Task 1: Data Validation, Cleaning, and Exploratory Analysis Report

1. Column-Wise Analysis

Dataset Overview

The dataset consists of **100 records and 54 columns**, capturing vehicle repair transactions, dealer information, cost metrics, manufacturing attributes, and unstructured text describing customer complaints and corrective actions.

The data types includes **categorical, numerical, datetime, and free-text fields**

#	Column	Non-Null Count	Dtype
0	VIN	100 non-null	object
1	TRANSACTION_ID	100 non-null	int64
2	CORRECTION_VERBATIM	100 non-null	object
3	CUSTOMER_VERBATIM	100 non-null	object
4	REPAIR_DATE	100 non-null	datetime64[ns]
5	CAUSAL_PART_NM	95 non-null	object
6	GLOBAL_LABOR_CODE_DESCRIPTION	100 non-null	object
7	PLATFORM	100 non-null	object
8	BODY_STYLE	100 non-null	object
9	VPPC	100 non-null	object
10	PLANT	99 non-null	object
11	BUILD_COUNTRY	100 non-null	object
12	LAST_KNOWN_DLR_NAME	100 non-null	object
13	LAST_KNOWN_DLR_CITY	100 non-null	object
14	REPAIRING_DEALER_CODE	100 non-null	object
15	DEALER_NAME	100 non-null	object
16	REPAIR_DLR_CITY	100 non-null	object
17	STATE	98 non-null	object
18	DEALER_REGION	100 non-null	int64
19	REPAIR_DLR_POSTAL_CD	98 non-null	object
20	REPAIR_AGE	100 non-null	int64
21	KM	100 non-null	int64
22	COMPLAINT_CD_CSI	100 non-null	int64
23	COMPLAINT_CD	100 non-null	object
24	VEH_TEST_GRP	98 non-null	object
25	COUNTRY_SALE_ISO	100 non-null	object
26	ORD_SELLING_SRC_CD	100 non-null	int64
27	OPTN_FAMLY_CERTIFICATION	90 non-null	object
28	OPTF_FAMLY_EMISSIOF_SYSTEM	95 non-null	object
29	GLOBAL_LABOR_CODE	100 non-null	int64
30	TRANSACTION_CATEGORY	100 non-null	object
31	CAMPAIGN_NBR	0 non-null	float64
32	REPORTING_COST	100 non-null	float64
33	TOTALCOST	100 non-null	float64
34	LBRCOST	100 non-null	float64
35	ENGINE	100 non-null	object
36	ENGINE_DESC	100 non-null	object

```
 37  TRANSMISSION                  100 non-null    object
 38  TRANSMISSION_DESC             100 non-null    object
 39  ENGINE_SOURCE_PLANT            88 non-null    object
 40  ENGINE_TRACE_NBR               88 non-null    object
 41  TRANSMISSION_SOURCE_PLANT      88 non-null    float64
 42  TRANSMISSION_TRACE_NBR         88 non-null    object
 43  SRC_TXN_ID                    100 non-null    int64
 44  SRC_VER_NBR                   100 non-null    int64
 45  TRANSACTION_CNTR              100 non-null    int64
 46  MEDIA_FLAG                    100 non-null    object
 47  VIN_MODL_DESGTR               100 non-null    object
 48  LINE_SERIES                    99 non-null    object
 49  LAST_KNOWN_DELVRY_TYPE_CD      98 non-null    float64
 50  NON_CAUSAL_PART_QTY           100 non-null    int64
 51  SALES_REGION_CODE             100 non-null    int64
dtypes: datetime64[ns](1), float64(6), int64(12), object(33)
```

---

2. Data Cleaning Summary

a. Handling Missing / Invalid Values

- Columns with **100% missing values** were removed.
- Columns with **low missing percentages** were imputed using **mode** (categorical) or **median** (numerical).
- Columns with **high missing rates** (e.g., traceability fields) were retained and populated with **logical placeholders** such as `UNKNOWN`

b. Categorical Consistency

- Standardized text by converting to uppercase and trimming whitespace.
- Removed hidden nulls (empty strings).
- Ensured consistent naming conventions across categorical fields.

c. Numerical Validation & Outliers

- Converted numeric columns to proper data types.
- Identified extreme outliers in cost fields using the **IQR method** and treated them to maintain analytical reliability.

This approach ensured **data completeness, consistency, and business validity** without distorting underlying patterns.

---

3. Identifying Critical Columns

Top 5 Critical Columns & Rationale

1. **TOTALCOST** – Direct indicator of warranty and repair expenditure.
2. **LBRCOST** – Major contributor to total repair cost.
3. **REPAIR_AGE** – Indicates when failures occur in the vehicle lifecycle.
4. **KM** – Helps distinguish early defects from wear-and-tear issues.
5. **CAUSAL_PART_NM** – Identifies failed components and supports root-cause analysis.

Visualizations Generated

- **Box Plot of TOTALCOST** – Shows cost distribution and outliers.
- **Bar Plot of Average TOTALCOST by DEALER_REGION** – Highlights regional cost variations.
- **Bar Plot of Top Causal Parts** – Identifies frequently failing components.

These visualizations provide **clear, stakeholder-friendly insights** into cost drivers and failure patterns.

---

4. Generating Tags / Features from Free Text

Free-text fields were processed to generate **structured tags** representing:

- **Repair actions** (e.g., replace, inspect)
- **Components** (e.g., steering_wheel, engine, transmission)

A rule-based, domain-aware approach was used to normalize variations and diagnostic messages into **canonical component tags** (e.g., any steering-related reference mapped to `steering_wheel`).

This transformed unstructured text into **machine-readable and business-interpretable features**, enabling faster analysis and trend identification.

---

5. Overall Synthesis & Key Takeaways

a. Discrepancies Identified & Resolution

- Missing values due to upstream data capture limitations were handled using logical placeholders.
- Encoding issues in text fields were cleaned to remove corrupted characters.
- Formatting inconsistencies were standardized across categorical columns.

b. Summary of Generated Tags

- Majority of issues were related to **steering wheel components**.
- Repair actions were dominated by **replace** and **inspect**, indicating part-level failures rather than minor adjustments.

c. Actionable Recommendations

- Focus quality checks on **steering wheel assemblies**, as they dominate failure trends.
- Investigate high-cost repairs linked to labor to optimize service procedures.
- Use tagged text features for **predictive maintenance and recall analysis**.

d. Additional Observations

- Cost variability is more strongly influenced by **repair complexity** than by vehicle age alone.
- Text-based tagging significantly enhances interpretability compared to raw complaint descriptions.

Conclusion

The cleaned, validated, and enriched dataset provides **reliable, actionable insights** for stakeholders. By combining structured data analysis with free-text feature generation, the analysis supports **cost optimization, quality improvement, and data-driven decision-making**