

Assignment5

Duo Zhou

11/8/2020

###Instructions:

- Total number of points is 30. The assignment's final grade will be multiplied by 1/5 to calculate its weight on the final grade.
- Mark the question number and your final answer clearly (use a textbox.)
- Remember to show and explain your work (If you can't explain it, you don't understand it.)
- Please submit your solution through Canvas. For this assignment, you will use a Manufacturer's Stocks of evaporated and sweetened condensed milk 1971/1 – 1980/12 dataset from FPP package. (Data set: condmilk.rda).

(2 points) Question 1:

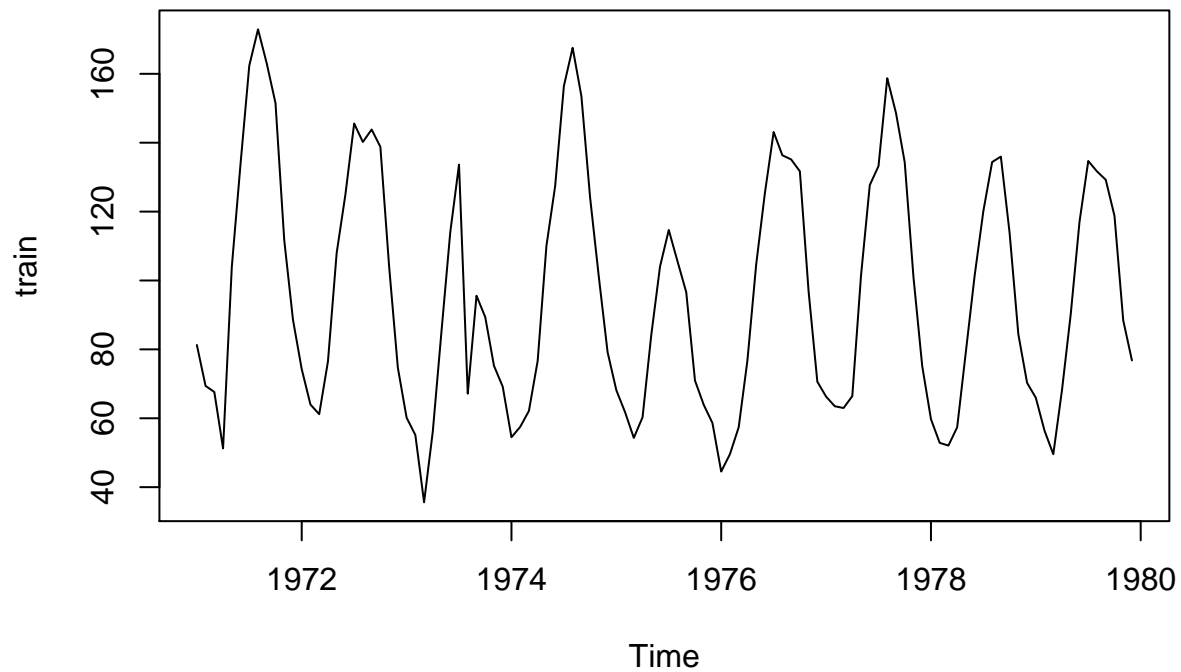
Load the condmilk.rda dataset and split it into a training dataset (1971/1 – 1979/12) and a test dataset (1980/1 – 1980/12)

```
train<-window(condmilk, c(1971, 1), c(1979,12))
test<-window(condmilk, c(1980, 1), c(1980,12))
```

(3 points) Question 2:

Plot the training dataset. Is Box-Cox transformation necessary for this data?

```
plot(train)
```

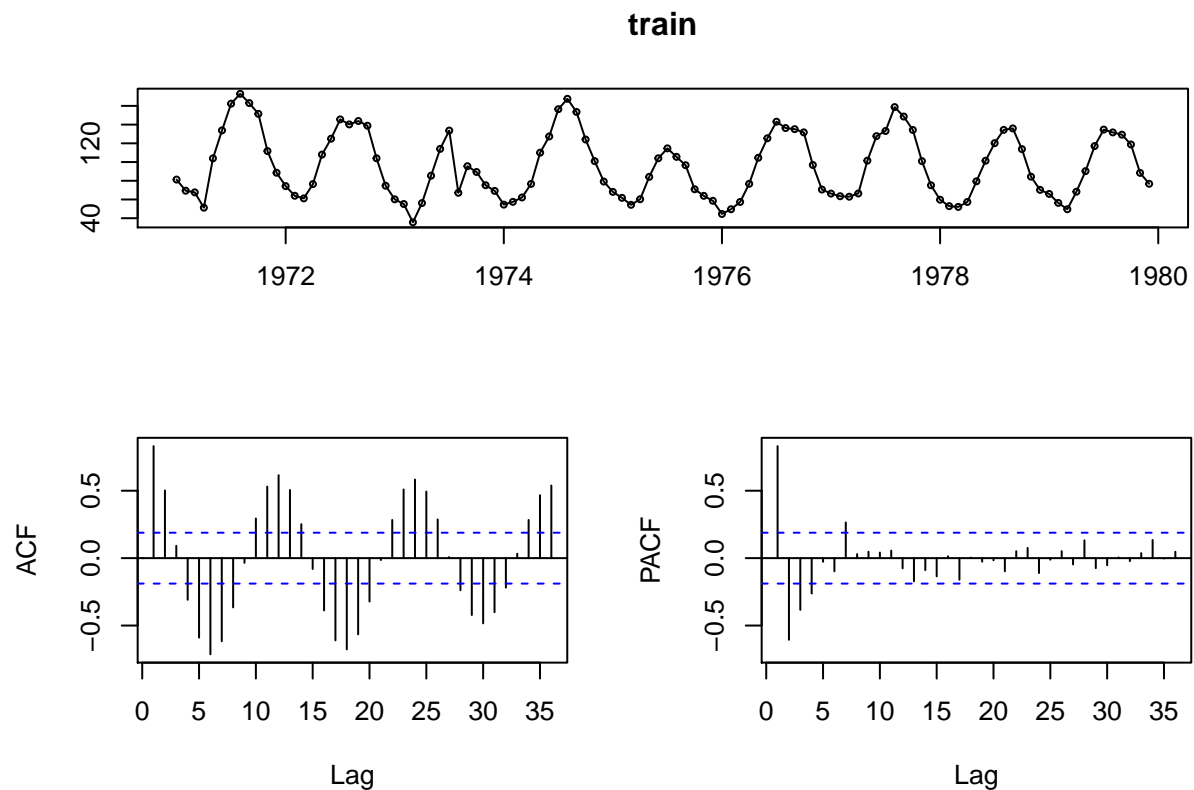


There is a not much variance change throughout the time-series. Box-Cox Transformation is Not necessary for the data.

(5 points) Question 3:

Is the training dataset stationary? If not, find an appropriate differencing which yields seasonal and trend stationary training dataset. Plot the ACF and PACF to determine if the detrended and deseasonalized time series is stationary.

```
tsdisplay(train)
```



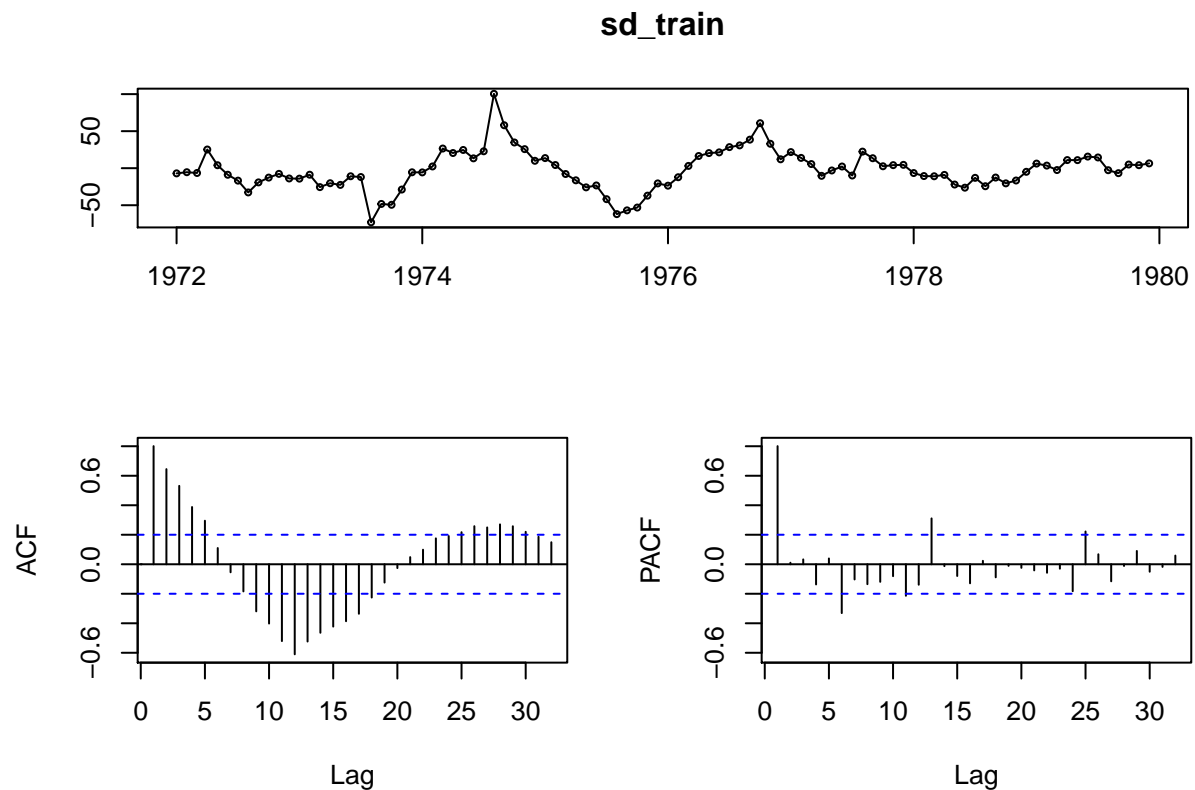
```
adf.test(train,k=12)
```

```
##
## Augmented Dickey-Fuller Test
##
## data: train
## Dickey-Fuller = -3.0469, Lag order = 12, p-value = 0.1424
## alternative hypothesis: stationary
```

For the original data, We see that the ACF does not vanish quickly indicating non-stationarity and ADF test also verified that the series is non-stationary. From the ACF plot, we can see that the seasonality occurs every 12 months.

Taking seasonal differencing of the data at lag=12

```
sd_train<-diff(train,lag=12)
tsdisplay(sd_train)
```

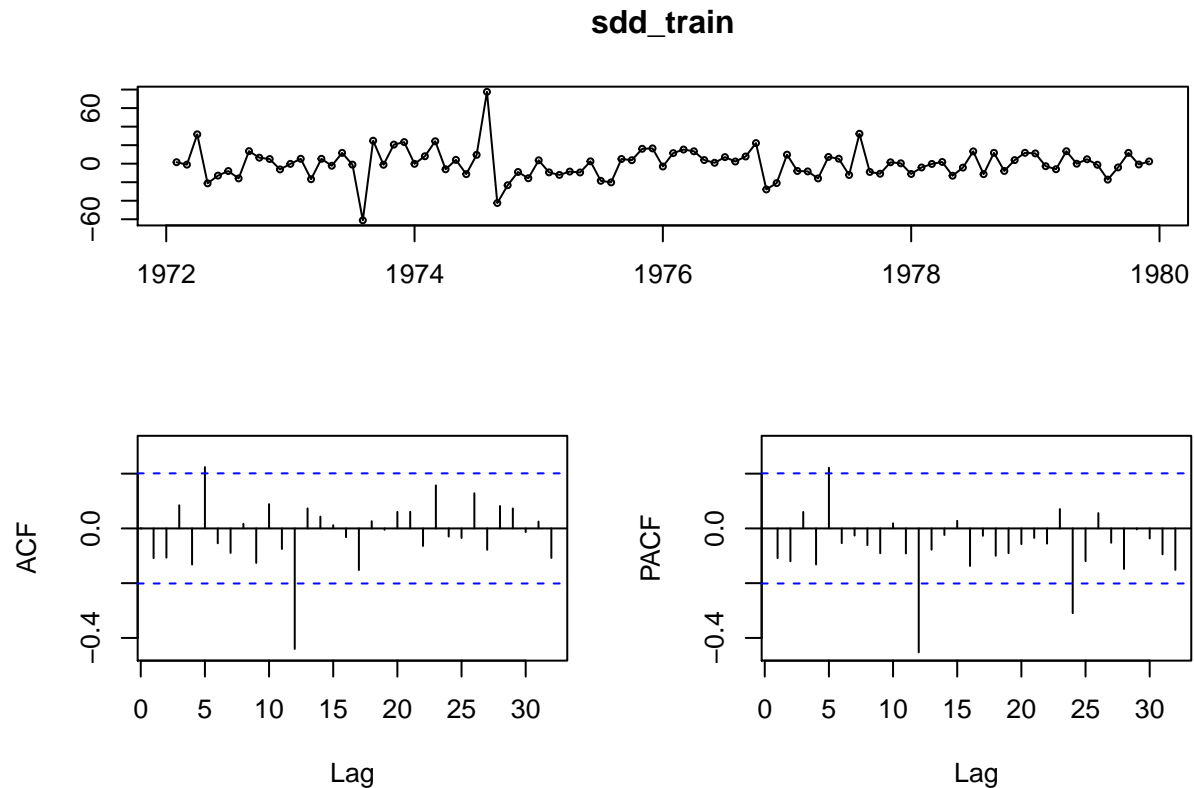


```
adf.test(sd_train,k=12)
```

```
##
## Augmented Dickey-Fuller Test
##
## data: sd_train
## Dickey-Fuller = -2.816, Lag order = 12, p-value = 0.2392
## alternative hypothesis: stationary
```

For the seasonal differenced(lag=12) data, We see that the ACF does not vanish quickly either indicating non-stationarity and ADF test also verified that the series is non-stationary. We can take another first order differencing on top of the seasonal differencing.

```
sdd_train<-diff(diff(train,lag=12))
tsdisplay(sdd_train)
```



```
adf.test(sdd_train,k=24)
```

```
##
## Augmented Dickey-Fuller Test
##
## data: sdd_train
## Dickey-Fuller = -3.5894, Lag order = 24, p-value = 0.03819
## alternative hypothesis: stationary
```

For the first order and seasonal differenced data, We see that the ACF vanishes quickly indicating stationarity and ADF test also verified that the series is stationary. We can conclude that A first order differencing on top of lag12 seasonal differencing is the appropriate differencing which yields seasonal and trend stationary training dataset.

(5 points) Question 4:

Build two ARIMA(p,d,q)(P,D,Q)_s models using the training dataset and `auto.arima()` function. • Model 1: Let the `auto.arima()` function determine the best order of non-seasonal and seasonal differencing. • Model 2: Set the order of seasonal-differencing d to 1 and D to 1. Report the resulting p,d,q,P,D,Q_s and the coefficients values for all cases and compare their AICc and BIC values.

```
model1<-auto.arima(train, seasonal=TRUE)
model2<-auto.arima(train,d=1, D=1)
```

Model1 is ARIMA(1,0,0)(2,1,0)[12] and the the coefficients values and AICc and BIC are reported in the following summary.

```
model1

## Series: train
## ARIMA(1,0,0)(2,1,0)[12]
##
## Coefficients:
##          ar1      sar1      sar2
##          0.7625  -0.7745  -0.5032
## s.e.    0.0676   0.1007   0.0989
##
## sigma^2 estimated as 143.5:  log likelihood=-378.83
## AIC=765.67   AICc=766.11   BIC=775.93
```

Model2 is ARIMA(1,1,1)(2,1,0)[12] and the the coefficients values and AICc and BIC are reported in the following summary.

```
model2

## Series: train
## ARIMA(0,1,1)(2,1,0)[12]
##
## Coefficients:
##          ma1      sar1      sar2
##          -0.1672  -0.7597  -0.4931
## s.e.    0.1104   0.1004   0.0997
##
## sigma^2 estimated as 160:  log likelihood=-379.49
## AIC=766.98   AICc=767.42   BIC=777.19
```

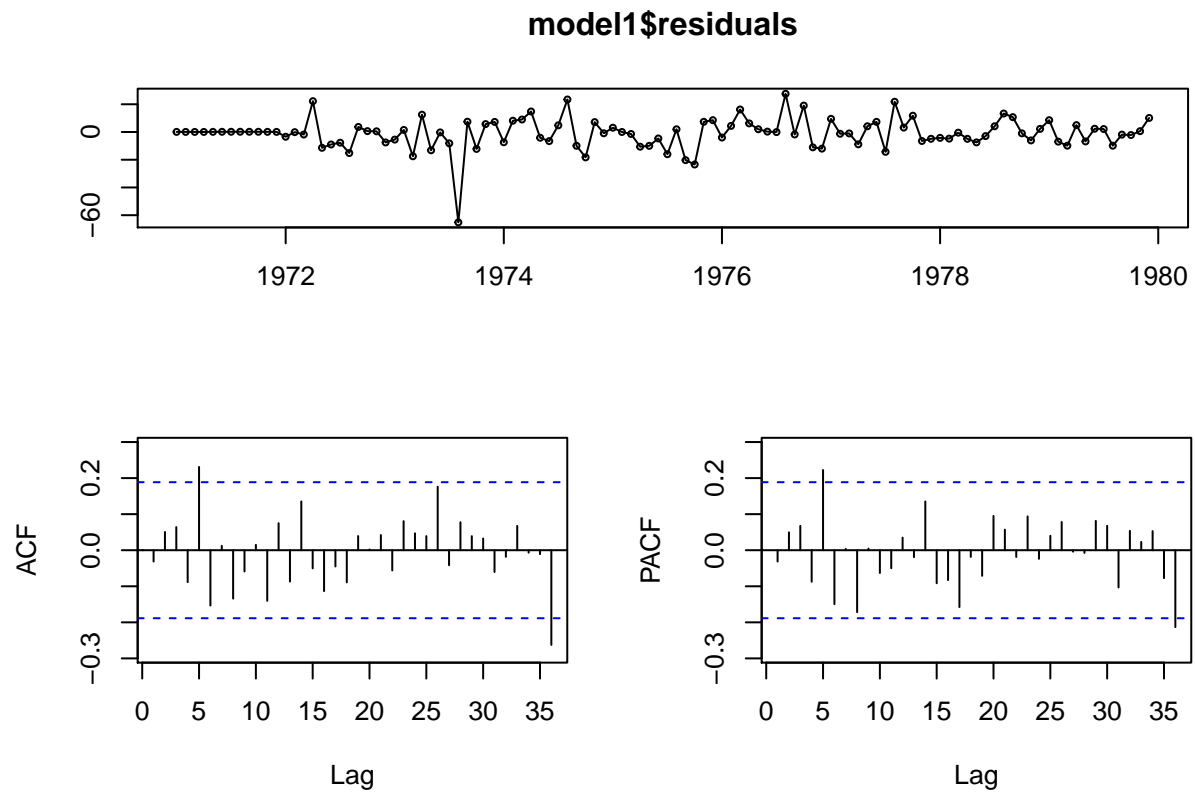
Both AICc and BIC are lower in Model1.

(2 points) Question 5:

Plot the residuals ACF of both models from part 4 and use the Ljung-Box Test with lag 12 to verify your conclusion.

Model 1

```
tsdisplay(model1$residuals)
```

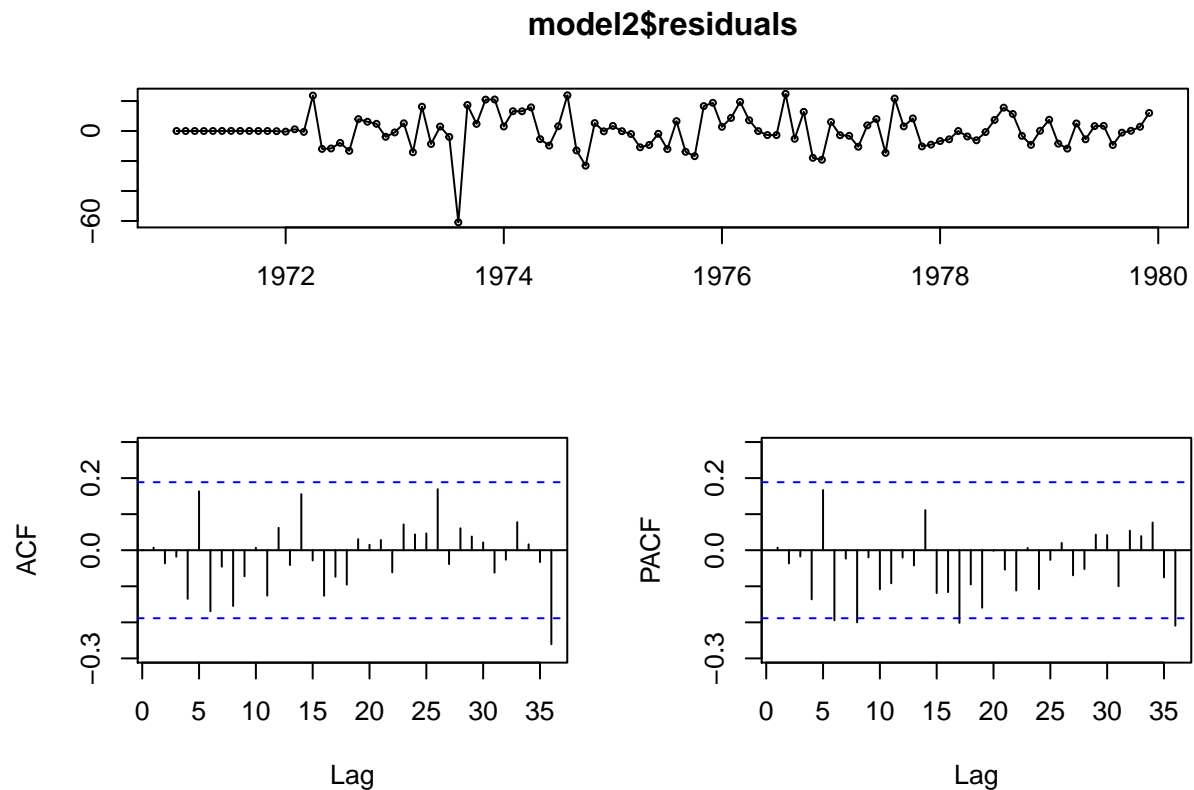


```
Box.test(model1$residuals, type="Ljung-Box",lag=12)
```

```
##
##  Box-Ljung test
##
## data:  model1$residuals
## X-squared = 16.366, df = 12, p-value = 0.1751
```

Model 2

```
tsdisplay(model2$residuals)
```



```
Box.test(model2$residuals, type="Ljung-Box",lag=12)
```

```
##
## Box-Ljung test
##
## data: model2$residuals
## X-squared = 14.81, df = 12, p-value = 0.252
```

Both residuals ACF plots show that there is no significant autocorrelation in the residuals from either model. Box-Ljung test also verified that at 5% confidence level, H_0 : there is no autocorrelation up to lag 12 in the residuals, should be accepted. Both models did a good job of fitting all the important auto-correlations in the time series.

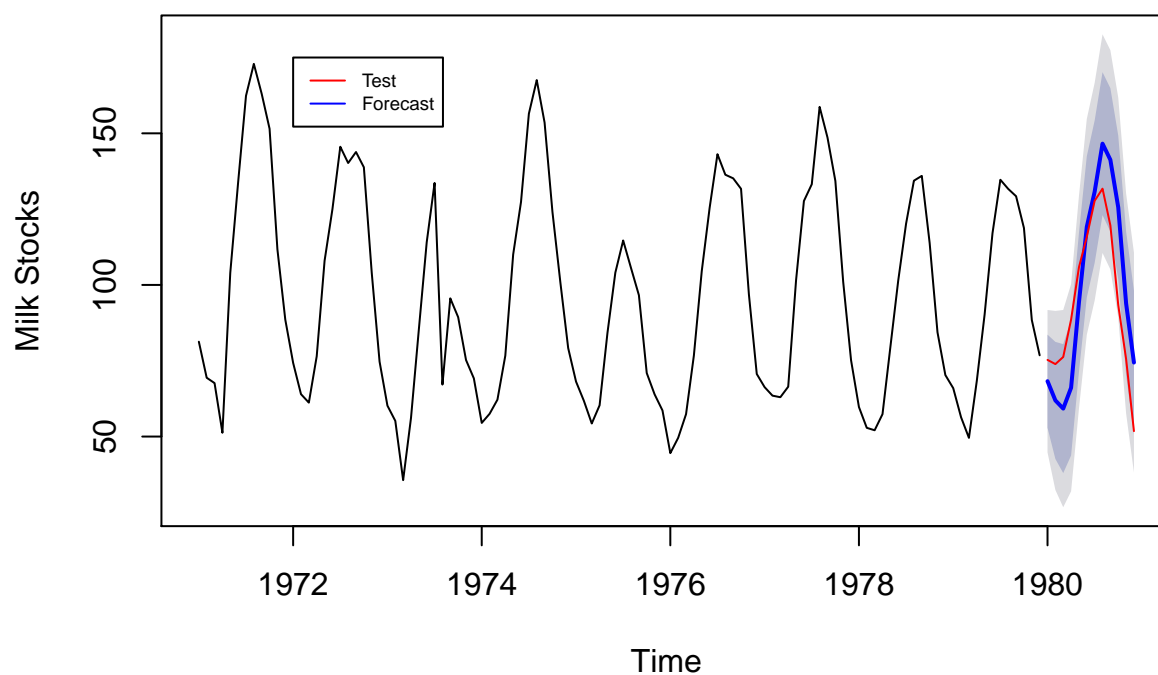
(5 points) Question 6:

Use both models from part 4 and the `h`-period argument in the `forecast()` function to forecast each month of 1980 (i.e., Jan, Feb, ..., Dec.) Plot the test dataset and forecasted values.

Forecast using model 1

```
plot(forecast(model1, 12), xlab="Time",
     ylab="Milk Stocks",main="Model1 Forecast")
lines(x=c(time(forecast(model1, 12)$mean)), y=ts(test, frequency = 12) , col="red")
legend(1972, 175, legend=c("Test", "Forecast"),
      col=c("red", "blue"), lty=1,cex=0.6)
```

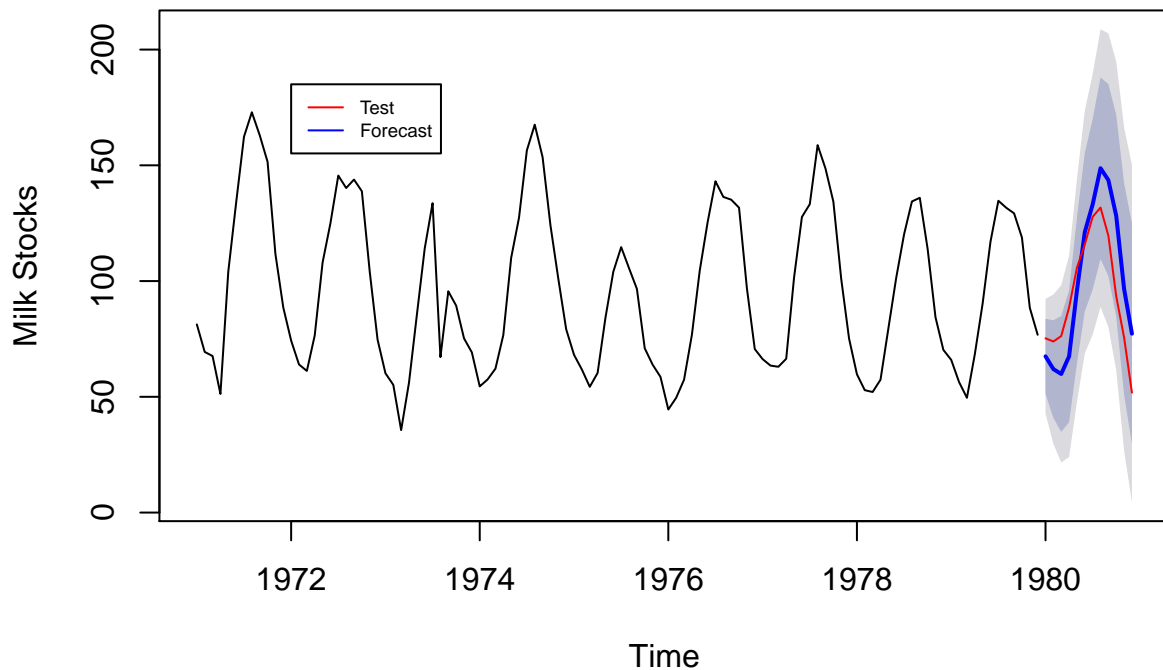

Model1 Forecast



Forecast using model 2

```
plot(forecast(model2, 12), xlab="Time",
     ylab="Milk Stocks",main="Model2 Forecast")
lines(x=c(time(forecast(model2, 12)$mean)), y =ts(test, frequency = 12) , col="red")
legend(1972, 185, legend=c("Test", "Forecast"),
      col=c("red", "blue"), lty=1,cex=0.6)
```

Model2 Forecast



(5 points) Question 7:

Compare the forecast with the actual test data by calculating the Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE). Which model is better to forecast the Manufacturer's Stocks for each month of 1980 (i.e., Jan, Feb, ..., Dec)?

```
m1.fc<-forecast(model1,h=12)$mean
m2.fc<-forecast(model2,h=12)$mean
m1.MSE<-sum((m1.fc-test)^2)/12
m2.MSE<-sum((m2.fc-test)^2)/12
m1.MAPE<-sum((abs(m1.fc-test)/test))/12
m2.MAPE<-sum((abs(m2.fc-test)/test))/12
cbind(m1.MSE,m2.MSE,m1.MAPE,m2.MAPE)
```

```
##          m1.MSE  m2.MSE  m1.MAPE  m2.MAPE
## [1,] 309.5345 352.806 0.1842871 0.1975673
```

Based on MSE and MAPE, model 1 is a better model.