4. Kết luận

Có nên chuyển hệ số IQR từ 1.5 lên 2.5 hoặc 3.0?

Có, nên chuyển sang IQR 3.0, ưu tiên sử dụng HSG_log:

- **Lý do**:
  - IQR 1.5 dẫn đến quá nhiều ngoại lai (`HSG` : 27.58%), làm nội suy tuyến tính thay đổi bản chất dữ liệu (mất biến động ngắn hạn, ý nghĩa tài chính).
  - IQR 3.0 giảm ngoại lai xuống 12.07% (`HSG`), và `HSG_log` chỉ còn 4.99%, bảo toàn dữ liệu gốc, giữ xu hướng ngành thép.
  - Phân phối `HSG_log` (skewness = -0.40) gần đối xứng hơn, phù hợp cho phân tích, dù không chuẩn (p-value = 0.0000).
- **Hành động đề xuất**:
  1. **Sử dụng IQR 3.0 với HSG_log**:

```python
import pandas as pd
import numpy as np
df = pd.read_csv('6.VNINDEX.csv', parse_dates=[0], index_col=0)
df['HSG_log'] = np.log1p(df['HSG'])
Q1 = df['HSG_log'].quantile(0.25)
Q3 = df['HSG_log'].quantile(0.75)
IQR = Q3 - Q1
outliers = df['HSG_log'][(df['HSG_log'] < Q1 - 3.0*IQR) | (df['HSG_log'] > Q3 + 3.
print(f"HSG_log - IQR 3.0: {len(outliers)} ngoại lai ({len(outliers)/1342*100:.2f}
```

  2. **Thay thế ngoại lai bằng nội suy tuyến tính**:
     - Chỉ thay thế 67 điểm (4.99%) của `HSG_log` , giảm tác động so với 371 điểm (27.58%) của `HSG` .

```python
df['HSG_log_cleaned'] = df['HSG_log'].where(
    (df['HSG_log'] >= Q1 - 3.0*IQR) & (df['HSG_log'] <= Q3 + 3.0*IQR),
    df['HSG_log'].interpolate(method='linear')
)
df.to_csv('VNINDEX_HSG_log_cleaned.csv')
```

  3. **Kiểm tra tác động**:
     - So sánh `HSG` gốc, `HSG_log` , và `HSG_log_cleaned` :

```python
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
plt.plot(df['HSG'], label='HSG gốc')
plt.plot(np.expm1(df['HSG_log']), label='HSG_log (chuyển ngược)')
plt.plot(np.expm1(df['HSG_log_cleaned']), label='HSG_log sau nội suy')
```

```
plt.title('So sánh HSG trước và sau xử lý')
plt.legend()
plt.grid(True)
plt.savefig('hsg_compare_log_cleaned.png')
plt.close()
```

4. **Đánh giá ngữ cảnh**:
   - Kiểm tra ngày ngoại lai của `HSG` gốc (IQR 1.5) để xác nhận ý nghĩa tài chính:

```python
Q1 = df['HSG'].quantile(0.25)
Q3 = df['HSG'].quantile(0.75)
IQR = Q3 - Q1
outliers_dates = df['HSG'][(df['HSG'] < Q1 - 1.5*IQR) | (df['HSG'] > Q3 + 1.5*IQR)
print("Ngày có ngoại lai của HSG (IQR 1.5):", outliers_dates)
```

   - Nếu ngoại lai tập trung vào 2016–2017, cân nhắc giữ một số điểm bằng IQR 3.0.

- **Lưu ý**:
  - Kết quả 0 ngoại lai (`DatetimeIndex([])`) có thể do dữ liệu đã nội suy hoặc hệ số IQR lớn. Xác nhận mã kiểm tra để đảm bảo không bỏ sót ngoại lai.
  - Phân phối không chuẩn (`HSG_log` p-value = 0.0000) không ảnh hưởng lớn, vì các mô hình như ARIMA không yêu cầu chuẩn.

---

5. Trích dẫn

- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.