Time Series Analysis for Truck Sales

Project Report

Professor Zinovy Radovilsky

BAN 673 – 01 Time Series Analytics

Group Members:

Krutarth Sompura

Manan Upadhyay

Hemanth Varma Pericherla

Priyadarshini Madhusudanan

Vaishnavi Karingala

March 10, 2024

Executive Summary

The data used for this project is from a truck selling company. The data comprises the number of trucks sold for each month starting from the year 2003 to 2014. Using time series analytics techniques, we have forecasted the sale of trucks for 24 periods in the future (2015 and 2016). The data set indicates an upward linear trend with multiplicative seasonality. The number of trucks sold keeps increasing until the middle of each year, and by the end of the year, the sale of trucks decreases significantly. Also, the autocorrelation coefficient for all 12 lags is statistically significant, indicating high autocorrelation in the data.

To conduct the sales forecast, the following models have been applied:

1. Two-level forecasting model using regression model with linear trend and seasonality along with trailing moving average
2. Holt-Winter’s model
3. Quadratic trend and seasonality model
4. Auto ARIMA model

For each model the fit with training and validation partitions, and the fit with the entire data set have been analyzed. The accuracy measures for each model have also been compared. Comparing Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), the best model for forecasting truck sales for 24 periods in future (2015 to 2016) has been determined.

Based on this analysis, the best model for forecasting truck sales for the period of January 2015 to December 2016 is Holt-Winter’s model.

Introduction

This dataset, procured from Kaggle, meticulously details the truck sales of a specific company from January 2003 to December 2014. Comprising monthly records, the dataset is strategically curated for time series analysis, facilitating a comprehensive exploration of temporal patterns and trends in truck sales. Boasting 144 observations, it serves as a substantial foundation for in-depth examinations, providing a valuable resource for forecasting and understanding the intricacies of sales dynamics within the specified timeframe.

The dataset on truck sales from January 2003 to December 2014 offers invaluable insights for strategic decision-making. Analyzing temporal patterns allows the company to discern seasonal fluctuations, cyclical trends, and long-term sales patterns. Through rigorous time series analysis, questions regarding peak sales months or industry-specific events can be answered. Additionally, the dataset supports forecasting future sales, enabling proactive decision-making. Accurate projections empower the company to adjust production schedules, optimize inventory, and efficiently manage production planning. This strategic approach ensures market responsiveness, allowing swift adjustments to changing conditions influenced by factors such as fuel prices and regulations. Moreover, the dataset provides a competitive edge, enabling the company to introduce innovative models and strategically respond to market shifts. It serves as a baseline for risk mitigation, preparing contingency plans for economic shocks or supply chain disruptions. Ultimately, the dataset transforms decision-making from reactive to forward-thinking, guiding the company toward sustainable growth and operational excellence in the dynamic landscape of truck sales.

8 Steps for forecasting process

Step 1: Define goal

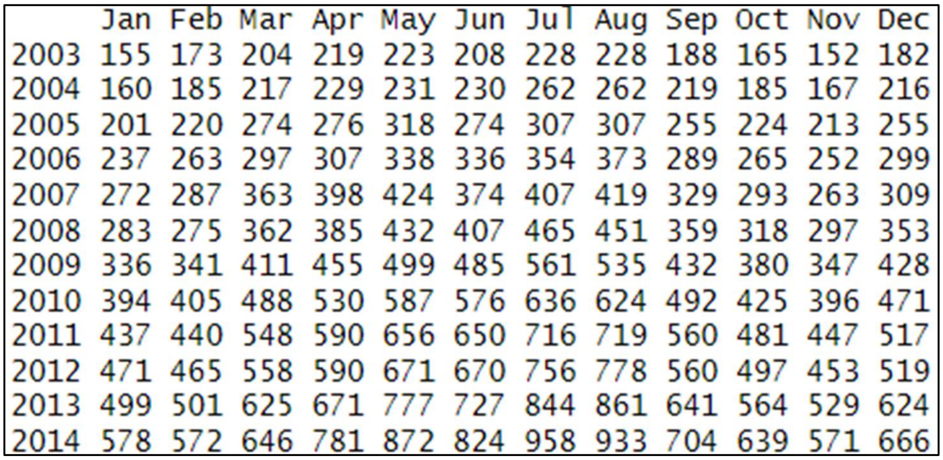
Our objective is to leverage historical data encompassing the monthly sales figures of trucks from January 2003 to December 2014. The aim is to forecast truck sales for a subsequent period spanning 24 months from 2015 to 2016. In this process, diverse forecasting models will be systematically tested, and forecasts will be generated and evaluated for accuracy using metrics such as RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error). The goal is to identify the most effective model for predicting truck sales accurately. This data-driven approach empowers the truck-selling company to make well-informed decisions, optimizing its sales strategies based on reliable forecasts and enhancing overall operational efficiency.

Step 2: Get data

The dataset, sourced from Kaggle, is meticulously designed for facilitating time series analysis. It is structured as a CSV file, featuring entries for the end of each month along with corresponding data on the number of trucks sold. Comprising a total of 144 monthly data points, the dataset spans from January 2003 to December 2014. To optimize model training and validation, the data is partitioned into distinct segments. The training partition encompasses the period from January 2003 to December 2012, while the validation partition spans from January 2013 to December 2014. Additionally, the dataset includes a future period from January 2015 to December 2016, providing a scope for forecasting and assessing model performance beyond the historical data. This thoughtful segmentation ensures robust analysis, effective model training, and reliable validation for comprehensive insights into the truck sales dynamics.

Step 3: Explore and visualize series

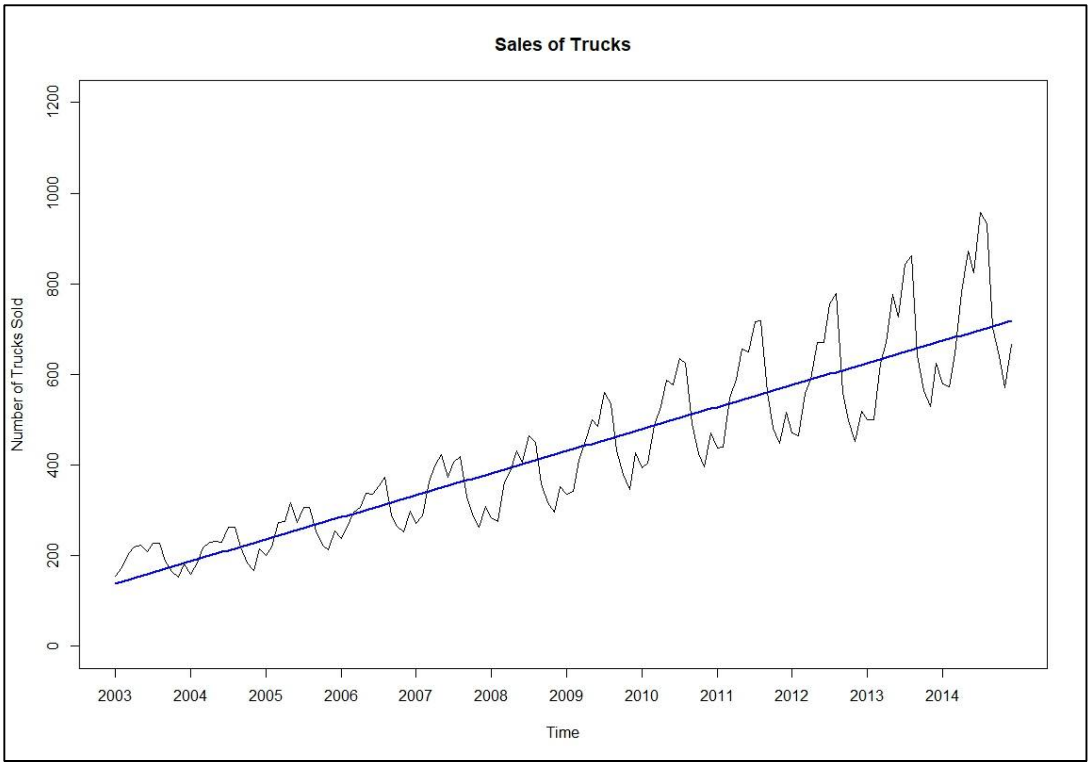
Time Series Data:



```
      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2003 155 173 204 219 223 208 228 228 188 165 152 182
2004 160 185 217 229 231 230 262 262 219 185 167 216
2005 201 220 274 276 318 274 307 307 255 224 213 255
2006 237 263 297 307 338 336 354 373 289 265 252 299
2007 272 287 363 398 424 374 407 419 329 293 263 309
2008 283 275 362 385 432 407 465 451 359 318 297 353
2009 336 341 411 455 499 485 561 535 432 380 347 428
2010 394 405 488 530 587 576 636 624 492 425 396 471
2011 437 440 548 590 656 650 716 719 560 481 447 517
2012 471 465 558 590 671 670 756 778 560 497 453 519
2013 499 501 625 671 777 727 844 861 641 564 529 624
2014 578 572 646 781 872 824 958 933 704 639 571 666
```

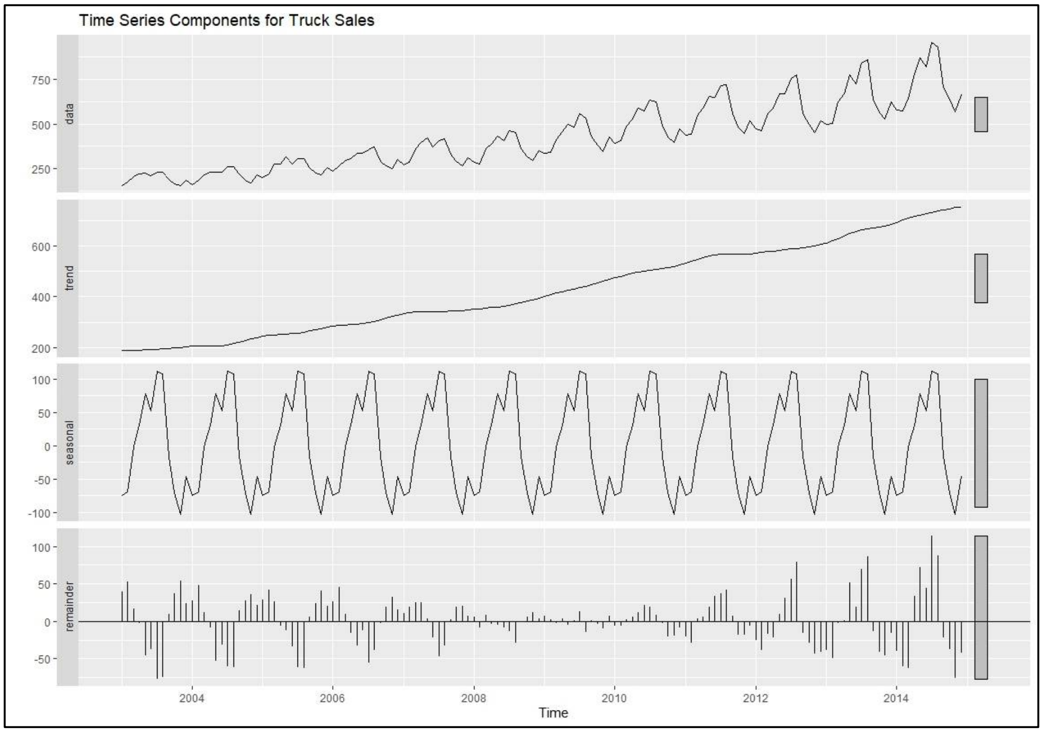
The truck sales data has been converted to time series data using the ts() function.

Visualizing the truck sales data:



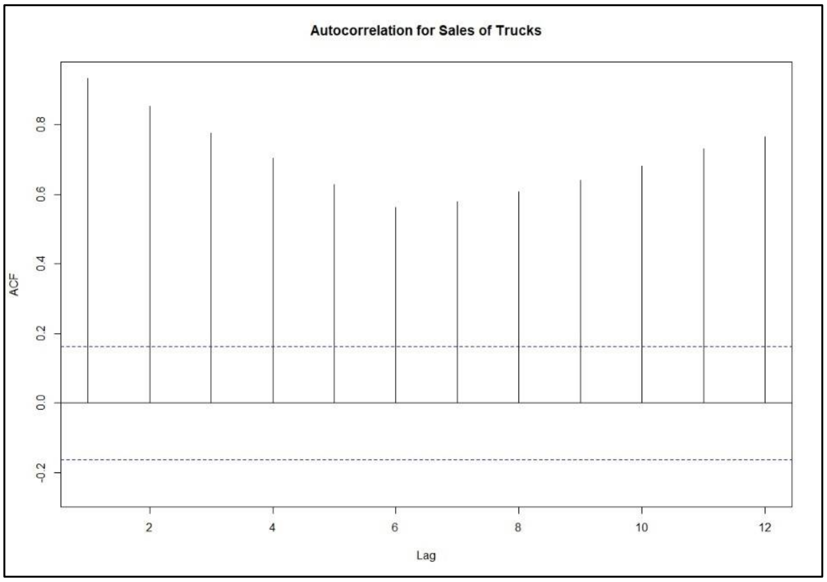
The plot above indicates an upward trend for the number of trucks sold from 2003 to 2014.

Time Series Components for the data:



The plot above shows the trend, seasonality, and level components of the data. It indicates an upward trend in the data and has a constant trend with additive seasonality. The sales of trucks start to increase each year, peak around the middle of the year, and then decrease significantly by the end of the year. This pattern repeats every year.

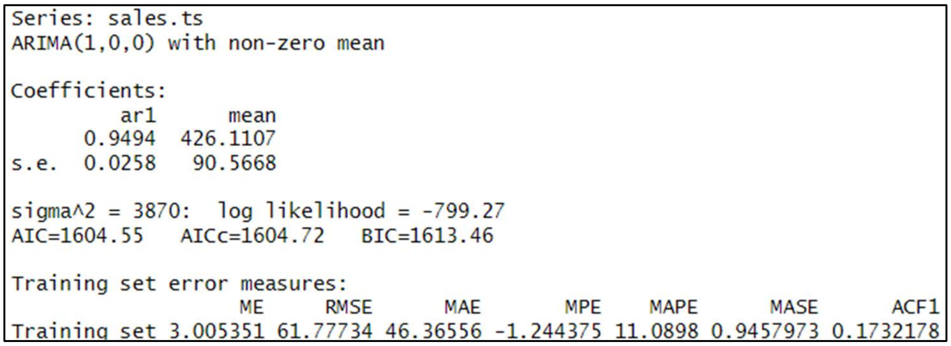
Autocorrelation for truck sales:



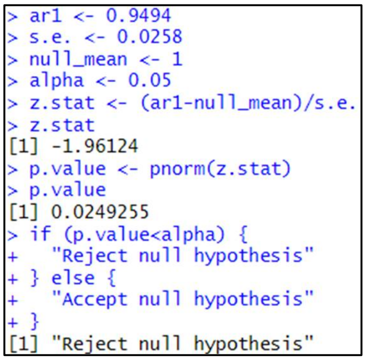
The analysis reveals a compelling pattern within the dataset, as all lags from 1 to 12 exhibit positive and statistically significant values. This pattern signifies a consistent upward trend in the data, suggesting a progressive increase over the monthly intervals. Notably, the positive and statistically significant autocorrelation at lag 12 further emphasizes the presence of monthly seasonality in the dataset. This implies that there is a recurring pattern or cycle every 12 months, aligning with a potential influence from external factors that cyclically impact the truck sales data. The collective evidence of positive and significant lags across this range strongly indicates a robust autocorrelation within the dataset, indicating a noteworthy degree of interdependence between successive monthly observations.

Predictability test:

Using AR1 approach



```
Series: sales.ts
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1      mean
      0.9494  426.1107
s.e.  0.0258   90.5668

sigma^2 = 3870:  log likelihood = -799.27
AIC=1604.55   AICc=1604.72   BIC=1613.46

Training set error measures:
                   ME     RMSE      MAE       MPE    MAPE      MASE      ACF1
Training set 3.005351 61.77734 46.36556 -1.244375 11.0898 0.9457973 0.1732178
```

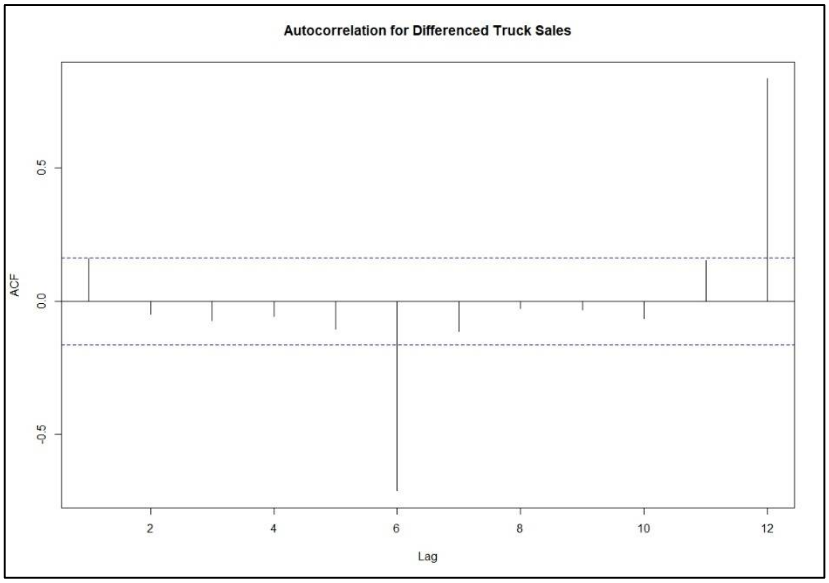


```
> ar1 <- 0.9494
> s.e. <- 0.0258
> null_mean <- 1
> alpha <- 0.05
> z.stat <- (ar1-null_mean)/s.e.
> z.stat
[1] -1.96124
> p.value <- pnorm(z.stat)
> p.value
[1] 0.0249255
> if (p.value<alpha) {
+   "Reject null hypothesis"
+ } else {
+   "Accept null hypothesis"
+ }
[1] "Reject null hypothesis"
```

Here, the z-statistic is -1.96124. The calculated p-value is 0.0249. Given a significance level (α) of 0.05, the decision criterion for rejecting the null hypothesis is when the p-value is less than α.

In this case, the obtained p-value (0.0249) is smaller than α (0.05), leading to the rejection of the null hypothesis. Hence, the data is not a random walk and is predictable.

First difference approach

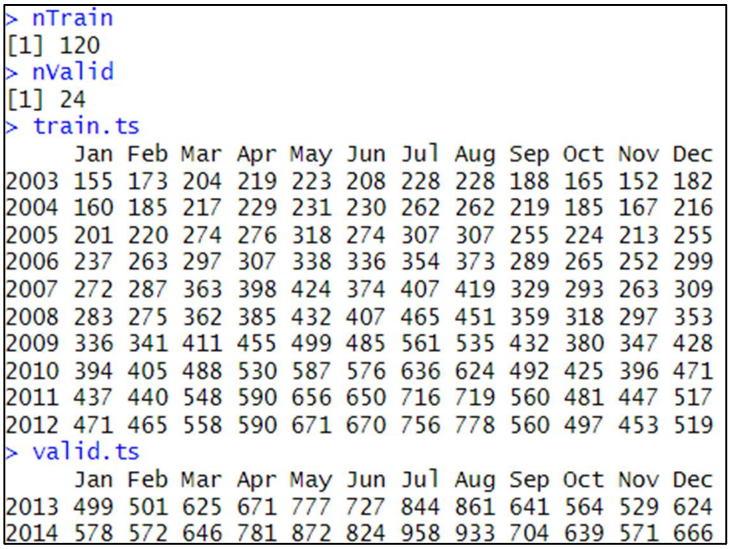


The presence of statistically significant autocorrelation at specific lags, particularly lag 12 indicates monthly seasonality, suggesting that the data exhibits a structured pattern. A random walk, characterized by a lack of pattern or predictability, is less likely when significant autocorrelation is observed. Hence, the data is not a random walk and is predictable.

Step 4: Data preprocessing

To ensure precision and consistency in time series forecasting, a partial subset of the original dataset was utilized for analysis. Specifically, sales data spanning from January 2003 to December 2014 was considered. As part of the data preparation procedure, the date format was adjusted to reflect the actual end-of-month dates in real-time, while retaining the corresponding sales figures for each respective month. This adjustment was crucial due to the original dataset dynamically altering the year for each month. The modification of the date format was undertaken to facilitate the accurate utilization of historical data, thereby enhancing the reliability of forecasts for the designated timeframe.

Step 5: Data partitioning



```
> nTrain
[1] 120
> nValid
[1] 24
> train.ts
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2003 155 173 204 219 223 208 228 228 188 165 152 182
2004 160 185 217 229 231 230 262 262 219 185 167 216
2005 201 220 274 276 318 274 307 307 255 224 213 255
2006 237 263 297 307 338 336 354 373 289 265 252 299
2007 272 287 363 398 424 374 407 419 329 293 263 309
2008 283 275 362 385 432 407 465 451 359 318 297 353
2009 336 341 411 455 499 485 561 535 432 380 347 428
2010 394 405 488 530 587 576 636 624 492 425 396 471
2011 437 440 548 590 656 650 716 719 560 481 447 517
2012 471 465 558 590 671 670 756 778 560 497 453 519
> valid.ts
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2013 499 501 625 671 777 727 844 861 641 564 529 624
2014 578 572 646 781 872 824 958 933 704 639 571 666
```

Prior to commencing the forecasting procedure, it is essential to partition the data into distinct training and validation periods. In this analysis, the dataset has undergone such division. Out of the total 144 data points, 120 points (representing approximately 85% of the dataset) have been assigned to the training partition, while the remaining 24 data points (constituting 17% of the dataset) have been designated for the validation partition.

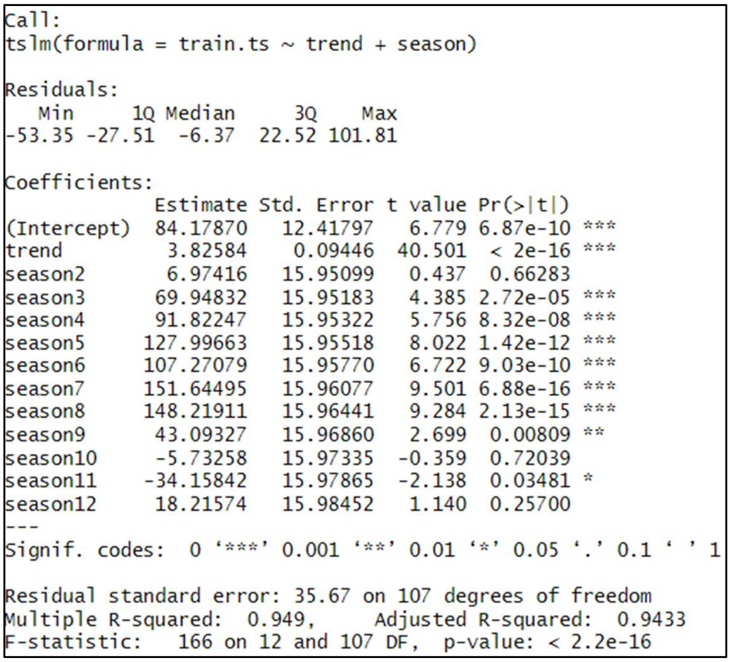
Step 6: Applying forecasting models

Model 1: Two-level forecasting using regression model with linear trend and seasonality along with trailing moving average (k = 4)

Two level forecasting model with linear trend and seasonality and trailing MA forecast with a window width of 4 is a combination of levels of models to produce a forecast. It is a time series forecasting technique with linear trend and seasonality and trailing MA window width 4 for residuals.

The first level of the model uses linear trend and seasonality models to estimate the underlying patterns in the time series. The trend component captures any long-term changes in the data, while the seasonality component captures any recurring patterns that repeat over a fixed period (monthly frequency). The second level of the model: forecasting the residuals, which are the differences between the actual values and the level 1 forecast.

Below is the summary for the regression model with linear trend and seasonality generated on training data:



```
Call:
tslm(formula = train.ts ~ trend + season)

Residuals:
   Min     1Q Median     3Q    Max 
-53.35 -27.51  -6.37  22.52 101.81 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  84.17870   12.41797   6.779 6.87e-10 ***
trend         3.82584    0.09446  40.501  < 2e-16 ***
season2       6.97416   15.95099   0.437  0.66283    
season3      69.94832   15.95183   4.385 2.72e-05 ***
season4      91.82247   15.95322   5.756 8.32e-08 ***
season5     127.99663   15.95518   8.022 1.42e-12 ***
season6     107.27079   15.95770   6.722 9.03e-10 ***
season7     151.64495   15.96077   9.501 6.88e-16 ***
season8     148.21911   15.96441   9.284 2.13e-15 ***
season9      43.09327   15.96860   2.699  0.00809 ** 
season10     -5.73258   15.97335  -0.359  0.72039    
season11    -34.15842   15.97865  -2.138  0.03481 *  
season12     18.21574   15.98452   1.140  0.25700    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.67 on 107 degrees of freedom
Multiple R-squared:  0.949,	Adjusted R-squared:  0.9433 
F-statistic:   166 on 12 and 107 DF,  p-value: < 2.2e-16
```

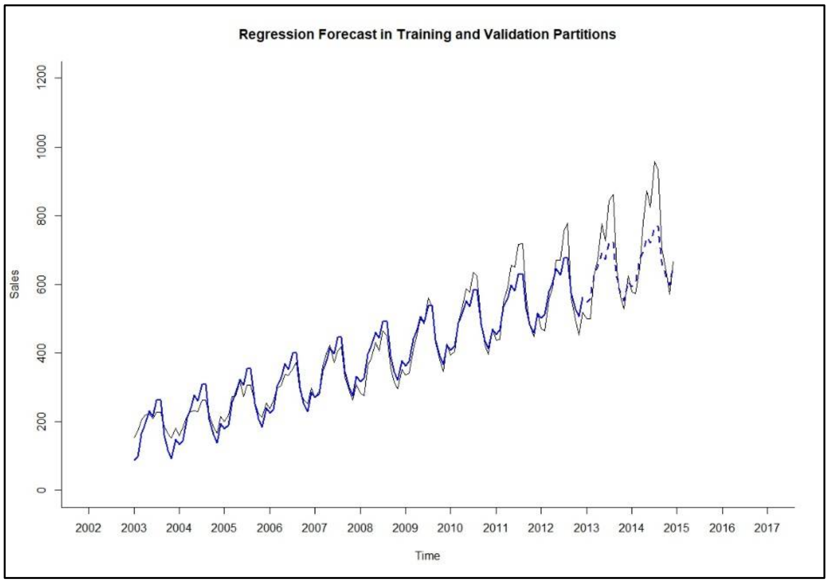
The equation for the above model is:

yt = 84.18 + 3.82 t + 6.97 D_2 + 69.95 D_3 + ……. + 18.22 D_{12}

where t is the time period, y_t is the output variable and D2, D3…D12 indicate binary variables for February, March…December respectively. In case all of these D2 – D12 are 0, it indicates January.

Since the p-values of coefficients for season 2, season 10 and season 12 are significantly larger than 0.05, some of the binary variable coefficients are not statistically significant. However, all the numerical variable coefficients are statistically significant with a p value lower than 0.05 and the adjusted R2 value (0.94) is also high. Thus, this model is a good fit and may be used for forecasting sales.

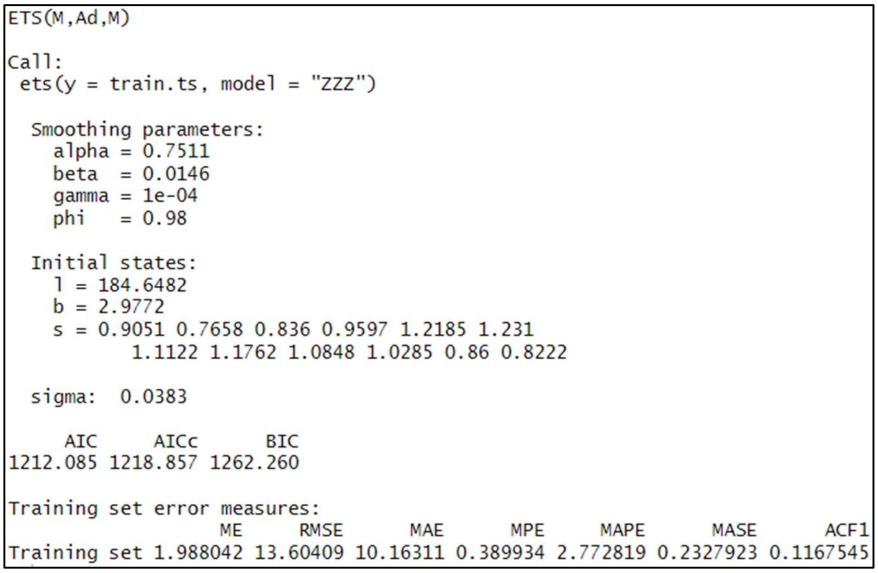
Below is the plot for the model in training and validation data:



We then find the regression residuals and run a trailing MA model with window width of 4 to forecast the residuals. This incorporates autocorrelation of residuals, if any.

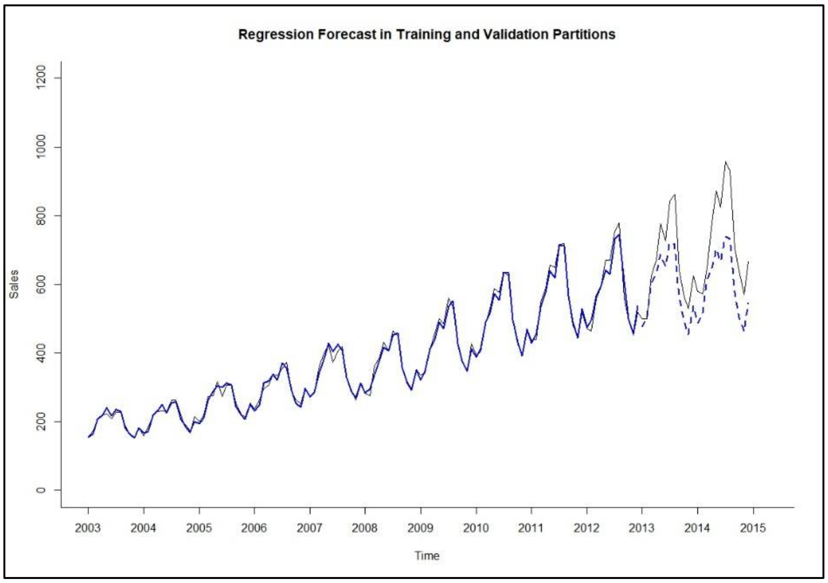
Model 2: Holt-Winter's Model

Holt-Winter's model is a widely used time series forecasting method designed for short-term predictions. An extension of the exponential smoothing model, it adeptly manages data with both trends and seasonality. This model comprises three key components: the level (or intercept), trend, and seasonality, each assigned weight updated according to new observations. Below is a concise summary of the Holt-Winter's Model, including options for error, trend, and seasonality.



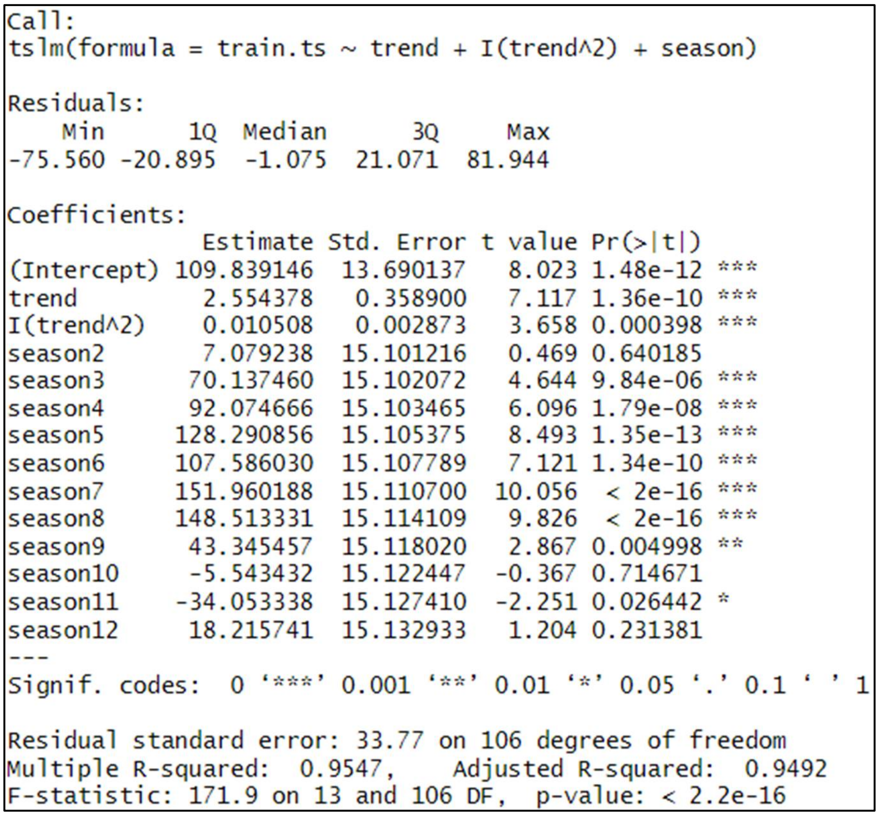
```
ETS(M,Ad,M)

Call:
 ets(y = train.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.7511
    beta  = 0.0146
    gamma = 1e-04
    phi   = 0.98

  Initial states:
    l = 184.6482
    b = 2.9772
    s = 0.9051 0.7658 0.836 0.9597 1.2185 1.231
           1.1122 1.1762 1.0848 1.0285 0.86 0.8222

  sigma:  0.0383

     AIC     AICc      BIC
1212.085 1218.857 1262.260

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE      ACF1
Training set 1.988042 13.60409 10.16311 0.389934 2.772819 0.2327923 0.1167545
```

The model has produced an ETS(M,Ad,M) configuration, representing multiplicative error, additive trend with damping effect, and multiplicative seasonality. The automated best-fit parameters yielded by the model are alpha = 0.7511, beta = 0.0146, gamma = 0.00001 and phi= 0.98. Below is a plot illustrating the model's good fit to both the training and validation partitions, encompassing historical data.



Model 3: Quadratic trend and seasonality

The quadratic trend and seasonality model is a widely used time series forecasting technique that combines a quadratic trend with a seasonal component. It is particularly effective for forecasting time series data with both a long-term trend and recurring seasonal patterns. Below is a summary of the regression model with a quadratic trend and seasonality.



```
Call:
tslm(formula = train.ts ~ trend + I(trend^2) + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-75.560 -20.895  -1.075  21.071  81.944 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 109.839146  13.690137   8.023 1.48e-12 ***
trend         2.554378   0.358900   7.117 1.36e-10 ***
I(trend^2)    0.010508   0.002873   3.658 0.000398 ***
season2       7.079238  15.101216   0.469 0.640185    
season3      70.137460  15.102072   4.644 9.84e-06 ***
season4      92.074666  15.103465   6.096 1.79e-08 ***
season5     128.290856  15.105375   8.493 1.35e-13 ***
season6     107.586030  15.107789   7.121 1.34e-10 ***
season7     151.960188  15.110700  10.056  < 2e-16 ***
season8     148.513331  15.114109   9.826  < 2e-16 ***
season9      43.345457  15.118020   2.867 0.004998 ** 
season10     -5.543432  15.122447  -0.367 0.714671    
season11    -34.053338  15.127410  -2.251 0.026442 *  
season12     18.215741  15.132933   1.204 0.231381    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.77 on 106 degrees of freedom
Multiple R-squared:  0.9547,	Adjusted R-squared:  0.9492 
F-statistic: 171.9 on 13 and 106 DF,  p-value: < 2.2e-16
```

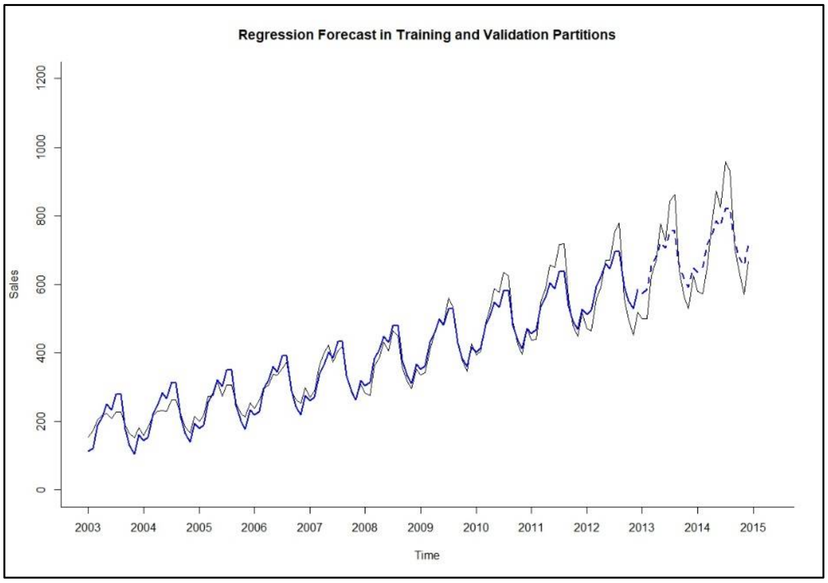
The model equation for quadratic trend and seasonality regression model is:

$yt = 109.84 + 2.55\ t + 0.01t^2 + 7.08\ D_2 + 70.14\ D_3 + \text{……….} + 18.22\ D_{12}$

where t is the time period, yt is the output variable and D2, D3 …D12 indicate binary variables for February, March…December. In case all these D2 – D12 are 0, it indicates January.

The adjusted R2 value is very high and except season 2, 10 and 12, all the coefficients are statistically significant and hence this model is a good fit and can be used for forecasting.

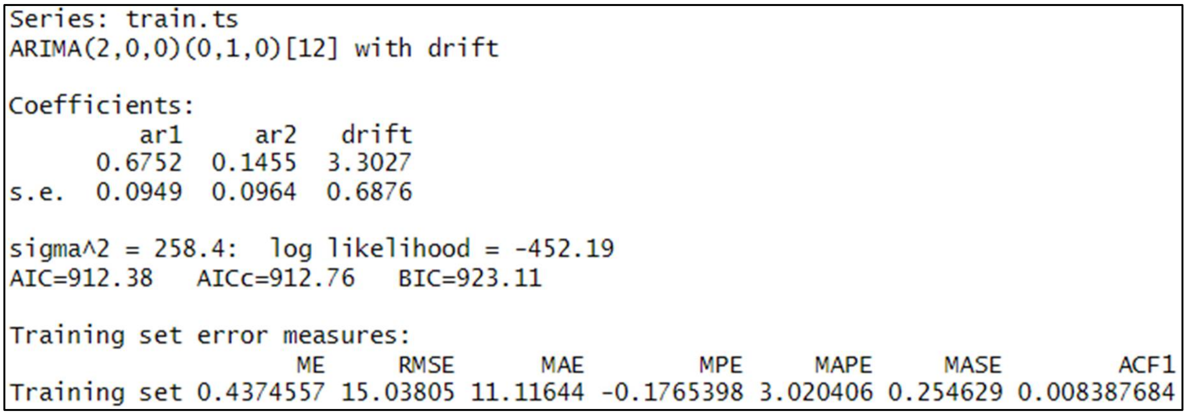
The plot below shows that the model fits well into the training and validation partition:



Model 4: Auto ARIMA

The Auto ARIMA model is an automated variant of the ARIMA (Autoregressive Integrated Moving Average) model, a widely used technique for time series forecasting. It employs a computational algorithm to automatically determine the optimal parameters for the ARIMA model based on the provided data.

Below is the summary of Auto ARIMA model using the training data set:



```
Series: train.ts
ARIMA(2,0,0)(0,1,0)[12] with drift

Coefficients:
         ar1     ar2   drift
      0.6752  0.1455  3.3027
s.e.  0.0949  0.0964  0.6876

sigma^2 = 258.4:  log likelihood = -452.19
AIC=912.38   AICc=912.76   BIC=923.11

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE        ACF1
Training set 0.4374557 15.03805 11.11644 -0.1765398 3.020406 0.254629 0.008387684
```

As can be seen from the above summary of auto-ARIMA model, we obtain the parameters in the form: ARIMA(2,0,0)(0,1,0)[12]

ARIMA(2,0,0)(0,1,0)[12]is in the form ARIMA(p,d,q)(P,D,Q)[m] and it denotes that (p,d,q) are the non-seasonal component parameters and the seasonal component parameters are (P,D,Q). The model parameters are explained below:

p = 2, order 2 autoregressive model for non-seasonal component

d = 0, no differencing to remove linear trend

q = 0, no moving average terms for error lags

P = 0, no autoregressive part for seasonal component

D = 1, order 1 differencing to remove linear trend for seasonal component

Q = 0, no moving average terms for seasonal component

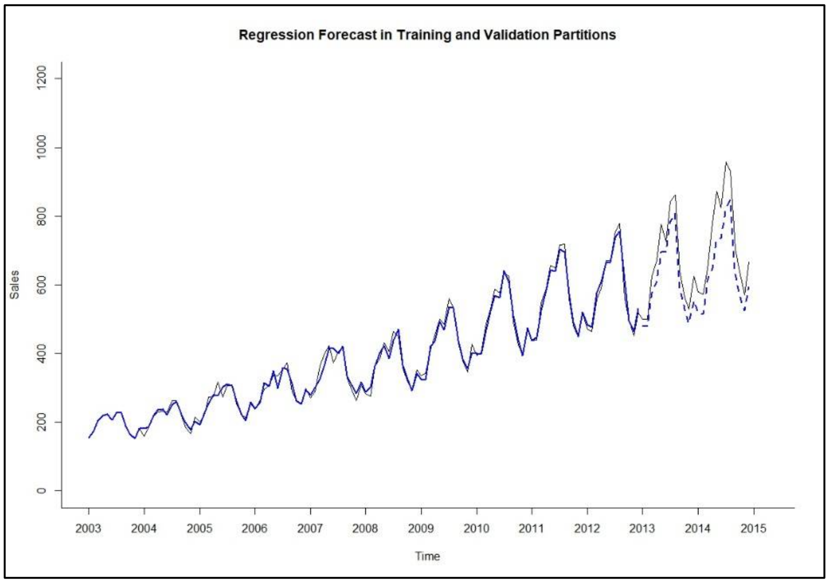
m = 12, for the number of seasons / monthly seasonality

The model equation is:

$$\mathbf{y_t - y_{t-1} = 3.3027 + 0.6752\ (y_{t-1} - y_{t-2}) + 0.1455\ (y_{t-2} - y_{t-3})}$$

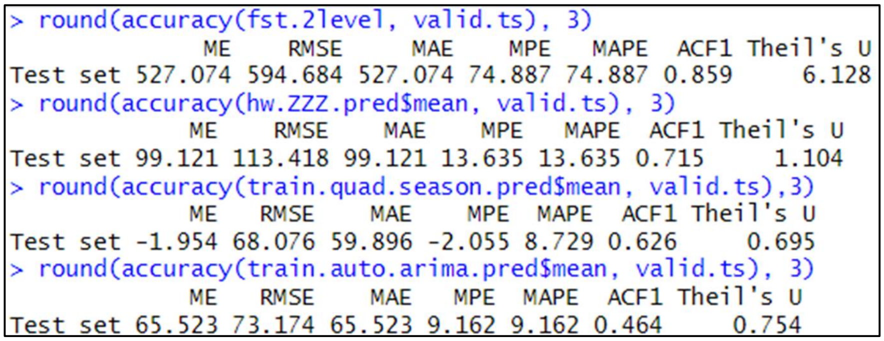
where yt is the value of series at time period t

Below is the plot for the training and validation partition which shows that the model fits well into the historical data.



Step 7: Comparing performance of validation data

In this step, we compare the accuracies of all the 4 models together to understand which model performs better in validation data:



```
> round(accuracy(fst.2level, valid.ts), 3)
              ME    RMSE     MAE    MPE   MAPE  ACF1 Theil's U
Test set 527.074 594.684 527.074 74.887 74.887 0.859     6.128
> round(accuracy(hw.ZZZ.pred$mean, valid.ts), 3)
             ME    RMSE    MAE    MPE   MAPE  ACF1 Theil's U
Test set 99.121 113.418 99.121 13.635 13.635 0.715     1.104
> round(accuracy(train.quad.season.pred$mean, valid.ts),3)
             ME   RMSE    MAE    MPE  MAPE  ACF1 Theil's U
Test set -1.954 68.076 59.896 -2.055 8.729 0.626     0.695
> round(accuracy(train.auto.arima.pred$mean, valid.ts), 3)
             ME   RMSE    MAE   MPE  MAPE  ACF1 Theil's U
Test set 65.523 73.174 65.523 9.162 9.162 0.464     0.754
```

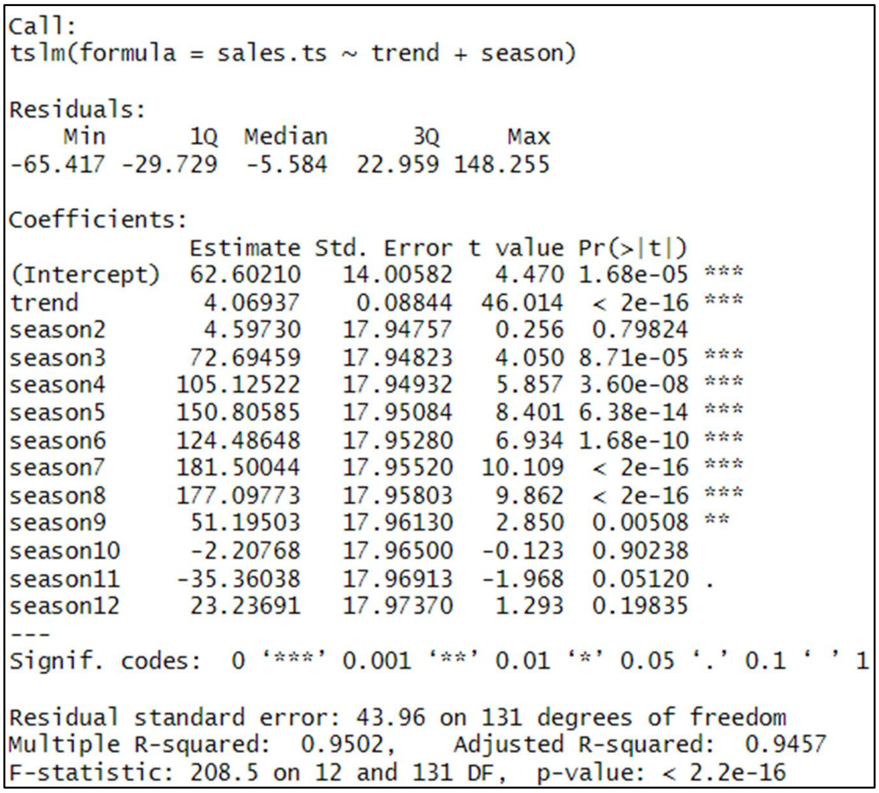
As shown above, from the MAPE and RMSE values, we can conclude that for the validation data, the Quadratic trend with seasonality model is the best model to forecast the sales.

Running models for the entire data set

To forecast for the future periods, we first need to combine the training data and the validation data and run the models again on the entire dataset. We then compare the performance of each of these models and depending on the accuracy of the models, we pick the best model for forecasting sales for the future 24 periods i.e., months of 2016 and 2017.

In the below sections, the models are re-run on the entire dataset.

Model 1: Two-level forecasting using regression model with linear trend and seasonality along with trailing moving average (k = 4)



```
Call:
tslm(formula = sales.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.417 -29.729  -5.584  22.959 148.255 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  62.60210   14.00582   4.470 1.68e-05 ***
trend         4.06937    0.08844  46.014  < 2e-16 ***
season2       4.59730   17.94757   0.256  0.79824    
season3      72.69459   17.94823   4.050 8.71e-05 ***
season4     105.12522   17.94932   5.857 3.60e-08 ***
season5     150.80585   17.95084   8.401 6.38e-14 ***
season6     124.48648   17.95280   6.934 1.68e-10 ***
season7     181.50044   17.95520  10.109  < 2e-16 ***
season8     177.09773   17.95803   9.862  < 2e-16 ***
season9      51.19503   17.96130   2.850  0.00508 ** 
season10     -2.20768   17.96500  -0.123  0.90238    
season11    -35.36038   17.96913  -1.968  0.05120 .  
season12     23.23691   17.97370   1.293  0.19835    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.96 on 131 degrees of freedom
Multiple R-squared:  0.9502,	Adjusted R-squared:  0.9457 
F-statistic: 208.5 on 12 and 131 DF,  p-value: < 2.2e-16
```

The equation for the above model is:

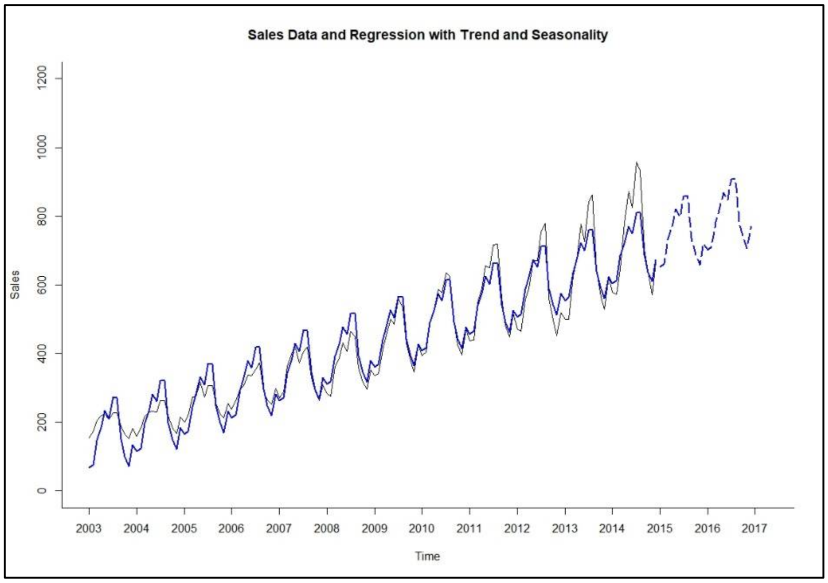
$y_t = 62.6 + 4.07\ t + 4.60\ D_2 + 72.69\ D_3 + \ldots\ldots + 23.24\ D_{12}$

where t is the time period, yt is the output variable and D2, D3…D12 indicate binary variables for February, March…December. In case all these D2 – D12 are 0, it indicates January.

The p-values of some of the seasons 2, 10, 11, 12 are greater than 0.05. However, all the numerical variable coefficients are statistically significant with a p value lower than 0.05 and the adjusted R^2 value (0.945) is also high. Thus, this model is a good fit and may be used for forecasting sales.

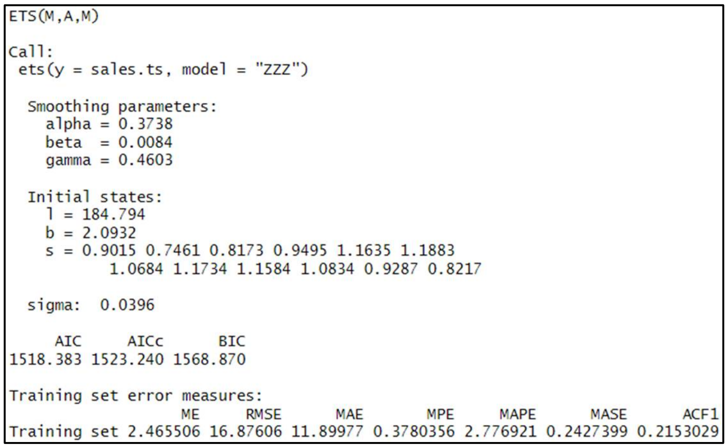
We then find the regression residuals and run a trailing MA model with window width of 4 to forecast the residuals. This incorporates autocorrelation of residuals, if any.

The plot below shows the historical data as well as the forecast for future periods using this model:



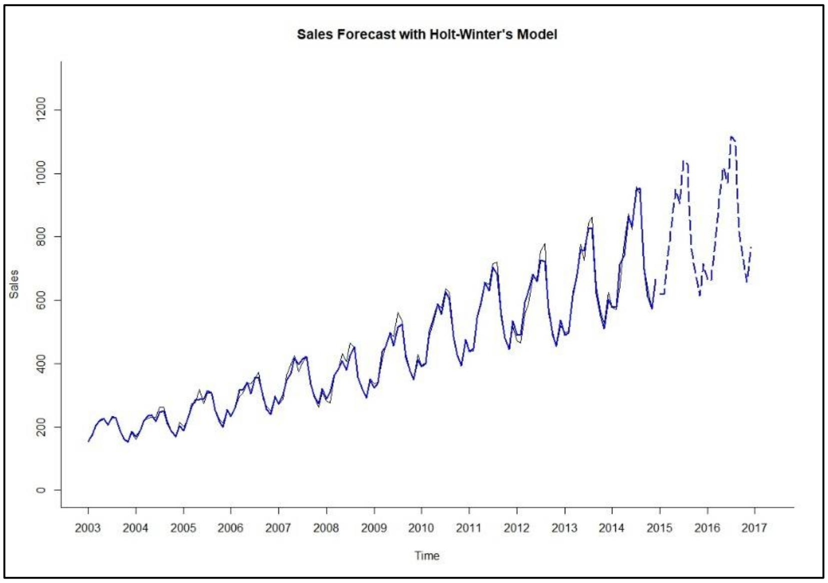
Model 2: Holt-Winter’s Model

Below is the summary for Holt Winter's Model with error, trend, and seasonality options for entire dataset:



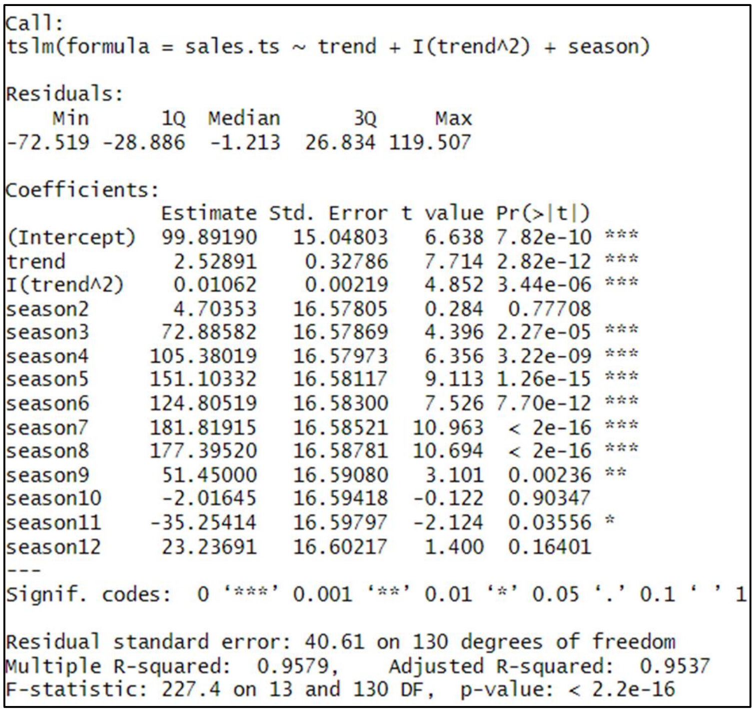
```
ETS(M,A,M)

Call:
 ets(y = sales.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.3738
    beta  = 0.0084
    gamma = 0.4603

  Initial states:
    l = 184.794
    b = 2.0932
    s = 0.9015 0.7461 0.8173 0.9495 1.1635 1.1883
           1.0684 1.1734 1.1584 1.0834 0.9287 0.8217

  sigma:  0.0396

     AIC     AICc      BIC
1518.383 1523.240 1568.870

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 2.465506 16.87606 11.89977 0.3780356 2.776921 0.2427399 0.2153029
```

The model has generated M,A,M model which denotes multiplicative error, additive trend, and multiplicative seasonality. The automated best fit values generated by the model are alpha = 0.3738, beta = 0.0084, gamma = 0.4603. The plot for the same in the historical and future periods is shown below:



Model 3: Quadratic trend and seasonality

Below is the summary of regression model with quadratic trend and seasonality for entire dataset:



```
Call:
tslm(formula = sales.ts ~ trend + I(trend^2) + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-72.519 -28.886  -1.213  26.834 119.507 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  99.89190   15.04803   6.638 7.82e-10 ***
trend         2.52891    0.32786   7.714 2.82e-12 ***
I(trend^2)    0.01062    0.00219   4.852 3.44e-06 ***
season2       4.70353   16.57805   0.284  0.77708    
season3      72.88582   16.57869   4.396 2.27e-05 ***
season4     105.38019   16.57973   6.356 3.22e-09 ***
season5     151.10332   16.58117   9.113 1.26e-15 ***
season6     124.80519   16.58300   7.526 7.70e-12 ***
season7     181.81915   16.58521  10.963  < 2e-16 ***
season8     177.39520   16.58781  10.694  < 2e-16 ***
season9      51.45000   16.59080   3.101  0.00236 ** 
season10     -2.01645   16.59418  -0.122  0.90347    
season11    -35.25414   16.59797  -2.124  0.03556 *  
season12     23.23691   16.60217   1.400  0.16401    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.61 on 130 degrees of freedom
Multiple R-squared:  0.9579,	Adjusted R-squared:  0.9537 
F-statistic: 227.4 on 13 and 130 DF,  p-value: < 2.2e-16
```

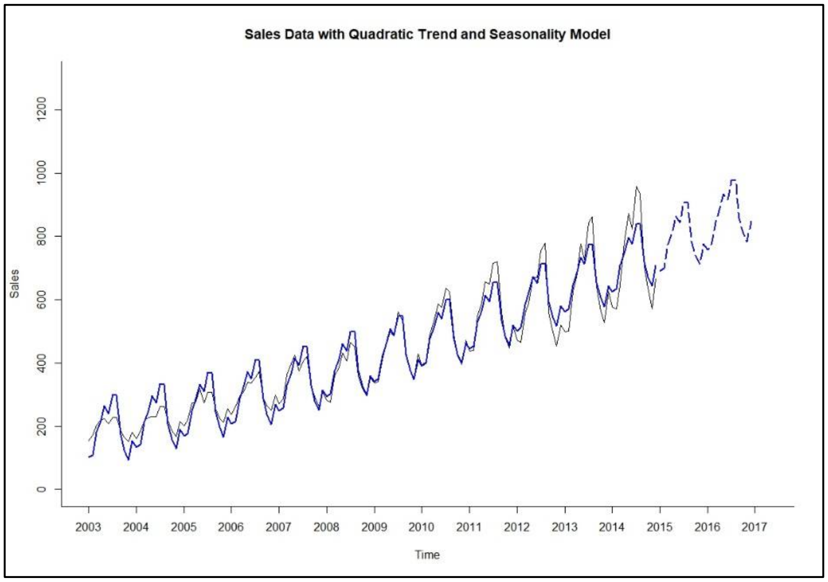
The model equation for quadratic trend and seasonality regression model is:

$\mathbf{y_t = 99.89 + 2.52\ t + 0.0106\ t^2 + 4.7\ D_2 + 72.89\ D_3 + \ldots\ldots\ldots + 23.24\ D_{12}}$

where t is the time period, yt is the output variable and D2, D3 …D12 indicate binary variables for February, March…December. In case all these D2 – D12 are 0, it indicates January.

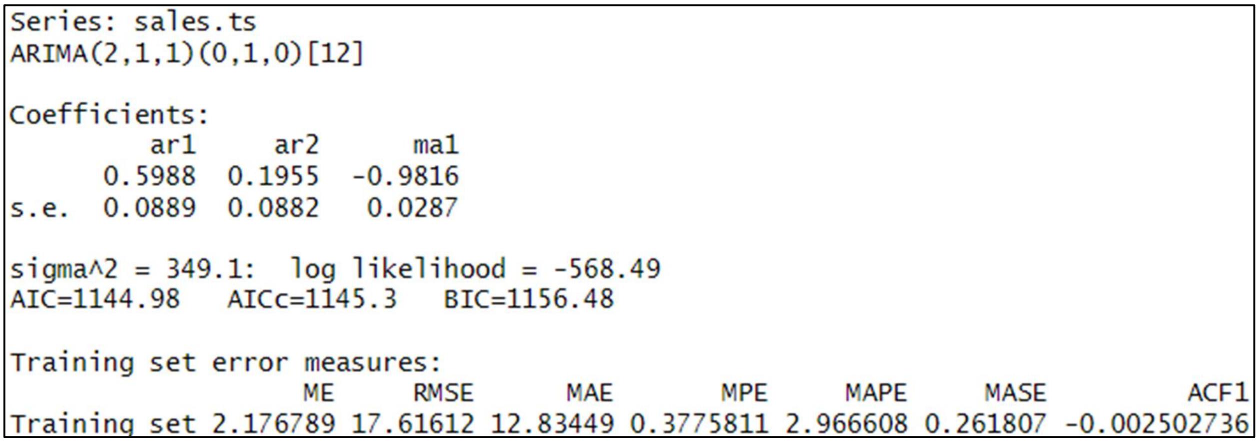
The adjusted R2 value is very high and except season 2,10 and 12, all the coefficients (especially numeric variable coefficients) are statistically significant and hence this model is a good fit and can be used for forecasting.

The plot below shows the historical data and future forecasts based on this model:



Model 4: Auto ARIMA

Below is the summary of Auto ARIMA model using the entire data set:



```
Series: sales.ts
ARIMA(2,1,1)(0,1,0)[12]

Coefficients:
         ar1     ar2      ma1
      0.5988  0.1955  -0.9816
s.e.  0.0889  0.0882   0.0287

sigma^2 = 349.1:  log likelihood = -568.49
AIC=1144.98   AICc=1145.3   BIC=1156.48

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE     MASE         ACF1
Training set 2.176789 17.61612 12.83449 0.3775811 2.966608 0.261807 -0.002502736
```

As can be seen from the above summary of auto-ARIMA model, we obtain the parameters in the form: ARIMA(2,1,1)(0,1,0)[12]

ARIMA(2,1,1)(0,1,0)[12] is in the form ARIMA(p,d,q)(P,D,Q)[m] and it denotes that (p,d,q) are the non-seasonal component parameters and the seasonal component parameters are (P,D,Q). The model parameters are explained below:

p = 2, order 2 autoregressive model for non-seasonal component

d = 1, order 1 differencing to remove linear trend

q = 1, order 1 moving average terms for error lags

P = 0, no autoregressive part for seasonal component

D = 1, order 1 differencing to remove linear trend for seasonal component

Q = 0, no moving average terms for seasonal component

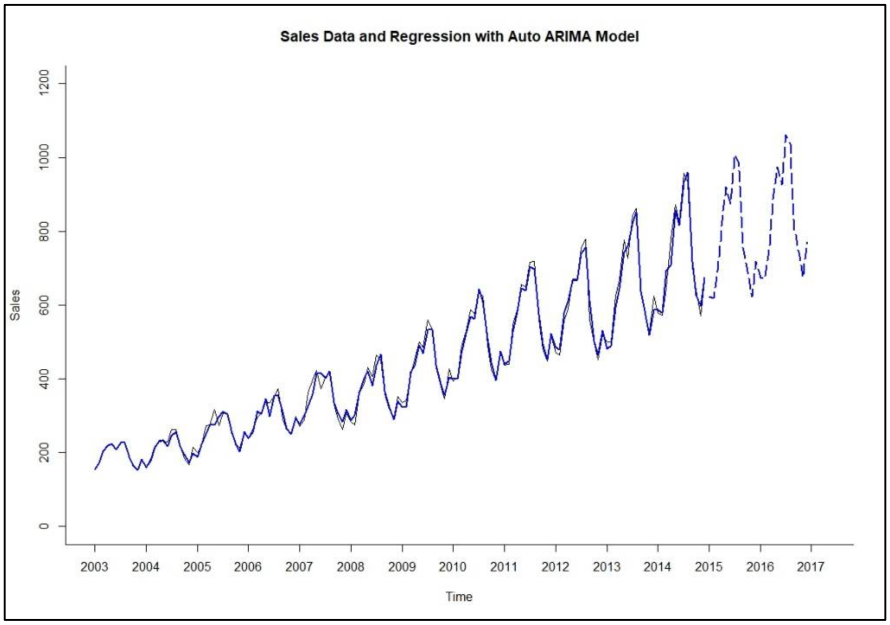
m = 12, for the number of seasons / monthly seasonality

The model equation is:

$y_t = 0.5988\ y_{t-1} + 0.1955\ y_{t-2} - 0.981\ \varepsilon_{t-1}$

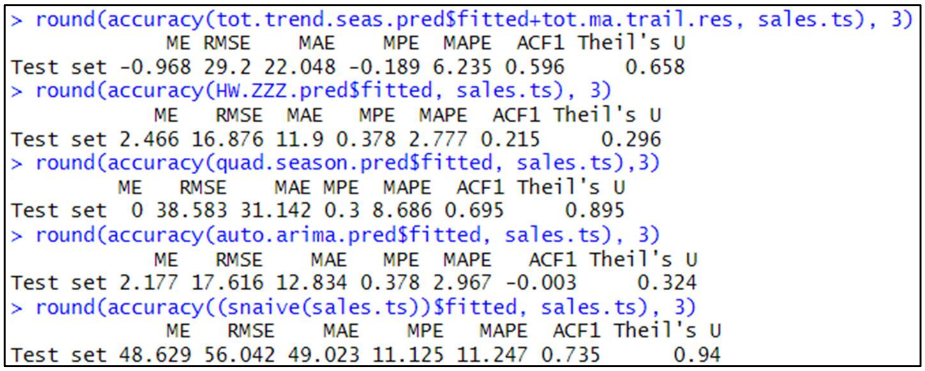
where yt is the value of series at time period t

Below is the plot for the training and validation partition which shows that the model fits well into the historical data.



Comparing performance on entire dataset:

We evaluate the accuracies of the previously executed models to identify the most effective one for predicting sales in 2015 and 2016. Additionally, we benchmark these models against a seasonal naïve baseline to determine whether the forecasting models offer better performance or if the baseline model suffices. Below are the accuracies assessed across the entire dataset for each model used:

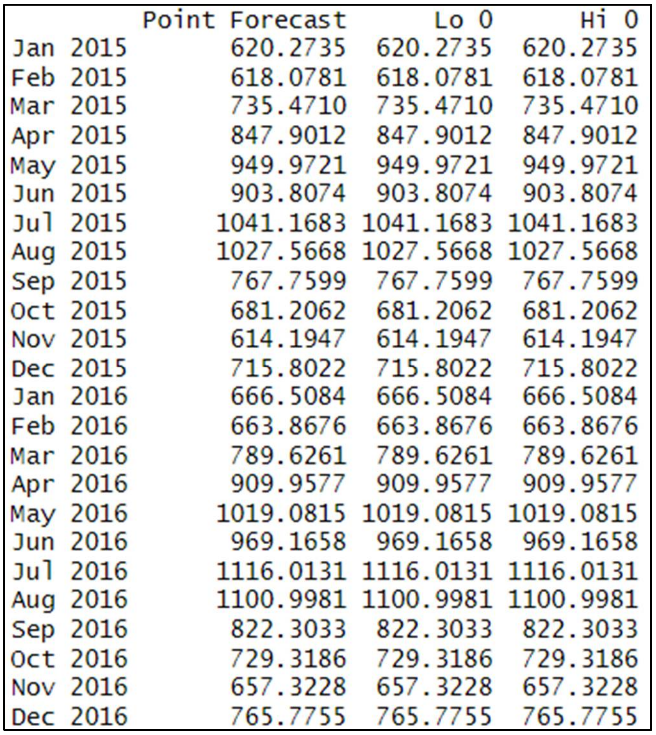


```
> round(accuracy(tot.trend.seas.pred$fitted+tot.ma.trail.res, sales.ts), 3)
             ME RMSE    MAE    MPE  MAPE  ACF1 Theil's U
Test set -0.968 29.2 22.048 -0.189 6.235 0.596     0.658
> round(accuracy(HW.ZZZ.pred$fitted, sales.ts), 3)
            ME   RMSE  MAE   MPE  MAPE  ACF1 Theil's U
Test set 2.466 16.876 11.9 0.378 2.777 0.215     0.296
> round(accuracy(quad.season.pred$fitted, sales.ts),3)
         ME   RMSE    MAE MPE  MAPE  ACF1 Theil's U
Test set  0 38.583 31.142 0.3 8.686 0.695     0.895
> round(accuracy(auto.arima.pred$fitted, sales.ts), 3)
            ME   RMSE    MAE   MPE  MAPE   ACF1 Theil's U
Test set 2.177 17.616 12.834 0.378 2.967 -0.003     0.324
> round(accuracy((snaive(sales.ts))$fitted, sales.ts), 3)
             ME   RMSE    MAE    MPE   MAPE  ACF1 Theil's U
Test set 48.629 56.042 49.023 11.125 11.247 0.735      0.94
```

As can be seen from the accuracies mentioned above, on the basis of MAPE and RMSE values, we get the best accuracy using Holt-Winter’s model(M,A,M) since the MAPE and RMSE values are the lowest for this.

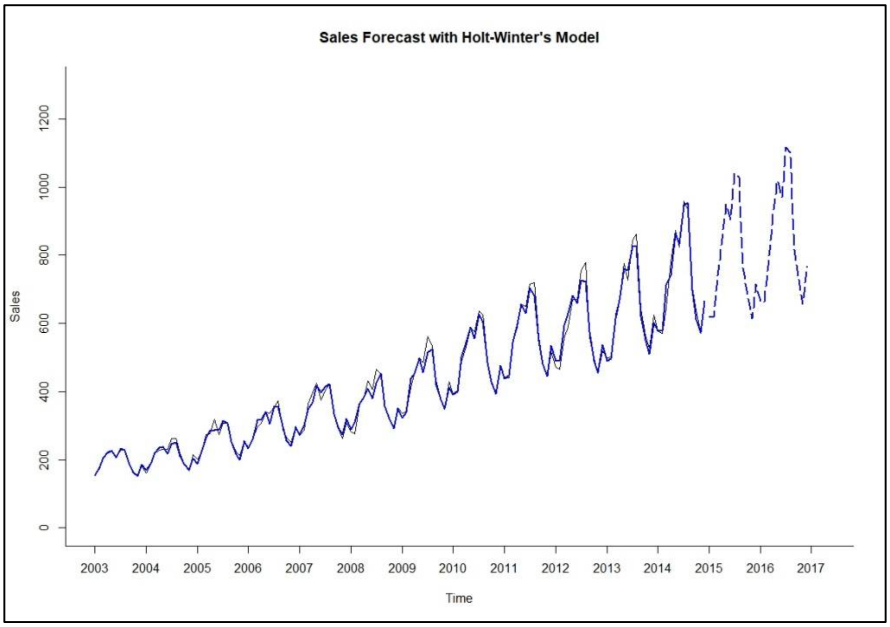
Step 8: Implementation

The best model with the lowest MAPE and RMSE is **Holt-Winter’s model**(M,A,M). On forecasting the Sales trends for the next two years (2015 and 2016) using HW-ZZZ, we obtain the below forecast:



```
         Point Forecast      Lo 0      Hi 0
Jan 2015       620.2735  620.2735  620.2735
Feb 2015       618.0781  618.0781  618.0781
Mar 2015       735.4710  735.4710  735.4710
Apr 2015       847.9012  847.9012  847.9012
May 2015       949.9721  949.9721  949.9721
Jun 2015       903.8074  903.8074  903.8074
Jul 2015      1041.1683 1041.1683 1041.1683
Aug 2015      1027.5668 1027.5668 1027.5668
Sep 2015       767.7599  767.7599  767.7599
Oct 2015       681.2062  681.2062  681.2062
Nov 2015       614.1947  614.1947  614.1947
Dec 2015       715.8022  715.8022  715.8022
Jan 2016       666.5084  666.5084  666.5084
Feb 2016       663.8676  663.8676  663.8676
Mar 2016       789.6261  789.6261  789.6261
Apr 2016       909.9577  909.9577  909.9577
May 2016      1019.0815 1019.0815 1019.0815
Jun 2016       969.1658  969.1658  969.1658
Jul 2016      1116.0131 1116.0131 1116.0131
Aug 2016      1100.9981 1100.9981 1100.9981
Sep 2016       822.3033  822.3033  822.3033
Oct 2016       729.3186  729.3186  729.3186
Nov 2016       657.3228  657.3228  657.3228
Dec 2016       765.7755  765.7755  765.7755
```

The plot for the above forecasts is shown below:



Conclusion

We applied four models to determine the best model to forecast the sale of trucks for the years 2015 and 2016. Thes models are: Two-level forecasting model using regression model with linear trend and seasonality along with trailing moving average, Holt-Winter's model, Quadratic trend and seasonality model, and Auto ARIMA model

Based on the RMSE and MAPE accuracy measures for the entire data set, we concluded that **Holt-Winter's model** is the most prominent model to apply for forecasting sales for 24 periods in future (January 2015 to December 2016).

References

https://www.kaggle.com/datasets/ddosad/dummy-truck-sales-for-time-series/data