**Forecasting Exam Exercise 2**

First of all we will split our data set into train and test sets train set from 1973 to 2010 and test set from 2011 onwards.

#Set Working Directory

setwd("C:/Users/mmajid1/Desktop/Forecasting")

data\_Houses<-read\_excel("DataSets.xlsx", sheet="Houses")

data\_Houses$average\_house\_prices<- data\_Houses$Houseprices/data\_Houses$Houses

avghouprc <- ts(data\_Houses[,4], frequency = 1, start = 1973)

# Split the data in training and test set

avghouprc1 <- window(avghouprc, end=2010)

avghouprc2 <- window(avghouprc, start=2011)

# Retrieve the length of the test set

h <- length(avghouprc2)

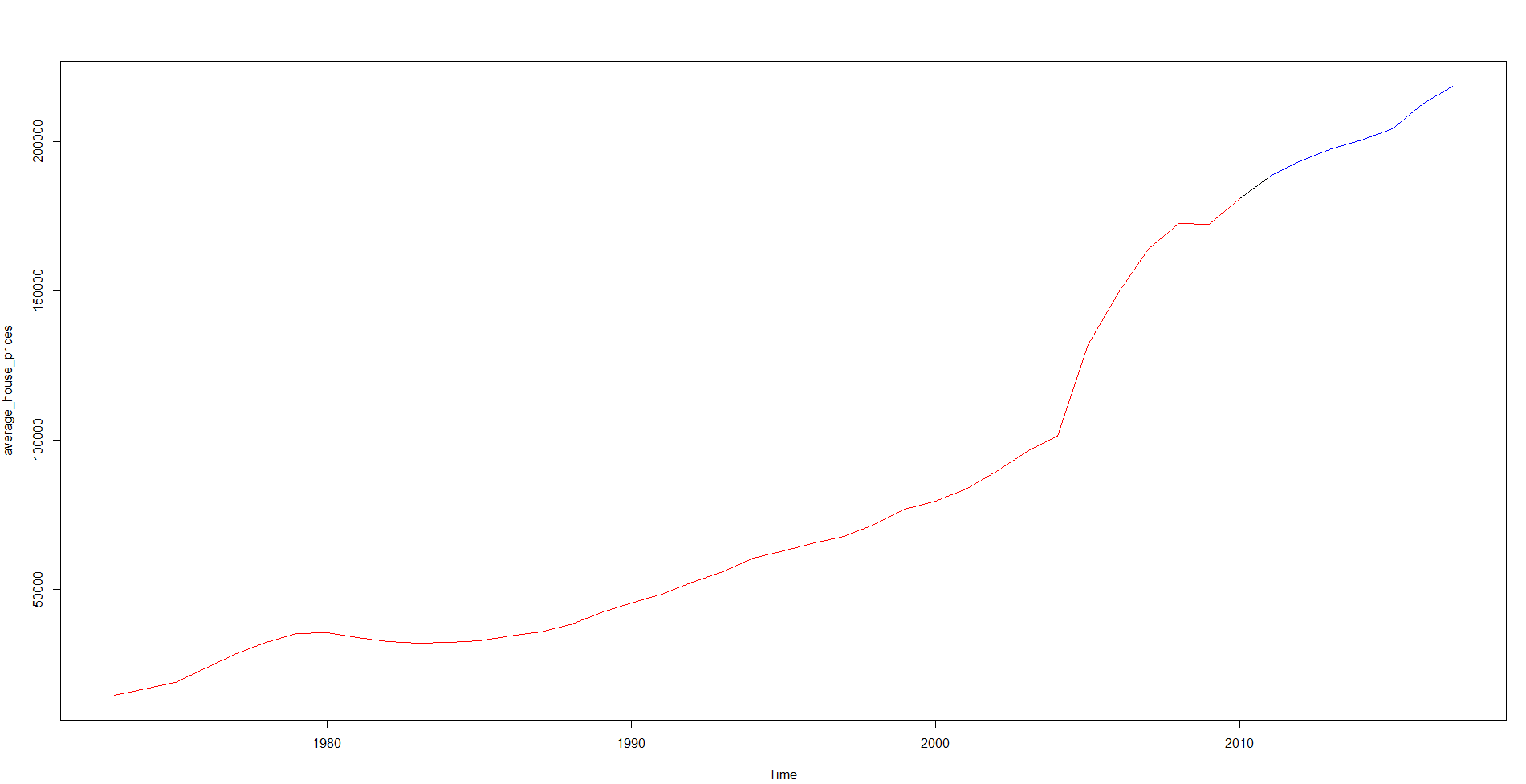
1. **Data Exploration and Visualization**

#Data Visualization

plot(avghouprc)

lines(avghouprc1, col="red")

lines(avghouprc2, col="blue")



We can observe a trend with no seasonality in the graph above.

1. **Forecasts using appropriate naïve method.**

f1 <- meanf(avghouprc1, h=h) #mean

f2 <- rwf(avghouprc1, h=h) #naive

f3 <- rwf(avghouprc1, drift=TRUE, h=h) #drift

accuracy(f1,avghouprc2)

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 4.967442e-12 46890.11 36795.66 -53.50653 79.47985 7.814272 0.8939366 NA

Test set 1.351759e+05 135532.41 135175.95 66.75574 66.75574 28.707231 0.5295402 25.83867

accuracy(f2,avghouprc2)

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 4500.453 7310.767 4708.777 6.410535 6.986804 1.000000 0.5499946 NA

Test set 21244.188 23405.435 21244.188 10.294740 10.294740 4.511615 0.5295402 4.607145

accuracy(f3,avghouprc2)

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 5.901345e-13 5761.357 3409.851 -3.327243 5.890129 0.7241480 0.5499946 NA

Test set 3.242378e+03 3660.853 3242.378 1.585778 1.585778 0.6885818 0.2458799 0.6955479

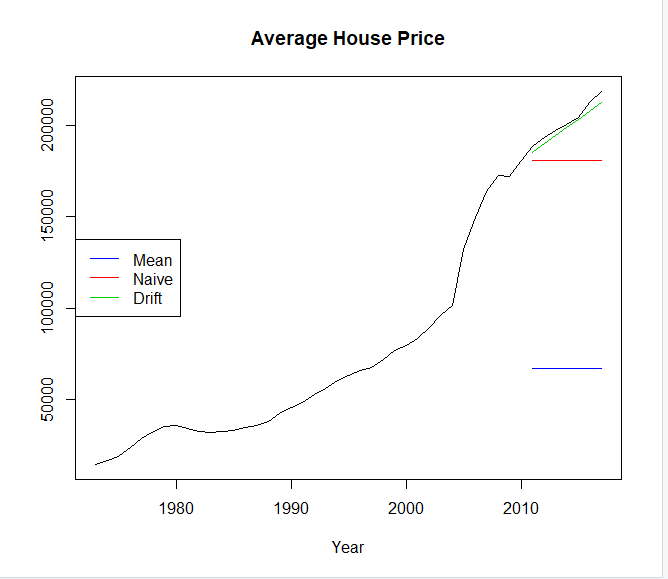
plot(avghouprc,main="Average House Price", ylab="",xlab="Year")

lines(f1$mean,col=4)

lines(f2$mean,col=2)

lines(f3$mean,col=3)

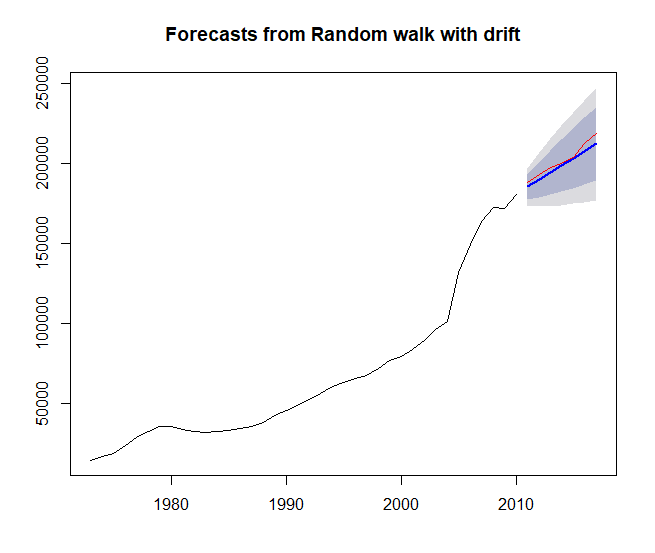
legend("left",lty=1,col=c(4,2,3),legend=c("Mean","Naive","Drift"))



Drift performs the best it has least RMSE moreover, Its more close to the actual and we can clearly see it by visualization.

plot(f3)

lines(avghouprc, col="red")



Now we will check the residuals

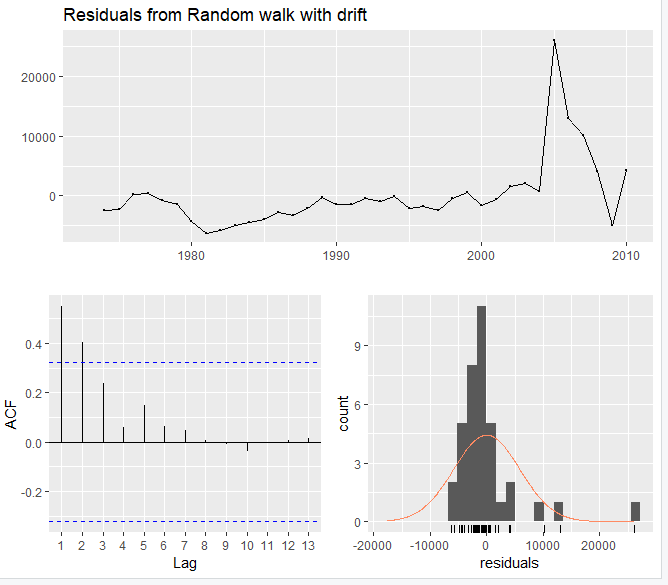
> checkresiduals(f3)

Ljung-Box test

data: Residuals from Random walk with drift

Q\* = 22.775, df = 6.6, p-value = 0.001393

Model df: 1. Total lags used: 7.6



We can observe that there is no white noise in the residuals.

Now we will check residuals from naive

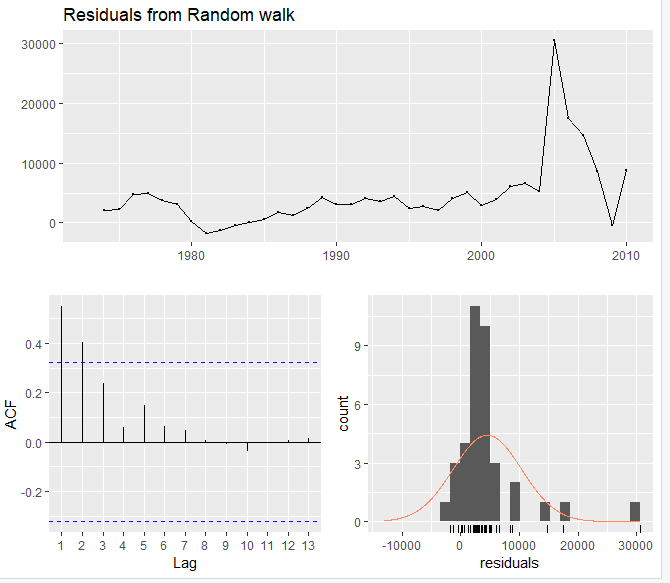
checkresiduals(f2)

Ljung-Box test

data: Residuals from Random walk

Q\* = 22.775, df = 7.6, p-value = 0.002822

Model df: 0. Total lags used: 7.6



Here we can also see no white noise.

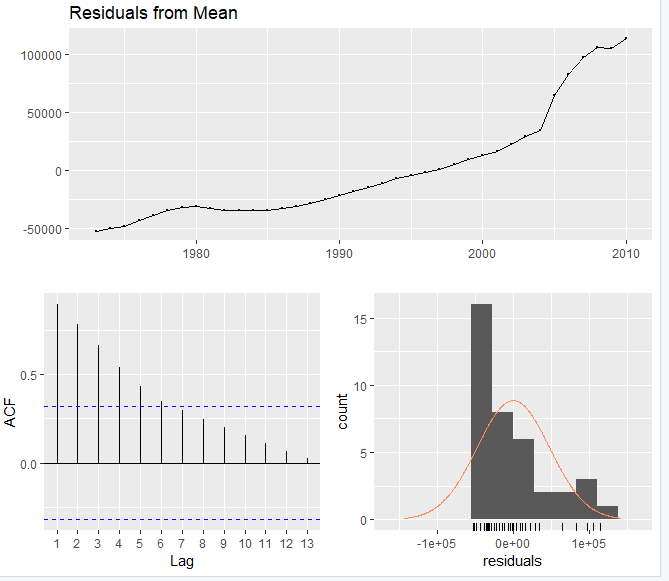
checkresiduals(f1)

Ljung-Box test

data: Residuals from Mean

Q\* = 109.5, df = 6.6, p-value < 2.2e-16

Model df: 1. Total lags used: 7.6

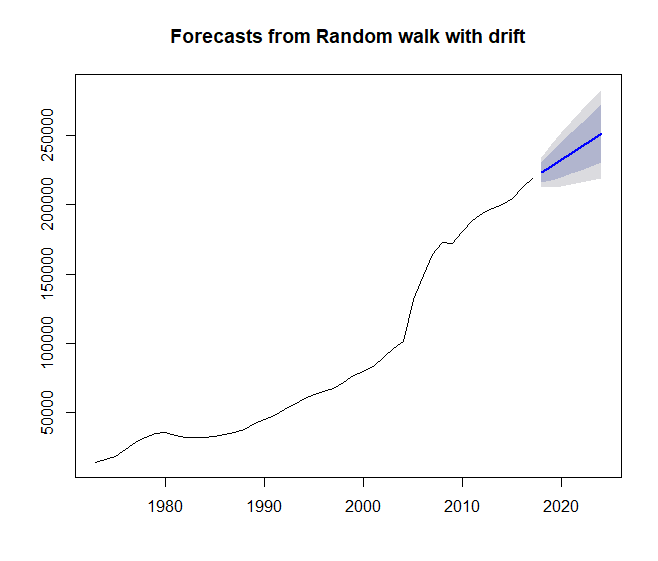


So based on accuracy and white noise in residuals Drift method is the best of all 3.

#now lets forecast on the original dataset.

F3 <- rwf(avghouprc, drift=TRUE, h=h) #drift

plot(f3)



1. **Forecasts using relevant exponential smoothing methods.**

We will use Normal Holt Method(h1), h2(Damped without exponential component),h3 (Exponential with Damping)

h1 <- holt(avghouprc1, h=100)

h2 <- holt(avghouprc1, h=100, damped=TRUE)

h4 <- holt(avghouprc, h=100, exponential=TRUE, damped=TRUE)

plot(h1, type="l", ylab="Average House Prices",

xlab="Year", fcol="white", shadecols="white")

lines(fitted(h1), col=2)

lines(fitted(h2), col=3)

lines(fitted(h3), col=4)

lines(fitted(h4), col=5)

lines(h1$mean, col=2, type="l")

lines(h2$mean, col=3, type="l")

lines(h4$mean, col=5, type="l") #almost the same as h2

legend("bottomleft", lty=1, col=c(2,3,4,5),c("Holt's Linear", "Additive Damped", "Multiplicative Damped"), cex=0.75)

here we can see that h4 is the best performing by visualization.

Now we will check the accuracy measure of all 3 models h1,h2 and h4

accuracy(h1, avghouprc2)[,c(2,3,5,6)]

RMSE MAE MAPE MASE

Training set 4985.364 2394.670 3.530889 0.5085545

Test set 9167.426 7722.326 3.720775 1.6399856

accuracy(h2,avghouprc2)[,c(2,3,5,6)]

RMSE MAE MAPE MASE

Training set 4874.444 2297.444 3.453124 0.4879067

Test set 7828.681 6381.759 3.064461 1.3552902

accuracy(h4, avghouprc2)[,c(2,3,5,6)]

RMSE MAE MAPE MASE

Training set 5049.69 2841.268 5.00893 0.6033983

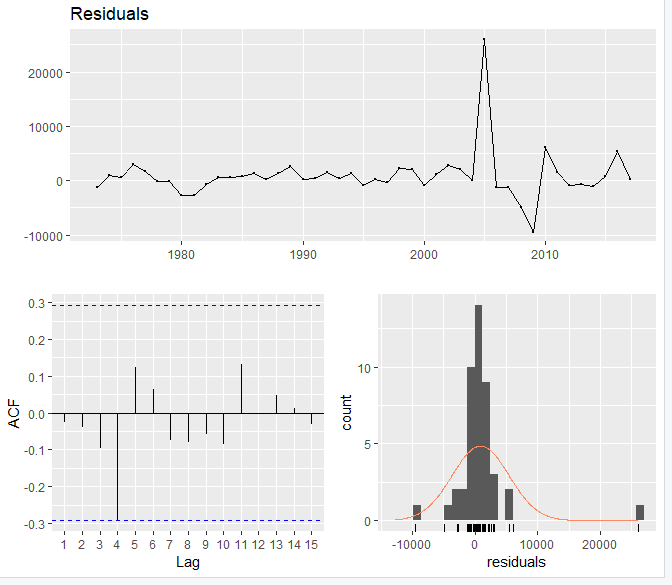
Test set 28184.65 24241.094 11.67866 5.1480661

We can see that h2 has the best accuracy measure here.

Lets check white noise for h2 the best performing model.

res <- residuals(h2)

checkresiduals(res)



LjungBox(res, lags=seq(1,15,1), order=length(h2$model$par))

lags statistic df p-value

1 0.03406617 0 NA

2 0.10463587 0 NA

3 0.56994848 0 NA

4 4.94749604 0 NA

5 5.76085092 0 0.00000000

6 5.98158070 1 0.01445604

7 6.29481845 2 0.04296329

8 6.65877434 3 0.08360681

9 6.86484736 4 0.14320514

10 7.30859763 5 0.19868242

11 8.41606974 6 0.20917752

12 8.41698299 7 0.29726358

13 8.57275134 8 0.37960861

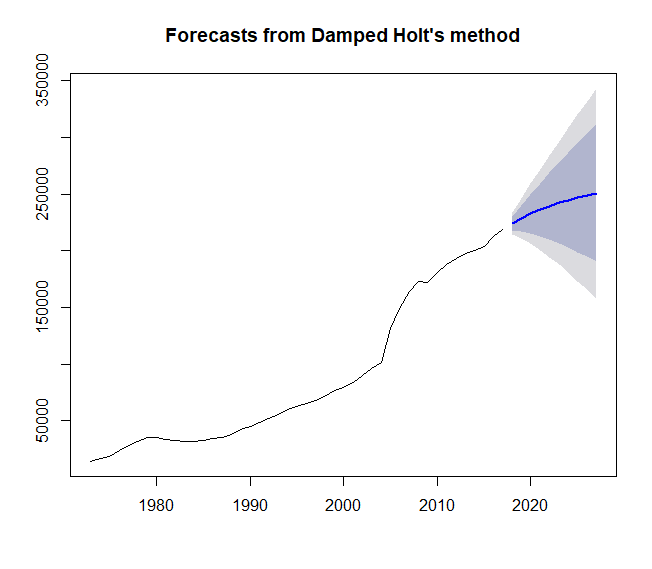
14 8.58551359 9 0.47637965

15 8.64729014 10 0.56587175

So we can see white noise in the residuals of best model h2 so we will go on with it.

h2\_on\_total\_data = holt(avghouprc,damped = TRUE)

> plot(h2\_on\_total\_data)



Accuracy Measure of this model is

Test set 7828.681 6381.759 3.064461 1.3552902

1. **Generating Forecasts using ETS.**

As there is no seasonality so we will use only below models.

#Different ETS models evaluation

e1 <- ets(avghouprc1,"AAN")

e2 <- ets(avghouprc1,"MNN")

e3 <- ets(avghouprc1,"ANN")

e4 <- ets(avghouprc1,"MAN")

e5 <- ets(avghouprc1,"MMN")

e6 <- ets(avghouprc1, model = "AAN", damped = TRUE)

e7 <- ets(avghouprc1, model = "MAN", damped = TRUE)

e8 <- ets(avghouprc1, model = "MMN", damped = TRUE)

auto\_ets = ets(avghouprc1)

Now lets do the forecasting for these all ets models.

f1 = forecast(e1,h=h)

> accuracy(f1,avghouprc2)[,c(2,3,5,6)]

RMSE MAE MAPE MASE

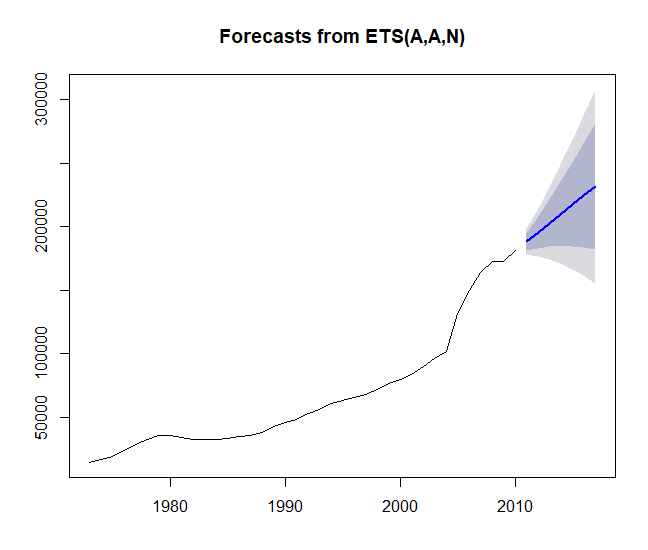
Training set 4985.364 2395.119 3.530972 0.5086499

Test set 9158.589 7714.848 3.717184 1.6383974

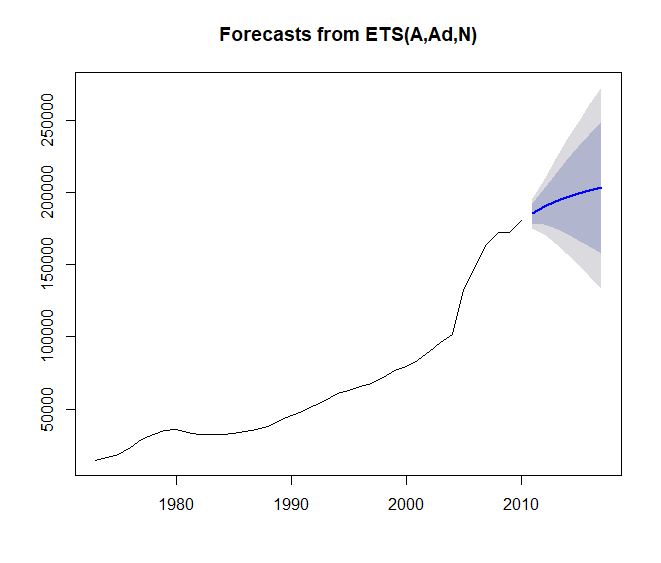
|  |
| --- |
| f2 = forecast(e2,h=h)  > accuracy(f2,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 7500.269 4917.907 9.098052 1.044413  Test set 23406.228 21245.062 10.295173 4.511801 |
|  |
| > f3 = forecast(e3,h=h)  > accuracy(f3,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 7500.269 4917.907 9.098052 1.044413  Test set 23406.228 21245.062 10.295173 4.511801  f4 = forecast(e4,h=h)  > accuracy(f4,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 5095.913 2602.828 3.574382 0.552761  Test set 8743.520 7361.704 3.547449 1.563400  f5 = forecast(e5,h=h)  > accuracy(f5,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 5161.543 2603.518 3.608358 0.5529074  Test set 2830.206 2376.248 1.149464 0.5046423  > f6 = forecast(e6,h=h)  > accuracy(f6,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 4874.445 2297.335 3.452873 0.4878835  Test set 7792.013 6347.321 3.047789 1.3479766  > f7 = forecast(e7,h=h)  > accuracy(f7,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 5035.41 2545.549 3.532507 0.5405968  Test set 3899.58 3225.074 1.574671 0.6849070  f8 = forecast(e8,h=h)  > accuracy(f8,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 5161.543 2603.518 3.608358 0.5529074  Test set 2830.206 2376.248 1.149464 0.5046423  > fauto\_ets = forecast(auto\_ets,h=h)  > accuracy(fauto\_ets,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 5095.913 2602.828 3.574382 0.552761  Test set 8743.520 7361.704 3.547449 1.563400   |  | | --- | |  | |

Model f1 and f6 are the best models as per there RMSE measures.

Plot(f1)

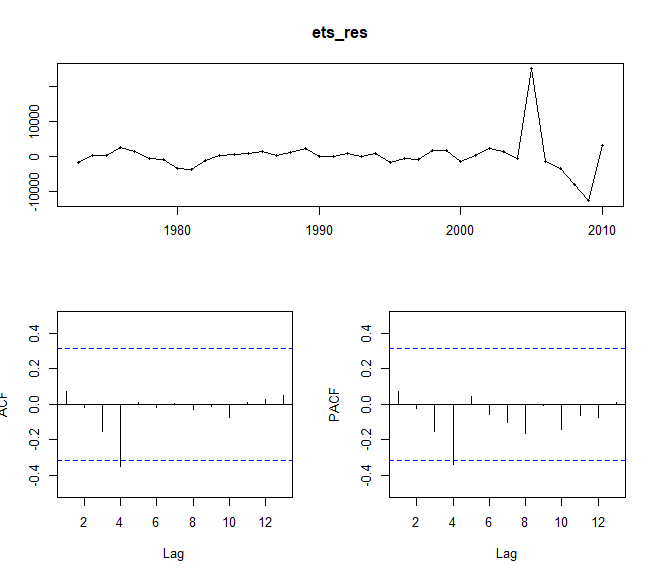


Plot(f6)



> ets\_res = residuals(e1)

> tsdisplay(ets\_res)



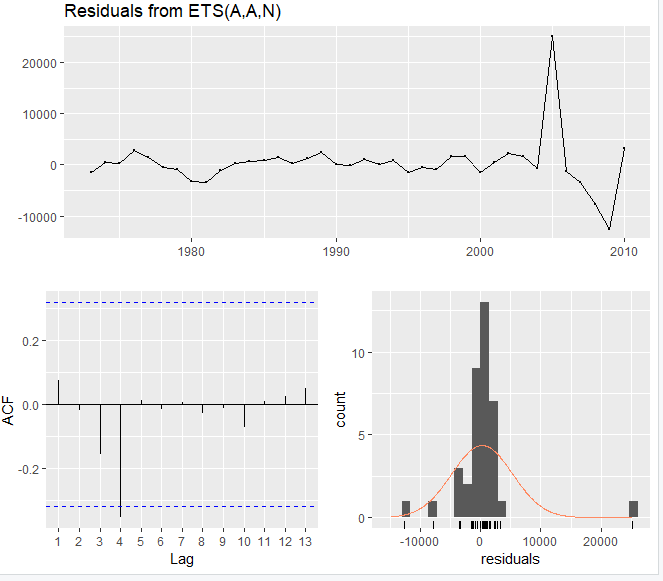
checkresiduals(e1)

Ljung-Box test

data: Residuals from ETS(A,A,N)

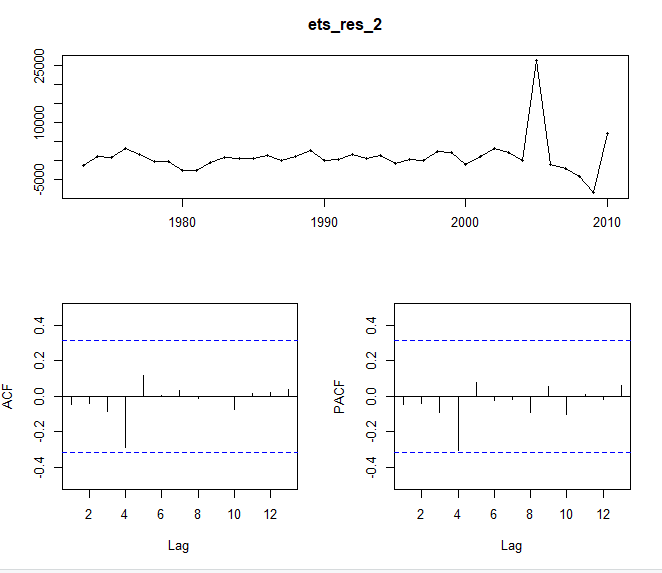
Q\* = 6.9231, df = 4, p-value = 0.14

Model df: 4. Total lags used: 8



> ets\_res\_2 = residuals(e6)

> tsdisplay(ets\_res\_2)



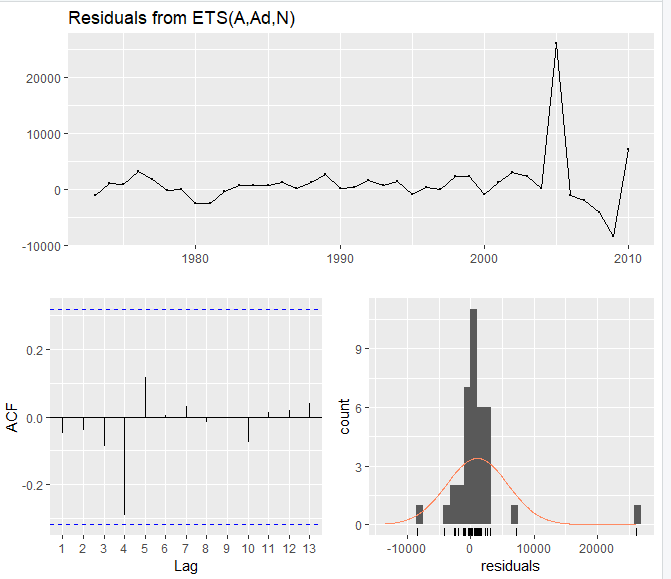
> checkresiduals(e6)

Ljung-Box test

data: Residuals from ETS(A,Ad,N)

Q\* = 4.9408, df = 3, p-value = 0.1762

Model df: 5. Total lags used: 8



> LjungBox(ets\_res,lags = seq(5,8,1),order=7) # for model 1

lags statistic df p-value

5 6.866430 0 NA

6 6.878617 0 NA

7 6.881159 0 0.000000000

8 6.923058 1 0.008509135

> LjungBox(ets\_res\_2,lags = seq(5,8,1),order=7) # for model 6

lags statistic df p-value

5 4.878658 0 NA

6 4.879233 0 NA

7 4.930209 0 0.00000000

8 4.940778 1 0.02623022

Now I will forcast on main dataset with model 6. ETS(A,Ad,N)

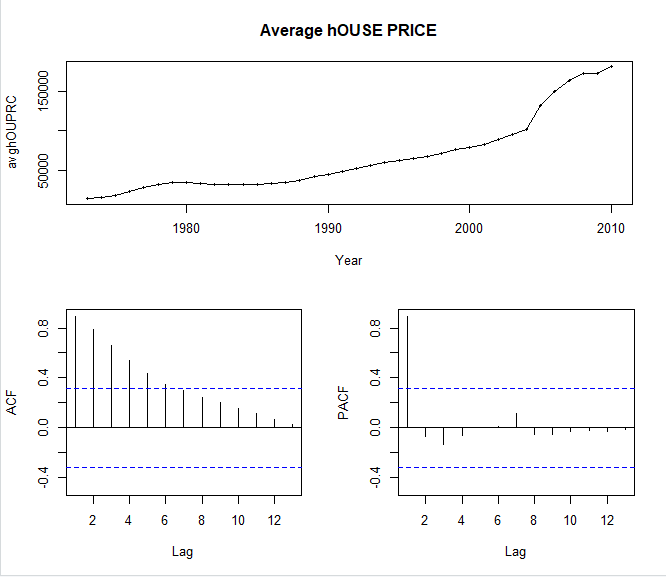
ets\_complete\_data = ets(avghouprc,model = "AAN", damped = TRUE)

|  |
| --- |
| summary(ets\_complete\_data)  ETS(A,Ad,N)  Call:  ets(y = avghouprc, model = "AAN", damped = TRUE)  Smoothing parameters:  alpha = 0.9999  beta = 0.6393  phi = 0.8842  Initial states:  l = 13610.0781  b = 2460.5  sigma: 4843.934  AIC AICc BIC  941.6930 943.9035 952.5330  Training set error measures:  ME RMSE MAE MPE MAPE MASE ACF1  Training set 874.6457 4566.904 2167.742 1.177223 3.023633 0.4500922 -0.02662851 |
|  |
| |  | | --- | | > forc2 = forecast(ets\_complete\_data,h=3)  > plot(forc2)    > accuracy(f6,avghouprc2)[,c(2,3,5,6)]  RMSE MAE MAPE MASE  Training set 4874.445 2297.335 3.452873 0.4878835  Test set 7792.013 6347.321 3.047789 1.3479766 | |

1. **Forecasting USING Arima**

We further investigate the characteristics of the time series.

tsdisplay(avghouprc1, main="Average hOUSE PRICE", ylab="avghOUPRC", xlab="Year")



We can see there is no seasonality and trending time series.

ACF shows that non stationary behavior is being caused by trend. we will try differencing

> ndiffs(avghouprc1)

[1] 2

> ndiffs(diff(avghouprc1))

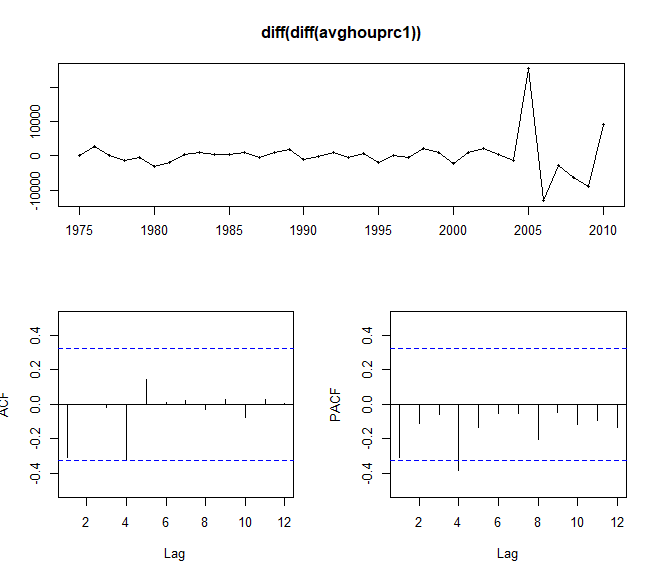
[1] 1

> ndiffs(diff(diff(avghouprc1)))

[1] 0

So we will do double differentiation here

tsdisplay(diff(diff(avghouprc1)))



Now we will apply auto.arima model on the data

> auto\_arima <-auto.arima(avghouprc1, seasonal=FALSE, allowdrift = TRUE)

> summary(auto\_arima)

Series: avghouprc1

ARIMA(1,2,1)

Coefficients:

ar1 ma1

0.4281 -0.8661

s.e. 0.2065 0.1187

sigma^2 estimated as 25432946: log likelihood=-357.31

AIC=720.62 AICc=721.37 BIC=725.37

Training set error measures:

ME RMSE MAE MPE MAPE MASE ACF1

Training set 669.5745 4770.303 2105.782 0.663473 2.834333 0.4472036 -0.05493343

So Arima(1,2,1) is suggested.

Lets forecast using this model.

> f\_auto =forecast(auto\_arima,h=h)

Now lets calculate the accuracy of the model on test data set.

> accuracy(f\_auto,avghouprc2)[,c(2,3,5,6)]

RMSE MAE MAPE MASE

Training set 4770.303 2105.782 2.834333 0.4472036

Test set 14422.719 12474.564 6.019946 2.6492155

Now lets check residuals.

> checkresiduals(auto\_arima)

Ljung-Box test

data: Residuals from ARIMA(1,2,1)

Q\* = 4.5148, df = 6, p-value = 0.6074

Model df: 2. Total lags used: 8

So there is a white noise.

Now we will select our own arima model.

> arima123 <- Arima(avghouprc1, order=c(0,2,1))

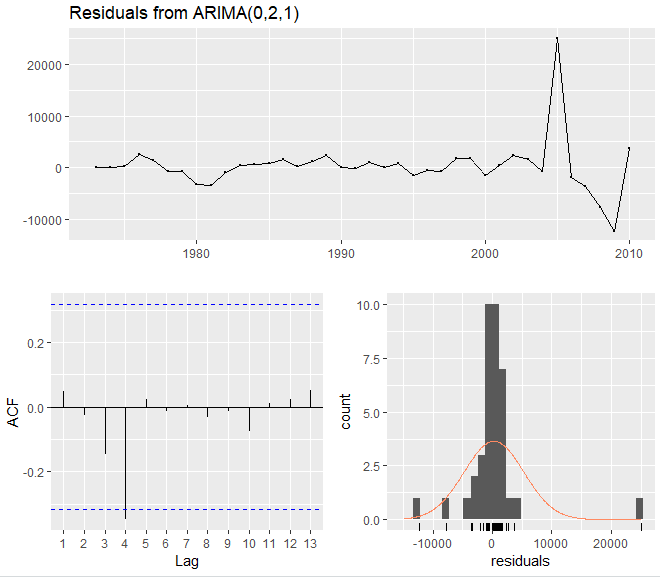
> checkresiduals(arima123)

Ljung-Box test

data: Residuals from ARIMA(0,2,1)

Q\* = 6.5724, df = 7, p-value = 0.4747

Model df: 1. Total lags used: 8



> forcastarima123 =forecast(arima123 ,h=h)

> accuracy(forcastarima123,avghouprc2)[,c(2,3,5,6)]

RMSE MAE MAPE MASE

Training set 4976.597 2361.837 3.135125 0.5015818

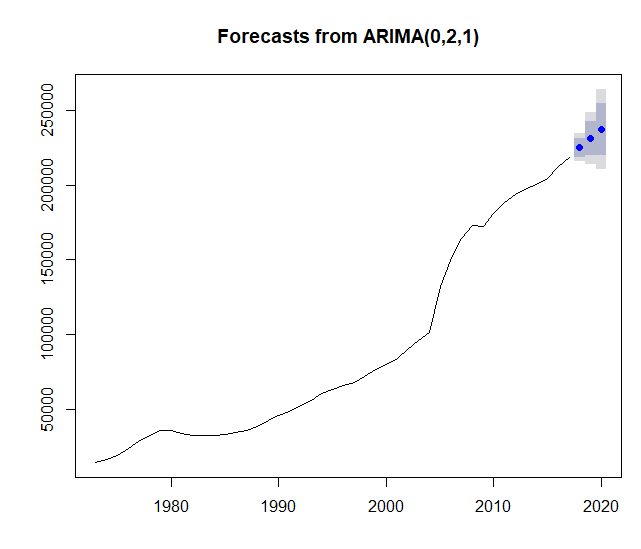
Test set 8693.973 7321.283 3.528208 1.5548164

So accuracy for arima123 is better than auto.arima model so we will forcast by using our own model.

> final = Arima(avghouprc, order = c(0,2,1))

> f\_final = forecast(final,h=3)

> plot(f\_final)



Now we will estimate the best of all models based on RMSE.

For Naïve with drift f3

Test set 21244.188 23405.435 21244.188 10.294740 10.294740 4.511615 0.5295402 4.607145

For Holt with damping h2

Test set 7828.681 6381.759 3.064461 1.3552902

For ets f6

Test set 7792.013 6347.321 3.047789 1.3479766

For Arima model

Test set 8693.973 7321.283 3.528208 1.5548164

So RMSE for h2 is the least hence holt with damping h2 is the best model. Below is the final forecast by using that method.

