Task 2 - Data Analysis & Insights Report

1. Column Analysis

The dataset contained **52 columns**, covering various aspects such as transaction details, customer complaints, repair actions, costs, and vehicle specifications. Key observations:

- **Data Types:** Mix of categorical, numerical, text, and date fields.
- **Missing Values:** Some columns, like `CAMPAIGN_NBR`, had 100% missing values and were dropped, while others like `ENGINE_TRACE_NBR` had partial missing data.
- **Unique Identifiers:** `VIN` and `TRANSACTION_ID` ensured record uniqueness.
- **Significant Columns for Analysis:** `CUSTOMER_VERBATIM`, `CORRECTION_VERBATIM`, `TOTALCOST`, `PLATFORM`, and `REPAIR_DATE` were selected for deeper insights.

2. Data Cleaning Summary

- **Dropped Columns:** Removed columns with excessive missing values (e.g., `CAMPAIGN_NBR`).
- **Filled Missing Values:**
  - Categorical values were replaced with the most frequent category.
  - Numerical values (e.g., `TOTALCOST`) were imputed using the median.
- **Standardization:** Converted text fields (e.g., `PLATFORM`, `BODY_STYLE`) to uppercase.
- **Outlier Handling:** Applied **IQR (Interquartile Range)** to remove extreme values in numerical fields.
- **Duplicate Removal:** Ensured no redundant records.

3. Key Visualizations & Insights

- **Repair Cost Distribution:** Most repairs cost between **$200-$500**, but some exceed **$3,000**, suggesting high-cost repair cases.
- **Most Repaired Vehicle Platforms: Full-Size Trucks** had the highest number of repairs, indicating potential reliability issues.
- **Repair Trends Over Time:** Spikes in repair costs suggest **seasonal effects or recall-related repairs**.

4. Generated Tags from Free-Text Data

From **`CUSTOMER_VERBATIM`** (customer complaints):

- **Common issues:** "steering failure," "heated seat not working," "sensor malfunction."

From **`CORRECTION_VERBATIM`** (repair actions):

- **Frequent fixes:** "software update," "component replacement," "realignment."

These tags help in identifying patterns in **vehicle defects and repair strategies**.

5. Discrepancies & Recommendations

Dataset Discrepancies

- **Null Values:** Missing data in `ENGINE_TRACE_NBR` and `TRANSMISSION_TRACE_NBR` was filled using mode.
- **Outliers:** High-cost repairs removed using statistical filtering.
- **Inconsistent Formats:** Standardized categorical values for uniformity.

Actionable Recommendations

- **Investigate High-Cost Repairs:** Identify recurring issues driving costs above **$3,000**.
- **Quality Control for Full-Size Trucks:** Address frequent repair cases to improve reliability.
- **Automate Issue Tagging:** Use generated tags for **early fault detection** and **predictive maintenance**.

Key Challenges & Solutions Implemented

Key Challenges:

- **High Missing Values:** Some categorical and numerical fields had significant gaps.
- **Inconsistent Data Formats:** Text fields had different capitalizations, and dates were in mixed formats.
- **Outliers in Repair Costs:** Some repair costs were unusually high, skewing analysis.
- **Extracting Insights from Free-Text Data:** The unstructured nature of customer complaints made pattern recognition difficult.

Solutions Implemented:

- **Systematic Missing Value Handling:** Used mode for categorical values and median for numerical values.
- **Data Standardization:** Converted categorical fields to uppercase and reformatted dates for consistency.
- **Outlier Removal:** Applied IQR filtering to improve analysis accuracy.
- **Text Processing for Insight Extraction:** Identified frequent terms in customer complaints and repair descriptions to generate structured tags.

7. Conclusion

This analysis provides valuable insights into repair trends, cost patterns, and failure conditions. These findings can help optimize maintenance, reduce costs, and enhance vehicle reliability. Key takeaways include:

- **Frequent high-cost repairs** should be investigated further to understand root causes and mitigate unnecessary expenses.
- **Full-Size Trucks require additional quality control measures** due to their high repair rates.
- **Automated issue tagging** can streamline fault detection and improve predictive maintenance.