Report: Comprehensive Data Analysis and Tagging Pipeline

Author: Naveen Dhawan

Date: 26 February, 2025

1. Introduction

This report presents the analysis and findings of the given dataset, which was shared in the **"Task2"** Excel sheet. The goal was to clean the data, analyze important columns, create meaningful tags from text data, and provide insights for better decision-making. The analysis was done using Python, and this report covers all key aspects, including data cleaning, visualization, and stakeholder recommendations.

2. Data Overview & Column-Wise Analysis

The dataset contains multiple columns with different types of information. A detailed column-wise analysis was done to understand:

- **Data types:** Whether a column contains numbers, text, or dates.
- **Unique values & distribution:** Checking for repeated values, missing data, or unusual patterns.
- **Significance for stakeholders:** Identifying which columns provide the most useful insights.

Some of the key observations include:

- CAMPAIGN_NBR: This column has 99% missing values, which means not affect the data
- Some other columns have only a few missing values means less than 5% of the total number of rows

3. Data Cleaning

- **Missing Values** :
  - Dropped **CAMPAIGN_NBR** (99% missing).
  - Imputed missing values in **PLANT**, **LAST_KNOWN_DELVRY_TYPE_CD**, and **LINE_SERIES** using **mode imputation**.

    - Filled missing **TOTAL COST** using **grouped mode/mean imputation** based on repair type and platform.

4. Critical Columns & Visual Insights

My understanding of the data is company sell some parts and if they provide a warranty then the company have to bare the cost of the repair.

Top 5 Critical Columns :

I identified the top 5 most critical columns based on their importance for insights. These columns were chosen considering factors like uniqueness, business impact, and data quality. The selected columns are:

1. **REPAIR_DATE**: Reveals seasonal spikes in repairs (e.g., January peaks).
2. **GLOBAL_LABOUR_CODE_DESCRIPTION**: Identifies frequent repairs (e.g., "Steering Wheel Replacement" accounts for more than 30% of records).
3. **PLATFORM**: Highlights "Full-Size Trucks" as the costliest platform.
4. **Symptom Condition 1**: Pinpoints "Heated" and "Dented" as top symptoms (44% of tagged issues).
5. **TOTAL COST**: Guides cost-saving priorities (e.g., "Heated" issues cost ₹19,182/month).

Note: All the suggestions depended on the units sold profitability per unit and repair cost.

5. Generating Tags/Features from Free Text

The dataset contained free text fields, which were processed to extract meaningful tags. Some of the common tags generated include:

Tags Generated from Free Text

- **Failure Modes** :
    - **Components**: Heated steering wheel, steering column, wiring harness.
    - **Issues**: Heating failure, dents, switches malfunction, rust.
- **Examples** :
    - "Heated Steering Wheel Module Replacement" tagged under **Heated** and **Failure**.

6. Key Insights & Takeaways

After analyzing the cleaned data, the following insights were drawn:

Useful Insights

- Out of 59 tagged repaired status, heated and dented were the most repaired issues. Both alone cost $23,311 (44%) of the total cost.
- 32 out of 59 cases are steering wheel replacement only because of heated and dented which cost $20,907 (39%).
- If we could solve 50% of this issue because of heat and dents, we would be able to save at least $10,000 in repair expenses each month.
- $1740 (3%) of the repair cost goes into the witches issue which if can be solved with better switches this cost can be avoided too.
- The "Full Size Trucks" platform incurs the highest costs ($24,077/month).

Actionable Recommendations :

1. **Prioritise "Heated" and "Dented" Repairs**: Investigate root causes (e.g., material defects) to reduce recurrence.
2. **Platform-Specific Quality Checks**: Focus on "Full-Size Trucks" for proactive maintenance.
3. **Enhance Component Durability**: Upgrade materials for steering wheel heaters and covers.

7. Recommendations for Stakeholders

Based on the findings, the following actions are suggested:

- **Improve Data Collection:** Standardize input methods to avoid inconsistencies in future data.
- **Monitor Key Trends:** Focus on critical factors like [mention key trend].
- **Enhance Customer Experience:** Address issues found in complaints and service delays.

8. Conclusion

The analysis highlights that "Heated" and "Dented" steering wheel issues account for **44% of total repair costs** ($23,311/month), with the "Full-Size Trucks" platform contributing 44% of expenses.

Additionally, "Switches Issue" contributes ₹1,740 (3%) to the repair expenses. Addressing these recurring problems through improved component durability and proactive quality checks could result in significant cost savings. For instance, resolving even 50% of the "Heated" and "Dented" issues could save approximately ₹10,000 monthly. Prioritising component durability upgrades and platform-specific quality checks can mitigate recurring failures, ensuring cost efficiency and customer satisfaction.

The cleaned and tagged records have been submitted in an Excel format, along with Python scripts used for the analysis.

Deliverables :

1. Updated Excel files: Handled missing values and standardised tags, and updated Excel where **I analysed the data by using Excel.** https://drive.google.com/drive/folders/1d5B5YNVhRmORdguIQMe6bysaHtvMSMG0?usp=drive_link
2. **Python Scripts**: With code for data cleaning, visualisation, and tagging. https://drive.google.com/drive/folders/1--XkW4G5eDHsEyKwi3ynqfnBf3xfErtZ?usp=drive_link
3. **Main Task Excel File**: https://drive.google.com/drive/folders/16HKFUSi4ZvcpAiINs5J3WYoDhUGeyr0T?usp=drive_link