Task 2 – Data Analysis & Insights Report

a. Column Analysis

The dataset we worked with contains **100 records and 52 columns**. To make the data easier to understand and analyze, I first grouped the columns based on their types:

- **Numeric Columns (9 total)**:
  - *Continuous*: These include `REPAIR_AGE`, `KM`, `REPORTING_COST`, `TOTALCOST`, and `LBRCOST`. They represent values that can vary across a wide range.
  - *Discrete*: Columns like `DEALER_REGION`, `COMPLAINT_CD_CSI`, `NON_CAUSAL_PART_QTY`, and `SALES_REGION_CODE` fall into this group — they mostly contain codes or counts.
- **Categorical Columns (42 total)**:
  These include vehicle features (`BODY_STYLE`, `ENGINE_DESC`, `TRANSMISSION`), location data (`STATE`, `REPAIR_DLR_CITY`), and unique identifiers.
- **Datetime Column (1)**:
  The `REPAIR_DATE` field helps track when each transaction or complaint took place.

This classification helped guide the cleaning and visualization steps by focusing on which columns carry business meaning vs. metadata or identifiers.

b. Data Cleaning Summary

To ensure the dataset was reliable and consistent, I performed the following cleaning steps:

- **Missing Values**:
  - Most missing values were within acceptable limits (under 15%) and were filled using suitable defaults — median for numeric columns and mode or "Unknown" for categorical ones.
  - One column, `CAMPAIGN_NBR`, had all values missing and was removed from the analysis.
- **Standardizing Categorical Data**:
  Categorical fields had inconsistent capitalization and spacing. These were cleaned up by converting everything to title case and removing extra spaces, so entries like `crew cab`, `Crew  cab`, and `Crew Cab` now appear consistently.
- **Outlier Handling**:
  For numeric fields like `KM` and `TOTALCOST`, I used box plots and IQR-based methods to detect and cap extreme values that could skew results.
- **Datetime Parsing**:
  The `REPAIR_DATE` field was successfully converted to datetime format, allowing for time-based filtering and future trend analysis.

c. Visualizations

To bring the data to life, I created several visual charts using Python:

- **State-wise Repairs**: A bar chart showing which states had the highest number of service events. States like `FL`, `CA`, and `TX` topped the list.
- **Body Style Distribution**: A bar plot of different body styles revealed that `Crew Cab` and `4 Door Utility` were the most common.
- **Transaction Type**: A count plot showed most transactions were marked as `Freight`, with a smaller proportion labeled as `Freight_Policy`.
- **Cost Insights**: A histogram of `TOTALCOST` helped identify the typical repair cost range and potential outliers.
- **Mileage Patterns**: A boxplot of `KM` showed how far vehicles had been driven at the time of repair.

These visuals helped uncover patterns and pointed us toward the key problem areas.

---

d. Generated Tags & Key Takeaways

I extracted key terms from two important free-text fields: `CUSTOMER_VERBATIM` and `CORRECTION_VERBATIM`.

Customer Complaint Tags:

The most frequent words were:

- `steering`, `wheel`, `heated`, `states`, `advise`

This tells us that customers frequently raise issues related to **vehicle control and comfort**.

Technician Correction Tags:

Top words included:

- `wheel`, `steering`, `replaced`, `verified`, `module`

These tags highlight that **steering systems and wheel assemblies** are not just frequently reported but also frequently repaired or replaced — indicating a real, recurring problem.

Final Insights & Recommendations

Based on the full analysis, here are some actionable takeaways:

1. **Investigate Steering & Wheel Issues**: These appear in both customer complaints and technician notes — likely a core quality issue that needs engineering review.

2. **Improve Spare Part Availability**: Ensure quick access to steering modules, heating systems, and wheels, especially in high-volume service regions.
3. **Standardize Technician Logs**: Encourage more descriptive and consistent entries (e.g., "Steering Module Replaced") for future traceability and better AI/ML analysis.
4. **Enhance Customer Support Scripts**: Equip service advisors with quick identifiers for steering and heating-related issues so they can offer faster resolutions.