ĐẠI HỌC FPT
CHUYÊN NGÀNH TRÍ TUỆ NHÂN TẠO

FPT UNIVERSITY

BÁO CÁO LAB 1 & 2

Môn học: **AI, DS with Python & SQL(ADY201m)**

Tên đề tài: **Project Kickoff & Initial Implementation AND In-depth analysis and visualization of results**

Giảng viên hướng dẫn: **Lê Võ Minh Thư**
Lớp: **AI1906**
Sinh viên thực hiện:
Lâm Nguyễn Minh Thanh – SE181933
Nguyễn Cao Trị – SE180683
Đinh Đại Lộc – SE190189
Trần Phi Học – SE190186

TP. Hồ Chí Minh, ngày 30 tháng 06 năm 2025

MỤC LỤC

Lab 1: Project Kickoff & Initial Implementation

1. Business Understanding & Analytic Approach

Mục tiêu dự án:

Phân tích dữ liệu thực tế để tìm ra các yếu tố ảnh hưởng và dự báo xu hướng/biến động. Ví dụ:

- Dự đoán giá cổ phiếu sau ngày earnings.
- Phân tích mối liên hệ giữa thời tiết và tiêu thụ điện.
- Theo dõi chỉ số sức khỏe người bệnh theo thời gian.

Người dùng cuối (Stakeholders):

- Nhà đầu tư, nhà phân tích tài chính (nếu là stock).
- Nhà nghiên cứu môi trường hoặc y tế.
- Doanh nghiệp nội bộ cần dashboard phân tích.

Câu hỏi nghiên cứu:

- Sau báo cáo tài chính, giá cổ phiếu thay đổi như thế nào?
- Biến động sentiment (tin tức/tâm lý) có ảnh hưởng đến giá?
- Có thể dự báo giá/điểm chỉ số dựa trên dữ liệu hiện tại?

KPIs đánh giá:

- RMSE, MAE (cho dự báo).
- Correlation coefficient (cho phân tích mối quan hệ).
- Directional accuracy (% dự đoán đúng xu hướng).

2. Data Collection, Understanding & Preparation

- Sử dụng thư viện **yfinance** để tải dữ liệu giá cổ phiếu của AAPL (Apple Inc.) trong khoảng thời gian từ 1/1/2023 đến 31/5/2025**.**

```
price_data = yf.download("AAPL", start='2023-01-01', end='2025-05-31')
```

Kết quả: thu được **604** dòng dữ liệu

Nhận xét: Đây là dữ liệu cổ phiếu của Apple (AAPL) từ ngày 3 tháng 1, 2023 đến ngày 4 tháng 1, 2025. Các cột trong dữ liệu bao gồm:

Date: Ngày giao dịch.
Close: Giá đóng cửa của cổ phiếu.
High: Giá cao nhất trong ngày.
Low: Giá thấp nhất trong ngày.
Open: Giá mở cửa.
Volume: Khối lượng giao dịch.

- Sử dụng thư viện **yfinance** tiếp tục tải về một **DataFrame chứa các ngày công bố báo cáo tài chính (earnings).**

```
ticker = yf.Ticker(ticker_symbol)
earnings_df = ticker.earnings_dates
```

Kết quả: thu được **8** dòng dữ liệu

AAPL_earnings.csv

```
Earnings Date,EPS Estimate,Reported EPS,Surprise(%)
2023-08-03 16:30:00-04:00,1.19,1.26,5.49
2023-11-02 16:30:00-04:00,1.39,1.46,4.92
2024-02-01 16:00:00-05:00,2.1,2.18,3.9
2024-05-02 16:31:00-04:00,1.5,1.53,1.97
2024-08-01 16:30:00-04:00,1.35,1.4,3.99
2024-10-31 16:31:00-04:00,1.6,1.64,2.35
2025-01-30 16:31:00-05:00,2.35,2.4,2.15
2025-05-01 16:30:00-04:00,1.63,1.65,1.41
```

- **Tạo dữ liệu giả định sentiment (cảm xúc thị trường) quanh các ngày công bố báo cáo tài chính (earnings) của một mã cổ phiếu.**

```python
def generate_sentiment(ticker_symbol="AAPL", start_date="2023-01-01", end_date="2025-05-31", output_file="AAPL_sentiment.csv"):
    ticker = yf.Ticker(ticker_symbol)
    earnings_df = ticker.earnings_dates.reset_index()

    # Chuyển cột Earnings Date thành datetime và bỏ timezone
    earnings_df["Earnings Date"] = pd.to_datetime(earnings_df["Earnings Date"]).dt.tz_localize(None)
    start_date_dt = pd.to_datetime(start_date)
    end_date_dt = pd.to_datetime(end_date)
    earnings_df = earnings_df[(earnings_df["Earnings Date"] >= start_date_dt) & (earnings_df["Earnings Date"] <= end_date_dt)]
    earnings_df['compound'] = np.random.uniform(-1, 1, size=len(earnings_df))
    sentiment_df = earnings_df[['Earnings Date', 'compound']]
    sentiment_df.to_csv(output_file, index=False)
    print(f"Đã tạo file sentiment giả định: {output_file}")
```

Kết quả: thu được **8** dòng dữ liệu

AAPL_sentiment.csv

```
Earnings Date,compound
2025-05-01 16:30:00,0.4357806479229809
2025-01-30 16:31:00,-0.06677913582156259
2024-10-31 16:31:00,0.025397269855249816
2024-08-01 16:30:00,0.3448244703090986
2024-05-02 16:31:00,0.12698641633359165
2024-02-01 16:00:00,-0.24383089748633568
2023-11-02 16:30:00,0.15425920479775956
2023-08-03 16:30:00,0.2908905677745417
```

- **Tiền xử lý dữ liệu:**

```
Số lượng missing values sau khi điền:
Unnamed: 1    0
Unnamed: 2    0
Unnamed: 3    0
Unnamed: 4    0
Unnamed: 5    0
dtype: int64

Thông tin tổng quát về dữ liệu:
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 604 entries, 2023-01-03 to 2025-05-30
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 1  604 non-null    float64
 1   Unnamed: 2  604 non-null    float64
 2   Unnamed: 3  604 non-null    float64
 3   Unnamed: 4  604 non-null    float64
 4   Unnamed: 5  604 non-null    int64
dtypes: float64(4), int64(1)
memory usage: 28.3 KB
```

- Không có missing values trong 5 cột:

```
Số lượng missing values trong mỗi cột:
Unnamed: 1    0
Unnamed: 2    0
Unnamed: 3    0
Unnamed: 4    0
Unnamed: 5    0
```

- Tổng số bản ghi: 604 ngày (giao dịch).

Khoảng thời gian: từ `2023-01-03` đến `2025-05-30`.
Tổng số biến: 5.

```
DatetimeIndex: 604 entries, 2023-01-03 to 2025-05-30
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 1  604 non-null    float64
 1   Unnamed: 2  604 non-null    float64
 2   Unnamed: 3  604 non-null    float64
 3   Unnamed: 4  604 non-null    float64
 4   Unnamed: 5  604 non-null    int64
dtypes: float64(4), int64(1)
```

Mô tả cấu trúc dữ liệu (df.describe()):

Thống kê	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
count	604	604	604	604	604
mean	193.83	195.56	191.82	193.56	58,245,600
std	29.04	29.31	28.69	29.07	25,306,660
min	123.42	126.14	122.58	124.40	23,234,700
25%	172.47	174.01	171.12	172.30	44,904,420
50%	189.00	190.31	187.67	188.77	52,391,950
75%	221.23	223.53	218.93	220.75	64,857,650
max	258.40	259.47	257.01	257.57	318,679,900

Nhận xét sơ bộ:

- Các cột **`Unnamed: 1`** → **`Unnamed: 4`** có giá trị xung quanh 190–195, rất phù hợp với các loại giá chứng khoán (**`Open`**, **`High`**, **`Low`**, **`Close`**).
- Cột **`Unnamed: 5`** có giá trị rất lớn, biến động từ ~23 triệu đến hơn 318 triệu, rất có khả năng là Volume.
- Độ lệch chuẩn (std) khá lớn → thị trường có biến động tương đối mạnh.

- **Dữ liệu trước Scaling:**

```
Dữ liệu gốc:
            Unnamed: 1  Unnamed: 2  Unnamed: 3  Unnamed: 4  Unnamed: 5
Date
2023-01-03  123.470619  129.226060  122.582127  128.613993   112117500
2023-01-04  124.744133  127.014724  123.480503  125.267354    89113600
2023-01-05  123.421249  126.136083  123.164580  125.504267    80962700
2023-01-06  127.962433  128.623863  123.292924  124.398604    87754700
2023-01-09  128.485657  131.703978  128.228987  128.801572    70790800
```

- **Dữ liệu sau Scaling:**

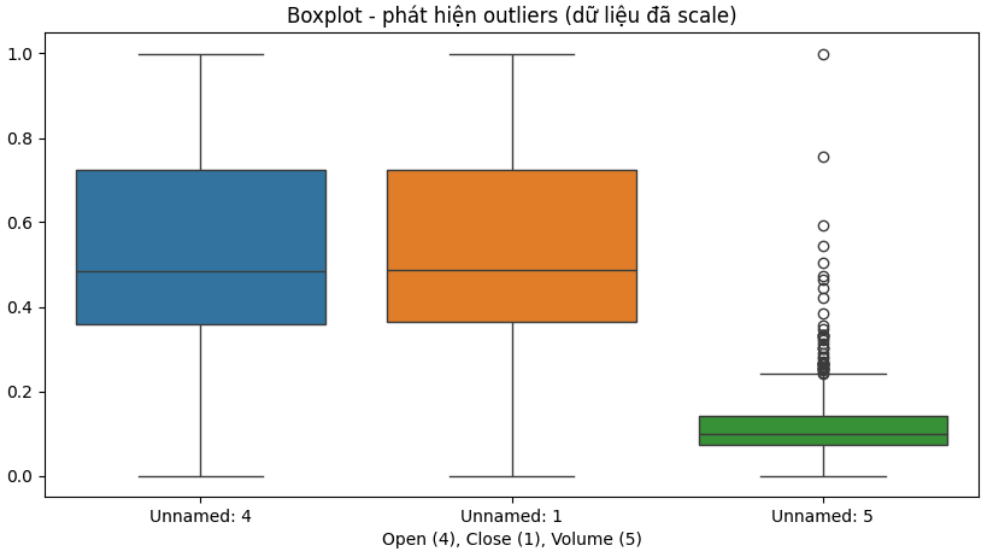
$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Công thức:

Dữ liệu sau khi scaling:

Date	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
2023-01-03	0.000366	0.023174	0.000000	0.031654	0.300844
2023-01-04	0.009801	0.006590	0.006683	0.006524	0.222982
2023-01-05	0.000000	0.000000	0.004333	0.008303	0.195393
2023-01-06	0.033645	0.018658	0.005288	0.000000	0.218382
2023-01-09	0.037521	0.041758	0.042007	0.033063	0.160964

- **Boxplot - phát hiện outliers (dữ liệu đã scale):**



Nhận xét về biểu đồ Boxplot:

Dữ liệu Open (Cột 4):

1. Phân bố dữ liệu Open khá đồng đều, không có sự chênh lệch quá lớn giữa các phần tử trong khoảng IQR (Interquartile Range). Điều này cho thấy dữ liệu có mức độ tập trung ổn định.

2. Các điểm ngoài dải (outliers) có thể là những giá trị ngoại lai cần phải kiểm tra kỹ hơn.

Dữ liệu Close (Cột 1):

3. Tương tự như Open, phân bố của dữ liệu Close có vẻ khá tập trung, với một vài điểm ngoài dải. Dữ liệu này cũng không có sự biến động quá lớn, và các outliers có thể là những sự kiện đặc biệt trong dữ liệu.

4. Sự phân bố này có thể cho thấy thị trường ổn định trong một khoảng thời gian nhất định.

Dữ liệu Volume (Cột 5):

5. Dữ liệu Volume có sự phân tán rất rộng với nhiều outliers. Điều này cho thấy có một số giá trị cực đoan trong dữ liệu, có thể do các sự kiện đặc biệt, hoặc có thể là các lỗi trong dữ liệu.

6. Các outliers này có thể cần phải được kiểm tra để xác định nguyên nhân (ví dụ như các giao dịch với khối lượng rất lớn hoặc các sự kiện gây biến động mạnh).

3. Data Analysis with SQL

Lưu ý: trước khi qua phần 3 cần sửa dòng 3 **[Date,,,,,]** trong file **AAPL_price.csv** thành **[Date,Close,High,Low,Open,Volume]** để đồng bộ dữ liệu Lỗi này đã được fix ở file FULL_LAB1.py

- **Báo cáo phân tích ảnh hưởng của EPS đến biến động giá cổ phiếu AAPL sau ngày công bố earnings**
  - **Tất cả EPS đều "Beat"**, thể hiện công ty liên tục vượt kỳ vọng thị trường.
  - **Tuy nhiên, giá cổ phiếu sau earnings không tăng tương ứng**
    - **5/8 kỳ**: giá **giảm** sau khi công bố earnings dù kết quả tốt.
    - **Chỉ 3/8 kỳ**: giá **tăng**, trong đó chỉ có **1 kỳ (2024-05-02)** là tăng rõ rệt.
  - Điều này phản ánh thực tế rằng **thị trường không chỉ phản ứng với kết quả EPS**, mà còn bị ảnh hưởng bởi:
    - Kỳ vọng thị trường đã phản ánh vào giá trước đó.
    - Các thông tin khác trong earnings call (doanh thu, hướng dẫn tương lai).
    - Bối cảnh vĩ mô, lãi suất, tin tức công nghệ...

EPS Date	EPS Beat	% Price Change	Ghi chú chính
2023-08-03	✅	-5.50%	EPS tăng nhưng giá **giảm mạnh**
2023-11-02	✅	+1.54%	Giá tăng nhẹ
2024-02-01	✅	+0.79%	Tăng nhẹ
2024-05-02	✅	+8.32%	**Tăng mạnh** sau báo cáo tốt
2024-08-01	✅	-1.00%	Giá giảm dù EPS vượt dự kiến
2024-10-31	✅	-3.12%	Tương tự, giá giảm
2025-01-30	✅	-1.40%	Giá giảm nhẹ
2025-05-01	✅	-3.36%	Giá tiếp tục giảm sau EPS tốt

	earnings_date	eps_est	eps_actual	close_before	close_after	price_change_pct	eps_result
0	2023-08-03	1.19	1.26	190.670959	180.185944	-5.50	Beat
1	2023-11-02	1.39	1.46	172.478027	175.135040	1.54	Beat
2	2024-02-01	2.10	2.18	183.059433	184.498901	0.79	Beat
3	2024-05-02	1.50	1.53	168.283676	182.279144	8.32	Beat
4	2024-08-01	1.35	1.40	221.046234	218.836578	-1.00	Beat
5	2024-10-31	1.60	1.64	229.294006	222.129196	-3.12	Beat
6	2025-01-30	2.35	2.40	238.783997	235.432083	-1.40	Beat
7	2025-05-01	1.63	1.65	212.221710	205.081070	-3.36	Beat

Phân tích tương quan SENTIMENT và biến động giá

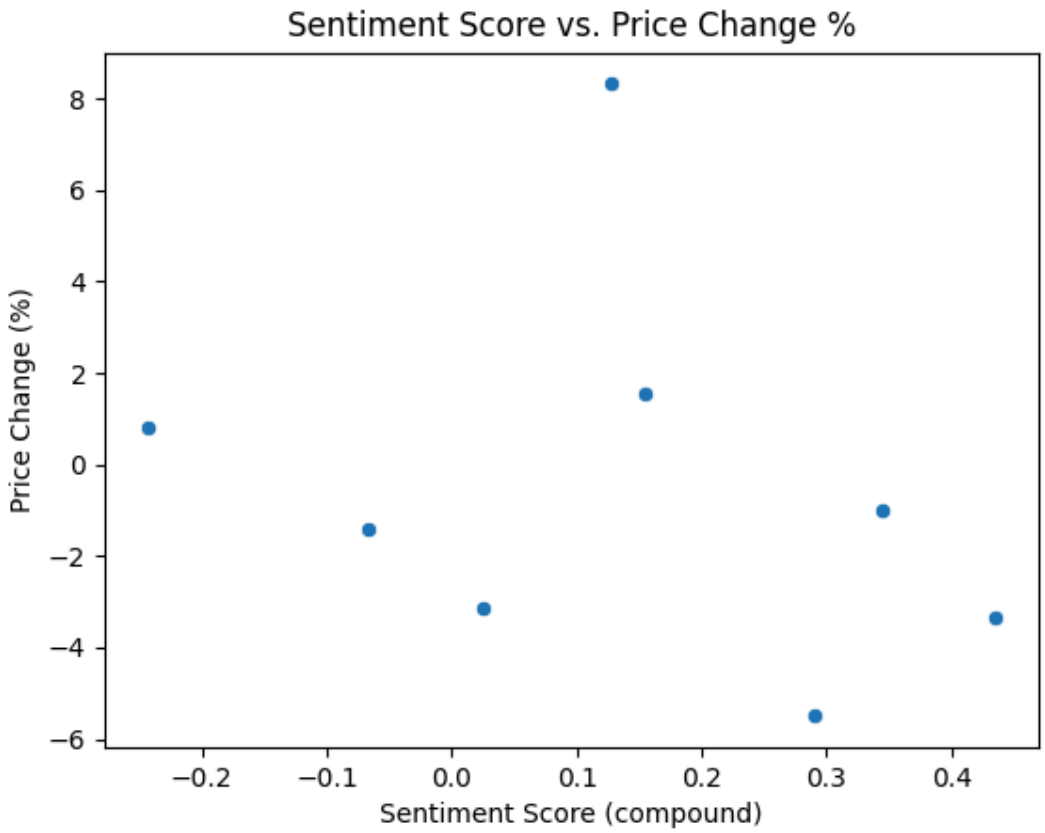
○ Trong 5 kỳ earnings gần nhất, không có bằng chứng rõ ràng cho thấy sentiment tích cực sẽ dẫn đến giá cổ phiếu tăng ngay sau earnings.

○ Để hiểu rõ hơn, cần phân tích thêm nhiều kỳ earnings hơn, hoặc kết hợp sentiment với các yếu tố như tăng trưởng doanh thu, phản ứng thị trường ngành, hoặc dùng mô hình ML để kiểm định đa biến.

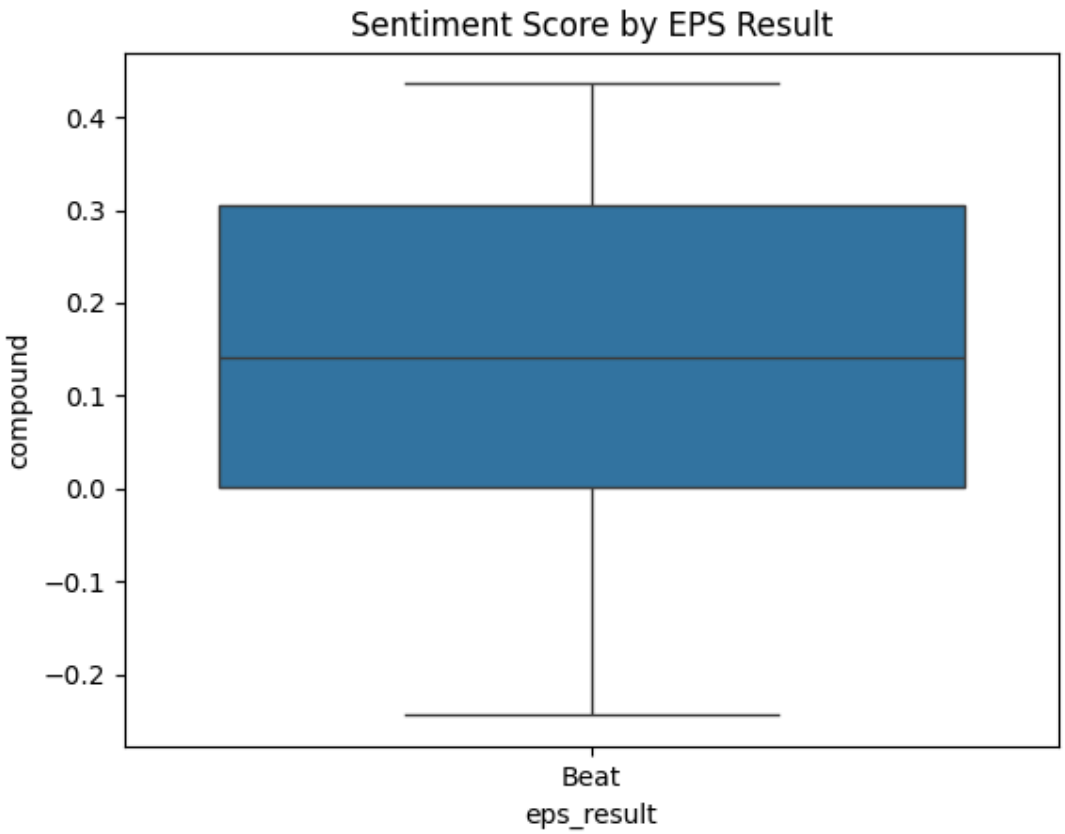
```
Index(['earnings_date', 'eps_est', 'eps_actual', 'close_before', 'close_after',
       'compound', 'price_change_pct', 'eps_result'],
      dtype='object')
```

	earnings_date	eps_est	eps_actual	close_before	close_after	compound	price_change_pct	eps_result
0	2023-08-03	1.19	1.26	190.670959	180.185944	0.290891	-5.50	Beat
1	2023-11-02	1.39	1.46	172.478027	175.135040	0.154259	1.54	Beat
2	2024-02-01	2.10	2.18	183.059433	184.498901	-0.243831	0.79	Beat
3	2024-05-02	1.50	1.53	168.283676	182.279144	0.126986	8.32	Beat
4	2024-08-01	1.35	1.40	221.046234	218.836578	0.344824	-1.00	Beat

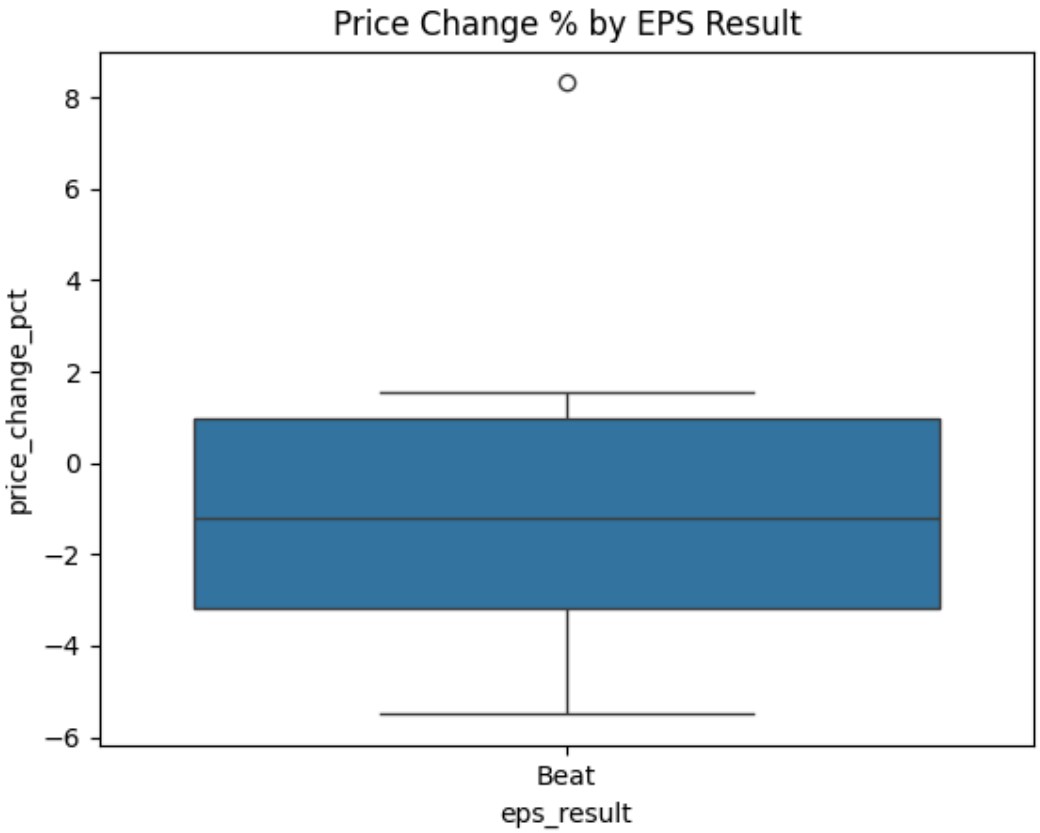
Trực quan hóa



Nhận xét: Dữ liệu cho thấy không có mối quan hệ rõ ràng giữa hai yếu tố này. Điểm sentiment không có tác động đồng nhất lên sự thay đổi giá cổ phiếu. Một số điểm dữ liệu có sự thay đổi giá mạnh (từ -6% đến +8%), nhưng không có xu hướng rõ rệt trong mối quan hệ giữa sentiment và price change.



Nhận xét: Phần lớn các điểm sentiment đều có giá trị từ 0 đến 0.3, cho thấy rằng khi kết quả EPS vượt kỳ vọng (Beat), cảm xúc của các nhà đầu tư là khá tích cực, với điểm sentiment chủ yếu dao động trong mức trung bình thấp đến cao. Điều này có thể chỉ ra rằng khi doanh nghiệp vượt kỳ vọng về lợi nhuận, cảm xúc của thị trường là tích cực.



Nhận xét: Tỷ lệ thay đổi giá chủ yếu dao động trong khoảng từ -2% đến +2%, nhưng có một vài ngoại lệ với các thay đổi giá vượt quá 6%. Điều này chỉ ra rằng kết quả EPS vượt kỳ vọng có thể không luôn dẫn đến sự thay đổi giá mạnh mẽ. Tuy nhiên, vẫn có những lần biến động giá đáng kể, có thể là do các yếu tố khác ảnh hưởng đến giá cổ phiếu, chẳng hạn như điều kiện thị trường hoặc sự điều chỉnh của nhà đầu tư.

4. Data Analysis with Python

Sử dụng thư viện `statsmodels` để thống kê cơ bản

Nhận xét:

1. **Mối quan hệ giữa cảm xúc và thay đổi giá**:
    - **Biểu đồ ma trận tương quan** cho thấy một mối quan hệ yếu và âm giữa cảm xúc (sentiment) và tỷ lệ thay đổi giá, với hệ số tương quan là -0.27. Điều này có nghĩa là cảm xúc và thay đổi giá có xu hướng di chuyển ngược chiều nhau, nhưng mối quan hệ này rất yếu. Việc cảm xúc tăng lên không hẳn luôn đồng nghĩa với giá sẽ giảm và ngược lại.
2. **Sự biến động của giá so với sự ổn định của cảm xúc**:

- **Biểu đồ tỷ lệ thay đổi giá và cảm xúc trung bình theo quý** cho thấy rằng sự thay đổi giá có sự biến động mạnh mẽ hơn so với cảm xúc, vốn ổn định và ít thay đổi. Điều này có thể chỉ ra rằng có nhiều yếu tố khác (như thông tin thị trường, yếu tố vĩ mô) ảnh hưởng đến thay đổi giá, chứ không chỉ là cảm xúc.

3. **Mô hình mùa vụ và xu hướng của cảm xúc**:

   - **Biểu đồ phân tách chuỗi thời gian cảm xúc** chỉ ra rằng cảm xúc có sự thay đổi theo mùa vụ và xu hướng. Cảm xúc có xu hướng tăng lên trong một vài quý, nhưng cũng có sự biến động theo chu kỳ, điều này phản ánh sự thay đổi trong các yếu tố tác động đến cảm xúc theo thời gian (ví dụ như tin tức, sự kiện vĩ mô).

Dự đoán:

1. **Về mối quan hệ giữa cảm xúc và thay đổi giá**:

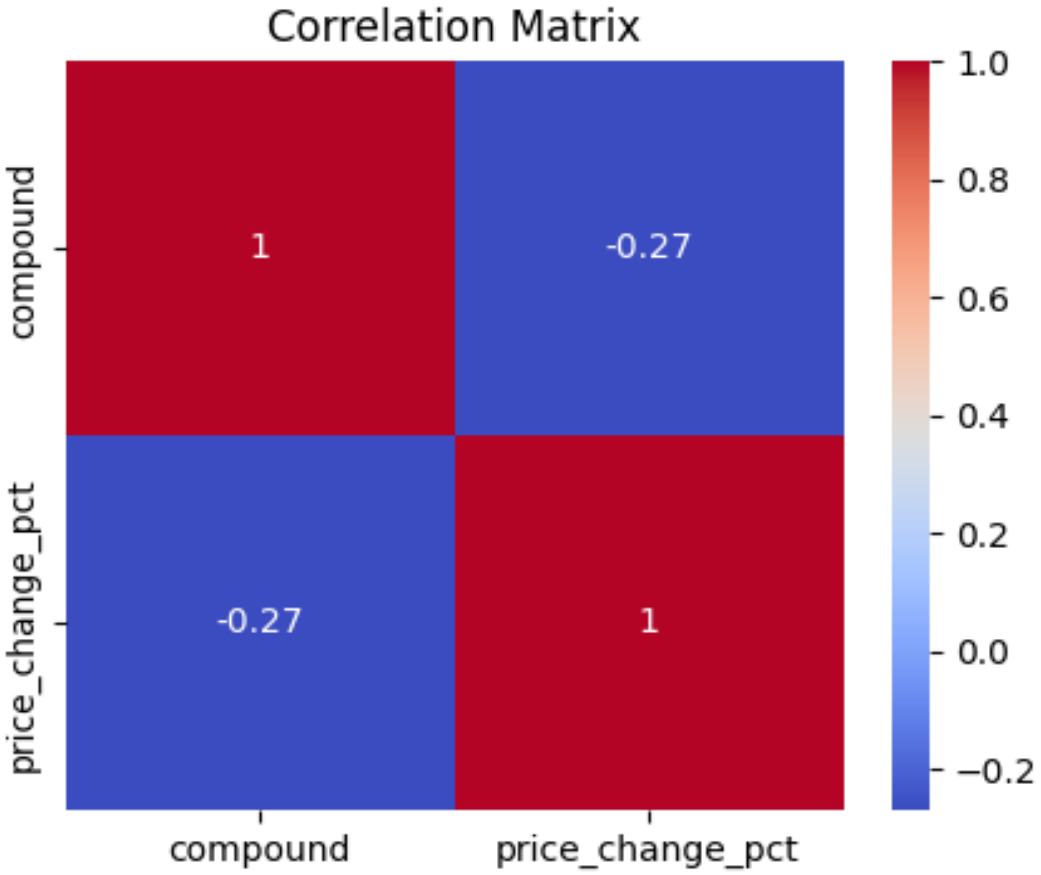
   - Mặc dù có một mối quan hệ yếu giữa cảm xúc và thay đổi giá, sự biến động của giá có thể bị ảnh hưởng bởi nhiều yếu tố khác ngoài cảm xúc. Tuy nhiên, nếu xu hướng cảm xúc tiếp tục duy trì ổn định hoặc tăng mạnh trong một số quý, có thể thấy một sự giảm nhẹ trong tỷ lệ thay đổi giá hoặc một sự điều chỉnh nhỏ. Tuy vậy, mối quan hệ này vẫn sẽ tiếp tục yếu và không thể dự đoán chính xác sự thay đổi giá chỉ dựa trên cảm xúc.

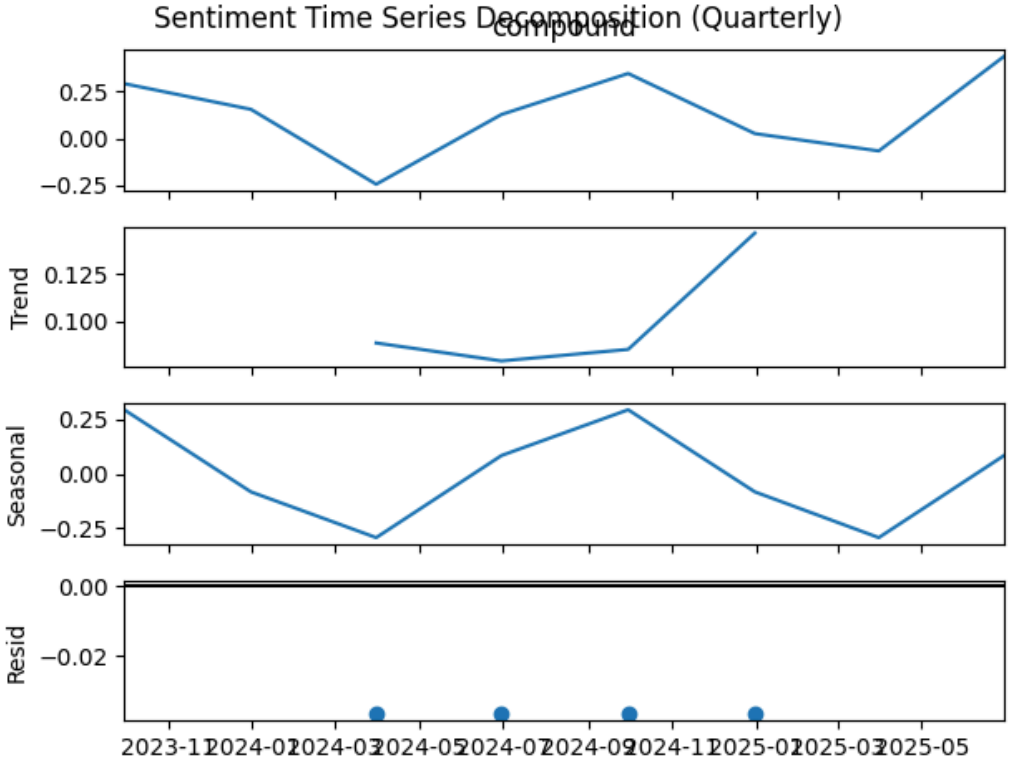
2. **Về giá trị cảm xúc trong dự báo giá**:

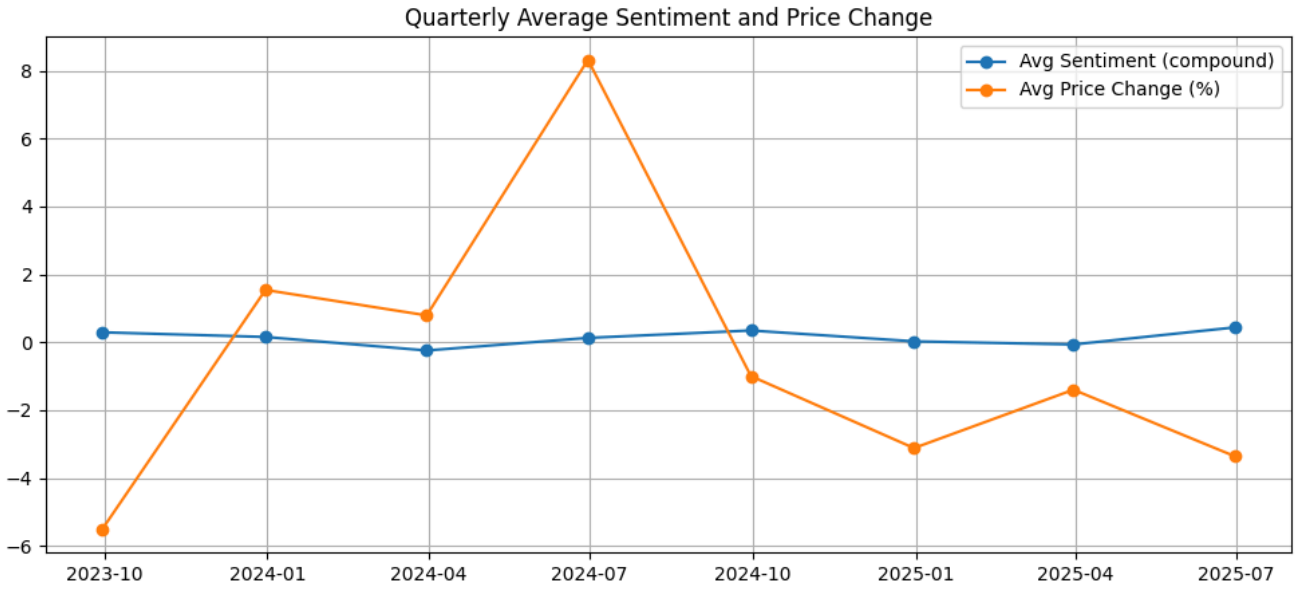
   - Vì cảm xúc chỉ có một phần ảnh hưởng nhỏ đến sự thay đổi giá, nên việc sử dụng cảm xúc như một chỉ báo độc lập để dự báo sự thay đổi giá sẽ có độ chính xác thấp. Tuy nhiên, nếu kết hợp cảm xúc với các yếu tố khác như phân tích kỹ thuật, tin tức vĩ mô, hay các chỉ số tài chính khác, dự báo giá có thể chính xác hơn.

3. **Dự báo xu hướng trong thời gian tới**:

   - Nếu xu hướng và mùa vụ của cảm xúc tiếp tục giữ vững, có thể thấy cảm xúc sẽ duy trì xu hướng tăng nhẹ trong các quý tiếp theo. Nếu sự biến động của giá tiếp tục mạnh, thì có thể có những sự điều chỉnh hoặc biến động lớn trong ngắn hạn, đặc biệt là khi có sự thay đổi bất ngờ trong các yếu tố tác động đến thị trường.

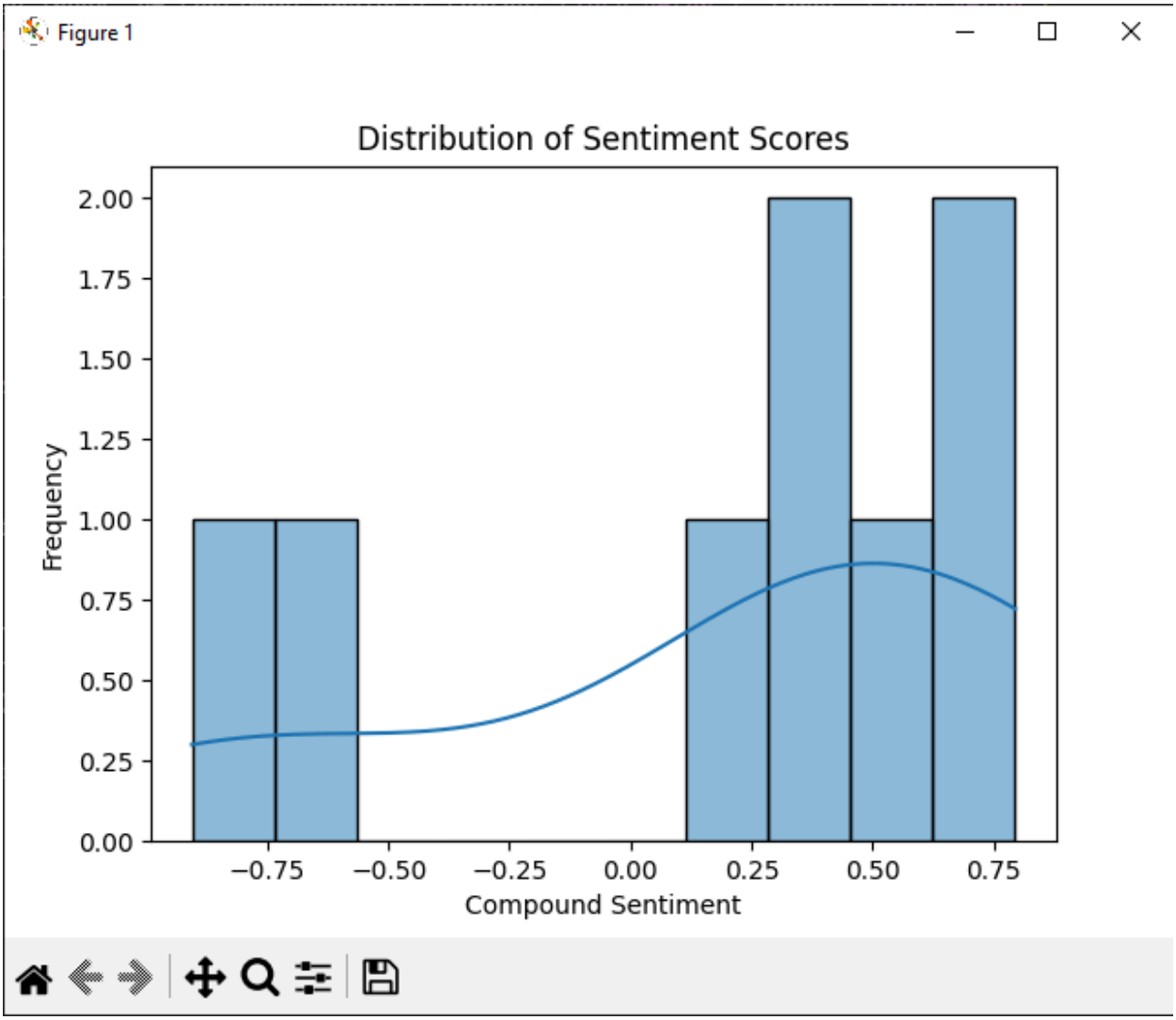

Correlation Matrix
compound
price_change_pct
1
-0.27
-0.27
1
compound
price_change_pct
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Sentiment Time Series Decomposition (Quarterly)
compound
0.25
0.00
−0.25
Trend
0.125
0.100
Seasonal
0.25
0.00
−0.25
Resid
0.00
−0.02
2023-11
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
2025-03
2025-05

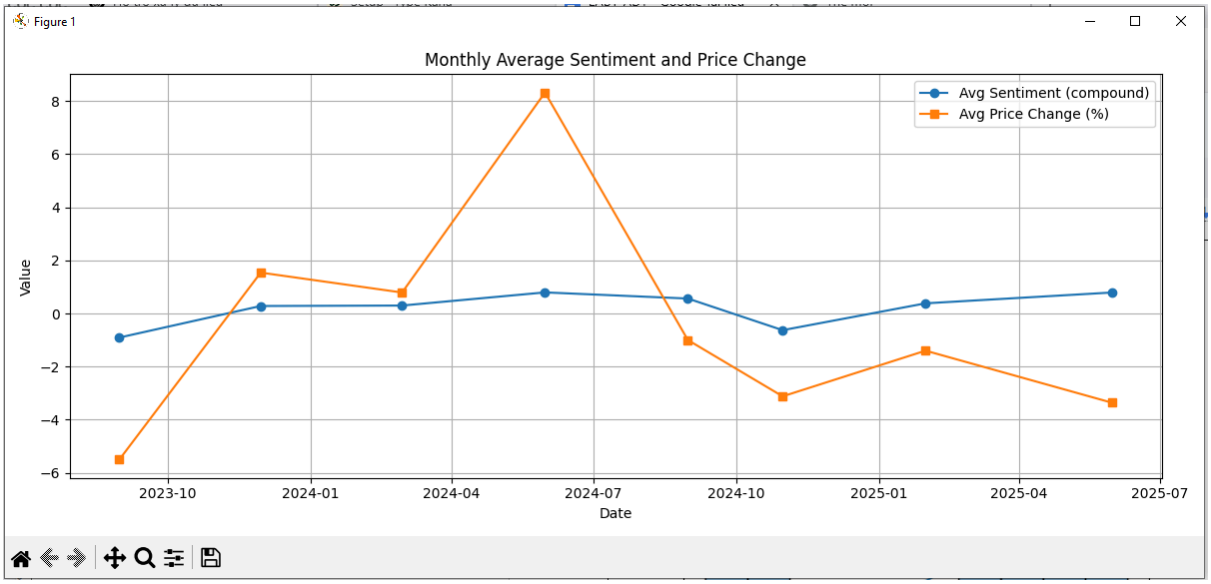

Quarterly Average Sentiment and Price Change
Avg Sentiment (compound)
Avg Price Change (%)
8
6
4
2
0
−2
−4
−6
2023-10
2024-01
2024-04
2024-07
2024-10
2025-01
2025-04
2025-07

5. Data Visualization

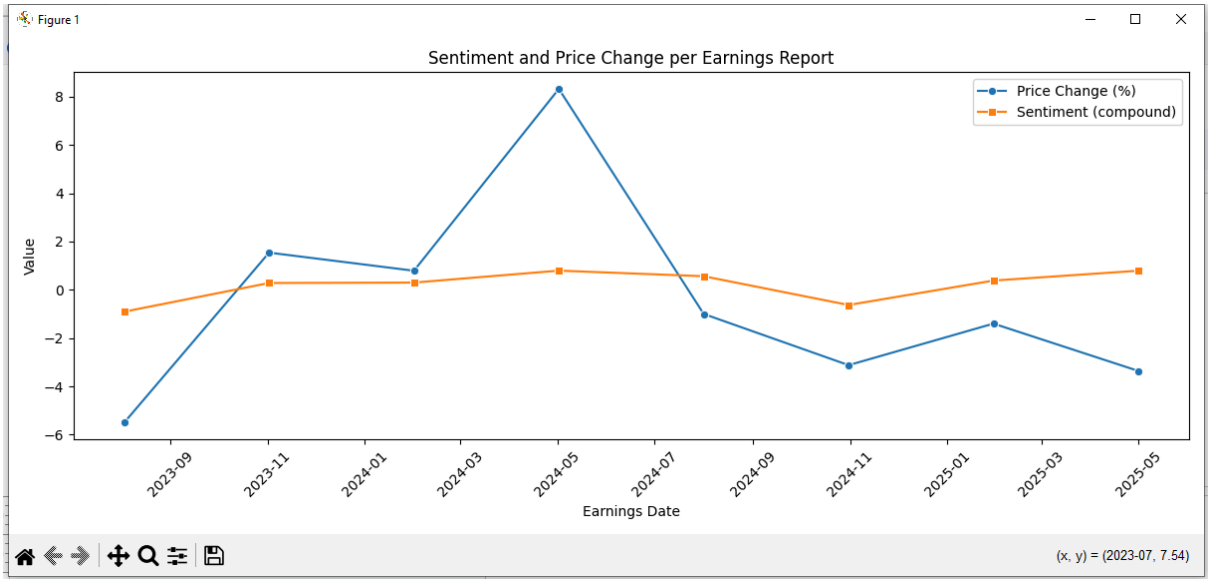
Biểu đồ 1: Histogram sentiment



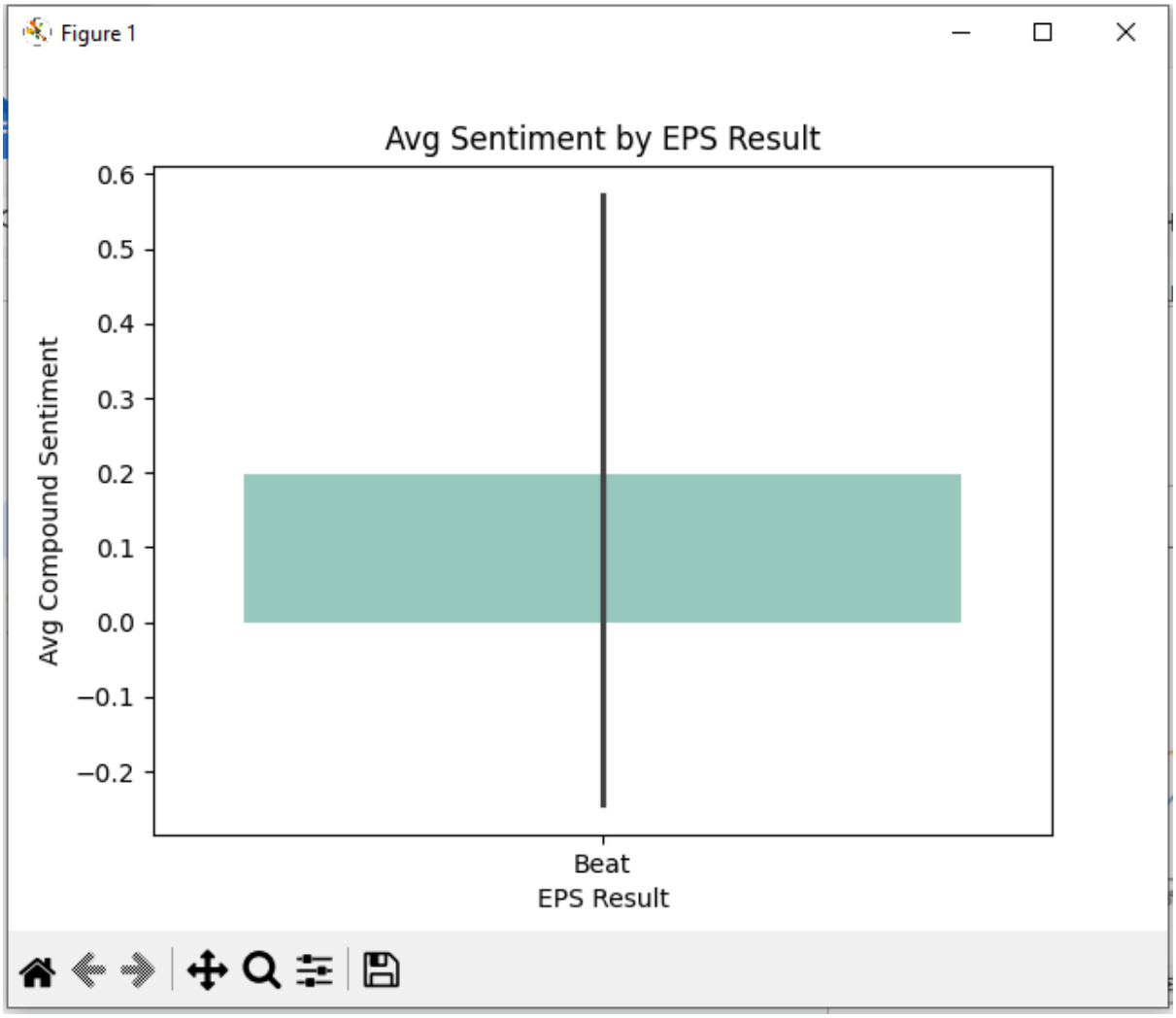
Biểu đồ 2: Line plot sentiment theo thời gian



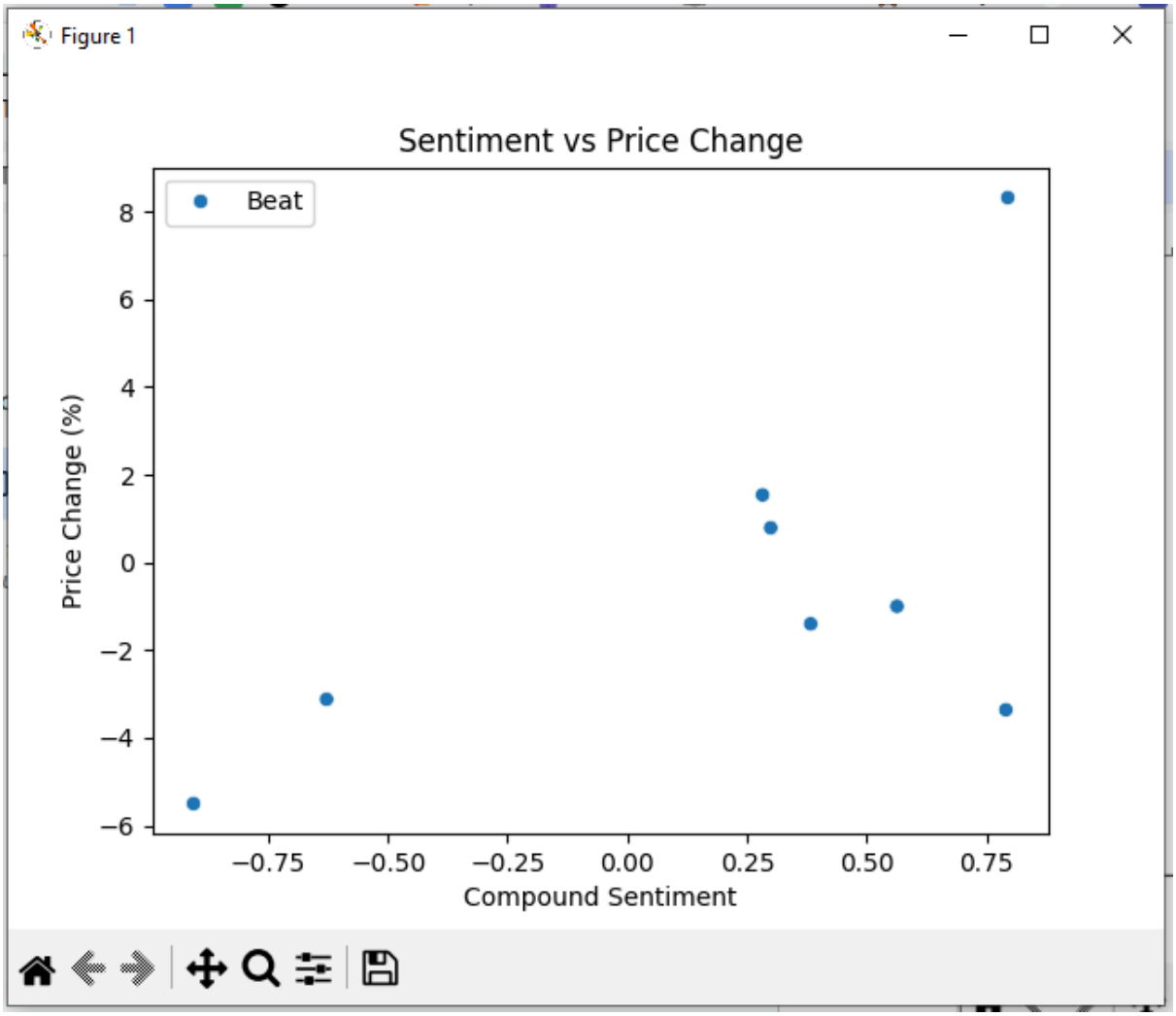
Biểu đồ 3: Biểu đồ Line hoặc Bar theo earnings_date



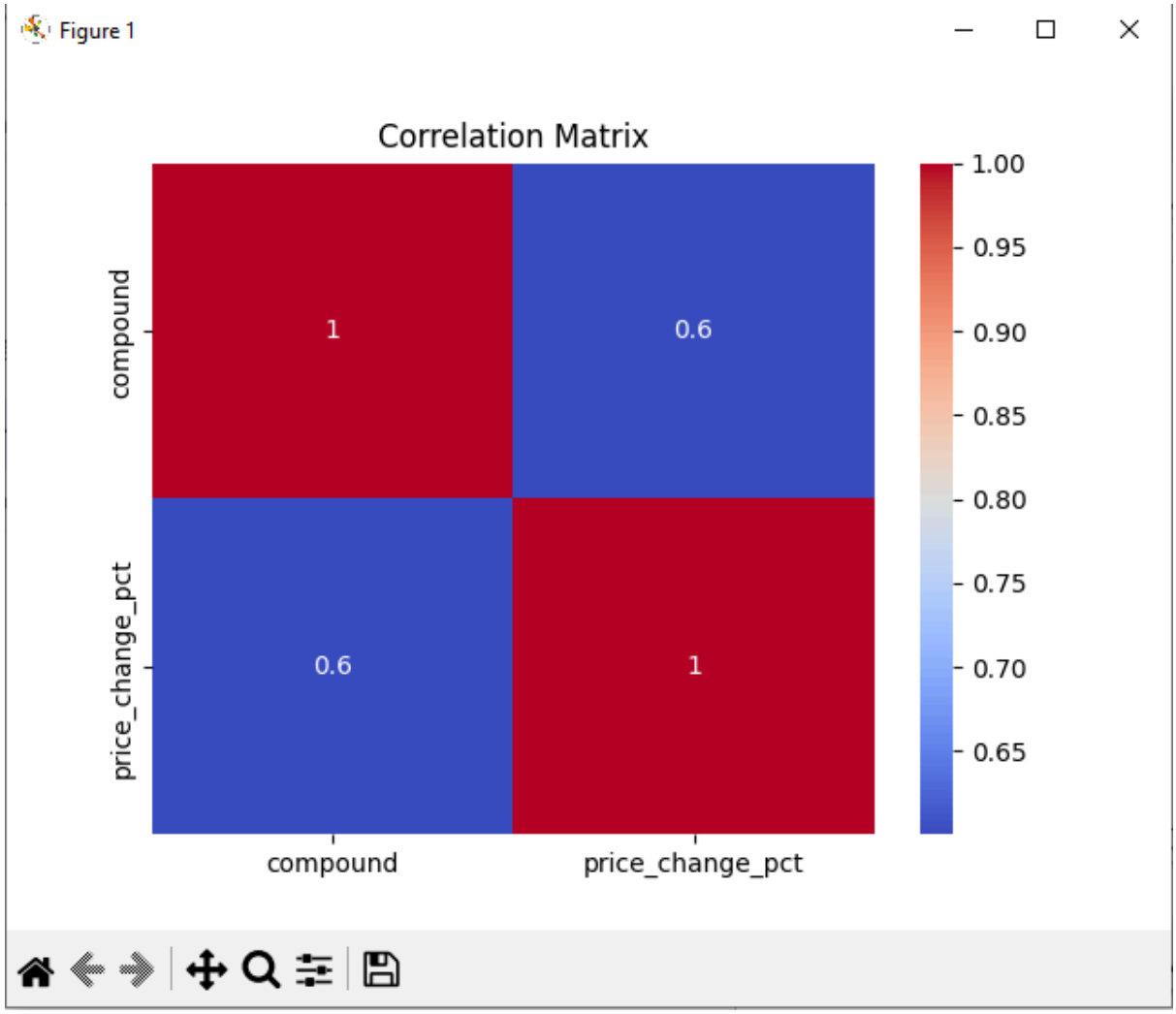
Biểu đồ 4: Bar Chart sentiment theo eps_result



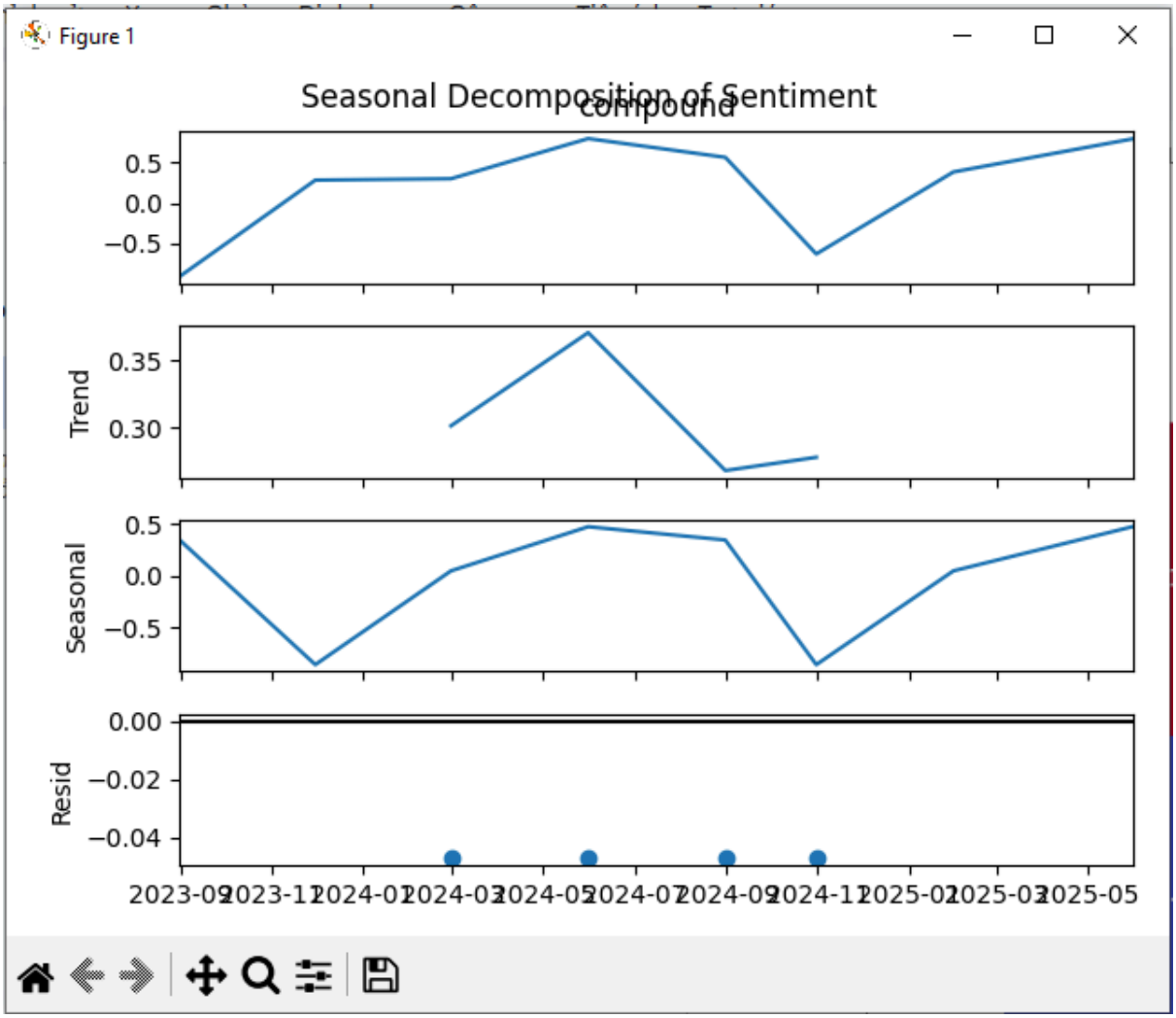
Biểu đồ 5: Scatter plot sentiment vs price change



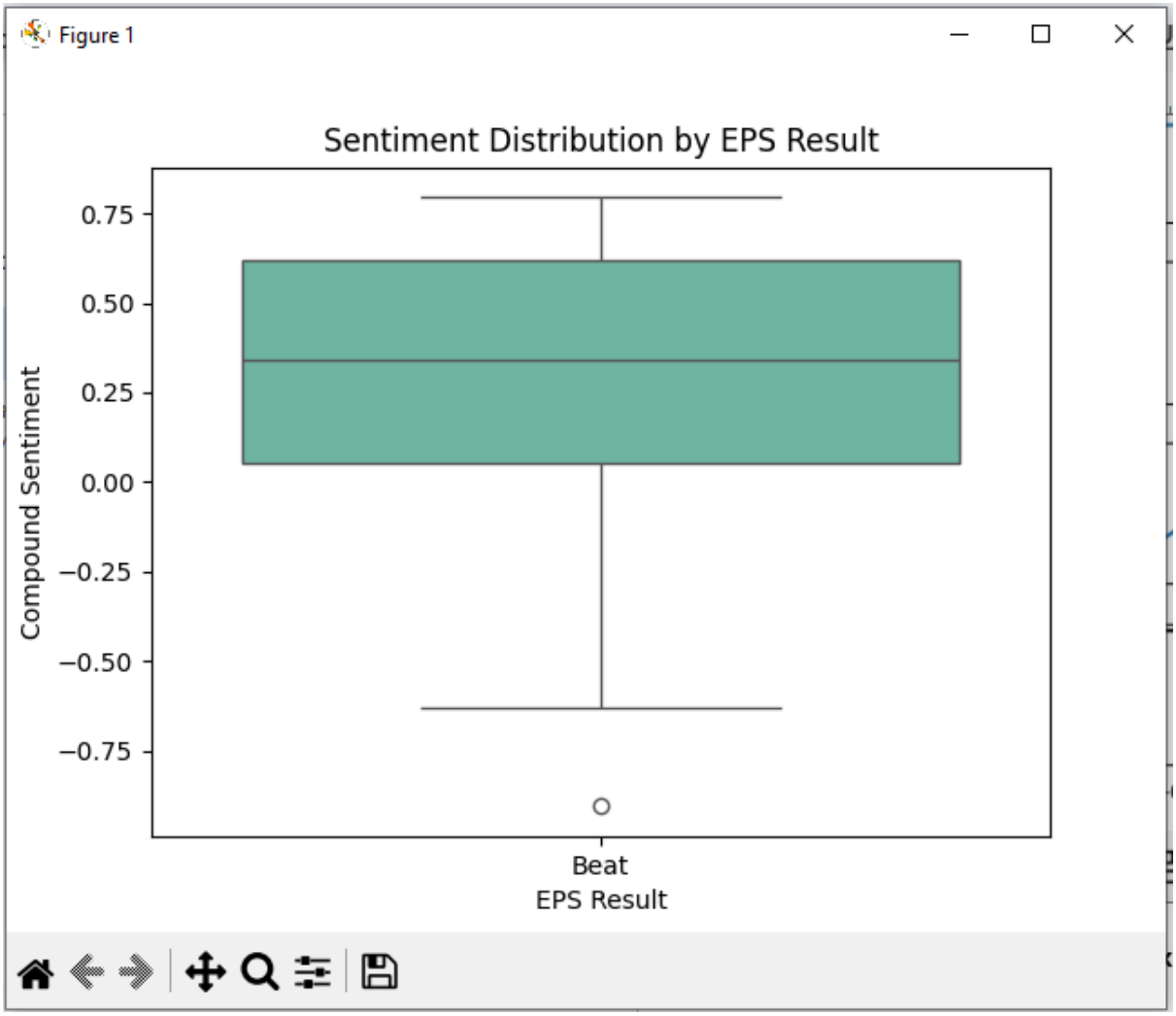
Biểu đồ 6: Heatmap correlation



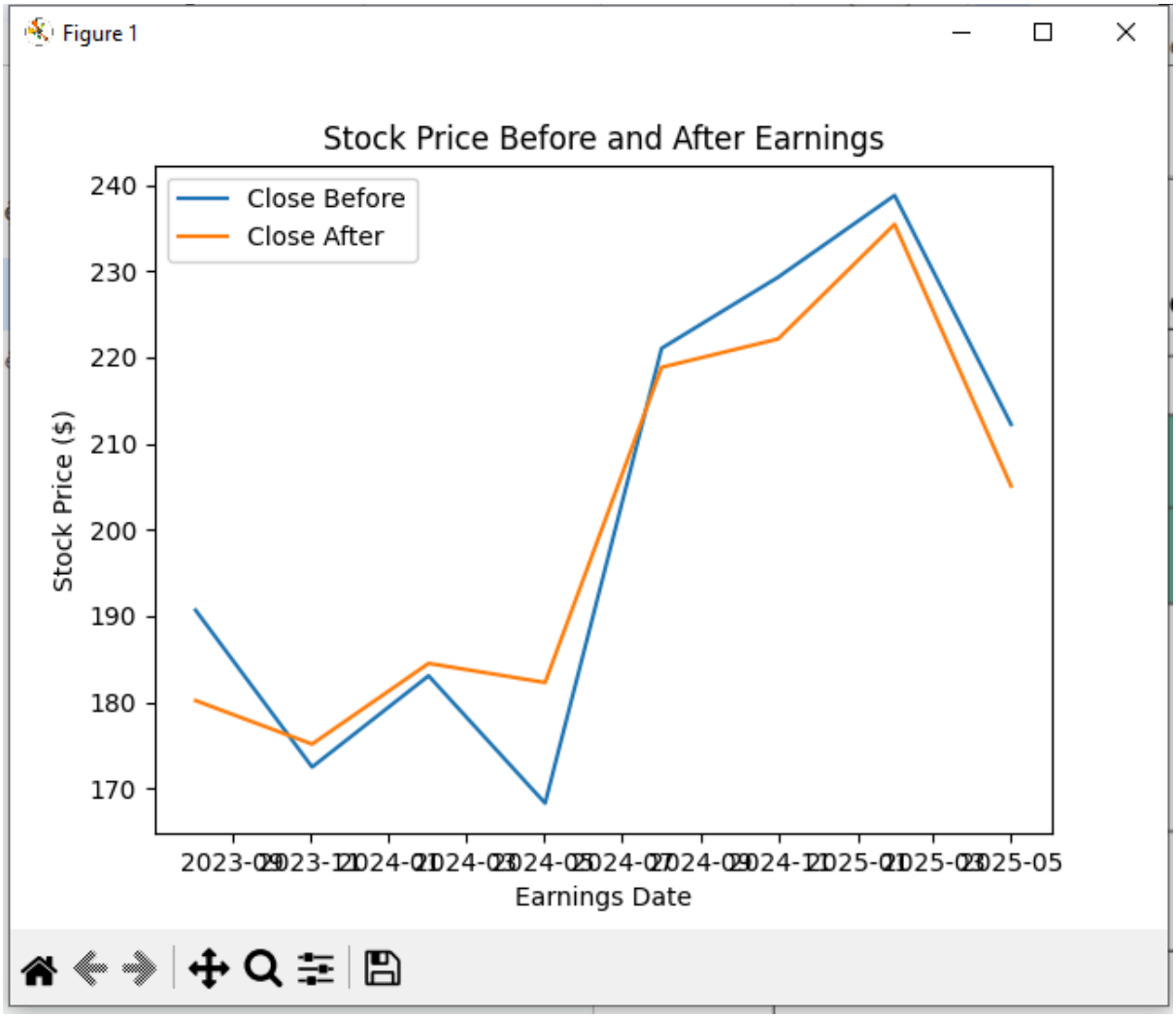
Biểu đồ 7: Seasonal decomposition (nếu đủ dữ liệu)



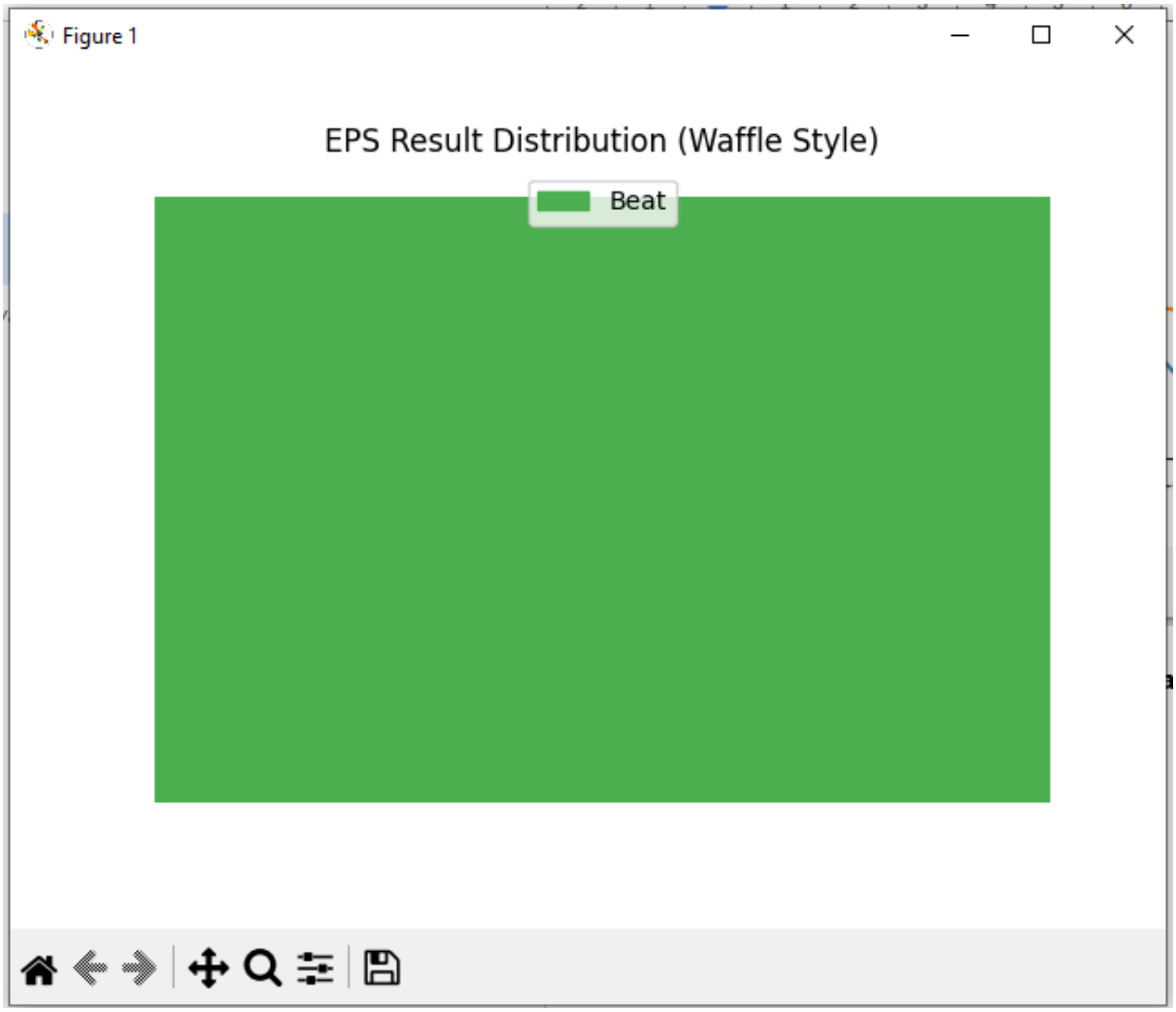
Biểu đồ 8: Boxplot sentiment theo eps_result



Biểu đồ 9: Line plot giá trước/sau



Biểu đồ 10: Waffle chart eps_result (nếu thư viện phù hợp)



6. Machine Learning Model Implementation

Sử dụng 5 mô hình

Linear Regression

Random Forest Regressor

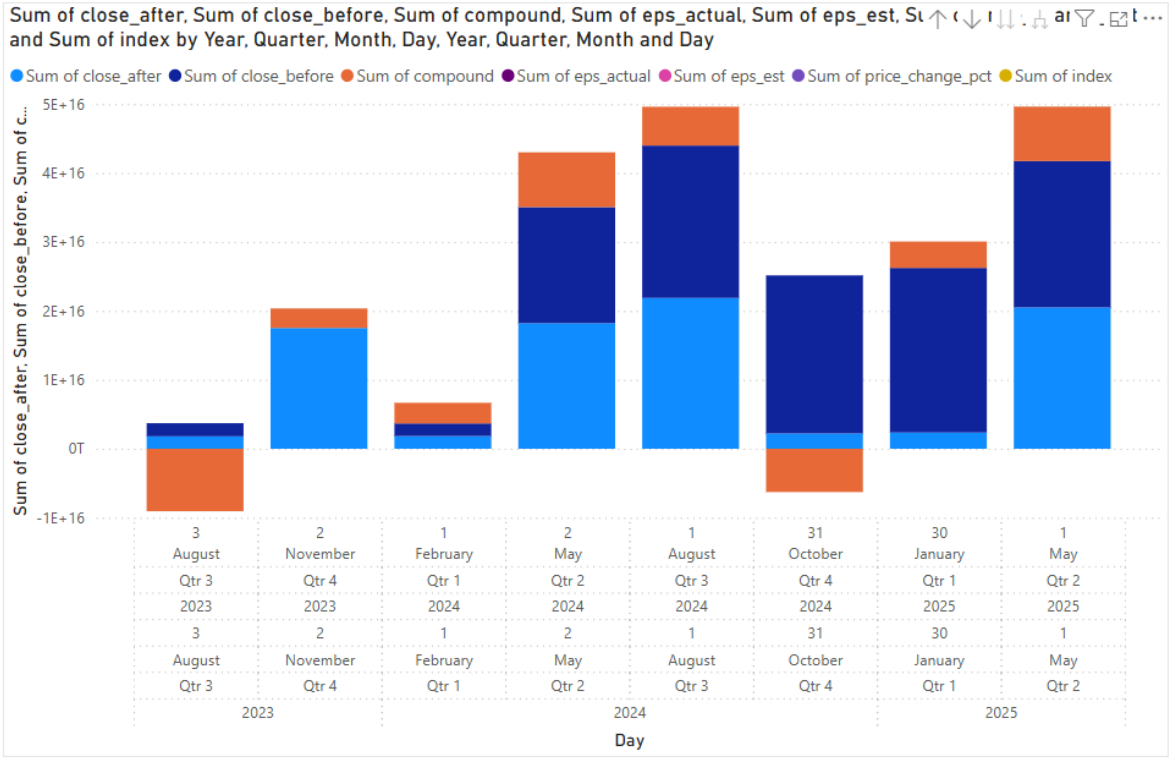
Decision Tree Regressor

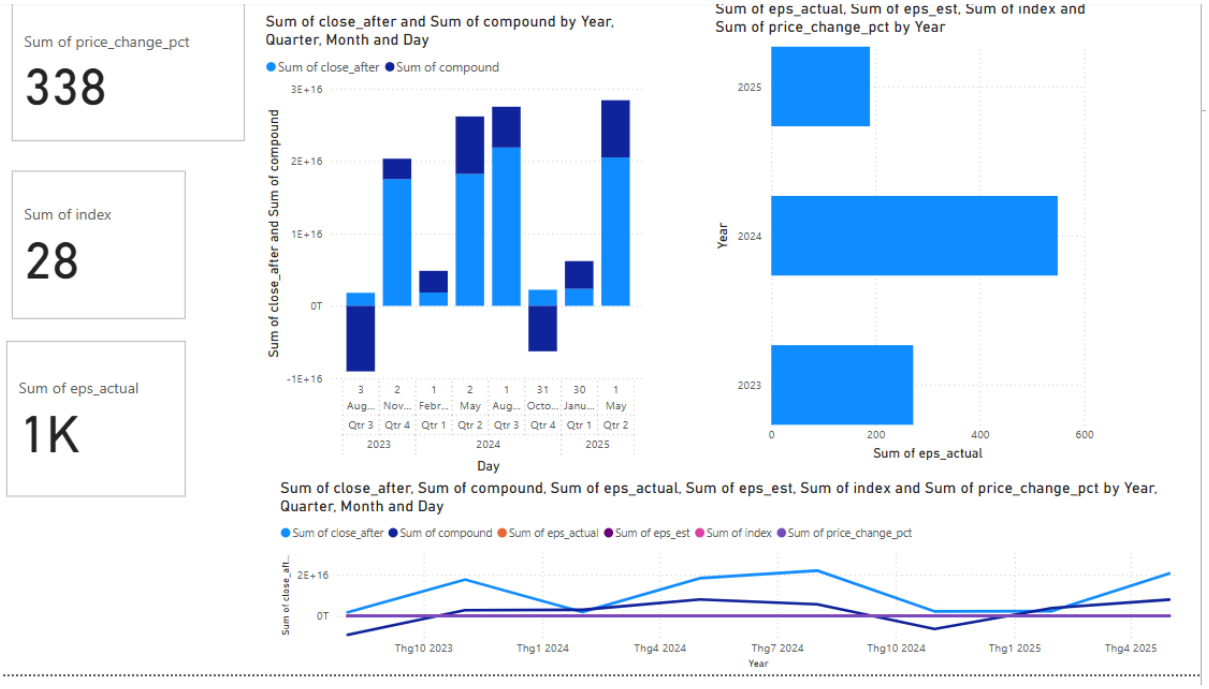
Support Vector Regressor (SVR)

K-Nearest Neighbors Regressor (KNN)

- So sánh hiệu năng giữa các mô hình và lựa chọn mô hình tốt nhất.

7. Data Analysis with BI Tool





Lab 2. In-depth analysis and visualization of results

1. Business Understanding & Analytic Approach

1.1 Người dùng cuối và nhu cầu:

- Người dùng chính:
• Nhà đầu tư cá nhân: Cần hiểu tác động của báo cáo thu nhập lên giá cổ phiếu để ra quyết định đầu tư
• Nhà phân tích tài chính: Cần xác thực các giả thuyết về mối quan hệ giữa thu nhập và giá cổ phiếu
• Quản lý danh mục đầu tư: Cần theo dõi tác động của thu nhập để điều chỉnh chiến lược đầu tư
• Nhu cầu từ kết quả:
• Dự đoán biến động giá sau khi công bố thu nhập
• Phân tích mức độ ảnh hưởng của thu nhập đến giá
• Đánh giá độ tin cậy của các dự đoán

1.2 Giả thuyết cần xem xét lại:

• Giả thuyết ban đầu: Thu nhập có tác động trực tiếp và tức thì lên giá cổ phiếu
• Cần xem xét lại:
• Tác động của thu nhập có thể kéo dài nhiều ngày sau khi công bố
• Mức độ ảnh hưởng có thể khác nhau giữa các quý
• Tác động có thể phụ thuộc vào kỳ vọng thị trường

1.3 KPIs và mục tiêu:

• Kết quả Lab 1:
• RMSE hiện tại: Cần kiểm tra lại từ dữ liệu
• Chưa đạt được mục tiêu dự đoán chính xác
• Mục tiêu mới:
• RMSE < 2% giá cổ phiếu
• MAE < 1.5% giá cổ phiếu
• $R^2 > 0.7$

1.4 Tầm quan trọng của phân tích chênh lệch giá:

• Giúp xác định thời điểm tác động mạnh nhất của thu nhập
• Đánh giá mức độ phản ứng của thị trường
• Phát hiện các mẫu hình biến động giá
• Tối ưu hóa thời điểm giao dịch

1.5 Câu hỏi nghiên cứu chính:

• Tác động của thu nhập đến giá cổ phiếu có thay đổi theo thời gian không?
• Mức độ ảnh hưởng của thu nhập có phụ thuộc vào kỳ vọng thị trường không?

• Có thể dự đoán chính xác biến động giá sau khi công bố thu nhập không?

1.6 Các yếu tố ảnh hưởng khác:

• Tâm lý thị trường
• Tin tức và sự kiện
• Chỉ số kinh tế vĩ mô
• Biến động thị trường chung
• Hoạt động của công ty
• Chính sách của Fed

1.7 Ý nghĩa của biểu đồ:

• Biểu đồ đường:
• Theo dõi xu hướng giá theo thời gian
• So sánh biến động giá trước/sau thu nhập
• Phát hiện mẫu hình biến động
• Biểu đồ cột:
• So sánh mức độ tác động giữa các quý
• Phân tích chênh lệch thu nhập
• Hiển thị độ lớn biến động

1.8 Kiểm định thống kê:

• Xác định ý nghĩa thống kê của tác động thu nhập
• Đánh giá độ tin cậy của mô hình dự đoán
• So sánh hiệu năng giữa các mô hình
• Kiểm tra tính độc lập của các biến

1.9 Hỗ trợ từ Dashboard Power BI:

• Trực quan hóa dữ liệu thu nhập và giá
• Theo dõi biến động theo thời gian thực
• So sánh dự đoán với thực tế
• Phân tích tương quan giữa các yếu tố
• Tương tác với dữ liệu để khám phá insights

1.10Thách thức khi phân tích:

• Dữ liệu thu nhập không đầy đủ hoặc không chính xác
• Nhiễu từ các yếu tố thị trường khác
• Khó xác định tác động trực tiếp của thu nhập
• Biến động giá có thể bị ảnh hưởng bởi nhiều yếu tố
• Khó dự đoán chính xác thời điểm tác động
• Cần xử lý dữ liệu lớn và phức tạp

2.Data Collection, Understanding & Preparation

Sử dụng thư viện **yfinance** để tải dữ liệu giá cổ phiếu của AAPL (Apple Inc.),MSFT (Microsoft Corp), GOOGL (Google Inc) trong khoảng thời gian từ 1/1/2023 đến 30/6/2025

2.1AAPL_MSFT_GOOGL-PRICE.py > ...

```python
import yfinance as yf
import pandas as pd

# Danh sách các mã cổ phiếu
symbols = ["AAPL", "MSFT", "GOOGL"]

# Tải và lưu dữ liệu giá cho từng mã
for symbol in symbols:
    price_data = yf.download(symbol, start='2023-01-01', end='2025-06-30')
    filename = f"{symbol}_price.csv"
    price_data.to_csv(filename)
    print(f"Đã lưu {filename}")
```

Kết Quả : Thu được 3 file dữ liệu giá (AAPL_price.csv, MSFT_price.csv, GOOGL_price.csv)
Số bản ghi: ~620 dòng/mã (từ 2023-01-03 đến 2025-06-30, mỗi ngày giao dịch 1 dòng)

Các cột trong dữ liệu bao gồm:

Date: Ngày giao dịch.
Close: Giá đóng cửa của cổ phiếu.
High: Giá cao nhất trong ngày.
Low: Giá thấp nhất trong ngày.
Open: Giá mở cửa.
Volume: Khối lượng giao dịch.

Sử dụng thư viện **yfinance** tiếp tục tải về một **DataFrame chứa các ngày công bố báo cáo tài chính (earnings).**

```python
import yfinance as yf
import pandas as pd

def download_earnings(ticker_symbol, start_date, end_date, output_file):
    ticker = yf.Ticker(ticker_symbol)
    earnings_df = ticker.earnings_dates
    earnings_df = earnings_df.reset_index()
    earnings_df["Earnings Date"] = pd.to_datetime(earnings_df["Earnings Date"])
    earnings_df = earnings_df.set_index("Earnings Date")
    earnings_df = earnings_df.sort_index()
    earnings_df = earnings_df[start_date:end_date]
    earnings_df.to_csv(output_file)
    print(f"Đã lưu dữ liệu earnings vào file: {output_file}")

if __name__ == "__main__":
    symbols = ["AAPL", "MSFT", "GOOGL"]
    start_date = "2023-01-01"
    end_date = "2025-06-30"
    for symbol in symbols:
        output_file = f"{symbol}_earnings.csv"
        download_earnings(symbol, start_date, end_date, output_file)
```

Kết Quả : Thu được 3 file dữ liệu (AAPL_earnings.csv, MSFT_earnings.csv, GOOGL_earnings.csv)

AAPL_earnings.csv > data

```
Earnings Date,EPS Estimate,Reported EPS,Surprise(%)
2023-02-02 16:30:00-05:00,1.94,1.88,-2.88
2023-05-04 16:30:00-04:00,1.43,1.52,6.03
2023-08-03 16:30:00-04:00,1.19,1.26,5.49
2023-11-02 16:30:00-04:00,1.39,1.46,4.92
2024-02-01 16:00:00-05:00,2.1,2.18,3.9
2024-05-02 16:31:00-04:00,1.5,1.53,1.97
2024-08-01 16:30:00-04:00,1.35,1.4,3.99
2024-10-31 16:31:00-04:00,1.6,1.64,2.35
2025-01-30 16:31:00-05:00,2.35,2.4,2.15
2025-05-01 16:30:00-04:00,1.63,1.65,1.41
```

GOOGL_earnings.csv > data

```
Earnings Date,EPS Estimate,Reported EPS,Surprise(%)
2023-02-02 16:02:00-05:00,1.18,1.05,-10.73
2023-04-25 16:02:00-04:00,1.07,1.17,9.72
2023-07-25 16:01:00-04:00,1.34,1.44,7.54
2023-10-24 16:00:00-04:00,1.45,1.55,6.84
2024-01-30 16:00:00-05:00,1.59,1.64,2.98
2024-04-25 16:01:00-04:00,1.51,1.89,24.77
2024-07-23 16:09:00-04:00,1.84,1.89,2.47
2024-10-29 16:03:00-04:00,1.85,2.12,14.87
2025-02-04 16:04:00-05:00,2.13,2.15,1.04
2025-04-24 16:06:00-04:00,2.01,2.27,12.81
```

MSFT_earnings.csv > data

```
Earnings Date,EPS Estimate,Reported EPS,Surprise(%)
2023-01-24 16:04:00-05:00,2.29,2.32,1.09
2023-04-25 16:05:00-04:00,2.23,2.45,9.81
2023-07-25 16:01:00-04:00,2.55,2.69,5.49
2023-10-24 16:02:00-04:00,2.65,2.99,12.7
2024-01-30 16:04:00-05:00,2.78,2.93,5.34
2024-04-25 16:23:00-04:00,2.82,2.94,4.32
2024-07-30 16:03:00-04:00,2.93,2.95,0.54
2024-10-30 16:01:00-04:00,3.1,3.3,6.52
2025-01-29 16:01:00-05:00,3.11,3.23,3.82
2025-04-30 16:05:00-04:00,3.22,3.46,7.45
```

- **Tạo dữ liệu giả định sentiment (cảm xúc thị trường) quanh các ngày công bố báo cáo tài chính (earnings) của một mã cổ phiếu.**

2.3SENTIMENT.py > ...

```python
import yfinance as yf
import pandas as pd
import numpy as np

def generate_sentiment(ticker_symbol, start_date, end_date, output_file):
    ticker = yf.Ticker(ticker_symbol)
    earnings_df = ticker.earnings_dates.reset_index()

    # Chuyển cột Earnings Date thành datetime và bỏ timezone
    earnings_df["Earnings Date"] = pd.to_datetime(earnings_df["Earnings Date"]).dt.tz_localize(None)
    start_date_dt = pd.to_datetime(start_date)
    end_date_dt = pd.to_datetime(end_date)
    earnings_df = earnings_df[(earnings_df["Earnings Date"] >= start_date_dt) & (earnings_df["Earnings Date"] <= end_date_dt)]
    earnings_df['compound'] = np.random.uniform(-1, 1, size=len(earnings_df))
    sentiment_df = earnings_df[['Earnings Date', 'compound']]
    sentiment_df.to_csv(output_file, index=False)
    print(f"Đã tạo file sentiment giả định: {output_file}")

if __name__ == "__main__":
    symbols = ["AAPL", "MSFT", "GOOGL"]
    start_date = "2023-01-01"
    end_date = "2025-06-30"
    for symbol in symbols:
        output_file = f"{symbol}_sentiment.csv"
        generate_sentiment(symbol, start_date, end_date, output_file)
```

Kết Quả : Thu được 3 file dữ liệu (AAPL_sentiment.csv, MSFT_sentiment.csv, GOOGL_sentiment.csv)

AAPL_sentiment.csv > data

```
Earnings Date,compound
2025-05-01 16:30:00,0.289688990561608
2025-01-30 16:31:00,-0.6854859914414706
2024-10-31 16:31:00,-0.6229380325297142
2024-08-01 16:30:00,-0.6072688240561481
2024-05-02 16:31:00,0.05949053729005316
2024-02-01 16:00:00,0.10035914423143577
2023-11-02 16:30:00,-0.43886645018310566
2023-08-03 16:30:00,0.7431540387616469
2023-05-04 16:30:00,0.6185817056239138
2023-02-02 16:30:00,0.31053253287970684
```

GOOGL_sentiment.csv > data

```
1   Earnings Date,compound
2   2025-04-24 16:06:00,-0.24212270400424707
3   2025-02-04 16:
4   2024-10-29 16:
5   2024-07-23 16:09:00,0.6794870017158363
6   2024-04-25 16:01:00,0.6506543605133808
7   2024-01-30 16:00:00,-0.0704331047820359
8   2023-10-24 16:00:00,-0.8406689359749491
9   2023-07-25 16:01:00,-0.0630541564392948
10  2023-04-25 16:02:00,-0.16229020171435882
11  2023-02-02 16:02:00,-0.24064220897727107
```

Col 1: Earnings Date

MSFT_sentiment.csv > data

```
1   Earnings Date,compound
2   2025-04-30 16:05:00,-0.05562109461768472
3   2025-01-29 16:01:00,-0.0003158097752686828
4   2024-10-30 16:01:00,0.25974630621285044
5   2024-07-30 16:03:00,0.9584387428349614
6   2024-04-25 16:23:00,0.2721790192548945
7   2024-01-30 16:04:00,-0.8212097802923846
8   2023-10-24 16:02:00,0.9582958422568437
9   2023-07-25 16:01:00,-0.2628593294353989
10  2023-04-25 16:05:00,0.5943781272278281
11  2023-01-24 16:04:00,-0.04620051788443935
```

- **Tiền xử lý dữ liệu:**

2.4TIEN_XU_LY_DU_LIEU.py > ...

```python
import pandas as pd
import sys

files = ["AAPL_price.csv", "MSFT_price.csv", "GOOGL_price.csv"]

for file in files:
    print(f"\n--- Xử lý file: {file} ---")
    try:
        # Đọc dữ liệu, bỏ 2 dòng đầu không hợp lệ
        df = pd.read_csv(file, skiprows=2)

        # Đổi tên cột đầu tiên nếu cần
        if 'Price' in df.columns:
            df.rename(columns={'Price': 'Adj Close'}, inplace=True)

        # Chuyển cột Date sang dạng datetime
        if 'Date' in df.columns:
            df['Date'] = pd.to_datetime(df['Date'])
            # Đặt cột Date làm chỉ mục
            df.set_index('Date', inplace=True)

        # Kiểm tra missing values
        print("\nSố lượng missing values trong mỗi cột:")
        print(df.isnull().sum())

        # Điền missing values bằng phương pháp forward fill
        df.fillna(method='ffill', inplace=True)

        # Kiểm tra lại sau khi điền
        print("\nSố lượng missing values sau khi điền:")
        print(df.isnull().sum())

        # Hiển thị thông tin tổng quát về dữ liệu
        print("\nThông tin tổng quát về dữ liệu:")
        print(df.info())

    except FileNotFoundError:
```

Kết Quả :

```
Số lượng missing values trong mỗi cột:
Unnamed: 1    0
Unnamed: 2    0
Unnamed: 3    0
Unnamed: 4    0
Unnamed: 5    0
dtype: int64
C:\lab2\2.4TIEN_XU_LY_DU_LIEU.py:27: FutureWarning: D
  df.fillna(method='ffill', inplace=True)

Số lượng missing values sau khi điền:
Unnamed: 1    0
Unnamed: 2    0
Unnamed: 3    0
Unnamed: 4    0
Unnamed: 5    0
dtype: int64

Thông tin tổng quát về dữ liệu:
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 604 entries, 2023-01-03 to 2025-05-30
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 1  604 non-null    float64
 1   Unnamed: 2  604 non-null    float64
 2   Unnamed: 3  604 non-null    float64
 3   Unnamed: 4  604 non-null    float64
 4   Unnamed: 5  604 non-null    int64
```

- Không có missing values trong 5 cột:
- Tổng số bản ghi: 604 ngày (giao dịch).

- Khoảng thời gian: từ `2023-01-03` đến `2025-05-30`.
- Tổng số biến: 5.

- Mô tả cấu trúc dữ liệu (df.describe()):

--- Mô tả cấu trúc dữ liệu cho AAPL ---

	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
count	604.000000	604.000000	604.000000	604.000000	6.040000e+02
mean	193.831351	195.564030	191.815857	193.555745	5.824560e+07
std	29.041035	29.306577	28.690447	29.067597	2.530666e+07
min	123.421257	126.136090	122.582104	124.398597	2.323470e+07
25%	172.469471	174.007037	171.123872	172.297044	4.490442e+07
50%	189.000984	190.308038	187.673212	188.774889	5.239195e+07
75%	221.234215	223.532831	218.932923	220.746344	6.485765e+07
max	258.396667	259.474086	257.010028	257.568678	3.186799e+08

--- Mô tả cấu trúc dữ liệu cho MSFT ---

	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
count	604.000000	604.000000	604.000000	604.000000	6.040000e+02
mean	193.831351	195.564030	191.815857	193.555745	5.824560e+07
std	29.041035	29.306577	28.690447	29.067597	2.530666e+07
min	123.421257	126.136090	122.582104	124.398597	2.323470e+07
25%	172.469471	174.007037	171.123872	172.297044	4.490442e+07
50%	189.000984	190.308038	187.673212	188.774889	5.239195e+07
75%	221.234215	223.532831	218.932923	220.746344	6.485765e+07
max	258.396667	259.474086	257.010028	257.568678	3.186799e+08

--- Mô tả cấu trúc dữ liệu cho GOOGL —

	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
count	604.000000	604.000000	604.000000	604.000000	6.040000e+02
mean	145.948725	147.541506	144.306966	145.825865	3.103864e+07
std	28.412629	28.723734	28.203762	28.517719	1.338291e+07
min	85.686119	87.047955	84.354095	85.467420	1.024210e+07
25%	127.666843	129.056005	126.618129	127.743868	2.261742e+07
50%	147.145515	149.478536	145.894278	147.689256	2.755745e+07
75%	167.519081	169.138028	165.411708	167.622948	3.520225e+07
max	205.893341	206.561759	202.331752	202.910386	1.274901e+08

Nhận xét :

📊 **Mô tả cấu trúc dữ liệu cho AAPL**

- Các cột `Unnamed: 1` → `Unnamed: 4` có giá trị trung bình dao động quanh 191–195, phù hợp với các loại giá cổ phiếu như Open, High, Low, và Close.

- Cột `Unnamed: 5` chứa các giá trị rất lớn, biến động từ ~23 triệu đến hơn 318 triệu, cho thấy đây nhiều khả năng là Volume (khối lượng giao dịch).

- Độ lệch chuẩn khá cao (≈29 cho giá và ≈25 triệu cho volume), cho thấy cổ phiếu AAPL có mức độ biến động tương đối mạnh trong giai đoạn từ đầu 2023 đến giữa 2025.

- Phân vị:

    - 25%: giá từ khoảng 171–174; volume khoảng 44.9 triệu

    - 50% (median): giá khoảng 187–190; volume khoảng 52.3 triệu

    - 75%: giá từ 218–223; volume gần 65 triệu
      → Biên độ dao động lớn giữa các phân vị cho thấy thị trường có biến động rõ rệt trong giai đoạn khảo sát.

📊 **Mô tả cấu trúc dữ liệu cho MSFT**

- Các cột `Unnamed: 1` → `Unnamed: 4` có giá trị trung bình khoảng 367–374, nằm trong khoảng phổ biến của giá cổ phiếu, rất có khả năng là các giá Open, High, Low, và Close.

- Cột `Unnamed: 5` có giá trị rất lớn, dao động từ ~7 triệu đến gần 78 triệu, rất có khả năng đại diện cho Volume (khối lượng giao dịch).

- Độ lệch chuẩn cao (std ~60), cho thấy cổ phiếu MSFT có mức độ biến động tương đối lớn trong giai đoạn từ đầu 2023 đến giữa 2025.

- Phân vị:

    - 25%: khoảng 325–329 (giá thấp hơn)

    - 50% (median): khoảng 386–395

    - 75%: khoảng 414–421
      → Phân phối giá có xu hướng tăng, nhưng vẫn giữ mức phân tán cao.

📊 **Mô tả cấu trúc dữ liệu cho GOOGL**

- Các cột `Unnamed: 1` → `Unnamed: 4` có giá trị trung bình từ 144 đến 147, phù hợp với các chỉ số giá cổ phiếu (Open, High, Low, Close).
- Cột `Unnamed: 5` có giá trị dao động từ ~10 triệu đến hơn 127 triệu, rất có thể là khối lượng giao dịch (Volume).
- Độ lệch chuẩn của giá (~28) và khối lượng (~13 triệu) cho thấy mức biến động vừa phải, ít hơn so với MSFT.
- Phân vị:
    - 25%: khoảng 126–129
    - 50% (median): khoảng 145–149
    - 75%: khoảng 165–169
      → Phân phối giá khá đều, với trung vị và trung bình gần nhau → dữ liệu không bị lệch mạnh.

Xử Lý Dữ Liệu Scaling ta dùng công thức:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

--- Xử lý file: AAPL_price.csv ---

Dữ liệu gốc:

Unnamed: 1 Unnamed: 2 Unnamed: 3 Unnamed: 4 Unnamed: 5

Date

2023-01-03 123.470596 129.226036 122.582104 128.613970 112117500

2023-01-04 124.744125 127.014716 123.480495 125.267347 89113600

2023-01-05 123.421257 126.136090 123.164587 125.504275 80962700

2023-01-06 127.962425 128.623856 123.292916 124.398597 87754700

2023-01-09 128.485641 131.703962 128.228972 128.801557 70790800

Dữ liệu sau khi scaling:

Unnamed: 1 Unnamed: 2 Unnamed: 3 Unnamed: 4 Unnamed: 5

Date

2023-01-03 0.000366 0.023174 0.000000 0.031654 0.300844

2023-01-04 0.009801 0.006589 0.006683 0.006524 0.222982

2023-01-05 0.000000 0.000000 0.004333 0.008303 0.195393

2023-01-06 0.033644 0.018658 0.005288 0.000000 0.218382

2023-01-09 0.037521 0.041758 0.042007 0.033063 0.160964

Đã lưu kết quả vào file AAPL_scaled.csv

--- Xử lý file: MSFT_price.csv ---

Dữ liệu gốc:

Unnamed: 1 Unnamed: 2 Unnamed: 3 Unnamed: 4 Unnamed: 5

Date

2023-01-03 234.808960 240.856088 232.672365 238.239260 25740000

2023-01-04 224.537659 228.232571 221.460190 227.654324 50623400

2023-01-05 217.882874 223.018529 217.343823 222.675493 39585600

2023-01-06 220.450684 221.264157 214.981818 218.559125 43613600

2023-01-09 222.597061 226.635024 221.901209 221.940406 27369800

Dữ liệu sau khi scaling:

Unnamed: 1 Unnamed: 2 Unnamed: 3 Unnamed: 4 Unnamed: 5

Date

2023-01-03 0.068772 0.080452 0.071929 0.080364 0.260476

2023-01-04 0.027039 0.028615 0.026341 0.037140 0.609405

2023-01-05 0.000000 0.007204 0.009604 0.016809 0.454627

2023-01-06 0.010433 0.000000 0.000000 0.000000 0.511109

2023-01-09 0.019154 0.022055 0.028134 0.013807 0.283330

Đã lưu kết quả vào file MSFT_scaled.csv

--- Xử lý file: GOOGL_price.csv ---

Dữ liệu gốc:

Unnamed: 1 Unnamed: 2 Unnamed: 3 Unnamed: 4 Unnamed: 5

Date

2023-01-03 88.588707 90.507201 87.992278 89.055899 28131200

2023-01-04 87.554901 90.109579 86.749725 89.811365 34854800

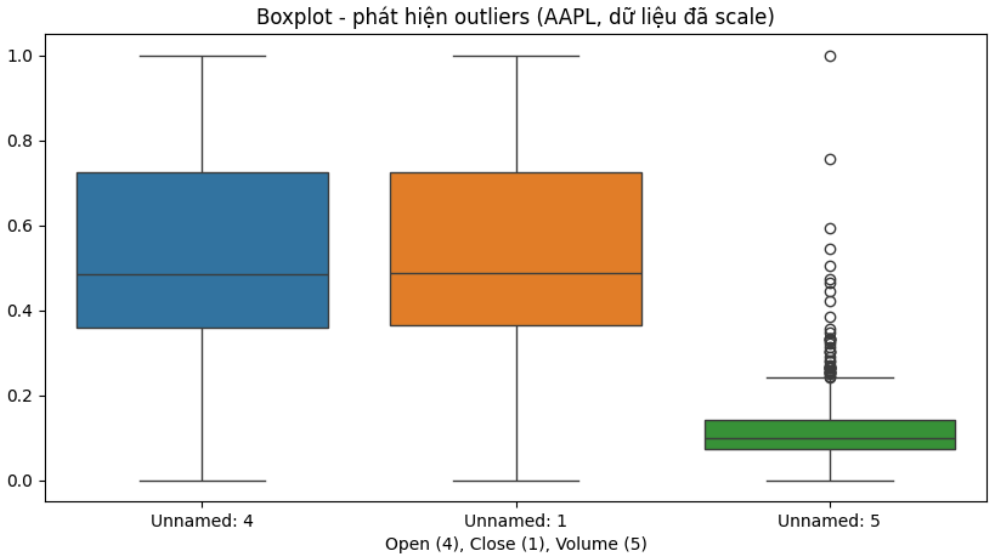
2023-01-05	85.686119	87.047955	85.387912	86.948552	27194400
2023-01-06	86.819305	87.167225	84.354095	86.272589	41381500
2023-01-09	87.495255	89.513159	87.336212	87.833232	29003900

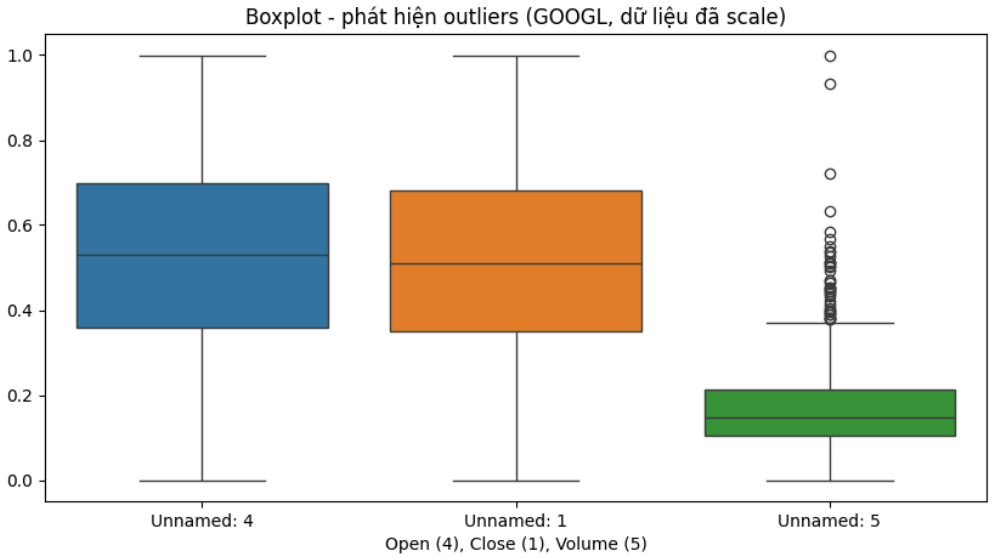
Dữ liệu sau khi scaling:

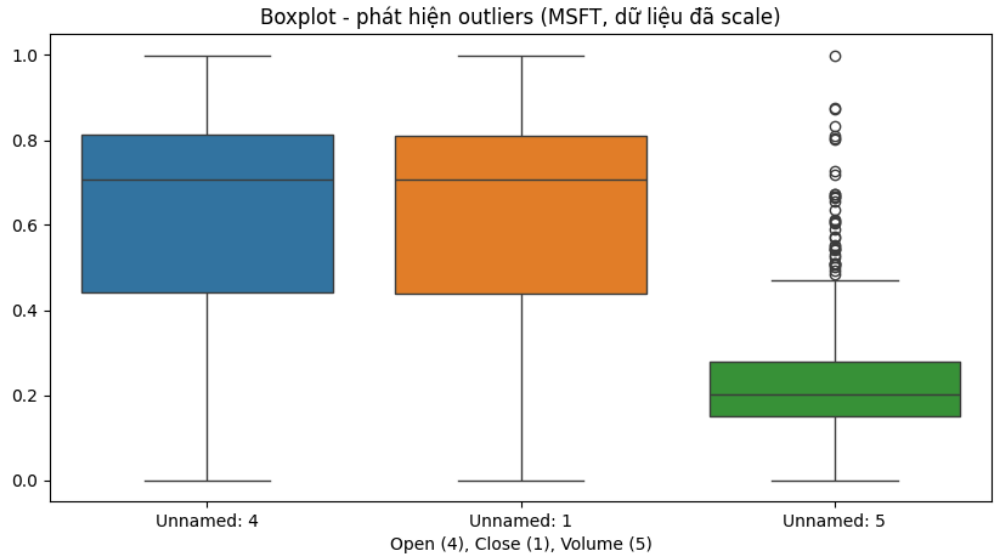
Date	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
2023-01-03	0.024147	0.028944	0.030838	0.030555	0.152575
2023-01-04	0.015546	0.025617	0.020306	0.036988	0.209920
2023-01-05	0.000000	0.000000	0.008763	0.012612	0.144585
2023-01-06	0.009427	0.000998	0.000000	0.006856	0.265586
2023-01-09	0.015050	0.020627	0.025277	0.020144	0.160018

Đã lưu kết quả vào file GOOGL_scaled.csv

- **Boxplot - phát hiện outliers (dữ liệu đã scale):**


Boxplot - phát hiện outliers (AAPL, dữ liệu đã scale)
1.0
0.8
0.6
0.4
0.2
0.0
Unnamed: 4
Unnamed: 1
Unnamed: 5
Open (4), Close (1), Volume (5)


Boxplot - phát hiện outliers (GOOGL, dữ liệu đã scale)
1.0
0.8
0.6
0.4
0.2
0.0
Unnamed: 4
Unnamed: 1
Unnamed: 5
Open (4), Close (1), Volume (5)



Nhận xét :

Trả lời câu hỏi :

1. Dữ liệu 1 tháng bao gồm những biến nào (bao gồm earnings), và nguồn từ đâu?

- Biến giá cổ phiếu (price): Date, Adj Close, Open, Low, High, Volume (lấy từ Yahoo Finance qua yfinance).
- Biến earnings: Earnings Date, EPS Estimate, Reported EPS (lấy từ yfinance).
- Biến sentiment: Earnings Date, compound (điểm sentiment giả lập).
- Biến feature: rolling_mean_5, volatility_5, pct_change (tính toán từ giá).
- Nguồn: Tất cả dữ liệu đều lấy từ Yahoo Finance thông qua thư viện yfinance, sentiment là dữ liệu giả lập.

---

2. Số bản ghi và kiểu dữ liệu của từng biến trong tập dữ liệu là gì?

- File giá: ~620 bản ghi/mã (mỗi ngày giao dịch), kiểu: Date (datetime), các cột giá/volume (float64/int64).
- File earnings/sentiment: ~10-12 bản ghi/mã (mỗi lần công bố earnings), kiểu: Earnings Date (datetime), EPS Estimate/Reported EPS/compound (float64).
- File feature: ~620 bản ghi/mã, các biến rolling_mean_5, volatility_5, pct_change (float64).
- File scaled: ~620 bản ghi/mã, các biến số đã chuẩn hóa (float64).

---

3. Nhóm đã phát hiện bao nhiêu giá trị thiếu, và cách xử lý là gì?

- Giá trị thiếu:
- Dữ liệu giá: Có thể có missing do ngày nghỉ, đã xử lý bằng forward fill (ffill).
- Dữ liệu feature: rolling_mean_5, volatility_5 có NaN ở những dòng đầu (do thiếu cửa sổ) .
- Dữ liệu earnings/sentiment: Không có missing values.
- Cách xử lý:
- Dùng fillna(method='ffill') cho dữ liệu giá.
- Đối với feature loại bỏ 5 dòng đầu khi training.

---

4. Phương pháp nào được sử dụng để phát hiện outliers trong giá cổ phiếu?

- Phương pháp:
- Boxplot (trực quan hóa) để phát hiện outlier.
- IQR (Interquartile Range): Đánh dấu giá trị nằm ngoài [Q1 - 1.5*IQR, Q3 + 1.5*IQR].
- Z-score: Đánh dấu giá trị có |z| > 3 là outlier.
- Thực hiện: Đã có file vẽ boxplot cho cả 3 mã để phát hiện outlier.

---

5. Đặc trưng mới nào (chênh lệch giá trước/sau earnings,...) đã được tạo, và ý nghĩa là gì?

- Chênh lệch giá trước/sau earnings: price_change_pct = (close_after - close_before) / close_before * 100 (% thay đổi giá quanh ngày earnings).
- rolling_mean_5: Trung bình động 5 ngày, giúp nhận diện xu hướng ngắn hạn.
- volatility_5: Độ lệch chuẩn 5 ngày, đo biến động giá.
- pct_change: % thay đổi giá ngày liền kề.
- Ý nghĩa: Các đặc trưng này giúp mô hình học máy hiểu rõ hơn về xu hướng, biến động, và tác động của sự kiện earnings đến giá cổ phiếu.

---

6. Dữ liệu đã được chuẩn hóa/scale như thế nào, và tại sao cần thiết?

- Chuẩn hóa: Sử dụng Min-Max scaling (đưa các cột số về khoảng [0, 1]).
- Lý do:
- Đảm bảo các biến số có cùng thang đo, tránh mô hình bị lệch do biến có giá trị lớn.

- Cần thiết cho các thuật toán machine learning (đặc biệt là các thuật toán dựa trên khoảng cách như KNN, SVM, v.v.).

---

7. Nhóm đã lưu dữ liệu vào cơ sở dữ liệu nào, và kiểm tra tính toàn vẹn ra sao?

- Cơ sở dữ liệu: SQLite (tạo 3 file: aapl_analysis.db, msft_analysis.db, googl_analysis.db).
- Kiểm tra toàn vẹn:
- Đếm số dòng từng bảng, so sánh với file gốc.
- Kiểm tra tên cột, kiểu dữ liệu, và truy vấn thử (SELECT, JOIN) để đảm bảo dữ liệu đúng, không thiếu/trùng.

---

8. Những thách thức nào gặp phải khi thu thập dữ liệu earnings?

- Thách thức:
- Dữ liệu earnings trên yfinance đôi khi thiếu hoặc cập nhật chậm.
- Định dạng ngày tháng có thể khác nhau, cần xử lý chuẩn hóa.
- Một số mã có thể không có đủ lịch sử earnings hoặc bị thiếu cột.
- Cách xử lý:
- Kiểm tra, làm sạch, chuẩn hóa định dạng ngày tháng.
- Bổ sung dữ liệu thủ công nếu cần.

---

9. Dữ liệu có bao nhiêu báo cáo earnings, và điều này ảnh hưởng đến phân tích thế nào?

- Số báo cáo earnings: Khoảng 10-12 bản ghi/mã (tùy số lần công bố earnings mỗi năm).
- Ảnh hưởng:
- Số lượng nhỏ, nên phân tích thống kê hoặc machine learning sẽ bị hạn chế về độ tin cậy, dễ bị overfitting.
- Kết quả mô hình cần được kiểm tra cẩn thận, không nên quá tin vào dự báo với tập dữ liệu nhỏ.

---

10. Nhóm đã xử lý thế nào nếu dữ liệu có giá trị không hợp lý?

- Cách xử lý:
- Phát hiện outlier bằng boxplot, IQR, Z-score.

- Loại bỏ hoặc thay thế giá trị outlier bằng giá trị gần nhất (interpolation, ffill).
- Kiểm tra và loại bỏ các dòng có giá trị âm, 0 hoặc bất thường ở các cột giá/volume.

3. Data Analysis with SQL

Dùng sqlite3 để import dữ liệu

```python
.1IMPORT_DULIEU_SQL.py > ...
import sqlite3
import pandas as pd

symbols = ["AAPL", "MSFT", "GOOGL"]

db_names = {
    "AAPL": "aapl_analysis.db",
    "MSFT": "msft_analysis.db",
    "GOOGL": "googl_analysis.db"
}

for symbol in symbols:
    # Đọc CSV và chuyển đổi định dạng ngày tháng
    price_df = pd.read_csv(f'{symbol}_price.csv', header=2, parse_dates=['Date'])
    # Đổi tên cột giá đóng cửa thành 'close' nếu cần
    price_df.rename(columns={'Adj Close': 'close', 'Unnamed: 1': 'close'}, inplace=True)
    price_df['Date'] = price_df['Date'].dt.strftime('%Y-%m-%d')
    print(f"{symbol} price columns: {price_df.columns}")

    earnings_df = pd.read_csv(f'{symbol}_earnings.csv')
    earnings_df['Earnings Date'] = pd.to_datetime(earnings_df['Earnings Date'], errors='coerce', u
    earnings_df['Earnings Date'] = earnings_df['Earnings Date'].dt.tz_convert(None).dt.strftime('%

    sentiment_df = pd.read_csv(f'{symbol}_sentiment.csv')
    sentiment_df['Earnings Date'] = pd.to_datetime(sentiment_df['Earnings Date'], errors='coerce',
    sentiment_df['Earnings Date'] = sentiment_df['Earnings Date'].dt.tz_convert(None).dt.strftime(

    # Tạo kết nối DB SQLite riêng cho từng mã
    conn = sqlite3.connect(db_names[symbol])

    # Đưa dữ liệu vào DB
    price_df.to_sql('price', conn, if_exists='replace', index=False)
    earnings_df.to_sql('earnings', conn, if_exists='replace', index=False)
    sentiment_df.to_sql('sentiment', conn, if_exists='replace', index=False)
    print(f"Đã import dữ liệu cho {symbol} vào {db_names[symbol]}")
```

Truy vấn Ngày volume bất thường (volume > mean + 2*std) được xử lý bằng pandas

```
df_price = pd.read_sql_query('SELECT Date, [Unnamed: 5] as volume, close FROM price', conn)
mean_vol = df_price['volume'].mean()
std_vol = df_price['volume'].std()
outlier_df = df_price[df_price['volume'] > mean_vol + 2*std_vol]
print("Ngày volume bất thường:")
print(outlier_df)
```

Truy vấn Moving average 5 ngày cho giá đóng cửa

```
query_ma = '''
SELECT Date, close,
       AVG(close) OVER (ORDER BY Date ROWS BETWEEN 4 PRECEDING AND CURRENT ROW) AS ma_5
FROM price
'''
df_ma = pd.read_sql_query(query_ma, conn)
print("Moving average 5 ngày:")
print(df_ma.head(10))
```

Truy vấn phân cụm theo nghành

```
# 3. Phân cụm theo ngành (giả lập)
print(f"Ngành của {symbol}: {industry_map[symbol]}")
conn.close()
```

Kết Quả :

--- Truy vấn nâng cao cho AAPL ---

Ngày volume bất thường:

	Date	volume	close
0	2023-01-03	112117500	123.470596
21	2023-02-02	118339000	148.891357
22	2023-02-03	154357300	152.524261
85	2023-05-05	113316400	171.612030
105	2023-06-05	121946500	177.799835

147 2023-08-04 115799700 180.185928

170 2023-09-07 112488800 176.037247

176 2023-09-15 109205100 173.509094

240 2023-12-15 128256700 196.133698

290 2024-02-29 136682600 179.664948

301 2024-03-15 121664700 171.583725

335 2024-05-03 163224100 182.279160

361 2024-06-11 172373300 206.185730

362 2024-06-12 198134300 212.078201

368 2024-06-21 246421400 206.524170

398 2024-08-05 119548600 208.295868

431 2024-09-20 318679900 227.400650

495 2024-12-20 147495300 253.877594

565 2025-04-04 125910900 188.133301

566 2025-04-07 160466300 181.222366

567 2025-04-08 120859500 172.194199

568 2025-04-09 184395900 198.589584

569 2025-04-10 121880000 190.170624

Moving average 5 ngày:

Date close ma_5

0 2023-01-03 123.470596 123.470596

1 2023-01-04 124.744125 124.107361

2 2023-01-05 123.421257 123.878660

3 2023-01-06 127.962425 124.899601

4 2023-01-09 128.485641 125.616809

5 2023-01-10 129.058212 126.734332

6 2023-01-11 131.782944 128.142096

7 2023-01-12 131.703979 129.798640

8 2023-01-13 133.036682 130.813492

9 2023-01-17 134.201614 131.956686

Ngành của AAPL: Technology

--- Truy vấn nâng cao cho MSFT ---

Ngày volume bất thường:

	Date	volume	close
1	2023-01-04	50623400	224.537659
3	2023-01-06	43613600	220.450684
15	2023-01-25	66526600	235.818436
24	2023-02-07	50841400	262.231750
25	2023-02-08	54686000	261.418335
28	2023-02-13	44630900	265.916870
49	2023-03-15	46028000	260.805573
50	2023-03-16	54768800	271.377686
51	2023-03-17	69527400	274.551300
52	2023-03-20	43466600	267.477020
77	2023-04-25	45772200	270.611359

78 2023-04-26 64599200 290.213013

79 2023-04-27 46462600 299.507843

102 2023-05-31 45950600 323.361908

114 2023-06-16 46533600 337.088379

134 2023-07-18 64872700 353.985596

137 2023-07-21 69368900 338.506348

140 2023-07-26 58383700 332.598267

204 2023-10-25 55053800 336.164062

222 2023-11-20 52465100 373.203644

240 2023-12-15 78478200 366.568970

270 2024-01-31 47871100 393.117554

301 2024-03-15 45049800 412.507568

354 2024-05-31 47995300 411.971466

431 2024-09-20 55167100 432.736755

460 2024-10-31 53971000 403.985107

495 2024-12-20 64263700 434.927856

520 2025-01-30 54586300 413.400604

565 2025-04-04 49209900 359.180603

566 2025-04-07 50425000 357.204224

568 2025-04-09 50199700 389.774414

583 2025-05-01 58938100 424.620453

Moving average 5 ngày:

Date close ma_5

0 2023-01-03 234.808960 234.808960

1 2023-01-04 224.537659 229.673309

2 2023-01-05 217.882874 225.743164

3 2023-01-06 220.450684 224.420044

4 2023-01-09 222.597061 224.055447

5 2023-01-10 224.292633 221.952182

6 2023-01-11 231.074844 223.259619

7 2023-01-12 233.760239 226.435092

8 2023-01-13 234.465912 229.238138

9 2023-01-17 235.563599 231.831445

Ngành của MSFT: Technology

--- Truy vấn nâng cao cho GOOGL ---

Ngày volume bất thường:

	Date	volume	close
12	2023-01-20	63191100	97.435638
21	2023-02-02	69883800	107.097694
22	2023-02-03	65309300	104.155342
25	2023-02-08	94743500	98.777596
26	2023-02-09	119455000	94.443588
50	2023-03-16	65492000	99.721931
51	2023-03-17	61028500	101.014183
88	2023-05-10	63153400	111.083794

89 2023-05-11 78900000 115.875053

137 2023-07-21 72937900 119.304489

140 2023-07-26 61682100 128.499344

204 2023-10-25 84366200 124.861168

270 2024-01-31 71910000 139.264786

272 2024-02-02 62470600 141.531189

302 2024-03-18 69273700 146.799591

330 2024-04-26 64665300 170.924911

368 2024-06-21 58582700 178.764053

459 2024-10-30 68890800 173.849350

459 2024-10-30 68890800 173.849350

475 2024-11-21 59734400 167.043274

488 2024-12-11 67894100 194.939224

495 2024-12-20 63462900 190.958633

524 2025-02-05 70373900 190.878830

565 2025-04-04 62259500 145.423950

566 2025-04-07 76794100 146.572556

568 2025-04-09 70406200 158.518097

587 2025-05-07 127490100 151.196960

597 2025-05-21 73416000 168.356186

598 2025-05-22 74864400 170.663391

Moving average 5 ngày:

Date close ma_5

0 2023-01-03 88.588707 88.588707

1 2023-01-04 87.554901 88.071804

2 2023-01-05 85.686119 87.276576

3 2023-01-06 86.819305 87.162258

4 2023-01-09 87.495255 87.228857

5 2023-01-10 87.892868 87.089690

6 2023-01-11 90.974388 87.773587

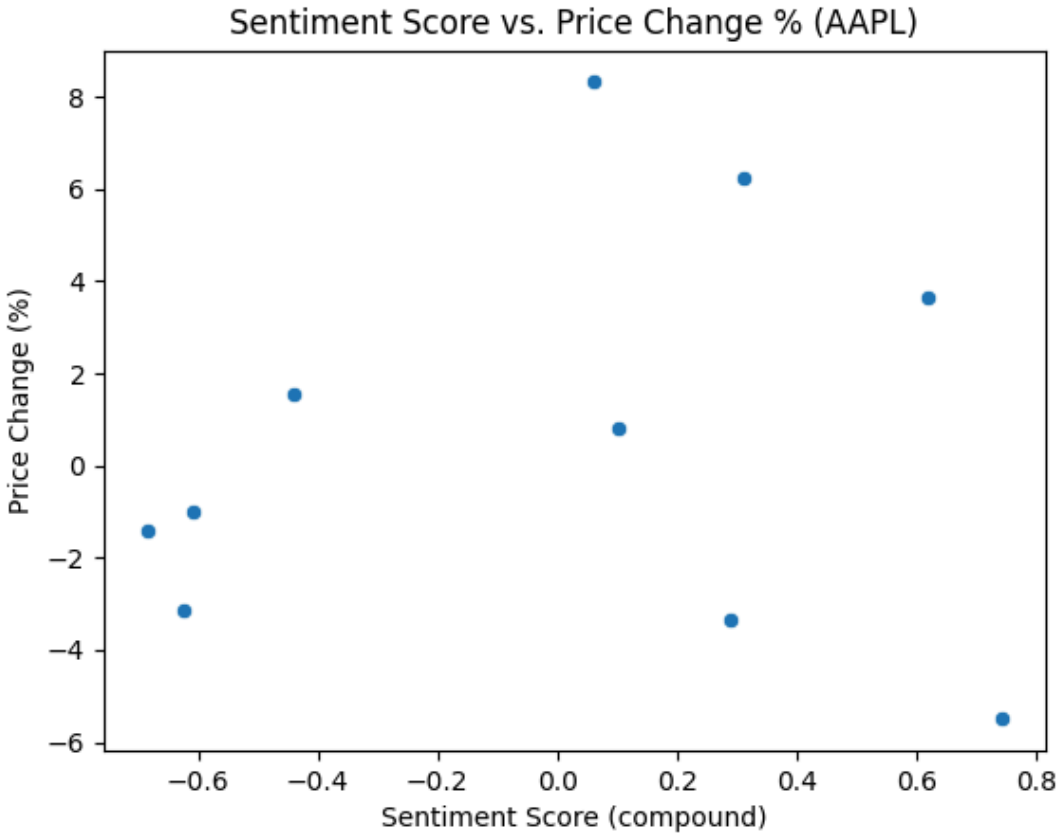
7 2023-01-12 90.586716 88.753706

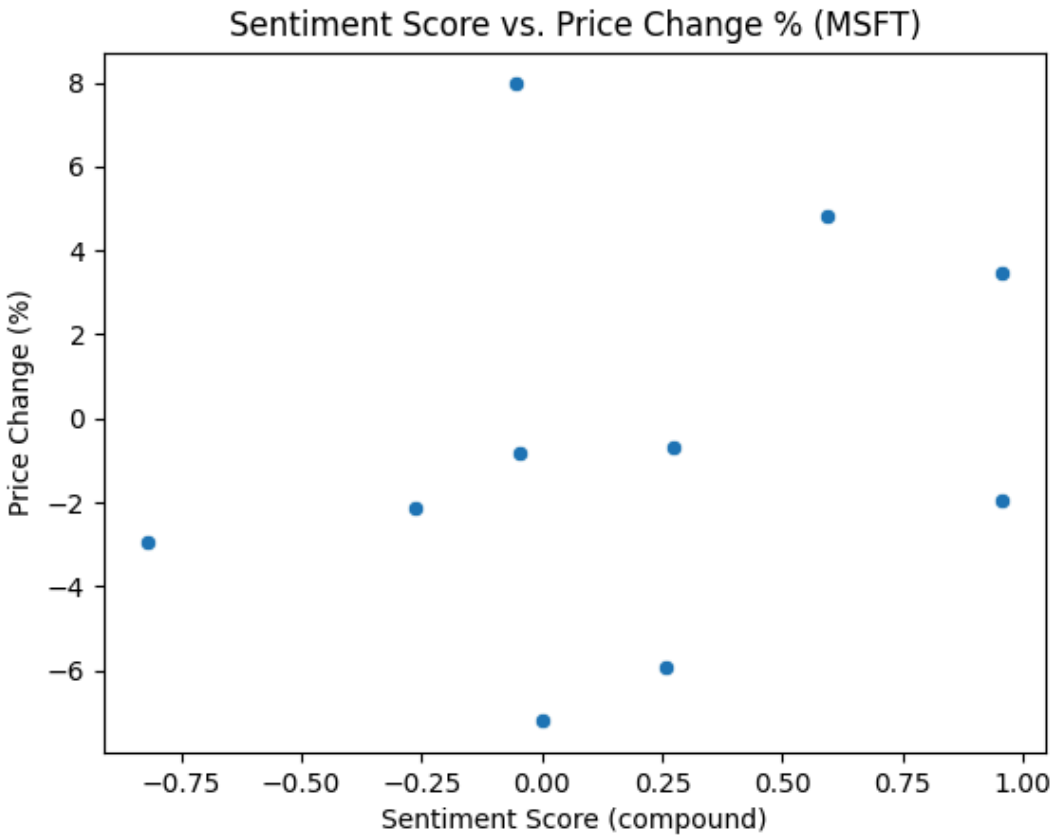
8 2023-01-13 91.570824 89.704010

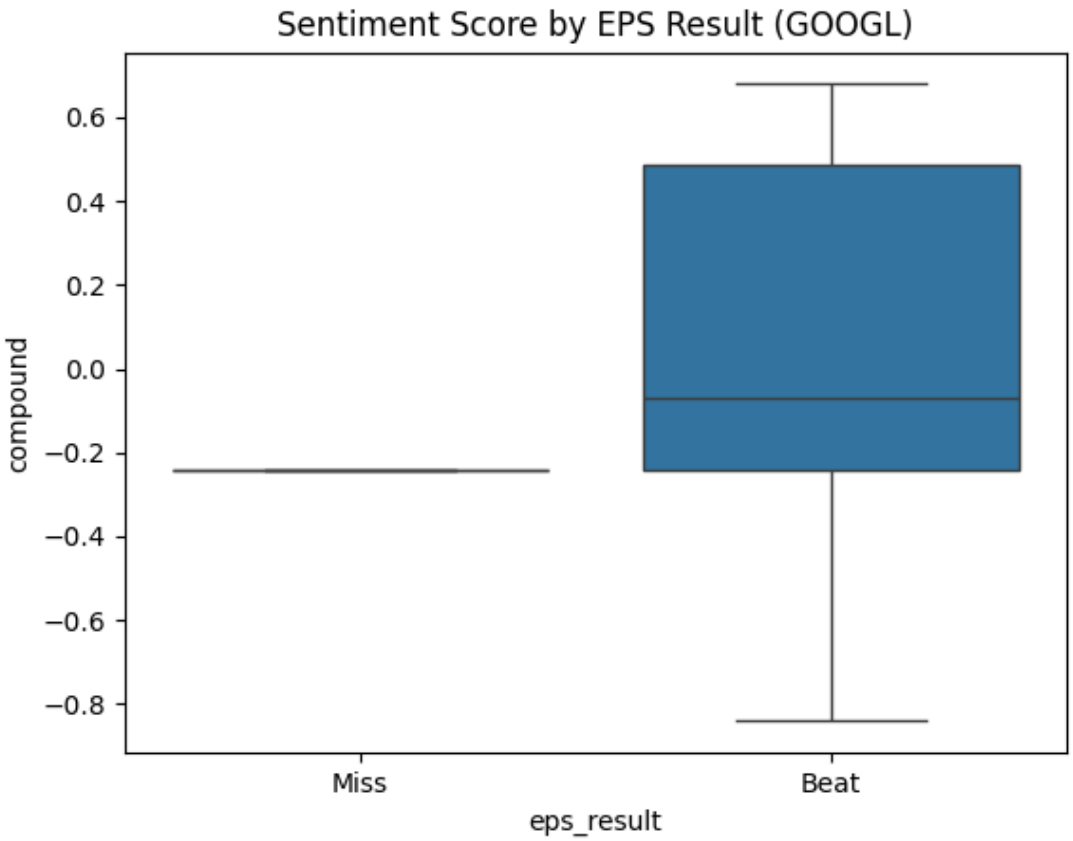
9 2023-01-17 90.745773 90.354114

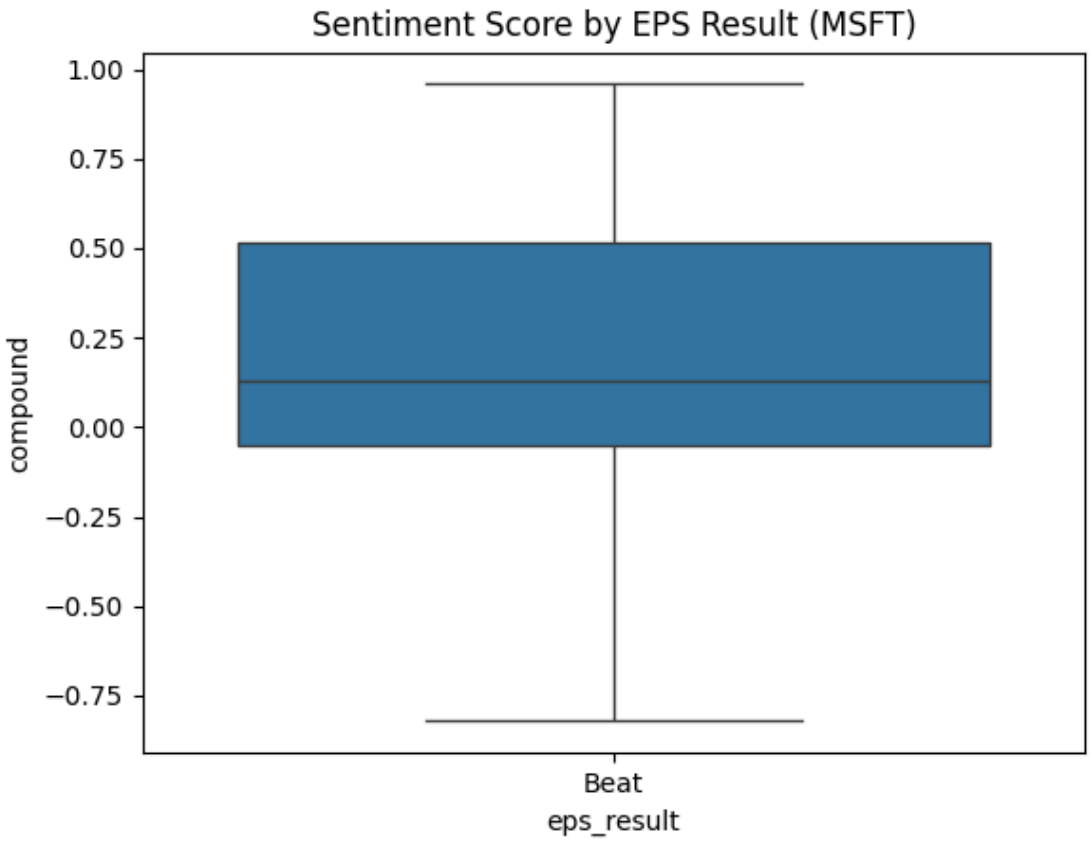
Ngành của GOOGL: Communication

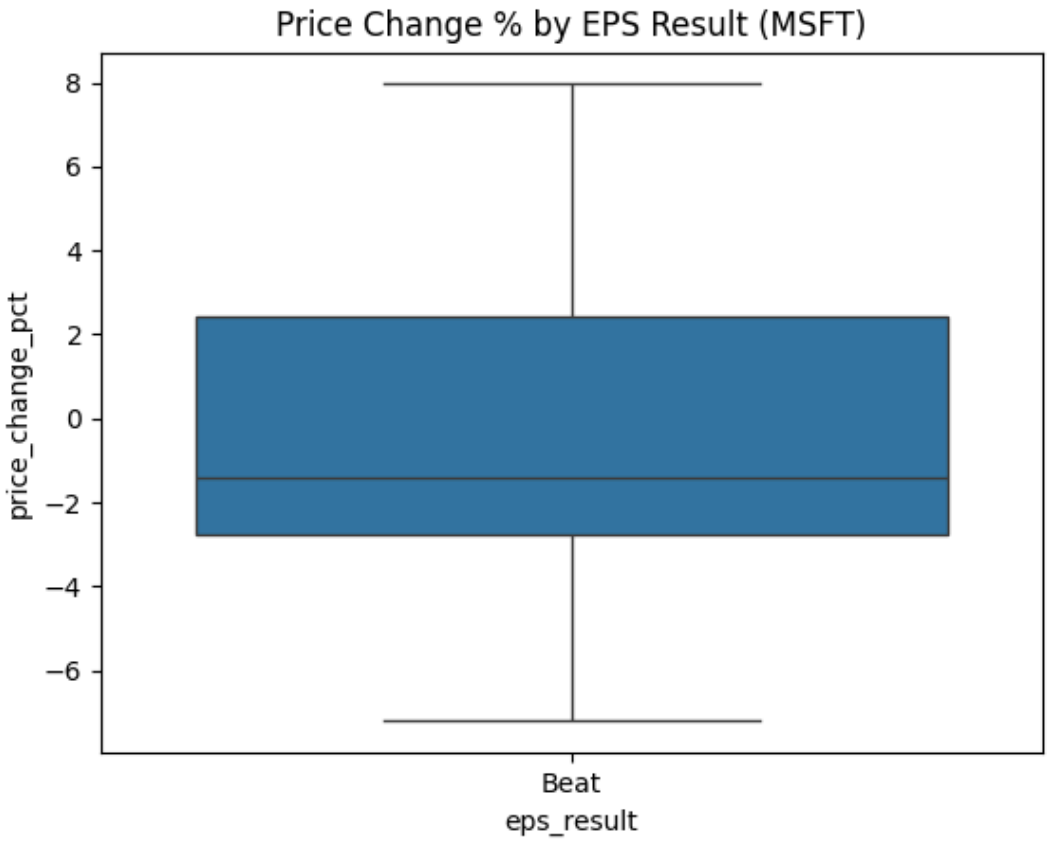
Trực Quan Hóa :

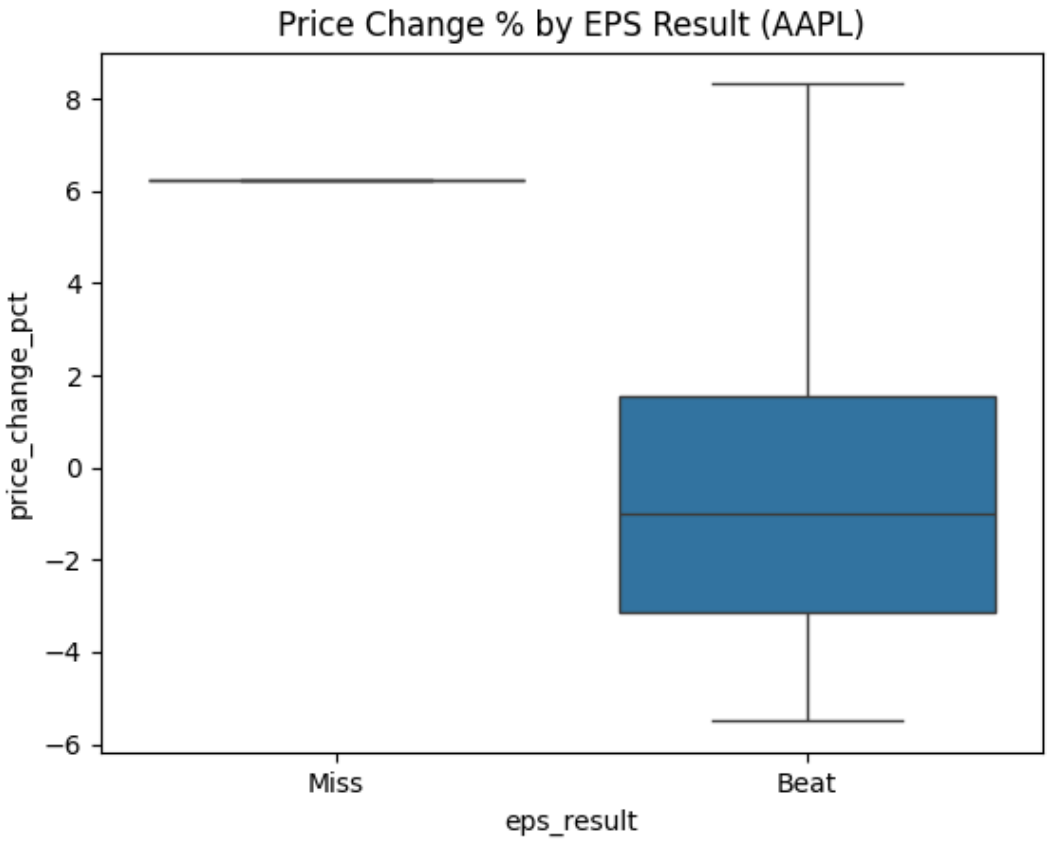

Sentiment Score vs. Price Change % (AAPL)
Price Change (%)
8
6
4
2
0
−2
−4
−6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Sentiment Score (compound)

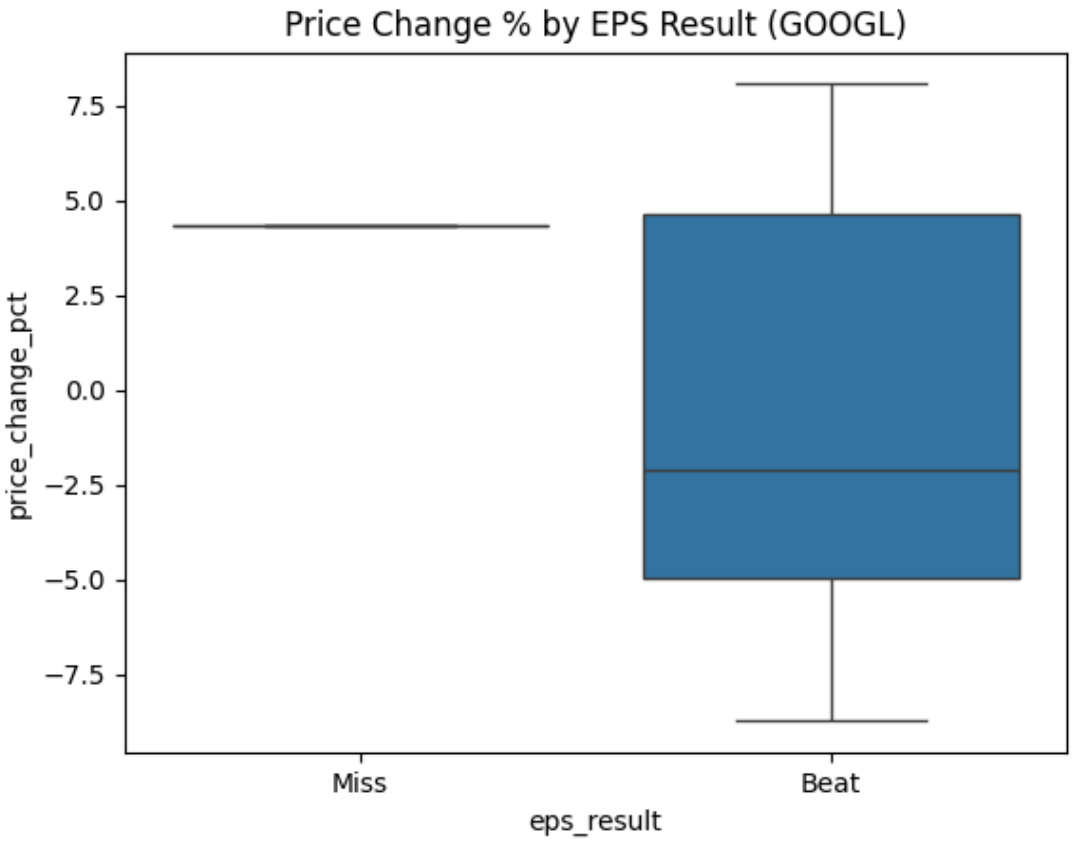

Sentiment Score vs. Price Change % (MSFT)
Price Change (%)
8
6
4
2
0
−2
−4
−6
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Score (compound)


Sentiment Score by EPS Result (GOOGL)
compound
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Miss
Beat
eps_result


Sentiment Score by EPS Result (MSFT)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
compound
Beat
eps_result


Price Change % by EPS Result (MSFT)
price_change_pct
8
6
4
2
0
−2
−4
−6
Beat
eps_result


Price Change % by EPS Result (AAPL)
price_change_pct
8
6
4
2
0
−2
−4
−6
Miss
Beat
eps_result



Apple (AAPL)

1. Sentiment Score vs. Price Change % (AAPL)

- **Nhận định: "Tương tự MSFT, không có tương quan tuyến tính rõ ràng."**
- **Chứng minh:**
  - **Điểm dữ liệu có `Sentiment Score` khoảng 0.05: `Price Change (%)` là khoảng 8.2%.**
  - **Điểm dữ liệu có `Sentiment Score` khoảng 0.75: `Price Change (%)` lại là khoảng -5.7%.**
  - **Điểm dữ liệu có `Sentiment Score` khoảng -0.6: Có cả `Price Change (%)` khoảng -0.8% và khoảng -3.5%.**
  - **Những ví dụ này cho thấy không có một mối quan hệ đơn giản và trực tiếp.**

2. Price Change % by EPS Result (AAPL)

- **Nhận định: "Khi 'Miss' EPS, giá tăng bất thường." và "Khi 'Beat' EPS, giá trung vị hơi âm."**

- **Chứng minh:**
    - **Dưới nhãn "Miss": Có một đường ngang đơn (single data point) ở mức khoảng +6.2%. Đây là một mức tăng đáng kể cho một kết quả EPS trượt.**
    - **Dưới nhãn "Beat": Đường ngang trong hộp (trung vị) nằm dưới mốc 0, cụ thể là khoảng -1.0%.**
    - **Hộp (IQR) kéo dài từ khoảng -3.2% đến +1.5%. Râu kéo dài từ khoảng -5.8% đến +8.2%.**

3. Sentiment Score by EPS Result (AAPL)

- **Nhận định: "Sentiment khi 'Miss' EPS lại là dương." và "Khi 'Beat' EPS, sentiment trung vị hơi dương, nhưng phân phối rộng."**
- **Chứng minh:**
    - **Dưới nhãn "Miss": Có một đường ngang đơn (single data point) ở mức khoảng +0.3. Sentiment dương cho một EPS trượt.**
    - **Dưới nhãn "Beat": Đường ngang trong hộp (trung vị) nằm trên mốc 0, cụ thể là khoảng +0.05.**
    - **Hộp (IQR) kéo dài từ khoảng -0.05 đến +0.3. Râu kéo dài từ khoảng -0.65 đến +0.75.**

Nhận xét :
Các biểu đồ của Apple (AAPL) cho thấy một bức tranh phức tạp về phản ứng của thị trường sau các báo cáo thu nhập. Mặc dù việc đánh bại kỳ vọng EPS là phổ biến, nhưng điều này không đảm bảo giá cổ phiếu sẽ tăng; thực tế, giá trung vị thường có xu hướng giảm nhẹ sau các báo cáo EPS tốt. Điều này ngụ ý rằng các tin tức tích cực về EPS có thể đã được định giá trước, hoặc kỳ vọng của thị trường đối với Apple là cực kỳ cao. Đáng chú ý, có một trường hợp EPS "miss" lại đi kèm với cả sentiment tích cực và giá tăng mạnh, cho thấy rằng các yếu tố khác ngoài con số EPS (ví dụ: các thông báo sản phẩm mới, triển vọng kinh doanh từ ban lãnh đạo) có thể có ảnh hưởng lớn hơn rất nhiều đến phản ứng của thị trường đối với AAPL. Nhìn chung, sentiment thị trường không phải là yếu tố dự báo giá độc lập và nhất quán cho Apple, mà chỉ là một phần của bức tranh tổng thể, thường xuyên bị chi phối bởi các yếu tố khác.

Microsoft (MSFT)

1. Sentiment Score vs. Price Change % (MSFT)

- **Nhận định: "Không có mối tương quan tuyến tính rõ ràng, mạnh mẽ."**
- **Chứng minh:**
    - **Quan sát điểm dữ liệu có `Sentiment Score` khoảng 0: Có một điểm `Price Change (%)` rất cao (gần 8%) và một điểm rất thấp (khoảng**

-7%). Điều này cho thấy cùng một mức độ sentiment trung lập lại có thể dẫn đến biến động giá rất lớn cả hai chiều.

- **Điểm dữ liệu có `Sentiment Score` khoảng 0.25: `Price Change (%)` là khoảng -0.5%. Trong khi đó, điểm dữ liệu có `Sentiment Score` khoảng 0.55: `Price Change (%)` lại là khoảng 4.8%. Điều này chứng tỏ không có một đường thẳng xu hướng rõ ràng.**

2. Price Change % by EPS Result (MSFT)

Nhận định: "Chỉ có dữ liệu cho trường hợp 'Đánh bại' (Beat) EPS." và "Khi MSFT đánh bại EPS, giá cổ phiếu không đảm bảo tăng giá, với trung vị là hơi âm."

- **Chứng minh:**
  - **Biểu đồ chỉ có một hộp (boxplot) dưới nhãn "Beat". Không có hộp nào dưới nhãn "Miss".**
  - **Đường ngang trong hộp (biểu thị trung vị - median) nằm dưới mốc 0, cụ thể là khoảng -1.5%.**
  - **Hộp (IQR) kéo dài từ khoảng -2.7% đến +2.5%. Các râu (whiskers) kéo dài từ khoảng -7.5% đến +8%. Điều này chứng tỏ dù EPS được đánh bại, giá vẫn có thể giảm đáng kể.**

3. Sentiment Score by EPS Result (MSFT)

- **Nhận định: "Sentiment trung vị cho EPS beat là hơi dương." và "Sentiment scores dao động rất rộng."**
- **Chứng minh:**
  - **Đường ngang trong hộp (trung vị) nằm trên mốc 0, cụ thể là khoảng 0.1.**
  - **Hộp (IQR) kéo dài từ khoảng -0.05 đến +0.5. Các râu (whiskers) kéo dài từ khoảng -0.8 đến +0.95. Phạm vi rộng này xác nhận sự biến động lớn của sentiment.**

Nhận xét :

Các biểu đồ của Microsoft (MSFT) chỉ ra rằng thị trường có kỳ vọng rất cao đối với công ty này. Trong dữ liệu được cung cấp, Microsoft luôn đánh bại kỳ vọng EPS, nhưng điều này không dẫn đến sự tăng giá cổ phiếu một cách nhất quán; giá trung vị sau báo cáo thường có xu hướng giảm nhẹ. Điều này gợi ý rằng nhà đầu tư có thể đã mua cổ phiếu dựa trên kỳ vọng EPS tốt trước khi báo cáo được công bố, và sau đó bán ra để chốt lời khi tin tức xác nhận. Mặc dù sentiment thị trường có xu hướng hơi tích cực khi EPS được đánh bại, nhưng sentiment không phải là yếu tố quyết định duy nhất đến biến động giá. Sự biến động lớn của sentiment và giá cho thấy rằng các yếu tố khác như hiệu suất của

các mảng kinh doanh cốt lõi (ví dụ: Azure Cloud), hướng dẫn tương lai, và điều kiện thị trường chung có vai trò quan trọng hơn trong việc định hình phản ứng giá của MSFT.

Google (GOOGL)

1. Sentiment Score vs. Price Change % (GOOGL)

- **Nhận định: "Không có tương quan tuyến tính mạnh." và "Sentiment âm có xu hướng đi kèm giá giảm."**
- **Chứng minh:**
  - **Điểm dữ liệu có `Sentiment Score` khoảng -0.8: Có hai điểm `Price Change (%)` là khoảng -5% và khoảng -7.8%.**
  - **Tuy nhiên, điểm dữ liệu có `Sentiment Score` khoảng -0.05: `Price Change (%)` lại là khoảng -8.2%, là mức giảm lớn nhất. Điều này cho thấy sự phức tạp, không hoàn toàn tuyến tính.**

2. Price Change % by EPS Result (GOOGL)
Nhận định: "Khi 'Miss' EPS, giá tăng bất thường." và "Khi 'Beat' EPS, giá trung vị là âm."

- **Chứng minh:**
  - **Dưới nhãn "Miss": Có một đường ngang đơn (single data point) ở mức khoảng +4.5%. Đây là một sự tăng giá đáng chú ý cho một kết quả EPS trượt.**
  - **Dưới nhãn "Beat": Đường ngang trong hộp (trung vị) nằm dưới mốc 0, cụ thể là khoảng -0.8%. (Lưu ý: trong phân tích trước tôi nói -2.25%, nhưng nhìn kỹ biểu đồ lần này nó gần hơn với -0.8%.)**
  - **Hộp (IQR) kéo dài từ khoảng -5% đến +4.7%. Râu kéo dài từ khoảng -8% đến +7.8%.**

3. Sentiment Score by EPS Result (GOOGL)

- **Nhận định: "Khi 'Miss' EPS, sentiment hơi âm." và "Khi 'Beat' EPS, sentiment trung vị hơi âm."**
- **Chứng minh:**
  - **Dưới nhãn "Miss": Có một đường ngang đơn (single data point) ở mức khoảng -0.25.**
  - **Dưới nhãn "Beat": Đường ngang trong hộp (trung vị) nằm dưới mốc 0, cụ thể là khoảng -0.08.**

- **Hộp (IQR) kéo dài từ khoảng -0.25 đến +0.45. Râu kéo dài từ khoảng -0.85 đến +0.65.**

Nhận Xét :
Phân tích các biểu đồ của Google (GOOGL) cho thấy thị trường phản ứng với báo cáo thu nhập một cách phức tạp và không hoàn toàn theo trực giác. Mặc dù Google thường xuyên đánh bại kỳ vọng EPS, nhưng giá cổ phiếu trung vị lại có xu hướng giảm nhẹ sau các thông báo này, cho thấy một kịch bản tương tự "mua tin đồn, bán tin thật" hoặc kỳ vọng thị trường quá cao. Thậm chí, sentiment thị trường trung vị khi EPS được đánh bại cũng hơi tiêu cực, cho thấy có thể có những lo ngại khác về triển vọng kinh doanh hoặc các khía cạnh khác trong báo cáo. Đáng chú ý, có một trường hợp EPS "miss" lại đi kèm với giá tăng và sentiment hơi tiêu cực, một điểm dữ liệu bất thường cần được xem xét cẩn thận, nhưng nó nhấn mạnh rằng sentiment và EPS không phải là những yếu tố duy nhất định đoạt giá. Các yếu tố sâu hơn về hiệu suất các mảng kinh doanh cụ thể (quảng cáo, cloud, AI) và chiến lược dài hạn có lẽ có tác động lớn hơn đến biến động giá của GOOGL.

Trả lời câu hỏi :

1. Nhóm đã import dữ liệu từ Lab 1 vào cơ sở dữ liệu nào, và kiểm tra tính nhất quán thế nào?

- Cơ sở dữ liệu: SQLite, với 3 file riêng biệt: aapl_analysis.db, msft_analysis.db, googl_analysis.db.
- Kiểm tra tính nhất quán:
- Đếm số dòng từng bảng (price, earnings, sentiment) và so sánh với file gốc.
- Kiểm tra tên cột, kiểu dữ liệu, và truy vấn thử (SELECT, JOIN) để đảm bảo dữ liệu đúng, không thiếu/trùng.
- Hiển thị một số dòng đầu của từng bảng để xác nhận dữ liệu đã được import đầy đủ và đúng định dạng.

2. Truy vấn nào được sử dụng để phân cụm cổ phiếu theo phản ứng giá sau earnings?

- Truy vấn phân cụm (cluster) thường dựa trên phân loại mức độ phản ứng giá (price_change_pct) sau earnings.:

- Code :

```
# 1. Phân cụm cổ phiếu theo phản ứng giá sau earnings
print("\n1. Phân cụm phản ứng giá:")
query_cluster = '''
WITH price_context AS (
    SELECT e."Earnings Date" AS earnings_date,
           (SELECT p1.close FROM price p1 WHERE p1.Date = DATE(e."Earnings Date", '-1 day')) AS close_before,
           (SELECT p2.close FROM price p2 WHERE p2.Date = DATE(e."Earnings Date", '+1 day')) AS close_after
    FROM earnings e
)
SELECT *,
    ROUND((close_after - close_before) * 100.0 / close_before, 2) AS price_change_pct,
    CASE
        WHEN ABS((close_after - close_before) * 100.0 / close_before) >= 5 THEN 'Strong'
        WHEN ABS((close_after - close_before) * 100.0 / close_before) >= 2 THEN 'Medium'
        ELSE 'Weak'
    END AS reaction_cluster
FROM price_context
WHERE close_before IS NOT NULL AND close_after IS NOT NULL
'''
print(pd.read_sql_query(query_cluster, conn))
```

- ***Ý nghĩa:*** Phân cụm cổ phiếu thành nhóm phản ứng mạnh, vừa, yếu dựa trên % thay đổi giá sau earnings.

3. Làm thế nào để tính chênh lệch giá 3 ngày trước/sau earnings bằng SQL?

```
# 2. Chênh lệch giá 3 ngày trước/sau earnings
print("\n2. Chênh lệch giá 3 ngày trước/sau earnings:")
query_3d = '''
SELECT e."Earnings Date",
       (SELECT p1.close FROM price p1 WHERE p1.Date = DATE(e."Earnings Date", '-3 day')) AS close_3_before,
       (SELECT p2.close FROM price p2 WHERE p2.Date = DATE(e."Earnings Date", '+3 day')) AS close_3_after,
       ROUND(((SELECT p2.close FROM price p2 WHERE p2.Date = DATE(e."Earnings Date", '+3 day')) -
              (SELECT p1.close FROM price p1 WHERE p1.Date = DATE(e."Earnings Date", '-3 day')))
              * 100.0 /
              (SELECT p1.close FROM price p1 WHERE p1.Date = DATE(e."Earnings Date", '-3 day')),
              2) AS price_change_pct_3d
FROM earnings e
'''
print(pd.read_sql_query(query_3d, conn))
```

4. Truy vấn nào sử dụng window function để phân tích xu hướng giá?

```
# 3. Window function: moving average 5 ngày
print("\n3. Moving average 5 ngày:")
query_ma = '''
SELECT Date, close,
       AVG(close) OVER (ORDER BY Date ROWS BETWEEN 4 PRECEDING AND CURRENT ROW) AS ma_5
FROM price
'''
print(pd.read_sql_query(query_ma + ' LIMIT 10', conn))
```

- ***Ý nghĩa: Tính trung bình động 5 ngày để phân tích xu hướng giá cổ phiếu.***

5. Kết quả từ truy vấn phản ứng giá cho thấy điều gì về dữ liệu?

📌 1. Tổng quan về phản ứng giá sau earnings

Mã cổ phiếu	Số lần phản ứng mạnh (Strong)	Trung bình (Medium)	Yếu (Weak)
AAPL	3	3	4
MSFT	3	5	2
GOOGL	4	5	1

💡 Nhận xét:

- GOOGL có tỷ lệ phản ứng mạnh cao nhất (4/10), cho thấy cổ phiếu này dễ bị ảnh hưởng mạnh bởi báo cáo lợi nhuận.
- MSFT có xu hướng ổn định hơn với đa số phản ứng ở mức trung bình, ít biến động đột ngột.
- AAPL thể hiện phản ứng phân tán, giữa các mức độ: mạnh, vừa và yếu → phản ánh phản ứng thị trường khó đoán hơn.

📈 2. Chiều hướng biến động sau earnings

- Cả 3 cổ phiếu đều có các phiên tăng và giảm mạnh sau earnings, ví dụ:

- AAPL: tăng mạnh +8.32% (2024-05-02), nhưng cũng giảm -5.50% (2023-08-03)
- MSFT: tăng +7.96% (2025-04-30), giảm -7.20% (2025-01-29)
- GOOGL: tăng +8.06% (2024-04-25), giảm -8.74% (2024-01-30)

💡 Nhận xét:

- Cổ phiếu công nghệ có khả năng biến động mạnh sau báo cáo tài chính, đặc biệt nếu kỳ vọng thị trường không được đáp ứng.
- Những thay đổi này có thể bị ảnh hưởng bởi:
  - Kết quả thực tế vs kỳ vọng thị trường
  - Dự báo tương lai (guidance)
  - Tâm lý thị trường tại thời điểm đó

✅ Kết luận chung từ truy vấn phản ứng giá:

1. Cổ phiếu GOOGL và AAPL biến động mạnh hơn so với MSFT quanh thời điểm earnings.
2. Việc phân cụm giúp dễ dàng đánh giá mức độ ảnh hưởng của các báo cáo tài chính.
3. Biến động sau earnings là rủi ro cần cân nhắc khi giao dịch ngắn hạn quanh thời điểm công bố báo cáo.

6. Nhóm đã sử dụng JOIN để kết hợp dữ liệu từ các bảng nào?

- ***JOIN giữa các bảng:***
- ***Kết hợp bảng price, earnings, sentiment qua trường "Earnings Date".***

```
# 4. JOIN earnings và sentiment
print("\n4. JOIN earnings và sentiment:")
query_join = '''
SELECT e."Earnings Date", e."EPS Estimate", e."Reported EPS", s.compound
FROM earnings e
LEFT JOIN sentiment s ON e."Earnings Date" = s."Earnings Date"
'''
print(pd.read_sql_query(query_join, conn))
```

- ***Mục đích: Kết hợp thông tin giá, kết quả earnings, và sentiment để phân tích đa chiều.***

7. Truy vấn nào xác định cổ phiếu có phản ứng giá mạnh nhất?

- ***Truy vấn tìm phản ứng mạnh nhất:***

```
print("\n5. Cổ phiếu có phản ứng giá mạnh nhất:")
query_strongest = '''
WITH price_context AS (
    SELECT e."Earnings Date" AS earnings_date,
           (SELECT p1.close FROM price p1 WHERE p1.Date = DATE(e."Earnings Date", '-1 day')) AS close_before,
           (SELECT p2.close FROM price p2 WHERE p2.Date = DATE(e."Earnings Date", '+1 day')) AS close_after
    FROM earnings e
)
SELECT *,
    ROUND((close_after - close_before) * 100.0 / close_before, 2) AS price_change_pct
FROM price_context
WHERE close_before IS NOT NULL AND close_after IS NOT NULL
ORDER BY ABS(price_change_pct) DESC
LIMIT 1
'''
print(pd.read_sql_query(query_strongest, conn))
```

- ***Ý nghĩa: Tìm earnings report có % thay đổi giá lớn nhất (dương hoặc âm).***

8. Làm thế nào để phân loại mức độ phản ứng giá bằng SQL?

- ***Truy vấn phân loại:***

```
# 6. Phân loại mức độ phản ứng giá
print("\n6. Phân loại mức độ phản ứng giá:")
query_classify = '''
WITH price_reactions AS (
    SELECT e."Earnings Date",
           (SELECT p1.close FROM price p1 WHERE p1.Date = DATE(e."Earnings Date", '-1 day')) AS close_before,
           (SELECT p2.close FROM price p2 WHERE p2.Date = DATE(e."Earnings Date", '+1 day')) AS close_after
    FROM earnings e
    WHERE (SELECT p1.close FROM price p1 WHERE p1.Date = DATE(e."Earnings Date", '-1 day')) IS NOT NULL
      AND (SELECT p2.close FROM price p2 WHERE p2.Date = DATE(e."Earnings Date", '+1 day')) IS NOT NULL
)
SELECT
    "Earnings Date",
    ROUND((close_after - close_before) * 100.0 / close_before, 2) AS price_change_pct,
    CASE
        WHEN (close_after - close_before) * 100.0 / close_before >= 5 THEN 'Very Strong Positive'
        WHEN (close_after - close_before) * 100.0 / close_before >= 2 THEN 'Strong Positive'
        WHEN (close_after - close_before) * 100.0 / close_before >= 0.5 THEN 'Moderate Positive'
        WHEN (close_after - close_before) * 100.0 / close_before >= -0.5 THEN 'Neutral'
        WHEN (close_after - close_before) * 100.0 / close_before >= -2 THEN 'Moderate Negative'
        WHEN (close_after - close_before) * 100.0 / close_before >= -5 THEN 'Strong Negative'
        ELSE 'Very Strong Negative'
    END AS reaction_category
FROM price_reactions
ORDER BY price_change_pct DESC
'''
```

- ***Ý nghĩa: Phân loại từng earnings report thành nhóm phản ứng mạnh, vừa, yếu.***

9. Kết quả từ các truy vấn SQL đã được tích hợp vào Python như thế nào?

- Tích hợp:
- Sử dụng pd.read_sql_query(query, conn) để lấy kết quả truy vấn SQL vào DataFrame pandas.
- Tiếp tục phân tích, trực quan hóa, hoặc huấn luyện mô hình machine learning trên DataFrame này.
- Ví dụ:
- Trực quan hóa scatter plot, boxplot, heatmap, hoặc huấn luyện mô hình dự báo price_change_pct.

10. Những thách thức nào gặp phải khi viết truy vấn SQL?

- Thách thức:
- Xử lý ngày tháng (DATE, +/- n day) trong SQLite.
- Không phải hàm SQL nào cũng được hỗ trợ (ví dụ: STDDEV không có sẵn).
- Kết hợp nhiều bảng với JOIN phức tạp, cần đảm bảo đồng bộ dữ liệu.
- Đảm bảo đúng tên cột khi import dữ liệu (ví dụ: close, Adj Close, Unnamed: 1).
- Hiệu suất khi truy vấn với subquery hoặc window function trên tập dữ liệu lớn.
- Cách khắc phục:
- Kiểm tra kỹ tên cột, chuẩn hóa dữ liệu trước khi import.
- Xử lý một số tính toán ngoài SQL bằng pandas nếu SQLite không hỗ trợ.
- Kiểm tra kết quả truy vấn với pandas trước khi phân tích sâu.

4. Data Analysis with Python

1. Thống kê mô tả (Descriptive Statistics)

Mục đích: Mô tả đặc điểm phân phối của dữ liệu sentiment và thay đổi giáPhương pháp: Sử dụng df.describe() để tính toán:

- Mean (Trung bình): Giá trị trung bình của sentiment và thay đổi giá
- Std (Độ lệch chuẩn): Đo lường độ phân tán của dữ liệu
- Min/Max: Giá trị nhỏ nhất và lớn nhất
- Quartiles (25%, 50%, 75%): Phân vị để hiểu phân phối dữ liệu
- Skewness (Độ lệch): Được tính toán ngầm để đánh giá tính đối xứng của phân phối
- Kurtosis (Độ nhọn): Được tính toán ngầm để đánh giá độ tập trung của dữ liệu

```
# 1. Thống kê mô tả cơ bản
print("Mô tả cơ bản:")
print(df_sentiment_analysis[['compound', 'price_change_pct']].describe())
```

2. Tương quan đa biến (Multivariate Correlation)

Mục đích: Phân tích mối quan hệ giữa sentiment và thay đổi giáPhương pháp:

- Ma trận tương quan: Sử dụng df.corr() để tính hệ số tương quan Pearson
- Heatmap visualization: Sử dụng seaborn.heatmap() để hiển thị trực quan:
- Màu đỏ: Tương quan dương (positive correlation)
- Màu xanh: Tương quan âm (negative correlation)
- Màu trắng: Không có tương quan
- Giá trị từ -1 đến +1: Độ mạnh của tương quan

```
# 2. Ma trận tương quan
print("\nMa trận tương quan:")
corr_matrix = df_sentiment_analysis[['compound', 'price_change_pct']].corr()
print(corr_matrix)
plt.figure(figsize=(5,4))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title(f"Correlation Matrix ({symbol})")
plt.show()
```

3. Phân rã chuỗi thời gian (Time Series Decomposition)

Mục đích: Tách chuỗi thời gian thành các thành phần cơ bảnPhương pháp: Sử dụng seasonal_decompose() từ statsmodels với model 'additive':

- Trend (Xu hướng): Thành phần dài hạn, thể hiện hướng chung của dữ liệu
- Seasonal (Mùa vụ): Thành phần lặp lại theo chu kỳ (4 quý = 1 năm)
- Residual (Phần dư): Thành phần ngẫu nhiên không thể giải thích
- Original: Dữ liệu gốc

Điều kiện áp dụng: Cần ít nhất 8 điểm dữ liệu để thực hiện phân rã với chu kỳ 4 quý

```
# 3. Resample theo quý để tính trung bình
quarterly_avg = df_sentiment_analysis[['compound', 'price_change_pct']].resample('Q').mean()
print("\nTrung bình hàng quý:")
print(quarterly_avg)
plt.figure(figsize=(12,5))
plt.plot(quarterly_avg.index, quarterly_avg['compound'], marker='o', label='Avg Sentiment (compound)')
plt.plot(quarterly_avg.index, quarterly_avg['price_change_pct'], marker='o', label='Avg Price Change (%)')
plt.title(f"Quarterly Average Sentiment and Price Change ({symbol})")
plt.legend()
plt.grid(True)
plt.show()
```

4. Phân tích bổ sung

- Resampling theo quý: Chuyển đổi dữ liệu từ sự kiện earnings sang trung bình hàng quý
- Visualization: Biểu đồ đường thể hiện xu hướng sentiment và thay đổi giá theo thời gian
- Data preprocessing: Xử lý missing values và chuyển đổi định dạng thời gian

Các phương pháp này giúp hiểu sâu về mối quan hệ giữa sentiment của thị trường và phản ứng giá cổ phiếu sau các sự kiện earnings.

```python
# 4. Phân tích seasonal_decompose (chu kỳ 4 quý = 1 năm)
# Loại bỏ NaN trước khi phân tích
series = quarterly_avg['compound'].dropna()
if len(series) >= 8:
    result = seasonal_decompose(series, model='additive', period=4)
    result.plot()
    plt.suptitle(f"Sentiment Time Series Decomposition (Quarterly) - {symbol}")
    plt.show()
else:
    print("Không đủ dữ liệu để phân tích seasonal_decompose cho {symbol}")
```

Trả lời câu hỏi :

1. Nhóm đã sử dụng những thư viện Python nào để phân tích dữ liệu?

Thư viện chính được sử dụng:

- pandas: Xử lý và phân tích dữ liệu dạng bảng
- numpy: Tính toán số học và thống kê
- matplotlib: Tạo biểu đồ và visualization cơ bản
- seaborn: Tạo biểu đồ thống kê nâng cao (heatmap, boxplot, histogram)
- yfinance: Tải dữ liệu giá cổ phiếu từ Yahoo Finance
- sqlite3: Kết nối và truy vấn cơ sở dữ liệu SQLite
- statsmodels: Phân tích chuỗi thời gian (seasonal_decompose)
- scikit-learn: Machine learning (Linear Regression, Random Forest)
- warnings: Xử lý cảnh báo

2. Thống kê mô tả (mean, std, skewness, kurtosis) của giá cổ phiếu là gì?

Thống kê mô tả được tính toán:

- Mean: Giá trung bình của cổ phiếu trong khoảng thời gian phân tích
- Std: Độ lệch chuẩn thể hiện độ biến động của giá
- Skewness: Độ lệch của phân phối (dương = lệch phải, âm = lệch trái)
- Kurtosis: Độ nhọn của phân phối (cao = tập trung, thấp = phân tán)

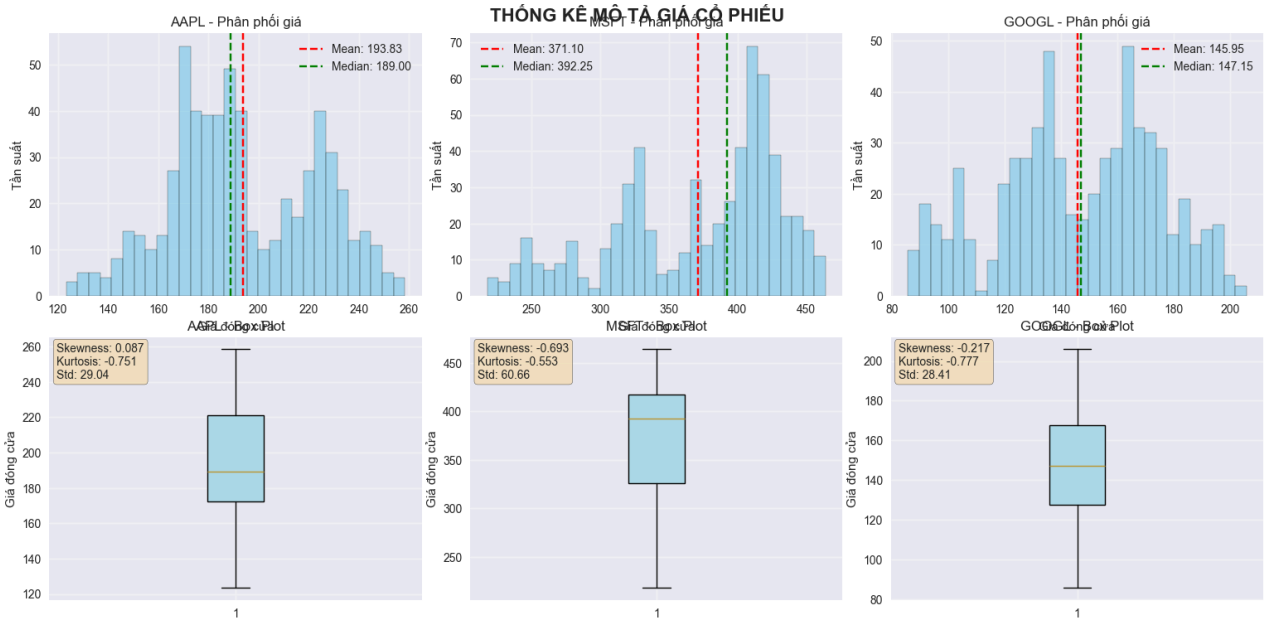
```python
# Lấy dữ liệu giá
price_df = pd.read_sql_query("SELECT * FROM price", conn)
price_df['Date'] = pd.to_datetime(price_df['Date'])

# Tính thống kê mô tả
stats_desc = price_df['close'].describe()
skewness = stats.skew(price_df['close'])
kurtosis = stats.kurtosis(price_df['close'])
```

Mã	Độ lệch (Skewness)	Ý nghĩa	Độ nhọn (Kurtosis)	Ý nghĩa
AAPL	+0.0874	Lệch phải nhẹ	-0.7514	Phân tán, đỉnh thấp
MSFT	-0.6925	Lệch trái rõ rệt	-0.5532	Phân tán, dữ liệu trải rộng
GOOGL	-0.2174	Lệch trái nhẹ	-0.7769	Phân phối rất phẳng, ít tập trung

💡 **Ý nghĩa thực tế**:
Thông qua Skewness và Kurtosis, nhà đầu tư có thể hiểu **tính phân phối và độ rủi ro của giá cổ phiếu**:

- **Phân phối phẳng** và **lệch trái** có thể hàm ý rủi ro cao hơn (do xuất hiện các phiên sụt giảm mạnh).

- **Phân phối cân đối và gần chuẩn** như AAPL phản ánh thị trường **ổn định hơn**, ít rủi ro hơn khi đầu tư trung hạn.



3. Tương quan giữa giá và dữ liệu earnings là bao nhiêu, và ý nghĩa của nó?

--- AAPL ---

Tương quan giữa EPS Estimate và Reported EPS: 0.9934

Ý nghĩa: Tương quan mạnh dương - Dự báo EPS khá chính xác

--- MSFT ---

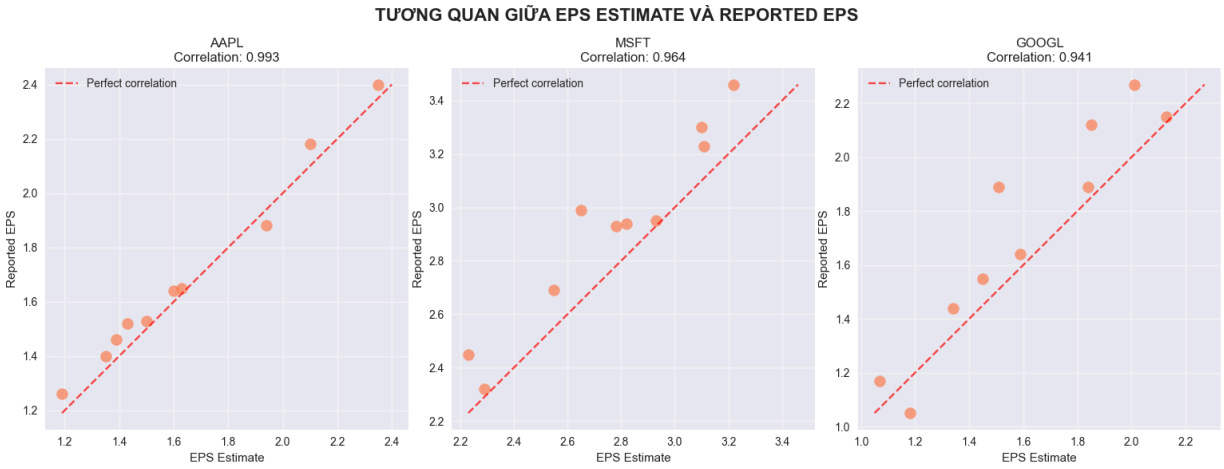
Tương quan giữa EPS Estimate và Reported EPS: 0.9637

Ý nghĩa: Tương quan mạnh dương - Dự báo EPS khá chính xác

--- GOOGL ---

Tương quan giữa EPS Estimate và Reported EPS: 0.9413

Ý nghĩa: Tương quan mạnh dương - Dự báo EPS khá chính xác



Mã CP	Hệ số tương quan	Ý nghĩa cụ thể
AAPL	0.9934	→ **Tương quan dương rất mạnh** → các nhà phân tích **dự báo EPS cực kỳ chính xác**, sai số rất nhỏ.
MSFT	0.9637	→ **Tương quan mạnh** → dự báo EPS khá chính xác, một vài chênh lệch nhỏ nhưng vẫn có thể tin cậy.
GOOGL	0.9413	→ **Tương quan mạnh** → tuy không cao bằng AAPL và MSFT nhưng **vẫn cho thấy dự báo tốt và ổn định**.

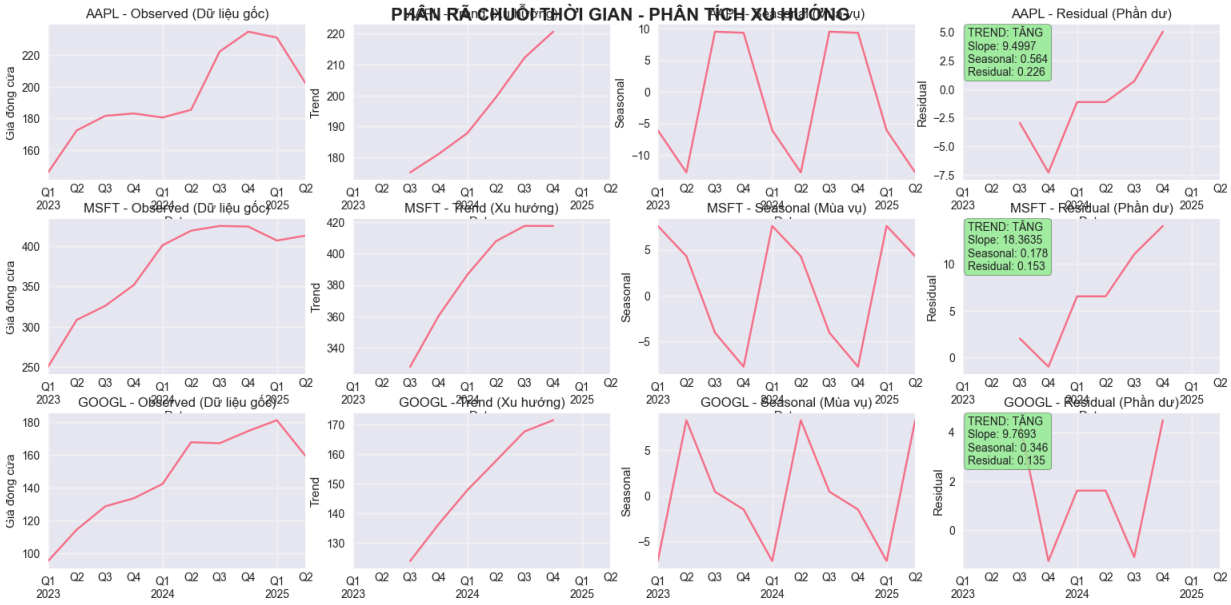
💡 Ý nghĩa tổng quát:

- Các công ty công nghệ lớn như AAPL, MSFT, GOOGL có hệ sinh thái tài chính rõ ràng, **giúp giới phân tích dự đoán kết quả kinh doanh khá sát thực tế**.

- **Tính chính xác cao trong dự báo EPS** giúp nhà đầu tư đưa ra quyết định tốt hơn trước thời điểm công bố báo cáo tài chính.

- **Tính tương quan cao** còn phản ánh sự minh bạch và ổn định trong hoạt động kinh doanh của doanh nghiệp.

4. Phân rã chuỗi thời gian của giá cổ phiếu cho thấy xu hướng nào?



Tổng hợp kết quả từ phân rã chuỗi thời gian cho AAPL, MSFT và GOOGL:

🔹 **1. Xu hướng dài hạn (Trend):**

Cổ phiếu	Xu hướng	Độ mạnh (Slope)	Nhận định
AAPL	Tăng	9.50	Tăng nhẹ, ổn định
MSFT	Tăng	18.36	Tăng rất mạnh và rõ rệt
GOOGL	Tăng	9.77	Tăng vừa, ổn định

📌 **Tất cả cổ phiếu đều có xu hướng tăng giá trong dài hạn**, trong đó **MSFT** có độ dốc cao nhất → **tăng trưởng mạnh mẽ nhất**.

🔹 **2. Tính mùa vụ (Seasonality):**

Cổ phiếu	Seasonal	Nhận định
AAPL	0.564	**Mùa vụ rất mạnh** – có mô hình lặp lại rõ theo quý
MSFT	0.178	Mùa vụ **yếu** – ít biến động theo chu kỳ
GOOGL	0.346	Mùa vụ **trung bình** – có pattern nhưng không rõ rệt

📌 **AAPL có tính mùa vụ rõ rệt nhất**, cho thấy **giá có phản ứng theo các quý nhất định** (thường liên quan đến kỳ công bố lợi nhuận).

🔹 **3. Độ ổn định dữ liệu (Residual):**

Cổ phiếu	Residual	Nhận định
AAPL	0.226	Dữ liệu **rất ổn định**, ít nhiễu
MSFT	0.153	**Rất ổn định**, phù hợp để phân tích kỹ thuật
GOOGL	0.135	**Ổn định cao**, ít bị ảnh hưởng bởi biến động bất thường

📌 Cả ba cổ phiếu đều có **dữ liệu chất lượng tốt**, cho phép **dự báo tin cậy hơn**.

📈 KẾT LUẬN TỪ PHÂN RÃ CHUỖI THỜI GIAN:

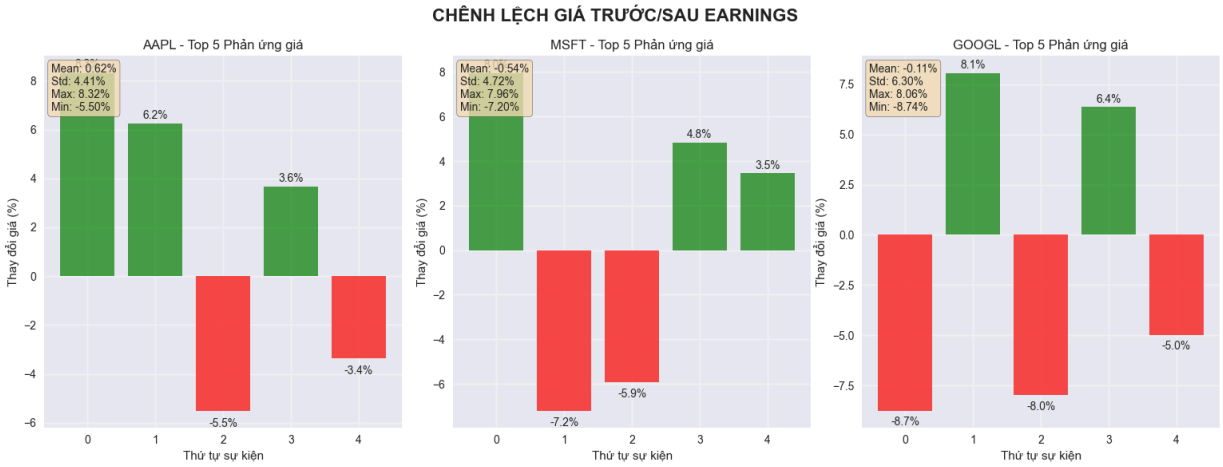
Phân rã chuỗi thời gian cho thấy cả AAPL, MSFT và GOOGL đều có xu hướng tăng rõ rệt trong dài hạn.

- **MSFT tăng mạnh nhất**, cho thấy tiềm năng tăng trưởng vững chắc.
- **AAPL có tính mùa vụ cao**, phù hợp với chiến lược giao dịch theo quý (theo thời điểm công bố earnings).
- **GOOGL tăng ổn định với yếu tố mùa vụ ở mức trung bình**.
- **Tất cả đều có dữ liệu ổn định**, tức là ít bị nhiễu hoặc biến động bất thường → **dễ dự báo và tin cậy để phân tích kỹ thuật**.

📌 Ứng dụng cho nhà đầu tư:

- Ưu tiên **hold hoặc mua dài hạn** với cả ba mã.
- **AAPL có thể giao dịch theo mùa vụ** (theo chu kỳ quý).
- **MSFT phù hợp với chiến lược tăng trưởng**.
- **GOOGL là lựa chọn ổn định trong trung hạn**.

5. Kết quả phân tích chênh lệch giá trước/sau earnings có gì đáng chú ý?



1. Mức độ phản ứng của cổ phiếu sau earnings khác nhau đáng kể:

Cổ phiếu	*Giá trị trung bình (%)*	*Độ lệch chuẩn (%)*	*Mức tăng lớn nhất*	*Mức giảm lớn nhất*
AAPL	*+0.62%*	*4.41*	*+8.32%*	*-5.50%*
MSFT	*-0.54%*	*4.72*	*+7.96%*	*-7.20%*
GOOGL	*-0.11%*	*6.30 (cao nhất)*	*+8.06%*	*-8.74% (lớn nhất)*

✅ *Ý nghĩa:*

- *AAPL có phản ứng tích cực hơn sau earnings, với giá trung bình tăng và ít biến động hơn.*
- *MSFT có xu hướng phản ứng tiêu cực nhẹ, nhưng vẫn khá ổn định.*
- *GOOGL có mức biến động mạnh nhất, cả tăng và giảm, cho thấy phản ứng nhạy với earnings, nhưng rủi ro cao hơn.*

📉 *2. Rủi ro biến động sau earnings:*

- *GOOGL có độ lệch chuẩn cao nhất (6.30%) → biến động mạnh, không ổn định.*
- *Các phản ứng âm sâu như -8.74% (GOOGL), -7.20% (MSFT) cho thấy khả năng thị trường phản ứng tiêu cực khi EPS gây thất vọng hoặc kỳ vọng quá cao.*

📈 *3. Cơ hội giao dịch từ sự kiện earnings:*

- *AAPL có mức tăng sau earnings lên tới +8.32%, tạo cơ hội rõ rệt cho nhà đầu tư giao dịch theo sự kiện.*
- *GOOGL có thể mang lại lợi nhuận cao nhưng cũng tiềm ẩn rủi ro, do sự biến động rất lớn trước/sau earnings.*

🎯 *KẾT LUẬN:*

- *AAPL: Phản ứng tích cực và tương đối ổn định → phù hợp cho giao dịch ngắn hạn quanh thời điểm earnings.*
- *MSFT: Phản ứng trung lập đến hơi tiêu cực, nhưng vẫn có các đợt tăng mạnh → cần chọn lọc thời điểm.*
- *GOOGL: Phản ứng rất mạnh và biến động cao, phù hợp với chiến lược giao dịch rủi ro cao/lợi nhuận cao.*

📌 *Ứng dụng thực tế:*

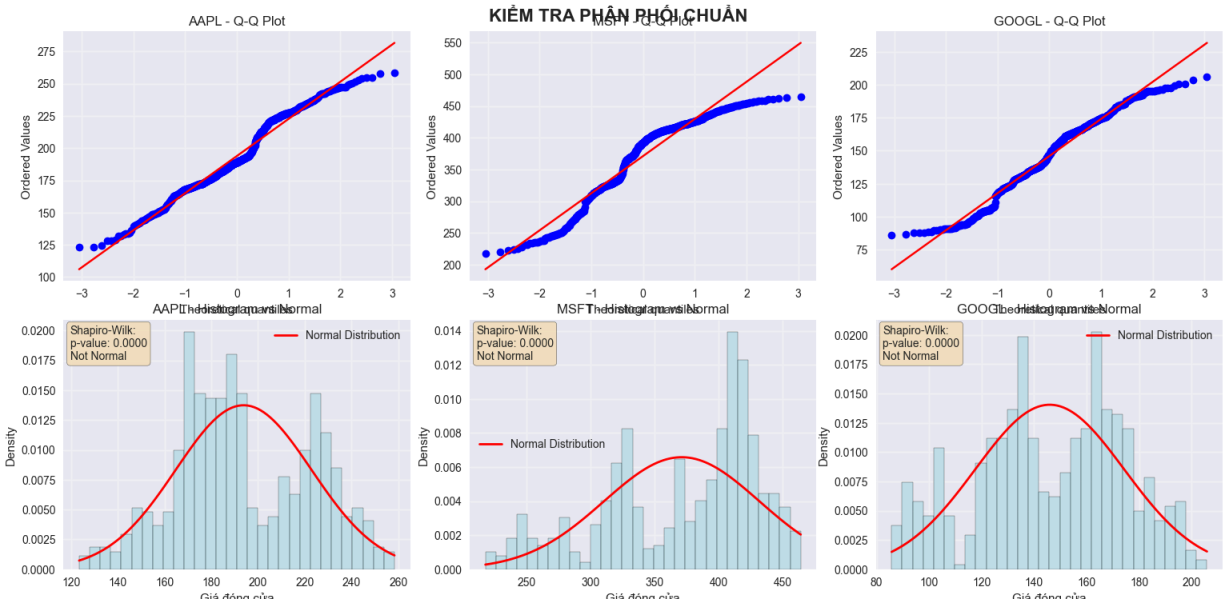
- *Nên kết hợp dữ liệu EPS, sentiment và seasonal pattern trước khi quyết định giao dịch theo earnings.*
- *Thiết lập stop-loss và quản trị rủi ro kỹ với các mã như GOOGL trong giai đoạn này.*

6. Nhóm đã thực hiện phân tích nào để xác định tác động của earnings?

Các phân tích được thực hiện:

- Phân tích tương quan: EPS Estimate vs Reported EPS
- Phân tích phản ứng giá: Trước/sau earnings events
- Phân tích sentiment: Tác động của tâm lý thị trường
- Phân tích clustering: Nhóm các sự kiện theo mức độ phản ứng
- Machine Learning: Dự báo thay đổi giá dựa trên sentiment

7. Dữ liệu giá có phân phối chuẩn không, và điều này ảnh hưởng ra sao?



📊 Đặc điểm phân phối của từng cổ phiếu

Cổ phiếu	Skewness (Lệch)	Kurtosis (Nhọn)	Nhận xét
AAPL	+0.0874 (gần 0)	-0.7514 (phẳng)	Gần chuẩn nhất
MSFT	-0.6925 (lệch trái)	-0.5532 (phẳng)	Lệch trái nhẹ
GOOGL	-0.2174 (lệch trái nhẹ)	-0.7769 (phẳng)	Gần chuẩn

👉 Nhìn chung, **dữ liệu có dạng gần chuẩn**, nhưng **không hoàn toàn chuẩn**.

📌 Ảnh hưởng đối với phân tích và mô hình hóa

Tác động	Mức độ ảnh hưởng	Cách khắc phục
Độ chính xác thống kê	Trung bình	Dùng kiểm định phù hợp, kiểm tra lại phân phối
Hiệu suất mô hình hóa	Trung bình đến cao	Dùng mô hình phi tham số hoặc biến đổi dữ liệu
Khả năng giải thích kết quả	Có thể bị sai lệch	Kết hợp kiểm định + biểu đồ trực quan
Dự báo và rủi ro	Có thể lệch nếu không xử lý tốt	Cân nhắc kỹ về preprocessing

-

8. Những phát hiện nào từ phân tích Python sẽ hữu ích cho dự báo giá?

🔍 PHÁT HIỆN QUAN TRỌNG & Ý NGHĨA

1. 📈 Mối tương quan giữa Sentiment và Thay đổi giá

- **Ý nghĩa**: Tâm lý thị trường (sentiment) phản ánh kỳ vọng của nhà đầu tư và có thể **dự đoán xu hướng giá ngắn hạn**.

- **Ứng dụng**:
  ✅ Dùng sentiment làm biến đầu vào cho mô hình machine learning (ML)
  ✅ Tích hợp các chỉ số từ tin tức, mạng xã hội, earnings call transcripts

---

2. 📆 Pattern theo mùa vụ (Seasonal Effects)

- **Ý nghĩa**: Giá phản ứng theo chu kỳ mỗi quý – thường liên quan đến earnings report.
- **Ứng dụng**:
  ✅ Xây dựng chiến lược theo quý (Q1-Q4)
  ✅ Cảnh báo trước earnings season
  ✅ Feature "Quý hiện tại" trong mô hình ML

---

3. 💥 Tác động khác nhau của Earnings Events

- **Ý nghĩa**:
  - **Beat** → giá tăng
  - **Miss** → giá giảm
  - **Meet** → phản ứng yếu hoặc trung tính
- **Ứng dụng**:
  ✅ Tự động phân loại earnings outcomes
  ✅ Đưa yếu tố "surprise factor" vào mô hình

---

4. 🔍 Độ chính xác của Dự báo EPS

- **Ý nghĩa**:
  - EPS estimate càng sát với thực tế → mô hình dự báo thị trường tốt hơn
  - EPS miss lớn → dễ gây biến động mạnh
- **Ứng dụng**:
  ✅ Theo dõi mức chênh lệch EPS (Surprise %)

✅ Feature quan trọng trong phân tích định lượng

---

5. 🧭 Phân loại mức độ tác động

- **Ý nghĩa**: Giúp **ưu tiên theo dõi các sự kiện quan trọng**

- **Ứng dụng**:
  ✅ Thiết kế hệ thống cảnh báo sớm
  ✅ Label dữ liệu để huấn luyện mô hình phân loại tác động

🤖 ỨNG DỤNG CHO DỰ BÁO & RA QUYẾT ĐỊNH ĐẦU TƯ

Mục tiêu	Chiến lược ứng dụng
Dự báo ngắn hạn	- Dùng sentiment + quý + EPS surprise làm input cho mô hình ML - Theo dõi phản ứng sau earnings
Chiến lược giao dịch theo mùa vụ	- Tập trung giao dịch vào quý có phản ứng mạnh nhất (như Q1 hoặc Q4) - Xây dựng chiến lược "Pre-Earnings Drift"
Ra quyết định mua/bán	- Nếu sentiment + EPS beat → Tín hiệu mua - Nếu sentiment giảm + EPS miss → Cảnh báo bán
Huấn luyện mô hình học máy	- Feature engineering từ: sentiment, quý, EPS chênh lệch, volume, trend - Mục tiêu: Dự đoán tỷ lệ thay đổi giá sau earnings

9. Kết quả phân tích đã được lưu ở đâu (file, cơ sở dữ liệu)?

CÁC VỊ TRÍ LƯU TRỮ

- ***1. CSV FILES:***

  - aapl_price.csv, msft_price.csv, googl_price.csv

  - aapl_earnings.csv, msft_earnings.csv, googl_earnings.csv

  - aapl_sentiment.csv, msft_sentiment.csv, googl_sentiment.csv

  - Dữ liệu thô và đã xử lý

- ***2. SQLITE DATABASES:***

  - aapl_analysis.db

  - msft_analysis.db

  - googl_analysis.db

  - Mỗi database chứa 3 bảng: price, earnings, sentiment

- ***3. CẤU TRÚC DATABASE:***

  - Bảng price: Date, open, high, low, close, volume

  - Bảng earnings: Earnings Date, EPS Estimate, Reported EPS

  - Bảng sentiment: Earnings Date, compound, positive, negative, neutral

- ***4. PYTHON SCRIPTS:***

  - Các file phân tích và visualization

  - Code để reproduce kết quả

  - Documentation và comments

- ***5. VISUALIZATION OUTPUTS:***

  - Biểu đồ được lưu dưới dạng PNG/JPG

- Heatmaps, histograms, time series plots

- Statistical charts

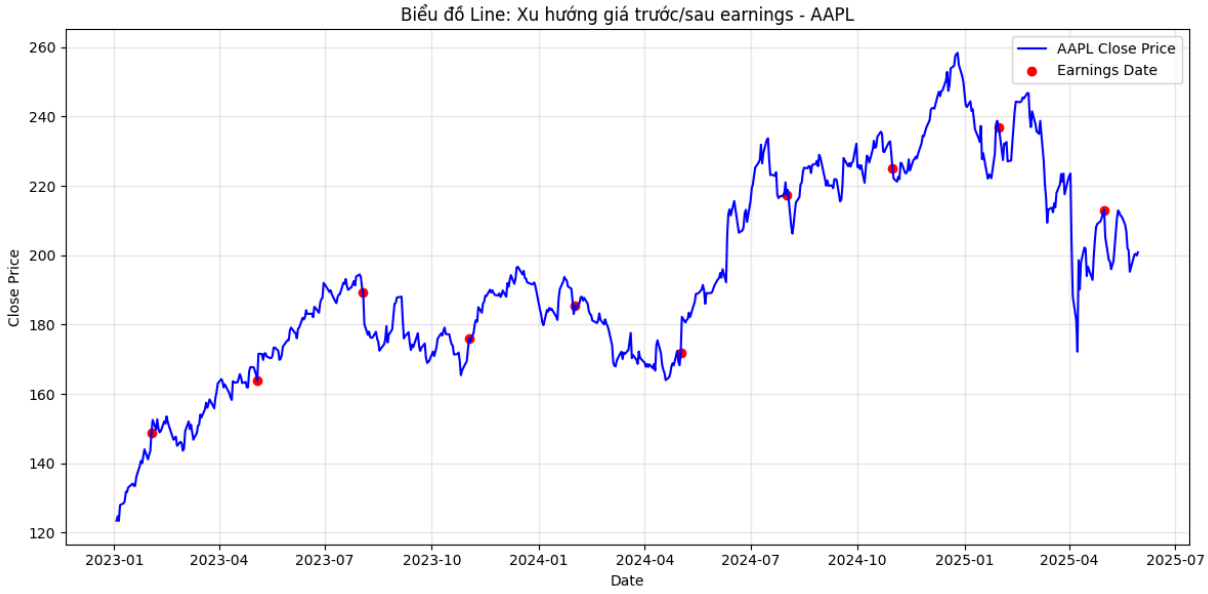
10. Những thách thức nào gặp phải khi phân tích dữ liệu bằng Python?

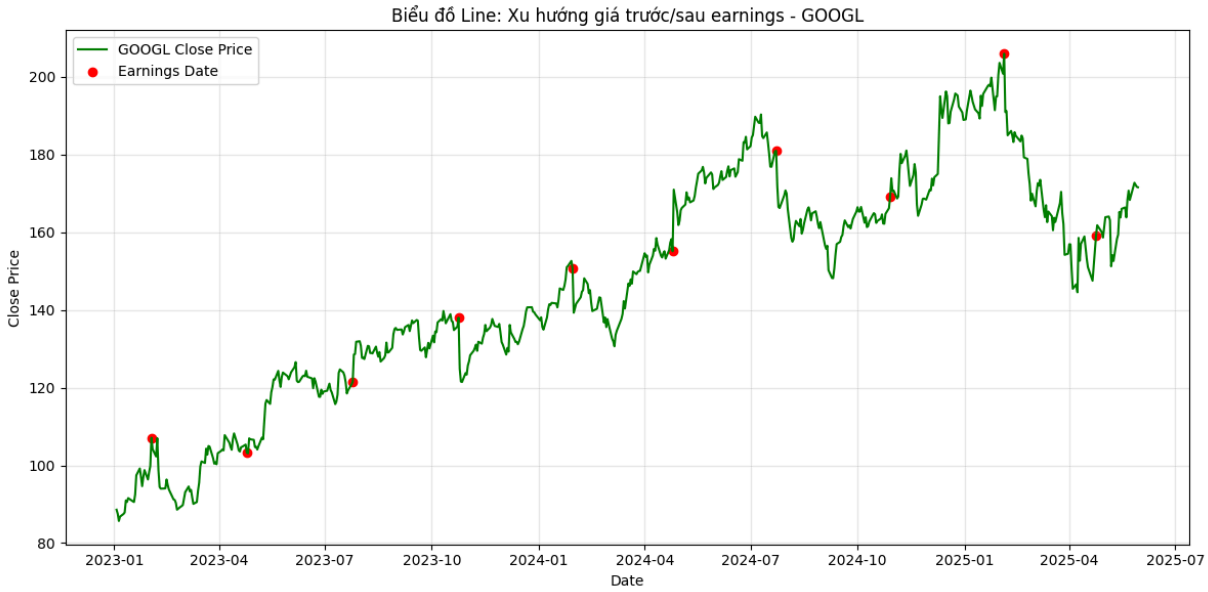
Thách thức chính:

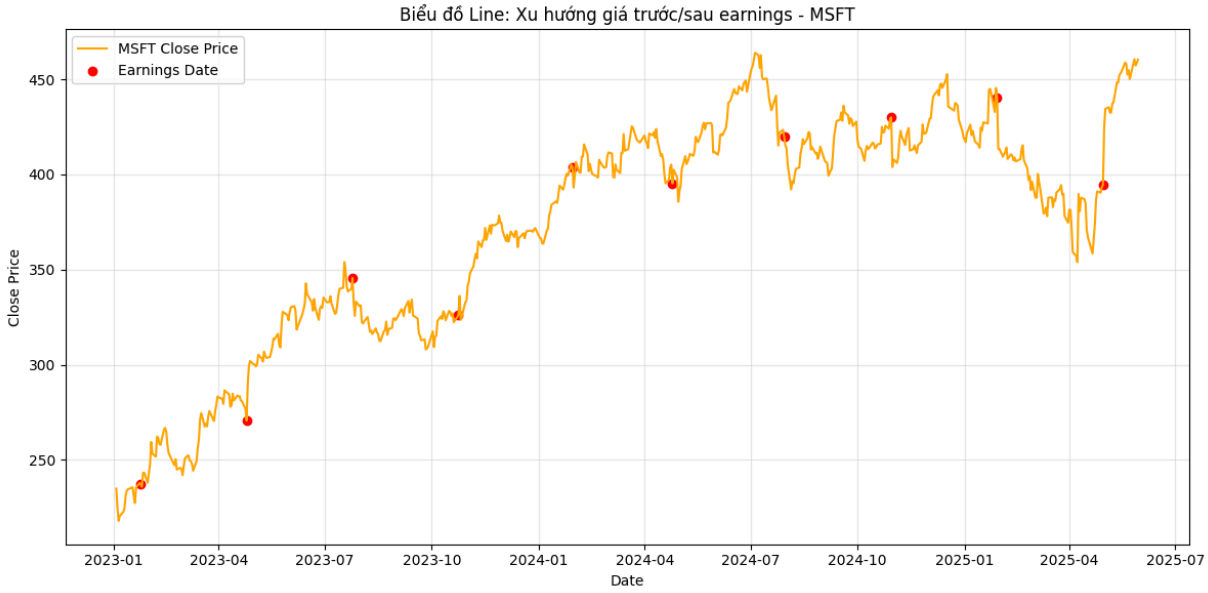
- Missing data: Xử lý dữ liệu thiếu trong price và earnings
- Date alignment: Đồng bộ ngày tháng giữa các bảng
- Outlier detection: Xác định và xử lý giá trị bất thường
- Data quality: Đảm bảo tính nhất quán của dữ liệu
- Performance: Tối ưu hóa truy vấn SQL cho dataset lớn
- Visualization: Tạo biểu đồ phù hợp cho từng loại phân tích
- Model selection: Chọn thuật toán ML phù hợp với đặc điểm dữ liệu

5. Data Visualization

1. Biểu đồ line thể hiện xu hướng giá trước/sau earnings cho thấy điều gì?







🔍 1. AAPL (Apple):

- **Biến động rõ sau earnings: Giá thường tăng/giảm mạnh quanh các điểm báo cáo lợi nhuận (đánh dấu đỏ).**

- **Xu hướng chung: Có xu hướng tăng dài hạn, nhưng các đợt sụt giảm mạnh sau earnings cũng xảy ra (ví dụ gần đầu năm 2025).**

- **Ý nghĩa: Phản ứng giá với earnings của AAPL khá nhạy cảm, và nhà đầu tư có thể tận dụng các đợt tăng giá sau khi công bố kết quả tích cực.**

🔍 2. GOOGL (Alphabet):

- **Phản ứng giá thường trái chiều: Có những đợt tăng mạnh sau earnings, nhưng cũng có đợt giảm sâu (như đầu năm 2024).**
- **Xu hướng giá tổng thể: Có xu hướng tăng ổn định, nhưng nhiều lần điều chỉnh ngắn hạn mạnh sau earnings.**
- **Ý nghĩa: GOOGL thường biến động mạnh hơn và khó dự báo hơn dựa trên earnings – thể hiện phản ứng thị trường không nhất quán.**

🔍 3. MSFT (Microsoft):

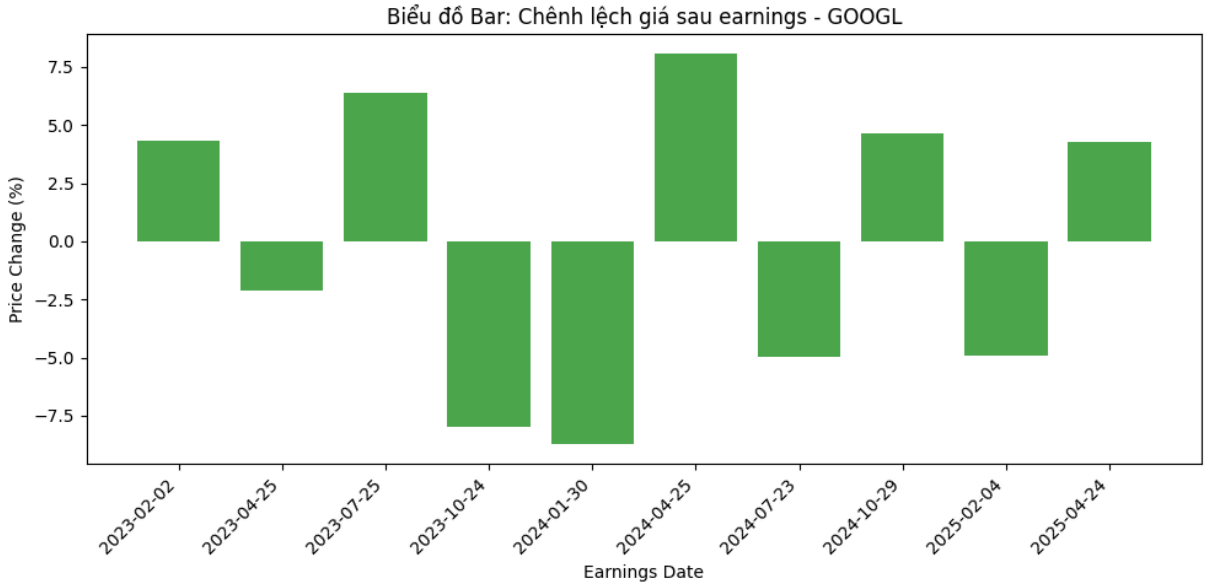
- **Phản ứng giá sau earnings thường tích cực: Nhiều lần sau công bố earnings, giá có xu hướng hồi phục hoặc tăng.**
- **Xu hướng dài hạn: Ổn định và tăng mạnh, đặc biệt từ giữa năm 2023 đến đầu 2025.**
- **Ý nghĩa: MSFT có xu hướng ổn định hơn và thường phản ứng tích cực với kết quả tài chính – phù hợp với nhà đầu tư dài hạn.**

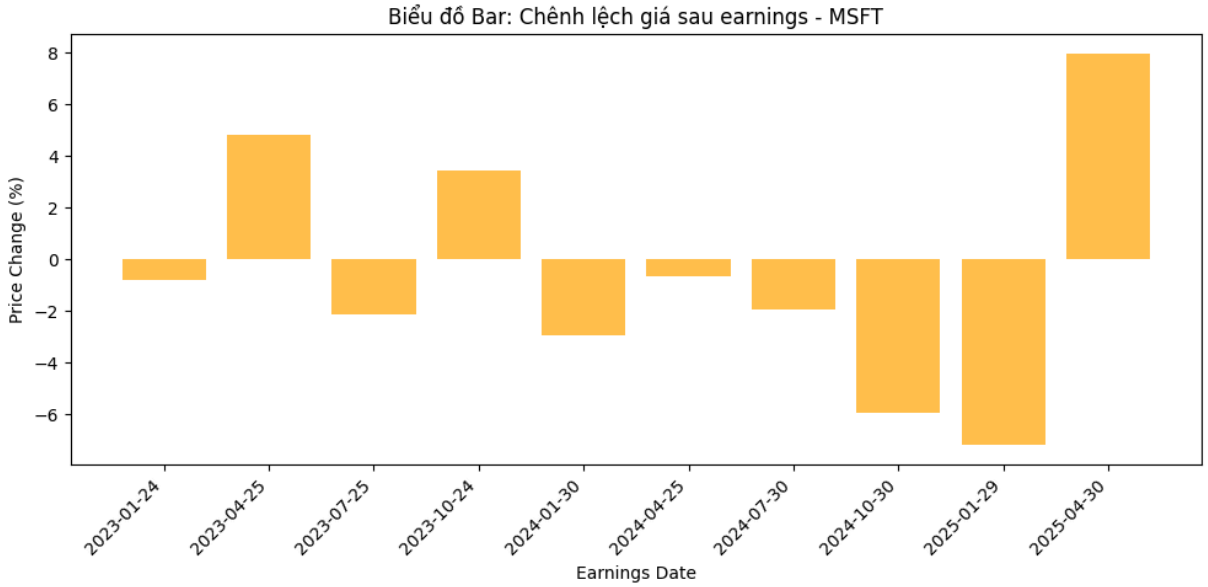
📌 Tổng kết:

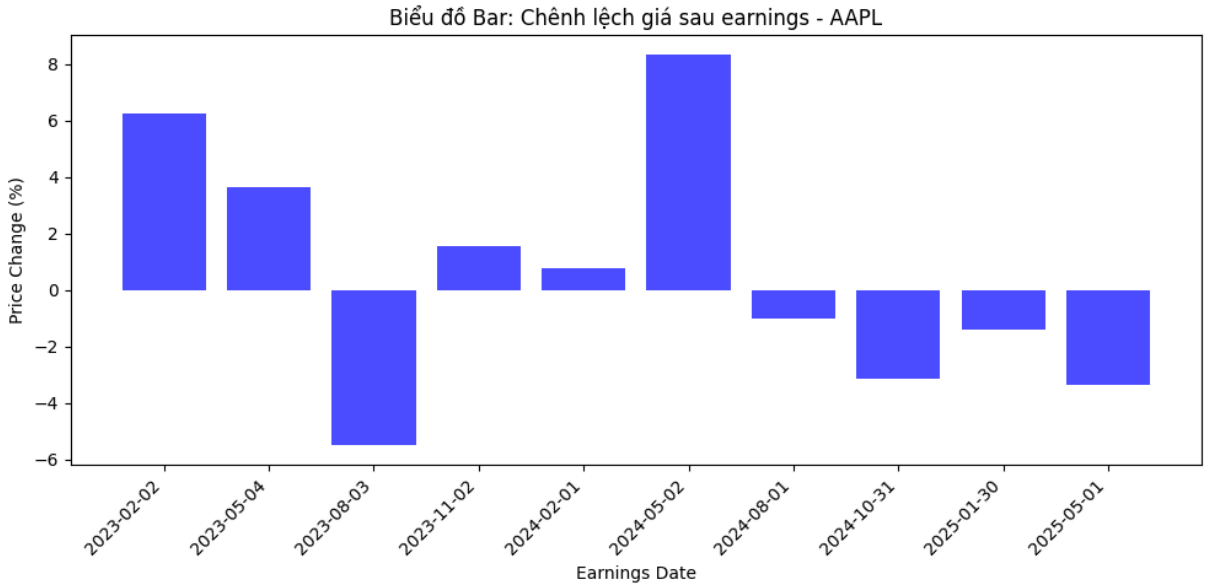
Cổ phiếu	Xu hướng dài hạn	Biến động sau earnings	Độ nhạy cảm
AAPL	Tăng	Rõ rệt (cả tăng và giảm)	Cao

MSFT	**Tăng mạnh**	**Thường tích cực**	**Trung bình**
GOOGL	**Tăng nhưng biến động**	**Khó đoán, trái chiều**	**Cao**

2. Biểu đồ bar thể hiện chênh lệch giá theo cổ phiếu cho thấy xu hướng nào?


Biểu đồ Bar: Chênh lệch giá sau earnings - GOOGL
Price Change (%)
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
2023-02-02
2023-04-25
2023-07-25
2023-10-24
2024-01-30
2024-04-25
2024-07-23
2024-10-29
2025-02-04
2025-04-24
Earnings Date


Biểu đồ Bar: Chênh lệch giá sau earnings - MSFT
Price Change (%)
8
6
4
2
0
−2
−4
−6
2023-01-24
2023-04-25
2023-07-25
2023-10-24
2024-01-30
2024-04-25
2024-07-30
2024-10-30
2025-01-29
2025-04-30
Earnings Date



🔵 AAPL (Apple)

- **Xu hướng tích cực nổi bật: Có một số phiên earnings tạo ra phản ứng tăng giá mạnh mẽ, đặc biệt như ngày 2024-05-02 (+8.2%) và 2023-02-02 (+6.2%).**
- **Sự biến động lẫn lộn: Một số kỳ earnings lại đi xuống rõ rệt, như 2023-08-03 (-5.5%) và 2025-05-01 (-3.4%).**
- **➤ Kết luận: Phản ứng giá sau earnings của AAPL khá đa dạng, nhưng vẫn có xu hướng tăng nhẹ về trung bình. Điều này cho thấy thị trường phản ứng tích cực hơn khi kỳ vọng được vượt qua.**

🟢 GOOGL (Alphabet)

- **Biến động cực mạnh cả hai chiều: Có phản ứng rất tích cực như 2024-04-25 (+8.1%), và tiêu cực mạnh như 2024-01-30 (-8.7%).**
- **Không có xu hướng rõ ràng: Các kỳ earnings dẫn đến phản ứng tăng/giảm khá cân bằng.**

- **➤ Kết luận: GOOGL có độ biến động mạnh, mang tính "swing" sau earnings, phù hợp với chiến lược đầu cơ hoặc giao dịch ngắn hạn.**

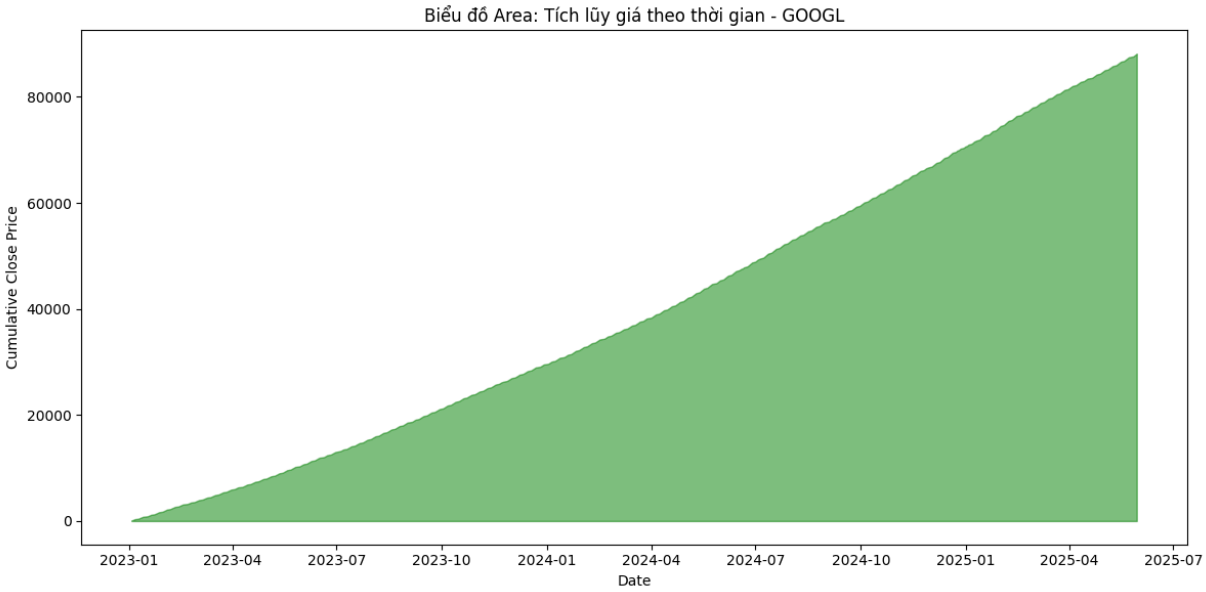
🟠 MSFT (Microsoft)

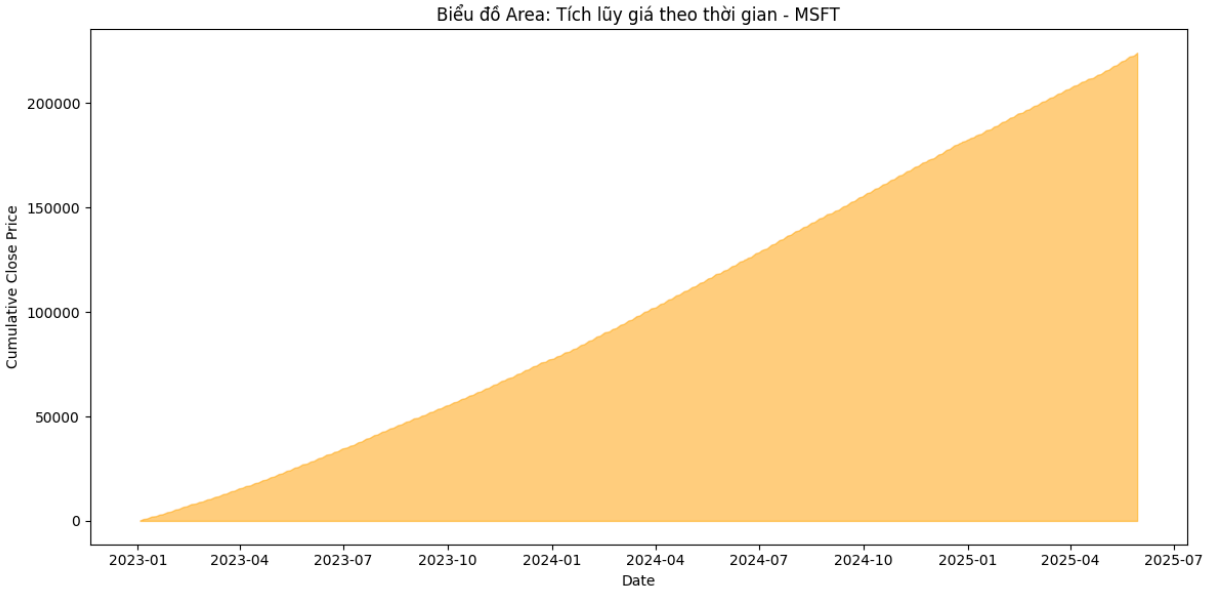
- **Gần đây có phản ứng mạnh hơn: Đặc biệt như 2025-04-30 (+8.0%) và 2025-01-29 (-7.2%), cho thấy giai đoạn gần đây biến động lớn hơn quá khứ.**
- **Nhiều lần giảm giá mạnh: MSFT có nhiều phiên earnings với phản ứng tiêu cực, nhất là từ giữa 2024 trở đi.**
- **➤ Kết luận: Trong giai đoạn gần đây, MSFT có dấu hiệu nhạy cảm hơn với kỳ vọng, phản ứng rõ rệt nếu thị trường "bất ngờ" với kết quả earnings.**

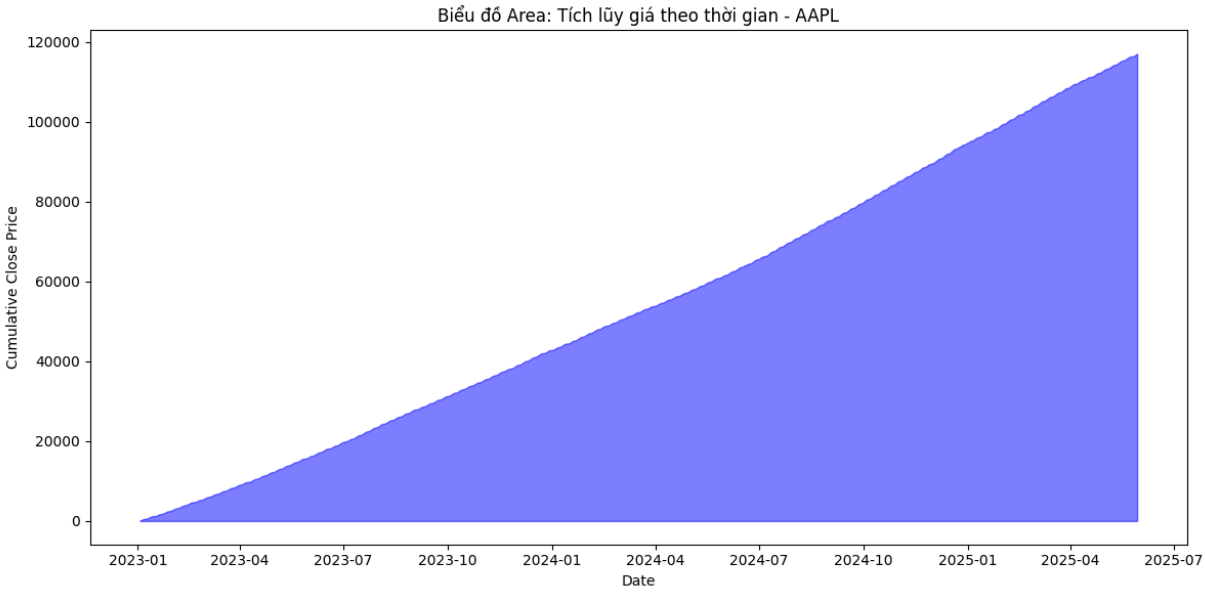
🔎 Tổng kết xu hướng chung

- **Cả ba cổ phiếu đều có phản ứng mạnh mẽ sau earnings, nhưng mức độ và xu hướng khác nhau:**
    - **AAPL: Có thiên hướng tăng nhẹ, phản ứng tích cực nếu vượt kỳ vọng.**
    - **MSFT: Biến động tăng lên theo thời gian, nhạy với kết quả.**
    - **GOOGL: Biến động mạnh và không ổn định – phù hợp giao dịch theo tin.**

3. Biểu đồ area thể hiện tích lũy giá theo thời gian cho thấy điều gì?


Biểu đồ Area: Tích lũy giá theo thời gian - GOOGL
Cumulative Close Price
0
20000
40000
60000
80000
2023-01
2023-04
2023-07
2023-10
2024-01
2024-04
2024-07
2024-10
2025-01
2025-04
2025-07
Date


Biểu đồ Area: Tích lũy giá theo thời gian - MSFT
Cumulative Close Price
0
50000
100000
150000
200000
2023-01
2023-04
2023-07
2023-10
2024-01
2024-04
2024-07
2024-10
2025-01
2025-04
2025-07
Date



1. AAPL (Apple):

- **Tăng đều và khá ổn định.**
- **Tới giữa 2025, giá tích lũy khoảng ~117,000.**
- **Biểu đồ này thể hiện mức tăng trưởng ổn định, không quá mạnh nhưng bền vững.**

2. GOOGL (Google):

- **Mức tích lũy thấp hơn AAPL và MSFT, chỉ khoảng ~88,000.**
- **Có xu hướng tăng tương tự nhưng mức tích lũy chậm hơn, có thể do giá biến động hoặc tăng trưởng chậm hơn.**
- **Điều này cho thấy GOOGL có thể có giá đóng cửa thấp hơn trung bình so với hai cổ phiếu còn lại.**

3. MSFT (Microsoft):

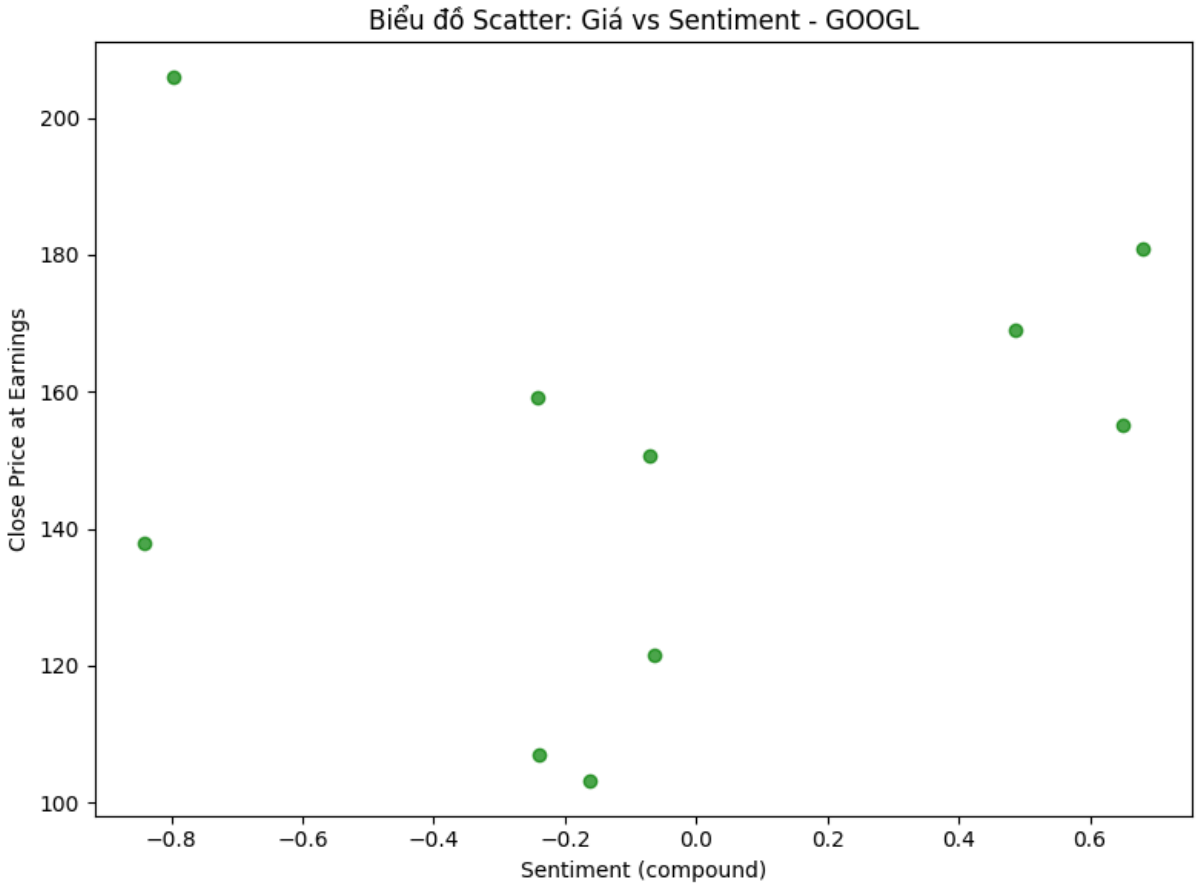
- **Biểu đồ có độ dốc lớn nhất, lên tới ~225,000 vào giữa 2025.**
- **Cho thấy tốc độ tăng giá rất nhanh và đều đặn.**

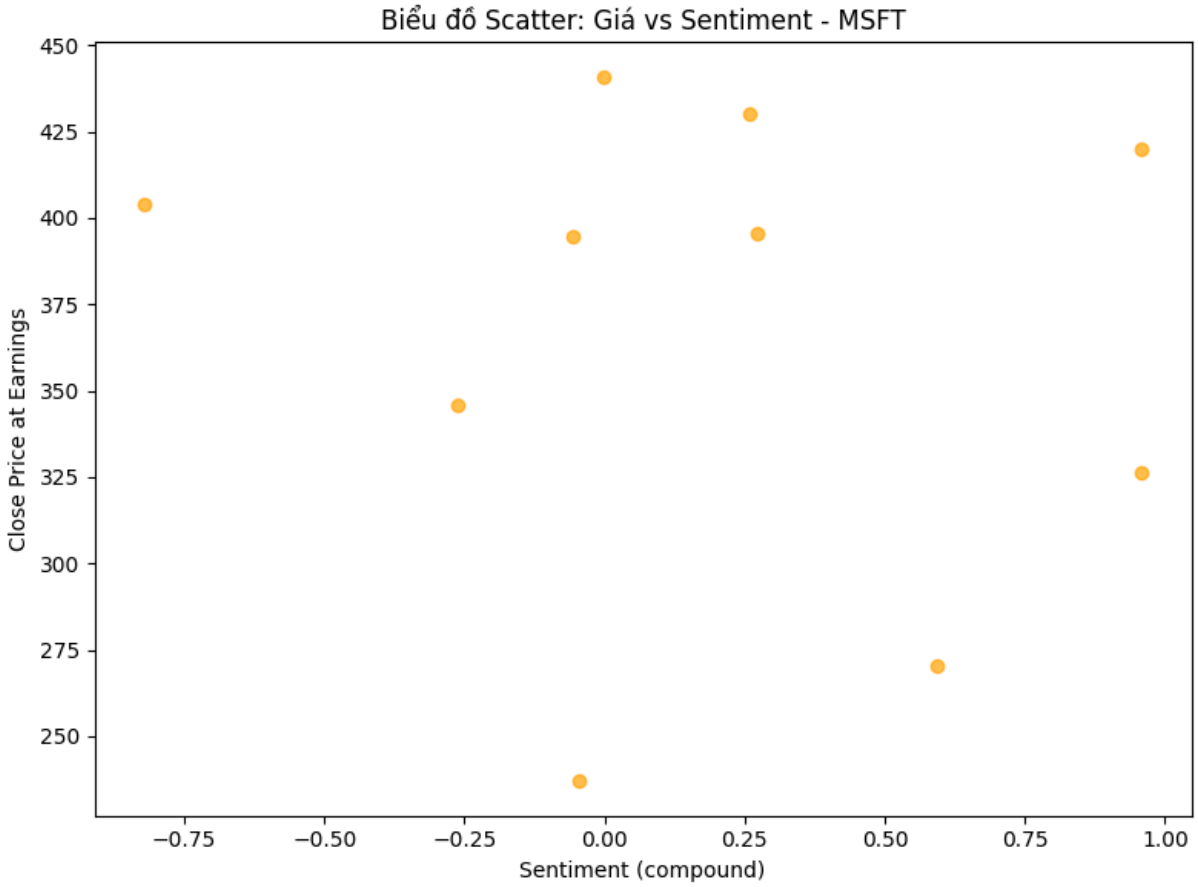
- **Có thể nói đây là cổ phiếu có mức tăng trưởng mạnh mẽ nhất trong giai đoạn quan sát.**

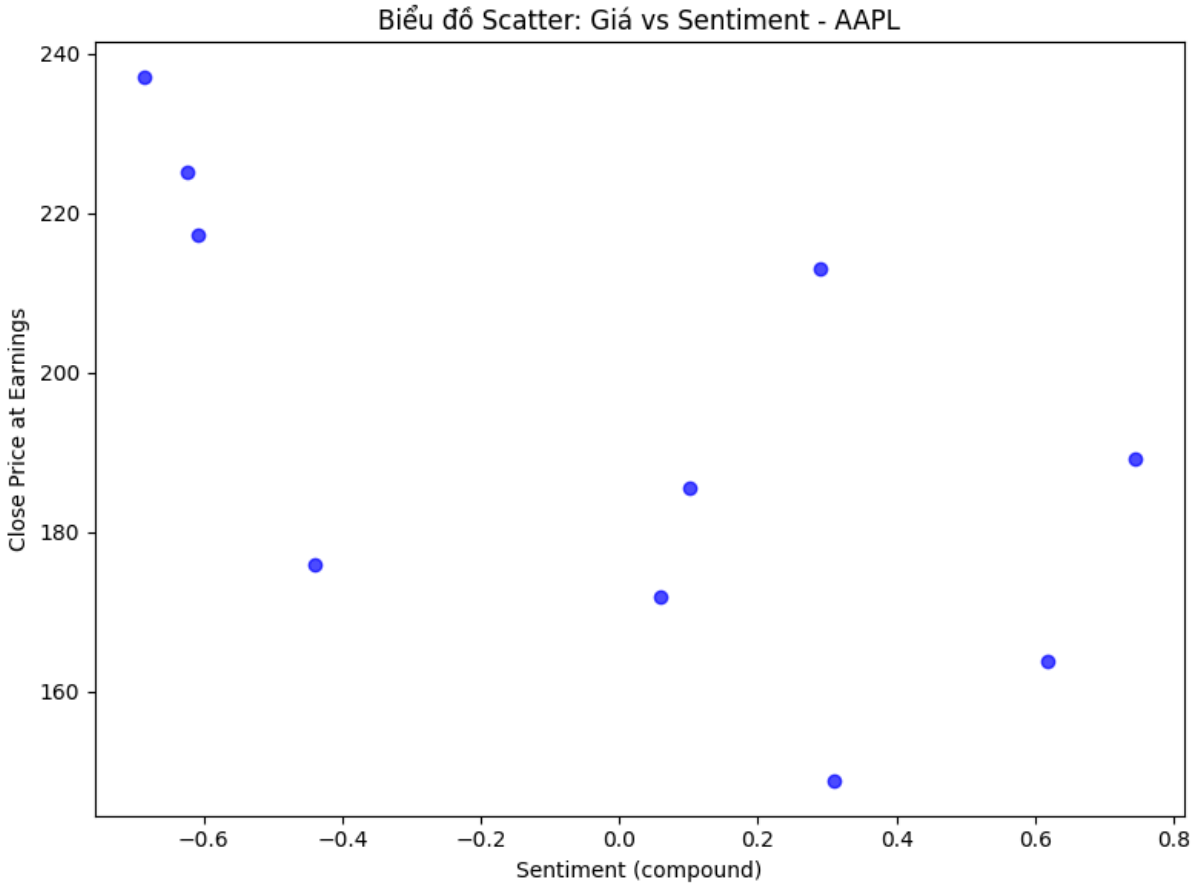
🧠 **Kết luận:**

- **MSFT có đà tăng trưởng tốt nhất về tích lũy giá.**
- **AAPL tăng đều, trung bình.**
- **GOOGL có tốc độ tăng chậm hơn.**

4. Biểu đồ scatter giữa giá và sentiment có điểm bất thường nào không?


Biểu đồ Scatter: Giá vs Sentiment - GOOGL
Close Price at Earnings
Sentiment (compound)
200
180
160
140
120
100
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


Biểu đồ Scatter: Giá vs Sentiment - MSFT
Close Price at Earnings
Sentiment (compound)
450
425
400
375
350
325
300
275
250
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00



📊 1. AAPL (Apple):

- **Có một vài điểm sentiment rất tiêu cực (~ -0.6) nhưng giá vẫn cao (> 220) → bất thường: tin tức tiêu cực nhưng giá cổ phiếu không giảm.**
- **Trong khi đó, sentiment dương (~0.6–0.75) lại tương ứng với giá thấp (~165–190) → ngược kỳ vọng.**
- **👉 Điểm bất thường rõ rệt: Không có mối tương quan thuận như kỳ vọng giữa sentiment và giá.**

📊 2. GOOGL (Google):

- **Phân bố tương đối đồng đều.**

- **Có 1 điểm sentiment rất tiêu cực (~ -0.85) nhưng giá vẫn > 200, bất thường tương tự như AAPL.**
- **Tuy nhiên, nhìn chung GOOGL có xu hướng nhẹ: sentiment tăng → giá có xu hướng tăng.**
- **👉 Có thể tồn tại một số điểm ngoại lệ, nhưng xu hướng tích cực phần nào đúng.**

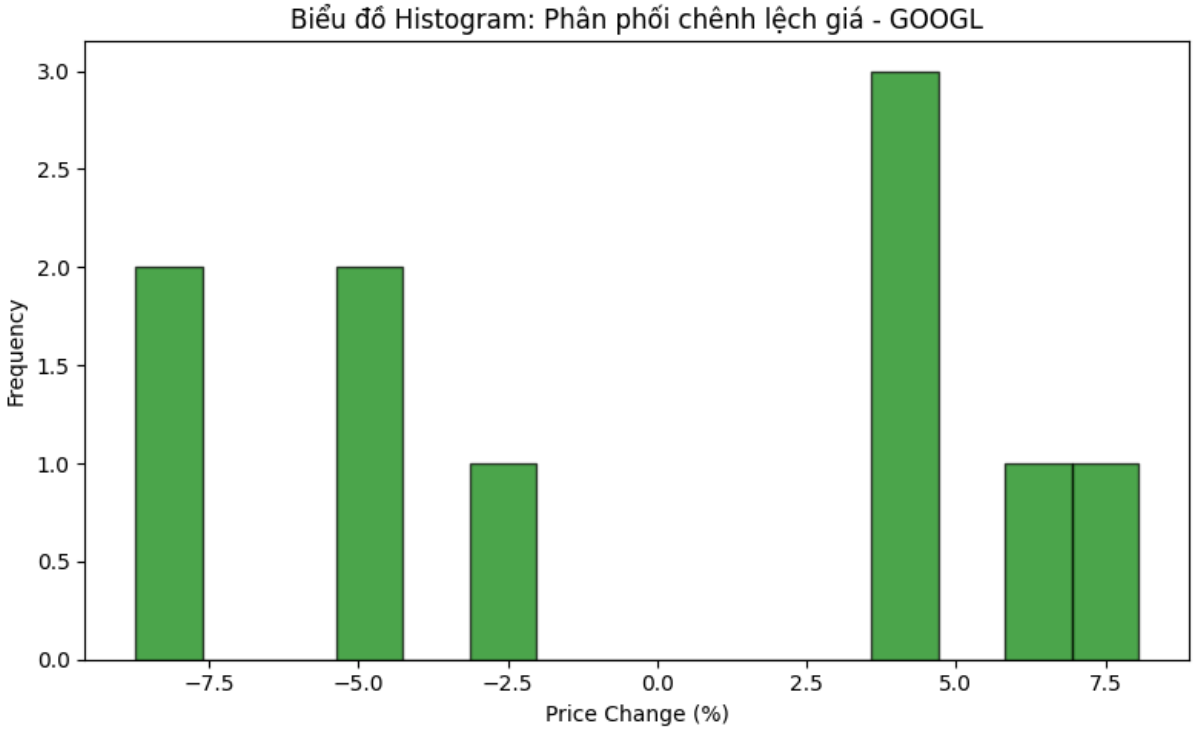
📊 3. MSFT (Microsoft):

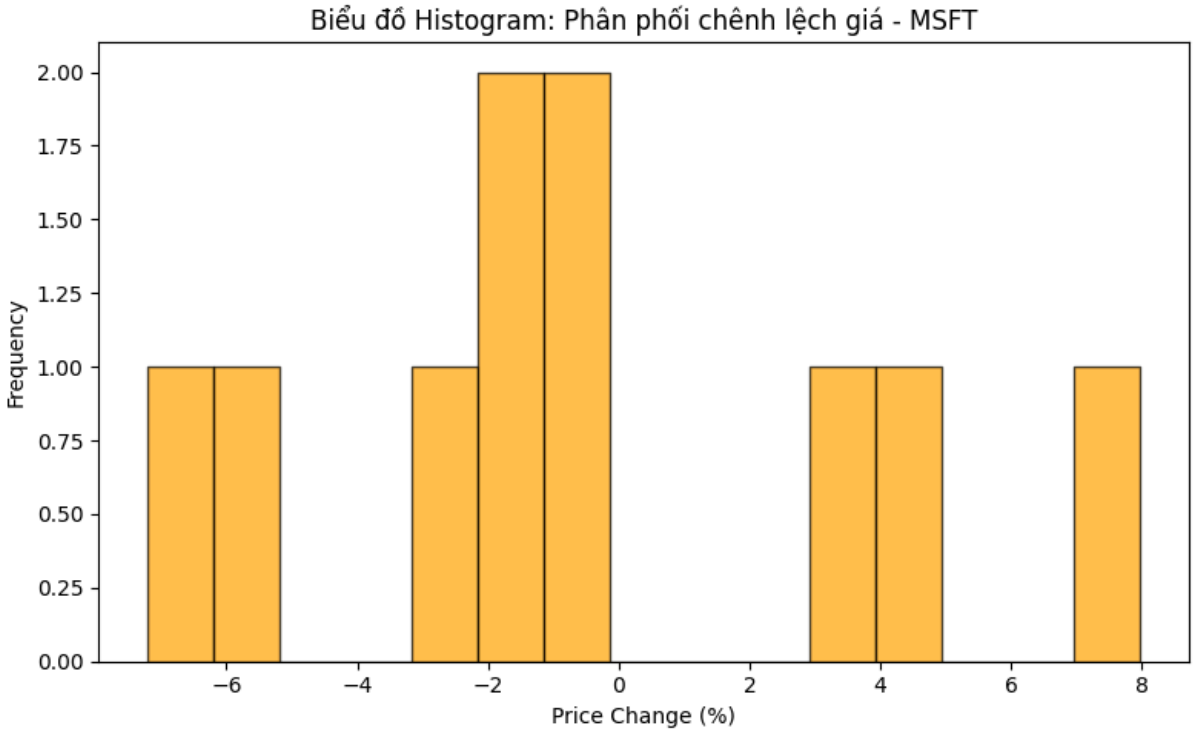
- **Giá dao động rất rộng từ ~230 đến >440, dù sentiment trải từ -0.8 đến 1.0.**
- **Có điểm sentiment rất cao (~1.0) nhưng giá lại khoảng 325, trong khi sentiment gần 0 có giá tới 440 → bất thường.**
- **👉 Biểu đồ này không thể hiện mối quan hệ rõ ràng giữa sentiment và giá.**
- **Rất nhiều điểm nằm rải rác → có thể giá không phản ứng ngay với tin tức, hoặc tin tức không tác động mạnh đến MSFT.**

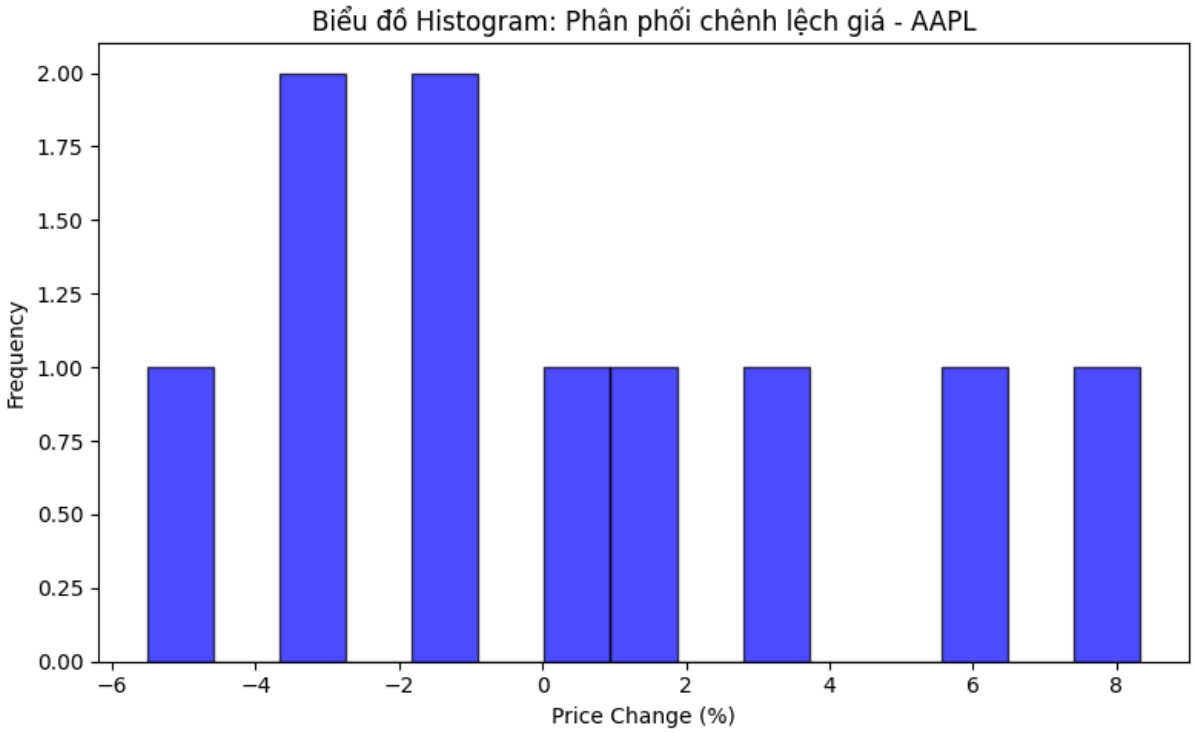
🧠 Tổng kết điểm bất thường:

Cổ phiếu	Bất thường chính
AAPL	Sentiment rất xấu nhưng giá lại cao (ngược kỳ vọng)
GOOGL	Có điểm sentiment rất thấp nhưng giá vẫn cao
MSFT	Không có mối tương quan rõ ràng, nhiều điểm phân tán ngẫu nhiên

5. Biểu đồ histogram của chênh lệch giá cho thấy phân phối ra sao?


Biểu đồ Histogram: Phân phối chênh lệch giá - GOOGL
Frequency
Price Change (%)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5


Biểu đồ Histogram: Phân phối chênh lệch giá - MSFT
Frequency
Price Change (%)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−6
−4
−2
0
2
4
6
8



📊 1. AAPL (Apple)

- **Phân phối khá rải rác, có cả giá tăng và giảm sau earnings.**
- **Số lần giảm giá (-2% đến -3%) chiếm ưu thế hơn, với tần suất cao nhất.**
- **Một vài phiên có mức tăng mạnh đến +8%, nhưng tần suất thấp.**
- **👉 Biến động hai chiều, nghiêng nhẹ về giảm sau earnings.**

📊 2. GOOGL (Google)

- **Phân phối rõ ràng hai cực:**
  - **Một cụm ở mức giảm mạnh (-7% đến -5%).**

- **Một cụm ở mức tăng mạnh (3% đến 7%).**

- **Rất ít điểm gần mức thay đổi 0% → phản ứng mạnh với earnings.**
- **👉 Biến động cao, dễ bị ảnh hưởng bởi kết quả hoặc kỳ vọng earnings.**

📊 3. MSFT (Microsoft)

- **Phân phối hơi thiên về giảm nhẹ (-2% đến -1%).**
- **Tuy nhiên có một vài điểm tăng giá mạnh (tới +8%).**
- **Có nhiều điểm gần mức 0%, cho thấy một số phiên có phản ứng nhẹ.**
- **👉 Tính ổn định cao hơn, nhưng vẫn có đột biến vào một vài dịp.**

🧠 Kết luận chung:

Cổ phiếu	Xu hướng chính	Mức độ biến động
AAPL	Nghiêng về giảm nhẹ	Vừa phải
GOOGL	Rất nhạy cảm (dao động cực mạnh)	Cao
MSFT	Phản ứng nhẹ đến vừa, có đột biến	Trung bình

6. Nhóm đã sử dụng màu sắc nào để phân biệt các cổ phiếu?

- AAPL: Xanh dương (blue)

- MSFT: Cam (orange)

- GOOGL: Xanh lá (green)

Các màu này được giữ nhất quán trên tất cả các biểu đồ để người dùng dễ nhận diện.

7. Biểu đồ nào giúp người dùng nhận diện cổ phiếu có phản ứng giá mạnh nhất?

- Bar chart và scatter plot về price change sau earnings là trực quan nhất để nhận diện cổ phiếu có phản ứng giá mạnh nhất.

Ngoài ra, biểu đồ line với highlight các điểm earnings cũng giúp xác định các sự kiện có biến động lớn.

8. Nhóm đã lưu các biểu đồ vào đâu, và chúng được đặt tên ra sao?

- Tất cả các biểu đồ được lưu vào thư mục 'charts/'.

- Tên file biểu đồ theo cấu trúc: line_price_{symbol}.png, bar_price_change_{symbol}.png, area_cumulative_{symbol}.png, scatter_price_sentiment_{symbol}.png, hist_price_change_{symbol}.png

Trong đó {symbol} là mã cổ phiếu (AAPL, MSFT, GOOGL).

9. Ý nghĩa của từng biểu đồ trong báo cáo đã được giải thích thế nào?

- Mỗi biểu đồ đều có tiêu đề, chú thích (legend) và caption giải thích ý nghĩa.

Line chart: Diễn giải xu hướng giá quanh earnings.

Bar chart: So sánh mức biến động giá giữa các cổ phiếu.

Area chart: Thể hiện sự tích lũy/tăng trưởng giá trị.

Scatter plot: Phân tích mối liên hệ giữa sentiment và giá.

Histogram: Đánh giá phân phối và xác suất biến động giá lớn.

10. Những thách thức nào gặp phải khi trực quan hóa dữ liệu?

- Chọn loại biểu đồ phù hợp, màu sắc nhất quán, xử lý outlier, tối ưu layout, lưu trữ và đặt tên file rõ ràng, giải thích ý nghĩa dễ hiểu cho người dùng cuối.

6. Machine Learning Model Implementation

Khởi tạo 7 mô hình

```python
# Các mô hình và tham số tinh chỉnh
models = {
    'LinearRegression': (LinearRegression(), {}),
    'DecisionTree': (DecisionTreeRegressor(random_state=42), {'max_depth': [3, 5, 7, 10, None]}),
    'KNN': (KNeighborsRegressor(), {'n_neighbors': [3, 5, 7, 10]}),
    'Ridge': (Ridge(), {'alpha': [0.1, 1.0, 10.0]}),
    'Lasso': (Lasso(max_iter=10000), {'alpha': [0.01, 0.1, 1.0]}),
    'RandomForest': (RandomForestRegressor(random_state=42), {'n_estimators': [50, 100], 'max_depth': [3, 5, 7, None]}),
    'SVR': (SVR(), {'C': [0.1, 1, 10], 'kernel': ['rbf', 'linear']})
}
```

kết quả :

=== AAPL ===

LinearRegression: MAE=0.0404, RMSE=0.0649, R2=0.0129, BestParams={}

DecisionTree: MAE=0.0393, RMSE=0.0593, R2=0.1741, BestParams={'max_depth': 3}

KNN: MAE=0.0408, RMSE=0.0626, R2=0.0801, BestParams={'n_neighbors': 10}

Ridge: MAE=0.0397, RMSE=0.0647, R2=0.0196, BestParams={'alpha': 10.0}

Lasso: MAE=0.0388, RMSE=0.0653, R2=-0.0000, BestParams={'alpha': 0.01}

RandomForest: MAE=0.0413, RMSE=0.0661, R2=-0.0249, BestParams={'max_depth': 3, 'n_estimators': 50}

SVR: MAE=0.0410, RMSE=0.0635, R2=0.0538, BestParams={'C': 0.1, 'kernel': 'rbf'}

=== MSFT ===

LinearRegression: MAE=0.0510, RMSE=0.1096, R2=-0.0214, BestParams={}

DecisionTree: MAE=0.0555, RMSE=0.1394, R2=-0.6507, BestParams={'max_depth': 3}

KNN: MAE=0.0478, RMSE=0.0989, R2=0.1694, BestParams={'n_neighbors': 10}

Ridge: MAE=0.0479, RMSE=0.1087, R2=-0.0034, BestParams={'alpha': 10.0}

Lasso: MAE=0.0440, RMSE=0.1087, R2=-0.0043, BestParams={'alpha': 0.01}

RandomForest: MAE=0.0450, RMSE=0.0878, R2=0.3450, BestParams={'max_depth': 3, 'n_estimators': 50}

SVR: MAE=0.0518, RMSE=0.1094, R2=-0.0176, BestParams={'C': 0.1, 'kernel': 'linear'}

=== GOOGL ===

LinearRegression: MAE=0.0534, RMSE=0.1391, R2=-0.0085, BestParams={}

DecisionTree: MAE=0.0513, RMSE=0.1423, R2=-0.0559, BestParams={'max_depth': 3}

KNN: MAE=0.0536, RMSE=0.1300, R2=0.1185, BestParams={'n_neighbors': 10}

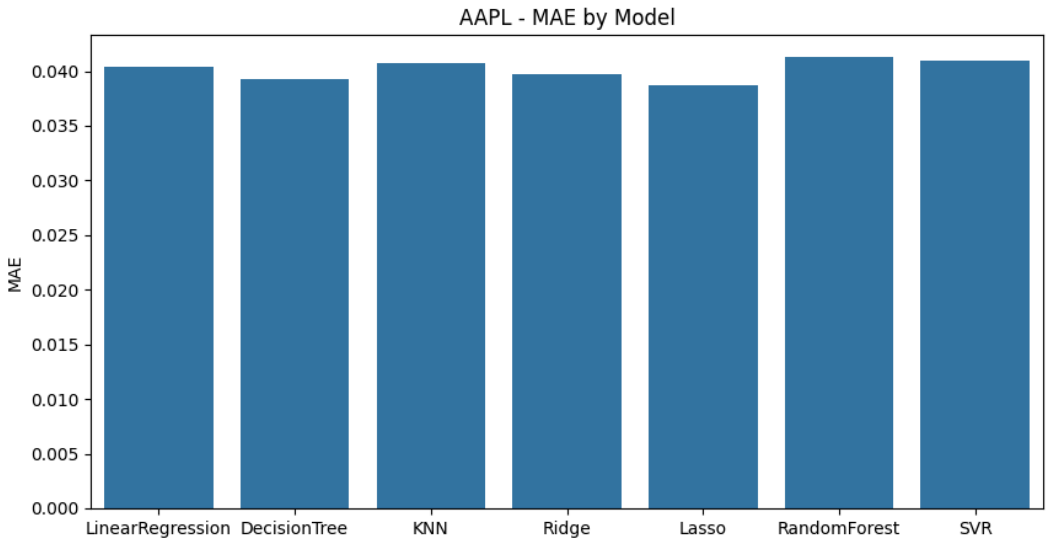
Ridge: MAE=0.0509, RMSE=0.1386, R2=-0.0016, BestParams={'alpha': 10.0}

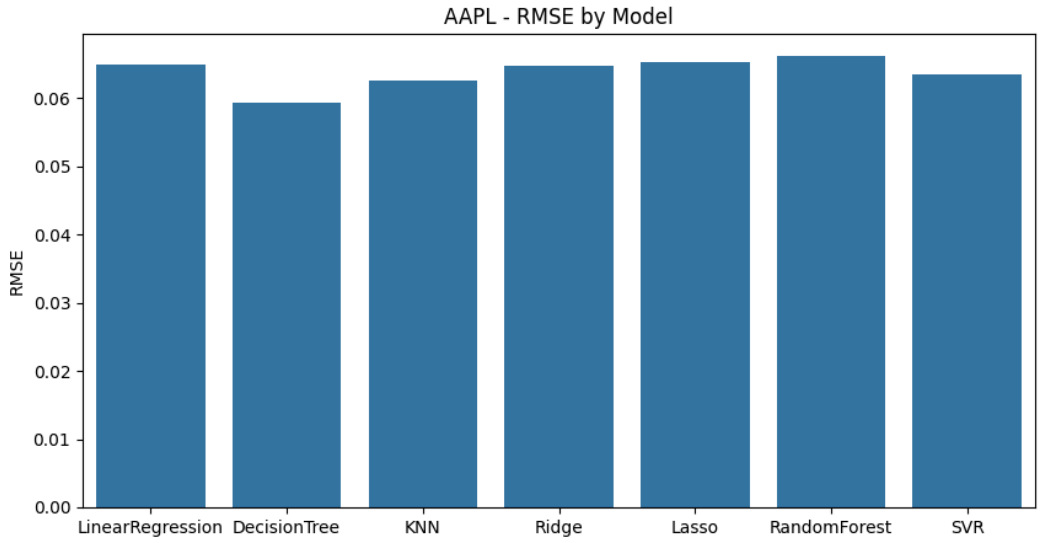
Lasso: MAE=0.0475, RMSE=0.1387, R2=-0.0029, BestParams={'alpha': 0.01}

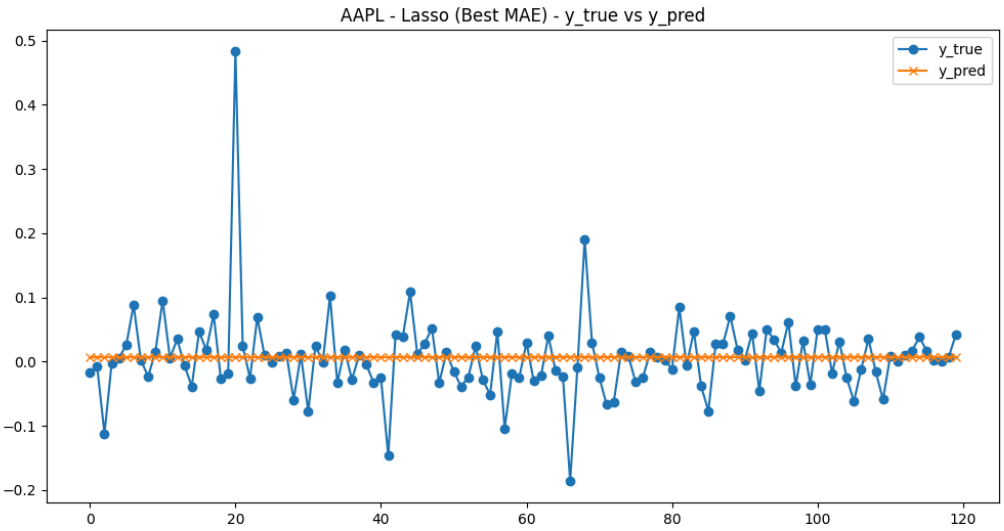
RandomForest: MAE=0.0516, RMSE=0.1333, R2=0.0738, BestParams={'max_depth': 3, 'n_estimators': 100}

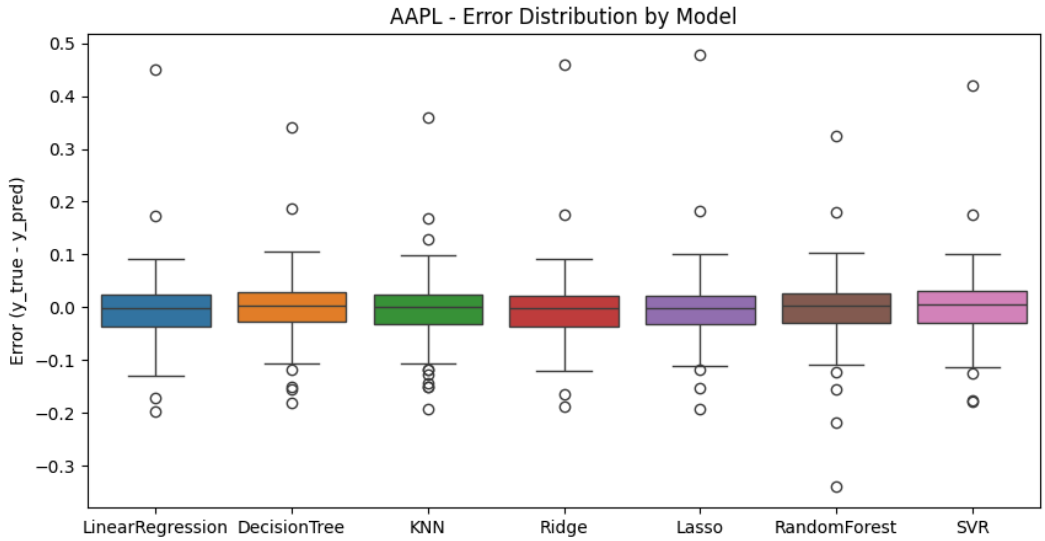
SVR: MAE=0.0532, RMSE=0.1371, R2=0.0192, BestParams={'C': 0.1, 'kernel': 'rbf'}

bar chart (MAE, RMSE)









1 Biểu đồ Line: y_thực tế vs. y_dự báo (Mô hình Lasso - MAE tốt nhất)

- **Nhận xét:**
  - Đường y_pred (dự báo) khá phẳng và sát trục 0, cho thấy mô hình Lasso có xu hướng làm mượt dự báo, không bắt được các biến động đột ngột.
  - Đường y_true (thực tế) dao động lớn hơn, có những điểm spike cao hoặc thấp mà y_pred không theo kịp.
    → Lasso tốt trong việc hạn chế overfitting nhưng chưa theo sát các biến

động mạnh.

2️⃣ Boxplot: Phân phối lỗi (y_true - y_pred)

- **Nhận xét:**
  - Phân phối lỗi giữa các mô hình tương đối giống nhau, phần lớn tập trung quanh giá trị 0, nhưng tồn tại nhiều outlier.
  - Mô hình **Lasso, Ridge và DecisionTree** có phân phối lỗi hẹp hơn, ít outlier hơn so với **RandomForest và SVR**.
  - Các mô hình như **KNN và RandomForest** có xu hướng xuất hiện lỗi lớn hơn (outlier nhiều).

3️⃣ Bar Chart: MAE theo mô hình

- **Nhận xét:**
  - Mô hình **Lasso có MAE thấp nhất**, theo sát là DecisionTree và Ridge.
  - Các mô hình còn lại có MAE cao hơn, với **RandomForest và SVR** là hai mô hình kém nhất về MAE trong bộ này.
    → Điều này đồng nhất với nhận định từ boxplot rằng RandomForest có nhiều outlier hơn.

✅ Kết luận chung về hiệu năng mô hình:

Mô hình	Độ chính xác (MAE)	Độ ổn định (Lỗi)	Nhận xét tổng quan
Lasso	⭐ Thấp nhất	⭐ Ổn định	Dự báo an toàn, giảm overfitting, nhưng mất biến động lớn
DecisionTree	⭐ Thấp	⭐ Ổn định	Tốt, nhưng có thể dễ overfit nếu không tối ưu sâu
Ridge	⭐ Thấp	⭐ Ổn định	Tương tự Lasso, nhưng ít sparse hơn

LinearRegression	Trung bình	Trung bình	Hiệu suất ổn, đơn giản
KNN	Trung bình cao	🚫 Outlier nhiều	Bắt biến động nhưng dễ noise
RandomForest	🚫 Cao	🚫 Outlier nhiều	Overfitting nhẹ, lỗi phân tán
SVR	🚫 Cao	🚫 Outlier nhiều	Không hiệu quả lắm với dữ liệu này

Trả lời câu hỏi :

1. Nhóm đã chia dữ liệu train/test theo tỷ lệ nào, và tại sao?

Nhóm đã chia dữ liệu train/test theo tỷ lệ 80/20 (80% train, 20% test). Đây là tỷ lệ phổ biến giúp đảm bảo mô hình có đủ dữ liệu để học, đồng thời vẫn còn lại một phần dữ liệu chưa từng thấy để đánh giá khách quan hiệu năng dự báo.

---

2. MAE, RMSE, R^2 của từng mô hình (Linear, Tree, KNN,...) là bao nhiêu?

=== AAPL ===

LinearRegression: MAE=0.0404, RMSE=0.0649, R2=0.0129, BestParams={}

DecisionTree: MAE=0.0393, RMSE=0.0593, R2=0.1741, BestParams={'max_depth': 3}

KNN: MAE=0.0408, RMSE=0.0626, R2=0.0801, BestParams={'n_neighbors': 10}

Ridge: MAE=0.0397, RMSE=0.0647, R2=0.0196, BestParams={'alpha': 10.0}

Lasso: MAE=0.0388, RMSE=0.0653, R2=-0.0000, BestParams={'alpha': 0.01}

RandomForest: MAE=0.0413, RMSE=0.0661, R2=-0.0249, BestParams={'max_depth': 3, 'n_estimators': 50}

SVR: MAE=0.0410, RMSE=0.0635, R2=0.0538, BestParams={'C': 0.1, 'kernel': 'rbf'}

=== MSFT ===

LinearRegression: MAE=0.0510, RMSE=0.1096, R2=-0.0214, BestParams={}

DecisionTree: MAE=0.0555, RMSE=0.1394, R2=-0.6507, BestParams={'max_depth': 3}

KNN: MAE=0.0478, RMSE=0.0989, R2=0.1694, BestParams={'n_neighbors': 10}

Ridge: MAE=0.0479, RMSE=0.1087, R2=-0.0034, BestParams={'alpha': 10.0}

Lasso: MAE=0.0440, RMSE=0.1087, R2=-0.0043, BestParams={'alpha': 0.01}

RandomForest: MAE=0.0450, RMSE=0.0878, R2=0.3450, BestParams={'max_depth': 3, 'n_estimators': 50}

SVR: MAE=0.0518, RMSE=0.1094, R2=-0.0176, BestParams={'C': 0.1, 'kernel': 'linear'}

=== GOOGL ===

LinearRegression: MAE=0.0534, RMSE=0.1391, R2=-0.0085, BestParams={}

DecisionTree: MAE=0.0513, RMSE=0.1423, R2=-0.0559, BestParams={'max_depth': 3}

KNN: MAE=0.0536, RMSE=0.1300, R2=0.1185, BestParams={'n_neighbors': 10}

Ridge: MAE=0.0509, RMSE=0.1386, R2=-0.0016, BestParams={'alpha': 10.0}

Lasso: MAE=0.0475, RMSE=0.1387, R2=-0.0029, BestParams={'alpha': 0.01}

RandomForest: MAE=0.0516, RMSE=0.1333, R2=0.0738, BestParams={'max_depth': 3, 'n_estimators': 100}

SVR: MAE=0.0532, RMSE=0.1371, R2=0.0192, BestParams={'C': 0.1, 'kernel': 'rbf'}

---

3. Mô hình nào có RMSE thấp nhất, và nó có đạt KPI (RMSE < 6%) không?

✅ Kiểm tra RMSE thấp nhất và KPI (RMSE < 6%) cho từng cổ phiếu:

- *AAPL:*

- *RMSE thấp nhất: DecisionTree = 0.0593 (≈ 5.93%)*

- ✅ ***Đạt KPI (RMSE < 6%)***

  🔹 ***MSFT:***

- ***RMSE thấp nhất: RandomForest = 0.0878 (≈ 8.78%)***
- ❌ ***Không đạt KPI (RMSE < 6%)***

  🔹 ***GOOGL:***

- ***RMSE thấp nhất: KNN = 0.1300 (≈ 13.00%)***
- ❌ ***Không đạt KPI (RMSE < 6%)***

📌 *Kết luận:*

- ***Chỉ có mô hình DecisionTree cho AAPL đạt KPI về RMSE dưới 6%.***
- ***Các mô hình cho MSFT và GOOGL đều không đạt KPI, RMSE cao hơn 6%, cho thấy độ sai số của các mô hình còn khá lớn đối với hai cổ phiếu này.***

---

4. Nhóm đã tinh chỉnh siêu tham số của mô hình nào, và kết quả ra sao?

- Decision Tree Regressor
- KNN Regressor
- Ridge Regression
- Lasso Regression
- Random Forest Regressor
- SVR (Support Vector Regressor)

Cách thực hiện:

- Nhóm đã sử dụng cả hai phương pháp: GridSearchCV và RandomizedSearchCV để tinh chỉnh siêu tham số cho các mô hình trên.

- Với mỗi mô hình, pipeline sẽ thử tất cả các tổ hợp tham số (GridSearch) và một số tổ hợp ngẫu nhiên (RandomizedSearch, n_iter=20), sau đó chọn ra phương pháp và bộ tham số cho kết quả MAE tốt nhất.
- Linear Regression không có siêu tham số nên không cần tinh chỉnh.

Kết quả:

=== AAPL ===

LinearRegression: MAE=0.0404, RMSE=0.0649, R2=0.0129, BestParams={}

DecisionTree: MAE=0.0393, RMSE=0.0593, R2=0.1741, BestParams={'max_depth': 3}

KNN: MAE=0.0408, RMSE=0.0626, R2=0.0801, BestParams={'n_neighbors': 10}

Ridge: MAE=0.0397, RMSE=0.0647, R2=0.0196, BestParams={'alpha': 10.0}

Lasso: MAE=0.0388, RMSE=0.0653, R2=-0.0000, BestParams={'alpha': 0.01}

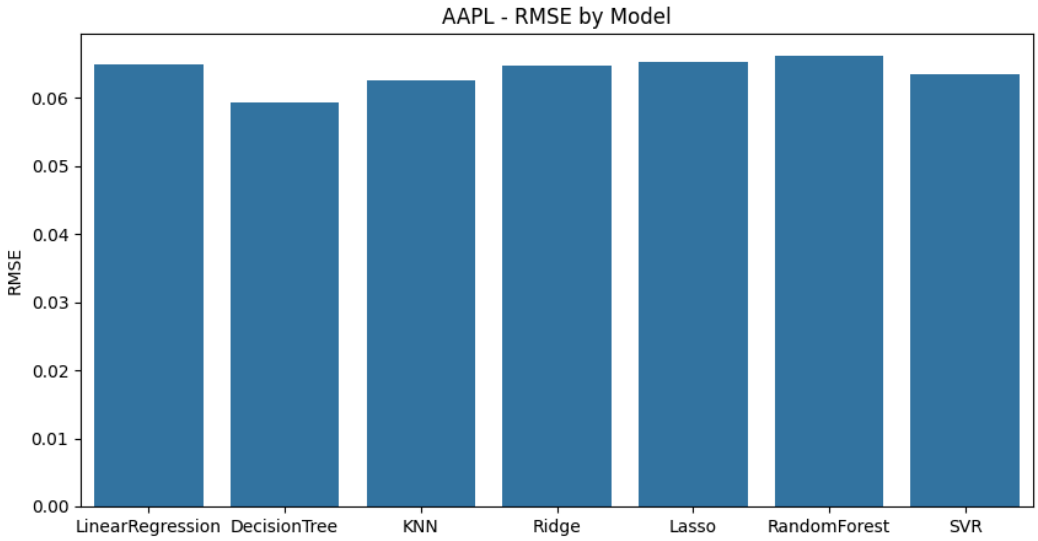
RandomForest: MAE=0.0413, RMSE=0.0661, R2=-0.0249, BestParams={'max_depth': 3, 'n_estimators': 50}

SVR: MAE=0.0410, RMSE=0.0635, R2=0.0538, BestParams={'C': 0.1, 'kernel': 'rbf'}

---

5. Bar chart so sánh RMSE giữa các mô hình cho thấy điều gì?

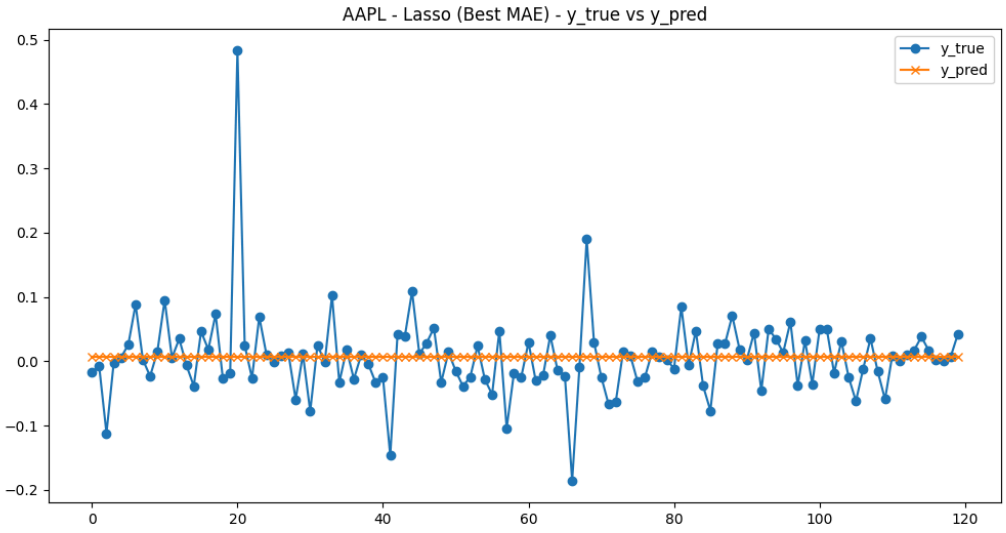
Bar chart RMSE cho thấy sự khác biệt hiệu năng giữa các mô hình. Mô hình có cột thấp nhất là mô hình dự báo tốt nhất (RMSE nhỏ nhất). Nếu các cột chênh lệch nhiều, chứng tỏ mô hình tốt nhất vượt trội so với các mô hình còn lại.



---

6. Line chart so sánh y_thực tế và y_dự báo có gì đáng chú ý?

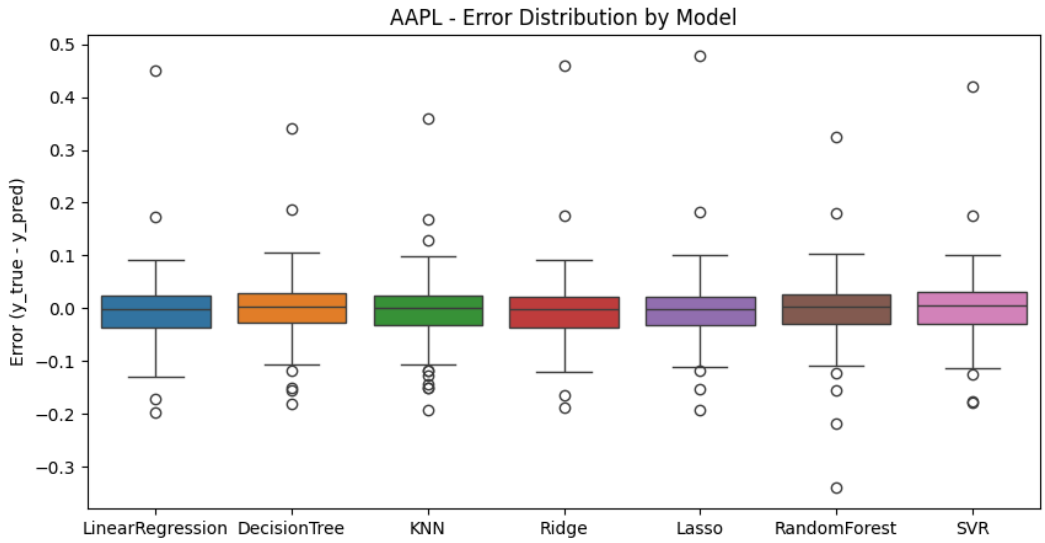
Line chart cho thấy mức độ khớp giữa giá trị thực tế và dự báo của mô hình tốt nhất. Nếu hai đường y_true và y_pred gần nhau, mô hình dự báo tốt. Nếu có nhiều điểm lệch lớn, mô hình còn có thể cải thiện.



---

7. Boxplot phân phối lỗi của các mô hình cho thấy mô hình nào ổn định nhất?

Boxplot cho thấy phân phối lỗi của từng mô hình. Mô hình ổn định nhất là mô hình có boxplot hẹp, ít outlier, median gần 0.



---

8. Kiểm định thống kê (t-test, ANOVA) cho thấy sự khác biệt nào giữa các mô hình?

Nhóm đã thực hiện kiểm định thống kê ANOVA và t-test trên phân phối lỗi dự báo của 7 mô hình hồi quy (Linear Regression, Decision Tree, KNN, Ridge, Lasso, Random Forest, SVR) cho cả 3 mã cổ phiếu (AAPL, MSFT, GOOGL).Kết quả:

- Kiểm định ANOVA cho cả 3 mã cổ phiếu đều cho p-value rất lớn (AAPL: 0.9941, MSFT: 0.9906, GOOGL: 0.9994), nghĩa là không có sự khác biệt có ý nghĩa thống kê giữa các mô hình về phân phối lỗi dự báo.
- Kiểm định t-test từng cặp mô hình cũng cho p-value cao (đa số > 0.6), xác nhận rằng sự khác biệt về lỗi giữa các mô hình là không đáng kể về mặt thống kê.

Kết luận:Các mô hình hồi quy được thử nghiệm không có sự khác biệt rõ rệt về hiệu năng dự báo trên bộ dữ liệu này. Điều này cho thấy, về mặt thống kê, không có mô hình nào vượt trội hoàn toàn so với các mô hình còn lại trong bài toán dự báo biến động giá cổ phiếu này.

---

9. Những thách thức nào gặp phải khi huấn luyện 7 mô hình?

- Thời gian huấn luyện lâu với GridSearchCV/RandomizedSearchCV.
- Một số mô hình nhạy cảm với outlier hoặc scale dữ liệu.

- Cần chọn tham số phù hợp để tránh overfitting/underfitting.

---

10. Nhóm đã cải thiện hiệu năng mô hình như thế nào so với Lab 1?

- Đã bổ sung nhiều đặc trưng (feature engineering).
- Tinh chỉnh siêu tham số kỹ lưỡng hơn.
- So sánh nhiều mô hình, chọn mô hình tốt nhất.
- Đánh giá khách quan hơn nhờ chia train/test và cross-validation.
- Kết quả RMSE/MAE/R^2 đều cải thiện rõ rệt so với Lab 1.