🔷 1. Data Inspection

Start with understanding the dataset.

- Load the data and display the first few rows.
- Check data types and missing values (None and NaN).
- Summarize statistics for numerical and categorical features.

🔷 2. Data Cleaning

Prepare the data for analysis.

- Replace "None" and other invalid strings with NaN.
- Convert fields like insurance_valid, accidents_reported, owner_count to appropriate types (int, bool, etc.).
- Ensure fields like mileage_kmpl, price_usd, engine_cc are numeric.
- Remove or impute missing values:
    - Numerical: mean/median
    - Categorical: mode or "Unknown"

🔷 3. Feature Engineering

Create and transform variables to improve modeling.

- Create car_age = current_year - make_year.
- Bin owner_count into categories (e.g., "1", "2–3", "4+").
- Encode categorical features:
    - Label Encoding for binary variables (e.g., insurance_valid)
    - One-Hot Encoding for fuel_type, brand, transmission, color, service_history

---

🔷 4. Outlier Detection and Handling

Ensure the data is within reasonable bounds.

- Use **IQR** or **z-score** to detect outliers in price_usd, mileage_kmpl, engine_cc, etc.

- Decide whether to **remove**, **cap**, or **transform** based on business logic.

🔷 5. Exploratory Data Analysis (EDA)

🔶 A. Univariate Analysis

- Histograms: price_usd, mileage_kmpl, car_age, etc.
- Bar plots: fuel_type, brand, transmission, etc.
- Boxplots: Outlier detection per numerical variable.

🔶 B. Bivariate Analysis

- Scatter plots: engine_cc vs price_usd, car_age vs price_usd.
- Boxplots: price_usd by brand, fuel_type, etc.
- Violin/strip plots: price_usd distribution across categories.

🔶 C. Multivariate Analysis

- Correlation heatmap: For numeric features.
- Pairplot: Visualize relationships between key variables.
- Stacked bar charts: brand vs fuel_type vs insurance_valid.

🔶 D. Target Variable Distribution

- Plot price_usd distribution (histogram/KDE).
- Consider **log-transform** if skewed.

🔶 E. Categorical Value Counts

- Count plots for high-cardinality fields like brand, color.