Hira Athar
August 22, 2025

Data Engineering Report

- **Software Used: Excel, Databricks, Github**
- **Task Overview:**
  On August 18, 2025, I was assigned an assessment by Middleby, with a one-week deadline. The task involved working with four distinct datasets that were interconnected in various ways. The objective was to clean, manipulate, and transform the data for analysis.

  I began the task by performing **data profiling** on each dataset. I used Excel to examine the structure and contents of each table (the Excel sheets are attached in the GitHub repository under dataset). After this initial analysis, I created a new notebook in **Azure Databricks** to conduct further exploration. The provided CSV files were uploaded to **DBFS**, allowing me to directly access and manipulate the data within the notebook. I applied the **Medallion Architecture** (bronze, silver, gold layers) to organize and analyze the data efficiently.

  Most tasks were completed quickly, except for the **description column in the item table**. To handle this, I referred to a Python tutorial on Udemy to extract size information from the descriptions and correct inaccurate size values. Also, I had initially planned to use **Delta Live Tables** for the task; however, limited access on Azure prevented me from implementing this approach.

- **Data Cleaning & Transformation**:
  Based on my previous experiences, it was not as messy and inaccurate dataset as I have seen before. This whole task took me around 1 hours and 30 mins.
  I had performed data cleaning and transformation in sales, promotions, and item tables. Transformations include removing extra spacing, filling

empty rows, dropping duplicates, converting to lowercase, and adding accurate information based on the description.

- **Business Insights**:

  I had created two business insights that stakeholders are mostly looking for:

- Find 5 items that were sold the most in each province
  I had performed the insights by joining sales and items first and then sales to the promotion table to get the top 5 items being sold in Province 1 and Province 2.

  2. Promotion effectiveness overall in the supermarkets

  For this insight, I had joined the sales with promotion table on code, supermarket, week, and province data to get total units sold and total amount to see how promotion brought revenue to the company.