23rd June @ 3:00

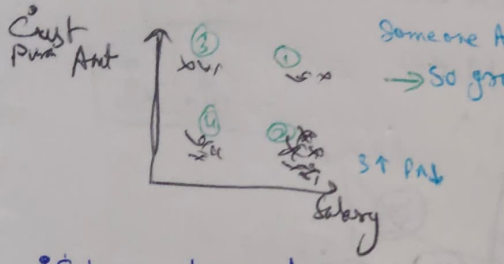
UnSupervised learning

- Target variable is not given (y)
- USL → groups / patterns / segment in Data

Ex: you work as Zara Store as Data Scientist

Customer Purchase Amt	Customer Salary in Lakh
2500	10
⋮	⋮

→ How increase the ~~Sale~~ Sales
What you do.
↑
You increase Sales then
What you do.



- Store does not matter what customer come often or not!

→ It depends you and ~~you~~ management to interpret the group & brings business.

→ In higher dimension you cannot identify / groups Pattern normally → So you need USL

→ Motivation

3:20
- Need to launch a Campaign based on income / sales
- Customer Segmentation

→ Different way of money laundering / fraud
→ Group of images / patterns of specific img.
→ Cohort / batch analysis.

Algo

1. K means clustering
2. Hierarchical
3. DBSCAN

→ Silhoutte Score

June-29 → K-Means, DBSCAN, Silhoulte Score

① USL → Kmeans
→ Hierarchical
→ DBSCAN

② Anomaly detection → Isolation forest
→ DBSCAN
→ Local outlier F

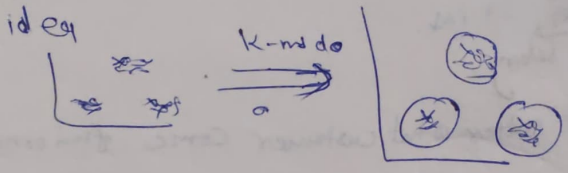
③ Time Series → Smoothing models
→ Autoregressive models

① K-means clustering

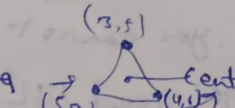
K — means (mean / Average)

ex:-

f_1	f_2	f_3
5	3	2
2	4	3
3	1	5
10/3	8/3	10/3



what Avg
Avg → Arithmetic Center of data



$x_{cen} = \frac{5+3+4}{3} = \frac{12}{3}$

$y_{cent} = \frac{2+5+6}{3} \Rightarrow \frac{13}{3}$

ex:-

How to decide the Klusters

→ Steps - ① Initialize centroid (k)
- can be a random dp
- Centroid of all dps
- A random new dp's

→ ② Points nearer to the centroid willbe labelled as the group

→ ③ Recalculate the centroid.

again sift centroid

again ↳ Step2 → will be repeated until Centroid doesn't change.

(any iteration Avg will be centroid)

→ How to Calculate distance

① Euclidean ⇒ $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

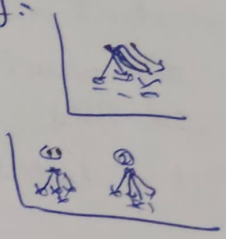
② Manhattan Distance ⇒ $|x_2-x_1| + |y_2-y_1|$

→ How to decide K

↳ use WCSS (within Clusters Sum of Square distance)

$$WCSS = \sum_{i=1}^{n} (\text{distance b/w Points to nearest Centroid})$$

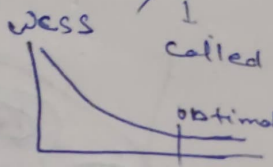
eg:

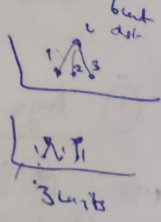


→ As you increases k, wcss decreases.

→ At optimal k, wcss will not change

called Elbow methods.



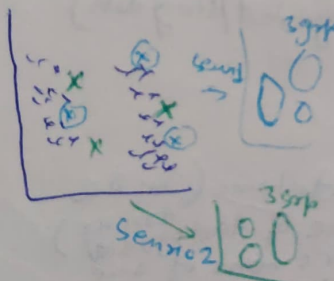


Stiff class label not changing → Centroid not Change.

↳ if all the dp are classified to the nearest centroid class.

- How K is decided → ① Elbow methods
  ② Business team also

→ Random initialization K-trap!



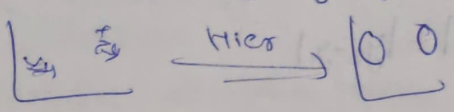
→ what you see

* No of Clusters depends on the initialization centroid.

② Hierarchical clustering

Hierarchical Cl says you give me data I will give you Cluster

1:29



→ but in HC will not any K

- So HC says buddy I will give you complete picture then you decide K ok

types
① Agglomerative clustring (combining)
② divisive clustring

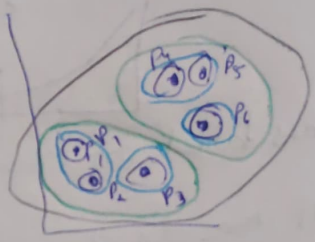
① Agglomerative (Combining)

1:39 ⓐ Steps:

ⓐ Each point is a cluster in its own

ⓑ find the nearest point and create a new cluster

ⓒ Keep on doing step 2 until we get a single group



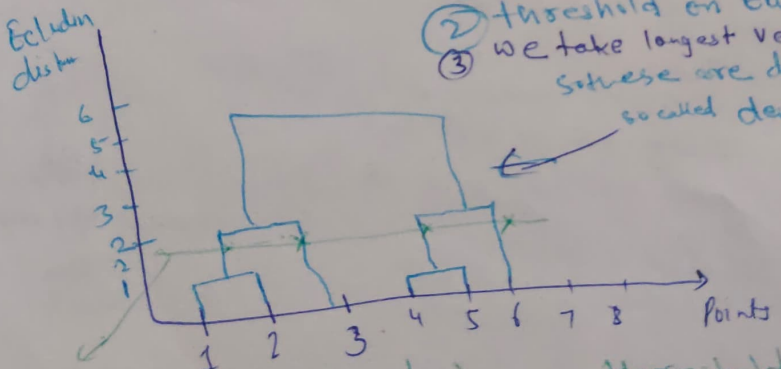
→ So How you are combining?

use ① Euclidean distance
② Manhatten distance
③ Cosine Similarity → To Calculate distance b/w two vectors (Categorical / String data)

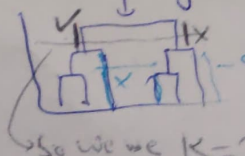
↳ $\cos\theta = \frac{a \cdot b}{\|a\| \|b\|}$

→ Can we Visualize
→ Here How do you decide no. of K
① by business team
② threshold on Euclidean distance (can be tricky)
③ we take longest vertical line of dendrogram

Solves are doubt where none of the horizontal line is passing.



so called dendrogram

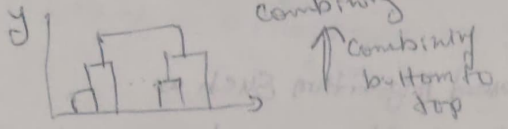


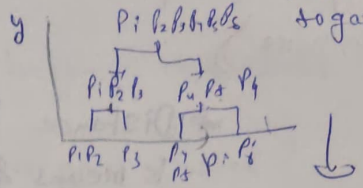
→ So we we K=2

take this Eclidean distance threshold → so 4 grop

• but second Man take another Euclidean distance

→ Agglomerative





it say same thing Put all data points togathers.

→ Here keep deviding untill every individual dp's becomes cluser in its own

→ In Industry Most time use Agglomerative
because → In divisive get all data in 1 go it will be memory consume.
↳ but in Aggl → combine based

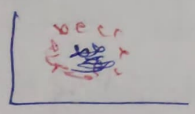
→ kmeans Vs Hrreracial

2:07

#DBSCAN (Density Based spatial clustering of application with noise)

- Why come

eg :-



we need

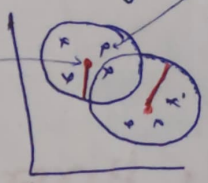
* Distance based Algorithm such as K means & hierarical can not be used here with Non linear Cluster)

So what we do

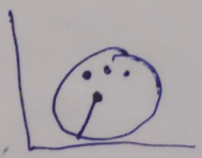
- → Core datapoint
- → border datapoint
- → outlier

* min no of dp - 4 } → hyperparameter
* radius (ε) - 1

→ Core → the dp have atleast minimum no of dp (4) in its radius (ε = 1unit)



② Border → No of dp within its boundary its boundary (ε) will be less than minimum dp.



So how many dp (3) but required dp is 4 So this is Border dp.

③ Outlier → No dp in radius (ε)

→

→ Why DBSCAN? use

- find Patterns in Non-linear data
- also find outlier

→ DBSCAN wiki Pedia

→ Shilhottaue Score / Gefficient :-

(-1 to 1) → ~~1 +1~~

More near to 1 better the cluster would be.

WCSS
↳ within cluster distance

OCSS
↳ outside of the distance of b/w clusters

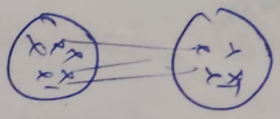
(Q) When we say a given cluster is a good cluster ?

↳ Point 1 :- Within cluster dp's are compact (tightly packed)

↳ So WCSS → should be minimum ~~intra cluster~~

→ Intra cluster distance should be minimum

Point 2 :- Outside cluster Sum of Square should be as maximum as possible (OCSS) from nearest clusters.
(Inter cluster distance)



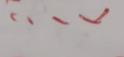
* Separate cluster / good cluster.

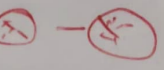
WCSS → (a(i)) → min^m

OCSS → (b(i)) → max^m

Si :

-1 → 

0 → 

1 → 

$$\text{Silhouette Score } (S_i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

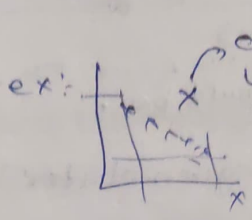
i is any dp

Lect 30 June
* Anomaly Detection & Time Series

Anomaly detection (USL)

||

Abnormality

ex:- 

eg:- ~~fraud~~ transcation
(2) gmail spam or not
(3) fraud IP

⇒ Three methods
(1) Isolation forest Anomaly detection
(2) DBSCAN
(3) Local outlier factor Anomaly detection

⇒ USECASE
↳ Specifically Solving an. outlier/Anomaly Problems

↳ DBSCAN → eg:- find only the transcation which are abnormal/outliers.

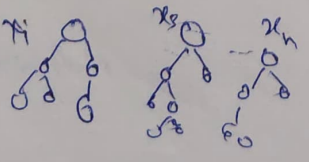
(1) Isolation forest (USL)

→ multiple Isolation trees
→ Isolation is like a decision tree.

$x_1 | x_2 | x_3 | x_4$

Construction of Isolation tree :-

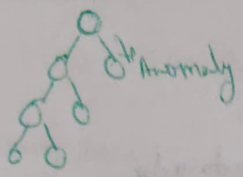
↳ Select a feature randomly x.
↳ Randomly choose a split value with range of x_q
↳ Repeat the process recursively to build the tree.

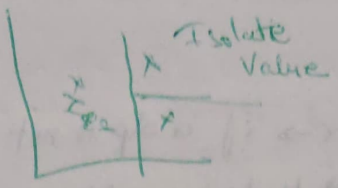


(each of these isolation tree will be built to leaf node/pure node)

⇒ How you will get Outlier?

Ans Anomaly due to their distinctiveness tend to end up in leaf node at shortest path.





→ Since Outliers are different from normal datapoints during Construction of Isolated trees it will be isolated as a seperated leaf node.

↓
How much confidence ??
↓
use
↓

* **Anomaly Score:** $S(x,m) = 2^{-\frac{E(h(x))}{c(m)}}$

for one dp →

m = total no of datapoints.

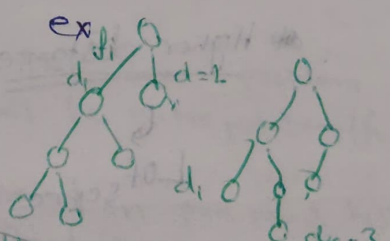
x = datapoint for which you want to check anomaly score.

E(h(x)) – Average ~~score~~ Search depth of x in all isolated tree.

C(m) ⇒ Average depth of all datapoints in all isolation tree.

$E(h(x)) \ll c(m)$

→ Since c(m) is avg depth of all dp's so C(m) be far greater then E(h(x))



$S(x,m) \in 2^{\frac{E(h(x)) \to \text{very small}}{C(m) \to \text{very large}}} \to$ very very small

So

$2^{-\text{very small no}} = \frac{1}{2^{\text{very sm}}} \approx 1$

if $(S(x,m) \geq 0.5 \Rightarrow$ Outlier

↳ threshold

$< 0.5 \to$ Normal dp

$\frac{1}{2^{\text{big}}} = 0.5$

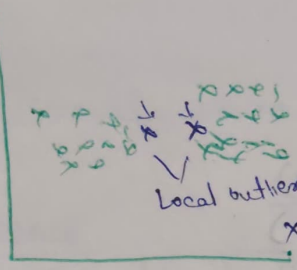
② DBSCAN

↳ Studied in USL

→ Also detects the outlier → if a dp is not core/boundary dp
↳ thats an outlier.

Problem → if the outlier dp are near then

③ Local Outlier factor



Local outlier (near to any cluster)

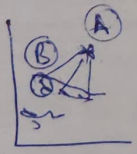
x → global outlier (far from all the cluster)
↳ easy to detects.

How to detect local outlier
↓↓
Local Outlier factor

⇒ ↳ Calculates distance of a dp with its k-nearest neighbour

* dp's which are far will have more distance from its neighbour & vice-versa.

* Higher distance ⇒ density less ⇒ Outlier.



↳

$LOf_{score} \Rightarrow \frac{\text{Interested dp distance}}{\text{All dp's distance of neighbood dp}}$

$Lof \rightarrow 0 \rightarrow \infty$

~1 ~ Normal
> 1 → outlier
< 1 → Normal

Time Series

→ A series with time Component

Time Series

Time Stamp
↳ Month
day
years
minute
hours
seconds

Previously solve
↳ x_1 Area of house | x_2 No of Room | x_3 Price of house

↳ MLR (Multi linear Regression)

Month	Sales
Jan	60K
Feb	40K
Mar	50K
Apr	80K
⋮	⋮

↳ Time component is involved

Jan → 15k, Feb 20k, Mar 60k ≠ Mar 60k, Jan 15k, Feb 20k

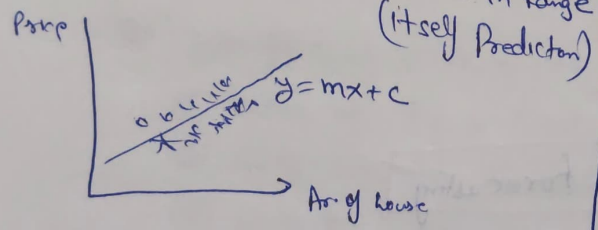
Here order Matters
↳ because the current time period value depends on the previous time period.

→ So

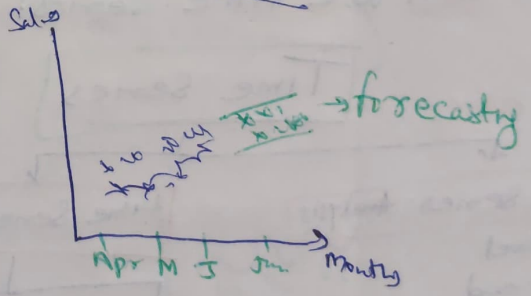
→ So Can we use Regression model for time series data?

→ There is 2 concept

① Interpolation → to find the value in range (it self Prediction)



② Extrapolation



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

Eg:- Training (Area) = 1000 – 10000
testing (Area) = ''

* if outside training range ⇒ Wrong Prediction

* Time Series problem will be extrapolation (forecasting)

* Based on previous history Predict next 6 months.

① Why not Linear Regression for time series?

↳ Time Component is involved.

↳ Because of Extrapolation, Prediction may be wrong.

→ LR → assume linear relationship but it time series the current observation depends on previous observation → which is not true for non-time series data.

Motivation

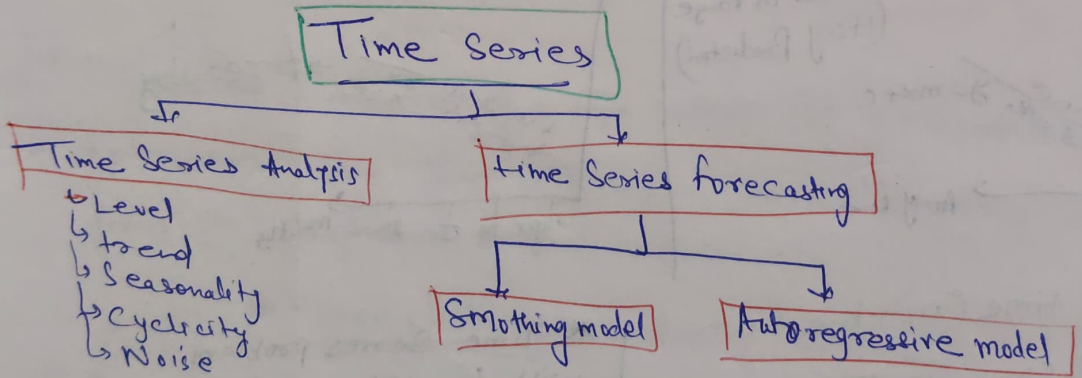
① Weather forecasting → weather pattern. day wise/month/year.

② Medical

③ Stock Mkt

④ Ecommerce/finance → Sales, bond Price.

Whenever time component is there in data, you can use time series!



⇒ Components of time Series

① Level

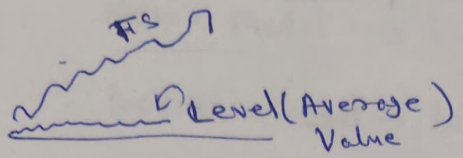
② trend

③ Season/Seasonality

④ cycle/cyclicity

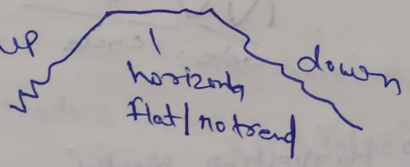
⑤ Noise

① **Level** :- The base value of a time series on which other components are added.



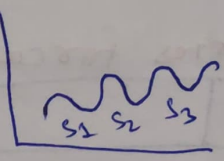
② **trend** : Long term moment or Direction in data over a long period of time.

(a) Upward
(b) downward
(c) Horizontal

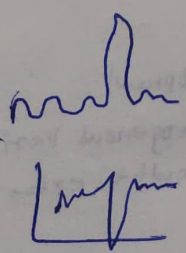


③ **Season** :- Frequent repeation in a data over regular interval (daily, monthly, annually)

ex:- ① traffic in peak hours
② Sales of TV in diwali
③ No of tourist in peak season
④ No of Ice-cream Sales.



④ **Noise / Anomly** :- Some uncertainty / randomness in time series data because of unexpected reason.

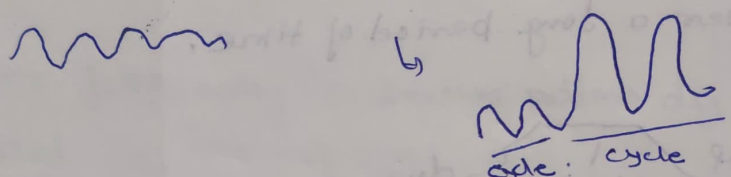


reason → news
- Reports
- Pandemic
- war
- Election
- Influencers.

⑤ Cycle → The fluctuation in the data over a longer period of time. These periods are not fixed and can vary

↳ Cyclicity = Season + Noise

Q ① Pharmacute industries stock price



② tractor making company → Harvesting Months.

Time Series

TS Analy
TS forcesting

Time Series forcasting

Smoothing Mode

① Simple moving Avg (SMA) (window Avg)
② Cumulative Moving Avg (CMA)
③ Exponential Weighted Moving Avg (EWMA)

Autoregressive Model

↳ AR
↳ MA (moving Avg
↳ ARIMA
↳ SARIMA
↳ SARIMAX (Optimal)
↳ Exogenous Variable (outlier, extra effect)

Stationary
↳ over the time, value is not varying (change)

Smoothening

Smoothing Models :- Smoothing of TS → Removing fluctuations

i Naive Model → last observed value is forecast

ii Average Model → Average Value is the prediction

(1) Simple Moving Avg (SMA)

$$\text{Avg} = \frac{\text{Sum of All nos}}{\text{no of values}}$$

eg: 2, 3, 4, 5 → $\frac{2+3+4+5}{4} = \frac{14}{4} = 3.5$

MA :→ moves over the time axis in a specific Moving Avg window (Window size → Avg Value)

eg:

Window 3

Av

Jan	50	SMA
f	65	N1
m	70	N2
A	85	62
Ma	90	74
J	100	82
J	110	80
		90
		95
		100

$\frac{50+65+70}{3} = 62$

2nd win = $\frac{65+70+85}{3} = 74$

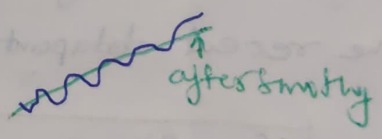
(2) Why Smoothing

→ To check trend of data

→ To Remove all the effect from data

→ reduce the effect of outlier

→ Visualization



(2) Cumulative MA :-

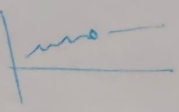
→ find all the Avg. if all the datapoint upto given timestamp.

→ for long period of time

→ Exponetial trend

eg:

	CMA
Jan 10	10
Feb 12	$\frac{10+12}{2}$ →
Mar 15	$\frac{10+12+15}{3}$
Apr 14	$\frac{10+12+15+14}{4}$



SMA, CMA, Naive, Avg → It gives equal priority to all the value.

but ↓

→ In Time Series Current observation is highly influnced by last few observation

∴ you need to give priority to recent observation.

(3) EMA or EWMA (Exponention Weighted MA)

→ we give more weightage / importance / priority to the recent datapoint / time stamp

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

V_t = EMA at time t

β = $0 < \beta < 1$ → generally its 0.9 (Weight)

V_{t-1} = EMA at previous ts.

θ_t → Datapt at current stam.

eg:-

$\beta = 0.9$ (To reduce the previous time stamp effect ↳ $\frac{1}{1-\text{Beta}}$)

V_{t-1}	Month	Value	Error
V_0	D_1	25	0
V_1	D_2	13	1.3
V_2	D_3	17	2.87
V_3	D_4	31	
V_4	D_5	43	

$V_1 = \beta_0 V_0 + (1-\beta)\theta$
$= 0.9 \cdot 0 + (1-0.9) \times 13$
$= 0.1 \times 13 \Rightarrow 1.3$

$V_2 \Rightarrow \beta \times V_1 + (1-\beta)\theta$
$= 0.9 \times 1.3 + 0.1 \times 17$
$= 2.87$

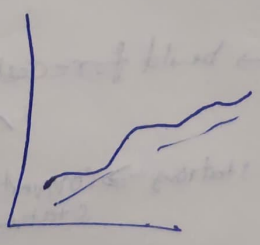
In Time series 2 way to Calculate

Additive

y_t = Trend + Season + Noise

↳ Linear over time
→ Constance Variance
→ Increased trend at same difference

eg Day1 100
Day2 200
Day3 300
Day4 400



Multiplication

y_t = Trend × Seasonality × Noise

→ Non linear
→ Non Constant Variance

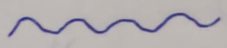
So TS → forecasting →
↙ Smoothing ↘ Autoregression

→ You can build a Smoothening time series model on Any time series data.

* To build Autoregressive model, statistical properties of a time series like mean/ Variance should be constant (Not change over time)

eg: Scenario 1

Scenario 2



→ It will be easy to build a forecasting model in SC-2
→ less variance in data.

* Non Stationary → Mean-Variance will not be constant.

* Stationarity → Over the time, value is not varying (changing)

* So → To build a Autoregressive model make the time series stationary.

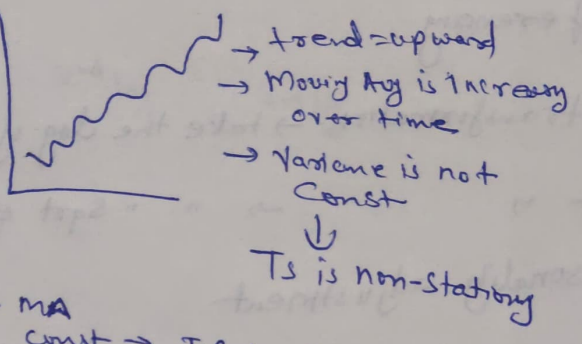
In ML Do → Data Ingestion → Analysis → Preprocessing → Model building

→ Time Ser → ... → TSA → TS is Stationary → build forecasting Model
↳ if non stationary → convert to stationary TS

⇒ To Check if a TS is Stationary or not

b.

1. Visualization



2. Statistical test

a. ADF (Augmented dickey fuller test)

H_0 = TS data is Non Stationary

H_A = TS " " Stationary

pValue ≤ 0.05 → reject H_0

Conclude → Ts data is Statn

b. KPSS test (kwitkowski - philips-Schmidt-Schin test)

H_0: ts data is Stationary

How to convert non-st TS to st TS

1. Differencing
2. log transformation → take the log of Value
3. root " → " " Sqrt of Value
4. Seasonality adjustment
5. Box-Cox transformation (to make variance constat → +ve)
6. Yeo-Johnson ,, (All TS data)

⇒ (1) Differencing → Difference $(y_t - y_{t-1})$
current, previous

eg:-

Month	Price	1st diff	2nd order	3rd
J	5	NA	NA	
F	10	10-5=5	NA	
M	6	6-10=-4	-4-(5)=-1	
A	8	8-6=2	2-(-4)=6	
M	15	7		
J	7	-8		

↳ Check st → 3rd order

↳ After diff check if TS is Stationary

→ Statistical test (ADF)
→ Visualization (using MA)

★ if Stationary → build the forecasting model
else → Again do differencing.

* ACF → Auto Correlation function
* PACF → Partial Autocorrection function
* Autoregression & ACF → Auto + Correlation
  - Auto → Correlation itself in feature
  - Correlation → Relationship b/w 2 variable

* ACF measures the correlation b/w time series & its lag value.

* Month y_t 1st lag 2nd lag 3rd lag