ECON_104_Project_2

Marc Luzuriaga, Takuya Sugahara, Daniel Day, Shabib Alam

2023-11-17

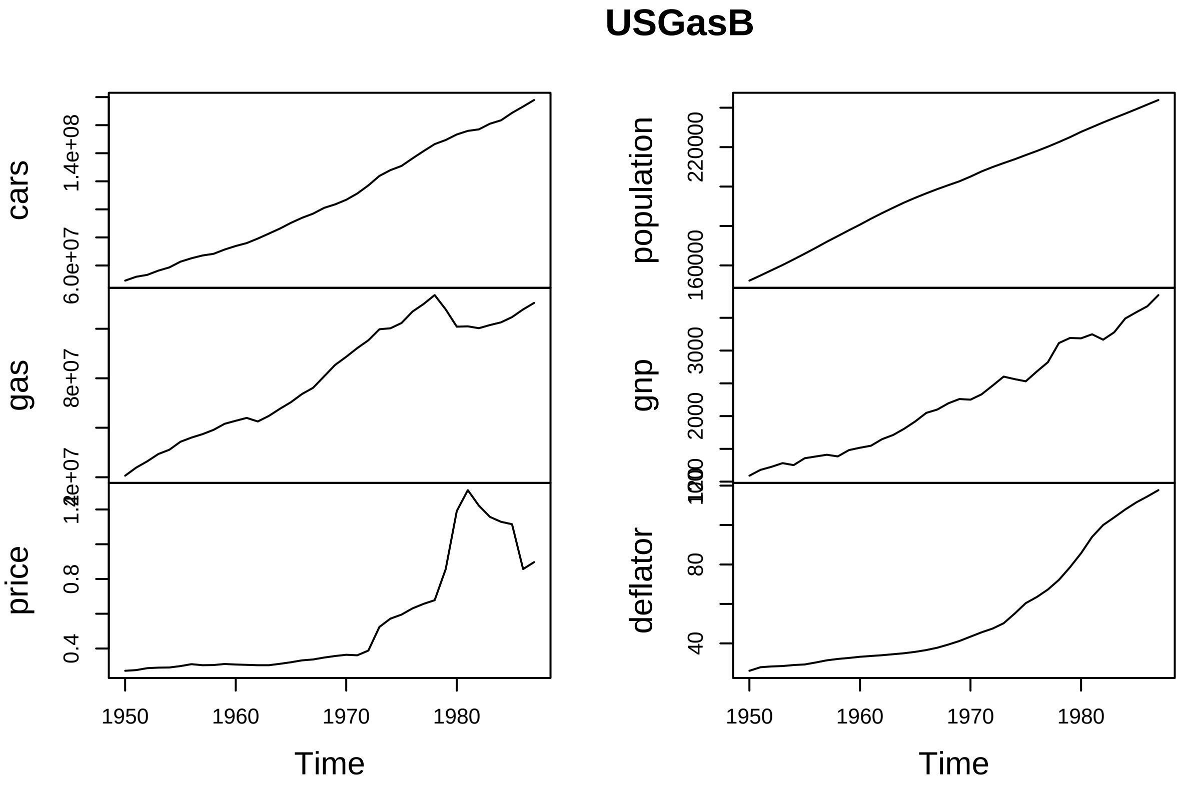
```
data("USGasB")
```

1 Introduction: USGasB

1(a) Question

Let us plot the relationship between time and the variables in our data set below:

```
plot(USGasB)
```



For the purposes of this project, we will ask "What is the effect of lagged values of price and lagged values of gas consumption on the last observation (t=38)?"

We can express our question more precisely using an ARDL model. We proceed to define the model as follows:

Suppose $t = 38$ represents the year 1987, y_t = price of gasoline at time period 38 and $x_{t,gasoline}$ = gasoline at time period 38. We also define $P \in 1, 2, 3, ..., l$ and $Q \in 1, 2, 3, ..., n$, where l = maximum lag periods for auto regressive model n = maximum lag periods for distributed finite model. Suppose $q \in Q$. Thus, we define our ARDL model as follows:

$y_t = \gamma + \sum_{p \in P} \alpha_p y_{t-p} + \sum_{q \in Q} \beta_q x_{t,gasoline-q}$

The goal of this project is to find the correct estimates and regression to the model above.

1(b) Data Set Summary

In this paper, we will be analyzing the USGasB Data Set from the AER Package. The data set holds 38 observations of annual time-series data from 1950 to 1987 on the US Gasoline Market. The dataset uses time as an explanatory variable and contains the following 6 dependent variables:

(1) cars: Stock of cars.

(2) gas: Consumption of motor gasoline (in 1000 gallons).

(3) price: Retail price of motor gasoline.

(4) population: Population.

(5) gnp: Real gross national product (in 1982 dollars).

(6) deflator: GNP deflator (1982 = 100).

1(c) Checking For Completeness and Consistency

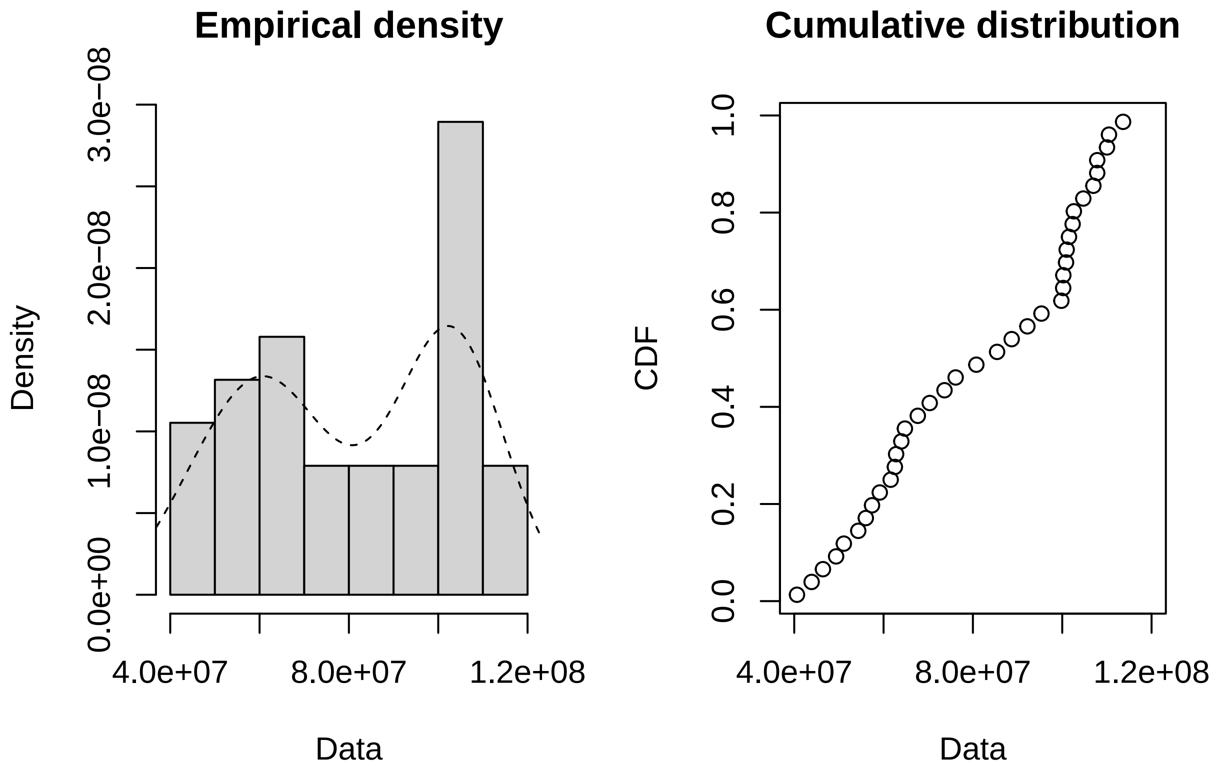
By observing the table of dependent variables, we observe that the data is complete and consistent for all the variables throughout the 38 time periods.

1(d) Descriptive Analysis of Variables

1(d).1 Graphs

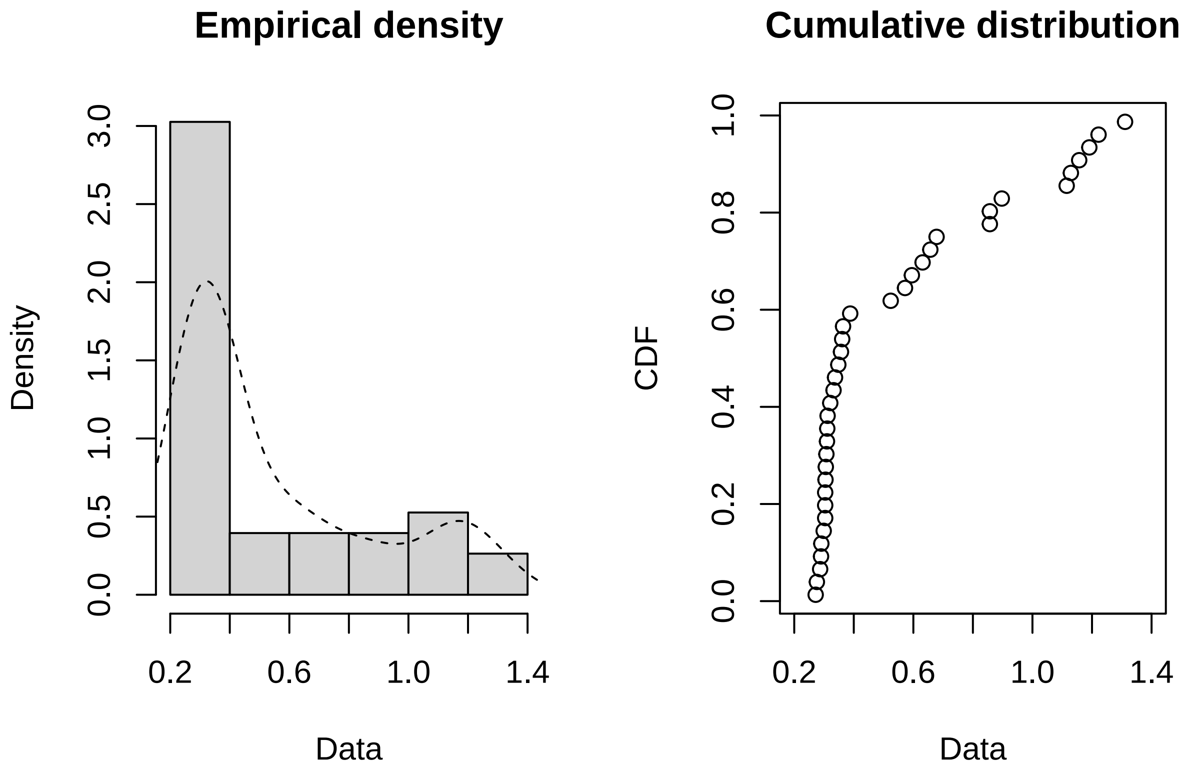
The graph below illustrates a histogram displaying the frequencies with respect to Gas Consumption. The histogram portrays the fact that the majority of our observations will fall between 1.0e+08 and 1.1e+08 gallons of consumption. The data also appears to be bi-modal distributed (Having two peaks)

```r
#Histogram of Gas
df_x <- as.data.frame(USGasB)
plotdist(df_x$gas, histo = TRUE, demp = TRUE)
```



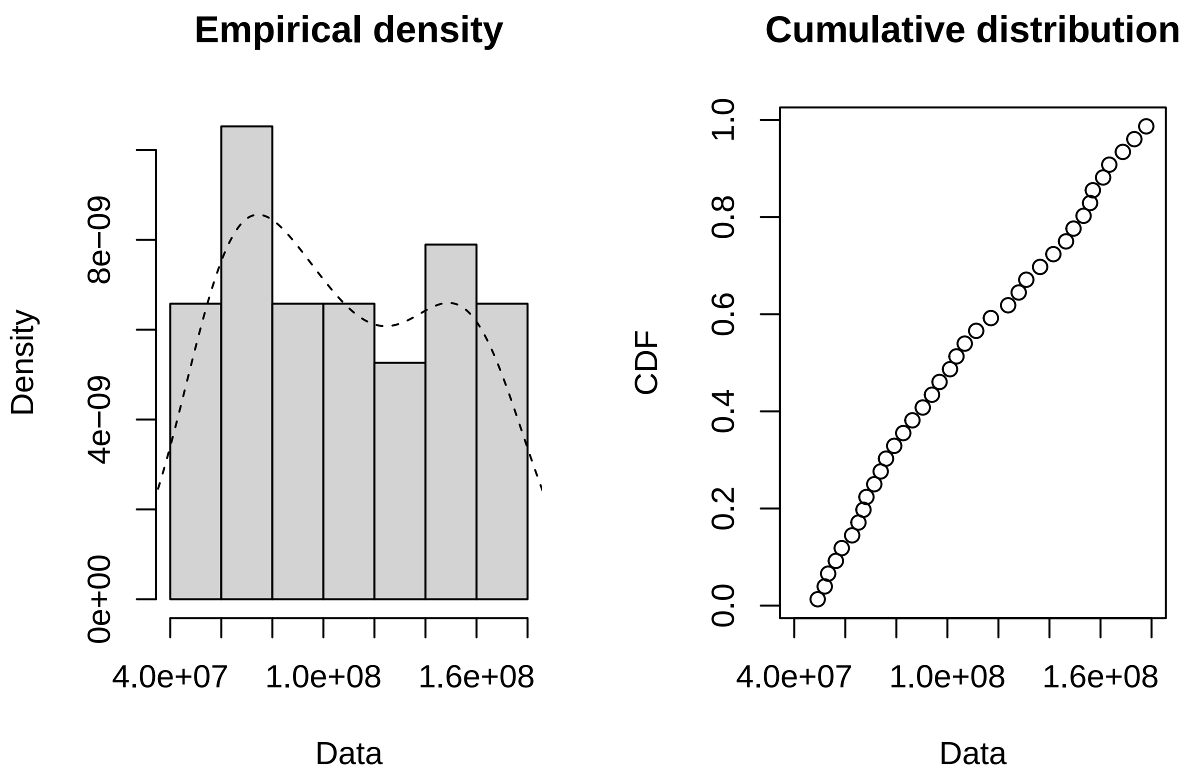
Next, we observe the histogram displaying the frequencies with respect to Gas Price. The histogram portrays the fact that the majority of our observations will fall between 0.2 and 0.4 dollars per gallon. The histogram is right-skewed with a long tail towards the right.

```r
plotdist(df_x$price, histo = TRUE, demp = TRUE)
```

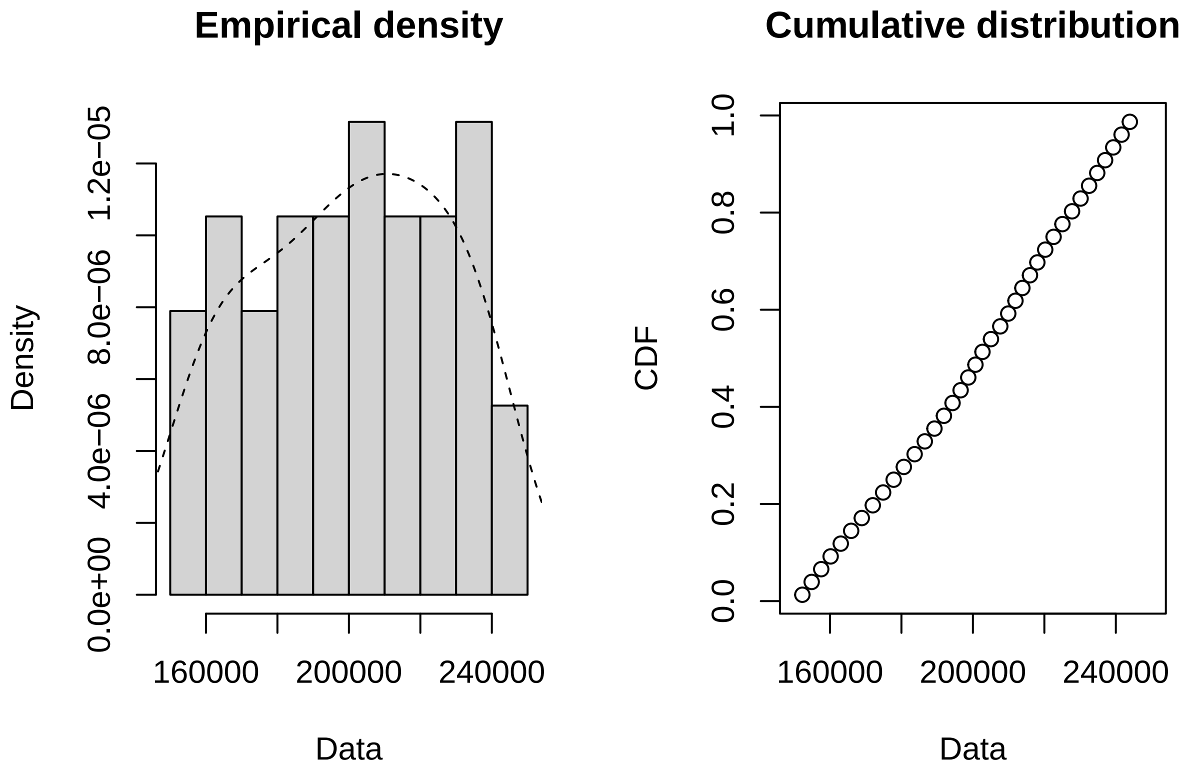


The other variables' histograms are presented below:

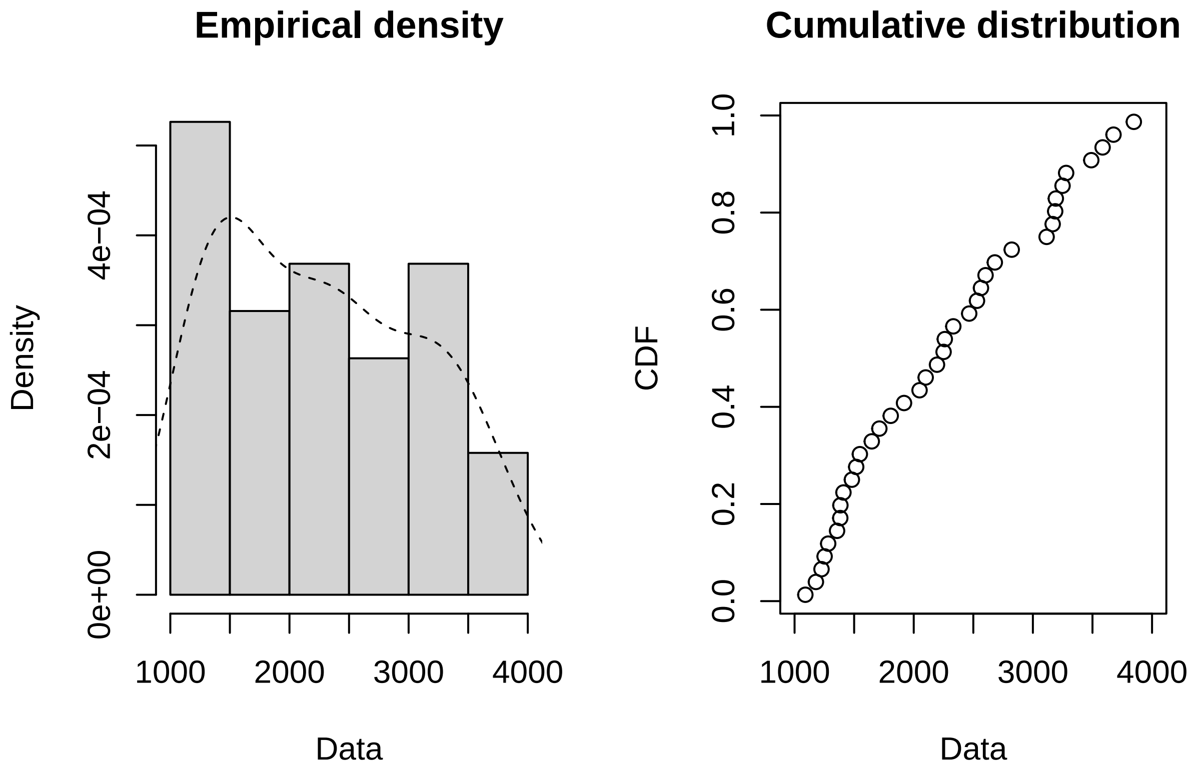
```
plotdist(df_x$cars, histo = TRUE, demp = TRUE)
```



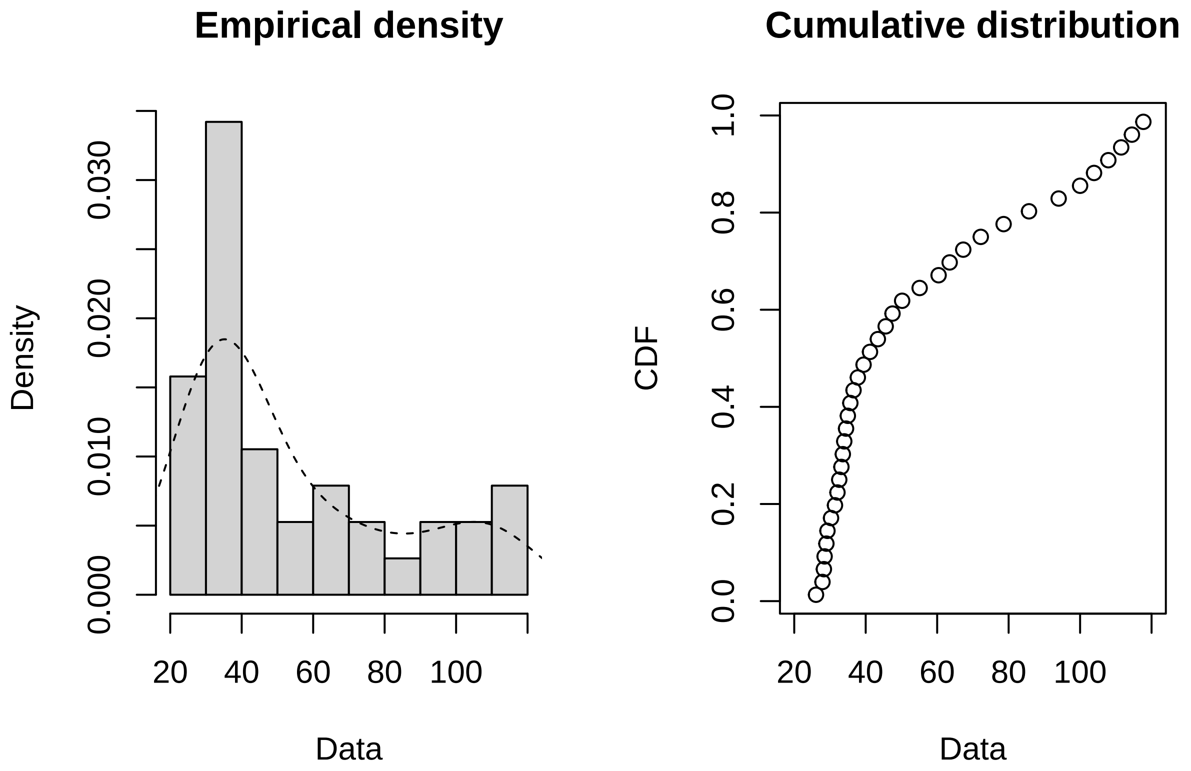
```r
plotdist(df_x$population, histo = TRUE, demp = TRUE)
```



```
plotdist(df_x$gnp, histo = TRUE, demp = TRUE)
```

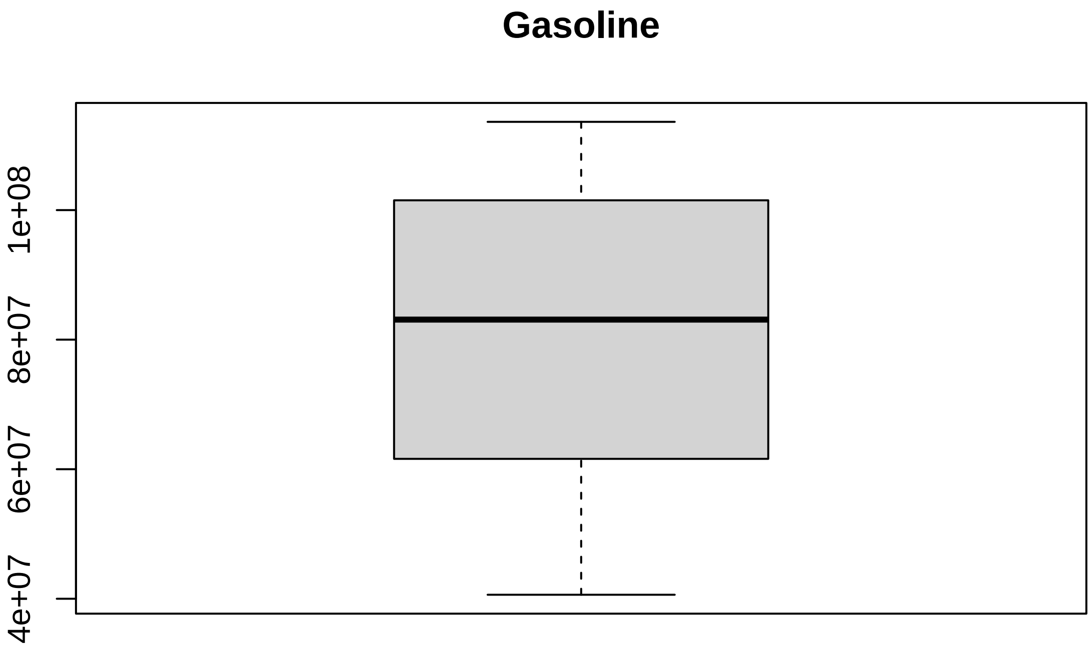


```
plotdist(df_x$deflator, histo = TRUE, demp = TRUE)
```



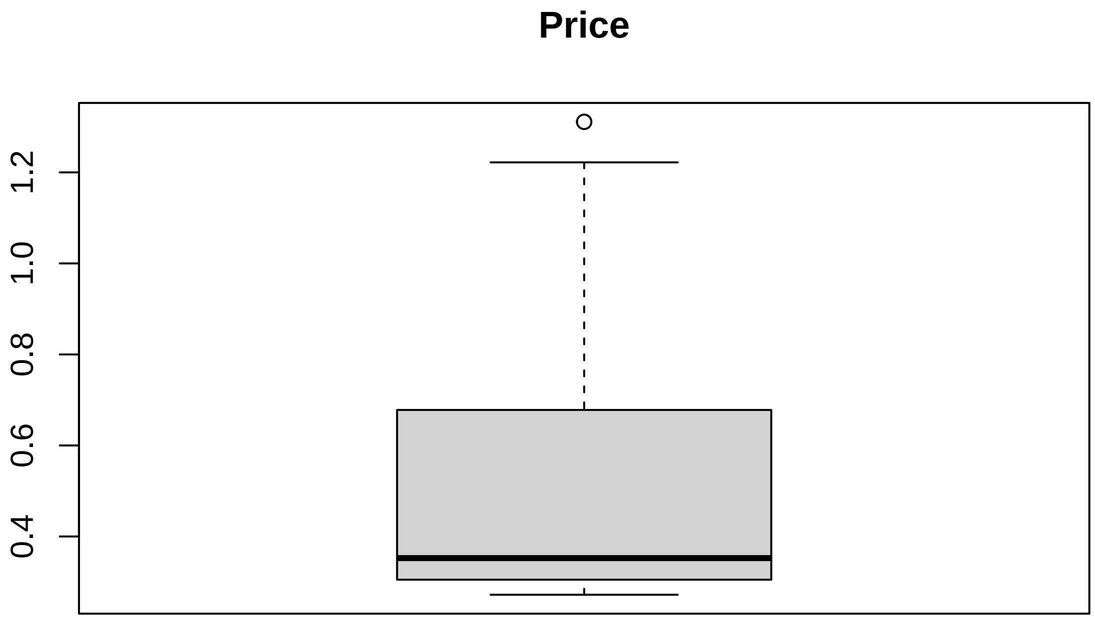
We present a Box Plot for the gasoline variable below. An interesting insight that the box plot provides us is that the data set's median is approximately 8e+07 gallons of consumption.

```r
#Box Plot For MurderRates
boxplot(df_x$gas, main="Gasoline")
```



We also present a Box Plot for the price variable below. An interesting insight that the box plot provides us is that a single outlier exists at price 1.311.

```r
#Box Plot For MurderRates
boxplot(df_x$price, main="Price")
```



Finally, we present the correlation matrix. The correlation between the price and the consumption of gas is 0.74328, a positive relationship. The correlation between the price and cars is 0.8756421. The correlation between the price and population is 0.8255576. The correlation between the price and gnp is 0.8708782. The correlation between the price and deflator is 0.9352551.

```
#Correlation Matrix
my_data <- df_x[, c(1,2,3,4,5,6)]
cor(my_data)
```

```
##                 cars       gas     price population       gnp  deflator
## cars       1.0000000 0.9576097 0.8756421  0.9875662 0.9956616 0.9414115
## gas        0.9576097 1.0000000 0.7432761  0.9620880 0.9438944 0.8134937
## price      0.8756421 0.7432761 1.0000000  0.8255576 0.8708782 0.9352551
## population 0.9875662 0.9620880 0.8255576  1.0000000 0.9813013 0.8984903
## gnp        0.9956616 0.9438944 0.8708782  0.9813013 1.0000000 0.9465645
## deflator   0.9414115 0.8134937 0.9352551  0.8984903 0.9465645 1.0000000
```

2 Data PreProcessing

2(a) Testing for Stationarity

In section 2(a), we will be exploring whether the time series data set is stationary. First, we must create the individual time series objects for our data set by running the following executable cells below:

```r
head(USGasB) # For exploring the dataset
```

```
## Time Series:
## Start = 1950
## End = 1955
## Frequency = 1
##          cars      gas price population    gnp deflator
## 1950 49195212 40617285 0.272     152271 1090.4     26.1
## 1951 51948796 43896887 0.276     154878 1179.2     27.9
## 1952 53301329 46428148 0.287     157553 1226.1     28.3
## 1953 56313281 49374047 0.290     160184 1282.1     28.5
## 1954 58622547 51107135 0.291     163026 1252.1     29.0
## 1955 62688792 54333255 0.299     165931 1356.7     29.3
```

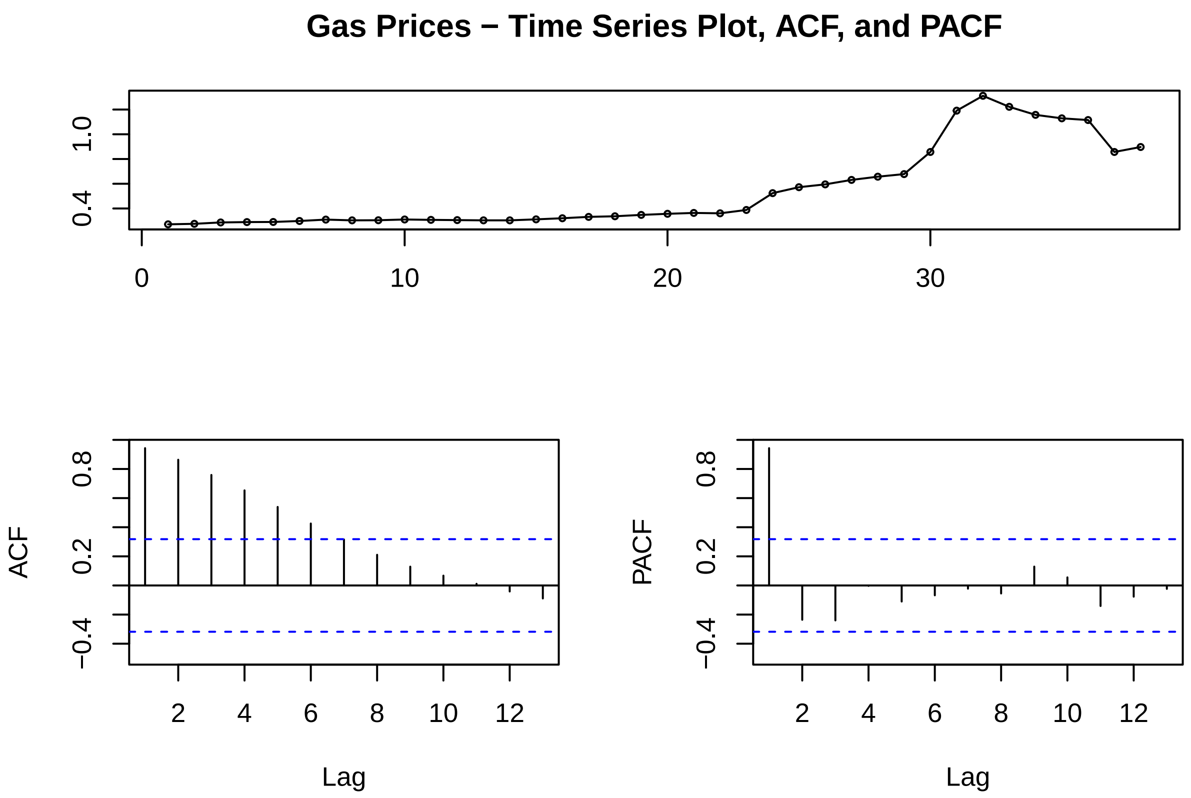
```r
summary(USGasB)
```

```
##       cars                 gas                 price          population
##  Min.   : 49195212   Min.   : 40617285   Min.   :0.2720   Min.   :152271
##  1st Qu.: 71982986   1st Qu.: 61830254   1st Qu.:0.3053   1st Qu.:178540
##  Median :102300566   Median : 83094370   Median :0.3525   Median :201692
##  Mean   :107634304   Mean   : 80901846   Mean   :0.5442   Mean   :200256
##  3rd Qu.:145244051   3rd Qu.:101384955   3rd Qu.:0.6727   3rd Qu.:221999
##  Max.   :177922000   Max.   :113625960   Max.   :1.3110   Max.   :243915
##       gnp          deflator
##  Min.   :1090   Min.   : 26.10
##  1st Qu.:1490   1st Qu.: 32.75
##  Median :2223   Median : 40.30
##  Mean   :2259   Mean   : 54.62
##  3rd Qu.:3042   3rd Qu.: 70.97
##  Max.   :3847   Max.   :117.70
```

```r
# Creating individual time series objects for specific variables
price_ts <- ts(USGasB[, "price"])
cars_ts <- ts(USGasB[, "cars"])
gas_ts <- ts(USGasB[, "gas"])
population_ts <- ts(USGasB[, "population"])
gnp_ts <- ts(USGasB[, "gnp"])
deflator_ts <- ts(USGasB[, "deflator"])
```

We then proceed to plot the time series plot for the “Price” variable:

```r
# Creating a time series display for the price where time series
# plot, ACF, and PACF will be displayed.
tsdisplay(price_ts, main = "Gas Prices - Time Series Plot, ACF, and PACF")
```



Upon visual inspection of the time series plot, it is evident that the data deviates from a consistent pattern after observation 20. This divergence indicates non-constant mean and variance over time, suggesting that the time series is non-stationary.

Formally, we will proceed to perform a unit root test, also known as an ADF test, to precisely check whether our time series is stationary. The test proposes a null hypothesis of non-stationary and an alternative hypothesis of stationary.

```r
# Conducting an  ADF test on the Price time series
adf_test_price <- adf.test(price_ts)
adf_test_price
```

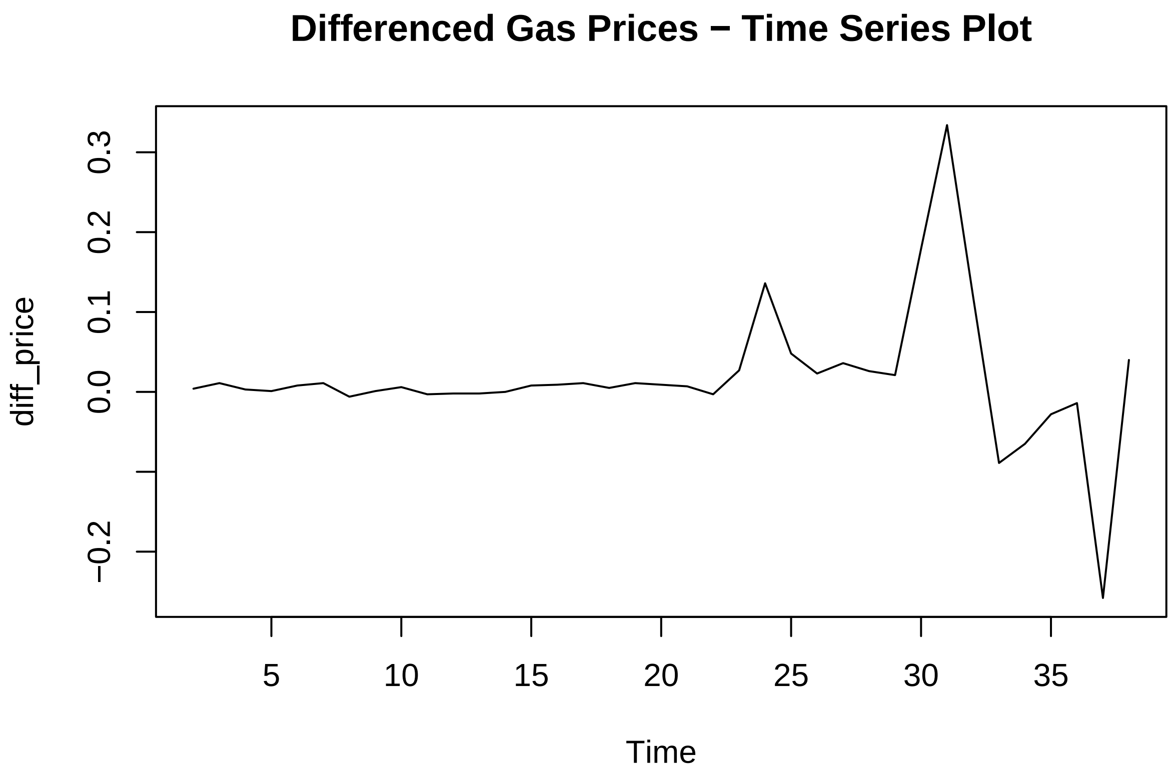
```
##
##  Augmented Dickey-Fuller Test
##
## data:  price_ts
## Dickey-Fuller = -2.1996, Lag order = 3, p-value = 0.4953
## alternative hypothesis: stationary
```

Since the p-value is greater than the common significance level of 0.05, we fail to reject the null hypothesis. Thus, the data does not provide enough evidence to conclude that the time series is stationary, confirming our intuition that the series was non-stationary from the visual plot.

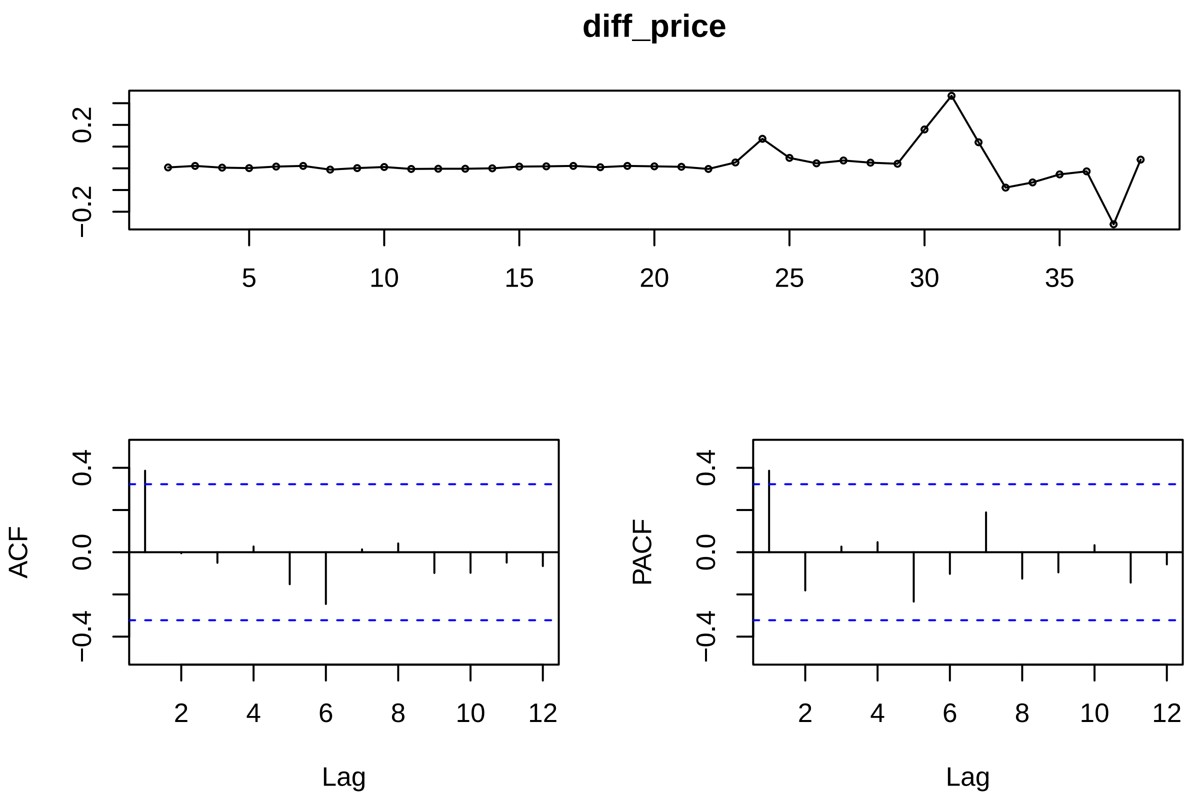
2(b) Correcting for Non-Stationary Data

In section 2(b), we will proceed to correct for the fact that our data showed evidence of non-stationary in section 2(a). The method we will use to correct for the non-stationary data is the method of differencing. We proceed to difference our time series once, plot our data, and run the unit root test again:

```
#Difference 1 Time
diff_order_price <- ndiffs(price_ts)
diff_price <- diff(price_ts, lag = 1,1)
plot(diff_price, main = "Differenced Gas Prices - Time Series Plot")
```



```
# Differenced time series, ACF and PACF
tsdisplay(diff_price)
```

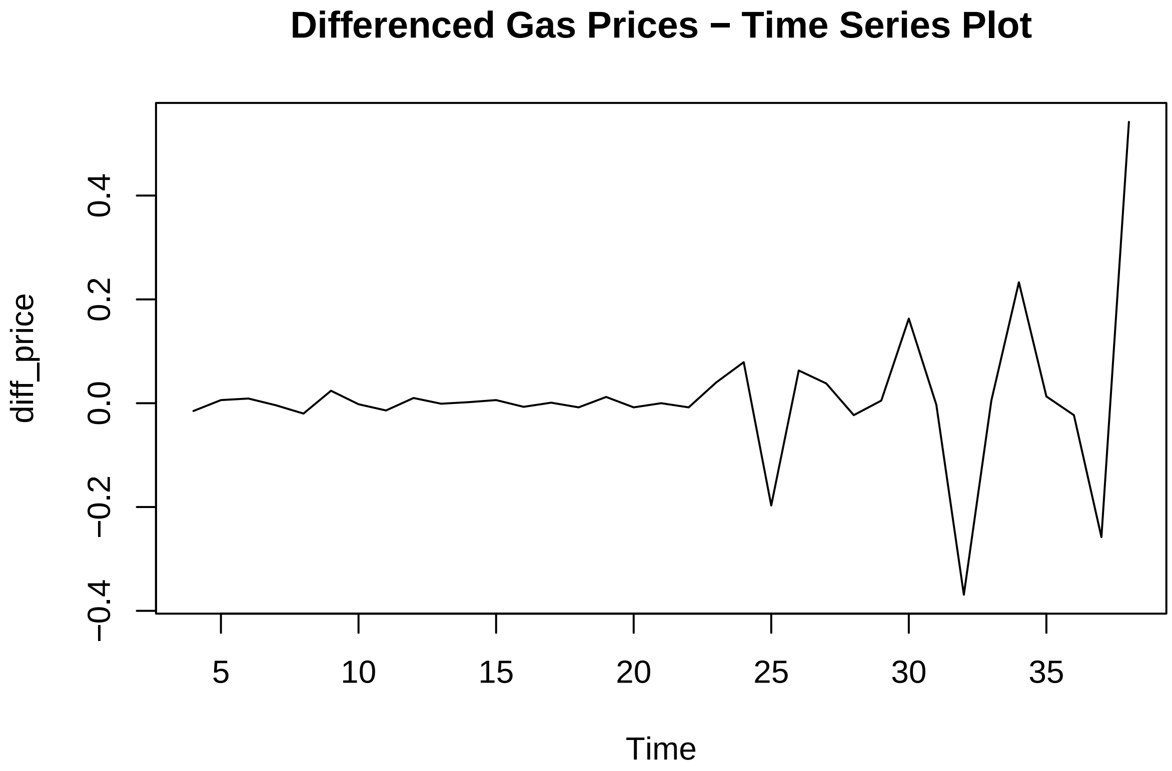


```r
# Re-doing DFT by adding lag to see that there has any stationary or not
adf_test_diff_price <- adf.test(diff_price)
adf_test_diff_price
```

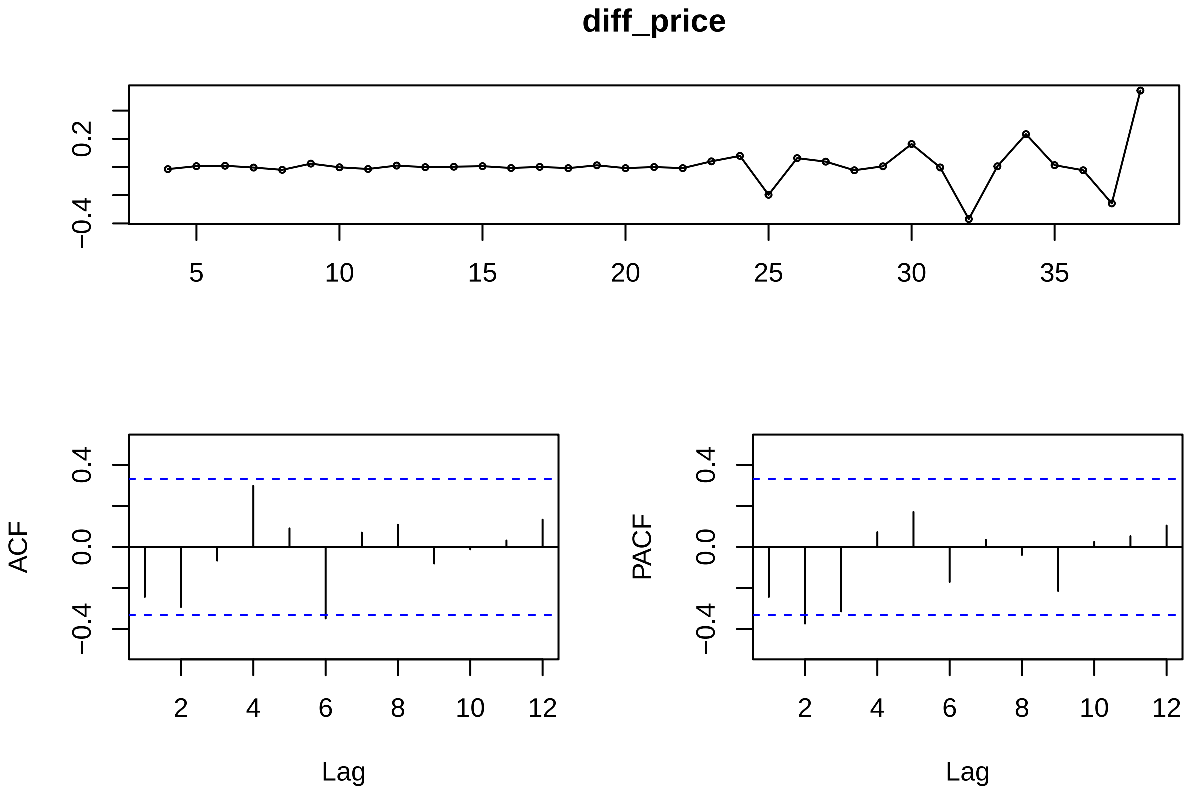
```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff_price
## Dickey-Fuller = -2.4369, Lag order = 3, p-value = 0.4028
## alternative hypothesis: stationary
```

Since the unit-root test suggests that we accept the null hypothesis after the first differencing, the test still suggest that our time-series is non-stationary. Thus, we proceeded to difference again. Yet, after the second difference, non-stationary persisted based on the unit root test. Thus, we needed to difference yet again. We proceeded to run the following executable to test whether our plot differenced three times suggested evidence of stationary using the unit root test:

```r
# Since it was not stationary, performing a unit-root test
#to determine the level of difference
diff_order_price <- ndiffs(price_ts)
diff_price <- diff(price_ts, lag = 1,3)
plot(diff_price, main = "Differenced Gas Prices - Time Series Plot")
```



```r
# Differenced time series, ACF and PACF
tsdisplay(diff_price)
```



```
# Re-doing DFT by adding lag to see that there has any stationary or not
adf_test_diff_price <- adf.test(diff_price)
adf_test_diff_price
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff_price
## Dickey-Fuller = -3.8433, Lag order = 3, p-value = 0.02897
## alternative hypothesis: stationary
```

By applying the third difference, the p-value of the unit root test is less than the common significance level of 0.05. Therefore, after differencing the time series 3 times, we can finally reject the null hypothesis, and the data provides enough evidence to conclude that the time series is stationary after 3 differences.

3 Future Generation, Model Testing and Forecasting

3(a) Autoregressive Lag Model

In this section, we will fit an AR(p) model to the data using AIC. We procede to determine AIC to assess which order p would fit our model best:

```r
library(dynlm)
```

```
## Warning: package 'dynlm' was built under R version 4.3.2
```

```r
Reg<- dynlm(diff_price ~ L(diff_price)) #AR(1)

Reg1<- dynlm(diff_price ~ L(diff_price, 1:2)) #AR2)
Reg2<- dynlm(diff_price ~ L(diff_price, 1:3)) #AR2)
Reg3<- dynlm(diff_price ~ L(diff_price, 1:6))
for(i in 1:14)
{Reg1<- dynlm(diff_price ~ L(diff_price, 1:i)) #AR}
print(AIC(Reg1))
}
```

```
## [1] -37.08874
## [1] -41.8477
## [1] -46.57812
## [1] -41.83213
## [1] -37.16001
## [1] -52.18145
## [1] -48.37966
## [1] -43.0902
## [1] -37.88404
## [1] -35.60565
## [1] -38.32668
## [1] -51.44783
## [1] -48.53827
## [1] -42.28961
```

Since we observe an AIC value of -52.18145 at order p = 6 lags, the lowest AIC vale suggests that we will be using an AR(6).

3(b) Testing Autocorrelation for AR(p)

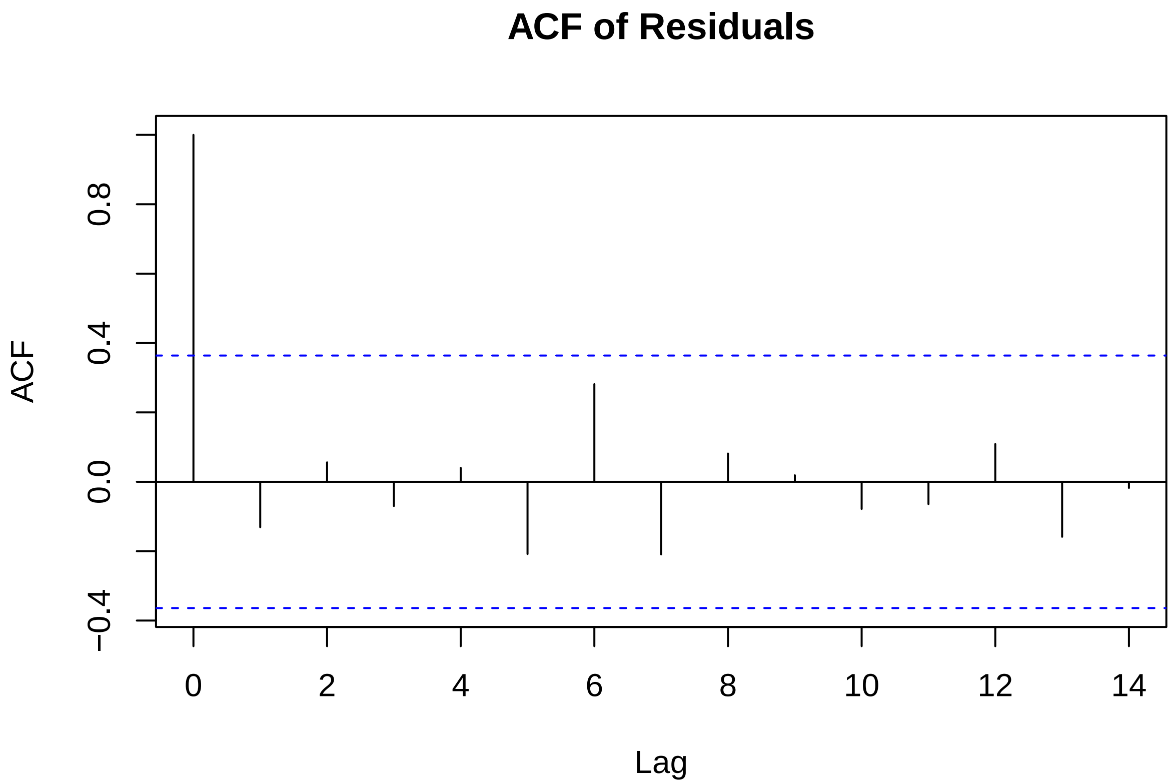
In section 3(b), we will test whether our AR(6) plot shows evidence of autocorrelation. Specifically, we will be using the Ljung-Box test as a formal method to test whether the autocorrelation of our time series are jointly different from zero. First, we plot the ACF of our residuals from the Reg3 Model obtained from the previous section.

```r
summary(Reg3)
```

```
##
## Time series regression with "ts" data:
## Start = 10, End = 38
##
## Call:
## dynlm(formula = diff_price ~ L(diff_price, 1:6))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.191347 -0.018948  0.004142  0.017593  0.248135
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.01011    0.01642  -0.616 0.544438
## L(diff_price, 1:6)1 -1.04841    0.21963  -4.774 9.13e-05 ***
## L(diff_price, 1:6)2 -1.25861    0.25859  -4.867 7.27e-05 ***
## L(diff_price, 1:6)3 -1.35851    0.34396  -3.950 0.000682 ***
## L(diff_price, 1:6)4 -1.23394    0.36446  -3.386 0.002661 **
## L(diff_price, 1:6)5 -0.68947    0.31482  -2.190 0.039413 *
## L(diff_price, 1:6)6 -1.33158    0.28781  -4.627 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08575 on 22 degrees of freedom
## Multiple R-squared:  0.7426, Adjusted R-squared:  0.6724
## F-statistic: 10.58 on 6 and 22 DF,  p-value: 1.495e-05
```

```r
Reg3<- dynlm(diff_price ~ L(diff_price, 1:6))
residuals <- residuals(Reg3)
acf(residuals, main = "ACF of Residuals")
```



As we can see, there is no subsequent spikes except zero lag model. So this would imply that the residuals are random.

We proceed to run the Box-Ljung test to test for autocorrelation. The null hypothesis of the Box-Ljung test states that no autocorrelation exists within our model, whereas the alternative hypothesis suggests that autocorrelation exists within our model.

```r
rersiduals <- residuals(Reg3)
Box.test(residuals, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  residuals
## X-squared = 0.55129, df = 1, p-value = 0.4578
```

Since we observed a p-value greater than the significance level (>0.05), this indicates that we fail to reject the null hypothesis, suggesting that there is no significant autocorrelation in the residuals. This is the desired outcome, implying that our model has captured the autocorrelation structure of the data adequately.

3(c) Autoregressive Finite Distributed Lag Model

In section 3(c), we will fit an ARDL(p,q) model to the data using AIC. We procede to determine AIC to assess which order p would fit our model best:

```r
diff_gas <- diff(gas_ts, lag = 1,3)
for (i in 1:6){
  Reg1<- dynlm(diff_price ~ L(diff_price, 1:i)+L(diff_gas, 1:i))
  print(AIC(Reg1))
}
```

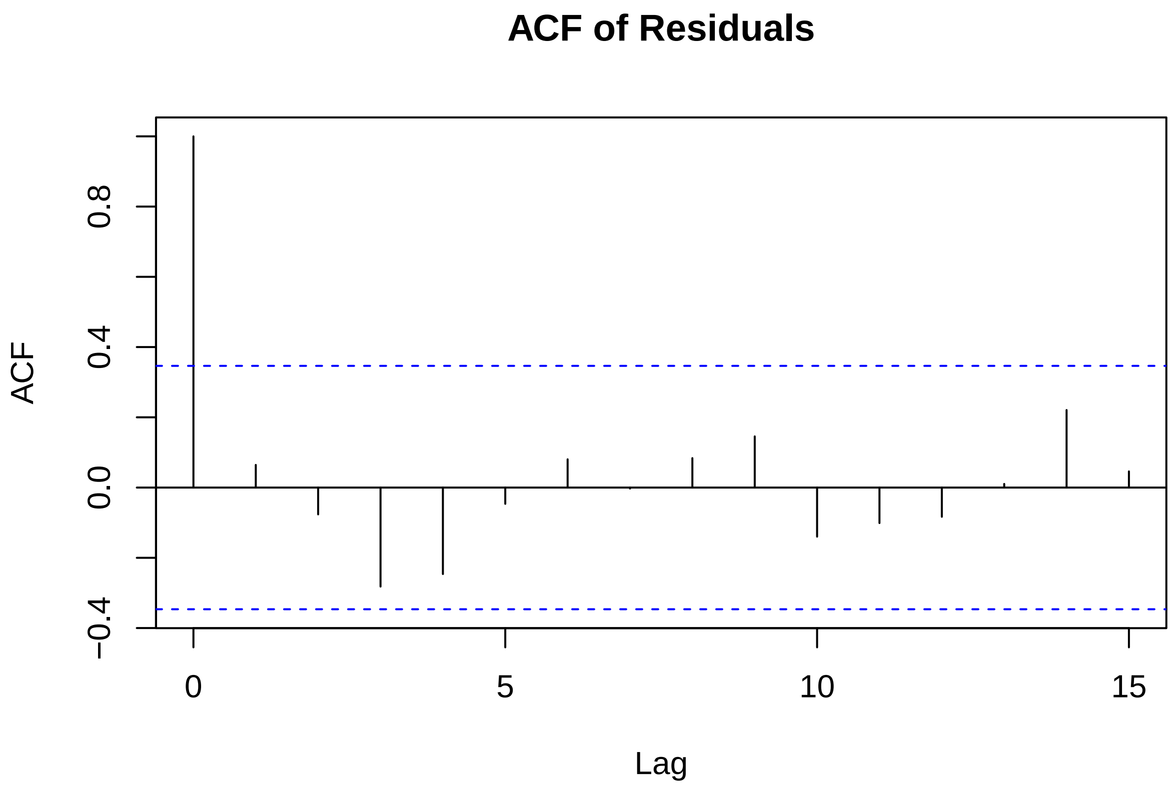
```
## [1] -40.22535
## [1] -42.32682
## [1] -53.86989
## [1] -51.17825
## [1] -44.92482
## [1] -53.83056
```

Based on the AIC value an ARDL(3,3) models seems to be best since it had the lowest AIC value.

3(d) Testing Autocorrelation for ARDL(p,q)

In this section, we will test whether our ARDL(3,3) shows evidence of autocorrelation.

```r
  Reg1<- dynlm(diff_price ~ L(diff_price, 1:3)+L(diff_gas, 1:3))
residuals <- residuals(Reg1)
acf(residuals, main = "ACF of Residuals")
```



Based on the ACF plot of the residuals there are no significant lags. Thus informally there is no serial correlation.

```
Resid1 <- residuals(Reg1)
Box.test(Resid1, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  Resid1
## X-squared = 0.14588, df = 1, p-value = 0.7025
```

Since we observed high p-value(>0.05) indicates that we fail to reject the null hypothesis, suggesting that there is no significant autocorrelation in the residuals. This is the desired outcome, implying that our model has captured the autocorrelation structure of the data adequately.

3(e) AR or ARDL?

In this section, we will determine whether we should use an AR or ARDL model

Both models do not have serial correlation. The AR(6) model has an AIC of -52.18145 while the AIC of the ARDL(3,3) has an AIC of -53.86989. Thus the lower AIC is the better model because it balances the prediction error and model fit.

4 Conclusion

4(a) Limitations

This section will explore the limitations found with our model.

The first potential limitation is Reverse Causality Bias. Specifically, changes in the dependent variable results in changes in one of the covariates. The problem with this is the fact that Reverse Causality Bias could lead to endogeneity in our model, where the errors and variables of our model are contemporaneously correlated. If endogenous, our model will result in biased and inconsistent estimates. Therefore, we will proceed to run the Granger Causality test for two models, Y~X and X~Y, at an alpha (critical level) of 5%.

The Null Hypothesis for the Granger Causality test is that Granger Causality does not exist, whereas the alternative Hypothesis for the Granger Causality test is that Granger Causality does exist.

First, we run the Granger Causality test for Y~X below:

```r
#Granger(Y~X)
grangertest(price_ts~gas_ts)
```

```
## Granger causality test
##
## Model 1: price_ts ~ Lags(price_ts, 1:1) + Lags(gas_ts, 1:1)
## Model 2: price_ts ~ Lags(price_ts, 1:1)
##   Res.Df Df      F Pr(>F)
## 1     34
## 2     35 -1 5.9694 0.0199 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Since the p-value of the F-statistic is less than the significance level of 5%, we would reject the null hypothesis that price does not Granger cause the consumption of gas. Therefore, the test implies that price Granger causes the consumption of gas.

Second, we run the Granger Causality test for X~Y below:

```r
#Granger(X~Y)
grangertest(gas_ts~price_ts)
```

```
## Granger causality test
##
## Model 1: gas_ts ~ Lags(gas_ts, 1:1) + Lags(price_ts, 1:1)
## Model 2: gas_ts ~ Lags(gas_ts, 1:1)
##   Res.Df Df      F Pr(>F)
## 1     34
## 2     35 -1 1.7537 0.1942
```

Since the p-value of the F-statistic is less than the significance level of 5%, we would fail to reject the null hypothesis that the consumption of gas does not Granger cause the price of gas. Therefore, the test implies that the consumption of gas does not Granger cause the price of gas.

The fact that our model exhibited Granger Causality for Y~X is a limitation we should further explore because it suggests reverse causality bias that leads to endogeneity. We would possibly need to include an instrumental variable that spits our endogenous variable into the part that it is correlated with the error term and the part that is uncorrelated with the error term in order to correct for this evidence of endogeneity.

Another limitation of our model is the potential for our time-series being a moving average.

Recall in section 2(a), we found that our time series was non-stationary through the unit root test. We proceeded to correct for the non-stationarity by the method of differencing. However, we needed to difference three times in order for the unit root test to indicate that the model was stationary. Although differencing is a way to correct for non-stationary, the trade off was that we lost one observation every difference that we performed. Since we had to difference three times, this also suggests that our series is a moving average and is a cause of concern.

4(b) Answering Question Presented in Beginning

To conclude this project, we will be answering the question we presented in the beginning of the paper: "What is the effect of lagged values of price and lagged values of gas consumption on the last observation (t=38)?"

We discussed in the beginning of this paper that we could express our question more precisely using an ARDL model. We proceeded to define the model as follows:

Suppose $t = 38$ represents the year 1987, y_t = price of gasoline at time period 38 and $x_{t,gasoline}$ = gasoline at time period 38. We also define $P \in 1, 2, 3, ..., l$ and $Q \in 1, 2, 3, ..., n$, where l = maximum lag periods for auto regressive model n = maximum lag periods for distributed finite model. Suppose $q \in Q$. Thus, we define our ARDL model as follows:

$$y_t = \gamma + \sum_{p \in P} \alpha_p y_{t-p} + \sum_{q \in Q} \beta_q x_{t,gasoline-q}$$

In section 3(c), we fitted an ARDL(p,q) model into our time-series using AIC values. Therefore, the coefficients of the lagged values of price and lagged values of gas on the last observation (t=38) are presented in the R markdown below:

```
Reg1<- dynlm(diff_price ~ L(diff_price, 1:3)+L(diff_gas, 1:3))
Reg1
```

```
##
## Time series regression with "ts" data:
## Start = 7, End = 38
##
## Call:
## dynlm(formula = diff_price ~ L(diff_price, 1:3) + L(diff_gas,
##     1:3))
##
## Coefficients:
##          (Intercept)  L(diff_price, 1:3)1  L(diff_price, 1:3)2
##            3.928e-03           -1.164e+00           -1.086e+00
## L(diff_price, 1:3)3    L(diff_gas, 1:3)1    L(diff_gas, 1:3)2
##           -9.019e-01           -2.054e-08           -1.917e-08
##   L(diff_gas, 1:3)3
##           -1.476e-08
```