**Task 2 - Data Analysis and Insights Generation**

**Neha Kumari**

## **1. Column-Wise Analysis**

The dataset was analyzed column by column to understand its structure and key attributes. Below are the observations:

* **Data Types**: The dataset contains numerical, categorical, and text-based columns.
* **Unique Values**: Some categorical fields have high cardinality, while others have repeated values.
* **Missing Values**: Several columns had missing data, with some exceeding 50%.
* **Outliers**: Certain numerical fields contained extreme values, requiring careful handling.

## **2. Data Cleaning Summary**

The dataset underwent multiple cleaning steps to ensure accuracy and consistency:

* **Handling Missing Values**:
  + Columns with more than 50% missing data were dropped.
  + Numerical columns were imputed with their median values.
  + Categorical fields were filled with "Unknown" where applicable.
* **Standardizing Categorical Data**:
  + Text fields were converted to lowercase to maintain consistency.
* **Outlier Removal**:
  + The interquartile range (IQR) method was applied to detect and remove outliers from numerical fields.

## **3. Key Visualizations and Insights**

Three key visualizations were generated to extract meaningful insights:

### **1. Distribution of Repair Costs**

* The repair cost distribution revealed a right-skewed pattern, indicating the presence of high-cost repairs.
* This suggests that cost control measures might be necessary to optimize service expenses.

### **2. Top Platforms with Most Repairs**

* A count plot of repair occurrences per platform showed that a few platforms account for a majority of the repairs.
* This insight helps stakeholders focus on platforms requiring better maintenance strategies.

### **3. Trends in Repairs Over Time**

* A time-series analysis revealed seasonal variations in repairs, with peaks in certain months.
* This finding supports inventory management and workforce allocation strategies.

## **4. Generated Tags & Key Takeaways**

Using **TF-IDF and NMF topic modeling**, meaningful tags were generated from customer feedback. These tags highlight:

* **Common Failure Conditions** (e.g., "battery issue", "engine failure")
* **Component Issues** (e.g., "brake pad wear", "transmission error")
* **Operational Problems** (e.g., "software glitch", "safety sensor fault")

## **5. Actionable Recommendations**

Based on the analysis, the following recommendations are proposed:

1. **Cost Optimization**: Address high-repair-cost categories by exploring alternative parts or service improvements.
2. **Preventative Maintenance**: Focus on platforms with frequent repairs to reduce breakdown occurrences.
3. **Seasonal Demand Planning**: Align inventory and workforce based on peak repair trends.
4. **Customer Feedback Insights**: Use generated tags to enhance predictive maintenance and service quality.

## **6. Discrepancies & Data Handling Approach**

* Missing values and outliers were carefully treated to ensure data integrity.
* Inconsistent categorical values were standardized for better analysis.
* Generated insights were validated against observed trends to improve accuracy.

### **Final Deliverables**

* **Cleaned Dataset:** cleaned\_task2\_data.csv
* **Python Script for Reproducibility**
* **Key Visualizations and Insights Report** (this document)

This structured approach ensures stakeholders gain actionable insights for improving operational efficiency and service quality.