Time Series

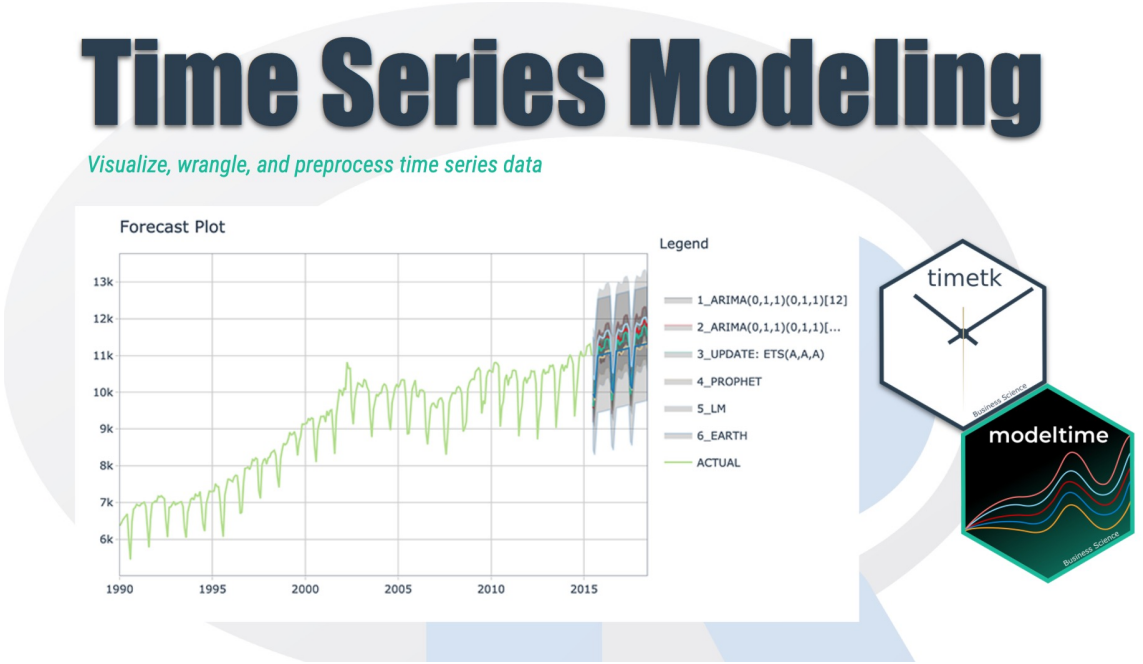
Multiple Linear Regression

Agenda

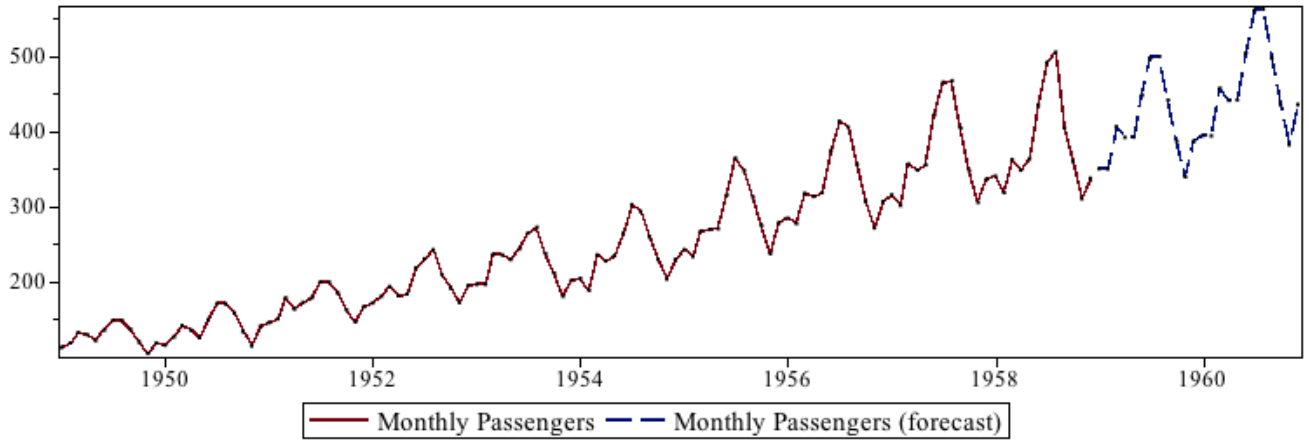
- Introduction Time Series
- Correlation vs AutoCorrelation
- Assumptions
- Linear regression in Time Series
- Regression vs AutoRegression
- Symbolic Regression

Time Series Introduction

- We're going to discover how it differs from other types of data you've previously encountered and why
- Time series is a sequence of information which attaches a time period to each value
- The value can be pretty much anything measurable.
- It depends on time in some way, like prices, humidity or number of people.
- As long as the values we record are unambiguous, any medium could be measured with Time series.
- There aren't any limitations regarding the total time span of our Time series.
- It could be a minute, a day, a month or even a century.
- All we need is a starting and an ending point.



Time Series Basics



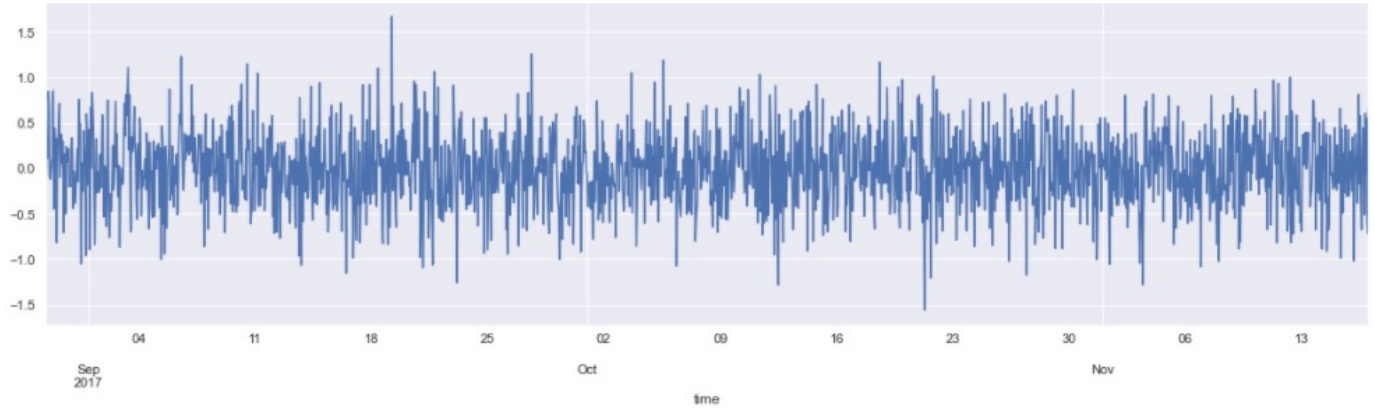
- Chronological Data
- Cannot be shuffled
- Each row indicate specific time record
- Train – Test split happens chronologically
- Data is analyzed univariately and multivariately (for given use case)
- Nature of the data represents if it can be predicted or not

time	IOT_Sensor_Reading	Error_Present	Sensor_2	Sensor_Value
2017-08-29 11:00:00	-0.015871	0.353986	-0.787655	0.008144
2017-08-29 12:00:00	-0.101576	0.353986	-0.787655	-0.029860
2017-08-29 13:00:00	-0.118241	0.353986	-0.787655	-0.021717
2017-08-29 14:00:00	-0.214262	0.353986	-0.787655	0.008144
2017-08-29 15:00:00	-0.249972	0.353986	-0.787655	-0.108583

Data Examining & Preprocessing

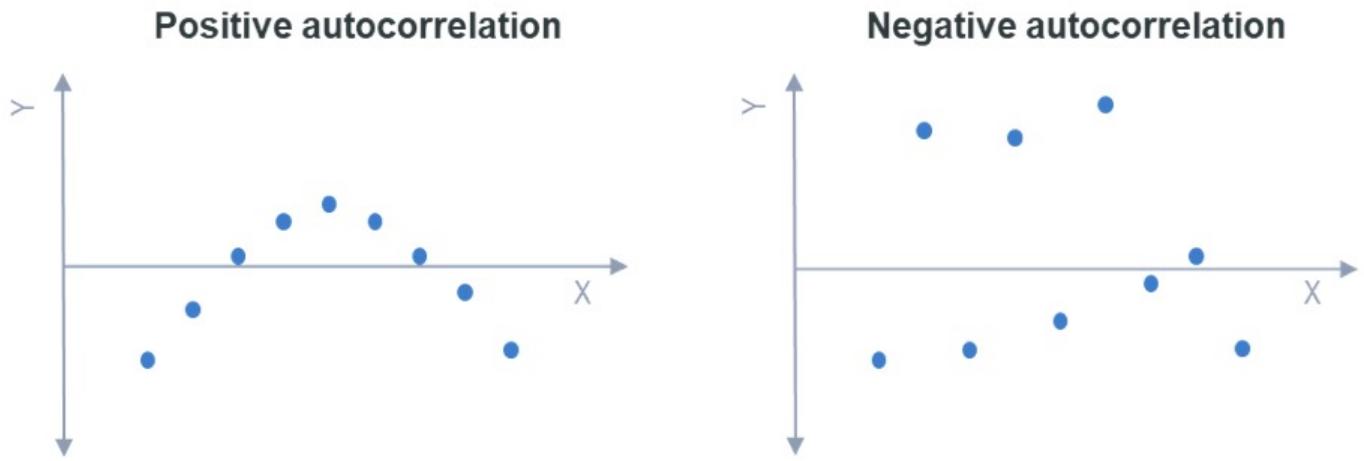
- Reading Data using Pandas – describe, head
- Check for null values
- Line plot for each feature
- Convert Date from String to Date time feature
- Setting the desired frequency
- Handling missing Values
- QQ plots

time	IOT_Sensor_Reading	Error_Present	Sensor_2	Sensor_Value
2017-08-29 11:00:00	-0.015871	0.353986	-0.787655	0.008144
2017-08-29 12:00:00	-0.101576	0.353986	-0.787655	-0.029860
2017-08-29 13:00:00	-0.118241	0.353986	-0.787655	-0.021717
2017-08-29 14:00:00	-0.214262	0.353986	-0.787655	0.008144
2017-08-29 15:00:00	-0.249972	0.353986	-0.787655	-0.108583



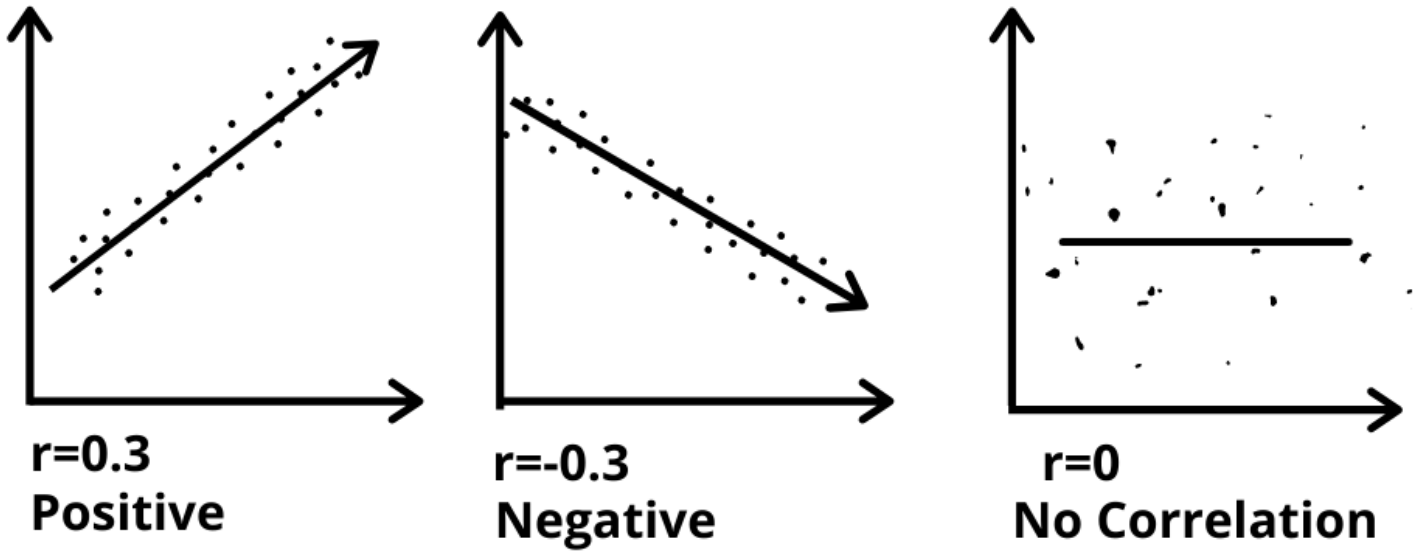
Correlation vs AutoCorrelation

- Correlation is the tendency for the distribution of two variables to follow one another, so that a unit rise or fall in one variable accompanies a proportionate rise or fall (not necessarily in the same direction) in the other.
- Autocorrelation is a form of correlation. It applies specifically to time series — variables that are measures of a quantity at differing points in time.
- Autocorrelation is the tendency for a time series to be correlated with the same measure one or more periods ahead or behind, for instance, the correlation between GDP in a given month with GDP the month before.



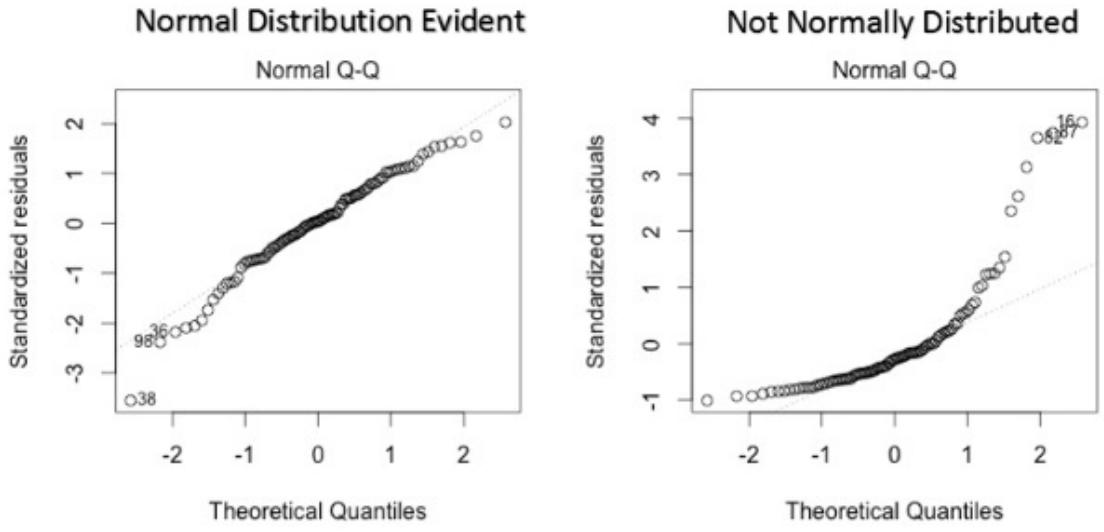
Correlation vs AutoCorrelation

- Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.
- A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker.
- **Auto-correlation** refers to the case when your errors are correlated with each other. In layman terms, if the current observation of your dependent variable is correlated with your past observations, you end up in the trap of auto-correlation.



Assumptions

- Linear Relationship between the features and target
- Normal distribution
- Little or No autocorrelation in the residuals

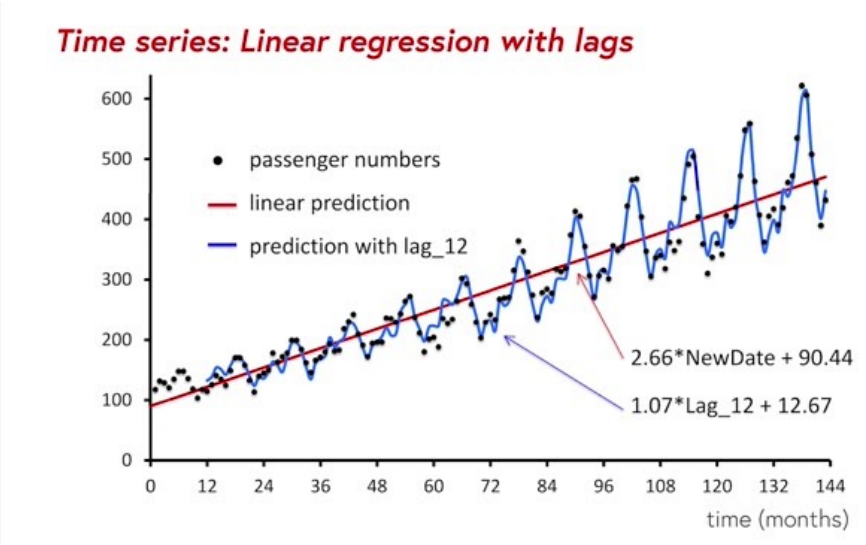


Linear Regression Time Series

- Linear regression is widely used in practice and adapts naturally to even complex forecasting tasks.
- The **linear regression** algorithm learns how to make a weighted sum from its input features. For two features, we would have:

  target = weight_1 * feature_1 + weight_2 * feature_2 + bias

- During training, the regression algorithm learns values for the parameters weight_1, weight_2, and bias that best fit the target. (This algorithm is often called *ordinary least squares* since it chooses values that minimize the squared error between the target and the predictions.)
- The weights are also called *regression coefficients* and the bias is also called the *intercept* because it tells you where the graph of this function crosses the y-axis.
- Time-step features let you model **time dependence**. A series is time dependent if its values can be predicted from the time they occurred.



Linear Regression Time Series

- To make a **lag feature** we shift the observations of the target series so that they appear to have occurred later in time.
- One of the assumptions of linear regression is that the residues are not correlated
- With time series data, **this is often not the case**
- If there are autocorrelated residues, then linear regression will not be able to "capture all the trends" in the data
- When linear regression is used but observations are correlated (as in time series data) you will have a biased estimate of the variance

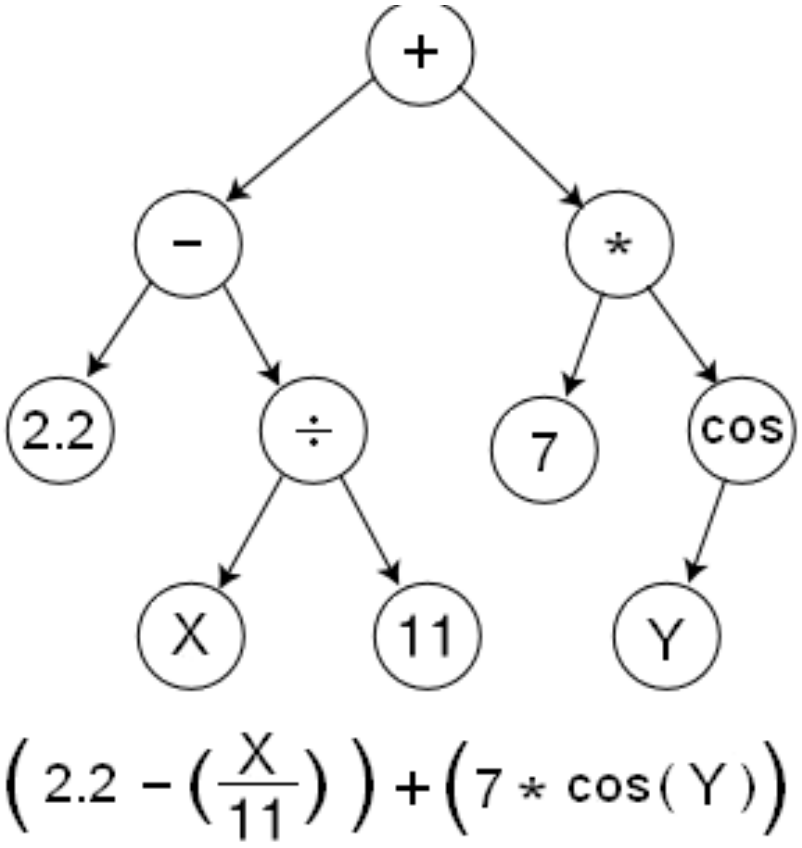
Regression Vs AutoRegression

- Regression models forecast a variable using a linear combination of predictors, whereas autoregressive models use a combination of past values of the variable.
- Autoregressive models base their predictions only on past information, they implicitly assume that the fundamental forces that influenced the past prices will not change over time.
- An autoregression model is a **linear regression model** that uses lagged variables as input variables

Symbolic Regression

Symbolic Regression (SR) is a type of Regression Analysis that searches the space of mathematical expressions to find the model that best fits a given dataset, both in terms of accuracy and simplicity.

The task of identifying a mathematical expression that best fits a provided dataset of input and output values. Due to the richness of the space of mathematical expressions, symbolic regression is generally a challenging problem.



- This approach has the disadvantage of having a much larger space to search, because not only the search space in symbolic regression is infinite, but there are an infinite number of models which will perfectly fit a finite data set (provided that the model complexity isn't artificially limited). This means that it will possibly take a symbolic regression algorithm longer to find an appropriate model and parametrization, than traditional regression techniques.
- This can be attenuated by limiting the set of building blocks provided to the algorithm, based on existing knowledge of the system that produced the data; but in the end, using symbolic regression is a decision that has to be balanced with how much is known about the underlying system.
- Nevertheless, this characteristic of symbolic regression also has advantages: because the evolutionary algorithm requires diversity in order to effectively explore the search space, the end result is likely to be a selection of high-scoring models (and their corresponding set of parameters). Examining this collection could provide better insight into the underlying process and allows the user to identify an approximation that better fits their needs in terms of accuracy and simplicity.