



# DATA CLEANING

Using Microsoft SQL

# DATASET OVERVIEW

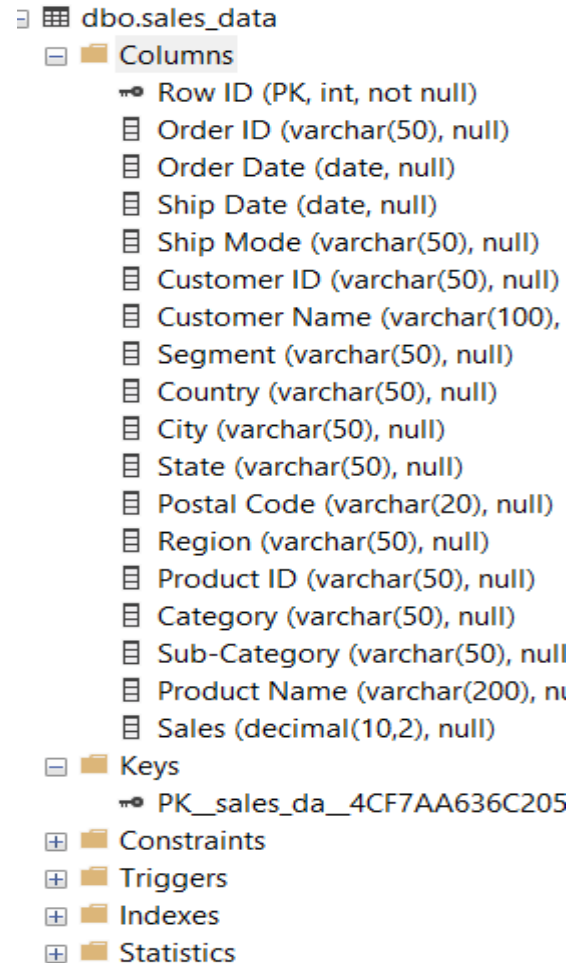
the Superstore Sales Dataset contains 9,800 rows and 18 columns, including details about orders, customers, products, shipping, and sales. Key columns include : Order Date and Ship Date for analyzing timelines

Category and Sub Category for product segmentation

Sales for performance analysis by region, segment, or category

Customer ID, Region, and Ship Mode for customer behavior and operational insights

# IMPORTING THE DATABASE TO MICROSOFT SQL



dbo.sales\_data

- Columns
  - Row ID (PK, int, not null)
  - Order ID (varchar(50), null)
  - Order Date (date, null)
  - Ship Date (date, null)
  - Ship Mode (varchar(50), null)
  - Customer ID (varchar(50), null)
  - Customer Name (varchar(100), null)
  - Segment (varchar(50), null)
  - Country (varchar(50), null)
  - City (varchar(50), null)
  - State (varchar(50), null)
  - Postal Code (varchar(20), null)
  - Region (varchar(50), null)
  - Product ID (varchar(50), null)
  - Category (varchar(50), null)
  - Sub-Category (varchar(50), null)
  - Product Name (varchar(200), null)
  - Sales (decimal(10,2), null)
- Keys
  - PK\_sales\_da\_\_4CF7AA636C205
- Constraints
- Triggers
- Indexes
- Statistics

## 1. Table Creation

The first step was to create the sales\_data table with the necessary fields to store all relevant sales information

```
CREATE TABLE sales_data (  
    "Row ID" INT IDENTITY PRIMARY KEY,  
    "Order ID" VARCHAR(50),  
    "Order Date" DATE,  
    "Ship Date" DATE,  
    "Ship Mode" VARCHAR(50),  
    "Customer ID" VARCHAR(50),  
    "Customer Name" VARCHAR(100),  
    "Segment" VARCHAR(50),  
    "Country" VARCHAR(50),  
    "City" VARCHAR(50),  
    "State" VARCHAR(50),  
    "Postal Code" VARCHAR(20),  
    "Region" VARCHAR(50),  
    "Product ID" VARCHAR(50),  
    "Category" VARCHAR(50),  
    "Sub-Category" VARCHAR(50),  
    "Product Name" VARCHAR(200),  
    "Sales" DECIMAL(10, 2)  
);
```

## 2. Handling Missing Postal Codes

To handle missing or empty values in the Postal Code column,  
I replaced any empty strings or NULL values with actual NULL entries

```
UPDATE sales_data  
SET "Postal Code" = NULL  
WHERE "Postal Code" IS NULL OR "Postal Code" = '';
```

### 3. Standardizing Date Formats

I standardized the Order Date and Ship Date columns to ensure they are in a proper DATE format.

```
ALTER TABLE sales_data  
  ALTER COLUMN "Order Date" DATE;  
  
ALTER TABLE sales_data  
  ALTER COLUMN "Ship Date" DATE;
```

## 4.Text Formatting

I capitalized the first letter of some entries while making the remaining letters lowercase

```
UPDATE sales_data
SET "Ship Mode" = UPPER(LEFT("Ship Mode", 1)) + LOWER(SUBSTRING("Ship Mode", 2, LEN("Ship Mode"))),
    "Category" = UPPER(LEFT("Category", 1)) + LOWER(SUBSTRING("Category", 2, LEN("Category"))),
    "Region" = UPPER(LEFT("Region", 1)) + LOWER(SUBSTRING("Region", 2, LEN("Region")));
```

## 5. Removing Duplicate Records

I ensured that the dataset does not contain redundant rows

```
WITH cte AS (  
  SELECT *,  
         ROW_NUMBER() OVER (PARTITION BY [Order ID], [Product ID] ORDER BY (SELECT NULL)) AS row_num  
  FROM sales_data  
)  
DELETE FROM cte WHERE row_num > 1;
```



## 6. Previewing Data

I previewed the first 10 rows of the sales\_data table.

```
SELECT TOP 10 * FROM sales_data;
```

## 8. Adding Shipping Delay Column

I added a new column called **Shipping Delay** to the table to provide insights into shipping delays for each order

```
= SELECT COLUMN_NAME  
   FROM INFORMATION_SCHEMA.COLUMNS  
  WHERE TABLE_NAME = 'sales_data' AND COLUMN_NAME = 'Shipping Delay';  
= ALTER TABLE sales_data  
   ADD [Shipping Delay] INT;  
= UPDATE sales_data  
   SET Shipping Delay = DATEDIFF(DAY, [Order Date], [Ship Date]);
```

## 9. Cleaning Sales and Region Data

I removed rows where the Sales value was NULL

, I corrected a misspelling in the Region column by updating "Northen" to the correct "Northern "

```
= DELETE FROM sales_data  
  WHERE [Sales] IS NULL OR [Sales] = 0;  
= UPDATE sales_data  
  SET [Region] = 'Northern'  
  WHERE [Region] = 'Northen';
```