# ESTIMATING STOCK KEEPING UNIT USING ML

## Raw Data Sources

**Dataset File Name:** train_0irEZ2H.csv

**Source Location:**

Stored in Google Drive directory: `/content/drive/MyDrive/dataset/`

Loaded into Google Colab for preprocessing and analysis

**File Format:** CSV (Comma-Separated Values)

**Column Descriptions:**

- `record_ID`: Unique identifier for each transaction row
- `week`: The week of the transaction (originally in YY/MM/DD format)
- `store_id`: Identifier for the store where the product was sold
- `sku_id`: Identifier for the stock keeping unit (product)
- `total_price`: Final price paid (after discounts/promotions)
- `base_price`: Original price before discounts
- `is_featured_sku`: 1 if the SKU was featured that week, else 0

- `is_display_sku`: 1 if the SKU was given display space, else 0
- `units_sold`: Number of units sold (target variable)

**Data Source Origin:**

The data appears to simulate or represent weekly SKU-level sales data from a retail or e-commerce system

**Preprocessing Summary:**

- Loaded using `pd.read_csv()`
- One missing value detected in `total_price` and removed
- Converted `week` to datetime for time-series analysis
- Sorted by `store_id`, `sku_id`, and `week` to preserve chronological order

**Additional Generated Features:**

While the raw dataset contained only 9 columns, further features were generated from this data:

- Lag features: day_1 to day_7
- Aggregates: rolling_mean_3, expanding_mean
- Interactions: lag1_lag2_interaction, lag1_plus_lag2
- Encoded averages for `store_id` and `sku_id`

This dataset served as the foundation for feature engineering, model training, and final application deployment in the SKU forecasting pipeline.