**Retail Data Cleaning & Transformation Project using SQL**

**Problem Statement**

A national retail chain has shared its raw operational data in a single CSV file, which contains order transactions along with embedded customer, product, regional, and sales details.

However, the dataset suffers from the following issues:

* Duplicate entries
* Inconsistent date formats
* Unclean strings (e.g., extra spaces, inconsistent casing)
* Missing values in key fields
* All data stored in a single flat, unstructured format

**Project Objective**

The objective is to clean, normalize, and structure the dataset to support business analytics and reporting.

**Steps Involved**

**1. Data Ingestion**

Import the CSV file into a relational database management system such as SQL Server, MySQL, or PostgreSQL.

**2. Data Cleaning**

Use SQL queries to:

* Remove duplicate rows
* Standardize all date formats to YYYY-MM-DD
* Handle or replace NULL values appropriately
* Clean string data by trimming spaces and fixing casing
* Format numeric values consistently (e.g., for prices, discounts)

**3. Data Normalization**

Break down the flat file into multiple related tables using normalization techniques. The final data model will consist of four structured tables:

**Final Database Design**

|  |  |  |
| --- | --- | --- |
| **Table Name** | **Type** | **Description** |
| Orders | Fact Table | Contains transaction-level sales records |
| Customers | Dimension | Contains customer information |
| Products | Dimension | Contains product details |
| Regions | Dimension | Contains geographic and shipping information |

**Table Structure Summary**

|  |  |  |
| --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Keys** |
| Orders | Order\_ID | Customer\_ID, Product\_Key, Region\_ID |
| Customers | Customer\_ID | - |
| Products | Product\_Key | - |
| Regions | Region\_ID | - |

**Dataset**

Dataset used: [Superstore Dataset (Kaggle)](https://www.kaggle.com/datasets/vivek468/superstore-dataset-final)