TASK 2 & 3

1. Primary Key Identification

a. Identified Primary Key

Primary Key: Primary Key
(Combination of Order No and Segment Number, e.g., SO0005588-1)

b. Justification

- Present in **both datasets**
- Uniquely identifies a **service order segment**
- Stable and suitable for integration without data loss

Potential Challenges

- Hidden formatting issues (spaces, inconsistent string formatting)
- High similarity but not exact string matches
- Requires standardization before merging

2. Data Cleaning

Data Inspection & Cleaning

i. Missing Values & Duplicates

- Optional free-text fields (Cause, Correction) contained missing values
  → Replaced with meaningful placeholders (e.g., Unknown)
- Informational fields (Coverage) standardized
- Duplicate records removed where applicable
- No missing values found in key identifier columns

ii. Format Correction (Consistent Data Types)

- Dates converted to proper datetime format (including Excel serial dates)
- Numeric fields (Cost, Revenue, Actual Hours, Segment Total) converted to numeric by removing currency symbols

- Invalid values (e.g., Model Year = 0) treated as missing
- Primary Key standardized to consistent string format

iii. Language Translation

- Multilingual free-text (German) identified in Correction and Cause
- Translated to English to ensure consistency and enable text analysis

3. Data Integration

a. Merge Approach

The datasets were merged using the identified **Primary Key** to create a comprehensive view combining:

- Work order details (equipment, complaint, failure)
- Repair and part information (fix action, cost, labor)

4. Rationale for Join Type Selection

Selected Join Type: Left Join

Why Left Join?

- Work Order data represents the **complete set of service events**
- Repair data is **conditional** (only present when parts or costs are recorded)
- Left join ensures **all work orders are retained**, even if no repair occurred

Implications of Other Join Types

- **Inner Join:** Drops valid work orders without repair data → under-reports service volume
- **Right Join:** Risks orphan repair records
- **Outer Join:** Adds complexity with limited analytical value

5. Identified Trends & Investigation

5.1 Trend Analysis (Visual Insights)

Trend 1: Failed Component vs Cost

- Electrical and cooling-related failures show **higher average repair cost**
- Indicates complex or high-impact repairs

Trend 2: Failed Component vs Actual Hours

- Cooling, alternator, and mechanical issues require **more labor hours**
- Highlights labor-intensive failure types

Trend 3: Cost vs Actual Hours

- Positive relationship observed
- Outliers distinguish **expensive parts** vs **labor-driven repairs**

5.2 Root Cause Identification

Failure Condition Trends

- High frequency of **“Not Mentioned” failures**
- Among identified issues, **hydraulic leaks** (hose, steering, axle, boom) dominate
- Cooling and suspension issues appear less frequently but are operationally significant

Fix Condition Trends

- Majority of fixes are **“No Component Mentioned”**

Root Cause Interpretation

- Many issues are **operational or diagnostic**, not component failures
- Strong indication of **documentation gaps**

Stakeholder-Level Summary

Analysis shows that most service issues are resolved without component replacement, indicating operational or transient problems rather than widespread part failures. Where components are replaced, issues are primarily related to hydraulic leaks and cooling systems. Improving diagnostic documentation and focusing preventive maintenance on hydraulic and electrical subsystems can reduce downtime, labor effort, and long-term costs.