**Report**

**CASE- 2**

**Column-Wise Analysis**

**1. Column Data Types:**

* **VIN:** Alphanumeric, unique identifier for vehicles.
* **TRANSACTION\_ID:** Alphanumeric, unique transaction reference.
* **CORRECTION\_VERBATIM & CUSTOMER\_VERBATIM:** Text, unstructured feedback from technicians and customers.
* **REPAIR\_DATE:** Date, indicates the date of repair.
* **CAUSAL\_PART\_NM:** Text, names of parts causing issues.
* **GLOBAL\_LABOR\_CODE\_DESCRIPTION:** Text, description of labor codes.
* **PLATFORM, BODY\_STYLE, VPPC:** Categorical, vehicle-specific attributes.
* **PLANT, BUILD\_COUNTRY:** Text, manufacturing details.
* **DEALER\_NAME, DEALER\_CITY, STATE, REGION:** Categorical, dealer details.
* **REPAIR\_AGE, KM:** Numeric, vehicle age and mileage.
* **COST FIELDS (e.g., REPORTING\_COST, TOTALCOST, LBRCOST):** Numeric, cost-related metrics.
* **TRANSACTION\_CATEGORY, MEDIA\_FLAG:** Categorical, transaction classification and media flags.

**2. Unique Values and Distribution:**

* High cardinality: VIN, TRANSACTION\_ID.
* Categorical fields (e.g., BODY\_STYLE, REGION) show well-defined groupings.
* Numeric fields (e.g., KM, COSTS) have skewed distributions indicating outliers.
* Free text fields (VERBATIMs) show wide variation, needing summarization.

**3. Overall Significance:**

* **VIN and TRANSACTION\_ID:** Key for identifying records.
* **REPAIR\_DATE and AGE/KM:** Crucial for time-based analysis.
* **COST FIELDS:** Direct impact on financial metrics.
* **VERBATIM FIELDS:** Provide actionable insights into recurring issues.

**Data Cleaning**

**1. Handling Missing Values:**

* **VIN/TRANSACTION\_ID:** Records with missing values dropped.
* **COST FIELDS:** Imputed with median values.
* **TEXT FIELDS:** Missing entries flagged for review.

**2. Addressing Inconsistencies:**

* Standardized categorical values (e.g., dealer names).
* Corrected date formats and ensured numeric columns were free from strings.

**3. Outlier Treatment:**

* Applied interquartile range (IQR) method for numeric columns (e.g., KM, COSTS).

**Identifying Critical Columns**

**Selected Columns:**

1. **REPAIR\_AGE:** Age impacts failure likelihood.
2. **KM:** Indicates usage levels.
3. **COST FIELDS:** Highlight financial impact.
4. **VERBATIM FIELDS:** Summarize recurring issues.
5. **PLATFORM/BODY\_STYLE:** Identify trends by model.

**Visualizations**

1. **Histogram**
2. **Bar Plot**
3. **Line chart**
4. **Scatter plot**

**Generating Tags from Free Text**

**Methodology:**

* Applied Natural Language Processing (NLP) techniques.
* Extracted terms related to "failure conditions" and "causal components."
* Grouped similar issues using clustering algorithms.

**Example Tags:**

* "Engine Overheating"
* "Brake Failure"
* "Electrical Malfunction"

**Summary and Insights**

**Key Insights:**

* Frequent failures in components such as brakes and engines.
* High repair costs associated with older vehicles (>5 years).
* Discrepancies in reporting costs between regions.

**Actionable Recommendations:**

1. Focus on preventive maintenance for high-failure components.
2. Standardize repair reporting processes.
3. Launch targeted customer education for older vehicles.

**Dataset Discrepancies:**

* Null values in numeric columns addressed via imputation.
* Inconsistent text formatting resolved with normalization scripts.

**Report Summary**

1. **Column Analysis:** Identified critical columns and their significance.
2. **Data Cleaning:** Handled missing values, resolved inconsistencies.
3. **Visualizations:** Highlighted repair age, costs, and text-based insights.
4. **Generated Tags:** Provided structured summaries of unstructured feedback.
5. **Recommendations:** Proposed actionable strategies to improve operational efficiency.