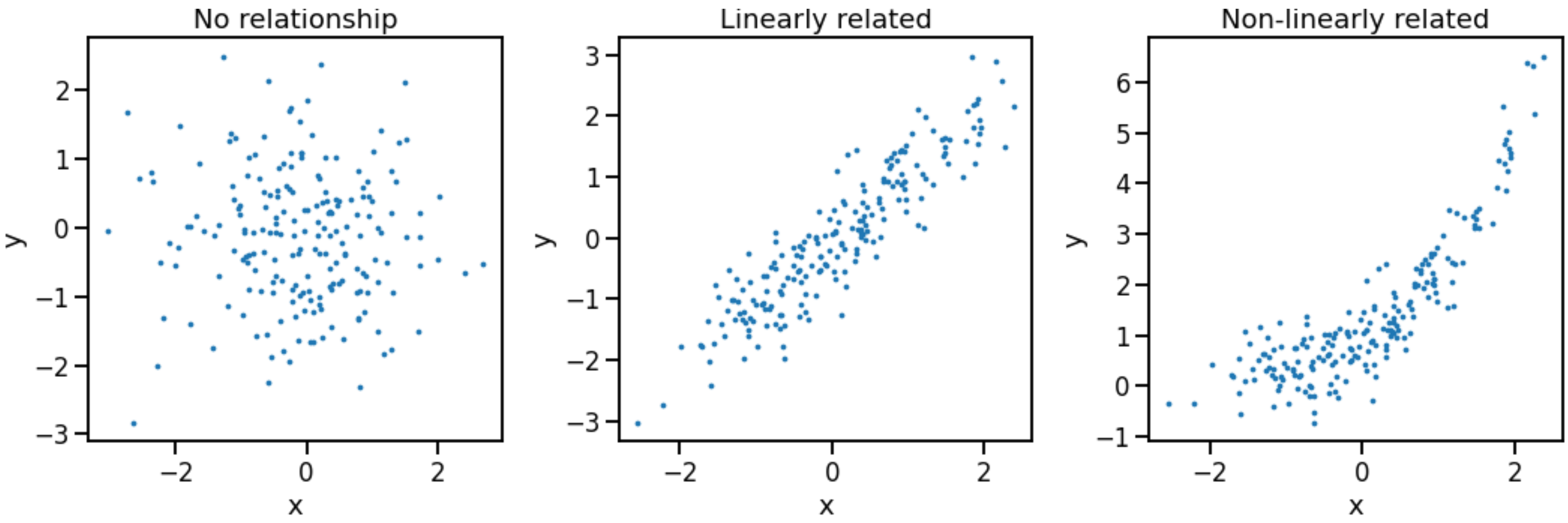
Lag plots

Lag features

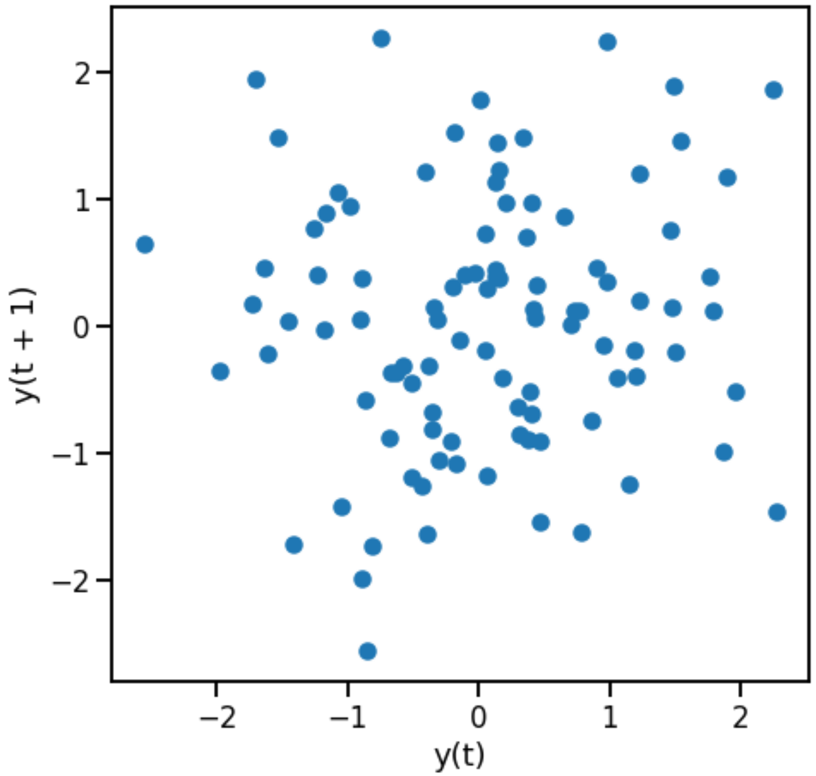
Contents

Scatterplots can help identify if two variables are related



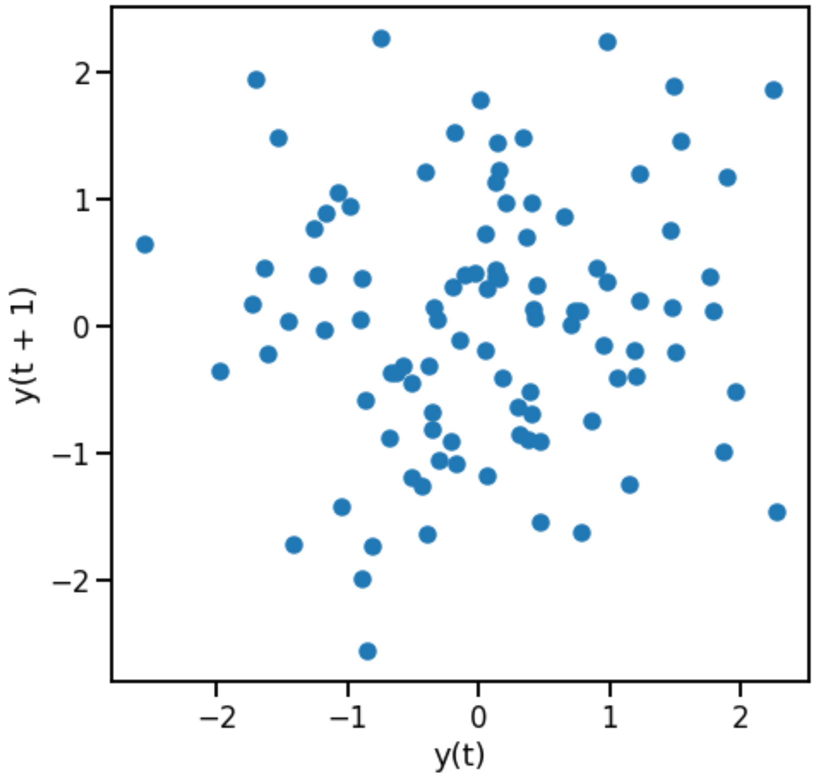
Lag plot is a scatter plot of a time series against a lagged version of itself

Date	y	y Lag 1
2020-02-12	23	NaN
2020-02-13	30	23
2020-02-14	35	30
2020-02-15	30	35



What can we learn from a lag plot?

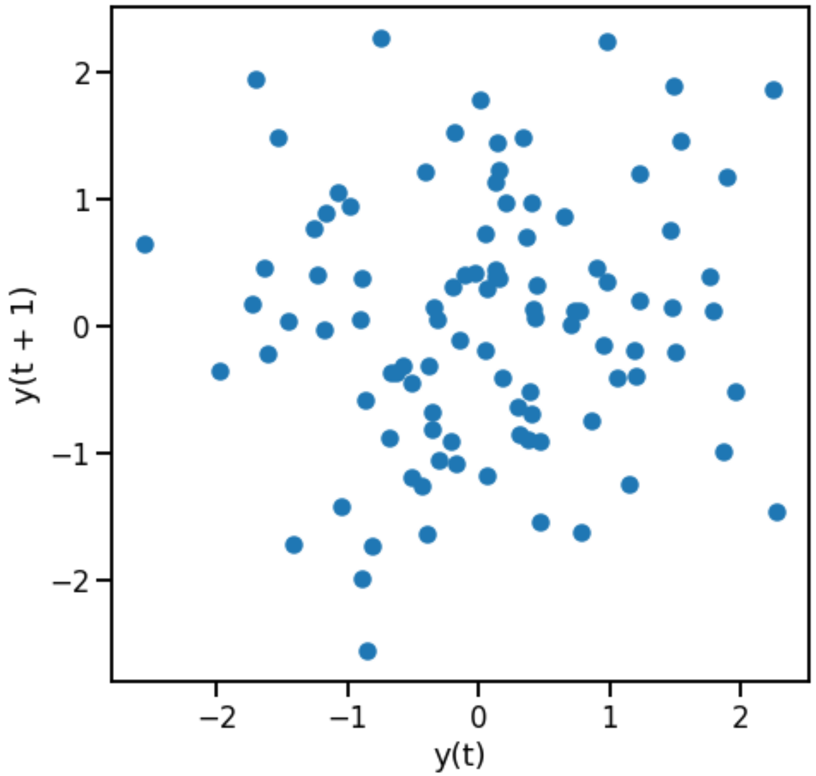
- A lag plot is a visual tool which can help show if y_t shows a non-random relationship with y_{t-k}.
- If it does then a lag of k could be a useful feature for forecasting.
- Let's look at lag plots for different types of time series to understand what signatures they leave on the lag plot.



What can we learn from a lag plot?

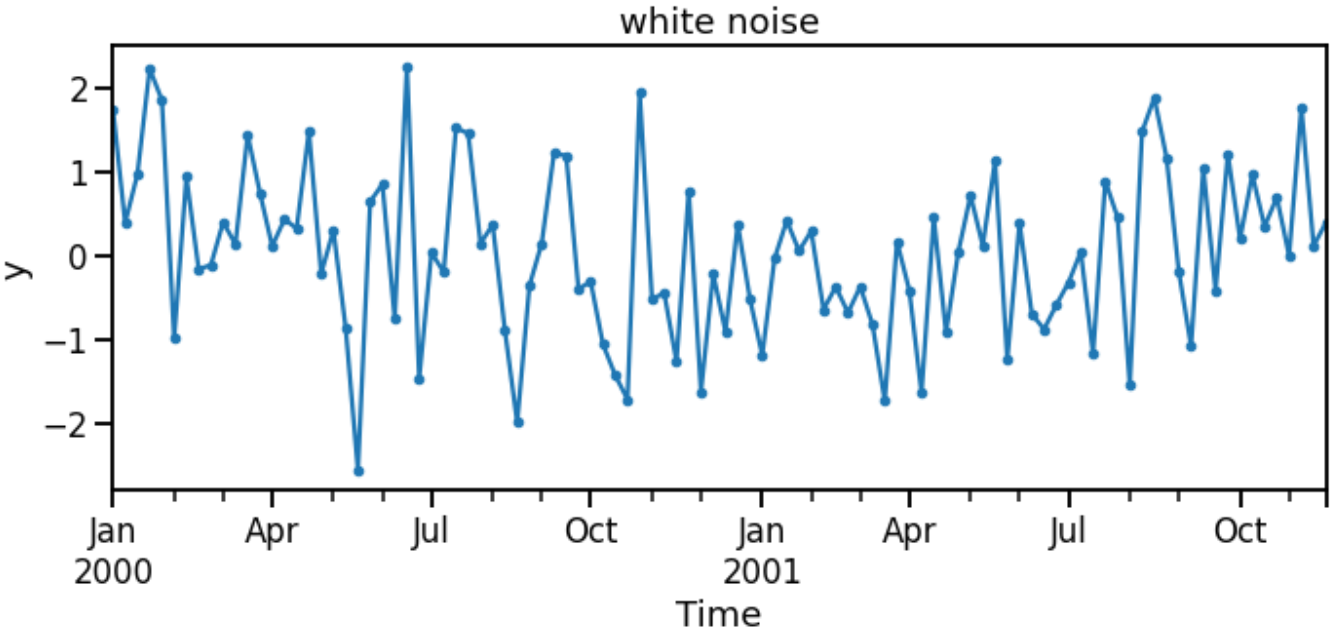
- We shall look at time series with properties:
  - white noise
  - AR(1) process
  - completely periodic (just seasonality)
  - trend
  - trend and seasonality

- We shall illustrate how the lag plot can help identify useful lags as we go along.



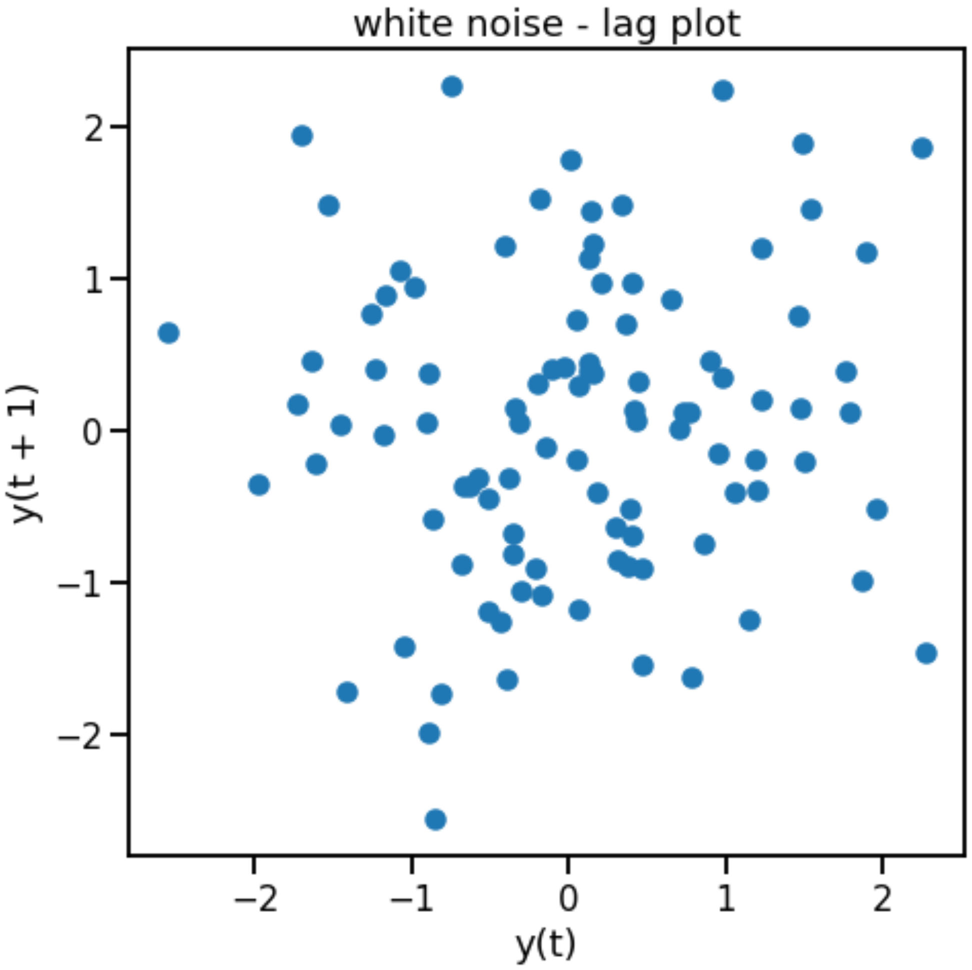
White noise

- $y_t = \epsilon_t$ where $\epsilon_t \sim N(0,1)$
- A random timeseries with no correlation between points.
- There is no predictive information in the historic data.



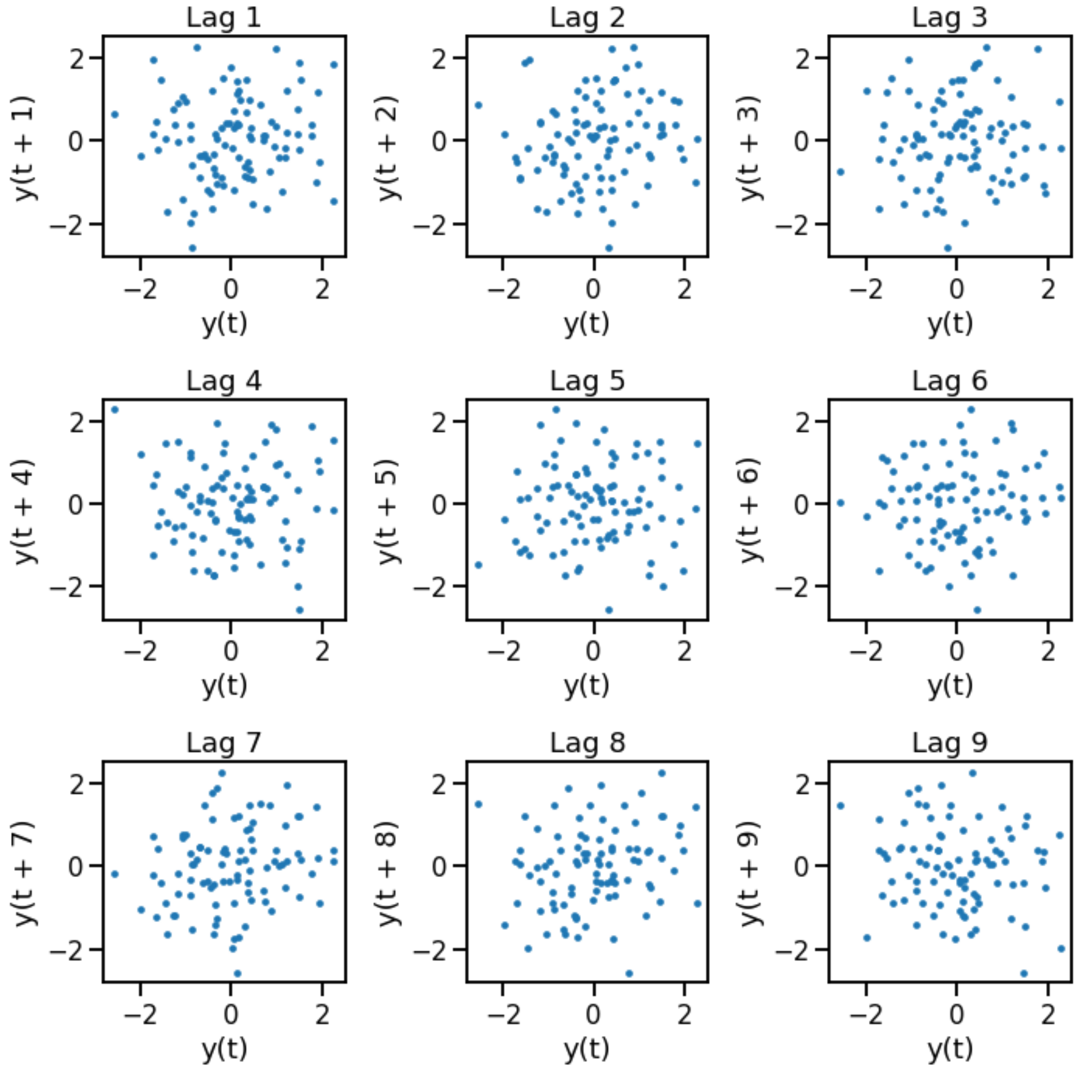
Lag plot: white noise

- No strong relationship in the lag plot, as expected from white noise.
- Let's look at additional lags.



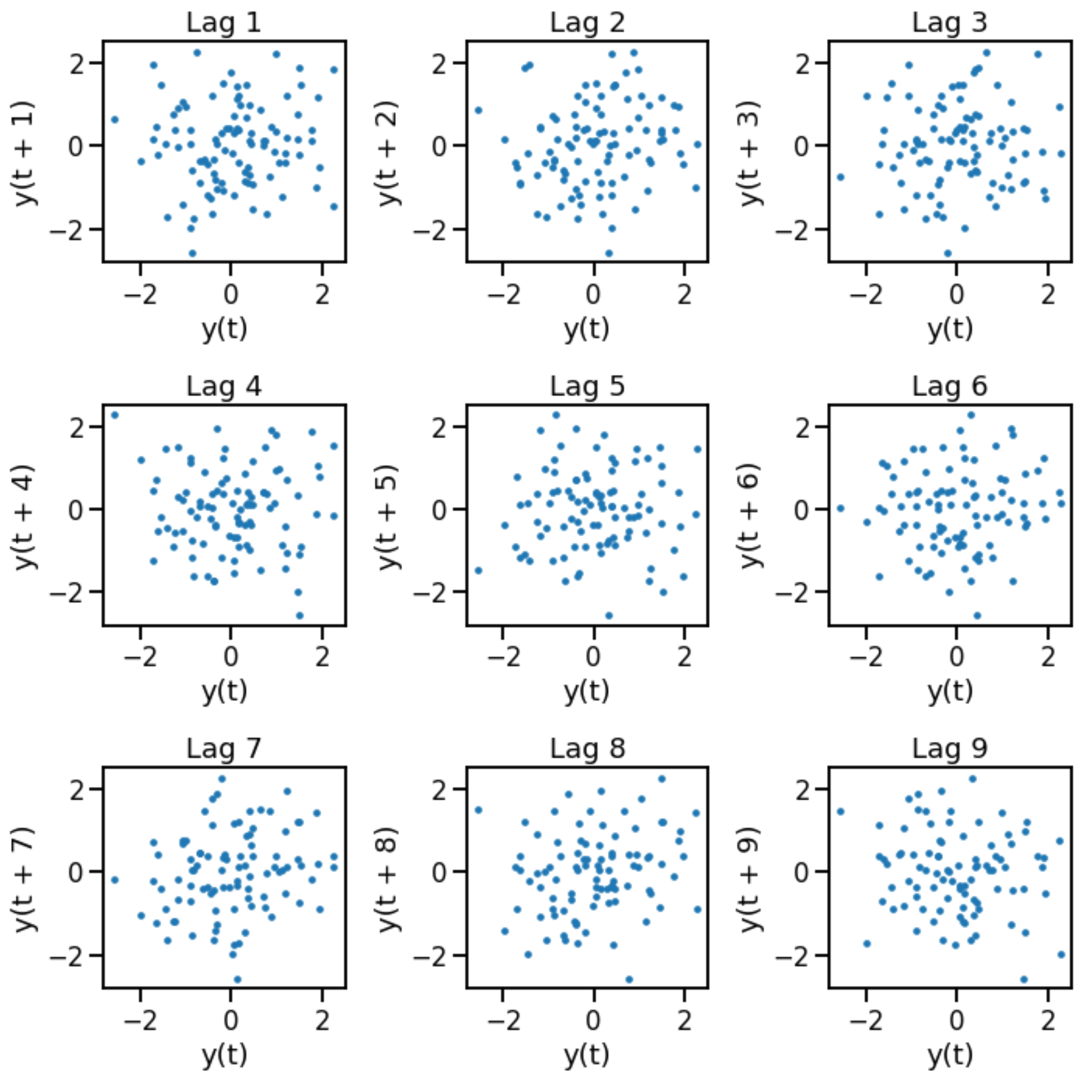
Lag plot: white noise

- No strong relationship in the lag plot, as expected from white noise.
- Let's look at additional lags.



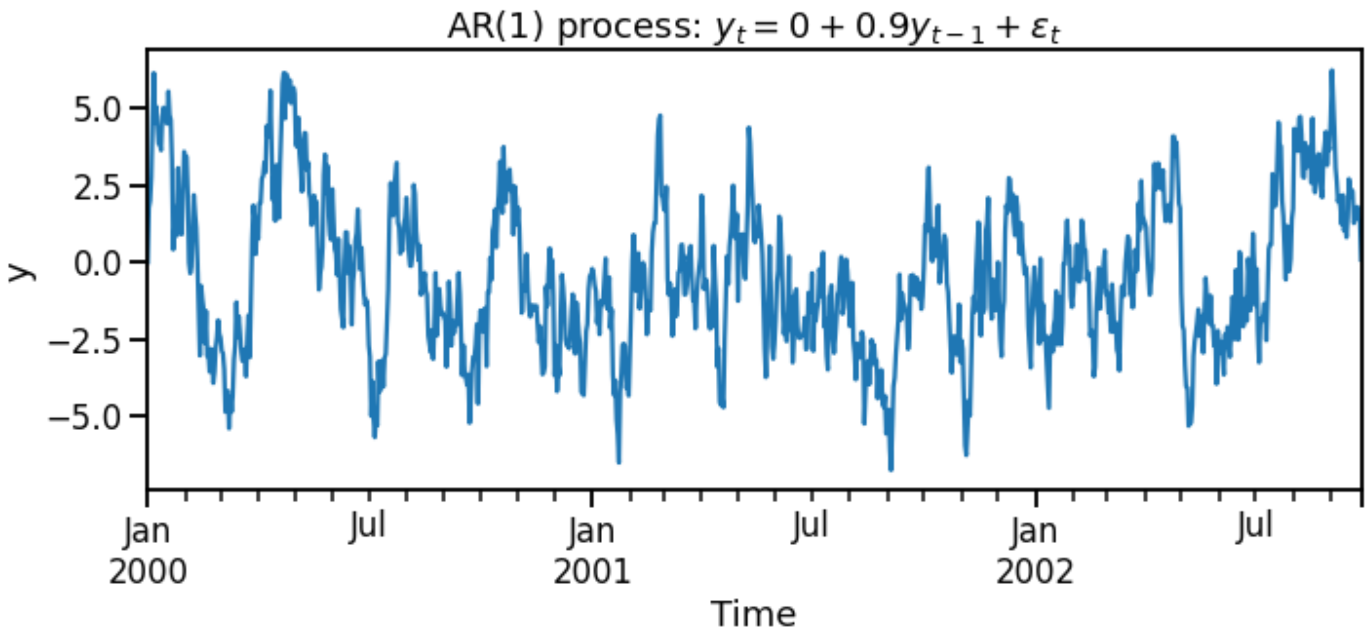
Lag plot: white noise

- No strong relationship in the lag plot, as expected from white noise.
- Let's look at additional lags.
- This shows us what to look for when determining when a lag may **not** be useful.



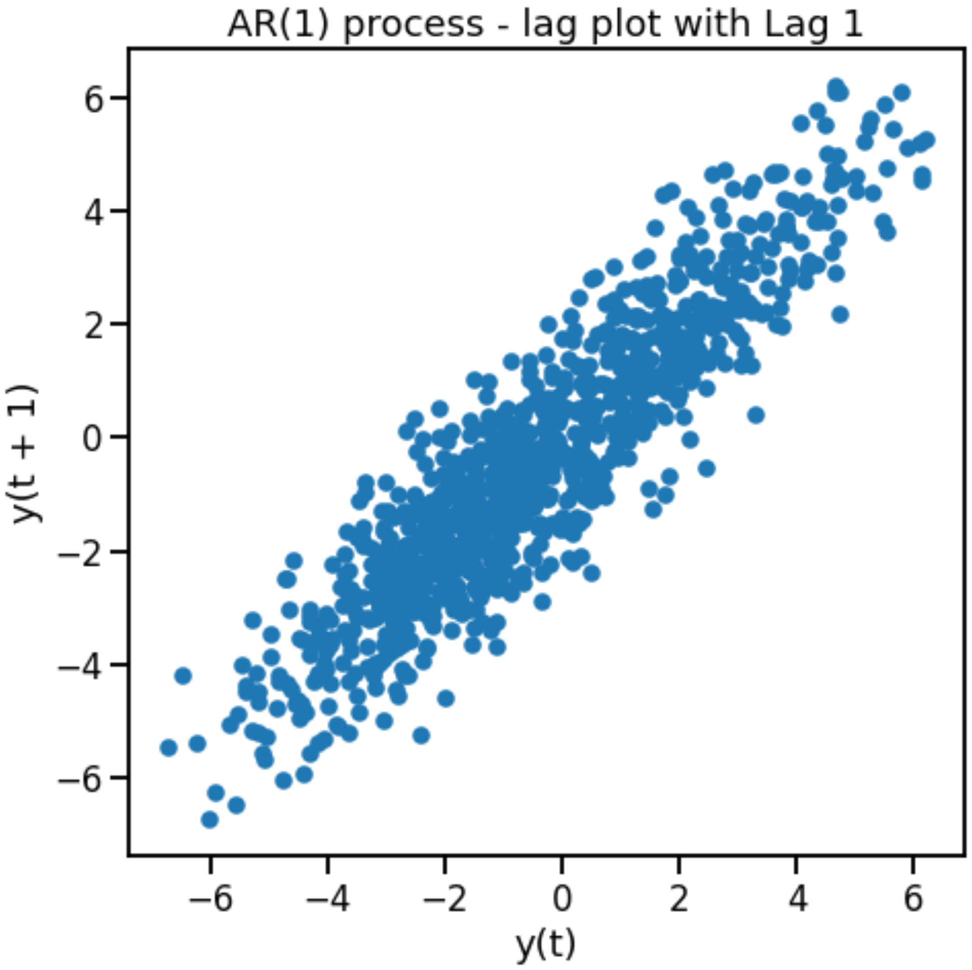
AR(1) process

- $y_t = c + \phi_1 y_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0,1)$
- The timeseries is determined by the previous lag (i.e., lag 1).
- So we expect this time series to be correlated to lagged values.



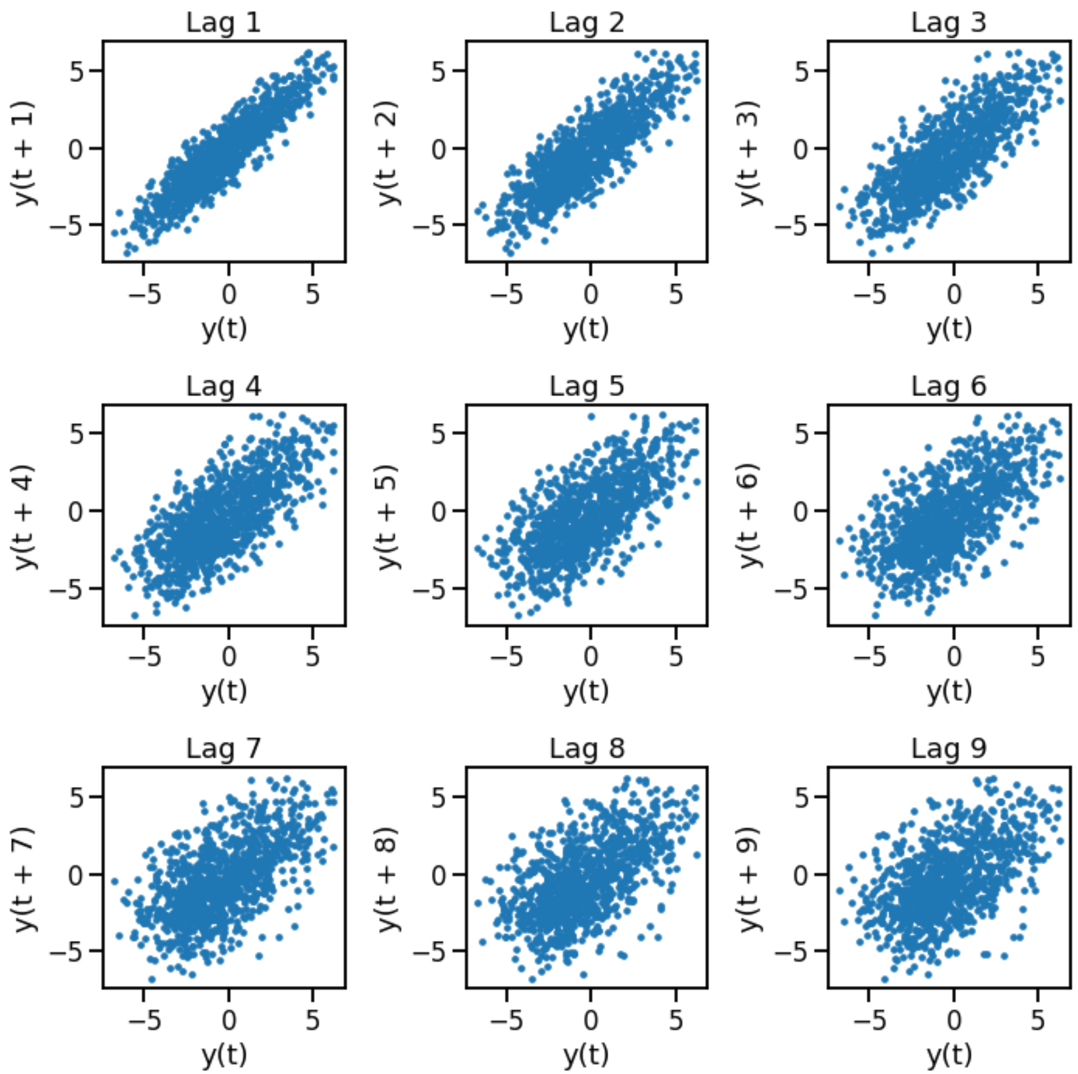
Lag plot: AR(1) process

- We see a strong linear correlation between y_t and y_{t-1}.
- Let's look at additional lags.



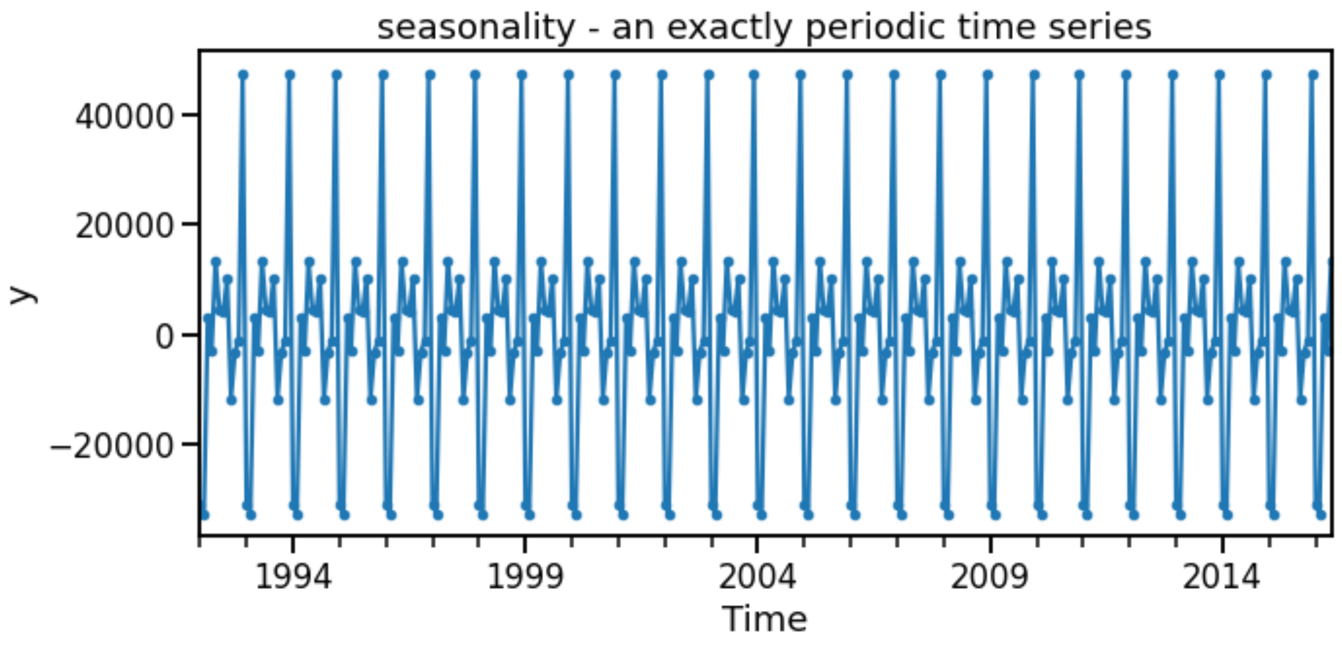
Lag plot: AR(1) process

- We see the correlation between y_t and its lagged values decay as we look at larger lags.
- This shows us that a time series which is determined only by a small number of previous lags (1 in this case) can generate correlations at multiple lags.
- When we discuss the partial autocorrelation function (PACF) we will see how we can identify that lag 1 is the most important.



Seasonality

- $y_t = seasonal_t$
- A time series which repeats exactly every 12 months.
- Any multiple lag of 12 should be the most predictive of future values as the time series is exactly periodic every 12 months.

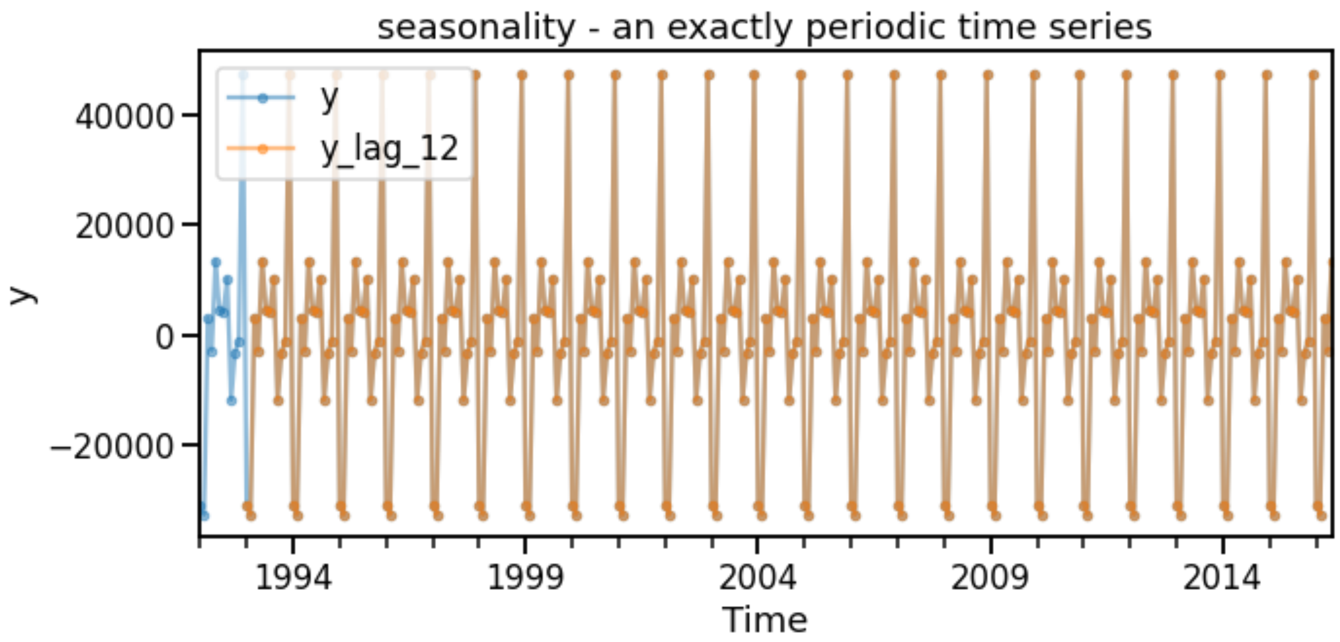


Lag plot: Seasonality

- $y_t = seasonal_t$

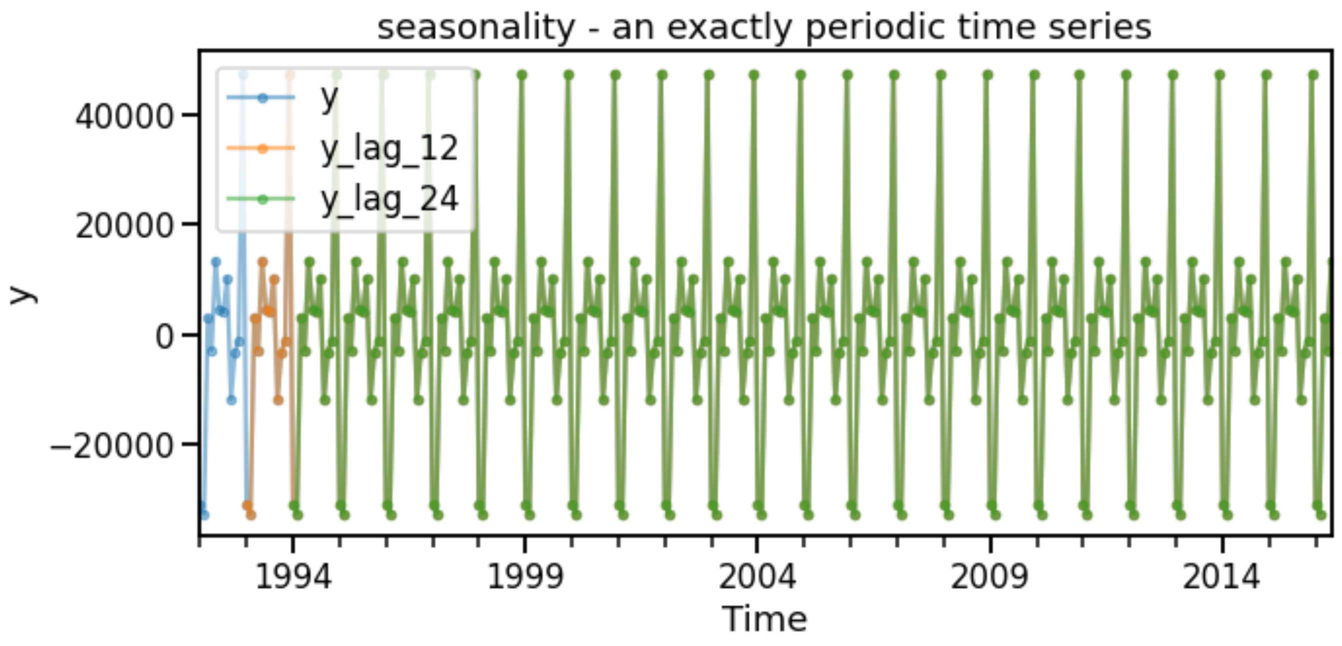
- A time series which repeats exactly every 12 months.

- Any multiple lag of 12 should be the most predictive of future values as the time series is exactly periodic every 12 months.



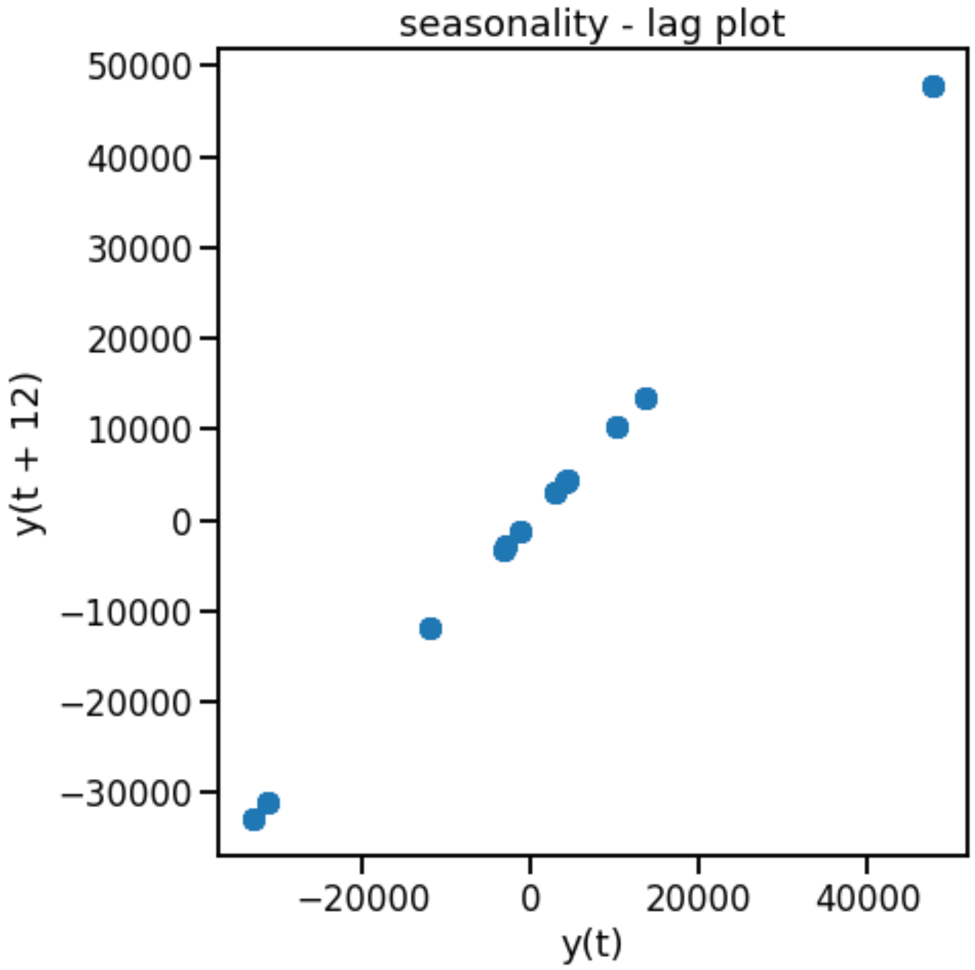
Lag plot: Seasonality

- $y_t = seasonal_t$
- A time series which repeats exactly every 12 months.
- Any multiple lag of 12 should be the most predictive of future values as the time series is exactly periodic every 12 months.



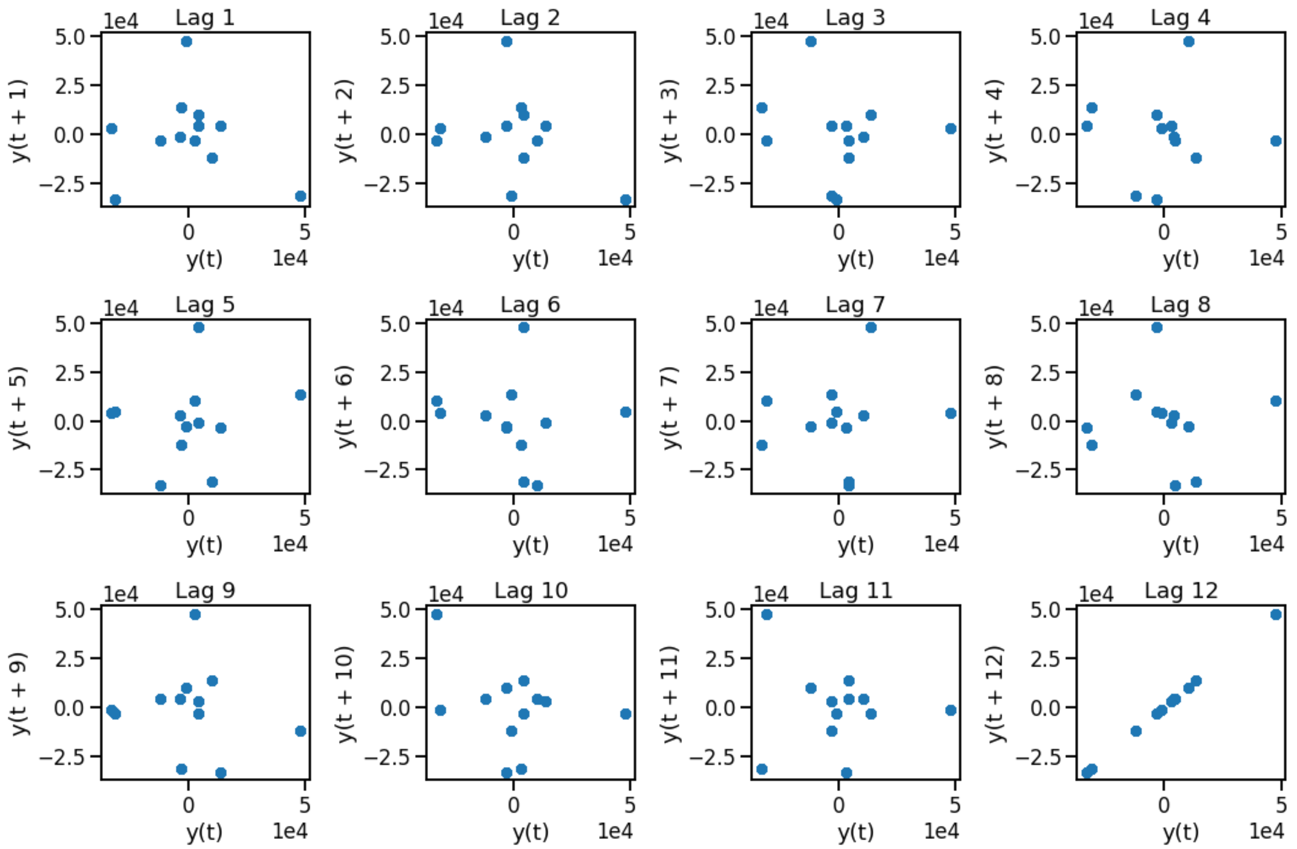
Lag plot: Seasonality

- We see strong seasonality at lag 12 indicated by the strong linear correlation.
- This shows that a lag of 12 could be a helpful feature.
- Let’s look at additional lags.



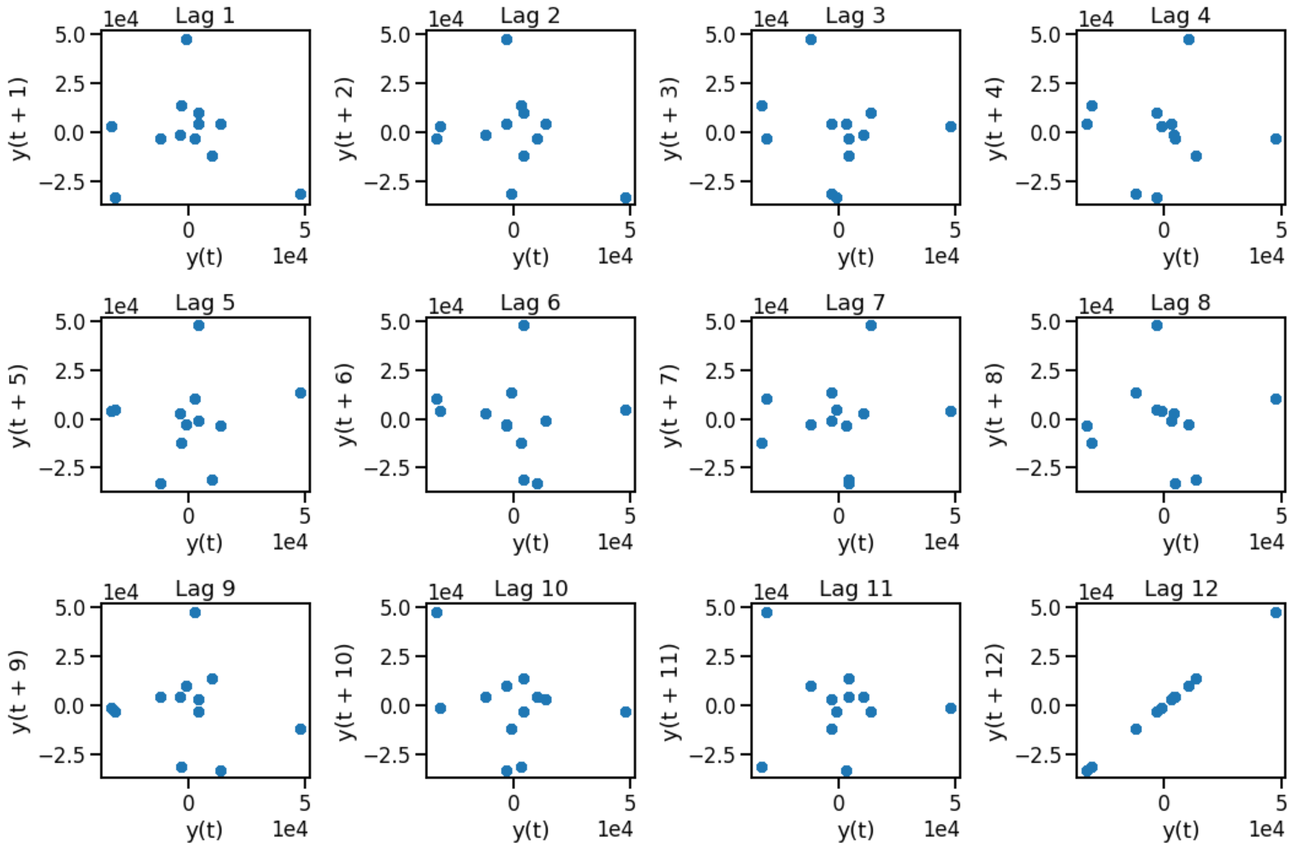
Lag plot: Seasonality

- y_t is exactly periodic with a seasonal period of 12.
- Therefore, there are only 12 unique values that y_t can take.
- This causes every lag plot to only have 12 data points in different configurations which repeat every 12 lags.
- This will be important when we discuss the autocorrelation function.



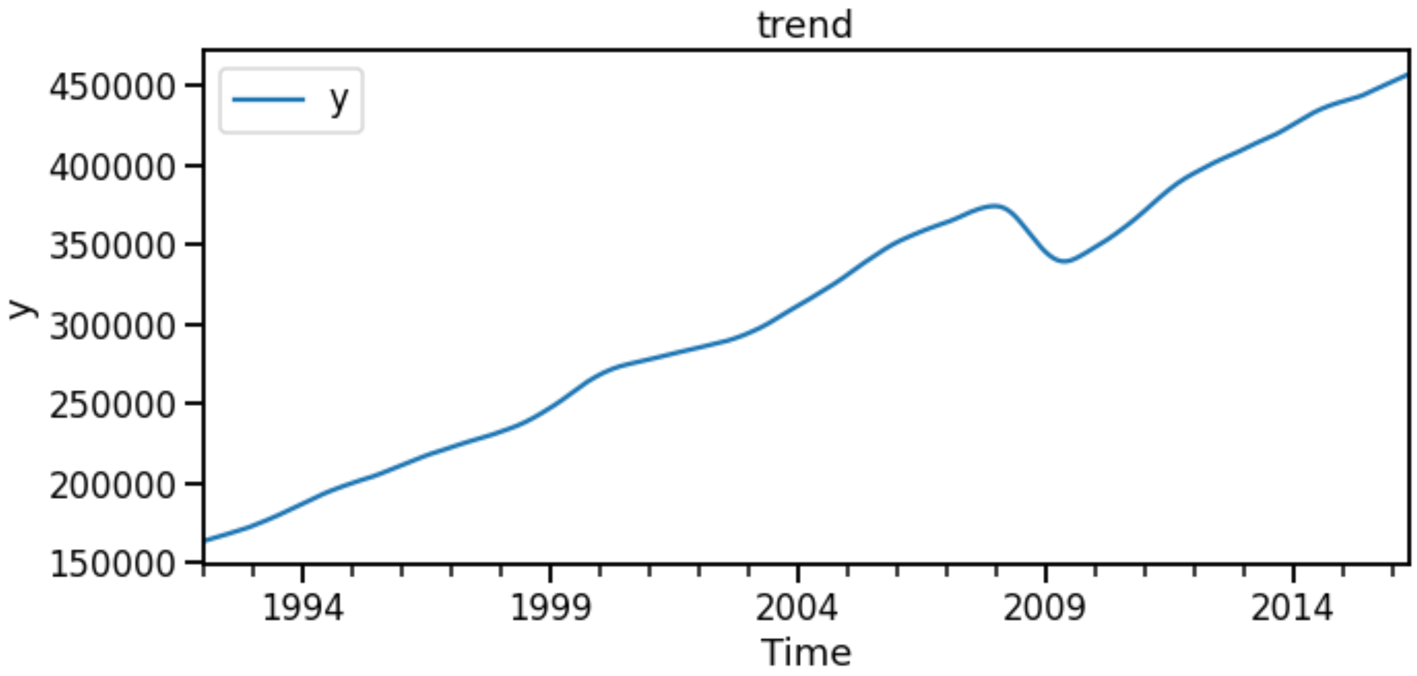
Lag plot: Seasonality

- This shows us that any seasonal patterns will appear as a strong linear correlation on the lag plot. This occurs at multiples of the seasonal period (E.g., every 12 lags for yearly seasonality).



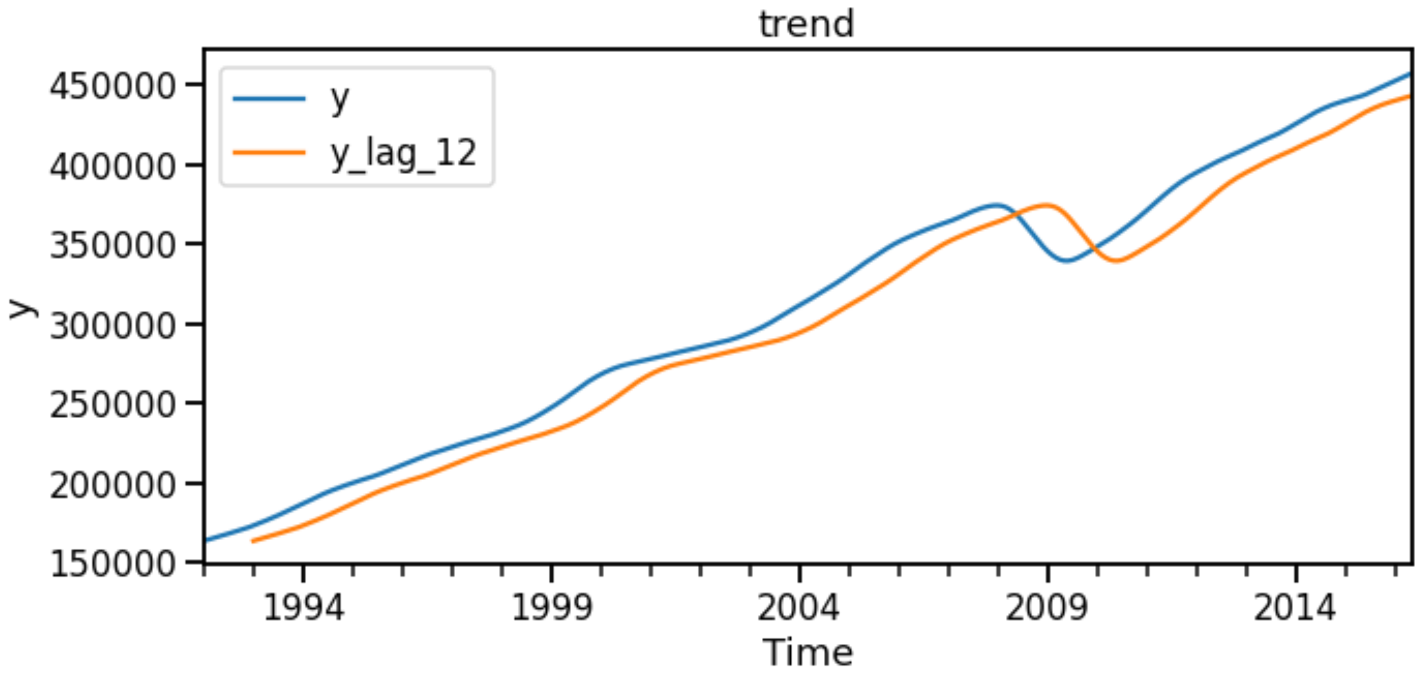
Trend

- $y_t = trend_t$
- There is a lot of predictive information in historic data.



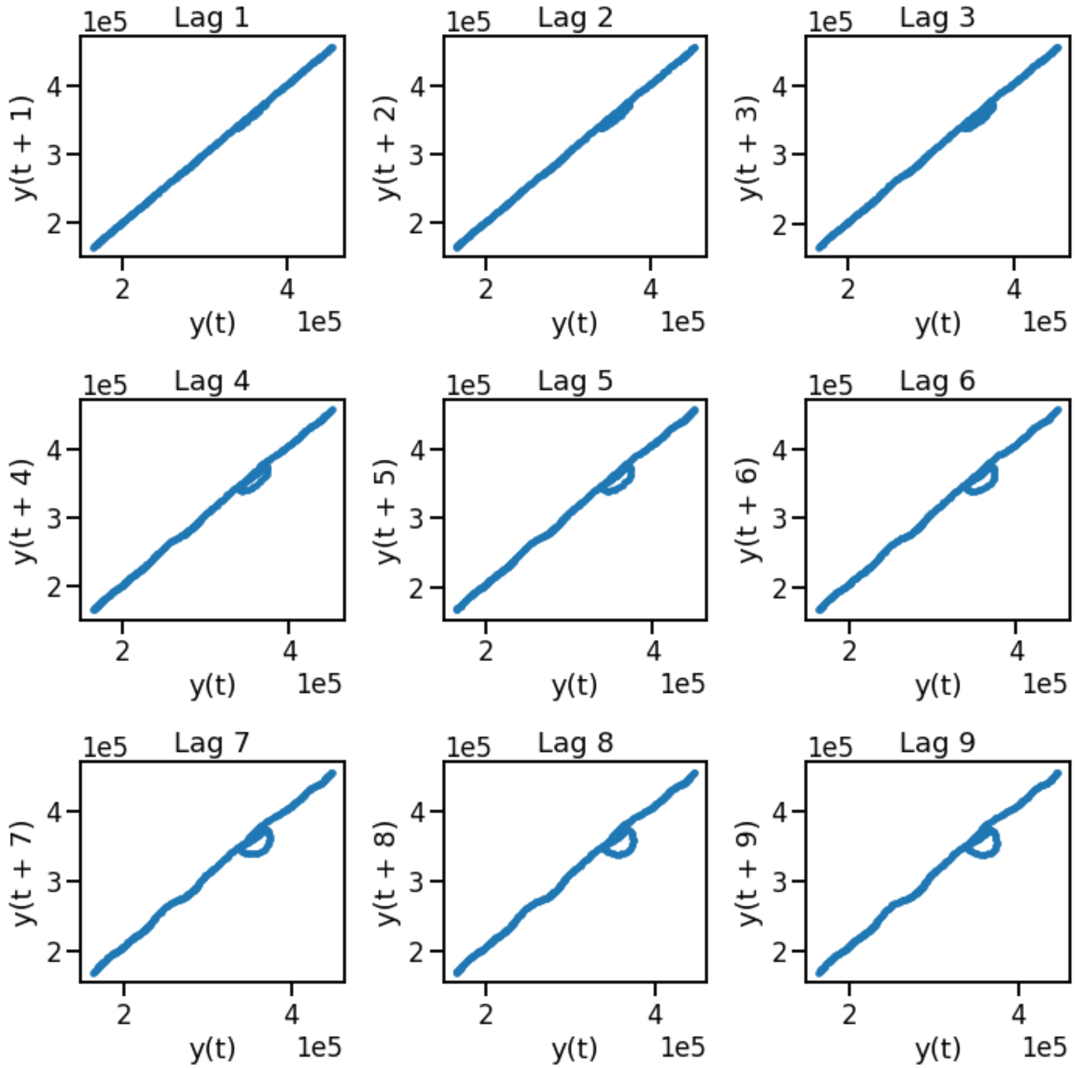
Trend

- $y_t = trend_t$
- There is a lot of predictive information in historic data.
- When y_t is small so is y_{t-k}. When y_t is large so is y_{t-k}. Expect a large positive relationship!



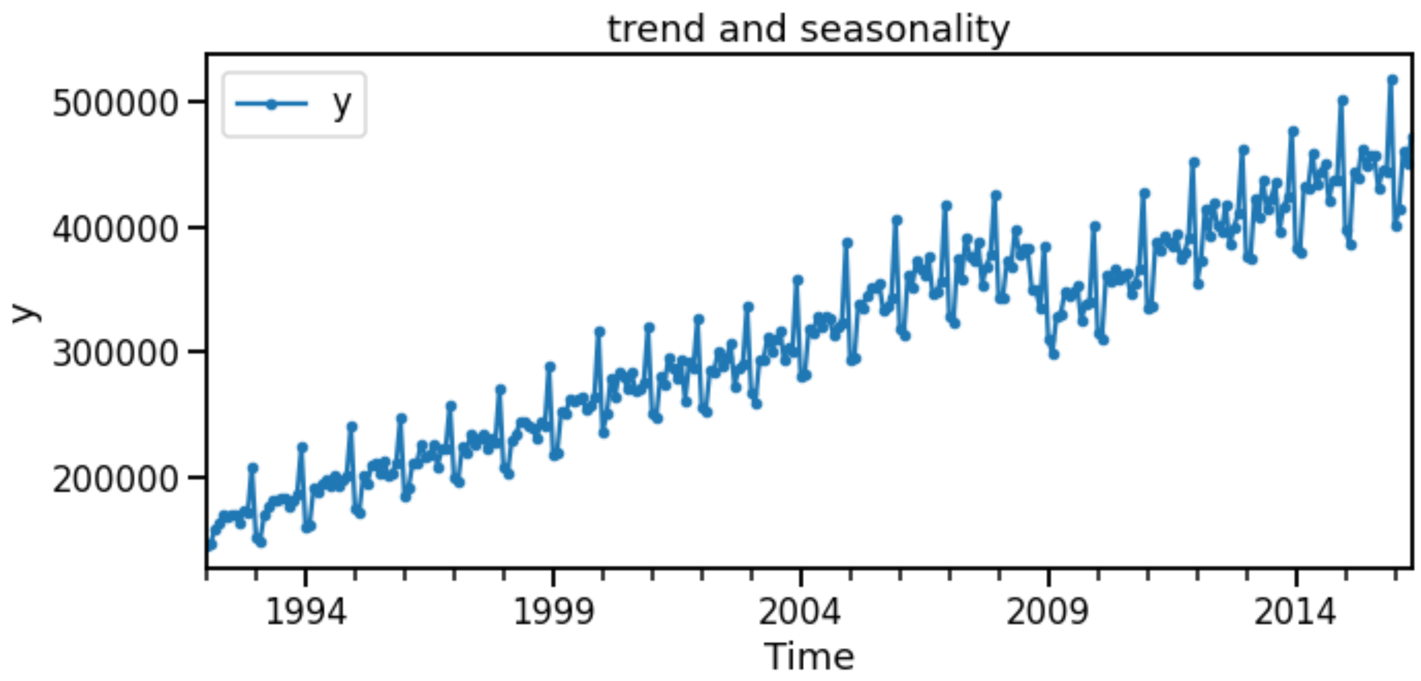
Lag plot: trend

- Strong linear relationship in the lag plot, as expected.
- This is seen across multiple lags because of the overall shape of the original time series means that when y_t is relatively large so is y_{t-k}.
- The trend causes correlations at many lags. This can make it difficult to identify patterns (e.g., seasonality) which appear as strong correlations only at specific lags.
- This shows that the lag plot does not provide much information about whether a specific lag will be helpful.



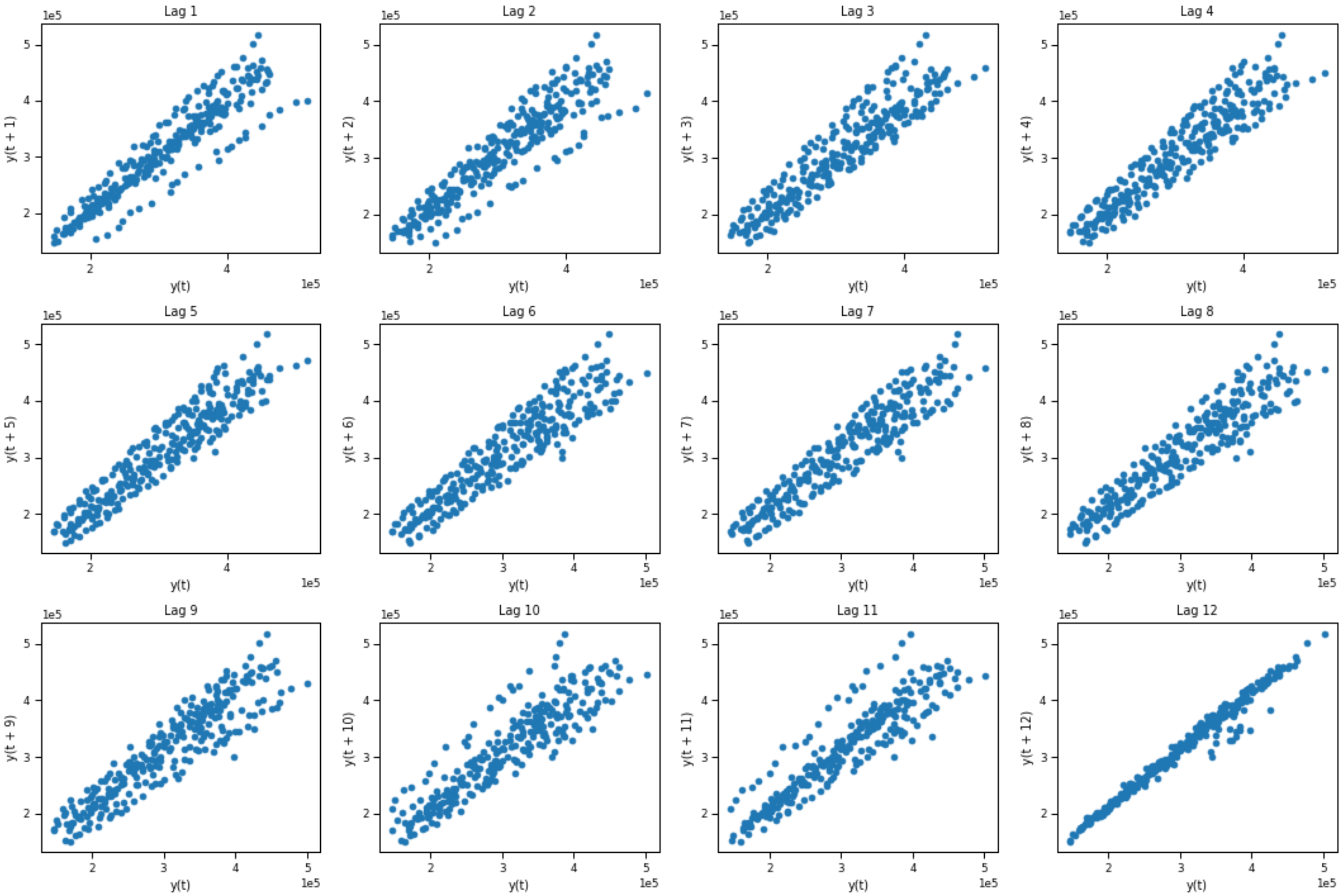
Lag plot: trend and seasonality

- $y_t = trend_t + seasonal_t + residual_t$
- There is a lot of predictive information in historic data.
- Will get a combination of effects from the trend, seasonality, and noise in the lag plots.



Lag plot: trend and seasonality

The strong trend results in linear relationships across many lags.



The seasonal component is seen in the seasonal lag as a much stronger linear relationship.

Lag plot implementation in Pandas

pandas.plotting.lag_plot

`pandas.plotting.lag_plot(series, lag=1, ax=None, **kwds)` ¶ [source]

Lag plot for time series.

Parameters:

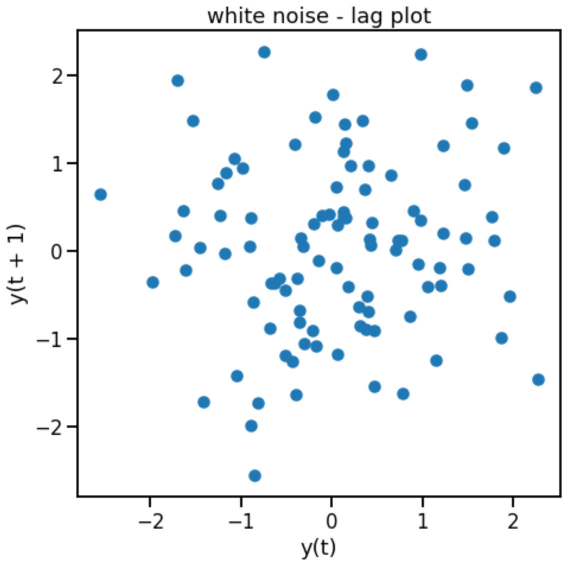
- **series** : *Time series*
- **lag** : *lag of the scatter plot, default 1*
- **ax** : *Matplotlib axis object, optional*
- ****kwds**

  Matplotlib scatter method keyword arguments.

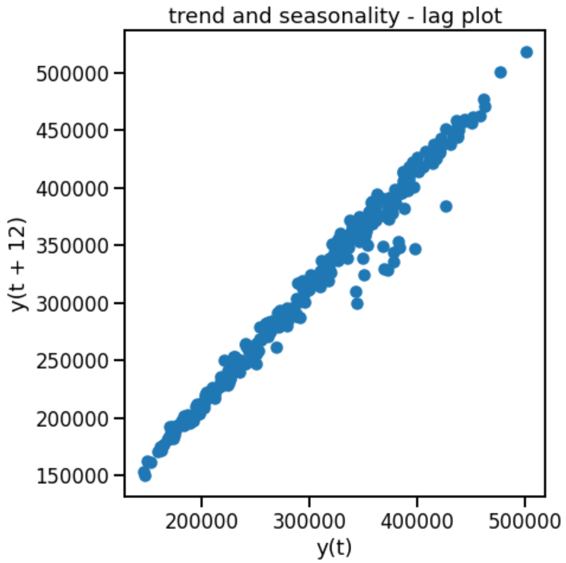
Returns: class:*matplotlib.axis.Axes*

Lag plot limitations and what's next

- Lag plots are a visual tool which can help identify useful lags but are not scalable.
- If we quantify when y_{t-k} is highly correlated with y_t then it would be easier to identify useful lags.
- Autocorrelation is a method to quantify the correlation of a time series with itself and can be used to understand properties of a time series including useful lags (coming up next!).



Low autocorrelation



High autocorrelation

Summary

Scatterplots can help identify if two variables are related.

Lag plot is a scatter plot of a time series against a lagged version of itself.

Lag plots can identify lags which are strongly related to the original time series.

Trend, seasonality, autocorrelation, and noise leave their own signatures on a lag plot.