David Aragón 21053
Renatto Guzmán 21646
Data Science - 10

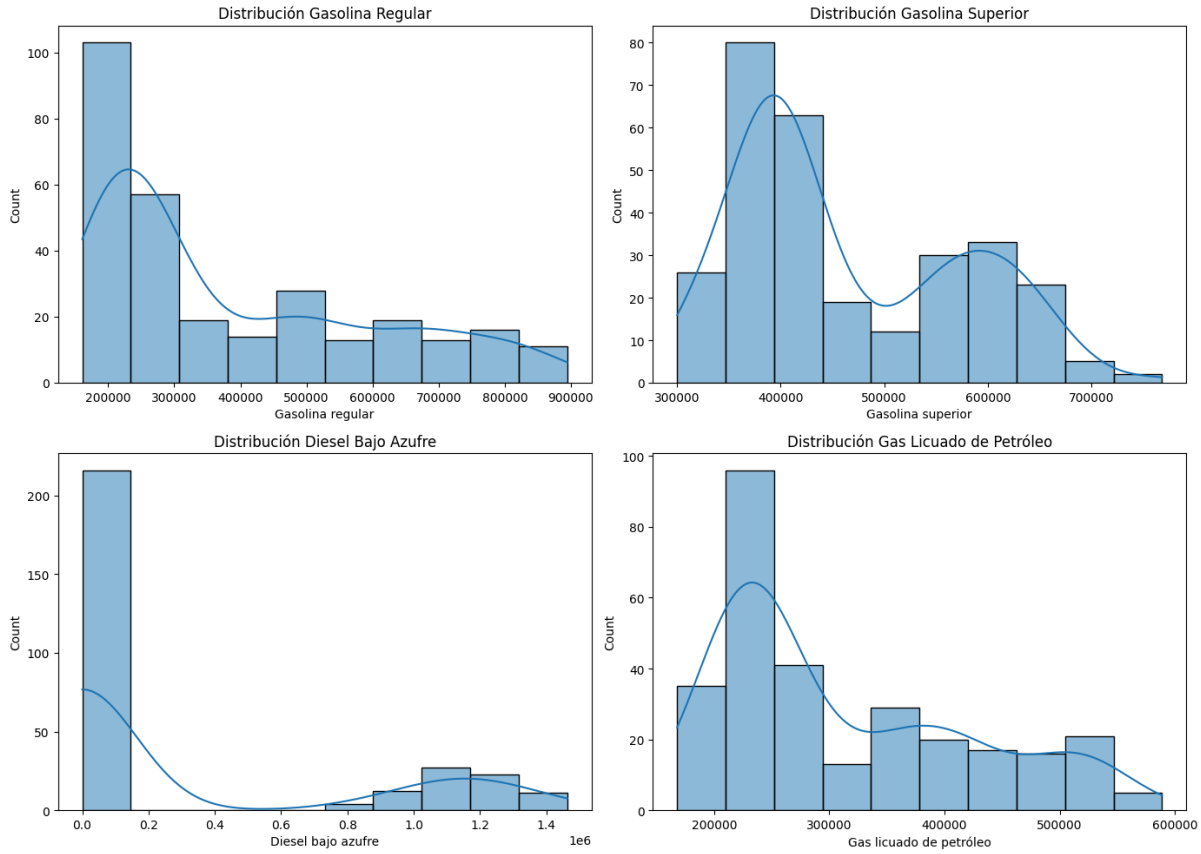
Laboratorio 2

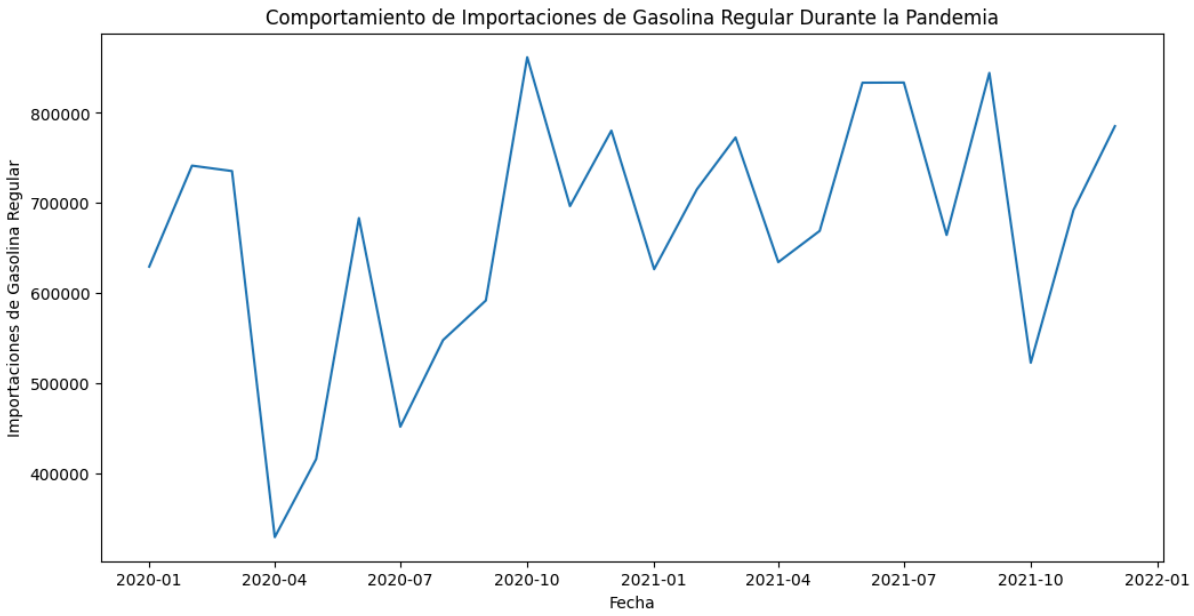
Link al repositorio: https://github.com/AragonD19/Lab2DataScince

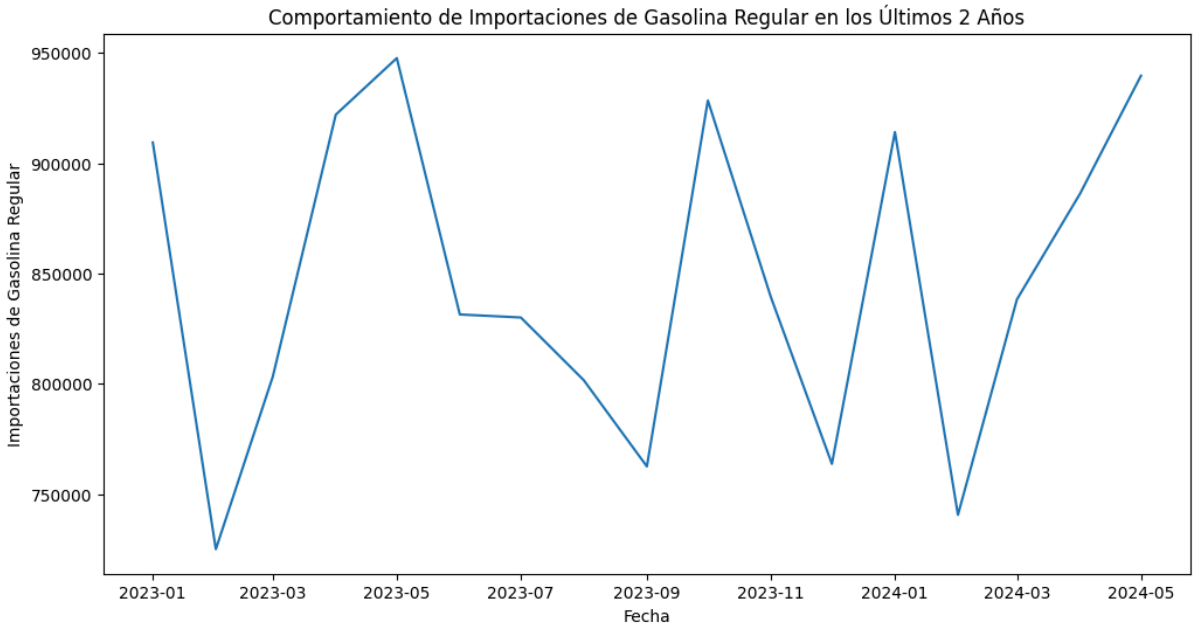
1. Análisis Exploratorio de Datos

Realice un análisis exploratorio de los datos proporcionados, centrándose únicamente en las columnas de gasolinas regular, gasolina superior, diésel y gas licuado de todos los conjuntos. Siga estas sugerencias para su análisis:

- **Exploración de Variables**: Investigue el comportamiento de las variables y determine si están distribuidas normalmente en el caso de las variables continuas.
- **Importaciones**: Analice los meses con mayores importaciones, los picos en importaciones por año por tipo de combustible, el comportamiento en los últimos X años, y el comportamiento durante la pandemia, entre otros aspectos.



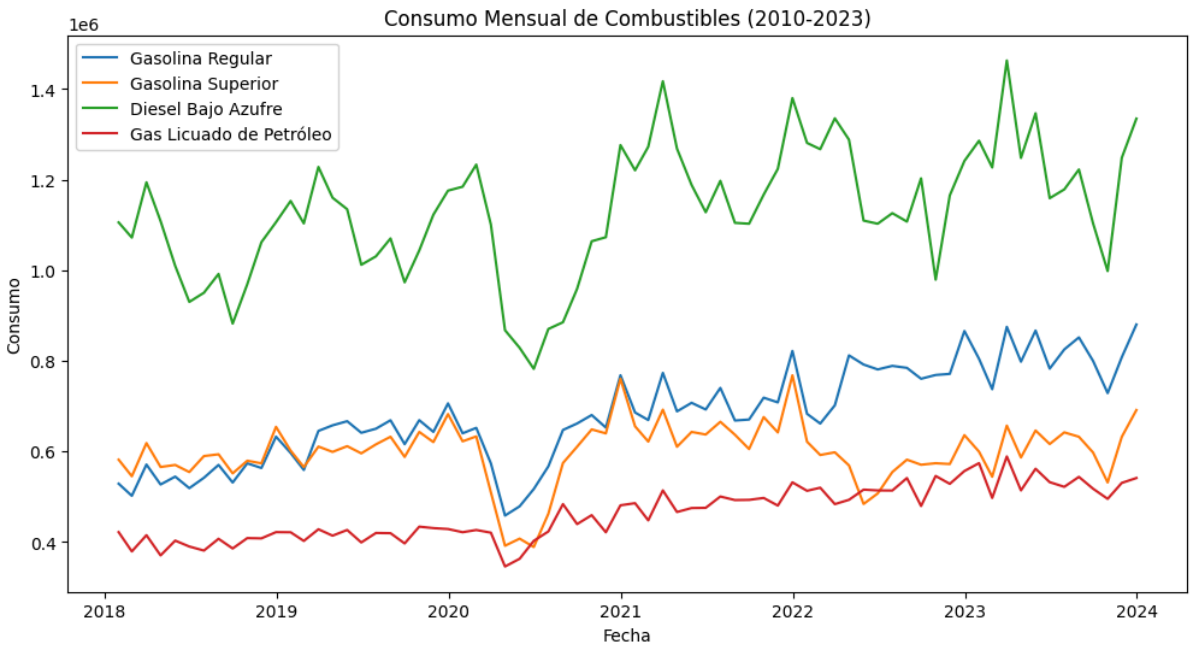




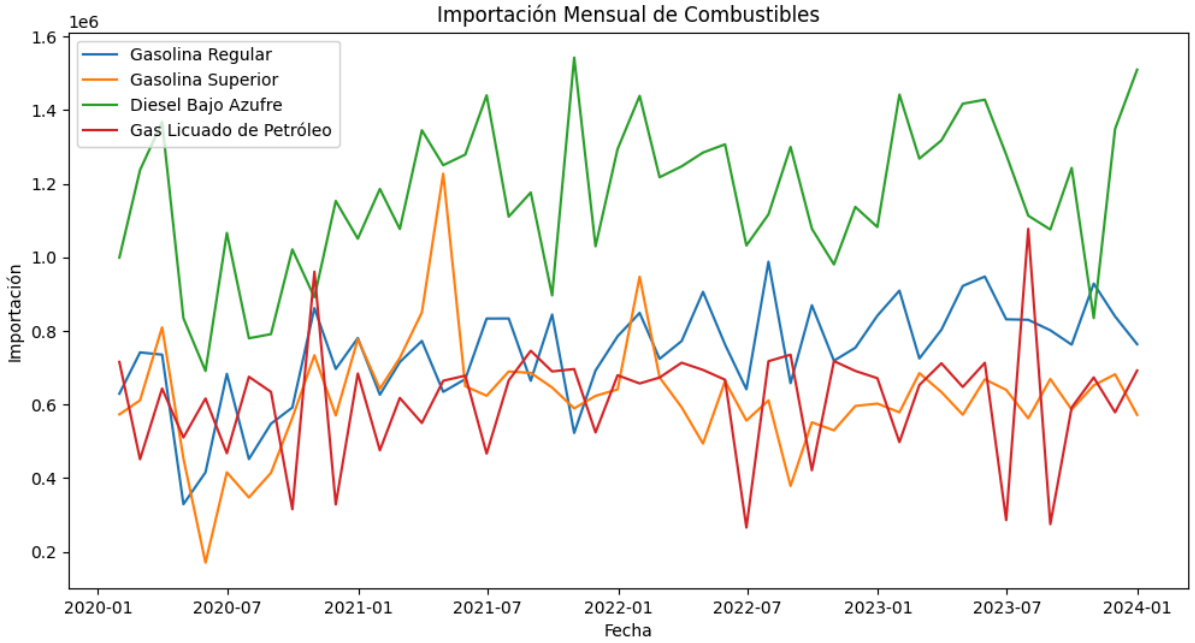
2. Selección y Análisis de Series Temporales

Trabaje con tres de las siguientes series, asegurándose de que al menos una de ellas sea para estimar precios:

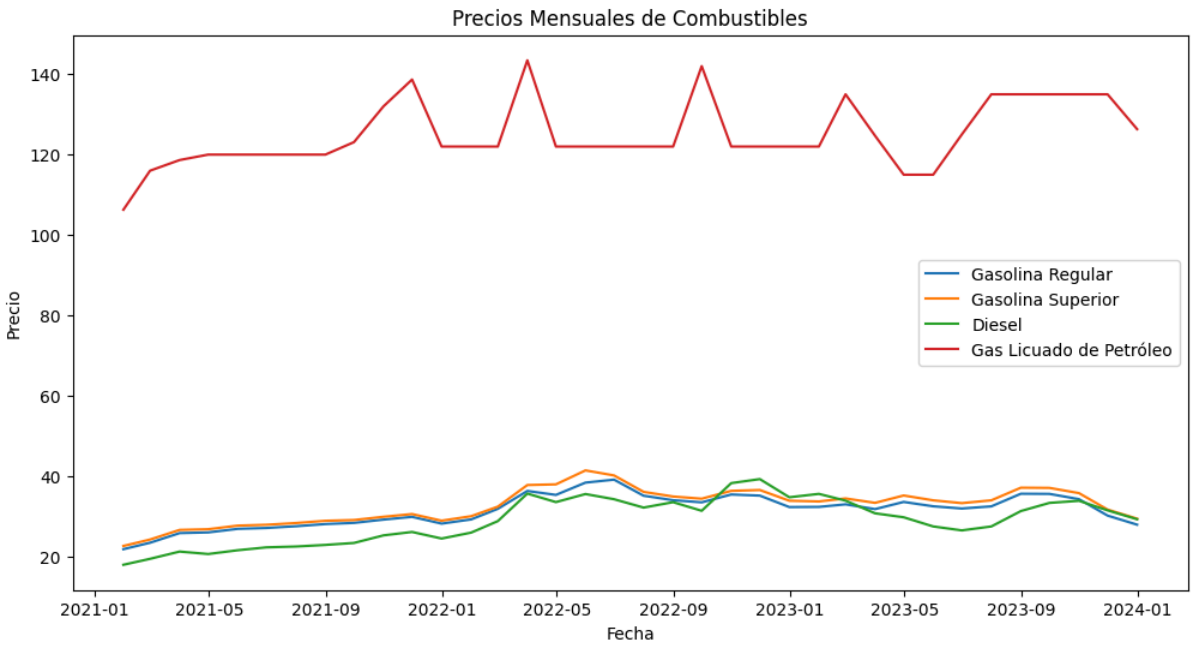
- **a.** Consumo mensual de gasolina superior, gasolina regular, diésel y gas licuado.



- **b.** Importación mensual de gasolina superior, gasolina regular, diésel y gas licuado.



- **c.** Precios en ciudad capital de gasolina superior, gasolina regular, diésel y gas licuado (puede usar precios de autoservicio o de servicio completo).

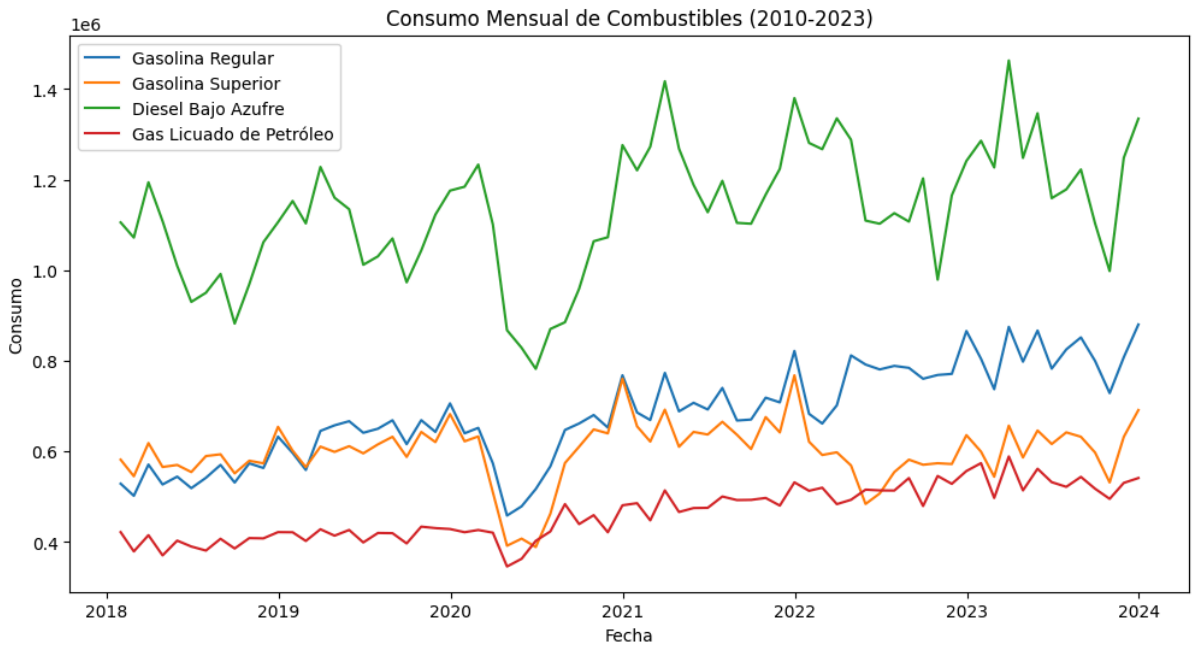


Para cada serie seleccionada:

a. Especifique el inicio, fin y frecuencia de la serie.

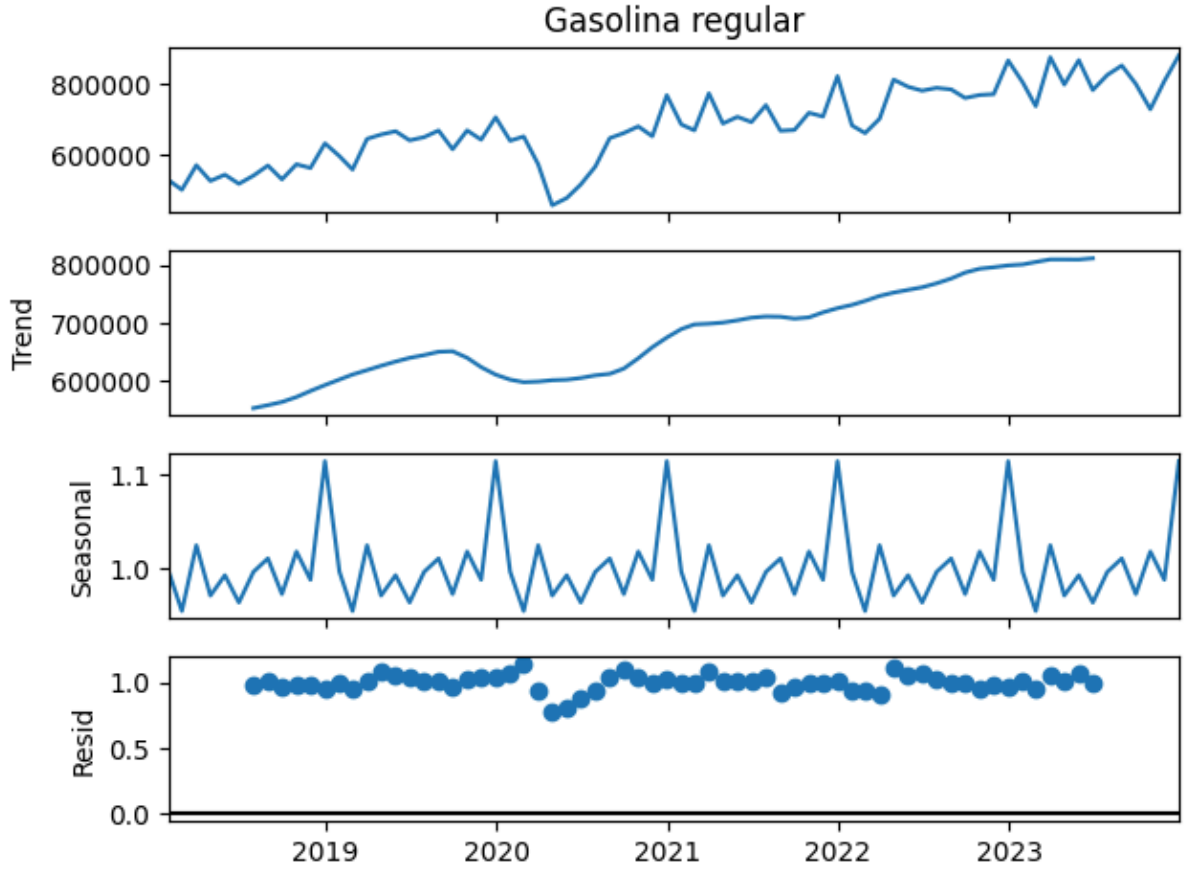
```
            Gasolina regular  Gasolina superior  Diesel bajo azufre  \
Fecha
2018-01-31         528763.30          581802.98         1105628.890
2018-02-28         502041.67          545210.00         1072071.670
2018-03-31         571052.13          618402.18         1194136.180
2018-04-30         526979.67          565665.72         1108452.970
2018-05-31         544395.17          570129.66         1009750.145

            Gas licuado de petróleo
Fecha
2018-01-31           421848.920476
2018-02-28           379239.067143
2018-03-31           415125.896429
2018-04-30           370434.605238
2018-05-31           403094.744048
            Gasolina regular  Gasolina superior  Diesel bajo azufre  \
Fecha
2023-08-31         851754.02          632333.40          1222453.19
2023-09-30         799852.73          597670.26          1103622.25
2023-10-31         728741.16          531542.58           998113.39
2023-11-30         807947.98          632180.03          1248223.75
2023-12-31         880233.12          691389.62          1334692.15

            Gas licuado de petróleo
Fecha
2023-08-31               543825.64
2023-09-30               517699.48
2023-10-31               495261.14
2023-11-30               530542.42
2023-12-31               541263.18
Consumo Mensual - Inicio: 2018-01-31 00:00:00, Fin: 2023-12-31 00:00:00, Frecuencia: ME

```

b. Realice un gráfico de la serie y explique qué información puede obtener a primera vista.



Como se puede ver el consumo tiene muchas subidas y bajadas, siendo la bajada más pronunciada en las fechas de la pandemia.

c. Descomponga la serie. Con base en el diagrama de la serie y sus componentes, discuta si es posible hablar de estacionariedad en media y en varianza.



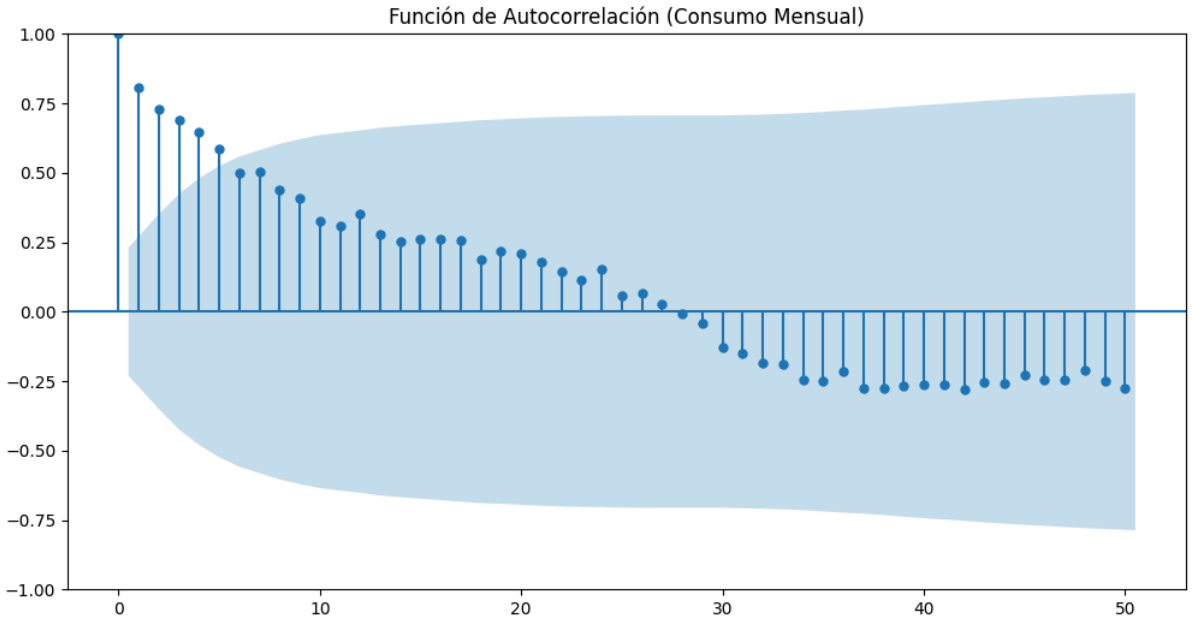
d. Determine si es necesario transformar la serie y explique el motivo.

```
ADF Statistic: -1.7049252959839365
p-value: 0.4285766195937699
ADF Statistic (Importación): -4.894919100205006
p-value (Importación): 3.5677293323226274e-05
ADF Statistic (Precios): -2.3130474869001283
p-value (Precios): 0.16779442124591654
```

```
Opción A: Consumo Mensual

ADF Statistic: -1.7049252959839365

p-value: 0.4285766195937699

La serie de consumo mensual no es estacionaria, ya que el p-valor es mayor que 0.05.
```

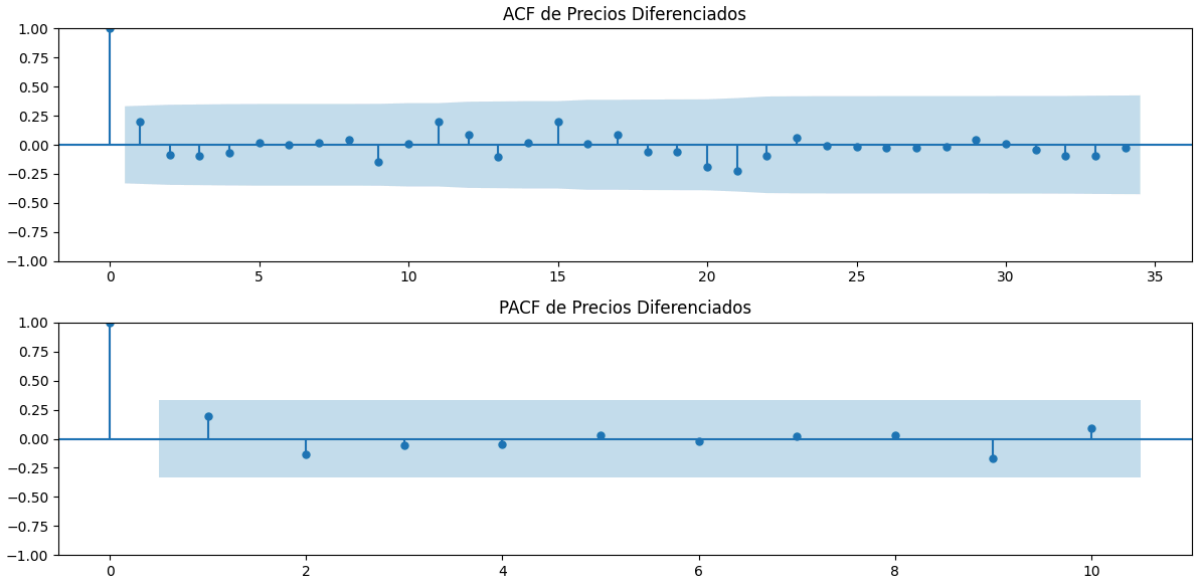
e. Explique la no estacionariedad en media. Para ello:

- **i.** Realice un gráfico de autocorrelación y utilícelo para explicar la no estacionariedad en media.
- **ii.** Basado en los valores de estadísticos como la prueba de Dickey-Fuller Aumentada, confirme la no estacionariedad en media. ¿Qué se necesita hacer para lograr la estacionariedad en media si no se cumple?



```
ADF Statistic (Diff Consumo): -11.9110185750175
p-value (Diff Consumo): 5.296569752096612e-22
ADF Statistic (Diff Precios): -4.489371301300771
p-value (Diff Precios): 0.00020552088002325476
```

f. Una vez analizada la serie, elija los parámetros ppp, qqq y ddd del modelo ARMA o ARIMA para predecir. Explique cómo determinó estos valores basándose en las funciones de autocorrelación y autocorrelación parcial. Si usa la función `auto.arima` de R, determine y explique si el modelo propuesto tiene sentido.



g. Realice varios modelos ARIMA y determine cuál es el mejor para estimar los datos de la serie. Para ello, analice los residuos y las métricas AIC y/o BIC.

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=150.870, Time=0.09 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=145.042, Time=0.01 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=145.456, Time=0.01 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=145.219, Time=0.02 sec
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=147.192, Time=0.06 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=146.736, Time=0.01 sec

Best model:  ARIMA(0,0,0)(0,0,0)[0]
Total fit time: 0.248 seconds
Modelo sugerido por auto_arima:                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   35
Model:                        SARIMAX   Log Likelihood                -71.521
Date:                Sun, 04 Aug 2024   AIC                            145.042
Time:                        19:15:33   BIC                            146.597
Sample:                    02-28-2021   HQIC                           145.579
                         - 12-31-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         3.4870      0.806      4.325      0.000       1.907       5.067
===================================================================================
Ljung-Box (L1) (Q):                   1.50   Jarque-Bera (JB):                 0.27
Prob(Q):                              0.22   Prob(JB):                         0.87
Heteroskedasticity (H):               2.72   Skew:                            -0.19
Prob(H) (two-sided):                  0.10   Kurtosis:                         3.21
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
ARIMA(1,1,1) - AIC: 145.90025474531092, BIC: 152.00569684377555
ARIMA(1,1,2) - AIC: 147.60936361595142, BIC: 155.24116623903222
ARIMA(2,1,1) - AIC: 147.2503616795875, BIC: 154.88216430266831
ARIMA(2,1,2) - AIC: 149.36095146051383, BIC: 158.5191146082108
```

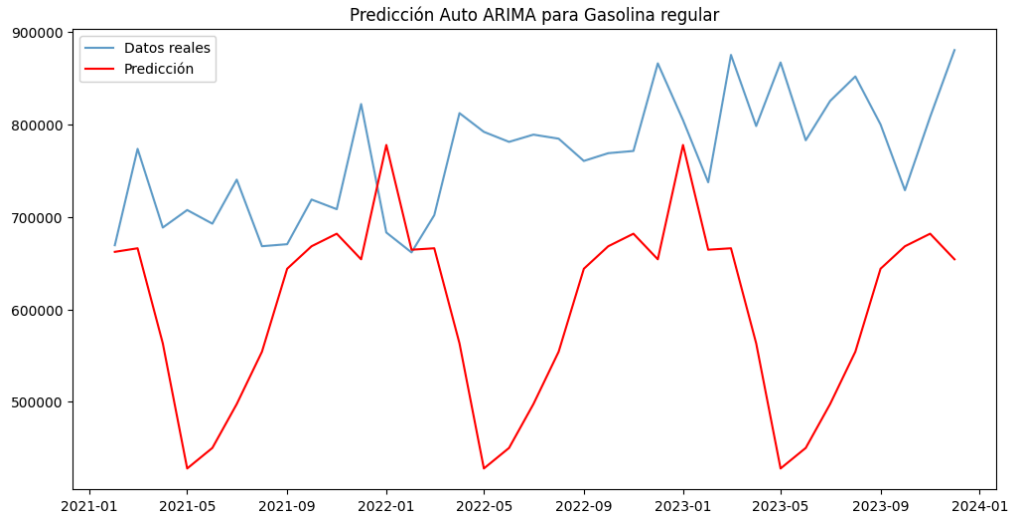
h. Elabore un modelo utilizando el algoritmo Prophet de Facebook o redes neuronales. Compare este modelo con los modelos del inciso anterior. ¿Cuál funcionó mejor?

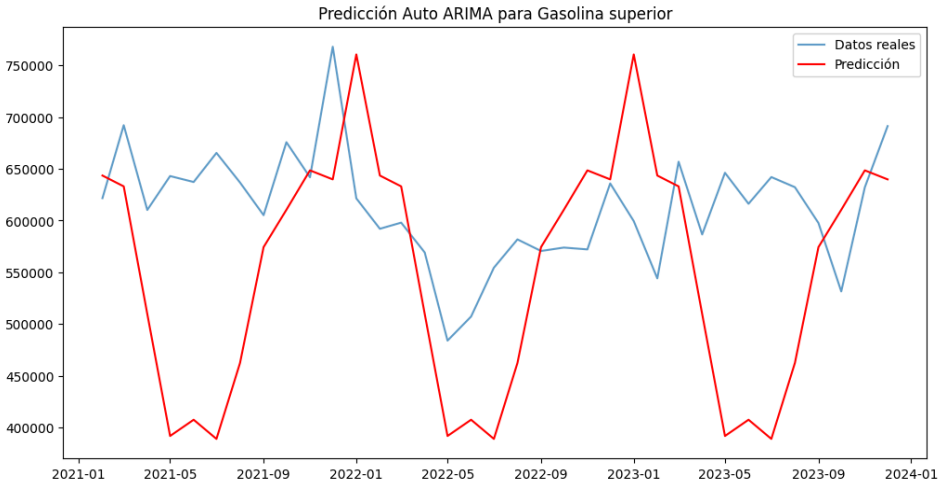
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   36
Model:               SARIMAX(0, 1, 0)   Log Likelihood                -71.521
Date:                Sun, 04 Aug 2024   AIC                            145.042
Time:                        19:20:35   BIC                            146.597
Sample:                  01-31-2021   HQIC                           145.579
                       - 12-31-2023
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         3.4869      0.806      4.325      0.000       1.907       5.067
===================================================================================
Ljung-Box (L1) (Q):                   1.50   Jarque-Bera (JB):                 0.27
Prob(Q):                              0.22   Prob(JB):                         0.87
Heteroskedasticity (H):               2.72   Skew:                            -0.19
Prob(H) (two-sided):                  0.10   Kurtosis:                         3.21
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

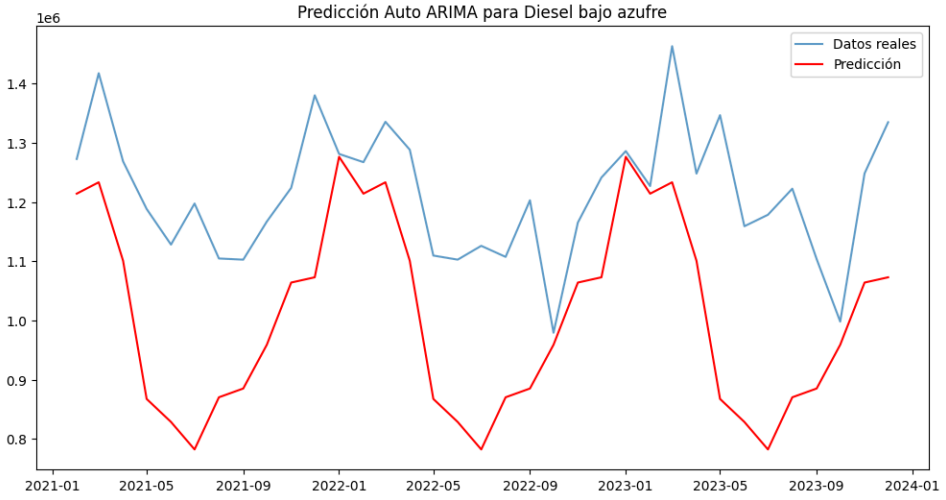
3. Predicción y Evaluación

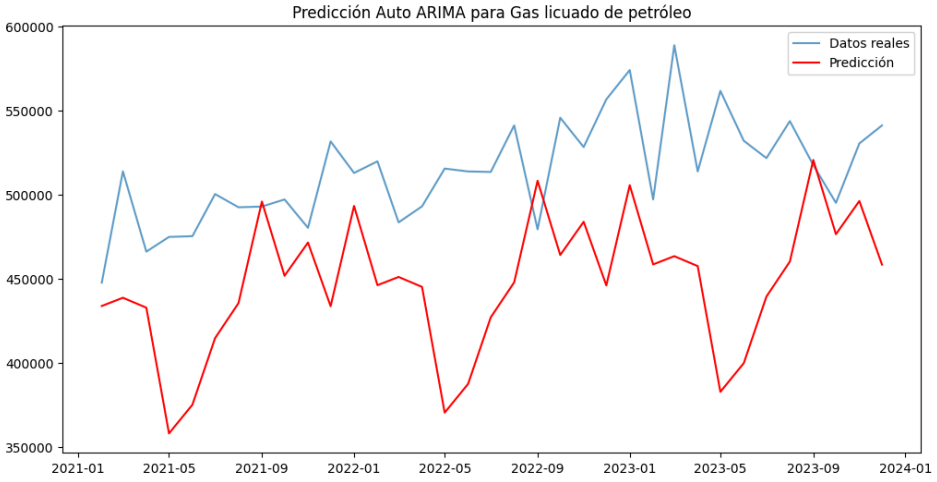
a. Realice una predicción de los valores de la serie para los últimos 3 años. Explique la calidad de su predicción. En el caso de la serie de precios, use los datos de 2024 como conjunto de prueba.

Se usó la serie de consumo de combustibles



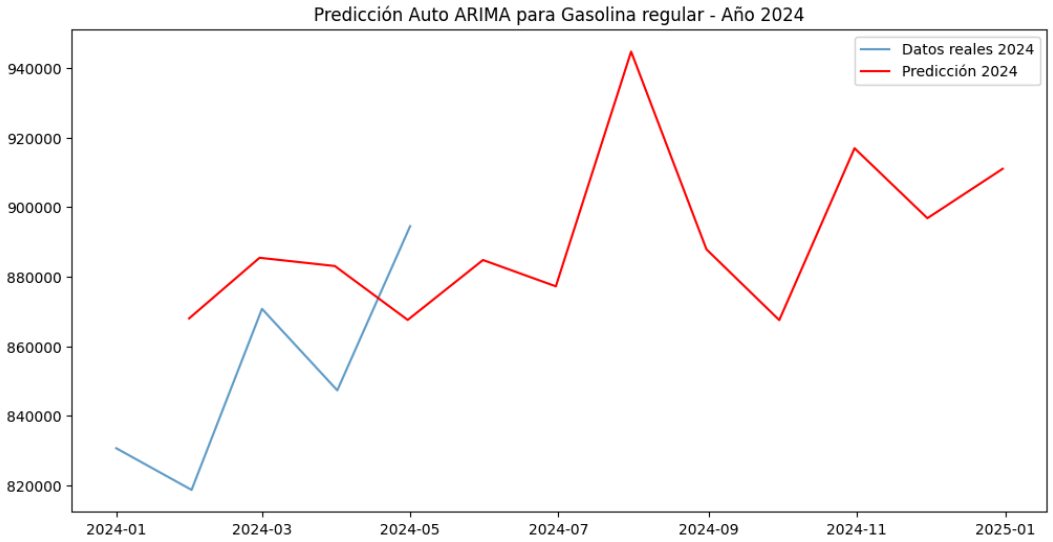


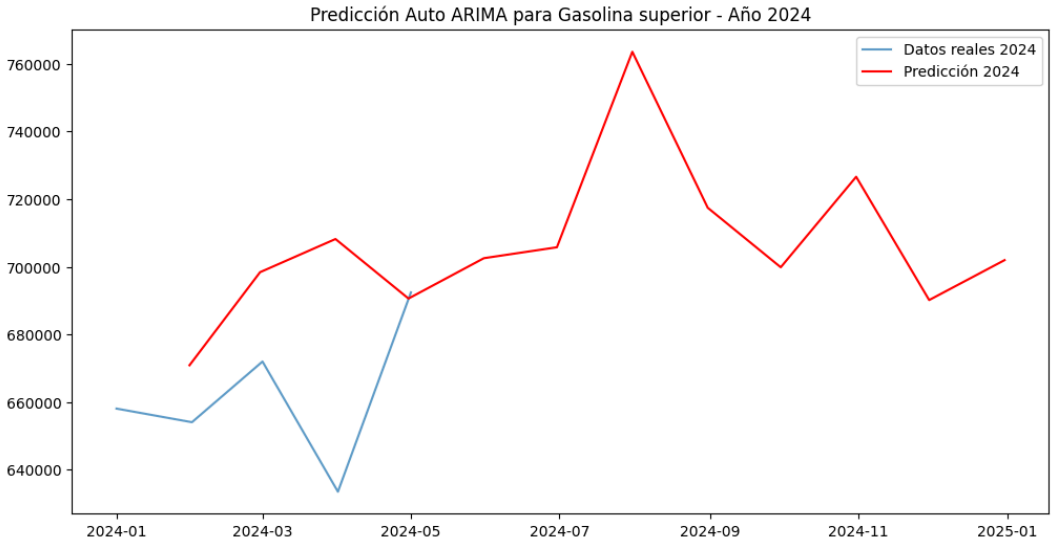


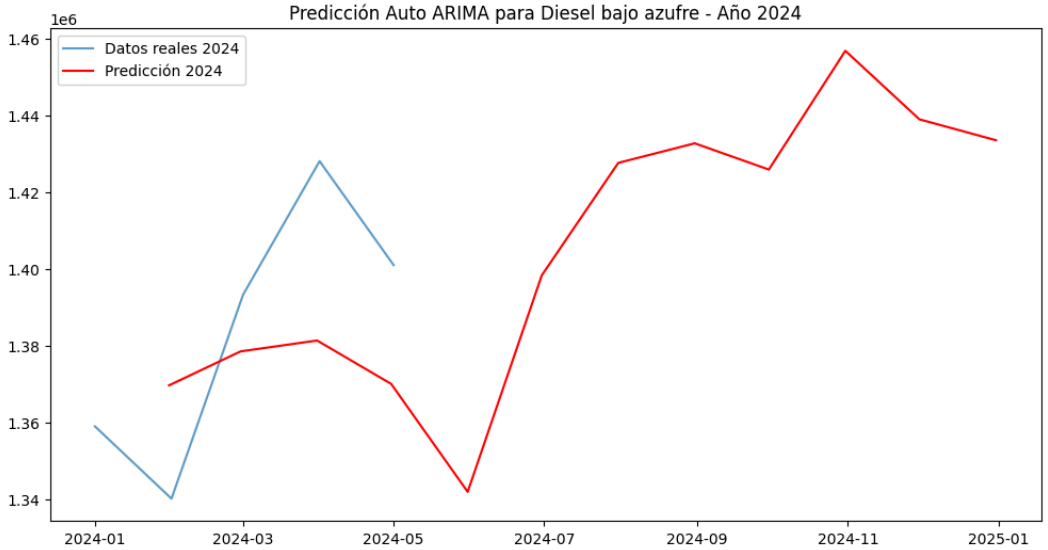


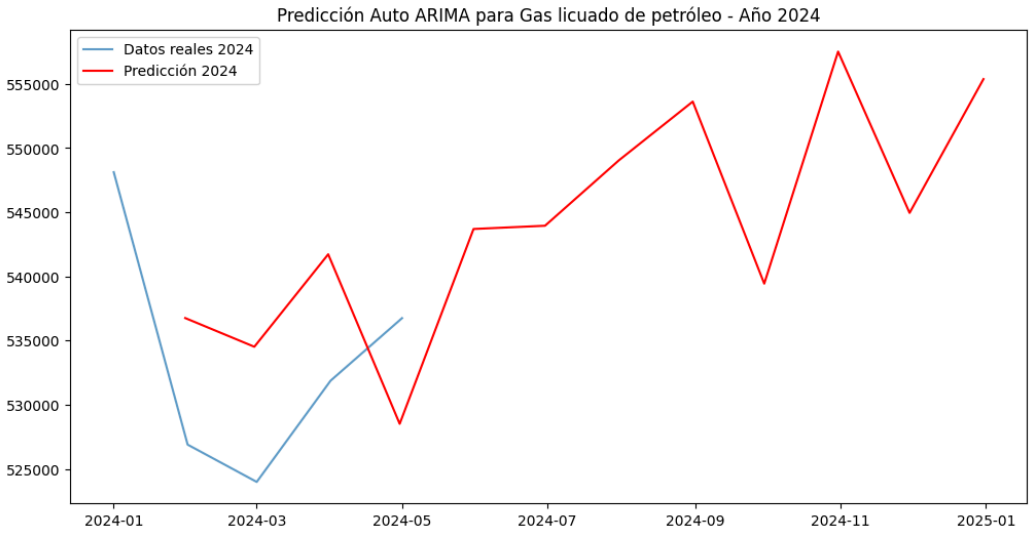
En general, la calidad de las predicciones es razonable, pero se pueden notar que parecen capturar bien los patrones estacionales, especialmente visible en el caso del diesel y el gas licuado de petróleo. También cabe notar que las predicciones parecen seguir las tendencias generales a largo plazo, pero tienen dificultades para capturar cambios repentinos o recuperaciones rápidas.

b. Realice una predicción para todo el año 2024 y determine qué tan ajustada está a la realidad, considerando los meses anteriores del año actual.









La predicción para el gas licuado de petróleo parece ser la más precisa de todas. La línea de predicción sigue muy de cerca la tendencia histórica y se alinea bien con los datos reales de 2024.

En general para las demás predicciones se puede ver que siguen de cerca la tendencia de los datos históricos y parece alinearse mejor con los datos reales de 2024. Sin embargo, hay una ligera subestimación en los primeros meses de 2024.

c. Discuta el comportamiento de las series durante la pandemia. ¿Qué impacto tuvo en las predicciones? ¿Se recuperaron los valores habituales en el regreso a la "nueva normalidad"?

Se observa una clara caída en el consumo de todos los combustibles durante el período de la pandemia (alrededor de 2020). Esta caída fue especialmente pronunciada para la gasolina regular y superior, probablemente debido a las restricciones de movilidad.

La pandemia creó una anomalía significativa en los datos históricos, lo que probablemente dificultó la capacidad de los modelos para hacer predicciones precisas. Esto podría explicar por qué algunas predicciones para 2024 parecen subestimar el consumo real, ya que los modelos pueden estar dando demasiado peso a los datos del período de la pandemia.

Basándonos en los datos de 2024, parece que el consumo de combustibles está volviendo a niveles pre-pandemia o incluso superándolos en algunos casos. Esto sugiere una recuperación hacia una "nueva normalidad", aunque los patrones de consumo pueden haber cambiado ligeramente en comparación con la era pre-pandemia.