*Raghunath Singh*

# Elevate Labs Internship

## Task 1: Data Cleaning and Pre-processing

## Summary of Data Cleaning and Preprocessing for Superstore Sales Data

1. **Original Data Source: Sample - Superstore.csv** (Downloaded from Kaggle)
2. **Tools Used:** Microsoft Excel, Power Query
3. **Original Dataset:**
    - **Rows: 9995**
    - **Columns: 21**

**Specific Changes Made:**

- **Column Renaming:** Standardized column headers to snake_case (e.g., 'Order ID' to 'order_id', 'Product Name' to 'product_name').
- **Data Type Conversion:**
    - Converted 'order_date' and 'ship_date' from Text to Date data type using 'English (United States)' locale.
    - Ensured 'row_id', 'postal_code', 'quantity' are Whole Numbers.
    - Ensured 'sales', 'discount', 'profit' are Decimal Numbers.
- **Missing Values Handling:** No explicit missing values were identified in this dataset (0% empty cells), so no specific handling steps were required.
- **Duplicate Records:** Checked for and removed any exact duplicate rows across all columns to ensure data uniqueness. No duplicate rows were found.
- **Data Standardization:** Applied consistent casing (e.g., 'Capitalize Each Word') to categorical text columns such as 'ship_mode', 'segment', 'category', and 'sub_category' for uniformity.
- **Date Format Consistency**: Ensured 'Order Date' and 'Ship Date' were consistently handled as Date data types within Power Query, ready for flexible formatting in Excel.

**Result:** The Superstore Sales dataset is now clean, structured, and prepared for further analysis or visualization, addressing common data quality issues.

**Cleaned Dataset: Task_1_Superstore_Sales_Data_Cleaned.xlsx**

- **Final row count: 9,994**
- **Final columns count: 21**

**Final Dataset Quality**

- ☑ Clean column headers
- ☑ No duplicate rows
- ☑ No missing values
- ☑ Standardized text values
- ☑ Consistent date format (dd-mm-yyyy)
- ☑ Proper data types