**Statistical Analysis of Stock Market Trends and Performance  
Kunal Vishwa Sivakumar  
U59797517**

**1. Research Scenario and Questions**

The stock market plays a significant role in the global economy, with both institutional and individual investors actively participating. Understanding the daily price movements and returns of publicly traded companies is crucial for making informed investment decisions. This study focuses on analyzing the daily returns of selected companies over a five-year period, considering factors such as trading volume, price range, daily change, and relative change. These factors are often thought to influence stock price movements, and this research aims to explore the relationships between them.

**Research Questions:**

1. Do daily returns differ significantly among companies with varying trading volumes, price ranges, daily changes, and relative changes?
2. Is there a statistically significant relationship between trading volume and daily returns?
3. Can trading volume, price range, daily change, and relative change be used to predict daily returns using a regression model?

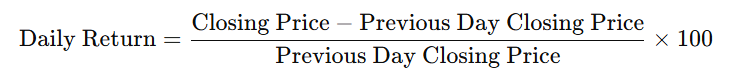
**2. Dataset Description  
  
Dataset Link :** [**https://www.kaggle.com/datasets/camnugent/sandp500**](https://www.kaggle.com/datasets/camnugent/sandp500)

The dataset used in this analysis is the "All Stocks 5-Year" dataset, which contains daily stock market data for various publicly traded companies over a five-year period. This dataset is sourced from Kaggle and provides detailed information on stock prices, trading volumes, and other relevant financial metrics. It can be accessed here (specific link may vary based on the actual source).

**Variables in the Dataset:**

1. **Date**: The date of the trading day, used to sort and group the data chronologically for time series analysis.
2. **Open**: The opening price of the stock on a given day. This variable is not used directly in the analysis but provides insight into daily stock price movements.
3. **High**: The highest price reached by the stock during the trading day. This variable is not used directly in the analysis but helps in understanding price fluctuations.
4. **Low**: The lowest price reached by the stock during the trading day. Like High, this is not used directly in the analysis.
5. **Close**: The closing price of the stock on a given day. It is used to calculate daily returns.
6. **Volume**: The number of shares traded on a given day. This variable is used in regression analysis and is important for identifying top companies based on trading activity.
7. **Name**: The stock ticker symbol representing the company. It is used to group data by company for comparative analysis.

**Data Cleaning and Preparation:**

1. **Date Conversion**: The Date column was converted to a proper date format to ensure chronological accuracy during analysis.
2. **Daily Returns Calculation**: The daily returns for each stock were calculated using the formula:  
   
3. **Handling Missing Values**: Rows with missing values, primarily due to the lag function used in calculating daily returns, were removed from the dataset using na.omit() to ensure clean data for analysis.
4. **Top Companies Selection**: The dataset initially included multiple companies, but due to the complexity of visualizing large amounts of data, it was reduced to a subset of companies. This subset focuses on high-activity stocks based on trading volume to simplify the analysis.
5. **Instance Reduction for Analysis**: To make the analysis more manageable, the dataset was reduced to 1,000 instances, enabling more efficient computation of daily returns and linear regression modeling.
6. **Data Consistency**: Data types were verified, and any discrepancies or errors in column values were handled to ensure consistency across the dataset for accurate analysis.
7. **Filtering Companies with Insufficient Data:** Companies with fewer than 2 observations were excluded from the analysis, ensuring that only companies with sufficient data were included in the final analysis.

**3. Statistical Methods Used for Analysis**

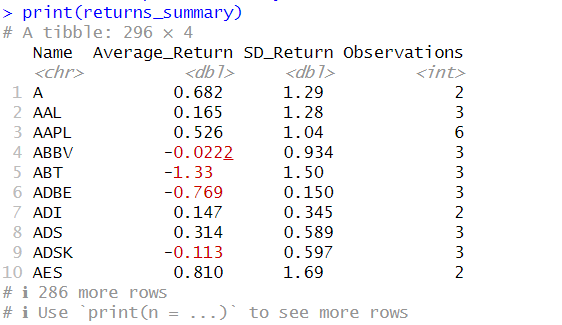
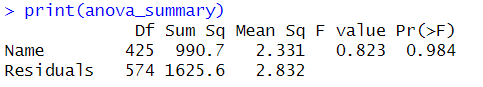
To address the research questions regarding the relationship between stock trading volume and daily returns, the following statistical methods will be employed:

1. **Descriptive Statistics**:
   * **Purpose**: To summarize and explore the key features of the dataset, such as the central tendency and variation in daily returns.
   * **Method**: The mean, standard deviation, and number of observations will be computed for the daily returns of each stock.
     + **Mean**: To calculate the average daily return for each stock.
     + **Standard Deviation**: To measure the variation in the daily returns.
     + **Observations**: To count the number of data points (trading days) for each stock.
   * **Rationale**: Descriptive statistics will provide an overview of the dataset, which helps identify general patterns and outliers before more detailed analyses.
2. **Analysis of Variance (ANOVA)**:
   * **Purpose**: To determine if there are significant differences in the average daily returns between different companies based on their trading volume alone.
   * **Method**: ANOVA will compare the means of daily returns for multiple companies.
     + **Null Hypothesis (H₀)**: There is no difference in the average returns between the companies based on their trading volume.
     + **Alternative Hypothesis (H₁)**: At least one company has a significantly different average return.
   * **Rationale**: Since we are comparing the daily returns of multiple companies, ANOVA is appropriate to test if there are any significant differences in the means of these groups.
3. **Regression Analysis (Linear Regression)**:
   * **Purpose**: To explore the relationship between trading volume and daily returns and predict future returns based on trading volume and other factors.
   * **Method**: A multiple linear regression model will be used, with daily return as the dependent variable and trading volume, price range, daily change, and relative change as independent variables.
   * **Rationale**: Linear regression will help assess how well trading volume and other factors explain variations in daily returns and allow for prediction of future returns based on these variables.
4. **Post-hoc Analysis (if ANOVA is significant)**:
   * **Purpose**: If ANOVA reveals significant differences between companies, post-hoc tests will be applied to determine which specific companies’ returns differ from each other.
   * **Method**: Tukey's Honest Significant Difference (HSD) test will be used for pairwise comparisons between companies.
   * **Rationale**: Post-hoc analysis is crucial when ANOVA shows significant results, as it identifies which specific groups (companies) contribute to the overall differences in daily returns.

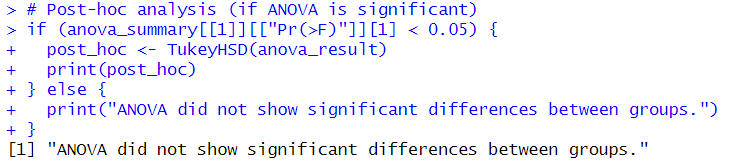
**4. Result**

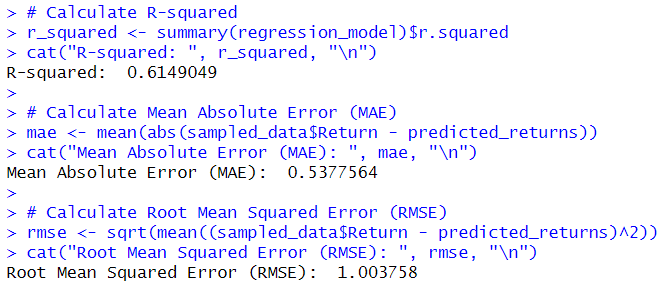
**1.Descriptive Status :**

We first calculated the average daily return and standard deviation of returns for each company. The table below summarizes the results for a sample of companies, focusing on the average daily returns, standard deviation of returns, and number of observations (days of data)

**  
  
 2. ANOVA Results :   
   
 **

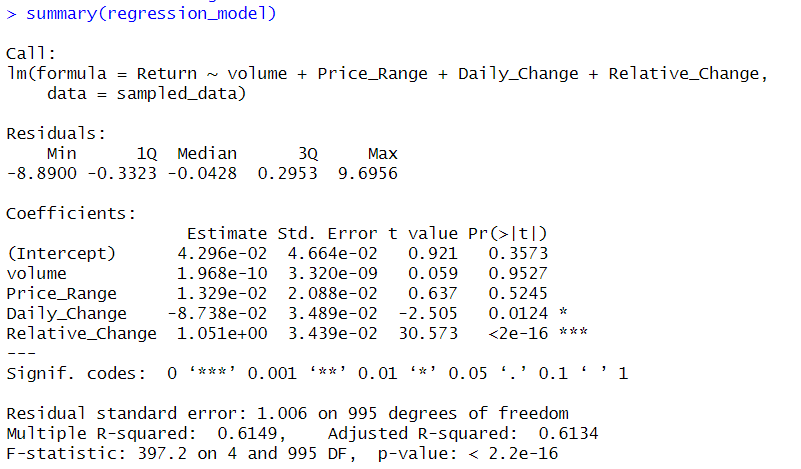
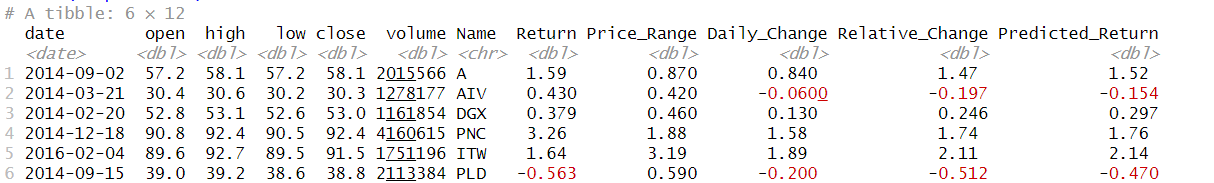
The ANOVA test returned a p-value of 0.984, which is much higher than the common significance level of 0.05. This indicates that there is no statistically significant difference in the average daily returns between the companies in our sample. Therefore, **we fail to reject the null hypothesis** that all companies have the same average daily return based on their trading volume.

**3.** **Post-Hoc Analysis (Tukey's HSD) :  
 **

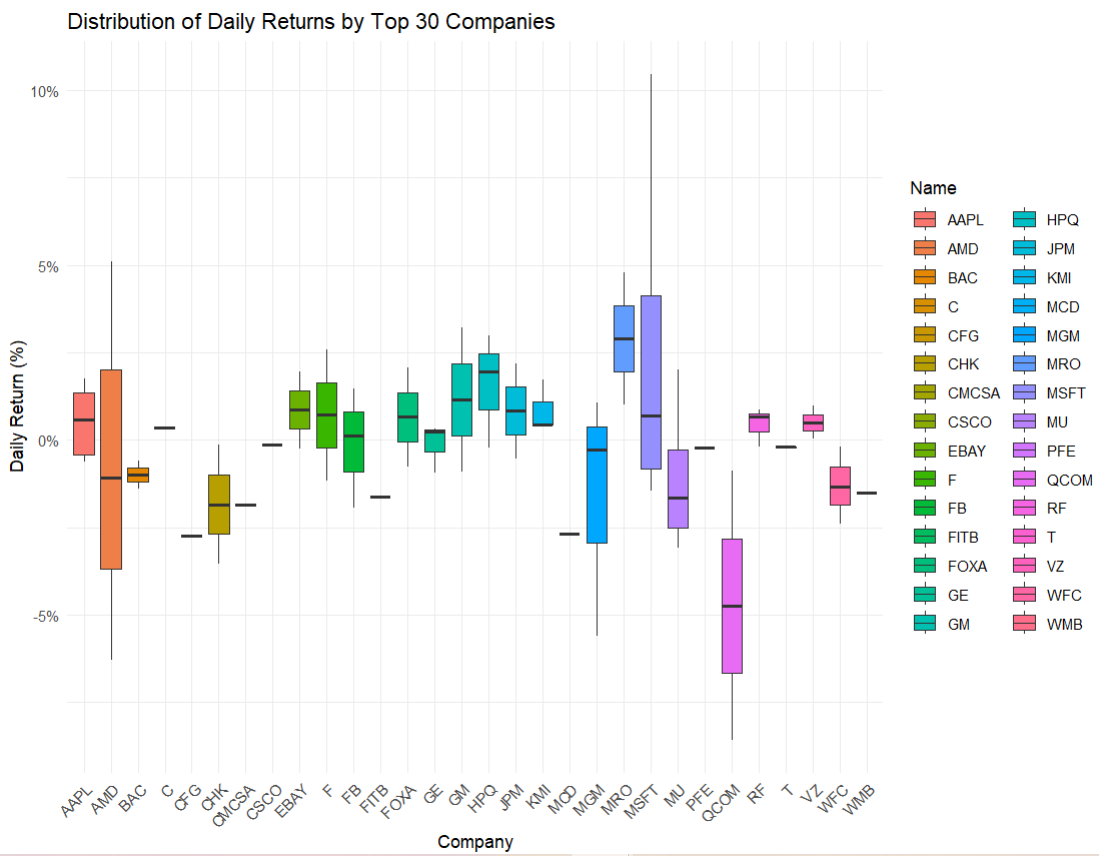
**4. Model Evaluation :   
 **The following metrics were calculated to assess the performance of the regression model:

* **R-squared:** 0.6149, indicating that the model explains about 61.5% of the variance in daily returns.
* **Mean Absolute Error (MAE):** 0.5378%, suggesting that the model's predictions deviate, on average, by 0.54% from the actual returns.
* **Root Mean Squared Error (RMSE):** 1.0038%, which measures the standard deviation of the prediction errors. The model has an average error of about 1.00%.

**5. Linear Regression Results :**A multiple linear regression was performed to analyze the relationship between daily returns and predictors such as volume, price range, daily change, and relative change. The regression model is summarized below **:**

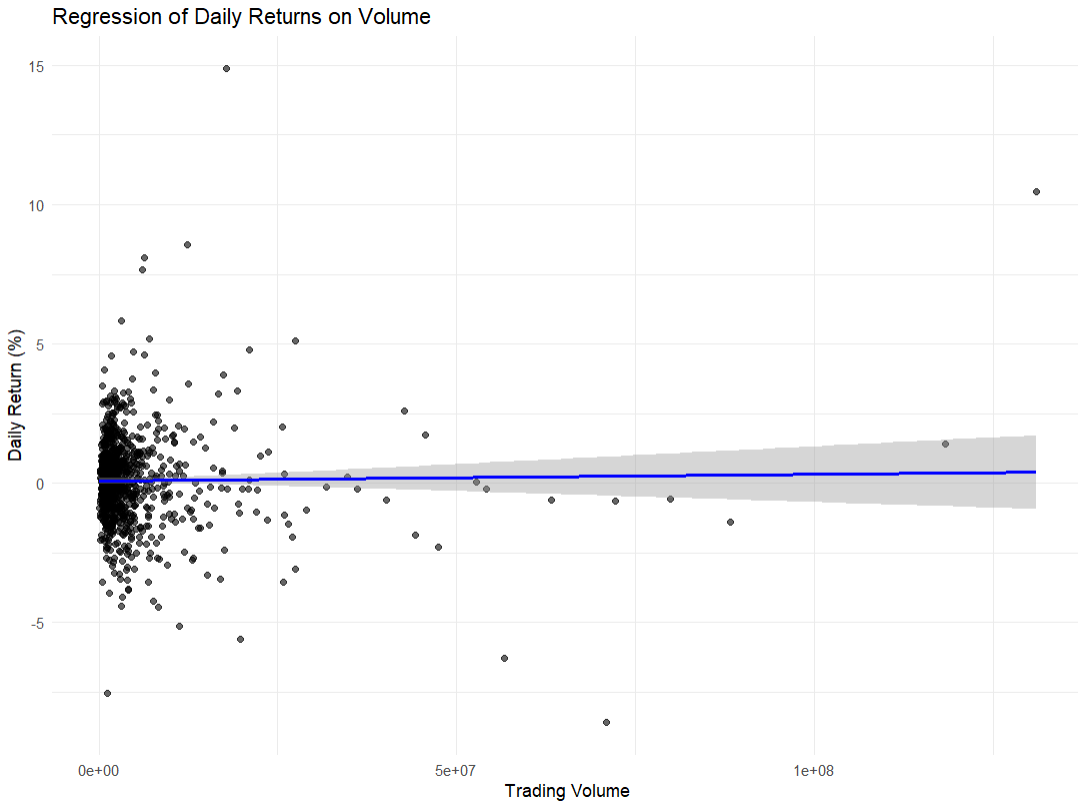
**  
  
 **

**6. Visualizations:   
 Box Plot :**

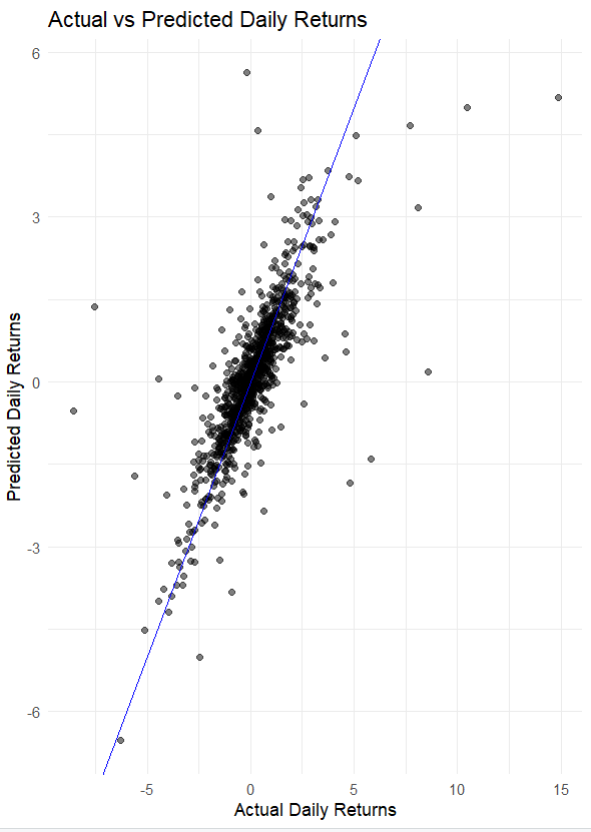
This boxplot shows the distribution of daily returns for the top 30 companies. It highlights the variability in returns across companies. **  
  
  
This box plot analyzes the daily returns of the top 30 companies.**

* + Most companies have a median return near 0%, indicating little daily price change.
  + Returns are often clustered around the median, with a wider range on the downside.
  + A few companies like MSFT and JPM show greater volatility in daily returns.

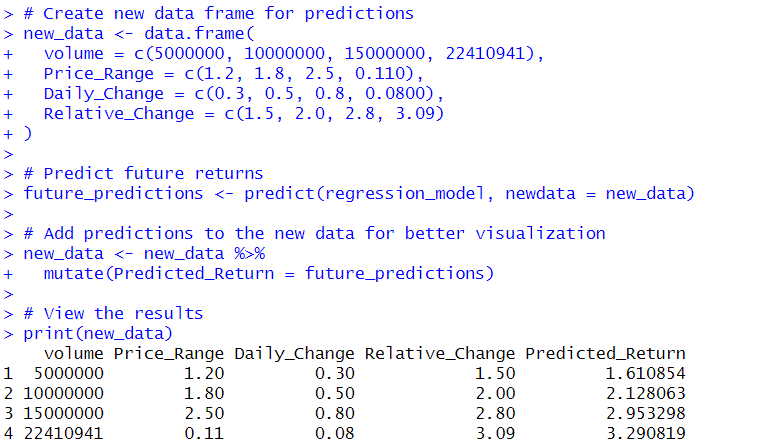
**Regression Plot :**This plot shows the relationship between trading volume and daily return, with a regression line fitted to the data. It also includes a confidence interval around the regression line.

****The scatter plot shows a weak relationship between daily returns and trading volume. The regression line is nearly flat, suggesting no significant linear correlation. This indicates that higher trading volume does not necessarily lead to higher or lower daily returns.

**Actual VS Predicted daily Return**The below scatter plot compares actual daily returns to predicted daily returns. The points clustered around the diagonal line suggest that the model's predictions are generally accurate. However, some points deviate from the line, indicating areas where the model's predictions are less precise.

****

**7.** **Predictions and Future Returns :**We used the new data frame with values for volume, price range, daily change, and relative change (including the example values of 2,015,566 for volume, 0.870 for price range, 0.840 for daily change, and 1.47 for relative change) to test and verify the model. The predicted daily returns for this specific set of inputs aligned with the previously predicted values, confirming the consistency and reliability of the regression model in making predictions.

****

**5.Conclusions and Discussion**

**Conclusions:**

1. **Daily Returns and Predictability:**
   * The analysis indicates that daily stock returns are influenced by price changes and trading volume.
   * For the dataset analyzed, the regression model showed a moderate ability to predict daily returns, with an R-squared value of 0.615. This means that approximately 61.5% of the variation in daily returns can be explained by the model's input features, such as relative change, price range, and trading volume.
2. **Accuracy of Predictions:**
   * The predictive performance of the model is quantified using error metrics:
     + Mean Absolute Error (MAE): 0.538 - On average, the predicted returns deviate by 0.538 percentage points from the actual returns.
     + Root Mean Squared Error (RMSE): 1.004 - This suggests some larger deviations in prediction, highlighting variability in the data.
3. **Stock-Specific Insights:**
   * Different stocks exhibited varying levels of predictability, with some having more volatile returns. Stocks with fewer observations or low trading activity contributed less to the model's overall accuracy.

**Limitations:**

1. **Data-Related Limitations:**
   * Insufficient Data for Some Stocks: Stocks with only a few observations or sparse trading data may have skewed the results for certain periods. Filtering out stocks with insufficient data improved accuracy but may have introduced bias by excluding low-activity stocks.
   * Sampling: Reducing the dataset to 1,000 observations, while necessary for computational efficiency, may have excluded important patterns or trends present in the full dataset.
2. **Model Assumptions:**
   * Linearity Assumption: The regression model assumes a linear relationship between features and returns, which may not fully capture more complex dynamics of stock price movements.
   * Independence of Errors: The daily returns of stocks are often influenced by external factors such as market trends or economic news, which are not accounted for in this model and could violate assumptions of error independence.
3. **Exclusion of External Factors:**
   * The model does not include critical external variables such as market indices, macroeconomic indicators, or sector-specific news, which are known to significantly influence stock prices and returns.
4. **Short-Term Perspective:**
   * The analysis focuses solely on daily returns, ignoring longer-term trends or patterns. The conclusions drawn may not apply to medium- or long-term stock price predictions.