Business Forecasting

NAME- Prabhat Johl RUID- 182002401

Import Data

Please do the following steps once the csv file is on your desktop.

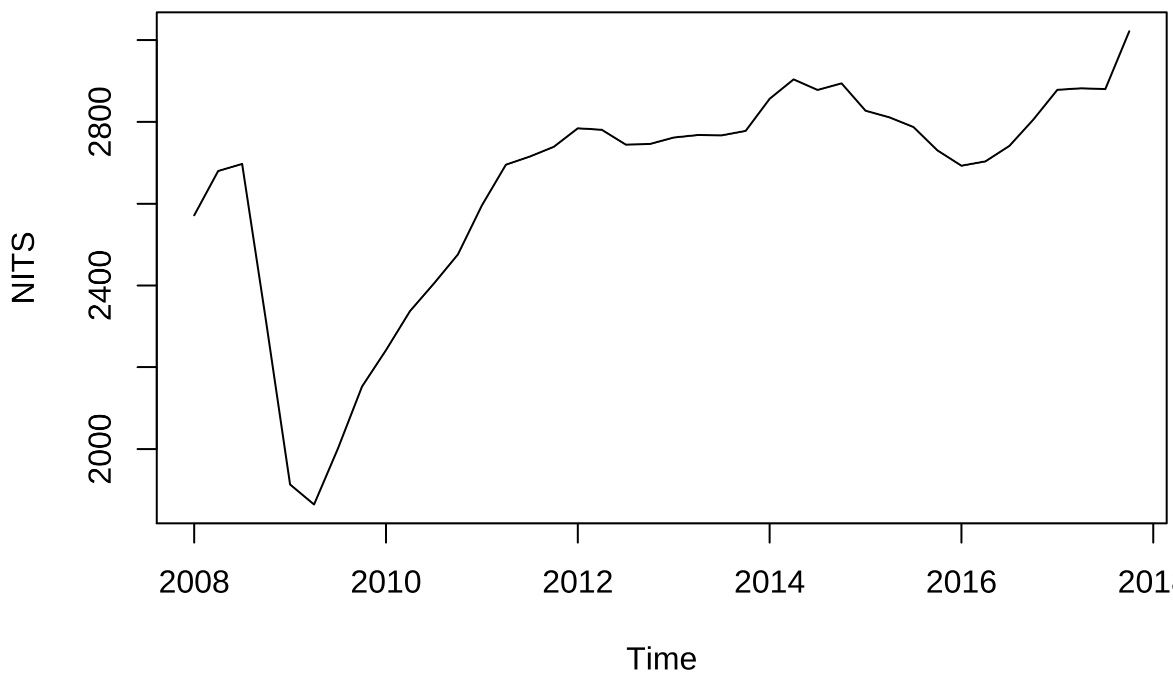
```
library(readr)
    Data_Spring_2018_NetImports <- read_csv("/Users/prabhatjohl/Documents/BF/Data_Spring_2018_NetImports
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Quarter = col_character(),
##   Imports = col_double()
## )
```

```
netImport <- Data_Spring_2018_NetImports
NITS <- ts(netImport$Imports,start=c(2008,01),frequency = 4)
```

Plot and Inference • Show a time series plot.

```
plot(NITS)
```



• Please summaries your observations of the times series plot From the first look, we can observe the absence of the seeasonality and hence, trend is driving the time series however we can see asubstancial drop at year 2009 which implies of some global cause. The import stated increasing substancially from year 2018 and keep continuing growing but in year 2016 we can see a small drop once again.

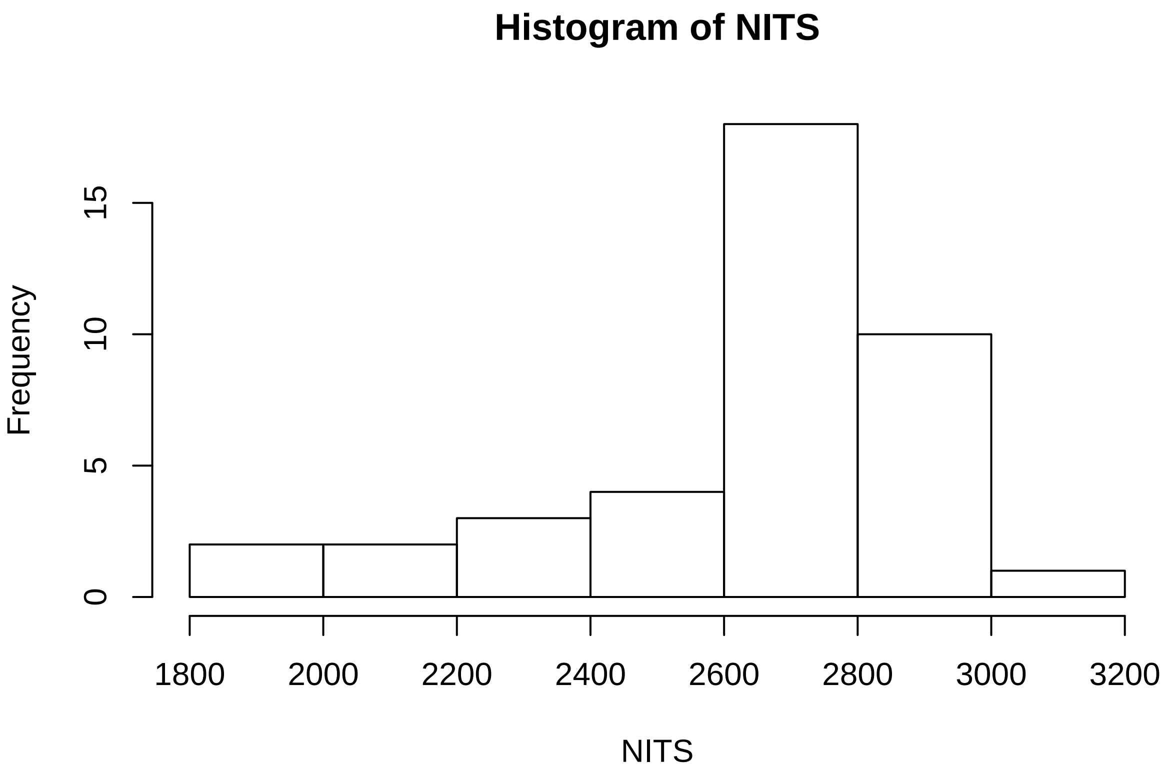
Central Tendency • What are the min, max, mean, median, 1st and 3rd Quartile values of the times series?

```
summary(NITS)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

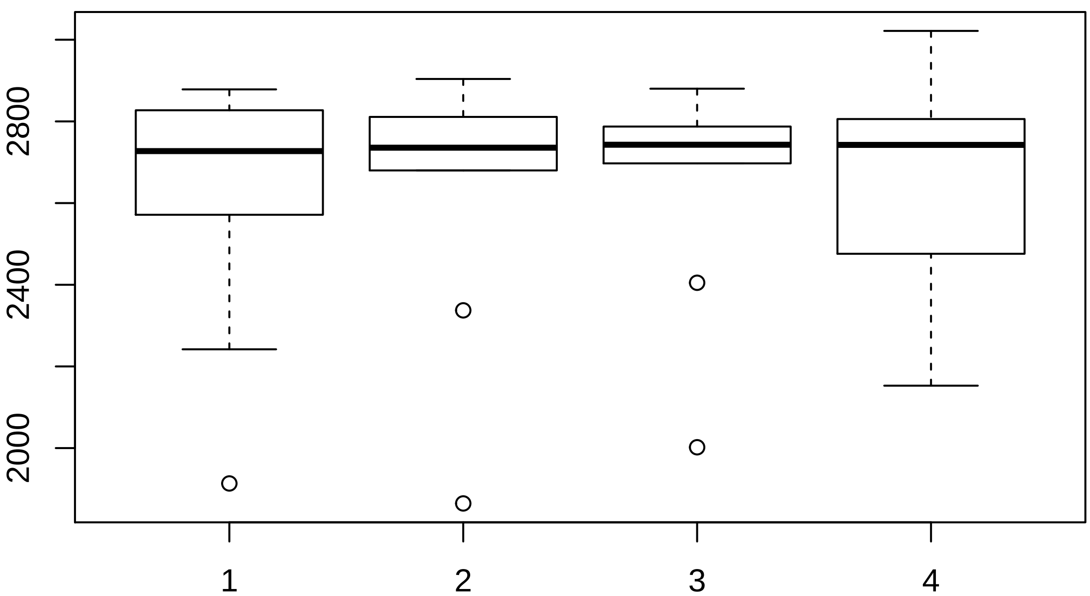
```
## 1864 2590 2743 2646 2807 3022
```

```r
hist(NITS)
```



- Show the box plot.

```r
boxplot(NITS~cycle(NITS))
```



- Can you summarize your observation about the time series from the summary stats and box plot? ->From the Histogram we can see that it is right skewed.

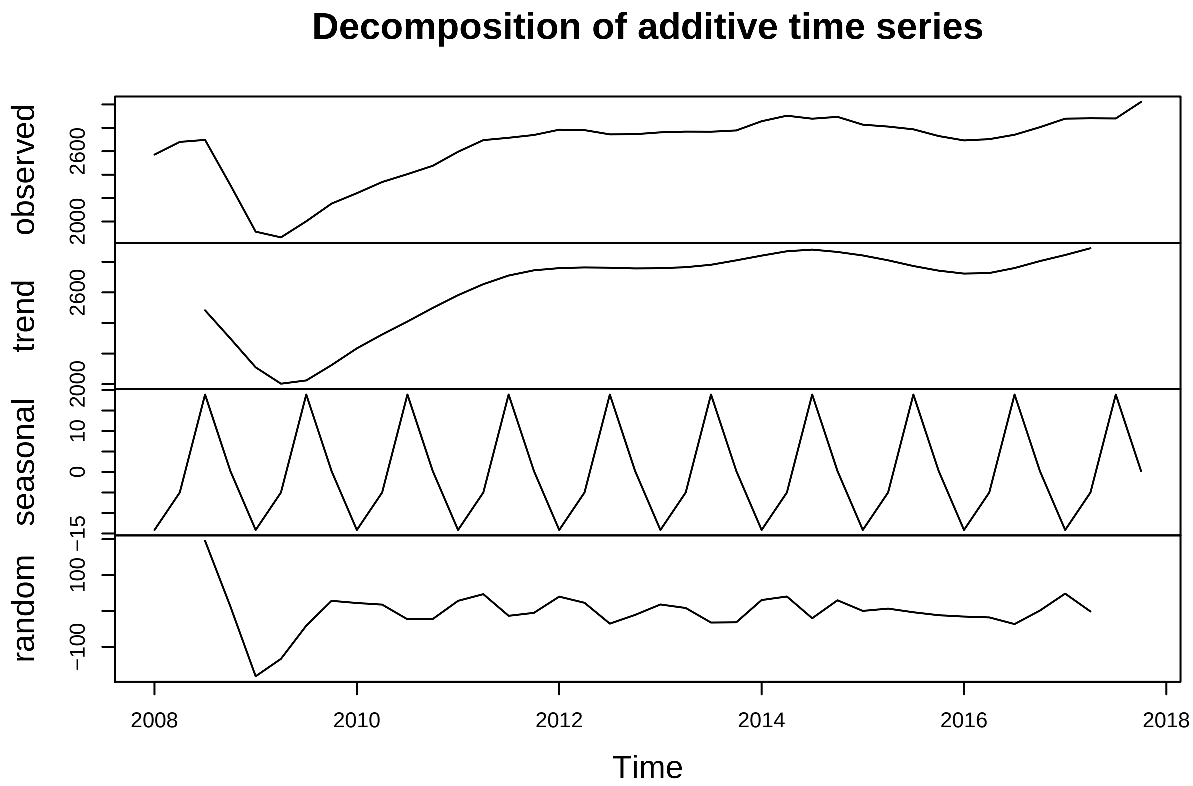
->From the Summary, we got information on minimum, maximum and mean of the US net import in given time frame.

->From the boxplot,It is interesting to see mean of the import stayed somewhat same among all the quaters. The outlines which are seem for the 1st 2nd and the 3rd quaters due to the drop in year 2009.

Decomposition

- Plot the decomposition of the time series.

```
NITS_decom <- decompose(NITS)

plot(NITS_decom)
```



- Is the times series seasonal?

The seasonal component is there in the series but it is not driving the time series as it is contributing 25 units of the variation in the searies whereas trend contributes 900 units. hence seasonality is not significantly contributing in the series.

- Is the decomposition additive or multiplicative?

```
NITS_decom$type
```

```
## [1] "additive"
```

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

```
NITS_decom$seasonal
```

```
##             Qtr1        Qtr2        Qtr3        Qtr4
## 2008 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2009 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2010 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2011 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2012 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2013 -14.1503472  -4.9961806  18.9038194   0.2427083
```

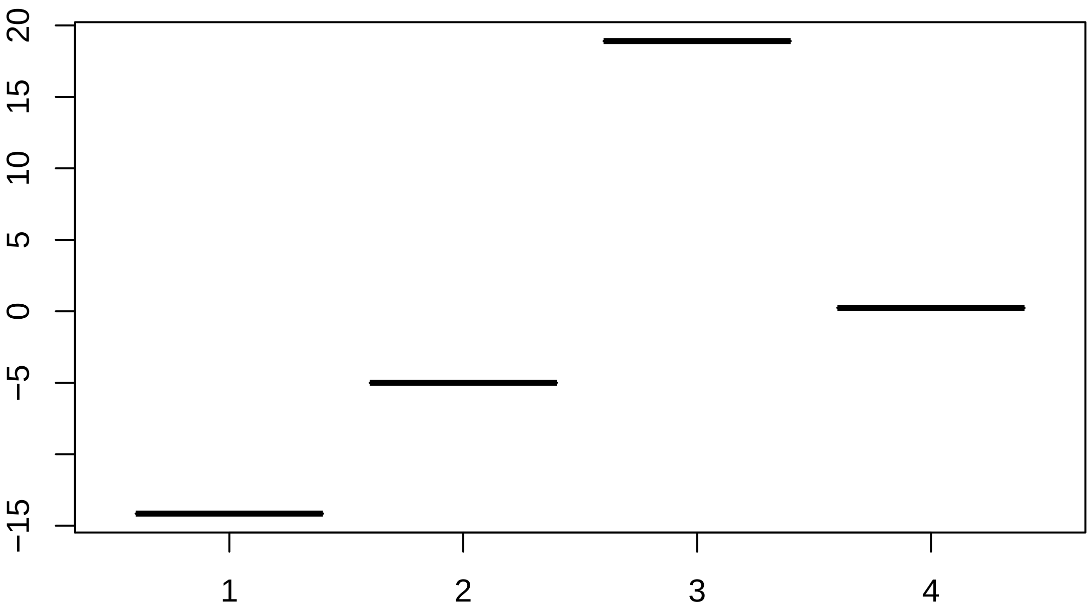
```
## 2014 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2015 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2016 -14.1503472  -4.9961806  18.9038194   0.2427083
## 2017 -14.1503472  -4.9961806  18.9038194   0.2427083
```

```r
NITS_decom$figure
```

```
## [1] -14.1503472  -4.9961806  18.9038194   0.2427083
```

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

```r
boxplot(NITS_decom$seasonal~cycle(NITS_decom$seasonal))
```

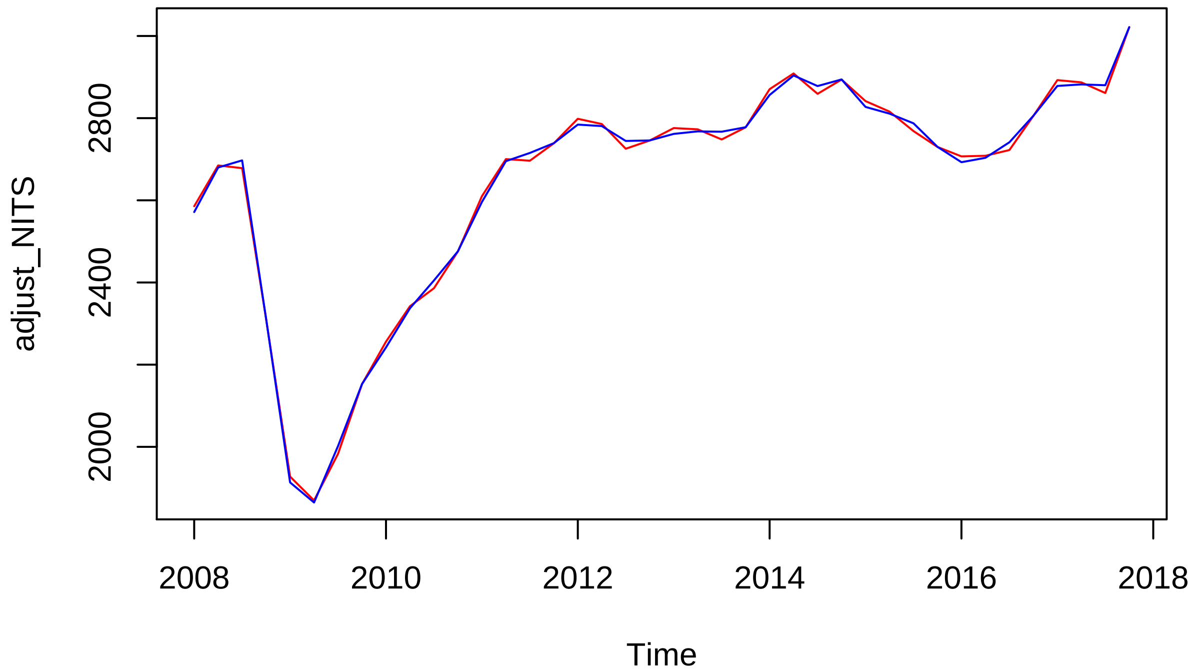


- Can you think of the reason behind the value being high in those Quater and low in those months?

The top 3 imports of states are Electrical machinery equipments, Vehicles and organic fossile fuels which summed to more than 45 percent of the total. It is due to the summer time in states hence building new infrastructure is more likely and preparing for the impending winter season which starts from the 4th quater of the year by importing more fuel as energy demand increases and Another possible reason is New finacial year, which results in new funds for the purchase. 1st quarter remains low due to the peak winters.

- 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?

```r
adjust_NITS = NITS - NITS_decom$seasonal
plot(adjust_NITS, col='red')
lines(NITS, col='blue')
```



Inference: No, There is no significant fluctuation.Hence seasonality is not effecting the time series.

Naïve Method • Output

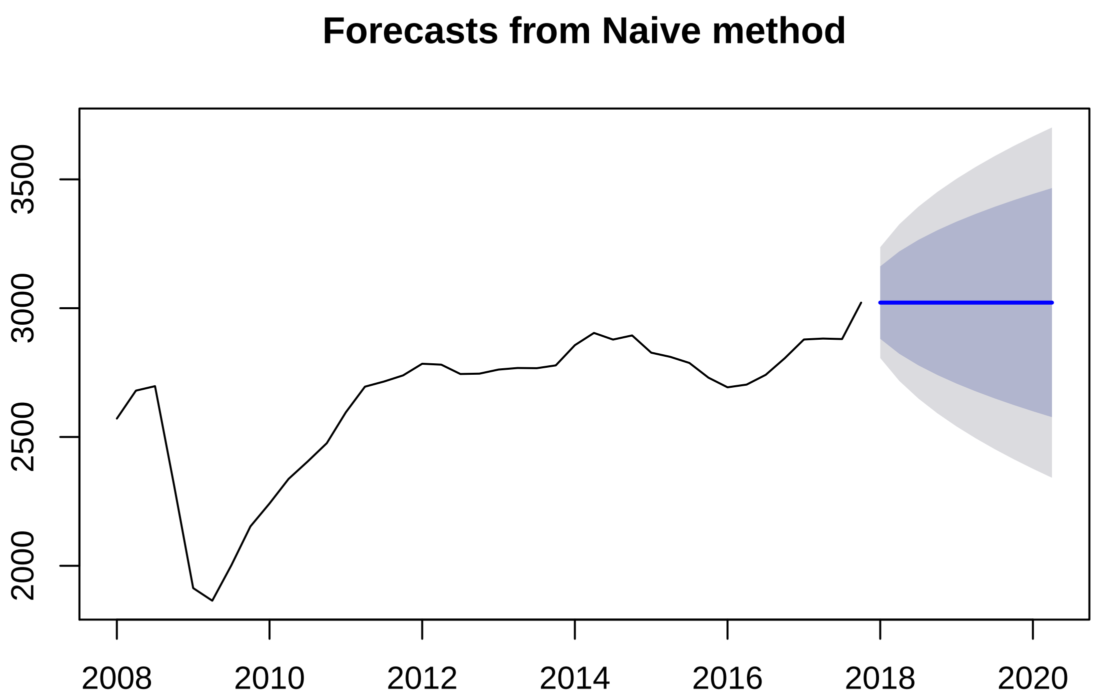
```r
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.4.4

## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/New_York'
```

```r
naive_forecast <- naive(NITS)
naive_forecast
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1         3021.6 2881.006 3162.194 2806.580 3236.620
## 2018 Q2         3021.6 2822.770 3220.430 2717.516 3325.684
## 2018 Q3         3021.6 2778.084 3265.116 2649.174 3394.026
## 2018 Q4         3021.6 2740.412 3302.788 2591.560 3451.640
## 2019 Q1         3021.6 2707.222 3335.978 2540.800 3502.400
## 2019 Q2         3021.6 2677.216 3365.984 2494.910 3548.290
## 2019 Q3         3021.6 2649.623 3393.577 2452.710 3590.490
## 2019 Q4         3021.6 2623.940 3419.260 2413.431 3629.769
## 2020 Q1         3021.6 2599.818 3443.382 2376.540 3666.660
## 2020 Q2         3021.6 2577.002 3466.198 2341.647 3701.553
```

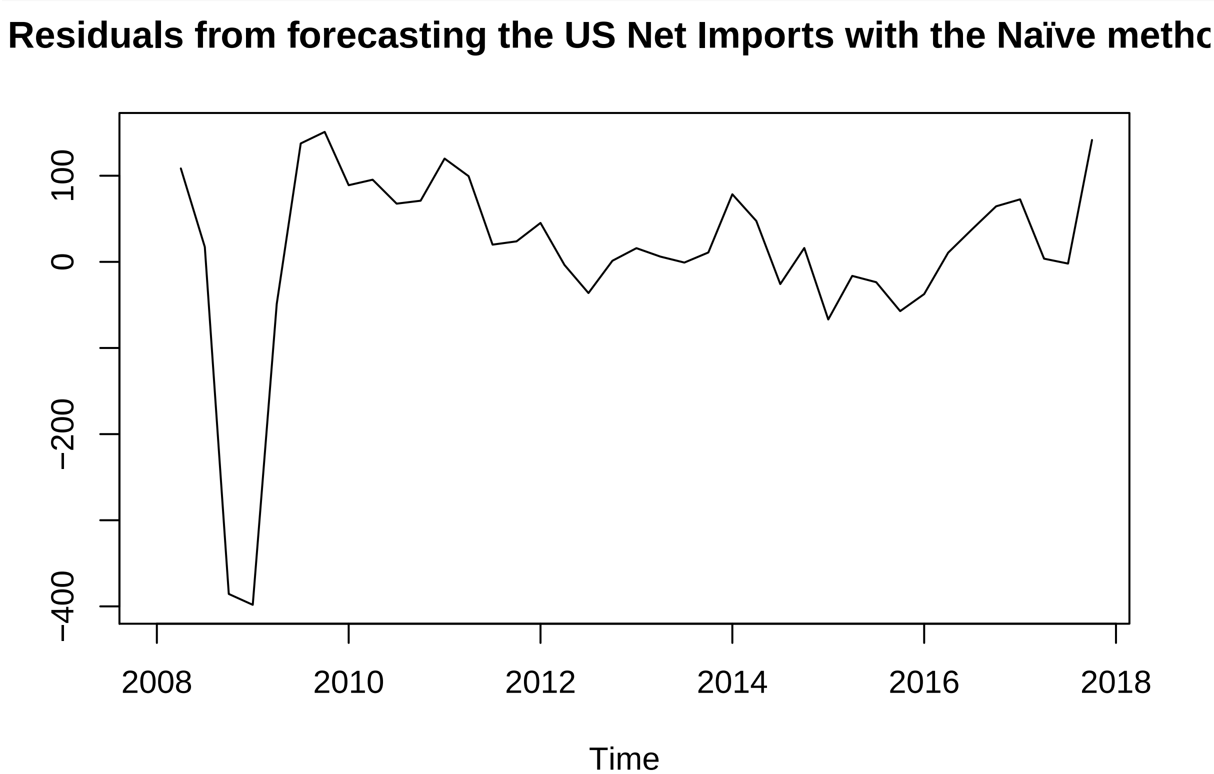
```r
plot(naive_forecast)
```



- Perform Residual Analysis for this technique.

o Do a plot of residuals. What does the plot indicate?

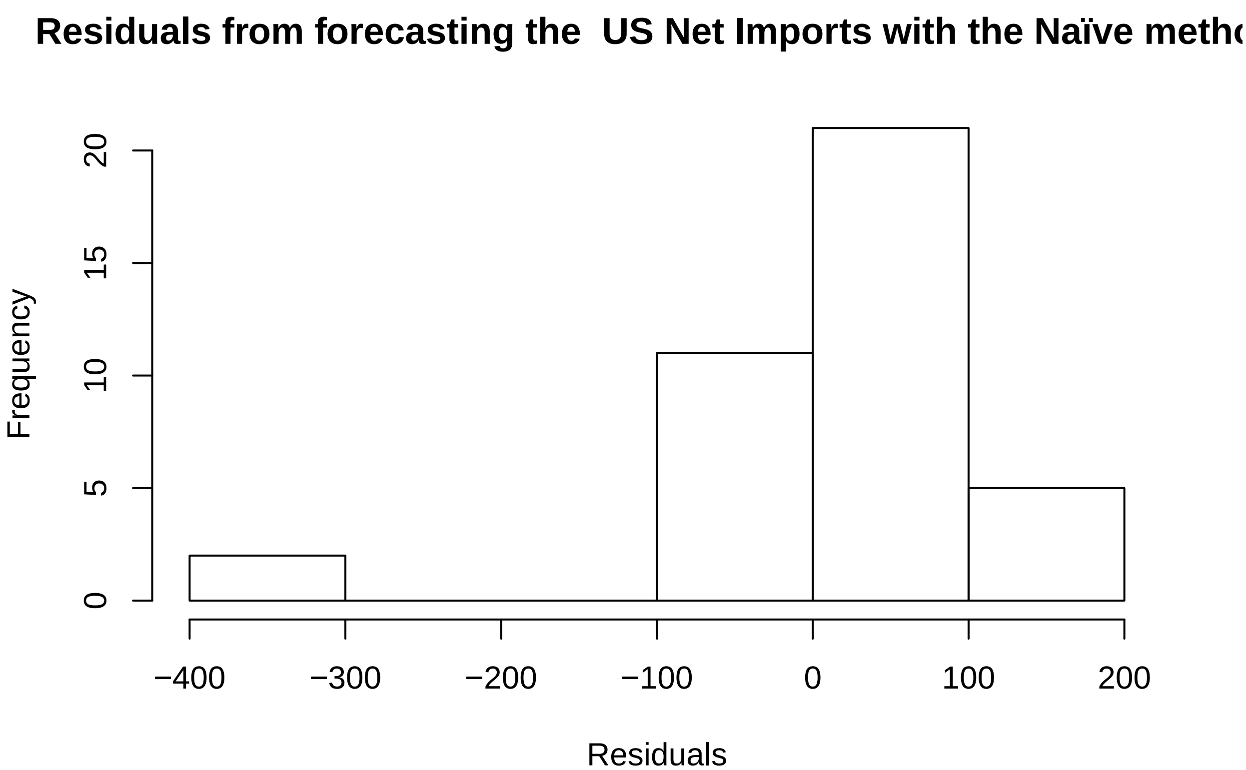
```r
plot(naive_forecast$residuals,main="Residuals from forecasting the US Net Imports with the Naïve method
     ylab="", xlab="Time")
```



Inference: ->From the time plot, we can see that the mean of variation of the residuals stayed somewhat constant through out the time frame. ->there is a big drop in year 2009 thats due to the slow world market and global recession.

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

```
hist(naive_forecast$residuals,main='Residuals from forecasting the  US Net Imports with the Naïve metho
```



->The residuals are not symmetric around the 0. which is due to massive drop in year 2009. Hence. Residuals are not uniformly distributed which shows biasness. if we ignore the activities of year 2009, I believe we could have a uniformly distributed residuals.

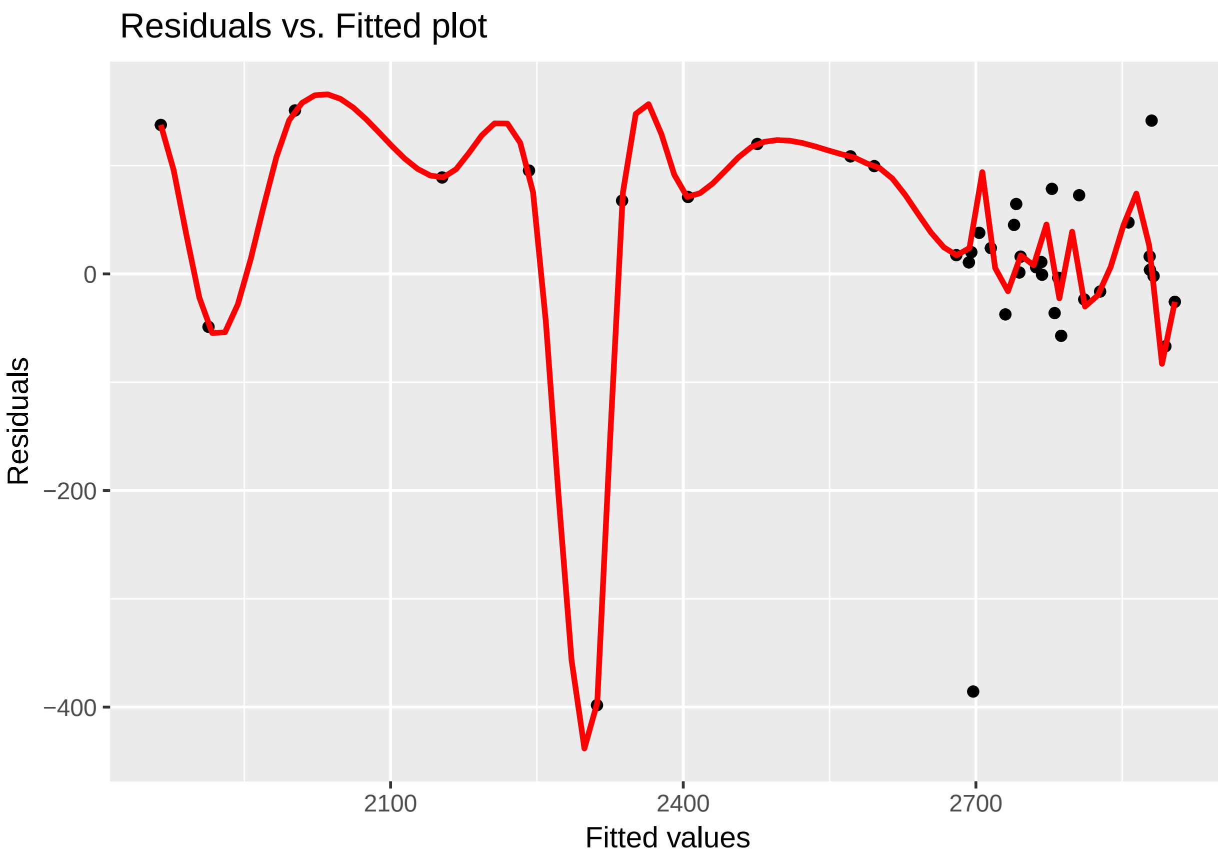
o Do a plot of fitted values vs. residuals. What does the plot indicate?

```
library(ggplot2)
qplot(y = naive_forecast$residuals, x = naive_forecast$fitted,
      ylab = "Residuals", xlab = "Fitted values",
      main = " Residuals vs. Fitted plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1859.3

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 142.7

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 728.84
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



->In the above plot of Residuals VS Fitted values shows residuals has pattern and they are not randomly distributed among themself until 2700 which is due to the 2009 year recession but afterwards it looks uniformly distributed and points are dense. ->Previous fitted values before 2700 are very small and we can see the value around 2300 dragging it down.

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

```
library(ggplot2)
qplot(y = naive_forecast$residuals, x = NITS,
      ylab = "Residuals", xlab = "Actual values",
      main = " Residuals vs. Actual plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```

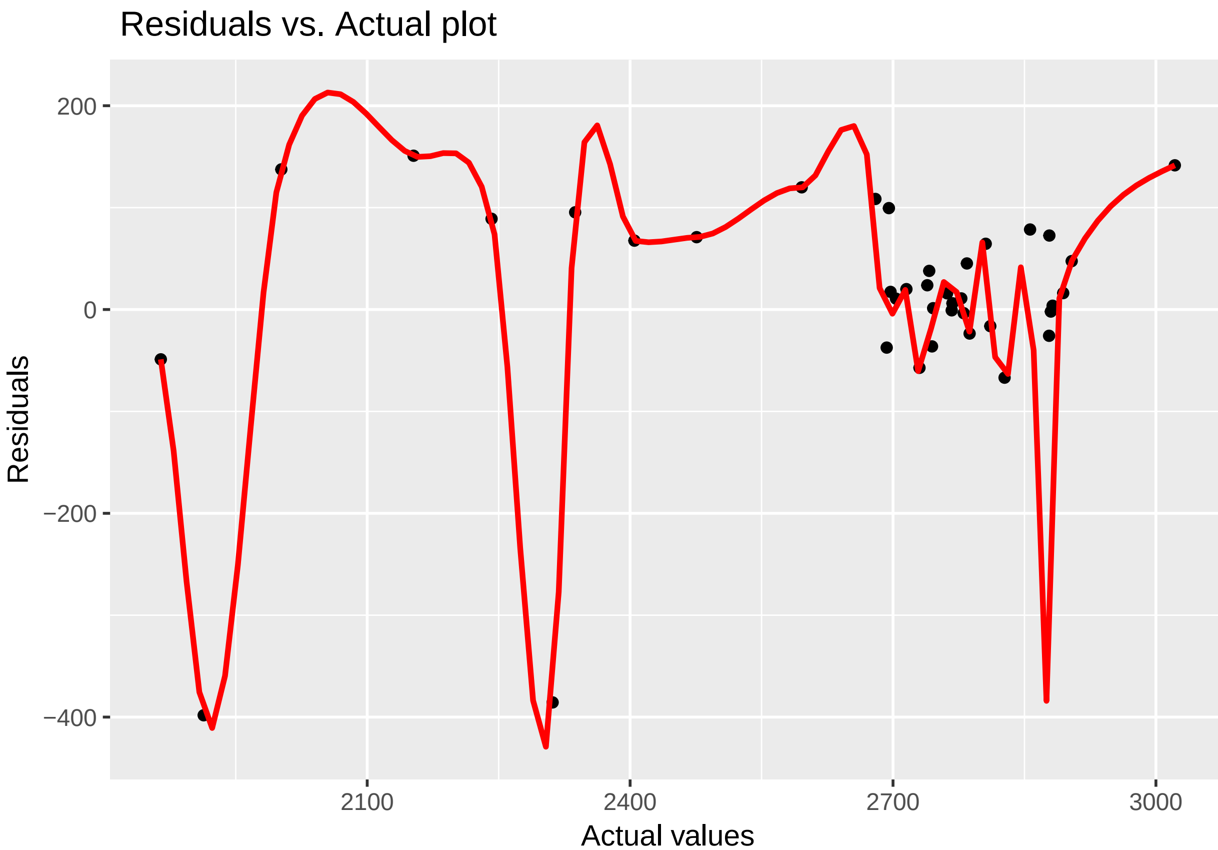
```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1858.7
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 143.29
```

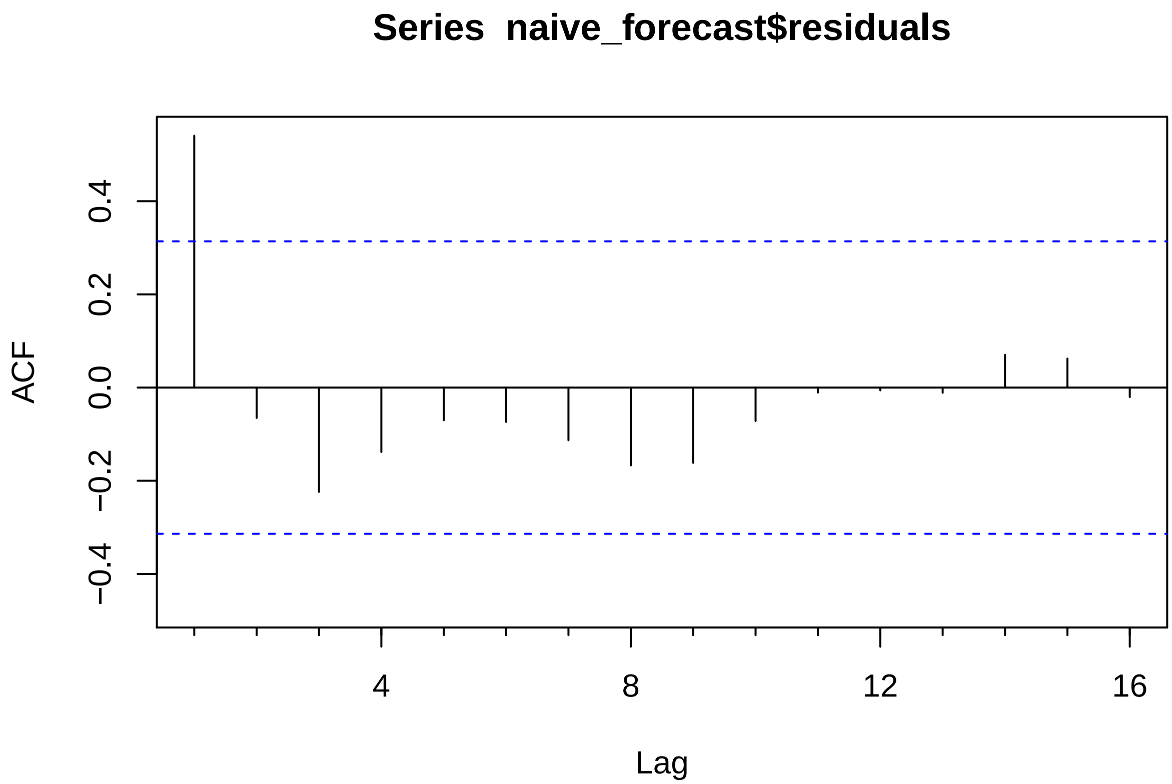
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 17738

## Warning: Removed 1 rows containing missing values (geom_point).
```



->In the above plot of Residuals VS Actual values shows residuals has pattern and they are not randomly distributed among themself.Hence the model does not fits well.

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

```
Acf(naive_forecast$residuals)
```



>Spikes shows the values of Autocorrelation with each lags. We can observe that amplitude of all the spikes from 2nd lag are in the blue segment which implies they are insignificant however it is significant for the lag at 1.

- Print the 5 measures of accuracy for this forecasting technique

```
naive_accuracy <- accuracy(naive_forecast)
naive_accuracy
```

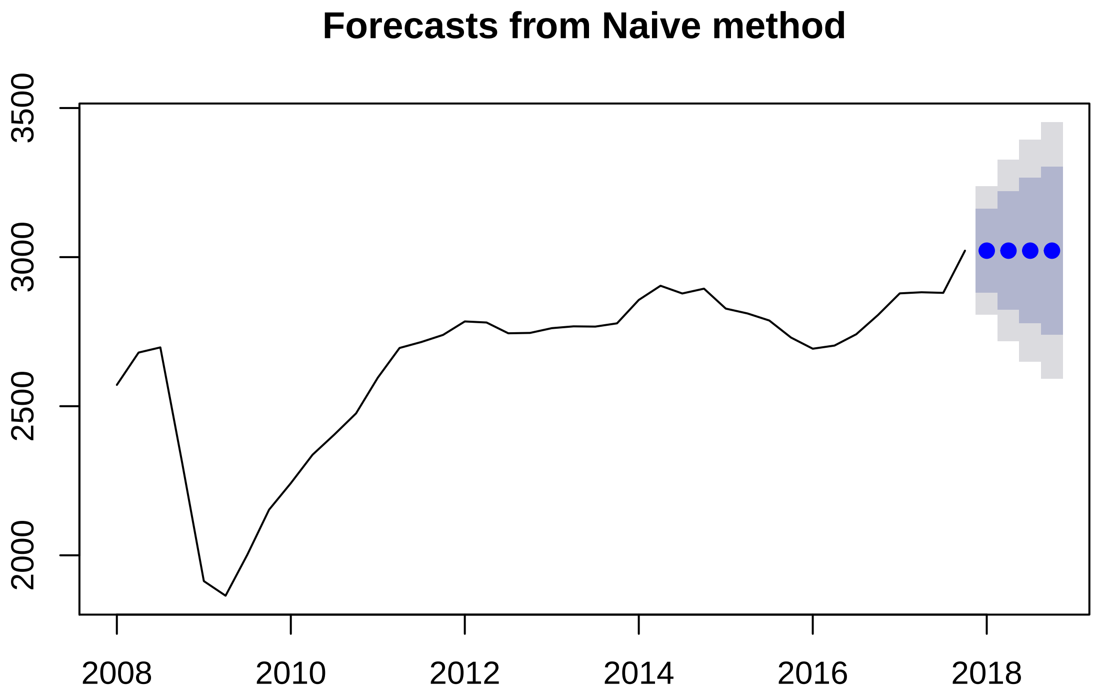
```
##                    ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 11.54359 109.7054 68.08205 0.2970542 2.852688 0.3286562
##                   ACF1
## Training set 0.5405785
```

- Forecast o Time series value for next year. Show table and plot

```
naive_forecast <- naive(NITS,4)
naive_forecast
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1         3021.6 2881.006 3162.194 2806.580 3236.620
## 2018 Q2         3021.6 2822.770 3220.430 2717.516 3325.684
## 2018 Q3         3021.6 2778.084 3265.116 2649.174 3394.026
## 2018 Q4         3021.6 2740.412 3302.788 2591.560 3451.640
```

```
plot(naive_forecast)
```



- Summarize this forecasting technique

```
summary(naive_forecast)
```

```
##
## Forecast method: Naive method
##
## Model Information:
## Call: naive(y = NITS, h = 4)
##
## Residual sd: 109.7062
##
## Error measures:
##                    ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 11.54359 109.7054 68.08205 0.2970542 2.852688 0.3286562
##                   ACF1
## Training set 0.5405785
##
## Forecasts:
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1         3021.6 2881.006 3162.194 2806.580 3236.620
## 2018 Q2         3021.6 2822.770 3220.430 2717.516 3325.684
## 2018 Q3         3021.6 2778.084 3265.116 2649.174 3394.026
## 2018 Q4         3021.6 2740.412 3302.788 2591.560 3451.640
```

o How good is the accuracy?

->Error measures are not that high but it could perform well. We can use other model to get better predictions.
o What does it predict the value of time series will be in one year?

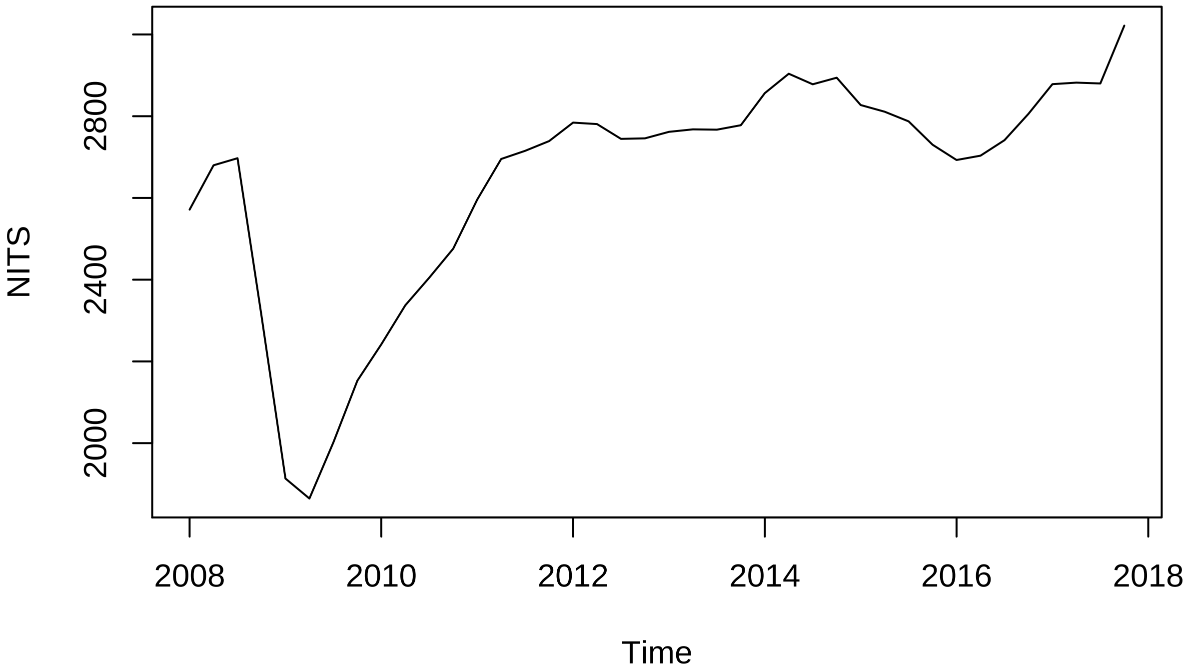
->It has Point Forecasted 3021.6 for the whole year but highs and lows for the 80 and 95 percent increases while going ahead in time, which can be observed in above table.

o Other observation

->I believe, WIth point Forecast for prediction for over an year. It would not be a great idea to predict far in the future.

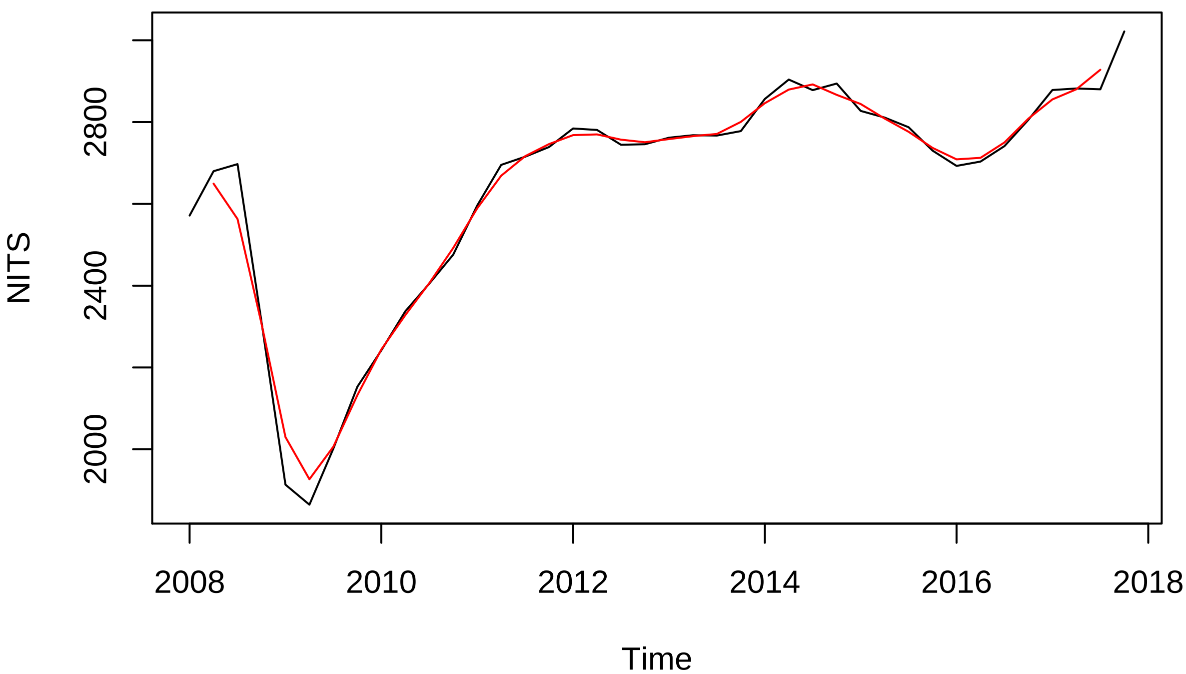
Simple Moving Averages • Plot the graph for time series.

```r
plot(NITS)
```



- Show the Simple Moving average of order 3 on the plot above in Red

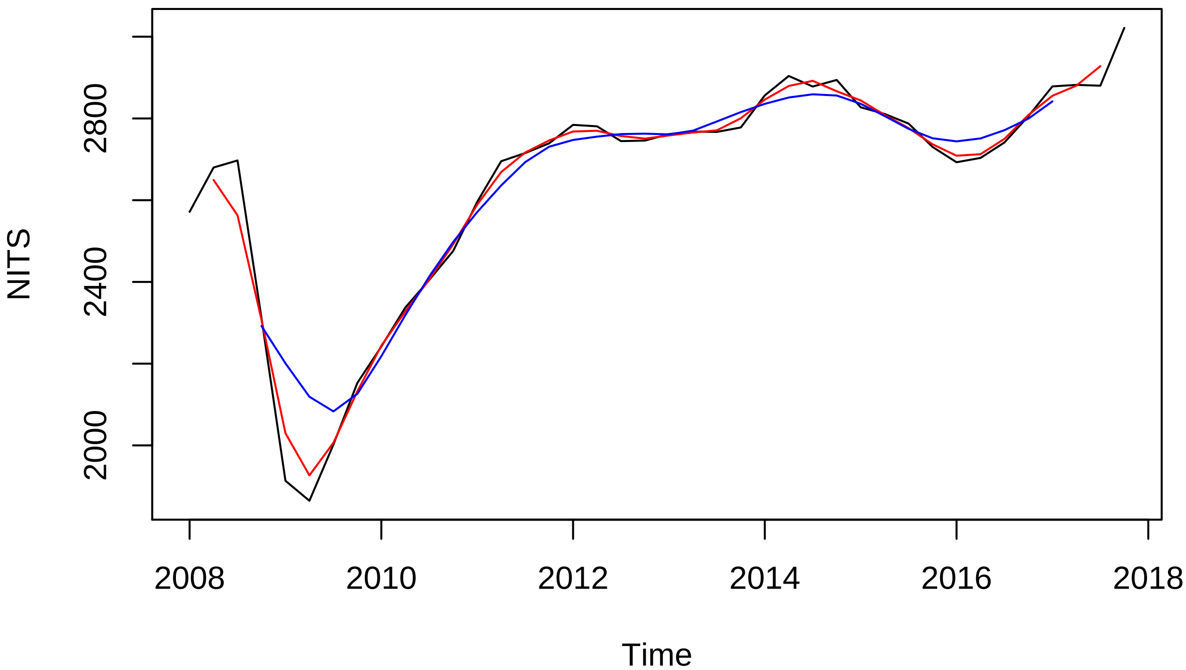
```r
MA3_forecast <- ma(NITS,order=3)
plot(NITS)
lines(MA3_forecast, col='Red')
```



•

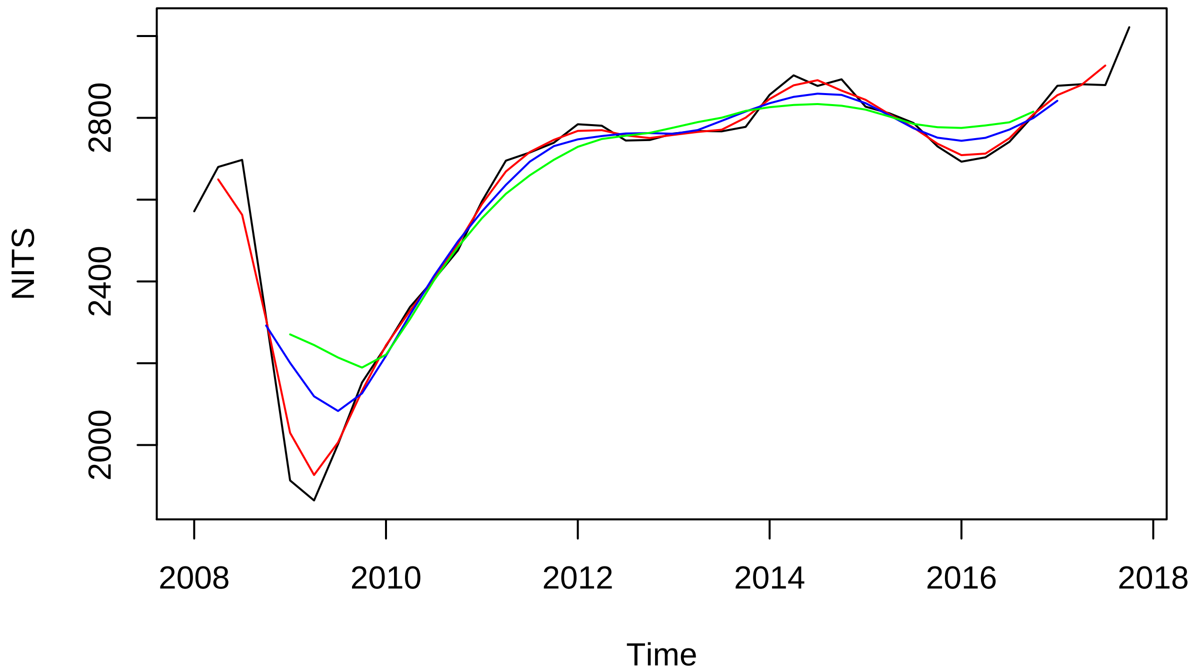
Show the Simple Moving average of order 6 on the plot above in Blue

```
MA6_forecast <- ma(NITS,order=6)
plot(NITS)
lines(MA3_forecast, col='Red')
lines(MA6_forecast, col='Blue')
```



- Show the Simple Moving average of order 9 on the plot above in Green

```
MA9_forecast <- ma(NITS,order=9)
plot(NITS)
lines(MA3_forecast, col='Red')
lines(MA6_forecast, col='Blue')
lines(MA9_forecast, col='Green')
```



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

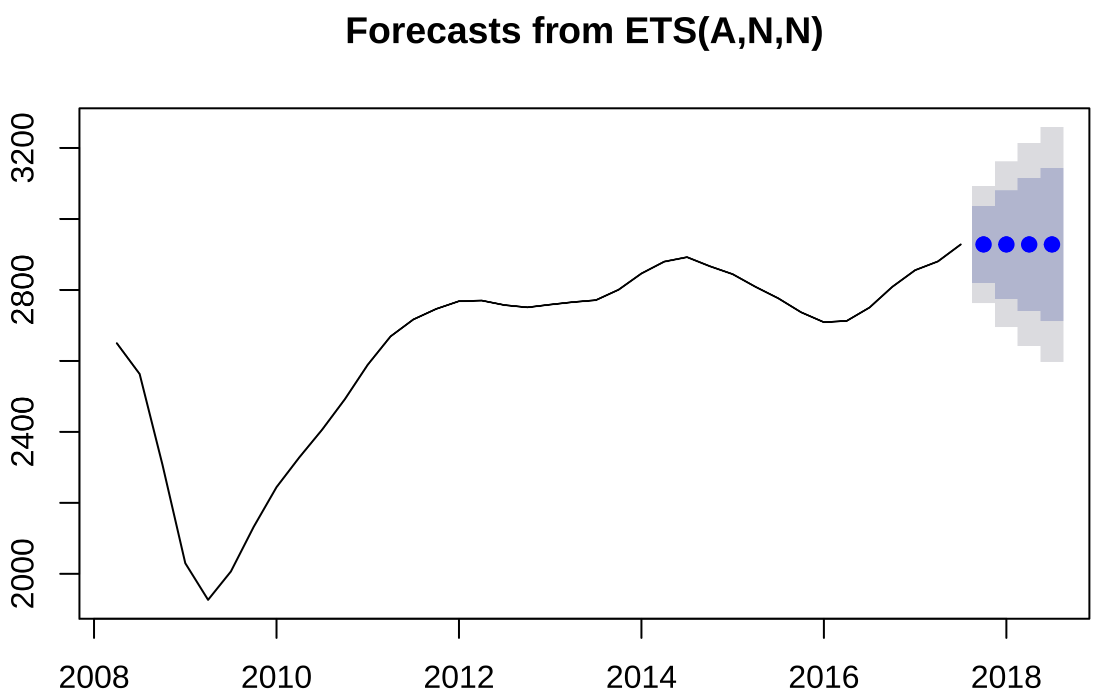
```
ma_forcast= forecast(object=MA3_forecast, h= 4 )
```

```
## Warning in ets(object, lambda = lambda, biasadj = biasadj,
## allow.multiplicative.trend = allow.multiplicative.trend, : Missing values
## encountered. Using longest contiguous portion of time series
```

```
ma_forcast
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2017 Q4       2927.929 2819.974 3035.883 2762.826 3093.031
## 2018 Q1       2927.929 2775.265 3080.592 2694.450 3161.407
## 2018 Q2       2927.929 2740.958 3114.899 2641.982 3213.875
## 2018 Q3       2927.929 2712.036 3143.821 2597.749 3258.108
```

```
plot(ma_forcast)
```



->I choose MA of order 3 for the forecast because it overlaps best in all of the orders used for the prediction here. Hence, It makes the better predictions as compare to other orders.

- What are your observations of the plot as the moving average order goes up? -> As the order goes up in moving average. It starts approching towards the mean of whole forecast. which can be observed in the above plot of order 9 which is much near to the mean of time series whereas of order 3 is overlapping best among of all.

Smoothing • Perform a smoothing forecast for next 12 months for the time series.

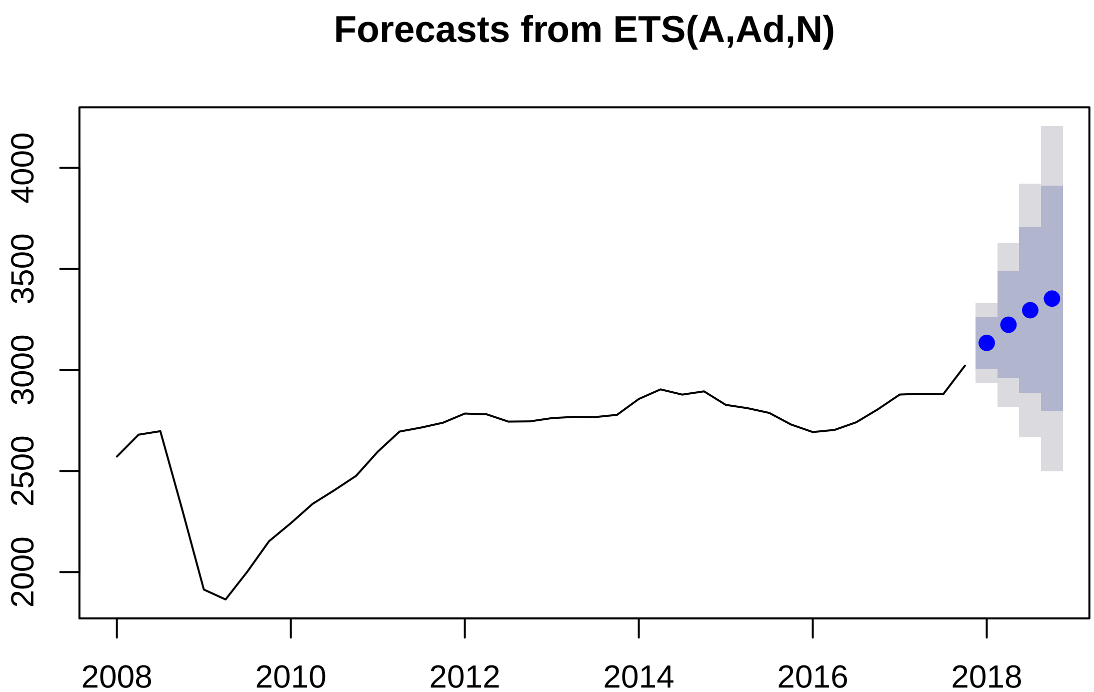
```
ets<-ets(NITS)
ets
```

```
## ETS(A,Ad,N)
##
## Call:
##  ets(y = NITS)
##
##   Smoothing parameters:
##     alpha = 0.9999
##     beta  = 0.9925
##     phi   = 0.8
##
##   Initial states:
##     l = 2576.5503
##     b = 7.0378
##
##   sigma:  100.5086
##
##      AIC     AICc      BIC
## 523.0334 525.5788 533.1666
```

```
forecast_ets <- forecast.ets(ets, h=4)
forecast_ets
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3133.926 3005.119 3262.733 2936.933 3330.919
## 2018 Q2       3223.799 2959.257 3488.341 2819.216 3628.382
## 2018 Q3       3295.698 2885.967 3705.429 2669.069 3922.327
## 2018 Q4       3353.218 2795.526 3910.910 2500.302 4206.134
```

```r
plot(forecast_ets)
```



o What is the value of alpha? What does that value signify?

->alpha = 0.9999

The value of alpha is extremly high and close to 1, which means the model is giving weight to the last lag values than to the past value.The value of alpha lies between 0 to 1.

o What is the value of initial state?

```r
ets$initstate
```

```
##           l           b
## 2576.550310    7.037826
```

o What is the value of sigma? What does the sigma signify?

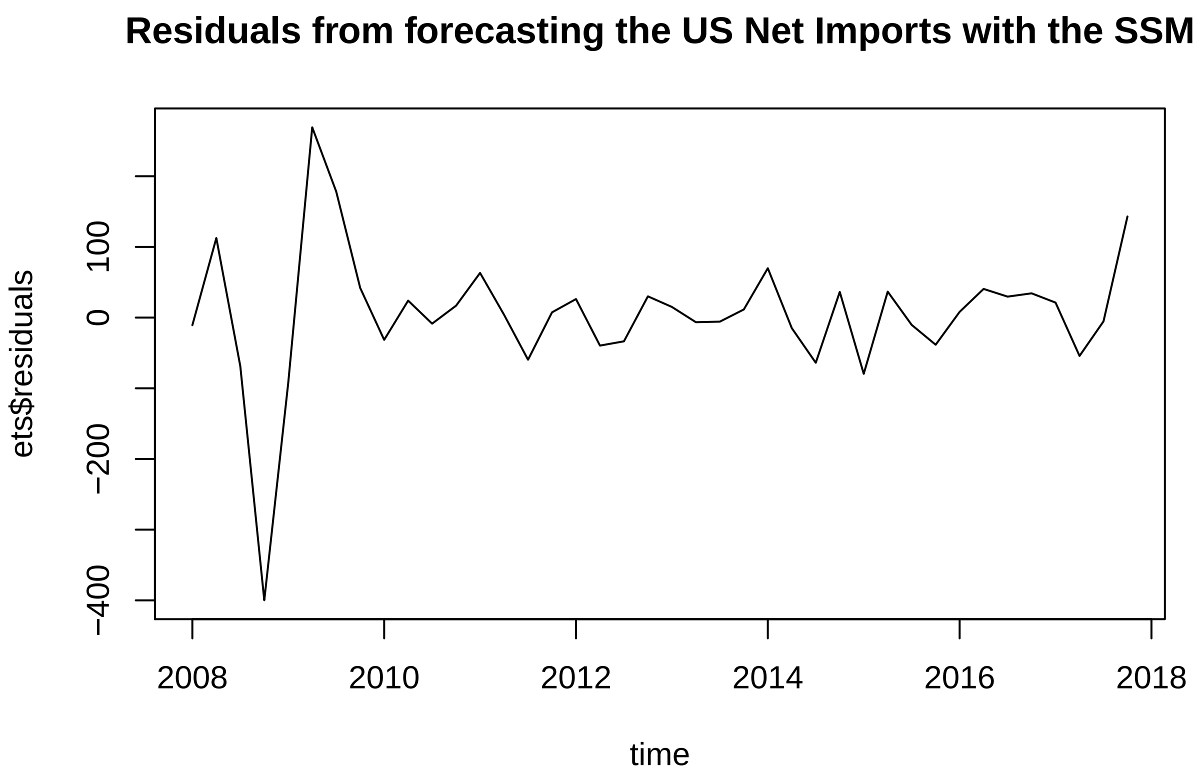
->The value of sigma : 108.8364

->It signifies the standard deviation of the residuals. Standard deviation is the extent of deviation for a group as a whole.

- Perform Residual Analysis for this technique.

o Do a plot of residuals. What does the plot indicate?

```r
plot(ets$residuals,xlab = "time", main="Residuals from forecasting the US Net Imports with the SSM")
```



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

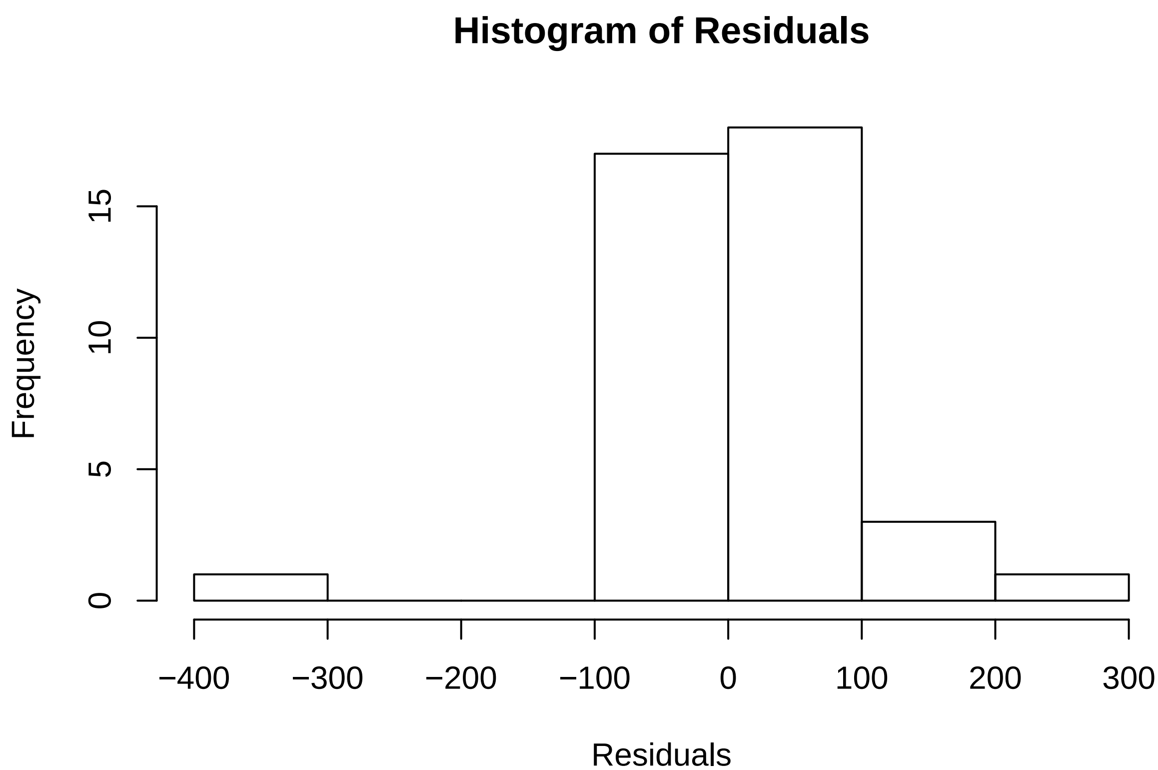
-> By observing the above graph of residuals,we can see that residuals did not stay same accross the historical data.Hence It can not be considered constant over time.

-> There is a dip observed in the residues(2009), which can be due to some special event occurred i.e global slow down which impacted the US net import.

-> By observing the above graph of residuals,we can see that average residuals stayed near to zero with ups and downs after year 2010 i.e. after cover coming global slowdown.

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

```r
hist(ets$residuals, xlab = "Residuals", main="Histogram of Residuals")
```



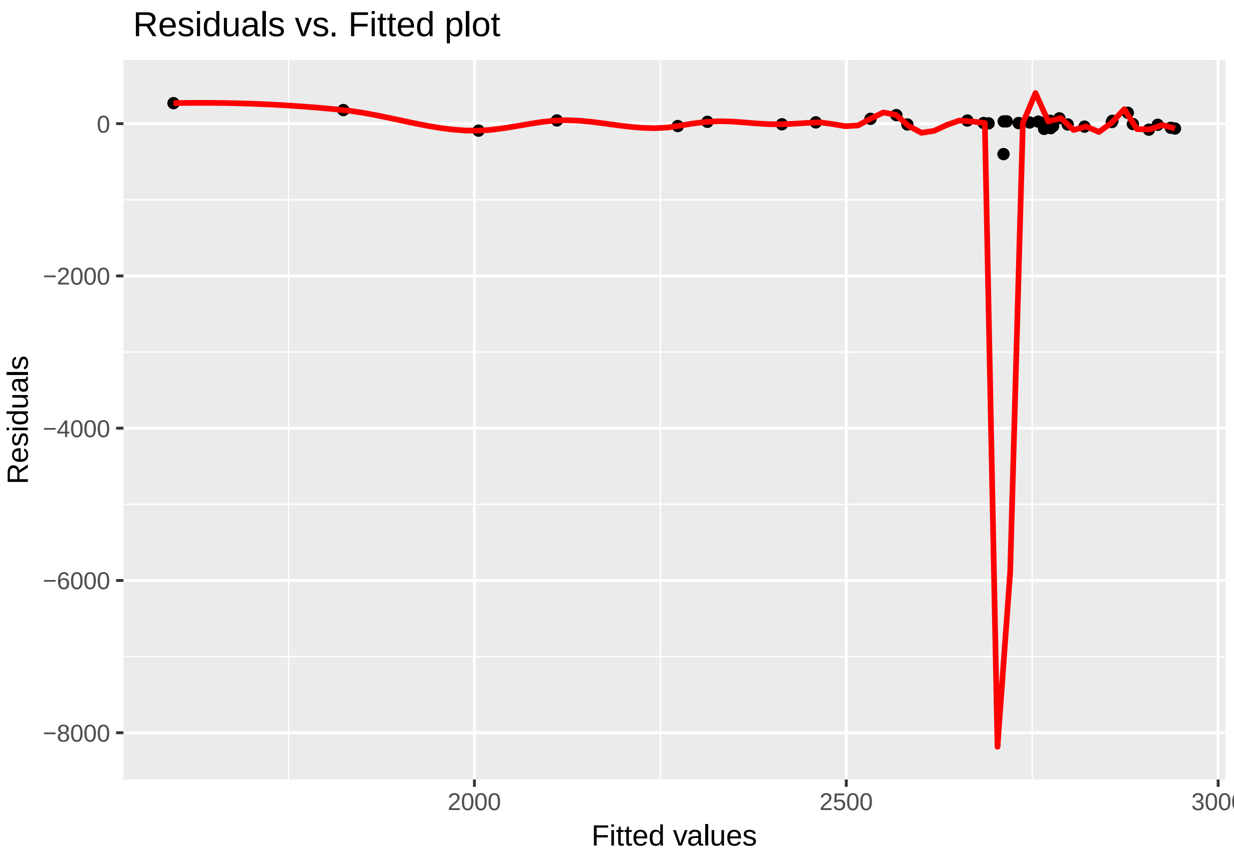
->The histogram of the residuals shows the distribution of the residuals for all observations.The model doesn't fit the data well because of the massive drop in year 2009, the residuals are not random and the histogram is not symmatric about the mean.Hence, it is not normally distributed which implies model doesn't fit well.

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

```
library(ggplot2)
qplot(y = ets$residuals, x = forecast_ets$fitted,
      ylab = "Residuals", xlab = "Fitted values",
      main = " Residuals vs. Fitted plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```

```
## Warning in sqrt(sum.squares/one.delta): NaNs produced
```



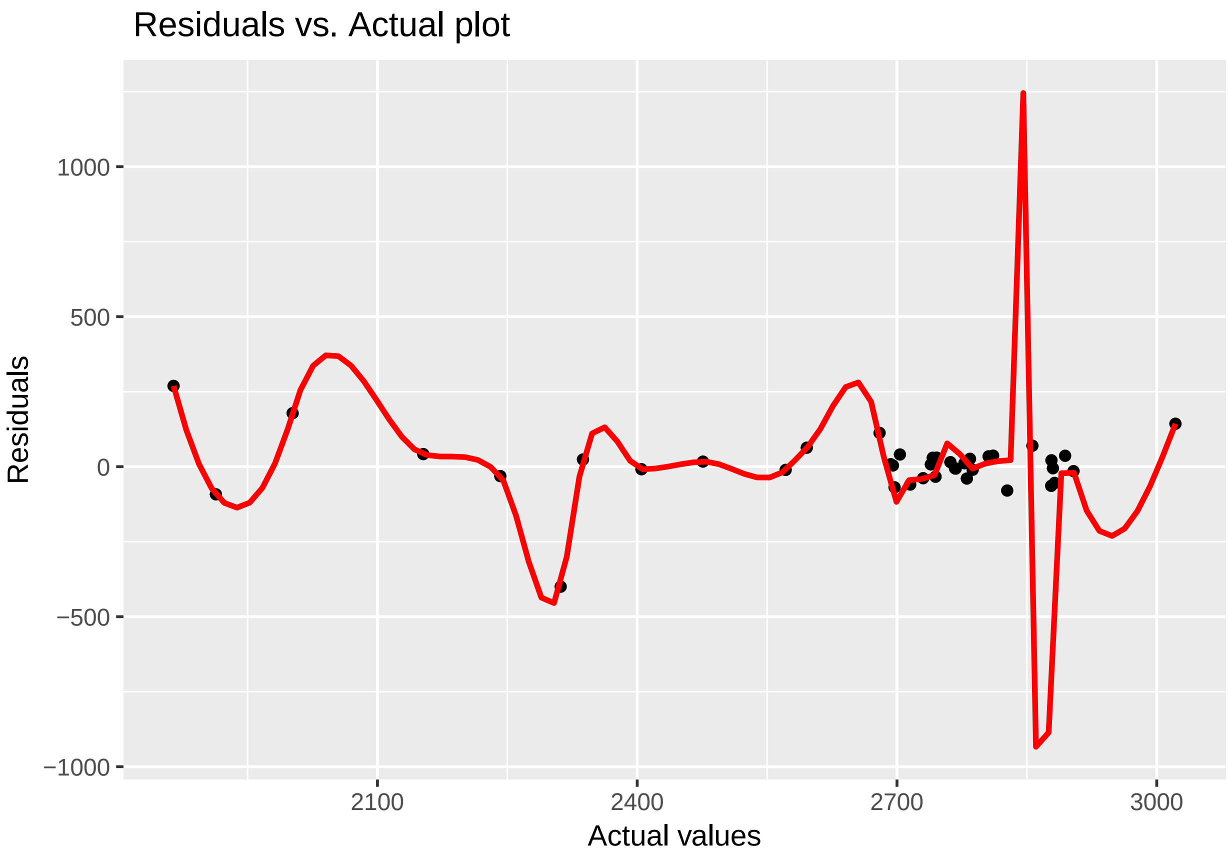
->In the above plot of Residuals VS Fitted values shows residuals has pattern and they are not randomly distributed among themself until 2700 which is due to the 2009 year recession but afterwards it looks uniformly distributed and points are dense.

->Previous fitted values before 2700 are very small and we can see the value around 2300 dragging it down.

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

```
library(ggplot2)
qplot(y = ets$residuals, x = NITS,
      ylab = "Residuals", xlab = "Actual values",
      main = " Residuals vs. Actual plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

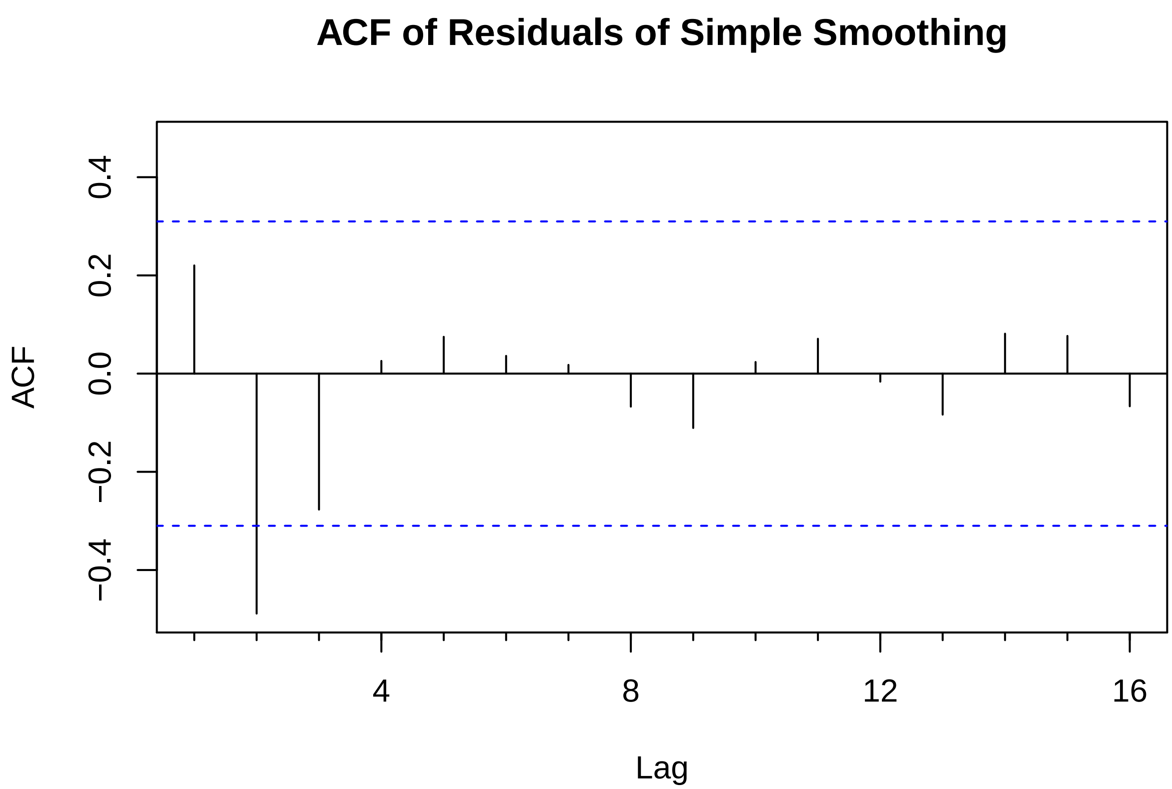
```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 2880.1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in sqrt(sum.squares/one.delta): NaNs produced
```



->In the above plot of Residuals VS Actual values shows residuals has pattern and they are not randomly distributed among themself.Hence the model does not fit well.

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

```
Acf(ets$residuals, main = "ACF of Residuals of Simple Smoothing")
```



->Spikes shows the values of Autocorrelation with each lags. We can observe that amplitude of all the spikes from 2nd lag are in the blue segment which implies they are insignificant however it is significant for the lag at 1.

- Print the 5 measures of accuracy for this forecasting technique

```
accuracy_ets <- accuracy(ets)
accuracy_ets
```

```
##                    ME     RMSE     MAE       MPE     MAPE      MASE
## Training set 4.922542 94.01716 56.0852 0.2528908 2.328826 0.2707432
##                   ACF1
## Training set 0.2203102
```

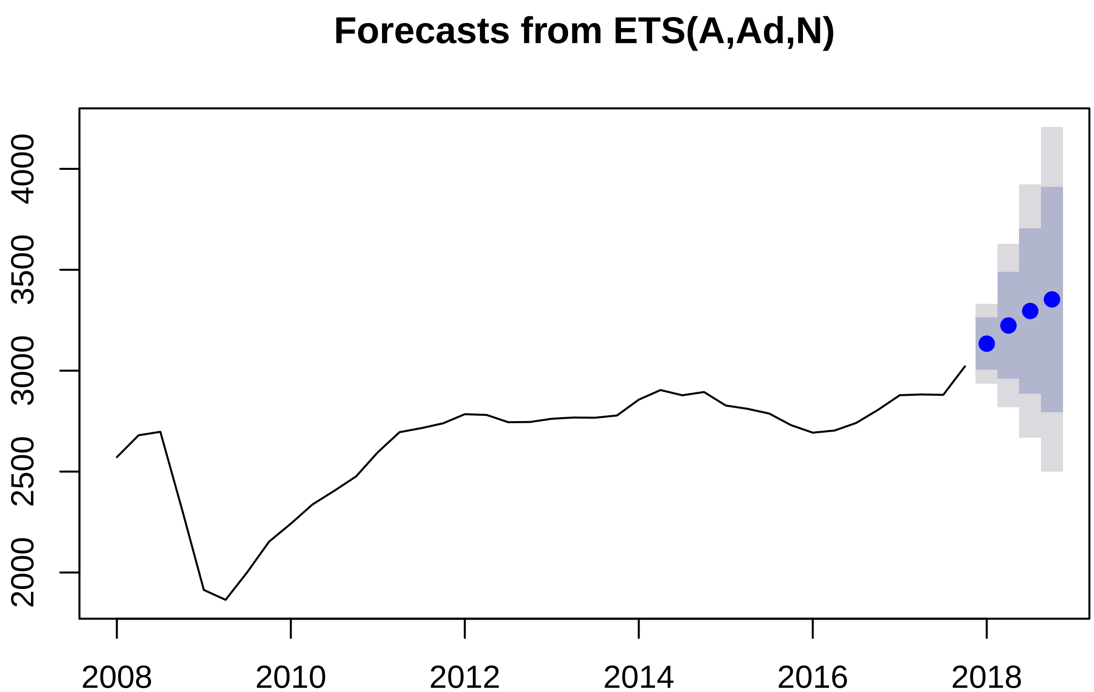
- Forecast

o Time series value for next year. Show table and plot

```
forecast_ets
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3133.926 3005.119 3262.733 2936.933 3330.919
## 2018 Q2       3223.799 2959.257 3488.341 2819.216 3628.382
## 2018 Q3       3295.698 2885.967 3705.429 2669.069 3922.327
## 2018 Q4       3353.218 2795.526 3910.910 2500.302 4206.134
```

```
plot(forecast_ets)
```



- Summarize this forecasting technique

```
summary(forecast_ets)
```

```
##
## Forecast method: ETS(A,Ad,N)
##
## Model Information:
## ETS(A,Ad,N)
##
## Call:
##  ets(y = NITS)
##
##   Smoothing parameters:
##     alpha = 0.9999
##     beta  = 0.9925
##     phi   = 0.8
##
##   Initial states:
##     l = 2576.5503
##     b = 7.0378
##
##   sigma:  100.5086
##
##      AIC     AICc      BIC
## 523.0334 525.5788 533.1666
##
## Error measures:
##                    ME     RMSE     MAE       MPE     MAPE      MASE
## Training set 4.922542 94.01716 56.0852 0.2528908 2.328826 0.2707432
##                   ACF1
## Training set 0.2203102
##
```

```
## Forecasts:
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3133.926 3005.119 3262.733 2936.933 3330.919
## 2018 Q2       3223.799 2959.257 3488.341 2819.216 3628.382
## 2018 Q3       3295.698 2885.967 3705.429 2669.069 3922.327
## 2018 Q4       3353.218 2795.526 3910.910 2500.302 4206.134
```

o How good is the accuracy?

-> Accuracy of the model is not that great and which is somewhat similar to simple smoothing.

o What does it predict the value of time series will be in one year?

->In one year the value of production would be 3021.586 with 95 percent confidence interval, It would be 2594.987 low to 3448.185 high

o Other observation

During the residual analysis, we observed that they are normally distributed.

Holt-Winters • Perform Holt-Winters forecast for next 12 months for the time series.

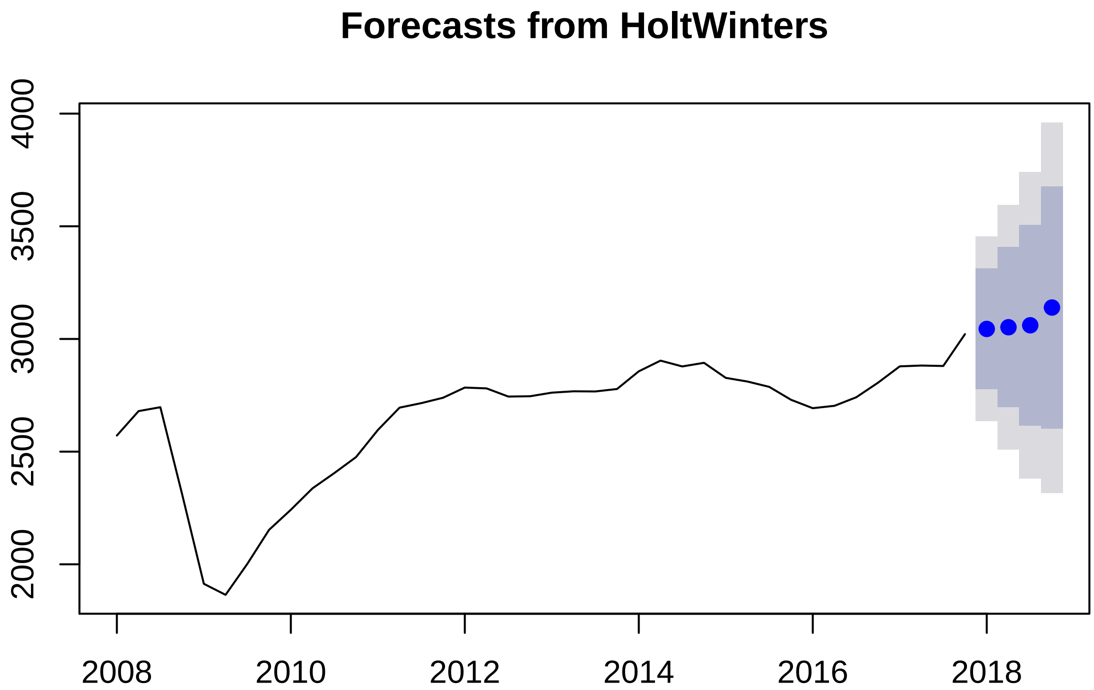
```r
HW <- HoltWinters(NITS)

HW
```

```
## Holt-Winters exponential smoothing with trend and additive seasonal component.
##
## Call:
## HoltWinters(x = NITS)
##
## Smoothing parameters:
##  alpha: 0.7428379
##  beta : 0.1730485
##  gamma: 1
##
## Coefficients:
##          [,1]
## a  2904.32028
## b    29.47630
## s1  110.70636
## s2   88.73853
## s3   67.93749
## s4  117.27972
```

```r
HW_forecast <- forecast(HW, h= 4)
HW_forecast
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3044.503 2776.792 3312.213 2635.075 3453.931
## 2018 Q2       3052.011 2696.923 3407.100 2508.950 3595.073
## 2018 Q3       3060.687 2615.999 3505.374 2380.596 3740.778
## 2018 Q4       3139.505 2601.904 3677.106 2317.315 3961.695
```

```r
plot(HW_forecast)
```



o What is the value of alpha? What does that value signify?

The value of alpha = 0.7428379 The estimated value of alpha is high, telling us that the estimate of the current value of the level is based mostly upon very recent data in the time series; i.e older values in the time series are weighted less as companre to the recent data.

o What is the value of beta? What does that value signify?

The value of beta = 0.1730485.

The value of beta is 0.1730485, Indicates the trend component is getting updated over the time series, beta takes care of level change. Here small value of beta tells that time series is giving weight to past data slope more then the current one.

o What is the value of gamma? What does that value signify? The value of gamma = 1 The valus of gamma is relatively extremely high which means that older values are not considered in the time series but the current lag is weighted more heavily.

o What is the value of initial states for the level, trend and seasonality? What do these values signify?

```
HW$coefficients
```

```
##          a          b         s1         s2         s3         s4
## 2904.32028   29.47630  110.70636   88.73853   67.93749  117.27972
```

->a,b,s1 contain the initial estimated values for the level, trend and seasonal components respectively.

o What is the value of sigma? What does the sigma signify?

```
sd(complete.cases(HW_forecast$residuals))
```

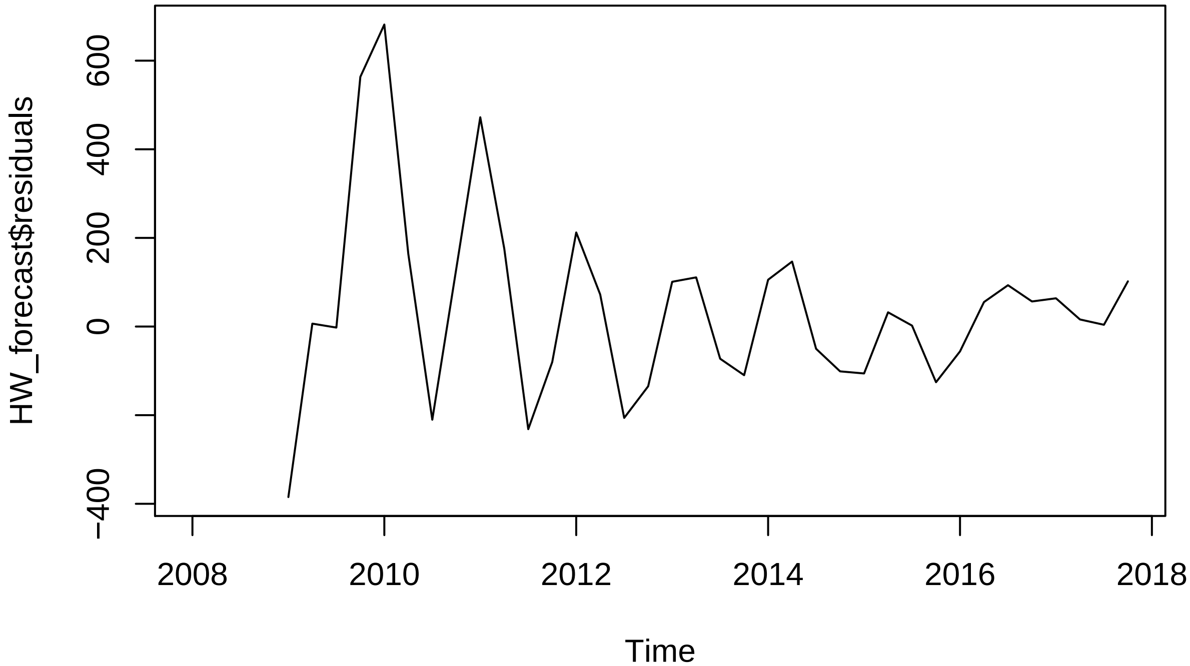
```
## [1] 0.3038218
```

->Above we calculated the Standard deviation of residuals which means it is small and implies goodness of the model.

- Perform Residual Analysis for this technique.

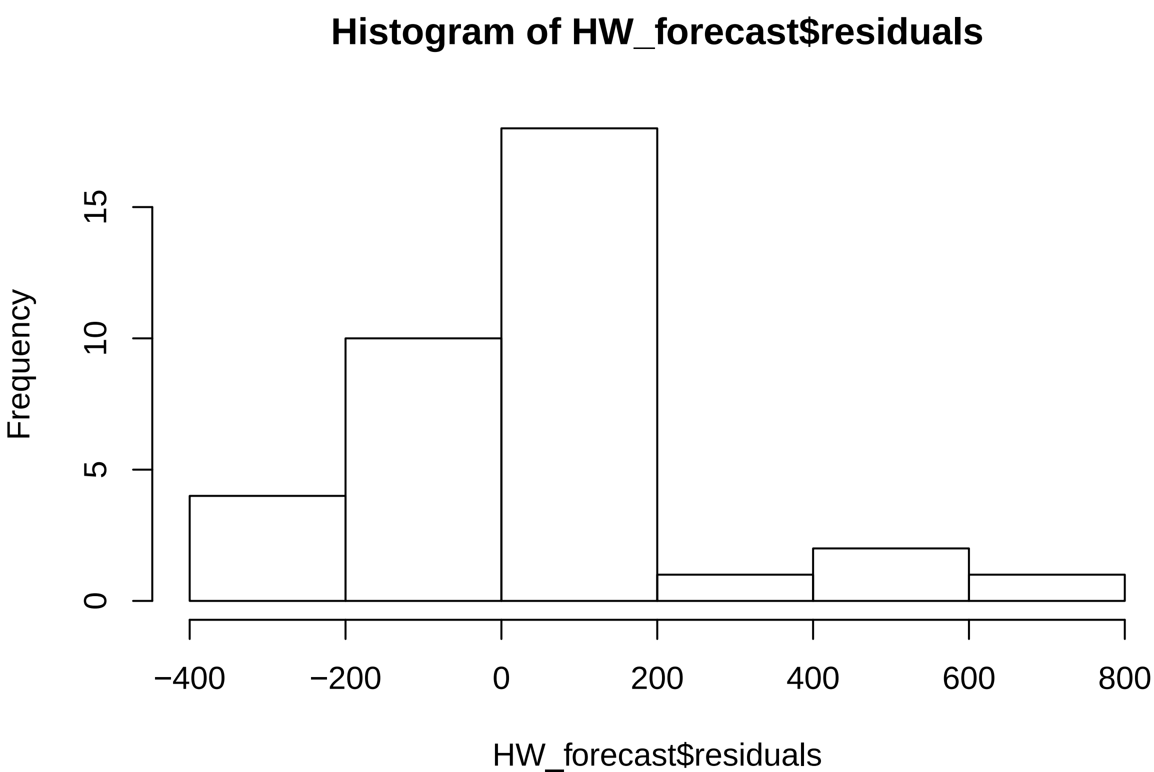
o Do a plot of residuals. What does the plot indicate?

```r
plot(HW_forecast$residuals)
```



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

```r
hist(HW_forecast$residuals)
```

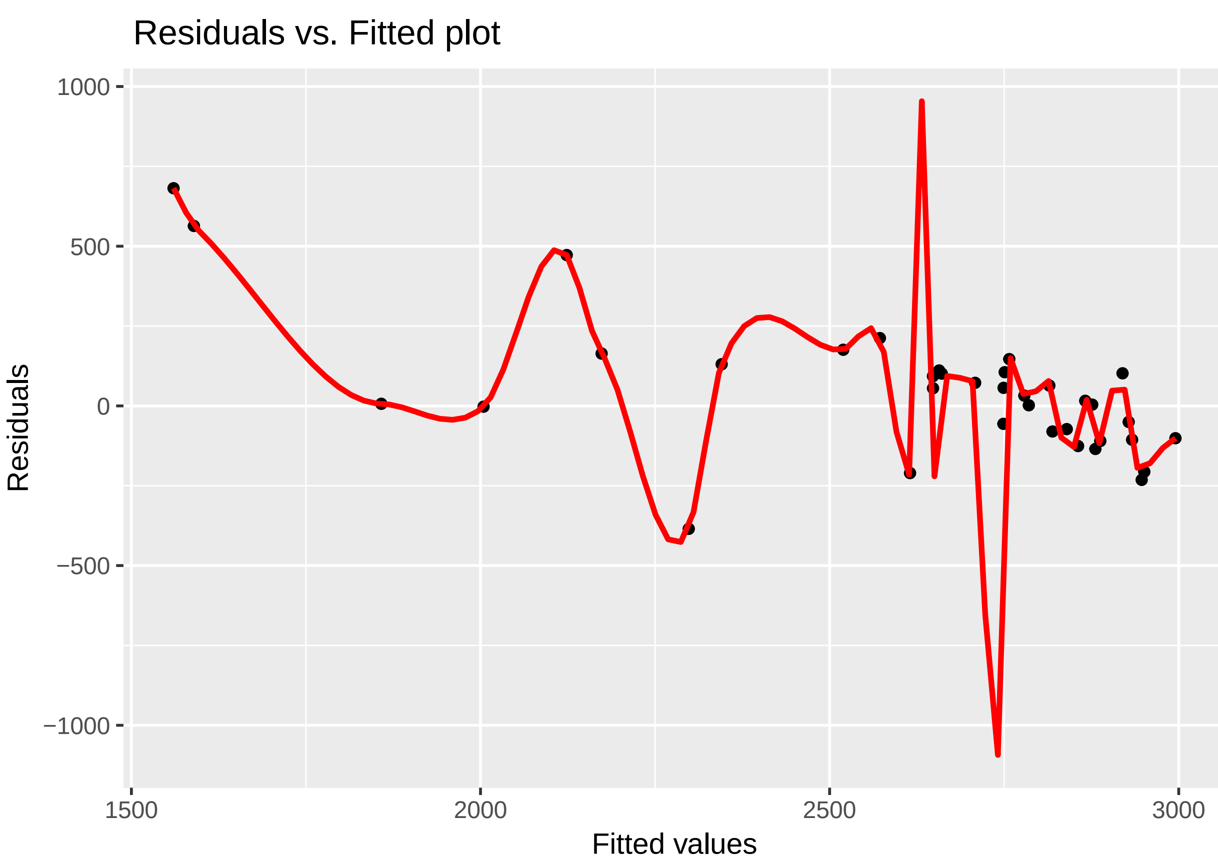


->The histogram plot of the residuals suggests that the residuals are not following normal distribution as they are right skewed.

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

```r
library(ggplot2)
qplot(y = HW_forecast$residuals, x = HW_forecast$fitted,
      ylab = "Residuals", xlab = "Fitted values",
      main = " Residuals vs. Fitted plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.

## Warning: Removed 4 rows containing non-finite values (stat_smooth).

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1553.2

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 304.75

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 3084.5

## Warning: Removed 4 rows containing missing values (geom_point).
```

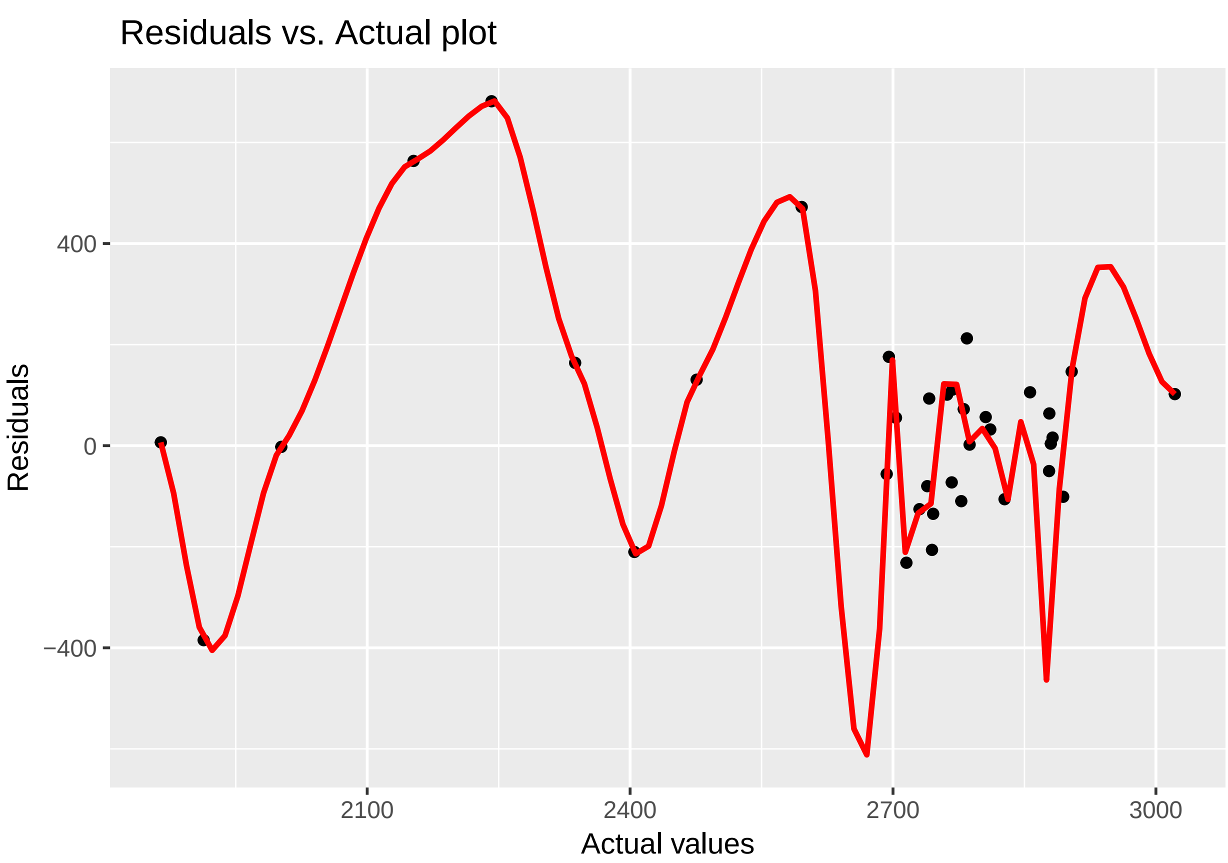


In the above plot of Residuals VS Fitted values shows residuals has a pattern initially, later they are randomly distributed among themself.

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

```
library(ggplot2)
qplot(y = HW_forecast$residuals, x = NITS,
      ylab = "Residuals", xlab = "Actual values",
      main = " Residuals vs. Actual plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

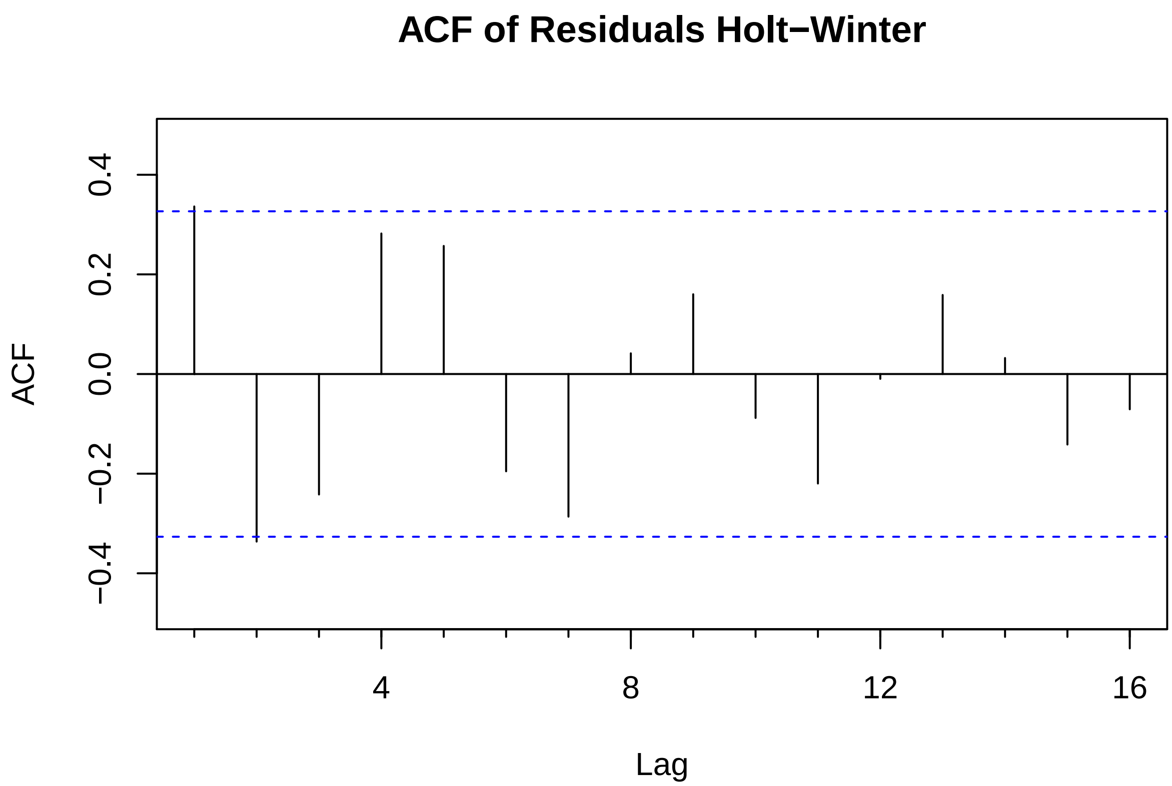
```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.

## Warning: Removed 4 rows containing non-finite values (stat_smooth).

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1858.7

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 143.29

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 17738

## Warning: Removed 4 rows containing missing values (geom_point).
```



In the above plot of Residuals VS Actual we can see that they are not spreaded sporadiclly and they dont have any pattern among them.There are drifted on a side from 2700.

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

```
Acf(HW_forecast$residuals, main = "ACF of Residuals Holt-Winter")
```



->Spikes shows the values of Autocorrelation with each lags. We can observe that amplitude of each spike is in the blue segment which implies they are insignificant.Hence the Autocorrelation is insignificant.

- Print the 5 measures of accuracy for this forecasting technique

```
accuracy_HW <- accuracy(HW_forecast)
accuracy_HW
```

```
##                    ME     RMSE      MAE     MPE     MAPE      MASE
## Training set 41.5499 210.1229 145.5452 1.67805 5.850816 0.7025981
##                   ACF1
## Training set 0.3363672
```

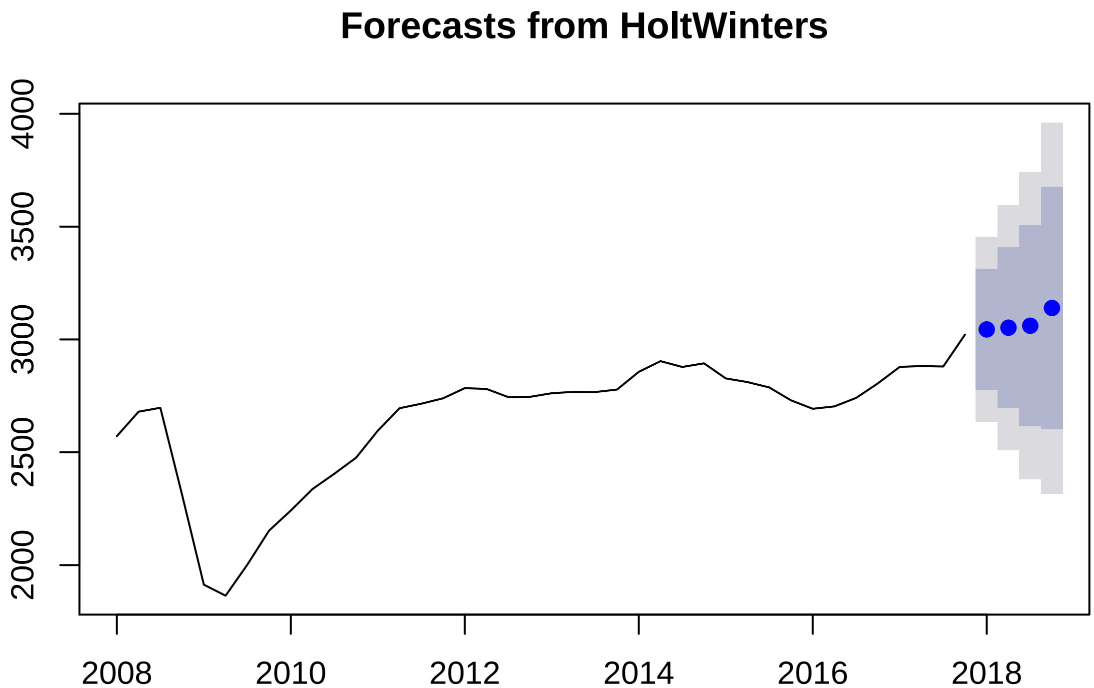
- Forecast

o Time series value for next year. Show table and plot

```
HW_forecast
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3044.503 2776.792 3312.213 2635.075 3453.931
## 2018 Q2       3052.011 2696.923 3407.100 2508.950 3595.073
## 2018 Q3       3060.687 2615.999 3505.374 2380.596 3740.778
## 2018 Q4       3139.505 2601.904 3677.106 2317.315 3961.695
```

```
plot(HW_forecast)
```



- Summarize this forecasting technique

```r
summary(HW_forecast)
```

```
##
## Forecast method: HoltWinters
##
## Model Information:
## Holt-Winters exponential smoothing with trend and additive seasonal component.
##
## Call:
## HoltWinters(x = NITS)
##
## Smoothing parameters:
##  alpha: 0.7428379
##  beta : 0.1730485
##  gamma: 1
##
## Coefficients:
##          [,1]
## a  2904.32028
## b    29.47630
## s1  110.70636
## s2   88.73853
## s3   67.93749
## s4  117.27972
##
## Error measures:
##                   ME     RMSE      MAE     MPE     MAPE      MASE
## Training set 41.5499 210.1229 145.5452 1.67805 5.850816 0.7025981
##                   ACF1
## Training set 0.3363672
##
```

```
## Forecasts:
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3044.503 2776.792 3312.213 2635.075 3453.931
## 2018 Q2       3052.011 2696.923 3407.100 2508.950 3595.073
## 2018 Q3       3060.687 2615.999 3505.374 2380.596 3740.778
## 2018 Q4       3139.505 2601.904 3677.106 2317.315 3961.695
```

o How good is the accuracy?

The accuracy is not that great. If we compare with the other models. Its performance is the worst among them.

o What does it predict the value of time series will be in one year? 3139.505 with 95 percent confidence interval, It would be 2317.315 low and 3961.695 high.

o Other observation During the residual analysis, we observed that they are biased and not normally distributed among them.

ARIMA or Box-Jenkins

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

```r
# ADF test says differences is required if p-value is > 0.05
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 3.4.4
```

```r
adf.test(NITS)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  NITS
## Dickey-Fuller = -2.907, Lag order = 3, p-value = 0.2175
## alternative hypothesis: stationary
```

```r
# Kipps test says differences is required if p-value is < 0.05
kpss.test(NITS)
```

```
## Warning in kpss.test(NITS): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  NITS
## KPSS Level = 1.1832, Truncation lag parameter = 1, p-value = 0.01
```

Inference: From the ADF and Kipps test, its evident that series is not stationary hence differences is required.

- How many differences are needed to make it stationary?

```r
# NSDIFFS only works for seasonal data
nsdiffs(NITS)
```

```
## [1] 0
```

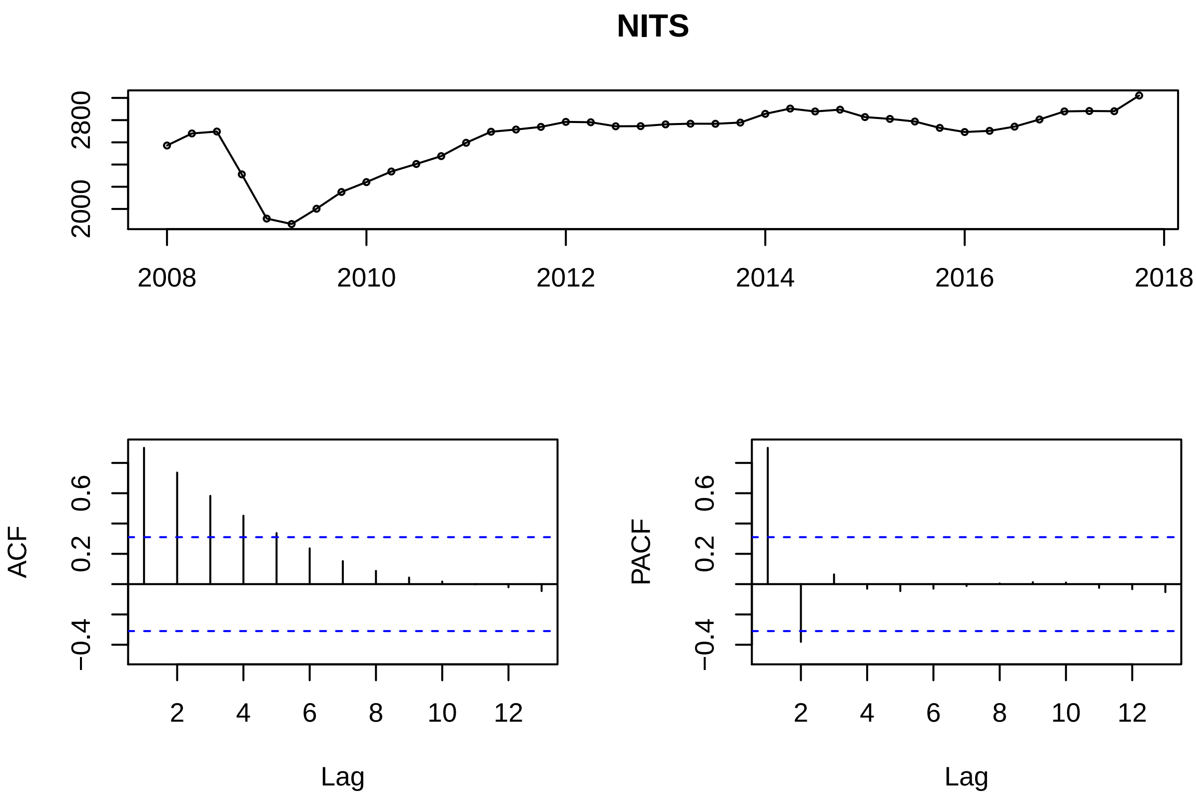
```r
#  NDIFFS works with non-seasonal data
ndiffs(NITS)
```

```
## [1] 1
```

Inference: From the test, we can conclude that 1 difference is required to make series stationary.

- Is Seasonality component needed?

```r
tsdisplay(NITS)
```




```r
fit <- tbats(NITS)
seasonal <- !is.null(fit$seasonal)
seasonal
```

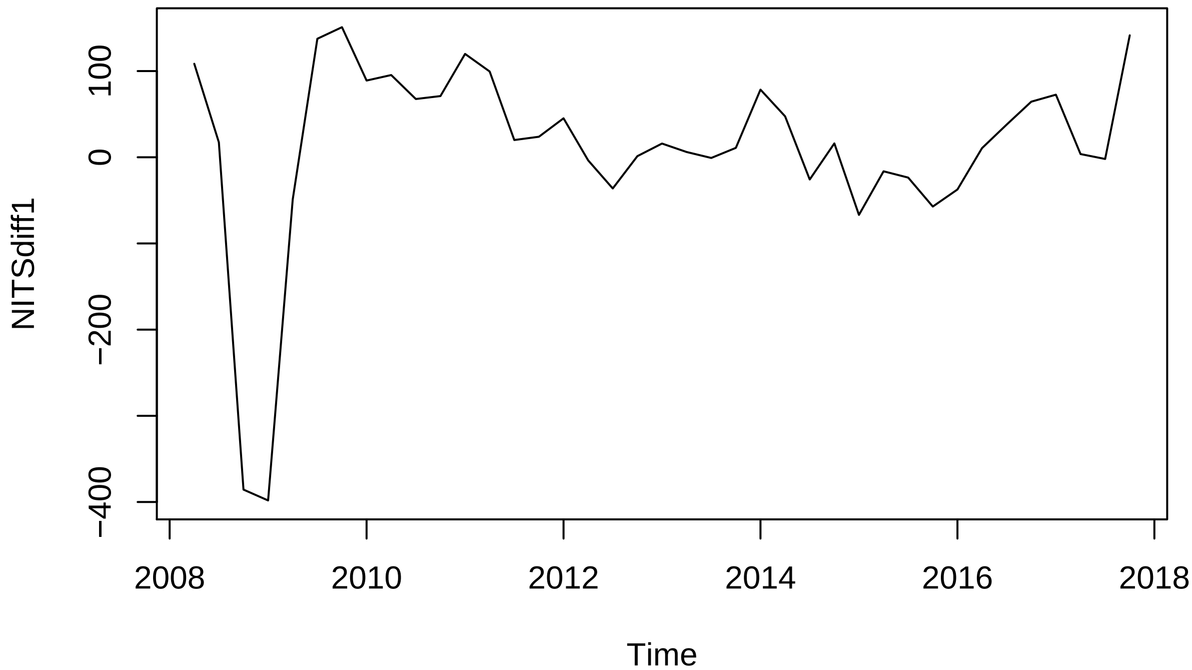
```
## [1] FALSE
```

```r
# seasonal will be TRUE if a seasonal model is chosen and otherwise FALSE
```

Inference: From the acf, we can see that autocorrelation indicated absence of seasonality in the model which is again proved by the above test because if it has selected a seasonal model for the series result could have been True.

- Plot the Time Series chart of the differenced series.

```r
NITSdiff1 <- diff(NITS, differences=1)
plot(NITSdiff1)
```




```r
adf.test(NITSdiff1)
```

```
## Warning in adf.test(NITSdiff1): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  NITSdiff1
## Dickey-Fuller = -4.9257, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(NITSdiff1)
```

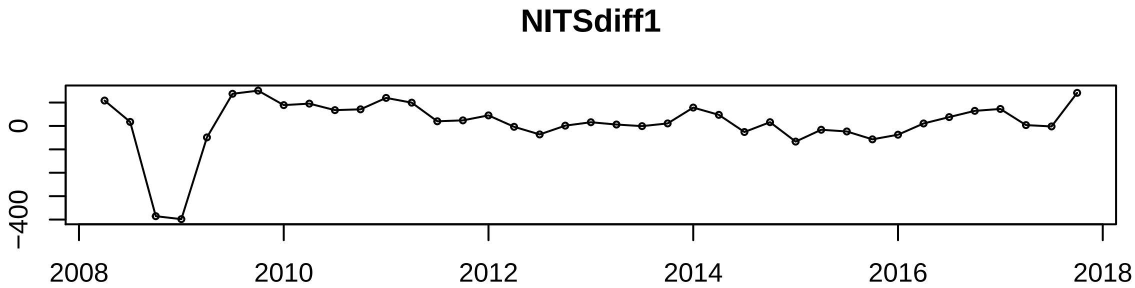
```
## Warning in kpss.test(NITSdiff1): p-value greater than printed p-value
```

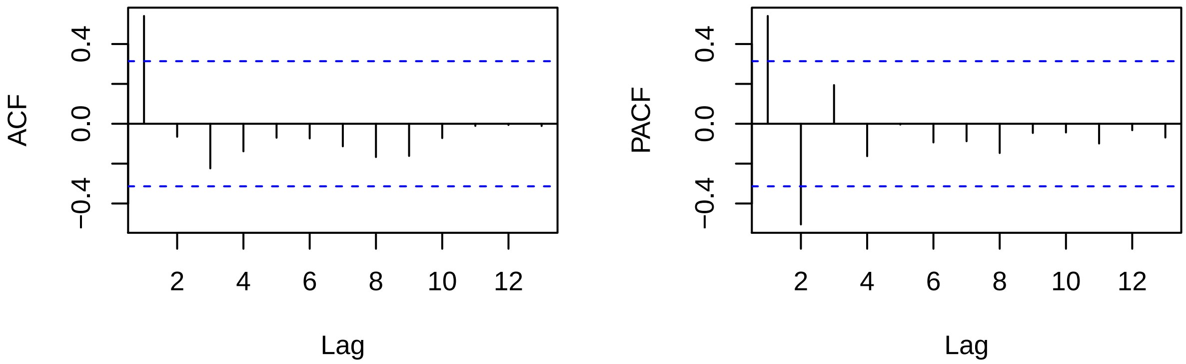
```
##
##  KPSS Test for Level Stationarity
##
## data:  NITSdiff1
## KPSS Level = 0.093613, Truncation lag parameter = 1, p-value = 0.1
```

Inference: The retest indicates that the series is now stationary and further difference is not required.

- Plot the ACF and PACF plot of the differenced series.

```r
tsdisplay(NITSdiff1)
```





- Based on the ACF and PACF, which are the possible ARIMA model possible?

From the ACF we can see that from 2nd lag autocorrelation is insignificant hence the possible value of q is 1. From the PACF we can see the first 2 lags are insignificant later none of them is significant the possible value of p could be 1.

Hence the possibe model could be (2,1,1),(1,1,1),(2,1,0)

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

```
NiTSarima1 <- arima(NITS, order=c(2,1,1))
NiTSarima1
```

```
##
## Call:
## arima(x = NITS, order = c(2, 1, 1))
##
## Coefficients:
##          ar1      ar2     ma1
##       0.5013  -0.3525  0.6409
## s.e.  0.1958   0.1824  0.1657
##
## sigma^2 estimated as 4912:  log likelihood = -221.93,  aic = 451.87
```

```
NiTSarima2 <- arima(NITS, order=c(1,1,1))
NiTSarima2
```

```
##
## Call:
## arima(x = NITS, order = c(1, 1, 1))
##
## Coefficients:
```

```
##          ar1     ma1
##       0.2973  0.7816
## s.e.  0.1682  0.0924
##
## sigma^2 estimated as 5372:  log likelihood = -223.55,  aic = 453.1
```

```
NiTSarima3 <- arima(NITS, order=c(2,1,0))
NiTSarima3
```

```
##
## Call:
## arima(x = NITS, order = c(2, 1, 0))
##
## Coefficients:
##          ar1      ar2
##       0.8955  -0.5622
## s.e.  0.1362   0.1321
##
## sigma^2 estimated as 5636:  log likelihood = -224.34,  aic = 454.68
```

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

Based on AIC, BIC and Siqma^2 we can see that model (2,1,1) is the best among all which lowest AIC value however we can validate our claim by running autocorrelation function to see what it proposes to us.

```
auto.arima(NITS)
```

```
## Series: NITS
## ARIMA(2,1,1)
##
## Coefficients:
##          ar1      ar2     ma1
##       0.5013  -0.3525  0.6409
## s.e.  0.1958   0.1824  0.1657
##
## sigma^2 estimated as 5322:  log likelihood=-221.93
## AIC=451.87   AICc=453.04   BIC=458.52
```

Hence The claimed model(2,1,1) came out as the best from the ARIMA function.

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

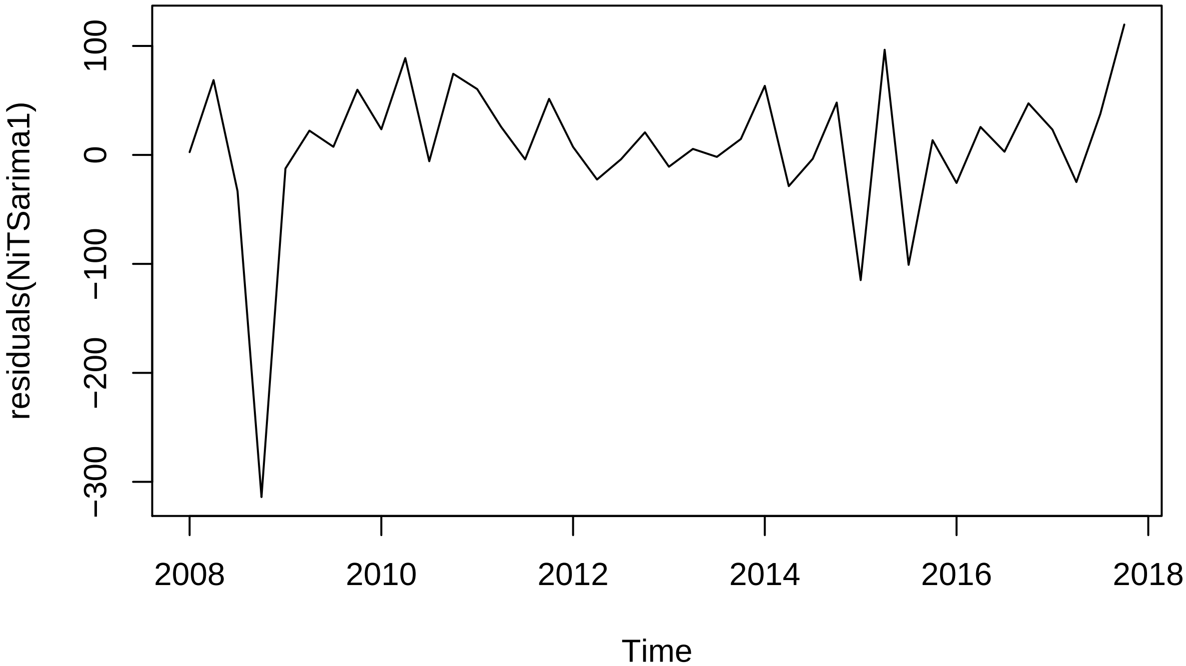
```
NiTSarima1 <- arima(NITS, order=c(2,1,1))
NiTSarima1
```

```
##
## Call:
## arima(x = NITS, order = c(2, 1, 1))
##
## Coefficients:
##          ar1      ar2     ma1
##       0.5013  -0.3525  0.6409
## s.e.  0.1958   0.1824  0.1657
##
## sigma^2 estimated as 4912:  log likelihood = -221.93,  aic = 451.87
```

- Perform Residual Analysis for this technique.

o Do a plot of residuals. What does the plot indicate?

```r
plot(residuals(NiTSarima1))
```

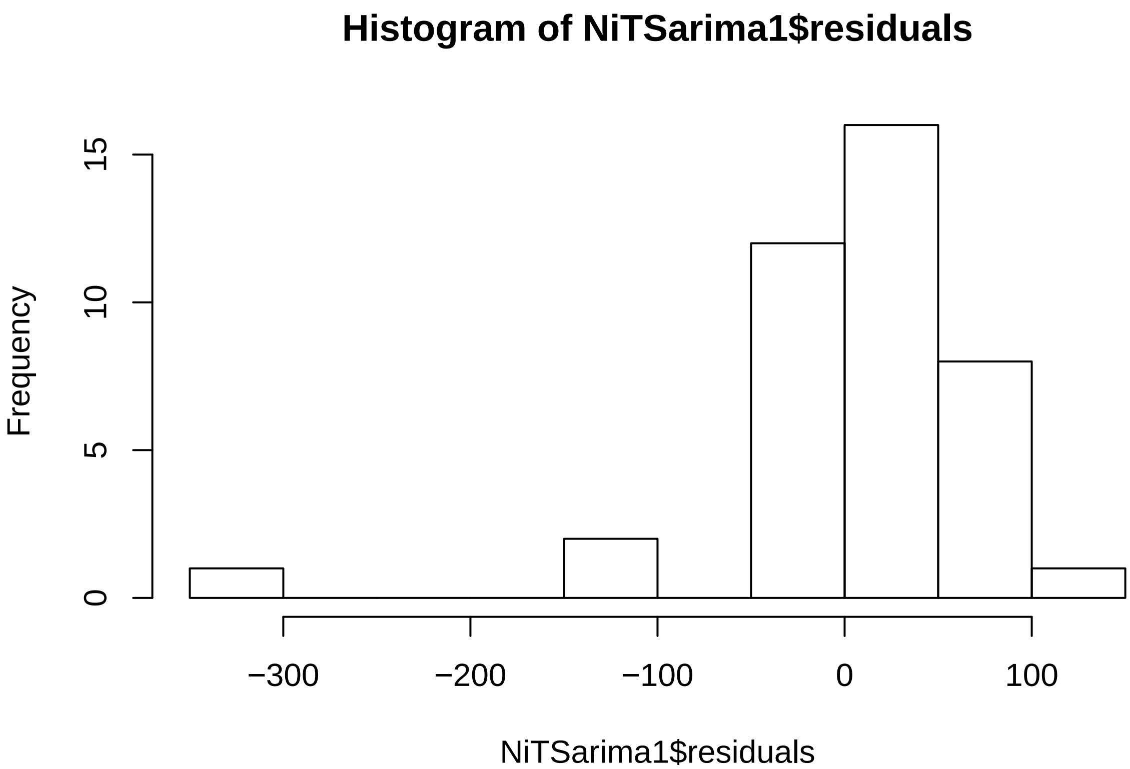


-> By observing the above graph of residuals,we can see that residuals stays same across the historical data.Hence the it can be considered constant over the time.

-> There are few spikes observed in the residues(2009,2015), which can be due to some special event occurance in the country or globally which influenced the import. ->2009 is due to the global slow down.

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

```r
hist(NiTSarima1$residuals)
```

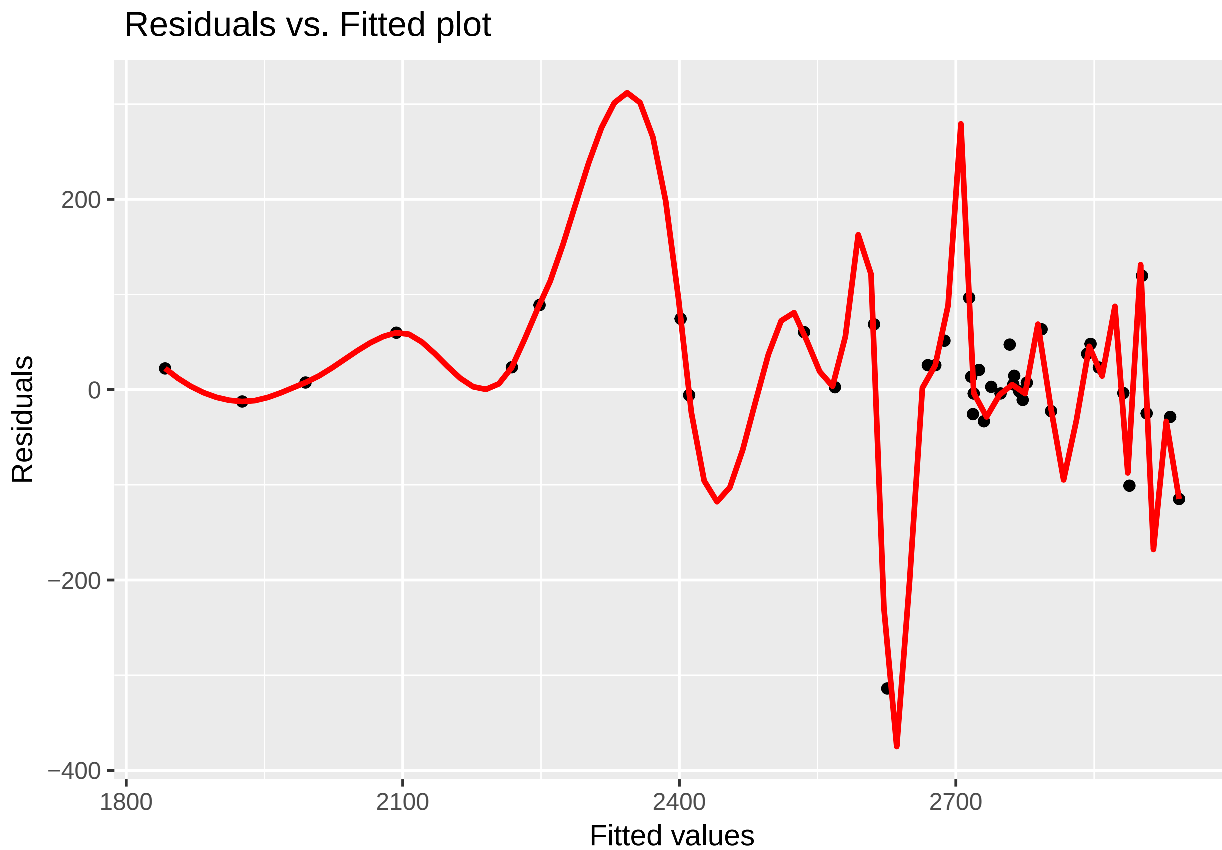


->->The histogram plot of the residuals suggests that the residuals are not normal distributed. They are little right skewed.

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

```
library(ggplot2)
qplot(y = NiTSarima1$residuals, x = fitted(NiTSarima1),
      ylab = "Residuals", xlab = "Fitted values",
      main = " Residuals vs. Fitted plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```

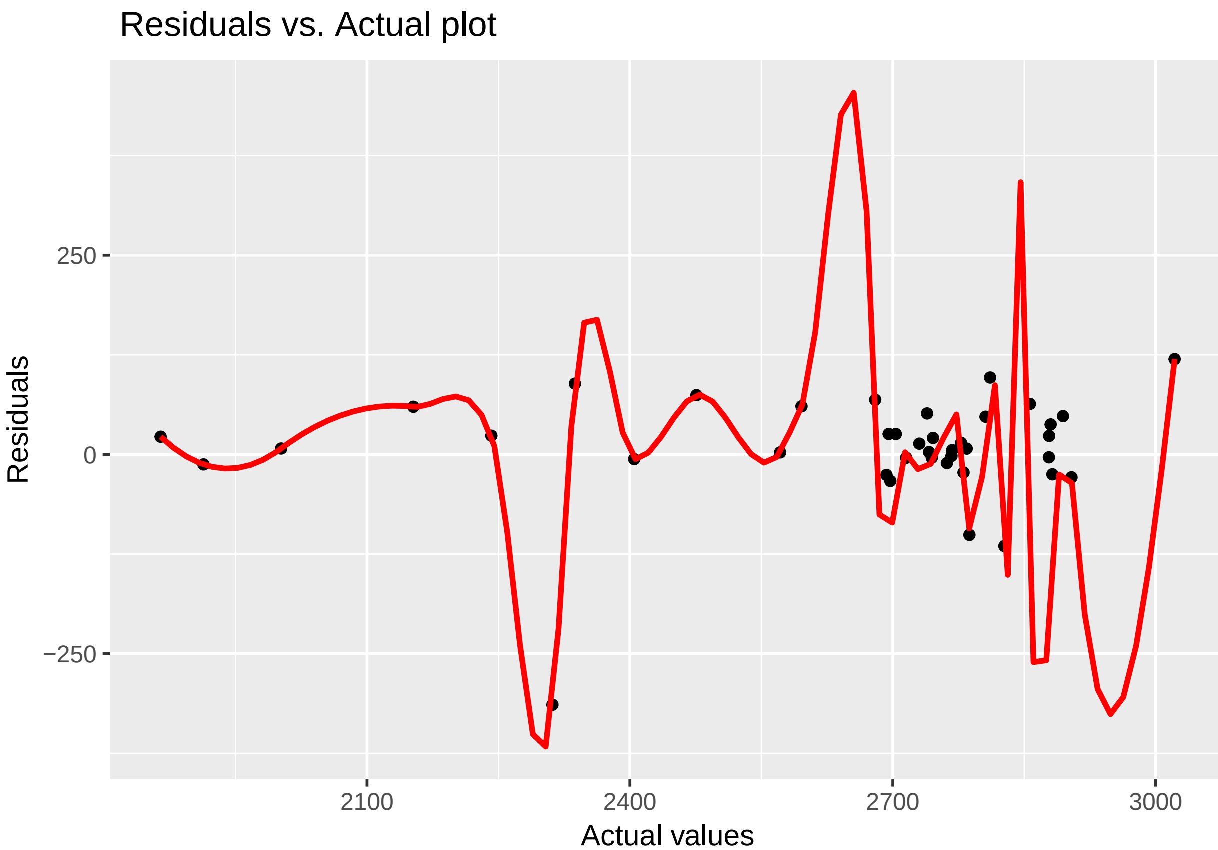


->In the above plot of Residuals VS Fitted values shows residuals has a pattern and they are not fully randomly distributed among themself. They are distributed randomly after 2600 .

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

```
library(ggplot2)
qplot(y = NiTSarima1$residuals, x = NITS,
      ylab = "Residuals", xlab = "Actual values",
      main = " Residuals vs. Actual plot") +
  stat_smooth(method = "loess", span = 0.1, colour = I("red"), se = FALSE)
```

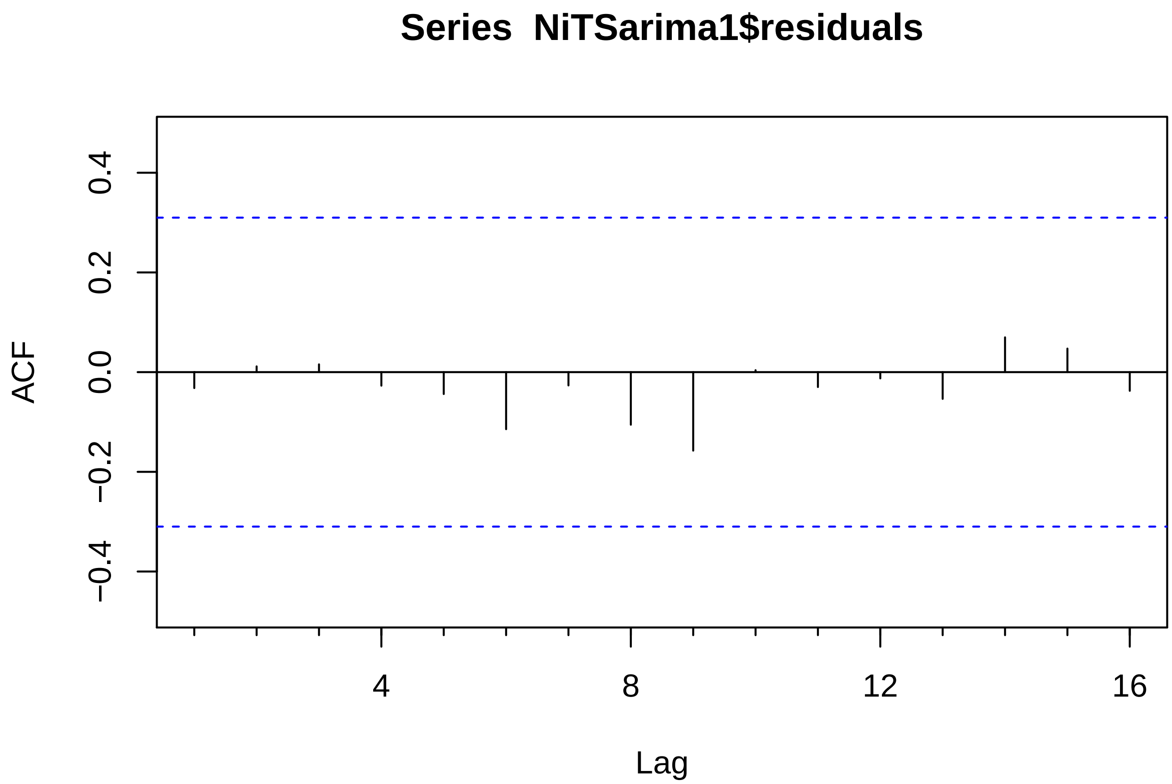
```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 2880.1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in sqrt(sum.squares/one.delta): NaNs produced
```



->In the above figure, we observe that the Residuals are not spreaded sporadiclly and they have any pattern among them.

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

```
Acf(NiTSarima1$residuals)
```



->Spikes shows the values of Autocorrelation with each lags. We can observe that amplitude of each spike is in the blue segment which implies they are insignificant.Hence the Autocorrelation is insignificant.

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

```
arimaaccuracy<-accuracy(NiTSarima1)
arimaaccuracy
```

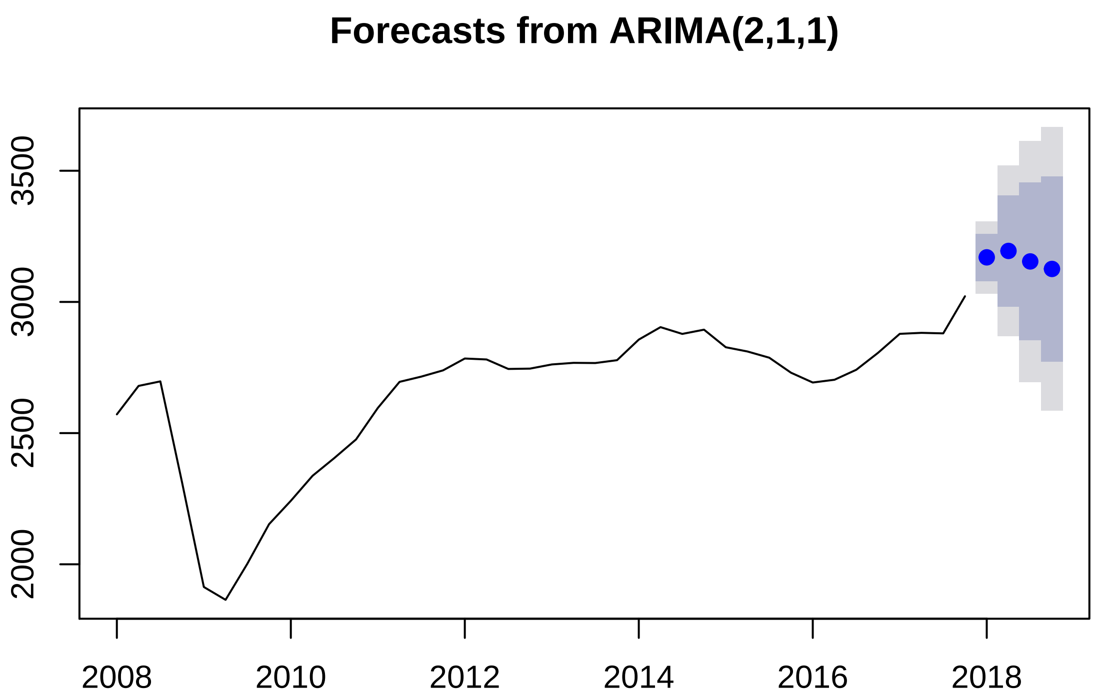
```
##                    ME     RMSE      MAE      MPE     MAPE      MASE
## Training set 7.607852 69.20748 42.97381 0.263098 1.657254 0.6312061
##                     ACF1
## Training set -0.03197029
```

- Forecast o Next one year. Show table and plot

```
forecast(NiTSarima1,h=4)
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3169.929 3080.107 3259.750 3032.559 3307.298
## 2018 Q2       3194.406 2982.057 3406.755 2869.646 3519.166
## 2018 Q3       3154.390 2854.198 3454.581 2695.287 3613.493
## 2018 Q4       3125.701 2772.594 3478.808 2585.670 3665.732
```

```
plot(forecast(NiTSarima1,h=4))
```

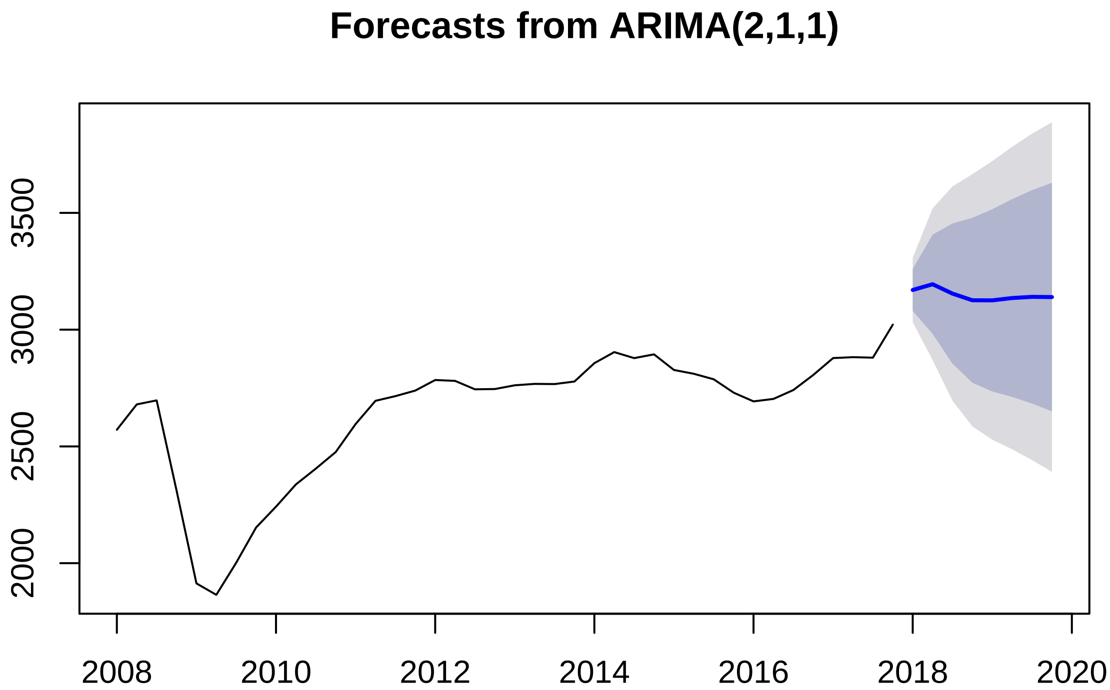


o Next two years. Show table and plot

```
forecast(NiTSarima1,h=8)
```

```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018 Q1       3169.929 3080.107 3259.750 3032.559 3307.298
## 2018 Q2       3194.406 2982.057 3406.755 2869.646 3519.166
## 2018 Q3       3154.390 2854.198 3454.581 2695.287 3613.493
## 2018 Q4       3125.701 2772.594 3478.808 2585.670 3665.732
## 2019 Q1       3125.426 2735.331 3515.521 2528.827 3722.024
## 2019 Q2       3135.400 2711.859 3558.941 2487.650 3783.150
## 2019 Q3       3140.498 2683.500 3597.496 2441.579 3839.416
## 2019 Q4       3139.537 2650.077 3628.997 2390.972 3888.102
```

```
plot(forecast(NiTSarima1,h=8))
```



- Summarize this forecasting technique:

o How good is the accuracy? It performed very well . It has the minimum MAPE 1.657254 among all of the meothods.

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

3125.701 with 95 percent confidence interval, It would be 2585.670 low and 3665.732 high. 3139.537 with 95 percent confidence interval, It would be 2390.972 low and 3888.102 high.

o Other observation

->During the residual analysis, we observed that they are residuals are not normally distributed.

Accuracy Summary

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

```
final_accuracy <- rbind(naive_accuracy,accuracy_ets,accuracy_HW,arimaaccuracy)

rownames(final_accuracy) <- c("Naive Method", "ETS", "Holt-Winter","ARIMA")

final_accuracy
```

```
##                     ME      RMSE       MAE       MPE     MAPE      MASE
## Naive Method 11.543590 109.70538  68.08205 0.2970542 2.852688 0.3286562
## ETS           4.922542  94.01716  56.08520 0.2528908 2.328826 0.2707432
## Holt-Winter  41.549897 210.12287 145.54515 1.6780501 5.850816 0.7025981
## ARIMA         7.607852  69.20748  42.97381 0.2630980 1.657254 0.6312061
##                     ACF1
## Naive Method  0.54057850
## ETS           0.22031024
## Holt-Winter   0.33636724
## ARIMA        -0.03197029
```

- Separately define each forecast method and why it is useful. Show the best and worst forecast method for each of the accuracy measures.

-> Naive Forecast: Naïve 1 forecasts are often used as a benchmark when assessing the accuracy of a set of forecasts. A ratio is obtained to show the upper bound of a forecasting method's accuracy relative to naïve 1 forecasts when the mean squared error is used to measure accuracy.It is know as no change forecast which has been observed while forecasting above.

->Simple Moving Average: It is the weighted average of the previous n data. It is used when recent observations influence more than the previous observations. As new data comes in , newest value is added and oldest value is dropped. Equal weights are assigned to each observation which is not considering seasonality and trend of the time series.

-> Simple Smoothing: When forecaster believes more-recent observations are likely to contain more information, this is the technique to use. This method is suitable for forecasting data with no trend or seasonal pattern. The main aim is to estimate the current level. The level estimate is then used to forecast future values. Since the most recent period's forecast was created based on the previous period's demand and the previous period's forecast, which was based on the demand for the period before that and the forecast for the period before that.

->Holt Winters: Holt Winters has levels which are level, trend ans seasonality. Hence it is called Triple Exponential Smoothing There is additive method and multiplicative method. It is used when forecast data points in a series, provided that the series is "seasonal", i.e. repetitive over some period.

->ARIMA: An ARIMA model produces forecasts based upon prior values in the time series (AR terms) and the errors made by previous predictions (MA terms). This typically allows the model to rapidly adjust for sudden changes in trend, resulting in more accurate forecasts.

We can select anyone of the method based on the business needs.

Best model forecast method for each of the accuracy measures: ME: Mean Error : -3.340844 - lowest -> ETS RMSE: Root Mean Squared Error: 52.12235 -> (Penalizes large errors) lowest -> ETS MAE: Mean Absolute Error: 39.17070 -> lowest -> ETS MPE: Mean Percentage Error: -0.4340073 -> closest to zero -> ETS MAPE: Mean Absolute Percentage Error: 2.741.6572549518 -> lowest -> ARIMA MASE: Mean Absolute Scaled Error: 0.6312061 -> lowest -> ARIMA ACF1: Autocorrelation of errors at lag 1: -0.03197029 -> lowest -> ARIMA

Worst model based on the accuracy measures: ME: Mean Error : 41.549897 - High -> Holt-Winter. RMSE: Root Mean Squared Error: 210.12287 -> (Penalizes large errors) High -> Holt-Winter MAE: Mean Absolute Error: 145.54515 -> high -> Holt-Winter MPE: Mean Percentage Error: 1.6780501 - high -> Holt-Winter MAPE: Mean Absolute Percentage Error: 5.850816 -> high -> Holt-Winter MASE: Mean Absolute Scaled Error: 0.7025981 -> High -> Holt-Winter ACF1: Autocorrelation of errors at lag 1: 0.54057850 -> High -> Naive Method

Conclusion

- Summarize your analysis of time series value over the time-period.

->It has been a great analysis as from the time series. I was able to understand that something abberant happened in year 2009 and i was able to relate the cause with the global slow down. -> Got to know about the quaters which has the minimum and maximum import. ->Maximum for the quarter 3rd which is influenced by the financial year start time and importing moreresources for impending winter season as the top 3 imports of states are Electrical machinery equipments, Vehicles and organic fossile fuels which summed to more than 45 percent of the total. ->Minimum for the winters as I believe it gets slow down due to the harsh weather.

- 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?

->For the next year, the series should increase and for the subsequent 2nd year should also increase as the market conditions are strong however It is also influenced by the global state and the political stability.

- Rank forecasting methods that best forecast for this time series based on historical values.

-> If we consider MAE, ETS performed well. -> If we consider MAPE, ARIMA performed well.

but overall I say ARIMA will performe well if we won't get all of sudden highs and lows due to unprecidented situations like global slowdown and tension among Russian and USA.