# Data Analytics Foundations

# (ECE 9063A)

# Assignment 1: Forecasting

Submitted by:
Student Name: Shivani Solanki
Student ID: 251056782
(ssolank8@uwo.ca)

# Introduction

**Data Analytics** is analysing the both qualitative and quantitative data using certain tools and techniques and turn it into meaningful insights to support decision making and process improvement.

**Forecasting** is a process of predicting a future event based on the data available, contributing factors and its affects with the assumptions that random fluctuations in the past to be ignored and the existing trends will continue.

There are several forecasting models like ARIMA , Linear Regression and Support Vector Regression which will be discussed in this report.

## Problem Description

In order to optimize sales and increase profits, the Cooperates need to establish a balance between their working capital and inventories. Inaccurate demand forecasting or absence of forecasting at all may lead to inventory pile up which will apparently result into loss of working capital, reduced financial liquidity, increased debts, dead stocks and so on. This reduced money circulation will directly affect the economic growth or the market value of that enterprise and on a wider scale, may affect the GDP of the region.

Hence by using various forecasting techniques and models, the future values of Inventory to sales ratios will be predicted to support Sales department and the Supply Chain department to formulate necessary policies to keep this ratio value low, liquidate inventory, capture and understand the customer demands to avoid order loss due to underestimation and to avoid surplus inventory (financial drain) due to overestimation.

## Dataset

The dataset used is time series and shows the inventories to sales ratios which represents the relationship between inventory available at the end of the month and the sales made that month. If this ratio increases with the increase in inventory and there isn't any rise in sales, then that means more inventory is getting accumulated without increased sales. For example, Ratio 1.5 shows that enough inventory is available on hand to fulfil the one and half months of sales.

The dataset consists of four attributes: realtime_start, value, date and realtime_end.

- Since the realtime_start and realtime_end are always the current dates due to daily updating, hence these two attributes have been dropped.
- Only attributes value and date will be used for forecasting.
- The attribute value represents the monthly sales to inventory ratios and date represents the month-for which the value is calculated.
- There are total 319 observations from the year 1992 till 2018 with monthly frequency.

The link for the data source can be found below:-

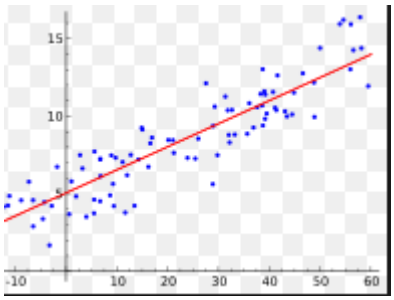
https://www.kaggle.com/census/total-business-inventories-and-sales-data

# Brief Overview of the Time Series Models

## 1. Time Series Simple Linear Regression

In this model , there is a single predictor variable that has a linear relationship with the forecast variable and can be represented as below:-

$y_t=\beta_0+\beta_1 x_t+\varepsilon_t$




y is the forecast variable and x is the predictor variable.
β0 and β1 are the coefficients. β0 is the value of y intercept when x=0 and β1 gives the slope.
$\varepsilon_t$ is the error accounts for the parameters other than x, which affect the predicted vale y.

There are few assumptions to be taken into account while using this model:-

1. It is assumed that linear equation ins satisfied by both x and y.
2. x is a random variable.
3. Errors are not autocorrelated.
4. Errors have zero mean.
5. Errors are unrelated to x.

## 2. Holt-Winters' Forecasting Model

It is a time series forecasting model which deals with the three aspects of the time i.e. value, trend and seasonality with smoothing parameters α,β and γ respectively. α,β,γ does exponentially smoothing of the value, trend and seasonality, hence Holt-Winters' forecasting model also known as triple exponential smoothing. If the smoothing parameters is near to zero, then the prediction is done on past values and when it is near to one, the prediction is done on the recent past value.
It employs two methods additive and multiplicate to handle the seasonal component. When the seasonal variations are almost same throughout the series, then additive model is used. When these variations change proportionally with the value of the series, then multiplicative model is used.

Additive method can be expressed as :-

$\hat{y}_{t+h|t}=\ell_t+hb_t+s_{t+h-m(k+1)}$

$\ell_t=\alpha(y_t-s_{t-m})+(1-\alpha)(\ell_{t-1}+b_{t-1})$

$b_t=\beta_*(\ell_t-\ell_{t-1})+(1-\beta_*)b_{t-1}$

$s_t=\gamma(y_t-\ell_{t-1}-b_{t-1})+(1-\gamma)s_{t-m},$

Multiplicative method can be expressed as:-

$\hat{}y_{t+h|t}=(\ell_t+hb_t)s_{t+h-m(k+1)}$

$$\ell_t = \alpha y_t / s_{t-m} + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*) b_{t-1}$$
$$s_t = \gamma y_t / (\ell_{t-1} + b_{t-1}) + (1-\gamma) s_{t-m}$$

## 3. ARIMA model (Autoregressive Integrated Moving Average)

It is a time series forecasting approach which is a combination of Autoregressive model and moving average model.

Autoregressive model:

The forecasting is done based on the past values. If $Y_t$ is the current value, it will depend only on its past values $Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots$ . The AutoRegressive model which depend on p of its past values, AR(p) is represented as below:-

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

Moving Average Model:

In this model, the forecasting is done by taking the past forecast error terms. When a current value is regressed with its immediate past value, we get an error ε. For example, if $Y_t$ is regressed with $Y_{t-1}$, we get an error $\varepsilon_t$. The moving average model of order q, MA(q) can be represented as below:-

$$Y_t = \beta_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots + \theta_q \varepsilon_{t-q}$$

ARMA model
Autoregression(p) + Moving Average model(q) = ARMA(p,q) model.
ARMA(p,q) is represented as below:-

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots + \theta_q \varepsilon_{t-q}$$

A series is said to be stationary if it is has no trend, no seasonality and has constant variance. If the time series is differentiated to achieve stationarity, it is called integration. It is denoted as I(d) where d is the order of differentiation.

**Autoregression(p) + differencing(d)+Moving Average model(q) = ARIMA(pdq) model. It is represented as below:**

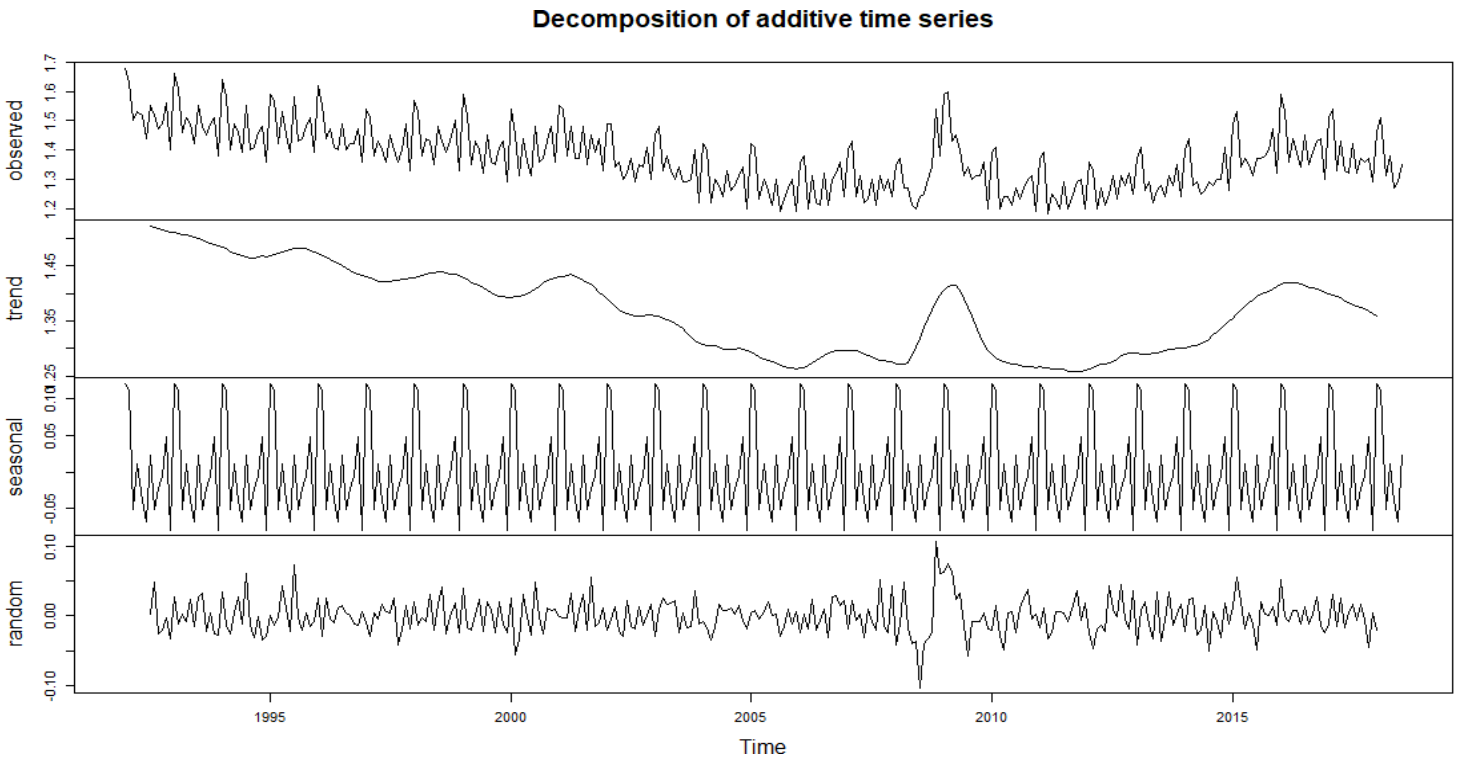
$$Y_t' = \beta_0 + \phi_1 Y'_{t-1} + \phi_2 Y'_{t-2} + \ldots + \phi_p Y'_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots + \theta_q \varepsilon_{t-q}$$

$Y_t'$ is the differentiated series.

# Algorithms

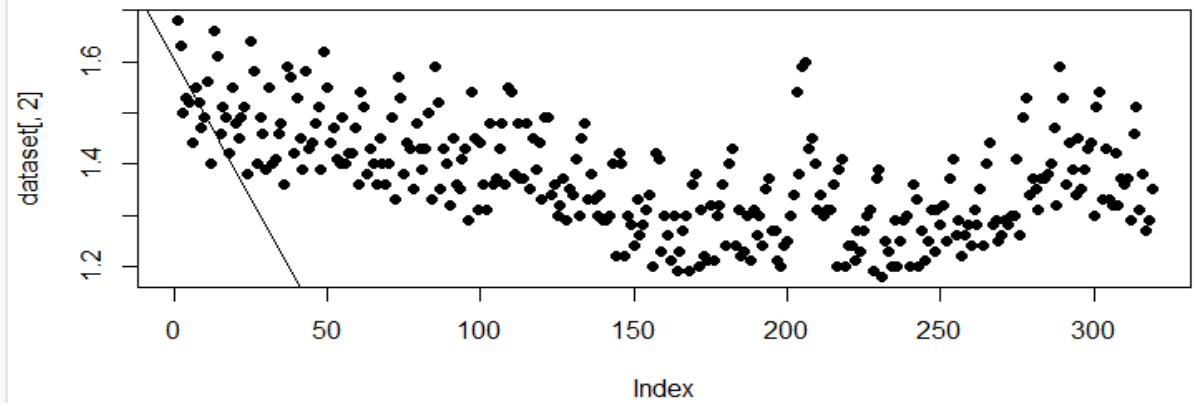
## Time Series Simple Linear Regression Algorithm

1. The complete available data is converted into time-series, plotted and is checked for the presence of seasonal and trend components by decomposing.

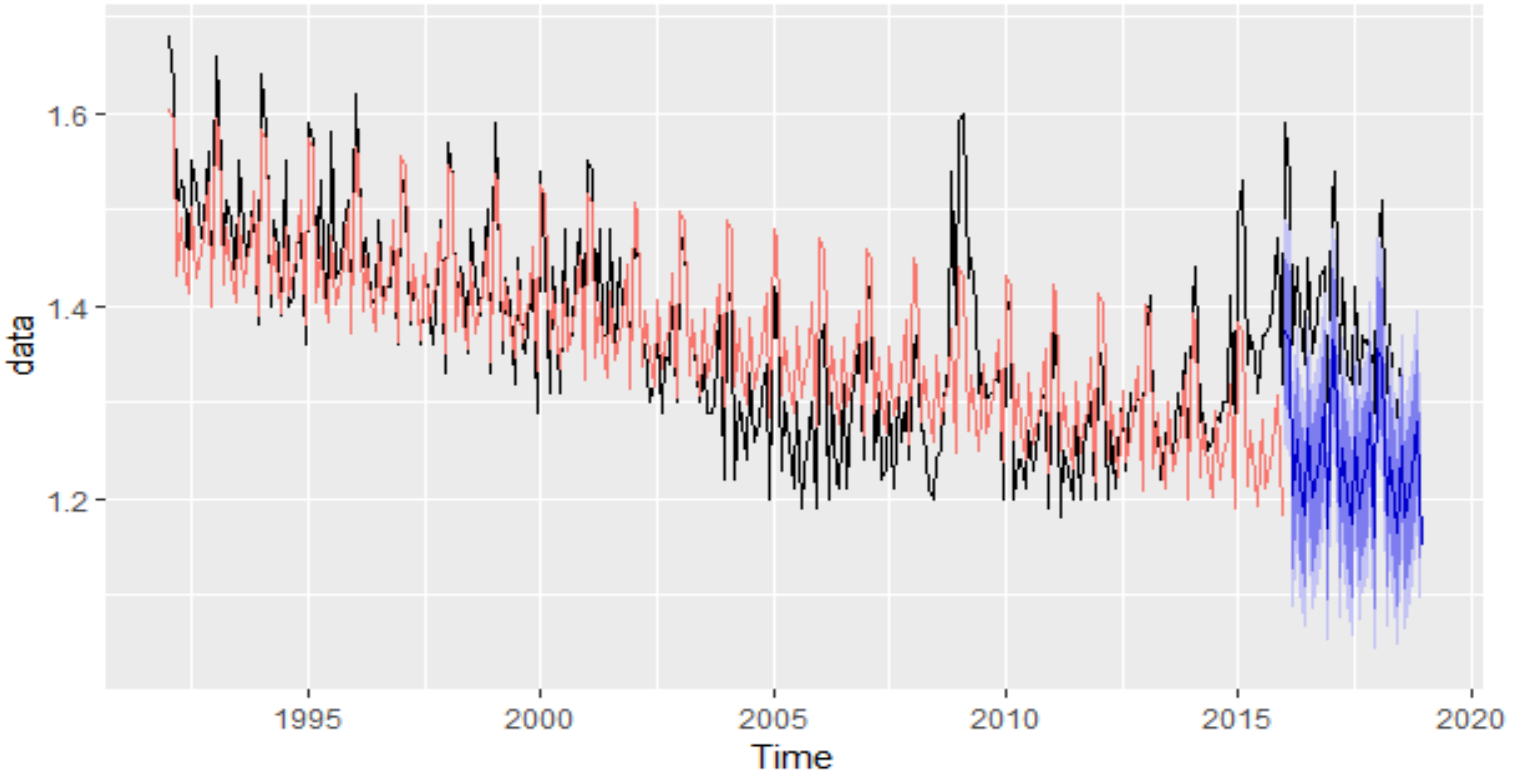



2. The data is split the into training set and the test set. Since this is time series, the sample set cannot be chosen randomly. So, we will take data from year 1992 to 2015 as trainset and 2016-2017 will be the test set.
3. By using tslm function, the model is built on the training set. As seasonality and trend both are present in the series, so both are taken into account while building the model. The tslm summary gives
    - the y intercept when x=0
    - the seasonal coefficients to make the Time Series seasonally adjusted.
    - the coefficient to remove the trend.

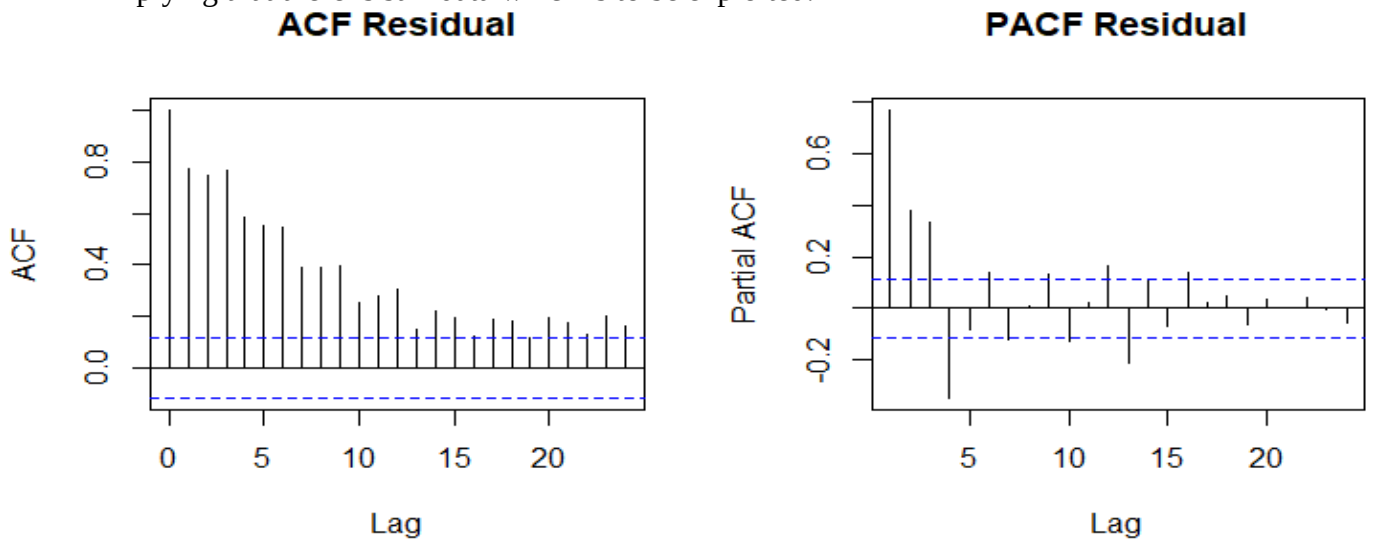
*(R warning > plotted only using the first two of 13 regression coefficients)*




4. The forecasting is done for next three years.
5. In the output plot below, Black plot shows the actual data, red shows the TSL model and blue shows the predicted value.




6. With time series data, it is possible that the predicted value may have the similar value to the past values. Therefore when using regression with time series , we can find autocorrelation in the residuals implying that there is still data which is to be exploited.

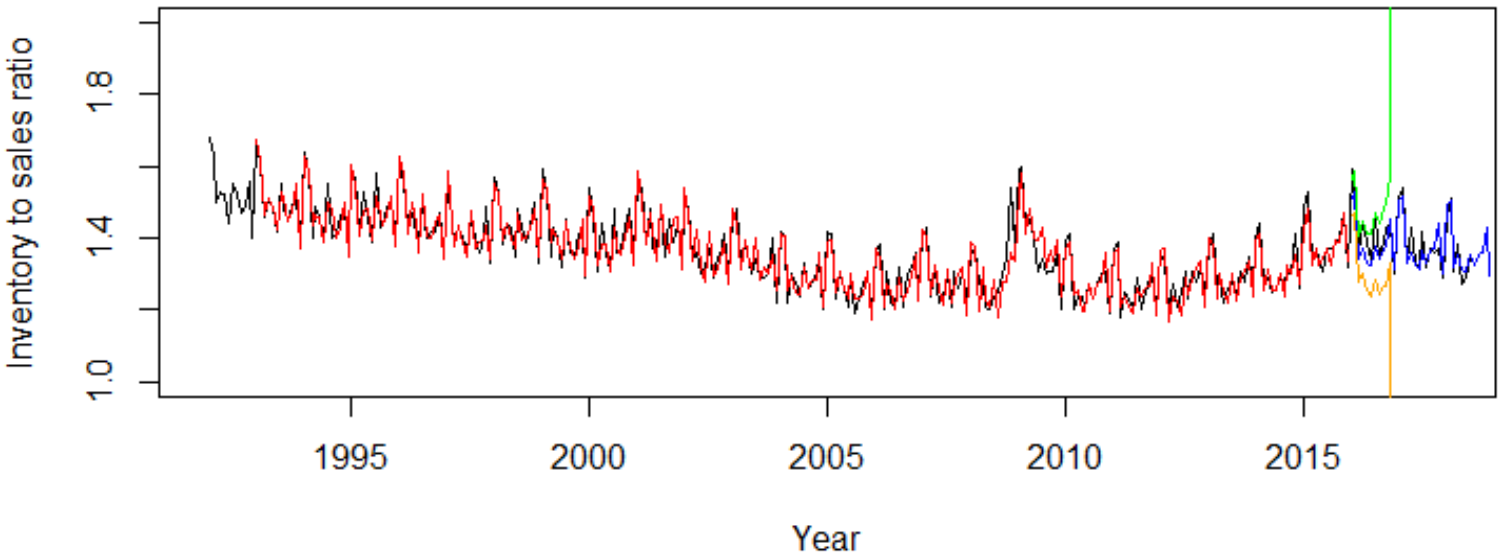



7. For accuracy measurement, Root mean squared error (RMSE) is measured.

## HoltWinters Algorithm

1. The complete available data is converted into time-series, plotted and is checked for the presence of seasonal and trend components by decomposing.
2. The data is split into trainset which will be from year 1992-2015 and rest will be test set i.e. from 2016-2018.
3. To build the model, we will then use HoltWinters function on the training set. The seasonality component is set to Additive by default and can be changed to Multiplicative manually.
   Due to presence of both trend and season, we have used HW function with Multiplicative model.

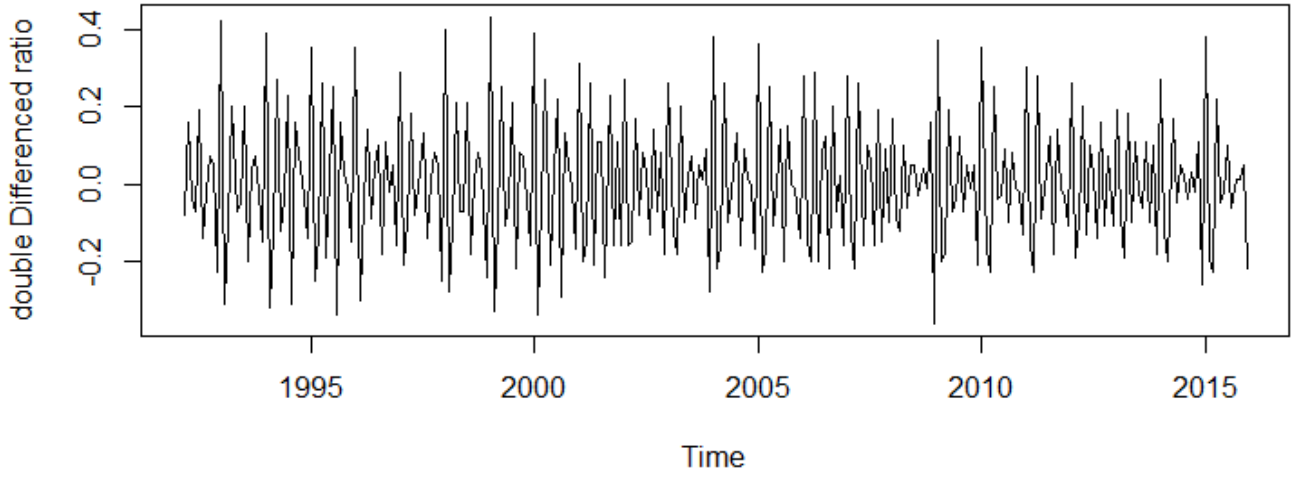
4. The model summary will give α,β,γ values for smoothing. It also gives the a and b coefficients which will be multiplied to the seasonal components to seasonally adjust the time series.
5. Now the forecast will be done for next three year (2016-2018).
6. In the graph below, black line specifies the actual data, red specifies the HoltWinters model built on training set, blue specifies the predicted value, green gives the upper limit of the expected error and orange gives the lower limit of the expected error.



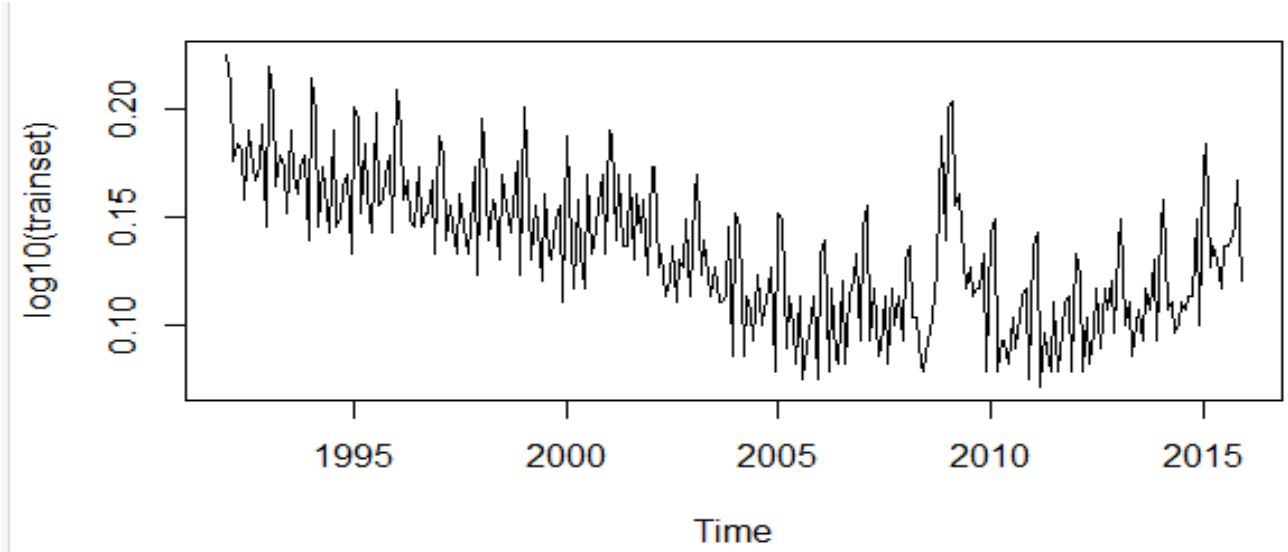

7. For accuracy measurement, Root mean squared error (RMSE) is measured.

## ARIMA Algorithm :

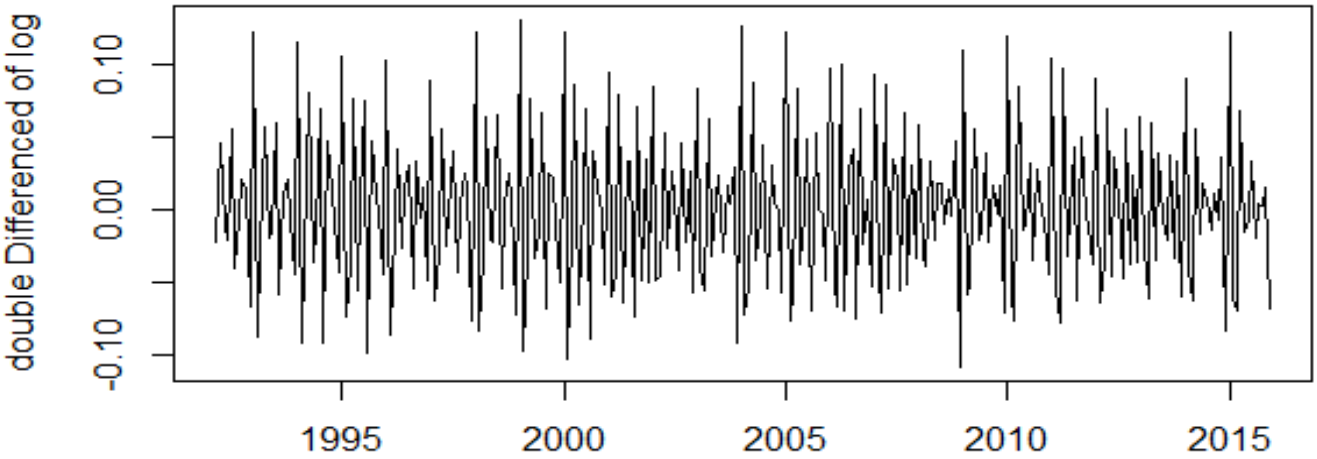
1. The complete available data is converted into time-series, plotted and is checked for the presence of seasonal and trend components by decomposing.
2. The data is split into trainset which will be from year 1992-2015 and rest will be test set i.e. from 2016-2018.
3. To make the series stationary, the trend is removed by 2$^{nd}$ order differencing.




4. To get the stationarity on variance, the log of the trainset is taken.




5. To see the stationarity on mean and variance , we will take 2nd order differentiation of log series.

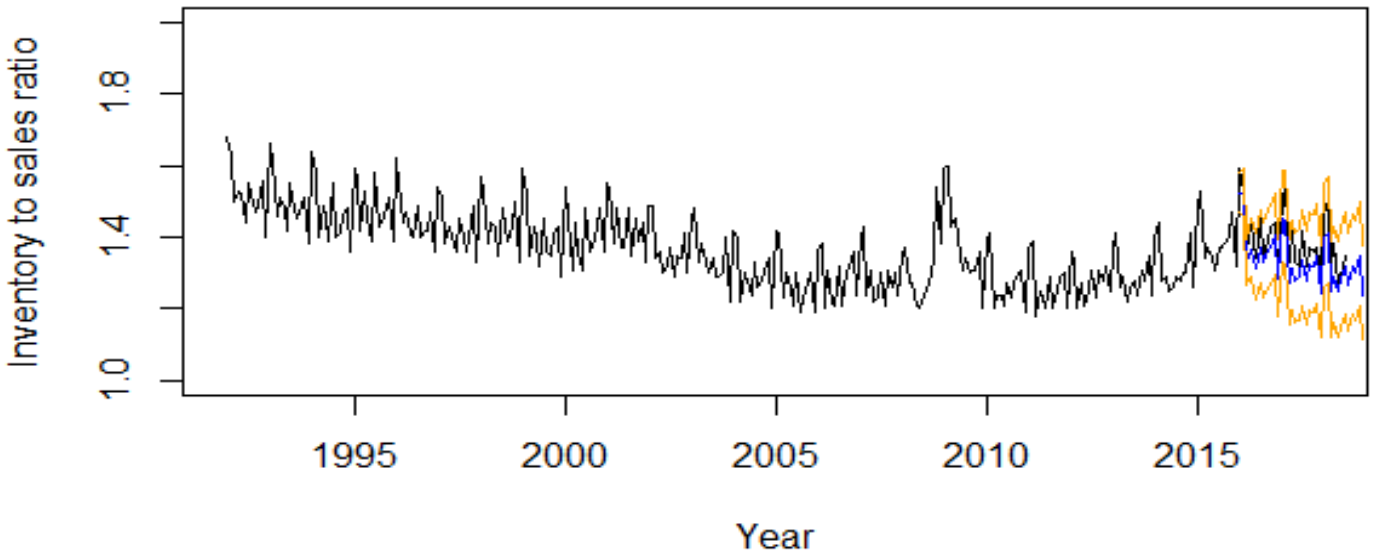



6. To build the ARIMA model, auto.arima function is applied on the Log of trainset.
   Below is the model suggested. ARIMA(2,0,3) being the non-seasonal and (2,1,2) are seasonal components.

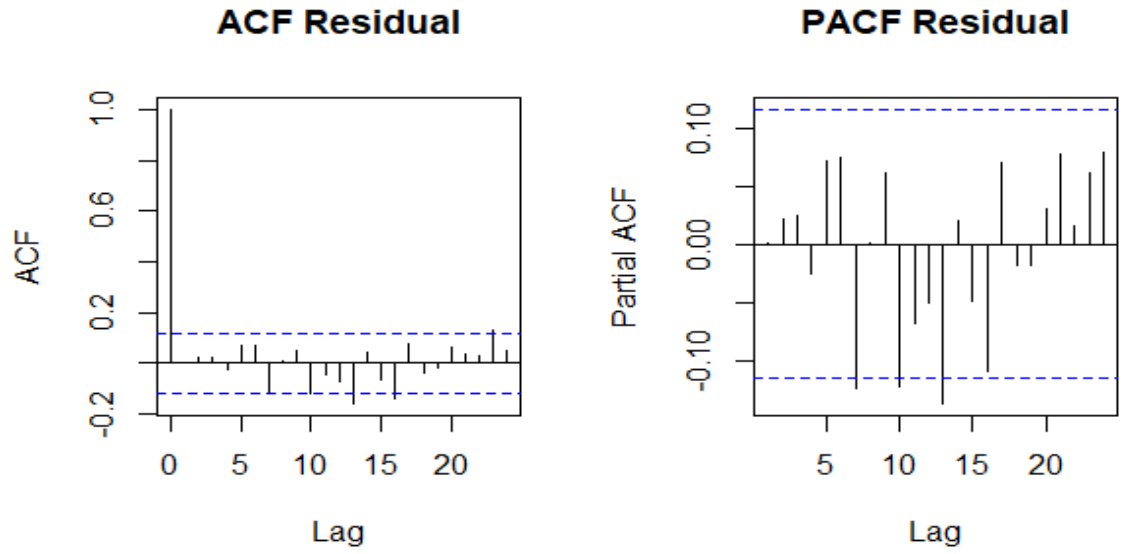
```
Series: log10(trainset)
ARIMA(2,0,3)(2,1,2)[12]

Coefficients:
         ar1     ar2     ma1      ma2     ma3    sar1     sar2     sma1    sma2
      0.5167  0.3442  0.1239  -0.0253  0.3964  0.8698  -0.6254  -1.4359  0.7355
s.e.  0.1490  0.1445  0.1393   0.0668  0.0586  0.0844   0.0667   0.1017  0.0811
```

7. We will forecast for the next 3 years 2016,2017 and 2018. The Values and plot will look as below. Black line is the actual data, blue line is predicted value and orange lines represents expected error range.




8. ACF and PACF of residuals are plotted. Most of the data is captured but still some spikes are out of the blue line zone that means there are some non-random residuals for which the data is not exploited.




9. For accuracy measurement, Root mean squared error (RMSE) is measured.

## Accuracy Comparison:-

The data is split into training set and test set. The models are were built on training set and tested on test set.

The Root Mean Squared Error has been calculated for the predicted values for all three models as below:-

| Model | RMSE value |
|---|---|
| Time Series Linear Regression | 0.05666996 |
| HoltWinter's Model | 0.0360995 |
| ARIMA Model | 1.383397 |

Since the HoltWinter's model has the least value of RMSE, it implies that it the best suited forecasting model for the chosen dataset. With ARIMA having the highest RMSE, makes ARIMA the least suited model for forecasting for the given dataset.

## Code

```
# To import data
dataset<- read.csv('C:/Users/solan/Downloads/total-business-inventories-to-sales-ratio.csv')
# converting into time series
data <- ts(dataset[,2],start= c(1992,1),frequency=12)
#to decompose data
dec<- decompose(data)
plot(dec)
#split data into trainset and test set
trainset <- ts(dataset[,2],start= c(1992,1), end= c(2015,12),frequency=12)
test<- ts(dataset[,2],start= c(2016,1), end= c(2018,12),frequency=12)
# create a data frame to use tslm function
df_ts<- data.frame(value=trainset, as.numeric(time(trainset)))
# to build tlsm model
tslmmodel <- tslm(value~season+trend,df_ts)
# to see linear relationship
plot(dataset[,2], pch=16)
abline(mymodel)
#forecasting
tslm_fc <- forecast(tslmmodel,h=36)
autoplot(data) + autolayer(tslmmodel$fitted.values) + autolayer(tslm_fc)
par(mfrow=c(1,2))
#plotting ACF and PCF
acf(ts(tslmmodel$residuals),main='ACF Residual')
pacf(ts(tslmmodel$residuals),main='ACF Residual')
# Calculating RMSE
rmse(tslm_fc$fitted,data)

#to build HoltWinters Model
HW1<- HoltWinters(trainset, seasonal='multiplicative')
HW1
#forecasting
HW1.pred<- predict(HW1,36, prediction.interval = TRUE)
# plotting actual data, predicted data and trainset
plot(data,xlim=c(1992,2018),ylim=c(1,2),xlab = 'Year',ylab = 'Inventory to sales ratio')
lines(HW1$fitted[,1], col='red')
lines(HW1.pred[,1], col='blue')
lines(HW1.pred[,2], col='green')
lines(HW1.pred[,3], col='orange')
# Calculating RMSE
```

```
rmse(HW1.pred[,1],data)

#to build ARIMA model
# 1st order differencing
plot(diff(trainset),ylab='single Differenced ratio')
# 2nd order differencing
plot(diff(diff((trainset))),ylab='double Differenced ratio')
#applying log
plot((log10(trainset)),ylab='log')
#2nd order differencing on log series
plot(diff(diff((log10(trainset)))),ylab='double Differenced of log series')
require(forecast)
#ARIMA model
ARIMAfit = auto.arima(log10(trainset))
summary(ARIMAfit)
#forecasting
pred = predict(ARIMAfit, n.ahead = 24)
pred
par(mfrow = c(1,1))
# plotting actual data, predicted data and the error range
plot(data,type='l',xlim=c(1992,2018),ylim=c(1,2),xlab = 'Year',ylab = 'Inventory to sales ratio')
lines(10^(pred$pred),col='blue')
lines(10^(pred$pred+2*pred$se),col='orange')
lines(10^(pred$pred-2*pred$se),col='orange')
par(mfrow=c(1,2))
#plotting ACF and PACF
acf(ts(ARIMAfit$residuals),main='ACF Residual')
pacf(ts(ARIMAfit$residuals),main='PACF Residual')
#calculating RMSE
rmse(test,pred$pred)
```

## References:-


Lecture Ppts by Katrina Grolinger
Forecasting: Principles and Practice by Rob J Hyndman and George Athanasopoulos
https://rpubs.com/sediaz/tslm
http://www-ist.massey.ac.nz/dstirlin/CAST/CAST/Hmultiplicative/multiplicative1.html
https://www.vividcortex.com/blog/holt-winters-forecasting-simplified
https://www.youtube.com/watch?v=cZRMFNTreQI