

=> Data Requirements, Data Collection, Data Processing, Data Cleaning, EDA, Modeling & Algo, Data Product, Communication /Visualization.

=> prob def, data prep, data analysis, representation of results.

=> value, velocity, veracity, variety, volume

=> Accuracy, Completeness, Consistency, timeliness, believability, Interpretability.

=> data - cleaning, integration, reduction, transformation (**data prep**)

=> Prob.def > d.coll > Model.Dev > Data.Analysis > R.C (**classical data analysis**)

=> P.D > D.C > D.A > M.D > Result.Communication (**exploratory data analysis**)

=> P.D > D.C > M.D > Prior Distribution > D.A > R.C (**Bayesian data analysis**)

=> **Line Chart:** relation b/w 2 or more continuous variable,(time series).

=> **Scatter plot:** relation b/w ( 1 indep 1 depend ) ( 2 indep(regressor)),(co-relation).

=> **Bar Chart:** categorical variables, ( comparison or ranking ) when large changes.

=> **xticks**= xlabels , **axis .text**(ha=center, va ='bottom')

=> **figure** = overall container, **axis** = actual bar area axis area

=> **pie chart** = part to whole ratio, composition

=> **Area and stacked plot** = area under line , cumulative effect of multiple variable,

=> **Table Chart:**

=> **Polar chart:** polar axis, multivariable(comparison b/w), angle & radius,

=> **Histogram** : distribution , continuous variable

=> **Lollipop Chart:** ranking ( ordered bar chart ) combine scatter or bar chart ,

=> **ax.set\_ylim**(0, 30)-(Y-axis starts from 0 to 30, prevent automatic scaling ).

=>**Drift** Method: linear trend from the first to the last point,Forecast increases based on the average change over time,Good when the data has a trend (upward or downward). (**SMA**) Averages recent values to smooth fluctuations,Does not assume a trend and instead it flattens spikes and dips,Useful when the data is erratic or you're unsure about a clear trend. Rolling (3) change over size of data.

=> **Additive Seasonal Model** : Forecast=Trend + Seasonality , when seasonal changes are constant in size (e.g., +10 sales every December), **Multiplicative** Forecast=Trend × Seasonality,when seasonal changes grow with the level (e.g., sales double every December).

=> **Holt Winters Smoothing** Forecasting with both Trend and Seasonality Seasonal: Level (overall average) Trend, Seasonality , Triple Exponential Smoothing,  $F = L + T + S$  &  $F = (L + T) \times S$ , smooth out all 3 components,

=> **MAE**: want average in same unit.**MSE**: big error are costly. **MAPE**: compare forecast across dataset.

=>  $WMA = (w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_n \cdot x_n) / (w_1 + w_2 + \dots + w_n)$ , In SMA lags more. WMA: responds faster to trends. near to trend line, react quickly to change in trend.

=> Choosing Time Period: 3,7 , larger m smooth forecast but more lag, small m more responsive but maybe noisy , 3 respond quickly to change in original data, ideal for short term insight.

=> Use high: track longterm trends,smooth out volatility ,seasonal & high noise

=> (3 Medium ,Very jagged, Good for alerts, when changes frequently, need quick signals)(7 High ,Smoother, Better for weekly trends)( 14 Lower, Very smooth ,May miss turning point)

=> **Merging** -> **concat**(no grouping, placing 1 on another), ignore\_index =true(creates new index), axis=0 (vertically, column wise), duplication occurs (remove/by other methods).

=> **Merge joins** -> **inner**(intersection),**outer**(union), **left**(common one and left one), **right** ( common one and right one ).

=> **Reshaping & Pivotong:** Transpose (R-><-Columns), Stacking (columns->rows, (.stack, taller), Unstacking (r->c, stacked.unstacked,wider, will create missing value when group are uneven/values aren't present in sub groups.

=> **Pivoting:** to group unique values,change the data representation to analyze hidden trend—**pivot table:** with aggregation of numeric data , create separate spread sheet, agr value nahi doga to tamam numeric value ko consider kry ga jis sa additionl level of hirerachy ho ge, multindex on rows, mean, sum, count, max, min, std, list – agg func, index, columns, values, agg fun—when:sorting not necessary, same func apply everywhere, want reshaping, ignore nan values, want summary table, when multiindexing(rows&col), less computataional,buil in agg F.

=> **Group by:** [ split,apply,combine], func manually, -when : sorting needed,no reshaping,less ideal for reshaping, need flexibility, need to apply multiple func, group rows, custom func, more programmatic, often create multiindex, keep nan values by default, need extra handling, group by multiple columns, group by multiple aggregation(column wise)[ columns:func]

**Time Series:** sequence of datapoints at regular time, order & time matters.

=>**Trend:**long term increase or decesss,**Forecasting, seasonality:**repeating patterns over time, **Cylic: not fixed duration** (long term fluctuation), **Noise:** random variation or error, **Abrupt change:** sharp increase or decrease, **Stationarity:** no trend/ no seasonality/ mean & variance do not change,

=> **Operations: Smoothing**( reduce noise, moving average), **Differencing**( remove trend & seasonality), **Decomposition**(split in to trend , seasonality, residual.--- ( line (trend), lag(auto correction), seasonal(for repeat pattern) plots).

=> **Models:** naïve , moving avg, exponential smoothing, ARIMA, LSTM, prophet.

=> **Univariate**(one var measu over time- monthly rainfall), **Multivariate**( two or more varuable measr, Depended ek dsry pa, then relation analyze krta hain,data should be synchronized, multiple time series,

=> **Trend: Local**(in specific period), **Global** ( over entire frame of dataset, stock).

=> **Seasonality:** data repeat pattern over time ( eid shopping )

=>**Residual**(capture noise or random fluctuation that model can't explained) , res= actual-fitted values, remainig values after trend.

=> **Outliers:** data sa hat ka yah bht uper yah bht necha ahije points, cause by error or event that are not predictable. **Additive**( one time, happen once only and effect few data points), **Seasonal additive**(repeating pattern of unusual values, happens at regular interval- icecream sales).

=> **Naïve Forecasting:** next is last observed values, use for evaluation as well, Simple and Easy, Last Value as Forecast, No Training Required, Benchmark for Comparison

=> **Simple Naive** (Random Walk): pechli he agli ho jaye ge.

=> **Seasonal Naive:** February 2025 forecast = February 2024 actual value.

=> **Drift Method:** draw line first and last obser and extend it,suitable when data show steady trends. ( $F_{t+h} = Y_t + Y_t - Y_{t-1} \times h$ ),  $Y_t$ : most recent actual value  $Y_1$ : first actual value ,t: number of time periods h: number of periods ahead to forecast.

=> **Moving Average:** smooth out the fluctuation in time series data.identify trends and make short term trend.

=> **Simple Moving Average** (SMA): equal weights to each value, even out short terms ups and down and revel long term pattern, designed to reduce noise and show general direction of the series not to follow trend.

(  $sma = Y_{t-1} + Y_{t-2} + Y_{t-3} + \dots + Y_{t-n} / n$  ) [  $Y_t$  : forecast for time t n : window size (e.g., 3-period moving average)]

=>**Weighted Moving Average:** More weight to recent values

=>**Exponential Moving Average:** Exponentially decreasing weights