**Certificate in Data Analysis – KCBF Group**

**Assignment:** Practical Data Cleaning Assignment **Name:** Outa Jeconia Wambogo Agunga **Admission No.** CDAV031-01-2025

**Unit**: Data Analysis

**Sector**: Supermarket Sector:- Sawa Supermarket- Nairobi outlets

1. **Problem Statement:**

Sawa Supermarket operates several retail outlets in and around the central business district of Nairobi. To enhance their operations, the supermarket collects sales data from each branch. However, the data gathered has several inconsistencies that hinder effective analysis and decision-making. The key issues identified include:

1. **Duplicate Transaction Records**

Some transactions are recorded multiple times, leading to inflated sales figures.

1. **Inconsistent Date Formats**:

The date entries vary with some recorded as mm/dd/yyyy and others as dd/mm/yyyy creating confusion in data interpretation.

1. **Variation in Product Names**

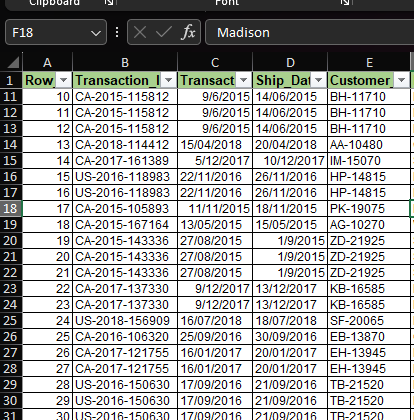
Product names are not standardized, with variations such as "saLT," "salt," "slt," and "SALT" appearing in the records. This inconsistency complicates inventory management and sales tracking.

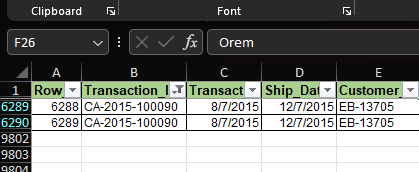
1. **Missing Product Categories and Prices**

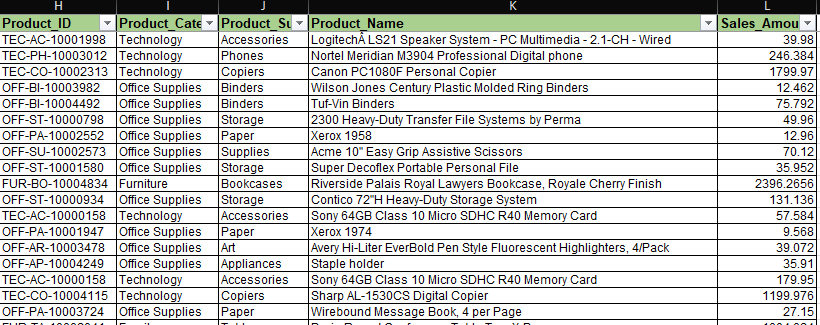
Some entries lack essential information, such as product categories and prices, which are crucial for accurate reporting and analysis.

1. **Inconsistent Currency Formatting**

While some records specify prices in KES, others either omit the currency symbol or use different formats, leading to discrepancies in financial reporting.







Addressing these anomalies is vital for Sawa Supermarket to improve data accuracy, enhance operational efficiency, and make informed business decisions

1. **Cleaning Techniques Applied**
2. **Removing Duplicates**

I eliminated duplicate transaction records using Excel's "Remove Duplicates" feature and the Pandas .drop\_duplicates() method. This ensured that each transaction is unique, providing a more accurate representation of sales.

1. **Standardizing Date Formats**

All date entries were standardized to the format DD-MM-YYYY. This uniformity simplifies date comparisons and analysis.

1. **Normalizing Product Names**

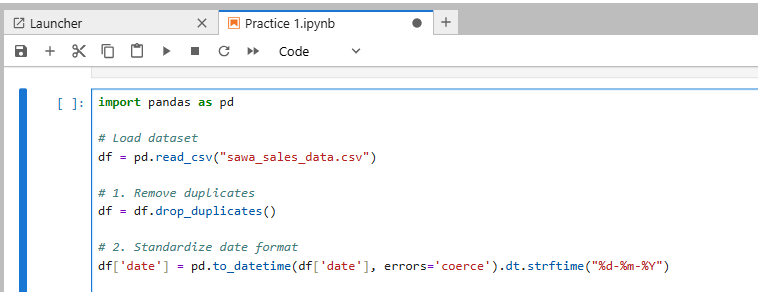
To create consistency in product naming, I applied string manipulation techniques, such as .str.strip().str.title() in Pandas and the Excel function PROPER(). This helped me in unifying variations like "saLT," "salt," and "SALT" into a standard format.

1. **Handling Missing Data**:

For entries with missing prices, I filled in the gaps using the median price of the available data. Rows missing product categories were removed to maintain data integrity.

1. **Ensuring Numeric Currency**

I ensured that all monetary amounts were numeric and formatted in KES to ensure accurate financial reporting and analysis.



1. **Challenges**
2. **Manual Corrections Needed**

**Certain product names required manual intervention for correction. For example, "SNY" needed to be changed to "Sony" to ensure accurate identification and consistency.**

1. **Currency Text in Prices**

**Some price entries included currency text, such as "KES 100," which needed to be stripped to ensure that all amounts were formatted correctly as numeric values.**

1. **Tools Used**
2. **Excel:**

I used Microsoft excel for quick duplicate removal and checking for missing values. Excel's user-friendly interface made it easy for me to perform the tasks efficiently.

1. **Python (Pandas)**

The Pandas library provided me with a powerful tool for handling data transformations, enabling me to standardize and clean the dataset effectively.

1. **Conclusion**

In conclusion, the data cleaning process for Sawa Supermarket's sales records was essential for improving data integrity and reliability. By addressing issues such as duplicate entries, inconsistent date formats, and variations in product names, I have laid the groundwork for more accurate analysis and reporting.

The challenges encountered, including the need for manual corrections and the removal of currency text, highlighted the complexities involved in data management. However, the combination of tools like Excel for quick checks and Python Pandas for automated cleanup proved effective in overcoming these challenges.

As a result, Sawa Supermarket can now leverage cleaner, more reliable data to make informed business decisions, optimize inventory management, and enhance overall operational efficiency. I would like to encourage the supermarket management to ensure consistent data practices which will be crucial for sustaining the quality of the information and supporting the supermarket's growth.