

**BU MET CS555**  
**Foundations of Machine Learning**  
**Final Project**

**Title: Stock Performance Analysis of AAPL, TSLA, and BA for Strategic Investment Decisions**

Sanjana K R Prasad  
U35186791

**A. Research Scenario**

This study aims to reveal statistically significant patterns in the daily stock returns of three selected influential individual stocks:

- Apple Inc. (AAPL) from the technology sector
- Tesla Inc. (TSLA) from the automotive sector and
- Boeing Co. (BA) from the aerospace sector.

Insights gained from the study are expected to contribute to a deeper understanding of their unique performance dynamics, relationships, and the effectiveness of trading strategies applied to each stock independently. The findings aim to provide valuable insights for investors focusing on these specific stocks. Investors may choose these stocks for their diversity across sectors.

The following research questions have been asked:

- *What role does trading volume play in predicting daily returns of these stocks?*
- *To what extent do lagged daily returns, trading volume, and lagged daily returns of the S&P 500 (SPY) predict daily returns?*
- *Do the daily returns of stocks from the technology (AAPL), automotive (TSLA), and aerospace (BA) sectors significantly differ?*
- *Is there a significant difference in the effectiveness of the 10-day SMA strategy between BA and TSLA?*
- *To what extent can stock type and trading volume predict buy signals, and how reliable is this prediction across different stocks?*

## B. Dataset

- 3 years worth of historical stock data (2020, 2021, 2022) has been downloaded from yahoo finance (<https://finance.yahoo.com/>) for the following stocks
  - AAPL (Apple)
  - TSLA (Tesla)
  - BA (Boeing)
  - SPY (Tracks the performance of the S&P 500, which is a stock market index comprising 500 of the largest publicly traded companies in the United States, and is often used for benchmarking.)

Each stock has the following columns - *Date*, *Open*, *High*, *Low*, *Close*, *Adj Close*, *Volume*.

- A column called “*Daily\_Returns*” has been added to each stock. The daily returns are calculated as follows:
$$\text{Daily return} = (\text{current closing price} - \text{previous day's closing price}) / \text{previous day's closing price}$$
- The data has been merged based on the “*Date*” column.
- The following columns have been selected as relevant to our analysis

**Date:** The date of the stock market data.

**Close\_AAPL:** The closing price of Apple Inc. stock on a particular date.

**Volume\_AAPL:** The total number of shares traded for Apple Inc. stock on a given date.

**Daily\_Returns\_AAPL:** The percentage change in the closing price of Apple Inc. stock from the previous day.

**Close\_TSLA:** The closing price of Tesla Inc. stock on a particular date.

**Volume\_TSLA:** The total number of shares traded for Tesla Inc. stock on a given date.

**Daily\_Returns\_TSLA:** The percentage change in the closing price of Tesla Inc. stock from the previous day.

**Close\_BA:** The closing price of The Boeing Company stock on a particular date.

**Volume\_BA:** The total number of shares traded for The Boeing Company stock on a given date.

**Daily\_Returns\_BA:** The percentage change in the closing price of The Boeing Company stock from the previous day.

**Close\_SPY:** The closing price of the S&P 500 index on a particular date.

**Daily\_Returns\_SPY:** The percentage change in the closing price of the S&P 500 index from the previous day.

First 5 columns of the dataset, (number of rows = 756)

	Date	Close_AAPL	Volume_AAPL	Daily_Returns_AAPL	Close_TSLA	Volume_TSLA	Daily_Returns_TSLA
1	2020-01-02	75.0875	135480400	NA	28.68400	142981500	NA
2	2020-01-03	74.3575	146322800	-0.0097220440	29.53400	266677500	0.0296332450
3	2020-01-06	74.9500	118387200	0.0079682482	30.10267	151995000	0.0192546557
4	2020-01-07	74.5975	108872000	-0.0047030422	31.27067	268231500	0.0388005156
5	2020-01-08	75.7975	132079200	0.0160862889	32.80933	467164500	0.0492048363

Close_BA	Volume_BA	Daily_Returns_BA	Close_SPY	Daily_Returns_SPY
333.32	4544400	NA	324.87	NA
332.76	3875900	-0.0016800582	322.41	-7.572232e-03
333.74	5355000	0.0029450053	323.64	3.815052e-03
337.28	9898600	0.0106070867	322.73	-2.811778e-03
331.37	8239200	-0.0175225451	324.45	5.329535e-03

### C. Statistical Methods

Briefly describe the statistical methods you will be utilizing to investigate your research question(s).

Statistical methods used to investigate research question:

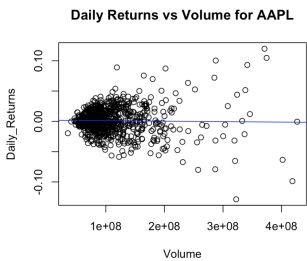
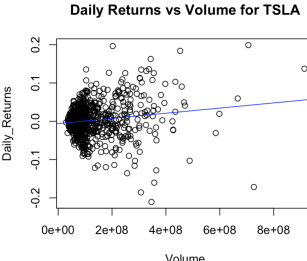
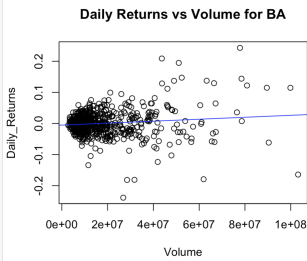
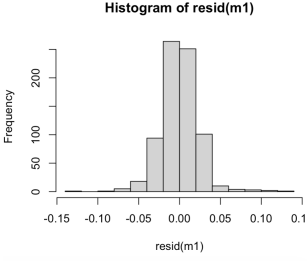
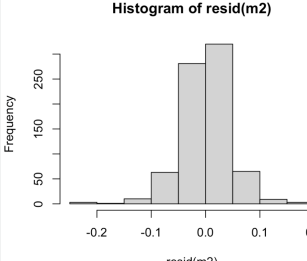
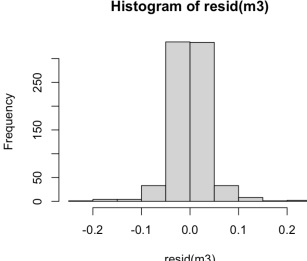
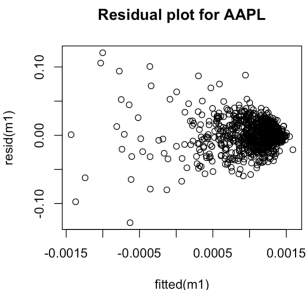
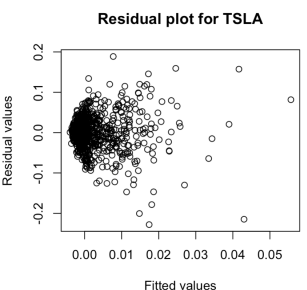
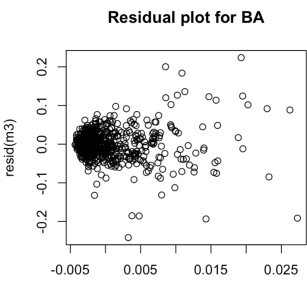
- **Hypothesis Testing:** Evaluates the validity of a claim or hypothesis based on sample data.
- **Correlation Tests:** Measures the strength and direction of the linear relationship between two variables.
- **Simple Linear Regression:** Models the relationship between a dependent variable and a single independent variable.
- **Multiple Linear Regression:** Models the relationship between a dependent variable and multiple independent variables.
- **ANOVA (Analysis of Variance):** Compares means among more than two groups to detect significant differences.
- **Two-sample Tests for Proportions:** Compares proportions between two independent groups.
- **Logistic Regression:** Models the relationship between a binary dependent variable and independent variables.

## D. Results

### 1. Simple Linear Regression

*Question: What role does trading volume play in predicting daily returns of these stocks?*

To check for assumptions:

	AAPL	TSLA	BA
<b>Correlation</b>	-0.019	0.148	0.12
<b>Scatterplot</b>			
<b>Histogram of residuals</b>			
<b>Residual plot</b>			

**Table 1.1:** Assumptions for simple linear regression

The following assumptions that are met:

- The correlation between daily returns of a stock and volume of a stock, is very small for all three stocks.
- The histogram of residuals for all 3 stocks are approximately normally distributed.

We now conduct hypothesis testing to check if there is a linear relationship between daily returns of a stock and trading volume of a stock. We define our hypothesis as follows:

H0: There is no linear association between daily returns of a stock and trading volume of a stock.

H1: There exists a linear association between daily returns of a stock and trading volume of a stock.

Alpha = 0.05

	<b>AAPL</b>	<b>TSLA</b>	<b>BA</b>
<b>F statistic</b>	0.2609	16.96	10.94
<b>F critical</b>	3.853838	3.853838	3.853838
<b>Result</b>	Fail to reject H0	Reject H0	Reject H0

**Table 1.2** : Hypothesis testing for simple linear regression

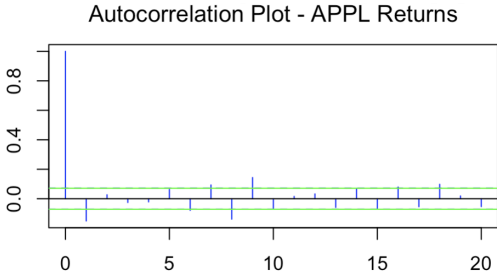
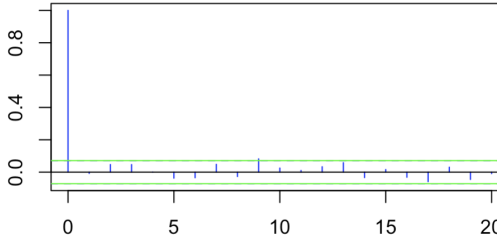
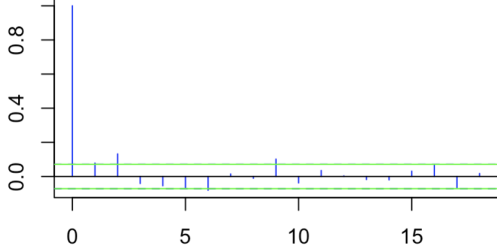
Therefore we can see that the volume of trades affects daily returns of a stock TSLA and BA.

## 2. Multiple Linear Regression

*Question: To what extent do lagged daily returns, trading volume, and lagged daily returns of the S&P 500 (SPY) predict daily returns?*

To select the optimal number of days of lag for our regression problem, we utilize the autocorrelation function. Autocorrelation is the correlation of a signal with a delayed copy of itself as a function of delay. The autocorrelation function (ACF) can help identify the appropriate lag for the lagged daily returns variable.

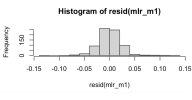
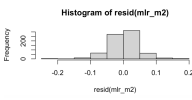
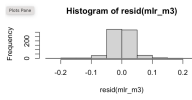
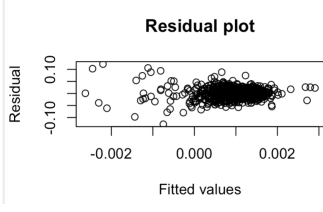
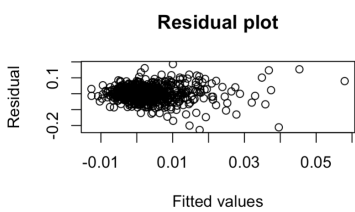
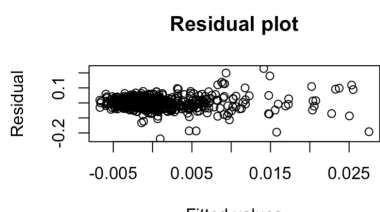
x axis: number of days  
y axis: autocorrelation

Stock	Autocorrelation plots
AAPL	 <p>Autocorrelation Plot - AAPL Returns</p> <p>This plot shows the autocorrelation of AAPL returns over 20 days. The y-axis represents autocorrelation from 0.0 to 0.8, and the x-axis represents the number of days from 0 to 20. A sharp peak is visible at day 0 (autocorrelation ~0.85). A smaller peak is observed at day 9 (autocorrelation ~0.15). Horizontal green dotted lines indicate the 95% confidence interval thresholds at approximately ±0.05.</p>
TSLA	 <p>Autocorrelation Plot - TSLA Returns</p> <p>This plot shows the autocorrelation of TSLA returns over 20 days. The y-axis represents autocorrelation from 0.0 to 0.8, and the x-axis represents the number of days from 0 to 20. A sharp peak is visible at day 0 (autocorrelation ~0.85). A smaller peak is observed at day 9 (autocorrelation ~0.1). Horizontal green dotted lines indicate the 95% confidence interval thresholds at approximately ±0.05.</p>
BA	 <p>Autocorrelation Plot - BA Returns</p> <p>This plot shows the autocorrelation of BA returns over 20 days. The y-axis represents autocorrelation from 0.0 to 0.8, and the x-axis represents the number of days from 0 to 20. A sharp peak is visible at day 0 (autocorrelation ~0.85). A smaller peak is observed at day 9 (autocorrelation ~0.1). Horizontal green dotted lines indicate the 95% confidence interval thresholds at approximately ±0.05.</p>

**Table 2.1:** Autocorrelation plots for the returns of stocks

The horizontal green dotted lines plot the significance thresholds for a two-sided 95% confidence interval. These lines help identify whether autocorrelation values in the plot are statistically significant. Values outside these lines may indicate statistically significant autocorrelation.

After analyzing all three autocorrelation plots, a lag of 9 days on daily return has been identified as having the most influence on the current daily returns.

	AAPL	TSLA	BA
<b>Correlation</b>	<pre> df.Daily_Returns_AAPL df.Lagged_returns_AAPL df.Lagged_returns_SPY df.Volume_AAPL df.Daily_Returns_AAPL 1.0000000 0.81597415 0.81597024 -0.81957004 df.Lagged_returns_AAPL 0.81597415 1.0000000 0.82590272 -0.85015065 df.Lagged_returns_SPY 0.81597024 0.82590272 1.0000000 -0.89795306 df.Volume_AAPL -0.81957004 -0.85015065 -0.89795306 1.0000000 </pre>	<pre> df.Daily_Returns_TSLA df.Lagged_returns_TSLA df.Lagged_returns_SPY df.Volume_TSLA df.Daily_Returns_TSLA 1.0000000 0.81132794 0.87047113 0.14181578 df.Lagged_returns_TSLA 0.81132794 1.0000000 0.5160552 0.87595783 df.Lagged_returns_SPY 0.87047113 0.5160552 1.0000000 -0.8723512 df.Volume_TSLA 0.14181578 0.87595783 -0.8723512 1.0000000 </pre>	<pre> df.Daily_Returns_BA df.Lagged_returns_BA df.Lagged_returns_SPY df.Volume_BA df.Daily_Returns_BA 1.0000000 0.83506394 -0.806814233 0.11973751 df.Lagged_returns_BA 0.83506394 1.0000000 0.646439366 0.82705627 df.Lagged_returns_SPY -0.806814233 0.64643937 1.0000000 0.81593685 df.Volume_BA 0.119737513 0.82705627 0.81593685 1.0000000 </pre>
<b>Histogram of residuals</b>			
<b>Residual plot</b>			

**Table 2.2:** Assumptions for multiple linear regression

We observe the following assumptions that are met:

- Pairwise correlation is very small for all three stocks
- The histogram of residuals for all 3 stocks are approximately normally distributed.

We now conduct hypothesis testing to check if lagged returns of the same stock, trading volume of a stock, and lagged returns of SPY are significant predictors of daily returns of a stock. We define our hypothesis as follows:

H0: Lagged returns of the same stock, trading volume of a stock, and lagged returns of SPY are not significant predictors of daily returns of a stock.

H1: Lagged returns of the same stock, trading volume of a stock, and lagged returns of SPY are significant predictors of daily returns of a stock.

Alpha = 0.05

	AAPL	TSLA	BA
<b>F statistic</b>	7.644	7.513	8.436
<b>F critical</b>	2.617	2.617	2.617
<b>Result</b>	Reject H0	Reject H0	Reject H0

**Table 2.3 :** Hypothesis testing for multiple linear regression

Therefore, lagged returns of the same stock, trading volume of a stock, and lagged returns of SPY are significant predictors of daily returns for stocks TSLA and BA.

### 3. One-Way Analysis of Variance (ANOVA)

*Question: Do the daily returns of stocks of AAPL, TSLA, and BA stocks significantly differ?*

We use ANOVA to check whether the means of returns of the three stocks are significantly different. The assumptions of ANOVA, i.e. independent, random samples from each stock, normal distribution of returns of each stock, and similar standard deviation have been met.

H<sub>0</sub>: means of returns of the three stocks are not significantly different.

H<sub>1</sub>: means of returns of the three stocks are significantly different.

Alpha = 0.05

```
> anova_result <- aov(return ~ stock, data = stacked_data)
> summary(anova_result) #F = 1.267
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
stock	2	0.0034	0.001723	1.267	0.282
Residuals	2262	3.0772	0.001360		

Upon using the aov() function in R, F statistic = 1.267. The value of F critical for df<sub>1</sub>=2, df<sub>2</sub>=2262 is 2.999. Since 1.267 is not greater than 2.999, we fail to reject H<sub>0</sub>.

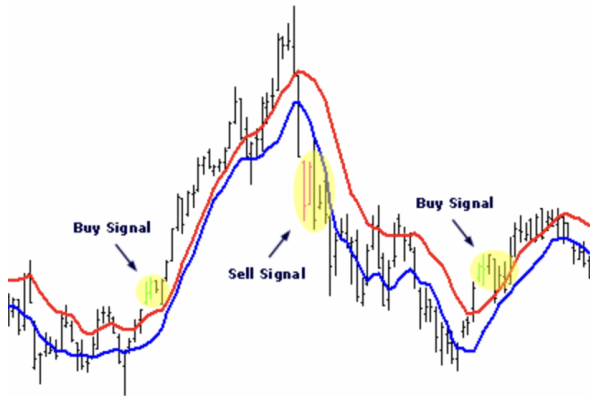
Based on the data and the chosen significance level, we do not have enough evidence to support the claim that there is a significant difference in the mean daily returns of stocks across these sectors. Therefore this method does not provide investors with useful insights.

### 4. Two Sample Test for Proportions

*Question: Is there a significant difference in the effectiveness of the 10-day SMA strategy between BA and TSLA?*

The simple moving average (SMA) strategy has been implemented on the stocks. The 10-day simple moving average (SMA) strategy involves tracking the average closing prices of a stock over the last 10 days.





We aim to test whether the proportion of stocks with buy signals is the same across BA and TSLA.

Signals have been defined as follows:

- When the current closing price crosses above the 10-day SMA, it signals a potential buying opportunity = 1
- When the current closing price crosses below the 10-day SMA, it signals a potential selling opportunity = 0

Summary of buy (1) and sell signals (0) for BA and TSLA.

	stock	signal	count
	<chr>	<dbl>	<int>
1	BA	0	371
2	BA	1	375
3	TSLA	0	326
4	TSLA	1	420

Assumptions met - We can see that both the number of failures (sell) and successes (buy) are greater than 10. They are also independent of each other.

Hypothesis testing:

H0: the proportion of stocks with buy signals is the same across BA and TSLA

H1: the proportion of stocks with buy signals is not the same across BA and TSLA

Alpha = 0.05

After conducting the proportion test, we see that  $p = 0.01954$  which is smaller than  $\alpha = 0.05$ , and we therefore reject H0. The proportion of stocks with buy signals is not the same across BA and TSLA.

2-sample test for equality of proportions without continuity correction

```
data:  c(375, 420) out of c(746, 746)
X-squared = 5.4525, df = 1, p-value = 0.01954
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.110861082 -0.009782349
sample estimates:
   prop 1    prop 2 
0.5026810 0.5630027
```

The percentage of buy signals in BA was 6% lower than those of TSLA. ( $0.56 - 0.50 = 0.06$ ) (Risk difference)

## 5. Multiple Logistic Regression

*Question: To what extent can stock type and trading volume predict buy signals, and how reliable is this prediction across different stocks?*

```
m_logreg <- glm(stacked_data2$signal ~ stacked_data2$stock + stacked_data2$volume, family=binomial)
```

Upon calling the `roc()` function in R, we can observe that the c statistic is 0.5255.

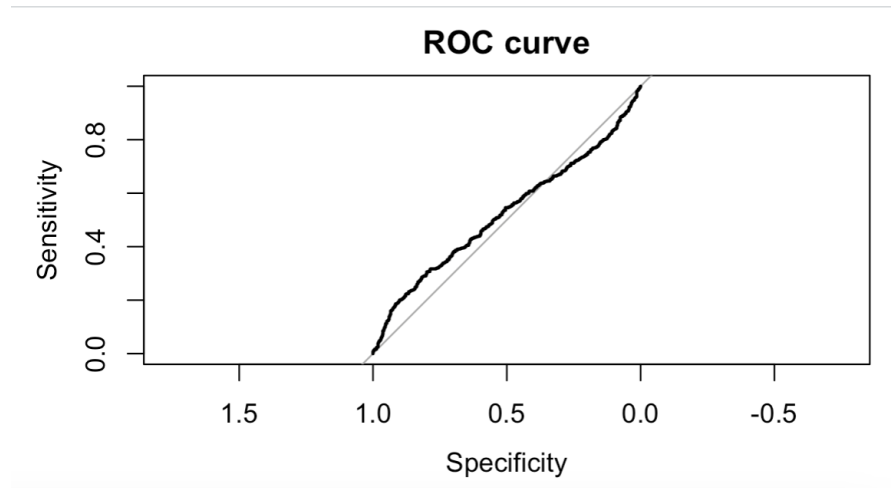
Call:

```
roc.formula(formula = stacked_data2$signal ~ stacked_data2$prob)
```

Data: stacked\_data2\$prob in 697 controls (stacked\_data2\$signal 0) < 795 cases (stacked\_data2\$signal 1).

Area under the curve: 0.5225

We also observe the following ROC curve:



A C-statistic of 0.5255 indicates very poor “goodness of fit”.

## **E. Conclusions**

### **Trading Volume Impact:**

Rejected null hypothesis for TSLA and BA suggests that changes in trading volume contribute to daily return variations for these stocks.

### **Predictive Factors in Multiple Linear Regression:**

Rejected null hypothesis for AAPL, TSLA and LA implies lagged returns, trading volume, and lagged returns of SPY are significant predictors of daily returns. Investors may use these factors for more accurate predictions.

### **Homogeneity in Daily Returns (ANOVA):**

Including stocks from diverse sectors in a portfolio allows the spread of risk across different industries. However the analysis, which indicates that the daily returns of the three stocks may not significantly differ, suggests that the risks associated with individual sectors within the portfolio might be balanced.

Note: It's essential to periodically review and rebalance the portfolio based on changing circumstances.

### **Effectiveness of SMA Strategy (Two Sample Test for Proportions):**

Variation in buy signal proportions between BA and TSLA suggests the 10-day SMA strategy performs differently for these stocks. Tailoring strategies based on sector-specific characteristics may be beneficial.

### **Limited Discriminatory Power (Multiple Logistic Regression):**

AUC of 0.5255 indicates the model's limited ability to discriminate between buy and sell signals based on stock type and trading volume. Caution is advised in relying on these predictors for buy signal predictions.

### **Limitations:**

**Outliers Inclusion:** Keeping outliers to capture market intricacies might introduce noise to the analysis, impacting the results.

**Residual Variances:** The linear regression model assumptions of constant variance are not fully met which might lead to some inconsistencies.

**Scatterplots and Linearity:** The analysis of scatterplots indicates that the relationship between variables might not be strictly linear.

**Data Complexity:** The complexity of stock data introduces challenges in capturing all relevant factors influencing stock returns.

**Generalization Caution for SMA strategy:** Findings are specific to selected stocks and the SMA strategy, cautioning against broad applications to other stocks/ strategies.

**Lurking Variables:** Unaccounted factors (such as geopolitical events, public statements) could be influencing the observed relationships.