Task 1 - Summary Report

1. Column Analysis

The data-set includes over 50 columns capturing various aspects of vehicle repair records, such as:

- Vehicle details: VIN, PLATFORM, BODY_STYLE
- Service metadata: REPAIR_DATE, PLANT, STATE, REPAIRING_DEALER_CODE
- Technical descriptions: CAUSAL_PART_NM, CORRECTION_VERBATIM, CUSTOMER_VERBATIM
- Cost and duration: TOTALCOST, LBRCOST, REPAIR_AGE, KM

Issues observed:

- Multiple columns had missing values (`CAUSAL_PART_NM`, `PLANT`, `CAMPAIGN_NBR`)
- Inconsistent text formatting in categorical fields
- Date fields were stored as strings
- Component names included extraneous details (e.g., colors, casing)

2. Data Cleaning Summary

- Column names were standardized to lowercase
- Missing values:
  - Non-critical fields filled with "unknown"
  - Rows missing `TRANSACTION_ID` (critical ID) were dropped

3. Visualizations

Top Failing Components

A bar chart showed the top 10 most frequent components involved in repairs (e.g., "WHEEL ASM-STRG", "HARNESS ASM-STRG").

4. Tag Generation & Observations

Tags were extracted from the CORRECTION_VERBATIM field using rule-based keyword matching.
New column added: **tags**

Common tags included:

- Overheating
- leakage
- wiring issue
- sensor failure
- motor fault
- battery issue

Note: These tags provide a structured summary of the free-text failure descriptions.

5. Missing Data Insights

- 6% of records had missing CAUSAL_PART_NM.
- All such entries were linked to just 3 plants: FLT, FTW, and SIL.
- Only 5 dealers were associated with these cases.
- The corresponding CORRECTION_VERBATIM field often contained hints of the actual component.

Thesis: Missing data likely stems from system-level capture issues, not technician error.

Task 1 - Conclusion

The data-set was successfully cleaned and prepared for analysis by addressing missing values, standardizing formats, and extracting tags from unstructured text. Key issues such as missing component data were traced to a small set of plants and dealers, with clues often present in free-text fields. Common repair tags included overheating, leakage, and wiring issues. Visualizations highlighted the most frequently failing components and significant variation in repair costs. These insights can support better quality monitoring, data entry validation, and automated tagging in future systems. The processed data-set is now ready for deeper diagnostic analysis or integration with related data sources.

References and links (Deliverables)

1. Problem Statement:

PO Assignment.pdf	LINK

2. Datasets:

Data For Task 1	LINK
cleaned_task1_with_tags.xlsx	LINK

3. Analysis:

task1_analysis.ipynb (Jupyter Notebook)	LINK
task1_analysis.pdf (PDF)	LINK
task1_analysis.py (Python Script)	LINK

To explore more of my work, visit my GitHub profile: https://github.com/priyanshubiswas-tech