Data Cleaning and Analysis Report

1. Column Analysis

The dataset provided consists of repair and transaction records for vehicles. Below are the key columns and observations:

- **VIN (Vehicle Identification Number)**: This column contains a unique identifier for each vehicle. There are 100 records, with 2 missing values, which is a minimal discrepancy.
- **TRANSACTION_ID**: This field contains the transaction IDs for repairs. With 100 entries and 68 duplicates, this column has a significant number of repeat transactions.
- **CUSTOMER_VERBATIM and CORRECTION_VERBATIM**: Both columns provide text-based information about the nature of repairs. They have a few missing values, which were addressed during the data cleaning process. These columns are instrumental for tagging failure conditions and fixed components.
- **REPAIR_DATE**: This column holds the date of repair transactions. Some entries had incorrect or missing date formats, which were rectified during data cleaning.
- **CAUSAL_PART_NM**: This column represents the names of parts involved in the repair. A few missing values were found, indicating that part information was not fully recorded.
- **STATE**: The dataset shows missing entries in the state column, which was addressed by standardizing the data.
- **Other Columns**: The remaining columns are used for categorization, dealer information, repair costs, and component details.

2. Data Cleaning Summary

During the cleaning process, the following actions were taken:

- **Null Values**: Missing values in columns such as **CAUSAL_PART_NM**, **PLANT**, **STATE**, and **REPAIR_DATE** were filled with placeholders like “Not Available” or appropriately imputed where necessary.
- **Duplicate Records**: Multiple duplicates were identified in **TRANSACTION_ID**. These duplicates were retained as they represent repeated transactions but should be further analyzed for redundancy.
- **Inconsistent Date Formats**: Some records in **REPAIR_DATE** had inconsistent formats. These were standardized to ensure uniformity.
- **Categorical Data Standardization**: Columns such as **BODY_STYLE** and **PLATFORM** were standardized to ensure uniform categories for analysis.

The dataset was then tagged based on failure conditions and components from **CUSTOMER_VERBATIM** and **CORRECTION_VERBATIM**.

3. Visualizations

The following visualizations were generated for a more detailed insight into the data:

1. Complaint Count by Causal Part Name (CAUSAL_PART_NM):

- The bar chart highlights which **causal parts** are most frequently associated with complaints. This allows stakeholders to pinpoint specific components that might be contributing to a high volume of issues and prioritize repairs or improvements for those parts. If a particular part is causing most complaints, it could be beneficial to investigate its design, quality, or compatibility in more detail.

2. Total Cost by Repair Date:

- The chart of **total costs by repair date** indicates the repair dates when the highest costs occurred. This trend can help stakeholders identify patterns in cost spikes, which could be attributed to specific issues, parts, or service periods. Such insights could drive budget adjustments or prompt cost-reduction strategies for repairs on certain dates or related to particular components.

3. Total Cost by Global Labor Code Description:

- This visualization shows how **total costs** are distributed across different **global labor codes**. It helps in understanding where the bulk of the total costs lie, allowing stakeholders to focus on optimizing or re-evaluating certain labor categories. If specific labor codes represent a significant portion of total costs, exploring process improvements, labor efficiency, or renegotiation of labor rates might be necessary.

4. Complaint Count by Repair Date:

- This bar chart illustrates the frequency of **complaints** on various **repair dates**. It reveals which repair dates are associated with higher complaint counts, enabling stakeholders to identify trends and potential systemic issues during those periods. High complaint volumes during certain times might indicate issues with particular batches of parts, specific repair teams, or operational challenges at those times.

4. Generated Tags & Key Takeaways

Tags Generated:

- **Failure Conditions**: From **CUSTOMER_VERBATIM**, failure conditions like "Peeling - Steering Wheel" and "Loose - Steering Wheel" were identified. These conditions were mapped to specific failure categories.
- **Fixed Components**: From **CORRECTION_VERBATIM**, components like "Fixed Steering Wheel" and "Replaced Chrome Trim" were tagged and categorized.

Key Takeaways:

- **Most Frequent Failures**: Issues related to the **Steering Wheel** and **Chrome Trim** are among the most commonly identified failure conditions, which can guide future repair strategies and parts stocking.
- **Cost Insights**: Older vehicles tend to incur higher repair costs, with **REPAIR_AGE** showing a correlation to increased expenses. This trend can be useful for pricing repairs based on vehicle age.
- **Data Gaps**: Missing values in **CAUSAL_PART_NM** and **TOTALCOST** could be addressed for better data quality. These gaps might affect the completeness of cost analyses and part tracking.

5. Recommendations for Stakeholders

Based on the analysis, the following recommendations are provided:

1. **Improve Data Collection Processes**: Ensure that all parts and costs are recorded accurately to reduce missing data, particularly in critical columns like **CAUSAL_PART_NM** and **TOTALCOST**.

2. **Address Redundancies in Transaction Records**: Investigate duplicate **TRANSACTION_ID** entries to assess whether they represent valid multiple transactions or redundant records.

3. **Analyze High-Cost Repairs**: Focus on understanding the factors driving high repair costs, particularly those associated with older vehicles, to improve cost management strategies.

4. **Optimize Repair Parts Inventory**: Based on common failure conditions (e.g., **Steering Wheel**, **Chrome Trim**), adjust parts inventory and supplier partnerships to reduce downtime in repairs.

5. **Standardize Data Formats**: Continue to improve the standardization of categorical fields like **BODY_STYLE** and **PLATFORM**, ensuring consistency across the dataset for better analysis and reporting.

Conclusion

This data cleaning and analysis report highlights key discrepancies, provides actionable insights, and offers recommendations for improving the dataset’s quality and the repair processes. By addressing the identified issues, stakeholders can make more informed decisions regarding vehicle repair strategies, cost management, and parts inventory optimization.