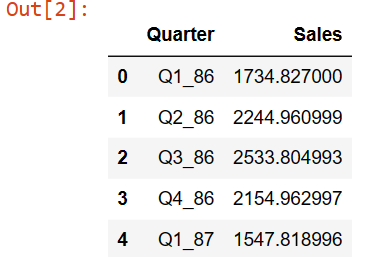
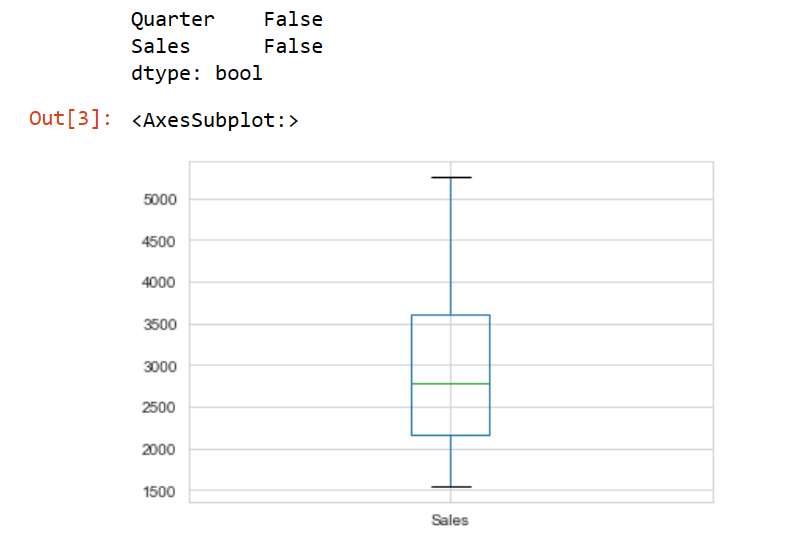
**Forecasting - CocaCola\_Sales\_Rawdata.xlsx**



Data contains two columns **Quarter** and **Sales**.

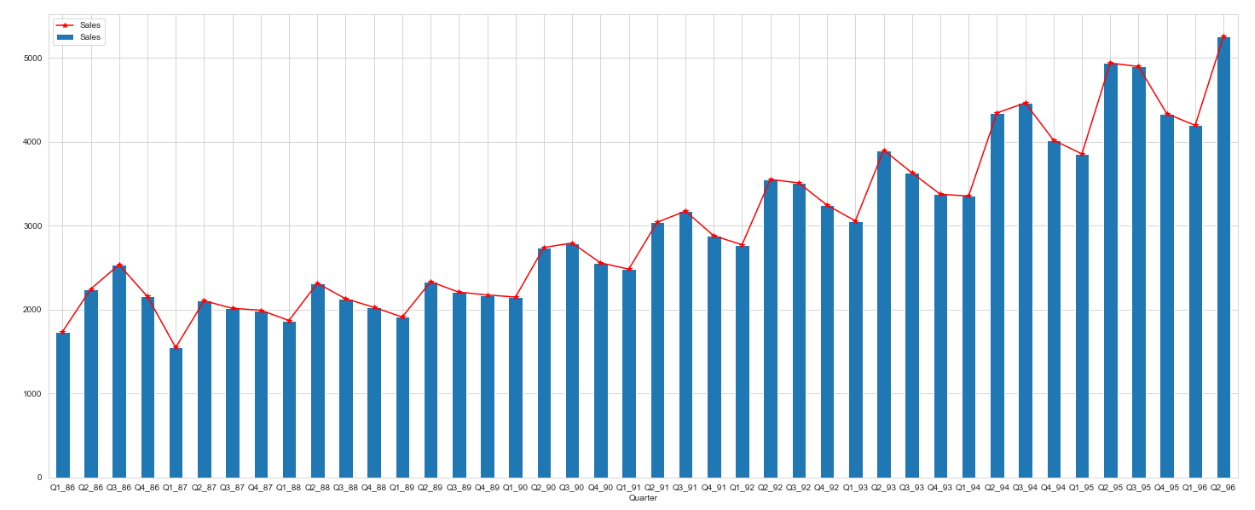
Data Analysis:

1. Check for missing values and outliers:



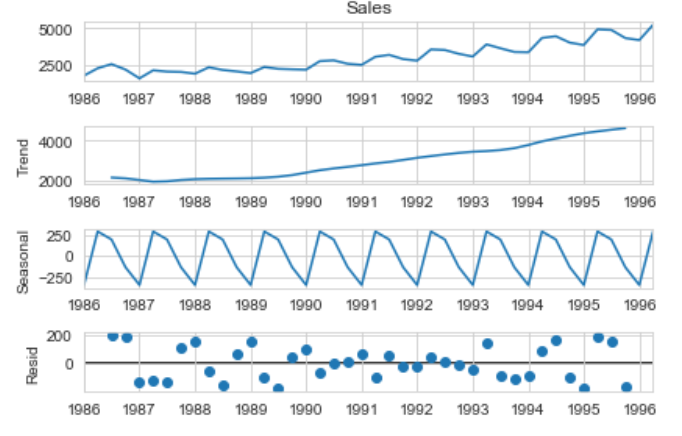
Data does not contain any missing value, and neither it contains outliers as observed through Boxplot.

1. Observe Pattern (Trend & Seasonality) in data:



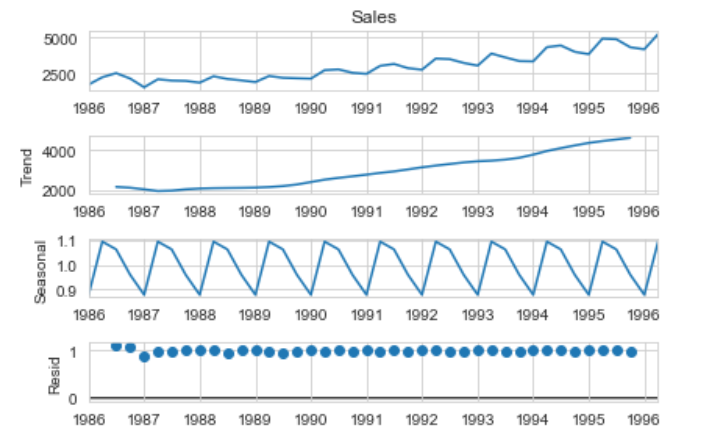
As it is clear from the above plot that there is an **upward trend** and **seasonality** in the data. We can cross-check the same using seasonal decomposition plots.

1. Seasonal Decomposition plots:
2. Seasonal decomposition using additive model.



As Trend and Seasonality is clearly visible, Residuals are randomly distributed, so better we will check the decomposition through multiplicative model.

1. Seasonal decomposition using multiplicative model

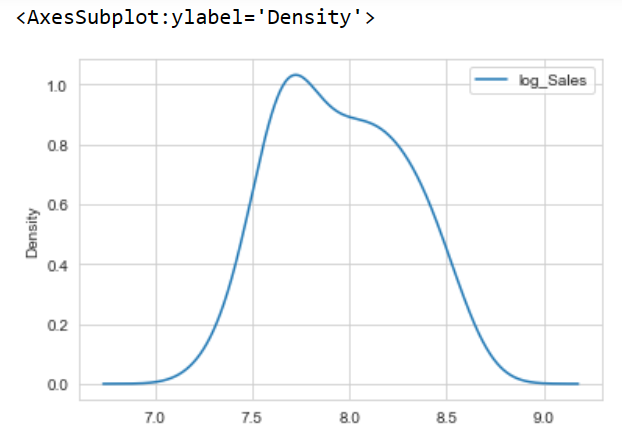


Residuals here are almost constant, thus confirming multiplicative seasonality.

1. We will do one more check of seasonality being additive or multiplicative.

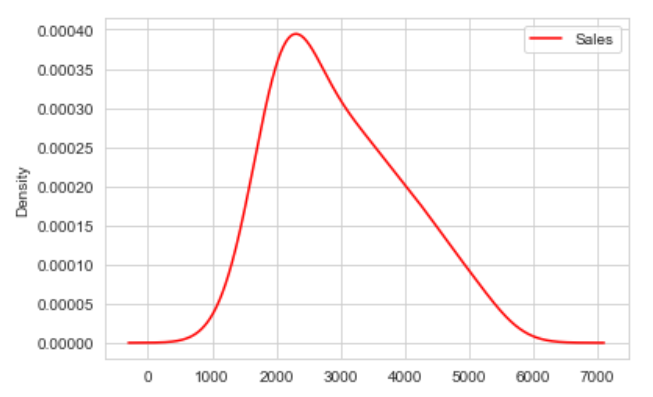
If the log value of sales is more normally distributed than actual values than it confirms multiplicativity of seasonality. This can be confirmed by either **plotting** the values or using **Shapiro-Wilk test of normality**.

1. Density Plot of original sales -





1. Density plot of log Sales-





As clearly visible from the plots that log values are better normally distributed than the actual values, Shapiro-Wilk test also confirming the same.

Shapiro-Wilk Test for Normality:

Null Hypothesis: Data is normally distributed.

1- For Original sales value p-value < 0.05 hence rejecting null hypothesis.

2- For log sales, p-value >> 0.05, so confirming data is almost normally distributed.

Hence Multiplicative seasonal model is more suitable.

Model Selection & Development: Since the data is showing both Trend and Seasonality, following methods can be considered.

1. Holt’s Winter method of Exponential Smoothing –

As though data has both Trend and Seasonality Component, Residuals are not varying significantly, so this method will work best. We will consider both trend (beta) and seasonal (gamma) component, for this model.

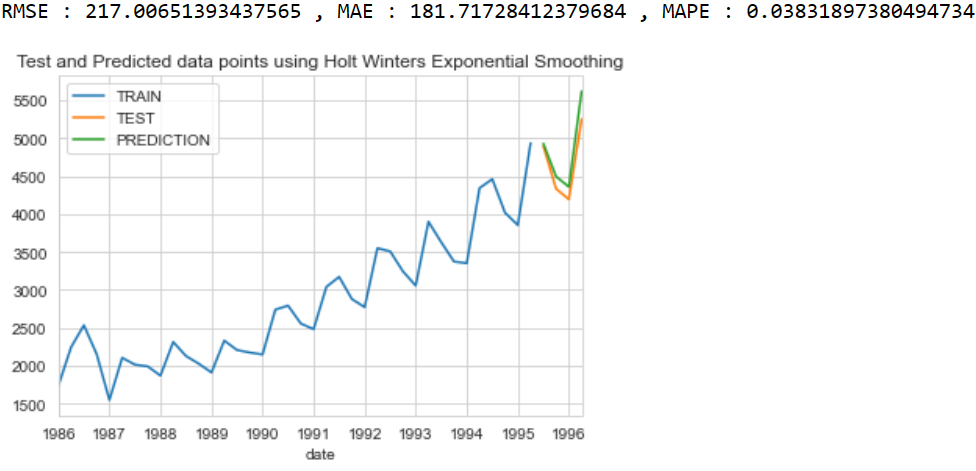
1. SARIMA – Seasonal Auto-Regressive Integrated Moving Average –

This model is more complex, usually used when data has both trend and season component and residuals are varying too much. Although this is not supposed to work as good as above method, I have also trained using this model just to showcase.

**Holt’s Winter method of Exponential Smoothing –**

1. Trained the model with both trend and seasonality as multiplicative.

Results are –

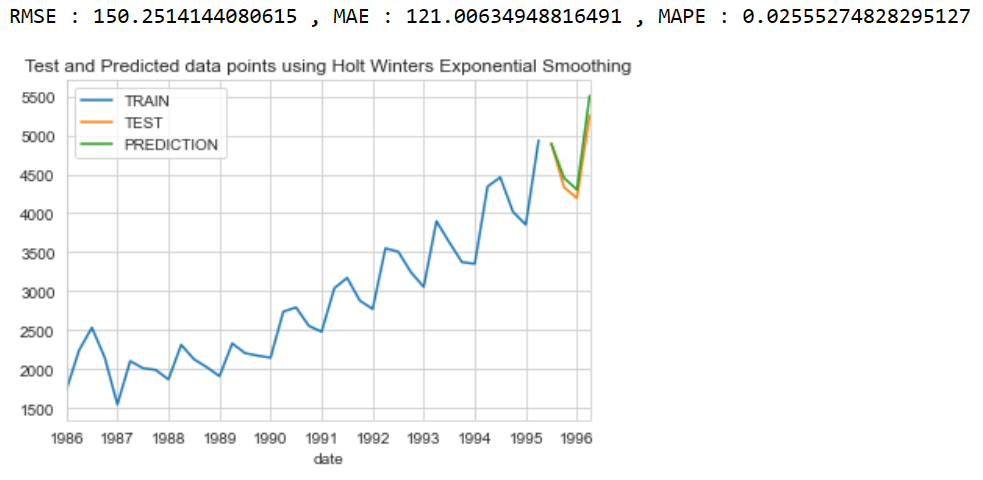
****

Model fits nicely as can be observed through plot, also errors are very less,

MAPE = 3.8% , showing 3.8% variation in actual and predicted values.

1. Trained the model with trend as additive and seasonality as multiplicative.

Results are –



Model fits even better, as MAPE reduces further to 2.5.

**Root Mean Square Error(RMSE), Mean Absolute Error(MAE) and Mean Absolute Percentage Error(MAPE) are used to Evaluate this model.**

**SARIMA – Seasonal Auto-Regressive Integrated Moving Average –**

SARIMA uses 7 parameters – (p, d, q) (P, D, Q, s)

**p** – Auto-Regressive component for trend

**d** – Order of differencing for trend

**q** – Moving average component for trend

**P** – Auto-Regressive component for Seasonality

**D** – Order of differencing for Seasonality

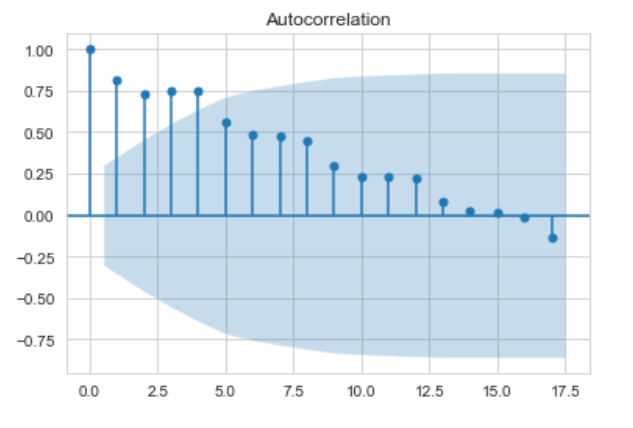
**Q** – Moving average component for Seasonality

**s** – Order of seasonality (Monthly=12, Quarterly=4, Yearly=1)

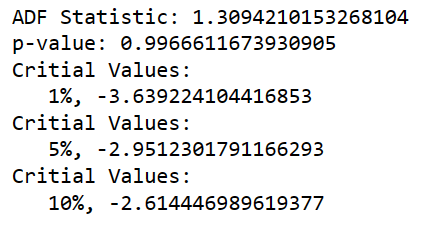
Steps for SARIMA model:

1. Check for Stationarity: Stationary data is the data without any trend or seasonality, so we can apply Simple regression to forecast the next value.

ACF plot or Augmented Dickey Fuller (ADF) Test can be used for checking stationarity.



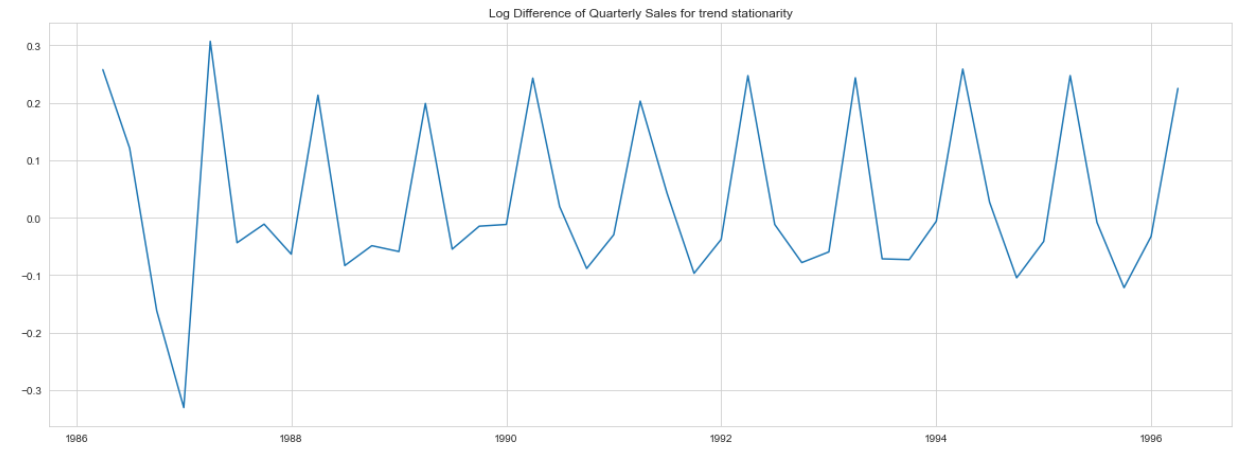
As is visible data is correlated with too many past lags so it is not stationary.



Augmented Dickey Fuller (ADF) Test- Null Hypothesis- Data is not stationary.

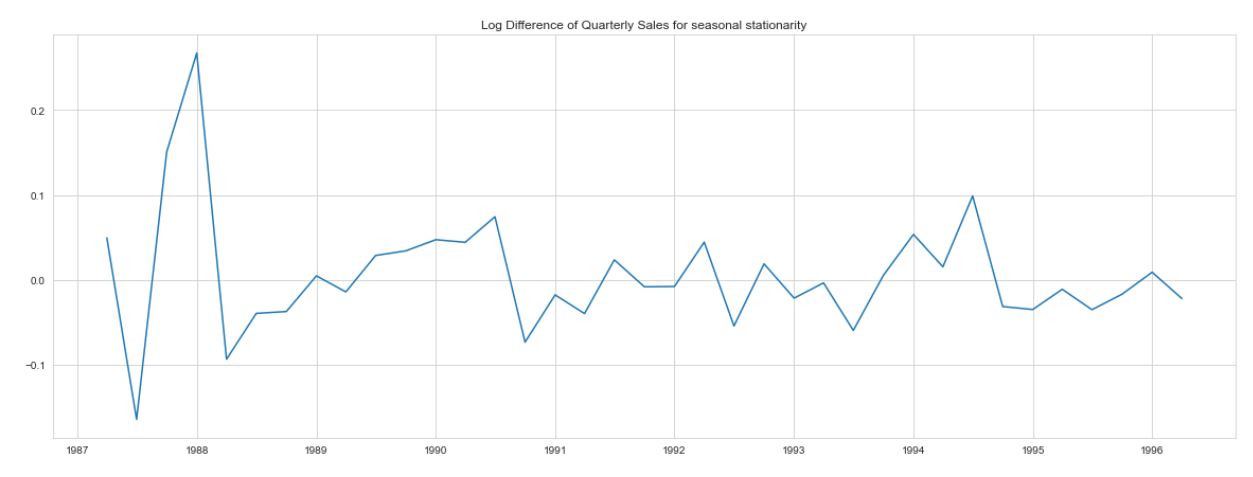
High p-value confirms that as we cannot reject null hypothesis.

1. So now our first task is to make Sales stationary by both trend and seasonality.
2. For this purpose, we will difference our data by lag 1 to remove trend.



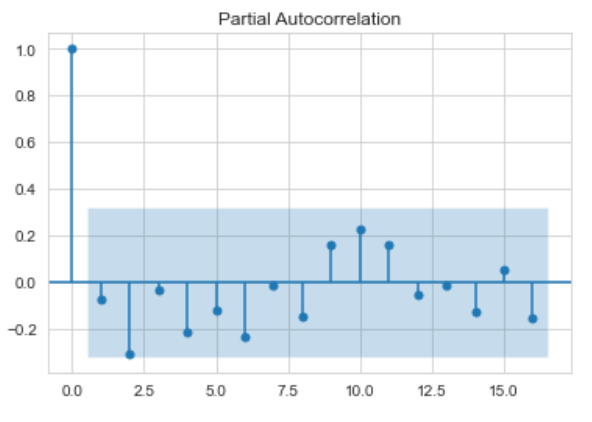
Data after differencing by lag 1, as visible does not have any trend but seasonal component is still present.

1. Differencing by lag 4 (Quarterly data), to remove seasonal component.

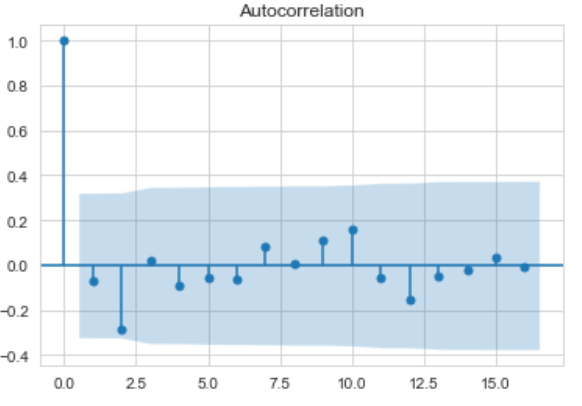


As clearly visible data now does not have either trend or seasonal component.

1. Plot ACF and PACF plots on differenced data to find q and p values.



p , P could be 1 as PACF is dampening soon after lag 1.



q , Q should be 1 as well as ACF is also dampening after lag 1.

1. We will still cross validate our model by using different combinations of p , q , P, Q values.

Which is as –

p = 0,1

d = 1

q = 0,1

P = 0,1

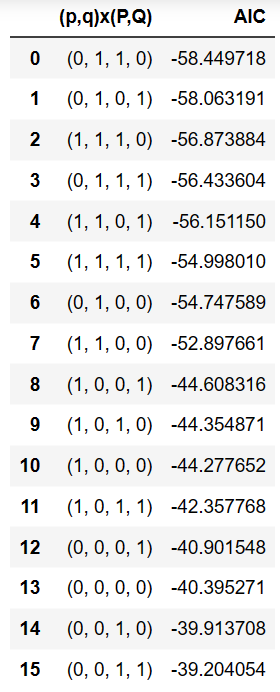
D = 1

Q = 0,1

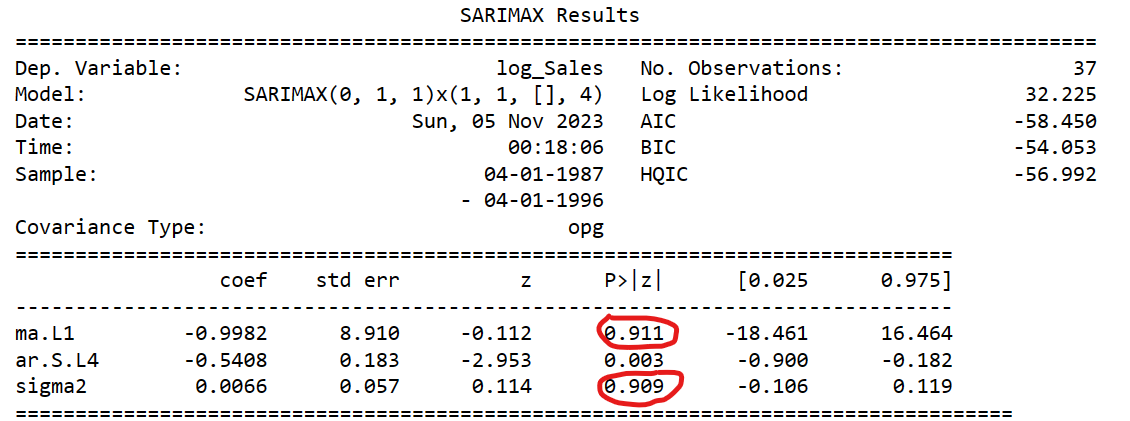
s = 4

Making it total of 16 combinations.

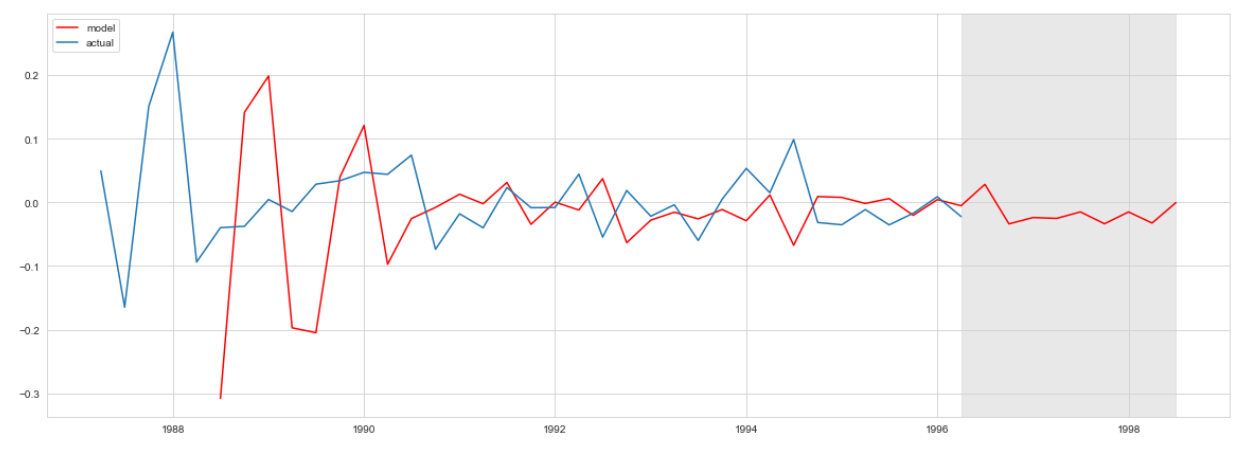
For each combination, we will calculate AIC value (Akaike Information Criterion). AIC is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. The lesser be the AIC, better is the model.



As shown, (p=0, q=1, P=1, Q=0) give the least AIC value we will use this model parameters to train out model. And the results are:



As two p-values are quite high shows model does not fit well with the data.



Also, the predicted plot (in red) does not fit with actual value plot (in blue) as good as with earlier HW method, shows SARIMA does not work as good as Holt’s Winter Exponential Smoothing.

**Conclusion**: So, as the matter-of-fact Holt’s Winter Exponential Smoothing is the most suitable model for this data and we can use it for model training for this data.