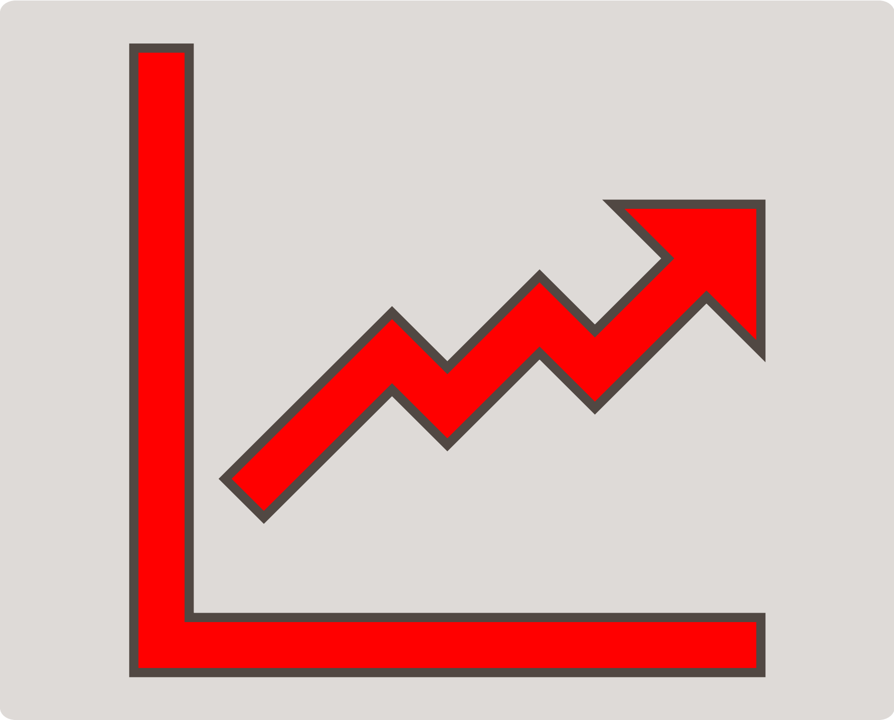


TIME SERIES & FORECASTING JARGON

By Siddharth Dixit

How different industries use time series forecasting

- **Energy** – Prices; demand; production schedules
- **Retail** – Sales; consumer demand for certain products
- **Marketing** – Ad views
- **Transportation** – Demand for future travel
- **Finance** – Stocks; market potential
- **Healthcare-** EEG signals and IOT devices

Describing vs. Predicting

Time Series Analysis

- Primary concern is the analysis of time series. This field of study seeks the why behind a time series.

Time Series Forecasting

- Making predictions about the future(extrapolation) is known as time series forecasting.
- Forecasting involves taking models fit on historical data and using them to predict future observations.

Univariate v/s Multivariate Time Series

- **Univariate Time Series-** Series with a single time-dependent variable.
- **Multivariate Time Series-** Series with more than 1 time-dependent variable.

Time	Temperature
5:00 am	59 °F
6:00 am	59 °F
7:00 am	58 °F
8:00 am	58 °F
9:00 am	60 °F
10:00 am	62 °F
11:00 am	64 °F
12:00 pm	66 °F
1:00 pm	67 °F
2:00 pm	69 °F
3:00 pm	71 °F
4:00 pm	71 °F
5:00 pm	71 °F
6:00 pm	69 °F
7:00 pm	68 °F
8:00 pm	65 °F
9:00 pm	64 °F

Time	Temperature	cloud cover	dew point	humidity	wind
5:00 am	59 °F	97%	51 °F	74%	8 mph SSE
6:00 am	59 °F	89%	51 °F	75%	8 mph SSE
7:00 am	58 °F	79%	51 °F	76%	7 mph SSE
8:00 am	58 °F	74%	51 °F	77%	7 mph S
9:00 am	60 °F	74%	51 °F	74%	7 mph S
10:00 am	62 °F	74%	52 °F	70%	8 mph S
11:00 am	64 °F	76%	52 °F	65%	8 mph SSW
12:00 pm	66 °F	80%	52 °F	60%	8 mph SSW
1:00 pm	67 °F	78%	52 °F	58%	10 mph SW
2:00 pm	69 °F	71%	52 °F	54%	10 mph SW
3:00 pm	71 °F	75%	52 °F	52%	11 mph SW
4:00 pm	71 °F	78%	52 °F	52%	11 mph SW
5:00 pm	71 °F	78%	52 °F	52%	12 mph SW
6:00 pm	69 °F	78%	52 °F	54%	11 mph SW
7:00 pm	68 °F	87%	53 °F	60%	12 mph SW
8:00 pm	65 °F	100%	54 °F	66%	11 mph SSW
9:00 pm	64 °F	100%	55 °F	72%	13 mph SSW

Components of Time Series

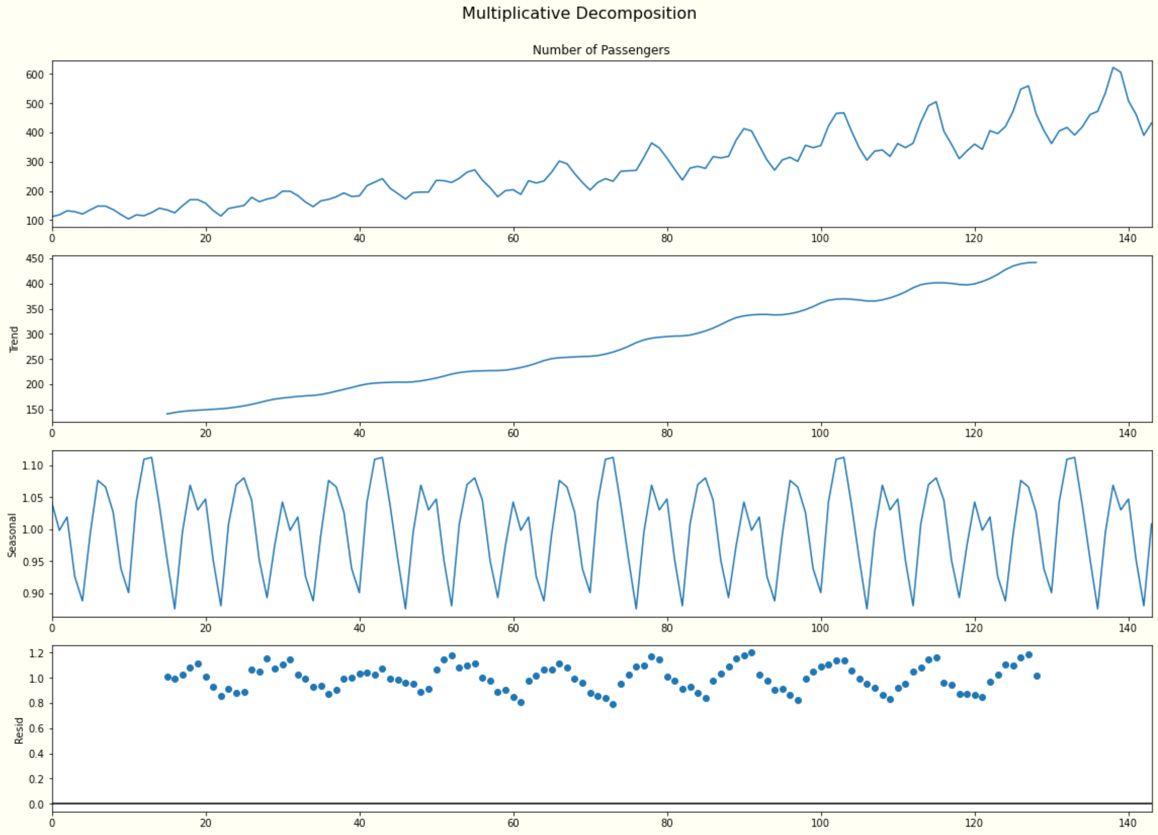
- **Level**: The average value in the series..
- **Trend**: The optional and often linear increasing or decreasing behavior of the series over time.
- **Seasonality**: The optional repeating patterns or cycles of behavior over time.
- **Noise**: The optional variability in the observations that cannot be explained by the model.

These constituent components can be thought to combine in some way to provide the observed time series.

Decomposition of a Time Series

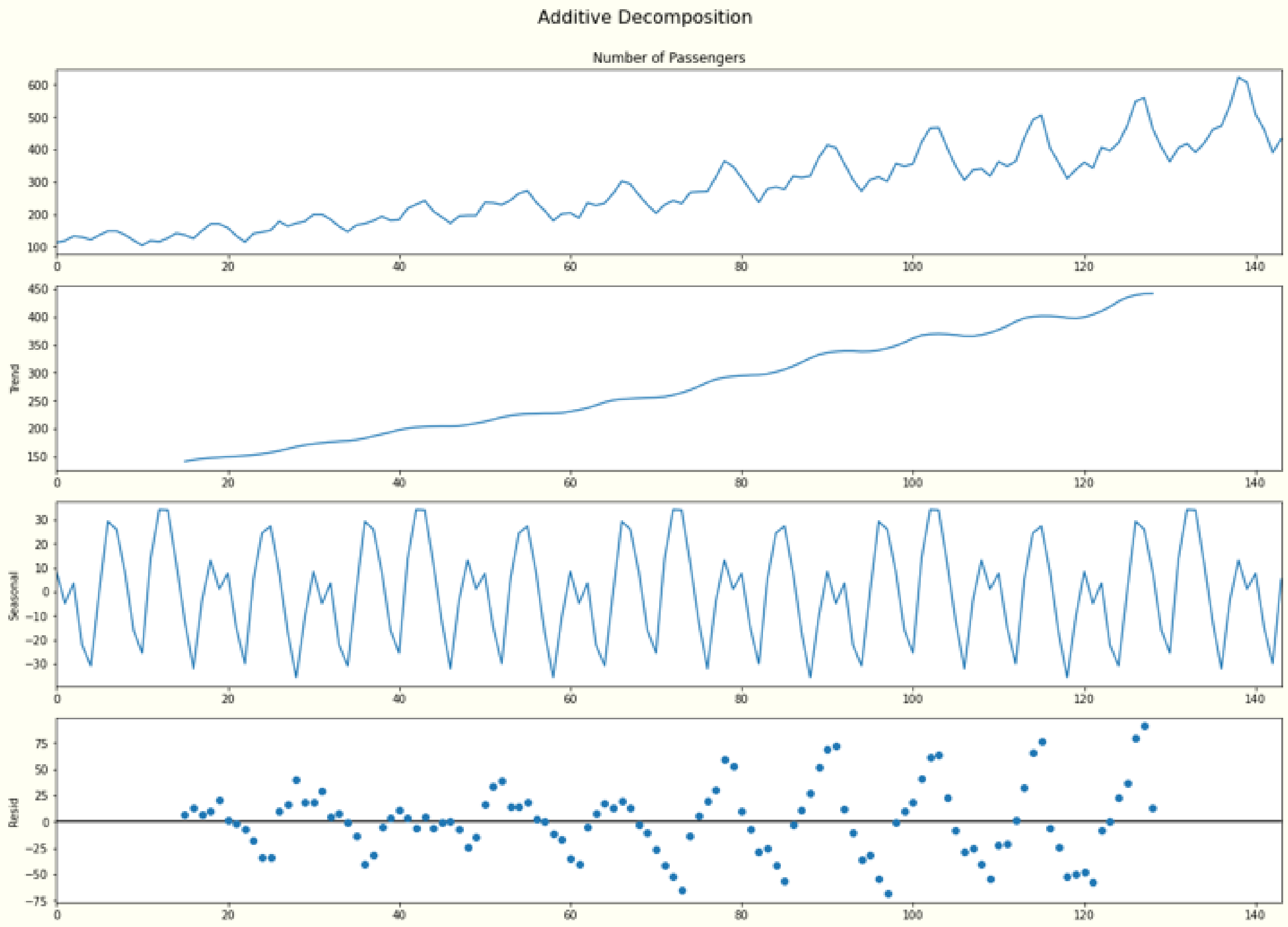
Decomposition of a time series can be performed by considering the series as an additive or multiplicative combination of the base level, trend, seasonal index and the residual term.

Multiplicative Time Series: *Value = Base Level x Trend x Seasonality x Error*



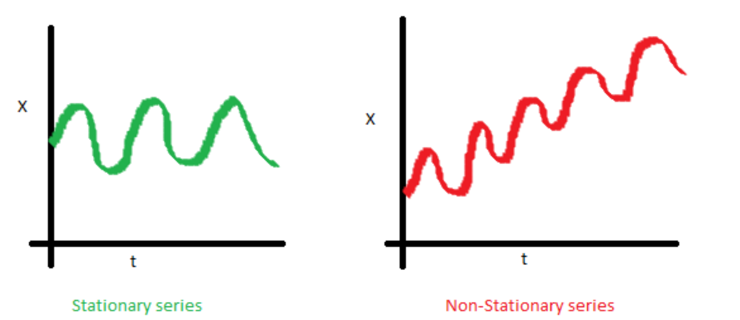
Decomposition of a Time Series

Additive time series: *Value = Base Level + Trend + Seasonality + Error*

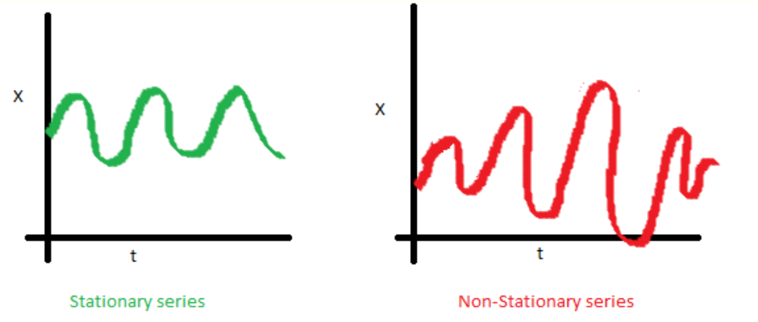


Stationary v/s Non-Stationary Time Series

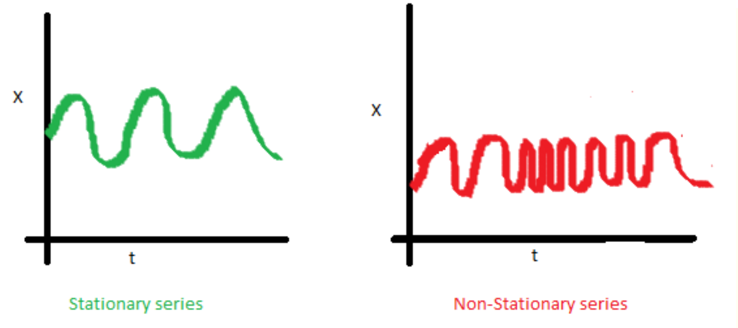
1.



2.



3.



If a process is stationary, that means it does not change its statistical properties over time, namely its mean and variance. (The constancy of variance is called *homoscedasicity*) . The covariance function should also not depend on time; it should only depend on the distance between observations.

1. The red graph is not stationary because the mean increases over time.

2. Varying spread of values over time => Non-Constant Variance

3. Covariance of the *i term* and the *(i + m) term* should not be a function of time. In the following graph, you will notice that the spread becomes closer as time increases. Hence, the covariance is not constant with time in the red graph

How to check whether a given Time Series is stationary?

There are many methods to check whether a time series is stationary or non-stationary:-

- **Look at Plots:** You can review a time series plot of your data and visually check if there are any obvious trends or seasonality.
- **Summary Statistics:** You can review the summary statistics for your data for seasons or random partitions and check for obvious or significant differences.
- **Statistical Tests:** You can use statistical tests to check if the expectations of stationarity are met or have been violated. A popular test for checking stationarity is Augmented Dickey Fuller Test.

So why is Stationarity so important?

- It is easy to make predictions on a stationary series since we can assume that the future statistical properties will not be different from those currently observed.
- Most of the time-series models, in one way or the other, try to predict those properties (mean or variance, for example).
- Future predictions would be wrong if the original series were not stationary. Unfortunately, most of the time series in real world are non-stationary, but we can (and should) change this.

You can make series stationary by:

- Differencing the Series (once or more)
- Take the log of the series
- Take the nth root of the series
- Combination of the above

Concerns of Forecasting

1. How much data do you have available and are you able to gather it all together?

More data is often more helpful, offering greater opportunity for exploratory data analysis, model testing and tuning, and model fidelity.

2. What is the time horizon of predictions that is required? Short, medium or long term?

Shorter time horizons are often easier to predict with higher confidence.

3. Can forecasts be updated frequently over time or must they be made once and remain static?

Updating forecasts as new information becomes available often results in more accurate predictions.

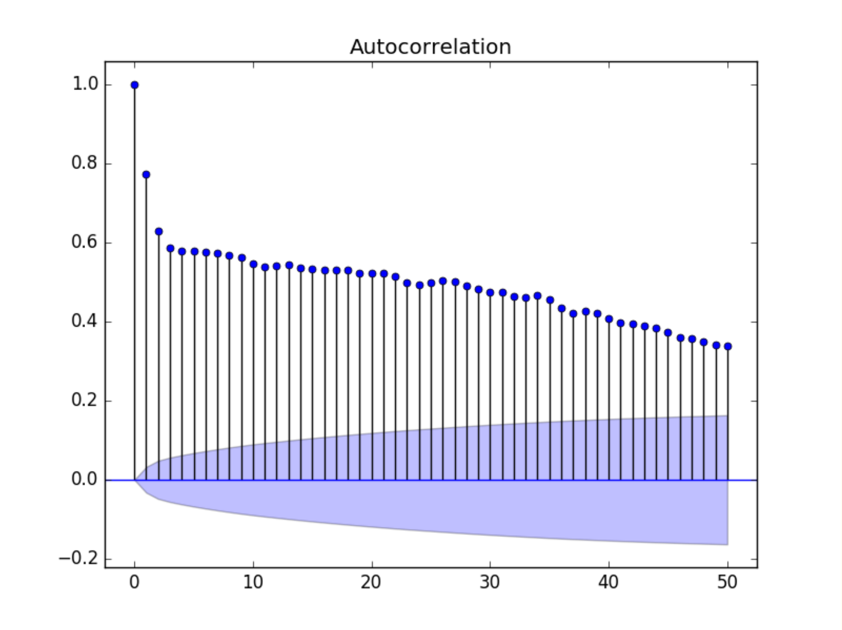
4. At what temporal frequency are forecasts required?

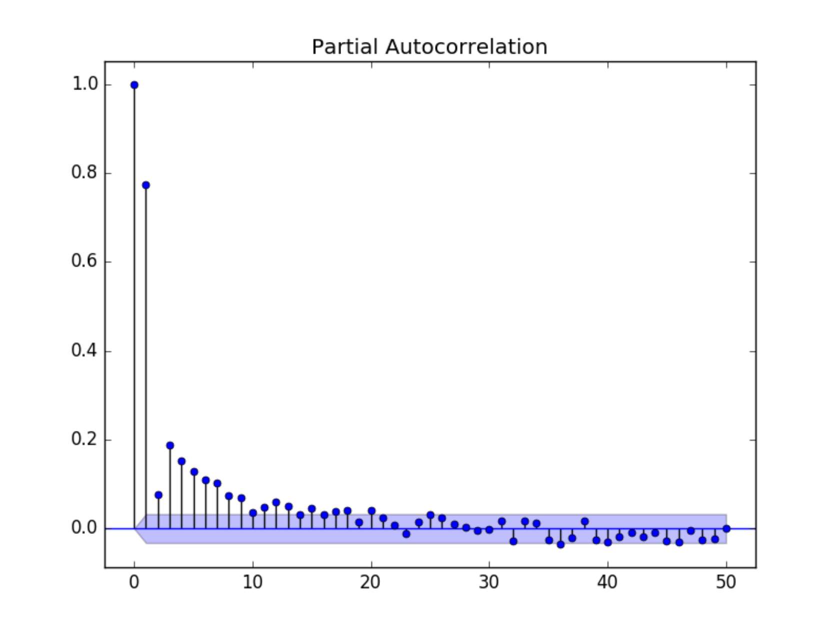
Often forecasts can be made at a lower or higher frequencies, allowing you to harness down-sampling, and up-sampling of data, which in turn can offer benefits while modeling.

ACF v/s PACF

- ACF:- A plot of the autocorrelation of a time series by lag is called the Auto-Correlation Function (ACF). Autocorrelation for an observation and an observation at a prior time step is comprised of both the *direct correlation* and *indirect correlations*. These indirect correlations are a linear function of the correlation of the observation, with observations at intervening time steps. It is these indirect correlations that the partial autocorrelation function seeks to remove.

- PACF:- A plot of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.





Econometric Models for Time Series Forecasting

- <u>Autoregression</u>:-

AR(p) - Autoregression model, i.e. regression of the time series onto itself. The basic assumption is that the current series values depend on its previous values with some lag (or several lags). The maximum lag in the model is referred to as p. To determine the initial p, you need to look at the *PACF* plot and find the biggest significant lag after which **most** other lags become insignificant.

AR(3) model: $Y_t = a + bY_{t-1} + cY_{t-2} + dY_{t-3} + \varepsilon_t$

- <u>Moving Average</u>:-

MA(q) - This models the error of the time series, again with the assumption that the current error depends on the previous with some lag, which is referred to as q. The initial value can be found on the PACF plot with the same logic as before.

MA(1) model: $Y_t = \mu + a.\epsilon_t + b.\epsilon_{t-1}$

It translates to Today's Sales = mean + today's noise + yesterday's noise

Econometric Models for Time Series Forecasting

- **Autoregressive Moving Average**:- *AR(p) + MA(q) = ARMA(p,q)*

$R_t = \mu + \varphi R_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$

Basically, Today's Sales = Mean + Yesterday's Sales + noise + yesterday's noise

- **Autoregressive Integrated Moving Average**:- ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

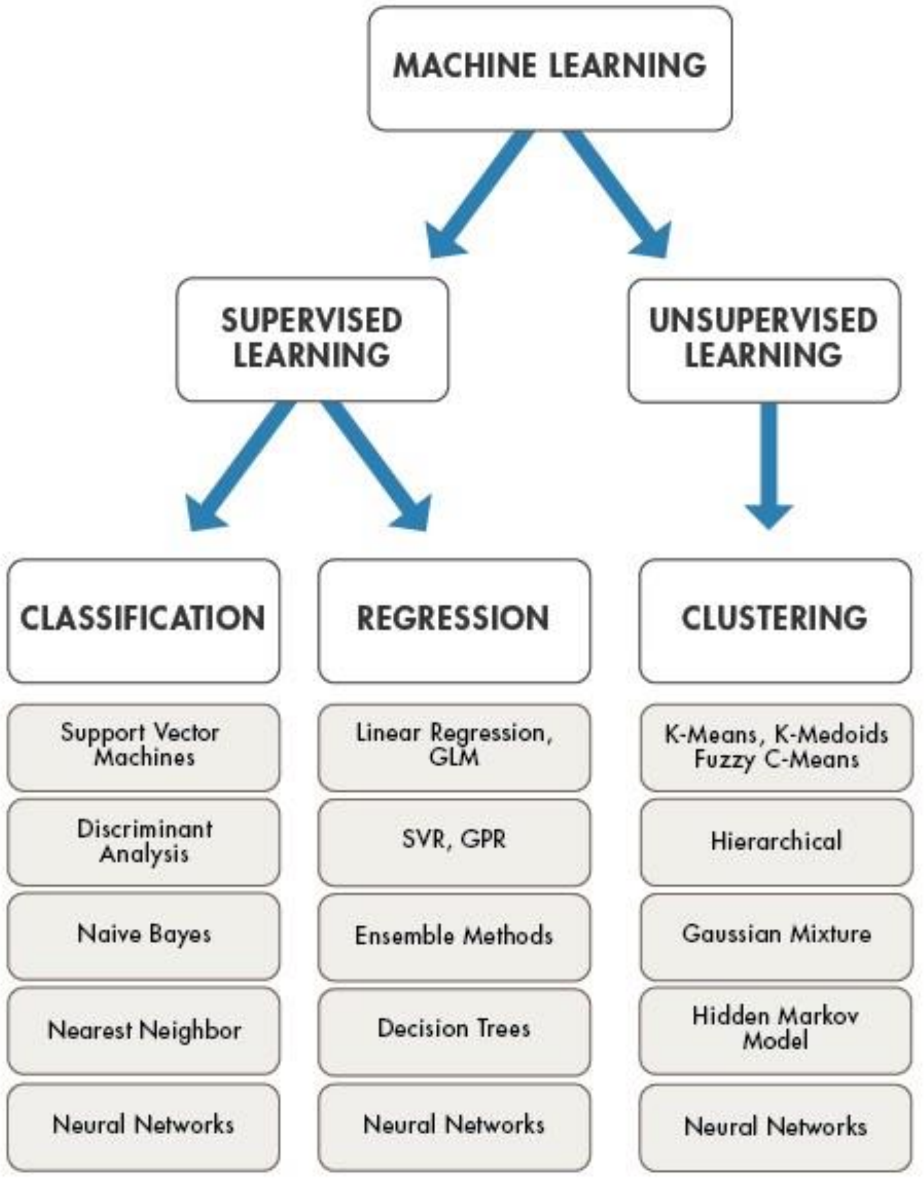
ARIMA(p,d,q): *p* is *AR* parameter, *d* is *differential* parameter, *q* is *MA* parameter

ARIMA(1,0,1): $y_t = a_1 y_{t-1} + \epsilon_t + b_1\epsilon_{t-1}$

ARIMA(1,1,1): $\Delta y_t = a_1 \Delta y_{t-1} + \epsilon_t + b_1\epsilon_{t-1}$ *where* $\Delta y_t = y_t - y_{t-1}$

And many more… such as SARIMA, ARCH, GARCH, VAR etc.

CAN WE MAKE USE OF MACHINE LEARNING METHODS FOR FORECASTING TIME SERIES?



Time Series as Supervised Learning Problem-I (Sliding Window Method)

Day	Sales
1	1000
2	900
3	300
4	120
5	260
6	1100
7	?

→

Day	S_lag1	S_lag2	Sales
1	N.A.	N.A.	1000
2	1000	N.A.	900
3	900	1000	300
4	300	900	120
5	120	300	260
6	260	120	1100
7	1100	260	?

- **One-step Forecast:** This is where the next time step (t+1) is predicted.
- **Multi-step Forecast:** This is where two or more future time steps are to be predicted.

Time Series as Supervised Learning Problem- II (Feature Engineering)

Broadly speaking there are three classes of features that can be created from our time series dataset:-

1. **Date Time Features:** Components of the time step itself for each observation.
2. **Lag Features:** Values at prior time steps.
3. **Rolling Window Features:** Summary of values over a fixed window of prior time steps.

	Time	Ads
0	2017-09-13 00:00:00	80115
1	2017-09-13 01:00:00	79885
2	2017-09-13 02:00:00	89325
3	2017-09-13 03:00:00	101930
4	2017-09-13 04:00:00	121630

Time	Ads	hour_of_day	Ads_lag1	Ads_lag2	Ads_lag3	Ads_lag4	Ads_lag5	Ads_lag6	rolling_mean_3	rolling_std_3	rolling_max_3	rolling_min_3
2017-09-13 10:00:00	131030	10	116125.0	108055.0	102795.0	106495.0	116475.0	121630.0	105781.666667	2701.579785	108055.0	102795.0
2017-09-13 11:00:00	149020	11	131030.0	116125.0	108055.0	102795.0	106495.0	116475.0	108991.666667	6714.181509	116125.0	102795.0
2017-09-13 12:00:00	157590	12	149020.0	131030.0	116125.0	108055.0	102795.0	106495.0	118403.333333	11655.717839	131030.0	108055.0
2017-09-13 13:00:00	150715	13	157590.0	149020.0	131030.0	116125.0	108055.0	102795.0	132058.333333	16471.592465	149020.0	116125.0
2017-09-13 14:00:00	149295	14	150715.0	157590.0	149020.0	131030.0	116125.0	108055.0	145880.000000	13555.556056	157590.0	131030.0

References:-

- Machine Learning Mastery by Jason Brownlee
- Introduction to Time Series and Forecasting- Richard A. Davis
- Forecasting Principals and Practice- Rob J. Hyndman
- Kaggle