**Axion Ray**

**TASK 1**

1. **Column-Wise Analysis**: We analyzed each column in the dataset to understand its data type, unique values, distribution, and significance for stakeholders. This step helps in identifying the structure and quality of the data, which is crucial for further analysis.
2. **Data Cleaning**: The script includes functions to: -
   1. **Handle missing values.**

**Used Mode for Categorical Data:**

* **Preserves Data Distribution**
* **Maintains Data Integrity**
* **Avoids Artificial Variability**

**Used Median for Numerical Data:**

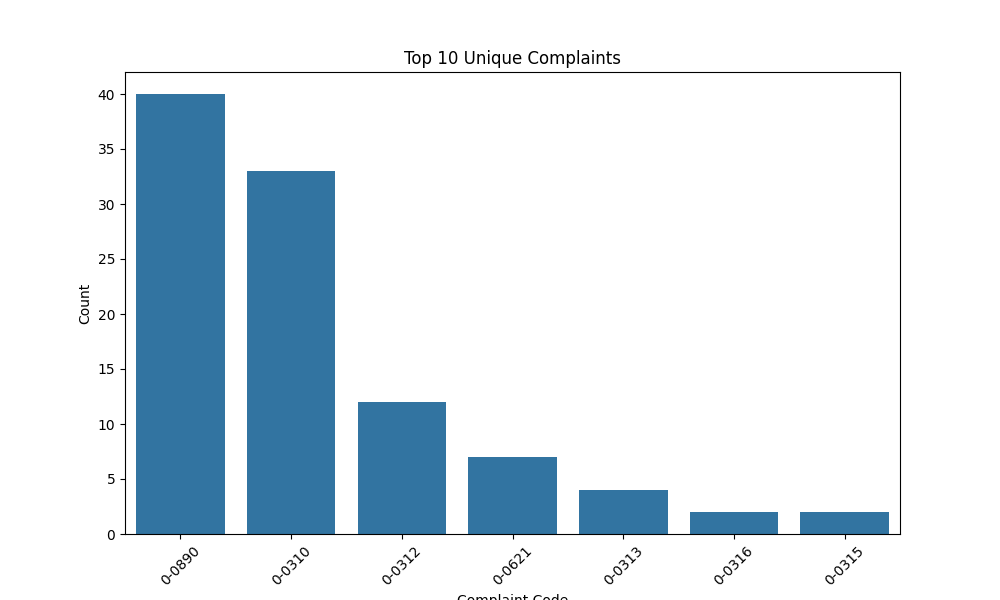
* **Robust to Outliers**
* **Preserves Data Distribution**
* **Avoids Skewed Data**
  1. **Clean categorical columns to maintain data integrity**

Convert categorical columns to string, strip whitespace and standardize capitalization.

* 1. **Format numerical columns and free from outliers**

Ensure the numeric columns to be numeric type, handling outliers using IQR method.

1. **Identifying Critical Columns:**
   1. **Top 5 critical columns selected are:**VIN, REPAIR\_DATE, COMPLAINT\_CD, GLOBAL\_LABOR\_CODE, TOTALCOST
   2. **Reasoning for selection:**
      1. **VIN:** This column is crucial for tracking individual vehicles and their repair history, allowing stakeholders to analyze patterns specific to certain vehicles.
      2. **REPAIR\_DATE:** Understanding when repairs occur can help identify trends, peak seasons, and the effectiveness of maintenance schedules.
      3. **COMPLAINT\_CD:** Analyzing complaint codes can reveal common issues faced by customers, guiding product improvements and customer service enhancements.
      4. **GLOBAL\_LABOR\_CODE:** This column helps in assessing the types of labor performed, which can inform training needs and operational efficiency.
      5. **TOTALCOST:** Financial insights are vital for stakeholders to manage budgets, pricing strategies, and profitability.
   3. **Generating Visualizations:**
   * **A bar plot showing the count of unique complaint codes:** This bar plot shows the most common complaint codes, helping stakeholders identify prevalent issues that need attention.

****

* + **A line plot showing the trend of repairs over time:** This line plot illustrates trends in repair activity over time, allowing stakeholders to see seasonal patterns or changes in repair frequency.

**A graph showing a line

AI-generated content may be incorrect.**

* + **A histogram showing the distribution of total repair costs:** This histogram displays the distribution of repair costs, helping stakeholders understand the cost structure and identify any outliers or trends in pricing.

**A graph of a distribution of repair costs

AI-generated content may be incorrect.**

1. **Generating tags/features from free text available:**

**vectorizer = CountVectorizer(stop\_words='english')**

A CountVectorizer object is created. This object is used to convert a collection of text documents into a matrix of token counts. The parameter stop\_words='english' specifies that common English stop words (like "the", "is", "and", etc.) should be ignored during the tokenization process. This helps in focusing on more meaningful words.

**X = vectorizer.fit\_transform(df[col].dropna())**

This method tokenizes the text, counts the occurrences of each word, and creates a sparse matrix **X** where each row corresponds to a document (in this case, each entry in the column) and each column corresponds to a unique word (token) found in the text.

**tags.extend(vectorizer.get\_feature\_names\_out())**

This method retrieves the unique words (features) that were identified during the tokenization process.

* Overally, we extracted meaningful tags from free text fields using natural language processing techniques. This helps in summarizing information and identifying common themes or issues.

**5. Overall Synthesis/key takeaways:**

**Recommendations for stakeholders**:

- Focus on improving the quality of parts that frequently require repairs.

- Enhance customer service training based on common complaints.

- Monitor repair costs closely to identify potential areas for cost savings.