School of Applied Mathematics

# TỔNG QUAN VỀ CHUỖI THỜI GIAN

TS. Nguyễn Mạnh Hùng

Đại học Giao thông Vận tải, 2021

School of Applied Mathematics

# NỘI DUNG

- Giới thiệu về các thư viện của Python
- Dự báo và chuỗi thời gian
- Các thuộc tính của chuỗi thời gian
- Cơ sở thống kê cho dự báo

School of Applied Mathematics

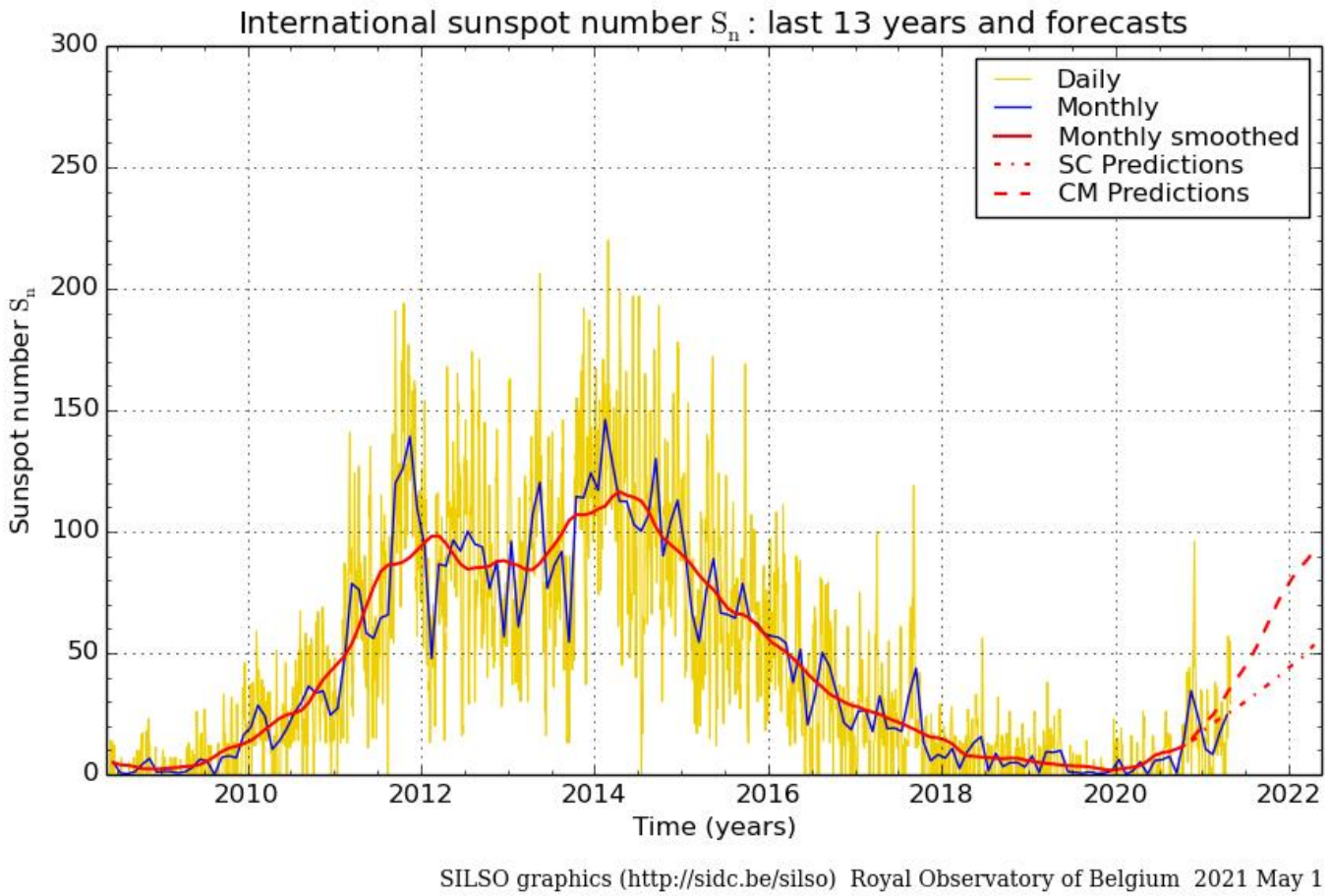
# Giới thiệu về các thư viện Python

- numpy
- pandas
- matplotlib
- scikit-learn

Ghi chú: học trên Jupyter Notebook

School of Applied Mathematics

# Dự báo số điểm đen mặt trời



# Dự báo

**Dự báo là tiên đoán về sự xuất hiện của một sự kiện nào đó trong tương lai**

- Diễn ra trong tất cả các lĩnh vực
- Phân loại:
  - Ngắn hạn (ngày, tuần, tháng)
  - Trung hạn (1-2 năm)
  - Dài hạn (nhiều năm)
- Dự báo ngắn/trung hạn: xác định vấn đề, mô hình hóa và suy luận về các hình mẫu trong dữ liệu quá khứ.

# Chuỗi thời gian

**Một tập các quan sát về giá trị của một biến nhận được tại những thời điểm khác nhau**

- Dữ liệu để dự báo thường có "*quán tính*" và không biến đổi quá nhanh.
- Dữ liệu thường được thu thập tại các thời điểm cách đều nhau:
  - Hàng ngày (chứng khoán, hàng hóa, nhiệt độ)
  - Hàng tuần (cung tiền)
  - Hàng tháng, quý, năm, … (tỉ lệ thất nghiệp, chỉ số lạm phát)

School of Applied Mathematics

# Biểu đồ: Tỉ giá VND/USD

# Ứng dụng của phân tích chuỗi thời gian

- **Quản trị vận hành**: lên kế hoạch sản xuất, tồn kho, logistic, nhân sự, năng lực sản xuất, …
- **Marketing**: quảng cáo, khuyến mại, thay đổi biểu giá …
- **Tài chính và quản trị rủi ro**: quản lý danh mục đầu tư …
- **Quản lý kinh tế**: định hướng chính sách tiền tệ
- **Điều khiển các quá trình công nghiệp**: thay đổi, dừng sản xuất, đại tu máy móc, …
- **Điều tra nhân khẩu học**: lập kế hoạch chính sách, dịch vụ xã hội.

School of Applied Mathematics

School of Applied Mathematics

# Các kỹ thuật dự báo phổ biến

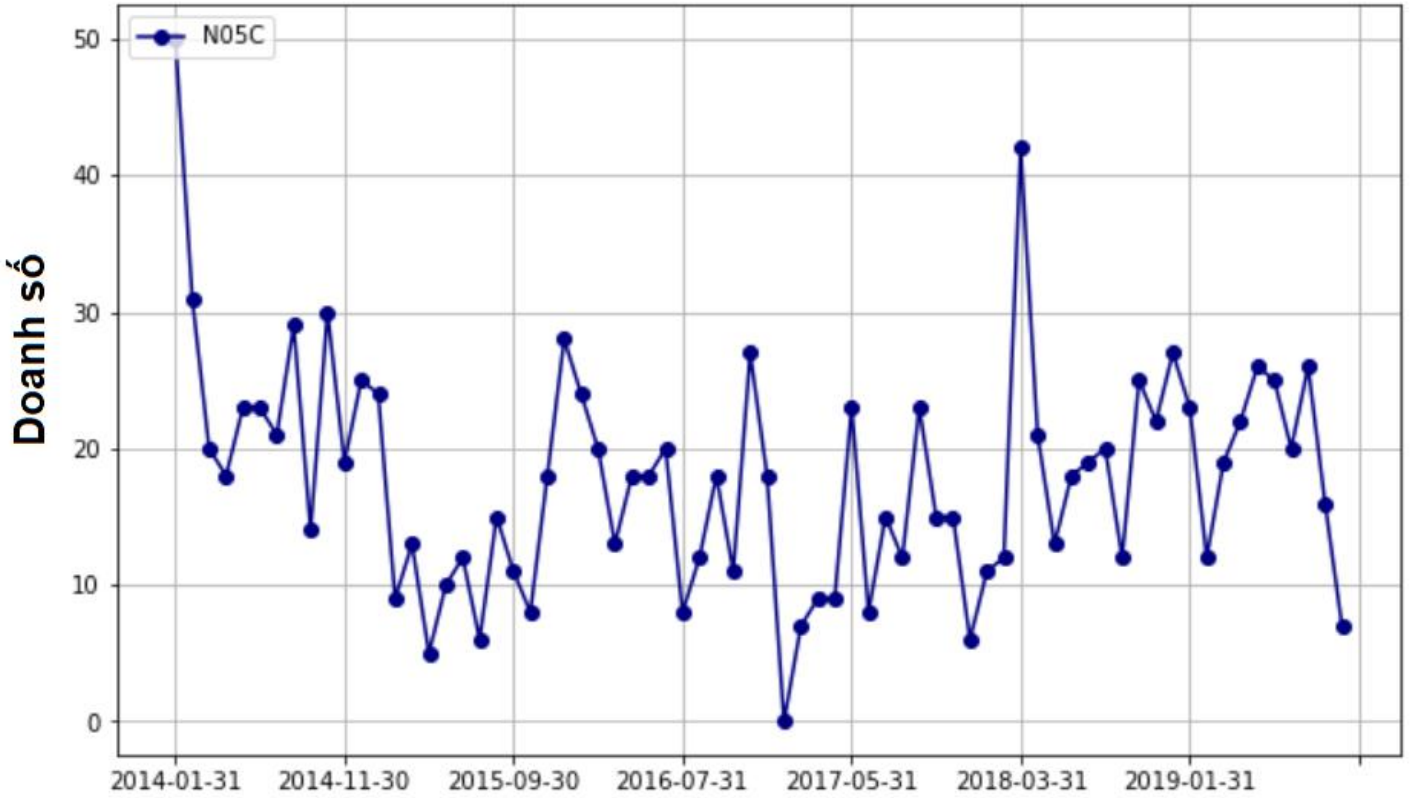
1. **Phương pháp định tính**
   - ❖ Đánh giá từ chuyên gia
   - ❖ Có ít hoặc không có dữ liệu làm cơ sở (Ví dụ: phát triển sản phẩm mới, tham khảo ý kiến nhân viên tiếp thị)
2. **Phương pháp định lượng**
   - ❑ Sử dụng dữ liệu quá khứ và một mô hình dự báo
   - ❑ Mô hình hồi quy, mô hình trơn, mô hình chuỗi thời gian

# Các thuộc tính của chuỗi thời gian

- Ngẫu nhiên
- Tự tương quan
- Chuyển động chu kỳ
- Mùa
- Dừng và không dừng
- Xu thế
- Biến động

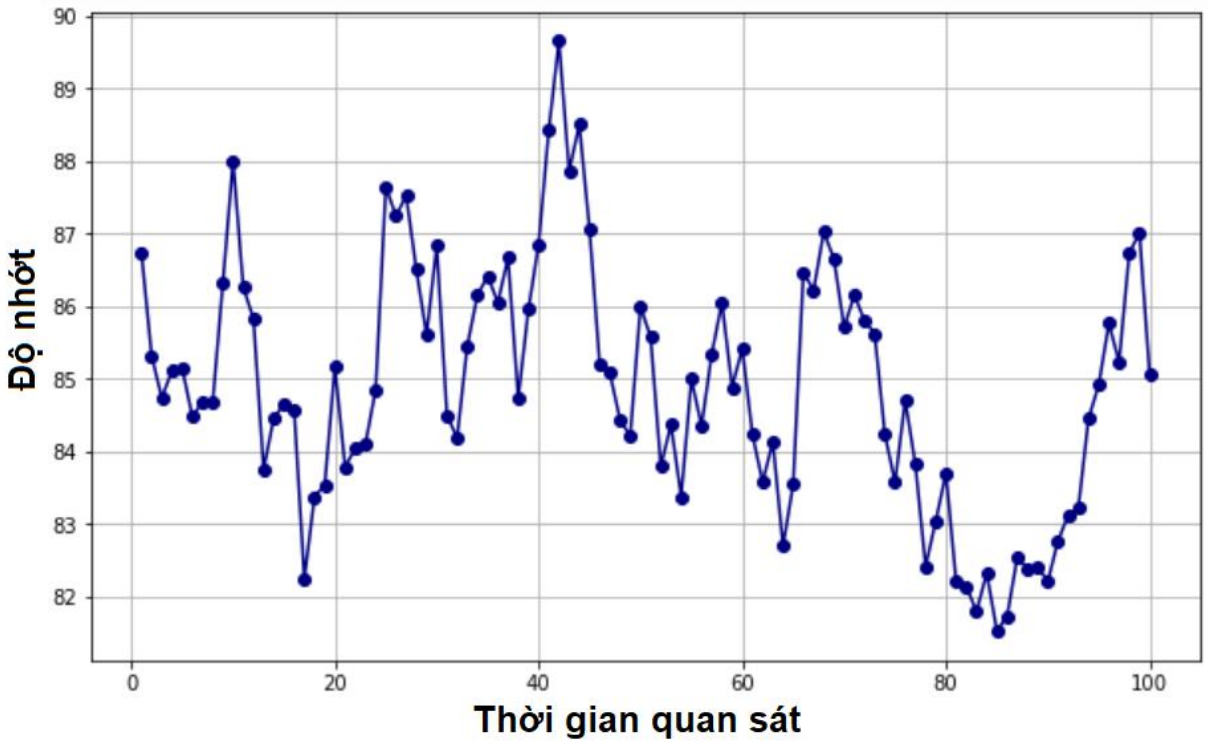
# Ngẫu nhiên

- Doanh số bán thuốc “N05C” theo tháng được thể hiện trên hình
- Không có một hình mẫu rõ ràng cho sự biến động doanh số.



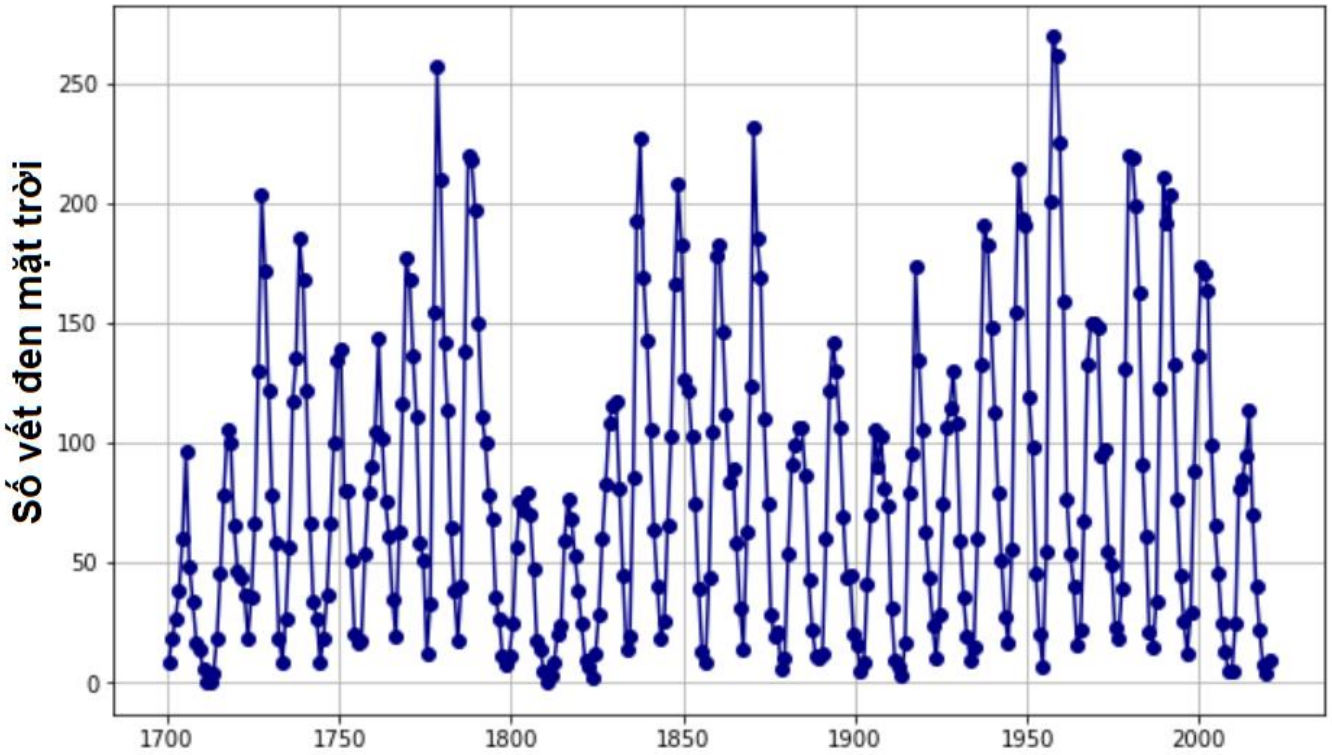
# Tự tương quan

- Giá trị của biến có tương quan hay có xu hướng biến thiên theo các giá trị khác của chuỗi.
- Do bản chất liên tục của quá trình sản xuất, độ nhớt có xu hướng đạt mức trung bình dài hạn là 85 (cP).



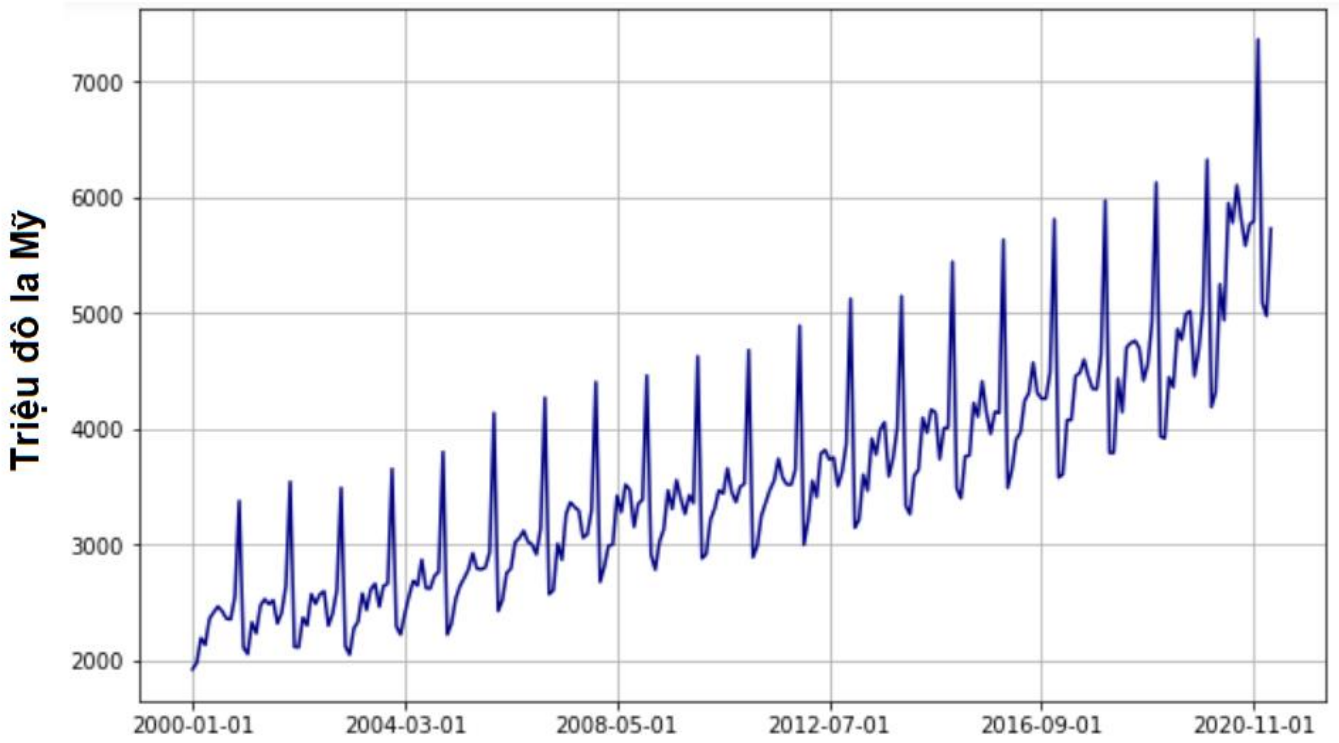
# Chuyển động chu kỳ

- Trong biểu đồ, cứ sau khoảng 11 năm thì số vết đen mặt trời lại lập đỉnh.



# Mùa

- Dữ liệu được quan sát theo tháng, theo quý có hình mẫu được lặp lại.
- Dữ liệu về doanh số bán lẻ rượu bia: quý 1 thấp, quý 4 cao.



# Dừng và không dừng

- Chuỗi thời gian dao động quanh một giá trị cố định là một đặc trưng của chuỗi dừng.
- Biểu đồ tỉ giả Eur/Usd thể hiện tính dừng.

# Xu thế

- Xu thế được hiểu là sự chuyển động đơn điệu lên hoặc xuống trong dài hạn.
- Biểu đồ về sự tiêu thụ rượu cho thấy xu hướng tăng mạnh sau 1950.

# Biến động

- Phần trăm thay đổi tỉ giá có biến động khác nhau theo thời gian.
- Brexit (2016): biến động mạnh
- Phương sai không đổi là một đặc trưng của chuỗi dừng.

School of Applied Mathematics

# Cơ sở thống kê cho dự báo

- ❑ Biểu diễn dữ liệu bằng đồ thị
- ❑ Các mô tả số học
- ❑ Biến đổi dữ liệu và hiệu chỉnh
- ❑ Phương pháp mô hình hóa chuỗi thời gian
- ❑ Các tiêu chuẩn đánh giá mô hình

# Vẽ biểu đồ chuỗi thời gian

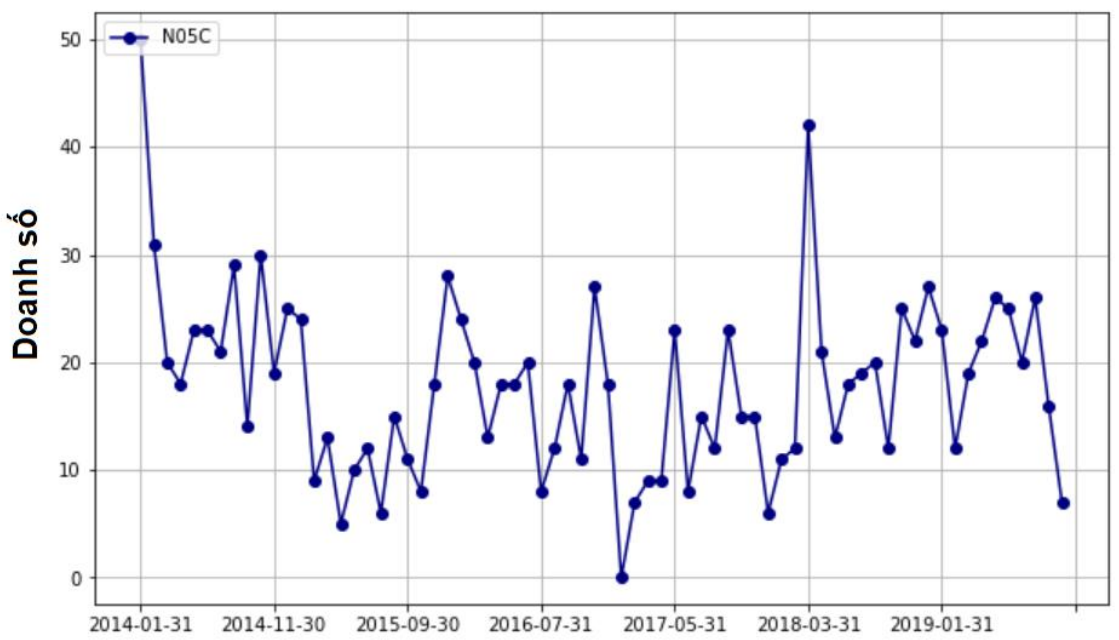
**Sử dụng thư viện Python**

```
import matplotlib.pyplot as plt
import pandas as pd
```

**Đọc dữ liệu từ tệp**

```
sales = pd.read_csv("salesmonthly.csv")
sales.head()
```

| | datum | M01AB | M01AE | N02BA | N02BE | N05B | N05C |
|---|---|---|---|---|---|---|---|
| 0 | 2014-01-31 | 127.69 | 99.090 | 152.100 | 878.030 | 354.0 | 50.0 |
| 1 | 2014-02-28 | 133.32 | 126.050 | 177.000 | 1001.900 | 347.0 | 31.0 |
| 2 | 2014-03-31 | 137.44 | 92.950 | 147.655 | 779.275 | 232.0 | 20.0 |

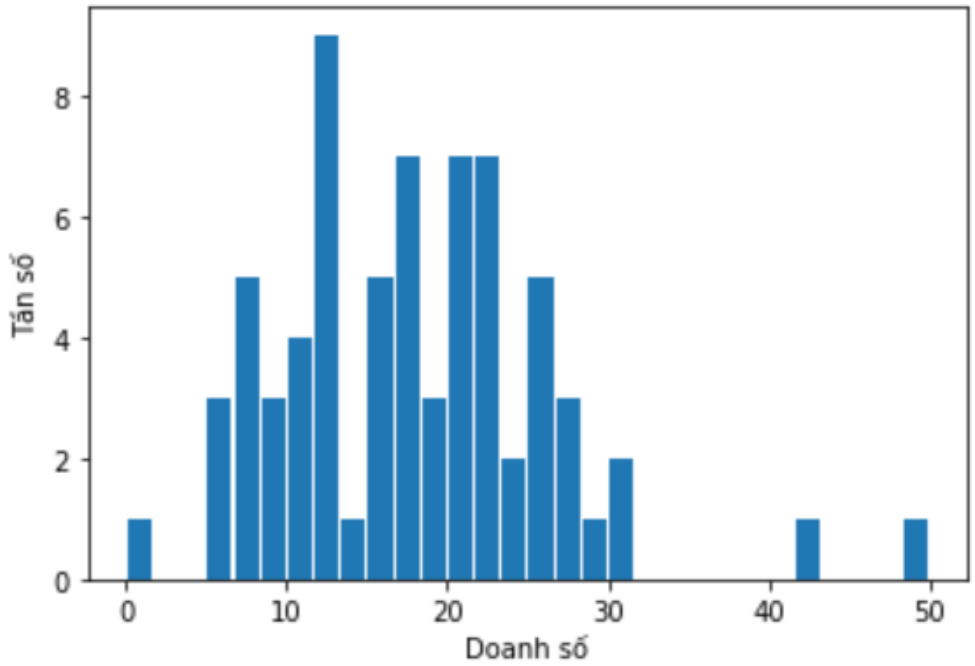



**Vẽ biểu đồ**

```
ax = sales.plot(x='datum', y='N05C', color='navy', style='-o', figsize=(10, 6))
ax.set_xlabel('Thời gian')
ax.set_ylabel('Doanh số')
#ax.set_title('Dữ liệu về doanh số bán thuốc theo tháng')
ax.grid(True)
ax.legend(loc='upper left');
plt.show()
```

# Vẽ biểu đồ histogram

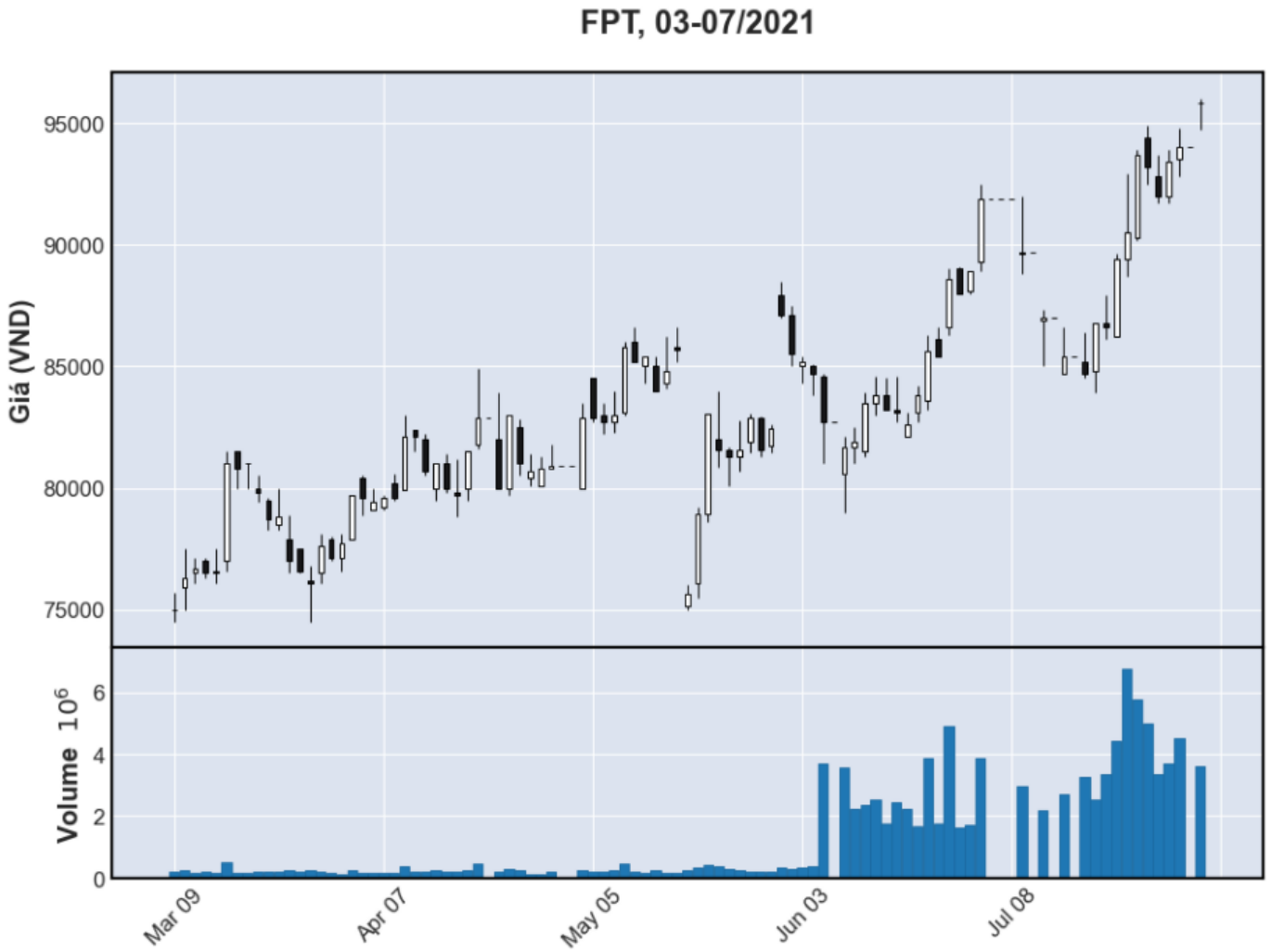
```
ax = sales.N05C.hist(bins=30,rwidth=0.9, figsize=(8, 6))
ax.set_xlabel('Doanh số')
ax.set_ylabel('Tần số')
ax.grid(False)
plt.show()
```



# Vẽ biểu đồ nến

| Date | High | Low | Open | Close | Volume |
|---|---|---|---|---|---|
| 2021-03-09 | 75700.0 | 74500.0 | 75000.0 | 75000.0 | 218240.0 |
| 2021-03-10 | 77500.0 | 75000.0 | 75900.0 | 76300.0 | 260240.0 |
| 2021-03-11 | 77100.0 | 76100.0 | 76500.0 | 76700.0 | 147780.0 |
| 2021-03-12 | 77100.0 | 76300.0 | 77000.0 | 76500.0 | 206280.0 |
| 2021-03-15 | 77500.0 | 76100.0 | 76600.0 | 76500.0 | 176090.0 |
| ... | ... | ... | ... | ... | ... |
| 2021-07-28 | 93700.0 | 91700.0 | 92800.0 | 92000.0 | 3346700.0 |
| 2021-07-29 | 93900.0 | 91700.0 | 92000.0 | 93400.0 | 3712900.0 |
| 2021-07-30 | 94800.0 | 92800.0 | 93500.0 | 94000.0 | 4525900.0 |
| 2021-08-02 | 94000.0 | 94000.0 | 94000.0 | 94000.0 | 0.0 |
| 2021-08-03 | 96000.0 | 94700.0 | 95800.0 | 95800.0 | 3607200.0 |

99 rows × 6 columns



```
# Cài đặt thư viện
! pip install pandas_datareader
! pip install mplfinance
```

```
# Sử dụng thư viện
import datetime as dt
import pandas_datareader as web
import mplfinance as fplt
```

```
# Tải dữ liệu
start = dt.datetime(2021,3,10)
end = dt.datetime.now()
df = web.DataReader('FPT', 'yahoo', start, end)
df
```

```
# Vẽ biểu đồ nến
fplt.plot(  df,
            type='candle',
            volume=True,
            title='FPT, 03-07/2021',
            ylabel='Giá (VND)',
            figscale=1.5    )
```

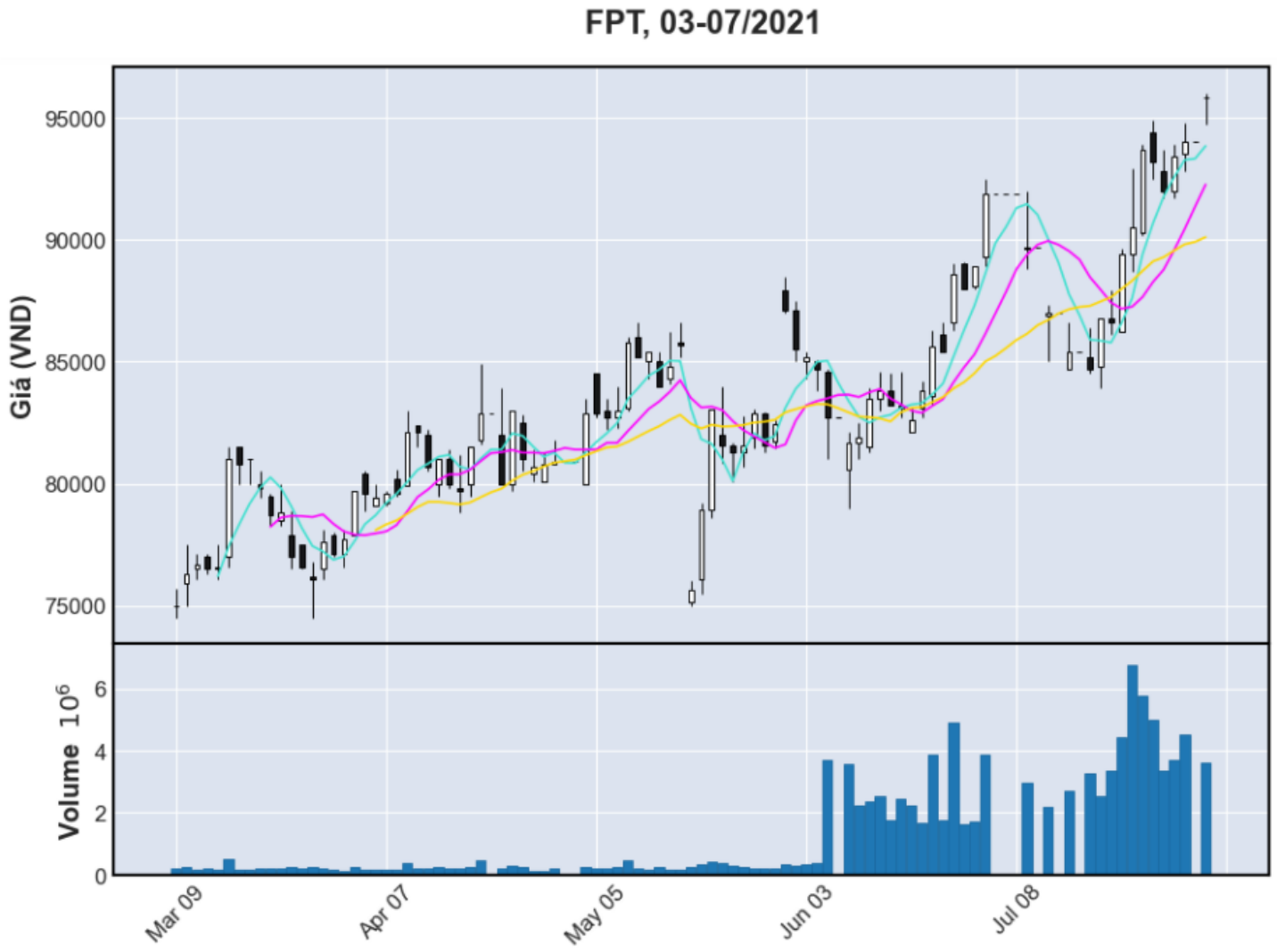
# Đường trung bình trượt (MA)

Giá trị $M_T$ là trung bình trượt tại thời điểm $T$ của $N$ quan sát gần nhất $y_T, y_{T-1}, \ldots, y_{T-N+1}$ được xác định bởi:

$$M_T = \frac{y_T + y_{T-1} + \cdots + y_{T-N+1}}{\mathrm{N}}$$

Nếu $\mathrm{Var}(y_t) = \sigma^2$ thì $\mathrm{Var}(M_T) = \frac{\sigma^2}{N}$

School of Applied Mathematics

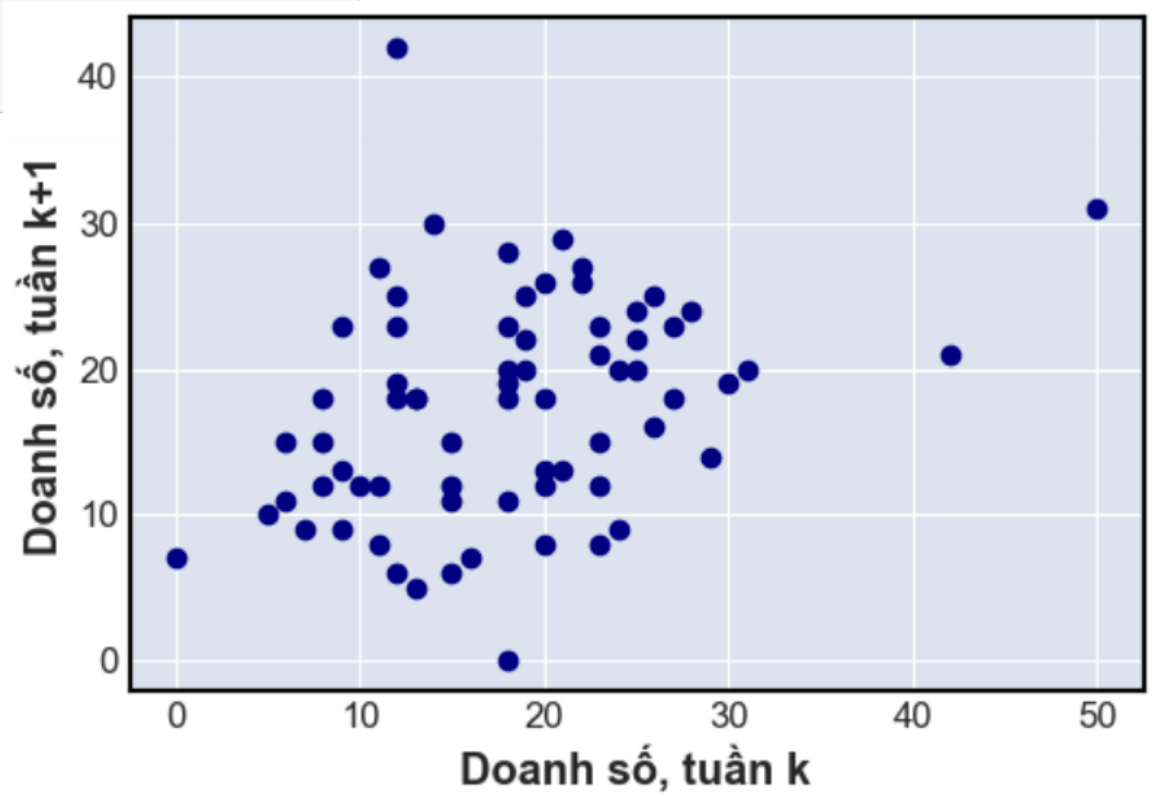



**Vẽ biểu đồ nến với các đường trung bình trượt**

```
fplt.plot(  df,
            type='candle',
            volume=True,
            title='FPT, 03-07/2021',
            ylabel='Giá (VND)',
            figscale=1.5,
            mav=(5,10,20)          # đường MA5, MA10, MA50
         )
```

# Các mô tả số học

```
sales = pd.read_csv("salesmonthly.csv")
plt.figure(figsize=(6,4))
plt.scatter(sales.N05C[:-1],sales.N05C[1:], color='navy')
plt.xlabel('Doanh số, tuần k')
plt.ylabel('Doanh số, tuần k+1')
plt.show()
```




Các hàm

- Tự hiệp phương sai
- Tự tương quan

School of Applied Mathematics

# Tự hiệp phương sai và tự tương quan

- Hiệp phương sai giữa quan sát $y_t$ và $y_{t+k}$ được gọi là **tự hiệp phương sai** với **độ trễ $k$** của chuỗi thời gian $y_t$:

$$\gamma_k = Cov(y_t, y_{t+k}) = E[(y_t - \mu)(y_{t+k} - \mu)]$$

$$\gamma_0 = Cov(y_t, y_{t+0}) = Var(y_t) = \sigma_y^2$$

- Hệ số **tự tương quan** với **độ trễ $k$** của chuỗi $y_t$ là:

$$\rho_k = \frac{Cov(y_t, y_{t+k})}{\sqrt{Var(y_t).Var(y_{t+k})}} = \frac{\gamma_k}{\gamma_0}$$

School of Applied Mathematics

# Hàm tự tương quan

- Tập hợp các giá trị $\rho_k$ với $k = 0,1,2, \ldots$ được gọi là hàm tự tương quan (**A**uto**C**orrelaltion **F**unction – **ACF**).
- Chú ý rằng:

$$\rho_0 = 1$$

$$\rho_k = \rho_{-k}$$

School of Applied Mathematics

# Ước lượng ACF

Giả sử các giá trị quan sát về chuỗi gồm: $y_1, y_2, \dots, y_T$. Khi đó:

$$\gamma_k \approx c_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})\,(y_{t+k} - \bar{y})$$

$$\rho_k \approx r_k = \frac{c_k}{c_0} \quad ; \quad se(r_k) \approx \frac{1}{\sqrt{T}}$$

với $k = 0,1,2, \dots, K$. Thông thường, số quan sát tối thiểu phải bằng 50 và $K < T/4$.

# Ví dụ

Sử dụng dữ liệu trong file '*salesmonthly.csv*' để tính hệ số tự tương quan mẫu tại độ trễ $k = 1$:

$$c_0 = \frac{1}{T}\sum_{t=1}^{T-0}(y_t - \bar{y})\,(y_{t+0} - \bar{y}) = 70.903878$$

$$c_1 = \frac{1}{T}\sum_{t=1}^{T-1}(y_t - \bar{y})\,(y_{t+1} - \bar{y}) = 20.627810$$

$$\Rightarrow r_1 = \frac{c_1}{c_0} = \frac{20.627810}{70.903878} = 0.290926$$

School of Applied Mathematics

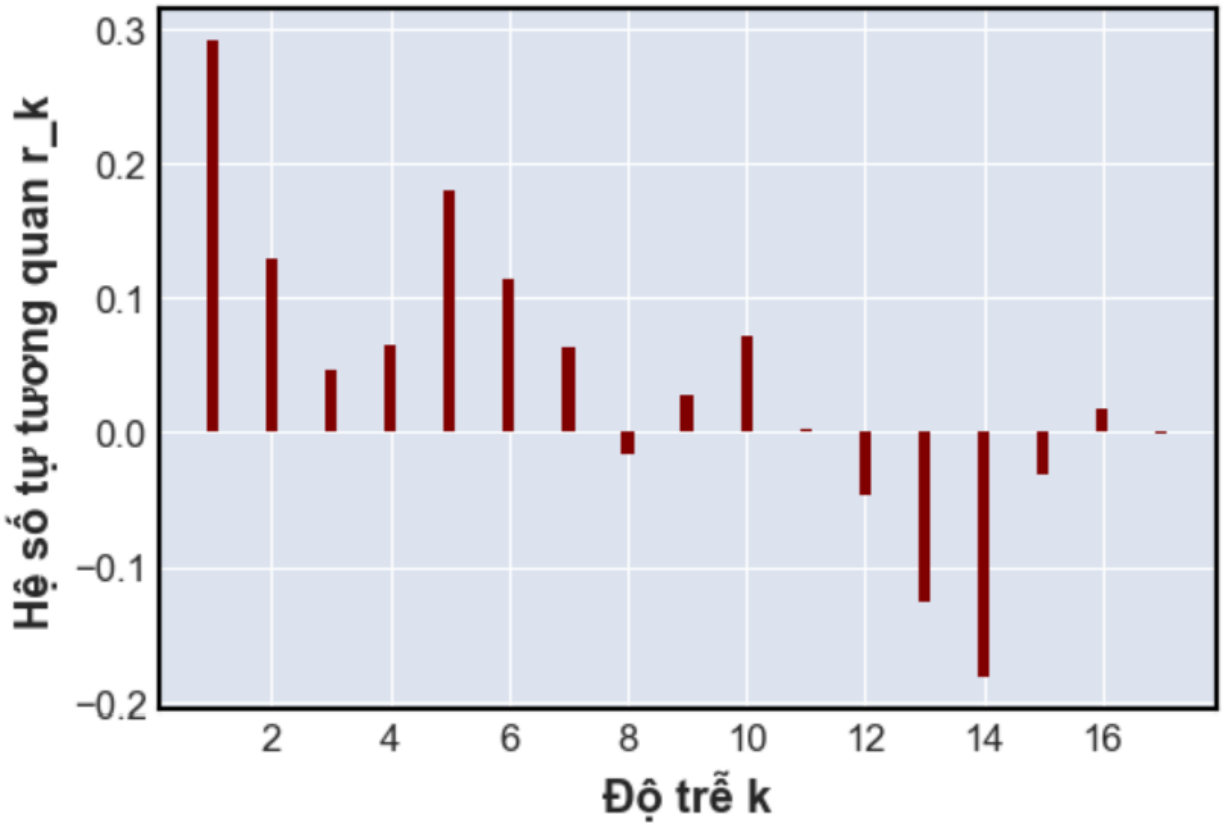
# Python code

```
# Sử dụng thư viện numpy
import numpy as np

# Tính toán hàm tự tương quan mẫu
yseries=sales[['N05C']].to_numpy()
T = len(yseries)
y_tb = yseries.mean()
K = T//4
y_res = yseries-y_tb
ACF = np.zeros((K+1,3))
for i in range(K+1):
    ACF[i,0] = i
    ACF[i,1] = 1/T*np.sum(y_res[:T-i]*y_res[i:T])
    ACF[i,2] = ACF[i,1]/ACF[0,1]
ACF = pd.DataFrame(ACF, columns = ['lag-k', 'c_k', 'r_k'])
ACF
```

# Bảng và biểu đồ ACF

| | lag-k | c_k | r_k |
|---|---|---|---|
| **0** | 0.0 | 70.903878 | 1.000000 |
| **1** | 1.0 | 20.627810 | 0.290926 |
| **2** | 2.0 | 9.117050 | 0.128583 |
| **3** | 3.0 | 3.251391 | 0.045856 |
| **4** | 4.0 | 4.646548 | 0.065533 |
| **5** | 5.0 | 12.762114 | 0.179992 |
| **6** | 6.0 | 8.061149 | 0.113691 |
| **7** | 7.0 | 4.459367 | 0.062893 |
| **8** | 8.0 | -1.182210 | -0.016673 |
| **9** | 9.0 | 1.939886 | 0.027359 |
| **10** | 10.0 | 5.129942 | 0.072351 |
| **11** | 11.0 | 0.178569 | 0.002518 |
| **12** | 12.0 | -3.346478 | -0.047197 |
| **13** | 13.0 | -8.947443 | -0.126191 |
| **14** | 14.0 | -12.914122 | -0.182136 |
| **15** | 15.0 | -2.264883 | -0.031943 |
| **16** | 16.0 | 1.233131 | 0.017392 |
| **17** | 17.0 | -0.092120 | -0.001299 |



# Biến đổi dữ liệu

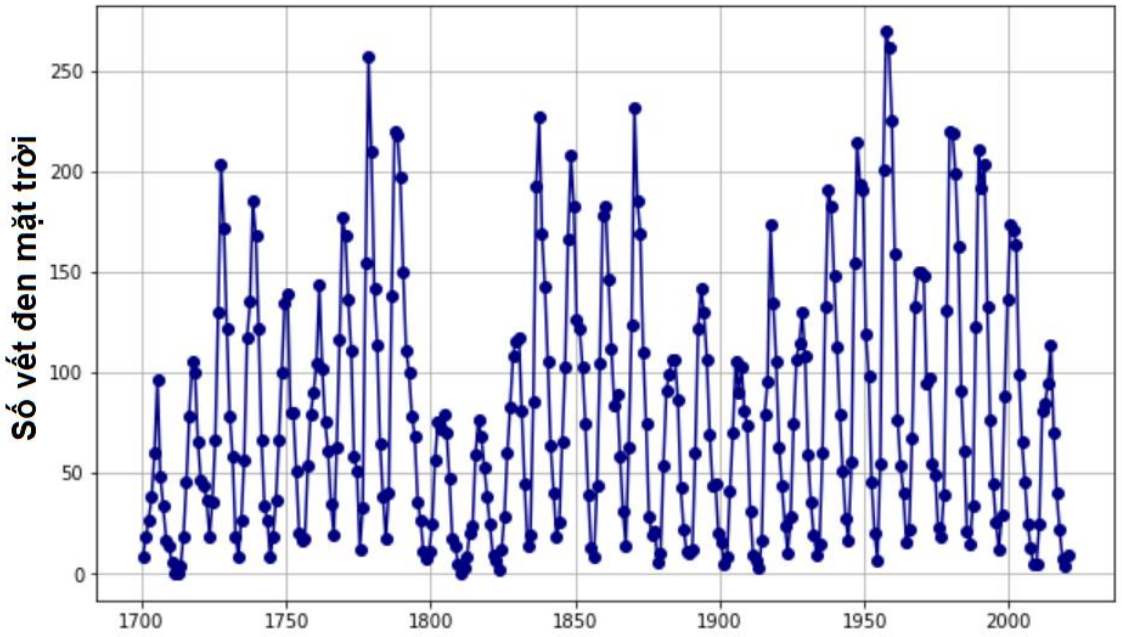
- Thường được sử dụng trong phân tích chuỗi thời gian nhằm ổn định hóa biến động của dữ liệu
- Biến đổi họ lũy thừa (**power family of transformation**)

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda\, \dot{y}^{\lambda-1}} & , \lambda \neq 0 \\ \dot{y}\, \ln(y) & , \lambda = 0 \end{cases}$$

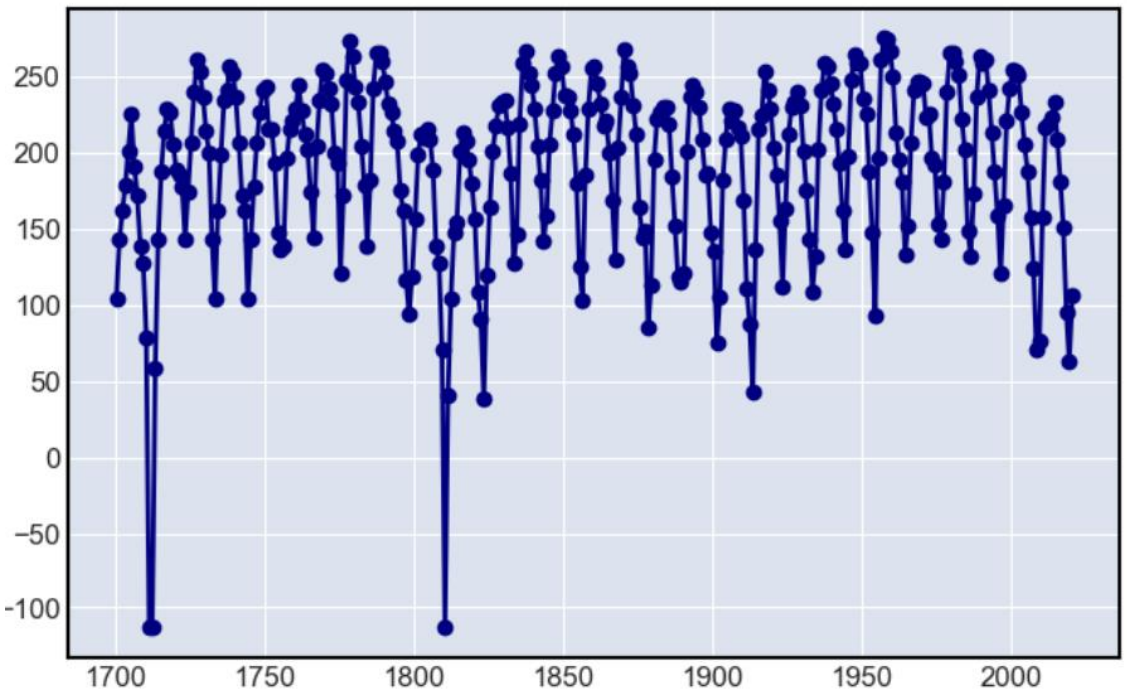
$$\text{với} \qquad \dot{y} = \exp\left( \frac{1}{T} \sum_{t=1}^{T} \ln(y_t) \right)$$

là trung bình nhân của các giá trị quan sát.

# Biến đổi logarit cho dữ liệu '*sunspot.xls*'




$$x = \dot{y}\, ln(y)$$



School of Applied Mathematics

# Hiệu chỉnh xu thế bằng hồi quy

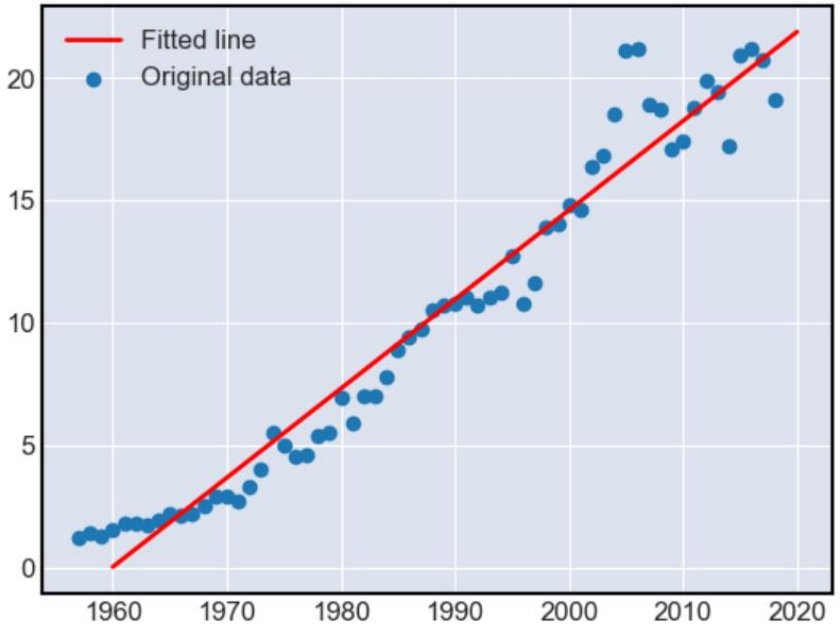
- Chuỗi thời gian có tính xu thế là chuỗi không dừng.
- Thành phần xu thế được xấp xỉ bởi một mô hình hồi quy, sau đó loại bỏ khỏi dữ liệu.
- Phần dư còn lại sẽ không còn yếu tố xu thế.

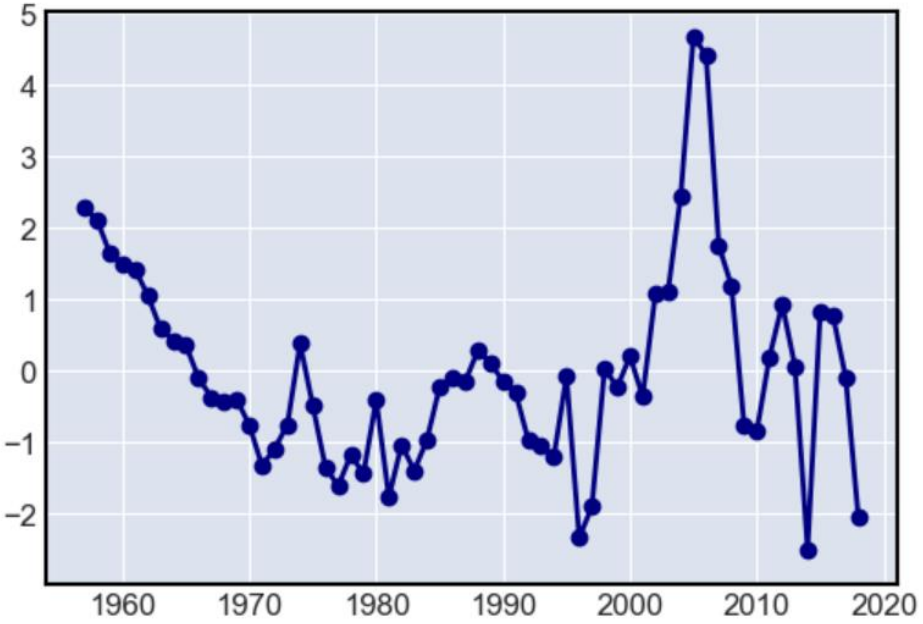
# Dữ liệu tiêu thụ rượu 1957-2018

Mô hình:

$y=-714.311808+0.364451x$

| | Year | Liters per capita | residual |
|---|---|---|---|
| 0 | 1957 | 1.2 | 2.281874 |
| 1 | 1958 | 1.4 | 2.117423 |
| 2 | 1959 | 1.3 | 1.652973 |
| 3 | 1960 | 1.5 | 1.488522 |
| | | 1.8 | 1.424071 |



School of Applied Mathematics

# Hiệu chỉnh xu thế bằng sai phân hóa

- Áp dụng sai phân lên chuỗi thời gian để thu được dữ liệu mới:

$$x_t = y_t - y_{t-1} = \nabla y_t = (1 - B)y_t$$

với $\nabla$ và $B$ là toán tử **sai phân** và toán tử **lùi**.

- Quy tắc lũy thừa:

$$B^d y_t = y_{t-d}$$
$$\nabla^d = (1 - B)^d$$

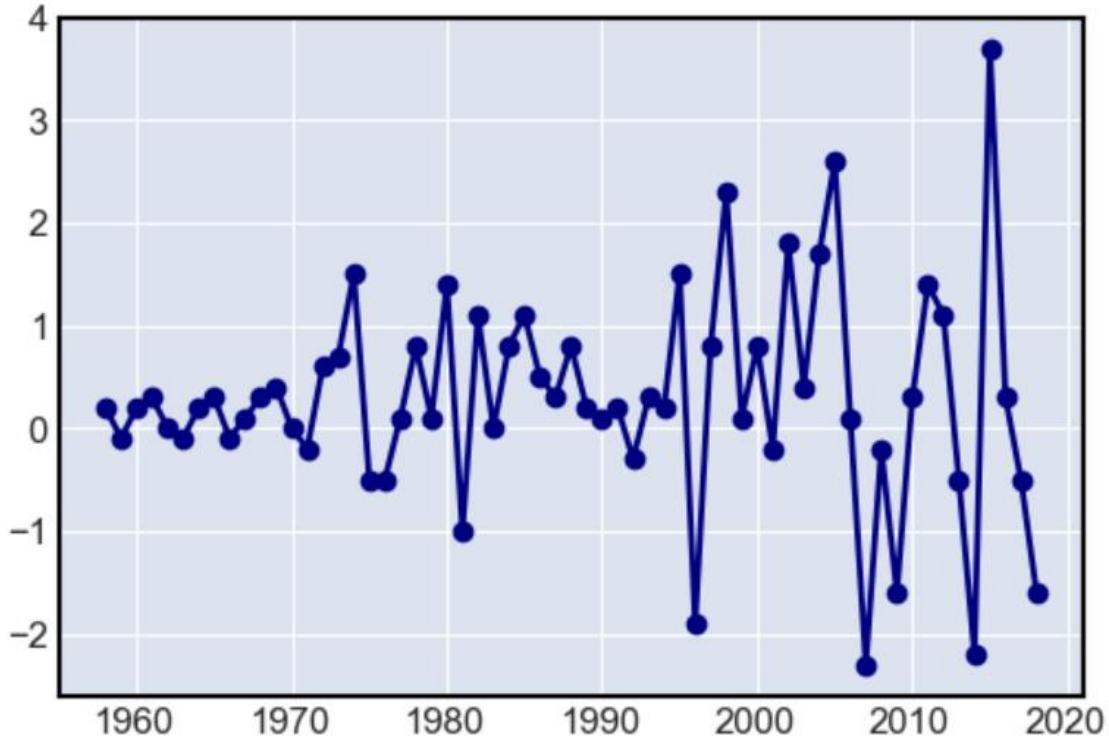
- Sai phân có lợi thế đơn giản (không phải ước lượng cái gì) và linh hoạt khi xu thế thay đổi theo thời gian.

# Dữ liệu tiêu thụ rượu 1957-2018 (tiếp)

Áp dụng sai phân một lần:

$$x_t = (1 - B)y_t$$

| | Year | Liters per capita | d=1 |
|---|---|---|---|
| 0 | 1957 | 1.2 | NaN |
| 1 | 1958 | 1.4 | 0.2 |
| 2 | 1959 | 1.3 | -0.1 |
| 3 | 1960 | 1.5 | 0.2 |
| 4 | 1961 | 1.8 | 0.3 |



School of Applied Mathematics

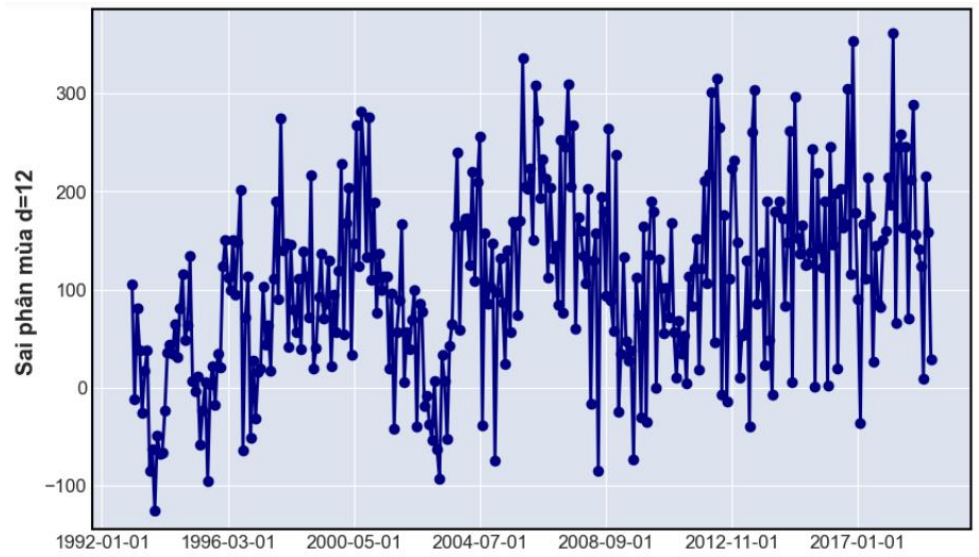
# Hiệu chỉnh mùa (+ xu thế) bằng sai phân

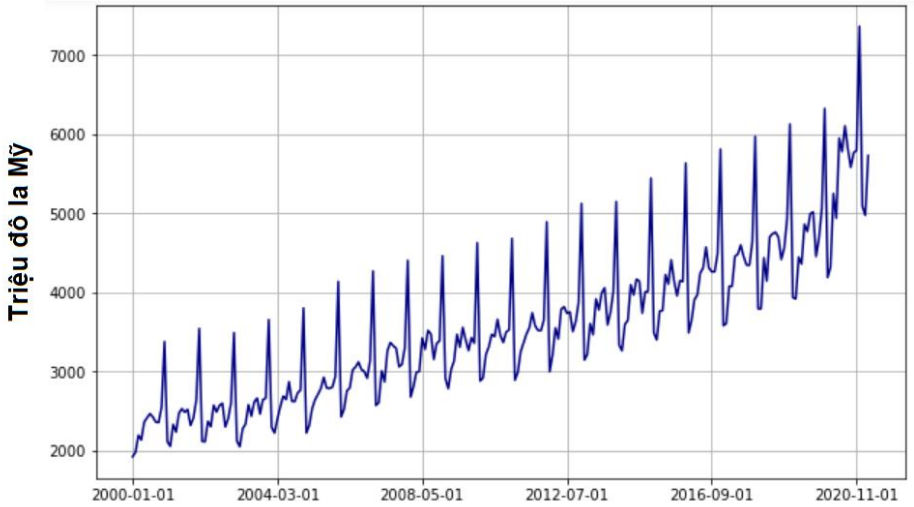
- Loại bỏ yếu tố mùa bằng toán tử sai phân mùa độ trễ d:

$$\nabla_d y_t = \left(1 - B^d\right) y_t = y_t - y_{t-d}$$

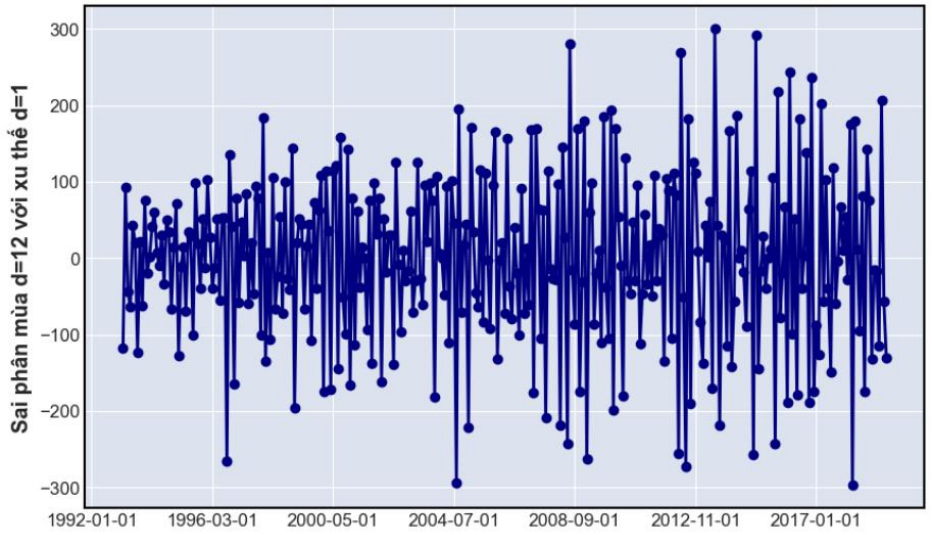
- Khi cả 2 yếu tố (xu thế + mùa) xuất hiện trong chuỗi thời gian:
  - Đầu tiên lấy sai phân mùa để loại bỏ yếu tố mùa,
  - Sau đó lấy sai phân 1 hoặc 2 lần để loại bỏ xu thế.

# Dữ liệu doanh số bán lẻ đồ uống có cồn




| | DATE | RETAILSALES | SEASON(k=12) | TREND(k=1) |
|---|---|---|---|---|
| **10** | 1992-11-01 | 1831 | NaN | NaN |
| **11** | 1992-12-01 | 2511 | NaN | NaN |
| **12** | 1993-01-01 | 1614 | 105.0 | NaN |
| **13** | 1993-02-01 | 1529 | -12.0 | -117.0 |
| **14** | 1993-03-01 | 1678 | 81.0 | 93.0 |



# Phương pháp mô hình hóa và dự báo

1. Vẽ đồ thị và xác định các thuộc tính cơ bản
2. Loại bỏ thành phần mùa và xu thế nếu có
3. Xây dựng mô hình dự báo cho phần dư
4. Đánh giá mô hình => chia dữ liệu (train + validation)
5. Dự báo => đảo ngược các biến đổi và hiệu chỉnh
6. Xây dựng và thực hiện quy trình theo dõi dữ báo => phát hiện suy giảm hiệu suất

School of Applied Mathematics

# Các tiêu chuẩn đánh giá mô hình

Các tiêu chuẩn đánh giá sử dụng sai số dự báo một bước:

$$e_t(1) = y_t - \hat{y}_t(t-1)$$

trong đó $\hat{y}_t(t-1)$ là dự báo của $y_t$ tại thời điểm $(t-1)$. Các số đo độ chính xác được dùng là:

**MAE** (mean absolute error) và **MSE** (mean squared error):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t(1)| \quad ; \quad MSE = \frac{1}{n}\sum_{t=1}^{n}[e_t(1)]^2$$

School of Applied Mathematics

# Bài tập

1. Viết code để vẽ biểu đồ về độ nhớt hóa chất được cho trong file '*viscosity.csv*'.
2. Chọn một mã cổ phiếu trên thị trường chứng khoán Việt Nam. Vẽ biểu đồ nến, đường trung bình MA20 và thử đánh giá về xu hướng giá cổ phiếu.
3. Viết hàm tính các hệ số tự tương quan ACF. Tham số của hàm gồm dữ liệu chuỗi thời gian (data) và độ trễ (K) tối đa. Hàm sẽ trả về mảng chứa các giá trị ước lượng của ACF. Áp dụng với dữ liệu '*salesmonthly*'.

School of Applied Mathematics

# Bài tập

4. Viết code để thực hiện biến đổi dữ liệu về số vết đen mặt trời trong file '*sunspot.xls*'.
5. Viết code để thực hiện hiệu chỉnh xu thế bằng phương pháp hồi quy và sai phân đối với dữ liệu về tiêu thụ rượu từ 1957 đến 2018 trong '*wineconsumption.csv*'.
6. Viết code để thực hiện hiệu chỉnh xu thế và mùa bằng sai phân đối với dữ liệu về doanh số bán lẻ đồ uống trong '*beveragesales.csv*'.