Time Series Forecasting

- → Problem Statement [20 mis]
- → Missing Values / Anomalies [15 mins]
- → Breakdown of TS [60 mins]
- → Simple Methods for forecasting [20 mins]

CASE STUDY

Problem statement

Imagine you are a Data Scientist at MobiPlus, a mobile manufacturing company

You need to forecast their future sales for better planning and revenue.

- **Agenda 1:** We want to understand the patterns in demand to be able to be better plan for factory maintenance / staffing requirements.
- **Agenda 2:** We need a certain level of accuracy. The management requires that the Mean Absolute Percentage Error (MAPE) is not more than 5%.
- **Agenda 3:** Need a range forecast to supplement the point forecast to make educated trade-off wherever needed.

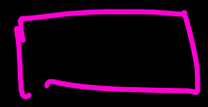
Over the next few lectures, we will be completing these tasks.

Q: Why forecast?

↳ under / over?

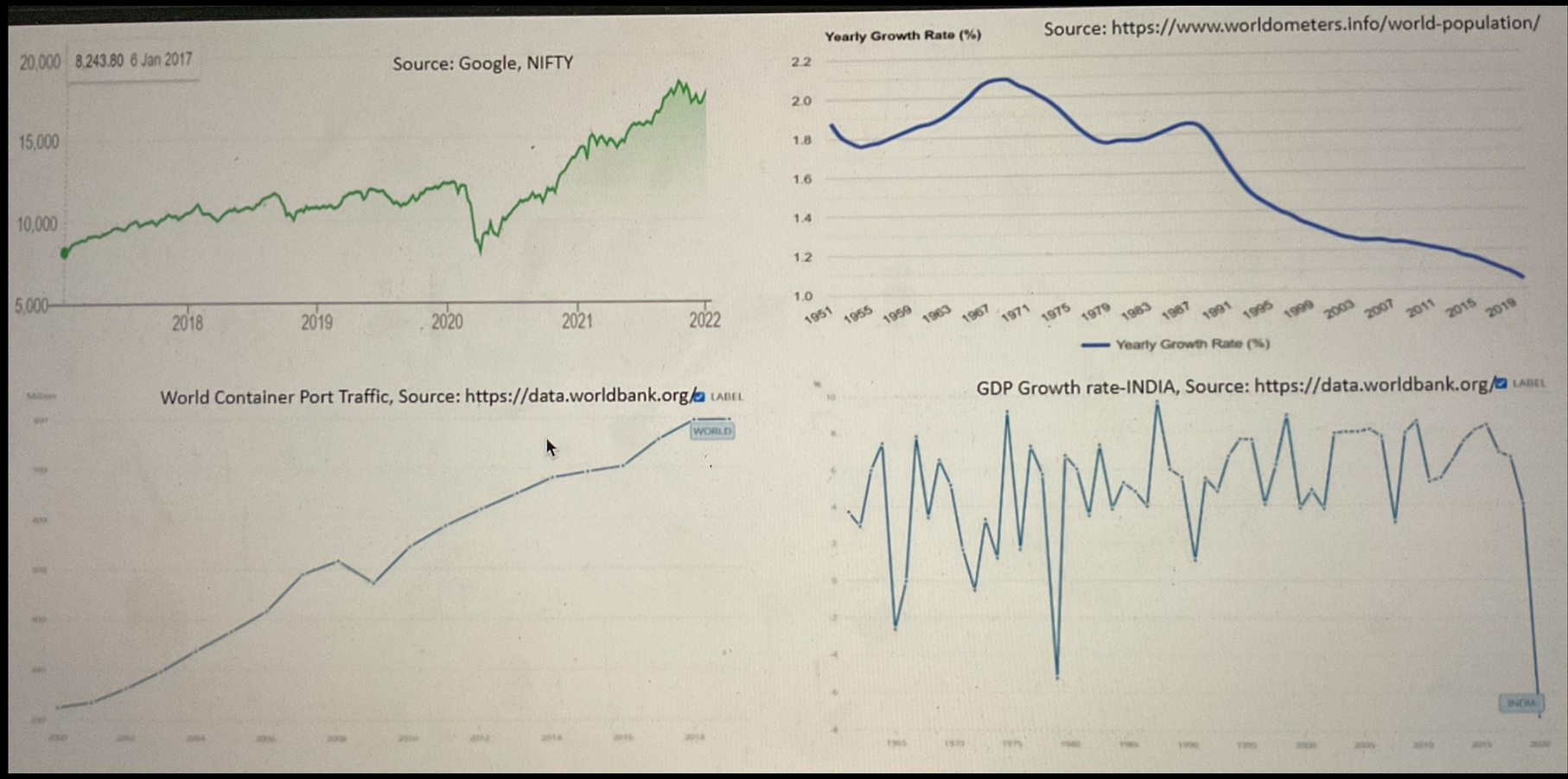
	DATE	Sales
0	2001-01-01	6519.0
1	2001-02-01	6654.0
2	2001-03-01	7332.0
3	2001-04-01	7332.0
4	2001-05-01	8240.0

Time Series Data:
Any "Signal", indexed by an ordered time stamp
- Yearly
- Monthly
- hourly
- etc.

 ? Continue the pattern!

A) Regression B) Classification C) Clustering

Examples

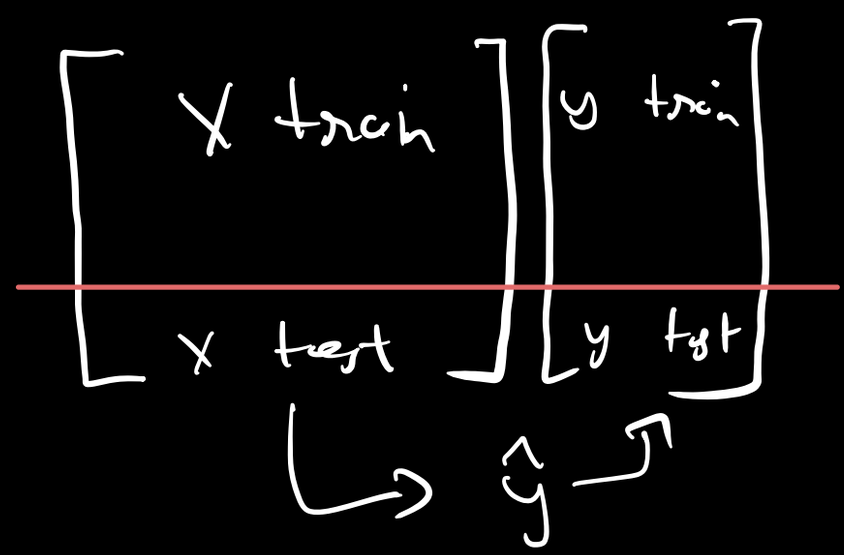


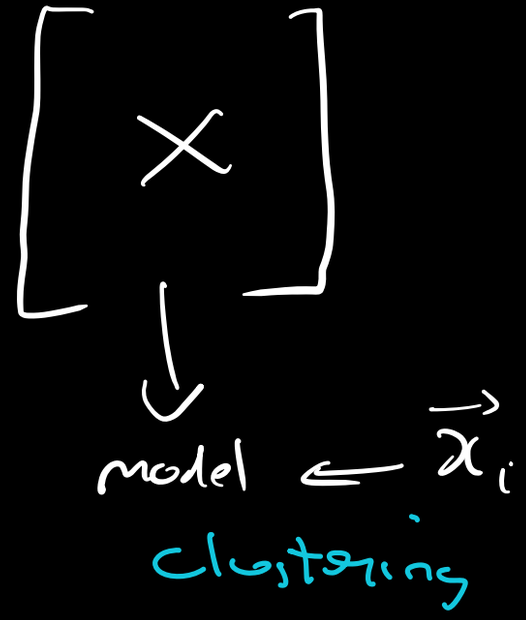
Forecasting

given : $y_1, y_2 \cdots y_{t-1}, y_t$

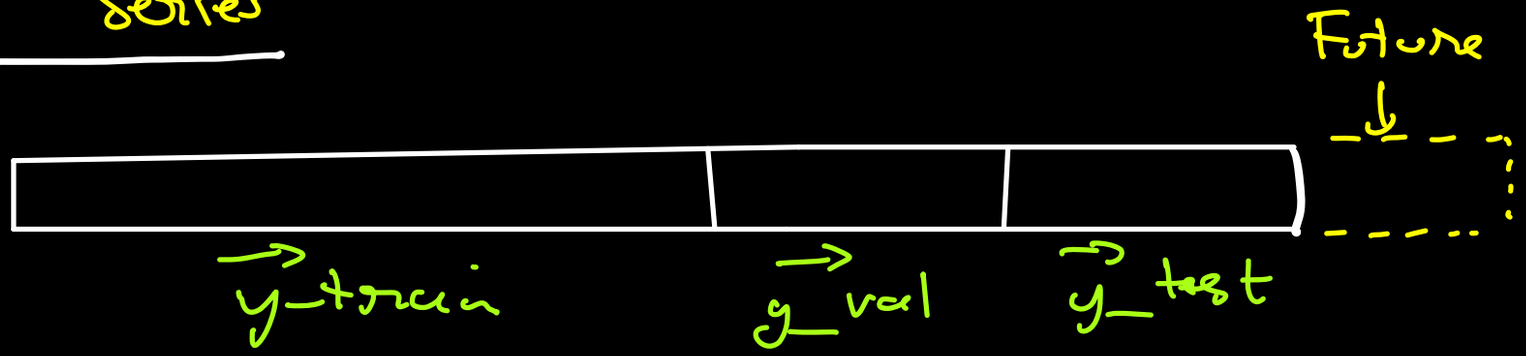
Future → predict $y_{t+1}, y_{t+2} \cdots$

Earlier Setups:





Time Series

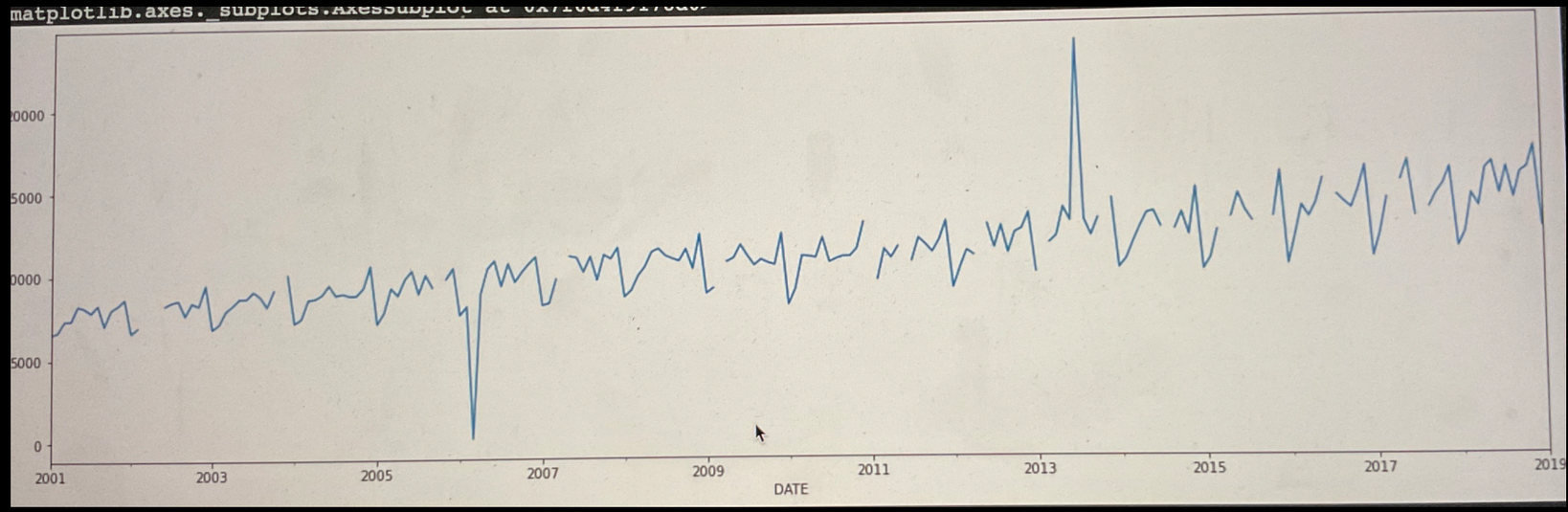


→ EDA

→ # rows : 18 years of monthly data
+ 1 month of 2019
Total = 18 * 12 + 1 = 217 months

df.date.nunique() =

→ Set index → Looking at the plot



→ Any challenges / wrong stuff?

→ Anomaly

→ Missing values.

Q: Can I do percentile based / IQR based detection?

→ Yes, in most cases that works

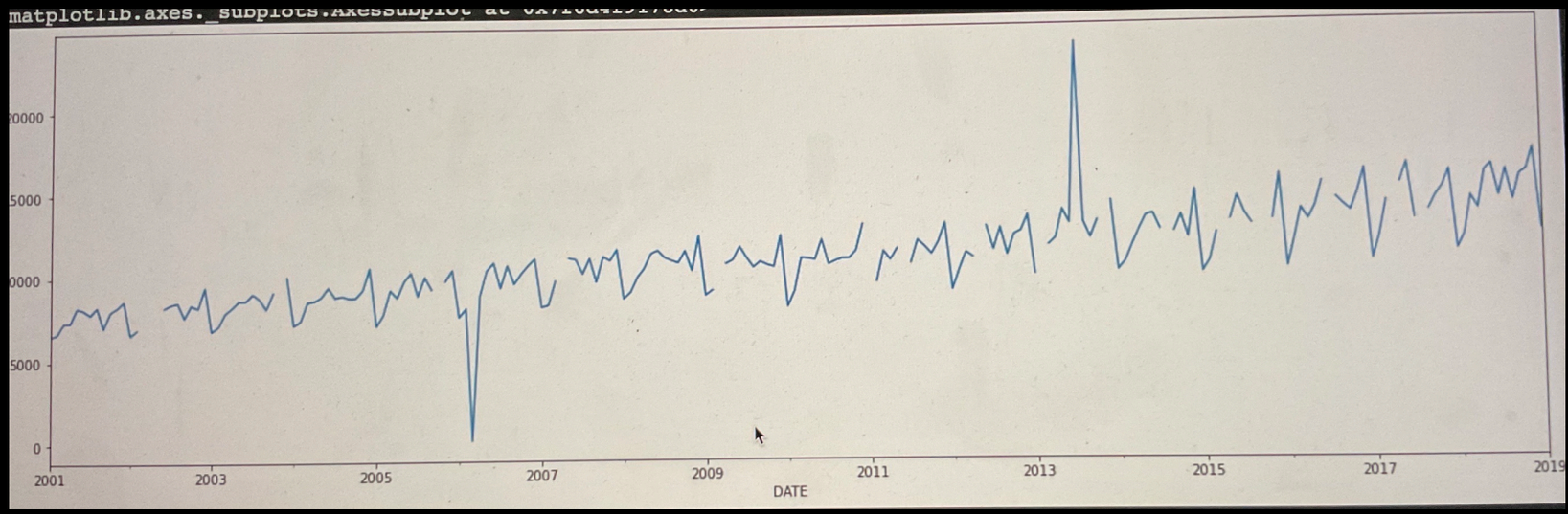
Q: Is it okay to delete / remove outlier data?

→ No, can't have break in the pattern.

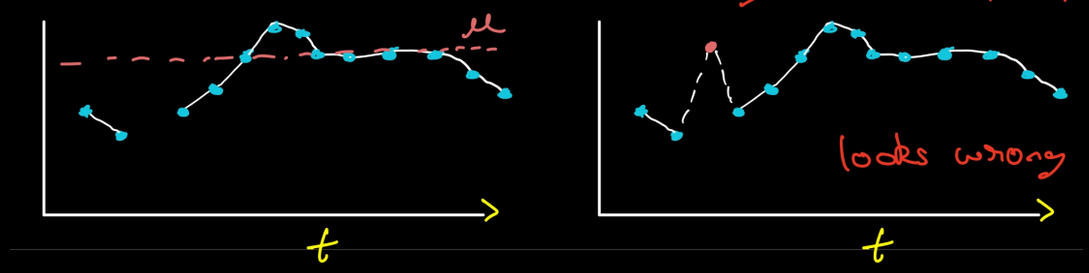
Imputation

Q: Any suggestions?

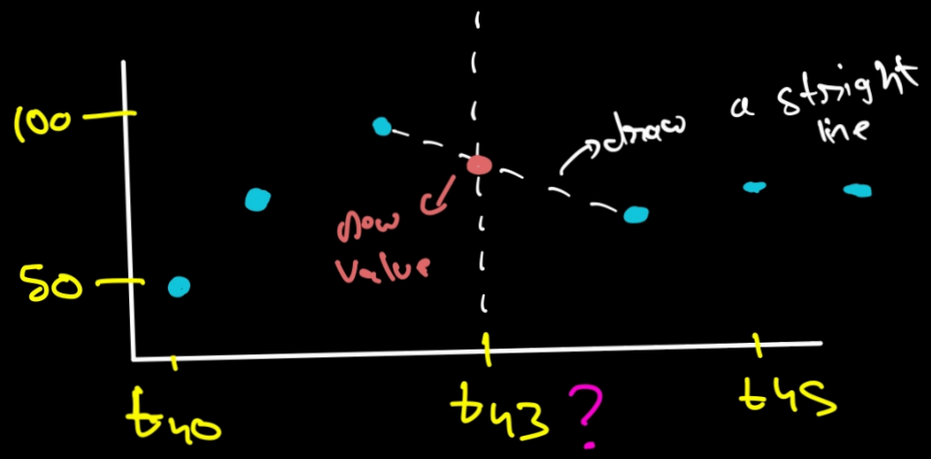
→ mean ✗ → median ✗ → zero ✗



→ Mean imputation

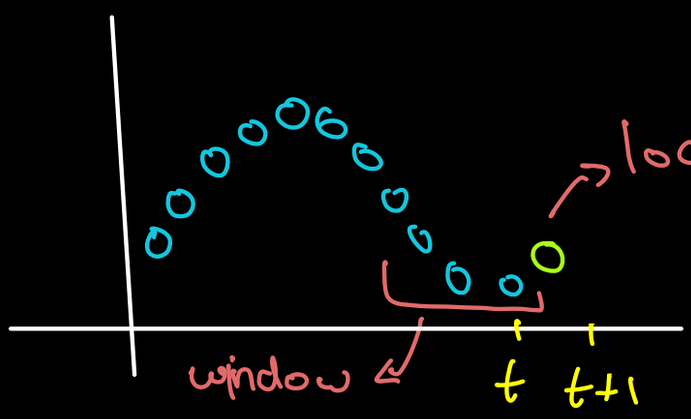


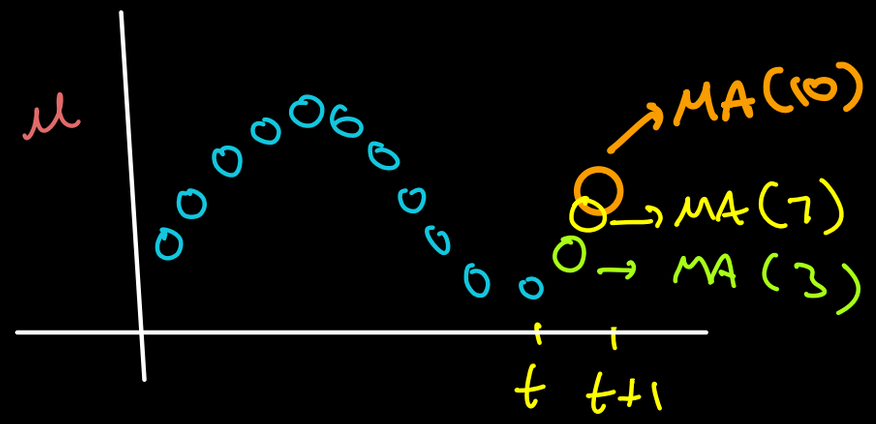
Interpolation



→ Another way to impute is called centered MA.

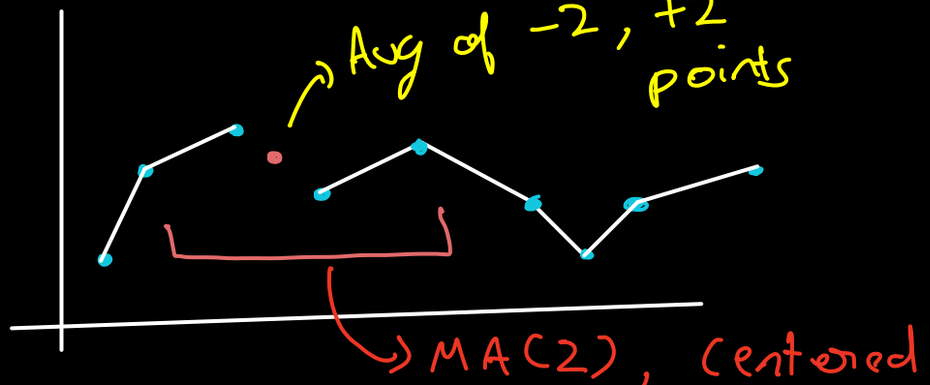
Moving Average





$$\hat{y}_t = \frac{y_{t-1} + y_{t-2} + y_{t-3} \cdots y_{t-m}}{m} = \frac{1}{m} \sum_{i=t-m}^{t} y_i$$

Centered MA



$$\hat{y}_t = \frac{1}{2m+1} \sum_{j=t-m}^{t+m} y_j$$

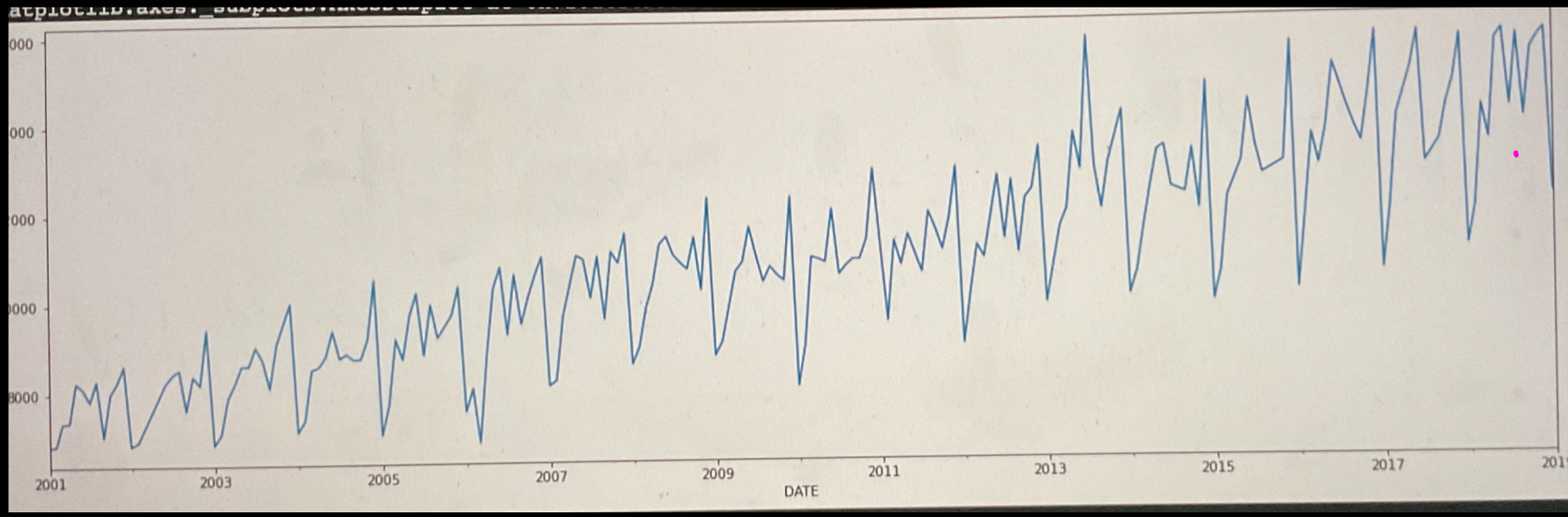
Weighted MA

$$\hat{y}_{t+1} = \sum_{i=t-m}^{t} \alpha_i y_i \quad \text{OR} \quad \sum_{j=t-m}^{t+m} \alpha_i y_i$$

Such that $\sum \alpha_i = 1$

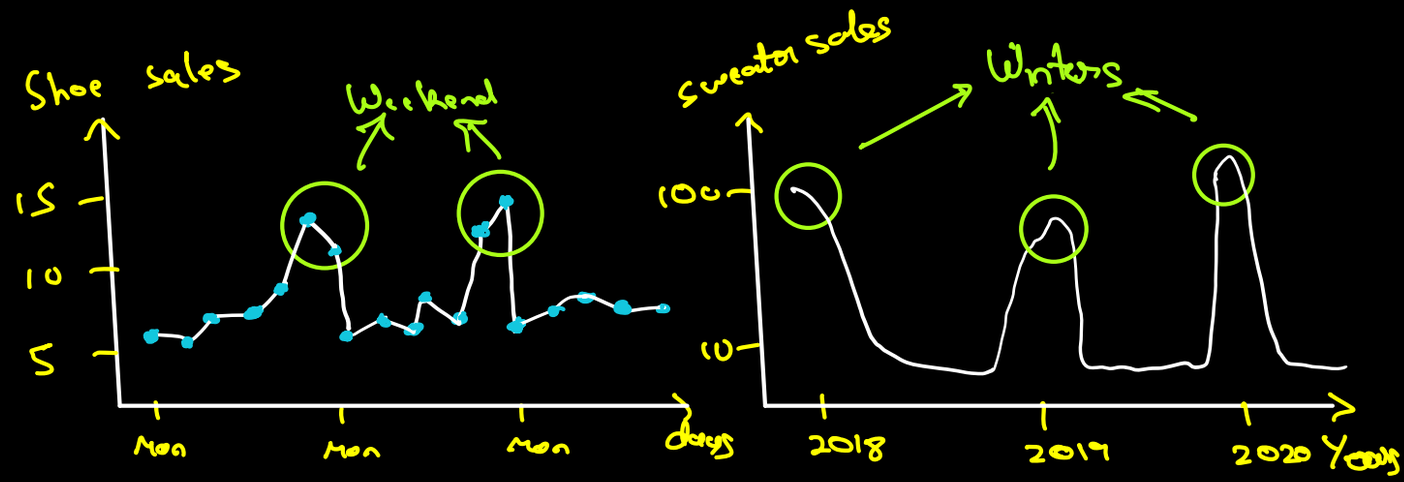
Components of a time series

Can you notice some patterns in this plot ?



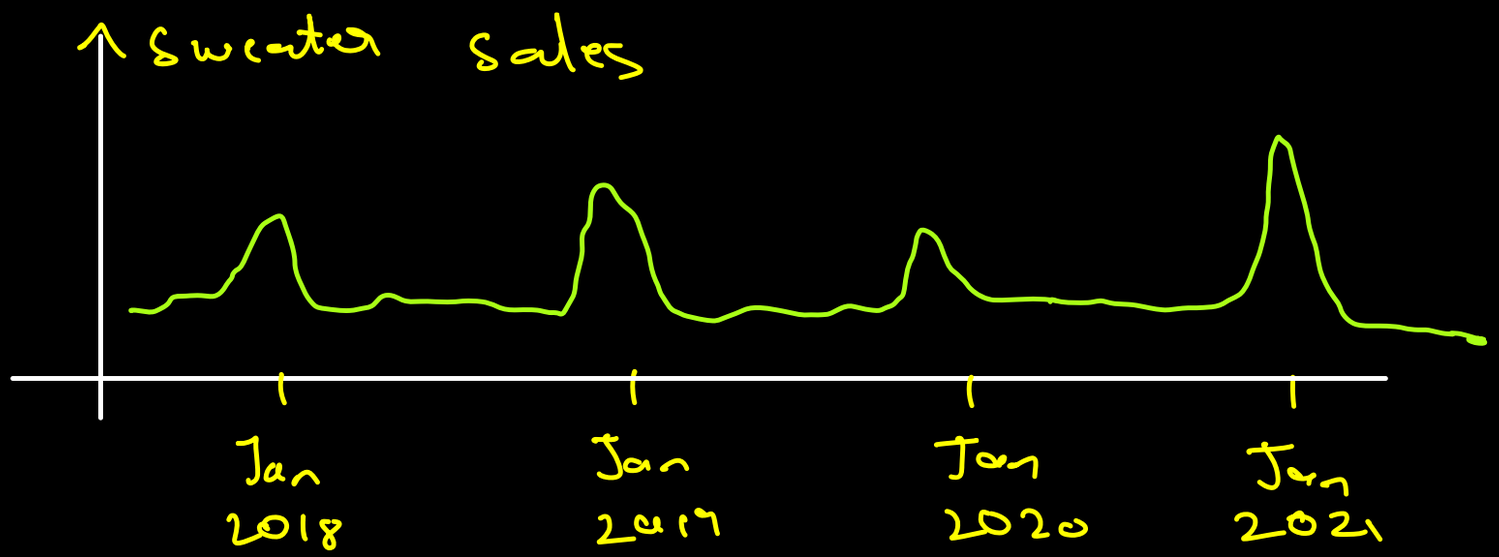
→ Business is growing [trend]

→ Some patterns repeat [seasonality]

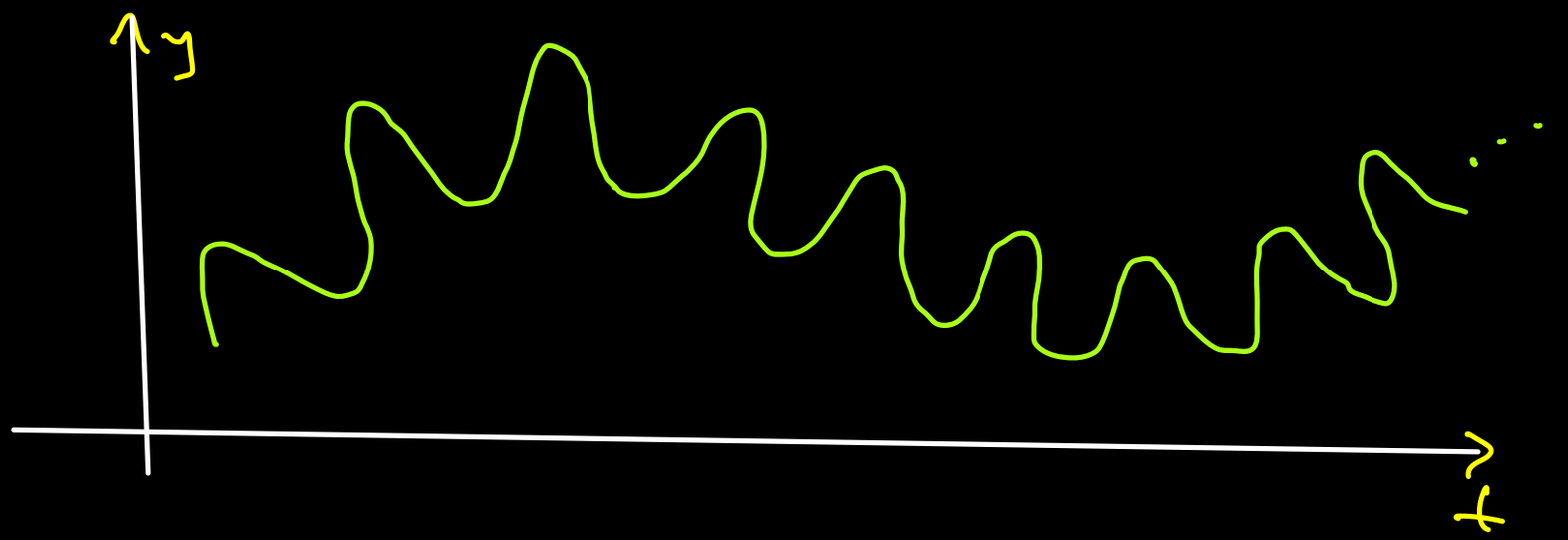

Shoe sales
Weekend
15
10
5
Mon
Mon
Mon
days
Sweater sales
Winters
100
10
2018
2019
2020
years

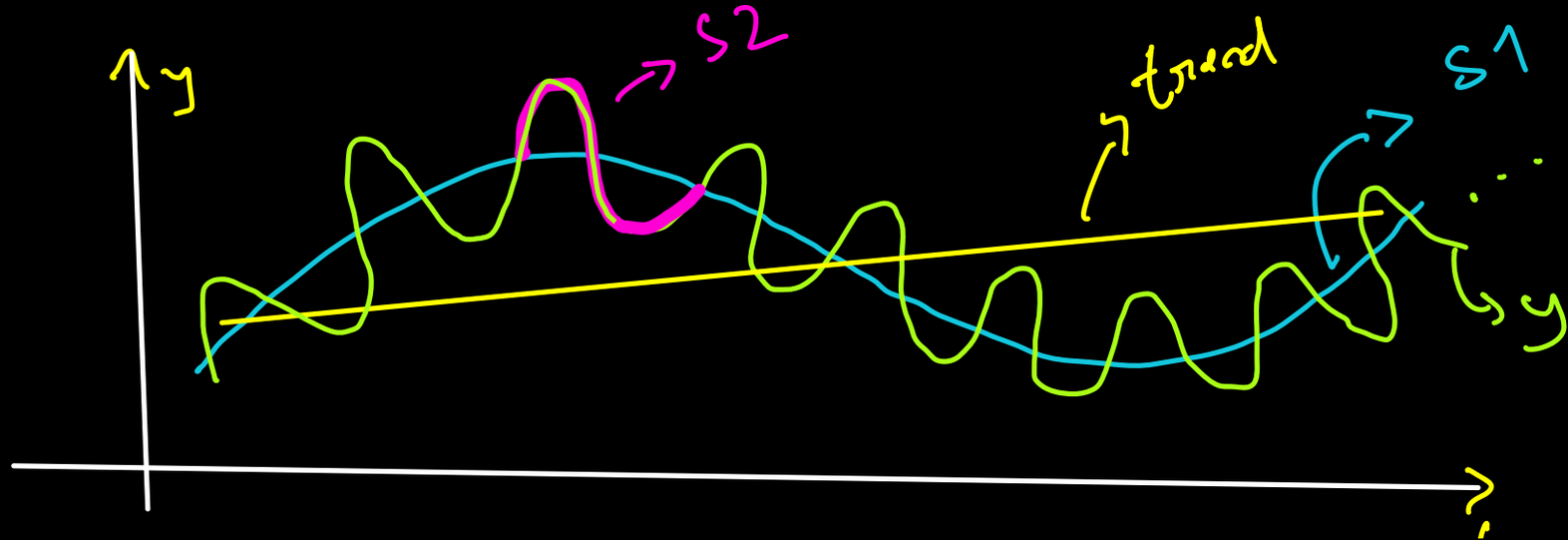
Trend

Seasonality

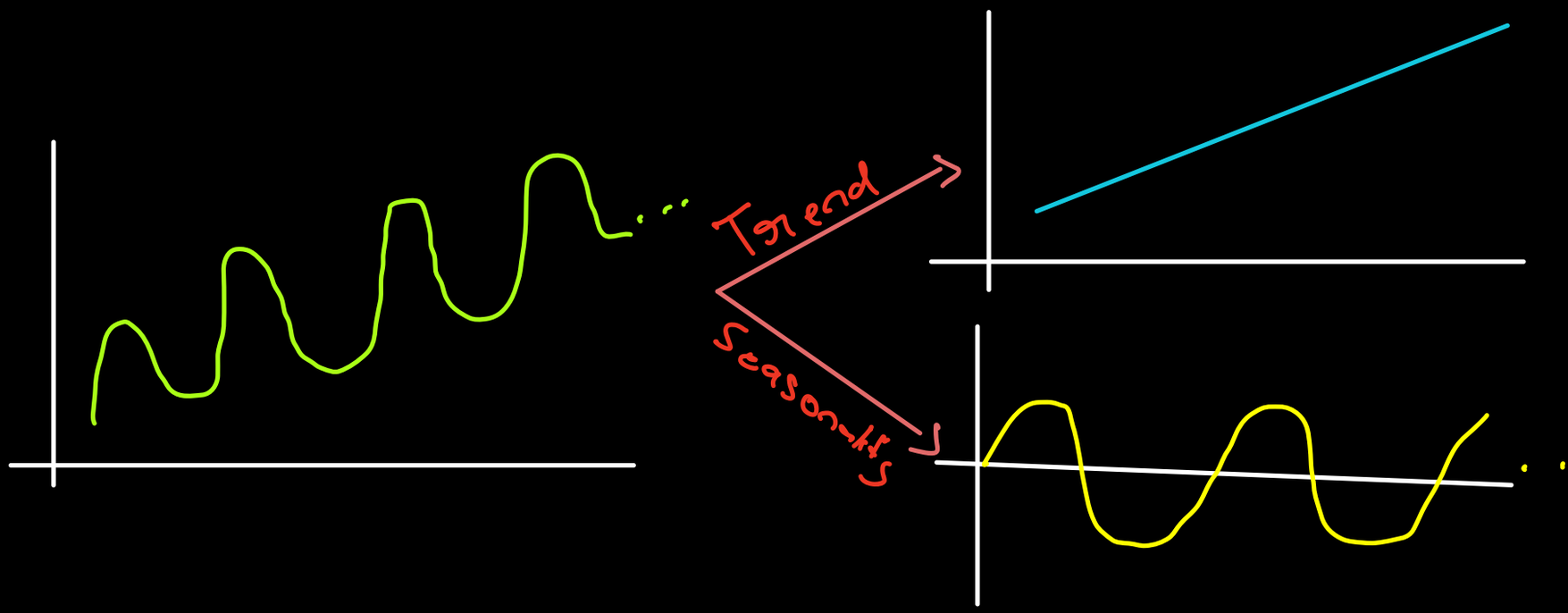


Multiple Seasonality?



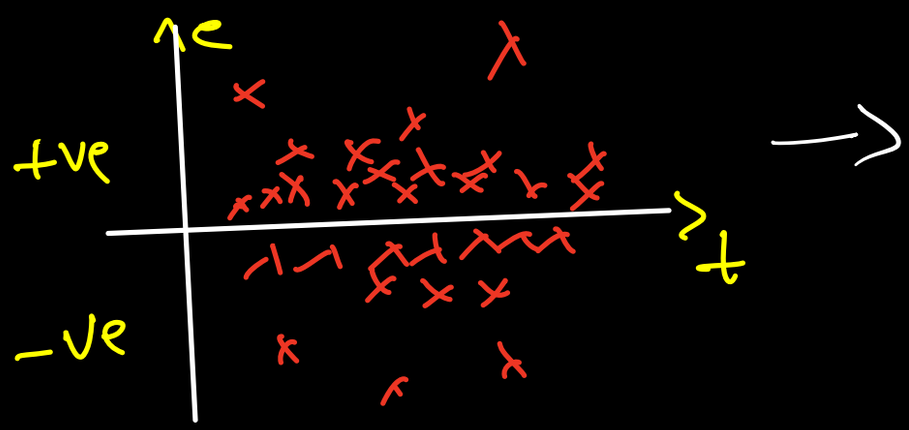


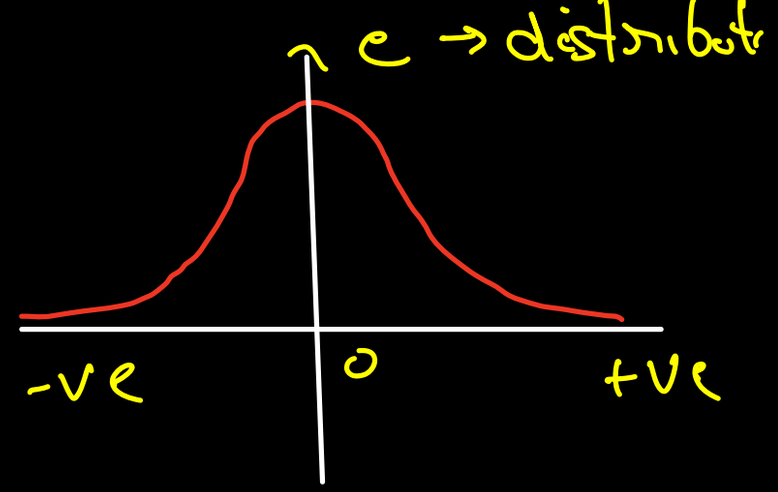
Time Series Decomposition

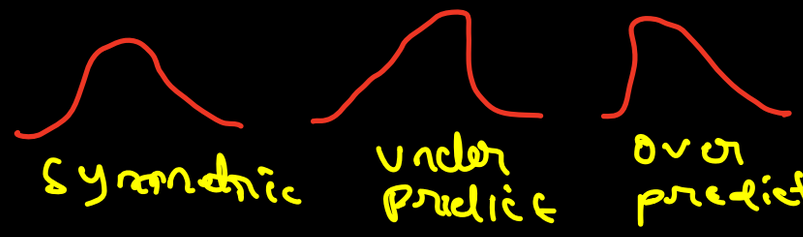


$$y(t) = \underbrace{b(t)}_{\text{trend}} + \underbrace{S(t)}_{\text{season}} + \underbrace{e(t)}_{\text{noise / residual}}$$

$$e(t) = y(t) - \underbrace{[b(t) + s(t)]}_{\hat{y}(t)}$$







How to extract ?

→ Take centered rolling avg with large window.
this is 1st estimate of trend. $b^*(t)$

→ Subtract this trend from original data.

$$s^*(t) = y(t) - b^*(t)$$

→ Take avg of all time periods for seasonality.

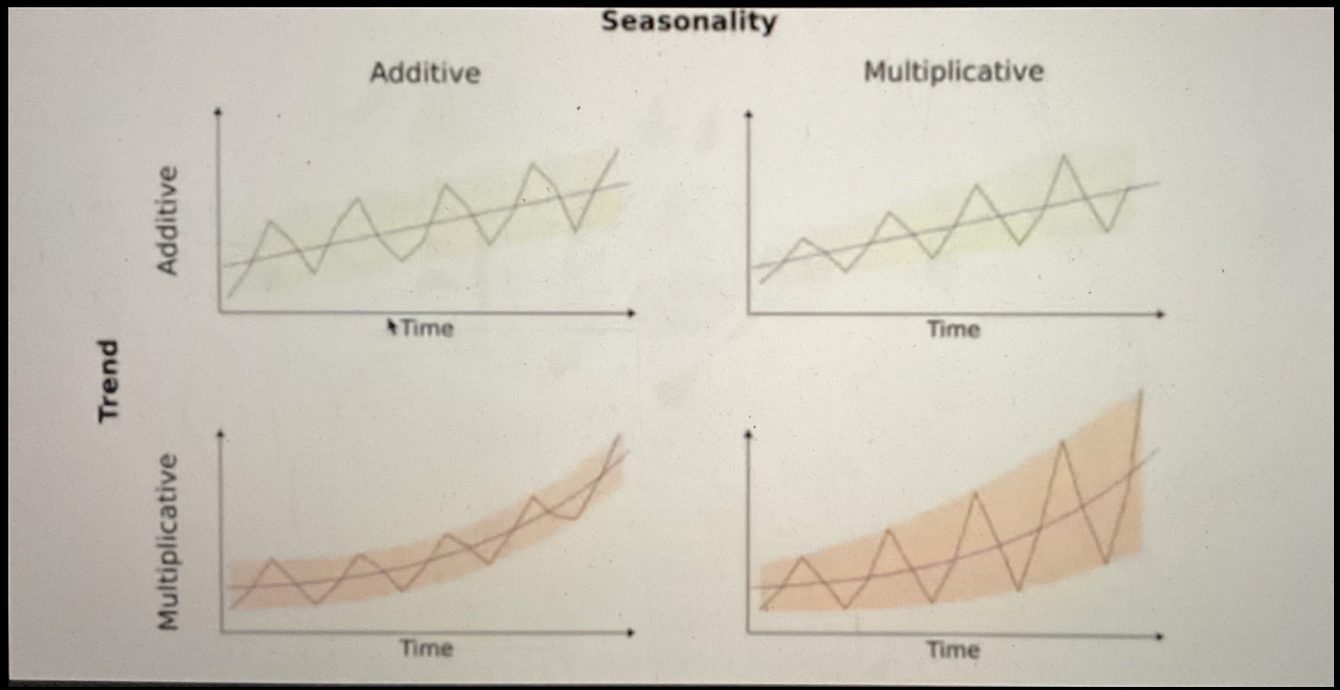
$$s(t) = s^*(t).\text{groupby}(\text{period}).\text{mean}()$$

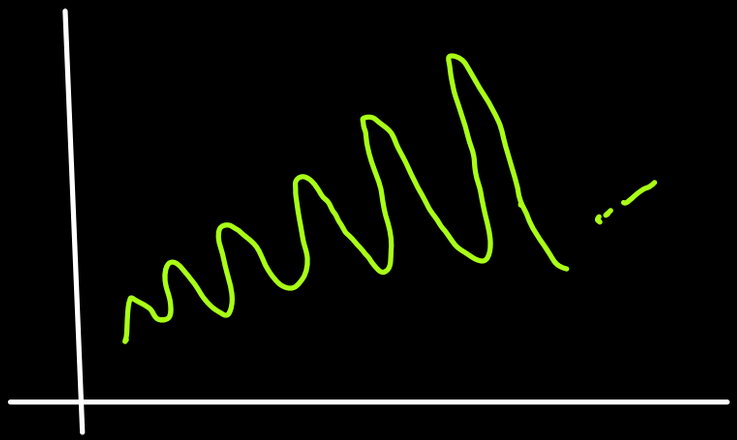
→ Subtract seasonality and find trend

$$b(t) = \text{rolling}(y(t) - s(t)).\text{mean}()$$

→ $e(t) = y(t) - b(t) - s(t)$

Types of decomposition models





$$y(t) = b(t) \cdot s(t) \cdot e(t)$$

$$e(t) = \frac{b(t) \cdot s(t)}{y(t)} = \hat{y}(t)$$