### ****1. Introduction****

This project focuses on building a complete data pipeline, starting from extracting a large e-commerce dataset, transforming and cleaning the data, storing it in a relational database, and visualizing key insights. The goal is to gain practical experience in ETL (Extract, Transform, Load) processes using various technologies.

### ****2. Data Source Identification & Extraction****

**Large E-commerce Dataset:**

* The dataset was downloaded from [Kaggle/UCI Repository].
* It contains over 1 million rows with information about customers, sales, products, and transactions.
* The dataset was loaded into a Pandas DataFrame for analysis.

### ****3. Data Transformation****

To ensure data quality, the following cleaning techniques were applied:

* **Handling Missing Values:** Rows with crucial missing data were removed or filled appropriately.
* **Removing Duplicates:** Duplicate customer entries and transactions were identified and removed.
* **Formatting Data Types:** Corrected data types (e.g., converting customer IDs to strings, dates to datetime format).
* **Handling Errors:** Checked for invalid values and corrected inconsistencies.

### ****4. Data Loading****

**Database:** PostgreSQL was used for storing processed data.

**Schema Design:**

Table 1: customers (customer\_id, unique\_id, city, state)

Table 2: orders (order\_id, customer\_id, order\_status, order\_date)

Table 3: products (product\_id, product\_category, price)

Data was successfully inserted into the PostgreSQL database.

### ****5. Data Visualization & Insights****

**Tool Used:** Microsoft Power BI

**Dashboards Created:**

Sales trends over time.

Customer segmentation based on location.

Popular product categories.

**Insights Found:**

Sales peaked during certain periods (e.g., holidays, promotions).

Certain cities/states had higher purchasing activity.

Some products were more popular than others.

### ****6. Conclusion & Business Insights****

The data transformation ensured clean and structured information for analysis.

PostgreSQL efficiently stored and managed the large dataset.

Power BI visualizations provided meaningful insights into customer behavior and sales trends.

This pipeline can help businesses understand their customers better and improve sales strategies.

### ****7. Documentation & Code Quality****

The ETL pipeline was well-documented with clear comments and explanations.

The code follows best practices and handles errors efficiently.

All steps have been pushed to GitHub for submission.

**Final Submission:** The completed project, including code, database schema, and visualizations, has been documented and uploaded to GitHub.