EDA Report: Vehicle Repair Data

By: Altamash Ansari

This report provides a streamlined overview of the Exploratory Data Analysis performed on the vehicle repair dataset. The primary goal was to uncover significant patterns, assess data quality, and derive actionable insights for stakeholders, presented in a descriptive and concise manner.

Understanding the Data Landscape

The dataset, comprising 100 entries across 52 columns, offers a rich view into vehicle repair incidents. Central to our understanding are columns like VIN (Vehicle Identification Number), which reliably identifies each unique vehicle, and the TRANSACTION_ID, marking individual repair events.

Crucially, the CUSTOMER_VERBATIM and CORRECTION_VERBATIM fields provide invaluable unstructured text, capturing the customer's direct complaint and the technician's repair description. These narratives are pivotal for grasping the nuanced nature of reported issues. Structured columns such as CAUSAL_PART_NM pinpoint the specific components responsible for failures, while GLOBAL_LABOR_CODE_DESCRIPTION outlines the repair actions taken. Financial implications are captured by TOTALCOST, and vehicle-specific trends are discernible through PLATFORM and other descriptive vehicle attributes.

Streamlining Data Quality

Initial examination revealed several data quality considerations, primarily in the form of missing values. Notably, CAMPAIGN_NBR was entirely empty, rendering it unusable and thus excluded from the analysis. Other columns, including CAUSAL_PART_NM, PLATFORM, and TOTALCOST, had a small percentage of missing entries.

Our approach to these discrepancies involved a pragmatic cleaning strategy. For categorical fields with missing data, we imputed values using the most frequent occurrence (mode), preserving their inherent distribution. For numerical fields like TOTALCOST, missing values were filled with the column's average (mean) to maintain statistical integrity. Furthermore, all textual data was standardized to lowercase and stripped of extraneous spaces, ensuring consistency and preparing it for effective text analysis. This meticulous cleaning process ensured the dataset's readiness for robust analytical exploration.

Unveiling Insights Through Visuals and Tags

While direct interactive plots are beyond this format, the analytical process heavily relied on visualizations to illuminate key trends. Bar plots vividly illustrated the most common causal parts and labor descriptions, consistently highlighting "Steering Wheel Replacement" as a predominant repair. Histograms of TOTALCOST revealed the distribution of repair expenditures, indicating that while most repairs are within a typical range, a few instances incur significantly higher costs, warranting closer inspection. Analyzing average costs by PLATFORM also provided insights into potential platform-specific vulnerabilities or cost drivers.

To further distill information from the free-text fields, a tag generation process was employed. By cleaning and analyzing the CUSTOMER_VERBATIM and CORRECTION_VERBATIM entries, we identified frequent terms related to components (e.g., 'wheel', 'steering', 'engine') and conditions (e.g., 'noise', 'leak', 'failure'). These tags effectively summarize the qualitative data, allowing for rapid

identification of prevalent issues and components. For instance, a high co-occurrence of 'steering' and 'noise' tags immediately signals a common problem area.

Key Takeaways:

The analysis consistently points to **steering-related issues** as a significant concern, evident from both the frequency of CAUSAL_PART_NM and the generated text tags. The **financial impact of repairs varies widely**, with a few high-cost outliers suggesting complex or severe failures. Furthermore, understanding the **distribution of repairs across different vehicle platforms** is crucial for targeted engineering and quality improvements. The ability to extract meaningful tags from customer and correction narratives bridges the gap between unstructured feedback and actionable insights, revealing the true voice of the customer and the nature of repairs.

Actionable Recommendations

Based on these findings, we recommend the following for stakeholders:

- **Prioritize Steering System Investigation:** Given the recurring nature of steering wheel replacements and related verbatim, a deep dive into the underlying causes of these issues is paramount. This could involve engineering reviews, supplier quality checks, and detailed failure analysis.
- **Analyze High-Cost Repairs:** Investigate the factors contributing to unusually high repair costs. Understanding these outliers can reveal critical failure modes, inefficient repair processes, or specific component vulnerabilities that drive up expenses.
- **Leverage Textual Data for Proactive Measures:** Continuously monitor and analyze customer and correction verbatim to identify emerging issues or subtle patterns not captured by structured codes. This "voice of the customer" can inform proactive maintenance campaigns, product improvements, and service training.
- **Tailor Strategies by Platform:** Utilize platform-specific insights to develop targeted maintenance schedules, part inventory management, and engineering improvements. Issues endemic to certain platforms require specialized attention.

This concise report aims to provide a clear, descriptive understanding of the dataset's core insights, enabling stakeholders to make informed decisions for product quality, service optimization, and cost management.