Gold Price Analysis Report (2013 - 2023)

Team members:

سجى سمير زريقات

Dataset Overview

This comprehensive dataset offers a decade's worth of insights into gold price trends, spanning from 2013 to 2023. It meticulously captures:

- 📈 Daily opening and closing prices
- 🔺 Highs and lows for each day
- 🔁 Trading volume

Such a wealth of information is instrumental for those seeking to analyze or visualize market dynamics over this ten-year period.
Source: Kaggle

🛠️ Data Preprocessing

- Removed punctuation and converted textual values in 'Price', 'Open', 'High', 'Low', 'Vol.', and 'Change %' columns to numerical values.
- Filled missing values in trading volume using the mean.
- Added useful features:
  - **Price Range:** Difference between high and low of the day.
  - **Date Features:** Year, Month, Day, Day of Week.
- Sorted data chronologically.
- Identified and analyzed outliers using the IQR method.

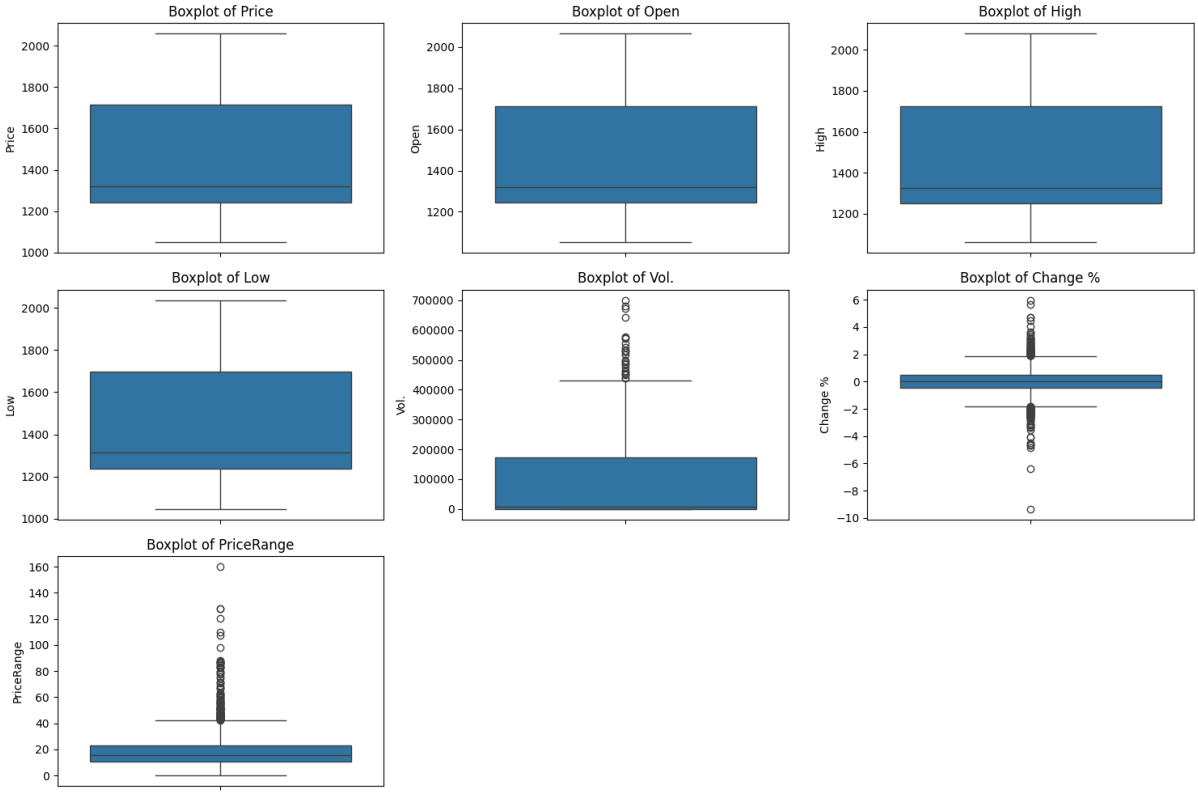
📉 Normalization

To prepare the data for machine learning models, **MinMaxScaler** was applied on numerical features, which scaled all values between **0 and 1**.
This normalization step is essential to ensure all features contribute equally to the model and avoid dominance due to different scales.

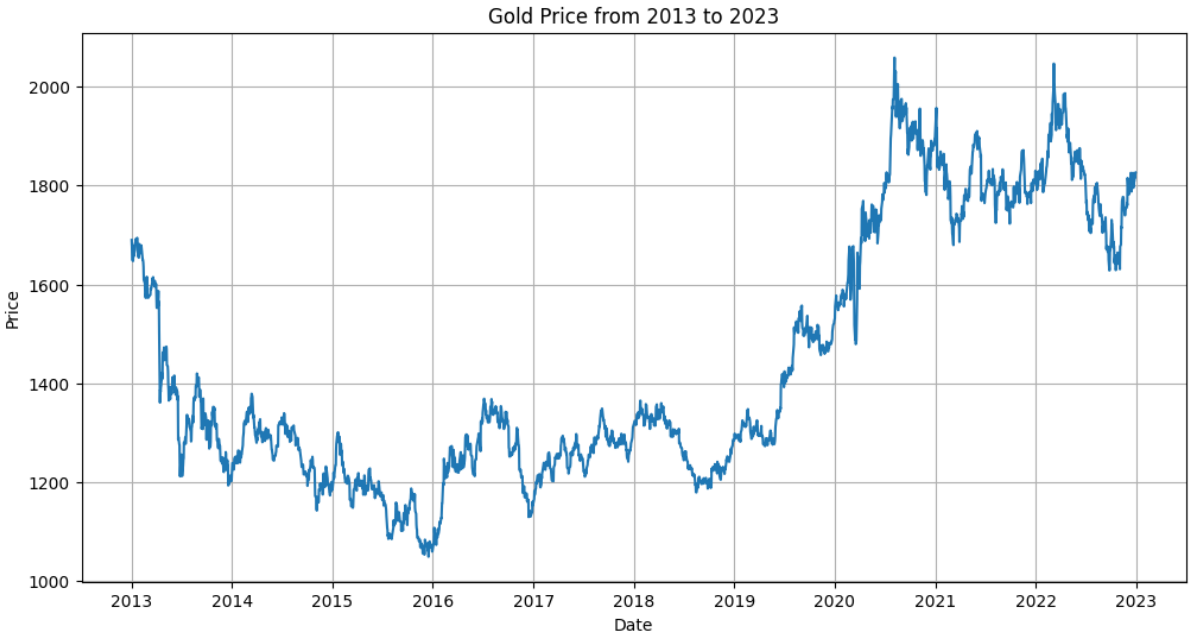
✅ **Note:** Normalization was applied **only before training the model**, and not on visualization data to keep plots interpretable.

📈 Exploratory Data Analysis

OUTLIERS

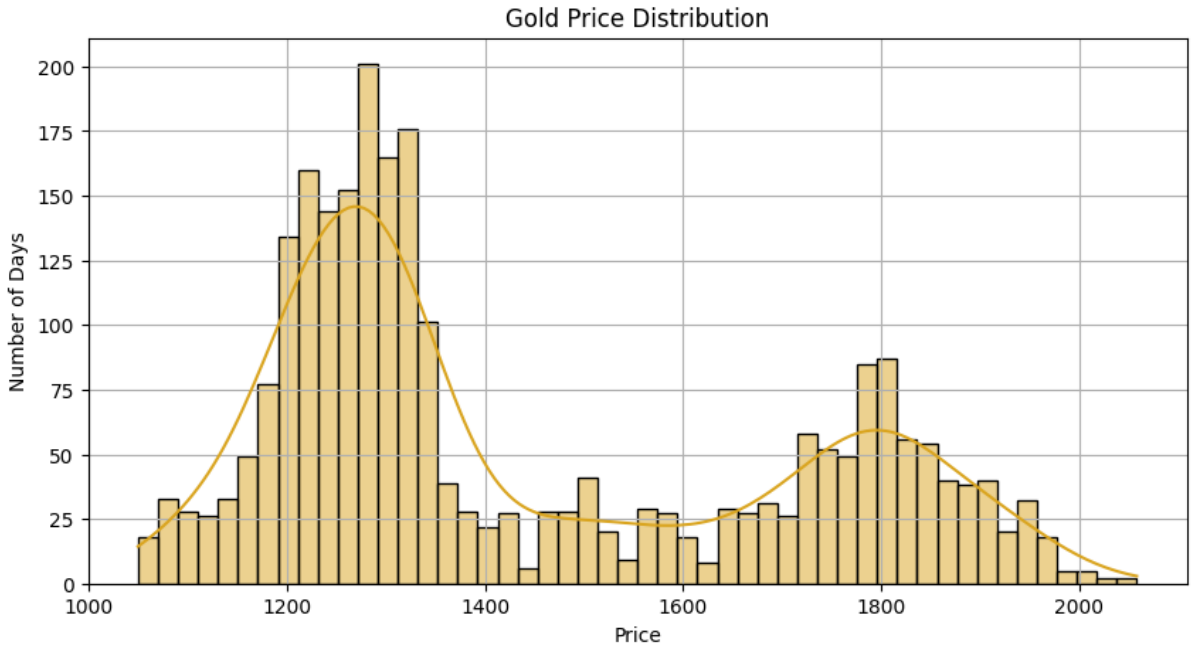


- **Trend Visualization:**

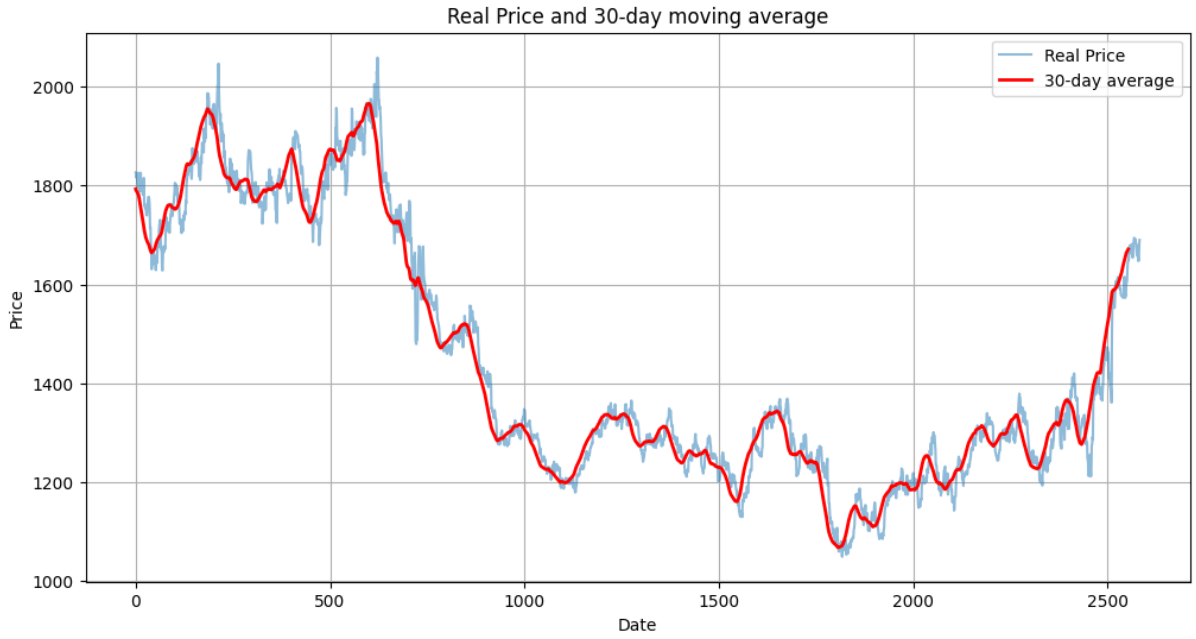


- The chart shows that gold prices experienced significant fluctuations between 2013 and 2023, beginning with a gradual decline until 2015, followed by relative stability until 2018. From 2019, prices began to rise sharply, peaking in 2020 with the COVID-19 pandemic, when they exceeded $2,000. After that, they maintained high levels with continued fluctuations due to inflation and global crises such as the

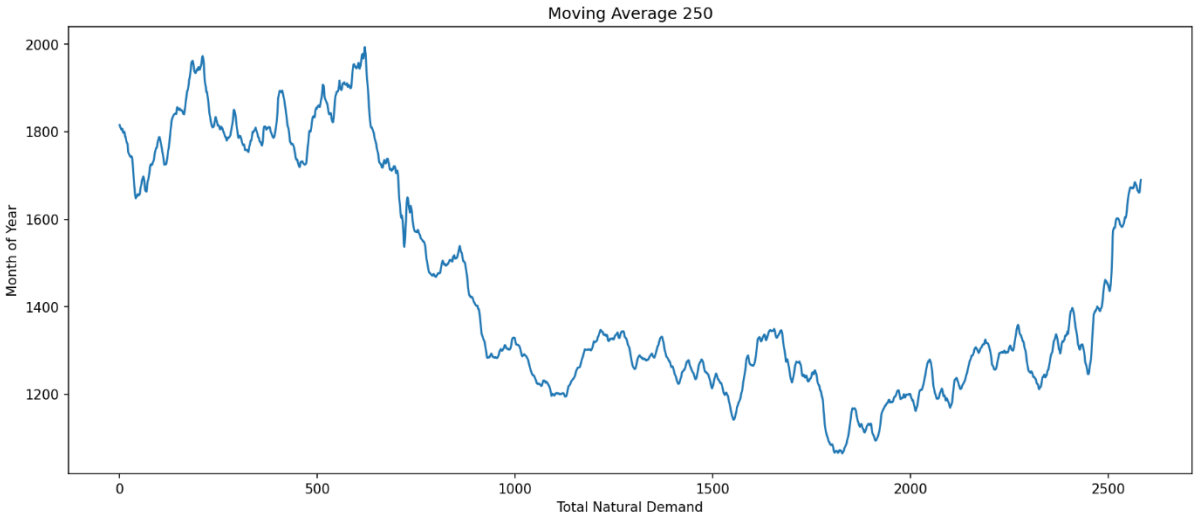
Russia-Ukraine war. The overall trend reflects a gradual increase in the value of gold during this period.



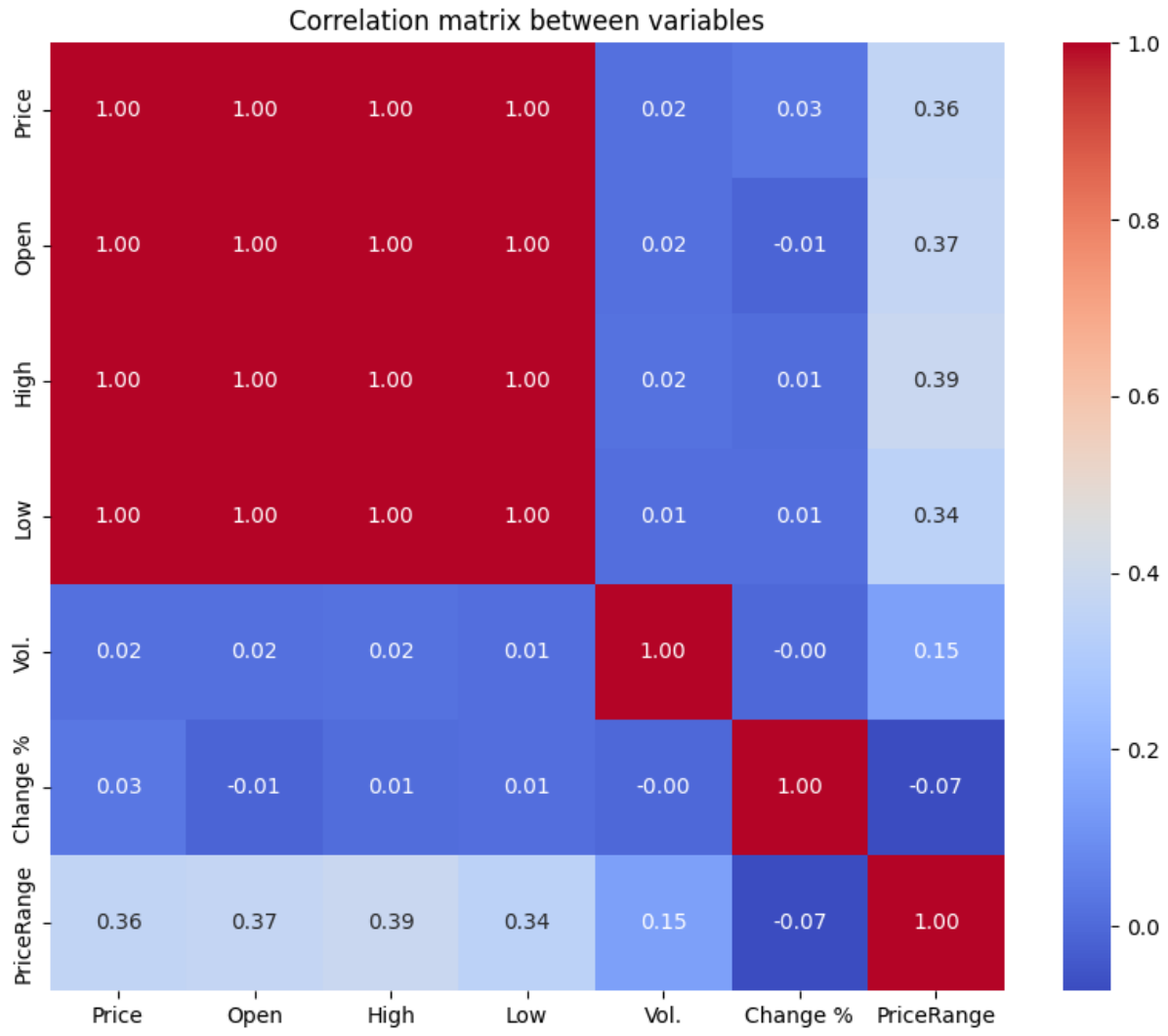
- The chart shows the distribution of gold prices over the period from 2013 to 2023. It shows that prices were concentrated mainly in the $1,200-$1,350 range, indicating that this range was the most frequent over the past ten years. Another peak in the distribution appears around $1,750-$1,850, reflecting the highs that have occurred in recent years. The shape of the distribution is asymmetrical and skewed to the right (positive skewness), indicating fewer days with very high prices. These periods are often associated with global economic crises.**EMA (Exponential Moving Average):** Captures recent trends more sensitively.



- The chart clearly shows that gold prices fluctuate significantly in the short term, but the red line (moving average) helps show long-term trends: Periods of gradual rise or steady decline can be observed. This average helps identify the overall market trend and identify entry or exit points for investors.



- The EMA250 chart shows that gold is in a strong long-term uptrend, especially since 2019. The recent trend indicates a revival in demand and the potential for this uptrend to continue, albeit with greater volatility than in the past.



- **Correlation Matrix:** Strong correlations observed between 'Price', 'Open', 'High', and 'Low'.

-

Model building

1- **Split the data:**

We don’t split the data randomly since it’s not allowed the future to predict the past (we take subset of the data from the last dates)

Data distribution:

Training: 2323 10% of it for validation

Testing: 260

window size of 60 use to create sliding window (The use pf prior time steps to predict the next time step in this way time series can be expressed as supervised learning)

the training shape after applying sliding window: X_train ->(60, 2263, 1)

y_train -> (2263, 1)

the testing shape: X_test -> (60, 200, 1)

y_test (200, 1)

2_model architecture:

Model: "functional"

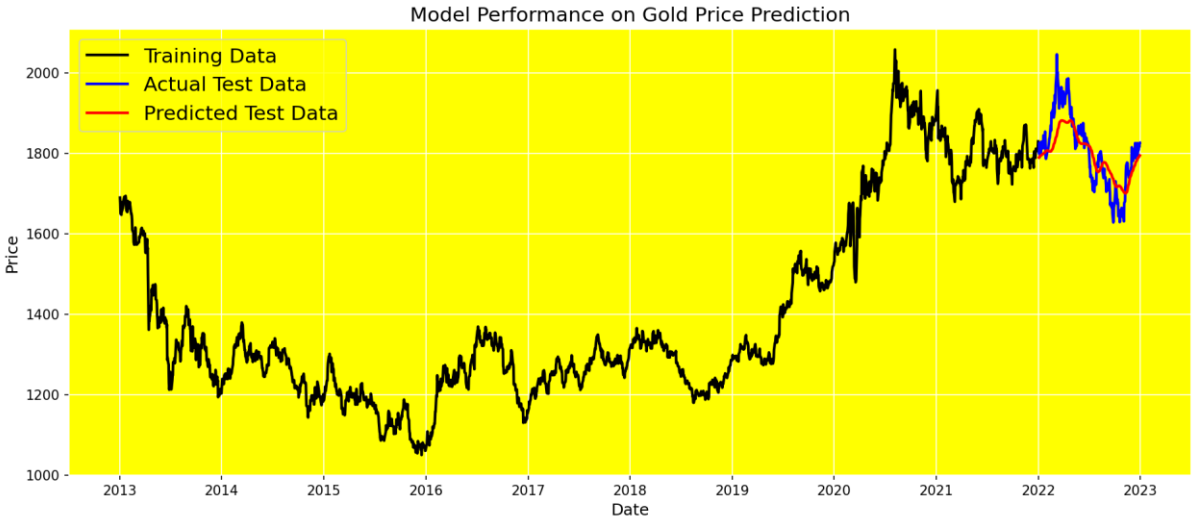
Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 60, 1)	0
lstm (LSTM)	(None, 60, 64)	16,896
dropout (Dropout)	(None, 60, 64)	0
lstm_1 (LSTM)	(None, 60, 64)	33,024
dropout_1 (Dropout)	(None, 60, 64)	0
lstm_2 (LSTM)	(None, 64)	33,024
dropout_2 (Dropout)	(None, 64)	0
dense (Dense)	(None, 32)	2,080
dense_1 (Dense)	(None, 1)	33

Total params: 85,057 (332.25 KB)
Trainable params: 85,057 (332.25 KB)
Non-trainable params: 0 (0.00 B)

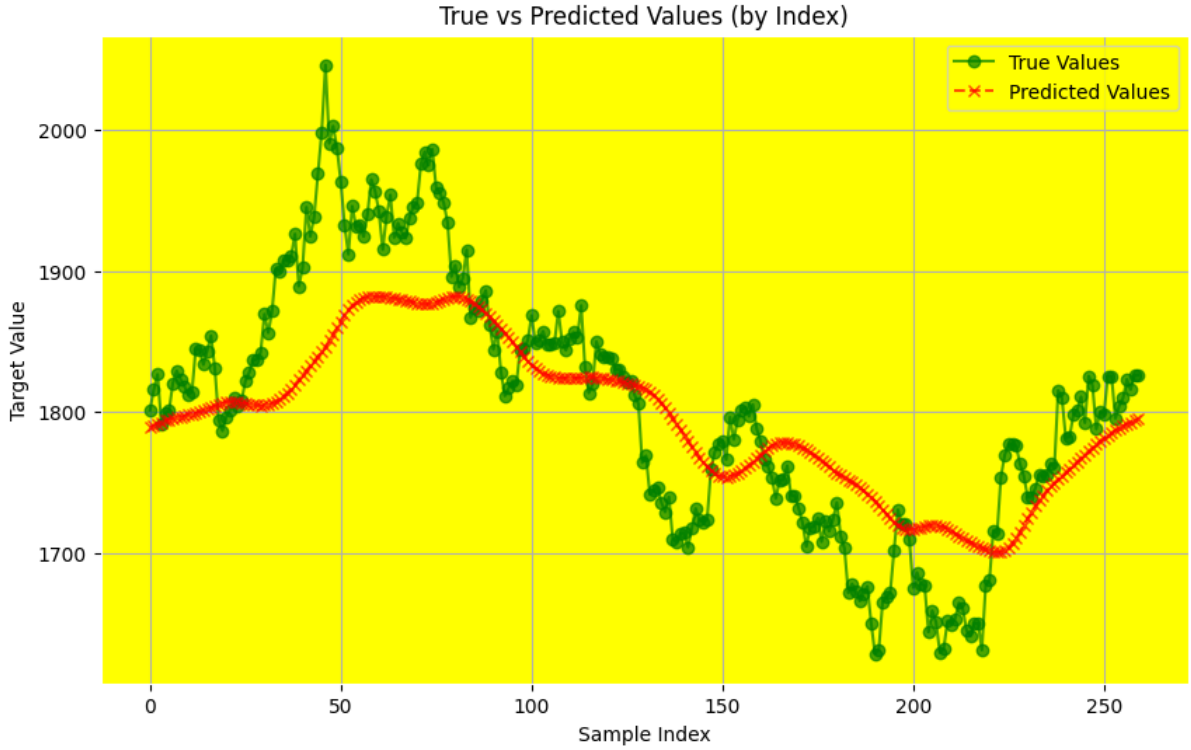
3_Model evaluation on the test set:

Metric	Value	Interpretation
Test Loss (MSE)	0.0029	Very low error (excellent fit). Predictions are, on average, ±0.054 units off (√MSE).
Test MAPE	5.82%	Strong accuracy. Predictions are off by ~5.8% on average.
Test Accuracy	94.18%	High accuracy (100% – MAPE). Model explains most variability.
R^2 Score	0.6469	Decent. ~64.7% of price variance is explained by the model.
Mean Residual	+9.92	Slight underprediction bias (model consistently predicts ~10 units below true values).
Std Dev of Residuals	53.19	~68% of predictions fall within ±53.19 units of true values.
Max Residual	+200.01	Worst overprediction: 200 units above true value (check outliers).
Min Residual	-108.42	Worst underprediction: 108 units below true value.

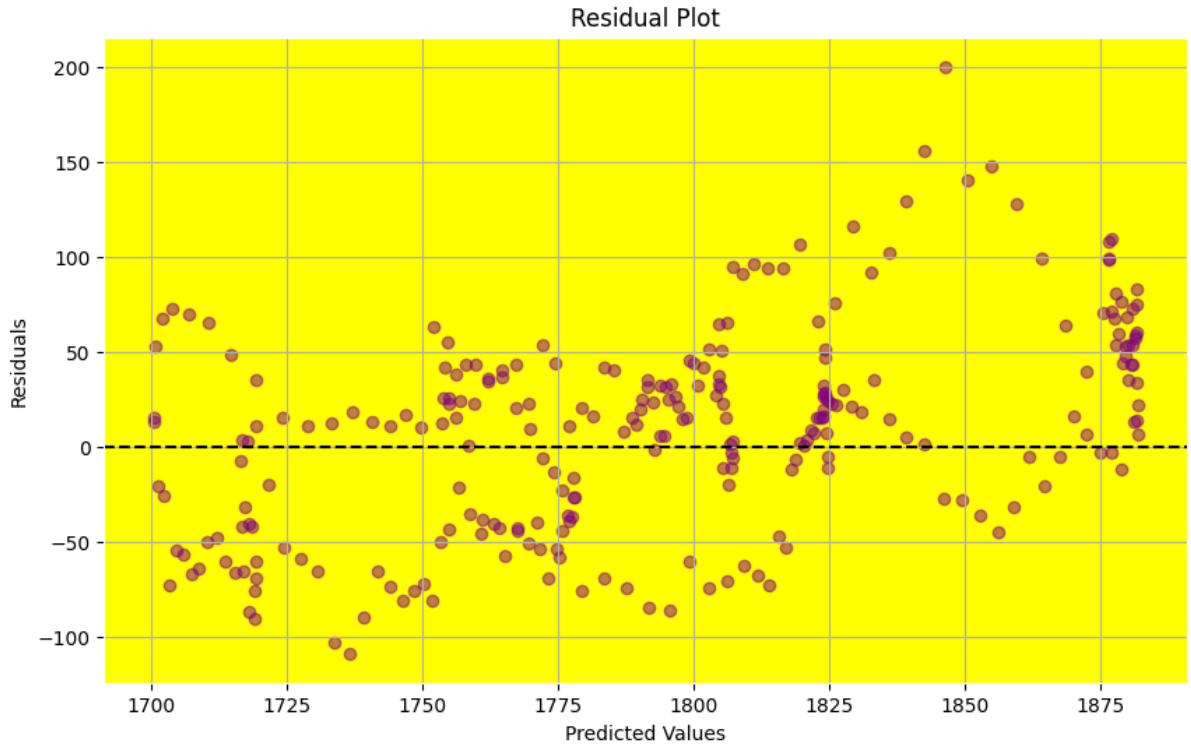
4_ Visualizing the result:



More closer version of the prediction:



Plotting the residual to see that it's scattered and there's no relationship between them:



Plotting the distribution of actual value and the predicted one to see to which extent they intersected the more the better