PSTAT 174 Final Project
Time Series Analysis

Milk Production January 1962 - December 1975

Authors:
Whitney Pemberton
Melanie Flandes
Tomio Sugawara
Runchen Zhang
Jensen Newlin

Supervised By:
Professor Sudeep Bapat

Contents

Abstract

Due to the necessity and mass consumption of milk by people in U.S. society the production of Milk has always been important topic to consider. The goal of this project is to apply time series analysis on milk production data, to predict future monthly production. Exploratory analysis methods such as Box Cox and differencing is conducted on the data. A SARIMA model is fitted for the data and used to predict future observations.

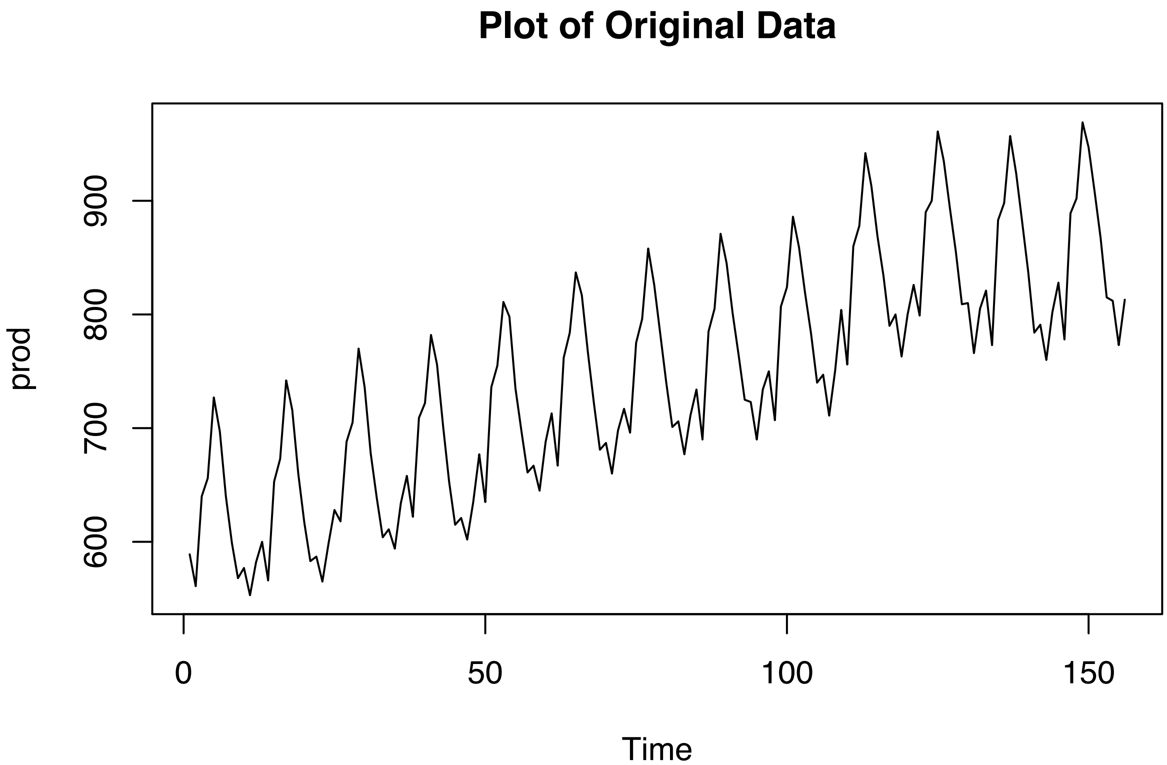
Introduction

Over the years, the United States population has increased drastically. Due to having a fast growing population, the production rates for all food products have also increased in order to meet the needs of the population. One product that is in constant production due to its high demand is milk. It is important to analyze any trends and predict any outcomes for future reference. The data we have found, provides us with the monthly milk production (pounds per cow) in the United States starting from January 1962 to December 1975. We are interested in conducting time series analysis on this milk production dataset because we want to create forecasts for future milk production. Forecasting for milk production is important because it is necessary for producers to have a grasp on how much milk is needed to be produced in order to meet the future demand of consumers. To conduct this analysis we use R and R studio since it has many built in and installable statistical package which include functions needed for conducting time series analysis. We make use of Box-Cox transformation, detrending and deseasonalizing to help us create find a stationary model. After which we fit a SARIMA model that is used for forecasting. A SARIMA(1, 1, 1) x (1, 1, 1)12 model is fitted on our 156 month milk production data.

Exploratory Analysis

Initial Data

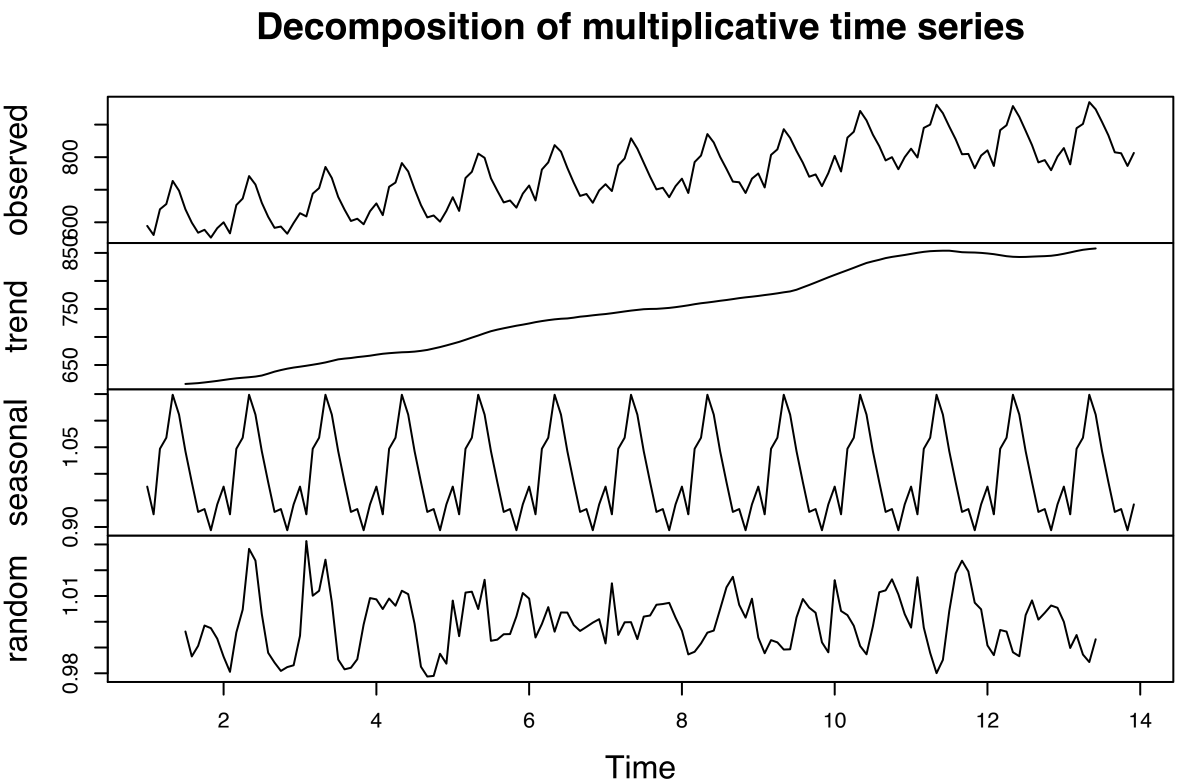
The data set we are using has two variables that are being analyzed in the data set are: date on a monthly basis and pounds per cow of milk produced. There are 168 observations in the data set but we leave out the last 12 months of data so we can access performance later on. Here is the plot of the original non transformed data:



The data shows a presence of variations that occur at specific regular intervals which suggests there is seasonality. And you can see a positive linear trend and a changing variance

Data Decompisition

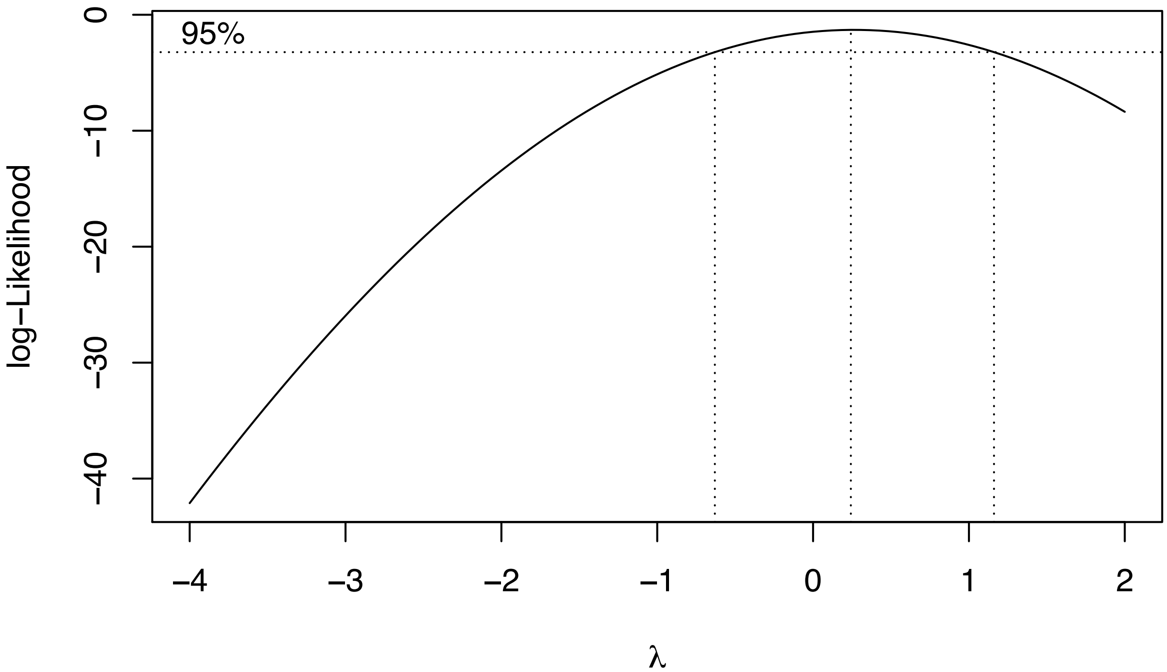
The trend and seasonality is further confirmed by our decomposition plots below.



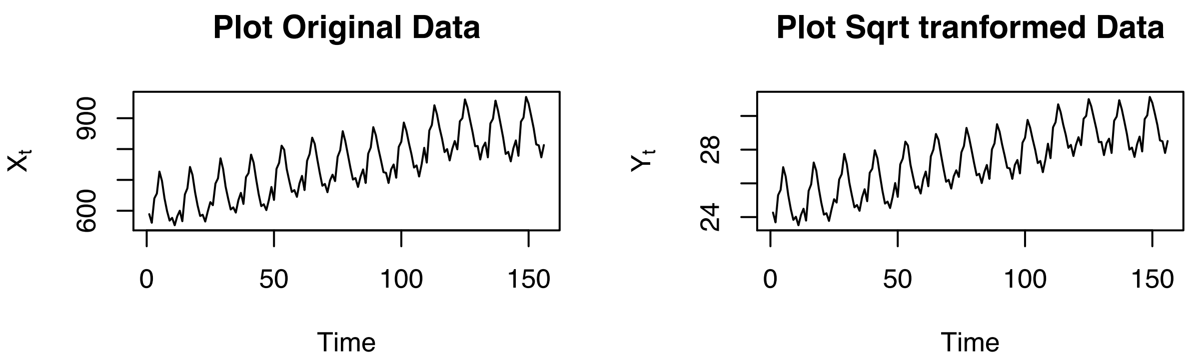
Data Transformation

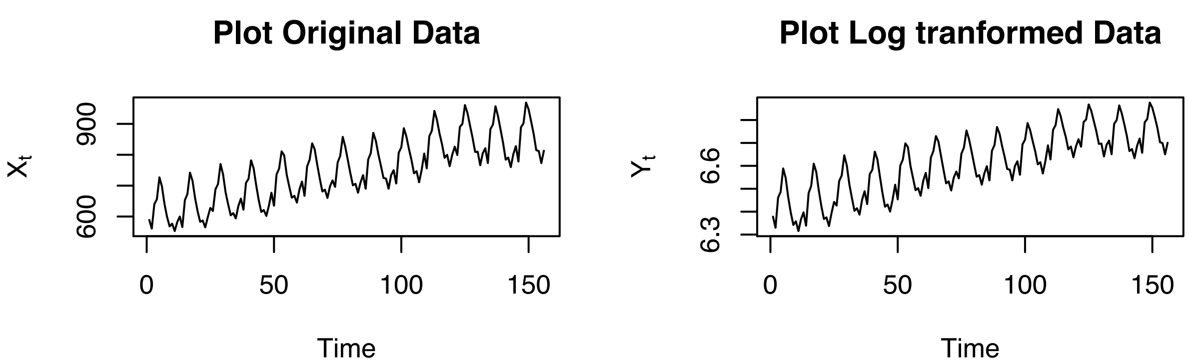
Changing Variance

We are going to transform the data to try and make it more stationary. Also we will be removing trend by differencing the data and take the log of the seasonal difference to remove seasonality. Even though the variance looks to be constant we apply the box cox transformation. After taking the box cox transformation, the lambda value calculated is .24 which is close to 0.5 and 0. So we thought we should try to do a sqrt transformation or a log transformation. Here is the plot of the original data vs. Sqrt transformed data and as you can see there is not much of a difference between them. The same goes with the log transformed data.



$\lambda = 0.242424242424242$





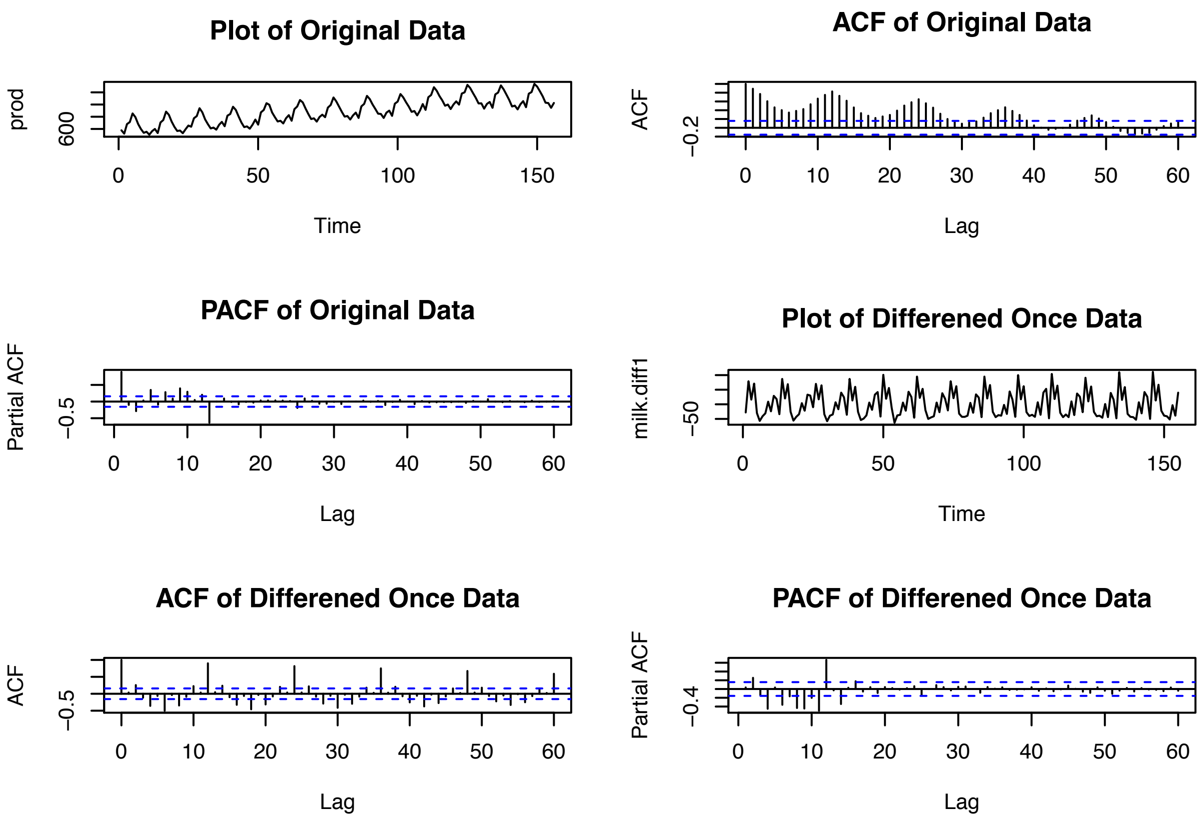
Since both of the transformed data is not much different from the original. We are going to stop pursuing these transformations and just work with the original non transformed data set

Detrending

We plot the ACF and PACF for our data first, and both the plots suggest non-stationarity of our data, implying that we need to apply differencing techniques to detrend and deseasonalize our raw data. The time series plot of our original data shows an obvious upward trend. So we decide to difference our data at lag=1. Comparing the original data and the differenced data, the variance goes down from 10055.58 to 2042.924, and this is a good sign that we control the range of our data to some extent and that we are heading towards stationarity. We plot the new time series plot, noticing that the upward trend vanishes, but there still exists a strong periodic trend, namely seasonality. We also plot the detrended data's ACF and PACF, from which we find a spike pattern existing at lag 12, 24, 36,… in the ACF plot and numbers of spikes before lag 12 in the PACF plot. The two findings suggest that the seasonal component has a period of 12.

$$Variance(Original) = 10055.5842431762$$

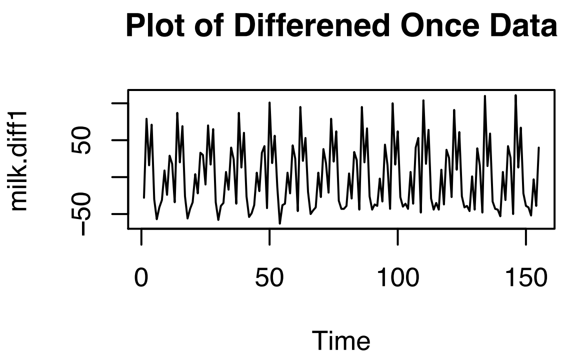
$$Variance(DifferencedOnce) = 2042.92392124005$$

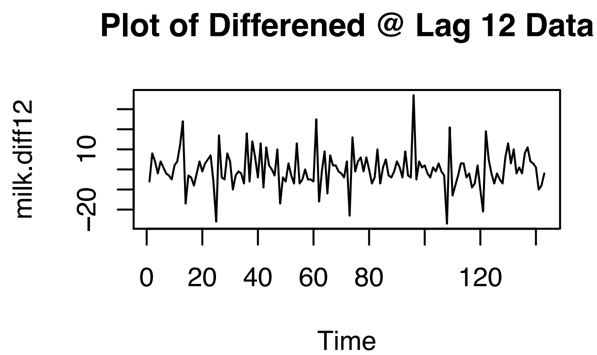


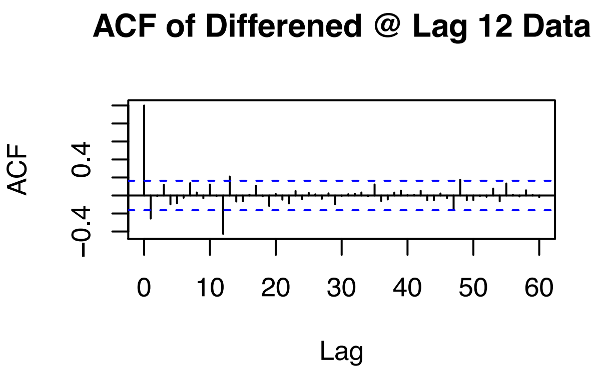
Deseasonalizing

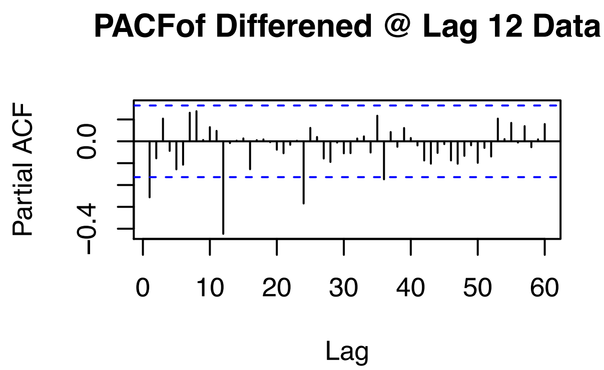
$$Variance(DifferencedOnce) = 2042.92392124005$$

$$Variance(Differenced@Lag12) = 83.0422535211268$$









We then proceed to try and remove seasonality from our data. To do this we difference our data again at lag=12. The variance goes down significantly, from 2042.924 to 83.04225, indicating that the differencing was a reasonable option. Plotting the data differenced twice, we now see that the seasonality is removed and the spread of observations is much narrower than it is before differencing.

Now, through obtaining ACF and PACF plots for the detrended and deseasonalized data, we may come to the conclusion that the ACF suggests that the data is now stationary and that the cut-off is at lag 12 and at lag 1 so that we obtain a Q = q = 1. Also, the PACF suggests that there exist cut-off at lag 1 so that p = 1 and cut-off at lag 12, 24, and 36 so that P = 1,2 or 3.

Model Selection

Our decision of parameters will be: d = 1, D = 1, P = (1,2,3), Q = 1, p = 1, q = 1. And the selection of parameters brings up our three final model candidates, stated as below:

$$SARIMA(1,1,1)x(1,1,1)_{12}$$

$$SARIMA(1,1,1)x(2,1,1)_{12}$$

$$SARIMA(1,1,1)x(3,1,1)_{12}$$

We transform the above expressions into R-code and name them as fit1, fit2, and fit3 correspondingly. When calling them separately, they report their own AIC of 991.93, 992.83 and 994.87, with fit1 having the smallest AIC. Intuitively, we reckon that fit1 = SARIMA (1,1,1) x (1,1,1) 12 will be our choice of final model. But before we go any further on determining the final model, we would like to check the causality and invertibility of the models. With all models proven to be both causal and invertible, we may now proceed to choose the final model.

$fit1 : SARIMA(1, 1, 1)x(1, 1, 1)_{12}$

```
##
## Call:
## arima(x = prod, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1), period = 12),
##     method = "ML")
##
## Coefficients:
##           ar1      ma1     sar1     sma1
##       -0.0937  -0.1723  -0.0468  -0.5845
## s.e.   0.2732   0.2668   0.1214   0.0978
##
## sigma^2 estimated as 53.97:  log likelihood = -490.97,  aic = 991.93
```

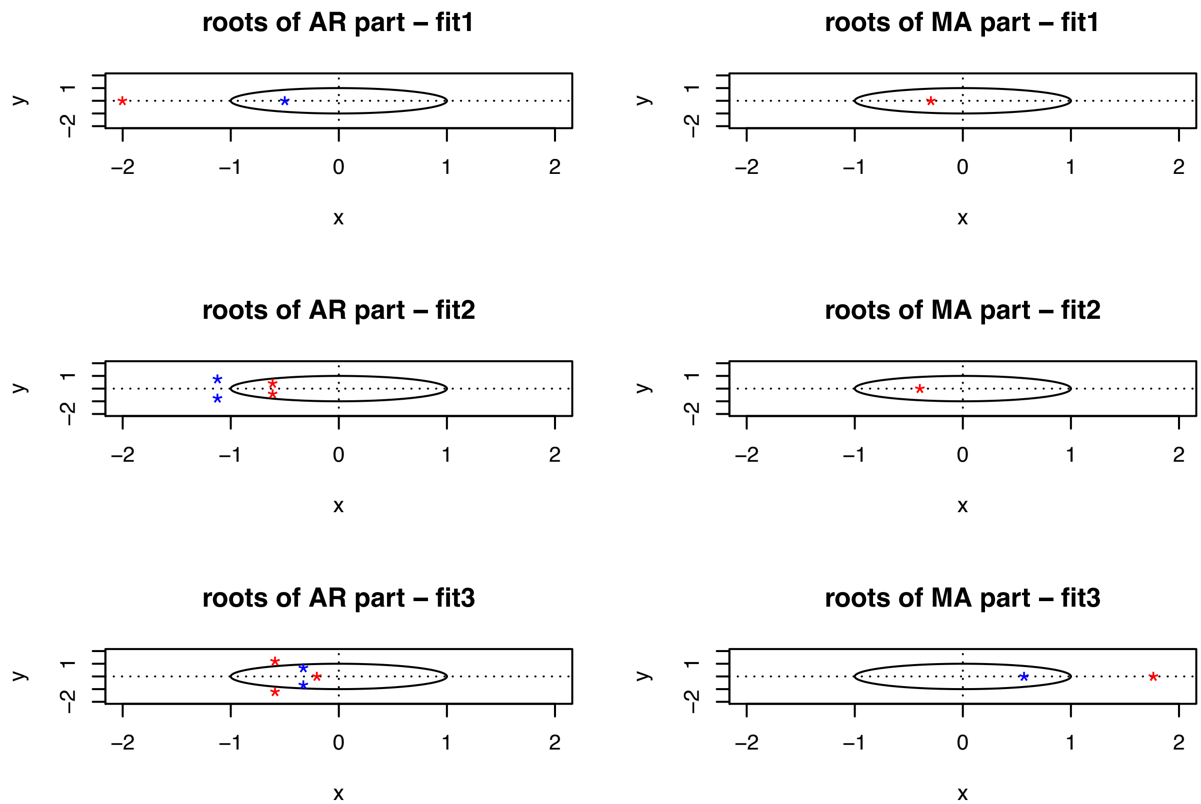
$fit2 : SARIMA(1, 1, 1)x(2, 1, 1)_{12}$

```
##
## Call:
## arima(x = prod, order = c(1, 1, 1), seasonal = list(order = c(2, 1, 1), period = 12),
##     method = "ML")
##
## Coefficients:
##           ar1      ma1     sar1     sar2     sma1
##       -0.0721  -0.1880  -0.1612  -0.1317  -0.4741
## s.e.   0.2346   0.2351   0.1612   0.1196   0.1531
##
## sigma^2 estimated as 53.39:  log likelihood = -490.42,  aic = 992.83
```

$fit3 : SARIMA(1, 1, 1)x(3, 1, 1)_{12}$

```
##
## Call:
## arima(x = prod, order = c(1, 1, 1), seasonal = list(order = c(3, 1, 1), period = 12),
##     method = "ML")
##
## Coefficients:
##           ar1      ma1   sar1     sar2     sar3     sma1
##       -0.1241  -0.0805  -0.33  -0.1079  -0.1422  -0.3547
## s.e.      NaN      NaN    NaN   0.0019   0.0043      NaN
##
## sigma^2 estimated as 53.34:  log likelihood = -490.44,  aic = 994.87
```

Causaulity and Invertibility



$$AICc(fit1) = 992.369515059726$$
$$AICc(fit2) = 993.447969188918$$
$$AICc(fit3) = 995.701022126925$$

With all models proven to be both causal and invertible, we may now proceed to choose the final model. We adopt the Second-order Akaike Information Criterion ($AICc$) to further help us with choosing a best model. From running the AICc function we found that our AICc for Model 1 (fit1) to be 992.36. For Model 2 AICc (fit2) was 993.447 and for Model 3 (fit3) AICc was 995.70. We found that AIC and AICc were lowest for Model1 thus we have decided to use $fit1 : SARIMA(1,1,1)x(1,1,1)_{12}$ as our final model.
$Model : SARIMA(1,1,1)x(1,1,1)_{12}$

$$(1 - .0937B)(1 - .0468B^{12}) \bigtriangledown_{12} \bigtriangledown X_t = (1 - .1723B)(1 - .5845B^{12})Z_t$$

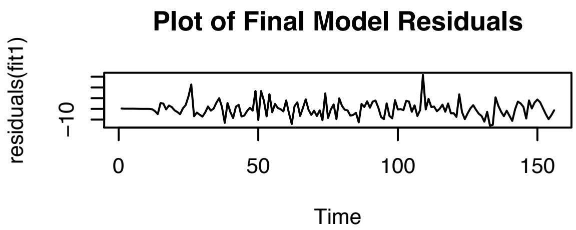
$$where\ Z_t \sim WN(0, 49.814)$$

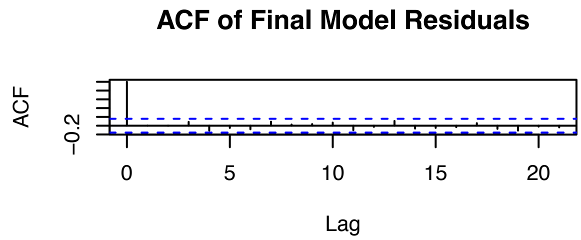
Diagnostics

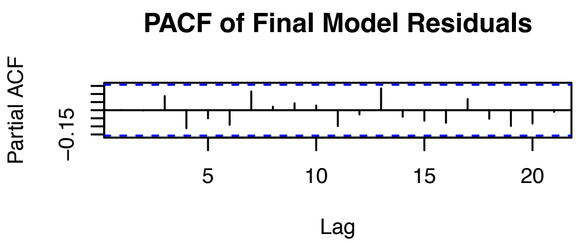
Normality of Residuals

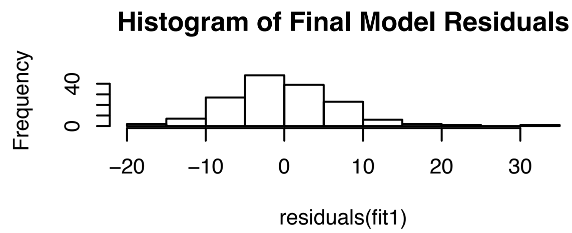
To ensure that our final model satisfies all necessary assumptions, we conduct the diagnostics check by taking the characteristics of residuals of fit1 into account. We plot the ACF and PACF of residuals and both plots seem satisfying. Then we plot the histogram of the residuals, from which we are able to find out if there are any problems regarding the distribution of residuals. The histogram indicates a slight skew towards the right, but for the most part it looks like the residuals are normally distributed.

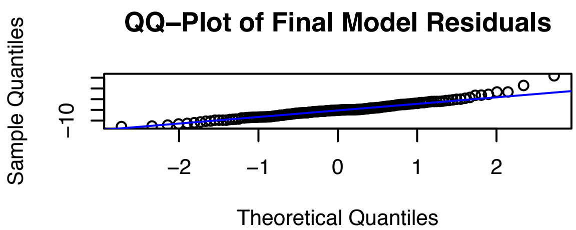
To find out if the residuals are indeed normally distributed, we then apply a more accurate way to see if the normality assumption is met, i.e. Normal Q-Q Plot. Similar to the histogram, the plot seems a bit off the trend around the right tail part, with a few observations deviating from the main track of the observations. However, after discreet consideration, we deem these observations as simple outliers, considering that the major portion of the points is normally distributed.











Diagnostic Tests

We would like to adopt the Box-Pierce and the Ljung-Box tests on our model. Before testing, it is crucial that we set up and declare relevant parameters. We decide the parameters as follow: lag = h = sqrt(n), n = 156, sqrt(156) = approx 12, p = 1, q = 1 , fitdf = 1+1 = 2.

We first test the Box-Pierce test with lag = 12, fitdf = 2 and the reported test statistics, namely the p-value shows a value of 0.558. Then we test the Ljung-Box test with lag = 12, fitdf = 2, with a resulting p-value = 0.5091. Finally, we tested the Ljung-Box test with a fitdf = 0, and this time the p-value = 0.682. To conclude, we have p- values for all the box tests greater than .05, indicating that the autocorrelations are within the lags 1 through 11. Also, we may conclude that the residuals are independent.

What is more, the Shapiro Wilk test reports a p-value of 0.0002331, which is less than .05, indicating a rejection of assumption of normality but eventually regarded acceptable considering our histogram and Normal Q-Q Plot are a bit skewed.

```
##
##  Box-Pierce test
##
## data:  residuals(fit1)
## X-squared = 8.7289, df = 10, p-value = 0.558

##
```

```
##  Box-Ljung test
##
## data:  residuals(fit1)
## X-squared = 9.2438, df = 10, p-value = 0.5091

##
##  Box-Ljung test
##
## data:  residuals(fit1)
## X-squared = 9.2438, df = 12, p-value = 0.682

##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit1)
## W = 0.96122, p-value = 0.0002332
```

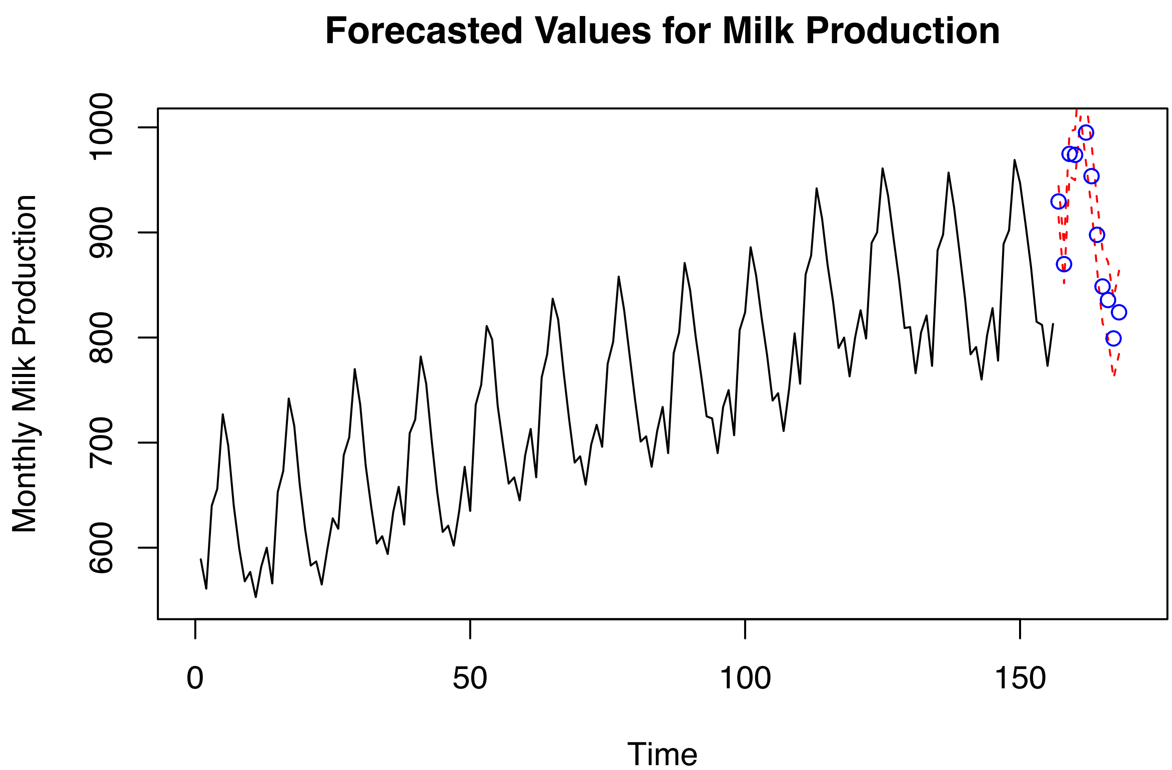
Forcasting

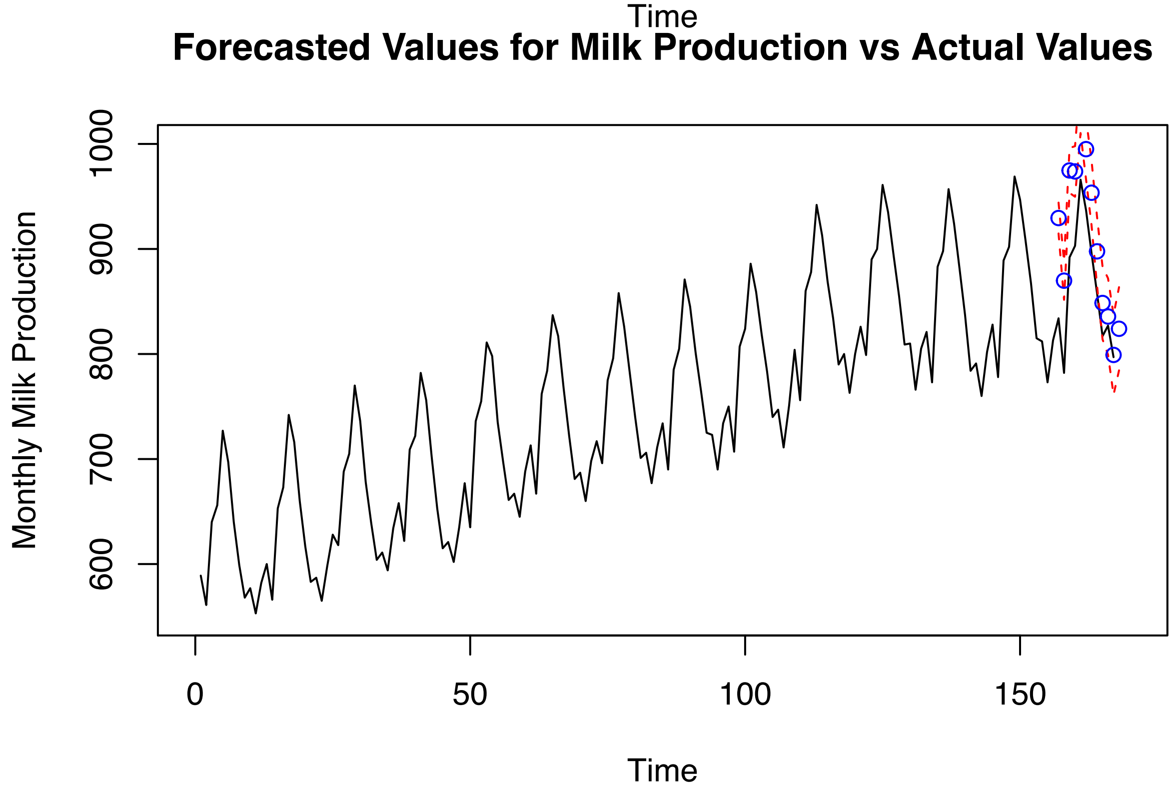
After checking for normality of the residuals and performing diagnostic tests on our model, we deemed that we could now perform forecasting for our original data using our $Model : SARIMA(1,1,1)x(1,1,1)_{12}$

$$(1 - .0937B)(1 - .0468B^{12}) \bigtriangledown_{12} \bigtriangledown X_t = (1 - .1723B)(1 - .5845B^{12})Z_t$$

$$where\ Z_t \sim WN(0, 49.814)$$

To perform forecasting we used the predict function in R and set the function to predict for the next 12 months of data: January 1975 to December 1975. We then calculated a 95 percent confidence interval which included predictions for lower and upper bounds of the next 12 months of data. We then plotted our predicted values. The red lines indicate the bounds of our confidence interval and the blue dots indicate the actual predicted values. To assess how well our model actually predicted the data we plotted our predictions versus the actual 12 months of data for January 1975 to December 1975. We can see from our plot that our model tended to overestimate a bit compared to actual data.


Forecasted Values for Milk Production
Monthly Milk Production
600
700
800
900
1000
0
50
100
150
Time


Forecasted Values for Milk Production vs Actual Values
Monthly Milk Production
600
700
800
900
1000
0
50
100
150
Time

Conclusion

The goal of our project was to fit a model that would forecast milk production. To accomplish this we used the box jenkins approach to model building. We started by plotting the milk production. The data has not been transformed or altered in any way. We noticed issues with our original data from plotting such as seasonality, trend, and high variance. To remedy the variance issue we first look into transforming the data. To figure out what transformation is best to use on our data we used the box cox method. Using box cox we found our lambda value. Based on our lambda value we decided a sqrt or log transformation would be best. We plotted our transformed data and found that there was not a significant difference between our transformed data vs our non transformed data. Therefore we decided that it would not be worthwhile to continue with our transformed data. We then proceeded to try and difference on our original data in hopes of removing the trend component. We differenced twice in our project, once to remove the trend and again to remove the seasonality. By differencing we also noticed that the variance had went down significantly, so we decided that at this point our data was stationary and could be used for model fitting. We came up with three possible SARIMA models for our data. To compare models and access which model was the best we used diagnosis such as AIC and AICc. After settling on one specific model we proceeded to perform further diagnosis checking on our model.

$Model : SARIMA(1, 1, 1)x(1, 1, 1)_{12}$

$$(1 - .0937B)(1 - .0468B^{12}) \bigtriangledown_{12} \bigtriangledown X_t = (1 - .1723B)(1 - .5845B^{12})Z_t$$

$$where\ Z_t \sim WN(0, 49.814)$$

This included plotting its roots to determine invertibility and causality as well as residuals to assess normality. We also used traditional diagnostic checking tests such as Box-Pierce, Box-Ljung, and Shapiro Wilk normality test. Our model passed all diagnostic tests except for Shapiro Wilk normality test. From this we deemed that our model was appropriate and could be used for forecasting. We then proceeded to forecast the next 12 months of data from January 1975 to December 1975. Comparing our forecasts to the actual data from those 12 months showed that our model had for the most part accurately predicted milk production. The weaknesses in our model were that it tended to overestimate milk product across the 12 months. In the end our goal of forecasting for January 1975 to December 1975 was achieved. However, there is definitely room to improve upon our model. We also should look into testing our model and forecasting for more recent observations. Perhaps data within the past 10 years: 2010-2019.

Future Study

As stated earlier, our models forecasts tended to overestimate the production of milk when compared to the actual data. Over estimation is a weakness of our model, but also our model does not account for significant world events such as possible disease outbreak among cows. To remedy this, we need to use implement model fitting methods that incorporate more variables to account for these possible variables. Also our model was fitted using data from 1962 to 1975; if we were to want to use our model to predict forecasts for say 2020,2021,. . . . we would need to probably expand the dataset we are using to include observations from 1976-2019.It would also be interesinting to see how our model fares for predicting milk production in other countries not just the U.S. This would require collecting more data from third party studies and adding it our current dataset. Overall, global population is continuing to rise rapidly and food products such as milk will need to be continued to be produced in mass. Therefore, the forecasting of milk production will remain relvant and important in determining the needs of society.

References

1. Milk Production 1962-1975, https://datamarket.com/data/list/?q=provider:tsdl DataMarket

Appendix

R Code:

```
library(MASS)
library(MuMIn)
library(sarima)
library(astsa)
#source("spec.arma.R")
source("plot.roots.R")
milk <-read.csv("milk.csv")
colnames(milk)[2] <- "production"
full.prod<- milk$production[-168]
prod<- milk$production[-c(157:170)] #Subset data so that we leave out last 12 months of data
#That way we can acess performance of our model later on
ts.plot(prod, main = "Plot Initial Data")
#Data is not stationary(clear upward trend) and seasonality component
#Variance looks to be constant
#Apply box cox even though variance looks to be constant




#decompose()
ts <- ts(prod, frequency = 12)
decompose <- decompose(ts, "multiplicative")

plot(as.ts(decompose$seasonal))
plot(as.ts(decompose$trend))
plot(as.ts(decompose$random))

plot(decompose)




#Box Cox
require(MASS)
bctrans <- boxcox(prod~as.numeric(1:length(prod)),
                  plotit=TRUE,
                  lambda=seq(-4,2,0.1))
lambda <- bctrans$x[which(bctrans$y==max(bctrans$y))]
print(paste("Lambda is equal to:",lambda))
#lamda = .42 which is close to .5 so should try sqrt transformation
#zero is also in the interval so lets try log transformation as well
prod.sqrt <- sqrt(prod)
prod.log <- log(prod)

# Plot original data vs Box-Cox transformed data
par(mfrow = c(2, 2))
ts.plot(prod,main = "Original data",ylab = expression(X[t]))
ts.plot(prod.sqrt,main = "Sqrt tranformed data", ylab = expression(Y[t]))
#Does not seem to be much of a difference between sqrt transformed and orginal
#Try plotting orginal vs log transformed
ts.plot(prod,main = "Original data",ylab = expression(X[t]))
ts.plot(prod.log,main = "Log tranformed data", ylab = expression(Y[t]))
```

```
#Not much of a difference in this case either
#Decide to not pursue transformation further
#Just work with original non transformed data



par(mfrow = c(3,2))

#Lets plot acf and pacf of milk production data
acf(prod, lag.max=60, main="ACF: No Differencing")
#From acf can tell that data is defintely non stationary
pacf(prod ,lag.max=60, main="PACF: No Differencing")
#pacf also suggests data is non stationary

#Differencing once at lag 1 to remove trend
milk.diff1 <- diff(prod, lag = 1)
print(paste("The variance with no differencing is: ", var(prod)))
print(paste("The variance after differencing is: ", var(milk.diff1)))#Variance goes down which is good
ts.plot(milk.diff1, main= "Plot of data after differencing once") #No trend is evident now, seasonality
acf(milk.diff1 ,lag.max=60, main="ACF after differencing once")
#Seasonal component is apparent with ACF, notice pattern of spikes at 12, 24,...
pacf(milk.diff1 ,lag.max=60, main="PACF after differencing once")
#Lots of spikes in PACF before 12



par(mfrow=c(2,2))
#Difference again at lag 12 to remove seasonality
milk.diff12 <- diff(milk.diff1, lag=12)
print(paste("The variance after differencing once is: ", var(milk.diff1)))#Variance goes down which is

plot.ts(milk.diff1, main = "Plot for data difference once")
plot.ts(milk.diff12, main = "Plot for data differnced twice")
print(paste("The variance after differencing twice is: ", var(milk.diff12)))#Variance goes down which i
#Patttern of spikes is gone which suggests seasonality is removed
#Also notice that variance is significantly lower
acf(milk.diff12 ,lag.max=60, main="ACF: Seasonality and Trend Removed")
#Acf suggests that data is stationary
#Possible cut off at 12, so Q= 1, cut off at 1 so q = 1
pacf(milk.diff12 ,lag.max=60, main="PACF: Seasonality and Trend Removed")
#Pacf suggests that data is stationary
#Cut off at 12, ,24 ,36  so P = 1, 2 or 3
#Cut off at 1 so p = 1
#d = 1, D = 1, P = (1,2,3), Q=1, p=1, q  = 1



#Sarima (1,1,1) x (1,1,1) 12
#Sarima (1,1,1) x (2,1,1) 12
#Sarima (1,1,1) x (3,1,1) 12

fit1 <- arima(x=prod, order=c(1,1,1),seasonal=list(order=c(1,1,1),period=12),method="ML")
fit1 #AIC = 991.93
```

```
fit2 <- arima(x=prod, order=c(1,1,1),seasonal=list(order=c(2,1,1),period=12),method="ML")
fit2  #AIC  = 992.83
fit3 <- arima(x=prod, order=c(1,1,1),seasonal=list(order=c(3,1,1),period=12),method="ML")
fit3 #AIC = 994.87

# Model 1 has lowest AIC value of 991.93

#Before continuing make sure to check causaulity and invertiblity of the models
par(mfrow=c(3,2))

plot.roots(NULL,polyroot(c(-.0937,-0.0468)),main="roots of AR part - fit1")
plot.roots(NULL,polyroot(c(-.1723,-.5845)),main="roots of MA part - fit1")
#Roots are outside unit circle. Process is causal and invertible.

plot.roots(NULL,polyroot(c(-.0718,-0.1612,-.1317)),main="roots of AR part - fit2")
plot.roots(NULL,polyroot(c(-.1883,-0.4741)),main="roots of MA part - fit2")
#Roots are outside unit circle. Process is causal and invertible.

plot.roots(NULL,polyroot(c(-.1258,-0.7035,-.4769,-.3451)),main="roots of AR part - fit3")
plot.roots(NULL,polyroot(c(-.0970,0.055)),main="roots of MA part - fit3")
#Roots are outside unit circle. Process is causal and invertible.

aic1<-AICc(arima(prod, order = c(1,1,1), seasonal=list(order=c(1,1,1), period=12), method="ML"))
print(paste("The AICc value for model 1 is : ", aic1))
#Model 1 AICc value is 992.1965
aic2<-AICc(arima(prod, order = c(1,1,1), seasonal=list(order=c(2,1,1), period=12), method="ML"))
print(paste("The AICc value for model 2 is : ", aic2))
#Model 2 AICc value is 993.2303
aic3<-AICc(arima(prod, order = c(1,1,1), seasonal=list(order=c(3,1,1), period=12), method="ML"))
print(paste("The AICc value for model 3 is : ", aic3))
#Model 3 AICc value is 995.4352


#Model 1 had the lowest AIC value, and also the lowest AICc value
#Will go with model 1



### Analyze residuals for our model
par(mfrow= c(3,2))
ts.plot(residuals(fit1), main = "Plot of residuals of Final Model")

acf(residuals(fit1), main= "ACF for residuals of Final Model")
pacf(residuals(fit1), main = "PACF for residuals of Final Model")

hist(residuals(fit1), main = "Histogram of residuals of Final Model") #Looks normally distributed.
#Slight skew on histogram towards the right.
#However for the most part looks normally distributed
qqnorm(residuals(fit1), main = "QQ-Plot of Residuals of Final Model")
qqline(residuals(fit1),col="blue")
#The qq plot looks to be a little off especially around the right tail of the plot.
#However it looks like these are just outliers considering it is only a couple points.
```

```r
###Diagnostic Checking
#lag = h = sqrt(n), n = 156, sqrt(156) =approx 12,  p =1, q =1 , fitdf = 1+1 = 2
Box.test(residuals(fit1), lag = 12, type = c("Box-Pierce"), fitdf = 2)
Box.test(residuals(fit1), lag = 12 , type = c("Ljung-Box"), fitdf = 2)
Box.test(residuals(fit1), lag = 12 , type = c("Ljung-Box"), fitdf = 0)
#P- values for all the box tests are greater than .05
#This indicates that autocorrelations are within the lags 1-11.
#Also indicates that residuals are independent
shapiro.test(residuals(fit1))
#Shapiro Wilk test p-value is less than .05 which indicates a rejection of assumption of normality
#However this is expected considering our histogram and qq-plot were a bit skewed



###Forecasting
par(mfrow=c(1,1))
fitfinal<- arima(x=prod, order=c(1,1,1),seasonal=list(order=c(1,1,1),period=12),
                 method =c("ML"), xreg=1:length(prod))
pred.prod<- predict(fitfinal, n.ahead = 12,
                        newxreg=(length(prod)+1) : length(prod)+12)
U.tr= pred.prod$pred + 2*pred.prod$se # upper bound for the C.I. for data
L.tr= pred.prod$pred - 2*pred.prod$se #Lower bound
#Forecasted data for months of January 1975: December 1975
ts.plot(prod,xlim=c(0, 170), ylim=c(550,1000),
        ylab = "Monthly Milk Production",
        main = "Forecasted Values for Milk Production")
lines(U.tr, col="red", lty="dashed")
lines(L.tr, col="red", lty="dashed")
points((length(prod)+1):(length(prod)+12), pred.prod$pred, col="blue")

#Compare forcasted data to actual data for months of January to Decemeber 1975
ts.plot(full.prod,xlim=c(0, 170), ylim=c(550,1000),
        ylab = "Monthly Milk Production",
        main = "Forecasted Values for Milk Production vs Actual Values")
lines(U.tr, col="red", lty="dashed")
lines(L.tr, col="red", lty="dashed")
points((length(prod)+1):(length(prod)+12), pred.prod$pred, col="blue")
#Looks like our forecasted data slightly overestimated production
#For the most part follows our actual data
```