Assignment: Stock Price Prediction using Regression

Objective:

The goal of this assignment is to apply the knowledge of regression modeling to predict the stock prices of a company. Students will use historical stock data to train, validate, and test a regression model that forecasts future prices based on features like date, opening price, high price, low price, and volume.

Tasks:

1. Data Collection:
- Download historical stock data for a chosen company (e.g., Google, Apple) from a source like Yahoo Finance, Quandl, or Alpha Vantage. Ensure that the dataset contains columns such as:
  - Date
  - Open price
  - High price
  - Low price
  - Close price
  - Adjusted close price
  - Volume

2. Exploratory Data Analysis (EDA):
- Perform EDA to understand the data:
  - Check for missing values and handle them appropriately.
  - Plot trends of stock prices over time using matplotlib or seaborn.
  - Generate summary statistics (mean, median, etc.) of the dataset.
  - Create scatter plots of features like volume, open price, and high price against the closing price to visualize relationships.

3. Feature Selection:
- Identify features that could be useful in predicting the stock price.
- Justify your choice of features based on correlation analysis or visual inspection.
- (Optional) Try to create new features like moving averages or price changes to improve the model.

4. Data Preprocessing:
- Split the data into training and testing sets.
- Scale the features using StandardScaler or MinMaxScaler if necessary.

- Ensure that the dates are properly formatted for time-series forecasting (if applicable).

5. Model Building:
- Choose a regression algorithm (Linear Regression, Ridge, Lasso, etc.) to build your model.
- Implement the regression model using Python's scikit-learn.
- Train the model using your training dataset.

6. Model Evaluation:
- Evaluate the model's performance using appropriate metrics such as:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R-squared (R^2)
- Analyze the model's residuals (the difference between predicted and actual values).

7. Model Improvement:
- Based on the evaluation metrics, try to improve the model by:
  - Testing different regression algorithms (e.g., Decision Tree Regressor, Random Forest Regressor).
  - Tuning hyperparameters using GridSearchCV or RandomizedSearchCV.
  - (Optional) Adding lagged features or using moving averages for time-series prediction.

8. Conclusion:
- Summarize the findings of your model:
  - What features were most influential in predicting stock prices?
  - How accurate is the model?
  - What could be done to improve the model's performance further?

Submission Requirements:

A Jupyter Notebook containing:
- Detailed steps of EDA, feature selection, preprocessing, model building, and evaluation.
- Code comments to explain each step.
- Visualizations showing key insights from the data.
- A 1-2 page report summarizing your findings and the model's performance.

Grading Criteria:

- Clarity of Code (15%): Code should be well-documented, with clear explanations and justifications for each step.
- EDA and Visualizations (25%): EDA should be comprehensive, and visualizations should provide meaningful insights.
- Model Implementation (30%): Proper implementation of regression model(s) with well-justified feature selection and preprocessing steps.
- Evaluation and Improvement (20%): Thorough evaluation of the model using appropriate metrics and attempts to improve the model's performance.
- Report and Conclusion (10%): Clear and concise report with insights on the data and model performance.