As a data engineer, I am initiating a project focused on efficient data handling and analysis for historical sales data. The main objectives are outlined below:

A. Data Extraction from CSV Files

- I will use Python libraries, especially **Pandas**, to read and extract data from CSV files.
- The process will be optimized to manage large datasets smoothly and efficiently.

B. Data Cleaning and Transformation

- Using **NumPy** and **Pandas**, I'll clean the data by removing duplicates, filling in missing values, and standardizing formats.
- I will also transform the data through techniques like normalization, encoding, and feature engineering to get it ready for analysis.

C. Data Storage in PostgreSQL

- The cleaned and processed data will be stored in a **PostgreSQL** database.
- I will design and implement database schemas that reflect the data structure and relationships accurately.

D. Version Control with GitHub

- A **GitHub** repository will be set up to track changes, collaborate, and maintain version control.
- I'll use **GitHub Desktop** to manage code updates, create branches, and merge contributions efficiently.

Summary

By executing these steps, I aim to build a robust data pipeline for Yanki Ecommerce. The project will leverage tools like Python, Pandas, NumPy, GitHub, and PostgreSQL to enhance data analysis and decision-making—supporting the company's growth and performance.