FORECASTING SALES REVENUE FOR LIFE SCIENCES MANUFACTURING COMPANY IN US

Table of Contents

INTRODUCTION

BACKGROUND OF THE LIFE SCIENCES MANUFACTURING COMPANY

The Life sciences company from which we collected the data deals with a variety of products where there is extensive research in diagnostic tools which helps the customers in calculating scientific measurements. It also has dedicated teams working on Life sciences research and development in the fields of chemical analysis and food research.

BUSINESS PROBLEM

The company wants to ensure it is keeping reasonable forecasts so that their sales and marketing team can understand the targets to be achieved and to ensure the overall growth of the company stays positive. Due to different product portfolios it may not be reasonable to give a similar forecast for all the products where products related to Research incur more expenses so we have classified the products into 2 groups diagnostics and Research. In this project we focussed on diagnostics product group to forecast the revenue so that it can support the expenses internal research department of the company and also to maintain overall positive outlook of the company with relevant forecast targets and give the executives required information to plan the expenses and understand the projected revenue.

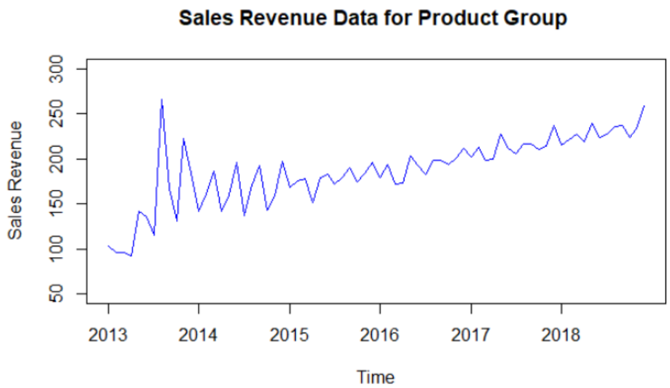
INSIDE THE DATA

In this project, we would like to extract the data from a well-known US based life sciences company, and we will mask (we are not relieving the exact name of the product group and hence giving it a generic name as product group) the data due to the data privacy concerns. This data contains monthly revenue information of 5 fiscal years of firm which starts from 1st January 2013 to 31st December 2018 and the revenue was measured in millions of USD. The analysts for their tracking purposes have decided to group the similar products into product groups and the goal is to identify the best forecasting method to predict monthly revenue for each product group for the following year 2019. For this project we are considering one of the product groups and we are forecasting the sales revenue for that product group.

EXPLORATORY DATA ANALYSIS

TIMES SERIES DATA PLOT

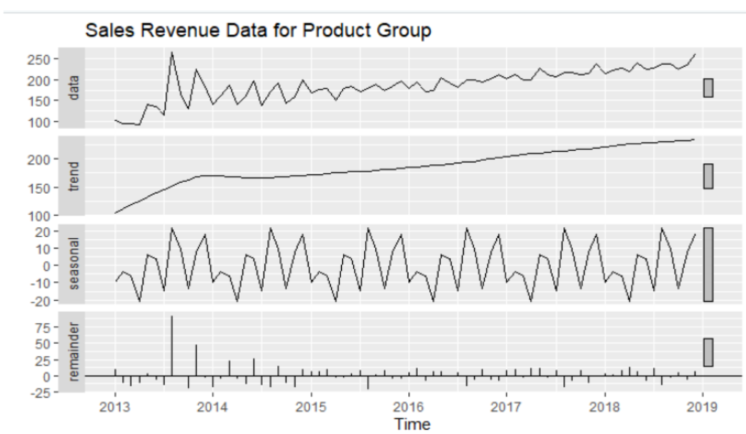
We started our data analysis by plotting the time series data using the plot() function in R to get a sense of how the data for Sales Revenue of the company for the product group.



From the time series data plot, we can see that the pattern is repeating itself every year. The peaks are growing towards the more recent periods suggesting a bit of upward trend in the data.

TIMES SERIES COMPONENTS

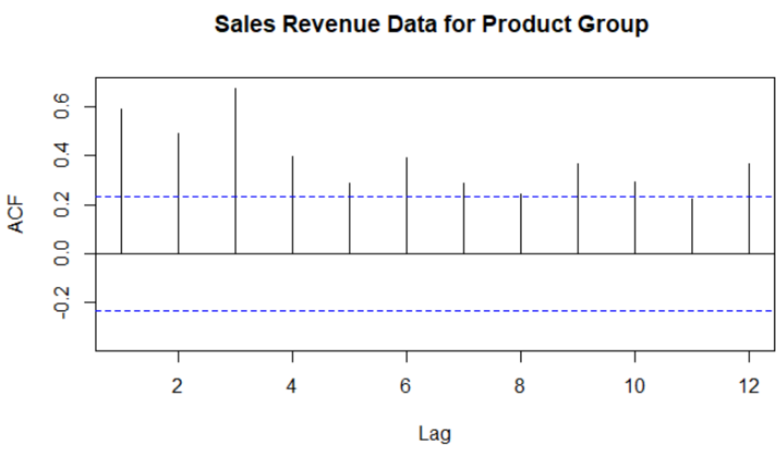
Every Time Series data comprises of systematic (level, trend and seasonality) and non-systematic components (noise). We used the stl() function to visualize these components for Sales Revenue Data.



From the plot above, in the trend component we can see that there is global upward trend and it increases along with the time period. The seasonal component shows that there is cyclic behavior in the data. These last component remainder is basically a combination of level and noise.

AUTOCORRELATION CHART

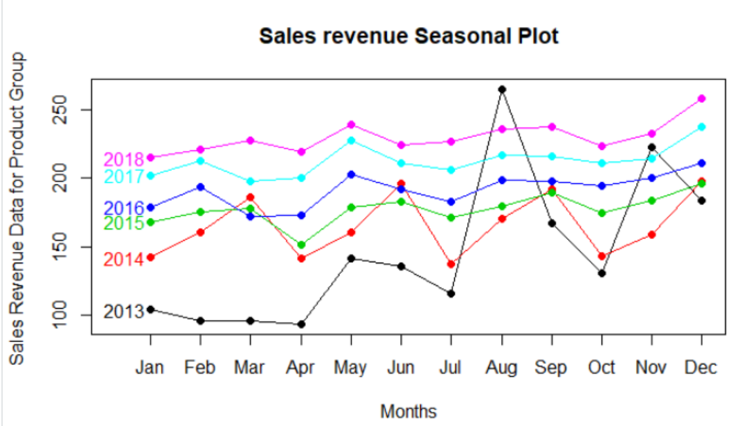
Autocorrelation represents the correlation between a random variable (time series data) itself and the same variable lagged one or more periods. We used the acf() function to plot the autocorrelation chart, to evaluate if the values in the neighboring periods are correlated.



From the correlogram it can be observed that there is positive autocorrelation coefficient from Lag 1 to lag 12 and all the lags have autocorrelation above significance threshold and only lag 11 is not beyond the significance threshold. This shows that the data has seasonality(lag 12) for monthly data.

SALES REVENUE SEASONAL CHART

The visualizations above indicated towards months seasonality in the data, therefore we plotted a seasonal chart for the data to see how Sales Revenue vary with the months.



The chart above shows that year after year there is a significant increase in the Sales Revenue data of the company. We can observe that in the first quarter the sales revenue gradually increases. The Last 3 months have shown an increase in Sales Revenue. There is generally an upward trend in the data model.

Testing Predictability Of Data

Before we attempt to forecast a time series data, it is important for us to know if the data is even predictable and subsequently forecastable. We need to know that the data is not a random walk and if forecasting whether the effort will be useful or not and should we even go beyond the Naïve forecast. In order to test the predictability of the Sales Revenue data, we have used two approaches: 1) Fit an AR(1) model to test the hypothesis that the slope coefficient β1 = 1 and 2) Examine differenced series: Y_2-Y_1, Y_3-Y_2, … , Y_T -Y_{T-1} which is the mathematical equivalent of approach 1.

Approach 1:

A partial output of the AR(1) model for Sales Revenue time series data is presented below. ARIMA(1, 0, 0) is an autoregressive (AR) model with order 1, no differencing, and no moving average model.

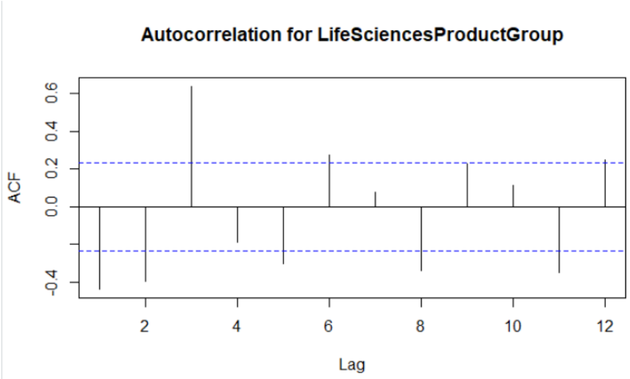
```
Series: LifeSciencesProductGroup1.ts
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1      mean
      0.6549  186.8438
s.e.  0.0959    9.7609

sigma^2 estimated as 884.3:  log likelihood=-345.68
AIC=697.36   AICc=697.72   BIC=704.19

Training set error measures:
                  ME     RMSE      MAE      MPE    MAPE      MASE       ACF1
Training set 1.03946 29.32094 20.59429 -2.38479 12.2956 0.8893311 -0.2086952
```

The *AR(1)* model equation is:

$\boldsymbol{Y_t = 186.8438 + 0.6549\ Y_{t-1}}$

The coefficient of the *ar1* (Y_{t-1}) variable, *0.6549*, is well below 1. With the confidence of 95%, the upper value of this coefficient (the population value of this coefficient) will be *0.6549* + 2*0.0959= 0.8467, which is still below 1. Therefore, *LifeSciencesProductGroup1.ts* time series is not a random walk and is predictable.

Approach 2:



In the chart above, several autocorrelation coefficients of the first differenced data are statistically significant as they are above significance threshold except for Lag 4,7 and 10. Therefore, using the first differencing approach, we can confirm that *LifeSciencesProductGroup1.ts* is not a random walk and is predictable.

Forecasting Methodology

Partitioning

Partition Series is an important preliminary step before applying any forecasting method. It comes from the need to be able to test how well any selected model performs with the new data not included in the model development. Therefore, we created a data partition of 60 records for training period which includes data points from January 2013 to December 2017. The data for the most recent 1 year i.e. from January 2018 to December 2018 was considered for validation period with a total of 12 records. We have built various forecasting models using the training data and measured its performance using the validation data.

Forecast Performance Baseline

A baseline in forecast performance provides a point of comparison. It is a point of reference for all other modeling techniques on the problem. If a model achieves performance at or below the baseline, the technique should be improved or abandoned completely. The most common baseline method for time series forecasting is Naïve Model approach.

NAÏVE & SEASONAL NAIVE FORECAST:

It is the simplest form of model. In this approach forecast for any period equals the previous period's actual value. Since, this model gives full weight to the last period original value it is not able to capture the features of data series.

For highly seasonal time series data, Seasonal Naïve Forecast can be a good baseline. Seasonal Naïve method is like the naive method but predicts the last observed value of the same season of the year.

Before diving into sophisticated algorithms, we evaluated the performance of both Naïve and Seasonal Naïve forecast to set a baseline and the results are as follows:

```
> # Comparing Accuracy for Seasonal Naive and Naive
> round(accuracy(LifeSciencesProductGroup1.naive.pred$fitted,LifeSciencesProductGroup1.ts) ,3)
            ME   RMSE    MAE   MPE   MAPE   ACF1 Theil's U
Test set 2.173 31.822 20.798 -0.25 11.478 -0.438         1
> round(accuracy(LifeSciencesProductGroup1.snaive.pred $fitted,LifeSciencesProductGroup1.ts) ,3)
             ME   RMSE    MAE   MPE   MAPE  ACF1 Theil's U
Test set 16.822 29.379 23.157 8.694 12.464 0.163     0.748
```

We can see that the MAPE and RMSE values for Naïve are substantially better than the Seasonal Naïve forecast. Therefore, we can say that it will be worthwhile to evaluate the performance of more sophisticated forecasting methods.

EXPONENTIAL SMOOTHING

Exponential Smoothing is a data-driven approach to perform time series forecasting. Exponential smoothing estimates time series components (level, trend, and seasonality) directly from the data without a predetermined structure. They smooth out the noise in a time series data to uncover the data patterns. We have performed advanced smoothing through Holt-Winters model that incorporates both trend and seasonality in the data.

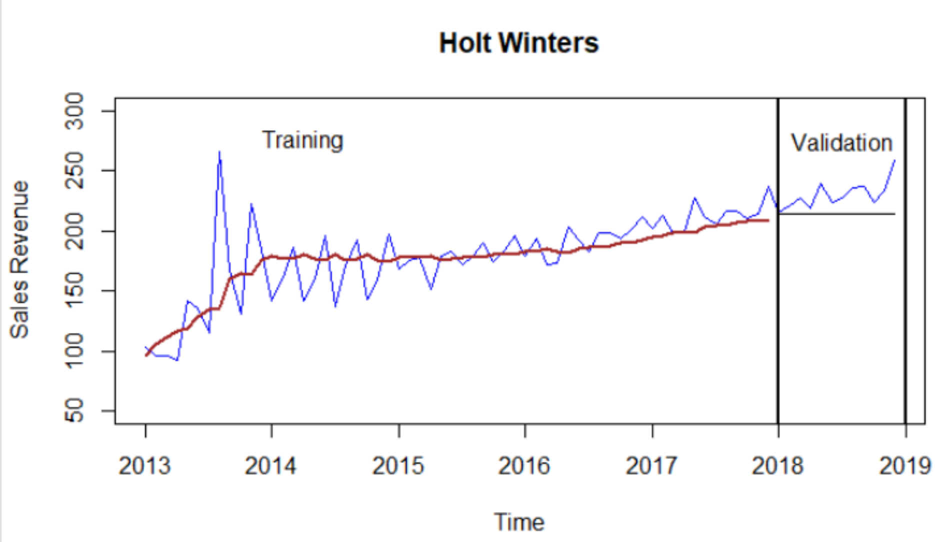
HW MODEL ON TRAINING DATASET:

Holt-Winters model for the training dataset along with the automated selection of the model options and optimal smoothing parameters:

```
ETS(A,Ad,N)

Call:
 ets(y = LifeSciencesProductGroup1.train.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.1494
    beta  = 1e-04
    phi   = 0.909

  Initial states:
    l = 87.5082
    b = 10.0178

  sigma:  25.8928

     AIC     AICc      BIC
642.9156 644.5005 655.4816
```

This Holt-Winters model has the (A,Ad,N) options, i.e., Additive error, Additive damped trend, and No seasonality. The optimal value for exponential smoothing constant (alpha) is 0.1494, The optimal smoothing constant for trend estimate (beta) is 0.0001, and there is no smoothing constant. The alpha value of this model indicates that the model's level component tends to be more global, while the trend is globally adjusted as beta is close to zero. The no gamma value indicates that no seasonal component is weighed in a forecast.

The Forecast plot for Holt-Winters model on the training and validation dataset is given below and the table is in the appendix.



HW MODEL ON HISTORICAL DATASET:

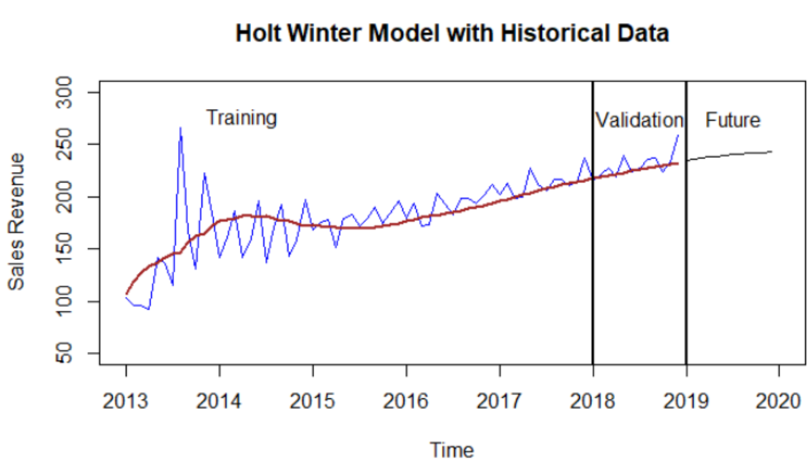
In order to forecast future values of the series using the Holt-Winters model, the training and validation periods were recombined into the entire (historical) time series dataset.

The summary for the Holt-Winters model for the entire dataset along with the automated selection of the model options and optimal smoothing parameters is given below:

```
ETS(A,Ad,N)

Call:
 ets(y = LifeSciencesProductGroup1.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.0338
    beta  = 0.0338
    phi   = 0.8701

  Initial states:
    l = 94.0164
    b = 15.4289

  sigma:  23.4771

     AIC     AICc      BIC
769.2055 770.4978 782.8655
```

This Holt-Winters model has the (A,Ad,N) options, i.e., Additive error, Additive damped trend, and No seasonality. The optimal value for exponential smoothing constant (alpha) is 0.0338, The optimal smoothing constant for trend estimate (beta) is 0.0338, and there is no smoothing constant. The alpha value of this model indicates that the model's level component tends to be more global and the trend is also global. The no gamma value indicates that no seasonal component is weighed in a forecast.

The Forecast plot for 12 periods into the future using the Holt-Winters model on the entire dataset is given below and the table is in the appendix.



Accuracy Measures :

```
> # Accuracy for Holt-Winter's model for the validation period.
> round(accuracy(HW.ZZZ.Pred$mean,LifeSciencesProductGroup1.valid.ts),3)
            ME   RMSE   MAE   MPE  MAPE ACF1 Theil's U
Test set 16.31 19.748 16.31 6.881 6.881 0.06     1.615
>
> # Accuracy for Holt-Winter's model for Entire historical dataset
> round(accuracy(Hist.HW.ZZZ.Pred$fitted,LifeSciencesProductGroup1.ts),3)
            ME   RMSE    MAE    MPE  MAPE   ACF1 Theil's U
Test set 2.524 22.647 14.852 -0.689 8.793 -0.05     0.776
> round(accuracy((naive(LifeSciencesProductGroup1.ts))$fitted,LifeSciencesProductGroup1.ts),3)
            ME   RMSE    MAE   MPE   MAPE   ACF1 Theil's U
Test set 2.173 31.822 20.798 -0.25 11.478 -0.438         1
> round(accuracy((snaive(LifeSciencesProductGroup1.ts))$fitted,LifeSciencesProductGroup1.ts),3)
             ME   RMSE    MAE   MPE   MAPE  ACF1 Theil's U
Test set 16.822 29.379 23.157 8.694 12.464 0.163     0.748
>
```

From the above accuracy measures , we are comparing the accuracy measures of Holt-winters with naive and seasonal naive forecast. We can clearly see that the MAPE and RMSE values are better for Holt-winters model when compared to naive and seasonal naive forecast.

REGRESSION-BASED MODELS

Regression-based models fall under the model-based approach of Time Series forecasting and are useful for Data visualization and Multi-period forecasting.

Based on the Acf plot, data plots and visualized time series components we now like to use the regression models with trend , seasonality, i.e.regression models with trend , seasonality , Quadratic Trend + Seasonality and see how the data is being forecasted.

REGRESSION MODEL WITH SEASONALITY FOR TRAINING DATASET:

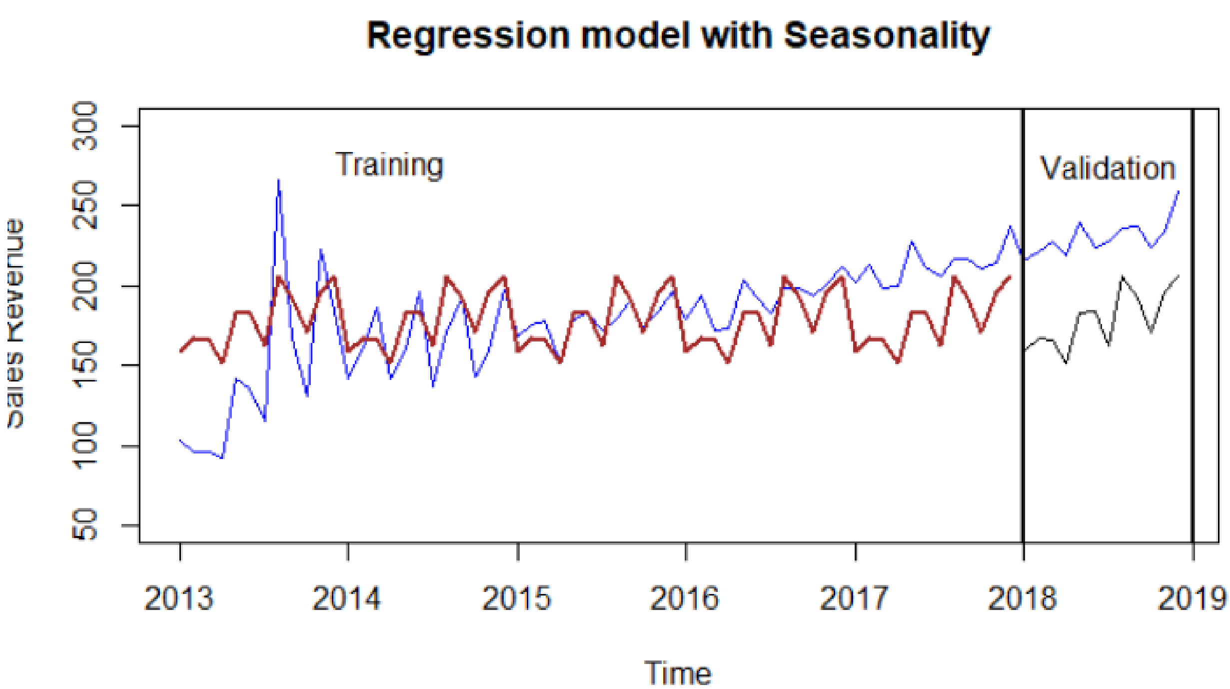
We developed a regression model with seasonality and the summary is as follows:

```
Call:
tslm(formula = LifeSciencesProductGroup1.train.ts ~ season)

Residuals:
   Min     1Q Median     3Q    Max 
-71.52 -21.76   4.81  20.70  59.19 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  158.917     15.167  10.478 5.38e-14 ***
season2        8.656     21.450   0.404   0.6883    
season3        7.134     21.450   0.333   0.7409    
season4       -7.226     21.450  -0.337   0.7377    
season5       23.262     21.450   1.085   0.2836    
season6       24.630     21.450   1.148   0.2566    
season7        3.729     21.450   0.174   0.8627    
season8       47.132     21.450   2.197   0.0329 *  
season9       33.784     21.450   1.575   0.1218    
season10      11.792     21.450   0.550   0.5850    
season11      37.005     21.450   1.725   0.0909 .  
season12      46.219     21.450   2.155   0.0362 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.91 on 48 degrees of freedom
Multiple R-squared:  0.251,	Adjusted R-squared:  0.07941 
F-statistic: 1.463 on 11 and 48 DF,  p-value: 0.1772
```

The Regression equation is

$$Y_t = 158.917 + 8.656D_2 + 7.134D_3 - 7.226D_4 + 23.262D_5 + 24.630D_6 + 3.729D_7 + 47.132D_8 + 33.784D_9 + 11.792D_{10} + 37.005D_{11} + 46.219D_{12}$$

This model has a very less R squared of 0.251 and Adjusted R Squared of 0.97941. The F-statistics p-value is greater than 0.01 which means overall the model is statistically not significant and we can see in the below graph that the forecast is underfitting.



REGRESSION MODEL WITH TREND FOR TRAINING DATASET:

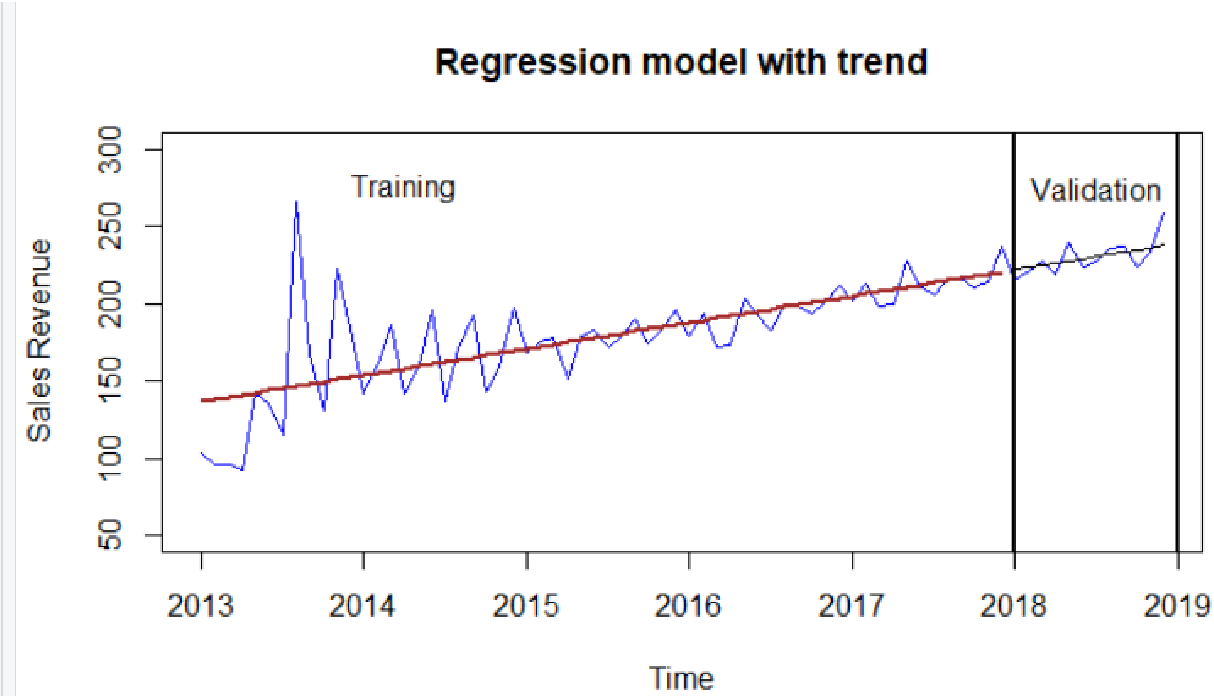
We developed a regression model with trend and the summary is as follows:

```
Call:
tslm(formula = LifeSciencesProductGroup1.train.ts ~ trend)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.907  -9.476  -1.297   6.375 118.602 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 135.2706     6.6403  20.371  < 2e-16 ***
trend         1.4204     0.1893   7.503 4.19e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.4 on 58 degrees of freedom
Multiple R-squared:  0.4925,    Adjusted R-squared:  0.4838 
F-statistic: 56.29 on 1 and 58 DF,  p-value: 4.191e-10
```

The Regression equation is :

$\boldsymbol{Y_t = 135.2706 + 1.4204t}$

This model has an R squared value of 0.4925 and Adjusted R Squared of 0.4838. The F-statistics p-value is less than 0.01 which means overall the model is statistically significant. So this model can be used for forecasting.



REGRESSION MODEL WITH QUADRATIC TREND AND SEASONALITY FOR TRAINING DATASET:

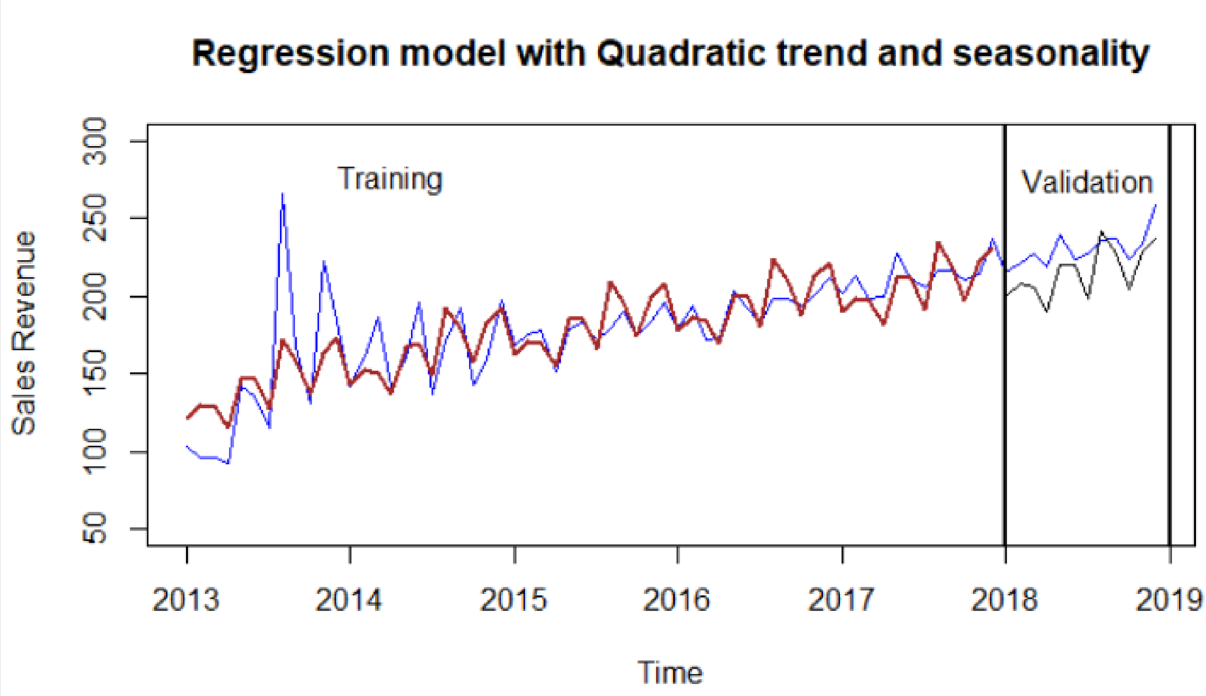
To detrend and deseasonalize the training data set, we developed a regression model with quadratic trend and seasonality and the summary is as follows:

```
Call:
tslm(formula = LifeSciencesProductGroup1.train.ts ~ trend + I(trend^2) +
    season)

Residuals:
    Min      1Q  Median      3Q     Max
-34.077 -12.081  -1.103   7.478  93.431

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.95303   13.05194   9.114 7.15e-12 ***
trend         2.00365    0.69710   2.874  0.00611 **
I(trend^2)   -0.01109    0.01106  -1.003  0.32128
season2       7.21799   14.52327   0.497  0.62156
season3       4.27991   14.52736   0.295  0.76962
season4     -11.47279   14.53367  -0.789  0.43393
season5      17.64383   14.54189   1.213  0.23120
season6      17.66178   14.55182   1.214  0.23105
season7      -4.56552   14.56335  -0.313  0.75532
season8      37.53276   14.57647   2.575  0.01331 *
season9      22.90179   14.59129   1.570  0.12337
season10     -0.35072   14.60799  -0.024  0.98095
season11     23.62406   14.62689   1.615  0.11313
season12     31.62224   14.64835   2.159  0.03612 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.96 on 46 degrees of freedom
Multiple R-squared:  0.671,	Adjusted R-squared:  0.578
F-statistic: 7.217 on 13 and 46 DF,  p-value: 2.159e-07
```

The Regression model equation is:

$$Y_t = 118.95303 + 2.00365t - 0.01109t^2 + 7.21799D_2 + 4.27991D_3 - 11.47279D_4 + 17.64383D_5 + 17.66178D_6 - 4.56552D_7 + 37.53276D_8 + 22.90179D_9 - 0.35072D_{10} + 23.62406D_{11} + 31.62224D_{12}$$

This model too has a very high R squared of 0.671 and Adjusted R Squared of 0.578. The F-statistics p-value is lower than 0.01 which means overall the model is statistically significant. However, the regression coefficients for Season 2, 3, 4, 5, 6, 7, 9, 10 and 11 are not statistically significant (p-value is greater than 0.01). The regression coefficients for Season 8 and 12 have value less and are statistically significant. We used the forecast() function to predict the values in the validation period, the forecast plot is as follows and the table is in the appendix.



ACCURACY COMPARISON FOR ALL 3 REGRESSION MODELS:

We used accuracy() function to evaluate the better model from the three regression models we built above.

```
> round(accuracy(LifeSciencesProductGroup1.train.season.pred $mean, LifeSciencesProductGroup1.ts),3)
              ME   RMSE    MAE    MPE   MAPE   ACF1 Theil's U
Test set 51.422 52.538 51.422 22.473 22.473 -0.079     4.141
> round(accuracy(LifeSciencesProductGroup1.train.trend.pred$mean,LifeSciencesProductGroup1.ts),3)
            ME  RMSE   MAE    MPE  MAPE   ACF1 Theil's U
Test set 0.287 8.544 6.873 -0.031 2.935 -0.231     0.675
> round(accuracy(LifeSciencesProductGroup1.quad.train.trend.season.pred$mean,LifeSciencesProductGroup1.ts),3)
             ME  RMSE    MAE   MPE  MAPE   ACF1 Theil's U
Test set 14.831 17.79 15.699 6.489 6.857 -0.303     1.431
```

From observing the above accuracy measures, Regression model with trend have substantially better MAPE (2.93) that Regression Model with Quadratic Trend and Seasonality (6.857). Taking into consideration the superiority of MAPE, we conclude that Regression model with trend is a more accurate model among these regression models. Also the RMSE value is better for Regression model with trend compared to other 2 regression models.

REGRESSION MODEL WITH SEASONALITY ON HISTORICAL DATA:

Before attempting to forecast future values of the series, the training and validation periods were recombined into the entire (historical) time series dataset. The chosen model regression model with seasonality was then run on the entire historical dataset.

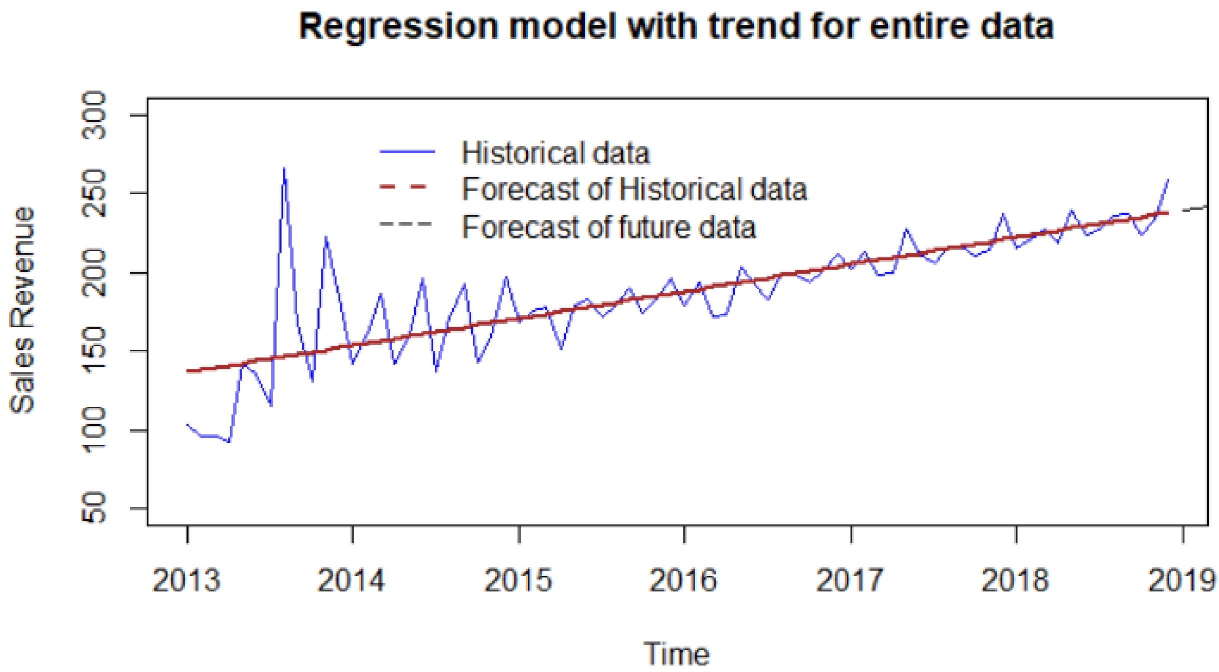
Here we are using regression model with seasonality on original data and the summary is as follows:

```
Call:
tslm(formula = LifeSciencesProductGroup1.ts ~ trend)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.722  -9.080  -1.635   5.705 118.758 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 135.0573     5.5701   24.25   <2e-16 ***
trend         1.4276     0.1326   10.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.39 on 70 degrees of freedom
Multiple R-squared:  0.6234,	Adjusted R-squared:  0.618 
F-statistic: 115.9 on 1 and 70 DF,  p-value: < 2.2e-16
```

The Regression equation for the above model is

$\boldsymbol{Y_t = 135.0573 + 1.4276t}$

Both R squared(0.6234) and Adjusted R squared(0.618) values are relatively moderate for this model and overall model also seems to be statistically significant as their F-statistic P-Value is less than 0.01. Also the trend coefficient is also statistically significant. The forecast for 12 periods into the future was done using the forecast() function. The plot is as follows and the table is in the appendix.



TWO-LEVEL MODEL

To improve the predictive performance, we combine multiple forecasting methods where the first method uses original time series to predict the future, and the second method uses forecast residuals from the first method to generate forecast for errors, and then combine two forecasts together. Here we have used Regression Model with trend to create the forecast for the training data and used Trailing Moving Average to forecast the residuals of this model. Below is the table for Two Level Model - Regression with trend and Trailing MA for residuals and total forecast for the validation data.

```
> total.reg.train.ma.pred
   Regression.Forecast Residuals.Forecast Combined.Forecast
1             158.9166         -0.0385309          221.8771
2             167.5726         -0.0385309          223.2975
3             166.0502         -0.0385309          224.7179
4             151.6911         -0.0385309          226.1383
5             182.1791         -0.0385309          227.5588
6             183.5462         -0.0385309          228.9792
7             162.6459         -0.0385309          230.3996
8             206.0490         -0.0385309          231.8200
9             192.7006         -0.0385309          233.2404
10            170.7086         -0.0385309          234.6608
11            195.9216         -0.0385309          236.0812
12            205.1358         -0.0385309          237.5016
```

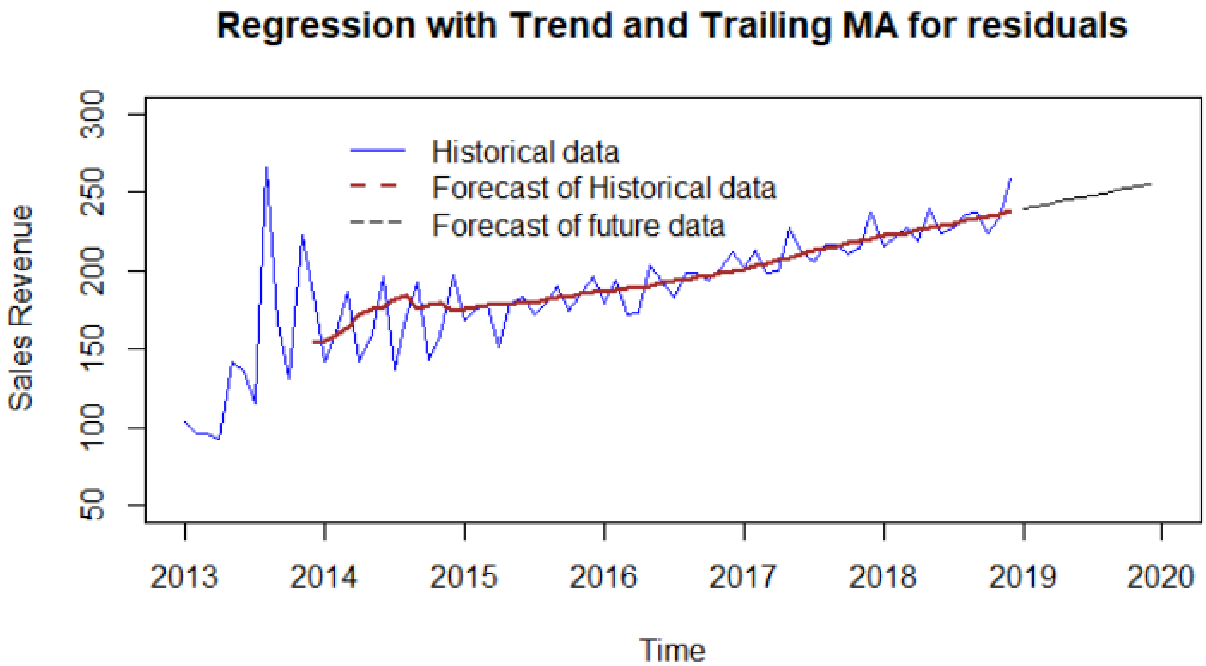
We recombined the partitioned data into one dataset and reapplied the two-level model built above. The table below contains Sales revenue for the year 2019 including regression forecast, trailing MA forecasts for residuals and the total forecasts that combines first two forecasts.

	Regression.Forecast	Residuals.Forecast	Combined.Forecast
1	168.2222	-41173.83	239.2942
2	176.4574	-77858.22	240.7217
3	176.2324	-114542.61	242.1493
4	162.8855	-151227.01	243.5769
5	191.7268	-187911.40	245.0044
6	190.3181	-224595.79	246.4320
7	173.3504	-261280.18	247.8596
8	211.0797	-297964.57	249.2871
9	200.1109	-334648.97	250.7147
10	179.5504	-371333.36	252.1423
11	202.0832	-408017.75	253.5698
12	213.9440	-444702.14	254.9974

```
> round(accuracy(LifeSciencesProductGroup1.trend.pred$fitted + ma.trailing.trend.res_12,LifeSciencesProductGroup1.ts),3)
             ME   RMSE   MAE    MPE MAPE   ACF1 Theil's U
Test set -1.251 13.437 9.777 -1.322 5.46 -0.267     0.611
> round(accuracy((naive(LifeSciencesProductGroup1.ts))$fitted,LifeSciencesProductGroup1.ts),3)
            ME   RMSE    MAE   MPE   MAPE   ACF1 Theil's U
Test set 2.173 31.822 20.798 -0.25 11.478 -0.438         1
> round(accuracy((snaive(LifeSciencesProductGroup1.ts))$fitted,LifeSciencesProductGroup1.ts),3)
             ME   RMSE    MAE   MPE   MAPE  ACF1 Theil's U
Test set 16.822 29.379 23.157 8.694 12.464 0.163     0.748
>
```

From the above accuracy measures , we are comparing the accuracy measures of Two level regression model (Regression model with trend + Trailing MA for residuals) with naive and seasonal naive forecast. We can clearly see that the MAPE (5.46) and RMSE (13.437) values are better for Two level regression model (Regression model with trend + Trailing MA for residuals) when compared to naive and seasonal naive forecast.

The forecast plot is as follows:



AUTOREGRESSIVE & MOVING AVERAGE MODELS

Autoregressive models are model-based approach of Time Series Forecasting and they model the existing relationship between successive datasets or autocorrelation directly in regression model, using past observations as predictors. They are similar to the linear regression models, except that the predictors are the past values of the series. We built multiple models to compare and find the most accurate one.

AUTOREGRESSIVE (AR) MODEL OF ORDER 2 ON TRAINING DATA:

We build an Autoregressive model of order 2 on the Training Data and below is the summary.

```
Series: LifeSciencesProductGroup1.train.ts
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1     ar2      mean
      0.4179  0.2466  177.5684
s.e.  0.1265  0.1341   10.9041

sigma^2 estimated as 918.3:  log likelihood=-288.52
AIC=585.04   AICc=585.77   BIC=593.42

Training set error measures:
                   ME    RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 1.581526 29.5368 20.73926 -2.226132 12.79515 0.8429324 -0.1845659
```

As can be seen from the above summary the β_1 coefficient is 0.4179 , β_2 coefficient is 0.2466 with the intercept is177.5684. And, the Model equation is as follows:

$\boldsymbol{y_t = 177.5684 + 0.4179\, y_{t-1} + 0.2466\, y_{t-2}}$

We then applied the forecast() function to make predictions for Validation period and the results are shown in the appendix.

MOVING AVERAGE (MA) MODEL OF ORDER 2 ON TRAINING DATA:

Moving Average model uses past forecast residuals (errors) of q autocorrelation lags in a regression-like model. We used Arima() function to fit MA(2) model with order = c(0,0,2) for Moving average and the results are as follows:

```
Series: LifeSciencesProductGroup1.train.ts
ARIMA(0,0,2) with non-zero mean

Coefficients:
         ma1      ma2      mean
      0.7159  -0.0890  178.5235
s.e.  0.1143   0.0954    6.1592

sigma^2 estimated as 915:  log likelihood=-288.65
AIC=585.31   AICc=586.04   BIC=593.69

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 0.3620217 29.48362 21.60395 -3.036568 13.69859 0.8780774 -0.05107523
```

The model equation is as follows:

$\boldsymbol{y_t = 178.5235 + 0.7159\, e_{t-1} - 0.0890 e_{t-2}}$

We used forecast() function to make predictions for Training data with the above MA model in validation set and the results are in the appendix.

AUTOREGRESSIVE MOVING AVERAGE (ARMA) MODEL OF ORDER 2 ON TRAINING DATA:

We also used ARMA model that incorporates time series lags ('Autoregressive' part) and lags of forecast residuals ('Moving Average' part) to capture all forms of autocorrelation in the data. We used ARMA(2,2) model with order = c(2,0,2) in ARIMA function and the results are as follows.

```
> summary(LifeSciencesProductGroup1.train.arma2)
Series: LifeSciencesProductGroup1.train.ts
ARIMA(2,0,2) with non-zero mean

Coefficients:
         ar1     ar2     ma1      ma2      mean
      0.3962  0.5733  0.1795  -0.6776  171.5610
s.e.  0.1408  0.1356  0.1190   0.0943   33.1025

sigma^2 estimated as 680.8:  log likelihood=-279.36
AIC=570.72   AICc=572.3   BIC=583.28

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 5.1507 24.98179 15.83542 1.185187 9.126962 0.6436193 -0.2342755
```

The model equation is as follows:

$$y_t = 171.5610 + 0.3962\, y_{t-1} + 0.5733\, y_{t-2} + 0.1795 e_{t-1} - 0.6776\, e_{t-2}$$

We applied forecast() function to make predictions for validation period with ARMA model and the results are in the appendix.

ARIMA MODELS

Autoregressive Integrated Moving Average (ARIMA) is a class of popular models in time series forecasting and is also referred to as Box-Jenkins methodology or Box-Jenkins approach.

ARIMA Models are capable of presenting any time series component – level (stationary), trend, and seasonality – or a combination of these components. These are model-based approach for forecasting and are flexible to forecast any time series. We built multiple versions of ARIMA Models to forecast the Sales Revenue times series data.

In order to incorporate trend and seasonality components we built Seasonal ARIMA Model. We used Arima() function to fit ARIMA(2,1,2)(1,1,2) model for trend and seasonality.

ARIMA (p, d, q) (P, D, Q)m model is used to forecast data with level, trend, and seasonality components. In addition to the (p, d, q) parameters, it includes seasonal parameters.

ARIMA (2, 1, 2) (1, 1, 2)12 means the following:

- p = 2, order 2 autoregressive model AR(2)
- d = 1, order 1 differencing to remove linear trend
- q = 2, order 2 moving average MA(2) for error lags
- P = 1, order 1 autoregressive model AR(1) for seasonality
- D = 1, order 1 differencing to remove linear trend
- Q = 2, order 2 moving average MA(2) for error lags
- m = 12, for monthly seasonality

The summary of the model is as follows:

```
> summary(LifeSciencesProductGroup1.train.arima)
Series: LifeSciencesProductGroup1.train.ts
ARIMA(2,1,2)(1,1,2)[12]

Coefficients:
          ar1      ar2      ma1     ma2     sar1     sma1     sma2
      -0.5074  -0.7965  -0.2377  0.4871  -0.0356  -0.0557  -0.0056
s.e.   0.1829   0.1101   0.2873  0.2139      NaN      NaN      NaN

sigma^2 estimated as 522.1:  log likelihood=-211.04
AIC=438.09   AICc=441.88   BIC=452.89

Training set error measures:
                     ME     RMSE     MAE        MPE     MAPE      MASE       ACF1
Training set -0.9403081 18.65581 10.0603 -0.7133607 5.758261 0.4088939 0.06456839
```

$$\boldsymbol{y_t - y_{t-1} = -0.5074\,(y_{t-1} - y_{t-2}) - 0.7965(y_{t-2} - y_{t-3}) - 0.2377\, e_{t-1}}$$

$$\boldsymbol{+\, 0.4871 e_{t-2} - 0.0356\,(y_{t-1} - y_{t-13}) - 0.0557\, r_{t-1} - 0.0056 r_{t-2}}$$

We used forecast() function to make predictions with ARIMA model in the validation set and the results are in the appendix.

AUTO ARIMA MODEL ON THE TRAINING DATASET:

Since ARIMA model has a complex structure, automated ARIMA Model development is the best way to go, therefore to forecast Sales Revenue of the company for Validation periods , we applied the auto.arima() function in R on the training dataset and the summary is as follows:

```
> summary(LifeSciencesProductGroup1.train.auto.arima)
Series: LifeSciencesProductGroup1.train.ts
ARIMA(2,1,3)(1,0,0)[12]

Coefficients:
          ar1      ar2     ma1     ma2     ma3    sar1
      -0.6614  -0.5532  0.1153  0.0134  0.5331  0.5269
s.e.   0.1397   0.1456  0.1441  0.1836  0.1434  0.1668

sigma^2 estimated as 386.1:  log likelihood=-259.41
AIC=532.81   AICc=535.01   BIC=547.35

Training set error measures:
                   ME     RMSE     MAE       MPE    MAPE      MASE        ACF1
Training set 1.48338 18.46748 11.9983 0.3120948 6.75537 0.4876626 -0.06758239
```

The Model equation is as follows:

$y_t - y_{t-1} = -0.6614(y_{t-1} - y_{t-2}) -0.5532(y_{t-2} - y_{t-3}) + 0.1153\, e_{t-1} + 0.0134 e_{t-2} + 0.5331 e_{t-3} + 0.5269\, (y_{t-1} - y_{t-13})$

We applied forecast() function to make predictions for the validation period and the results are in the appendix.

ACCURACY COMPARISON FOR AUTOREGRESSIVE AND ARIMA MODELS

We then compared models using accuracy() function to identify common performance measures for future period forecast and the results are as follows:

```
> #Use accuracy() function to identify common accuracy measures for validation period forecast:
> # (1) AR(2) model; (2) MA(2) model; (3) ARMA(2,2) model; (4) ARIMA(2,1,2)(1,1,2) model; and
> # (5) Auto ARIMA model.
> round(accuracy(LifeSciencesProductGroup1.train.ar2.pred, LifeSciencesProductGroup1.valid.ts), 3)
                 ME   RMSE    MAE    MPE   MAPE  MASE   ACF1 Theil's U
Training set  1.582 29.537 20.739 -2.226 12.795 0.843 -0.185        NA
Test set     40.747 45.193 40.747 17.389 17.389 1.656  0.451     3.692
> round(accuracy(LifeSciencesProductGroup1.train.ma2.pred, LifeSciencesProductGroup1.valid.ts), 3)
                 ME   RMSE    MAE    MPE   MAPE  MASE   ACF1 Theil's U
Training set  0.362 29.484 21.604 -3.037 13.699 0.878 -0.051        NA
Test set     49.414 51.963 49.414 21.239 21.239 2.008  0.051     4.273
> round(accuracy(LifeSciencesProductGroup1.train.arma2.pred, LifeSciencesProductGroup1.valid.ts), 3)
                 ME   RMSE    MAE   MPE  MAPE  MASE   ACF1 Theil's U
Training set  5.151 24.982 15.835 1.185 9.127 0.644 -0.234        NA
Test set     17.432 22.318 18.983 7.321 8.043 0.772  0.203     1.808
> round(accuracy(LifeSciencesProductGroup1.train.arima.pred, LifeSciencesProductGroup1.valid.ts), 3)
                 ME   RMSE    MAE    MPE  MAPE  MASE   ACF1 Theil's U
Training set -0.940 18.656 10.060 -0.713 5.758 0.409  0.065        NA
Test set     -2.098  5.090  4.075 -0.958 1.797 0.166 -0.137     0.405
> round(accuracy(LifeSciencesProductGroup1.train.auto.arima.pred, LifeSciencesProductGroup1.valid.ts), 3)
                 ME   RMSE    MAE    MPE  MAPE  MASE   ACF1 Theil's U
Training set  1.483 18.467 11.998  0.312 6.755 0.488 -0.068        NA
Test set     -2.563  6.519  5.412 -1.228 2.364 0.220 -0.016     0.495
```

As can be seen above the best values of MAPE and RMSE are for the ARIMA(2,1,2)(1,1,2) model on the validation and training set.

AUTO ARIMA MODEL ON THE ENTIRE DATASET:

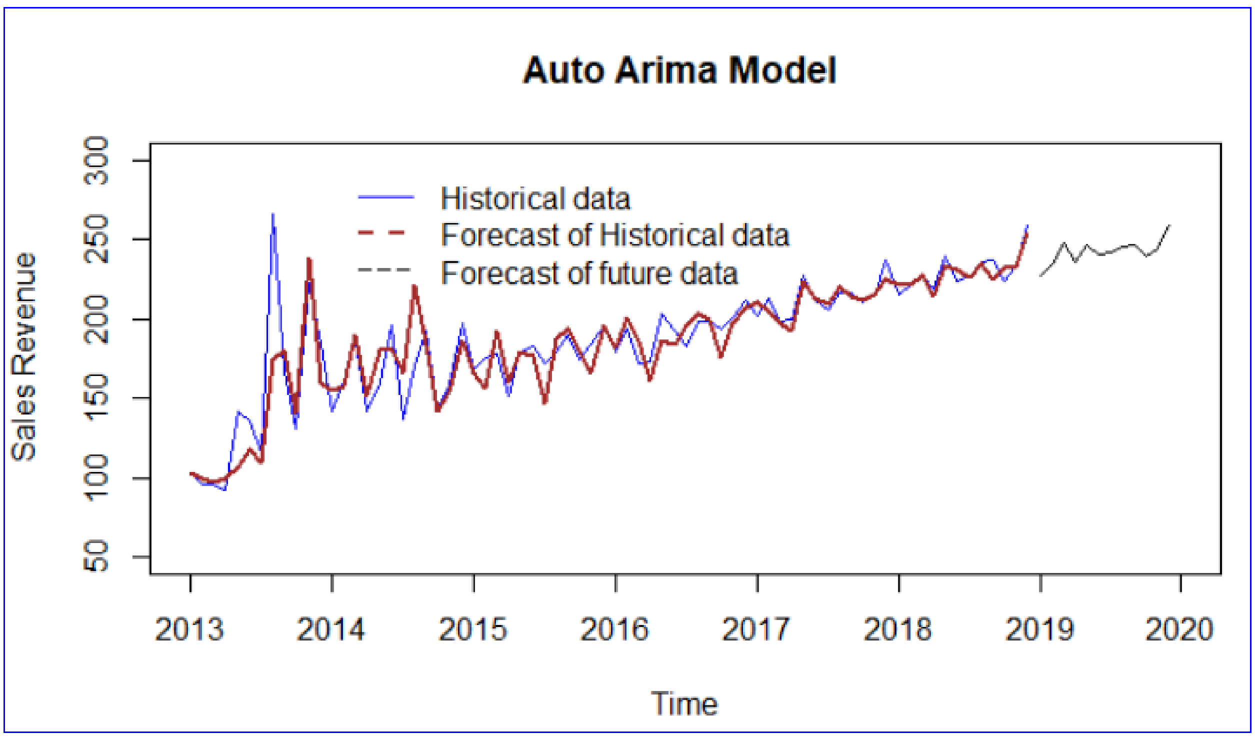
Since the auto ARIMA model resulted into the best accuracy measures, we used it to forecast 12 periods into the future using the entire (historical) dataset and the summary is as follows:

```
> summary(LifeSciencesProductGroup1.auto.arima)
Series: LifeSciencesProductGroup1.ts
ARIMA(2,1,4)(1,0,0)[12]

Coefficients:
          ar1      ar2      ma1     ma2     ma3      ma4    sar1
      -0.3100  -0.4423  -0.3236  0.0838  0.5822  -0.4530  0.5729
s.e.   0.2343   0.1664   0.2551  0.1303  0.1075   0.2359  0.1357

sigma^2 estimated as 304.5:  log likelihood=-304.07
AIC=624.14   AICc=626.46   BIC=642.24

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE     MASE         ACF1
Training set 1.644273 16.45106 9.961356 0.4873807 5.527798 0.430165 -0.008290879
```

The model equation is:

$$y_t - y_{t-1} = -0.3100(y_{t-1} - y_{t-2}) - 0.4423(y_{t-2} - y_{t-3}) - 0.3236\, e_{t-1} + 0.0838 e_{t-2} + 0.5822 e_{t-3} - 0.4530 e_{t-4} + 0.5729\,(y_{t-1} - y_{t-13})$$

We applied forecast() function to make predictions using auto ARIMA model for the future 12 periods, the plot is as follows and the table in the appendix.



CONCLUSION

This is the most important step in Time Series Forecasting, we need to generate performance measures that evaluate the models we built and find out the most accurate one. Here, we used accuracy() function to evaluate the performance of the models we built on the Company historical data. We are focusing on the MAPE and RMSE measures of forecasting models for comparison. Where MAPE is percentage score of how forecast deviates from actual values and RMSE is standard deviation of residuals. In Business terms however, MAPE is simply referred to as the margin of error. The MAPE and RMSE values are shown below:

```
> #-----------------------Model Comparison--------------------------#
>
> #Comparison of Models using accuracy()
> round(accuracy(LifeSciencesProductGroup1.auto.arima.pred$fitted, LifeSciencesProductGroup1.ts),3)
            ME   RMSE   MAE   MPE  MAPE   ACF1 Theil's U
Test set 1.644 16.451 9.961 0.487 5.528 -0.008     0.572
> round(accuracy(Hist.HW.ZZZ.Pred$fitted,LifeSciencesProductGroup1.ts),3)
            ME   RMSE    MAE    MPE  MAPE  ACF1 Theil's U
Test set 2.524 22.647 14.852 -0.689 8.793 -0.05     0.776
> round(accuracy(LifeSciencesProductGroup1.trend.pred$fitted,LifeSciencesProductGroup1.ts ),3)
         ME   RMSE    MAE    MPE  MAPE  ACF1 Theil's U
Test set  0 23.059 14.227 -2.058 8.701 0.046     0.823
> round(accuracy(LifeSciencesProductGroup1.trend.pred$fitted + ma.trailing.trend.res_12,LifeSciencesProductGroup1.ts),3)
             ME   RMSE   MAE    MPE MAPE   ACF1 Theil's U
Test set -1.251 13.437 9.777 -1.322 5.46 -0.267     0.611
> round(accuracy((naive(LifeSciencesProductGroup1.ts))$fitted,LifeSciencesProductGroup1.ts),3)
            ME   RMSE    MAE   MPE   MAPE   ACF1 Theil's U
Test set 2.173 31.822 20.798 -0.25 11.478 -0.438         1
> round(accuracy((snaive(LifeSciencesProductGroup1.ts))$fitted,LifeSciencesProductGroup1.ts),3)
             ME   RMSE    MAE   MPE   MAPE  ACF1 Theil's U
Test set 16.822 29.379 23.157 8.694 12.464 0.163     0.748
>
```

Model	**MAPE**	**RMSE**
Holt Winters	8.793	22.647
Regression Model with trend	8.701	23.059
Two-level Model (Regression with Seasonality + Trailing MA for residuals)	5.46	13.437
Auto ARIMA	5.528	16.451
Seasonal Naïve	12.464	29.379
Naïve	11.478	31.822

The above accuracy measures indicate that Two-level Model (Regression with Seasonality + Trailing MA for residuals) has substantially better **MAPE(5.46)** and **RMSE(13.437)** than all other models. Hence we conclude that Two-level Model (Regression with Seasonality + Trailing MA for residuals) is more accurate model and is our final choice of the forecasting model to project sales revenue of the company in 2019.

LIMITATIONS & WAY FORWARD

Currently in the project as per the scope established in the introduction, we focussed on forecasting the revenue . But as seen in our visualizations the forecast is going towards an upward trend. This company does a lot of acquisitions and as seen in our visualizations we did notice sudden spikes. We also need to account for opportunities which are getting converted into revenue to predict possible downtrends in future.

We also can do a forecasting on expenses for Research to help the company to manage the expenses to keep them under control. We believe utilizing these expenses forecast and including opportunities will help to further improve the quality of the forecast.

APPENDIX

BIBLIOGRAPHY

LECTURE MATERIALS:

By Dr. Zinovy Radovilsky, Professor of Management, California State University, East Bay

FORECAST TABLES

Training Data Set

```
> LifeSciencesProductGroup1.train.ts
           Jan       Feb       Mar       Apr       May       Jun       Jul       Aug       Sep       Oct       Nov       Dec
2013 103.72687  96.05653  95.81561  93.04564 141.59212 135.47498 115.58756 265.23586 167.39477 130.66058 222.80031 183.48213
2014 142.11643 160.64808 186.47087 141.59212 160.10455 196.12356 137.52067 170.73270 192.15251 143.27271 158.82130 197.72592
2015 168.05952 175.25715 178.03699 150.95670 178.99119 182.95533 171.53152 179.39030 189.82510 174.56586 183.72492 195.96890
2016 178.76842 193.46102 172.14753 173.05271 202.65133 191.90861 182.52019 198.29617 198.01024 194.07987 200.23117 211.30695
2017 201.91198 212.44002 197.78009 199.80820 227.55609 211.26846 206.06948 216.58986 216.12055 210.96381 214.03027 237.19535
```

Validation Data Set :

```
> LifeSciencesProductGroup1.valid.ts
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2018 214.7499 220.8819 227.1435 218.8575 239.4658 224.1777 226.8731 236.2334 237.1623 223.7594 232.8913 257.9845
```

HW Model Point forecast for Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	213.5605	213.5605	213.5605
Feb 2018	213.5955	213.5955	213.5955
Mar 2018	213.6273	213.6273	213.6273
Apr 2018	213.6562	213.6562	213.6562
May 2018	213.6825	213.6825	213.6825
Jun 2018	213.7064	213.7064	213.7064
Jul 2018	213.7281	213.7281	213.7281
Aug 2018	213.7478	213.7478	213.7478
Sep 2018	213.7657	213.7657	213.7657
Oct 2018	213.7821	213.7821	213.7821
Nov 2018	213.7969	213.7969	213.7969
Dec 2018	213.8104	213.8104	213.8104

HW Model Point Forecast on Historical Dataset

	Point Forecast	Lo 0	Hi 0
Jan 2019	234.1322	234.1322	234.1322
Feb 2019	235.5021	235.5021	235.5021
Mar 2019	236.6941	236.6941	236.6941
Apr 2019	237.7312	237.7312	237.7312
May 2019	238.6335	238.6335	238.6335
Jun 2019	239.4186	239.4186	239.4186
Jul 2019	240.1018	240.1018	240.1018
Aug 2019	240.6961	240.6961	240.6961
Sep 2019	241.2133	241.2133	241.2133
Oct 2019	241.6633	241.6633	241.6633
Nov 2019	242.0548	242.0548	242.0548
Dec 2019	242.3954	242.3954	242.3954

Regression Model with Seasonality for Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	158.9166	158.9166	158.9166
Feb 2018	167.5726	167.5726	167.5726
Mar 2018	166.0502	166.0502	166.0502
Apr 2018	151.6911	151.6911	151.6911
May 2018	182.1791	182.1791	182.1791
Jun 2018	183.5462	183.5462	183.5462
Jul 2018	162.6459	162.6459	162.6459
Aug 2018	206.0490	206.0490	206.0490
Sep 2018	192.7006	192.7006	192.7006
Oct 2018	170.7086	170.7086	170.7086
Nov 2018	195.9216	195.9216	195.9216
Dec 2018	205.1358	205.1358	205.1358

Regression Model with Trend for Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	221.9156	221.9156	221.9156
Feb 2018	223.3361	223.3361	223.3361
Mar 2018	224.7565	224.7565	224.7565
Apr 2018	226.1769	226.1769	226.1769
May 2018	227.5973	227.5973	227.5973
Jun 2018	229.0177	229.0177	229.0177
Jul 2018	230.4381	230.4381	230.4381
Aug 2018	231.8585	231.8585	231.8585
Sep 2018	233.2789	233.2789	233.2789
Oct 2018	234.6993	234.6993	234.6993
Nov 2018	236.1198	236.1198	236.1198
Dec 2018	237.5402	237.5402	237.5402

Regression Model with Quadratic Trend and Seasonality for Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	199.8998	199.8998	199.8998
Feb 2018	207.7571	207.7571	207.7571
Mar 2018	205.4361	205.4361	205.4361
Apr 2018	190.2782	190.2782	190.2782
May 2018	219.9676	219.9676	219.9676
Jun 2018	220.5360	220.5360	220.5360
Jul 2018	198.8370	198.8370	198.8370
Aug 2018	241.4415	241.4415	241.4415
Sep 2018	227.2945	227.2945	227.2945
Oct 2018	204.5037	204.5037	204.5037
Nov 2018	228.9181	228.9181	228.9181
Dec 2018	237.3337	237.3337	237.3337

Autoregressive (AR) Model of Order 2 Point Forecast for Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	211.4783	211.4783	211.4783
Feb 2018	206.4434	206.4434	206.4434
Mar 2018	197.9975	197.9975	197.9975
Apr 2018	193.2263	193.2263	193.2263
May 2018	189.1497	189.1497	189.1497
Jun 2018	186.2695	186.2695	186.2695
Jul 2018	184.0606	184.0606	184.0606
Aug 2018	182.4272	182.4272	182.4272
Sep 2018	181.1999	181.1999	181.1999
Oct 2018	180.2842	180.2842	180.2842
Nov 2018	179.5989	179.5989	179.5989
Dec 2018	179.0867	179.0867	179.0867

Moving Average (MA) Model Point Forecast for Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	207.2964	207.2964	207.2964
Feb 2018	174.6795	174.6795	174.6795
Mar 2018	178.5235	178.5235	178.5235
Apr 2018	178.5235	178.5235	178.5235
May 2018	178.5235	178.5235	178.5235
Jun 2018	178.5235	178.5235	178.5235
Jul 2018	178.5235	178.5235	178.5235
Aug 2018	178.5235	178.5235	178.5235
Sep 2018	178.5235	178.5235	178.5235
Oct 2018	178.5235	178.5235	178.5235
Nov 2018	178.5235	178.5235	178.5235
Dec 2018	178.5235	178.5235	178.5235

Autoregressive Moving Average (ARMA) Model Point forecast on Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	224.0562	224.0562	224.0562
Feb 2018	211.4873	211.4873	211.4873
Mar 2018	217.4777	217.4777	217.4777
Apr 2018	212.6447	212.6447	212.6447
May 2018	214.1645	214.1645	214.1645
Jun 2018	211.9957	211.9957	211.9957
Jul 2018	212.0078	212.0078	212.0078
Aug 2018	210.7691	210.7691	210.7691
Sep 2018	210.2852	210.2852	210.2852
Oct 2018	209.3833	209.3833	209.3833
Nov 2018	208.7486	208.7486	208.7486
Dec 2018	207.9800	207.9800	207.9800

ARIMA(2, 1, 2) (1, 1, 2)[12] Model Point Forecast for Validation Period

	Point Forecast	Lo 0	Hi 0
Jan 2018	219.0844	219.0844	219.0844
Feb 2018	229.3174	229.3174	229.3174
Mar 2018	220.8983	220.8983	220.8983
Apr 2018	219.6580	219.6580	219.6580
May 2018	243.7667	243.7667	243.7667
Jun 2018	232.3650	232.3650	232.3650
Jul 2018	227.6014	227.6014	227.6014
Aug 2018	234.7704	234.7704	234.7704
Sep 2018	235.6461	235.6461	235.6461
Oct 2018	232.8451	232.8451	232.8451
Nov 2018	234.0553	234.0553	234.0553
Dec 2018	255.3460	255.3460	255.3460

Auto Arima Model Point forecast on Validation Dataset

	Point Forecast	Lo 0	Hi 0
Jan 2018	222.3239	222.3239	222.3239
Feb 2018	226.0509	226.0509	226.0509
Mar 2018	232.8328	232.8328	232.8328
Apr 2018	225.3145	225.3145	225.3145
May 2018	237.5894	237.5894	237.5894
Jun 2018	235.3087	235.3087	235.3087
Jul 2018	229.6994	229.6994	229.6994
Aug 2018	233.6550	233.6550	233.6550
Sep 2018	236.0453	236.0453	236.0453
Oct 2018	232.4620	232.4620	232.4620
Nov 2018	233.1916	233.1916	233.1916
Dec 2018	246.4625	246.4625	246.4625

Auto Arima Model Point Forecast in the Future

	Point Forecast	Lo 0	Hi 0
Jan 2019	226.7186	226.7186	226.7186
Feb 2019	234.9816	234.9816	234.9816
Mar 2019	247.2941	247.2941	247.2941
Apr 2019	235.9938	235.9938	235.9938
May 2019	245.9730	245.9730	245.9730
Jun 2019	240.6793	240.6793	240.6793
Jul 2019	241.9578	241.9578	241.9578
Aug 2019	245.8702	245.8702	245.8702
Sep 2019	246.9696	246.9696	246.9696
Oct 2019	239.7564	239.7564	239.7564
Nov 2019	244.5930	244.5930	244.5930
Dec 2019	258.8861	258.8861	258.8861