1. Xác nhận: Đã xử lý ngoại lai với IQR, có cần Mean Centering và Z-score cho dữ liệu không?

Đáp án: **Có**, bạn vẫn cần thực hiện **cả Mean Centering và Z-score** cho dữ liệu trong `VNINDEX_iqr_filled_HSG_Adjust.csv` sau khi đã xử lý ngoại lai với IQR. Lý do:

- **Mean Centering**:
  - **Bắt buộc** cho PCA: PCA dựa trên ma trận hiệp phương sai, yêu cầu dữ liệu phải có trung bình bằng 0 để đảm bảo ma trận hiệp phương sai chỉ phản ánh sự biến động, không bị lệch bởi giá trị trung bình của các cột.
  - Với tập dữ liệu của bạn: Các cột như `VNINDEX`, `HSG_log`, `BVH`, ... có trung bình khác 0, nên mean centering là cần thiết để đưa trung bình của mỗi cột về 0.
- **Z-score (:
  - **Được khuyến nghị**: Z-score (bao gồm mean centering và chia cho độ lệch chuẩn σ_j) đưa các biến về cùng thang đo (trung bình = 0, độ lệch chuẩn = 1). Điều này đặc biệt quan trọng với tập dữ liệu của bạn vì:
    - Các cột có thang đo khác nhau (ví dụ: `VNINDEX` ~ hàng trăm, `HSG_log` ~ đơn vị, `KDC`, `VNM` ở các phạm vi khác nhau).
    - Nếu không áp dụng z-score, 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`).
  - Z-score đảm bảo tất cả các biến đóng góp đồng đều vào PCA, giúp các thành phần chính phản ánh mối quan hệ tuyến tính giữa các biến, thay vì bị chi phối bởi thang đo.
- **Sau xử lý IQR**:
  - Xử lý ngoại lai với IQR (hệ số 3.0, đã áp dụng cho `HSG_log` và các cột khác) giúp loại bỏ hoặc giảm ảnh hưởng của các giá trị cực đại, làm trung bình và độ lệch chuẩn ổn định hơn.
  - Tuy nhiên, IQR không thay thế mean centering hoặc z-score, vì:
    - IQR chỉ xử lý ngoại lai, không đảm bảo trung bình = 0 hay độ lệch chuẩn = 1.
    - Mean centering và z-score là các bước chuẩn hóa cần thiết để chuẩn bị dữ liệu cho PCA.
- **Kết luận cho câu 1**:
  - Bạn cần thực hiện **Mean Centering** để đưa trung bình mỗi cột về 0.
  - Bạn nên thực hiện **Z-score** để chuẩn hóa độ lệch chuẩn về 1, đặc biệt vì các cột trong `VNINDEX_iqr_filled_HSG_Adjust.csv` có thang đo khác nhau.
  - Quy trình: **IQR (đã xong) → Mean Centering → Z-score**.

---

2. Xác nhận: Sau khi thực hiện Mean Centering và Z-score, có tính ma trận hiệp phương sai không?

Đáp án: **Có**, sau khi thực hiện Mean Centering và Z-score, bạn sẽ tính **ma trận hiệp phương sai** (thực chất là ma trận tương quan trong trường hợp z-score) để sử dụng trong PCA. Cụ thể:

- **Quy trình trong PCA**:
  1. **Chuẩn hóa dữ liệu**:
     - Mean Centering: Đưa trung bình mỗi cột về 0.
     - Z-score: Đưa độ lệch chuẩn mỗi cột về 1.
  2. **Tính ma trận hiệp phương sai (hoặc ma trận tương quan)**:

    - Nếu chỉ mean centering: Ma trận hiệp phương sai giữ nguyên phương sai gốc của các biến.
    - Nếu áp dụng z-score: Ma trận hiệp phương sai trở thành **ma trận tương quan**, vì:

$$\varrho_{jk} = \frac{\mathrm{Cov}(X_j, X_k)}{\sigma_j \sigma_k}$$

      Trong đó ϱ_{jk} là hệ số tương quan giữa biến j và k.
3. **Áp dụng PCA**: Sử dụng ma trận hiệp phương sai/tương quan để tìm các thành phần chính (eigenvectors) và giá trị riêng (eigenvalues).

- **Trường hợp của bạn**:
    - Vì bạn áp dụng **z-score** (như khuyến nghị ở câu 1), ma trận hiệp phương sai sẽ là **ma trận tương quan**.
    - Ma trận tương quan phù hợp hơn cho tập dữ liệu của bạn, vì:
        - Các cột (`VNINDEX` , `HSG_log` , `BVH` , ...) có thang đo khác nhau.
        - Ma trận tương quan đảm bảo các biến đóng góp đồng đều vào PCA, tránh hiện tượng các cột có thang đo lớn (như `VNINDEX`) chi phối các thành phần chính.
    - Ma trận tương quan được tính từ dữ liệu đã chuẩn hóa z-score:

$$C'_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} Z_{ij} Z_{ik}, \quad \text{với} \quad Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

- **Kết luận cho câu 2**:
    - Sau khi thực hiện Mean Centering và Z-score, bạn sẽ tính **ma trận tương quan** (một dạng đặc biệt của ma trận hiệp phương sai) từ dữ liệu đã chuẩn hóa.
    - Ma trận này sẽ được sử dụng trong PCA để tìm các thành phần chính và tỷ lệ phương sai giải thích.

Tóm tắt quy trình

Dựa trên xác nhận của bạn, đây là các bước cần thực hiện với `VNINDEX_iqr_filled_HSG_Adjust.csv` :

1. **Đã hoàn thành**:
    - Xử lý ngoại lai với IQR (hệ số 3.0).
    - Nội suy dữ liệu thiếu.
    - Thay `HSG` bằng `HSG_log` (log1p).
    - Kết quả: `VNINDEX_iqr_filled_HSG_Adjust.csv` .
2. **Cần thực hiện**:
    - **Mean Centering và Z-score**:
        - Chuẩn hóa dữ liệu để mỗi cột có trung bình = 0, độ lệch chuẩn = 1.
    - **Tính ma trận tương quan**:
        - Sử dụng dữ liệu đã chuẩn hóa z-score để tính ma trận tương quan (thay vì ma trận hiệp phương sai gốc).
    - **Chạy PCA**:
        - Áp dụng PCA trên ma trận tương quan để tìm thành phần chính.