GameZone Dirty Dataset Cleaning Using Excel

Data Cleaning Framework using CLEAN

🧼 Project Title:

GameZone Dirty Data Cleaning Using the CLEAN Framework

🔍 Objective:

To implement a structured and well-documented data cleaning process using Microsoft Excel, based on the **CLEAN** methodology. This approach is designed to improve data quality, ensure consistency, and prepare raw datasets for effective analysis.

🚦 Framework Overview – CLEAN

Step	Name	Description
C	Conceptualize	Understand the dataset and define the end goals
L	Locate	Identify and classify all data issues (solvable/unsolvable)
E	Evaluate	Assess severity, impact, and decide on cleaning strategy
A	Augment	Clean, enrich, and transform the data for better usability
N	Note	Document each step, changes, assumptions, and decisions made

---

1️⃣ C - Conceptualize the Data

✔️ Goal:

Understand the dataset’s structure, purpose, and business context.

🧾 Actions Taken:

- Reviewed the dataset containing product-level and sales-related information.
- Clarified the intended outcome: prepare a clean, consistent dataset for reporting and analysis.
- Identified initial concerns such as:
    - Missing or incomplete data
    - Duplicate records
    - Inconsistent formats (text/date/number)
    - Outliers affecting data reliability

2️⃣ L - Locate the Issues (Solvable or Unsolvable)

🧾 Actions Taken:

- Explored the dataset visually and manually to identify:
    - Empty or null cells
    - Repeating rows or entries
    - Formatting inconsistencies in date and number fields
    - Values that appear unrealistic or outside expected ranges
- Categorized issues as:
    - **Solvable**: Inconsistent formats, missing values (with reference), duplicates
    - **Unsolvable**: Rows with completely missing category or business context

---

3️⃣ E - Evaluate the Issues

🧾 Actions Taken:

- Prioritized issues based on:

  - Their impact on downstream reports and dashboards
  - Data reliability and completeness
- Decided which values can be imputed, which records to remove, and which issues to ignore or escalate
- Reviewed with stakeholders for validation

4️⃣ A - Augment the Data

🧾 Actions Taken:

- Replaced missing values using business logic or averages from relevant groups
- Removed duplicate entries while retaining the most complete record
- Standardized inconsistent formats:
  - Aligned dates to one consistent style
  - Reformatted textual fields (e.g., capitalization, spacing)
- Removed or flagged outlier values based on defined thresholds
- Created new calculated fields where needed
- Enriched data using reference lists or mappings

5️⃣ N - Note and Document

🧾 Actions Taken:

- Created a dedicated documentation sheet within the workbook
  - Listed all cleaning activities step-by-step
  - Mentioned column-level actions, replacements, and reasons
  - Tracked version changes and dates
- Highlighted all modified fields for easy traceability

- Ensured future users can understand the transformation history without ambiguity

✅ Final Output:

A cleaned, consistent, and analysis-ready dataset with:

- No missing critical fields
- No duplicate records
- Consistent formatting
- Documented changes and business rules for transparency

📁 Files Delivered:

1. **Cleaned Dataset (.xlsx)**
2. **Cleaning Log Sheet (included inside dataset file)**
3. **CLEAN Framework Summary (optional presentation or PDF)**

📌 Conclusion:

The CLEAN framework provided a simple, structured method to manage data cleaning tasks efficiently in Excel. It ensured high data quality, traceability of changes, and improved confidence in the results.