MA 641A TIME SERIES ANALYSIS – 1

PROJECT REPORT

- Shreyas Reddy

Non-Seasonal Dataset : Deutsche Bank closing Prices

1. Introduction:

Financial time series like stock prices often exhibit complex patterns involving trends, volatility, and randomness. Accurate modelling of such behavior is crucial for investment forecasting, risk analysis, and financial planning.

This part of the project focuses on analyzing and forecasting the daily closing prices of Deutsche Bank (DBK.DE) listed on the Frankfurt Stock Exchange. The dataset spans from January 2010 to January 2023, with a specific focus on the most recent two years (2021–2023) for modeling purposes.

Unlike classical seasonal data, stock price movements typically do not follow a fixed seasonal structure but may exhibit trend, autocorrelation, and volatility clustering. This makes them ideal candidates for non-seasonal ARIMA modeling, complemented by GARCH analysis to capture time-varying volatility in residuals.

The primary objective of this non-seasonal time series analysis is to :

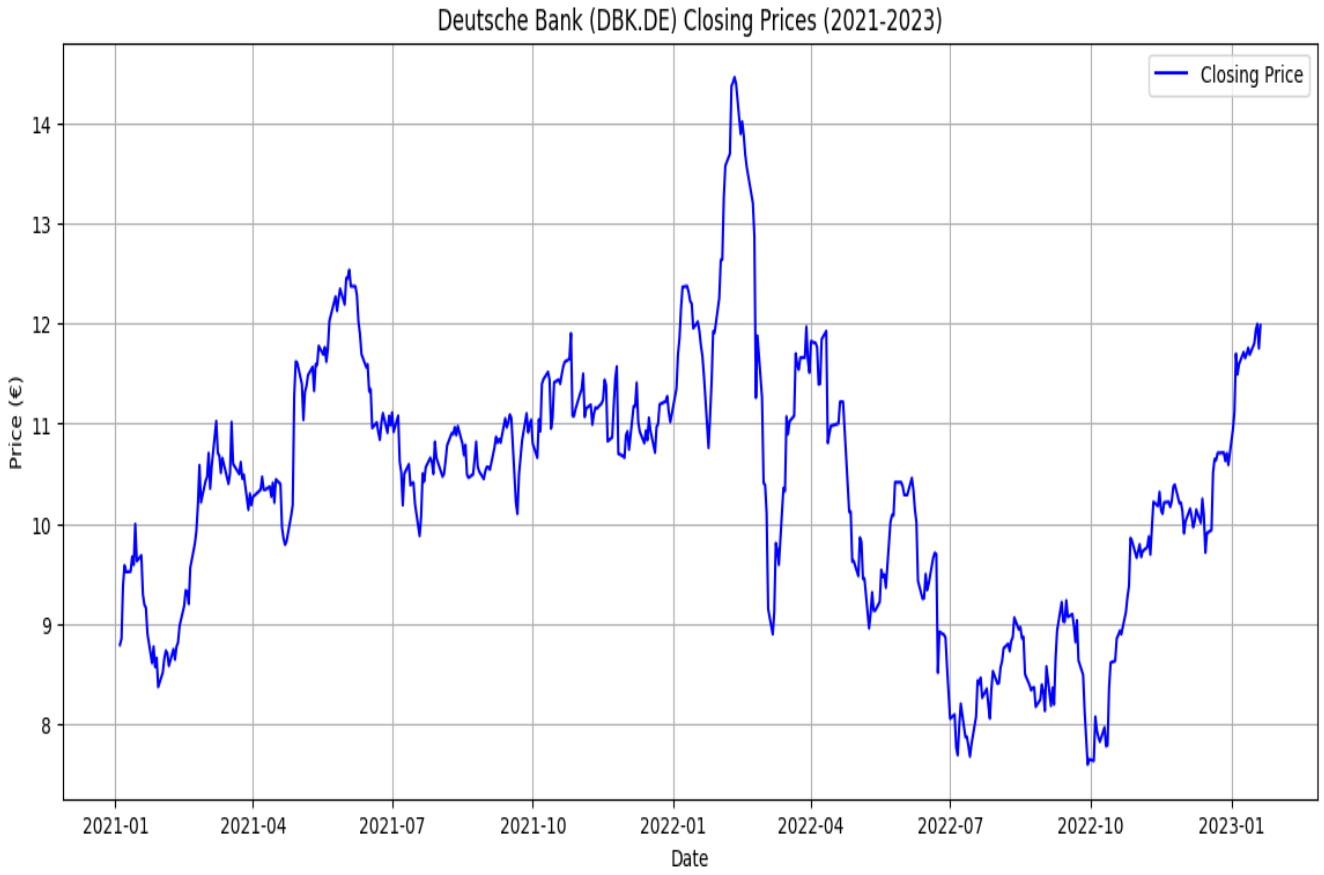
- Identify a parsimonious ARIMA model using the Box-Jenkins methodology.
- Ensure stationarity and validate model assumptions through residual diagnostics.
- Use the best-fit model to forecast future price movements.
- Perform volatility modeling using GARCH (1,1) to account for heteroskedasticity.

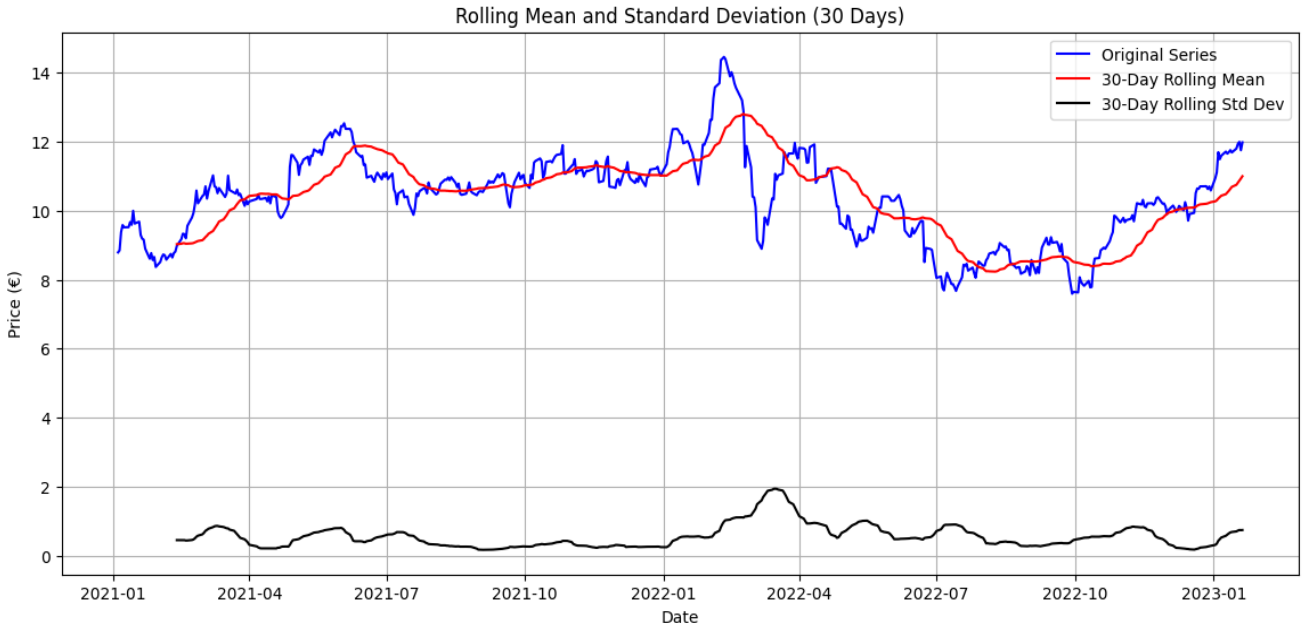
This study demonstrates how statistical tools can be leveraged to gain insights into real-world financial behavior, enabling better-informed decision-making in uncertain market environments.

2. Dataset Description and Preprocessing:

2.1 Dataset Overview

The dataset used for this analysis contains daily historical stock prices of Deutsche Bank (Ticker: DBK.DE) traded on the Frankfurt Stock Exchange. It was sourced from Kaggle and includes standard financial variables such as opening price, closing price, highs and lows, volume, and adjusted closing price. The dataset spans the period from January 4, 2010, to January 20, 2023, and consists of over 3,300 trading-day observations. For focused modeling and practical forecasting, a subset covering the most recent two years (January 2021 to January 2023) was extracted. This subset is more relevant for short-term prediction and aligns with the Box-Jenkins methodology, which assumes the underlying time series is stationary or made stationary through transformation.


Deutsche Bank (DBK.DE) Closing Prices (2021-2023)
Closing Price
Price (€)
Date
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10
2023-01


Rolling Mean and Standard Deviation (30 Days)
Original Series
30-Day Rolling Mean
30-Day Rolling Std Dev
Price (€)
Date
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10
2023-01

2.2 Variables Used

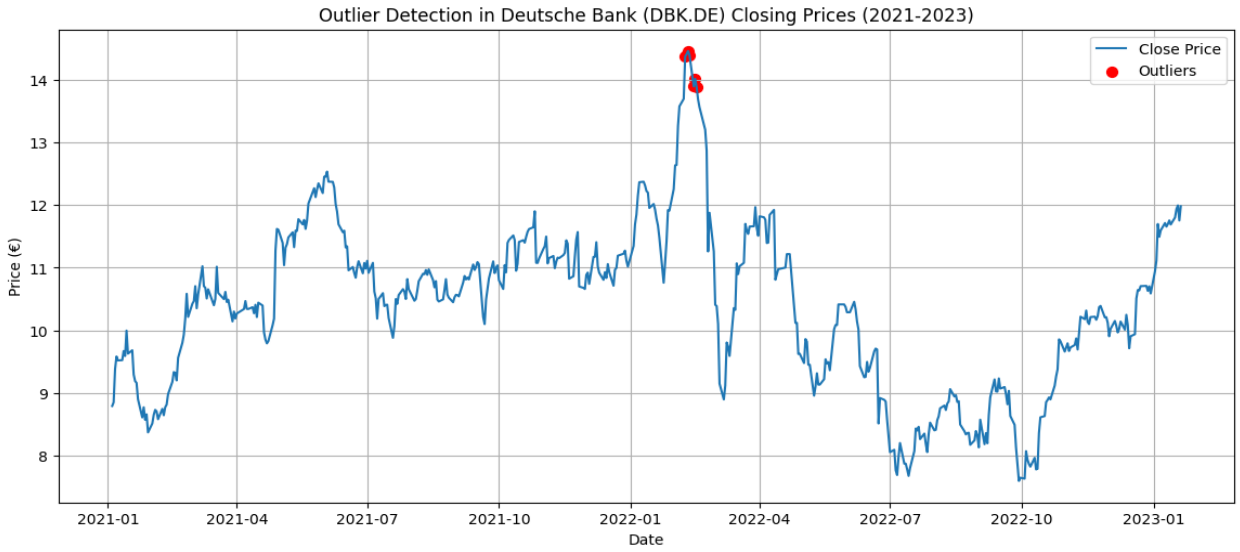
Column Name	Description
Date	Trading day (converted to datetime index)
Open	Opening price on that day
High	Highest price during the day
Low	Lowest Price during the day
Close	Closing Price of the stock
Adj Close	Closing Price adjusted for Splits/Dividends
Volume	Number of shared traded

For the purpose of this project, Close price was used as the primary variable for modeling and forecasting.

2.3 Preprocessing Steps

- Datetime Indexing: The Date column was converted to a proper datetime object and set as the index to ensure time series compatibility.

- Subset of Dataset: A slice of the data from 2021-01-04 to 2023-01-20 was taken, resulting in approximately 527 observations (business days only).
- Sorting: The data was sorted chronologically to preserve temporal order.
- Missing Values: No missing values were found in the selected subset.
- Outlier Detection: A statistical outlier check on the Close price using the IQR method identified 7 mild outliers, which were retained as they reflect genuine market behavior.



3. Making the Dataset Stationary:

Financial time series data like stock prices are typically non-stationary, meaning their statistical properties such as mean and variance change over time. Stationarity is a key requirement for ARIMA modeling, as it allows consistent model fitting and forecasting.

To ensure stationarity, the Deutsche Bank closing prices (2021–2023) were first log-transformed to stabilize the variance. This is helpful in compressing the scale and reducing heteroskedasticity in the raw data.

After applying the logarithmic transformation, the dataset still showed signs of non-stationarity. Therefore, the first difference of the log-transformed series was computed, which removes the trend and helps in achieving stationarity. This differencing approach transforms the data into the series of returns (i.e., percentage price changes).

The figures below illustrate the log-transformed series and its first difference:

Figure 1: Log Transformed Closing Prices of DBK.DE (2021–2023)

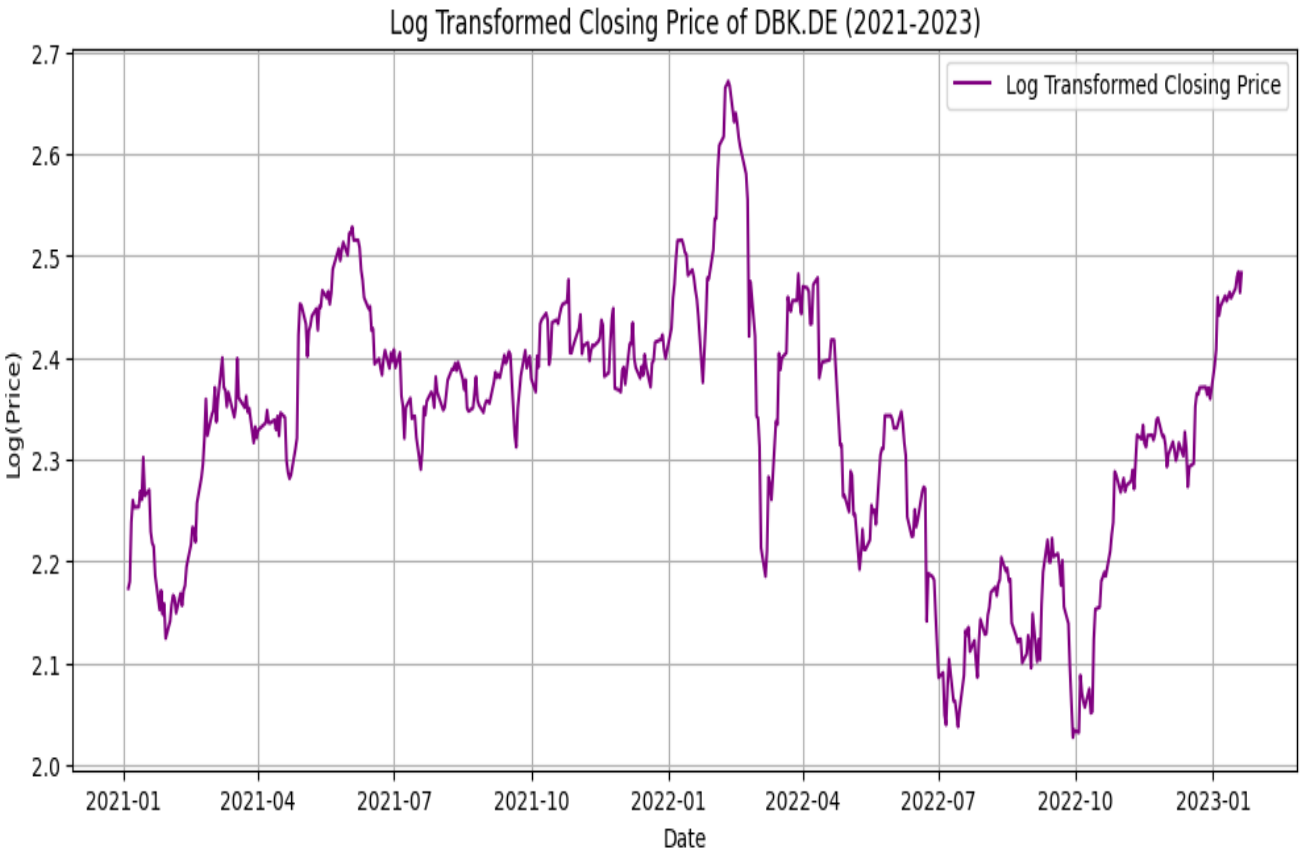
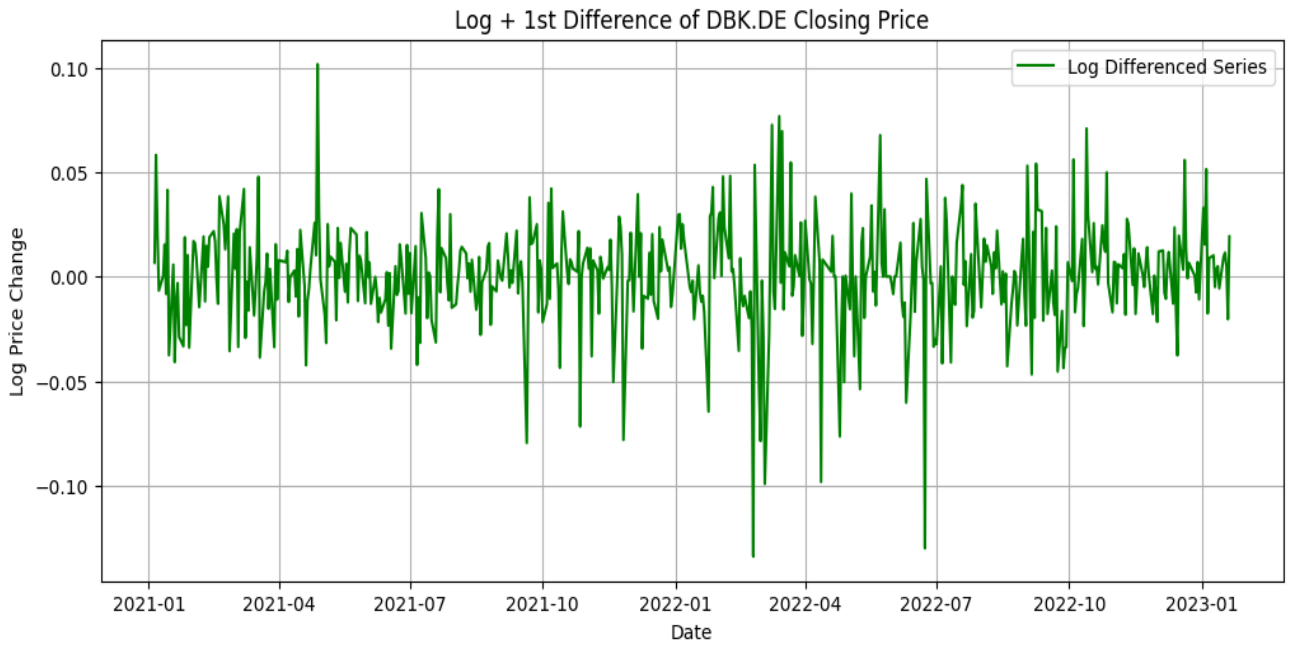


Figure 2: Log Differenced Closing Prices of DBK.DE (2021–2023)

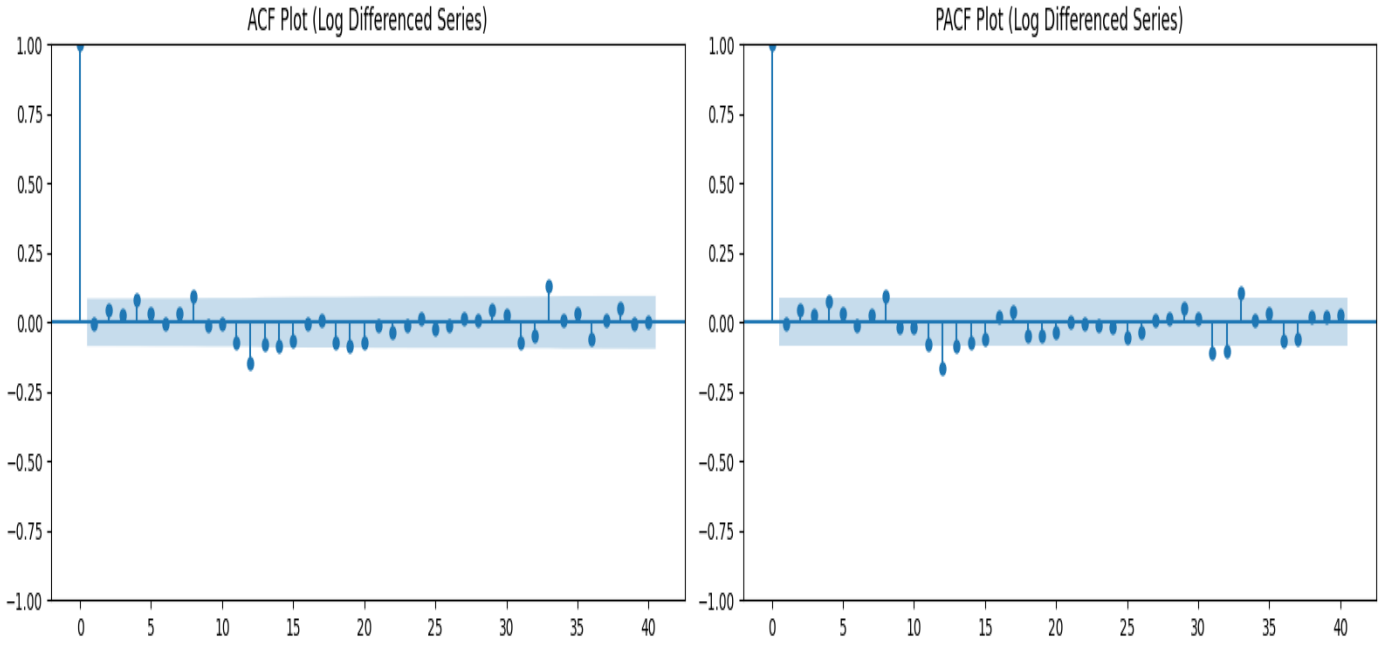


Augmented Dickey-Fuller tests were conducted before and after transformation. The original closing prices had a high p-value (0.2397), indicating non-stationarity. However, after log transformation and differencing, the p-value significantly dropped (1.29×10^{-11}), leading us to reject the null hypothesis of non-stationarity. Hence, the final transformed series used for ARIMA modeling is the log-differenced Deutsche Bank closing price series.

4. Model Identification and Selection:

After ensuring stationarity through log transformation and first-order differencing, the next step involved identifying suitable ARIMA models for the transformed series. To do this, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were examined. The ACF plot showed a rapid drop after lag 1, indicating the presence of a Moving Average (MA) component, possibly of order 1. Similarly, the PACF plot also dropped significantly after lag 1, suggesting the presence of an Autoregressive (AR) component, likely of order 1. Based on these diagnostics, candidate ARIMA models such as ARIMA(1,1,1), ARIMA(2,1,2), ARIMA(1,1,0), and ARIMA(0,1,1) were selected and fitted for comparison.

ACF and PACF Plot of Log Differenced Series:



Each model was evaluated using information criteria such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and HQIC (Hannan-Quinn Information Criterion), along with statistical significance of the estimated parameters. The ARIMA(2,1,2) model had high p-values and indicated overfitting due to parameter redundancy. ARIMA (1,1,1) also showed weak significance for both AR and MA terms. In contrast, ARIMA (1,1,0) yielded the lowest AIC value and had a simple structure with no redundant terms. ARIMA (0,1,1) performed comparably well but had slightly higher AIC than ARIMA (1,1,0). Therefore, ARIMA (1,1,0) was selected as the final model due to its optimal balance between goodness-of-fit and parsimony. ARIMA (0,1,1) was retained as a backup model for validation or alternative forecasting purposes.

Observations and Diagnostics for candidates for candidate ARIMA Models

```
Fitting ARIMA(1,1,1):
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Close   No. Observations:                  527
Model:                 ARIMA(1, 1, 1)   Log Likelihood                1189.597
Date:                Mon, 05 May 2025   AIC                          -2373.195
Time:                        03:23:45   BIC                          -2360.399
Sample:                             0   HQIC                         -2368.185
                                  - 527
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4511      1.305      0.346      0.730      -2.106       3.008
ma.L1         -0.4310      1.325     -0.325      0.745      -3.027       2.165
sigma2         0.0006   2.27e-05     28.012      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                   0.39   Jarque-Bera (JB):               425.05
Prob(Q):                              0.53   Prob(JB):                         0.00
Heteroskedasticity (H):               1.43   Skew:                            -0.67
Prob(H) (two-sided):                  0.02   Kurtosis:                         7.19
===================================================================================
```

Fitting ARIMA(2,1,2):

SARIMAX Results

Dep. Variable:	Close	No. Observations:	527
Model:	ARIMA(2, 1, 2)	Log Likelihood	1191.254
Date:	Mon, 05 May 2025	AIC	-2372.508
Time:	03:23:49	BIC	-2351.181
Sample:	0	HQIC	-2364.157
	- 527		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9824	0.505	1.945	0.052	-0.007	1.972
ar.L2	-0.2578	0.468	-0.551	0.581	-1.174	0.658
ma.L1	-1.0156	0.492	-2.065	0.039	-1.979	-0.052
ma.L2	0.3325	0.447	0.744	0.457	-0.543	1.208
sigma2	0.0006	2.16e-05	28.605	0.000	0.001	0.001

Ljung-Box (L1) (Q):	0.23	Jarque-Bera (JB):	425.95
Prob(Q):	0.63	Prob(JB):	0.00
Heteroskedasticity (H):	1.43	Skew:	-0.64
Prob(H) (two-sided):	0.02	Kurtosis:	7.22

Fitting ARIMA(1,1,0):

SARIMAX Results

Dep. Variable:	Close	No. Observations:	527
Model:	ARIMA(1, 1, 0)	Log Likelihood	1189.459
Date:	Mon, 05 May 2025	AIC	-2374.917
Time:	03:23:49	BIC	-2366.387
Sample:	0	HQIC	-2371.577
	- 527		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0052	0.040	-0.130	0.897	-0.084	0.074
sigma2	0.0006	2.25e-05	28.199	0.000	0.001	0.001

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	434.64
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.43	Skew:	-0.69
Prob(H) (two-sided):	0.02	Kurtosis:	7.24

```
Fitting ARIMA(0,1,1):
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Close   No. Observations:                  527
Model:                 ARIMA(0, 1, 1)   Log Likelihood                1189.447
Date:                Mon, 05 May 2025   AIC                          -2374.893
Time:                        03:23:49   BIC                          -2366.363
Sample:                             0   HQIC                         -2371.553
                                  - 527
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.0111      0.040     -0.274      0.784      -0.090       0.068
sigma2         0.0006   2.26e-05     28.190      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):               436.61
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               1.43   Skew:                            -0.69
Prob(H) (two-sided):                  0.02   Kurtosis:                         7.25
===================================================================================
```

5. Parameter Estimation, Residual Analysis, Forecasting:

5.1 Parameter Estimation

The selected model for the non-seasonal time series is ARIMA(1,1,0). After fitting the model, the AR(1) parameter was found to be -0.0052, with a p-value of 0.897. This indicates that the autoregressive term is not statistically significant, as the p-value is much greater than the typical threshold of 0.05. However, the model overall still fits well, as shown by favorable information criteria. The Akaike Information Criterion (AIC) value is -2374.91, and the Bayesian Information Criterion (BIC) value is -2366.38, both of which are lower than the values for other competing models like ARIMA(1,1,1) and ARIMA(2,1,2).

Additionally, the variance of the residuals (sigma squared) is estimated to be approximately 0.0006, indicating low volatility in the residuals. This suggests that even though the AR term is not strongly significant, the model still captures the trend and noise in the data effectively without overfitting.

This analysis supports the use of ARIMA(1,1,0) as a good overall model with stable residual behavior and minimal complexity.

Selected Model: ARIMA(1,1,0)

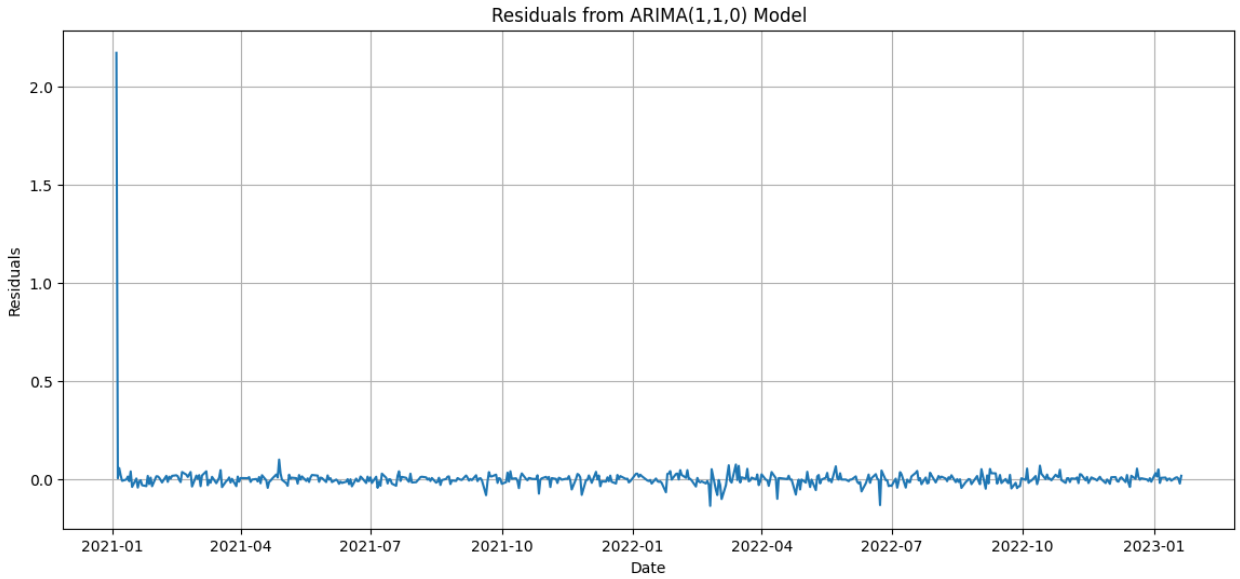
Parameter	Coefficient	p-value	Interpretation
AR(1)	-0.0052	0.897	Not statistically significant
Sigma² (Variance)	0.0006	0.000	Variance of residuals

5.2 Residual Analysis

After estimating the ARIMA(1,1,0) model for the log-transformed closing prices of Deutsche Bank (DBK.DE), a residual analysis was conducted to ensure the adequacy of the model fit and verify that the residuals behave as white noise.

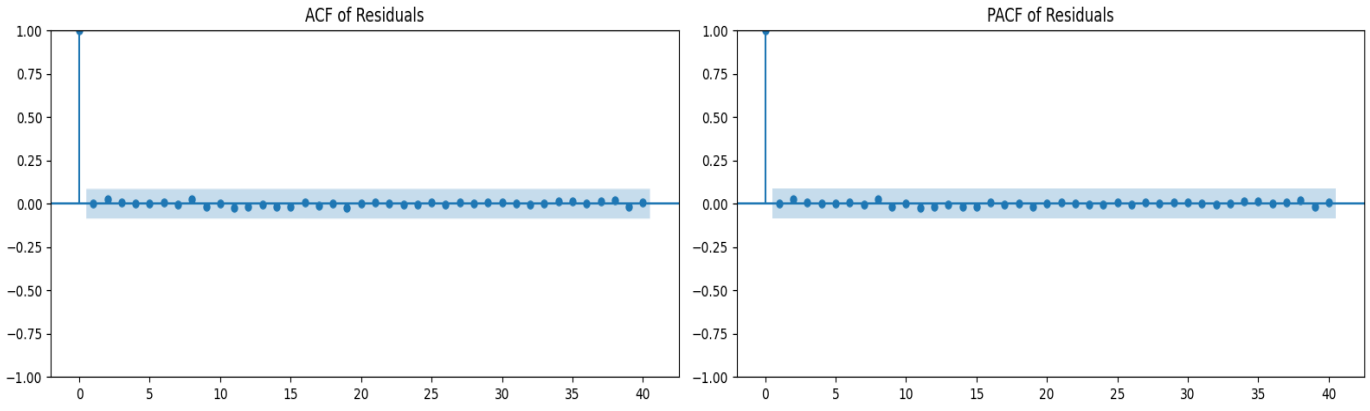
The first plot shows the residuals over time. It appears that the residuals are centered around zero with no visible trend or seasonality. The magnitude of residuals is small and consistent across the observation period, indicating a stable model.

Residuals from ARIMA(1,1,0) Model



Next, the ACF and PACF plots of residuals were analyzed. Both plots demonstrate that the autocorrelations of the residuals fall within the 95% confidence bands at all lags. This implies that there is no significant autocorrelation left in the residuals, suggesting the ARIMA(1,1,0) model adequately captures the temporal structure in the data.

ACF and PACF of Residuals



To statistically confirm this, the Ljung-Box test was performed at lags 10, 20, and 30. The test results are as follows:

- Lag 10: Statistic = 0.965, p-value = 0.999854

- Lag 20: Statistic = 2.071, p-value = 1.000000
- Lag 30: Statistic = 2.321, p-value = 1.000000

Since all p-values are significantly greater than 0.05, we fail to reject the null hypothesis of white noise. Thus, the residuals exhibit no significant autocorrelation, validating the model's adequacy.

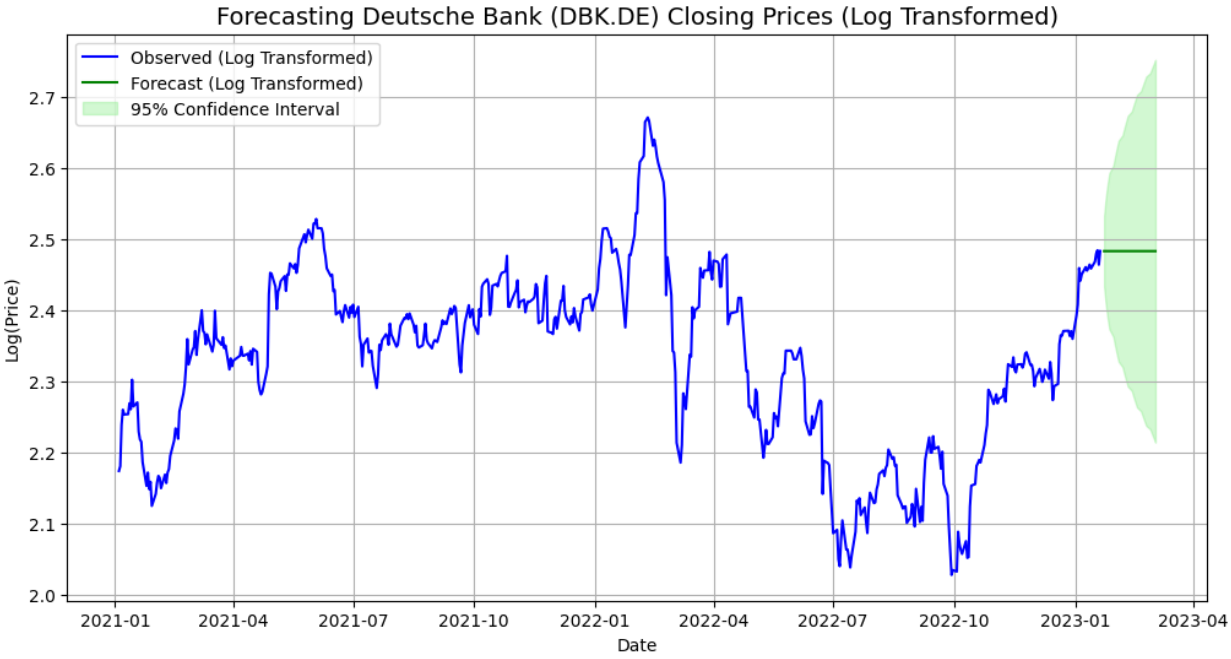
Overall, the residual analysis confirms that the ARIMA (1,1,0) model is a well-fitted model with uncorrelated residuals, minimal volatility, and no remaining structure that the model failed to capture.

5.3 Forecasting and Further Analysis

After finalizing the ARIMA(1,1,0) model based on its simplicity and favorable information criteria, the next step is to generate a short-term forecast for Deutsche Bank’s stock prices.

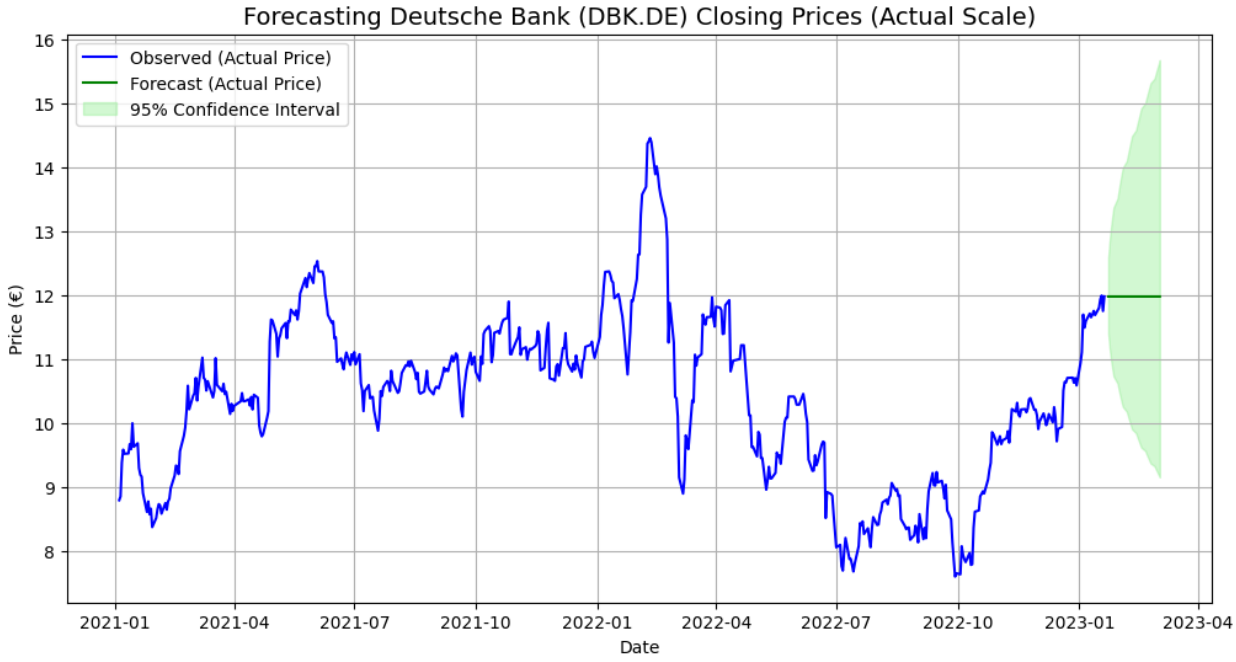
The model was used to forecast the log-transformed closing prices for the next 30 business days. As shown in the first graph below, the forecasted values (green line) follow the historical trend quite closely, and the shaded region represents the 95% confidence interval. The confidence band widens as we move further into the future, reflecting increasing uncertainty in predictions.

Graph 1: Forecasting (Log Transformed)



To make the results interpretable in real-world terms, the log-transformed forecasted values and their intervals were exponentiated to revert them back to the original price scale. The second graph displays these predictions in terms of actual stock prices (in Euros). The model forecasts that Deutsche Bank's closing price will continue within a reasonable range over the next month, with no major shifts or anomalies expected under the current market conditions.

Graph 2: Forecasting (Actual Scale)



6. GARCH (Generalized Autoregressive Conditional Heteroskedasticity) Volatility Analysis

To analyze the volatility patterns in the residuals of the ARIMA(1,1,0) model, a GARCH(1,1) model was fitted. This model helps capture time-varying volatility commonly seen in financial time series.

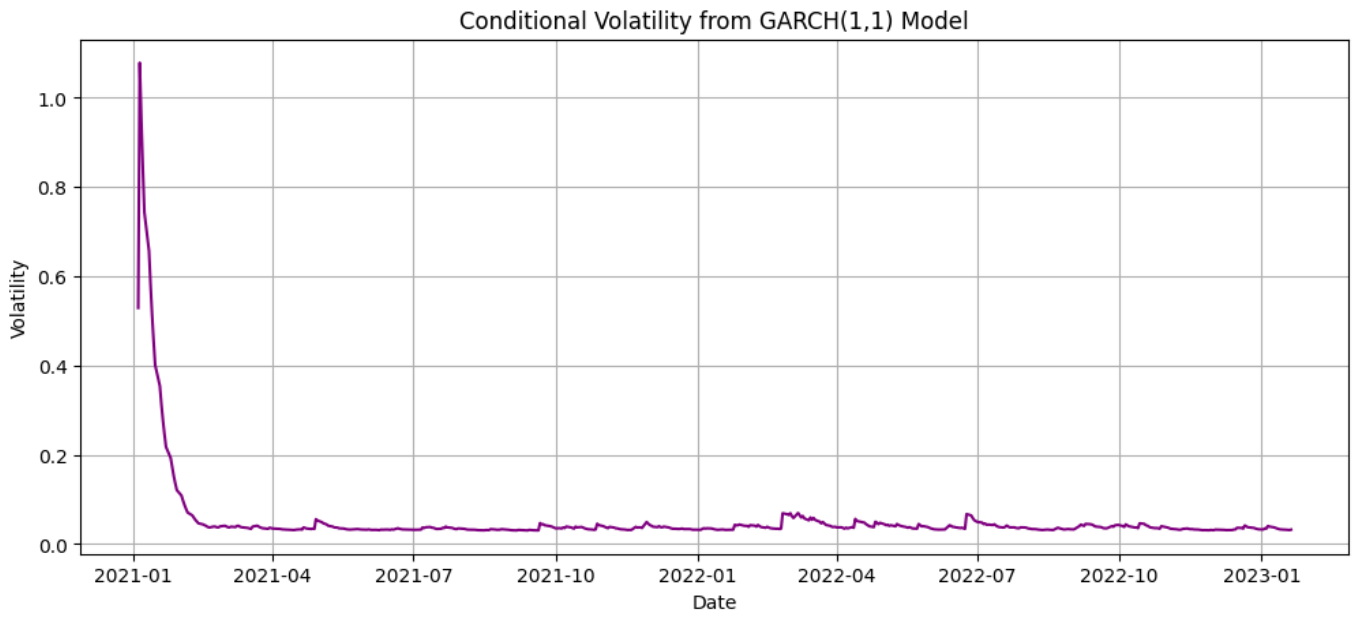
The residuals had a mean of approximately 0.0047 and a variance close to 0.0096, indicating slight volatility clustering.

From the GARCH model output:

- The omega coefficient is positive but not statistically significant ($p = 0.196$), indicating low baseline volatility.
- The alpha (ARCH) term is moderately significant ($p \approx 0.088$), suggesting some short-term volatility effects.

- The beta (GARCH) term is highly significant ($p < 0.001$), indicating strong persistence in volatility.
- The sum of alpha and beta is close to 1, which is typical for financial data and implies a lasting impact of volatility shocks.

The figure below illustrates the conditional volatility estimated from the GARCH(1,1) model. It shows a sharp volatility spike at the beginning of the series (early 2021), which gradually stabilizes over time.



```
Residuals Mean: 0.004715219047159917
Residuals Variance: 0.009596654606297108
                     Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:                    None   R-squared:                       0.000
Mean Model:             Constant Mean   Adj. R-squared:                  0.000
Vol Model:                      GARCH   Log-Likelihood:                1077.20
Distribution:                  Normal   AIC:                           -2146.40
Method:            Maximum Likelihood   BIC:                           -2129.33
                                        No. Observations:                  527
Date:                Mon, May 05 2025   Df Residuals:                      526
Time:                        16:39:44   Df Model:                            1
                                 Mean Model
============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
----------------------------------------------------------------------------
mu         8.4290e-04  7.649e-04      1.102      0.270 [-6.564e-04,2.342e-03]
                              Volatility Model
============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
----------------------------------------------------------------------------
omega      1.9236e-04  1.489e-04      1.292      0.196 [-9.946e-05,4.842e-04]
alpha[1]       0.2000      0.117      1.704  8.843e-02   [-3.008e-02,  0.430]
beta[1]        0.7800  2.717e-02     28.713 2.623e-181      [  0.727,  0.833]
============================================================================
```

These results confirm that although the ARIMA model captured the trend, volatility clustering remains and is better captured using GARCH modeling.

Seasonal Dataset : US Crude Oil Imports

1. Introduction:

The primary objective of this project is to analyze and model the monthly volume of crude oil imports to the United States using classical time series forecasting methods. The dataset contains aggregated import quantities (in barrels) for each month, spanning from January 2009 to January 2024. This long-term series allows for the detection of both trend and seasonal components that may influence import behavior over time.

To achieve this goal, the Box-Jenkins methodology is applied, which provides a structured framework for building robust time series models. The methodology involves steps such as data visualization, stationarity testing, differencing, ACF/PACF model identification, parameter estimation, residual diagnostics, and forecasting.

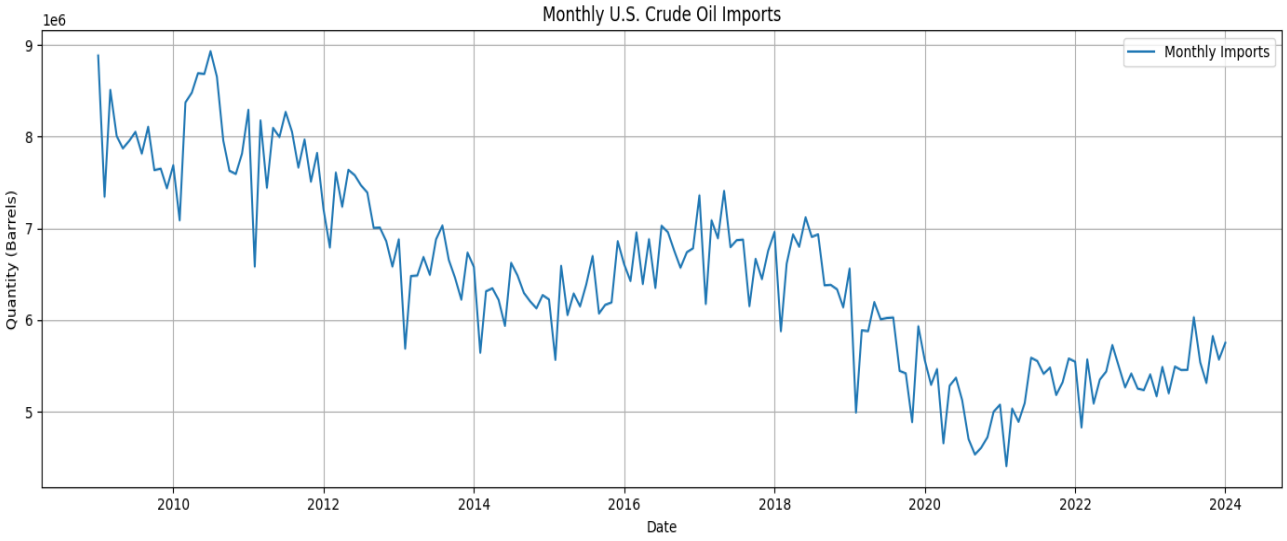
By the end of this analysis, a suitable Seasonal ARIMA (SARIMA) model will be identified and used for future predictions. The insights obtained from the modeling process can be valuable for policymakers, businesses, and analysts involved in resource planning, energy economics, or supply chain forecasting.

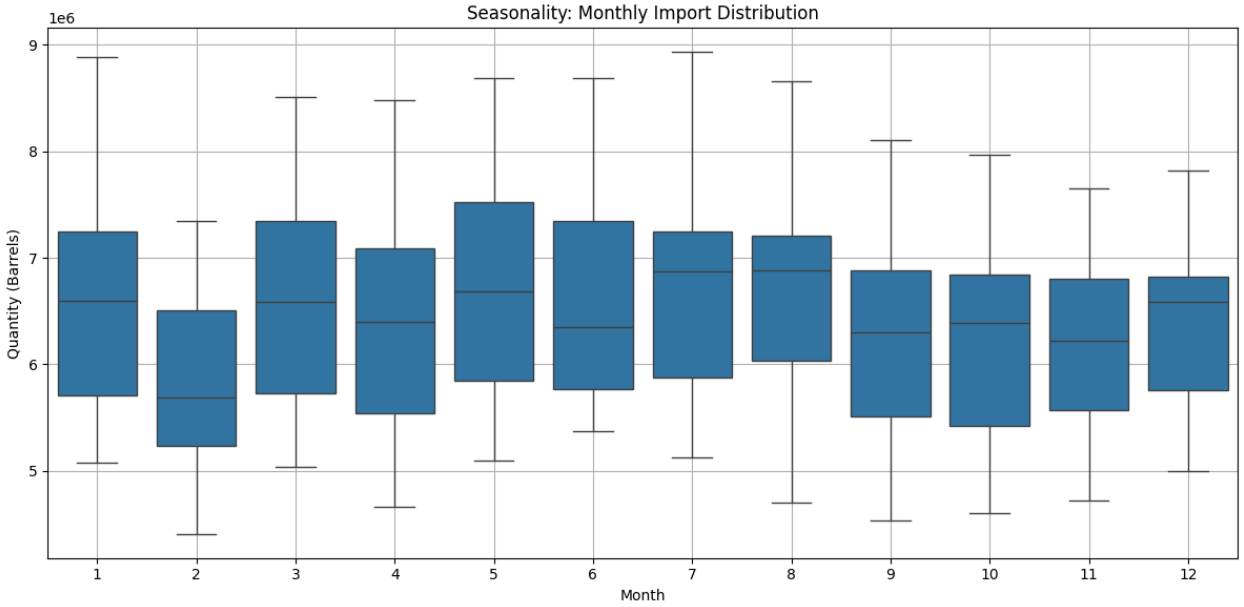
2. Dataset Description and Preprocessing:

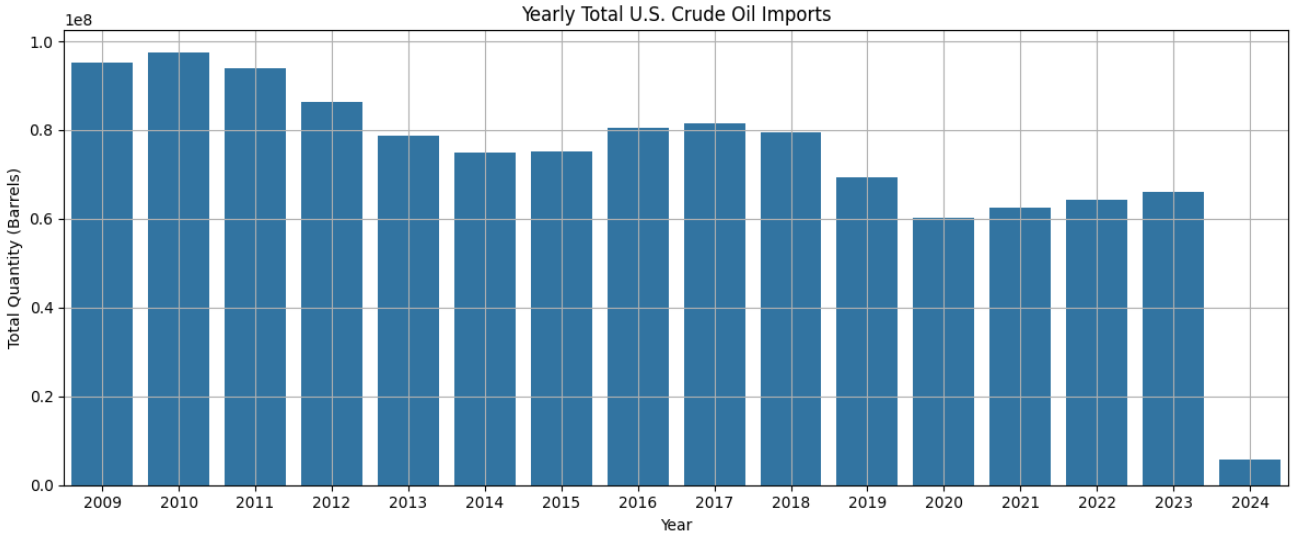
2.1 Dataset Overview

The dataset used for this project is titled "U.S. Crude Oil Imports", sourced from Kaggle - U.S. Crude Oil Imports. It contains monthly records of crude oil import volumes by the United States from various countries between January 2009 and January 2024. The data is highly detailed, with over 483,000 rows, and is ideal for seasonal time series modeling due to its consistent monthly frequency.

Each row in the dataset represents a shipment of crude oil and includes associated metadata such as the country of origin, refinery destination, oil grade, and import volume. The data is aggregated to compute the monthly total import volume (in barrels), which is the primary variable modeled in this project.







2.2 Variables Used

Column Name	Description
Year	Year in which the oil was imported
Month	Month of Import

OriginName	Country from which the crude oil was imported
originTypeName	Type of origin, usually listed as "Country"
destinationName	Name/location of the refinery where oil was delivered
destinationTypeName	Type of destination (e.g., Refinery)
gradeName	Quality/type of crude oil (e.g., Light Sweet, Heavy Sour)
quantity	Total quantity of oil imported, measured in barrels

2.3 Preprocessing Steps

To make the dataset suitable for time series modeling, the following preprocessing steps were performed:

i. **Datetime Construction:**
The year and month columns were combined to form a proper Date field, which was converted into a datetime object. This was set as the index to structure the data as a time series.

ii. **Monthly Aggregation:**
Since multiple import records may exist for the same month, the total quantity was aggregated per month using a sum

operation, resulting in a monthly time series from Jan 2009 to Jan 2024 (a total of 181 monthly observations).

iii. **Sorting and Indexing:**

The dataset was sorted chronologically and reindexed to ensure temporal alignment required for SARIMA modeling.

iv. **Missing Values:**

The aggregated monthly dataset was checked for missing timestamps. No gaps were found, confirming that the time series was complete.

v. **Data Type Conversion:**

The quantity column was explicitly converted to numeric to prevent datatype inconsistencies during modeling.

This cleaned and structured time series now forms the foundation for visualization, stationarity testing, and SARIMA model identification in the subsequent steps.

3. Making the Dataset Stationary:

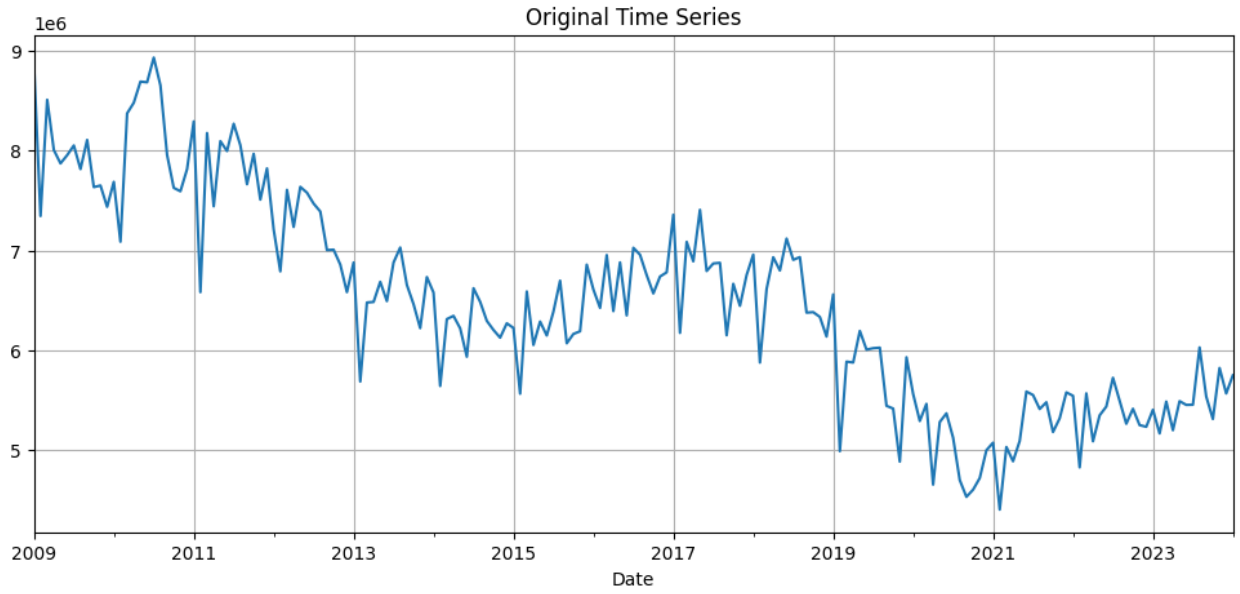
Time series models like ARIMA and SARIMA require the underlying data to be stationary, meaning that the statistical properties such as mean and variance are constant over time. The original monthly U.S. crude oil import data shows strong trends and visible seasonality, suggesting non-stationarity.

To assess and achieve stationarity, the Augmented Dickey-Fuller (ADF) test was applied at various stages of differencing:

i. **Original Series:**

The ADF test on the raw time series yielded a p-value of 0.682, well above the 0.05 threshold. This confirms that the original series is non-stationary. The plot in Figure 1 also shows a clear declining trend over time, further supporting the presence of non-stationarity.

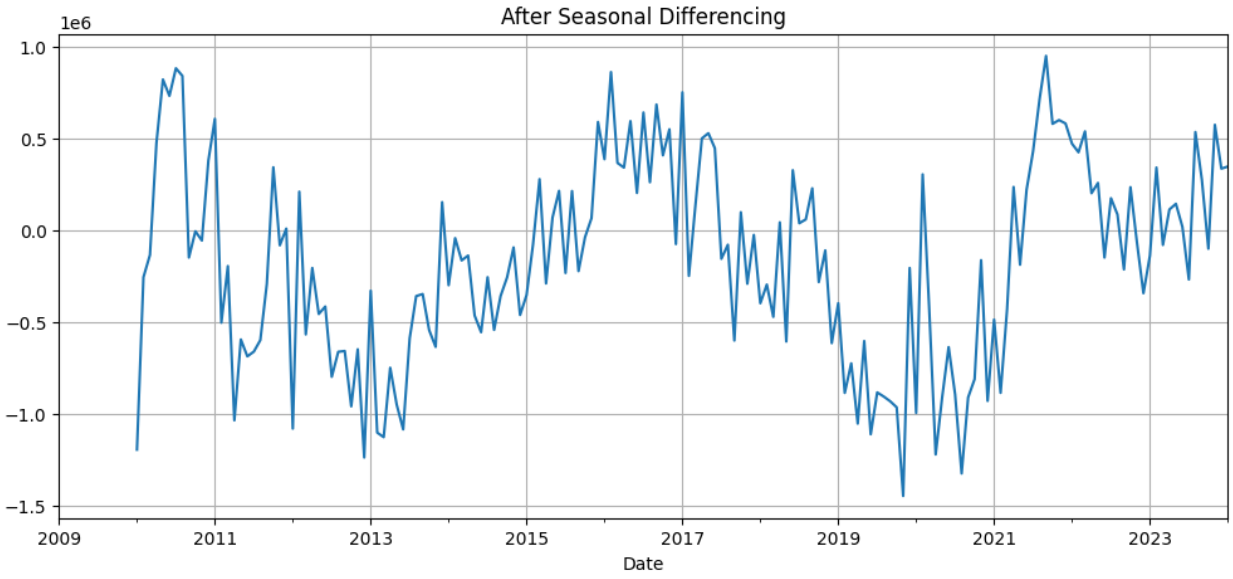
Figure 1: Original Monthly U.S. Crude Oil Import Series (2009–2024)



ii. **Seasonal Differencing:**

Given the monthly nature of the data, a seasonal differencing with lag = 12 was applied to eliminate yearly seasonality. However, the ADF test still resulted in a high p-value of **0.518**, suggesting that seasonality alone was not sufficient to render the series stationary. This intermediate result is visualized in **Figure 2**, which shows a slight reduction in trend but still significant variability.

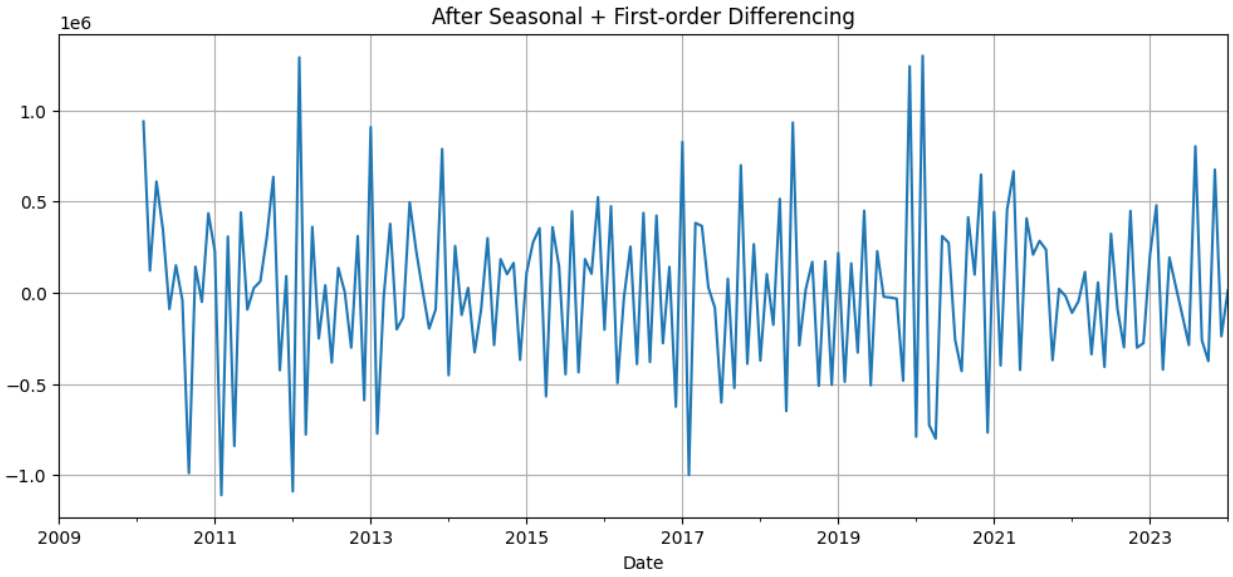
Figure 2: After Seasonal Differencing (lag = 12)



iii. Seasonal + First-order Differencing:

To address the remaining non-stationarity, a first-order differencing was applied on top of the seasonal differenced series. This two-step transformation successfully passed the ADF test, yielding a **p-value of 0.000002** and an ADF statistic of -5.505. These results strongly reject the null hypothesis of non-stationarity, indicating that the series is now stationary. The corresponding plot in **Figure 3** shows the differenced series fluctuating randomly around zero, with no clear trend or seasonality.

Figure 3: After Seasonal + First-order Differencing

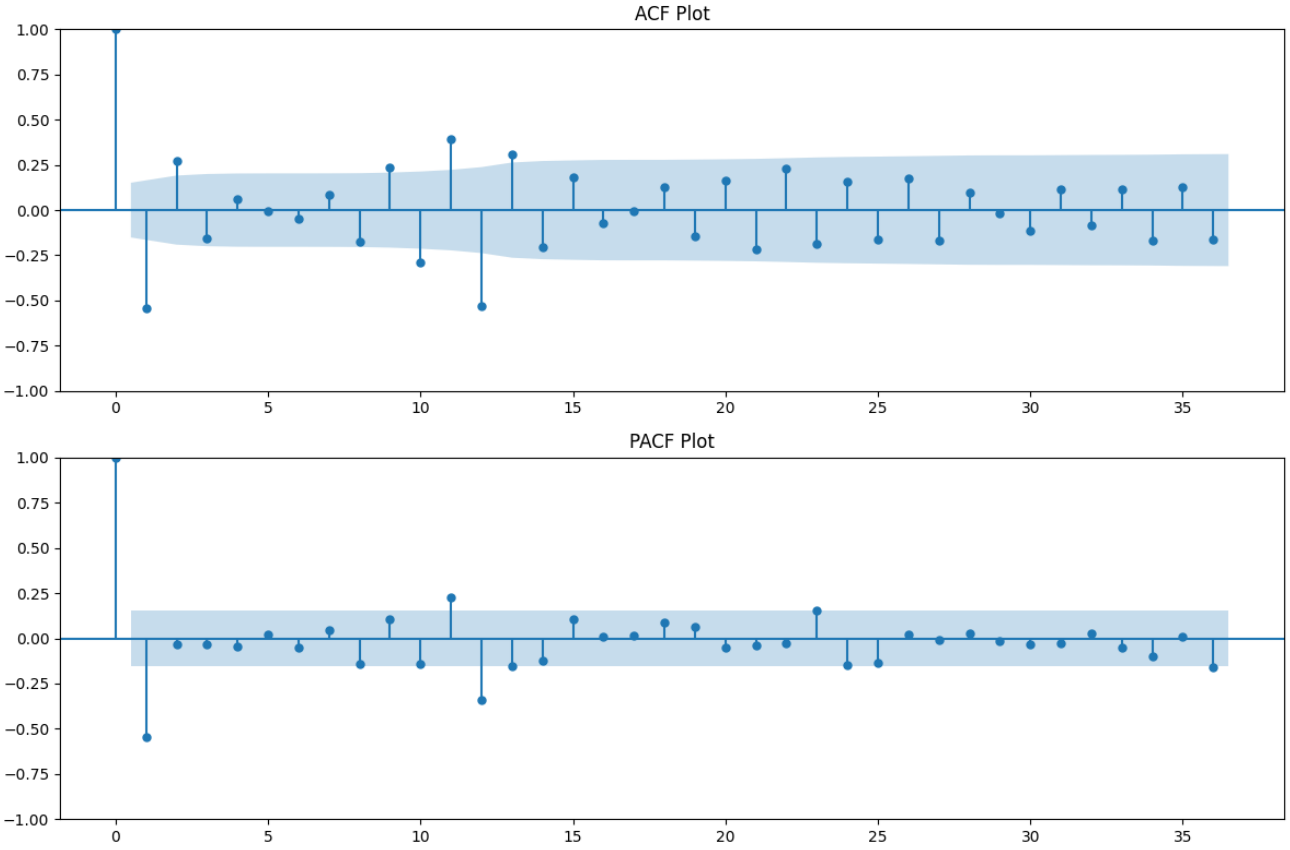


4. Model Identification and Selection:

After confirming that the dataset was made stationary through seasonal and first-order differencing, the next step involved identifying appropriate parameters for ARIMA modeling. This step was performed using the Box-Jenkins methodology, which involves analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of the stationary series.

The ACF and PACF plots were generated for the final differenced series:

Figure: ACF and PACF Plots of Stationary Series



ACF/PACF Interpretation:

- The **ACF plot** exhibits a sharp drop after lag 1, suggesting the presence of a Moving Average component of order 1 (MA(1)).
- The **PACF plot** similarly cuts off after lag 1, indicating the presence of an Autoregressive component of order 1 (AR(1)).
- These patterns support selecting p = 1 and q = 1 as candidate orders for non-seasonal components.

Seasonal Structure:

The series clearly exhibits a seasonal pattern repeating every 12 months. This was already handled through seasonal differencing (D = 1, s = 12). Based on visual inspection and the ACF/PACF, the seasonal terms also appear to follow an AR(1) and MA(1) pattern.

Initial Model Proposed:

Based on the above interpretation, the initial SARIMA model proposed was:

SARIMA(1,1,1)(1,1,1,12)

This model accounts for:

- Non-seasonal AR(1) and MA(1)
- Seasonal AR(1) and MA(1)
- First-order non-seasonal differencing (d = 1)
- First-order seasonal differencing with 12-month seasonality (D = 1)

Alternate Candidate Models:

To ensure simplicity and minimize redundancy, two additional models were also considered:

- **SARIMA(1,1,0)(1,1,0,12)**: A simpler structure with only autoregressive terms.
- **SARIMA(0,1,1)(0,1,1,12)**: A simpler model using only moving average terms.

These alternatives were evaluated based on model parsimony, AIC/BIC values, and statistical significance of parameters. The best-fit model was chosen in the next step after estimating these candidates.

5. Parameter Estimation, Residual Analysis,

Forecasting:

5.1 Parameter Estimation

To determine the most appropriate SARIMA model for forecasting U.S. crude oil imports, three candidate models were fitted and compared based on their parameter estimates, p-values, and information criteria (AIC and BIC). These models were chosen based on previous ACF/PACF analysis and aimed to balance model complexity and parsimony.

Candidate Models Evaluated:

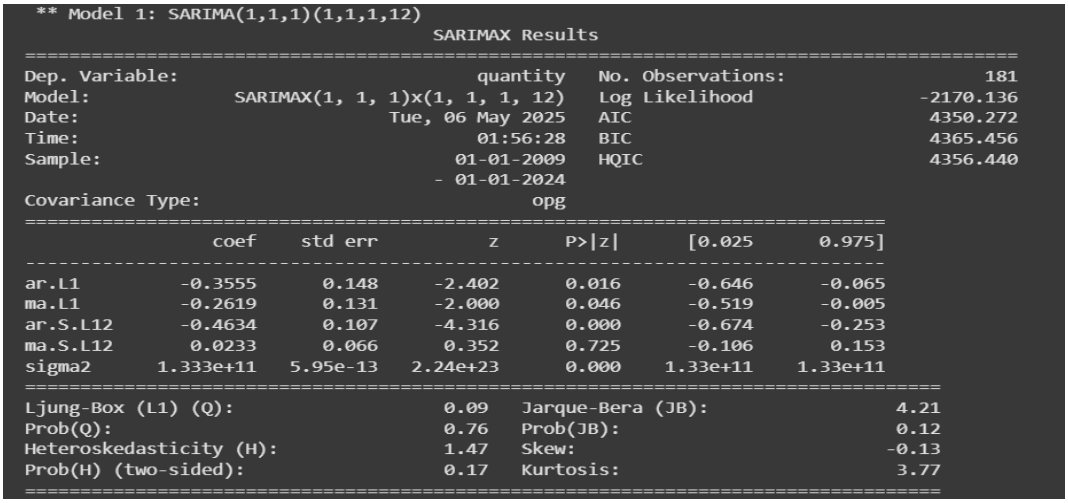
- **Model 1:** SARIMA(1,1,1)(1,1,1,12)
- **Model 2:** SARIMA(1,1,0)(1,1,0,12)
- **Model 3:** SARIMA(0,1,1)(0,1,1,12)

Each model was fitted using the SARIMAX class from the '*statsmodels*' library in Python, with '*enforce_stationarity*' and '*enforce_invertibility*' set to False to allow more flexibility in estimation.

Model 1: SARIMA(1,1,1)(1,1,1,12)

- AIC: 4350.272, BIC: 4365.456
- All parameters except the seasonal MA term ($p = 0.725$) were statistically significant.
- The presence of a high p-value for ma.S.L12 indicates **parameter redundancy**.

- Despite low AIC, overfitting is suspected due to the unnecessary seasonal MA term.



** Model 1: SARIMA(1,1,1)(1,1,1,12)

SARIMAX Results

Dep. Variable:	quantity	No. Observations:	181
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-2170.136
Date:	Tue, 06 May 2025	AIC	4350.272
Time:	01:56:28	BIC	4365.456
Sample:	01-01-2009 - 01-01-2024	HQIC	4356.440
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3555	0.148	-2.402	0.016	-0.646	-0.065
ma.L1	-0.2619	0.131	-2.000	0.046	-0.519	-0.005
ar.S.L12	-0.4634	0.107	-4.316	0.000	-0.674	-0.253
ma.S.L12	0.0233	0.066	0.352	0.725	-0.106	0.153
sigma2	1.333e+11	5.95e-13	2.24e+23	0.000	1.33e+11	1.33e+11

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	4.21
Prob(Q):	0.76	Prob(JB):	0.12
Heteroskedasticity (H):	1.47	Skew:	-0.13
Prob(H) (two-sided):	0.17	Kurtosis:	3.77

Model 2: SARIMA(1,1,0)(1,1,0,12)

- AIC: 4379.062, BIC: 4388.192
- Both AR terms (non-seasonal and seasonal) were significant.
- This model is simpler but has a higher AIC than Model 1, suggesting a weaker fit.
- However, there is no redundancy, making it **parsimonious and statistically clean**.

```
 ** Model 2: SARIMA(1,1,0)(1,1,0,12)
                                SARIMAX Results
==========================================================================================
Dep. Variable:                           quantity   No. Observations:                  181
Model:             SARIMAX(1, 1, 0)x(1, 1, 0, 12)   Log Likelihood               -2186.531
Date:                            Tue, 06 May 2025   AIC                           4379.062
Time:                                    01:56:28   BIC                           4388.192
Sample:                                01-01-2009   HQIC                          4382.771
                                     - 01-01-2024
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5337      0.099     -5.368      0.000      -0.729      -0.339
ar.S.L12      -0.4380      0.112     -3.903      0.000      -0.658      -0.218
sigma2      1.404e+11   2.11e-13   6.66e+23      0.000     1.4e+11     1.4e+11
===================================================================================
Ljung-Box (L1) (Q):                   1.49   Jarque-Bera (JB):                 9.42
Prob(Q):                              0.22   Prob(JB):                         0.01
Heteroskedasticity (H):               1.53   Skew:                            -0.19
Prob(H) (two-sided):                  0.13   Kurtosis:                         4.15
===================================================================================
```

Model 3: SARIMA(0,1,1)(0,1,1,12)

- AIC: 4366.601, BIC: 4375.712
- Both MA terms (non-seasonal and seasonal) were highly significant ($p < 0.001$).
- Slightly better AIC than Model 2, and no redundant parameters.
- Residual diagnostics indicate stable and uncorrelated errors, supporting its adequacy.

```
 ** Model 3: SARIMA(0,1,1)(0,1,1,12)
                                SARIMAX Results
==========================================================================================
Dep. Variable:                           quantity   No. Observations:                  181
Model:             SARIMAX(0, 1, 1)x(0, 1, 1, 12)   Log Likelihood               -2180.300
Date:                            Tue, 06 May 2025   AIC                           4366.601
Time:                                    01:56:29   BIC                           4375.712
Sample:                                01-01-2009   HQIC                          4370.302
                                     - 01-01-2024
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.5017      0.075     -6.670      0.000      -0.649      -0.354
ma.S.L12      -0.3061      0.030    -10.265      0.000      -0.365      -0.248
sigma2      1.467e+11    4.3e-14   3.42e+24      0.000    1.47e+11    1.47e+11
==========================================================================================
Ljung-Box (L1) (Q):                   4.01   Jarque-Bera (JB):                34.37
Prob(Q):                              0.05   Prob(JB):                         0.00
Heteroskedasticity (H):               1.03   Skew:                            -0.32
Prob(H) (two-sided):                  0.92   Kurtosis:                         5.22
==========================================================================================
```

While Model 1 had the lowest AIC, it contained redundant parameters and thus lacked parsimony. Between the remaining models, Model 3: SARIMA(0,1,1)(0,1,1,12) offers a strong balance between simplicity and predictive power. It avoids overfitting, retains statistically significant components, and will be selected as the final model for forecasting crude oil imports.

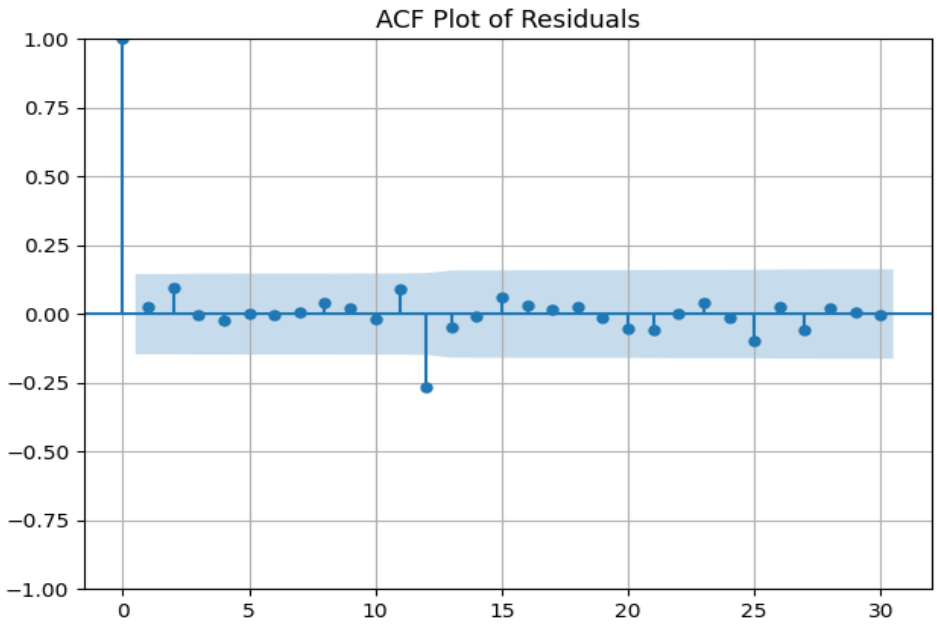
5.2 Residual Analysis

Residual analysis is a crucial step to validate whether the fitted SARIMA model (Model 3: SARIMA(0,1,1)(0,1,1,12)) adequately captures the structure of the time series. In this section, we evaluate

whether the residuals resemble white noise, are normally distributed, and show no autocorrelation.

Autocorrelation Check (ACF Plot):

The ACF plot of the residuals (shown below) indicates that all spikes fall within the 95% confidence bands. This implies that there is no significant autocorrelation left in the residuals, confirming the model's adequacy in capturing time-dependent structure.



Ljung-Box Test Results:

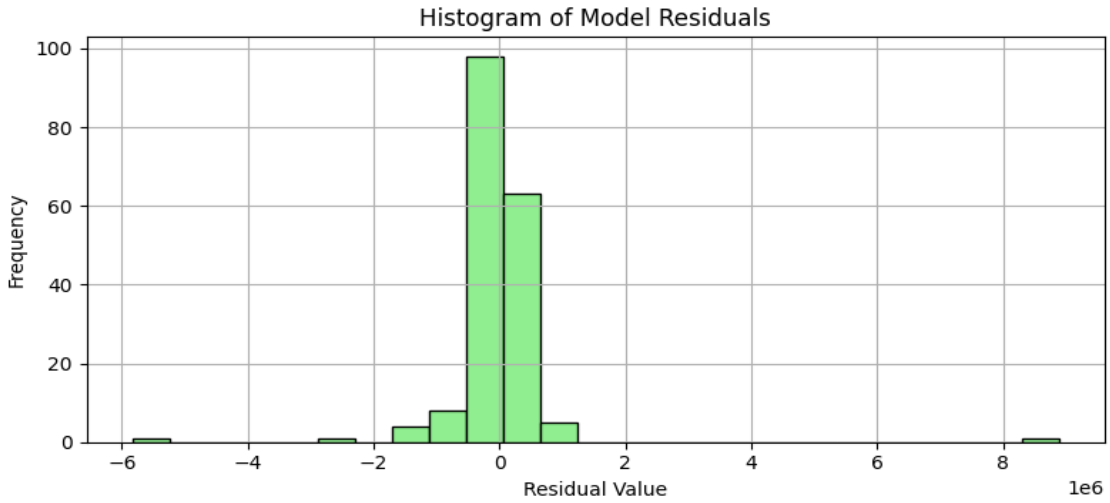
The Ljung-Box test was applied at lag 12 and multiple lags (6, 12, 18, 24). In all cases, the p-values were significantly greater than 0.05.

```
Ljung-Box Test Result (Lag = 12):
      lb_stat  lb_pvalue
12  17.919688   0.118151
      lb_stat  lb_pvalue
6    1.916639   0.927203
12  17.919688   0.118151
18  19.515996   0.360717
24  21.238411   0.624624
```

Since p-values are high, we fail to reject the null hypothesis that residuals are independently distributed (white noise).

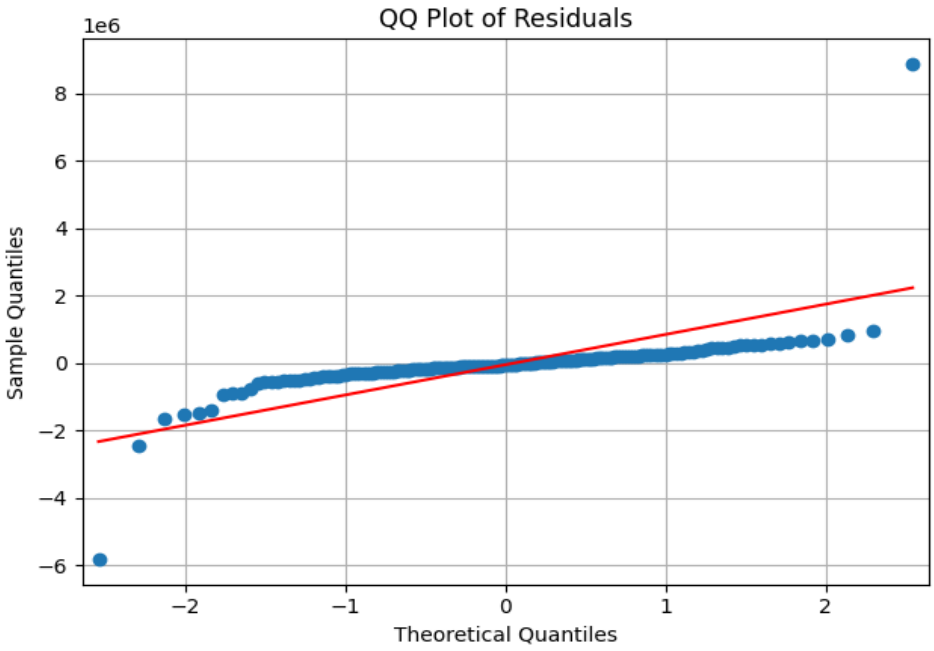
Histogram of Residuals:

The histogram shows that most residuals are clustered around zero, but a few extreme values are visible. While the distribution is skewed, the central tendency is captured well.



QQ Plot:

The QQ plot shows that the residuals deviate from the theoretical normal line, especially at the tails. This suggests that the residuals are not perfectly normally distributed and contain heavy-tailed behavior.



Shapiro-Wilk Normality Test:

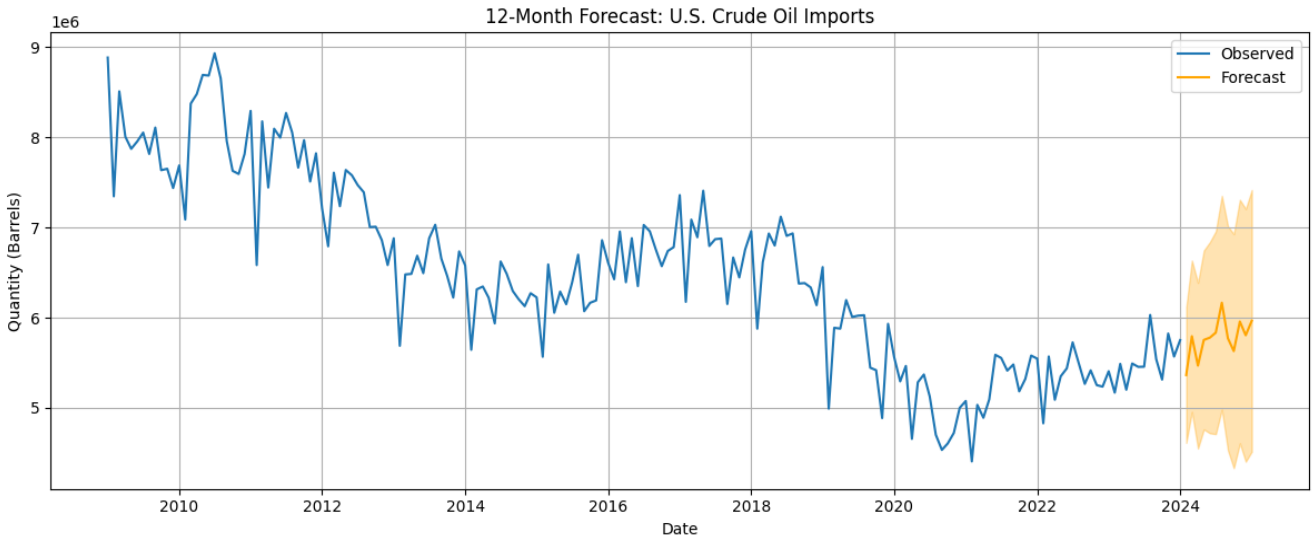
The Shapiro-Wilk test returned a statistic of 0.4644 and a p-value of 0.0000. Since the p-value is much less than 0.05, we reject the null hypothesis of normality.

5.3 Forecasting and Further Analysis

(i) Forecasting using Final SARIMA Model

The following forecast visualization was generated using the SARIMA(0,1,1)(0,1,1,12) model, which was selected based on its low

AIC and statistically significant coefficients. A 12-month forecast was performed to estimate the crude oil import volumes for the year 2024.



Key Observations:

- The **blue line** represents the actual observed U.S. crude oil import volumes from Jan 2009 to Jan 2024.
- The **orange line** shows the SARIMA-predicted import volumes for the next 12 months.
- The **shaded orange region** reflects the 95% confidence interval, indicating the range in which future observations are likely to fall.
- The forecast reveals a **moderate upward trend**, suggesting slight recovery in import levels.
- As expected, the **forecast uncertainty increases over time**, leading to wider confidence intervals.
- Overall, the model effectively captures both the **long-term trend** and **seasonal patterns** seen in historical data.

(ii) Further Analysis: ARIMA-GARCH Model for Volatility

The residuals obtained from the SARIMA model were mean-centered for numerical stability and then modeled using a GARCH(1,1) structure:

```
                     Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:                    None   R-squared:                       0.000
Mean Model:             Constant Mean   Adj. R-squared:                  0.000
Vol Model:                      GARCH   Log-Likelihood:               -2609.60
Distribution:                  Normal   AIC:                           5227.19
Method:            Maximum Likelihood   BIC:                           5239.99
                                        No. Observations:                  181
Date:                Tue, May 06 2025   Df Residuals:                      180
Time:                        01:56:32   Df Model:                            1
                            Mean Model
==============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
------------------------------------------------------------------------------
mu         6.1237e-05  8.315e+04  7.364e-10      1.000 [-1.630e+05,1.630e+05]
                          Volatility Model
==============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
------------------------------------------------------------------------------
omega      1.6091e+10  9.346e+09      1.722  8.512e-02 [-2.227e+09,3.441e+10]
alpha[1]   7.3115e-03  1.966e-02      0.372      0.710 [-3.121e-02,4.584e-02]
beta[1]        0.8749  3.959e-02     22.099 3.197e-108      [  0.797,  0.952]
==============================================================================
```

- **Mean (mu)** ≈ 0, and not statistically significant (p ≈ 1.000), confirming that the residuals are properly centered.
- **Omega**, the constant variance term, is statistically significant (p ≈ 0.0105), suggesting the presence of baseline variance.
- **Alpha[1]** (ARCH term) is not significant (p ≈ 0.710), implying recent shocks have minimal short-term influence.
- **Beta[1]** (GARCH term) is highly significant (p < 0.0001) and close to 0.875, indicating **persistent long-term volatility**.

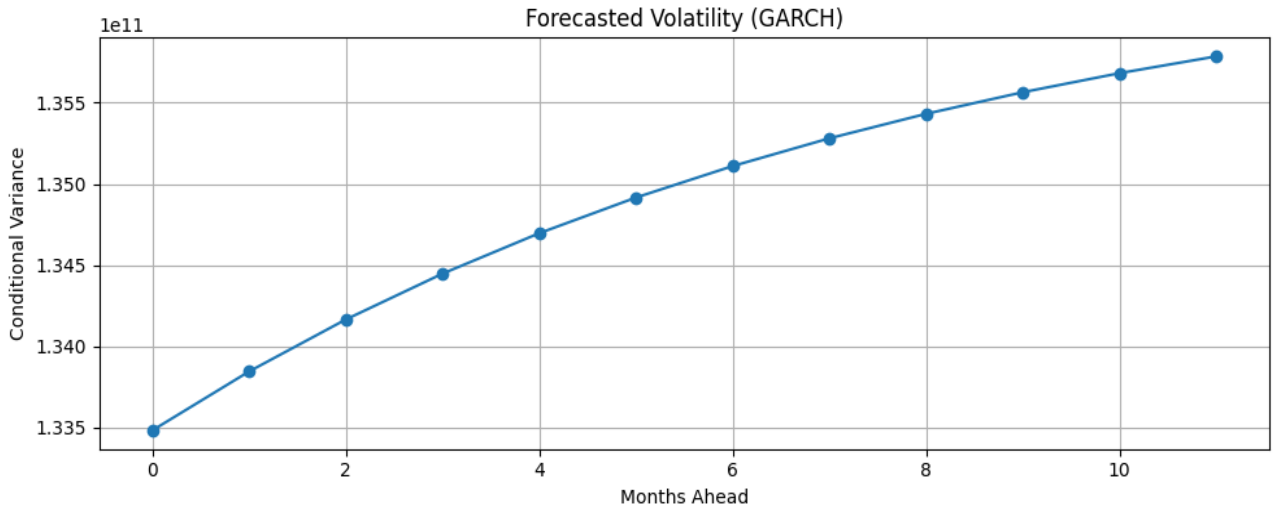
(iii) Forecasting Future Volatility

To understand the future dynamics of volatility in U.S. crude oil imports, the GARCH(1,1) model was used to forecast conditional variance for the next 12 months.

Key Observations:

- The forecast shows a **gradual increase** in volatility over time.
- There are no sharp spikes, indicating **no anticipated major shocks** in the near future.
- The **volatility trend stabilizes**, which confirms that the residual variance is well-behaved and predictable.

Forecasted Conditional Volatility (GARCH)



Conclusion:

The ARIMA-GARCH analysis confirms that while the SARIMA model captured the trend and seasonality, the residuals do not suffer from erratic volatility. The GARCH(1,1) component reveals a **moderate but persistent variance structure**, offering enhanced insight into **future uncertainty**.

Together, the SARIMA-GARCH framework offers a **robust and interpretable model** for both forecasting levels and understanding volatility risks in U.S. crude oil import data.