Reflection on Data Preprocessing (Lesson 3 Assignment)

In this assignment, I developed a reproducible preprocessing pipeline to clean and prepare a messy car dataset for modeling.

Target Cleaning (Price).
The Price column contained mixed numeric and currency-formatted strings. I converted it to numeric values to ensure consistency and calculated skewness to understand the distribution. To address positive skew, I created a log-transformed target (LogPrice) as an alternative for modeling.

Categorical Cleaning (Location).
The Location column had typos and placeholders such as "??". I normalized typos to consistent categories ("City," "Suburb," "Rural") and replaced unknowns with missing values. This step ensured imputation would be meaningful.

Imputation Choices.

- Odometer_km: filled with **median** to reduce the effect of outliers.
- Doors and Accidents: filled with **mode**, as these discrete values are best represented by the most common category.
- Location: also imputed with **mode**, since the most frequent location is a reasonable estimate for missing entries.

Duplicates and Outliers.
Duplicates were removed to improve dataset integrity. For continuous variables (Price and Odometer_km), I applied **IQR capping** to limit the influence of extreme values without discarding rows, ensuring robust modeling.

Feature Engineering.
I added three meaningful features:

- CarAge = CURRENT_YEAR − Year, to capture vehicle age.
- Km_per_year = Odometer_km ÷ CarAge (with safe handling for zero age), to standardize mileage per year.
- Is_City = binary indicator for urban cars.
  These features enhance predictive power while avoiding data leakage.

Encoding and Scaling.
Location was one-hot encoded into 0/1 dummies. Continuous numeric features (Odometer_km,

CarAge, Km_per_year) were standardized using StandardScaler for mean ≈ 0 and std ≈ 1. Target variables and dummy columns were excluded from scaling to preserve interpretability.

Final Checks.
Assertions confirmed no missing values remained, targets were numeric, dummy columns existed, and scaled features met standardization expectations. The cleaned dataset was saved as car_l3_clean_ready.csv for reproducible use.

Conclusion.
Each step was guided by the data type and distribution, ensuring that missing values, outliers, and categorical inconsistencies were handled appropriately. Feature engineering and scaling choices were made to maximize modeling readiness while maintaining data integrity and interpretability.