Autoregression

A regression model, such as linear regression, models an output value based on a linear combination of input values

$$yhat = b0 + b1*X1$$

Where yhat is the prediction, b0 and b1 are coefficients found by optimizing the model on training data, and X is an input value.

Technique can be used on time series where input variables are taken as observations at previous time steps, called lag variables.

$$X(t+1) = b0 + b1*X(t-1) + b2*X(t-2)$$

The **regression model** uses data from the same input variable at previous time steps, it is referred to as an **autoregression** (regression of self).

Auto regressive (AR) process , a time series is said to be AR when present value of the time series can be obtained using previous values of the same time series i.e the present value is weighted average of its past values. Stock prices and global temperature rise can be thought of as an AR processes.

The AR process of an order p can be written as,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Autocorrelation

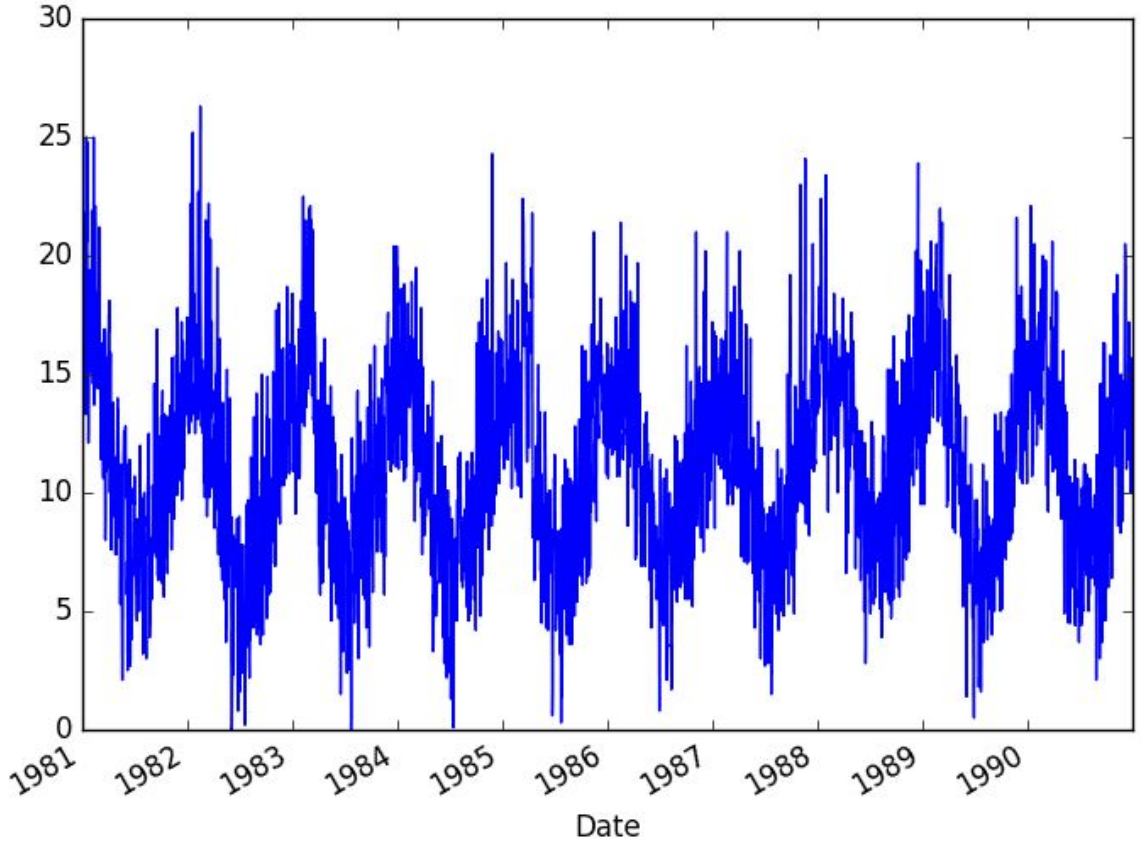
An autoregression model makes an assumption that the observations at previous time steps are useful to predict the value at the next time step.

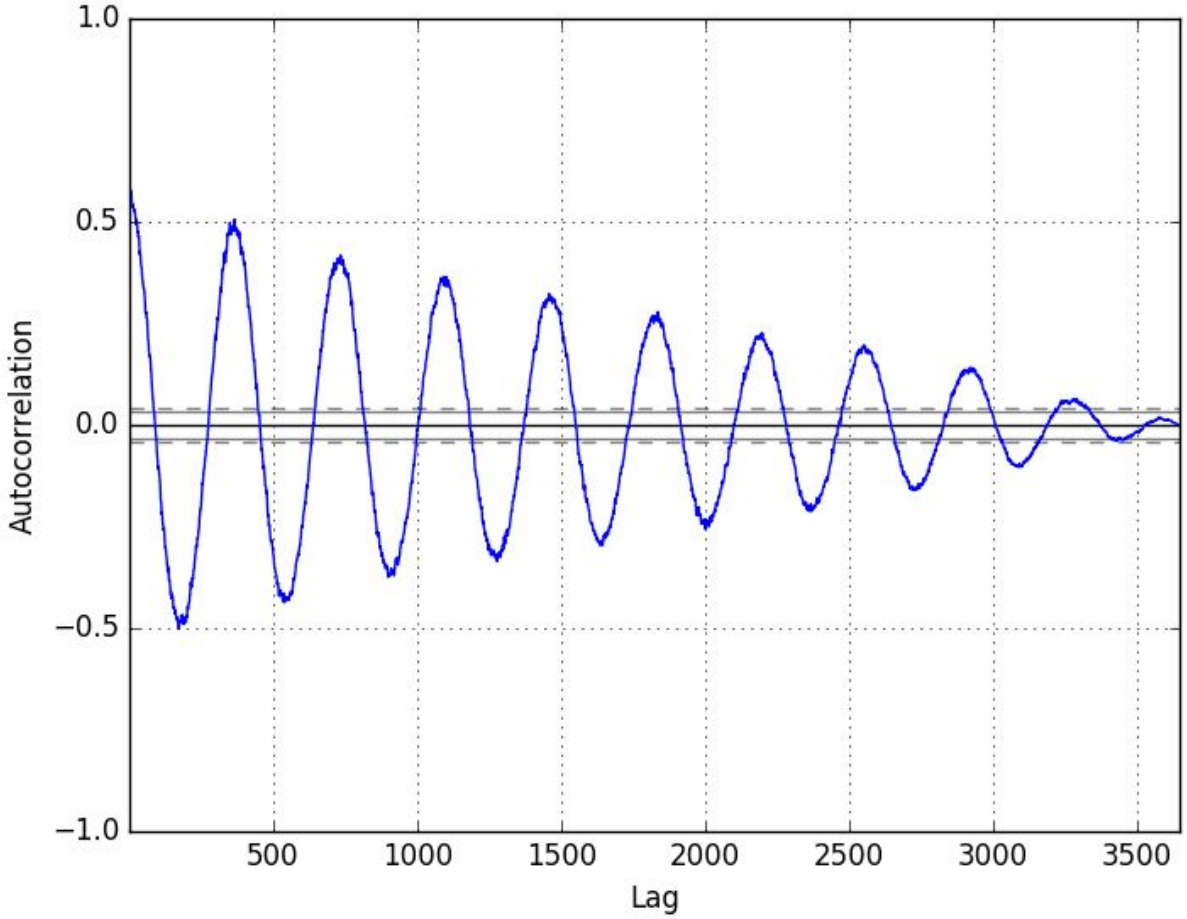
This relationship between variables is called **correlation**.

If both variables change in the same direction (e.g. go up together or down together), this is called a **positive correlation**. If the variables move in opposite directions as values change (e.g. one goes up and one goes down), then this is called **negative correlation.**

We can use statistical measures to calculate the correlation between the output variable and values at previous time steps at various different lags. The stronger the correlation between the output variable and a specific lagged variable, the more weight that autoregression model can put on that variable when modeling

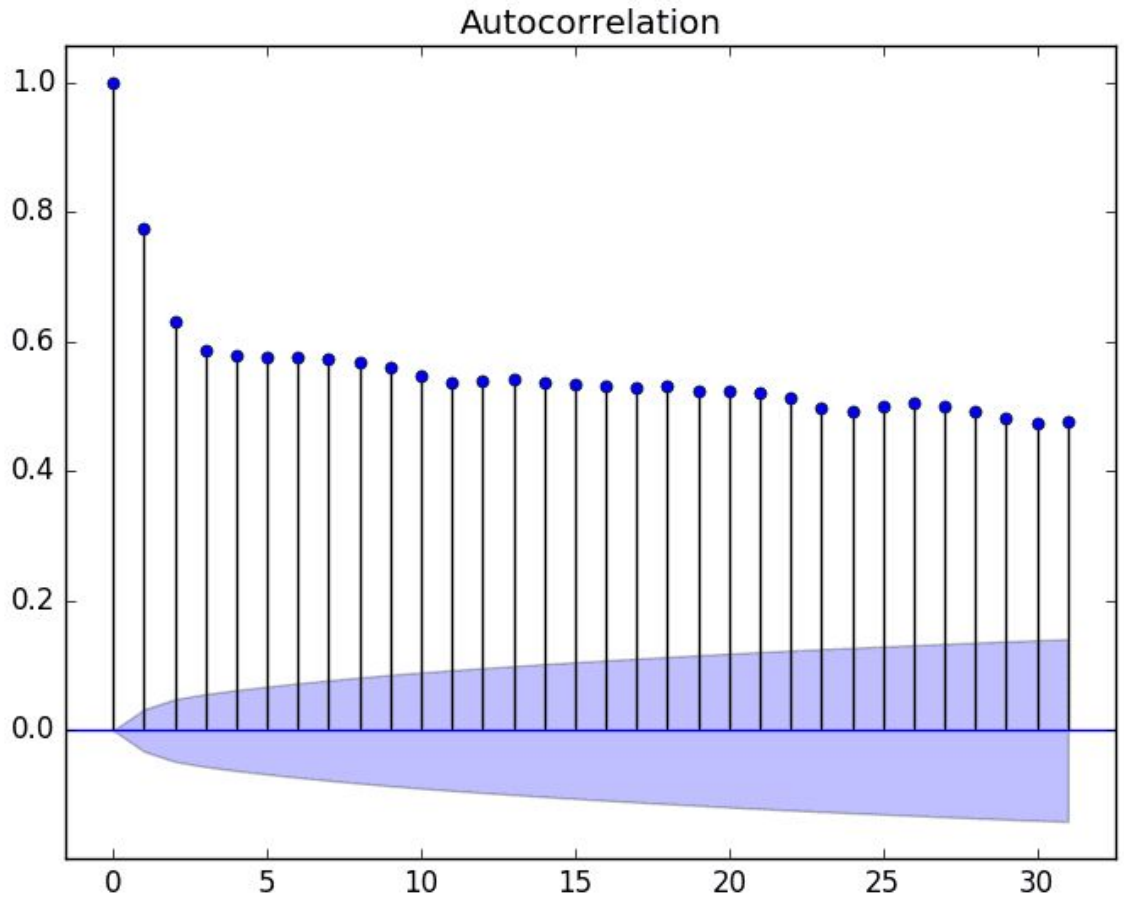
The correlation is calculated between the variable and itself at previous time steps, it is called an **autocorrelation**. It is also called serial correlation because of the sequenced structure of time series data

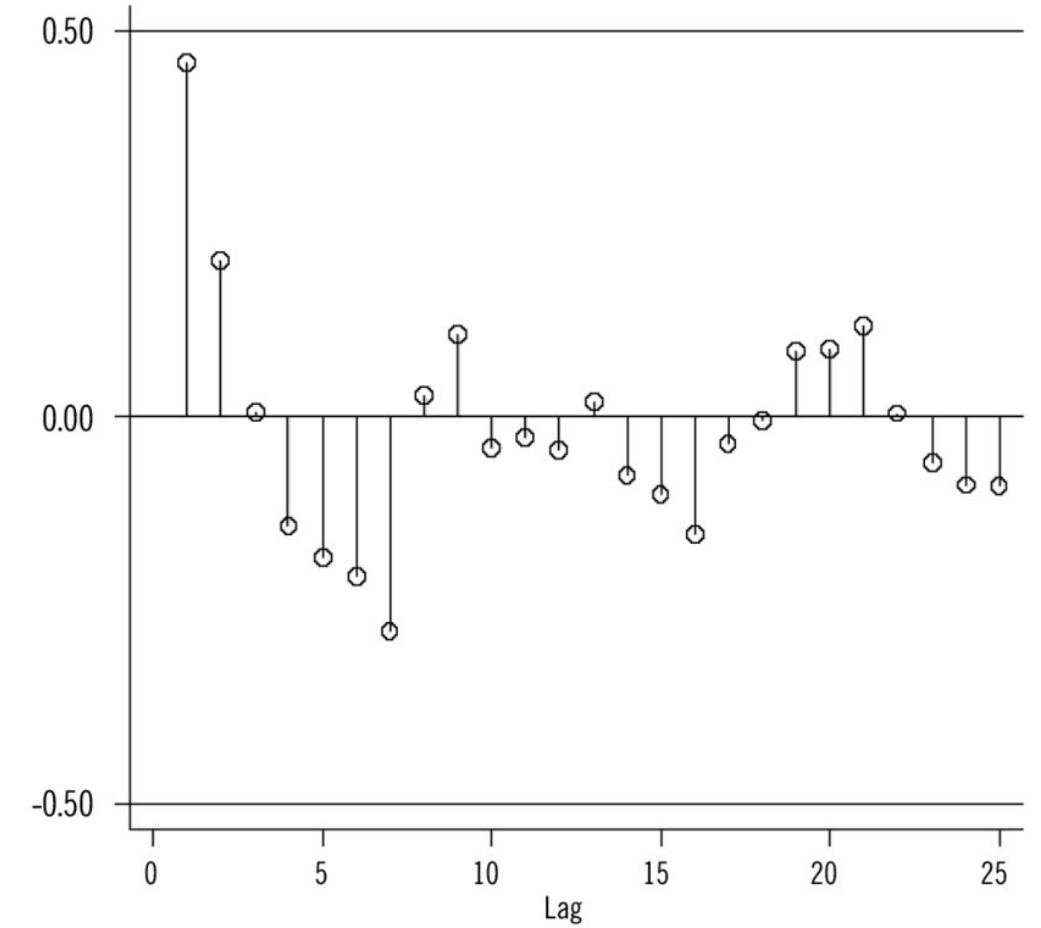

30
25
20
15
10
5
0
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
Date


1.0
0.5
0.0
−0.5
−1.0
Autocorrelation
500
1000
1500
2000
2500
3000
3500
Lag

ACF

Autocorrelation Function


Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30



Bartlett's formula for MA(q) 95% confidence bands

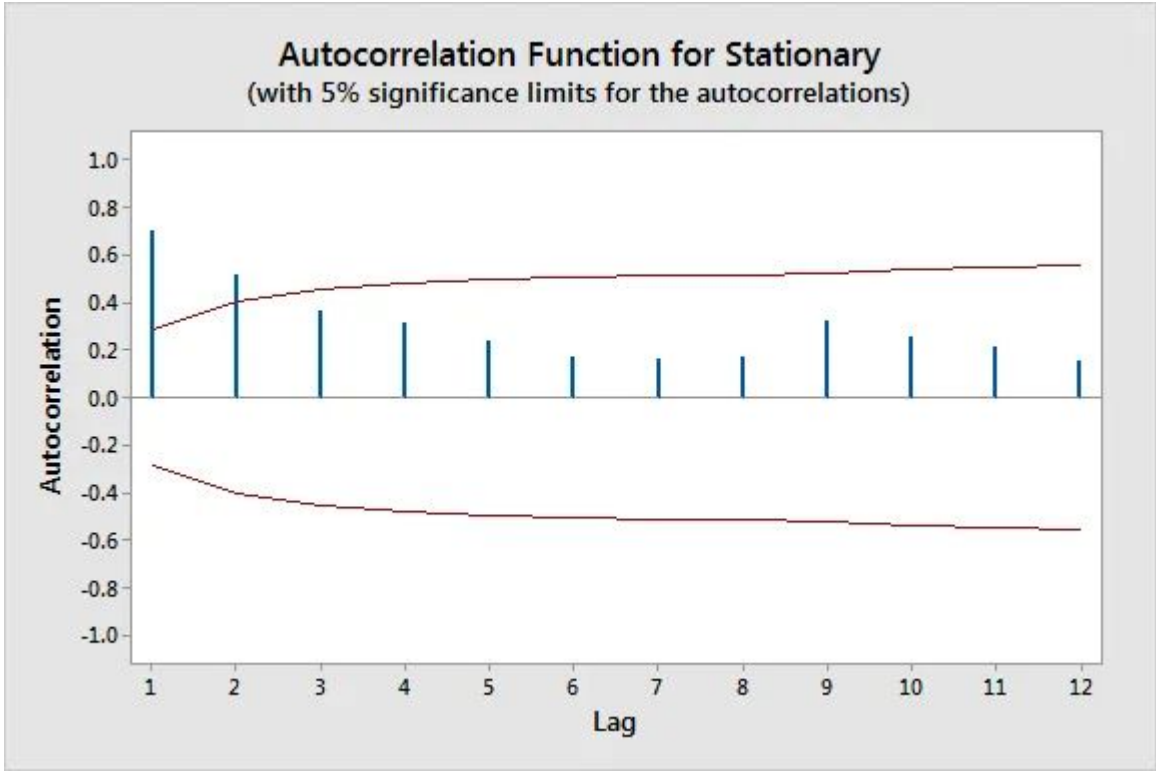
Autocorrelation Function (ACF)

Use the autocorrelation function (ACF) to identify which lags have significant correlations, understand the patterns and properties of the time series, and then use that information to model the time series data. From the ACF, you can assess the randomness and stationarity of a time series. You can also determine whether trends and seasonal patterns are present.

In an ACF plot, each bar represents the size and direction of the correlation. Bars that extend across the red line are statistically significant

Stationarity

Stationarity means that the time series does not have a trend, has a constant variance, a constant autocorrelation pattern, and no seasonal pattern. The autocorrelation function declines to near zero rapidly for a stationary time series. In contrast, the ACF drops slowly for a non-stationary time series



Additive Time Series

- Value = Base Level + Trend + Seasonality + Error

Multiplicative Time Series

- Value = Base Level * Trend * Seasonality * Error

A stationary time series has statistical properties or moments (e.g., mean and variance) that do not vary in time. Stationarity, then, is the status of a **stationary time series**.
Conversely, **nonstationarity** is the status of a time series whose statistical properties are changing through time.

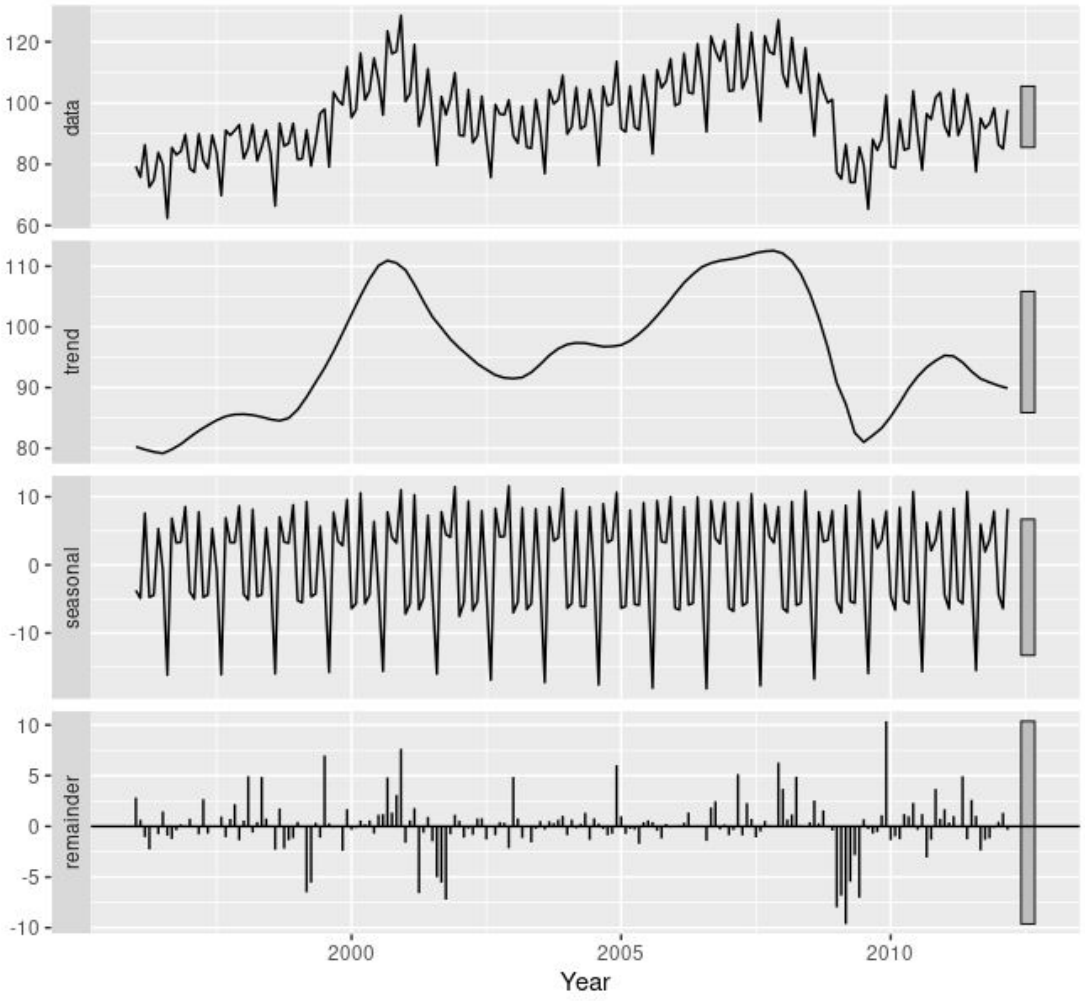
A **stationary time series** is one whose properties do not depend on the time thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times. On the other hand, a white noise series is stationary — it does not matter when you observe it, it should look much the same at any point in time.

1) Seasonality
Seasonality is a simple term that means while predicting a time series data there are some months in a particular domain where the output value is at a peak as compared to other months. for example if you observe the data of tours and travels companies of past 3 years then you can see that in November and December the distribution will be very high due to holiday season and festival season. So while forecasting time series data we need to capture this seasonality.

2) Trend
The trend is also one of the important factors which describe that there is certainly increasing or decreasing trend time series, which actually means the value of organization or sales over a period of time and seasonality is increasing or decreasing.

3) Unexpected Events
Unexpected events mean some dynamic changes occur in an organization, or in the market which cannot be captured. for example a current pandemic we are suffering from, and if you observe the Sensex or nifty chart there is a huge decrease in stock price which is an unexpected event that occurs in the surrounding.

Methods and algorithms are using which we can capture seasonality and trend But the unexpected event occurs dynamically so capturing this becomes very difficult


data
120
100
80
60
trend
110
100
90
80
seasonal
10
0
-10
remainder
10
5
0
-5
-10
2000
2005
2010
Year

When we make a model for forecasting purposes in time series analysis, we require a stationary time series for better prediction. So the first step to work on modeling is to make a time series stationary.

Testing for stationarity is a frequently used activity in **autoregressive modeling**

ADF (Augmented Dickey-Fuller) test is a statistical significance test which means the test will give results in hypothesis tests with null and alternative hypotheses. As a result, we will have a p-value from which we will need to make inferences about the time series, whether it is stationary or not.

Unit Root Test

LINK

A unit root test tests whether a time series is not stationary and consists of a unit root in time series analysis. The presence of a unit root in time series defines the null hypothesis, and the alternative hypothesis defines time series as stationary.

Mathematically the unit root test can be represented as

$$y_t = D_t + z_t + \varepsilon_t$$

Where,

- Dt is the deterministic component.
- z_t is the stochastic component.
- ε_t is the stationary error process.

The unit root test's basic concept is to determine whether the z_t (stochastic component) consists of a unit root or not.

ARIMA MODEL

Moving average (MA) process, a process where the present value of series is defined as a linear combination of past errors. We assume the errors to be independently distributed with the normal distribution. The MA process of order q is defined as ,

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

Here ϵt is a white noise. To get intuition of MA process lets consider order 1 MA process which will look like,

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

A **moving average** term in a time series model is a past error (multiplied by a coefficient).

Let $w_t \overset{iid}{\sim} N(0, \sigma_w^2)$, meaning that the w_t are identically, independently distributed, each with a normal distribution having mean 0 and the same variance.

The **1st order moving average** model, denoted by MA(1) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1}$$

The **2nd order moving average** model, denoted by MA(2) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$$

The **qth order moving average** model, denoted by MA(q) is:

$$x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}$$

https://online.stat.psu.edu/stat510/lesson/2/2.1

A moving average process, or the moving average model, states that the current value is linearly dependent on the current and past error terms. Again, the error terms are assumed to be mutually independent and normally distributed, just like white noise.

A moving average model is denoted as MA(q) where q is the order. The model expresses the present value as a linear combination of the mean of the series (*mu*), the present error term (*epsilon*), and past error terms (*epsilon*). The magnitude of the impact of past errors on the present value is quantified using a coefficient denoted with *theta*. Mathematically, we express a general moving average process as follows:

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

General equation of the MA(q) model

The order q of the moving average model determines the number of past error terms that affect the present value. For example, if it is of order one, meaning that we have a MA(1) process, then the model is expressed as follows:

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1}$$

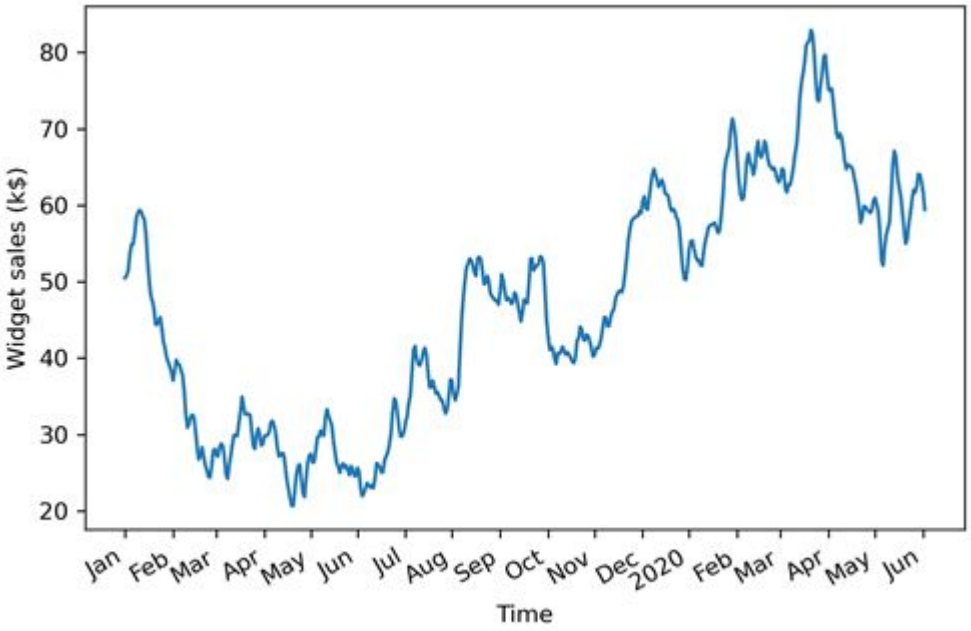
Equation of the MA(1) model

If we have a moving average process of order two, or MA(2), then we express the equation like this:

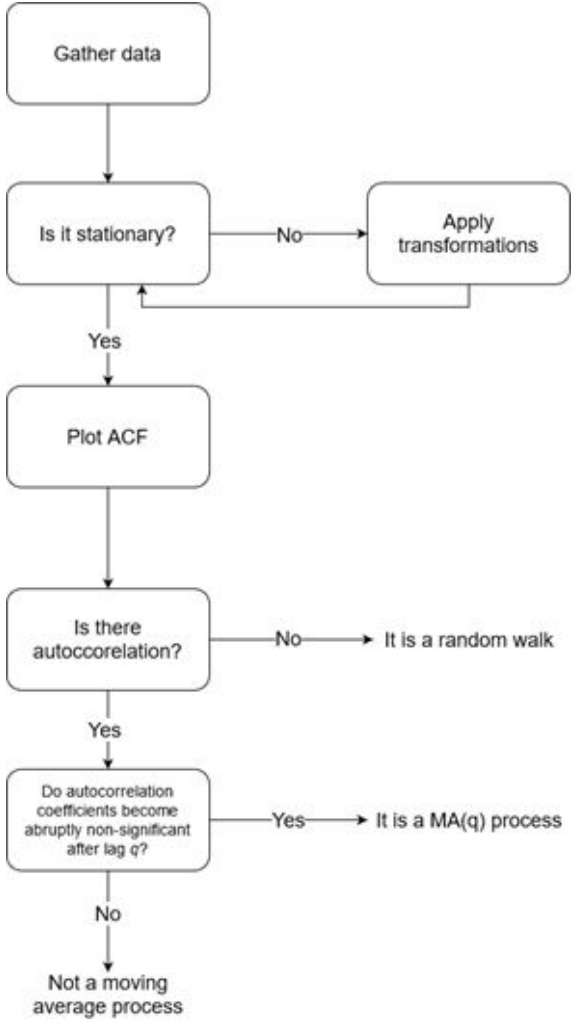
$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

Equation of the MA(2) model

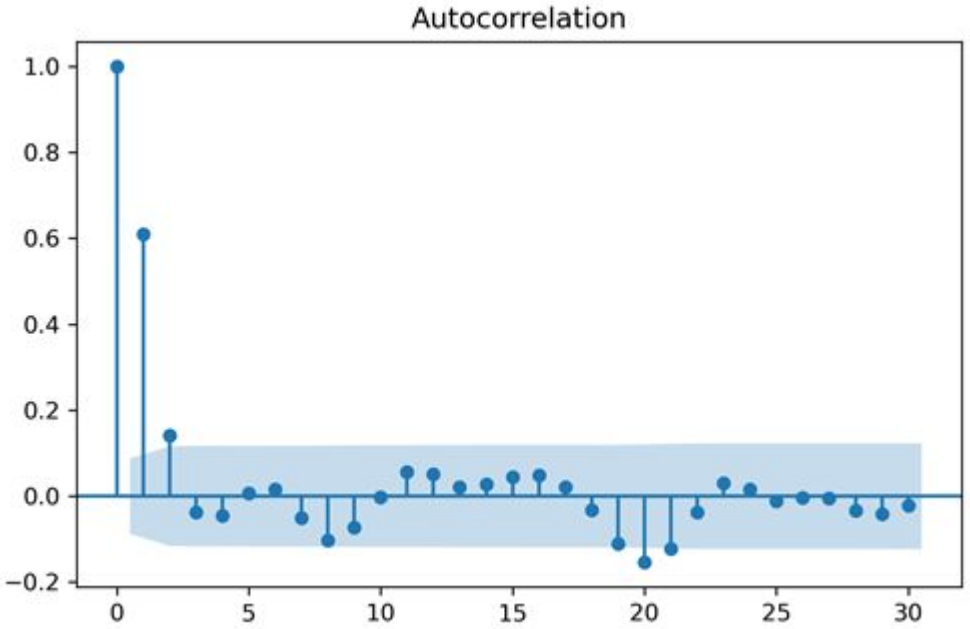
Hence, we can see how the order q of the MA(q) process affects the number of past error terms that must be included in the model. The larger q is, the more past error terms affect the present value. Therefore, it is important to determine the order of the moving average process in order to fit the appropriate model, meaning that if we have a second-order moving average process, then a second-order moving average model will be used for forecasting.


Widget sales (k$)
80
70
60
50
40
30
20
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2020
Feb
Mar
Apr
May
Jun
Time

1. Our series is not stationary, we apply transformations, such as differencing, until the series is stationary.
2. we plot the ACF and look for significant autocorrelation coefficients. In the case of a random walk, we will not see significant coefficients
3.



We have significant autocorrelation coefficients up until lag 2, this means that we have a stationary moving average process of order 2. Therefore, we can use a second-order moving average model, or MA(2) model, to forecast our stationary time series.



Autoregressive Integrated Moving Average Model

AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model, also called the lag order.
d: The number of times that the raw observations are differenced, also called the degree of differencing.
q: The size of the moving average window, also called the order of moving average.

Use the ACF plot to determine the order of an MA(q) model.