PROJECT REPORT FOR THE AWARD OF BACHELOR OF SCIENCE IN STATISTICS

MIDNAPORE COLLEGE

(AUTONOMOUS)

SESSION: 2021-2022

SUBMITTED BY: GUNJAN CHAKRABORTY

SUBJECT: STATISTICS(HONS.) (UG)

PAPER: STSC-DSE-IV

ROLL NO.- 1612

REGISTRATION NO: 2019-248

TOPIC NAME

TIME SERIES ANALYSIS OF GOLD PRICE

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project.All that I have done is only due to such supervision and assistance and I would not regret to thank them.

I respect and thank Prof. Moumita Roy for providing me an opportunity to do the project time series analysis of gold price and giving me all support and guidance, which helped me to complete the project work.

I am extremely thankful to Mrs. Piyali Kundu for providing such a nice support and guidance as my project guide. She guides me all along till the completion of my project work by providing all the necessary information for developing a good system.

I am fortunate enough to get constant encouragement, support and guidance from all teaching staffs of our department and I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

INDEX

➢ SUMMERY

Time Series in R is used to see how an object behaves over a period of time. In R, it can be easily done by ***ts()*** function with some parameters. Time series takes the data vector and each data is connected with timestamp value as given by the user. This function is mostly used to learn and forecast the behaviour of an asset in business for a period of time. For example, sales analysis of a company, inventory analysis, price analysis of a particular stock or market, population analysis, etc.

In this project I have taken gold prices from the year 2000 to year 2021 and using R programming I have forecasted an estimation of gold price of next 5 years for each month (i.e. next 60 months).

➢ MODEL MOTIVATION

The gold rates are always fluctuating and is difficult to predict. It can easily be shown by time series analysis. Before we begin the analysis, I will do the steps that we have to do. The steps are like this,

- First, We have to gather and pre-process the data, and also, we should know the domain knowledge of the data that we use,
- Then, We analyse the time series, visually and statistically,
- Then, We identify the perfect model based on its autocorrelation,
- Then, We diagnose the model whether it meets the independence assumption and finally,
- We can use the model for doing forecasting

➢ DATA DESCRIPTION

The data is consisted of two columns which are the date(dd-mm-yyyy) and the monthly average value of gold per gram in Indian rupees. It contains monthly average values of gold from year 2000 to year 2021.

➢ INTRODUCTION

❖ THE GOLD

Gold is a chemical element. Its symbol is *Au* and atomic number 79. It is a bright, slightly reddish yellow, dense, soft, malleable, and ductile metal. According to the World Gold Council, India is one of the largest markets for Gold, and growing affluence is driving growth in demand. Gold has a central role in the country's culture, considered a store of value, a symbol of wealth and status. Roughly 50% of new gold produced goes into jewellery, 40% is bought for investment purposes and 10% goes into practical uses. The price of gold is primarily determined by a combination of factors like supply, demand, and investor behaviour. However, current gold prices not only factor immediate supply and demand but also expectations of future supply and demand.

❖ PROJECT PROBLEM

The aim of the project is to see whether the price of gold will increase or decrease in future.

❖ WORK DONE IN DIFFERENT SECTION

Firstly I collected data from various websites and choose the suitable data for the project. Then I downloaded it in *.CSV* format and read the data in R software. Then I converted the data into a timeseries data and plotted it. I imported two library functions ,i.e. *tseries* and *forecast* to help the analysis of the data. Then I checked the data for stationarity by acf ,pacf and adf-test. Then I saw that the data was not stationary so I created a new model and applied auto,arima function, which automatically gives the best values of p, d, q. now I checked for stationarity again and it was stationary. Then I created a new model with the forecasted values and plotted it in a graph. Lastly I checked the data again with "Ljung-Box" test for any autocorrelation issue after forecasting and if the test was valid.

➢ METHODOLOGY

❖ TIME SERIES

Time series data is quantitative data in a series of particular periods or intervals. Time series data occurred when monitoring industrial processes or tracking corporate business metrics. Time series data can be defined as an ordered sequence of values of a quantitative variable at equally spaced time intervals. A time series has some quantitative value that is measured sequentially in time over some interval. Time series analysis has been exploited for many applications, e.g., sales forecasting, economic forecasting, budgetary analysis, yield projections, stock market analysis, census analysis, process, and quality control, inventory studies, workload projections, utility studies.

Time series analysis attempts to understand the chronicle data and predict future data. The time-series has four necessary components trend, seasonal, cyclical, and irregular. A trend is an overall directional movement of a time series. The trend will either be deterministic or stochastic. From a modelling viewpoint, the trend is the essential component of a time series. Many time series contain seasonal variation, and the seasonal component is measured on a monthly or quarterly basis. The cyclical component is considered overa long-time prospect, typically one year or longer. That occurs when time-series observations that are close together in time tend to be correlated. Irregular effects are the impact of any random events such as earthquakes, strikes, and disasters. The nature of Irregular effects is entirely unpredictable.

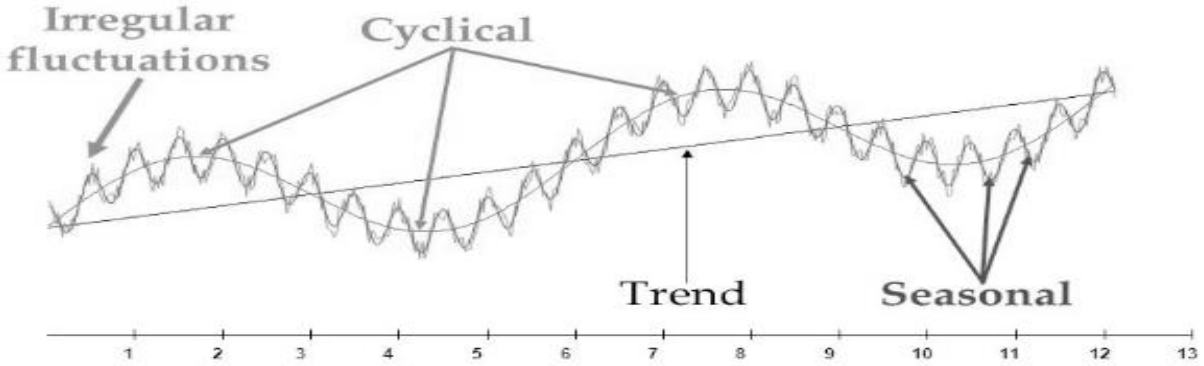


Figure: Time Series components

Time series data consists of four components. These are a trend (T), seasonality (S), cyclical (C), and Irregularity (noise)

Each data point (*Yt*) at time *t* in a Time Series can be expressed as either a sum or a product of 3 components, namely, *Seasonality (St)*, *Trend (Tt)* and *Error (et)* (a.k.a White Noise).

For Additive Time Series,

$$Y_t = S_t + T_t + \epsilon_t$$

For Multiplicative Time Series,

$$Y_t = S_t \times T_t \times \epsilon_t$$

❖ ARIMA MODEL IN R

ARIMA model of time series introduced by Box and Jenkins introduced in 1970. This type of model includes three models of time series, the ***Autoregressive*** (AR), ***Moving Average*** (MA) and ***Autoregressive Moving Average*** (ARMA). The general notation for the ARIMA model is ARIMA (p, d, q). ARIMA models are classified by three factors:

p = Number of autoregressive terms (AR)

d = How many non-seasonal differences are needed to achieve stationarity (I)

q = Number of lagged forecast errors in the prediction equation (MA)

The idea for this analysis is to identify the time series components which are:

- Trend
- Seasonality
- Random behaviour of data

Then, we'll forecast the values based on historical data.

Assumptions of ARIMA model

- Data should be stationary – by stationary it means that the properties of the series don't depend on the time when it is captured. A white noise series and series with cyclic behaviour can also be considered as stationary series.
- Data should be univariate – ARIMA works on a single variable. Auto-regression is all about regression with the past values.

Steps to be followed for ARIMA modelling

- Exploratory analysis
- Fit the model
- Diagnostic measures

Trend, seasonal, ACF, PACF are some useful plots for analysis and visualization purposes of time series. ***The autocorrelation function*** **(ACF)** gives values of auto-correlation of any series with its lagged values. An ACF plot describes the relation of the present value of the series with past values. ***The partial autocorrelation function*** **(PACF)** is a partial autocorrelation function and finds a correlation of the residuals with the next lag value.

R is a software environment with a programming language most commonly used for statistical computing and machine learning.It is maintaining by the R Foundation. In 1993, R was developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. Now R is available as open-source software. R possesses an extensive catalogue of statistical and graphical methods. R is not only entrusted by academics, even some reputed business organizations, including Uber, Google, Airbnb, Facebook, also use. The scientific community has also recognized computation using R. The enormous help and reference resources of R available on the internet.

In R, a time-series object similar to a data frame is used to store data of time series An R function **ts()** is used to create a time- series object.

The basic syntax for ts() function in time series analysis is:

time-series.object.name <- ts(data, start, end, frequency)

where,

- **data** represents the data vector
- **start** represents the first observation in time series
- **end** represents the last observation in time series
- **frequency** represents number of observations per unit time. For example, frequency=12 for monthly data, 4 for quarterly data and 1 for yearly data.

❖ STATIONARY TIME SERIES

A time series is said to be stationary if it holds the following conditions true.

1. The mean value of time-series is constant over time, which implies, the trend component is nullified.
2. The variance does not increase over time.
3. Seasonality effect is minimal.

This means it is devoid of trend or seasonal patterns, which makes it looks like a random white noise irrespective of the observed time interval.

Use ***Augmented Dickey-Fuller Test*** **(adf test)**. A p-Value of less than 0.05 in adf.test() indicates that it is stationary.

➢ DATA ANALYSIS

First we have to take/create a dataset (.CSV) which consist of time and the corresponding values of gold (per gram) at that time. Then we read the dataset in the R programming.

COMMANDS IN R:

```
> gold_data<-read.csv(file="C:/Users/Gunjan/Desktop/
project/monthly.csv")
> gold_data
Date    Gram
1   31-01-2000  398.13
2   29-02-2000  420.41
3   31-03-2000  401.30
4   28-04-2000  392.42
5   31-05-2000  389.03
6   30-06-2000  410.45
7   31-07-2000  405.37
8   31-08-2000  403.08
9   29-09-2000  403.75
10  31-10-2000  402.24
11  30-11-2000  400.02
12  29-12-2000  408.01
13  31-01-2001  397.14
14  28-02-2001  391.64
15  30-03-2001  394.18
16  30-04-2001  391.45
17  31-05-2001  410.81
18  29-06-2001  408.28
19  31-07-2001  405.39
20  31-08-2001  412.66
21  28-09-2001  434.09
22  31-10-2001  436.92
23  30-11-2001  426.13
```

24	31-12-2001	424.56
25	31-01-2002	437.14
26	28-02-2002	462.28
27	29-03-2002	460.57
28	30-04-2002	475.81
29	31-05-2002	495.13
30	28-06-2002	505.23
31	31-07-2002	490.95
32	30-08-2002	484.41
33	30-09-2002	496.82
34	31-10-2002	492.03
35	29-11-2002	494.83
36	31-12-2002	513.55
37	31-01-2003	549.60
38	28-02-2003	550.96
39	31-03-2003	521.52
40	30-04-2003	499.86
41	30-05-2003	538.34
42	30-06-2003	535.01
43	31-07-2003	521.53
44	29-08-2003	531.13
45	30-09-2003	558.32
46	31-10-2003	552.85
47	28-11-2003	570.89
48	31-12-2003	596.02
49	30-01-2004	604.27
50	27-02-2004	588.77
51	31-03-2004	587.40
52	30-04-2004	569.00
53	31-05-2004	557.21
54	30-06-2004	573.65
55	30-07-2004	588.78
56	31-08-2004	596.19
57	30-09-2004	599.93

58	29-10-2004	618.24
59	30-11-2004	636.46
60	31-12-2004	624.28
61	31-01-2005	595.53
62	28-02-2005	594.19
63	31-03-2005	609.50
64	29-04-2005	603.17
65	31-05-2005	589.68
66	30-06-2005	603.08
67	29-07-2005	593.52
68	31-08-2005	613.61
69	30-09-2005	643.37
70	31-10-2005	676.76
71	30-11-2005	699.90
72	30-12-2005	749.05
73	31-01-2006	782.04
74	28-02-2006	789.80
75	31-03-2006	795.74
76	28-04-2006	881.35
77	31-05-2006	984.41
78	30-06-2006	882.08
79	31-07-2006	946.75
80	31-08-2006	946.41
81	29-09-2006	886.82
82	31-10-2006	856.11
83	30-11-2006	904.95
84	29-12-2006	903.97
85	31-01-2007	898.97
86	28-02-2007	943.57
87	30-03-2007	925.46
88	30-04-2007	919.88
89	31-05-2007	875.73
90	29-06-2007	859.57
91	31-07-2007	864.53

92 31-08-2007 873.27
93 28-09-2007 923.87
94 31-10-2007 958.88
95 30-11-2007 1022.79
96 31-12-2007 1018.98
97 31-01-2008 1125.94
98 29-02-2008 1179.00
99 31-03-2008 1255.99
100 30-04-2008 1171.32
101 30-05-2008 1205.05
102 30-06-2008 1225.29
103 31-07-2008 1294.49
104 29-08-2008 1159.04
105 30-09-2008 1217.44
106 31-10-2008 1259.77
107 28-11-2008 1200.05
108 31-12-2008 1278.12
109 30-01-2009 1349.80
110 27-02-2009 1496.07
111 31-03-2009 1523.57
112 30-04-2009 1434.76
113 29-05-2009 1450.18
114 30-06-2009 1452.49
115 31-07-2009 1455.22
116 31-08-2009 1475.19
117 30-09-2009 1550.45
118 30-10-2009 1567.18
119 30-11-2009 1687.60
120 31-12-2009 1700.24
121 29-01-2010 1650.70
122 26-02-2010 1632.03
123 31-03-2010 1629.16
124 30-04-2010 1643.26
125 31-05-2010 1780.26
126 30-06-2010 1846.05

127 30-07-2010 1796.18
128 31-08-2010 1821.01
129 30-09-2010 1879.62
130 29-10-2010 1916.52
131 30-11-2010 1981.71
132 31-12-2010 2018.48
133 31-01-2011 1982.44
134 28-02-2011 2005.29
135 31-03-2011 2058.95
136 29-04-2011 2103.84
137 31-05-2011 2183.20
138 30-06-2011 2203.93
139 29-07-2011 2245.77
140 31-08-2011 2560.36
141 30-09-2011 2711.90
142 31-10-2011 2638.09
143 30-11-2011 2837.89
144 30-12-2011 2784.76
145 31-01-2012 2719.54
146 29-02-2012 2756.22
147 30-03-2012 2711.41
148 30-04-2012 2748.13
149 31-05-2012 2776.65
150 29-06-2012 2880.41
151 31-07-2012 2841.31
152 31-08-2012 2904.18
153 28-09-2012 3056.64
154 31-10-2012 2982.23
155 30-11-2012 3035.66
156 31-12-2012 2965.01
157 31-01-2013 2914.74
158 28-02-2013 2816.99
159 29-03-2013 2787.25
160 30-04-2013 2597.00

161 31-05-2013 2498.82
162 28-06-2013 2517.49
163 31-07-2013 2473.96
164 30-08-2013 2724.58
165 30-09-2013 2769.30
166 31-10-2013 2608.48
167 29-11-2013 2568.56
168 31-12-2013 2437.21
169 31-01-2014 2485.76
170 28-02-2014 2602.87
171 31-03-2014 2618.42
172 30-04-2014 2519.87
173 30-05-2014 2457.13
174 30-06-2014 2457.39
175 31-07-2014 2531.95
176 29-08-2014 2537.01
177 30-09-2014 2424.32
178 31-10-2014 2412.97
179 28-11-2014 2333.73
180 31-12-2014 2421.10
181 30-01-2015 2501.37
182 27-02-2015 2448.36
183 31-03-2015 2367.50
184 30-04-2015 2417.09
185 29-05-2015 2459.34
186 30-06-2015 2425.50
187 31-07-2015 2312.40
188 31-08-2015 2336.41
189 30-09-2015 2394.60
190 30-10-2015 2425.64
191 30-11-2015 2310.30
192 31-12-2015 2288.68
193 29-01-2016 2377.31
194 29-02-2016 2632.69

195 31-03-2016 2684.04
196 29-04-2016 2655.29
197 31-05-2016 2710.27
198 30-06-2016 2761.81
199 29-07-2016 2889.62
200 31-08-2016 2884.89
201 30-09-2016 2846.41
202 31-10-2016 2717.06
203 30-11-2016 2685.04
204 30-12-2016 2512.02
205 31-01-2017 2611.43
206 28-02-2017 2660.17
207 31-03-2017 2606.65
208 28-04-2017 2625.94
209 31-05-2017 2579.84
210 30-06-2017 2611.38
211 31-07-2017 2561.76
212 31-08-2017 2637.69
213 29-09-2017 2724.96
214 31-10-2017 2676.75
215 30-11-2017 2674.34
216 29-12-2017 2607.63
217 31-01-2018 2724.50
218 28-02-2018 2757.17
219 30-03-2018 2771.03
220 30-04-2018 2820.04
221 31-05-2018 2829.81
222 29-06-2018 2793.52
223 31-07-2018 2736.77
224 31-08-2018 2688.83
225 28-09-2018 2782.49
226 31-10-2018 2876.05
227 30-11-2018 2820.49
228 31-12-2018 2843.57
229 31-01-2019 2939.98

230 28-02-2019 3023.02
231 29-03-2019 2907.31
232 30-04-2019 2870.97
233 31-05-2019 2882.71
234 28-06-2019 3033.50
235 31-07-2019 3124.07
236 30-08-2019 3426.45
237 30-09-2019 3468.08
238 31-10-2019 3414.64
239 29-11-2019 3378.10
240 31-12-2019 3376.76
241 31-01-2020 3577.09
242 28-02-2020 3671.60
243 31-03-2020 3812.19
244 30-04-2020 4122.19
245 29-05-2020 4175.07
246 30-06-2020 4218.62
247 31-07-2020 4442.60
248 31-08-2020 4726.84
249 30-09-2020 4542.88
250 30-10-2020 4489.58
251 30-11-2020 4450.51
252 31-12-2020 4395.41
253 29-01-2021 4387.96
254 26-02-2021 4231.58
255 31-03-2021 4021.66
256 30-04-2021 4223.46
257 31-05-2021 4361.07
258 30-06-2021 4338.09
259 30-07-2021 4330.67
260 31-08-2021 4254.76
261 30-09-2021 4204.75
262 29-10-2021 4283.61
263 30-11-2021 4359.25
264 31-12-2021 4335.99

Now we have to convert this data to a time-series data.

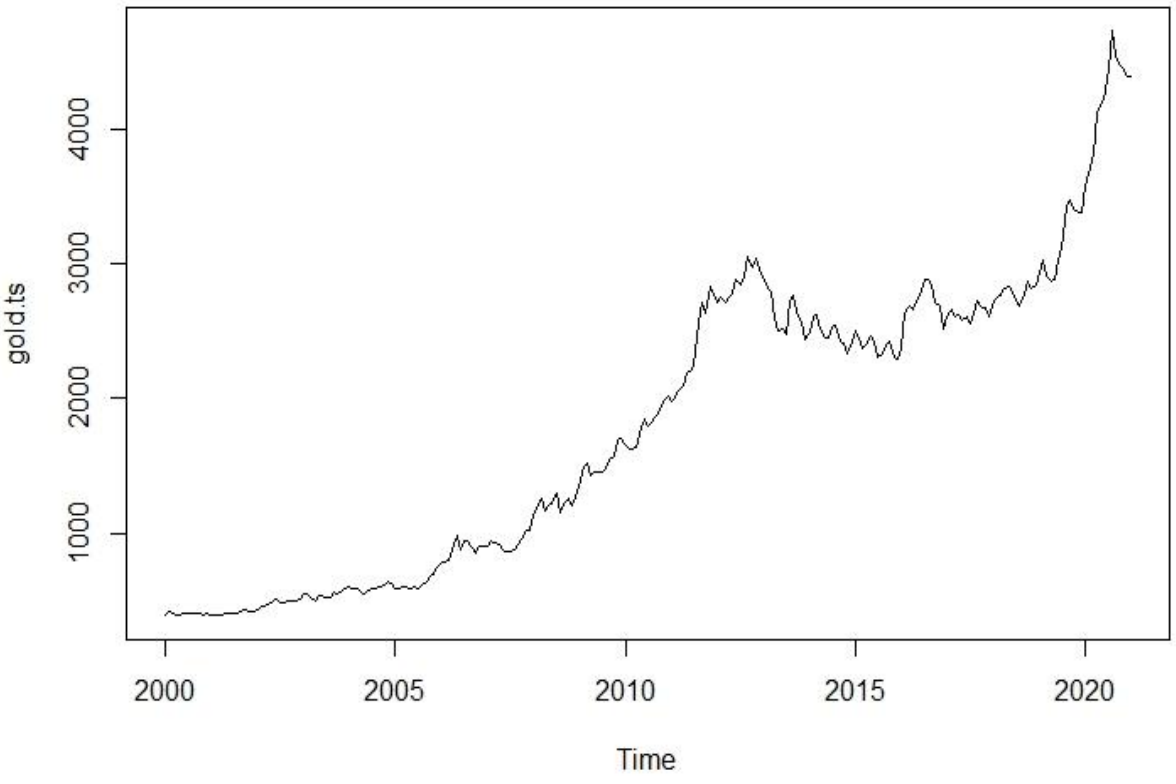
COMMANDS IN R:

```
> gold.ts=ts(gold_data$gram,start=c(2000,1,1),end=c(2021,1,1),
  freq=12)
```

After converting this data,we can plot it into a line diagram and see how the data looks. But before that we have to import two libraries, that are Forecast() and Tseries()

COMMANDS IN R:

```
> library(forecast)
> library(tseries)
> plot(gold.ts)
```

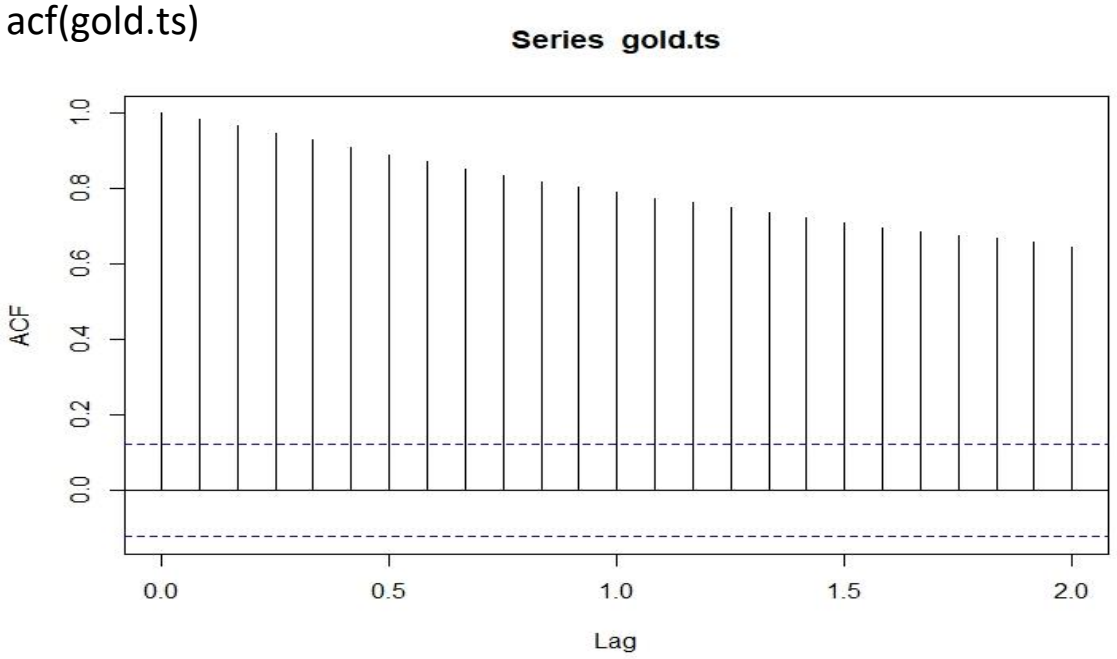


We can see that from the graph that there is an upward trend.

Now we have to check if the data is stationary or not, for that we will check the auto-correlation function.

COMMANDS IN R:

> acf(gold.ts)

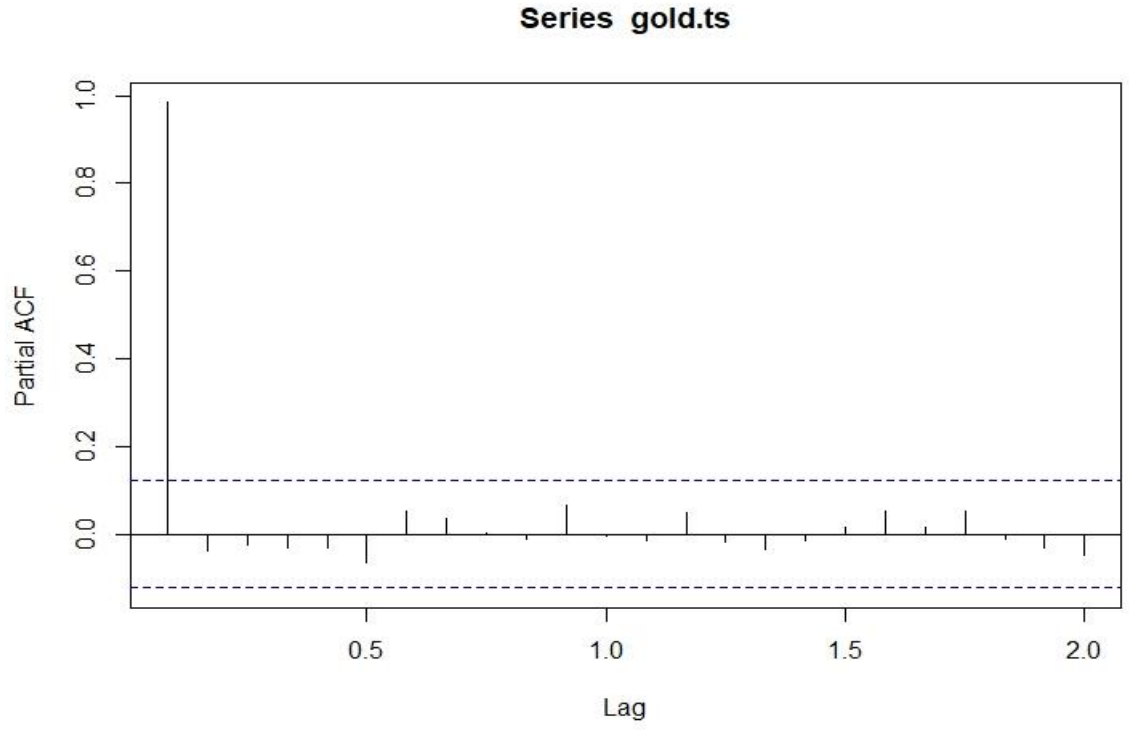


as we can see the data is not stationary and highly auto-correlated as the spikes are going above the blue dotted line .

now we check the partial auto-correlation function.

COMMANDS IN R:

> pacf(gold.ts)



Now the last test is the Augmented Dickey-Fuller Test (adf test)

COMMANDS IN R:

```
> adf.test(gold.ts)
```

Augmented Dickey-Fuller Test

data: gold.ts

Dickey-Fuller = -1.8783, Lag order = 6, p-value = 0.6271

alternative hypothesis: stationary

we can see that p-value is greater than 0.05, so the the data is not stationary.

So, now we have to convert the non stationary data to a stationary data.

There is a function called auto.arima function which automatically evaluates the data and gives us the values of p,d,q. It uses akaike information criterion to determine these valuese.

So for that, we need to create a new model say gold.model

COMMANDS IN R:

```
> gold.model=auto.arima(gold.ts,ic="aic",trace=TRUE)
```

Fitting models using approximations to speed things up...

ARIMA(2,1,2)(1,0,1)[12] with drift : 2898.018

ARIMA(0,1,0) with drift : 2890.833

ARIMA(1,1,0)(1,0,0)[12] with drift : 2894.412

ARIMA(0,1,1)(0,0,1)[12] with drift : 2882.061

ARIMA(0,1,0) : 2899.635

ARIMA(0,1,1) with drift : 2884.42

ARIMA(0,1,1)(1,0,1)[12] with drift : 2891.521

ARIMA(0,1,1)(0,0,2)[12] with drift : 2877.481

ARIMA(0,1,1)(1,0,2)[12] with drift : 2890.9

ARIMA(0,1,0)(0,0,2)[12] with drift : 2884.296

ARIMA(1,1,1)(0,0,2)[12] with drift : 2879.83

ARIMA(0,1,2)(0,0,2)[12] with drift : 2878.911

ARIMA(1,1,0)(0,0,2)[12] with drift : 2879.574

ARIMA(1,1,2)(0,0,2)[12] with drift : 2881.232

ARIMA(0,1,1)(0,0,2)[12] : 2880.282

Now re-fitting the best model(s) without approximations...

ARIMA(0,1,1)(0,0,2)[12] with drift : 2887.203

Best model: ARIMA(0,1,1)(0,0,2)[12] with drift

We can see the akaike information criterion value is lowest in the : ARIMA(0,1,1)(0,0,2)[12] with drift model, so we choose this model to work with.

Now before forecasting we check whether the data is stationary or not again.

COMMANDS IN R:

> gold.model

Series: gold.ts

ARIMA(0,1,1)(0,0,2)[12] with drift

Coefficients:

ma1	sma1	sma2	drift
0.1972	0.1580	0.1702	16.4759
s.e. 0.0650	0.0678	0.0712	7.1652

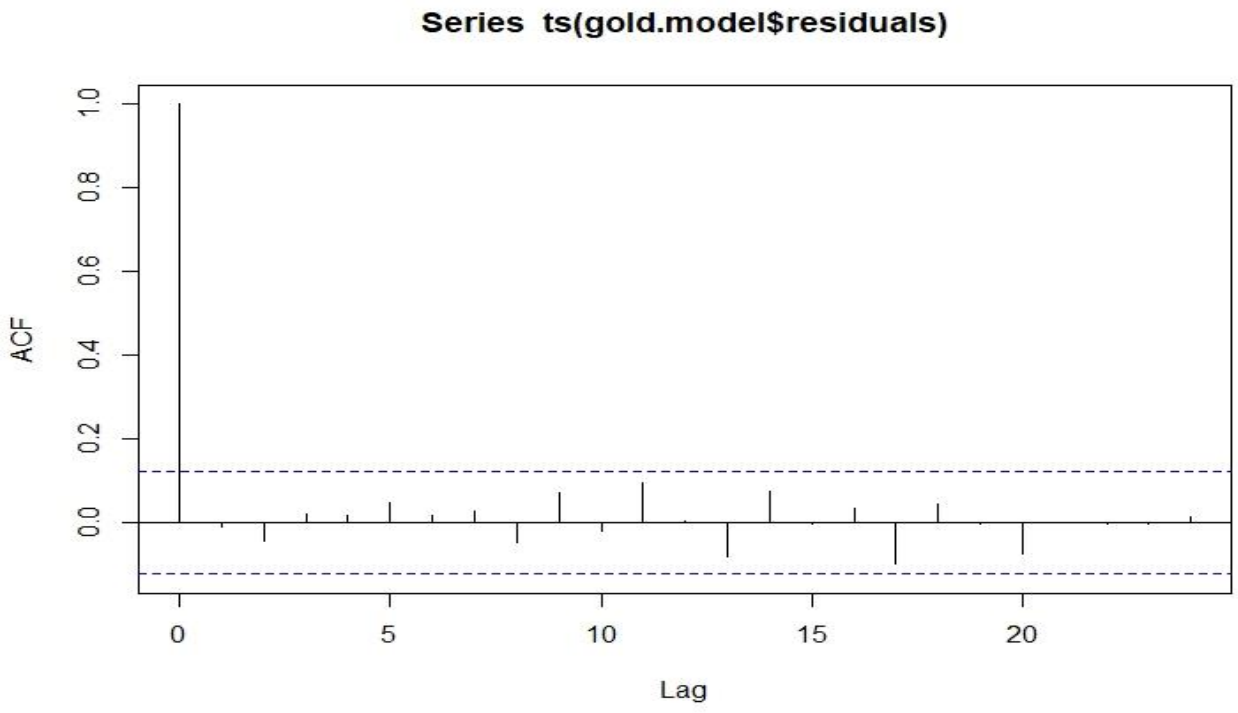
sigma^2 = 5387: log likelihood = -1438.6

AIC=2887.2 AICc=2887.45 BIC=2904.85

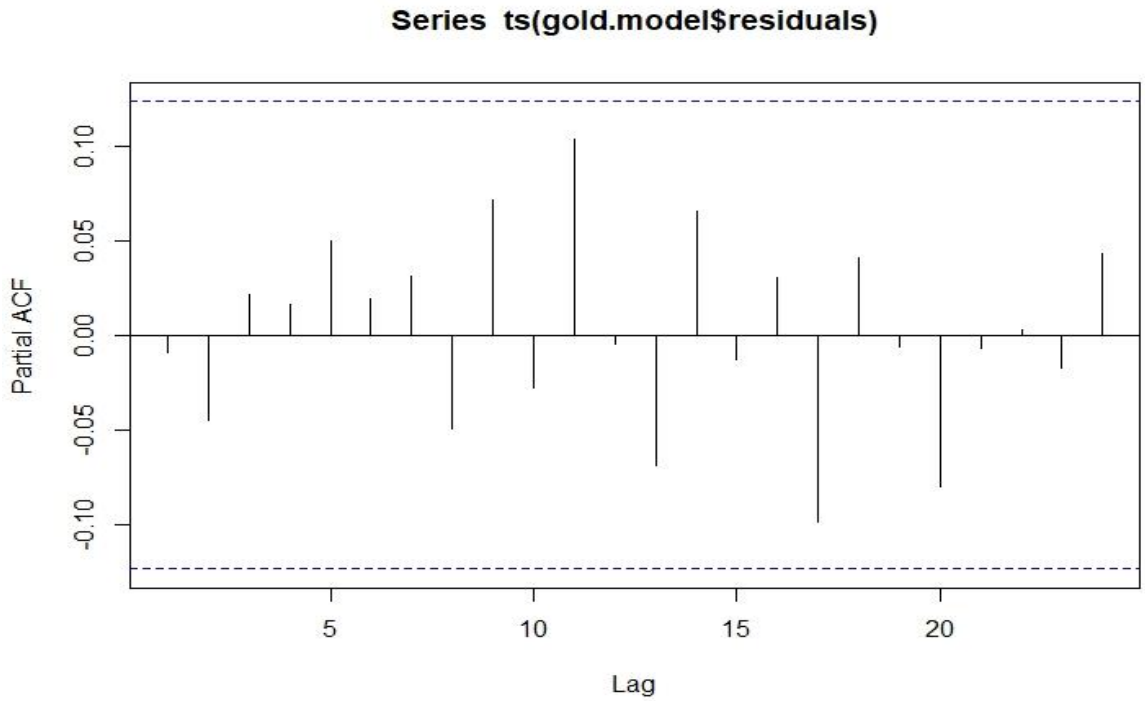
now we check the acf(autocorrelation function) and pacf(partial autocorrelation function) for the residuals.

COMMANDS IN R:

> acf(ts(gold.model$residuals))



> pacf(ts(gold.model$residuals))



We can see from these graphs that the data is stationary and the the values of p,d,q are 0,1,1 .

Now for the forecast we will use level of confidence =95 and 12 months each for next 5 years

COMMANDS IN R:

> my_gold_forecast=forecast(gold.model,level=c(95),h=5*12)

> my_gold_forecast

	Point Forecast	Lo 95	Hi 95
Feb 2021	4418.579	4274.723	4562.434
Mar 2021	4437.030	4212.631	4661.429
Apr 2021	4490.291	4207.420	4773.161
May 2021	4513.820	4182.646	4844.995
Jun 2021	4556.857	4183.578	4930.136
Jul 2021	4622.601	4211.508	5033.695
Aug 2021	4724.949	4279.238	5170.660

Sep 2021	4708.105	4230.277	5185.932
Oct 2021	4699.045	4191.128	5206.962
Nov 2021	4702.454	4166.132	5238.775
Dec 2021	4703.757	4140.462	5267.052
Jan 2022	4738.345	4149.310	5327.379
Feb 2022	4765.541	4145.083	5386.000
Mar 2022	4806.439	4154.715	5458.164
Apr 2022	4873.276	4191.719	5554.833
May 2022	4895.745	4185.608	5605.882
Jun 2022	4914.847	4177.236	5652.458
Jul 2022	4966.435	4202.338	5730.533
Aug 2022	5022.892	4233.196	5812.588
Sep 2022	5003.051	4188.560	5817.541
Oct 2022	5006.578	4168.026	5845.131
Nov 2022	5016.745	4154.802	5878.688
Dec 2022	5020.116	4135.401	5904.831
Jan 2023	5026.554	4119.638	5933.470
Feb 2023	5041.518	4107.367	5975.670
Mar 2023	5057.994	4096.242	6019.747
Apr 2023	5074.470	4085.887	6063.053
May 2023	5090.946	4076.242	6105.650
Jun 2023	5107.422	4067.252	6147.592
Jul 2023	5123.898	4058.871	6188.925
Aug 2023	5140.374	4051.057	6229.691
Sep 2023	5156.850	4043.773	6269.926
Oct 2023	5173.326	4036.986	6309.666

Nov 2023 5189.802 4030.665 6348.938

Dec 2023 5206.278 4024.785 6387.771

Jan 2024 5222.754 4019.319 6426.188

Feb 2024 5239.230 4014.247 6464.212

Mar 2024 5255.705 4009.547 6501.864

Apr 2024 5272.181 4005.201 6539.162

May 2024 5288.657 4001.191 6576.123

Jun 2024 5305.133 3997.503 6612.764

Jul 2024 5321.609 3994.120 6649.098

Aug 2024 5338.085 3991.031 6685.139

Sep 2024 5354.561 3988.221 6720.901

Oct 2024 5371.037 3985.680 6756.393

Nov 2024 5387.513 3983.397 6791.629

Dec 2024 5403.989 3981.361 6826.617

Jan 2025 5420.465 3979.563 6861.367

Feb 2025 5436.941 3977.993 6895.888

Mar 2025 5453.417 3976.644 6930.189

Apr 2025 5469.892 3975.508 6964.277

May 2025 5486.368 3974.577 6998.160

Jun 2025 5502.844 3973.844 7031.845

Jul 2025 5519.320 3973.303 7065.338

Aug 2025 5535.796 3972.946 7098.646

Sep 2025 5552.272 3972.770 7131.775

Oct 2025 5568.748 3972.767 7164.729

Nov 2025 5585.224 3972.932 7197.516

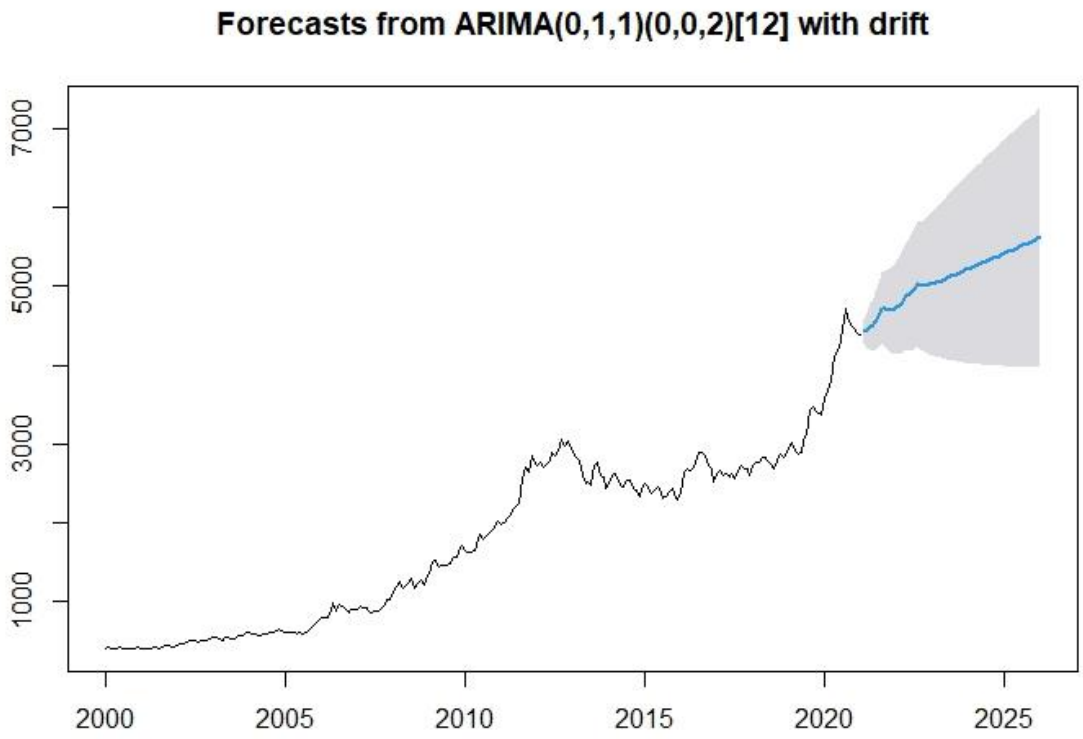
Dec 2025 5601.700 3973.261 7230.139

Jan 2026 5618.176 3973.748 7262.604

Now we plot this data into the graph

COMMANDS IN R:

> plot(my_gold_forecast)



Now for the last step we have to validate the data. So we use Box-test.

COMMANDS IN R:

> Box.test(my_gold_forecast$resid, lag=5, type= "Ljung-Box")

Box-Ljung test

data: my_gold_forecast$resid

X-squared = 1.3175, df = 5, p-value = 0.9331

```
> Box.test(my_gold_forecast$resid, lag=15, type= "Ljung-Box")

        Box-Ljung test

data:  my_gold_forecast$resid
X-squared = 9.3681, df = 15, p-value = 0.8575

> Box.test(my_gold_forecast$resid, lag=25, type= "Ljung-Box")

        Box-Ljung test

data:  my_gold_forecast$resid
X-squared = 14.527, df = 25, p-value = 0.9517
```

As we can see from the Box-test the p-value is always greater than level of significance (p-value>0.05), we conclude the data do not face any autocorrelation issue after forecasting and the test is valid.

❖ CONCLUSION

At first the data was not stationary. After making the data stationary we plotted it as a time series data and saw that there was a upward trend. Now, after the ARIMA process we got the predicted values of next 5 years. When we plotted the future prediction values in the graph, it was also giving a upward line. So we conclude that in the future the gold price will increase as it shows in the graph.

❖ REFERENCES

- Website

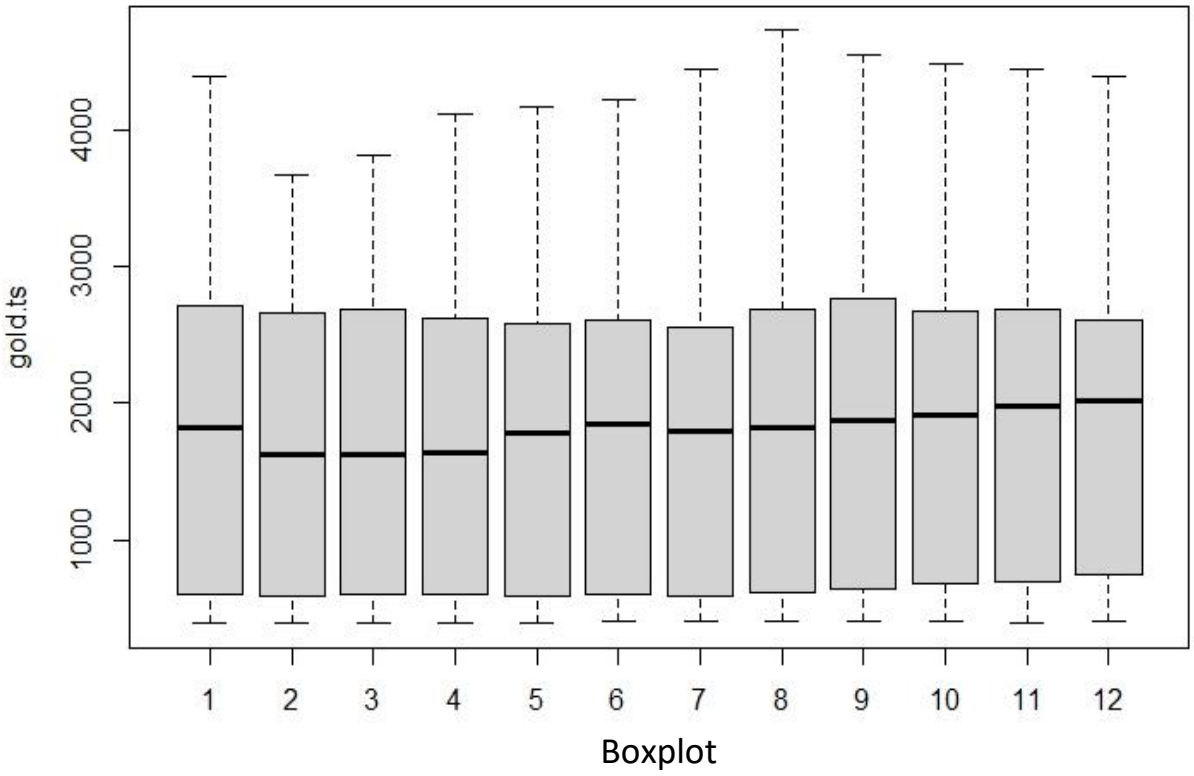
https://www.indexmundi.com/commodities/?commodity=gold
https://www.gold.org/goldhub/data/gold-prices
https://www.tutorialspoint.com/r/index.htm
https://en.wikipedia.org/wiki/Gold

- Books

Fundamentals of applied statistics –S. C. Gupta , V. K. Kapoor .

❖ SCOPE

When I was analysing the data for forecasting I also saw that over the span of last 21 years the gold rate stays low around the starting of the year (mainly in February and march) and gradually increases throughout the year(the price is highest around august). The comparison is shown in the boxplot bellow



Boxplot

❖ APPENDIX

Date	Gram	Date	Gram	Date	Gram	Date	Gram
31-01-2000	398.13	31-10-2003	552.85	31-07-2007	864.53	29-04-2011	2103.84
29-02-2000	420.41	28-11-2003	570.89	31-08-2007	873.27	31-05-2011	2183.2
31-03-2000	401.3	31-12-2003	596.02	28-09-2007	923.87	30-06-2011	2203.93
28-04-2000	392.42	30-01-2004	604.27	31-10-2007	958.88	29-07-2011	2245.77
31-05-2000	389.03	27-02-2004	588.77	30-11-2007	1022.79	31-08-2011	2560.36
30-06-2000	410.45	31-03-2004	587.4	31-12-2007	1018.98	30-09-2011	2711.9
31-07-2000	405.37	30-04-2004	569	31-01-2008	1125.94	31-10-2011	2638.09
31-08-2000	403.08	31-05-2004	557.21	29-02-2008	1179	30-11-2011	2837.89
29-09-2000	403.75	30-06-2004	573.65	31-03-2008	1255.99	30-12-2011	2784.76
31-10-2000	402.24	30-07-2004	588.78	30-04-2008	1171.32	31-01-2012	2719.54
30-11-2000	400.02	31-08-2004	596.19	30-05-2008	1205.05	29-02-2012	2756.22
29-12-2000	408.01	30-09-2004	599.93	30-06-2008	1225.29	30-03-2012	2711.41
31-01-2001	397.14	29-10-2004	618.24	31-07-2008	1294.49	30-04-2012	2748.13
28-02-2001	391.64	30-11-2004	636.46	29-08-2008	1159.04	31-05-2012	2776.65
30-03-2001	394.18	31-12-2004	624.28	30-09-2008	1217.44	29-06-2012	2880.41
30-04-2001	391.45	31-01-2005	595.53	31-10-2008	1259.77	31-07-2012	2841.31
31-05-2001	410.81	28-02-2005	594.19	28-11-2008	1200.05	31-08-2012	2904.18
29-06-2001	408.28	31-03-2005	609.5	31-12-2008	1278.12	28-09-2012	3056.64
31-07-2001	405.39	29-04-2005	603.17	30-01-2009	1349.8	31-10-2012	2982.23
31-08-2001	412.66	31-05-2005	589.68	27-02-2009	1496.07	30-11-2012	3035.66
28-09-2001	434.09	30-06-2005	603.08	31-03-2009	1523.57	31-12-2012	2965.01
31-10-2001	436.92	29-07-2005	593.52	30-04-2009	1434.76	31-01-2013	2914.74
30-11-2001	426.13	31-08-2005	613.61	29-05-2009	1450.18	28-02-2013	2816.99
31-12-2001	424.56	30-09-2005	643.37	30-06-2009	1452.49	29-03-2013	2787.25
31-01-2002	437.14	31-10-2005	676.76	31-07-2009	1455.22	30-04-2013	2597
28-02-2002	462.28	30-11-2005	699.9	31-08-2009	1475.19	31-05-2013	2498.82
29-03-2002	460.57	30-12-2005	749.05	30-09-2009	1550.45	28-06-2013	2517.49
30-04-2002	475.81	31-01-2006	782.04	30-10-2009	1567.18	31-07-2013	2473.96
31-05-2002	495.13	28-02-2006	789.8	30-11-2009	1687.6	30-08-2013	2724.58
28-06-2002	505.23	31-03-2006	795.74	31-12-2009	1700.24	30-09-2013	2769.3
31-07-2002	490.95	28-04-2006	881.35	29-01-2010	1650.7	31-10-2013	2608.48
30-08-2002	484.41	31-05-2006	984.41	26-02-2010	1632.03	29-11-2013	2568.56
30-09-2002	496.82	30-06-2006	882.08	31-03-2010	1629.16	31-12-2013	2437.21
31-10-2002	492.03	31-07-2006	946.75	30-04-2010	1643.26	31-01-2014	2485.76
29-11-2002	494.83	31-08-2006	946.41	31-05-2010	1780.26	28-02-2014	2602.87
31-12-2002	513.55	29-09-2006	886.82	30-06-2010	1846.05	31-03-2014	2618.42
31-01-2003	549.6	31-10-2006	856.11	30-07-2010	1796.18	30-04-2014	2519.87
28-02-2003	550.96	30-11-2006	904.95	31-08-2010	1821.01	30-05-2014	2457.13
31-03-2003	521.52	29-12-2006	903.97	30-09-2010	1879.62	30-06-2014	2457.39
30-04-2003	499.86	31-01-2007	898.97	29-10-2010	1916.52	31-07-2014	2531.95
30-05-2003	538.34	28-02-2007	943.57	30-11-2010	1981.71	29-08-2014	2537.01
30-06-2003	535.01	30-03-2007	925.46	31-12-2010	2018.48	30-09-2014	2424.32
31-07-2003	521.53	30-04-2007	919.88	31-01-2011	1982.44	31-10-2014	2412.97
29-08-2003	531.13	31-05-2007	875.73	28-02-2011	2005.29	28-11-2014	2333.73
30-09-2003	558.32	29-06-2007	859.57	31-03-2011	2058.95	31-12-2014	2421.1

Date	Gram	Date	Gram	Date	Gram	Date	Gram
30-01-2015	2501.37	31-10-2016	2717.06	31-07-2018	2736.77	30-04-2020	4122.19
27-02-2015	2448.36	30-11-2016	2685.04	31-08-2018	2688.83	29-05-2020	4175.07
31-03-2015	2367.5	30-12-2016	2512.02	28-09-2018	2782.49	30-06-2020	4218.62
30-04-2015	2417.09	31-01-2017	2611.43	31-10-2018	2876.05	31-07-2020	4442.6
29-05-2015	2459.34	28-02-2017	2660.17	30-11-2018	2820.49	31-08-2020	4726.84
30-06-2015	2425.5	31-03-2017	2606.65	31-12-2018	2843.57	30-09-2020	4542.88
31-07-2015	2312.4	28-04-2017	2625.94	31-01-2019	2939.98	30-10-2020	4489.58
31-08-2015	2336.41	31-05-2017	2579.84	28-02-2019	3023.02	30-11-2020	4450.51
30-09-2015	2394.6	30-06-2017	2611.38	29-03-2019	2907.31	31-12-2020	4395.41
30-10-2015	2425.64	31-07-2017	2561.76	30-04-2019	2870.97	29-01-2021	4387.96
30-11-2015	2310.3	31-08-2017	2637.69	31-05-2019	2882.71	26-02-2021	4231.58
31-12-2015	2288.68	29-09-2017	2724.96	28-06-2019	3033.5	31-03-2021	4021.66
29-01-2016	2377.31	31-10-2017	2676.75	31-07-2019	3124.07	30-04-2021	4223.46
29-02-2016	2632.69	30-11-2017	2674.34	30-08-2019	3426.45	31-05-2021	4361.07
31-03-2016	2684.04	29-12-2017	2607.63	30-09-2019	3468.08	30-06-2021	4338.09
29-04-2016	2655.29	31-01-2018	2724.5	31-10-2019	3414.64	30-07-2021	4330.67
31-05-2016	2710.27	28-02-2018	2757.17	29-11-2019	3378.1	31-08-2021	4254.76
30-06-2016	2761.81	30-03-2018	2771.03	31-12-2019	3376.76	30-09-2021	4204.75
29-07-2016	2889.62	30-04-2018	2820.04	31-01-2020	3577.09	29-10-2021	4283.61
31-08-2016	2884.89	31-05-2018	2829.81	28-02-2020	3671.6	30-11-2021	4359.25
30-09-2016	2846.41	29-06-2018	2793.52	31-03-2020	3812.19	31-12-2021	4335.99

THANK YOU