**Stock Price Prediction using Linear Regression**

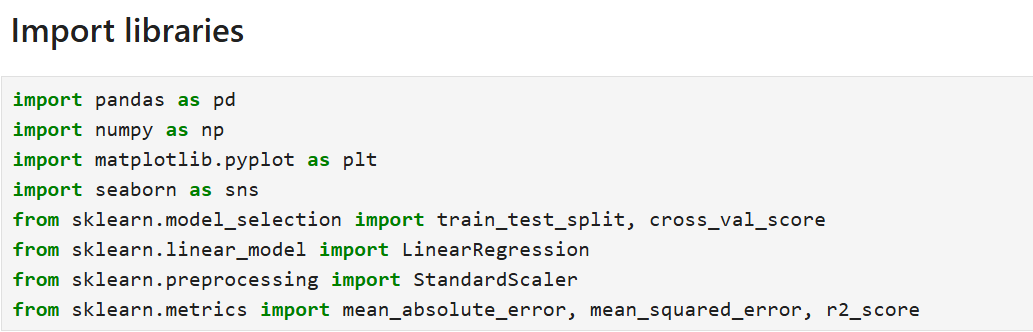
**By Shaila Magline**

# Introduction

This project aims to predict Amazon stock prices using linear regression, leveraging historical data on commodity prices and trading volumes. The objective is to build a model that accurately forecasts stock prices by analyzing the impact of various features on Amazon’s stock. Accurate stock price predictions are vital for informed investment decisions and financial planning. The dataset includes historical stock prices and relevant economic indicators, providing a comprehensive basis for analysis. The project encompasses data exploration, preprocessing, model development, and evaluation, ultimately offering insights into the factors influencing stock prices and enhancing practical machine learning applications in finance.

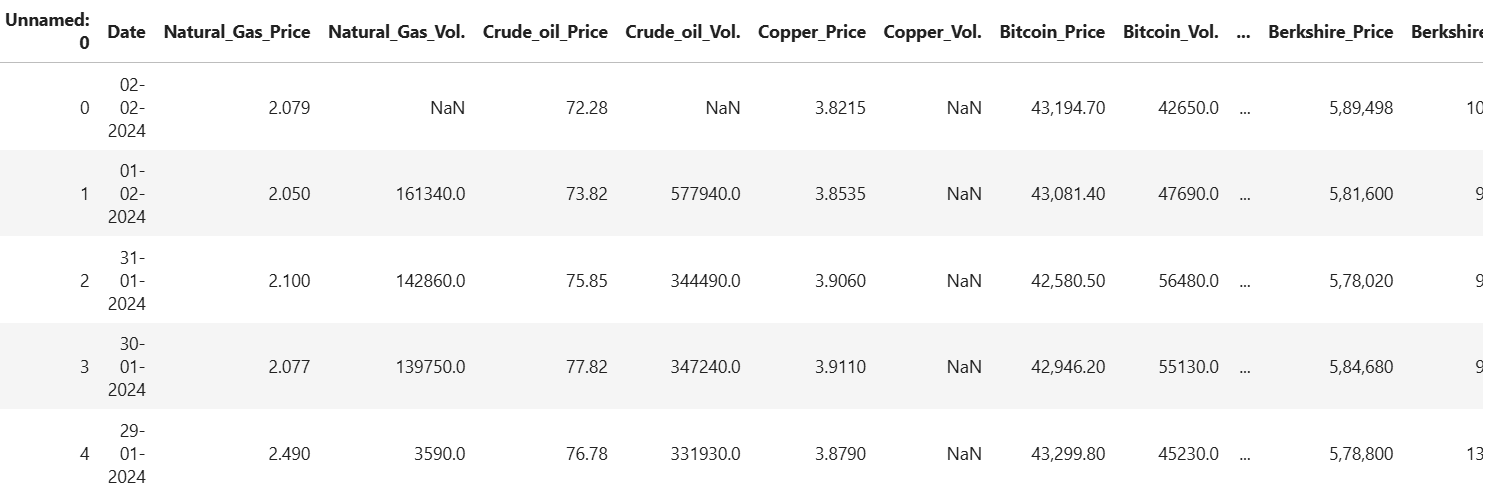
# Data Exploration and Preprocessing

## Import library and dataset

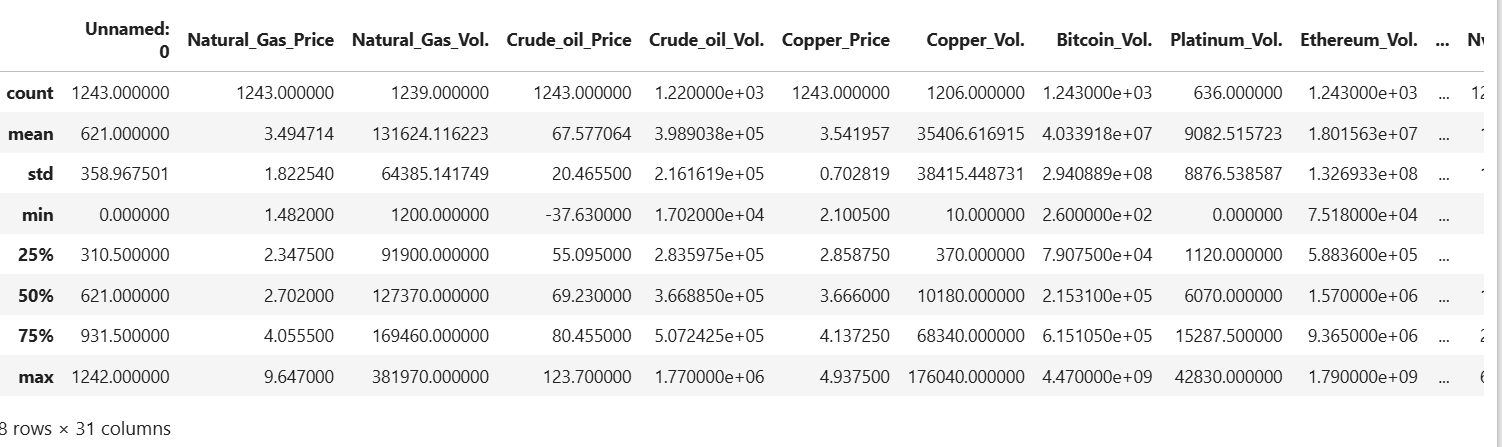


*Fig 1: Import Dataset*

The first step in the data exploration process is to import the dataset into the working environment. This involves loading the historical stock prices, commodity prices, and trading volumes from the provided data source. We will use libraries such as Pandas to read the dataset and examine its structure.



*Fig 2: Stock Market dataset*

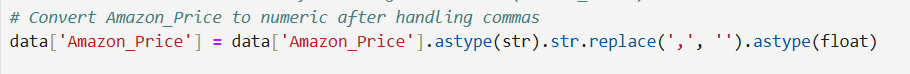


*Fig 3: Data Exploration*

## Data Preprocessing

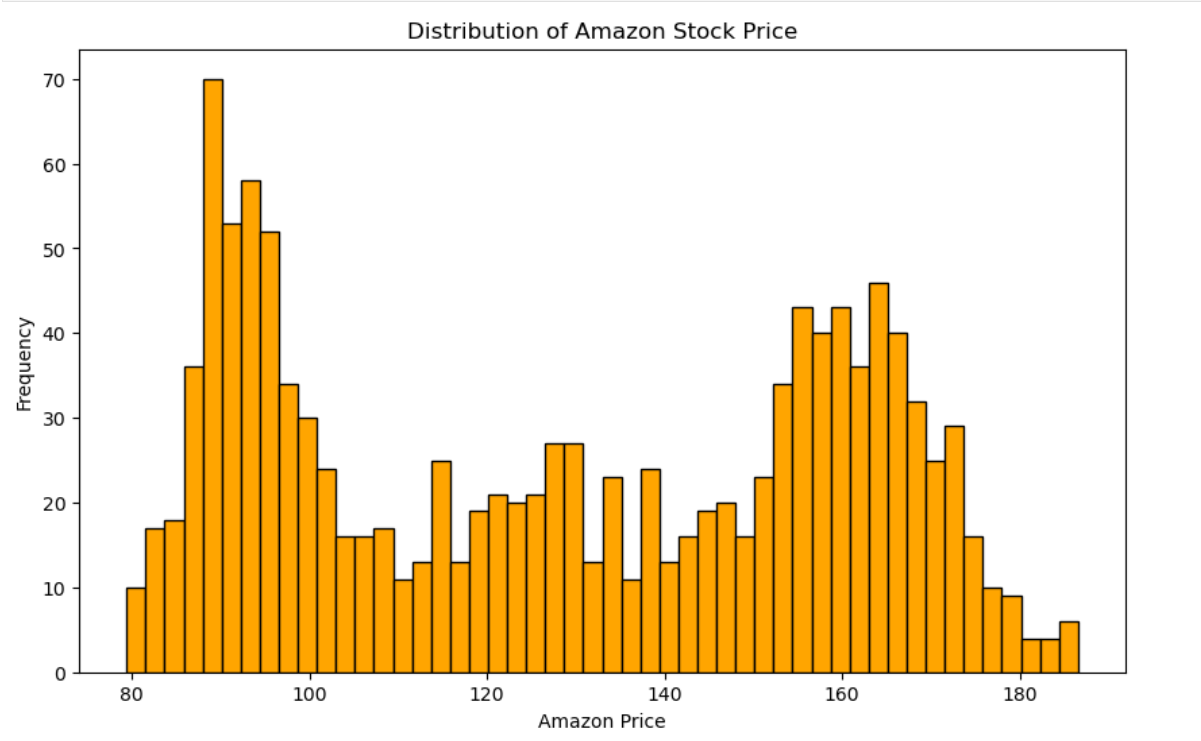
Data preprocessing involves cleaning and preparing the data for analysis. This includes handling missing values, converting data types, and standardizing the data. Key steps include:

* Handling Missing Values: Fill or interpolate missing values to ensure completeness.
* Data Type Conversion: Convert data types as needed for analysis, such as converting date columns to datetime format.
* Normalization/Standardization: Scale features to ensure they are on a comparable scale, often using techniques such as StandardScaler.



*Fig 4: Data Preprocessing*

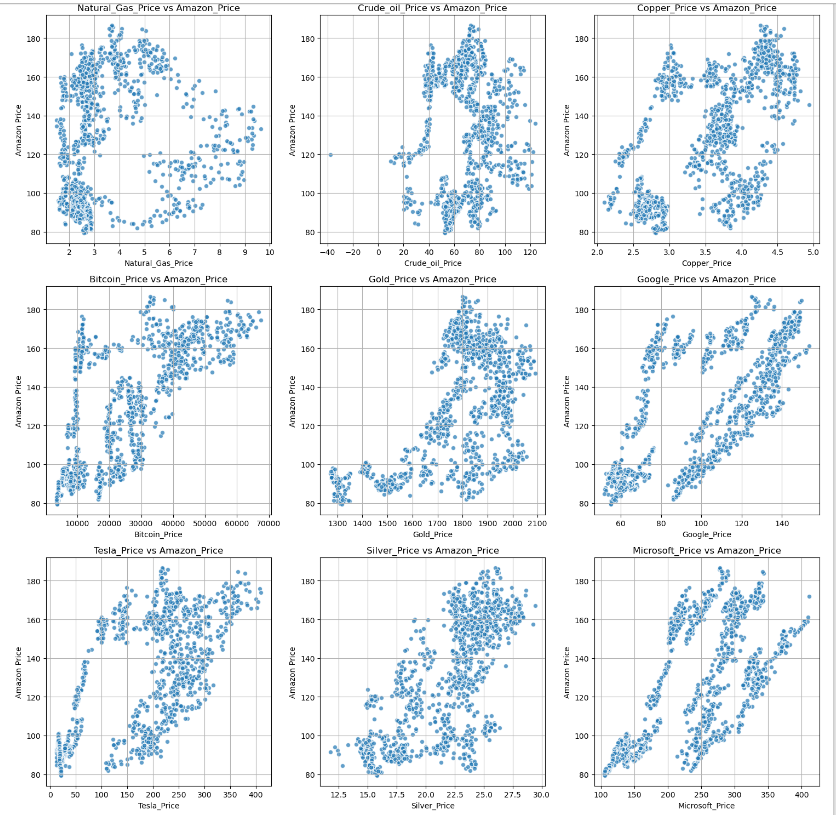
## Exploratory Data Analysis (EDA)



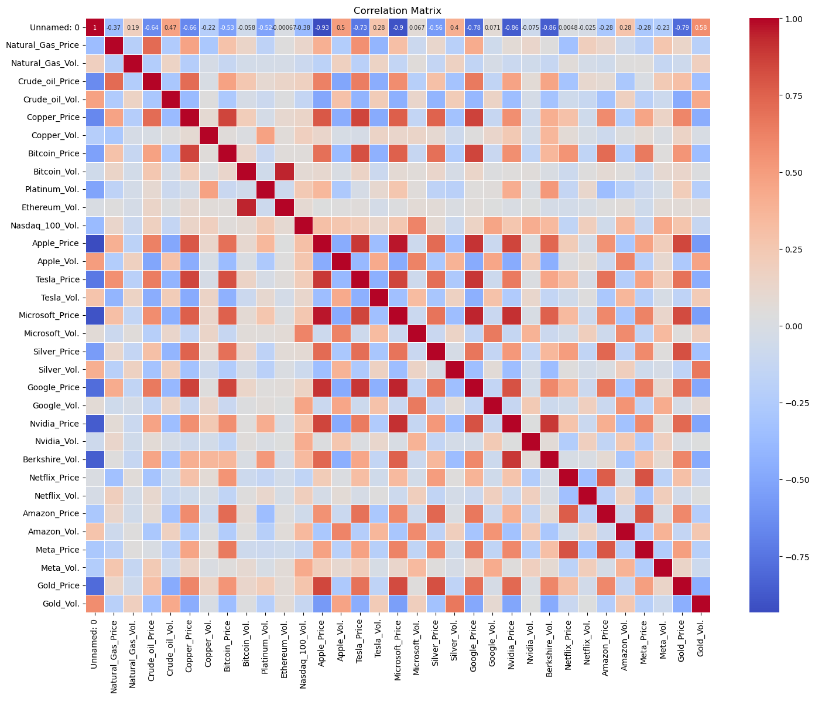
*Fig 4: Distribution of Amazon Price*

Exploratory Data Analysis (EDA) is crucial for understanding the dataset’s characteristics and relationships. Key EDA tasks include:

* Distribution Analysis: Visualize the distribution of the target variable (Amazon stock prices) to understand its range and variability.
* Correlation Analysis: Compute and visualize the correlation matrix to identify relationships between features and the target variable.
* Scatter Plots: Analyze the relationship between Amazon stock prices and other features through scatter plots.

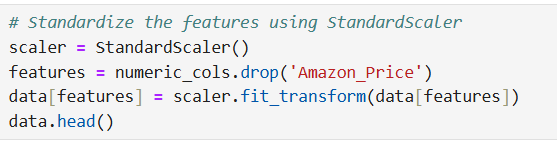


*Fig 6: Relationship with features*

**

*Fig 7: Correlation matrix*

## Feature Engineering



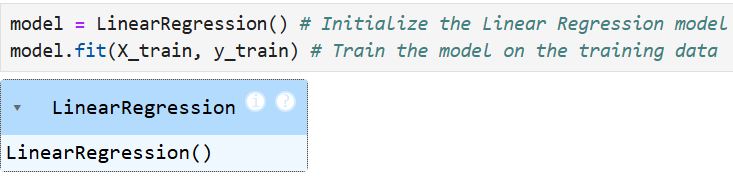
*Fig 8: Standardizing the features*

Feature engineering involves transforming raw data into meaningful features that improve model performance. This includes creating new features, handling missing values, and scaling features to ensure they contribute effectively to the model. For instance, we handle missing values by imputing with column means and standardize features to have a mean of 0 and standard deviation of 1, which helps in improving the model's accuracy and convergence.

# Model Development

## Split and train the model

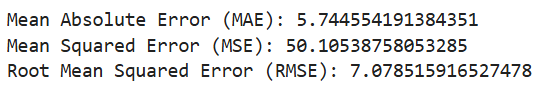
To develop a robust model, the dataset is divided into training and testing sets. This step ensures that the model is trained on one portion of the data and evaluated on another, providing a clear measure of its performance. We use an 80-20 split ratio, where 80% of the data is allocated for training and 20% for testing. The linear regression model is then initialized and trained using the training data.



*Fig 9: Training the model*

## Evaluate the model

Once the model is trained, its performance is evaluated using the testing set to determine how well it predicts unseen data. The evaluation involves calculating several metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide insights into the accuracy and reliability of the model's predictions.



*Fig 10: Model Evaluation*

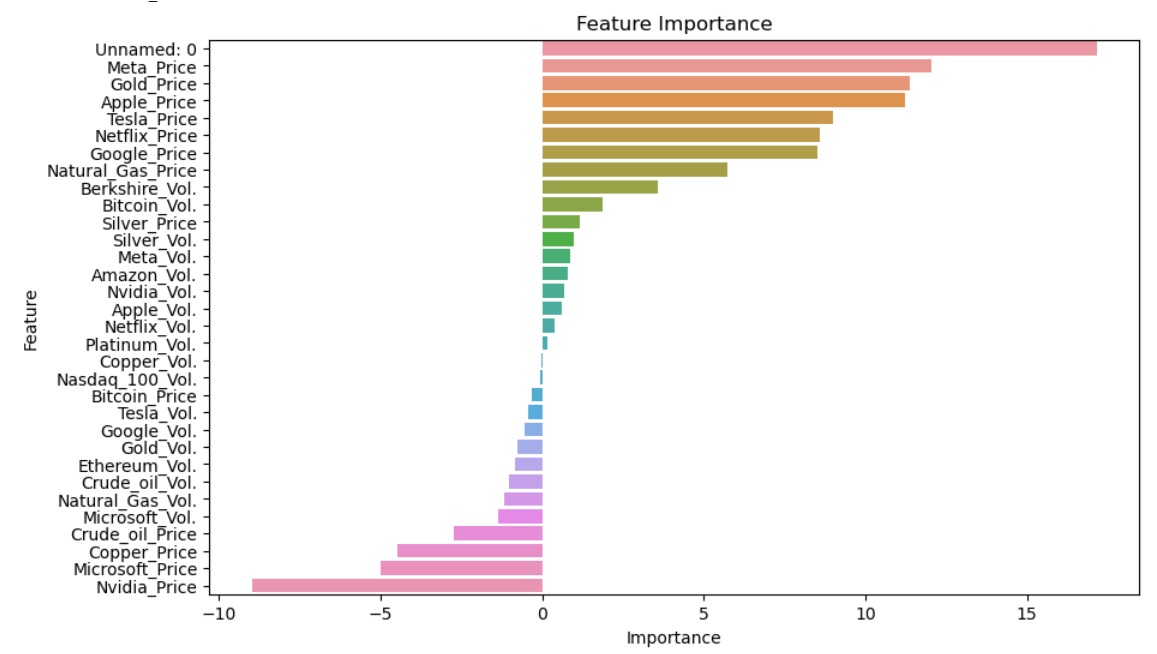
# Cross-Validation of Model performance

## Cross Validation

Cross-validation is a technique used to assess the performance of a model by partitioning the data into several subsets or "folds." The model is trained on some folds and tested on the remaining fold(s), rotating through all folds. This helps ensure that the model's performance is consistent and not overly dependent on a specific subset of data. We use 5-fold cross-validation for this project to evaluate the model's performance more robustly.

## 4.2 Feature Importance

Feature importance helps us understand the impact of each feature on the model's predictions. In linear regression, this is determined by examining the coefficients of the model. Larger coefficients indicate a greater impact on the target variable. We can visualize feature importance using a bar plot to better interpret which features are most influential.



*Fig 11: Feature Importance*

# Insights

* **Feature Importance**: The analysis revealed that features like 'Unnamed: 0', 'Meta\_Price', and 'Gold\_Price' have the highest importance, suggesting they have significant influence on the model's predictions. Conversely, features such as 'Nvidia\_Price' and 'Microsoft\_Price' have negative coefficients, indicating a negative relationship with Amazon stock prices.
* **Model Performance**: While the model performs reasonably well, the high cross-validation mean MSE and its standard deviation suggest that the model’s performance is inconsistent across different data subsets. This variability indicates room for improvement in model generalization.

# Recommendations

* **Feature Engineering**: Explore additional feature engineering techniques to capture non-linear relationships and interactions between features that may enhance model performance.
* **Advanced Models**: Consider using more sophisticated models, such as Ridge or Lasso regression, and ensemble methods like Random Forests or Gradient Boosting, which might offer better accuracy and robustness compared to linear regression.
* **Hyperparameter Tuning**: Systematically tune hyperparameters to optimize the model and reduce overfitting or underfitting.
* **Data Enrichment**: Incorporate external data sources, such as macroeconomic indicators or market sentiment, to provide a more comprehensive view and improve predictive accuracy.

# Future Work

* **Refine Model Evaluation**: Implement additional performance metrics and validation techniques, such as k-fold cross-validation or time-series cross-validation, to better assess model performance and reliability.
* **Feature Selection**: Conduct a thorough feature selection process to eliminate irrelevant or redundant features, potentially using methods like Recursive Feature Elimination (RFE) or feature importance from advanced models.
* **Real-Time Data Integration**: Develop a system for integrating and analyzing real-time market data to enhance the model's adaptability and relevance to current market conditions.
* **Model Deployment**: Explore deploying the model in a real-world scenario, such as a trading platform or financial analysis tool, to assess its practical applicability and performance in live settings.