TASK 2: DATA ANALYSIS AND INSIGHTS GENERATION USING PYTHON

Submit the cleaned and tagged records in CSV or Excel format.

EXCEL LINK- Task2FinalOutput

Submit a detailed report covering the above pointers, including (Maximum 2 page):

- Column analysis
- Data cleaning summary and
- Visualizations
- Generated tags & Key takeaways

Column_Wise Analysis:

Field Name	Description / Insight
VIN	Few duplicates (indicating multiple repairs on the same unit), 98 unique rows
Transaction_Id	Multiple records are linked to the same transactions
Customer_verbatim	Unstructured text
Correction_verbatim	Unstructured text
Repair_Date	Ranges from 2 Jan to 7 Feb 2024
Casual_part_nm	Issues related to Steering Wheel Assembly
Campaign_nbr	No campaign repairs recorded
Repair_age	Mean = 14.8 months, within warranty period
Total_Cost	Mean = 561

Data Cleaning Summary:

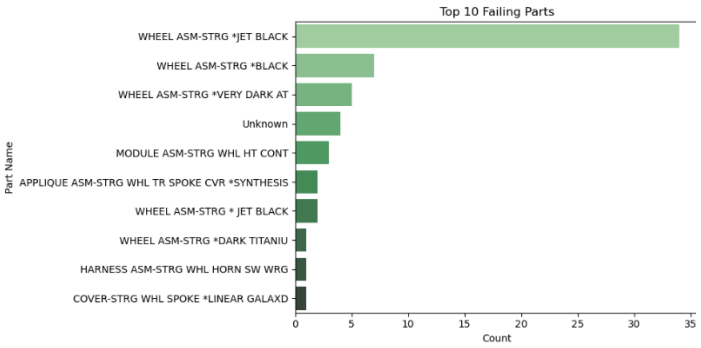
Missing Values Treatment

- Missing values in key categorical fields such as:
  - CAUSAL_PART_NM, PLANT, STATE, VEH_TEST_GRP,
  - OPTN_FAMLY_CERTIFICATION, OPTF_FAMLY_EMISSIOF_SYSTEM,
  - ENGINE_SOURCE_PLANT, ENGINE_TRACE_NBR,
  - TRANSMISSION_SOURCE_PLANT, TRANSMISSION_TRACE_NBR,
  - LINE_SERIES

  were imputed with the label 'Unknown' to maintain data integrity for categorical grouping.
- TOTALCOST:
  - Missing values were filled with the column median, not zero, to avoid skewing cost-based analysis.
  - Zero values were kept, assuming they may reflect genuine warranty or free services.
- Fields like REPAIR_DLR_POSTAL_CD and LAST_KNOWN_DELVRY_TYPE_CD were filled with 0, assuming placeholder numeric consistency.
- All textual data (e.g., in CORRECTION_VERBATIM, CUSTOMER_VERBATIM, and dealer/location names) was converted to lowercase, punctuation removed, and extra whitespace trimmed.
- CORRECTION_VERBATIM text was normalized using .str.lower().strip() to support consistent NLP analysis and tag extraction.
- Values like '5' and 'K' in ENGINE_SOURCE_PLANT were mapped to 'Unknown', as they didn't align with valid source plant codes.
- The same field was cleaned further to standardize both numeric and textual identifiers.
- Rows 11, 24, 37, 46, 50, and 51 were dropped entirely due to excessive null values in crucial technical identifiers:
  ENGINE_SOURCE_PLANT, ENGINE_TRACE_NBR, TRANSMISSION_SOURCE_PLANT, and TRANSMISSION_TRACE_NBR.

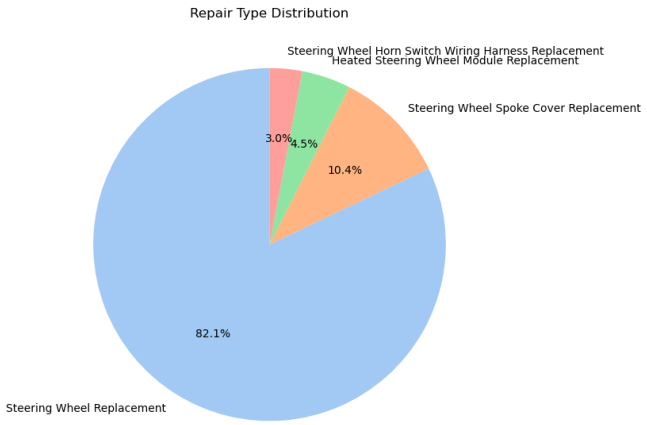
- Outliers in key numeric fields were removed using the IQR method to avoid bias in cost and usage analytics. Cleaned columns included:
    - REPAIR_AGE
    - KM
    - TOTALCOST
    - LBRCOST
    - REPORTING_COST
    - NON_CAUSAL_PART_QTY

Data Visualization



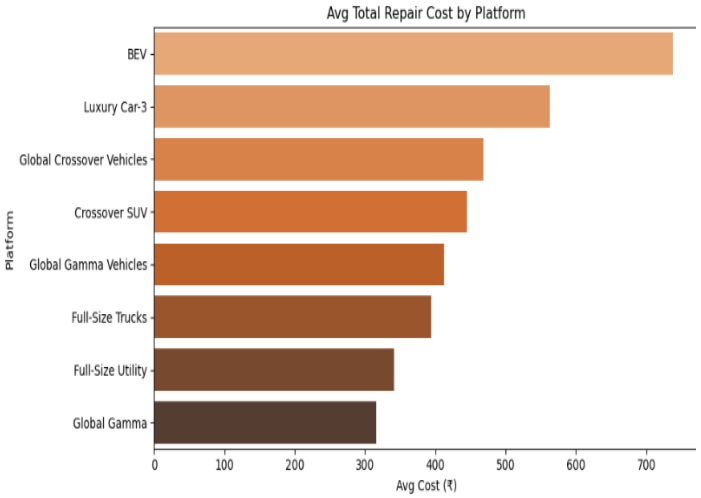
Top 10 Failing Parts (Bar Chart)

- Most common failing part: WHEEL ASM-STRG 3F1T BLACK.
- Steering-related parts dominate the top failures (e.g., wheel assemblies, harnesses, modules).
- Indicates a systemic issue with steering components.

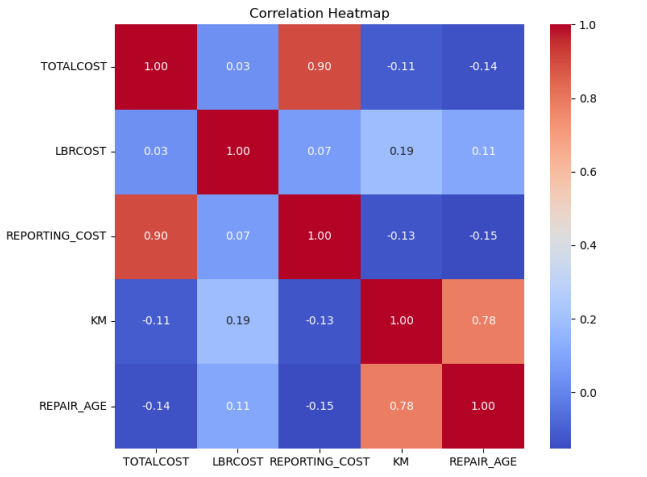


Repair Type Distribution (Pie Chart)

- Dominant repair: Steering Wheel Replacement.
- Others: Spoke Cover Replacement (10.4%), Module Replacement, and Wiring Harness.
- Steering system failures are the major issue across platforms.



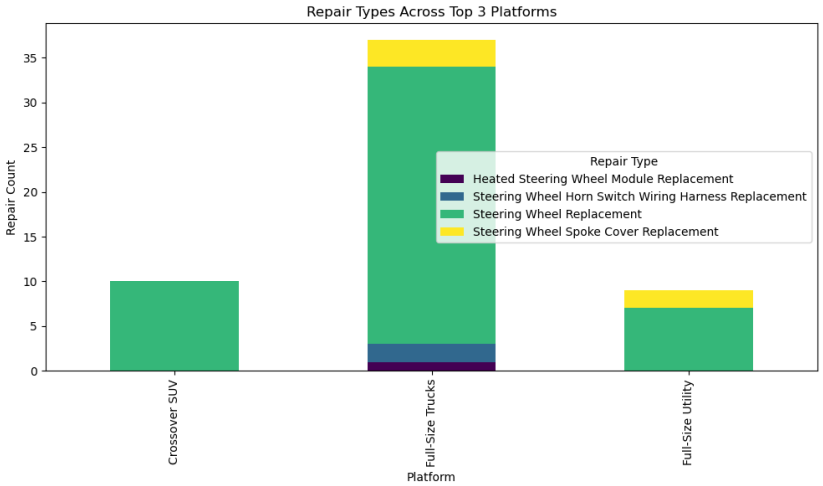
- Highest average repair cost: Battery Electric Vehicles (BEVs) and Luxury Cars.
- Lowest: Global Gamma platforms.
- Insight: High-tech or luxury platforms incur higher costs, likely due to part prices and complexity.



Correlation Heatmap

Shows relationships among various repair-related metrics:

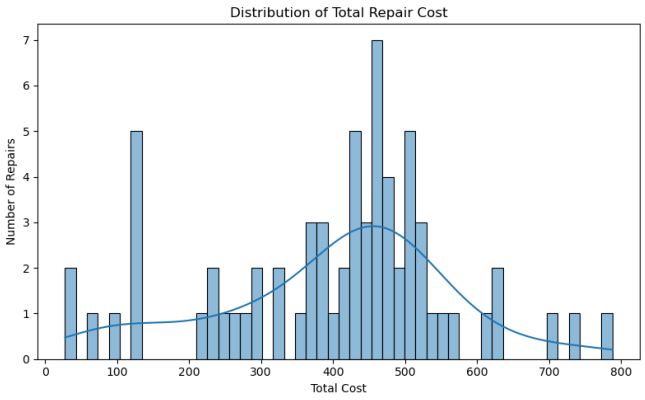
- KM and Repair Age (0.78): Older vehicles have higher mileage.
- Repair Age vs. Total Cost: Weak correlation (0.14), suggesting age isn't a strong cost driver.



Repair Types Across Top 3 Platforms (Stacked Bar Chart)

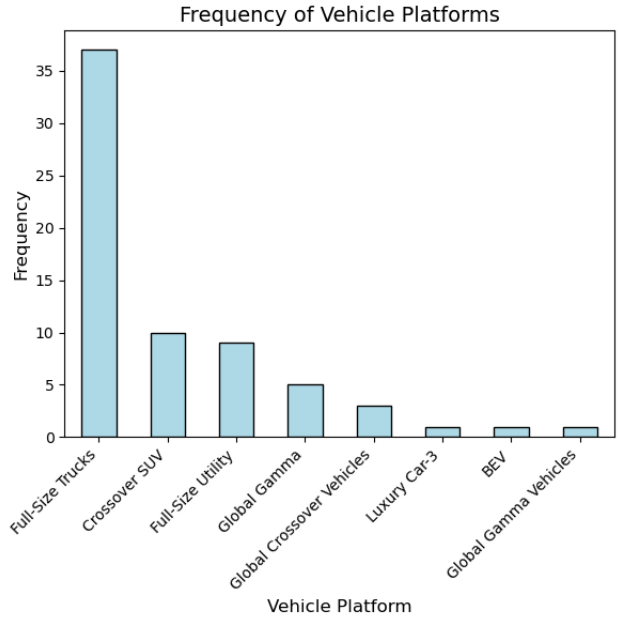
Platforms: Crossover SUV, Full-Size Truck, Full-Size Utility.

- Full-Size Trucks have the highest repair count, mainly steering wheel replacement.
- Full-Size Utility and Crossover SUV follow with more variety (like spoke covers and module replacements).



Distribution of Total Repair Cost (Histogram)

- Shows a bell-shaped curve (somewhat normal), centered around 400–500.
- Most repairs cost between 300–600.
- There's a wide spread, indicating variable repair complexities.



Frequency of Vehicle Platforms (Bar Chart)

- Full-Size Trucks dominate the data (35+ records).
- Followed by Crossover SUV and Full-Size Utility.
- Least common: BEVs, Global Gamma Vehicles, and Luxury Cars.

Generated tags & Key takeaways

We extracted three types of tags from the CUSTOMER_VERBATIM and CORRECTION_VERBATIM fields using GPT-based text parsing:

1. Symptom / Condition
    - Capturees the issue faced by the customer or identified by the technician.
    - Examples:
        - *Steering wheel coming apart*
        - *Horn and switches inoperable*
        - *Heated steering wheel not functioning*
2. Component / Part
    - Identifies the vehicle part or subsystem related to the complaint.
    - Examples:
        - *Steering wheel*
        - *Wire harness*
        - *Spoke cover*
3. Action Taken / Fix
    - Documents the repair performed to resolve the issue.
    - Examples:
        - *Replaced steering wheel*
        - *Adjusted trim and retested*
        - *Removed and replaced harness*

PS:
Common repairs like steering wheel replacement or harness replacement indicate repetitive issues possibly related to design quality. The consistent language in responses suggest we should standarised codes/ actions for easier reporting.Most Cases were related to steeering wheel indicating a frequently failing component.Frequent symptoms like steering noise can be flagged to quality teams for design review.This system could be automated to feed into warranty claims dashboards or predictive models.

Python_File: Task2

By, Diya Haldankar