Stationarity.

A time Sevies is stationavy when its statistical behavior doesn't change over time. Concretely that usually means. the mean, variance, and autoco-variance stay same across time. If this properties drift, change volatility a have seasonal patterns, the Sevies is non-stationavy Stationavy matters because many models assume that the relationship they learn is stable if the underlying distribution shifts. models can fail (or) be biased.

types of Stationarity:

⇒ Strict : the joint distribution of any set of time points is invariant under time shifts.

⇒ Weak Stationarity: Mean is constant, variance is finite and constant, and co-variance depend only on lag not absolute time.

⇒ Raw Stock prices are typically non-Stationavy

Estimate trend (T_t) $\boxed{T_t = \frac{1}{m} \sum_{i=-K}^{K} y_{t+i}}$

where m is the window size often equal to

Estimate Seasonality (S_t)

* Remove the trend to isolate the Seasonal Pattern

Additive $Y_t - T_t$. Multiplicative Y_t / T_t

Group by Season (month, quater) and average to estimate Seasonal

Estimate Residuals (R_t):

$R_t = Y_t - T_t - S_t$ $\quad R_t = \frac{Y_t}{T_t \times S_t}$

Descriptive: Ex EDA

Predictive: what likely to happen

Prescribtive: what we need to do albive the Predictive.

Forecasting: It about time
And Predictive may (or) may not abut time.

Data types in time Servics:

1. Cross Sectional data: Data is collected at Single Point in time on one (or) more variables. Here data not sequential and usualy, data Points are independent of one another. Regression, random forest, neural network methods have been applied widely on these data.

2. Time Series data: onivariate (or) multivariate data is observed across time in a sequential manner at pre determined and equall-spaced time interval Ex yearly montly .... ordering among data points is important and cannot be destroyed.

3. combination of cross sectional & time Series data:
this is Complex data where information on Same variables are collected our various points of time.

Time Series:

A collection of observations that has been observed at regular time intervals for a certain variables over given duration: is called time series.

*All observation are dependent: In time series data, each observation is expected to depend on the past data observation.

*Missing data must be imputed: Because all the data points are sequential in time series. If any data point is missing it must be imputed before the actual analysis process commences, otherwise proper ordering is not preserved

*two different types of intervals cannot be mixed.

Time series data is observed on the same variables over a given period of time with fixed and regular time intervals. Though data can be collected at various intervals such as yearly, weekly, hourly etc any specific time interval, The time-interval must remain same throughout the entire range.

*objective.

Time series forcasting is applied to extract information from historical series and is used to predict feature behaviours of the same series based on past pattern.

*Approaches used for time series:

1. Decomposition= this method is based on extraction of individual compones of time series.
2. Regression: this method is based on regression on past observations:

Decomposition: what part of the past projects to the future and what part does not.

The part that can be forecast which is systamtic and not an error. thar part which is not systematic (or) random error we are going to ignore.

$$y = a + bx + (\varepsilon_0) \rightarrow \text{residual}$$

In forecast we ignore the residual Because the residual is applied to current data set not the future data set.

① Part that is predictible
② Part that is not predictible. (Ignore)

* what we are going to project (or) Predict the future there is nothing to do with the past data.

* If we are tryng to get any information about the any situation so we need gtt information from similar situations already happend

* If nothing like this Exist Then ??
Here we make hypothesis
we guess this is what is usslay happen, if not this happen we keep the safty margins aroundbit to ~~process~~ ~~posted~~ which is confidence intervals suppose to do.

what is happing long term period and shat term Period.

$y = f(x)$

$y_{t+1} = f(y_t, y_{t-1}, y_{t-2})$ → test Situation

future as a function of the past.

$y_t = f(y_{t-1}, y_{t-2}, y_{t-3} \ldots)$ → train Situation.

Supervised algo → time Series forecast algo

Past observations as inputs to the supervised algo. (Neural Network).

we create the Regression on the present and Past to forcast the future

y_t = trend + Seasonality + Residual

y_t = trend x Seasonality x Residual

trend → t, Long term upward/downward movement

Ex: Rising Stock prices over years.

Seasonality S_t → Regular pattern repeating over time

ex: monthly energy usage

cyclical c_t → fluctuations due to business/economic cycles

Boom - bust cycles.

Residual/Irregular → Random, unexplained variation

Noise, Errors, Sudden shocks.

a) additive model:

Used when the magnitude of fluctuations is constant over time.

Ex: temperature data, Sales without exponential growth

$$Y_t = T_t + S_t + R_t$$

b) Multiplicative Model.

Used when seasonal fluctuations increase (or) decrease proportionally with the level of the series.

$$Y_t = T_t \times S_t \times R_t$$

Ex: Product sales where peaks growth

Trend: when the series increases (or) decreases over the entire length of time. the price of a share may increase (or) decrease linearly over a period of time.

Seasonality: when a series is observed with more frequenty than a year, the series are subjected to rythmic fluctuations which are stable and repetable in each year.

Normal dataset It's called the EDA

& the time series data its called as the data decomposhay.

where it gives the full picture of the trend & seasonality and noise.

Methods:- STL, X-11/X-13 ARIMA, Wavelet, EMD, VMD, Fouier

Stationary:

⇒ No Long term trend & no seasonality

⇒ constant mean and constant variance

How to transform the data to - stationary.

1. Logarithm transform (impacts variance)
2. Differencing (impacts mean)

ADF test: (Augmented Dickey-fuller) test

Determines if a time series is stationary or not.

null hypothesis H_0 = non stationary
Alternative hypothesis H_a = Stationary.
Significance level $\alpha = 0.05$

If P value $< \alpha = 0.05$ suggests rejecting the null hypothesis, meaning the series is stationary.

KPSS Test: (Kwiatkowski-Phillips-Schmidt-Shin) test
Determines if time series is Stationary.

null hypothesis H_0 = Stationary, H_1 = non Stationary.
$\alpha = 0.05$

P-value < 0.05 suggests rejecting the null hypothesis meaning the series is non-stationary.

ACF: Autocorrelation function

Displays the correlation between the time series and its own past values (a) lags.

X-axis represents lag (time difference)
Y-axis represents correlation coefficient (-1 to 1)

How to interpret:

Gradual Decay : indicates non-stationary (trend is present in the data).

Sharp cutoff : indicates stationarity and potential use of an autoregressive (AR) model.

Repeated peaks: Suggests seasonality in the data
Eg: correlation repeats every month

PACF: Partial auto correlation function:

Measures the correlation between the series and its lag while controlling for the influence of intermediate lags.

A slowly decaying PACF plot may indicate the presence of the seasonality in the data.

If the PACF does not show a clear cut-off it might suggest that an AR model may not be a good fit or data may benefit from additional preprocessing or transformation.

Covariance always measures the relationship between two variables.

for same variable: Then we need to do auto-covariance which measures how the same variable relates to itself at different time lags.

data Points: [100, 102, 101, 103, 105, 107, 106, 108, 110, 111]

Autocovariance of lagk for points $X_1, X_2, X_3 \ldots\ldots X_m$

$$r_k = \frac{1}{m-k} \sum_{t=k+1}^{m} (X_t - \bar{x})(X_{t-k} - \bar{x})$$

m = total no of data points

$\bar{x}$ = mean of the series.

X_t = current value

X_{t-k} = value k steps before.

Lag1 $r_1 = 1/9 \sum_{t=2}^{10} (Xt - 105.3)(Xt-1 - 105.3)$

Lag1 = 9.36 Lag2 = 6.35 Lag3 2.20

Lag1 is high because the each day's price is strongly related to the previous day's.

Further days (2,3...) became weaker because the connection fades older days influence the present is less.

we can capture the direct and indirect co-variance of the previous days.

ACF

$$\rho_k = \frac{cov(x_t, x_{t-k})}{Var(k)}$$

Here Acf at lag k tells us how much does today's value x_t move together with the value k days ago x_{t-k}

It tells the overall linear relationship it includes both

Direct effect of x_{t-k} on x_t
indirect effect of all the values between.

x_t is strongly related to x_{t-1}

x_{t-1} is strongly related to x_{t-2}

Then even if x_t never talks directly to x_{t-2} they will still show correlation because of the indirect path $x_t \leftarrow x_{t-1} \leftarrow x_{t-2}$

PACF: Removing the indirect paths.

PACF measures the direct correlation between x_t and x_{t-k} after removing the effects of all intermediate lags

Mathamatically it's the correlation between the residuals of two regressions:

Regress x_t on $x_{t-1}, x_{t-2}, \cdots x_{t-(k-1)}$

Regress x_{t-k} on the same set of intermediate terms.