Data cleaning is one of the most crucial yet time-consuming aspects of **Data Science**.
Raw datasets often have missing values, inconsistencies, duplicate records, and incorrect formats.
If not handled properly, these issues can lead to misleading insights and incorrect conclusions. S
this article, has 3 challenging datasets to master Data Cleaning for real-world challenges.

Dataset 1: E-Commerce Logistics Data

This dataset consists of multiple files that contain logistics and shipping information for an e-commerce business. The files include:

1. **Courier Company – Rates.csv**: Contains different courier charges based on shipping zones.
2. **Invoice.csv**: Includes order details such as Order ID, warehouse PIN code, customer PIN code, shipment type, and billing amount.
3. **Order Report.csv**: Lists SKUs (Stock Keeping Units) and order quantities.
4. **pincodes.csv**: Maps warehouse and customer pin codes with their respective shipping zones.
5. **SKU Master.csv**: Provides SKU details along with their weights.

This dataset presents multiple challenges. Some files contain missing and unnecessary columns, while others have inconsistent data types that make numerical calculations difficult. Integrating data from different files is complicated due

to varying formats and naming conventions. Duplicate records often appear, leading to data redundancy and incorrect analysis. Merging these files based on key identifiers like Order ID or SKU can be complex due to mismatches and missing values.
Find this dataset with resources to solve the data cleaning challenges **here**.

Dataset 2: Food Delivery Orders Data

This dataset contains records of food delivery orders from various restaurants. The key columns include:

1. **Order Details**: Order ID, Customer ID, Restaurant ID
2. **Timestamps**: Order Date & Time, Delivery Date & Time
3. **Financial Metrics**: Order Value, Delivery Fee, Commission Fee, Payment Processing Fee, Refunds
4. **Discounts & Payment Methods**

The food delivery orders dataset presents several challenges, particularly with timestamp inconsistencies. Some delivery times are recorded before order times due to logging errors, requiring careful verification. Discounts and promotions are often stored as unstructured text, making it difficult to extract numerical values for analysis. Duplicate order entries can distort revenue calculations and customer insights, requiring rigorous deduplication techniques.
Find this dataset with resources to solve the data cleaning challenges **here**.

Dataset 3: Cricket Match Deliveries Data

This dataset contains ball-by-ball details of a cricket match. The key columns include:

1. **Team & Player Information**: team, batter, bowler, non_striker
2. **Delivery Details**: over, runs_batter, runs_extras, runs_total
3. **Wicket Information**: player_out, wicket_kind, fielders

This data presents unique challenges due to its structure and inconsistencies. The dataset contains missing values in the player_out and wicket_kind columns since not every ball results in a wicket. Standardizing player names is another major challenge, as different formats (e.g., “V Kohli” vs. “Virat Kohli”) can cause discrepancies in performance tracking. Some records may have incorrect run totals, where runs_total does not match the sum of runs_batter and runs_extras, requiring validation.

Find this dataset with resources to solve the data cleaning challenges **here**.