Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 10
Time Series Analysis

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Time series data

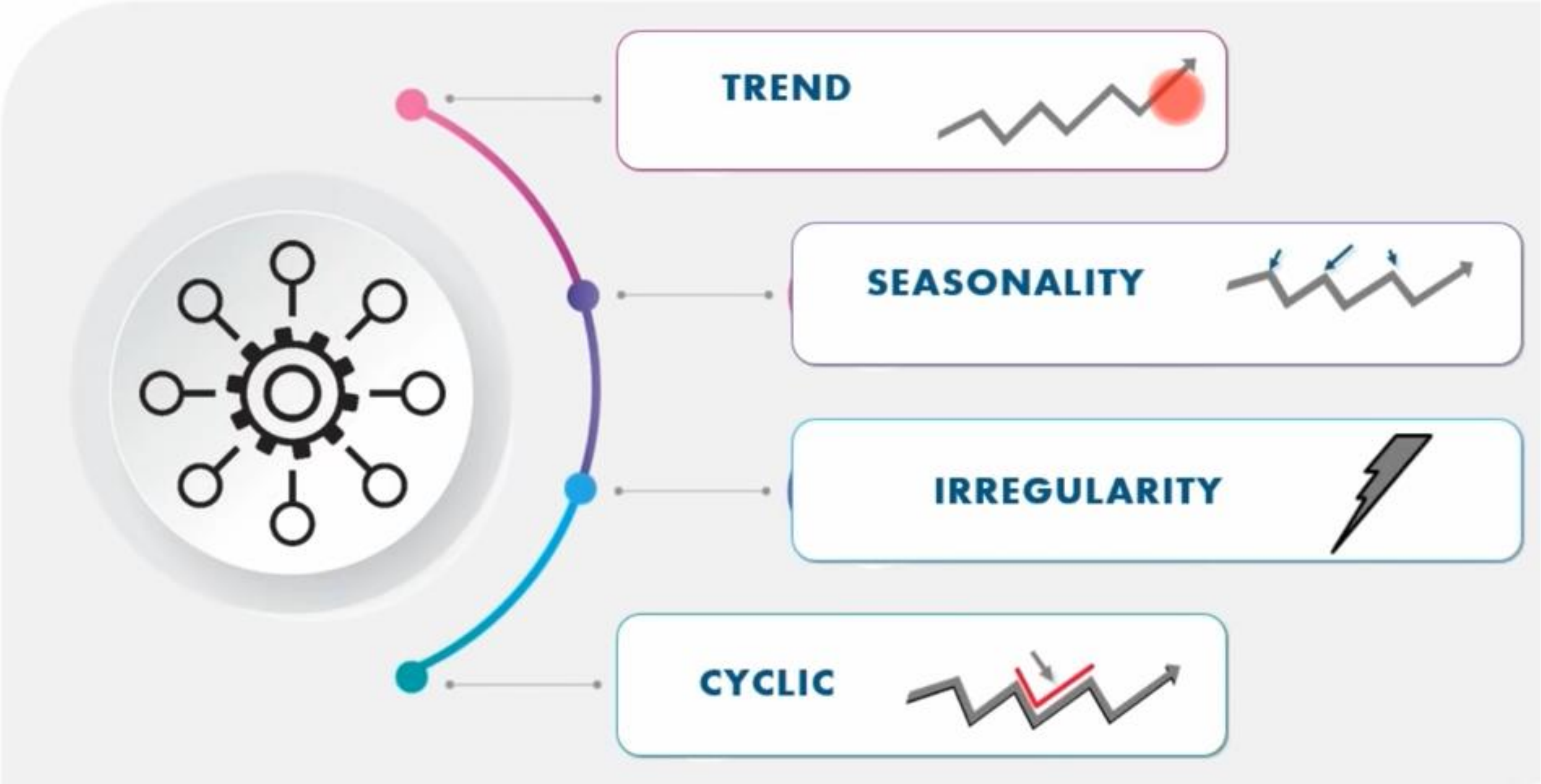
Time series data – an ordered sequence of values of a variable at equally spaced time intervals. For example, daily exchange rates, the average daily stock price of a company. Time Series Analysis is often focused on identifying underlying trends and patterns, describing them mathematically, and ultimately making a prediction or forecast about what will happen next.

Time series modelling allows to replicate every element of the process by decomposing the mathematical process into a combination of signals (e.g. year-on-year growth, seasonal variability, etc.) and noise (random probabilistic processes), without necessarily knowing the underlying causes for each. Time series decomposition involves thinking of a series as a combination of trend, seasonality, cycles and noise components.

Time series data

Field	Example Topics
Economics	Gross Domestic Product (GDP), Consumer Price Index (CPI), S&P 500 Index, and unemployment rates
Social Sciences	Birth rates, population, migration data, political indicators
Epidemiology	Disease rates, mortality rates, mosquito populations
Medicine	Blood pressure control, weight management, cholesterol measurements, heart rate monitoring
Physics	Global temperatures, monthly observations of sunspots, pollution levels.

Time Series Components



Time Series Components

- **Trend** – overall long-term direction of the series.
- **Seasonality** – repeated behavior in the data which occurs at regular intervals. Is related to seasonal natural or human behavior. For example, prices for agricultural products in summer are higher than in winter, the unemployment rate in resort towns in winter is higher than in summer.
- **Cycles** occur when a series follows up-and-down pattern that is not seasonal. The cycle can be of varying length which make it more difficult to detect than seasonality. For example, product life cycle, economic waves, solar activity.
- **Unexplained** (or Random) **Variation** exists in most time series. It is the component of time series that is obtained after the first three patterns have been 'extracted' out of the series. Some time series could be rather regular with little random variation.

Trend

Example 1: Downward sloping Curve

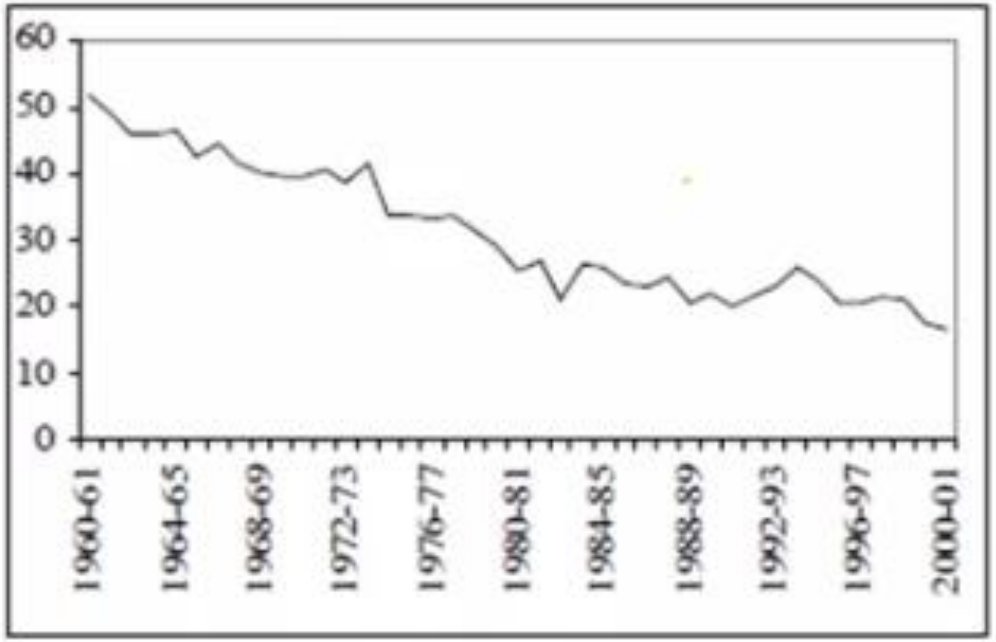


Figure 1.1 Share of agriculture in GSDP for the state of Tamilnadu

Trend and Seasonality

Example 2: Seasonal pattern around an upward sloping trend

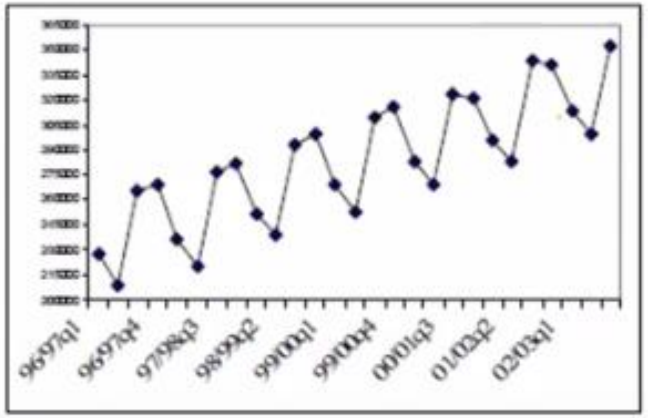


Figure 1.2 Quarterly GDP series for India

Example 3: Seasonal pattern around quadratic trend (Fig 1.3)
Seasonal pattern around linear trend (Fig 1.4)

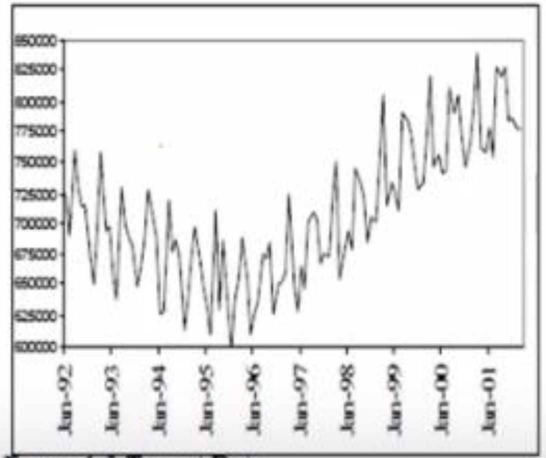


Figure 1.3 Transit Data

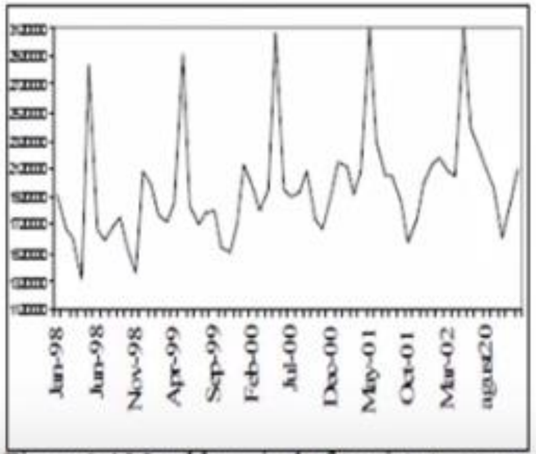
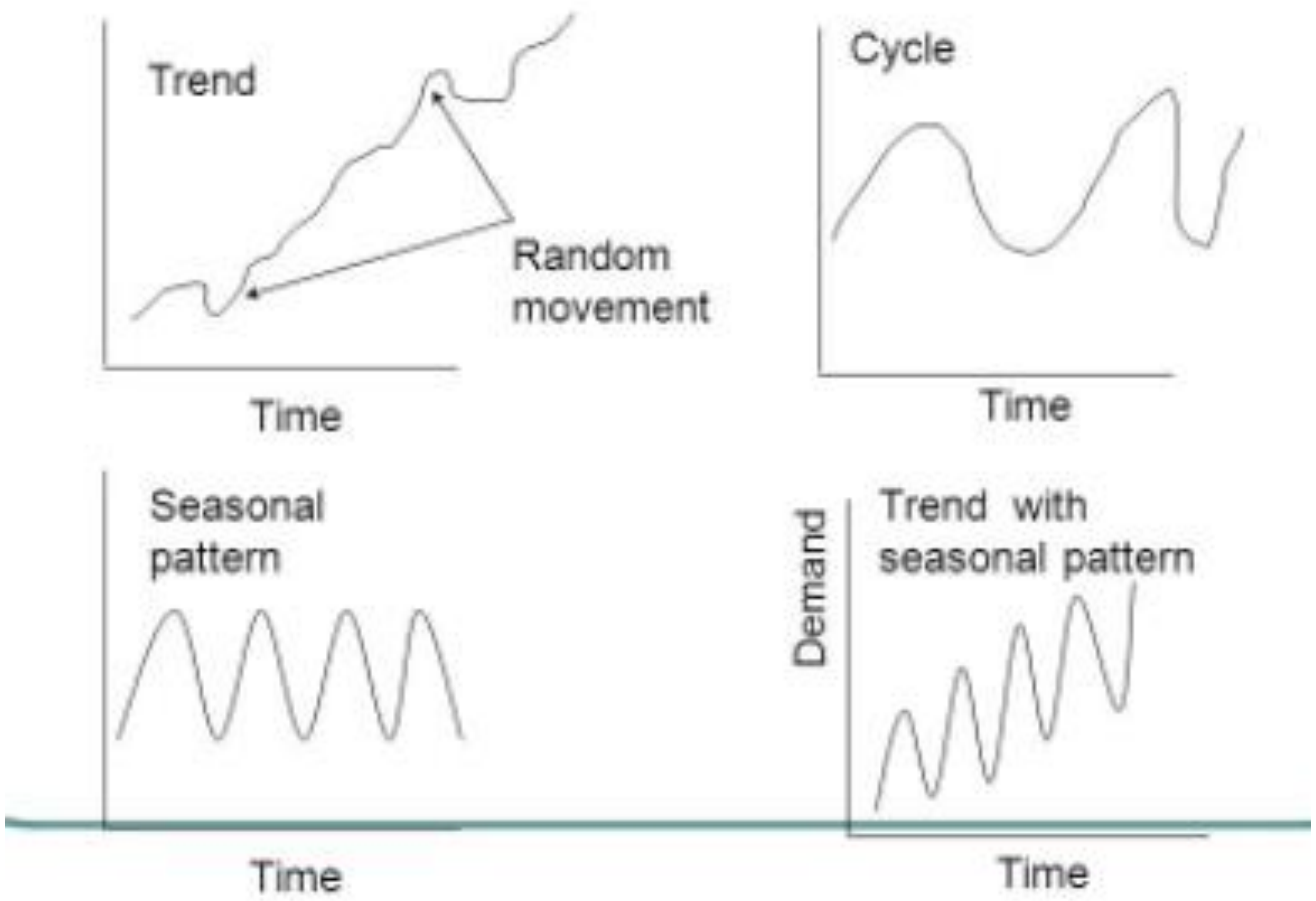
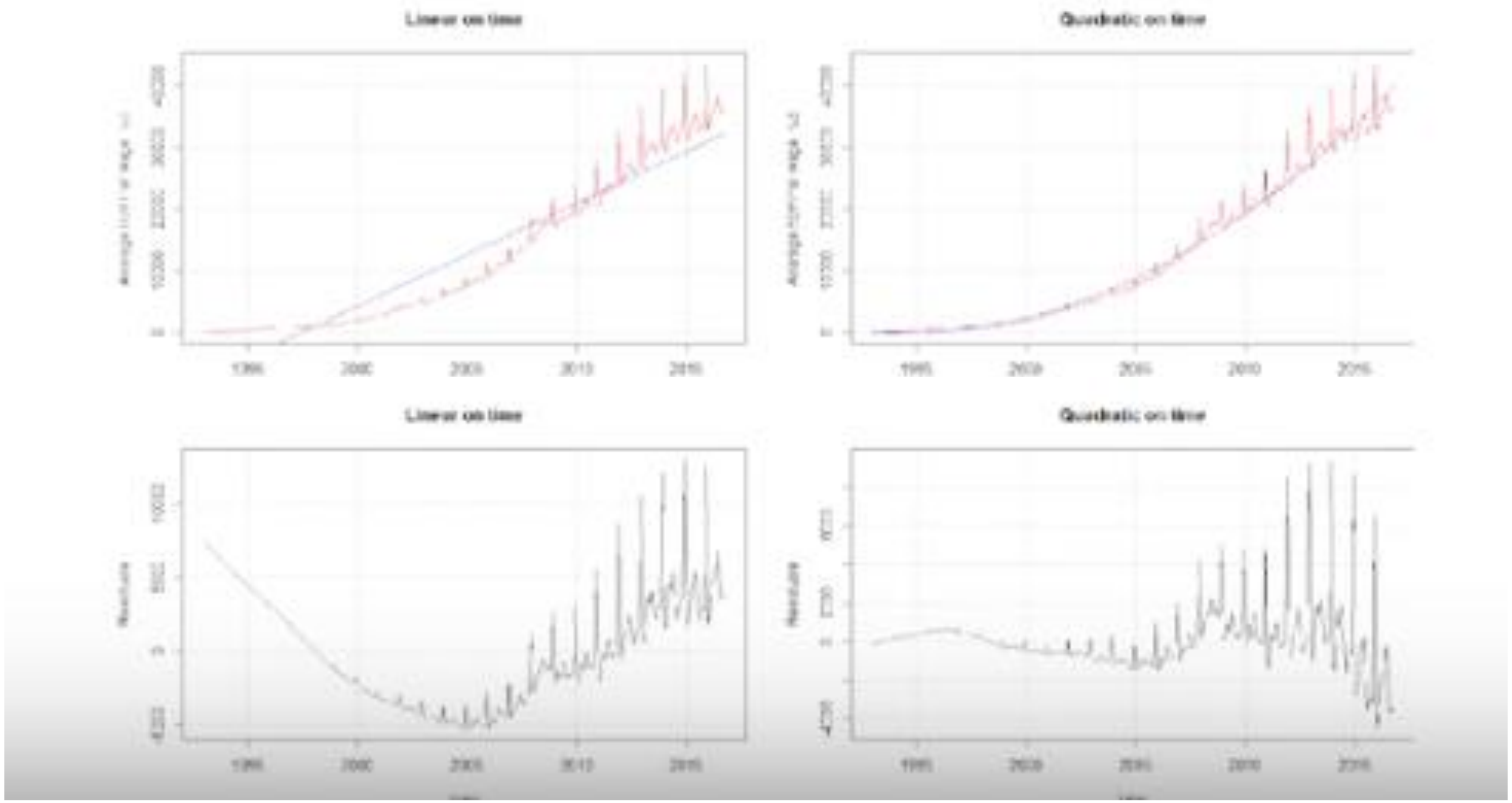


Figure 1.4 Monthly arrival of tourists to a town

Time Series Components



If we use regression



Source: https://www.youtube.com/watch?v=u433nrxdf5k

Univariate and Multivariate Time Series

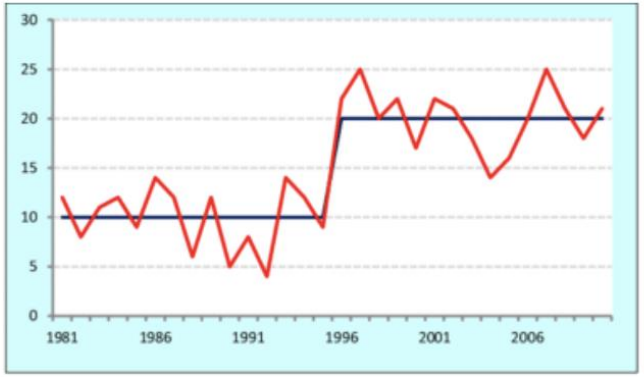
Univariate Time Series: only one variable is varying over time. For example, GDP in Russia.

Multivariate Time Series: multiple variables are varying over time. For example, GDP, population, unemployment rate and inflation in Russia.

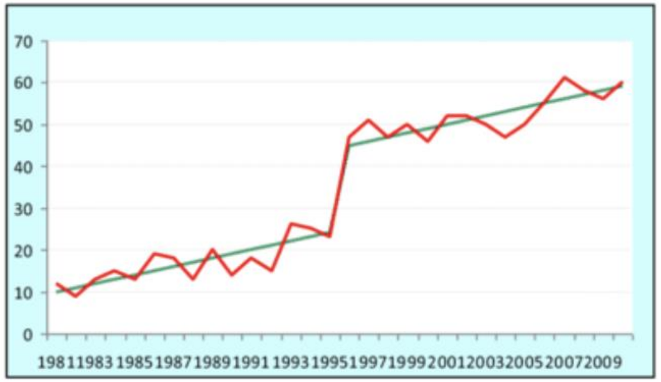
Structural Break

Sudden change in behavior at a specific point in time. For example, many macroeconomic indicators experienced a drastic change in 2008 following the onset of the global financial crisis. These sudden changes are often referred to as structural breaks or nonlinearities.

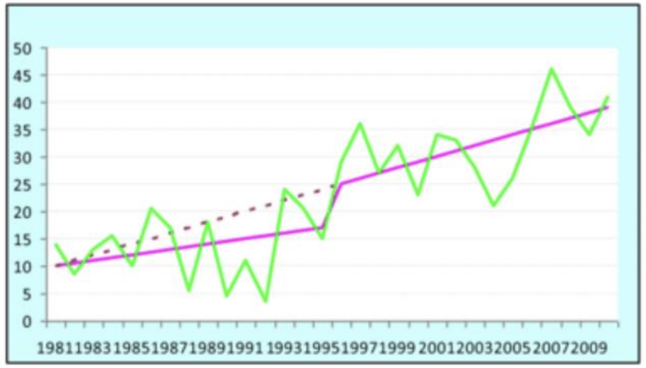
a. Structural break in intercept,
b. Structural break in intercept,
c. Structural break in trend,
d. Structural break in intercept and trend.



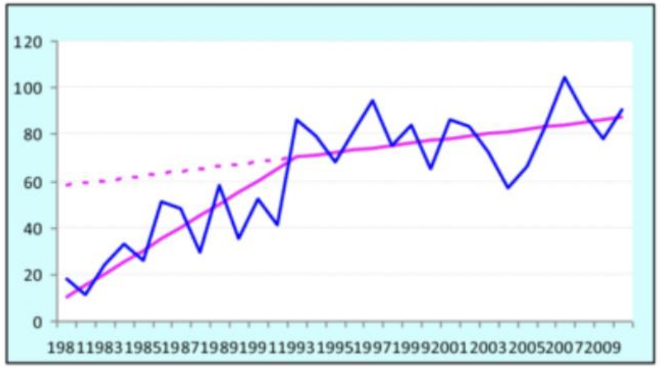
a



b



c



d

Stationary Time Series

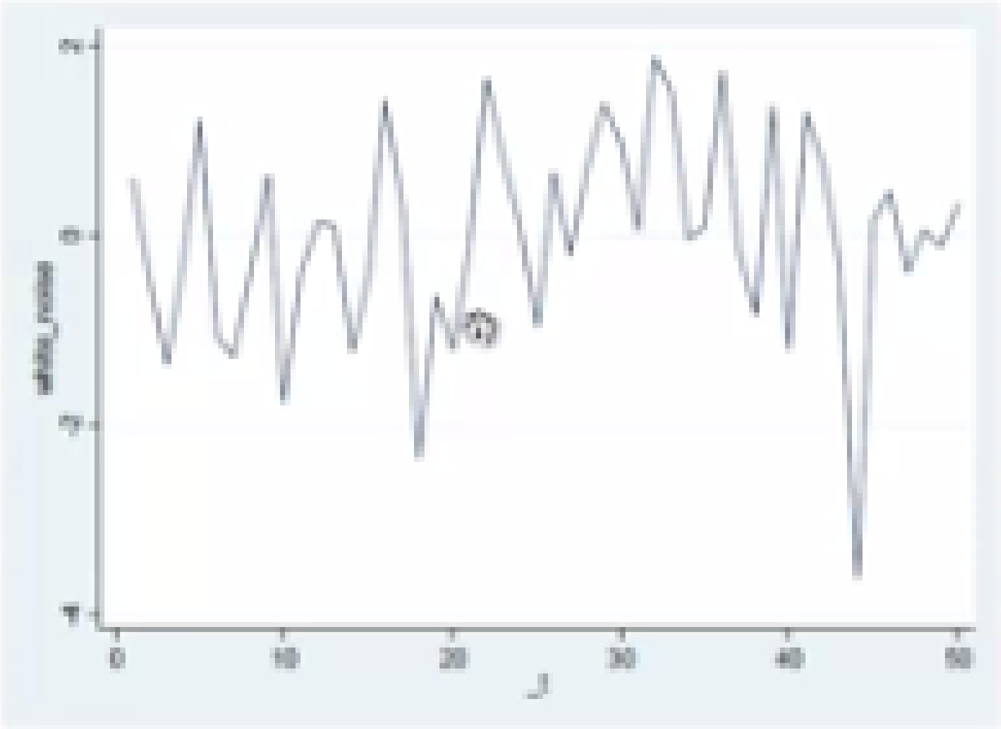
A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary – the trend and seasonality will affect the value of the time series at different times.

For stationary time series:

- The mean is constant,
- The variance is constant,
- Autocovariance does not depend on time.

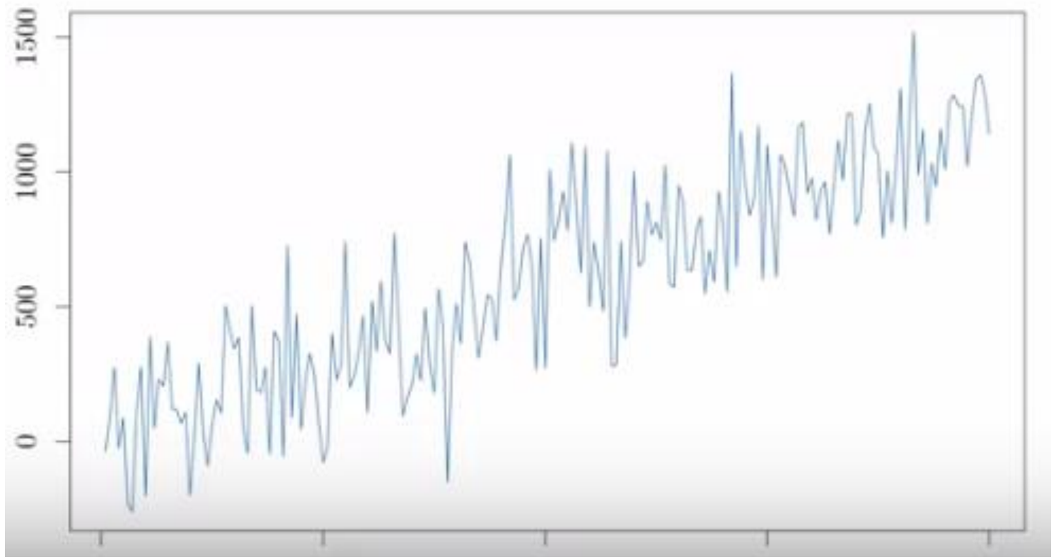
Example of stationary process: White Noise

White Noise is a random process, whose samples are regarded as a sequence of serially uncorrelated random variables with zero mean and finite variance.

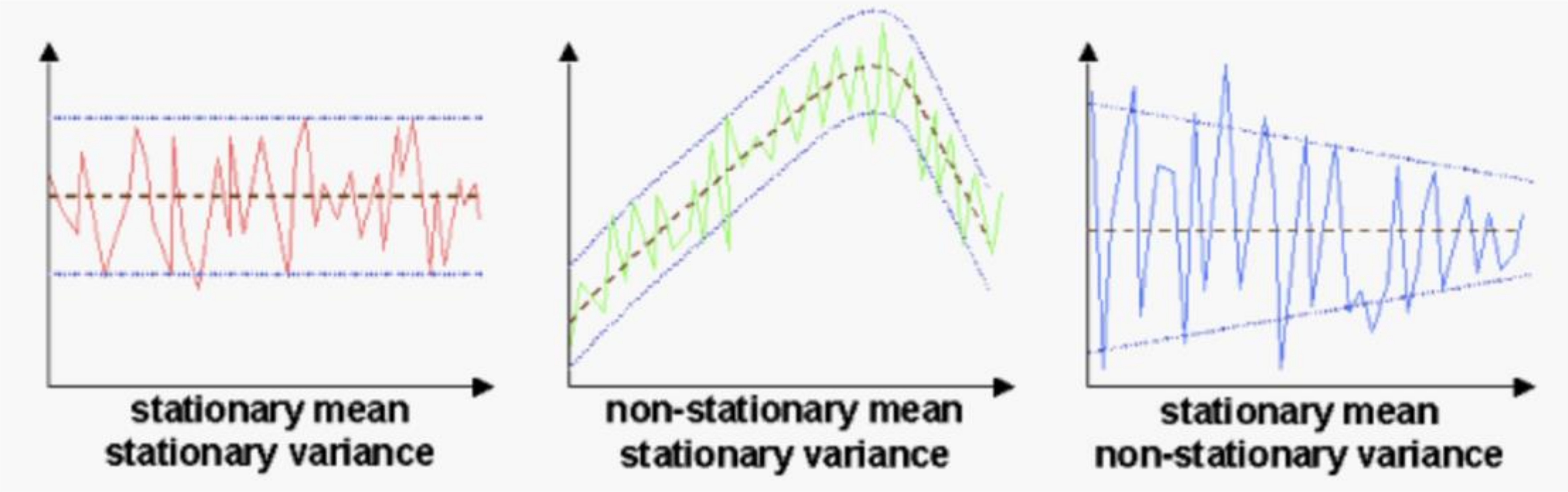


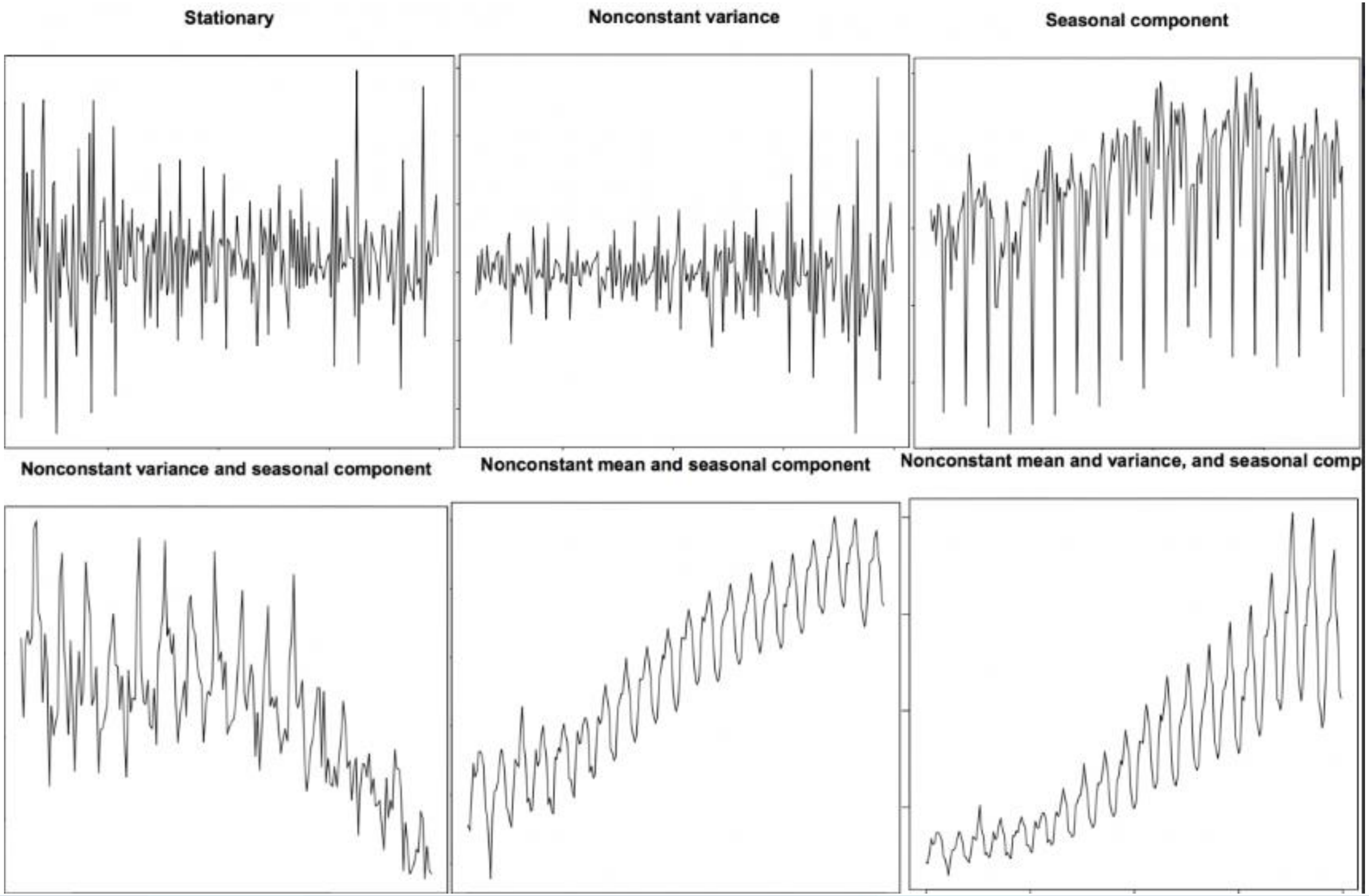
Example of non-stationary process

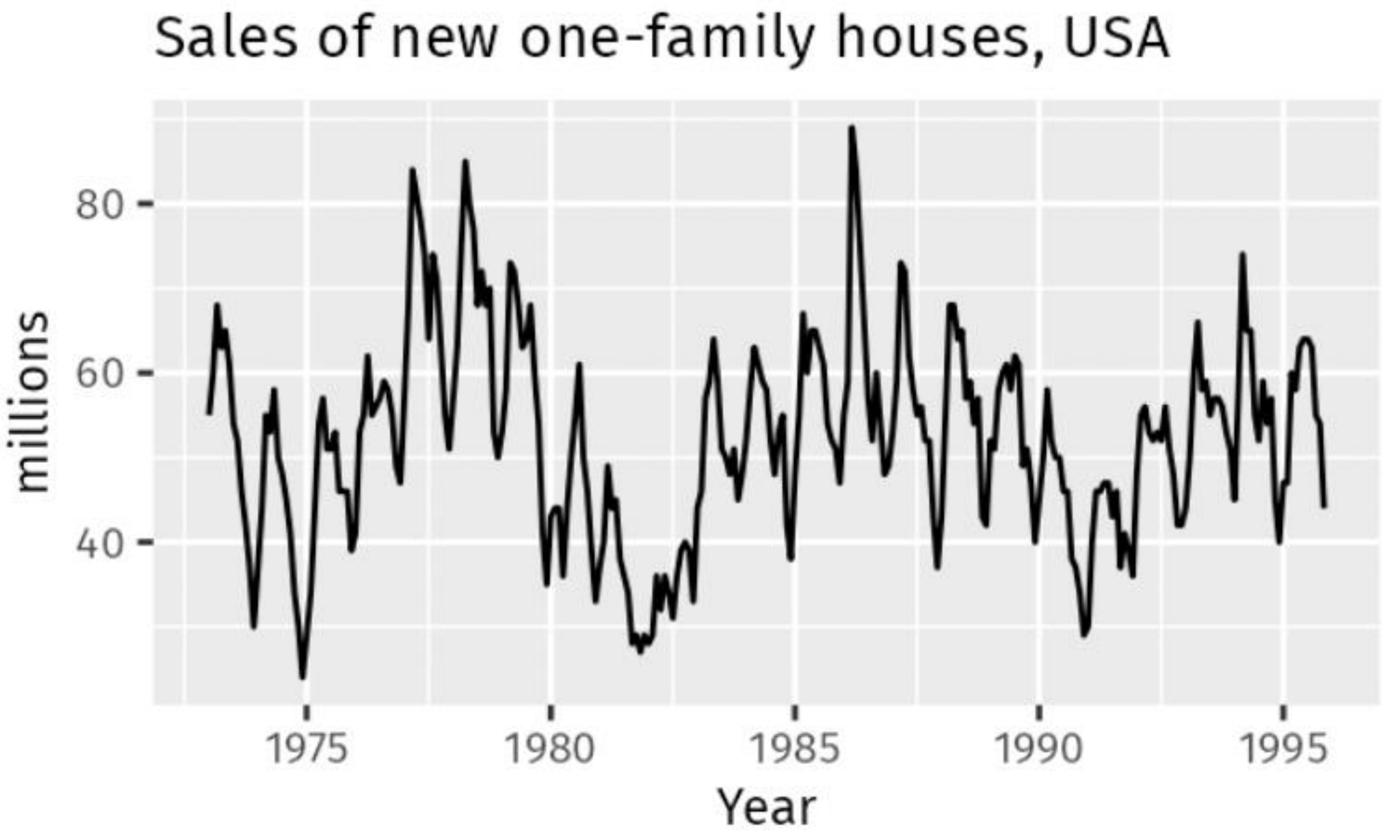
If there is a trend or seasonality in the data, then the series is not stationary.



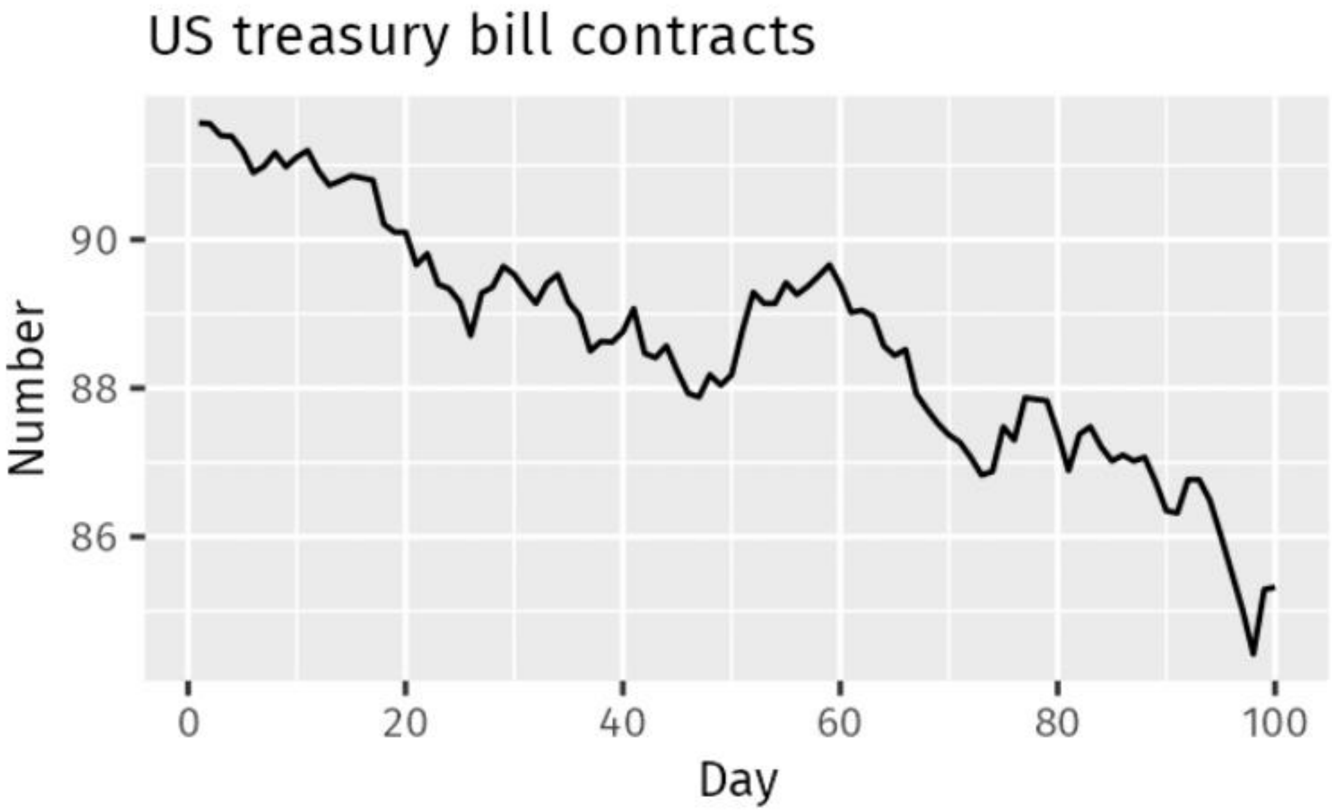
Process with a deterministic trend.


stationary mean
stationary variance
non-stationary mean
stationary variance
stationary mean
non-stationary variance

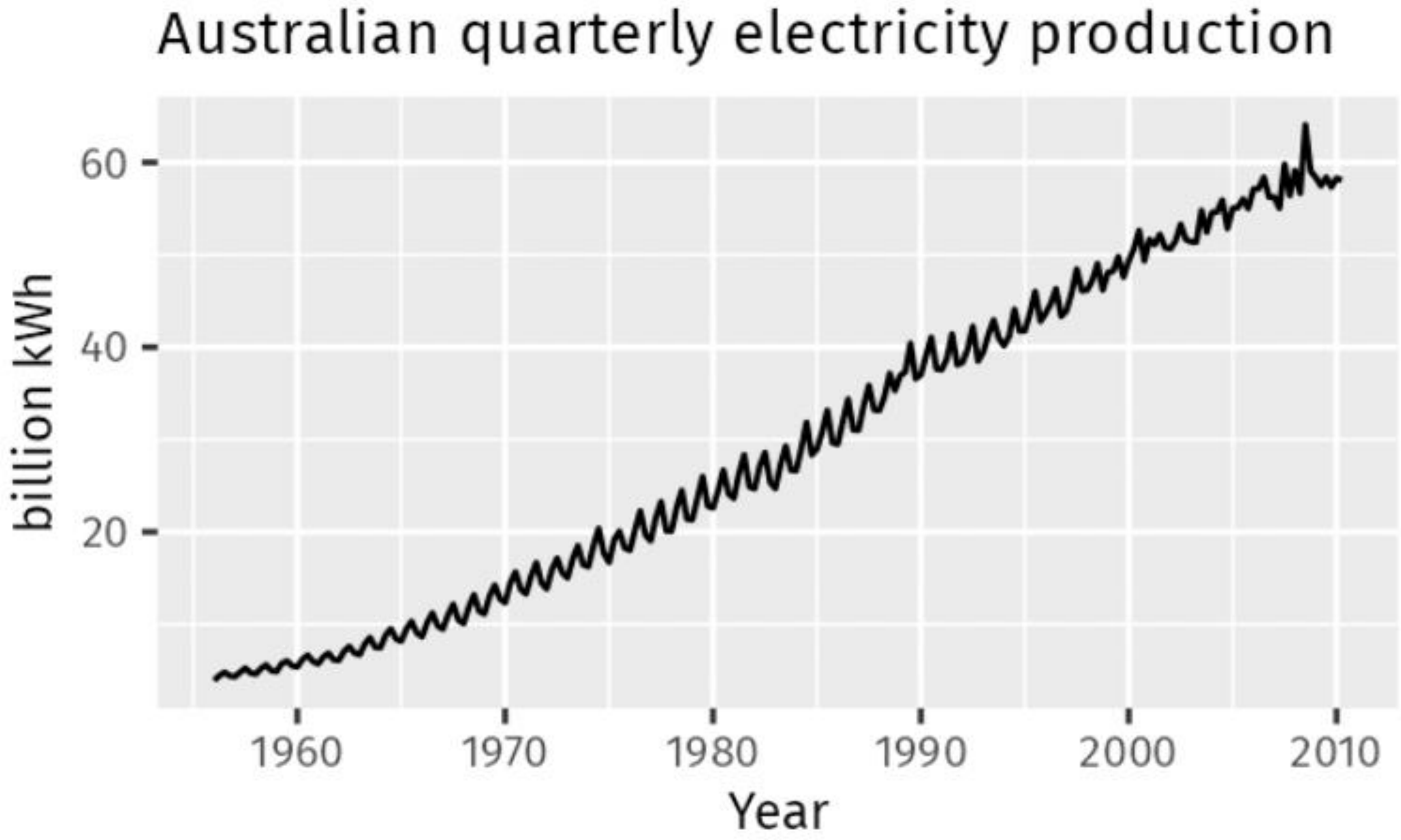

Stationary
Nonconstant variance
Seasonal component
Nonconstant variance and seasonal component
Nonconstant mean and seasonal component
Nonconstant mean and variance, and seasonal comp



Monthly home sales data show seasonality within each year, as well as cyclicality over a period of about 6-10 years. There is no clear trend in the data.

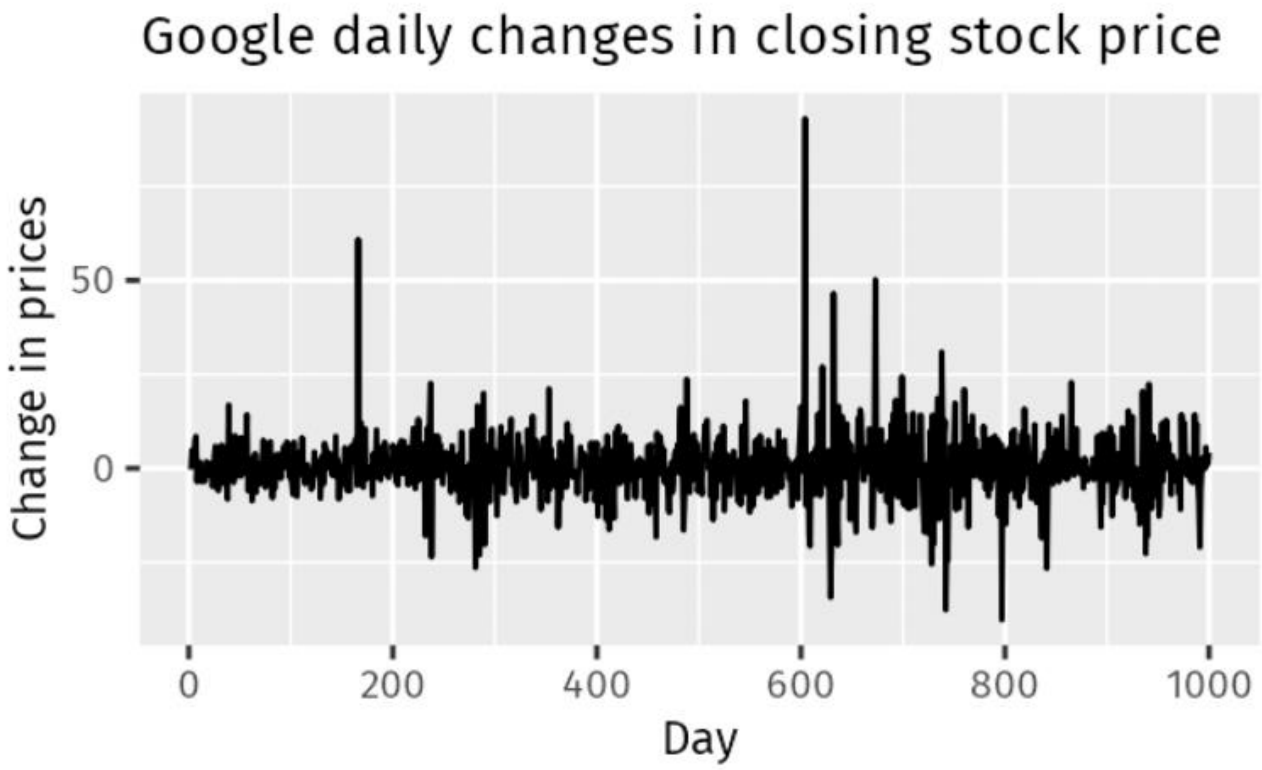


There is no seasonality, but there is a downward trend. Perhaps if we had a longer series, we would see that this trend is part of a long cycle.



Upward trend and seasonality.

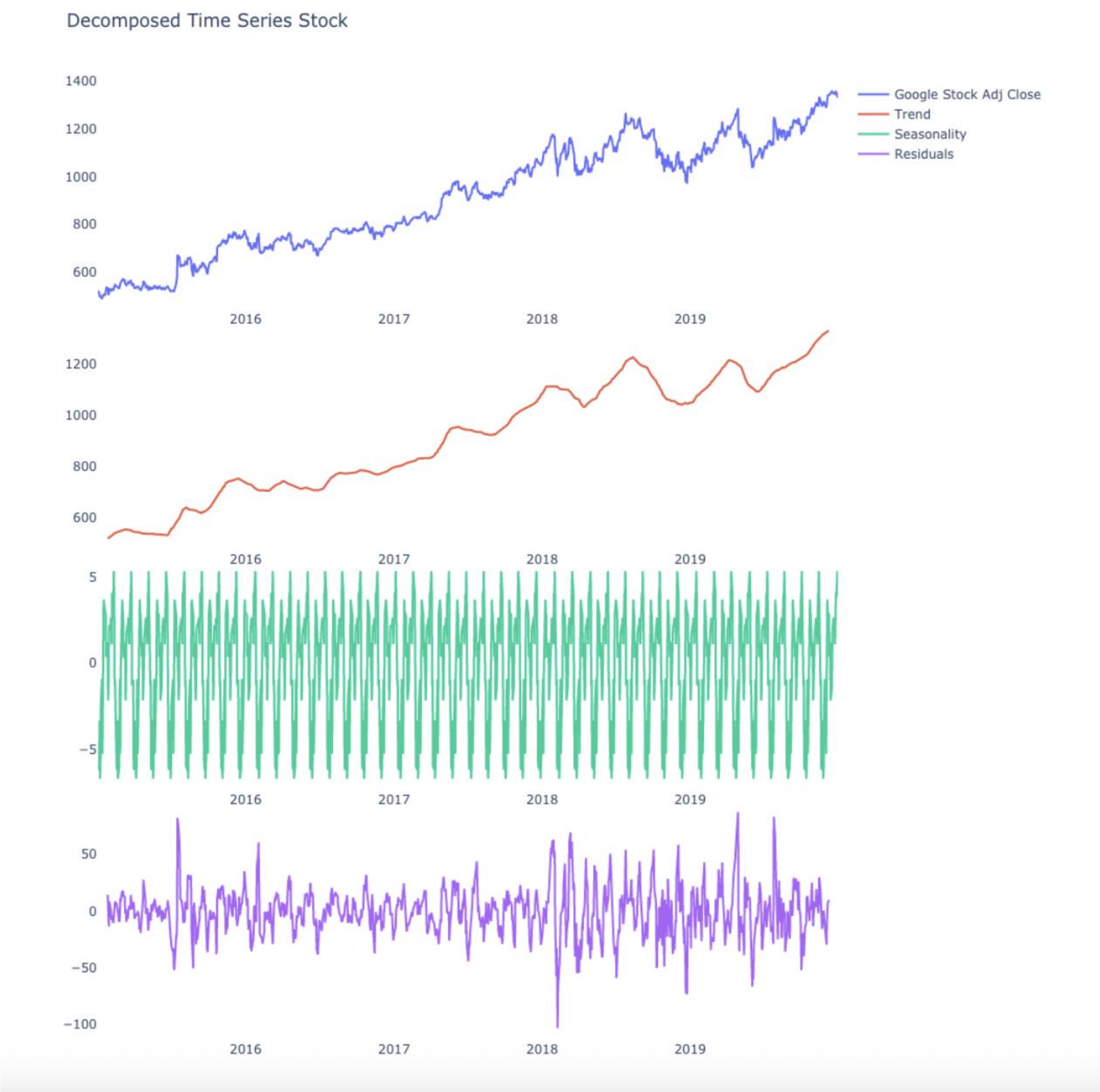
Day



There is no trend, seasonality or cycles.

Time series decomposition

We break down a time series into its constituent components: trend, seasonality and residuals.



Time series decomposition models

1. Additive decomposition

The formula is saying is that if you take the trend and add it to the seasonal component, and add that to the residuals, you get back the original time series. We use an additive decomposition when seasonal variation is constant over time.

Series = Trend + Seasonality + Residuals

2. Multiplicative decomposition

Multiplicative model is used when seasonal variation scales with the trend. That is, when you see that the size of the seasonal effects become larger as the overall series moves up or down. It’s a form of heteroscedasticity.

Series = Trend × Seasonality × Residuals

Time series decomposition stages

1. Trend Estimation

The first step is to identify the trend component, which represents the long-term movement or direction in the data (upward, downward, or stable). We can use Moving averages for this.

2. Seasonal Component Estimation

After removing the trend, the next step is to estimate the seasonal component, which captures periodic fluctuations or repeating patterns (e.g., monthly, quarterly, yearly).

Typically done by averaging the detrended data over equivalent time periods (e.g., all Januarys, all Mondays).

3. Residual (Irregular) Component Calculation

Once both the trend and seasonal components are estimated, the residuals are calculated.

These represent the irregular or random variations that cannot be explained by trend or seasonality.

Autocorrelation

If there is a trend or cycles in the time series, the values of each subsequent level of the series depend on the previous one. The correlation dependence between successive levels of the time series is called autocorrelation of the levels of the series. It can be measured using a linear correlation coefficient between a time series and itself at different “lags” (shifts in time).

Autocorrelation Coefficient

The first-order autocorrelation coefficient measures the relationship between the adjacent levels of the series y_t and y_{t-1}. Similarly, we can define the autocorrelation coefficients of the second and higher orders. Thus, the second-order autocorrelation coefficient characterizes the strength of the relationship between the levels y_t and y_{t-2}. The number of periods for which the autocorrelation coefficient is calculated is called lag. As the lag increases, the number of pairs of values used to calculate the autocorrelation coefficient decreases. It is recommended that the maximum lag is no more than n / 4.

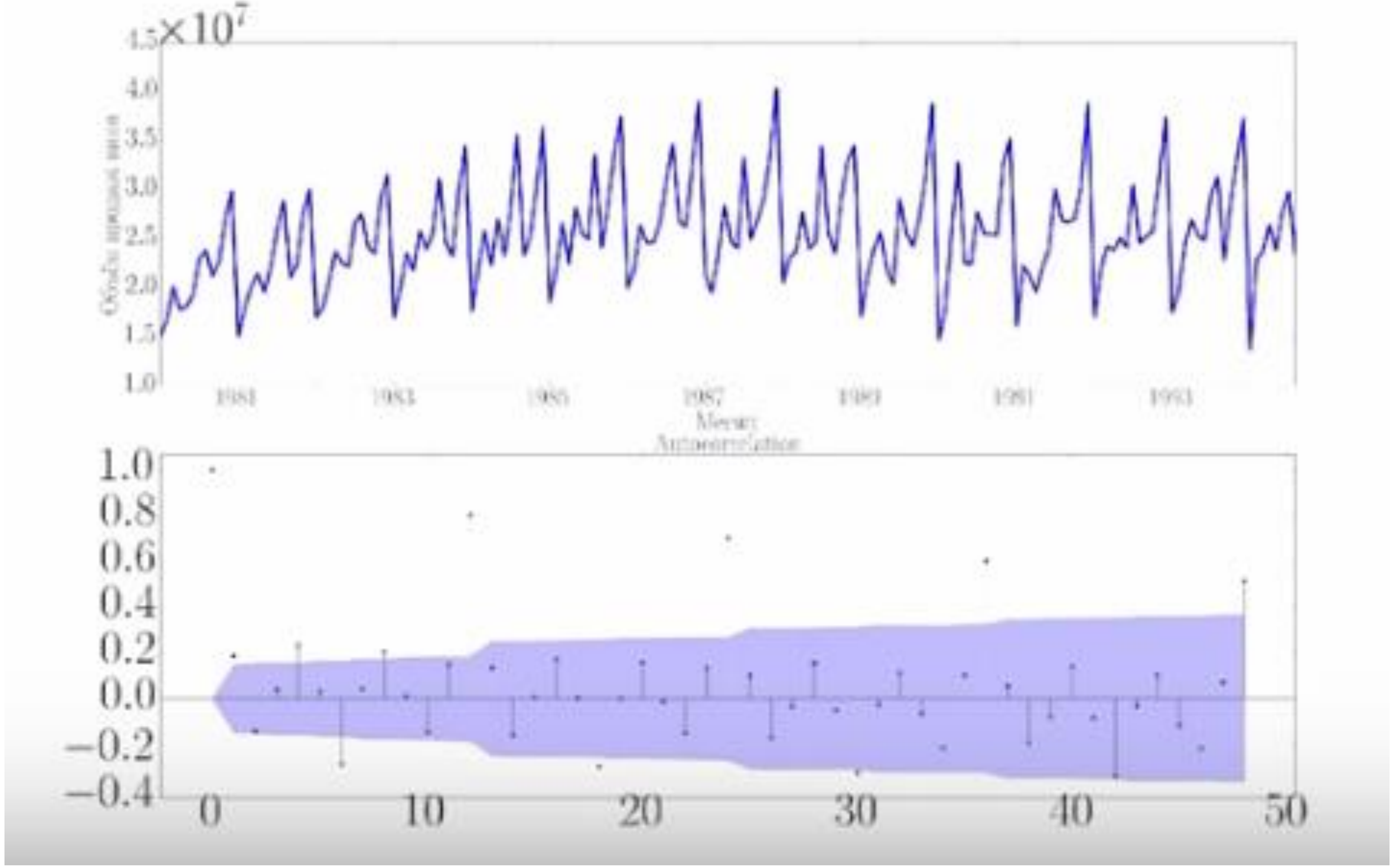
Autocorrelation Function

Sequence of autocorrelation coefficients for the levels of the first, second, etc. orders is called the autocorrelation function of the time series. The graph of the dependence of autocorrelation coefficients values on the lag value (of the order of the autocorrelation coefficient) is called correlogram. By analyzing the autocorrelation function and the correlogram, it is possible to reveal the structure of the series.

If the first order autocorrelation coefficient turned out to be the highest, the series contains only a trend. If the autocorrelation coefficient of the order of X turned out to be the highest, then the series contains cyclical fluctuations with a periodicity at X points in time. If none of the autocorrelation coefficients is significant, then either the series does not contain trends and cyclical fluctuations, or the series contains a strong non-linear trend, which requires additional analysis to be identified.

Autocorrelation Function

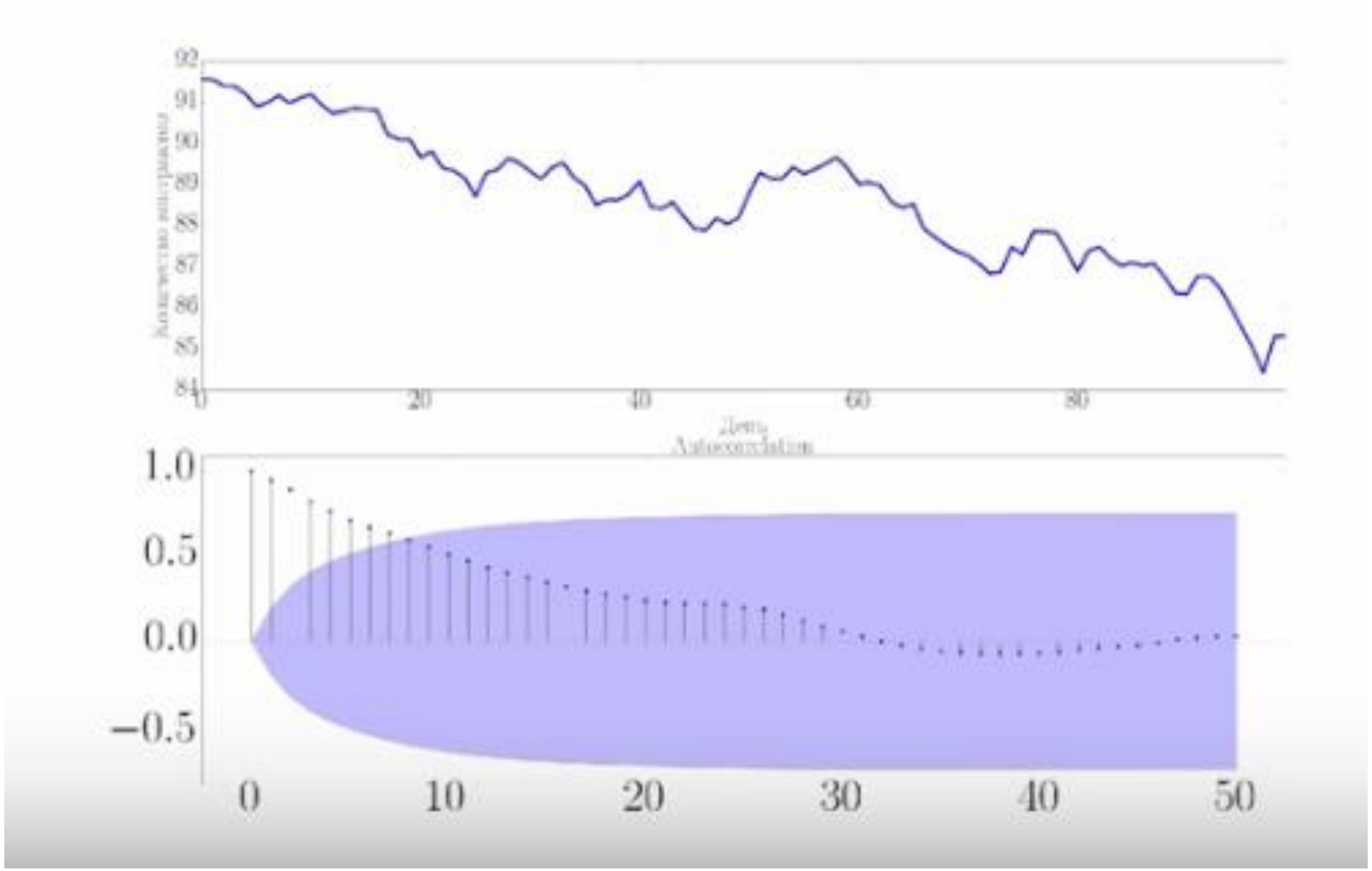
Autocorrelation plot for different lags. The graphs show peaks on the lags of the seasonal period (12, 24, 36, 48 months).



Source: https://www.youtube.com/watch?v=u433nrxdf5k

Autocorrelation Function

Series with a strong trend without seasonality.



AutoRegressive Model (AR)

Autoregressive model: $y_t = b*y_{t-1} + E$

AR model is one in which y_t depends only on its own past values y_{t-1}, y_{t-2}, y_{t-3}, etc.

Model AR(p):
p – number of lags
AR(0): $y_t = b_0$
AR(1): $y_t = b_0 + b_1 * y_{t-1}$

$$\boxed{Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \dots + \beta_p Y_{t-p} + \varepsilon_t}$$

Moving Average Model (MA)

A Moving Average model is one when y_t depends only on the random error terms.

A common representation of moving average model where it depends on q of its past values is called MA(q) model and is represented below.

$y_t = b_0 + b_1 * y_{t-1} + E_1$
$y_{t-1} = b_0 + b_1 * y_{t-2} + E_2$

$$\boxed{Y_t = \beta_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \dots + \phi_q \varepsilon_{t-q}}$$

MA(q) q – number of lags

AutoRegressive Moving Average Model (ARMA)

There are situations where the time-series may be represented as a mix of both AR and MA models referred as ARMA (p,q):

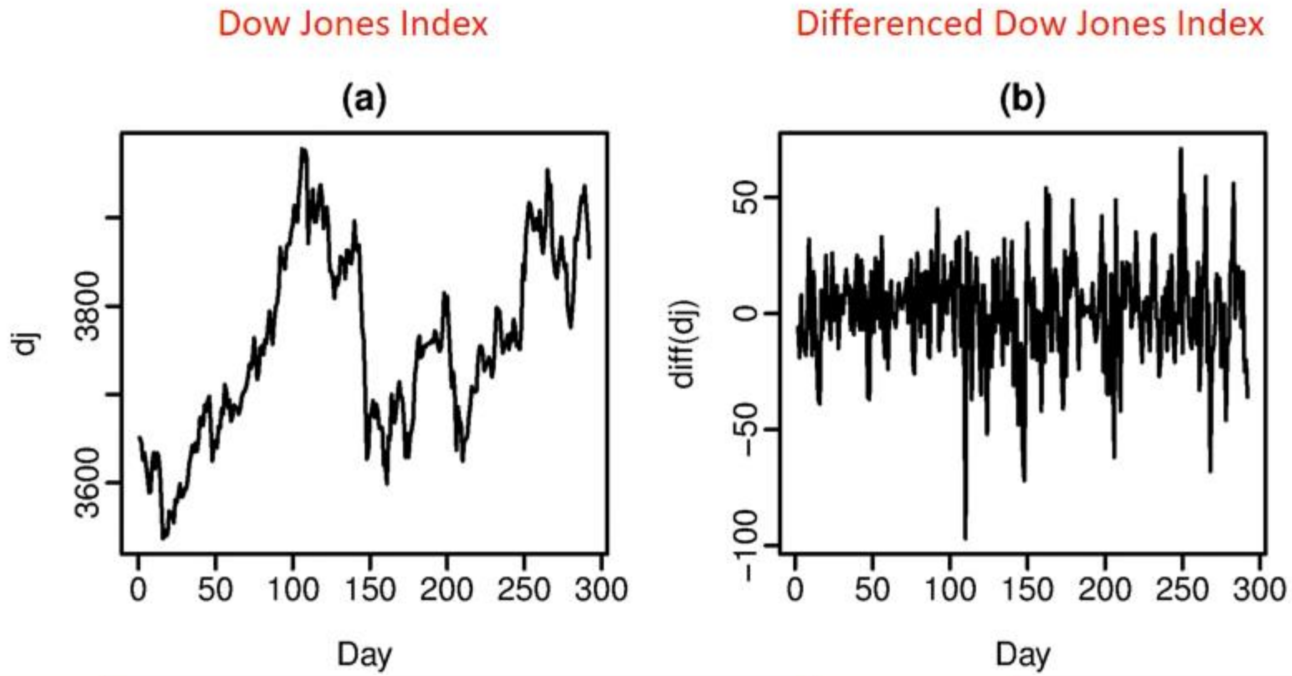
- p – number of autoregressive components
- q – number of moving average components

$$\boxed{\begin{aligned}Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \\ \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \ldots + \phi_q \varepsilon_{t-q} \ .\end{aligned}}$$

To apply the model, the series must be stationary (i.e., the mean and variance do not change over time and there is no trend). According to the Wold's theorem, any stationary series can be described by the ARMA (p, q) model with any accuracy.

Differencing

With the help of differencing, it is possible to transform a non-stationary series into a stationary one. We calculate the difference between the values in current and previous moment in time. Differentiation can stabilize the average value of the series and get rid of the trend and seasonality. For monotonically varying variances, the logarithmic transformation can be used.



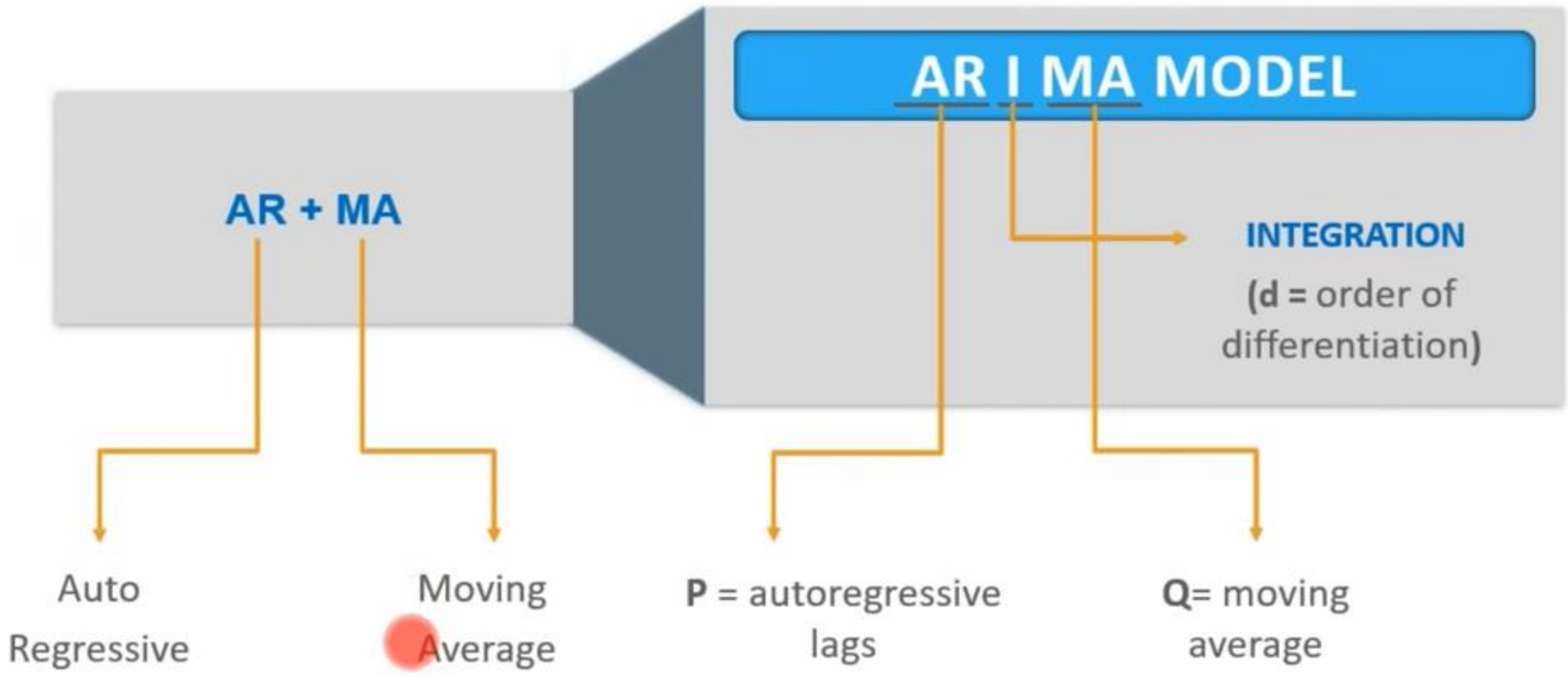
ARIMA

A series is described by ARIMA (p, d, q) model if the d times differentiated series is described by the ARMA (p, q) model.

ARIMA (auto regression, integrated, moving average).

ARIMA (p, d, q) is an ARMA model with p autoregressive lags, q moving average lags and differentiation of order d.

ARIMA MODEL


AR + MA
AR I MA MODEL
INTEGRATION
(d = order of differentiation)
Auto Regressive
Moving Average
P = autoregressive lags
Q= moving average

Box-Jenkins methodology for selecting appropriate ARIMA model

Identification step:

1) Examining the timeline of the series:
- identify outliers, missing values and structural gaps in data,
- estimate of the stationarity of the series,
- transform the series if it is necessary to bring it to a stationary form (logarithmic transformation, differentiation).

2) Studying the autocorrelation function and the partial autocorrelation function to determine the values of the ARIMA model - p, d, q.

Estimation step:

Finding the simplest model with significant coefficients and the highest model quality scores.

Diagnostic Checking:

If the model is good, then the residuals should be distributed as white noise, outliers can skew the result.

Comparison of models: information criteria

An information criterion is a measure of the quality of a statistical model. It takes into account:

- how well the model fits the data;
- the complexity of the model.

Information criteria are used to compare alternative models fitted to the same data set. These criteria penalize the number of parameters and reward goodness of fit. A lower AIC or BIC value indicates a better fit.

$$AIC = 2k - 2\ln(L) \qquad BIC = 2k - \ln(n)\ln(L)$$

SARIMA

SARIMA or Seasonal Autoregressive Integrated Moving Average is an extension of the traditional ARIMA model, specifically designed for time series data with seasonal patterns. While ARIMA is great for non-seasonal data, SARIMA introduces seasonal components to handle periodic fluctuations and provides better forecasting capabilities for seasonal data.

Before applying SARIMA, seasonal differencing is often required to make the data stationary. This process involves subtracting the current observation from one that corresponds to the same season in the previous cycle. Seasonal differencing helps remove the seasonal pattern from the data, enabling more accurate forecasting.

SARIMAX

SARIMAX model is a statistical time series forecasting method that extends the ARIMA model by incorporating both seasonality and exogenous variables. It is used to model and forecast data with repeating patterns over time, like monthly or yearly cycles, by using past values of the time series (Autoregressive) and past forecast errors (Moving Average). It also uses differencing to make the data stationary and includes external factors (Exogenous Regressors) that can influence the time series.

Components of SARIMAX

- **Seasonal** (S): Accounts for repeating patterns that occur at fixed intervals, such as weekly, monthly, or yearly cycles.
- **Autoregressive** (AR): Uses the relationship between an observation and a number of lagged observations (past values).
- **Integrated** (I): Uses differencing to make the time series stationary (removing trend and seasonality).
- **Moving Average** (MA): Uses the relationship between an observation and a residual error from a moving average model applied to lagged observations.
- **Exogenous Regressors** (X): Includes external predictor variables that influence the time series but are not influenced by it.

Simple Exponential Smoothing

Simple Exponential Smoothing (SES) is a classic time series forecasting method used to predict future values by applying exponentially decreasing weights to past observations. In other words, recent data points are given more importance than older ones. It's particularly effective for time series data without trend or seasonality, where the underlying level of the series is relatively stable.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

where $0 \leq \alpha \leq 1$ is the smoothing parameter.

$$Overall\ Equation : \hat{y}_{t+h} = l_t$$

$$Level\ Equation : l_t = \alpha y_t + (1-\alpha) l_{t-1}$$

h is the time step we are forecasting

Simple Exponential Smoothing: smoothing parameter

	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
y_T	0.2000	0.4000	0.6000	0.8000
y_{T-1}	0.1600	0.2400	0.2400	0.1600
y_{T-2}	0.1280	0.1440	0.0960	0.0320
y_{T-3}	0.1024	0.0864	0.0384	0.0064
y_{T-4}	0.0819	0.0518	0.0154	0.0013
y_{T-5}	0.0655	0.0311	0.0061	0.0003

For any α between 0 and 1, the weights attached to the observations decrease exponentially as we go back in time, hence the name "exponential smoothing". If α is small (i.e., close to 0), more weight is given to observations from the more distant past. If α is large (i.e., close to 1), more weight is given to the more recent observations.

Holt’s Linear Trend Model

Holt’s Linear Trend model, also is called Double Exponential Smoothing, extends SES to handle data with a trend, that does not have seasonality. It uses alpha α to smooth the level of the series and beta β to smooth the trend or rate of change.

We want to estimate two things at each time step:

- **Level** – the current value of the series.
- **Trend** – how fast the value is increasing or decreasing.

$$Overall\ Equation : \hat{y}_{t+h} = l_t + hb_t$$

$$Level\ Equation : l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$Trend\ Equation : b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

Holt-Winters Model

The Holt-Winters Exponential Smoothing model (also called Triple Exponential Smoothing) is a smoothing method used to predict time series data with both a trend and seasonal component. New smoothing parameter, gamma (γ), is used to control the effect of seasonal component.

The technique uses exponential smoothing applied three times:

(α) the level,

(β) the trend,

(γ) the seasonal component.

$$Overall\ Equation: \hat{y}_{t+h} = l_t + hb_t + s_{t+h-m}$$

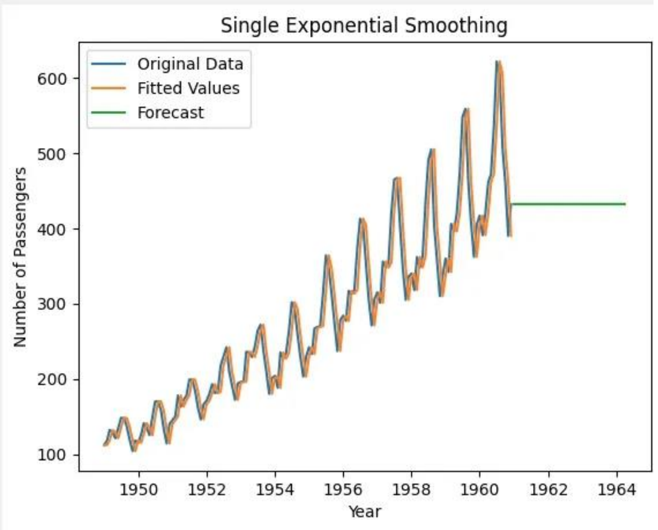
$$Level\ Equation: l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$

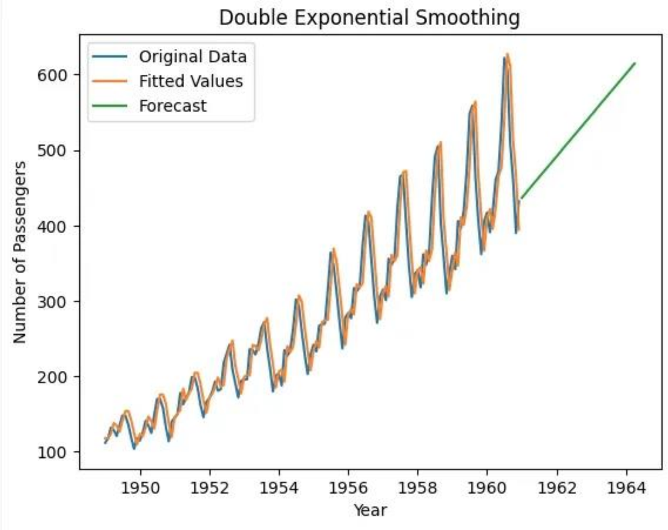
$$Trend\ Equation: b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

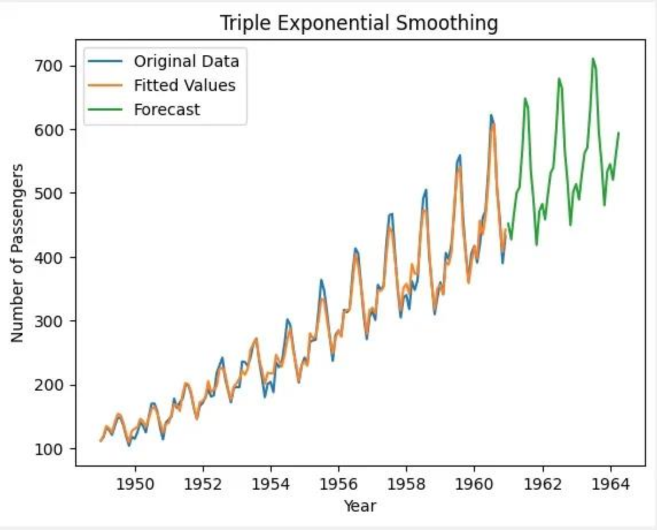
$$Seaonality\ Equation: s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

Where ***m*** is the seasonality of the time series,

Comparing the Exponential Smoothing models







Cross-validation

We can gradually expand the train sample.

Cross-validation

Window size is fixed.

Useful links

- https://www.machinelearningplus.com/time-series/time-series-analysis-python/
- https://www.kaggle.com/code/kashnitsky/topic-9-part-1-time-series-analysis-in-python/notebook
- https://seaborn.pydata.org/generated/seaborn.lineplot.html
- https://builtin.com/data-science/time-series-python
- https://otexts.com/fpp2/tspatterns.html
- https://blogs.cisco.com/analytics-automation/arima1
- https://www.geeksforgeeks.org/machine-learning/sarima-seasonal-autoregressive-integrated-moving-average/

Thank you for your attention!