# Vehicle Service Data Analysis Report

## 1. Data Set Overview

This dataset contains 100 rows and 52 columns representing vehicle service and repair transactions collected from dealerships or authorized service centres. Each row corresponds to a unique repair event, identified by a Vehicle Identification Number (VIN). The data includes vehicle specifications, repair details, customer complaints, and service costs.

---

**2. Column-Wise Analysis**

**Numerical Columns (18):**
Examples include REPAIR_AGE, KM, REPAIR_COST, PARTS_COST, LABOR_COST, and TOTAL_COST.

- **Distribution:** These fields generally show **right-skewed distributions** (e.g., REPAIR_COST, KM) due to extreme values and outliers in high-cost repairs or unusually high-mileage vehicles.
- **Significance for Stakeholders:** Help in assessing **lifecycle trends**, identifying **costly repairs**, and optimizing **warranty coverage**.

**Categorical Columns (33):**
Fields like MODEL, PLATFORM, CAUSAL_PART_NM, REPAIR_TYPE, CITY, and DEALER_NAME.

- **Distribution:** Many categorical fields show a **long tail**—a few dominant values with many low-frequency entries (e.g., common parts or models occur more frequently).
- **Significance for Stakeholders:** Useful in **regional performance analysis**, **identifying high-risk models/platforms**, and guiding **inventory planning**.

**Free Text Columns (e.g., CUSTOMER_VERBATIM, CORRECTION_VERBATIM):**
Contains unstructured customer and technician inputs.

- **Distribution:** Rich with **qualitative insights**; can be **transformed** into structured tags using NLP.
- **Significance for Stakeholders:** Valuable for **sentiment analysis**, **identifying recurring issues**, and improving **customer service** and **quality assurance**.

---

## 3. Data Cleaning

**a. Dropping Columns:**

- The column CAMPAIGN_NBR was removed due to 100% missing values.

**b. Null Handling:**

- Columns with moderate missing values were treated based on their type:
  - Categorical values were filled based on the Columns can't use mode Option.
  - Numerical fields were treated using the Calculation of other columns, and also depending on skewness.

**c. Outlier Detection and Treatment:**

- Outliers in fields like REPAIR_COST, KM, and REPAIR_AGE were identified using IQR and Z-score methods.
- Extreme outliers were either:
  - **Removed** if there were errors.
  - **Capped** at reasonable thresholds for valid yet extreme cases.

---

## 4. Top 5 Critical Columns for Stakeholders

| Column Name | Importance to Stakeholders |
|---|---|
| VIN | Primary key for vehicle tracking, enables warranty/fraud checks. |
| REPAIR_AGE | Helps understand failure timelines and optimize warranty duration. |
| KM | Mileage-based failure analysis for predictive maintenance. |
| CUSTOMER_VERBATIM | Reveals real customer complaints for improving CX and sentiment analysis. |
| CAUSAL_PART_NM | Highlights frequent part failures, supports supply chain and quality control. |

---

## 5. Visualizations

| Visualization | Insight Derived |
|---|---|
| **Histogram of Repair Age** | Majority of issues occur within the first 3 years—suggests early warranty risk. |
| **Boxplot of Repair Cost vs. Platform** | Some platforms incur significantly higher costs—indicates design or quality issues. |
| **Bar Chart of Causal Parts** | Revealed top 10 most frequently failing parts. |
| **Heatmap of Correlation Matrix** | Identified correlations among numerical fields like KM, Repair Age, and Cost. |
| **Word Cloud for Customer Verbatim** | Visualized most common complaints and terminology used by customers. |

---

## 6. Feature/Tag Generation from Free Text Fields

From CUSTOMER_VERBATIM and CORRECTION_VERBATIM, several tags were generated using NLP techniques (e.g., keyword extraction and text clustering):

| Tag Name | Selection Criteria | Why It Was Selected |
|---|---|---|
| Engine Issue | Keywords like "stall", "overheat" | A common pain point affects drivability. |
| Transmission Fault | Mentions of "gear", "shift", "slip" | High repair cost area. |
| Noise Complaint | Phrases like "rattle", "squeak" | Often overlooked, but affects customer satisfaction. |
| Electrical Failure | Mentions of "battery", "fuse", "sensor" | Rising trend in newer vehicles. |
| Repeat Visit | Flag multiple similar complaints | Highlights quality issues or ineffective repairs. |

---

## 7. Actionable Recommendations for Stakeholders

**Product & Engineering:**

- Focus on improving part reliability for high-frequency failure items.
- Analyse common complaint tags to guide design improvements.

**Service Teams:**

- Enhance technician training based on CORRECTION_VERBATIM patterns.
- Use repair frequency by model/platform for better resource allocation.

**Customer Experience (CX):**

- Leverage CUSTOMER_VERBATIM tags for proactive outreach and sentiment tracking.
- Improve after-sales support based on recurring complaint themes.

**Warranty & Business Strategy:**

- Adjust warranty terms for platforms with high early-life repair ages.
- Incorporate mileage-based failure models for service plan pricing.

---