CSC 735 – Data Analytics

Time Series Fundamentals

Introduction

- A **time series** is a dataset with observations in an ordered sequence with explicit attributes that indicate a temporal value
- In contrast, an **event sequence** is an ordered series of observations with implicit temporal attributes: e.g., A B C B C B

Introduction

- A **time series** is a dataset with observations in an ordered sequence with explicit attributes that indicate a temporal value
- In contrast, an **event sequence** is an ordered series of observations with implicit temporal attributes: e.g., A B C B C B
- If the observations are equally spaced in time, it is a **regular time series;** otherwise it is **irregular.**
- Time series data can be used to forecast future values of a time series using forecasting models

Time Series Data

Machine learning data set →

Sensor ID	Value
Sensor_1	20
Sensor_1	21
Sensor_2	22
Sensor_2	23

Time series data set →

Sensor ID	Time Stamp	Value
Sensor_1	01/01/2020	20
Sensor_1	01/02/2020	21
Sensor_2	01/01/2020	22
Sensor_2	01/02/2020	23

Time Series Data

In a time series, the chronological arrangement of data is captured in a specific column that is often denoted as *time stamp*, *date*, or simply *time*.

Machine learning data set →

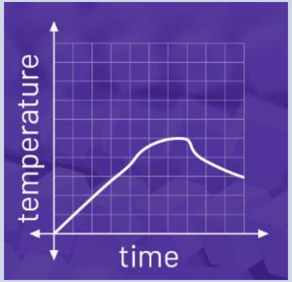
Sensor ID	Value
Sensor_1	20
Sensor_1	21
Sensor_2	22
Sensor_2	23

Time series data set →

Sensor ID	Time Stamp	Value
Sensor_1	01/01/2020	20
Sensor_1	01/02/2020	21
Sensor_2	01/01/2020	22
Sensor_2	01/02/2020	23

Discrete vs. Continuous Values

The time value can be discrete or continuous (default is discrete)

Discrete	Continuous
Specific values	Any values (within a range)
“Counted”	“Measured”
For example, the child’s allowance increases $20 each month. 	For example, the temperature measured 18.523…°C at 7AM. 

https://www.g2.com/articles/discrete-vs-continuous-data

Time Series Examples

- Stock market prices (day, month, etc.)
- Temperatures (minute, hour, etc.)
- Number of patients (day, week, etc.)
- GDP (quarter, year, etc.)
- Sensor reading (every millisecond, second, etc.)

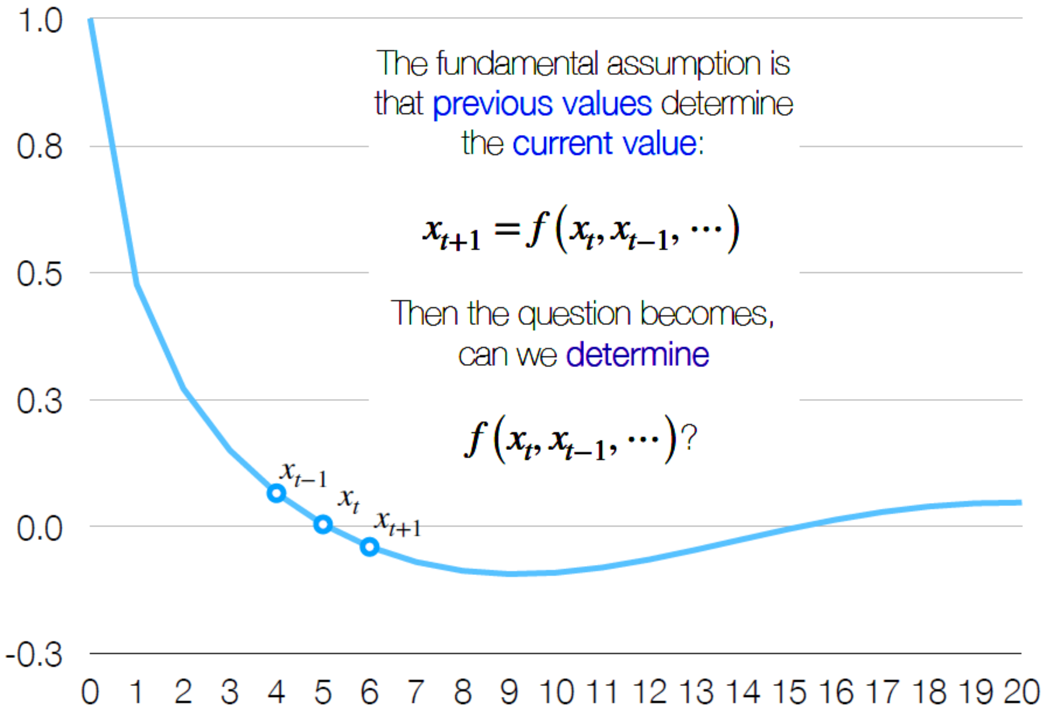
Notation Introduction

- The value of the time series at time t is given by x_t
- Regular time series: Observations are ordered by equal time intervals and, thus, can be represented as: $x_1, x_2, x_3, \ldots$
- In a time series, time is the primary way of structuring your dataset

Notation Introduction

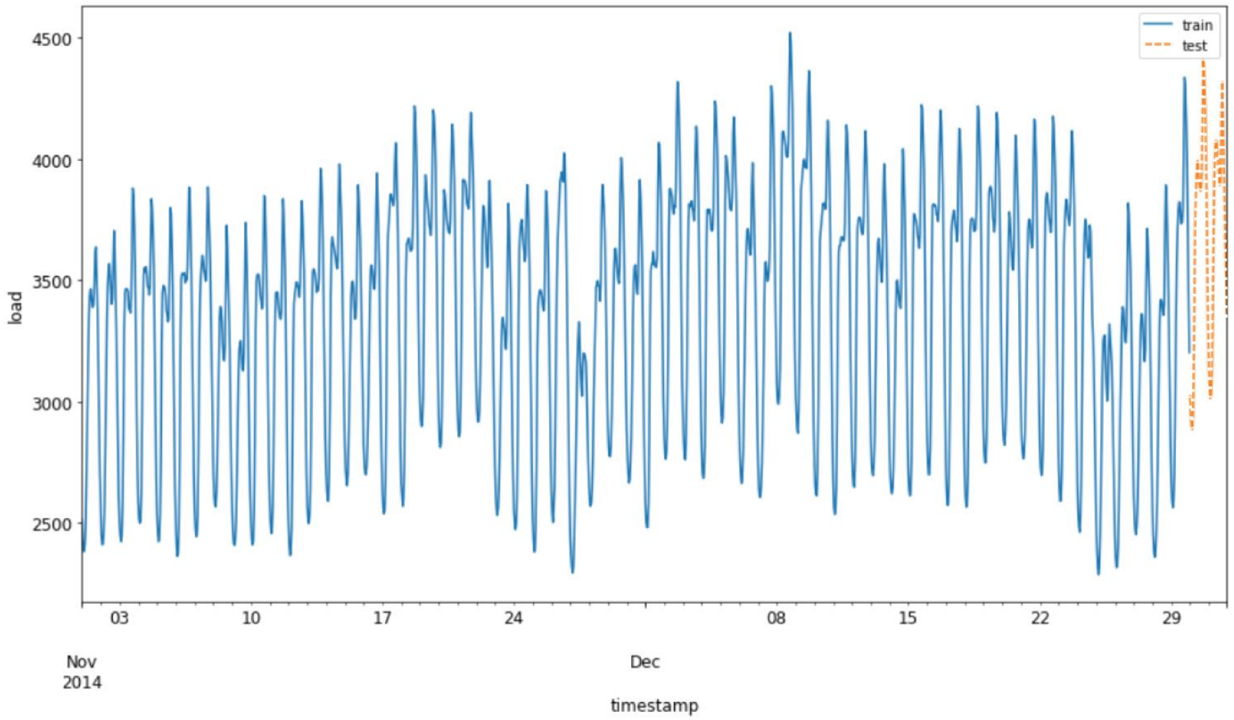
- **Trends** and **seasonality** can be determined to assist in modelling the behaviour of the time series for future planning
- **Forecasting** is the action of predicting future values based on past behaviour

Notation Introduction



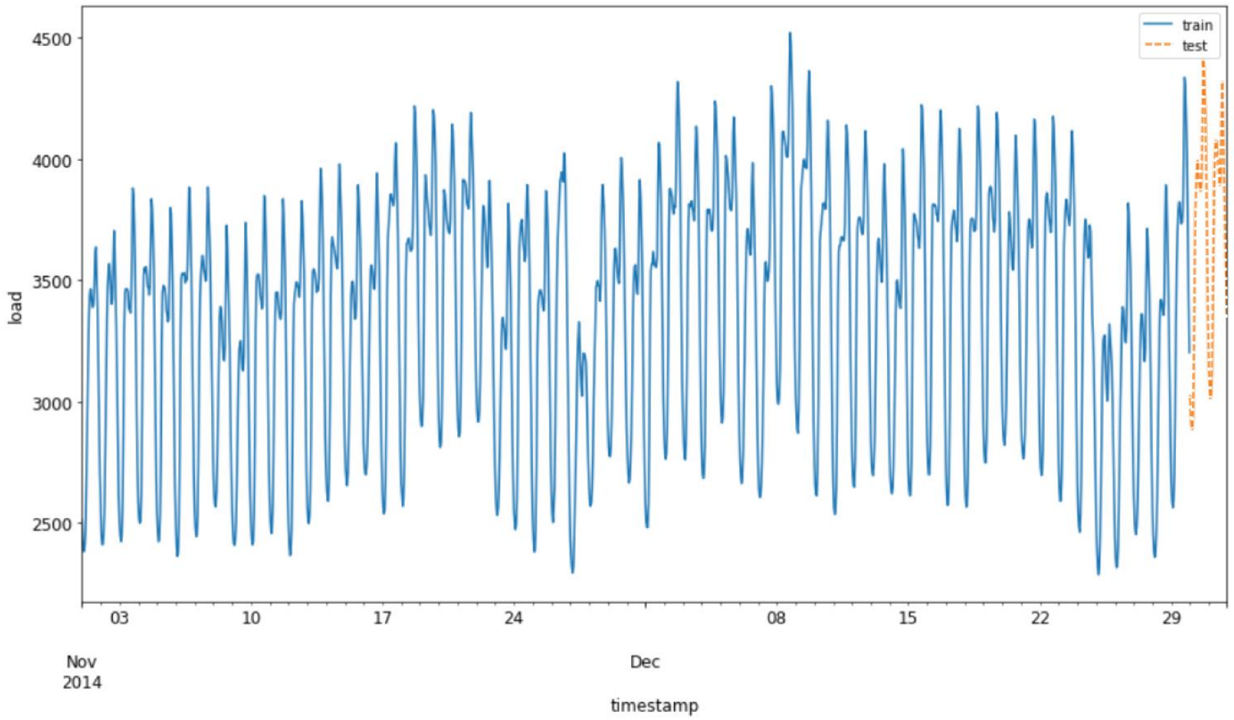
Time Series Problems

- What are the expected sales volumes of thousands of food groups in different grocery stores next quarter?
- What are the resale values of vehicles after leasing them out for three years?
- What is the future electricity load in an energy supply chain infrastructure, so that suppliers can ensure efficiency and prevent energy waste and theft?



Note:
“test” refers to forecasted values

Example of time series forecasting applied to the energy load use case



Note:
“test” refers to forecasted values

What are some observations you have about this graph?

Analysis vs. Forecasting

- **Time series analysis** consists of analyzing and preparing time series data prior to constructing a time series model.
- It is used for many applications such as process and quality control, utility studies, and census analysis
- It is usually considered the first step prior to the modeling step, which is properly called ***time series forecasting***

Analysis vs. Forecasting

- **Time series forecasting** involves taking machine learning models, training them on historical time series data, and applying them to predict future values.

Time series analysis on historical and/or current time stamps and values

Sensor ID	Time Stamp	Values
Sensor_1	01/01/2020	60
Sensor_1	01/02/2020	64
Sensor_1	01/03/2020	66
Sensor_1	01/04/2020	65
Sensor_1	01/05/2020	67
Sensor_1	01/06/2020	68
Sensor_1	01/07/2020	70
Sensor_1	01/08/2020	69
Sensor_1	01/09/2020	72
Sensor_1	01/10/2020	?
Sensor_1	01/11/2020	?
Sensor_1	01/12/2020	?

Time series forecasting on future time stamps to generate future values

Difference between time series analysis historical input data and time series forecasting output data

Time series analysis on historical and/or current time stamps and values

Sensor ID	Time Stamp	Values
Sensor_1	01/01/2020	60
Sensor_1	01/02/2020	64
Sensor_1	01/03/2020	66
Sensor_1	01/04/2020	65
Sensor_1	01/05/2020	67
Sensor_1	01/06/2020	68
Sensor_1	01/07/2020	70
Sensor_1	01/08/2020	69
Sensor_1	01/09/2020	72
Sensor_1	01/10/2020	?
Sensor_1	01/11/2020	?
Sensor_1	01/12/2020	?

Time series forecasting on future time stamps to generate future values

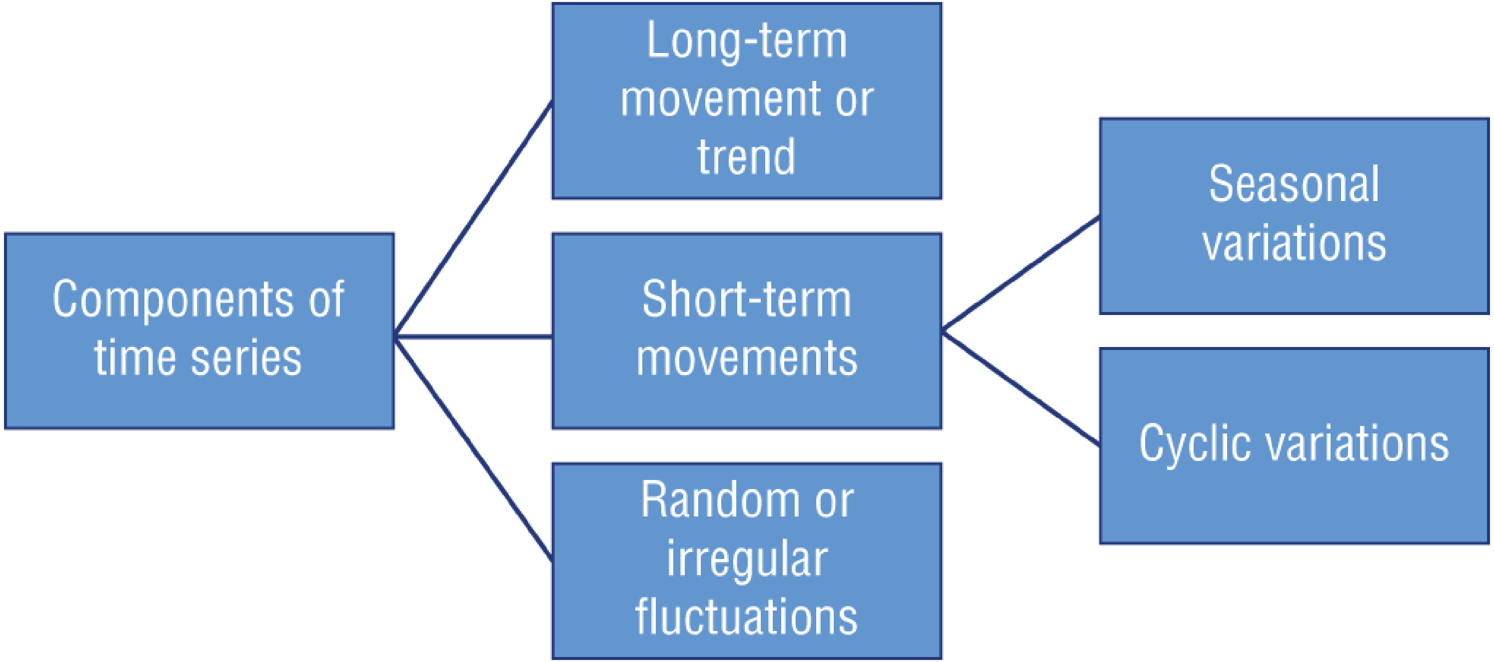
This dataset can be regarded as univariate because Sensor ID is always the same and the time stamps are regular (and could be replaced by an implicit parameter). Thus, Values is the only variable of interest.

Difference between time series analysis historical input data and time series forecasting output data

Components of Time Series

- Different historical and current **phenomena** may affect the values of the data in a time series, and these events can be diagnosed as components of a time series
- We should recognize these components and decompose the time series in order to separate them from other aspects of the data values

Components of Time Series



Components of Time Series

- ***Long-term movement*** or ***trend*** refers to the overall movement of time series values to increase or decrease during a prolonged time interval

Components of Time Series

- There are two different types of ***short-term movements***:
  - ***Seasonal variations*** *(regular period length)*
  - ***Cyclic variations*** *(irregular period length)*

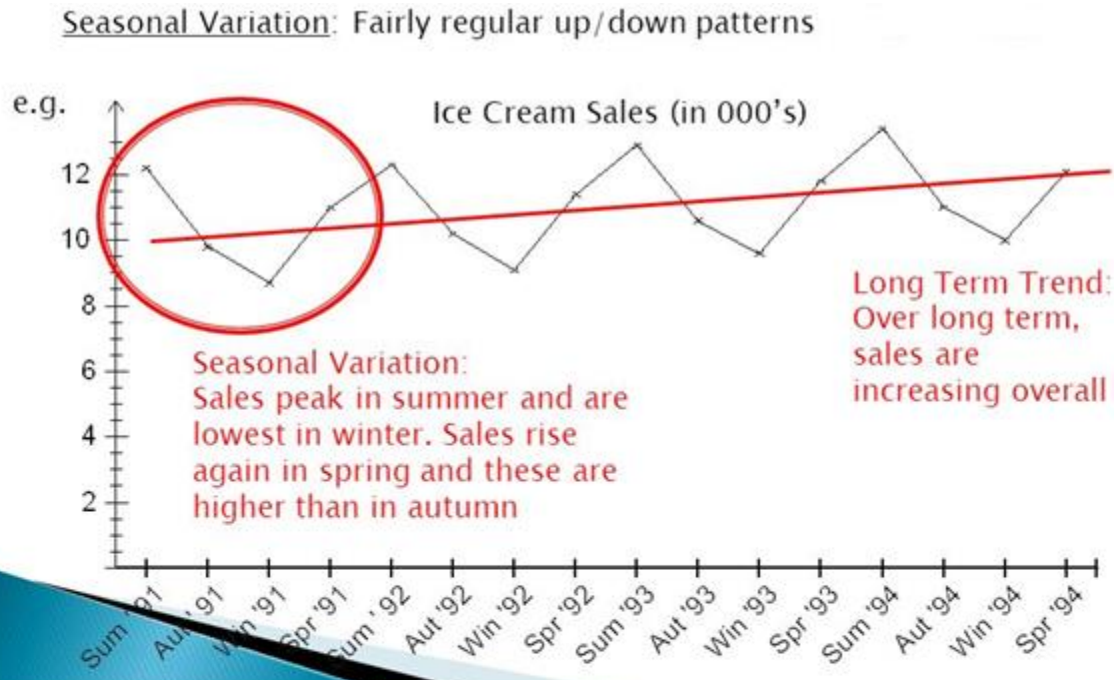
Seasonal Variations (Seasonality)

- ***Seasonal variations*** are periodic temporal fluctuations that show the same variation and usually recur over a period of a year or less
- Seasonality is always of a fixed and known period
- Most of the time, this variation will be present in a time series if the data is recorded hourly, daily, weekly, quarterly, or monthly

Seasonal Variations (Seasonality)

- Different social conventions (such as holidays and festivities), weather seasons, and climatic conditions play an important role in seasonal variations
- For example, the sale of umbrellas and raincoats in the rainy season and the sale of air conditioners in summer seasons
- The term *seasonality* is used even if the time period does not correspond to seasons in the normal sense

Seasonal Variations (Seasonality)



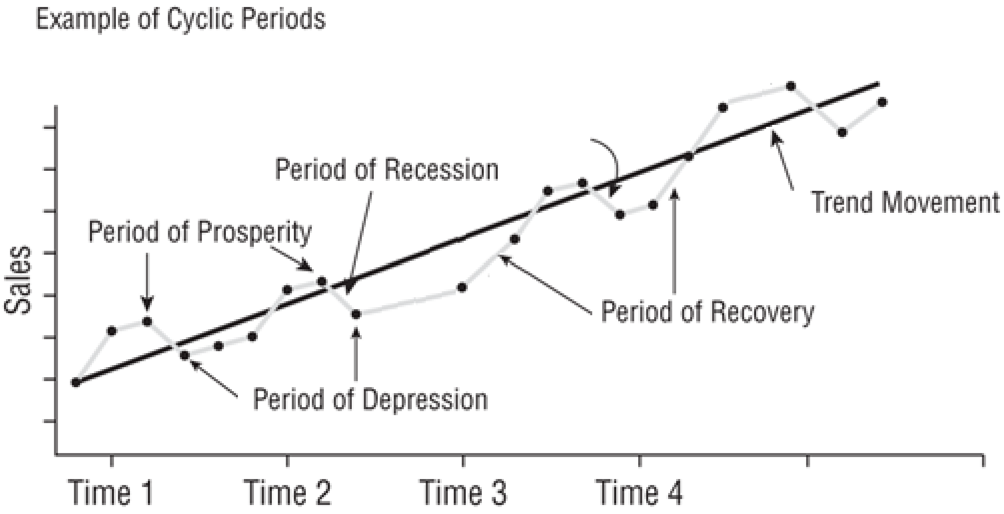
https://slideplayer.com/slide/9945281/

Cyclic Variations

- ***Cyclic variations***, on the other hand, are recurrent patterns that exist when data exhibits rises and falls that are <u>not</u> of a fixed period

Example of Cyclic Variations

- Each economic cycle consists of prosperity, recession, depression, and recovery
- The lengths of the cycles differ



Components of Time Series

- ***Random*** or ***irregular fluctuations*** are the last element to cause variations in time series data
  - These fluctuations are uncontrollable, unpredictable, and erratic, such as earthquakes, wars, floods, or any other natural disasters

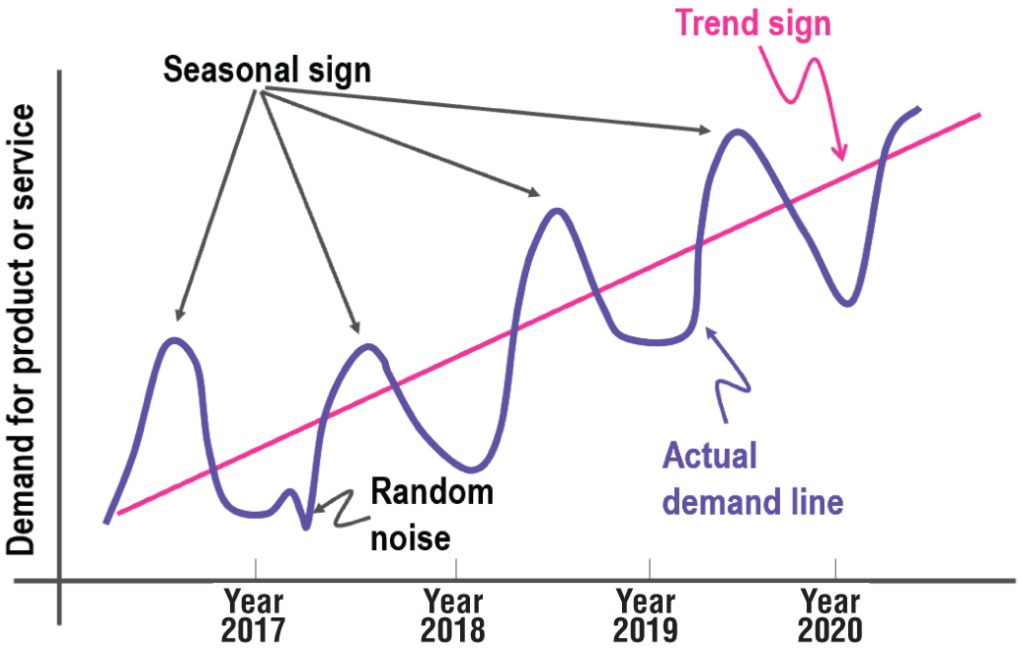
Components of Time Series

- Data scientists often refer to long-term movements, seasonal short-term movements, and cyclic short-term movements as ***signals*** in time series data because they actually are deterministic indicators that can be derived from the data itself

Components of Time Series

- Random or irregular fluctuations are an arbitrary variation of the values in your data that you cannot really predict, because each data point of these random fluctuations is independent of the other signals above, such as long-term and short-term movements
- Data scientists often refer to them as **noise**, because they are triggered by latent variables difficult to observe

Components of Time Series



Actual representation of time series components

Stationarity

- Understanding these four time series components and how to identify and remove them represents a strategic first step for building any time series forecasting solution because they can help with another important concept in time series that may help increase the predictive power of your machine learning algorithms: **stationarity**

Stationarity

- **Stationary** means that statistical parameters of a time series do not change over time
- Basic properties of the time series data distribution, like the mean and variance, remain constant over time
- Therefore, stationary time series processes are easier to analyze and model because the basic assumption is that their properties are not dependent on time and will be the same in the future

Stationarity

- A time series is defined as having a **strong stationarity** when **all** its statistical parameters do not change over time
- A time series is defined as having a **weak stationarity** when its mean and auto-covariance functions do not change over time

Nonstationarity

- Time series that exhibit changes in the values of their data, such as a trend or seasonality, are clearly not stationary, and as a consequence, they are more difficult to predict and model
- For accurate and consistent forecasted results to be received, the nonstationary data needs to be **transformed** into stationary data

Nonstationarity

- Another important reason for trying to render a time series stationary is to be able to obtain meaningful sample statistics such as means, variances, and correlations with other variables
- These sample statistics can be used to gain more insights from your data and can be included as additional features in your time series data set

Nonstationarity

- However, there are cases where unknown nonlinear relationships cannot be determined by classical methods, such as auto-regression, moving average, and autoregressive integrated moving average methods

Nonstationarity

- In reality, many economic time series are far from stationary when visualized in their original units of measurement
- Even after seasonal adjustment they will typically still exhibit trends, cycles, and other nonstationary characteristics

Forecasting Aspects

- Before building a forecasting solution, you should define the following aspects of the forecasting model:
- *Inputs and outputs*
- *Granularity level*
- *Endogenous and exogenous features*
- *Structured or unstructured features*
- *Univarate or multivariate*
- *Single-step or multi-step structure*

Forecasting Aspects

- *The inputs and outputs of your forecasting model*

– For data scientists who are about to build a forecasting solution, it is critical to think about the data they have available to make the forecast and what they want to forecast about the future

Forecasting Aspects

Granularity level of your forecasting model

- **Granularity** in time series forecasting represents the lowest detailed level of values captured for each time stamp. Granularity is related to the frequency at which time series values are collected
- Example: time series data that have been collected by sensors every few seconds

Endogenous Features

- An **endogenous feature** is an input variable of a forecasting model that has values that are determined by other variables in the system, and the output variable depends on it.

Endogenous Features

- Example:
  - predict daily gas prices
  - a binary variable named Holiday

Exogenous Features

- An **exogenous feature** is an input variable that is not influenced by other variables in the system and on which the output variable depends.
- In the gas pricing example,
- oil reserve prices or disasters such as oil tanker accidents

Exogenous Features

Exogenous variables present some common characteristics (Glen 2014), such as:

- fixed when they enter the model.
- taken as a given in the model.
- not determined by the model.
- not explained by the model.
- They may influence endogenous variables

Forecasting Aspects

The structured or unstructured features of your forecasting model:

- Structured data comprises clearly defined data types whose pattern makes them easily **searchable**
- Unstructured data comprises data that is usually not as easily searchable, including formats like audio, video, and social media postings
- Structured data usually resides in relational databases

Forecasting Aspects

The univariate or multivariate nature of your forecasting model:

- **Univariate** data is characterized by a single variable. It does not deal with causes or relationships. Its descriptive properties can be identified in some estimates such as central tendency (mean, mode, median), dispersion (range, variance, maximum, minimum, quartile, and standard deviation), and the frequency distributions

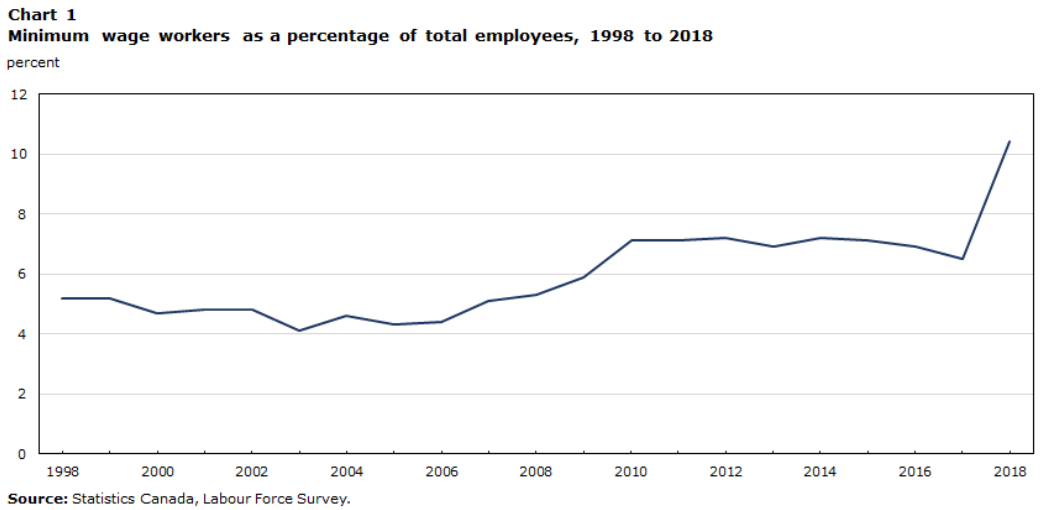
Types of Time Series

- **Univariate Time Series** – observing one variable at equally spaced time intervals

https://www150.statcan.gc.ca/n1/pub/75-004-m/75-004-m2019003-eng.htm

Types of Time Series

- **Univariate Time Series** – observing one variable at equally spaced time intervals

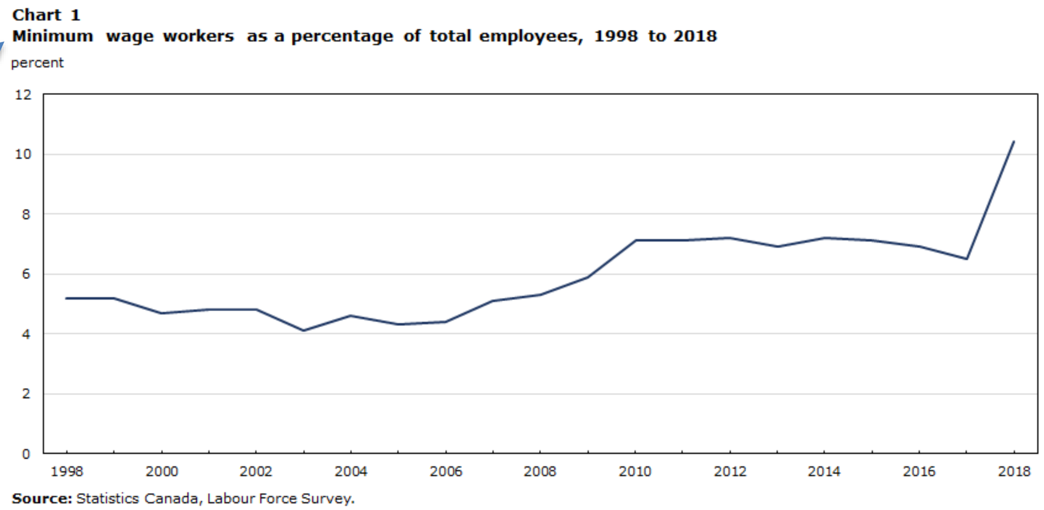


https://www150.statcan.gc.ca/n1/pub/75-004-m/75-004-m2019003-eng.htm

Types of Time Series

- **Univariate Time Series** – observing one variable at equally spaced time intervals

The single variable...



across equally spaced time intervals

https://www150.statcan.gc.ca/n1/pub/75-004-m/75-004-m2019003-eng.htm

Types of Time Series

- **Univariate Time Series** – observing one variable at equally spaced time intervals

Time	Software Price Index
January 2000	150.3
February 2000	149.8
March 2000	150.4
April 2000	151.3
May 2000	147.5

Types of Time Series

- Univariate data analysis is known for its limitation
- Insufficient in any analysis involving more than one variable

Types of Time Series

- **Multivariate Time Series** – observing multiple variables at equally spaced time intervals
- You may have multiple variables as input to predict only one of the variables as output

Types of Time Series

Multivariate time series essentially involves more than one input variable.

Types of Time Series

	Wed 11 am	Wed 12 pm	Wed 1 pm	Wed 2 pm	Wed 3 pm	Wed 4 pm
	A mix of sun and clouds	A mix of sun and clouds	A mix of sun and clouds	A mix of sun and clouds	Risk of a thunderstorm	Chance of a shower
	22°	23°	24°	24°	23°	23°
Feels like	22	23	24	24	23	23
POP	20 %	20 %	20 %	20 %	40 %	40 %
Rain	-	-	-	-	<1 mm	<1 mm
Wind (km/h)	26 s	25 s	24 s	23 s	22 s	22 s
Wind gust (km/h)	39	38	36	35	33	33

You may use multiple input variables to predict a single output variable (e.g. temperature).

https://www.theweathernetwork.com/us/weather/missouri/springfield

Types of Time Series

- **Multivariate Time Series** – observing multiple variables at equally spaced time intervals

Time	Potato Prices	Carrot Prices
January 2000	30.3	45.6
February 2000	39.8	44.9
March 2000	40.4	40.3
April 2000	31.3	39.9
May 2000	29.5	43.5

Forecasting Aspects

- *Single-step or multi-step structure of your forecasting model*
- **Single-step forecast**: predict the observation at the next time step.
- **Multi-step forecast**: predict a sequence of values in a time series

Multi-Step Forecasts

- Four commonly used strategies for making multi-step forecasts (Brownlee 2017):
  - Direct multi-step forecast
  - Recursive multi-step forecast
  - Direct-recursive hybrid multi-step forecast
  - Multiple output forecast

Multi-Step Forecasts

- *Direct multi-step forecast*: The direct method requires creating a separate model for each forecast time stamp

- *Recursive multi-step forecast*: a single time series model is created to forecast the next time stamp, and the following forecasts are then computed using previous forecasts

Multi-Step Forecasts

- *Direct-recursive hybrid multi-step forecast*: The direct and recursive strategies can be combined to offer the benefits of both methods (Brownlee 2017)

- *Multiple output forecast*: The multiple output strategy requires developing one model that is capable of predicting the entire forecast sequence. Example: ?

Multi-Step Forecasts

- *Direct-recursive hybrid multi-step forecast*: The direct and recursive strategies can be combined to offer the benefits of both methods (Brownlee 2017)

- *Multiple output forecast*: The multiple output strategy requires developing one model that is capable of predicting the entire forecast sequence. Example: a neural network with multiple output variables.

Forecasting Aspects

Contiguous or noncontiguous time series

- A time series that present a **consistent** temporal interval between each other (for example, every five minutes, every two hours, or every quarter).
- Time series that are **not uniform** over time may be defined as noncontiguous: very often the reason behind noncontiguous time series may be missing or corrupt values