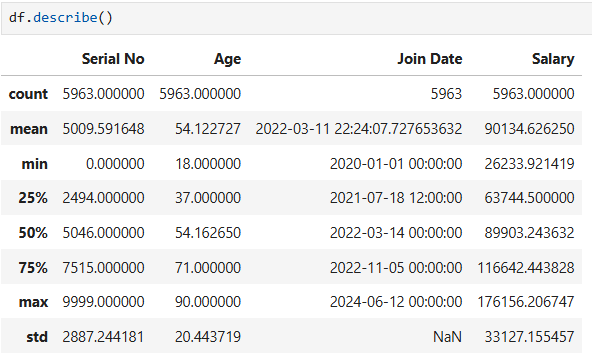
df[df['Name'].str.contains('Taylordaughter', case=False)]

df[df['Name'].str.contains('Lamb', case=False)]

We can see difference between surnames in df[‘Name’] and df[‘Name New’]

1. **Handle Salary Noise:**

* Salary column was verified



* Mean and Median(50%) is approximately same. Giving a normally distributed effect
* Minimum value is 26233.921419 and Maximum value is 176156.206747 which does not show any visible noise presence.

1. **Save the dataset**

* Data save to cleaned\_dataset.csv