Market Data Pipeline

Data Acquisition, Cleaning, EDA, Data Preprocessing, and Feature Engineering

Team Members: Qingyue Wang, Gujie Li, Wenyang Zu, Lucas Meadows

Github links: https://github.com/Taylorwq/Applied-Data-Science-Project-1

1. Introduction and Dataset Description

Financial markets are shaped not only by individual asset movements but also by broader economic conditions. Factors such as interest rates, inflation levels, and labor market trends can change investors expectations and risk assessments. Therefore, relying only on market prices makes it difficult to present a complete picture and conduct thorough analysis. This project's dataset integrates financial market variables with macroeconomic indicators to capture the overall financial landscape more comprehensively.

Our dataset is from January 2017 to February 2026, including approximately eight years of diverse economic conditions. This period includes low interest rate environments, pandemic market volatility, rising inflation, and subsequent monetary tightening policies. Covering numerous events impacting the economy, studying this timeframe allows us to observe how market behavior evolves across different economic cycles.

To represent market activity, we selected three financial series: the S&P 500 index, the USD/JPY exchange rate, and gold prices. The S&P 500 provides a broad view of equity market performance and investor sentiment. USD/JPY reflects global capital flows and differences in monetary policy across major economies. Gold was included as a traditional safe asset that often shows differently from equities during periods of uncertainty. We also include these market variables with macroeconomic indicators obtained from the Federal Reserve Economic Data (FRED) database, including Treasury yields (2-year, 5-year, and 10-year), inflation, and unemployment. Treasury yields capture expectations about interest rates and economic growth, inflation measures price pressure in the economy, and unemployment reflects labor market conditions. Together, these variables allow us to connect market movements.

2. Data Acquisition Methodology

Data acquisition in this project mainly include steps like selecting variables, retrieving data using APIs and official downloads, saving raw files, and check the data before any preprocessing. We built the dataset to from real world dataset website, so it’s complex， the data comes from multiple sources, has different time frequencies, and includes many kinds of quality issues.

The dataset is relatively complex because it combines information from different sources, reporting frequencies, and data structures. Instead of relying on a single prepared dataset, we integrate market price obtained through an API with macroeconomic variable released by official institutions. Market variables are recorded daily, while macroeconomic indicators are monthly or released with gaps. In addition, Treasury yield series contain missing values related to non trading days, and macroeconomic data may revised after publication. Market prices display high volatility and short term dependence, whereas macroeconomic indicators evolve gradually and reflect economic conditions. Integrating variables with different dynamics increases complexity. These factors position the dataset at a high level of complexity consistent with real financial data environments rather than simplified academic datasets.

Variable selection was based on the goal of linking market behavior with economic conditions. The S&P 500 was used as a measure of equity market performance and situation. The USD/JPY exchange rate was included to capture global capital flows, but exchange rates can also be affected by political events. Gold was selected as a safe asset that often behaves differently from equities, but its relationship with macro variables may change across regimes. Macroeconomic indicators were chosen to represent key economic mechanisms. Treasury yields provide information about interest rate expectations, inflation measures price pressure, and unemployment reflects labor market conditions. These variables are important and it will introduce challenges for lower frequency and reporting lag.

Market price series were retrieved programmatically using the Yahoo Finance API through the Python yfinance library. A loop pipeline downloads historical price data for each ticker within date range. Each dataset was saved as a raw CSV prior to any transformation. Macroeconomic indicators were obtained from the FRED database as CSV files. This approach involves a manual step, we need to obtain the data one by one. The files were organized into a folder to separate macroeconomic data from market data. The selected FRED datasets include Treasury yields, inflation, and unemployment. All datasets stored as csv and file name are standardized.

Following acquisition, we verify the stucture and identify integration challenges. The inspection examined observation counts, date coverage, column names, and data types. Market datasets contain roughly 2,300 daily observations, with slight differences across assets due to trading calendars and data availability. The macroeconomic indicators are reported less frequently and may contain gaps related to release schedules. There are also some missing values in Treasury yield series. Inspection also revealed differences in date formatting and variable naming across sources, necessitating standardization before merging.

Overall, this acquisition process creates a dataset that combines multiple data sources and reflects different aspects of financial conditions. This provides basis for the data cleaning, exploratory analysis, and feature engineering steps that follow.

Dataset	Source	Frequency	Time Range	Advantages	Limitations
S&P 500	Yahoo Finance API	Daily	Jan 2017 – Feb 2026	Broad market proxy	Aggregated signal
USD/JPY	Yahoo Finance API	Daily	Jan 2017 – Feb 2026	Global flow indicator	Noise
Gold (GC=F)	Yahoo Finance API	Daily	Jan 2017 – Feb 2026	Safe signal	Regime instability

Dataset	Source	Frequency	Time Range	Advantages	Limitations
Treasury Yields (2Y,5Y,10Y)	FRED	Daily (gaps)	Jan 2017 – Feb 2026	Rate expectations	Missing days, revisions
Inflation	FRED	Monthly	Jan 2017 – Feb 2026	Price pressure signal	Release lag
Unemployment	FRED	Monthly	Jan 2017 – Feb 2026	Economic cycle indicator	Low frequency

3. Data Cleaning, Integration, and Handling Inconsistencies

The datasets used in this project come from multiple sources, including daily financial market data (USD/JPY exchange rate, gold price, S&P 500 index), US Treasury yields (2-year, 5-year, and 10-year), and monthly macroeconomic indicators (unemployment rate and CPI). Because this data comes from different providers and is released at varying frequencies, several data cleaning steps were required before analysis. Column names were standardized to remove special characters and unify formatting; all date variables were converted to date-time formats for time series alignment. The datasets were sorted chronologically before merging.

A key challenge was reconciling the frequency differences, as financial variables are recorded daily, while macroeconomic indicators are reported monthly. To address this, we used forward imputation interpolation to resample the monthly macroeconomic data to daily frequencies, ensuring that each trading day reflects the latest available macroeconomic information. Missing values in the Treasury yield data (typically occurring on non-trading days) were also forward imputed to maintain data continuity.

Duplicate observations were checked and no duplicate are found. Potential outliers were identified through descriptive statistical analysis. All observed extreme values were consistent with known economic events, such as pandemic related surges in unemployment and post pandemic interest rate increases. Since these fluctuations reflected real market conditions rather than data errors, no observations were removed. Following these steps, the final dataset remained consistent in time sequence, contained no duplicate entries, and had a consistent structure for all variables, providing a reliable foundation for subsequent analysis.

After cleaning, we combined all datasets into a single data using the date as reference. We performed several checks after merging. We also confirmed that no duplicate records were introduced and to make sure the values were reasonable and consistent with major economic events during the sample period.

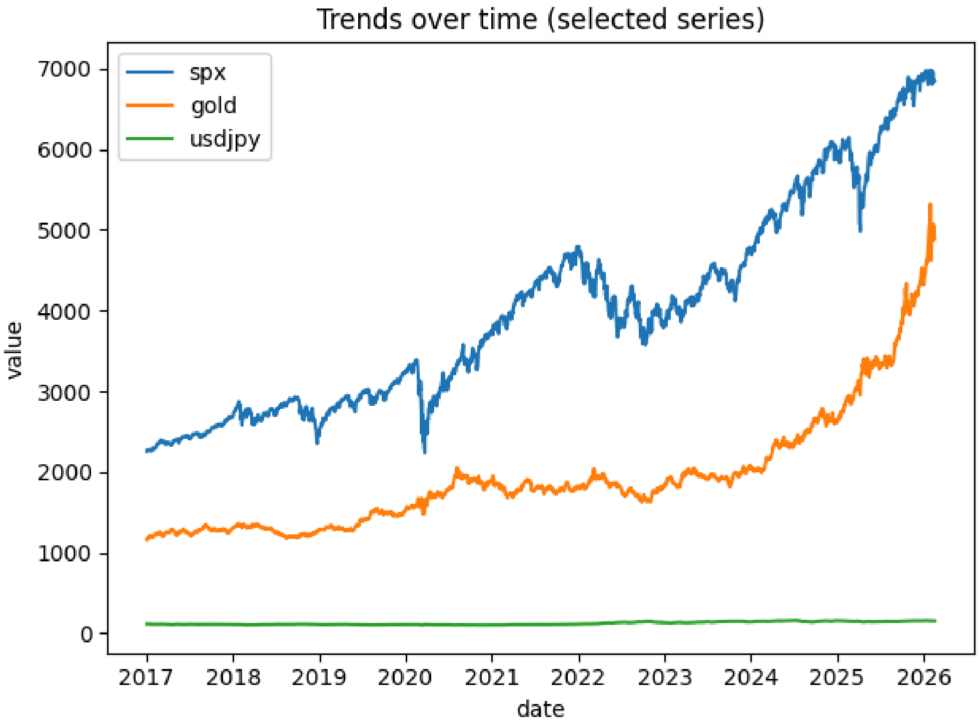
4.Exploratory Data Analysis (EDA)

After cleaning and integrating the dataset, we conducted exploratory data analysis to understand the main characteristics of the data. The purpose of this stage is to examine trends, distributions, relationships across variables, and potential regime changes that may influence modeling decisions.

Because the dataset combines financial market variables with macroeconomic indicators, exploratory analysis is important for identifying non stationarity, multicollinearity, and time vary risk. The analysis focuses on time series trends, distributional behavior, correlation structure, and several derived indicators such as returns, yield curve spread, and rolling volatility.

Time Series Trends for SPX, GOLD, USDJPY

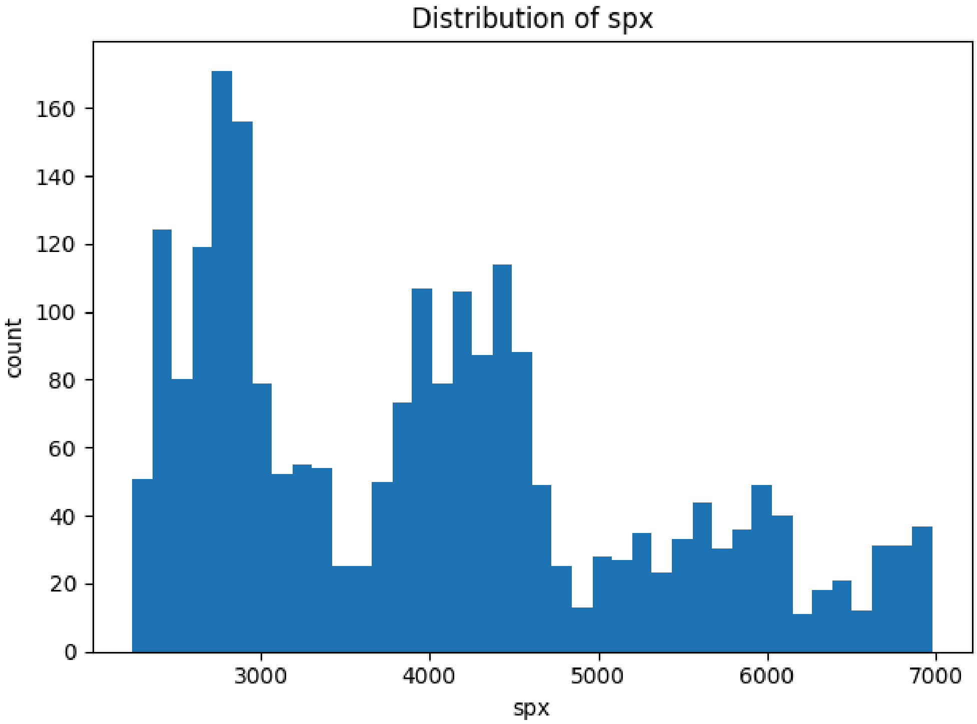
We first visualize the long-term behavior of the S&P 500, gold prices, and the USD/JPY exchange rate.



From 2017 to 2026, SPX and gold show a clear upward long-run trend. With noticeable drawdowns and volatility clustering, there is a sharp drop around early 2020. Another visible drop is around 2022. These synchronized regime changes suggest that macro shocks and risk sentiment may simultaneously affect multiple asset classes. In contrast, USDJPY is comparatively range bound relative to SPX/gold in level scale, but it still exhibits a notable regime shift upward after 2022. It is consistent with a higher rate environment and changing FX dynamics. Overall, the time series plots indicate that the variables are non-stationarity in levels trends and structural breaks, implying that later modeling may benefit from transformations such as returns/differences rather than raw levels.

Distribution Analysis of SPX

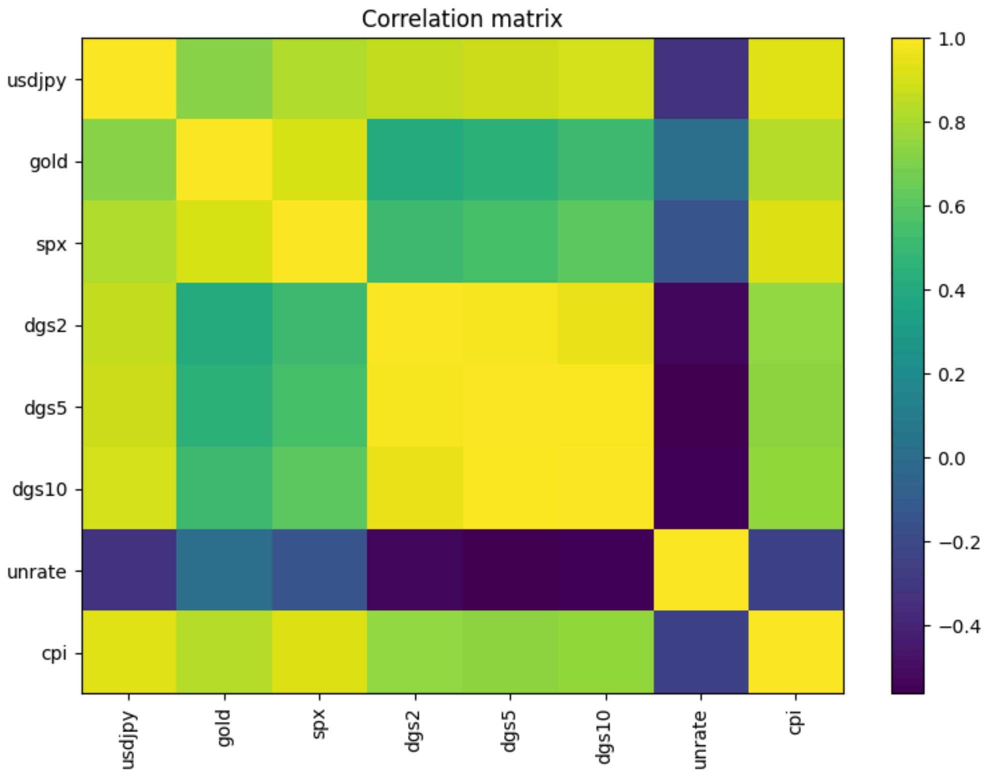
We examine the distribution of SPX levels using a histogram.



The SPX level distribution is not symmetric and appears multimodal. This reflects that the index spent long periods in different price regimes. This is expected for a trending financial time series. The histogram mixes multiple market regimes rather than representing a single stable distribution. The wide spread and heavier right tail also indicate that levels are non stationary, so modeling in levels may overemphasize time trend rather than underlying relationships. A practical implication is that return based features may better capture a stable distributional behavior for predictive tasks.

Covariance

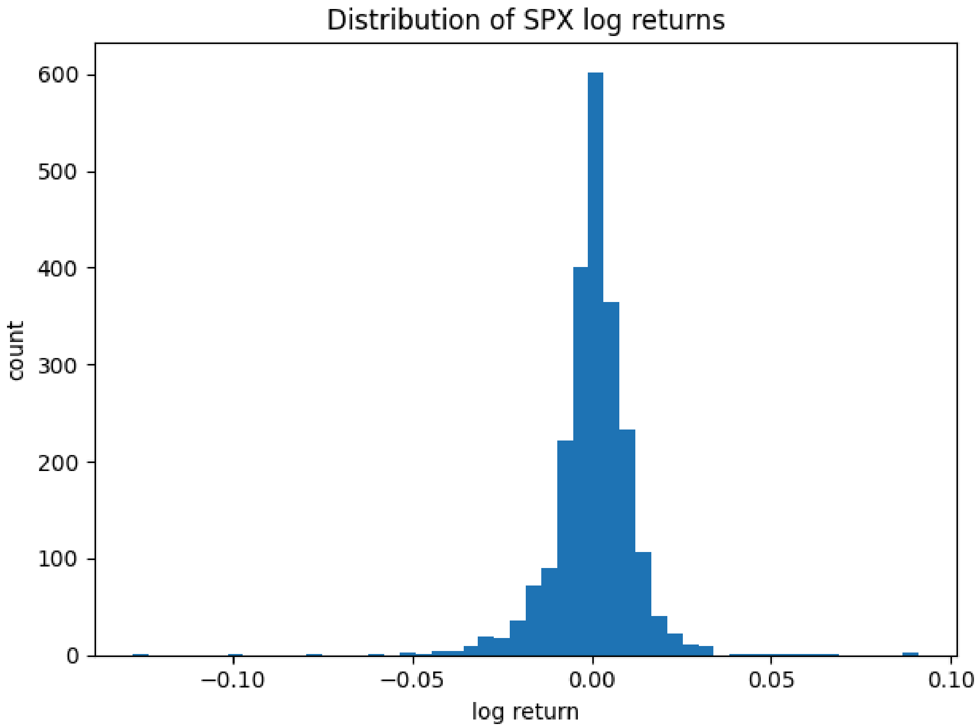
We next analyze relationships across variables using a correlation matrix.



The correlation matrix shows very strong positive correlations among interest rate variables DGS2/DGS5/DGS10, which indicates that the collinearity across the yield curve is substantial. The unemployment rate UNRATE is negatively correlated with yields and with market variables, which is consistent with counter-cyclically moving trend of unemployment. CPI shows positive correlations with several level variables, which is plausible since CPI is an index that trends upward over time and can move toghther with other trending level series. Overall, the matrix suggests that shared time trends can inflate correlations in levels, so subsequent modeling should consider detrending.

Return Dynamics

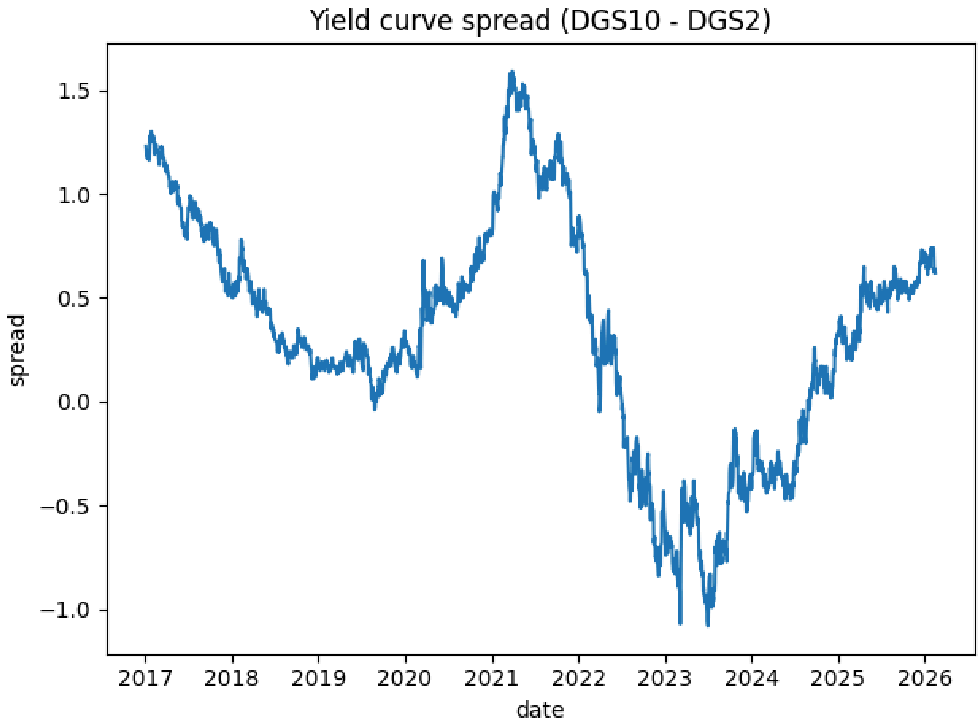
Because financial price levels are strongly trending, we compute log returns for SPX to examine a more stable representation of daily movements.



The histogram of SPX log returns is highly concentrated around zero, indicating that most daily movements are small. At the same time, the distribution shows fat tails, rare but large positive/negative returns, which is typical for financial returns and suggests the presence of extreme market events. We plot logrized returns rather than price levels because the SPX level is strongly trending and non-stationary. Log returns produces a more stable series and makes patterns and anomalies more comparable over time. This observation fuels later preprocessing and feature choices such as using return based targets and considering volatility related features.

Yield Curve Spread

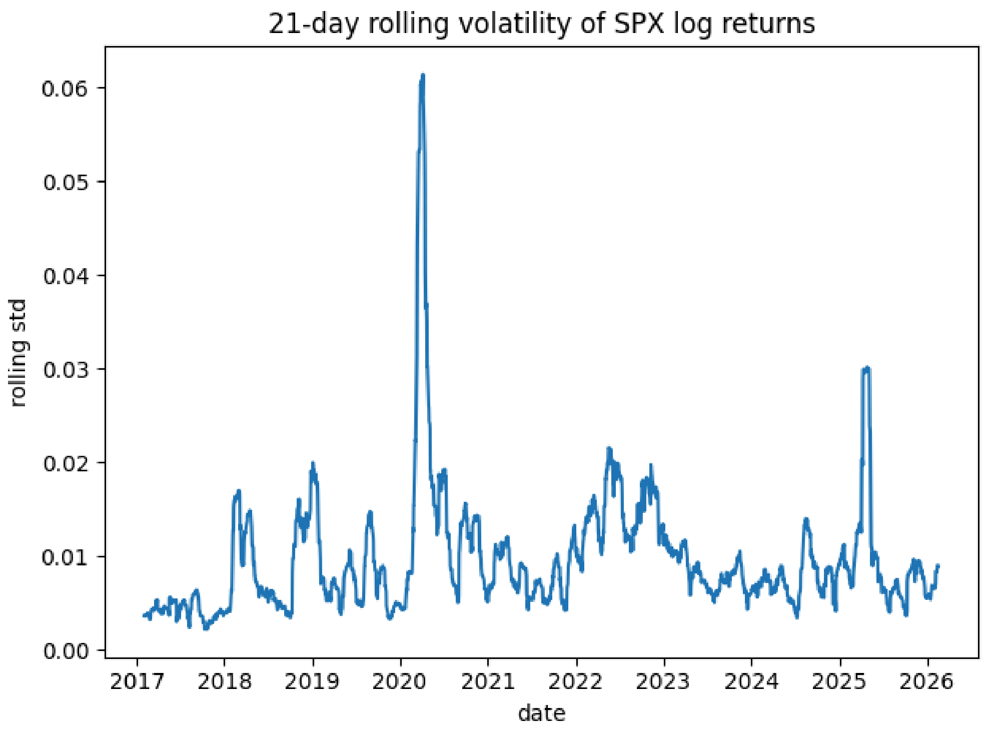
To summarize information across the yield curve, we construct the yield spread defined as DGS10 minus DGS2.



We construct the yield curve spread as DGS10 – DGS2 to summarize the slope of the yield curve in a single macro financial indicator. The series transitions from positive values normal upward sloping curve to negative territory, as know as yield curve inversion for a sustained period. This indicates a regime shift in market expectations about future growth and monetary policy. This visualization is useful because the individual yield variables DGS2, DGS5, DGS10 are highly correlated, so using a spread can reduce redundancy and multicollinearity while maintains economic interpretability. In later stages, this spread can serve as a compact feature that captures changes in the macro regime that may relate to movements in risk assets such as equities and FX.

Volatility Regimes

We compute a 21-day rolling standard deviation of SPX log returns as a proxy for short-term market volatility.



This plot shows the 21 day rolling standard deviation of SPX log returns, which serves as a simple proxy for short term market volatility. The series exhibits clear volatility clustering: long periods of relatively low volatility are interrupted by sharp spikes, which indicates that regime shifts in market uncertainty. The largest spike occurs around early 2020, consistent with a major market shock period, and there is another noticeable spike later in the sample, a episodic riskoff events. We include this visualization because it highlights that the return distribution is not constant over time. Instead, market risk varies by regime. In later stages, rolling volatility can be used as an engineered feature to capture market state and to help models account for time varying risk.

Overall, the EDA shows that the data is strongly influenced by market regimes, time trends, and changing volatility rather than stable patterns. Price levels trend upward, returns display volatility clustering, and yield variables are highly correlated. Instead, transformed features such as returns, yield spreads, and rolling volatility are more suitable for capturing meaningful relationships. These findings guide the next stage of preprocessing and feature engineering, where the goal is to build features that better reflect real market dynamics.

5. Feature Engineering

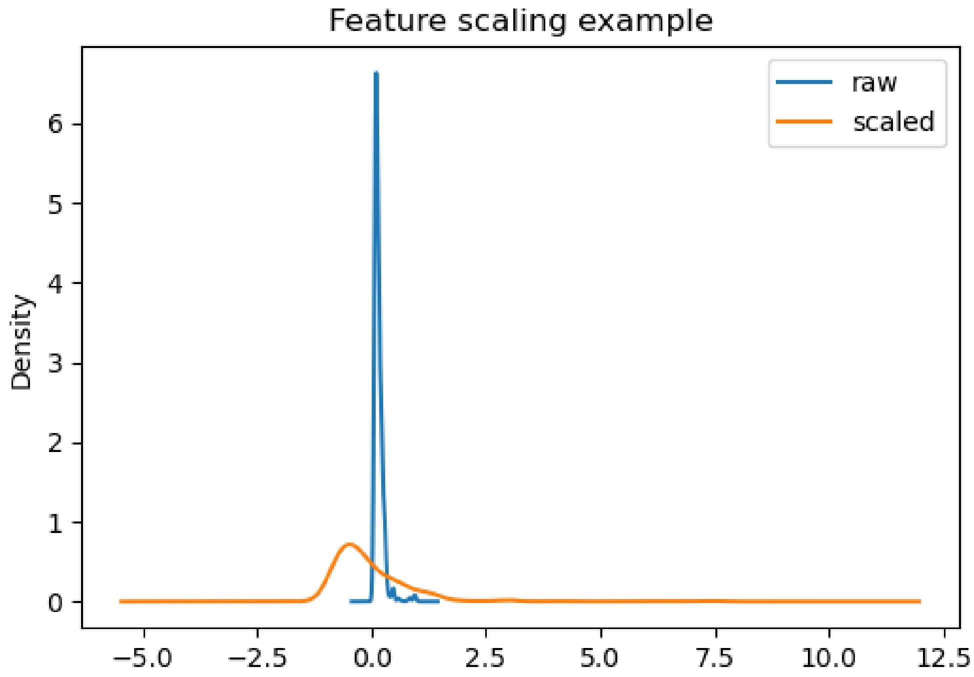
Given that this dataset contains both daily financial variables and monthly macroeconomic indicators, we first ensured temporal consistency before constructing additional features. All variables were sorted chronologically, and because CPI and unemployment data are released monthly and contain missing values in later sample periods, we implemented an as-of merge strategy. This ensures that each trading day is matched to the most recently available macroeconomic data, allowing the dataset to reflect only information that would have been accessible at that point in time. Establishing this time alignment was essential before building more advanced financial and structural features.

To better capture multiple dimensions of financial markets, we computed logarithmic returns for stock, gold, and exchange rate series, transforming non-stationary price levels into return-based measures more suitable for modeling. A 21-day rolling volatility metric was introduced to characterize short-term risk environments and shifts in uncertainty. Yield spreads between different-maturity government bonds (e.g., 10-year vs. 2-year and 10-year vs. 5-year) were constructed to summarize the term structure of interest rates, along with an inverted yield curve indicator to rapidly identify potential recession risk states. We also created percentage change variables for inflation and unemployment, reflecting the idea that financial markets respond to new information rather than static levels. Event indicators were defined around macroeconomic releases and extended into five-day rolling windows to capture short-term market adjustment dynamics. Momentum indicators and moving average ratios were further constructed to reflect cumulative price changes and deviations from recent trends, while a volatility state indicator was introduced using the upper quantile of rolling volatility to distinguish high-risk environments from normal conditions.

Building on these domain-specific financial features, we further enhanced the dataset using broader structural feature engineering techniques. To improve the analytical depth of our dataset and its readiness for modeling, we applied additional targeted transformations. Numerical variables with high skewness were log-transformed to reduce the influence of extreme values and stabilize their variance. Furthermore, we generated interaction terms and second-degree polynomial features to capture potential nonlinear and multiplicative relationships observed during our exploratory analysis. These steps allow the dataset to represent more complex structural patterns beyond simple linear effects.

To incorporate contextual information, we constructed group-level features, including category-based means and standard deviations for relevant numerical variables. We also created relative features that measure the deviation of each observation from its group average. These provide insight into how individual observations perform relative to their peers and enhance the interpretability and explanatory power of the dataset. Finally, categorical variables were encoded to ensure compatibility with modeling algorithms, and numerical features were standardized using z-score normalization to ensure comparability across scales and prevent high-magnitude variables from dominating results. Overall, this combined feature engineering process captures short-term risk conditions, trend persistence, macroeconomic shifts, nonlinear structural relationships, and contextual positioning within categories. Rather than relying on raw variables, the constructed features transform the dataset into a richer representation capable of reflecting dynamic market behavior and supporting robust analytical and predictive exploration.

We choose one of the feature to check the standardization, the feature values are centered around zero and shows similar dispersion, so this enhance the numerical stability.



Summary of key findings

Overall, analysis indicates that financial markets are frequently influenced by diverse factors rather than long-term stable relationships. Asset prices have risen significantly, yet yields exhibit volatility clustering and fat-tail characteristics, indicating that risk gradually evolves over time. Macroeconomic data analysis reveals that interest rate expectations, inflation, and labor market conditions also significantly influence market behavior. The high correlation among yield variables suggests that raw data shares many similar characteristics. Therefore, feature engineering is necessary to observe economically meaningful changes using features like yield spreads, rather than analyzing the raw data directly. Exploratory analysis reveals that transformed models incorporating yield, rolling volatility, and momentum indicators provide a more stable reflection of market dynamics than raw level data. Thus, datasets constructed through EDA better explain market behavior than simple data collection. Processed datasets facilitate superior analysis of market conditions, short-term risk, and macroeconomic information relationships, laying the foundation for subsequent forecasting modeling and financial analysis tasks.

Challenges faced and future recommendations

Due to variations of monthly and daily data, there are many challenges throughout the data processing pipeline. The primary difficulty lies in reconciling differing reporting frequencies, because daily market variables must combine with monthly macroeconomic while ensuring the authenticity and usability. Also handling missing values in government bond yields and gaps in indicator requires careful interpolation strategies to maintain data continuity. Another challenge is from non-stationarity and state transitions, which complicate data interpretation as correlations and distributions evolve over time rather than remaining stable. Furthermore, integrating multiple data sources introduces practical issues such as inconsistent naming conventions, varying date formats.

Future research can expand the dataset across multiple dimensions to enhance analytical depth and modeling potential. Incorporating additional asset classes—such as commodities, sector indices, or broader macroeconomic indicators—will provide a more comprehensive view of market conditions and cross-asset interactions. Utilizing higher-frequency or real-time data can better capture information accuracy while reducing requirements for feature insertion and processing. Feature engineering workflows can also be extended to include other risk metrics, sentiment indicators, or nonlinear transformations capable of capturing complex market behaviors. We can also perform more predictive and classification tasks using linear regression or other machine learning methods, enabling more accurate and efficient data utilization while improving overall data efficiency.

Team Contribution:

Qingyue - Data Acquisition, Data Preprocessing and Feature Engineerin. Lucas - Data Preprocessing and Feature Engineering. Wenyang - EDA. Gujie - Data Cleaning.

Each member writes the report section corresponding to their part