**Data Tagging Summary – Task 1**

**1. Approach**

To ensure a structured and consistent tagging of the dataset, I adopted a **keyword-based matching approach** aligned with a predefined taxonomy of categories. The tagging process was divided into three major components:

* **Root Cause Tagging**:  
  I analysed the text in the **“Cause”** column to identify and extract specific keywords. These keywords were then mapped to a set of predefined root cause categories from the taxonomy. This allowed for standardized classification of failure reasons such as installation errors, wear and tear, or design flaws.
* **Symptom Condition & Component Tagging**:  
  The **“Complaint”** field was processed to detect descriptive terms and phrases that indicated the condition and component involved. Using a curated list of known symptom categories (e.g., 'leak', 'broken', 'movement issue') and component tags (e.g., 'wheel', 'seat', 'tray'), I matched complaint text to generate accurate symptom tags.
* **Fix Condition & Component Tagging**:  
  The **“Correction”** field was evaluated to identify the corrective actions taken. Based on the action verbs and terminology (e.g., 'replaced', 'tightened', 'aligned'), corresponding fix conditions and associated components were tagged. This helped in understanding the actual repair efforts made for each case.

**2. Challenges Encountered**

* **Ambiguity in Descriptions**:  
  A notable portion of the complaints lacked sufficient detail, making it difficult to conclusively determine the root cause or exact symptom. In such cases, only general tagging was possible.
* **Multi-symptom and Multi-fix Entries**:  
  Several records included multiple issues or fixes within a single entry. Disentangling this required additional logic to avoid misclassification and ensure all relevant tags were captured.
* **Data Inconsistencies**:  
  The dataset exhibited formatting and terminology variations—such as misspellings, shorthand notations, and inconsistent phrasing—which necessitated a preprocessing step to normalize the text for better keyword matching and accuracy.

**3. Key Insights**

* **Common Root Causes**:  
  The most frequently tagged root causes were **"Not Installed"** and **"Loose"**, indicating that improper assembly or incomplete installation procedures are the primary contributors to product issues.
* **Trending Complaints**:  
  A significant number of complaints revolved around **leakages, missing parts, and faulty component connections**. These highlight critical quality control areas that need immediate attention.
* **Data Quality Concerns**:  
  Approximately **20% of the dataset was found to be incomplete or too ambiguous** to allow for accurate tagging. This underscores the importance of improving the data collection process—specifically in how customer complaints and technician notes are recorded.
* **Implications for Quality Control**:  
  This tagging exercise provides actionable insights into recurring defects and repair patterns. By identifying the most common issues and failure points, manufacturers can refine their **defect tracking systems, enhance product assembly protocols, and implement targeted quality improvement measures.**