

# Báo cáo bài tập Hiểu dữ liệu: Phân tích cổ phiếu Công ty Cổ phần Sữa Việt Nam (VNM)

Nguyễn Thị Minh Ly  
23020399@vnu.edu.vn

Đặng Minh Nguyệt  
23020407@vnu.edu.vn

## 1 Quan sát để hiểu doanh nghiệp và dữ liệu

### 1.1 Tổng quan về doanh nghiệp

Vinamilk hiện là doanh nghiệp sữa lớn nhất Việt Nam, nắm giữ hơn 55% thị phần cùng danh mục sản phẩm phong phú và hệ thống phân phối trải dài khắp cả nước. Bên cạnh đó, công ty cũng dẫn đầu về vốn hóa, giá trị thương hiệu và năng lực sản xuất khi sở hữu trang trại sữa hữu cơ đầu tiên cũng như hai siêu nhà máy sữa hiện đại. Trong bối cảnh cạnh tranh ngày càng gay gắt, Vinamilk tiếp tục mở rộng quy mô, nâng cao năng lực sản xuất và đa dạng hóa nguồn nguyên liệu để duy trì tăng trưởng.

Việc phân tích cổ phiếu Vinamilk vì vậy sẽ giúp chúng ta hiểu rõ hơn:

- Mức độ bền vững trong nền tảng kinh doanh của doanh nghiệp
- Khả năng duy trì tăng trưởng trong ngành sữa cạnh tranh
- Triển vọng đầu tư trong ngắn hạn lẫn dài hạn

Từ đây, báo cáo được chia thành ba phần chính, đó là Thực hiện thống kê miêu tả, Tiền xử lý dữ liệu và Lựa chọn các yếu tố nguy cơ tiềm năng cho mô hình tài chính.

### 1.2 Thực hiện thống kê miêu tả

#### 1.2.1 Chỉ số tài chính

Trước tiên, ta thực hiện một vài bước xử lý để bắt đầu đọc dữ liệu.

Listing 1: Tìm header row, tức dòng bắt đầu chứa dữ liệu thực

```
1 # looking for header row
2 for i, row in df_fin.iterrows():
3     if row.astype(str).str.contains("CHI TIEU", case=False, na=False).
4         any():
5         header_row = i
6         break
7
8 # block rows before header
9 df_fin = pd.read_excel("data\Simplify_VNM_FinancialIndicator_20250315.
10 xlsx", header=header_row)
df_fin.columns = [str(c).strip() for c in df_fin.columns]
df_fin = df_fin.dropna(how='all')
```

Listing 2: Loại bỏ các category row

```

1 # identify quarter columns
2 quarter_cols = [c for c in df_fin.columns if any(q in str(c) for q in
   ['Q', '/202'])]
3
4 # get rid of category rows
5 col_name = 'CHI TIÊU'
6
7 df_fin[col_name] = df_fin[col_name].astype(str).str.strip()
8 df_fin['_is_category'] = df_fin[quarter_cols].isna().all(axis=1)
9
10 df_fin = df_fin[df_fin['_is_category'] == False].drop(columns=['
   _is_category'])

```

Listing 3: Loại bỏ dòng NaN và chuẩn hóa các giá trị số

```

1 # get rid of rows with no data in quarter columns
2 df_fin = df_fin[df_fin[quarter_cols].notna().any(axis=1)]
3
4 # normalize quarter columns
5 def clean_value(x):
6     if pd.isna(x): return np.nan
7     x = str(x).strip()
8     if x.endswith('%'):
9         try:
10             return float(x.replace('%', '').replace(',','.')) / 100
11         except: return np.nan
12     try:
13         return float(x.replace(',','').replace(' ','.'))
14     except:
15         return np.nan
16
17 for col in quarter_cols:
18     df_fin[col] = df_fin[col].apply(clean_value)

```

**Bảng Chỉ số tài chính sau khi đã được xử lý :**

CHỈ TIÊU	Q4/2024	Q3/2024	Q2/2024	Q1/2024	Q4/2023	Q3/2023	
Doanh thu thuần	15477073125441.00	15537337313473.00	16655787772473.00	14112411317058.00	156187109444490.00	15636987684682.00	1519482
Tăng trưởng doanh thu	-0.01	-0.01	0.10	0.01	0.04	-0.03	
Lợi nhuận gộp	6209690644985.00	6401445454290.00	7067518779284.00	5911521444565.00	6441612591490.00	6554900643375.00	615019
Tăng trưởng lợi nhuận gộp	-0.04	-0.02	0.15	0.10	0.10	0.03	
Lợi nhuận thuần từ HĐKD	2581919225696.00	2970748116715.00	3325773210703.00	2715537085273.00	2725822436482.00	3086270738943.00	277674

Hình 1: Bảng Chỉ số tài chính (5 dòng đầu và 5 quý gần nhất)

## Thống kê miêu tả của bảng Chỉ số tài chính :

	count	average	std	min	Q1	median	Q3
<b>CHI TIÊU</b>							
Doanh thu thuần	40.00	13756068660577.47	1871424029095.31	8716035913568.00	12424439584226.50	14015401621849.00	15481613283780.50
Tăng trưởng doanh thu	40.00	0.06	0.07	-0.07	0.01	0.05	0.10
Lợi nhuận gộp	40.00	6015609450533.85	883079667533.60	3108971976990.00	5574577547434.25	6186138028562.00	6618902131330.75
Tăng trưởng lợi nhuận gộp	40.00	0.09	0.16	-0.13	-0.02	0.05	0.16

Hình 2: Thống kê miêu tả của bảng Chỉ số tài chính (5 dòng đầu và một phần các chỉ số)

### 1.2.2 Lịch sử giá

Ta cũng thực hiện một vài bước xử lý đối với dữ liệu Lịch sử giá:

Listing 4: Tìm header row, tức dòng bắt đầu chứa dữ liệu thực

```

1 for i, row in df_price.iterrows():
2     if row.astype(str).str.contains("GIA MO CUA", case=False, na=False).any():
3         header_row = i
4         break
5
6 df_price = pd.read_excel("data\Simplize_VNM_PriceHistory_20250315.xlsx", header=header_row)
7 df_price.columns = [str(c).strip() for c in df_price.columns]
8 df_price = df_price.dropna(how='all')
```

Listing 5: Tìm header row, tức dòng bắt đầu chứa dữ liệu thực

```

1 # keeps valid dates
2 df_price = df_price[pd.to_datetime(df_price['NGAY'], dayfirst=True, errors='coerce').notna()]
3
4 # turn into datetime values
5 df_price['NGAY'] = pd.to_datetime(df_price['NGAY'], dayfirst=True)
```

## Bảng Lịch sử giá sau khi đã được xử lý :

	NGÀY	GIÁ MỞ CỬA	GIÁ CAO NHẤT	GIÁ THẤP NHẤT	GIÁ ĐÓNG CỬA	THAY ĐỔI GIÁ	% THAY ĐỔI	KHỐI LƯỢNG
0	2025-03-14	62400.00	62700.00	62100.00	62100.00	100.00	0.00	2696700.00
1	2025-03-13	62300.00	62900.00	61900.00	62000.00	-200.00	-0.00	5100400.00
2	2025-03-12	62500.00	62800.00	62200.00	62200.00	-200.00	-0.00	2755400.00
3	2025-03-11	62300.00	62500.00	62100.00	62400.00	-100.00	-0.00	2287600.00
4	2025-03-10	63000.00	63000.00	62500.00	62500.00	-200.00	-0.00	2529500.00

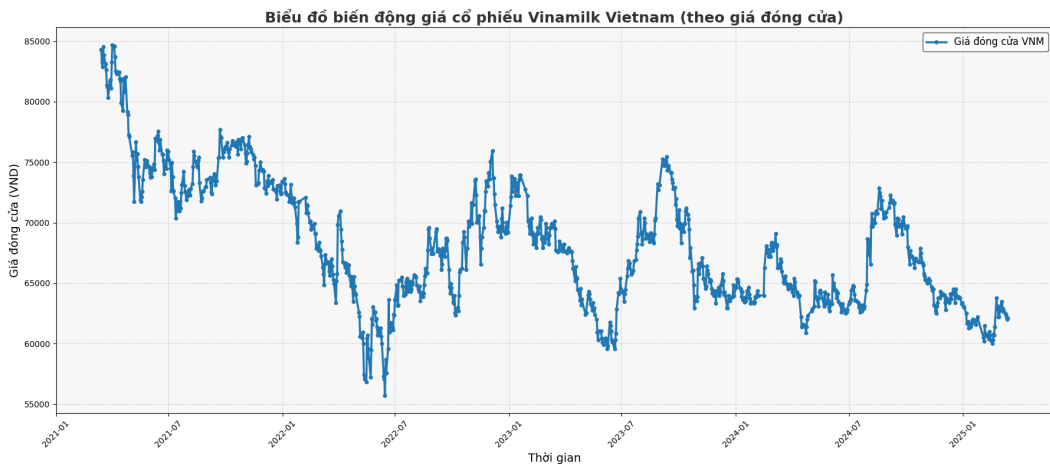
Hình 3: Bảng Lịch sử giá (5 dòng đầu)

### Thống kê miêu tả của bảng Lịch sử giá :

	count	average	min	Q1	median	Q3	max	std	variance
NGÀY	1000	2023-03-13 01:56:38.400000	2021-03-15 00:00:00	2022-03-14 18:00:00	2023-03-14 12:00:00	2024-03-13 06:00:00	2025-03-14 00:00:00	NaN	NaN
GIÁ MỞ CỬA	1000.00	68190.56	56214.81	64018.15	67571.00	72110.90	86174.49	5260.29	27670701.21
GIÁ CAO NHẤT	1000.00	68783.36	57423.73	64495.34	68099.15	72710.85	86174.49	5284.93	27930451.27
GIÁ THẤP NHẤT	1000.00	67603.14	55696.70	63611.65	67037.38	71477.98	84277.83	5192.61	26963152.25
GIÁ ĐÓNG CỬA	1000.00	68111.67	55696.70	63992.72	67571.00	72051.66	84690.15	5246.75	27528360.66
THAY ĐỔI GIÁ	547.00	-14.99	-2500.00	-500.00	-100.00	400.00	3900.00	811.30	658199.77
% THAY ĐỔI	547.00	-0.00	-0.03	-0.01	-0.00	0.01	0.06	0.01	0.00
KHOI LƯỢNG	1000.00	3131697.22	691300.00	1932900.00	2702950.00	3882825.00	21564900.00	1819471.43	3310476277282.17

Hình 4: Thống kê miêu tả của bảng Lịch sử giá

### 1.3 Theo dõi sự biến động giá cổ phiếu



Hình 5: Biểu đồ biến động giá cổ phiếu Vinamilk Vietnam (theo giá đóng cửa)

**Nhận xét:** Từ biểu đồ biến động giá cổ phiếu (5), ta nhận thấy xu hướng chung là giảm dài hạn. Từ đầu 2021 đến đầu 2025, giá cổ phiếu VNM giảm từ 82.000-85.000 VND xuống chỉ còn khoảng 62.000-65.000 VND. Đây là mức giảm khoảng 20-25% trong 5 năm - xu hướng giảm tương đối mạnh so với nhóm hàng tiêu dùng thiết yếu. Đặc biệt, ta thấy được các giai đoạn biến động nổi bật như sau:

#### 1. Năm 2021: Giai đoạn giảm mạnh nhất trong 5 năm

Đây là năm Vinamilk đối mặt với bão giá nguyên liệu, đặc biệt sữa bột và đường tăng tới 35% (theo báo cáo của Công ty chứng khoán VCBS). Lợi nhuận giảm, biên lãi gộp co hẹp mạnh, nhà đầu tư mất kỳ vọng, khiến giá giảm liên tục từ 82.000 xuống khoảng 72.000 rồi tiếp tục lao về 65.000. Biểu đồ cho thấy mức sụt giảm khủng khiếp trong 2021 là điểm đánh dấu sự chuyển pha từ tăng trưởng sang chững - giảm.

#### 2. 2022–2023: nỗ lực phục hồi và tái định vị thương hiệu

Có những nhịp hồi nhỏ (lên 75.000 vào đầu 2023), nhưng không giữ được vì nguyên liệu tiếp tục duy trì mức cao toàn cầu và cạnh tranh thị trường sữa tăng mạnh (đặc biệt từ Mộc Châu Milk, TH True Milk).

Tuy nhiên, vào khoảng tháng 7/2023, Vinamilk triển khai chiến dịch tái định vị thương hiệu - thời điểm ngay sau khi giá cổ phiếu chạm mức thấp nhất trong năm (dù vẫn chưa thấp bằng đáy cùng kỳ năm trước). Chiến dịch này tạo hiệu ứng tích cực ban đầu khi giá cổ phiếu hồi phục trở lại, nhưng mức tăng nhìn chung chỉ tương đương với diễn biến của cùng kỳ năm trước, chưa tạo ra sự bứt phá rõ rệt.

### 3. 2024–2025: Kế hoạch cải tổ 2026 chưa phản ánh vào giá

Vinamilk đặt mục tiêu đến năm 2026 đạt lợi nhuận 16.000 tỷ đồng và doanh thu 86.200 tỷ đồng, song kỳ vọng này vẫn chưa được thể hiện trên diễn biến giá cổ phiếu. Mặc dù công ty đã công bố chiến lược tái định vị và các bước tái cấu trúc, kết quả hiện tại vẫn cho thấy tốc độ cải thiện chậm, biên lợi nhuận chưa tăng mạnh, khiến giá cổ phiếu tiếp tục duy trì trạng thái ảm đạm.

Ngoài ra, một đặc điểm rất rõ trên biểu đồ là:

- **Tháng 7 hằng năm:** Giá thường chạm đáy.
- **Cuối năm – đầu tháng 1:** Giá thường đạt đỉnh.

Đây không phải là sự trùng hợp mà phản ánh *chu kỳ theo mùa* của ngành sữa và ngành hàng tiêu dùng nhanh (FMCG).

#### 1. Yếu tố mùa vụ trong tiêu thụ

Giai đoạn tháng 7–9 là mùa thấp điểm tiêu dùng: trẻ em nghỉ hè, chương trình sữa học đường tạm gián đoạn và nhu cầu các sản phẩm sữa giảm do thời tiết nóng. Điều này khiến doanh thu và lợi nhuận quý 3 thường thấp hơn các quý còn lại, kéo theo giá cổ phiếu có xu hướng giảm và tạo đáy trong thời gian này.

Ngược lại, cuối năm là mùa cao điểm tiêu thụ khi thời tiết lạnh hơn, nhu cầu dinh dưỡng tăng và sức mua phục vụ lễ Tết tăng mạnh. Nhờ đó, lợi nhuận của VNM được đẩy lên, khiến giá cổ phiếu thường tăng mạnh, và lập đỉnh vào tháng 12 hoặc tháng 1.

#### 2. Tác động của dòng tiền cuối năm

Cuối năm là thời điểm các quỹ đầu tư thực hiện chiến lược window dressing, ưu tiên gom các cổ phiếu phòng thủ như VNM để làm đẹp danh mục, tạo lực cầu hỗ trợ giá. Đồng thời, với đặc tính là cổ phiếu beta thấp, VNM thường được nhà đầu tư ưu tiên trong giai đoạn thị trường biến động, góp phần tăng cầu rõ rệt trong những tháng cuối năm.

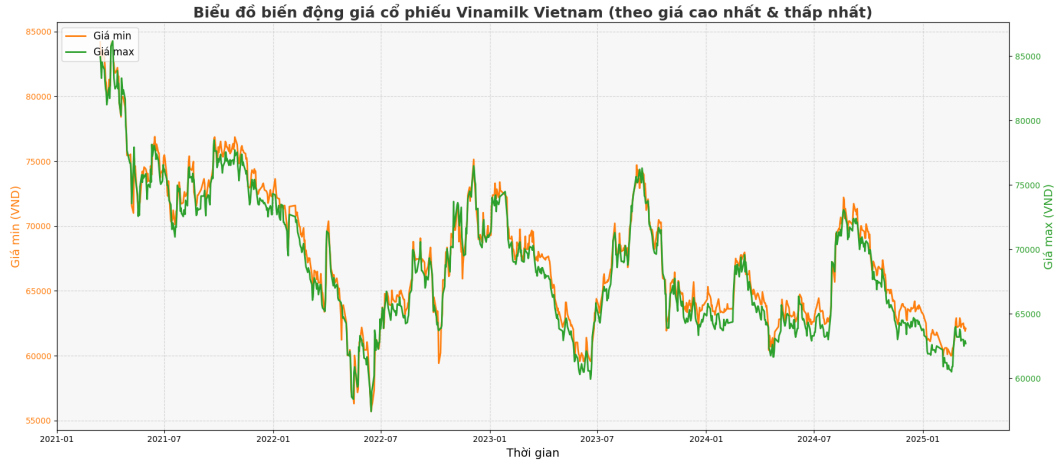
Việc phân tích giá cổ phiếu theo Giá nhỏ nhất và Giá lớn nhất trong ngày cũng cho ta kết quả tương tự (6)

## 2 Tiền xử lý dữ liệu

### 2.1 Bảng Chỉ số tài chính

Listing 6: Thực hiện lấy các chỉ số được yêu cầu

```
1 # get required indices
2 df_fin_mod = df_fin.loc[['Bien loi nhuan gop', 'Bien loi nhuan rong',
3                          'P/E', 'EPS (VND/CP)',
4                          'Tang truong EPS', 'ROE LTM', 'No phai tra /
5                          Von chu so huu',
6                          'Kha nang thanh toan tong quat', 'Vong quay
7                          tai san (vong)', 'Gia tri so sach (VND/
8                          CP)']]
9 df_fin_mod = df_fin_mod.iloc[:, :16]
```



Hình 6: Biểu đồ biến động giá cổ phiếu Vinamilk Vietnam (theo giá cao nhất và thấp nhất)

## 2.2 Bảng Lịch sử giá

Listing 7: Thực hiện lấy các chỉ số được yêu cầu

```
1 # get required indices
2 df_price_mod = df_price[['NGAY', 'GIA DONG CUA', 'THAY DOI GIA', '%
   THAY DOI']].copy()
3 df_price_mod['QUY'] = df_price_mod['NGAY'].dt.quarter
4 df_price_mod['NAM'] = df_price_mod['NGAY'].dt.year
5 df_price_mod[:5]
```

Listing 8: Tính trung bình các chỉ số theo từng quý

```
1 # get the last day of each quarter
2 last_days = df_price_mod.groupby(['NAM', 'QUY'])['NGAY'].max().
   reset_index()
3 last_days = last_days.sort_values(by=['NAM', 'QUY'], ascending=[False,
   False]).reset_index(drop=True)
4
5 average_values = []
6
7 for _, quarter_row in last_days.iterrows():
8     # get the average values for existing indices
9     end_day = quarter_row['NGAY']
10    window_mask = df_price_mod['NGAY'].between(end_day - pd.Timedelta(
        days=14), end_day + pd.Timedelta(days=14))
11    window_df = df_price_mod.loc[window_mask]
12
13    avg_close = window_df['GIA DONG CUA'].mean()
14    avg_change = window_df['THAY DOI GIA'].mean()
15    avg_pct_change = window_df['% THAY DOI'].mean()
16
17    average_values.append({
18        'NAM': quarter_row['NAM'],
19        'QUY': quarter_row['QUY'],
20        'Ngày cuối quý': end_day,
21        'Giá đóng cửa': avg_close,
22        'Thay đổi giá': avg_change,
23        '% Thay đổi': avg_pct_change
24    })
```

	NĂM	QUÝ	Ngày cuối quý	Giá đóng cửa	Thay đổi giá	% Thay đổi	QUÝ/NĂM
0	2025	1	2025-03-14	62581.82	-9.09	-0.00	Q1/2025
1	2024	4	2024-12-31	63009.81	-145.00	-0.00	Q4/2024
2	2024	3	2024-09-30	68946.07	-157.14	-0.00	Q3/2024
3	2024	2	2024-06-28	63419.11	-28.57	-0.00	Q2/2024
4	2024	1	2024-03-29	64743.29	-138.10	-0.00	Q1/2024
5	2023	4	2023-12-29	64141.01	-35.00	-0.00	Q4/2023
6	2023	3	2023-09-29	71375.98	-171.43	-0.00	Q3/2023
7	2023	2	2023-06-30	63729.72	333.33	0.00	Q2/2023
8	2023	1	2023-03-31	67674.94	-171.43	-0.00	Q1/2023
9	2022	4	2022-12-30	71131.20	360.00	0.00	Q4/2022
10	2022	3	2022-09-30	65269.36	NaN	NaN	Q3/2022
11	2022	2	2022-06-30	62667.03	NaN	NaN	Q2/2022
12	2022	1	2022-03-31	66900.81	NaN	NaN	Q1/2022
13	2021	4	2021-12-31	72629.58	NaN	NaN	Q4/2021
14	2021	3	2021-09-30	75974.19	NaN	NaN	Q3/2021
15	2021	2	2021-06-30	74468.89	NaN	NaN	Q2/2021
16	2021	1	2021-03-31	82612.85	NaN	NaN	Q1/2021

Hình 7: Bảng Lịch sử giá sau khâu tiền xử lý đầu

**Dự đoán các giá trị còn thiếu** Sau khi thống kê, nhóm nhận ra các giá trị Thay đổi giá và % Thay đổi từ Q1/2021 - Q3/2022 bị thiếu. Để đảm bảo tính nhất quán của dữ liệu, nhóm sử dụng phương pháp **Time Series Regression (dựa trên Autoregressive model AR(1))** để dự đoán giá trị thiếu. Theo đó, đây là một phương pháp trong phân tích chuỗi thời gian, trong đó giá trị hiện tại của biến phụ thuộc tuyến tính vào các giá trị trước đó trong chuỗi. Cụ thể, AR(1) giả thiết:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$$

trong đó  $y_t$  là giá đóng cửa tại thời điểm  $t$ ,  $\phi_0$  là hằng số,  $\phi_1$  là hệ số autoregressive bậc 1, và  $\epsilon_t$  là sai số ngẫu nhiên. Khi giá trị tại một quý bị thiếu, phương pháp này cho phép:

- **Forward filling:** Nếu giá quý trước đã có, sử dụng công thức AR(1) để ước lượng giá hiện tại.
- **Backward filling:** Nếu quý đầu tiên hoặc quý trước cũng thiếu, sử dụng giá quý sau để suy ngược lại theo hệ số AR(1).

Listing 9: Sử dụng Time Series Regression để dự đoán giá trị thiếu

```

1 from statsmodels.tsa.ar_model import AutoReg
2
3 # order date in ascending order
4 df_price_avg = df_price_avg.sort_values(["NAM", "QUY"], ascending=[True
5     , True]).reset_index(drop=True)
6
7 # fit AR(1) in the existing data
8 train = df_price_avg["Giá đóng cửa"].dropna()
9 ar_model = AutoReg(train, lags=1, old_names=False).fit()
10
11 # prediction function

```

```

11 def fill_missing_prices(series, model):
12     prices = series.copy()
13     for i in range(len(prices)):
14         if np.isnan(prices[i]):
15             # if i > 0 & previous GIA_fill exists -> forward AR(1)
16             if i > 0 and not np.isnan(prices[i-1]):
17                 prices[i] = model.params[0] + model.params[1]*prices[i-1]
18             # if i == 0 or previous GIA_fill NaN -> backward AR(1)
19             else:
20                 j = i + 1
21                 while j < len(prices) and np.isnan(prices[j]):
22                     j += 1
23                 if j < len(prices):
24                     prices[i] = (prices[j] - model.params[0])/model.params[1]
25     return prices
26
27 df_price_avg["Gia_fill"] = fill_missing_prices(df_price_avg["Gia dong
    cua"].values, ar_model)
28
29 # calculate
30 df_price_avg["Thay doi gia"] = df_price_avg["Gia_fill"].diff()
31 df_price_avg["% Thay doi"] = df_price_avg["Thay doi gia"] /
    df_price_avg["Gia_fill"].shift(1)

```

Lưu ý rằng Q1/2021 vẫn là NaN, và đây là điều bình thường vì không có dữ liệu quý trước để AR(1) có thể dự đoán (không thể forward hay backward filling cho điểm đầu tiên của chuỗi). Phương pháp này vẫn đảm bảo các giá trị còn lại trong chuỗi được điền một cách hợp lý, duy trì tính liên tục và tương quan tự nhiên theo thời gian.

	NĂM	QUÝ	Giá_fill	Thay đổi giá	% Thay đổi
0	2021	1	82612.85	NaN	NaN
1	2021	2	74468.89	-8143.96	-0.10
2	2021	3	75974.19	1505.30	0.02
3	2021	4	72629.58	-3344.61	-0.04
4	2022	1	66900.81	-5728.77	-0.08
5	2022	2	62667.03	-4233.77	-0.06
6	2022	3	65269.36	2602.33	0.04
7	2022	4	71131.20	5861.84	0.09
8	2023	1	67674.94	-3456.27	-0.05
9	2023	2	63729.72	-3945.21	-0.06
10	2023	3	71375.98	7646.25	0.12
11	2023	4	64141.01	-7234.96	-0.10
12	2024	1	64743.29	602.28	0.01
13	2024	2	63419.11	-1324.18	-0.02
14	2024	3	68946.07	5526.96	0.09
15	2024	4	63009.81	-5936.26	-0.09
16	2025	1	62581.82	-427.99	-0.01

Hình 8: Bảng Lịch sử giá sau khi đã được điền các giá trị còn thiếu



## 2.3 Bảng đã gộp

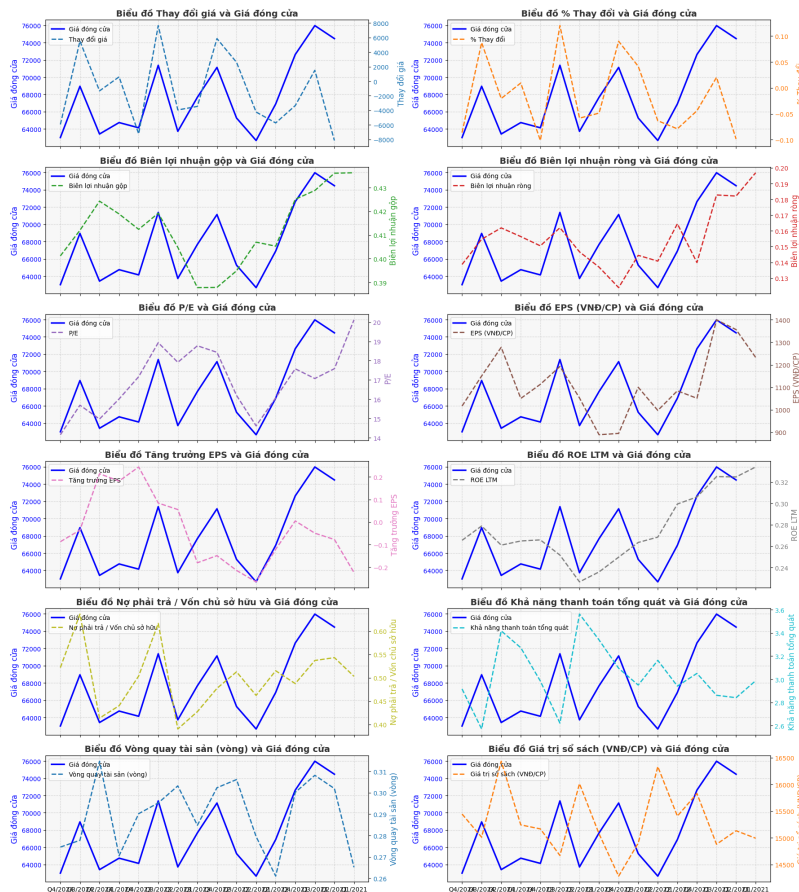
Sau đó, ta tiến hành gộp 2 bảng và được kết quả như sau:

	Q4/2024	Q3/2024	Q2/2024	Q1/2024	Q4/2023	Q3/2023	Q2/2023	Q1/2023	Q4/2022	Q3/2022	Q2/2022	Q1/2022
Thời gian												
Giá đóng cửa	63009.81	68946.07	63419.11	64743.29	64141.01	71375.98	63729.72	67674.94	71131.20	65269.36	62667.03	66900.81
Thay đổi giá	-5936.26	5526.96	-1324.18	602.28	-7234.96	7646.25	-3945.21	-3456.27	5861.84	2602.33	-4233.77	-5728.77
% Thay đổi	-0.09	0.09	-0.02	0.01	-0.10	0.12	-0.06	-0.05	0.09	0.04	-0.06	-0.08
Biên lợi nhuận gộp	0.40	0.41	0.42	0.42	0.41	0.42	0.40	0.39	0.39	0.39	0.41	0.41
Biên lợi nhuận ròng	0.14	0.15	0.16	0.16	0.15	0.16	0.15	0.14	0.12	0.14	0.14	0.16

Hình 9: Kết quả gộp hai bảng Chỉ số tài chính và Lịch sử giá (5 dòng đầu và 12 cột đầu)

## 3 Lựa chọn các yếu tố nguy cơ tiềm năng cho mô hình tài chính

### 3.1 Biểu đồ đường tương quan giữa giá đóng cửa và các chỉ số tài chính khác



Hình 10: Biểu đồ đường tương quan giữa giá đóng cửa và các chỉ số tài chính khác

**Nhận xét** Dựa trên tổng hợp 12 biểu đồ, có thể thấy giá cổ phiếu nhìn chung biến động cùng nhịp với các chỉ tiêu tài chính quan trọng, cụ thể:

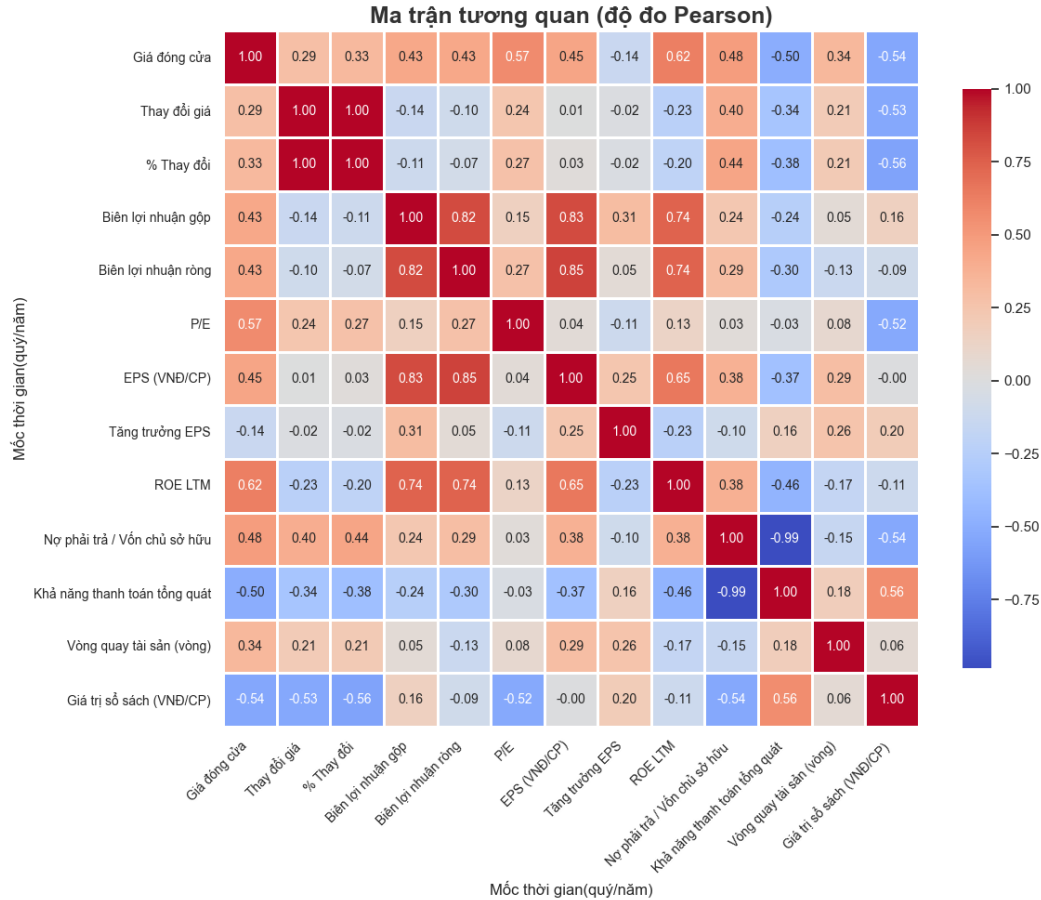
1. Các biểu đồ về thay đổi giá tuyệt đối và % thay đổi giá cho thấy giá cổ phiếu khá nhạy với các biến động ngắn hạn. Những chỉ tiêu như biên lợi nhuận gộp và biên lợi nhuận ròng có xu hướng cải thiện vào các kỳ giá cổ phiếu tăng mạnh, cho thấy nhà đầu tư đánh giá tích cực các yếu tố về hiệu quả hoạt động.
2. P/E tăng - giảm cùng nhịp với giá cổ phiếu, đặc biệt tăng mạnh khi giá tăng nhanh hơn tốc độ cải thiện lợi nhuận. Điều này phản ánh kỳ vọng thị trường được mở rộng theo giá. EPS và tăng trưởng EPS tăng trong các kỳ giá cổ phiếu đi lên, cho thấy lợi nhuận thực tế là động lực quan trọng cho đà tăng.
3. Chỉ số ROE LTM cũng có xu hướng cải thiện trong các kỳ giá tăng, cho thấy khả năng tạo lợi nhuận trên vốn chủ sở hữu là một yếu tố được nhà đầu tư quan tâm. Ngược lại, chỉ tiêu Nợ phải trả/Vốn chủ sở hữu thường gia tăng tại những giai đoạn giá suy giảm, phản ánh mức đòn bẩy cao hơn có thể làm thị trường trở nên thận trọng.
4. Các chỉ số về khả năng thanh toán tổng quát và vòng quay tài sản cũng diễn biến tích cực khi giá tăng. Cuối cùng, giá trị sổ sách đi lên từ từ, đồng hành với xu hướng giá cổ phiếu, thể hiện sự tăng trưởng ổn định của nền tảng tài sản.

Như vậy, **EPS – tăng trưởng lợi nhuận, ROE và biên lợi nhuận** là các yếu tố tác động mạnh nhất và đồng chiều với giá cổ phiếu. Trong khi đó, **tỷ lệ nợ/vốn chủ sở hữu** là yếu tố ảnh hưởng tiêu cực và có thể làm hạn chế đà tăng giá.

### 3.2 Ma trận tương quan giữa các chỉ số tài chính

Ma trận tương quan tính theo độ đo Pearson, ta có bảng tổng kết mức độ tương quan giữa giá cổ phiếu và các chỉ số tài chính như sau:

Bảng 1: Độ tương quan Pearson với Giá đóng cửa	
Chỉ số	Độ tương quan Pearson
Giá đóng cửa	1.00
ROE LTM	0.62
P/E	0.57
Nợ phải trả / Vốn chủ sở hữu	0.48
EPS (VNĐ/CP)	0.45
Biên lợi nhuận ròng	0.43
Biên lợi nhuận gộp	0.43
Vòng quay tài sản (vòng)	0.34
% Thay đổi	0.33
Thay đổi giá	0.29
Tăng trưởng EPS	-0.14
Khả năng thanh toán tổng quát	-0.50
Giá trị sổ sách (VNĐ/CP)	-0.54



Hình 11: Ma trận tương quan giữa các chỉ số tài chính

**Nhận xét** Dựa trên hệ số tương quan Pearson, giống như với việc phân tích biểu đồ đường, các yếu tố có khả năng ảnh hưởng mạnh nhất đến giá đóng cửa là ROE LTM và P/E. Điều này cho thấy thị trường phản ứng rất mạnh với khả năng sinh lời trên vốn và kỳ vọng định giá. Các chỉ tiêu khác như Nợ phải trả/Vốn chủ sở hữu, EPS, biên lợi nhuận ròng và gộp có tương quan dương trung bình, thể hiện mức độ ảnh hưởng nhưng không mạnh bằng ROE và P/E.

Ở chiều ngược lại, có ba chỉ số mang tương quan âm khá rõ: Giá trị sổ sách, Khả năng thanh toán tổng quát và Tăng trưởng EPS. Điều này cho thấy giá cổ phiếu tăng không nhất thiết đi kèm sự cải thiện trong các yếu tố an toàn tài chính hay tăng trưởng lợi nhuận ngắn hạn; thậm chí, trong một số giai đoạn, giá càng tăng thì các chỉ số này có xu hướng giảm.

Như vậy, giá cổ phiếu vận động chủ yếu theo **khả năng sinh lời** và **kỳ vọng thị trường**, hơn là theo các chỉ tiêu an toàn tài chính hay giá trị sổ sách.

## 4 Kết luận

Trong báo cáo này, nhóm đã thực hiện quá trình **hiểu dữ liệu** trên bộ dữ liệu tài chính và lịch sử giá cổ phiếu của Vinamilk (VNM). Thông qua các bước thống kê miêu tả, tiền xử lý dữ liệu và phân tích tương quan, nhóm đã rút ra một số kết luận quan trọng:

1. **Hiểu rõ dữ liệu và doanh nghiệp:** Phân tích các chỉ số tài chính cùng với lịch sử giá cổ phiếu giúp nhận thức được mối quan hệ giữa hiệu quả kinh doanh, sức khỏe tài chính của doanh nghiệp và biến động thị trường chứng khoán. Ví dụ, các chỉ số ROE, EPS và biên lợi nhuận gộp/ròng cho thấy tác động trực tiếp tới giá cổ phiếu.

2. Tiền xử lý và chuẩn hóa dữ liệu: Việc xác định header, loại bỏ các category row, dòng NaN, chuẩn hóa giá trị số và dự đoán giá trị thiếu đã minh họa cách làm sạch dữ liệu thực tế, tạo nền tảng để phân tích tiếp theo và xây dựng các mô hình dữ liệu.
3. Khám phá quan hệ giữa các biến: Qua các biểu đồ tương quan và ma trận Pearson, nhóm đã xác định những yếu tố tài chính quan trọng nhất ảnh hưởng đến giá cổ phiếu, bao gồm ROE LTM, P/E, biên lợi nhuận và tỷ lệ nợ/vốn chủ sở hữu. Điều này giúp sinh viên thấy rõ mối liên hệ giữa các biến định lượng trong dữ liệu thực tế.
4. Ứng dụng kiến thức chuyên ngành: Báo cáo đã kết hợp kiến thức về phân tích chuỗi thời gian (Time Series Regression - AR(1)) để dự đoán giá trị còn thiếu, minh họa khả năng ứng dụng các phương pháp data mining và thống kê cơ bản vào dữ liệu tài chính thực tế.

Như vậy, thông qua bài tập này, nhóm đã làm quen với quy trình **Data Understanding** và rèn luyện khả năng khai thác thông tin quan trọng từ dữ liệu tài chính thực tế, từ đó có thể áp dụng các phương pháp thống kê, khai thác dữ liệu và trực quan hóa để giải quyết các vấn đề thực tế trong lĩnh vực chứng khoán và phân tích tài chính.