Monthly Natural Gas Price Forecast

Gerald Ayers

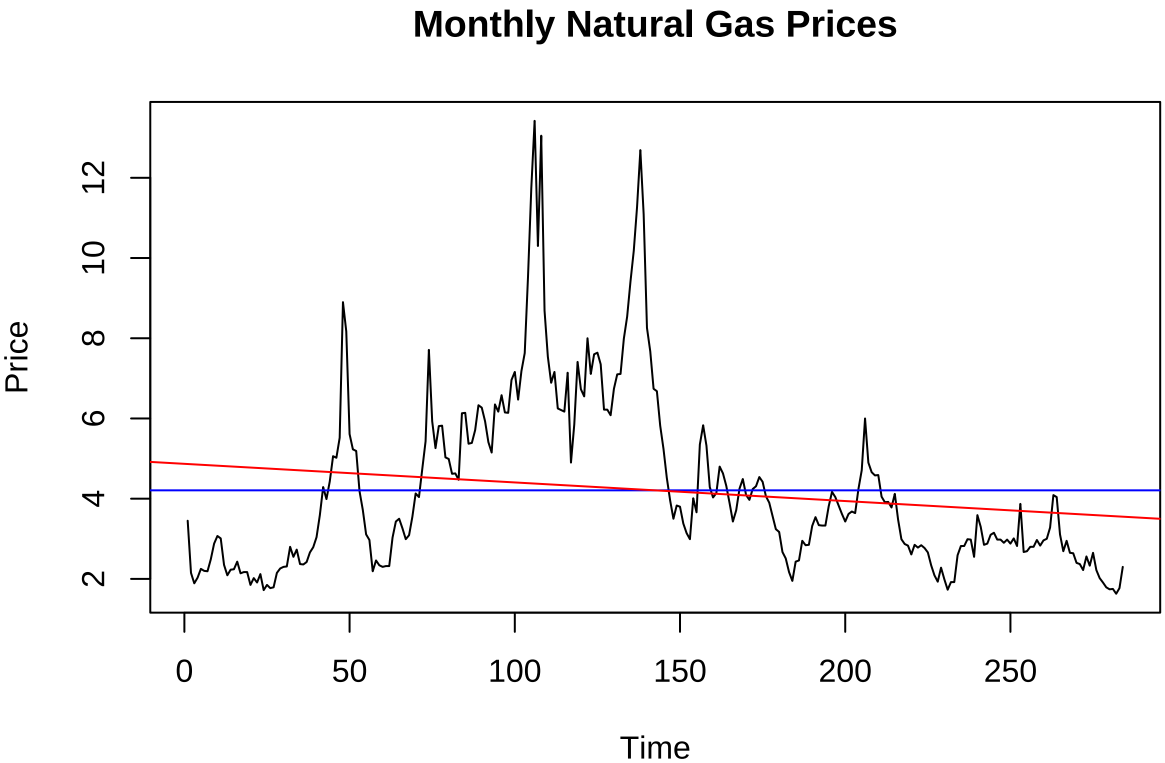
2025-03-03

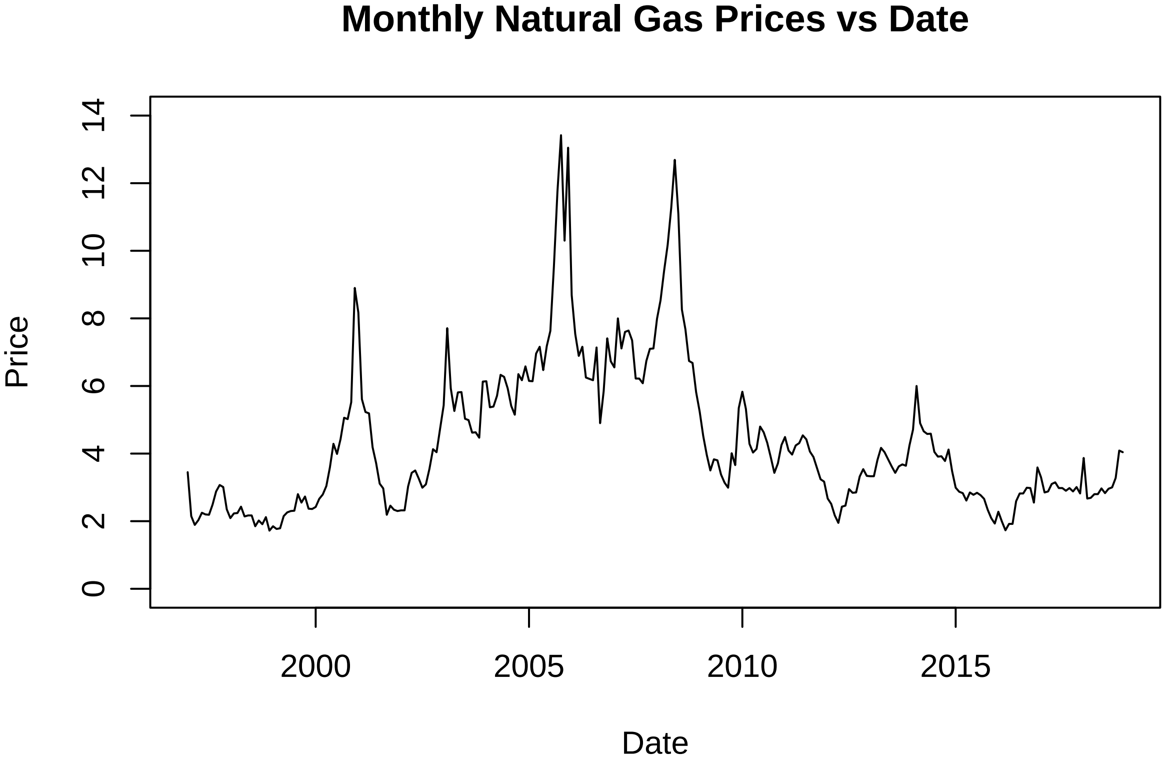
Abstract:

This project examines monthly natural gas prices from January 1997 to August 2020 using time series analysis techniques to fit the data to the correct model and test forecast accuracy. The data was split into train/test sets to check the accuracy of the model forecast on the test data. I performed a Box-Cox Transformation of the data to stabilize the variance, and then differenced at lag one to remove trend from the data so that the data became stationary. Using this stationary data, I fit the transformed data to SARIMA models to find which model gave the best fit. From this a $SARIMA(0,1,0)x(1,0,0)$ model was identified as the best, and the corresponding coefficient was estimated. I performed diagnostic checks to ensure my model was a correct fit for the data and finally forecast the future values. The forecast values were within the 95% prediction interval of the model, although due to high uncertainty, the prediction interval was quite wide.

Introduction:

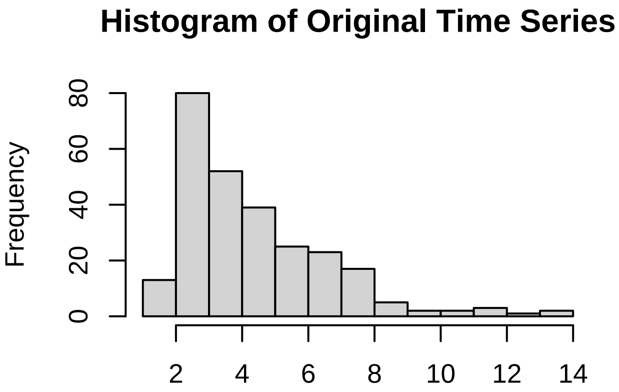
This project examines monthly natural gas prices from January 1997 to August 2020 using time series analysis techniques to fit the data to the correct model and test forecast accuracy. The data has 284 observations in total in which 264 observations, which correspond to the prices from the years 1997-2018, are going to be used to train the time series model, while the remaining 20 observations, from January 2019-August 2020, will be used to test the accuracy of the forecast for the fitted model. Analyzing natural gas prices is important since natural gases are a cleaner source of energy than other fossil fuels and can be used for a more sustainable energy plan in the future and the burning of natural gas is already used for many everyday things such as heating systems in homes and cooking appliances like stoves and ovens. This natural gas data was fit to a SARIMA model and was used to forecast the future natural gas prices, however due to the nature of the data, a heavy-tailed model would be more appropriate for this model rather than a SARIMA model. Due to this fact, there will be high prediction uncertainty characterized by wide prediction intervals. This natural gas data was sourced from Kaggle.com from the following link: https://www.kaggle.com/datasets/joebeachcapital/natural-gas-prices

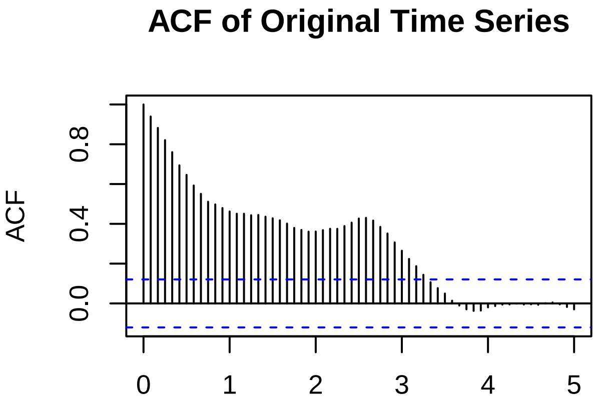

Monthly Natural Gas Prices

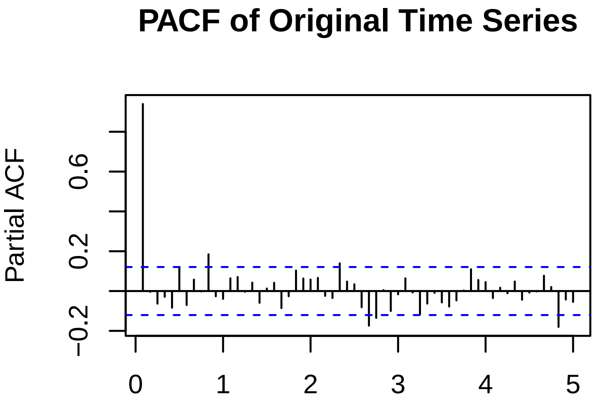

Monthly Natural Gas Prices vs Date

The initial plot of the training set for this time series shows evidence of an unstable variance since there isn't a consistent distance between observations which is an indicator of an unstable variance and a noticeable trend as shown by the slope of red line on the graph of the time series. This issue can be corrected by a Box-Cox transformation of the data to create stable variance. I'm going to check skew of the histogram of the data and the variance to confirm the need for a transformation.

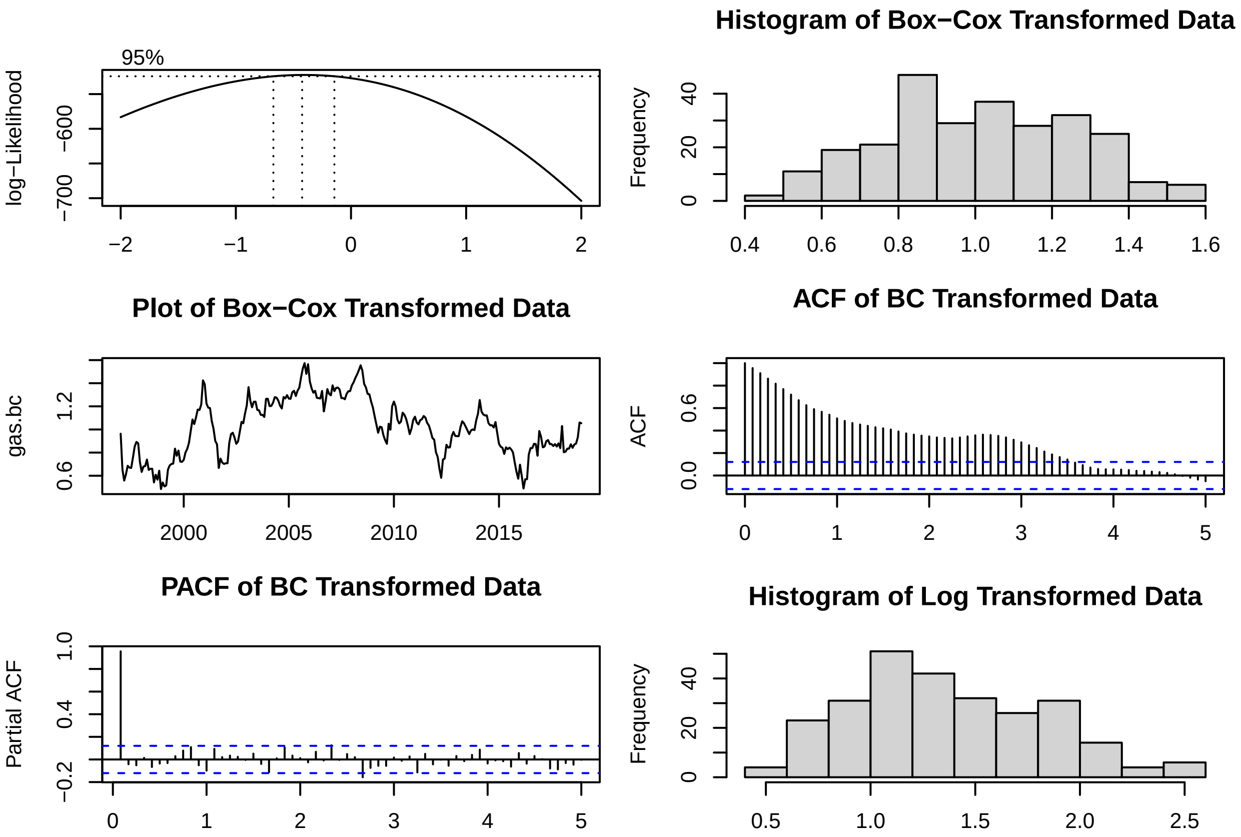
```
## [1] "Variance of Original Data 4.81391468919231"
```







The histogram of the time series seems skewed and looks to have a non-constant and high variance, therefore we want to perform a Box-Cox transformation on the data which should stabilize and reduce the variance. The ACF also shows a pattern which is a sign of trend or seasonality in the data, so we will apply differencing.

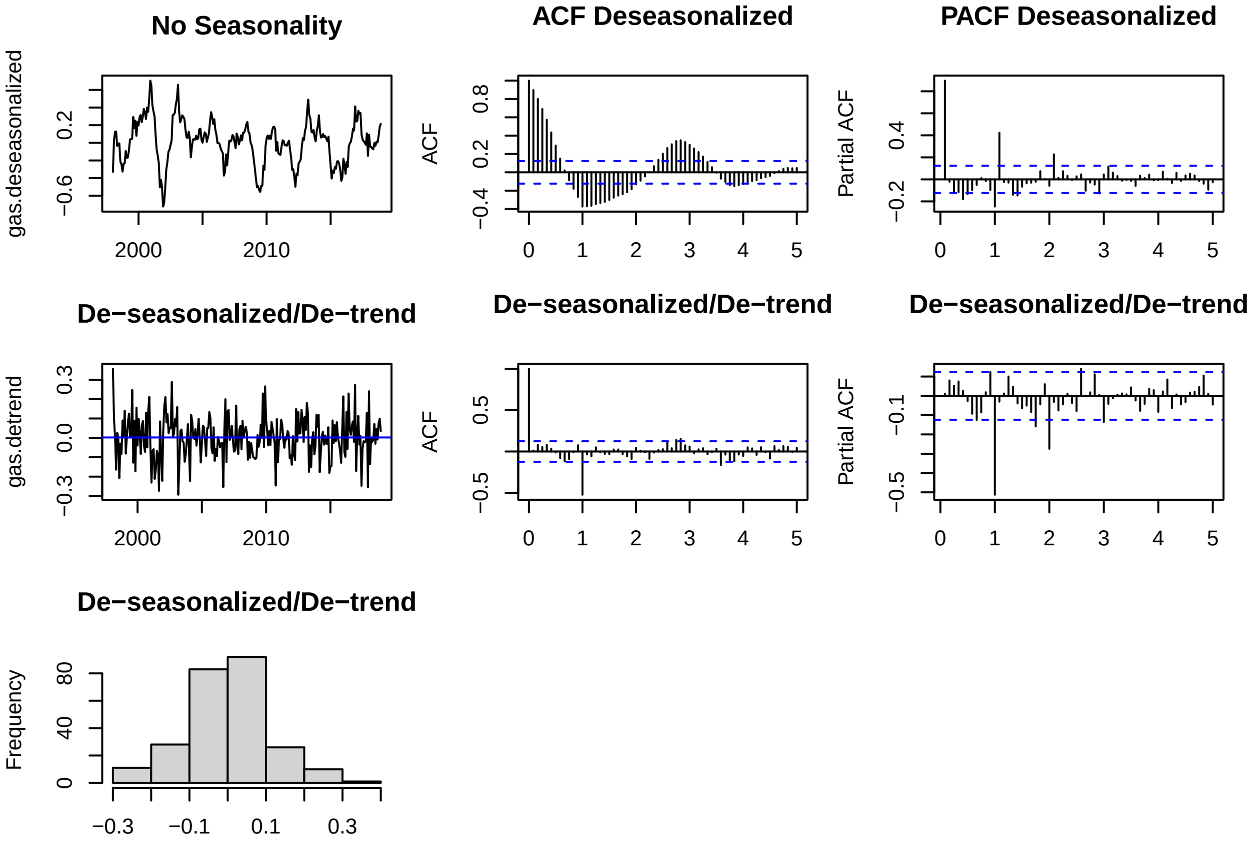


```
## [1] "Variance of Box-Cox Transformed Data 0.0625968123765194"
```

After performing the Box-Cox transformation of the data, we see a significant reduction in the variance and the histogram looks more Gaussian. I also tested a log transformation of the data, but the Box-Cox Transformation gave a more symmetric histogram with more even variance. Now I need to eliminate trend and seasonality from the data.

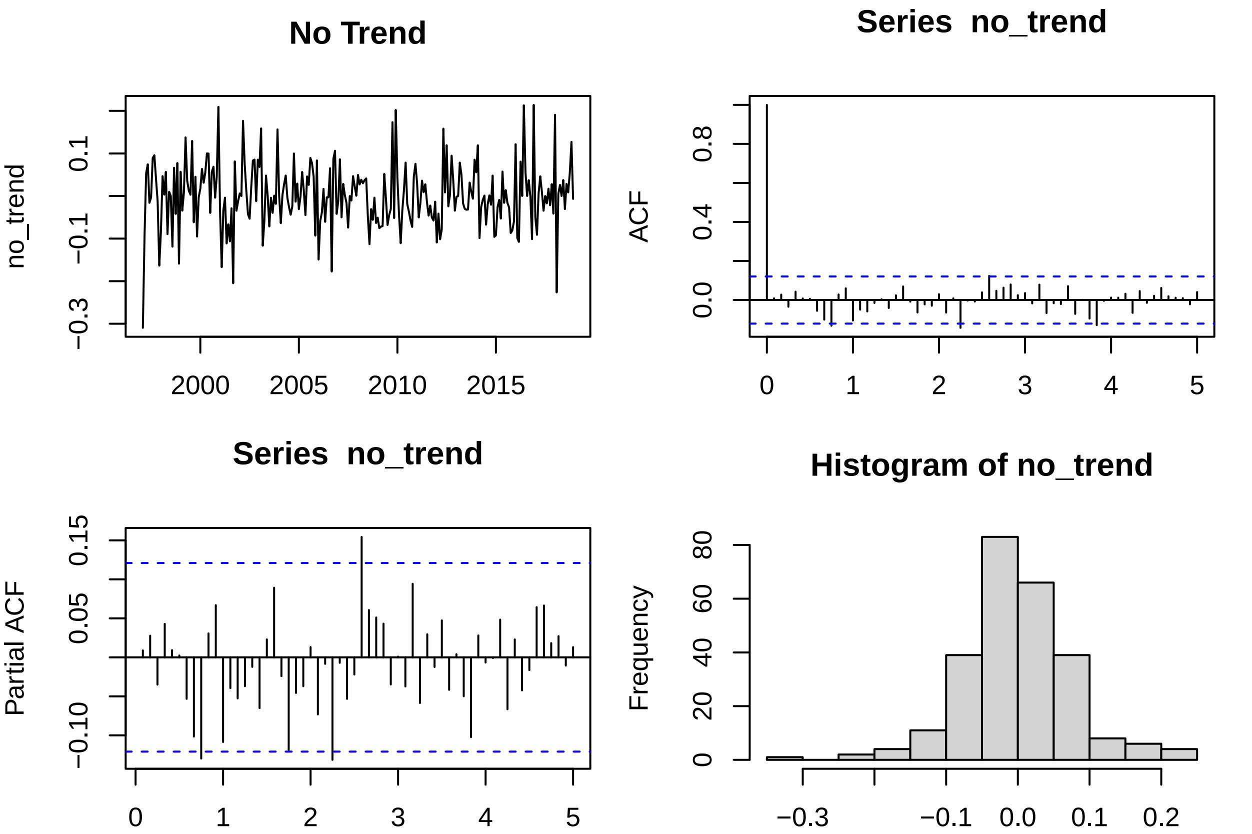
```
## [1] "Variance of De-Seasonalized Data 0.0592708357938336"
```

```
## [1] "Variance of De-seasonalized and De-trended Data 0.0116555128201719"
```



The ACF and PACF seemed to show seasonality in the data so, to remove the seasonality from the data, I differenced at lag 12, and see a further reduction in the variance, showing seasonality was removed. After removing seasonality, we see the ACF still has a pattern, therefore we difference at lag 1 to remove the trend. Following this, we see another reduction in variance, the plot of the series looks stationary, and the ACF and PACF now show evidence of a SMA(1) component. The histogram of the data after the transformation and differencing applied now appears almost Gaussian so we now have stationary data. From the plots of ACF and PACF, we see the data seems to be pure SMA(1) since the ACF is 0 after lag 1, and the PACF decays towards zero, therefore the choices for p and q are P = p = 0, q = 0, and Q = 1. After initial testing of this model, I found that the lowest AICc model of SMA(1) had a coefficient of -1 indicated by model 5, meaning the unit root shows overdifferencing in the data. To correct this I return to the Box-Cox transformed data and difference only at lag 1 to remove trend which achieves stationarity.

```
## [1] "Variance of Only De-trended Data 0.00551309486951797"
```



The ACF and PACF plots show SAR(1) behavior, so now I assume my P = 1, p = 0, q = 0, and Q = 1 for the best fitting model, and I'll test other possible candidate models to see which achieves the lowest AICc value.

```
##
## Call:
## arima(x = gas.bc, order = c(0, 1, 0), seasonal = list(order = c(0, 1, 1), period = 12),
##     method = "ML")
##
## Coefficients:
##           sma1
##        -1.0000
## s.e.    0.0692
##
## sigma^2 estimated as 0.00541:  log likelihood = 280.37,  aic = -556.75

## [1] -556.7318

##
## Call:
## arima(x = gas.bc, order = c(0, 1, 1))
##
## Coefficients:
##           ma1
##        0.0091
## s.e.  0.0618
##
## sigma^2 estimated as 0.005492:  log likelihood = 311.21,  aic = -618.42

## [1] -618.4064
```

```
##
## Call:
## arima(x = gas.bc, order = c(1, 1, 0))
##
## Coefficients:
##          ar1
##       0.0096
## s.e.  0.0637
##
## sigma^2 estimated as 0.005492:  log likelihood = 311.21,  aic = -618.42
```

```
## [1] -618.4077
```

```
##
## Call:
## arima(x = gas.bc, order = c(3, 1, 0), seasonal = list(order = c(1, 0, 0), period = 12),
##     method = "ML")
##
## Coefficients:
##          ar1     ar2      ar3     sar1
##       0.0111  0.0181  -0.0587  -0.1313
## s.e.  0.0636  0.0643   0.0650   0.0672
##
## sigma^2 estimated as 0.005398:  log likelihood = 313.38,  aic = -616.75
```

```
## [1] -616.5995
```

```
##
## Call:
## arima(x = gas.bc, order = c(0, 1, 0), seasonal = list(order = c(1, 0, 0), period = 12),
##     method = "ML")
##
## Coefficients:
##          sar1
##       -0.1231
## s.e.   0.0662
##
## sigma^2 estimated as 0.005417:  log likelihood = 312.91,  aic = -621.82
```

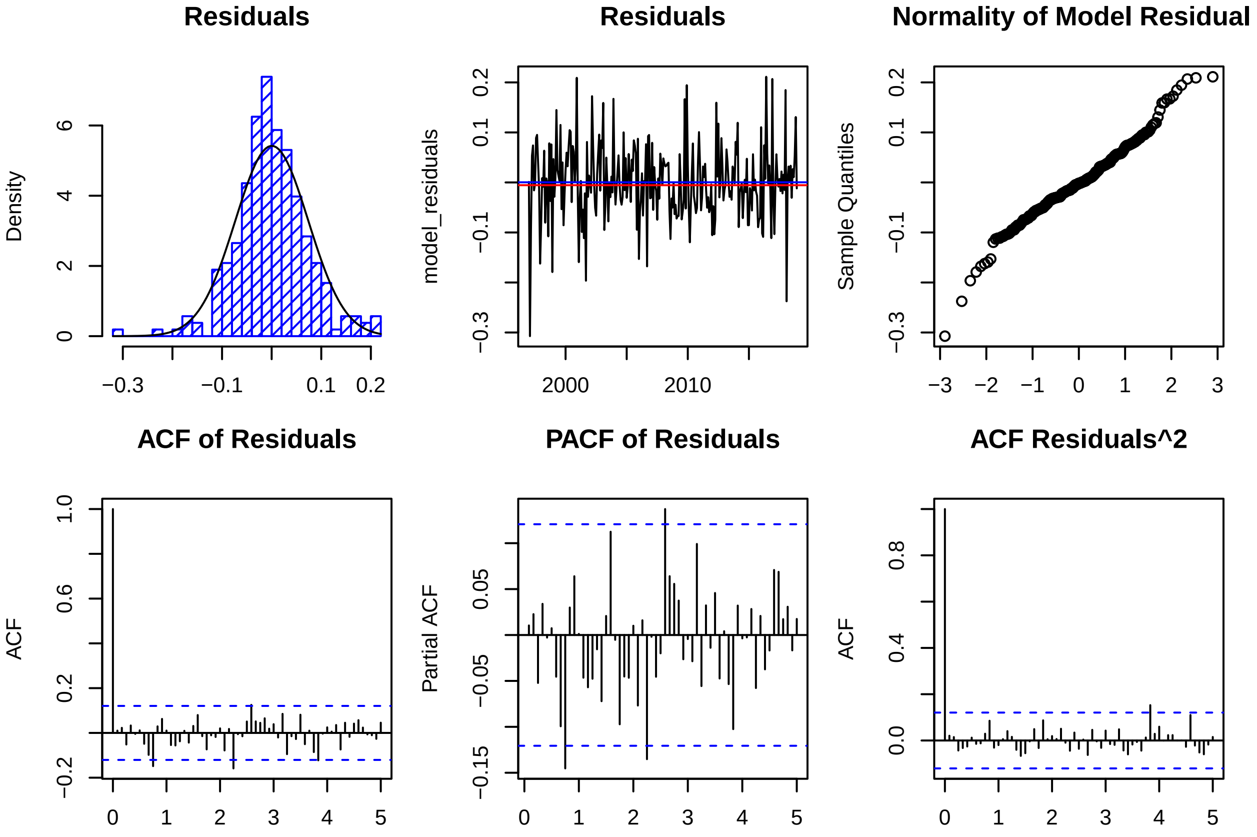
```
## [1] -621.809
```

The best fitted model which is characterized by the lowest AICc value is $SARIMA(0,1,0)x(1,0,0)_{12}$, therefore I continue to use the only de-trended Box-Cox transform data. This is the model as suggested by the plots of ACF/PACF since those plots seem to be consistent with behavior of a SAR(1) with seasonality = 12 process as the PACF's are significant at lags consistent with seasonality 12. The fitted model in algebraic form is $(1-B)(1+0.1231B^{12})X_t = Z_t$ which expanded is of the form $X_t - X_{t-1} - 0.1231X_{t-12} + 0.1231X_{t-13} = Z_t$, where X_t represents the Box-Cox transformation of the original data. The root of the SAR(1) component is outside the unit circle since $\Phi_1 = |-0.1231| < 1$, therefore the model is stationary and invertible. I will now perform diagnostic checks on my model to confirm it's satisfactory.

```
##
##  Shapiro-Wilk normality test
##
## data:  model_residuals
## W = 0.97799, p-value = 0.0004164
```

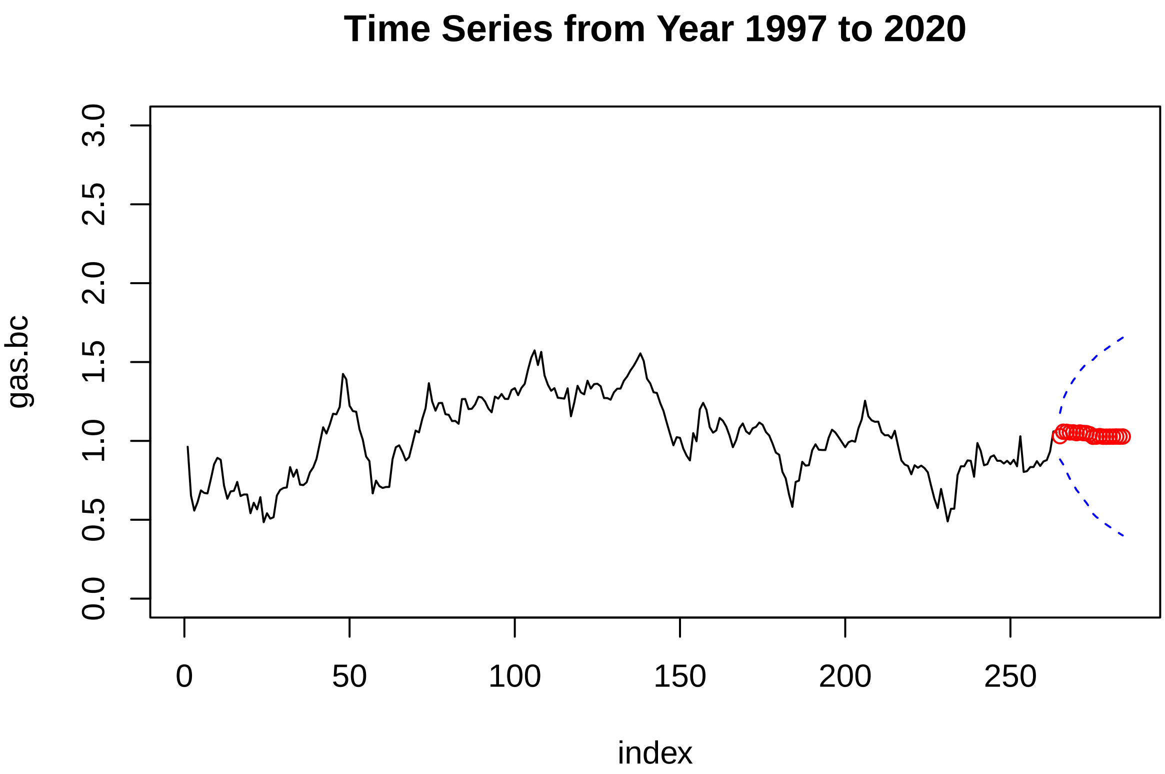
```
##
##  Box-Pierce test
```

```
##
## data:  model_residuals
## X-squared = 13.47, df = 15, p-value = 0.566

##
##  Box-Ljung test
##
## data:  model_residuals
## X-squared = 14.058, df = 15, p-value = 0.5212

##
##  Box-Ljung test
##
## data:  model_residuals^2
## X-squared = 4.9079, df = 16, p-value = 0.9962
```

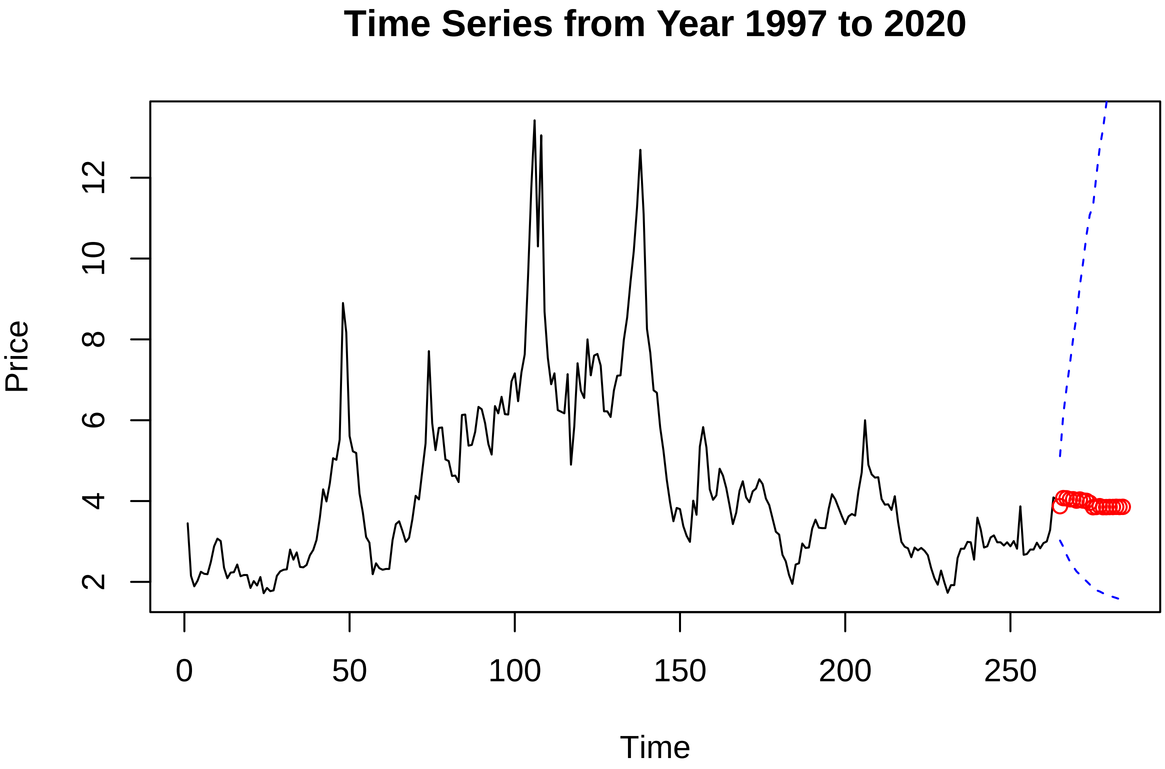


```
##
## Call:
## ar(x = model_residuals, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  0.005417
```

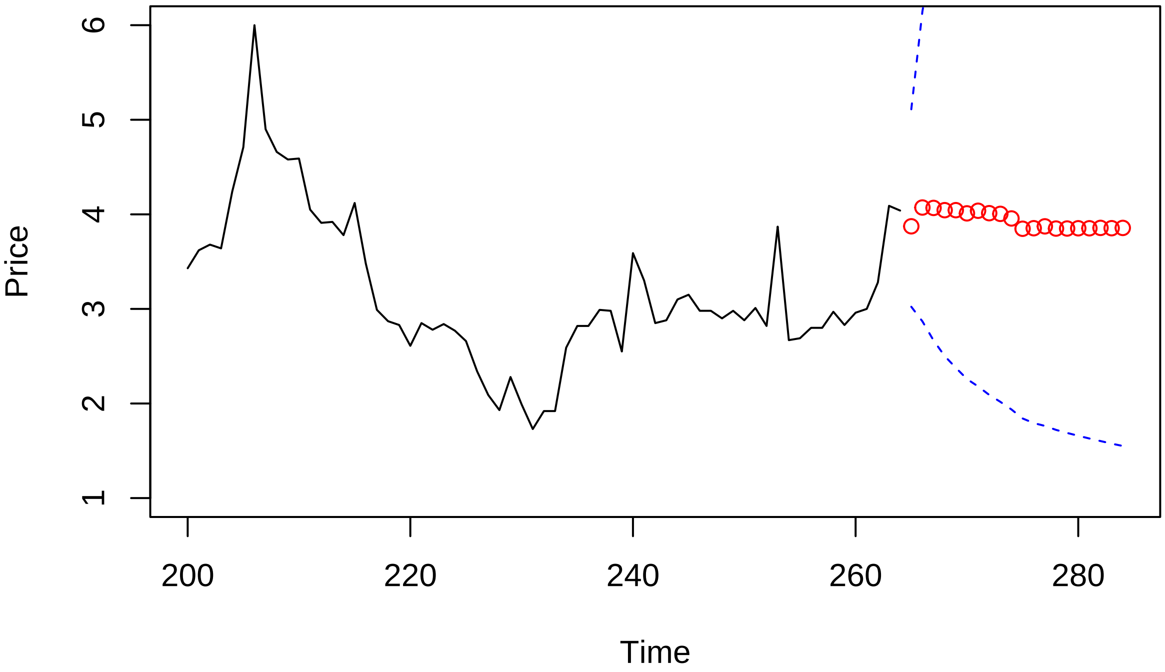
My chosen model passes all of the diagnostic checks except the Shapiro-Wilk test for normality since the resultant p-value < 0.05. This qq normality plot of the residuals seems to show deviation from normality in the tails. This suggests that a heavy-tailed model would work better for fitting the data in this case and explains the failure of the Shapiro-Wilk normality test. With the model now checked, I can use it for forecast.



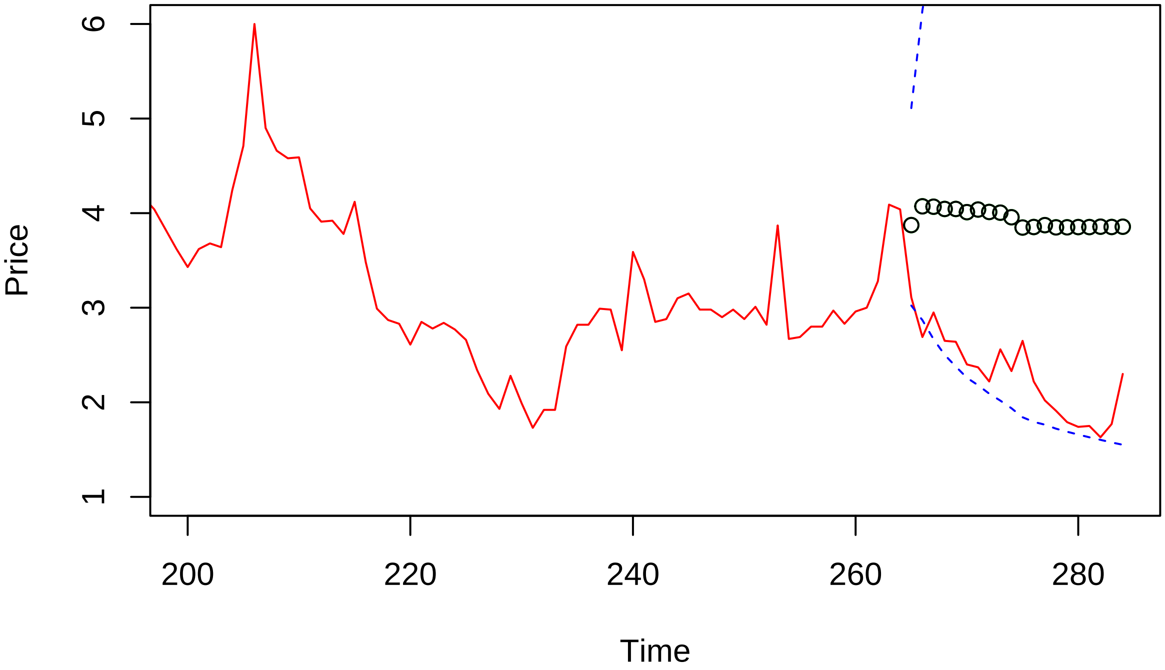
This plot examines the forecast values for the next 20 months on the Box-Cox Transform of the data. The confidence intervals seem to be quite wide which means there's high prediction uncertainty.



This plot shows the forecast values transformed and plotted back onto the original time series data. The transformation back to the original data causes the prediction intervals to be even wider than before.



This graphs shows the same as the previous except zoomed in to show only values starting at index 200 for better visualization of the forecast points.



This graph includes the training and test set so we can see how accurate our forecast was to the original data. The test set is within the prediction intervals, however the forecast values for the data points are off due to high prediction uncertainty of the observations.

Conclusion:

The natural gas time series data was able to be fit to a $SARIMA(0,1,0)x(1,0,0)_{12}$ model with the equation: $X_t - X_{t-1} - 0.1231X_{t-12} + 0.1231X_{t-13} = Z_t$, where X_t represents the Box-Cox transformation of the original data. This model was able to be used to forecast future natural gas prices as it was shown that the 95%

prediction interval contained the true values, however the data had high prediction uncertainty. The model can be further improved by using a heavy-tailed model as was shown by the failure of the Shapiro-Wilk normality test, and the deviation from normality shown by the qq plot of the model residuals.

Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
nat_gas_prices = read.csv("/Users/jerryayers/Downloads/archive (3)/monthly_csv.csv")

nat_gas_train = nat_gas_prices[c(1:264),]
nat_gas_test = nat_gas_prices[c(265:284),] #last year and 8 months since last year of data is incomplet


plot.ts(1:284, nat_gas_prices$Price, main = "Monthly Natural Gas Prices", ylab = "Price", xlab = "Time"
abline(h = mean(nat_gas_prices$Price), col = "blue")
fit <- lm(nat_gas_prices$Price ~ as.numeric(1:length(nat_gas_prices$Price))); abline(fit, col="red")

nat_gas_ts = ts(nat_gas_train$Price, start = c(1997, 01), frequency = 12)

plot.ts(nat_gas_ts, main = "Monthly Natural Gas Prices vs Date", xlab = "Date", ylab = "Price", ylim =


library(MASS)
par(mfrow = c(2,2), mar = c(2,4,4,1))
hist(nat_gas_ts, main = "Histogram of Original Time Series")
acf(nat_gas_ts, lag.max = 60, main = "ACF of Original Time Series")
pacf(nat_gas_ts, lag.max = 60, main = "PACF of Original Time Series")
print(paste("Variance of Original Data", var(nat_gas_ts)))

par(mfrow = c(3,2), mar = c(2,4,4,1))
t = 1:length(nat_gas_ts)
fit = lm(nat_gas_ts ~ t)
bcTransform = boxcox(nat_gas_ts ~ t,plotit = TRUE)

lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
gas.bc = (1/lambda)*(nat_gas_ts^lambda-1)

hist(gas.bc, main = "Histogram of Box-Cox Transformed Data")
plot(gas.bc, main = "Plot of Box-Cox Transformed Data")

acf(gas.bc, lag.max = 60, main = "ACF of BC Transformed Data")
pacf(gas.bc, lag.max = 60, main = "PACF of BC Transformed Data")

hist(log(nat_gas_ts), main = "Histogram of Log Transformed Data")


print(paste("Variance of Box-Cox Transformed Data", var(gas.bc)))
par(mfrow = c(3,3), mar = c(2,4,4,1))
gas.deseasonalized = diff(gas.bc, 12)
print(paste("Variance of De-Seasonalized Data", var(gas.deseasonalized)))
plot(gas.deseasonalized, main = "No Seasonality")
acf(gas.deseasonalized, lag.max = 60, main = "ACF Deseasonalized")
```

```
pacf(gas.deseasonalized, lag.max = 60, main = "PACF Deseasonalized")

gas.detrend = diff(gas.deseasonalized, 1)
plot(gas.detrend, main = "De-seasonalized/De-trend")
abline(h=mean(gas.detrend), col = "blue")

acf(gas.detrend, lag.max = 60, main = "De-seasonalized/De-trend")
pacf(gas.detrend, lag.max = 60, main = "De-seasonalized/De-trend")

hist(gas.detrend, main = "De-seasonalized/De-trend")
print(paste("Variance of De-seasonalized and De-trended Data", var(gas.detrend)))
par(mfrow= c(2,2), mar = c(2,4,4,1))
no_trend = diff(gas.bc, lag = 1)
print(paste("Variance of Only De-trended Data", var(no_trend)))
plot(no_trend, main = "No Trend")
acf(no_trend, lag.max = 60)
pacf(no_trend, lag.max = 60)
hist(no_trend)
library(qpcR)
set.seed(444)

model_5 = arima(gas.bc, order = c(0,1,0), seasonal = list(order = c(0,1,1), period = 12), method = "ML"
model_5
AICc(model_5)

model_6 = arima(gas.bc, order = c(0,1,1))
model_6
AICc(model_6)

model_7 = arima(gas.bc, order = c(1,1,0))
model_7
AICc(model_7)

model_8 = arima(gas.bc, order = c(3,1,0), seasonal = list(order = c(1,0,0), period = 12), method = "ML"
model_8
AICc(model_8)

model_9 = arima(gas.bc, order = c(0,1,0), seasonal = list(order = c(1,0,0), period = 12), method = "ML"
model_9
AICc(model_9) #correct model
par(mfrow= c(2,3), mar = c(2,4,4,1))
model_correct = model_9
model_residuals = residuals(model_9)
hist(model_residuals,density=20,breaks=20, col="blue", xlab="", prob=TRUE, main = "Residuals")
mean <- mean(model_residuals)
std <- sqrt(var(model_residuals))
curve( dnorm(x,mean,std), add=TRUE )
plot.ts(model_residuals, main = "Residuals")
fit_model <- lm(model_residuals ~ as.numeric(1:length(model_residuals))); abline(fit_model, col="red")
abline(h=mean(model_residuals), col="blue")

qqnorm(model_residuals, main = "Normality of Model Residuals")
acf(model_residuals, lag.max = 60, main = "ACF of Residuals")
```

```r
pacf(model_residuals, lag.max = 60, main = "PACF of Residuals")
shapiro.test(model_residuals)
Box.test(model_residuals, lag = 16, type = c("Box-Pierce"), fitdf = 1)
Box.test(model_residuals, lag = 16, type = c("Ljung-Box"), fitdf = 1)
Box.test(model_residuals^2, lag = 16, type = c("Ljung-Box"), fitdf = 0)
acf(model_residuals^2, lag.max = 60, main = "ACF Residuals^2")
ar(model_residuals, aic = TRUE, order.max = NULL, method = c("yule-walker"))
library(forecast)
train_prediction = predict(model_correct, n.ahead = 20)
upper = train_prediction$pred+2*train_prediction$se
lower = train_prediction$pred-2*train_prediction$se

plot(1:length(gas.bc),gas.bc, main =
"Time Series from Year 1997 to 2020", type = 'l',xlab='index', xlim = c(1, 284), ylim = c(0,3))
index = 1: length(gas.bc)
lines(265:284, upper, col="blue", lty="dashed")
lines(265:284, lower, col="blue", lty="dashed")
points((length(gas.bc)+1):(length(gas.bc)+20), train_prediction$pred, col="red")
forecast_original = InvBoxCox(train_prediction$pred, lambda)
upper_original = InvBoxCox(upper, lambda)
lower_original = InvBoxCox(lower, lambda)

plot(1:length(nat_gas_ts),nat_gas_ts, main =
"Time Series from Year 1997 to 2020", type = 'l', xlim = c(1, 284), ylab = "Price", xlab = "Time")
index = 1: length(nat_gas_ts)
lines(265:284, upper_original, col="blue", lty="dashed")
lines(265:284, lower_original, col="blue", lty="dashed")
points((length(nat_gas_ts)+1):(length(nat_gas_ts)+20), forecast_original, col="red")


plot(200:264, nat_gas_ts[c(200:264)], xlim = c(200,length(nat_gas_ts)+20), xlab = "Time", ylab = "Price
lines(265:284, upper_original, col="blue", lty="dashed")
lines(265:284, lower_original, col="blue", lty="dashed")
points((length(nat_gas_ts)+1):(length(nat_gas_ts)+20), forecast_original, col="red")

plot.ts(1:284, nat_gas_prices$Price, type = "l", xlim = c(200,length(nat_gas_ts)+20), col="red", xlab =
lines(265:284, upper_original, col="blue", lty="dashed")
lines(265:284, lower_original, col="blue", lty="dashed")
points((length(nat_gas_ts)+1):(length(nat_gas_ts)+20), forecast_original, col="green")
points((length(nat_gas_ts)+1):(length(nat_gas_ts)+20), forecast_original, col="black")
```