Forecasting US gasoline prices

Franck A Olilo

MS.Applied Statistics

Abstract

We show that certain factors causing the rise in gasoline prices discussed by many observers may have been overstated. Some would only briefly raise weekly gasoline prices, before fading rather slowly. However, the short-run effects on the overall price would be sizable. Arima model is used for the prediction of average gasoline price in the US which goes from $4.035 in 2020/4 to $3.68 nine months later before the price starts to rise again. This model performs well in predicting the average price for the upcoming weeks with 95% confidence intervals.

Introduction

Inflation affected nearly any product or service, from expenses such as housing, food, medical care, utilities, to oil and gas. In the aftermath of the disruptions caused by certain factors, natural gas prices rose to record-high levels. Because natural gas is an important energy source for the U.S. economy, there was widespread concern that these high prices might keep rising and further disrupt the economy. In this project, I will focus on Gasoline prices that have risen significantly in recent months. Especially, on What's behind the price spike. Is it the cancelation of the Keystone XL pipeline? Is it the supply disruptions from COVID-19? Is it Russia invading Ukraine? There are many opinions on the issue. I forecasted the weekly price of gas using time series analysis techniques. I first came up with possible models to describe the change in the gas price. We then picked the best model by comparing the performance of each model.

Possible Causes

The Keystone XL Pipeline Cancelation

Initially called TransCanada Keystone Pipeline and later abbreviated here as Keystone pipeline is supposed to transport crude oil from the oil sands region of Alberta, Canada to the existing Keystone Pipeline in Steele City, Nebraska. It is owned by Canadian company TC Energy Corp and the government of Alberta. The KXL would have transported 830,000 barrels per day of heavy oil-sands crude oil. From there existing pipelines would carry the oil to points in the U.S., including refineries in the Gulf South. On November 19, 2014, the U.S. Senate voted against the passage of the bill which would allow the Keystone XL Pipeline to proceed. Was the KXL cancelation a contributing factor in the recent spike in oil and gasoline? Canceling the pipeline had a minimal effect on current prices. The Keystone XL crude oil pipeline wasn't yet operational when it was canceled in 2021, and wasn't expected to be running until 2023. Then high gas prices are due to other factors such as the global spike in the cost of crude oil and increased demand after pandemic lockdowns ended.

COVID-19

The global pandemic caused a significant disruption in global supply chains, including oil. Even though the global supply of oil may have declined during COVID-19, demand was also muted as people were afraid to travel. This kept oil and gasoline prices low. As the world began to emerge from the pandemic, demand increased. Even so, prices remained at the low end.

Patrick De Haan is a graduate of DePaul University with a degree in Business Economics. He has analyzed and tracked oil markets and fuel prices for nearly two decades. On March 10, 2022, he tweeted the following graph explaining the effect of the pandemic on US gas prices.

From the graph, we can clearly see that, during the pandemic, the price of gas continued to declaine as demand for gas declined due to the lockdown. Prices started rising when vaccins were discivered and many people got vaciinated. As people get the vaccin, they return back to work and demand for gas increases; and so the prices.

Russia Invades Ukraine

There has been tensions between Russia and Ukraine since March 2014, Russia invaded and subsequently annexed the Crimean Peninsula from Ukraine. In 2021, following an unsuccessful ceasefire, Ukrainian President Volodymyr Zelenskyy sought to bring Ukraine into NATO, which angered Russian President Putin. Then in January 2022, in response, Russia sent troops to the Ukrainian border, which prompted international governments to speak out on the matter. Although oil and gasoline prices rose during 2021, in 2022, oil has risen 58% and retail gas has risen 24%. Because Russia has been consistently in the top three in global oil production it was feared that the supply might be disrupted. Therefore, demand increased as the pandemic faded, while at the same time Russia invaded Ukraine. Crude oil is the main input cost in the production of gasoline, and changes in crude oil price, along with changes in gasoline market conditions, drive changes in wholesale and retail gasoline prices.

ARIMA Model

ARIMA stand for Auto Regressive Integrated Moving Average and is actually a class of models that explains a given time series based on its own past values. It is specified with three parameters p, q and d. where p=number of autoregressive terms, d= number of non seasonal differences integrated and q= number of lagged forecast error in the equation. As a first step, we explore the ACF and PACF for the series. An autocorrelation function (ACF) and a partial autocorrelation function (PACF) for the time series variable. ACFs and PACFs will help us understand the temporal dynamics of an individual time series.

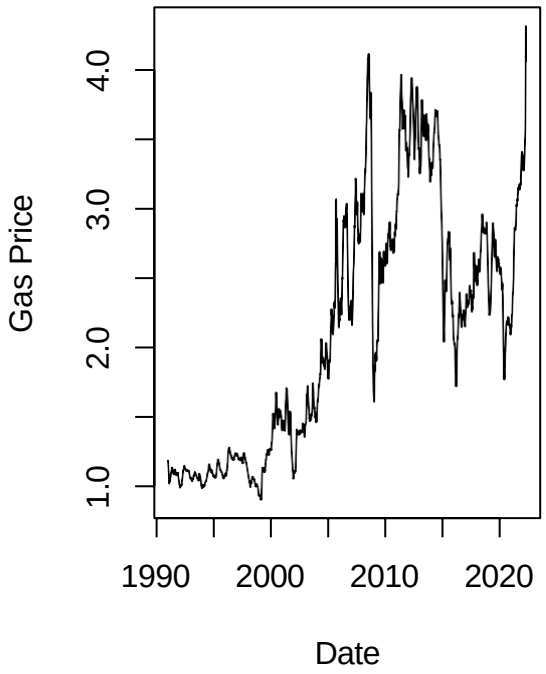
The following figure shows the general process for my forecasting:

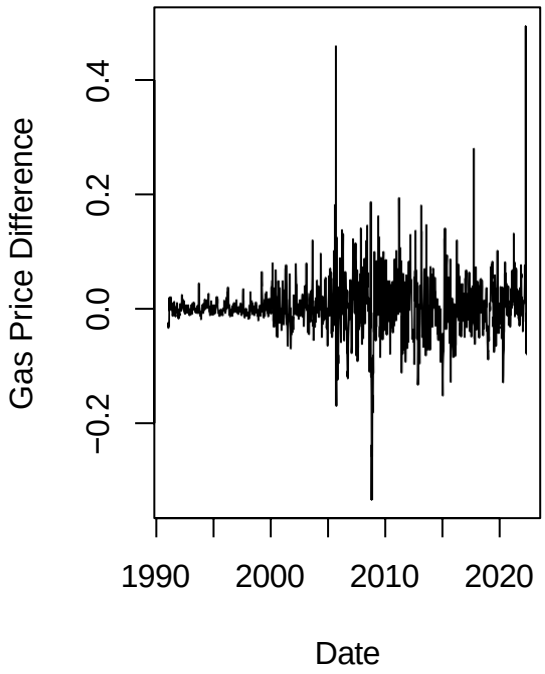
Modeling

I used weekly gas price in the U.S. from January, 1991 to April, 2022.As a first step, we explore the graph of the gasoline price.

Gas Price & Gas Price Difference Plot

```
# Plot the data
par(mfrow=c(1,2))
plot(data$Date,data$Price, type="l", ylab="Gas Price", xlab = "Date")
plot(data$Date[-1],diff(data$Price), type="l", ylab="Gas Price Difference", xlab = "Date")
```






```
adf.test(data$Price)
```

```
##
##   Augmented Dickey-Fuller Test
##
## data:   data$Price
## Dickey-Fuller = -3.2362, Lag order = 11, p-value = 0.08168
## alternative hypothesis: stationary
```

```
adf.test(diff(data$Price))
```

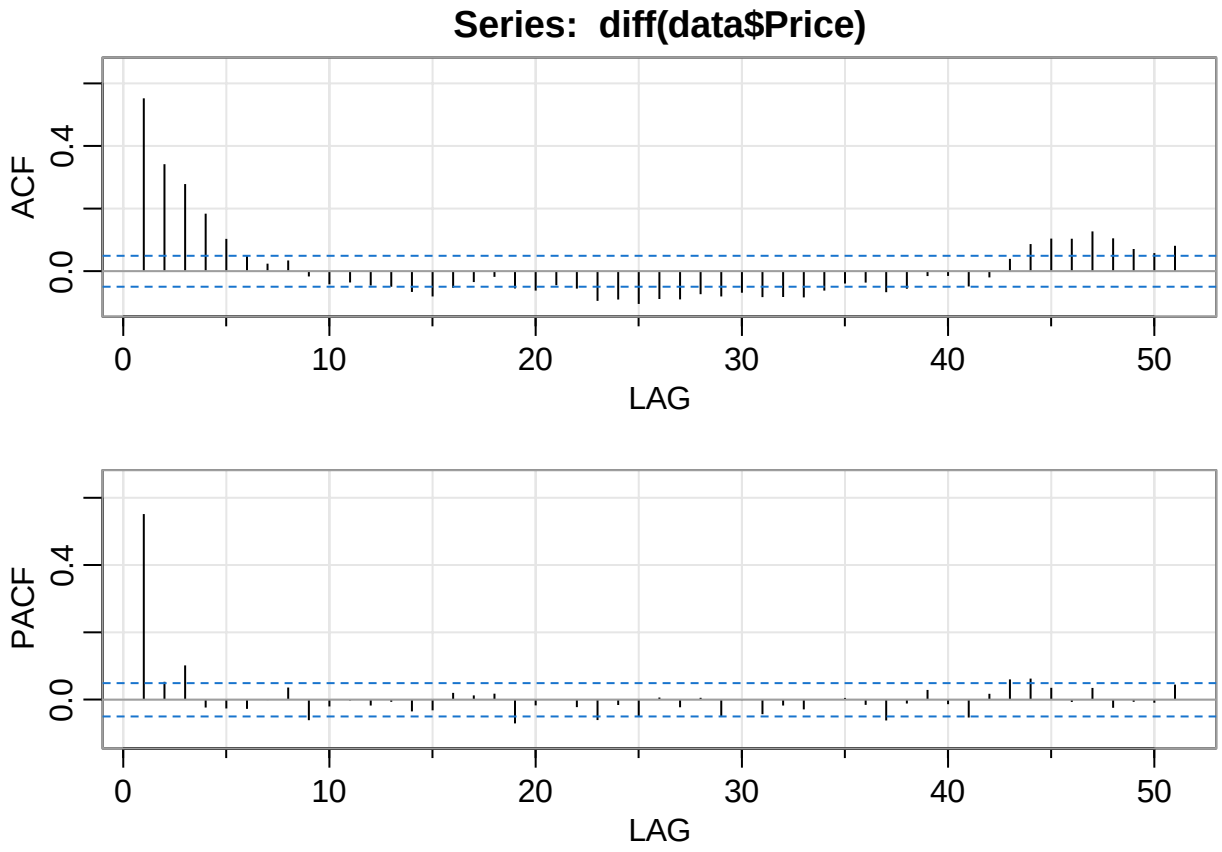
```
## Warning in adf.test(diff(data$Price)): p-value smaller than printed p-value
```

```
##
##   Augmented Dickey-Fuller Test
##
## data:   diff(data$Price)
## Dickey-Fuller = -10.987, Lag order = 11, p-value = 0.01
## alternative hypothesis: stationary
```

When we looked at the gas price graph, it did not look stationary because there is an upward trend. we also ran the ADF (Augmented Dickey-Fuller) test and the p-value was 0.082, indicating that it is not a stationary process. For that, I took the difference in gas prices. When I ran the ADF test of difference in gas prices, the p-value was less than 0.05, indicating a stationary process

Model Selction

```
acf2(diff(data$Price))
```



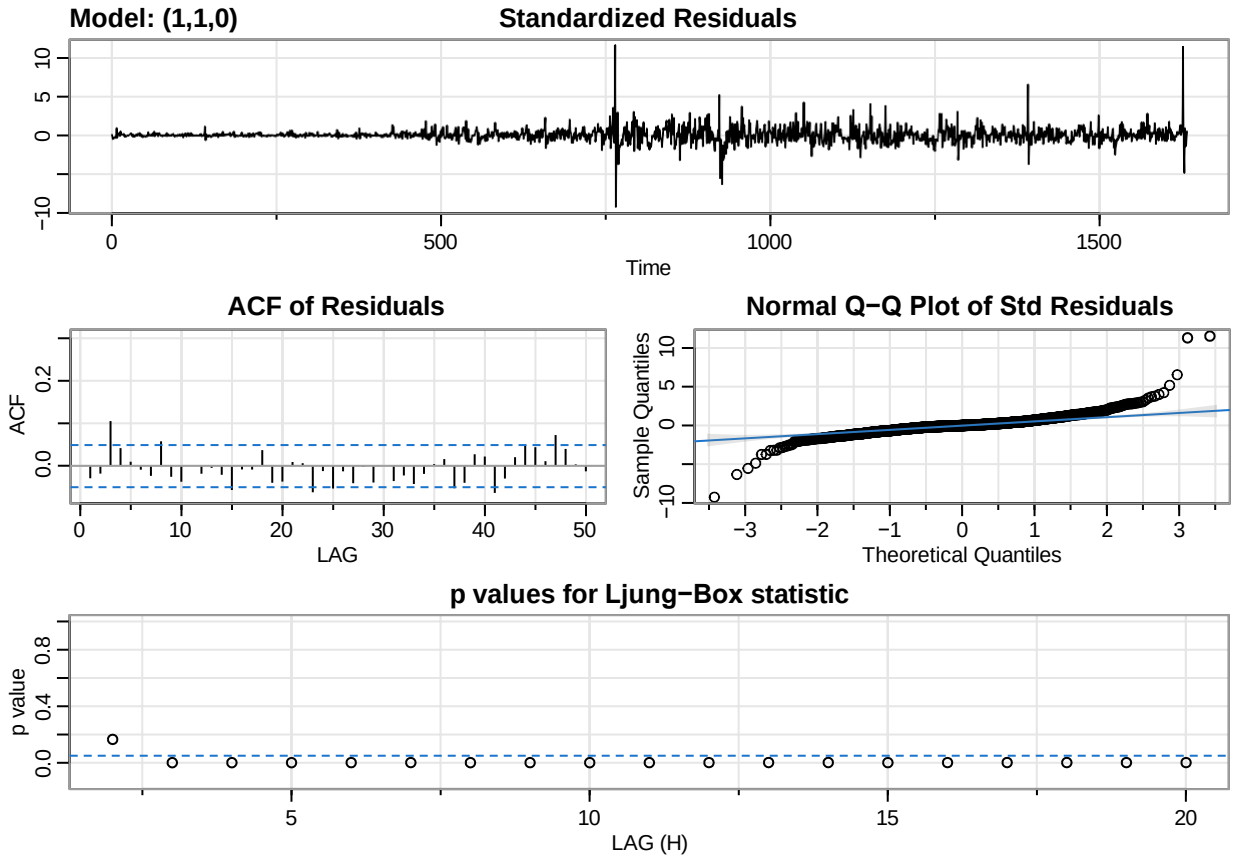

```
##       [,1] [,2] [,3]  [,4]  [,5]  [,6] [,7] [,8]  [,9] [,10] [,11] [,12] [,13]
## ACF   0.55 0.34 0.28  0.18  0.10  0.05 0.02 0.03 -0.02 -0.04 -0.04 -0.05 -0.05
## PACF  0.55 0.05 0.10 -0.02 -0.03 -0.03 0.00 0.04 -0.06 -0.02  0.00 -0.02 -0.01
##       [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## ACF  -0.07 -0.08 -0.05 -0.04 -0.02 -0.06 -0.06 -0.05 -0.06 -0.10 -0.09 -0.10
## PACF -0.04 -0.03  0.02  0.01  0.02 -0.07 -0.02  0.00 -0.02 -0.06 -0.02 -0.05
##       [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
## ACF  -0.09 -0.09 -0.07 -0.08 -0.07 -0.08 -0.08 -0.08 -0.06 -0.04 -0.04 -0.07
## PACF  0.01 -0.02  0.01 -0.05  0.00 -0.04 -0.02 -0.03  0.00  0.00 -0.02 -0.06
##       [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49]
## ACF  -0.06 -0.02 -0.02 -0.05 -0.02  0.04  0.09  0.10  0.10  0.13  0.10  0.07
## PACF -0.01  0.03 -0.01 -0.05  0.02  0.06  0.06  0.03 -0.01  0.03 -0.02 -0.01
##       [,50] [,51]
## ACF   0.06  0.08
## PACF -0.01  0.04
```

The ACF is tailing off rapidly and PACF cut off after the first lag shown above are typical of a series that has autocorrelation at the first lag only. This is often called an autoregressive process at one lag, or simply an AR(1) process. This is illustrated by the fairly quick decay toward zero in the ACF and the single spike at the first lag followed by small apparently random values after the first lag for the PACF.

AR(1) model

```
sarima(data$Price, 1,1,0)
```

```
## initial  value -3.042088
## iter   2 value -3.223767
## iter   2 value -3.223767
## iter   2 value -3.223767
## final  value -3.223767
## converged
## initial  value -3.223873
## iter   1 value -3.223873
## final  value -3.223873
## converged
```




```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##       xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##           REPORT = 1, reltol = tol))
##
## Coefficients:
##             ar1    constant
```

```
##          0.5518       0.0018
## s.e.     0.0206       0.0022
##
## sigma^2 estimated as 0.001584: log likelihood = 2942.04,    aic = -5878.09
##
## $degrees_of_freedom
## [1] 1628
##
## $ttable
##              Estimate       SE  t.value  p.value
## ar1            0.5518   0.0206  26.7401   0.0000
## constant       0.0018   0.0022   0.8026   0.4223
##
## $AIC
## [1] -3.606188
##
## $AICc
## [1] -3.606183
##
## $BIC
## [1] -3.596256
```

The results from AR(1) model are not the best. Usually, in the basic ARIMA model, we need to provide the p,d, and q values which are essential. We use statistical techniques to generate these values by performing the difference to eliminate the non-stationarity and plotting ACF and PACF graphs. Instead, we will use Auto ARIMA. In Auto ARIMA, the model itself will generate the optimal p, d, and q values based on AIC which would be suitable for the data set to provide better forecasting.

Mymodel

```
auto.arima(data$Price,ic="aic",trace = TRUE)
```

```
##
##  Fitting models using approximations to speed things up...
##
## ARIMA(2,1,2) with drift         : -5884.761
## ARIMA(0,1,0) with drift         : -5282.092
## ARIMA(1,1,0) with drift         : -5871.657
## ARIMA(0,1,1) with drift         : -5737.499
##  ARIMA(0,1,0)                   : -5281.868
##  ARIMA(1,1,2) with drift        : -5883.658
##  ARIMA(2,1,1) with drift        : -5877.756
##  ARIMA(3,1,2) with drift        : -5890.487
##  ARIMA(3,1,1) with drift        : -5886.311
##  ARIMA(4,1,2) with drift        : Inf
##  ARIMA(3,1,3) with drift        : -5884.775
##  ARIMA(2,1,3) with drift        : -5887.294
##  ARIMA(4,1,1) with drift        : -5895.32
##  ARIMA(4,1,0) with drift        : -5885.681
##  ARIMA(5,1,1) with drift        : -5892.244
##  ARIMA(3,1,0) with drift        : -5887.715
##  ARIMA(5,1,0) with drift        : -5883.826
```

```
##   ARIMA(5,1,2) with drift         : -5895.161
##   ARIMA(4,1,1)                    : Inf
##
##   Now re-fitting the best model(s) without approximations...
##
##   ARIMA(4,1,1) with drift         : -5904.693
##
##   Best model: ARIMA(4,1,1) with drift

## Series: data$Price
## ARIMA(4,1,1) with drift
##
## Coefficients:
##             ar1      ar2     ar3      ar4      ma1    drift
##          1.4783  -0.4952  0.1149  -0.123  -0.9668  0.0015
## s.e.     0.0313   0.0452  0.0441   0.025   0.0203  0.0013
##
## sigma^2 = 0.001556:  log likelihood = 2959.35
## AIC=-5904.69   AICc=-5904.62   BIC=-5866.92
```

From the analyses above, we can see that the best model is ARIMA(4,1,1) with drift. we will use that model for forecasting.

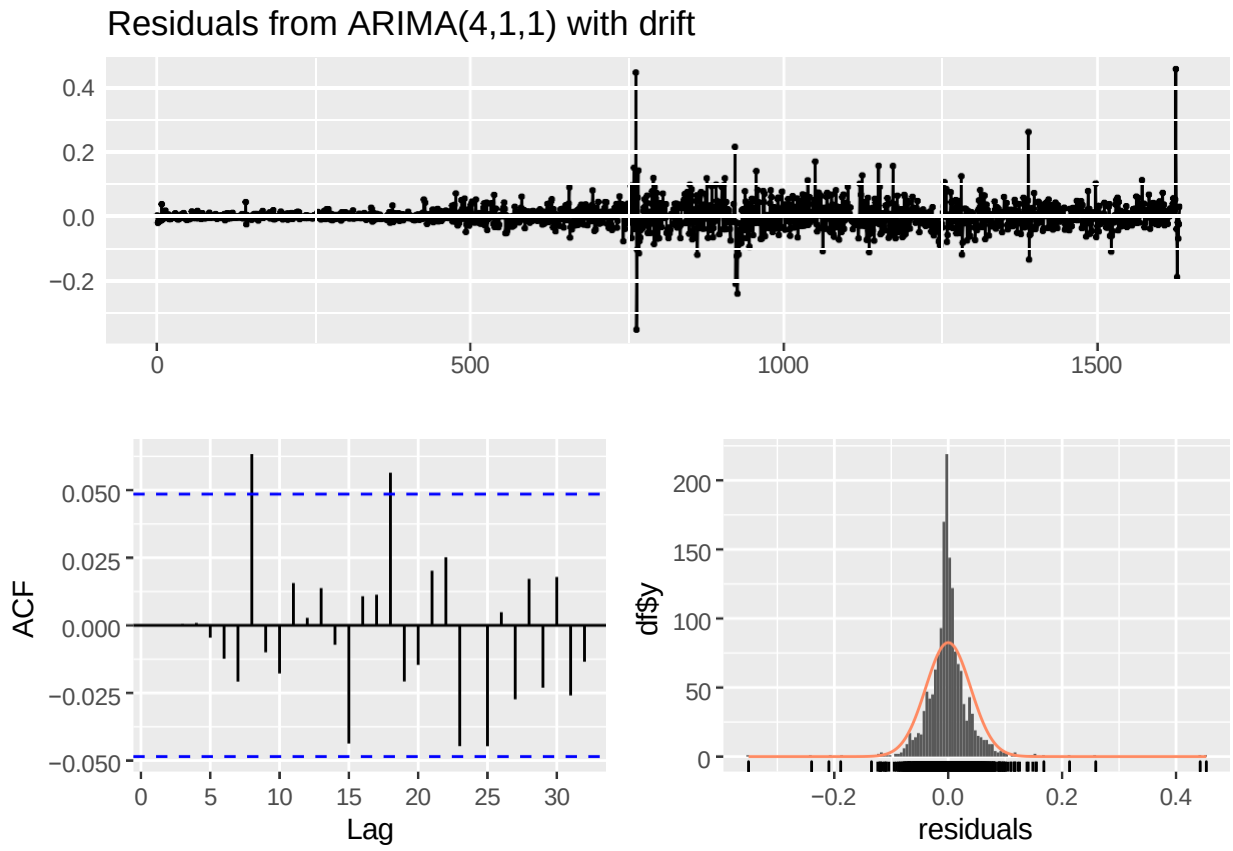
arima(4,1,1) Model Forecast Test

```
model<-auto.arima(data$Price,ic="aic",trace = TRUE)
```

```
##
##  Fitting models using approximations to speed things up...
##
## ARIMA(2,1,2) with drift         :  -5884.761
## ARIMA(0,1,0) with drift         :  -5282.092
## ARIMA(1,1,0) with drift         :  -5871.657
## ARIMA(0,1,1) with drift         :  -5737.499
## ARIMA(0,1,0)                    :  -5281.868
## ARIMA(1,1,2) with drift         :  -5883.658
## ARIMA(2,1,1) with drift         :  -5877.756
## ARIMA(3,1,2) with drift         :  -5890.487
##  ARIMA(3,1,1) with drift        :  -5886.311
##  ARIMA(4,1,2) with drift        :  Inf
##   ARIMA(3,1,3) with drift       : -5884.775
##   ARIMA(2,1,3) with drift       : -5887.294
##   ARIMA(4,1,1) with drift       : -5895.32
##   ARIMA(4,1,0) with drift       : -5885.681
##   ARIMA(5,1,1) with drift       : -5892.244
##   ARIMA(3,1,0) with drift       : -5887.715
##   ARIMA(5,1,0) with drift       : -5883.826
##   ARIMA(5,1,2) with drift       : -5895.161
##   ARIMA(4,1,1)                  : Inf
##
##   Now re-fitting the best model(s) without approximations...
```

```
##
##   ARIMA(4,1,1) with drift         : -5904.693
##
##   Best model: ARIMA(4,1,1) with drift
```

```
checkresiduals(model)
```

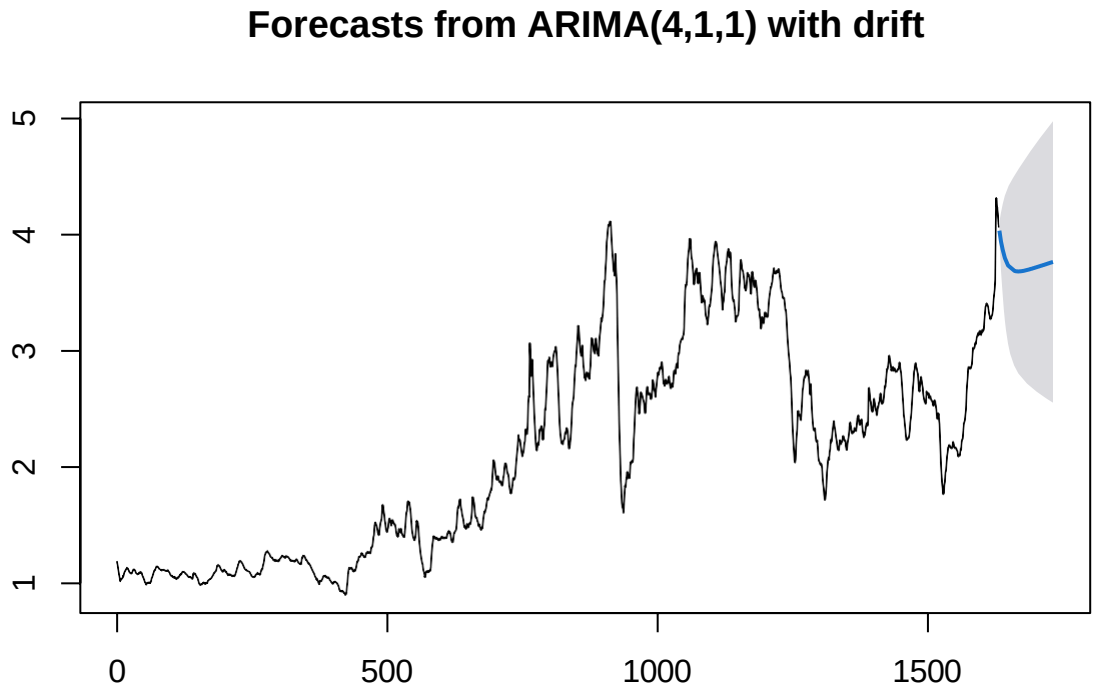



```
##
##   Ljung-Box test
##
## data:   Residuals from ARIMA(4,1,1) with drift
## Q* = 8.2598, df = 4, p-value = 0.08251
##
## Model df: 6.      Total lags used: 10
```

The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, apart from certain outliers, and therefore the residual variance can be treated as constant. The mean of the residuals is close to zero and there is no significant correlation in the residuals series. The ACF plot of the residuals from the ARIMA(4,1,1) model shows that most autocorrelations are within the threshold limits, indicating that the residuals are behaving like white noise. This can also be seen on the histogram of the residuals. The histogram suggests that the residuals may be normal.

arima(4,1,1) Model Forecast Test

```
Mforecast<-forecast(model,  level=c(.95),h=100)
plot(Mforecast)
```



We used data and the selected model to predict the gas price. The plot suggested, our forecast was not far from the actual gas price. Our model predicted a decrease in gas prices, and the gas price dropped in reality.

Conclusion

In this project, I looked at certain scenarios as of why the gasoline price was high, and then used time series analysis techniques to find the 'best' model to forecast the price change of the gas. My model succeeds to predict a significant drop in the gas price. In the future, it would improve our forecast significantly if we could find ways to include other factors that may influence the gas price change.