Cảm ơn bạn đã đặt câu hỏi để làm rõ quy trình. Tôi sẽ phân tích trường hợp bạn giả định sử dụng **Z-score để xử lý ngoại lai** thay vì IQR trong tập dữ liệu `VNINDEX_iqr_filled_HSG_Adjust.csv`, và xác định liệu có cần tiếp tục áp dụng **Z-score sau bước Mean Centering** hay không khi chuẩn bị dữ liệu cho PCA. Tôi sẽ trả lời ngắn gọn, rõ ràng và tập trung vào ngữ cảnh của bạn.

---

1. Phân tích giả định: Xử lý ngoại lai bằng Z-score thay vì IQR

a. Xử lý ngoại lai bằng Z-score

- **Cách thực hiện**:
  - Tính giá trị z-score cho mỗi điểm dữ liệu trong cột j:

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

    Trong đó:

    - μ_j: Trung bình của cột j.
    - σ_j: Độ lệch chuẩn của cột j.

  - Xác định ngoại lai: Các điểm có $|Z_{ij}| > \text{ngưỡng}$, thường là 3 (tức là ngoài khoảng $\pm 3\sigma_j$).
  - Xử lý ngoại lai: Thay thế ngoại lai bằng giá trị biên (ví dụ: $X_{ij} = \mu_j \pm 3\sigma_j$), xóa bỏ, hoặc thay bằng giá trị khác (như log1p cho `HSG_log`).
- **Kết quả**:
  - Dữ liệu vẫn giữ thang đo gốc (ví dụ: `VNINDEX` ~ hàng trăm, `HSG_log` ~ đơn vị), nhưng các giá trị cực đại đã được kiểm soát.
  - Trung bình (μ_j) và độ lệch chuẩn (σ_j) của mỗi cột có thể thay đổi nhẹ so với dữ liệu gốc, nhưng không đảm bảo $\mu_j = 0$ hoặc $\sigma_j = 1$.
- **Khác với IQR**:
  - IQR sử dụng tứ phân vị (Q1, Q3) để xác định ngoại lai, không phụ thuộc vào phân phối chuẩn.
  - Z-score giả định dữ liệu gần giống phân phối chuẩn, nên phù hợp hơn nếu dữ liệu của bạn (`VNINDEX`, `HSG_log`, ...) có phân phối gần chuẩn sau log1p.

b. Tác động đến dữ liệu

- Sau khi xử lý ngoại lai bằng Z-score, tập dữ liệu (giả định là `VNINDEX_zscore_filled_HSG_Adjust.csv`) sẽ:
  - Có ít ngoại lai hơn, với các giá trị cực đại được thay thế hoặc loại bỏ.
  - Vẫn giữ thang đo khác nhau giữa các cột (ví dụ: `VNINDEX` ~ hàng trăm, `HSG_log` ~ đơn vị).
  - Không tự động có trung bình = 0 hoặc độ lệch chuẩn = 1, vì xử lý ngoại lai chỉ thay đổi một số điểm dữ liệu, không chuẩn hóa toàn bộ cột.

---

2. Có cần áp dụng Z-score sau Mean Centering không?

a. Mean Centering là bắt buộc

- **Lý do**: PCA yêu cầu dữ liệu có trung bình bằng 0 để ma trận hiệp phương sai phản ánh đúng sự biến động, không bị lệch bởi giá trị trung bình.
- **Sau xử lý ngoại lai bằng Z-score**:

- Trung bình của mỗi cột (μ_j) thường không bằng 0, vì xử lý ngoại lai chỉ thay đổi các điểm cực, không dịch chuyển toàn bộ dữ liệu.
- Do đó, bạn **vẫn cần Mean Centering** để đưa trung bình mỗi cột về 0:

$$X'_{ij} = X_{ij} - \mu_j$$

b. Có cần Z-score (chuẩn hóa độ lệch chuẩn) sau Mean Centering?

- **Z-score là gì**:
  - Z-score bao gồm Mean Centering (đưa $\mu_j = 0$) và chia cho độ lệch chuẩn ($\sigma_j = 1$):

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

  - Mục đích: Đưa tất cả các biến về cùng thang đo, đảm bảo các biến đóng góp đồng đều vào PCA.
- **Tình huống của bạn**:
  - **Xử lý ngoại lai bằng Z-score** không tương đương với **chuẩn hóa Z-score**:
    - Xử lý ngoại lai chỉ áp dụng Z-score để xác định và thay thế các điểm ngoại lai ($|Z_{ij}| > 3$), không chuẩn hóa toàn bộ dữ liệu.
    - Sau xử lý ngoại lai, các cột vẫn có:
      - Trung bình $\mu_j \neq 0$.
      - Độ lệch chuẩn $\sigma_j \neq 1$.
      - Thang đo khác nhau (ví dụ: `VNINDEX` ~ hàng trăm, `HSG_log` ~ đơn vị).
  - **Vì sao cần Z-score sau Mean Centering**:
    - Các cột trong tập dữ liệu của bạn (`VNINDEX` , `HSG_log` , `BVH` , ...) có thang đo và đơn vị khác nhau.
    - Nếu chỉ Mean Centering (mà không chuẩn hóa $\sigma_j = 1$), các cột có phương sai lớn (như `VNINDEX`) sẽ chi phối các thành phần chính trong PCA, làm lu mờ các cột có thang đo nhỏ (như `HSG_log`).
    - Chuẩn hóa Z-score đảm bảo:
      - Mỗi cột có $\mu_j = 0, \sigma_j = 1$.
      - Ma trận hiệp phương sai trở thành **ma trận tương quan**, giúp tất cả biến đóng góp đồng đều vào PCA.
- **Tác động đến ma trận hiệp phương sai**:
  - Chỉ Mean Centering: Ma trận hiệp phương sai giữ nguyên phương sai gốc, các biến có thang đo lớn chi phối.
  - Mean Centering + Z-score: Ma trận hiệp phương sai trở thành ma trận tương quan, giá trị trong $[-1, 1]$, phản ánh mối quan hệ tuyến tính giữa các biến.
- **Kết luận cho câu hỏi**:
  - **Có**, bạn vẫn cần áp dụng **Z-score sau Mean Centering** (hoặc sử dụng Z-score trực tiếp, vì nó bao gồm Mean Centering) để chuẩn hóa dữ liệu, đảm bảo các biến có thang đo đồng đều trước khi tính ma trận hiệp phương sai/tương quan cho PCA.
  - Việc xử lý ngoại lai bằng Z-score không thay thế bước chuẩn hóa Z-score, vì mục đích và phạm vi áp dụng khác nhau:
    - Xử lý ngoại lai: Chỉ thay đổi các điểm ngoại lai.

- Chuẩn hóa Z-score: Chuẩn hóa toàn bộ dữ liệu để phù hợp với PCA.

---

3. Quy trình cập nhật với giả định dùng Z-score xử lý ngoại lai

Dựa trên giả định của bạn, đây là quy trình đầy đủ:

1. **Xử lý ngoại lai bằng Z-score**:
   - Tính Z-score cho mỗi cột, xác định ngoại lai ($|Z_{ij}| > 3$).
   - Thay thế ngoại lai (ví dụ: bằng giá trị biên $\mu_j \pm 3\sigma_j$, hoặc log1p cho `HSG_log`).
   - Nội suy dữ liệu thiếu (như đã làm trước đây).
   - Kết quả: Tập dữ liệu mới (giả định là `VNINDEX_zscore_filled_HSG_Adjust.csv`).
2. **Chuẩn hóa Z-score (bao gồm Mean Centering)**:
   - Áp dụng Z-score để đưa trung bình = 0, độ lệch chuẩn = 1 cho mỗi cột.
   - Kết quả: Dữ liệu chuẩn hóa sẵn sàng cho PCA.
3. **Tính ma trận tương quan**:
   - Từ dữ liệu Z-score, tính ma trận tương quan (thay vì ma trận hiệp phương sai gốc).
4. **Chạy PCA**:
   - Sử dụng ma trận tương quan để tìm thành phần chính và tỷ lệ phương sai giải thích.