Forecasting Assignment

By : SHANBHAG Nitesh

July 16th, 2019

Contents

0.0.1 Exercise 1 : Forecasting of Road Fatalities in Belgium

This data set contains the road fatalities in Belgium from January 1995 to December 2017. The whole data is split into Train and Test, The Training set contains the information January 2001 up to December 2015 while the test set from January 2016 up to December 2017. The historical data is used to forecast the fatalities in advance so the authorities can take action and save lives.

Before starting the analysis, as a student having lived in Lille very close to Belgium, I made the hypothesis that the officials in Belgium would have taken the necessary steps and precautions which would definitely lead to decreasing trend in the fatalities. But would like to know if seasonality is a factor in road fatalities in Belgium.

0.0.1.1 1) Explore the data using relevant graphs

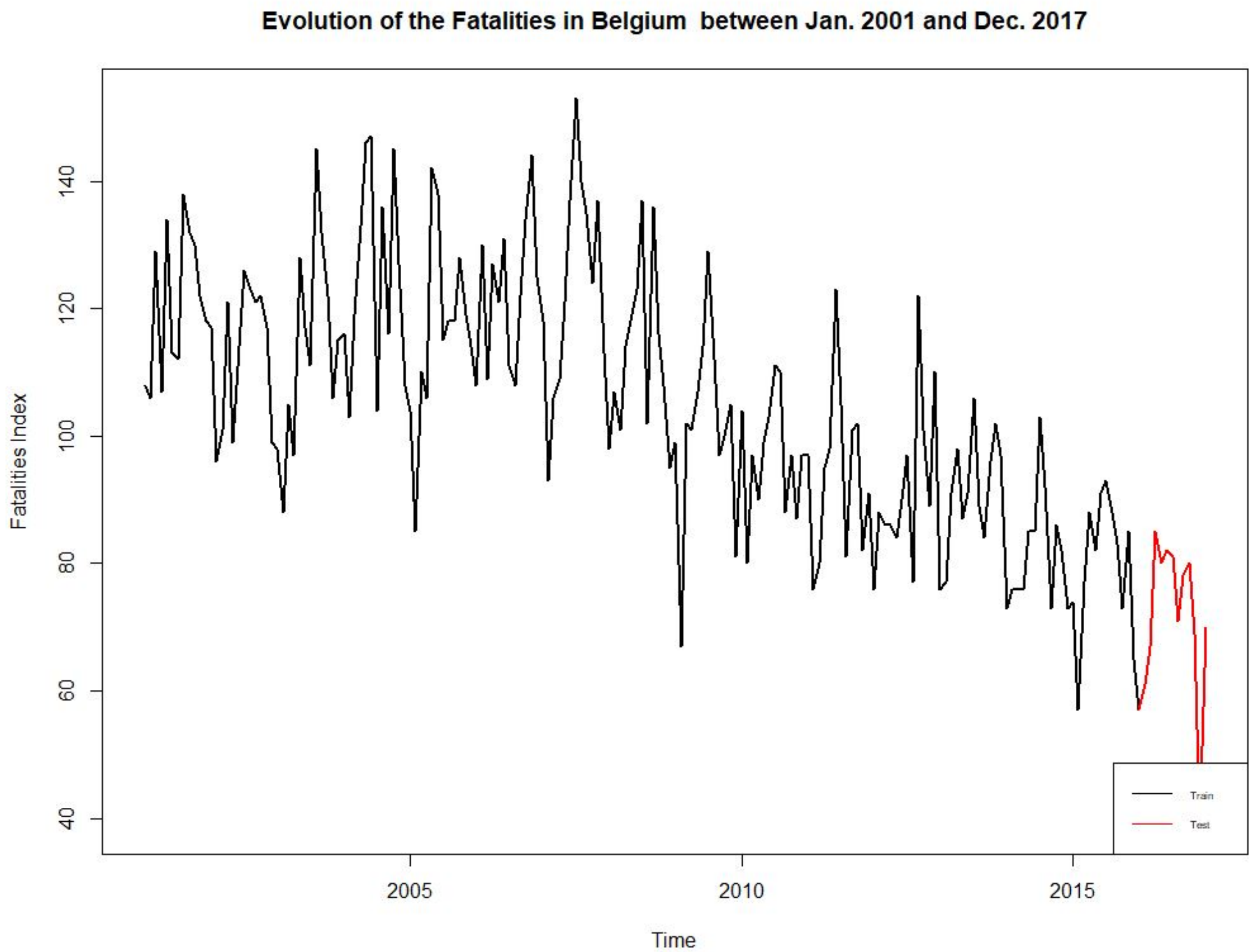
First I would like to explore data using appropriate graphs. For time series data, the best graph to start an analysis would be a classic time plot. This time plot will allow me to see if my previous null hypothesis are confirmed and will allow me to detect seasonality, cyclicity and a trend.

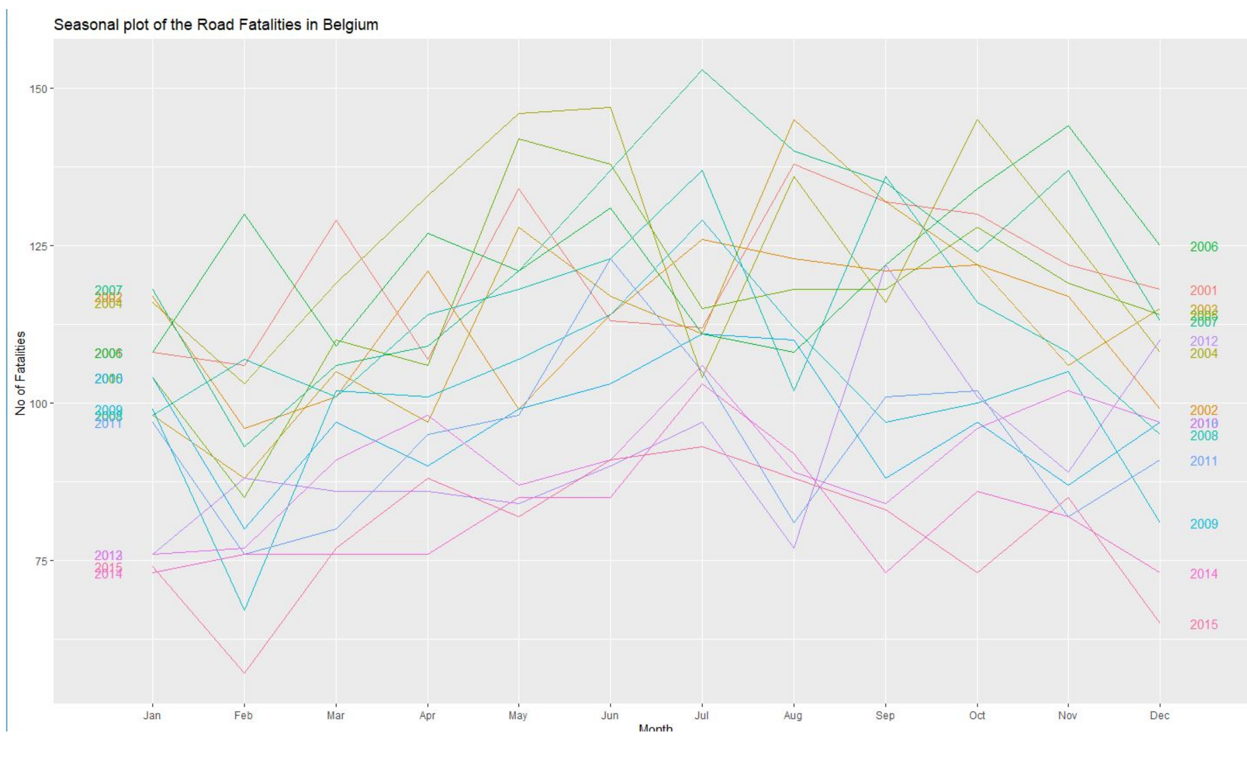
With the below first graph, we can observe a clear Decreasing trend. This trend seems to be constant in the time (the same slope). There is also a seasonal pattern that seems to slightly decrease with the years.

So further analysis on the Fatalities data will have to keep in mind the decreasing trend and seasonality.

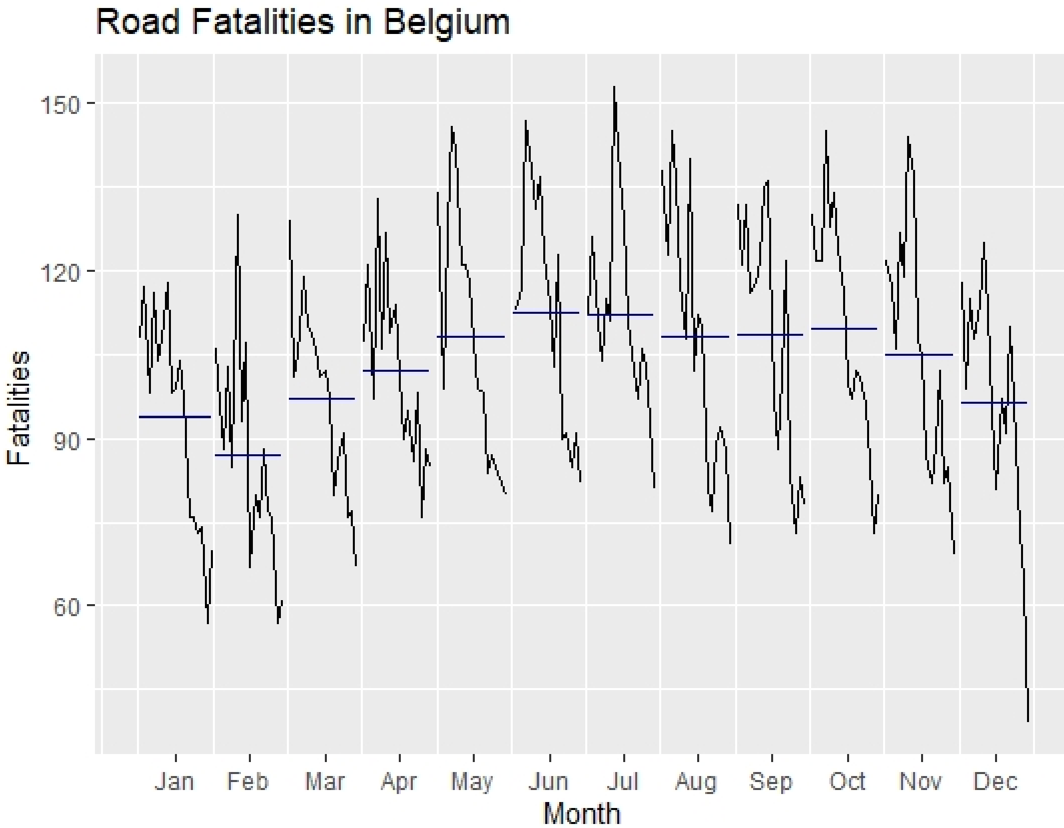
To deeper display the seasonality, I have used the second plot below:

In the First plot we can see a clear Decreasing trend over the years and in the second plot if we look closely there is lot less fatalities during (Dec,Jan,Feb) with an exception in 2006 but generally these 3 months have very less fatalities while compared to other. and a bit higher during the summer that is in June and July.


Evolution of the Fatalities in Belgium between Jan. 2001 and Dec. 2017
Fatalities Index
Time
Train
Test


Seasonal plot of the Road Fatalities in Belgium
No of Fatalities
Month

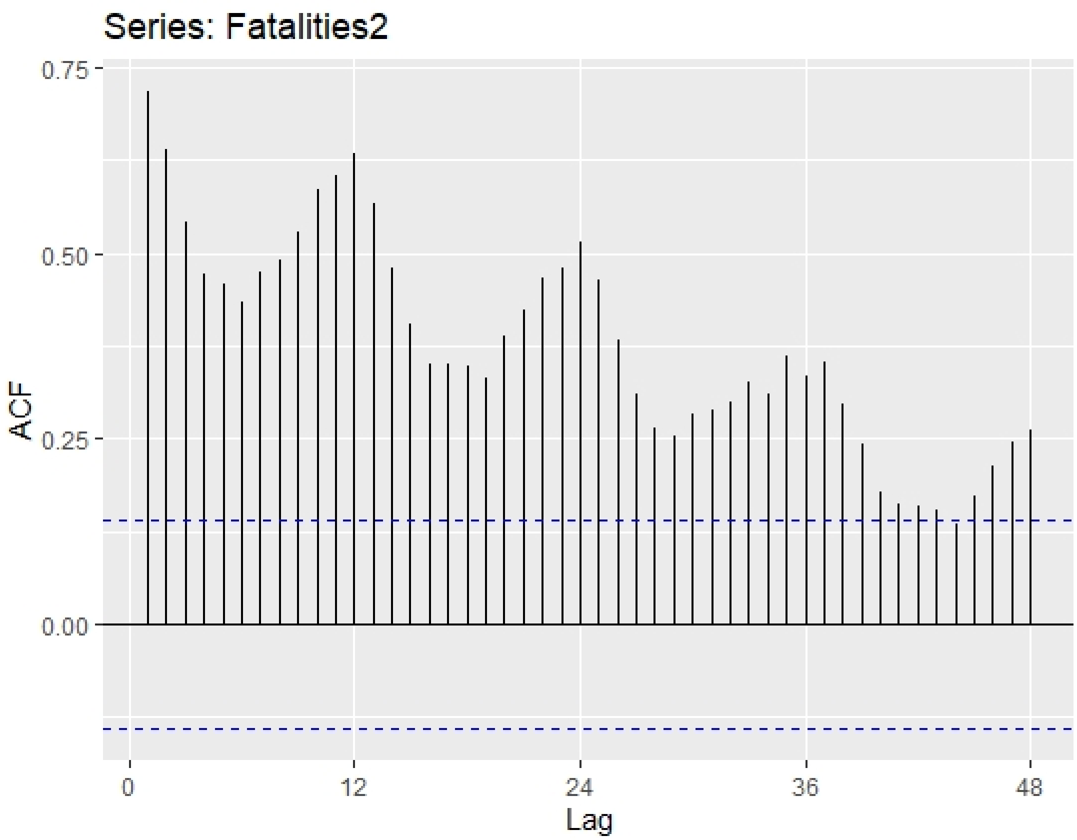
In the below plot we can clearly see the above mentioned observation about less fatalities in December,Jan,Feb and a bit higher in June and July.



The two graphs above confirm my hypothesis. The Number of Fatalities in Belgium is indeed a bit higher during Summer (May, June, July) and lower during Winter(Dec,Jan,Feb) and the same pattern seems to come back every year: Moreover, the trows graphs also confirm the Decreasing trend: The Road Fatalities in Belgium has decreased with the years.

Finally, the following ACF graph confirm the trend and the seasonality:

As explained by Hyndman R. & Athanasopoulos G., when data are seasonal, the autocorrelation will be larger for the seasonal lags than the other lag. Therefore, since we have a peak at lag 12,24,36 etc. one can conclude that the seasonality of yearly. Moreover, the ACF of trended time series tend to have positive values that slowly decrease of the lags increase which is also the case here. Thus, the ACF graph confirm the seasonality and the trend in the data.



0.0.1.2 2) Create forecasts using the seasonal naive method. Check the residual diagnostics and the forecast accuracy

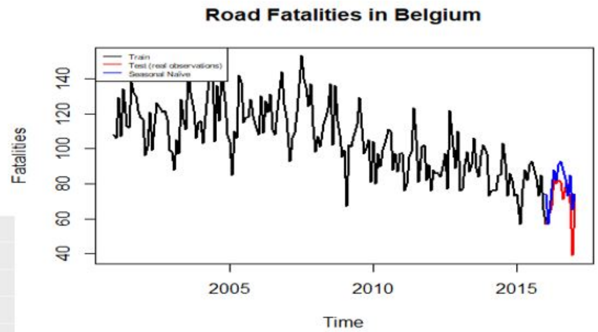
The seasonal naive method sets each forecast to be equal to the last observed value from the same season of the year.

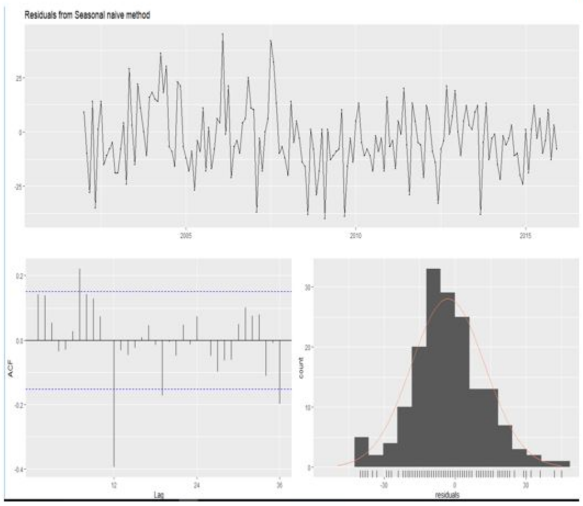
On the graph below, we can see the seasonal naive forecast (blue line) and we can compare it to the real value of the observation. One can see that since the seasonal naive method doesn't take the trend into consideration, the results aren't really good. The accuracy of both the training and the test set can also be found here below. They will be useful to compare the performance of the several models constructed.

Moreover, we can check the quality of the residuals:

The residuals are not white noise since the p-value of the Ljungbox test is lower the the threshold .05.

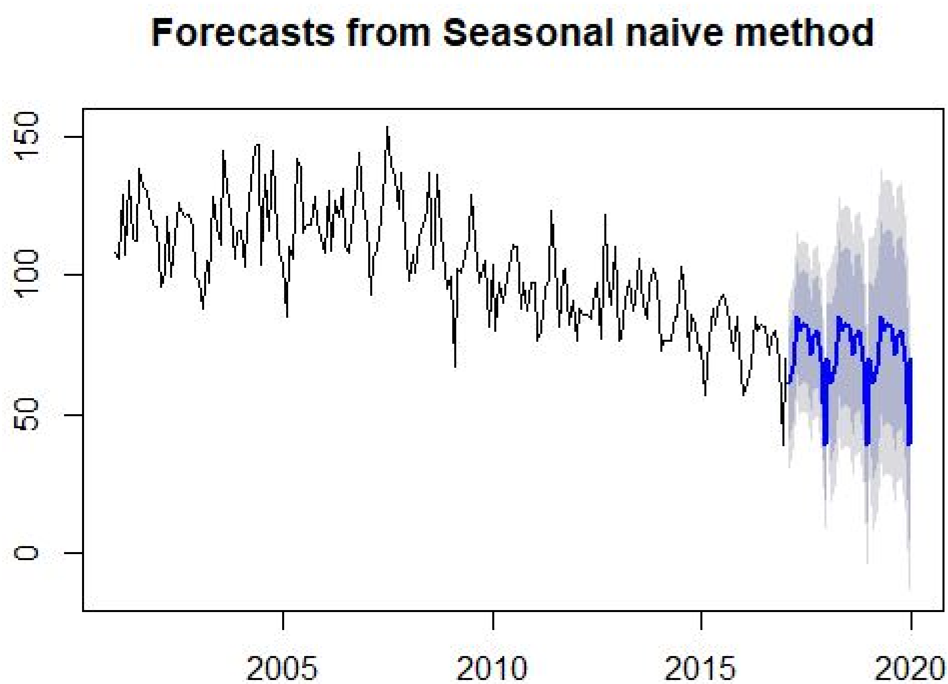
```
                 RMSE    MAE   MAPE    MASE
Training set 15.789 12.470 12.212 1.00000
Test set     12.260 10.154 16.754 0.81425
>
```





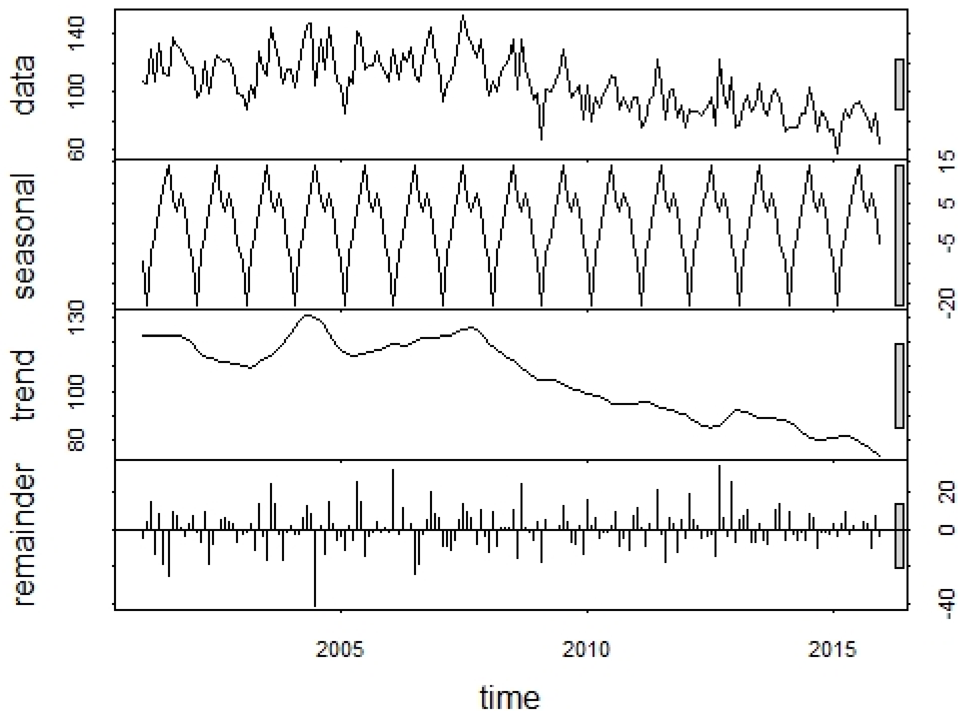

```
        Ljung-Box test

data:  Residuals from Seasonal naive method
Q* = 61.3, df = 24, p-value = 0.000042

Model df: 0.   Total lags used: 24
```

Final Forecasts until 2020 from Seasonal Naive Method:



0.0.1.3 3) Use an STL decomposition to forecast the time series. Use the appropriate underlying methods to do so. Check the residual diagnostics and the forecast accuracy.

In this exercise, since during the exploratory phase, we discover that there was a decreasing and constant trend, Hence I will use a random walk with drift model to forecast the seasonally adjusted time series. First, let's have a look to the STL decomposition graphs:

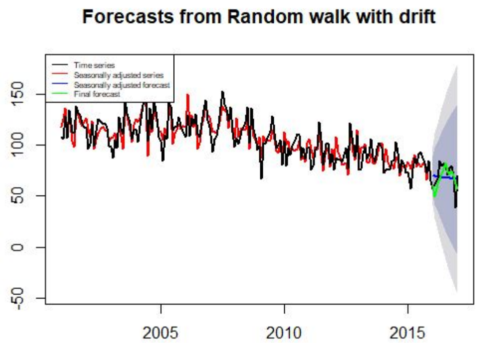


These graphs confirm what I previously said: 1) Decreasing trend 2) Seasonality Let's now construct the model with the STL decomposition.

Graphically, the results seem to be a bit better than the seasonal naive method since this time the negative trend is taken into account. To confirm this, one can have a look to the accuracy measures shown in the image below:

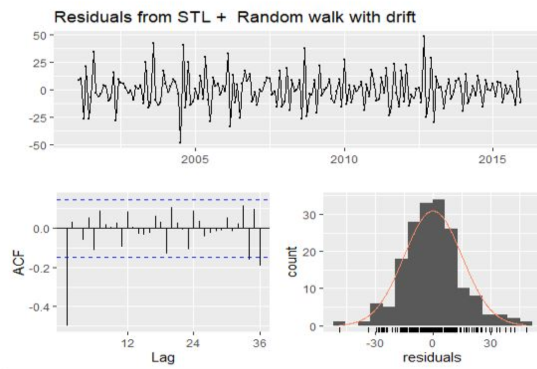
Indeed, the four accuracy measures for the test set are better than the ones of the seasonal naive method. We can also check the residual quality:

We can also check the residual quality: The residuals are not white noise since the p-value is lower than .05.

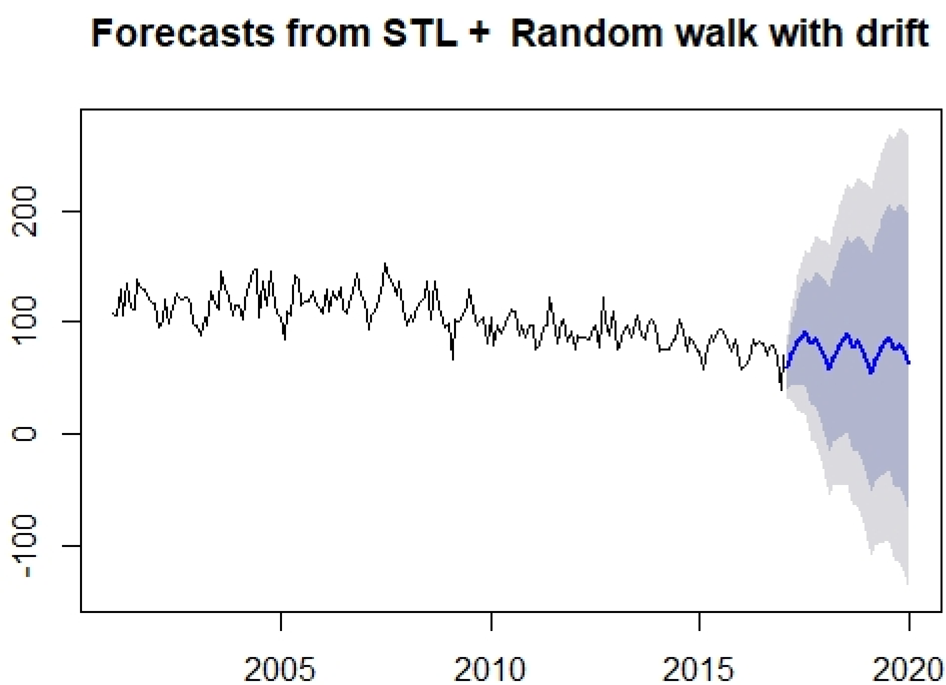


	RMSE	MAE	MAPE	MASE
Training set	14.978	11.4361	10.965	0.91707
Test set	10.381	8.1717	13.222	0.65530

```
> checkresiduals(fit2rwd)

        Ljung-Box test

data:  Residuals from STL +  Random walk with drift
Q* = 65.3, df = 23, p-value = 0.0000064

Model df: 1.   Total lags used: 24

Warning message:
In checkresiduals(fit2rwd) :
  The fitted degrees of freedom is based on the model used for the seasonally adjusted data.
```



Finally, the forecast till 2020 on the complete data set is the following one:



0.0.1.4 4) Generate forecasts using Holt-Winters' method. Check the residual diagnostics and the forecast accuracy.

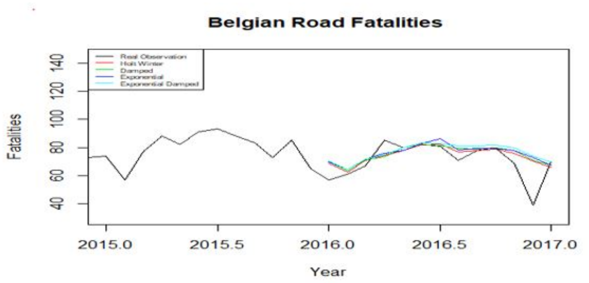
Holt-Winter's seasonal method is an extension of the Holt's method to capture seasonality. As said in the first point, it seems that the seasonality is decreasing with the years but not fully significant. That's why the multiplicativity of the seasonality will be used as well as additive seasonality will be tried. Therefore, I'll build six different Holt-Winter's model. One without any additional parameters, one with only the damped parameter, one with only the exponential parameter, and one with both the damped and the exponential parameter. One with additive non damped and last with additive seasonality and damped.

On the following graph, one can find the 4 estimations of the 6 models build:

By looking at the graph(Graph1), it seems that the Holt-Winter's model seems to be the best one. To confirm this, let's have a look to the accuracy matrix:

By Looking at Accuracy of Test as well Holt Winter Seems to be the best so hence I will check residuals for this model only.

Again the p-value is lower than .05, therefore I cannot consider that the residuals are white noise.

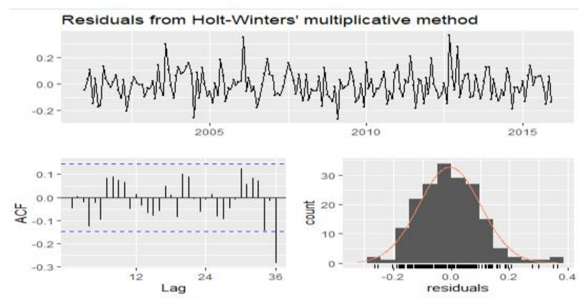


	RMSE	MAE	MAPE	MASE
Train_HW	11.308	8.8573	8.5300	0.71027
Train_HW_Damped	11.228	8.9796	8.6343	0.72008
Train_HW_Exp	11.797	9.2816	8.9257	0.74430
Train_HW_Exp_Damp	11.264	8.9001	8.6238	0.71371
Train_Additive_NonDamped	11.205	8.8463	8.4863	0.70939
Train_Additive_Damped	11.266	8.9761	8.6617	0.71980

	RMSE	MAE	MAPE	MASE
Test_HW	10.187	6.1849	11.638	0.49598
Test_HW_Damped	10.711	6.5078	12.219	0.52187
Test_HW_Exp	11.024	7.0911	13.143	0.56864
Test_HW_Exp_Damp	11.594	7.3395	13.613	0.58856
Train_Additive_NonDamped	10.397	6.9391	12.549	0.55646
Train_Additive_Damped	10.856	7.2958	13.158	0.58506

```
        Ljung-Box test

data:  Residuals from Holt-Winters' multiplicative method
Q* = 20.6, df = 8, p-value = 0.0083

Model df: 16.   Total lags used: 24
```

Holt Winter



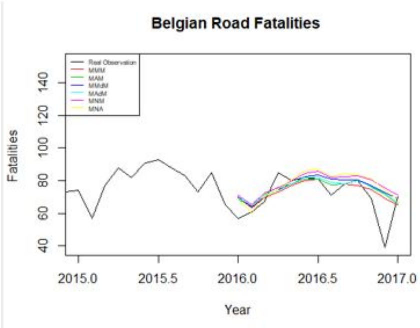
0.0.1.5 5) Generate forecasts using ETS. First select the appropriate model(s) yourself and discuss their performance. Compare these models with the results of the automated ETS procedure. Check the residual diagnostics and the forecast accuracy for the various ETS models you've considered.

From the precious Holt-Winter's , since the trend look to be really constant,the damped parameter doesn't seem to be useful here. Best ETS model will not need a damped parameter. Since I observe that the seasonality is not yet confirmed. So I would also like to check whether seasonality multiplicative and additive exists for this data so hence I will try with no seasonality as well.

So the below Models will be Tried:

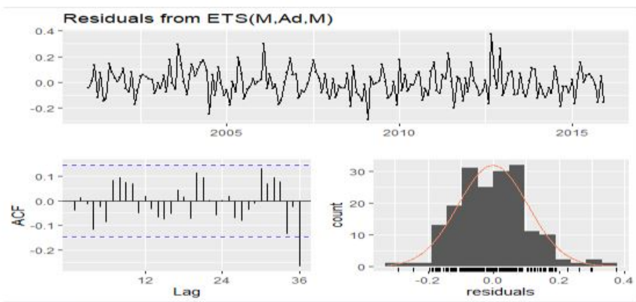
1) MMM (Trend Multiplicative Seasonality Multiplicative)
2) MAM (Trend Additive Seasonality Multiplicative)
3) MMdM (Tred Multiplicative Damped Seasonality Multiplicative)
4) MAdM (Trend Additive Damped Seasonality Multiplicative)
5) MNM (No Trend Seasonality Multiplicative)
6) MNA (No Trend Seasonality Additive)

If we see the Accuracies on the Test set MAM seems to be the best model. But if we check residuals then its white noise. Hence MADM the second best in terms of accuracies was selected and checked for residuals and p value of 0.0059 which < 0.05 and also there is information if we look at the residuals plot. So Additive Damped with Seasonality multiplicative proves to be better.



Test Accuracy

	RMSE	MAE	MAPE	MASE
MMM	11.322	8.9064	8.5561	0.71421
MAM	11.247	8.7954	8.4150	0.70532
MMdM	11.332	8.8787	8.5756	0.71199
MAdM	11.291	8.9003	8.5619	0.71372
MNM	11.296	8.8404	8.5950	0.70892
MNA	11.321	8.9155	8.6593	0.71494



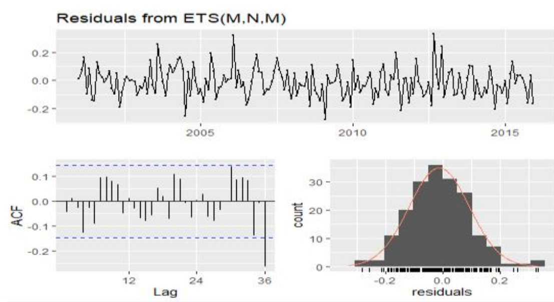

```
Ljung-Box test

data:  Residuals from ETS(M,Ad,M)
Q* = 19.9, df = 7, p-value = 0.0059

Model df: 17.   Total lags used: 24
```

Now looking at Auto ETs we get MNM to be the best. So checking the Accuracies of Auto ETS and residuals from that to compare with MAdM.However, the AUTO.ETS function just want to minimize the AIC measure.

	RMSE	MAE	MAPE	MASE
Training set	11.296	8.8404	8.595	0.70892
Test set	12.340	8.2525	15.063	0.66178

AUTO ETS

```
Ljung-Box test

data:  Residuals from ETS(M,N,M)
Q* = 21.1, df = 10, p-value = 0.02

Model df: 14.   Total lags used: 24
```



If we compare MAdM stands to be best among all ETS tried.

0.0.1.6 6) Generate forecasts using ARIMA. First select the appropriate model(s) yourself and discuss their performance. Compare these models with the results of the auto.arima procedure. Check the residual diagnostics and the forecast accuracy for the various ARIMA models you've considered.

Before using ARIMA we need stationary data. As a stationary time series is one whose properties do not depend on the time at which the series is observed. In our case, due to the seasonality and the trends, there is a seasonality of the data. It's confirmed with the ACF graph of the first point.

It's not really clear at this point if I have to do another difference or not. So for trial purpose i just tried making a couple of difference and the spikes and graphs resulted in same. So i took the First one for further analysis.

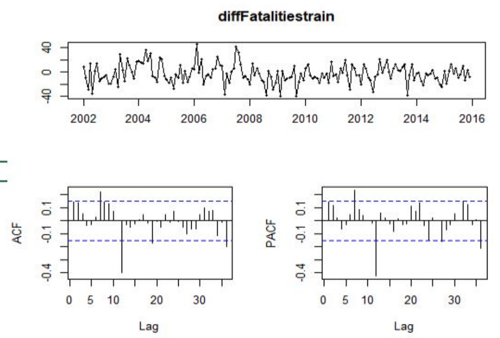
The significant value at lag 12 in the ACF suggest a seasonal MA(1) term. In the non seasonal lags, there are around 5 high peaks that are not significant in the PACF. That's why I'll start with the ARIMA (5,0,0)(0,1,1)[12]. According to the AUTO.ARIMA function the ARIMA that minimizes the AIC values is the ARIMA(0,1,3)(2,0,0).

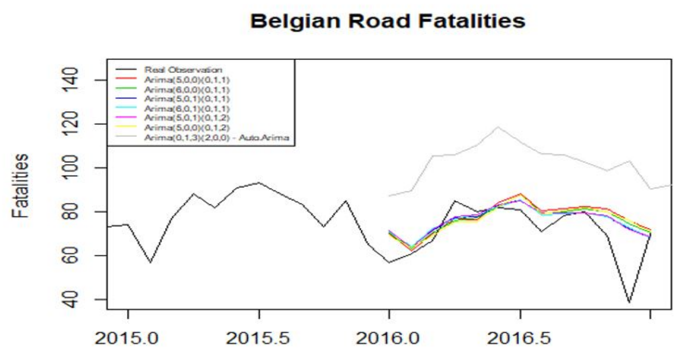
Finally I felt its better to try different models near to both at the end to select a best model.

In the plot below we can clearly see that The Autoarima is no where close to the real observation. So Finding out ACcuracies for all the models to finally select the best one.

The best one in term of RMSE is the Arima(5,0,1)(0,1,2)[12]. Let's have a look to the residuals of this model.

For this ARIMA model, as for all the other models tested, the residuals aren't white noise since the p§value of the LjungBox test is higher than .05.



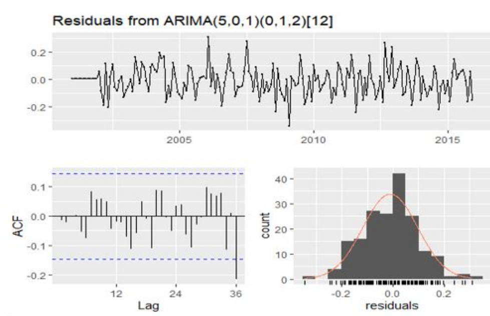


Train

	RMSE	MAE	MAPE	MASE
Arima(5,0,0)(0,1,1)	12.154	9.2845	9.0089	0.74453
Arima(6,0,0)(0,1,1)	12.062	9.1703	8.8885	0.73537
Arima(5,0,1)(0,1,1)	11.457	8.7871	8.5080	0.70465
Arima(6,0,1)(0,1,1)	11.434	8.8056	8.5325	0.70613
Arima(5,0,1)(0,1,2)	11.374	8.7185	8.4388	0.69914
Arima(5,0,0)(0,1,2)	12.102	9.2707	8.9857	0.74343
Arima(0,1,3)(2,0,0) - Auto.Arima	4.281	3.1861	3.8448	0.61761

Test

	RMSE	MAE	MAPE	MASE
Arima(5,0,0)(0,1,1)	12.247	8.0135	14.646	0.64261
Arima(6,0,0)(0,1,1)	11.747	7.3910	13.707	0.59269
Arima(5,0,1)(0,1,1)	10.991	6.9068	12.940	0.55386
Arima(6,0,1)(0,1,1)	11.032	6.8946	12.949	0.55289
Arima(5,0,1)(0,1,2)	10.930	6.9074	12.905	0.55391
Arima(5,0,0)(0,1,2)	12.067	7.7215	14.230	0.61919
Arima(3,0,0)(0,1,1) - Auto.Arima	33.671	31.9164	50.107	6.18688

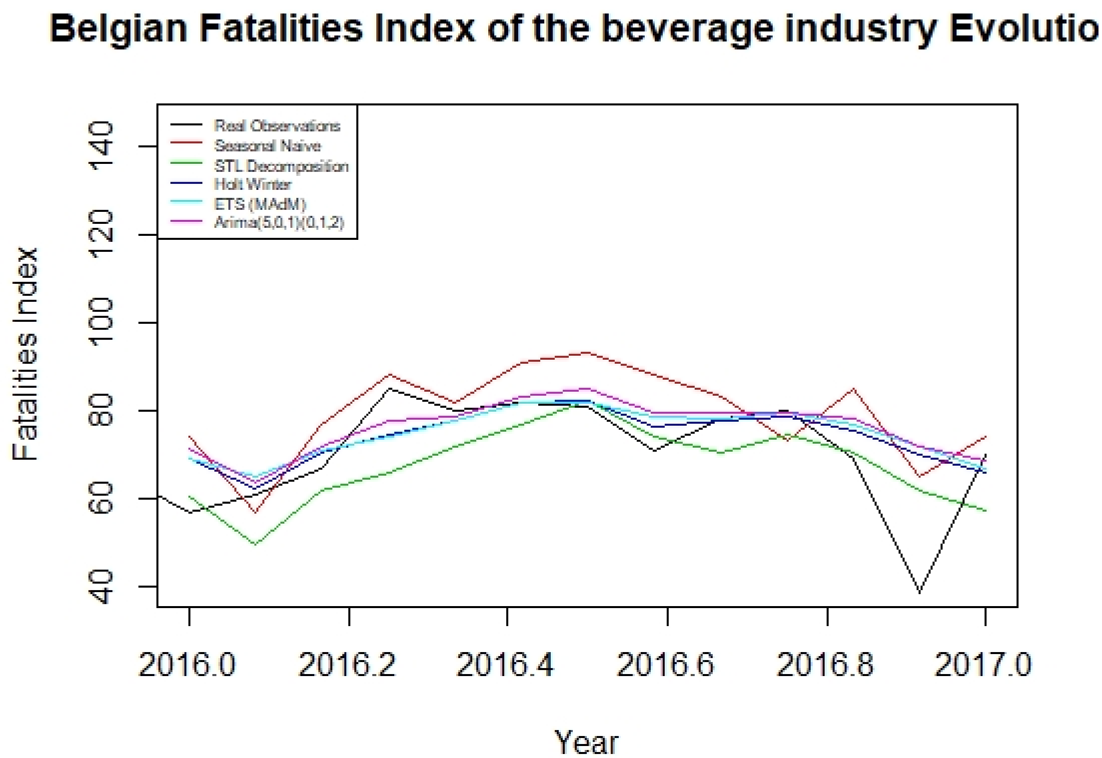



```
        Ljung-Box test

data:  Residuals from ARIMA(5,0,1)(0,1,2)[12]
Q* = 16.4, df = 16, p-value = 0.43

Model df: 8.   Total lags used: 24
```

0.0.1.7 7) Generate forecasts using ARIMA. First select the appropriate model(s) yourself and discuss their performance. Compare these models with the results of the auto.arima procedure. Check the residual diagnostics and the forecast accuracy for the various ARIMA models you've considered.

Several model were tested and it's now time to select the one I'll use to do my final forecast. To select it I'll first plot all the selected models and I will compare their accuracy measures. In this residuals wont be tested again since for all the models its already tested for this datab above, the residuals were never considered as white noise.

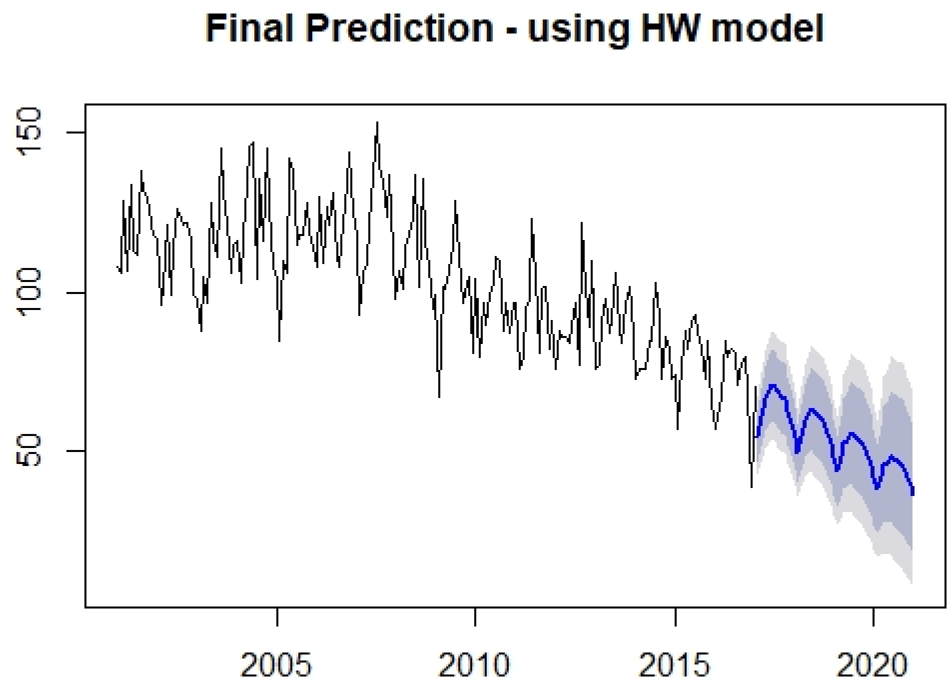
So all the Models that were selected best among others were plotted now and in final Selection Seasonal Naive, STL Decomposition,Holt Winter, ETS(MAdM),Arima(5,0,1)(0,1,2) and by the graph below we can see that Holt Winter follows the same pattern as Test set.



Now Testing the Accuracies on Test as well Holt Winter Proves to be the best performing forecasts.

	RMSE	MAE	MAPE	MASE
SNaive	12.260	10.1538	16.754	0.81425
STL Decomp.	10.381	8.1717	13.222	0.65530
Holt Winter	10.187	6.1849	11.638	0.49598
ETS	10.858	6.6798	12.562	0.53566
Arima	10.930	6.9074	12.905	0.55391

Therfore Final Forecast using the Holt Winter Method :



Displaying below is the Forecast until 2020 and it would be really helpful to check till June 2019 with real data to see how close our forecast is to the real road fatalities in Belgium.

	Point Forecast		
Jan-18	54.38	Jan-19	48.297
Feb-18	49.23	Feb-19	43.671
Mar-18	54.353	Mar-19	48.157
Apr-18	59.412	Apr-19	52.575
May-18	60.628	May-19	53.583
Jun-18	63.379	Jun-19	55.942
Jul-18	62.736	Jul-19	55.302
Aug-18	61.385	Aug-19	54.039
Sep-18	59.968	Sep-19	52.719
Oct-18	59.385	Oct-19	52.133
Nov-18	55.824	Nov-19	48.937
Dec-18	52.523	Dec-19	45.976
	Jan-20	42.214	
	Feb-20	38.112	
	Mar-20	41.961	
	Apr-20	45.737	
	May-20	46.538	
	Jun-20	48.506	
	Jul-20	47.868	
	Aug-20	46.692	
	Sep-20	45.47	
	Oct-20	44.882	
	Nov-20	42.05	
	Dec-20	39.429	

0.0.2 Exercise 2 : Forecasting of Renewable Energy Prices in E.U

This data contains the yearly evolution of the shows the yearly gross inland consumption of renewable energies in the European Union, in thousand tonnes of oil equivalent (TOE) from 1990 up to 2016.

The dataset has been separated into two datasets: the train set contains the data from 1990 to 2010, while the test set contains the information from 2011 and 2016.

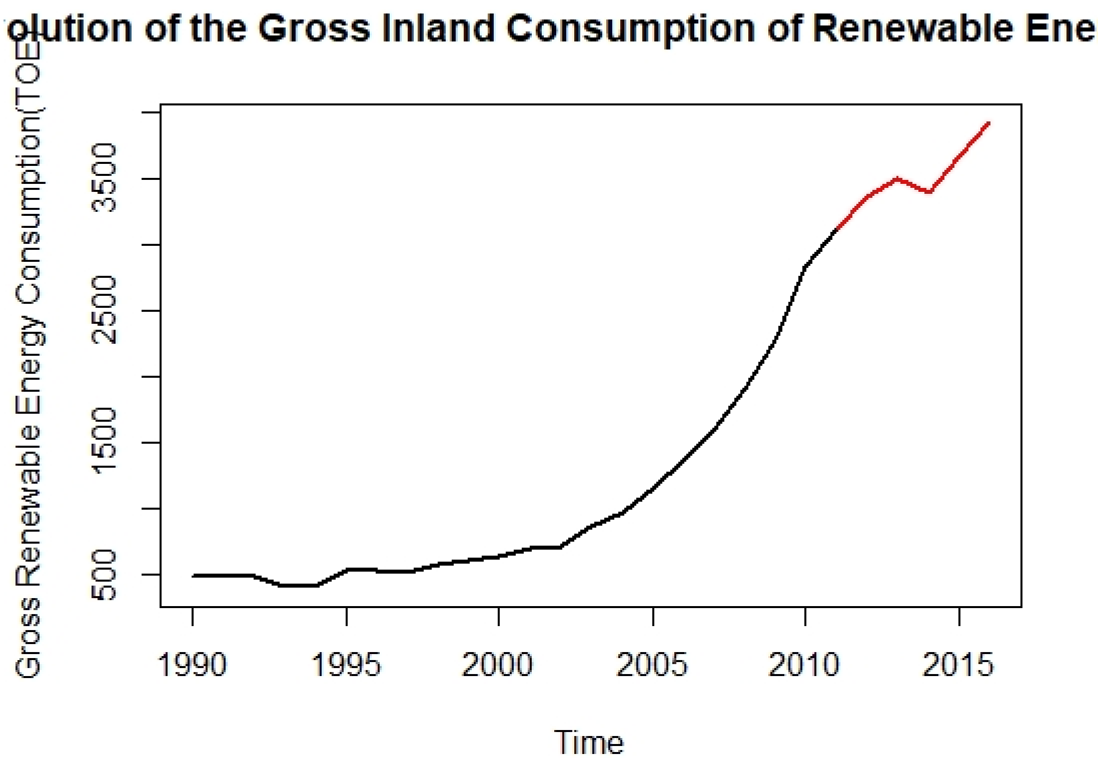
Before starting the analysis in general I would like to have a null hypothesis that the data will definetely be increasing trend with so much increase in demand of renewables energy consumption.

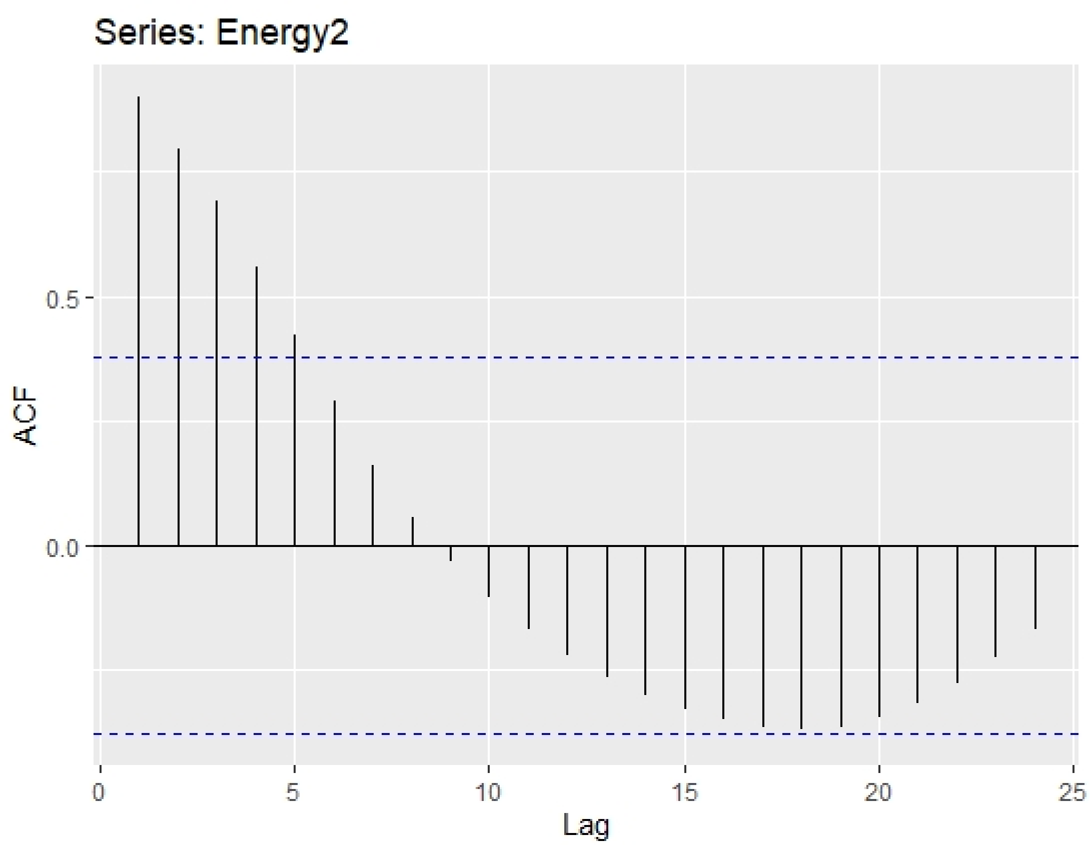
0.0.2.1 1) Explore the data using relevant graphs

Starting the forecasting analysis by plotting the timeseries plot.

As expected, we can see a positive trend that is not really constant through the years. The trend was quite constant until around 2012 and the increased a lot from 2000 to around 2012. However, from 2014 to 2016, the slope seems to be constant again.

The ACF graph of the data confirms the trend:

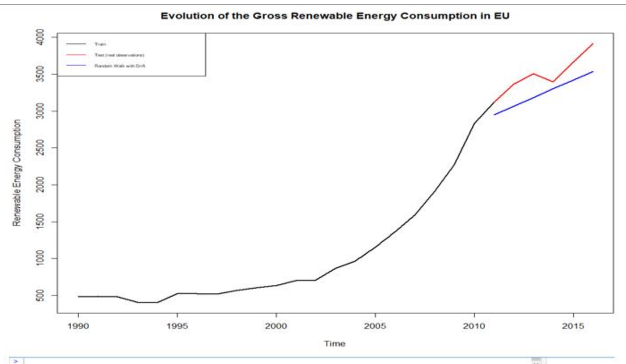

olution of the Gross Inland Consumption of Renewable Ener
Gross Renewable Energy Consumption(TOE
500
1500
2500
3500
1990
1995
2000
2005
2010
2015
Time


Series: Energy2
ACF
0.0
0.5
0
5
10
15
20
25
Lag

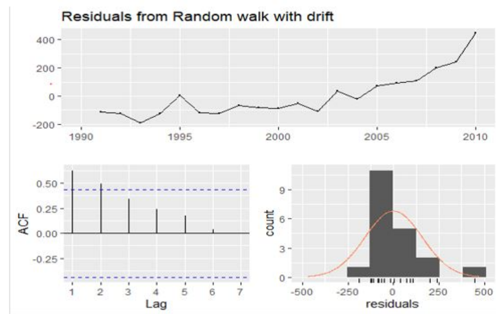
0.0.2.2 2) Create forecasts using the most appropriate naive method. Check the residual diagnostics and the forecast accuracy.

The appropriate naive method for this dataset would be the random walk drift method. It's a variation of the naive method. It allows the forecast to increase or decrease over time. In our case, since we observe a clear positive trend, it's the most appropriate naive method.

On the 1st graph below, one can see the random walk drift forecast. The results seems to be quite good for this kind of simple model. The accuracy measures of this model are also displayed below:



	RMSE	MAE	MAPE	MASE
Training set	152.82	120.46	14.8512	0.95131
Test set	267.98	250.62	7.0841	1.97917




```
	Ljung-Box test

data:  Residuals from Random walk with drift
Q* = 19.6, df = 3, p-value = 0.0002

Model df: 1.   Total lags used: 4
```

Moreover, we can check the residuals above:

The residuals are not white noise since the p-value of the Ljungbox test is lower they the threshold .05.

0.0.2.3 3) Generate forecasts using the relevant exponential smoothing methods. Check the residual diagnostics and the forecast accuracy.

The relevant Exponential Smoothing Methods for this data would be the Holt's method. As, the Holt's method is an extension of the SES method that allow the forecasting of data with a trend. Since in this case there is a clear increasing trend, the Holt's method is the one to use. four different Holt's models will be tried.

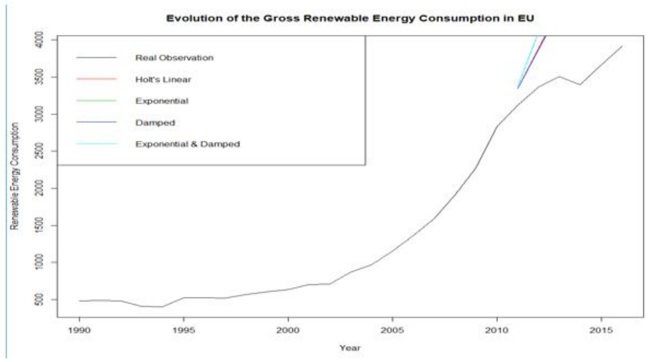
1) without any additional parameters
2) with only the dampedparameter
3) with only the exponential parameter
4) with both the damped and the exponential parameter.

In the below, The 4 estimations of the 4 models build are displayed:

By Looking at the First graph itself its very clear that the Forecasts are not anywhere close to real test values hence there is going to be a significant decrease in accuracy measures as well as there is a chance of white noise.

But in the second graph the accuracy measures have been calculated and damped proved to be the best among all.

So calculating Residuals only to the best performing among the 4 that is damped we can see that The residuals can now be considered as white noise since the p-value of the LjunBox test is higher than .05. We can also see that in the ACF graph since all the peaks are within the confidence level (between the two blue lines). So there is no information in the Forecast.



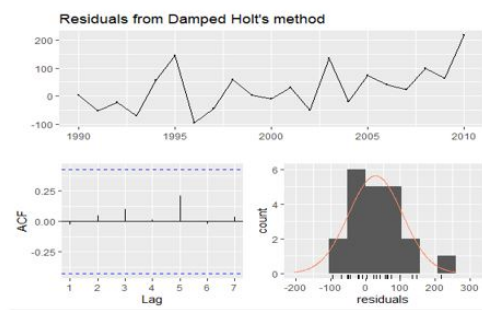
Train

	RMSE	MAE	MAPE	MASE
Holt Linear	78.327	58.315	7.3819	0.46051
Exp.	58.439	39.127	5.1942	0.30899
Damped	80.863	62.472	8.0381	0.49334
Exp & Damped	64.624	46.478	6.3546	0.36704

Test

	RMSE	MAE	MAPE	MASE
Holt Linear	1396.5	1208.1	33.487	9.5403
Exp.	2870.6	2323.9	63.708	18.3522
Damped	1298.2	1131.1	31.396	8.9326
Exp & Damped	2686.5	2197.1	60.322	17.3506

```
        Ljung-Box test

data:  Residuals from Damped Holt's method
Q* = 1.75, df = 3, p-value = 0.63

Model df: 5.   Total lags used: 8
```



0.0.2.4 4) Generate forecasts using ETS. First select the appropriate model(s) yourself and discuss their performance. Compare these models with the results of the automated ETS procedure. Check the residual diagnostics and the forecast accuracy for the various ETS models you've considered.

In this case we face non seasonal data. That's why the third term of all my ETS models will be equal to N.

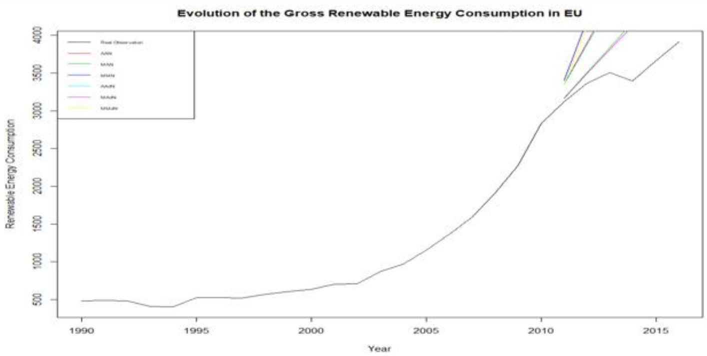
6 models were tried: . AAN Damped and non-damped . MAN Damped and non-damped . MMN Damped and non-damped

Also tried the AMN model but since it is an unstable model it will not be taken further. Here are the result I got with these 6 models.

As you can see below none of the forecasts are close to test. So i already assume that this is not information but white noise but anyways trying to find out the accuracies of all models and choose the best to perform the Ljung box test.

As you can see in the below graph MAdN seems to be best in terms of accuracy parameters hence Finding the residuals on that:

We can clearly see that there is only white noise both with graph as well as p-value > 0.05.

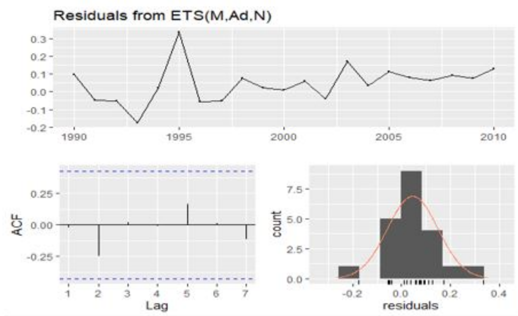


Train

	RMSE	MAE	MAPE	MASE
AAN	79.116	61.628	8.0734	0.48668
MAN	105.114	76.039	7.8434	0.60048
MMN	74.548	64.335	8.2552	0.50806
AAdN	80.863	62.472	8.0381	0.49334
MAdN	106.151	76.854	7.9327	0.60692
MMdN	78.877	67.126	8.3727	0.53010

Test

	RMSE	MAE	MAPE	MASE
AAN	1409.31	1220.55	33.837	9.6387
MAN	621.76	512.51	14.155	4.0473
MMN	3243.62	2617.32	71.711	20.6690
AAdN	1298.24	1131.13	31.396	8.9326
MAdN	563.02	465.63	12.879	3.6771
MMdN	2802.95	2276.46	62.435	17.9773

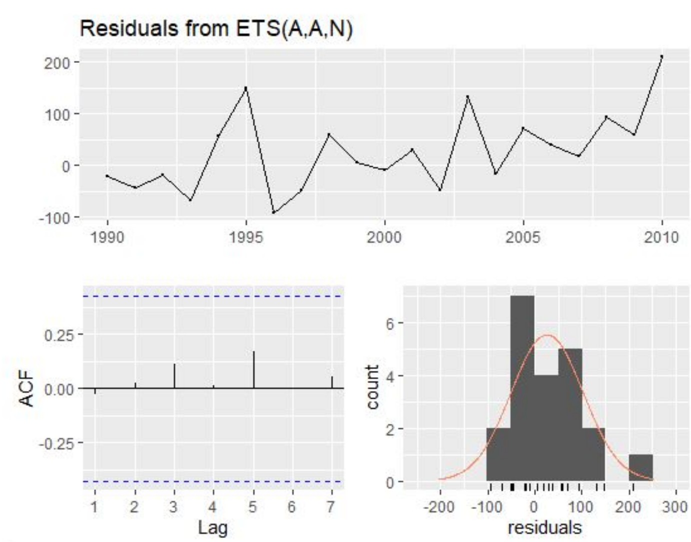



```
        Ljung-Box test

data:  Residuals from ETS(M,Ad,N)
Q* = 3.59, df = 3, p-value = 0.31

Model df: 5.   Total lags used: 8
```

By Trying the AUTO ETS Procedure we get AAN model in terms of lower AICC. Lets see if Auto ETS gives us any information.

But if we see the plots and p-value its again whitenoise.

```
> checkresiduals(auto_ets)

        Ljung-Box test

data:  Residuals from ETS(A,A,N)
Q* = 1.34, df = 3, p-value = 0.72

Model df: 4    Total lags used: 7
```

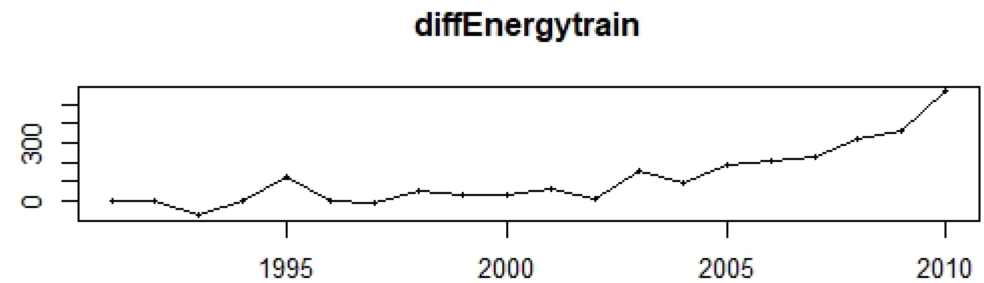


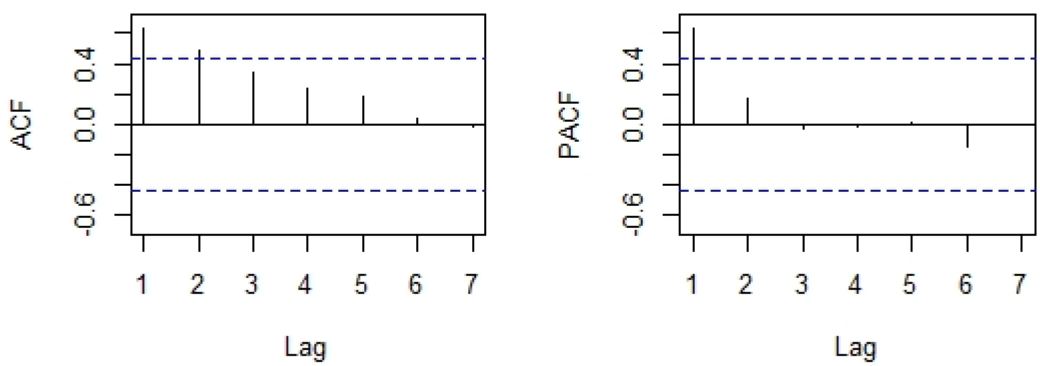
Though there is white noise but in terms of Accuracy measures MAdN can be concluded to be the best performing ETS model based on accuracy measures on test.

0.0.2.5 5) Generate forecasts using ARIMA. First select the appropriate model(s) yourself and discuss their performance. Compare these models with the results of the auto.arima

procedure. Check the residual diagnostics and the forecast accuracy for the various ARIMA models you've considered.

Before starting the ARIMA we need stationary data. In our case lets check if the data is stationery The seasonally differences data are shown in the following graph:





The graph shown above is after 1 differentiation but Even before one Differentiation the data looked stationery so we can take minimum to be no differentiation. After just one differentiation the data seems to be stationary.

As there is exponetial decay in the ACF graph and one significant peak in the PACF graph, We can say that this could be AR(1). And in PACF as there is 1 significant spike at lag1 we could also say that its MA(1). Hence I will start my Tests with ARIMA(1,0,1) and I will try around many models related to and around this one.

According to the AUTO.ARIMA function the ARIMA that minimizes the AIC values is the ARIMA(0,2,0)

The ARIMA(1,0,1) failed the white noise test because the residuals were not white noise and in ARIMA modelling we need the residuals to be white noise.

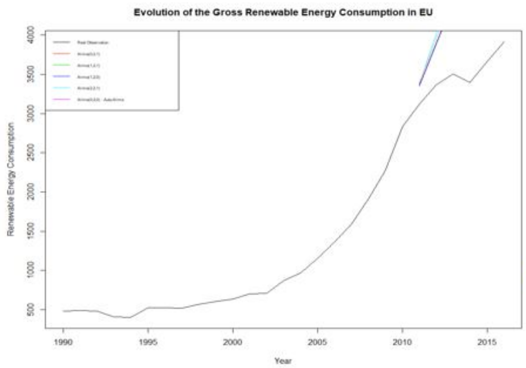
Hence I ended up trying the below models.

Finally, I tested the following ARIMA models: . Arima(0,2,1) . Arima(1,2,1) . Arima(1,2,0) . Arima(2,2,1) . Arima(0,2,0) - Auto.Arima

From the First graph ARIMA(1,2,0) seems to be the best, To confirm that the Train and Test ACcuracies are also plotted in Graph 2 below and we can see on Test ARIMA(1,2,0) performs best.

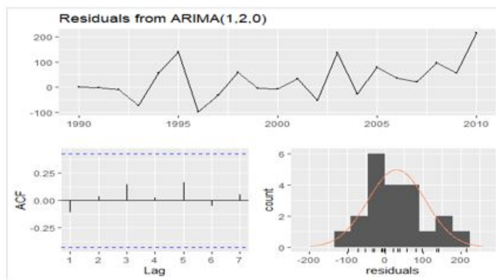
Also after model is finalized the White noise is checked again and the plot 3 and 4 clearly confirms that it is white noise for the selected model ARIMA(1,2,0) and autoARIMA(0,2,0) gives white noise.

That's why to do the final forecast I decided to keep only the model that give the best accuracy scores. This model was the Arima(1,2,0).



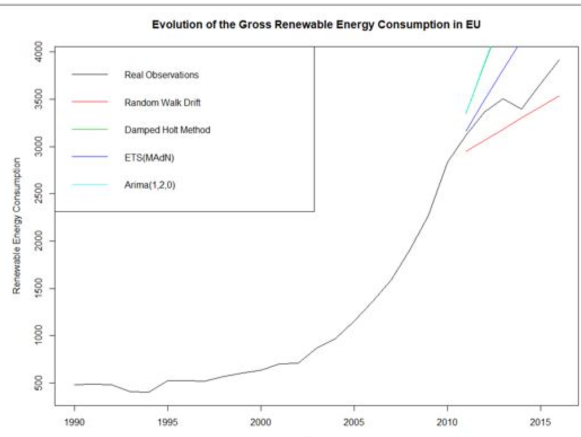

```
> #Train
> acc <- rbind(accTrain.m1,accTrain.m2,accTrain.m3,accTrain.m4,accTrain.auto)
> rownames(acc) <- c("Arima(0,2,1)","Arima(1,2,1)","Arima(1,2,0)",
+                    "Arima(2,2,1)","Arima(0,2,0) - Auto.Arima")
> acc
                              RMSE    MAE   MAPE    MASE
Arima(0,2,1)                79.647 58.914 7.3345 0.46525
Arima(1,2,1)                79.175 58.861 7.4018 0.46482
Arima(1,2,0)                79.282 58.679 7.3086 0.46339
Arima(2,2,1)                73.386 54.402 7.4111 0.42961
Arima(0,2,0) - Auto.Arima   80.601 59.507 7.5832 0.46993
> #Test
> acc <- rbind(accTest.m1,accTest.m2,accTest.m3,accTest.m4,accTest.auto)
> rownames(acc) <- c("Arima(0,2,1)","Arima(1,2,1)","Arima(1,2,0)",
+                    "Arima(2,2,1)","Arima(0,2,0) - Auto.Arima")
> acc
                              RMSE    MAE   MAPE    MASE
Arima(0,2,1)                1383.2 1202.7 33.365  9.4977
Arima(1,2,1)                1382.9 1200.5 33.295  9.4804
Arima(1,2,0)                1365.9 1185.7 32.886  9.3638
Arima(2,2,1)                1888.8 1607.5 44.413 12.6941
Arima(0,2,0) - Auto.Arima   1505.8 1313.3 36.443 10.3710
> |
```

```
        Ljung-Box test

data:  Residuals from ARIMA(1,2,0)
Q* = 0.929, df = 3, p-value = 0.82

Model df: 1.   Total lags used: 4
```



0.0.2.6 6) Compare the different models in terms of residual diagnostics, model fit, and forecast accuracy. Analyse your results and select your final model. Finally, generate out of sample forecasts up to 2020, based on the complete time series. Discuss your results.

Several model were tested and it's now time to select the one I'll use to do my final forecast. To select it I'll first plot all the selected models and I will compare their accuracy measures. In this case I will not really pay attention to the residuals again since for all the models were aleady tested for this data and HOlt and ETS failed with white noise. Random Walk and Arima have passed it and lets see which model fits best.

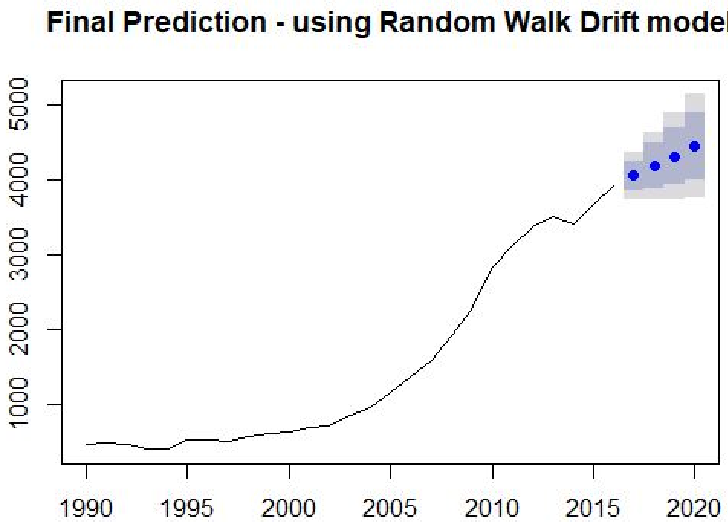


Accuracy on Test :

	RMSE	MAE	MAPE	MASE
RandomW Walk Drift	267.98	250.62	7.0841	1.9792
Damped Holt	1298.24	1131.13	31.3963	8.9326
ETS MAdN	563.02	465.63	12.8789	3.6771
Arima(1,2,0	1365.92	1185.73	32.8864	9.3638

From the Above graph and accuracy measures on test, Random Walk with Drift proves to be the best model.

Therefore, I can do the final forecast using the Random Walk with Drift model.



	Point Forecast
2017	4048.0
2018	4180.1
2019	4312.2
2020	4444.3

We can see that the increasing trend is expected to continue and the Forecast values are also displayed on the right till 2020.

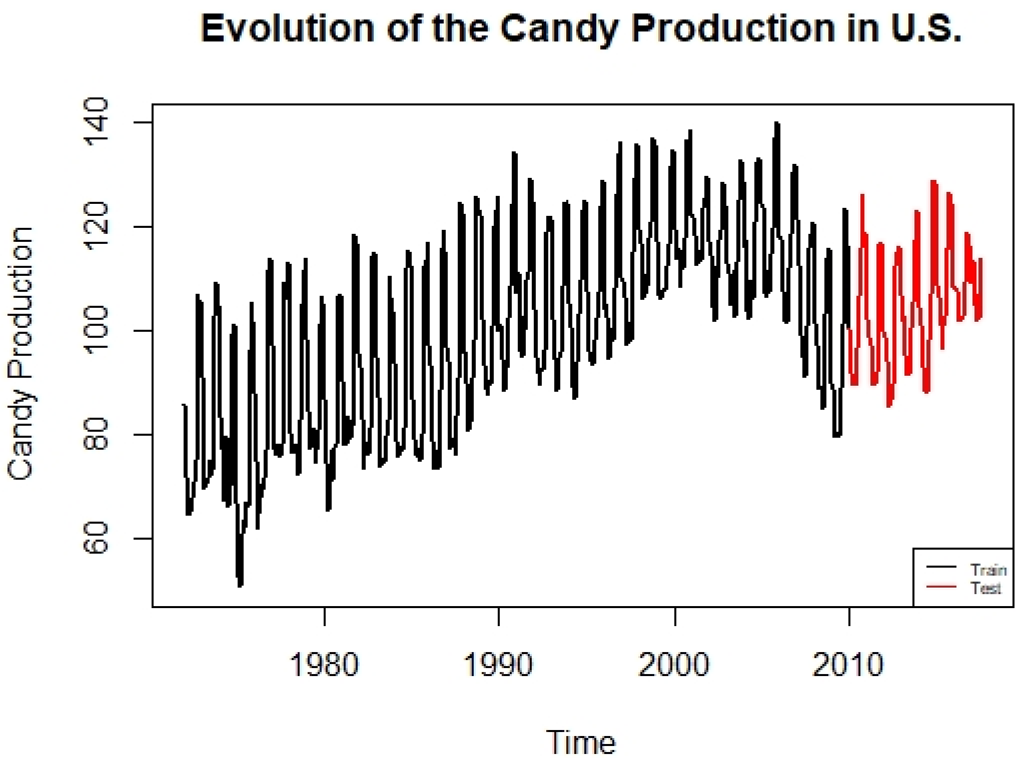
0.0.3 Exercise 3 : Forecasting Candy Production in U.S

This data deals with the Candy production of all relevant establishments located in the United States, regardless of their ownership, but not those located in U.S. territories. This dataset tracks industrial production every month from January 1972 to August 2017.

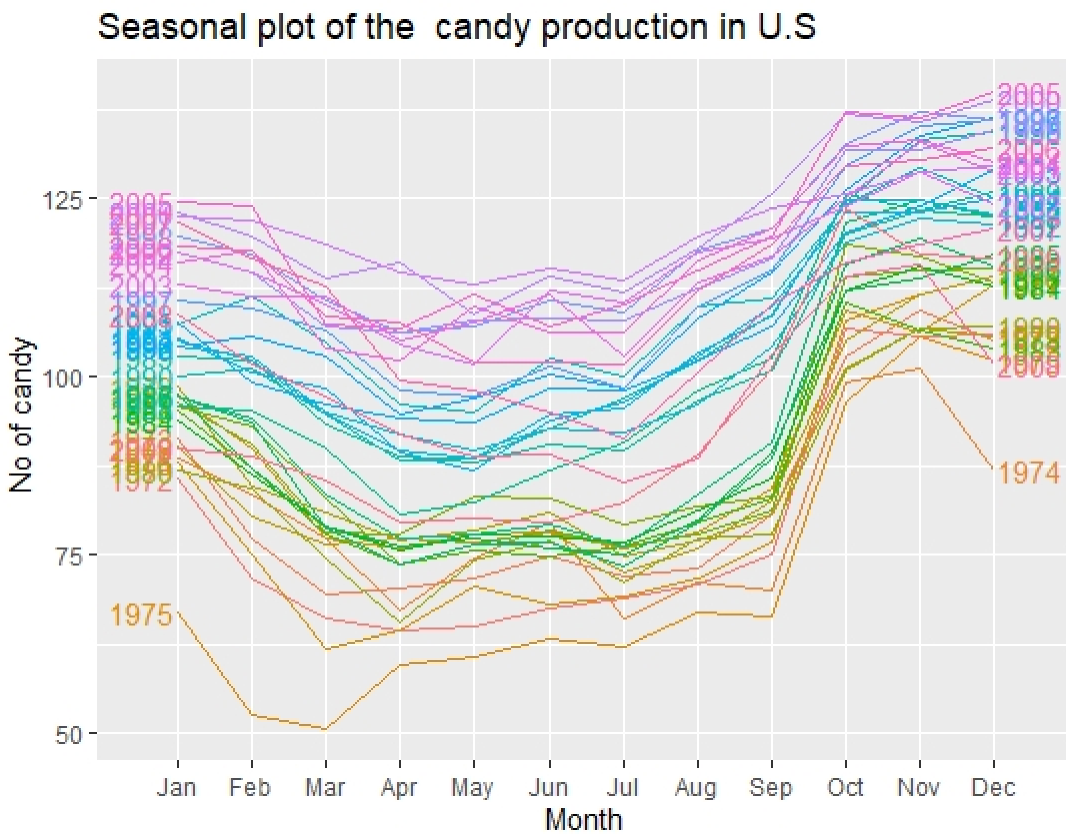
Data Source is Kaggle competition : https://www.kaggle.com/rtatman/us-candy-production-by-month References : https://www.kaggle.com/goldens/forecast-simple-exponential-arima/data

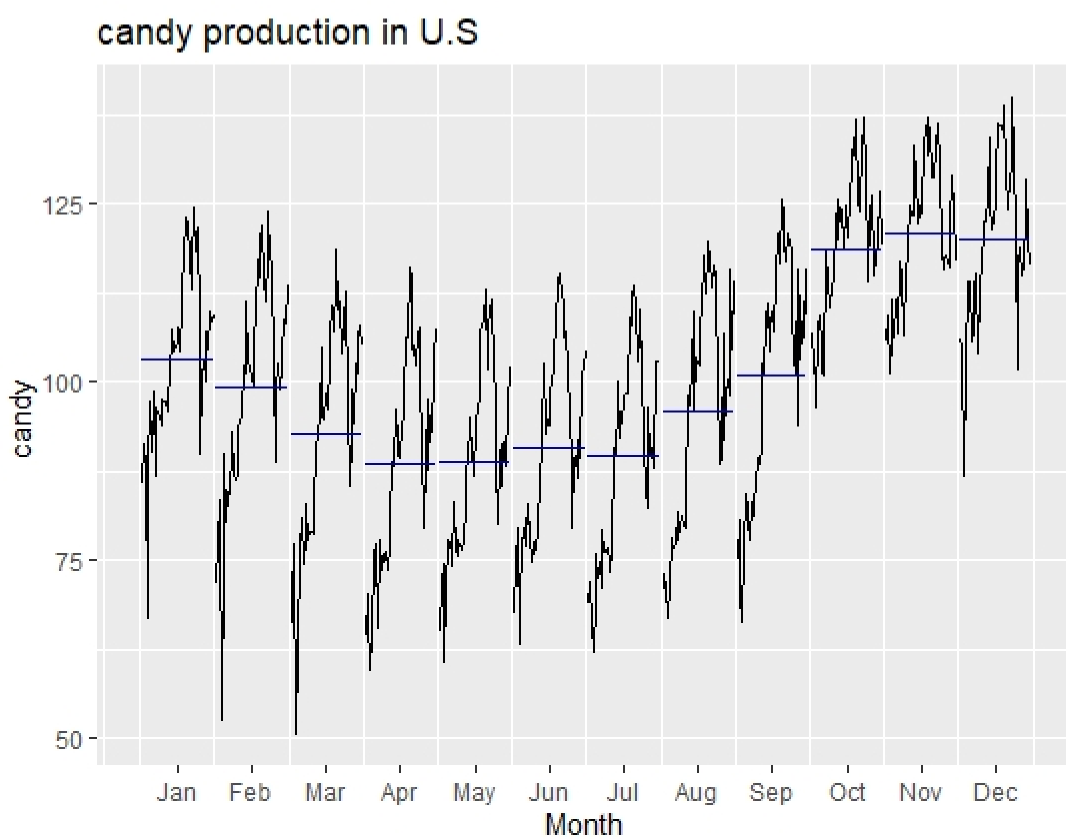
0.0.3.1 1) Explore the data using relevant graphs

First, I will explore the database using the appropriate graph. For time series data, the obvious graph to start an analysis is a classic time plot. This time plot will allow me to see if my previous hypothesis are confirm and will allow me to detect seasonality, cyclicity and a trend.



With this graph, we can observe a clear increasing trend. One can also see an increasing seasonality effect through the years. We can see that clearly in October to December months there is clear increase in production may be due to the fact that due to Christmas the demand is higher and in February March months a bit lower


Seasonal plot of the candy production in U.S
No of candy
125
100
75
50
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
1975
1974
1972

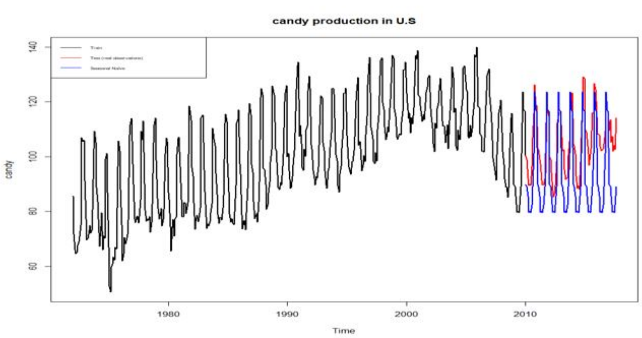

candy production in U.S
candy
125
100
75
50
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

The above two graphs above confirm my hypothesis. The number of Candy production increases during the October to December. It's at its highest during Christmas period(October & December). However, the number of candy Production seems to be particularly low in February,March and April. We can also see that the number of candy production increases year by year and on these two plots one can see the effect of the seasonality and increasing trend. This hints at multiplicative time series as effect of seasonality changes during time.

Let's now starting the forecasting part.

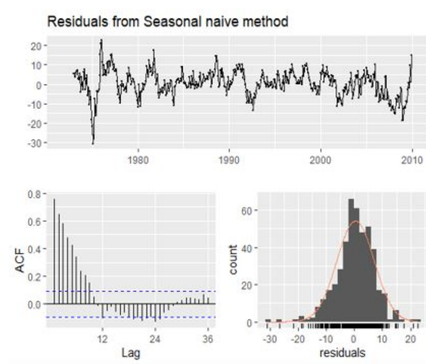
0.0.3.2 2) Forecasting using Seasonal Naive Method.

The seasonal naive method sets each forecast to be equal to the last observed value from the same season of the year.



	RMSE	MAE	MAPE	MASE
Training set	6.6505	5.0341	5.3609	1.0000
Test set	14.0034	12.0943	11.7122	2.4025

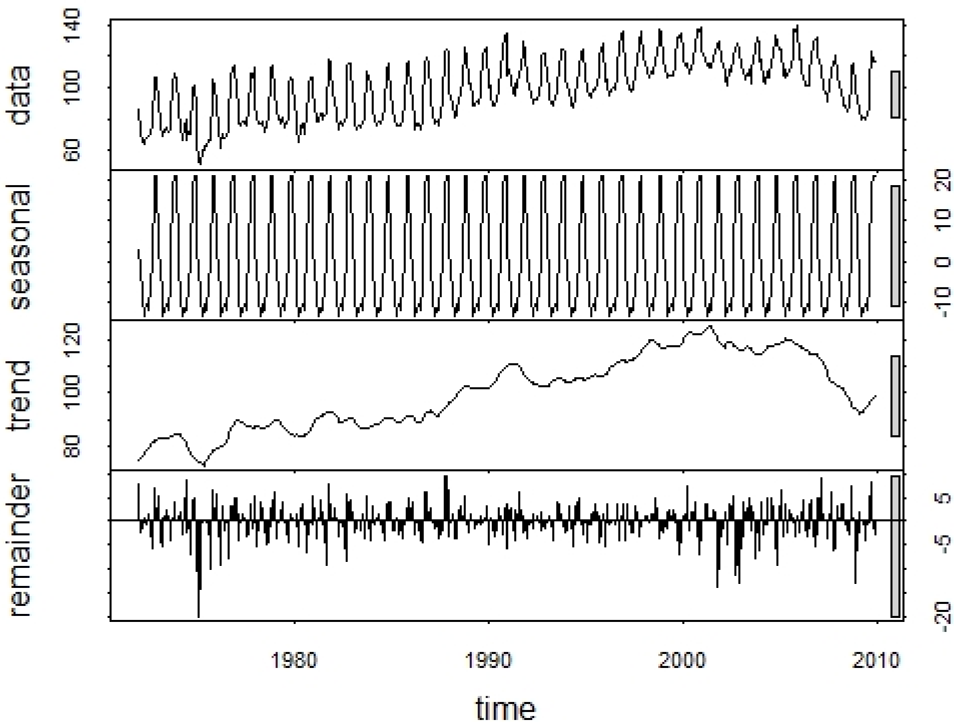
```
        Ljung-Box test

data:  Residuals from Seasonal naive method
Q* = 937, df = 24, p-value <2e-16

Model df: 0.   Total lags used: 24
```



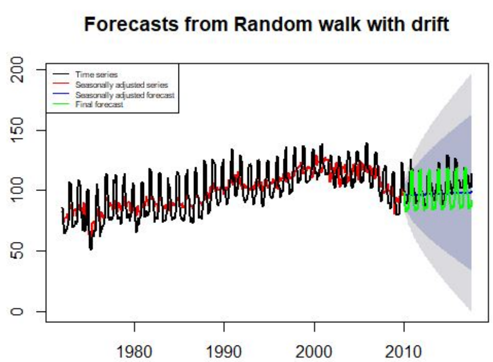
On the First graph above, one can see the seasonal naive forecast (blue line) and we can compare it to the real value of the observation. The accuracy of both the training and the test set can also be found here above. They will be useful to compare the performance of the several models constructed. Moreover, we can check the quality of the residuals: The residuals are not white noise since the p-value of the Ljungbox test is lower the the threshold .05.

0.0.3.3 2) Forecasting using STL decomposition.

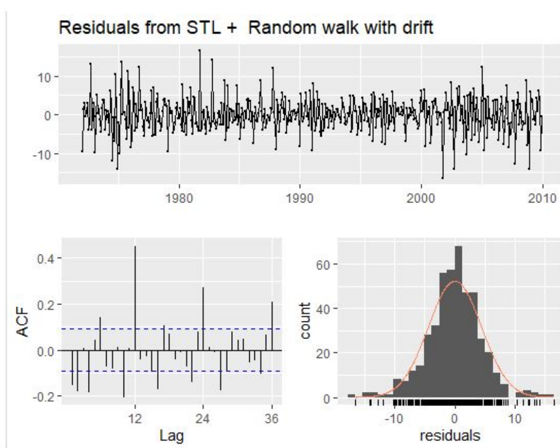
In this exercise, since during the exploratory phase, we discover that there was a positive and constant trend, I will use a random walk with drift model to forecast the seasonally adjusted time series. First, let's have a look to the STL decomposition graphs:



These graphs confirm what I previously said: . Positive trend . Seasonality Let's now construct the model with the STL decomposition.



	RMSE	MAE	MAPE	MASE
Training set	4.442	3.3380	3.4413	0.66307
Test set	10.601	8.6869	8.2772	1.72561




```
        Ljung-Box test

data:  Residuals from STL +  Random walk with drift
Q* = 253, df = 23, p-value <2e-16

Model df: 1.   Total lags used: 24
```

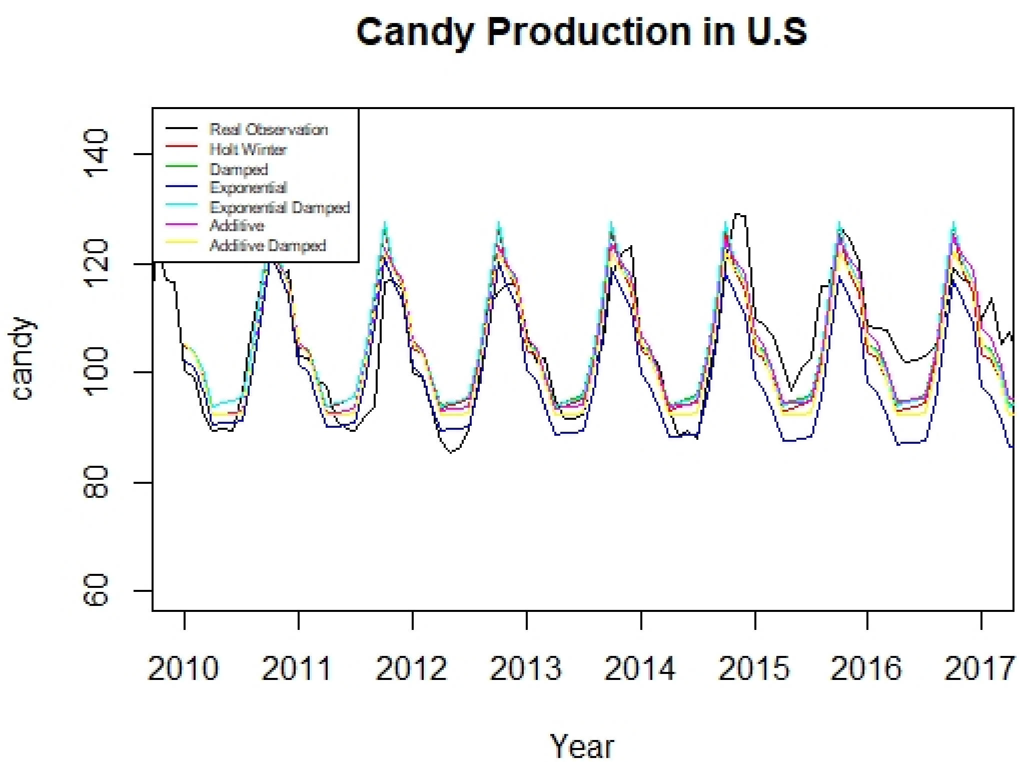
We can see in the First graph that the results are better than previous seasonal naive method. Also the accuracy measures are better. So while checking the residual quality, The residuals are not white noise since the p-value is lower than .05.

So this model can be taken into account later for final comparision.

0.0.3.4 3) Forecasting using Holt-Winters' Method.

Holt-Winter's seasonal method is an extension of the Holt's method to capture seasonality. As said in the first point, it seems that the seasonality is increasing with the years. That's why I will only use the multiplicativity of the seasonality. Therefore, I'll build four different Holt-Winter's model. One without any additional parameters, one with only the damped parameter, one with only the exponential parameter, and one with both the damped and the exponential parameter.

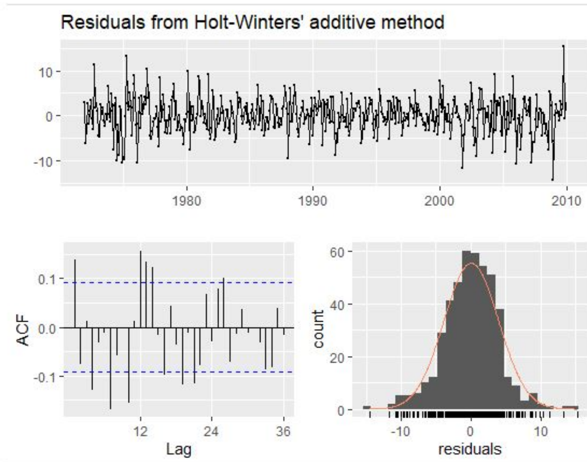
On the following graph, one can find the 4 estimations of the 4 models build:



By looking at the above graph one can say that exponetial and also additive non damped may be a bit closer to real observation but lets test the accuracy measures as its not fully evident.

	RMSE	MAE	MAPE	MASE
Test_HW	6.7120	5.5423	5.3395	1.10094
Test_Hw_Damped	6.3728	5.2634	5.1032	1.04555
Test_Hw_Exp	9.0858	6.9660	6.5134	1.38376
Test_Hw_Exp_Damp	6.4729	5.3408	5.1700	1.06092
Test_Additive_NonDamped	5.5507	4.6344	4.4673	0.92060
Test_Additive_Damped	6.2768	5.0109	4.7793	0.99539

```
        Ljung-Box test

data:  Residuals from Holt-Winters' additive method
Q* = 97.7, df = 8, p-value <2e-16

Model df: 16.   Total lags used: 24
```



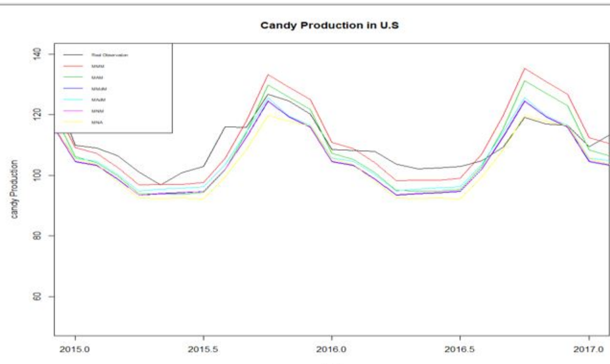
By looking at the accuracy measures the Additive Method with non damped is best in terms of accuracy measures and hence used that for residuals test and we found out that p value less than 0.05 and its not white noise and this can be used for final forecasts.

0.0.3.5 4) Forecasting using ETS Method.

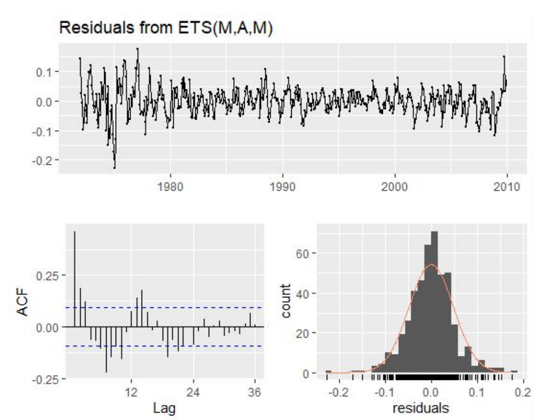
Since I observe that the seasonality is additive in the above process I want to try the following model (with and without the damped parameter): . AMM . MMM . MAM . MMdM . MAdM
. MNM . MNA and also Auto ETS.

The graphs are ploteed below with each model and we can see that all the models are performing consistently well and its hard to choose just by looking at the graph and hence we will look at the accuracy measures. And we can see that MAM performs the best in terms of accuracy.

So MAM is chosen as final model and checked for white noise and we can see that as the p value < 0.05 there is no white noise and there is information. Hence this can be used for forecasting.



	RMSE	MAE	MAPE	MASE
MMM	5.7667	4.4779	4.2501	0.88951
MAM	5.6711	4.4514	4.1942	0.88425
MMdM	6.2264	5.1460	4.9556	1.02223
MAdM	6.2241	5.1863	5.0533	1.03023
MNM	6.1826	5.1122	4.9254	1.01552
MNA	6.3023	4.9426	4.7100	0.98183



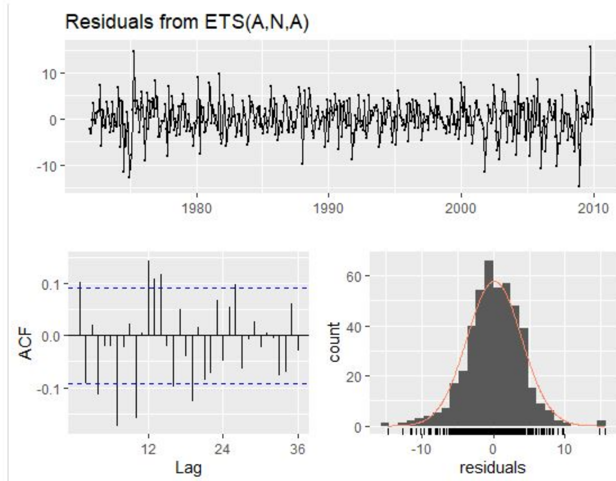

```
        Ljung-Box test

data:  Residuals from ETS(M,A,M)
Q* = 241, df = 8, p-value <2e-16

Model df: 16.   Total lags used: 24
```

Also tried the AUTO ETS Proceudre and found out that A.N.A was chosen from Auto ETS.

Checked the Forecast Accuracy on test as well as Residual test. We can see that the p value is less than 0.05 and there is no white noise also we can see the accuracy measures.

```
                  RMSE    MAE   MAPE    MASE
Training set 3.8203 2.9140 3.0259 0.57886
Test set     6.1776 4.9643 4.7500 0.98613
```

```
        Ljung-Box test

data:  Residuals from ETS(A,N,A)
Q* = 87.8, df = 10, p-value = 0.000000000000015

Model df: 14.   Total lags used: 24
```

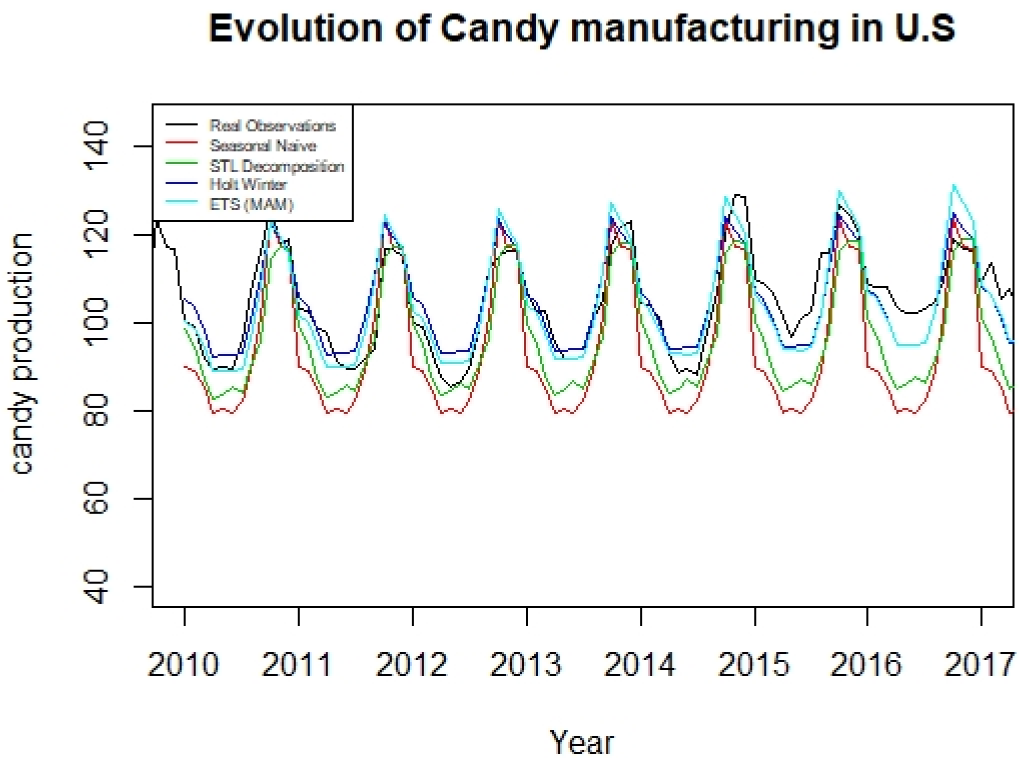


Looking at the Accuracy Measures we can conclude that M.A.M is the best among the ETS models for this data.

0.0.3.6 4) Comparison of the different models.

Several model were tested and it's now time to select the one I'll use to do my final forecast. To select it I'll first plot all the selected models and I will compare their accuracy measures. In this case I will not really pay attention to the residuals since for all the models tested for this data, the residuals were never considered as white noise.

Plotting all the best models tried till now.



By Looking at the graph Holt and ETS both look very good as per the real observation.

Now lets test the Accuracy measures

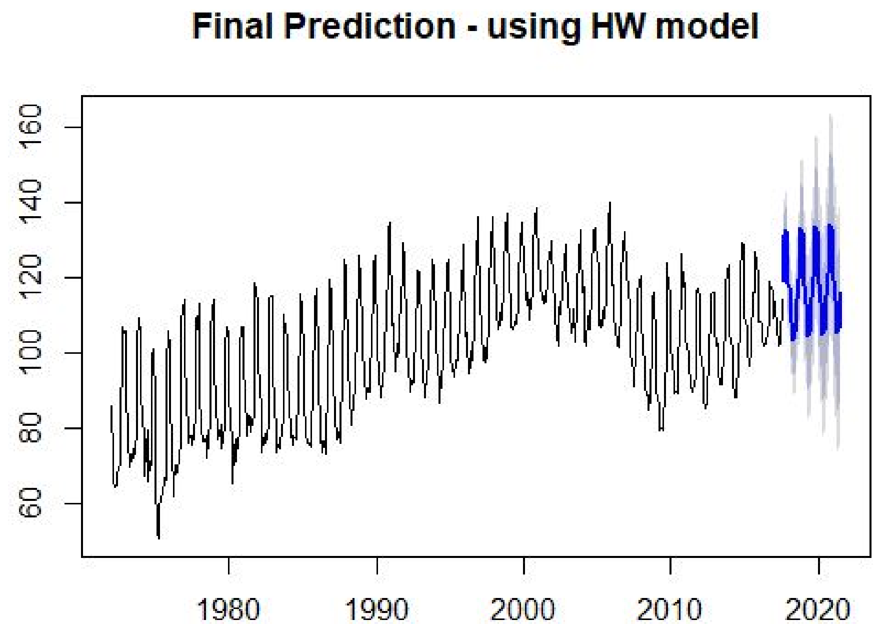
Test Accuracy

	RMSE	MAE	MAPE	MASE
SNaive	14.0034	12.0943	11.7122	2.40249
STL Decomp.	10.6010	8.6869	8.2772	1.72561
Holt Winter	5.5507	4.6344	4.4673	0.92060
ETS	5.6711	4.4514	4.1942	0.88425

We can see that Holt Winter performs the best in terms of accuracy eventhough ETS was pretty close.

Chosing Holt to do the final Forecast.

	Point Forecast
Sep-17	119
Oct-17	131.3
Nov-17	132.45
Dec-17	131.42
Jan-18	118.55
Feb-18	117.47
Mar-18	112.37
Apr-18	108.4
May-18	103.66
Jun-18	105.04
Jul-18	105.26
Aug-18	114.07
Sep-18	119.61
Oct-18	131.91
Nov-18	133.06
Dec-18	132.03



The above is the forecast until 2018 December it would be good if we could see the exact values of candies produced so we could get an idea if my forecast would have helped them if it had been done earlier.