Forecasting Using ARIMA model

ARIMA stands for: Auto Regressive (AR) + Integration(I) + MA (Moving Average)

Pre-requisites for using ARIMA Model

To use ARIMA, the time series should be stationary. A Series is said to be stationary when both the mean and variance do not depend on time at which the series is observed is a stationary time series. In other words, both mean and variance are constant.

The series is stationary or not, this is decided by looking at the graph or using a statistical function in R **kpss.test.**

In a time, series, the variance is made constant using **transformation** and mean is made constant using a process of **differencing** the time series.

Detailed Steps involved in predicting the sales of Souvenir data for the next five years

1. Converting the data into time series

   Below is the R code used to convert the data into time series:

   ```
   Shopsalets<-ts(Shopsale, start = c(1987,01),end = c(1993,12), frequency = 12)
   str(Shopsale)
   Shopsalets
   ```

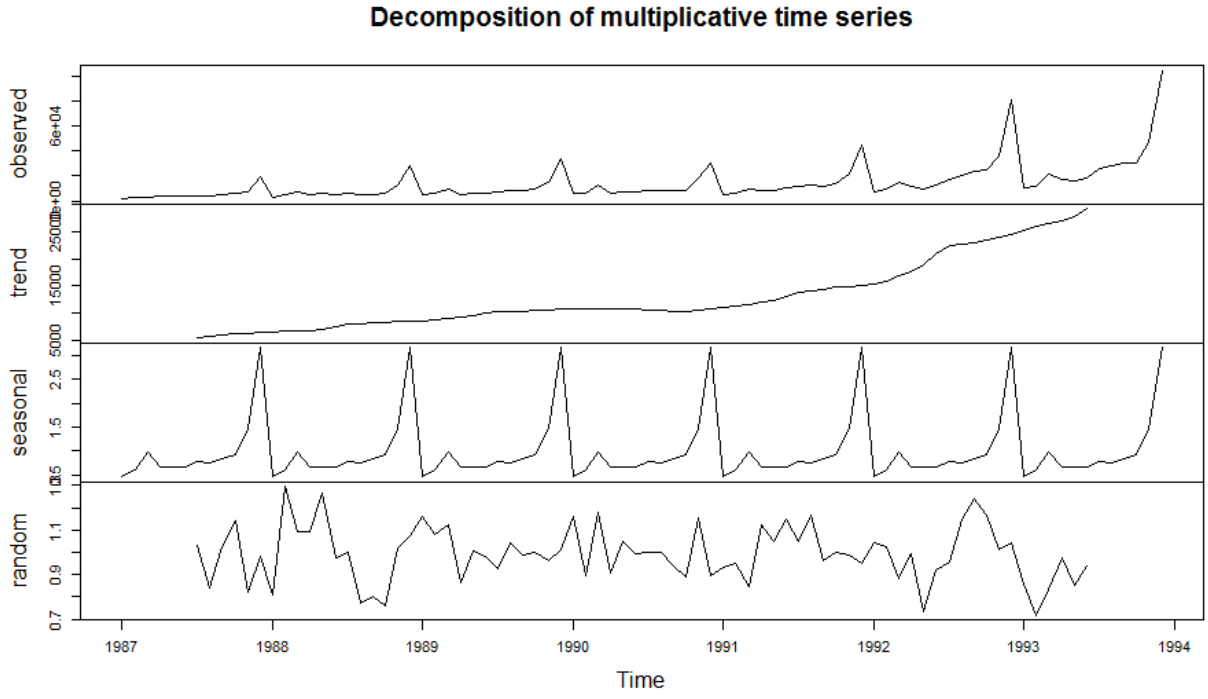
   Data set after conversion

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1987	1664.81	2397.53	2840.71	3547.29	3752.96	3714.74	4349.61	3566.34	5021.82	6423.48	7600.6	19756.21
1988	2499.81	5198.24	7225.14	4806.03	5900.88	4951.34	6179.12	4752.15	5496.43	5835.1	12600.08	28541.72
1989	4717.02	5702.63	9957.58	5304.78	6492.43	6630.8	7349.62	8176.62	8573.17	9690.5	15151.84	34061.01
1990	5921.1	5814.58	12421.25	6369.77	7609.12	7224.75	8121.22	7979.25	8093.06	8476.7	17914.66	30114.41
1991	4826.64	6470.23	9638.77	8821.17	8722.37	10209.48	11276.55	12552.22	11637.39	13606.89	21822.11	45060.69
1992	7615.03	9849.69	14558.4	11587.33	9332.56	13082.09	16732.78	19888.61	23933.38	25391.35	36024.8	80721.71
1993	10243.24	11266.88	21826.84	17357.33	15997.79	18601.53	26155.15	28586.52	30505.41	30821.33	46634.38	104660.7

2. Components of Time series

   Below is the R code used for breaking the time series in four components

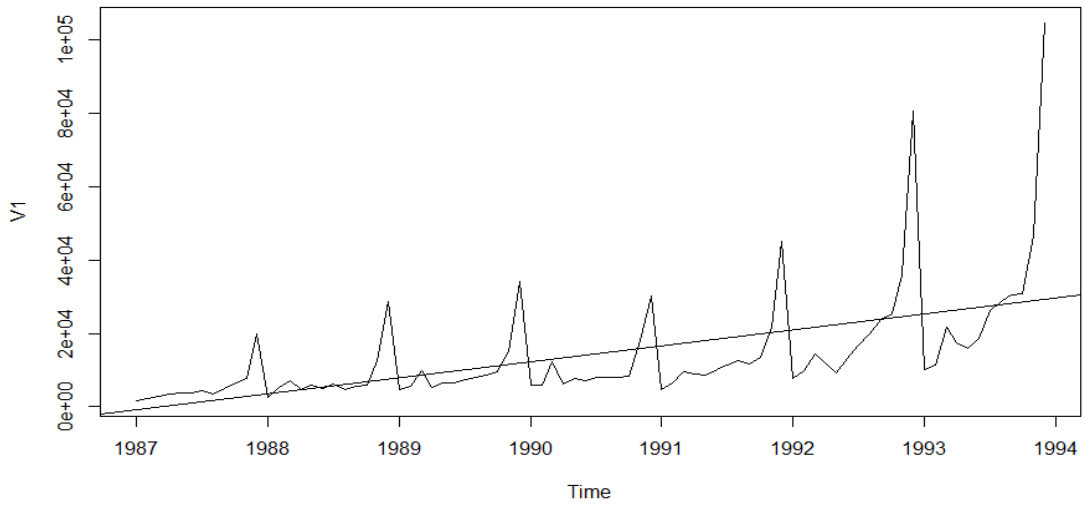
```
decom = decompose(Shopsalets, type = "multiplicative")
plot(decom)
```



3. Check for the stationarity of the time series

   Below is the R code used:

```
plot(Shopsalets)
abline(reg=lm(Shopsalets~time(salesdnew)))
```



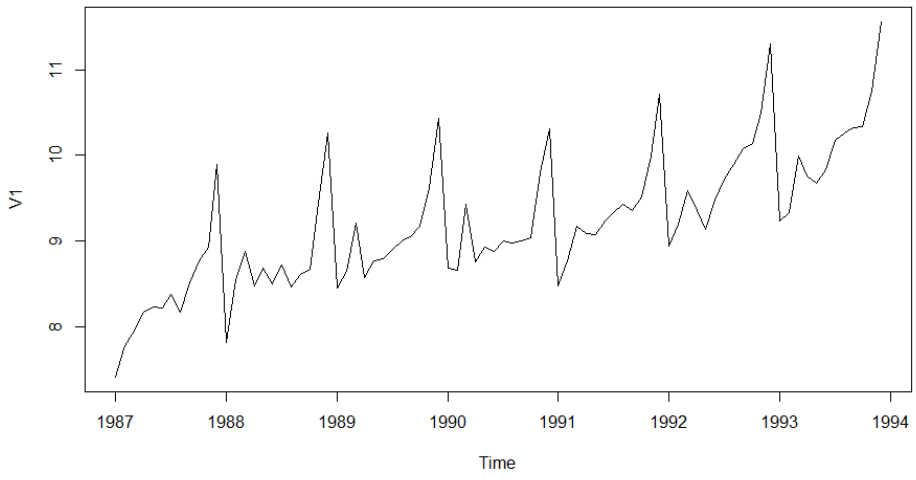
In the above graph, both the mean and variance are not constant. The mean is not parallel to the x-axis while the variance is increasing over the time.

4. Making variance constant using process of transformation

   To make the variance constant we can use transformation functions like log, sin, cos, tan etc. Here we have used the best suitable transformation function i.e. log. This will help in compressing the values of variance and thus making it constant.

   Below is the R code used for transformation:

   plot(log(Shopsalets))



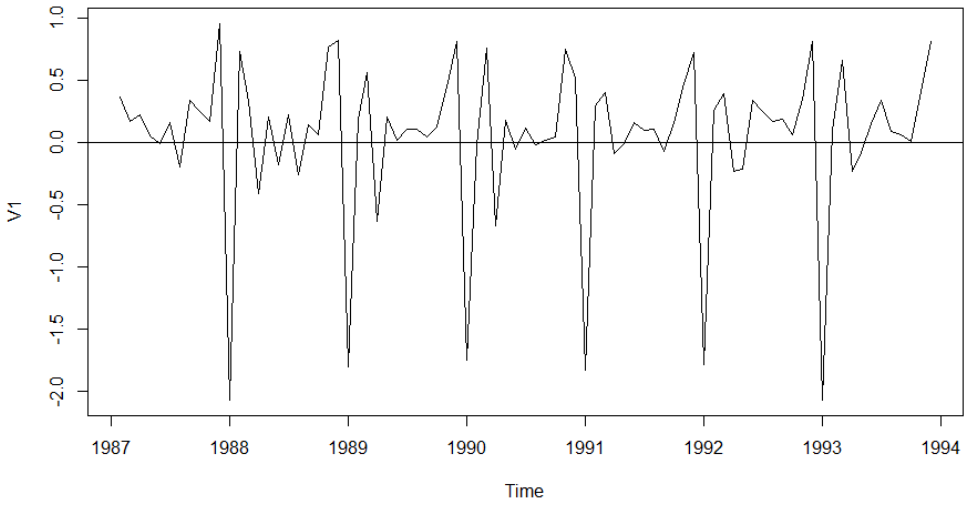
5. Making mean constant using process of Differencing

   Using differencing we get a new series called difference series through which mean becomes constant. If we do differencing once it is called first order differencing. In case the mean does not become constant then can go for second order differencing and so on. Also, with each differencing we lose one observation.

   Here, we have used first order of differencing to make the mean constant.

   Below is the R code used

```
plot(diff(log(Shopsalets)))
abline(h=0)
```

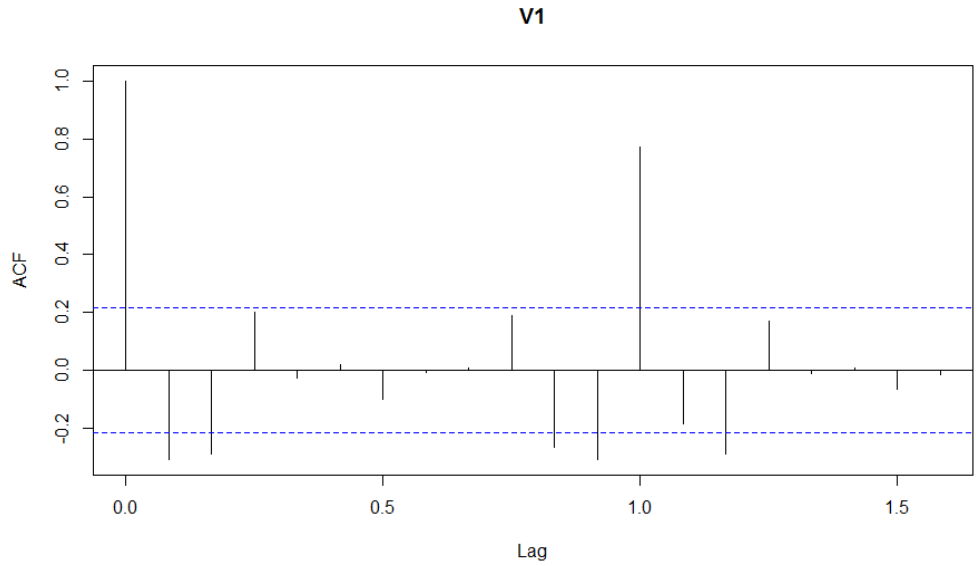


6. ACF (Auto-correlation function) Plot

   In ARIMA, the moving average takes care of the irregularity which is because of the correlation between the errors.
   This is denoted by **q** value which we get from the ACF plot. In an ACF plot correlation is plotted between the actual series and the lagged series.

   Below is the R code used:

   acf(diff(log(Shopsalets)))



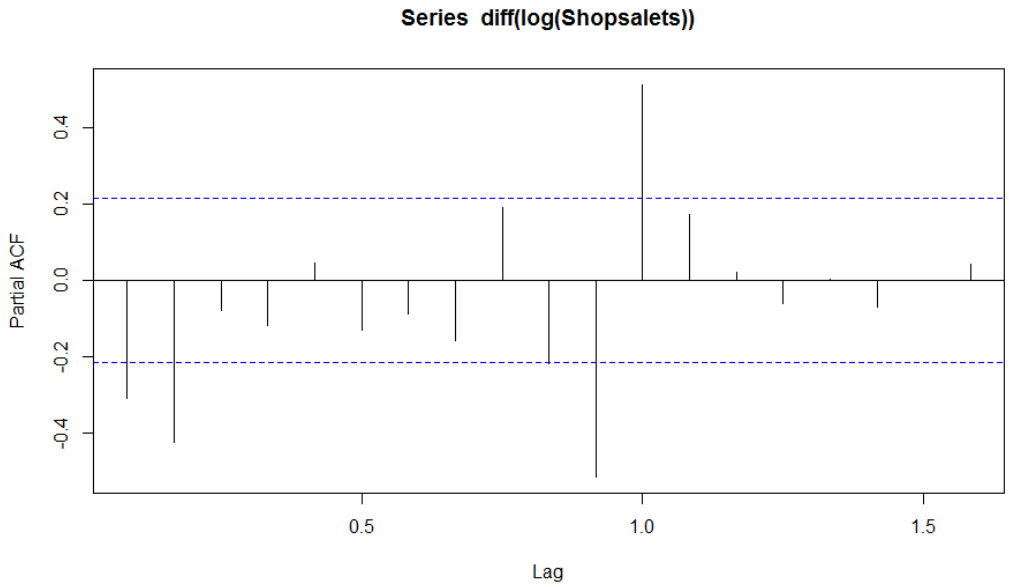
In the above graph the q value is 2 which will be put in the ARIMA model later. The area between the blue dotted line is called the significance zone where the correlation is acceptable.
Since the ACF starts from zero, thus the q value is (3-1) i.e. 2

7. PACF (Partial-correlation function) Plot

   In the ARIMA model, the auto regressive (AR) means that the forecasted value of dependent variable(y) is based on the its historical data. The number of periods which is to be taken in ARIMA model depends on the **p-value** which is obtained from PACF plot.

   Below is the R code used:

   pacf(diff(log(Shopsalets)))



In the above graph the P value is 2 which will be put in the ARIMA model later. PACF unlike ACF always start from one.

So, the final coordinates for ARIMA are [2(p), 1(d), 2(q)].

8. ARIMA Model

Below is the R code used for predicting the sale for next five years

```
model <- arima(log(Shopsalets), c(2,1,2), seasonal = list(order = c(2,1,2), period = 12))

pred <- predict(model, n.ahead = 60)

predf<-2.718^pred$pred

predf
```

	Predicted values											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1994	14819.18	18242.87	31479.38	23686.53	22892.89	25955.89	33544.61	35441.06	38740.07	40574.46	65323.84	140434.9
1995	20386.3	24963.15	42174.68	32775.17	31950.61	36352.16	45925.64	48929.21	53089.59	56928.14	88543.36	193215.1
1996	28395.82	35303.24	59663.54	44752.12	43640.79	49436.18	62422.54	65774.22	72060.38	76366	123379.2	263990.7
1997	38617.11	47415.52	79797.41	61847.9	60379.93	68673.85	86392.43	91928	99853.92	107214.2	167278.6	364248.8
1998	53588.97	66604.38	112452.5	84487.65	82417.26	93380.28	117792.2	124167.7	135985.9	144267.9	232686.5	498248.4

Check the Stationarity of the series through kpss.test

The null hypothesis for kpss.test is that the series is stationary. P-value should be greater than 0.05 to accept the null hypothesis.

As per the test the p-value is 0.1.

9. Check stability of the model

   Below is the R code used:

   shapiro.test(model$residuals)

   mean(model$residuals, na.rm = TRUE)

   Box.test(model$residuals,type = "Ljung-Box")

   - Result of Shapiro test

     p-value = 0.2637

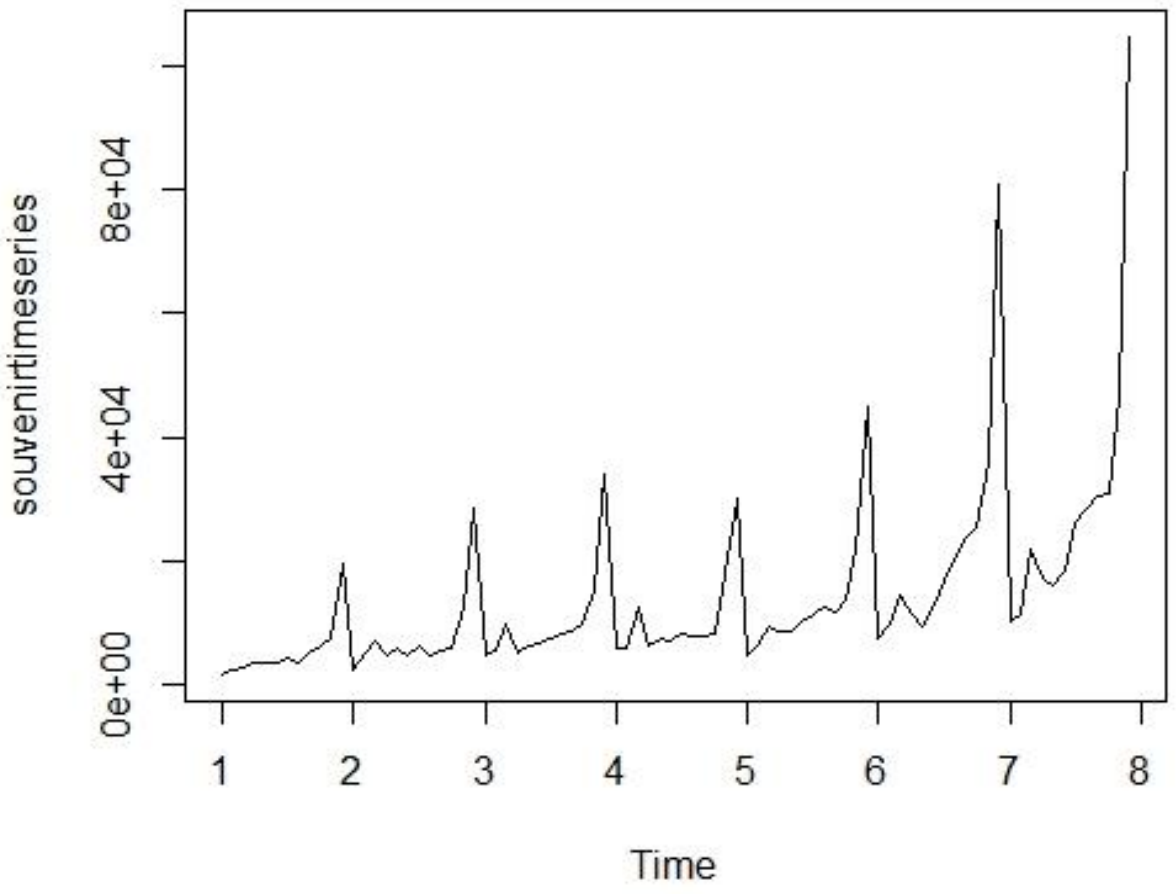
   - Result of Box Test

     X-squared = 0.030893, df = 1, p-value = 0.8605

Forecasting Using Winter-Holts model

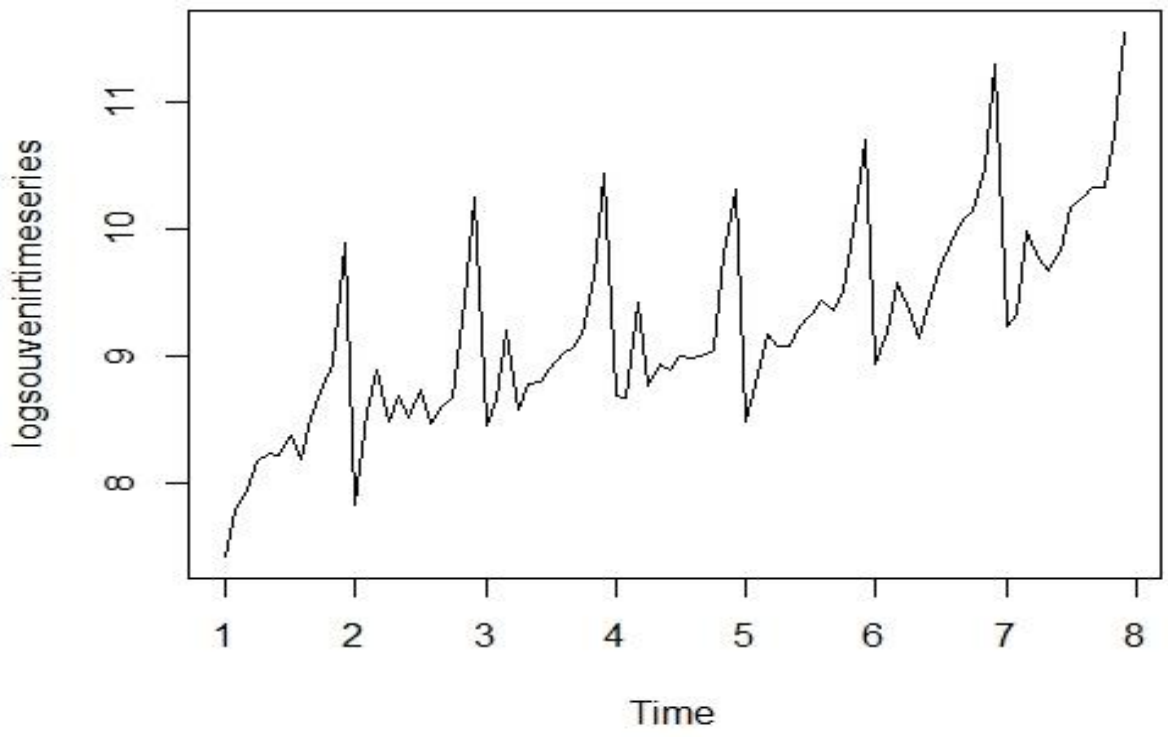
Steps involved in predicting the sales of Souvenir data for the next five years using Winter-Holts model

```
souvenir<-scan("C:/Users/Admin/Documents/R/fancy (1).txt")

library("forecast")

souvenirtimeseries<-ts(souvenir,frequency=12)

plot(souvenirtimeseries)
```



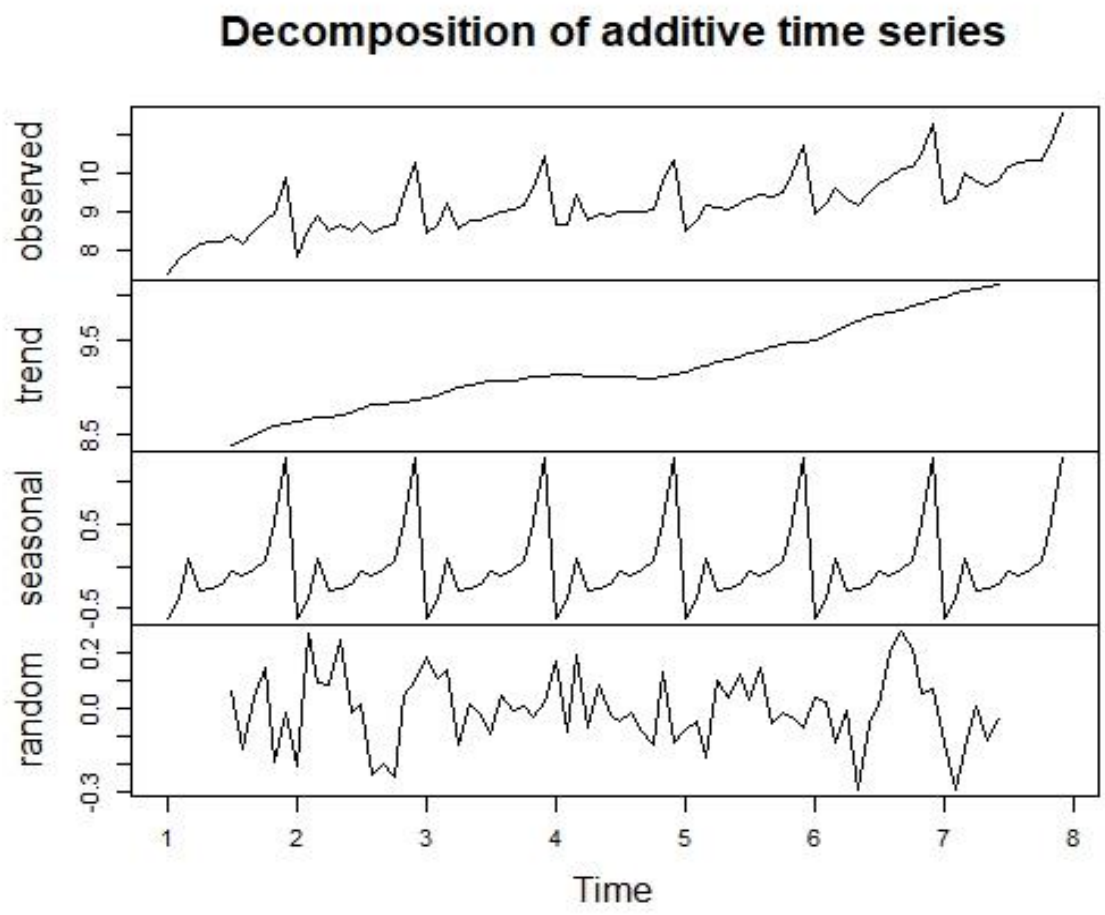
This can be noted from the above plot that there seems to be seasonal variation in the sales in one year. The sales are high in the end of the year and mostly steady for the first 10 months each year. As seen in the plot above, the seasonal sales are changing over time, and the random fluctuations also seem to be change in size over time. Additive model cannot be used to explain this model. Thus, we will transform the original data to its natural log.

```
logsouvenirtimeseries<-log(souvenirtimeseries)
plot.ts(logsouvenirtimeseries)
```



The plot above, it can be seen that the seasonal change seems constant over time and the random fluctuations also seem constant. I may say that the log model is an additive model.

```
> souvenir<-scan("C:/Users/Admin/Documents/R/fancy (1).txt"
)
Read 84 items
> library("forecast")
> souvenirtimeseries<-ts(souvenir,frequency=12)
> logsouvenirtimeseries<-log(souvenirtimeseries)
> plot.ts(logsouvenirtimeseries)
> logsouvenirtimeseries.component<-decompose(logsouvenirtim
eseries)
> names(logsouvenirtimeseries.component)
[1] "x"        "seasonal" "trend"    "random"   "figure"
"type"
> plot(logsouvenirtimeseries.component)
```



souvenirtimeseriesforecasts <- HoltWinters(logsouvenirtimeseries)

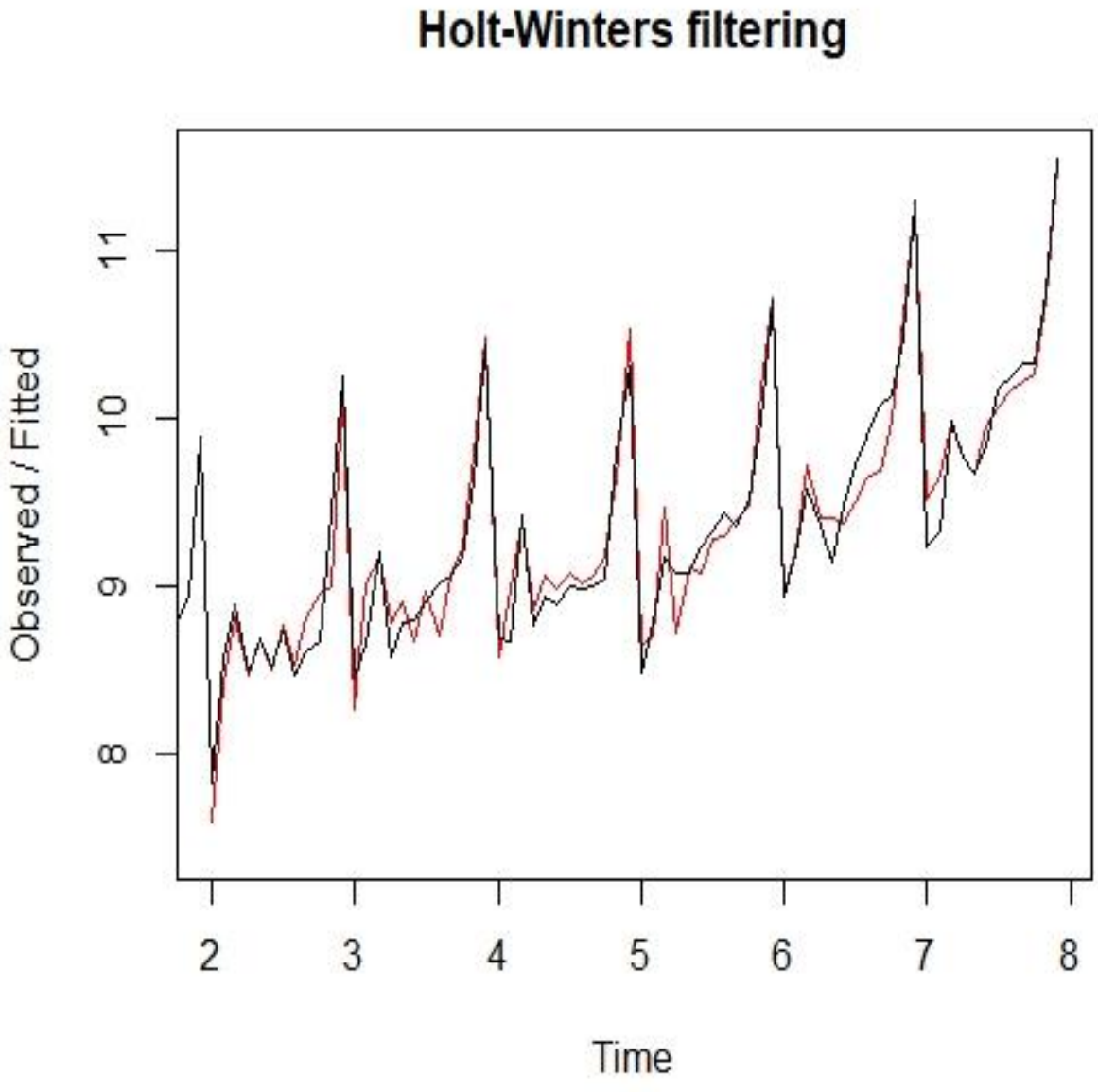
souvenirtimeseriesforecasts

```
Holt-Winters exponential smoothing with trend and additive seasonal
component

Call:
HoltWinters(x = logsouvenirtimeseries)

Smoothing parameters:
 alpha: 0.413418
 beta : 0
 gamma: 0.9561275


Coefficients:
           [,1]
a   10.37661961
b    0.02996319
s1  -0.80952063
s2  -0.60576477
s3   0.01103238
s4  -0.24160551
s5  -0.35933517
s6  -0.18076683
s7   0.07788605
s8   0.10147055
s9   0.09649353
s10  0.05197826
s11  0.41793637
s12  1.18088423
```

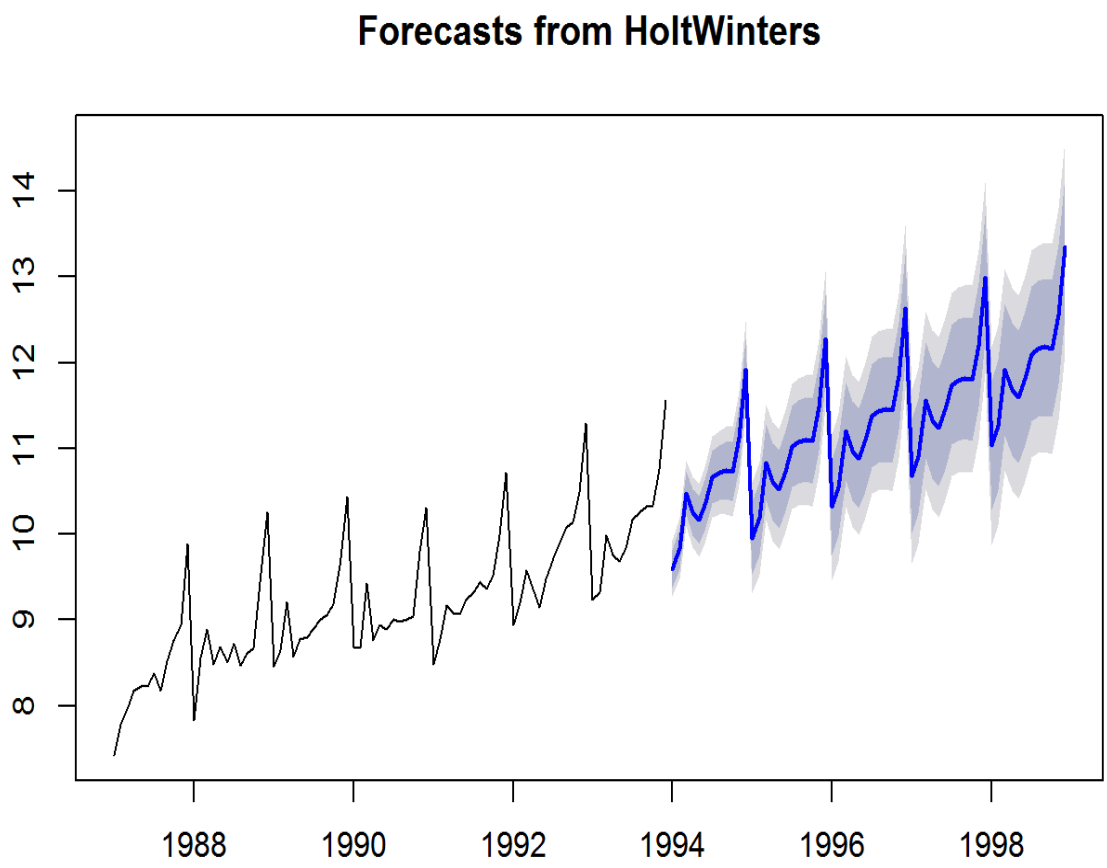
The estimated values of alpha, beta and gamma are 0.41, 0.00, and 0.96, respectively. The value of alpha (0.41) is relatively low, indicating that the estimate of the level at the current time point is based upon both recent observations and some observations in the more distant past. The value of beta is 0.00, indicating that the estimate of the slope b of the trend component is not updated over the time series, and instead is set equal to its initial value. In contrast, the value of gamma (0.96) is high, indicating that the estimate of the seasonal component at the current time point is just based upon very recent observations.

plot(souvenirtimeseriesforecasts)

To see the result compared to the original data, we will plot both data in one plane, red as the forecast data and black is the original data.



Plot of Holt Winters for next 5 years



The blue line is the forecast value while the dark grey and grey shows the 80% and 95% prediction intervals.

```
> souvenirtimeseriesforecasts
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Jan 1994       9.597062  9.381514  9.812611  9.267409  9.926715
Feb 1994       9.830781  9.597539 10.064024  9.474068 10.187495
Mar 1994      10.477542 10.227856 10.727227 10.095680 10.859403
……(cutted for simplicity)
Oct 1997      11.806904 11.091167 12.522642 10.712278 12.901531
Nov 1997      12.202826 11.481562 12.924089 11.099748 13.305903
Dec 1997      12.995737 12.268989 13.722485 11.884272 14.107202
```