Report Summary

1. Column Analysis
The dataset consisted of **52 columns and 100 records** (before outlier removal). These included:

- **Text fields**: CUSTOMER_VERBATIM, CORRECTION_VERBATIM
- **Dates**: REPAIR_DATE
- **Categorical**: STATE, ENGINE, PLATFORM, TRANSMISSION
- **Numerical**: REPAIR_AGE, KM, TOTALCOST, LBRCOST, etc.

Critical identifiers like VIN and TRANSACTION_ID were treated as primary keys.

2. Data Cleaning Summary
- **Missing Values**: Addressed using:
  Deletion for rows with too many missing critical fields
  Simple imputation (like "Unknown") for non-critical categorical fields
- **Inconsistencies**:
  Standardized categorical columns using .str.lower() and .str.strip()
  Corrected typos and formatting (e.g., capitalization)
- **Outlier Removal**:
  Outliers removed from key numerical fields (like KM, REPAIR_AGE, TOTALCOST, LBRCOST) using the **IQR method**, which resulted in a cleaner dataset of **69 records**.

3. Visualizations
- **Top 10 types of Complaints**
  One complaint was the most frequent one resulting in identifying root cause of QA defects or Manufacturing anomalies
- **Top 10 States by Repair Volume**
  Bar chart showing geographic distribution. Post-outlier removal, states like FL and OH had higher representation.
- **Average Cost Repairs by State**
  Gave insights into which state had the highest average of Total Cost in repairs.

4. Tags Generated (From Free Text Fields)
Tags were extracted from CUSTOMER_VERBATIM and CORRECTION_VERBATIM using a basic keyword-matching technique. Keywords were grouped into:

- **Component Tags**: e.g., steering wheel, transmission, engine
- **Condition Tags**: e.g., not working, loose, heating, peeling

Example Tags: *"steering wheel", "heating", "replaced", "loose", "cover"*
These tags help summarize the repair in a structured format.

5. Key Takeaways & Recommendations
- **Tag-based Insights**:
  Steering-related complaints are the most common.
  Heating and cosmetic issues dominate failures.
  Many replacements happened without error codes, implying possible quality gaps.

- **Recommendations**:
  Perform root cause analysis for frequently tagged components.
  Enhance pre-delivery checks for cosmetic/comfort features.
  Consider sentiment analysis on verbatim fields for early failure signals.

- **Discrepancies Found**:
  Foreign language text – considered for translation in future iterations.
  Encoding issues – handled using utf-8 and cleaned during preprocessing.
  Multiple representations for the same issue – normalized via tagging.

Deliverables Summary
- Cleaned file with tags
- Python script used for cleaning and analysis

Use Case for Tags:
ALL_TAGS column will support in future NLP or predictive modeling tasks — for example, predicting failure types based on symptoms.