**Task 2**

**1. Column-Wise Analysis:** (IN PYTHON FILE IN CODE FORMAT)

**2. Data Cleaning:** (CODES ARE IN PYTHON FILE)

1. **Handle Missing/Invalid Values**
   * Drop rows with excessive missing data.
   * **Numeric**: Fill with mean/median.
   * **Categorical**: Fill with mode or "Unknown".
2. **Fix Categorical Inconsistencies**
   * Standardize text: lowercase/title case.
   * Trim whitespace.
   * Correct typos (manual/fuzzy matching).
3. **Validate & Format Numerics**
   * Convert columns to numeric dtype.
   * Remove/flag outliers.
   * Standardize units.

**3. Identifying Critical Columns:**

○**top 5 critical columns** that might be most insightful for stakeholders

T**op 5 Critical Columns & Insights**

1. PLATFORM → Reveals failure trends by vehicle type (e.g., EVs vs. trucks).

2. CAUSAL\_PART\_NM → Identifies most frequent failing parts for quality improvement.

3. REPAIR\_AGE vs. TOTALCOST → Highlights cost impact of early failures (defects vs. wear).

4. GLOBAL\_LABOR\_CODE\_DESCRIPTION → Shows most common repairs for service optimization.

5. STATE → Uncovers geographic failure patterns (e.g., cold-weather states).

**○ Provide reasoning for your selection.**

Combines failure causes, costs, service trends, and regional insights for data-driven decisions.

○ Generate visualizations (e.g., bar plots etc) using Python to represent these

insights effectively. **(atleast 3)**

**(Done in Python Jupyter notebook)**

**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.

(DONE IN EXCEL SHEET ATTACHED -> excel file name= DA-task 2 cleaned\_tagged)

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.

**5. Summary of Tags & Insights, Plus Recommendations**

**Key Insights (Tag Distribution)**

* **Steering\_Wheel\_Issue (65%)** — Dominant issue; indicates systemic steering concerns.
* **Module\_Failure (18%)** — Significant ECU/control-unit faults; check hardware vs. firmware.
* **Loose\_Assembly (12%)** — Assembly line torque/spec lapses; tighten QC.
* **Cruise\_Control\_Fault (8%)** — Some sensor or calibration instability.
* **Leather\_Replacement (6%)** — Minor interior wear; review supplier/material quality.

**Actionable Recommendations**

* **Steering:** Launch cross-functional review of design, sensors, and UI feedback.
* **Modules:** Perform firmware audits, life tests, and ensure proper ECU calibration.
* **Assembly:** Add torque checkpoints and digital torque logs to catch loose fasteners.
* **Leather:** Engage suppliers to verify specs and improve stitching/bonding quality.
* **Cruise Control:** Update manuals, infotainment prompts, and log faults for OTA diagnostics.

**Discrepancies:**

* Some key fields have missing values — e.g. CAUSAL\_PART\_NM (5), PLANT (1), STATE (2), TOTALCOST (6), CAMPAIGN\_NBR (100), and a few traceability columns (~12 each).
* **Primary keys:**  
  VIN and TRANSACTION\_ID are complete — no missing or duplicate IDs.
* **How handled:**  
  Filled missing **text fields** with 'NOT MENTIONED', imputed **numeric gaps** like TOTALCOST with median, and flagged trace columns for review. Gaps like CAMPAIGN\_NBR were left as whole column is blank so not taken into consideration.