1. Phân tích kết quả IQR (hệ số 3.0)

Kết quả IQR với hệ số 3.0 cho thấy số lượng và tỷ lệ ngoại lai tiếp tục giảm so với các hệ số trước đó:

Cột	IQR 1.5	IQR 2.0	IQR 2.5	IQR 3.0
FLC	0 (0%)	48 (3.58%)	38 (2.83%)	33 (2.46%)
HSG	371 (27.58%)	278 (20.72%)	217 (16.17%)	162 (12.07%)
KDC	138 (10.26%)	29 (2.16%)	1 (0.07%)	0 (0.00%)
PPC	241 (17.92%)	104 (7.75%)	0 (0.00%)	0 (0.00%)

- **FLC**:
  - Ngoại lai: 33 (2.46%)
  - **So sánh**: Giảm từ 38 (2.83%) với IQR hệ số 2.5, 48 (3.58%) với hệ số 2.0, và 36 (2.68%) với Z-score. Tỷ lệ 2.46% rất thấp, phù hợp với phân phối lệch phải mạnh (skewness = 2.40) và đuôi nặng (kurtosis = 8.02).
  - **Ý nghĩa**: `FLC` ổn định, chỉ có một số giá trị bất thường (giá tăng đột biến), dễ xử lý.
- **HSG**:
  - Ngoại lai: 162 (12.07%)
  - **So sánh**: Giảm từ 217 (16.17%) với IQR hệ số 2.5, 278 (20.72%) với hệ số 2.0, và 371 (27.58%) với hệ số 1.5. Z-score không phát hiện ngoại lai (0%), do phân phối lệch phải (skewness = 0.89) và độ lệch chuẩn lớn (7.12).
  - **Ý nghĩa**: Tỷ lệ 12.07% vẫn tương đối cao, nhưng đã cải thiện đáng kể. Biến động của `HSG` (thép) có thể phản ánh thị trường thép 2016–2017, nhưng vẫn cần đánh giá thêm.
- **KDC**:
  - Ngoại lai: 0 (0.00%)
  - **So sánh**: Giảm từ 1 (0.07%) với IQR hệ số 2.5, 29 (2.16%) với hệ số 2.0, và 138 (10.26%) với hệ số 1.5. Z-score không phát hiện ngoại lai (0%).
  - **Ý nghĩa**: Không có ngoại lai, `KDC` (thực phẩm) cực kỳ ổn định, lý tưởng cho phân tích.
- **PPC**:
  - Ngoại lai: 0 (0.00%)
  - **So sánh**: Không đổi so với IQR hệ số 2.5 (0%), giảm từ 104 (7.75%) với hệ số 2.0, và 241 (17.92%) với hệ số 1.5. Z-score không phát hiện ngoại lai (0%).
  - **Ý nghĩa**: Không có ngoại lai, `PPC` (điện) rất ổn định, phù hợp cho phân tích.

Tóm tắt:

- IQR hệ số 3.0 làm giảm đáng kể ngoại lai, đặc biệt ở `HSG` (12.07%), `KDC` (0%), và `PPC` (0%).
- `FLC` (2.46%) và `KDC` (0%) có tỷ lệ ngoại lai thấp hoặc không có, rất ổn định.
- `PPC` (0%) không có ngoại lai, ổn định hoàn toàn.
- `HSG` vẫn có tỷ lệ ngoại lai cao nhất (12.07%), cần đánh giá thêm với biến đổi log và ngữ cảnh.

---

2. Phân tích biến đổi log cho HSG

- **Skewness HSG_log**: -0.40
  - **So sánh**: Skewness gốc của `HSG` là 0.89 (lệch phải vừa phải). Sau biến đổi log, skewness giảm xuống -0.40 (lệch trái nhẹ), cho thấy phân phối trở nên đối xứng hơn.
  - **Ý nghĩa**: Biến đổi log (`np.log1p`) đã giảm độ lệch phải, làm dữ liệu `HSG` gần chuẩn hơn, phù hợp hơn cho các phương pháp phân tích giả định phân phối chuẩn.
- **HSG_log - IQR (hệ số 3.0)**: 67 ngoại lai (4.99%)

- **So sánh**: Giảm mạnh từ 162 (12.07%) với `HSG` gốc (IQR hệ số 3.0). Tỷ lệ 4.99% là rất thấp, cho thấy biến đổi log không chỉ giảm skewness mà còn giảm đáng kể số ngoại lai.
- **Ý nghĩa**: `HSG_log` ổn định hơn nhiều so với `HSG` gốc, có thể sử dụng thay thế trong phân tích để tránh loại bỏ cột.

---

3. Phân tích ma trận tương quan

Ma trận tương quan (đã cung cấp trước đó):

```text
          FLC       HSG       KDC       PPC
FLC  1.000000 -0.285902 -0.227923 -0.222676
HSG -0.285902  1.000000  0.904930  0.533702
KDC -0.227923  0.904930  1.000000  0.678608
PPC -0.222676  0.533702  0.678608  1.000000
```

- **FLC**: Tương quan thấp, âm với `HSG` (-0.286), `KDC` (-0.228), `PPC` (-0.223), cho thấy hành vi độc lập, mang giá trị đa dạng hóa.
- **HSG**: Tương quan rất cao với `KDC` (0.905), trung bình với `PPC` (0.534), có thể trùng lặp thông tin với `KDC`.
- **KDC**: Tương quan cao với `HSG` (0.905) và `PPC` (0.679), là trung tâm của nhóm tương quan.
- **PPC**: Tương quan vừa phải với `HSG` (0.534) và `KDC` (0.679), mang giá trị độc lập hơn `HSG`.

Ý nghĩa:

- Tương quan cao giữa `HSG` và `KDC` (0.905) cho thấy chúng có thể cung cấp thông tin tương tự. Nếu cần loại bỏ, `KDC` được ưu tiên do không có ngoại lai (0% với IQR hệ số 3.0).
- `FLC` độc lập, nên giữ lại.
- `PPC` ổn định, có tương quan vừa phải, nên giữ lại.

---

4. Đánh giá việc loại bỏ cột

Dựa trên **tỷ lệ ngoại lai (IQR hệ số 3.0)**, **HSG_log**, **tương quan**, **phân phối**, và **ý nghĩa tài chính**, tôi đánh giá từng cột:

Cột FLC

- **Ngoại lai**: 2.46% (thấp, ổn định).
- **Tương quan**: Thấp, độc lập (-0.286 với `HSG`, -0.228 với `KDC`, -0.223 với `PPC`).
- **Phân phối**: Lệch phải mạnh (skewness = 2.40), đuôi nặng (kurtosis = 8.02), nhưng tỷ lệ ngoại lai thấp, dễ xử lý (nội suy trong `VNINDEX_iqr_cleaned.csv`).
- **Ý nghĩa tài chính**: `FLC` (bất động sản) đại diện cho phân khúc quan trọng, hành vi độc lập là giá trị.
- **Kết luận**: **Giữ lại**, vì tỷ lệ ngoại lai thấp, độc lập, và ý nghĩa tài chính.

Cột HSG

- **Ngoại lai**: 12.07% (vẫn cao, dù giảm từ 27.58%, 20.72%, 16.17%).
- **HSG_log**: Skewness = -0.40, ngoại lai 4.99% (rất thấp, ổn định).
- **Tương quan**: Rất cao với `KDC` (0.905), trung bình với `PPC` (0.534), có thể trùng lặp thông tin.
- **Phân phối**: Lệch phải vừa (skewness = 0.89), đuôi gần chuẩn (kurtosis = 0.05), biến động lớn (độ lệch chuẩn = 7.12). Sau log, phân phối gần chuẩn hơn.
- **Ý nghĩa tài chính**: `HSG` (thép) phản ánh ngành biến động mạnh, quan trọng trong phân tích thị trường 2012–2017 (giá thép tăng 2016–2017).
- **Kết luận**: **Không loại bỏ, sử dụng HSG_log thay thế**:
  - Tỷ lệ ngoại lai 12.07% với `HSG` gốc vẫn cao, nhưng `HSG_log` giảm xuống 4.99% và skewness cải thiện (-0.40), cho thấy dữ liệu ổn định hơn.

- Tương quan cao với `KDC` (0.905) là rủi ro trùng lặp, nhưng `HSG` mang giá trị ngành thép.
- **Hành động**: Sử dụng `HSG_log` (cột sau biến đổi log) thay vì `HSG` gốc trong phân tích để giảm ngoại lai và cải thiện phân phối. Không cần loại bỏ cột.

Cột KDC

- **Ngoại lai**: 0.00% (không có, rất ổn định).
- **Tương quan**: Rất cao với `HSG` (0.905), khá cao với `PPC` (0.679).
- **Phân phối**: Gần đối xứng (skewness = 0.37), đuôi nhẹ (kurtosis = -0.17), lý tưởng.
- **Ý nghĩa tài chính**: `KDC` (thực phẩm) đại diện cho ngành ổn định, giá trị cao (trung bình = 23.72).
- **Kết luận**: **Giữ lại**, vì không có ngoại lai, phân phối tốt, và ý nghĩa tài chính.

Cột PPC

- **Ngoại lai**: 0.00% (không có, rất ổn định).
- **Tương quan**: Trung bình với `HSG` (0.534), khá cao với `KDC` (0.679).
- **Phân phối**: Lệch trái vừa (skewness = -0.81), đuôi nhẹ (kurtosis = -0.42), chấp nhận được.
- **Ý nghĩa tài chính**: `PPC` (điện) đại diện cho ngành ổn định, biến động thấp (độ lệch chuẩn = 4.08).
- **Kết luận**: **Giữ lại**, vì không có ngoại lai, tương quan vừa phải, và ý nghĩa tài chính.

---

5. Kết luận cuối cùng

- **Không loại bỏ cột nào**:
  - Tất cả các cột (`FLC` , `HSG` , `KDC` , `PPC`) đều có ý nghĩa tài chính, đại diện cho các ngành quan trọng (bất động sản, thép, thực phẩm, điện).
  - `FLC` (2.46%), `KDC` (0%), `PPC` (0%) có tỷ lệ ngoại lai thấp hoặc không có, rất ổn định.
  - `HSG` có tỷ lệ ngoại lai giảm từ 27.58% (IQR 1.5) xuống 12.07% (IQR 3.0), và `HSG_log` chỉ còn 4.99% ngoại lai với skewness cải thiện (-0.40).
- **Giải pháp cho HSG**:
  - Sử dụng **HSG_log** (cột sau biến đổi log) thay vì `HSG` gốc trong phân tích:
    - Tỷ lệ ngoại lai 4.99% là chấp nhận được.
    - Skewness -0.40 cho thấy phân phối gần chuẩn hơn, phù hợp cho mô hình hóa.
  - Lý do: Biến đổi log giữ được giá trị tài chính của `HSG` (ngành thép) mà không cần loại bỏ cột, đồng thời giảm biến động bất thường.