Shampoo Sales Forecast

Sai Jyothi Gurram

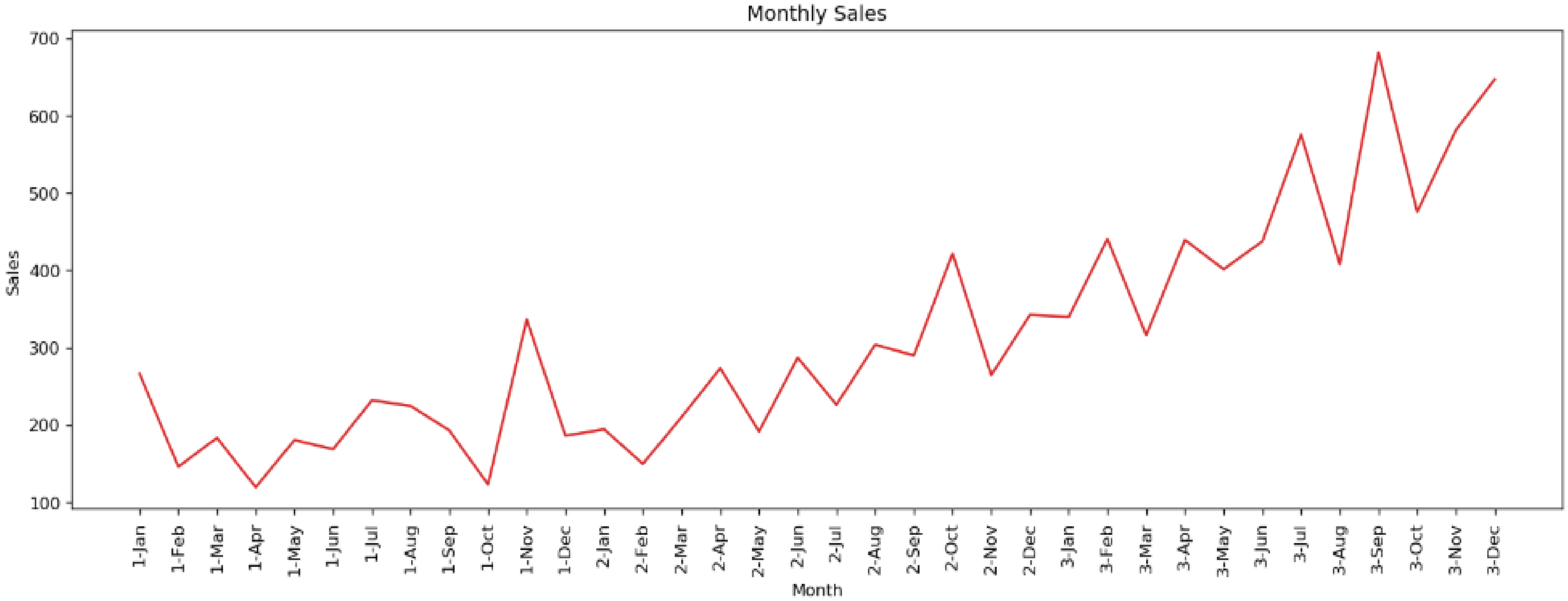
Data Overview

- The shampoo sales dataset has the sales information at Month level for three consecutive years.
- The dataset doesn’t contain any missing values.
- For easier analysis, Month has been set as the index.

Objective

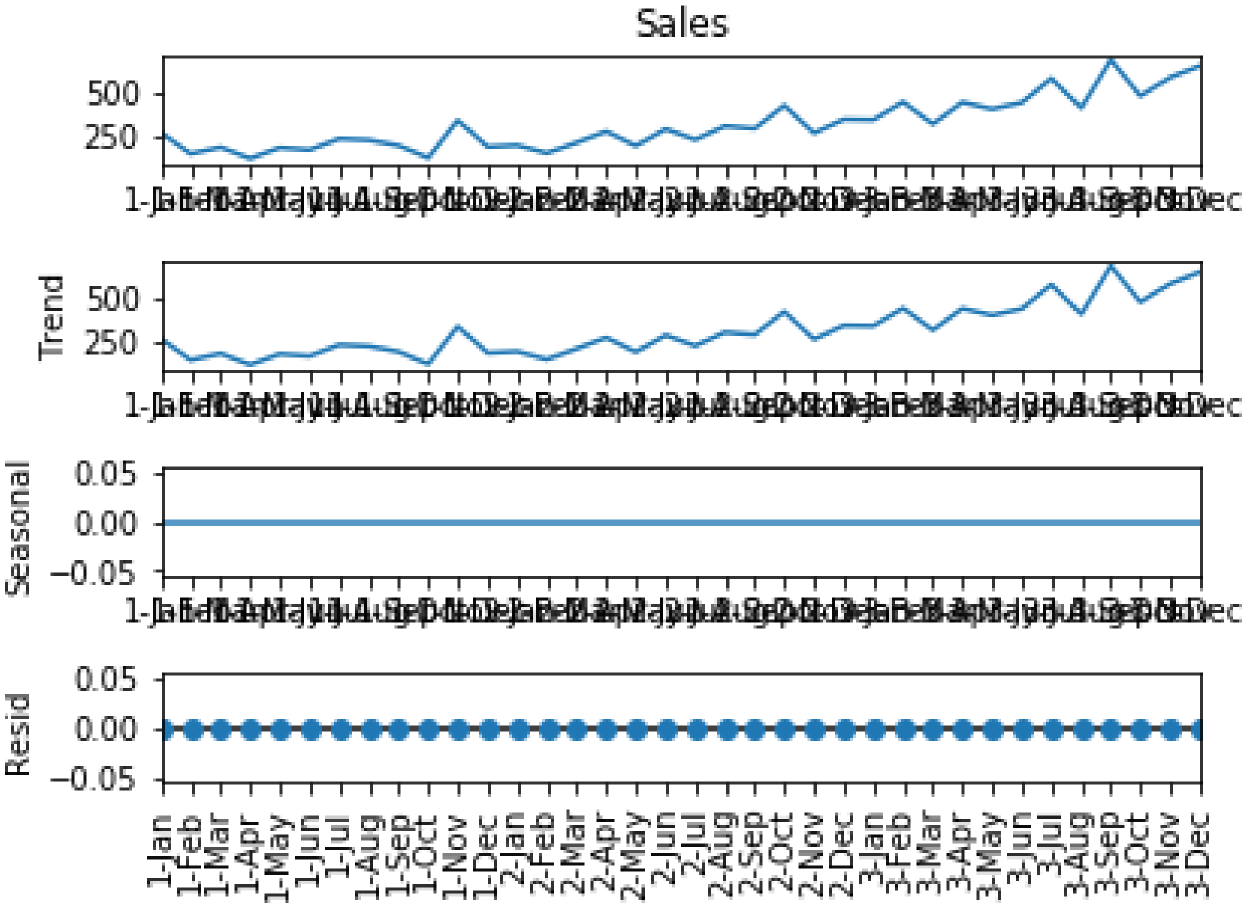
- Forecast the shampoo sales for any given period.

Shampoo Sales Distribution - EDA



The Sales distribution has an upward trend but cannot identify any seasonality

Seasonal Decomposition



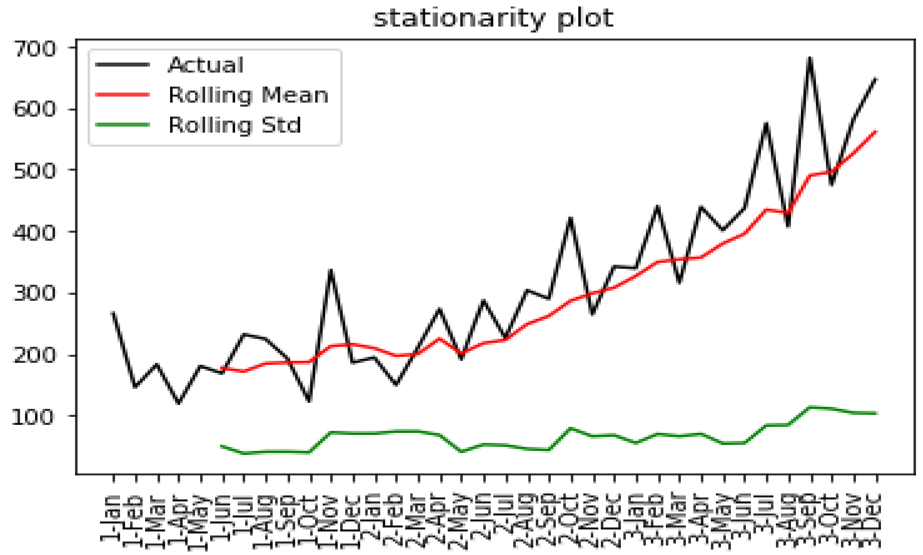
There is an upward trend but there is no seasonality and no noise in the residuals. There is no difference between using an additive or multiplicative model here.

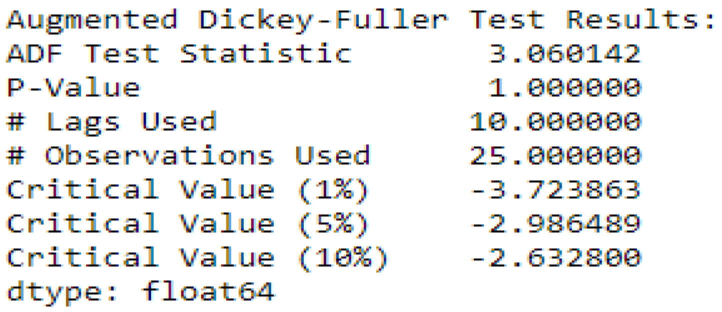
Stationarity

- Stationarity of the original data is checked by using rolling statistics and augmented Dickey-Fuller test.

- The mean has an upward trend while the standard deviation is approximately flat.

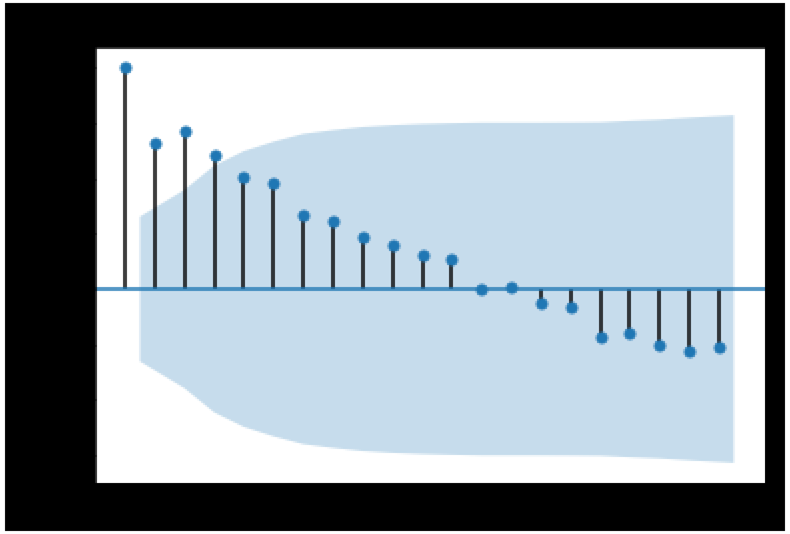
- The test statistic proves that the data is not stationary.

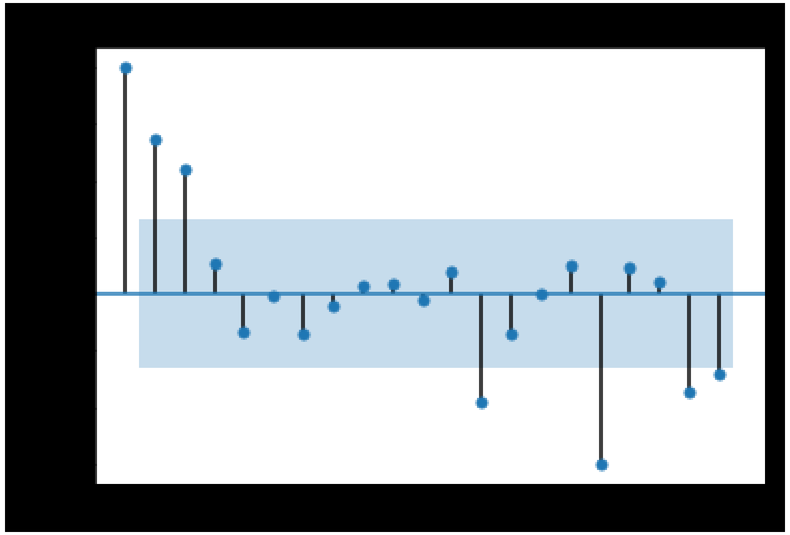




```
Augmented Dickey-Fuller Test Results:
ADF Test Statistic         3.060142
P-Value                    1.000000
# Lags Used               10.000000
# Observations Used       25.000000
Critical Value (1%)       -3.723863
Critical Value (5%)       -2.986489
Critical Value (10%)      -2.632800
dtype: float64
```

ACF and PACF plot

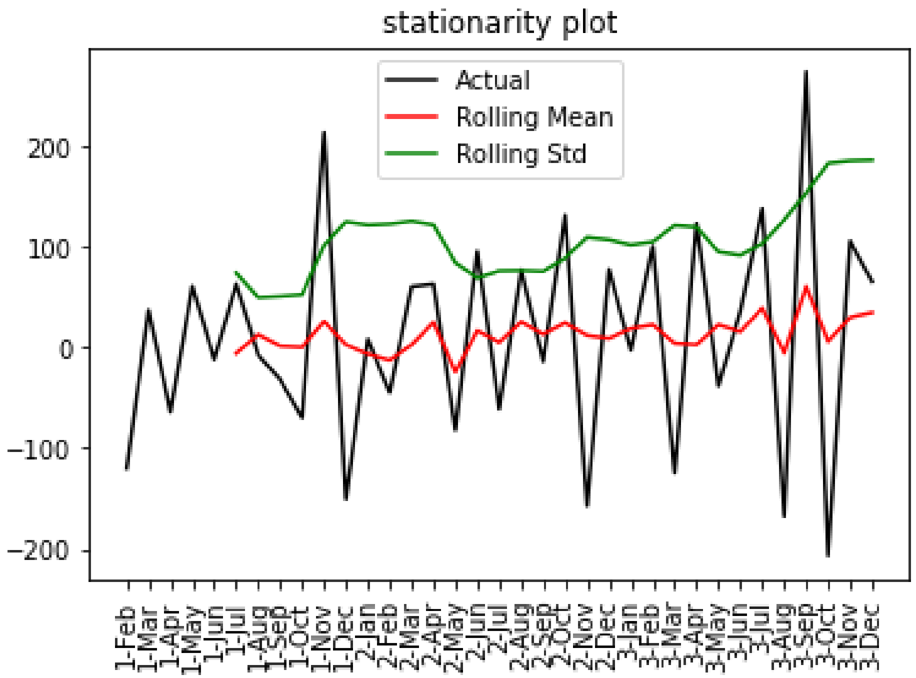


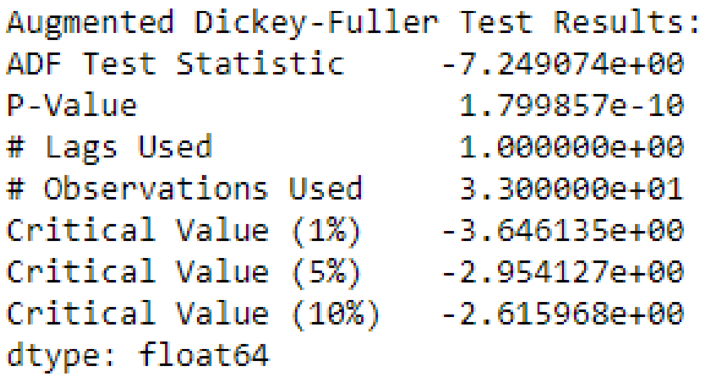


- The PACF plot clearly indicates that there is correlation until 3 lags.
- The ACF plot indicates that there is correlation until 4 lag errors.

Stationarity

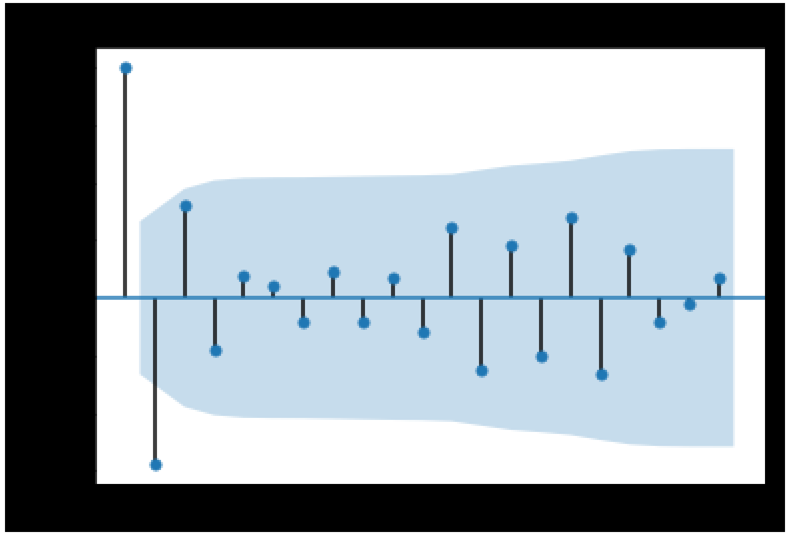
- Stationarity of the first order differenced data is checked by using rolling statistics and augmented Dickey-Fuller test.
- The mean and the standard deviation are approximately flat.
- The test statistic proves that the data is stationary.



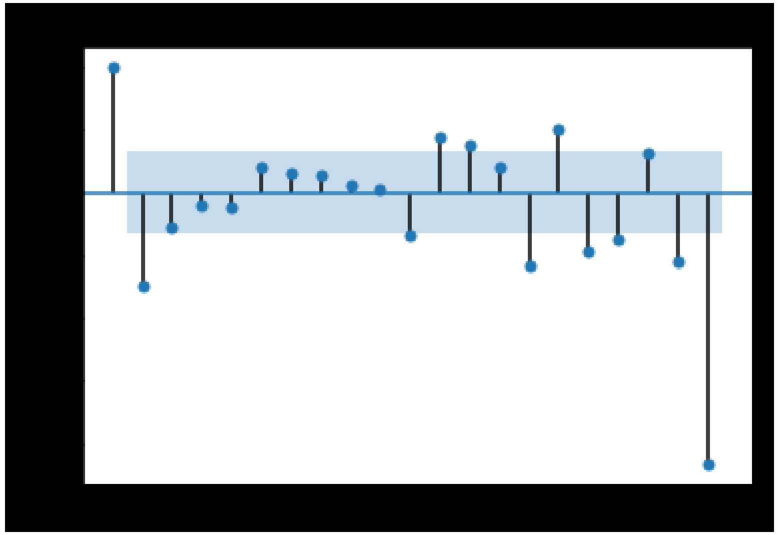


```
Augmented Dickey-Fuller Test Results:
ADF Test Statistic      -7.249074e+00
P-Value                  1.799857e-10
# Lags Used              1.000000e+00
# Observations Used      3.300000e+01
Critical Value (1%)     -3.646135e+00
Critical Value (5%)     -2.954127e+00
Critical Value (10%)    -2.615968e+00
dtype: float64
```

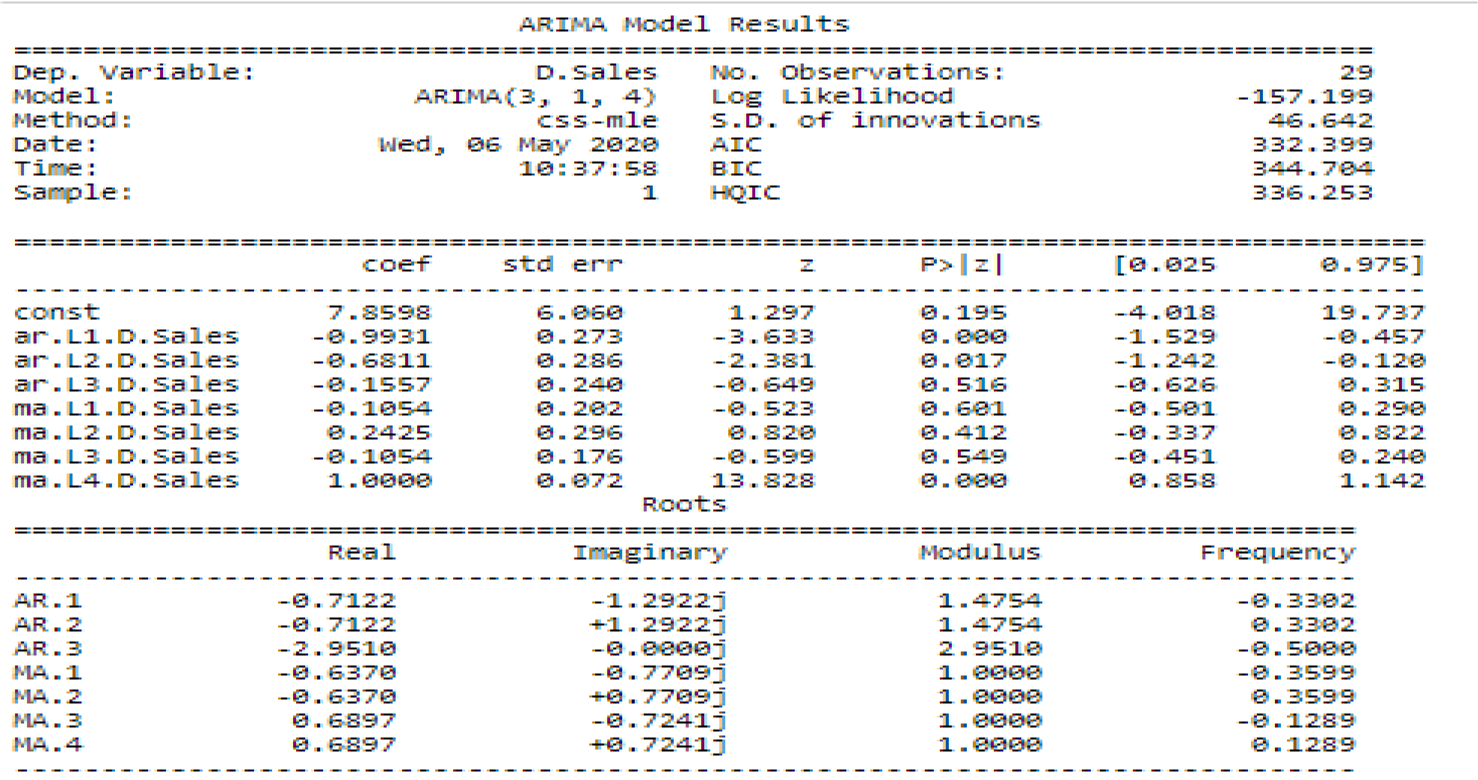
ACF and PACF plot - 1st order data



- The PACF plot clearly indicates that there is correlation until 2 lags.
- The ACF plot indicates that there is correlation until 2 lag errors.

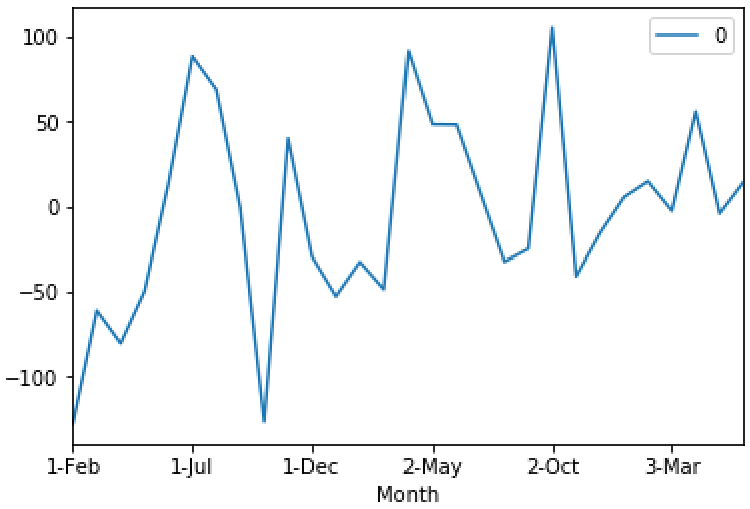


Base Model



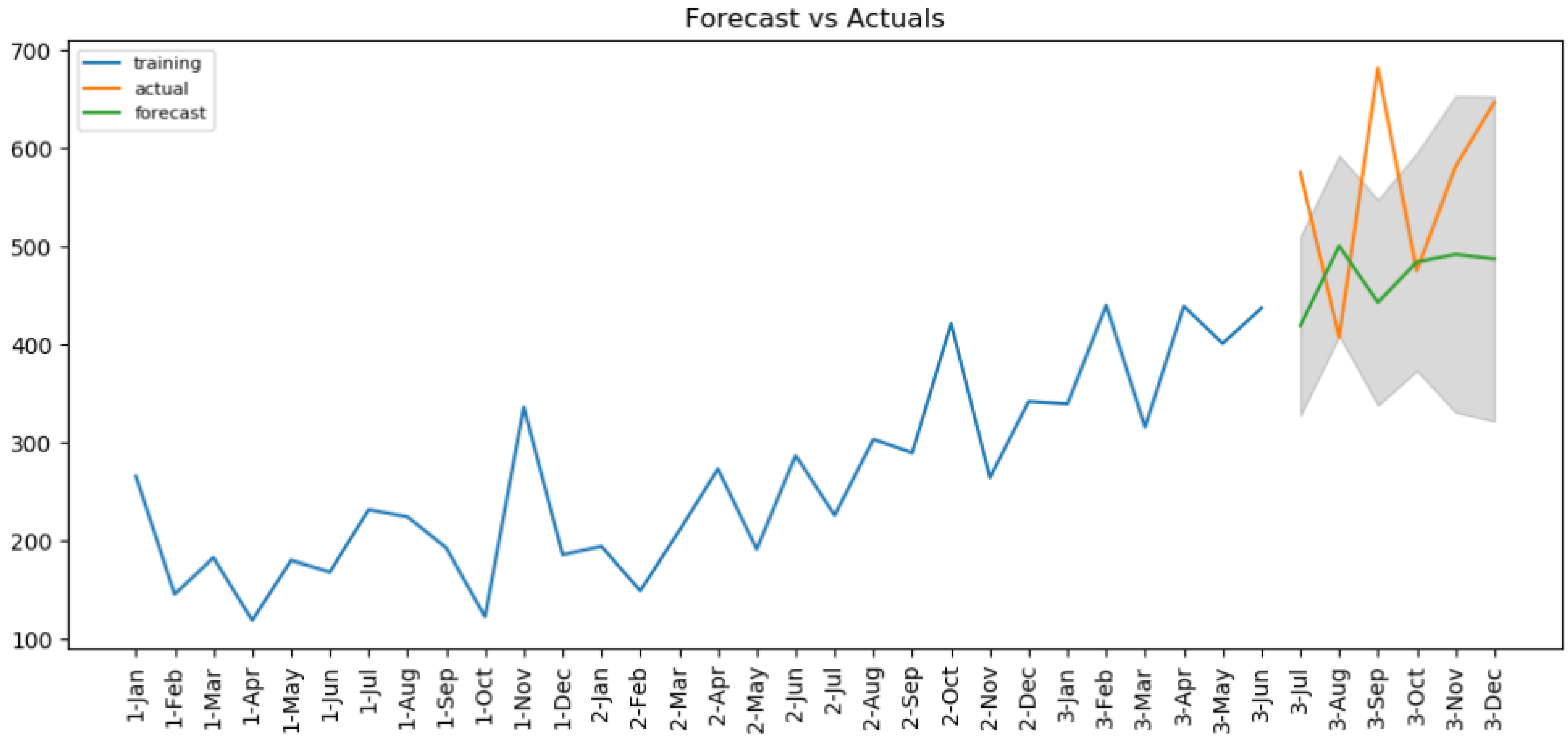
```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                D.Sales   No. Observations:                   29
Model:                 ARIMA(3, 1, 4)   Log Likelihood                -157.199
Method:                       css-mle   S.D. of innovations             46.642
Date:                Wed, 06 May 2020   AIC                            332.399
Time:                        10:37:58   BIC                            344.704
Sample:                             1   HQIC                           336.253

=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             7.8598      6.060      1.297      0.195      -4.018      19.737
ar.L1.D.Sales    -0.9931      0.273     -3.633      0.000      -1.529      -0.457
ar.L2.D.Sales    -0.6811      0.286     -2.381      0.017      -1.242      -0.120
ar.L3.D.Sales    -0.1557      0.240     -0.649      0.516      -0.626       0.315
ma.L1.D.Sales    -0.1054      0.202     -0.523      0.601      -0.501       0.290
ma.L2.D.Sales     0.2425      0.296      0.820      0.412      -0.337       0.822
ma.L3.D.Sales    -0.1054      0.176     -0.599      0.549      -0.451       0.240
ma.L4.D.Sales     1.0000      0.072     13.828      0.000       0.858       1.142
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.7122           -1.2922j            1.4754           -0.3302
AR.2           -0.7122           +1.2922j            1.4754            0.3302
AR.3           -2.9510           -0.0000j            2.9510           -0.5000
MA.1           -0.6370           -0.7709j            1.0000           -0.3599
MA.2           -0.6370           +0.7709j            1.0000            0.3599
MA.3            0.6897           -0.7241j            1.0000           -0.1289
MA.4            0.6897           +0.7241j            1.0000            0.1289
-----------------------------------------------------------------------------
```



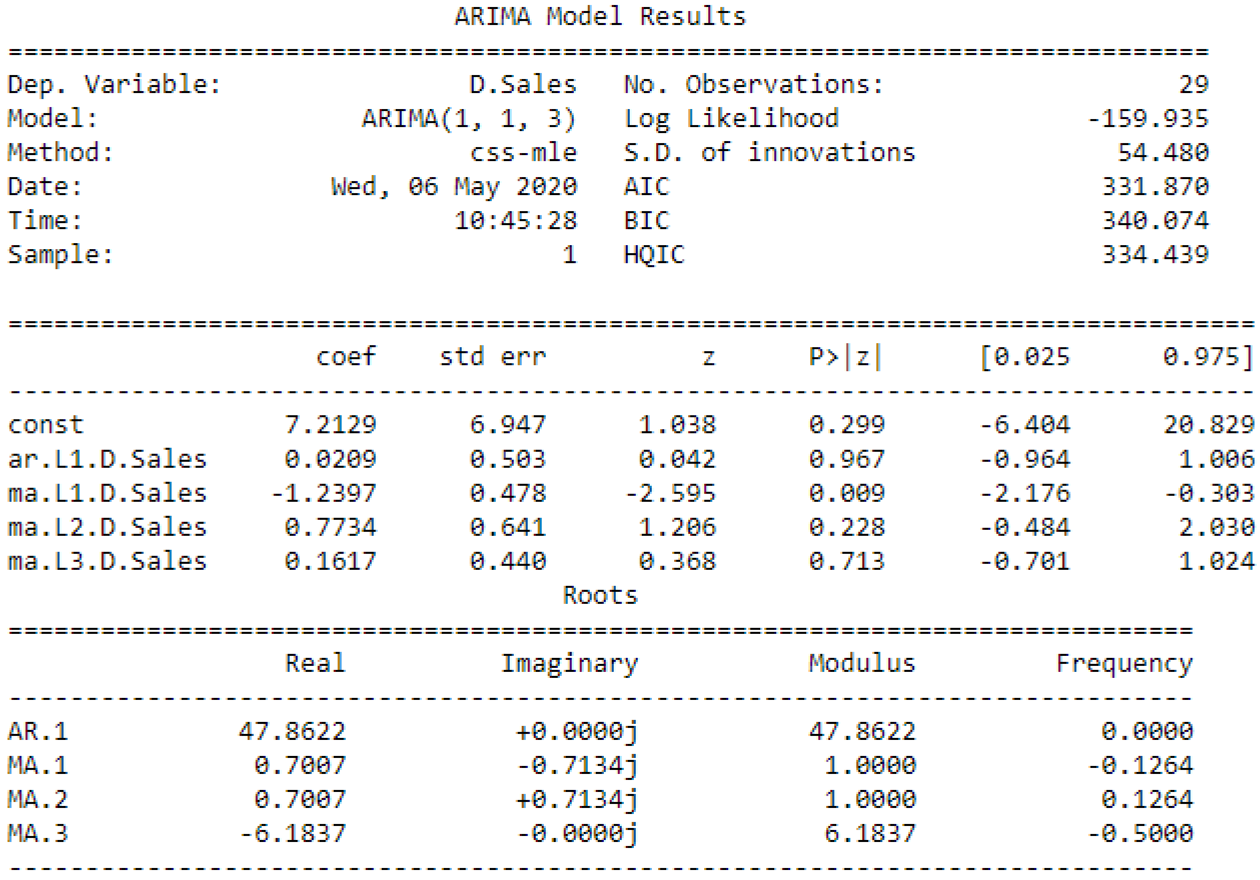
- The coefficients of the AR and MA terms is low . The AIC value is 332.
- The residuals still have slight upward trend and the mean value is not close to 0.

Base Model Forecasts VS Actuals



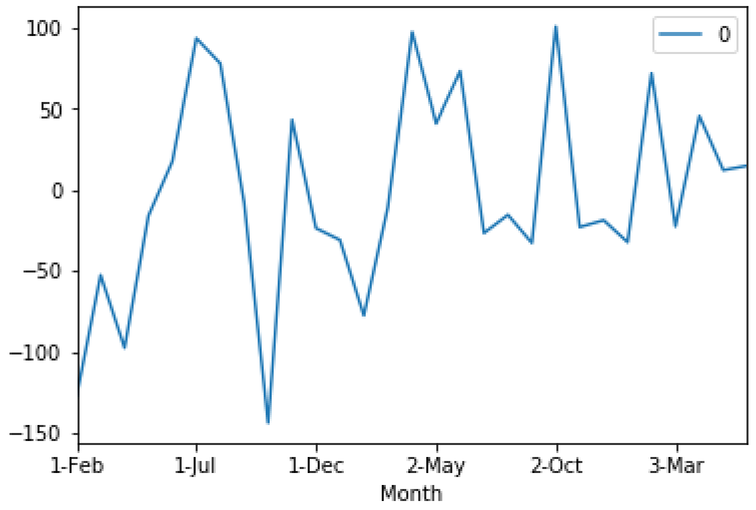
- The upward trend is captured but the AIC value and the error values should be reduced.
- The MAPE error of the base model is 0.211

Final Model



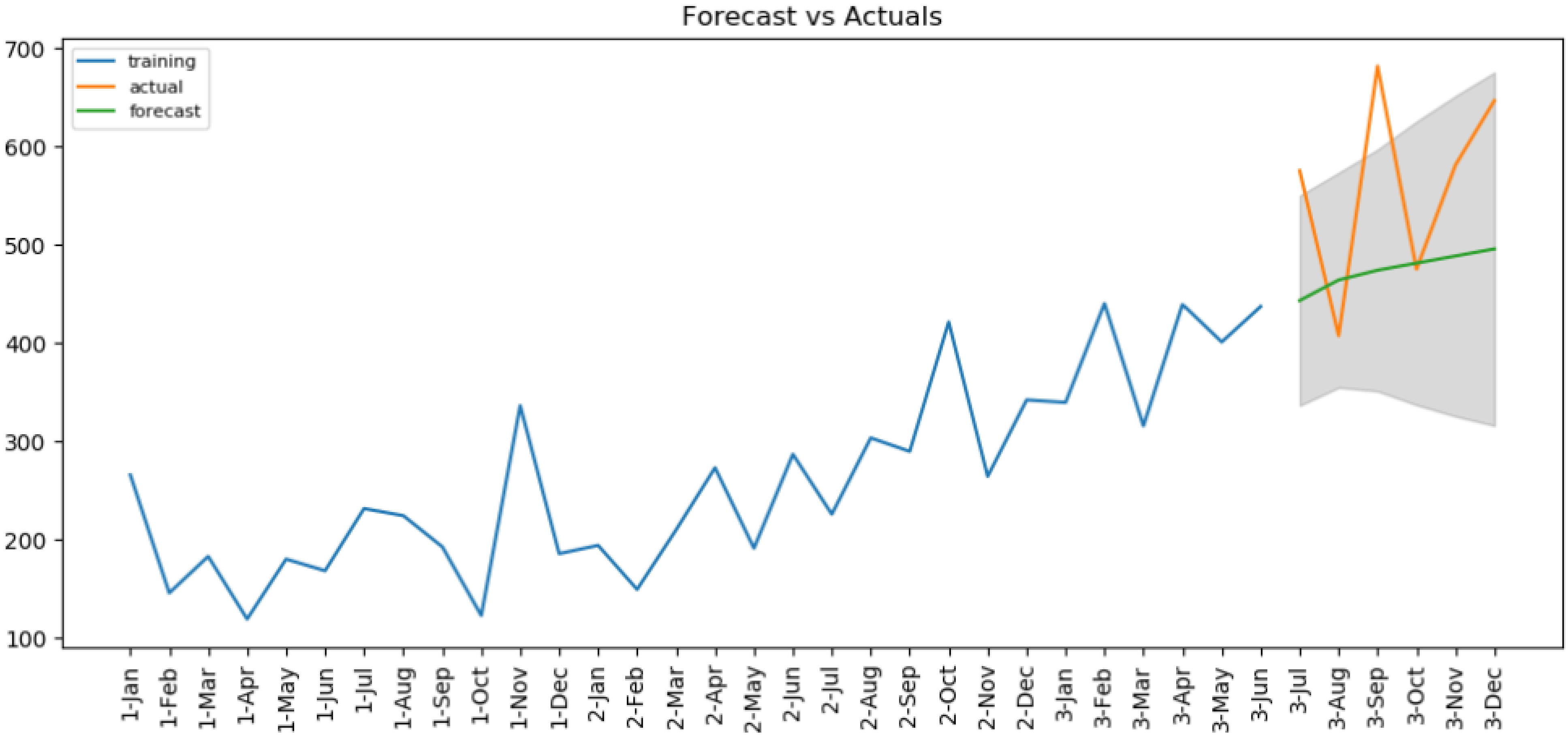
```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                D.Sales   No. Observations:                   29
Model:                 ARIMA(1, 1, 3)   Log Likelihood                -159.935
Method:                       css-mle   S.D. of innovations             54.480
Date:                Wed, 06 May 2020   AIC                            331.870
Time:                        10:45:28   BIC                            340.074
Sample:                             1   HQIC                           334.439

=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             7.2129      6.947      1.038      0.299      -6.404      20.829
ar.L1.D.Sales     0.0209      0.503      0.042      0.967      -0.964       1.006
ma.L1.D.Sales    -1.2397      0.478     -2.595      0.009      -2.176      -0.303
ma.L2.D.Sales     0.7734      0.641      1.206      0.228      -0.484       2.030
ma.L3.D.Sales     0.1617      0.440      0.368      0.713      -0.701       1.024
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           47.8622           +0.0000j           47.8622            0.0000
MA.1            0.7007           -0.7134j            1.0000           -0.1264
MA.2            0.7007           +0.7134j            1.0000            0.1264
MA.3           -6.1837           -0.0000j            6.1837           -0.5000
-----------------------------------------------------------------------------
```



- The coefficients of the AR term is low but the MA terms are now very significant. The AIC value is 331 which is slightly less than the base model.
- The residuals still have slight upward trend and the mean value is -2 which is closer to 0 compared to the base model.

Final Model Forecasts VS Actuals



- The upward trend is captured.
- The MAPE error of the base model is 0.17.

Summary

- This model can now be used to forecast the shampoo sales for any given period.
- Applied exponential smoothing techniques to make the series stationary but first order differentiation worked better.

Areas of Improvement:

- The MAPE error here is still 0.17 in the final model. This error can further be reduced.
- Along with log transformation and rolling statistics other differentiation techniques can be explored.
- The test statistic of the AR and MA can be improved to make it statistically significant.
- Other algorithms like LSTM, Prophet can be used to optimize the forecasts.