Test sets will be the most recent end of the data.

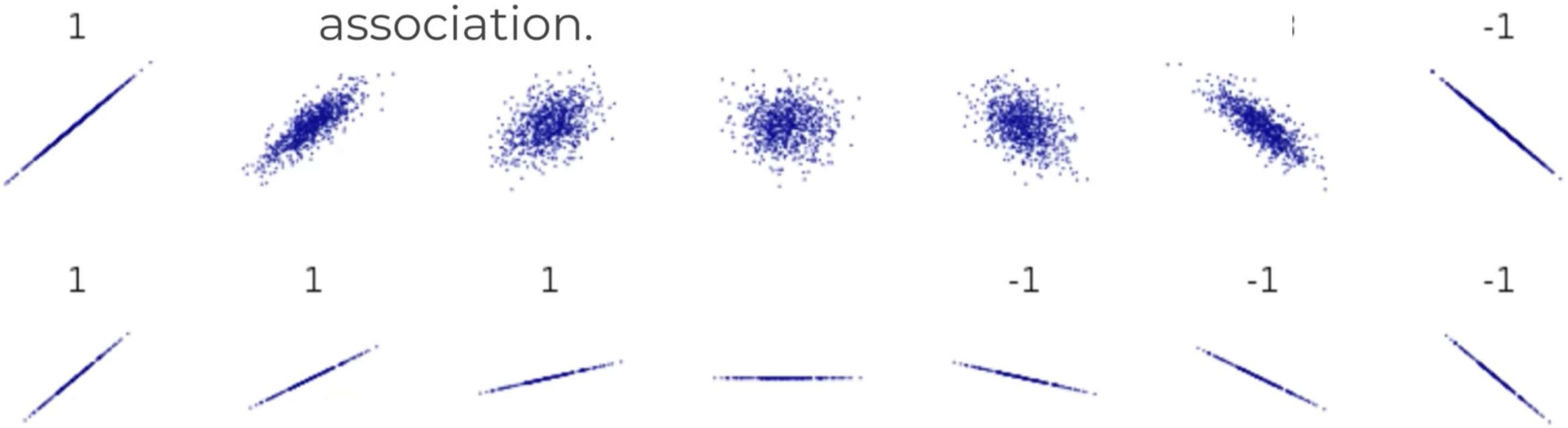
The test set should ideally be at least as large as the maximum forecast horizon required.
Keep in mind, the longer the forecast horizon, the more likely your prediction becomes less accurate.

- Forecasting Procedure
  - Choose a Model
  - Split data into train and test sets
  - Fit model on training set
  - Evaluate model on test set
  - Re-fit model on entire data set
  - Forecast for future data

- You may have heard of some evaluation metrics like accuracy or recall.
- These sort of metrics aren't useful for time series forecasting problems, we need metrics designed for **continuous** values!

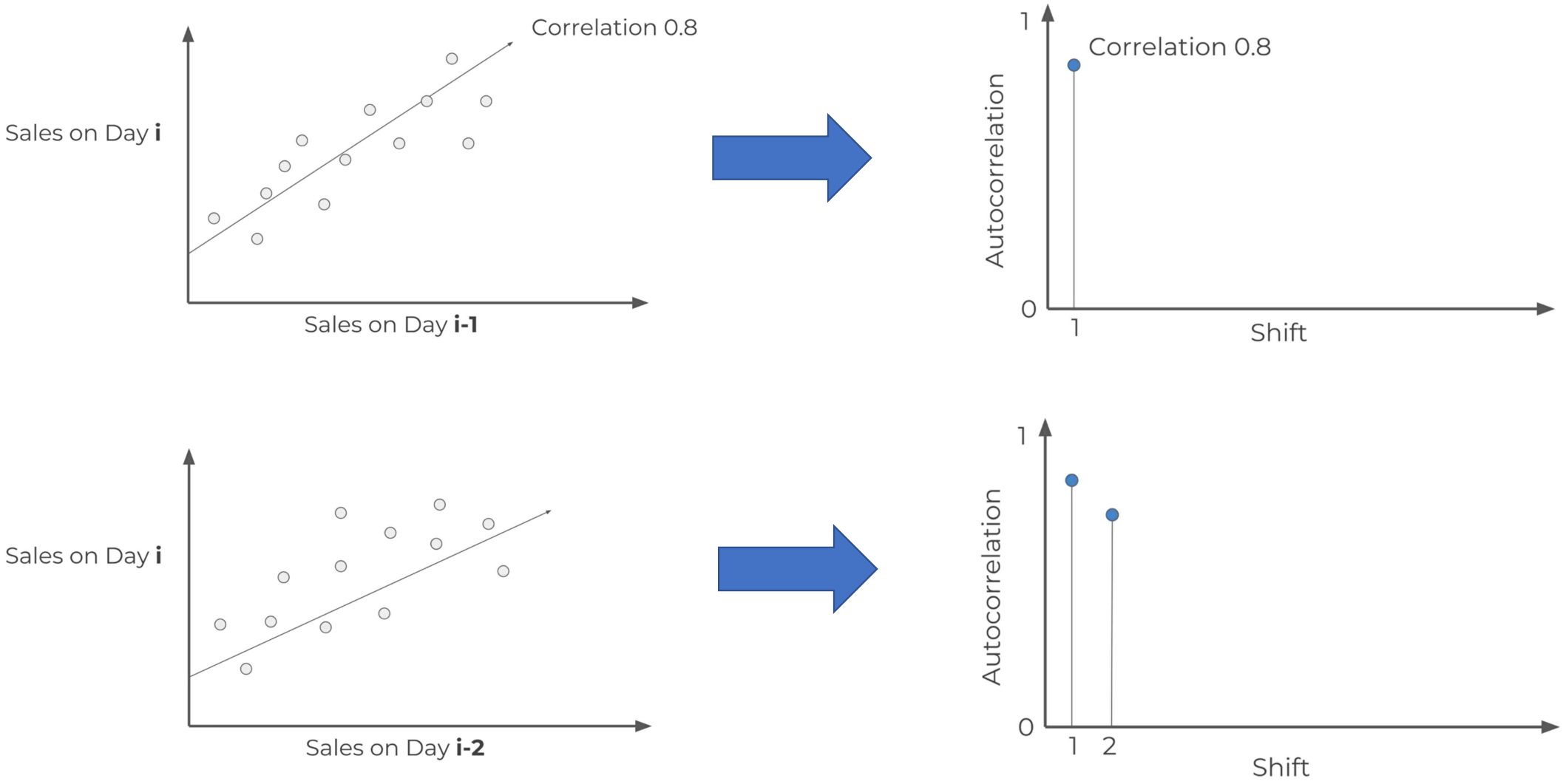
- You will need to use your own judgement and compare the RMSE to the average values in your data set’s test set.
- Then make a decision for the acceptability of the error.

- The closer the correlation is to +1, the stronger the positive linear relationship
- The closer the correlation is to -1, the stronger the negative linear relationship.
- And the closer the correlation is to zero, the weaker the linear relationship, or association.



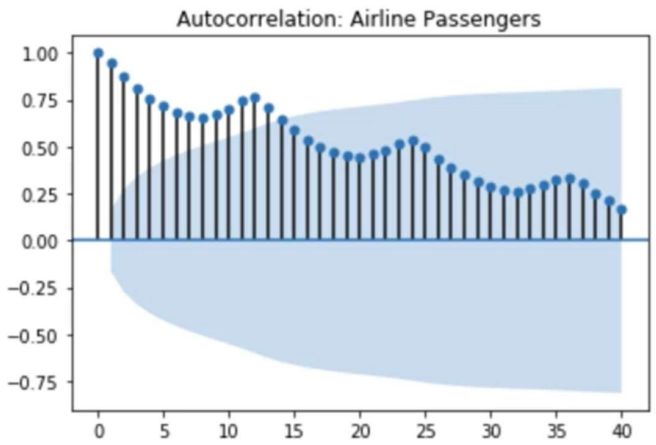
- An autocorrelation plot (also known as a Correlogram) shows the correlation of the series with itself, lagged by x time units.
- So the y axis is the correlation and the x axis is the number of time units of lag.

Imagine we had some sales data.
We can compare the standard sales data against the sales data shifted by 1 time step.
This answers the question, “How correlated are today’s sales to yesterday’s sales?”

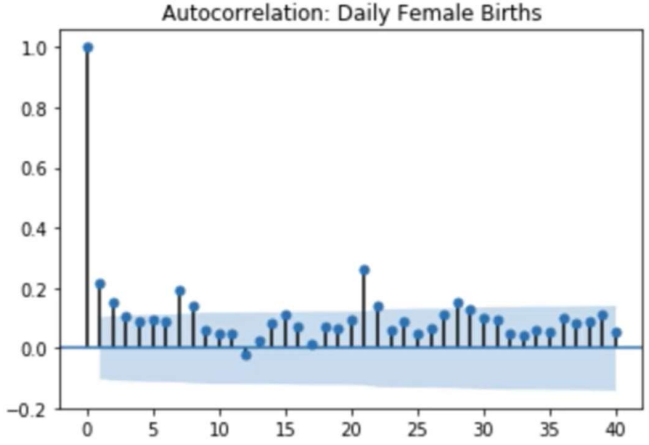

Correlation 0.8
Sales on Day i
Sales on Day i-1
1
Autocorrelation
Correlation 0.8
0
1
Shift
Sales on Day i
Sales on Day i-2
1
Autocorrelation
0
1
2
Shift

An autocorrelation plot shows the correlation of the series with itself, lagged by x time units.
You go on and do this for all possible time lags x and this defines the plot.

Gradual Decline

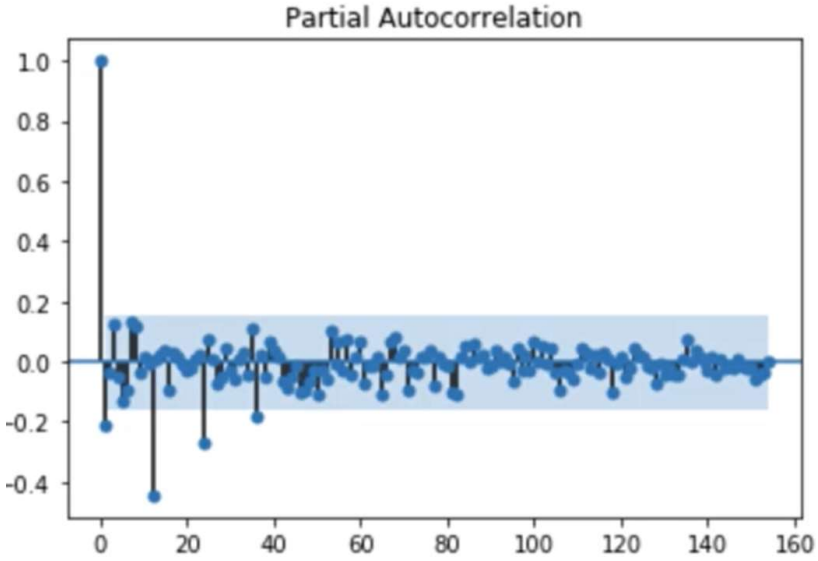


Sharp Drop-off



The actual interpretation and how it relates to ARIMA models can get a bit complicated, but there are some basic common methods we can use for the ARIMA model.

partial autocorrelation



The PACF only describes the direct relationship between an observation and its lag.

Many models are based off the ARIMA model, which stands for AutoRegressive Integrated Moving Average

- It is important to understand that ARIMA is not capable of perfectly predicting any time series data.
- Beginner students often want to directly apply ARIMA to time series data that is not directly a function of time, such as stock data.

Stock price data for example has so many outside factors that much of the information informing the price of the stock won't be available with just the time stamped price information.

ARIMA performs very well when working with a time series where the data is directly related to the time stamp, such as the airline passenger data set.
In that data we saw clear growth and seasonality based on time.

But it is important to keep in mind that an ARIMA model on that data wouldn't be able to understand any outside factors, such as new developments in jet engines, if those effects weren't already present in the current data.

AutoRegressive Integrated Moving Average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model.

Both of those models (ARIMA and ARMA) are fitted to time series data either to better understand the data or to predict future points in the series (forecasting).

- ARIMA (Autoregressive Integrated Moving Averages)
  - Non-seasonal ARIMA
  - Seasonal ARIMA (SARIMA)
- Also understanding SARIMA with exogenous variables, such as SARIMAX.

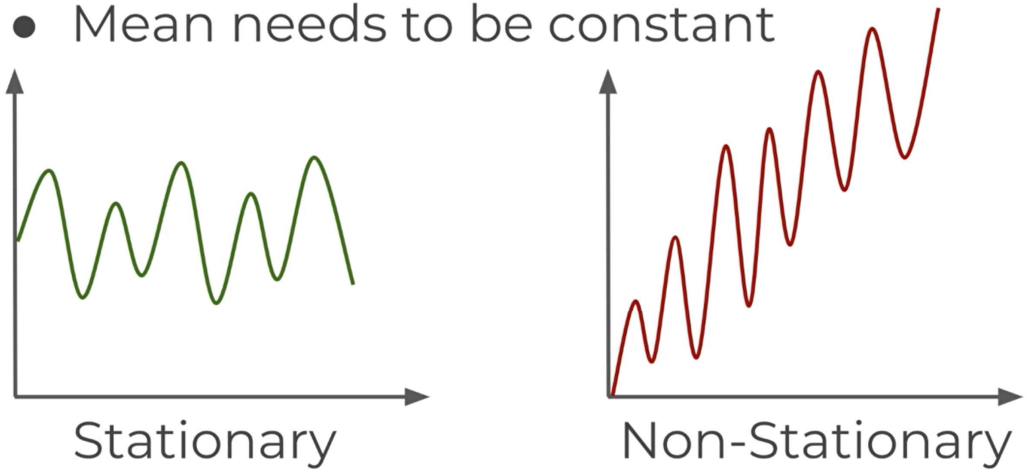
ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

Non-seasonal ARIMA models are generally denoted ARIMA(p,d,q) where parameters p, d, and q are non-negative integers.

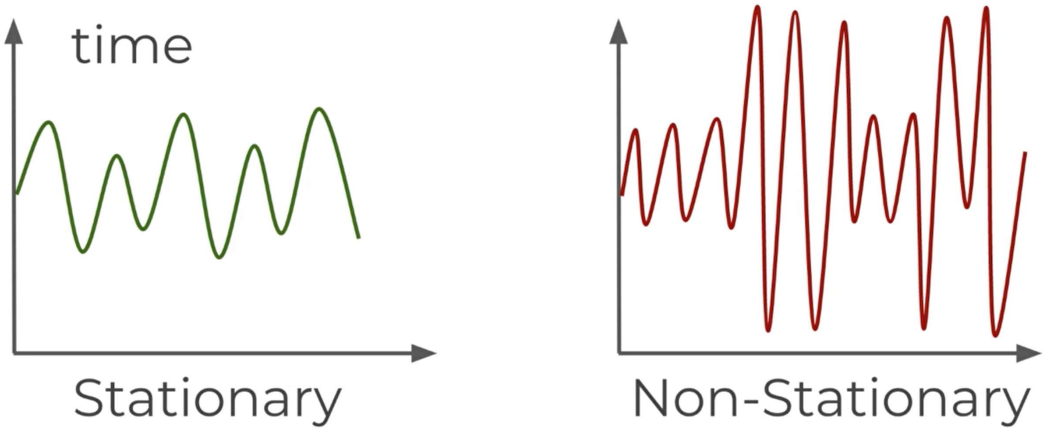
- Parts of ARIMA model
- AR (p): Autoregression
    - A regression model that utilizes the dependent relationship between a current observation and observations over a previous period
- I (d): Integrated.
    - Differencing of observations (subtracting an observation from an observation at the previous time step) in order to make the time series stationary.

- MA (q): Moving Average.
    - A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.
- Stationary vs Non-Stationary Data
    - To effectively use ARIMA, we need to understand Stationarity in our data.
    - So what makes a data set Stationary?
        - A Stationary series has constant mean and variance over time.

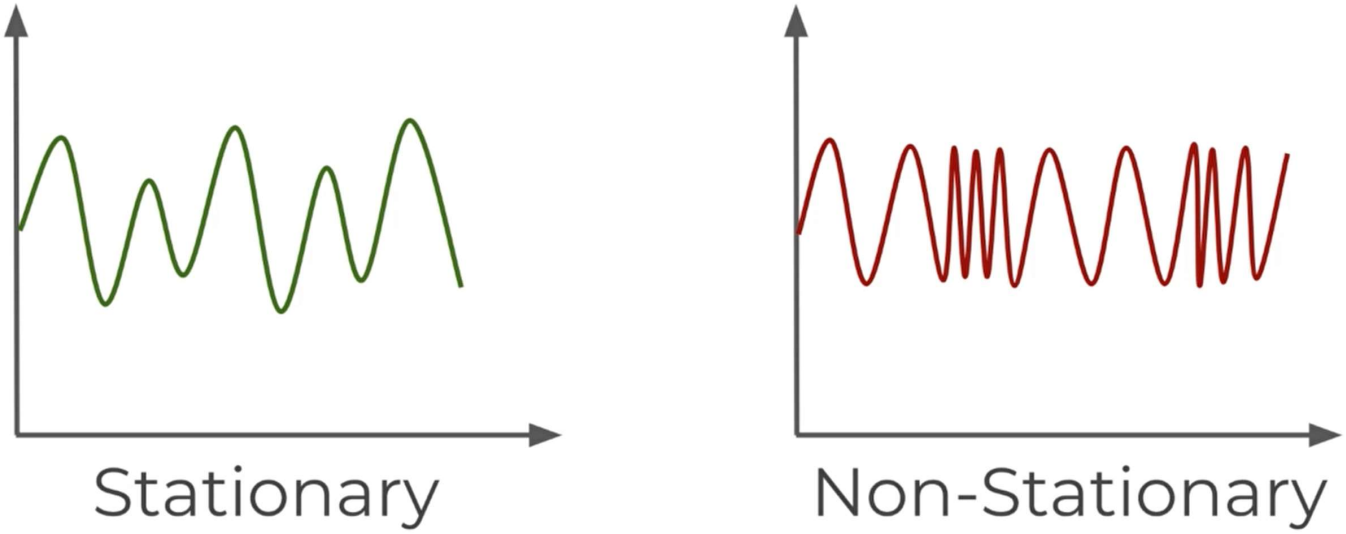
- Mean needs to be constant



- Variance should not be a function of time



Covariance should not be a function of time



There are also mathematical tests you can use to test for stationarity in your data.
A common one is the Augmented Dickey–Fuller test (we will see how to use this with Python's statsmodels)

If you've determined your data is not stationary (either visually or mathematically), you will then need to transform it to be stationary in order to evaluate it and what type of ARIMA terms you will use.

Original Data

Time1	10
Time2	12
Time3	8
Time4	14
Time5	7

First Difference

Time1	NA
Time2	2
Time3	-4
Time4	6
Time5	-7

Second Difference

Time1	NA
Time2	NA
Time3	-6
Time4	10
Time5	-13

You can continue differencing until you reach stationarity (which you can check visually and mathematically)

Each differencing step comes at the cost of losing a row of data!

- For seasonal data, you can also difference by a season.
- For example, if you had monthly data with yearly seasonality, you could difference by a time unit of 12, instead of just 1.

Another common technique with seasonal ARIMA models is to combine both methods, taking the seasonal difference of the first difference.

With your data now stationary it is time to go back and discuss the p,d,q terms and how you choose them.
There are two main ways to choose these p,d, and q terms.

Method One (Difficult):

- AutoCorrelation Plots and Partial AutoCorrelation Plots.
- Using these plots we can choose p,d and q terms based on viewing the decay in the plot.
- These plots can be very difficult to read, and often even when reading them correctly, the best performing p,d, or q value may be different than what is read.

Method Two (Easy but takes time):

- Grid Search
- Run ARIMA based models on different combinations of p, d, and q and compare the models for on some evaluation metric.

Method Two (Easy but takes time):

- Due to computational power becoming cheaper and faster, its often a good idea to use the built-in automated tools that search for the correct p, d, and q terms for us!

- SARIMA is very similar to ARIMA, but adds another set of parameters (P, D, and Q) for the seasonal component.

ARIMA stands for AutoRegression Integrated Moving Average.
If we drop the Integrated and Moving Average components, then we're only left with AR.

In an autoregression model, we forecast using a linear combination of past values of the variable. The term autoregression describes a regression of the variable against itself. An autoregression is run against a set of lagged values of order **p**.