Task 1: Linear Regression Model Report

1. Task Description

This task involved building a linear regression model to predict a continuous variable. The objective was to load and preprocess a dataset, train a linear regression model using scikit-learn, interpret the model coefficients, and evaluate the model using R-squared and Mean Squared Error (MSE).

2. Dataset Used

The dataset used for this task was `Stock Prices Data Set.csv`. It contains historical stock price data with columns such as `symbol`, `date`, `open`, `high`, `low`, `close`, and `volume`.

3. Model Implemented

A Linear Regression model from `scikit-learn` was implemented.

4. Features and Target

- **Target Variable (y):** `close` (the closing price of the stock)
- **Features (X):** `open`, `high`, `low` (the opening, highest, and lowest prices of the stock for the same day)

5. Preprocessing Steps

1. **Loading Data:** The dataset was loaded using pandas `read_csv`.
2. **Handling Missing Values:** Rows with any missing values were dropped using `dropna()`.

6. Model Training and Evaluation

Training Data Size:

397968 samples

Testing Data Size:

99493 samples

Model Coefficients:

- `open` : -0.5381
- `high` : 0.7872
- `low` : 0.7510
- `Intercept` : -0.0084

Evaluation Metrics:

- **Mean Squared Error (MSE):** 0.5160
- **R-squared (R2):** 1.0000

Interpretation of Results:

The Mean Squared Error (MSE) of 0.5160 indicates that, on average, the squared difference between the predicted and actual closing prices is very small. The R-squared value of 1.0000 suggests a perfect fit, meaning the model explains 100% of the variance in the 'close' price. This exceptionally high R-squared is expected when predicting the 'close' price using 'open', 'high', and 'low' prices from the same day, as these values are inherently highly correlated. This model effectively demonstrates the relationship between these intra-day price movements.

7. Script Execution and Output

Below is the console output from the execution of `linear_regression_model.py` after the modifications to correctly identify the target variable as 'close'. This output demonstrates the data loading, preprocessing, model training, and evaluation steps.

```
Dataset loaded successfully. First 5 rows:
  symbol        date      open      high       low     close    volume
0    AAL  2014-01-02   25.0700   25.8200   25.0600   25.3600   8998943
1   AAPL  2014-01-02   79.3828   79.5756   78.8601   79.0185  58791957
2    AAP  2014-01-02  110.3600  111.8800  109.2900  109.7400    542711
3   ABBV  2014-01-02   52.1200   52.3300   51.5200   51.9800   4569061
4    ABC  2014-01-02   70.1100   70.2300   69.4800   69.8900   1148391

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 497472 entries, 0 to 497471
Data columns (total 7 columns):
 #   Column  Non-Null Count   Dtype
---
 0   symbol  497472 non-null  object
 1   date    497472 non-null  object
 2   open    497461 non-null  float64
 3   high    497464 non-null  float64
 4   low     497464 non-null  float64
 5   close   497472 non-null  float64
 6   volume  497472 non-null  int64
dtypes: float64(4), int64(1), object(2)
memory usage: 26.6+ MB

Missing values before preprocessing:
symbol     0
date       0
open      11
high       8
low       8
close      0
volume     0
dtype: int64

Missing values after preprocessing:
symbol    0
date      0
open      0
high      0
low       0
close     0
volume    0
dtype: int64
```

```
Features (X) shape: (497461, 3)
Target (y) shape: (497461,)
Target variable identified as: 'close'

Training data size: 397968
Testing data size: 99493

Model training complete.

Model Coefficients:
open: -0.5381
high: 0.7872
low: 0.7510
Intercept: -0.0084

Mean Squared Error (MSE): 0.5160
R-squared (R2): 1.0000
```