INTRODUCTION TO TIME SERIES FORECASTING AND ANALYSIS

Thursday, May 11, 2023 12:17 PM

Q what kinds of Data have we seen?

Ans Panel Data (Pooled Data) | Cross Sectional Data | Time Series Data

Also called longitudinal data.
→ cross-sectional + time series data.
Multiple entities and multiple time intervals

Panel

City	Date	Temperature	Humidity	Wind
NYC	01-01-2015	50		
NYC	01-01-2014	30		
NYC	01-01-2013	40		
SFO	01-01-2015	70		
SFO	01-01-2014	75		
SFO	01-01-2013	65		
BOSTON	01-01-2015	30		
BOSTON	01-01-2014	20		
BOSTON	01-01-2013	15		

Cross Sectional

City	Date	Temperature	Humidity	Wind
NYC	01-01-2015	50		
SFO	01-01-2015	70		
Boston	01-01-2015	30		
Chicago	01-01-2015	30		

Cross Sectional Data is a collection of observations for multiple entities at single point of time

It is basically a snapshot at a given time ⇒ Forecasting / Prediction can't be done on it.

Time Series

City	Date	Temperature	Humidity	Wind
NYC	01-01-2015	50		
NYC	01-01-2014	30		
NYC	01-01-2013	40		

Time series is a collection of observations for an entity at different time intervals
{Equally Spaced}

→ When we talk about intervals they are supposed to be more or less even.
→ The order of the data has to be chronologically.

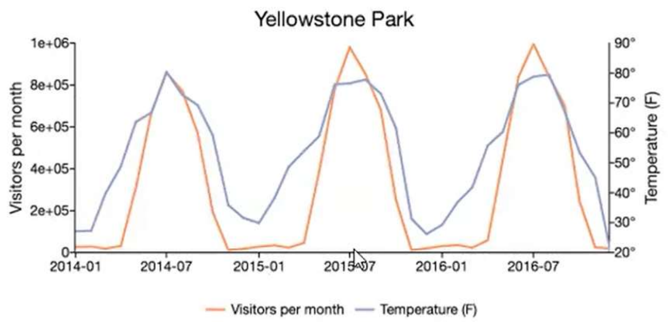
Importance of Time Series Analysis

Real time applications:

Most businesses work on time-series data to analyze or forecast:

1.) Sales Numbers
2.) Website Traffic
3.) Position w.r.t. Competition Products
4.) Demand & Supply of Product
5.) Census Analysis
6.) Budget Analysis
7.) Stock Market Analysis
etc.

* Stock Market data is one of the most volatile data and if you are forecasting on it, then always bear in mind that any real world event can easily thwart your forecast.
eg Dogecoin price due to Elon Musk



The above graph → It is a plot of visitors per month to avg. monthly temperatures.

Terminology:

1.) Time Series Frequency: The time interval at which data collection is generally reffered to as the time series frequency.

→ The dates range from Jan 2014 to Dec 2016 and is collected at a monthly frequency.

Challenges in Time Series (w.r.t to a Regression Problem)

1.) Time dependence of time series: In linear regression the basic assumption that the observations are independent doesnot hold in this case.

2.) Seasonality of time series: Along with an increasing or decreasing trend, most time series have some sort of seasonal trends'. i.e. variations specific to a time frame.

Univariate V/S MultiVariate Time Series

→ Univariate Time Series models are used when the dependent variable is a single time-series.
Eg: Modelling an individuals heart rate using only past observations as exogeneous variables.

→ Multi Variate Time Series models are used when there are multiple dependent variables.
It means in addition to depending on their own past values, each series may depend on past & present value of other series as well.
Eg modelling U.S. GDP, inflation and unemployment together as endogeneous variable.

Regression V/S Time Series Forecasting

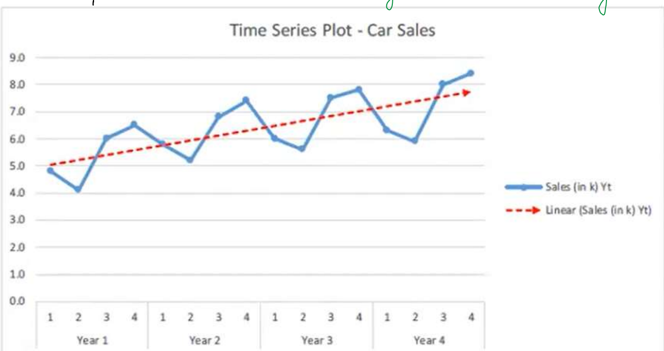
→ Time-Series Forecasting is Extrapolation {Extrapolation → Just like you extrapolate on a Jigsaw Puzzle}

→ Regression is Intrapolation
↳ {Finding values based on what the model learnt}

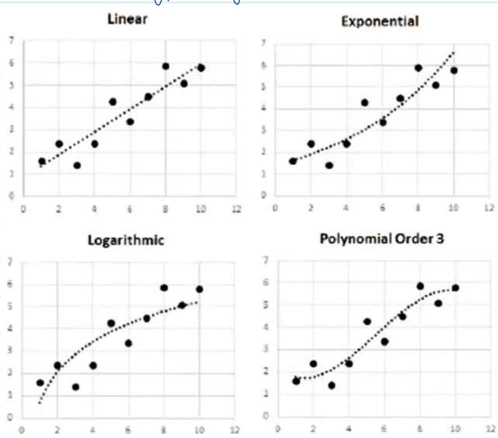
Time-series refers to an ordered series of data. Time-series models usually forecast what comes next in the series - much like our childhood puzzles where we extrapolate and fill patterns.
Time-series may or may not be accompanied with other companion series which usually can be seen as occurring together. Sometimes, the prediction is also applied for these companion series... Such problems are referred to as 'Multivariate Timeseries'
Apart from all these, Time-series could also be accompanied by Exogeneous variables which are very much like companion series.... but they are not predicted because it is something exogeneous to the System. Their future values will be specified when we are making the prediction for the Target series. For e.g. while doing sales forecast, the variable whether we will be running promotions on TV at that time is an Exogeneous variable. The predictions can be made by specifying different values for them.
Regression can be applied to Time-series problems as well. e.g. Auto-regression
But Regression can also be applied to non-ordered series where a target variable is dependent on values taken by other variables. These other variables are called as Features. When making a prediction, new values of Features are provided and Regression provides an answer for the Target variable. Essentially, Regression is a kind of intrapolation technique.

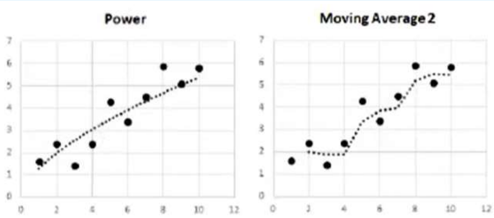
Components of Time Series Data

1.) Trend Component: Trend is a general direction of data series.

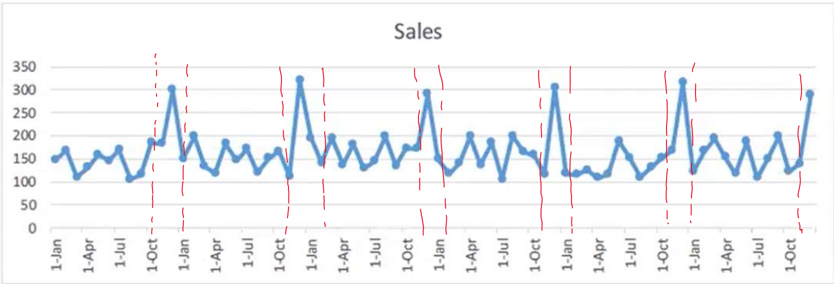


Types of Trends ↓





Imp 2.) Seasonality Component : Pattern that repeats at regular time interval.



→ Seasonality comes into existence when a series is influenced by seasonal factors
Eg. the quarter of the year, the month or day of the week.

→ Can also be due to holidays, festivals or an extravaganza event.

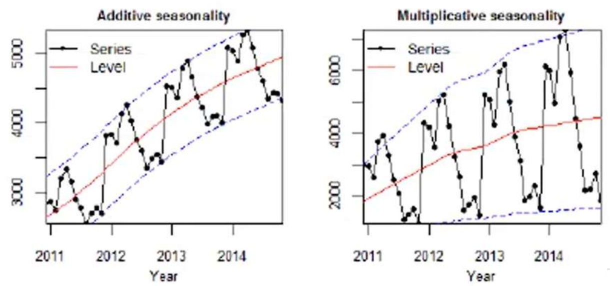
→ Seasonal variation is also known as periodic.

→ Seasonality can be additive or multiplicative.

→ Additive seasonality is when the values in different seasons vary by a constant amount.

→ Multiplicative seasonality is when the values in different seasons vary by a constant degree.

* Multiplicative seasonality increases as the level increases.



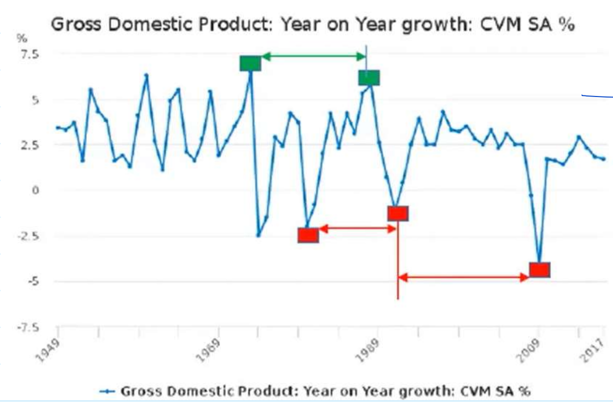
→ Seasonality is often expressed as
↳ low (< 20%)
↳ intermediate (20 – 50%)
↳ High (>50%)
in terms of range.

where the range is defined as

$$range = \left(\frac{max - min}{min}\right)\%$$

3.) Cyclical Component : a cyclic pattern exists when data exhibit rises and falls that are not of fixed period.

This graph shows rises & falls but not at fixed intervals.



→ Medium term variation caused by circumstances repeating at irregular interval.

→ Eg 5 years of economic growth followed by 2 years of fall and again 7 years of growth.

→ Cyclacity may or may not be present in the data.

→ Avg. length of a cycle is usually longer than that of seasonality and the avg. magnitude of cycle is more variable than that of seasonality.

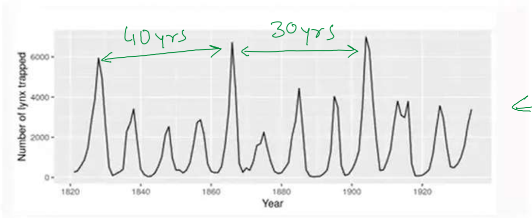
Cyclicity V/S Seasonality

→ A cyclical component means the pattern is repeated at irregular intervals and the period when it reoccurs is over a year and the outcome may change from one cycle to another.

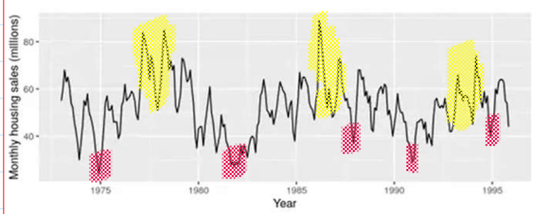
→ A seasonal component is in which a certain pattern is repeated after a regular period of time and the recurrence is usually less than a year.

Conclusion: → Seasonal Component is more predictable than cyclic one.
→ Seasonal Component is constant more or less compared to the cyclic.

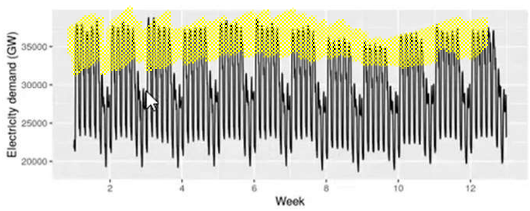
eg of seasonal component → Diwali Sale on Amazon.
eg of cyclic component → Savings of Person before and after marriage.
↓ monthly



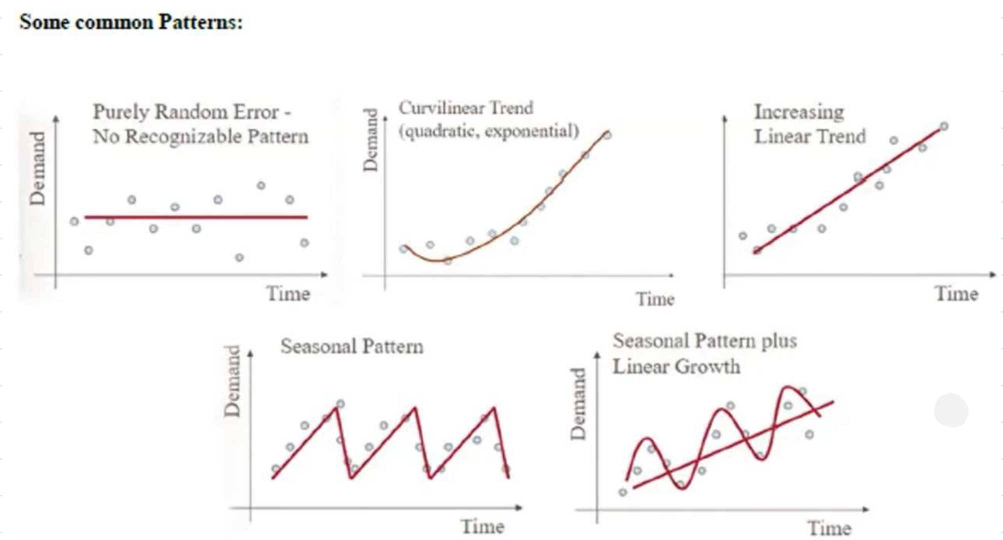
← Cyclicity Eg.



← Seasonality
← Cyclicity



← Here we can see seasonality pattern → daily & weekly both



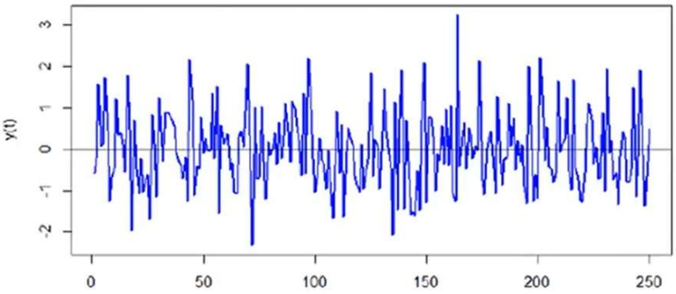
4.) **Irregular Component:** It is the residual time-series after the trend cycle and the seasonal components (including the calendar effects) have been removed.

↓

The variation in time-series caused by the changing number of particular week days. Eg. working days

5.) **White Noise:**
→ It has mean as Zero and has no correlations.
→ It doesn't help in predictions
→ Often used in order to reduce other noises in the data.

{ White noise may not be counted as a component of time-series but is a very important concept }



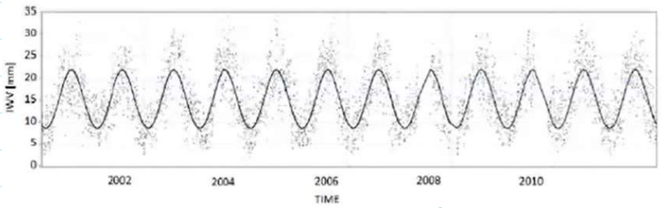
☆ Eg. of white noise ↓

When a time series doesnot follow a pattern or seasonality. Just a random spike or random data on which you can't do prediction.

☆ **Stationarity**

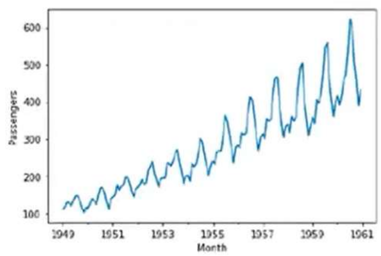
→ Time series requires the data to be stationary and most of the time-series models work by assuming the data is stationary.

→ Stationarity means:
↳ Constant mean according to time
↳ Constant variance [at different time intervals]
↳ Covariance doesn't depend on time. It should be constant over time.



Stationary series [mean is constant]

Non-Stationary Series:



Increasing trend or non-constant **mean**

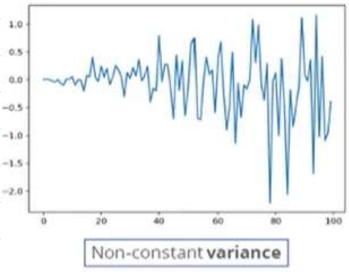
✓ Linear Pattern / Trend
✓ Seasonality
} ⇒ Non-stationary.

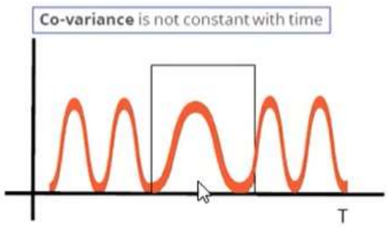
☆ Before working on a time-series data you need to ensure its stationarity and then proceed further.

1.) Check if the time-series is stationary or not. ⟶ yes ⟶ Good to go

2.) If not, then first convert it to stationary and then go for modelling.

Analogy → you are standardizing your data and then extrapolating it, i.e. the forecast happens on the original dataset only





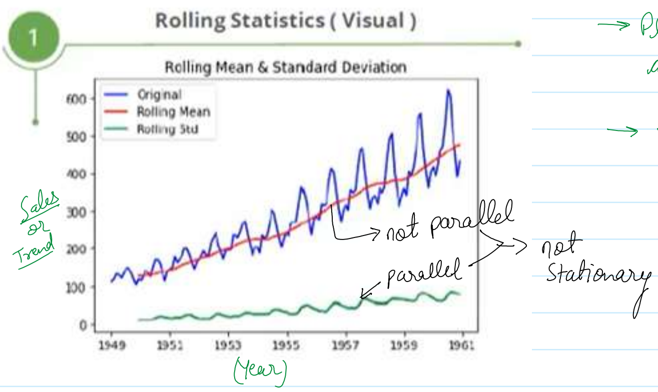
mean ≈ constant
variance & covariance → not constant with time

mean ≈ constant
variance ≈ constant
covariance → variable with time

☆ Plotting Rolling Statistics → To check for stationarity

Method-1 → Plotting the graph visually and checking manually

Method-2 → Plotting Rolling Statistics



→ Plot the rolling mean and rolling standard deviation along with the original data.

→ For the series to be stationary

↳ BOTH mean and standard Deviation have to be constant with time i.e. parallel to X-axis

Method-3 : Dickey-Fuller Test

→ Statistical test to check for stationarity

→ Null hypothesis : Time Series is non-stationary.

→ Test result comprise of : Test Statistic & Some Critical Values for different confidence levels.

→ If the "Test Statistic" is less than the "Critical Value", we can reject the null hypothesis and say that the series is stationary.

Dickey Fuller test (Statistical) 2

```
Test Statistic                  0.815369
p-value                         0.991880
#Lags Used                     13.000000
Number of Observations Used   130.000000
Critical Value (1%)            -3.481682
Critical Value (5%)            -2.884042
Critical Value (10%)           -2.578770
dtype: float64
```

→ Not Stationary

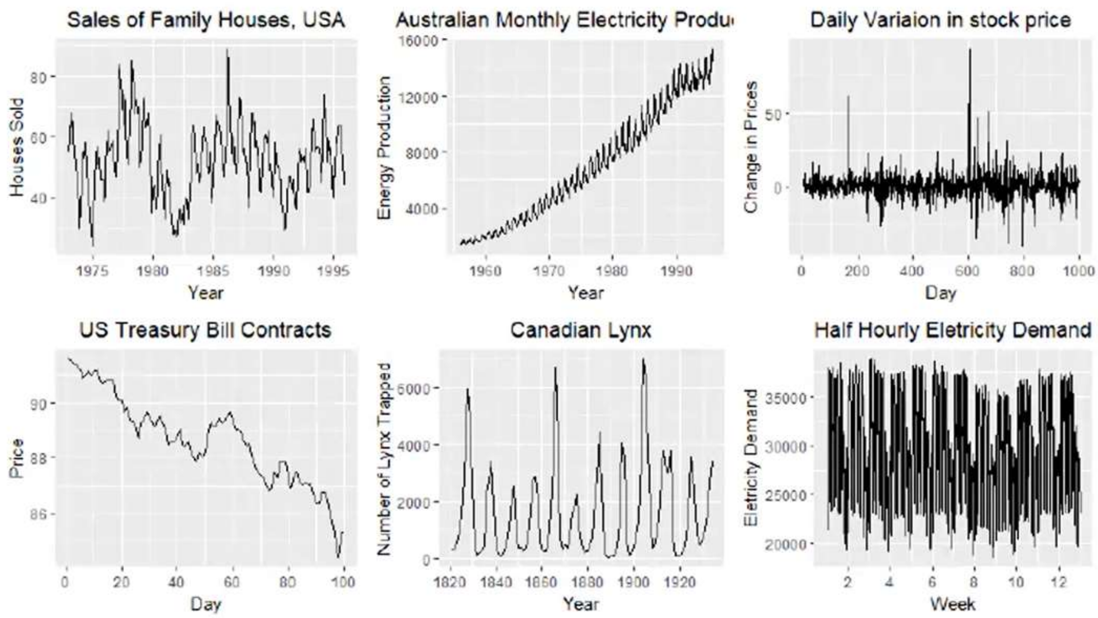
if $p < 0.05$ → Reject the null hypothesis

Null Hypothesis = TS is non-stationary

If 'Test Statistic' < 'Critical Value', Reject the null hypothesis

---

Practice



→ Identify the time series components for each graph.

Graph 1 → The monthly housing sales
- → Strong seasonality within each year
- → Strong Cyclic behaviour with a period of about 6-10 years
- → No apparent trend in the data over this period.

Graph 2 → Australian Energy Production
- → Increasing trend with strong seasonality
- → No apparent cyclicity.

Graph 3 → Daily Variation of stock Price
- → Pure noise
- → No trend, seasonality or cyclicity

Graph 4 → U.S. Treasury Bill
→ No seasonality
→ Downward trend in 100 days

Graph 5 → Canadian lynx
→ Cycles of variable length on average approx 10 years

Graph 6 → Half Hourly Electricity Demand
→ Daily & weekly seasonality.

£ Graph 1 – 6 → None of them have stationarity → as there is some trend or seasonality present

\$ Auto Correlation

aka → Lagged Correlation, Serial Correlation

Autocorrelation → It is a mathematical representation of the degree of similarity between a given time-series and the lagged version of itself over successive time intervals.

In other words, instead of calculating the correlation between two different series, we calculate the correlation of the series with an "x" unit lagged version ($x \in N$) of itself.

* The value of auto correlation varies between +1 & −1.

* If the auto correlation of a series is a very small value that does not mean, there is no correlation; the correlation could be non-linear.

	A	B	C	D
1	ORIGINAL DATA	1 - UNIT LAG	2 - UNIT LAG	3 - UNIT LAG
2	9.08			
3	12.63	9.08		
4	15	12.63	9.08	
5	20.73	15	12.63	9.08
6	2.2	20.73	15	12.63
7	18	2.2	20.73	15
8	7.16	18	2.2	20.73
9	18.28	7.16	18	2.2
10	21	18.28	7.16	18
11	19.68	21	18.28	7.16
12	15.54	19.68	21	18.28
13	24	15.54	19.68	21
14	16.1	24	15.54	19.68
15	11.93	16.1	24	15.54
16	27	11.93	16.1	24
17	12.51	27	11.93	16.1
18	20.04	12.51	27	11.93
19	30	20.04	12.51	27
20	12.41	30	20.04	12.51
21	14.33	12.41	30	20.04
22	33	14.33	12.41	30
23	22.11	33	14.33	12.41
24	17.91	22.11	33	14.33
25	36	17.91	22.11	33

} will be blank

\$ How to make a Time-Series Stationary?

→ We can achieve stationarity through data transformation like taking $\log_{10}$, $\log_e$, square, square root, cube, cube root, exponential decay, time shift.

1.) Differencing: The p-value (>0.05) indicates that we cannot reject the null hypothesis and hence series is non-stationary.

Differencing is performed by subtracting the previous observation from the current observation

or

by subtracting previous day demand from current day demand.

$$\boxed{\Delta y_t = y_t - y_{t-1}}$$

$y_t \rightarrow$ 'y' is the demand function depending on time 't'

* By differencing, stationarity can be achieved easily.

⇒ Time series does not depend on time.

i.e. its like white noise, no matter when it is observed it looks same at any point of time.

* Trends and seasonality affect the time-series at different times.

* Stationary time series does not have any predictable pattern.

* Once you apply differencing, you re-check with rolling statistics to see if your time-series is stationary or not. If not then you again apply differencing which is called Double Differencing and you repeat the same process.

2.) Decomposition of time-series:

Decomposition removes the trending and seasonality pattern by decomposing any non-stationary time-series into trend, seasonal and some random error[having zero mean and correlated over time].

→ We analyze the random error or irregular pattern as stationary component.

They are of 2 types:

a.) Additive Decomposition

$$y_t = T_t + S_t + C_t + I_t$$

where: $T_t \rightarrow$ Trend

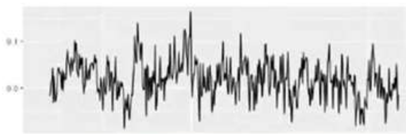
$S_t \rightarrow$ Seasonal

$C_t \rightarrow$ Cyclic

$I_t \rightarrow$ Irregular Pattern

b.) Product Decomposition $y_t = T_t \times S_t \times C_t \times I_t$

Seasonality with increasing trend



Seasonally decomposed series

Plotting ACF and PACF

1.) Auto-Correlation Function (ACF): It refers to the way the observations in a time-series are related to each other.

ACF is the coefficient of correlation in time-series b/w the value of the point at current time and its value at lag k.

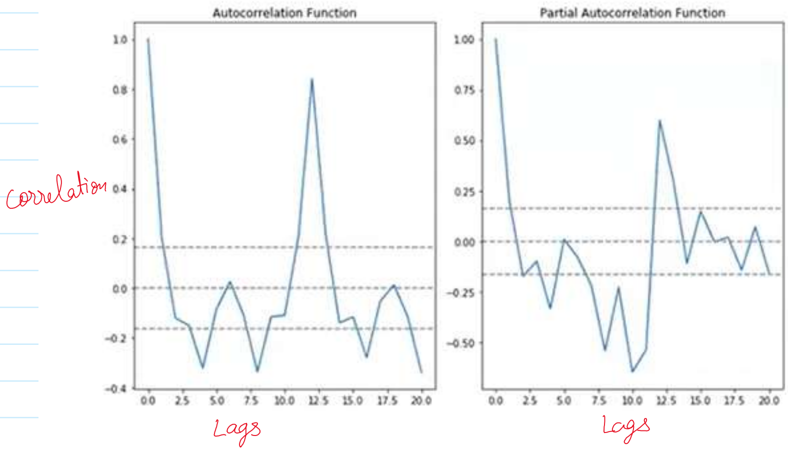
In Layman terms → ACF tells how well is your present value correlated with the past value.

i.e. correlation between $y(t)$ and $y(t-k)$.

* ACF identifies the order of Moving Average (MA) process.

2.) Partial Auto-Correlation Function (PACF) (aka. Auto Regressor): It is the same as ACF, but the intermediate lags between $y(t)$ and $y(t-k)$ are removed (or partial out).

i.e. correlation between $y(t)$ and $y(t-k)$ with $(k-1)$ lags removed.



From the ACF graph, we see that curve touches y=0.0 line at x=2. Thus, from theory, Q = 2 From the PACF graph, we see that curve touches y=0.0 line at x=2. Thus, from theory, P = 2.

Q → Moving Average (MA) → for ACF graph

P → Auto Regressor → for PACF graph

and together it becomes Auto Regressor Moving Average

Interpreting ACF plots

ACF Shape	Indicated Model
Exponential, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to identify the order of the autoregressive model
Alternating positive and negative, decaying to zero Autoregressive model.	Use the partial autocorrelation plot to help identify the order.
One or more spikes, rest are essentially zero	Moving average model, order identified by where plot becomes zero.
Decay, starting after a few lags	Mixed autoregressive and moving average (ARMA) model.
All zero or close to zero	Data are essentially random.
High values at fixed intervals	Include seasonal autoregressive term.
No decay to zero	Series is not stationary