Task 1

Summary Report – Data Tagging

Approach to Tagging Each Field

For this task, we aimed to categorize **free-text data** from the **Complaint, Cause, and Correction** columns into predefined categories provided in the **Taxonomy** sheet. Below is the step-by-step process followed for tagging each field:

1. Root Cause

- **Source:** Extracted from the **Cause** column.
- **Method:** Used **XLOOKUP with SEARCH and ISNUMBER** to find relevant keywords in the **Cause** text that matched entries in the **Root Cause category** from the Taxonomy.
- **Challenge:** Some causes contained multiple root causes; in such cases, we selected the most relevant or dominant one.

2. Symptom Condition

- **Source:** Extracted from the **Complaint** column.
- **Method:** Matched keywords from the **Symptom Condition category** in the Taxonomy against the Complaint text.
- **Challenge:** Some complaints were vague, requiring manual validation.

3. Symptom Component

- **Source:** Extracted from the **Complaint** column.
- **Method:** Similar approach as **Symptom Condition**, but focused on identifying the affected component (e.g., battery, sensor).
- **Challenge:** Some components were referenced indirectly, requiring contextual understanding.

4. Fix Condition

- **Source:** Extracted from the **Correction** column.
- **Method:** Matched with **Fix Condition category** to determine the type of correction applied.
- **Challenge:** Some corrections involved multiple steps, making it hard to assign a single fix condition.

5. Fix Component

- **Source:** Extracted from the **Correction** column.
- **Method:** Identified components that were repaired/replaced using Taxonomy keywords.
- **Challenge:** Some cases referenced indirect fixes (e.g., “updated firmware” instead of naming the exact component).

Potential Insights & Observations (Bonus Marks)

1. Common Root Causes

- **Frequent failure patterns** emerged, such as **battery overheating** and **sensor malfunctions**, indicating systemic issues in these areas.

2. Predictive Maintenance Opportunity

- Many **symptoms** and **causes** were reported **multiple times**, suggesting that AI-driven **predictive maintenance models** could be implemented to **detect early failure warnings**.

3. Fix vs. Root Cause Misalignment

- Some fixes did **not directly address** the root cause. For example, **temporary resets** were common instead of **permanent component replacements**, which could lead to recurring failures.

4. Automation Potential

- With a **well-structured taxonomy and consistent tagging**, **AI-driven text classification models** can automate this process, improving **efficiency and accuracy**.

Conclusion

By systematically tagging the dataset using **logical reasoning and Excel functions**, we identified patterns in failures and fixes. These insights can guide **engineering improvements, predictive maintenance strategies, and automation** for handling future complaints more effectively.

TASK 2

Detailed Report: Data Analysis and Insights Generation

---

a. Column Analysis

The dataset contains a variety of columns that describe vehicle repairs, their associated costs, causes, and related transactional information. A column-wise analysis was performed to understand the characteristics of each column.

1. **Data Types and Summary**:
    - **VIN** (Vehicle Identification Number): A unique identifier for each vehicle. Data type is string, and it is crucial for tracking vehicle-specific repair history.
    - **TRANSACTION_ID**: Unique ID for each transaction. Essential for tracking individual repairs.

- **REPAIR_DATE**: Date of repair. Helps in analyzing repair trends over time.
- **CAUSAL_PART_NM**: Describes the part causing the repair. This free text column contains varied values like ‘brake’, ‘engine’, and ‘battery’, providing insights into common failure conditions.
- **TOTALCOST**: The total cost of the repair. Numerical, used for cost analysis.
- **COMPLAINT_CD_CSI & COMPLAINT_CD**: Codes representing repair complaints, which help in categorizing the nature of repairs.
- **REPAIR_AGE**: Age of the vehicle when repaired. Useful for understanding how repair frequency varies with vehicle age.

2. **Unique Values and Distribution**:
   - Columns like **VIN**, **TRANSACTION_ID**, and **REPAIR_DATE** contain unique identifiers, while **CAUSAL_PART_NM** contains repeated categories.
   - **TOTALCOST** shows a range of values, highlighting significant variance in repair costs, with some repairs being much more expensive than others.

This analysis provides an overview of each column’s role and importance in understanding the dataset.

---

b. Data Cleaning Summary

Several data cleaning steps were performed to address missing values, inconsistencies, and outliers:

1. **Handling Missing Values**:
   - **Forward Imputation** was used to handle missing values (fillna(method='ffill')). This method propagates the previous valid value to fill missing data, which is suitable for time-series or transaction-based data.
2. **Categorical Data Normalization**:
   - **Textual Consistency**: Inconsistent capitalization in categorical columns (e.g., CAUSAL_PART_NM) was normalized to lowercase for consistency (str.lower().str.strip()).
3. **Numerical Columns**:
   - **Outlier Removal**: Outliers were detected and removed by calculating values beyond 3 standard deviations from the mean in numerical columns like **TOTALCOST**.

By ensuring the dataset was free of missing, inconsistent, or extreme values, the analysis was able to provide more reliable insights.

---

c. Visualizations

Several visualizations were generated to communicate key findings from the dataset:

1. **Distribution of Total Repair Cost**:

- A **histogram** with a **kernel density estimate (KDE)** was created to visualize the distribution of repair costs. This shows that most repairs are inexpensive, but some outliers (more expensive repairs) exist.

2. **Frequency of Repairs by Vehicle Age**:
   - A **count plot** was used to examine the relationship between vehicle age and repair frequency. It shows that older vehicles tend to have more repairs, indicating a higher repair frequency for aging vehicles.
3. **Frequency of Repairs by Causal Part**:
   - A **bar plot** was created to show the frequency of repairs by part type, with **brakes**, **engines**, and **batteries** being the most common causes of repairs.

These visualizations helped highlight key trends in repair frequency, cost distribution, and part-specific issues.

---

d. Generated Tags & Key Takeaways

1. **Generated Tags from Free Text**:
   - From the **CAUSAL_PART_NM** column, tags like **battery**, **brake**, **engine**, and **transmission** were identified as common terms. A **Word Cloud** was generated to visualize the frequency of these terms, revealing the most frequent repair parts.

Top Tags:

- **Battery**, **Engine**, **Brake**, **Transmission**, **Exhaust**, **Clutch**, etc.

2. **Key Takeaways**:
   - **Common Failures**: The most frequent failure parts are **batteries** and **brakes**, which suggests that regular maintenance or warranties might be beneficial for these components.
   - **Vehicle Age and Repair Frequency**: Older vehicles have a higher frequency of repairs, indicating that maintenance becomes more frequent as the vehicle ages. This insight could inform service strategies for older vehicles.
   - **Cost Analysis**: Repair costs vary significantly, with some repairs being highly expensive (especially engine-related). Stakeholders should consider pricing strategies for high-cost repairs and possibly introduce package deals for frequent repair types.
3. **Discrepancies and Approach**:
   - **Missing Primary Keys**: The dataset had missing primary keys (e.g., **VIN** or **TRANSACTION_ID**), which were handled by forward imputation.
   - **Inconsistent Data**: Textual fields were normalized to lowercase to resolve capitalization inconsistencies.

By cleaning the dataset and analyzing it with these methods, valuable insights were generated to help stakeholders improve vehicle maintenance strategies, pricing, and service offerings.

Colab Notebook

https://colab.research.google.com/drive/1I_p5ku6uXQYoby4uFW9BnGXi1zCRQUYv?usp=sharing