# Chương 1.

# Tổng quan về dữ liệu chuỗi thời gian

# (Time Series Data)

**PhD. Nguyễn Thị Khánh Tiên**

**email: tiennntk@ukr.net**

## Nội dung chính trong chương 1:

- Các khái niệm quan trọng của dữ liệu chuỗi thời gian
- Quy trình khai phá Time Series theo từng bước
- Các ứng dụng quan trọng trong các lĩnh vực

## Mục tiêu thực hành chương 1:

- Hiểu cấu trúc dữ liệu chuỗi thời gian
- Biết cách tiền xử lý, trực quan hóa và phân tích cấu trúc chuỗi
- Biết tạo đặc trưng cơ bản và đánh giá dữ liệu
- Sẵn sàng cho các bài thực hành nâng cao (ARIMA, LSTM, Transformer)

# Các khái niệm cơ bản

## Dữ liệu chuỗi thời gian (Time Series Data)

- là tập hợp các quan sát được ghi nhận theo trình tự thời gian, thường cách đều nhau.
- ví dụ: nhiệt độ mỗi giờ, doanh số mỗi ngày, lượng điện tiêu thụ mỗi tháng, giá cổ phiếu theo từng phút

## Đặc trưng quan trọng nhất:

- thứ tự thời gian không được xáo trộn, vì tính phụ thuộc giữa các điểm dữ liệu.

## Timestamp

- là thời điểm xảy ra một quan sát. Nó phải được làm sạch và chuẩn hóa trước khi phân tích (ví dụ chuyển về cùng timezone, cùng định dạng).

## Các loại chuỗi thời gian:

- ***Univariate Time Series****:* chỉ có 1 biến mục tiêu
- ***Multivariate Time Series***: nhiều biến diễn biến theo thời gian
- ***Irregular Time Series***: dữ liệu không lấy mẫu đều (cần resampling)
- ***High-frequency Time Series****:* dữ liệu theo mili giây, micro giây (tài chính)

# Các khái niệm cơ bản

## Stationarity (Tính dừng).

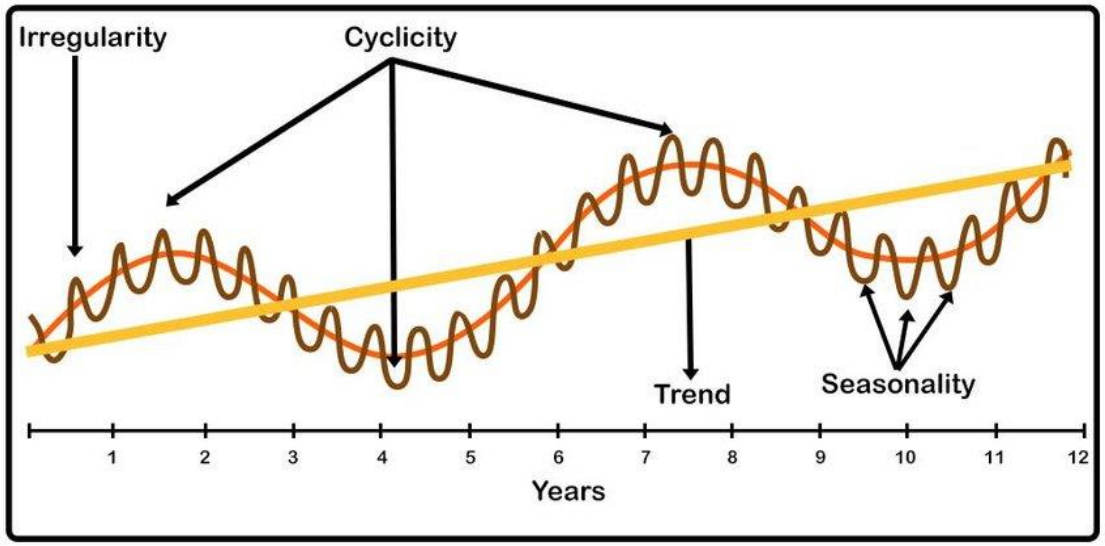
- Một chuỗi được xem là dừng nếu:
  - Giá trị kỳ vọng không đổi
  - Phương sai không đổi
  - Tự tương quan không thay đổi theo thời gian
  - Tính dừng rất quan trọng trong các mô hình truyền thống như ARIMA.

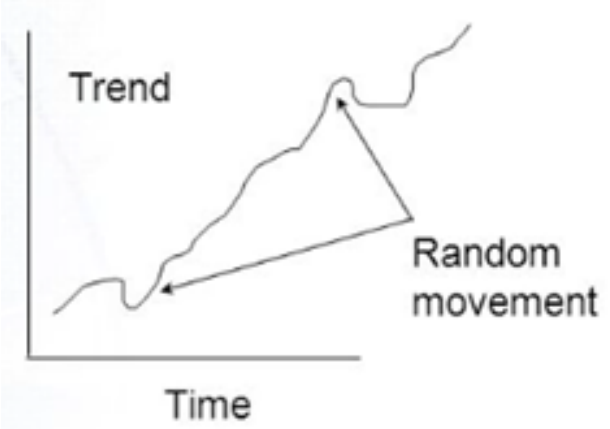
## Autocorrelation & Partial Autocorrelation

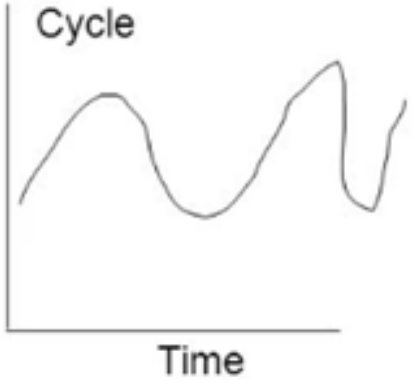
- **ACF (Autocorrelation Function)**: đo mức độ tương quan giữa dữ liệu hiện tại và các giá trị quá khứ.
- **PACF (Partial Autocorrelation Function)**: đo tương quan "thuần", không bị ảnh hưởng các lag trung gian.

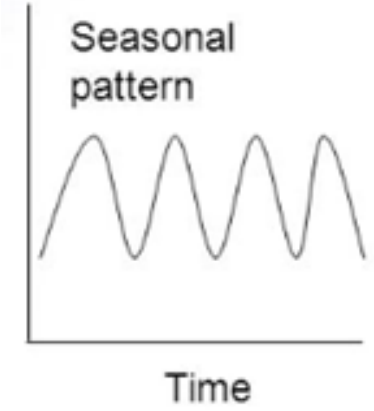
# Các thành phần chính của chuỗi thời gian

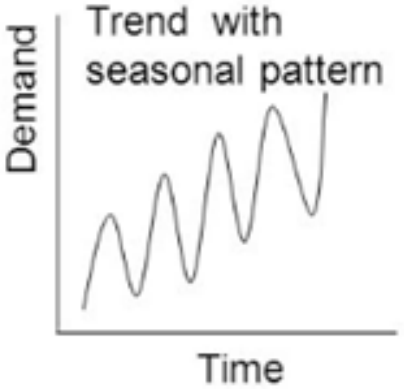
- **Trend (Xu hướng).**
  - Biểu thị xu hướng tăng hoặc giảm trong thời gian dài.
  - Ví dụ: nhu cầu điện tăng theo năm.
- **Seasonality (Mùa vụ).**
  - Các mẫu lặp lại theo chu kỳ như giờ, ngày, tuần, quý, năm.
  - Ví dụ: doanh số tăng vào cuối tuần.
- **Cyclic (Tính chu kỳ không đều).**
  - Các biến động dài hạn nhưng không có chu kỳ cố định (khác seasonality).
  - Ví dụ: chu kỳ kinh tế (khủng hoảng – phục hồi).
- **Noise (Nhiễu).**
  - Những dao động ngẫu nhiên, khó dự đoán.

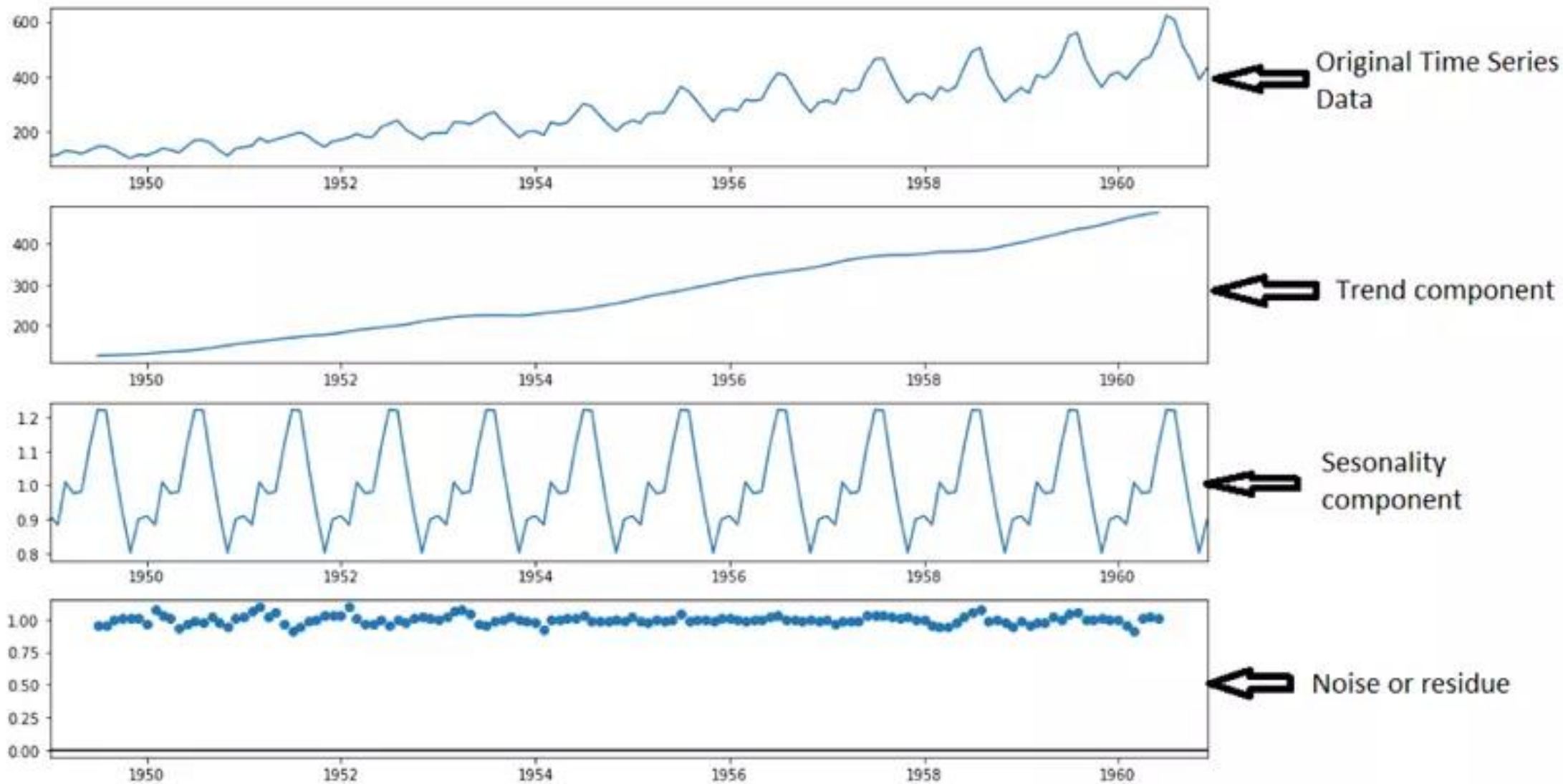

Irregularity
Cyclicity
Trend
Seasonality
1
2
3
4
5
6
7
8
9
10
11
12
Years


Trend
Random movement
Time


Cycle
Time

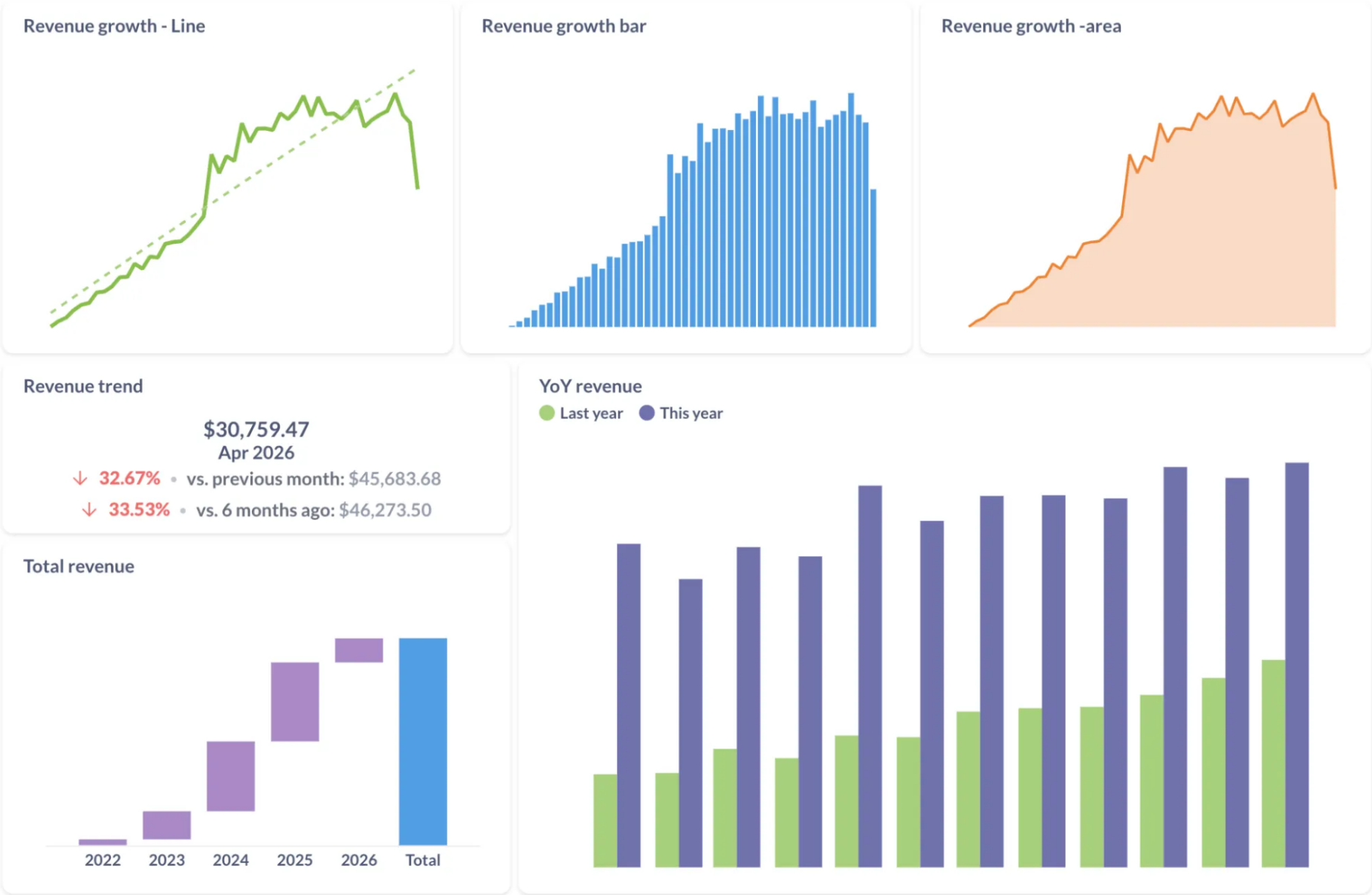

Seasonal pattern
Time


Demand
Trend with seasonal pattern
Time


600
400
200
1950
1952
1954
1956
1958
1960
Original Time Series Data
400
300
200
Trend component
1.2
1.1
1.0
0.9
0.8
Sesonality component
1.00
0.75
0.50
0.25
0.00
Noise or residue

# Các loại biểu đồ dùng để trực quan hóa dữ liệu chuỗi thời gian

| | |
|---|---|
| **Biểu đồ đường (Line Chart):** | • Dùng để theo dõi xu hướng hoặc sự thay đổi của dữ liệu theo thời gian.<br>→ Phù hợp nhất cho dữ liệu liên tục, muốn xem tăng/giảm ra sao theo từng mốc thời gian. |
| **Biểu đồ cột (Bar Chart):** | • Dùng để so sánh giá trị giữa các khoảng thời gian khác nhau, đặc biệt hữu ích với dữ liệu dạng rời rạc theo tháng, quý, hoặc năm. |
| **Biểu đồ vùng (Area Chart):** | • Thể hiện giá trị tích lũy hoặc tổng thể theo thời gian, nhấn mạnh vào mức độ "lấp đầy" dưới đường biểu diễn để thấy quy mô tổng. |
| **Biểu đồ xu hướng (Trend Chart):** | • Dùng để phân tích sự thay đổi hiệu suất bằng cách so sánh giá trị hiện tại với các giai đoạn trước.<br>→ Hay dùng trong báo cáo tài chính, KPI hoặc đánh giá tăng trưởng. |
| **Biểu đồ thác nước (Waterfall Chart):** | • Thể hiện các thay đổi liên tiếp của dữ liệu qua từng bước hoặc từng thời kỳ.<br>→ Ít phổ biến hơn nhưng rất hữu ích khi muốn xem từng yếu tố đóng góp vào thay đổi cuối cùng như thế nào. |

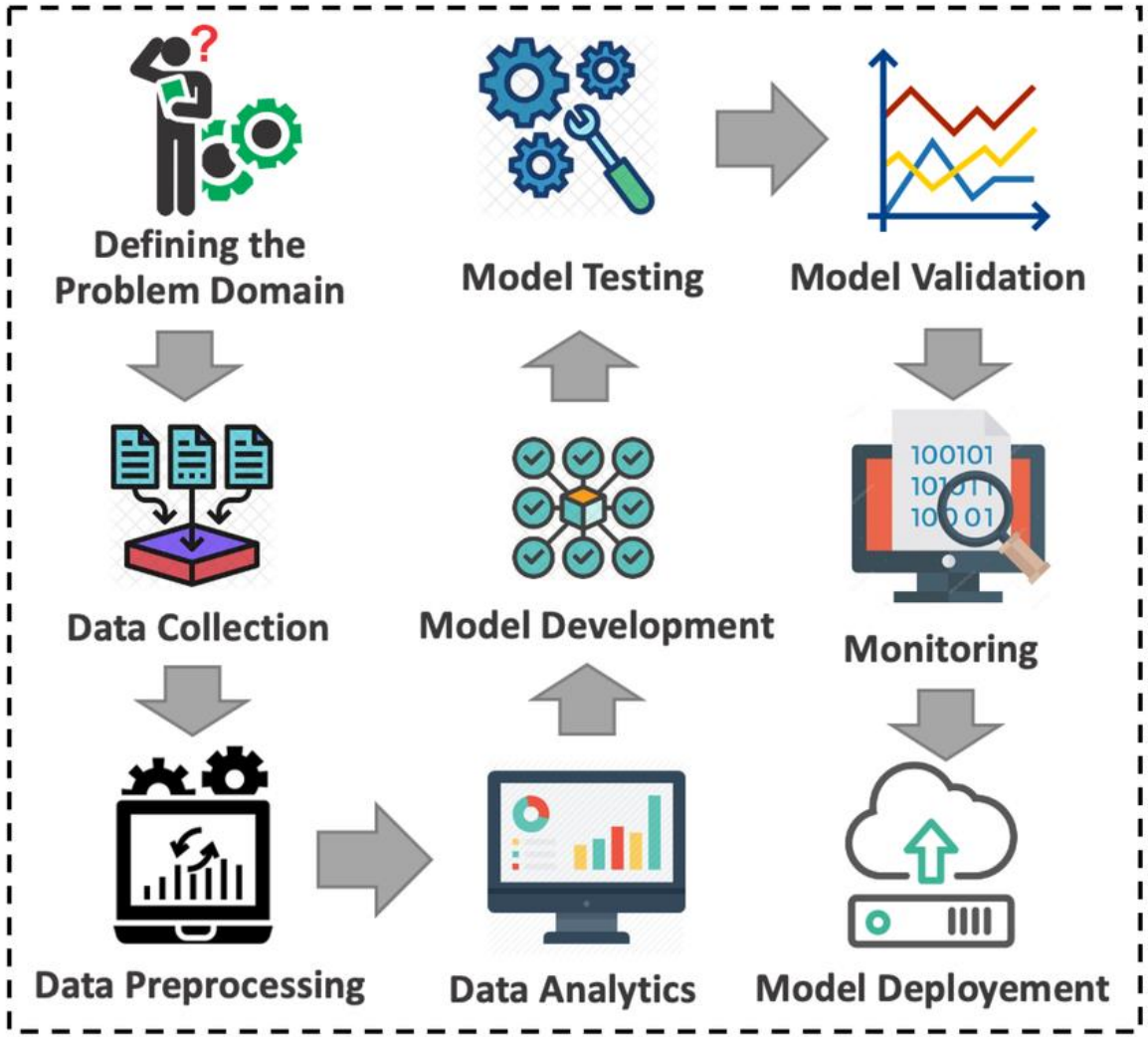



https://metabase-public.metabaseapp.com/public/dashboard/8d4961e2-15bd-4a8d-9cc1-d6451c570346?tab=199-chart-types-for-time-series

## Revenue by category

| | | 2022 | | 2023 | | 2024 | | 2025 | |
|---|---|---|---|---|---|---|---|---|---|
| **Product → Category** | **Product added** | Total | Quantity | Total | Quantity | Total | Quantity | Total | Quantity |
| Doohickey | 2022 | $5,700.49 | 3.73 | $16,907.08 | 3.87 | $25,410.45 | 3.71 | $25,643.65 | |
| | 2023 | $1,955.04 | 4.05 | $18,785.52 | 4.26 | $40,534.53 | 3.59 | $40,953.54 | |
| | 2024 | $938.06 | 4.07 | $5,121.60 | 4.08 | $24,208.05 | 3.74 | $27,611.67 | |
| | 2025 | $437.96 | 4.27 | $2,254.94 | 3.61 | $8,362.44 | 3.78 | $16,113.33 | |
| **Totals for Doohickey** | | **$9,031.56** | **3.86** | **$43,069.14** | **4.05** | **$98,515.47** | **3.67** | **$110,322.19** | |
| Gadget | 2022 | $6,221.79 | 4.62 | $17,322.80 | 4.06 | $29,602.17 | 4.18 | $28,701.18 | |
| | 2023 | $2,069.55 | 5.46 | $22,375.05 | 3.76 | $50,523.62 | 3.88 | $51,824.32 | |
| | 2024 | $1,699.26 | 3.9 | $10,068.58 | 4.02 | $37,139.37 | 4.46 | $51,662.65 | |
| | 2025 | $682.02 | 3.5 | $5,194.19 | 4.68 | $16,546.53 | 4.52 | $28,314.54 | |
| **Totals for Gadget** | | **$10,672.63** | **4.63** | **$54,960.62** | **3.99** | **$133,811.69** | **4.17** | **$160,502.69** | |
| Gizmo | 2022 | $5,535.74 | 4.46 | $15,689.72 | 3.45 | $21,969.88 | 3.71 | $19,352.90 | |
| | 2023 | $2,579.40 | 4.8 | $20,318.05 | 3.72 | $53,947.63 | 3.83 | $50,507.97 | |
| | 2024 | $1,449.95 | 6.3 | $9,613.43 | 3.8 | $52,029.84 | 3.63 | $60,212.04 | |
| | 2025 | $264.22 | 4.22 | $2,509.51 | 3.85 | $9,822.45 | 3.47 | $14,971.04 | |

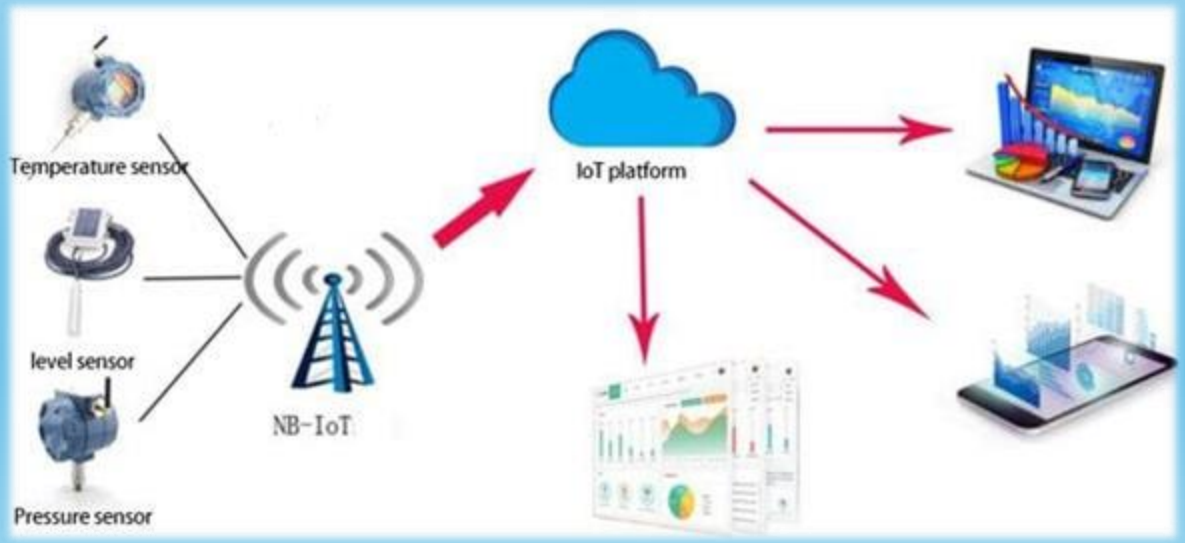
# Quy trình khai phá dữ liệu chuỗi thời gian (tương tự CRISP-DM)



# Bước 1: Thu thập dữ liệu (Data Collection)

## Dữ liệu chuỗi thời gian có thể đến từ:

| | | | | |
|---|---|---|---|---|
| Sensor IoT | API tài chính | Logs server | CSDL doanh nghiệp | Hệ thống ERP / CRM |





***Lưu ý:***

Đồng nhất timezone

Loại bỏ duplicate timestamp

Kiểm tra missing intervals

# Bước 2: Tiền xử lý dữ liệu (Preprocessing)

## a. Làm sạch dữ liệu

- Xử lý thiếu (impute bằng forward-fill, interpolate, rolling mean)
- Xử lý outliers
- Lọc nhiễu bằng moving average, exponential smoothing

## b. Chuẩn hóa thời gian

- Chuyển timestamp về chuẩn ISO
- Đảm bảo tần suất đều (resampling)

## c. Chuyển đổi & biến đổi

- Log transform để làm mượt chuỗi
- Differencing để làm dừng chuỗi

# Bước 3: Phân tích khám phá dữ liệu (EDA)

## Các biểu đồ thường dùng

- Line plot
- ACF, PACF
- Seasonal decomposition (trend / seasonality / residual)
- Heatmap seasonality theo giờ/ngày/tháng
- Boxplot theo ngày trong tuần

## Mục tiêu của EDA:

- Hiểu cấu trúc chuỗi
- Kiểm tra tính dừng
- Nhận diện seasonality
- Đánh giá biến động và bất thường

# Bước 4: Tạo đặc trưng (Feature Engineering)

| | |
|---|---|
| **a. Lag features:** | • lag-1, lag-7, lag-30,… |
| **b. Rolling features:** | • rolling mean, std, min, max |
| **c. Các đặc trưng theo thời gian (time features):** | • hour, day, week, month, quarter, holiday flag |
| **d. Cross-features:** | • tương tác giữa các biến trong multivariate time series |

# Bước 5: Lựa chọn mô hình

## Mô hình thống kê truyền thống ARIMA

- SARIMA (có mùa vụ)
- Holt-Winters (ETS)
- Ưu điểm: giải thích tốt. Nhược điểm: khó mở rộng với dữ liệu lớn, multivariate.

## Mô hình Machine Learning

- Random Forest
- Gradient Boosting (XGBoost, LightGBM)
- Linear Regression với lag features
- Ưu điểm: dễ triển khai, mạnh mẽ. Nhược điểm: không tự modeling được tính tuần tự.

## Mô hình Deep Learning

- RNN,LSTM, GRU
- Temporal Convolutional Network (TCN)
- Transformer (mạnh nhất cho chuỗi dài)
- DeepAR, N-BEATS, Temporal Fusion Transformer (TFT)
- Ưu điểm: modeling được long-term dependency, non-linear. Nhược điểm: cần dữ liệu lớn & GPU.

# Bước 6: Huấn luyện và đánh giá

## Train–validation split đúng cách

- Không được shuffle dữ liệu.
- 80% train – 20% test
- TimeSeries Cross Validation (rolling windows)

## Các chỉ số đánh giá

- MAE
- RMSE
- MAPE
- SMAPE
- MASE

## Backtesting

- Kiểm tra mô hình bằng nhiều cửa sổ dự báo.

# Bước 7: Dự báo & triển khai

- Dự báo ngắn hạn (short-term)
- Dự báo dài hạn (long-term)
- Triển khai bằng REST API, batch, hoặc streaming
- Theo dõi drift & tái huấn luyện định kỳ

# Ứng dụng của dữ liệu chuỗi thời gian


## Tài chính – Ngân hàng

- Dự báo giá cổ phiếu
- Dự báo rủi ro tín dụng
- Phát hiện gian lận
- Phân tích hành vi giao dịch

## Kinh doanh – Marketing

- Dự báo doanh số
- Dự báo tồn kho (Inventory Forecasting)
- Dự báo nhu cầu khách hàng
- Pricing optimization

## Sản xuất – IoT – Công nghiệp

- Predictive maintenance (dự đoán hỏng hóc)
- Phân tích dữ liệu cảm biến
- Kiểm soát chất lượng theo thời gian

# Ứng dụng của dữ liệu chuỗi thời gian


## Giao thông – Vận tải – Logistics

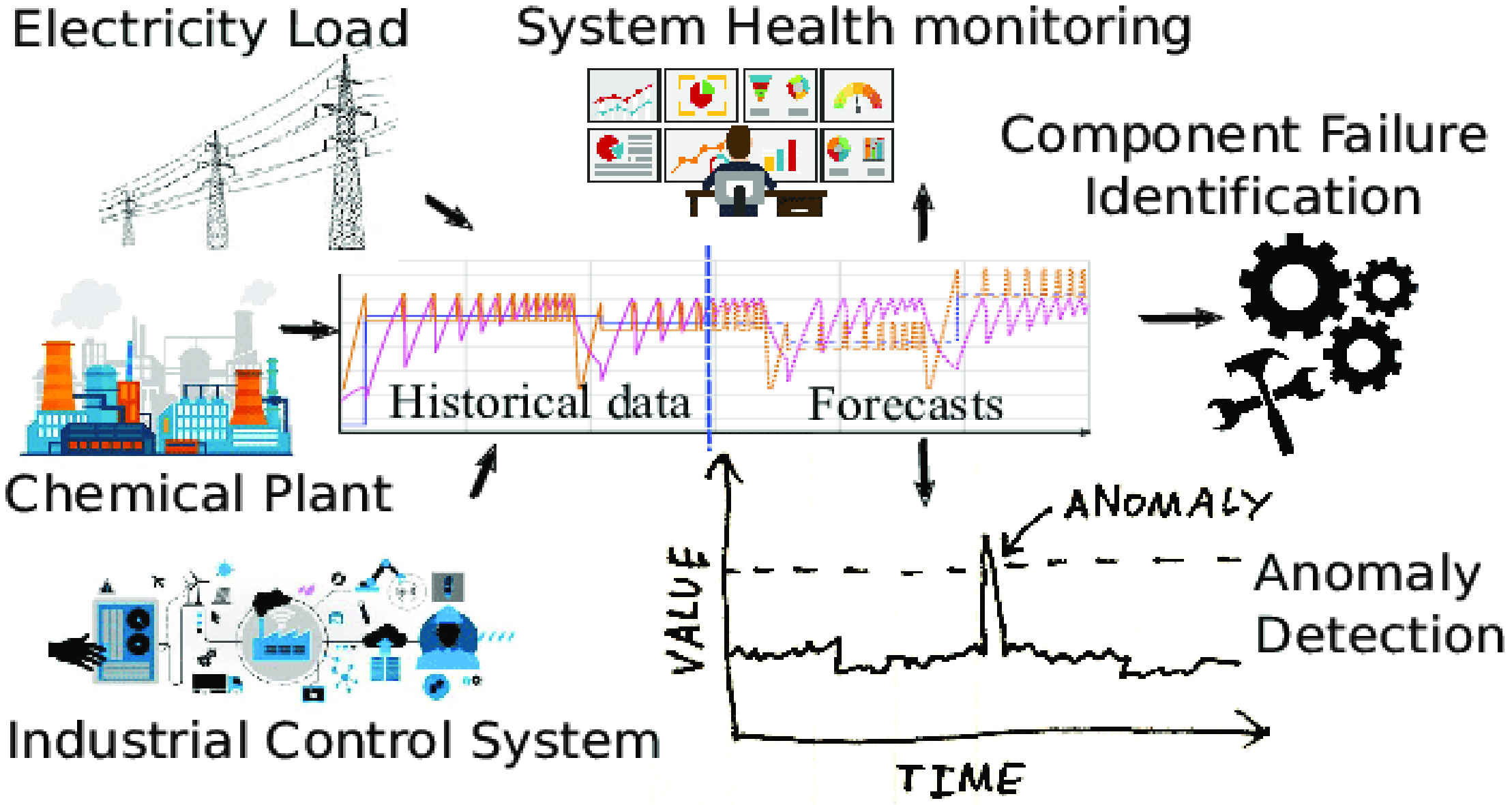
- Dự báo lưu lượng xe
- Thời gian giao hàng
- Dự báo nhu cầu vận tải

## Y tế

- Phân tích tín hiệu ECG, EEG
- Theo dõi sức khỏe bệnh nhân
- Dự báo tiến triển bệnh

## Năng lượng & Môi trường

- Dự báo nhu cầu điện
- Khí tượng: nhiệt độ, mưa, bão
- Chất lượng không khí, ô nhiễm


Electricity Load
System Health monitoring
Component Failure Identification
Historical data
Forecasts
Chemical Plant
Industrial Control System
ANOMALY
VALUE
TIME
Anomaly Detection

# Python Framework & Libraries for Time Series Data Analysis

## Forecasting

PROPHET ★17,9k statsmodels ★9,6k SKTIME ★7,5k

Darts ★7,4k Kats ★4,8k Time Series Library (TSLib) ★4,7k

★3,7k Statistical Forecast ★3,7k MERLION ★3,3k

★1,8k skforecast ★1k ETNA ★840 FEDOT ★610

## Changepoint detection

RUPTURES ★1,5k PROPHET ★17,9k Time Series Library (TSLib) ★4,7k

Kats ★4,8k MERLION ★3,3k ALIBI DETECT ★2,1k ★1,8k

luminol ★1,2k Anomaly Detection Toolkit (ADTK) ★1,1k

## Outlier detection

Python Outlier Detection (PyOD) ★8k TODS ★1,3k

ETNA ★840 Luminaire ★750 PySAD ★220

## Classification

SKTIME ★7,5k Time Series Library (TSLib) ★4,7k tslearn ★2,8k

Deep Learning for Time Series Classification(dl-4-tsc) ★1,5k

pyts ★1,7k ETNA ★840

## Clustering

SKTIME ★7,5k tslearn ★2,8k

## Pattern detection

stumpy ★3,1k

## Aggregation (feature extraction)

tsfresh ★8,2k Kats ★4,8k tsfel ★800 tsflex ★370

## Augmentation and data generation

tsai ★4,8k RGAN ★630 tsaug ★330

TimeSynth ★330 Time Series Augmentation ★320

# Python Framework & Libraries for Time Series Data Analysis

- **Pandas** → xử lý chuỗi thời gian
- **Statsmodels** → ARIMA, SARIMA
- **Scikit-learn** → ML cho time series
- **XGBoost/LightGBM** → forecasting kiểu tabular
- **Prophet** → forecast tự động
- **Darts / GluonTS** → deep learning cho chuỗi thời gian
- **TensorFlow/PyTorch** → mô hình TS tự xây
- **Plotly/Matplotlib** → trực quan hóa

# Python Framework & Libraries for Time Series Data Analysis

| Nhóm thư viện | Tên thư viện | Mục đích chính | Ưu điểm | Hạn chế |
|---|---|---|---|---|
| **Xử lý dữ liệu** | **Pandas** | Resample, rolling window, xử lý time index | Dễ dùng, mạnh cho xử lý dữ liệu | Không dự báo |
| | **NumPy** | Tính toán số học, hỗ trợ tạo feature | Nhanh, tối ưu | Không hỗ trợ time series trực tiếp |
| **Trực quan hóa** | **Matplotlib** | Vẽ biểu đồ time series cơ bản | Linh hoạt, phổ biến | Code dài, không tương tác |
| | **Seaborn** | Biểu đồ nâng cao, đẹp hơn | Dễ dùng, trình bày đẹp | Không chuyên cho time series |
| | **Plotly** | Biểu đồ tương tác | Rất trực quan, phù hợp dashboard | Hơi nặng |

# Python Framework & Libraries for Time Series Data Analysis

| Nhóm thư viện | Tên thư viện | Mục đích chính | Ưu điểm | Hạn chế |
|---|---|---|---|---|
| **Machine Learning** | **Scikit-learn** | ML (Random Forest, SVM, SVR) cho time series | Pipeline mạnh, đơn giản | Không hỗ trợ forecasting tự động |
| | **XGBoost** | Forecast dạng supervised learning | Cực mạnh cho tabular, xử lý phi tuyến | Cần tự tạo lag/rolling |
| | **LightGBM** | ML boosting tối ưu tốc độ | Rất nhanh, chính xác cao | Tuning khó nếu dữ liệu nhiễu |
| | **CatBoost** | ML boosting tối ưu categorical | Không cần one-hot, mạnh mẽ | Chậm hơn LightGBM |
| **Dự báo tự động** | **Prophet** | Dự báo trend + seasonality tự động | Dễ dùng, ít cần tuning | Không tốt với chuỗi nhiễu/phi tuyến mạnh |

# Python Framework & Libraries for Time Series Data Analysis

| Nhóm thư viện | Tên thư viện | Mục đích chính | Ưu điểm | Hạn chế |
|---|---|---|---|---|
| **Deep Learning chuyên Time Series** | **Darts** | Tập hợp mô hình: ARIMA, Prophet, RNN, LSTM, TCN, N-BEATS, Transformer | Full-stack, dễ dùng, multivariate tốt | Cài đặt nặng |
| | **GluonTS** | DeepAR, Transformer, seq2seq | Chuẩn research, mạnh với forecasting | Khá phức tạp, cần GPU |
| | **Kats (Meta/Facebook)** | Forecasting, anomaly detection | Nhiều thuật toán tích hợp sẵn | Chưa phổ biến rộng |
| **Framework DL tự xây** | **TensorFlow/Keras** | LSTM, GRU, TCN, Transformer | Tự do, mạnh nhất khi customise | Cần nhiều code & tuning |
| | **PyTorch** | Mô hình chuỗi nâng cao, research | Linh hoạt, mạnh cho mô hình phức tạp | Không phải thư viện chuyên TS |

# Reference

- https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python#1