

# **TASK-1: DATA IMMERSION, CLEANING & TRANSFORMATION REPORT**

- **Superstore Retail Sales Dataset**

## **1. Executive Summary**

First Task primarily deals with immersing, analyzing, cleaning, and transforming the Superstore Retail Sales data using Python. The task aimed to transform the raw retail sales data into a clean, structured, and ready-to-analyze format for exploratory data analysis (EDA), visualization, and further business intelligence analysis.

The dataset provides retail sales transaction data for various regions in the United States. The data includes customer information, product information, shipping information, and sales metrics. A comprehensive analytical process was adopted to analyze the dataset structure, data quality, data correction, data formatting, outlier identification, and the development of relevant business variables.

The final result is a cleaned dataset saved as `cleaned_superstore.csv`, ready for further sophisticated analysis and reporting.

---

## **2. Dataset Overview**

The dataset is a representation of transaction-level retail sales data, where each row represents a product sold in an order.

### **Dataset Properties**

- Total Records: 9,800+ records
- Total Features: 18 variables (initial)
- Dataset Type: Transactional retail sales data
- Geographic Distribution: United States
- Entities Described: Customers, Orders, Products, Regions, Sales

Each Order ID can be repeated several times since an order can consist of more than one product.

---

## **3. Data Immersion & Familiarization**

An exploration process was followed to gain insights into the data.

The following activities were carried out:

- First five rows were displayed to have a glimpse of the data structure
- Dataset dimensions (number of rows and columns) were analyzed
- Column names were viewed
- Data types were analyzed

- Descriptive statistics for numerical columns were obtained
- Unique values for each column were assessed

This phase enabled a thorough understanding of:

- Data structure
  - Categorical variability
  - Data type uniformity
- 

## **4. Data Dictionary**

The following is a detailed description of each column, its meaning, and importance.

### **Row ID**

Meaning: A unique and sequential identifier for each row.

Business Importance: It has importance only for indexing purposes and is not of much use for analysis.

### **Order ID**

Meaning: A unique identifier for each order transaction.

Business Importance: It allows for grouping various products into a single order and facilitates calculations for order-level revenues.

### **Order Date**

Meaning: The date of the order transaction.

Business Importance: It facilitates time-series analysis, such as monthly trends and annual revenue, among others.

### **Ship Date**

Meaning: The date of shipment of the order.

Business Importance: It facilitates analysis of the time taken for shipping and operational efficiency.

### **Ship Mode**

Meaning: The shipping mode chosen (Standard Class, First Class, Same Day, and so on).

Business Importance: It facilitates analysis of customer shipping preferences and cost-effectiveness.

### **Customer ID**

Meaning: A unique identifier for each customer.

Business Importance: It facilitates customer-level analysis and repeat business analysis.

**Customer Name**

Meaning: The full name of the customer.

Business Importance: It has importance only for identification purposes, and analysis focuses on Customer ID.

**Segment**

Meaning: Customer segment (Consumer, Corporate, Home Office).

Business Importance: Allows revenue comparison across customer segments.

**Country**

Meaning: Customer's country.

Business Importance: Minimal variation in this data set (all from United States).

**City**

Meaning: Customer's city.

Business Importance: Aids in analyzing sales performance at the city level.

**State**

Meaning: Customer's state.

Business Importance: Allows revenue comparison at the state level.

**Postal Code**

Meaning: ZIP code of customer location.

Business Importance: Aids in geographic mapping and regional analysis.

**Region**

Meaning: Geographic region classification (East, West, South, Central).

Business Importance: Aids in regional performance comparison.

**Product ID**

Meaning: Unique identifier for each product.

Business Importance: Allows product-level tracking and analysis.

**Category**

Meaning: High-level product grouping (Furniture, Office Supplies, Technology).

Business Importance: Aids in strategic category performance analysis.

**Sub-Category**

Meaning: Detailed product grouping within category.

Business Importance: Allows granular performance analysis.

## **Product Name**

Meaning: Full product description.

Business Importance: Used to determine top-performing products.

## **Sales**

Meaning: Revenue generated from each transaction.

Business Importance: Primary financial metric used for revenue analysis and performance evaluation.

---

## **5. Data Quality Assessment**

A systematic quality check was conducted to ensure reliability.

---

### **5.1 Handling Missing Values**

Missing values were detected in the Postal Code field.

Missing postal codes were replaced with a neutral value (0) to maintain data integrity and prevent the removal of genuine transactions.

### **5.2 Handling Duplicate Values**

Duplicate values were detected using in-built duplicate detection tools. The data was reviewed to ensure that no spurious duplication occurred.

### **5.3 Handling Data Type Issues**

The Order Date and Ship Date fields were initially stored as string data types.

The data type was corrected to facilitate:

- Year extraction
- Month extraction
- Calculation of shipping times
- Analysis of time series data

### **5.4 Handling Text Data**

The City and State fields were cleaned to maintain uniformity in formatting:

- Leading and trailing spaces were removed
- Text was converted to title case

## **5.5 Handling Outliers**

Outliers were detected using a box plot visualization technique on the Sales field. The data revealed some high-value sales. However, these sales were genuine bulk sales and not errors. Since retail sales data is known to be right-skewed due to high-value sales, no outliers were removed.

---

## **6. Data Cleaning & Transformation**

After quality assessment, structured transformations were applied.

---

### **6.1 Removal of Non-Analytical Column**

The Row ID column was removed since it was non-analytical.

### **6.2 Sales Value Standardization**

The sales values were rounded to two decimal places to standardize the currency format.

### **6.3 Feature Engineering**

New features were developed to improve analytical power.

#### **Shipping Days**

This feature was derived by subtracting the Order Date from the Ship Date.

Purpose:

To analyze delivery efficiency and performance.

#### **Order Year**

This feature was derived by extracting the year from the Order Date.

Purpose:

To analyze annual trends and revenue growth.

#### **Order Month**

This feature was derived by extracting the month from the Order Date.

Purpose:

To analyze seasonal and monthly performance.

## **Sales Category**

The sales were categorized as follows:

- LOW (Sales < 50)
- MEDIUM (50 ≤ Sales < 200)
- HIGH (Sales ≥ 200)

Purpose:

To simplify revenue analysis and group analysis.

---

## **7. Final Output**

The cleaned and processed dataset was saved as:

cleaned\_superstore.csv

The final dataset is ready and optimized for:

- Revenue trend analysis
  - Regional performance analysis
  - Customer segmentation
  - Shipping efficiency analysis
  - Business dashboard development
- 

## **8. Conclusion**

In Task-1, the raw retail transaction data was successfully transformed into a structured and analysis-ready dataset. Through data immersion and analysis, inconsistencies in the data were removed, and new business attributes were created.

The dataset is now ready for:

- Revenue trend analysis
- Regional performance analysis
- Customer segmentation
- Shipping efficiency analysis
- Business dashboard development