DEPI Graduation Project
The Analytix Team

Task1: Build Data Model, Data Cleaning and Preprocessing:

Dataset Link :
[https://drive.google.com/file/d/1c67mQSxNTcpED9MbkB_qlFMErMoAExFk/view?usp=drive_link]

Notion Link : [https://www.notion.so/Superstore-Sales-DEPI-Graduation-Project-19e07a7c03c1808f82a9e345a8dcd816?pvs=4]

Data Cleaning and Preprocessing

1. **Importing Libraries**
   - Loads essential Python libraries like pandas, numpy, and plotly.express for data manipulation and visualization.
2. **Loading the Dataset**
   - Reads the Superstore dataset from an Excel file into a Pandas DataFrame.
3. **Initial Data Exploration**
   - Displays the first few rows.
   - Checks dataset structure (info()).
   - Identifies duplicate rows and missing values.
   - Displays rows containing missing values.
4. **Outlier Detection**
   - Creates box plots for 'Sales' and 'Profit' using plotly.express to visualize outliers.
5. **Handling Missing Values**
   - 'Sales' & 'Profit' → Filled with median to avoid outlier influence.
   - 'Quantity' & 'Discount' → Filled with mean.
6. **Data Type Conversion**
   - Converts 'Quantity' to integer (as it represents whole numbers).
   - Converts 'Order Date' & 'Ship Date' to datetime format for proper date handling.

8. Final Data Verification

- Displays dataset information after modifications.
- Ensures there are no missing values left.

Deliverables:

o Data preprocessing notebook : [https://drive.google.com/file/d/1x3N10Ls_xdqB5vqXfSczmII4c-ophaWb/view?usp=drive_link]

o Cleaned dataset ready for analysis : [https://docs.google.com/spreadsheets/d/1-fiXkkbJ788iHX6i1ys6O9TZ9qcJRB0L/edit?usp=drive_link&ouid=112242095008648923771&rtpof=true&sd=true]

o SQL query : [https://drive.google.com/file/d/1n8ew8ft-uf9lSI9-5PG19GBBEsLRhMtS/view?usp=drive_link]

Build A Data Model 

The Star Schema is a data modeling approach designed for efficient querying and reporting. It consists of:

☑ Fact Table – Stores transactional data (e.g., sales, quantity, profit)

☑ Dimension Tables – Provide descriptive attributes (e.g., customers, products, dates)

📌 Converting Superstore Data to Star Schema

1. Fact Table: Fact_Sales (Central Table)
   - Contains sales transactions, including sales amount, quantity, discount, and profit.
   - Has foreign keys linking to dimensions: customers, products, ship mode, and dates.
2. Dimension Tables (Lookup Tables for Filtering & Grouping):
   - Dim_Customers: Stores customer details (ID, name, segment, location).
   - Dim_Products: Contains product info (ID, name, category, sub-category).
   - Dim_ShipModes: Stores shipping mode types.

- **Dim_Dates: Provides date-based attributes (year, month, day) for time analysis.**

🚀 Why Use a Star Schema?

✓ **Faster Queries:** Reduces complex joins, improving performance.

✓ **Better Readability:** Organized structure for easy analysis.

✓ **Scalability:** Suitable for large datasets & BI tools like Power BI/Tableau.

This schema enables efficient sales analysis, customer segmentation, and trend forecasting 🚀 💡