CSC 735 – Data Analytics

Time Series Fundamentals

Reference

Lazzeri, Francesca, ``Machine Learning for Time Series Forecasting with Python'', Wiley, 2020.

Book Resources: https://github.com/FrancescaLazzeri/Machine-Learning-for-Time-Series-Forecasting

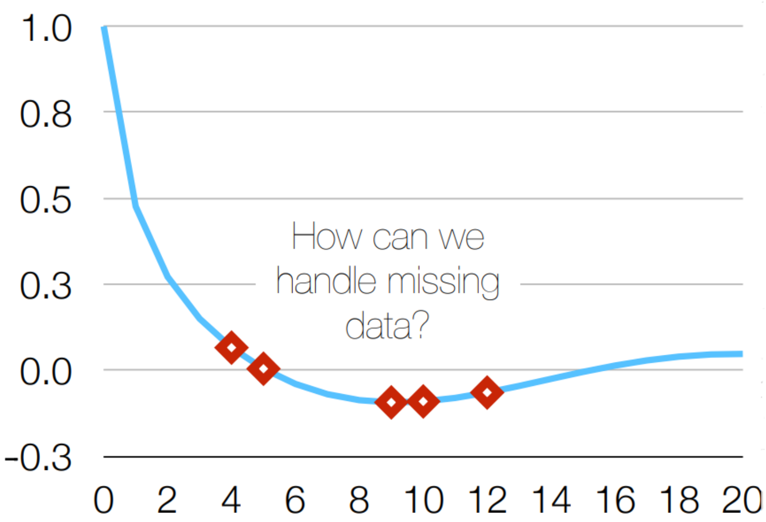
Forecasting Aspects

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

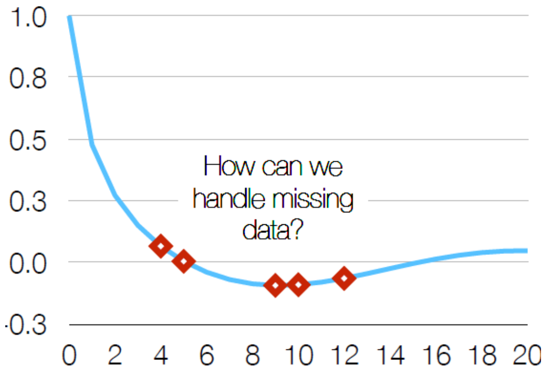
Missing Data

- Data points can be missing due to data corruption, data collection problems, etc.
- Missing values are often represented as NaN.



Missing Data

- Three common reasons for missing data include:
  - *Missing completely at random (MCAR)*
  - *Missing at random (MAR)*
  - *Missing not at random (MNAR)*



Missing Data

Scenario: A survey collects data on individuals' **income levels** and various demographic factors. However, missing data occurs for some participants' income information.

Missing Data

Missing completely at random: the missingness of data is unrelated to both observed and unobserved data. The missingness occurs randomly, without any systematic pattern; the probability of data being missing is the same for all observations.

Example: some participants accidentally skip the income question due to a formatting error in the survey software. The missingness is unrelated to any participant characteristics or their actual income values. It is completely random.

Missing Data

- ***Missing at random***: the propensity for a data point to be missing is not related to the missing data but it is related to some of the observed data.

- Example: participants with higher education levels are more likely to provide their income information, while those with lower education levels tend to skip the question due to a lack of confidence in reporting. Here, the missingness of income is related to the observed variable (education level), indicating a systematic pattern. However, the missingness is not directly related to the unobserved income values themselves.

https://stefvanbuuren.name/fimd/sec-MCAR.html

Missing Data

- ***Missing not at random***: the missingness of data is related to unobserved (missing) data values themselves. The missingness is non-random and can be influenced by factors that are not observed in the dataset.

- Example: participants with higher income are less likely to disclose their income information due to privacy concerns. The missingness of income is directly related to the unobserved values (higher income). This missingness pattern is MNAR as it depends on unobserved data, indicating a non-random pattern.

https://stefvanbuuren.name/fimd/sec-MCAR.html

Missing Data

- It is safe to remove data with missing values in the first two cases:
  - *Missing completely at random*
  - *Missing at random*

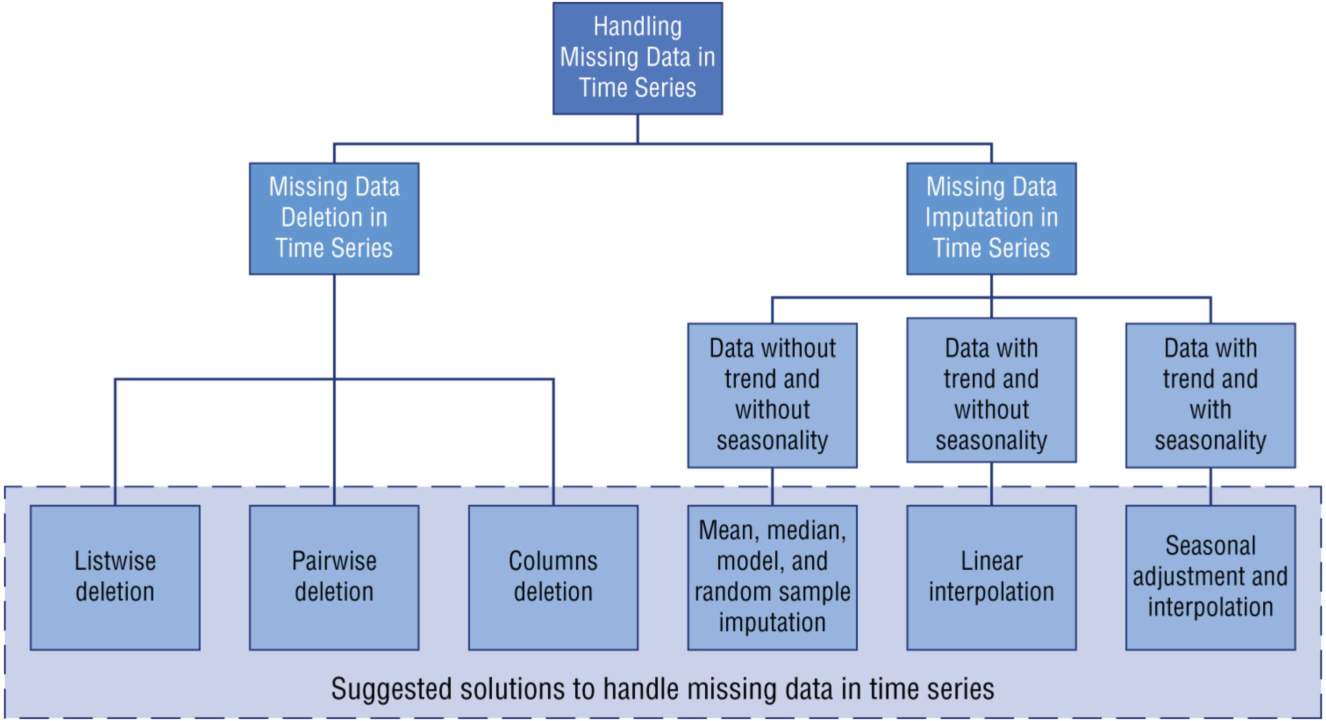
Missing Data

- If data is not missing at random, the removal of missing values could produce a bias in your forecasting model

Data Imputation

- **Data imputation** is where missing data is replaced with a specific value
- There are different solutions for data imputation depending on the problem you are solving

Handling Missing Data


Handling Missing Data in Time Series
Missing Data Deletion in Time Series
Missing Data Imputation in Time Series
Data without trend and without seasonality
Data with trend and without seasonality
Data with trend and with seasonality
Listwise deletion
Pairwise deletion
Columns deletion
Mean, median, model, and random sample imputation
Linear interpolation
Seasonal adjustment and interpolation
Suggested solutions to handle missing data in time series

Deletion in Time Series

- **Listwise deletion** removes all data for every observation that has one or more missing values
  - Not ideal in most cases
  - Listwise deletion methods produce biased parameters and estimates

Deletion in Time Series

- **Pairwise deletion** analyzes all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis
  - If you delete pairwise, then you'll end up with different numbers of observations contributing to different parts of your model, which can make interpretation difficult

Deletion in Time Series

- **Deleting a column** is another option
- Based on the idea that there is no remaining value left in the values in the column (perhaps because too many are missing)
- Generally, it is better to keep data than to discard it

Imputation in Time Series

- **Mean, median, and mode:** Computing the overall mean, median, or mode is a very **basic** imputation method; it is the only tested function that takes no advantage of the time series characteristics or relationship between the variables.
- It is fast
- mean imputation is suitable for variables with a roughly symmetric distribution

Imputation in Time Series

- **Linear interpolation:** This method works well for a time series with some trend but is not suitable for seasonal data

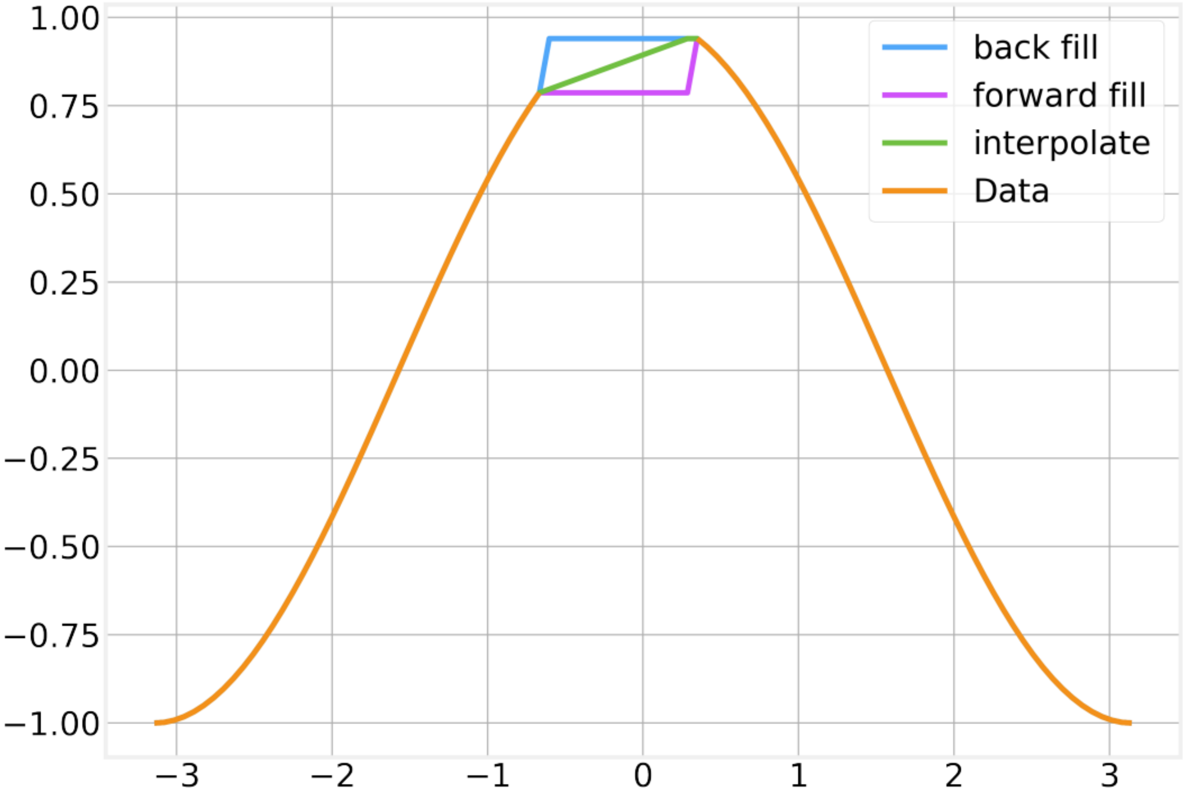
Imputation in Time Series

- **Seasonal adjustment and linear interpolation:** This method works well for data with both trend and seasonality

Imputation Methods

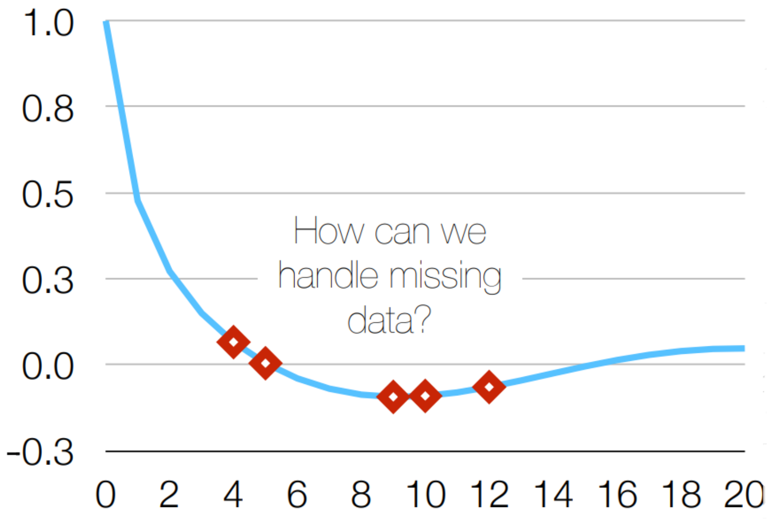
- **data imputation methods** (**data filling methods**): any technique for adding values based on what we expect the missing values to be
  - **Forward fill** – substitute the last valid value
  - **Back fill** – substitute the next valid value
  - **Interpolate** – add values by interpolating between the previous and the next value (can be linear interpolation, cubic interpolation, etc.)

Fill Methods



Multiple Imputation Methods

- Use an imputation technique that draws from an appropriate distribution to choose a value to substitute for any missing value
- Repeat *m* times giving *m* complete data sets
- Perform the analysis *m* times
- Average the overall results from the *m* analyses to get the final result.



Time Series Models

- We will cover the following time series models:

1. Persist
2. Exponential Smoothing (Exponential Running Average)
3. Moving Average (MA)
4. Autoregressive (AR)
5. Autoregressive Integrated Moving Average (ARIMA)

Persist Model

- A simple time series model in the context of univariate time series
- Recall that a univariate time series observes one variable at equally spaced time intervals
- Univariate models are easy to understand since they do not deal with inter-variable relationships

Persist Model

- The **Persist model** uses the last value of a time series to predict the next value
- This model assumes that the next value of the series will be the same as the previous value of the series

Persist Model

Date	End of Day Gas Price in Springfield ($)
2022-06-10	1.85
2022-06-11	1.90
2022-06-12	1.85
2022-06-13	1.87
2022-06-14	?

1.87

For example, the Persist Model predicts that the next day's gas price will be the same as the previous day's gas price.

NOTE: This is synthetic data for illustration purposes.

Persist Model

- The **Persist model's** predicted value will lag by 1 data point for all data points
- Recall that the value of the time series at time t is given by x_t and represented as: $x_1, x_2, x_3, \dots$
- Let x'_t represent a predicted value at time t
- Thus, the equation for the Persist model is represented as follows:

$$x'_t = ?$$

Persist Model

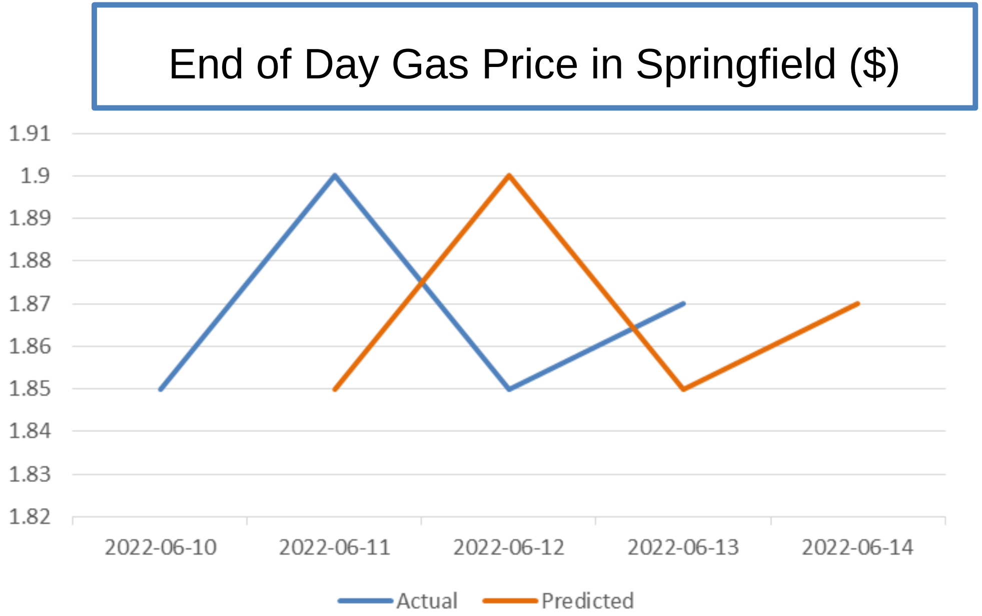
- The **Persist model's** predicted value will lag by 1 data point for all data points
- Recall that the value of the time series at time t is given by x_t and represented as: $x_1, x_2, x_3, \ldots$
- Let x'_t represent a predicted value at time t
- Thus, the equation for the Persist model is represented as follows:

$$x'_t = x_{t-1}$$

Persist Model

End of Day Gas Price in Springfield ($)

t	Date	Actual x_t	Predicted $x'_t = x_{t-1}$
1	2022-06-10	1.85	-
2	2022-06-11	1.90	1.85
3	2022-06-12	1.85	1.90
4	2022-06-13	1.87	1.85
5	2022-06-14	?	1.87



NOTE: This is synthetic data for illustration purposes.

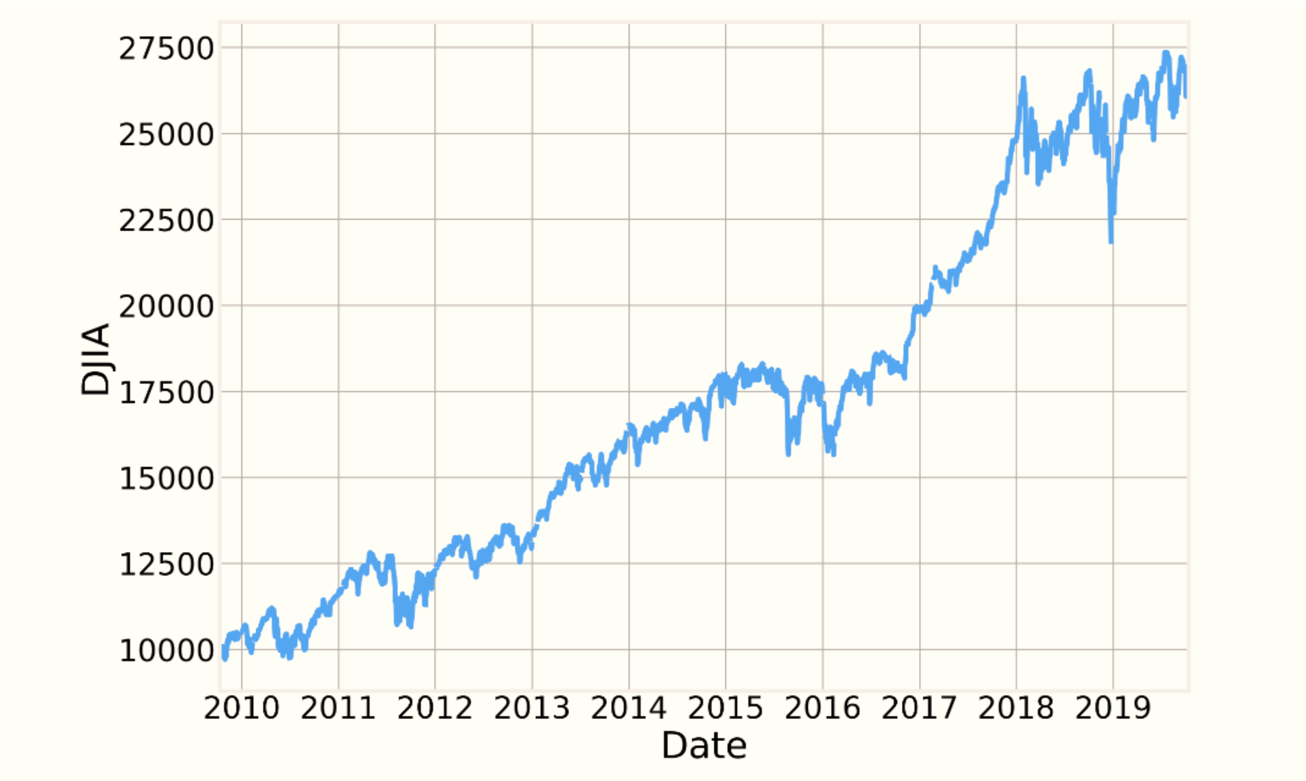
Time Series Review

- A set of values measured **sequentially** in time
- Values are typically (but not always) measured at **equal intervals**, $x_1, x_2, x_3, x_4,$ etc.
- Values can be:
  - **continuous**
  - **discrete** or **symbolic** (words)
- Associated with empirical observations of time-varying phenomena:
  - Stock market prices (day, hour, minute, tick, etc.)
  - Temperature (day, minute, second, etc.)
  - Number of patients (week, month, etc.)
  - GDP (quarter, year, etc.)
- Forecasting requires predicting future values based on past behavior

Mathematical Conventions

- The value of the time series at time $\boldsymbol{t}$ is given by $\boldsymbol{x_t}$
- The values at a given lag $\boldsymbol{l}$ are given by $\boldsymbol{x_{t-l}}$
- The mean of the overall signal is $\boldsymbol{\mu}$ and the corresponding running value is $\boldsymbol{\mu_t^w}$
- The standard deviation of the overall signal is $\boldsymbol{\sigma}$ and the running value is $\boldsymbol{\sigma_t^w}$
- **Running values** are calculated over a window of width $\boldsymbol{w}$
- A time series is represented by the set $\boldsymbol{X}$
- A forecasted time series is represented by the set $\boldsymbol{X'}$
- A forecasted value of the time series at time t is given by $\boldsymbol{x'_t}$

Stock Market – Dow Jones Industrial Average (DJIA)



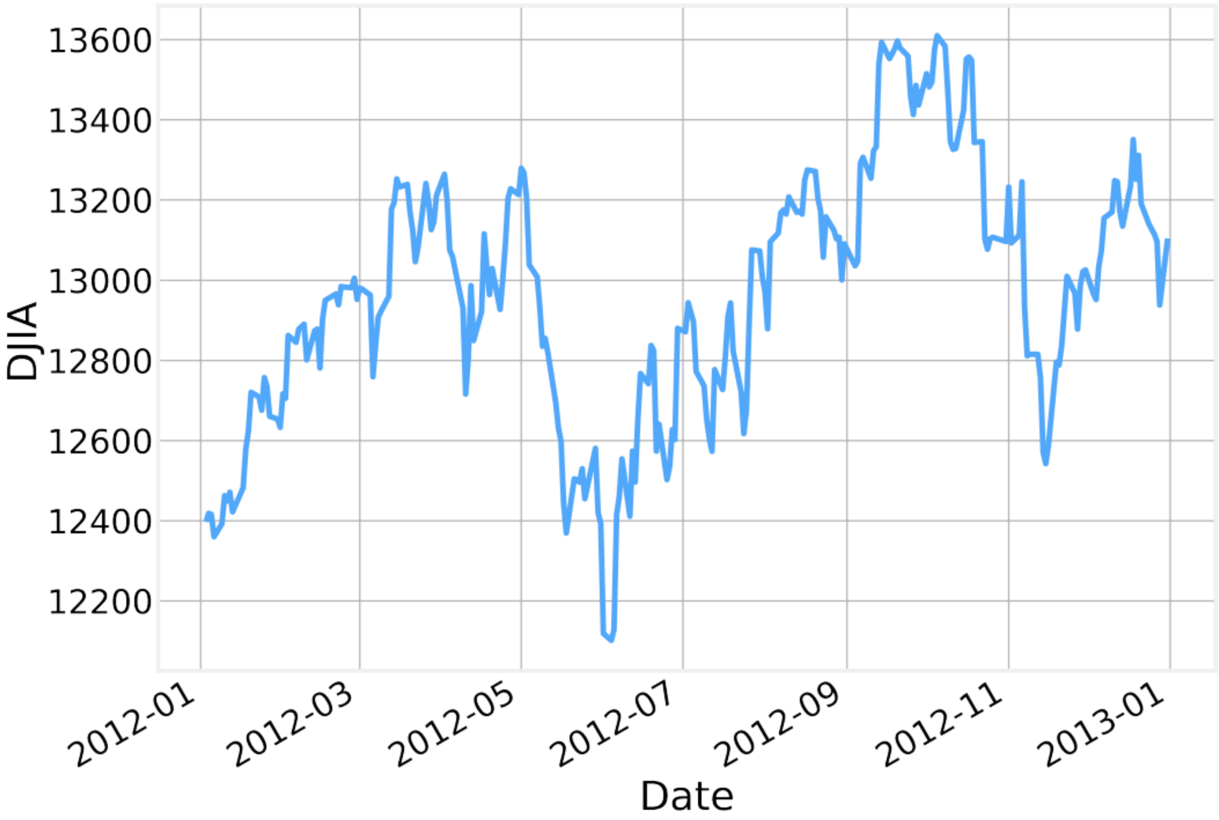
Lagged Values

- While analyzing time series, we often refer to values that our time series took $\mathbf{1}, \mathbf{2}, \mathbf{3}$, etc. time steps in the past
- These are known as **lagged values** and are denoted by:

$$\boldsymbol{x_{t-l}}$$

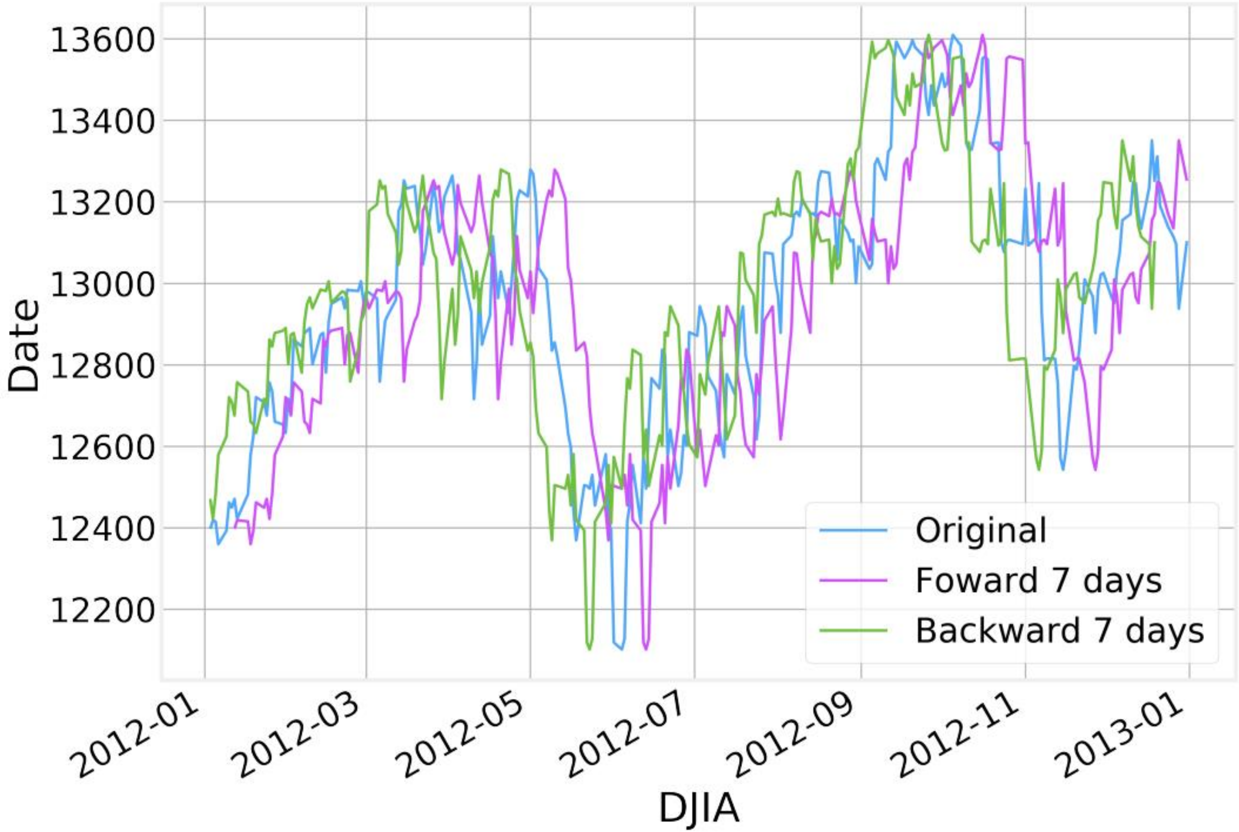
where $\boldsymbol{l}$ is the value of the lag we are considering

Lagged Values



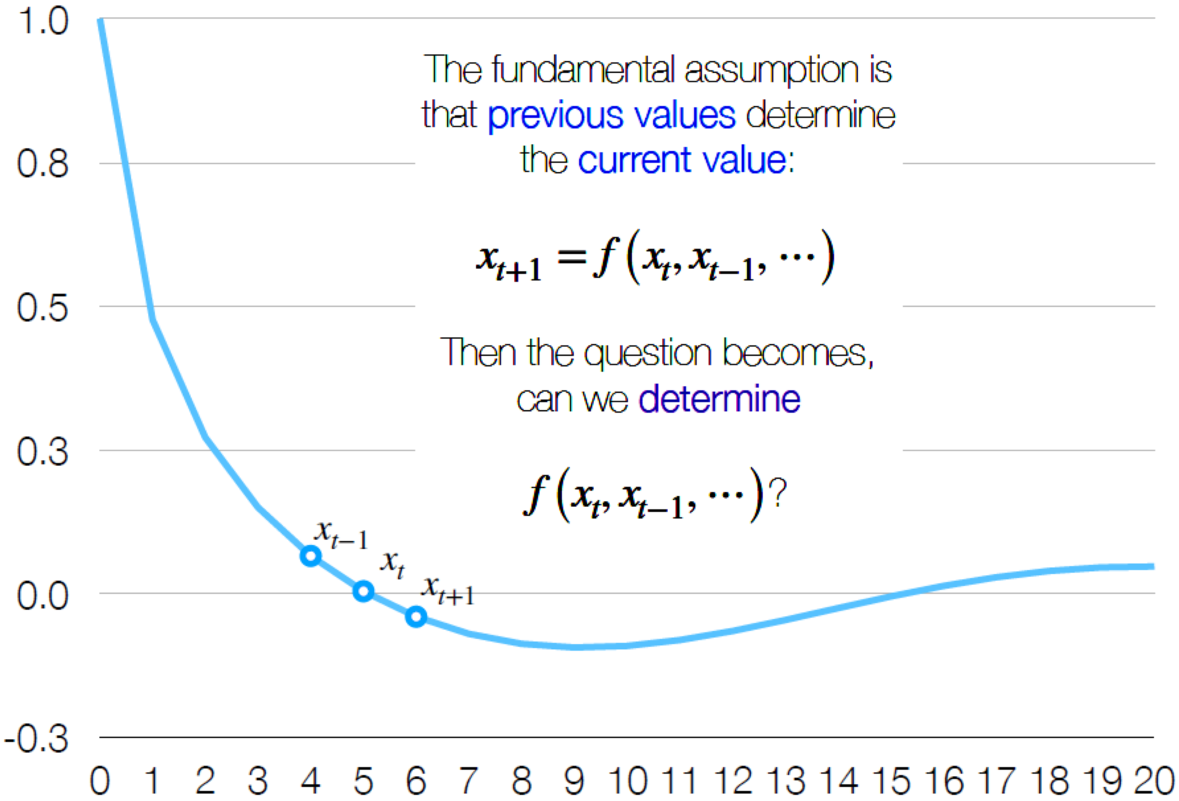
Consider the DJIA variable that gives the average stock price

Lagged Values



Imagine defining two new variables such that their values for some time *t* were identical to our variable’s values at time *t*+7 or time *t*–7

Time Series Analysis: Recall the Goal



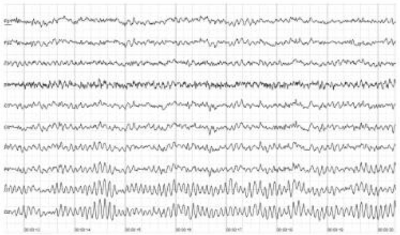
The fundamental assumption is that previous values determine the current value:

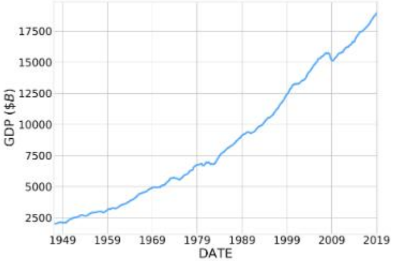
$$x_{t+1} = f\left(x_t, x_{t-1}, \cdots\right)$$

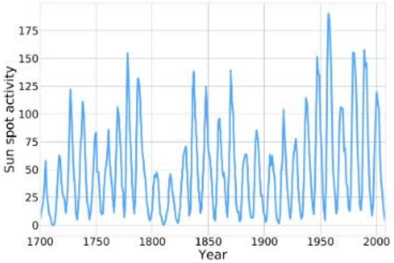
Then the question becomes, can we determine

$$f\left(x_t, x_{t-1}, \cdots\right)?$$

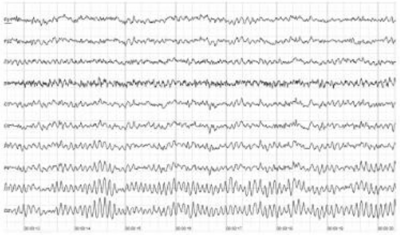
Three Fundamental Behaviours





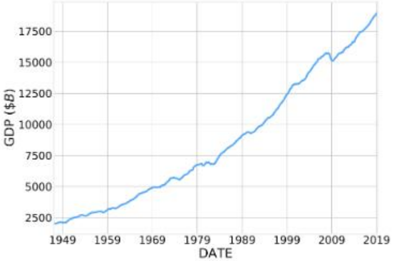


Three Fundamental Behaviours



Stationarity

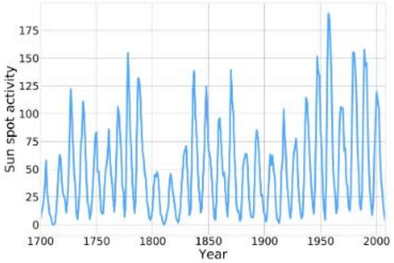
$$\langle x_t \rangle \approx constant$$



Trend

$$\langle x_t \rangle \approx ct$$

(continuous increase or decrease)



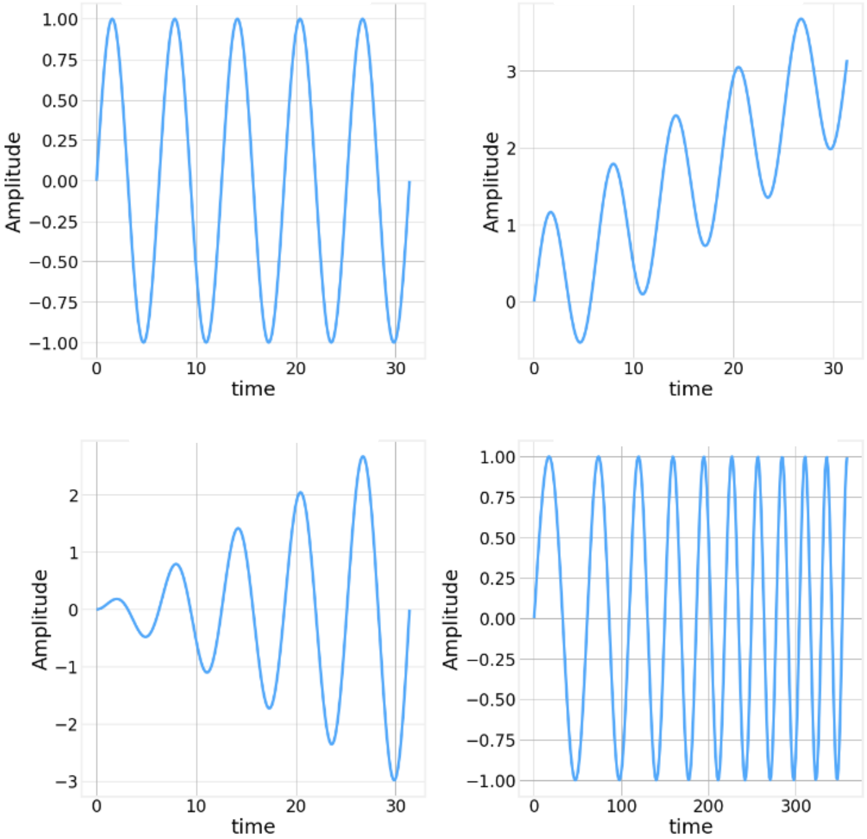
Seasonality

$$x_{t+T} \approx x_t$$

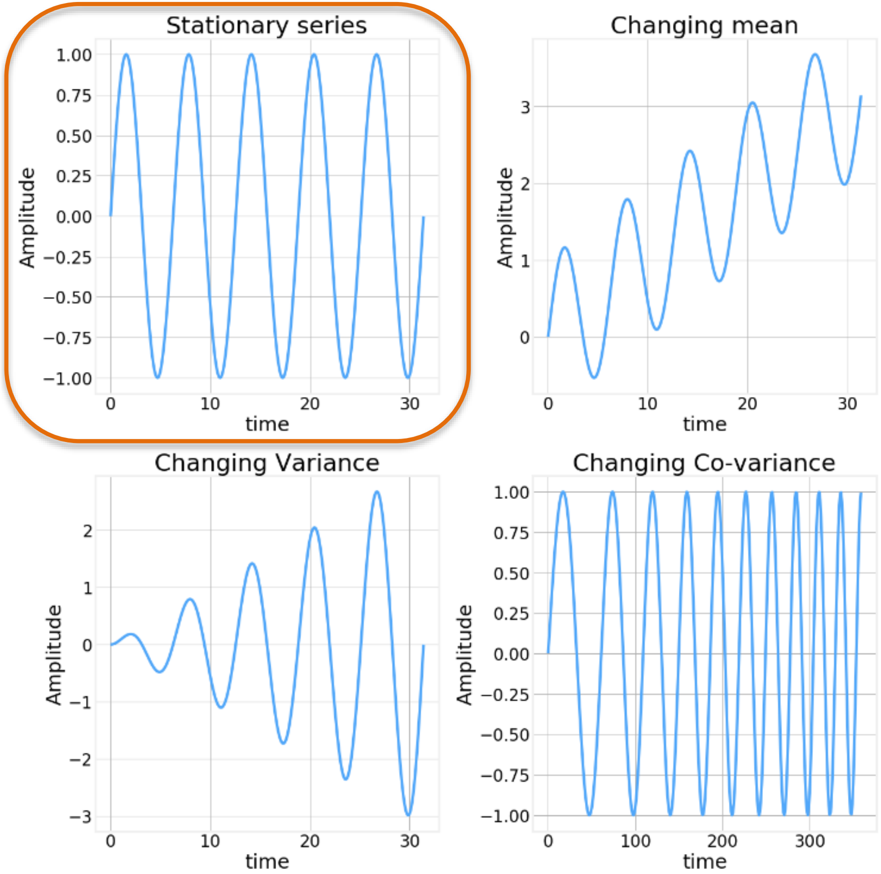
Stationarity

- A time series is said to be stationary if its basic statistical properties are **independent of time**
- In particular:
  - **Mean** – Average value stays constant
  - **Variance –** The width of the curve is bounded
  - **Covariance** – Correlation between points is independent of time
- Stationary processes are **easier** to analyze
- Many time series analysis algorithms assume the time series to be stationary
- Several rigorous tests for stationarity have been developed such as the **(Augmented) Dickey-Fuller** and **Hurst Exponent** tests
- Typically, the first step of any analysis is to transform the series to make it stationary

Stationarity

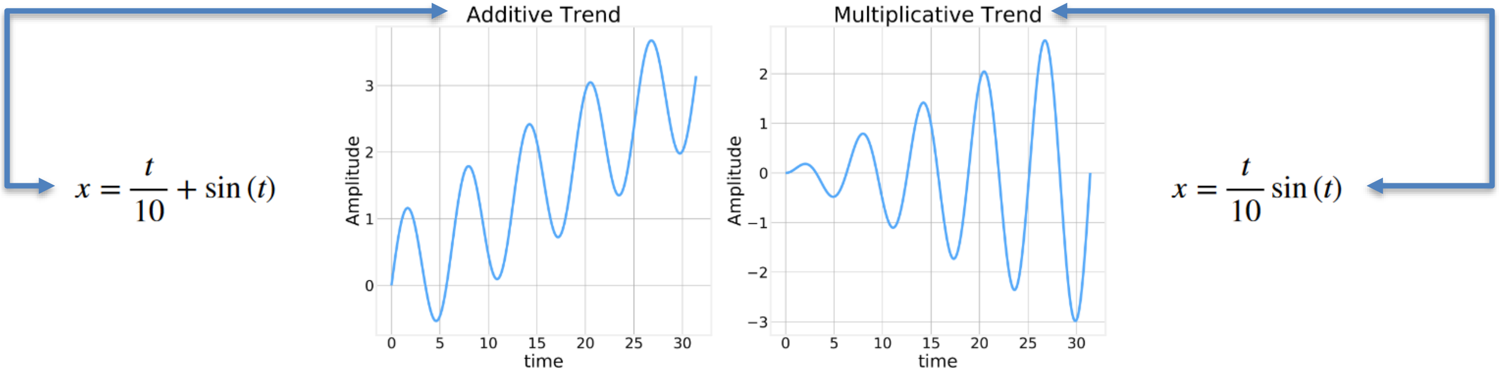


Stationarity



Trend

- Many time series have a clear trend or tendency:
  - Stock market indices tend to go up over time
  - Number of cases of preventable diseases tends to go down over time
- Trend can be **additive** or **multiplicative**:



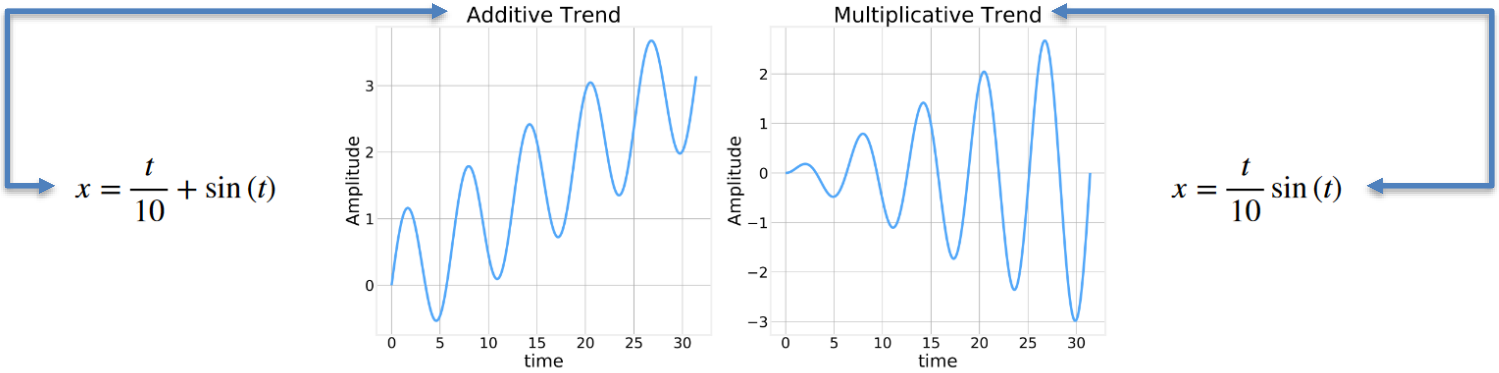
$$x = \frac{t}{10} + \sin(t)$$

$$x = \frac{t}{10} \sin(t)$$

- Such trends can be removed by **subtraction** or **division** of the correct values
- One simple way to determine the trend is to calculate a **running average** over the series
- $m_t = (x_{t-k+1} + x_{t-k+2} + \ldots + x_t)/k.$

Trend

- Many time series have a clear trend or tendency:
    - Stock market indices tend to go up over time
    - Number of cases of preventable diseases tends to go down over time
- Trend can be **additive** or **multiplicative**:

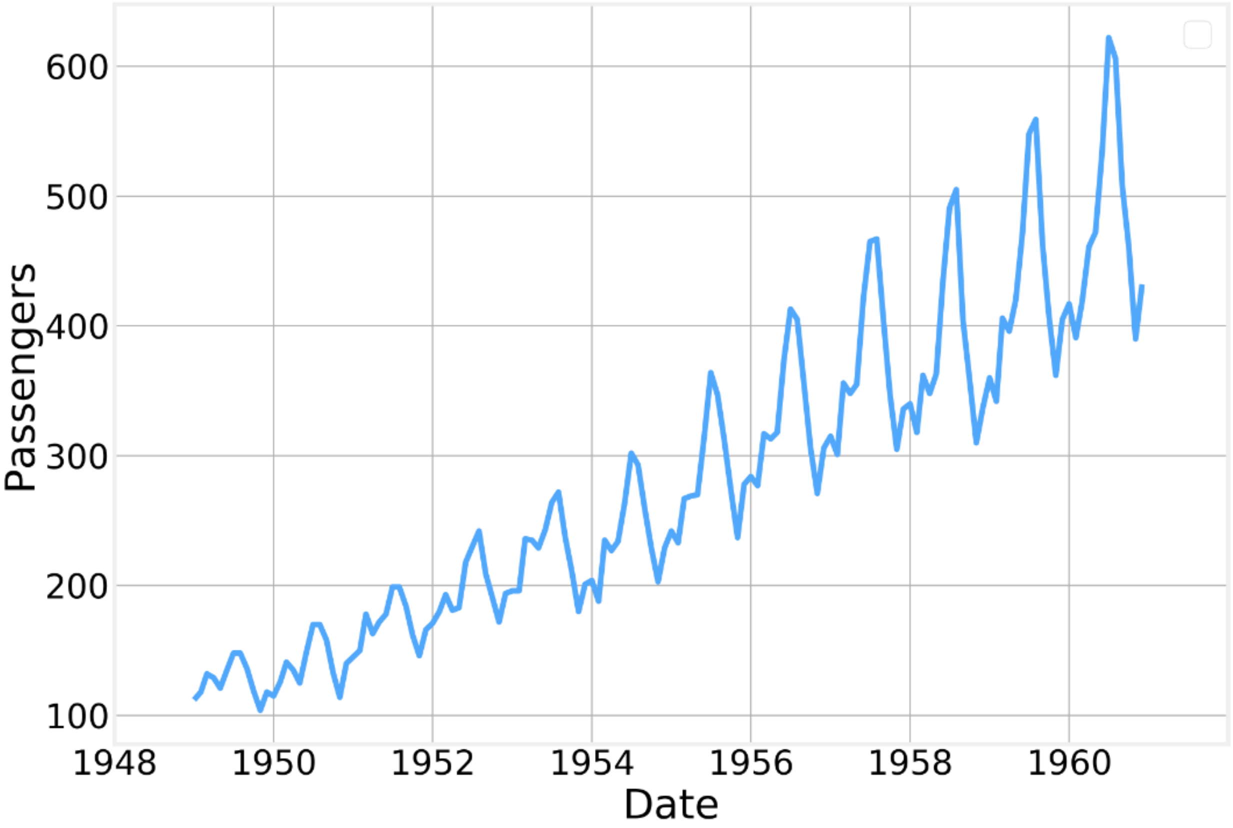


$$x = \frac{t}{10} + \sin(t)$$

$$x = \frac{t}{10}\sin(t)$$

- Such trends can be removed by **subtraction** or **division** of the correct values
- One simple way to determine the trend is to calculate a **running average** over the series
- $m_t = (x_{t-k+1} + x_{t-k+2} + \ldots + x_t)/k$. If $t = 16$, $k = 3$, then m_{16} = (x_{14} + x_{15} + x_{16}) / 3

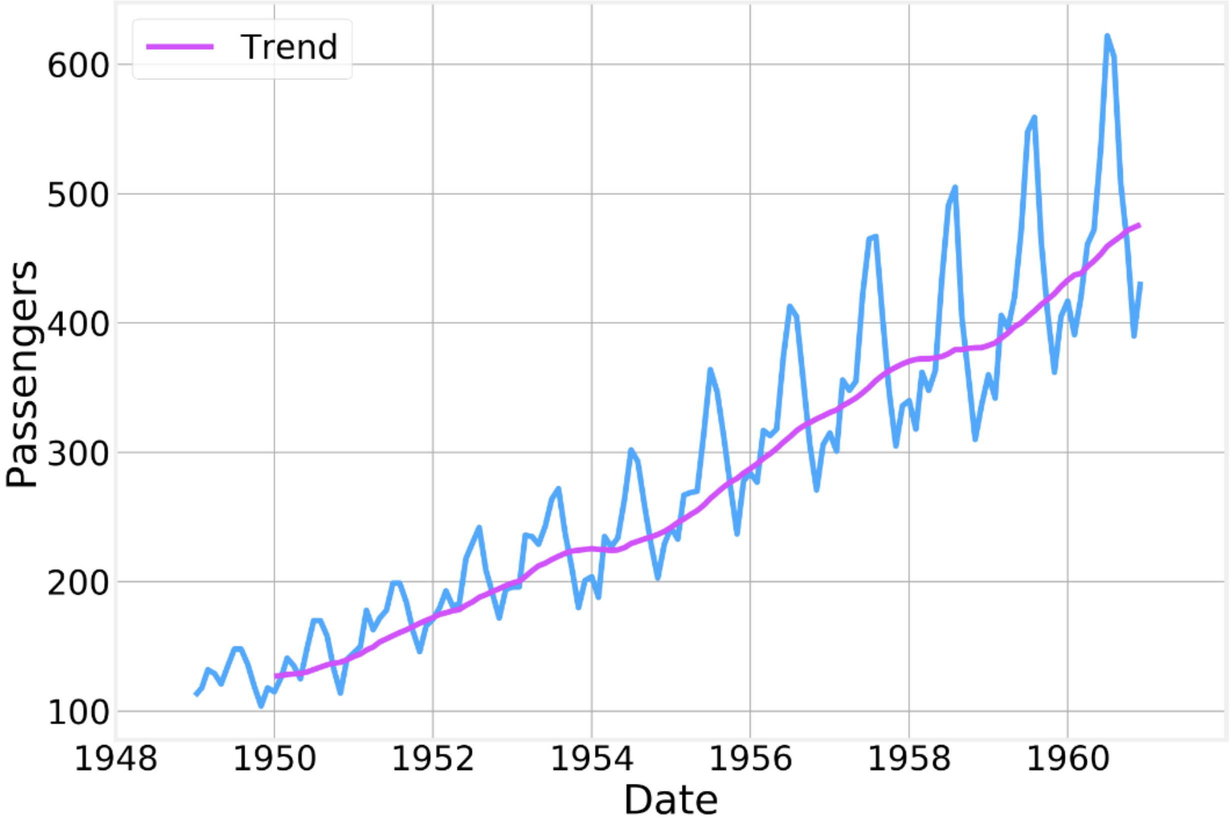
Trend



Consider this graph of the number of air traffic passengers

https://www.kaggle.com/datasets/chirag19/air-passengers

Trend



If we calculate the running average, we readily see the trend.

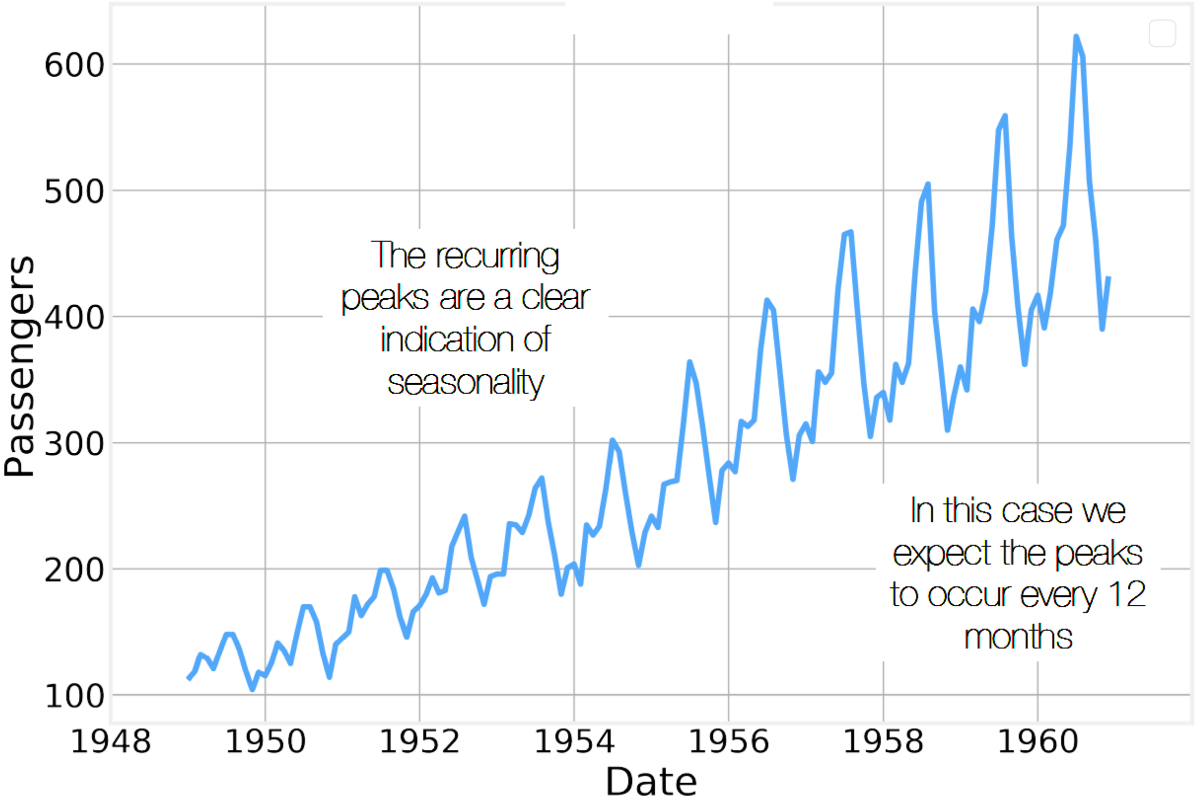
https://www.kaggle.com/datasets/chirag19/air-passengers

Seasonality

- Many of the phenomena we might be interested in, vary in time in a **seasonal** or **cyclical** fashion:
  - Ice-cream sales peak in the **summer** and drop in the **winter**
  - Number of cell phone calls made is larger during the day than during the night
  - Many types of crime are more frequent at **night** than during the **day**
  - Visits to museums are more frequent on **weekend** days than on **weekdays**
  - The stock market generally grows during **bull** periods and shrinks during **bear** periods
- Understanding the seasonality of a time series provides important information about its long-term behaviour and is extremely useful in predicting future values
- If the period is **fixed** it’s called seasonality, while if the period is **irregular** it’s called cyclical

Seasonality

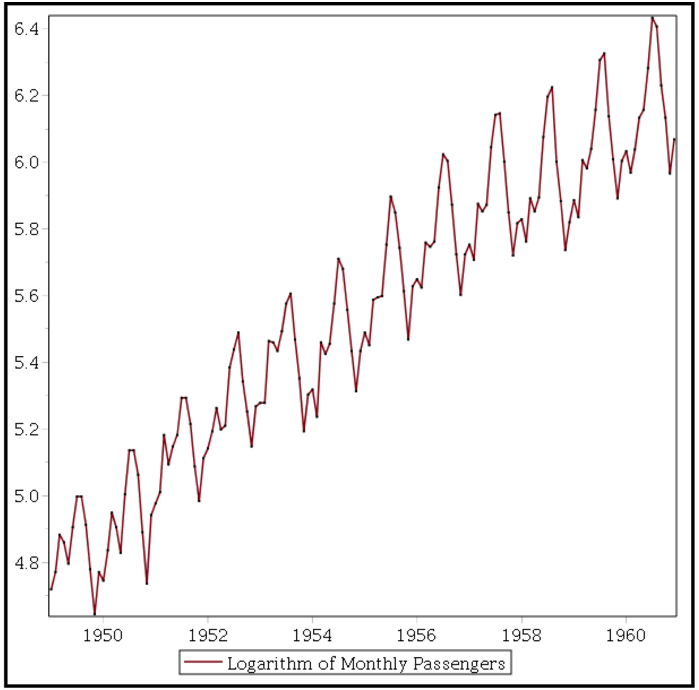
https://www.kaggle.com/datasets/chirag19/air-passengers



The recurring peaks are a clear indication of seasonality

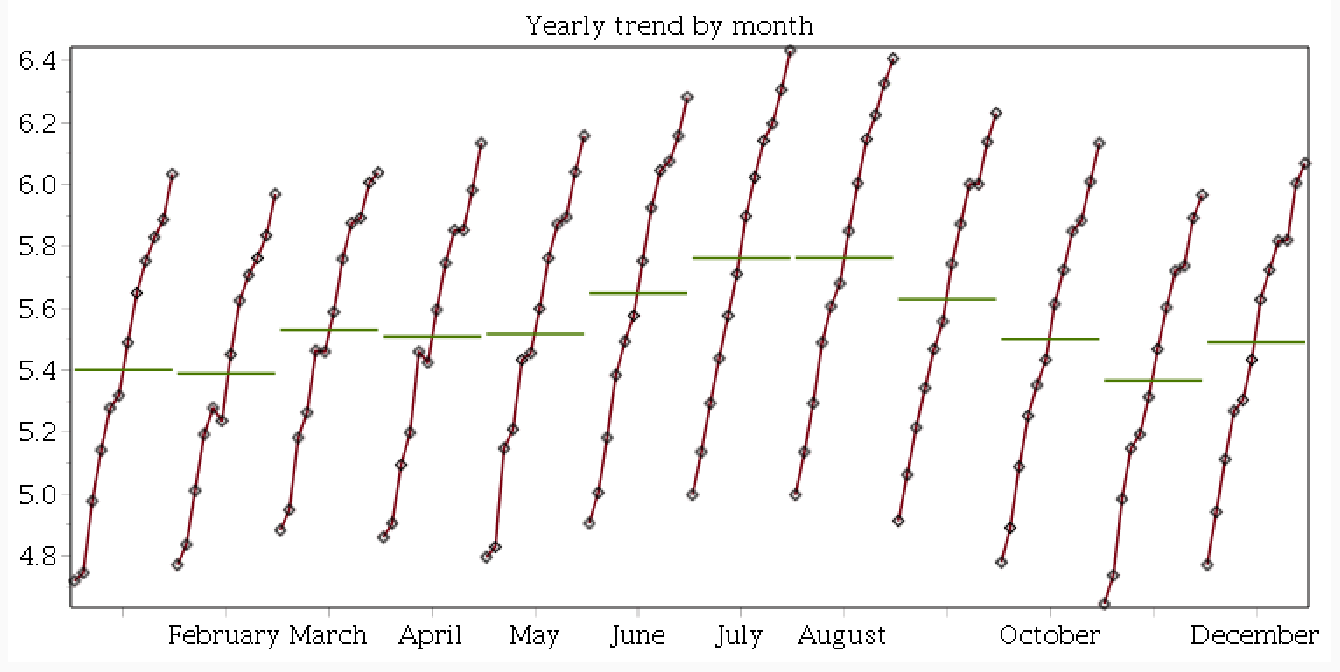
In this case we expect the peaks to occur every 12 months

Logarithm of Air Passengers



https://www.maplesoft.com/support/help/maple/view.aspx?path=examples/AirPassengers

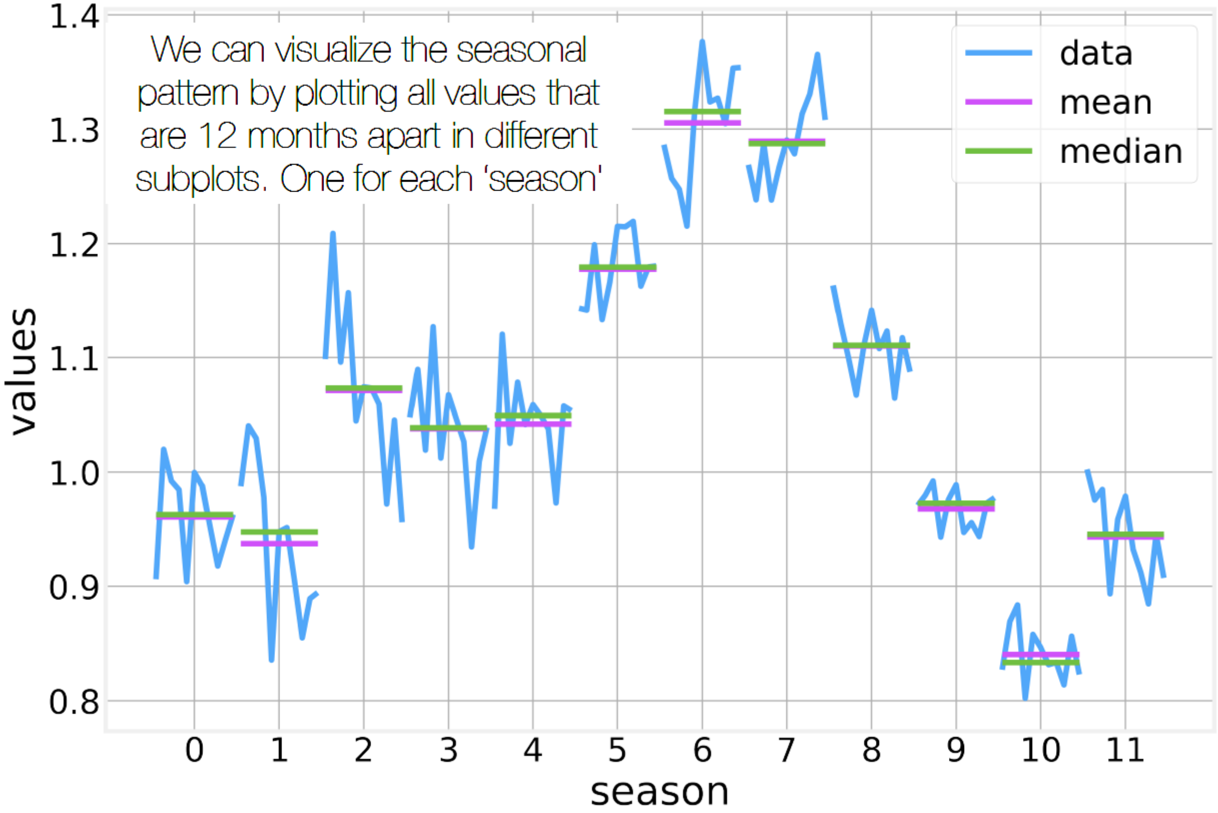
Seasonality of Logarithms of Air Passengers



Here a season is one month long.

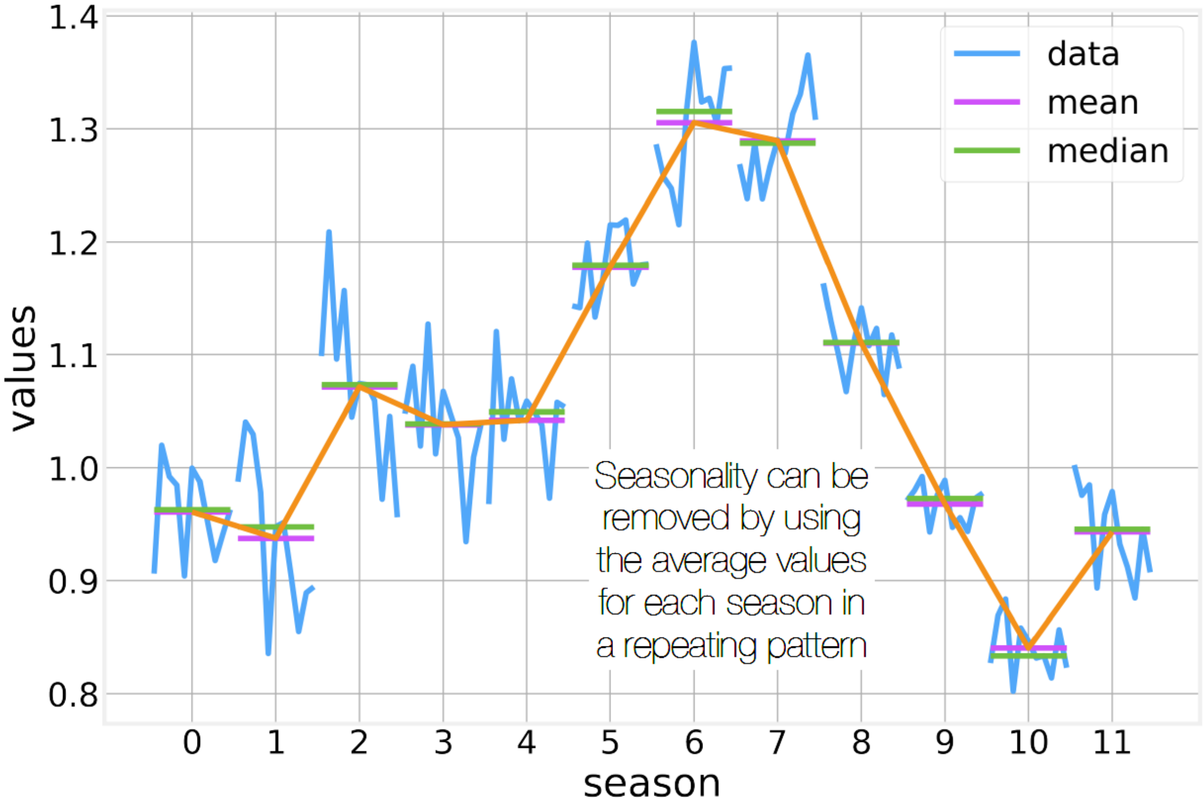
https://www.maplesoft.com/support/help/maple/view.aspx?path=examples/AirPassengers

Seasonality



Seasonality of differences of logarithms

Seasonality

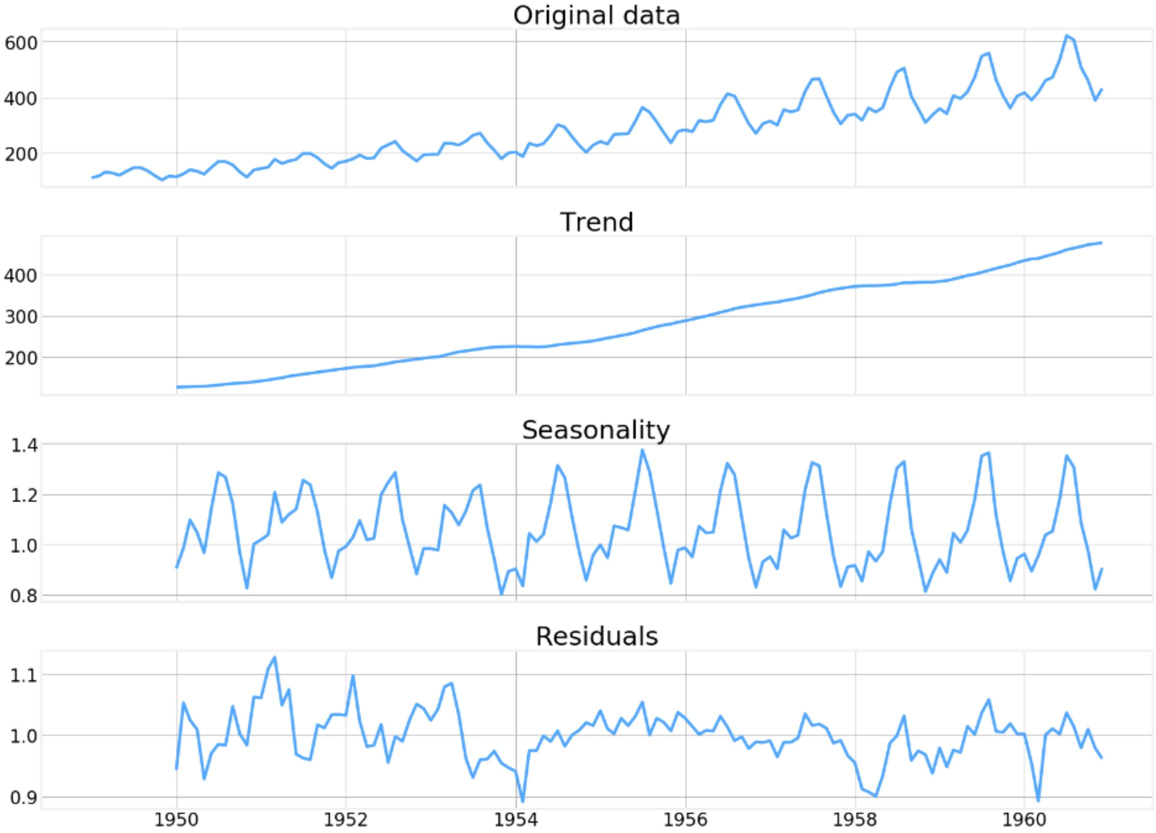


Seasonality can be removed by using the average values for each season in a repeating pattern

Time Series Decomposition

- A time series can be decomposed into **three components**:
  - Trend, $\boldsymbol{T_t}$
  - Seasonality, $\boldsymbol{S_t}$
  - Residuals, $\boldsymbol{R_t}$
- Decompositions can be:
  - additive: $\boldsymbol{x_t = T_t + S_t + R_t}$
  - multiplicative: $\boldsymbol{x_t = Tt \bullet St \bullet R_t}$
- The residuals are simply what is left of the original signal after we **remove the trend and the seasonality**
- Residuals are typically **stationary**

Time Series Decomposition



Differences

- Perhaps the most common use case for lagged values is for the calculation of **differences** of the form:

$$\boldsymbol{x_t - x_{t-l}}$$

where $\boldsymbol{l \geq 1}$ is the value of the lag we are interested in

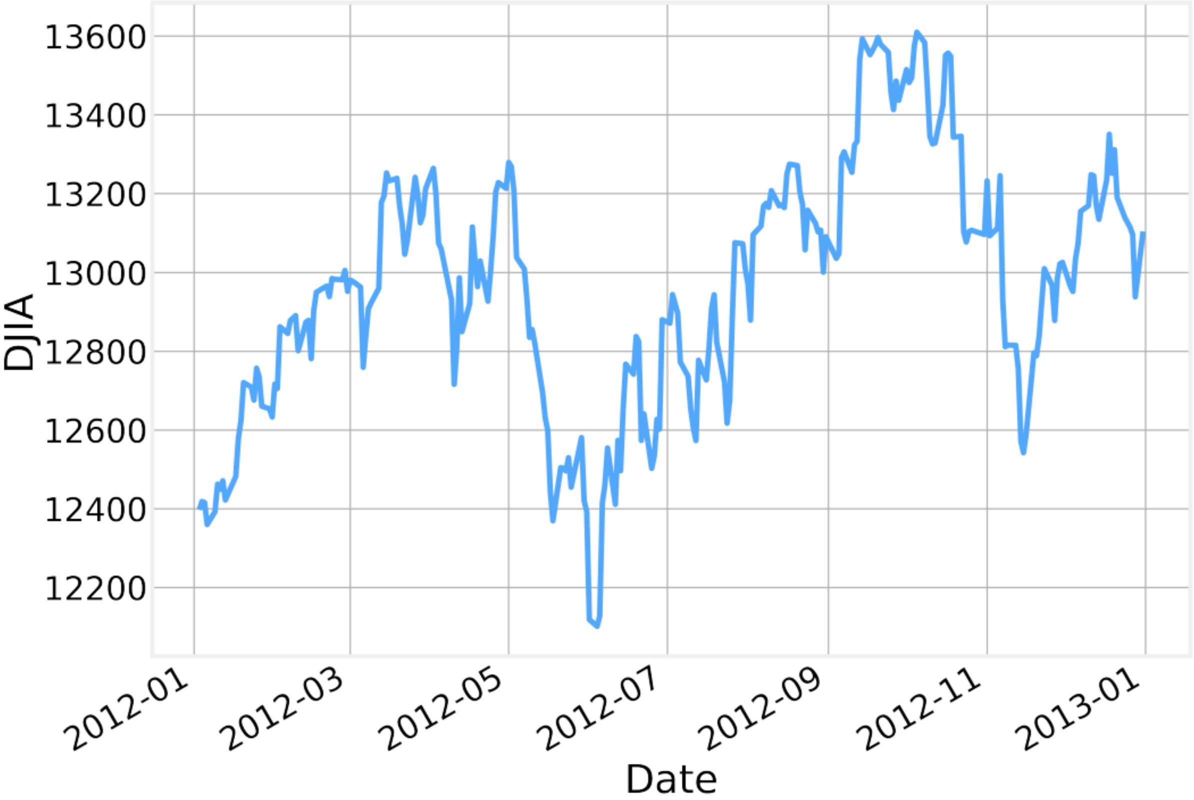
- Naturally, higher-order differences can also be used, in which case, the difference of the difference is calculated:

$$\boldsymbol{y_t = x_t - x_{t-l}}$$

$$\boldsymbol{z_t = y_t - y_{t-l} \equiv x_t - 2x_{t-l} + x_{t-2l}}$$

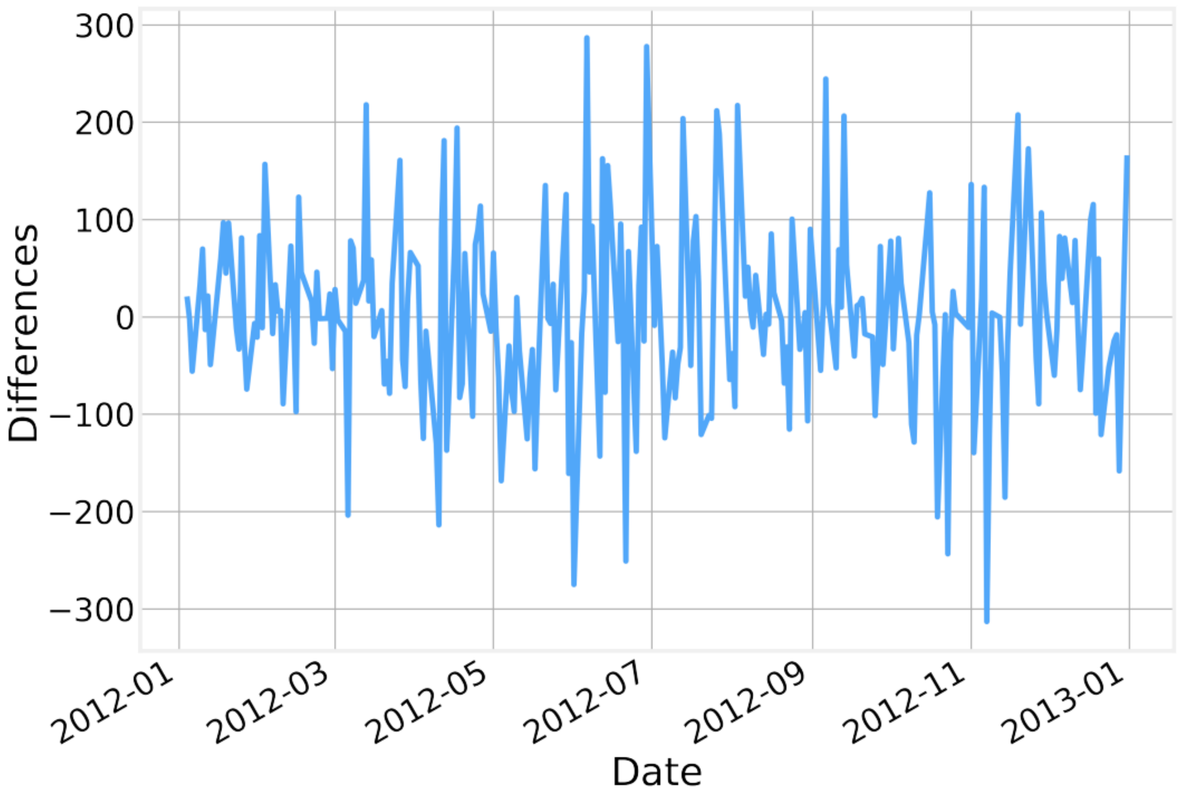
- This can be thought of as a discrete version of the usual derivative of a function
- Differences are also a particularly simple way to **detrend** a time series

Differences



Consider again the DJIA variable that gives the average stock price

Differences



The differences look more like a stationary series.