Introduction to Time Series and Forecasting

Forecasting

- **Definition:** Process of using historical data to predict future values, akin to regression in supervised learning, focusing on real-valued number predictions.
- **Key Difference from Regression:** Unlike regression's multiple inputs for one output, forecasting uses past signal values to predict future values.
- **Business Application:** Essential for reducing risks and uncertainties in business operations by predicting future events.

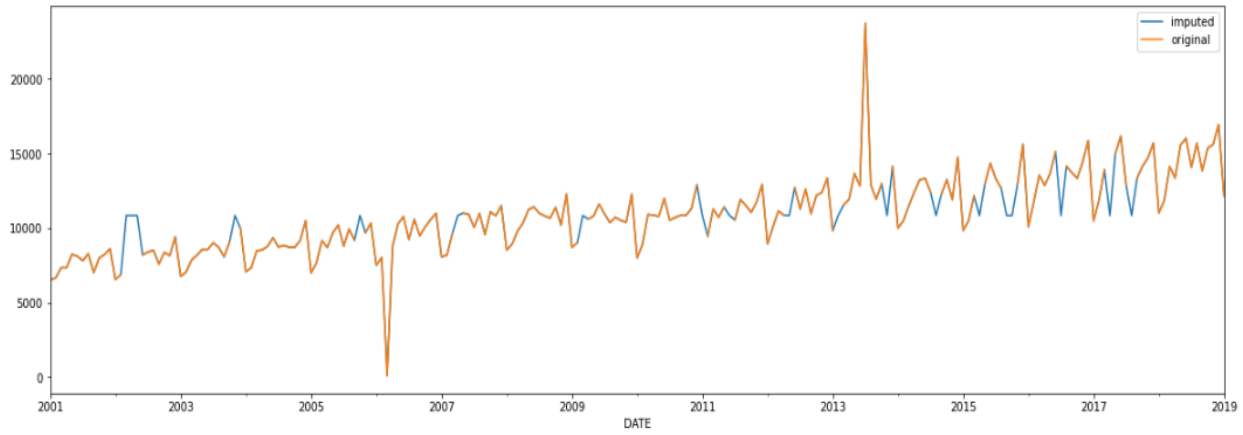
Time Series Data

- **Definition**: Measurements of a variable over time, indexed by timestamps.
- **Requirements:** Date/timestamp (t) and quantity (y).
- **Observations:** Includes sales, revenues, inventories.
- **Types:**
    - **Univariate:** Single variable data.
    - **Multivariate:** Multiple variables data.
- **Dependency:** Order is crucial; altering it changes the data structure.

Handling Missing Values:

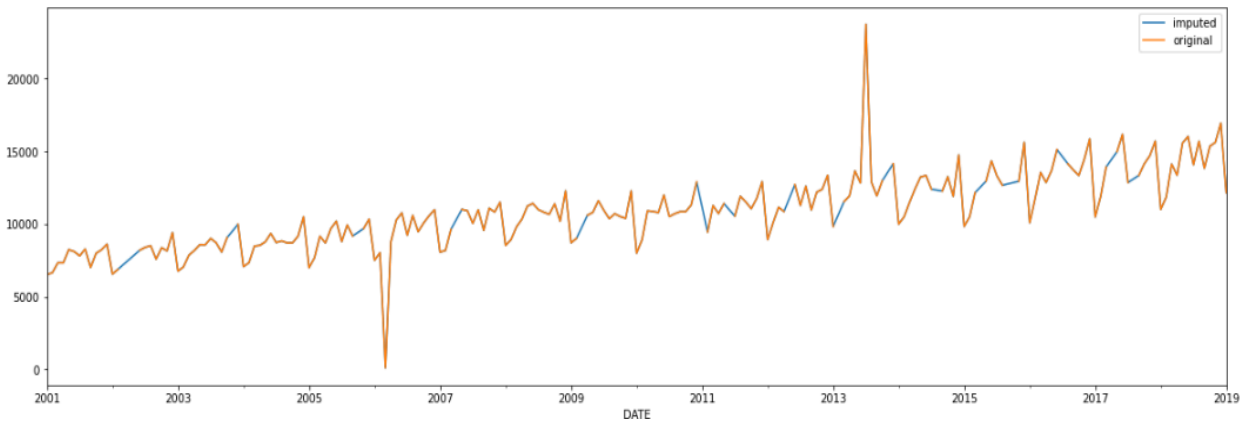
- **Rule:** Time series data must not have missing values; observations need to be contiguous.
- **Imputation Methods:** Mean/Median, Interpolation, Moving Averages.

Method 1: Imputing Using Mean/Median:



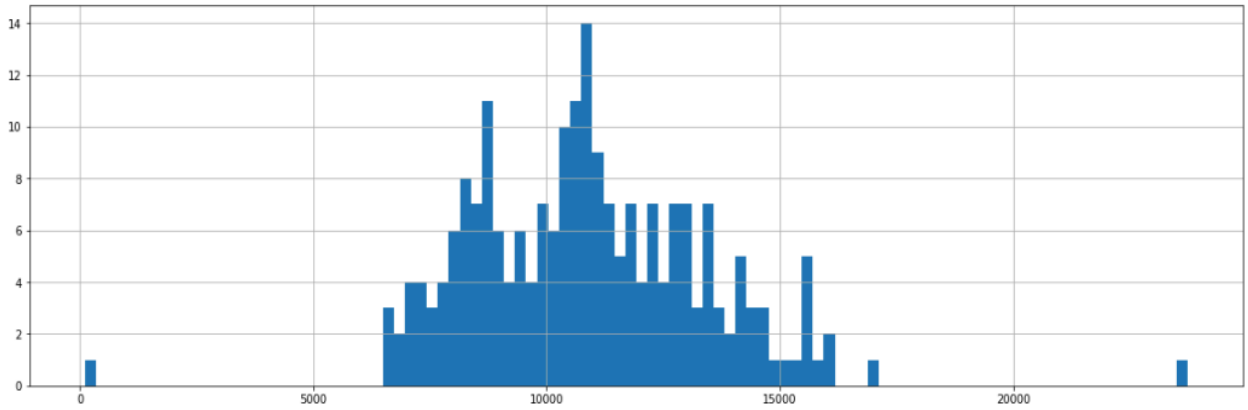
- **Strategy:** Fill missing values with mean or median of entire dataset.
- **Outcome:** Sharp increments/decrements at imputed points due to global mean/median imputation.

Method 2: Linear Interpolation:



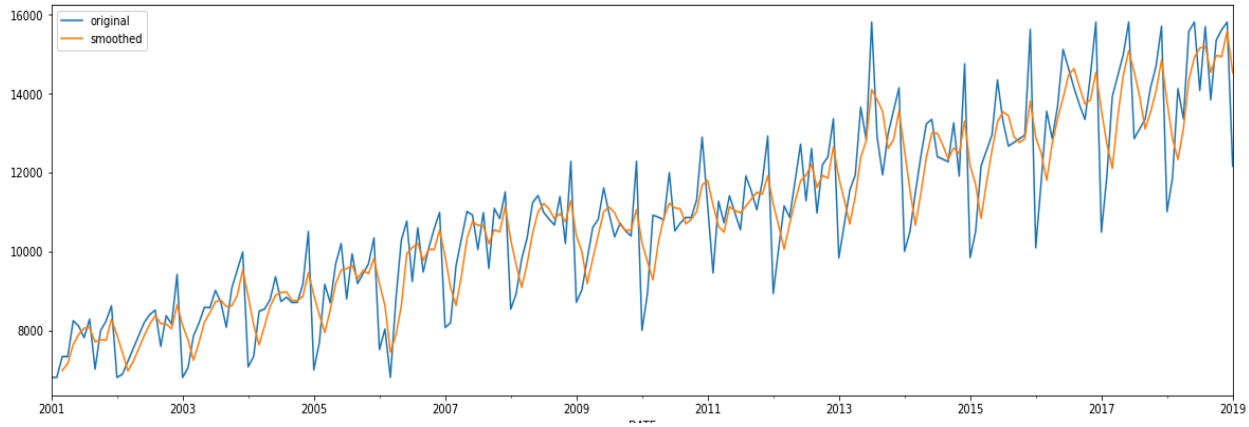
- **Technique:** Average of points before and after missing value for imputation.
- **Advantage:** Prevents under/overestimation by using adjacent values.
- **Outcome:** Imputed values fall naturally within two finite points, avoiding forced appearance.

Handling Anomalies:



- **Definition:** Anomalies are abnormal data points that stand out.
- **Causes:** Incorrect entries or valid but rare one-time events.
- **Goal:** Remove to prevent model bias towards rare events.
- **Methods:** Fixed number replacement, quantiles, robust scaling.
- **Identification:** Histograms help visualize and identify outliers.
- **Quantile Approach:** Exclude observations beyond 95th percentile or below 5th percentile to remove anomalies.

Moving Averages:



Moving averages with k=3

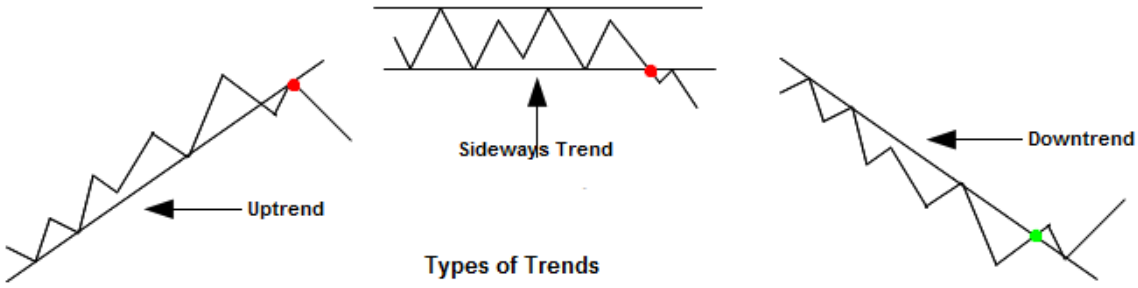
- **Definition:** Average of last k data points used to predict the next point.
- **Window Size (k):** Acts as a hyperparameter; determines smoothness of moving average curve.
- **Properties:**
    - Smoothness increases with larger k.
    - Reflects trends in time series (upward or downward).
- **Types:**
    - **Simple Moving Average:** Equal weights to past observations.

    - **Weighted Moving Average:** Different weights to past observations.
    - **Centered Moving Average:** Averages points before and after current point; not suitable for forecasting due to future data requirement.
- **Applications:** Forecasting, anomaly detection, handling missing values.

Time Series Decomposition:

- **Purpose:** Splits time series into components, each representing underlying patterns.
- **Components:**
    - **Trend:** General movement over time.
    - **Seasonality:** Patterns within seasonal periods.
    - **Residual:** Variations not explained by trend or seasonality.
- **Characteristics:** Trend and seasonality are systematic; Residual is irregular.

Trend:



Types of Trends

- **Definition:** Linear increase or decrease in series over time, non-repetitive.
- **Types:** Uptrend, downtrend, fluctuating.
- **Trend Line:** Smooth function tracing series trend, aids in future predictions.
- **Calculation:** Rolling/moving average or Linear Regression fitting.

Seasonality:

- **Definition:** Regular pattern in data at specific intervals (e.g., yearly, weekly).
- **Characteristics:** Occurs due to seasonal factors like time of year or week.
- **Calculation Steps:**
    - Subtract trend from original series.
    - Average results over each period (e.g., monthly averages over years).
- **Note:** Yearly series lack seasonality; possible to have multiple seasonalities (short-term and long-term).

Irregular Component:

- **Definition:** Random fluctuation in time series not explained by trend or seasonality.
- **Characteristics:** Assumed normal distribution, 0 mean, constant variance.
- Also Known As: Error, White Noise, Remainder.

Additive Decomposition:

- **Formula:** Time Series = Trend (b(t)) + Seasonality (s(t)) + Error (e(t)).
- **Characteristics:**
    - Error (e(t)) computed from actual values, scattered around zero if estimates are accurate.
    - Seasonal fluctuations independent of trend, constant seasonality.

Multiplicative Decomposition:

- **Principle**: Seasonal amplitude increases with trend.
- **Formula:** Time Series = Trend (b(t)) * Seasonality (s(t)).
- **Characteristic:** Seasonal fluctuations dependent on trend.