1.Column analysis :

1. **VIN (Vehicle Identification Number)**: A unique code assigned to each vehicle, which helps in identifying specific vehicles. It’s like a fingerprint for cars, used to track the vehicle’s history and specifications.
2. **TRANSACTION\_ID**: This is a unique identifier for each repair or service transaction. It helps to track each repair event in the dataset.
3. **CORRECTION\_VERBATIM**: Detailed description of the repair or correction performed. This column provides a free-text explanation of what was fixed or corrected in the vehicle.
4. **CUSTOMER\_VERBATIM**: This column records the customer's complaint or issue with the vehicle, written in their own words.
5. **REPAIR\_DATE**: The date when the repair or service was completed. This helps in understanding the timeline of repairs.
6. **CAUSAL\_PART\_NM**: This column contains the name of the part that caused the issue, leading to the repair. It identifies the root cause of the problem.
7. **GLOBAL\_LABOR\_CODE\_DESCRIPTION**: A description of the labor operation performed during the repair, based on a global coding system. It gives details about the type of labor involved.
8. **PLATFORM**: The vehicle’s underlying design or architecture. It indicates the base design that the vehicle is built on, such as the type of vehicle platform.
9. **BODY\_STYLE**: The type or design of the vehicle’s body. For example, this could be sedan, SUV, truck, etc.
10. **VPPC (Vehicle Product Program Code)**: A code used to identify the vehicle in a product program. This is used for tracking specific vehicle models and configurations.
11. **PLANT**: The manufacturing plant where the vehicle was assembled. It helps in tracing the vehicle's production history.
12. **BUILD\_COUNTRY**: The country in which the vehicle was manufactured. This gives geographical information about where the vehicle was produced.
13. **LAST\_KNOWN\_DLR\_NAME**: The name of the last dealer who interacted with or serviced the vehicle.
14. **LAST\_KNOWN\_DLR\_CITY**: The city where the last known dealer is located. It helps in identifying where the vehicle was last serviced.
15. **REPAIRING\_DEALER\_CODE**: A code identifying the dealer that performed the most recent repair or service on the vehicle.
16. **DEALER\_NAME**: The name of the dealer performing the current repair.
17. **REPAIR\_DLR\_CITY**: The city where the repair dealer is located.
18. **STATE**: The state in which the repair dealer is located.
19. **DEALER\_REGION**: The region associated with the dealer performing the repair. It helps in understanding the geographical service area.
20. **REPAIR\_DLR\_POSTAL\_CD**: The postal code of the repair dealer’s location.
21. **REPAIR\_AGE**: The age of the vehicle at the time of repair, usually expressed in months or years. This can indicate whether older vehicles are more likely to need repairs.
22. **KM**: The distance traveled by the vehicle, recorded in kilometers. It helps in tracking how much the vehicle has been used and whether higher mileage is associated with more repairs.
23. **COMPLAINT\_CD\_CSI**: Customer Satisfaction Index code for the complaint. It categorizes the complaints based on their severity or nature.
24. **COMPLAINT\_CD**: A code that categorizes the type of complaint made by the customer. It helps in analyzing the most common issues vehicles face.
25. **VEH\_TEST\_GRP**: This refers to the vehicle test group, which could indicate the vehicle’s emissions or compliance with specific vehicle regulations.
26. **COUNTRY\_SALE\_ISO**: The ISO code of the country where the vehicle was sold. This helps to identify the geographical sales market for the vehicle.
27. **ORD\_SELLING\_SRC\_CD**: A code representing the source from which the vehicle order or sale originated.
28. **OPTN\_FAMLY\_CERTIFICATION**: Certification information related to the vehicle’s option families. Option families refer to the different configuration options available for the vehicle.
29. **OPTF\_FAMLY\_EMISSIOF\_SYSTEM**: This column indicates the emission system associated with the vehicle’s option family.
30. **GLOBAL\_LABOR\_CODE**: A global labor operation code that identifies specific labor tasks performed during repairs.
31. **TRANSACTION\_CATEGORY**: This categorizes the transaction, such as warranty, service, or repair. It helps in understanding the nature of the transaction.
32. **CAMPAIGN\_NBR**: The number of the associated service campaign or recall. This column is useful for identifying vehicles that are part of a recall or repair campaign.
33. **REPORTING\_COST**: The cost reported for the repair or service performed. It reflects the expense associated with the repair.
34. **TOTALCOST**: The total cost incurred for the repair, including parts and labor. It helps in understanding the financial aspect of repairs.
35. **LBRCOST**: The labor cost incurred during the repair. This is part of the total cost of the repair.
36. **ENGINE**: The engine code or identifier. It helps in identifying the specific engine model in the vehicle.
37. **ENGINE\_DESC**: A description of the engine type or specifications.
38. **TRANSMISSION**: The transmission code or identifier for the vehicle’s transmission system.
39. **TRANSMISSION\_DESC**: A description of the transmission system in the vehicle.
40. **ENGINE\_SOURCE\_PLANT**: The plant where the engine was manufactured.
41. **ENGINE\_TRACE\_NBR**: A traceability number for the engine, used for tracking its history.
42. **TRANSMISSION\_SOURCE\_PLANT**: The plant where the transmission was manufactured.
43. **TRANSMISSION\_TRACE\_NBR**: A traceability number for the transmission system.
44. **SRC\_TXN\_ID**: The source transaction ID, which links the current transaction to the original record.
45. **SRC\_VER\_NBR**: The version number of the source transaction record. This helps in tracking different versions of a transaction record.
46. **TRANSACTION\_CNTR**: A counter for the number of transactions associated with a specific vehicle. This helps in understanding how many times a vehicle has been serviced or repaired.
47. **MEDIA\_FLAG**: An indicator for whether media (images, videos) is attached to the transaction record. This can be useful in visual documentation of repairs.
48. **VIN\_MODL\_DESGTR**: This column links the VIN to the model designation, indicating which vehicle model is associated with the VIN.
49. **LINE\_SERIES**: This refers to the series or product line of the vehicle, which helps in categorizing the vehicle based on its model range.
50. **LAST\_KNOWN\_DELVRY\_TYPE\_CD**: A code that indicates the type of delivery used for the last known delivery of the vehicle.
51. **NON\_CAUSAL\_PART\_QTY**: The quantity of parts used in the repair that were not identified as the causal part. This helps in understanding what additional parts were replaced during the repair process.
52. **SALES\_REGION\_CODE**: A code representing the sales region of the vehicle, which helps in understanding the geographic area in which the vehicle was sold.

#### **Steps Taken in the Analysis**

1. **Fixing Missing Values**:
   * Missing values in some columns were filled with appropriate replacements:
     + **Text Columns** (e.g., ENGINE\_SOURCE\_PLANT and OPTN\_FAMLY\_CERTIFICATION): Filled with "Unknown" or "Undefined."
     + **Numbers**:
       - TOTALCOST: Filled with the value from REPORTING\_COST if available.
       - LBRCOST and KM: Converted to numbers.
       - REPAIR\_AGE: Changed to whole numbers (integers).
2. **Data Type Conversion**:
   * Changed columns like REPORTING\_COST, TOTALCOST, LBRCOST, KM, and REPAIR\_AGE to numbers so calculations could be done.
   * REPAIR\_DATE was converted to a date format to analyze repair timing.
3. **Finding Outliers**:
   * Boxplots (charts that show data spread) were used to check for unusually high or low values in TOTALCOST, LBRCOST, KM, and REPAIR\_AGE.
   * No changes were made to outliers, but they were noted.
4. **Checking Relationships**:
   * A heatmap (colored chart) showed how strongly numbers like KM, TOTALCOST, REPAIR\_AGE, and others are related to each other.
5. **Data Visualizations**:
   * **Top 10 Failed Parts**: Pie chart showing the most common parts causing issues.
   * **Cost by Platform**: Bar charts showing average total cost, labor cost, and repair age for each vehicle platform.
   * **Body Style Insights**:
     + Bar chart showing the number of repairs for different body styles.
     + Another chart showing body styles grouped by platform.
   * **Vehicle Test Groups**: Bar chart showing repair counts for each test group.
   * **State-wise Costs**: Bar chart showing the average total cost in each state.

### **Tags and Takeaways**

* **Tags**:
  1. The code adds tags (labels) to each record based on conditions like:
     + High repair costs (TOTALCOST).
     + Older vehicles (REPAIR\_AGE).
     + Specific parts (CAUSAL\_PART\_NM).
     + Vehicle platform or body style.
  2. These tags help to organize and analyze the data in detail.
* **Key Findings**:
  1. **High Costs**: Some vehicles have very high repair costs and are older. These need further investigation.
  2. **Platform Issues**: Vehicles from Platform 'A' seem to cost more to repair than others.
  3. **Engine Problems**: Engine-related issues could have specific patterns worth exploring.
  4. **Body Style Trends**: Some body styles seem to need repairs more often. This might be related to their design or usage.

### **Recommendations**

1. **Look into High Repair Costs**:
   * Check records tagged with "HighRepairCost" and cross-check them with other factors like the vehicle's age, platform, or body style.
2. **Preventive Maintenance**:
   * If certain parts fail often, consider better-quality parts or a plan for regular maintenance.
3. **Investigate Platform 'A'**:
   * Study why repairs for this platform are more expensive. This could help in reducing future costs.
4. **Body Style Study**:
   * Explore why certain body styles have more repairs. It might be related to design flaws or how they're used.

### **Suggestions for Improvement**

* Use smarter ways to fill missing values, like averaging or predicting based on other data, instead of just filling with placeholders.
* Handle outliers by adjusting extreme values or using methods that reduce their effect.
* Make tag generation more precise to capture more useful patterns in the data.