Business Forecasting

Final Exam: Forecasting US Tourism

Ronak Parikh

Business Forecasting Final Exam 3

Introduction 3

Import Data 3

Plot and Inference 3

Central Tendency 3

Decomposition 4

Naïve Method 4

Simple Moving Averages 4

Simple Smoothing 5

Holt-Winters 5

ARIMA or Box-Jenkins 5

Accuracy Summary 6

Conclusion 6

Final Question 6

# Business Forecasting Final Exam

## Introduction

US Department of Tourism maintains statistics on visitors to America. They have very details datasets. We are going to focus on just a general statistic that keeps track of monthly visitors. See details at http://travel.trade.gov/research/monthly/arrivals/index.asp

## Import Data

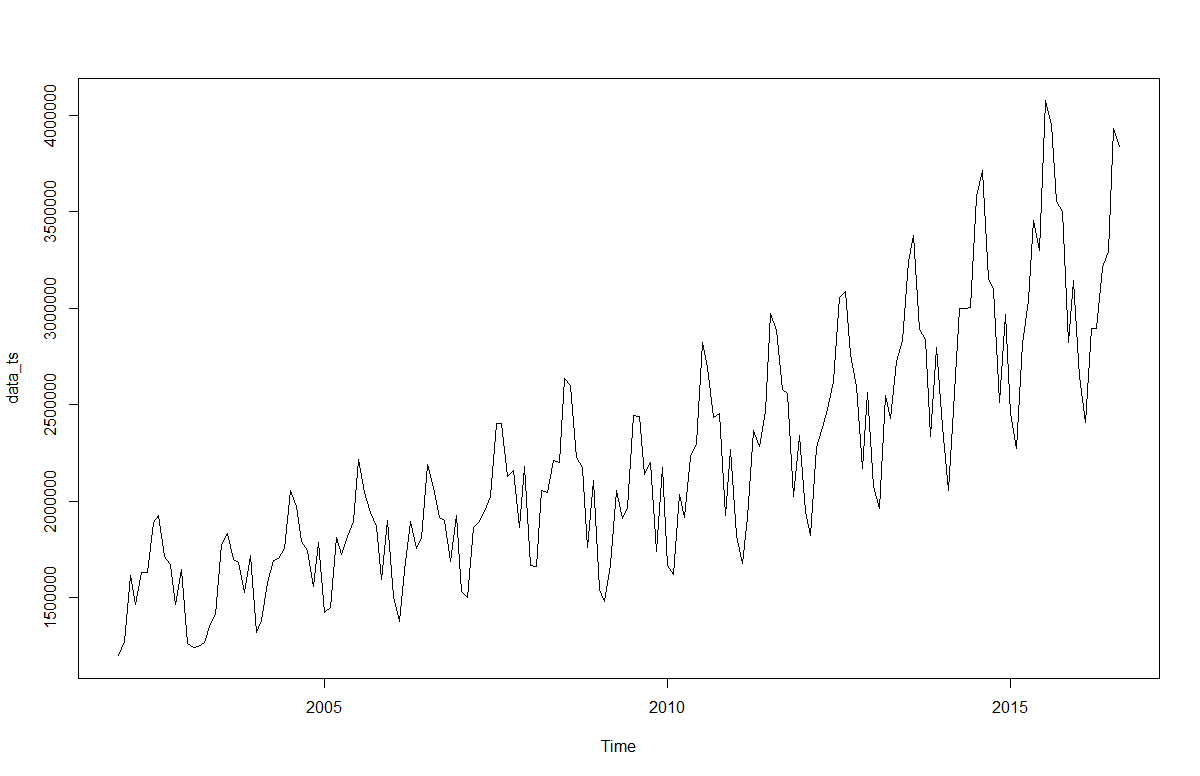
Please do the following steps:

* Final\_Travel <- read\_csv("~/Desktop/Final\_Travel.csv")
* travel <- Final\_Travel$Value
* plot(travel)
* travel\_ts <- ts(travel,start=c(1999,1),frequency = 12)
* plot(travel\_ts)

**Note**   
Use Window function to take subset of the data to remove values you do not want in your analysis. See ?window for help on the function. If you feel comfortable to take values out in csv and then import in R, please do so.

## Plot and Inference

* Show a time series plot.



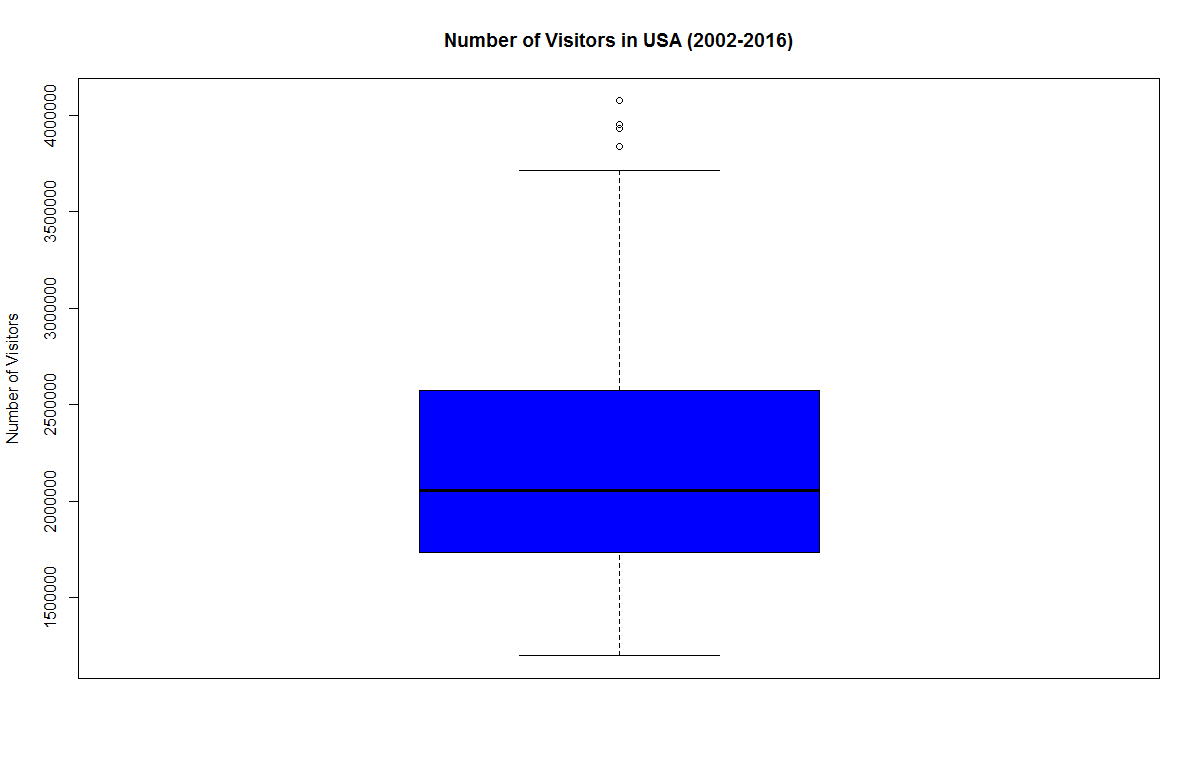
* Please summaries your observations of the times series plot
* The plot shows visitors to USA from the January 2002 to December 2016 with a period of one month. *The original dataset has modified to only include the data from the year 2002, as there was an erratic dip in the number of visitors to USA considering the 9/11 terror attacks on USA in the year 2001*
* The data shows an upward trend from 2002 to 2017
* There is also a seasonality factor where the number of visitors are minimum at the start of the year(winters), increases over the spring months to reach the maximum in the month of July(summers), reduces to month of November (fall) and then again increase in the month of December(winters) probably due to holidays & festive season
* It seems that this time series could probably be described using an additive model, as the seasonal fluctuations are roughly constant in size over time and do not seem to depend on the level of the time series, and the random fluctuations also seem to be roughly constant in size over time.

## Central Tendency

* What are the min, max, mean, median, 1st and 3rd Quartile values of the times series?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 1197000 | 1735000 | 2056000 | 2206000 | 2569000 | 4075000 |

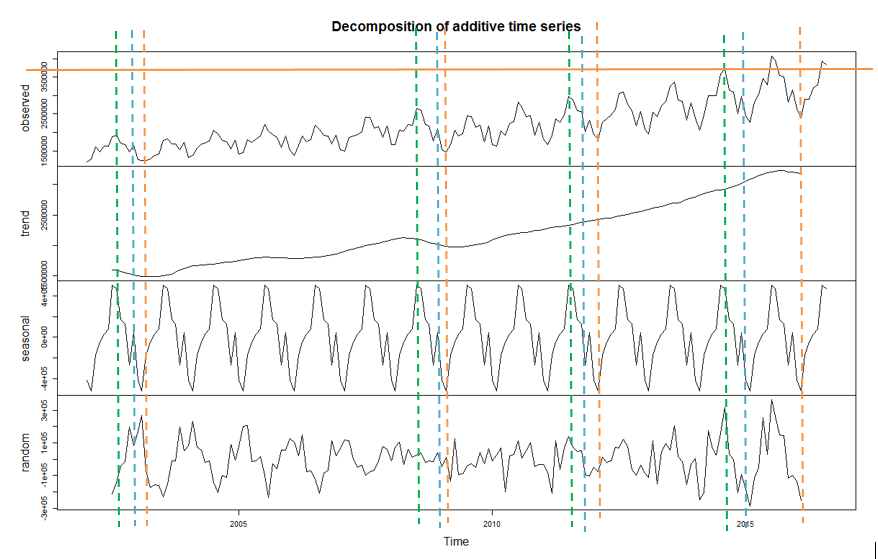
* Show the box plot.



* Can you summaries your observation about the time series from the summary stats and box plot?
* The 1st quartile value 1735000 is the value that cuts off the first 25% of number of visitors in USA when sorted in ascending order. Whereas the 3rd quartile value 2569000 is the value that cuts off the first 75% of number of visitors in USA when sorted in ascending order. The central rectangle spans the 1st quartile to the third quartile (the interquartile range or IQR).
* The segment inside the rectangle shows the median value of 2056000 represents the midpoint of the data, "whiskers" above and below the box show the locations of the minimum value of 1197000 and maximum value of 4075000 for the number of claims respectively.
* The mean value for the time series data is 2206000, that lies above the median line in the box plot. The data is not symmetrical, as most of the observations are on the lower end of the scale and also the mean is above the median, the data is skewed towards the right.

## Decomposition

* Plot the decomposition of the time series.



* Is the times series seasonal?
* The table provides:

1. % contribution by seasonality in the data at the maximum number of visitors (highlighted by green)
2. % contribution by seasonality in the data at the moderate number of the visitors (highlighted by blue)
3. % reduction by seasonality in the data at the minimum numbers of visitor (highlighted by orange)

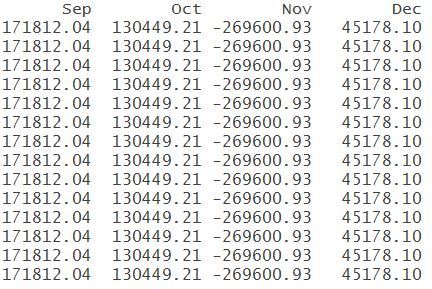
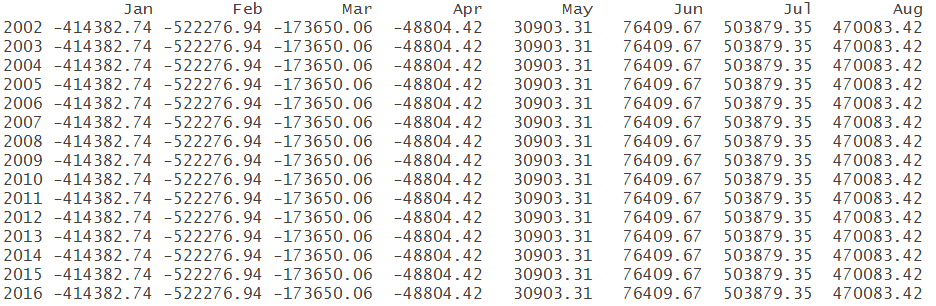
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data | Trend | Seasonality | Random Noise | % Contribution by Seasonality (ABS) | % Contribution by Random (ABS) |
| 1924890 | 1596882 | 470083.42 | -142075.088 | 24.42131343 | 7.380945831 |
| 2601265 | 2093439 | 470083.42 | 26340.4445 | 18.07133914 | 1.01260135 |
| 2882766 | 2340004 | 470083.42 | 72678.3701 | 16.30667977 | 2.521133179 |
| 3712864 | 2926500 | 470083.42 | 316281.0368 | 12.66093829 | 8.51851931 |
| Average |  |  |  | 17.86506766 | 4.858299917 |
| 1647288 | 1517553 | 45178.1 | 84557.0665 | 2.742574462 | 5.133107659 |
| 2104777 | 2019541 | 45178.1 | 40058.1915 | 2.14645542 | 1.903203594 |
| 2340148 | 2399903 | 45178.1 | -104932.975 | 1.93056593 | 4.48403157 |
| 2968315 | 3018154 | 45178.1 | -95017.5585 | 1.522011646 | 3.201060484 |
| Average |  |  |  | 2.085401865 | 3.680350827 |
| 1239906 | 1495468 | -522276.94 | 266714.7749 | 42.12230121 | 21.5108867 |
| 1477601 | 1987036 | -522276.94 | 12841.6082 | 35.34627684 | 0.869084969 |
| 1821812 | 2422088 | -522276.94 | -77999.2251 | 28.66799318 | 4.281409119 |
| 2272927 | 3081880 | -522276.94 | -286675.6 | 22.97816604 | 12.61261801 |
| Average |  |  |  | 35.37885708 | 9.8184997 |
| Total Average |  |  |  | 18.44310887 | 6.119050148 |

1. The total average indicates that the seasonality affects the number of visitors by 18.4431%. Therefore it is evident that the time series is seasonal with trend as dominant component
2. The total average indicates that the randomness/noise affects the number of visitors by 6.12 %. Therefore it is evident that the time series is somewhat affected by randomness/noise

* Is the decomposition additive or multiplicative?

Additive, as the seasonal variation looks constant. It doesn’t change with the time

* If seasonal, what are the values of the seasonal monthly indices?



* For which month is the value of time series high and for which month is it low?

1. The estimated seasonal factors are given for the months January-December and are the same for each year.
2. The largest seasonal factor is for July (503879.35), and the lowest is for February (-414382.74), indicating that there seems to be a peak in the number of visitors in July and a trough in number of visitors in February each year.
3. The time series data value is highest for July, and the lowest is for February.

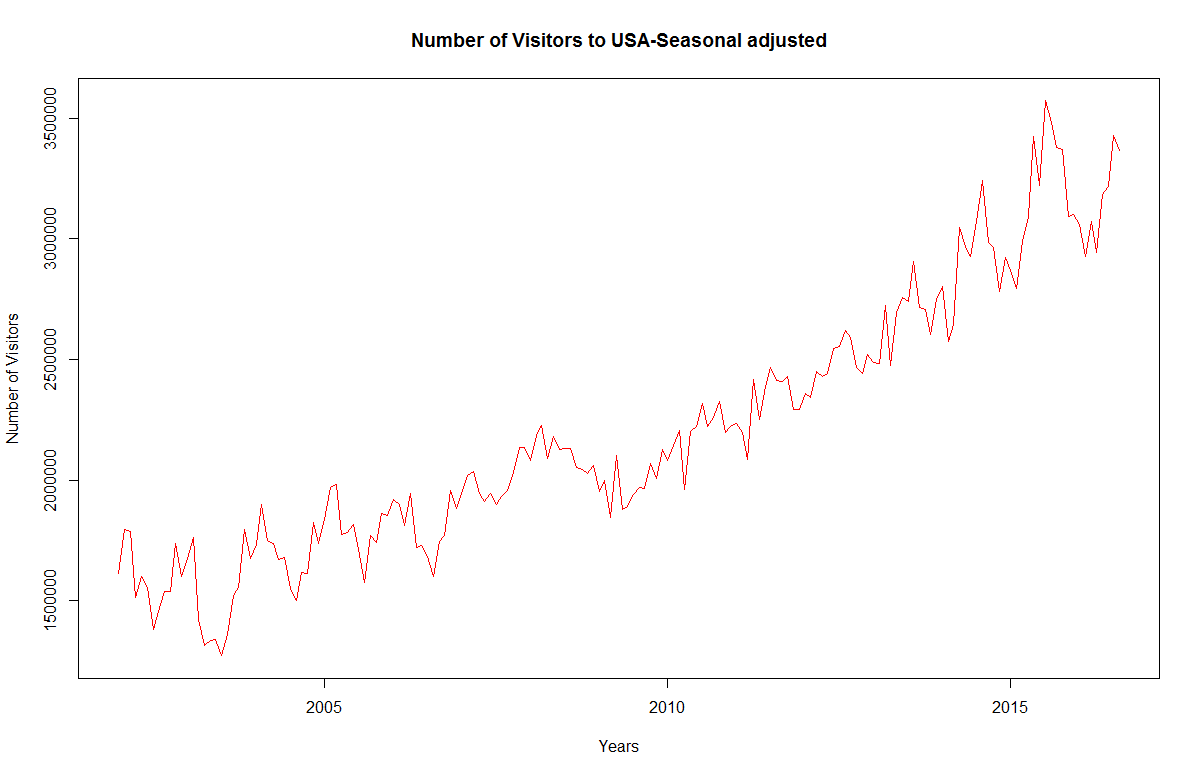
* Can you think of the reason behind the value being high in those months and low in those months?
* Number of visitors are high in July (Summer) because of following reasons:

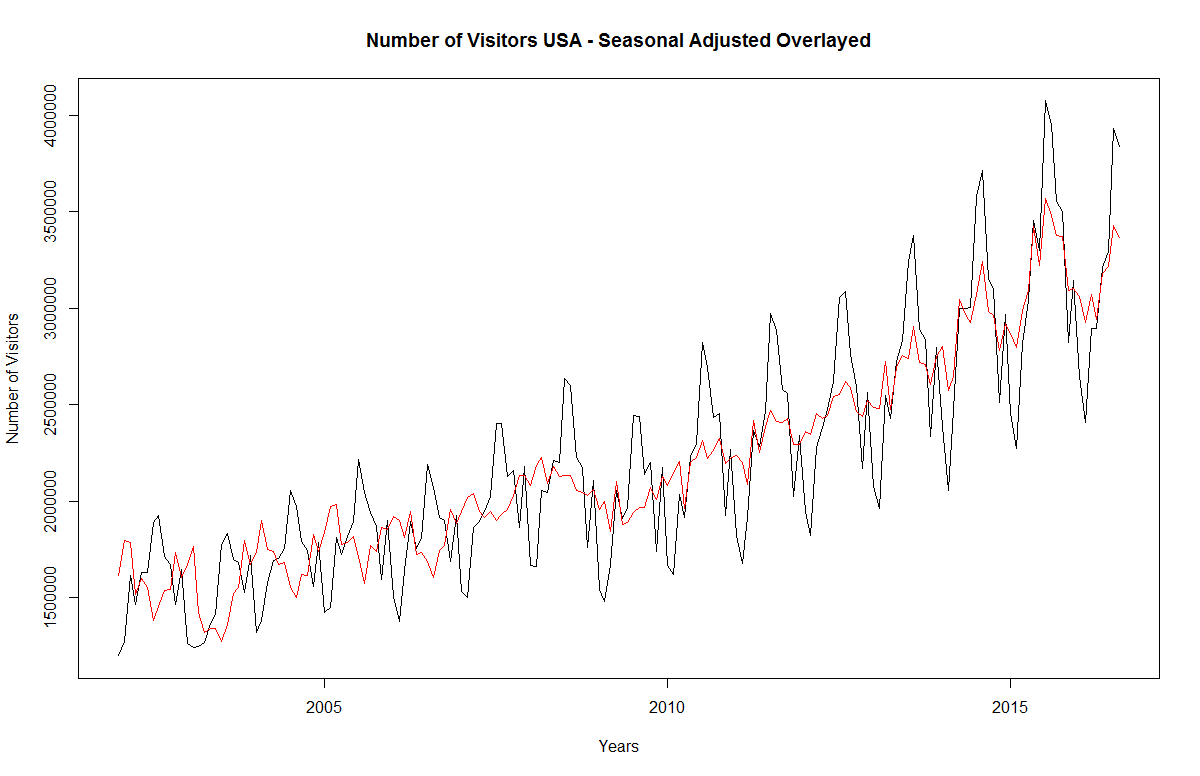
1. Tourists visit USA to spend the summer vacations
2. International students travel to USA in July and August (summer months) just before the start of Fall semester in Universities

* Number of visitors are low in February (Winter) because of following reasons:

1. A large part of USA gets harsh winter, therefore less number of tourists visit the country in winter season, they tend to travel to other warmer countries of the world

* There is also a rise in number of visitors in month of December due to:
  1. Festive & holidays season, people visit their families during Christmas & New Year
  2. International students travel to USA in December just before the start of spring semester in Universities
* Show the plot for time series adjusted for seasonality. Overlay this with the line for actual time series? Does seasonality have big fluctuations to the value of time series?





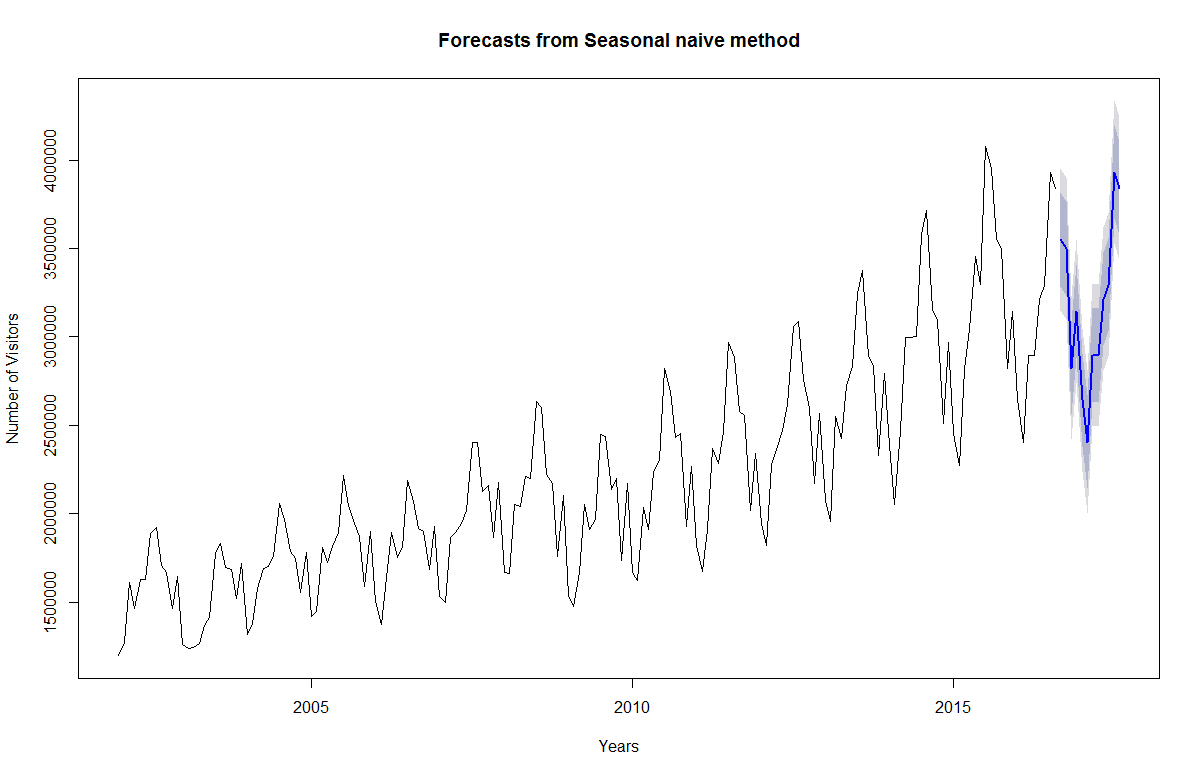
Yes, seasonality does have big fluctuations to the value of time series, the number of visitors follows the peak and depression of the seasonal data to hit the highest and lowest values.

The data is following the seasonality fluctuations in the starting two years i.e. 2002 & 2003 and from 2009 to 2016, for the years 2004 to 2008 seasonality shows an inverse effect on the data.

## Naïve Method

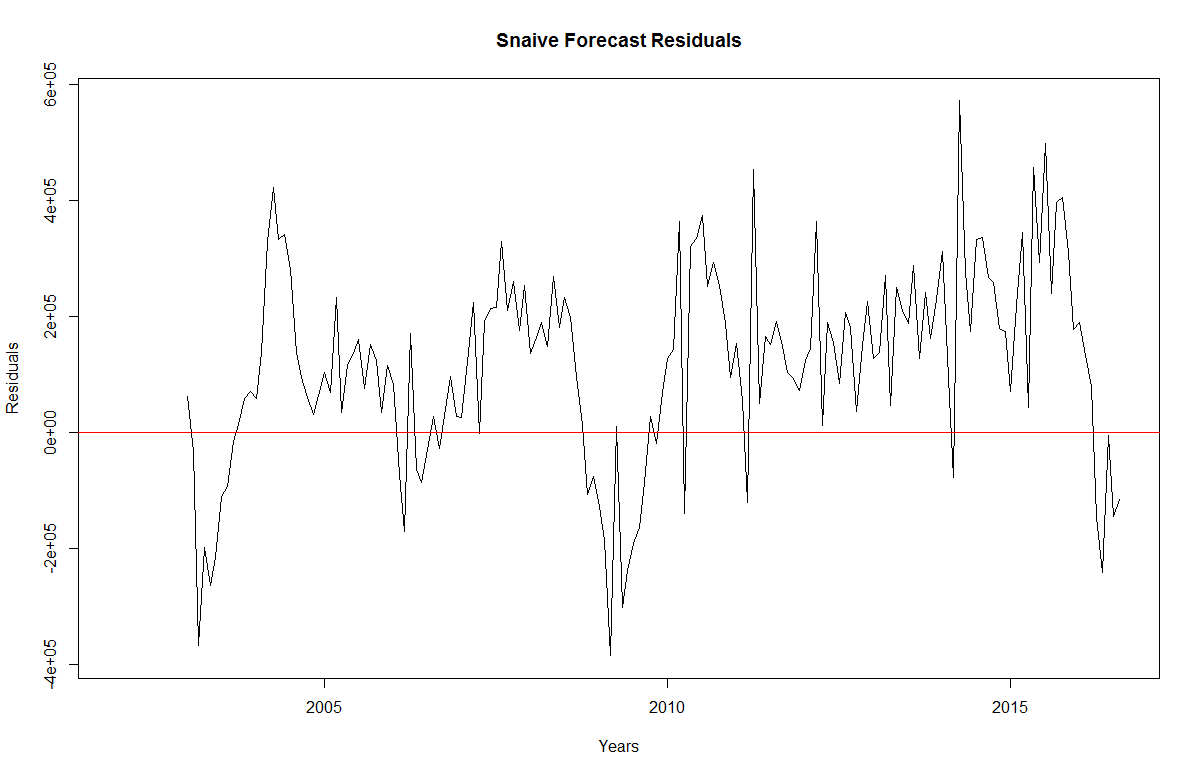
* Output

As there is seasonal component in the data, we will use snaive forecasting method, with forecasting period = 12



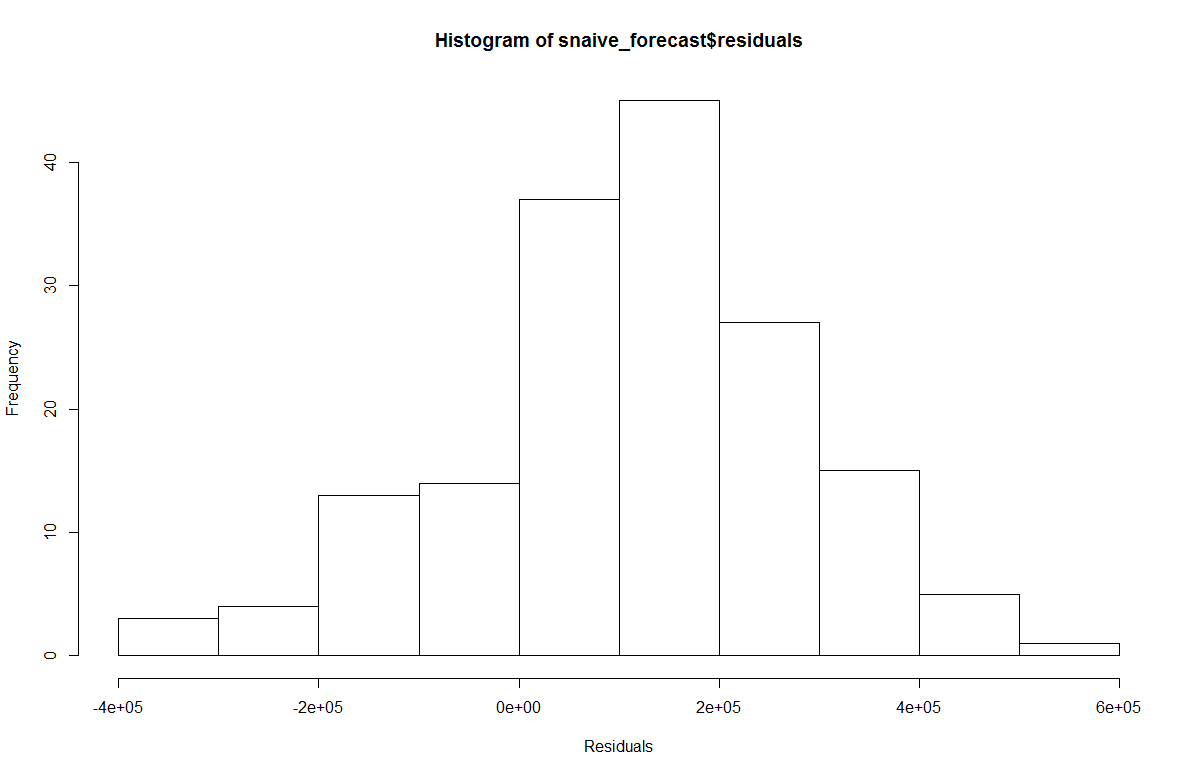
* Perform Residual Analysis for this technique.
  + Do a plot of residuals. What does the plot indicate?

The plot indicates forecast errors or residuals which are calculated as the observed values minus predicted values, for each time point. We can only calculate the forecast errors for the time period covered by our original time series, which is 2002-2016 for the data.



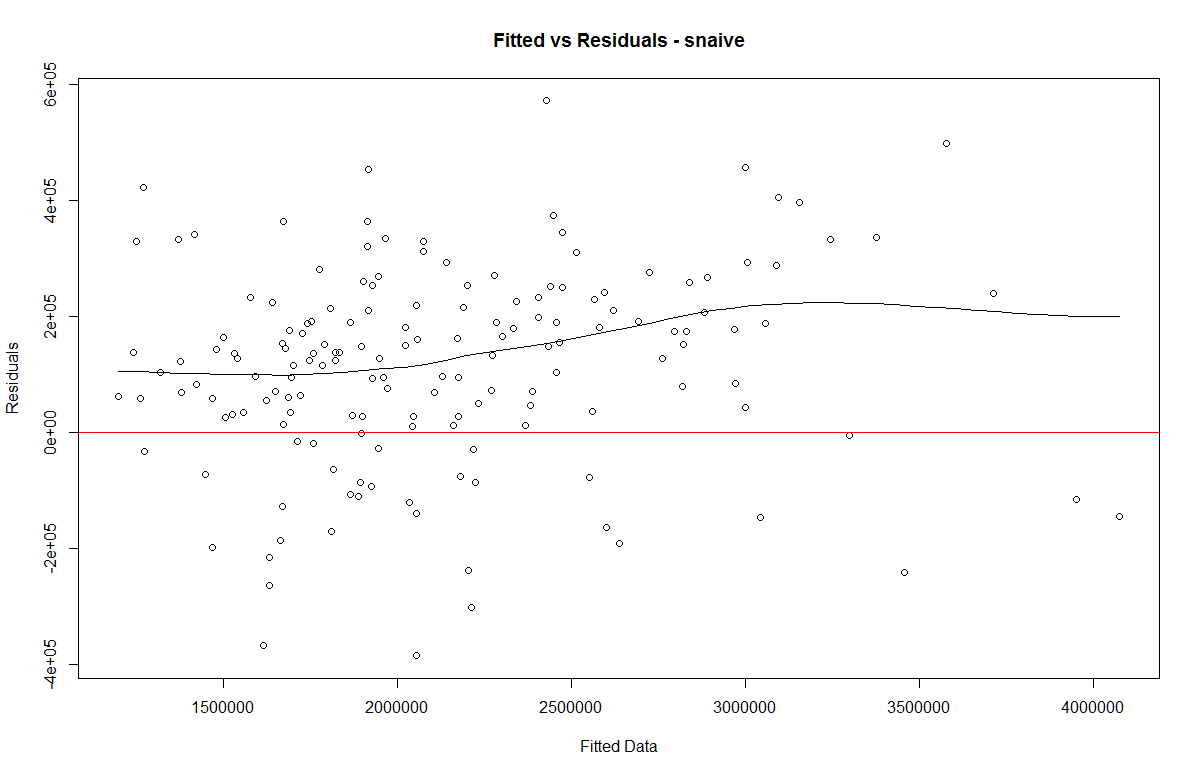
The plot shows that the forecast errors or residuals have roughly constant variance over the time.

* + Do a Histogram plot of residuals. What does the plot indicate?



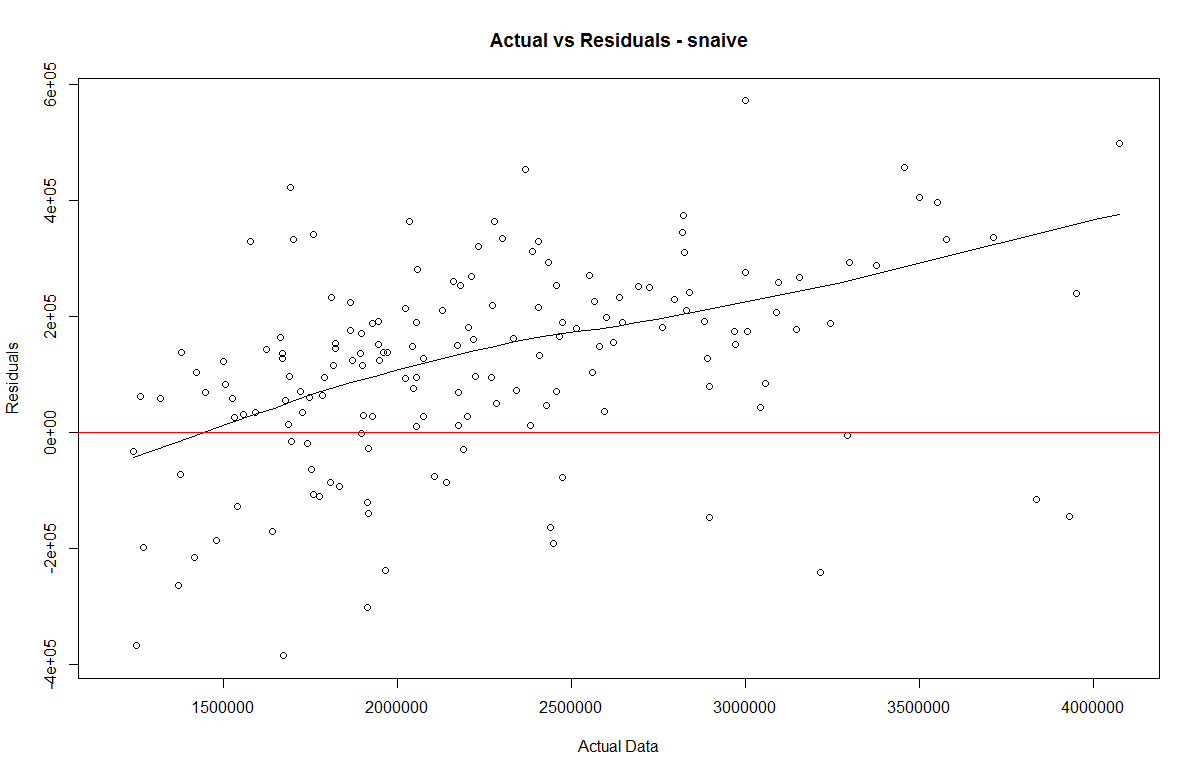
The histogram of forecast errors show that it is *likely* that the forecast errors are normally distributed with mean zero and constant variance. The data is skewed towards the left.

* + Do a plot of fitted values vs. residuals. What does the plot indicate?



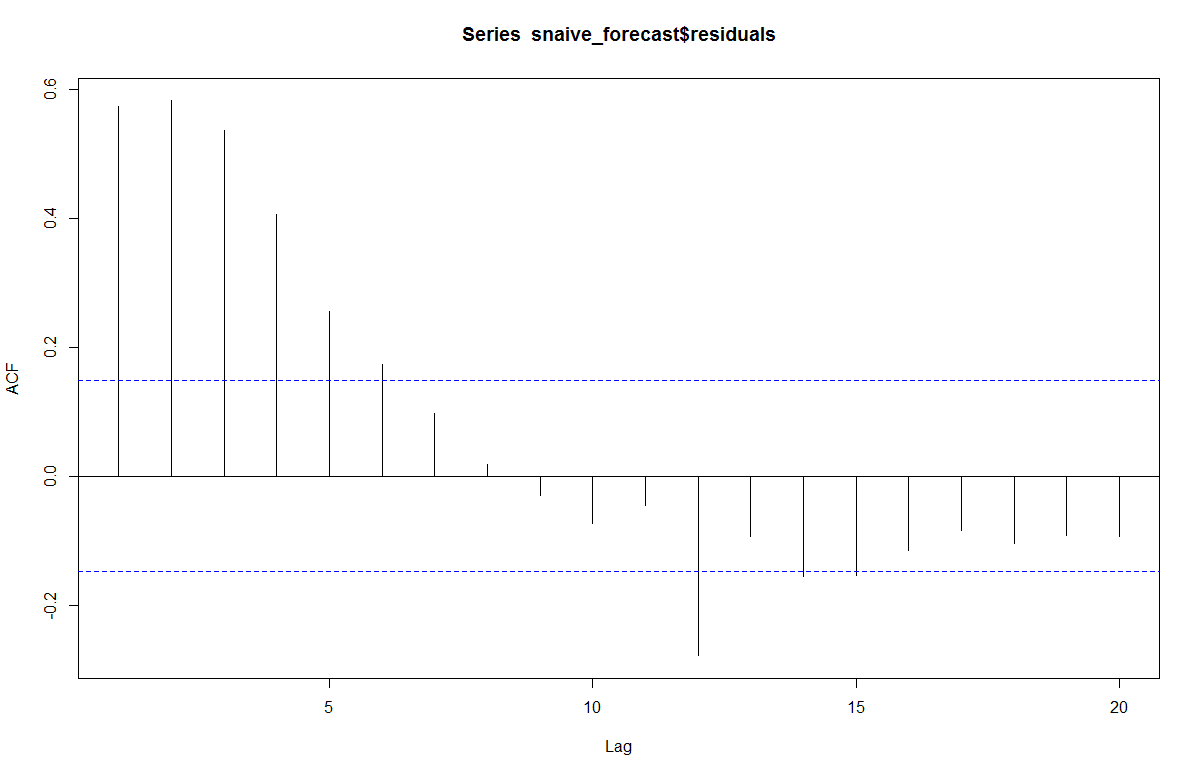
The errors have constant variance, with the residuals scattered randomly around zero. The variation in the residuals does not seem to change with the size of the fitted value.

* + Do a plot of actual values vs. residuals. What does the plot indicate?



The errors have constant variance, with the residuals scattered randomly around zero. There is a slight linearity, as the value of actual data is increasing the residuals appears to be more on the +ive side.

* + Do an ACF plot of the residuals? What does this plot indicate?

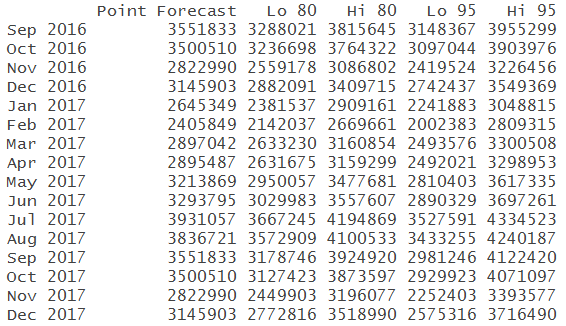


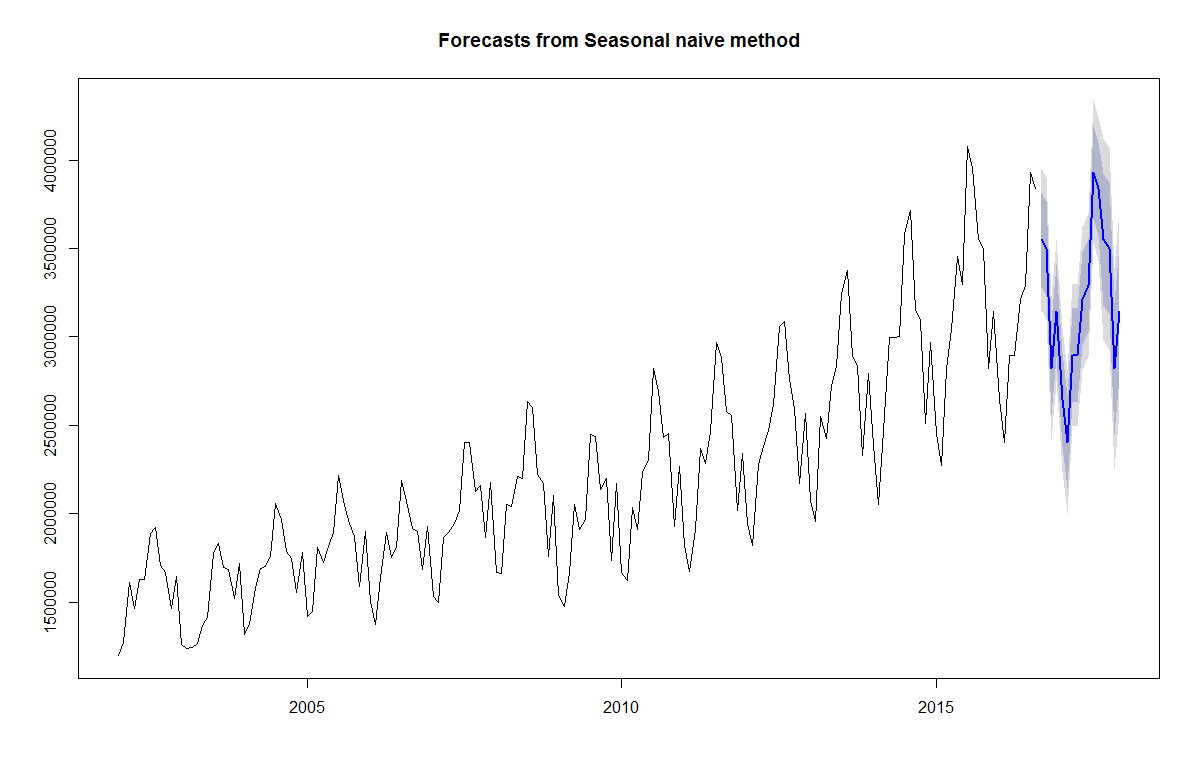
The correlogram provides the correlations between forecast errors for successive predictions. As autocorrelations at lag 1 – 6 crosses the significance bounds, this shows that there is correlation between successive forecast errors for successive predictions. Snaive forecast can by another forecasting technique.

* Print the 5 measures of accuracy for this forecasting technique

|  |  |
| --- | --- |
| ME | Mean Error |
| RMSE | Root Mean Square Error |
| MAE | Mean Absolute Error |
| MPE | Mean Percentage Error |
| MAPE | Mean Absolute Percentage Error |
| MASE | Mean Absolute Scaled Error |

* Forecast
  + Time series value for next year. Show table and plot

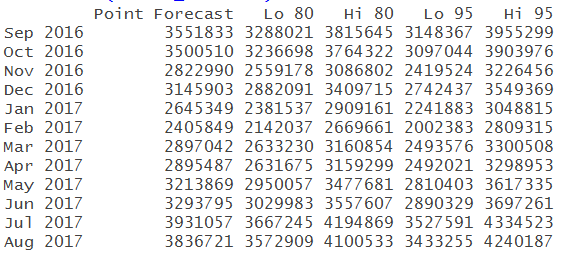


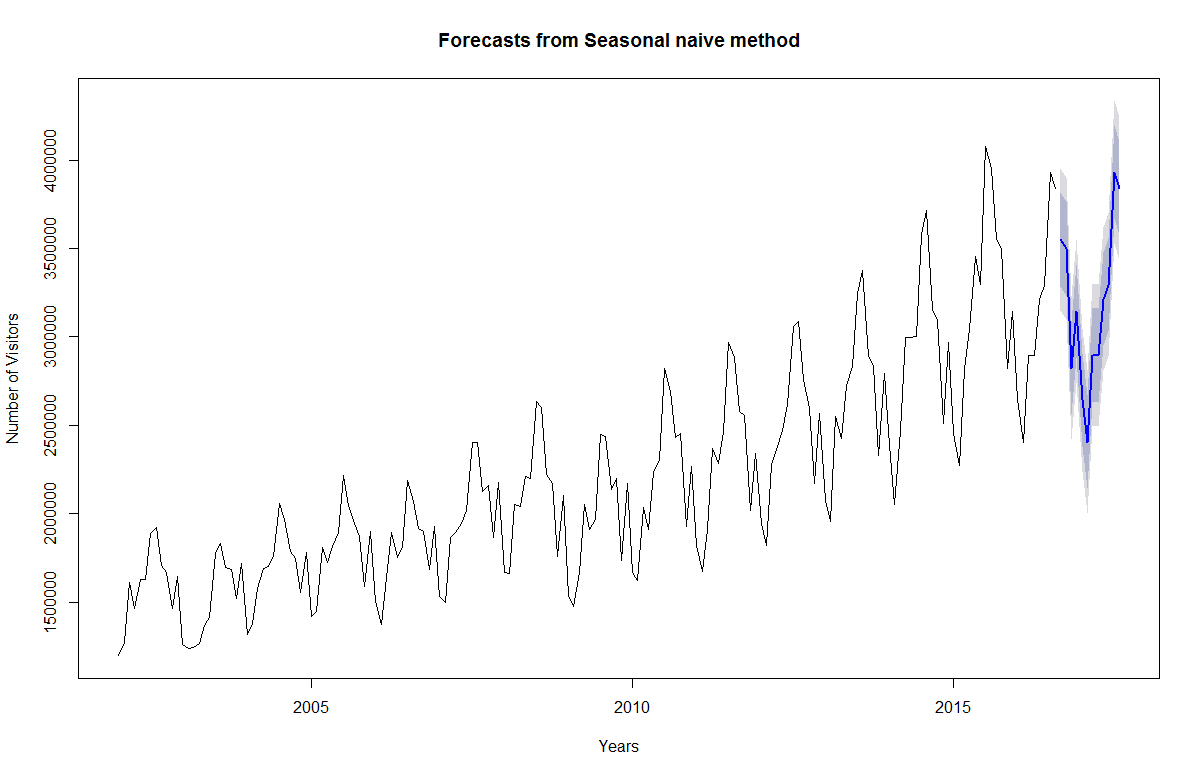


* Summarize this forecasting technique
  + How good is the accuracy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
| 115998.2 | 205853.8 | 172182.5 | 4.595344 | 7.741379 | 1 | 0.57322 |

* As RMSE is very high we see that explanatory/predictive information is in the error.
  + What does it predict the value of time series will be in one year



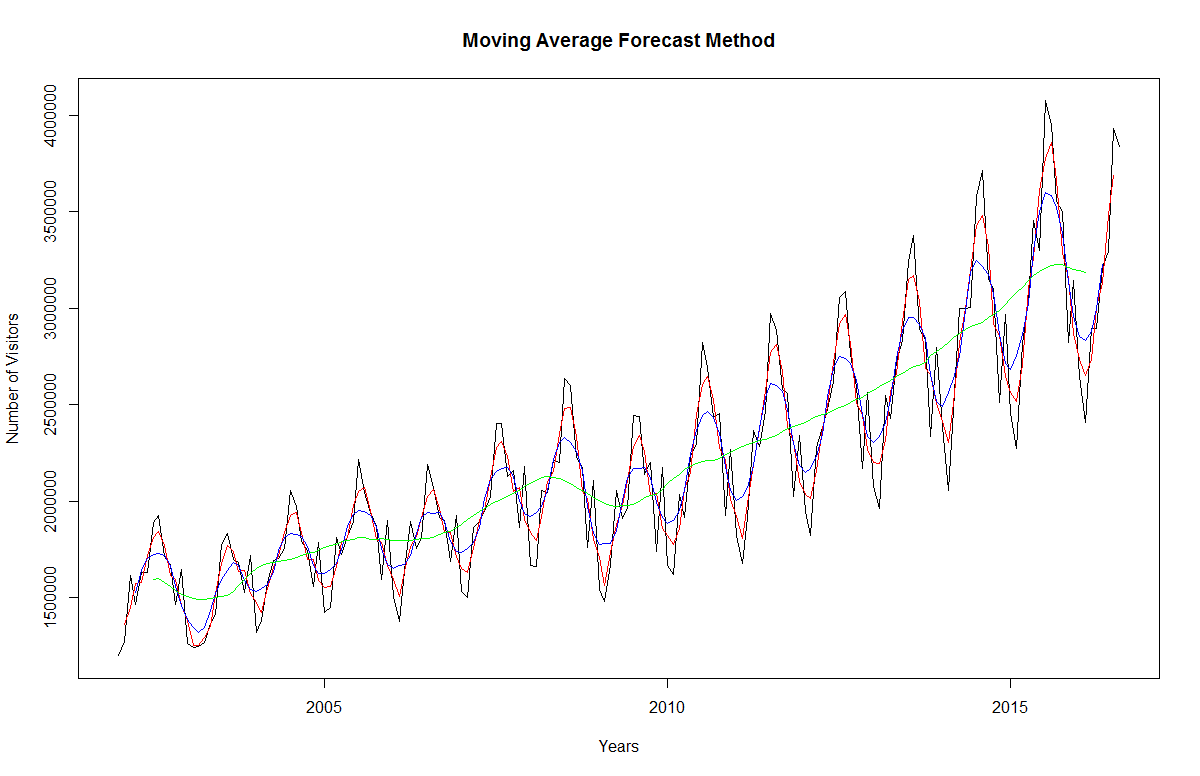


* + Other observation

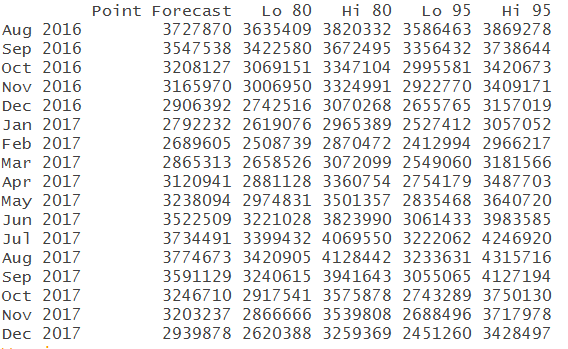
The forecast can be improved using other models as RMSE is quiet high

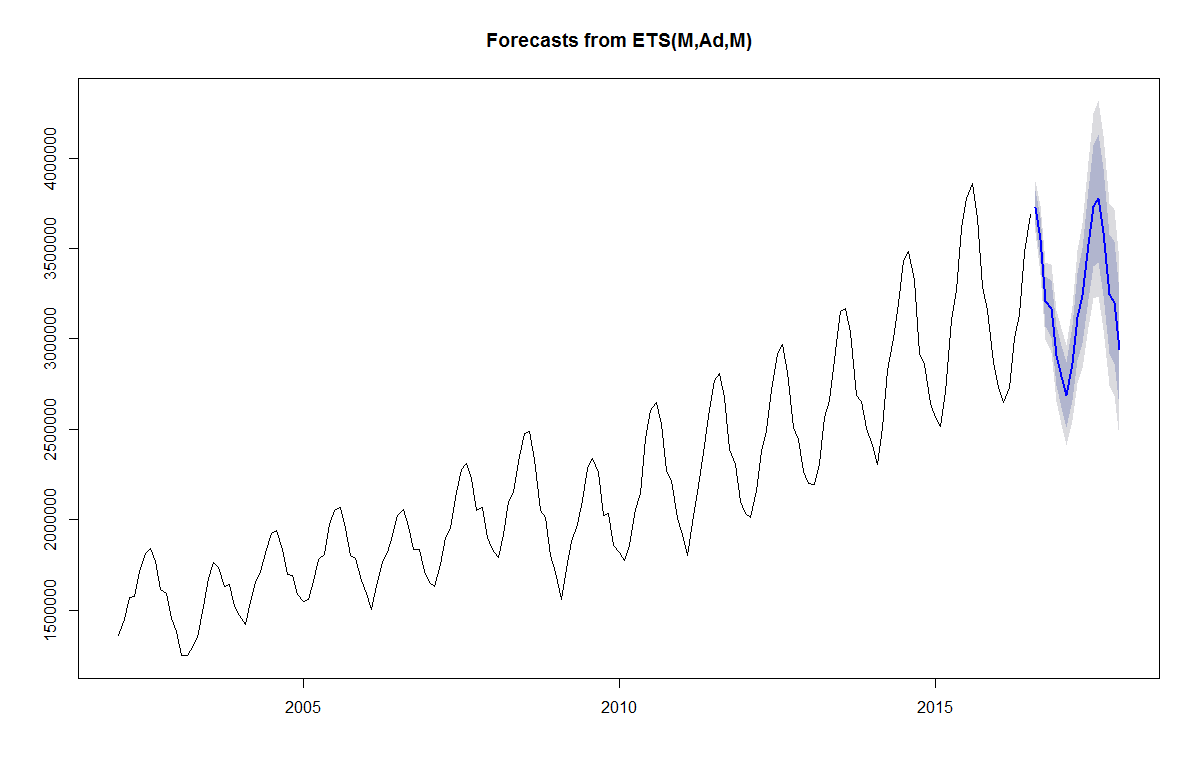
## Simple Moving Averages

* Plot the graph for time series.
* Show the Simple Moving average of order 3 on the plot above in Red
* Show the Simple Moving average of order 6 on the plot above in Blue
* Show the Simple Moving average of order 12 on the plot above in Green



* (Bonus) show the forecast of next 12 months using one of the simple average order that you feel works best for time series

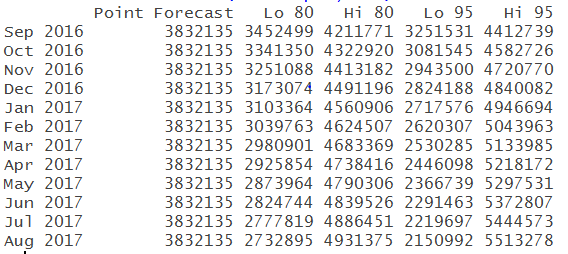


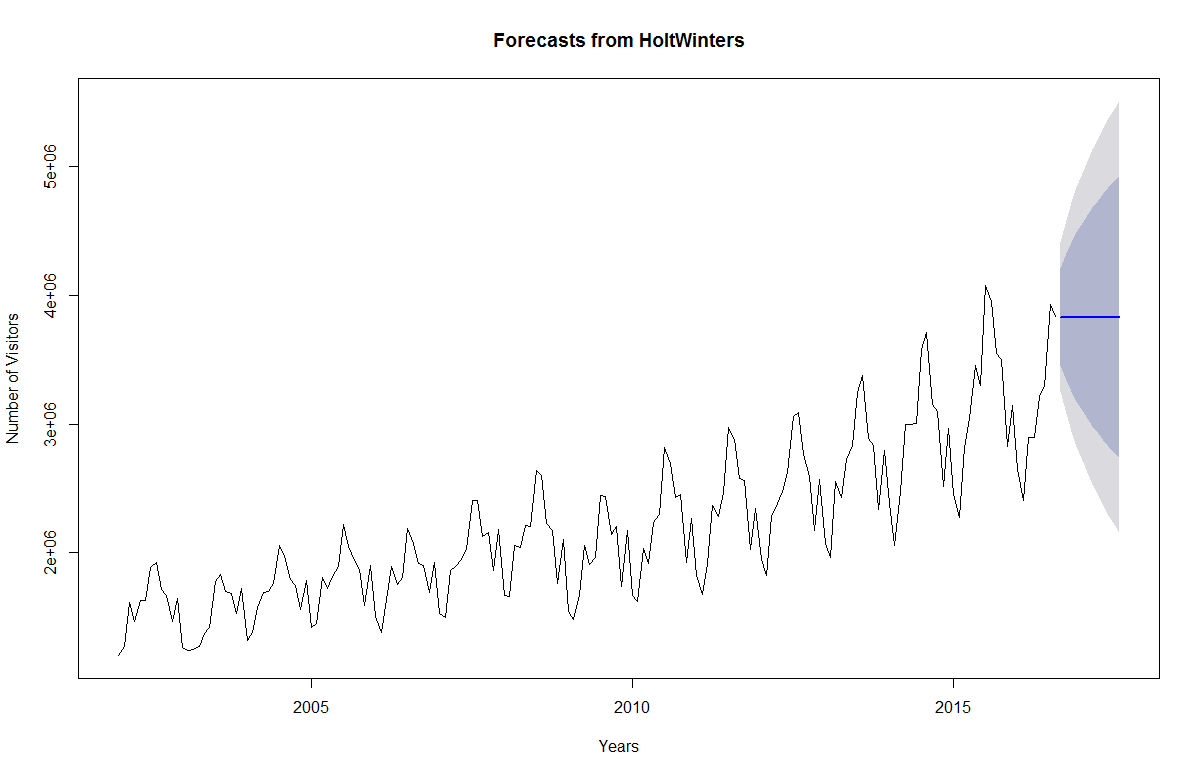


* What are your observations of the plot as the moving average order goes up?
* Smaller N in our case i.e. 3, tracks shifts in the level of a time series more quickly(randomness), while larger N’s in our case 12, are more effective in smoothing out random fluctuations over time.
* So as N increase from 3 to 12 it includes more observations to smooth out the random fluctuations in the data, thus providing a smooth line.
* Moreover, the more weight we would like to give to more recent data, the smaller N should be; correspondingly, the more past values are considered relevant, the larger N should be. So while our data is not stationary and has a upwards trend so more recent value shall be relevant and thus N should be small.

## Simple Smoothing

* Perform a simple smoothing forecast for next 12 months for the time series.





As we see from the forecasted graph that SSE just provides a steady straight line and does not seem to consider seasonal variation in the data, so it can be evidently that SSE is not a good forecasting method for our data and we shall focus on ETS.

* + What is the value of alpha? What does that value signify?

Smoothing factor or level factor or alpha(α) lies within 0 <= α <= 1

Notes:

* Large values of α yields forecasts which react quickly to shifts in the demand pattern, but exhibit more variation (less smoothing) from period to period.
* Small values of the smoothing constant α give greater weight to historical data (like large values of N in the moving average model), and hence exhibit relatively little sensitivity to variation.

Choose the alpha that gives smallest residual error.

SSE:

* Smoothing parameters:
* alpha: 0.8193105
* beta : FALSE
* gamma: FALSE
* As the value of alpha is very high the estimate of the current value of the are based mostly upon very recent observations in the time series.

ETS:

* Smoothing parameters:
* alpha = 0.4147
* gamma = 1e-04
* As the value of alpha is low the estimate of the current value of the are based mostly upon historic observations in the time series.
  + What is the value of initial state?

SSE:

* It is common in simple exponential smoothing to use the first value in the time series as the initial value for the level.
* ETS:

Initial states:

l = 1615326.4368

s=1.0256 0.8825 1.0596 1.0823 1.2087 1.2262

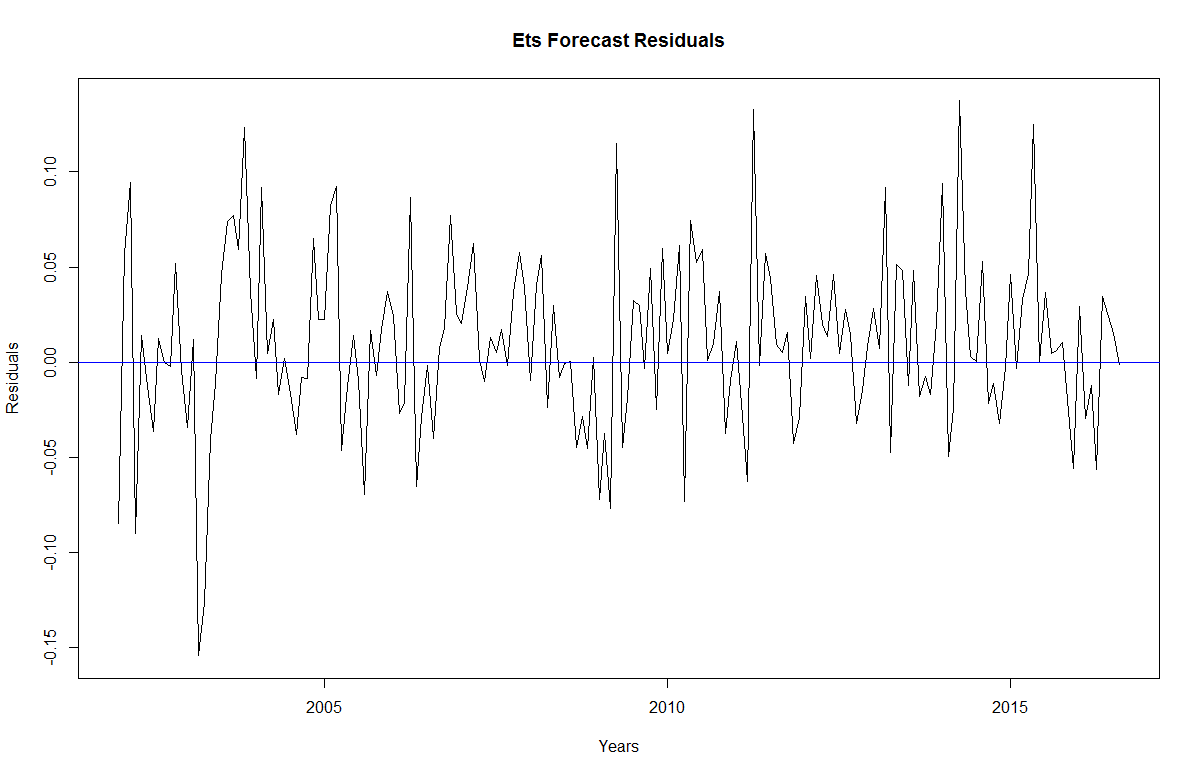
1.0286 1.0079 0.972 0.9245 0.7721 0.81

* + What is the value of sigma? What does the sigma signify?

sigma: 0.0469

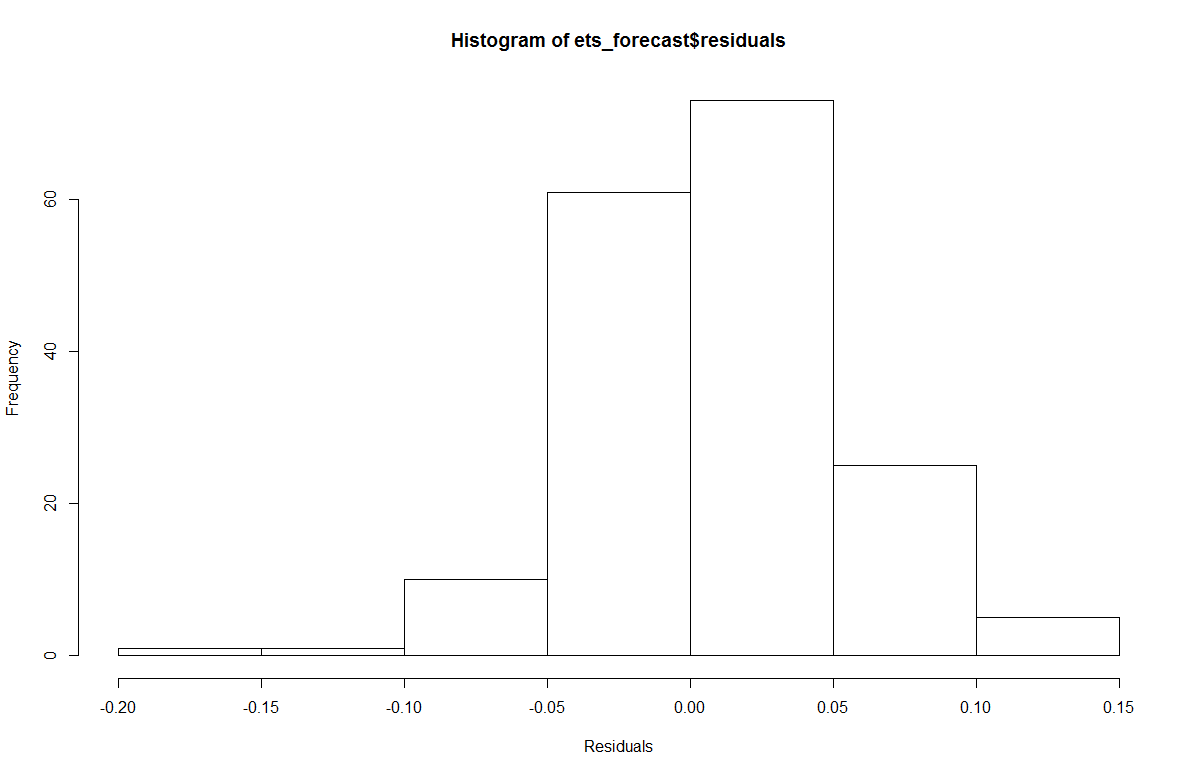
Standard deviation of residuals from the mean

* Perform Residual Analysis for this technique.
  + Do a plot of residuals. What does the plot indicate?

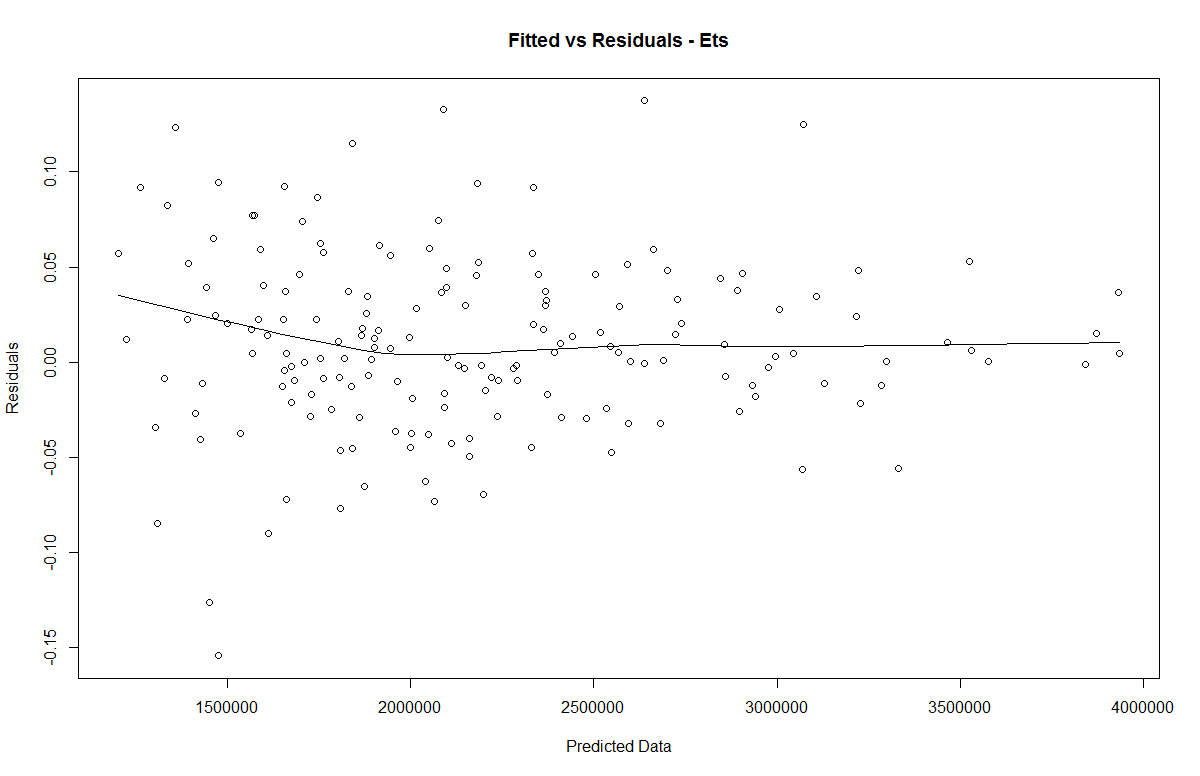


The plot shows that the forecast errors or residuals have constant variance over the time.

* + Do a Histogram plot of residuals. What does the plot indicate?

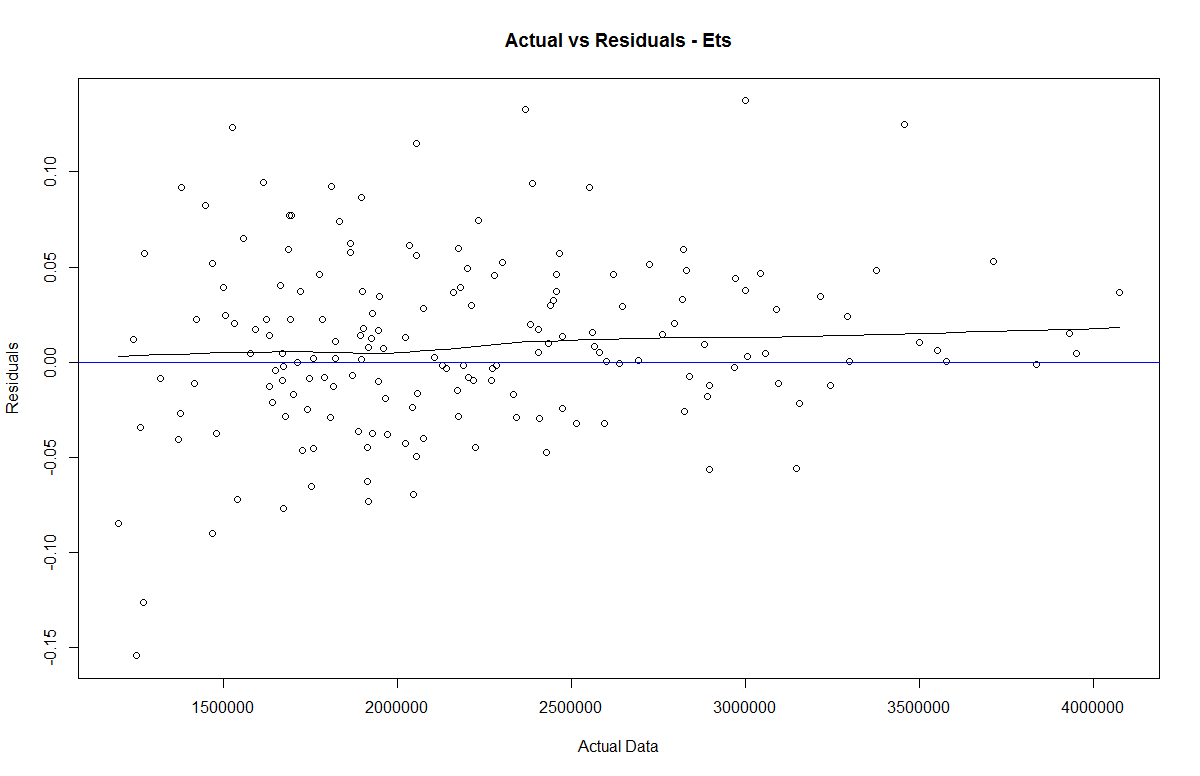


* The histogram of forecast errors show that the forecast errors are normally distributed with mean zero and constant variance.
* The data is slightly skewed towards the left.
  + Do a plot of fitted values vs. residuals. What does the plot indicate?

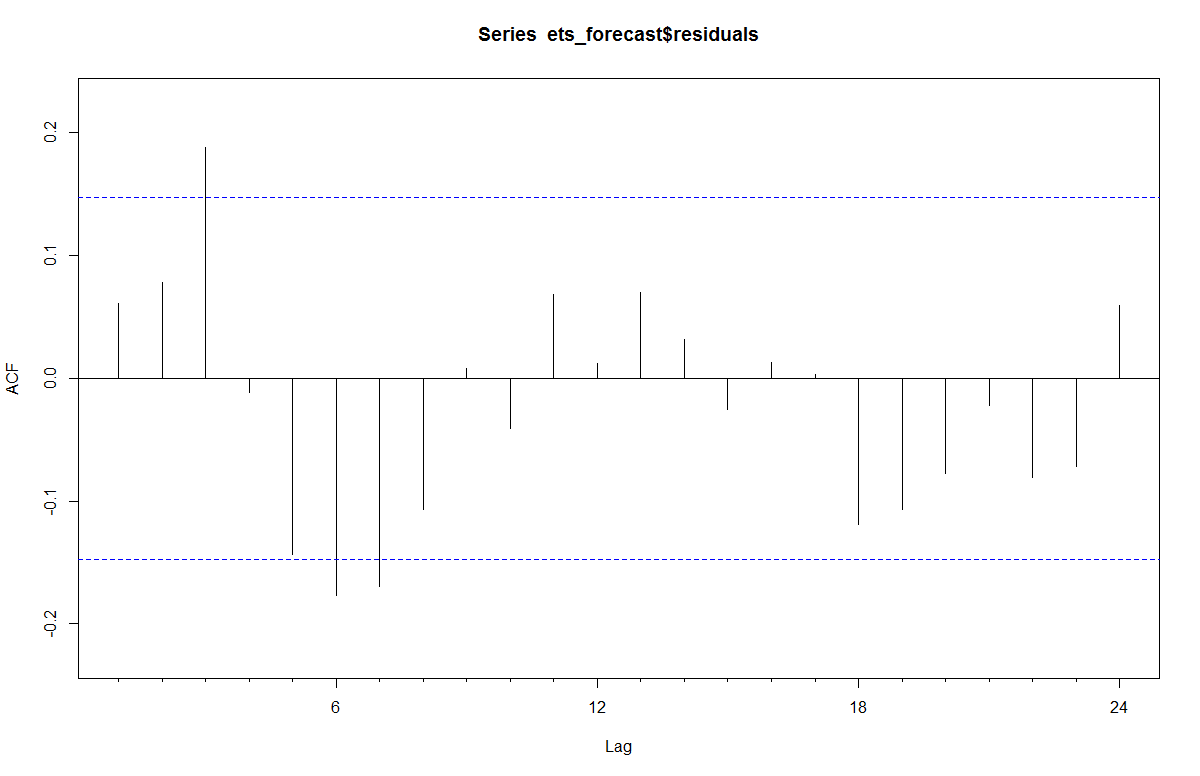


The errors have constant variance, with the residuals scattered randomly around zero. Though the variation of residuals decreases when the predicted data increases, the +ive and –ve values of residuals cancel each other to make mean error equal to zero.

* + Do a plot of actual values vs. residuals. What does the plot indicate?



* The errors have constant variance for the initials values of actual data, as the data is increased the +ve and –ve values of residuals cancel each other to make mean error equal to zero.
* For the higher values of the data the residuals are +ive and thereby indicating that the residuals are having some information.
* The plot also indicates presence of outliers.
  + Do an ACF plot of the residuals? What does this plot indicate?



* The correlogram provides the correlations between forecast errors for successive predictions. As autocorrelations at lag 3, 6 & 7crosses the significance bounds, this shows that there is correlation between successive forecast errors for successive predictions.
* The residuals are not random and there is some information available in the residuals.
* To test whether there is significant evidence for non-zero correlations at lags 1-20, we can carry out a Ljung-Box test. This can be done in R using the “Box.test()”, function.

Box-Ljung test

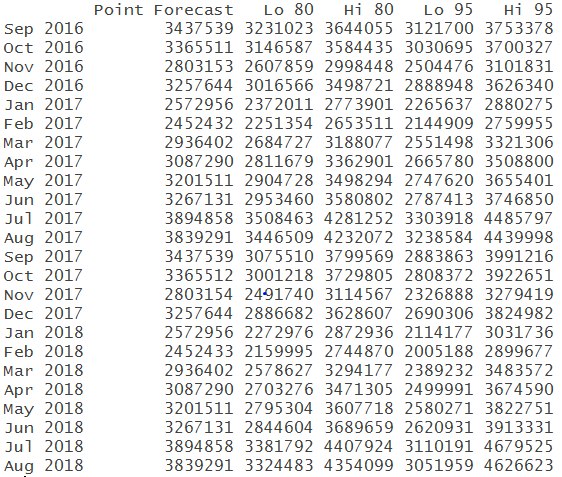
data: ets\_forecast$residuals

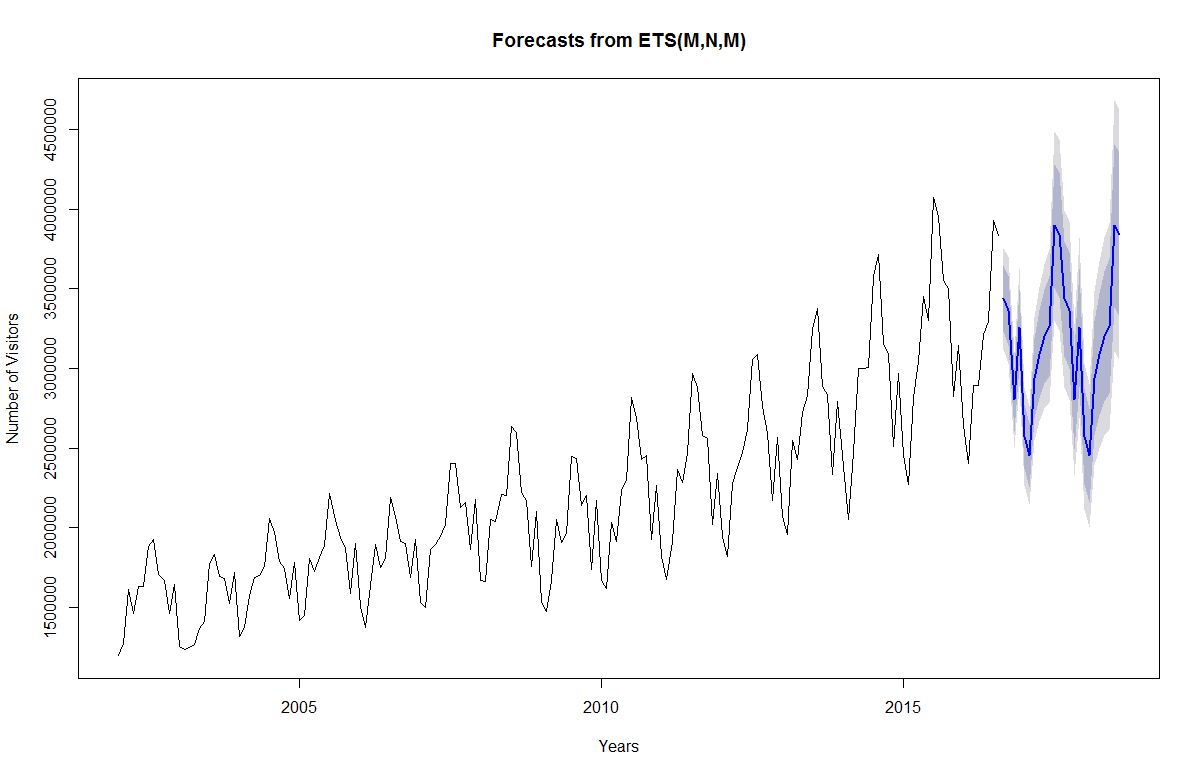
X-squared = 0.66496, df = 1, p-value = 0.4148

* Here the Ljung-Box test the p-value is 0.41 > 0.05, so there is little evidence of non-zero autocorrelations in the in-sample forecast errors
* Print the 5 measures of accuracy for this forecasting technique

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
| 21543.29 | 96382.89 | 72295.69 | 0.756152 | 3.456318 | 0.419878 | 0.0209 |

* Forecast
  + Time series value for next year. Show table and plot

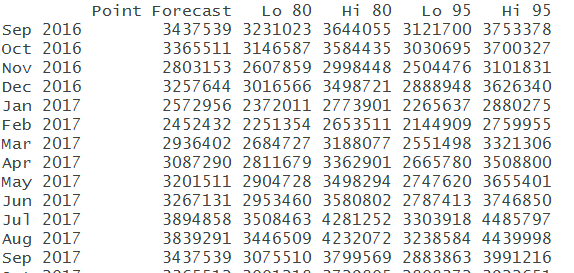


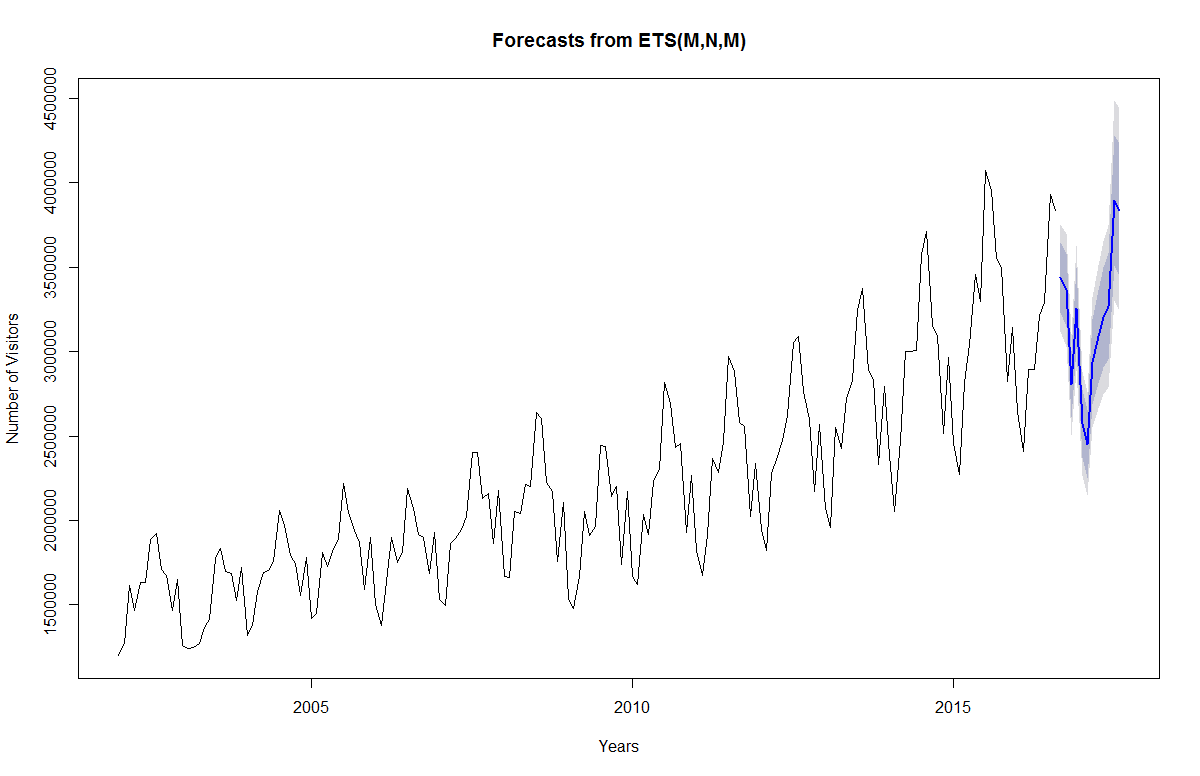


* Summarize this forecasting technique
  + How good is the accuracy

ETS has very good accuracy, residuals have constant variation, there is limited presence of any information in the error. Also the RMSE is low 96382.89.

* + What does it predict the value of time series will be in one year





* + Other observation

## Holt-Winters

* Perform Holt-Winters forecast for next 12 moths for the time series.
  + What is the value of alpha? What does that value signify?

Smoothing factor or level factor or alpha(α) = 0.3272698 0 <= α <= 1

It is the estimate of the level at the current time point.

The value of α=0 .3272698is 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.

Smoothing parameters:

alpha: 0.3272698

Coefficients:

[,1]

a 3009108.503

b 9420.709

s1 339813.279

s2 249777.147

s3 -387797.749

s4 1381.222

s5 -465808.327

s6 -668332.884

s7 -149157.471

s8 -15436.641

s9 313995.641

s10 314927.905

s11 929179.097

s12 824051.837

* + What is the value of beta? What does that value signify?
* Trend factor or beta: 0.02480278
* Determines the degree of ascent or decent value that should be used to adjust the forecast. This tell the system how much to focus on recent trend
* The value of beta is 0.02480278, indicates that the estimate of the slope of the trend component is not updated over the time series, and instead is set equal to its initial value
  + What is the value of gamma? What does that value signify?

Seasonality factor or gamma: 0.7684698

* Seasonal component of the forecast. Higher the parameter more the recent seasonal component is weighted.
* The value of gamma=0.76846 is relatively high, indicating that the estimate of the seasonal component at the current time point is based highly upon the recent observations and on few observations in the distant past.
  + What is the value of initial states for the level, trend and seasonality? What do these values signify?

alpha 3009108.503

beta 9420.709

seasonality 339813.279

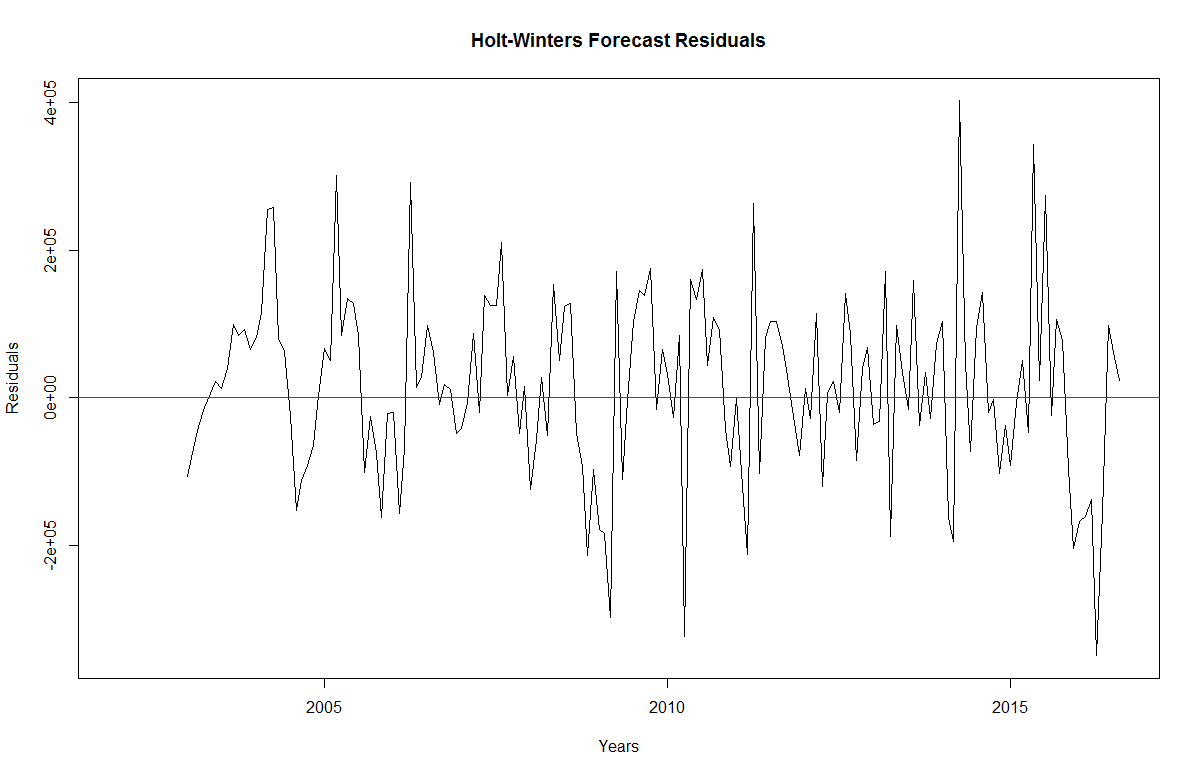
This values are the initial values used by the forecasting algorithm which

are eventually changed to reduce the error

* + What is the value of sigma? What does the sigma signify?

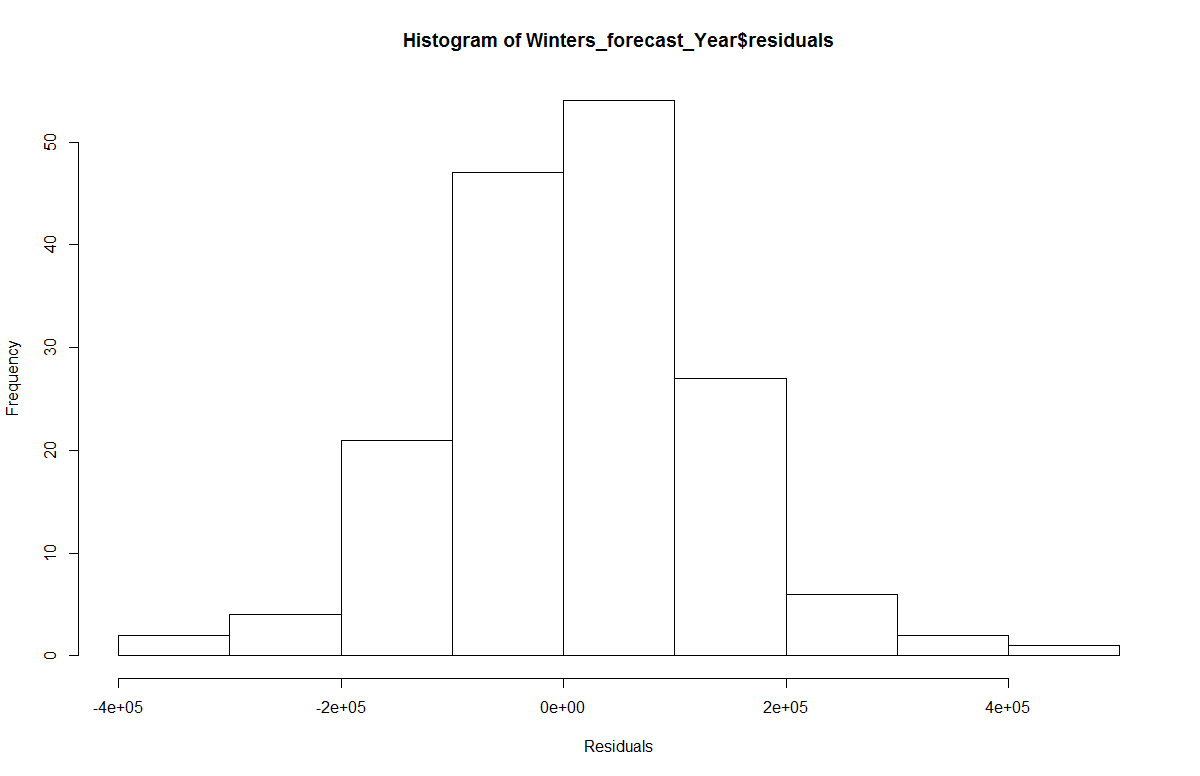
N/A

* Perform Residual Analysis for this technique.
  + Do a plot of residuals. What does the plot indicate?



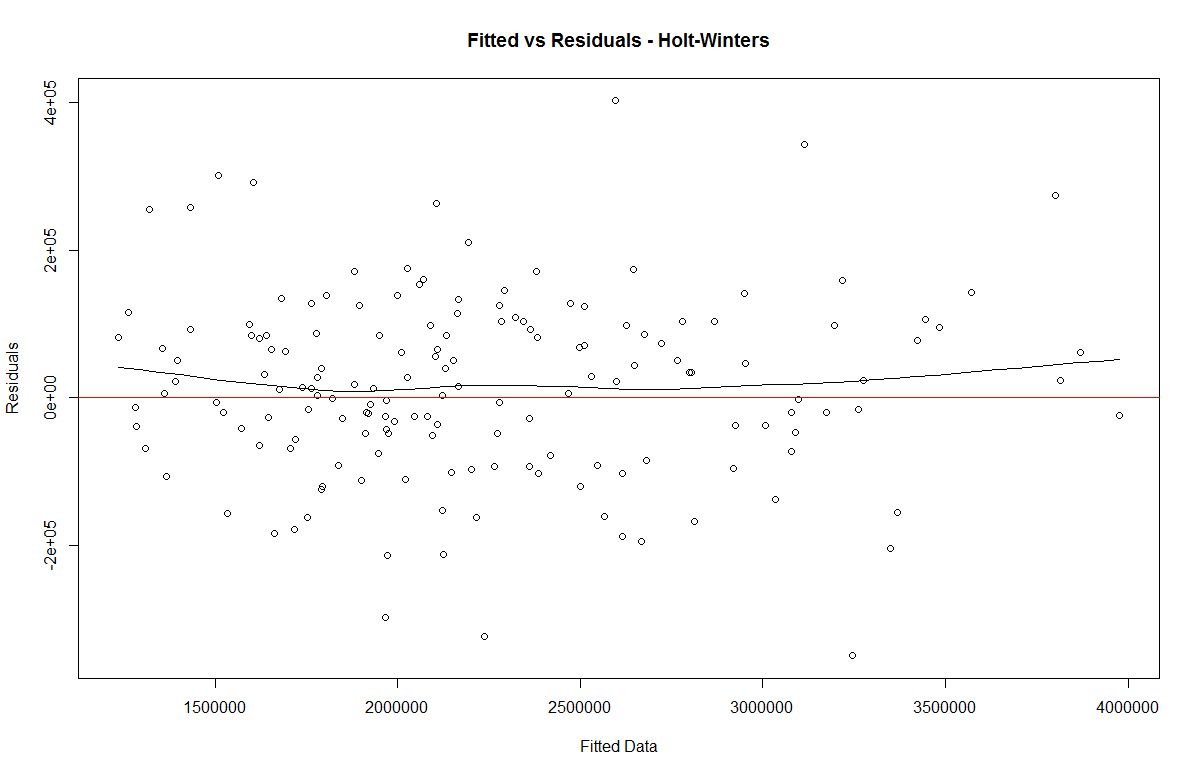
The plot shows that the forecast errors or residuals have roughly constant variance over the time.

* + Do a Histogram plot of residuals. What does the plot indicate?

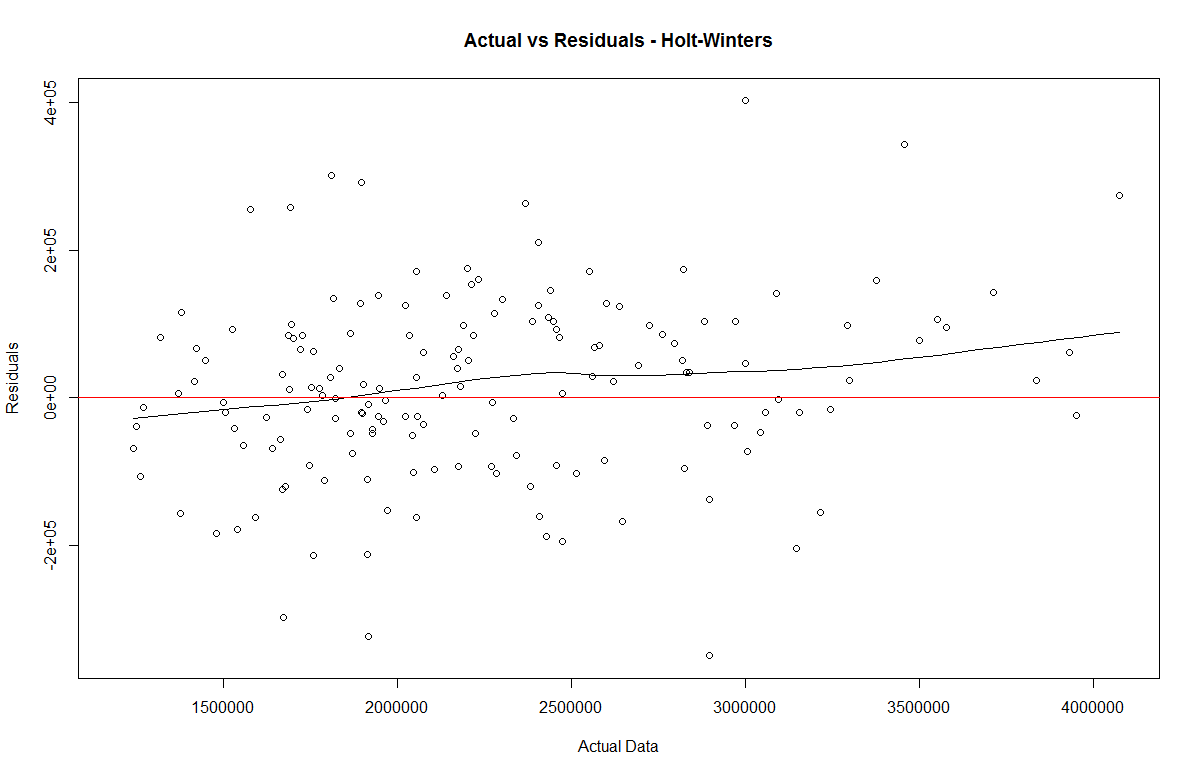


The histogram of forecast errors show that the forecast errors are normally distributed with mean zero and constant variance.

* + Do a plot of fitted values vs. residuals. What does the plot indicate?

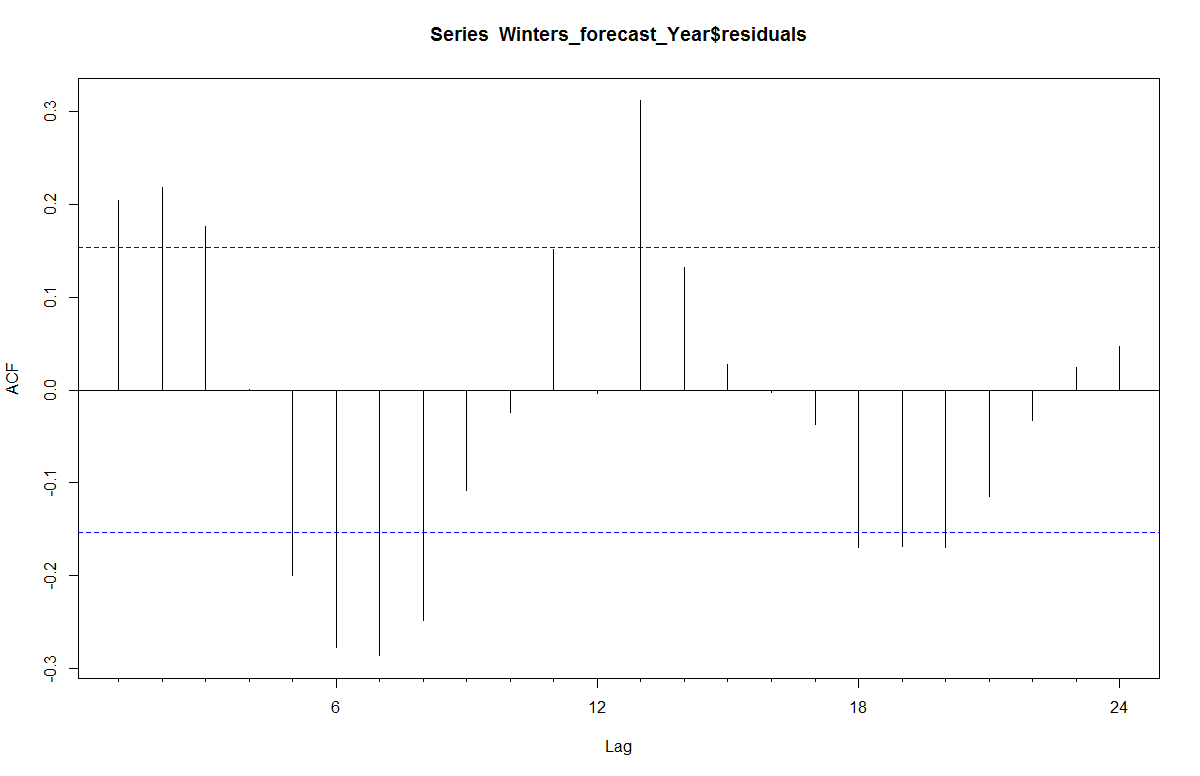


* The errors have constant variance, with the residuals scattered randomly around zero.
* Though the variation of residuals decreases when the predicted data increases, the +ve and –ve values of residuals cancel each other to make mean error equal to zero. The plot also indicates presence of outliers.
  + Do a plot of actual values vs. residuals. What does the plot indicate?



The errors have constant variance, with the residuals scattered randomly around zero. The plot also indicates presence of outliers.

* + Do an ACF plot of the residuals? What does this plot indicate?



* The correlogram provides the correlations between forecast errors for successive predictions. As autocorrelations at lag1,2,3 – 5,6,7,8 & 13 crosses the significance bounds, this shows that there is correlation between successive forecast errors for successive predictions.
* To test whether there is significant evidence for non-zero correlations at lags 1-20, we can carry out a Ljung-Box test. This can be done in R using the “Box.test()”, function.

Box-Ljung test

data: holts\_f$residuals

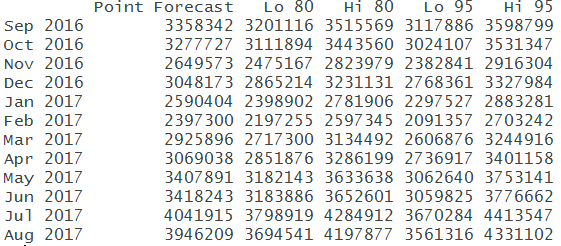
X-squared = 6.9391, df = 1, p-value = 0.008433

As p-value (0.008) is less than 0.05 i.e. there is enough evidence of the correlation between the residuals.

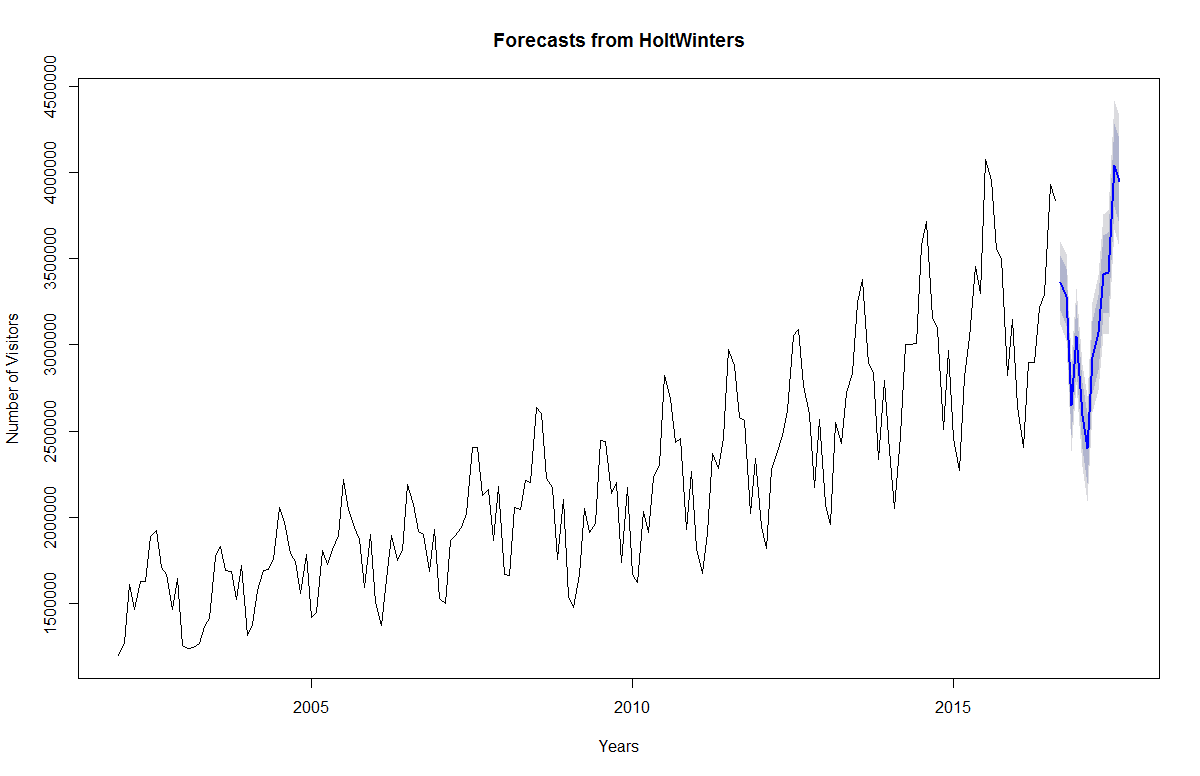
* Print the 5 measures of accuracy for this forecasting technique

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
| 15175.03 | 123247.5 | 95446.26 | 0.446107 | 4.438431 | 0.554332 | 0.20383 |

* Forecast
  + Time series value for next year. Show table and plot



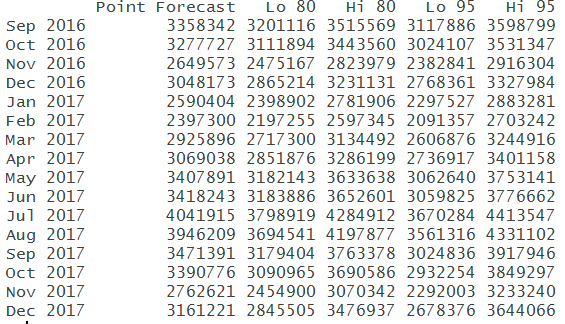
Forecast for a year, an 80% prediction interval for the forecast, and a 95% prediction interval for the forecast. For example, the forecasted visitors for Jan 2017 is about 2590404 claims, with a 95% prediction interval of (2297527, 2883281).

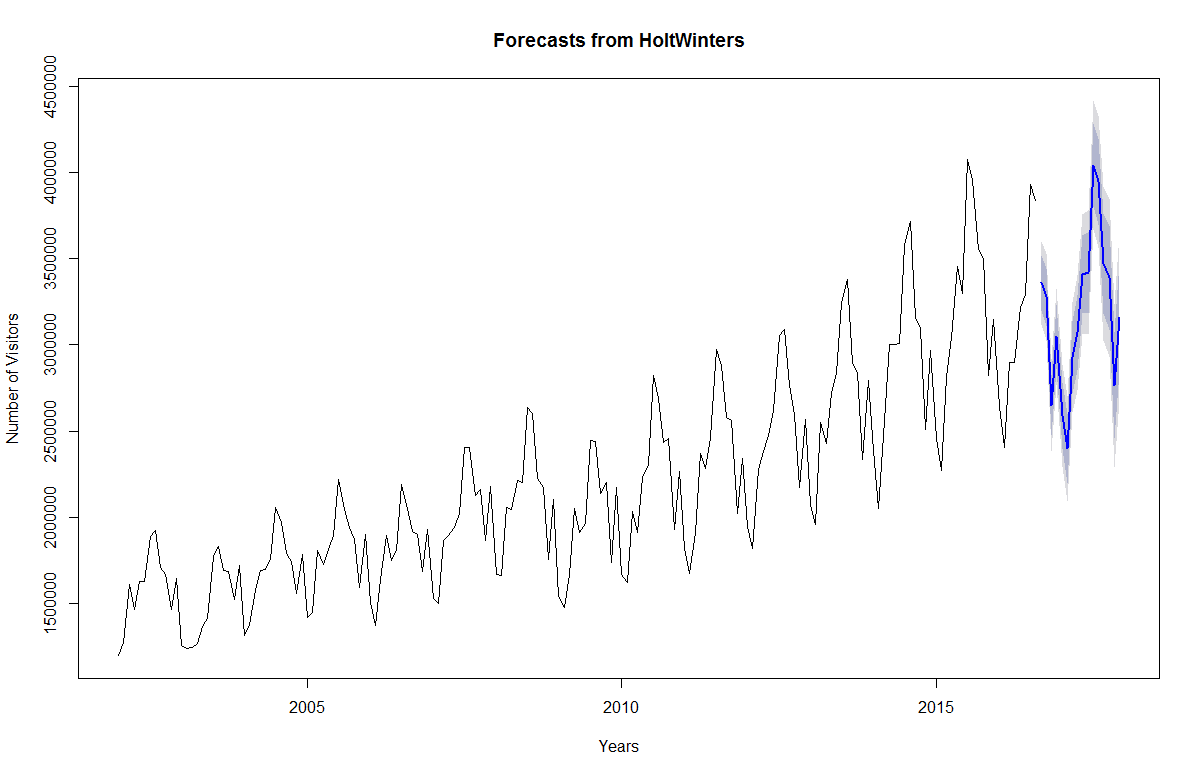


* Summarize this forecasting technique
  + How good is the accuracy?

The accuracy of HoltWinters forecasting method is fairly good (RMSE= 123247.5) but less as compared to ETS, the values for smoothing constants can be changed to reduce the error.

* + What does it predict the value of time series will be in one year





* + Other observation

## ARIMA or Box-Jenkins

* Is Time Series data Stationary? How did you verify? Please post the output from one of the test.

The time series is not stationary as it has upward trend and seasonality as discussed in the previous sections. This can be verified by KPSS test for Level Stationarity, as the p-value (0.01) is less than 0.05, therefore differences are required to make the data stationary.

KPSS Test for Level Stationarity

data: data\_ts

KPSS Level = 3.5156, Truncation lag parameter = 3, p-value = 0.0

* How many differences are needed to make it stationary?

nsdiffs(data\_ts)

[1] 1

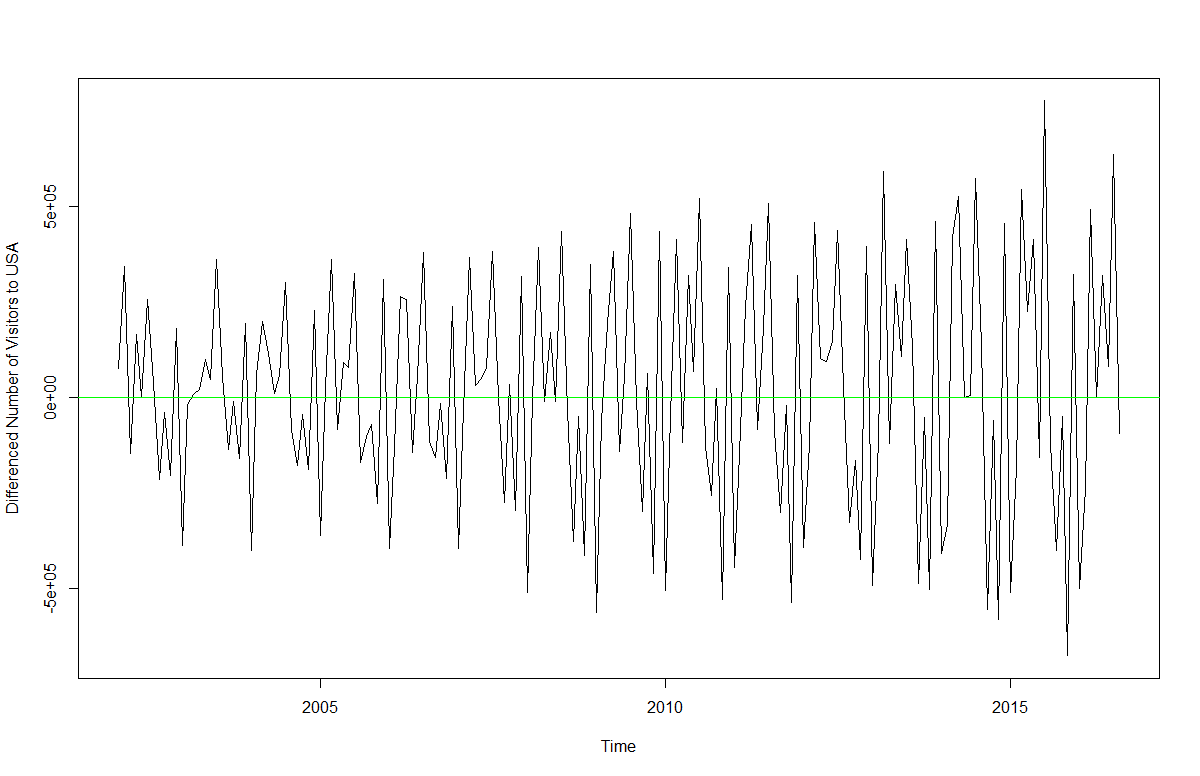
* Is Seasonality component needed?

No the seasonality component is not required for stationary data

* Plot the Time Series chart of the differenced series.

##### Difference data to make data stationary on mean (remove trend)

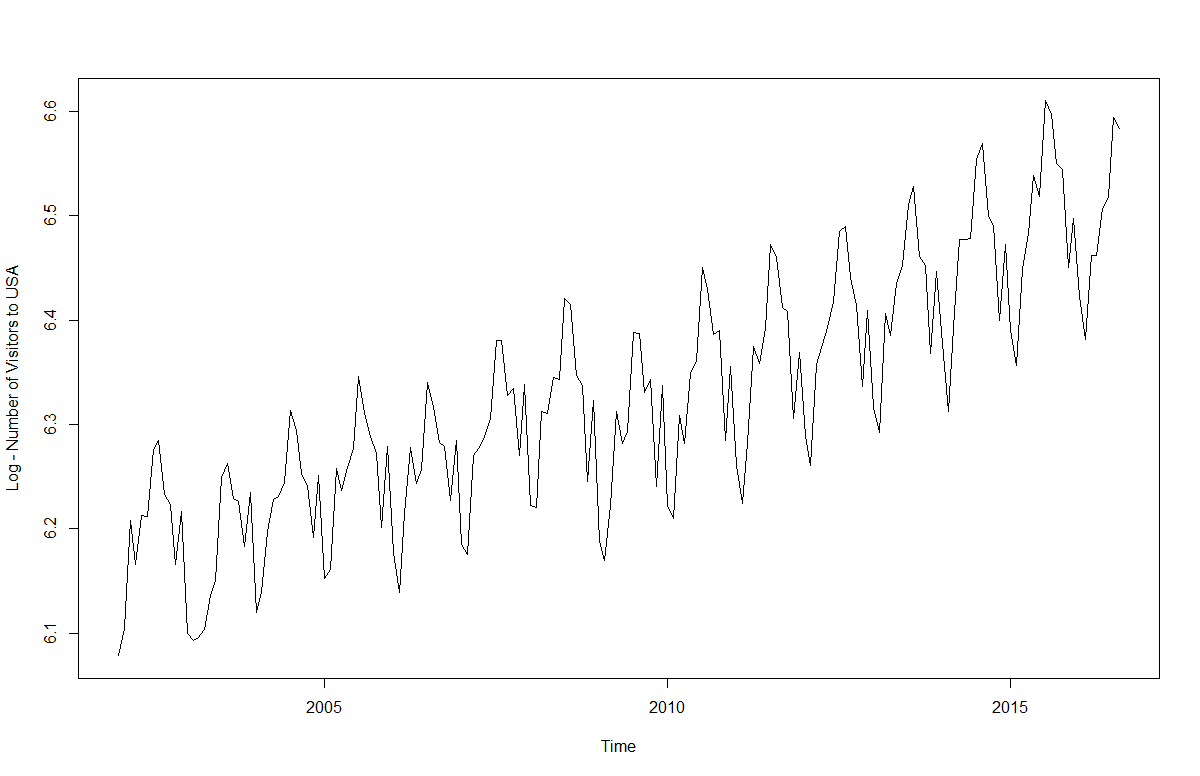
plot(diff(data\_ts),ylab="Differenced Number of Visitors to USA")



The above series is not stationary on 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 variance to produce reliable forecasts through ARIMA models.

##### log transform data to make data stationary on variance

plot(log10(data\_ts),ylab="Log - Number of Visitors to USA")

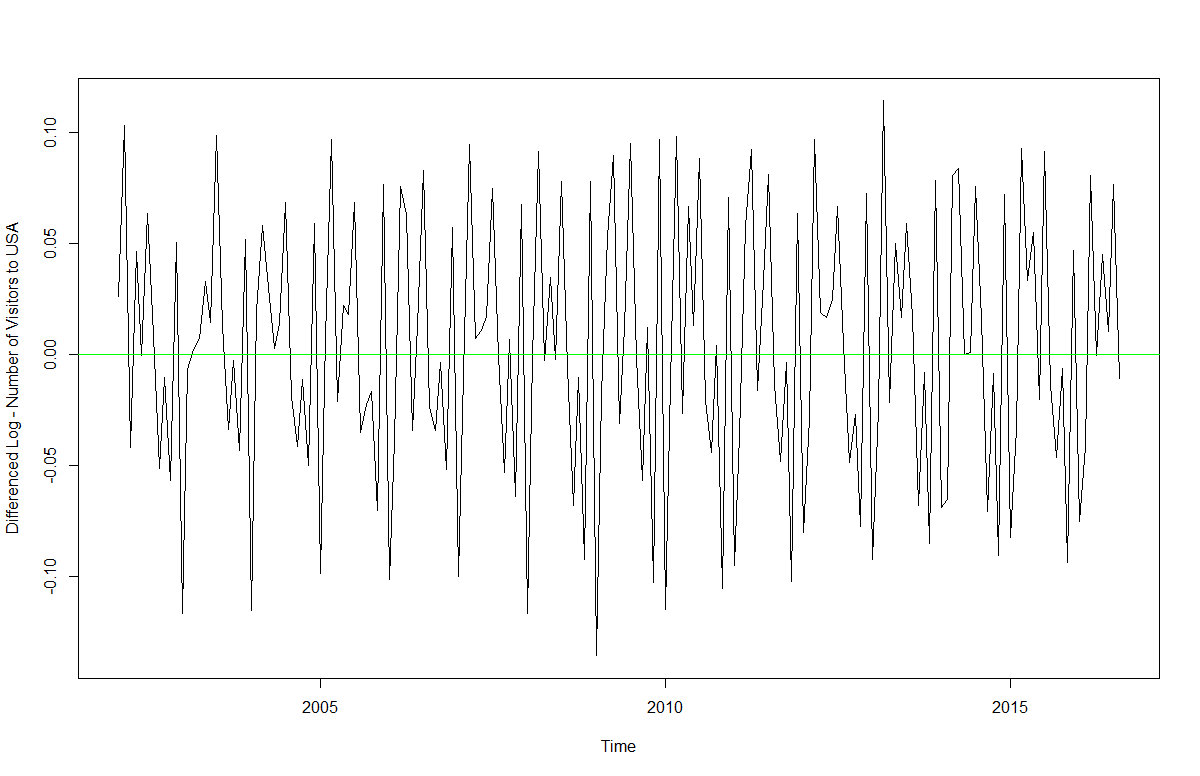


Now the series looks stationary on variance.

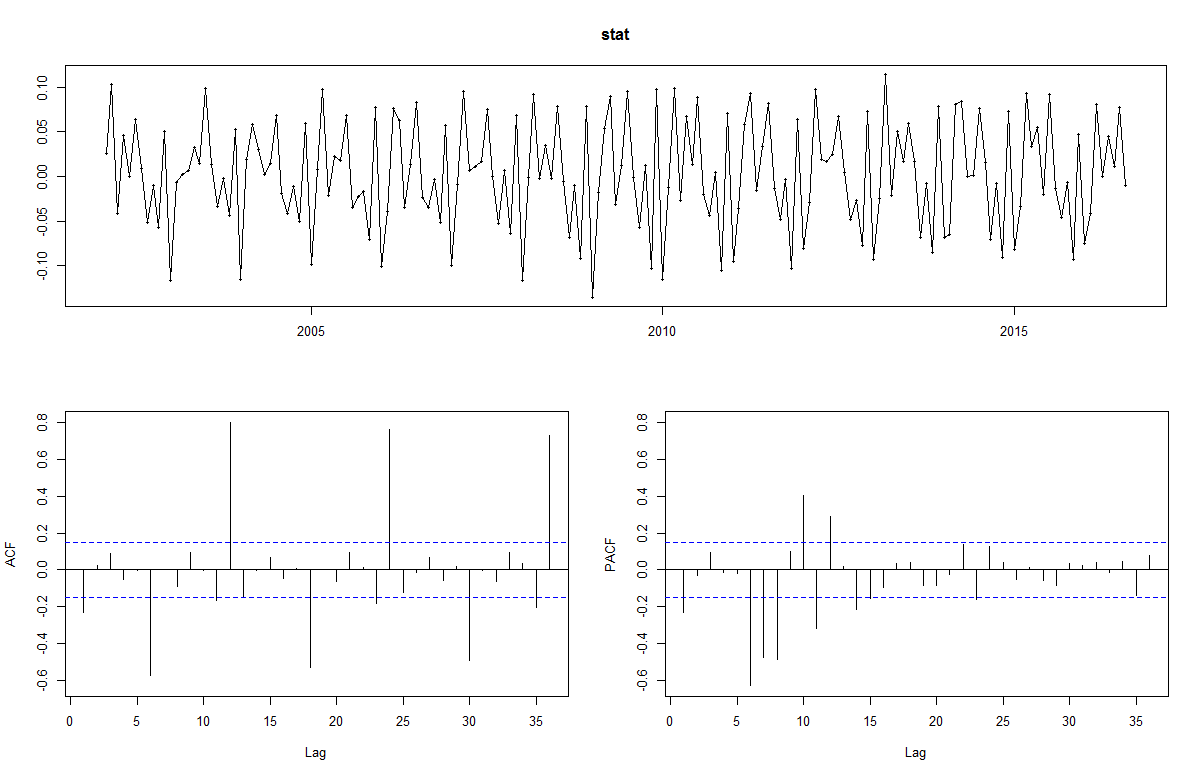
##### Difference log transform data to make data stationary on both mean and variance

stat = log10(diff(data\_ts))

plot(stat,ylab="Differenced Log - Number of Visitors to USA")



* Now the series looks stationary on both mean and variance.
* This also gives us the hint that I or integrated part of our ARIMA model will be equal to 1 as 1st difference is making the series stationary.
* Plot the Acf and Pacf plot of the differenced series.



* Since there are a number of spikes in the plots crossing the significant values (dotted horizontal lines) we can conclude that the residuals are not random and are autocorrelated. 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 12,24 & 36 represented by spikes at corresponding lags. This makes sense since we are analyzing monthly data that tends to have seasonality of 12 months.
* Based off the ACF and PACF, which are the possible ARIMA model possible?

Model 1:

ARIMA(0,1,2)(2,1,1) [12]

p = 0 (as the only spike is crossing the significant level in PACF

initial lag values so either 0 or 1)

P = 2 (as more than 2 spikes are crossing significant level at 6 ,7,8 & 10,11,12 in PACF)

d,D = 1 as difference of 1 is required

q = 2 (as a spike is crossing significant level at 1 so 1 or 2)

Q = 1 (as spikes are crossing significant levels at 6,18,30 & 12 ,24,36 in ACF so 1 or 2)

Model2:

ARIMA(1,1,2)(2,1,2)[12]

p = 1 (as only spike is crossing the significant level in PACF

initial lag values so either 0 or 1)

P = 2 (as more than 2 spikes are crossing significant level at 6 ,7,8 & 10,11,12 in PACF)

d,D = 1 as difference of 1 is required

q = 2 (as a spike is crossing significant level at 1 so 1 or 2)

Q = 2 (as spikes are crossing significant levels at 6,18,30 & 12 ,24,36 in ACF so 1 or 2)

* Show the AIC, BIC and Sigma^2 for the possible models?

Model 1:

Series: data\_ts

ARIMA(0,1,2)(2,1,1)[12]

Coefficients:

ma1 ma2 sar1 sar2 sma1

-0.5586 0.2425 -1.1258 -0.6276 0.5921

s.e. 0.0908 0.1114 0.0910 0.0624 0.1163

sigma^2 estimated as 1.022e+10: log likelihood=-2112.69

AIC=4237.39 AICc=4237.93 BIC=4255.95

Model2:

Series: data\_ts

ARIMA(1,1,2)(2,1,2)[12]

Coefficients:

ar1 ma1 ma2 sar1 sar2 sma1 sma2

0.3058 -0.8711 0.4949 -1.1237 -0.6055 0.5942 -0.0347

s.e. 0.1420 0.1436 0.1718 0.1184 0.1592 0.1566 0.2338

sigma^2 estimated as 1.018e+10: log likelihood=-2111.5

AIC=4238.99 AICc=4239.93 BIC=4263.74

* Based on the above AIC, BIC and Sigma^2 values, which model will you select?

I will select the model with least AIC & BIC, so I will select model 1

* What is the final formula for ARIMA with the coefficients?

Series: data\_ts

ARIMA(0,1,3)(2,1,1)[12]

Coefficients:

ma1 ma2 ma3 sar1 sar2 sma1

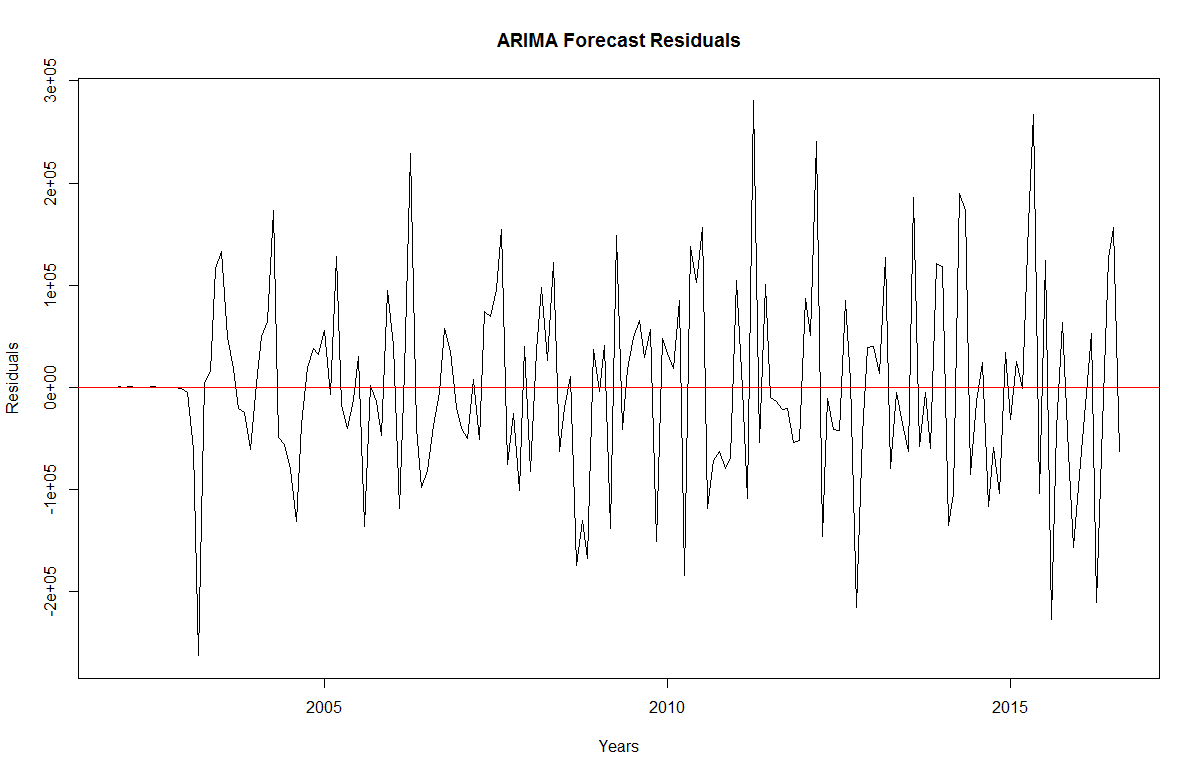
-0.5191 0.2406 0.3137 -1.1627 -0.6272 0.6237

s.e. 0.0746 0.0834 0.0940 0.0943 0.0615 0.1258

sigma^2 estimated as 9.774e+09: log likelihood=-2108.86

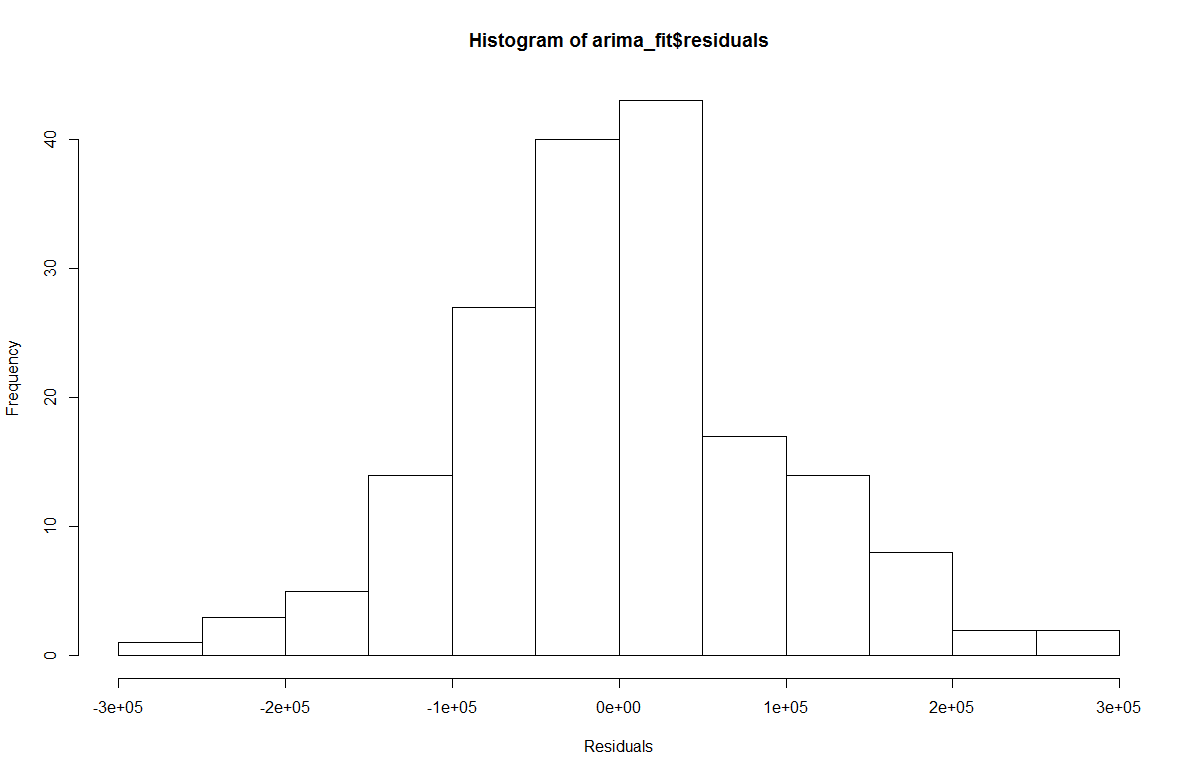
AIC=4231.73 AICc=4232.45 BIC=4253.38

* Perform Residual Analysis for this technique.
  + Do a plot of residuals. What does the plot indicate?



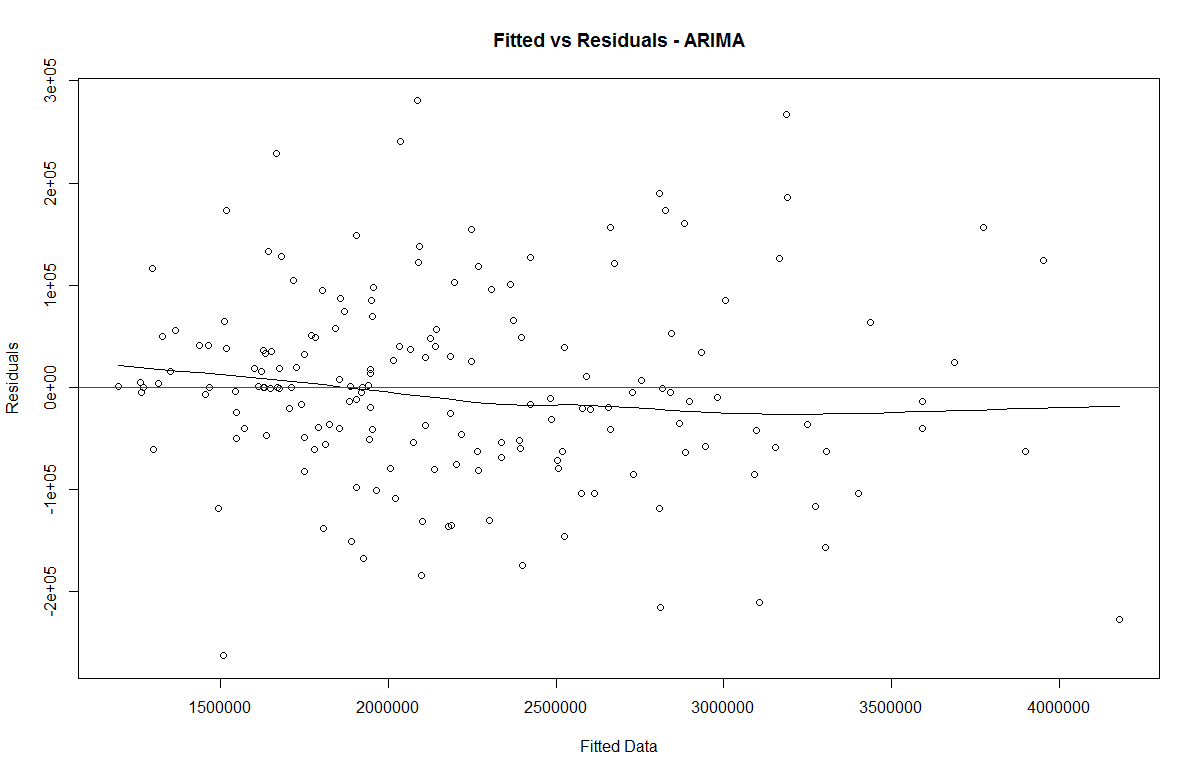
The plot shows that the forecast errors or residuals have constant variance over the time

* + Do a Histogram plot of residuals. What does the plot indicate?



The histogram of forecast errors show that the forecast errors are normally distributed with mean zero and constant variance.

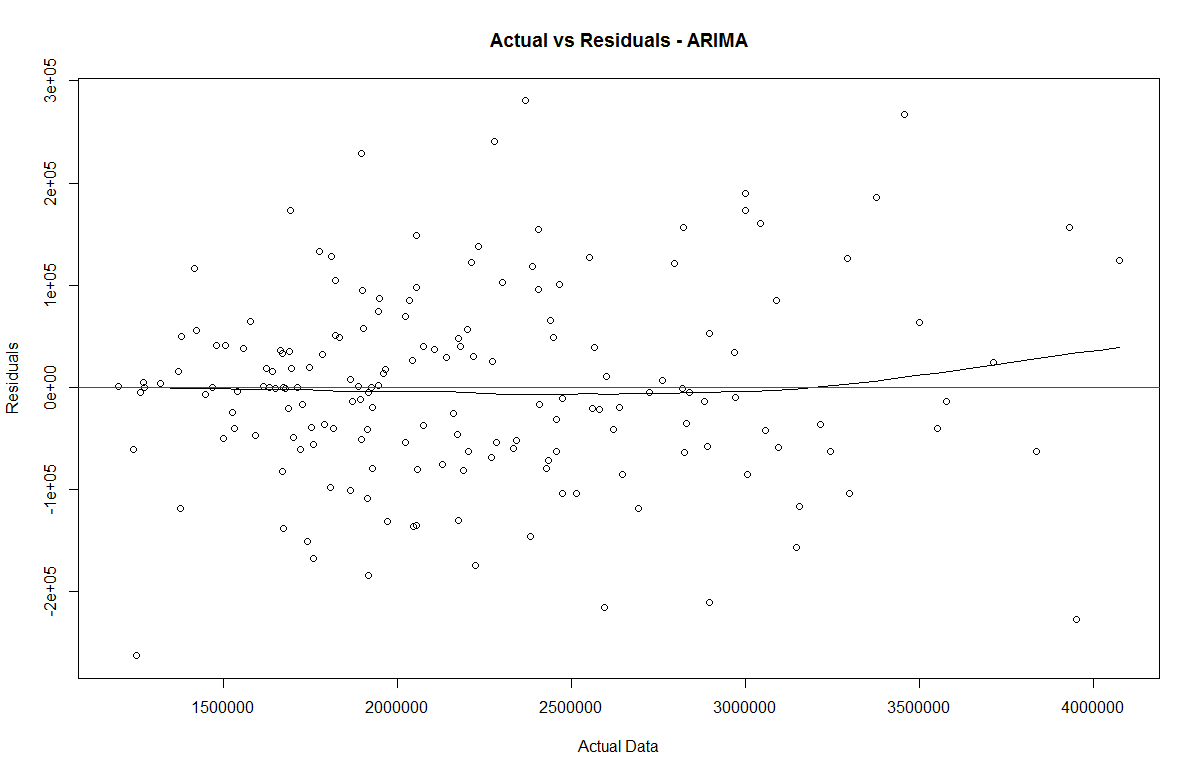
* + Do a plot of fitted values vs. residuals. What does the plot indicate?



The errors have constant variance, with the residuals scattered randomly around zero.

The plot also indicates presence of outliers.

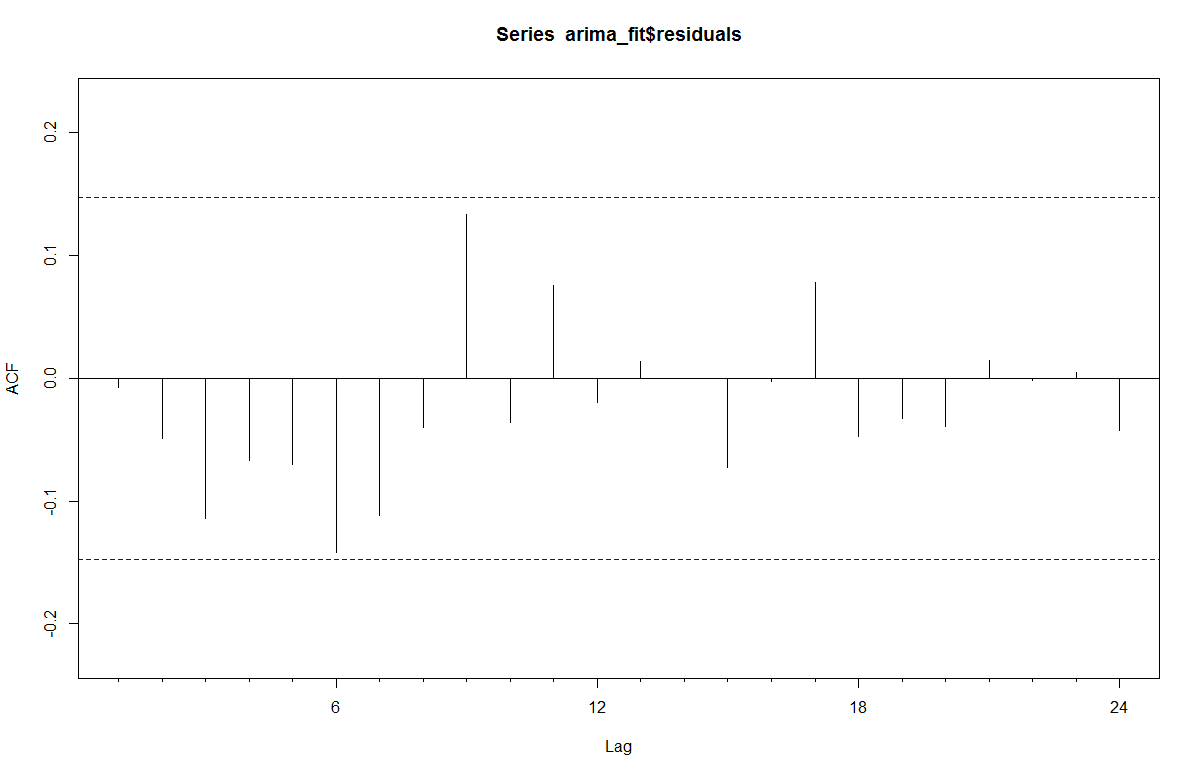
* + Do a plot of actual values vs. residuals. What does the plot indicate?



The errors have constant variance, with the residuals scattered randomly around zero.

The plot also indicates presence of outliers.

* + Do an ACF plot of the residuals? What does this plot indicate?

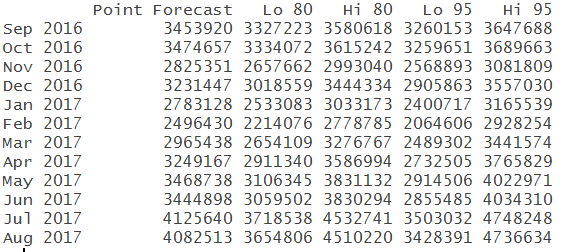


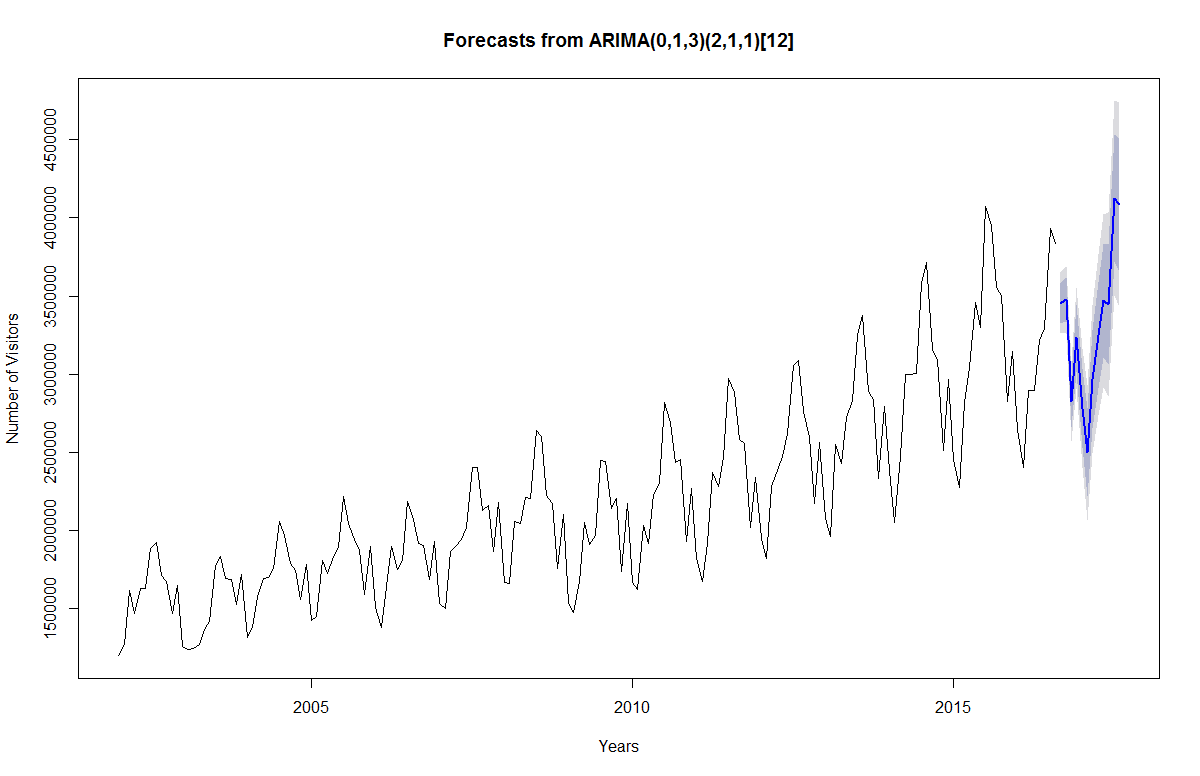
As no autocorrelations crosses the significance bounds, this shows that there is no correlation between successive forecast errors for successive predictions.

* Print the 5 measures of accuracy for this forecasting technique.

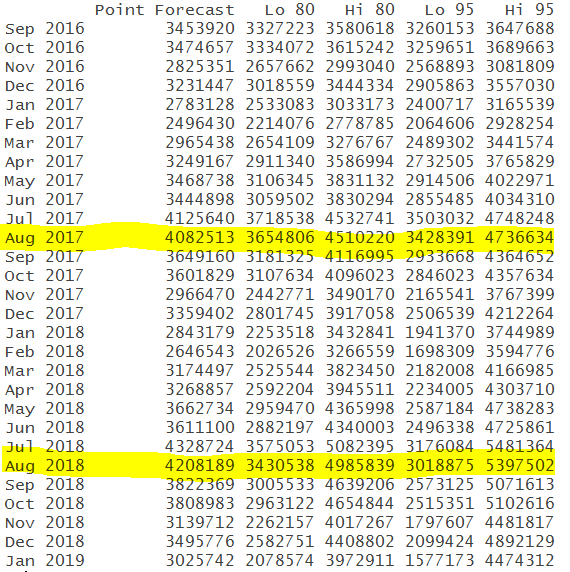
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
| 186.607 | 93373.77 | 70007.5 | -0.08475952 | 3.192901 | 0.406589 | -0.00709 |

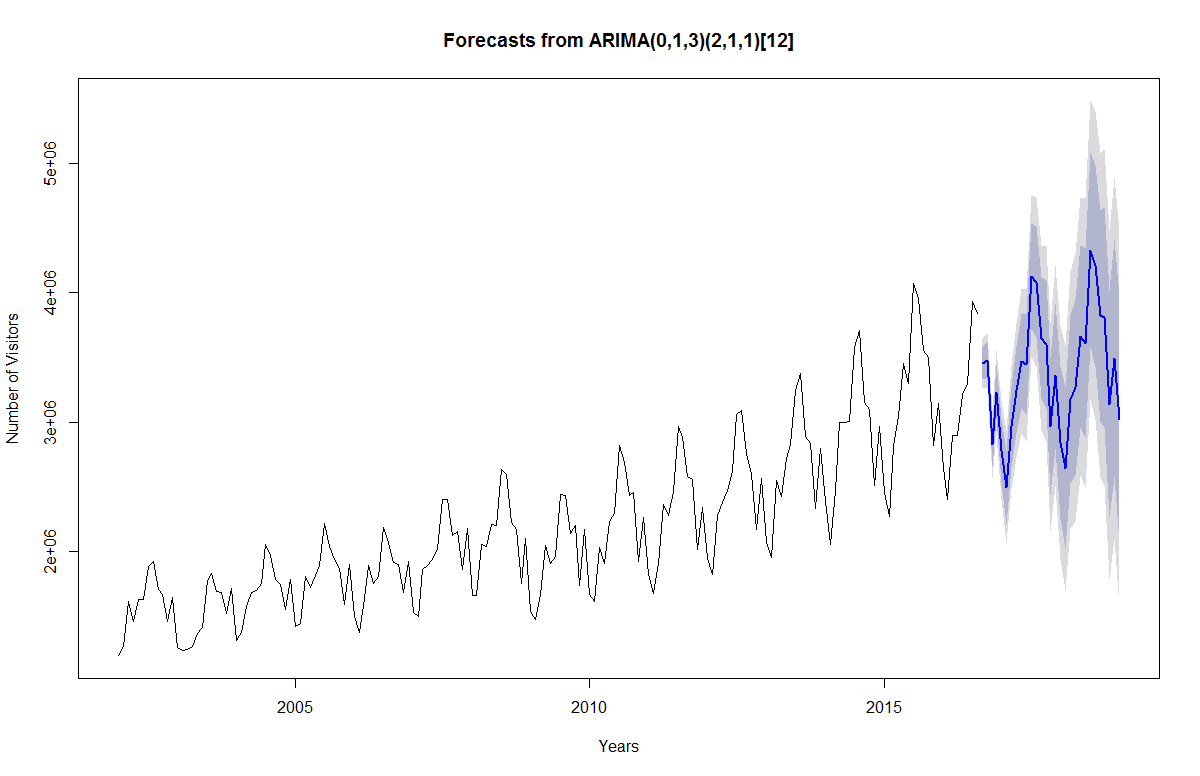
* Forecast
  + Next one year. Show table and plot





* + Next two years. Show table and plot

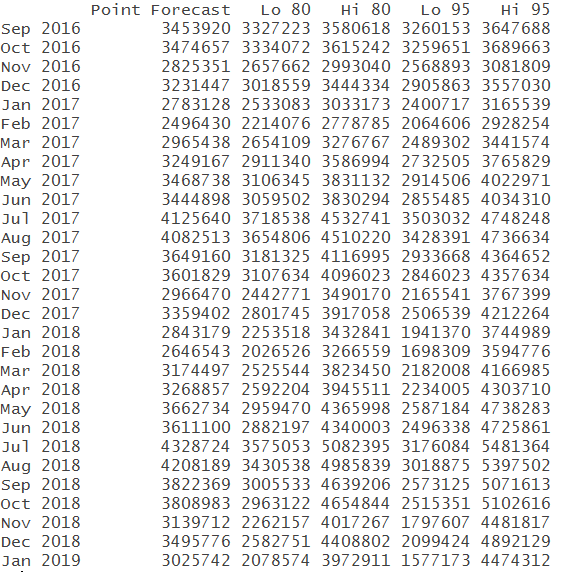


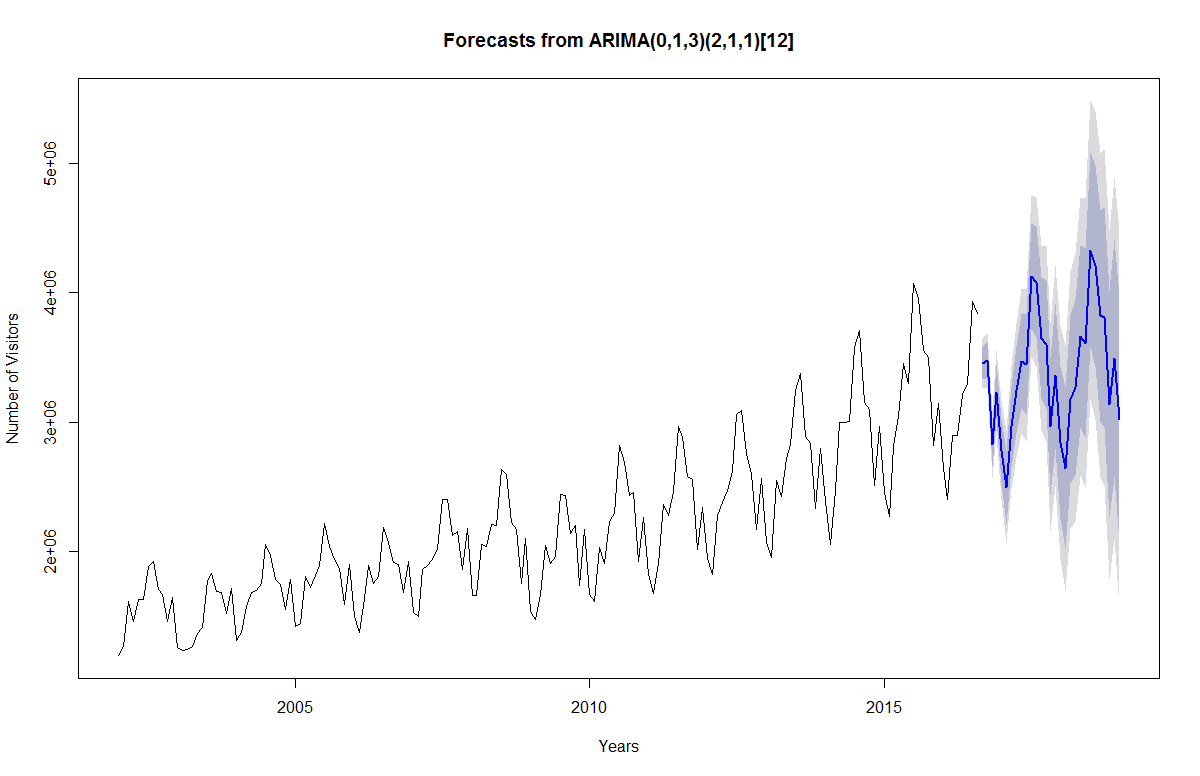


* Summarize this forecasting technique
  + How good is the accuracy?

The accuracy of Arima forecasting method is very good (RMSE=93373.77) least as compared to Ets, Holt-Winters and Snaive.

* + What does it predict time series will be in one year and next two years?





* + Other observation

## Accuracy Summary

* Show a table of all the forecast method above with their accuracy measures.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| F Methods | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
| Snaive | 115998.2 | 205853.8 | 172182.5 | 4.595344 | 7.741379 | 1 | 0.57322 |
| ETS | 21543.29 | 96382.89 | 72295.69 | 0.756152 | 3.456318 | 0.41987 | 0.0209 |
| Holt-Winters | 15175.03 | 123247.5 | 95446.26 | 0.4461 | 4.438431 | 0.55433 | 0.20383 |
| Arima | 186.607 | 93373.77 | 70007.5 | -0.08475 | 3.192901 | 0.40658 | -0.00709 |

* Separately define each forecast method and why it is useful. Show the best and worst forecast method for each of the accuracy measures.
* Naïve Forecast: Naïve models are often used when no trend is apparent in the data. They are descriptive techniques which generate only a point estimate for the forecast, not a confidence interval reflecting a range of values and an associated level of confidence
* Simple Smoothing: In simple smoothing model we use a series of decreasing weights, giving the most weight to the most recent observation. Unlike the moving average model every data point, not just the last N data points, gets a weight, with more recent points getting more weight and the weights decreasing (geometrically) with increasing age of the data, as we move back in time.
* Holts Winter: It is appropriate when there is a trend in data. It is also known as double exponential smoothing.
* Ets takes care of the seasonality in the data, it also seems to be fair forecasting method as highlighted in the table able as it has least value for MPE.
* Non-seasonal ARIMA models are generally denoted ARIMA(*p*,*d*,*q*) where [parameters](https://en.wikipedia.org/wiki/Parameter) *p*, *d*, and *q* are non-negative integers, *p* is the order (number of time lags) of the [autoregressive model](https://en.wikipedia.org/wiki/Autoregressive_model), *d* is the degree of differencing (the number of times the data have had past values subtracted), and *q* is the order of the [moving-average model](https://en.wikipedia.org/wiki/Moving-average_model). Seasonal ARIMA models are usually denoted ARIMA(*p*,*d*,*q*)(*P*,*D*,*Q*)*m*, where *m* refers to the number of periods in each season, and the uppercase *P*,*D*,*Q* refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Forecasting Methods | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
| Best | Arima | Arima | Arima | ETS | Arima | Arima | ETS |
| Worst | Snaive | Snaive | Snaive | Snaive | Snaive | Snaive | Snaive |

## Conclusion

* Summarize your analysis of time series value over the time period.
* The plot shows visitors to USA from the January 2002 to December 2016 with a period of one month. *The original dataset has modified to only include the data from the year 2002, as there was an erratic dip in the number of visitors to USA considering the 9/11 terror attacks on USA in the year 2001*
* The data shows an upward trend from 2002 to 2017
* There is also a seasonality factor where the number of visitors are minimum at the start of the year(winters), increases over the spring months to reach the maximum in the month of July(summers), reduces to month of November (fall) and then again increase in the month of December(winters) probably due to holidays & festive season
* It seems that this time series could probably be described using an additive model, as the seasonal fluctuations are roughly constant in size over time and do not seem to depend on the level of the time series, and the random fluctuations also seem to be roughly constant in size over time.
* Based on your analysis and forecast above, do you think the value of the time series will increase, decrease or stay flat over the next year? How about next 2 years?
* There shall be an upwards/increasing trend in the number of visitors coming to USA
* The estimated seasonal factors are given for the months January-December, and will be similar for coming 2 years.
* The largest seasonal factor will be for July and the lowest will be for February indicating that there will be a peak in number of visitors in July and a trough in number of visitors in February each year.
* The time series data value will highest for July, and the lowest for February for the next and coming 2 years.
* There will also be a rise in number of visitors in month of December for the coming years.
* Rank forecasting methods that best forecast for this time series based on historical values.

|  |  |
| --- | --- |
| Forecasting Methods | Rank |
| SNaive | 4 |
| Ets | 2 |
| HoltsWinter | 3 |
| ARIMA | 1 |

## Final Question

* If you were me, what final grade would you give yourself for this class?
* Indicate the reasons why you gave yourself this grade?

Well, this is the most difficult question.

* I wish I could be you with the knowledge and excellence you have in the forecasting domain. But still if I were you, I would have deducted some marks on late submission, but I would like to highlight the fact on why I did late submission.
* So the time I started actually understanding the true interpretation of the graphs it was already 11 pm. I did write the explanations which were fairly right but still, it was not up to the mark/quality of what you taught in class. So I decided to stay up overnight and go through the internet to see numerous sites and videos to correct understanding and then wrote the entire explanations again.
* This did cost me to miss the deadline but surely increased the quality of work (which I believe to be paramount in any task). My entire motive was to learn and understand and then submit exam with quality.
* I could have chosen the easier path to copy and paste from internet/classmates but that is not integrity.
* So considering the reason I stated above, my class participation, dedication with assignments and integrity, I would say that I should rank amongst the top students in the class.