Time Series Analysis

- Time series analysis is extensively used to forecast company sales, product demand, stock market trends, agricultural production etc.

- The fundamental idea for time series analysis is to decompose the original time series (sales, stock market trends, etc.) into several independent components.

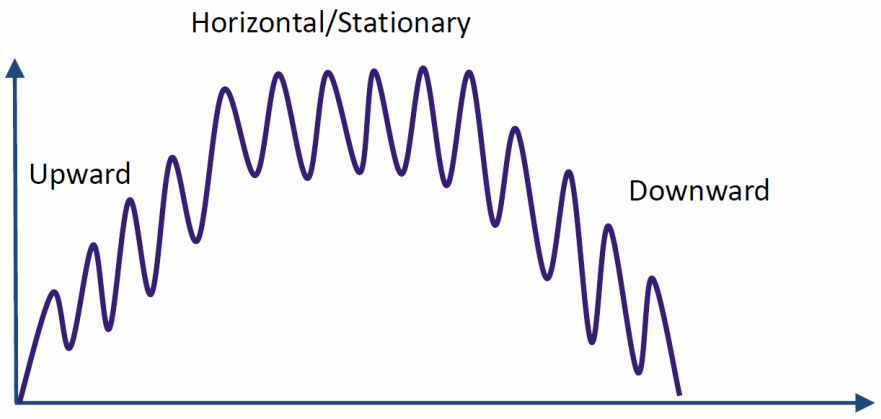
Time Series Analysis

- Typically, business time series are divided into the following four components:
- **Trend** – overall direction of the series i.e. upwards, downwards etc.
- **Seasonality** – monthly or quarterly patterns
- **Cycle** – long-term business cycles, they usually come after 5 or 7 years
- **Irregular remainder** – random noise left after extraction of all the components

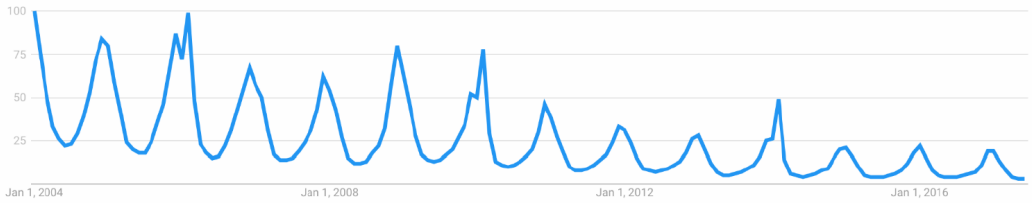
Time Series Analysis

- Interference of these components produces the final series.
- Why decomposing the original / actual time series into components?
- It is much easier to forecast the individual regular patterns produced through decomposition of time series than the actual series.

- Trends



- Seasonality - Repeating trends



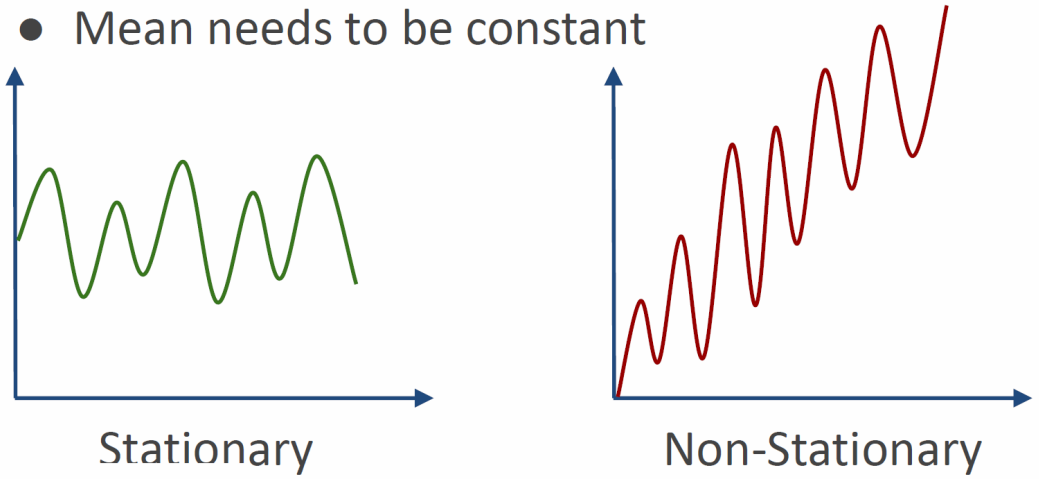
Google Trends - “Snowboarding”

- Cyclical - Trends with no set repetition.

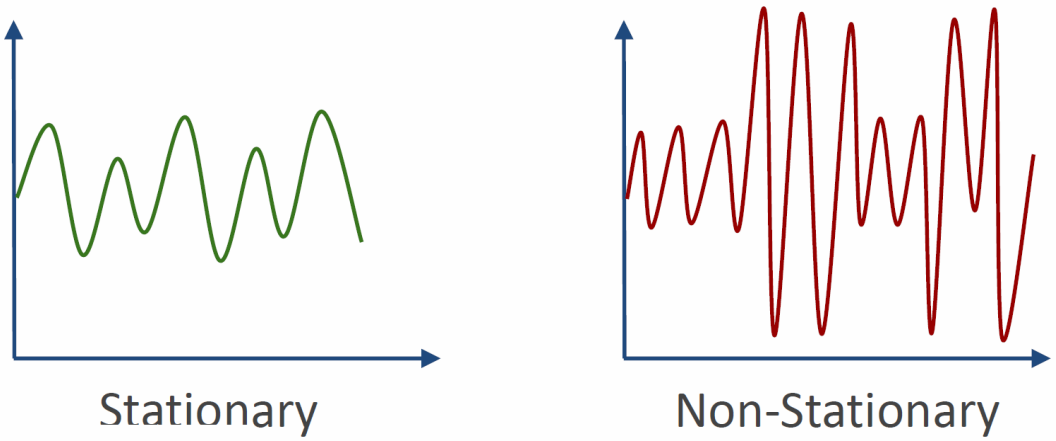
SP500

- Stationary vs Non-Stationary Data
  - To effectively use ARIMA, we need to understand Stationarity in our data.
  - So what makes a data set Stationary?
    - A Stationary series has constant mean and variance over time.

- Mean needs to be constant



- Variance should not be a function of time



- A Stationary data set will allow our model to predict that the mean and variance will be the same in future periods.

- There are also mathematical tests you can use to test for stationarity in your data.
- A common one is the Augmented Dickey–Fuller test

- If we've determined your data is not stationary (either visually or mathematically), we will then need to transform it to be stationary in order to evaluate it and what type of ARIMA terms you will use.

- One simple way to do this is through "differencing".

Original Data

Time 1	10
Time 2	12
Time 3	8
Time 4	14
5	

First Difference

Time 1	NA
Time 2	2
Time 3	-4
Time 4	6
Time 5	-7

Second Difference

Time 1	NA
Time 2	NA
Time 3	-6
Time 4	10
Time 5	-13

- You can continue differencing until you reach stationarity (which you can check visually and mathematically)
- Each differencing step comes at the cost of losing a row of data.

- For seasonal data, we can also difference by a season.
- For example, if we had monthly data with yearly seasonality, we could difference by a time unit of 12, instead of just 1.

- With our data now stationary it is time the p,d,q terms and how we choose them.
- A big part of this are AutoCorrelation Plots and Partial AutoCorrelation Plots.

Trend

- From the plots it is obvious that there is some kind of increasing trend in the series along with seasonal variation.

- Stationarity is a vital assumption we need to verify if our time series follows a stationary process or not.

Trend

- We can do by
  - Plots: review the time series plot of our data and visually check if there are any obvious trends or seasonality
  - Statistical tests: use statistical tests to check if the expectations of stationarity are met or have been violated.

Trend using MAs

- Moving averages over time
  - One way to identify a trend pattern is to use moving averages over a specific window of past observations.
  - This smoothens the curve by averaging adjacent values over the specified time horizon (window).

Seasonality

- People tend to go on vacation mainly during summer holidays.

- At some time periods during the year people tend to use aircrafts more frequently. We can check the hypothesis of a seasonal effect

Noise

- To understand the underlying pattern in the number of international airline passengers, we assume a multiplicative time series decomposition model

- Purpose is to understand underlying patterns in temporal data to use in more sophisticated analysis like Holt-Winters seasonal method or ARIMA.

Noise

- Noise - is the residual series left after removing the trend and seasonality components

Stationarize a Time series

- Before models forecasting can be applied, the series must be transformed into a stationary time series.

- The Augmented-Dickey Fuller Test can be used to test whether or not a given time series is stationary.

Stationarize a Time series

- If the test statistic is smaller than the critical value, the hypothesis is rejected, the series would be stationary, and no further transformations of the data would be required.

Residuals Serial Correlation

- When the residuals (errors) in a time series are correlated with each other it is said to exhibit serial correlation.

- Autocorrelation is a better measurement for the dependency structure, because the autocovariacne will be affected by the underlying units of measurement for the observation.

White Noise & ACF & PACF

- Random process is white noise process
- Errors are serially uncorrelated if they are independent and identically distributed (iid).
- It is important because if a time series model is successful at capturing the underlying process, residuals of the model will be iid and resemble a white noise process.

White Noise

- Part of time series analysis is simply trying to fit a model to a time series such that the residual series is indistinguishable white noise.

ACF & PACF

- The plots of the Autocorrelation function (ACF) and the Partial Autorrelation Function (PACF) are the two main tools to examine the time series dependency structure.

- The ACF is a function of the time displacement of the time series itself.

- It is the similarity between observations as a function of the time lag between them.

PACF

- The PACF is the conditional correlation between two variables under the assumptions that the effects of all previous lags on the time series are known.

Random Walk

- What is special about the random walk is, that it is non-stationary, that is, if a given time series is governed by a random walk process it is unpredictable.
- It has high ACF for any lag length
- The normal QQ plot and the histogram indicate that the series is not normally distributed

Random Walk

- The random walk is a first order autoregressive process that is, this causes the process to be non-stationary.
- The process can be made stationary

Auto Regressive Model – AR(p)

- The random walk process belongs to a more general group of processes, called autoregressive process

- The current observation is a linear combination of past observations.

- An AR(1) time series is one period lagged weighted version of itself.

The Moving Average Model - MA(q)

- The moving average model MA(q) assumes that the observed time series can be represented by a linear combination of white noise error terms.

- The time series will always be stationary.

ARIMA Forecasting

- An autoregressive integrated moving average (ARIMA) model is an generalization of an autoregressive moving average (ARMA) model.

- Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting).

ARIMA Forecasting

- ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

- There are three parameters (p, d, q) that are used to parametrize ARIMA models. Hence, an ARIMA model is denoted as $ARIMA(p, d, q)$

- Each of these three parts is an effort to make the time series stationary, i. e. make the final residual a white noise pattern.

Optimal Parameter Selection

- To fit the time series data to a seasonal ARIMA model with parameters ARIMA(p, d, q)(P, D, Q)s$ the optimal parameters need to be found first.

- This is done via grid search, the iterative exploration of all possible parameters constellations.

Optimal Parameter Selection

- Depending on the size of the model parameters $(p, d, q)(P, D, Q)s$ this can become an extremely costly task with regard to computation. We start of by generating all possible parameter constellation we'd like to evaluate.

Akaike Information Criterion (AIC).

- For all possible parameter constellations from both lists pdq and seasonal_pdq the alogrithm will create a model and eventually pick the best one to proceed.

- The best model is chosen based on the Akaike Information Criterion (AIC).

Akaike Information Criterion (AIC).

- The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data.

- Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.

- It measures the trade-off between the goodness of fit of the model and the complexity of the model (number of included and estimated parameters).

One step ahead prediction

- The get_prediction and conf_int methods calculate predictions for future points in time for the previously fitted model and the confidence intervals associated with a prediction, respectively.

- The dynamic=False argument causes the method to produce a one-step ahead prediction of the time series

MSE

- To quantify the accuracy between model fit and true observations we use the mean squared error (MSE).

- The MSE computes the squared difference between the true and predicted value.

Out of sample Prediction

- To put the model to the real test with a 24-month-head prediction.

- This requires to pass the argument dynamic=False when using the get_prediction method.'''

Long term forecasting

- Finally, a 10 year ahead forecast, leveraging a seasonal ARIMA model trained on the complete time series y.

- Grid search found the best model to be of form SARIMAX(2, 1, 3)(1, 2, 1)12 for the data vector y.