Modul 3

Time Series Forecasting – ARIMA and SARIMA

Purwadhika
Startup and Coding School

Time Series

- Time series is a sequence of observations recorded at a regular time
- The frequency could be Yearly, Monthly, Daily or even milliseconds
- The data analysis for time series is inherently different compared to the other data because:
    - It is time dependent
    - Time series could contain trend-cycle and seasonality

Time Series Advantages

- Given a time series data, if modelled right the data could provide a massive business advantage
- E.g. forecasting the sales or demand of the company in the next month

Time Series Forecasting Types

- Commonly, Time Series Forecasting are divided into 2 type:
  - Univariate Time Series Forecasting

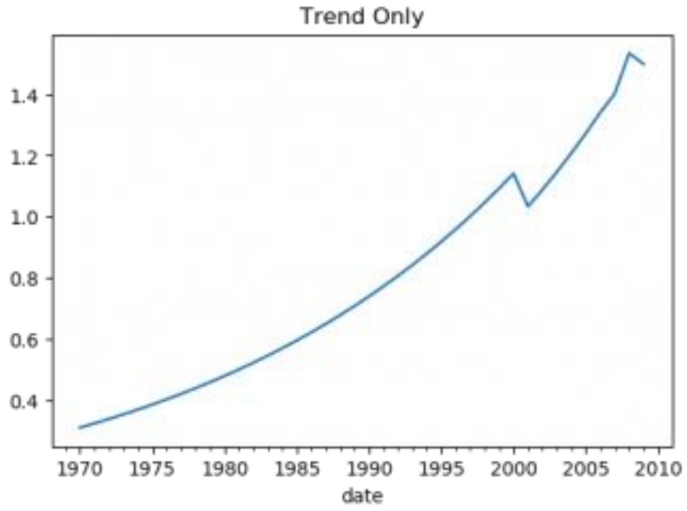
    Forecasting by relying only to the previous values of the time series
  - Multi Variate Time Series Forecasting

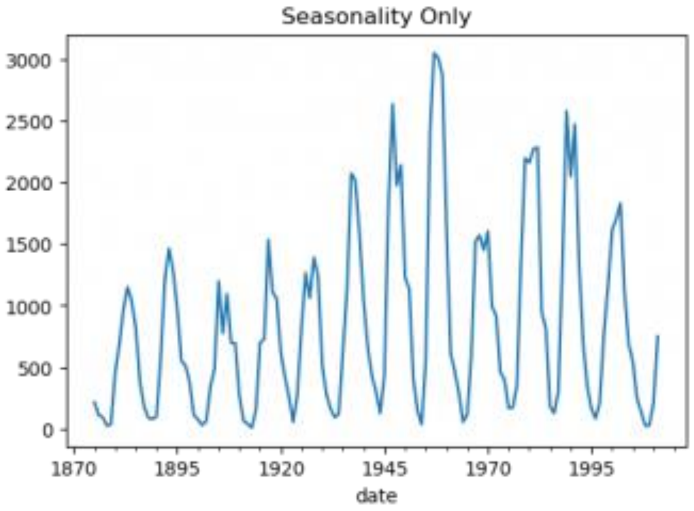
    Forecasting by an additional variable other than the series value (Exogenous Variable)

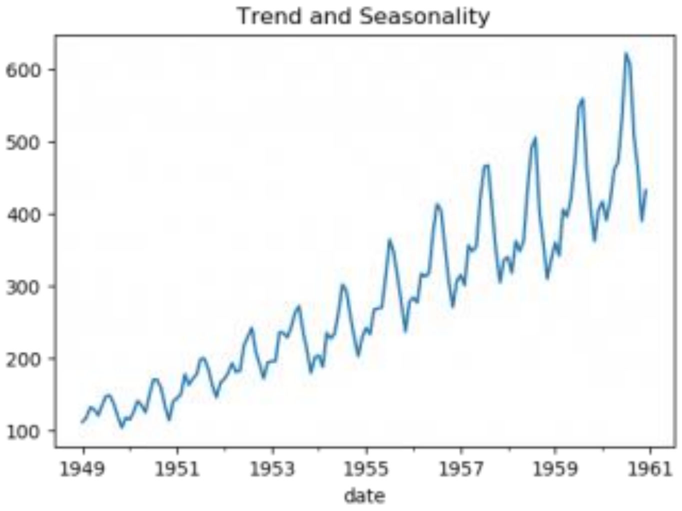
Time Series Pattern

- Time series could be divided into 4 parts. Base, Trend, Seasonality, and Error
- Trend is an increasing or decreasing slope observed in the time series
- Seasonality is a distinct repeated pattern observed between regular intervals due to seasonal factors. It could be because of the month of the year, the day of the month, weekdays or even time of the day.
- Time series is not always having Trend and/or Seasonality term in the series

Time Series Pattern



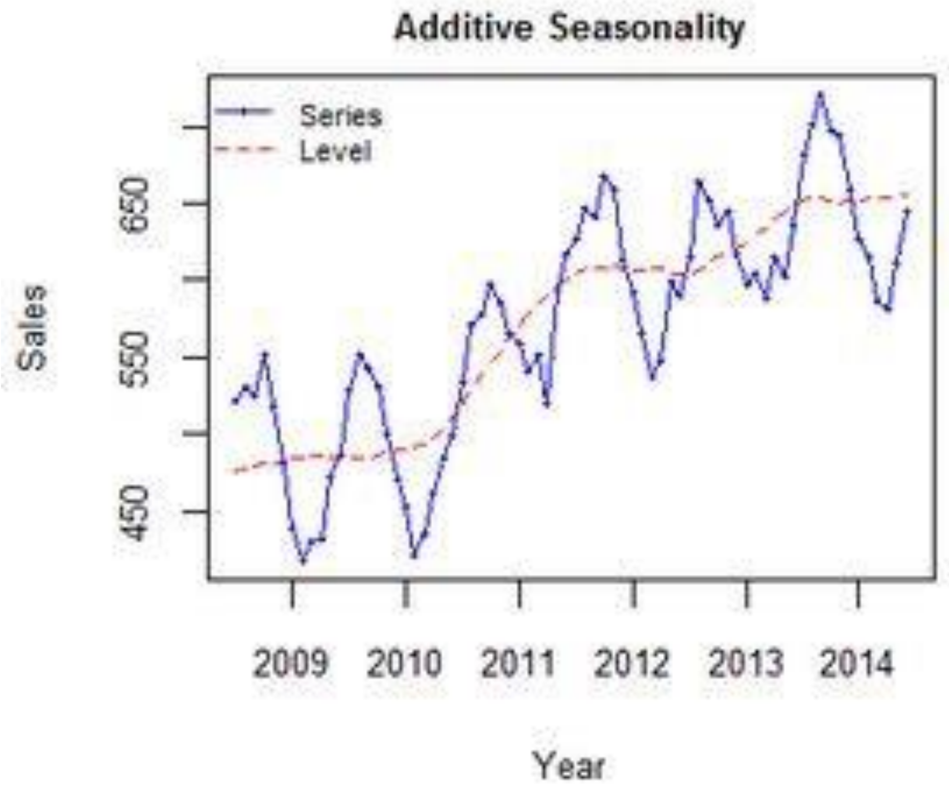


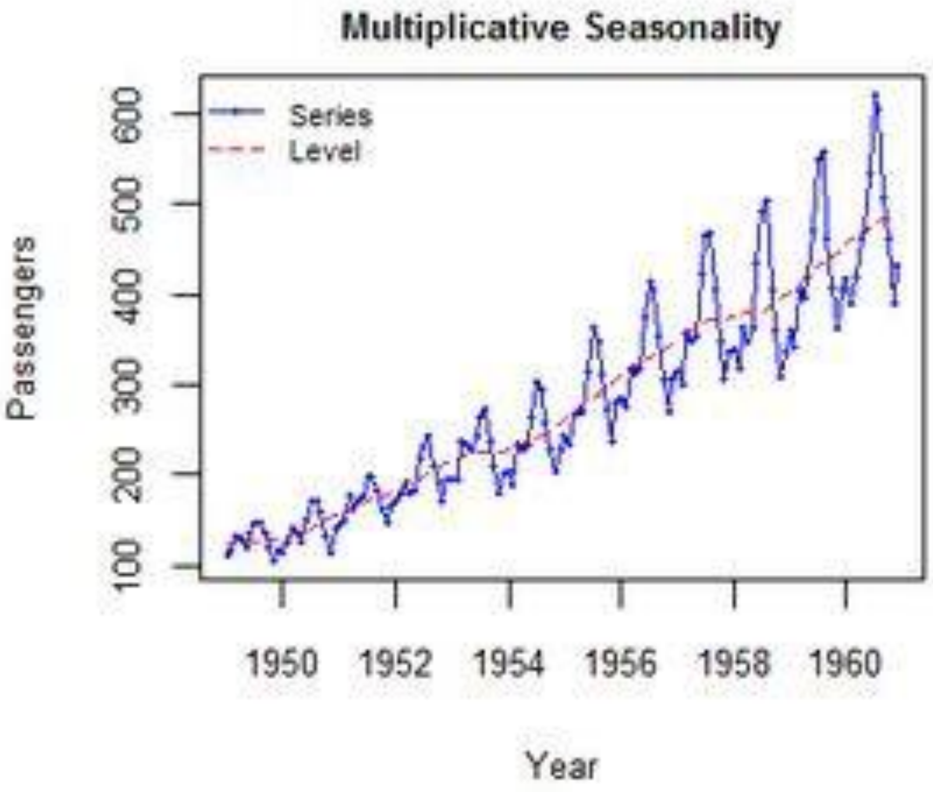


Time Series Modelling

- Time series could be modelled as an additive model or multiplicative model, depend on the nature of the trend and/or the seasonality.
- Additive model is where each observation is the sum of the components, whereas Multiplicative model is the product of the component

Time Series Modelling


Additive Seasonality

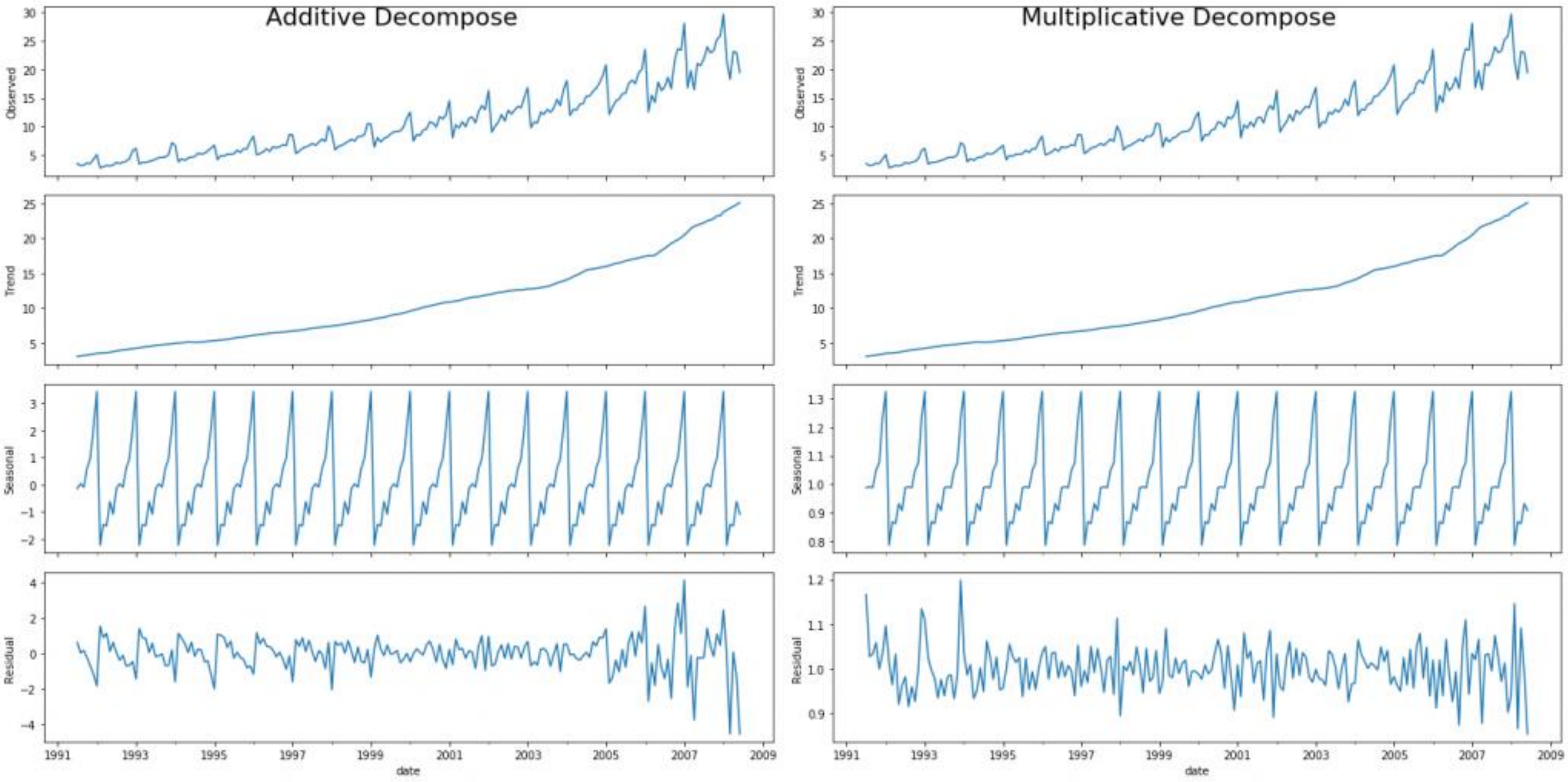

Multiplicative Seasonality

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

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

Time Series Decompose

- We can decompose Time Series to acquire the decomposed plot

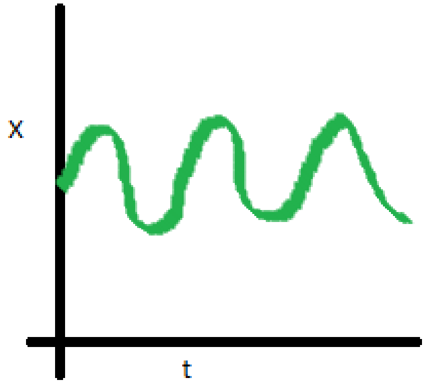


Stationary Time Series

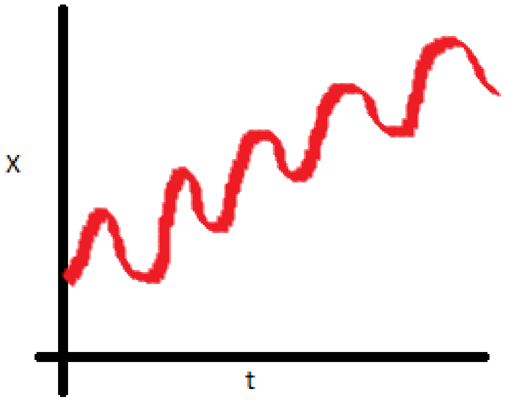
- To build a time series model, we need to have a stationary series.
- Stationary time series basically a series without function of time. It also devoid of seasonal effect
- The important reason is because most of the time series model is essentially a linear regression models that utilize the series itself as predictors.
- Linear regression works best if the predictors (X variables) are not correlated against each other. Stationarizing the series solves this problem since it removes any persistent autocorrelation, thereby making the predictors in the forecasting models nearly independent.

Stationary Time Series – Mean Constant

- The mean is constant, and not be a function of time



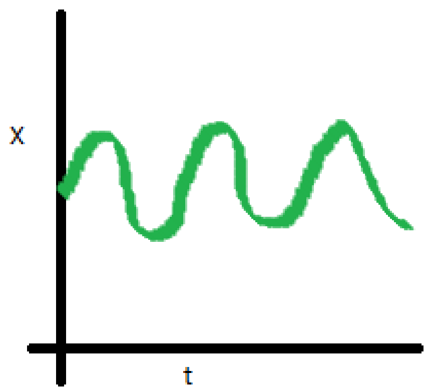
Stationary series



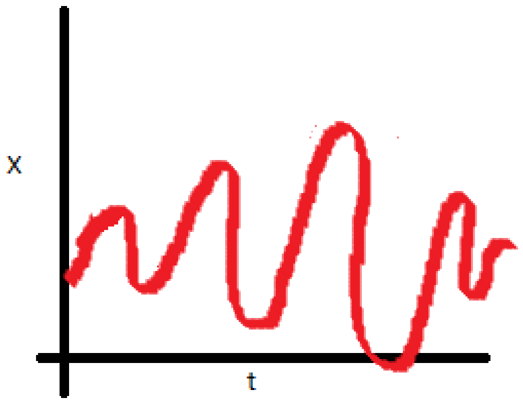
Non-Stationary series

Stationary Time Series – Variance Constant

- The variance constant , and not be a function of time (homoscedasticity)



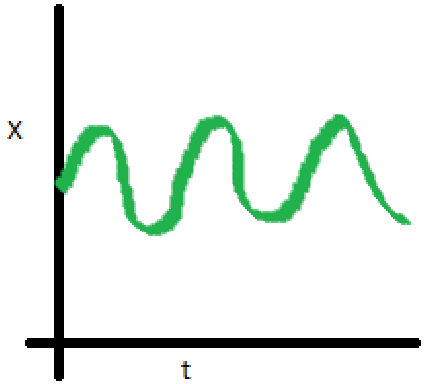
Stationary series



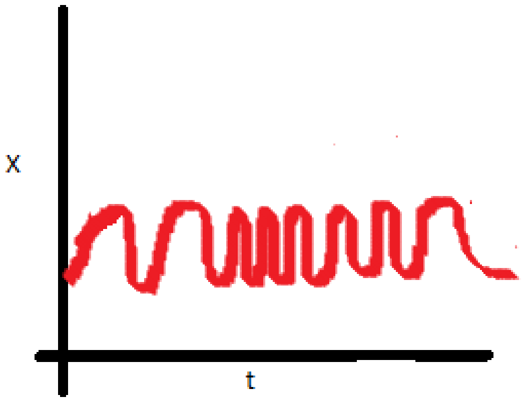
Non-Stationary series

Stationary Time Series – Covariance does not depend on time

- The covariance of the i-th term and the-(i + m)th term should not be a function of time.



Stationary series



Non-Stationary series

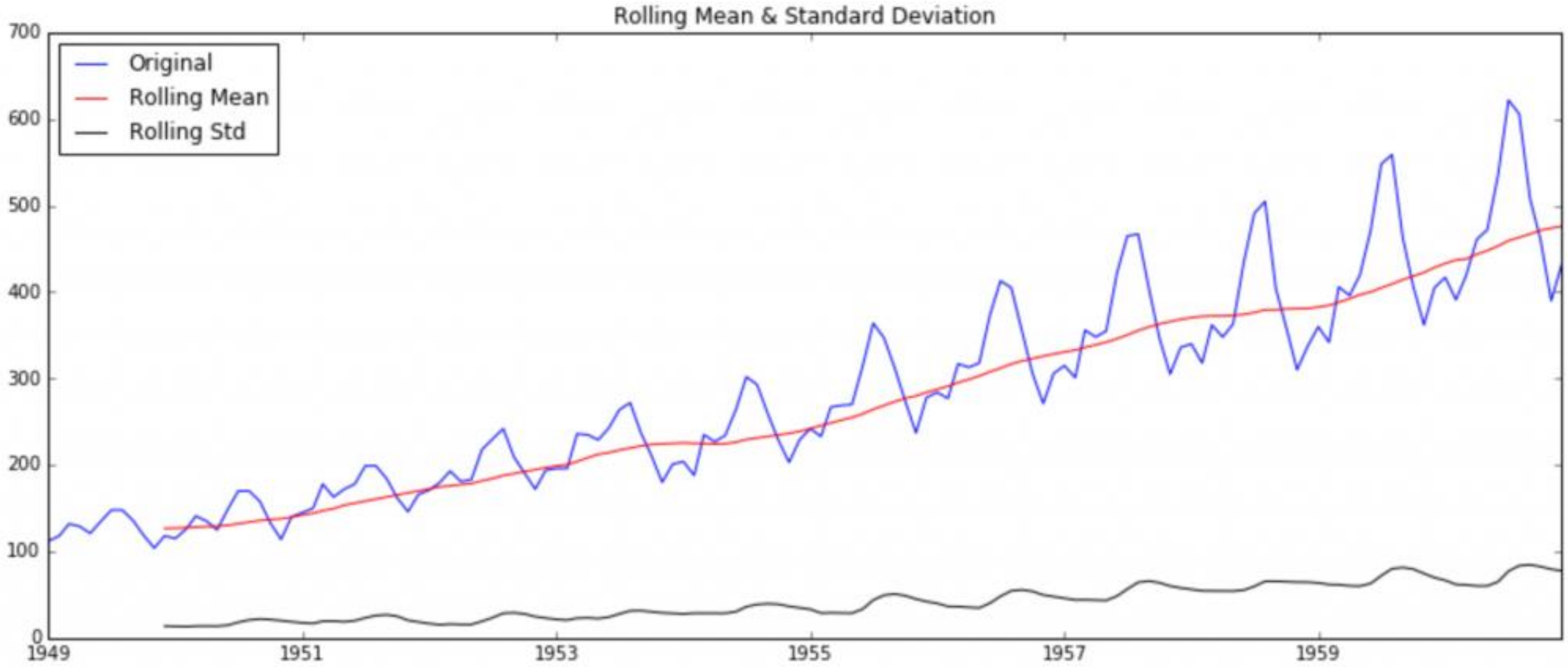
Time Series stationary validation

- Few techniques we could use to check Time Series stationary is shown below:
  - Plotting Rolling Statistics
  - Augmented Dickey-Fuller Statistic (ADF) Test

Rolling Statistic

- We can plot the moving average and/or moving variance to see if it varies within time. Moving average and moving variance are calculation of several subset of data that were calculated several time. E.g. We have data for year 2010, 2011, 2012, 2013. We could calculate 2-year moving average by getting the average of the data at 2010/2011, 2011/2012, and 2012/2013.
- In time series, using rolling statistic basically subset it by the data in the t-1 time. For example, for d
- Rolling statistic is best to be visualized, as it is a visual technique

Rolling Statistic

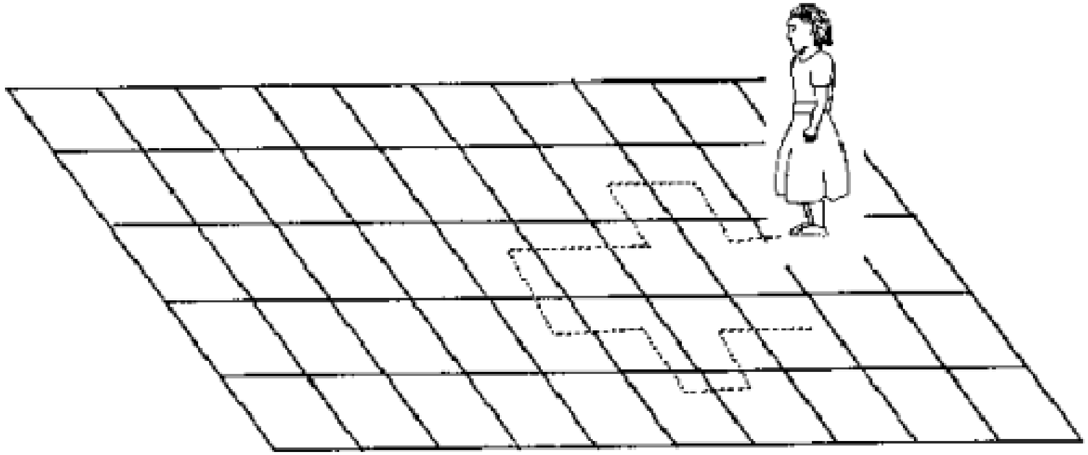


This is an example of Time series with Constant Variance but the Mean is a function of time (mean increase with time). This Time series is Non-Stationary Time Series

Random Walk

- Random walk is a mathematical random process consist of discrete fixed steps sequence with certain lengths. In each period the variable takes a random step away from its previous value, and the steps are independently and identically distributed in size

Random Walk



Imagine there is a girl walking around in the floor tiles. She step exactly to one other floor tile when she walk. This mean, her next position would only dependent on the last position.

Now, imagine we in other room and want to predict the position of the girl with time. The accuracy of our prediction would definitely become less accurate with time. At initial time (t=0) we know where the girl is, but as she take her first step; there are 8 possible tiles she could walk on hence the probability would be 1/8 at t=1. The next step she take (t =2), our probability to accurately predict where the girl now would become less accurate.

We could formulate this problem as equation:

X(t) = X(t-1) + Er(t)

If we put all the X, it would be formulated as:

X(t) = X(0) + Sum(Er(1),Er(2),Er(3).....Er(t))

Where X(t) is value of X during time t, X(0) value of X initially and Er(t) is the error during time t

Augmented Dickey-Fuller Statistic

- ADF test is a test to check whether the series is stationary or not.
- Dickey-Fuller Test is based on the random walk which involve fitting to the model to create a formula like below:

$$\mathbf{X(t) = \rho X(t\text{-}1) + Er(t)}$$

Which could be expand as:

$$\mathbf{X(t) - X(t\text{-}1) = (\rho - 1)\ X(t - 1) + Er(t)}$$

- Augmented Dickey-Fuller test fit the model even further that including the lags. The point of the test is to test if the **(ρ - 1)** is significantly different than zero or not. If the null hypothesis gets rejected, our time series is stationary time series

Create a stationary time series

- Time series need the series to be stationary, but almost all the series are not-stationary series. It is impossible to create a perfect stationary series from non-stationary series but we could create it as close as possible.
- There are 2 properties that create a series as stationary time series:
  1. Trend (Varying mean over time)
  2. Seasonality (Variation only at specific time)

- To create a stationary time series, we need to remove this 2 properties

Eliminating Trend and Seasonality

- There are few ways to eliminate trend and seasonality, but method that work with high seasonality often classified as these 2 method:
  1. Differencing – Take a difference between current value and the previous value
  2. Decomposition – Model both trend and seasonality, then remove them from the model

Time Series Forecasting

- One of the most popular technique for time series forecasting is to use statistical model **ARIMA (Auto-Regressive Integrated Moving Average)**
- ARIMA is a regression linear model that depend on the (p,d,q) parameter where p is the number of AR terms, q is the number of MA terms, and d is the number of the non-seasonal differences.
- ARIMA model that had been differenced at least once would be formulated as below:

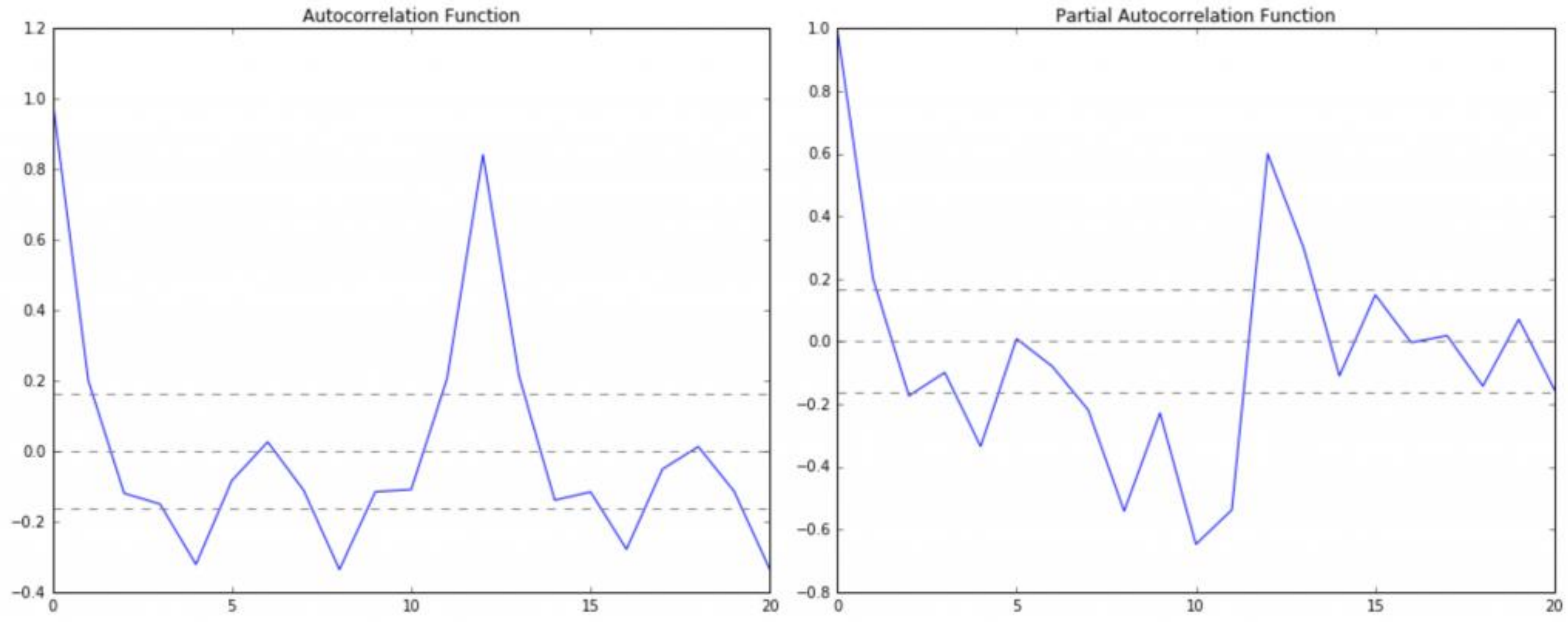
$$Y_t=\alpha+\beta_1 Y_{t-1}+\beta_2 Y_{t-2}..+\beta_p Y_{t-p}+\epsilon_t+\phi_1\epsilon_{t-1}+\phi_2\epsilon_{t-2}+..+\phi_q\epsilon_{t-q}$$

where, Y_{t-1} is the lag 1 of the series, $\beta 1$ is the coefficient of lag 1 that the model estimates, ϕ_1 is the coefficient of the error term, α is the intercept term, and ϵ_t is the error term

ARIMA parameter estimation

- An important step in ARIMA is to estimate the p and q value. To determine this value, we could use 2 different plot called ACF (Autocorellation Function) and PACF (Partial Autocorellation Function).
- ACF plot is a measured of the correlation between the time series and their own lags
- PACF plot measures the correlation between the time series with their own but after eliminating the variations such as Trend and Seasonality.

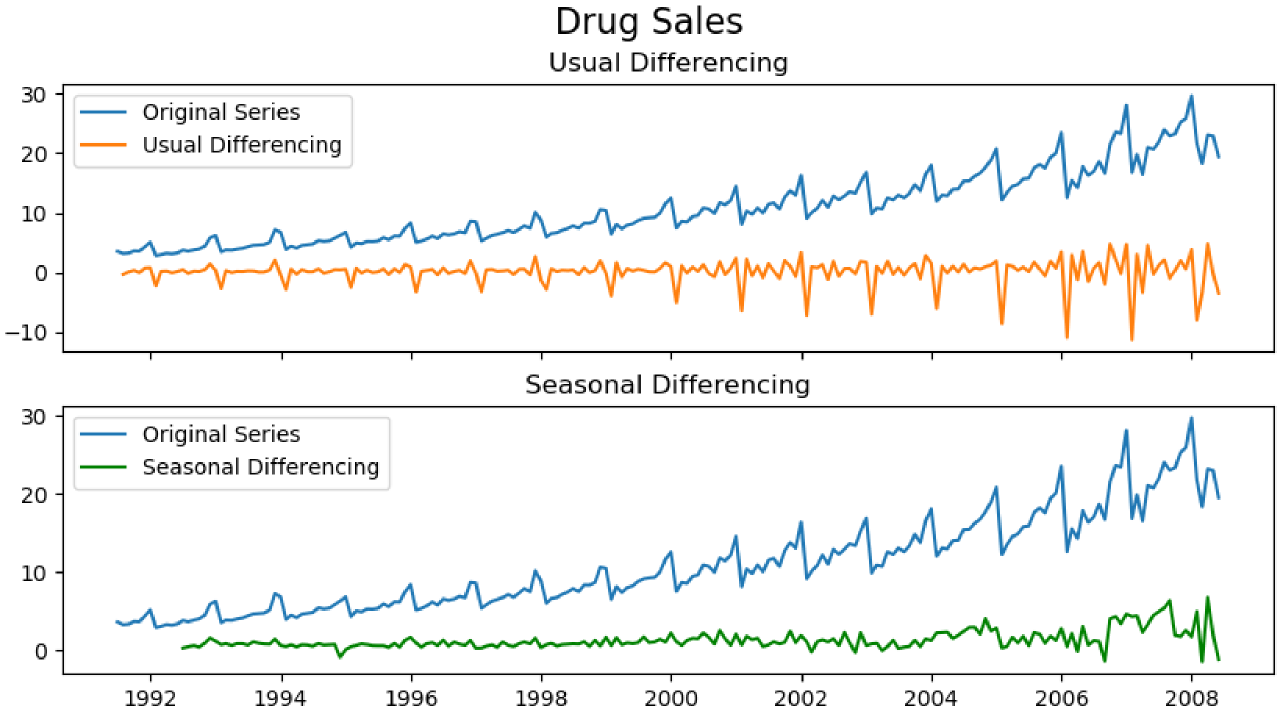
Autocorellation Function and Partial Autocorellation Function plot



In this plot, the two dotted lines on either sides of 0 are the confidence intervals. These can be used to determine the ‘p’ and ‘q’ values as:
p – The lag value where the **PACF** chart crosses the upper confidence interval for the first time. In this case p=2.
q – The lag value where the **ACF** chart crosses the upper confidence interval for the first time. In this case q=2.

Seasonal ARIMA

- The problem with ARIMA model is that it could not capture the seasonality effect from the data. We use SARIMA model to have the seasonal difference



- Instead of subtracting by the previous value, seasonal difference subtract the previous season value. From the above plot we could see the difference between regular and seasonal differencing

Seasonal ARIMA

- In SARIMA, we still need to find our usual parameter (p,d,q). As addition, we have seasonal parameter P, D, Q, x that if we represented in the model would be as SARIMA(p,d,q)x(P,D,Q,x), where, P, D and Q are Seasonal AR, order of seasonal differencing, and Seasonal MA terms respectively and 'x' is the frequency of the time series (12 for yearly, 4 for quarterly).

SARIMAX

- There are 2 type of time series model; univariate and multivariate. Univariate use their own values to predict the future values, and multivariate present another external variable other than their own value called 'Exogenous Variable'. In ARIMA class model, there is SARIMAX model that could use this exogenous variable as additional variable to help us forecast future value.