**<https://www.linkedin.com/advice/0/what-common-data-cleaning-techniques-time-series-ojnff#feature-engineering>**

**Why do we need to remove seasonality?**

The analysis and preprocessing of data around its seasonality is specific to time series data. Time series may have patterns which recur at a fixed period(season) & are predictable. This is called as time series seasonality. As opposed to that, cyclic patterns repeat themselves at a regular frequency and may not be seasonal. Though the seasonal component gives us additional information, "seasonal adjustment" or "deseasoning" is important for getting a clearer relationship between input and output data (Methods like differencing can be explored) After removing the seasonal component, the data becomes stationary. In stationary time series, the statistical measures like mean, median, variance do not vary with time.

Common data cleaning techniques for time series data in machine learning include handling missing values through imputation or seasonal decomposition, removing outliers using statistical methods or visual inspection, encoding categorical data with methods like label encoding, feature engineering to summarize, lag, or compute rolling statistics, normalization and scaling, differencing for stationarity, removing duplicates, handling time zones, resampling, dealing with irregular timestamps, and addressing seasonality and trends. The choice of techniques depends on data characteristics and modeling requirements, often requiring a combination of methods for effective data preparation.

**Feature engineering** is a data enhancement technique aimed at improving data quality and relevance for machine learning models. It involves creating new features or modifying existing ones to extract more insights and enhance model accuracy. Techniques include aggregation, decomposition, lagging, and combining data using methods like mean, median, moving average, lag, and polynomial expansion. These techniques help capture valuable information, temporal relationships, and nonlinear patterns in the data.

In time series data, categorical or ordinal features need to be converted into numerical values for machine learning models. Common encoding techniques include: Label Encoding: Assigning a unique integer value to each category. One-Hot Encoding: Creating a binary vector for each category, representing its presence or absence. Ordinal Encoding: Assigning numerical values based on inherent or logical order. Target Encoding: Assigning numerical values based on the mean or median of the target variable.

Data transformation is essential for preparing time series data for machine learning models. It involves applying mathematical or statistical functions to alter the data's scale, shape, or distribution. Common techniques include: Scaling or **Normalizing:** Reducing the range or variance of data values to ensure consistent scales across features. Logarithmic or Exponential Transformation: Addressing skewness by applying logarithm or exponentiation functions to the data. Differencing or Detrending: Removing trends or seasonality from the data to achieve stationarity. Power or Box-Cox Transformation: Linearizing the data and improving homoscedasticity through power transformations.

To remove outliers from a dataset, several techniques can be employed. Outliers are data points that significantly deviate from the normal range of values and can distort data patterns and machine learning model performance. Descriptive statistics like mean, standard deviation, quartiles, and interquartile range can be used to identify outliers by setting a threshold or criterion for their removal. Additionally, graphical methods such as box plots, scatter plots, or histograms can visually detect outliers based on data distribution and shape.

Handling missing values in time series data is crucial for maintaining data quality in machine learning models. Strategies for addressing missing values include: Deletion: Removing rows or columns with missing values, but this may lead to loss of valuable data. Imputation: Filling missing values with a constant (e.g., zero) or statistical measures like mean, median, or mode. Interpolation: Estimating missing values using linear or non-linear functions based on adjacent data points.