Time Series

Time Series ⟶ A series with time component.

Area of House	No of rooms	Price of House
—	—	—
—	—	—
—	—	—

You can use MLR

Regression Problem statement

Area of house	Price of house
1000	60 L
950	50 L
1100	75 L

⇓ (Rearrangement of rows)

1100	75 L	X_train was selected randomly
1060	60L	
950	50	

(timestamp) month ↓

(Month, day, year, secs, Mins, Hours)

Month	Sales
Jan	60k
Feb	50k
March	40k
—	—
—	—
—	—

Can we use MLR?

⇓

No because its time series problem

Time Series Problem

Month	Sales
Jan	50k
Feb	40k
Mar	30k

X rearrangement is not possible

Feb	40k
Jan	50k
Mar	30k

⇓ because

⇒ Time Component is involved

⇒ Here order matters.

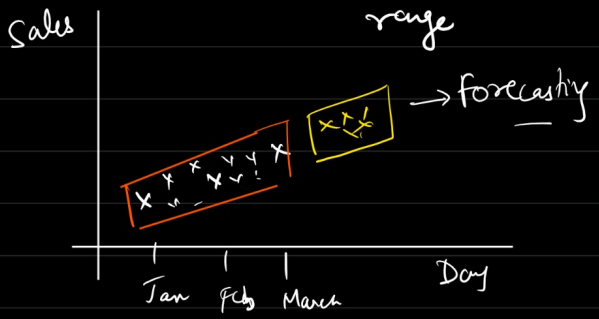
⇒ because current row depends on prev rows

i.e - March depends on Feb, Feb depends on Jan and so on.

* Interpolation ⇒ To find out the value in range itself (Prediction)

Extrapolation →

To find out the value out of the range



Most of the time (99% of the time), the test data will come in the training range

training (Area) = 1000 - 10000
testing (Area) = " (Most of times)

* Time series problem statement will extrapolation
* Based on previous History, forecast the future value.

* If outside training range ⇒ wrong prediction

* Why Not Linear Regression for time series?

→ time component is involved.
→ Because of Extrapolation, it may lead to wrong Prediction.
→ LR - Assumes linear relationship but in time series, the current observation depends on previous obs. → which is not true for non-time series data.

* Motivation

① Weather forecasting ⇒ weather, pattern: day wise, month wise, season wise
② Medical → based on previous medical history, predict future
③ finance/ — Sales, Bond price, Stock price
④ Ecommerce / Economics — Interest rate, GDP etc.

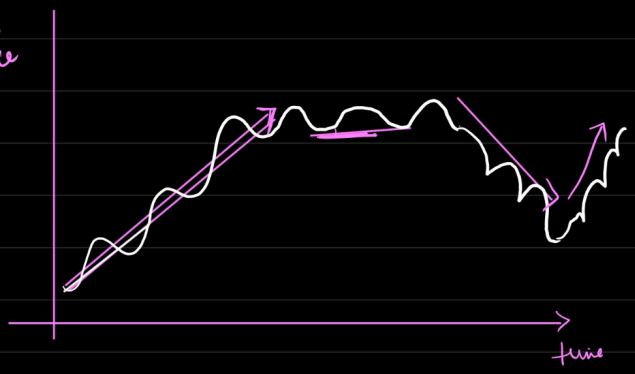
* You can use time series in every domain wherever you have a time dependent data.

Components of Time series

Analysis ↔ forecasting — time series → Airline passenger.

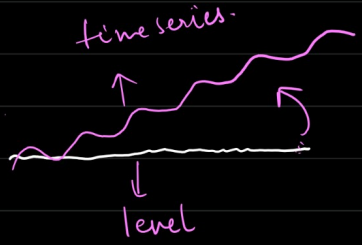
Time series Analysis

1. level
2. trend
3. season / seasonality
4. cycle / cyclicity
5. Noise



① level → the base value of a time series on which other components are added.

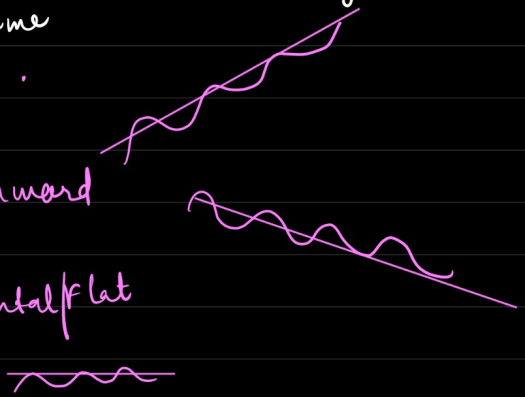
② trend — Long term movement or direction in data over a long period of time



① Upward.

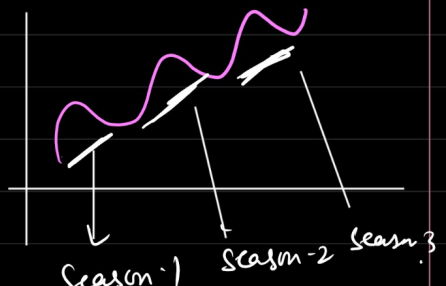
② Downward

③ Horizontal/flat



③ season :- Frequent repetation in a data over regular interval such as daily, monthly, annualy

Example → Sales of ice-cream increases in summer
→ Sales of electronic gadgets in diwali.
→ traffic in peak hours
→ No of tourist visiting every year.

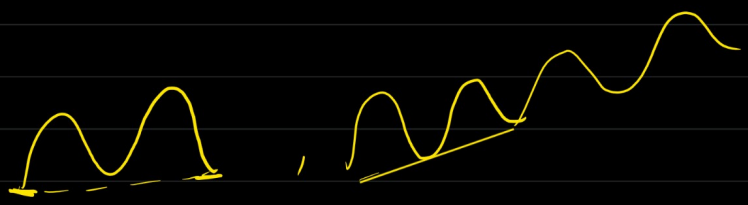


④ Cycle / Cyclicity → The fluctuation in the data over a longer period of time. These periods are not fixed and can vary.

Cycle → Noise (Anomalous behaviours, Unexpected behaviour)

Cyclicity – Season + Noise over a long period of time.

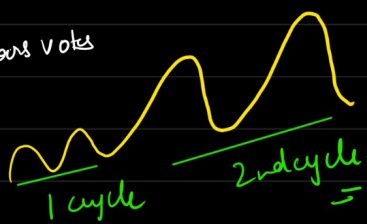
∴ Noise/Anomaly



ex. Stock price

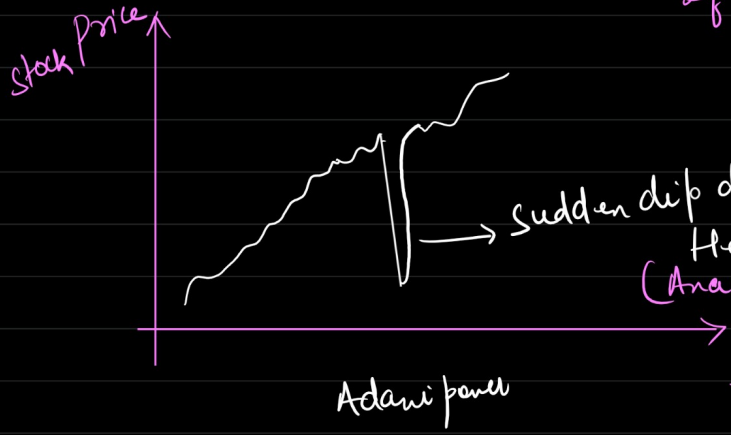
ex. Politics – every 5 years votes

ex. Economics – GDP.



⑤ Noise

↳ Anomaly / Outlier → Some uncertainty / randomness in time series data because of unexpected reason.



reasons

→ News
→ Reports
→ Pandemic
→ wars
→ elections.
→ Influencer

* Dogecoin always use to move unexpectedly after Elon Musk's tweet.

Time series

Analysis

Forecasting

* Based on components of a time series, there are two types of time series.

time series

Additive

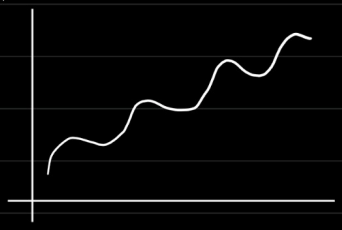
→ y_t = Trend + Seasm + Noise

→ Linear over time

→ Constant variance

→ Increased trend at same difference

Day 1	100	→ diff
Day 2	200	
Day 3	300	
Day 4	400	

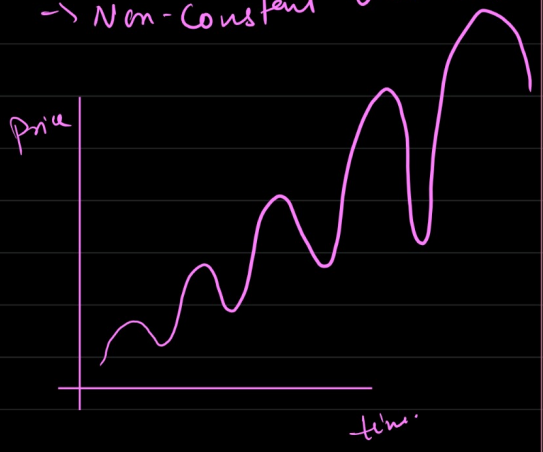


Multiplicative

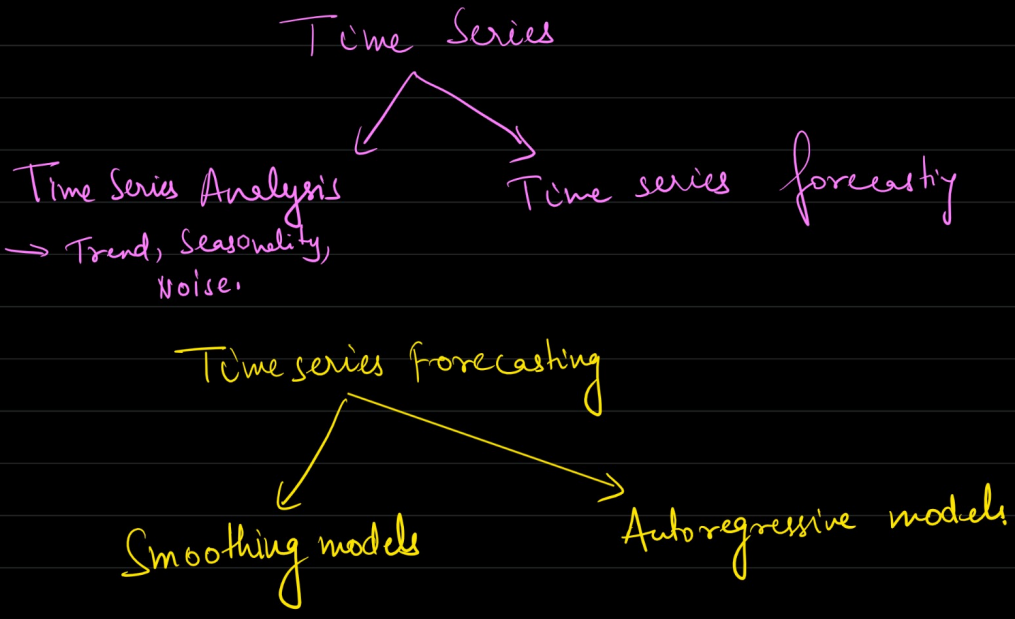
→ y_t = Trend × Seasm × Noise

→ Non-linear

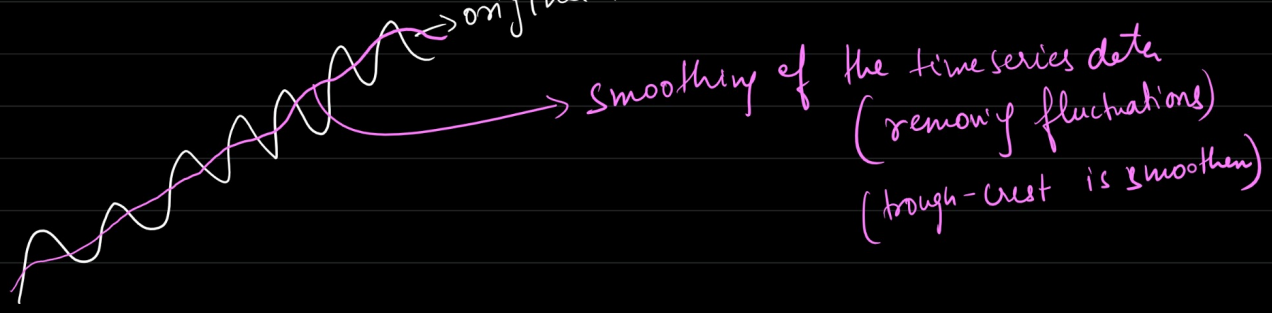
→ Non-constant variance.



Time series forecasting



* Smoothing models.



① Simple moving avg (SMA)
② Cumulative moving avg (CMA)
③ Exponential weighted Moving Average (EWMA)

① Simple moving Average

$$\text{Avg} = \frac{\text{Sum of all nos}}{\text{No of values.}}$$

$2, 3, 4, 5 \Rightarrow \frac{2+3+4+5}{4} = \frac{14}{4} = 3.5$

Moving avg.
⇓
move over the time axis in a specific window.
→ window size
→ avg value.

① Window = 3
② Avg.

		SMA
Jan	50	NA
Feb	65	NA
Mar	70	62
Apr	85	74
May	90	82
Jun	100	=
July	110	=

→ 1st avg = $\frac{50+65+70}{3}$ ⇒ 62

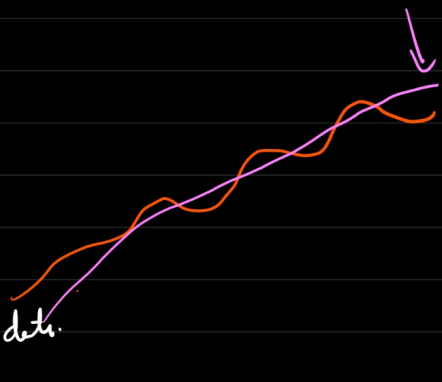
→ 2nd avg = $\frac{65+70+85}{3}$ ⇒ 74

→ 3rd avg = $\frac{70+85+90}{3}$ = 82

If window 2

→ Due to simple moving avg, the time series will smoothen.

after smoothy.



Why smoothing

→ To remove all the effects from data

→ you want to see trend of data
→ reduce effect of outlier
→ pattern recognition from the data.
→ visualisation

② Cumulative Moving Avg.

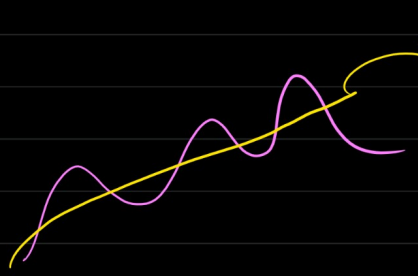
→ Find out the avg of all the datapoints upto the given time stamp

		CMA
Jan	10	10
Feb	12	$\frac{10+12}{2}$
March	15	$\frac{10+12+15}{3}$
Apr	14	$(10+12+15+14)/4$
May	16	( ) /5
June	17	( ) /6
July	18	( )/7

When?

→ for long time period.

→ give you exponential trend



③ EMA or EWMA

→ We give more weightage/ importance / priority to the recent data point / timestamp.

$$V_t = \beta V_{t-1} + (1-\beta)\theta_t$$

V_t = EMA at time t

$\beta \Rightarrow 0 < \beta < 1 \Rightarrow$ generally its 0.9.

V_{t-1} = EMA at previous timestamp

$\theta_t \rightarrow$ Data at current timestamp t.

$D_1 \rightarrow$ less weight
D_2
D_3
D_4
$D_5 \rightarrow$ recent obs / more weight to be given

SMA or CMA

disadvantage

→ It gives equal priority all the time period values.

→ In time series current observation is highly influenced by recent observation.

Jan 10
Feb 15
Mar 16

∴ You need to give priority to recent observations.

			EMA
V_0	D_1	25	0 or 25 ($V_0=0$ or $V_0=25$)
V_1	D_2	13	1.3
V_2	D_3	17	2.87
V_3	D_4	31	-
V_4	D_5	43	-

$V_1 = \beta \times V_0 + (1-\beta)\theta$ ($\beta = 0.9$)

$V_1 = \beta \times 0 + (1-0.9) \times 13$

$= 0 + 0.1 \times 13$

$= 1.3$

$V_2 = \beta \times V_1 + (1-\beta)\theta$

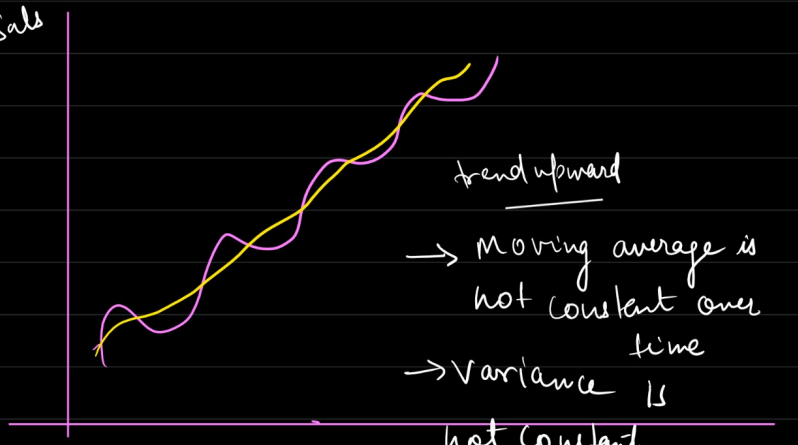
$= 0.9 \times 1.3 + 0.1 \times 17$

$= 1.17 + 1.7$

$V_2 = 2.87$

Stationary and Non stationary time series

Month	Sales
Jan	40
Feb	50
Mar	60
	70
	80
	90



trend upward

→ Moving average is not constant over time

→ Variance is not constant over time.

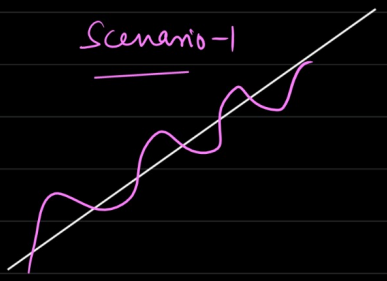
* You can build a smoothening time series on any time series data.

Time series forecasting

- Smoothing models
- Autoregressive model

* To build autoregressive models, statistical properties of a time series like mean / variance should be constant (not change over time)

Scenario-1



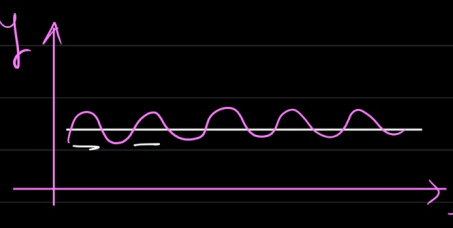
Scenario-2

→ It will be easy to build a Autoregressive forecasting model in Scenario-2

Non stationary → Mean - variance will not be constant

Stationary → Mean - variance will be constant

Constant → Over time, value is not varying (changing)



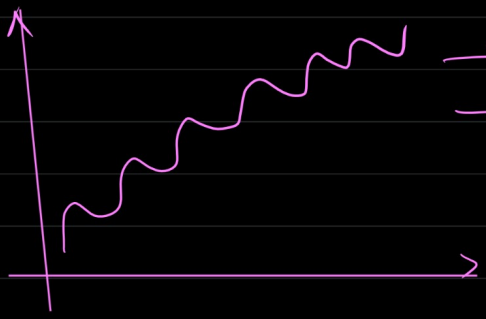
→ trend is flat | Horizonal
→ variance is constant

ML → Data ingestion → Analysis → Preprocessing | transformation → Model building

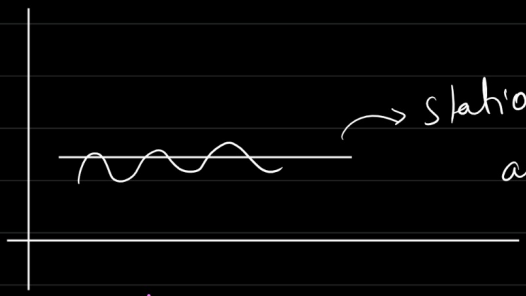
Time series → Time series is
Stationary (mean | variance constant) ⇒ Model (ARIMA, SARIMA)
or
Non-stationary
↑ Check.
if non stationary → convert to stationary time series

* To check if a time series data is stationary or Non-St

① Visualization



→ trend ⇒ upward
→ Moving average is increase over time
→ variance is not constant
⇓
TS is non stationary



→ stationary as moving avg is constant.

② Statistical based test

ⓐ ADF (Augmented dickey fuller test)
H_0: TS data is non stationary
H_A: TS data is stationary

Pvalue ≤ 0.05 → reject the H_0
conclusion ⇒ TS data is stationary

② KPSS test
(Kwiatkowski-Phillips-Schmidt-Shin test)
H_0 → my data is stationary.

* How to Convert non-stationary time series data to Stationary time series data.

① Differencing
② log transformation → take the log of value
③ root → take the sq root of value
④ seasonal adjustment.

ARIMA
I → Integrated stands for differencing

Differencing = $y_t - y_{t-1}$
(current) (previous)

Month	Price (y)	1st difference	2nd order differencing	3rd order difference
Jan	5	NA	NA	
Feb	10	10-5 = 5	NA	
Mar	6	6-10 = -4	-4-5 = 9	
Apr	8	8-6 = 2	2-(-4) = 6	
May	15	7	7-2 = 5	
June	7	-8	-15	

→ check stationary

→ After differencing check if TS is Stationary
- Stats test (ADF)
- Visualisation.

→ if stationary build the model
else do differencing again and keep doing so untill you get a stationary TS.

ACF, PACF and Autoregression

ACF → Auto correlation function
PACF → Partial Auto Correlation function.

ACF → Auto + Correlation

Auto → Correlation itself in the feature.

Correlation → Relationship b/w two feature.

Date	y
Jan	10
Feb	20
Mar	30
Apr	40

↓ time series data

* ACF measures the correlation between time series & its lag value

Month	y_t	1st lag	2nd lag	3rd lag.
Jan	10	NA	Na	Na
Feb	25	10 (Jan)	Na	Na
Mar	35	25 (Feb)	10 (Jan)	Na
Apr	42	35 (Mar)	25 (Feb	10 (Jan)
May	50	42	35 (Mar)	25 (Feb)
June	55	50	42	35
July	62	55	50	42

y_t | y_{t-1} | y_{t-2} | y_{t-3} ...

Auto correlation

{ Pearson, Spearman rank, Kendall }

Corr (y_t , y_{t-1})

Cor (y_t , y_{t-2})

Corr (y_t , y_{t-3})

Partial Autocorrelation Function (PACF)

	Month	y_t	$y_t - y_{t-2}$
t-6	Jan	10	
t-5	feb	18 X	
t-4	Mar	25	15
t-3	April	29 X	
t-2	May	35	10
t-1	June	45 X	
t	July	50	15

25	15
35	10
50	15

PACF

$y_t - y_{t-3}$

$y_t - y_{t-4}$

* Auto regression

Auto regression

regression with itself.

regression

x — y

x — IV

y — DV

$y = mx + c$

slope dip intercept

MLR ⇒ $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \dots \theta_n x_n$

$\theta_0 \theta_1 \theta_2$, = Coeff.

$y_t = \psi y_{t-1} + c$

$y_t = \psi_1 t_{-1} + \psi_2 y_{t-2} + c$

y_t — value at current timestamp

ψ — Coeff

$y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + \psi_3 y_{t-3} + c$

c — constant.

ε → Error.

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varphi_3 y_{t-3} + \cdots \varphi_n y_{t-n} + c$$

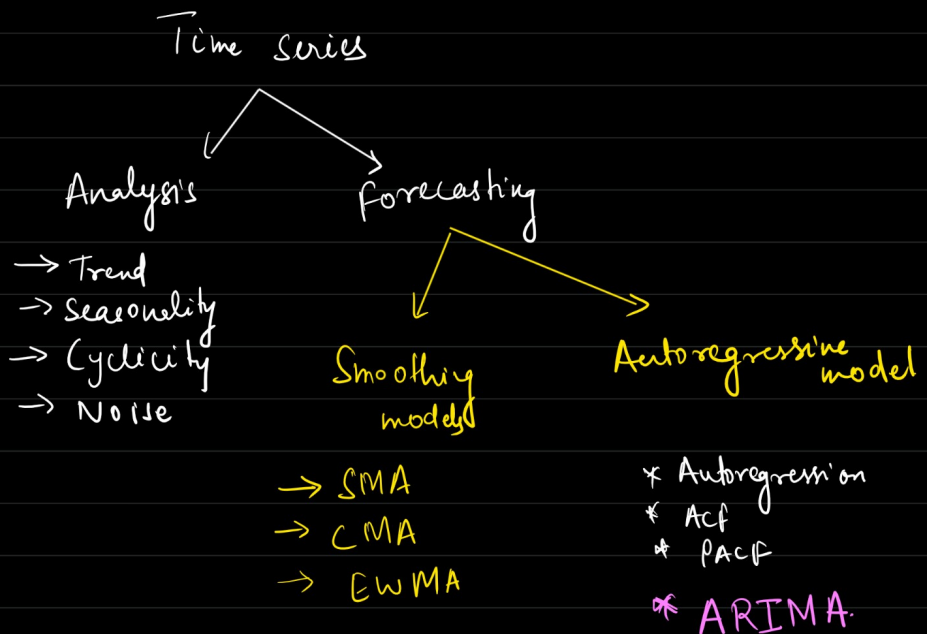
How to decide if you have to go till $t-1$, or $t-2$ or $t-3$??

⇓

ACF | PACF

ARIMA

ARIMA



ML → Data → LR, SVR, DTR and so on.

TS data → Forecasting models. (Smoothing models already we have seen)

- → AR
- → MA
- → ARIMA
- → SARIMA
- → SARIMAX

AR

(Autoregression)

⇓

p

(0, 1, 2, ... n)

lag value

PACF

Partial autocorrelation function

(Correlelogram)

I

(Integrated)

⇓

q

(0, 1, 2, 3 .. n)

lag value

Differencing

(Stationary)

MA. model.

(Moving average)

⇓

q

(0, 1, 2, 3 .. n)

lag value.

ACF

(Autocorrelation function)

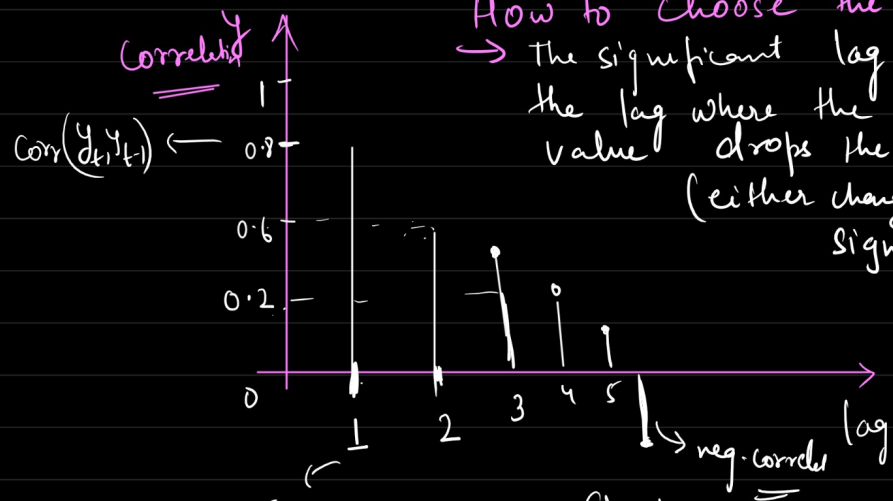
(Correlelogram)

ACF (for MA)

$Corr(y_t, y_{t-1})$ ✓

$Corr(y_t, y_{t-2})$

$Corr(y_t, y_{t-3})$

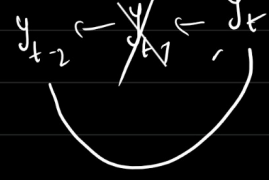


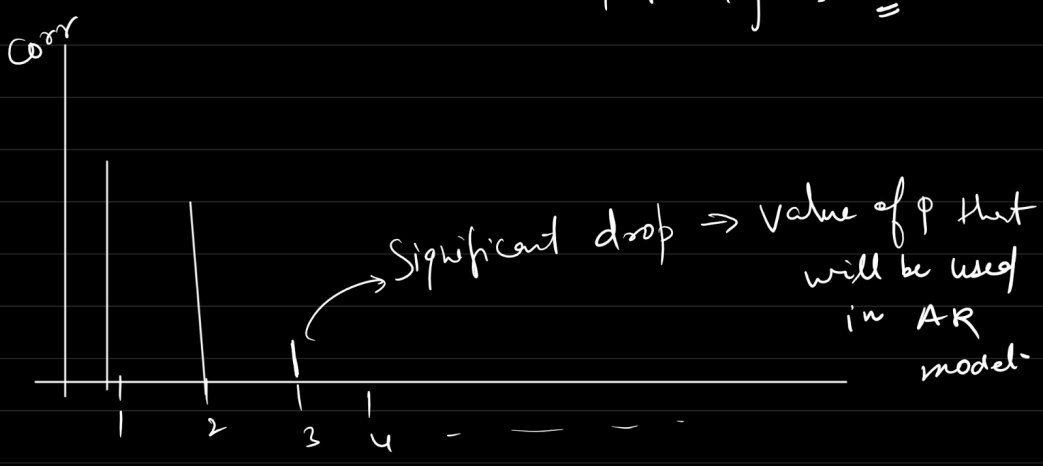
How to choose the lag?

→ The significant lag will be the lag where the correlation value drops the most. (either changes the sign, dropped significant)

Significant lag is - 5.

PACF → p → AR





* Autoregression

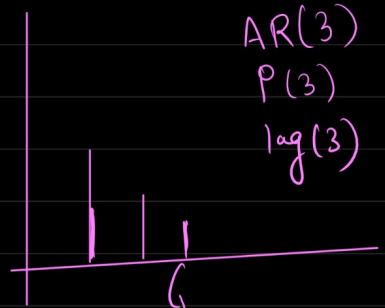
$$y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots \psi_n y_{t-n} + C$$

Till how much lag we have to model??

PACF ⇒ Significant lag → P

Day	value	y_{t-1}	y_{t-2}	y_{t-3}
1	10			
2	20	↓		
3	30			
4				
5	40	20		
6	~~50~~			
7	60	20		

(y_t, y_{t-1}) , (y_t, y_{t-2}) (y_t, y_{t-3})



AR(3)
P(3)
lag(3)

$$y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + \psi_3 y_{t-3} + C$$

for 2 lags ($P=2$)

$$y_t = \psi_1 t_{-1} + \psi_2 y_{t-2} + c$$

* We decide p (PACF) of AR, by seeing PACF plot. and q (ACF) of MA by seeing ACF plot.

ACF	lag 1	lag 2
10	Na	Na
20	10	Na
30	20	10
40	30	20
50	40	30

* Integrated - Differencing

D_0

D_1 D_2 D_3 D_4 D_5

d_1 d_1 d_1

D_2

y_t y_{t-1}

$(y_t - y_{t-1}) \rightarrow D = 1$

$(y_t - y_{t-2}) = D = 2.$

* Moving Error :-

(No same as Smoothy model average)

→ It models the error.

$$y_t = \varepsilon_{t-1} \psi + c$$

		Pred
D_1	10	8
D_2	15	?
D_3	20	
D_4	25	
D_5	30	

Last error $= 10 - 8 = 2$

$$y_t = \varepsilon_{t-1} \psi_{t-1} + \varepsilon_{t-2} \psi_{t-2} + \varepsilon_{t-3} \psi_{t-3} \cdots \varepsilon_{t-n} \psi_{t-n} + c$$

$\varepsilon \rightarrow$ error

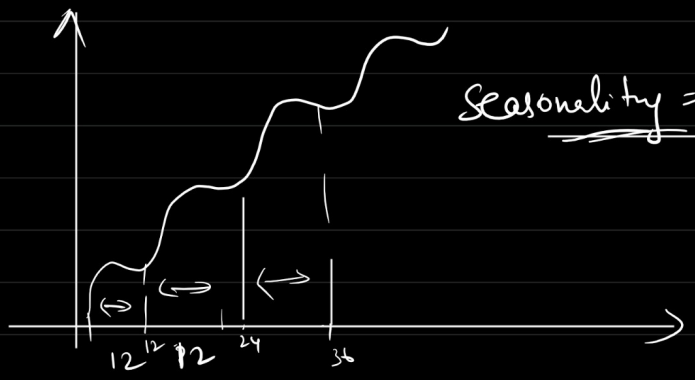
$$ARIMA = \underbrace{y_{t-1}\psi_{t-1} + y_{t-2}\psi_{t-2} + \cdots + y_{t-n}\psi_{t-n} + c}_{AR(p)} + \underbrace{(y_{t-1} - y_{t-2} - y_{t-3} \cdots - y_{t-n})}_{\text{Integrated } d} + \underbrace{\varepsilon_{t-1}\psi_{t-1} + \varepsilon_{t-2}\psi_{t-2} \cdots \varepsilon_{t-n}\psi_{t-n} c}_{MA(q) \rightarrow \text{Acf.}}$$

AR(1)
I(2)
MA(1)
} → ARIMA

$$y_t = (y_{t-1}\psi_{t-1} + c) + (y_t - y_{t-1} - y_{t-2}) + \varepsilon_{t-1}\psi_{t-1} + c$$

SARIMA.

Seasonality



$$(p,d,q)(P,D,Q)_s$$

$$\Rightarrow (1,1,2)(1,0,2)_{12}$$

→ Same p, d, q for seasonal component.

SARIMAX
↓
Seasonal

Exogenous variable (outlier, extra effect)
Noise

$$(p,d,q)(P,D,Q)_s$$
$$(0,1,2)(1,0,2)_{12} \cdot y_t \text{ (some extra variable)}$$