Gold Price Prediction

Rhea Mall (509), Akriti Arora (515), Anjuli Jain (548),
Parul Nagar (945), Maitreyee Katre (967)

Contents

Introduction

- The World Gold Council report in 2022 stated that Indian households own a gold stock of about **24000** to **25000** metric tons.
- According to another report, India's Gems and jewellery industry contributes to around **7%** of its GDP and constitutes **13.30%** of total merchandising exports.
- Gold is looked upon as one of the best options when it comes to security and savings for a good percentage of the Indians.

Even lawyers, bankers, politicians and others purchase gold jewellery during festivals or special occasions. They consider it a fool-proof financial strategy, and since the real estate and capital markets are losing promise off late, the wealthy section of Indian society is now a new class of gold investors on its own. As well as direct and indirect jobs, gold mining also brings foreign direct investment and tax revenues to countries.

Problem Statement

In recent times, the price of gold has become a hot topic for everyone.

All investors are very keen to know about the trend of gold prices - whether the prices will rise or fall.

In this study we propose a time series model for forecasting the monthly average price of 10g gold in India.

Data Description

- The data has been obtained from the official RBI website.
- The data consists of the monthly average prices (in Indian Rupees) per 10 grams of gold, over the time period April 2018 to September 2022.

Methodology

Softwares used: We performed the analysis using SPSS and R.

Check if the variable is stationary or not. If not, then make it stationary by differencing.

Select possible values of the parameters p, d, q by using Correlogram of ACF and PACF.

Estimate the parameters of selected models and by using values of AIC select the best fitted model.

By using the best fitted model, forecast the next values of the variable.

ARIMA Model

ARIMA is an acronym that stands for **A**uto**R**egressive **I**ntegrated **M**oving **A**verage. This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

- **AR**: Autoregressive. A model that uses the dependent relationship between an observation and some number of lagged observations.
- **I**: Integrated. The use of differencing of raw observations (i.e. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
- **MA**: Moving Average. A model that uses the dependency between an observation and residual errors from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter.

A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used. The parameters of the ARIMA model are defined as follows:

p	d	q
The number of lag observations included in the model, also called the lag order.	The number of times that the raw observations are differenced, also called the degree of differencing.	The size of the moving average window, also called the order of moving average.

The equation for the model is given by:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Box-Jenkins Method

The Box-Jenkins methos refers to the iterative application of the following three steps:

1 Identification

Using the plot of data, autocorrelations, partial autocorrelations, and other information, a class of simple ARIMA model is selected. This amounts to estimating appropriate values for p, d, and q.

2 Estimation

The phis and thetas of the selected model are estimated using MLE techniques etc.

3 Diagnostic Checking

The fitted model is checked for inadequacies by considering the autocorrelations of the residual series (the series of residuals, or error, values).

Autocorrelation Function

The stationary assumption allows us to make simple statements about the correlation between two successive values, X_t and X_{t+k} . This correlation is called the autocorrelation of lag k of the series. The autocorrelation function displays the autocorrelation on the vertical axis for successive values of k on the horizontal axis.

Partial Autocorrelation Function

The Partial Autocorrelation Function is a second function that expresses information useful in determining the order of an ARIMA model.

This function is constructed by calculating the partial correlation between X_t and X_{t-1} , X_t and X_{t-2} , and so on, statistically adjusting out the influence of intermediate lags. For example, the partial autocorrelation of lag four is the partial correlation between X_t and X_{t-4} after statistically removing the influence of X_{t-1} , X_{t-2} , and X_{t-3} from both X_t and X_{t-4} . The autoregressive order, p, is estimated as the lag of the last large partial autocorrelation.

Assumptions

The assumptions of the ARIMA model are as follows:

The data should be stationary – by stationary it means that the properties of the series doesn't depend on the time when it is captured.

Data should be univariate – ARIMA works on a single variable.

Analysis & Discussion

Steps

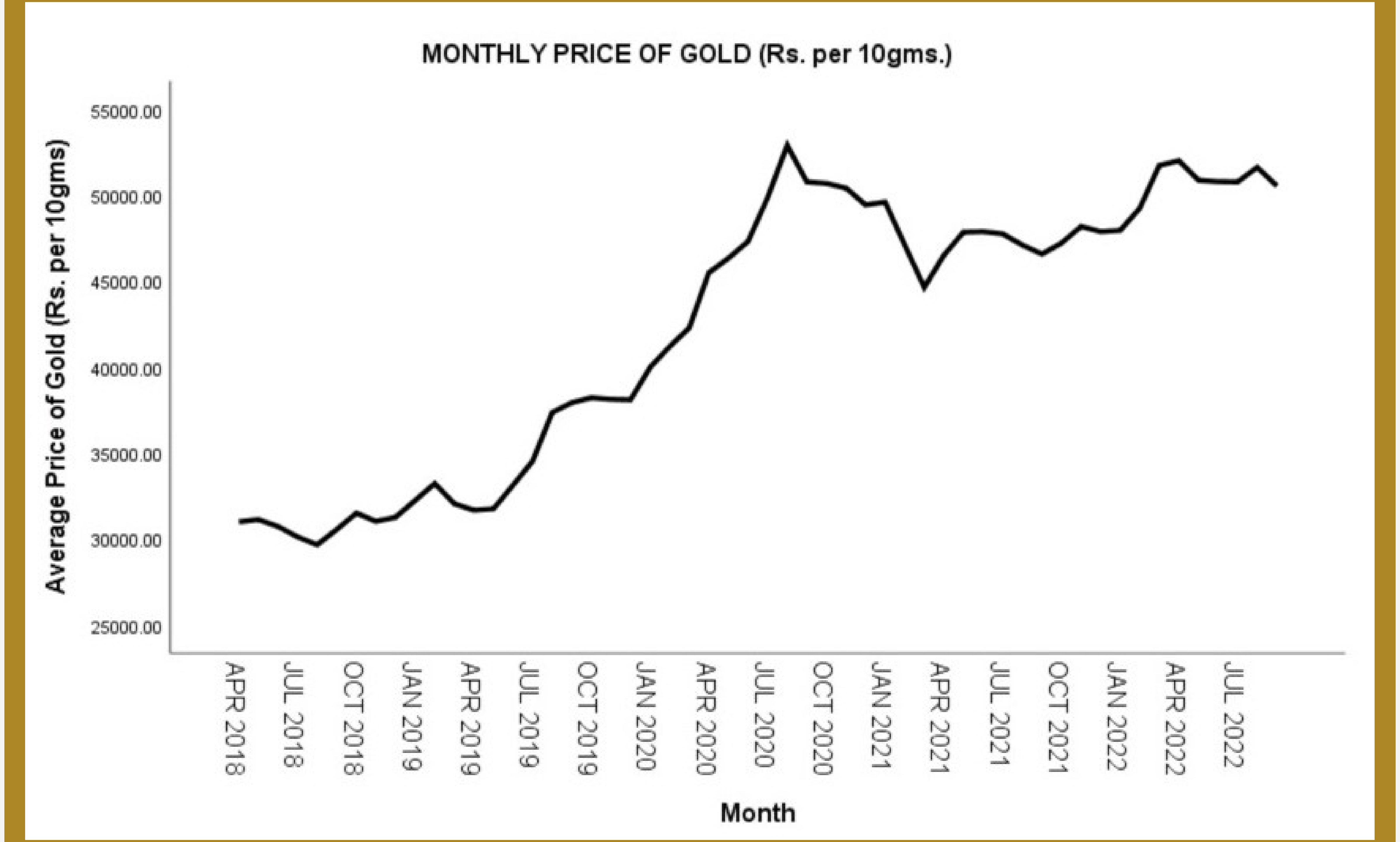
For testing the null hypothesis "data is not stationary", we use Augmented Dickey-Fuller test (ADF). If the significant value of ADF test is less than desired level of significance which is mostly 0.05, we conclude that data is stationary.

The linear relationship between two values of same variable is termed as autocorrelation (ACF). If we calculate ACF after removing any linear dependency from the lag values, it will be called partial autocorrelation (PACF). ACF and PACF both range from -1 to +1. In correlogram (graphical representation) ACF shows the order of moving average model (q) and PACF shows the order of autoregressive model (p).

By using Akaike Information Criterion (AIC), best model selected is selected in Box-Jenkins approach. The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.

By using best fitted model, the average monthly gold prices for next months are forecast.

We plot the given data of average monthly Gold prices in India



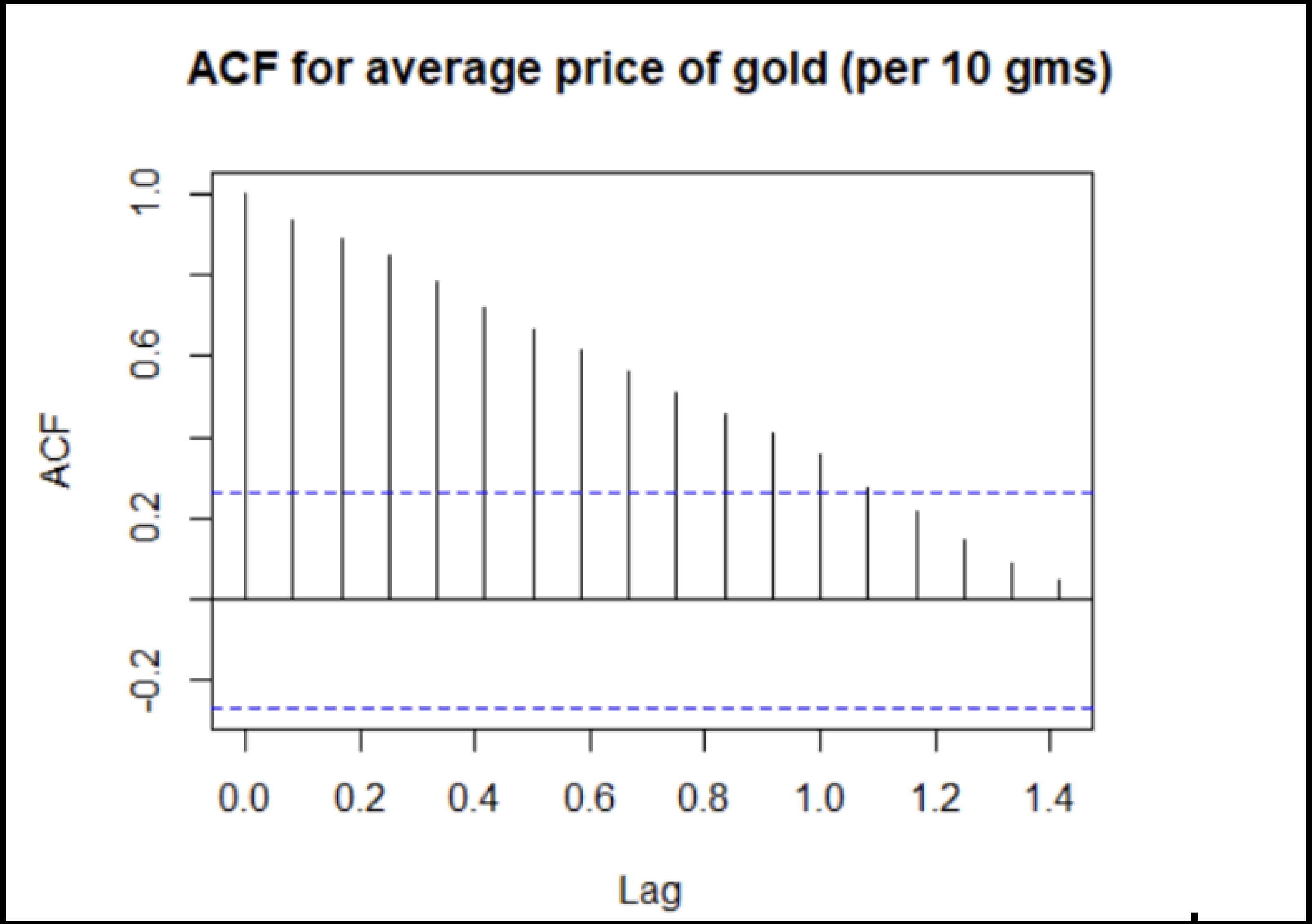
Checking for Seasonality

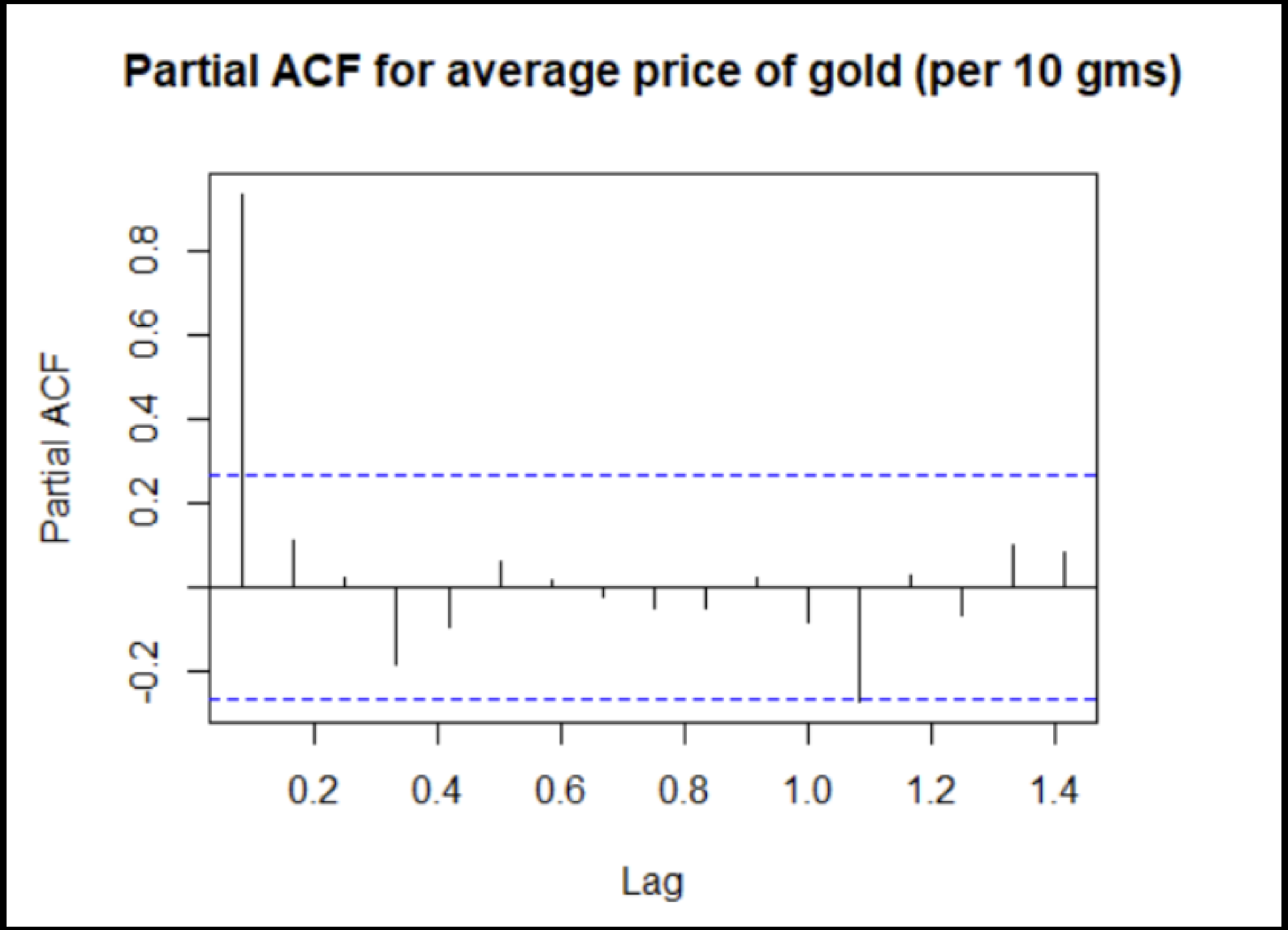
```
t_s <- read.csv("ts1proj.csv",header=FALSE)
ts_ <- ts(t_s,start=c(2018,4),end=c(2022,9),frequency=12)

isSeasonal(ts_)   #FALSE
```

From the above test we can see that seasonality is absent in the data we have taken.

Autocorrelation Graphs

To check the data for stationarity, we plot ACF and PACF graph





Checking the data for Stationarity

We apply the Augmented Dickey-Fuller Test to the data to check for stationarity.

```
adf.test(ts_)
Ho: the data is not stationary
H1: the data is stationary

OUTPUT
Augmented Dickey-Fuller Test

data:  ts_
Dickey-Fuller = -1.5691, Lag order = 3, p-value = 0.748
alternative hypothesis: stationary
```

The p-value > 5% which means we have insufficient evidence to reject Ho and hence the conclusion is that data is non-stationary.

Since the data is not stationary, we thus use differencing to get stationary data.

Differencing the Data

DIFFERENCING THE DATA **ONCE**

```
ts1 <- diff(ts_,lag=1,differences = 1)

adf.test(ts1)
```

OUTPUT

Augmented Dickey-Fuller Test

data: ts1
Dickey-Fuller = -2.9786, Lag order = 3, p-value = 0.1806
alternative hypothesis: stationary

The p-value > 5% which means we have insufficient evidence to reject Ho and hence the conclusion is that data is non-stationary.

Differencing once more

```
adf.test(ts1)
```

Augmented Dickey-Fuller Test

data: ts2
Dickey-Fuller = -5.4467, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary

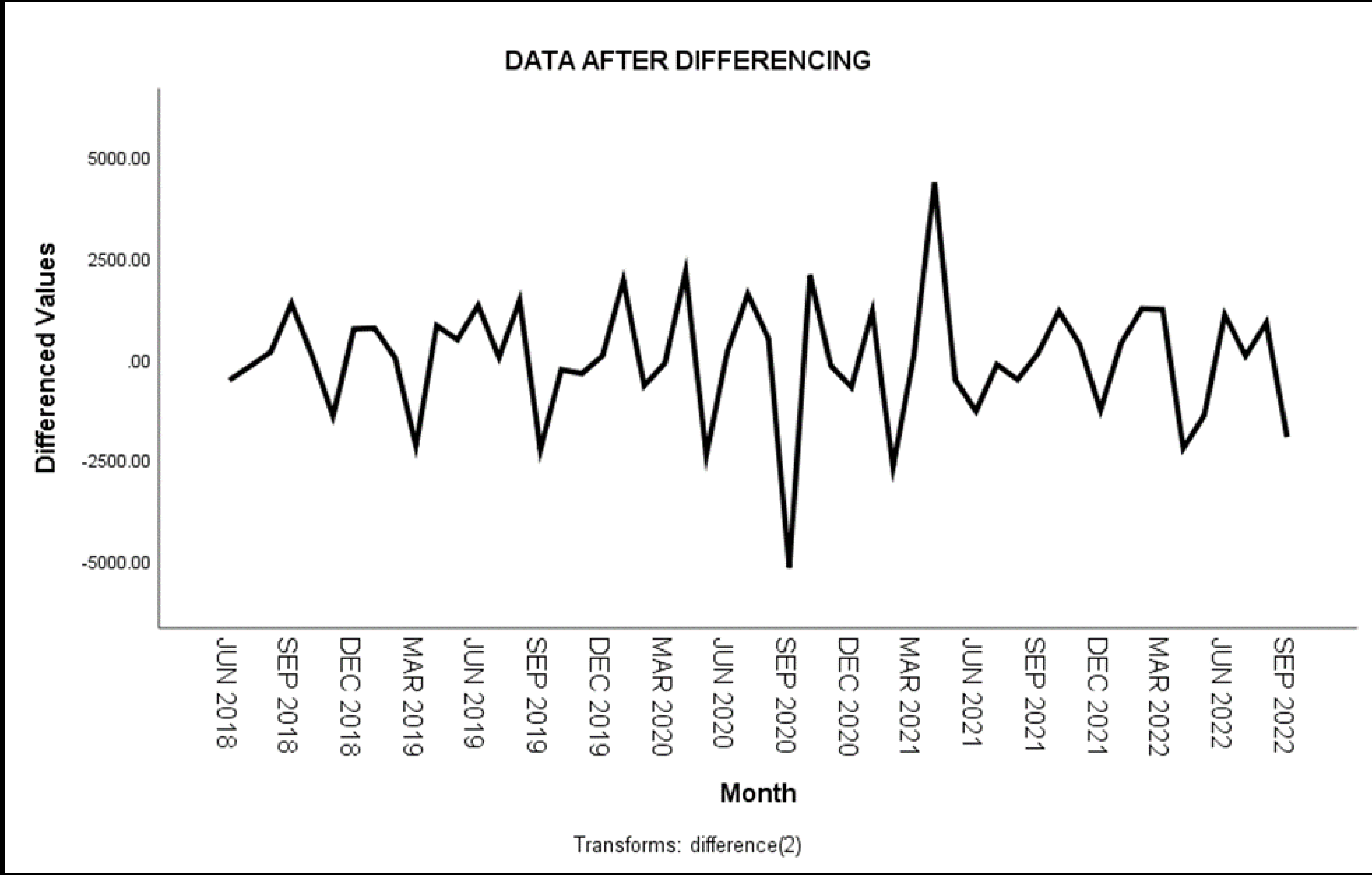
Data becomes stationary now.

After differencing the data, we repeat the Augmented Dickey-Fuller Test. Since the p-value is still greater than 5% we conclude that data is non-stationary.

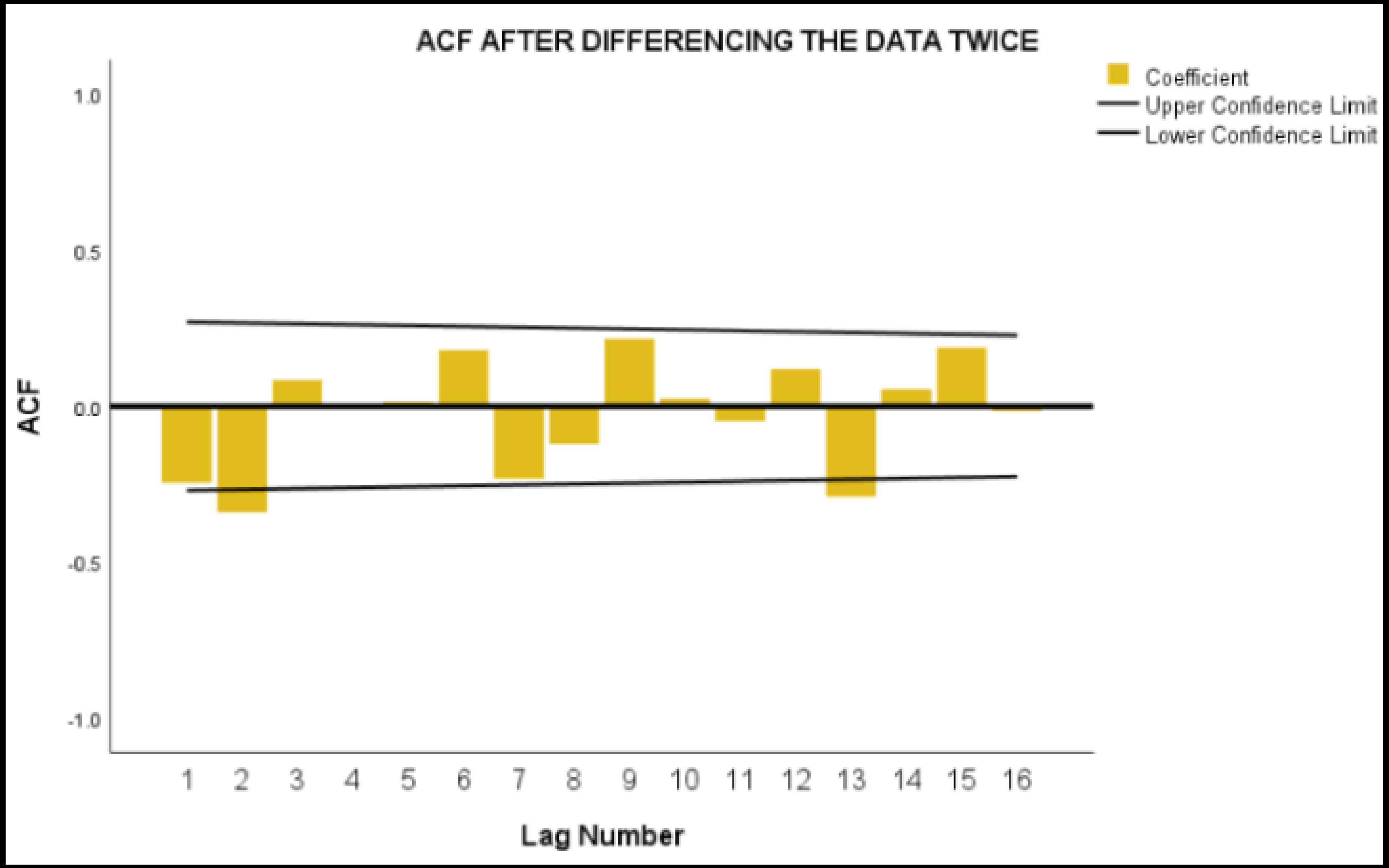
We difference the data one more time and run the Augmented Dickey-Fuller Test. Here, the p value comes out to be less than 5%.

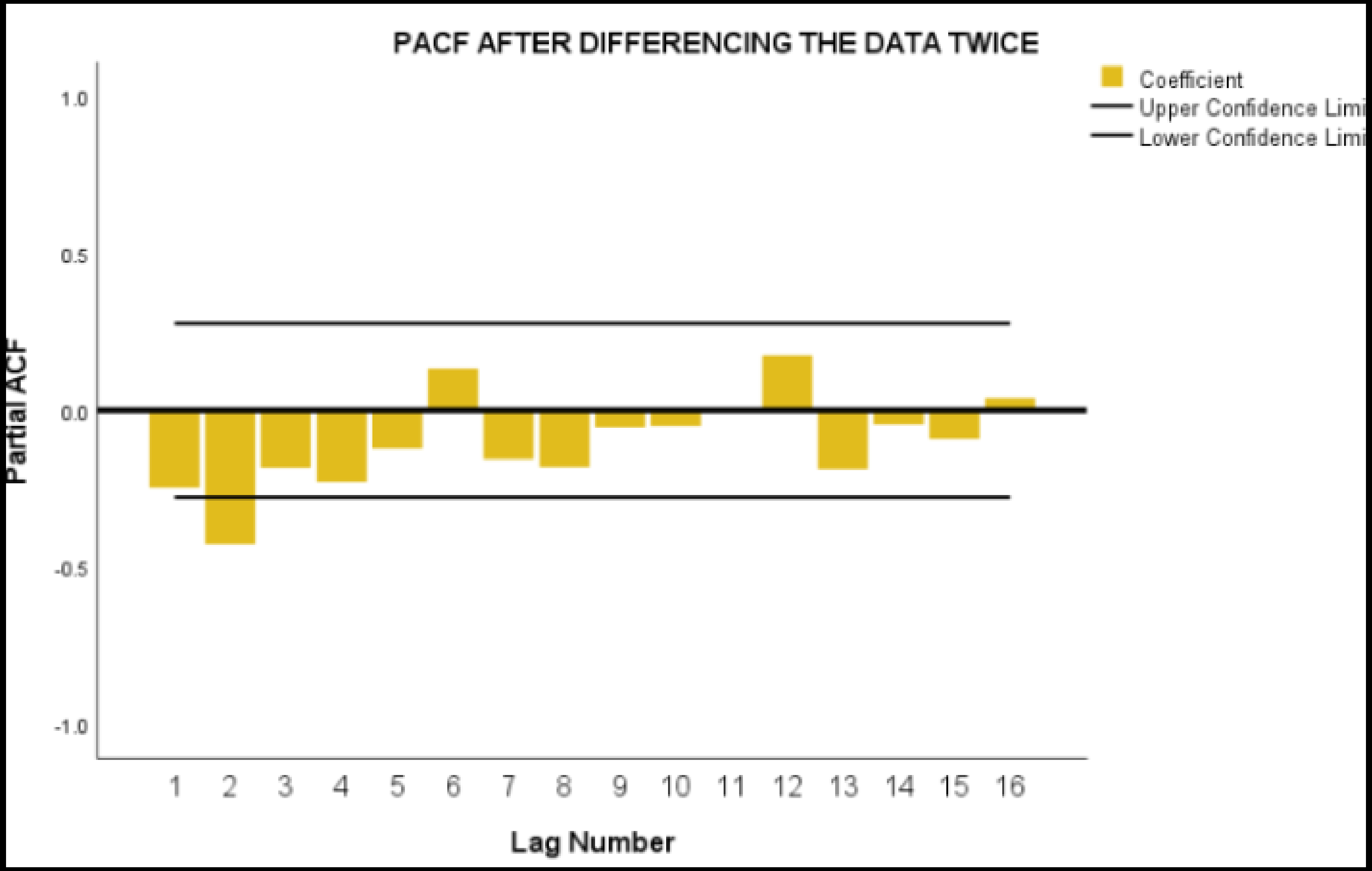
Thus, we conclude that the data is stationary.

Graph after Differencing the Data



Correlograms after Differencing the Data





From the above graphs, we can see that the graphs cut off after 1 lag

This shows that the value of p and q both will be 1

DETERMINING THE APPROPRIATE ARIMA MODEL

After differencing twice, our data became stationary. So, d = 2
To determine the appropriate ARIMA model, we'll observe the AIC values.

```
d=2
arima(ts_,order=c(2,2,0))  #955.35
arima(ts_,order=c(1,2,0))  #966.85
fitted <- arima(ts_,order=c(1,2,1))  #944.5 CHOSEN
arima(ts_,order=c(2,2,1))  #946.44
fits <- arima(ts_,order=c(1,2,2))  #946.47
arima(ts_,order=c(2,2,2))  #948.37
fitted <- arima(ts_,order=c(0,2,2))#945.2
arima(ts_,order=c(0,2,1))  #948.52
arima(ts_,order=c(0,2,0))  #993.85
```

Model	AIC VALUE
ARIMA(2,2,0)	955.35
ARIMA(1,2,0)	966.85
ARIMA(1,2,1)	**944.5**
ARIMA(2,2,1)	946.44
ARIMA(1,2,2)	946.47
ARIMA(2,2,2)	948.37
ARIMA(0,2,2)	945.2
ARIMA(0,2,1)	948.52
ARIMA(0,2,0)	993.85

Least value

Thus, the appropriate ARIMA model is ARIMA(1,2,1).

Predicted Average Price Of Gold (Rs. per 10gms.)

Case Summaries[a]

	Date. Format: "MMM YYYY"	Average Price of Gold (Rs. per 10gms.)	Predicted Average price of Gold (Rs.per 10gms.)
1	APR 2018	30985.50	.
2	MAY 2018	31116.14	.
3	JUN 2018	30707.14	31246.78
4	JUL 2018	30097.05	30475.41
5	AUG 2018	29644.29	29702.79
6	SEP 2018	30537.78	29292.40
7	OCT 2018	31489.29	30853.70
8	NOV 2018	31010.37	31915.34
9	DEC 2018	31250.50	30833.28
10	JAN 2019	32230.00	31366.18
11	FEB 2019	33216.50	32685.29
12	MAR 2019	32036.32	33722.29
13	APR 2019	31659.26	31658.95
14	MAY 2019	31737.55	31557.38
15	JUN 2019	33121.26	31804.69
16	JUL 2019	34522.26	33723.68
17	AUG 2019	37356.15	35180.38
18	SEP 2019	37926.74	38633.41
19	OCT 2019	38213.55	38388.52
20	NOV 2019	38122.20	38568.81
21	DEC 2019	38091.57	38325.46
22	JAN 2020	39983.83	38304.55
23	FEB 2020	41195.21	40932.54
24	MAR 2020	42285.33	41921.78
25	APR 2020	45504.44	42985.43
26	MAY 2020	46343.28	47035.37
27	JUN 2020	47314.73	47031.20
28	JUL 2020	49878.43	48058.62
29	AUG 2020	52917.19	51233.36
30	SEP 2020	50783.64	54493.04
31	OCT 2020	50697.85	50461.90
32	NOV 2020	50428.65	51086.10
33	DEC 2020	49443.86	50733.47
34	JAN 2021	49611.95	49464.20
35	FEB 2021	47107.10	50032.08
36	MAR 2021	44647.57	46526.93
37	APR 2021	46517.21	44030.80
38	MAY 2021	47860.25	47452.48
39	JUN 2021	47890.86	48625.62
40	JUL 2021	47763.52	48187.24
41	AUG 2021	47108.95	47995.14
42	SEP 2021	46584.14	47138.13
43	OCT 2021	47218.79	46644.65
44	NOV 2021	48192.85	47690.34
45	DEC 2021	47889.83	48792.23
46	JAN 2022	47960.25	48031.13
47	FEB 2022	49254.00	48228.11
48	MAR 2022	51750.29	49963.27
49	APR 2022	52022.53	52909.24
50	MAY 2022	50879.43	52400.54
51	JUN 2022	50804.23	50741.57
52	JUL 2022	50784.00	51033.89
53	AUG 2022	51639.00	51027.72
54	SEP 2022	50550.00	52194.45
Total N	54	54	52

Predicting the Parameters of ARIMA Model

ARIMA Model Parameters

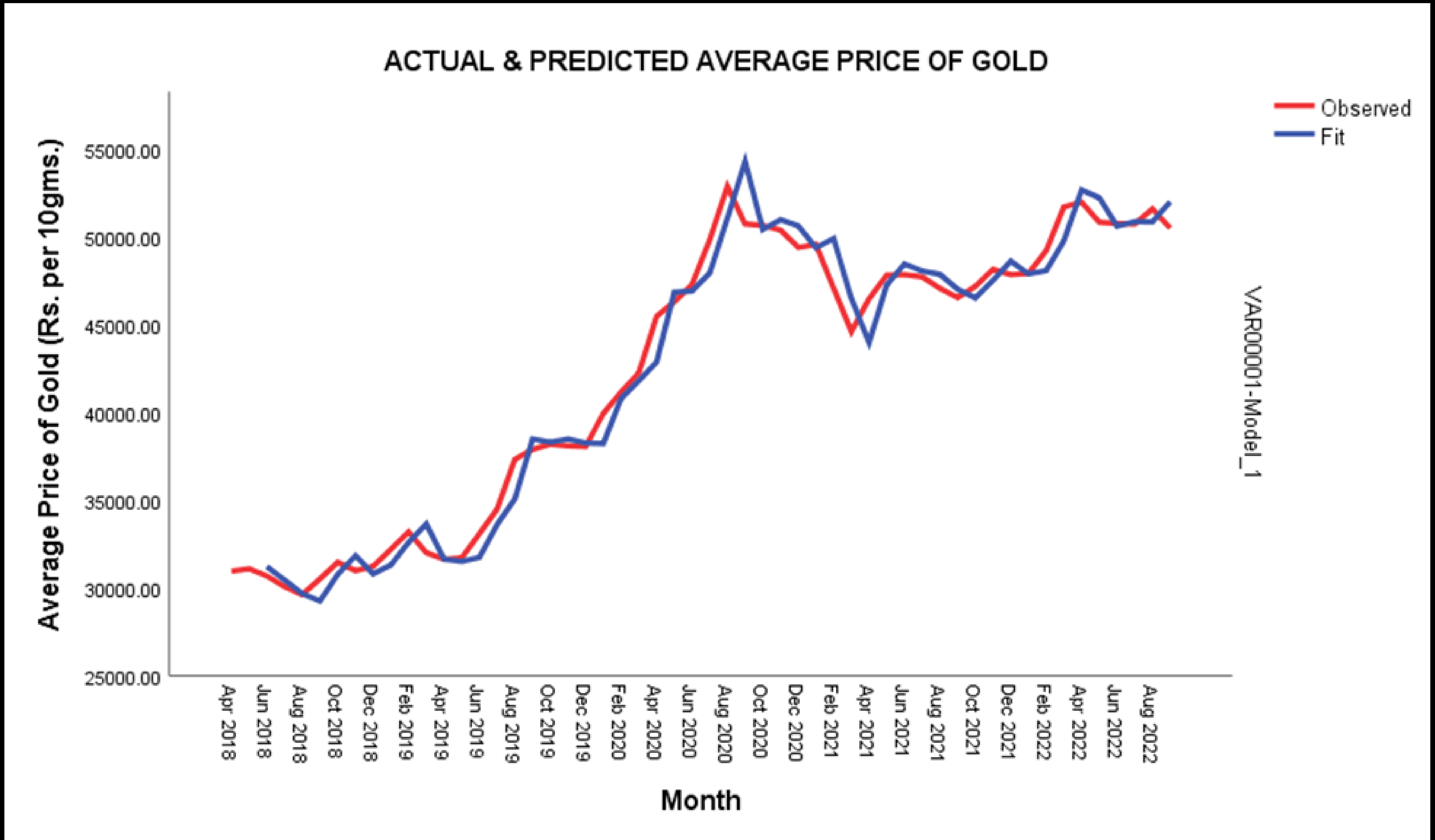
					Estimate	SE	t	Sig.
VAR00001-Model_1	VAR00001	No Transformation	Constant		-7.511	16.228	-.463	.646
			AR	Lag 1	.314	.160	1.962	.055
			Difference		2			
			MA	Lag 1	.999	6.782	.147	.883

The equation of ARIMA model is:

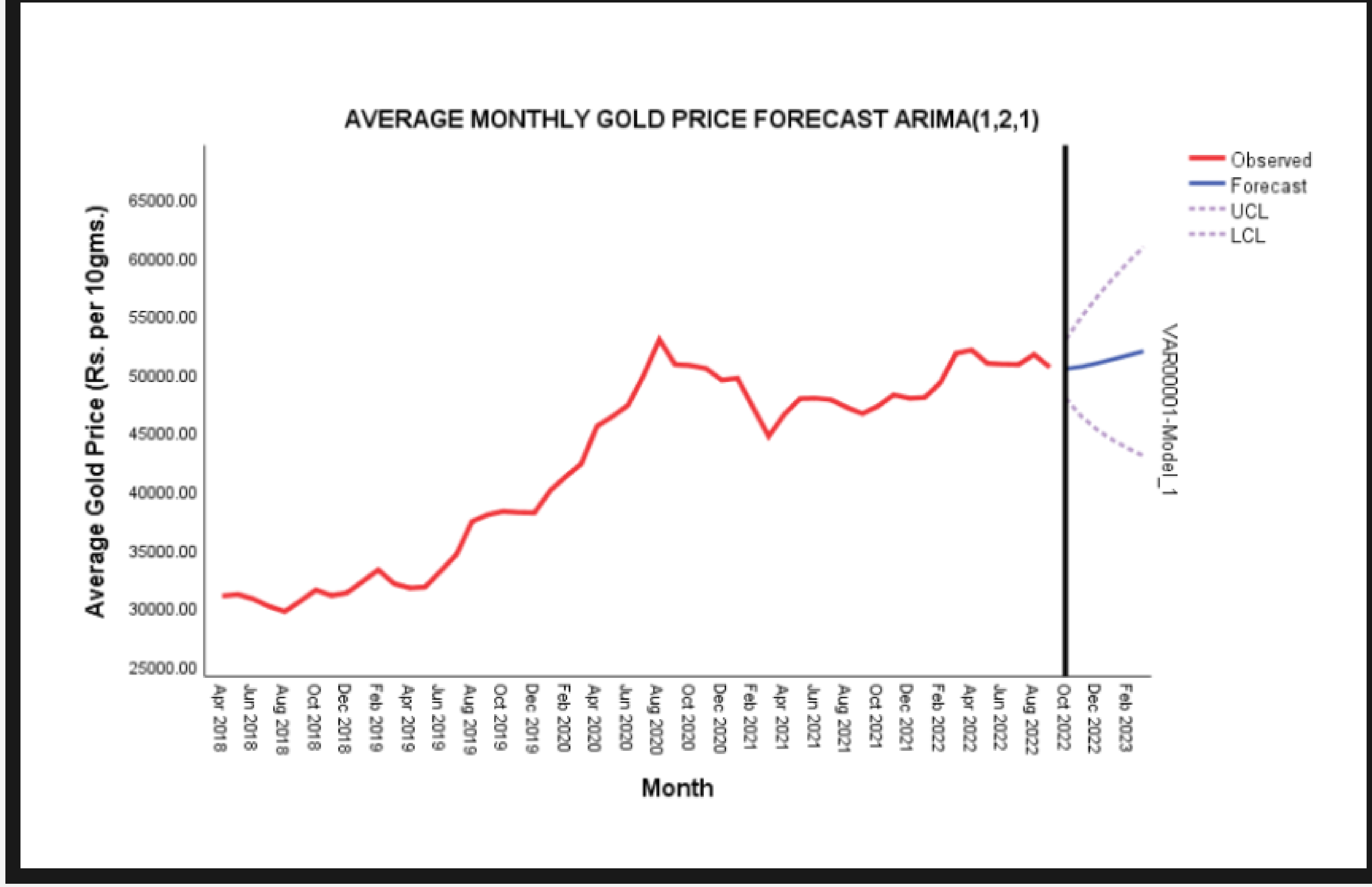
$$W_t = -7.511 + 0.314W_{t-1} + a_t - 0.9998a_{t-1}$$

where $W_t = \Delta^2 y_t$

Graph representing Actual and Predicted Price of Gold



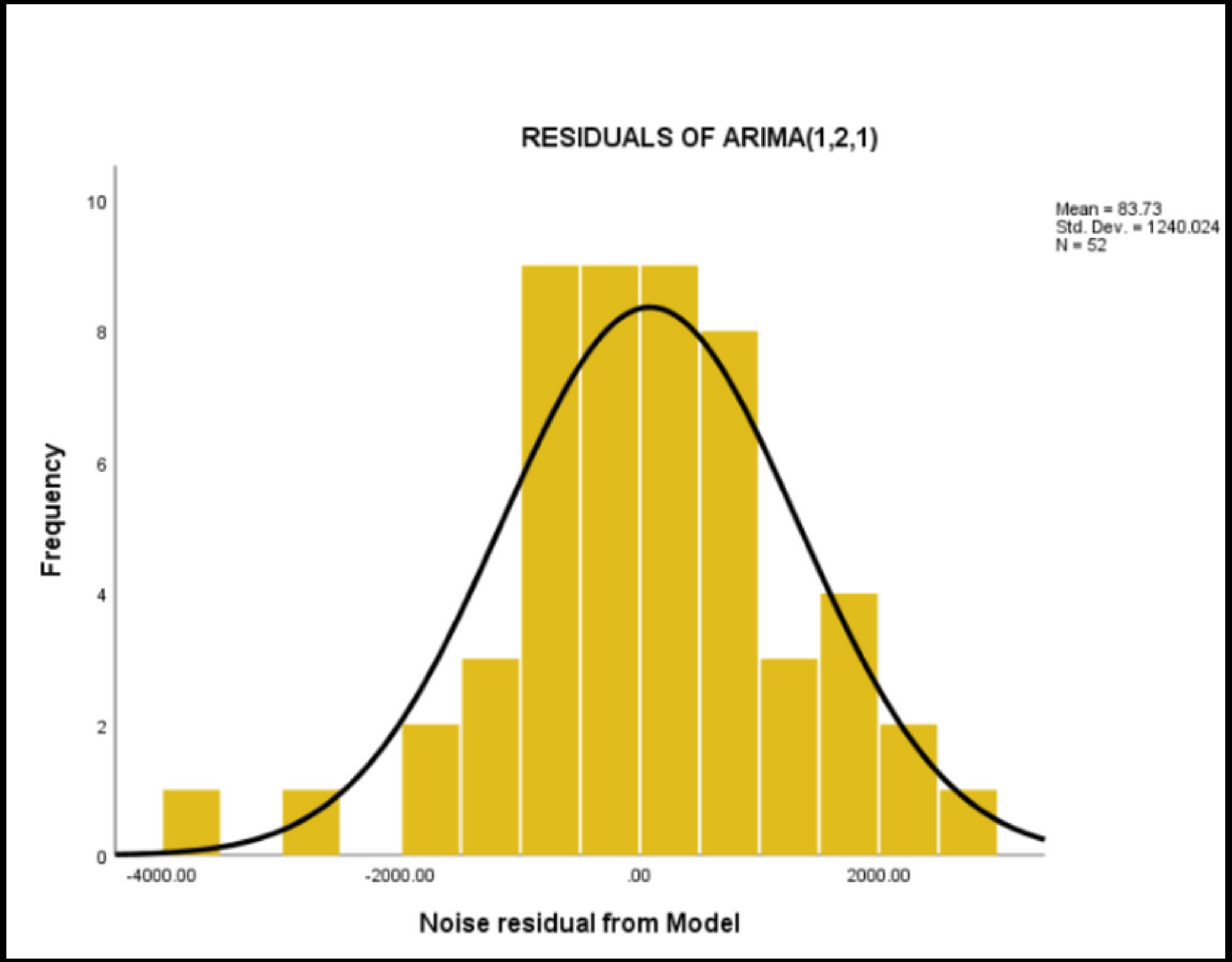
Forecasted Values

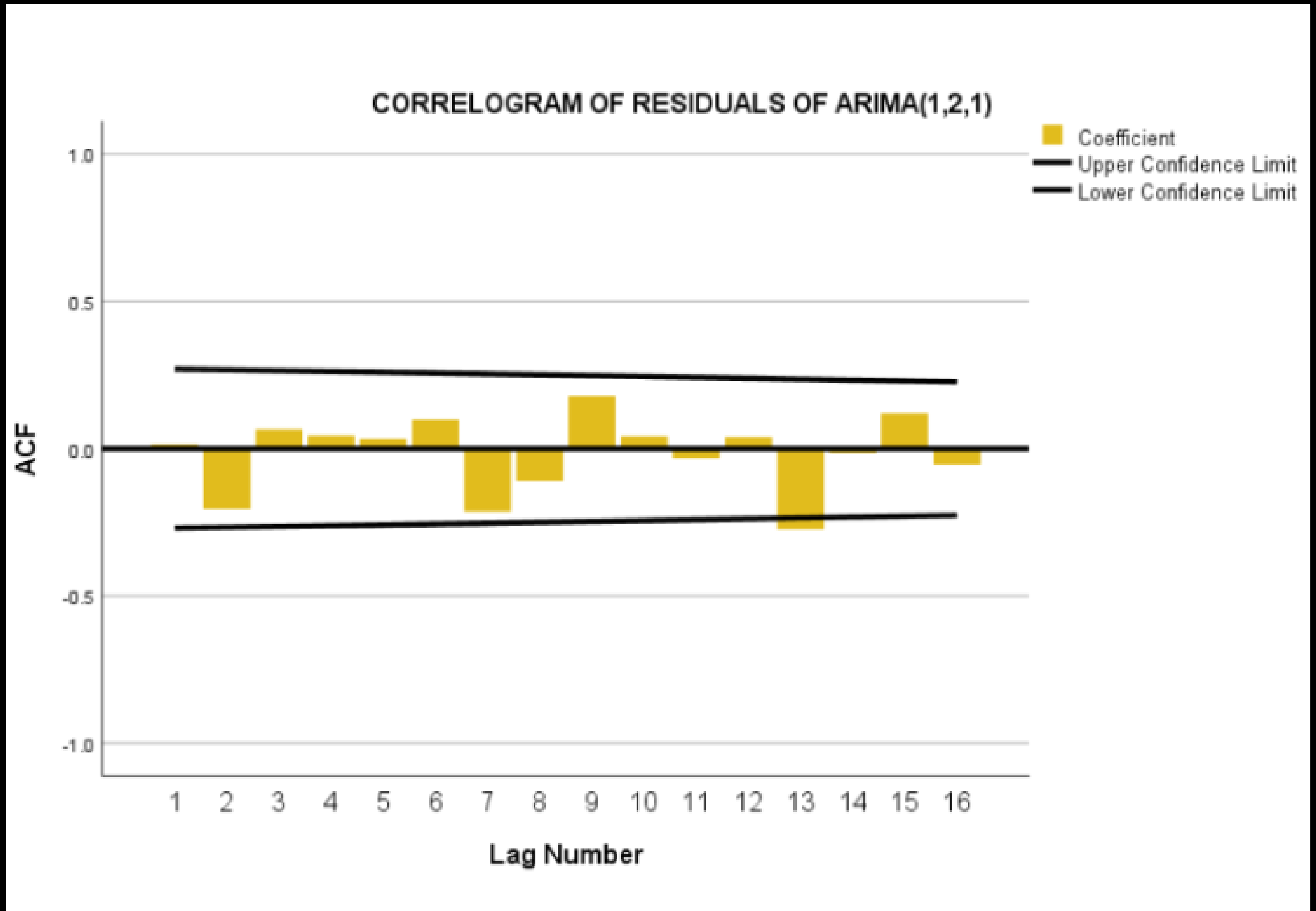


FORECASTED VALUES	
OCT 2022	50408.22
NOV 2022	50591.33
DEC 2022	50885.89
JAN 2023	51218.67
FEB 2023	51564.56
MAR 2023	51914.95

Residual Analysis

In Box Jenkins methodology, residuals of best fitted model must be independently identically normally distributed (IID). For justifying this assumption, Histogram and Correlogram have been utilized.





This graph shows that the residuals of ARIMA (1,2,1) are normally distributed.

Results

The forecasted values of the Gold prices obtained are :

- October 2022 - 50408.22
- November 2022 - 50591.33
- December 2022 - 50885.89
- January 2023- 51218.67
- February 2023 - 51564.67
- March 2023 - 51914.95

The fitted ARIMA (1,2,1) is a good fit since the distribution of residuals of the ARIMA model seems to follow a normal distribution.

Summary & Conclusions

- In this study, a univariate time series model was selected using the data of monthly Gold prices in India per 10 grams. We applied Box Jenkins methodology for forecasting gold prices for the upcoming months.
- Using the correlogram of ACF and PACF and the Dickey-Fuller test, we found that our data is stationary at the 2nd difference.
- By comparing AIC values, we determined the appropriate ARIMA model which came out to be ARIMA(1,2,1).
- We estimated the parameters of ARIMA(1,2,1) and then forecasted values for upcoming months. Looking at the forcasted valued we can see that the values are going in a upward trend.

Remarks

- When the expected or actual returns on bonds, equities, and real estate fall, the interest in gold investing can increase, driving up its price
- Gold is used in India as a form of tackling inflation, and holding an item with an intrinsic value because of its rarity is a good way to counter the fluctuations in fiat currency.

References

1. https://towardsdatascience.com/
2. NCSS Statistical Software
3. Data taken from the website of RBI: https://m.rbi.org.in//Scripts/AnnualPublications.aspx?head=Handbook%20of%20Statistics%20on%20Indian%20Economy
4. Link to the R codes: https://docs.google.com/document/d/1WjUXvNtIcTTK5tPGjLoSTS8kP_MlBZOvGBE_W4qzyrA/edit

Thank you