KI Service Zentrum by Hasso-Plattner-Institut

HPI

Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

Time Series Forecasting

1.4 Missing Data and Outliers

Mario Tormo Romero

Design IT.
Create Knowledge.

www.hpi.de

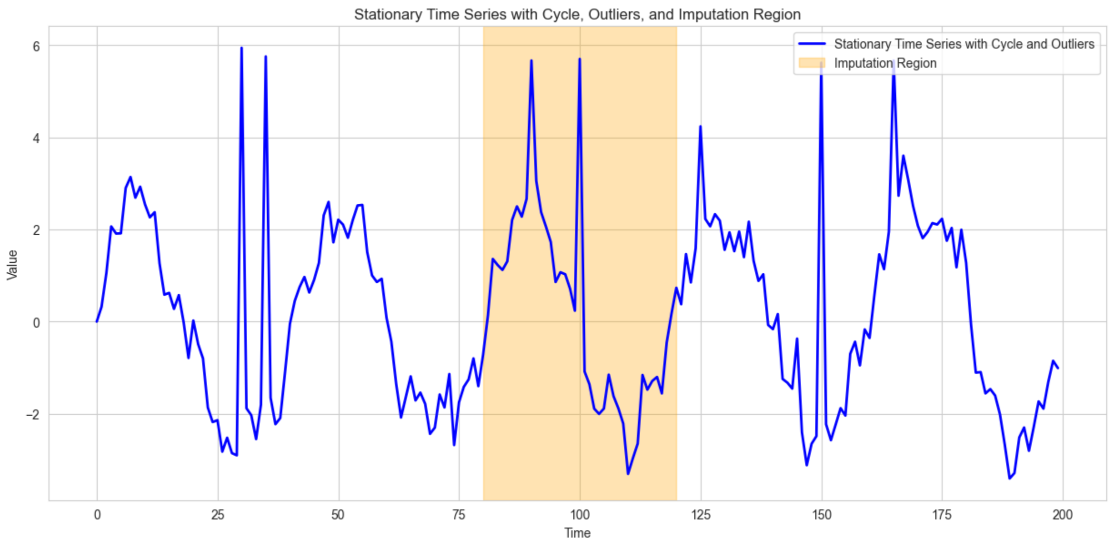
What we`ll cover in this video

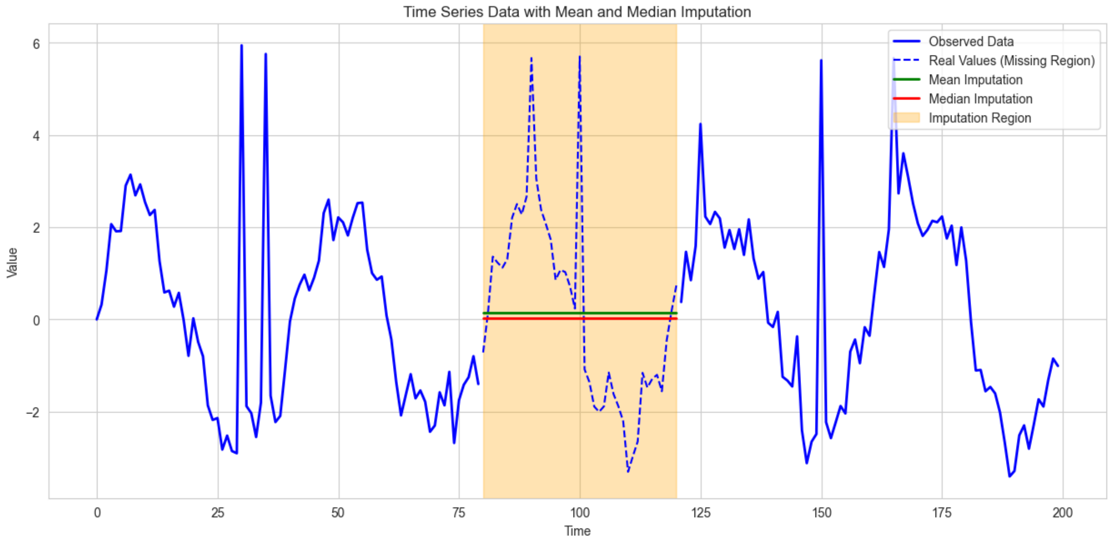
- Dealing with Missing Data
- Detecting Outliers

Dealing with Missing Data

Classical Methods: Mean & Median Imputation

- Replace missing values with the average or median of available data, either globally or within a time window (e.g., daily, weekly)
- Simple, quick to compute
- Median imputation is more robust and suitable when data is skewed or contains extreme values
- It doesn’t preserve the temporal dynamics of the data



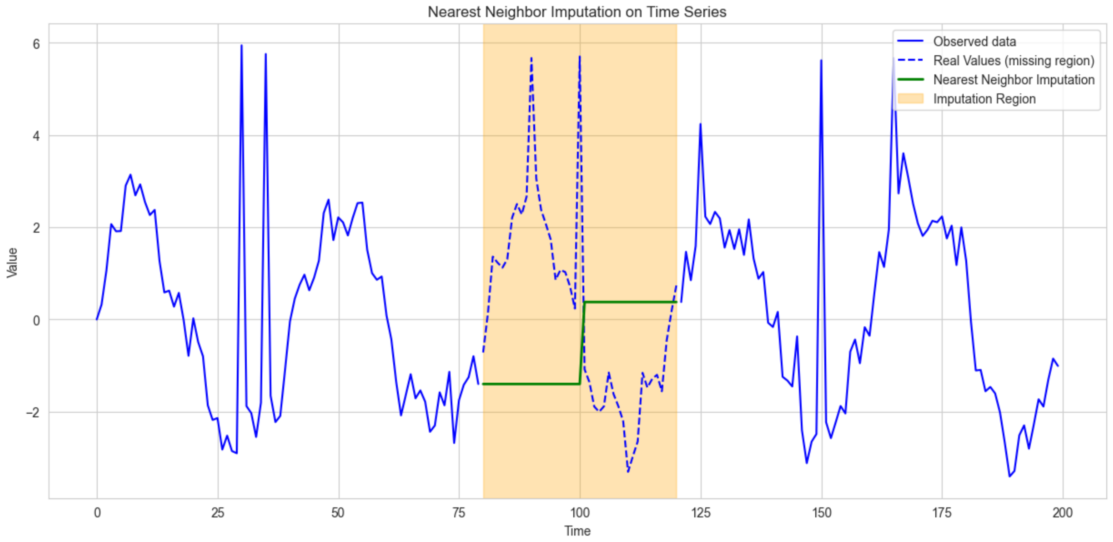


Dealing with Missing Data

Fordward Fill and Backward Fill Imputation

- **Last Observation Carried Forward (LOCF)**
  - Also known as Forward Fill
  - Replaces missing values with the last known observation.
  - Assumes that the previous value is a reasonable estimate for the missing point.

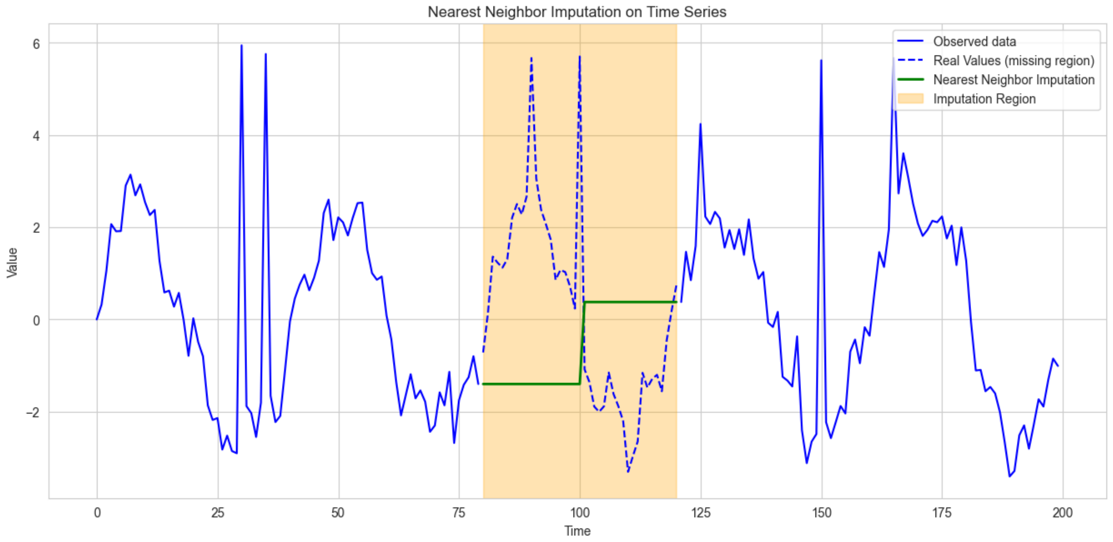
- **Next Observation Carried Backward (NOCB)**
  - Also known as Backward Fill
  - Replaces missing values with the next known observation.
  - Useful when future values are more representative.



Dealing with Missing Data

Nearest Neighbour Imputation

- Fills each missing value with the nearest observed value in time, whether before or after the gap.
- It can be understood as a more flexible mix of Forward Fill and Backward Fill Imputation.
- Maintains existing data levels without creating new intermediate values.
- Results in a step-like pattern.



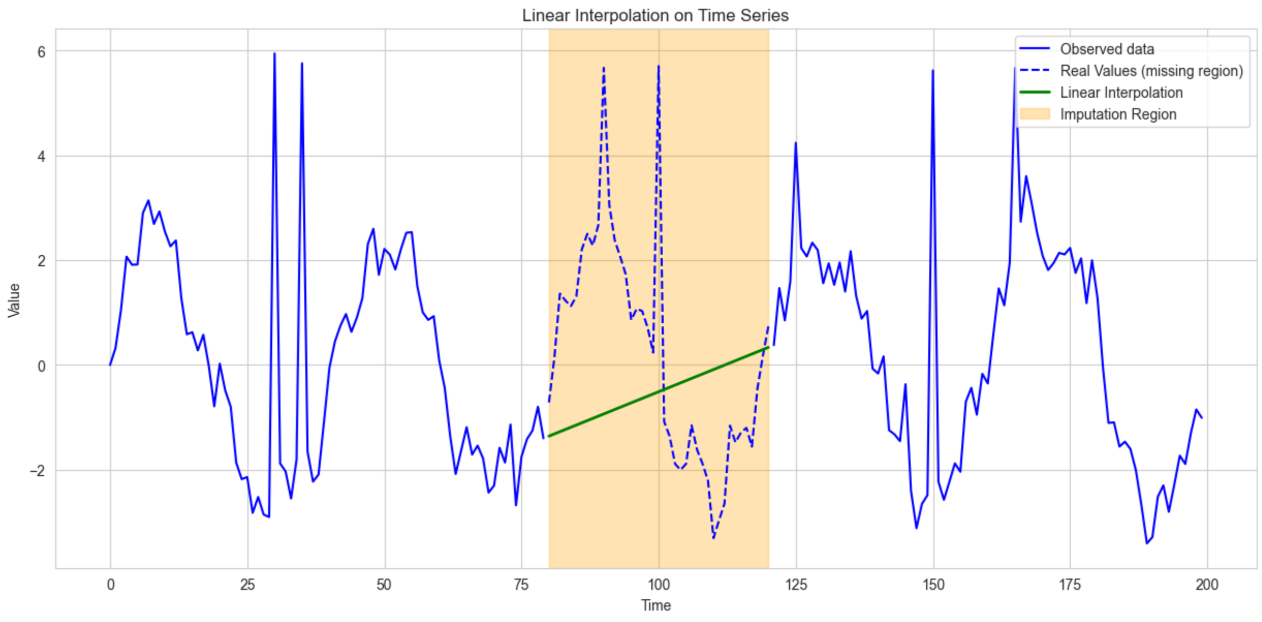
Dealing with Missing Data

Linear Interpolation

- Fills missing values by connecting the nearest known points with a straight line and estimating intermediate values along that line.

- Creates a smooth transition between observed data points.

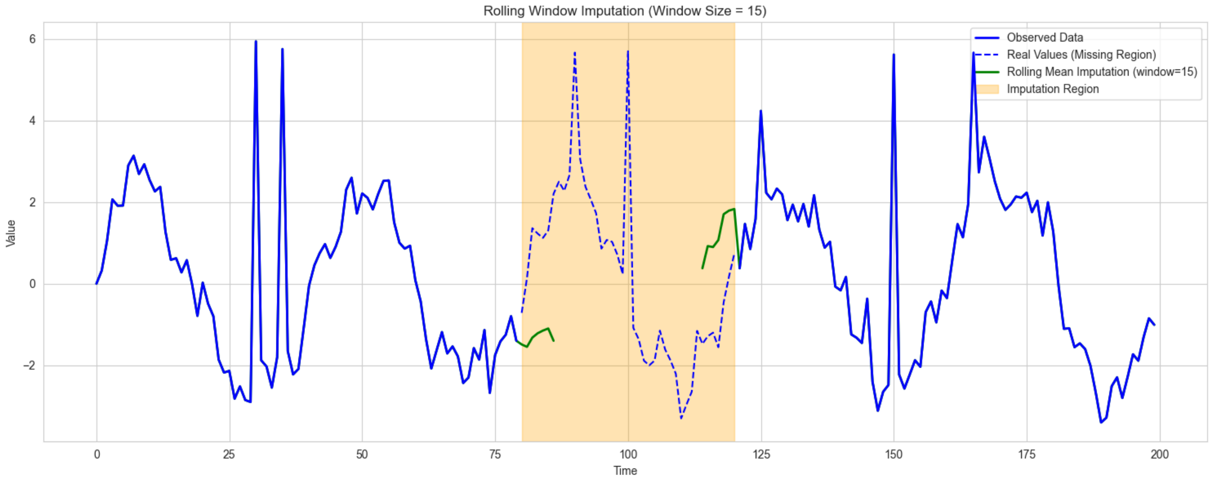
- Assumes changes between points happen at a constant rate.



Dealing with Missing Data

Rolling Window Imputation

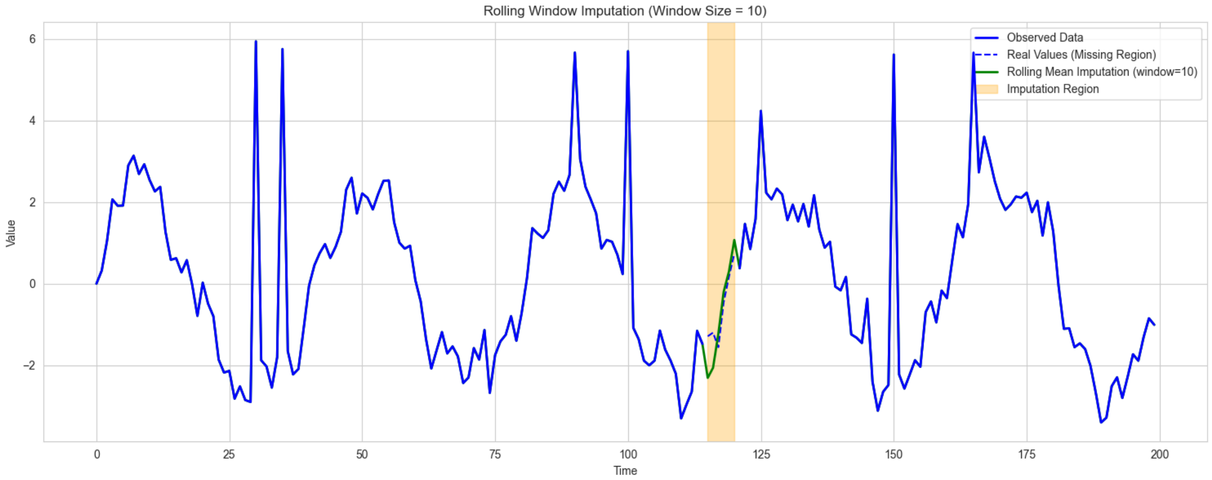
- Missing values are filled by averaging neighboring points within a window around the gap.
- If the missing gap is larger than the rolling window size, only the values near the edges of the gap can be imputed.
- The middle of a large gap remains unimputed, because there are no valid neighbors within the window to calculate the average.
- Rolling window imputation works well for individual missing values or small gaps but struggles with large continuous missing segments..

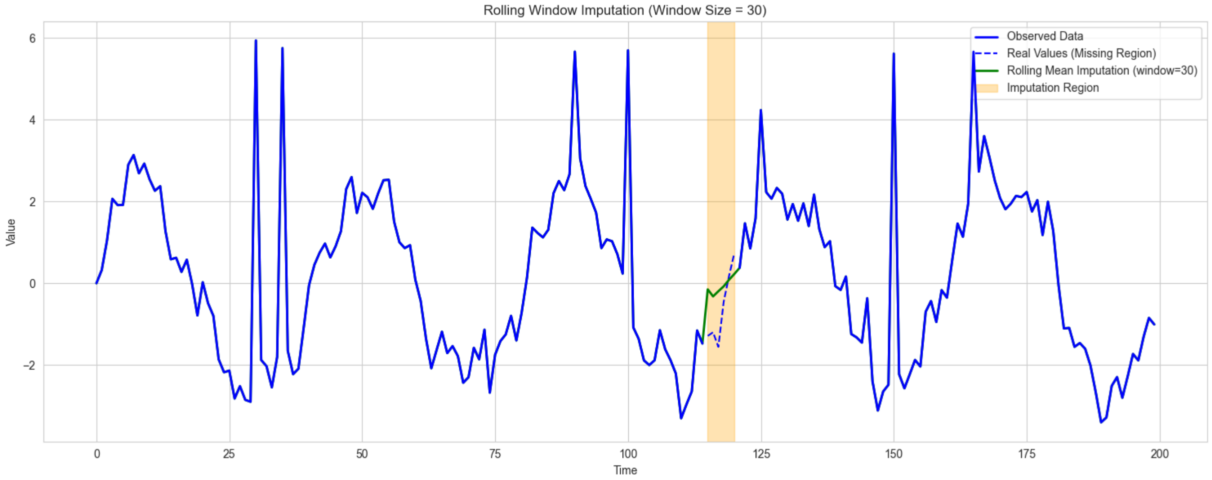


Dealing with Missing Data

Rolling Window Imputation

- When the gap size is smaller or equal to the window size, all missing points can be imputed.
- The rolling average smooths over the gap, using neighboring data on both sides. This helps maintain local trends and patterns better than mean or median imputation.
- Large windows can oversmooth, while small windows might not capture enough context.

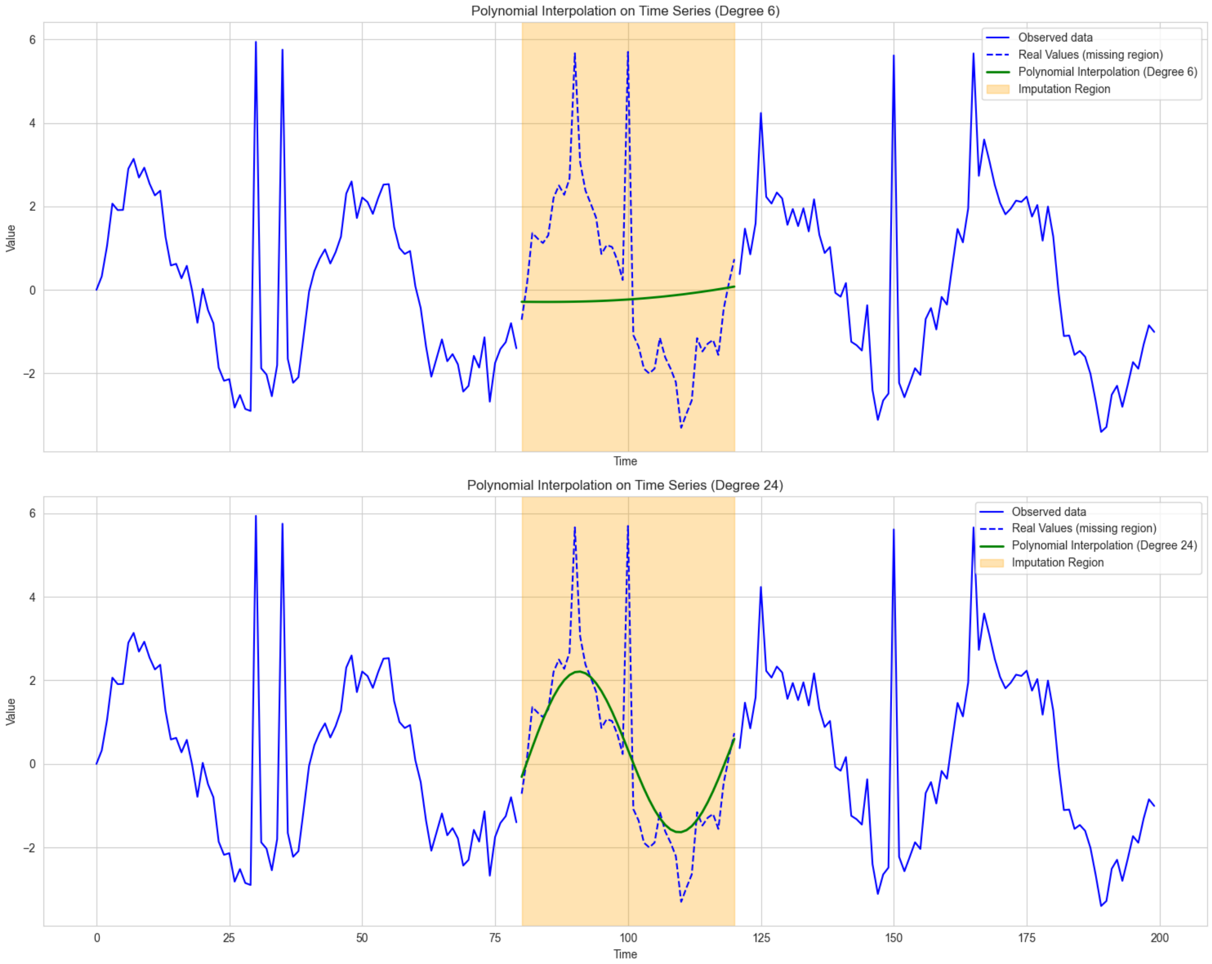




Dealing with Missing Data

Polynomial Interpolation

- Fills missing values by fitting a polynomial curve through the known data points to estimate the missing ones, capturing non-linear relationships between data points
- The degree of the polynomial controls the curve's flexibility — higher degrees allow more complex shapes
- Suitable when data is expected to follow a smooth, curved trend or on small to moderate gaps, if the underlying pattern is non-linear
- Overfitting risk: High-degree polynomials can introduce unrealistic oscillations
- Sensitive to noise and outliers. **Splines** are an extension that use piecewise polynomials, reducing oscillations and improving stability across longer series.



Detecting Outliers

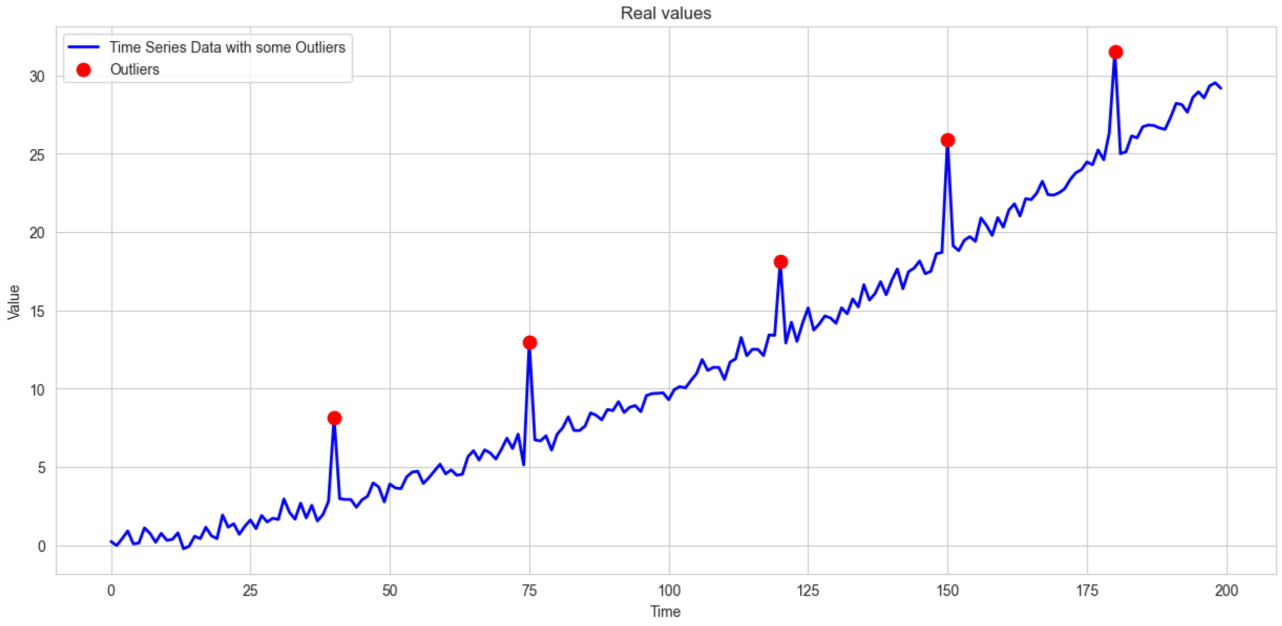
Outliers in Time Series

- Outliers are data points that significantly differ from the overall pattern.

- They can indicate errors, unusual events, or important signals.

- Identifying outliers is essential to improve the accuracy of models and analyses.

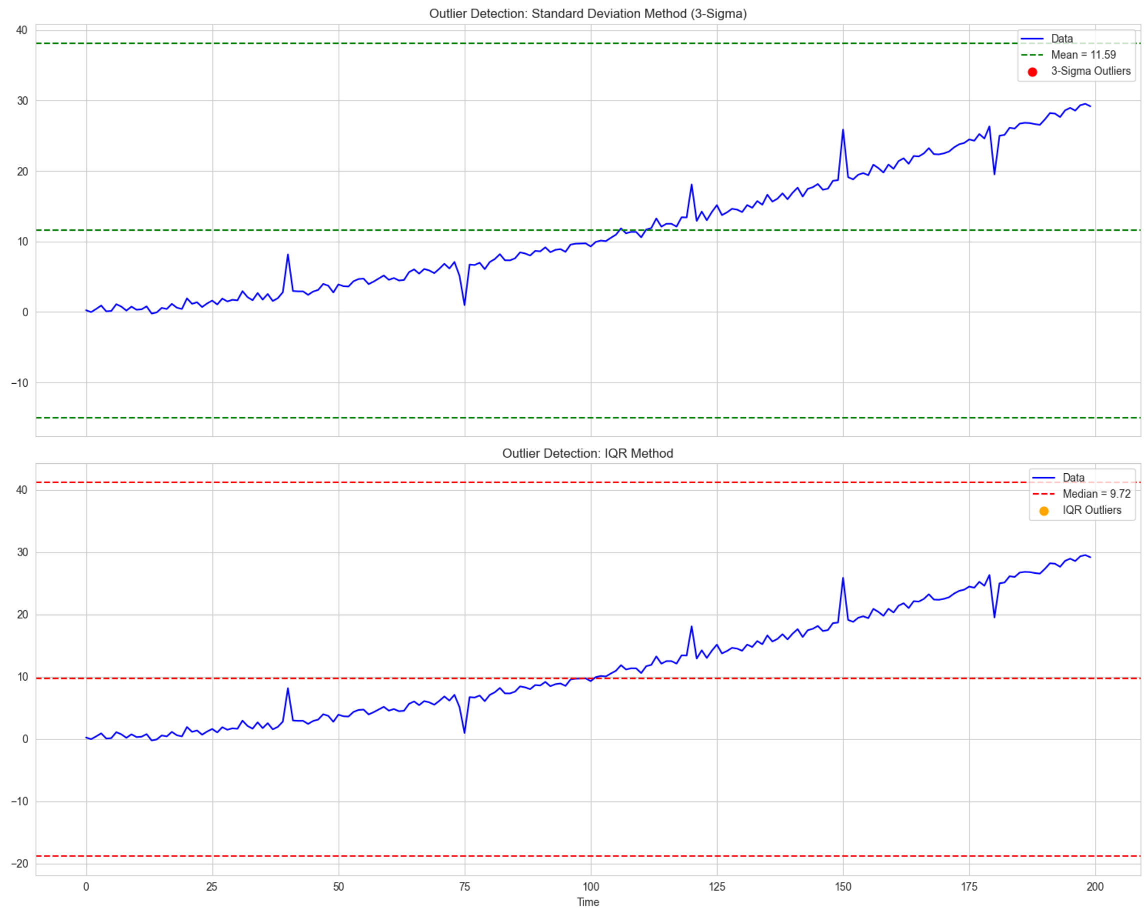
- Different methods exist, ranging from simple statistical rules to advanced machine learning techniques.



Detecting Outliers

Statistical Methods: Standard Deviation and IQR

- **Standard Deviation Method**
  - Flags points that fall beyond a defined number of standard deviations from the mean (commonly ±3σ).
  - Best for normally distributed data.
  - Simple and fast, but sensitive to skewed data and existing outliers, and can miss contextual anomalies in time series with trends or seasonality.
- **Interquartile Range (IQR) Method**
  - Detects outliers outside the range:
  - [Q1 - 1.5 IQR, Q3 + 1.5 IQR], where IQR = Q3 - Q1.
  - More robust to skewed data compared to standard deviation, best for detecting global outliers, not tailored for temporal patterns.
  - Non-parametric: doesn't assume a specific data distribution.



Detecting Outliers

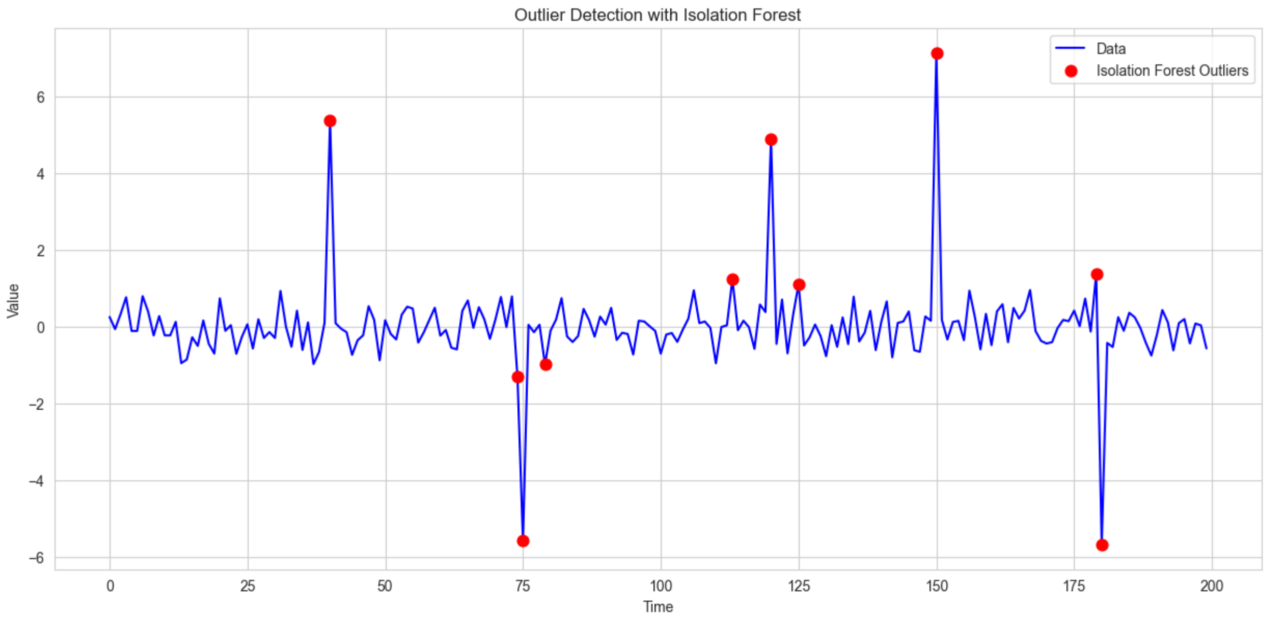
Statistical Methods: Standard Deviation and IQR

- **Standard Deviation Method**
  - Flags points that fall beyond a defined number of standard deviations from the mean (commonly ±3σ).
  - Best for normally distributed data.
  - Simple and fast, but sensitive to skewed data and existing outliers, and can miss contextual anomalies in time series with trends or seasonality.
- **Interquartile Range (IQR) Method**
  - Detects outliers outside the range:
  - [Q1 - 1.5 IQR, Q3 + 1.5 IQR], where IQR = Q3 - Q1.
  - More robust to skewed data compared to standard deviation, best for detecting global outliers, not tailored for temporal patterns.
  - Non-parametric: doesn’t assume a specific data distribution.

Detecting Outliers

Isolation Forest

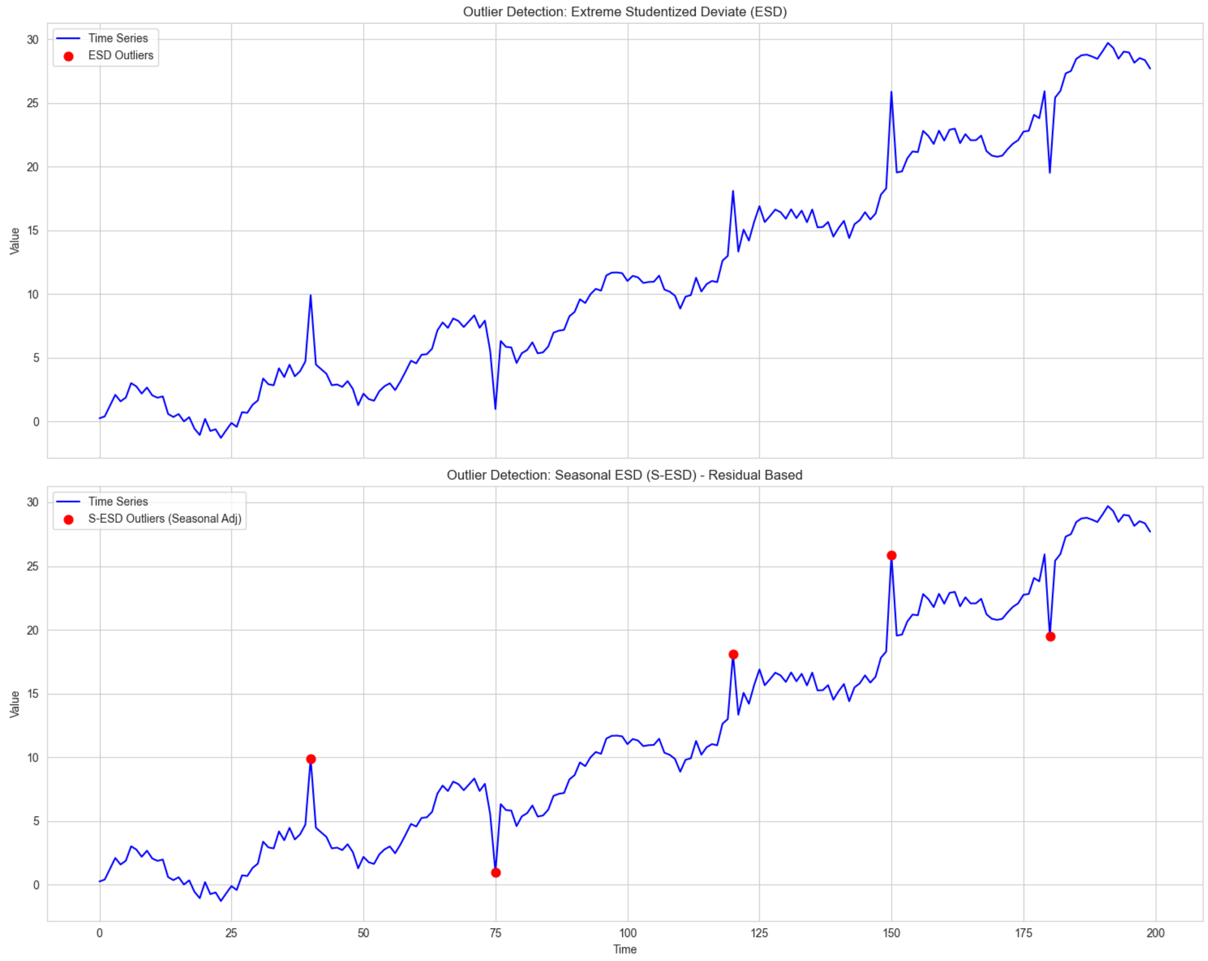
- Works by randomly partitioning data; anomalies are isolated faster due to their uniqueness.
- No assumptions about data distribution.
- Effective for high-dimensional and complex data.
- Detects both global and contextual anomalies.
- Can be applied to time series after feature extraction (e.g., trend, seasonality, cycles).



Detecting Outliers

Extreme Studentized Deviate (ESD) & Seasonal ESD (S-ESD)

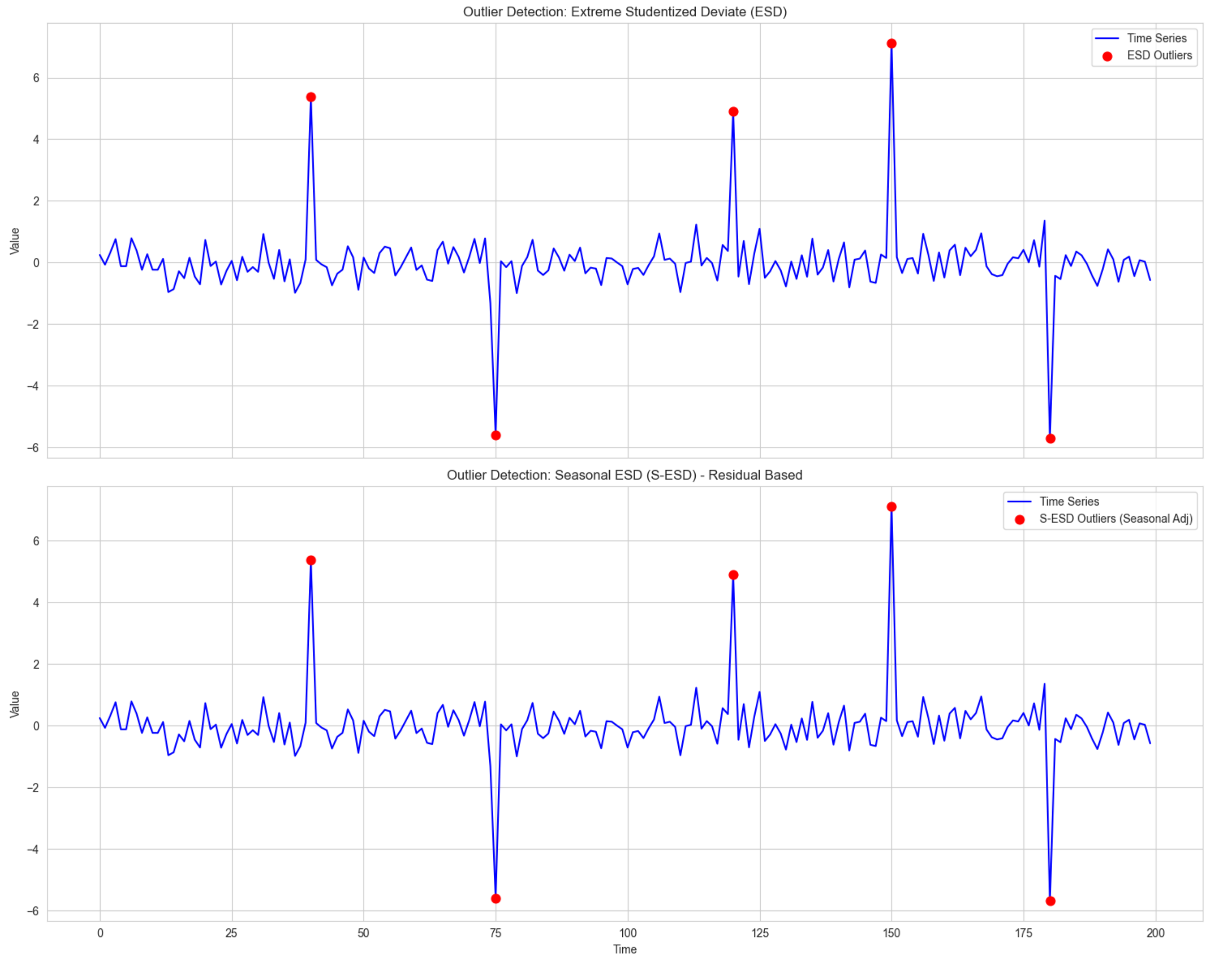
- **Extreme Studentized Deviate (ESD)**:
  - Statistical test designed to detect one or more outliers in a univariate dataset.
  - Iteratively tests whether the most extreme data point is an outlier.
  - Assumes data is approximately normally distributed and stationary.
  - Useful for detecting global anomalies.
- **Seasonal ESD (S-ESD)**:
  - Extension of ESD tailored for time series with seasonality.
  - Applies ESD on seasonally adjusted residuals to detect anomalies that repeat or vary with seasons.
  - Helps identify outliers while accounting for recurring patterns.



Detecting Outliers

Extreme Studentized Deviate (ESD) & Seasonal ESD (S-ESD)

- **Extreme Studentized Deviate (ESD)**:
  - Statistical test designed to detect one or more outliers in a univariate dataset.
  - Iteratively tests whether the most extreme data point is an outlier.
  - Assumes data is approximately normally distributed and stationary.
  - Useful for detecting global anomalies.
- **Seasonal ESD (S-ESD)**:
  - Extension of ESD tailored for time series with seasonality.
  - Applies ESD on seasonally adjusted residuals to detect anomalies that repeat or vary with seasons.
  - Helps identify outliers while accounting for recurring patterns.



What we‵ve learnt

- Handling Missing Data:
  - Multiple strategies: from simple (mean, median, forward/backward fill) to more advanced (interpolation, rolling methods).
  - No one-size-fits-all: Choose based on gap size, data patterns, and modeling goals.
- Outlier Detection:
  - Classical methods: Standard Deviation, IQR. Useful but sensitive to trends and seasonality.
  - More robust approaches: Isolation Forest, ESD, and S-ESD. Adaptable to structured time series with preprocessing.
- Understanding your data’s structure (trend, seasonality, noise) is essential before applying these methods.
- Preprocessing like detrending and deseasonalizing often improves detection and imputation results.