Time Series Analysis and Forecasting

AIM:

To develop an ARIMA model to forecast sale / demand for next year.

DATA USED:

<https://drive.google.com/open?id=1sluidX3PbuOzqjgT8ehDyNFzjPDj0fXn>

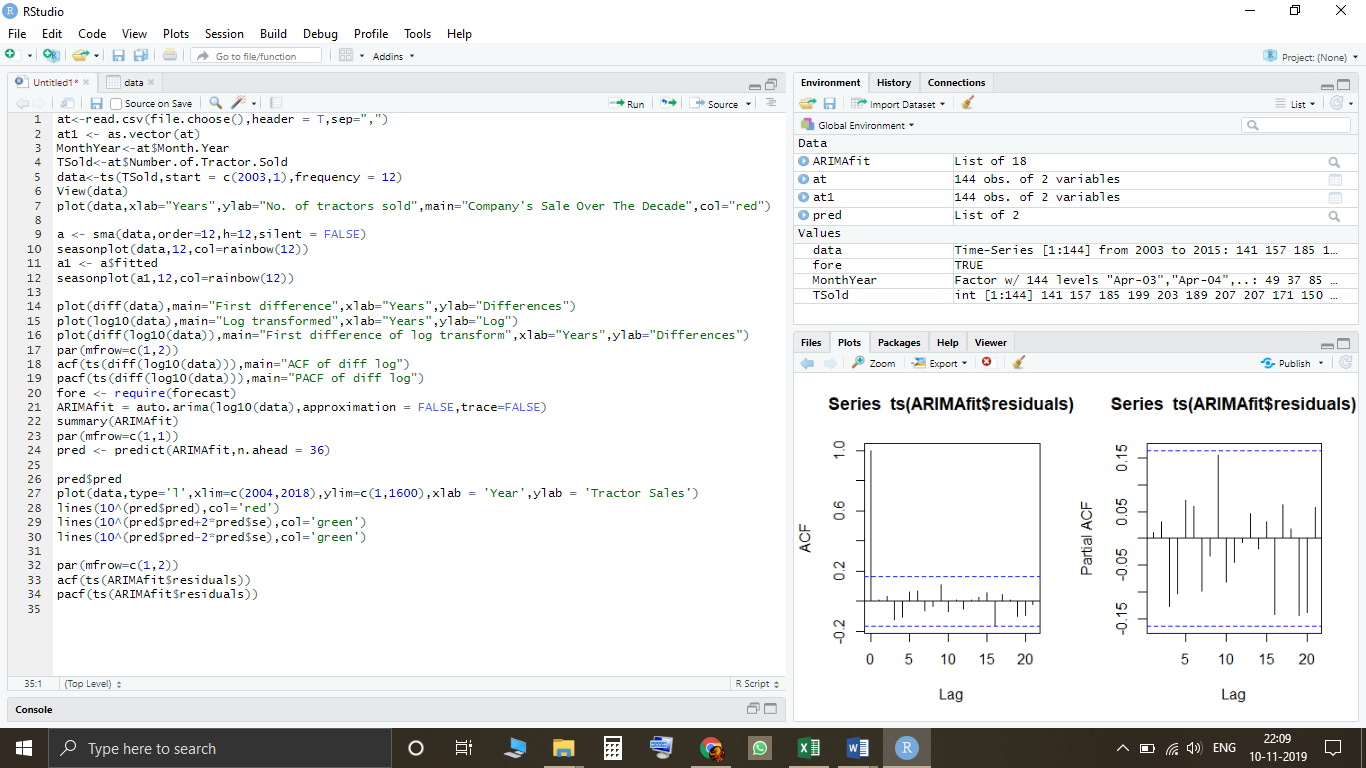
PowerHorse, a tractor and farm equipment manufacturing company, was established a few years after World War II. The company has shown a consistent growth in its revenue from tractor sales since its inception. However, over the years the company has struggled to keep it’s inventory and production cost down because of variability in sales and tractor demand. The management at PowerHorse is under enormous pressure from the shareholders and board to reduce the production cost. Additionally, they are also interested in understanding the impact of their marketing and farmer connect efforts towards overall sales. In the same effort, they have hired you as a data science and predictive analytics consultant.

You will start your investigation of this problem in the next part of this series using the concept discussed in this article. Eventually, you will develop an ARIMA model to forecast sale / demand for next year. Additionally, you will also investigate the impact of marketing program on sales by using an exogenous variable ARIMA model.

DATA:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Month-Year | Jan-03 | Feb-03 | Mar-03 | Apr-03 | May-03 | Jun-03 | Jul-03 | Aug-03 | Sep-03 | Oct-03 | Nov-03 | Dec-03 |
| Number of Tractor Sold | 141 | 157 | 185 | 199 | 203 | 189 | 207 | 207 | 171 | 150 | 138 | 165 |
| Month-Year | Jan-04 | Feb-04 | Mar-04 | Apr-04 | May-04 | Jun-04 | Jul-04 | Aug-04 | Sep-04 | Oct-04 | Nov-04 | Dec-04 |
| Number of Tractor Sold | 145 | 168 | 197 | 208 | 210 | 209 | 238 | 238 | 199 | 168 | 152 | 196 |
| Month-Year | Jan-05 | Feb-05 | Mar-05 | Apr-05 | May-05 | Jun-05 | Jul-05 | Aug-05 | Sep-05 | Oct-05 | Nov-05 | Dec-05 |
| Number of Tractor Sold | 183 | 200 | 249 | 251 | 289 | 249 | 279 | 279 | 232 | 204 | 194 | 232 |
| Month-Year | Jan-06 | Feb-06 | Mar-06 | Apr-06 | May-06 | Jun-06 | Jul-06 | Aug-06 | Sep-06 | Oct-06 | Nov-06 | Dec-06 |
| Number of Tractor Sold | 215 | 239 | 270 | 279 | 307 | 305 | 322 | 339 | 263 | 241 | 229 | 272 |
| Month-Year | Jan-07 | Feb-07 | Mar-07 | Apr-07 | May-07 | Jun-07 | Jul-07 | Aug-07 | Sep-07 | Oct-07 | Nov-07 | Dec-07 |
| Number of Tractor Sold | 247 | 261 | 330 | 362 | 385 | 340 | 370 | 381 | 299 | 266 | 239 | 281 |
| Month-Year | Jan-08 | Feb-08 | Mar-08 | Apr-08 | May-08 | Jun-08 | Jul-08 | Aug-08 | Sep-08 | Oct-08 | Nov-08 | Dec-08 |
| Number of Tractor Sold | 257 | 250 | 329 | 350 | 393 | 370 | 423 | 410 | 326 | 289 | 270 | 321 |
| Month-Year | Jan-09 | Feb-09 | Mar-09 | Apr-09 | May-09 | Jun-09 | Jul-09 | Aug-09 | Sep-09 | Oct-09 | Nov-09 | Dec-09 |
| Number of Tractor Sold | 305 | 310 | 374 | 414 | 454 | 441 | 510 | 486 | 393 | 345 | 315 | 389 |
| Month-Year | Jan-10 | Feb-10 | Mar-10 | Apr-10 | May-10 | Jun-10 | Jul-10 | Aug-10 | Sep-10 | Oct-10 | Nov-10 | Dec-10 |
| Number of Tractor Sold | 358 | 368 | 444 | 482 | 534 | 524 | 578 | 567 | 447 | 386 | 360 | 428 |
| Month-Year | Jan-11 | Feb-11 | Mar-11 | Apr-11 | May-11 | Jun-11 | Jul-11 | Aug-11 | Sep-11 | Oct-11 | Nov-11 | Dec-11 |
| Number of Tractor Sold | 397 | 400 | 498 | 536 | 596 | 591 | 651 | 654 | 509 | 437 | 406 | 470 |
| Month-Year | Jan-12 | Feb-12 | Mar-12 | Apr-12 | May-12 | Jun-12 | Jul-12 | Aug-12 | Sep-12 | Oct-12 | Nov-12 | Dec-12 |
| Number of Tractor Sold | 428 | 423 | 507 | 536 | 610 | 609 | 687 | 707 | 509 | 452 | 412 | 472 |
| Month-Year | Jan-13 | Feb-13 | Mar-13 | Apr-13 | May-13 | Jun-13 | Jul-13 | Aug-13 | Sep-13 | Oct-13 | Nov-13 | Dec-13 |
| Number of Tractor Sold | 454 | 455 | 568 | 610 | 706 | 661 | 767 | 783 | 583 | 513 | 481 | 567 |
| Month-Year | Jan-14 | Feb-14 | Mar-14 | Apr-14 | May-14 | Jun-14 | Jul-14 | Aug-14 | Sep-14 | Oct-14 | Nov-14 | Dec-14 |
| Number of Tractor Sold | 525 | 520 | 587 | 710 | 793 | 749 | 871 | 848 | 640 | 581 | 519 | 605 |

R-Script:



R-Console:

> at1 <- as.vector(at)

> MonthYear<-at$Month.Year

> TSold<-at$Number.of.Tractor.Sold

> data<-ts(TSold,start = c(2003,1),frequency = 12)

> View(data)

> plot(data,xlab="Years",ylab="No. of tractors sold",main="Company's Sale Over The Decade",col="red")

>

> a <- sma(data,order=12,h=12,silent = FALSE)

> seasonplot(data,12,col=rainbow(12))

> a1 <- a$fitted

> seasonplot(a1,12,col=rainbow(12))

>

> plot(diff(data),main="First difference",xlab="Years",ylab="Differences")

> plot(log10(data),main="Log transformed",xlab="Years",ylab="Log")

> plot(diff(log10(data)),main="First difference of log transform",xlab="Years",ylab="Differences")

> par(mfrow=c(1,2))

> acf(ts(diff(log10(data))),main="ACF of diff log")

> pacf(ts(diff(log10(data))),main="PACF of diff log")

> fore <- require(forecast)

> ARIMAfit = auto.arima(log10(data),approximation = FALSE,trace=FALSE)

> summary(ARIMAfit)

Series: log10(data)

ARIMA(0,1,1)(0,1,1)[12]

Coefficients:

ma1 sma1

-0.4047 -0.5529

s.e. 0.0885 0.0734

sigma^2 estimated as 0.0002571: log likelihood=354.4

AIC=-702.79 AICc=-702.6 BIC=-694.17

Training set error measures:

ME RMSE MAE MPE MAPE MASE ACF1

Training set 0.0002410698 0.01517695 0.01135312 0.008335713 0.4462212 0.2158968 0.01062604

> par(mfrow=c(1,1))

> pred <- predict(ARIMAfit,n.ahead = 36)

>

> pred$pred

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov

2015 2.754168 2.753182 2.826608 2.880192 2.932447 2.912372 2.972538 2.970585 2.847264 2.797259 2.757395

2016 2.796051 2.795065 2.868491 2.922075 2.974330 2.954255 3.014421 3.012468 2.889147 2.839142 2.799278

2017 2.837934 2.836948 2.910374 2.963958 3.016213 2.996138 3.056304 3.054351 2.931030 2.881025 2.841161

Dec

2015 2.825125

2016 2.867008

2017 2.908891

> plot(data,type='l',xlim=c(2004,2018),ylim=c(1,1600),xlab = 'Year',ylab = 'Tractor Sales')

> lines(10^(pred$pred),col='red')

> lines(10^(pred$pred+2\*pred$se),col='green')

> lines(10^(pred$pred-2\*pred$se),col='green')

>

> par(mfrow=c(1,2))

> acf(ts(ARIMAfit$residuals))

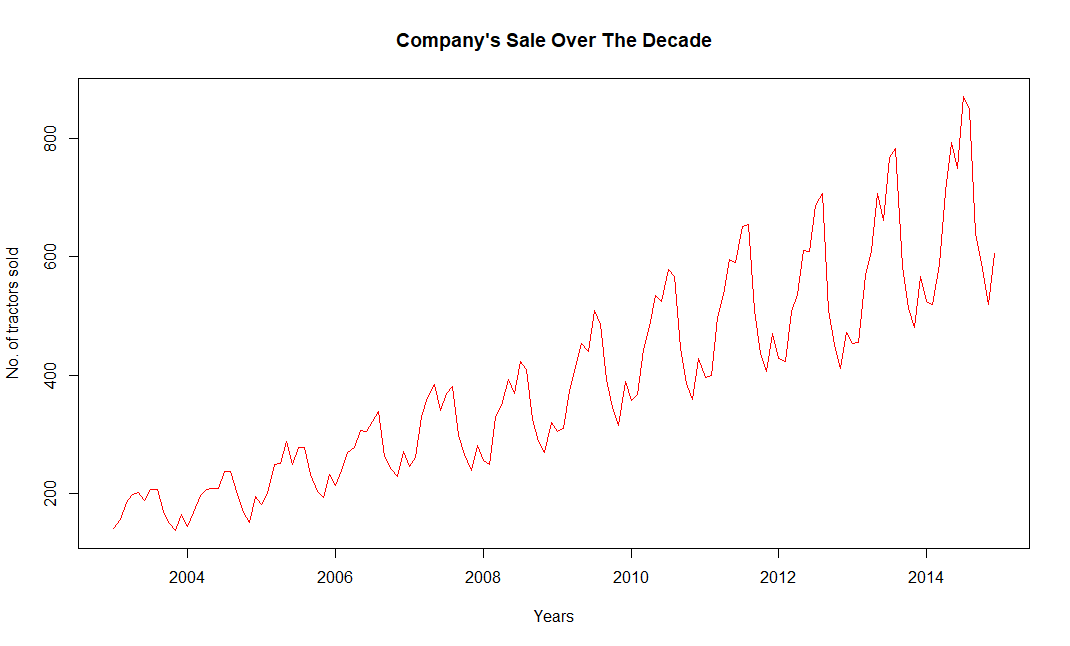
> pacf(ts(ARIMAfit$residuals))

STEP-BY-STEP ANALYSIS:

As a part of this Project, we assume:

1. The tractor manufacturing company’s sale depends on internal factors only and is not dependent on Market Competition.
2. Since we’ve taken dataset of about a decade, it is safe to assume that there is no or next to no Cyclic Variation in the Dataset.

# Time Series Decomposition-



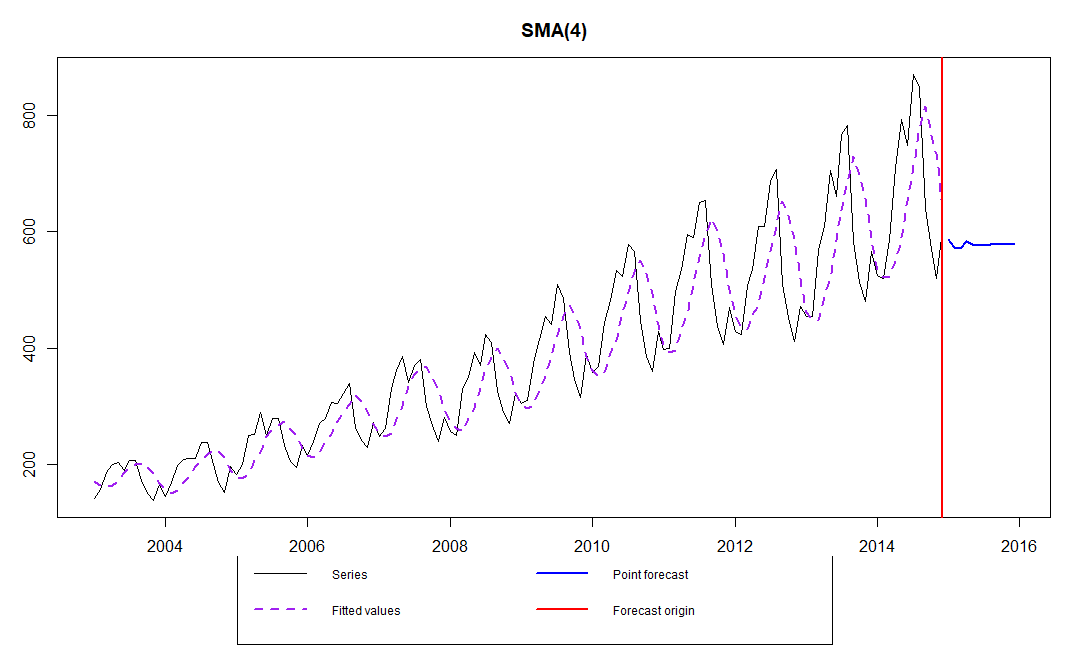
We came across a Time Series, showing Continual Changes over time giving an overall impression of Haphazard Movement.

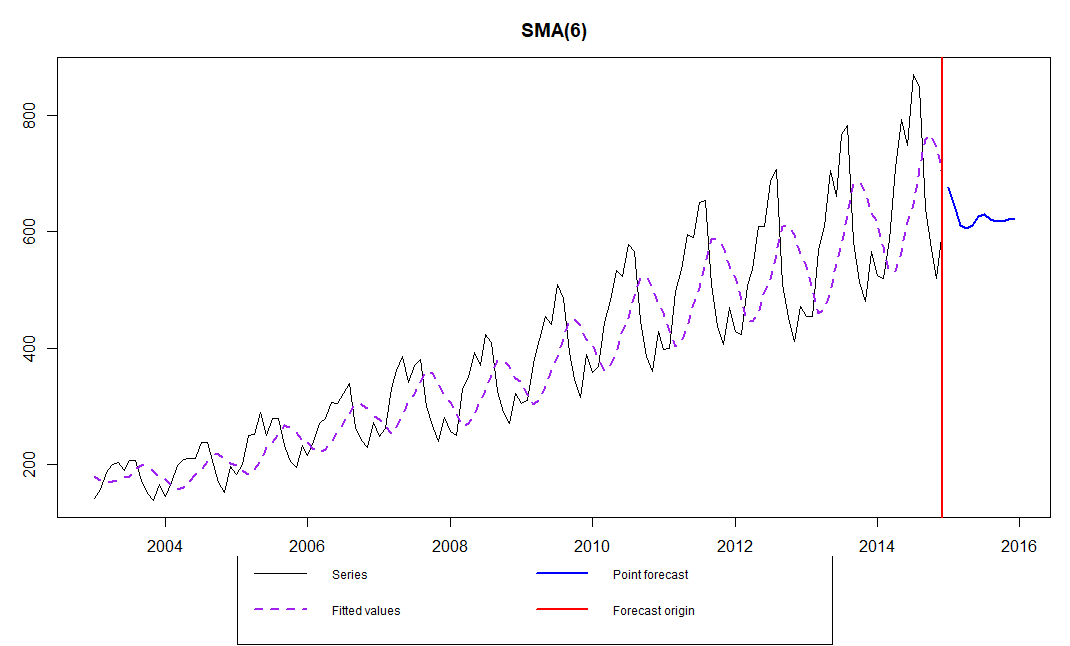
Now, We start with time series decomposition of this data to understand underlying patterns for tractor sales.

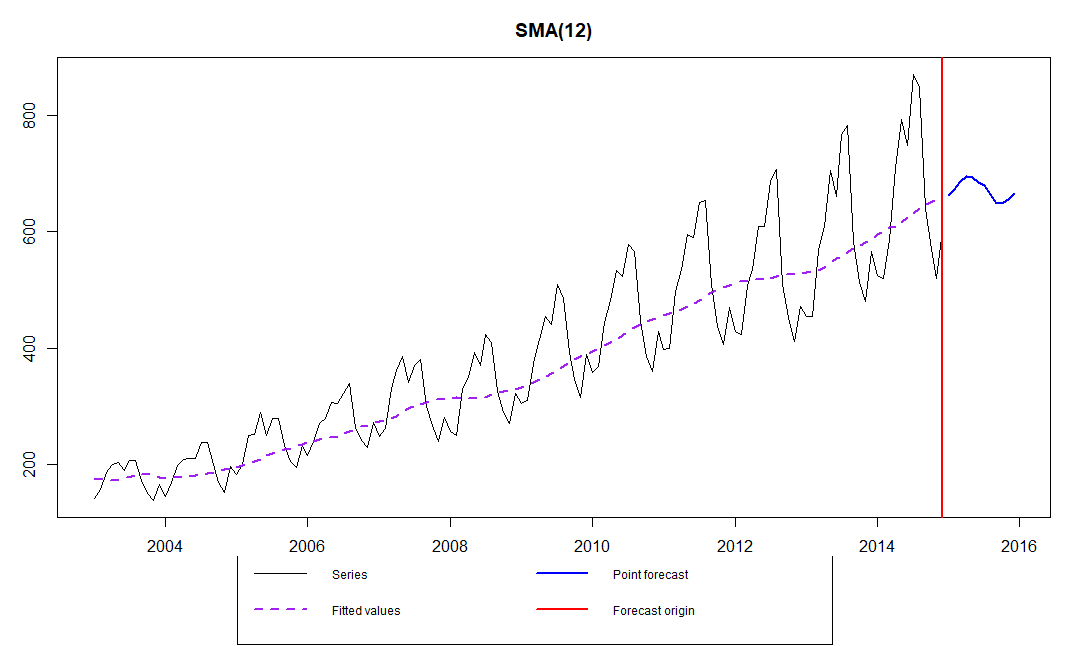
* **Trend** –  overall direction of the series i.e. upwards
* **Seasonality** – monthly or quarterly patterns
* **Cycle** –  long term business cycles
* **Irregular remainder**– random noise left after extraction of all the components

Now, to begin with We decipher trends embedded in the above tractor sales time series. One of the commonly used procedures to do so is moving averages. The idea with moving average is to remove all the zigzag motion from the time series to produce a steady trend through averaging adjacent values of a time period.

Now, We remove wrinkles from our time series using moving average. We will take moving average of different time periods i.e. 4, 6 and 12 months as shown below. Here, moving average is as shown below:



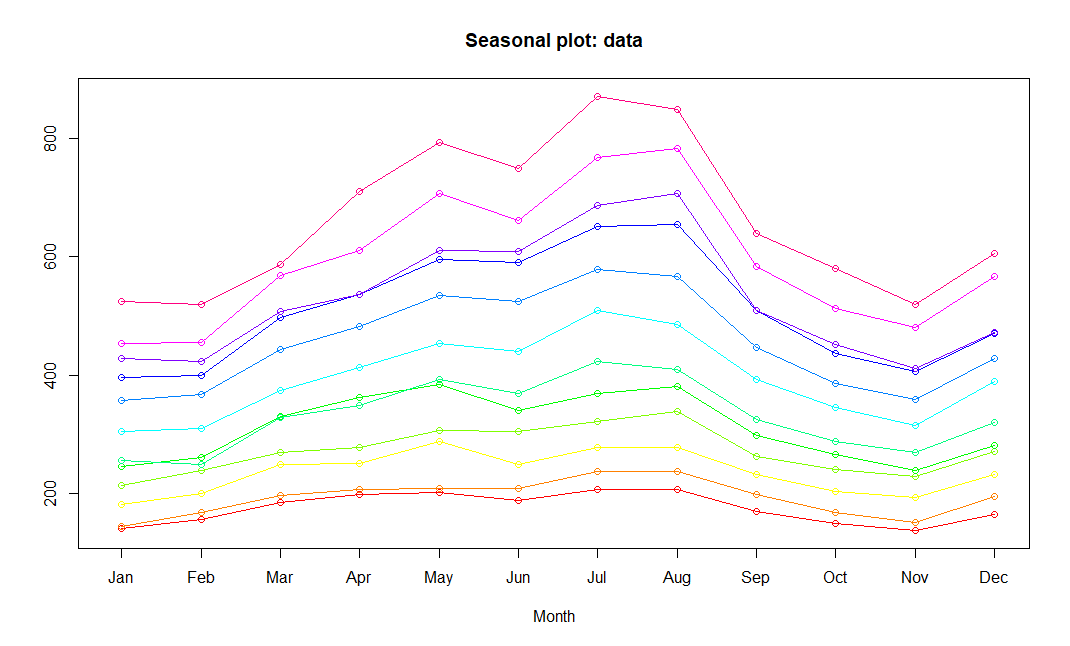




As it is seen in the above plots, 12-month moving average could produce a wrinkle free curve as desired. This on some level is expected since we are using month-wise data for our analysis and there is expected monthly-seasonal effect in our data. (Here, we used “sma” function in R to get an idea about the next year’s prediction).

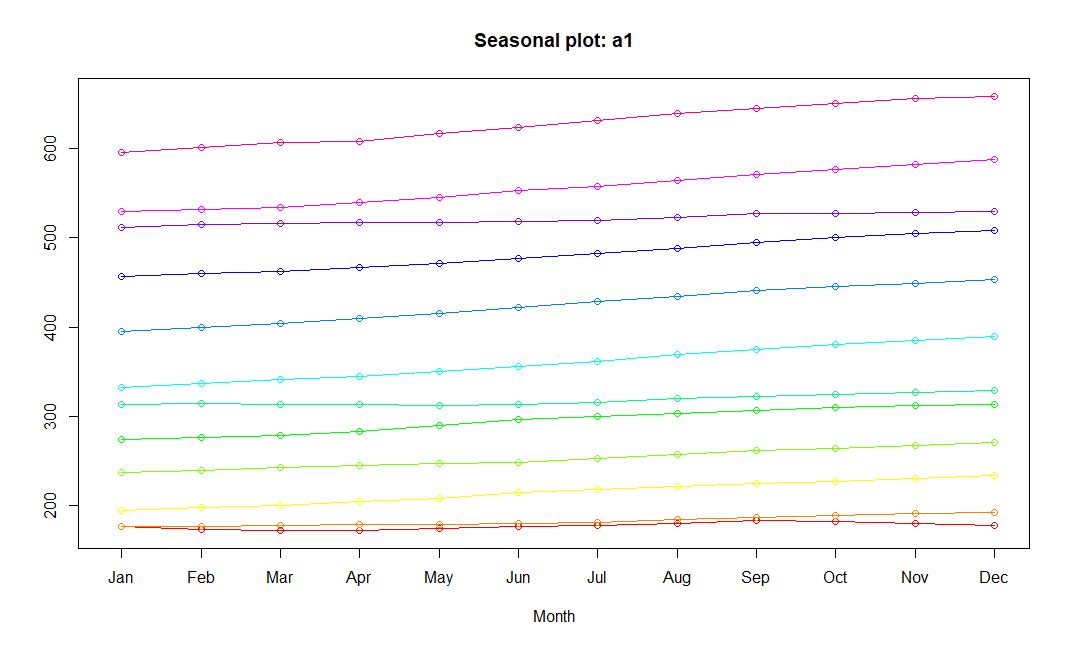
Now, we Decipher the Seasonal Component:

The first thing to do is to see how number of tractors sold vary on a month on month basis. We will plot a stacked annual plot to observe seasonality in our data.



As seen, there is a fairly consistent month on month variation with July and August as the peak months for tractor sales.

The monthly seasonal components are average values for a month after removal of trend.



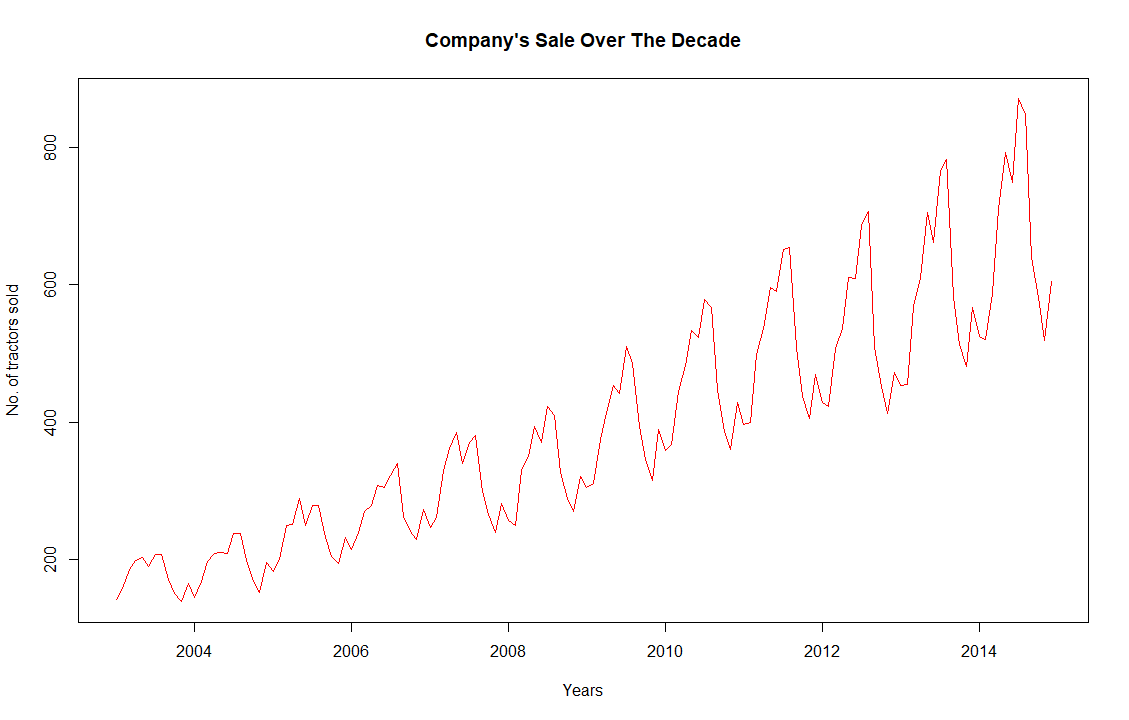
Now, we start with ARIMA modelling for forecasting. ARIMA is an abbreviation for  ***A***uto-***R***egressive ***I***ntegrated ***M***oving ***A***verage.

A convenient notation for ARIMA model is ARIMA(p,d,q). Here p,d, and q are the levels for each of the AR, I, and MA parts. Each of these three parts is an effort to make the final residuals display a white noise pattern (or no pattern at all).

**Step 1: Plot tractor sales data as time series**

**Step 2: Difference data to make data stationary on mean (remove trend)**

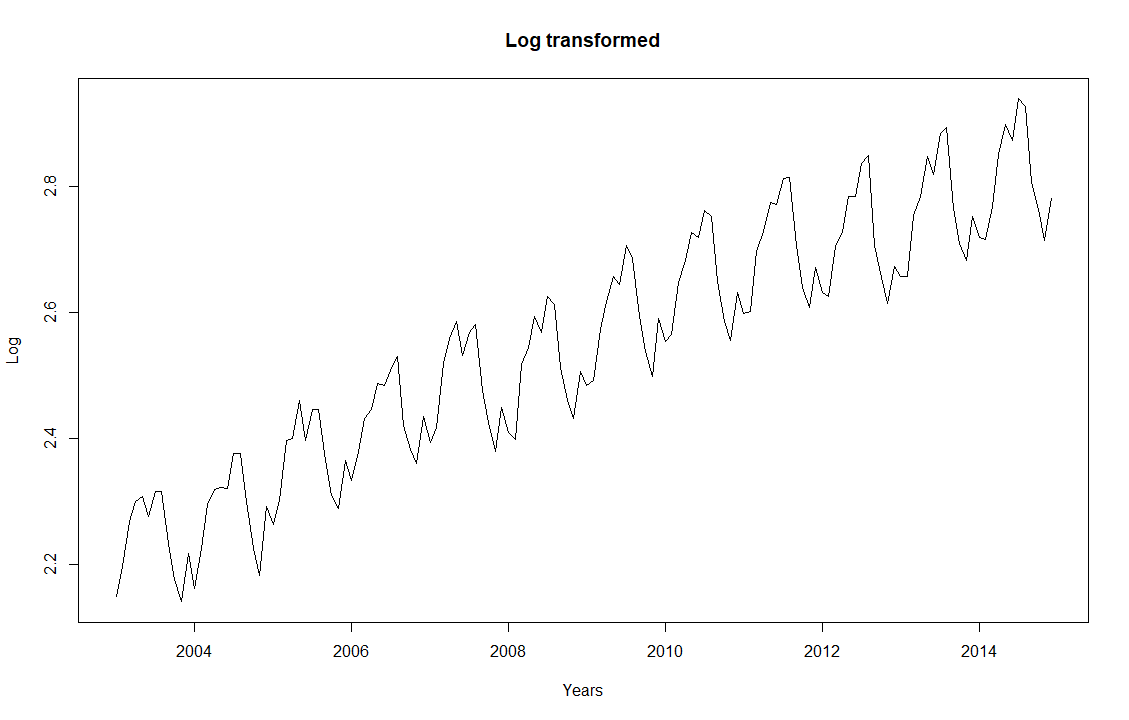
The next thing to do is to make this time-series stationary. This to remove the upward trend through 1st order differencing the series.



The above series is not stationary on Mean and Variance i.e. variation in the plot is increasing as we move towards the right of the chart. We need to make the series stationary on mean and variance to produce reliable forecasts through ARIMA models.

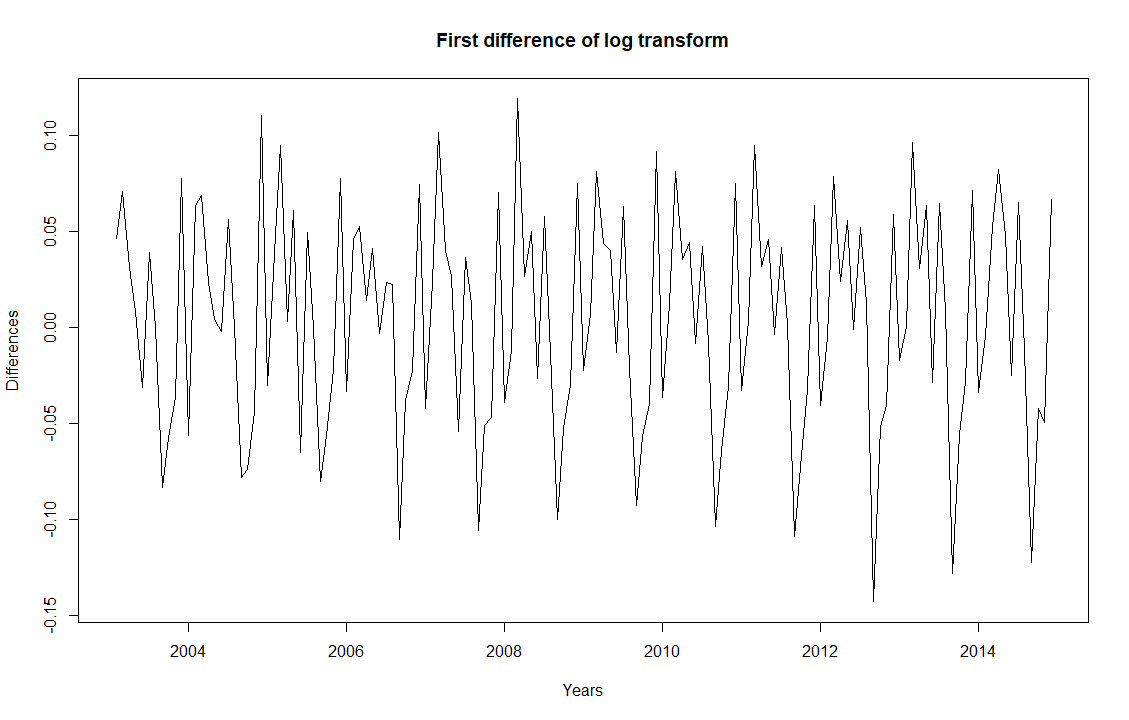
**Step 3: log transform data to make data stationary on variance**

One of the best ways to make a series stationary on mean and variance is through transforming the original series through log transform. We will go back to our original tractor sales series and log transform it to make it stationary on mean and variance.



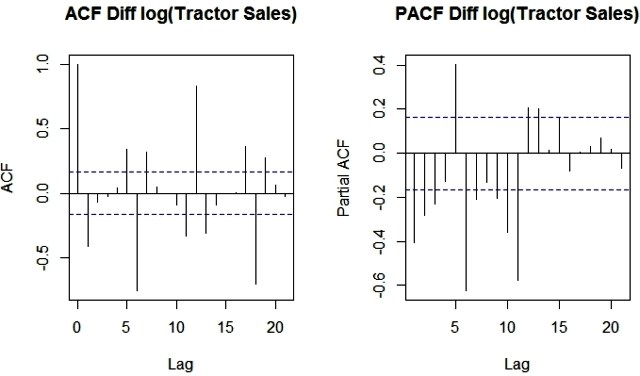
**Step 4: Difference log transform data to make data stationary on both mean and variance**

Now, we did First-order Differencing for log transformed series (and plotted) to reconfirm if the series is actually stationary on both mean and variance.



**Step 5: Plot ACF and PACF to identify potential AR and MA model**

Now, we create autocorrelation factor (ACF) and partial autocorrelation factor (PACF) plots to identify patterns in the above data which is stationary on both mean and variance. The idea is to identify presence of AR and MA components in the residuals.



Since, there are enough spikes in the plots outside the insignificant zone (dotted horizontal lines) we can conclude that the residuals are not random. This implies that there is juice or information available in residuals to be extracted by AR and MA models. Also, there is a seasonal component available in the residuals at the lag 12 (represented by spikes at lag 12). This makes sense since we are analyzing monthly data that tends to have seasonality of 12 months because of patterns in tractor sales.

**Step 6: Identification of best fit ARIMA model**

> summary(ARIMAfit)

Series: log10(data)

ARIMA(0,1,1)(0,1,1)[12]

Coefficients:

ma1 sma1

-0.4047 -0.5529

s.e. 0.0885 0.0734

sigma^2 estimated as 0.0002571: log likelihood=354.4

AIC=-702.79 AICc=-702.6 BIC=-694.17

Training set error measures:

ME RMSE MAE MPE MAPE MASE

Training set 0.0002410698 0.01517695 0.01135312 0.008335713 0.4462212 0.2158968

ACF1

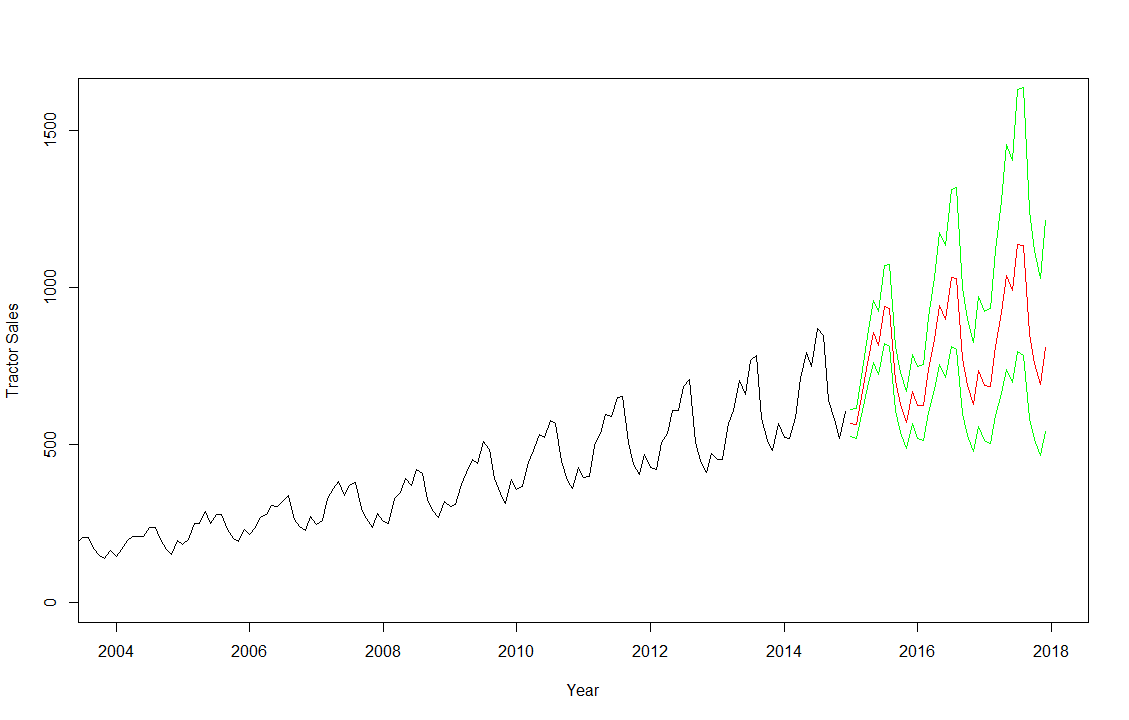
Training set 0.01062604

Our model has I (or integrated) component equal to 1. This represents differencing of order 1. There is additional differencing of lag 12 in the above best fit model. Moreover, the best fit model has MA value of order 1. Also, there is seasonal MA with lag 12 of order 1.

**Step 7: Forecast sales using the best fit ARIMA model**

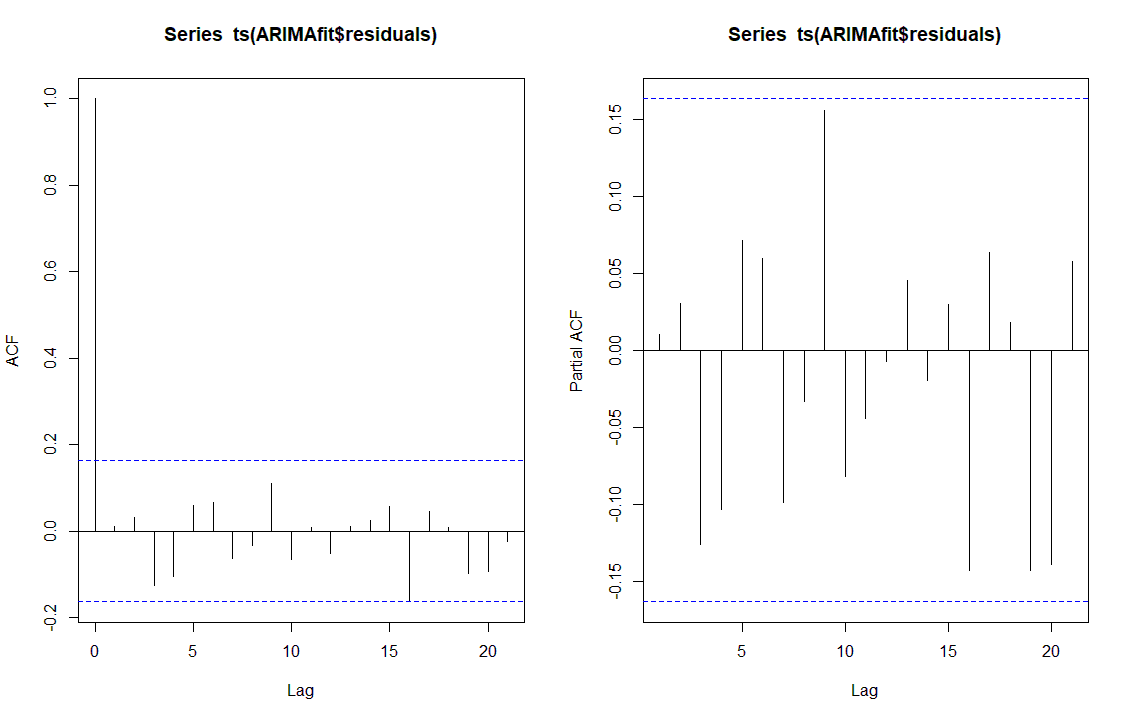
The next step is to predict tractor sales for next 3 years i.e. for 2015, 2016, and 2017 through the above model.

The following is the output with forecasted values of tractor sales in blue. Also, the range of expected error is displayed with Green line and predicted values with Red.



**Step 8: Plot ACF and PACF for residuals of ARIMA model to ensure no more information is left for extraction**

Finally, let’s create an ACF and PACF plot of the residuals of our best fit ARIMA model i.e. ARIMA(0,1,1)(0,1,1)[12].



These are the log10 Predicted values:

pred

$pred

Jan Feb Mar Apr May Jun Jul Aug Sep

2015 2.754168 2.753182 2.826608 2.880192 2.932447 2.912372 2.972538 2.970585 2.847264

2016 2.796051 2.795065 2.868491 2.922075 2.974330 2.954255 3.014421 3.012468 2.889147

2017 2.837934 2.836948 2.910374 2.963958 3.016213 2.996138 3.056304 3.054351 2.931030

Oct Nov Dec

2015 2.797259 2.757395 2.825125

2016 2.839142 2.799278 2.867008

2017 2.881025 2.841161 2.908891

$se

Jan Feb Mar Apr May Jun Jul

2015 0.01603508 0.01866159 0.02096153 0.02303295 0.02493287 0.02669792 0.02835330

2016 0.03923008 0.04159145 0.04382576 0.04595157 0.04798329 0.04993241 0.05180825

2017 0.06386474 0.06637555 0.06879478 0.07113179 0.07339441 0.07558934 0.07772231

Aug Sep Oct Nov Dec

2015 0.02991723 0.03140337 0.03282229 0.03418236 0.03549035

2016 0.05361850 0.05536960 0.05706700 0.05871534 0.06031866

2017 0.07979828 0.08182160 0.08379608 0.08572510 0.08761165

The Actual Predicted Values Are:

> 10^pred$pred

Jan Feb Mar Apr May Jun Jul Aug

2015 567.7645 566.4765 670.8226 758.9138 855.9482 817.2827 938.7239 934.5120

2016 625.2464 623.8280 738.7384 835.7481 942.6065 900.0265 1033.7626 1029.1243

2017 688.5479 686.9859 813.5300 920.3613 1038.0383 991.1474 1138.4233 1133.3154

Sep Oct Nov Dec

2015 703.5005 626.9879 571.9988 668.5363

2016 774.7246 690.4657 629.9094 736.2206

2017 853.1596 760.3701 693.6830 810.7573