## **What is Time Series Analysis**

**Definition**:

A time series is nothing but a sequence of various data points that occurred in a successive order for a given period of time

or

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.

Or

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

Or

Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals.

Or

A time series is a sequence of observations over a certain period. The simplest example of a time series that all of us come across on a day to day basis is the change in temperature throughout the day or week or month or year.

Facts:

* Time series analysis will provide the consequences and insights of features of the given dataset that changes over time.
* Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance.
* Time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Types of data in time series:

**Time series data:** A set of observations on the values that a variable takes at different times.

**Cross-sectional data:** Data of one or more variables, collected at the same point in time.

**Pooled data:** A combination of time series data and cross-sectional data.

## **How to analyze Time Series?**

* Collecting the data and cleaning it
* Preparing Visualization with respect to time vs key feature
* Observing the stationarity of the series
* Developing charts to understand its nature.
* Model building – AR, MA, ARMA and ARIMA
* Extracting insights from prediction

## What Are the Different Components of Time Series Analysis?

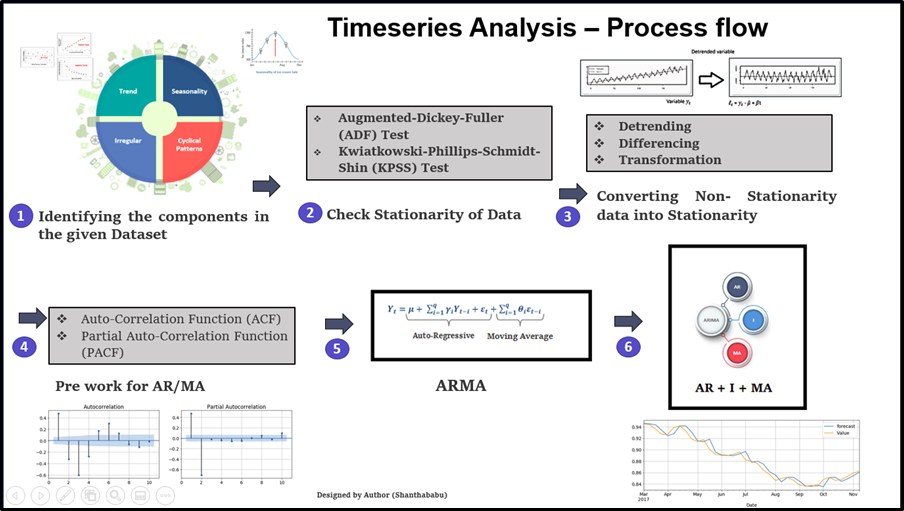
The different components of time series analysis

1. Trend: The Trend shows the variation of data with time or the frequency of data. Using a Trend, you can see how your data increases or decreases over time. The data can increase, decrease, or remain stable. Over time, population, stock market fluctuations, and production in a company are all examples of trends.
2. Seasonality: Seasonality is used to find the variations which occur at regular intervals of time. Examples are festivals, conventions, seasons, etc. These variations usually happen around the same time period and affect the data in specific ways which you can predict.
3. Irregularity: Fluctuations in the time series data do not correspond to the trend or seasonality. These variations in your time series are purely random and usually caused by unforeseeable circumstances, such as a sudden decrease in population because of a natural calamity.
4. Cyclic: Oscillations in time series which last for more than a year are called cyclic. They may or may not be periodic.
5. Stationary: A time series that has the same statistical properties over time is stationary. The properties remain the same anywhere in the series. Your data needs to be stationary to perform time-series analysis on it. A stationary series has a constant mean, variance, and covariance.

## **limitations of Time Series Analysis**

Time series has the below-mentioned limitations, we have to take care of those during our analysis,

* Similar to other models, the missing values are not supported by TSA
* The data points must be linear in their relationship.
* Data transformations are mandatory, so a little expensive.
* Models mostly work on Uni-variate data.



**Terms and concepts:**

**Dependence:** Dependence refers to the association of two observations with the same variable, at prior time points.

**Stationarity:** Shows the mean value of the series that remains constant over a time period; if past effects accumulate and the values increase toward infinity, then stationarity is not met.

**Differencing:** Used to make the series stationary, to De-trend, and to control the auto-correlations; however, some time series analyses do not require differencing and over-differenced series can produce inaccurate estimates.

**Specification:** May involve the testing of the linear or non-linear relationships of dependent variables by using models such as ARIMA, ARCH, GARCH, VAR, Co-integration, etc.

**Exponential smoothing in time series analysis:** This method predicts the one next period value based on the past and current value.  It involves averaging of data such that the nonsystematic components of each individual case or observation cancel out each other.  The exponential smoothing method is used to predict the short term predication.  Alpha, Gamma, Phi, and Delta are the parameters that estimate the effect of the time series data.  Alpha is used when seasonality is not present in data.  Gamma is used when a series has a trend in data.  Delta is used when seasonality cycles are present in data.  A model is applied according to the pattern of the data.

**Curve fitting in time series analysis:** Curve fitting regression is used when data is in a non-linear relationship. The following equation shows the non-linear behavior :

Dependent variable, where case is the sequential case number.

Curve fitting can be performed by selecting “regression” from the analysis menu and then selecting “curve estimation” from the regression option. Then select “wanted curve linear,” “power,” “quadratic,” “cubic,” “inverse,” “logistic,” “exponential,” or “other.”

**Assumptions:**

**Stationarity:** The first assumption is that the series are stationary.  Essentially, this means that the series are normally distributed and the mean and variance are constant over a long time period.

**Uncorrelated random error:** We assume that the error term is randomly distributed and the mean and variance are constant over a time period.  The Durbin-Watson test is the standard test for correlated errors.

**No outliers:** We assume that there is no outlier in the series.  Outliers may affect conclusions strongly and can be misleading.

**Random shocks (a random error component):** If shocks are present, they are assumed to be randomly distributed with a mean of 0 and a constant variance.

**ARIMA:**

ARIMA stands for autoregressive integrated moving average.  This method is also known as the Box-Jenkins method.

**Identification of ARIMA parameters:**

**Autoregressive component:** AR stands for autoregressive.  Autoregressive paratmeter is denoted by p.  When p =0, it means that there is no auto-correlation in the series.  When p=1, it means that the series auto-correlation is till one lag.

**Integrated:** In ARIMA time series analysis, integrated is denoted by d.  Integration is the inverse of differencing. When d=0, it means the series is stationary and we do not need to take the difference of it.  When d=1, it means that the series is not stationary and to make it stationary, we need to take the first difference.  When d=2, it means that the series has been differenced twice.  Usually, more than two time difference is not reliable.

**Moving average component:** MA stands for moving the average, which is denoted by q.  In ARIMA, moving average q=1 means that it is an error term and there is auto-correlation with one lag.

In order to test whether or not the series and their error term is auto correlated, we usually use W-D test, ACF, and PACF.

**Decomposition:** Refers to separating a time series into trend, seasonal effects, and remaining variability

## **Types of time series analysis**

Even within time series analysis, there are different types and models of analysis that will achieve different results.

* Classification: Identifies and assigns categories to the data.
* Curve fitting: Plots the data along a curve to study the relationships of variables within the data.
* Descriptive analysis: Identifies patterns in time series data, like trends, cycles, or seasonal variation.
* Explanative analysis: Attempts to understand the data and the relationships within it, as well as cause and effect.
* Exploratory analysis: Highlights the main characteristics of the time series data, usually in a visual format.
* Forecasting: Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points.
* Intervention analysis: Studies how an event can change the data.
* Segmentation: Splits the data into segments to show the underlying properties of the source information.

## **Classification and considerations**

While time series data is data collected over time, there are different types of data that describe how and when that time data was recorded. For example:

* Time series data is data that is recorded over consistent intervals of time.
* Cross-sectional data consists of several variables recorded at the same time.
* Pooled data is a combination of both time series data and cross-sectional data.

Further, time series data can be classified into two main categories:

* Stock time series data means measuring attributes at a certain point in time, like a static snapshot of the information as it was.
* Flow time series data means measuring the activity of the attributes over a certain period, which is generally part of the total whole and makes up a portion of the results.

In time series data, variations can occur sporadically throughout the data:

* Functional analysis can pick out the patterns and relationships within the data to identify notable events.
* Trend analysis means determining consistent movement in a certain direction. There are two types of trends: deterministic, where we can find the underlying cause, and stochastic, which is random and unexplainable.
* Seasonal variation describes events that occur at specific and regular intervals during the course of a year. Serial dependence occurs when data points close together in time tend to be related.

Time series analysis and forecasting models must define the types of data relevant to answering the business question. Once analysts have chosen the relevant data they want to analyze, they choose what types of analysis and techniques are the best fit.

## **Significance of Time Series and its types**

TSA is the backbone for prediction and forecasting analysis, specific to the time-based problem statements.

* Analyzing the historical dataset and its patterns
* Understanding and matching the current situation with patterns derived from the previous stage.
* Understanding the factor or factors influencing certain variable(s) in different periods.

With help of “Time Series” we can prepare numerous time-based analyses and results.

* Forecasting
* Segmentation
* Classification
* Descriptive analysis
* Intervention analysis

## When time series analysis is used and when it isn’t

Time series analysis is not a new study, despite technology making it easier to access. Many of the recommended texts teaching the subject’s fundamental theories and practices have been around for several decades. And the method itself is even older than that. We have been using time series analysis for thousands of years, all the way back to the ancient studies of planetary movement and navigation. Time series analysis is used for non-stationary data—things that are constantly fluctuating over time or are affected by time. Industries like finance, retail, and economics frequently use time series analysis because currency and sales are always changing. Stock market analysis is an excellent example of time series analysis in action, especially with automated trading algorithms. Likewise, time series analysis is ideal for forecasting weather changes, helping meteorologists predict everything from tomorrow’s weather report to future years of climate change. Examples of time series analysis in action include:

* Weather data
* Rainfall measurements
* Temperature readings
* Heart rate monitoring (EKG)
* Brain monitoring (EEG)
* Quarterly sales
* Stock prices
* Automated stock trading
* Industry forecasts
* Interest rates

Because time series analysis includes many categories or variations of data, analysts sometimes must make complex models. However, analysts can’t account for all variances, and they can’t generalize a specific model to every sample. Models that are too complex or that try to do too many things can lead to lack of fit. Lack of fit or overfitting models lead to those models not distinguishing between random error and true relationships, leaving analysis skewed and forecasts incorrect.

## **Data Types of Time Series**

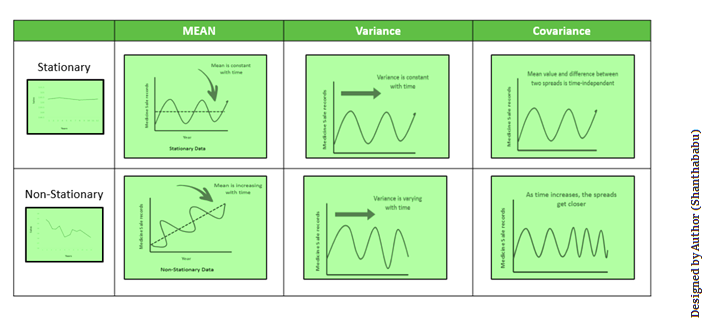
Let’s discuss the time series’ data types and their influence. While discussing TS data-types, there are two major types.

* Stationary
* Non- Stationary

**Stationary**: A dataset should follow the below thumb rules, without having Trend, Seasonality, Cyclical, and Irregularity component of time series

* The MEAN value of them should be completely constant in the data during the analysis
* The VARIANCE should be constant with respect to the time-frame
* The COVARIANCE measures the relationship between two variables.

**Non- Stationar**y: This is just the opposite of Stationary.



## **Methods to check Stationarity**

During the TSA model preparation workflow, we must access if the given dataset is Stationary or NOT. Using **Statistical and Plots test.**

**Statistical Test:** There are two tests available to test if the dataset is Stationary or NOT.

* Augmented Dickey-Fuller (ADF) Test
* Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

**Augmented Dickey-Fuller (ADF)** Test or Unit Root Test: The ADF test is the most popular statistical test and with the following assumptions.

* Null Hypothesis (H0): Series is non-stationary
* Alternate Hypothesis (HA): Series is stationary
  + p-value >0.05 Fail to reject (H0)
  + p-value <= 0.05 Accept (H1)

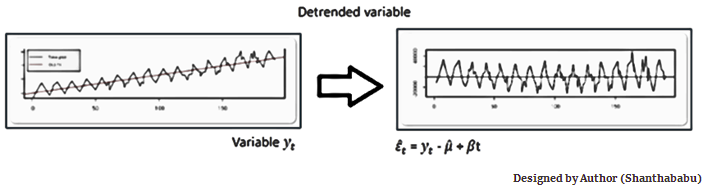
**Kwiatkowski–Phillips–Schmidt–Shin (KPSS):** these tests are used for testing a NULL Hypothesis (HO), that will perceive the time-series, as stationary around a deterministic trend against the alternative of a unit root. Since TSA looking for Stationary Data for its further analysis, we have to make sure that the dataset should be stationary.

## **Converting Non- stationary into stationary**

Let’s discuss quickly how to convert Non- stationary into stationary for effective time series modeling. There are two major methods available for this conversion.

* Detrending
* Differencing
* Transformation

**1. Detrending:** It involves removing the trend effects from the given dataset and showing only the differences in values from the trend. it always allows the cyclical patterns to be identified.

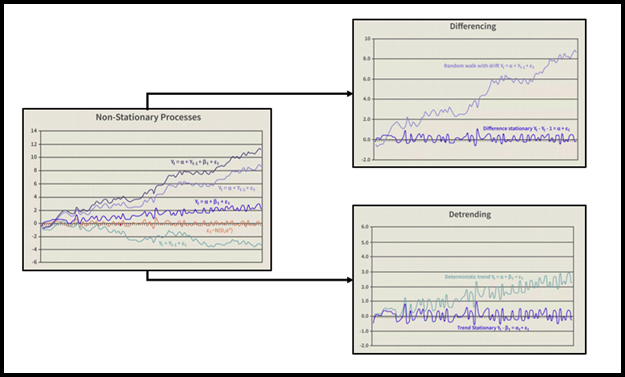


**2. Differencing:** This is a simple transformation of the series into a new time series, which we use to remove the series dependence on time and stabilize the mean of the time series, so trend and seasonality are reduced during this transformation.

Yt= Yt – Yt-1

Yt=Value with time

**Detrending and Differencing extractions**



**3. Transformation:** This includes three different methods they are Power Transform, Square Root, and Log Transfer., most commonly used one is Log Transfer.

## **Moving Average Methodology**

The commonly used time series method is Moving Average. This method is slick with random short-term variations. Relatively associated with the components of time series.

**The Moving Average (MA) (Or) Rolling Mean:** In which MA has calculated by taking averaging data of the time-series, within k periods.

Let’s see the types of moving averages:

* Simple Moving Average (SMA),
* Cumulative Moving Average (CMA)
* Exponential Moving Average (EMA)

**Simple Moving Average (SMA)**

The SMA is the unweighted mean of the previous M or N points. The selection of sliding window data points, depending on the amount of smoothing is preferred since increasing the value of M or N, improves the smoothing at the expense of accuracy.

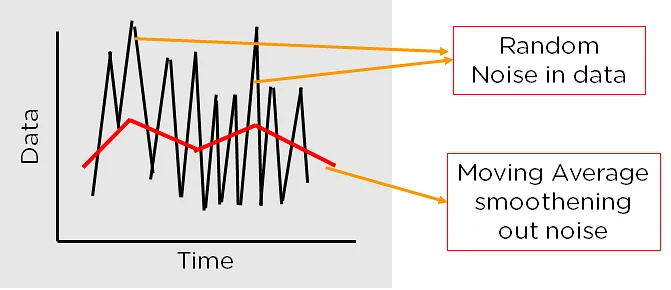
**Cumulative Moving Average (CMA)**

The CMA is the unweighted mean of past values, till the current time.

**Exponential Moving Average (EMA)**

EMA is mainly used to identify trends and to filter out noise. The weight of elements is decreased gradually over time. This means It gives weight to recent data points, not historical ones. Compared with SMA, the EMA is faster to change and more sensitive.

* It has a value between 0,1.
* Represents the weighting applied to the very recent period.



## **Time Series Analysis in Data Science and Machine Learning**

When dealing with TSA in Data Science and Machine Learning, there are multiple model options are available. In which the Autoregressive–Moving-Average (ARMA) models with [p, d, and q].

* P==> autoregressive lags
* q== moving average lags
* d==> difference in the order

Before we get to know about Arima, first you should understand the below terms better.

* Auto-Correlation Function (ACF)
* Partial Auto-Correlation Function (PACF)

**1. Auto-Correlation Function (ACF)**: ACF is used to indicate and how similar a value is within a given time series and the previous value. (OR) It measures the degree of the similarity between a given time series and the lagged version of that time series at different intervals that we observed.

This is used to identify a set of trends in the given dataset and the influence of former observed values on the currently observed values.

**2. Partial Auto-Correlation (PACF):** PACF is similar to Auto-Correlation Function and is a little challenging to understand. It always shows the correlation of the sequence with itself with some number of time units per sequence order in which only the direct effect has been shown, and all other intermediary effects are removed from the given time series.

**Interpret ACF and PACF plots**

|  |  |  |
| --- | --- | --- |
| **ACF** | **PACF** | **Perfect ML -Model** |
| Plot declines gradually | Plot drops instantly | Auto Regressive model. |
| Plot drops instantly | Plot declines gradually | Moving Average model |
| Plot decline gradually | Plot Decline gradually | ARMA |
| Plot drop instantly | Plot drop instantly | You wouldn’t perform any model |

Remember that both ACF and PACF require stationary time series for analysis.

**Auto-Regressive model**

This is a simple model, that predicts future performance based on past performance. mainly used for forecasting, when there is some correlation between values in a given time series and the values that precede and succeed (back and forth).

Auto-Regressive models predict future behavior using past behavior where there is some correlation between past and future data. The formula below represents the autoregressive model. It is a modified version of the slope formula with the target value being expressed as the sum of the intercept, the product of a coefficient and the previous output, and an error correction term.

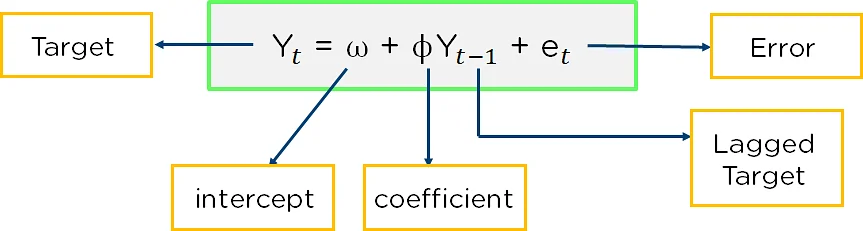
The equation for the AR model (Let’s compare Y=mX+c)

**Yt =C+b1 Yt-1+ b2 Yt-2+……+ bp Yt-p+ Ert**

**Key Parameters**

* p=past values
* Yt=Function of different past values
* Ert=errors in time
* C=intercept = w (in below equation)
* b1, b2 , … bp = coefficient

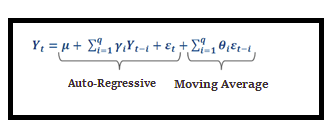
The equation can be written as



### Integration

Integration is the difference between present and previous observations. It is used to make the time series stationary.

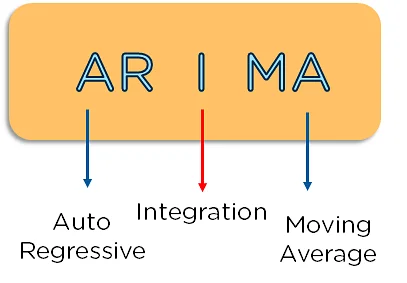
**ARMA** This is a combination of the Auto-Regressive and Moving Average model for forecasting. This model provides a weakly stationary stochastic process in terms of two polynomials, one for the Auto-Regressive and the second for the Moving Average. It is best for predicting stationary series.



**ARIMA** supports stationary as well as non-stationary.

* AR ==> Uses the past values to predict the future
* MA ==> Uses the past error terms in the given series to predict the future
* I==> uses the differencing of observation and makes the stationary data

**AR+I+MA= ARIMA**

Understand the Signature of ARIMA

* p==> log order => No of lag observations.
* d==> degree of differencing => No of times that the raw observations are differenced.
* q==>order of moving average => the size of the moving average window

## **Implementation steps for ARIMA**

Step 1: Plot a time series format

Step 2: Difference to make stationary on mean by removing the trend

Step 3: Make stationary by applying log transform.

Step 4: Difference log transform to make as stationary on both statistic mean and variance

Step 5: Plot ACF & PACF, and identify the potential AR and MA model

Step 6: Discovery of best fit ARIMA model

Step 7: Forecast/Predict the value, using the best fit ARIMA model

Step 8: Plot ACF & PACF for residuals of the ARIMA model, and ensure no more information is left.

**Some Q/As**

**WHAT ARE SEASONAL EFFECTS?**  
A seasonal effect is a systematic and calendar related effect. Some examples include the sharp escalation in most Retail series which occurs around December in response to the Christmas period, or an increase in water consumption in summer due to warmer weather. Other seasonal effects include trading day effects (the number of working or trading days in a given month differs from year to year which will impact upon the level of activity in that month) and moving holiday (the timing of holidays such as Easter varies, so the effects of the holiday will be experienced in different periods each year).  
**WHAT IS SEASONAL ADJUSTMENT AND WHY DO WE NEED IT?**  
Seasonal adjustment is the process of estimating and then removing from a time series influences that are systematic and calendar related. Observed data needs to be seasonally adjusted as seasonal effects can conceal both the true underlying movement in the series, as well as certain non-seasonal characteristics which may be of interest to analysts.  
  
  
**WHY CAN'T WE JUST COMPARE ORIGINAL DATA FROM THE SAME PERIOD IN EACH YEAR?**  
A comparison of original data from the same period in each year does not completely remove all seasonal effects. Certain holidays such as Easter and Chinese New Year fall in different periods in each year, hence they will distort observations. Also, year to year values will be biased by any changes in seasonal patterns that occur over time. For example, consider a comparison between two consecutive March months i.e. compare the level of the original series observed in March for 2000 and 2001. This comparison ignores the moving holiday effect of Easter. Easter occurs in April for most years but if Easter falls in March, the level of activity can vary greatly for that month for some series. This distorts the original estimates. A comparison of these two months will not reflect the underlying pattern of the data. The comparison also ignores trading day effects. If the two consecutive months of March have different composition of trading days, it might reflect different levels of activity in original terms even though the underlying level of activity is unchanged. In a similar way, any changes to seasonal patterns might also be ignored. The original estimates also contains the influence of the irregular component. If the magnitude of the irregular component of a series is strong compared with the magnitude of the trend component, the underlying direction of the series can be distorted.   
  
However, the major disadvantage of comparing year to year original data, is lack of precision and time delays in the identification of turning points in a series. Turning points occur when the direction of underlying level of the series changes, for example when a consistently decreasing series begins to rise steadily. If we compare year apart data in the original series, we may miss turning points occurring during the year. For example, if March 2001 has a higher original estimate than March 2000, by comparing these year apart values, we might conclude that the level of activity has increased during the year. However, the series might have increased up to September 2000 and then started to decrease steadily.   
**WHEN IS SEASONAL ADJUSTMENT INAPPROPRIATE?**  
When a time series is dominated by the trend or irregular components, it is nearly impossible to identify and remove what little seasonality is present. Hence seasonally adjusting a non-seasonal series is impractical and will often introduce an artificial seasonal element.  
  
**WHAT IS SEASONALITY?**  
The seasonal component consists of effects that are reasonably stable with respect to timing, direction and magnitude. It arises from systematic, calendar related influences such as:

* **Natural Conditions**

weather fluctuations that are representative of the season   
(uncharacteristic weather patterns such as snow in summer would be considered irregular influences)

* **Business and Administrative procedures**

start and end of the school term

* **Social and Cultural behaviour**

Christmas

It also includes calendar related systematic effects that are not stable in their annual timing or are caused by variations in the calendar from year to year, such as:

* **Trading Day Effects**

the number of occurrences of each of the day of the week in a given month will differ from year to year  
- There were 4 weekends in March in 2000, but 5 weekends in March of 2002

* **Moving Holiday Effects**

holidays which occur each year, but whose exact timing shifts  
- Easter, Chinese New Year

**HOW DO WE IDENTIFY SEASONALITY?**  
Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend.

**WHAT IS AN IRREGULAR?**  
The irregular component (sometimes also known as the residual) is what remains after the seasonal and trend components of a time series have been estimated and removed. It results from short term fluctuations in the series which are neither systematic nor predictable. In a highly irregular series, these fluctuations can dominate movements, which will mask the trend and seasonality.

**WHAT IS THE TREND?**  
The ABS trend is defined as the 'long term' movement in a time series without calendar related and irregular effects, and is a reflection of the underlying level. It is the result of influences such as population growth, price inflation and general economic changes.