Financial Data Analysis:
Gold Price Prediction in Sync with World Events

Vindhya Jain, Avanti Mittal, Khushi Bhardwaj

Introduction

Historically, gold had been used as a form of currency in various parts of the world including the USA. In present times, precious metals like gold are held with central banks of all countries to guarantee re-payment of foreign debts, and also to control inflation which results in reflecting the financial strength of the country. Recently, emerging world economies, such as China, Russia, and India have been big buyers of gold, whereas the USA, SoUSA, South Africa, and Australia are among the big seller of gold.

Forecasting rise and fall in the daily gold rates can help investors to decide when to buy (or sell) the commodity. But **Gold prices are dependent on many factors such as prices of other precious metals, prices of crude oil, stock exchange performance, Bonds prices, currency exchange rates, etc**.

We plan to use this Kaggle dataset to start off. It contains data from November 18th 2011 to January 1st 2019 from various sources. The data has 1718 rows in total and 80 columns in total. Data for attributes, such as Oil Price, Standard and Poor's (S&P) 500 index, Dow Jones Index US Bond rates (10 years), Euro USD exchange rates, prices of precious metals Silver and Platinum and other metals such as Palladium and Rhodium, prices of US Dollar Index, Eldorado Gold Corporation and Gold Miners ETF are present.

Data Reliability Framework

Designing a data reliability framework for financial data analysis in a Database Management System (DBMS) involves several key components to ensure data accuracy, consistency, and trustworthiness. We plan to use a couple of Kaggle datasets and incorporate them and assess inconsistencies.

1. Data Governance

1.1 Data Ownership: Assign responsibility for gold price data management and ensure accountability for its accuracy and consistency.

1.2 Data Quality: Enforce rules to prevent missing or duplicated price records and ensure data is accurate, complete, and timely.

1.3 Access Control: Implement role-based access to secure sensitive gold price data, restricting access to authorized personnel only.

1.4 Data Integrity: Use encryption and validation to protect the integrity of gold price data during collection, storage, and transmission.

1.5 Standardization: Ensure consistent units (e.g., per ounce) and date formats across data sources to maintain uniformity.

1.6 Metadata Management: Maintain metadata for tracking data sources, transformations, and changes to the gold price dataset.

1.7 Data Lifecycle: Define retention and deletion policies to comply with financial regulations and business needs.

1.8 Compliance: Align with financial regulations (e.g., SEC, MiFID II) and data protection laws (e.g., GDPR) to ensure lawful use of gold price data.

2. Data Collection Standards

2.1 Reliable Sources: Collect gold price data from trusted financial sources (e.g., NYSE, LBMA, COMEX).

2.2 Frequency & Timeliness: Ensure data is collected at consistent intervals (e.g., hourly or daily) for accurate time series forecasting.

2.3 Data Accuracy: Validate the accuracy of gold prices and related metrics (e.g., volume, bid-ask spread) across sources before use.

2.4 Consistency in Units: Collect data in consistent units (e.g., price per ounce) to avoid conversion errors.

2.5 Time Zone Standardization: Align timestamps to a single time zone (e.g., UTC) to avoid discrepancies when integrating global market data. (e.g., London, New York, and Shanghai).

2.6 Complete Historical Data: Ensure historical gold price data is complete for long-term trend analysis and prediction.

3. Data Cleaning and Validation

3.1 Handling Missing Data: Identify missing gold price data or market volume due to holidays or trading halts.

3.2 Mathematical Statement (forward-filling): For a series of gold prices P_t where P_t represents the price at time t: If $P_t = P_{t-1}$ if P_t =NA

3.3 Mathematical Statement (Linear interpolation): For a series of gold prices P_t , where P_t is missing and P_{t-1} and P_{t+k} are known: P_t =P_{t-1} + ((P_{t+k} − P_{t-1}) / k).

Where:

- P_t is the interpolated value at time t.
- P_{t-1} is the price before the missing data.
- P_{t+k} is the next available price after the missing data.
- k is the number of periods between the known values P_{t-1} and P_{t+k}.

3.4 Dealing with Outliers:

3.4.1 Define a threshold for price changes. For example, any day where the price increases or decreases by more than 5-10

3.4.2 Use financial context (e.g., significant economic events) to decide whether to keep or smooth out the outliers.

3.5 Feature Engineering:

3.5.1 Date Features: Create features like day of the week, month, or season (as gold prices often show cyclical behavior).

3.5.2 Lag Features: Generate lagged gold prices (e.g., previous day, week, or month) as predictors for future prices.

3.5.3 Evaluate patterns and set thresholds, dummy data and obvious patterns,

3.6 Normalization/Standardization: Normalize prices and trading volumes to account for scale differences when using machine learning algorithms.

3.7 Removing Duplicates: Check for and remove duplicate records from data sources, ensuring unique gold price entries for each timestamp.

4. Version control

We have decided to implement version control through database versioning.

4.1 Audit Tables are used to store historical versions of financial data. Each change (insert, update, delete) should be logged with metadata such as timestamp, user ID, and reason for modification.

4.2 Separate tables are maintained for each version of the dataset, with a version number column or suffix (e.g., dataset_v1, dataset_v2).

4.3 Using temporal tables that allow querying the data as it existed at any point in time.

5. Data Auditing and Monitoring

5.1 Audit Logs and Trails: Tracking who accessed the data, what changes were made, and when. Track changes INSERT, UPDATE, DELETE) in a dedicated audit table. Include metadata such as user ID, timestamp, query executed, and client IP.

5.2 Data Change Checking: Database triggers can be used to track any changes in the data. Before and after snapshots can be maintained.

5.3 Data Profiling: Continuously monitor data quality using profiling tools to check for issues like missing data, duplicates, or unexpected patterns. This can help catch issues beforehand.

6. Data Security Measures

Sensitive data includes historical and real-time gold price data, financial market data, and geopolitical event data. Incorporate security protocols to protect sensitive data from unauthorized access.

6.1 Role-Based Permissions (RBAC): Define roles (e.g., data engineer, analyst) and assign access accordingly. Restrict write access to raw data and grant read-only access for processed data.

6.2 Encryption: 6.2.1 AES-256 for Data Storage: AES-256 encryption ensures the protection of stored financial and event data by using a 256-bit key, providing strong resistance against brute-force attacks. 6.2.2 TLS for Data Transmission: TLS is used to secure data transmission across systems (e.g., APIs fetching gold price data) by encrypting the communication, preventing interception or tampering during transit.

6.3 Multi-Factor Authentication (MFA): Require MFA for access to critical systems, ensuring secure entry into databases and dashboards. In detail justifying

6.4 Vulnerability Assessments: Regularly conduct security audits and penetration testing to identify vulnerabilities, especially with third-party integrations like financial or news APIs.

6.4.1 Weak Authentication: If APIs lack strong authentication mechanisms (e.g., using simple API keys instead of OAuth), they are vulnerable to unauthorized access, allowing attackers to exploit sensitive data.

6.4.2 Insecure Data Transmission: Without proper encryption (e.g., not using TLS), API communication can be intercepted in transit, exposing sensitive data like financial information or credentials to attackers.

6.4.3 Outdated Software or Libraries: Third-party APIs might use outdated software or vulnerable libraries, which can be exploited through known security flaws, making the system susceptible to attacks like injection or remote code execution.

7. Data Backup and Recovery

A backup plan is designed to avoid data loss and recovery strategies are implemented in case of incidents.

7.1 3-2-1 Backup Rule: Maintain three copies of data — one active, one backup on a separate system (e.g., cloud), and one offsite.

7.2 Backup Frequency: Hourly backups for real-time data like gold prices and world events, and daily or weekly backups for historical datasets.

7.3 Backup Testing: Conduct quarterly tests to ensure successful data restoration, focusing on recovering real-time and historical data without corruption.

7.4 Document Backup Details: Maintain clear documentation of backup schedules and locations (e.g., cloud storage or physical offsite) to ensure recoverability and compliance with backup policies.

8. Data Lineage Tracking

8.1 The source for some of the Kaggle data is known: [Yahoo Finance:(https://finance.yahoo.com) MacroTrends: (https://www.macrotrends.net/1333/historical-gold-prices-100-year-chart]

8.2 Tools like Apache Atlas or Collibra can be used to further track the data's origins, transformations, and dependencies.

8.3 A pipeline orchestration tool like Apache Airflow, Dagster, or Prefect will be used to build data workflows.

8.4 These tools have built-in logging, which helps track the flow of datasets from ingestion to transformation and model training.

8.5 Each step of the Kaggle dataset processing is defined in an Airflow DAG, and Airflow will automatically log timestamps, execution success, and dependencies between tasks.

9. Access Control and Authentication

Robust access controls to safeguard access to financial and world events data.

9.1 Role-Based Access: Define roles (e.g., data scientist, analyst, system administrator) and limit access based on these roles. For example, only data engineers should have write access to raw data, while analysts may only have read access.

9.2 Periodic Review of Access Permissions: Schedule quarterly reviews of team access privileges, especially when roles change, to ensure no one has excess privileges.

9.3 Access Logs: Maintain logs of data access and modifications, ensuring each access attempt is logged for auditing and identifying potential security breaches.

10. Staff Training and Awareness

Ensure all team members are trained specifically on handling financial data and world events data, as it relates to predicting gold prices.

10.1 Training Sessions: Organize mandatory sessions focused on best practices for managing and analyzing sensitive financial data. These sessions should cover data governance, data hygiene, and specific legal regulations relevant to financial data.

10.2 Learning Materials: Develop and distribute materials explaining the importance of data quality (e.g., accurate world event tagging) and how poor data hygiene could lead to incorrect gold price predictions.

10.3 Promote Accountability: Foster a culture where every team member understands their responsibility in maintaining data integrity, by emphasizing that even small errors in world events data could result in substantial inaccuracies in gold price forecasts.

11. Data Quality and Metrics and KPIs

The following metrics (KPIs) are used to assess data quality:

11.1 Completeness: all data points must be present, missing values are handled

11.2 Accuracy: validate the dataset against authoritative sources (like World Gold Council or official financial platforms). Additionally, extreme outliers should not be present

11.3 Uniqueness: ensuring that no duplicate records are found

11.4 Timeliness: data freshness. Data must be up to date

11.5 Validity & Consistency: there should be uniformity in the format (measurement units for currency and weight) of the 2 - 3 Kaggle datasets we are using. If not, it should be implemented.

11.6 Prices should not be negative.

11.7 Correlation with External Markets: Measure the correlation of gold prices with stock markets or other commodities. This can help in understanding external factors influencing price movements.

12. Third Party Data Validation

External services to validate the data being used for gold price predictions, ensuring accuracy and robustness.

12.1 Identify Trusted Validators: Identify reliable third-party services like Thomson Reuters or Bloomberg for gold price data, and geopolitical analysis providers for event data validation.

12.2 API Integration: Integrate APIs from these services into your pipeline to automatically validate and cross-check your internal data. For instance, compare real-time gold price feeds from your primary source with a trusted third-party provider.

12.3 Document Validation Efforts: Keep a detailed log of when and how third-party validation was used. For instance, document discrepancies found between internal and third-party world event data and note corrective actions taken.

13. Real Time Data Updates

The gold dataset can be updated in real time in the following procedure:

13.1 Data Ingestion: Use Kafka or Kinesis to stream data from real-time sources into the pipeline.

13.2 Data Quality: Validate the incoming data using Great Expectations in real-time.

13.3 Data Storage: Use BigQuery or Snowflake to store and update the gold dataset with streaming inserts.

13.4 Trigger Updates: Use Airflow or AWS Lambda to trigger real-time updates and transformations on new data arrival.

14. Documentation and Metadata Management

Detailed documentation and metadata tracking for both financial and world events data to ensure transparency and traceability.

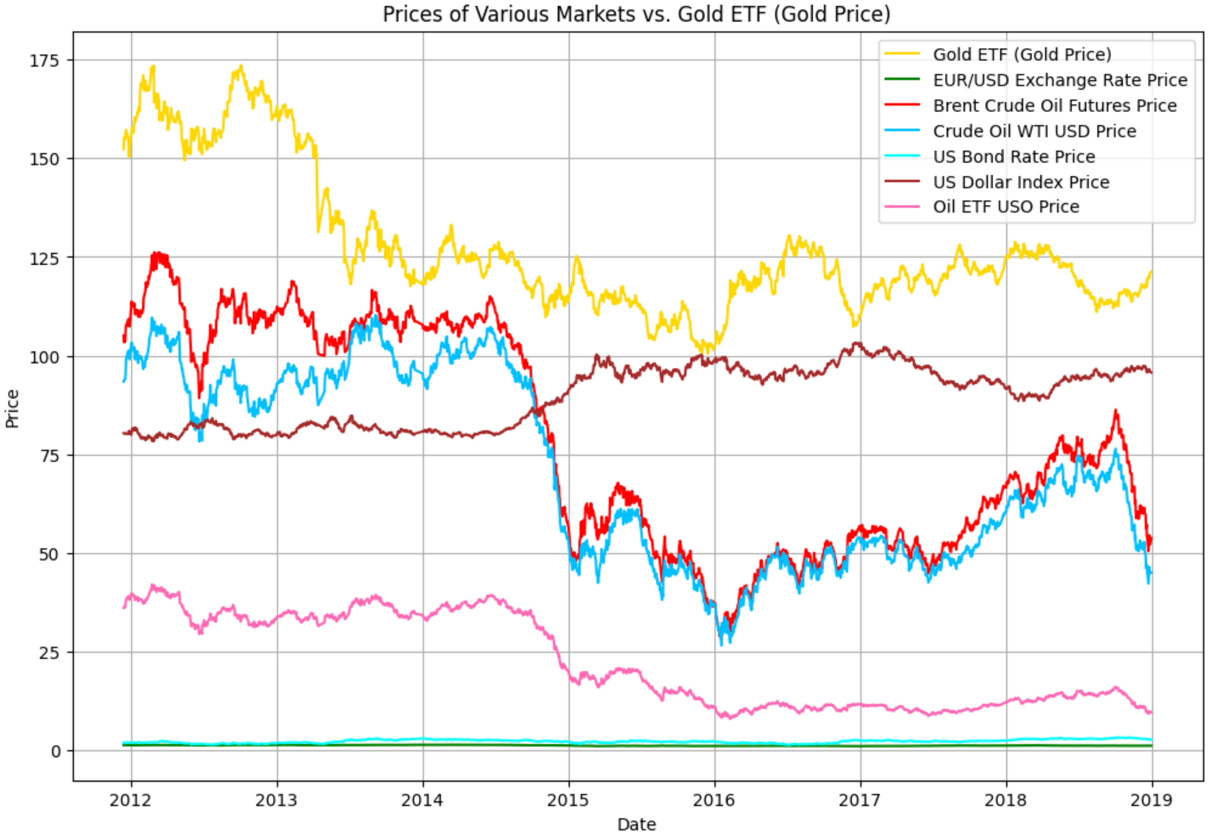
14.1 Metadata Management System: Develop a system to document key metadata such as data source (financial vs. world event), timestamp, collection method, and transformations applied. For instance, gold price data should be tagged with its source (e.g., stock exchange, API) and how it was processed.

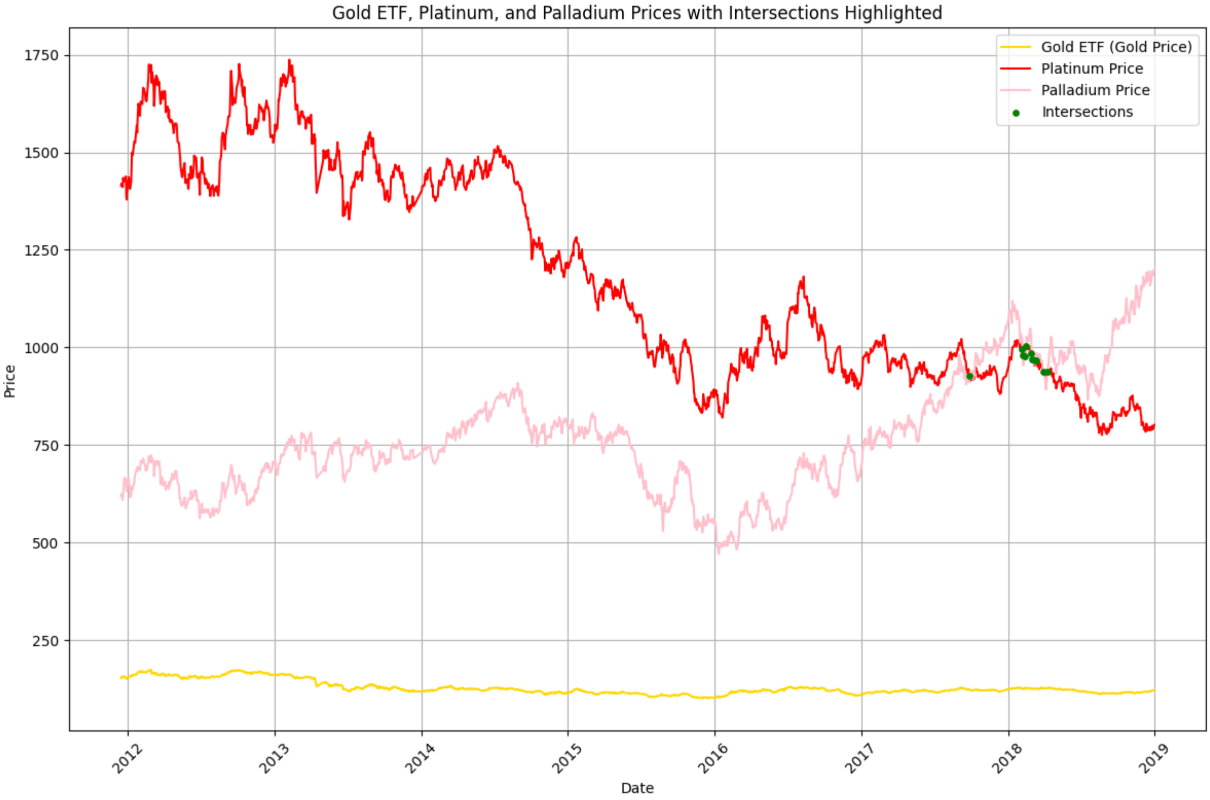
14.2 Regular Updates: Ensure that this documentation is updated regularly, especially when new data sources are integrated (e.g., a new world event feed) or when data collection methods change.

14.3 Accessible to Stakeholders: Ensure that all team members, especially analysts and data scientists, can easily access this documentation to understand the source and integrity of the data being used in their models.

Analysis of the Data

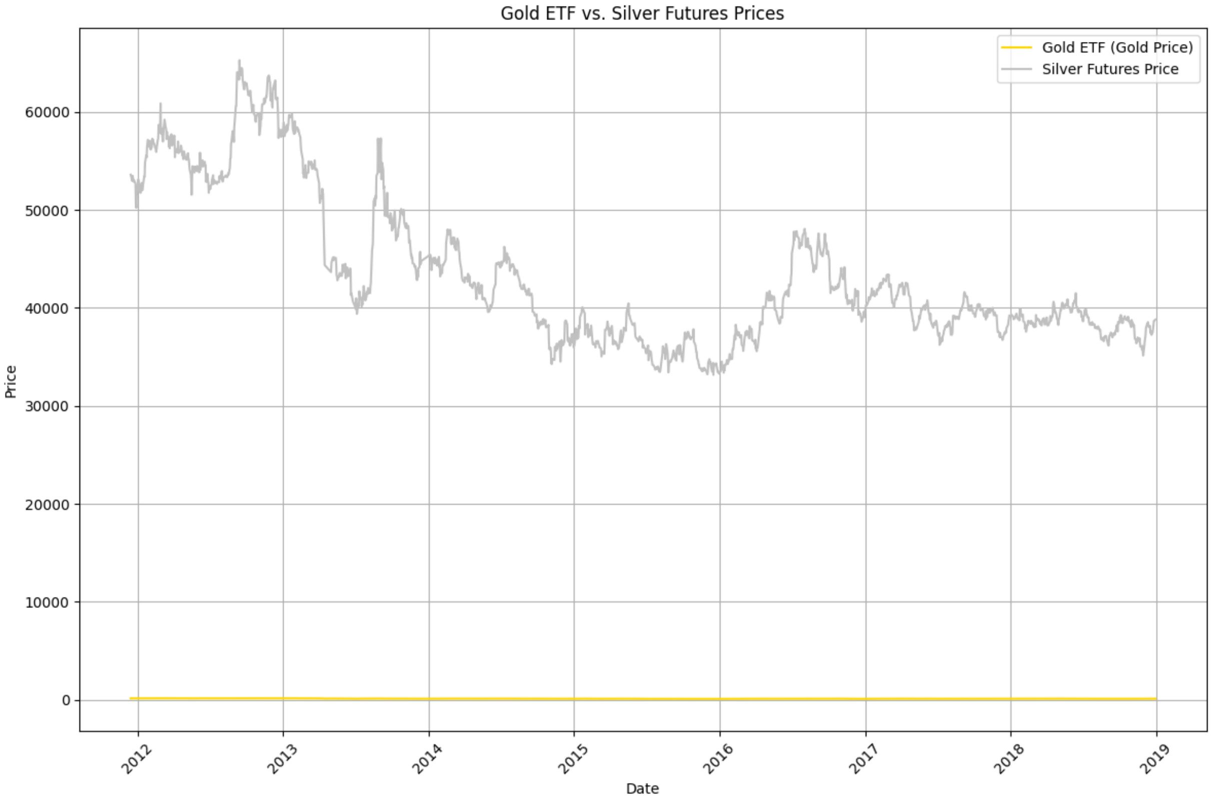
We have done some preliminary analysis of the available data by comparing and contrasting how the Gold Price (Adjusted Close) is varying with Adjusted Close Price in other markets. Link to Colab Notebook.

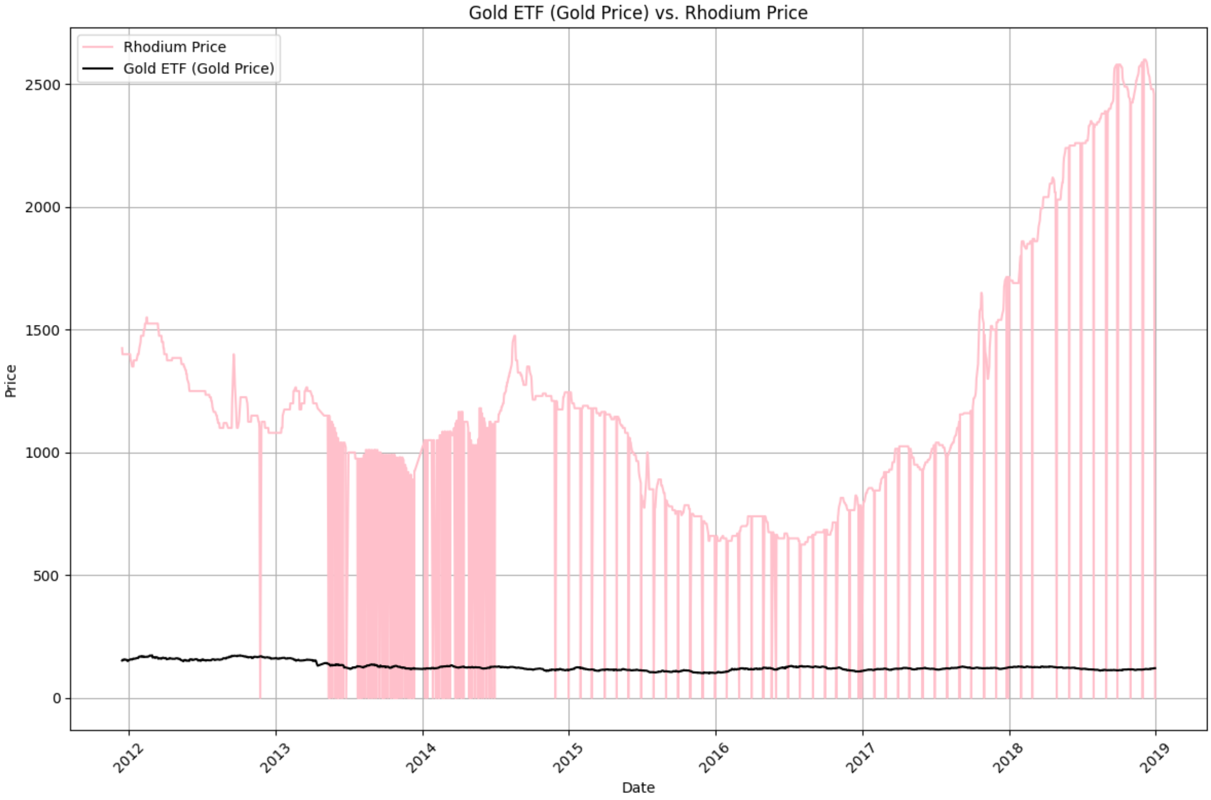




Palladium - Platinum Intersection Points (Date, Price): (2017-09-26, 928.45) (2018-02-05, 994.80) (2018-

02-08, 978.90) (2018-02-13, 978.10) (2018-02-15, 1004.20) (2018-02-28, 984.80) (2018-03-01, 970.00) (2018-03-09, 966.90) (2018-03-12, 966.10) (2018-03-29, 936.60) (2018-04-09, 937.00)





Observations:

1. The Kaggle data seems to be problematic as gold prices seem to be in the range of 100-200USD, which does not match with data from newspapers, which show the gold rate to be around 2600USD.

2. Furthermore, this dataset implies that the rate of silver is significantly higher than gold, which is not true.
3. Rhodium is often sold in over-the-counter (OTC) markets and not on centralized exchanges. As a result, price reporting can be delayed, irregular, or missing on certain days. Rhodium also has a much smaller market compared to other metals like gold. If there are days when no trading occurs due to illiquidity or market inactivity, some data sources might report the price as zero instead of leaving it blank or using the last recorded price. This is the reason the graph of rhodium has so many zeros.

Keeping this in mind, we plan to fully build the dataset from scratch by web scraping from reliable newspapers so that it depicts real market prices.

What the Attributes Represent

Date: The specific date on which the data was recorded, **Open**: The price of the asset at the start of the trading day, **High**: The highest price of the asset during the trading day, **Low**: The lowest price of the asset during the trading day, **Close**: The price of the asset at the end of the trading day, **Adj_Close**: The closing price adjusted for dividends, stock splits, or other factors, **Volume**: The total number of shares or contracts traded during the day.

Normalisation:

1NF:The table must contain atomic values, meaning each cell must contain a single value
All attributes in each table contain atomic values (e.g., Date, Open, High, Low, etc., are atomic). Therefore, all tables comply with 1NF
2NF:Every non-key attribute must be fully functionally dependent on the entire primary key. If a table has a composite primary key, no non-key attribute should depend on just part of it.
Date is the primary key across all tables, and all other attributes (Open, High, Low, Price, Trend, etc.) depend entirely on the Date. Hence, the tables comply with 2NF.
3NF: There must be no transitive dependency, meaning no non-key attribute should depend on another non-key attribute. All non-key attributes must depend solely on the primary key.
Attributes such as Open, High, Low, Volume, Trend are dependent only on Date, and not on other non-key attributes (e.g., Open does not determine High or Low). This eliminates transitive dependencies, ensuring 3NF.

Tables and Attributes

1. Gold_ETF Table
Attributes: Date (Primary Key),Open, High, Low, Close, Adj_Close, Volume
Purpose: Stores historical price and volume data for Gold ETF.

2. SP500 Table Attributes: Date (Primary Key, Foreign key),SP_Open, SP_High, SP_Low, SP_Close, SP_Adj_Close, SP_Volume
Purpose: Stores daily open, high, low, close, adjusted close, and volume data for the S&P 500 Index.

3.Dow_Jones Table
Attributes: Date (Primary Key, Foreign key),DJ_Open, DJ_High, DJ_Low, DJ_Close, DJ_Adj_Close, DJ_Volume
Purpose: Contains daily data for the Dow Jones Index.

4. Eldorado_Gold_Corp Table
Attributes: Date (Primary Key, Foreign key),EG_Open, EG_High, EG_Low, EG_Close, EG_Adj_Close, EG_Volume
Purpose: Stores financial data for Eldorado Gold Corporation (EGO).

5. EUR_USD_Exchange Table
Attributes: Date (Primary Key, Foreign key), EU_Price, EU_Open, EU_High, EU_Low, EU_Trend
Purpose: Tracks EUR/USD exchange rate fluctuations.

6. Brent_Crude_Oil_Futures Table
Attributes: Date (Primary Key, Foreign key),OF_Price, OF_Open, OF_High, OF_Low, OF_Volume, OF_Trend
Purpose: Contains data on Brent Crude oil futures prices.

7. Crude_Oil_WTI Table
Attributes: Date (Primary Key, Foreign key), OS_Price, OS_Open, OS_High, OS_Low, OS_Trend
Purpose:Stores WTI crude oil data.

8. Silver_Futures Table
Attributes: Date (Primary Key, Foreign key), SF_Price, SF_Open, SF_High, SF_Low, SF_Volume, SF_Trend
Purpose: Holds historical data for silver futures.

9. US_Bond_Rate Table
Attributes: Date (Primary Key, Foreign key), USB_Price, USB_Open, USB_High, USB_Low, USB_Trend
Purpose:Tracks US bond rate data.

10. Platinum_Price Table
Attributes: Date (Primary Key, Foreign key), PLT_Price, PLT_Open, PLT_High, PLT_Low, PLT_Trend
Purpose:Contains platinum price data.

11. Palladium_Price Table
Attributes: Date (Primary Key), PLD_Price, PLD_Open, PLD_High, PLD_Low, PLD_Trend
Purpose: Holds palladium price data.

12. Rhodium_Price Table
Attributes: Date (Primary Key), RHO_Price
Purpose: Stores rhodium price data.

13. US_Dollar_Index Table
Attributes: Date (Primary Key), USDI_Price, USDI_Open, USDI_High, USDI_Low, USDI_Volume, USDI_Trend
Purpose: Tracks the US Dollar Index.

14. Gold_Miners_ETF (GDX) Table
Attributes: Date (Primary Key), GDX_Open, GDX_High, GDX_Low, GDX_Close, GDX_Adj_Close, GDX_Volume
Purpose: Tracks performance of gold mining companies through ETFs.

15. Oil_ETF_USO Table
Attributes: Date (Primary Key), USO_Open, USO_High, USO_Low, USO_Close, USO_Adj_Close, USO_Volume
Purpose: Contains historical data for Oil ETF USO.

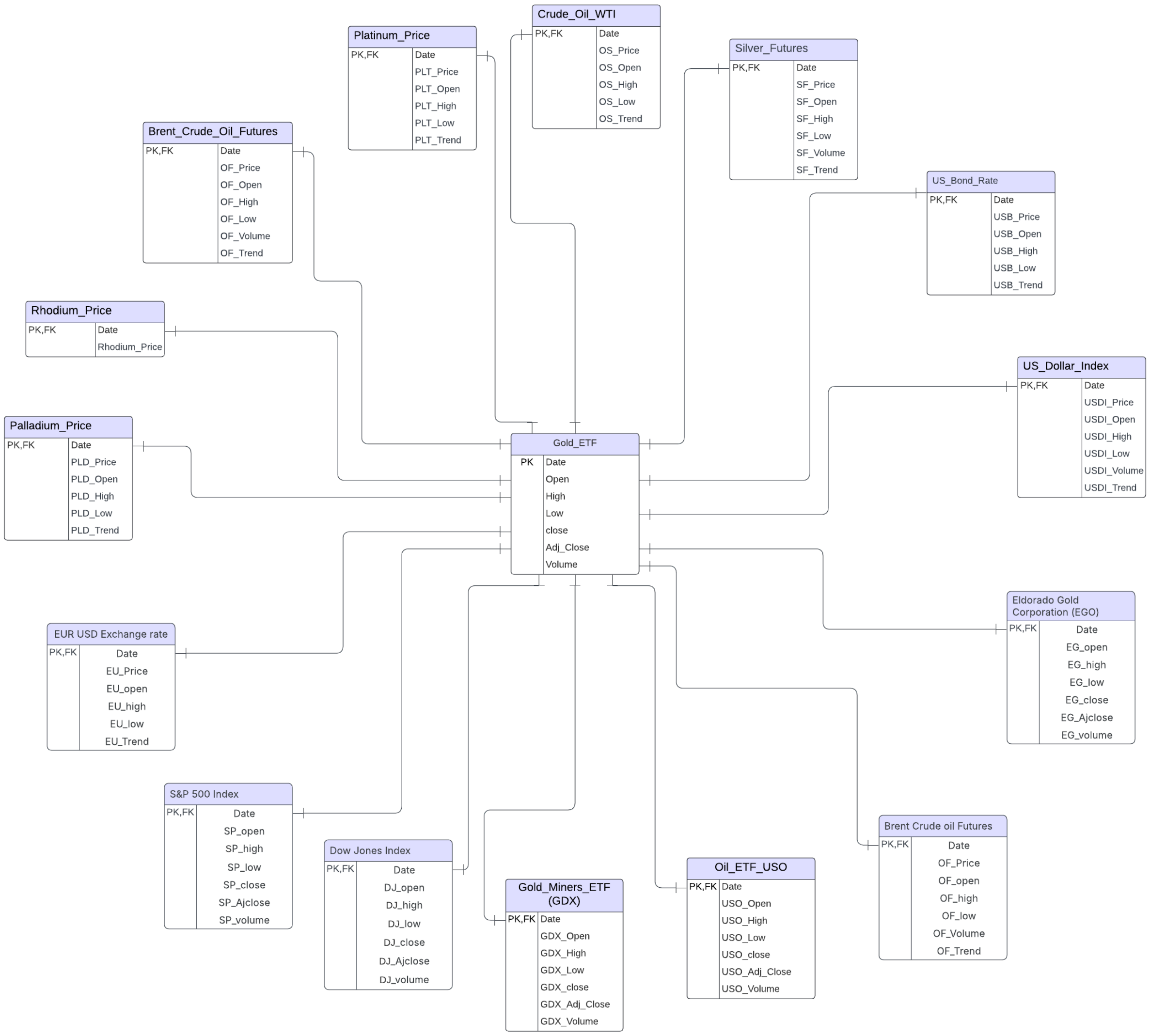


Figure 1: ER Diagram

normalisation

Survey Questions for Data Collection

Source 1: yahoo!finance

For each of the following markets: *Gold ETF*, *S&P 500 Index*, *Dow Jones Index*, *Eldorado Gold Corporation (EGO)*, *EUR/USD Exchange Rate*, *Brent Crude Oil Futures*, *Crude Oil WTI USD*, *Silver Futures*, *US Bond Rate*, *Platinum Price*, *Palladium Price*, *Rhodium Price*, *US Dollar Index*, *Gold Miners ETF (GDX)*, and *Oil ETF (USO)*, we are asking the following questions:

1. What is the Open price on the corresponding date?
2. What is the High price on the corresponding date?
3. What is the Low price on the corresponding date?
4. What is the Close price on the corresponding date?
5. What is the Adjusted Close price (if applicable) on the corresponding date?
6. What is the Volume traded (if applicable) on the corresponding date?

We plan to implement web scraping in the future to automatically expand and update the dataset daily.

Source 2: The Wall Street Journal

The Wall Street Journal (WSJ) can provide market summaries, news, and some day-to-day performance metrics for indices like the S&P 500 and Dow Jones, as well as commodities such as gold and oil. WSJ can be used to track the daily opening, high, and low values of these financial instruments. WSJ is primarily a financial news source that focuses on market analysis, breaking news, and some financial metrics, but it may not have the level of granularity and breadth of historical data across different markets.

Nevertheless, the following survey questions can be used to collect data:

1. What is the Open price on the corresponding date?
2. What is the High price on the corresponding date?
3. What is the Low price on the corresponding date?

Source 3: The Economic Times

The Economic Times (ET) provides in-depth market summaries, news, and daily performance metrics for key indices such as the S&P BSE Sensex, Nifty 50, and a wide array of commodities like gold, oil, and silver. It offers real-time updates on the daily opening, high, low, close, and volume of various financial instruments. As India's leading business daily, ET focuses on both local and global financial markets, offering extensive coverage of corporate news, market trends, and sectoral analysis.

While The Economic Times is a reliable source for current financial news and metrics, it may not offer the same breadth of detailed, historical financial data across different markets as specialized financial platforms like Bloomberg or Yahoo Finance. However, it excels in providing market insights and business analysis tailored for both Indian and global contexts.

The following survey questions can be used to collect data from the required markets:

1. What is the Open price on the corresponding date?
2. What is the High price on the corresponding date?
3. What is the Low price on the corresponding date?
4. What is the Close price on the corresponding date?
5. What is the Volume traded (if applicable) on the corresponding date?

Data Cleaning Considerations: The following data cleaning tasks will have to be handled to accurately clean and collect the data

- Missing Data is initialised as -9999 if present (so it is an obvious outlier).
- Dates should be aligned across all markets, accounting for trading holidays and other inconsistencies.
- Open, High, Low, Close, Adj Close are all in **USD**.
  Volume, which is the total number of shares traded during the day, is measured in **number of shares or units**.
- Remove any duplicate rows from the dataset.
- Ensure that Adjusted Close prices account for dividends, stock splits, etc.

Future Prospects

We plan to expand our dataset by integrating data from sources such as Yahoo Finance and Bloomberg, both of which are tailored to provide detailed financial metrics through web scraping techniques.

Additionally, we intend to incorporate data from significant global events, such as the COVID-19 pandemic, which will be stored separately or flagged with specific indicators, given that market behavior during such crises was influenced by unique factors. This approach will allow us to analyze the pandemic's impact on the gold market and better understand how external shocks affect financial instruments.