🧼 Sales Dataset – Data Cleaning Report

Prepared By: Pournima Jagdish More
Date: 21/04/2025
Tool Used: Microsoft Excel
Project: Sales Data Cleaning

📌 Objective

To clean and preprocess a raw Sales dataset by handling missing values, correcting inconsistencies, removing duplicates, and formatting the data for future analysis.

📋 Step-by-Step Data Cleaning Process

---

1 Identified Missing Values

- Used **Excel Filter** to check each column for missing/null values.
- Columns with missing values:
    - Item_Weight: 1463 missing entries.
    - Outlet_Size: 1606 missing entries.

---

2 Handled Missing Values

- **Item_Weight**: Filled missing values using the **mean** of available weights. Formula used: =AVERAGE(<column>) → Pasted manually where blanks were found.
- **Outlet_Size**: Filled missing values using the **mode** (most frequent value = "Medium"). Replaced blanks using filter and manual fill.

---

3 Standardized Categorical Values

- Cleaned inconsistent text values in Item_Fat_Content:
  - Mapped all variations to either **"Regular"** or **"Low Fat"**.
  - Replaced "reg", "Reg" → "Regular"; "low fat", "LF", "Low Fat" → "Low Fat".
- Applied Excel formulas:
  - =PROPER() → for consistent casing
  - =TRIM() → to remove extra spaces

---

4 Removed Duplicates

- Used **Data → Remove Duplicates** feature to eliminate any fully repeated rows.

---

5 Formatted Column Data

- Checked data types:
  - Numerical columns formatted as **Number**
  - Date/year column Outlet_Establishment_Year formatted as **Number**
  - Text columns standardized using PROPER() and text filters

---

6 Renamed & Cleaned Column Headers

- Ensured clean column names without extra spaces or inconsistent cases.
- Example: Item_Fat_Content was kept consistent across the sheet.

✅ Final Dataset Summary

Column Name	Action Taken
Item_Identifier	Standardized casing
Item_Weight	Missing values filled with mean
Item_Fat_Content	Text cleaned & standardized
Item_Visibility	Verified – No missing value
Item_Type	Verified – No missing value
Item_MRP	Verified – No missing value
Outlet_Identifier	Verified – No missing value
Outlet_Establishment_Year	Formatted as Number
Outlet_Size	Missing values filled using mode ("Medium")
Outlet_Location_Type	Verified – No missing value
Outlet_Type	Verified – No missing value

🧠 Learnings & Experience Gained

- Gained hands-on experience with **real-world data issues**
- Learned how to handle missing values and inconsistent categories
- Practiced using **Excel functions for cleaning & formatting**
- Improved understanding of **data preparation for analysis**