TASK-2 DATA ANALYSIS AND INSIGHTS GENERATION USING PYTHON

Data Analysis & Cleaning Report

1. Column Analysis

The dataset consists of **100 records** and **52 columns**, including:

- Categorical Fields (33): Ex- PLATFORM, BODY_STYLE, ENGINE_DESC.
- Numerical Fields (12): Ex- KM, REPAIR_AGE, TOTALCOST.
- Floating-Point Fields (6): Ex- REPORTING_COST, TRANSMISSION_SOURCE_PLANT.
- Date Field (1): REPAIR_DATE.

2. Data Cleaning Summary

- **Missing Data:**
  - CAMPAIGN_NBR was entirely null and removed.
  - Categorical columns filled with ‘UNKNOWN' where applicable.
  - Numerical columns filled using median imputation.
- **Inconsistencies:**
  - Fixed typos and capitalization mismatches.
  - Standardized text across categorical fields.
- **Outliers:**
  - Detected high repair costs ($3,000+), potentially indicating warranty-related cases.

3. Visualizations

- **Vehicle Mileage Distribution**
  - The histogram shows the distribution of vehicle mileage (KM).
  - A right-skewed pattern may indicate that most vehicles have lower mileage, while fewer have very high mileage.
  - The density curve (KDE) helps in identifying peaks in mileage frequency.
- **Repair Cost Distribution**
  - Displays the distribution of repair costs (TOTALCOST) across all records.
  - If the graph is right-skewed, it suggests that most repairs are low-cost, but some high-cost repairs exist as outliers.
  - Helps stakeholders understand common repair expenses.
- **Repair Age vs. Cost (Scatterplot)**
  - This scatter plot visualizes the relationship between repair age (REPAIR_AGE) and repair cost (TOTALCOST).

- If a trend is observed, older vehicles may require more expensive repairs.
- A random spread means no clear dependency between repair age and cost.

- **Distribution of Repair Age**
  - Histogram showing how vehicle age at the time of repair (REPAIR_AGE) is distributed.
  - Peaks in certain age ranges indicate common repair periods (e.g., cars aged 3-5 years may require frequent maintenance).
- **Correlation Heatmap**
  - Displays the correlation between numerical features.
  - Strong positive correlations (red areas) indicate direct relationships, while negative correlations (blue areas) show inverse trends.
  - Helps in identifying which variables impact repair costs the most

4. Generated Tags & Key Takeaways

Tags Beyond Failure Conditions & Components:

- ELECTRICAL_ISSUE → Battery, sensors, ECU failures.
- SAFETY_CONCERN → Brakes, airbags, or potential hazards.
- PERFORMANCE_PROBLEM → Engine stalling, power loss.
- COMFORT_FEATURES → Seat heating, infotainment, climate control.
- RECURRING_ISSUE → Repeat customer complaints.

Key Takeaways & Recommendations:

Proactive Maintenance: Identify high-mileage vehicles for preventive service.

Warranty Review: Optimize policies for **20+ month-old vehicle repairs**.

Customer Satisfaction: Address recurring complaints promptly.

Safety Focus: Prioritize brake & airbag failures for recalls.

Conclusion:

The analysis highlights key repair trends, cost patterns, and common vehicle issues. Most repairs are low-cost, but high-cost outliers suggest warranty claims or expensive fixes. Older vehicles may require costlier repairs, emphasizing the need for proactive maintenance. Recurring complaints and safety concerns, such as brake and airbag issues, should be prioritized. These insights can guide warranty optimizations, customer service improvements, and preventive maintenance strategies.