Background:

You have recently joined AxionRay, a leading technology company revolutionizing engineering quality and safety with AI-driven solutions. AxionRay's innovative tools leverage cutting-edge AI technologies, including Generative AI and Large Language Models (LLMs), to address complex challenges in data processing, automation, and analytics for next-generation products like electric vehicles and airplanes.

This assignment is designed to evaluate your skills in data validation, cleaning, integration, and exploratory data analysis (EDA) using Python, as well as your ability to derive actionable insights.

Task 1 - Data Tagging

Data for task 1

Guidelines:

The dataset consists of the following:

- Free-text data (Columns: Complaint, Cause, Correction) that needs to be tagged.
- Taxonomy Sheet: A reference list with predefined categories for Root Cause, Symptom_Condition, Symptom_Component, Fix_Condition, and Fix_Component.

Your task is to tag the data by applying logical reasoning and aligning it with the categories provided in the taxonomy.

*Note: Two Examples of how to tag the data are already provided in Task 1. Please refer to those examples.

Deliverables for Task 1:

1. Submit the tagged dataset in **CSV or Excel format**.
2. A summary report (200-400 words) explaining -
   a. How you approached the tagging of each given field (Root Cause, Symptom_Condition, Symptom_Component, Fix_Condition, and Fix_Component.)
   b. Any potential insights that can be generated ***(Food for thought and has bonus marks)***

Evaluation Criteria for Task 1:

- **Tagging Accuracy**: At least 50% tagging accuracy is expected.
- **Critical Thinking**: Ability to handle ambiguous cases and document observations.

Task 2 - Data Analysis and Insights Generation using Python

Data for task 2 (Attached in email)

Guidelines:

1. **Column-Wise Analysis:**
   - Perform a column-wise analysis of the provided dataset.
   - Describe each column in terms of its data type, unique values, distribution, and overall significance for stakeholders
2. **Data Cleaning:**
   - Handle missing or invalid values using appropriate methods (e.g., imputation, deletion).
   - Address inconsistencies in categorical columns (e.g., typos, inconsistent capitalization).
   - Ensure numerical columns are in the correct format and free from outliers, where applicable.
3. **Identifying Critical Columns:**
   - Select the **top 5 critical columns** that might be most insightful for stakeholders according to your data understanding.
   - Provide reasoning for your selection.
   - Generate visualizations (e.g., bar plots etc) using Python to represent these insights effectively. **(atleast 3)**
4. **Generating tags/features from free text available :**
   - Generate meaningful tags from the free text fields to summarize information, example - failure conditions and components etc..
5. **Summary and Insights** ***(Food for thought and has bonus marks)***
   - Write a summary of the tags generated, including potential insights derived from the dataset.
   - Provide actionable recommendations for stakeholders based on your analysis.
   - Highlight discrepancies in the dataset (e.g., null values, missing primary keys) and how did you approach.

Deliverables for Task 2:

1. Submit the cleaned and tagged records in **CSV or Excel format**.
2. Submit a detailed report covering the above pointers, including (Maximum 2 page):
   a. Column analysis
   b. Data cleaning summary and
   c. Visualizations
   d. Generated tags & Key takeaways
3. Attach Python scripts used for the analysis.

Evaluation Criteria for Task 2:

1. Demonstrates clear understanding of the dataset provided in summary reports.
2. Visualizations are relevant and easy to interpret, with well-supported insights derived from exploratory data analysis.
3. Outputs, including scripts, datasets, and reports, are well-organized, professional, and demonstrate attention to detail.
4. **Bonus Points**:
    - Unique and meaningful tags beyond failure conditions and components.
    - Comprehensive summary with actionable insights.