Topic: Forecasting – Time Series

Forecast the Coca-Cola prices and Airlines Passengers data set. Prepare a document for each model explaining how many dummy variables you have created and RMSE value for each model. Finally which model you will use for Forecasting.

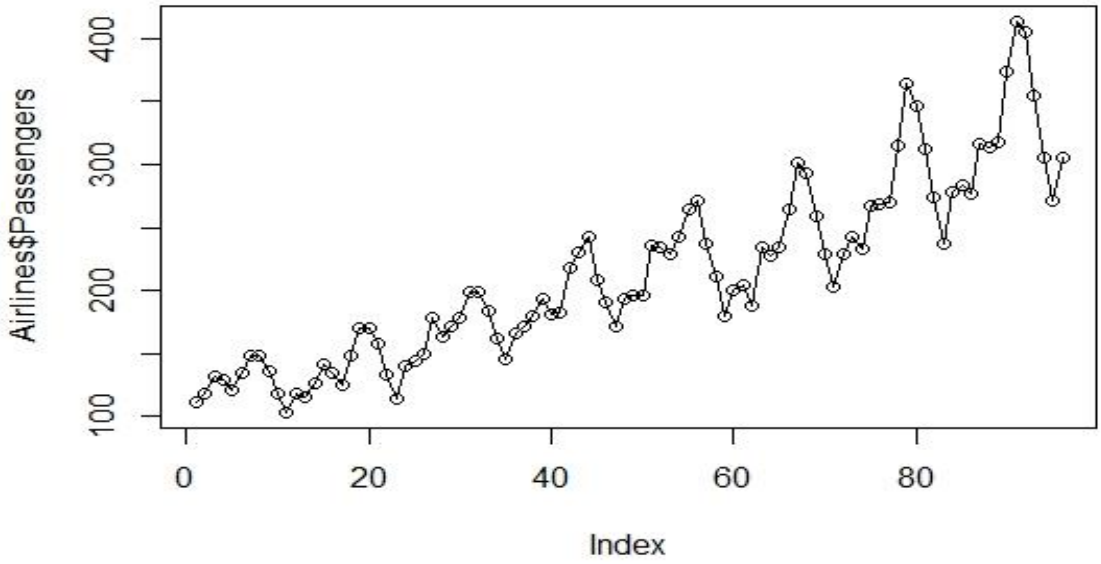
1.) Airlines.xlsx

	Month	Passengers
1	1995-01-01	112
2	1995-02-01	118
3	1995-03-01	132
4	1995-04-01	129
5	1995-05-01	121
6	1995-06-01	135
7	1995-07-01	148
8	1995-08-01	148
9	1995-09-01	136
10	1995-10-01	119
11	1995-11-01	104
12	1995-12-01	118
13	1996-01-01	115
14	1996-02-01	126
15	1996-03-01	141
16	1996-04-01	135

Firstly we need to convert the dataset to csv to remove the time stamp error in python and proceed with data preprocessing steps.

Data Preprocessing:

1) The dataset consists of 96 observations with 2 variables of Month-Year and passengers list.
2) Below is the graph that represents passengers list for the given period of time



3) Checking the NA values, as there are no NA values no further imputation is required.
4) As the data is of 12 months, so the frequency is taken as 12 and created 12dummy variables.
5) Now the months names are assigned to the columns names of dummy variables and combined with the airline data.
6) A time column “t” is assigned to the dataset and taking the log of passengers and square of t, so that the whole data is normalized.

Splitting the data:

1) Now the data is divided into training and test data with [1:70] and [71:96]
2) Calculating the RMSE value using different models, as we cannot directly tell what is the exact trend followed by the data
3) Only residual values are calculated for all the models to calculate the RMSE values

Calculating RMSE Using Different Models:

Linear Model:

- ✓ RMSE is 48.30 and Adjusted R2 Value is 0.7699

Exponential Model:

- ✓ RMSE is 43.47 and Adjusted R2 is 0.78
- ✓ As predicted values are logged values, we do exponential of expo_pred$fit to get the actual values.

Quadratic Model:

- ✓ RMSE is 43.898 and Adjusted R2 Value is 0.769

Additive Seasonality Model:

- ✓ RMSE is 124.97and Adjusted R2 is 0.083 Hence, it may not be additive seasonality model.

Additive Seasonality with Linear Model:

- ✓ RMSE is 34.502 and Adjusted R2 is 0.94

Additive Seasonality with Quadratic Model:

- ✓ RMSE is 30.393 and Adjusted R2 is 0.95

Multiplicative Seasonality Model:

- ✓ In multiplicative we multiply but we can't multiply directly hence we apply log
- ✓ RMSE is 129.6291 & Adjusted R2 is 0.07

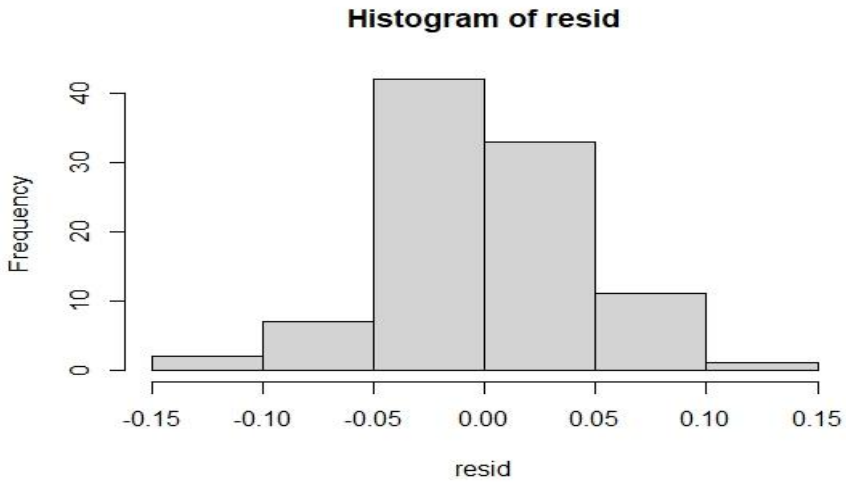
Multiplicative Seasonality Linear trend:

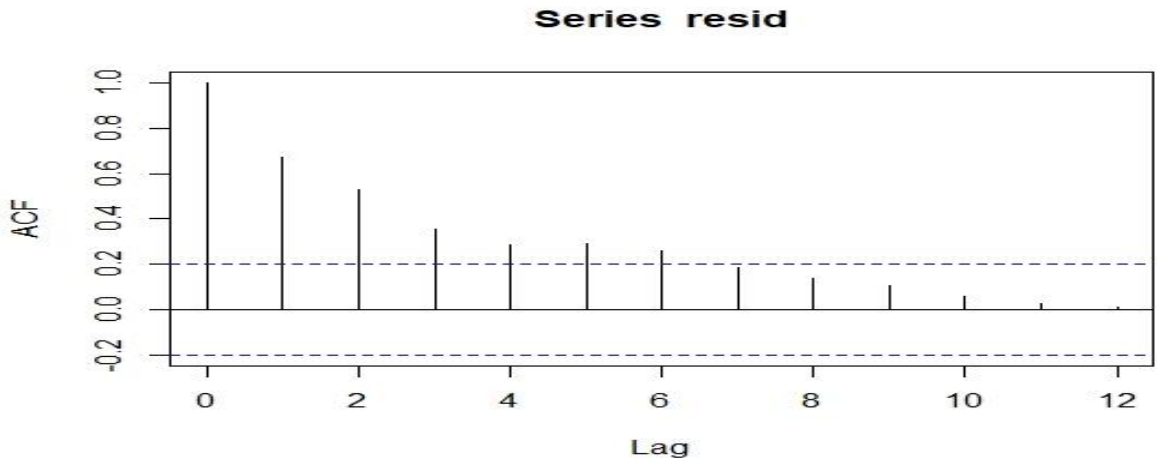
- ✓ RMSE is 11.72 and Adjusted R2 is 0.96
- ✓ This is the highest R2 & lowest RMSE

Below are the over all RMSE values for all models

	model	RMSE
1	rmse_linear	48.30986
2	rmse_expo	43.47847
3	rmse_Quad	43.89814
4	rmse_Add_season	124.97570
5	rmse_Add_sea_Quad	30.39304
6	rmse_multi_sea	129.62914
7	rmse_multi_add_sea	11.72479

- ✓ Now building the model with the whole dataset of airlines
- ✓ Histogram of residual values

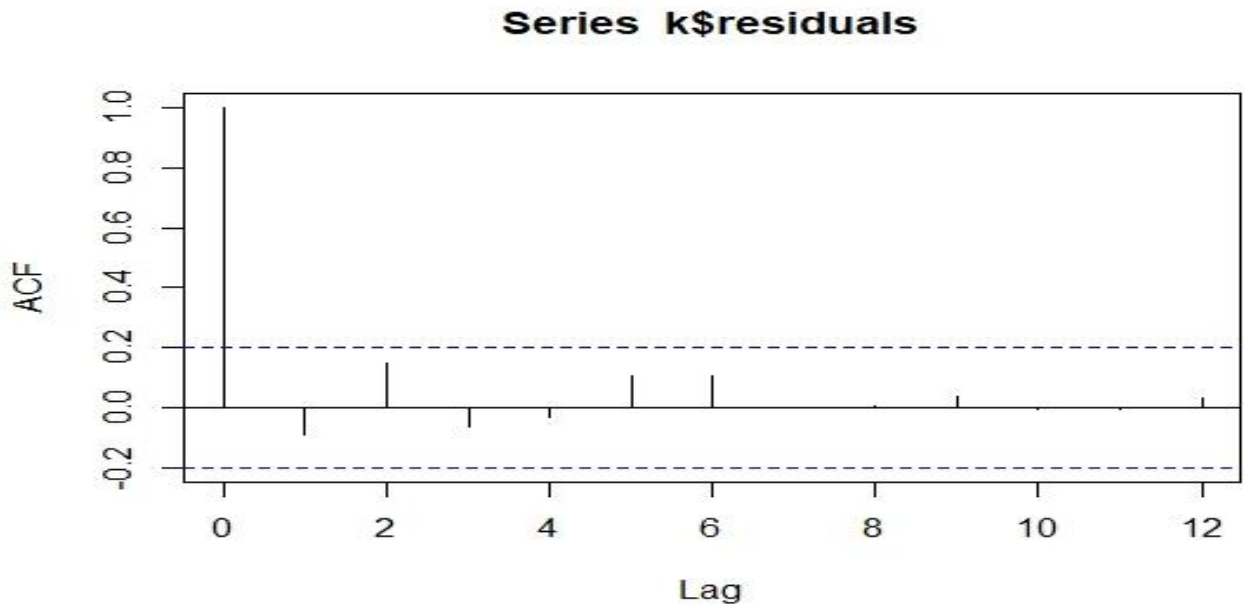




- ✓ From the above residual graph, we can say that the lag 1 to 4 are significant, so Arima can be built

Building the Arima Model:

- ✓ Auto regression is only used to forecast errors
- ✓ Performing auto regression with 2nd lag, p=2,d=0,q=0



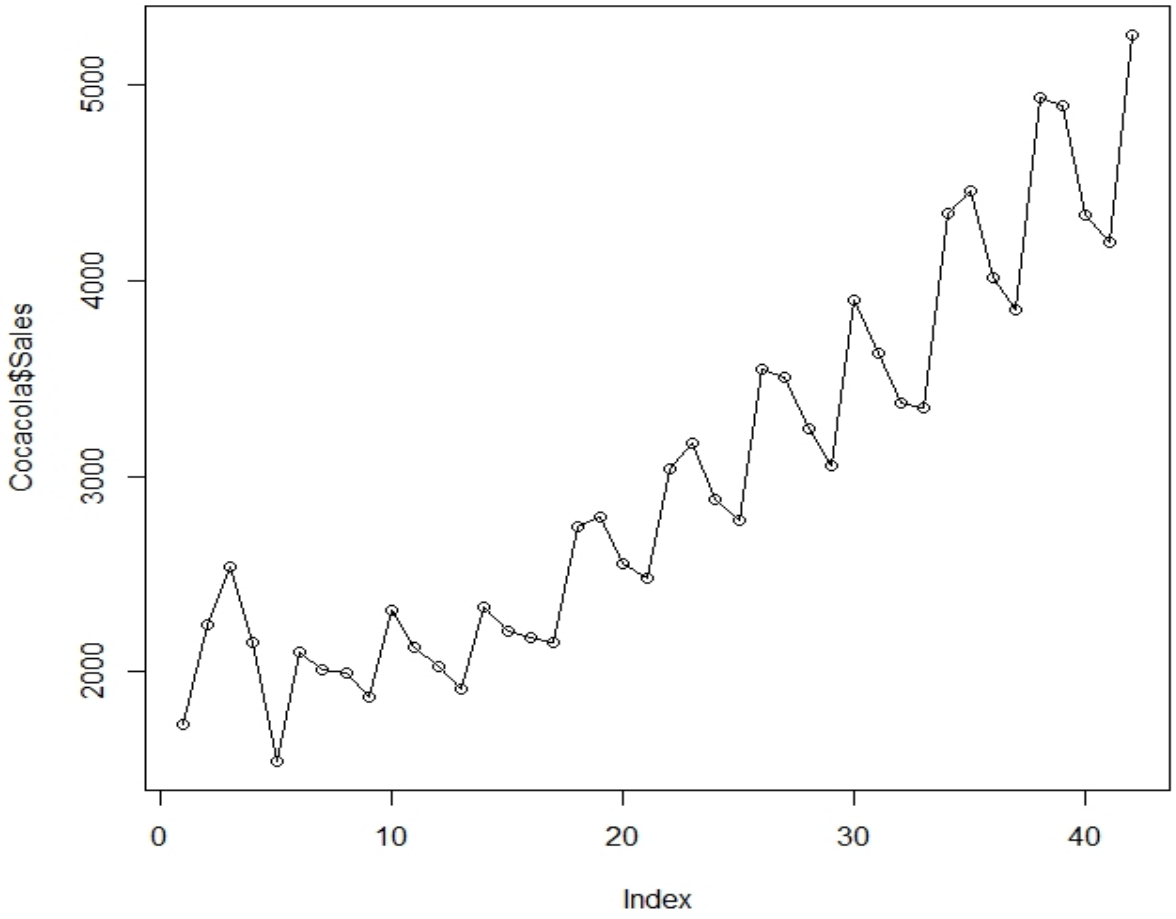
- ✓ From the above graph, we can say that significance problem is removed & all are below threshold ACF values.
- ✓ Using the function pred_res$pred then recalling function acf(k$residuals,lag.max = 12) we get the output of the problem statement.

2.) CocaCola_Sales_RawData.xlsx

	Quarter	Sales
1	Q1_86	1734.827
2	Q2_86	2244.961
3	Q3_86	2533.805
4	Q4_86	2154.963
5	Q1_87	1547.819
6	Q2_87	2104.412
7	Q3_87	2014.363
8	Q4_87	1991.747
9	Q1_88	1869.050
10	Q2_88	2313.632
11	Q3_88	2128.320
12	Q4_88	2026.829
13	Q1_89	1910.604
14	Q2_89	2331.165
15	Q3_89	2206.550
16	Q4_89	2173.968
17	Q1_90	2148.278

Data Preprocessing:

1) The dataset consists of 42 observations with 2 variables of Month-Year and Sales for the 4 quarters in a specific period.
2) Below is the graph that represents sales for the given period of time.



3) As the data is of 4 quarters, so the frequency is taken as 4 and created 4 dummy variables
4) Now the quarters names are assigned to the columns names of dummy variables and combined with the data.
5) A time column “t” is assigned to the dataset and taking the log of sales and square of t, so that the whole data is normalized.

Splitting the data:

1) Now the data is divided into training and test data with [1:30] and [31:42]
2) Calculating the RMSE value using different models, as we cannot directly tell what is the exact trend followed by the data
3) Only residual values are calculated for all the models to calculate the RMSE values

Calculating RMSE Using Different Models

Linear Model:

- ✓ RMSE is 714.014 and Adjusted R2 Value is 0.69

Exponential Model:

- ✓ RMSE is 552.28 and Adjusted R2 is 0.69
- ✓ RMSE has reduced of the exponential model than linear.
- ✓ As predicted values are logged values, we do exponential of expo_pred$fit to get actual values

Quadratic Model:

- ✓ RMSE is 646.27 and Adjusted R2 Value is 0.79

Additive Seasonality Model:

- ✓ RMSE is 1778.00 and Adjusted R2 is 0.05. Hence, it may not be additive seasonality model.

Additive Seasonality with Linear Model:

- ✓ RMSE is 637.94 and Adjusted R2 is 0.82

Additive Seasonality with Quadratic Model:

- ✓ RMSE is 586.05 and Adjusted R2 is 0.94

Multiplicative Seasonality Model:

- ✓ In multiplicative we multiply but we can't multiply directly hence we apply log
- ✓ RMSE is 1828.92 and Adjusted R2 is 0.07

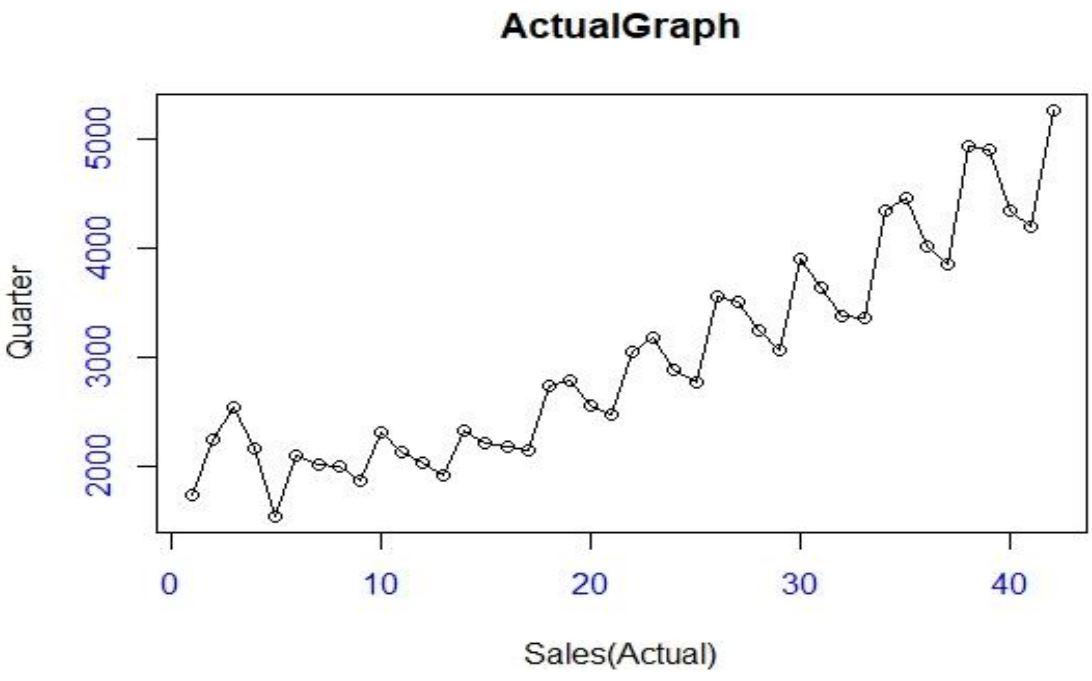
Multiplicative Seasonality Linear trend:

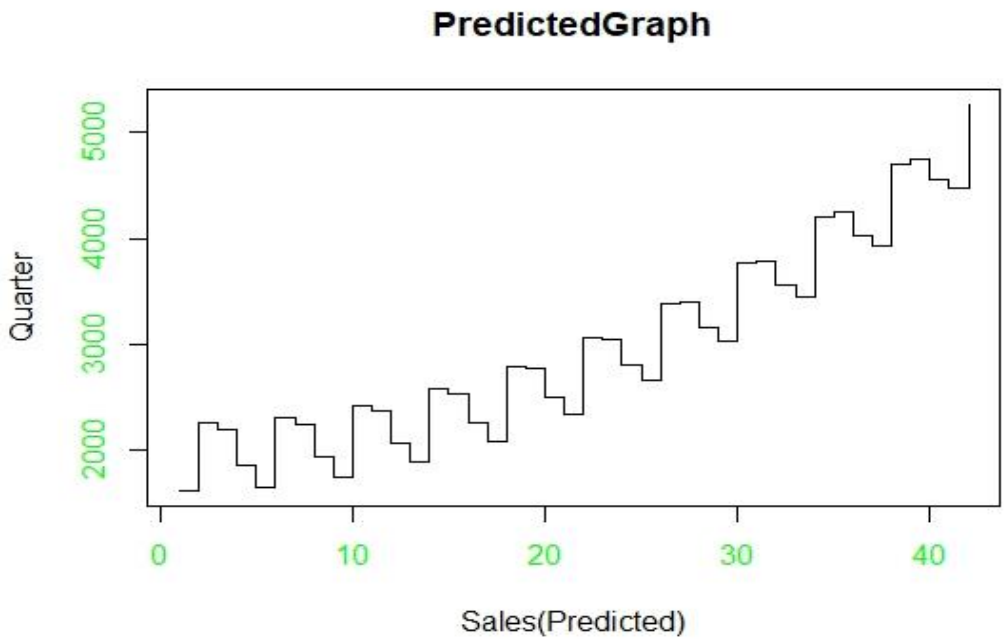
- ✓ RMSE is 410.24 and Adjusted R2 is 0.83
- ✓ This is the highest R2 & lowest RMSE

Below are the overall RMSE values for all models

	model	RMSE
1	rmse_linear	714.0144
2	rmse_expo	552.2821
3	rmse_Quad	646.2715
4	rmse_Add_season	1778.0065
5	rmse_Add_sea_Quad	586.0533
6	rmse_multi_sea	1828.9239
7	rmse_multi_add_sea	410.2497

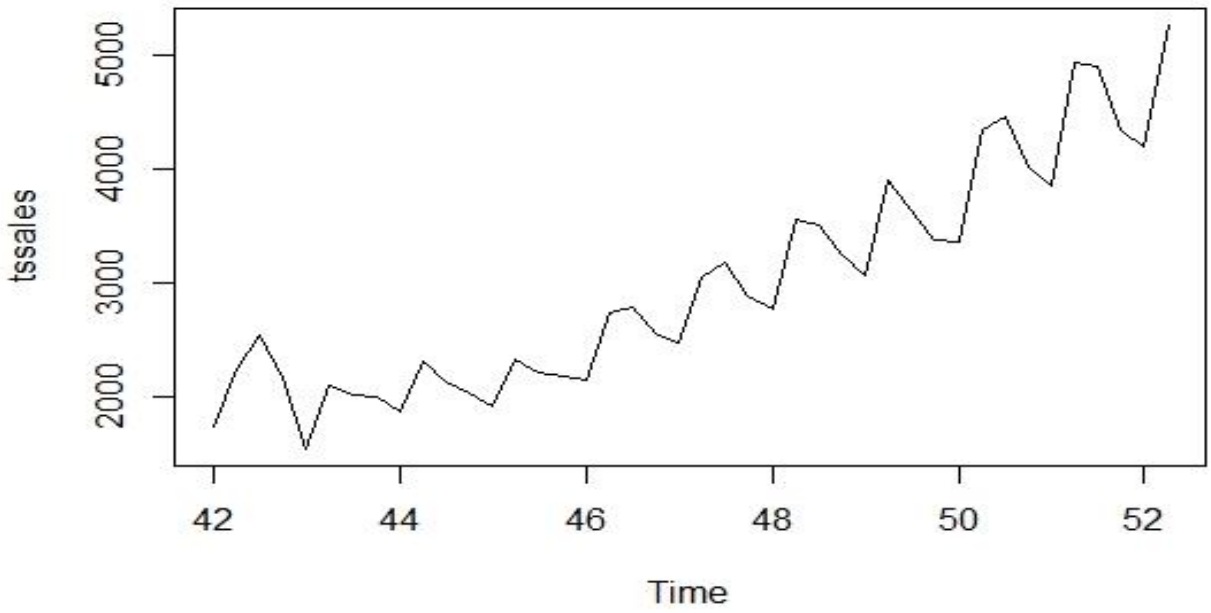
- ✓ Here we find that Multiplicative additional Seasonality with Linear trend which has least RMSE value of 410.24
- ✓ Now building the model with the whole dataset





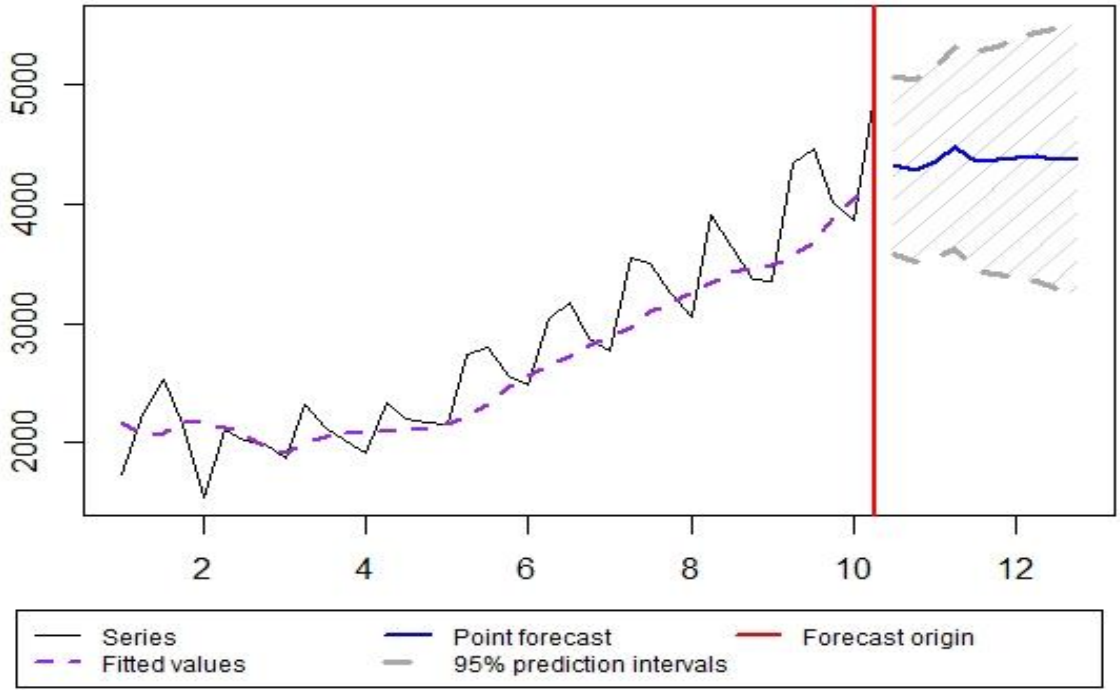
MAPE and MAE Graphical Representation:

Time Series Graph:



- ✓ Visualization shows that it has level, trend, seasonality => Additive seasonality

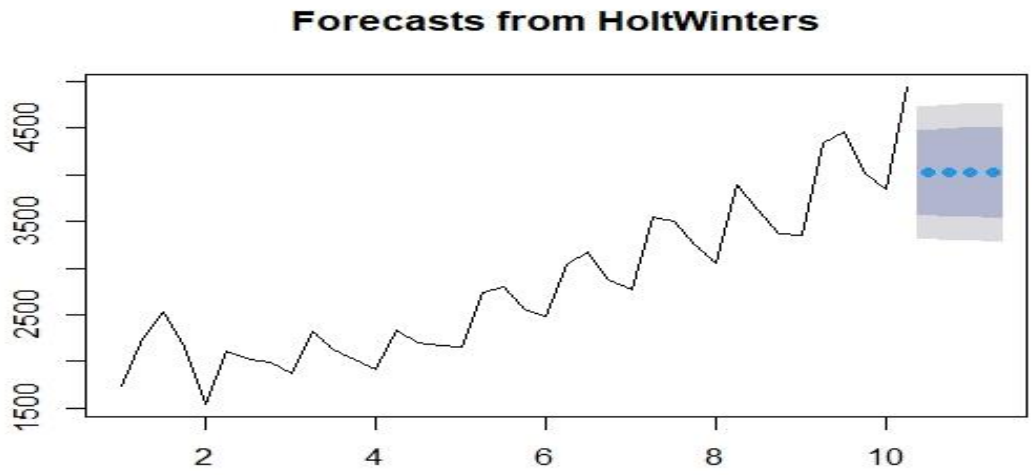
Moving Average Graph:

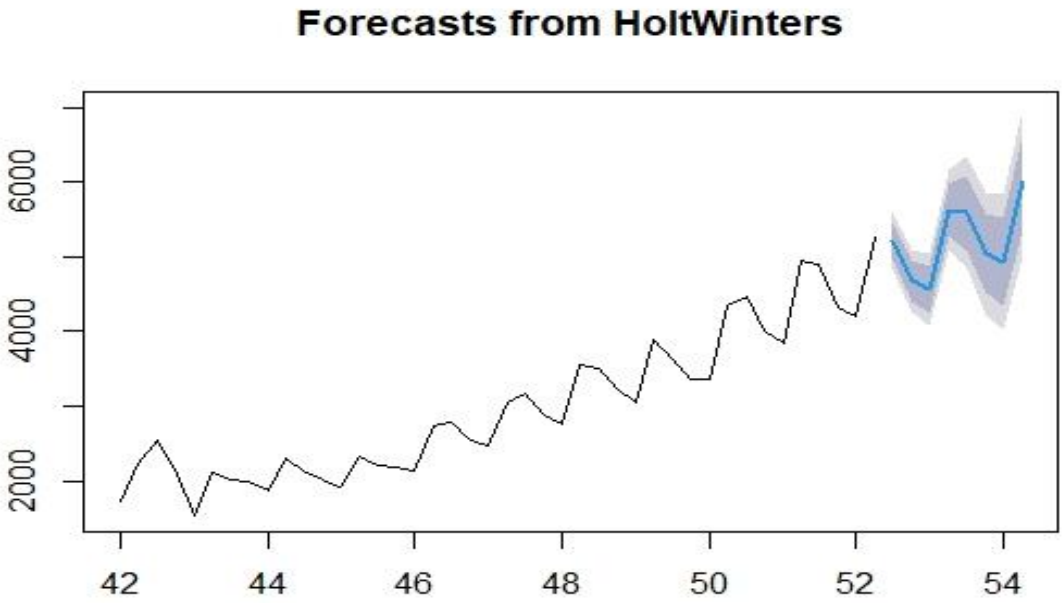


✓ The MAPE value for moving average is 8.908

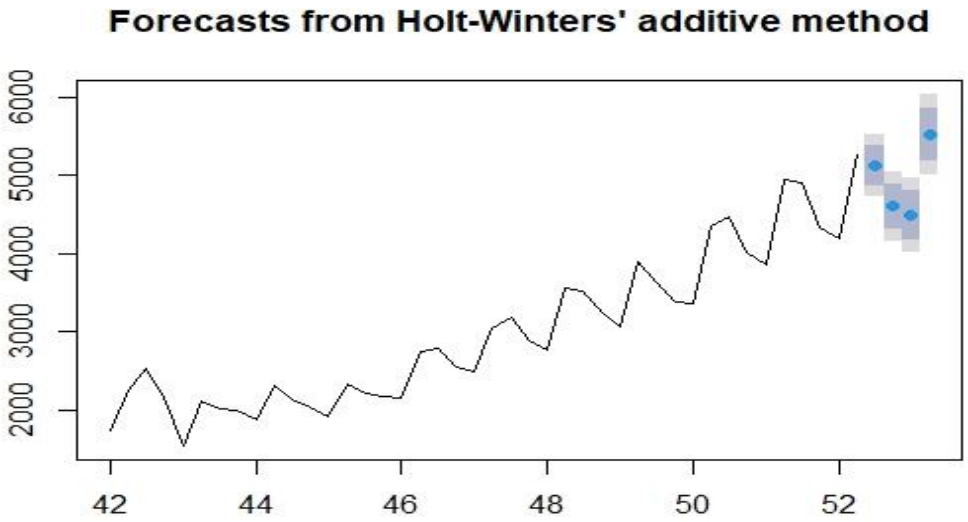
Using Holtwinters Function:

✓ Taking alpha value as 0.2 as a default and assuming time series data has only level parameter





Additive Graph:



- ✓ Based on the MAPE value who choose holts winter exponential technique which assumes the time series
- ✓ Data level, trend, seasonality characters with default values of alpha, beta and gamma

Hints:

1. Business Problem

    1.1. Objective
    1.2. Constraints (if any)

2. Data Pre-processing

    2.1 Feature Engineering, EDA etc.

3. Model Building

    3.1 Partition the dataset
    3.2 Model(s) – Work with all the models (linear, exponential, quadratic etc.)
    3.3 Model(s) Improvement steps
    3.4 Model Evaluation
    3.5 Python and R codes

4. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

Note:

1. For each assignment the solution should be submitted in the format

2. Research and Perform all possible steps for improving the model(s) accuracy & reduce the RMSE (also evaluate errors like MAPE, MAE etc.)

3. All the codes (executable programs) are running without errors

4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here