**Case Study: E-commerce Data Engineering and Analytics Project**

**Project Description:**

**Objective**: Utilize Python, SQL, PySpark, and Git to preprocess, analyze, and derive actionable insights from e-commerce transactional data to optimize inventory and marketing strategies.

**Problem Statement:**

An e-commerce company aims to enhance its inventory management and marketing campaigns by analyzing customer purchase data. The goals include increasing sales, improving customer satisfaction, and reducing inventory costs by identifying purchase patterns, customer preferences, and product performance.

**Data Set:**

You are provided with a dataset containing customer transactions over the past two years. The dataset includes the following features:

* **Transaction ID**: Unique ID of the transaction
* **Product ID**: Unique ID of the product
* **Product Category**: Category of the product
* **Product Subcategory**: Subcategory of the product
* **Product Price**: Price of the product
* **Quantity Sold**: Quantity sold in each transaction
* **Transaction Date**: Date of the transaction
* **Customer ID**: Unique ID of the customer
* **Customer Location**: Location of the customer
* **Payment Method**: Payment method used (e.g., credit card, PayPal)

**Tasks:**

1. **Data Preprocessing:**
   * **Python:**
     + Clean the data by handling missing values and correcting data types.
     + Normalize the product categories and subcategories.
     + Create new features such as Total Sales (Product Price \* Quantity Sold) and Transaction Month.
   * **SQL:**
     + Create a database and import the cleaned dataset.
     + Write SQL queries to validate the data and perform initial data checks.
2. **Data Analysis and Visualization:**
   * **Python:**
     + Analyze the distribution of sales across different product categories and subcategories.
     + Identify the top 5 best-selling products and the top 5 least-selling products.
     + Visualize the monthly sales trends and identify any seasonal patterns.
     + Analyze customer purchasing behavior by location and payment method.
   * **SQL:**
     + Write queries to extract insights such as top-performing products, sales trends, and customer behavior.
3. **Advanced Analysis:**
   * **Python:**
     + Use Python classes and data structures to manage and analyze the data.
     + Implement Python generators and iterators to handle large datasets efficiently.
     + Calculate profit margins for different products and identify the most profitable products.
   * **PySpark:**
     + Read the data from a CSV file into a PySpark DataFrame.
     + Perform transformations to clean and prepare the data.
     + Calculate aggregates such as total sales per category and customer segmentation.
     + Store the processed data back into a distributed file system.
4. **Version Control:**
   * **Git:**
     + Initialize a Git repository for the project.
     + Create branches for different tasks and merge them back to the main branch.
     + Document your progress and changes using Git commit messages and push the code to a remote repository.
5. **Reporting and Insights:**
   * **Python and SQL:**
     + Generate a summary report with key findings and actionable insights.
     + Recommend strategies for improving sales and optimizing inventory based on the analysis.
     + Provide visualizations to support your recommendations.

**Deliverables:**

* A cleaned and preprocessed dataset.
* Python scripts and Jupyter notebooks containing your analysis and visualizations.
* SQL scripts for data validation and analysis.
* PySpark scripts for distributed data processing.
* A Git repository with all the project files and version history.
* A summary report with key insights and recommendations.
* Visualizations of sales trends, product performance, and customer behavior.

**Instructions:**

1. **Data Preprocessing:**
   * Use pandas for data cleaning and preprocessing in Python.
   * Ensure the data is free from errors and inconsistencies.
   * Import the data into a SQL database and validate it using SQL queries.
2. **Data Analysis and Visualization:**
   * Use matplotlib or seaborn for creating visualizations in Python.
   * Perform descriptive statistics and exploratory data analysis (EDA) using both Python and SQL.
3. **Advanced Analysis:**
   * Leverage Python's object-oriented programming features to structure your code.
   * Utilize PySpark for handling large datasets and performing distributed data processing.
4. **Version Control:**
   * Use Git for version control, branching, and merging your code.
5. **Reporting:**
   * Summarize your findings in a clear and concise report.
   * Provide actionable recommendations based on your analysis.
   * Support your recommendations with visualizations.