Time Series Analysis of Gold Prices

Authors: Yuebi Li, Yuran Liao

Date: March 11, 2025

Abstract

This project is to analyze the historical trend and seasonal patterns in gold prices from 2014 to 2024 and forecast future prices. We applied a series of useful concepts from time series to extract trends from the dataset and fit an ARMA model for the residuals. The results of the model selection through the AIC algorithm showed that the ARMA (1,0) model provides the best residual model. Furthermore, we forecast gold prices for 255 trading day after data ends, and the results indicate that the price will decrease at first and then increase slightly.

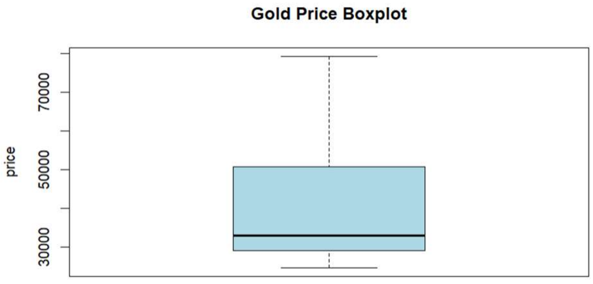
Introduction

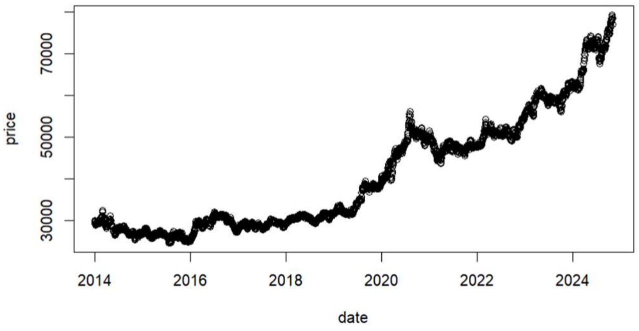
Our research question is to observe the trend of gold over the past eleven years and predict what the price of gold will be like in 2025. We find this research topic interesting and meaningful because the price of gold can reflect the state of the global economy, and it is also a large indicator of the financial markets. Being able to accurately predict the price of gold can help financial institutions and investors make better decisions. The price of gold usually expands or declines because of major events around the globe, so the data will have distinctive characteristics or seasonal trends, making it suitable for analyzing the dataset using time series. We perform time series analysis by using techniques such as Box-cox transformation, ACF and PACF plot analysis, linear regression, poly regression, fourier algorithm and Augmented Dickey-

Fuller (ADF test). This study will forecast the price of gold in 2025 and then discuss the commodity price fluctuations that may cause.

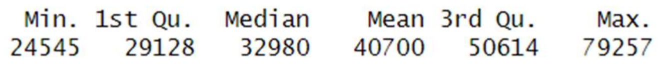
Data Description

For this project, we use a dataset containing gold prices for each trading day from January 1, 2014, to November 6, 2024. A trading day includes all regular business days except for weekends and holidays. The recorded unit in the dataset is in Indian Rupees (INR) per 10 grams of gold (e.g., 79,257 INR/10g). The dataset includes features such as final prices, opening price, closing price, and the lowest and highest prices of the day for each trading day. We use the final price as the daily price of gold to analyze the trend because it reflects the actual market value at the end of each trading day, making it more stable and representative. This dataset was acquired on Kaggle, and the Raw Data Source is from India's largest commodity trading website, MCX Market. Since MCX Market is an official and publicly available trading platform, the data is guaranteed to have a high level of reliability and has been collated and shared on Kaggle to make it more analyzable.





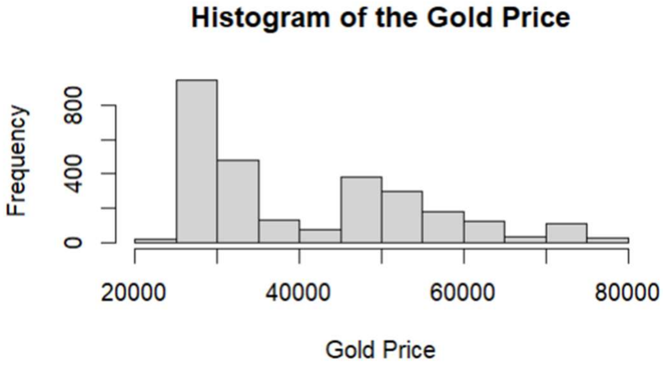
One of the interesting features that we found is that through boxplot analysis and IQR calculation, we found that there is no outlier in this dataset. Through scatter plot, we observed that there is a clear and steady upward trend in the price of gold over the period from 2014 to 2024. Also, by the Augmented Dickey-Fuller Test, the p-value is greater than alpha at 5% level, indicating that the gold price data is not stationary and requires further transformation or decomposition to remove the trend.



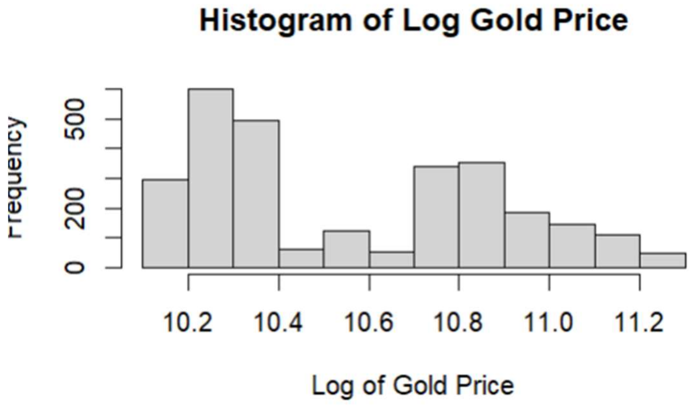
```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 24545   29128   32980   40700   50614   79257 
```

The summary statistics show that the gold prices range from 24,545 to 79,257, with a median of 32,980 and a mean of 40,700. The interquartile range (IQR) spans from 29,128 (1st quartile) to 50,614 (3rd quartile), indicating a right-skewed distribution.

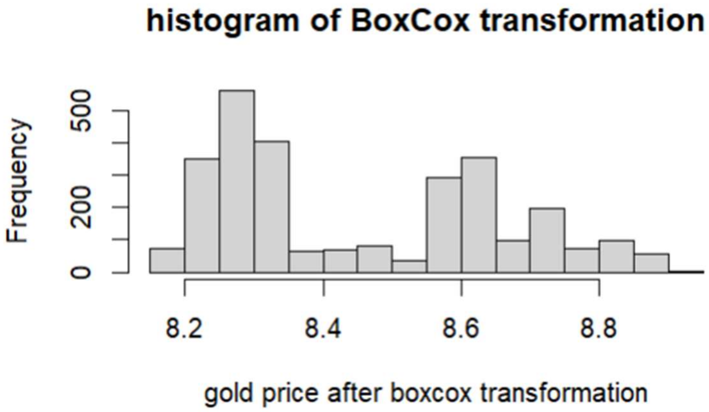
Data Analysis



We first plotted a histogram to check symmetricity. According to the histogram above and statistics shown in the data description section, this dataset is right-skewed. To stabilize variance and decrease skewness, we tried to transform the data and make it symmetrical.

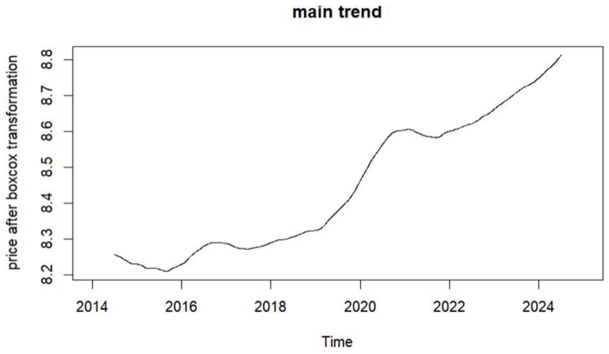


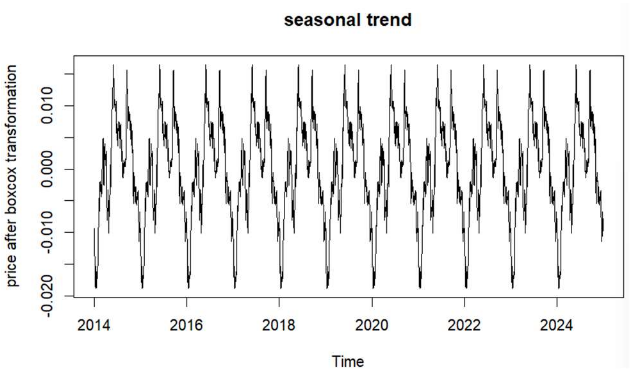
We tried log transformation at first. However, the histogram of log gold price shows that the dataset is still not symmetrical. We performed an augmented Dickey-Fuller test and got p-value = 0.3513. We cannot reject the H_0, which shows the data is not stationary after log transformation.

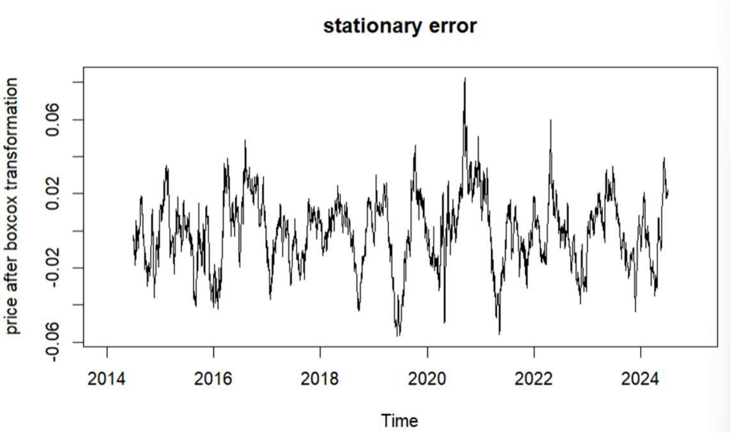


We then tried to use BoxCox transformation. The histogram is not symmetrical as well. We got p = 0.3328 by performing an augmented Dicky-Fuller test. We cannot reject the H_0 or conclude the data is stationary after BoxCox transformation. However, the histogram shows less skewness than not transforming data. So, we decided to keep the BoxCox transformation, since it makes the data less skewed and has a p-value smaller than log transformation.

The next step is performing classical decomposition. The dataset does not conclude all prices of 365 days a year but is recorded by trading days. Therefore, we calculated the mean trading days for 11 years, which is about 255 trading days a year. We then performed classical decomposition with 255 trading days a year.

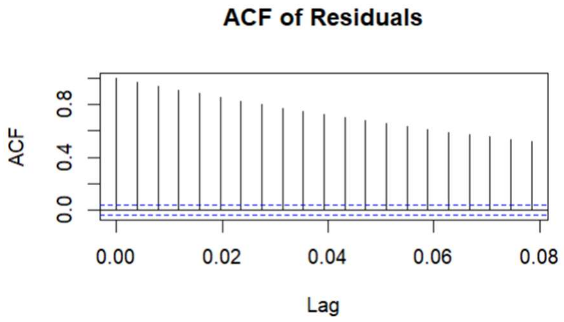


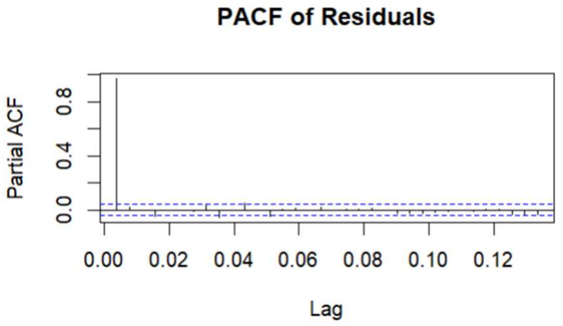




After the classical decomposition, the main trend shows an increasing trend. The increase of the main trend is relatively flat before 2019 and increases quickly after then. This increase may be caused by COVID-19, the viruses that swept the world. People may buy more gold for safe-haven purposes. The seasonal trend exhibits a repeating pattern. Gold prices are relatively lower at the beginning of the year and relatively higher in the middle of the year. This shows that gold trading may be off-season and high season.

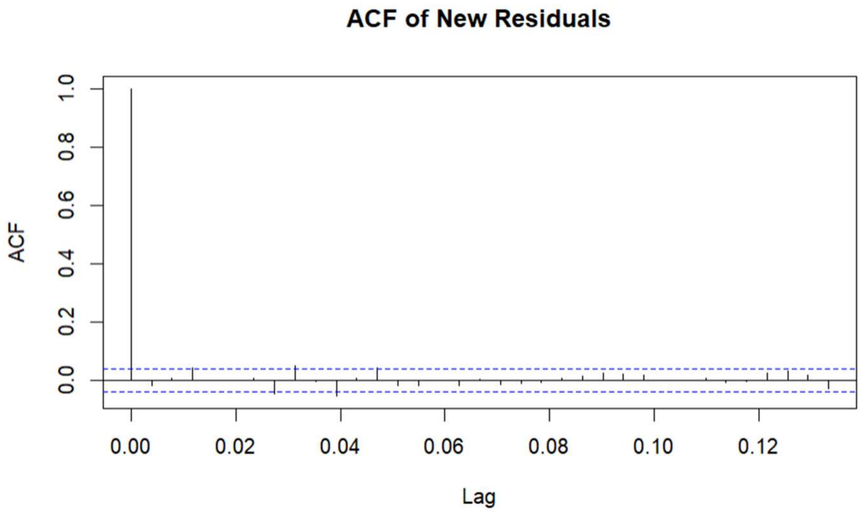
The “stationary error” graph shows that the decomposed residual fluctuates up and down around 0, with no obvious trend or seasonal structure, and the variance is roughly stable. This shows that after the Box-Cox transformation, seasonality and trends are stripped off well. The remaining noise terms are relatively stable, indicating that the decomposition effect is good. We performed the Augmented Dickey-Fuller Test and got p = 0.01 and agreed with the result we got from the plot. We are able to reject H_0 and conclude that the stationary error is stationary.

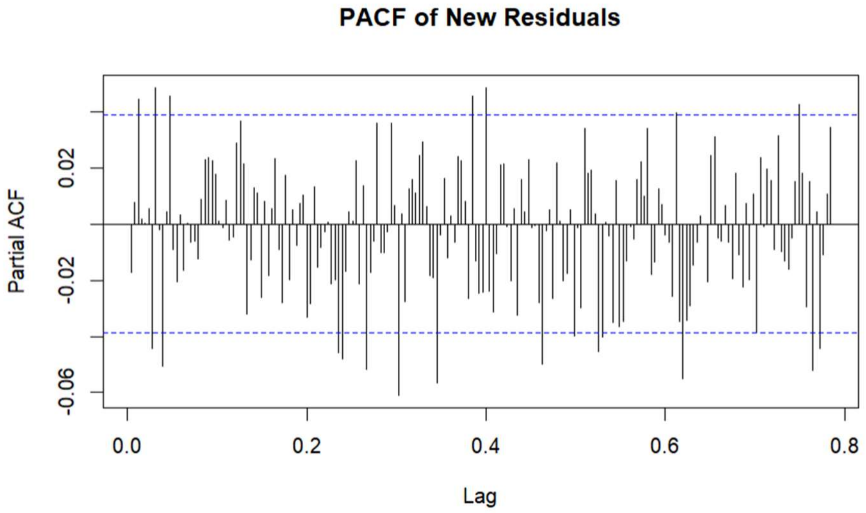




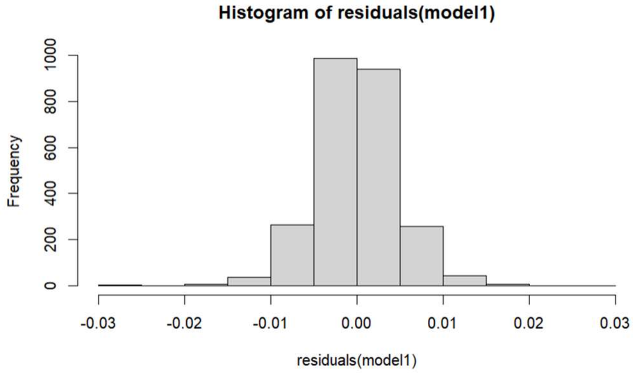
We plotted ACF and PACF plots to further detect possible autocorrelation. We found that the ACF plot tails off while the PACF plot cut off, which means that we can probably fit the possible pattern with an AR model. We performed the Box-Ljung test and got a p-value < 2.2e-16. We have strong evidence to reject the H0 and conclude that the residuals have autocorrelation. This result agrees with what we conclude from the ACF and PACF plots. So, we are going to fit an ARMA model.

By applying auto.arima(), we got ARMA(1, 0) model, which is AR(1) model. The result agrees with our guess.



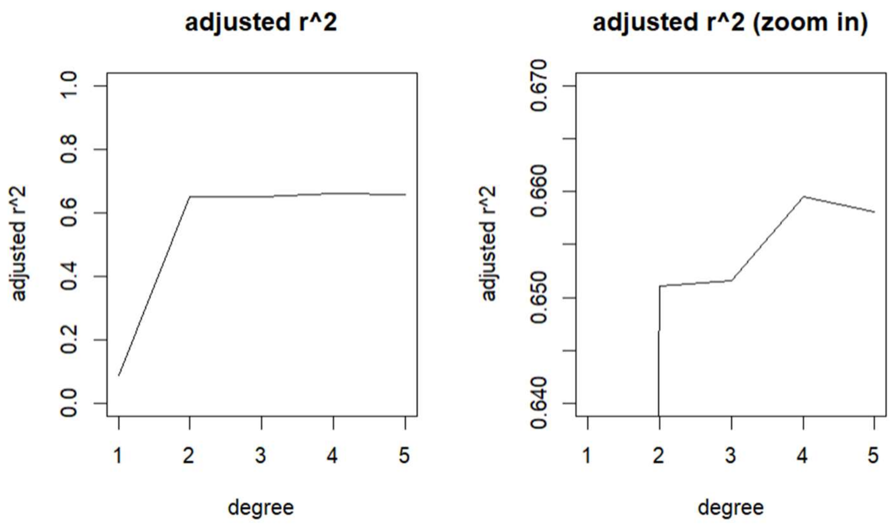


We plotted ACF and PACF plots for residuals of the ARMA(1, 0) model. The autocorrelation coefficients for most of the lags in the ACF plot are within the blue dashed line (roughly the 95% confidence interval), with no significant peaks or valleys above the threshold. This indicates that there is no significant autocorrelation in the residual series. The PACF generally fluctuates randomly within the blue confidence interval, which means there is no significant autocorrelation in the residuals. This may indicate that the residuals are white noise.

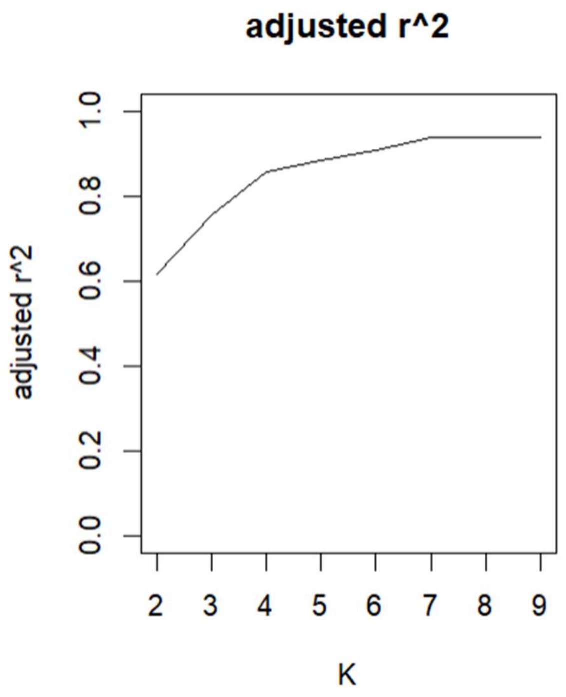


We plotted a histogram for the ARMA(1, 0) model, in which this distribution of residuals is approximately normal. We performed the Box-Ljung test and got a p-value = 0.3847. we cannot reject H_0, and with the above evidence, we can conclude that the residuals are white noise. The ARMA(1, 0) model performs well.

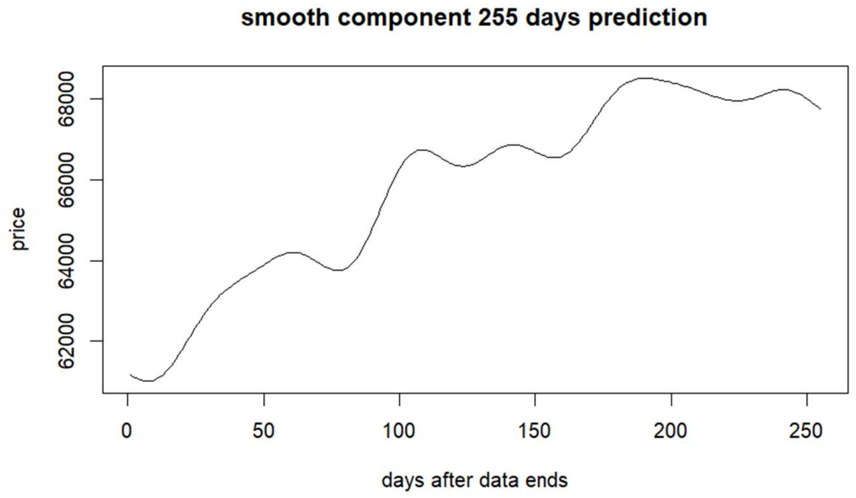
To make price predictions for the following 255 trading days (about a year), we need to fit the model for the main and seasonal trends. We fit a linear model for the main trend, with adjusted R-squared = 0.934. This shows that 93.4% of information in the main trend is explained by this model. This result meets our expectations.



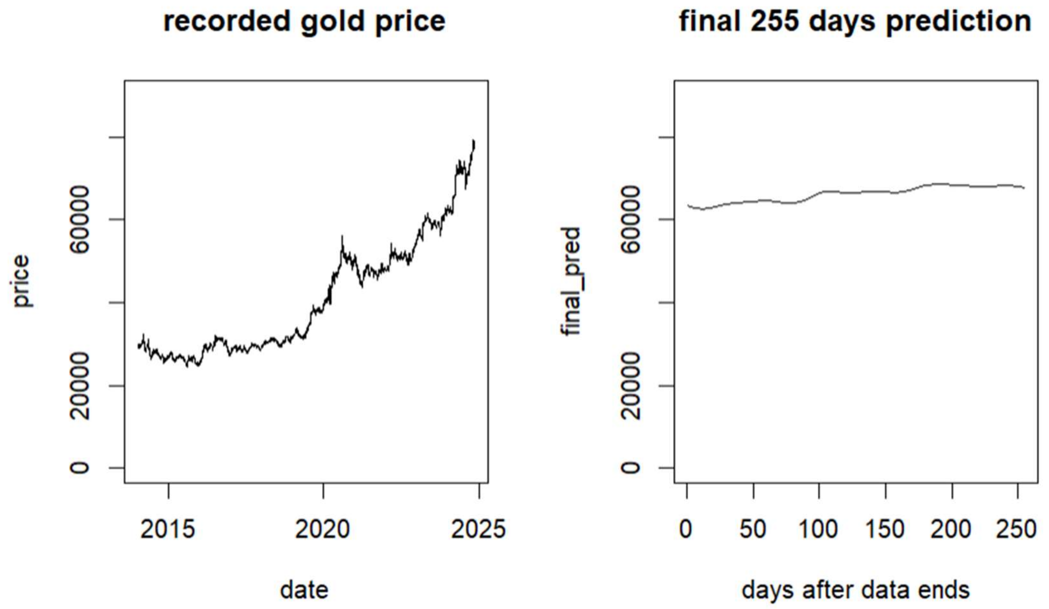
For the seasonal trend, we first tried to fit a linear model, and polynomial regression model from degree 2-5. The linear model has adjusted R^2 less than 0.1, which is very low. The adjusted R^2 increases quickly from degree 1 to degree 2 to about 0.7 and then increases relatively flat. We think linear or polynomial regression may not be the most reasonable way to explain seasonal trends, so we then tried Fourier algorithm.



From K = 2 to K = 9, we can see an obvious increasing trend, and the adjusted R^2 are more than 0.9 after K = 7. We decided to pick K = 7, since the adjusted R^2 increases slower after K = 7.



By adding the prediction results of main trends and seasonal trends, we are able to plot the prediction results of the smooth components for 255 trading days after the data ends on 11/06/2024.



The results show in the following 255 trading days, the price may decrease back to about 60,000 INR per 10 grams and then increase slowly.

Discussion

In the forecast section, we get the final prediction of the gold price for the year 2025 by adding up the smooth part prediction and the rough part prediction. Our predicted average gold price for 2025 is 66,149 INR, with a steady upward trend from a minimum value of 62,642 INR to a maximum value of 68,534 INR. However, by comparing the price of gold in the first three months of this year, we found that the price of gold has risen to more than 80,000 INR, which is still somewhat different from our forecast.

We have analyzed that this may have happened because the model we fitted to the main trend forecast was linear, so it might miss some detailed fluctuation change information on the trend. Another reason why our forecasted gold price is lower than the actual gold price is the impact of unexpected major external events. In the first three months of 2025, there were many major international events that could have affected the price of gold, such as the change of leadership in the US government, the change in the international cooperation situation, the end of the Russian-Ukrainian conflict, and so on, which were not included in our model's prediction.

We believe that the direction of improvement for this could be to re-fit the prediction model of the main trends with a polynomial model or with Fourier algorithm to observe the best performing degree. Alternatively, one could also consider introducing some macroeconomic factors, such as global commodity prices, to enhance the predictive skill of the model. Eder points out that when the price of gold rises, the prices of world commodities (both essential and non-essential) rise simultaneously and to varying degrees (Eder, 1938). This is an interesting pattern and can be an external factor that can be introduced into our model fit, thus giving the model more indicators to consider. In this way, we can make the main trend prediction model more able to capture more fluctuation information, thus making our overall prediction model more accurate.

Conclusion

Through the behavior analysis of the ACF and PACF plot, we have observed that ARMA(1,0) is probably the best model for this dataset, and the selection criteria of AIC supported this conclusion. In creating the forecasting model, we forecasted the main trend, the seasonal trend and the rough part separately, and the final combination formed the final forecast for the gold price in 2025. The main trend model achieved an R^2 of 93.41%, while the seasonal model also reached an R^2 of 94.07%. With effective overfitting control, we believe that this is a good smooth forecasting model that captures most of the information in the data and provides a reasonable prediction of future gold price trends.

The limitations of the data include the inability to fully predict future gold price changes based solely on the past 11 years of data, as well as the absence of the consideration of external factors in the analysis. Although the R^2 of the model is high, it is also possible that the model's over-reliance on historical data patterns may not be a good predictor of changes in the future. Our future improvements could be to try to improve the predictions by using machine learning methods or to incorporate external economic factors, such as inflation and other relevant indicators, into the model to improve the predictive power.

Reference

Eder, G. J. (1938). Effect of gold price changes upon prices for other commodities. Journal of the Royal Statistical Society, 101(1), 173. https://doi.org/10.2307/2980657

Appendix

The following appendix contains the full R code used in this project.