Date: 11/06/2023

Time Series

In time series based problem we are provided with time-based features that means o/p feature or forecasting to be done in time series based problem statement is dependent on time based I/P features.
For Eg; Day-wise Sales of a product.

Non time Series

Eg; Predicting price of house based in non time dependent features such as Size, Location and No. of bedrooms

House Price Prediction → Regression

Size	location	Bedrooms	Price

$Y = mx + c$ → linear eq

$Y = m_1x_1 + m_2x_2 + m_3x_3 + \cdots m_nx_n + c$

Time

time Series

For Eg; Day-wise Sales of a product.

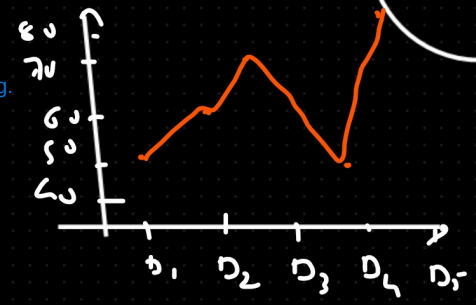
Time based feature can have any of the below timestamp format

hour, min, sec, Day, month, year → timestamp → Sales (corr)

Day 1	50lc
Day 2	60lc
Day 3	70lc
Day 4	40lc
Days	75lc

Intersection → Sales, Mongment

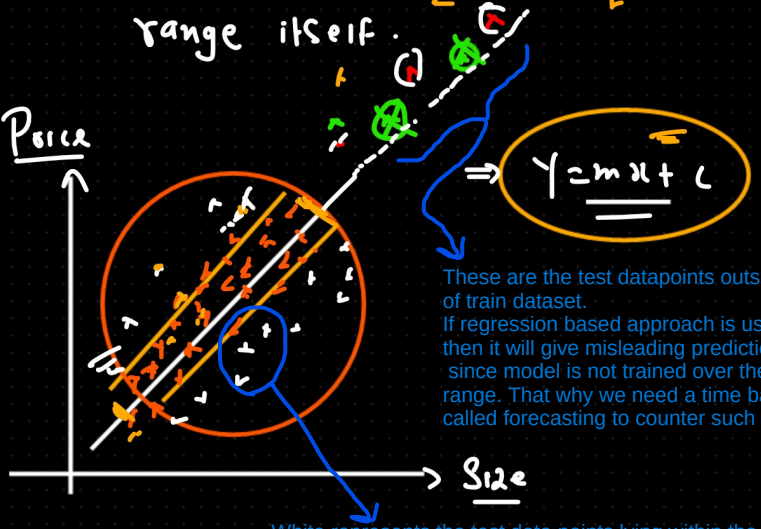
Graphical representation of time series based question can be termed as time based forecasting.



Ques Can we Solve this ts with linear regression?

To answer this we need to understand Interpolation and Extrapolation concepts

Interpolation ⇒ to find out the value in the range itself.



Best fit line in this case is infinite line. Training dataset captures only a portion of this infinite best fit line. In interpolation Test data points or new datapoints on which prediction is supposed to be made lie within the same range of training datapoint. Whereas, In Extrapolation these data points will lie outside the proximity of the train dataset often denoted by different time frame.

These are the test datapoints outside the range/proximity of train dataset.
If regression based approach is used for such datapoints then it will give misleading predictions since model is not trained over the datapoints lying in this range. That why we need a time based approach called forecasting to counter such cases.

White represents the test data points lying within the same range of train dataset.

Orange → training Data we have in the Particular range

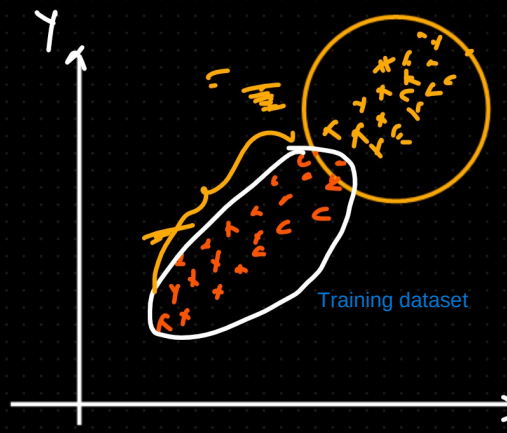
test data

Non-time series problem Data

Prediction and Forecasting may look similar but are different considering the range of Test dataset. If Test data set within the same time frame as train dataset then Interpolation else Extrapolation.

Since Interpolation consider same timeframe so the features are independent of time. Which in case of Extrapolation is not the case hence features in it is time dependent.

- **Extrapolation** ⇒ to find out the value out of Range.



Testing /New data points of new timestamp different from the one chosen while training the model.

Forecasting

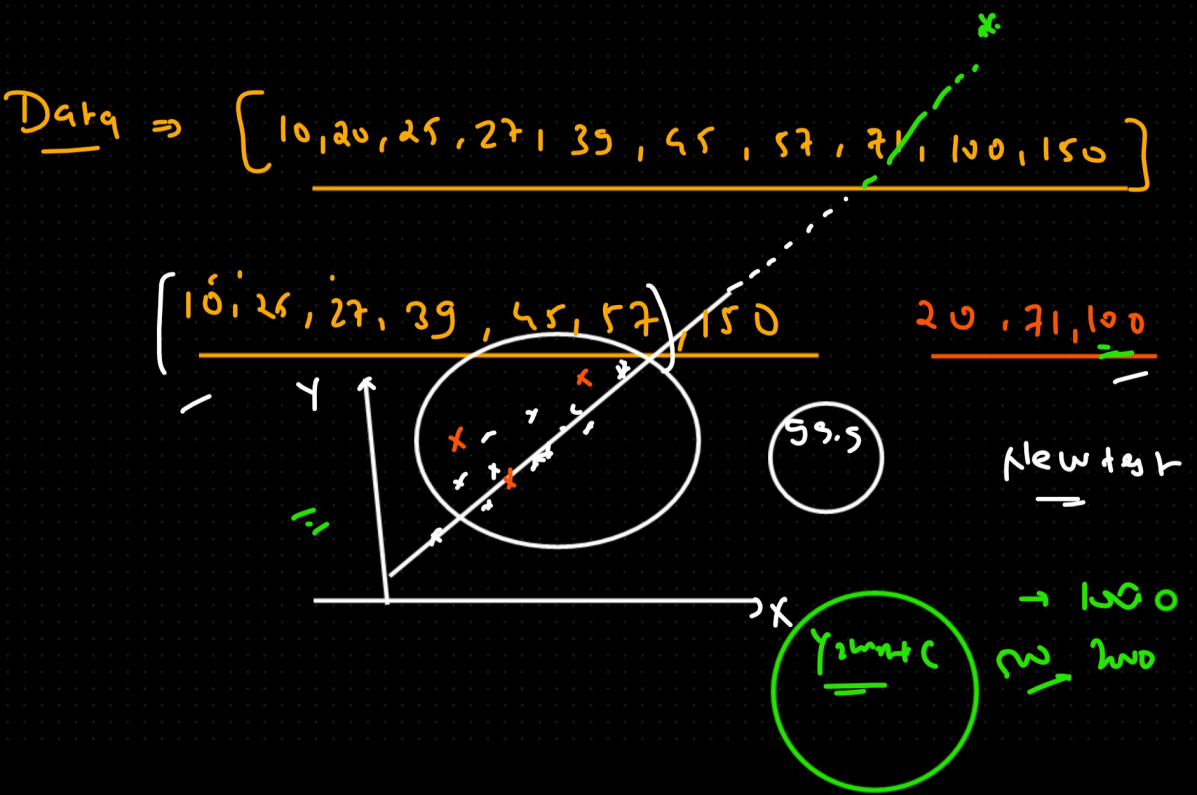
Timeseries data → extrapolation

time series Problem statement will be extrapolation

- Based on Previous history forecast the future value.

Ques Can we Solve this ts with linear regression?

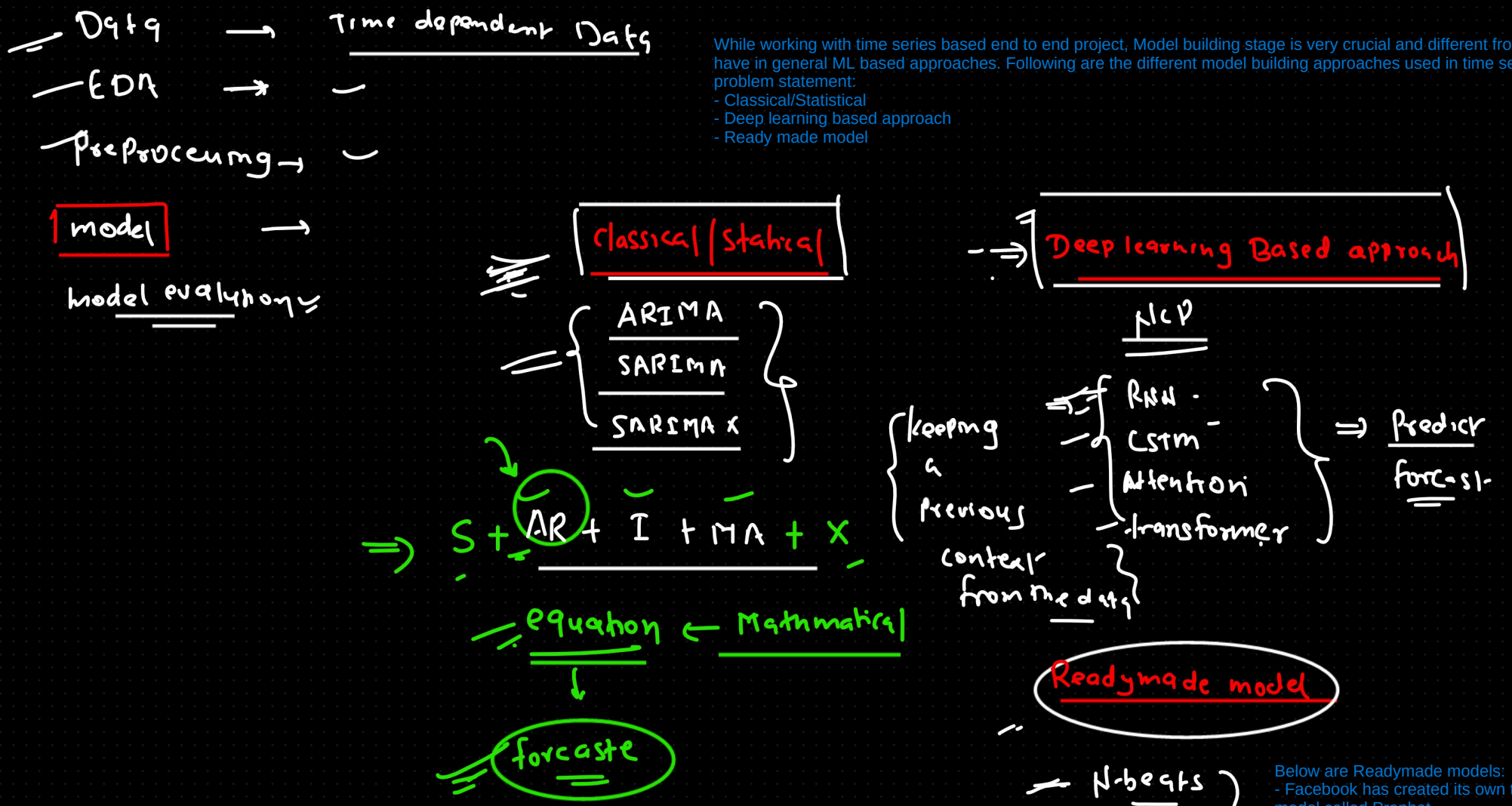
1. Because of extrapolation. we can not do.
2. Because of outlier we can not use.
3. Time Series TS Data is more complex
4. Linear reg. assume that there should linear realtionship but in time you won't get linear realtionship
5. In Non-TS Data there is no such effect of time but in TS — Data there is effect of time meaning current TS is Dependent on prev. time stamp.



Example

1. economics forcasting → GDP, interest rate, inflation rate
2. Finance → Sales, Bond Price
3. Weather forcasting ⇒ Pattern of Diff season, Prediction of weather
4. Medical ⇒ Based on prev. history of patance need to predict future condition.

Time series ⇒ time dependent data ⇒ any domain



While working with time series based end to end project, Model building stage is very crucial and different from what we have in general ML based approaches. Following are the different model building approaches used in time series based problem statement:
- Classical/Statistical
- Deep learning based approach
- Ready made model

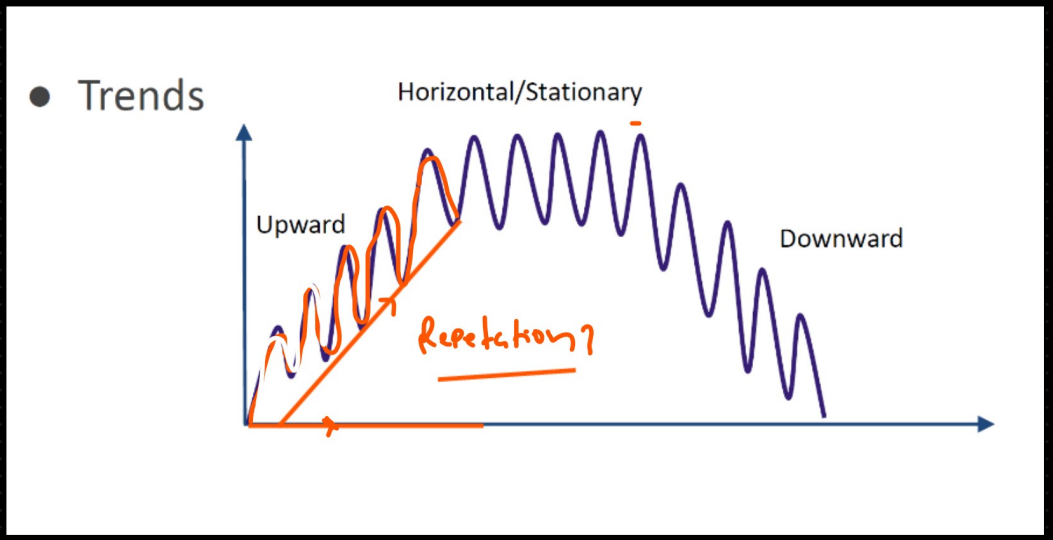
Below are Readymade models:
- Facebook has created its own time series model called Prophet.
- Similarly, Google has also created it's own time series based model called Temporal Fusion transformer.
- N-bets can also be used for time series based problem statement.

Google these to know more.

Lets, now discuss some of the common terminologies to be used while working with a time series based problem statement:
- Trend
- Season
- Noise
- Cycle

Timeseries ⇒

1. trend
2. Season
3. Noise
4. Cycle



Trend ⇒

1. Upward trend
2. Downward trend
3. flat (horizontal)

Trend tells us that how the data is behaving with the time. It has 3 types: Upward, Downward and horizontal or stationary.

Season ⇒ = frequent (repetation) (Daily, monthly, yearly)

1. Sales of the ice-creame in summer
2. traffic situation at 6 pm in my area
3. tourist in goa at the end of year (Dec, Jan)

Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year.

Noise ⇒ Some uncertainity or some randomness in my data. bcoz of unpredictable reason.

my data = our data; bcoz = because

UNPREDICTABLE REASON =>

Pandamic => corona

War => R|U

Report => hindenberg report

Eg; Ronaldo removed Pepsi bottle during interview that affected stock price of Pepsi company.

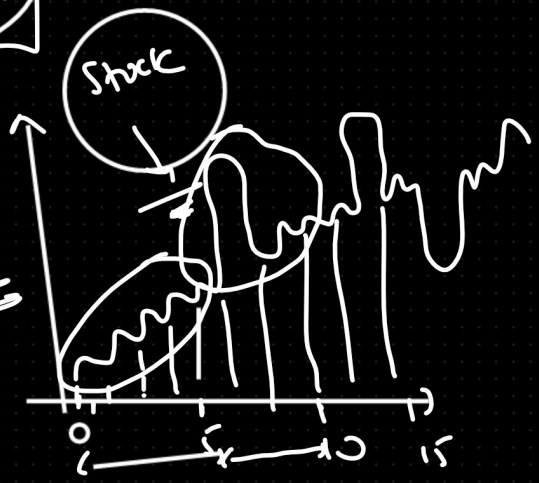
Eg; Stock market crashes at time of war, pandemic, publishing report that defames the company.

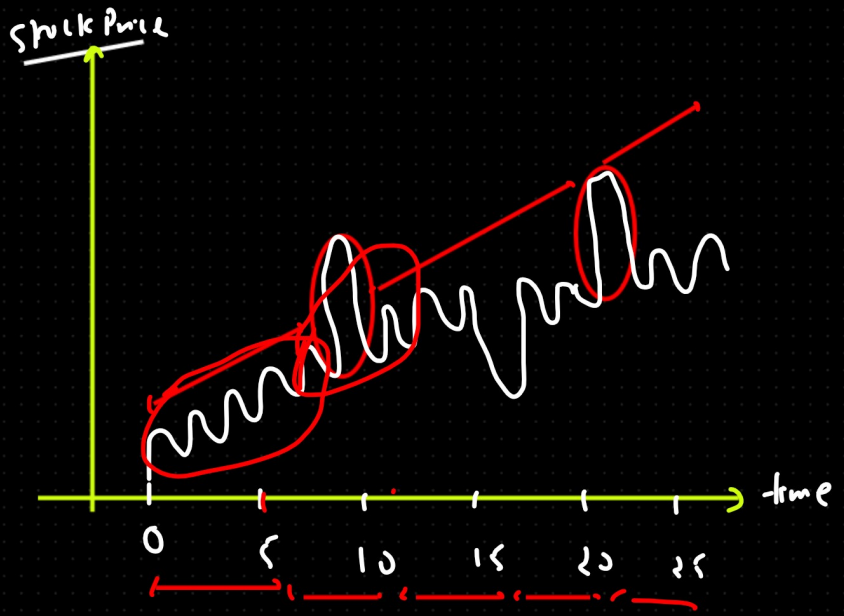
Cycle => time series behaviour over the long Period of time

Cycle=Season + Noise
When GDP report is published after every 1 year(Season) then based on the sentiment of the report, their is sudden fluctuation(Noise) on the price of stocks listed on stock market.

Cycle => seasonality + Noise(fluctation)

- Election => 5-year => cycle
- Census Counting => 10 year
- Economics => GDP





trend → upward
Seasonality → yes
Noise → yes
⇒ Non Stationary

Cycle → long period

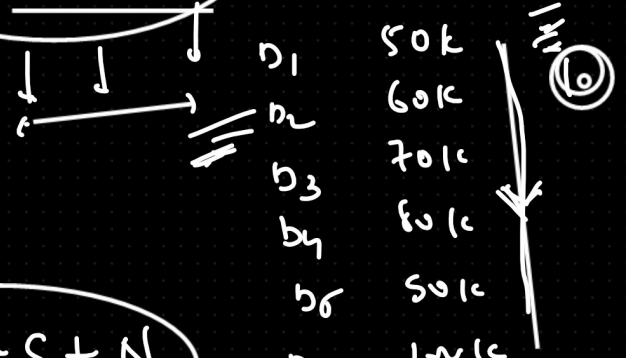
Seasonal effect + Noise

Two components of timeseries:
- Multiplicative
- Additive

Component ⇒ Multiplicative ⇒ $y_t = T \times S \times N$

Trend Season Noise

Applied when there is
1. Non-linear Pattern
2. Non_Constant var variation



Additive ⇒ $y_t = T + S + N$

Applied when there is
1. linear Pattern
2. It is having constant var. variation

Moving Average

1. Simple moving average (SMA)
2. Commulative moving average (CMA)
3. EMA or ELWMA

Why we Calculate MA?

SMOOTHING of TIME SERIES DATA

SMA =>

Simple moving average

Average of [10, 12, 15, 13, 11]

$$\frac{10+12+15+13+11}{5} = 12.5 \text{ Approx}$$

Moving => move over the time axis in given window size

Moving + average

Window Size → Calculating average

Move over the time axis in the given window size followed by calculating the average of the corresponding time dependent output feature for the respective windows.

Moving average =>

Lets understand the simple moving average using the below example:

Window Size = 3

D_1	10	NAN
D_2	12	NAN
D_3	15 →	12.33
D_4	13 →	13.33
D_5	14 →	14
D_6	16 →	14.33
D_7	17 →	15.66

representing 1st average upto D3. Above since there out not in the window size so the respective average value is replaced by NAN

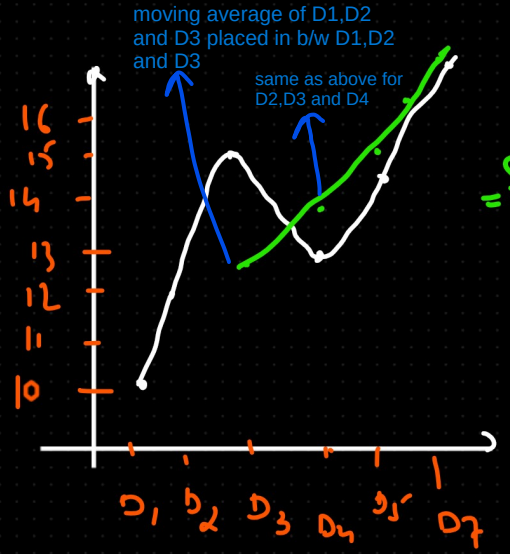
$$1st\ avg = \frac{D_1+D_2+D_3}{3} = \frac{10+12+15}{3} = 12.33$$

$$2nd\ avg = \frac{D_2+D_3+D_4}{3} = \frac{12+15+13}{3} = 13.33$$

$$3rd\ avg = \frac{D_3+D_4+D_5}{3} = \frac{15+13+14}{3} = 14$$

$$4th = \frac{D_4+D_5+D_6}{3} = \frac{13+14+16}{3} = 14.33$$

$$5th = \frac{D_5+D_6+D_7}{3} = \frac{14+16+17}{3} = 15.66$$



In above plot white line represents general time series plot whereas, green line represents the smoothen time series curve obtained using SMA.

Why we do Smoothing ?

TYSXN

1. to remove all the effect from the data
2. Pattren recognization from the data (to summarize the overall trend of the data)
3. You can easily analysis trend of Data
4. You Can reduce the effect of outlier
5. You can easily visulizing the data.

CMA => Find out the avg of all the Data Points upto the given time stamp

D_1	10	⇒	$D_1 = 10$
D_2	12	⇒	$\frac{D_1+D_2}{2} = 10+12/2 =$
D_3	15	⇒	$\frac{D_1+D_2+D_3}{3}$
D_4	14	⇒	$\frac{D_1+D_2+D_3+D_4}{5}$
D_5	16	⇒	$D_1 + \ldots D_5/5$
D_6	17	⇒	$D_1 + D_2 + \ldots D_6/6$
D_7	18	⇒	$D_1 + D_2 + D_3 \ldots D_7/7$

For nth datapoint CMA will be this =

$$\frac{D_1 + D_2 + D_3 + \ldots + D_n}{n}$$

Ans

1. It will give you the exponential trend.
2. We use CMA for the long time period based time series problem statement.

Now with the values of respective cumulative averages we can plot the smoothen curve of time series based plot.

EMA or EWMA ⇒ in EMA we give more weightage to the recent Data point or give More weightage to the recent time stamp.

(Exponential moving average)

Please note that SMA and CMA are the non weighted moving averages. Whereas, EMA is the weighted moving average technique.

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

V_t = EMA at time t

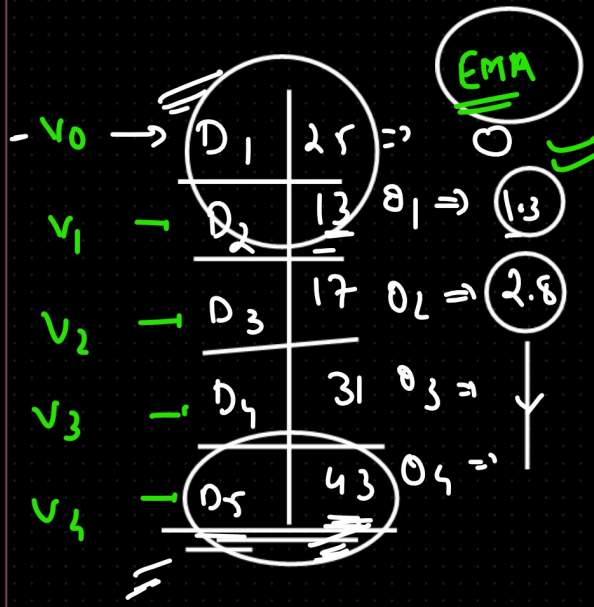
β = $0 < \beta < 1$ ⇒ 0.9

Beta is the parameter whose value range b/w 0 and 1. In most of the cases it is taken as 0.9

V_{t-1} = EMA at Previous time stamp

θ_t = Data value at t time stamp

V0 will be 0 or 25 based on the research paper documenting this approach



$V_0 = 0$ or $V_0 = 25$

$V_1 = \beta \times V_0 + (1-\beta)\theta_1$

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

$= 0 + 0.1 \times 13 = 1.3$

$V_2 = \beta \times V_1 + (1-\beta)\theta_2$

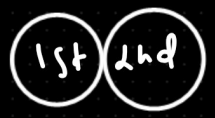
$= 0.9 \times 1.3 + (1-0.9) \times 17$

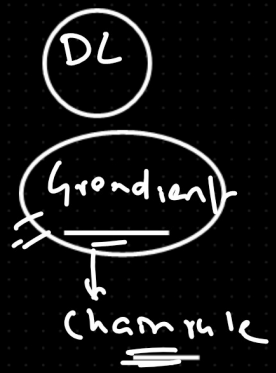
$= 1.17 + (0.1 \times 17)$

$= 1.17 + 1.7$

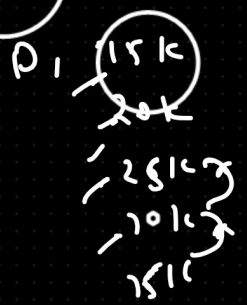
$= 2.8$

Now with the values of V0, V1, V2.... so on we can plot the smoothen curve of time series based plot.

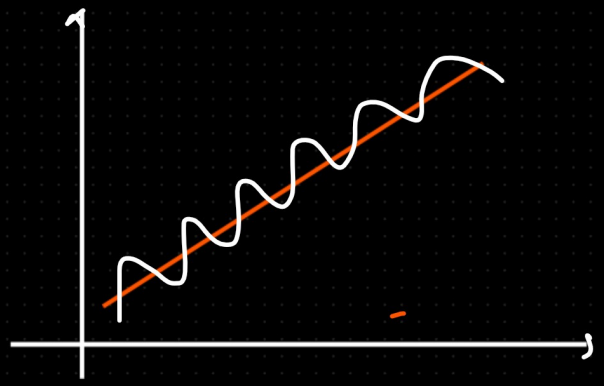




In SMA we are not assigning any weights to the dataset whereas in EMA we are assigning the weighting where more weight is given to the most recent dataset considering the fact that most recent datapoint will be more important for performing forecasting.



Stationary & Non-Stationary



trend = upward

MA (SMA)

Window average

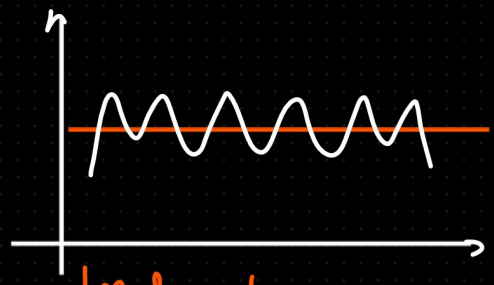
⇒ my moving is not constant over the time.

⇒ variance is not constant over the time.

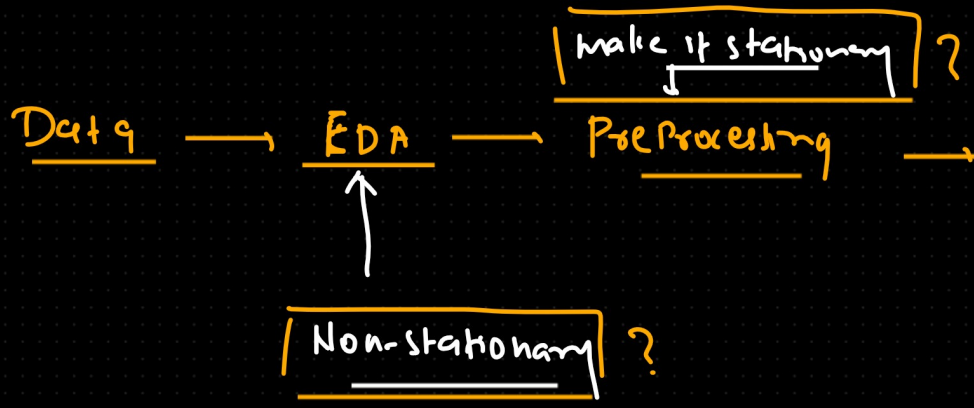
Non-stationary = Mean and var will not be const

Stationary = Mean and var will be constant

{the meaning is there is no change in Mean, var over the time axis}



trend is horizontal



While performing EDA if it is found that the time series bases dataset is Non Stationary then during the Pre Processing stage we need to convert it into Stationary time series based dataset.

This is because Stationary dataset will give better results over Non Stationary dataset.

Analysis =>
1. Visualization
2. & Stats based test

Ways to identify whether time series is Stationary/Non stationary
- Using Visualization techniques
- Using Stats based test

Preprocess => (Non-st to st) =>
1. Differencing
2. log
3. root
4. Adjustment of seasonal Data

These are the different approaches that can be used for converting the Non-Stationary time series based dataset into Stationary time series based dataset.

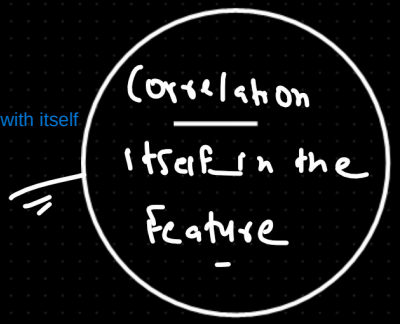
ACF and PACF and ARIMA

ACF → auto corr (correlation) fn (function)

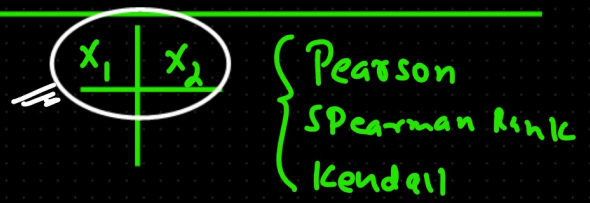
PACF → Partial auto corr fn

ACF ⇒ Auto + Correlation

Auto means correlation of time dependent feature with itself



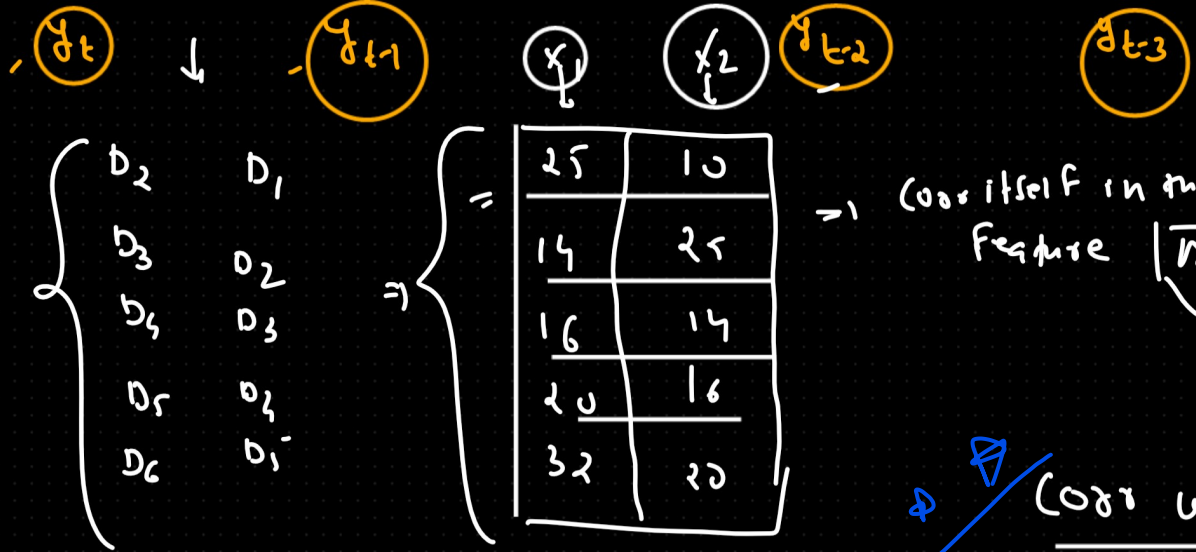
It is a relationship B/w two feature



ACF Measure the Corr B/w time Series and it's lag value

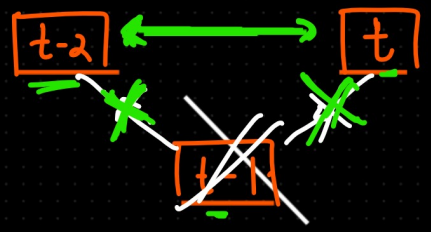
		1st lag	2nd lag	3rd lag
D_1	10	NA	NA	NA
D_2	25	D_1 → 10	NA	NA
D_3	14	D_2 → 25	D_1	NA
D_4	16	D_3 → 14	D_2	D_1
D_5	29	D_4 → 16	D_3	D_2
D_6	32	D_5 → 20	D_4	D_3

In 1st lag we are basically comparing or measuring the correlation b/w current and previous time series dataset. Since, we are comparing with the same column but in current to previous manner so that's why it is called auto corr fun^.


y_t
y_{t-1}
x_1
x_2
y_{t-2}
y_{t-3}
representing 1st lag, 2nd lag, 3rd lag.....
so on with Yt, Yt-1, Yt-2..... so on.
D_2 D_1
D_3 D_2
D_4 D_3
D_5 D_4
D_6 D_5
25 10
14 25
16 14
20 16
32 20
=> Corr itself in the Feature
Autocorr with lag1
Corr with 1st lag
$\rightarrow$ Corr (y_t, y_{t-1})
Corr with 2nd lag
Corr (y_t, y_{t-2})
Corr of 3rd lag
$=$ Corr (y_t, y_{t-3})

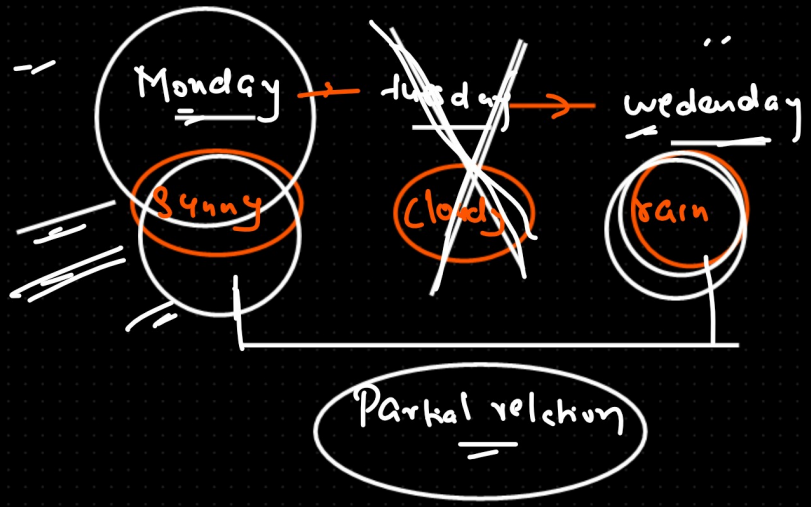
PACF => Partial auto corr

In auto we are first finding corr b/w t and t-1 and then corr b/w t-1 and t-2. By using partial we can directly find the corr b/w t and t-2.



AR

Mathematical



Let's understand with this example. If we attempt to predict forecast of Wednesday based on Monday by skipping Tuesday then this will lie under PACF.

Difference b/w auto correlation and correlation:

Correlation is correlation b/w 2 or more different features whereas, auto correlation is the correlation of a feature with itself.

Auto regression

AR

Auto — {regression itself in the var}

Regression

X → independ- feature
Y → dependent feat-

$y = mx + c$

m = slope
c = intercept

$$y = m_1x_1 + m_2x_2 + m_3x_3 \dots m_nx_n + c$$

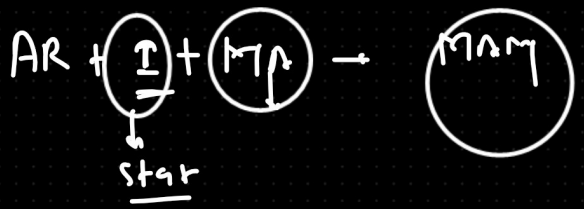
ϕ {
- y_t = value at the current time stamp
- ψ = coff term
- $\underline{c}$ = constant
- $\underline{\varepsilon}$ = error
}

Auto regression

AR eq. with lag 1 $\quad y_t = \psi y_{t-1} + c$

AR eq. with lag 2 $\quad y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + c$

AR eq. with lag n $\quad y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + \dots + \psi_n y_{t-n} + c$


AR
I
Star