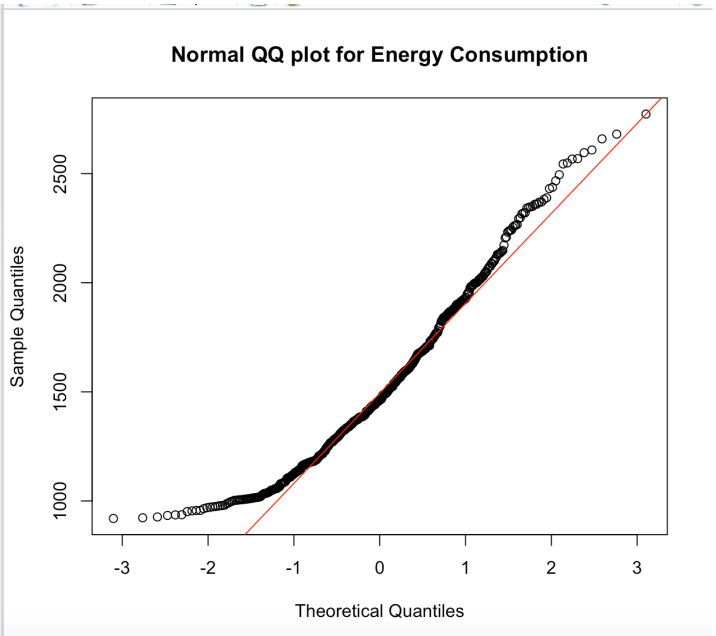
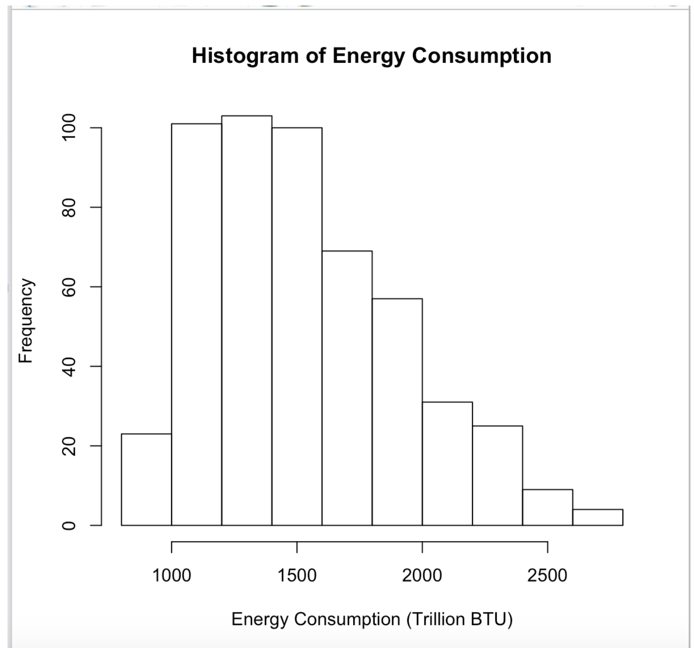
CSC 425
HW 4
Sarah Cummings

1) Energy Consumption Problem

a. Analysis of distribution of energy consumption





For this data, we have the following summary statistics:

Minimum	919.805000
Maximum	2772.805000
1. Quartile	1215.065000
3. Quartile	1771.142250
Mean	1529.458182
Median	1463.134000

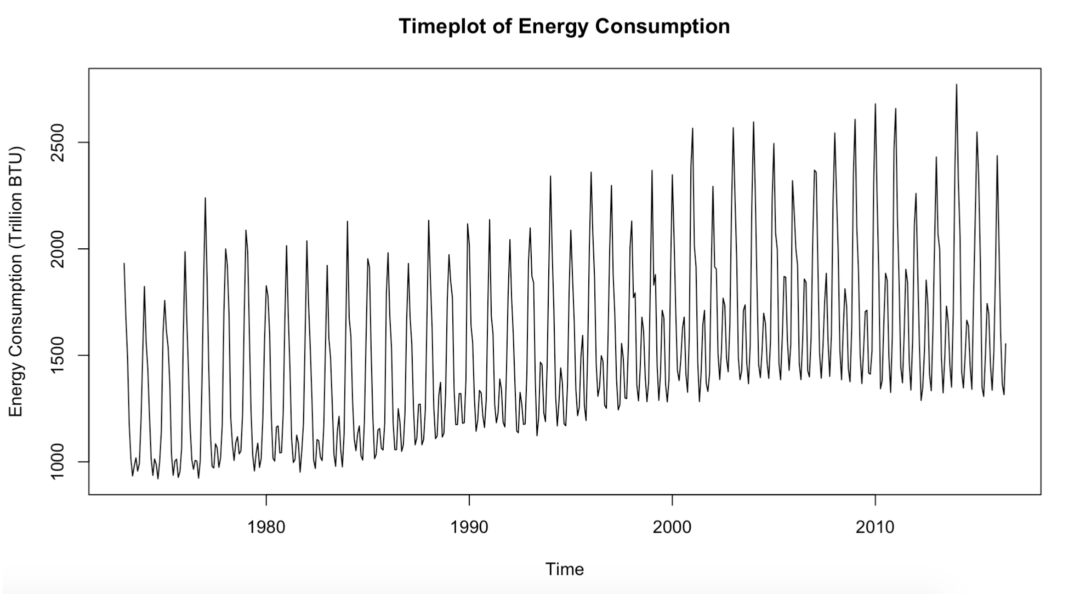
```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 38.5278
  P VALUE:
    Asymptotic p Value: 4.303e-09
```

Skewness	0.661541
Kurtosis	-0.114761

Looking at the histogram, there appears to be a right skew in the data. The Jarque-Bera Normality test has a significant statistic, so we reject the null hypothesis of a normal distribution.

b. Analysis of time plot of energy consumption

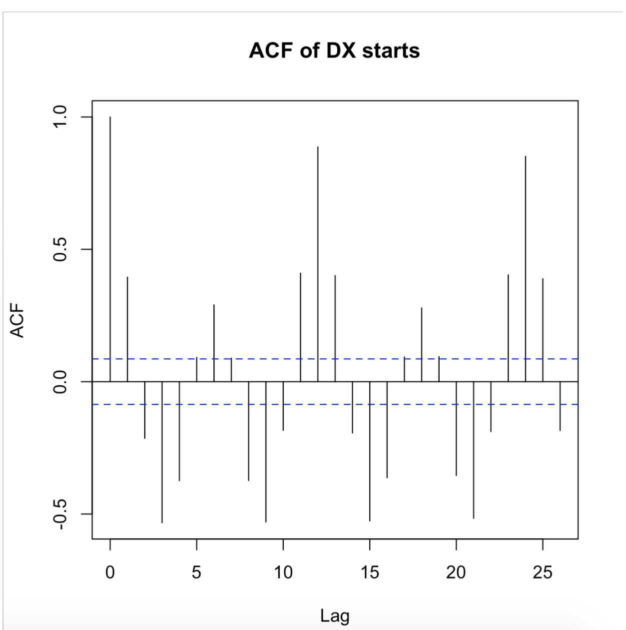


Looking at the time plot, the energy consumption appears to be increasing over time. There also is likely some seasonality, since there are repetitive flocculations that seem may coincide with them of year. The energy consumption appears to have regular cycles of high values (around 2000) and low values (around 1000) that likely correspond to the seasons where we use more or less energy due to weather. The variance seems seems to be increasing over time.

c. Should we transform the data?

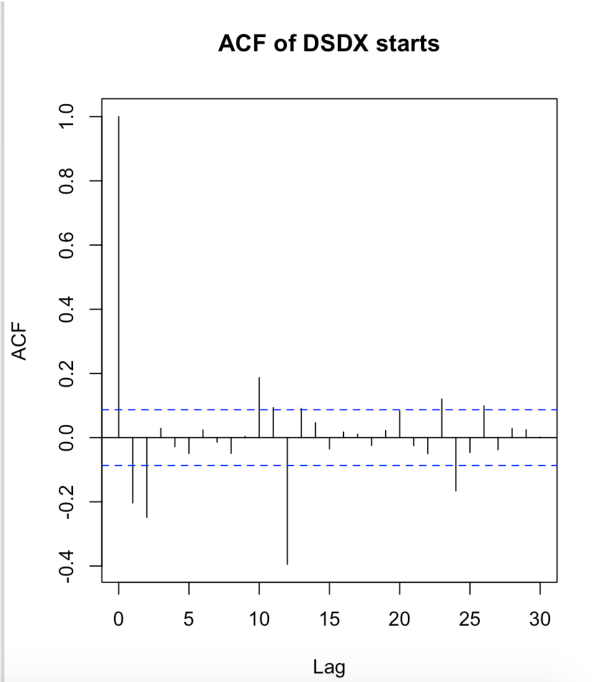
I would not transform the data.

d. Analyze the ACF of the first difference of the time series data



Looking at the ACF of the first difference, we see that lags that are multiples of three are significantly different than zero. Because of the patterns in the ACF, we can conclude there is seasonality in this data.

e. After de-trending and de-seasonalizing, do you obtain a stationary time series?



Box-Ljung test

data: sdx

X-squared = 53.276, df = 3, p-value = 1.601e-11

Box-Ljung test

data: sdx

X-squared = 55.225, df = 6, p-value = 4.175e-10

Yes. Given the small p values of both of our Box Ljung tests, we reject the null hypothesis of independence and conclude energy consumptions are serially correlated.

f. Find initial SARIMA model with autoartima:

Series: ts1
ARIMA(2,0,3)(0,0,2)[12] with non-zero mean

Coefficients:

	ar1	ar2	ma1	ma2	ma3	sma1	sma2	intercept
	0.6671	-0.5847	0.271	0.5025	0.3666	0.6980	0.3979	1527.4942
s.e.	0.1236	0.0824	0.123	0.0727	0.0543	0.0537	0.0433	28.9166

sigma^2 estimated as 19198: log likelihood=-3315.3
AIC=6648.6 AICc=6648.95 BIC=6686.91
z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
ar1	0.667120	0.123611	5.3969	6.780e-08	***
ar2	-0.584687	0.082392	-7.0964	1.281e-12	***
ma1	0.270975	0.122961	2.2037	0.02754	*
ma2	0.502496	0.072693	6.9126	4.758e-12	***
ma3	0.366597	0.054307	6.7505	1.474e-11	***
sma1	0.697964	0.053685	13.0011	< 2.2e-16	***
sma2	0.397880	0.043267	9.1960	< 2.2e-16	***

```
intercept 1527.494150   28.916581 52.8242 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

g. Find adequate model:

Alt model 1:

```
Series: ts1

ARIMA(0,1,1)(0,1,1)[12]
Coefficients:
         ma1     sma1
     -0.5208  -0.7819
s.e.   0.0892   0.0261
sigma^2 estimated as 8470:  log likelihood=-3028.84
AIC=6063.69   AICc=6063.74   BIC=6076.39

z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ma1  -0.520765   0.089178  -5.8396 5.233e-09 ***
sma1 -0.781927   0.026119 -29.9375 < 2.2e-16 ***
```

Alt model 2:

```
Series: ts1
ARIMA(0,1,2)(0,1,1)[12]

Coefficients:
         ma1      ma2     sma1
     -0.4898  -0.4307  -0.7983
s.e.   0.0370   0.0370   0.0278

sigma^2 estimated as 7099:  log likelihood=-2985.05
AIC=5978.1   AICc=5978.18   BIC=5995.03
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ma1  -0.489841   0.037049 -13.221 < 2.2e-16 ***
ma2  -0.430743   0.037013 -11.637 < 2.2e-16 ***
sma1 -0.798329   0.027772 -28.746 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Alt model 3:

```
Series: ts1
ARIMA(1,1,2)(0,1,1)[12]

Coefficients:
         ar1      ma1      ma2     sma1
      0.2928  -0.7329  -0.2195  -0.7959
s.e.  0.0941   0.0964   0.0887   0.0284

sigma^2 estimated as 6990:  log likelihood=-2980.74
AIC=5971.47   AICc=5971.59   BIC=5992.64
```

z test of coefficients:

```
      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.292757   0.094107   3.1109  0.001865 **
ma1  -0.732883   0.096445  -7.5990 2.984e-14 ***
ma2  -0.219543   0.088687  -2.4755  0.013306 *
sma1 -0.795913   0.028372 -28.0526 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. Model selected:

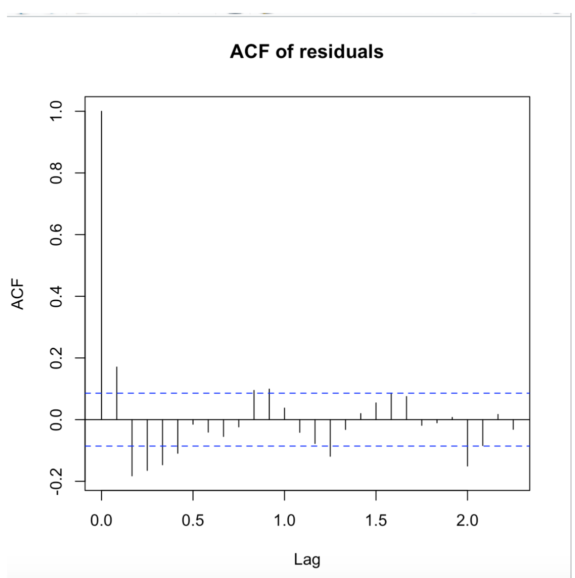
The autoarima and the alt models 2 and 3 do not pass the box-jung test on model residuals, thus I will proceed with alt model 1:
$(1-B^{12})(1-B)\, x_t = (1-0.5208\, B)(1-0.7819\ ^{2})a_t$, with $var(a_t) = 8470$.

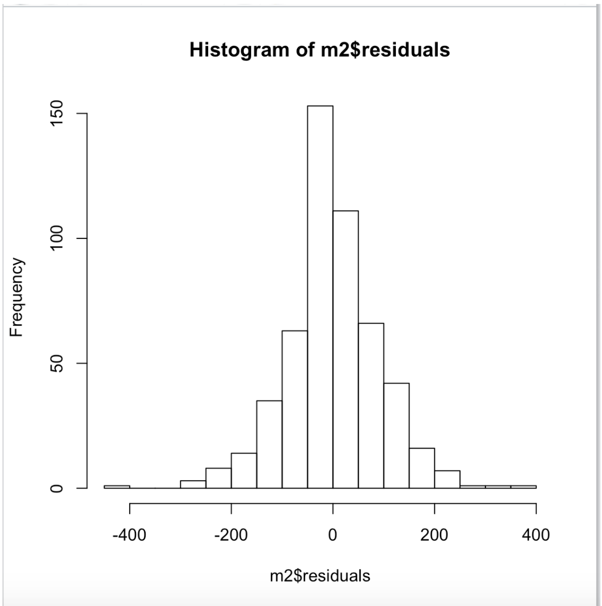
b. Does model provides an adequate explanation of the time process ? Analysis of model.
ARIMA(0,1,1)(0,1,1)[12]

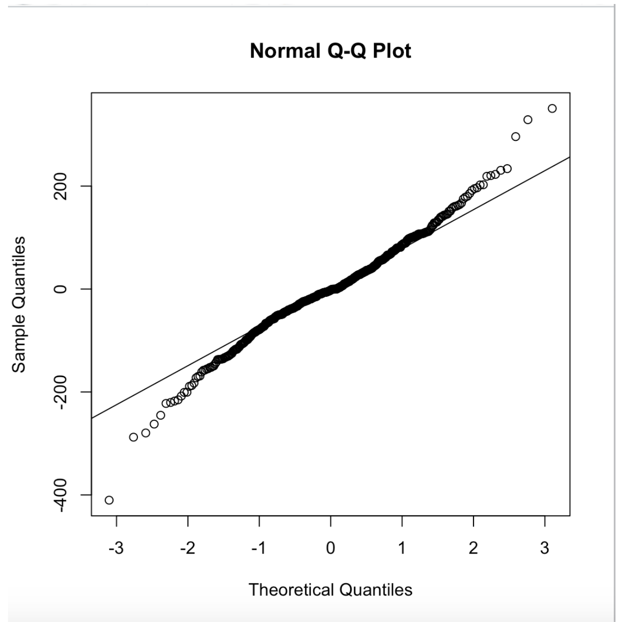
Box-Ljung test
data: m2$residuals
X-squared = 64.8, df = 4, p-value = 2.836e-13

Box-Ljung test
data: m2$residuals
X-squared = 72.359, df = 8, p-value = 1.664e-12

Box-Ljung test
data: m2$residuals
X-squared = 78.368, df = 10, p-value = 1.048e-12







This model has normally distributed residuals as seen above in the histogram and the QQ plot. It also passes the Box- Ljung test of residuals.

h. Compute forecasts for energy consumption for the next five months using the selected model.

f1

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jul 2016	**1732.559**	1614.613	1850.506	1552.176	1912.943
Aug 2016	**1689.049**	1558.257	1819.840	1489.021	1889.076
Sep 2016	**1421.324**	1278.841	1563.806	1203.415	1639.232
Oct 2016	**1318.842**	1165.557	1472.127	1084.413	1553.271
Nov 2016	**1581.857**	1418.482	1745.231	1331.997	1831.717

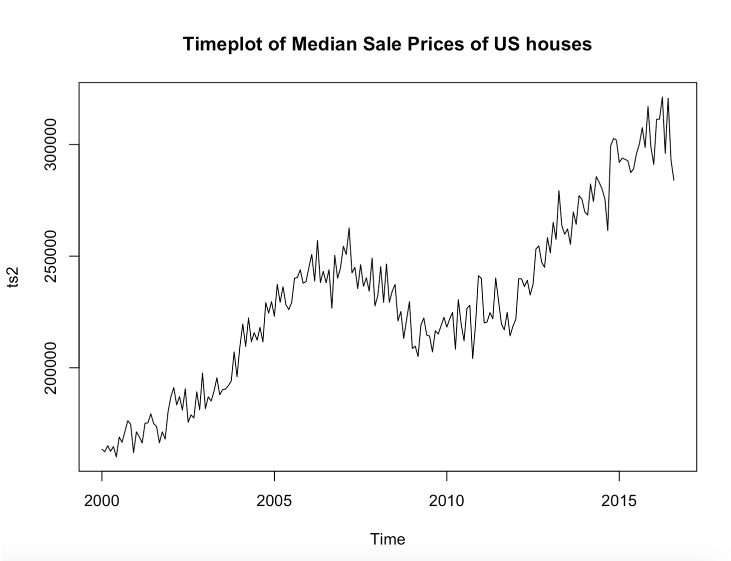
i. Apply backtesting to compute the MAPE for the fitted model. Discuss the accuracy of your model forecasts.

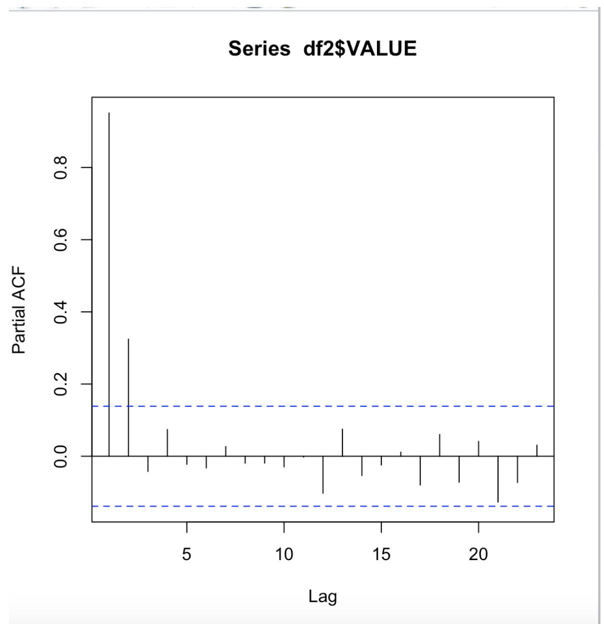
[1] "Mean Absolute Percentage error"
[1] 0.05096556

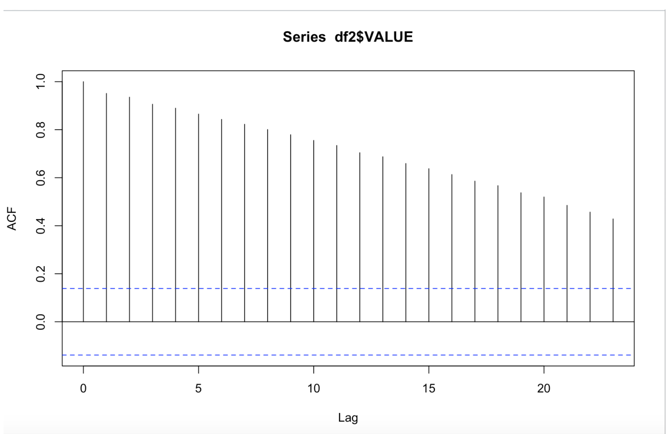
This is a fairly low MAPE, thus our model is fairly accurate.

2) Home sales problem

1. Plot observed time series and its ACFs, 20 lags







2. Analyze if it is stationary using the ACF and Dickey Fuller Test

Looking at the time plot, it looks like this is a unit root non stationary TS. The ACF plot decays, but it decays very slowly. It looks like the ACF plot does not reach zero.

The dickey fuller tests are as follows:

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 3
  STATISTIC:
    Dickey-Fuller: -0.9588
  P VALUE:
    0.6998

Description:
 Sat Nov  5 17:50:45 2016 by user:
```

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 5
  STATISTIC:
    Dickey-Fuller: -0.7586
  P VALUE:
    0.7742

Description:
 Sat Nov  5 17:50:50 2016 by user:
```

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 7
  STATISTIC:
    Dickey-Fuller: -0.6088
  P VALUE:
    0.8298

Description:
 Sat Nov  5 17:51:03 2016 by user:
```

Since we have large p values, we cannot reject the null hypothesis. Thus, the process can be considered non stationary and its dynamic behavior can be explained by an ARIMA(p,1,q) model.

3. Specify an ARIMA model that describes the behavior over time of median home prices.

Series: ts2
ARIMA(0,1,1) with drift

Coefficients:

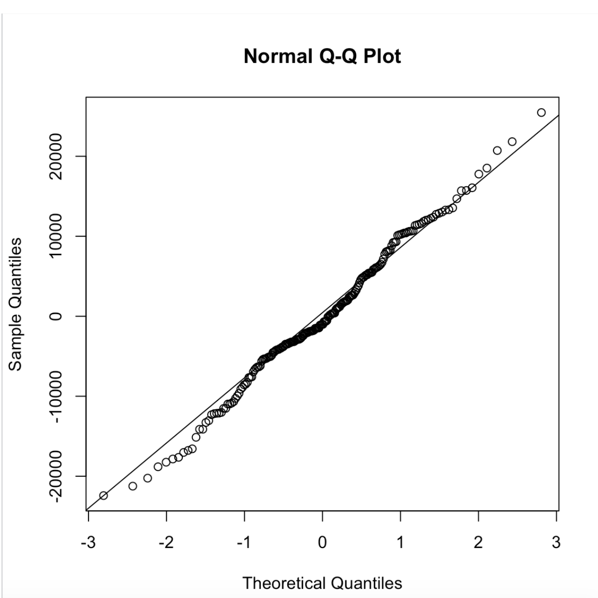
	ma1	drift
	-0.6245	680.3049
s.e.	0.0541	239.0084

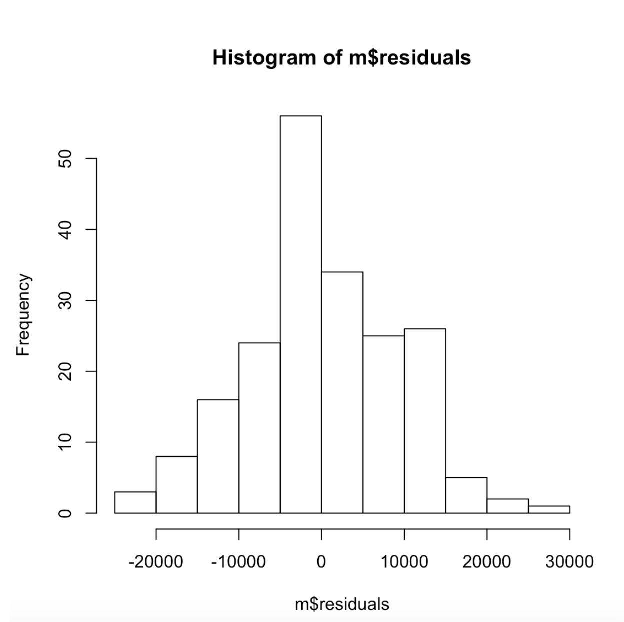
sigma^2 estimated as 79945151: log likelihood=-2092.2
AIC=4190.39   AICc=4190.52   BIC=4200.27

a. Write Down model

$(1-B)\,Xt = 680.3049 + a_t - 0.6245\, a_{t-1}$

b. Residual analysis:





Looking at the histogram and QQ plot of residuals, the residuals appear to be fairly normal. However, they do not pass the Box Ljung test.

Box-Ljung test

data: m$residuals
X-squared = 3.0165, df = 4, p-value = 0.5551

Box-Ljung test

data: m$residuals
X-squared = 6.9049, df = 8, p-value = 0.5469

Box-Ljung test

data: m$residuals
X-squared = 7.9727, df = 10, p-value = 0.6315

I tried did the same model with the natural log of the values hoping it would improve the residuals test, but that did not work.

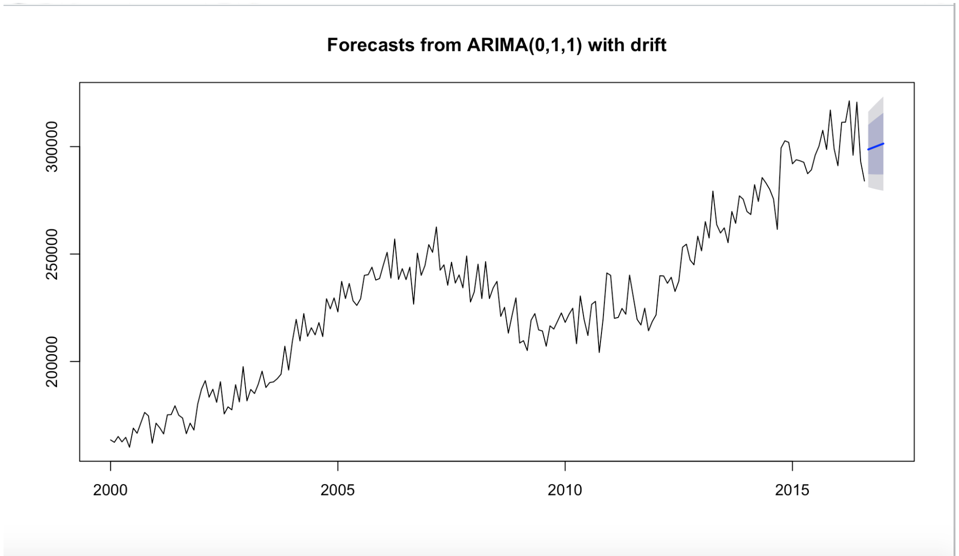
4. Modeling discussion

This data is unit root non stationary with unit-root of order 1. This means there is a zero intercept to the data, and the time series has a linear trend. The linear trend is described in the drift. The median home value in the US has a linear relationship with time. Home values are increasing over time.

5. 5-step ahead forecasts using the fitted model. Write down the forecasts and their standard errors. Do the forecasts show an increasing trend?

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Sep 2016	**298679.9**	287221.3	310138.6	281155.5	316204.4
Oct 2016	**299360.2**	287120.3	311600.2	280640.8	318079.7
Nov 2016	**300040.5**	287066.2	313014.9	280197.9	319883.2
Dec 2016	**300720.9**	287051.5	314390.2	279815.3	321626.4
Jan 2017	**301401.2**	287070.4	315731.9	279484.2	323318.1

The forecasts do continue an increasing trend. I'm not sure where to find the errors for the forecasts, but the error for the model is : 0.05407



6. Use backtesting procedures to compute the RMSE and the MAPE

[1] "RMSE of out-of-sample forecasts"

[1] 98.01205

[1] "Mean Absolute Percentage error"

[1] 0.04451203

Our model is expected to be off by about 4%. On average, the expected values = actual values ± 0.04(actual values)