Overview of demand for natural gas in the US in a time-series format.

Supervision

Dr: Mona Abd-Alshafi

Dr:Youssra Hassan

Presented by:

Verina Alfred Fahmy	5210407
Kerolous Marzouk Rady Farid	5210853

Table of Contents:

Introduction…………………………………………………………….3

Literature review of gold price……………………………..........4

Descriptive statistics……………………………………………........6

Data analysis "Time series analysis: modern approach…..6

Forecast future values of daily gold price…………………..……12

Auto ARIMA…………………………………………………..……………….15

Conclusion……………………………………………………………………….16

References…………………………………………………………..............16

❖ Introduction

In today's rapidly evolving energy landscape, understanding patterns and trends in US gas consumption is of paramount importance. As one of the primary sources of energy for residential, commercial, and industrial sectors, natural gas plays a crucial role in driving economic activities and meeting the nation's energy demands. Analyzing the time series data of US gas consumption not only sheds light on past consumption patterns but also serves as a vital tool for forecasting future demand, aiding policy decisions, and guiding strategic investments in the energy sector.

So;

We'll adopt a **modern approach** where we first ensure that our time series data is stable, using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. This helps us reduce the number of parameters we need to estimate and improves the accuracy of our model. We start by plotting the data and examining these ACF and PACF plots to see if the data shows consistent statistical patterns over time. If it doesn't, we make transformations to achieve stability. For instance, if there are issues with both trend and variance, we might apply a logarithmic transformation. If there's a trend problem, differencing operations can help remove it.

Once our time series is stable, we proceed to identify the suitable model—whether it's autoregressive (AR), moving average (MA), or a combination of both (ARMA)—again using ACF and PACF plots. With the model selected, we estimate its parameters and evaluate its quality, checking for invertibility. Finally, armed with these estimated parameters and past data, we're able to forecast future values, offering valuable insights into upcoming trends and patterns.

Importance of Analyzing Gas Consumption Data:

Insight into Economic Trends: Gas consumption trends often correlate with broader economic indicators. By analyzing gas consumption data over time, economists and policymakers can gain insights into economic growth, industrial production, and consumer behavior.

Energy Policy Formulation: Gas consumption data serves as a cornerstone for energy policy formulation at both the national and regional levels. Governments rely on accurate consumption forecasts to design energy policies, allocate resources, and promote sustainable energy practices.

Infrastructure Planning: Understanding consumption patterns helps in infrastructure planning and development. Forecasting future demand allows for the timely expansion or construction of pipelines, storage facilities, and distribution networks to ensure a reliable supply of natural gas to consumers.

Environmental Impact Assessment: Gas consumption directly impacts environmental factors such as greenhouse gas emissions and air quality. Analyzing consumption trends aids in assessing the environmental impact of gas usage and formulating strategies for mitigating emissions.

Significance of Time Series Forecasting:

Resource Allocation: Accurate forecasts enable energy companies to allocate resources efficiently, ensuring an adequate supply of gas to meet future demand without overstocking or shortages.

Risk Management: Forecasting helps in identifying potential supply-demand imbalances and price fluctuations, allowing stakeholders to hedge risks and make informed decisions in volatile energy markets.

Cost Optimization: By forecasting gas consumption, companies can optimize production schedules, storage utilization, and transportation logistics, leading to cost savings and improved operational efficiency.

Market Planning: Forecasting future gas consumption enables market participants to anticipate changes in demand patterns, adjust pricing strategies, and capitalize on emerging market opportunities.

❖ Background/Literature Review: Analyzing US Gas Consumption Time Series:

Domestic production of natural gas directly impacts supply levels within the US market. Factors influencing production include:

- **Technological advancements**: Innovations such as hydraulic fracturing (fracking) and horizontal drilling have enabled the extraction of gas from previously inaccessible shale formations, leading to a significant increase in domestic production.

- **Geological reserves**: The presence of abundant shale gas reserves in regions like the Marcellus and Permian basins has contributed to the growth of US natural gas production.

- **Regulatory environment**: Policies governing drilling permits, environmental regulations, and land access can affect production rates by influencing industry investment and operational decisions.

Market Dynamics: - Fluctuations in supply and demand dynamics can lead to volatility in gas prices and market conditions. Factors such as weather patterns, industrial activity, electricity generation trends, and geopolitical events can influence consumption patterns and drive shifts in supply-demand equilibrium.

- Price signals play a crucial role in incentivizing investment in production, infrastructure, and technology, helping to balance supply and demand over the long term.

Understanding these supply and demand dynamics is essential for stakeholders in the gas industry, including producers, consumers, policymakers, and investors, as they navigate market trends, anticipate future developments, and make informed decisions regarding production, investment, and trade.

Forecasting and Prediction Techniques: Researchers have explored various time series forecasting methods to predict gas consumption . These techniques include autoregressive integrated moving average (ARIMA), machine learning algorithms. Improved forecasting accuracy assists in effective planning, resource allocation, and decision-making.

TIME SERIES ANALYSIS

Modern approach

❖ Descripe statistic:

Our data is the US consumption of gas from 2001 to 2021.

The variable measured: Gas consumption

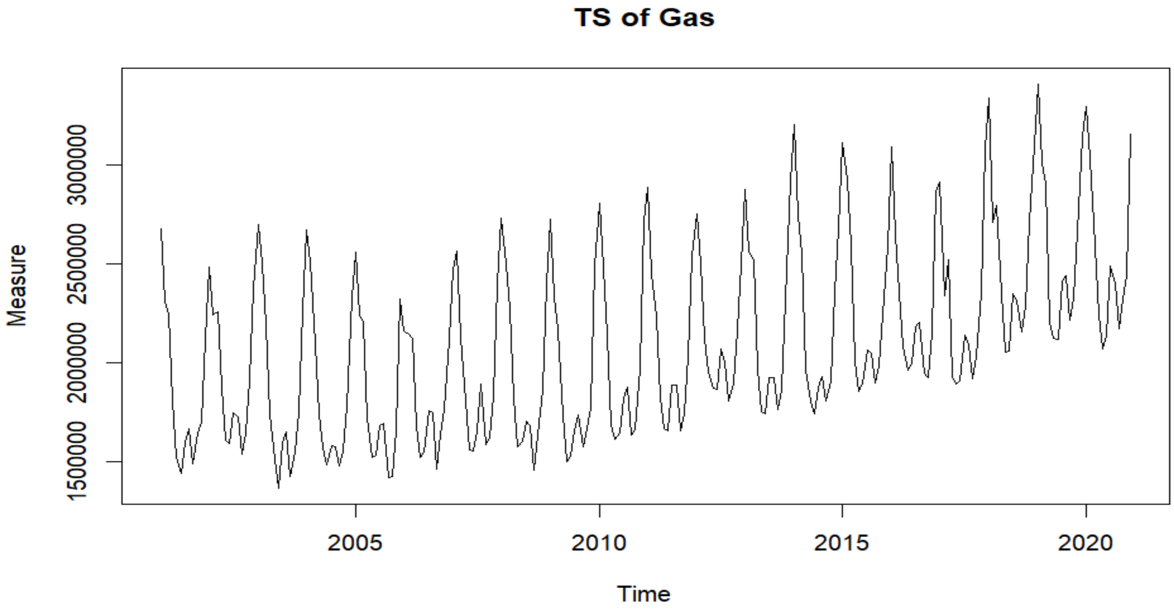
Its Scale unit: Million of cubic feet (MMCF)

The Sample size: 240 with no missing observations

Data source: https://www.kaggle.com/datasets/mexwell/usgas-dataset

❖ Time series analysis:

Presenting the data graphically:

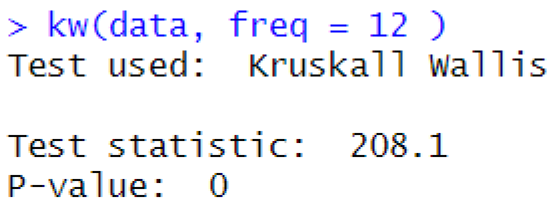


Figure(1):original time series plot

- **After we plot the gas consumption time series** we noticed that**:**

The data has seasonal variation as the time series repeat itself.
Also the data has a positive trend as it is increasing over time
“Non-stationarity in mean”
It doesn’t have outliers or turning points.

-Checking seasonality through Kruskal Wallis test:



```
> kw(data, freq = 12 )
Test used:  Kruskall Wallis

Test statistic:  208.1
P-value:  0
```

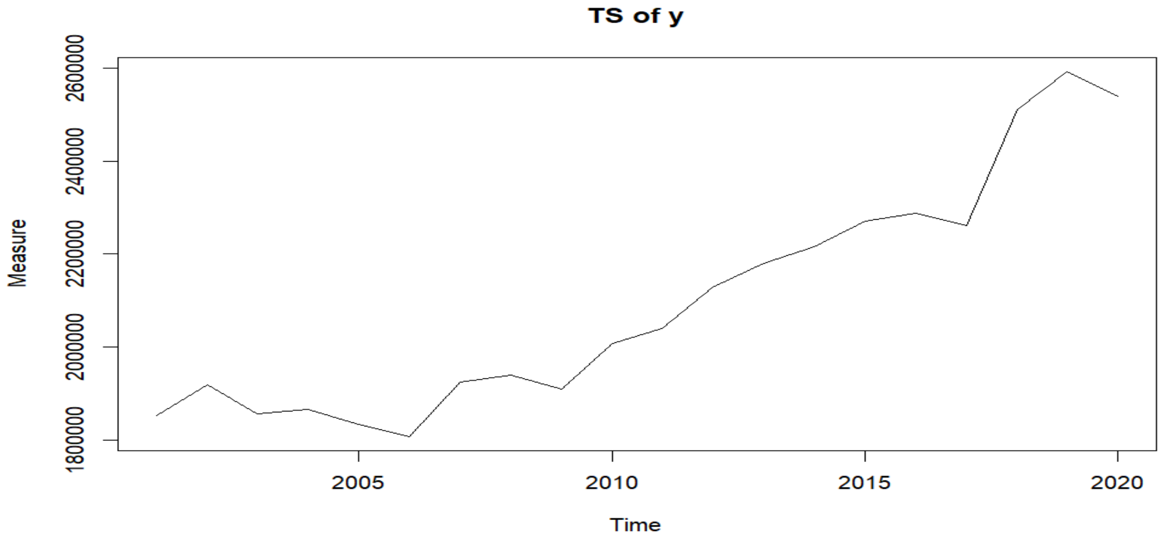
The P-value = 0 **which means** we will reject H_o at $\propto = \mathbf{0.01}$,so, there is a seasonality.

AS; H_o: no seasonality “there is no significance difference between the medians of the groups corresponding to different seasons”

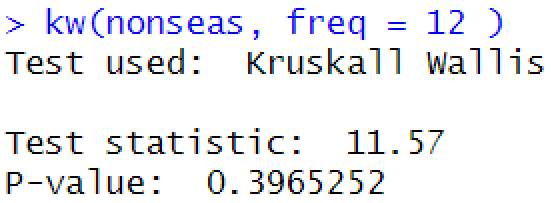
H_1: seasonality “there is significance difference between the medians of the groups corresponding to different seasons”

The Data after removing seasonality:

- We removed the seasonality by taking the average of each year to make our data Yearly instead of monthly; then :



Figure(2):The TS plot after removing seasonality



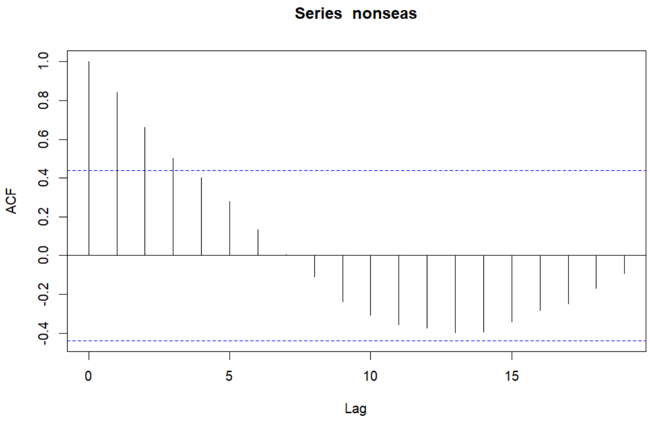
```
> kw(nonseas, freq = 12 )
Test used:  Kruskall Wallis

Test statistic:  11.57
P-value:  0.3965252
```

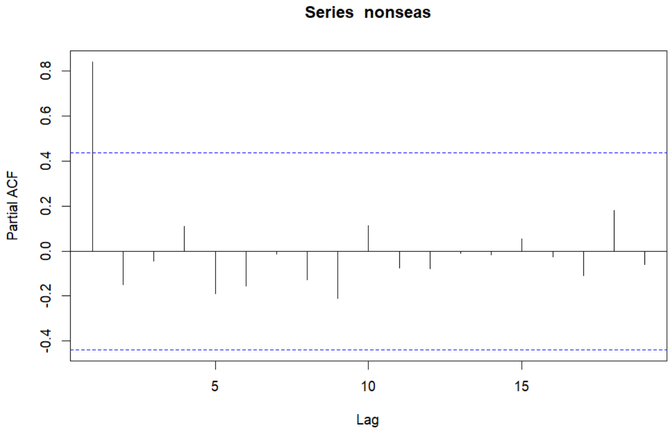
The Kruskal Wallis test is now referred to nonseasonal data as The P-value = 0.396 **which means** we will not reject H_o at $\propto = \mathbf{0.01}$.

Checking Stationarity:

The first step in model identification is to ensure the time series is stationary. The series must be converted to a stationary series to proceed, and this is accomplished by the differencing the time series using a lag in the variable.

From graph(2): The time series has a positive trend so it is nonstationary in mean.





-Checking Stationarity through “Dicky Fuller” test:
The P-value = 0.7743 **which means** we will will not

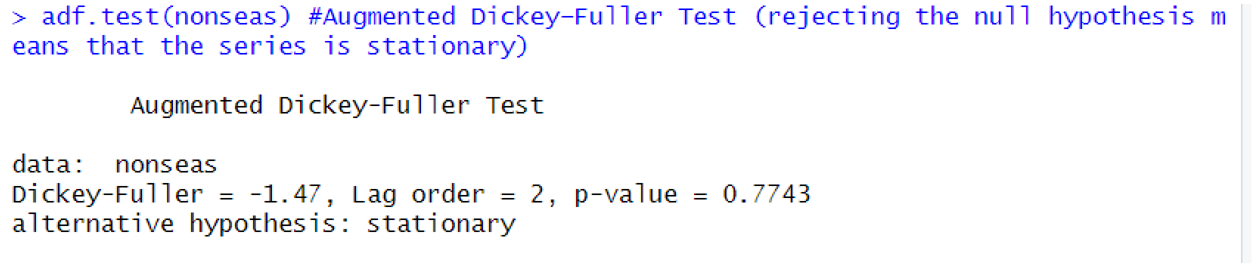
reject H_o at $\propto = 0.01$, the data is non-stationary in

mean.

As; H_o: non-stationarity

H_1: stationarity

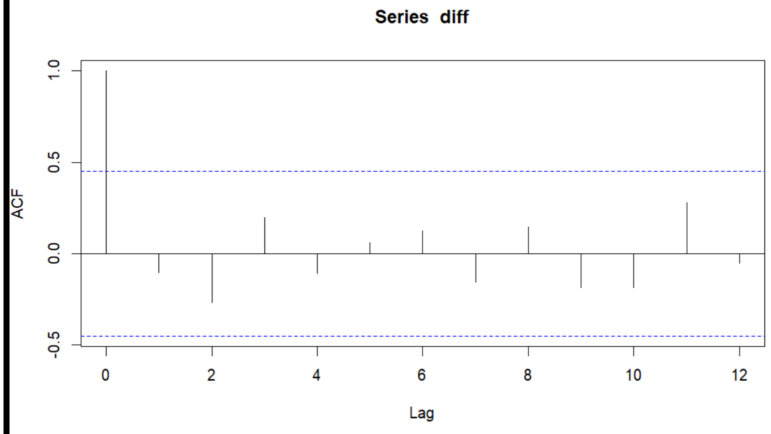
We will take the first difference it does not seem to be stationary. So we will take the second difference.

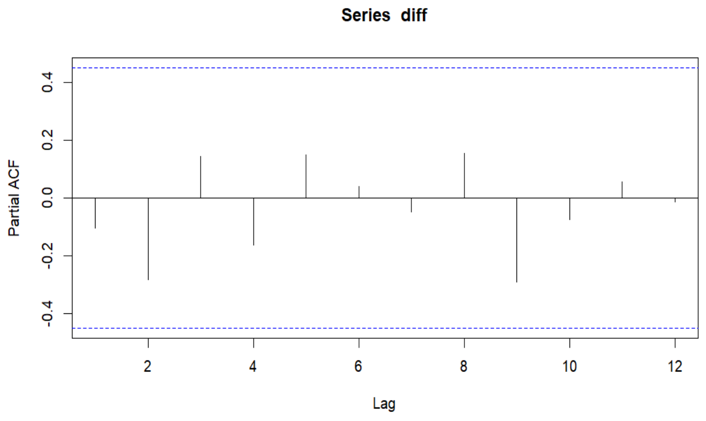


```
> adf.test(nonseas) #Augmented Dickey-Fuller Test (rejecting the null hypothesis m
eans that the series is stationary)

        Augmented Dickey-Fuller Test

data:  nonseas
Dickey-Fuller = -1.47, Lag order = 2, p-value = 0.7743
alternative hypothesis: stationary
```

Now it is stationary in mean and variance; Once we have a stationary time series “in variance & mean”, we can get ACF & PACF to the data, then we can identify the appropriate model.

First Difference:



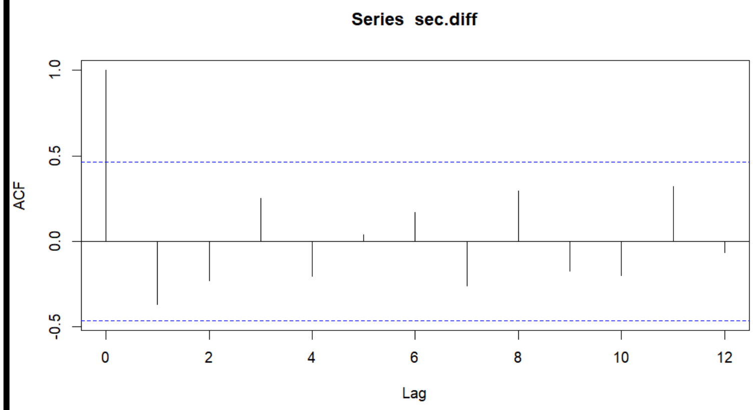


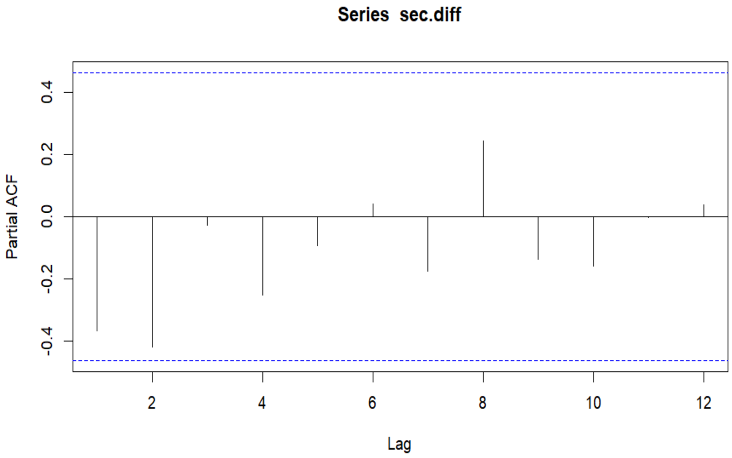
Second Difference:

From the acf and the pacf

We noticed that the data is White noise so we can’t determine ARIMA from them.

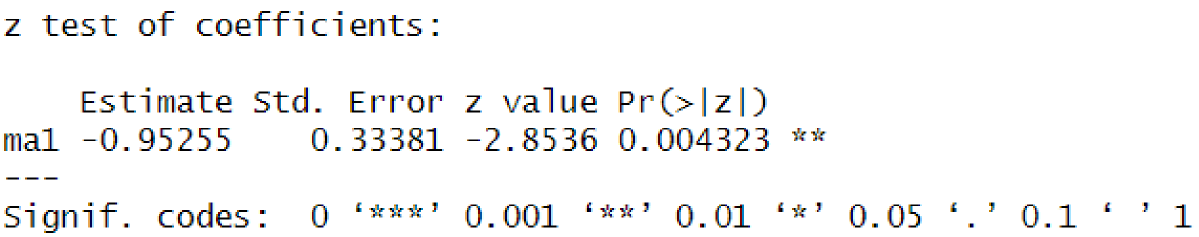
SO we try different combinations of AR & MA orders Models that satisfied the conditions ; Also if there is more than one model that achieves the conditions so we choose the best one for reducing the residuals..





Box & Jenkins Analysis (ARIMA):

Our best model is ARIMA (0,2,1)



```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ma1 -0.95255    0.33381 -2.8536 0.004323 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-Value for MA1 is lower than 0.01, so we will reject H_o at $\propto = 0.01$

ARIMA (0,2,1) estimation:

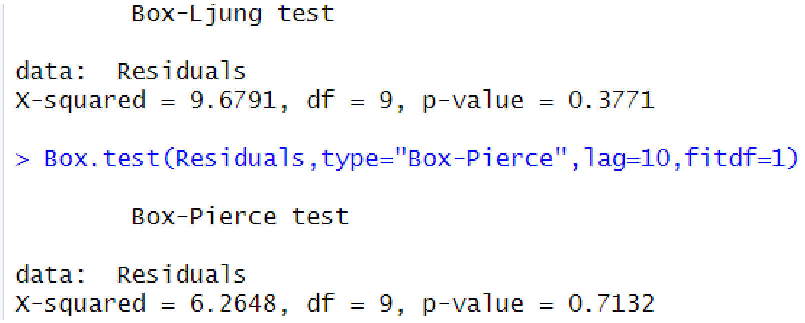
The estimation means that we will use ARIMA (0,2,1) to fit the time series data.

$$Yt = 2 * y_{t-1} - y_{t-2} + \varepsilon t - \theta * \varepsilon_{t-1}$$

$$Yt = 2 * y_{t-1} - y_{t-2} + \varepsilon t + 0.9525 * \varepsilon_{t-1}$$

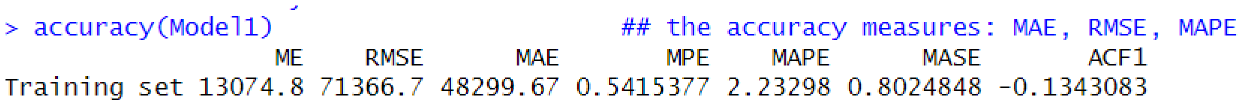
And we can conclude that ARIMA (0,2,1) is equivalent to ARMA (2,1)

o Diagnostic checking of fitted ARIMA (0,2,1) model using Box-Ljung & Box-Pierce tests:



```
        Box-Ljung test

data:  Residuals
X-squared = 9.6791, df = 9, p-value = 0.3771

> Box.test(Residuals,type="Box-Pierce",lag=10,fitdf=1)

        Box-Pierce test

data:  Residuals
X-squared = 6.2648, df = 9, p-value = 0.7132
```

p-value for both “greater than 0.01”, so we will not reject at ∝ =0.01, “which means that all two coefficients are not different from zero, insignificant”.



```
> accuracy(Model1)                          ## the accuracy measures: MAE, RMSE, MAPE
                    ME    RMSE      MAE       MPE    MAPE      MASE       ACF1
Training set 13074.8 71366.7 48299.67 0.5415377 2.23298 0.8024848 -0.1343083
```

the model is very good at predicting. It's making small errors on average, both in terms of actual values and percentage differences. Also, it seems like there's no clear pattern in the mistakes it's making. We compared between the different models using fulfills the conditions & the forecast errors. And we found that ARIMA (0, 2,1) is the optimal model.

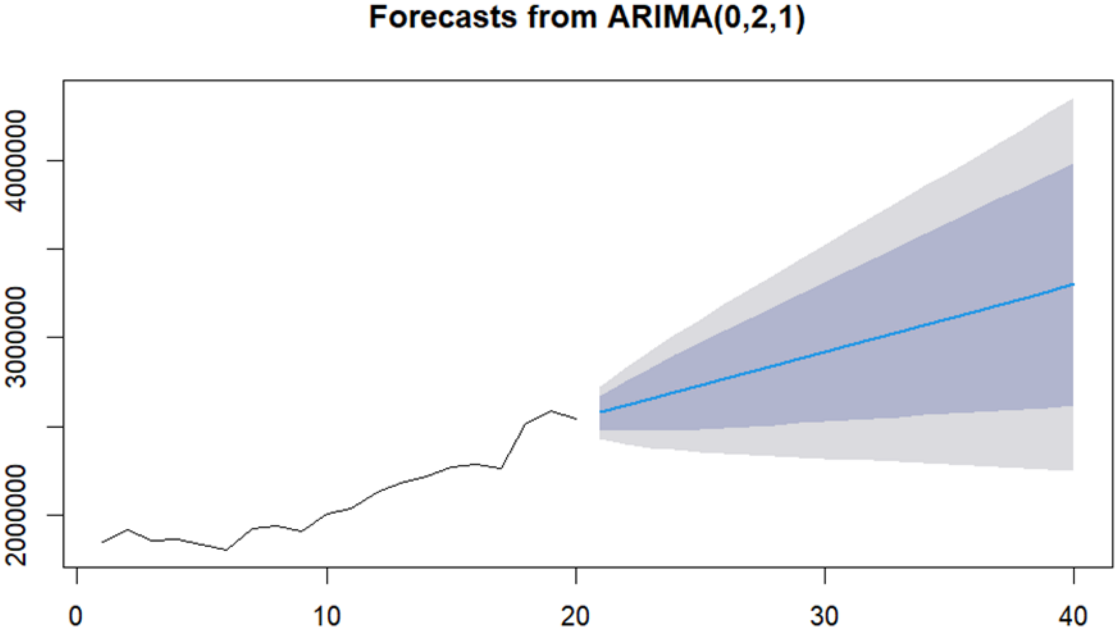
▪ **To Check the stationarity & invertibility of the model:**

$$Yt = 2 * y_{t-1} - y_{t-2} + \varepsilon t + 0.9525 * \varepsilon_{t-1}$$

- **The model is always stationarity**
- **Invertibility conditions are valid:**
  $|\theta 1| = 0.9525 < 1$

❖ Forecasting:

After a model is assured to be stationary, and fitted such that there is no information in the residuals, we can proceed to forecasting. Forecasting assesses the performance of the model against real data. Usually the utility of a specific model or the utility of several classes of models to fit actual data can be assessed by minimizing a value such as root mean square.



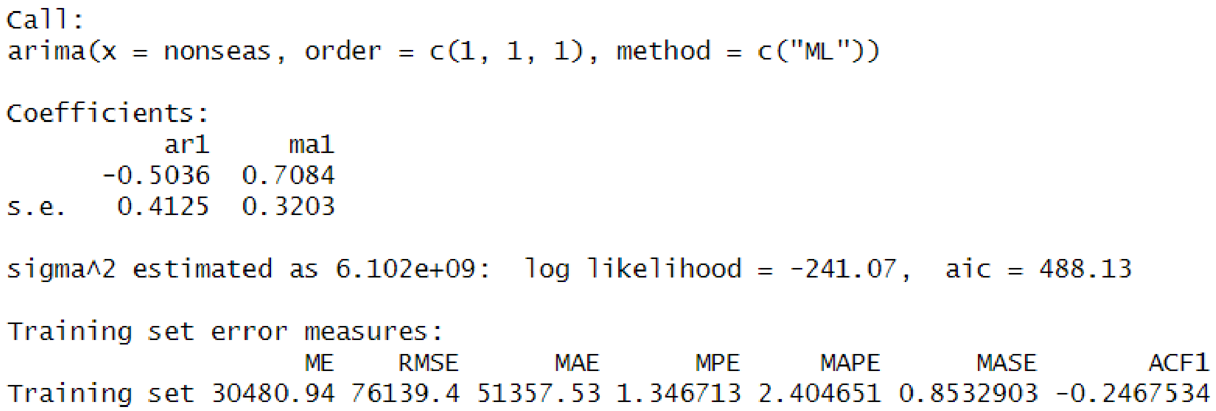
We can notice that expectations tend to increase in general, that is, on average, the price of gold for the coming years is expected to have increased or to be more or less constant.

Other models:

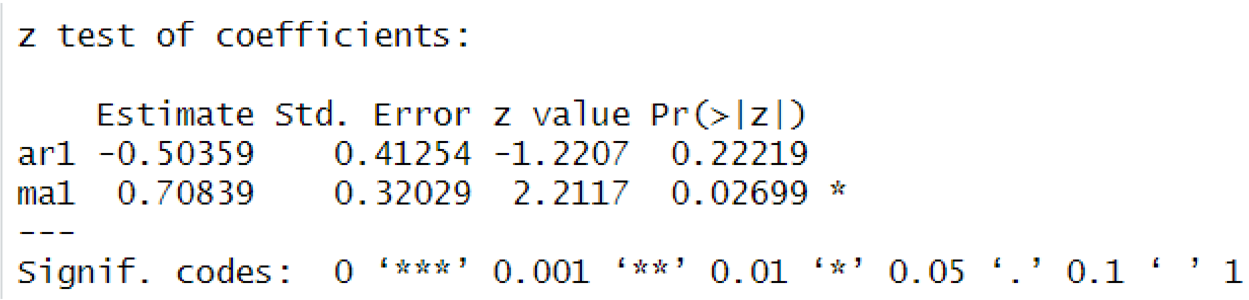
We tried other models with different combinations of (p, d, q) and check if there is another model that achieves the conditions and if the initial model is the best model for reducing the residuals or not.

Trials: Model 2 (1,1,1)

We need to check the two conditions to know if the model can be the other initial model or not.



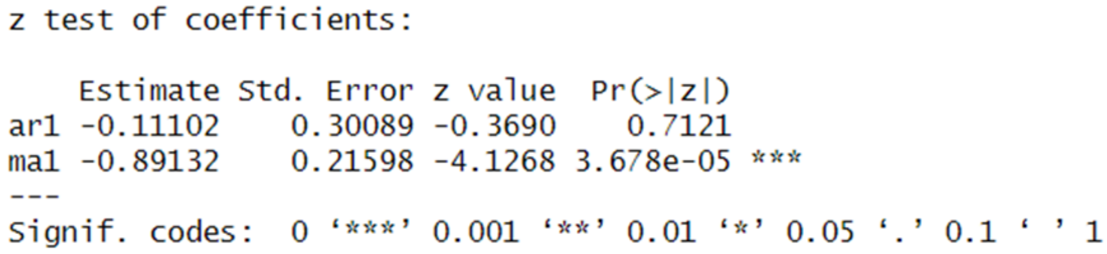
```
Call:
arima(x = nonseas, order = c(1, 1, 1), method = c("ML"))

Coefficients:
          ar1     ma1
      -0.5036  0.7084
s.e.   0.4125  0.3203

sigma^2 estimated as 6.102e+09:  log likelihood = -241.07,  aic = 488.13

Training set error measures:
                   ME    RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 30480.94 76139.4 51357.53 1.346713 2.404651 0.8532903 -0.2467534
```



```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ar1 -0.50359    0.41254 -1.2207  0.22219
ma1  0.70839    0.32029  2.2117  0.02699 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value for all coefficients is more than 0.01, so we will not reject Ho at ∝ =0.01 **“which means that all two coefficients are not statistically significantly different from zero”.**

<u>Trials: Model 3 (1,2,1)</u>



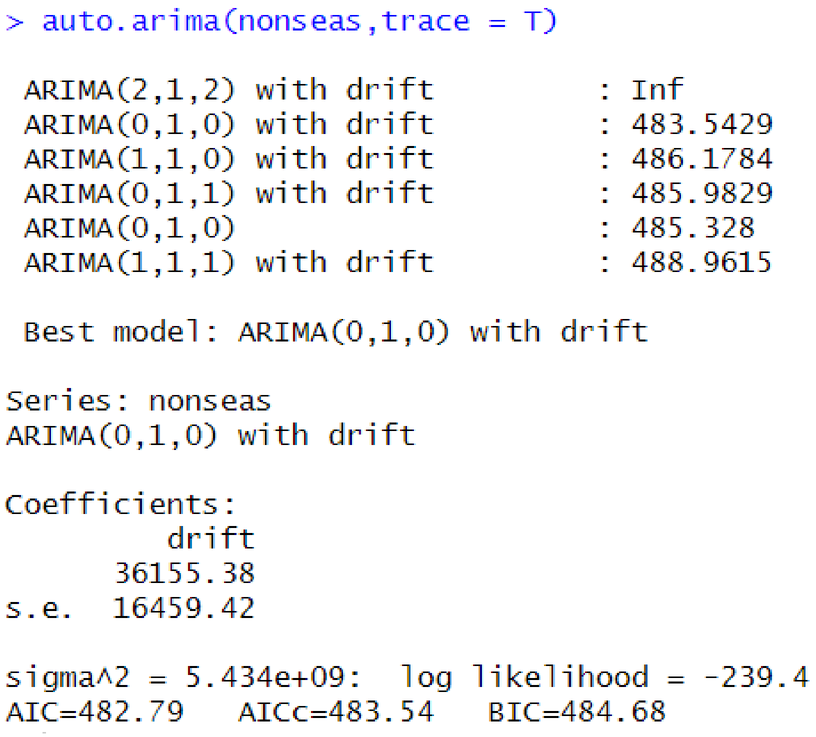
```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.11102    0.30089 -0.3690    0.7121
ma1 -0.89132    0.21598 -4.1268 3.678e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value for all coefficients is more than 0.01, so we will not reject Ho at ∝ =0.01_**“which means that all two coefficients are not statistically significantly different from zero.**

And also tried ARIMA (0,1,1) & ARIMA (2,2,2), both of them have not fulfills the conditions “coefficients must be significant & Residuals must be insignificant”, so, we cannot take them as the initial models, so we will take the initial model ARIMA (0,2,1) as the final model.

Auto-ARIMA:

Auto.arima is a function used for automatically selecting the optimal parameters for an ARIMA time series model. This function searches through different combinations of parameters to find the best-fitting ARIMA model for a given time series dataset. The auto. ARIMA function evaluates models based on criteria like AIC or BIC to determine the best model fit.



```
> auto.arima(nonseas,trace = T)

 ARIMA(2,1,2) with drift         : Inf
 ARIMA(0,1,0) with drift         : 483.5429
 ARIMA(1,1,0) with drift         : 486.1784
 ARIMA(0,1,1) with drift         : 485.9829
 ARIMA(0,1,0)                    : 485.328
 ARIMA(1,1,1) with drift         : 488.9615

 Best model: ARIMA(0,1,0) with drift

Series: nonseas
ARIMA(0,1,0) with drift

Coefficients:
         drift
      36155.38
s.e.  16459.42

sigma^2 = 5.434e+09:  log likelihood = -239.4
AIC=482.79   AICc=483.54   BIC=484.68
```

the function chooses ARIMA (0,1,0) as best model, because the data is white noise, and the mean equal zero.

So we will keep our initial model ARIMA (0,2,1) as the final model

In conclusion

We had US gas monthly consumption and found that our data was seasonality.

So first we take the average and converted it to a yearly data to remove seasonality.

Then we check stationarity and it wasn't stationary in mean so we took the first and second difference; we also noticed that our data is white noise from it's ACF and PACF

So we suggest an initial model ARIMA(0,2,1) model and we another combinations of (p,d,q) and check if there is another model that achieves the conditions and if the initial model is the best model for reducing the residuals or not. And then, we estimate the parameters of the selected model:

$$Yt = 2 * y_{t-1} - y_{t-2} + \varepsilon t + 0.9525 * \varepsilon_{t-1}$$

Then we check diagnostic of the fitted ARIMA (0,2,1) model by using Box-Ljung & Box-Pierce "Tests that checking the residuals", and we conclude that the ARIMA (0,2,1) model has fulfills the conditions

From accuracy measures, we found that the ARIMA (0,2,1) is the optimal model. And we checked the stationarity and invertibility of the model by their conditions and all of them are satisfied.

And finally, we used forecasting to assess the performance of the model against real data and we notice that the predicted values are too close to the actual ones, which means that the consumption of gas will increase.

References

https://www.kaggle.com/datasets/mexwell/usgas-dataset

https://github.com/RamiKrispin/USgas

https://www.eia.gov/

https://ramikrispin.github.io/USgas/