Time Series Analysis

Introduction

Dr. Ludovic Amruthalingam
ludovic.amruthalingam@hslu.ch

Informatik

Outline

- What is a time series?
- Time series examples
- Time series patterns
- Data generation process
- The challenge of time dependence
- Modeling time series
- Visualization of time series

Warm-up

Where do you encounter time-based data?

What is a time series?

A time series is a collection of **data points** observed **sequentially in time**.

A time series can be
- Continuous: continuous signal over a time interval e.g., analog audio signal
- **Discrete**: data collected at distinct & separate time points e.g., digital audio signal

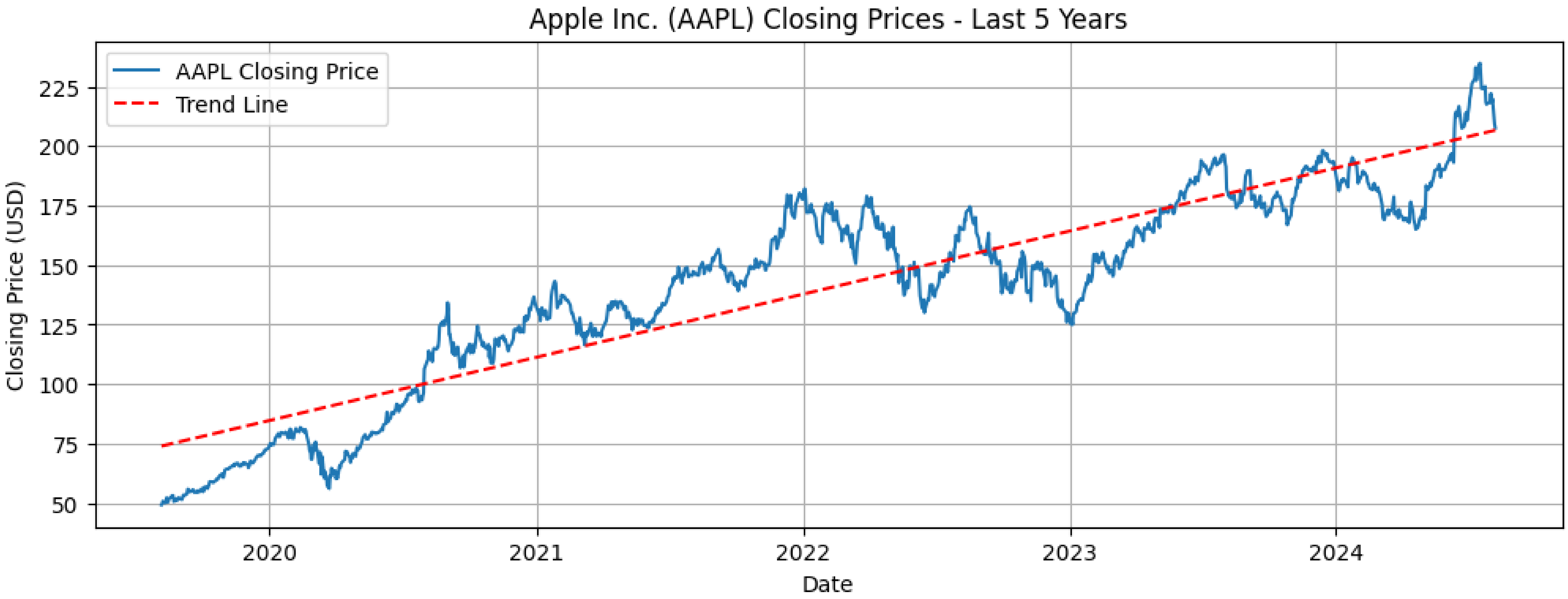
Interval between data points can be
- Irregular e.g., patient data collected during medical appointments
- **Regular** e.g., data collected every hours, months

Time interpretation is secondary, important is for the **indexing variable** to be **ordered**.

Time series examples – Finance

Yahoo Finance. (2024). Historical Stock Price Data. Retrieved from https://finance.yahoo.com/

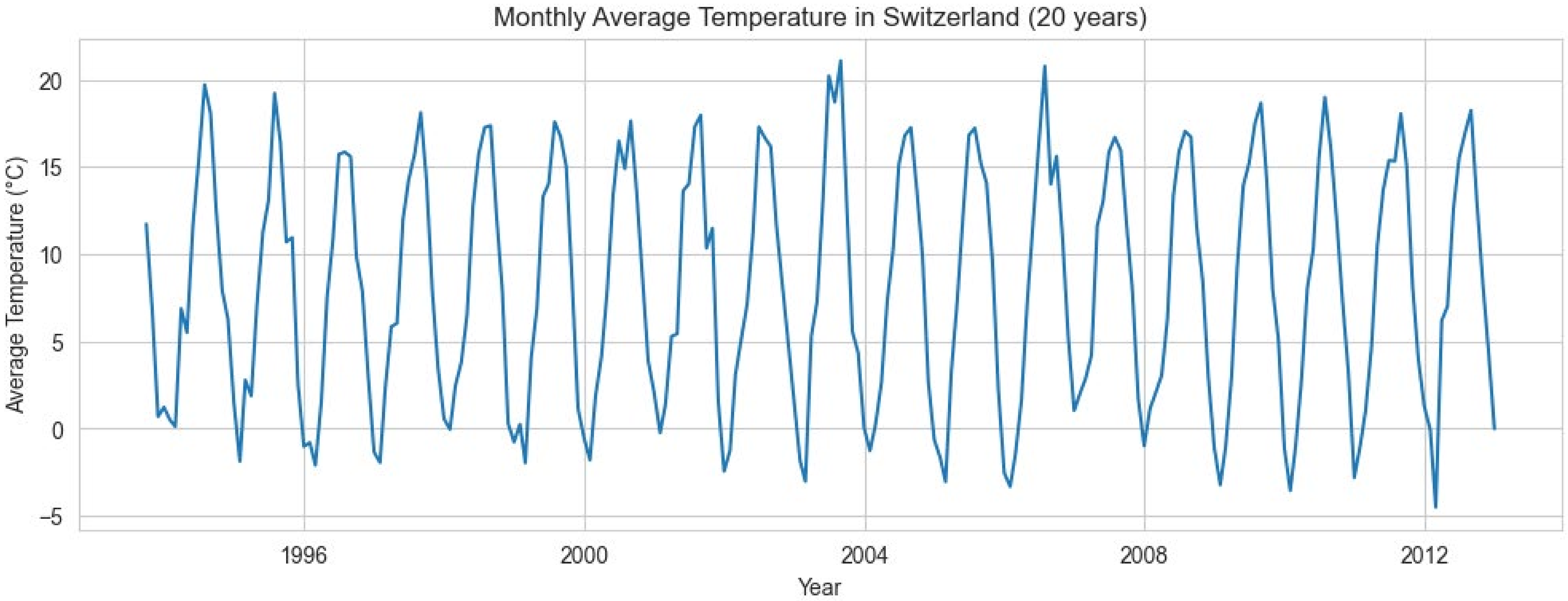
Time series patterns – Trend



A **trend** is a long-term movement or direction in a time series that shows a consistent upward or downward trajectory over time, i.e., a long-term change in the mean.

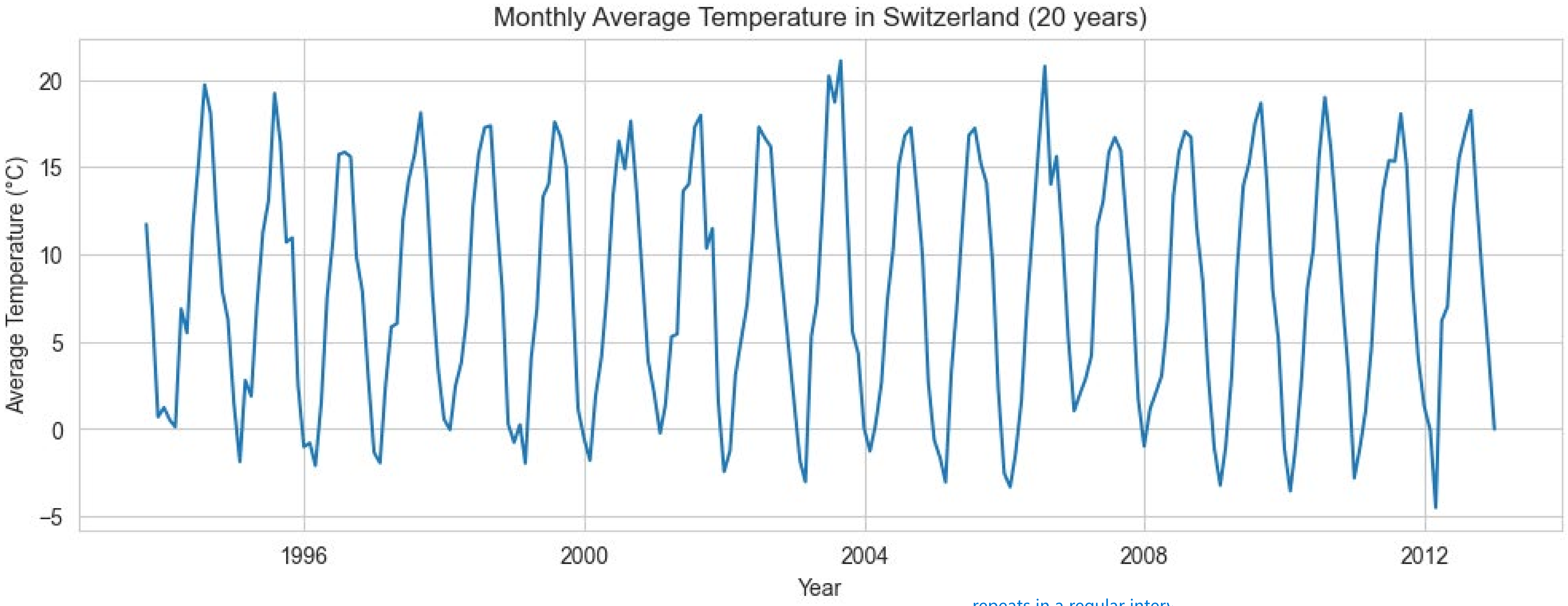
Yahoo Finance. (2024). Historical Stock Price Data. Retrieved from https://finance.yahoo.com/

Time series examples – Meteorology



Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Time series patterns – Seasonal

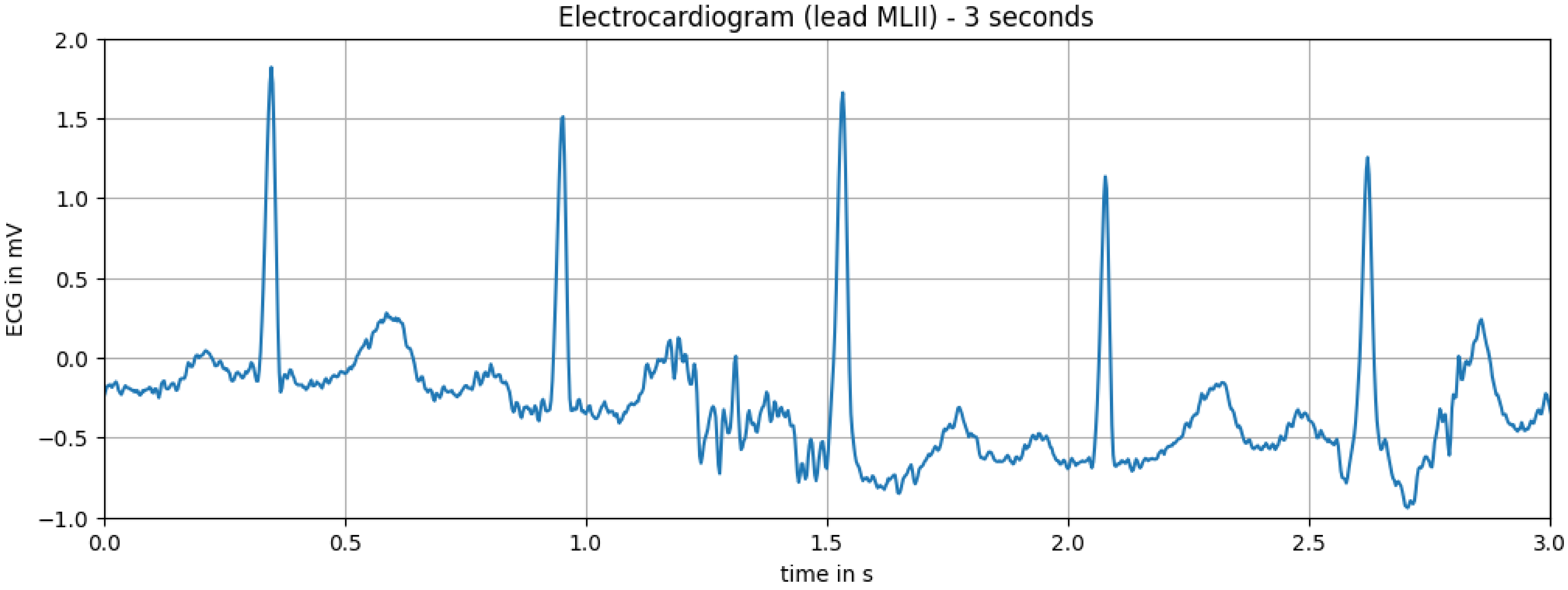


repeats in a regular interv

A **seasonal** behavior is a repeating pattern with fixed period, typically tied to the calendar, such as higher/lower temperatures in summer/winter or increase in retail sales at year-end.

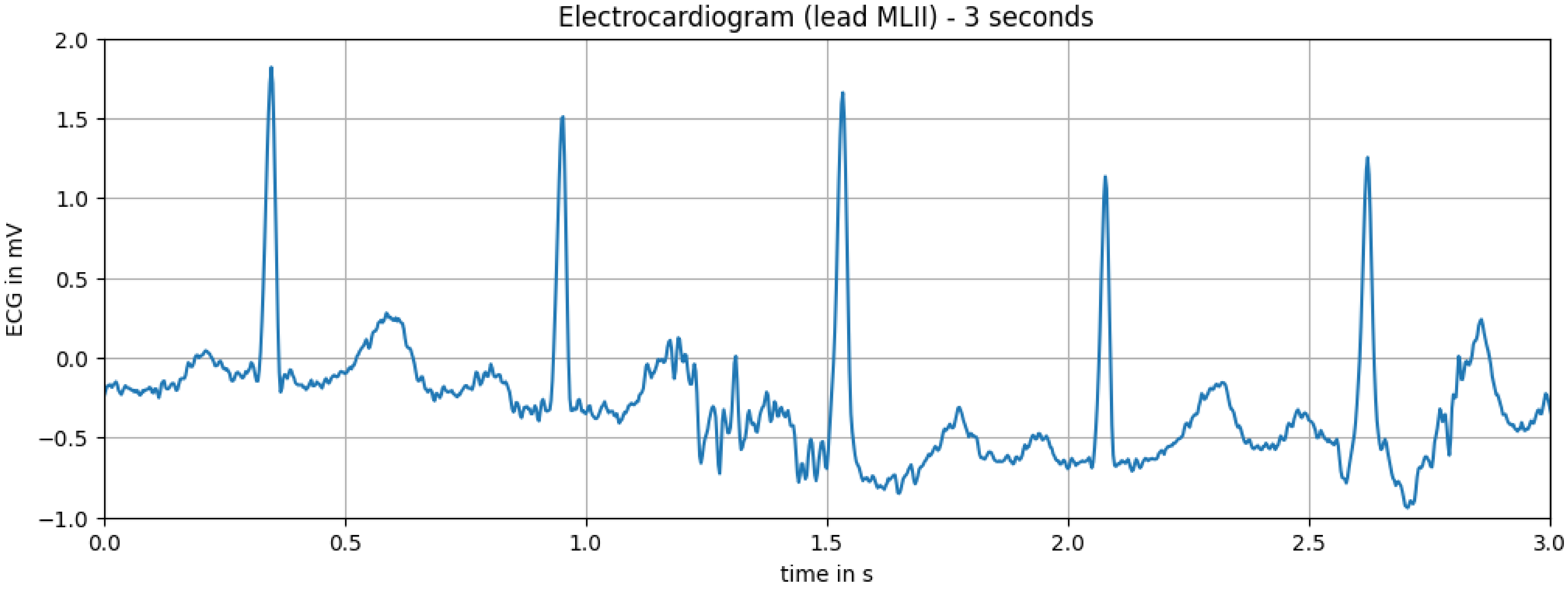
Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Time series examples – Healthcare



Moody et al. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209); DOI:10.13026/C2F305

Time series patterns – Cycle

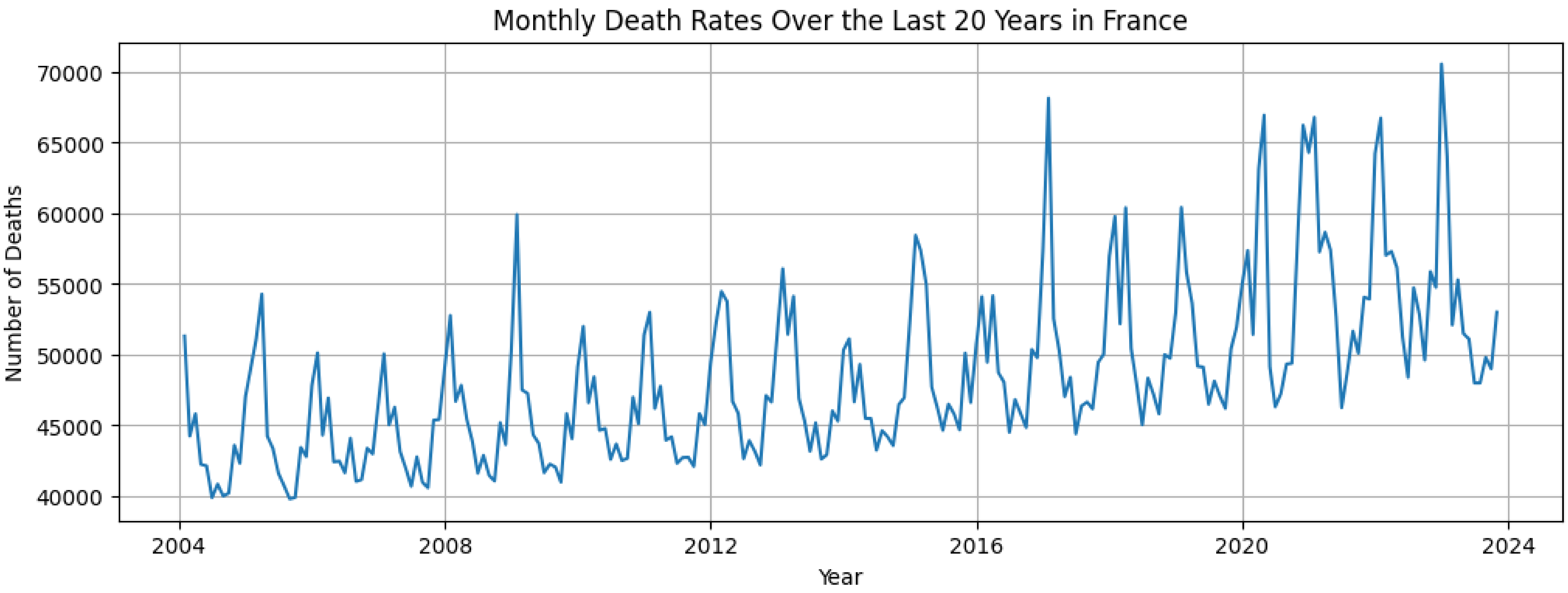


requires different statistical assumptions when modelling

A **cyclic** behavior is a pattern that repeats over ~~regular or~~ irregular intervals, such as heartbeats in an ECG or economic cycles of expansion and recession.

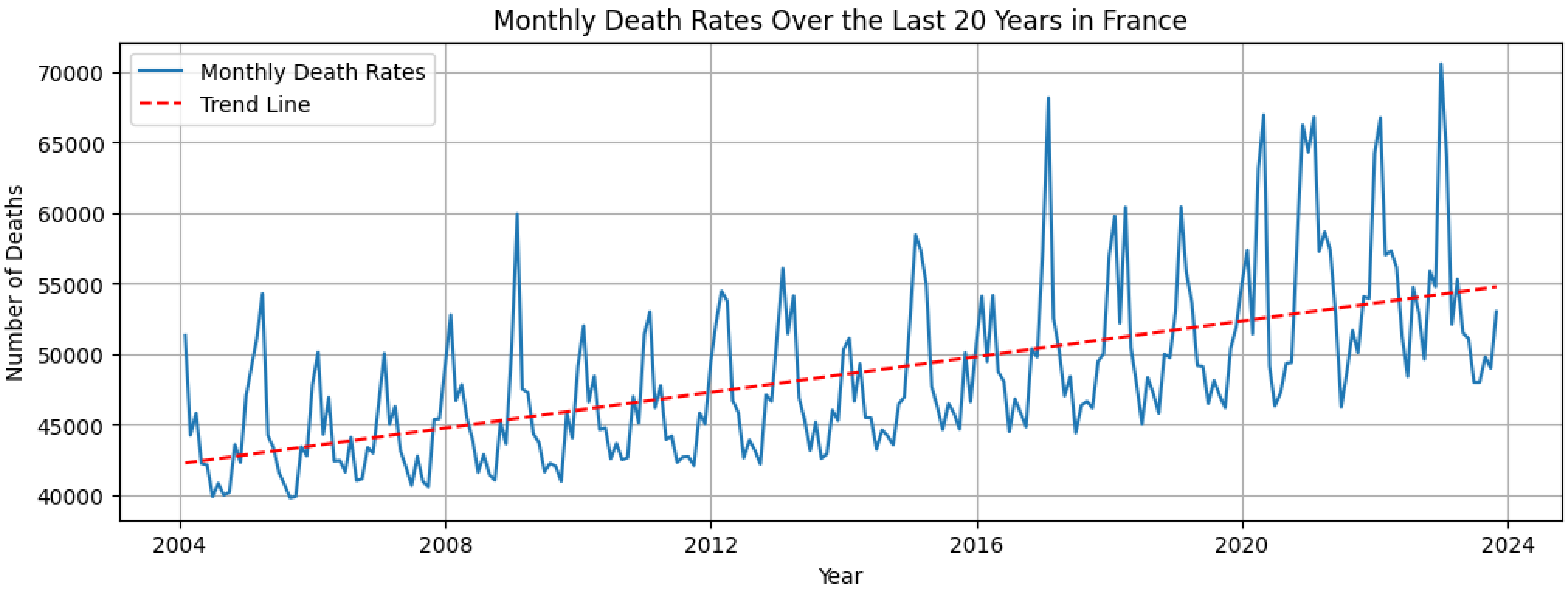
Moody et al. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209); DOI:10.13026/C2F305

Time series examples – Demographics



Eurostat. (2024). Monthly death rates by country. Retrieved from https://ec.europa.eu/eurostat

Time series patterns – Trend + Seasonal



Eurostat. (2024). Monthly death rates by country. Retrieved from https://ec.europa.eu/eurostat

Time series examples – Industry

(Hydraulic system, shown a few sensors of

Helwig et al. "Condition monitoring of a complex hydraulic system using multivariate statistics." 2015 IEEE International Instrumentation and Measurement Technology Conference.

Data generating process (DGP)

Expert knowledge required

Important to identify issues or potentialy biases, etc

The generation of time series data is influenced by:

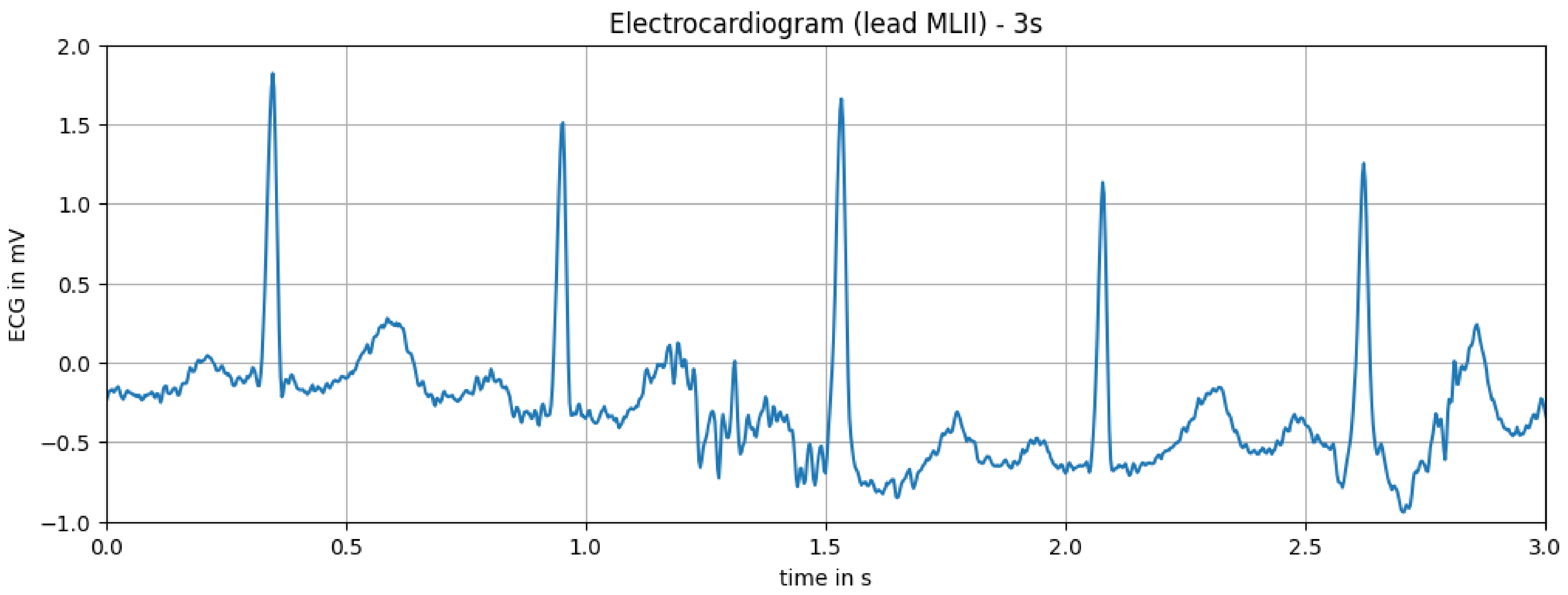
Deterministic components: known factors, laws, predictable patterns influencing data

- Predictor variables e.g., moon distance for tidal variations
- Trend e.g., long-term growth or decline
- Seasonal variations e.g., holiday season, weather patterns
- Cyclical patterns e.g., business cycles

Stochastic components: random fluctuations introducing uncertainty

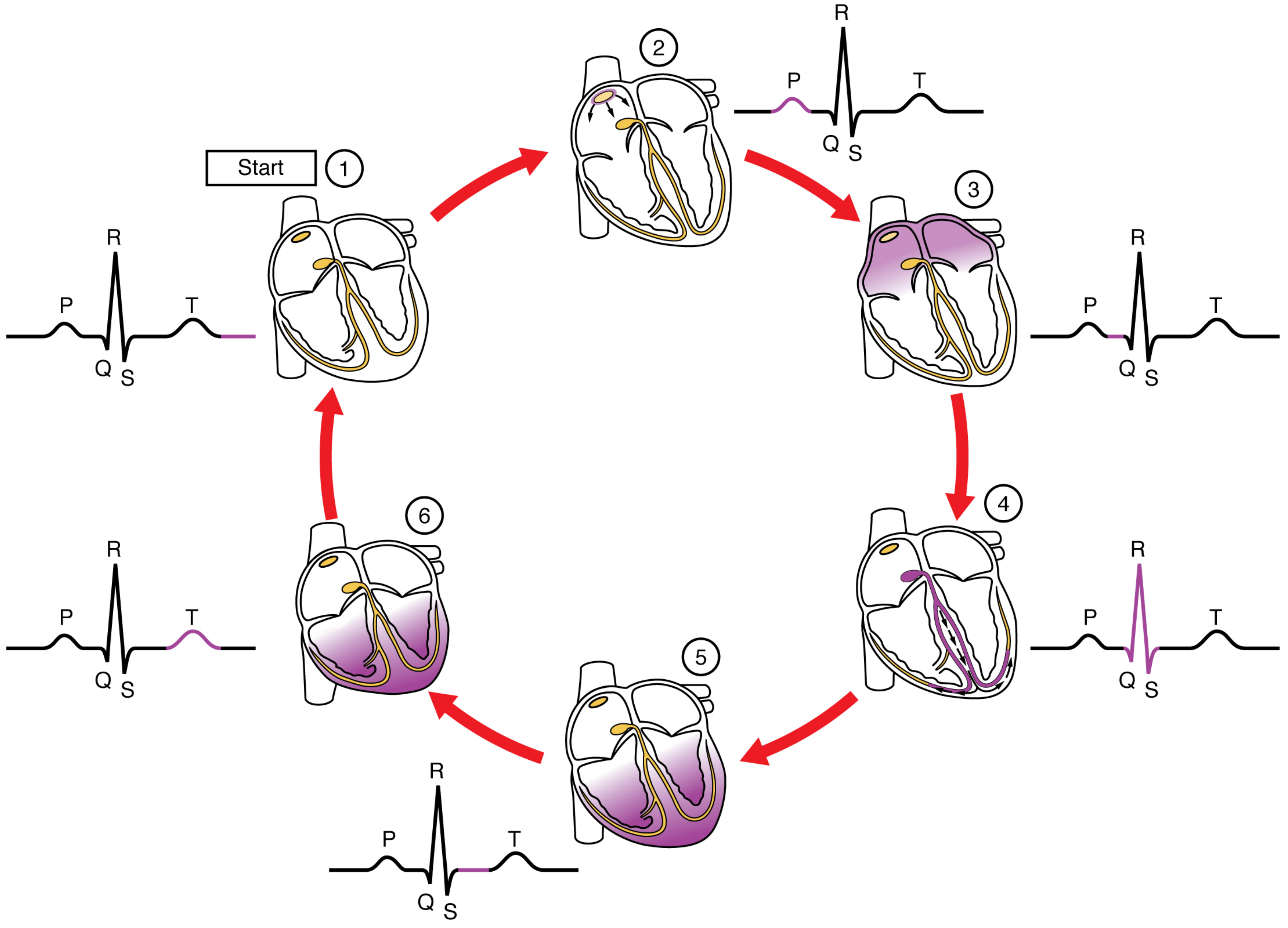
- Noise e.g., random errors in measurement or observation
- Irregular events e.g., natural disasters, economic crises

DGP example – Electrocardiogram (ECG)



Deterministic components

- Cardiac cycle (P wave, QRS complex, and T wave)
- Heart rate (based on physiological state e.g., sleep, exercise)



OpenStax College, Wikimedia

Stochastic components

- Electrical noise from the recording equipment
- Movement artifacts from patient motion
- Irregular heartbeats or arrhythmias

Modeling ECG data requires approaches that account for **both** the deterministic and stochastic components.

Moody et al. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209); DOI:10.13026/C2F305

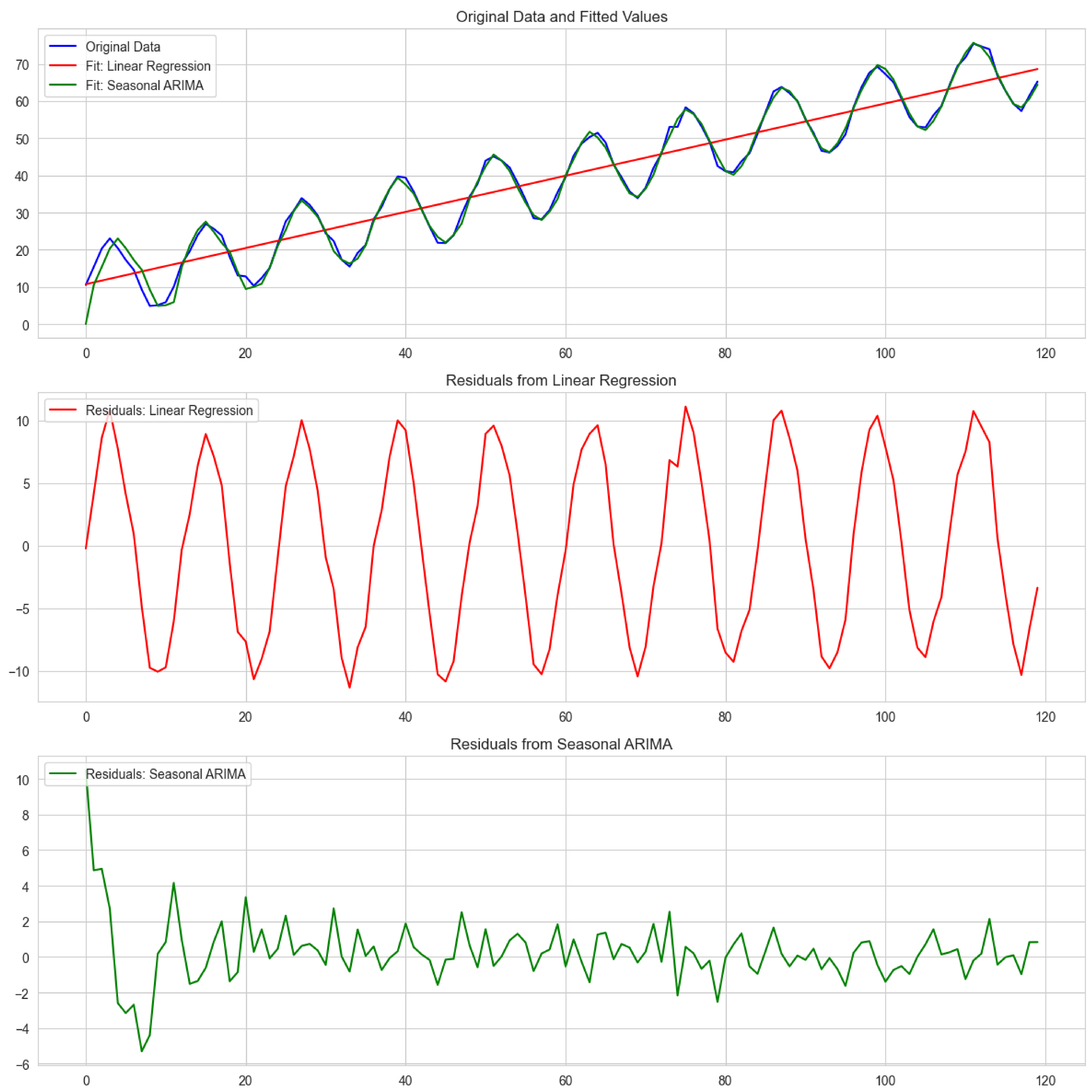
The challenge of time dependence

Most statistical methods assume data points to be **independent** and **identically distributed** (IID).

Time series data violates this assumption due to the **correlation between consecutive observations**.

Ignoring this dependence structure can lead to misleading inference and models with poor data fit.

Time series analysis provides a specialized conceptual framework to address these challenges.



Modeling time series

A. **Understand DGP** i.e., underlying mechanisms/factors that generate/influence the data

B. Collect **representative** data (quantity dependent on modeling objective and data)

In practice, make simplifying **assumptions** (based on the modeling objective and data) such as:

- Stationarity: The statistical properties of the data remain constant over time.
- Linearity: The relationship between variables is linear.
- Normality: The data follows a normal distribution.
- Decomposability: The time series can be broken down into trend, seasonal, and irregular components.

Modeling objectives:

- **Forecasting**: prioritize trends and seasonal patterns while downplaying stochastic variations
- **Classification**: assume extracted patterns and features can differentiate between categories
- **Anomaly detection**: capture deviations from assumed “normal” patterns

Modeling approaches

Time domain approaches: model correlations between adjacent observations
- Directly interpretable in the context of time
- Well-suited for small to moderate datasets.
- E.g., autoregressive integrated moving average, state-space models, Kalman filters

Frequency domain approaches: model periodicity of the time series
- Effective for handling cycles and seasonal effects
- Useful in signal processing
- E.g., Fourier analysis, spectral analysis

Machine learning and deep learning approaches: learn statistical patterns from data
- Model complex nonlinear relationships and patterns
- Require fewer assumptions about data
- E.g., random forest, recurrent/convolutional neural networks

Visualization – Preparing the data

Review data structure

- Ensure **consistent** time intervals
- Granularity of observations: up/down-sampling

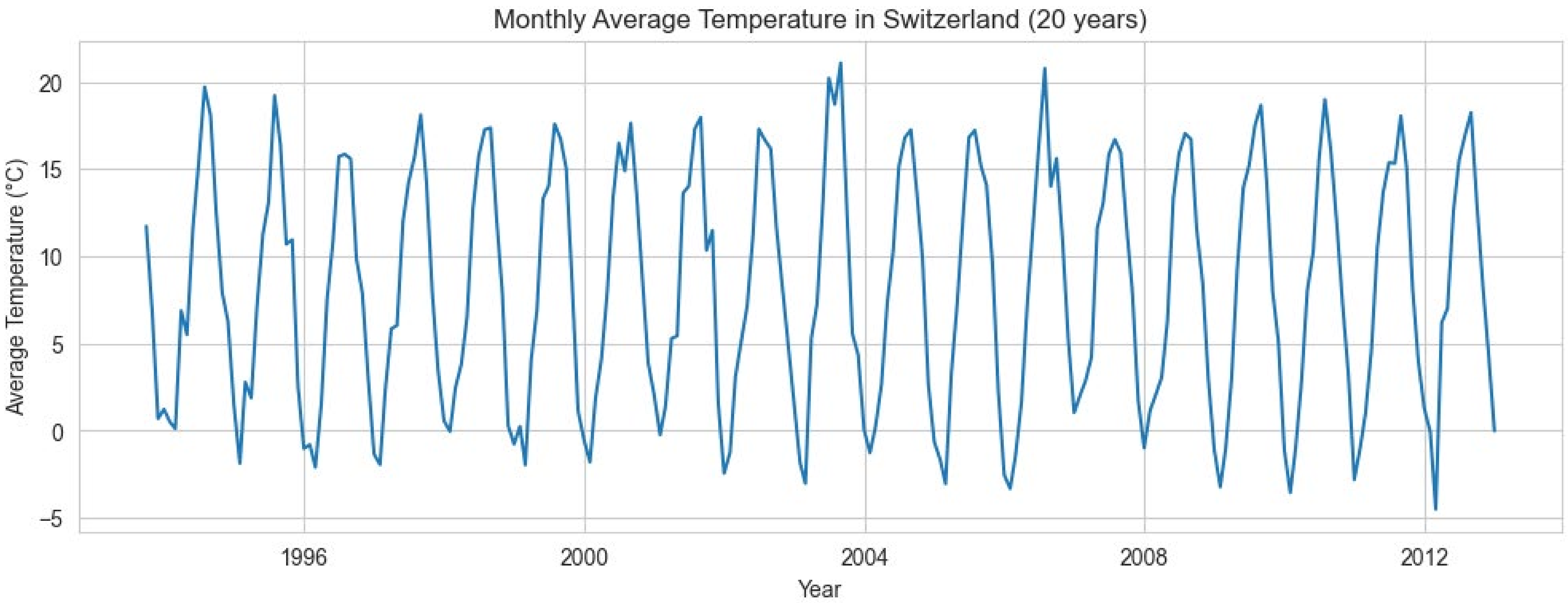
With **missing data**, start by asking the following questions

- What is the underlying **reason** for the missing data?
- Is the data **missing at random or is there a pattern?**
- How much data is missing, and could imputing it introduce **bias** or distort the analysis?
- Would **removing** the missing data be more appropriate in the context?

Based on the answers decide whether to proceed with **imputation**

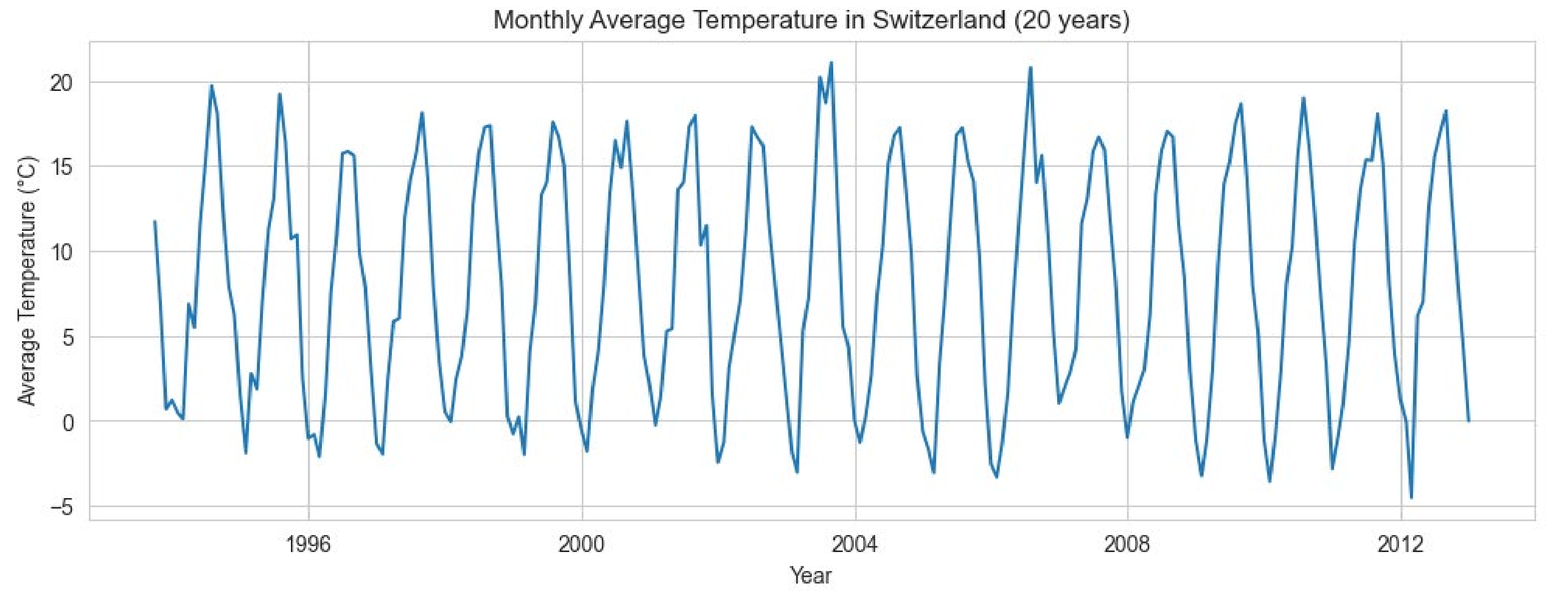
- Filling: forward, backward, mean
- Interpolation: nearest, linear, polynomial, seasonal

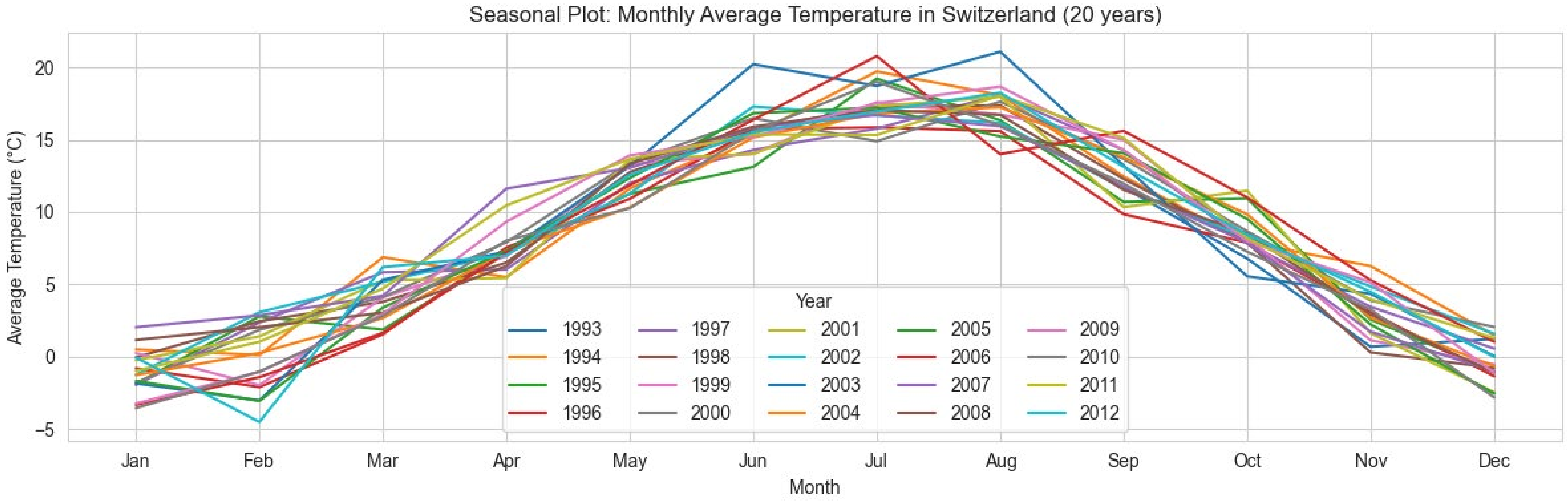
Visualization – Time plot



Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

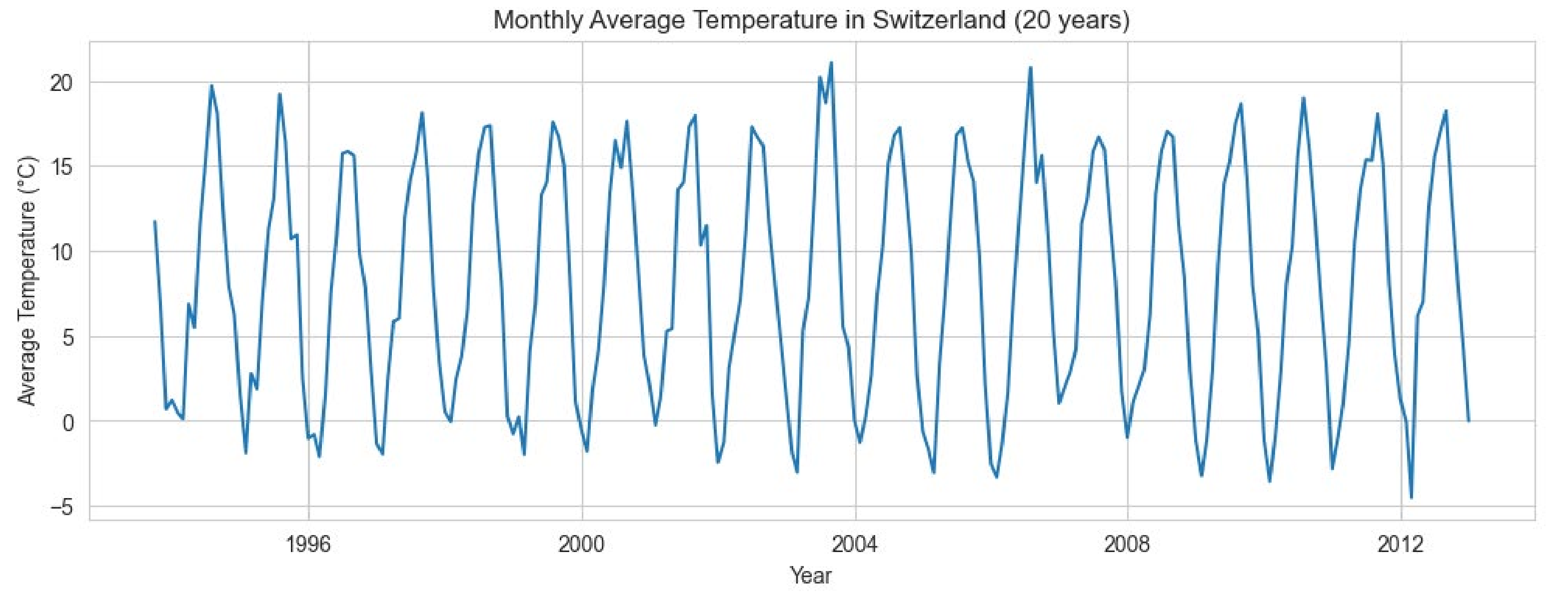
Visualization – Seasonal plot

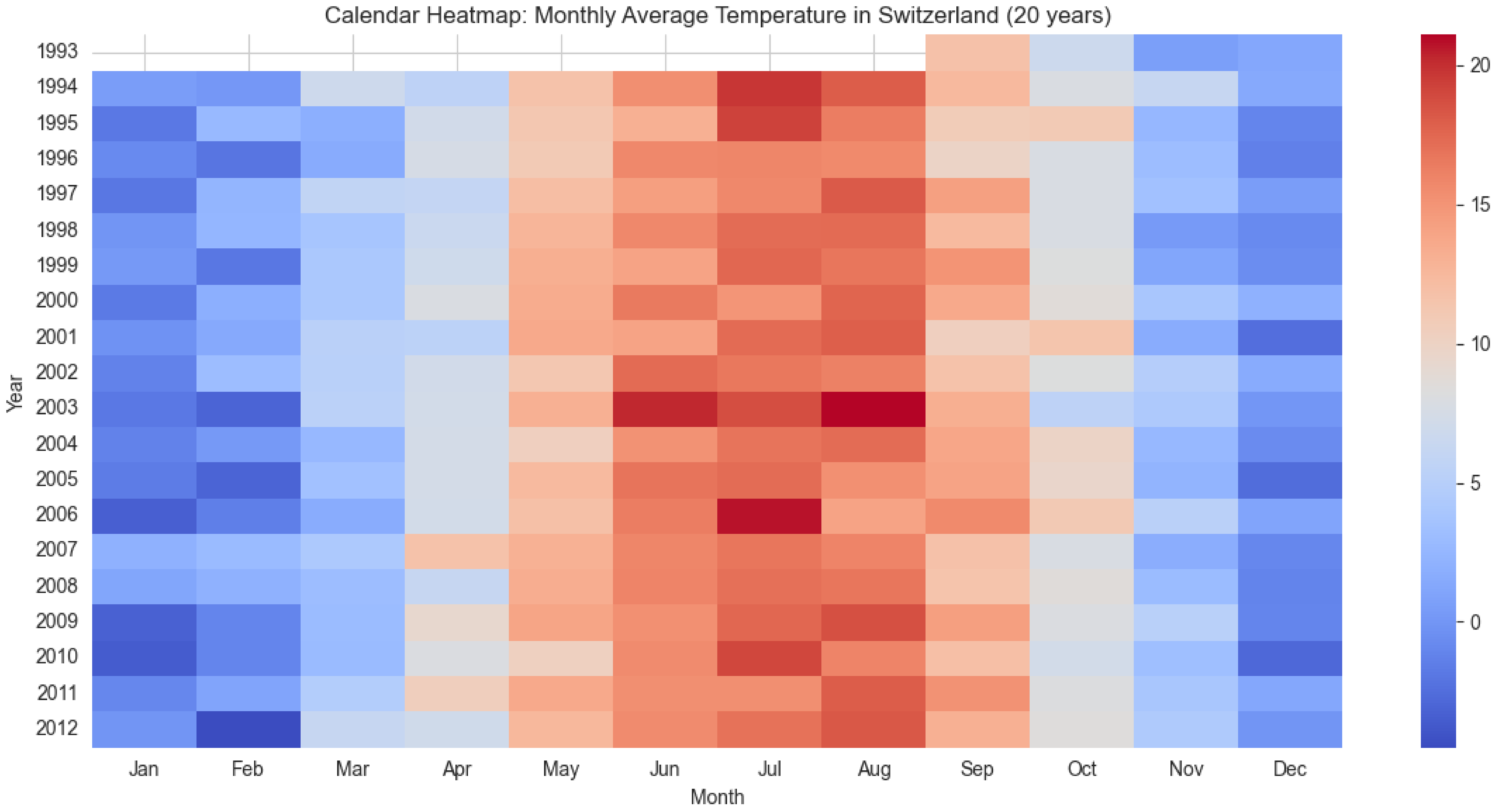




Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

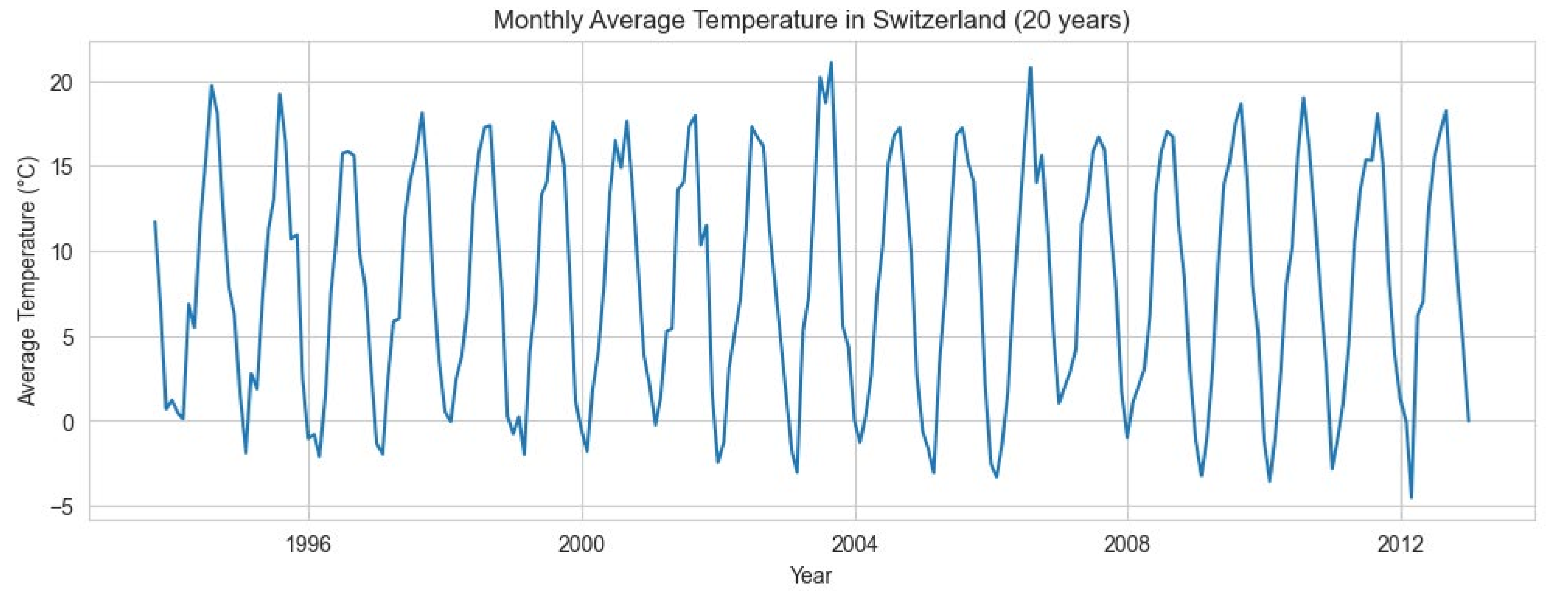
Visualization – Heatmap

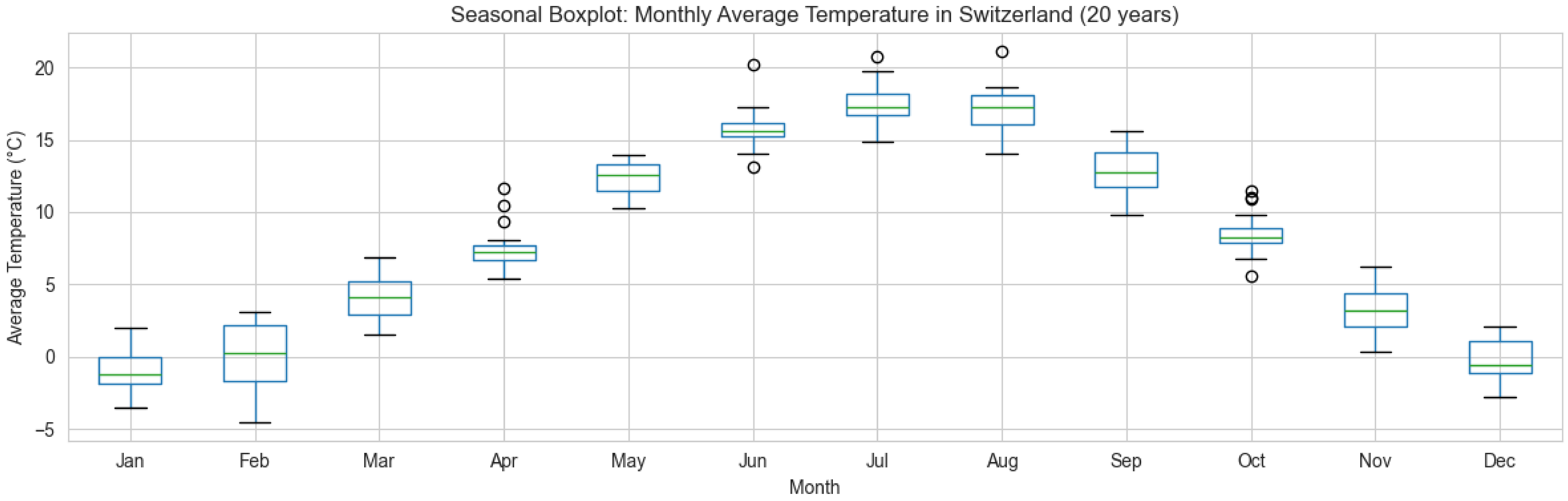




Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

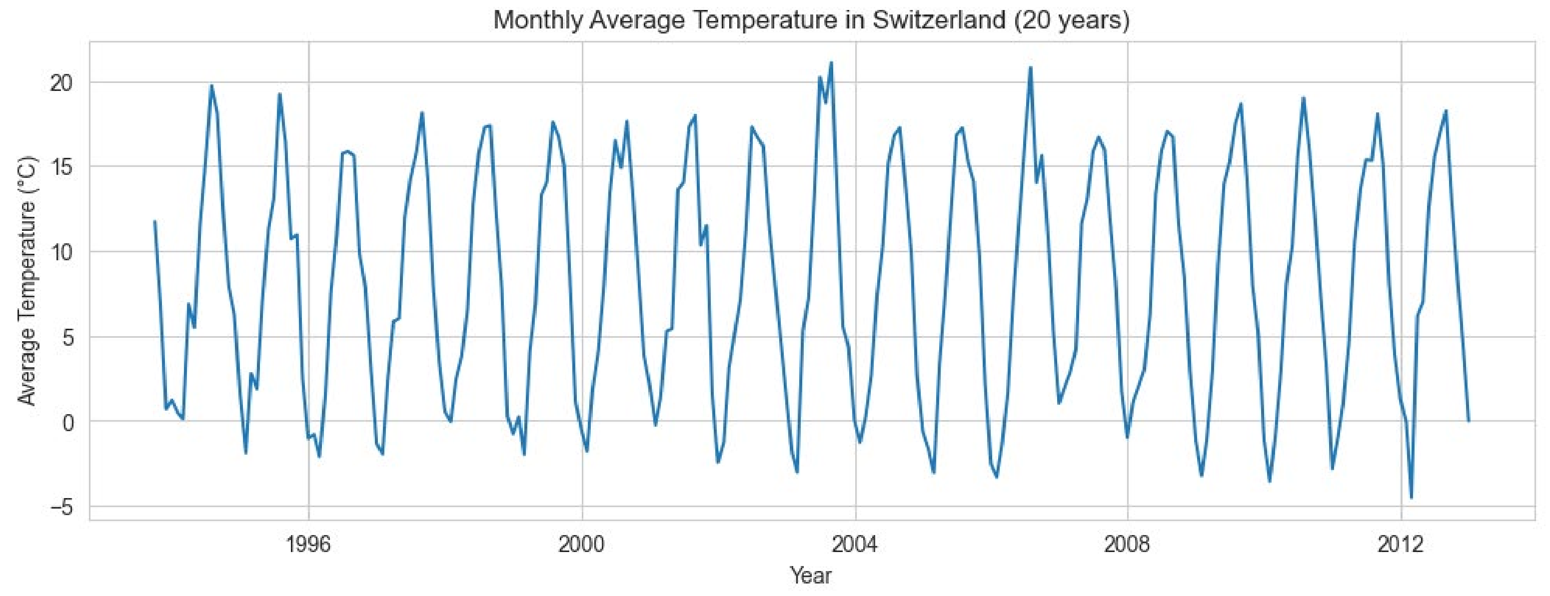
Visualization – Seasonal boxplot

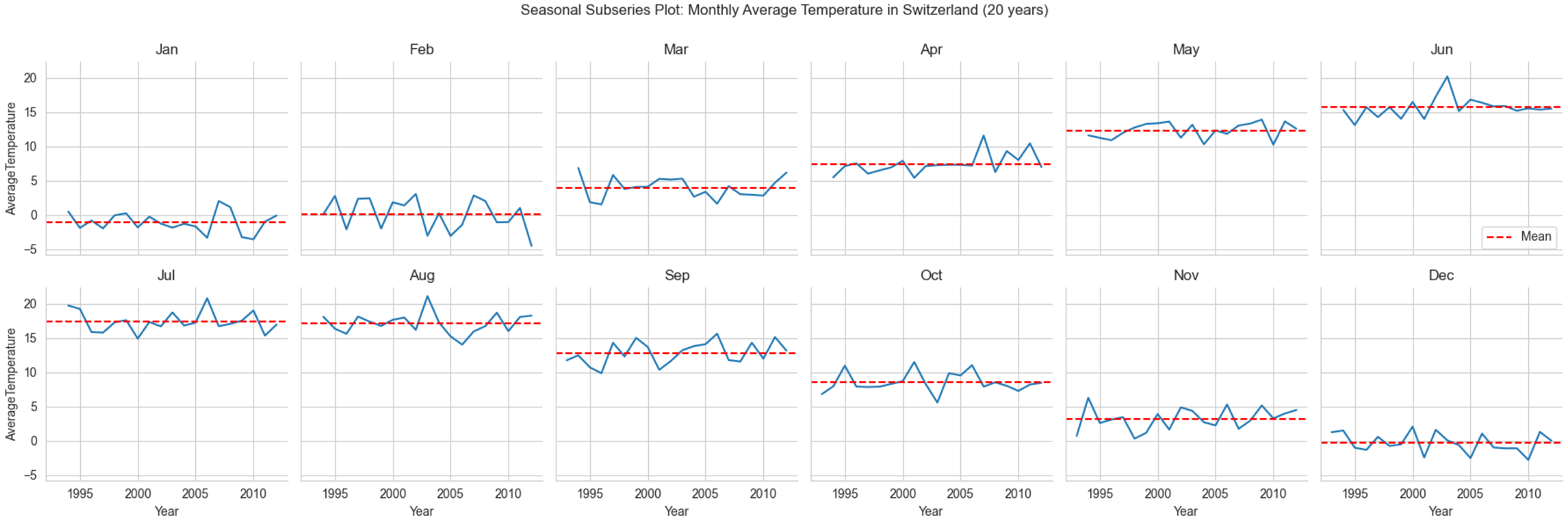




Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

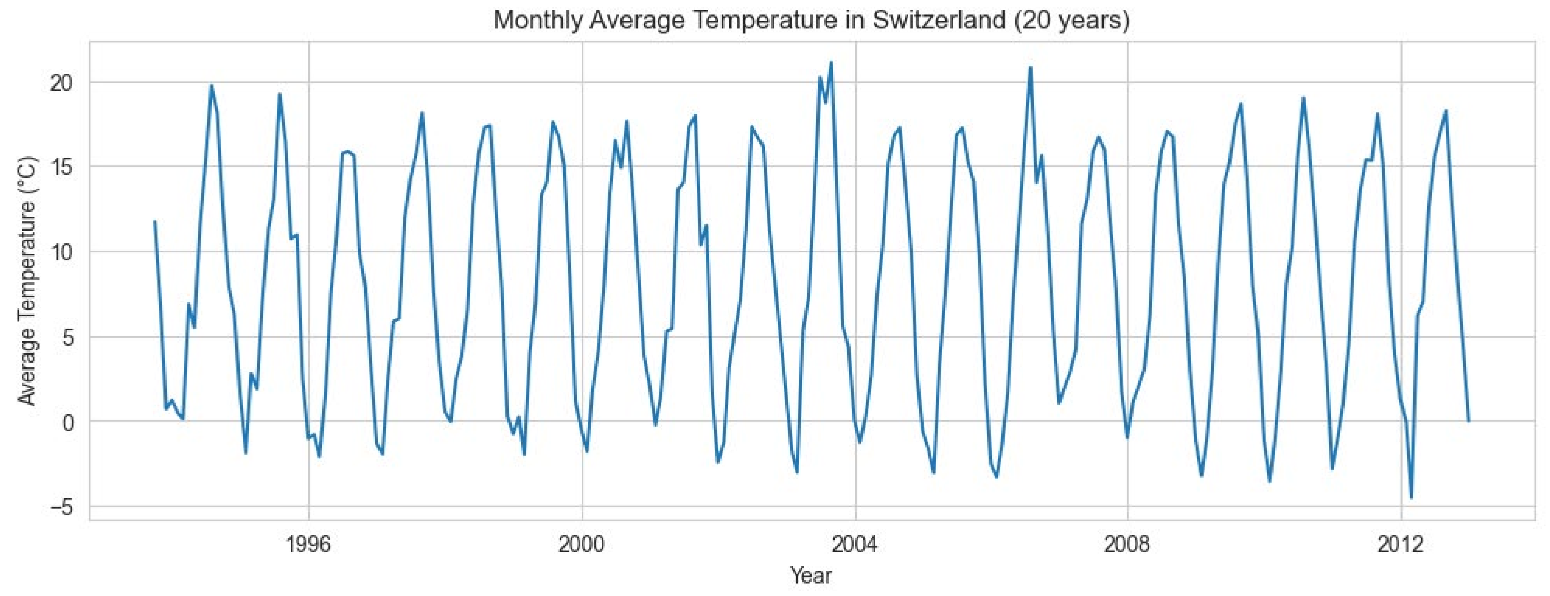
Visualization – Seasonal subseries plot





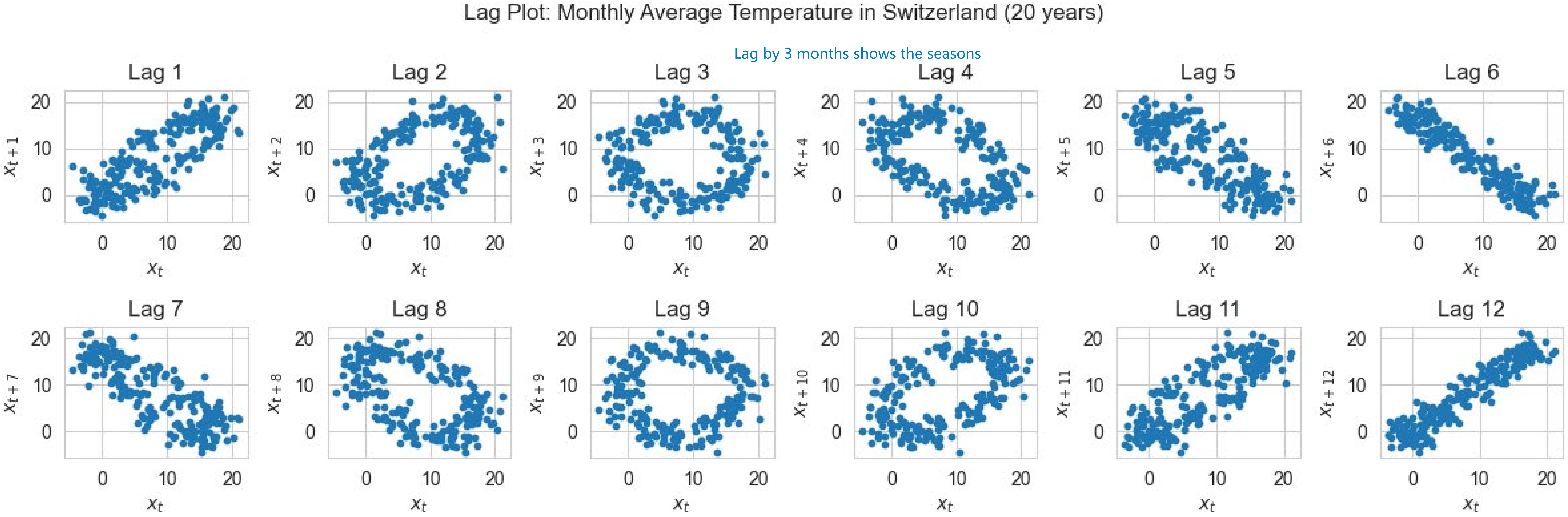
Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Visualization – Lag plot



Lag n: Shift time series data by n (to the left, in this case by n months)
Then plot time series against itself

Lag by 3 months shows the seasons



high linear correlation

Berkeley Earth. (2024). Earth Surface Temperature Data.
Retrieved from https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Exercise

Note: Worked on this & getting the python env up and running
same structure as folders on Ilias

Select 2-3 time series datasets: opendata.swiss, ec.europa.eu/Eurostat, Kaggle, etc.

Solve exercise in the exercises fold

Reflect on the data generation process:

- How was the data collected? Time interval, time range, etc.
- Which factors/laws influence the data?
- Which patterns do you expect to find?

We can solve the practical exercises as a group of 4 (Sofia, Dave, you & me)

Review notebooks of Modern Time Series Forecasting with Python – Chapter 2 (GitHub repository)

Load and **visualize** datasets with Python.

Analysis:

- Explain and interpret the plots. How noisy is the data?
- Review patterns, do they match your expectations given your understanding of the DGP?
- Review outliers, interpret them, and match with corresponding real-world events when applicable.