ESTIMATING STOCK KEEPING UNIT USING ML

Data Quality Report

Dataset Name: train_0irEZ2H.csv

Total Records: 150,150

Total Columns: 9 (initial), expanded to 17 (after feature engineering)

1. Missing Data Summary:

`total_price`: 1 missing value — removed from the dataset

All other fields: no missing values

```
Missing Values:
record_ID          0
week               0
store_id           0
sku_id             0
total_price        1
base_price         0
is_featured_sku    0
is_display_sku     0
units_sold         0
dtype: int64
```

2. Duplicate Rows:

Duplicate records: 0

Verified using `.duplicated().sum()`

3. Data Type Checks:

`week`: originally object → converted to datetime

`store_id`, `sku_id`: integers → used for encoding

`total_price`, `base_price`: float64

`is_featured_sku`, `is_display_sku`: binary (0/1)

`units_sold`: target variable, integer

4. Outlier Detection:

Used boxplots to detect outliers in `units_sold`, `total_price`, and `base_price`

High values were preserved as they may represent bulk orders

No transformations were applied due to the robustness of tree-based models

5. Data Consistency and Integrity:

Dates in `week` column span a consistent period

`store_id` and `sku_id` match known retail identifiers and follow expected value ranges

No invalid or corrupted entries observed

6. Feature Completeness:

After preprocessing, all generated features (lag, rolling, expanding, interaction) had no missing values post-cleaning

7. Final Structure:

Clean dataset with 17 relevant features, fully ready for model training

This report confirms the data is of high quality and well-prepared for building predictive models in a demand forecasting context.