Task: Retail Data Analytics

You are a data scientist working for a retail company that wants to improve its sales forecasting accuracy. The company has provided you with historical sales data for various products across multiple stores. Your task is to develop a machine learning model that can accurately predict future sales based on relevant features such as product type, store location, and promotional events. The company is particularly interested in predicting sales for the upcoming holiday season.

Task Details:

1. Module 1: Introduction to Python for Machine Learning

- Load the provided sales dataset (e.g., "sales_data.csv") into a Pandas DataFrame.
- Perform basic data cleaning and preprocessing tasks, such as removing duplicates and handling missing values.
- Explore the dataset to gain insights into the available features and their distributions.

2. Module 2: Environment Setup and Version Control

- Set up a development environment on your local machine using your preferred IDE (e.g., Visual Studio Code).
- Create a virtual environment using Anaconda or virtualenv to manage project dependencies.
- Initialize a Git repository for the project and commit the initial codebase.

3. Module 3: Mathematics and Statistics for Machine Learning (focus on regression)

- Perform exploratory data analysis to understand the relationships between different features and the target variable (sales).
- Split the dataset into training and testing sets, using a suitable ratio (e.g., 80% training, 20% testing).
- Develop a regression model (e.g., linear regression) to predict sales based on the available features.
- Train the model using the training dataset and evaluate its performance on the testing dataset.
- Measure the model's performance using appropriate evaluation metrics, such as mean squared error (MSE) or R-squared.

Sample Dataset:

The provided sales dataset ("sales_data.csv") contains the following columns:

- Product ID: Unique identifier for each product.
- Store ID: Unique identifier for each store.
- Date: Date of the sales record.
- Sales: The total sales for the corresponding product and store on the given date.
- Promotion: Indicates whether a promotional event was held on the date (0: No, 1: Yes).
- Holiday: Indicates whether the date is a holiday (0: No, 1: Yes).
- Store Location: The geographic location of the store.

Deliverables:

1. A clean and preprocessed version of the dataset with duplicates removed and missing values handled.
2. Exploratory data analysis results, including visualizations and insights about the dataset.
3. A trained regression model capable of predicting sales based on the available features.
4. Model evaluation metrics (e.g., MSE or R-squared) on the testing dataset.
5. Codebase organized in a Git repository, with clear commit history and proper documentation.

By completing this task, you will have gained hands-on experience in loading and preprocessing data, setting up a development environment, performing exploratory data analysis, implementing regression modeling, and utilizing version control techniques.