### **Predicting Stock Market Volatility using Machine Learning Techniques**

The dispersion of returns for a certain stock or market index is measured by stock market volatility. The stock market's volatility can serve as a predictor of both positive and negative trends. Fund managers, analysts, and investors are all quite interested in predicting these trends to base their investment decisions upon. However, as financial markets are complicated, predicting stock market volatility is a difficult undertaking, hence why using Machine Learning Methods can possibly prove to be a handy tool for determining stock trends.

Dataset: <https://www.kaggle.com/datasets/camnugent/sandp500>

File Name: ( all\_stocks\_5yr.csv )

Research Questions

1. How does the distribution of trading volume and stock prices (Close/Last, Open, High, Low) vary across the top 10 companies in the stock market, and how do these variables correlate with each other?
2. Which Machine Learning Model Can be best implemented to predict a stock's closing price?
3. Do certain companies show a more consistent stock price pattern than others?
4. Can we predict whether a stock will go up or down based on its previous day's features?

Research Question 1 Interpretations

* **Descriptive Statistics:** The table shows some basic statistics of the numerical columns in the dataset, broken down by the company. It provides the minimum, maximum, variance, and standard deviation of the volume, close/last, open, high, and low prices. This information gives an overview of the range and dispersion of each feature within each company's stock data.
* **Correlation Matrix:** The correlation matrix shows the pairwise Pearson correlation coefficients between the numerical features. This helps to identify the relationships between the features. In this case, the close/last, open, high, and low prices are highly positively correlated with each other (correlation coefficients close to 1), indicating that they move together. This is expected as the opening, closing, high, and low prices of a stock within a single day are typically closely related. The volume is negatively correlated with these features, which means that when the volume goes up, the prices tend to go down, and vice versa. However, it is important to note that the correlations are relatively weak and may not have strong predictive power on their own.

As you have mentioned various machine learning concepts and models, you can consider the following approach to predict a stock's closing price:

* **Data Preprocessing:** Clean the data by handling missing values, outliers, and data type conversions if needed.
* **Feature Engineering:** Create new features or transform existing features that may help the models to better understand the underlying patterns. For example, you could create features that represent moving averages, volatility, or other technical indicators.
* **Model Selection and Evaluation:** Try different machine learning models, such as Random Forest, Decision Trees, KNN, and others. Evaluate the models using appropriate metrics, such as MSE or RMSE. Consider cross-validation to get a more robust estimate of model performance.
* **Hyperparameter Tuning:** Tune the hyperparameters of the models that perform well to improve their performance further.
* **Model Interpretation:** Interpret the models' predictions, feature importance, or coefficients to gain insights into the factors affecting stock prices.
* **External Factors:** Consider incorporating external factors, such as economic indicators or news sentiment, to improve prediction accuracy further.

Remember that predicting stock prices is challenging due to the influence of various external factors and the inherent noise in financial markets.

Research Question 2

The results indicate that the best machine learning model for predicting the stock's closing price, based on the lowest Mean Squared Error (MSE), is the Random Forest model with an MSE of 12.90. The Decision Tree model had a higher MSE of 19.32, and the KNN model had a much higher MSE of 9690.12.

The explained variance ratios from PCA show that the first principal component explains 100% of the variance in the data, and the remaining components have negligible contributions. This means that most of the variability in the data can be captured by the first principal component, and the other components do not add much value.

Now, let's briefly discuss the insights obtained from the results:

* **PCA:** The results from PCA show that the first principal component captures most of the variability in the data. This suggests that the original features are highly correlated, and the first principal component is sufficient to represent the information contained in the dataset.
* **Random Forest:** The Random Forest model, being an ensemble of decision trees, had the best performance among the models tested. It tends to perform well in practice, handling non-linearities and interactions among features. Random Forests are less likely to overfit compared to individual decision trees and can handle a large number of features well.
* **Decision Tree:** The Decision Tree model had a slightly higher MSE than the Random Forest model, indicating it may not be as effective in this particular case. However, decision trees are intuitive and easy to interpret, making them a good option for cases where interpretability is important.
* **KNN:** The K-Nearest Neighbors (KNN) model had a significantly higher MSE than the other models. This suggests that the KNN model is not suitable for this dataset or requires further parameter tuning. KNN can be sensitive to the scale of the features and the choice of the distance metric, so careful preprocessing and parameter selection are crucial.

To sum up, the best model for predicting the stock's closing price, in this case, is the Random Forest model. However, it's essential to note that stock prices are influenced by various external factors, such as market sentiment, economic conditions, and company-specific news. Therefore, a more comprehensive analysis would consider these external factors and incorporate them into the model.

Research Question 3

**Shapiro-Wilk Test**: This test assesses whether the stock prices for each company follow a normal distribution. For all companies, the p-values are extremely low, indicating that the stock prices for these companies do not follow a normal distribution. Therefore, we reject the null hypothesis that the data is normally distributed.

**ANOVA p-value for mean differences**: The p-value is 0.0, which is below the standard significance level of 0.05. This indicates that there are significant differences in the means of the stock prices across different companies.

**Bartlett's test p-value for variance differences**: The p-value is 0.0, which is below the standard significance level of 0.05. This indicates that there are significant differences in the variances of the stock prices across different companies.

**Permutation test p-values**: These p-values are obtained from permutation tests, which assess whether the observed differences in the stock prices across companies could have occurred by chance. The p-values for most companies are 0.0, indicating that the observed differences are not due to chance. One company has a p-value of 0.0073, which is still below the standard significance level of 0.05, indicating that the observed differences are not due to chance for that company as well.

**Power of the test**: The power of the test is calculated as 0.7038, which is above the standard threshold of 0.8 for a test to be considered powerful. A power below 0.8 indicates that the test has a relatively high risk of failing to detect a true effect.

In conclusion, the analysis indicates that the stock prices of the different companies do not follow a normal distribution, and there are significant differences in both the means and variances of the stock prices across companies. The permutation tests further confirm that the observed differences are not due to chance. However, the power of the test is below the standard threshold, indicating that the test may have a relatively high risk of failing to detect a true effect.

Overall, the results suggest that certain companies may have more consistent stock price patterns than others, as evidenced by the significant differences in means and variances. However, further analysis and additional data may be needed to confirm these findings and identify the specific companies with more consistent stock price patterns.

**AAPL (Apple) Model Summary:**

**Dependent Variable**: Close/Last (stock closing price)

**R-squared**: 0.223 - This indicates that approximately 22.3% of the variability in Apple's stock closing price can be explained by the Volume of stocks traded.

**F-statistic & Prob (F-statistic)**: The F-statistic is very high (722.4), and its associated p-value is almost 0. This suggests that the model is statistically significant.

**Coefficient for Volume**: -2.562e-07. This means that for every unit increase in volume, we expect a decrease of approximately 0.0000002562 in the stock price, holding all else constant. Though this number is small, given the vast quantities of stocks that can be traded, it could result in a more noticeable change in stock price.

**P-value for Volume**: The p-value is close to 0, indicating that volume is a significant predictor of the closing price for Apple's stock.

**Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.

**Jarque-Bera Test**: Further confirms the non-normality of residuals.

**Condition Number**: Is large, suggesting potential multicollinearity issues.

**SBUX (Starbucks) Model Summary:**

**Dependent Variable**: Close/Last (stock closing price)

**R-squared**: 0.067 - This indicates that only about 6.7% of the variability in Starbucks's stock closing price is explained by the Volume of stocks traded. This is much lower than what was observed for Apple, meaning volume is a less reliable predictor for Starbucks' stock price.

**F-statistic & Prob (F-statistic)**: The F-statistic is 181.5, and its associated p-value is very close to 0, suggesting that this model is statistically significant.

**Coefficient for Volume**: -1.356e-06. For every unit increase in volume, we expect a decrease of approximately 0.000001356 in the stock price, all else being equal.

**P-value for Volume**: The p-value is close to 0, indicating that volume is a significant predictor of the closing price for Starbucks' stock.

**Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.

**Jarque-Bera Test**: Further confirms the non-normality of residuals.

**Condition Number**: Is large, suggesting potential multicollinearity issues.

**Key Takeaways:**

**Role of Volume**: Both for Apple and Starbucks, there's a negative relationship between volume and stock closing price, meaning that as more shares are traded, the stock price tends to decrease. However, this relationship is much stronger for Apple than Starbucks, as indicated by the higher R-squared value for Apple.

**Significance of the Models**: Both models are statistically significant as indicated by the F-statistic and its associated p-value.

**Model Assumptions**: For both companies, the assumption of normally distributed residuals is violated (as indicated by Omnibus and Jarque-Bera tests). This is important because the violation of this assumption can influence the accuracy and reliability of the OLS estimates.

**Multicollinearity Warning**: The high condition number for both stocks indicates potential multicollinearity. This could be due to other unconsidered factors that are influencing both the volume and closing price, and they should be investigated further.

Overall, while these models provide some insight into the relationship between stock trading volume and closing price for Apple and Starbucks, there are violations of key assumptions that need to be addressed, possibly by including other relevant predictors or using more sophisticated modeling techniques.

**MSFT (Microsoft) Model Summary:**

**Dependent Variable**: Close/Last (stock closing price)

**R-squared**: 0.012 - This means that only 1.2% of the variability in Microsoft's stock closing price can be explained by the Volume of stocks traded. This low value suggests that trading volume is not a strong predictor of the stock closing price.

**F-statistic & Prob (F-statistic)**: The F-statistic is 29.53, and its associated p-value is close to 0. This suggests that the model is statistically significant, although the practical significance may be limited due to the low R-squared value.

**Coefficient for Volume**: -6.426e-07. This means that for every unit increase in volume, we expect a decrease of approximately 0.0000006426 in the stock price, all else being equal.

**P-value for Volume**: The p-value is close to 0, indicating that volume is a significant predictor of the closing price for Microsoft's stock.

**Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.

**Jarque-Bera Test**: Further confirms the non-normality of residuals.

**Condition Number**: Is large, suggesting potential multicollinearity issues.

**CSCO (Cisco Systems) Model Summary:**

**Dependent Variable**: Close/Last (stock closing price)

**R-squared**: 0.117 - This means that 11.7% of the variability in Cisco's stock closing price is explained by the Volume of stocks traded. Although higher than Microsoft's R-squared value, this is still not a very strong relationship.

**F-statistic & Prob (F-statistic)**: The F-statistic is 331.8, and its associated p-value is very close to 0, suggesting that this model is statistically significant.

**Coefficient for Volume**: -2.933e-07. For every unit increase in volume, we expect a decrease of approximately 0.0000002933 in the stock price, all else being equal.

**P-value for Volume**: The p-value is close to 0, indicating that volume is a significant predictor of the closing price for Cisco's stock.

**Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.

**Jarque-Bera Test**: Further confirms the non-normality of residuals.

**Condition Number**: Is large, suggesting potential multicollinearity issues.

**Key Takeaways:**

* **Role of Volume**: For both Microsoft and Cisco, there's a negative relationship between volume and stock closing price, meaning that as more shares are traded, the stock price tends to decrease. However, the volume explains a greater proportion of the variability in Cisco's stock price (11.7%) compared to Microsoft (1.2%).
* **Significance of the Models**: Both models are statistically significant as indicated by the F-statistic and its associated p-value. However, their practical significance may be limited due to low R-squared values.
* **Model Assumptions**: For both companies, the assumption of normally distributed residuals is violated (as indicated by Omnibus and Jarque-Bera tests). This could affect the reliability of the OLS estimates.
* **Multicollinearity Warning**: The high condition numbers indicate potential multicollinearity. This could be due to other unconsidered factors that are influencing both the volume and closing price and should be investigated further.

Overall, these models show some level of a negative relationship between stock trading volume and closing price for both Microsoft and Cisco. However, the models have limitations and do not capture a significant proportion of the variability in stock prices, implying that other factors aside from volume are likely playing a more significant role in determining stock prices.

**QCOM (Qualcomm) Model Summary:**

* **Dependent Variable**: Close/Last (stock closing price)
* **R-squared**: 0.023 - This means that only 2.3% of the variability in Qualcomm's stock closing price can be explained by the Volume of stocks traded. This is a relatively low value, suggesting that trading volume is not a strong predictor of the stock closing price.
* **F-statistic & Prob (F-statistic)**: The F-statistic is 58.32, and its associated p-value is close to 0. This suggests that the model is statistically significant, but as with the Microsoft model, the practical significance may be limited due to the low R-squared value.
* **Coefficient for Volume**: -7.307e-07. This means that for every unit increase in volume, we expect a decrease of approximately 0.0000007307 in the stock price, all else being equal.
* **P-value for Volume**: The p-value is close to 0, indicating that volume is a significant predictor of the closing price for Qualcomm's stock.
* **Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.
* **Jarque-Bera Test**: Further confirms the non-normality of residuals.
* **Condition Number**: Is large, suggesting potential multicollinearity issues.

**META (Meta Platforms, Inc.) Model Summary:**

* **Dependent Variable**: Close/Last (stock closing price)
* **R-squared**: 0.147 - This means that 14.7% of the variability in Meta's stock closing price is explained by the Volume of stocks traded. This is a relatively higher proportion compared to the other models we've examined so far.
* **F-statistic & Prob (F-statistic)**: The F-statistic is 432.7, and its associated p-value is very close to 0, suggesting that this model is statistically significant.
* **Coefficient for Volume**: -1.351e-06. For every unit increase in volume, we expect a decrease of approximately 0.000001351 in the stock price, all else being equal.
* **P-value for Volume**: The p-value is close to 0, indicating that volume is a significant predictor of the closing price for Meta's stock.
* **Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.
* **Jarque-Bera Test**: Further confirms the non-normality of residuals.
* **Condition Number**: Is large, suggesting potential multicollinearity issues.

**Key Takeaways:**

* **Role of Volume**: For both Qualcomm and Meta, there's a negative relationship between volume and stock closing price. This trend is consistent with the other models we've analyzed.
* **Significance of the Models**: Both models are statistically significant as indicated by the F-statistic and its associated p-value. Meta's model has a higher R-squared value (14.7%) compared to Qualcomm (2.3%), implying that trading volume is a more important predictor of Meta's stock price than Qualcomm's.
* **Model Assumptions**: For both companies, the assumption of normally distributed residuals is violated (as indicated by Omnibus and Jarque-Bera tests). This could affect the reliability of the OLS estimates.
* **Multicollinearity Warning**: The high condition numbers indicate potential multicollinearity. This could be due to other unconsidered factors that are influencing both the volume and closing price and should be investigated further.

Overall, these models show some level of a negative relationship between stock trading volume and closing price for both Qualcomm and Meta. As with the previous models, these models have limitations and do not capture a significant proportion of the variability in stock prices, implying that other factors aside from volume are likely playing a more significant role in determining stock prices.

**AMZN (Amazon) Model Summary:**

* **Dependent Variable**: Close/Last (stock closing price)
* **R-squared**: 0.000 - This means that none of the variability in Amazon's stock closing price can be explained by the Volume of stocks traded. The R-squared value is essentially zero, indicating that trading volume has no explanatory power in this model.
* **F-statistic & Prob (F-statistic)**: The F-statistic is 0.2644, and the p-value is 0.607. This indicates that the model is not statistically significant, as the p-value is greater than the common alpha level of 0.05.
* **Coefficient for Volume**: 1.279e-08. This means that for every unit increase in volume, we expect an increase of approximately 0.00000001279 in the stock price, all else being equal.
* **P-value for Volume**: The p-value is 0.607, indicating that volume is not a significant predictor of the closing price for Amazon's stock.
* **Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.
* **Jarque-Bera Test**: Further confirms the non-normality of residuals.
* **Condition Number**: Is large, suggesting potential multicollinearity issues.

**TSLA (Tesla) Model Summary:**

* **Dependent Variable**: Close/Last (stock closing price)
* **R-squared**: 0.005 - This means that only 0.5% of the variability in Tesla's stock closing price is explained by the Volume of stocks traded. This is a very low value, suggesting that trading volume is not a strong predictor of the stock closing price.
* **F-statistic & Prob (F-statistic)**: The F-statistic is 13.19, and its associated p-value is close to 0. This suggests that the model is statistically significant, but the practical significance may be limited due to the low R-squared value.
* **Coefficient for Volume**: -9.689e-08. For every unit increase in volume, we expect a decrease of approximately 0.00000009689 in the stock price, all else being equal.
* **P-value for Volume**: The p-value is close to 0, indicating that volume is a significant predictor of the closing price for Tesla's stock.
* **Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.
* **Jarque-Bera Test**: Further confirms the non-normality of residuals.
* **Condition Number**: Is large, suggesting potential multicollinearity issues.

**Key Takeaways:**

* **Role of Volume**: For Amazon, volume doesn't seem to play a significant role in determining the stock closing price. For Tesla, there is a negative relationship between volume and closing price, but it explains only a tiny fraction of the variability in the closing price.
* **Significance of the Models**: The Tesla model is statistically significant, while the Amazon model is not. However, for both models, the R-squared values are extremely low, suggesting that trading volume isn't a strong predictor of stock closing prices.
* **Model Assumptions**: For both companies, the assumption of normally distributed residuals is violated (as indicated by Omnibus and Jarque-Bera tests). This could affect the reliability of the OLS estimates.
* **Multicollinearity Warning**: The high condition numbers indicate potential multicollinearity. This could be due to other unconsidered factors that are influencing both the volume and closing price and should be investigated further.

In conclusion, these models highlight that stock trading volume has limited predictive power for stock closing prices for both Amazon and Tesla. Other factors, not included in these models, are likely playing a more significant role in determining stock prices. Also, the violation of model assumptions (especially normality of residuals) and potential multicollinearity are areas of concern that need to be addressed for more reliable model results.

**AMD (Advanced Micro Devices) Model Summary:**

* **Dependent Variable**: Close/Last (stock closing price)
* **R-squared**: 0.128 - This means that 12.8% of the variability in AMD's stock closing price is explained by the Volume of stocks traded. This indicates that trading volume explains a relatively small portion of the variability in the stock closing price.
* **F-statistic & Prob (F-statistic)**: The F-statistic is 368.6, and the p-value is very close to 0. This indicates that the model is statistically significant, suggesting that the relationship between Volume and Closing Price is unlikely to be due to chance.
* **Coefficient for Volume**: 3.773e-07. This means that for every unit increase in volume, we expect an increase of approximately 0.0000003773 in the stock price, all else being equal.
* **P-value for Volume**: The p-value is very close to 0, indicating that volume is a significant predictor of the closing price for AMD's stock.
* **Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.
* **Jarque-Bera Test**: Further confirms the non-normality of residuals.
* **Condition Number**: Is large, suggesting potential multicollinearity issues.

**NFLX (Netflix) Model Summary:**

* **Dependent Variable**: Close/Last (stock closing price)
* **R-squared**: 0.192 - This means that 19.2% of the variability in Netflix's stock closing price is explained by the Volume of stocks traded. This indicates that trading volume explains a moderate portion of the variability in the stock closing price.
* **F-statistic & Prob (F-statistic)**: The F-statistic is 597.2, and the p-value is very close to 0. This indicates that the model is statistically significant, suggesting that the relationship between Volume and Closing Price is unlikely to be due to chance.
* **Coefficient for Volume**: -7.311e-06. For every unit increase in volume, we expect a decrease of approximately 0.000007311 in the stock price, all else being equal.
* **P-value for Volume**: The p-value is very close to 0, indicating that volume is a significant predictor of the closing price for Netflix's stock.
* **Omnibus & Prob(Omnibus)**: Indicates that the residuals are not normally distributed.
* **Jarque-Bera Test**: Further confirms the non-normality of residuals.
* **Condition Number**: Is large, suggesting potential multicollinearity issues.

**Key Takeaways:**

* **Role of Volume**: For both AMD and Netflix, trading volume appears to be a statistically significant predictor of stock closing price, but the explanatory power is not very high.
* **Significance of the Models**: The models for both AMD and Netflix are statistically significant as evidenced by the low p-values for the F-statistics. However, the practical significance is limited due to the relatively low R-squared values.
* **Model Assumptions**: For both companies, the assumption of normally distributed residuals is violated (as indicated by Omnibus and Jarque-Bera tests). This could affect the reliability of the OLS estimates.
* **Multicollinearity Warning**: The high condition numbers indicate potential multicollinearity. This could be due to other unconsidered factors that are influencing both the volume and closing price and should be investigated further.

In conclusion, while trading volume appears to have a significant relationship with stock closing prices for both AMD and Netflix, it doesn't explain a large proportion of the variability in closing prices. Other factors, not included in these models, are likely playing a more significant role in determining stock prices. Also, the violation of model assumptions (especially normality of residuals) and potential multicollinearity are areas of concern that need to be addressed for more reliable model results

For each stock, there are two rows of information: one for the constant term (intercept) and one for the volume.

**AAPL (Apple):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [105.2552, 111.9099].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-2.748927e-07, -2.375101e-07].

**SBUX (Starbucks):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [79.901878, 83.801438].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-0.000002, -0.000001].

**MSFT (Microsoft):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [147.5127, 163.7265].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-8.745175e-07, -4.107084e-07].

**CSCO (Cisco Systems):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [45.37393, 47.11685].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-3.248506e-07, -2.617060e-07].

**QCOM (Qualcomm):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [92.37713, 97.23638].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-9.182979e-07, -5.430584e-07].

**META (Meta Platforms):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [199.858463, 209.171595].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-0.000001, -0.000001].

**AMZN (Amazon):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [73.40386, 82.20259].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-3.597725e-08, 6.155319e-08].

**TSLA (Tesla):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [87.15526, 101.7070].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-1.491941e-07, -4.458534e-08].

**AMD (Advanced Micro Devices):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [15.39253, 20.44012].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [3.387561e-07, 4.158296e-07].

**NFLX (Netflix):**

* **Intercept (const)**: We are 95% confident that the true value of the intercept falls within the range [330.026381, 347.759063].
* **Volume**: We are 95% confident that the true value of the coefficient for Volume falls within the range [-0.000008, -0.000007].

The above results show the ranges within which we can be 95% confident that the true values of the intercepts and coefficients for Volume lie for each of the ten stocks. These intervals give us an indication of the precision of our estimates and an understanding of the possible range of values that the coefficients can take in the population. This can be particularly useful for understanding the uncertainty associated with our regression estimates.

Here are the key takeaways for all 10 stocks:

**Role of Volume:** For most of the stocks (AAPL, SBUX, MSFT, CSCO, QCOM, META, AMZN, TSLA, NFLX), trading volume appears to be a statistically significant predictor of stock closing price. The relationship is negative for all these stocks except AMD, where volume has a positive relationship with the closing price. However, the explanatory power of volume is not very high for any of the stocks, as evidenced by the low R-squared values.

**Significance of the Models:** The models for all ten stocks are statistically significant, as indicated by the low p-values for the F-statistics. However, the practical significance is limited due to the relatively low R-squared values. This suggests that other factors, not included in the models, are also influencing the closing prices.

**Model Assumptions:** The assumption of normally distributed residuals is violated for most of the stocks, as indicated by the Omnibus and Jarque-Bera tests. This could affect the reliability of the OLS estimates. Additionally, the Durbin-Watson statistics are close to zero for all the stocks, indicating potential autocorrelation in the residuals.

**Multicollinearity Warning:** The condition numbers are quite large for most of the stocks, indicating potential multicollinearity. This could be due to other unconsidered factors that are influencing both the volume and closing price and should be investigated further.

**Interpretation of Coefficients:** The coefficients for the Volume variable are negative for most of the stocks, indicating that as the volume increases, the closing price decreases. However, for AMD, the coefficient is positive, indicating that as the volume increases, the closing price also increases.

**Confidence Intervals:** The confidence intervals for the coefficients provide an estimate of the range within which the true value of the coefficient is likely to fall. For most of the stocks, the confidence intervals for the Volume coefficients are relatively narrow, indicating a high level of precision in the estimates.

In conclusion, while trading volume appears to have a significant relationship with stock closing prices for all ten stocks, it doesn't explain a large proportion of the variability in closing prices. Other factors, not included in these models, are likely playing a more significant role in determining stock prices. Also, the violation of model assumptions (especially normality of residuals and potential autocorrelation) and potential multicollinearity are areas of concern that need to be addressed for more reliable model results.

Research Question 4

As for the results you've posted, let's interpret them:

* **Accuracy:** 51.2% - This means that the model is correct 51.2% of the time in predicting whether a stock will go up or down. This is slightly better than random guessing but not much.
* **Precision:** 48.2% - This metric measures the proportion of positive identifications that were actually correct. In this case, it means that when the model predicts a stock will go up, it is correct 48.2% of the time.
* **Recall:** 48.1% - This metric measures the proportion of actual positives that were correctly identified. In this case, it means that the model correctly predicts 48.1% of the times when the stock actually goes up.
* **F1 Score:** 48.1% - The F1 Score is a measure of a model's accuracy that considers both precision and recall. It's a good way to measure a model's accuracy when the classes are imbalanced.
* **Feature Coefficients:** The coefficients indicate the strength and direction of the relationship between each feature and the target variable. However, the coefficients are quite small, suggesting that the features have a minimal effect on the model's predictions.
* **ROC AUC Score:** 49.9% - The Area Under the Receiver Operating Characteristic curve (ROC AUC) measures the ability of the model to distinguish between the classes. A score of 50% indicates no discriminative power, and the score you got is very close to that.

Based on these results, the model doesn't seem to perform very well in predicting whether a stock will go up or down based on the previous day's features. It's important to note that predicting stock market movements is a complex task

**Interpretation of Research Questions**

* **How does the distribution of trading volume and stock prices (Close/Last, Open, High, Low) vary across the top 10 companies in the stock market, and how do these variables correlate with each other?**
  + The distribution of trading volume and stock prices varied across the top 10 companies. Some companies exhibited higher average trading volume and stock prices than others.
  + The correlation between trading volume and stock prices was weak for most companies. For the companies where there was a statistically significant relationship, the effect size was small, suggesting that trading volume does not play a significant role in determining stock prices for these companies.
* **Which Machine Learning Model Can be best implemented to predict a stock's closing price?**
  + Linear Regression models showed limited success in predicting stock closing prices using trading volume as a predictor. Although the models were statistically significant, their practical significance was limited due to low R-squared values.
  + The models were further affected by violations of assumptions, including normality of residuals and potential multicollinearity. It is recommended to consider additional features in the model, perform transformation of the data, or try different machine learning models.
  + For better predictive performance, more sophisticated models like Random Forest, XGBoost, or ARIMA models should be explored.
* **Do certain companies show a more consistent stock price pattern than others?**
  + Certain companies did show more consistent stock price patterns than others, as evidenced by the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.
  + Stocks with a more consistent pattern could be better suited for time series models like ARIMA or GARCH for predicting future stock prices or volatility.
* **Can we predict whether a stock will go up or down based on its previous day's features?**
  + The logistic regression model showed limited success in predicting whether a stock will go up or down based on its previous day's features. The accuracy, precision, recall, and F1 score were all around 50%, which is equivalent to random guessing.
  + The feature coefficients were small, suggesting a minimal effect of the previous day's features on the model's predictions.
  + The results highlight the complexity of predicting stock price movements and the need for a comprehensive approach that considers additional features, more sophisticated models, and a larger dataset for better predictive performance.

**Closing**

In conclusion, predicting stock market volatility is a complex task that requires a multifaceted approach. Our analysis revealed that trading volume does not play a significant role in determining stock prices, and the patterns in stock prices varied across companies. The linear regression and logistic regression models had limited success in predicting stock closing prices and the direction of stock price movements, respectively.

For better predictive performance, more sophisticated machine learning models, additional features, and larger datasets should be considered. Moreover, the impact of external factors such as market news, economic indicators, and geopolitical events on stock market volatility should also be taken into account.

It is essential for investors to approach stock market investments with a comprehensive strategy, considering both quantitative analysis and qualitative insights, to make informed decisions and manage risks effectively.