* **Data Cleaning vs Data Processing**

**Data Cleaning**

Focuses on fixing or removing bad data:

* Handling missing values
* Removing duplicates
* Correcting inconsistent entries

"North Amer", "North America", "N. America" — same meaning, different spellings

"Petrol", "petrol", "GAS" — inconsistent capitalization

* Handling outliers
* Data type issues:

Make sure each column’s data type is correct:

Year → integer

Price → float

* Logical consistency:

Check relationships make sense, e.g.:

Newer cars shouldn’t have huge mileage.

Higher prices should roughly align with “High” sales classification.

**Data Preprocessing**

Goes beyond cleaning — it’s about preparing the data for modeling:

* **Encoding** categorical features (like Fuel\_Type, Region)
* **Scaling/normalizing** numeric columns (if your model needs it)
* **Feature engineering** (creating new columns from existing ones)
* **Splitting** data into training/testing sets
* **Balancing** target classes if needed

**In short:**

* If your dataset is clean → you can **skip cleaning** steps.
* But you still **must do preprocessing** steps like encoding, scaling, and splitting before training a mode
* **Choosing the target**

When we do **data preprocessing**, we usually prepare the data for a **machine learning task** — such as predicting something.  
Looking at your dataset, the column Sales\_Classification contains values like **“High”** and **“Low”**, which look like *labels*.

That suggests the dataset could be used for a **classification task**, for example:

“Predict whether a car model’s sales will be High or Low based on its features (price, engine size, region, etc.)”

In this case:

* The **input features (X)** are: all other columns (Model, Year, Region, etc.)
* The **target (y)** is: Sales\_Classification

That’s why I said it’s *likely the target* — because it’s what you might want the model to learn to predict.

* **When we do up/down sampling**

**Upsampling (or downsampling)** is done **only on the target variable** — more precisely, on the **rows of the dataset based on the target classes**.

**What to choose?**

It depends on your model and dataset size:

* Since your dataset is **fairly large**, **downsampling** is a valid choice — it’ll still leave you with plenty of data.
* But if you don’t want to lose data, **upsampling** (or even better, **SMOTE**) is a safer and more balanced approach.

**When you upsample (the "High" class)**

You don’t fill other columns manually — you **duplicate** the *entire rows* that belong to the minority class (High).

So every time a “High” row is copied, all its feature values (like Model, Engine\_Size\_L, Region, Year, etc.) are copied with it.

That way, your dataset remains **consistent** — the features always match their correct target label.