# INTRODUCTION

**Time Series** is a set of data points or observations taken at specified times usually at equal intervals (e.g. hourly, daily, weekly, quarterly, yearly, etc.).

**Time series analysis** comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

**Time series forecasting** is the use of a model to predict future values based on previously observed values.

Example: Predicting what would happen in the stock market tomorrow, volume of goods that would be sold in the coming week, whether or not price of an item would skyrocket in December, number of Uber rides over a period of time are some of the things we can do with Time Series Analysis.

## TYPE OF TIME SERIES DATA

The data is considered in three types:

* 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.

## WHY TIME SERIES ANALYSIS

Time series helps us understand past trends so we can forecast and plan for the future. For example, you own a coffee shop, what you’d likely see is how many coffees you sold every day or month and when you want to see how your shop has performed over the past six months, you’re likely going to add all the six-month sales. Now, what if you want to be able to forecast sales for the next six months or year. In this kind of scenario, the only variable known to you is time (either in seconds, minutes, days, months, years, etc.) — hence you need Time Series Analysis to predict the other unknown variables like trends, seasonality, etc.

## METHODS FOR ANALYSIS

Time series analysis techniques may be divided into parametric and non-parametric methods.

### PARAMETRIC APPROACH

The **parametric approaches** assume that the **underlying stationary stochastic process** has a certain structure which can be described using a small number of parameters (for example, using an **autoregressive or moving average model**). In these approaches, the task is to estimate the parameters of the model that describes the stochastic process.

In probability theory and related fields, a stochastic or random process is a mathematical object usually defined as a collection of random variables. Example: Stochastic process representing numerical values of some system randomly changing over time, such as the growth of a bacterial population

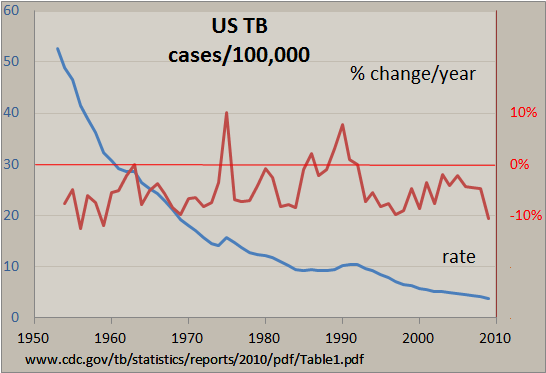
### NON-PARAMETRIC APPROACH

By contrast, non-parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure.

Methods of time series analysis may also be divided into linear and non-linear, and univariate and multivariate.

## EXPLORATORY ANALYSIS

The clearest way to examine a regular time series manually is with a line chart such as the one shown for tuberculosis in the United States, made with a spreadsheet program.



Other techniques include:

* **Autocorrelation analysis** to examine serial dependence
* **Spectral analysis** to examine cyclic behavior which need not be related to seasonality. For example, sun spot activity varies over 11-year cycles. Other common examples include celestial phenomena, weather patterns, neural activity, commodity prices, and economic activity.
* Separation into components representing trend, seasonality, slow and fast variation, and cyclical irregularity.

## 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.

## MODELS

Models for time series data can have many forms and represent different stochastic processes. When modeling variations in the level of a process, three broad classes of practical importance are the

* Autoregressive (AR) models.
* Integrated (I) models and
* Moving Average (MA) models. T

These three classes depend linearly on previous data points. Combinations of these ideas produce autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models.

Time series models are very useful models when you have serially correlated 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:

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

**AUTOREGRESSIVE COMPONENT:** AR stands for autoregressive. Autoregressive parameter 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.

# LOADING DATA IN PANDAS

We can specify below options while reading time series data in pandas:

**parse\_dates:** This specifies the column which contains the date-time information. As we say above, the column name is ‘Month’.

**index\_col:** A key idea behind using Pandas for TS data is that the index has to be the variable depicting date-time information. So, this argument tells pandas to use the ‘Month’ column as index.

**date\_parser:** This specifies a function which converts an input string into datetime variable. Be default Pandas reads data in format ‘YYYY-MM-DD HH:MM: SS’. If the data is not in this format, the format has to be manually defined. Something similar to the dataparse function defined here can be used for this purpose.

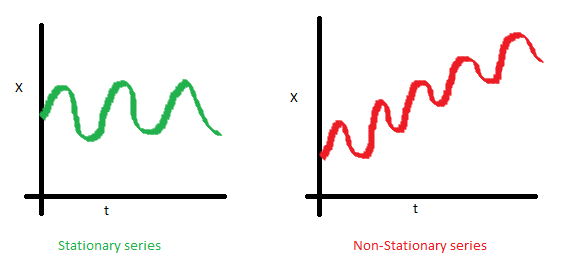
Import ant points to remember:

* Unlike numeric indexing, the end index is included here. For instance, if we index A list as A[:5] then it would return the values at indices – [0,1,2,3,4]. But here the index ‘1949-05-01’ was included in the output.
* The indices have to be sorted for ranges to work. If you randomly shuffle the index, this won’t work.

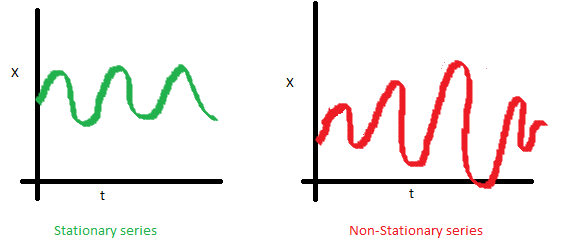
# STATIONARY SERIES

There are three basic criteria for a series to be classified as stationary series:

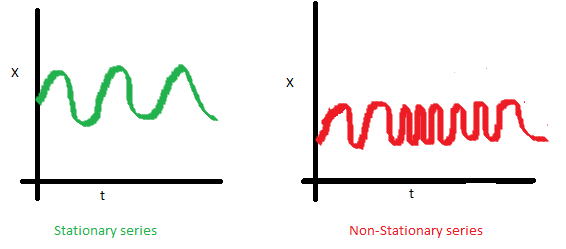
1. The mean of the series should not be a function of time rather should be a constant. The image below has the left-hand graph satisfying the condition whereas the graph in red has a time dependent mean.



1. The variance of the series should not a be a function of time. This property is known as homoscedasticity. Following graph depicts what is and what is not a stationary series. (Notice the varying spread of distribution in the right-hand graph)



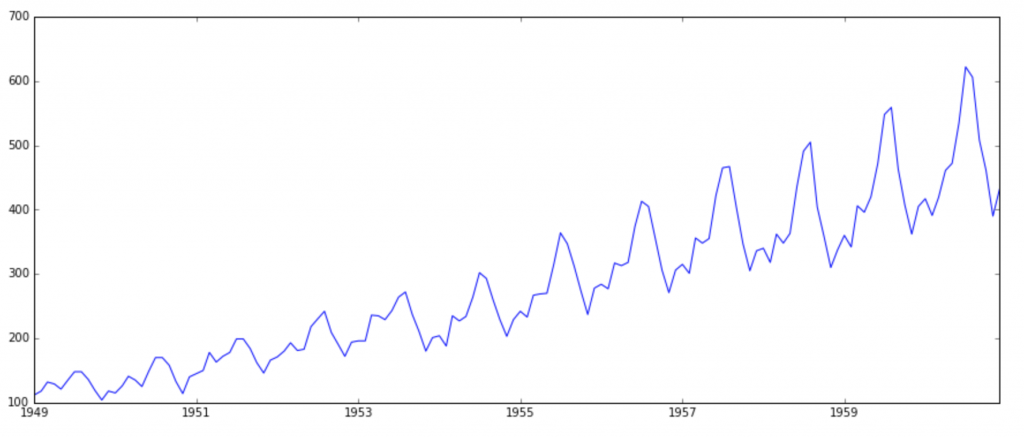
1. The covariance of the ith term and the (i + m) th term should not be a function of time. In the following graph, you will notice the spread becomes closer as the time increases. Hence, the covariance is not constant with time for the ‘red series.



Most of the Time Series models work on the assumption that the TS is stationary. Intuitively, we can say that if a Time Series has a particular behavior over time, there is a very high probability that it will follow the same in the future. Also, the theories related to stationary series are more mature and easier to implement as compared to non-stationary series.

## CHECK STATIONARITY OF A TIME SERIES

For testing stationarity, we can plot the data and analyze visually.



It is clearly evident that there is an overall increasing trend in the data along with some seasonal variations. More formally, we can check stationarity using the following:

* **PLOTTING ROLLING STATISTICS:** We can plot the moving average or moving variance and see if it varies with time. By moving average/variance I mean that at any instant ‘t’, we’ll take the average/variance of the last year, i.e. last 12 months. But again, this is more of a visual technique.
* **DICKEY-FULLER TEST:** This is one of the statistical tests for checking stationarity. Here the null hypothesis is that the TS is non-stationary. The test results comprise of a Test Statistic and some Critical Values for difference confidence levels. If the ‘Test Statistic’ is less than the ‘Critical Value’, we can reject the null hypothesis and say that the series is stationary. Refer this article for details.

## MAKING A TIME SERIES STATIONARY

Though stationarity assumption is taken in many Time Series models, almost none of practical time series are stationary. So, statisticians have figured out ways to make series stationary.

There are 2 major reasons behind non-stationarity of a Time Series:

1. Trend: varying mean over time. For e.g. in this case we saw that on average, the number of passengers was growing over time.

2. Seasonality: variations at specific time-frames. E.g. people might have a tendency to buy cars in a particular month because of pay increment or festivals.

The underlying principle is to model or estimate the trend and seasonality in the series and remove those from the series to get a stationary series. Then statistical forecasting techniques can be implemented on this series. The final step would be to convert the forecasted values into the original scale by applying trend and seasonality constraints back.

Some of the tricks to reduce trend are:

* One of the first tricks to reduce trend can be transformation
* Aggregation – taking average for a time period like monthly/weekly averages
* Smoothing – taking rolling averages
* Polynomial Fitting – fit a regression model

## ELIMINATING TREND AND SEASONALITY

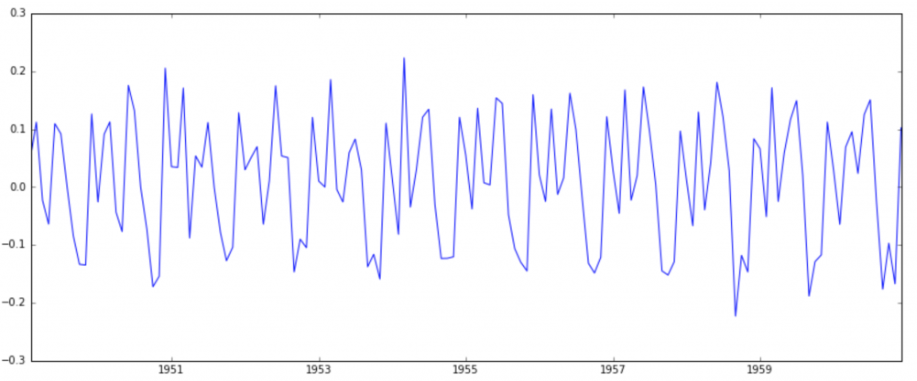
The simple trend reduction techniques discussed before don’t work in all cases, particularly the ones with high seasonality. Let’s discuss two ways of removing trend and seasonality:

* Differencing – taking the difference with a particular time lag
* Decomposition – modeling both trend and seasonality and removing them from the model.

### DIFFERENCING

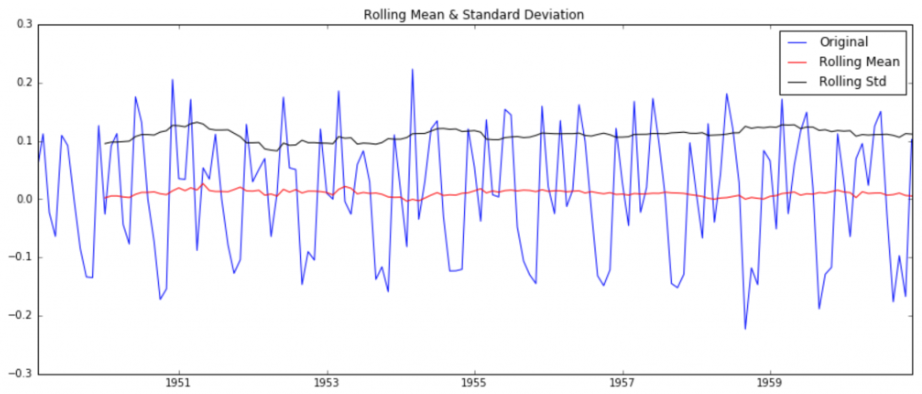
One of the most common methods of dealing with both trend and seasonality is differencing. In this technique, we take the difference of the observation at a particular instant with that at the previous instant. This mostly works well in improving stationarity. First order differencing can be done in Pandas as:

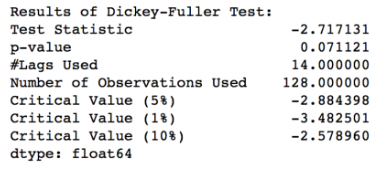




This appears to have reduced trend considerably. Let’s verify using our plots:





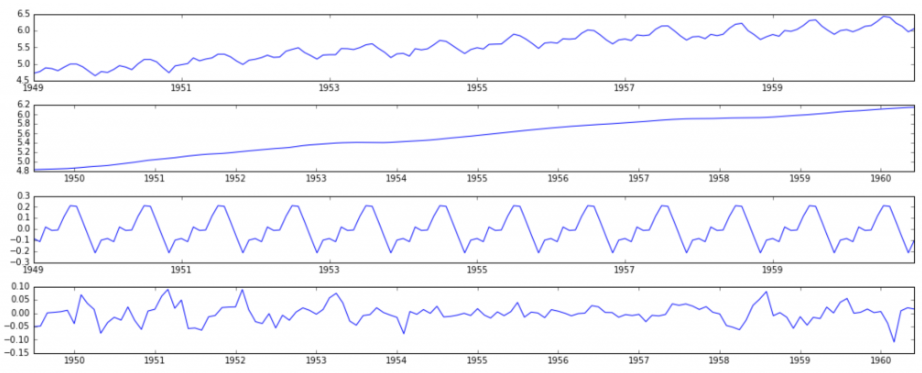


We can see that the mean and std variations have small variations with time. Also, the Dickey-Fuller test statistic is less than the 10% critical value, thus the TS is stationary with 90% confidence. We can also take second or third order differences which might get even better results in certain applications. I leave it to you to try them out.

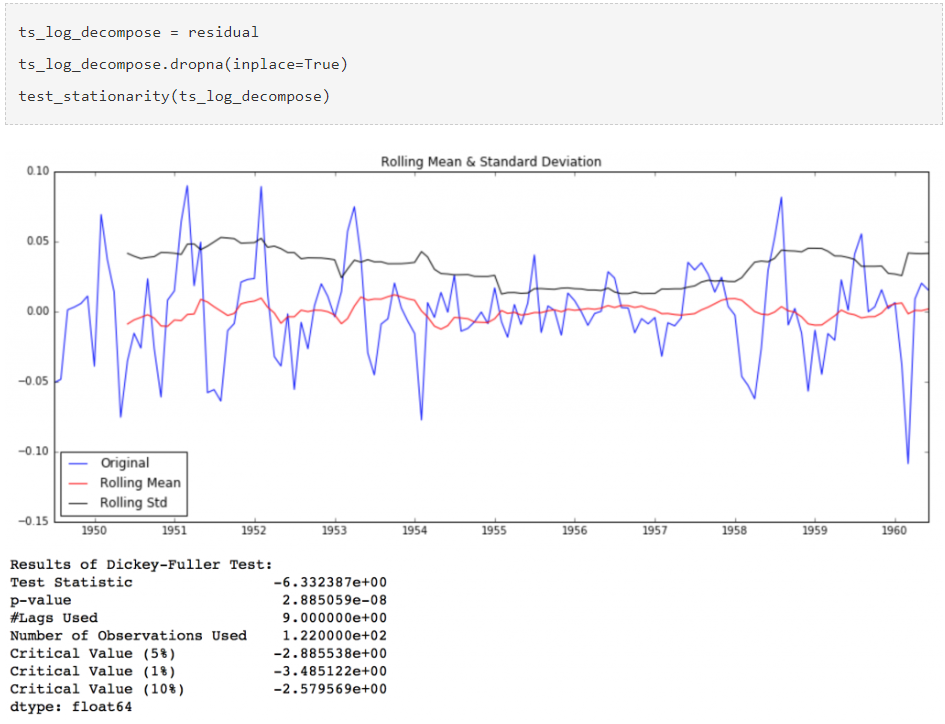
### DECOMPOSING

In this approach, both trend and seasonality are modeled separately and the remaining part of the series is returned.





Here we can see that the trend, seasonality is separated out from data and we can model the residuals. Let’s check stationarity of residuals:



The Dickey-Fuller test statistic is significantly lower than the 1% critical value. So, this TS is very close to stationary. We can try advanced decomposition techniques as well which can generate better results. Also, we should note that converting the residuals into original values for future data in not very intuitive in this case.

# FORECASTING A TIME SERIES

Having performed the trend and seasonality estimation techniques, there can be two situations:

* A strictly stationary series with no dependence among the values. This is the easy case wherein we can model the residuals as white noise. But this is very rare.
* A series with significant dependence among values. In this case we need to use some statistical models like ARIMA to forecast the data.

ARIMA stands for Auto-Regressive Integrated Moving Averages. The ARIMA forecasting for a stationary time series is nothing but a linear (like a linear regression) equation. The predictors depend on the parameters (p, d, q) of the ARIMA model:

* **Number of AR (Auto-Regressive) terms (p):** AR terms are just lags of dependent variable. For instance, if p is 5, the predictors for x(t) will be x(t-1) ………x(t-5).
* **Number of MA (Moving Average) terms (q):** MA terms are lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1 )….e(t-5) where e(i) is the difference between the moving average at ith instant and actual value.
* **Number of Differences (d):** These are the number of nonseasonal differences, i.e. in this case we took the first order difference. So, either we can pass that variable and put d=0 or pass the original variable and put d=1. Both will generate same results.

An importance concern here is how to determine the value of ‘p’ and ‘q’. We use two plots to determine these numbers. Let’s discuss them first.

* **AUTOCORRELATION FUNCTION (ACF):** It is a measure of the correlation between the TS with a lagged version of itself. For instance, at lag 5, ACF would compare series at time instant ‘t1’…’t2’ with series at instant ‘t1-5’…’t2-5’ (t1-5 and t2 being end points).
* **PARTIAL AUTOCORRELATION FUNCTION (PACF):** This measures the correlation between the TS with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. E.g. at lag 5, it will check the correlation but remove the effects already explained by lags 1 to 4.

The ACF and PACF plots for the TS after differencing can be plotted as:



# REFERENCE

<https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/>

<https://github.com/PacktPublishing/Practical-Time-Series-Analysis>

<https://github.com/wblakecannon/DataCamp/tree/master/22-introduction-to-time-series-analysis-in-python>