**Superstore Sales EDA – HIGH LEVEL DOCUMENT**

## 1. Project Objective

This project involved a complete *Exploratory Data Analysis (EDA)* of the Superstore Sales dataset. The goal was to systematically apply *data cleaning*, *feature engineering*, and *visualization* to identify key performance metrics and data-driven patterns. The final output is a **reusable** Google Colabr Notebook that documents the entire process.

## 2.Dataset Details

The dataset comprises **21 columns**, each providing specific information about sales transactions.

Column Names and Descriptions:

- **Row ID**: Unique identifier for each row.
- **Order ID**: An identifier for each order, including the year of the order.
- **Order Date**: Date when the order was placed.
- **Ship Date**: Date when the order was shipped.
- **Ship Mode**: Mode of shipping (e.g., Standard Class, First Class).
- **Customer ID**: Unique identifier for each customer.
- **Customer Name**: Full name of the customer (to be masked).
- **Segment**: The market segment of the customer (e.g., Consumer, Corporate).
- **Country**: Country where the order was placed.
- **City**: City where the order was placed.
- **State**: State where the order was placed.
- **Postal Code**: Postal code of the shipping address.
- **Region**: Geographic region of the shipping address.
- **Product ID**: Unique identifier for each product.
- **Category**: Product category (e.g., Furniture, Office Supplies).
- **Sub-Category**: Product sub-category (e.g., Chairs, Storage).
- **Product Name**: Name of the product.

- **Sales Price**: The final price at which the product was sold after applying any discounts.
- **Quantity**: Number of units sold.
- **Discount**: The discount applied to the original price of the product.
- **Profit**: Profit earned per quantity sold.

**Source**: Public dataset from [Kaggle]

**Size:** ~50K transactions, ~5K customers, ~2K products

**Time Range**: 2014– 2017

**3. Deliverables**

Your submission should include:

- **Cleaned Dataset**: The final dataset after all cleaning and transformation steps.
- **Colab Notebook**: A notebook detailing your complete EDA process, including all code, explanations, and visualizations.
- **Brief Report**: A summary report of your findings, discussing the steps taken, insights gained, and challenges encountered during the analysis. This will be discussed live.
- **Code Documentation**: HLD/ LLD
-

4. **Methodology Overview**

The analysis followed a structured workflow:

- **Data Cleaning and Preprocessing:** Initial data was scanned for quality issues. The process included *deduplication*, *date normalization*, correction of logical errors, *imputation* of nulls, *PII masking*, and *standardization* of categorical data. All modifications were tracked by row and order ID count.
- **Feature Engineering:** The cleaned dataset was enriched with new features to support analysis, including Total Sales, Total Profit, Original Price, Shipping Urgency, Days Since Last Order, and customer-level aggregates.

- **Outlier Handling and Segmentation:** A function was implemented to remove statistical *outliers* from Sales Price and Profit using the **3*IQR rule**. Customers were then segmented into *quintiles* based on aggregated sales and profit data.
- **Visualization and Analysis:** The prepared data was used to generate a series of plots and pivot tables to address the specific analytical questions from the case study.