Project 2 Econ 104

Sia Phulambrikar, Ahnaf Tamid, Sofia Giorgi, Michael Sorooshian

2023-11-17

Q1. Exploratory Analysis

(a) Briefly discuss the question you are trying to answer.

We are trying to answer how, from 1950 to 1987, the stock of cars and retail price of gasoline affect the consumption of gasoline.

(b) Cite the dataset and give a summary of what the dataset is about

The USGasB dataset can be found in the AER package, as shown in this document: https://cran.r-project.org/web/packages/AER/AER.pdf. It is a time-series dataset showing the stock of cars (cars), consumption of gasoline (gas), retail price of gasoline (price), population (population), real gross national product (gnp), and deflator (deflator), each year from 1950 to 1987.

(c) First check for completeness and consistency of the data (if there are NAs or missing observations, replace with the value of the previous observation; make a note of this)

```
sum(is.na(USGasB))
```

```
## [1] 0
```

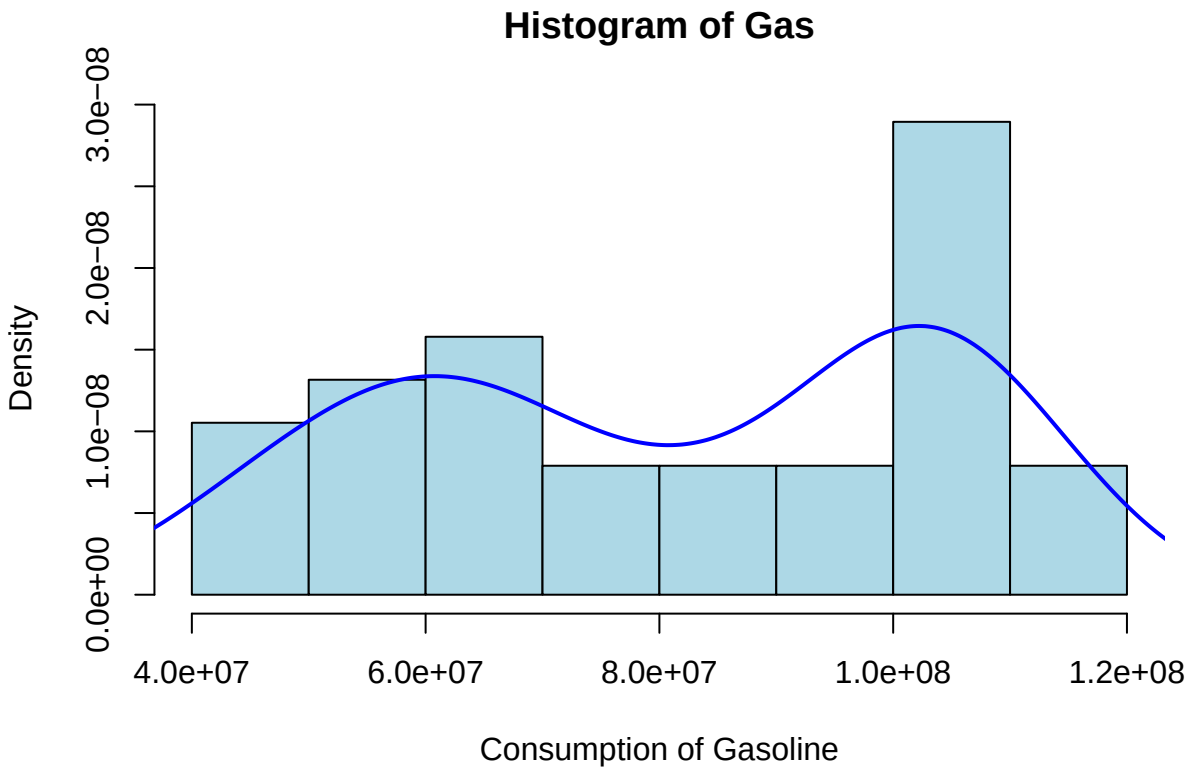
```
summary(USGasB)
```

```
##       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.:221998
##  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
```

There are no observations marked as NA, and the data presents as complete and consistent.

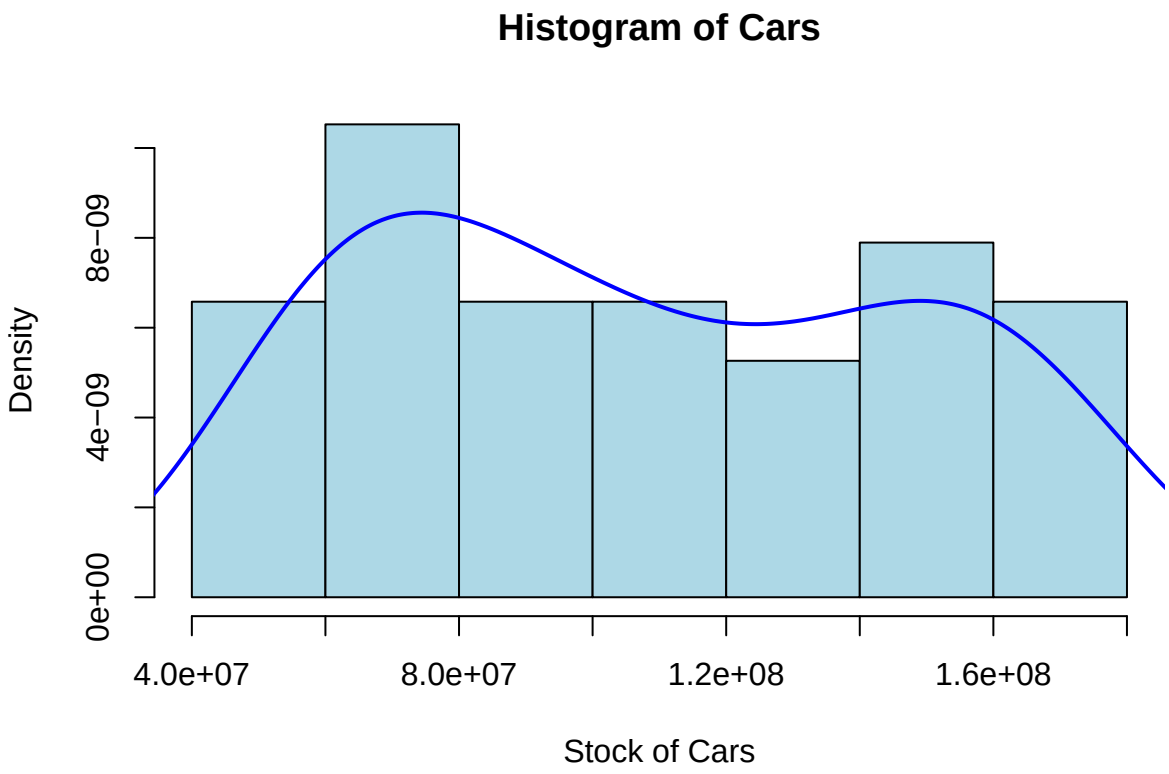
(d) Provide descriptive analyses of your variables. This should include the histogram with overlying density, boxplots, cross correlation. All figures/statistics must include comments.

```
hist(USGasB[,"gas"], prob = TRUE, col="lightblue", main = "Histogram of Gas",
     xlab = "Consumption of Gasoline")
lines(density(USGasB[,"gas"]), col = "blue", lwd=2)
```



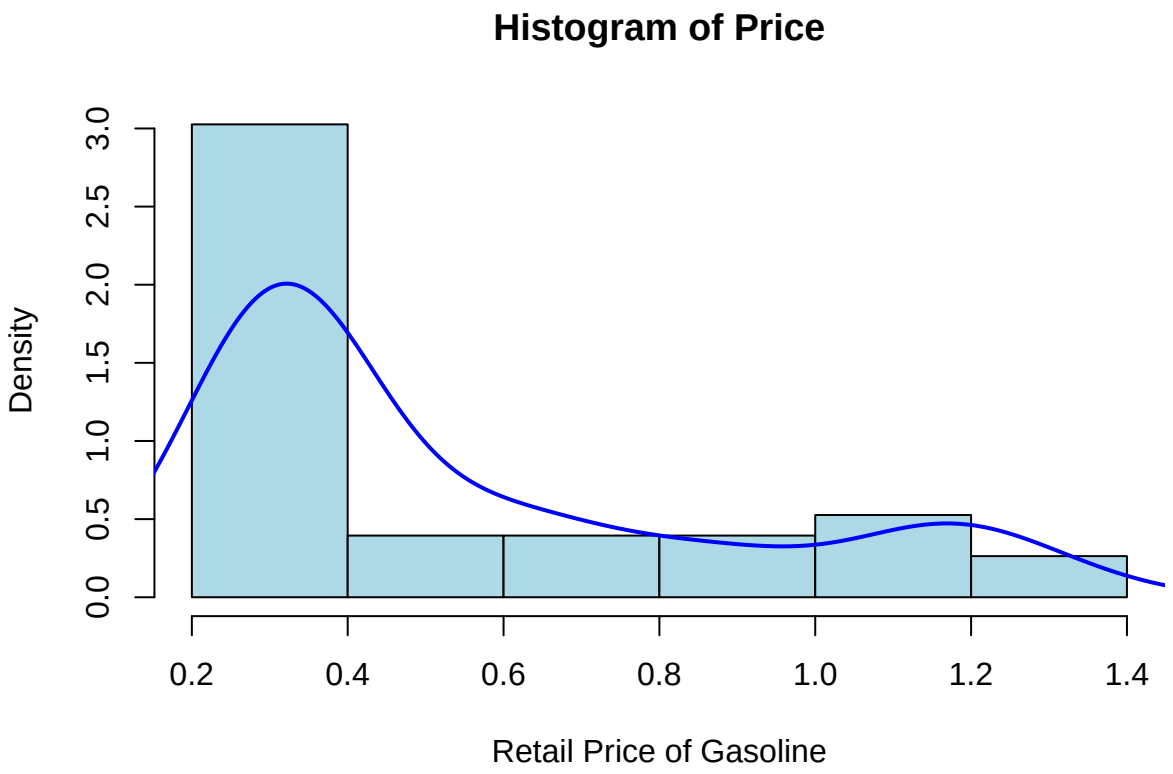
The consumption of gasoline forms a bimodal distribution. This indicates that there are roughly two different groups in the consumption of gasoline, one that possibly consumes less gas and one that consumes more.

```
hist(USGasB[,"cars"], prob = TRUE, col="lightblue", main = "Histogram of Cars",
     xlab = "Stock of Cars")
lines(density(USGasB[,"cars"]), col = "blue", lwd=2)
```



The stock of cars shows a bimodal distribution as well. The group with less cars seems to have more density.

```
hist(USGasB[,"price"], prob = TRUE, col="lightblue", main = "Histogram of Price",
     xlab = "Retail Price of Gasoline")
lines(density(USGasB[,"price"]), col = "blue", lwd=2)
```



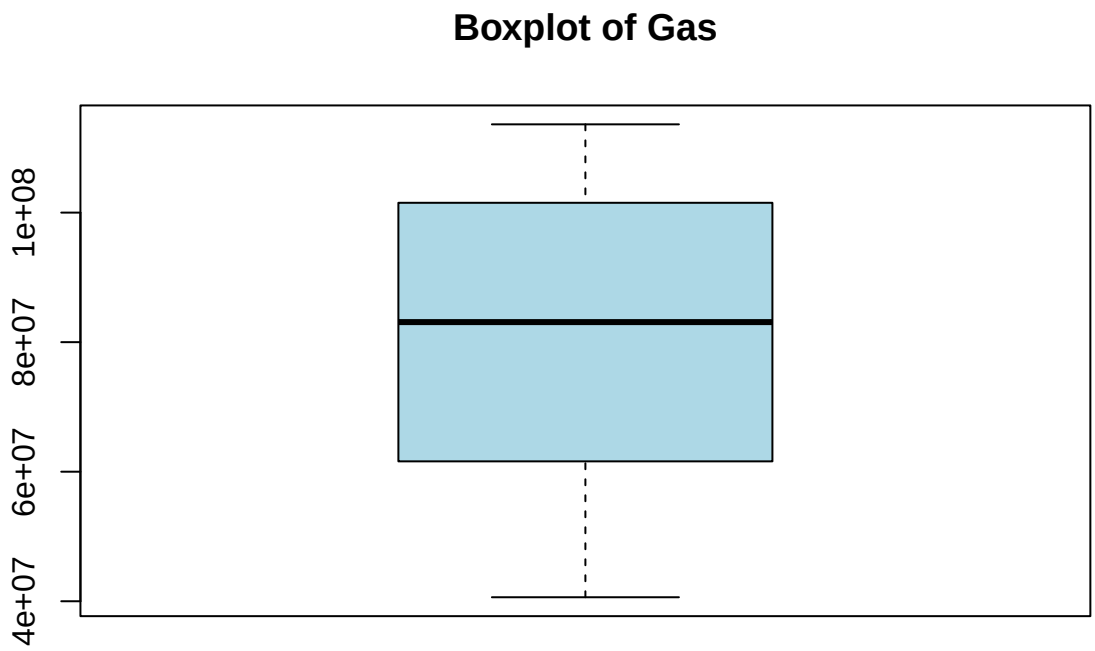
The retail price of gasoline is highly skewed to the right. This means that the retail price of gasoline tends to be on the lower end, but occasionally goes higher.

```
USGasB_vars <- USGasB[, c("gas", "cars", "price")]
summary(USGasB_vars)
```

```
##       gas                cars              price
##  Min.   : 40617285   Min.   : 49195212   Min.   :0.2720
##  1st Qu.: 61830254   1st Qu.: 71982986   1st Qu.:0.3053
##  Median : 83094370   Median :102300566   Median :0.3525
##  Mean   : 80901846   Mean   :107634304   Mean   :0.5442
##  3rd Qu.:101384955   3rd Qu.:145244051   3rd Qu.:0.6727
##  Max.   :113625960   Max.   :177922000   Max.   :1.3110
```

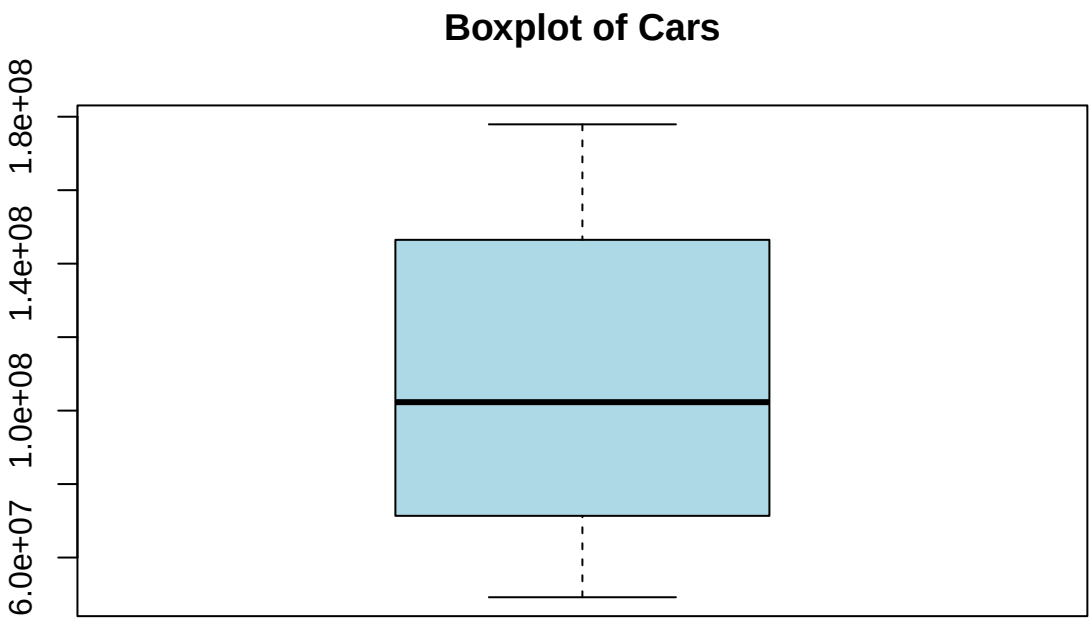
This summary of the key variables gas, cars, and price further demonstrates how each is bimodal, bimodal, and skewed, respectively.

```
boxplot(USGasB[,"gas"], main = "Boxplot of Gas", col="lightblue")
```



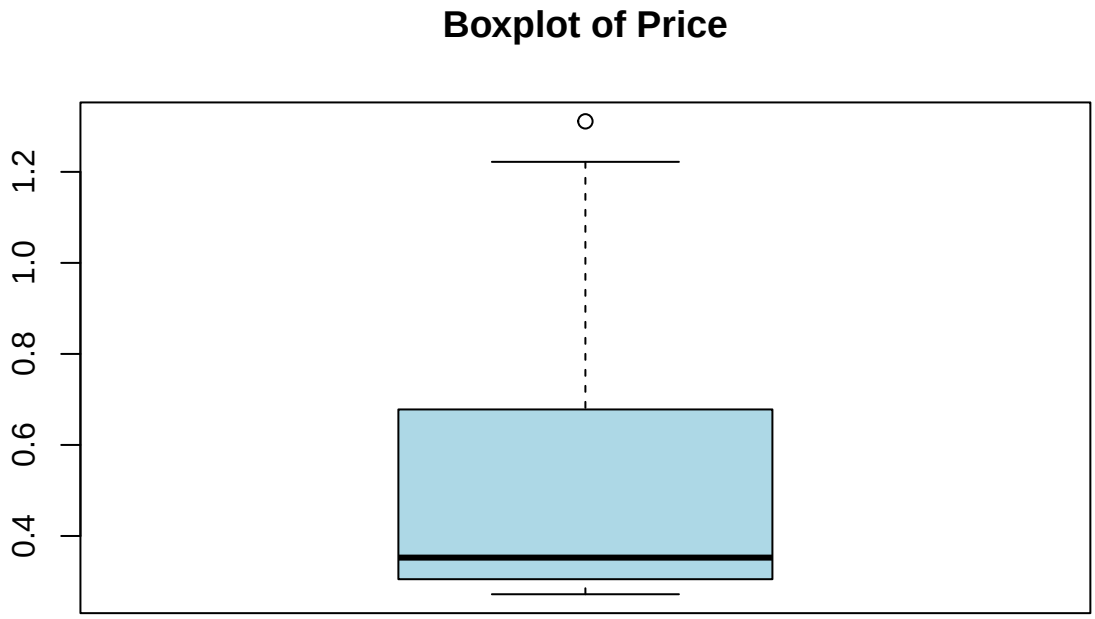
As a boxplot, gas's bimodal distribution appears more normal. The minimum is 40,617,285, the first quartile is 61,830,254, median is 83,094,370, third quartile is 101,384,955, and maximum is 113,625,960.

```
boxplot(USGasB[,"cars"], main = "Boxplot of Cars", col="lightblue")
```



As previously indicated in the histogram, the stock of cars is bimodal but appears more normal in the boxplot. The minimum is 49,195,212, first quartile is 71,982,986, median is 102,300,566, third quartile is 145,244,051, and the maximum is 177,922,000.

```
boxplot(USGasB[,"price"], main = "Boxplot of Price", col="lightblue")
```



Price has a highly skewed boxplot, with its minimum at 0.2720 and maximum at 1.3110, and yet the median is 0.3525.

```
matrix <- cor(USGasB_vars)
print(matrix)
```

```
##             gas      cars     price
## gas   1.0000000 0.9576097 0.7432761
## cars  0.9576097 1.0000000 0.8756421
## price 0.7432761 0.8756421 1.0000000
```

This correlation matrix is helpful in understanding that both cars and price are highly correlated with gas. Interestingly, cars and price are also very highly correlated with one another.

Data Pre-Processing

Gas = dependent variable Cars= independent variable price= independent variable

```
head(USGasB)
```

```
## 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
```

```
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.:221998
##  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
```

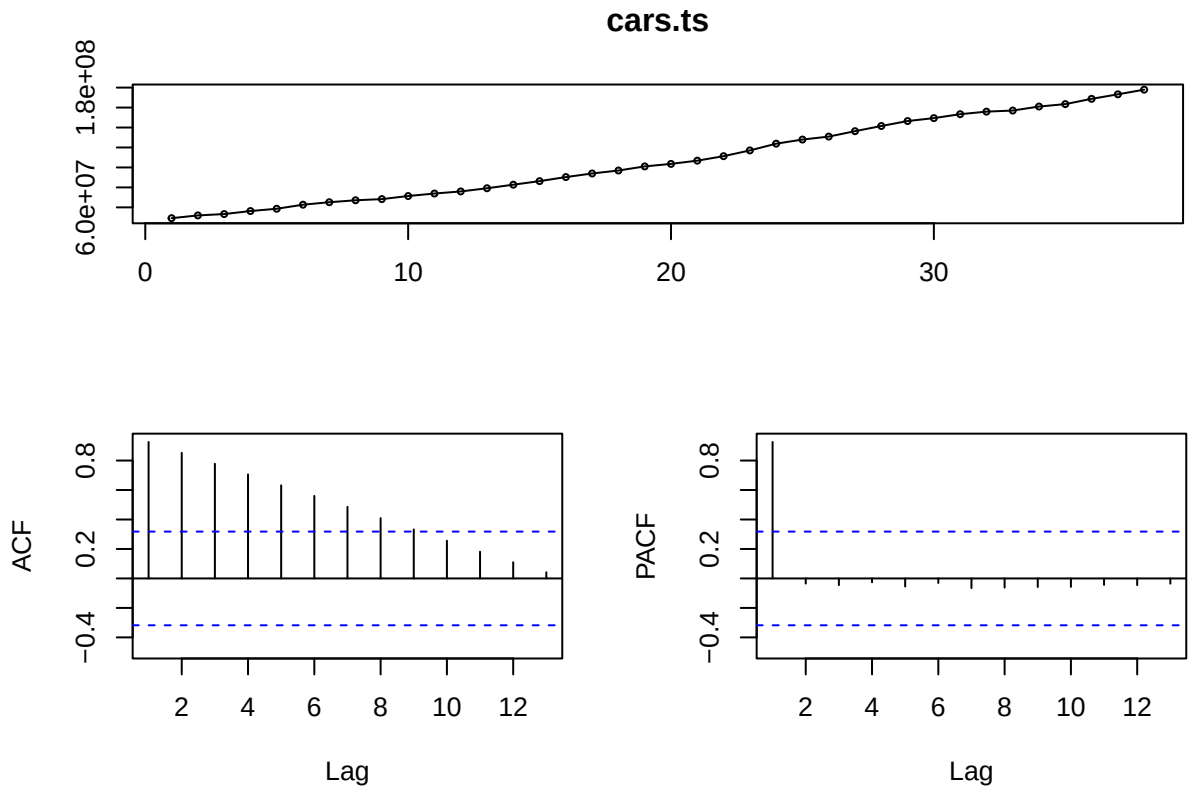
Is "USGasB" a time series data?

```
is.ts(USGasB)
```

```
## [1] TRUE
```

(a) With tsdisplay or ggtsdisplay, for each variable, use its time series plot, ACF and PACF to comment on its stationarity (you can also decompose the time series; note if there is seasonality). To supplement this, use the appropriate Dickey-Fuller (unit root) test, to determine whether or not it is stationary. Note using its PACF what the suspected order might be.

```
#tsdisplay of Cars
cars.ts <- ts(USGasB[,"cars"])
tsdisplay(cars.ts)
```

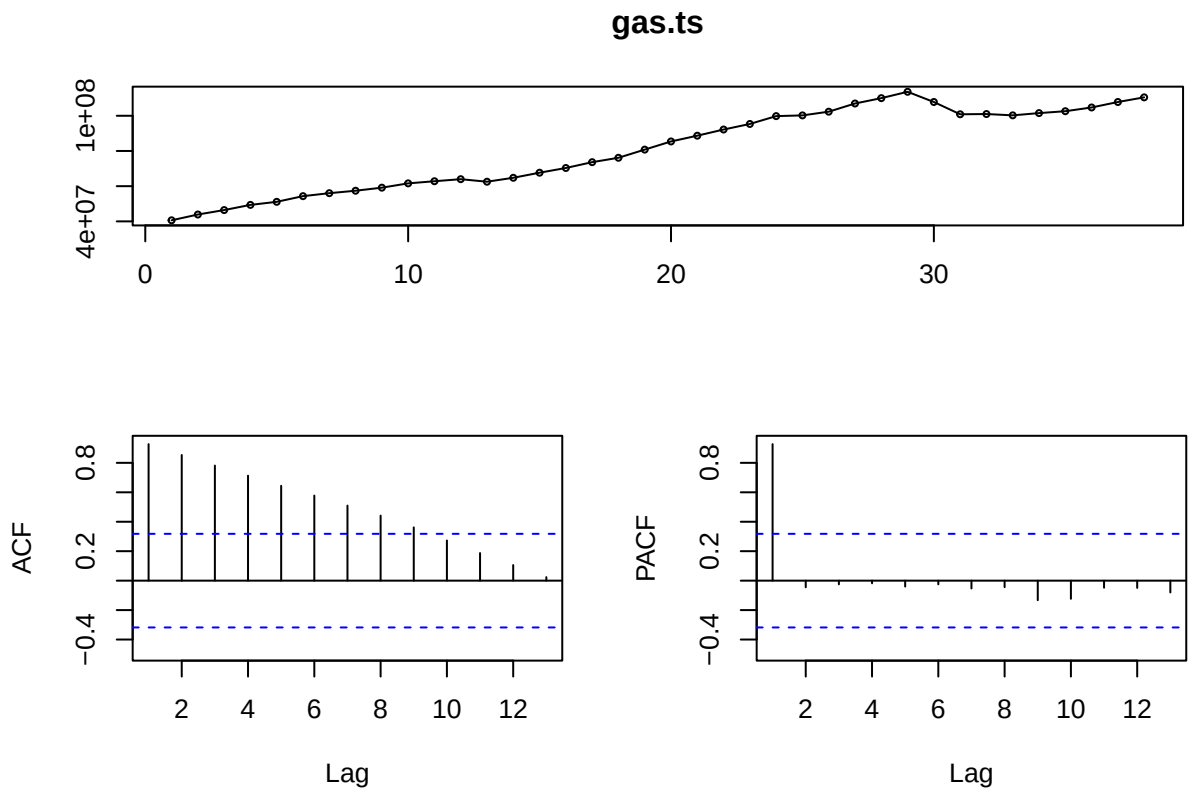


Stationary Analysis

Cars: Looking at the plot it looks to be trending therefore it is not stationary. The PACF suggests lag order of 1.

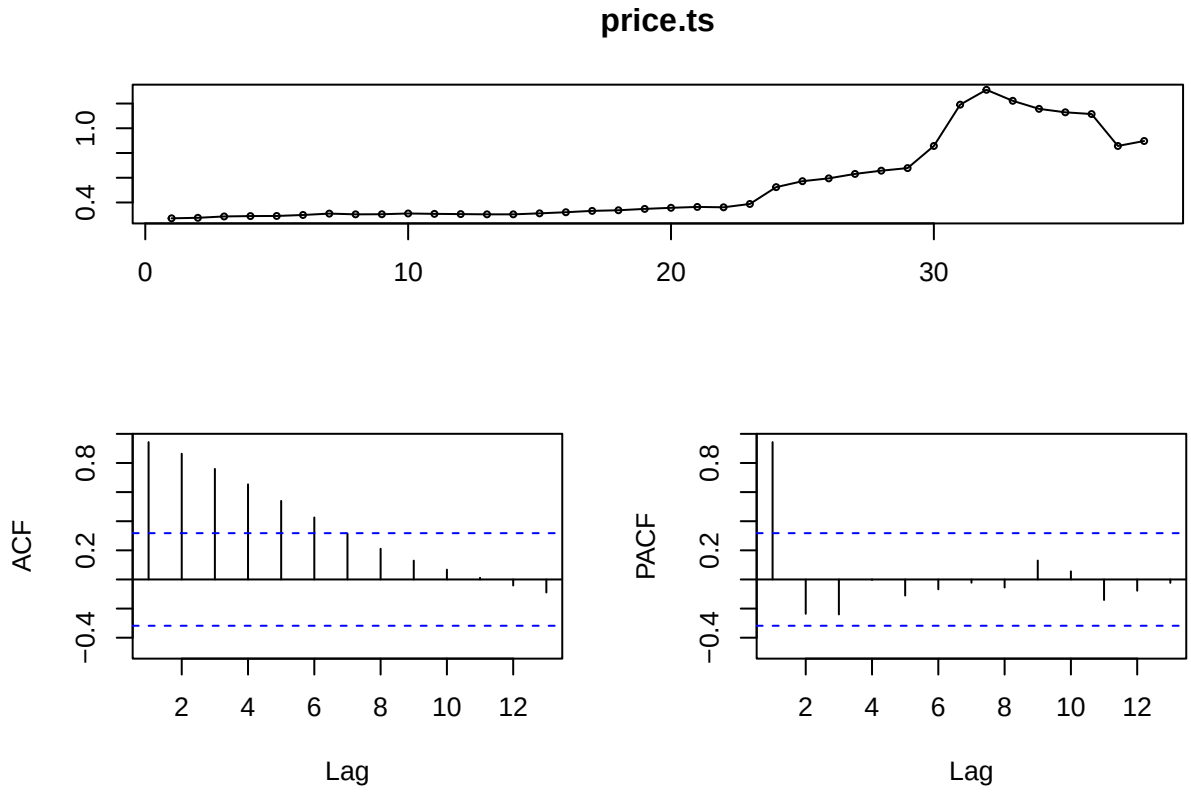
The time series plot looks to be trending because it increases with more observations, meaning this is an upward trend which suggests that as time increases so will the variable. So the mean and variance aren't constant.

```r
#tsdisplay of gas
gas.ts <- ts(USGasB[,"gas"])
tsdisplay(gas.ts)
```



Gas: Looking at the plot it looks to be mostly trending therefore it is not stationary. The PACF suggests lag order of 1, since that's where the plot shows it cuts off. The time series plot looks to be trending, looks a lot like the Cars plot of time series, meaning this is an upward trend which suggests that as time increases so will the variable. So the mean and variance aren't constant.

```r
#tsdisplay of price
price.ts <- ts(USGasB[,"price"])
tsdisplay(price.ts)
```

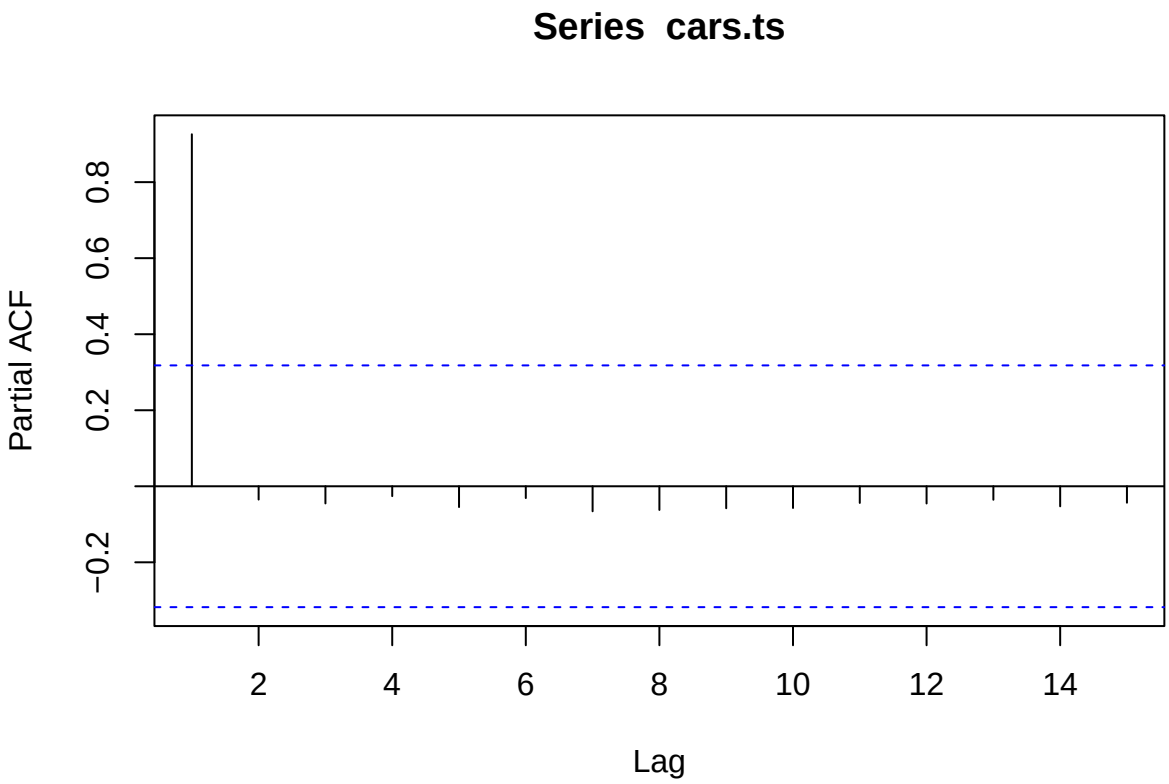


Stationary Analysis

Price: Looking at the plot it looks to be mostly trending with a bit of dip towards the end, therefore not stationary. The PACF suggests lag order of 1. It is almost consistent from observation 1 to 20 but after that it starts resembling the random walk plot, which suggests that is likely non stationary. So the mean and variance aren't constant.

PACF summary

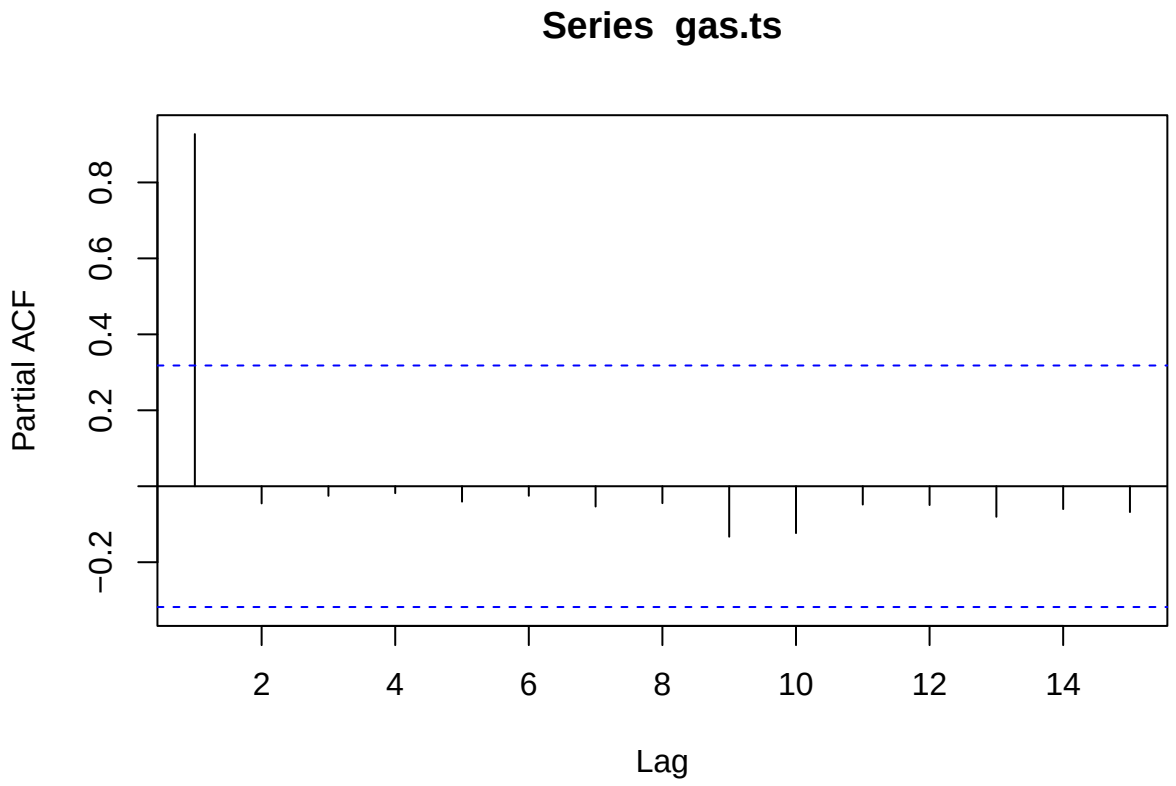
```
cars_pacf <- pacf(cars.ts)
```




```
cars_pacf
```

```
##
## Partial autocorrelations of series 'cars.ts', by lag
##
##      1      2      3      4      5      6      7      8      9     10     11
##  0.926 -0.035 -0.045 -0.026 -0.055 -0.031 -0.066 -0.062 -0.058 -0.057 -0.044
##     12     13     14     15
## -0.045 -0.036 -0.053 -0.043
```

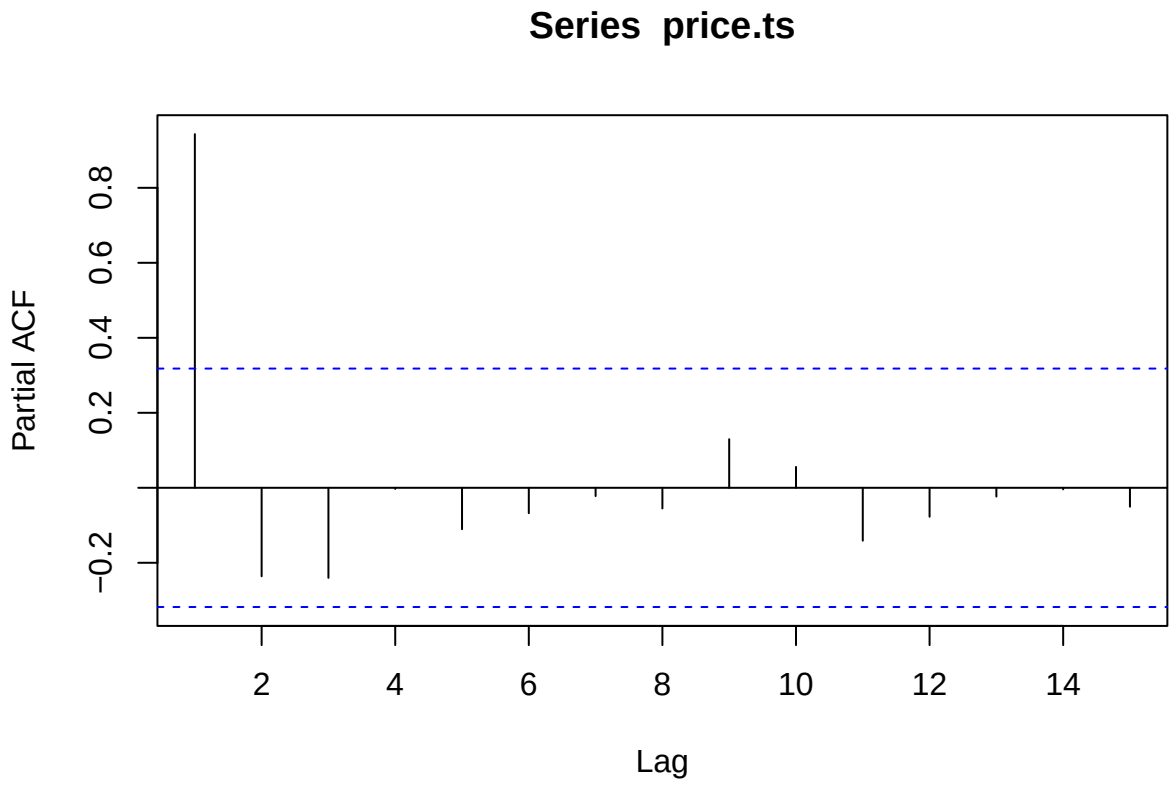
```
gas_pacf <- pacf(gas.ts)
```




```
gas_pacf
```

```
##
## Partial autocorrelations of series 'gas.ts', by lag
##
##      1      2      3      4      5      6      7      8      9     10     11
##  0.927 -0.045 -0.025 -0.018 -0.040 -0.025 -0.053 -0.045 -0.133 -0.123 -0.048
##     12     13     14     15
## -0.049 -0.080 -0.060 -0.068
```

```
price_pacf <- pacf(price.ts)
```




```
price_pacf
```

```
##
## Partial autocorrelations of series 'price.ts', by lag
##
##      1      2      3      4      5      6      7      8      9     10     11
##  0.943 -0.236 -0.240 -0.003 -0.110 -0.068 -0.022 -0.055  0.130  0.056 -0.141
##     12     13     14     15
## -0.077 -0.023 -0.004 -0.050
```

```r
#Dickey Fuller Unit Root test
#Unit root test for Cars
adf_test_cars <- adf.test(cars.ts)
adf_test_cars
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  cars.ts
## Dickey-Fuller = -2.0349, Lag order = 3, p-value = 0.5597
## alternative hypothesis: stationary
```

```r
#Unit root test for Gas
adf_test_gas <- adf.test(gas.ts)
adf_test_gas
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  gas.ts
## Dickey-Fuller = -1.819, Lag order = 3, p-value = 0.644
## alternative hypothesis: stationary
```

```r
#Unit root test for Price
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
```

Interpreting Stationarity Using Unit Root Test

Hypothesis:

$H0$: The data has unit root test and is non-stationary

$H1$: The data does not have unit root test and is stationary

For Cars: p-value = 0.5597 > 0.05: Fail to reject the null hypothesis (H0), the data has unit root and is non-stationary

For Gas: p-value = 0.644 > 0.05: Fail to reject the null hypothesis (H0), the data has unit root and is non-stationary

For Price: p-value = 0.4953 > 0.05: Fail to reject the null hypothesis (H0), the data has unit root and is non-stationary

Note: As this data is yearly data, we don't have to decompose it meaning there isn't seasonality involved in the data.

(b) If it is not stationary, determine the level of differencing to make our series stationary. We can use the ndiffs function which performs a unit-root test to determine this. After this, difference your data to ascertain a stationary time series. Re-do part a) for your differenced time series and comment on the time series plot, ACF and PACF. Recall that the time series models we've observed rely on stationarity.

```
#Differencing test for Cars
diff_cars <- ndiffs(cars.ts)
diff_cars
```

```
## [1] 2
```

For Cars: The value of 2, suggests that there needs to be a second order differencing

```
#Differencing test for gas
diff_gas <- ndiffs(gas.ts)
diff_gas
```

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

For Gas: The value of 1, suggests that there needs to be a first order differencing

```
#Differencing test for price

diff_price <- ndiffs(price.ts)
diff_price
```

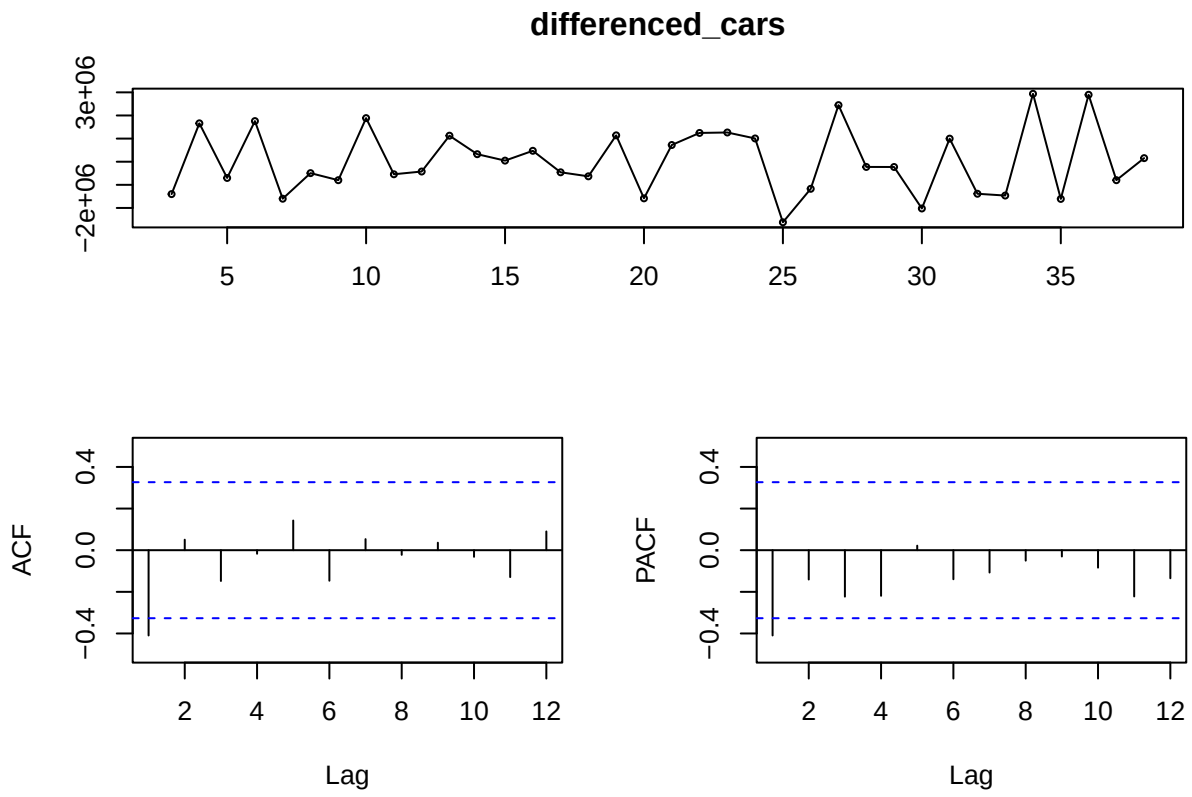
```
## [1] 1
```

For Price: The value of 1, suggests that there needs to be a first order differencing.

Note: Even though the "ndiffs" gave result of 2,1,1 for car, gas, price, respectively. I used second order differencing for all of them to make it consistent, and so it doesn't cause inconsistencies when it comes to creating AR models.

Using timeseries, ACF and PACF on differenced/stationary Variables

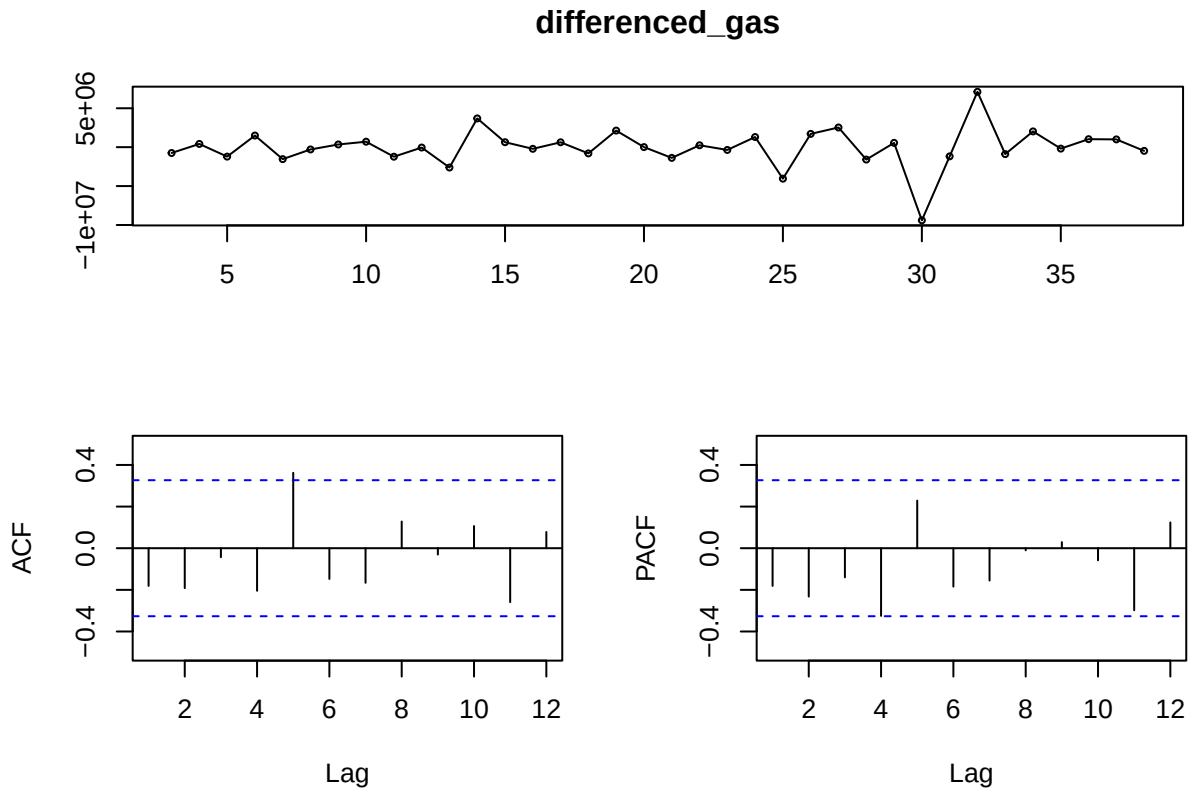
```
#timeseries, ACF and PACF of Cars
differenced_cars <- diff(cars.ts, lag = 1, difference = 2)
tsdisplay(differenced_cars)
```



Stationary Analysis of Differenced Variables

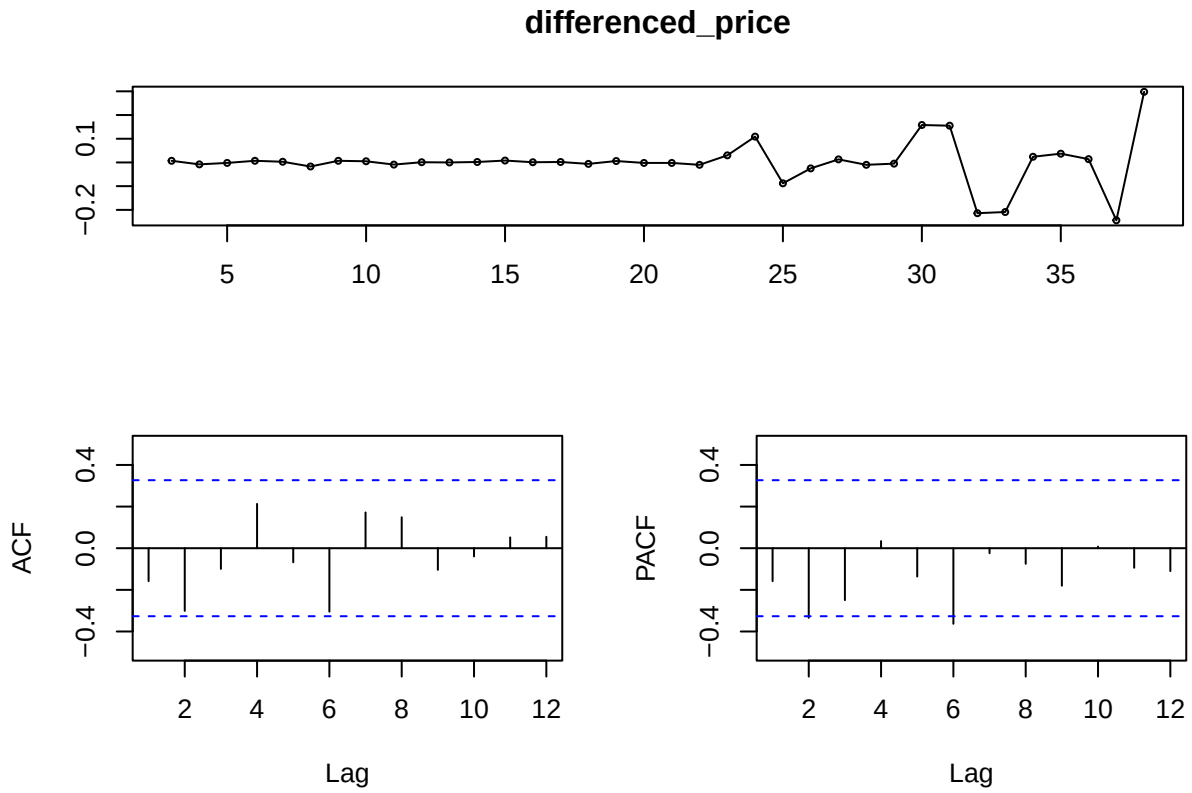
For Cars: Although the plot may look like it's a random walk, it seems to be mean reverting, meaning that the time series has a tendency to move toward a certain mean, so the mean and variance will be constant, therefore the plot suggests Stationarity. Around the 1970's alot of policies and oil related events caused inconsistencies in the stock of cars.

```r
#timeseries, ACF and PACF of gas
differenced_gas <- diff(gas.ts, lag = 1, difference = 2)
tsdisplay(differenced_gas)
```



For Gas: As seen with the plot of the Cars time series, the plot of gas time series is also very similar in the sense that it looks like a random walk, but it seems to be mean reverting, also note that there is huge dip around observation 30 but right after there's a huge spike followed by a dip, this could also mean, mean reverting meaning that the time series has a tendency to move toward a certain mean, so the mean and variance will be constant, therefore the plot suggests Stationarity. Another explanation for the dip and spike around observation 30 to 35 is that the arab oil embargo around 1973 caused fuel shortages and a spike in oil prices with long lines at gas stations, which is a real world explanation for the inconsistency

```r
#timeseries, ACF and PACF of price
differenced_price <- diff(price.ts, lag = 1, difference = 2)
tsdisplay(differenced_price)
```



For Price: The plot looks to be very consistent up until observation 25 but after that it seems to be mean reverting, so all in all, it suggests that the time series has a tendency to move toward a certain mean, so the mean and variance will be constant,therefore the plot suggests Stationarity. The real world explanation were the factors with OPEC and gas shortages in the U.S that caused it to go up and down in the time series plot, since they were unusual events.

Unit root test on differenced variables

```
#Dickey Fuller Unit Root test
#Unit root test for Cars
adf_test_cars_differenced <- adf.test(differenced_cars)
adf_test_cars_differenced
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  differenced_cars
## Dickey-Fuller = -4.197, Lag order = 3, p-value = 0.01343
## alternative hypothesis: stationary
```

```
#Unit root test for Gas
adf_test_gas_differenced <- adf.test(differenced_gas)
adf_test_gas_differenced
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  differenced_gas
## Dickey-Fuller = -4.522, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```
#Unit root test for Price
adf_test_price_differenced <- adf.test(differenced_price)
adf_test_price_differenced
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  differenced_price
## Dickey-Fuller = -2.9668, Lag order = 3, p-value = 0.197
## alternative hypothesis: stationary
```

Interpreting Stationary Using Unit Root Test after differencing

Hypothesis:

$H0$: The data has unit root test and is non-stationary $H1$: The data does not have unit root test and is stationary

For Cars: p-value = 0.01343 < 0.05: reject the null hypothesis (H0), The data does not have unit root test and is stationary

For Gas: p-value = 0.01 < 0.05: reject the null hypothesis (H0), The data does not have unit root test and is stationary

For Price: p-value = 0.197 > 0.05: Fail to reject the null hypothesis (H0), The data has unit root test and is non-stationary

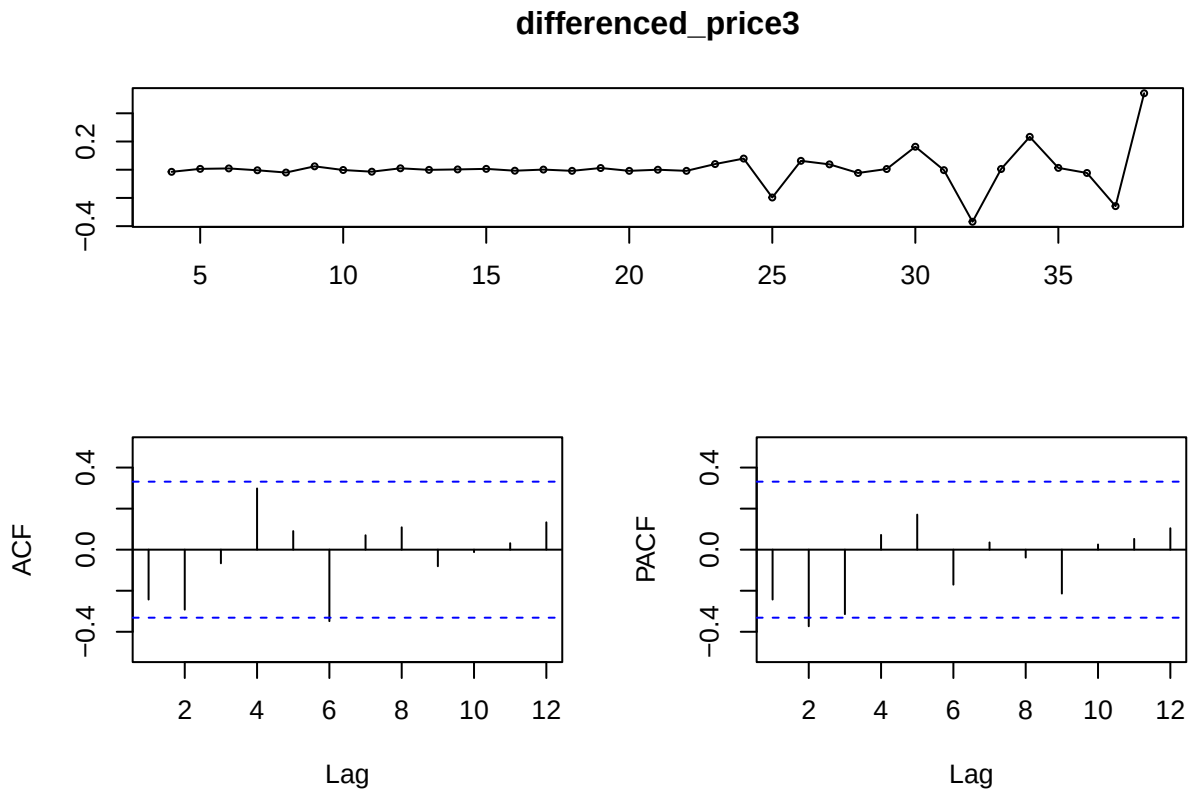
Note: For Price: *ProfessorBoswell* said it's ok that the p-value is greater than 0.05. Since sometimes the ADF doesn't capture all of the situation or detect contextual anomalies that might have occurred. It doesn't necessarily mean it is non-stationary.

Differencing means removing the trends and as seen from the time series plot before differencing that there was trending, we had to remove it in order to make the variables stationary. In order to confirm that they are stationary we had to use the Unit root test to confirm that the differenced variables do in fact have constant mean and variances, which means they are stationary. But ADF is not always accurate or cannot factor in the anomalous events so we got a p-value that would suggest that Price is non-stationary, but its not.

Extra Note:

For the sake of transparency: Referring to the p-value of Price and the fact that if we difference it 3 times, we end up with stationarity.

```
#timeseries, ACF and PACF of price
differenced_price3 <- diff(price.ts, lag = 1, difference = 3)
tsdisplay(differenced_price3)
```



```r
#Unit root test for Price
adf_test_price_differenced <- adf.test(differenced_price3)
adf_test_price_differenced
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  differenced_price3
## Dickey-Fuller = -3.8433, Lag order = 3, p-value = 0.02897
## alternative hypothesis: stationary
```

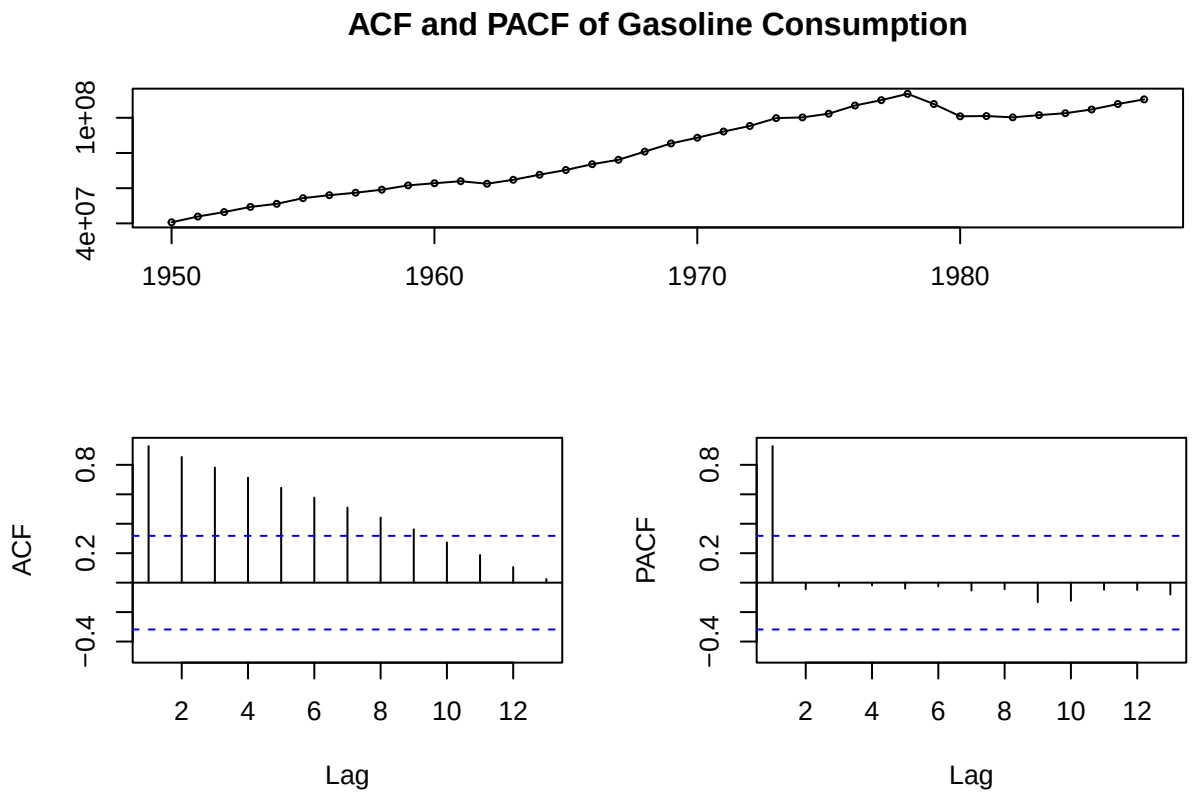
For Price: p-value = 0.02897 < 0.05: reject the null hypothesis (H0), The data does not have unit root test and is stationary.

The reason we don't want to difference it 3 times is because we might end up over-differencing which can make it difficult to identify meaningful patterns. We also learned from ndiff that we should difference it once but in-order to be consistent we differenced all the variable twice. But we don't want to do it 3 times because that will be over-differencing, which might take away any meaningful pattern. Also, the ADF will struggle to provide accurate results if the data provided is more complex and has anomalous characteristics.

Feature Generation, Model Testing and Forecasting

(a) Fit an AR(p) model to the data (using part 2(a), AIC or some built in R function)

```r
df <- USGasB[,c('cars','gas', 'price')]
tsdisplay(df[,'gas'], main = "ACF and PACF of Gasoline Consumption")
```

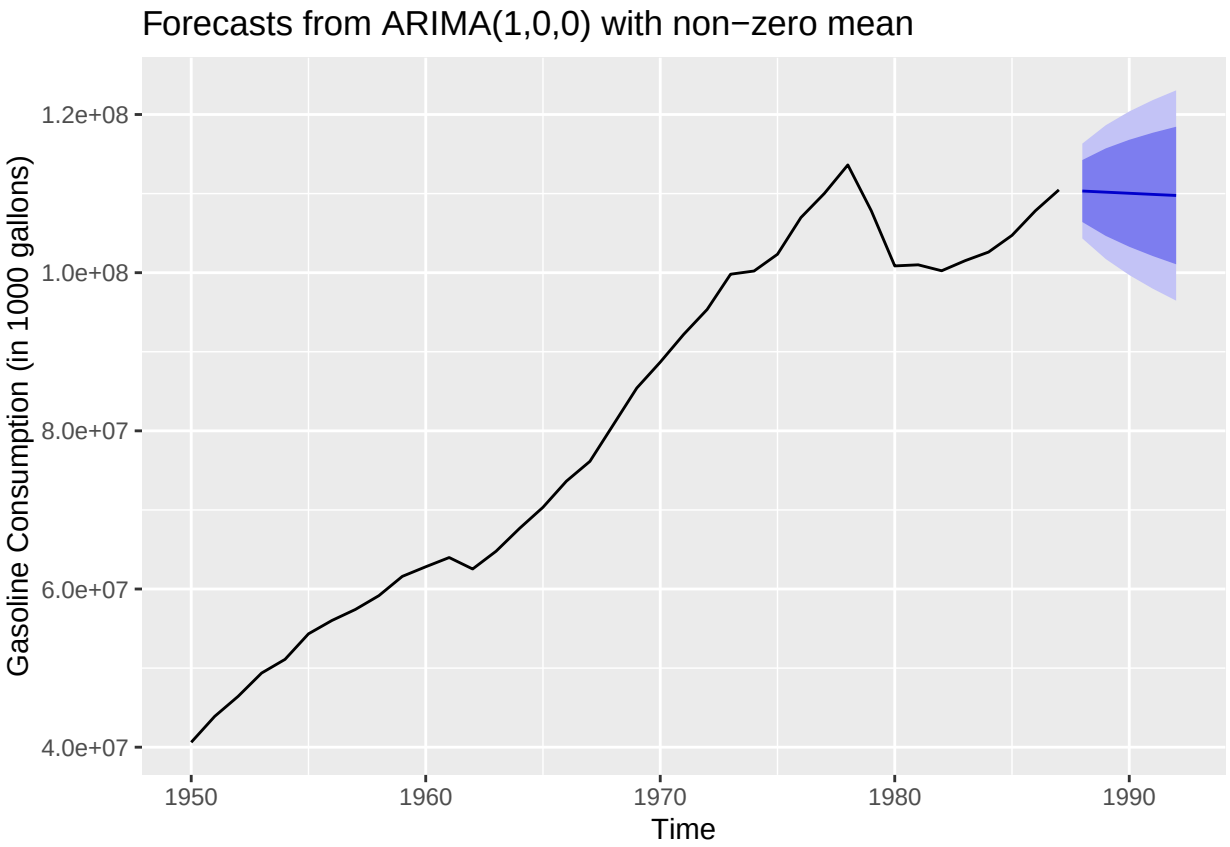


Using the `tsdisplay` ACF and PACF plots from part 2(a), we see that ACF is steadily decreasing to 0, while PACF exhibits one strong spike at lag 1 and cuts off after. This would be typical of an AR(1) process. Thus, we try fitting an AR(1) model as follows:

```
ar1 <- arima(df[,'gas'], order = c(1,0,0))
print(ar1)
```

```
##
## Call:
## arima(x = df[, "gas"], order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.9959   75906045
## s.e.  0.0058   27132115
##
## sigma^2 estimated as 9.36e+12:  log likelihood = -623.8,  aic = 1253.6
```

```
autoplot(forecast(ar1, h = 5), ylab = "Gasoline Consumption (in 1000 gallons)")
```



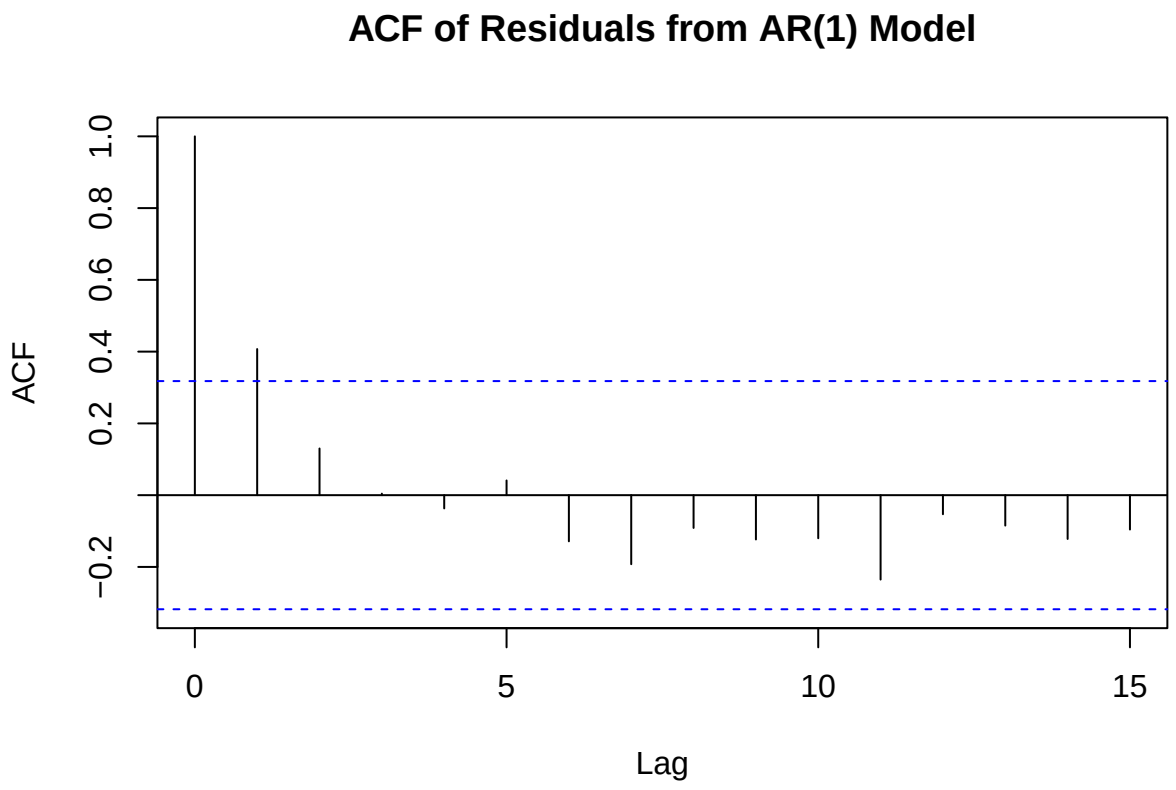
The model estimated is as follows:

$$Y_t = 75906045 + 0.9959Y_{t-1} + e_t$$

(b) Plot and comment on the ACF of the residuals of the model chosen in 3(a). If the model is properly fit, then we should see no autocorrelations in the residuals. Carry out a formal test for autocorrelation and comment on the results.

We can save the residuals from this model and check for autocorrelation in the residuals.

```
resid <- ar1$residuals
acf(resid, main = "ACF of Residuals from AR(1) Model")
```




```
#objective tests
res_mod <- lm(resid ~ lag(resid))
bgtest(res_mod, order = 1)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  res_mod
## LM test = 1.3208, df = 1, p-value = 0.2504
```

```
dwtest(res_mod)
```

```
##
##  Durbin-Watson test
##
## data:  res_mod
## DW = 2.3467, p-value = 0.842
## alternative hypothesis: true autocorrelation is greater than 0
```

The ACF plot shows two peaks at Lag 0 and at Lag 1, while the rest of the values are not statistically significant. While there seems to be no autocorrelation, we can obtain a better conclusion by using the Durbin Watson Test or the Breusch-Godfrey Test.

The Breusch-Godfrey Test returns a p-value of 0.2504. Since this is greater than $a = 0.05$, we fail to reject the null hypothesis, and cannot conclude that there is serial correlation between residuals.

We can also use results from the Durbin-Watson Test, which tests for serial correlation of order = 1. The p-value is much higher than our significance level of 5%, thus we cannot reject H_0 in favor of H_1, which states that true autocorrelation is greater than 0.

(c) Using the appropriate predictors, fit an ARDL(p,q) model to the data and repeat step (b) in part 3.

We first create a model with the first 3 lags of `gas`, `price`, and `cars` variables from the dataset. We can check which coefficients are close to 0 and remove them, retaining others.

```
mod_lag1 <- dynlm(gas ~ L(gas,1:3) + L(price, 0:3) + L(cars, 0:3), data = df)
mod_lag1
```

```
##
## Time series regression with "ts" data:
## Start = 1953, End = 1987
##
## Call:
## dynlm(formula = gas ~ L(gas, 1:3) + L(price, 0:3) + L(cars, 0:3),
##     data = df)
##
## Coefficients:
##    (Intercept)    L(gas, 1:3)1    L(gas, 1:3)2    L(gas, 1:3)3  L(price, 0:3)0
##      2.685e+06       1.065e+00      -2.638e-01       3.739e-02      -1.497e+07
## L(price, 0:3)1  L(price, 0:3)2  L(price, 0:3)3   L(cars, 0:3)0   L(cars, 0:3)1
##      1.159e+07       2.302e+06      -8.251e+06       9.704e-01      -1.043e+00
##  L(cars, 0:3)2   L(cars, 0:3)3
##      4.707e-03       2.072e-01
```

If we drop variables gas_{t-3} and $cars_{t-2}$, $cars_{t-3}$, we get the following regression:

$$gas_t = 2.685 * 10^6 + 1.065 gas_{t-1} - 0.264 gas_{t-2} - 1.497 * 10^7 price_t + 1.159 * 10^7 price_{t-1} +$$

$$2.302 * 10^6 price_{t-2} - 8.251 * 10^6 price_{t-3} + 9.704 cars_t - 1.043 cars_{t-1}$$

We can also experiment with models of different lags and check AIC/BIC values.

```
mod_lag2 <- dynlm(gas ~ L(gas,1:2) + L(price, 0:2) + L(cars, 0:2), data = df)
mod_lag3 <- dynlm(gas ~ L(gas,1:3) + L(price, 0:3), data = df)
mod_lag4 <- dynlm(gas ~ L(gas,1:3) + L(cars, 0:3), data = df)


AIC(mod_lag1, mod_lag2, mod_lag3, mod_lag4)
```

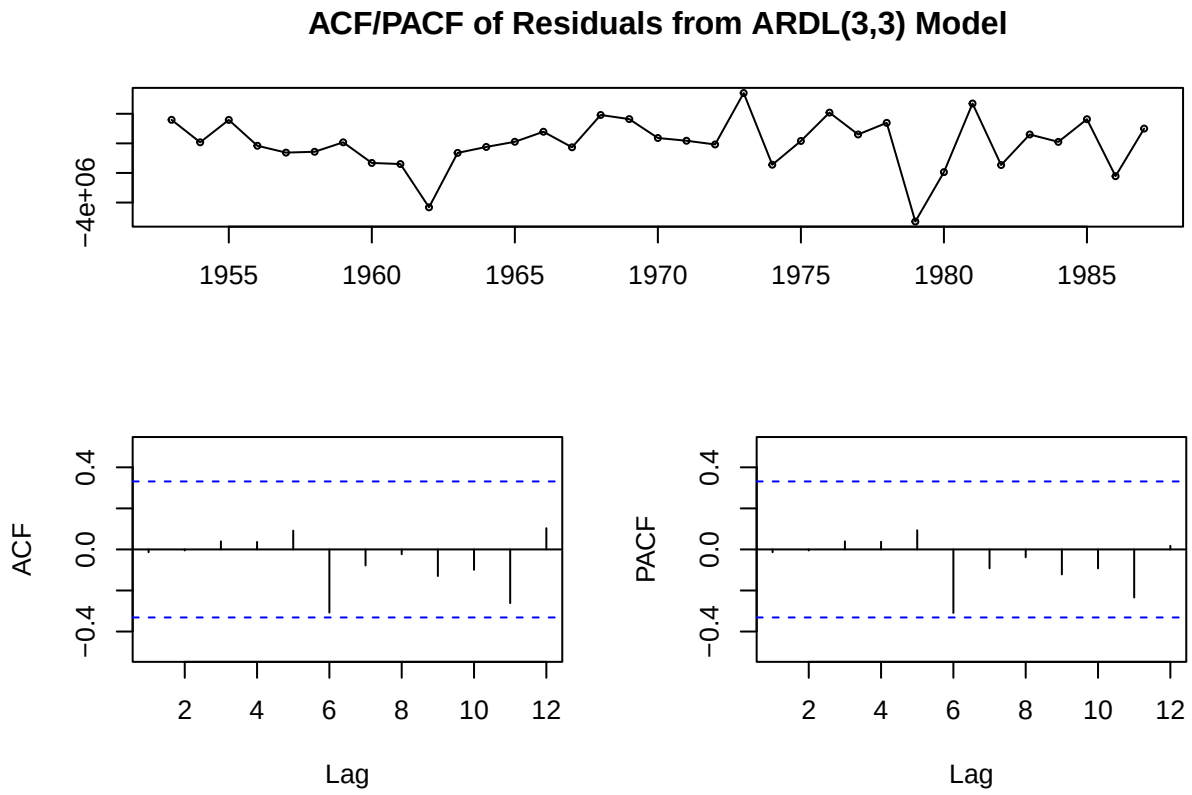
```
##          df      AIC
## mod_lag1 13 1117.714
## mod_lag2 10 1145.315
## mod_lag3  9 1123.284
## mod_lag4  9 1124.905
```

```
BIC(mod_lag1, mod_lag2, mod_lag3, mod_lag4)
```

```
##          df      BIC
## mod_lag1 13 1137.933
## mod_lag2 10 1161.150
## mod_lag3  9 1137.282
## mod_lag4  9 1138.903
```

We get the lowest value of AIC for `mod_lag1`, which was the ARDL (3,3,3) model with up to 3 lags of the dependant and independant variable. However, BIC gives us the lowest value for `mod_lag3`, which was the ARDL(3,3) model with only lags of gas and price. We should go with the result given by BIC and select `mod_lag3`, since it harshly penalizes models which include extra variables. If we need to include more variables, or more lags of the dependant variable, we will see this in the serial correlation of errors.

```
resid <- mod_lag3$residuals
tsdisplay(resid, main = "ACF/PACF of Residuals from ARDL(3,3) Model")
```




```
res_mod2 <- lm(resid ~ lag(resid, 1))
bgtest(res_mod2, order = 1)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  res_mod2
## LM test = 0.82125, df = 1, p-value = 0.3648
```

```
dwtest(res_mod2)
```

```
##
##  Durbin-Watson test
##
## data:  res_mod2
## DW = 2.2768, p-value = 0.7969
## alternative hypothesis: true autocorrelation is greater than 0
```

```
mod_lag3
```

```
##
## Time series regression with "ts" data:
## Start = 1953, End = 1987
##
## Call:
## dynlm(formula = gas ~ L(gas, 1:3) + L(price, 0:3), data = df)
##
## Coefficients:
##    (Intercept)    L(gas, 1:3)1    L(gas, 1:3)2    L(gas, 1:3)3  L(price, 0:3)0
##      1.068e+06       1.014e+00      -1.181e-01       1.664e-01      -1.957e+07
## L(price, 0:3)1  L(price, 0:3)2  L(price, 0:3)3
##      1.090e+07      -3.455e+05       2.647e+06
```

To check residuals for serial correlation, we used the ACF/PACF plots, the DW Test and the BG Test.
- ACF/PACF show no statistically significant peaks, thus there is no visual evidence of autocorrelation.
- `bgtest` gives us a p-value of r`bgtest(res_mod2, order = 1)$p.value`, which is greater than the significance level of 5%, and thus we fail to reject the null hypothesis of no autocorrelation.

- `dwtest` gives us a p-value of r`dwtest(res_mod2)$p.value`, which is also much higher than the significance level, and we do not have enough evidence to conclude autocorrelated errors.

Hence, we have created an ARDL(3,3) model with errors that are not autocorrelated. The model is:

$$gas_t = 1.068 * 10^6 + 1.014 gas_{t-1} - 0.118 gas_{t-2} + 0.166 gas_{t-3} - 1.957 * 10^7 price_t +$$

$$1.090 price_{t-1} - 3.455 * 10^5 price_{t-2} + 2.647 * 10^6 price_{t-3}$$

Q4. Provide a brief summary of your findings and state which model performs better.

```
AIC(ar1, mod_lag3)
```

```
##          df      AIC
## ar1       3 1253.600
## mod_lag3  9 1123.284
```

```
BIC(ar1, mod_lag3)
```

```
##          df      BIC
## ar1       3 1258.512
## mod_lag3  9 1137.282
```

In summary, we looked at two models—the ARDL(3,3) and the AR(1) model—for predicting gasoline consumption. Analysis of ACF/PACF plots, the Breusch-Godfrey Test, and the Durbin Watson test consistently demonstrated that the residuals of both models lacked noticeable autocorrelation. The ARDL(3,3) model outperformed the AR(1) model, because it gave lower AIC and BIC values than the AR(1) model. Thus, we can conclude that the AR(3,3) model performs better and strikes a good balance between model fit and complexity. Moreover, since residuals from the ARDL model are not serially correlated, this model is more suitable for making unbiased forecasts.

The ARDL(3,3) model uses lags of both dependant and independant variables predict future gasoline consumption. Intuitively, this can help decision-makers understand the factors influencing gasoline consumption. Thus, the ARDL(3,3) model works not only in statistical measures but is also a practical choice for predicting gasoline use in the real world.

Q5. Suggest any limitations faced or improvements which could've been made to the model based on your findings, which should be supplemented with statistical tests(eg. degree of freedom restrictions, reverse causality).

Some improvements we could have made to our model are:

- In future work, we should find a way to correct the skewness of the independent variables, including 'price' and 'cars'
- We should also have checked for collinearity, since we saw that 'cars' and 'price' were also highly correlated variables.

- Differencing our variables would make sure we don't violate the stationarity assumption, and this would be a useful method to avoid biased estimates, and hence inccurate forecasts.