Gold Forecasting

Time-Series Analysis

OUR TEAM

Summer Yu

Kelvin Hsueh

Peter Fu Chen

Reina Chen

PROBLEM STATEMENT

Gold has historically been a valuable asset, influenced by various factors,

and has served as a hedge against economic uncertainty and inflation

We are curious: By applying the time series analysis learned in class,

How accurately can we predict or analyze the gold price?

PROJECT DESIGN

The ARIMAX model integrates autoregressive and moving average components with exogenous variables, providing a robust framework for capturing complex relationships. The Drift method assumes a constant trend over time. The Prophet model leverages an additive approach to handle seasonality and holidays. Lastly, the GARCH model is used for volatility forecasting, capturing the time-varying nature of volatility in the data. These methods collectively offer comprehensive insights into the underlying patterns and future trends of the time series under study.

01. ARIMAX

02. DRIFT

03. PROPHET

Mean method

04. ARCH / GARCH

Volatility method

PROJECT FLOW

STEP-1

Data Source & Pre-Processing

DATA SOURCE

DATA SET-1: Daily Basis

Gold Daily Price

DATA SET-2: Month-End Basis (ARIMAX Only)

Interest Rate

- Inverse with gold prices, as rising interest rates increase the opportunity cost of holding non-yielding assets like gold

Recession

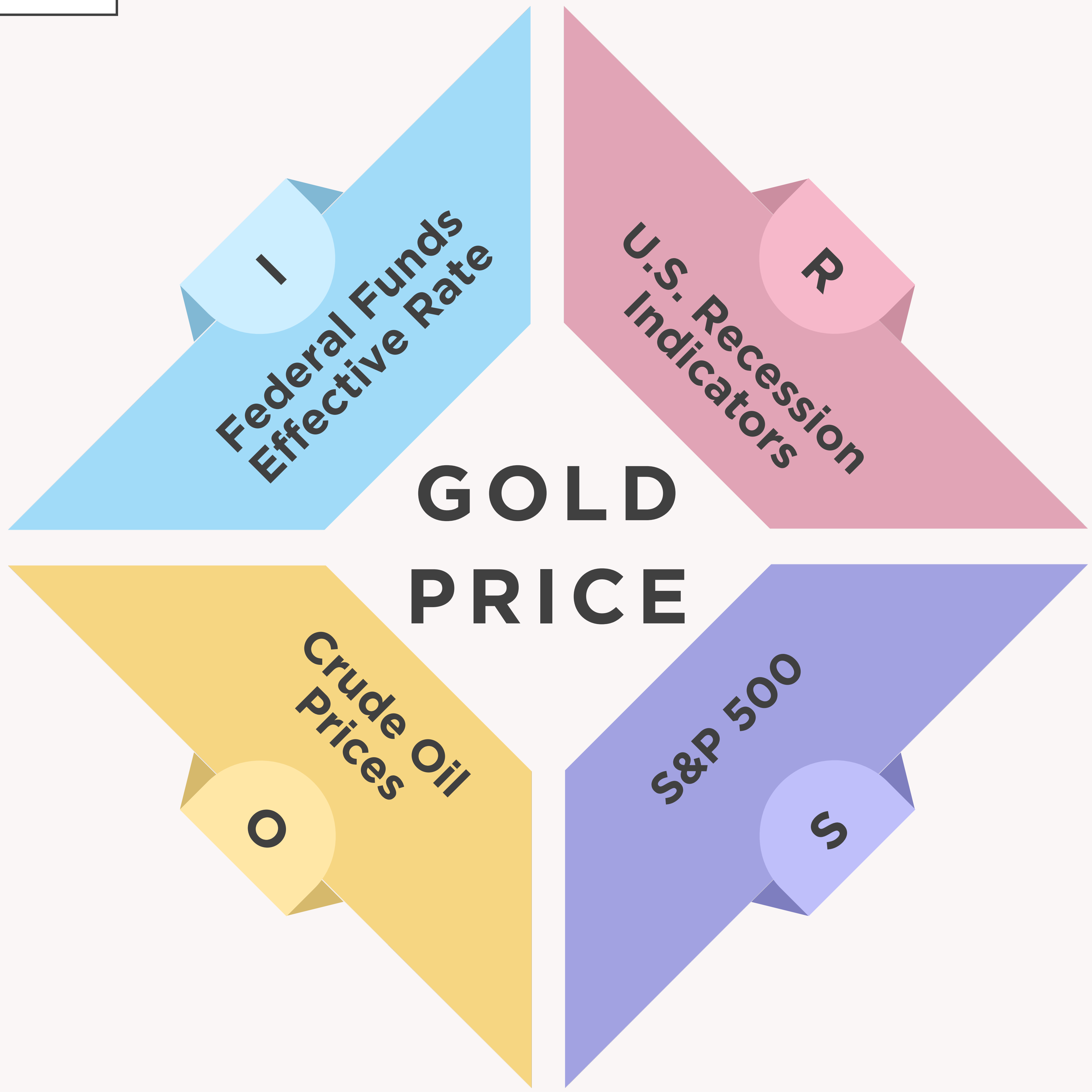
- Positive correlation; as gold is often seen as a safe haven during economic downturns

Stock Price

- Inverse with gold price; as stock returns increase, gold becomes less appealing

Oil Price

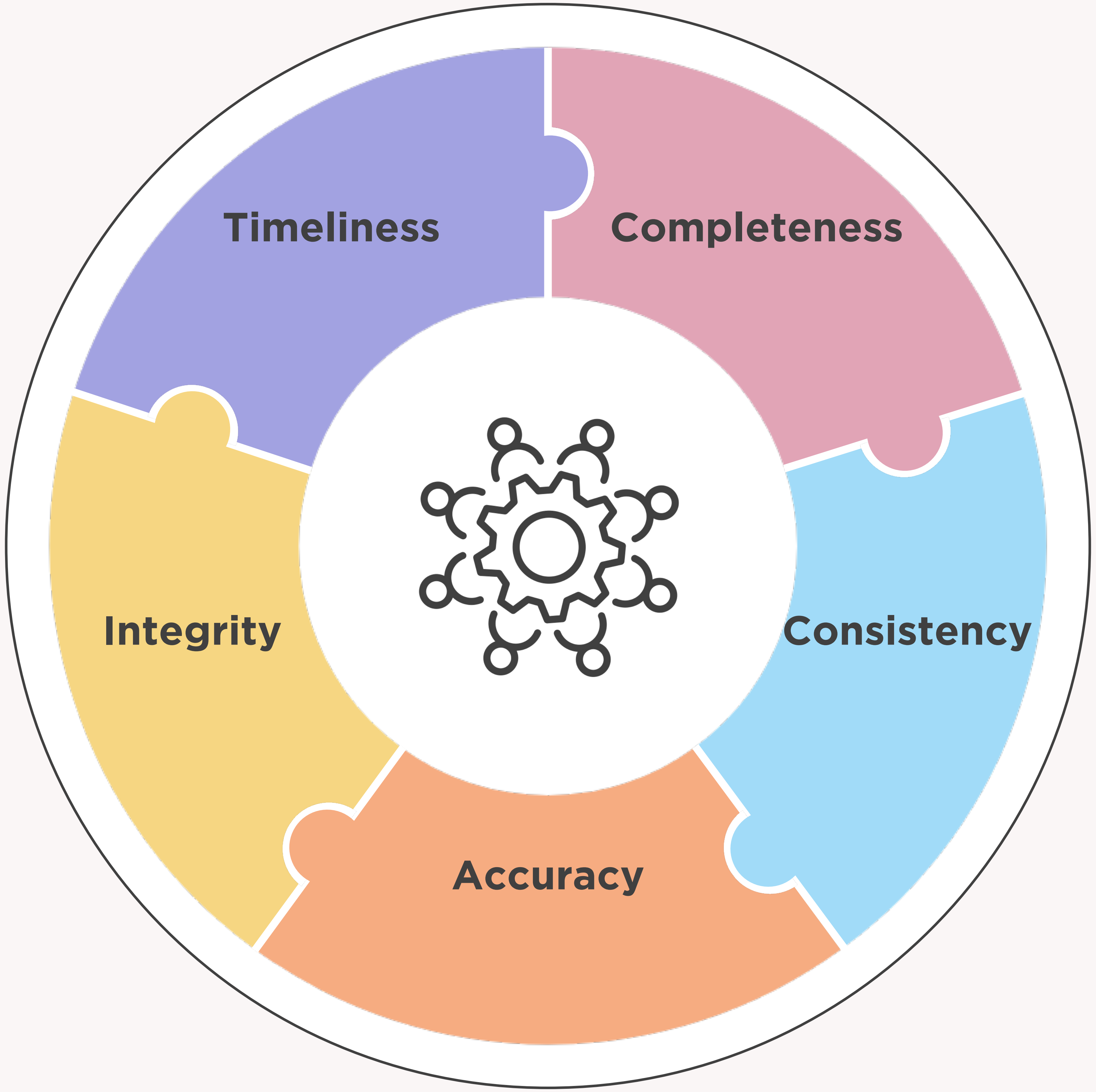
- As a major indicator affecting inflation and growth; rising oil prices potentially increasing gold demand as an inflation hedge



DATA ASSUMPTIONS

01. DATA COMPLETENESS AND QUALITY

- Data collected from reputable sources such as the World Gold Council, FRED, and others
- Assume the sources provide accurate, complete data, free from significant error



DATA PRE-PROCESSING

01. Formatting

02. NaN Values

- Ensure that the data's dates are month-end and ranked in order
- Deal with NaN values by using the fill-forward method to fill the values

Date	Gold_Price	Fed_Funds_Rate	CPI	SP500	WTI_OIL
2019-04-30	1286.45	2.45	2.099331	2945.83	63.83
2019-05-31	1283.95	2.40	1.916335	2752.06	53.49
2019-06-28	1359.04	NaN	NaN	NaN	NaN
2019-06-30	NaN	2.40	1.764365	2941.76	58.20
2019-07-31	1412.98	2.40	1.917843	2980.38	58.53

Date	Gold_Price	Fed_Funds_Rate	CPI	SP500	WTI_OIL
2019-04-30	1286.45	2.45	2.099331	2945.83	63.83
2019-05-31	1283.95	2.40	1.916335	2752.06	53.49
2019-06-28	1359.04	2.40	1.764365	2941.76	58.20
2019-06-30	1412.98	2.40	1.764365	2941.76	58.20
2019-07-31	1412.98	2.40	1.917843	2980.38	58.53
...	...	...	...	...	...
2023-08-31	1920.03	5.33	3.696583	4507.66	83.55
2023-09-29	1916.96	5.33	3.650365	4288.05	90.77
2023-09-30	1913.04	5.33	3.650365	4288.05	90.77
2023-10-31	1913.04	5.33	3.214305	4193.80	81.64
2023-11-30	1985.27	5.33	3.121570	4567.80	75.66

71 rows × 5 columns

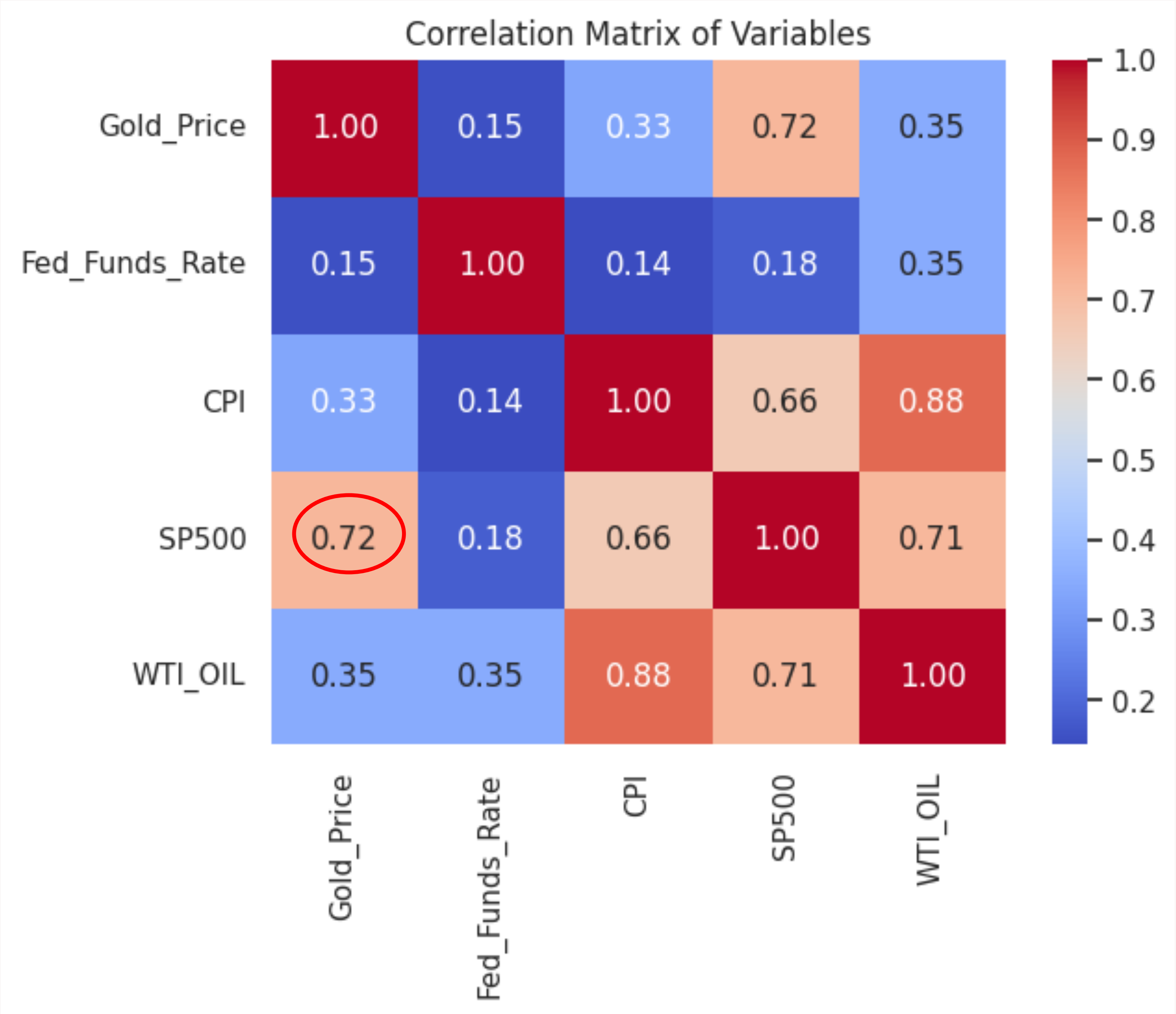
STEP-2

Preliminary Analysis

EDA

01. Correlation Matrix

- Gold Price and S&P 500 have a correlation of 0.72, indicating a strong positive correlation
- In contrast, Gold Price and Fed Funds Rate have a correlation of 0.15, showing a weak positive correlation
- All economic indicators have a positive relationship with Gold Price



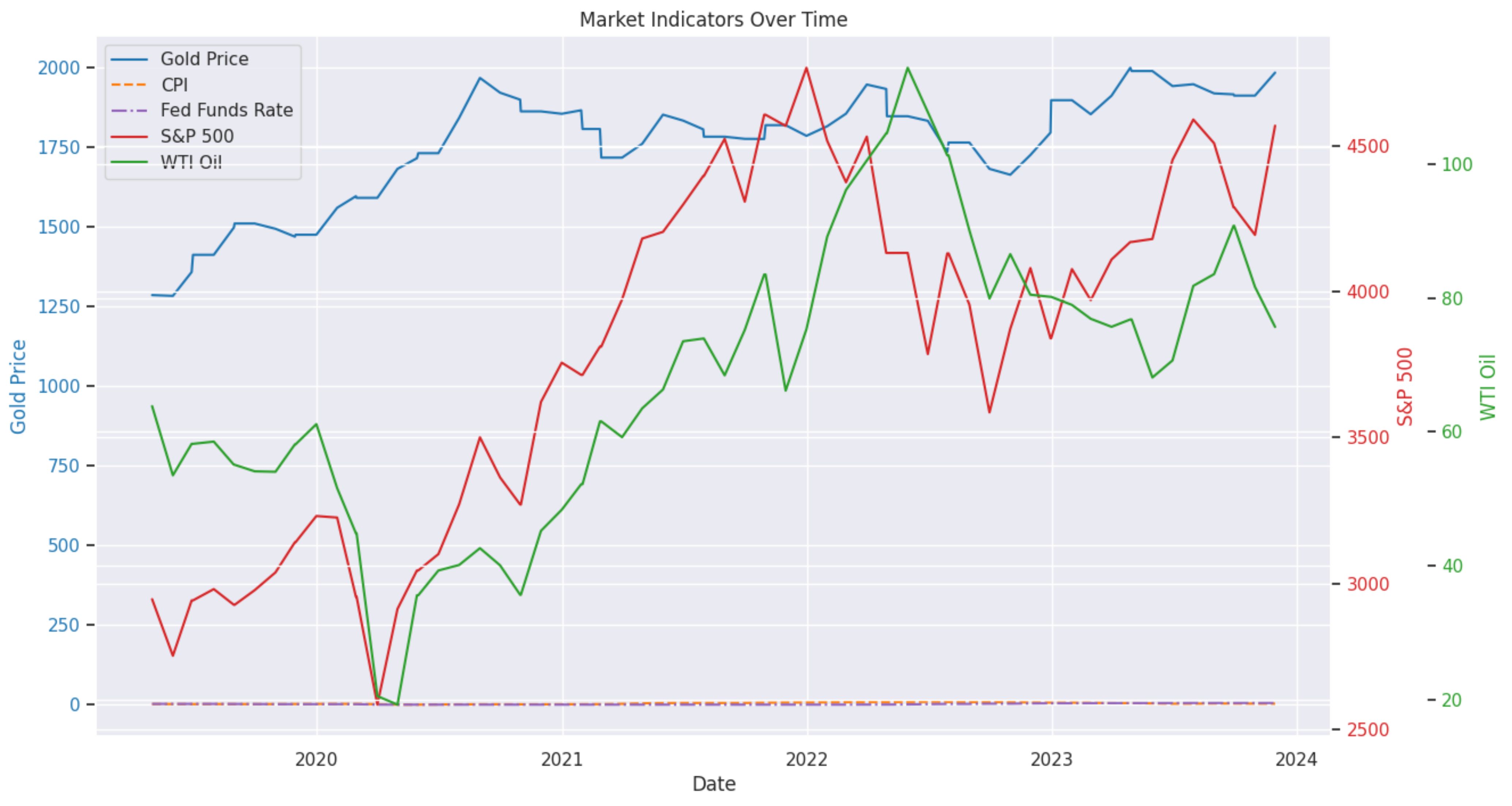
	Gold_Price	Fed_Funds_Rate	CPI	SP500	WTI_OIL
Gold_Price	1.000000	0.152112	0.330900	0.716368	0.345616
Fed_Funds_Rate	0.152112	1.000000	0.143392	0.177329	0.346727
CPI	0.330900	0.143392	1.000000	0.661722	0.878530
SP500	0.716368	0.177329	0.661722	1.000000	0.712973
WTI_OIL	0.345616	0.346727	0.878530	0.712973	1.000000

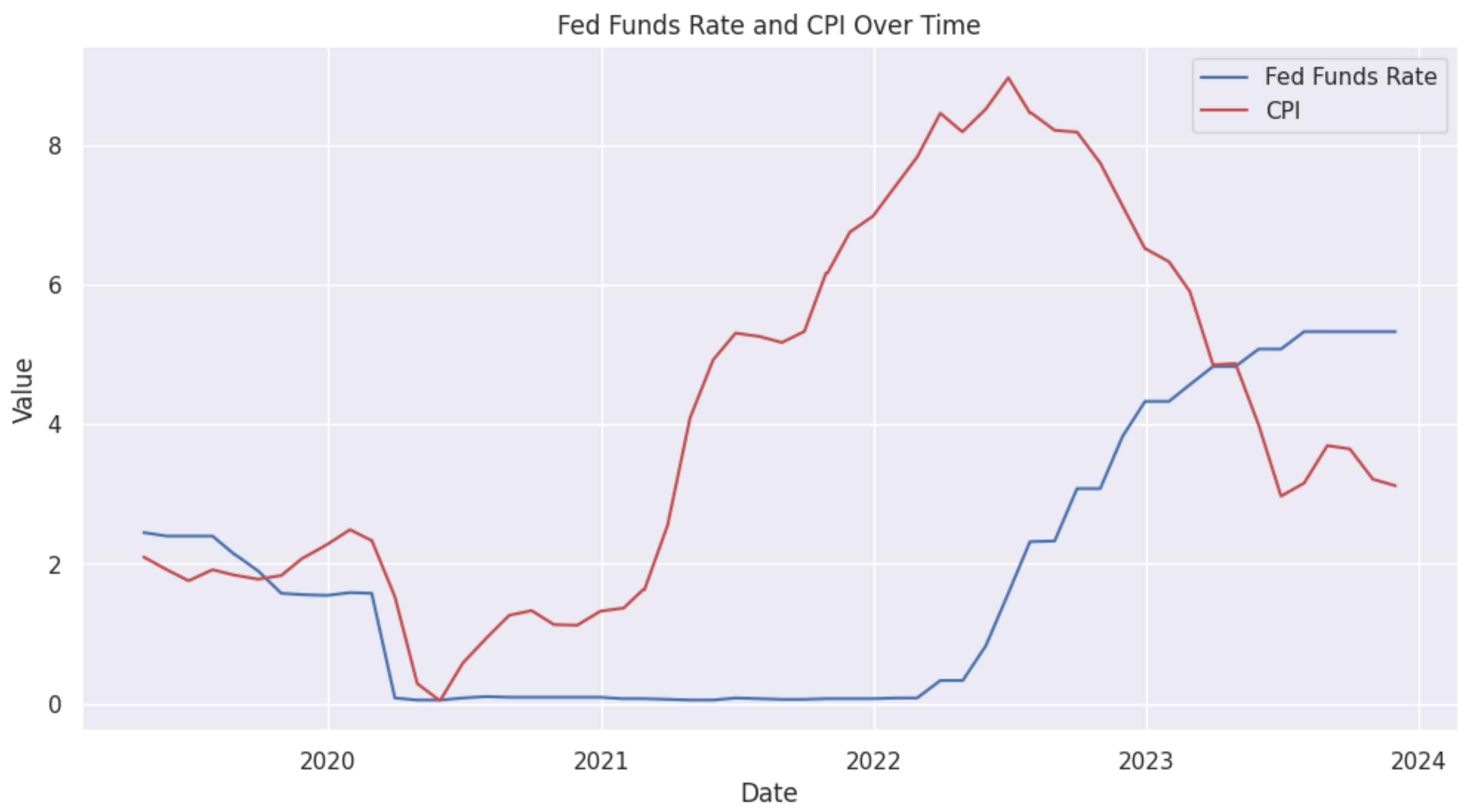
EDA

01. Correlation Matrix

02. Overall Trend (2019-23)

- Gold Price: Rises steeply through mid-2020 to 2021, peaks at around 2000, and then shows fluctuation with a slight overall decrease towards 2024
- CPI: Increased due to inflation, then started to decline
- Fed Funds Rate: The sharp increase in the Fed Funds Rate in 2022 is a clear indication of the Federal Reserve's aggressive monetary policy to curb inflation
- S&P500: General upward trend with some volatility; Shows a sharp decline early in 2021, followed by recovery and growth
- WTI Oil: High volatility, notable is the sharp decline early 2020, likely due to decreased demand during lockdowns, and subsequent recovery





STEP-3

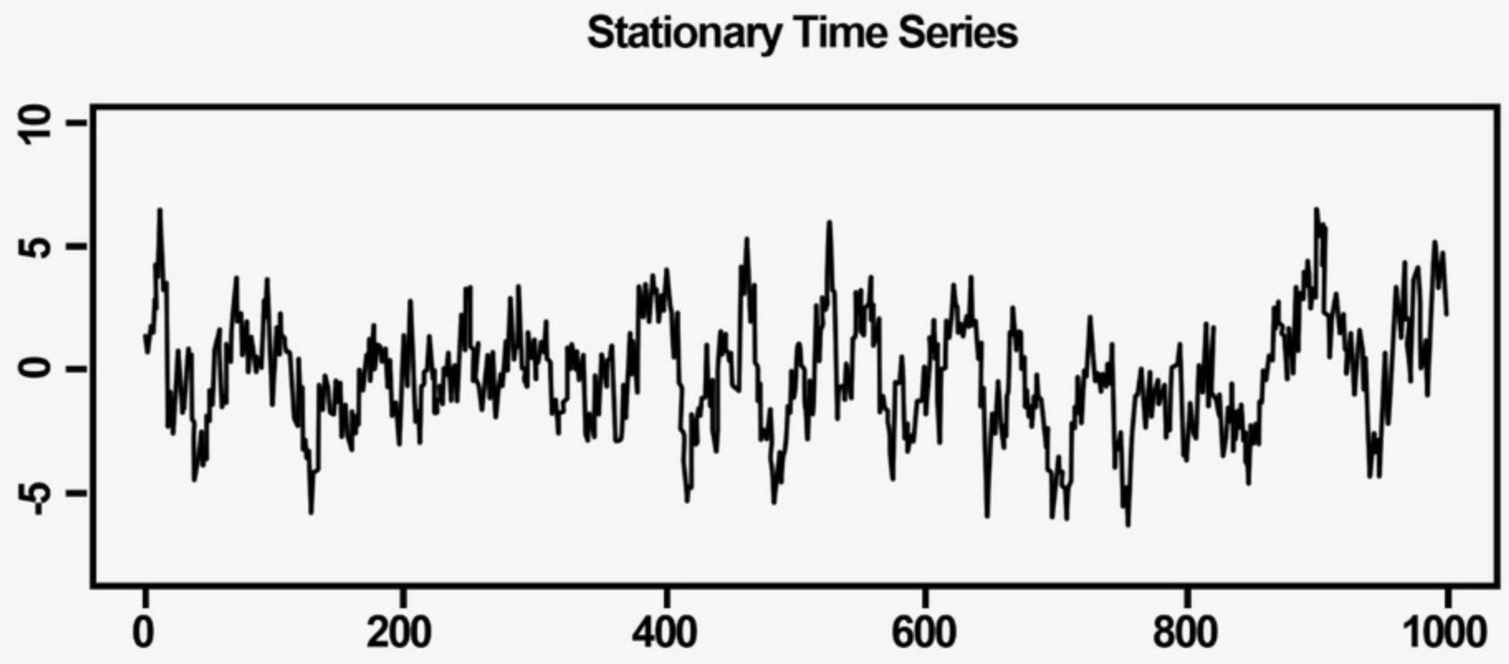
Data Processing

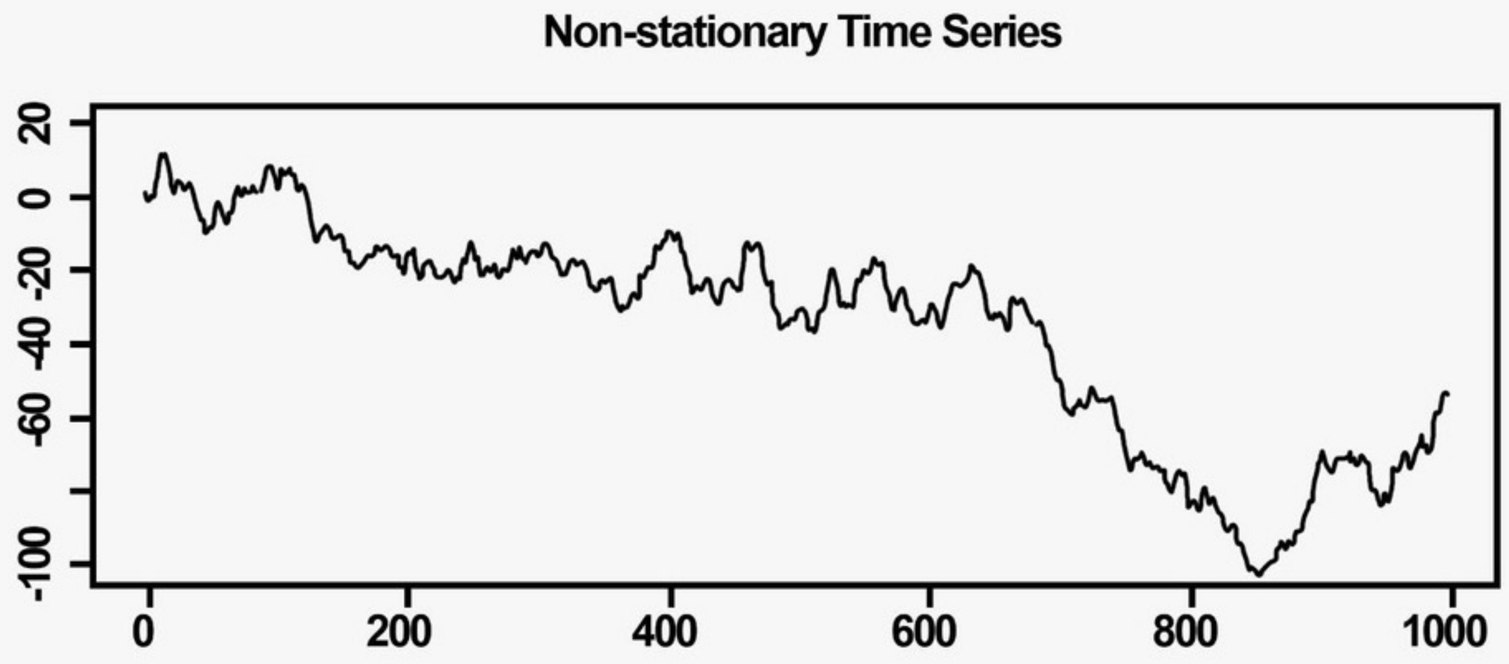
DATA PROCESSING

01. STATIONARITY

For effective time series analysis, we need to ensure that the data are stationary or have been properly transformed to be stationary. This means that the statistical properties, such as:

- constant mean,
- constant variance
- constant autocorrelation structure
- no seasonality

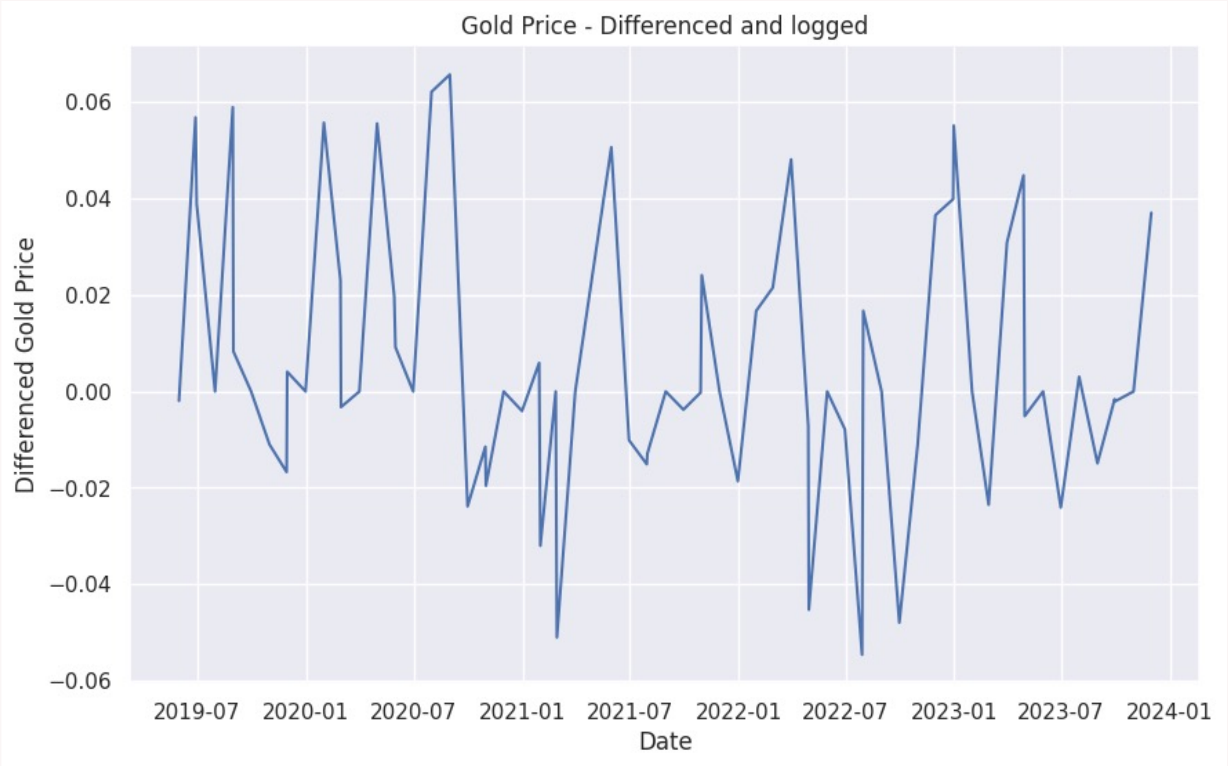


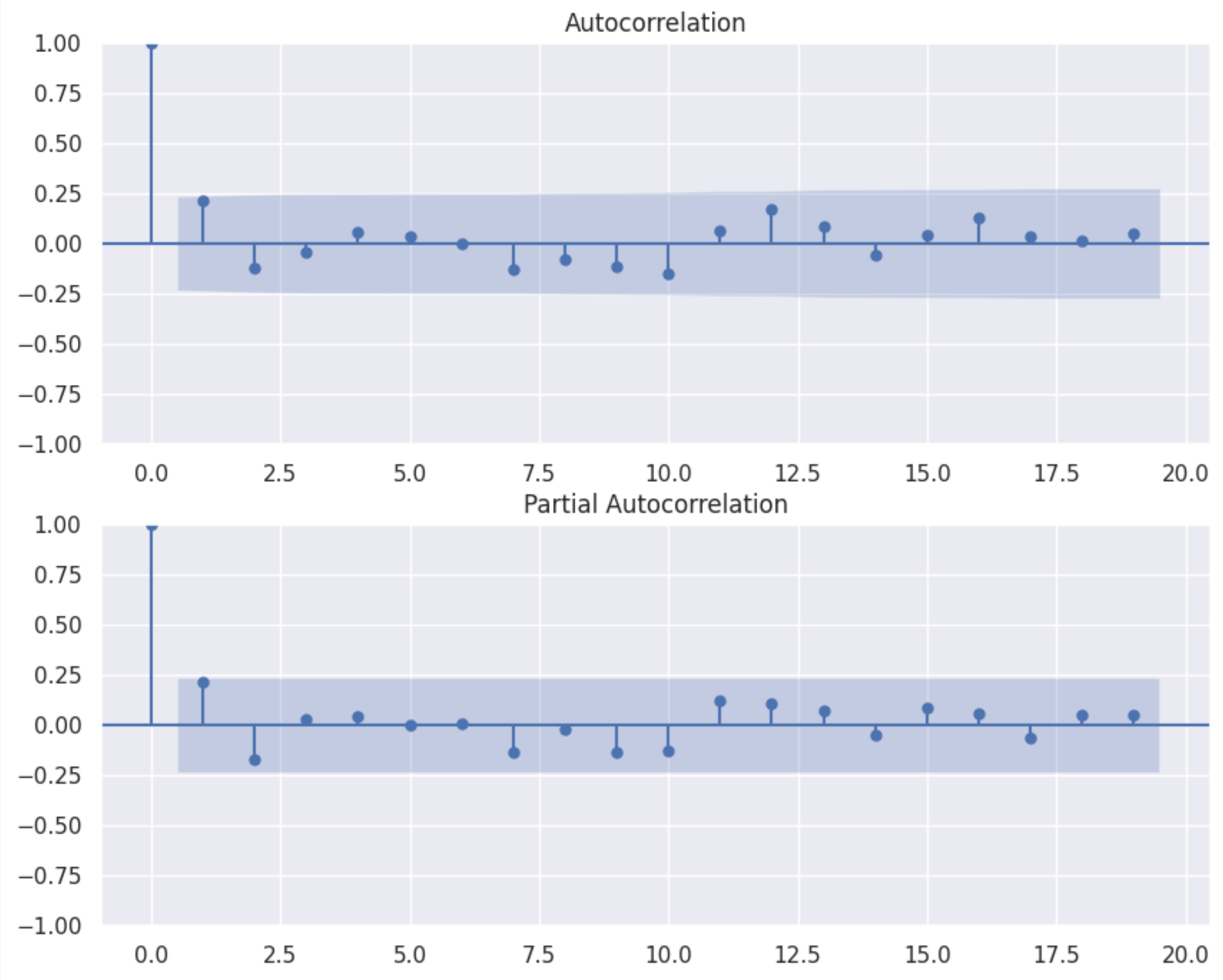


DATA PROCESSING

01. STATIONARITY

1. Log the data and then take the difference
2. Check the ADF test to ensure p-value < 0.05
3. Plot the data along with its ACF and PACF

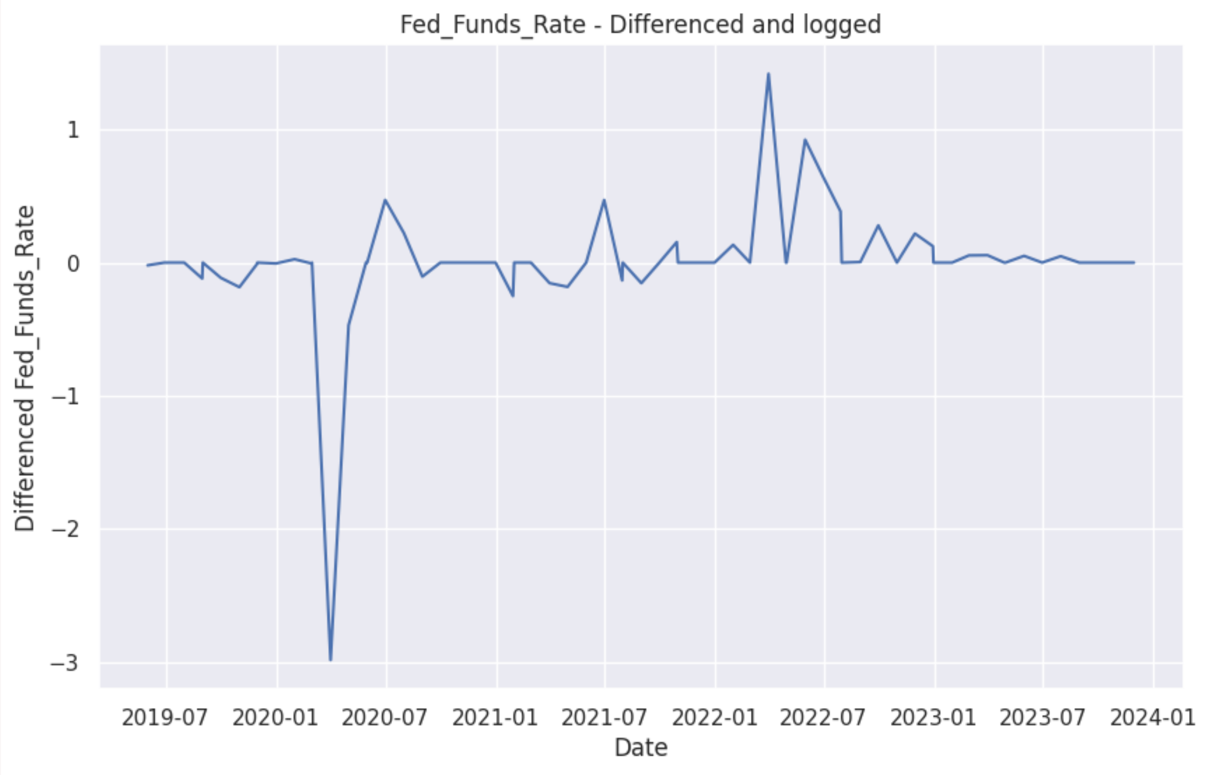


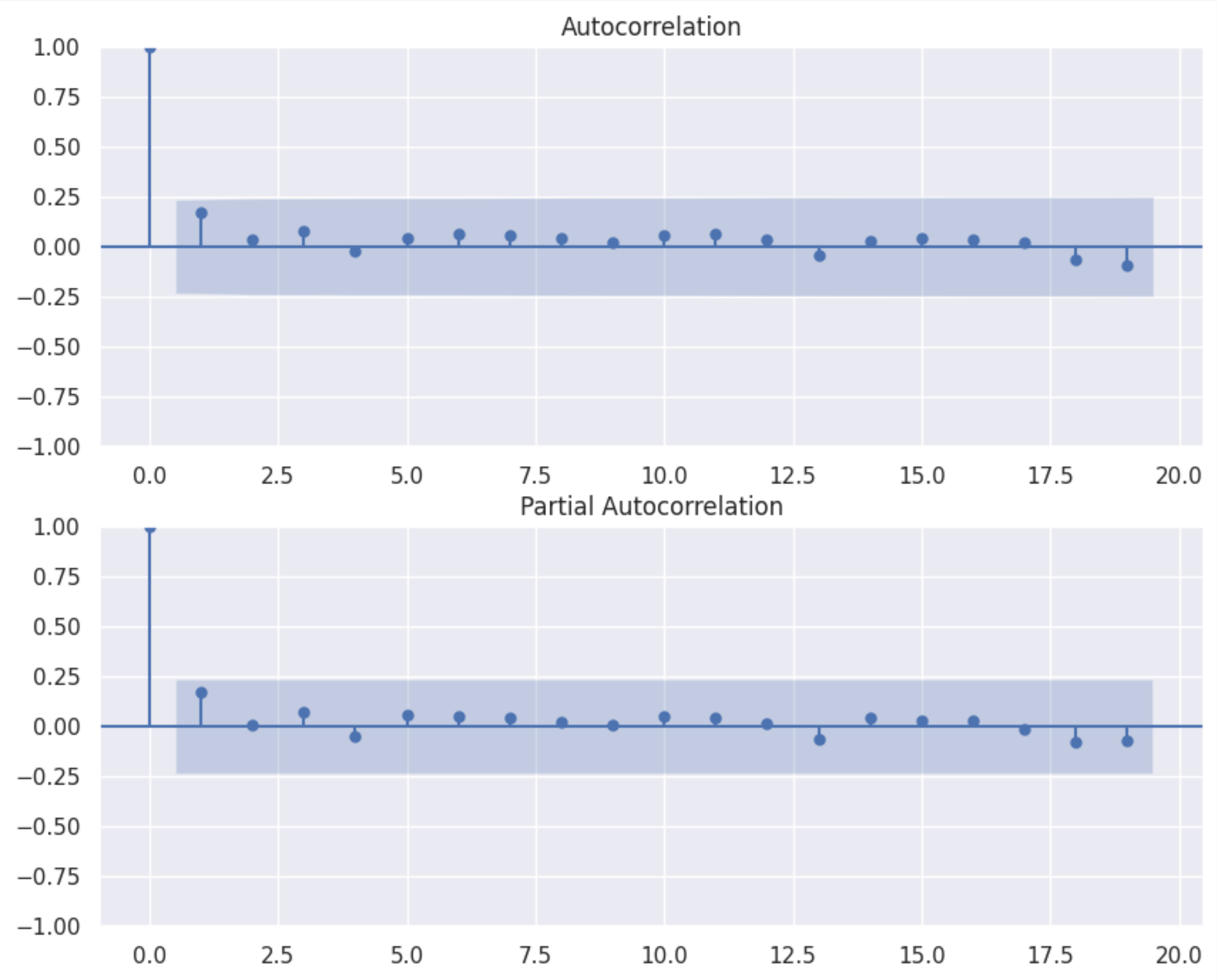



```
Check for Gold Stationarity after differencing and log
ADF Statistic: -6.140589
p-value: 0.000000
```

Dickey Fuller test: Unit root test indicates if there is any stochastic trend in the time series that drives it away from its mean value

If the test statistic is less than the critical value or if the p-value is less than a pre-specified significance level (e.g., 0.05), then the null hypothesis is rejected and the time series is considered stationary

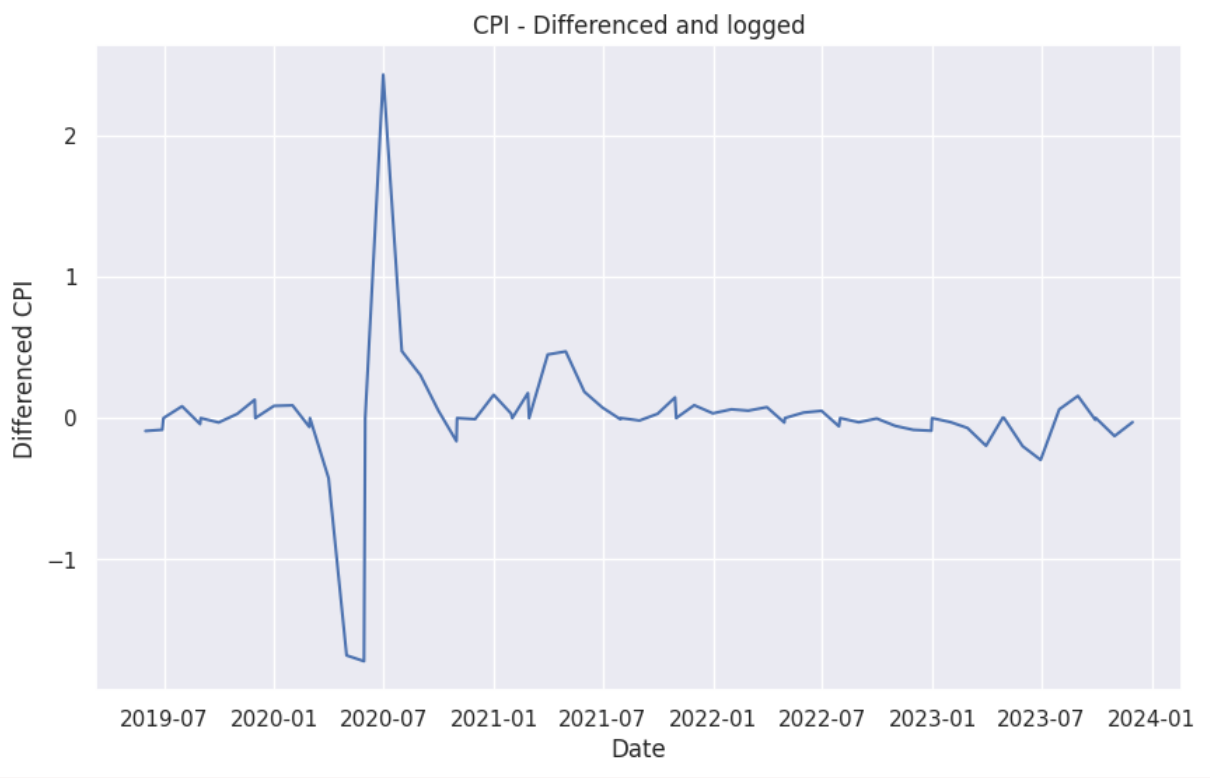


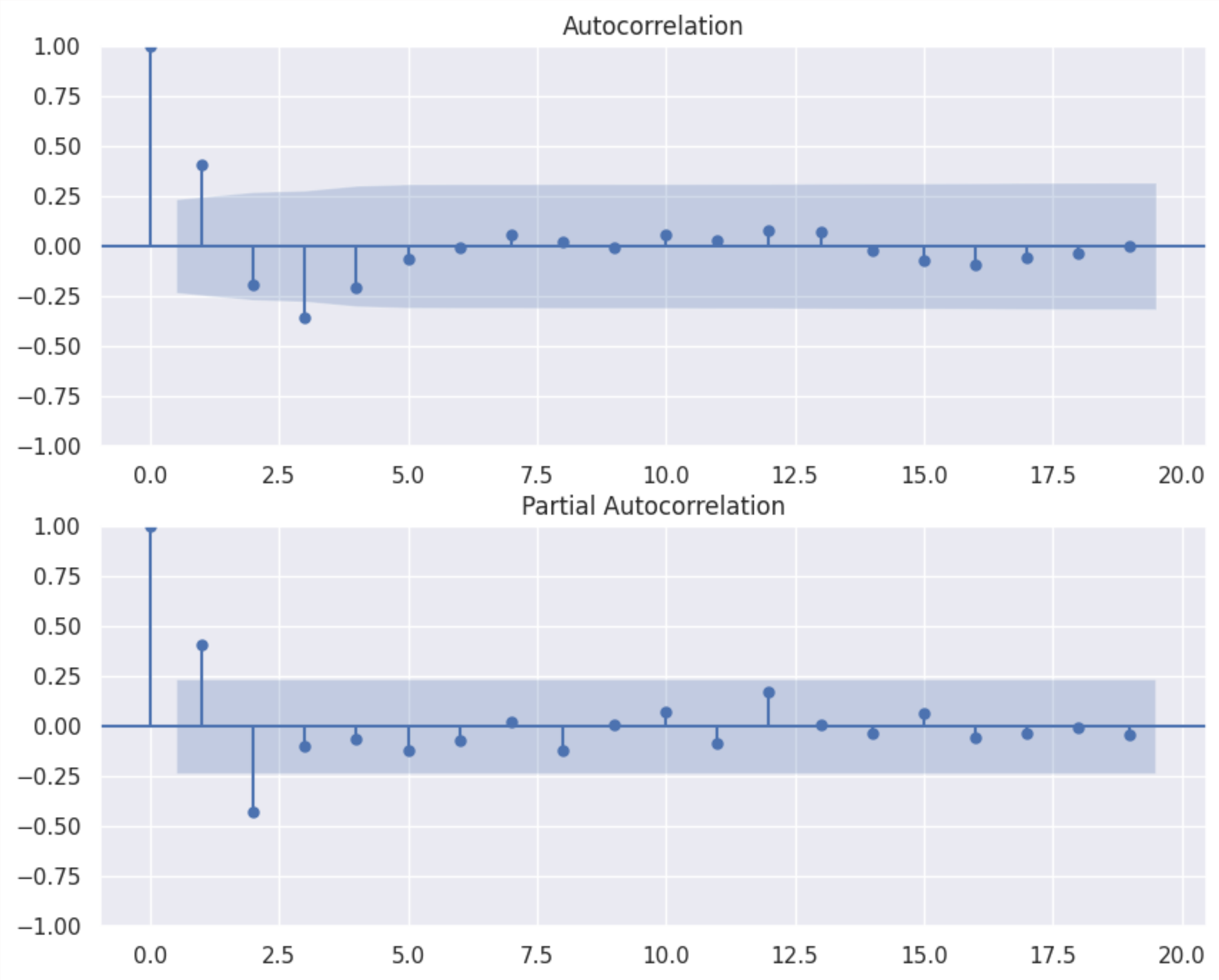


Index-1: Fed Funds Rate

- Differenced and Logged : Significant fluctuation around early 2020, likely reflecting policy responses to economic conditions
- ACF : The autocorrelation decreases rapidly and remains within a no significant autocorrelation after differencing and logging
- PACF Plot: Displays a sharp drop at the first lag and minimal thereafter

```
Check for Fed_Funds_Rate Stationarity after differencing
and log ADF Statistic: -6.869107
p-value: 0.000000
```





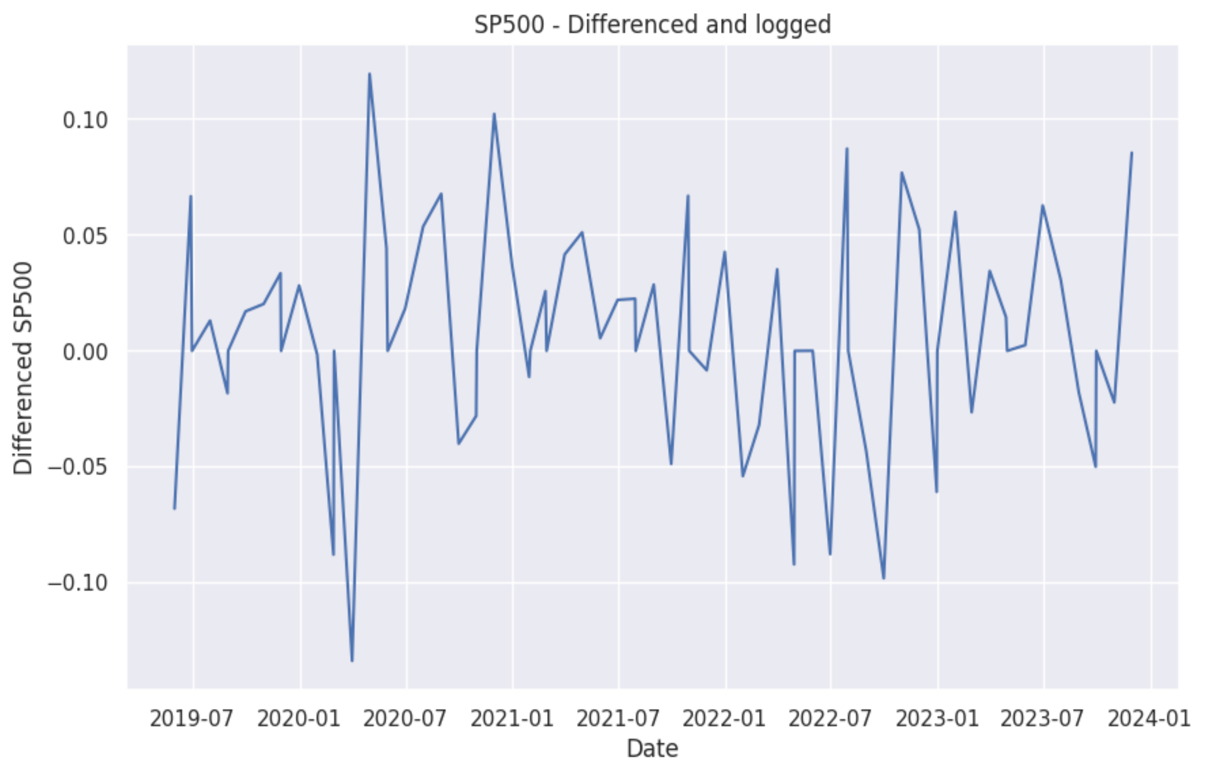
Index-2: CPI

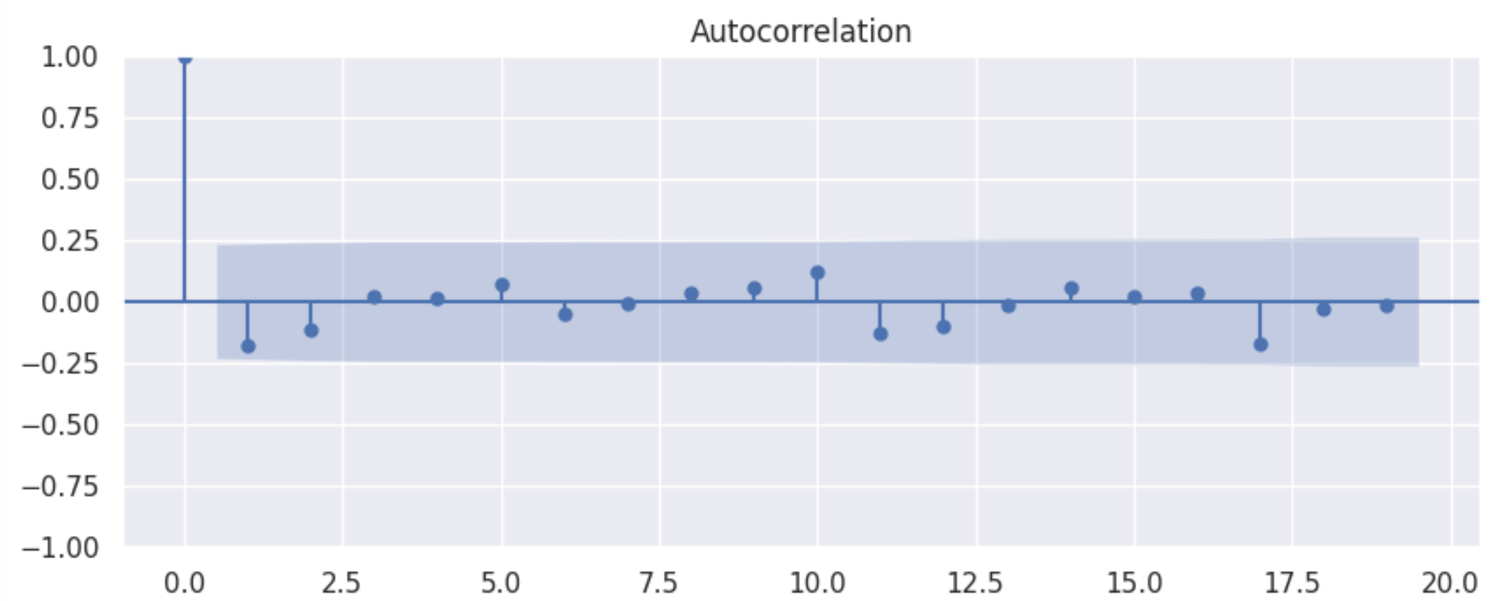
- Differenced and Logged : Sharp peak around early 2020, then stabilizes, which might indicate a response to external shocks, possibly inflationary pressures due to policy changes or economic events

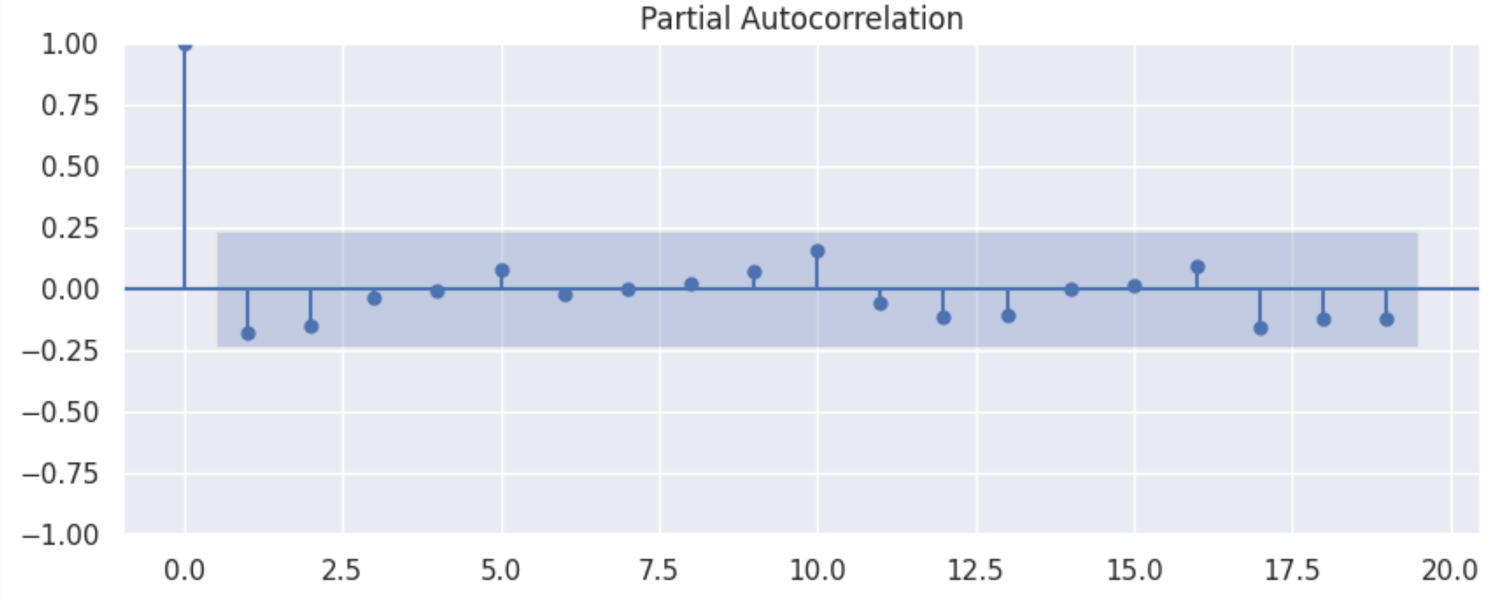
- ACF Plot: Slow decay in autocorrelation, implying that the series might still retain some memory of past values

- PACF Plot: Shows significance in the first few lags, which could suggest an AR process of order 1 or 2 might be appropriate

```
Check for CPI Stationarity after differencing and log
ADF Statistic: -6.965347
p-value: 0.000000
```



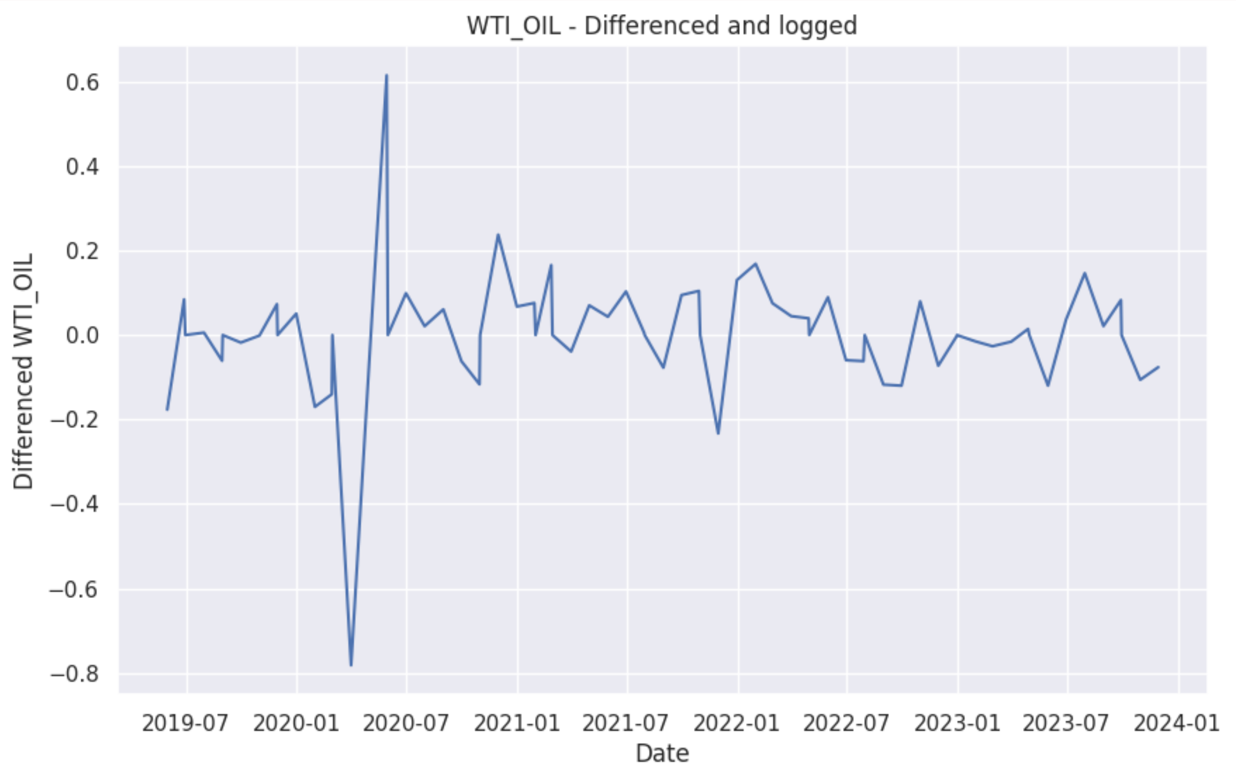


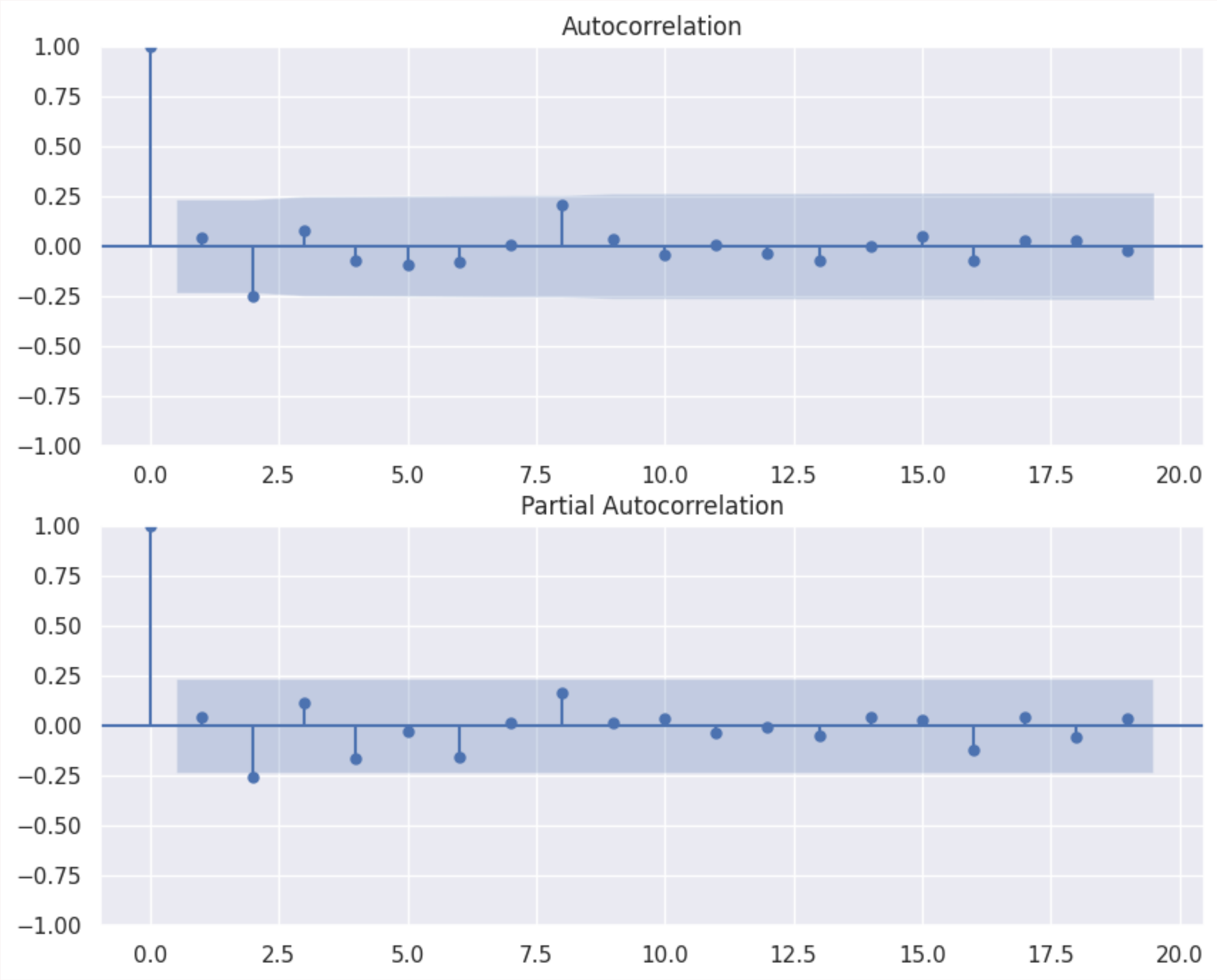


Index-3: S&P500

- Differenced and Logged : High volatility with noticeable peaks and troughs, characteristic of stock market indices
- ACF Plot: The autocorrelations are significant and slowly taper off, suggesting a long memory process
- PACF Plot: The initial significant lag and subsequent tapering off hint that an AR(1) model might be appropriate

```
Check for SP500 Stationarity after differencing and log
ADF Statistic: -9.853761
p-value: 0.000000
```





Index-4: WTI OIL

- Differenced and Logged : Similar to S&P 500, shows high volatility, likely due to external economic impacts on oil prices
- ACF Plot: Autocorrelation slowly decays
- PACF Plot: Significant at the first lag, suggesting an AR(2) process may be a good starting model

```
Check for WTI_OIL Stationarity after differencing and
log ADF Statistic: -7.220938
p-value: 0.000000
```

STEP-4

Modeling

For modeling, we utilize 80/20 train test split for fit our model

MODEL-1 ARIMAX

ARIMAX stands for

AutoRegressive Integrated Moving Average with eXogenous variables

It is an extension of the ARIMA model

that includes exogenous variables (predictors) to improve forecasting accuracy

What is ARIMAX?

AR (AutoRegressive) Component (p): The number of lag observations included in the model (lag order).

Example: If p=2, the model uses the last two observations to predict the current value.

I (Integrated) Component (d): The number of times the data needs to be differenced to achieve stationarity.

Example: If d=1, the model uses the first difference of the series.

MA (Moving Average) Component (q): The number of lagged forecast errors included in the model (order of the moving average).

Example: If q=2, the model uses the last two forecast errors to predict the current value.

X (Exogenous Variables): These are additional external predictors that can influence the time series.

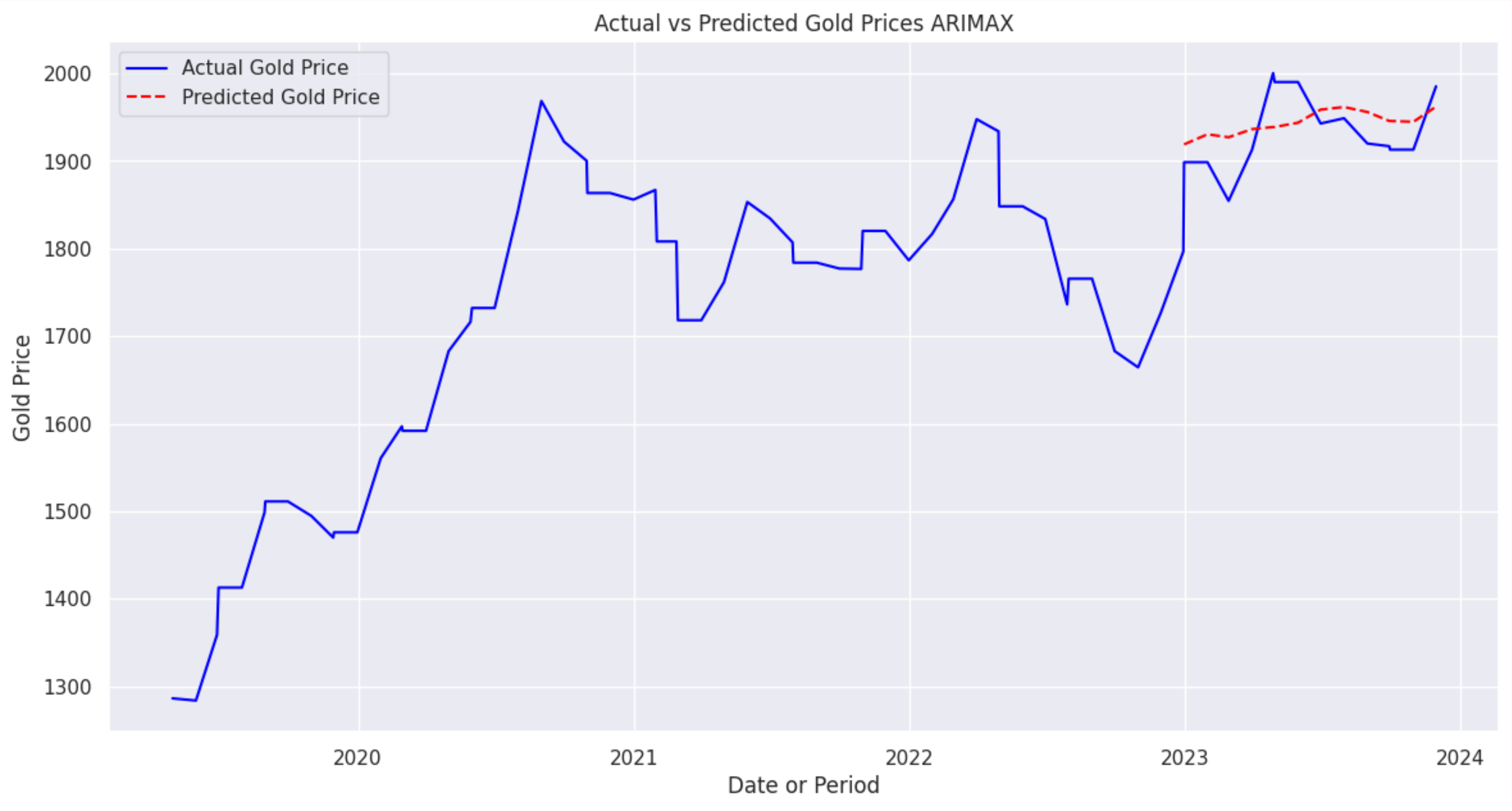
MODEL-1
ARIMAX

```python
combined_df_diff.dropna(inplace=True)

train, test = train_test_split(combined_df_diff, shuffle=False, test_size=0.2)

model=pm.auto_arima(train['Gold_Price_diff'],start_p=0,d=0,start_q=0,
          max_p=5,max_d=5,max_q=5, start_P=0,
          D=0, start_Q=0, max_P=5,max_D=5,
          max_Q=5, seasonal=False,
          error_action='warn',trace=True,
          supress_warnings=True,
          random_state=20,n_fits=200)
```

```
Performing stepwise search to minimize aic
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=-235.325, Time=0.08 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=-236.641, Time=0.02 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=-237.691, Time=0.06 sec
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=-235.831, Time=0.10 sec
 ARIMA(0,0,2)(0,0,0)[0]             : AIC=-235.901, Time=0.05 sec
 ARIMA(1,0,2)(0,0,0)[0]             : AIC=-233.947, Time=0.09 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=-237.259, Time=0.13 sec

Best model:  ARIMA(0,0,1)(0,0,0)[0]
Total fit time: 0.543 seconds
```

- Model Fitting with pm.auto_arima, automatically discovers the best parameters for an ARIMA model
- ARIMA(0,0,1)(0,0,0)[0] is the best among tested models with
  - Lowest AIC (-237.691)
  - Suggests using a simple MA(1) model
  - Without differencing (d=0)
  - Without any seasonal components
  - Indicates that only previous error influences current value

MODEL-1 ARIMAX




```
Mean Absolute Error (MAE): 34.93558811104661
Mean Squared Error (MSE): 1500.966767139674
Root Mean Squared Error (RMSE): 38.74231236180507
```

- ARIMAX model captured upward trend, anticipates an increase in gold prices based on the learned patterns and exogenous factors included in the model
- MAE and RMSE are relatively close, indicating a reasonable prediction error when considering the price range of gold
- MSE is larger than MAE and RMSE, suggests there are instances where the model errors are quite large
- Overall, the ARIMAX model appears to perform adequately with a typical error around $35 to $39

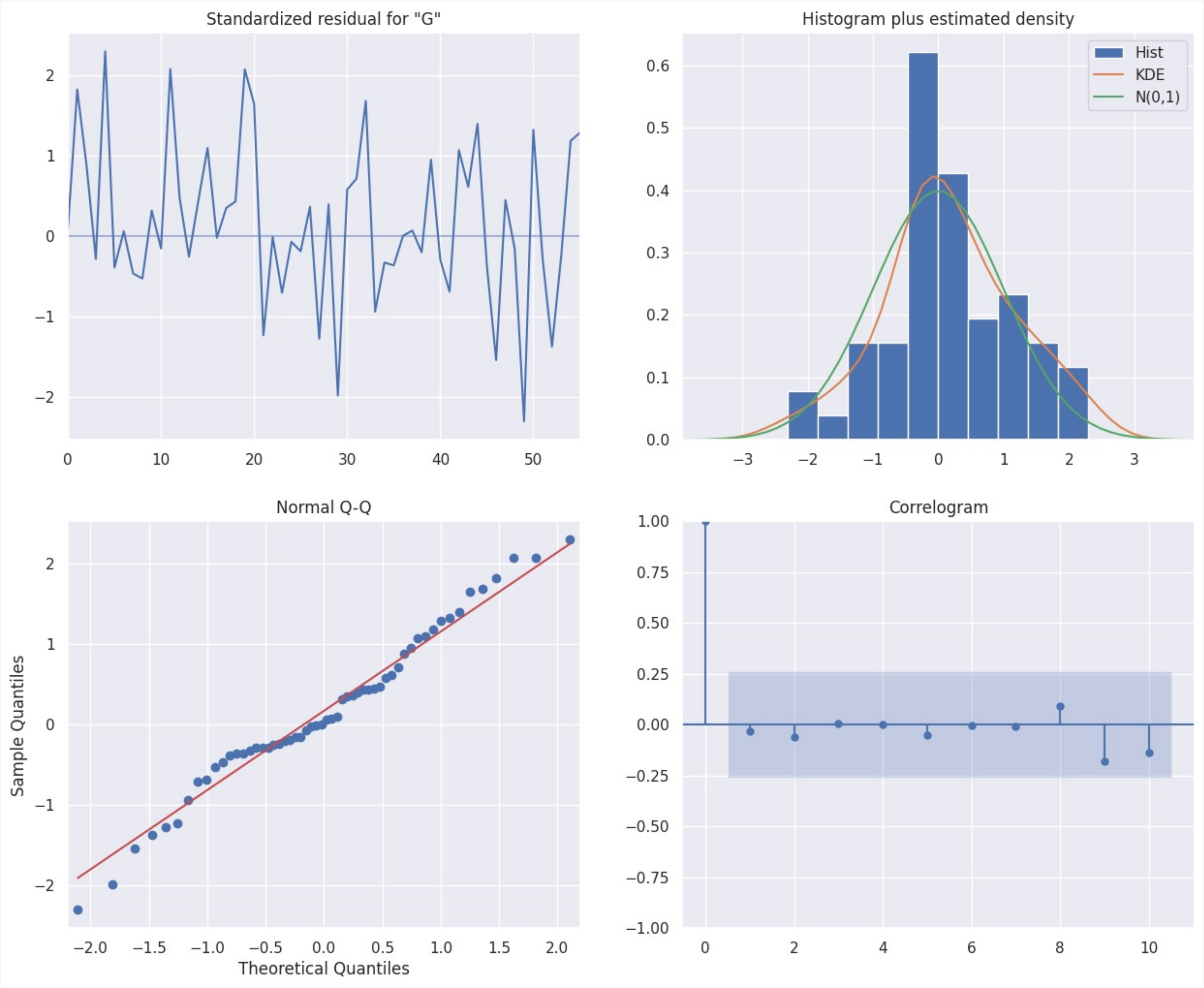
INDEPENDENCE OF ERRORS

Independence of Errors: The errors (residuals) from the model are not correlated with each other

Normality of Errors: Assuming that the errors are normally distributed

Zero Mean of Errors: The assumption that the errors have a mean of zero implies that the model is unbiased

MODEL-1 ARIMAX



- Standardized Residuals: Appear to be randomly distributed, but have some negative trend
- Histogram and Estimated Density: Residuals are roughly normally distributed but with some deviations.
- Normal Q-Q Plot: Residuals mostly follow the normal distribution line, with some deviations at the tails.
- Correlogram: No significant autocorrelation in the residuals.

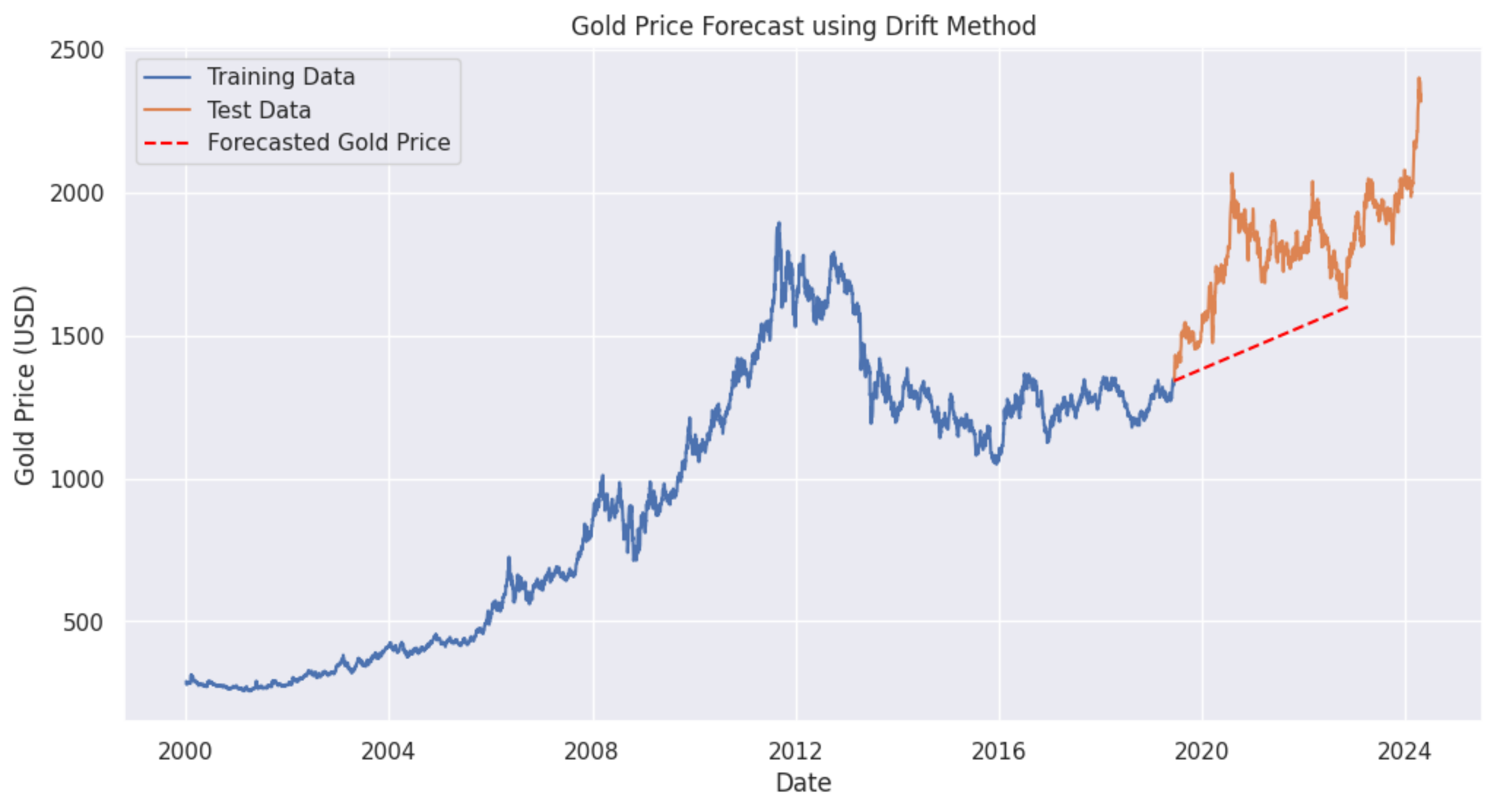
`Ljung-Box Test: 0.07`

- Null hypothesis (no autocorrelation) failed to be rejected, it suggests that there is some autocorrelation in the residuals of ARIMAX model

MODEL-2

DRIFT

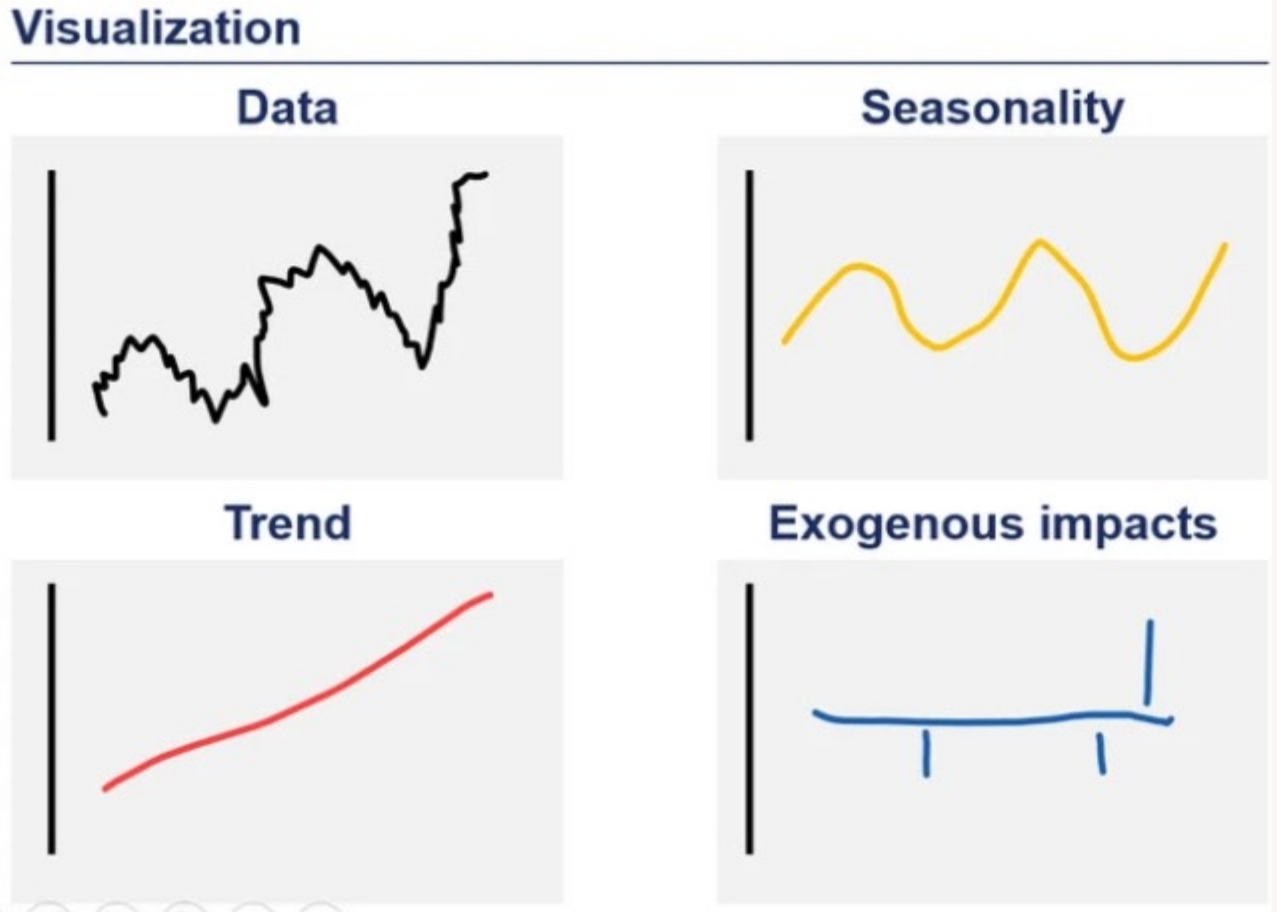
Assumes that the forecast for the next period is equal to the last observed value plus the average change (or drift) over the entire historical period



- MAE, MAPE, and RMSE value are relatively high
- Though drift can capture the general direction of movement, it does not account adequately for the volatility and the factors that might cause fluctuations

```
Mean Absolute Error (MAE): 336.07
Mean Absolute Percentage Error (MAPE): 18.08%
Root Mean Squared Error (RMSE): 360.54
```

MODEL-3 PROPHET



Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data.

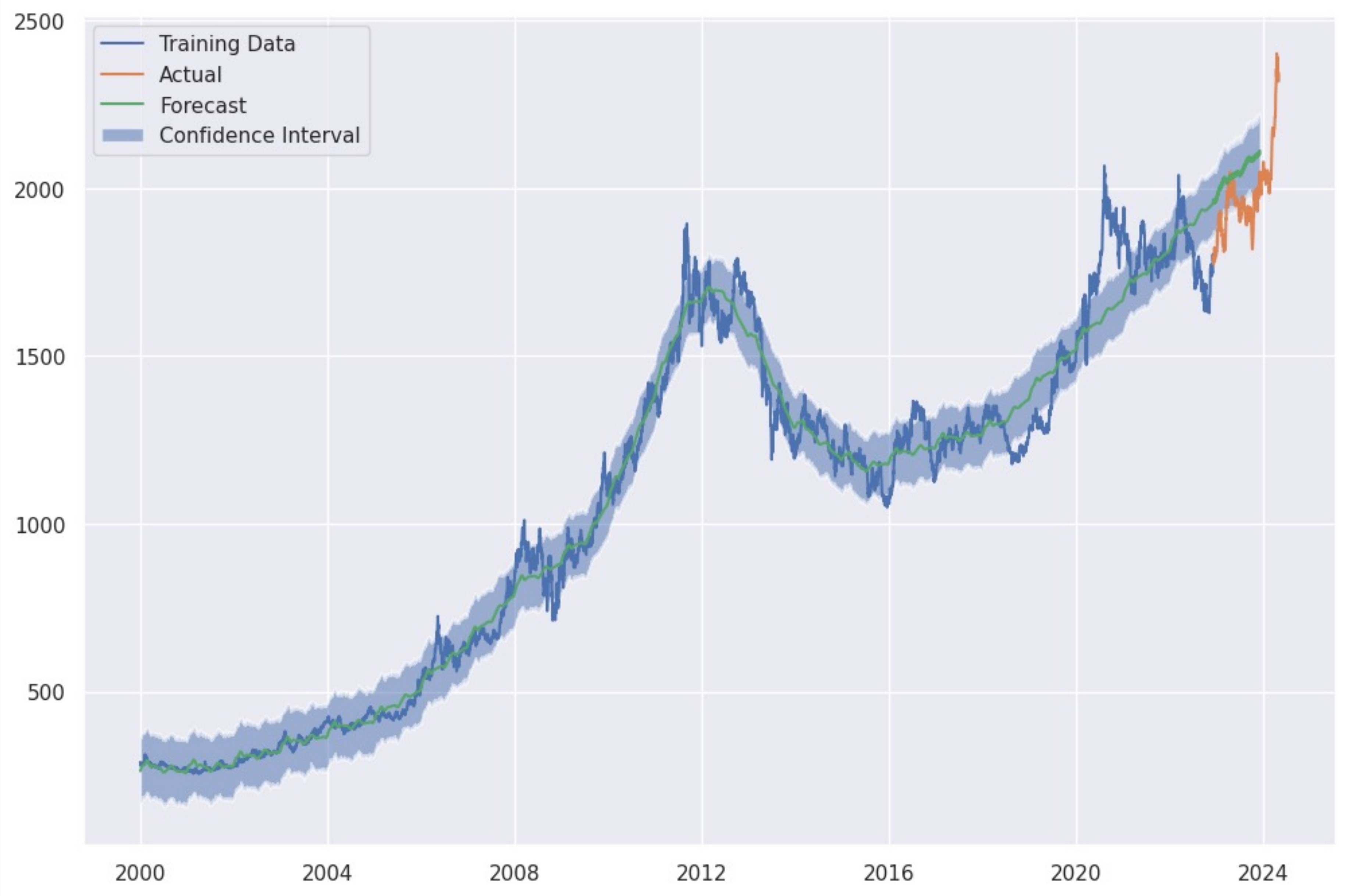
Methodological framework

$$y(t) = c(t) + s(t) + h(t) + x(t) + \epsilon$$

Where:

c(t)	Trend +
s(t)	Seasonality +
h(t)	Holiday effects +
x(t)	External regressors +
e	error

MODEL-3 PROPHET



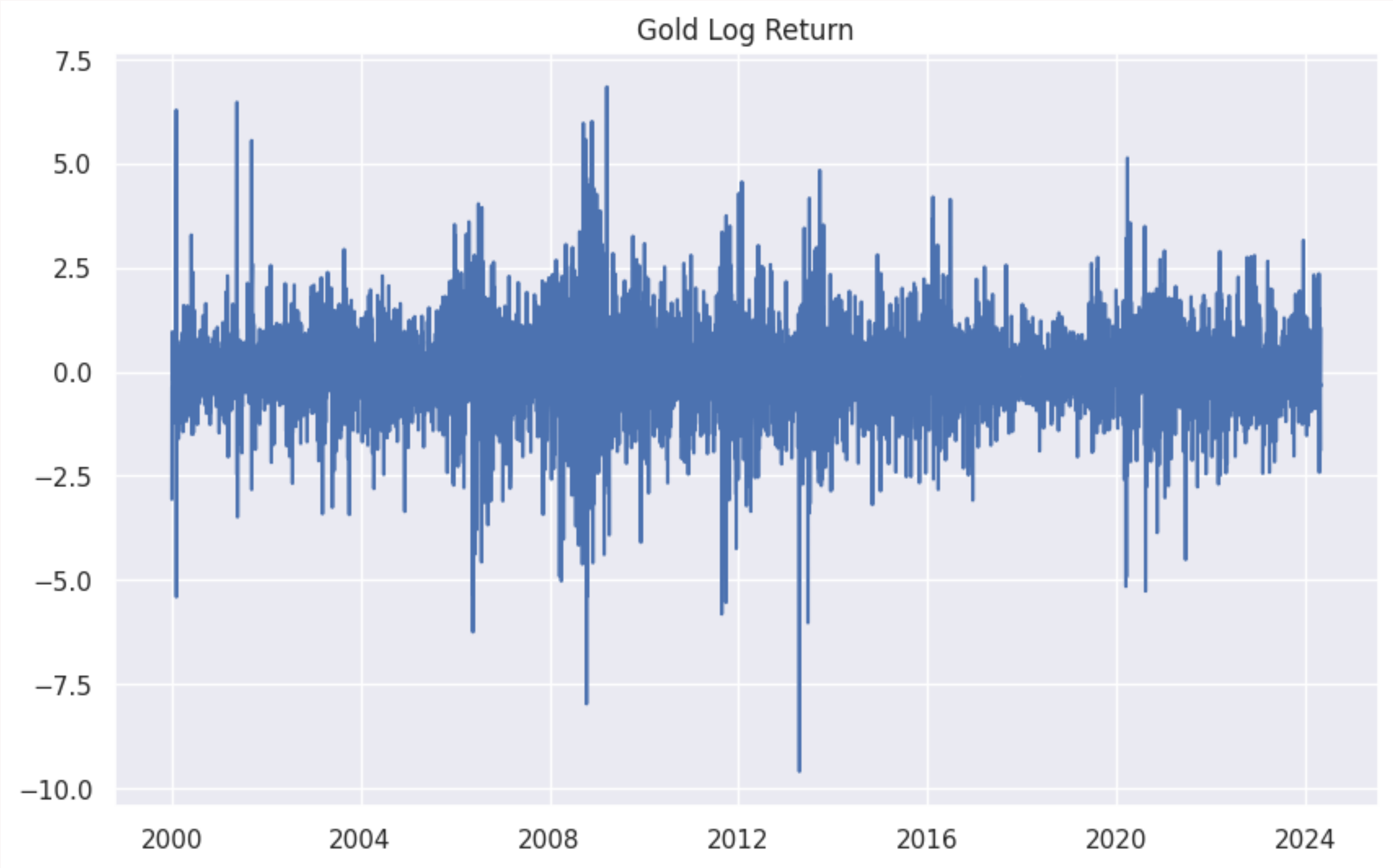

```
MAE: 96.21898888953442
RMSE: 112.19681650299636
MAPE: 4.895163998189626%
```

Observations:

- Alignment: Initial close alignment between forecast and actual values
- Divergence: Noticeable divergence towards the end of the test period, with actual values exceeding forecast and confidence intervals
- Confidence Interval: Narrow initially, widens towards the end, indicating increased uncertainty
- Trends & Seasonality: Long-term trends and seasonal patterns well-captured

MODEL-4
ARCH / GARCH

An approach to estimating the volatility of financial markets




```
McLeod-Li test statistic: 5719.68410349655
p-value: 0.0
```

McLeod-Li Test:

Tested for ARCH effects

Result: Significant ARCH effects

Model Selection:

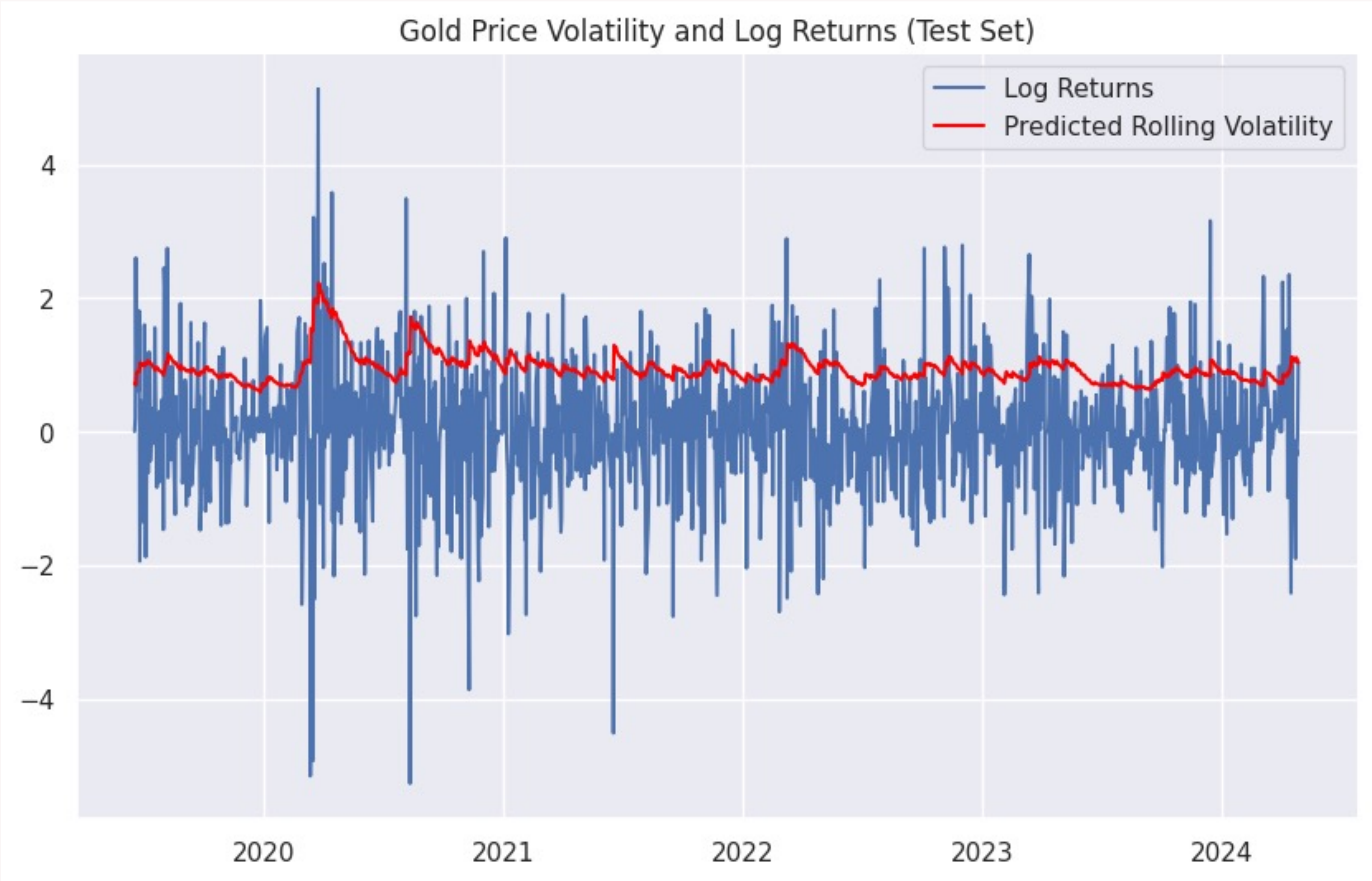
Criterion: Akaike Information Criterion (AIC)

Best Model: ARCH(1) with an AIC of 14788.245

What is ARCH(1): Suppose you have a time series of daily returns on a stock. The ARCH(1) model could be used to model the changing volatility of these returns. If a large return (positive or negative) occurs on one day, the model would predict higher volatility for the next day, reflecting the increased uncertainty.

MODEL-4
ARCH / GARCH

```
Jarque-Bera test statistic: 8199.073899191371
p-value: 0.0
```



Jarque-Bera Test:

Tested for normality of residuals

Result: Non-normal residuals

Rolling Predictions:

Generated rolling predictions for test set

```
RMSE: 1.332397072065138
MAE: 1.070060158518971
```

STEP-5

Conclusion

MODEL RECAP

ARIMAX

- Combines ARIMA with external predictors to enhance forecasting accuracy
- Captures linear relationships and trends in the time series data
- Effective for time series with additional influencing factors

DRIFT

- A simple model that extends the last observed value by a constant drift
- Assumes a linear trend over time
- Useful as a benchmark for more complex models

PROPHET

- Incorporates holidays and other seasonality patterns to improve forecast accuracy
- Can handle gaps and outliers/ nan in the data without significant loss of accuracy

ARCH/GARCH

- Focuses on modeling and forecasting time-varying volatility
- Captures clustering in volatility
- Essential for understanding and predicting market risk

EFFICIENT MARKET HYPOTHESIS

Financial markets are "informationally efficient," meaning that asset prices fully reflect all available information at any given time. It is impossible to consistently achieve higher returns than average market returns on a risk-adjusted basis, since asset prices should only respond to new information, which is inherently unpredictable. When applied to the gold market, this hypothesis suggests that all known factors influencing gold prices, such as economic indicators, geopolitical events, and market sentiment, are already incorporated into current prices.

THANK YOU SO MUCH