Assignment for Data QA & QC Internship @ Datahut

Data Cleaning Task Instructions

You have been provided with a dataset named messy_Data.csv. Your task is to clean this dataset and ensure it is ready for analysis. Follow the instructions below to guide your data cleaning process.

1. **Load the Data:**

- Load the dataset into a jupyter notebook. Download the data from here: Messy data

2. **Inspect the Data:**

- Examine the data to understand its structure and identify the types of errors and inconsistencies present.

3. **Record QA Issues:**

- Document data quality issues in the dataset field-wise to ensure clarity and organisation

4. **Handle Missing Values:**

- Identify and handle missing values in the dataset. You can choose to fill them with appropriate values (e.g., mean, median) or remove rows/columns with excessive missing values.

5. **Remove Duplicates**:

- Identify and remove duplicate rows to ensure each record is unique.

6. **Correct Email Formats:**

- Identify and correct invalid email formats. Ensure that all email addresses follow a standard format (e.g., username@domain.com).

- have only professional emails in the final list.

7. **Clean Name Fields:**

- Remove any noise added to names. Ensure names are consistently formatted and free from extraneous words.

8. **Standardise Date Formats:**

- Ensure all dates in the 'Join Date' column follow a consistent format (e.g., YYYY-MM-DD).

9. **Correct Department Names:**

- Identify and correct typos in the 'Department' column. Standardise the department names to ensure consistency (e.g., HR, Engineering, Marketing, Sales, Support).

10. **Handle Salary Noise:**

- Identify and handle any noise in the 'Salary' column. Ensure salary values are within a reasonable range and free from random fluctuations.

11. **Document Your Process:**

- Keep a record of the steps you took to clean the data. Document any assumptions you made and the methods you used to address specific issues and add to the Jupyter notebook

Submission:

1. Once you have completed the data cleaning task, save the cleaned dataset as \`cleaned_dataset.csv\`.
2. Provide a summary document detailing the steps you took to clean the data, including any assumptions and methodologies used.
3. Rename your GitHub project to “Datahut QA Assignment” and submit the repository URL for this project as outlined in the email.

Timeline:

1. Final Delivery date: as mentioned in the email

Tools:

1. Python
2. Jupyter Notebook

Evaluation Criteria:

1. Completeness of data cleaning
2. Accuracy of corrections
3. Consistency and standardisation
4. Clarity and thoroughness of documentation

Good luck with the task! If you have any questions or need further clarification, please do not hesitate to ask.