
Cleaning Strategy Summary for car_prices.csv Dataset

Columns:
make, model, year, price, mileage, fuel_type, transmission, engine_size, etc.

---

1. Handle Missing Values

- Drop columns with more than 50% missing values.
- Impute missing values:
    - Numeric columns: Fill missing values with the median.
    - Categorical columns: Fill missing values with the mode (most frequent value).

---

2. Standardize Categorical Data

- Convert all categorical columns (make, model, fuel_type, transmission, etc.) to lowercase.
- Strip any leading or trailing whitespace.

---

3. Convert Columns to Proper Data Types

- year → Convert to integer.
- price, mileage, engine_size → Convert to numeric, removing commas, symbols, or any non-numeric characters.
- Convert date-related columns (e.g., registration_date) → to datetime format.

---

4. Feature Engineering

- Create new features:
    - car_age = Current Year − year
- Bucket numerical features like engine_size or mileage if required (e.g., low, medium, high).
- Flag outliers in price with a new binary column, e.g., is_outlier.

---

5. Remove Duplicates

- Drop duplicate rows based on the combination of:
  make + model + year + mileage.
