Modeling and Forecasting on
California Natural Gas Net Generation Time Series Data

By: Claire Lee

Spring 2023

Summary

In this report, the chosen data to analyze is the Natural Gas Net Generation for all sectors in California from January 2001 to March 2023. One reason why this data was chosen is because I wanted to know if there is a trend when it comes to the amount of natural gas produced in California. Another question is "What are the future values of the Natural Gas Net Generation for all sectors in California?". To address these questions, we develop a time series model and forecast these future values in California. To choose one final model that we forecast on, we select possible candidate models, analyze them, and compare their results. In conclusion, there is a seasonal trend of 12 months, and according to the forecast, it is predicted that there will be a high natural gas net generation in the months of July, August, and September. Also, for certain months in 2024, it is predicted that the net generation will be lower than those same months in 2023.

Introduction

The dataset chosen is the Natural Gas Net Generation for all sectors monthly in California in thousand megawatt hours from January 2001 to March 2023. This dataset is provided by the U.S. Energy Information Administration (EIA) in the Electricity Data Browser on the EIA website. All sectors mean the sources of natural gas being generated include electric power, commercial, industrial, and residential. There are 267 observations in the dataset provided. It's important to analyze and be aware of the amount of natural gas being generated in California. Natural gas is a major factor of air pollution and air pollution is a major cause of environmental and health issues. Analyzing this data and predicting future amounts of natural gas generated can help state governments create new regulations and policies in order to decrease the risk of air pollution.

The first step to address this problem is to split the data into a training dataset and a test dataset. We split the data so that the training dataset has 255 observations from January 2001 to March 2022 and the test dataset has 12 observations from April 2022 to March 2023. We will use the training dataset throughout the process to select our final model time series. To first make sure that it has a normal distribution, we use the Box-Cox transformation to transform our data. To make sure that our time series is stationary, we use differencing to get rid of seasonality and trend. Then we plot the ACF and PACF in order to decide what our lowercase and uppercase p's and q's would be to choose our candidate models. Then we fit the models and analyze the coefficients and AICc values to decide what the final model would be. Once we choose our model, we check to make sure that it is stationary and invertible. The final technique

we use to make sure that our model is correct to use, is diagnostic checking where we analyze the residuals and perform tests to see if the residuals are White Noise. We use the model that satisfies the final technique to forecast in order to predict the future values.

Sections

1.1 Plot and analyze the time series

From the “Time Series on Natural Gas Trained Data” plot, we can see that there is an upward trend from 2001 to 2015 and then a downward trend from 2015 to near 2020, and then it increases again after 2020. Therefore, there is a trend. The fluctuations every 12 points indicate that there is a seasonal period of 12 months, which is one year. There are no sudden or sharp changes in behavior.

1.2 Transformation

We use the Box-Cox transformation technique in order to stabilize the variance and make our data have a Gaussian distribution. We plot the Box-Cox and the confidence interval to see our possible values for lambda. One lambda value found is 0.7070707. Another possible value is lambda=1 because it is also in the confidence interval. We use these two lambda values and compare the two transformed data. When we check the data with lambda=1, the histogram appears to have a normal distribution and the qq-norm plot has most of the points lying on a straight line. But once we perform the Shapiro-Wilk test, the p-value is 0.3539 which is less than 0.05, so we reject that hypothesis that the data is normally distributed. Then we check the transformed data with lambda 0.7070707. The histogram also appears to have a normal distribution and it is more symmetric than the previous transformed data. The qq-norm plot also has most of the points lying on a straight line. We also perform the Shapiro-Wilk test, and this time the p-value is 0.5551 which is greater than 0.05 so we can assume that this transformed data is normally distributed. Therefore, we will use the Box-Cox transformed data with the lambda value of 0.7070707.

We plot the time series and ACF plot after the Box-Cox transformation and see that the time series is still not stationary and there are seasonal peaks. So we will use the next technique differencing in order to remove the seasonality and trend. To remove the seasonality, we differenced at lag 12. To check that there is no overdifferencing, we compare the variances to make sure that the variance decreased. The variance before difference is 21305.88 and after differencing it becomes 10622.18, so it in fact decreased. The time series plot still shows a trend and the ACF plot shows that the ACF’s are decaying slowly. This means that it is still not stationary so we use differencing again. After differencing at lag 12, we difference at lag 1 to remove the trend. We compare the variances again to see if there is overdifferencing. The new variance becomes 7626.34 which again is lower than the previous variance mentioned earlier. From this plot we see no more trend and from the ACF plot we see that the acf’s decayed

quickly after lag 0. This indicates that the time series is stationary now, so we will use the Box-Cox transformed data with differencing at lag 12 and then lag 1.

1.3 Identify Model(s)

We continue to use the Box-Cox transformed data with differencing at lag 12 and then lag 1, and we use the ACF and PACF plots to preliminary identify the candidate models. First we know that our candidate models are SARIMA models because of the seasonality. Since we differenced at lag 12 to remove the seasonality, we use D=1 and s=12. Then we differenced at lag 1 to remove the trend so we use d=1. By using the ACF plot, we can see what our parameters q and Q would be for the MA part of the SARIMA models. From the ACF plot, we can see that between lags 0 and 12, the first lag is outside the confidence interval so q=1, and at lag 12, the ACF is outside the confidence interval so Q=1. By using the PACF plot, we can see what our parameters p and P would be for the AR part. From the PACF plot, we can see that between lags 0 and 12, lags 1 and 2 are outside the interval and lag 1 sticks out the most, so our possible parameters are p=1 and p=2. And at lag 12, the PACF is outside the confidence interval so P=1. Therefore we have two candidate models:
SARIMA (1,1,1) x (1,1,1)s=12 and SARIMA (2,1,1) x (1,1,1)s=12.

1.4 Fit Model(s)

We will identify the SARIMA (1,1,1) x (1,1,1)s=12 as Model A and SARIMA (2,1,1) x (1,1,1)s=12 as Model B. We fit both Model A and Model B, and analyze the estimated coefficients, standard errors, and AICc values. We want to choose the model with the lowest AICc value while also using the principle of parsimony, so the one with the least number of parameters possible. If the coefficient is less than 2 times the standard error then the estimated coefficient is inside the confidence interval and we can set this coefficient to zero. Once we fit Model A we obtain the AICc value of 2725.693, and only the coefficient for sar1 is within the confidence interval. So we fit the model again but with sar1 coefficient as zero which means Model A2 is SARIMA (1,1,1) x (0,1,1)s=12. After we fit Model A2, we see that the AICc value is 2725.721 which means the AICc increased but the P parameter decreased. We now fit Model B and see that the AICc value is 2727.772 which means the AICc increased even more. Because Model B q parameter increased and the AICc increased, we do not continue with Model B. Therefore, we choose Model A and Model A2 as the models that will continue to be used and checked. Model A in algebraic form is:
$(1-0.5556B)(1-0.1340B^{12})(1-B)(1-B^{12})X_t = (1-0.8639B)(1-0.8876B^{12})Z_t$
Model A2 in algebraic form is:
$(1-0.5499B)(1-B)(1-B^{12})X_t = (1-0.8538B)(1-0.8063B^{12})Z_t$

We then check Model A and Model A2 to see if both models are stationary and invertible. For Model A seasonal part, it is stationary because the sar1 coefficient is less than 1 and it is invertible because the absolute value of sma1 coefficient is less than 1. For the non-seasonal part of Model A, the root(s) of the polynomial for the AR part is 1.799856+0i which is outside the unit circle so it is stationary, and the root(s) of the polynomial for the MA part is 1.157541+0i which is outside the unit circle so it is invertible. For Model A2, since sar1 coefficient is set to zero, we only need to check invertibility for the seasonal part since all MA

models are stationary. The absolute value of the sma1 coefficient is 0.8063 which is less than 1 so it is invertible. For the non-seasonal part of Model A2, the root(s) of the polynomial for the AR part is 1.818512+0i which is outside the unit circle so it is stationary, and the root(s) of the polynomial for the MA part is 1.71234+0i which is outside the unit circle so it is invertible. This means both Model A and Model A2 are stationary and invertible.

Now we use the diagnostic checking technique on Model A and Model A2 and compare their results to choose the final model. For Model A, we can see from the residuals plot that there is no trend, no seasonality, a mean of about 0, and constant variance. Then we check for normality. The histogram of Model A residuals appears to have a normal distribution, and the QQ-Norm plot appears to have the residuals lie approximately on a straight line. We also run the Shapiro-Wilk test and find the p-value to be 0.7617 which is greater than 0.05 so we fail to reject the hypothesis that the residuals are normally distributed. From the ACF plot, all acf's of the residuals are within the confidence interval and can be counted as zero, and from the PACF plot, about all pacf's are within the confidence interval so they can be counted as zero. We can check this with the Portmanteau Tests: Box-Pierce and Ljung-Box and the McLeod-Li test. The lag is the square-root of the number of observations in the training dataset so the square root of 255 rounded is 16. For the Box-Pierce and Ljung-Box tests, fitdf is p + q so 1+1 so fitdf=2 and for the McLeod-Li test, the fitdf=0. For the Box-Pierce test, the p-value 0.5038 is greater than 0.05 so we fail to reject the hypothesis that the residuals are white noise. For the Box-Ljung test, the p-value 0.4516 is greater than 0.05 so we fail to reject the hypothesis that the residuals are white noise. For the McLeod-Li test, the p-value 0.2176 is greater than 0.05 so we fail to reject the hypothesis that the residuals are white noise. The residuals are uncorrelated and there is linear dependence. Therefore, Model A residuals resemble Gaussian White Noise.

For Model A2, we can see from the residuals plot that there is no trend, no seasonality, a mean of about 0, and constant variance. Then we check for normality. The histogram of Model A2 residuals also appears to have a normal distribution, and the QQ-Norm plot also appears to have the residuals lie approximately on a straight line. We also run the Shapiro-Wilk test and find the p-value to be 0.7501 which is greater than 0.05 so we fail to reject the hypothesis that the residuals are normally distributed. From the ACF plot, not all acf's of the residuals are within the confidence interval, and from the PACF plot, there are few pacf's that have strong peaks outside the confidence interval.

Since Model A passes more diagnostic checks than Model A2 does and also has a lower AICc value than Model A2 does, the final model chosen is Model A SARIMA (1,1,1) x (1,1,1)s=12 with the algebraic form (1-0.5556B)(1-0.1340B^12)(1-B)(1-B^12)X_t = (1-0.8639B)(1-0.8876B^12)Z_t

1.5 Forecasting

Using Model A SARIMA (1,1,1) x (1,1,1)s=12, we now predict the natural gas net generation from April 2022 to March 2023. These predicted values are called the forecasted values. We plot the forecasted values, the true values from the test dataset that we split in the beginning, and the confidence intervals on the original training dataset. From the plot, we can see that the forecasted values are very close to the true values. Both the forecasted and true values are within the confidence intervals. Therefore, the final model chosen, Model A, performs well and predicts pretty accurately forecasted natural gas net generation for all sectors monthly in California.

Conclusion

For this report, we are creating a model that can forecast future values based on a time series data. The data chosen is monthly natural gas net generation from all sources of generated natural gas in California in thousand megawatt hours from the U.S. Energy Information Administration (EIA). We study this dataset to see if there is a seasonal trend and to predict what the natural gas net generation is for the next 12 months. The model can be used to see if the predicted values are close to the true values from the whole original dataset provided by the EIA. To choose the right form of data, we use Box-Cox transformation in order to transform our data into a normally distributed one. Once we have our transformed time series, we use differencing to make our time series stationary. Then we use multiple different techniques to identify, select, and check our candidate models. In the end, the model we chose is SARIMA (1,1,1) x (1,1,1)s=12 with the math formula (1-0.5556B)(1-0.1340B^12)(1-B)(1-B^12)X_t = (1-0.8639B)(1-0.8876B^12)Z_t. We conclude that there is in fact a seasonal period of 12 months. This model was able to forecast future natural gas net generation that are very close to the actual values from the original EIA data. Therefore, it is safe to assume that this model can forecast pretty accurately for the next 12 months that have not been recorded. For the next 12 months, the forecasted values tell us that in March 2023, the natural gas net generation will be less than the recorded value in March 2022. But in the months of summer, July and August especially will have high peaks of natural gas net generation. This means that during this summer, there should be regulations to decrease natural gas net generation during the summer or at least awareness that these high amounts of natural gas net generation can pose risks for our health and the environment.

Acknowledgement to Professor Raisa Feldman and Thiha Aung for helping with this report.

References

EIA.gov Electricity Data Browser
https://www.eia.gov/electricity/data/browser/#/topic/0?agg=2,0,1&fuel=1&geo=000000000004&sec=g&freq=M&start=200101&end=202303&ctype=linechart&ltype=pin&rtype=s&maptype=0&rse=0&pin=

Claire Lee - 174 final project

2023-06-13

APPENDIX

```
# for AICc
library(qpcR)
```

```
## Loading required package: MASS
```

```
## Loading required package: minpack.lm
```

```
## Loading required package: rgl
```

```
## Loading required package: robustbase
```

```
## Loading required package: Matrix
```

Organizing the dataset and splitting data to get Training set

```
library(readr)
gas_ca_monthly <- read_csv("gas_ca_monthly.csv")
```

```
## Rows: 267 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Month
## dbl (1): California
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
gas_date <- gas_ca_monthly[nrow(gas_ca_monthly):1,] # reordering so that it's from earli
est to current date
gas_date
```

```
## # A tibble: 267 × 2
##    Month    California
##    <chr>         <dbl>
##  1 Jan 2001     10192.
##  2 Feb 2001      8871.
##  3 Mar 2001      9474.
##  4 Apr 2001      9209.
##  5 May 2001      9699.
##  6 Jun 2001      9350.
##  7 Jul 2001     10846.
##  8 Aug 2001     11506.
##  9 Sep 2001      9413.
## 10 Oct 2001      8943.
## # … with 257 more rows
```

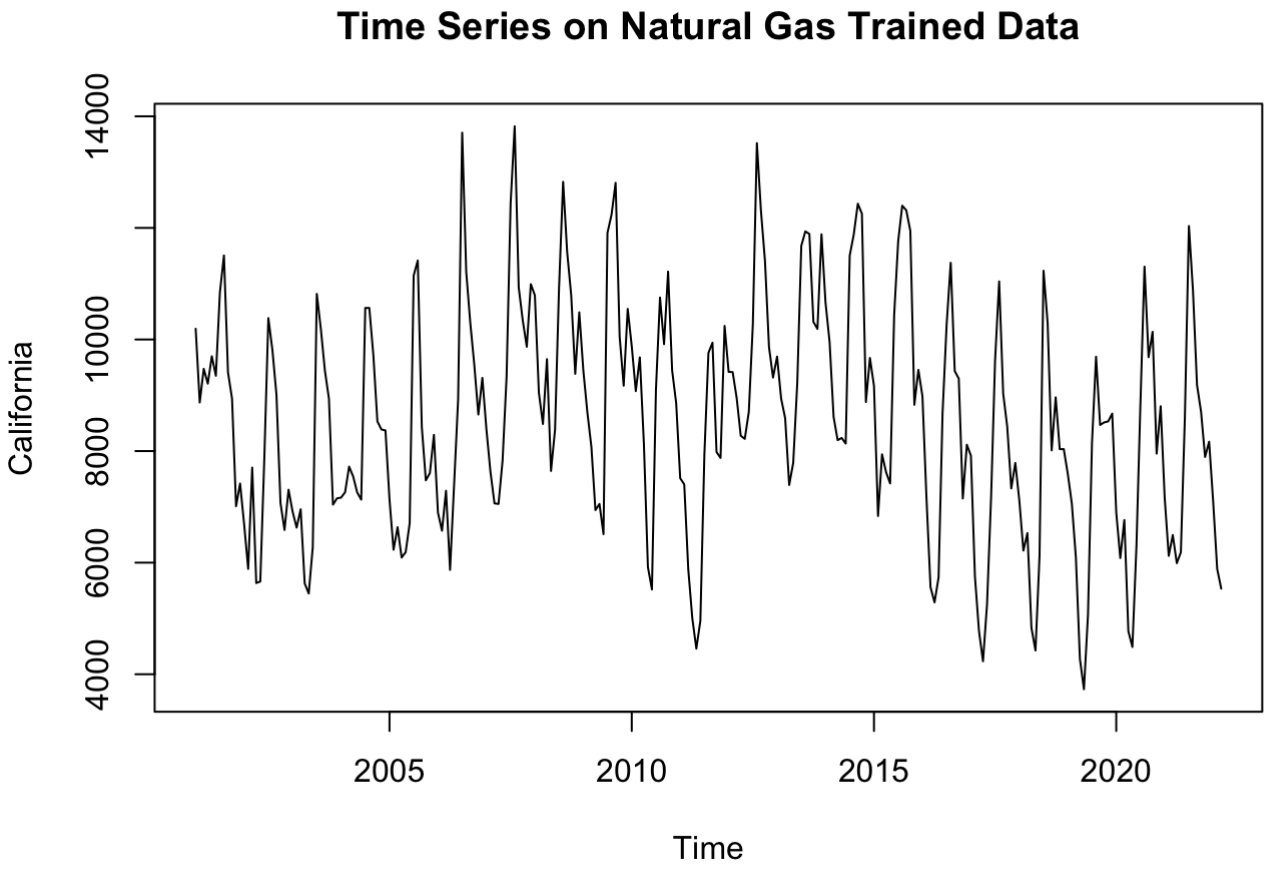
```
gasprod <- gas_date[,2]
gasprod
```

```
## # A tibble: 267 × 1
##    California
##         <dbl>
##  1     10192.
##  2      8871.
##  3      9474.
##  4      9209.
##  5      9699.
##  6      9350.
##  7     10846.
##  8     11506.
##  9      9413.
## 10      8943.
## # … with 257 more rows
```

```
gasprod_train <- gasprod[1:255,] # training data
gast <- ts(gasprod_train, start=c(2001,1), frequency=12)
```

This is the training dataset used for the rest of the project

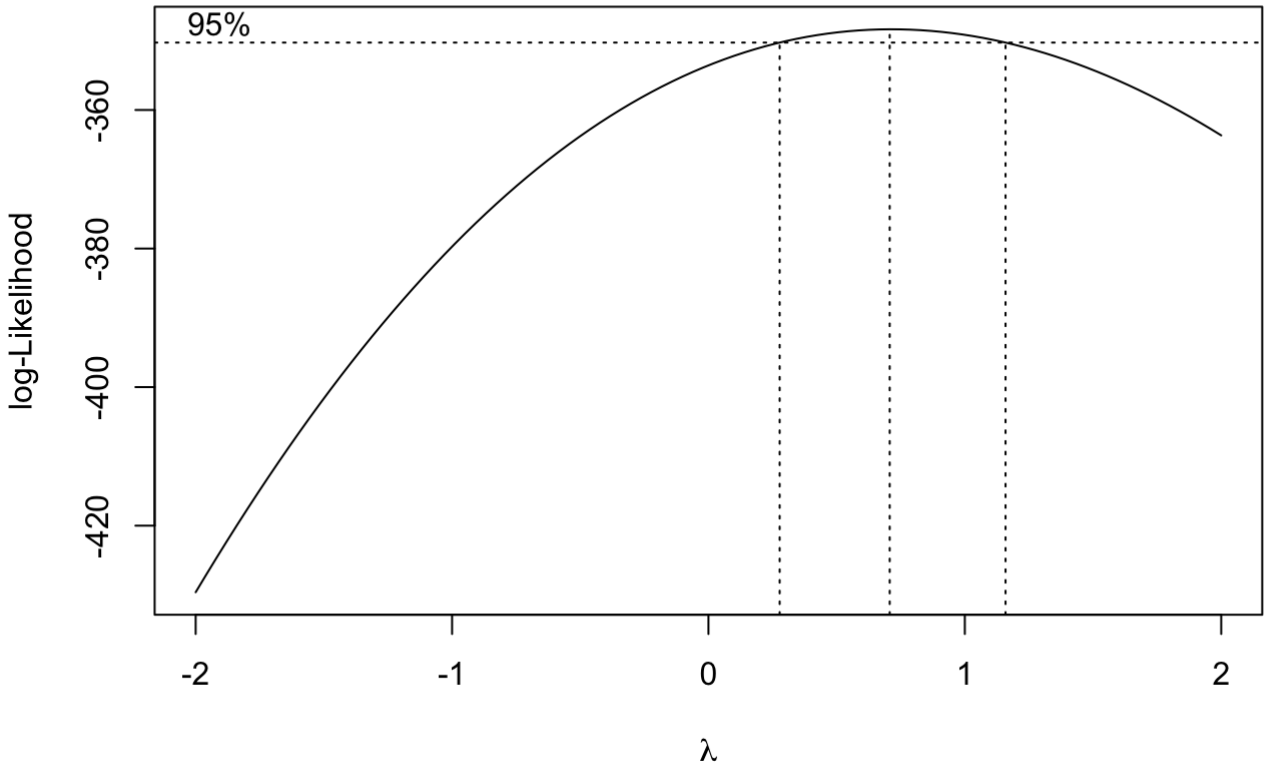
```
plot.ts(gast, main="Time Series on Natural Gas Trained Data")
```



This is a time series plot on the trained data. We must check if this time series data has a normal distribution. There is trend and seasonality as well so once after (if) transformation is needed, we will do differencing to make it stationary

Box-Cox Transformation

```
library(MASS)
bcTransform <- boxcox(gast ~ as.numeric(1:length(gast)))
```




```
lambda <- bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
lambda
```

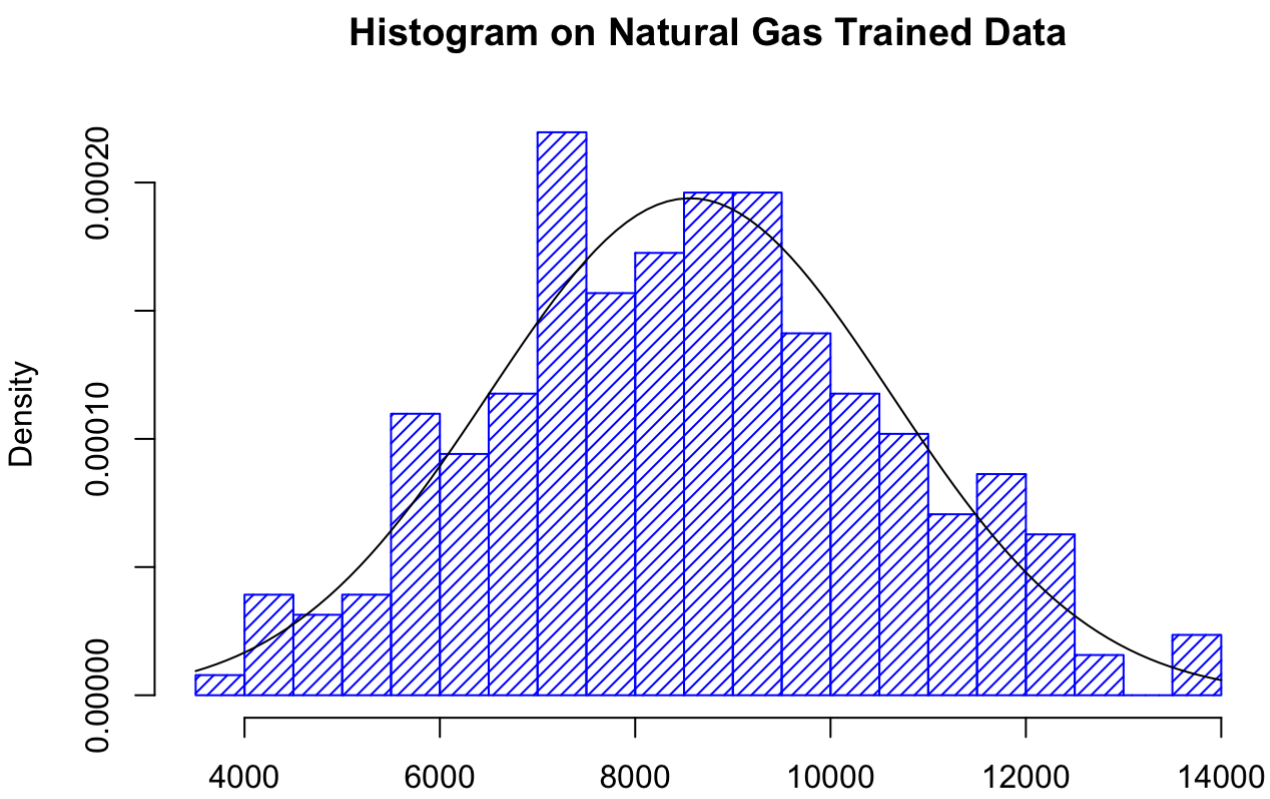
```
## [1] 0.7070707
```

```
gast_bc <- (1/lambda)*(gast^lambda-1)
```

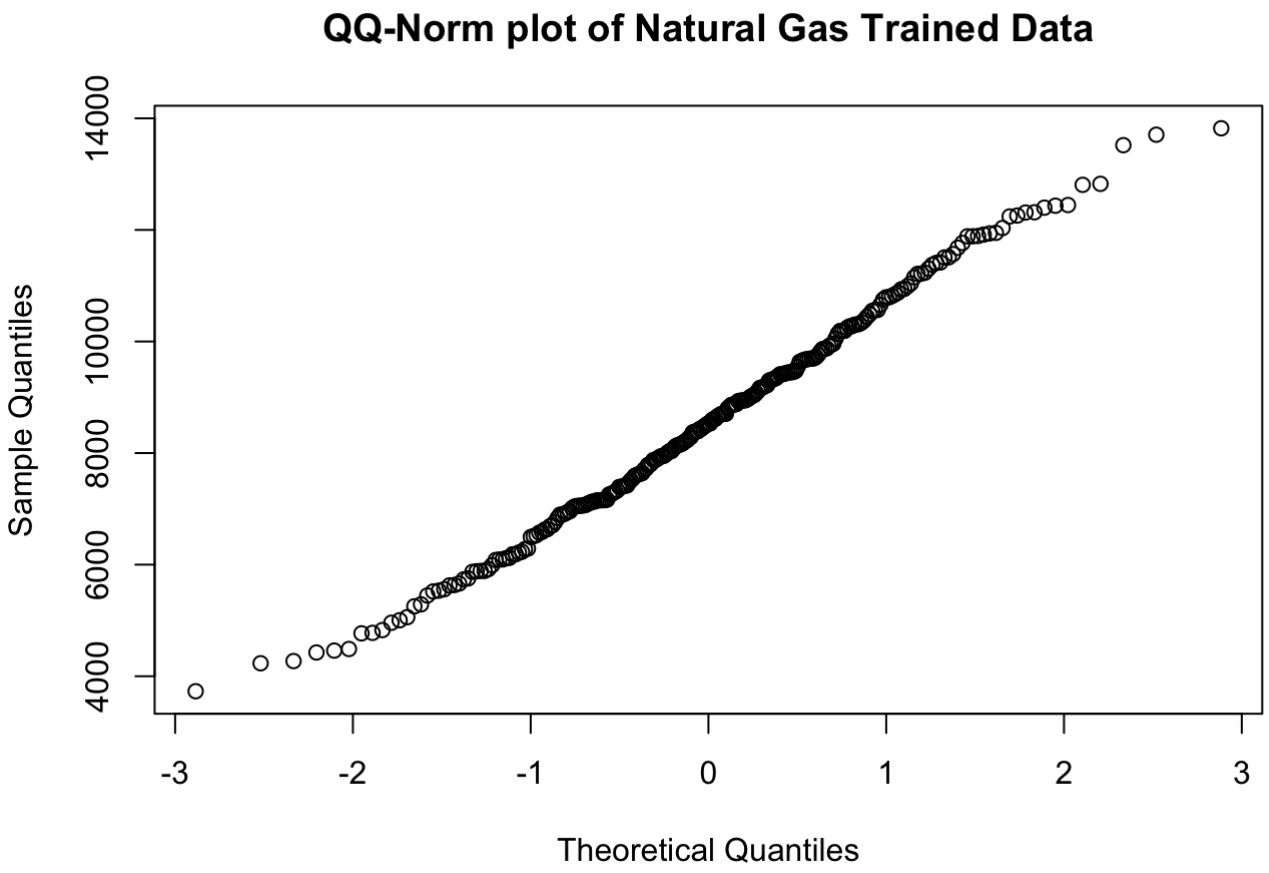
Lambda is 0.7070707 and the confidence interval includes lambda of 1. We will check if lambda=1 has the data with a normal distribution

Check if lambda = 1 is good (which is the original trained data)

```
hist(gast, density=20,breaks=20, col="blue", xlab="", prob=TRUE, main="Histogram on Natu
ral Gas Trained Data")
m<-mean(gast)
std<- sqrt(var(gast))
curve( dnorm(x,m,std), add=TRUE )
```




```
qqnorm(gast, main="QQ-Norm plot of Natural Gas Trained Data")
```



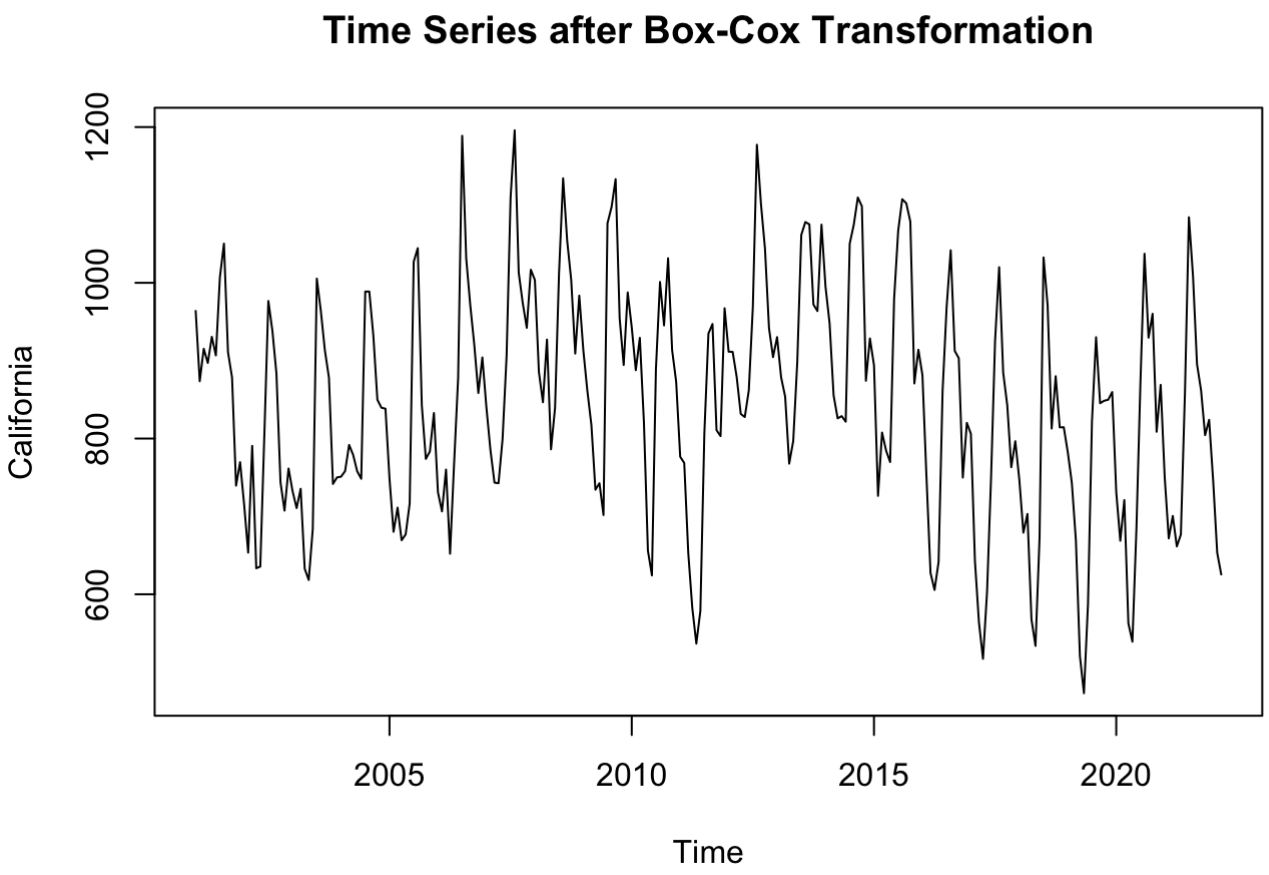

```
shapiro.test(gast)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gast
## W = 0.99364, p-value = 0.3539
```

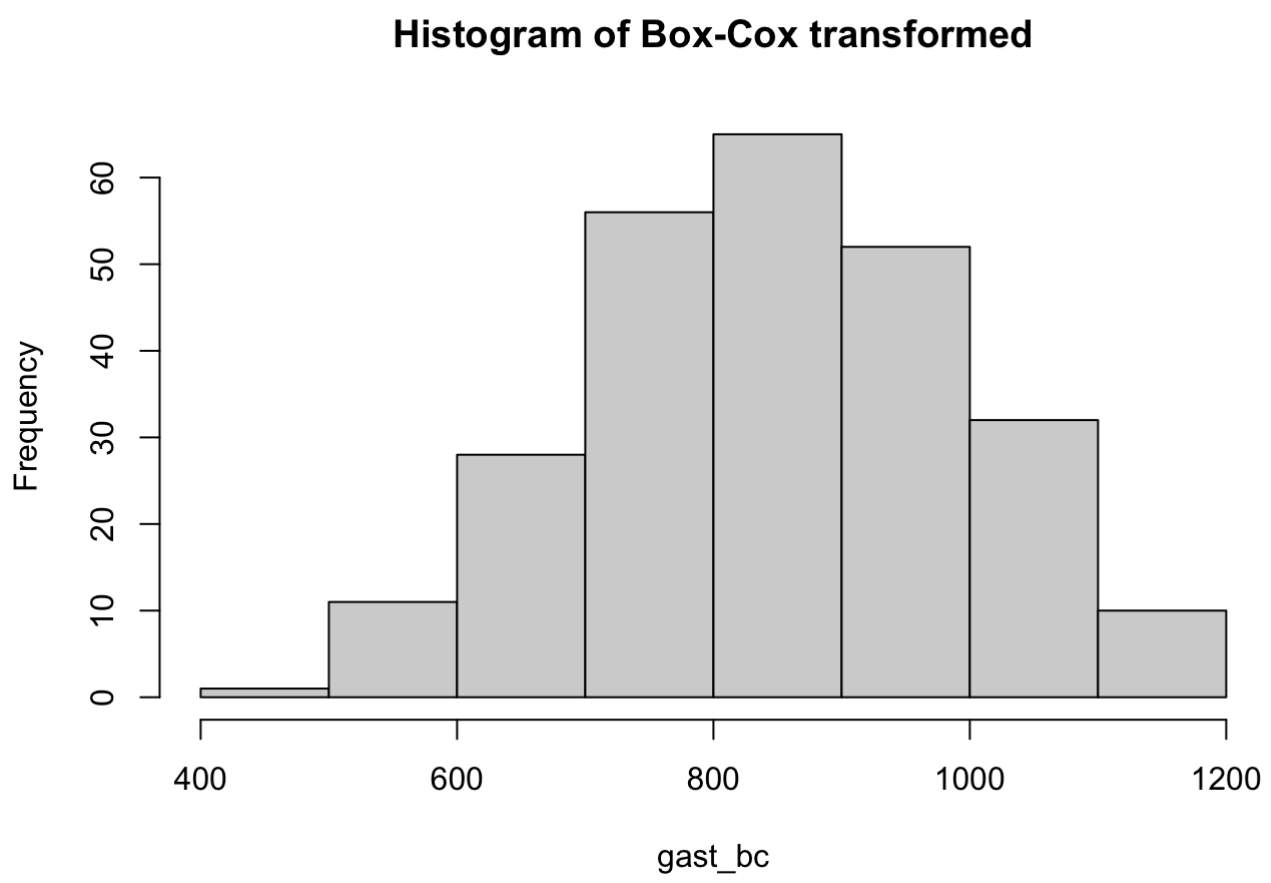
The histogram appears to have a normal distribution but has a few bars outside of the normal bell curve The qq-norm plot has most of the points lying on a straight line. The Shapiro-Wilk test: the p-value is 0.3539 which is less than 0.05, so we reject that hypothesis that the data is normally distributed. p-value is less than 0.5 –> assume not normal We can say that the lambda=1 original trained data is not good to use so we will check if the box-cox transformed data with lambda=0.7070707 is better

Check if box-cox transformed train is better

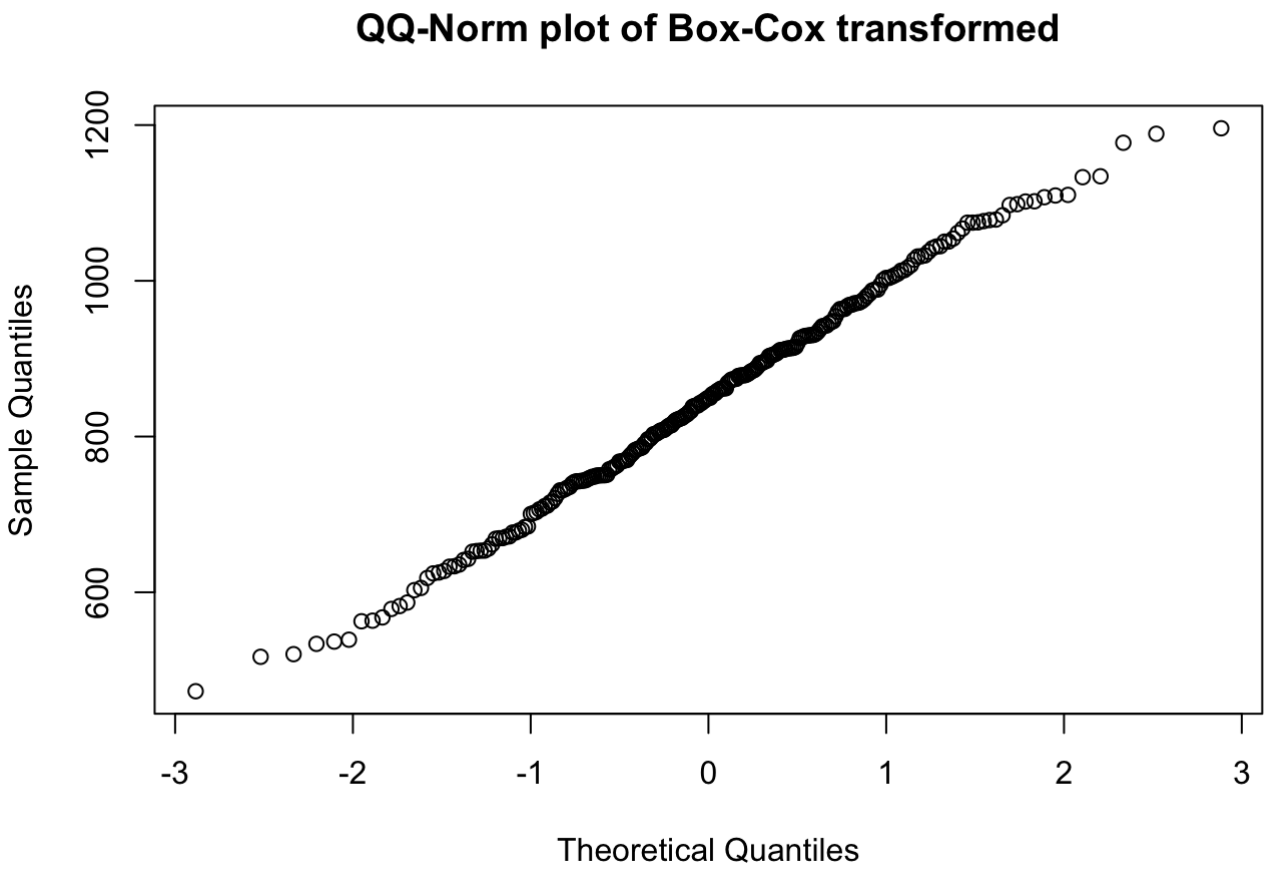
```
plot.ts(gast_bc, main="Time Series after Box-Cox Transformation")
```




```
hist(gast_bc, main="Histogram of Box-Cox transformed")
```




```
qqnorm(gast_bc, main="QQ-Norm plot of Box-Cox transformed")
```



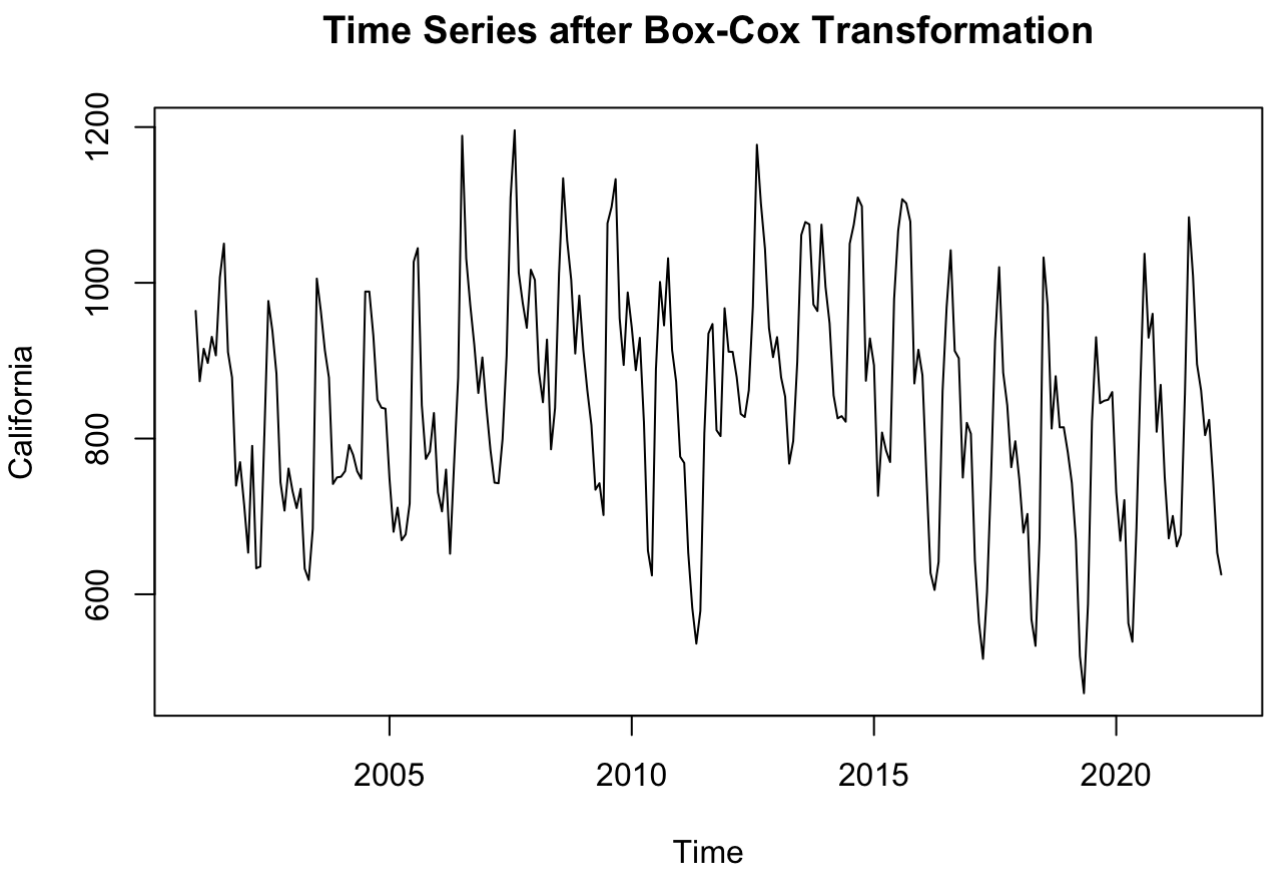

```
shapiro.test(gast_bc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gast_bc
## W = 0.99489, p-value = 0.5551
```

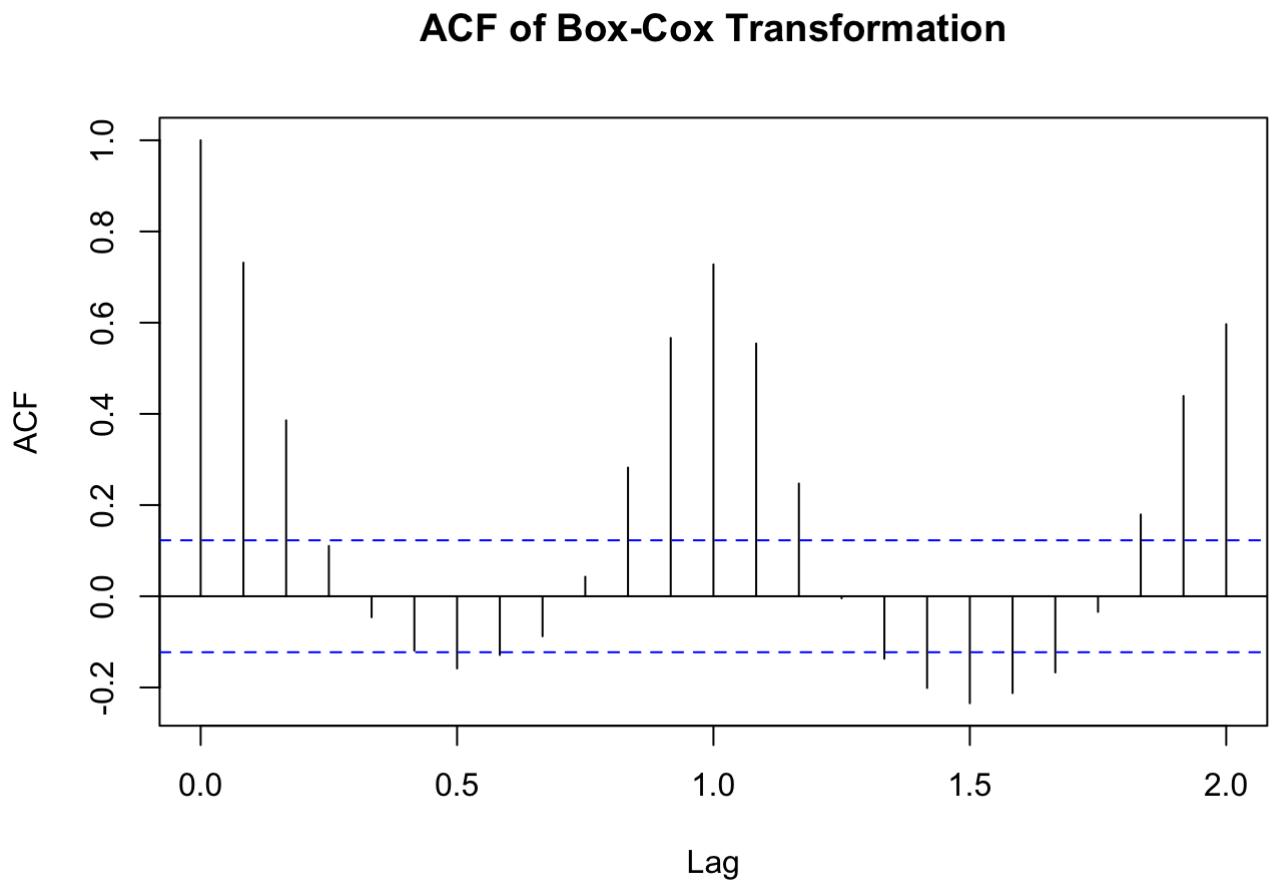
The histogram also appears to have a normal distribution and it is more symmetric than the previous transformed data. The qq-norm plot also has most of the points lying on a straight line. The Shapiro-Wilk test: p-value is 0.5551 which is greater than 0.05 so we can assume that this transformed data is normally distributed. Therefore, we will use the Box-Cox transformed data with the lambda value of 0.7070707.

Use Box-Cox transformed trained data for Differencing step

```
plot.ts(gast_bc, main="Time Series after Box-Cox Transformation")
```

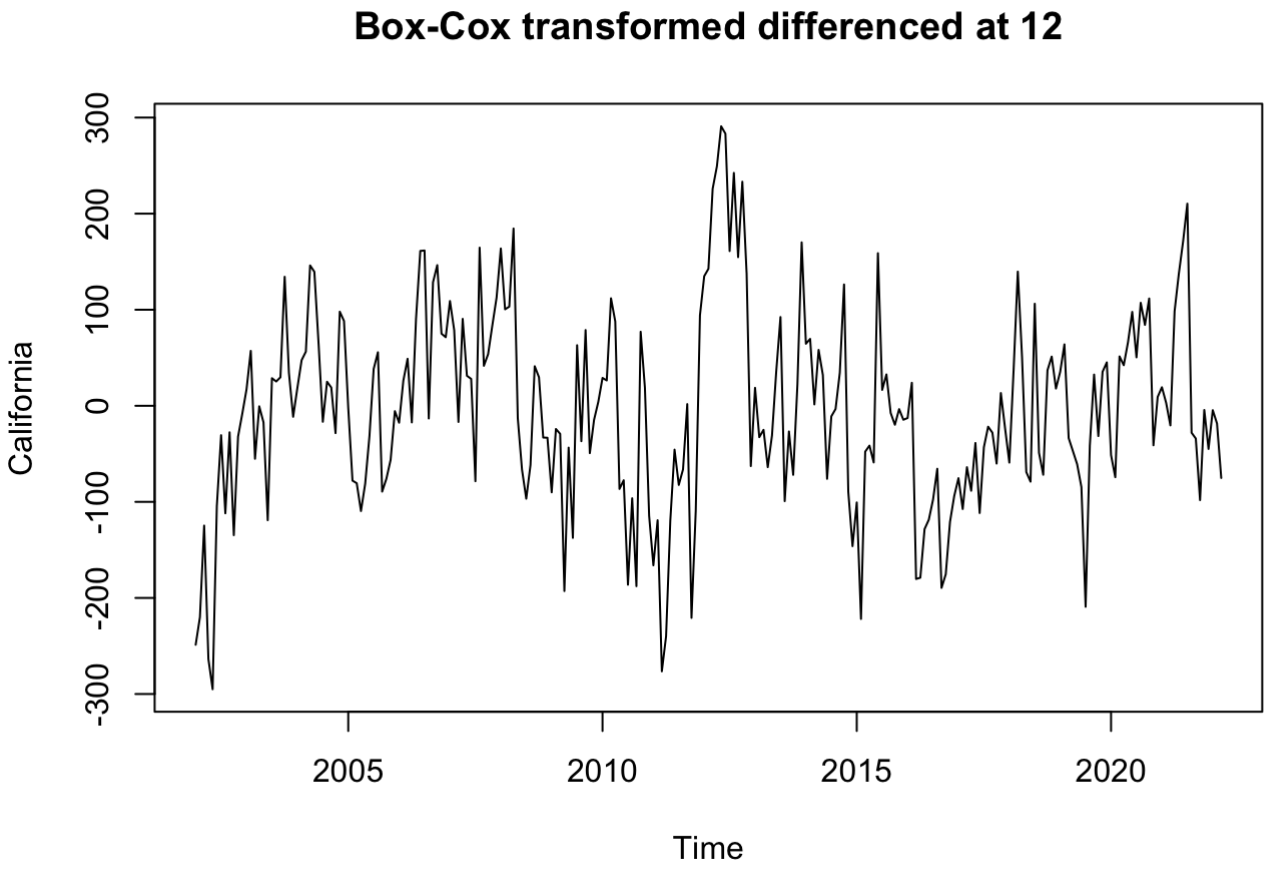



```
acf(gast_bc, main="ACF of Box-Cox Transformation")
```

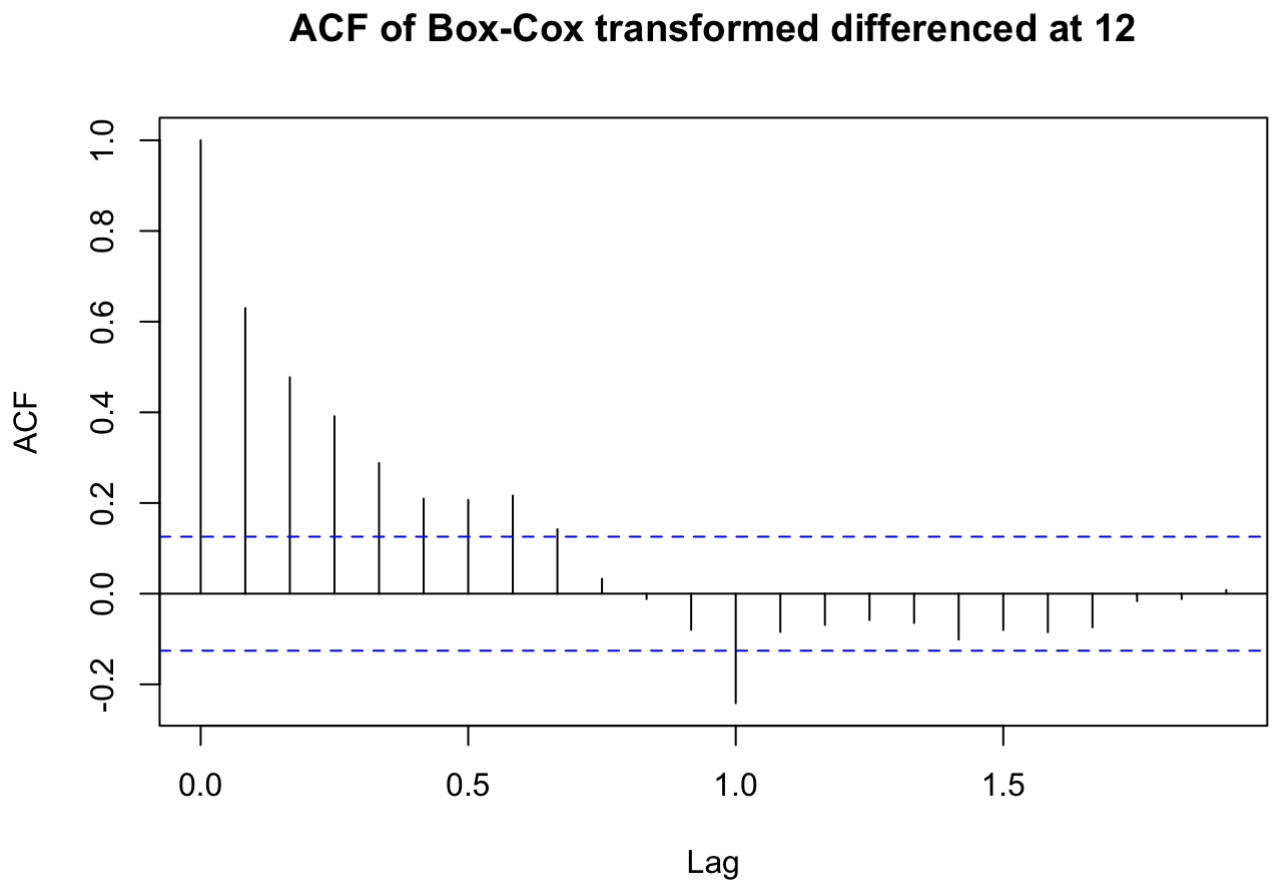


ACF plot shows seasonality where there is a fluctuation every 12 lags. So difference at lag 12

```
dgast <- diff(gast_bc, lag=12)
plot.ts(dgast, main="Box-Cox transformed differenced at 12" )
```



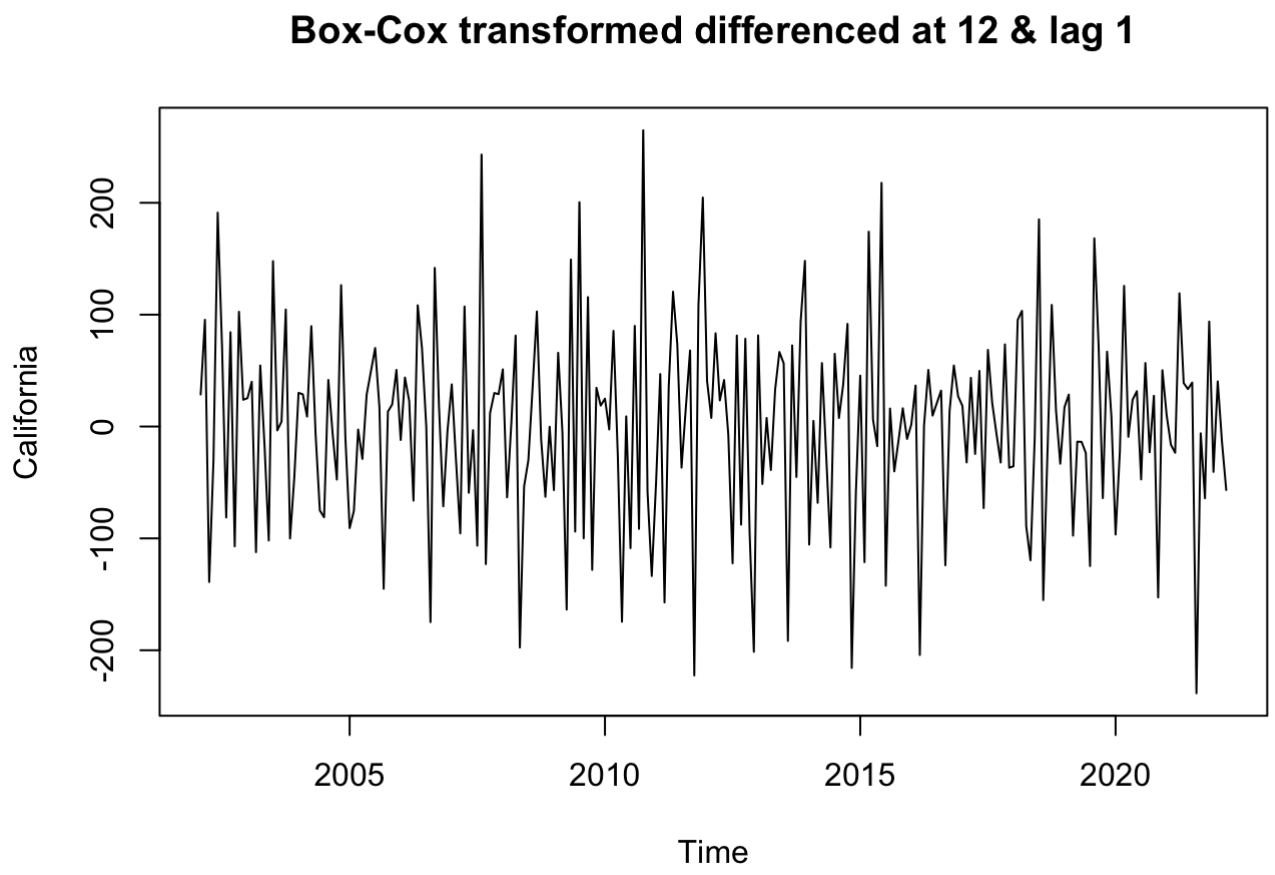

```
acf(dgast, main= "ACF of Box-Cox transformed differenced at 12")
```



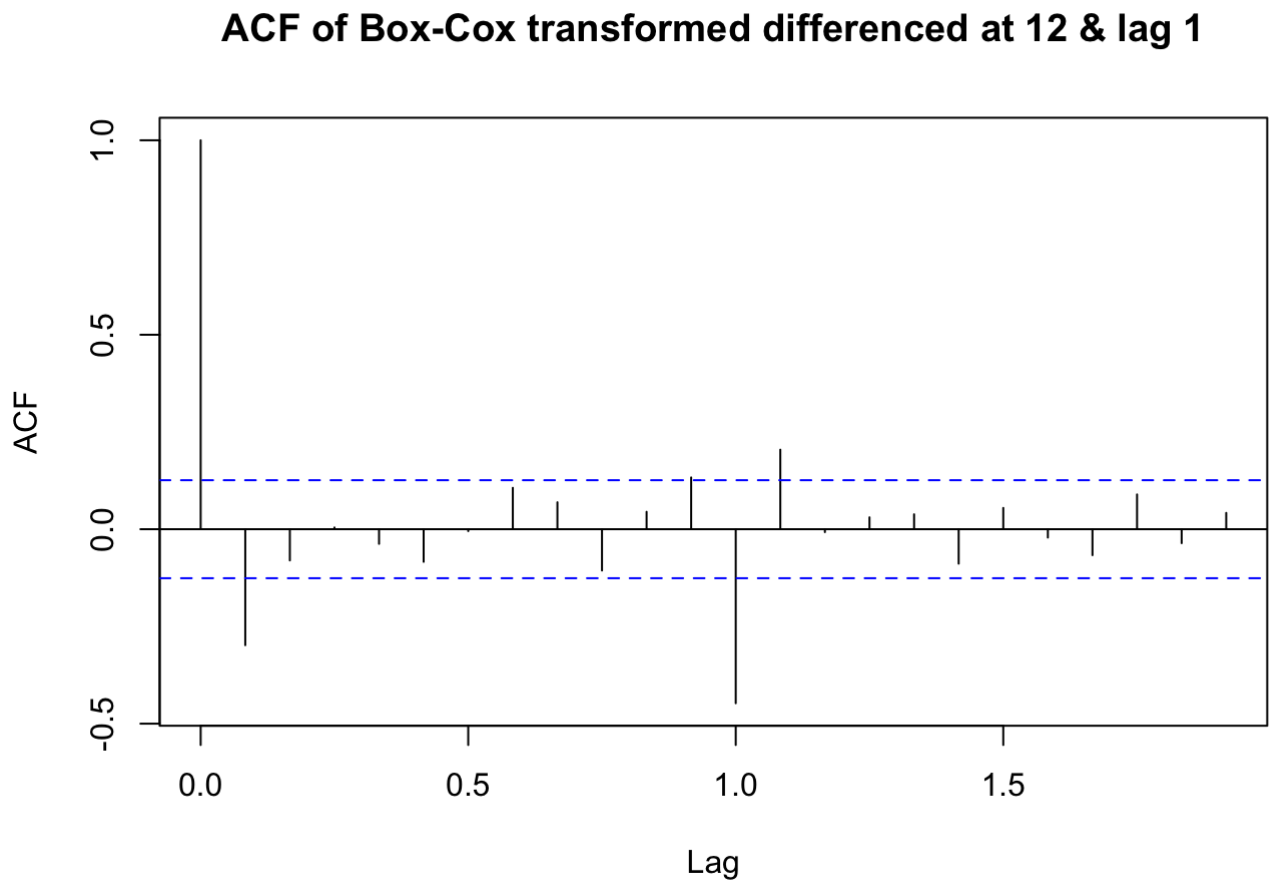
ACF still gradually decreasing so difference again, next at lag 1

Difference at lag 12 then lag 1

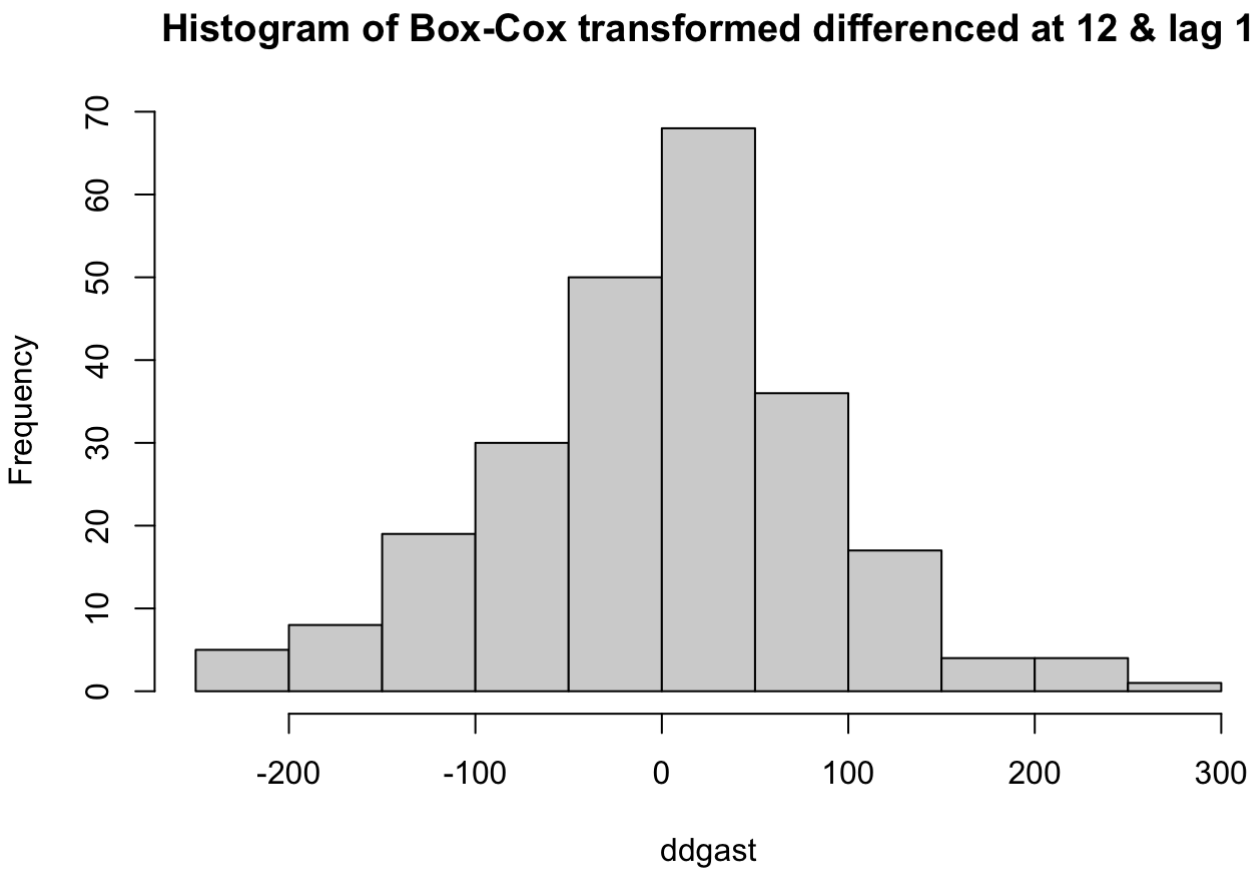
```
ddgast <- diff(dgast, lag=1)
plot.ts(ddgast, main ="Box-Cox transformed differenced at 12 & lag 1")
```




```
acf(ddgast, main="ACF of Box-Cox transformed differenced at 12 & lag 1")
```




```
hist(ddgast, main="Histogram of Box-Cox transformed differenced at 12 & lag 1")
```



Time Series plot of Box-Cox transformed differenced at 12 & lag 1 has no trend and no seasonality so it is stationary. ACF plot shows acf quickly decaying after lag 0 so it also shows that it is stationary

Check variances to see if there's overdifferencing

```
var(gast_bc)
```

```
##            California
## California   21305.88
```

```
var(dgast)
```

```
##            California
## California   10622.18
```

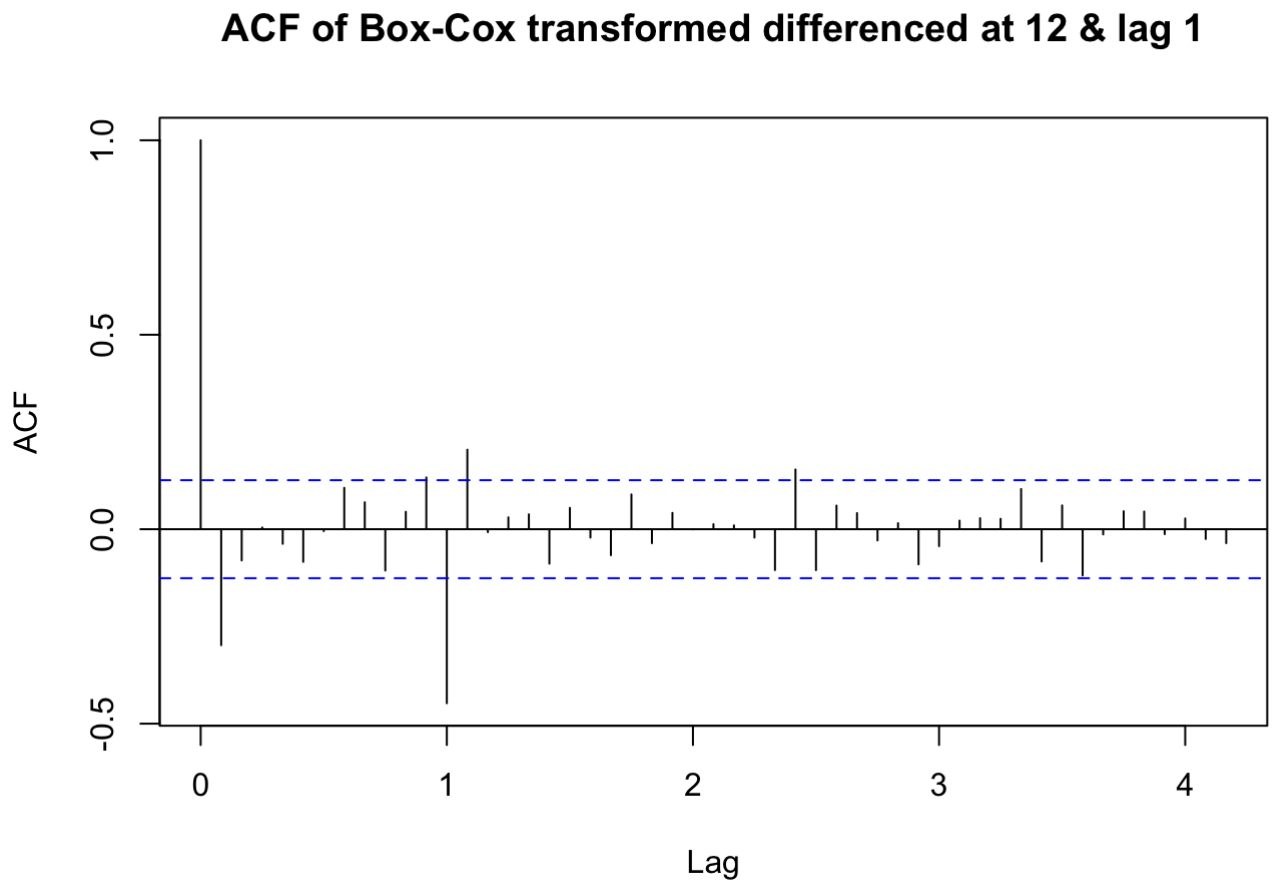
```
var(ddgast)
```

```
##            California
## California    7626.34
```

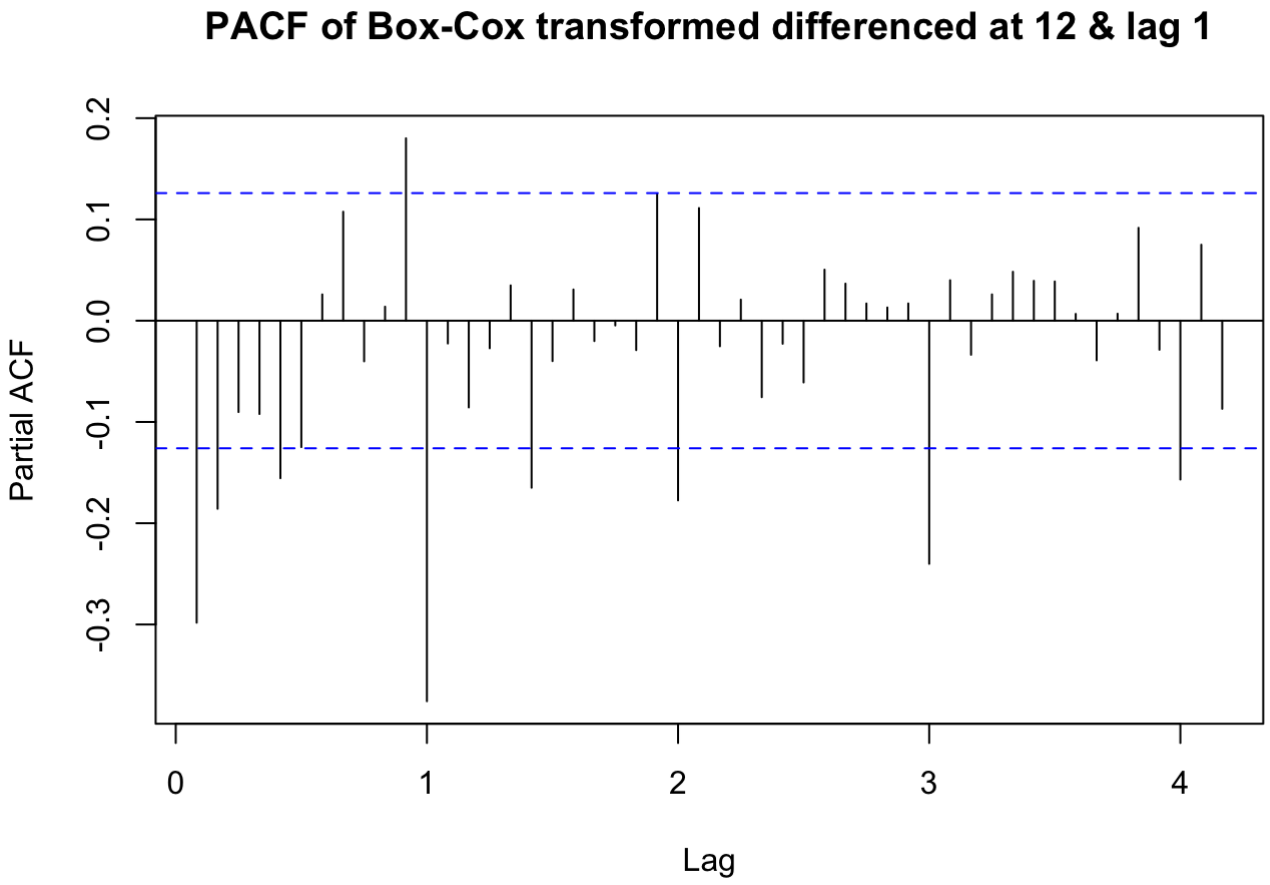
From box-cox transformation with no differencing to with differencing at lag 12 to with differencing at lag 12 and then lag 1, the variance decreases There is no overdifferencing

ACF and PACF of Box-Cox transformed differenced at 12 & lag 1

```
acf(ddgast, lag.max=50, main="ACF of Box-Cox transformed differenced at 12 & lag 1")
```




```
pacf(ddgast, lag.max=50, main="PACF of Box-Cox transformed differenced at 12 & lag 1")
```



Candidate models SARIMA because of seasonality s=12, D=1, d=1 From differencing at lag 12 and then lag 1

p and q ACF: between lags, acfs outside interval –> q = 1 PACF: between lags, pacfs outside interval –> p = 1 or 2

P and Q ACF: acfs outside around lag 12 –> Q = 1 PACF: pacfs outside around lag 12 –> P = 1

Candidate models: Model A SARIMA (1,1,1) x (1,1,1) s=12 Model B SARIMA (2,1,1) x (1,1,1) s=12

Fit these candidate models to estimate coefficients and find AICc values in order to help choose final model

SARIMA (1,1,1) x (1,1,1) s=12

```
arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=12), method='ML')
```

```
##
## Call:
## arima(x = gast_bc, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1), period = 1
2),
##     method = "ML")
##
## Coefficients:
##          ar1      ma1    sar1     sma1
##       0.5556  -0.8639  0.1340  -0.8876
## s.e.  0.1028   0.0691  0.0906   0.0781
##
## sigma^2 estimated as 4087:  log likelihood = -1357.77,  aic = 2725.53
```

```
AICc(arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=12), method='M
L'))
```

```
## [1] 2725.693
```

coefficient < 2(s.e) –> inside CI yes zero

For ar1: 0.5556 < 2(0.1028) false –> outside CI, no zero

For ma1: |-0.8639| < 2(0.0691) false –> outside CI

For sar1: 0.1340 < 2(0.0906) –> inside CI, yes zero

For sma1: |-0.8876| < 2(0.0781) false –> outside CI

only sar1 inside CI so set sar1 coefficient to zero as candidate model A2 AICc = 2725.693, used to compare other candidate models

SARIMA (1,1,1) x (0,1,1), set coef for sar1 to zero

```
arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(0,1,1), period=12),  method='ML')
```

```
##
## Call:
## arima(x = gast_bc, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 1
2),
##     method = "ML")
##
## Coefficients:
##          ar1      ma1     sma1
##       0.5499  -0.8538  -0.8063
## s.e.  0.0990   0.0657   0.0625
##
## sigma^2 estimated as 4176:  log likelihood = -1358.81,  aic = 2725.63
```

```
AICc(arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(0,1,1), period=12),  method='ML'))
```

```
## [1] 2725.721
```

AICc increased from 2725.693 to 2725.721 Even though AICc increased, the # of parameters decreased so this is a candidate model, named Model A2

Checking Model B to see if increasing p by 1 is a good model

SARIMA (2,1,1) x (1,1,1) s=12

```
arima(gast_bc, order=c(2,1,1), seasonal=list(order=c(1,1,1), period=12),  method='ML')
```

```
##
## Call:
## arima(x = gast_bc, order = c(2, 1, 1), seasonal = list(order = c(1, 1, 1), period = 12),
##     method = "ML")
##
## Coefficients:
##          ar1     ar2      ma1    sar1     sma1
##       0.5572  0.0047  -0.8672  0.1356  -0.8887
## s.e.  0.1038  0.0823   0.0853  0.0952   0.0801
##
## sigma^2 estimated as 4085:  log likelihood = -1357.77,  aic = 2727.53
```

```
AICc(arima(gast_bc, order=c(2,1,1), seasonal=list(order=c(1,1,1), period=12),  method='ML'))
```

```
## [1] 2727.772
```

Model A AICc is 2725.693 and Model B AICc is 2727.772 AICc increased and the # of parameters is higher so Model B is not a good model

So the final candidate models for diagnostic checking are Model A SARIMA (1,1,1) x (1,1,1) s=12 with lowest AICc = 2725.693 and Model A2 SARIMA (1,1,1) x (0,1,1)s=12 (Model A but with sar1 coefficient set to 0) with second lowest AICc = 2725.721

Analyzing the coefficients of Model A to write algebraic form

```
arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=12), method='ML')
```

```
##
## Call:
## arima(x = gast_bc, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1), period = 1
2),
##     method = "ML")
##
## Coefficients:
##          ar1      ma1    sar1     sma1
##       0.5556  -0.8639  0.1340  -0.8876
## s.e.  0.1028   0.0691  0.0906   0.0781
##
## sigma^2 estimated as 4087:  log likelihood = -1357.77,  aic = 2725.53
```

```
AICc(arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=12), method='M
L'))
```

```
## [1] 2725.693
```

(1-0.5556B)(1-0.1340B$^{12)(1-B)(1-B}$12)X_t = (1-0.8639B)(1-0.8876B^12)Z_t Differenced at lag 12 and then at lag 1

Checking stationarity and invertibility

For seasonal part Check SAR1 for stationarity 0.1340 < 1 –> stationary

Check SMA1 for invertibility | -0.8876 | < 1 –> invertible

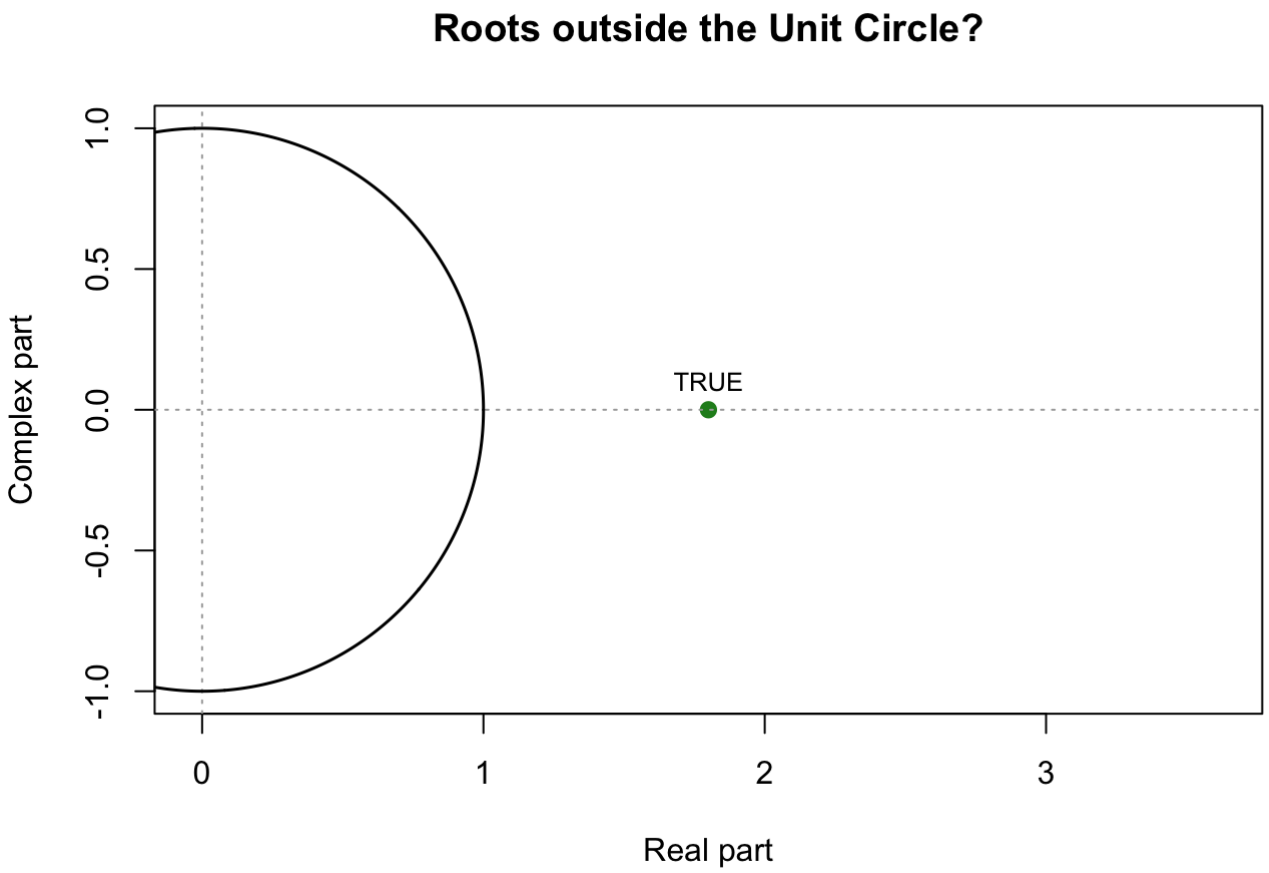
For non-seasonal part

```
# Check AR parts for stationary
# (1-0.5556B)
polyroot(c(1,-0.5556))
```

```
## [1] 1.799856+0i
```

```
library(UnitCircle)
uc.check(pol_ = c(1,-0.5556), plot_output = TRUE)
```

```
##       real complex outside
## 1 1.799856       0    TRUE
## *Results are rounded to 6 digits.
```



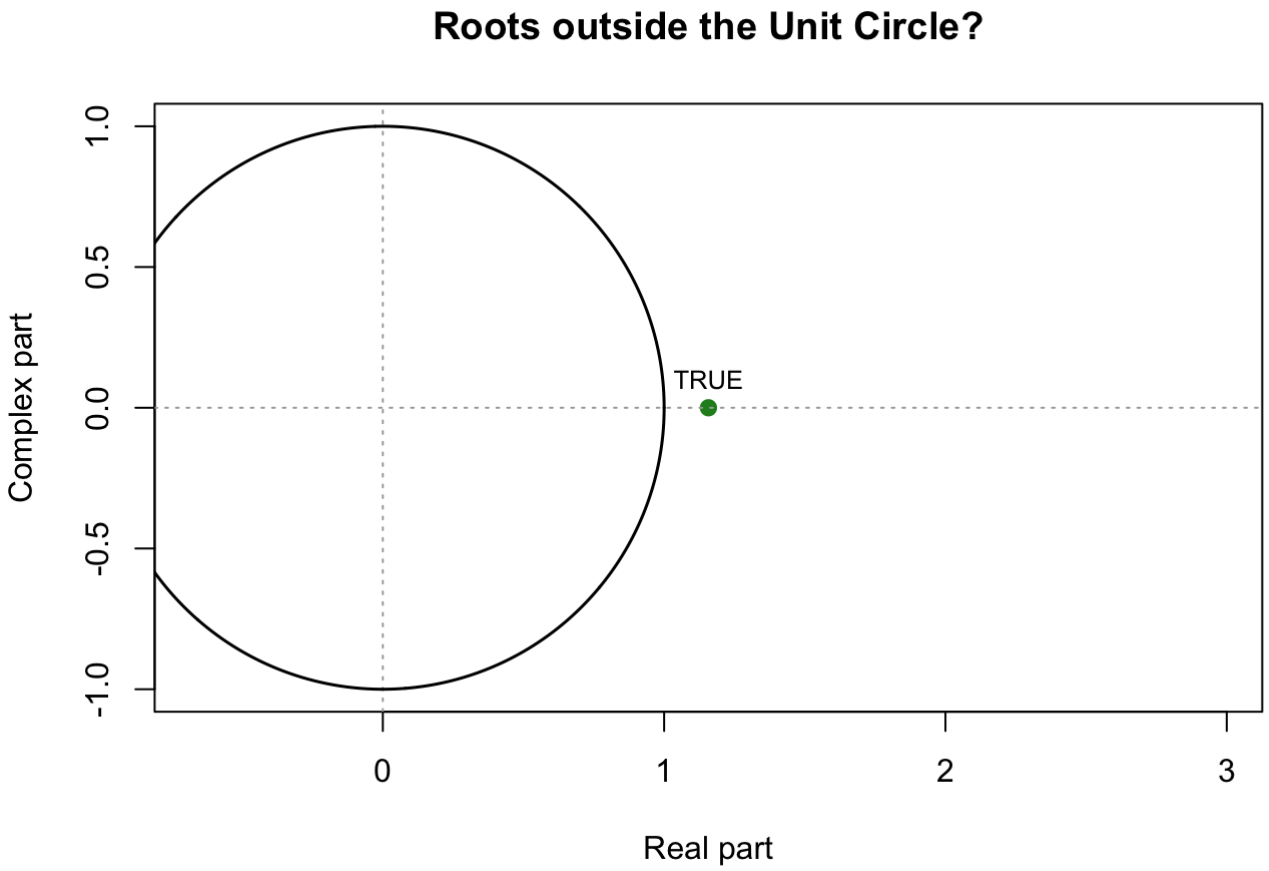
root of polynomial for AR part is outside unit circle –> stationary

```
# Check MA parts for invertibility
# (1-0.8639B)
polyroot(c(1,-0.8639))
```

```
## [1] 1.157541+0i
```

```
library(UnitCircle)
uc.check(pol_ = c(1,-0.8639), plot_output = TRUE)
```

```
##       real complex outside
## 1 1.157541       0    TRUE
## *Results are rounded to 6 digits.
```



root of polynomial for MA part is outside unit circle –> invertible

Model A is stationary and invertible Now use diagnostic checking to see if Model A is Gaussian White Noise and can be used as final model for forecasting

Diagnostic checking

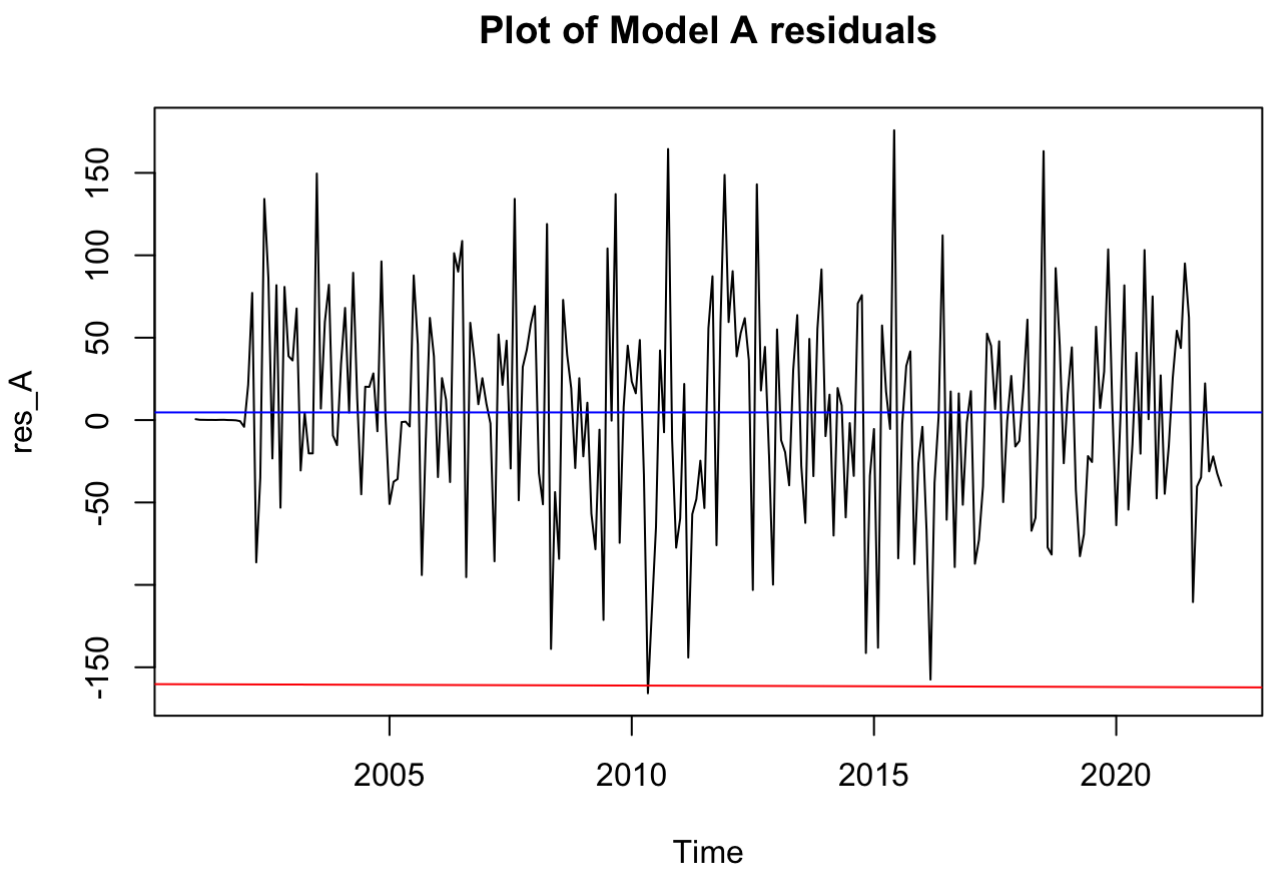
compare residuals reject hypothesis if p-value is < 0.05 i) plot residuals (should be WN) ii) plot histogram and qqplot (should be Gaussian) ii) run Shapiro Wilk test for normality iii) Box-Pierce, Ljung-Box, McLeod-Li tests

Fitting Model A named as fit_A and finding the residuals

```
fit_A <- arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=12),  metho
d='ML')

res_A <- residuals(fit_A)
m_res_A <- mean(res_A)
std_res_A <- sqrt(var(res_A))
```

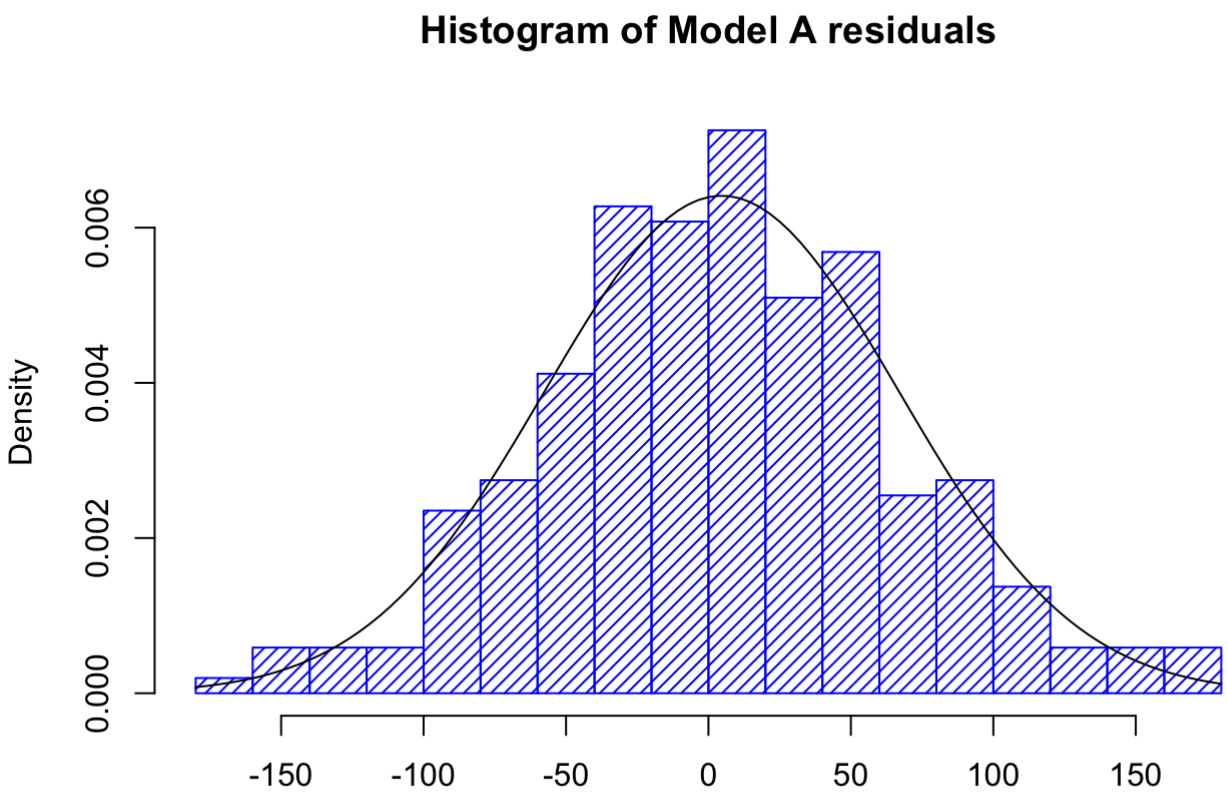
```
plot.ts(res_A, main='Plot of Model A residuals')
fit_line_A <- lm(res_A ~ as.numeric(1:length(res_A)))
abline(fit_line_A, col="red")
abline(h=m_res_A, col="blue")
```



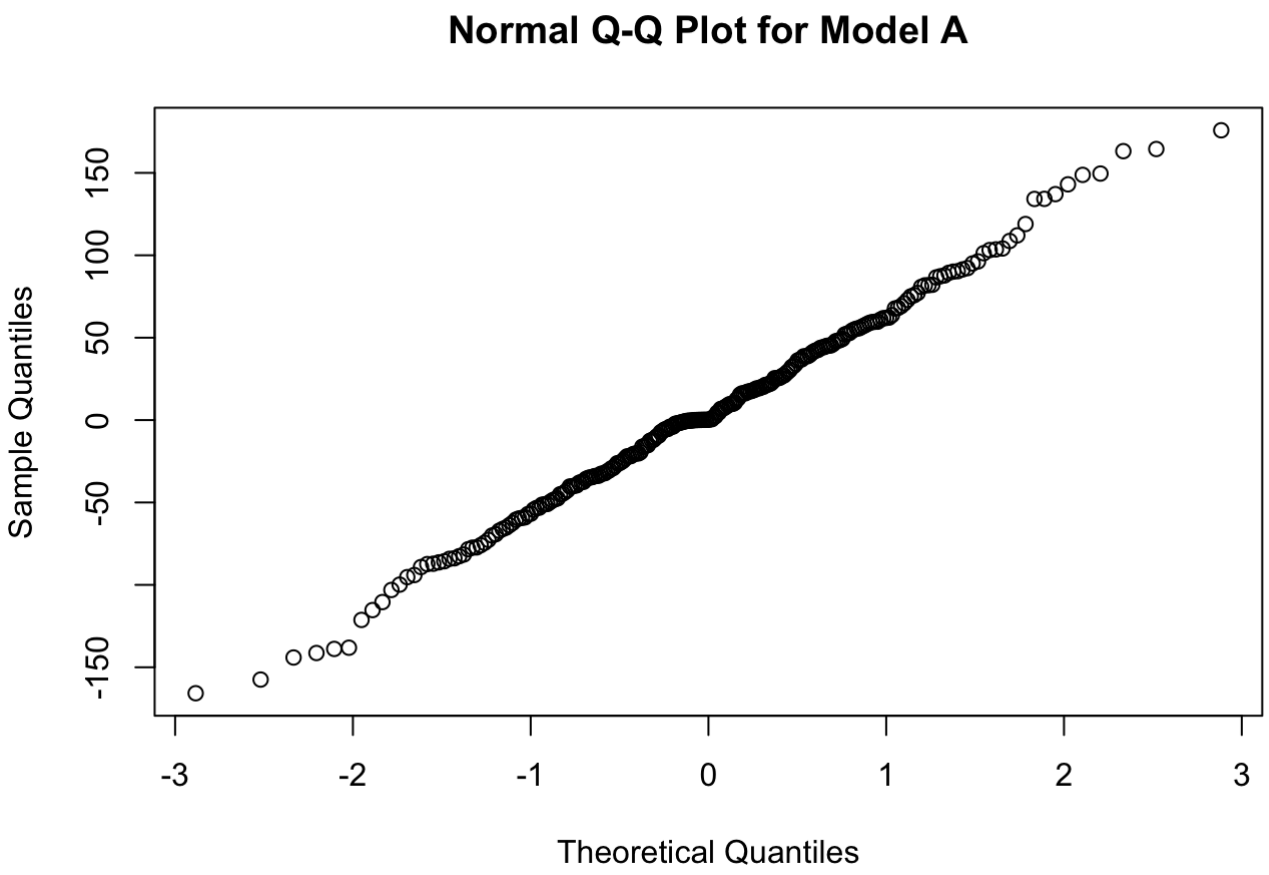
Residuals plot shows no trend, no seasonality, and has a constant mean and constant variance. Therefore it resembles White Noise

To check for normality

```
hist(res_A,density=20,breaks=20, col="blue", xlab="", prob=TRUE, main='Histogram of Mode
l A residuals')
curve(dnorm(x, m_res_A, std_res_A), add=TRUE)
```




```
qqnorm(res_A, main='Normal Q-Q Plot for Model A')
```



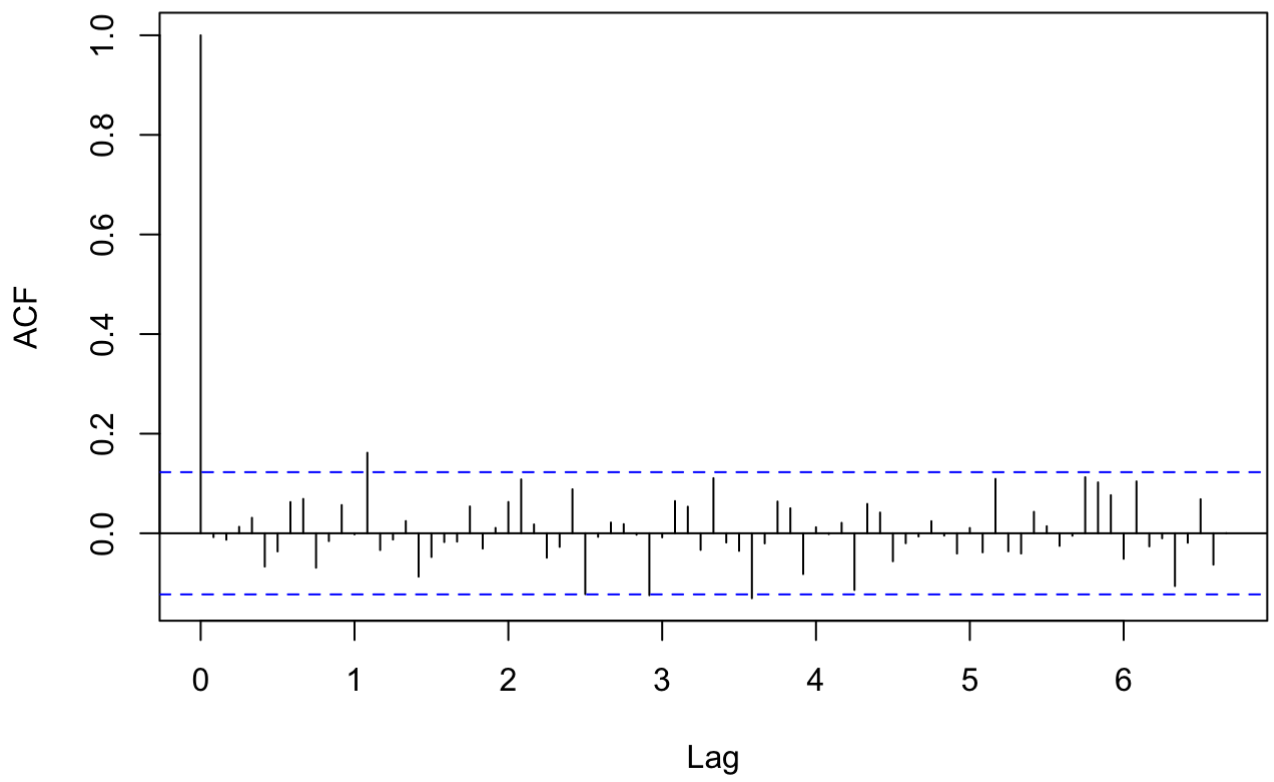

```
shapiro.test(res_A)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_A
## W = 0.99601, p-value = 0.7617
```

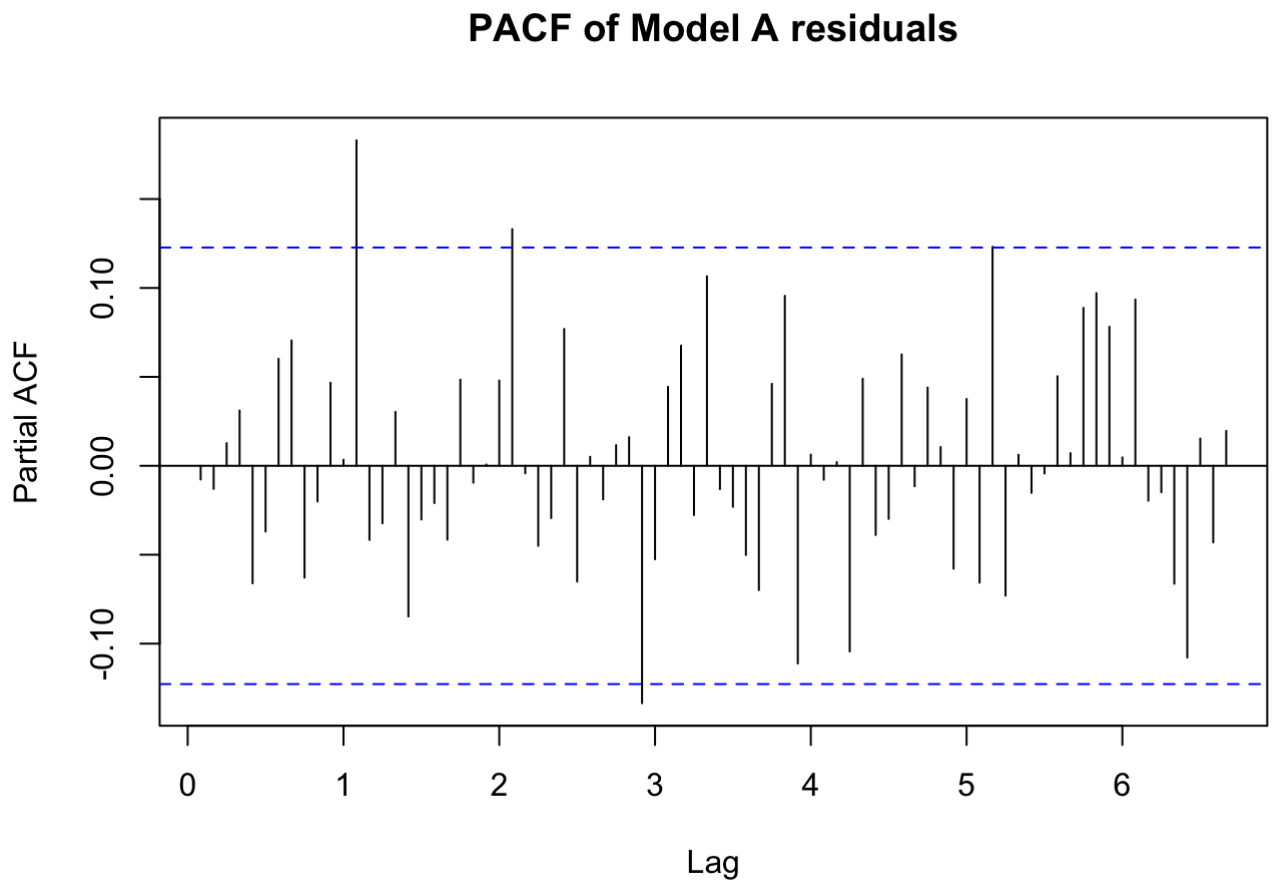
Histogram appears to have a normal distribution Normal Q-Q plot appears to have the residuals lie approximately on a straight line From the histogram and normal q-q plot, it appears that the residuals resemble Gaussian After running the Shapiro-Wilk test of normality, the p-value of 0.7617 is greater than 0.05, so we fail to reject the hypothesis that the residuals are normally distributed.

Check if sample acf/pacf of residuals resemble WN

```
acf(res_A, lag.max =80, main='ACF of Model A residuals')
```




```
pacf(res_A, lag.max=80, main='PACF of Model A residuals')
```



ACFs are within the confidence interval PACFs almost all are within the confidence interval

Tests to check if residuals resemble WN

```
n <- nrow(gast_bc)
n
```

```
## [1] 255
```

```
lag <- round(sqrt(n))
lag
```

```
## [1] 16
```

N = 255 observations in trained data lag = 16 fitdf = p + q = 1 + 1 = 2 McLeod Li : fitdf = 0

```
Box.test(res_A, lag=16 , type=c("Box-Pierce"), fitdf=2)
```

```
##
##  Box-Pierce test
##
## data:  res_A
## X-squared = 13.291, df = 14, p-value = 0.5038
```

For the Box-Pierce test, the p-value 0.5038 is greater than 0.05 so we fail to reject the hypothesis that the residuals are white noise

```
Box.test(res_A, lag=16 , type=c("Ljung-Box"), fitdf=2)
```

```
##
##  Box-Ljung test
##
## data:  res_A
## X-squared = 13.974, df = 14, p-value = 0.4516
```

For the Box-Ljung test, the p-value 0.4516 is greater than 0.05 so we fail to reject the hypothesis that the residuals are white noise

```
Box.test((res_A)^2, lag=16 , type=c("Ljung-Box"), fitdf=0)
```

```
##
##  Box-Ljung test
##
## data:  (res_A)^2
## X-squared = 20.059, df = 16, p-value = 0.2176
```

For the McLeod-Li test, the p-value 0.2176 is greater than 0.05 so we fail to reject the hypothesis that the residuals are white noise. The residuals are uncorrelated and there is linear dependence

All the diagnostic checkings passed for Model A Check other model A2 to see if it's good as Model A

Analyzing the coefficients of Model A2 to write algebraic form

Model A2 SARIMA (1,1,1) x (0,1,1), set coef for sar1 to zero

```
fit_A2 <- arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(0,1,1), period=12),  meth
od='ML')
fit_A2
```

```
##
## Call:
## arima(x = gast_bc, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 1
2),
##     method = "ML")
##
## Coefficients:
##          ar1      ma1     sma1
##       0.5499  -0.8538  -0.8063
## s.e.  0.0990   0.0657   0.0625
##
## sigma^2 estimated as 4176:  log likelihood = -1358.81,  aic = 2725.63
```

(1-0.5499B)(1-B)(1-B^12)X_t = (1-0.8538B)(1-0.8063B^12)Z_t Differenced at lag 12 and then at lag 1

Checking stationary and invertibility

For seasonal part ALl MA models are stationary

Check SMA1 for Invertibility | -0.8063 | < 1 –> invertible

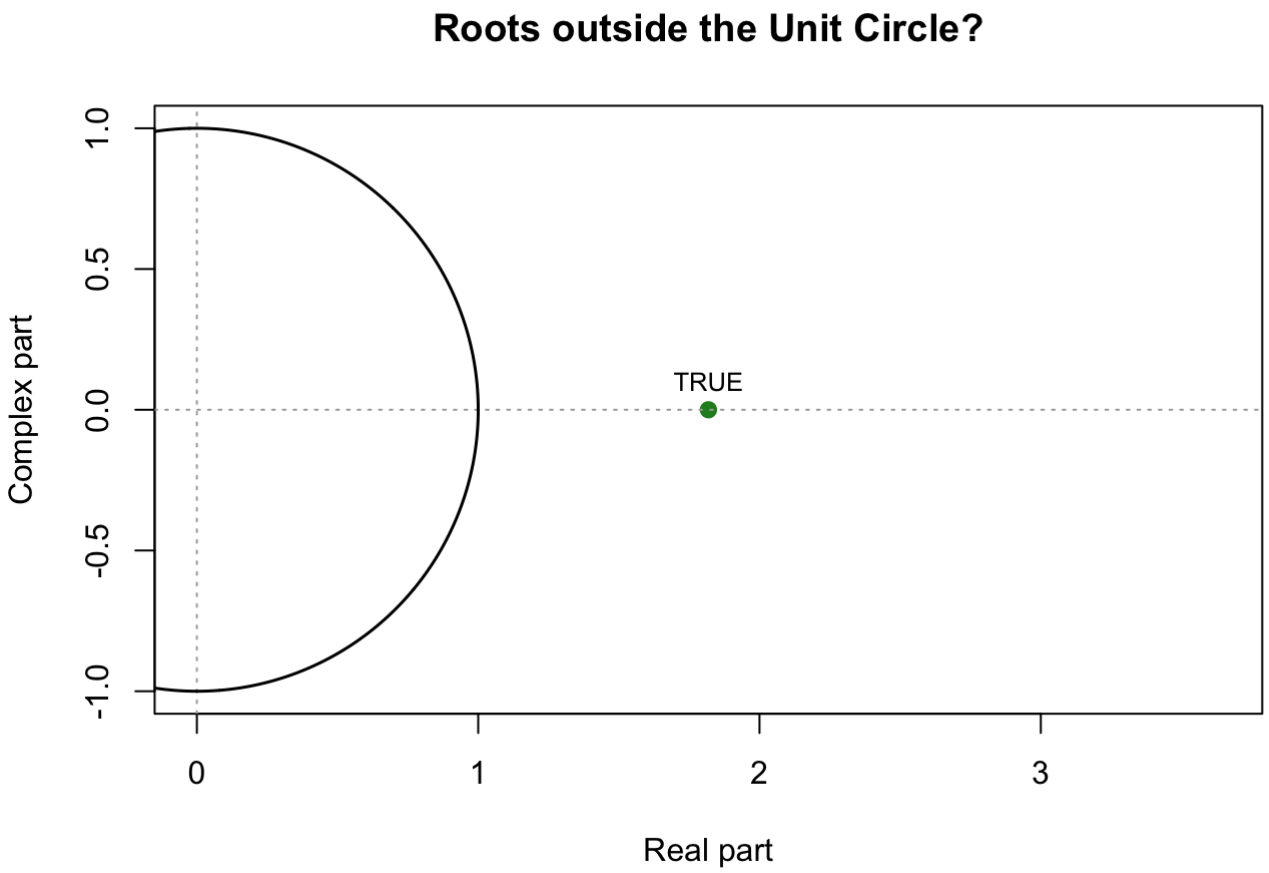
For non-seasonal part

```
# Check AR parts for stationary
# (1-0.5499B)
polyroot(c(1,-0.5499))
```

```
## [1] 1.818512+0i
```

```
library(UnitCircle)
uc.check(pol_ = c(1,-0.5499), plot_output = TRUE)
```

```
##       real complex outside
## 1 1.818512       0    TRUE
## *Results are rounded to 6 digits.
```



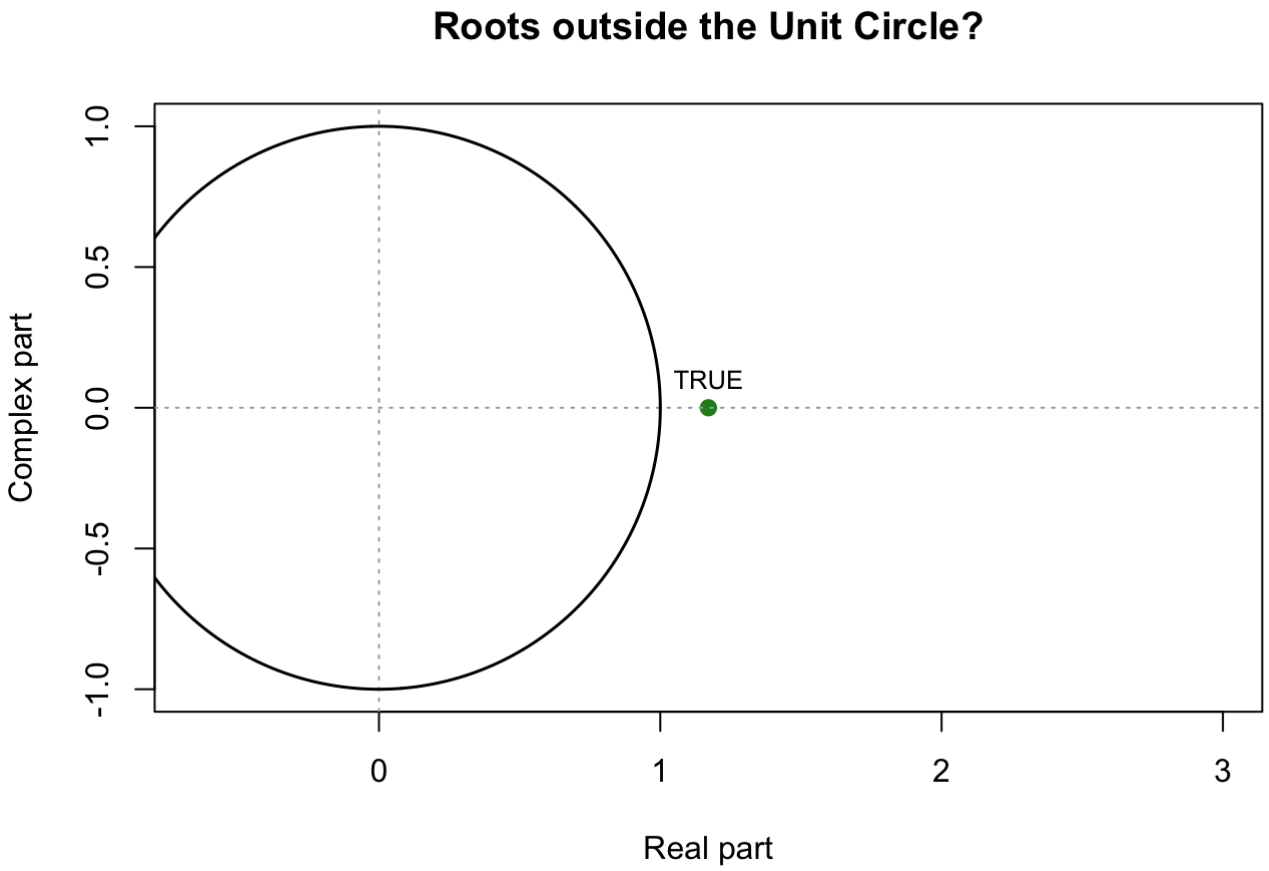
root of polynomial for AR part is outside of the unit circle –> stationary

```
# Check MA parts for invertibility
# (1-0.8538B)
polyroot(c(1,-0.8538))
```

```
## [1] 1.171234+0i
```

```
library(UnitCircle)
uc.check(pol_ = c(1,-0.8538), plot_output = TRUE)
```

```
##       real complex outside
## 1 1.171234       0    TRUE
## *Results are rounded to 6 digits.
```



root of polynomial for MA part is outside of unit circle –> invertible

Model A2 is both stationary and invertible

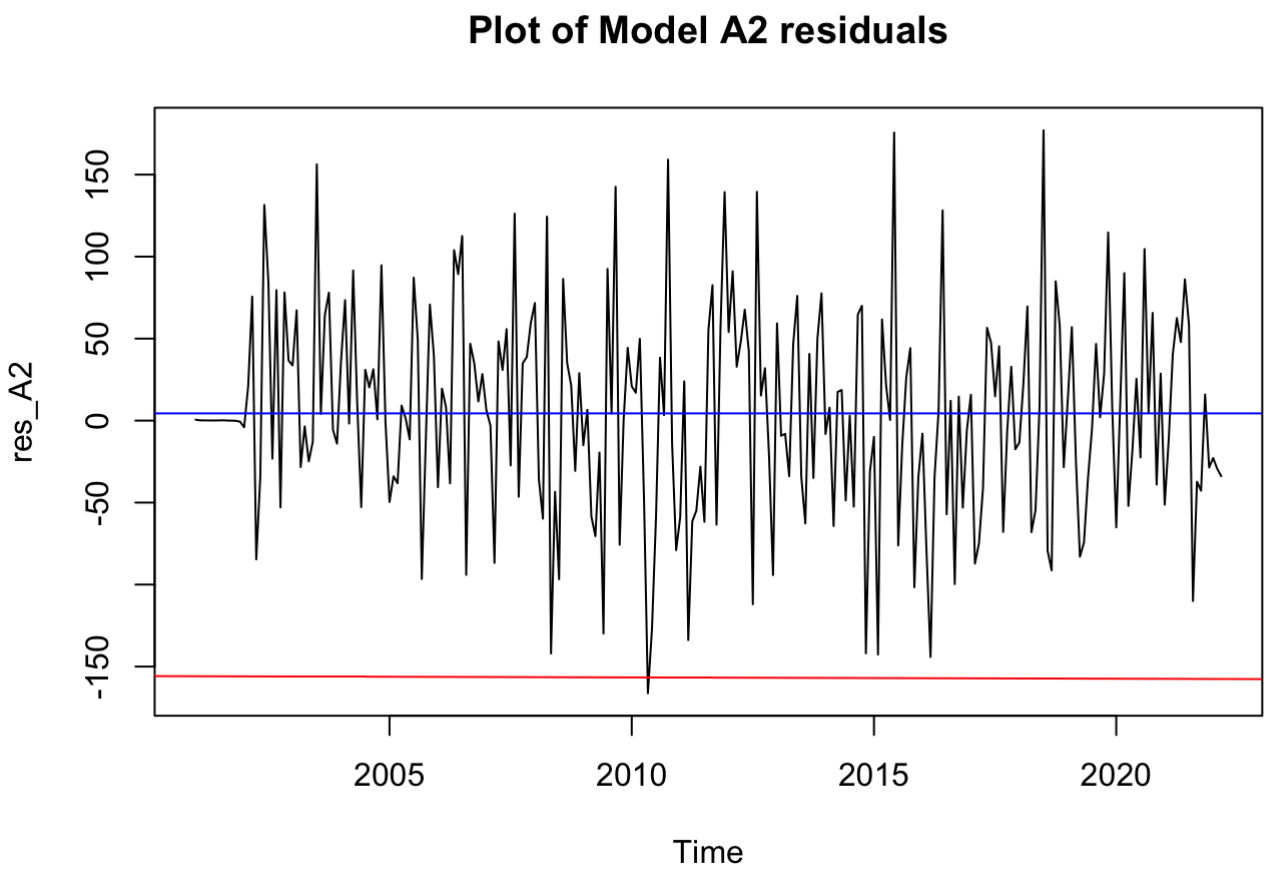
Diagnostic checking for A2

SARIMA (1,1,1) x (0,1,1), with sar1 coeff zero

```
fit_A2 <- arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(0,1,1), period=12),  meth
od='ML')

res_A2 <- residuals(fit_A2)
m_res_A2 <- mean(res_A2)
std_res_A2 <-- sqrt(var(res_A2))
```

```
plot.ts(res_A2, main='Plot of Model A2 residuals')
fit_line_A2 <- lm(res_A2 ~ as.numeric(1:length(res_A2)))
abline(fit_line_A2, col="red")
abline(h=m_res_A2, col="blue")
```

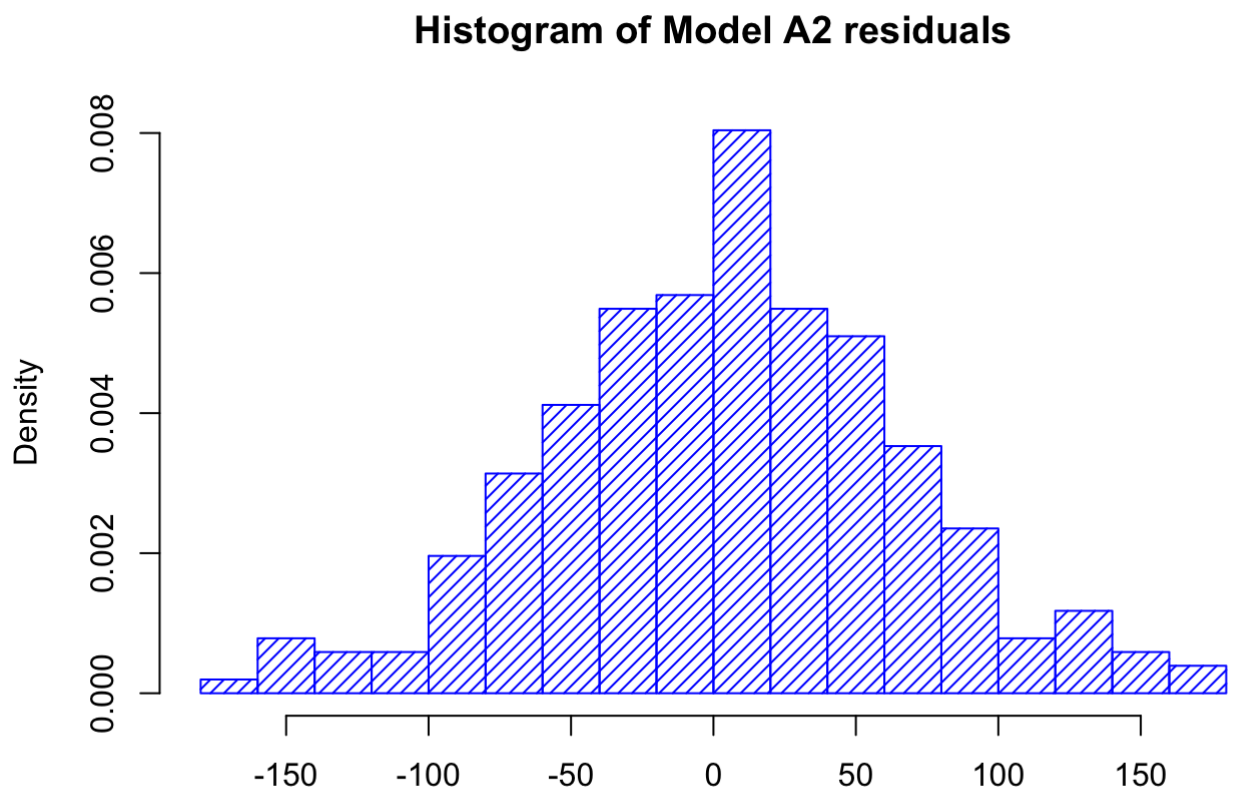


Residuals plot shows no trend, no seasonality, and has a constant mean and constant variance. Therefore it resembles White Noise

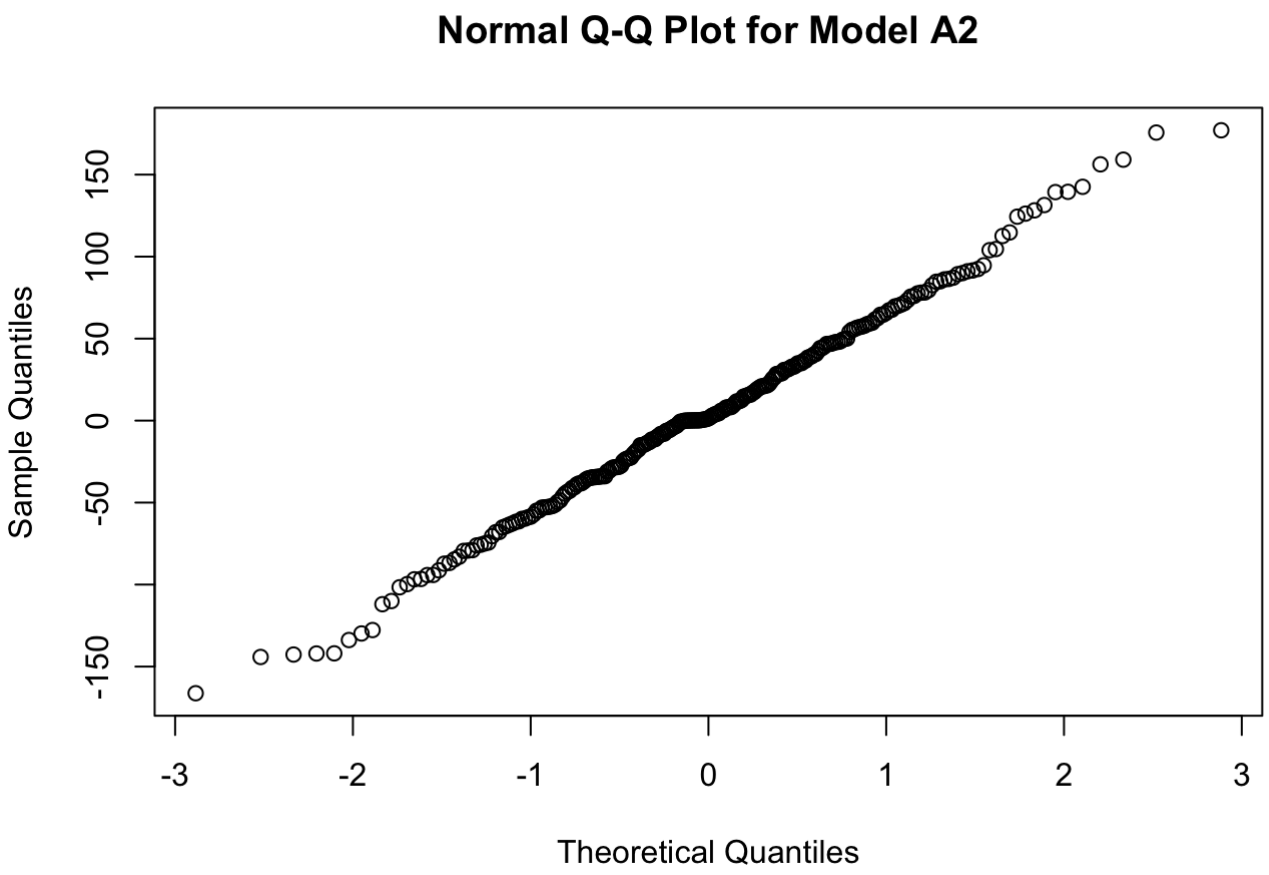
To check for normality

```
hist(res_A2,density=20,breaks=20, col="blue", xlab="", prob=TRUE, main='Histogram of Mod
el A2 residuals')
curve(dnorm(x, m_res_A2, std_res_A2), add=TRUE)
```

```
## Warning in dnorm(x, m_res_A2, std_res_A2): NaNs produced
```




```
qqnorm(res_A2, main='Normal Q-Q Plot for Model A2')
```

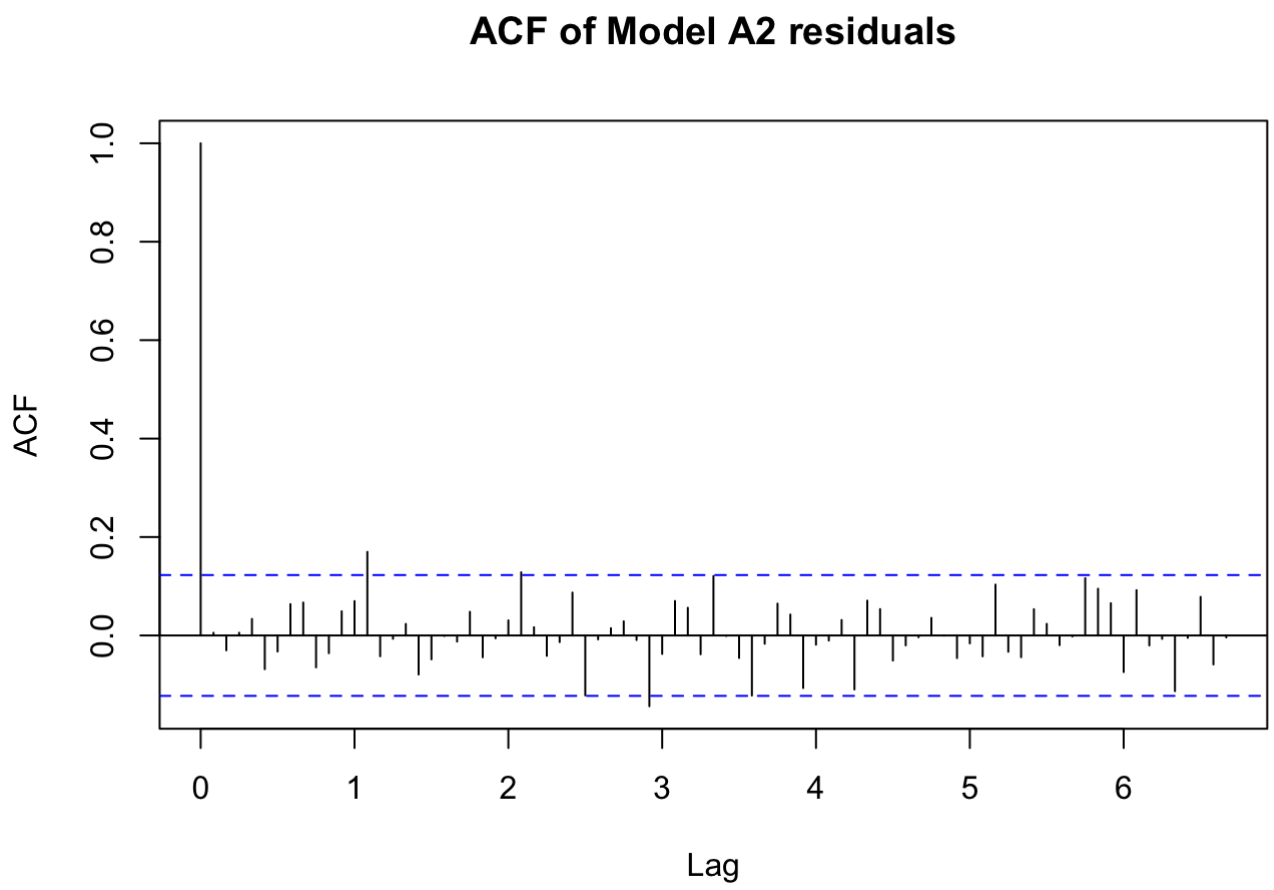



```
shapiro.test(res_A2)
```

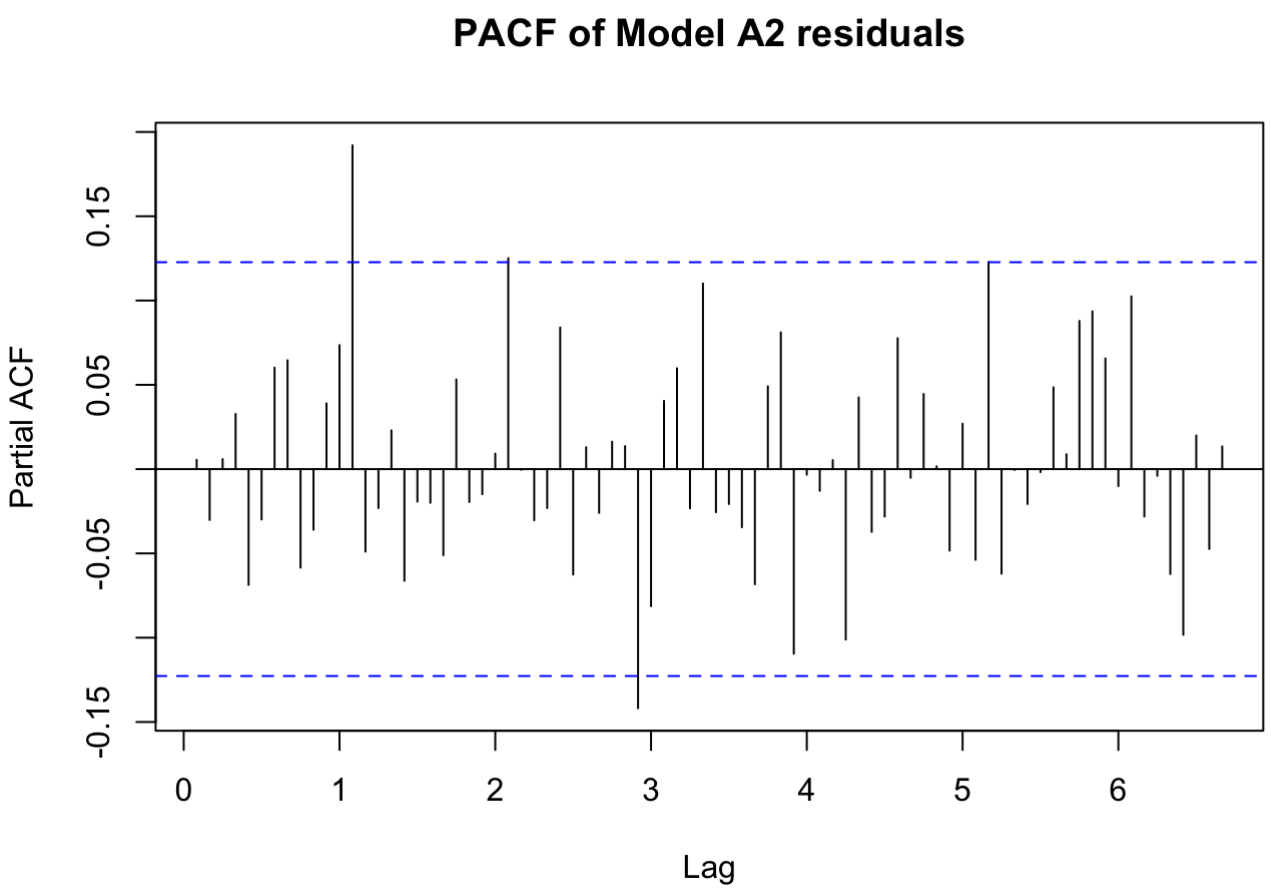
```
##
##  Shapiro-Wilk normality test
##
## data:  res_A2
## W = 0.99594, p-value = 0.7501
```

Histogram appears to have a normal distribution Normal Q-Q plot appears to have the residuals lie approximately on a straight line From the histogram and normal q-q plot, it appears that the residuals resemble Gaussian After running the Shapiro-Wilk test of normality, the p-value of 0.7501 is greater than 0.05, so we fail to reject the hypothesis that the residuals are normally distributed.

```
acf(res_A2, lag.max = 80, main='ACF of Model A2 residuals')
```




```
pacf(res_A2, lag.max = 80, main='PACF of Model A2 residuals')
```



Model A2 has ACF and PACF residuals outside the confidence interval Model A is better than Model A2 Choose Model A as final model

Forecasting using final chosen model Model A

use original data and add confidence/prediction intervals

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
fit_A <- arima(gast_bc, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=12),  metho
d='ML')
forecast(fit_A)
```

```
##          Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
## Apr 2022       578.6530 496.6731  660.6328 453.2756  704.0303
## May 2022       588.6917 489.0076  688.3759 436.2380  741.1455
## Jun 2022       715.5428 607.1092  823.9764 549.7079  881.3777
## Jul 2022       935.5926 821.6898 1049.4954 761.3933 1109.7919
## Aug 2022       966.7117 848.7865 1084.6369 786.3606 1147.0628
## Sep 2022       879.8180 758.5843 1001.0518 694.4070 1065.2291
## Oct 2022       853.2106 729.0515  977.3697 663.3257 1043.0955
## Nov 2022       770.2805 643.4177  897.1433 576.2606  964.3004
## Dec 2022       808.1237 678.6956  937.5519 610.1805 1006.0670
## Jan 2023       744.1337 612.2346  876.0328 542.4114  945.8560
## Feb 2023       666.6499 532.3515  800.9482 461.2583  872.0414
## Mar 2023       650.8865 514.2442  787.5288 441.9101  859.8629
## Apr 2023       583.5872 439.5799  727.5945 363.3471  803.8273
## May 2023       582.0948 432.8517  731.3379 353.8471  810.3424
## Jun 2023       695.9485 542.4961  849.4009 461.2633  930.6337
## Jul 2023       911.5137 754.3923 1068.6351 671.2173 1151.8102
## Aug 2023       954.9764 794.4925 1115.4604 709.5374 1200.4155
## Sep 2023       870.6625 707.0030 1034.3220 620.3669 1120.9582
## Oct 2023       844.3952 677.6837 1011.1067 589.4320 1099.3584
## Nov 2023       757.7625 588.0875  927.4375 498.2669 1017.2581
## Dec 2023       797.8660 625.2958  970.4363 533.9426 1061.7894
## Jan 2024       735.6838 560.2751  911.0925 467.4193 1003.9482
## Feb 2024       660.1429 481.9529  838.3329 387.6248  932.6610
## Mar 2024       645.9860 465.0606  826.9115 369.2844  922.6877
```

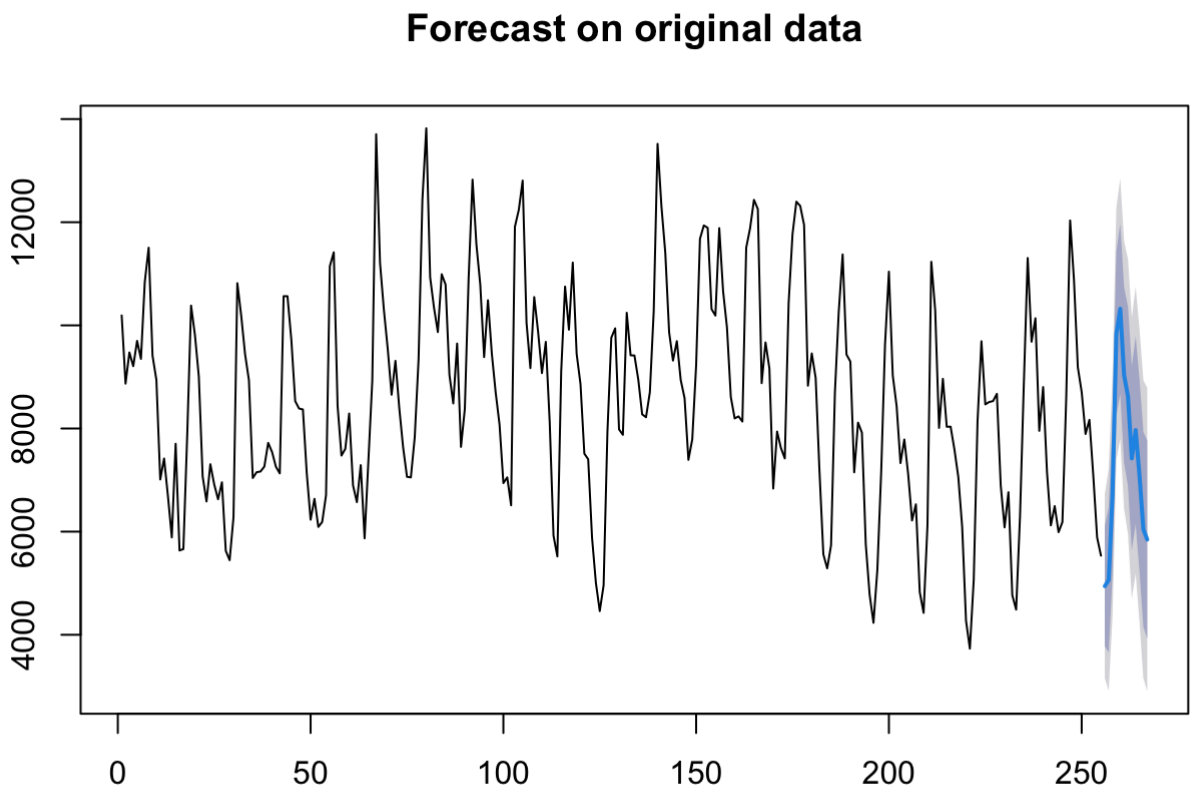
These are the forecasted values from April 2022 to March 2024. We will compare the forecasted valules from April 2022 to March 2023 with the true values from the test data (the rest of the original EIA data that wasn't in the trained data)

12 forecasts on Original data Plot

```
library(forecast)
library(ggplot2)

gas_orig_ts <- ts(gasprod, start=c(2001,1), frequency=12)
gas_orig_train <- gas_orig_ts[1:255,]
gas_orig_test <- gas_orig_ts[256:267,]

fit_orig <- arima(gas_orig_train, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=1
2),  method='ML')
my_forecast_orig <- forecast(fit_orig, h=12)
plot(my_forecast_orig, main="Forecast on original data")
```



This is the time series plot of the original data but with the forecasted values and confidence intervals. The forecasted values are within the confidence intervals.

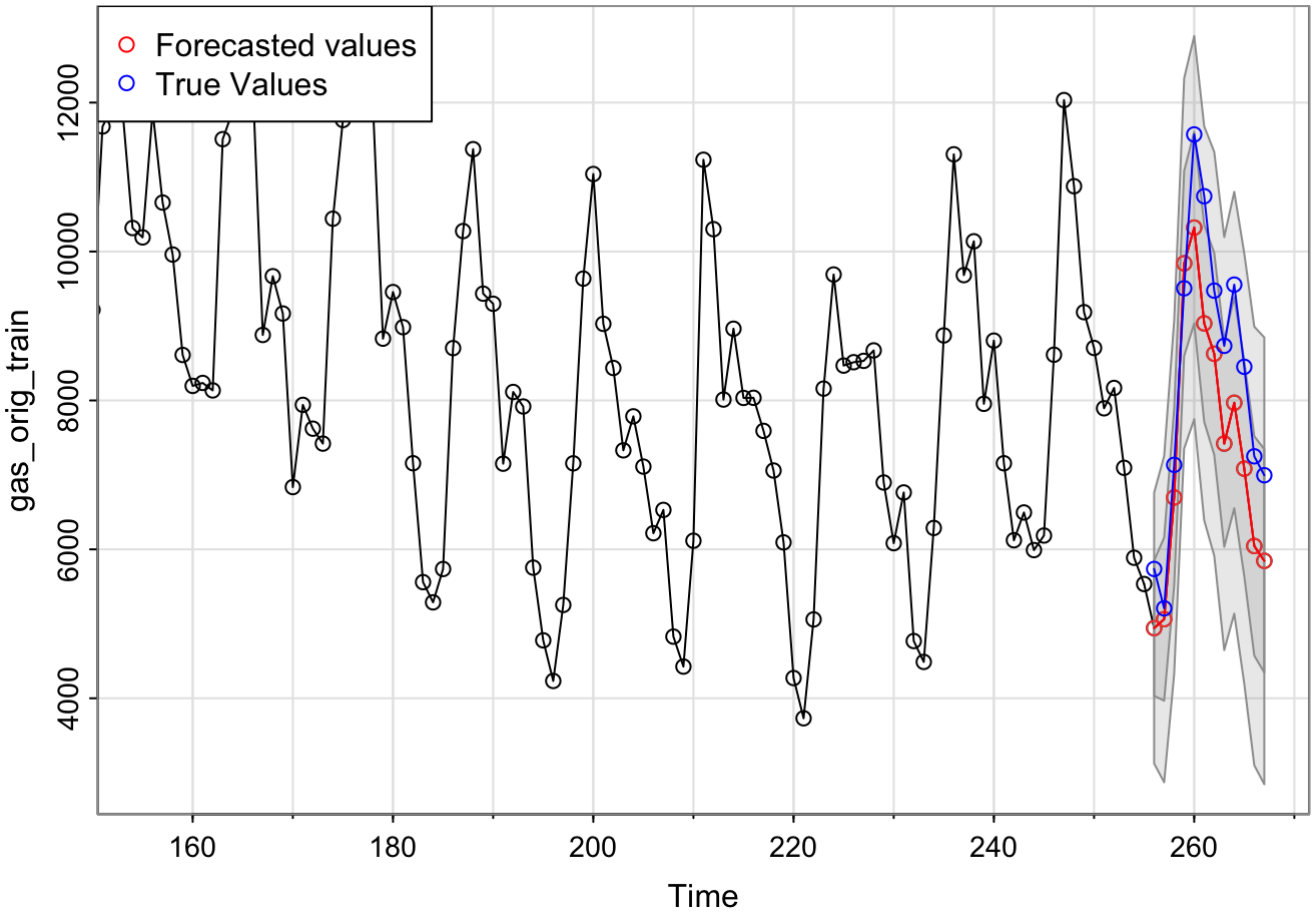
Forecast with original values Zoomed in plot

```
library(astsa)
```

```
##
## Attaching package: 'astsa'
```

```
## The following object is masked from 'package:forecast':
##
##     gas
```

```
pred.tr <- sarima.for(gas_orig_train, n.ahead=12, plot.all=F,
p=1, d=1, q=1, P=1, D=1, Q=1, S=12)
lines(256:267, pred.tr$pred, col="red")
lines(256:267, gas_orig_test, col="blue")
points(256:267, gas_orig_test, col="blue")
legend("topleft", pch=1, col=c("red", "blue"),
legend=c("Forecasted values", "True Values"))
```



This is the time series plot of the original data with forecasted values in red circles and true values in blue circles and confidence intervals in gray. The forecasted values are very close to the true values and they are both within the confidence intervals. This Model A performs well and can predict future values pretty accurately