Statistical Analysis of Volatility and Distribution Patterns of Global Top 10 Stocks

Final Report

Statistics for IT

Group Members [Group C]:

1. Dhakal Sagar	ID: M25W7443
2. Karki Anushka	ID: M25W7436
3. Ranhoti Pedige Suraj Rangana Sugathapla	ID: M25W7450
4. Nalla Handi Thushan Madushanka De Silva	ID: M25W7449
5. Umange Nethsara Dileesha De Silva	ID: M25W7452

Instructor's Name: Kaewfak Kwanjira

Institution Name: The Kyoto College of Graduate Studies for Informatics

Date: January 14, 2026

Submitted in partial fulfillment of the requirements for the course Statistics for IT.

Abstract

This report presents a statistical analysis of the volatility and return distributions of ten major global technology and industrial companies over a five year period. Utilizing historical data sourced via the Yahoo Finance API, we examine the daily returns of equities including NVIDIA, Alphabet, Apple, and Saudi Aramco. The study employs descriptive statistics, Shapiro-Wilk normality tests, and Pearson correlation coefficients to evaluate market risk and the relationship between trading volume and price volatility. Our findings reject the hypothesis of normality for all selected assets, revealing significant kurtosis and **fat-tail** risks, particularly in the technology sector. Furthermore, a strong positive correlation between volume and volatility is observed in US tech stocks, contrasting with industrial conglomerates. The complete analysis workflow is reproducible via the provided Google Colab notebook.

Contents

Abstract **1**

1 Introduction **1**
1.1 Objective and Statistical Questions 1
1.2 Dataset Overview 1

2 Methodology **3**
2.1 Tools and Environment 3
2.2 Data Availability and Reproducibility 3
2.3 Data Preprocessing 3
2.4 Statistical Techniques 3
2.4.1 Descriptive Statistics & Higher Moments 3
2.4.2 Hypothesis Testing: The Shapiro-Wilk Test 4
2.4.3 Correlation Analysis 4

3 Analysis and Results **5**
3.1 Descriptive Statistics 5
3.2 Visualization of Price Trends 5
3.3 Distribution Analysis 6
3.4 Hypothesis Testing Results 6
3.5 Volume and Volatility Correlation 7

4 Conclusion **9**
4.1 Key Insights 9

1 Introduction

In the modern Information Technology (IT) landscape, data is the most valuable asset. The ability to understand data, identify patterns, and make predictive analyses is the primary aim of **Statistics for IT**. Financial markets represent one of the richest data environments globally, generating millions of datasets daily regarding stock prices, trading volumes, and market sentiment.

This report focuses on the statistical analysis of an expanded subset of major global technology and industrial leaders. This analysis the diverse portfolio of 10 companies: **NVIDIA (NVDA), Alphabet (GOOGL), Apple (AAPL), Microsoft (MSFT), Amazon (AMZN), TSMC (TSM), Broadcom (AVGO), Saudi Aramco (2222.SR), Meta Platforms (META), and Tesla (TSLA)**.

The technology sector is characterized by high growth potential but also significant volatility. Understanding the statistical properties of these stock movements beyond simple averages is essential for algorithmic trading, risk management systems, and financial forecasting, all of which rely heavily on robust IT infrastructure.

1.1 Objective and Statistical Questions

The primary objective of this report is to analyze historical stock data to understand the distribution of daily returns and the quantitative relationship between trading volume and price volatility. Specifically, we aim to answer the following statistical questions:

1. **Descriptive Analysis:** What are the central tendencies (mean) and dispersion (standard deviation) of the daily returns? Furthermore, do the returns exhibit Skewness or Kurtosis indicative of **fat tails**?

2. **Distribution Analysis:** Do the daily returns of these technology stocks follow a Normal Distribution? This is a critical assumption in many legacy financial IT models.

3. **Correlation Analysis:** Is there a statistically significant relationship between the volume of shares traded and the magnitude of price volatility?

1.2 Dataset Overview

The analysis utilizes 5 years of historical data accessed via the `yfinance` API. Key data points include:

- **Date:** The specific trading day.

- **Close:** The closing price of the stock.

- **Volume:** The number of shares traded.

Data was cleaned and preprocessed to calculate daily returns and absolute returns for all 10 tickers.

2 Methodology

To ensure the statistical validity of our findings, a structured methodological approach was executed using Python.

2.1 Tools and Environment

The analysis was conducted using Python with the following libraries:

- **yfinance:** For real-time data acquisition.
- **Pandas:** For data manipulation and time-series management.
- **Matplotlib & Seaborn:** For generating data visualizations.
- **Scipy.stats:** For conducting hypothesis testing (Shapiro-Wilk) and calculating Skewness/Kurtosis.

2.2 Data Availability and Reproducibility

The complete source code, data extraction logic, and analysis workflow are hosted on Google Colab to ensure reproducibility. The notebook can be accessed at the following URL:

`https://colab.research.google.com/drive/1SpvK_HD41PsJs2ZU0wvb77B4sw0zueR7?usp=sharing`

2.3 Data Preprocessing

Raw stock price data is not suitable for direct statistical comparison due to varying price levels. We calculated the **Daily Return** (R_t), defined as:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{1}$$

Where P_t is the closing price. Additionally, to measure volatility magnitude against volume, we calculated the **Absolute Return** ($|R_t|$).

2.4 Statistical Techniques

2.4.1 Descriptive Statistics & Higher Moments

Beyond Mean (μ) and Standard Deviation (σ), we calculated:

- **Skewness:** A measure of the asymmetry of the probability distribution.

- **Kurtosis:** A measure of the **tailedness**. High kurtosis indicates frequent extreme deviations.

2.4.2 Hypothesis Testing: The Shapiro-Wilk Test

To formally test for normality:

- **Null Hypothesis (H_0):** The data is distributed normally.
- **Alternative Hypothesis (H_1):** The data is not distributed normally.
- **Alpha Level (α):** 0.05.

2.4.3 Correlation Analysis

To quantify the relationship between Volume and Volatility, we utilized the **Pearson Correlation Coefficient (r)**:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{2}$$

3 Analysis and Results

3.1 Descriptive Statistics

Table 1 summarizes the statistical properties of the daily returns of 10 companies.

Table 1: Descriptive Statistics of Daily Returns (5-Year Period)

Ticker	Mean	Std Dev	Skewness	Kurtosis
2222.SR	-0.0001	0.0097	0.37	5.19
AAPL	0.0007	0.0175	0.47	6.64
AMZN	0.0006	0.0221	0.14	5.15
AVGO	0.0020	0.0267	0.88	10.37
GOOGL	0.0013	0.0196	0.09	3.16
META	0.0011	0.0273	-0.19	21.04
MSFT	0.0008	0.0162	0.21	3.25
NVDA	0.0026	0.0328	0.51	4.53
TSLA	0.0011	0.0381	0.37	3.18
TSM	0.0010	0.0235	0.29	2.78

Interpretation: The data reveals striking differences in risk profiles. **Tesla (TSLA)** exhibits the highest standard deviation (0.0381), indicating it is the most volatile asset in the group. Conversely, **Saudi Aramco (2222.SR)** is the most stable with a deviation of only 0.0097.

Crucially, **Meta Platforms (META)** displays an extreme Kurtosis of **21.04**, significantly higher than the normal distribution value of 3.0. This indicates an extremely "fat-tailed" distribution, implying that Meta is subject to extreme price shocks far more often than statistical models would predict.

3.2 Visualization of Price Trends

Figure 3 illustrates the closing prices over the analyzed period.

Figure 1: Line chart showing stock trends for 10 tickers over 5 years.

3.3 Distribution Analysis

To visualize the nonnormality suggested by the Kurtosis values, we examine the distribution of returns.

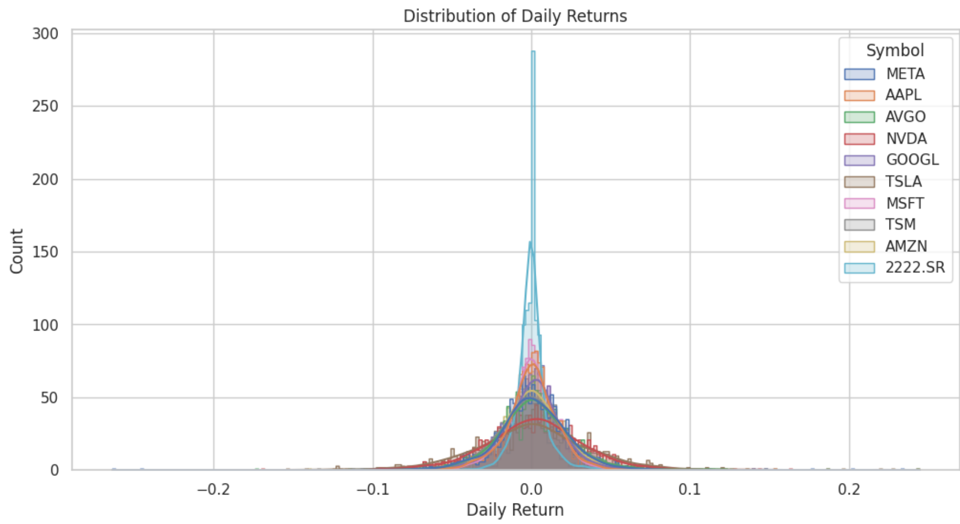

Figure 2: Histogram of Daily Returns with KDE vs. Normal Curve

3.4 Hypothesis Testing Results

The Shapiro-Wilk test results (Table 2) confirm our visual and descriptive analysis.

Table 2: Shapiro-Wilk Normality Test Results

Ticker	P-Value	Result
NVDA	1.32×10^{-17}	Reject H_0
GOOGL	5.52×10^{-17}	Reject H_0
AAPL	1.01×10^{-20}	Reject H_0
MSFT	2.98×10^{-16}	Reject H_0
AMZN	5.32×10^{-21}	Reject H_0
TSM	1.26×10^{-14}	Reject H_0
AVGO	2.27×10^{-26}	Reject H_0
2222.SR	5.24×10^{-26}	Reject H_0
META	1.33×10^{-33}	Reject H_0
TSLA	9.20×10^{-16}	Reject H_0

Conclusion: With p-values virtually at zero for every single asset, we emphatically **reject the Null Hypothesis**. None of the stock returns follow a normal distribution.

3.5 Volume and Volatility Correlation

We hypothesized that higher trading volume leads to higher volatility (absolute price change).

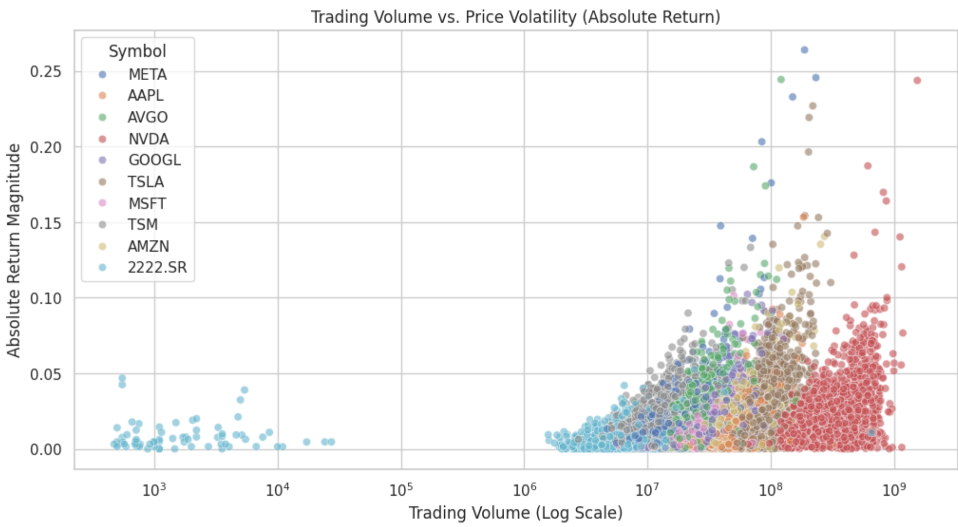

Figure 3: Scatter plot: Volume (log scale) vs Absolute Return

The Pearson correlation analysis yields significant results for all assets:

- **Highest Correlation:** META ($r = 0.6561$), indicating a very strong link between volume and price swings.
- **Lowest Correlation:** Saudi Aramco (2222.SR) ($r = 0.1313$), suggesting its price movement is less dependent on volume surges compared to tech stocks.

- **Significance:** All p-values are < 0.05, confirming statistically significant relationships across the board.

4 Conclusion

This expanded analysis of 10 global leaders reinforces the need for advanced statistical modeling in IT financial systems.

4.1 Key Insights

1. **Extreme Events (Fat Tails):** The extreme Kurtosis observed in META (21.04) and AVGO (10.37) proves that standard deviation is an insufficient measure of risk. IT systems must account for "black swan" events that occur far more frequently than Gaussian models predict.

2. **Universal Non-Normality:** Every single asset tested failed the Shapiro-Wilk normality test. Algorithms assuming normality are fundamentally flawed for this sector.

3. **Volume as a Predictor:** For US tech stocks, volume is a strong predictor of volatility ($r > 0.45$). However, this relationship weakens for non-tech/industrial giants like Saudi Aramco ($r \approx 0.13$), suggesting that algorithmic triggers must be tuned specifically to the asset class.

References

[1] Group C, *Statistical Analysis Code Repository*, Google Colab Notebook. Available at: https://colab.research.google.com/drive/1SpvK_HD4lPsJs2ZU0wvb77B4swOzueR7?usp=sharing

[2] Dhrubang Talukdar, *Fortune 500 Companies Stock Data*, Kaggle Dataset.

[3] Wes McKinney, *Python for Data Analysis*, O'Reilly Media, 2017.

[4] Bruce, P., & Bruce, A., *Practical Statistics for Data Scientists*, O'Reilly Media, 2018.

Appendix: Python Implementation Code

The following code was used to access data via `yfinance`, generate statistics, and produce visualizations.

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# 1. Data Acquisition (yfinance)
print("Initializing data download...")
try:
    import yfinance as yf
except ImportError:
    print("Error: yfinance not found.")
    exit()

# Define tickers
tech_tickers = [
    'NVDA', 'GOOGL', 'AAPL', 'MSFT', 'AMZN',
    'TSM', 'AVGO', '2222.SR', 'META', 'TSLA'
]

data_list = []
for ticker in tech_tickers:
    # Download 5 years of data
    ticker_data = yf.download(ticker, period="5y", progress=False, auto_adjust=False)
    ticker_data = ticker_data.reset_index()
    ticker_data['Symbol'] = ticker

    # Flatten MultiIndex columns if necessary
    if isinstance(ticker_data.columns, pd.MultiIndex):
        ticker_data.columns = ticker_data.columns.get_level_values(0)
        ticker_data['Symbol'] = ticker

    data_list.append(ticker_data)

df = pd.concat(data_list, ignore_index=True)

# 2. Data Cleaning & Preprocessing
df.columns = df.columns.str.strip()
df['Date'] = pd.to_datetime(df['Date'])
df.sort_values('Date', inplace=True)

```

```python
df_tech = df[df['Symbol'].isin(tech_tickers)].copy()
df_tech['Daily_Return'] = df_tech.groupby('Symbol')['Close'].pct_change
    ()
df_tech['Abs_Return'] = df_tech['Daily_Return'].abs()
df_tech.dropna(inplace=True)

# 3. Statistical Analysis
# A. Descriptive Statistics
desc_stats = df_tech.groupby('Symbol')['Daily_Return'].agg(
    ['mean', 'std', lambda x: stats.skew(x), lambda x: stats.kurtosis(x)
    ]
)
desc_stats.columns = ['Mean', 'Std Dev', 'Skewness', 'Kurtosis']
print("Descriptive Statistics:\n", desc_stats)

# B. Normality Test (Shapiro-Wilk)
print("\nShapiro-Wilk Normality Test Results:")
for ticker in tech_tickers:
    data = df_tech[df_tech['Symbol'] == ticker]['Daily_Return']
    stat, p = stats.shapiro(data)
    print(f"{ticker}: P-value={p:.4g}")

# C. Correlation Analysis
print("\nCorrelation Analysis (Volume vs Absolute Return):")
for ticker in tech_tickers:
    subset = df_tech[df_tech['Symbol'] == ticker]
    corr, p_val = stats.pearsonr(subset['Volume'], subset['Abs_Return'])
    print(f"{ticker}: Pearson r={corr:.4f}, p-value={p_val:.4g}")

# 4. Visualization
sns.set_theme(style="whitegrid")

# Stock Price Trend
plt.figure(figsize=(12, 6))
sns.lineplot(data=df_tech, x='Date', y='Close', hue='Symbol')
plt.title('Stock Closing Prices Trend (Last 5 Years)')
plt.savefig('trend.png')

# Distribution
plt.figure(figsize=(12, 6))
sns.histplot(data=df_tech, x='Daily_Return', hue='Symbol', kde=True,
    element="step")
plt.title('Distribution of Daily Returns')
plt.savefig('hist.png')

# Volume vs Volatility
plt.figure(figsize=(12, 6))
```

```python
sns.scatterplot(data=df_tech, x='Volume', y='Abs_Return', hue='Symbol',
    alpha=0.6)
plt.xscale('log')
plt.title('Trading Volume vs. Price Volatility')
plt.savefig('scatter.png')
```

Listing 1: Updated Python Analysis Script