**Detailed Preprocessing Documentation for Your Project**

**Dataset Used:**

* Source: segmented\_structured\_dataset.csv
* Initial Rows: **8,300**

**Features Available:**

| **Feature** | **Data Type** | **Description** |
| --- | --- | --- |
| oem | Categorical | Brand of the car (Maruti, Hyundai, etc.) |
| model | Categorical | Specific model name (Alto, Creta, etc.) |
| modelyear | Numerical (Year) | Manufacturing year of the car |
| Variant | Categorical | Specific trim/variant (e.g., VXi, ZXi) |
| transmission | Categorical | Manual / Automatic |
| Body Type | Categorical | Hatchback / Sedan / SUV etc. |
| Fuel Type | Categorical | Petrol / Diesel / CNG etc. |
| km | Numerical (Integer) | Kilometers driven by the car |
| ownerno | Numerical (Integer) | Number of previous owners |
| prize | Numerical (Float) | Used car price (target variable, in Lakhs) |
| city | Categorical | City of listing |
| signature | Categorical | Unique combination of oem, model, variant |

**Preprocessing Steps Applied:**

| **Step** | **Feature(s)** | **Method Used** | **Reason** |
| --- | --- | --- | --- |
| Handling Missing Values | modelyear, km, ownerno | Median Imputation | Robust against outliers |
| Handling Missing Values | oem, model, Variant, transmission, Body Type, Fuel Type, city, signature | Mode Imputation (Most Frequent) | Preserve category distributions |
| Handling Missing Values | prize | Median Imputation | No missing target values allowed |
| Encoding Categorical Features | oem, model, Variant, transmission, Body Type, Fuel Type, city, signature | One-Hot Encoding (get\_dummies) | To convert categorical to numerical for ML models |
| Scaling Numerical Features | modelyear, km, ownerno | Min-Max Scaling (0 to 1) | To balance different ranges and make training stable |
| No Scaling | prize | Left untouched (for now) | Target column, scaling only inside model pipeline if needed |
| Outlier Removal (IQR Method) | km | IQR Outlier Removal | Remove unrealistically high/low driven cars |
| Outlier Removal (IQR Method) | prize | IQR Outlier Removal | Remove wrong price listings |

**Effects of Preprocessing:**

| **Step** | **Effect** |
| --- | --- |
| Handling Missing Values | No NaN or empty entries left |
| Encoding Categorical Features | Model now understands brand, model, variant information numerically |
| Scaling Numerical Features | Features balanced on the same scale without losing real-world meaning |
| Outlier Removal | Reduced dataset to **7,400** rows, removed scrap/outlier cars |

**Changes to Columns:**

| **Column** | **Change Applied** |
| --- | --- |
| oem | One-Hot Encoded |
| model | One-Hot Encoded |
| modelyear | Scaled (0 to 1) |
| Variant | One-Hot Encoded |
| transmission | One-Hot Encoded |
| Body Type | One-Hot Encoded |
| Fuel Type | One-Hot Encoded |
| km | Scaled (0 to 1) + Outliers removed |
| ownerno | Scaled (0 to 1) |
| prize | Imputed missing values + Outliers removed (No scaling now) |
| city | One-Hot Encoded |
| signature | One-Hot Encoded |

**Final Result:**

| **Metric** | **Before** | **After** |
| --- | --- | --- |
| Rows | 8300 | 7400 |
| Columns | ~12 | ~300+ (after One-Hot encoding) |

Now fully clean, structured, ready for training.

**Extra note for documentation:**

* "During preprocessing, the prize field was carefully treated: Missing values were imputed using the median, but no normalization was performed at this stage to preserve the real-world interpretation of the price values."
* "Outlier removal was critical to prevent skewed model learning caused by anomalous records."