# **Forecasting**

* Forecasting is the process of analyzing historical data to predict future values. It is a supervised learning problem.
* It is more similar to regression since we try to predict real-valued numbers.
* Given the historic values of a feature **x** till the present time **t** (x**1**, x**2**, x**3** … x**t** ), we need to predict the value of this feature for the future (x**t+1**, x**t+2**, ...).
* In forecasting, Unlike regression, we do not have a set of inputs and one output. Rather, we have a signal and we are looking at some past values to predict some future value.
* Every business operates under risk and uncertainty. The forecast is necessary to lessen the adverse effects of the risks.

# **Time series data**

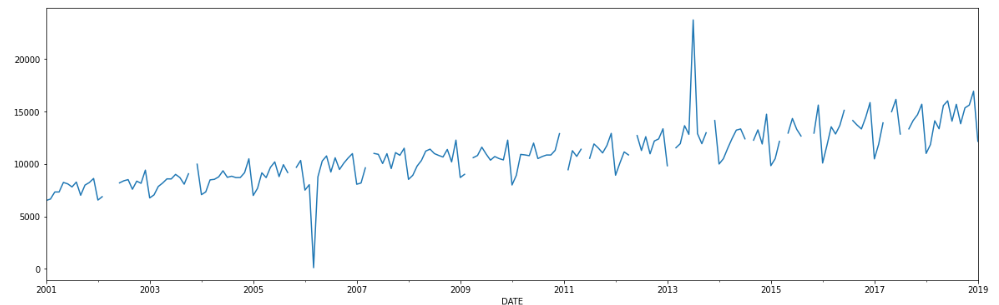
* A signal, indexed by an ordered timestamp is time-series data.
* A time series is a sequence of measurements on the same variable collected over time.
* It is a set of observations, each one being recorded at equally spaced time intervals.
* For our data to be a time series, we need a minimum of two things
  + Date/timestamp (denoted as **t**)
  + One quantity (denoted as **y**)
* The timestamp(t) can be in days, weeks, months, years, or even seconds.
* Data observations like sales, revenues, inventories, etc are commonly expressed as a time series.
* There are two types of time series present:
  + - Univariate time series - Data collected only for one variable over a period of time.
    - Multivariate time series - Data collected only for more than one variable over the same period of time.
* Time series data are not independent.
* Since there is dependency, the ordering of the time series data is most important. Changing the order will change the data structure.

# **Handling missing values**

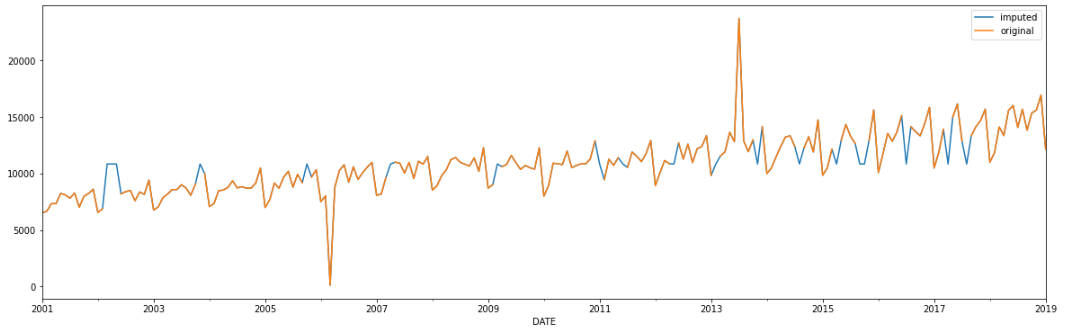
* Time series data does not admit missing values.
* All the observations in the data must be contiguous.
* There are many ways to impute the missing data:
  + - Mean / Median imputation
    - Interpolation
    - Moving Averages

**Imputing using mean/median:**

Let’s consider the sales dataset:



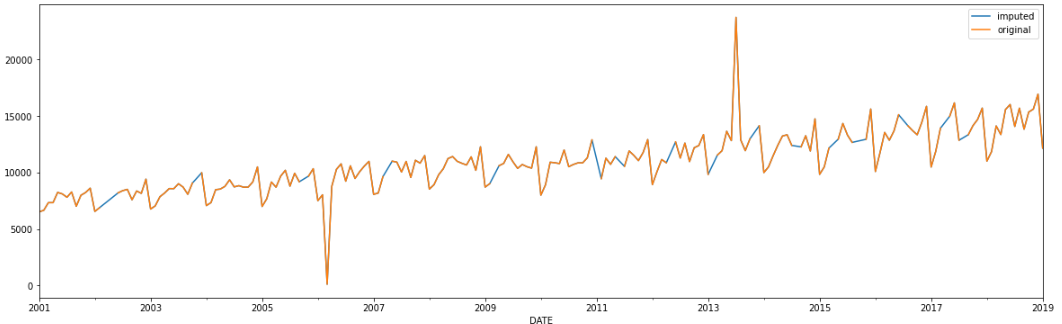
* One basic intuition is to fill these missing values with the **mean** or the **median** of the signal. We will get the following plot after the imputation:



* We can clearly observe that there are sharp increments and sharp decrements at some of the imputed places. This is because we imputed the local missing values with the mean of the entire data.
* Replacing with median would've also given the same result, as they have a similar value.

**Linear Interpolation**

* It is a technique for handling the missing values. We take the average of the first point before and the first point after the missing value and fill the missing value with this average.
* Interpolation is an estimation of a value within two known values in a sequence of values.
* It ensures that the missing values are not under or over-estimated by taking the average of the entire data.
* After using the linear interpolation for the imputation of missing values in the above data, we get the following plot:



* The imputed values fall within two finite points(preceding and succeeding points) because of which the imputed values don't seem to be forced.

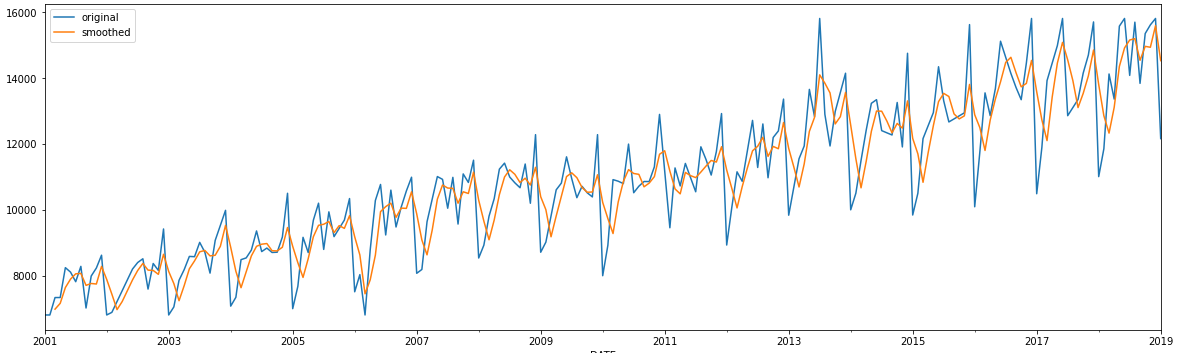
**Moving Averages**

* If we take the average of the last *k* data points in our series and use it to guess the next point at *t=k*. This approach is called the Moving Average.

i.e.

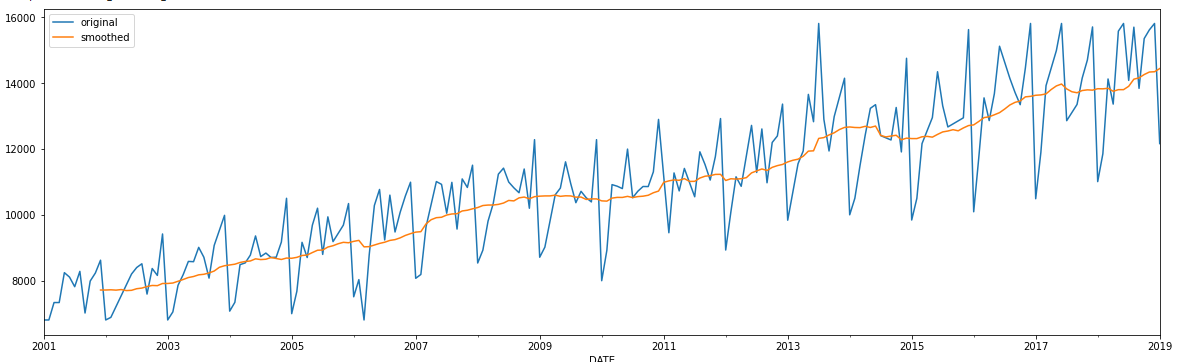


* The value of k acts as a hyperparameter, which we can set based on what works best for us. It is also called the **window size**.
* The moving average line has been shown below.



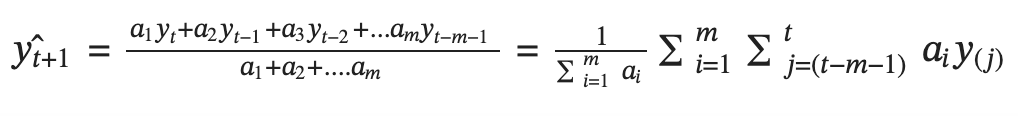
**Moving averages with k=3**

* If we increase the window, the moving average curve gets smoother.



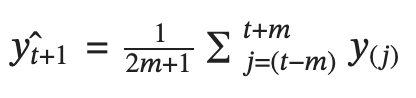
**Moving averages with k=12**

* Another important property of moving averages is that, when the time series goes up, the moving average also goes up and when the time series go down the moving average also goes down.
* Instead of giving all the previous observations equal weights, we can assign weights to the observations
* This is called **the weighted moving average.**

****

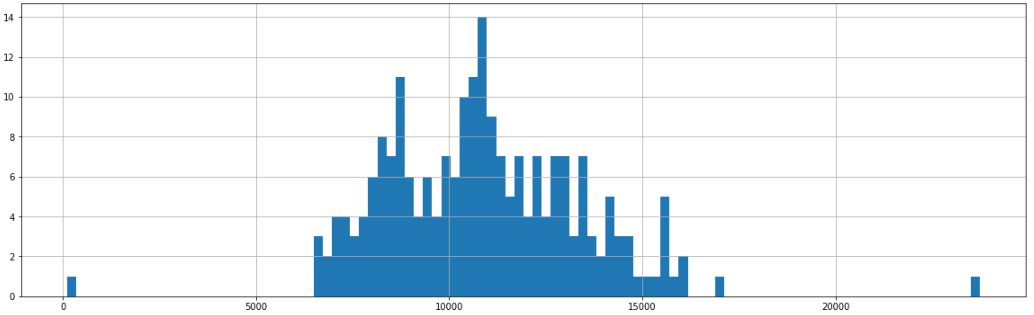
**Centered moving average**

* The average of n points before the current point, and n points after the current point are called the Centered moving average.
* We can't use this in forecasting as future values are not available.
* This can be used to deal with the missing values and also deals with the anomalies present in the time series data.

****

# **Handling Anomalies**

* An anomaly/outlier is an abnormal or unusual data point in the data set, which stands out of the data. It can simply be seen as a wrong entry in the data, and that can happen at times.
* Alternatively, it could be the correct entry but for a one-time event, which is not likely to repeat in the future. So, even if it's valid data, we would like to remove it from our training set, because we don't want our model to get biased by that one-time event that is not going to happen again.
* Methods to remove anomalies:
  + - Replace it with a fixed number
    - Using quantiles, which we will see now
    - Robust scaling is also a method that can be used.
* One of the best ways to identify the anomalies is by plotting a **Histogram**.



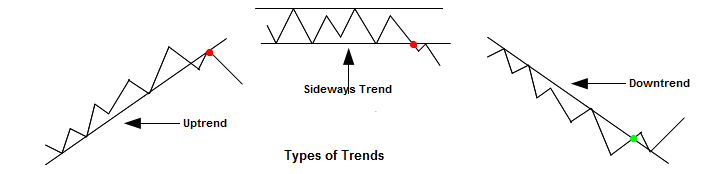
* From the plot, we can clearly see the anomalies and we can easily decide where to cut the data, in order to get rid of anomalies.
* If the above histogram plot is more continuous, we use the concept of **quantiles** to deal with anomalies. We rule out observation as an anomaly if it is greater than 95 percentile, or less than 5 percentile.

# **Time series decomposition**

* Graphs highlight the variety of patterns/ features in the time series.
* A time series can be split into several components, each representing one of the underlying categories of patterns.
* There are three important components which we want to decompose our signal into.
  + Trend - general movement over time
  + Seasonality - behaviors captured in individual seasonal periods
  + Residual - everything not captured by trend and seasonal components
* Trend, Seasonality are the systematic components of the time series data whereas Residual/ Error is an irregular component.

## **Trend**

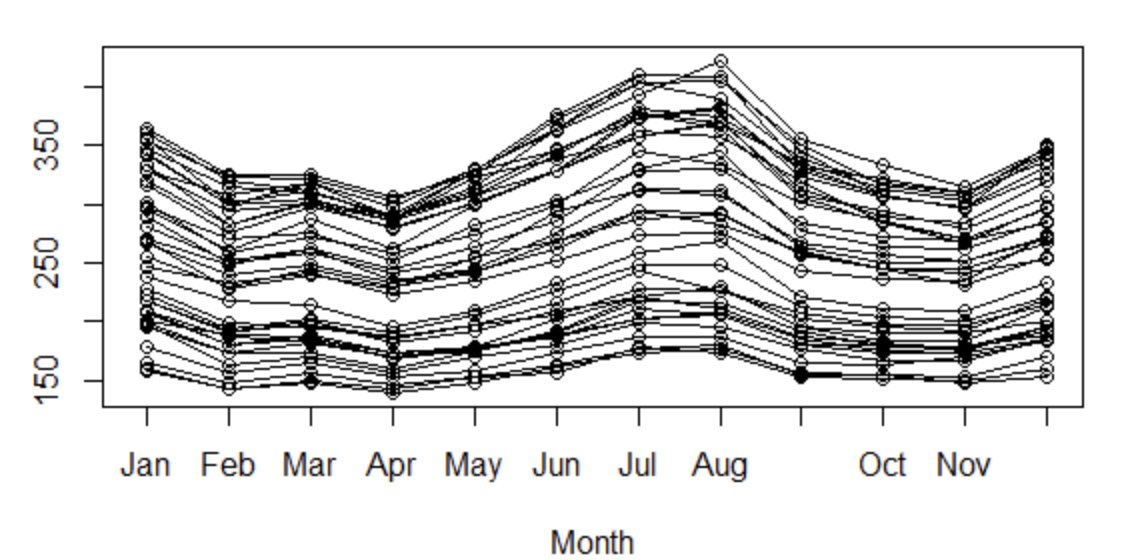
* Trend can be thought of as the linear increasing or decreasing behavior of the series over a long period of time. The trend usually happens for some time and then disappears, it does not repeat.
* A trend can be uptrend, downtrend, or can be up and down, need not be a straight line.
* Trend line is a smooth predictable function that traces the trend of a time series, and can help us predict the time series indefinitely in the future.



* Trend can be calculated by taking the rolling average (or moving average) over a long period of time or by just fitting a Linear Regression line on the points.

## **Seasonality**

* Seasonality in time-series data refers to a pattern that occurs at a regular interval.
* Making copies of seasonality can fetch us our time series.
* A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week.
* The steps for calculating the seasonality are:
  + Calculate the trend and subtract it from the original time series.
  + From the result, take the average across the period. If it is a monthly time series for 4 years, then you have 4 Jan, 4 Feb, 4 Dec, etc. Take the avg of all Jans, Febs, etc.
* A yearly series does not have seasonality.
* The signal can have multiple seasonalities one can be short-term seasonality and one can be long-term seasonality.



## **Irregular Component**

* It is the random fluctuation in the time series data that the above components(i.e, trend and seasonality) cannot explain.
* This component is assumed to have a normal distribution with 0 mean and constant variance.
* This is also called Error/ White Noise/ Remainder.

## **Additive** **Decomposition**

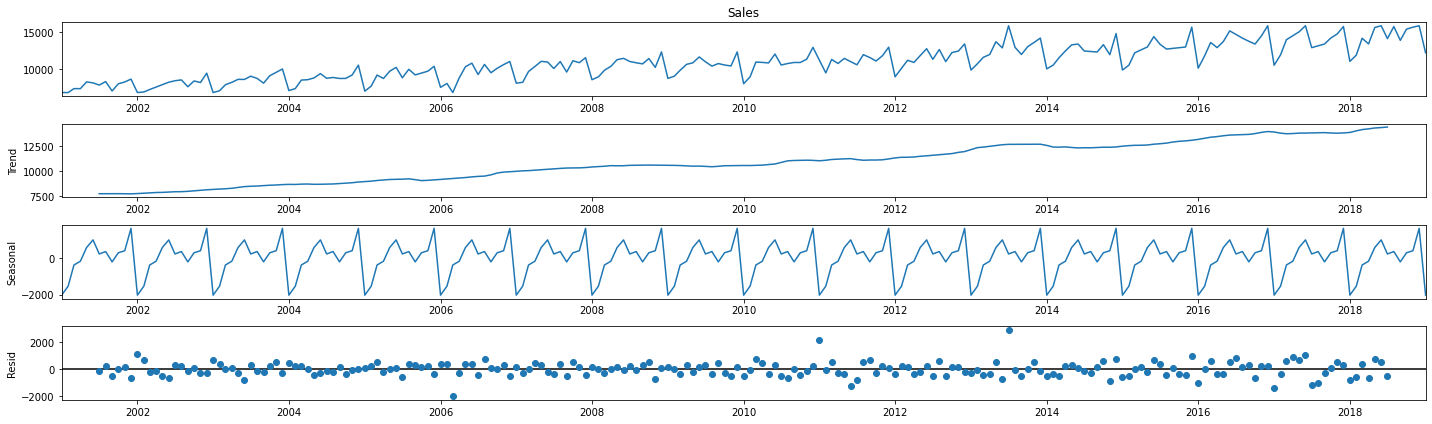


Where, b(t) -> trend of signal

s(t) -> seasonality

e(t) -> error term

* For the sales data, decomposition would be in the following way:



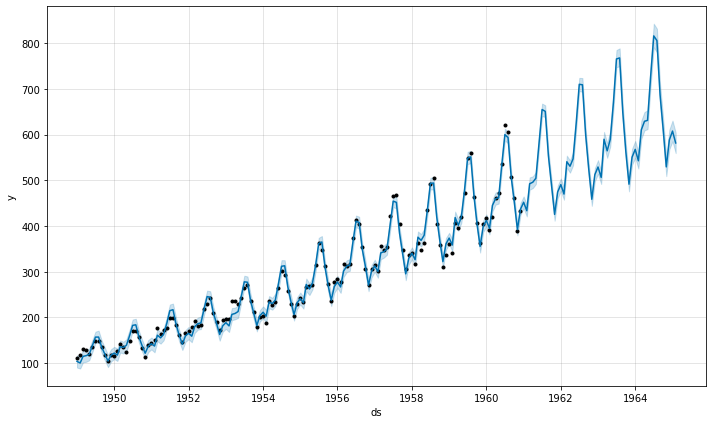
* The error term e(t) cannot be estimated since it is based on the real values that we have but can be computed using the same formulation.



* If the trend and seasonality components obtained are a good estimate, then the errors would be small and would be scattered around zero.
* The seasonal fluctuations are not dependent on the trend and the seasonality will be constant.

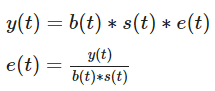
## **Multiplicative** **Decomposition**

* It is a time series in which the amplitude of the seasonal component is increasing with an increasing trend.



**Multiplicative seasonality Example**

* In multiplicative seasonality, we do not obtain the time series by adding the trend and seasonality components with each other. Instead, if we multiply them, then we obtain a time series in which the amplitude of the seasonal component is increasing with an increasing trend.
* Therefore, the **multiplicative seasonality decomposition** is written as



* The seasonal fluctuations are dependent on the trend.

**But how do we actually decompose the time series?**

* One of the methods is that we can take long-term moving averages and connect them to get a trend line.
* Another way can be by fitting a linear regression to the data points to get the trend line.
  + The line equation would be Y**(t+1)**= m(t) + c
  + You can also fit with the higher degree polynomial curve.
* Once you have the trend, you can subtract the trend from the current time series and take a group average to find the seasonal component.