DS JULY 2022 Batch
Module 24 : Time series forecasting

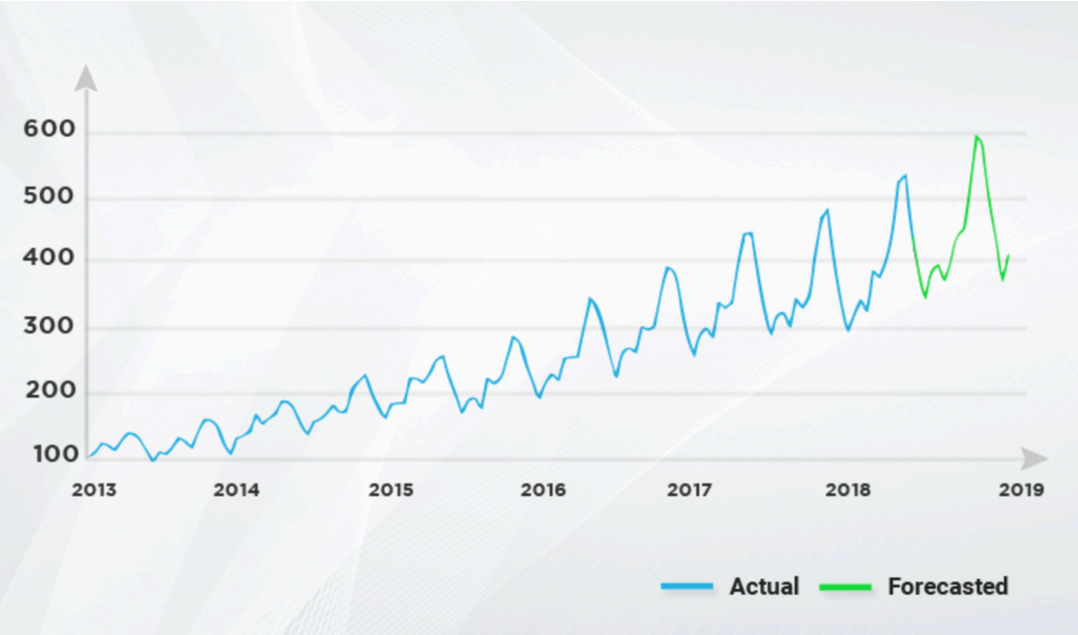
Topics

- Different Components of Time Series
- Statistical Models of time series forecasting AR MA ARMA ARIMA
- Time Series Forecasting using Statsmodel library
- Time Series Forecasting using Deep Learning

Time Series forecasting

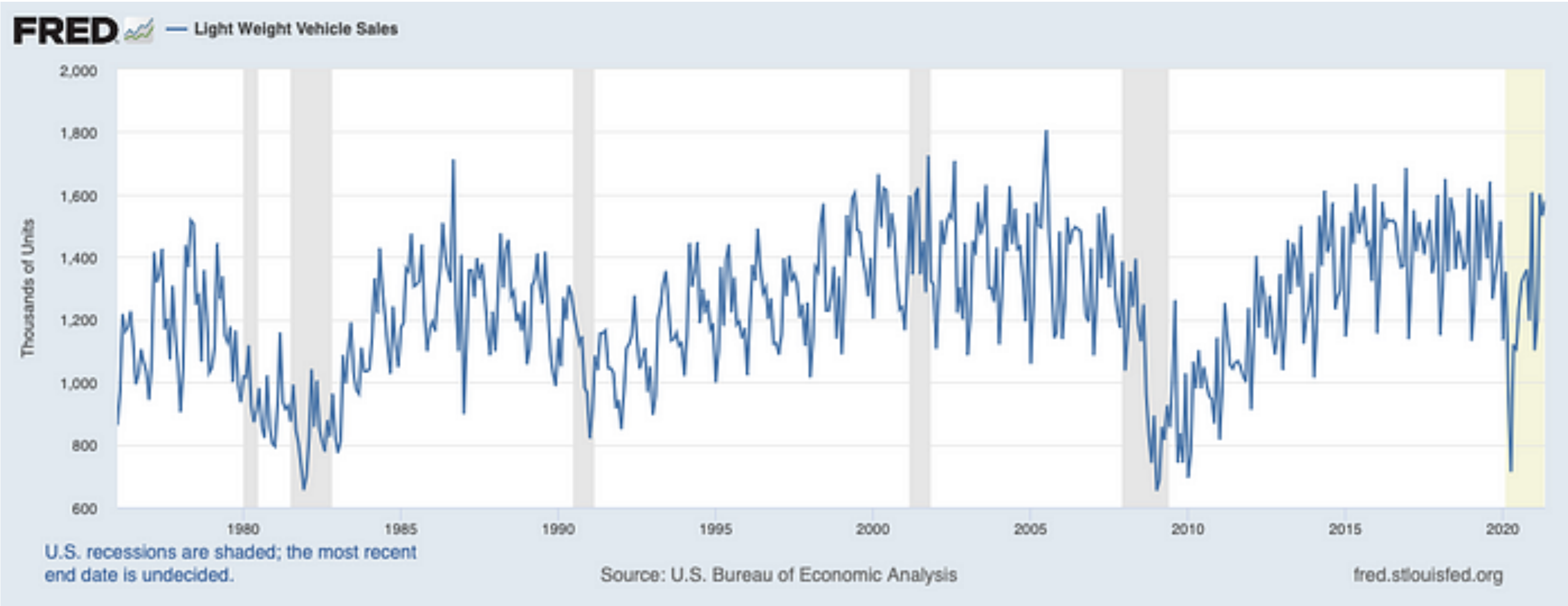
Time series

- A time series is nothing but a sequence of various data points that occurred in a successive order for a given period of time.
- Time series analysis will provide the consequences and insights of the given dataset's features that change over time.
- With the amount of data present in today's business world, it is easy to keep track of changes in patterns and trends.
- Stocks, sales, and census all have one thing in common, their data, which changes according to time, and hence, it is called time-series data.



Time series : Components

- Let's take a look at an example. The following figure shows light weight vehicle sales in thousands of units:

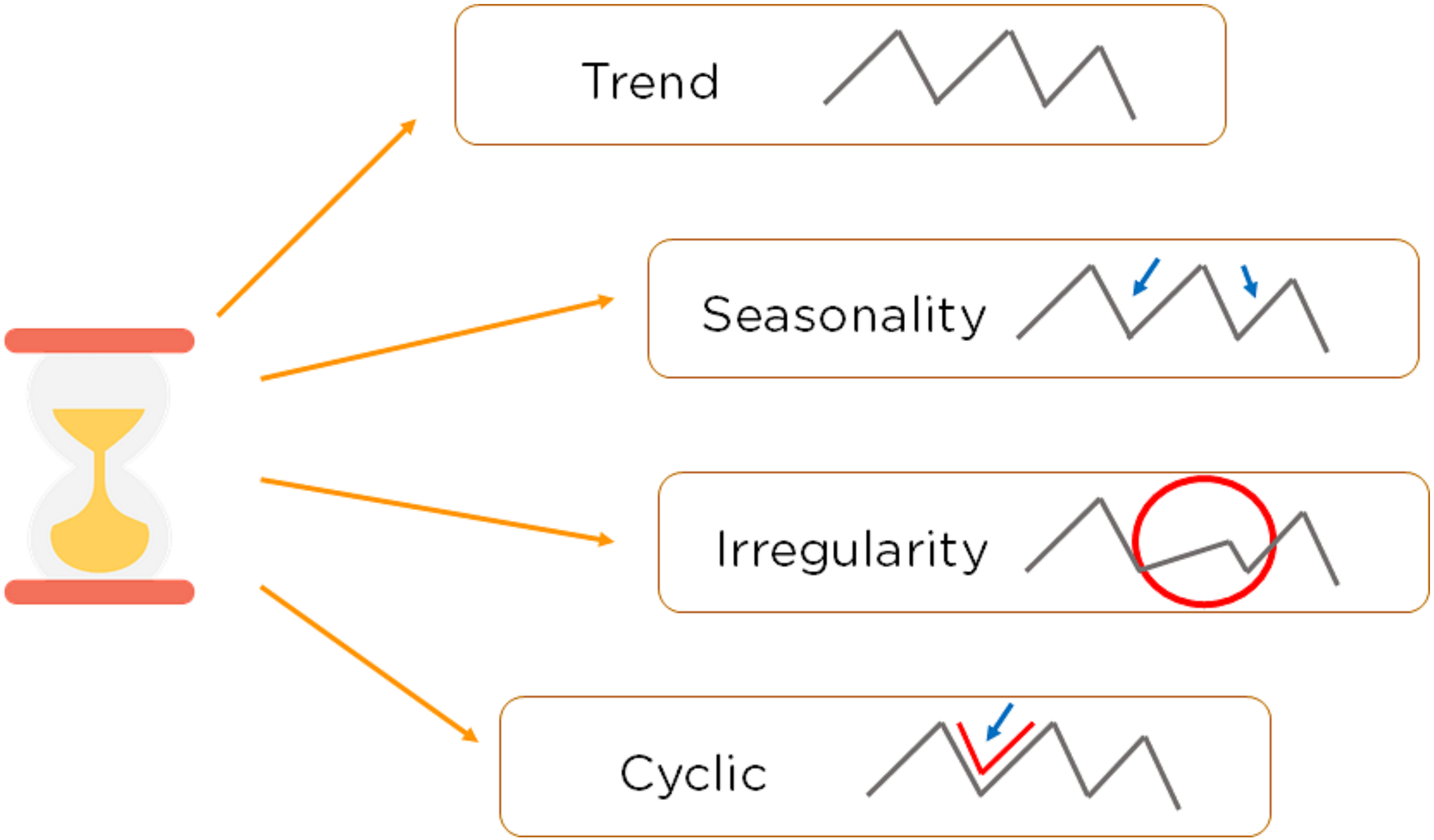


- This an excellent example of a time series. The data is collected for the last 40-something years at monthly intervals. The shaded areas represent US recessions.

Time series : Components

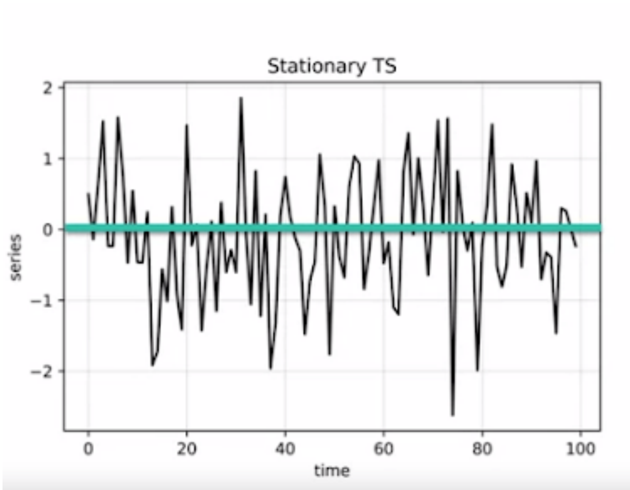
- **Trend**: A trend refers to a long-term and consistent upward or downward movement in a series. Unlike seasonal variation, a trend is unexpected and not immediately identifiable. A trend in which we can find the cause is called deterministic, while a trend that is unexplainable is called stochastic. For example, if a new author releases a book and the book skyrockets in sales, such a trend would be deterministic.
- **Cycle**: A cycle is an up and down movement that occurs around a trend. Unlike seasonal variation, a cycle does not have a precise and equal time between time periods, and is therefore not predictable.
- **Seasonality**: Unlike a trend, seasonality refers to variations that occur at a predictable and fixed frequency. For example, ice cream sales rise in the summer because the weather is warmer and more people crave a cool, sweet treat.
- **Irregularity**: Also referred to as noise, irregularity is what’s left over when you take seasonality or trends out of the dataset. Irregularities are random and unpredictable. A prime example of irregular variations are changes in stock prices.

Time series : Components



Time series : Stationary

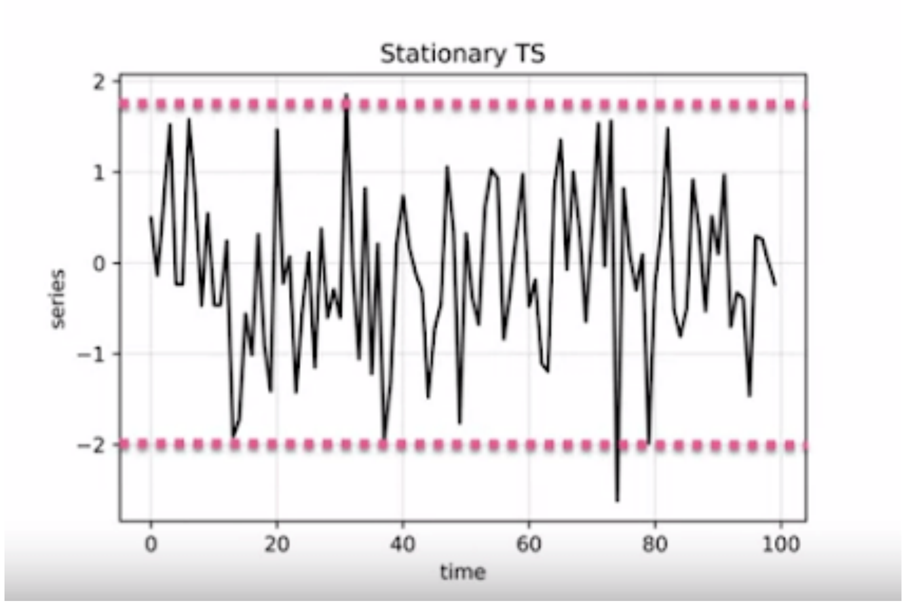
- For time series data to be stationary, the data must exhibit following properties over time:
- **Constant Mean:** A stationary time series will have a constant mean throughout the entire series.



- As an example, if we were to draw the mean of the series, this holds as the mean throughout all of the time.
- A good example where the mean wouldn't be constant is if we had some type of trend. With an upward or downward trend, for example, the mean at the end of our series would be noticeably higher or lower than the mean at the beginning of the series.

Time series : Stationary

- For time series data to be stationary, the data must exhibit following properties over time:
- **Constant Variance:** A stationary time series will have a constant variance throughout the entire series.



Time series : Stationary

- **Constant Autocorrelation :** A stationary time series will have a constant variance throughout the entire series.
- Autocorrelation simply means that the current time series measurement is correlated with a past measurement. For example, today's stock price is often highly correlated with yesterday's price.
- The time interval between correlated values is called LAG. Suppose we wanted to know if today's stock price correlated better with yesterday's price, or the price from two days ago.
- We could test this by computing the correlation between the original time series and the same series delayed by one time interval. So, the second value of the original time series would be compared with the first of the delayed. The third original value would be compared with the second of the delayed, and so on.
- Performing this process for a lag of 1 and a lag of 2, respectively, would yield two correlation outputs. This output would tell which lag is more correlated. That is autocorrelation in a nutshell.

Time Series
Statistical Model

Time series : Statistical models

- AR, MA, ARMA, and ARIMA models are used to forecast the observation at (t+1) based on the historical data of previous time spots recorded for the same observation.
- However, it is necessary to make sure that the time series is stationary over the historical data of observation overtime period. If the time series is not stationary then we could apply the differencing factor on the records and see if the graph of the time series is a stationary overtime period.
- **ACF (Auto Correlation Function)** : Auto Correlation function takes into consideration of all the past observations irrespective of its effect on the future or present time period. It calculates the correlation between the t and (t-k) time period. It includes all the lags or intervals between t and (t-k) time periods. Correlation is always calculated using the Pearson Correlation formula.
- **PACF(Partial Correlation Function)** : The PACF determines the partial correlation between time period t and t-k. It doesn’t take into consideration all the time lags between t and t-k. For e.g. let's assume that today's stock price may be dependent on 3 days prior stock price but it might not take into consideration yesterday's stock price closure. Hence we consider only the time lags having a direct impact on future time period by neglecting the insignificant time lags in between the two-time slots t and t-k.

Time series : Statistical models

- **How to differentiate when to use ACF and PACF?**
- Let's take an example of sweets sale and income generated in a village over a year. Under the assumption that every 2 months there is a festival in the village, we take out the historical data of sweets sale and income generated for 12 months. If we plot the time as month then we can observe that when it comes to calculating the sweets sale we are interested in only alternate months as the sale of sweets increases every two months. But if we are to consider the income generated next month then we have to take into consideration all the 12 months of last year.
- So in the above situation, we will use ACF to find out the income generated in the future but we will be using PACF to find out the sweets sold in the next month.

Time series : AR

- An Auto Regressive (AR) model is a type of time series model that is used to model the behaviour of a time series based on its past values.
- In an AR model, the value of the time series at any point in time is assumed to be a linear combination of its past values, and a stochastic term (error term) that captures any deviation from the predicted value.
- The order of an AR model is determined by the number of past values used to predict the current value.
- For example, an AR(1) model uses only the most recent past value to predict the current value, while an AR(p) model uses the p most recent past values.
- AR(p) models are assumed to depend on the last p values of the time series. Let's say p = 2, the forecast has the form:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \omega_t$$

Here, ω_t is the ***forecast error***, ϕ_1 and ϕ_2 are the (p=2) parameters (estimated by regression).

Time series : MA

- A Moving Average (MA) model is a type of time series model that is used to model the behavior of a time series based on its past errors.
- In an MA model, the value of the time series at any point in time is assumed to be a linear combination of its past errors (the difference between the predicted and actual values), and a stochastic term (error term) that captures any deviation from the predicted value.
- The order of an MA model is determined by the number of past errors used to predict the current value. For example, an MA(1) model uses only the most recent past error to predict the current value, while an MA(q) model uses the q most recent past errors.
- Ma(q) models are assumed to depend on the last q values of the time series. Let say q = 2, the forecast has the form:

$$X_t = \theta_2 \omega_{t-2} + \theta_1 \omega_{t-1} + \omega_t$$

ω_t is the forecast error, ω_{t-1} is the previous forecast error, etc. θ_1 and θ_2 are the (q=2) parameters.

Time series : ARMA

- An Auto Regressive Moving Average (ARMA) model is a type of time series model that combines the concepts of AR and MA models.
- In an ARMA model, the value of the time series at any point in time is assumed to be a linear combination of its past values and past errors, and a stochastic term (error term) that captures any deviation from the predicted value.
- The order of an ARMA model is determined by the number of past values and past errors used to predict the current value.
- To get our AR(p) and MA(q) models together, we combine the AR(p) and MA(P) to yield the ARMA(p,q) model. For p = 2 and q = 2 the ARMA (2,2) forecast will be:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \theta_2 \omega_{t-2} + \theta_1 \omega_{t-1} + \omega_t$$

ω_t is the forecast error, ϕ_1, ϕ_2, θ_1, and θ_2 are the ($p + q = 4$) parameters.

Time series : ARIMA

- **ARIMA (Auto Regressive Integrated Moving Average)** is a popular time series model used for forecasting and analysing time-dependent data. ARIMA combines three main components: Auto Regressive (AR), Integrated (I), and Moving Average (MA) models.
- It is a generalized form of the ARMA model that is suitable for modelling time series with non-stationary behaviour.
- **Auto Regressive (AR) component:** The AR component models the relationship between the current value of a time series and its past values. Specifically, it models the linear relationship between the current value and one or more past values of the same series. The order of the AR component is denoted by the letter "p," which represents the number of past values used in the model. The equation for an AR(p) model can be written as:
- $Yt = c + \Phi1Yt\text{-}1 + \Phi2Yt\text{-}2 + \ldots + \Phi pYt\text{-}p + \varepsilon t$
- where Yt is the value of the time series at time t, c is a constant term, Φ1, Φ2, ..., Φp are the parameters of the model, and εt is the error term at time t.

Time series : ARIMA

- **Integrated (I) component:** The I component is used to model the non-stationarity of the time series. A stationary time series has constant statistical properties such as the mean and variance over time, while a non-stationary time series has properties that change over time. The order of the I component is denoted by the letter "d," which represents the number of times the time series must be differenced to make it stationary. The equation for the I component is:
- $\Delta Yt = Yt - Yt\text{-}1$
- where ΔYt is the first difference of Yt, and Yt-1 is the lagged value of the time series at t-1.
- **Moving Average (MA) component:** The MA component models the relationship between the current value of the time series and its past errors (the difference between the predicted and actual values). The order of the MA component is denoted by the letter "q," which represents the number of past errors used in the model. The equation for an MA(q) model can be written as:
- $Yt = c + \varepsilon t + \theta 1\varepsilon t\text{-}1 + \theta 2\varepsilon t\text{-}2 + \ldots + \theta q\varepsilon t\text{-}q$
- where Yt is the value of the time series at time t, c is a constant term, $\theta 1, \theta 2, \ldots, \theta q$ are the parameters of the model, and εt is the error term at time t.

Time series : ARIMA

- The ARIMA model is denoted by ARIMA(p,d,q). It combines the AR, I, and MA components to model the behavior of the time series. The ARIMA model can be written as:
- $\Delta dYt = c + \Phi 1\Delta dYt\text{-}1 + \Phi 2\Delta dYt\text{-}2 + \ldots + \Phi p\Delta dYt\text{-}p + \varepsilon t + \theta 1\varepsilon t\text{-}1 + \theta 2\varepsilon t\text{-}2 + \ldots + \theta q\varepsilon t\text{-}q$
- where ΔdYt is the d-th difference of Yt, c is a constant term, $\Phi 1$, $\Phi 2$, …, Φp are the parameters of the AR component, $\theta 1$, $\theta 2$, …, θq are the parameters of the MA component, and εt is the error term at time t.


Thank you!