**Data Summary**

**Dataset Overview:** The dataset contains various columns, including categorical and numerical features. The goal of this analysis was to clean, explore, and derive meaningful insights to help stakeholders make informed decisions.

**Data Cleaning and Column Management**

**Dropped Columns:**

* CAMPAIGN\_NBR: Completely empty, hence removed.
* TRANSACTION\_CNTR: Contains only zeros, offering no variability, so it was removed.
* COMPLAINT\_CD\_CSI: Contains only ones, so it was also removed.
* DEALER\_REGION: Dropped due to a perfect correlation (1) with SALES\_REGION\_CODE, making it redundant.

**Columns with Strong Correlation:**

TOTALCOST and REPORTING\_COST have a very high correlation (0.97), indicating they follow the same pattern. Therefore, one of them can be considered for removal in the future to reduce multicollinearity.

**Outliers:**

Outliers were detected in the numerical columns, but due to the small size of the dataset, they were retained to preserve data integrity.

**Key Observations from Categorical Columns**

**BUILD\_COUNTRY:**

US has the highest frequency (73), followed by MX (23) and CA (4), indicating a skewed distribution toward US. This imbalance may need to be addressed in future modeling tasks.

**OPTN\_FAMLY\_CERTIFICATION**:

The values are mostly concentrated in FE9 (72), followed by NE1 (19) and YF5 (9), suggesting that FE9 dominates the dataset. Similar to BUILD\_COUNTRY, this imbalance should be considered when using this feature in predictive models.

**GLOBAL\_LABOR\_CODE\_DESCRIPTION**:

The frequency analysis showed that categories like Steering Wheel Replacement and Steering Wheel Spoke Cover Replacement dominate the dataset. Other categories, like Steering Wheel Horn Switch Wiring Harness Replacement, had significantly fewer occurrences.

**Visualizations and Distribution Insights**

**Histograms and Boxplots:**

Various numerical columns were analyzed using histograms and boxplots to assess their distributions and detect outliers. Outliers were kept due to the small dataset size.

**Bar Plots:**

Categorical columns, such as BUILD\_COUNTRY and OPTN\_FAMLY\_CERTIFICATION, were visualized using bar plots to understand the frequency distribution of values.

**Recommendations and Future Steps**

**Handling Imbalanced Data:**

The imbalance in categorical columns like BUILD\_COUNTRY and OPTN\_FAMLY\_CERTIFICATION could be addressed through resampling techniques (e.g., oversampling or under sampling) in the future.

**Outlier Treatment:**

Although outliers were kept in this analysis due to the small dataset, future larger datasets may warrant techniques like trimming, Winsorizing, or log transformations to handle outliers.

**Feature Selection:**

For predictive modeling, it's recommended to select critical columns like TOTALCOST and REPORTING\_COST, which are highly correlated, and remove redundant features.

**Additional Exploration**:

Further exploration of relationships between numerical and categorical columns, along with a deeper analysis of missing values, would help improve the dataset quality for future analyses.