UK Deaths From Bronchitis, Emphysema, and Asthma by Sex

By: Jose Maldonado

### Introduction:

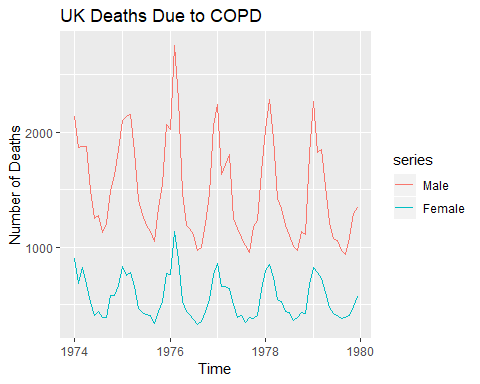
#### The following paper is on deaths in the United Kingdom from 1974 to 1979 due to bronchitis, emphysema, and asthma. All of these are respiratory problems that still exist today and were more severe in the past due to the lack of treatments. Furthermore, all of these illnesses are worsened by smoking. From the data given, we discuss the reasons for the increase and decreases in the deaths as well as our future predictions.

### Exploring the Dataset:

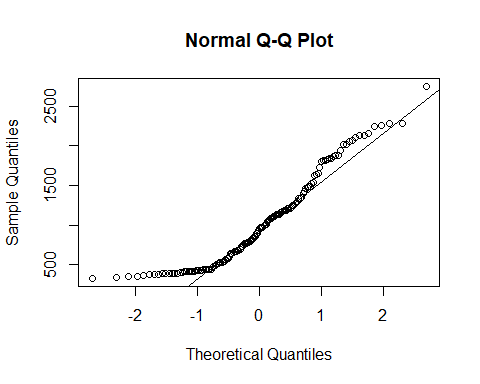
data = read.csv("uk-deaths-from-bronchitis-emphys.csv",header=TRUE, stringsAsFactors = FALSE)  
deathsts = ts(data[,-1], start = 1974, frequency = 12)  
  
#Validating the data, checking that it matches with the excel file, and that it's not out of the normal.  
#deathsts  
summary(deathsts)

## Male Female   
## Min. : 940 Min. : 330.0   
## 1st Qu.:1138 1st Qu.: 411.0   
## Median :1344 Median : 512.0   
## Mean :1496 Mean : 560.7   
## 3rd Qu.:1846 3rd Qu.: 681.5   
## Max. :2750 Max. :1141.0

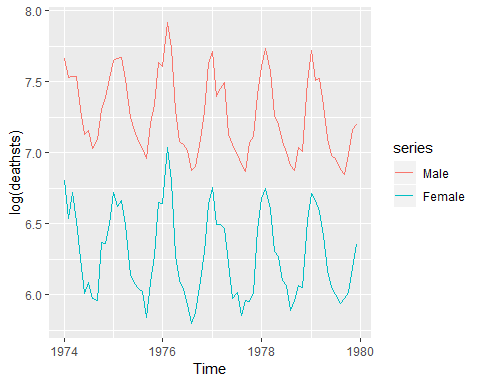
autoplot(deathsts, ylab = "Number of Deaths") + ggtitle("UK Deaths Due to COPD")



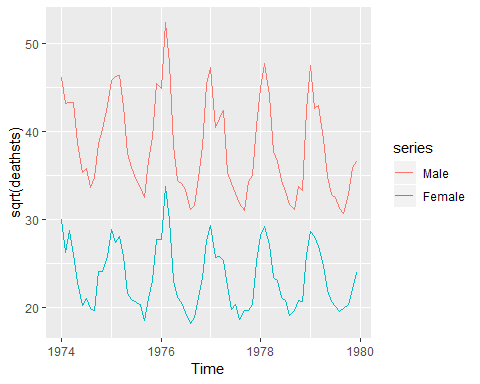
qqnorm(deathsts)  
qqline(deathsts)



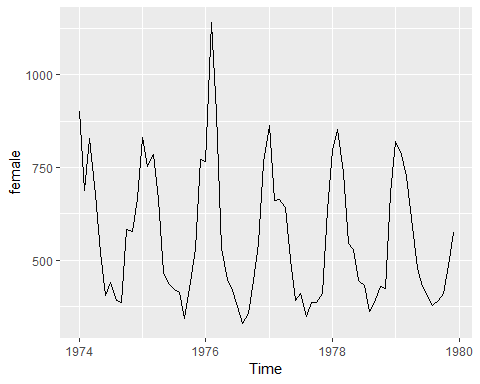
autoplot(log(deathsts))



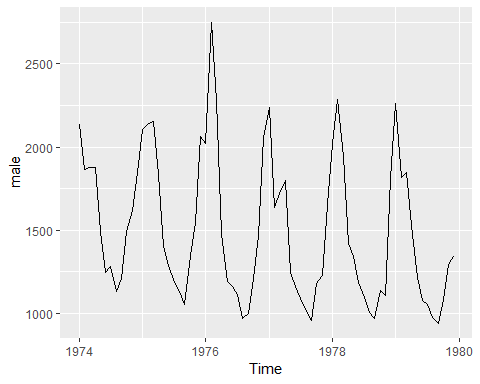
autoplot(sqrt(deathsts))



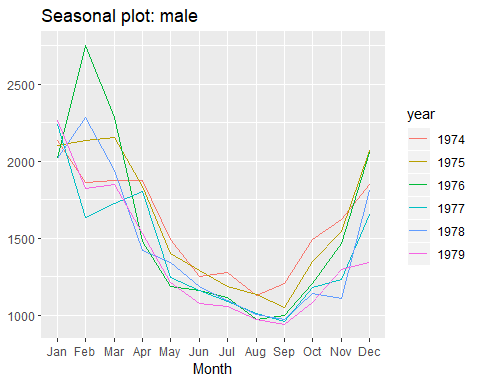
## From the plots, there does seem to be a decreasing trend. There also appears to be similar seasonalities for both the male and female models. The seasonal variation is consistent except for the spike in 1975. Also due to the spike in 1976 we can say that there isn't enough evidence of a cyclic behaviour.  
  
#Slpit the data set into male and female to be able to create appropriate forecast models.  
male = deathsts[, "Male"]  
female = deathsts[, "Female"]  
  
#Making sure the data/plots match the ones from before.  
autoplot(female)



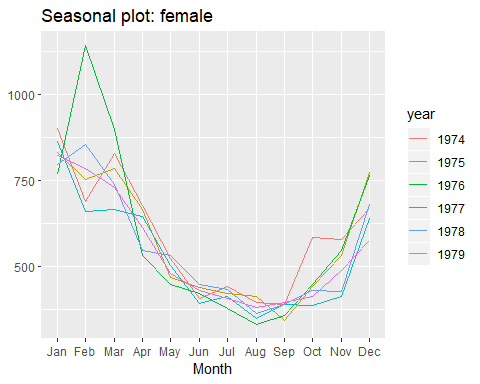
autoplot(male)



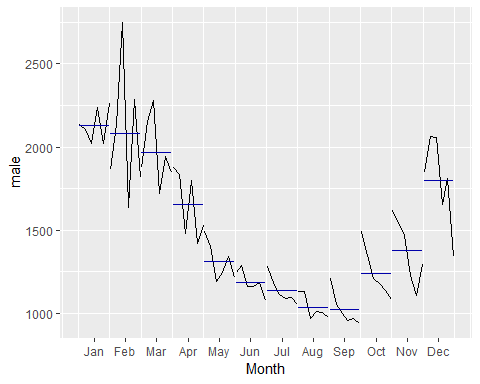
ggseasonplot(male)



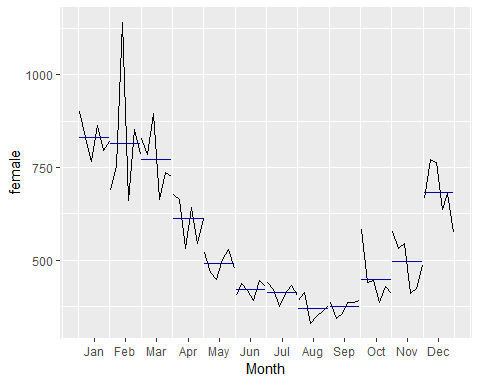
ggseasonplot(female)



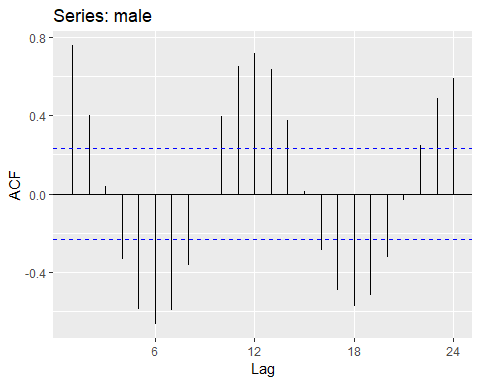
ggsubseriesplot(male)



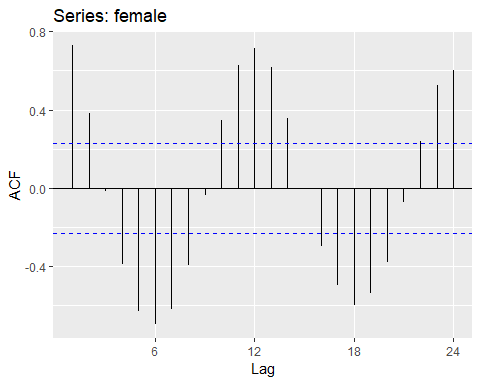
ggsubseriesplot(female)



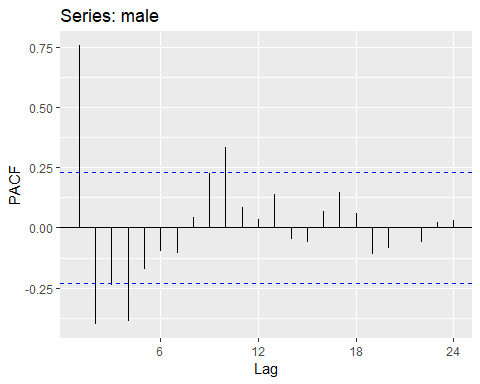
## From the seasonal and subseries plots, we can see that there is a peak in deaths in January and February and a trough in September and August. The data starts to increase in around September until its peak and then continues to decrease for the rest of the year. This could mean that the data would need to be seasonally differenced.  
  
ggAcf(male)



ggAcf(female)



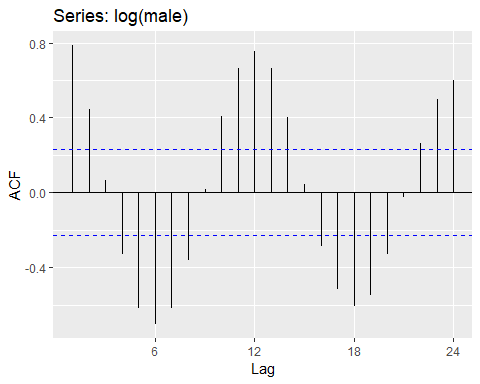
## Neither the male or female models are stationary as seen from the ACF plots therefore the data needs to be seasonally differenced to become stationary. Both the male and female ACF plots show the positive lags slowly decreasing over time indicating there is a decreasing trend.  
  
ggPacf(male)



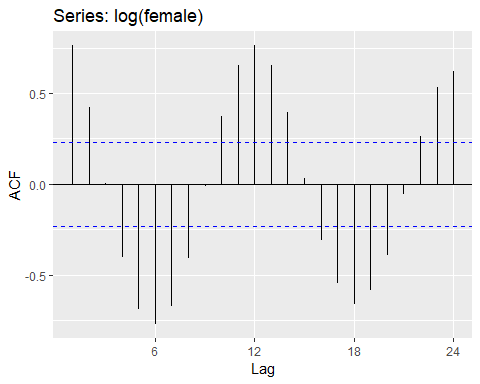
ggPacf(female)



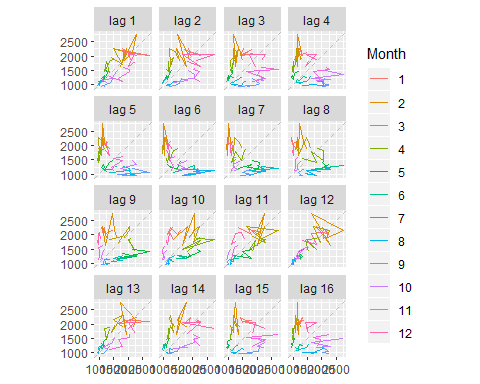
## The PACF plots show spikes at the start which will be helpful when checking our ARIMA model.  
  
ggAcf(log(male))



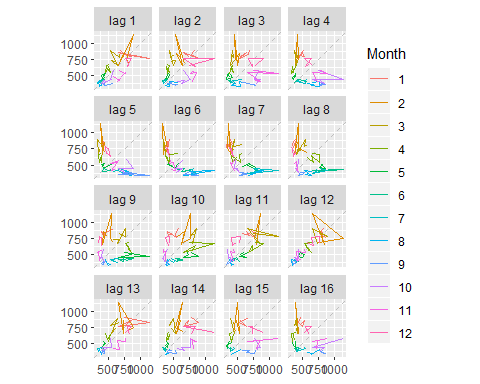
ggAcf(log(female))



gglagplot(male)



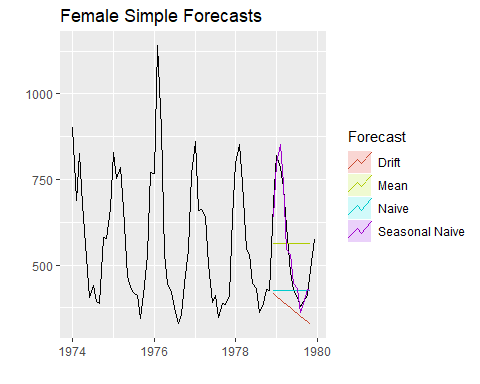
gglagplot(female)



## The lagplots provide more evidence about the seasonality in the winter that is occurring.

### Simple Forecasting Methods:

male.train = subset(male, end = 59)  
male.test = subset(male, start = 60)  
female.train = subset(female, end = 59)  
female.test = subset(female, start = 60)  
## Created a test and train of the dataset where the test set is the last year of data, which is 1979. Training set is everything before.  
  
driftfemale = rwf(female.train, drift = TRUE, h = 12)  
naivefemale = naive(female.train, h = 12)  
meanfemale = meanf(female.train, h = 12)  
snaivefemale = snaive(female.train, h = 12)  
autoplot(female)+  
 autolayer(meanfemale, series = "Mean",PI = FALSE) +   
 autolayer(naivefemale,series = "Naive",PI=FALSE) +   
 autolayer(driftfemale, series="Drift", PI=FALSE) +  
 autolayer(snaivefemale, series = "Seasonal Naive", PI = FALSE) + ggtitle("Female Simple Forecasts")+xlab("")+ylab("")+  
 guides(colour=guide\_legend(title="Forecast"))



## From the graph, the only method that looks reasonable is seasonal naive. However, we will still test all methods to see their results.  
  
accuracy(snaivefemale, female.test)

## ME RMSE MAE MPE MAPE MASE  
## Training set -10.382979 118.50595 75.87234 -3.2781080 12.690536 1.0000000  
## Test set 2.083333 40.39905 34.08333 0.2240799 6.211067 0.4492195  
## ACF1 Theil's U  
## Training set 0.1055140 NA  
## Test set -0.2168053 0.5579733

accuracy(meanfemale, female.test)

## ME RMSE MAE MPE MAPE MASE  
## Training set 5.202759e-14 183.7980 157.3352 -10.04150 29.41344 2.073684  
## Test set -1.202542e+01 158.1777 147.2542 -10.32946 28.27700 1.940816  
## ACF1 Theil's U  
## Training set 0.7190335 NA  
## Test set 0.8551148 2.460467

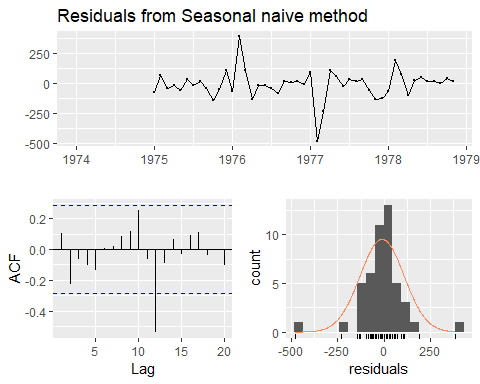
accuracy(naivefemale, female.test)

## ME RMSE MAE MPE MAPE MASE  
## Training set -8.206897 130.4186 97.31034 -3.402663 16.65038 1.282554  
## Test set 125.500000 201.5585 144.16667 16.643733 21.41445 1.900122  
## ACF1 Theil's U  
## Training set 0.1834675 NA  
## Test set 0.8551148 2.111634

accuracy(driftfemale, female.test)

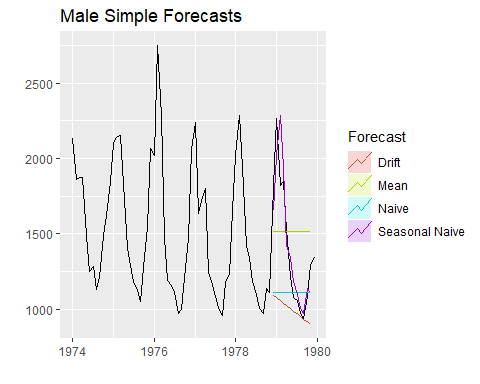
## ME RMSE MAE MPE MAPE MASE  
## Training set 6.143107e-11 130.1601 96.58621 -1.78525 16.38715 1.273009  
## Test set 1.788448e+02 223.9884 178.84483 28.31268 28.31268 2.357181  
## ACF1 Theil's U  
## Training set 0.1834675 NA  
## Test set 0.8377128 2.516585

## Using RMSE and MAPE, it is easy to see that the best method to use here would be seasonal naive.  
  
checkresiduals(snaivefemale)



##   
## Ljung-Box test  
##   
## data: Residuals from Seasonal naive method  
## Q\* = 10.44, df = 11.8, p-value = 0.5606  
##   
## Model df: 0. Total lags used: 11.8

## From these results, we can see that the residuals are not correlated with a p-value of 0.5606. They also have a zero mean and are normally distributed. This is an acceptable forecasting method for the female dataset.  
  
driftmale = rwf(male.train, drift = TRUE, h = 12)  
naivemale = naive(male.train, h = 12)  
meanmale = meanf(male.train, h = 12)  
snaivemale = snaive(male.train, h = 12)  
autoplot(male)+  
 autolayer(meanmale, series = "Mean",PI = FALSE) +   
 autolayer(naivemale,series = "Naive",PI=FALSE) +   
 autolayer(driftmale, series="Drift", PI=FALSE) +  
 autolayer(snaivemale, series = "Seasonal Naive", PI = FALSE) + ggtitle("Male Simple Forecasts")+xlab("")+ylab("")+  
 guides(colour=guide\_legend(title="Forecast"))



## Similar to the female model, the only reasonable method appears to be seasonal naive.   
  
accuracy(snaivemale, male.test)

## ME RMSE MAE MPE MAPE MASE  
## Training set -40.40426 274.7425 176.9574 -4.311901 10.916817 1.000000  
## Test set -22.00000 180.1324 137.5000 -2.190195 8.976911 0.777023  
## ACF1 Theil's U  
## Training set 0.1173297 NA  
## Test set -0.1468334 0.6802945

accuracy(naivemale, male.test)

## ME RMSE MAE MPE MAPE MASE  
## Training set -17.65517 289.2884 215.6552 -2.629016 13.90122 1.218684  
## Test set 299.00000 511.7159 369.5000 14.662431 21.82633 2.088073  
## ACF1 Theil's U  
## Training set 0.2860774 NA  
## Test set 0.7902708 1.777724

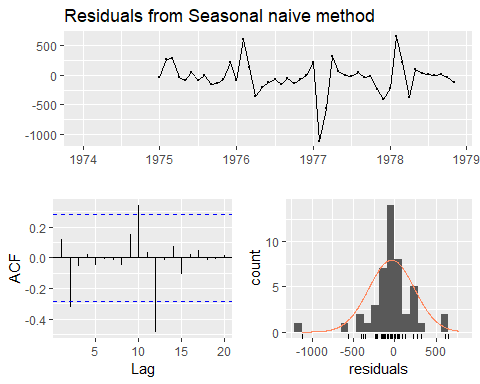
accuracy(meanmale, male.test)

## ME RMSE MAE MPE MAPE MASE  
## Training set 6.552574e-14 433.9905 375.3864 -7.898863 25.81201 2.121337  
## Test set -1.072542e+02 428.9005 389.0424 -16.570676 30.70980 2.198508  
## ACF1 Theil's U  
## Training set 0.7570063 NA  
## Test set 0.7902708 2.38249

accuracy(driftmale, male.test)

## ME RMSE MAE MPE MAPE MASE  
## Training set -1.883674e-11 288.7492 214.4376 -1.365251 13.66041 1.211803  
## Test set 4.137586e+02 552.7707 413.7586 24.499416 24.49942 2.338181  
## ACF1 Theil's U  
## Training set 0.2860774 NA  
## Test set 0.7607672 1.995659

## From these tests, using RMSE and MAPE, the best method, clearly, is seasonal naive.  
  
checkresiduals(snaivemale)

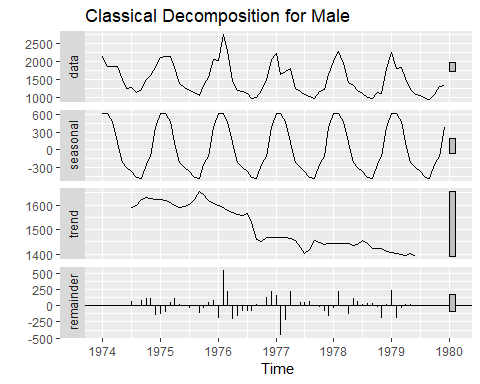


##   
## Ljung-Box test  
##   
## data: Residuals from Seasonal naive method  
## Q\* = 15.28, df = 11.8, p-value = 0.2146  
##   
## Model df: 0. Total lags used: 11.8

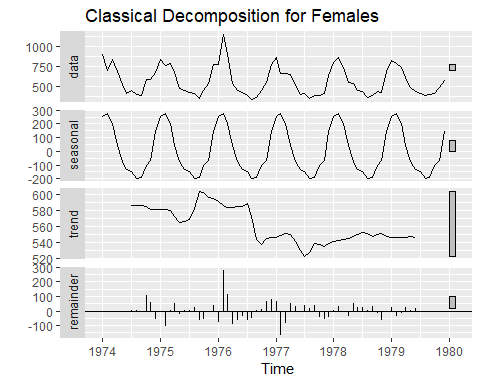
## With a p-value of 0.21 the residuals seem to be correlated even though they appear to have a 0 mean. Also, the residuals are some-what normally distributed. Ultimately this is not an acceptable forecasting method, there is a better one.

### Decomposition of Male and Female Model Along with STL Forecast:

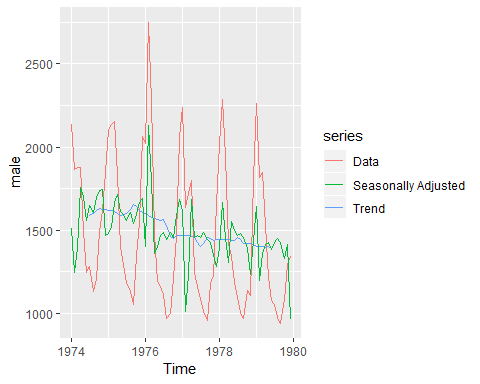
decomale = decompose(male, type = "additive")  
autoplot(decomale, main = "Classical Decomposition for Male")



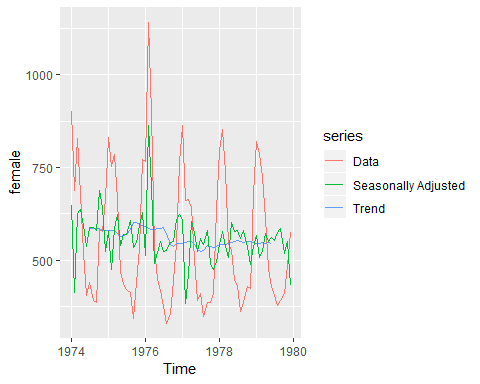
## From the classical decomposition, we can see that there is a small decreasing trend. There is more evidence of the seasonality. However, the important information from the decomposition is the spike in 1976 in the remainder component.  
decofemale = decompose(female, type = "additive")  
autoplot(decofemale, main = "Classical Decomposition for Females")



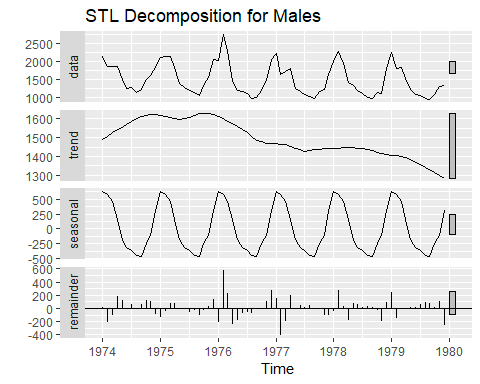
## The spike occurs in the remainder component in 1976 as well for the females.  
autoplot(male, series = "Data") +  
 autolayer(trendcycle(decomale), series = "Trend") +  
 autolayer(seasadj(decomale), series = "Seasonally Adjusted")



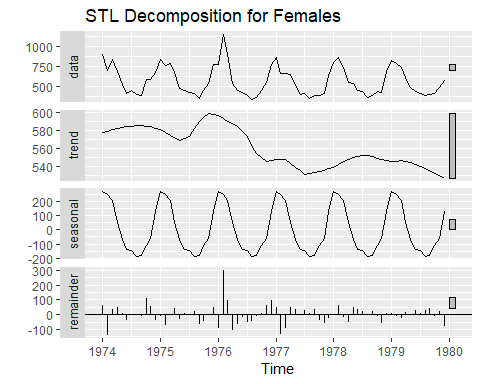
autoplot(female, series = "Data") +   
 autolayer(trendcycle(decofemale), series = "Trend") +  
 autolayer(seasadj(decofemale), series = "Seasonally Adjusted")



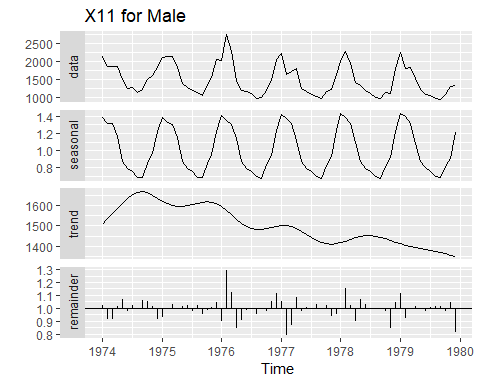
##This lets us see the decreasing trend and the seasonality along with the original data.  
  
stlmale = stl(male, s.window = "periodic")  
autoplot(stlmale, main = "STL Decomposition for Males")



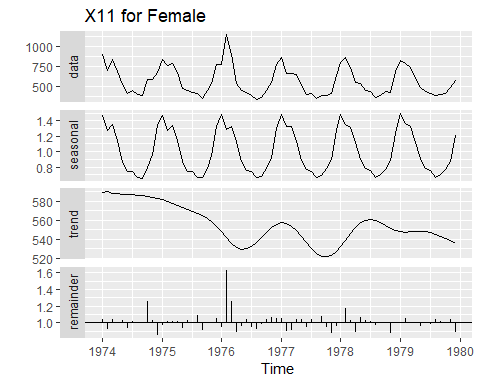
stlfemale = stl(female, s.window = "periodic")  
autoplot(stlfemale, main = "STL Decomposition for Females")



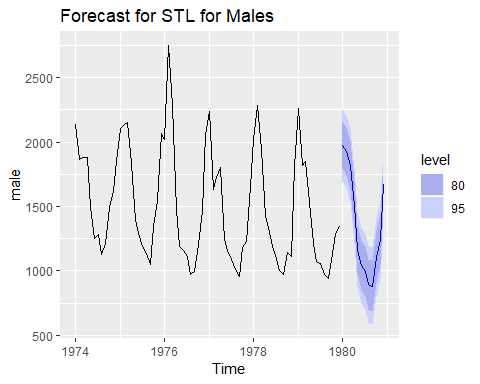
## The STL decomposition of male and female data sets shows us the same results we found in the classical decomposition. Decreasing trend and heavy seasonality patterns.  
  
x11male = seas(male, x11 = "")  
autoplot(x11male, main = "X11 for Male")



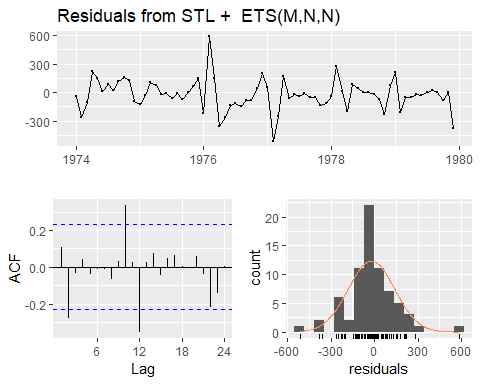
x11female = seas(female, x11 = "")  
autoplot(x11female, main = "X11 for Female")



## The X11 decomposition just more clearly shows the results already found by the decomposition methods we used above.  
  
stlfcmale = forecast(stlmale, h = 12)  
autoplot(stlfcmale, main = "Forecast for STL for Males")

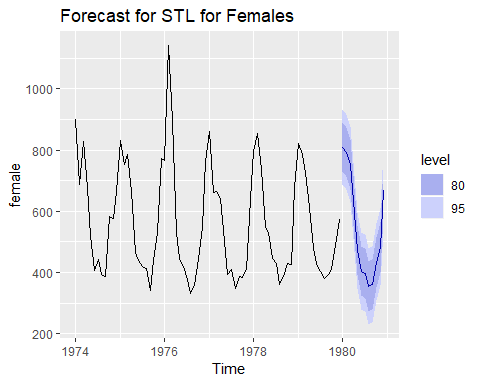


checkresiduals(stlfcmale, main = "Residuals STL Forecast for Males")

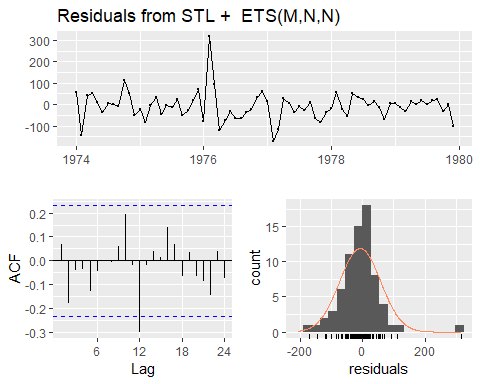


##   
## Ljung-Box test  
##   
## data: Residuals from STL + ETS(M,N,N)  
## Q\* = 28.763, df = 12.4, p-value = 0.005253  
##   
## Model df: 2. Total lags used: 14.4

## The STL forecast for males has residuals that are correlated so it is thrown out.  
  
stlfcfemale = forecast(stlfemale, h = 12)  
autoplot(stlfcfemale, main = "Forecast for STL for Females")



checkresiduals(stlfcfemale, main = "Residuals STL Forecast for Females")



##   
## Ljung-Box test  
##   
## data: Residuals from STL + ETS(M,N,N)  
## Q\* = 16.154, df = 12.4, p-value = 0.206  
##   
## Model df: 2. Total lags used: 14.4

## The STL forecast for females has uncorrelated residuals and a zero mean. However, the seasonal naive still appears to be a better forecast. Still, this is not a bad method.

### ARIMA Forecasting Method

ndiffs(male)

## [1] 0

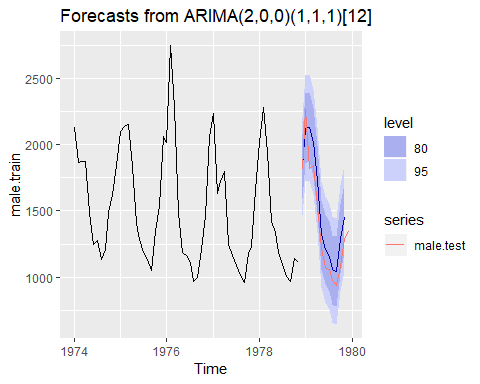
nsdiffs(male)

## [1] 1

## The data only needs to be seasonally differenced once.  
  
malearima = auto.arima(male, seasonal = TRUE)  
malearima

## Series: male   
## ARIMA(2,0,0)(1,1,1)[12] with drift   
##   
## Coefficients:  
## ar1 ar2 sar1 sma1 drift  
## 0.2795 -0.3450 -0.3652 -0.7604 -4.3863  
## s.e. 0.1351 0.1414 0.1766 0.3922 0.7055  
##   
## sigma^2 estimated as 26690: log likelihood=-397.11  
## AIC=806.21 AICc=807.8 BIC=818.78

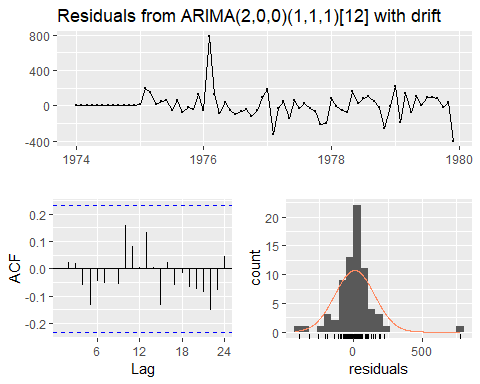
## The model found makes sense when looking at the PACF and ACF plots for males. The model is fairly simple so this should be suffice.  
  
malearimatrain = arima(male.train, order = c(2,0,0),   
 seasonal = c(1,1,1))  
malearimatest = Arima(male.test, model = malearimatrain)  
malearimatrain %>%  
 forecast(h = 12) %>%  
 autoplot() + autolayer(male.test)



## From the plot, it can be seen that the arima forecasts is very close to the actual data.  
  
accuracy(malearimatest)

## ME RMSE MAE MPE MAPE MASE  
## Training set -20.75855 79.54796 23.35977 -1.552671 1.737286 0.04959612  
## ACF1  
## Training set -0.006072987

checkresiduals(malearima)



##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(2,0,0)(1,1,1)[12] with drift  
## Q\* = 6.7477, df = 9.4, p-value = 0.6994  
##   
## Model df: 5. Total lags used: 14.4

## From these results, a p-value of 0.70 was found from the Ljung-Box test showing the residuals are uncorrelated. The residuals are not normally distributed but this appears to be the best forecasting method.  
  
ndiffs(female)

## [1] 0

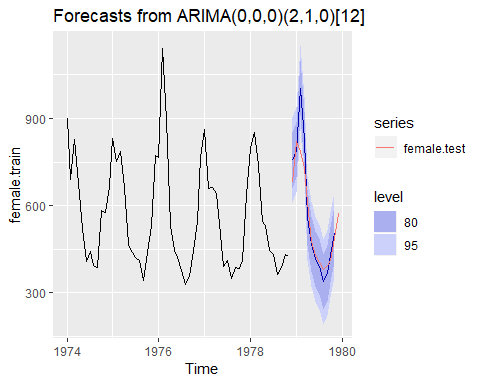
nsdiffs(female)

## [1] 1

## The data only needs to be seasonally differenced one time.  
  
femalearima = auto.arima(female, seasonal = TRUE)  
femalearima

## Series: female   
## ARIMA(0,0,0)(2,1,0)[12] with drift   
##   
## Coefficients:  
## sar1 sar2 drift  
## -0.8721 -0.4950 -0.9642  
## s.e. 0.1266 0.1298 0.4139  
##   
## sigma^2 estimated as 5887: log likelihood=-349.88  
## AIC=707.76 AICc=708.49 BIC=716.14

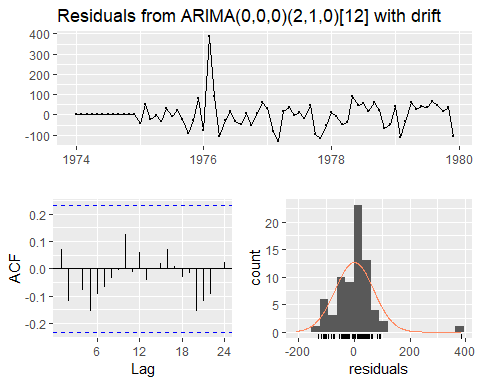
femalearimatrain = arima(female.train, order = c(0,0,0),   
 seasonal = c(2,1,0))  
femalearimatest = Arima(female.test, model = femalearimatrain)  
femalearimatrain %>%  
 forecast(h = 12) %>%  
 autoplot() + autolayer(female.test)



## From the plot, it can be seen that the arima forecasts is very close to the actual data.  
  
accuracy(femalearimatest)

## ME RMSE MAE MPE MAPE MASE  
## Training set -4.511335 18.10638 5.527641 -0.7821677 0.9667829 0.0526442  
## ACF1  
## Training set -0.005463314

checkresiduals(femalearima)

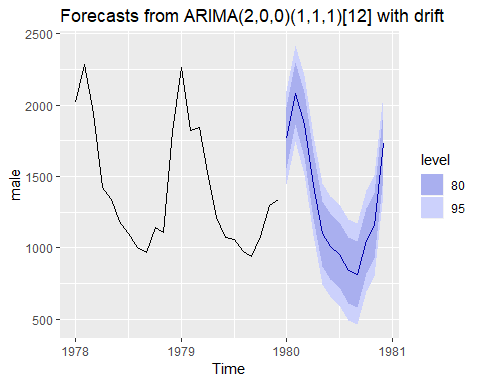


##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(0,0,0)(2,1,0)[12] with drift  
## Q\* = 6.8378, df = 11.4, p-value = 0.8363  
##   
## Model df: 3. Total lags used: 14.4

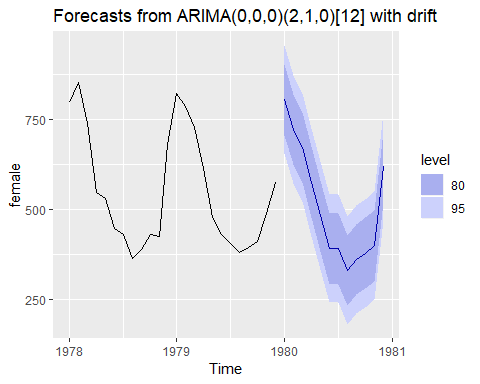
## From these results, we can see that the ARIMA model is very good with a p-value of 0.84. Also, the accuracy of the test was very good using MAPE and MASE making this our best model for the female dataset.

### Forecasts for the Next Year From ARIMA Model:

fitmale = forecast(malearima, h = 12)  
fitfemale = forecast(femalearima, h = 12)  
autoplot(include = 24, fitmale)



autoplot(include = 24, fitfemale)



### Conclusion:

#### The best forecasting methods for the both models is ARIMA. For the male model, we will use an ARIMA(2,0,0)(1,1,1)[12] with drift and the female is ARIMA(0,0,0)(2,1,0)[12] with drift. The models are pretty simple with both being only seasonally differenced once. From these forecasts, we expect the number of deaths to increase in the winter time as explained by the model because of flare ups of these respiratory problems. Furthermore, the only real difference between the male and female model is that males have a larger amount of deaths. Chronic Obstructive Pulmonary Disease (COPD) has been proven to worsen in the winter for patients causing an increase in the number of deaths. However, the reason for more male deaths than female deaths remains unknown. It does not appear to be the correlation of deaths from the Vietnam War because the war ended in 1975. Recent studies have shown that women are more likely to have respiratory diseases than men, which does not support this data. It may have been due to the generations of World War veterans dying and several soldiers were smokers. However, further research or more information would be needed to prove this. From research it was found that starting 1962 the Ministry of Health began to change their stance around smoking not leading to serious health problems. A combination of that and campaings being ran to educate the public about the harm that smoking has on your health also explain why the trend was decreasing.