FD-OS

Price Prediction for Gold in KSA [2020-2026] Using Time-Series Analysis & Confidence Intervals

Project Overview:

This project aims to predict daily gold prices (24K per Gram) in Saudi Arabia (KSA) for the year 2026. By analyzing historical data from 2020 to 2025, the project utilizes a hybrid time-series approach combining ARIMA (for trend analysis) and GARCH (for volatility clustering) to generate a robust forecast with risk-adjusted Confidence Intervals.

Objectives:

- **Data Extraction:** Scrape and clean historical gold price data from gold.sa/en.
- **Analysis:** Perform Exploratory Data Analysis (EDA) to understand volatility and trends.
- **Modeling:** Develop a time-series model to handle non-stationary data and volatility clustering.
- **Forecasting:** Predict 2026 prices with a 95% Confidence Interval to quantify risk for retailers and investors.

Technology Stack:

The project was developed using the following tools and libraries:

Category	Tools/Libraries	Purpose
Environment	VS Code, Jupyter Notebook	Interactive data analysis and coding.
Language	Python	Core scripting.

Data Handling	Pandas, Numpy, JSON, StringIO	Data manipulation and numerical operations.
Web Scraping	Requests	Handling HTTP headers and network requests.
Modeling	Statsmodels, Sklearn	ARIMA, GARCH, and statistical testing.
Visualization	Matplotlib, Seaborn	Static and statistical plotting.

1. Data Acquisition & Cleaning:

Source Selection via AI Assistance

Large Language Models (LLMs) were leveraged to identify potential data sources with historical KSA gold prices. A strict selection process involved checking robots.txt files to ensure ethical scraping compliance.

- **Selected Source:** https://gold.sa/en.
- **Reasoning:** Unlike competitors with restrictive API limits and clauses against automation, gold.sa allowed for full access (verified via robots.txt: Allow / and Privacy Policy review).

Technical Implementation: Reverse-Engineering

The website utilized dynamic content, making standard HTML parsing insufficient.

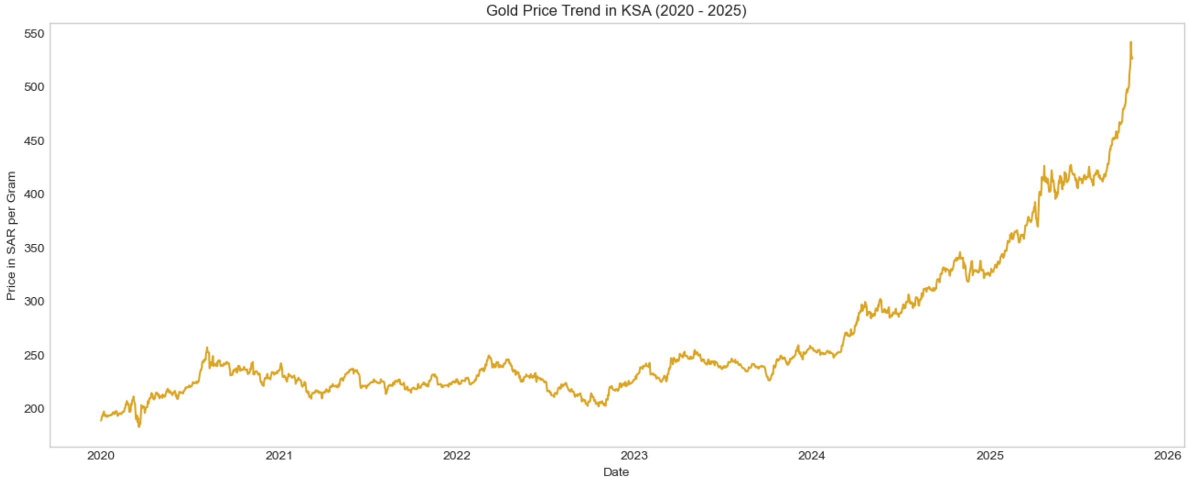
- **Method:** Intercepted network requests to identify the exact JSON endpoint.
- **Pagination:** Implemented a while loop to iteratively fetch historical data.
- **Ethical Constraints:** Added a 7-second delay (time.sleep(7)) between requests to prevent server throttling.
- **Cleaning:** Parsed nested JSON using Pandas lambda functions and converted dates using pd.to_datetime to ensure a valid time-series index.

2. Exploratory Data Analysis (EDA):

Dataset Status: 2,113 records (2020-2025) with zero nulls or duplicates.

Key Insights

- **Trend:** The data exhibits a clear long-term upward trend, confirming the series is **non-stationary**.



- **Historical Context:**
  - **2020:** Prices surged (~25%) due to the COVID-19 pandemic.
  - **2021-2022:** A gradual decrease due to economic recovery and interest rate hikes.
  - **2023-2025:** Volatility returned, driven by global tensions, establishing gold as a "safe haven" asset.
- **Risk Analysis:**
  - **Standard Deviation:** 66.13 SAR
  - **Interquartile Range (IQR):** 66.39 SAR
  - **Business Implication:** The high spread indicates a high-risk environment requiring dynamic pricing models.

3. Modeling Strategy:

Why ARIMA?

Linear Regression was rejected because it assumes independence between data points. Gold prices are time-dependent and non-stationary. **ARIMA** (AutoRegressive Integrated Moving Average) was selected to handle this time dependency and differencing)

Preprocessing & Parameters

1. **ADF Test:** The raw data was non-stationary (P-value: 1.0).
2. **Differencing:** Applied 1st order differencing (*d*=1). The re-test confirmed stationarity (P-value: 0.0).
3. **ACF/PACF Analysis:** showed no significant spikes after Lag 0, suggesting a model of *p*=0, *q*=0.
4. **Baseline Model: ARIMA(0, 1, 0)** (Random Walk).

The Limitation of Pure ARIMA

Upon evaluation using a **Chronological Split** (Last 90 days for testing) to avoid data leakage, the baseline ARIMA(0, 1, 0) model failed to capture market volatility.

- **Result:** It predicted a flat line (Random Walk).
- **Error Rate:** MAPE of 6.45%.
- **Conclusion:** While ARIMA modeled the *mean*, it failed to model the *variance* (risk).

The Solution: ARIMA + GARCH Hybrid

To address the "Fat Tails" and "Volatility Clustering" observed in the Log Returns, a **GARCH(1, 1)** model was added.

Final Model Performance:

- **Model:** Constant Mean - GARCH(1, 1)
- **RMSE:** 43.86 SAR
- **MAPE:** 6.45%.
- **Benefit:** This hybrid approach allows for the calculation of a **Risk-Adjusted Confidence Interval**.

4. Forecasting Results (2026):

The final model projected the price of 24K Gold per Gram for the next 365 days.

Parameter	Value	Interpretation
Prediction Horizon	365 Days	Full year forecast for 2026.
Average Price	525.49 SAR/g	Mean expected price.
Lower Bound (95%)	438.01 SAR/g	Best-case / Low-risk scenario.
Upper Bound (95%)	612.96 SAR/g	Worst-case / High-risk scenario.

Challenges & Solutions:

Challenge	Root Cause	Solution
Dynamic Data Scraping	Standard HTML parsing returned empty data.	Reverse-Engineered network requests to find the hidden JSON endpoint and replicated HTTP headers.

API Rate Limiting	Server blocked rapid requests (403 Forbidden).	Implemented ethical delays (time.sleep(7)) and mimicked browser User-Agents.
Model Flatlining	ARIMA(0,1,0) assumes constant variance.	Integrated GARCH(1,1) to model heteroskedasticity (volatility clustering).
Forecast Date Errors	Jupyter memory contamination (truncated variables).	Performed a Full Kernel Restart and migrated code to a clean notebook to reset the memory state.

License & Disclaimer:

- **Data Source:** https://gold.sa
- **Usage:** This project is for educational and analytical purposes. Financial decisions should not be based solely on this model.