. Phân tích tổng quan

Dữ liệu được phân tích:

- **Số điểm dữ liệu**: 1342 (cho mỗi cột), tương ứng với khoảng 1345 ngày giao dịch từ 2012-01-03 đến 2017-05-31 (trừ ngày nghỉ lễ/cuối tuần không có trong dữ liệu gốc).
- **Các chỉ số**:
  - **Trung bình** và **trung vị**: Cho biết giá trị trung tâm của dữ liệu.
  - **Độ lệch chuẩn**: Đo độ phân tán của dữ liệu.
  - **Skewness**: Đo độ lệch (dương: lệch phải; âm: lệch trái).
  - **Kurtosis**: Đo độ nhọn (cao: đuôi nặng; thấp: đuôi nhẹ).
  - **Shapiro-Wilk Test**: Kiểm tra tính chuẩn của phân phối (p-value < 0.05: không chuẩn).

Tóm tắt chung:

- **Tất cả các cột** (`FLC` , `HSG` , `KDC` , `PPC`) đều có **phân phối không chuẩn** (p-value = 0.0000 < 0.05 trong Shapiro-Wilk Test).
- **Skewness** khác nhau:
  - `FLC` và `HSG` : Lệch phải mạnh (skewness dương lớn).
  - `KDC` : Lệch phải nhẹ.
  - `PPC` : Lệch trái nhẹ.
- **Kurtosis** khác nhau:
  - `FLC` : Đuôi rất nặng (kurtosis cao).
  - `HSG` , `KDC` , `PPC` : Đuôi nhẹ hoặc trung bình.
- **Độ lệch chuẩn**:
  - `HSG` và `KDC` có độ lệch chuẩn cao, cho thấy biến động giá lớn.
  - `FLC` và `PPC` có độ lệch chuẩn thấp hơn, cho thấy biến động ít hơn.

---

2. Phân tích chi tiết từng cột

Cột FLC

- **Thống kê**:
  - Số điểm: 1342
  - Trung bình: 7.74
  - Trung vị: 7.05
  - Độ lệch chuẩn: 3.78
  - Skewness: 2.40 (lệch phải mạnh)
  - Kurtosis: 8.02 (đuôi rất nặng)
  - Shapiro-Wilk: Statistic=0.7849, p-value=0.0000 (không chuẩn)
- **Phân tích**:
  - **Trung bình > Trung vị**: Xác nhận phân phối lệch phải (giá trị lớn kéo trung bình lên).
  - **Skewness = 2.40**: Phân phối lệch phải mạnh, tức là có nhiều giá trị thấp (gần 7.05) và một số giá trị cao bất thường (có thể là các đỉnh giá).
  - **Kurtosis = 8.02**: Đuôi rất nặng, cho thấy nhiều giá trị ngoại lai hoặc các biến động cực đoan (giá tăng đột biến).
  - **Shapiro-Wilk**: p-value = 0.0000, xác nhận phân phối không chuẩn, phù hợp với skewness và kurtosis cao.
  - **Độ lệch chuẩn = 3.78**: Biến động giá vừa phải, nhưng các giá trị ngoại lai (đuôi nặng) làm tăng độ lệch chuẩn.
- **Liên hệ với ngoại lai**:

- **Z-score (36 ngoại lai, 2.68%)**: Z-score nhạy với các giá trị cực đoan trong phân phối đuôi nặng, giải thích tại sao nó phát hiện ngoại lai ở `FLC`. Các giá trị cao bất thường (do skewness và kurtosis cao) vượt ngưỡng Z-score (2.5).
- **IQR (0 ngoại lai)**: IQR dựa trên tứ phân vị, có thể không phát hiện ngoại lai vì khoảng `[Q1 - 1.5*IQR, Q3 + 1.5*IQR]` bao quát được nhiều giá trị lệch phải. Hệ số 1.5 có thể quá rộng với phân phối đuôi nặng của `FLC`.

Cột HSG

- **Thống kê**:
  - Số điểm: 1342
  - Trung bình: 12.29
  - Trung vị: 11.15
  - Độ lệch chuẩn: 7.12
  - Skewness: 0.89 (lệch phải vừa phải)
  - Kurtosis: 0.05 (đuôi gần chuẩn)
  - Shapiro-Wilk: Statistic=0.8764, p-value=0.0000 (không chuẩn)
- **Phân tích**:
  - **Trung bình > Trung vị**: Phân phối lệch phải, nhưng mức độ nhẹ hơn `FLC`.
  - **Skewness = 0.89**: Lệch phải vừa phải, cho thấy có một số giá trị cao (giá tăng) nhưng không quá cực đoan.
  - **Kurtosis = 0.05**: Gần phân phối chuẩn (kurtosis = 0), đuôi không quá nặng, ít giá trị cực đoan hơn `FLC`.
  - **Shapiro-Wilk**: p-value = 0.0000, xác nhận không chuẩn, có thể do skewness.
  - **Độ lệch chuẩn = 7.12**: Biến động giá lớn, phản ánh `HSG` (cổ phiếu ngành thép) có nhiều giai đoạn tăng/giảm mạnh.
- **Liên hệ với ngoại lai**:
  - **IQR (371 ngoại lai, 27.58%)**: IQR nhạy với phân phối lệch phải, phát hiện nhiều ngoại lai vì nhiều giá trị nằm ngoài `[Q1 - 1.5*IQR, Q3 + 1.5*IQR]`. Tỷ lệ cao (27.58%) cho thấy `HSG` có nhiều biến động giá lớn.
  - **Z-score (0 ngoại lai)**: Z-score không phát hiện ngoại lai, có thể vì phân phối lệch và độ lệch chuẩn lớn (7.12) làm giảm giá trị Z-score của các điểm bất thường, khiến chúng không vượt ngưỡng 2.5.

Cột KDC

- **Thống kê**:
  - Số điểm: 1342
  - Trung bình: 23.72
  - Trung vị: 23.12
  - Độ lệch chuẩn: 7.57
  - Skewness: 0.37 (lệch phải nhẹ)
  - Kurtosis: -0.17 (đuôi nhẹ hơn chuẩn)
  - Shapiro-Wilk: Statistic=0.9617, p-value=0.0000 (không chuẩn)
- **Phân tích**:
  - **Trung bình ≈ Trung vị**: Phân phối gần đối xứng, chỉ lệch phải nhẹ.
  - **Skewness = 0.37**: Lệch phải rất nhẹ, cho thấy phân phối khá cân bằng, ít giá trị cao bất thường.
  - **Kurtosis = -0.17**: Đuôi nhẹ hơn phân phối chuẩn (platykurtic), ít giá trị cực đoan.
  - **Shapiro-Wilk**: p-value = 0.0000, không chuẩn, nhưng statistic (0.9617) cao, cho thấy phân phối gần chuẩn hơn `FLC` và `HSG`.
  - **Độ lệch chuẩn = 7.57**: Biến động giá lớn, tương tự `HSG`, phản ánh tính chất cổ phiếu `KDC` (ngành thực phẩm).

- **Liên hệ với ngoại lai**:
  - **IQR (138 ngoại lai, 10.26%)**: IQR phát hiện ngoại lai do một số giá trị nằm ngoài khoảng tứ phân vị, dù phân phối gần đối xứng. Tỷ lệ 10.26% hợp lý với phân phối lệch nhẹ.
  - **Z-score (0 ngoại lai)**: Z-score không phát hiện ngoại lai, do phân phối gần chuẩn và độ lệch chuẩn lớn (7.57) làm các giá trị bất thường không vượt ngưỡng 2.5.

Cột PPC

- **Thống kê**:
  - Số điểm: 1342
  - Trung bình: 13.48
  - Trung vị: 14.32
  - Độ lệch chuẩn: 4.08
  - Skewness: -0.81 (lệch trái vừa phải)
  - Kurtosis: -0.42 (đuôi nhẹ hơn chuẩn)
  - Shapiro-Wilk: Statistic=0.8891, p-value=0.0000 (không chuẩn)
- **Phân tích**:
  - **Trung bình < Trung vị**: Phân phối lệch trái, tức là có nhiều giá trị cao (gần 14.32) và một số giá trị thấp bất thường.
  - **Skewness = -0.81**: Lệch trái vừa phải, cho thấy `PPC` có một số giai đoạn giá giảm mạnh.
  - **Kurtosis = -0.42**: Đuôi nhẹ hơn chuẩn, ít giá trị cực đoan.
  - **Shapiro-Wilk**: p-value = 0.0000, không chuẩn, do lệch trái.
  - **Độ lệch chuẩn = 4.08**: Biến động giá thấp hơn `HSG` và `KDC`, cho thấy `PPC` ổn định hơn.
- **Liên hệ với ngoại lai**:
  - **IQR (241 ngoại lai, 17.92%)**: IQR phát hiện nhiều ngoại lai, do các giá trị thấp (lệch trái) nằm ngoài `[Q1 - 1.5*IQR, Q3 + 1.5*IQR]`. Tỷ lệ 17.92% phản ánh biến động giá giảm.
  - **Z-score (0 ngoại lai)**: Z-score không phát hiện ngoại lai, do độ lệch chuẩn (4.08) và phân phối lệch làm các giá trị bất thường không vượt ngưỡng 2.5.

---

3. Giải thích sự khác biệt trong phát hiện ngoại lai (Z-score vs. IQR)

Dựa trên kết quả phân tích thống kê, sự khác biệt trong phát hiện ngoại lai giữa **Z-score** (chỉ `FLC`, 2.68%) và **IQR** (`HSG` 27.58%, `KDC` 10.26%, `PPC` 17.92%) có thể được giải thích như sau:

3.1. Z-score

- **Đặc điểm**:
  - Dựa trên trung bình và độ lệch chuẩn, giả định phân phối gần chuẩn.
  - Ngưỡng 2.5 loại bỏ ~2.5% dữ liệu nếu phân phối chuẩn.
- **Tại sao chỉ phát hiện ngoại lai ở `FLC`?**:
  - `FLC` có **skewness = 2.40** và **kurtosis = 8.02**, cho thấy phân phối lệch phải mạnh và đuôi rất nặng. Các giá trị cực đoan (giá cao bất thường) dễ vượt ngưỡng Z-score (2.5), dẫn đến 36 ngoại lai (2.68%).
  - `HSG`, `KDC`, `PPC` không có ngoại lai với Z-score, vì:
    - **Phân phối lệch**: Skewness (`HSG` 0.89, `KDC` 0.37, `PPC` -0.81) làm trung bình và độ lệch chuẩn không đại diện tốt, giảm giá trị Z-score của các điểm bất thường.
    - **Độ lệch chuẩn lớn**: `HSG` (7.12), `KDC` (7.57) có độ lệch chuẩn cao, khiến các giá trị bất thường không đủ xa trung bình để vượt ngưỡng 2.5.
- **Hạn chế**: Z-score kém hiệu quả với phân phối không chuẩn, đặc biệt khi dữ liệu lệch hoặc đuôi nhẹ (`HSG`, `KDC`, `PPC`).

3.2. IQR

- **Đặc điểm**:
  - Dựa trên tứ phân vị, không giả định phân phối chuẩn.
  - Ngưỡng `[Q1 - 1.5*IQR, Q3 + 1.5*IQR]` phát hiện các điểm ngoài khoảng trung tâm.
- **Tại sao phát hiện nhiều ngoại lai ở `HSG`, `KDC`, `PPC`?**:
  - `HSG` (skewness = 0.89, 371 ngoại lai, 27.58%): Lệch phải và biến động lớn (độ lệch chuẩn 7.12) khiến nhiều giá trị cao nằm ngoài khoảng IQR, đặc biệt trong các giai đoạn tăng giá mạnh (ngành thép 2016–2017).
  - `KDC` (skewness = 0.37, 138 ngoại lai, 10.26%): Lệch phải nhẹ, nhưng độ lệch chuẩn cao (7.57) và một số giá trị bất thường (giá cao/thấp) vượt ngưỡng IQR.
  - `PPC` (skewness = -0.81, 241 ngoại lai, 17.92%): Lệch trái, với nhiều giá trị thấp nằm ngoài khoảng IQR, phản ánh các giai đoạn giá giảm.
  - IQR nhạy với phân phối lệch (`HSG`, `PPC`) và các giá trị ngoài khoảng trung tâm, dẫn đến tỷ lệ ngoại lai cao.
- **Tại sao không phát hiện ngoại lai ở `FLC`?**:
  - Mặc dù `FLC` có skewness (2.40) và kurtosis (8.02) cao, khoảng `[Q1 - 1.5*IQR, Q3 + 1.5*IQR]` có thể bao quát được các giá trị lệch phải, do IQR dựa trên phần trung tâm dữ liệu.
  - Độ lệch chuẩn thấp (3.78) và phân phối đuôi nặng khiến các giá trị bất thường không đủ xa Q1/Q3 để được coi là ngoại lai.

3.3. So sánh Z-score và IQR

- **Z-score**:
  - Phù hợp với `FLC` do đuôi nặng (kurtosis = 8.02), phát hiện các giá trị cực đoan.
  - Không hiệu quả với `HSG`, `KDC`, `PPC` do phân phối lệch và độ lệch chuẩn lớn.
- **IQR**:
  - Phù hợp với `HSG`, `KDC`, `PPC` do phân phối lệch, phát hiện nhiều ngoại lai.
  - Không phát hiện ngoại lai ở `FLC`, có thể do hệ số 1.5 quá rộng với phân phối đuôi nặng.
- **Ngữ cảnh tài chính**:
  - `HSG` (thép): Biến động giá lớn, nhiều giai đoạn tăng mạnh, phù hợp với IQR.
  - `KDC` (thực phẩm): Biến động vừa phải, IQR phát hiện các giá trị bất thường nhẹ.
  - `PPC` (điện): Giá giảm trong một số giai đoạn, IQR nhạy với lệch trái.
  - `FLC` (bất động sản): Giá có một số đỉnh cao bất thường, phù hợp với Z-score.

---

4. Ý nghĩa của kết quả

1. **Phân phối không chuẩn**:
   - Tất cả các cột đều không chuẩn (Shapiro-Wilk p-value = 0.0000), phù hợp với dữ liệu tài chính (giá cổ phiếu thường lệch phải hoặc có đuôi nặng).
   - Điều này giải thích tại sao Z-score (giả định chuẩn) kém hiệu quả với `HSG`, `KDC`, `PPC`, trong khi IQR (không giả định chuẩn) phát hiện nhiều ngoại lai hơn.
2. **Skewness và Kurtosis**:
   - `FLC`: Lệch phải mạnh, đuôi nặng, dẫn đến các giá trị cực đoan phù hợp với Z-score.
   - `HSG`: Lệch phải vừa, đuôi gần chuẩn, IQR phát hiện nhiều giá trị cao bất thường.
   - `KDC`: Gần đối xứng, đuôi nhẹ, IQR phát hiện một số giá trị bất thường.
   - `PPC`: Lệch trái, đuôi nhẹ, IQR nhạy với các giá trị thấp.

3. **Tỷ lệ ngoại lai**:
    - IQR phát hiện tỷ lệ ngoại lai cao (`HSG` 27.58%, `PPC` 17.92%) có thể do hệ số 1.5 quá nghiêm ngặt với dữ liệu tài chính, coi nhiều biến động tự nhiên là ngoại lai.
    - Z-score phát hiện ít ngoại lai (`FLC` 2.68%) do ngưỡng 2.5 và phân phối lệch làm giảm độ nhạy.