# Vehicle Repair Data Analysis – Task 2 Report

## 1. Column-Wise Analysis

The dataset contains multiple columns related to vehicle repairs, including complaint details, vehicle information, dealer information, and cost breakdowns. I analyzed each column for data type, uniqueness, and missing values to assess their usefulness. Key columns include VIN, REPAIR\_DATE, CUSTOMER\_VERBATIM, CAUSAL\_PART\_NM, TOTALCOST, and STATE. These help identify patterns and provide value to stakeholders.

## 2. Data Cleaning Summary

Data cleaning involved formatting dates, converting numeric columns, and standardizing text. Specifically:  
 Converted 'REPAIR\_DATE' to proper date format  
 Filled missing complaints with 'Not provided'  
 Removed top 1% outliers in 'TOTALCOST' and 'KM'  
 Uppercased categorical columns like 'STATE' and 'PLANT'  
 Converted string-based numbers into numeric values

## 3. Visualizations

**Top 10 Failing Parts**: Bar chart showing most frequently replaced parts such as steering wheel assemblies.  
**Repairs Over Time**: Bar chart of repair frequency by year to detect trends.  
**Cost vs. Kilometers**: Scatter plot that shows if higher usage leads to more costly repairs.

## 4. Generated Tags from Free Text

We extracted meaningful tags from 'CUSTOMER\_VERBATIM' using keyword searches. Tags included:  
 Steering Issue  
 Heated Steering  
 Horn Problem  
 Driver Assist  
These tags summarize the main complaint themes and allow better grouping of customer issues.

## 5. Key Takeaways & Recommendations

**Insight**: Steering-related issues are the most common cause of repair.  
 **Insight**: Repairs happen even with low kilometers—indicating early part failure.  
 **Recommendation**: Investigate quality of steering wheel components across manufacturing plants.  
 **Recommendation**: Use tag-based clustering to speed up issue resolution at service centers.

## 6. Data Quality Observations

Missing values were found in date and complaint columns.  
 Outliers in cost and kilometers were removed to ensure fair analysis.  
 Inconsistent casing (e.g., 'mx', 'MX') in some columns was resolved.