Volatility Regime Prediction Dataset

Data Collection and Quality Assessment Report

Philipp D. Dubach

December 14, 2025

Abstract

This report documents the data collection methodology and quality assessment for a volatility regime prediction research project. We collect and integrate data from multiple sources including CBOE (VIX indices, SKEW index, put/call ratios, and futures term structure), Yahoo Finance (S&P 500 prices), and FRED (macroeconomic and volatility indicators). The final dataset spans from January 2006 to December 2025, covering 5,046 trading days with 139 features. We provide descriptive statistics, correlation analysis, and an assessment of data quality suitable for machine learning applications in volatility forecasting.

Contents

1 Introduction

Volatility regime prediction is a fundamental challenge in quantitative finance with applications in portfolio management, risk management, and derivatives pricing. The VIX index, often called the "fear gauge," captures market expectations of near-term volatility derived from S&P 500 index options. Understanding and predicting transitions between volatility regimes can provide valuable insights for systematic trading strategies.

This report documents our data collection infrastructure, which is designed with the following objectives:

1. **Comprehensiveness**: Integrate multiple data sources covering spot volatility, forward-looking volatility (VIX term structure), realized volatility, and macroeconomic conditions.
2. **Extensibility**: Implement a modular architecture that allows easy replacement of data sources (e.g., transitioning from free to premium data providers).
3. **Reproducibility**: Create a fully automated pipeline that can be re-run to update the dataset.

2 Data Sources and Methodology

2.1 CBOE Volatility Indices

The Chicago Board Options Exchange (CBOE) provides historical data for various volatility indices. We collect:

Table 1: CBOE Volatility Index Series

Series	**Description**	**Start Date**
VIX	30-day implied volatility	2006-01-03
VVIX	Volatility of VIX	2006-03-06
VIX9D	9-day implied volatility	2011-01-04
VIX3M	3-month implied volatility	2009-09-18
VIX6M	6-month implied volatility	2008-01-02
SKEW	Tail risk index	2006-01-03

Each index provides OHLC (Open, High, Low, Close) values, enabling analysis of intraday volatility dynamics.

2.2 CBOE SKEW Index

The CBOE SKEW Index measures perceived tail risk in the S&P 500 distribution. It is derived from out-of-the-money option prices and reflects the market's expectation of extreme negative returns:

- **SKEW = 100**: Normal distribution (no perceived tail risk)
- **SKEW > 100**: Elevated left-tail risk (crash protection demand)
- **Historical range**: Typically 100–150, with spikes during stress periods

Unlike VIX, which measures the overall level of implied volatility, SKEW captures the *asymmetry* in option pricing—specifically, the premium investors pay for downside protection.

2.3 Put/Call Ratios

We collect historical put/call ratio data from CBOE's options volume database (available through October 2019):

Table 2: CBOE Put/Call Ratio Series

Series	**Description**	**Coverage**
TOTAL_PC	All CBOE options put/call ratio	2006-11 to 2019-10
INDEX_PC	Index options put/call ratio	2006-11 to 2019-10
EQUITY_PC	Equity options put/call ratio	2006-11 to 2019-10
VIX_PC	VIX options put/call ratio	2006-02 to 2019-10

Each series includes:

- Call volume, put volume, and total volume
- Put/call ratio: PC = Put Volume/Call Volume

Interpretation:

- **PC > 1**: More puts traded than calls (bearish sentiment or hedging demand)
- **PC < 1**: More calls traded than puts (bullish sentiment)
- **Extreme readings**: Often used as contrarian indicators

2.4 VIX Futures Term Structure

We construct the VIX futures term structure by downloading individual contract files from CBOE. For each trading day, we identify the front-month through ninth-month contracts and calculate:

- VX1 to VX9: Settlement prices for each contract month
- Term structure slope: VX2 – VX1 and VX4 – VX1

The term structure slope is a key indicator of market sentiment:

- **Contango** (slope > 0): Normal market conditions; near-term volatility expected to be lower than future volatility
- **Backwardation** (slope < 0): Stressed conditions; elevated near-term volatility expectations

2.5 S&P 500 Index Data

We obtain S&P 500 (ticker: ˆGSPC) price data from Yahoo Finance via the `yfinance` library. This provides daily OHLCV data for computing realized volatility measures.

2.5.1 Realized Volatility

We compute multiple realized volatility estimators:

Standard Realized Volatility:

$$RV_t^{(n)} = \sqrt{\frac{252}{n} \sum_{i=0}^{n-1} r_{t-i}^2} \tag{1}$$

where $r_t = \log(P_t/P_{t-1})$ is the log return.

Parkinson Volatility:

$$\sigma_{P,t}^{(n)} = \sqrt{\frac{1}{4n \log 2} \sum_{i=0}^{n-1} \left(\log \frac{H_{t-i}}{L_{t-i}} \right)^2} \tag{2}$$

where H_t and L_t are daily high and low prices. The Parkinson estimator is more efficient than the close-to-close estimator when intraday data is available.

2.6 Macroeconomic Data (FRED)

We collect macroeconomic indicators from the Federal Reserve Economic Data (FRED) API:

Table 3: FRED Economic Series

Series	**Description**	**Frequency**
Interest Rates		
DFF	Federal Funds Effective Rate	Daily
DGS1, DGS2, DGS10, DGS30	Treasury Yields	Daily
T10Y2Y	10Y-2Y Treasury Spread	Daily
T10Y3M	10Y-3M Treasury Spread	Daily
TEDRATE	TED Spread (3M LIBOR - T-Bill)	Daily
Credit Spreads		
BAMLH0A0HYM2	High Yield Corporate Spread	Daily
BAMLC0A0CM	Investment Grade Corporate Spread	Daily
Financial Stress Indices		
NFCI	Chicago Fed Financial Conditions Index	Weekly
STLFSI4	St. Louis Fed Financial Stress Index	Weekly
Economic Uncertainty		
USEPUINDXD	Economic Policy Uncertainty Index	Daily

The TED spread (TEDRATE) captures credit risk in the banking system—the difference between 3-month LIBOR and the risk-free T-Bill rate. Spikes in the TED spread historically precede market stress (e.g., 2008 financial crisis).

The Economic Policy Uncertainty Index (USEPUINDXD) is a text-based index measuring policy-related economic uncertainty from newspaper coverage, tax code provisions, and economic forecaster disagreement.

3 Dataset Overview

3.1 Summary Statistics

Table 4: Dataset Summary Statistics

Metric	**Value**
Date Range	2006-01-03 to 2025-12-12
Trading Days	5,046
Total Features	139
VIX Statistics	
Mean	19.46
Standard Deviation	8.73
Median	17.10
Minimum	9.14
Maximum	82.69
Skewness	2.50
Kurtosis	9.36
SKEW Statistics	
Mean	121.8
Typical Range	110–135
Put/Call Ratios	
Total PC Mean	0.92
VIX PC Mean	0.65
S&P 500 Returns	
Annualized Mean Return	10.30%
Annualized Volatility	19.45%
Sharpe Ratio	0.53
Worst Daily Return	-11.98%
Best Daily Return	11.58%
Term Structure	
Contango Frequency (VX1 > VIX)	76.5%
Upward Slope (VX2 > VX1)	81.5%
Mean Slope (VX2-VX1)	0.88

3.2 VIX Distribution Characteristics

The VIX exhibits several well-documented statistical properties:

1. **Right-skewness** (skew = 2.50): Volatility spikes are more common than volatility crashes
2. **Excess kurtosis** (kurtosis = 9.36): Fat tails indicate frequent extreme values
3. **Mean reversion**: VIX tends to revert to its long-term mean around 19-20
4. **Regime-dependent behavior**: Distinct low, normal, and high volatility states

These characteristics motivate regime-switching models rather than simple linear forecasting approaches.

4 Feature Engineering

We compute the following derived features for use in predictive modeling:

4.1 Realized Volatility Features

- Rolling realized volatility: 5, 10, 21, 63, 126, 252-day windows
- Parkinson volatility: 5, 10, 21, 63, 126, 252-day windows
- Log returns and log realized volatility

4.2 Variance Risk Premium (VRP)

The Variance Risk Premium captures the difference between implied and realized volatility:

$$\text{VRP} = \sigma_{\text{IV}}^2 - \sigma_{\text{RV}}^2 \tag{3}$$

We compute both forward-looking VRP (comparing to future realized volatility, ex-post) and backward-looking VRP (comparing to historical realized, available in real-time).

Important: Forward VRP (`vrp_forward`) uses future realized volatility and **cannot** be used as a predictor in ML models. It is computed for ex-post analysis only. Use `vrp_backward` for real-time available features.

Key VRP statistics in our dataset:

- VRP backward is positive 84.1% of the time
- VRP forward is positive 82.0% of the time
- Mean VRP: ∼3.4 volatility points
- VRP turns sharply negative during volatility shocks

4.3 Term Structure Features

- VIX basis: VX1 – VIX
- Term slope: VX2 – VX1, VX4 – VX1
- Percentage slopes for standardization

4.4 Volatility of Volatility

- VVIX index level
- Rolling VIX return volatility: 5, 10, 21-day windows
- VIX change and range over multiple horizons

4.5 Regime Indicators

- VIX regime: 0 (low), 1 (normal), 2 (high) based on thresholds
- VIX percentile: Expanding window percentile rank (252-day minimum) – no look-ahead bias
- VIX z-score: Rolling 252-day standardized value
- Contango indicator: Binary flag for positive term structure slope (76.5% when VX1 data available)

4.6 SKEW Features

- SKEW z-score: Rolling 252-day standardized SKEW value
- SKEW percentile: Rolling 252-day percentile rank
- SKEW changes: 5-day and 21-day changes
- SKEW/VIX ratio: Relative tail risk vs. overall volatility
- High SKEW regime: Binary indicator for elevated tail risk (SKEW > 130)

4.7 Put/Call Sentiment Features

- Moving averages: 5-day and 21-day smoothed P/C ratios
- P/C z-score: Rolling 63-day standardized value
- Extreme P/C indicators: Binary flags for ratios > 1.2 (bearish) or < 0.6 (bullish)
- P/C spread: Difference between equity and index P/C ratios

5 Data Quality Assessment

5.1 Missing Data Analysis

Table 5: Missing Data Summary (Trading Days Only)

Series	**Missing %**	**Reason**
VIX (OHLC)	0%	Full coverage (trading day filter)
S&P 500	0%	Full coverage
VVIX	4.5%	Later start date (2006-03-26)
VIX9D	25.8%	Series started 2011
VIX3M	20.0%	Series started 2009
VIX6M	11.0%	Series started 2008
SKEW	2.8%	Small gaps in early data
Put/Call Ratios	27%	Data ends October 2019
VX1–VX5	31%	Futures started 2007
NFCI, STLFSI4	68–75%	Weekly frequency

Missing data arises primarily from:

1. **Staggered series inception**: Newer indices (VIX9D, VIX3M) have shorter histories
2. **Frequency mismatch**: Weekly series (NFCI) mapped to daily trading dates
3. **Data availability**: VIX futures electronic trading began 2007
4. **Historical data cutoff**: CBOE put/call ratio data ends October 2019

Handling Strategy: For modeling, we recommend either:

- Using a subset of features with complete data (VIX, S&P 500, Treasury yields)
- Restricting the analysis to 2012–2019 when futures and P/C data overlap
- Applying forward-fill for weekly economic indicators
- Using separate models for post-2019 without P/C features

5.2 Data Validation

We implement automated validation checks:

- Non-negative VIX values (all indices should be positive)
- Reasonable price ranges (VIX < 200, S&P 500 > 0)
- Monotonic date indices without gaps
- Consistent cross-series relationships (VX1 $\approx$ VIX)

6 Volatility Regime Analysis

6.1 Regime Classification

We define three volatility regimes based on VIX levels:

- **Low Volatility** (VIX < 15): Calm market conditions
- **Normal Volatility** ($15 \leq$ VIX < 25): Typical trading environment
- **High Volatility** (VIX ≥ 25): Elevated uncertainty/stress

6.2 Regime Statistics

Table 6: Volatility Regime Characteristics

Regime	Frequency	Avg. VIX	Avg. SPX Return
Low Vol (VIX < 15)	30.2%	12.5	Positive
Normal Vol ($15 \leq$ VIX < 25)	50.1%	18.9	Near zero
High Vol (VIX ≥ 25)	19.7%	34.2	Negative

Key observations:

- The market spends roughly half of the time in “normal” volatility conditions
- High volatility regimes, while less frequent, are associated with significant market stress
- Regime persistence is high, suggesting Markov-switching models are appropriate

7 Correlation Structure

Key correlations in the dataset:

- **VIX vs S&P 500 returns**: Strong negative correlation ($\rho \approx -0.72$ using % changes, $\rho \approx -0.81$ using point changes)
- **VIX vs VVIX**: Positive correlation ($\rho \approx 0.65$)
- **VIX vs High Yield spread**: Positive correlation ($\rho \approx 0.75$)
- **VIX vs SKEW**: Weak positive correlation; SKEW captures different information
- **VIX term structure**: VX1-VX9 highly correlated but with decreasing magnitude
- **VIX vs Treasury curve**: Weak negative correlation with slope (flattening associated with higher VIX)
- **Put/Call ratios vs VIX**: Moderate positive correlation (higher P/C with elevated VIX)

8 Technical Implementation

8.1 Architecture

The data pipeline implements a modular architecture:

```
src/
+-- data/
|   +-- base.py         # Abstract base class for data sources
|   +-- fred.py         # FRED API integration
|   +-- cboe.py         # CBOE web crawler
|   +-- yfinance_source.py  # Yahoo Finance wrapper
|   +-- data_manager.py # Orchestration and merging
+-- features/
|   +-- volatility.py   # Feature engineering
+-- analysis/
    +-- eda.py          # Exploratory analysis
```

8.2 Extensibility

The abstract `BaseDataSource` class provides:

- Standardized interface for all data sources
- Built-in caching with configurable expiration
- Validation framework
- Easy replacement of free sources with premium alternatives

8.3 Reproducibility

The pipeline can be run with:

```
python src/main.py
```

This will:

1. Download fresh data from all sources
2. Compute derived features
3. Save processed dataset to `data/processed/`
4. Generate data quality report

9 Conclusions

We have constructed a comprehensive volatility dataset suitable for regime prediction research. Key strengths:

1. **Comprehensive coverage**: Multiple volatility measures, SKEW tail risk, put/call sentiment, term structure data, and macro indicators
2. **Long history**: Nearly 20 years of data covering multiple market cycles (2006–2025)
3. **Research-ready features**: Pre-computed realized volatility, regime indicators, sentiment metrics, and term structure features
4. **Multi-dimensional sentiment**: VIX (level), SKEW (tail risk), and P/C ratios (positioning) capture different aspects of market fear
5. **Extensible design**: Easy to upgrade to premium data sources

9.1 Data Limitations

Important limitations to consider:

- **Forward-looking features**: `vrp_forward` and `vrp_vol_points_forward` use future data and cannot be used as predictors
- Put/call ratio data ends October 2019 (CBOE discontinued free distribution)
- Weekly economic indicators (NFCI, STLFSI) create alignment challenges
- VIX futures term structure data has shorter history than spot VIX (VX1 from 2013)
- Look-ahead bias must be carefully managed in feature engineering

9.2 Next Steps

Recommended modeling approaches for this dataset:

- Hidden Markov Models for regime detection
- Causal discovery methods (e.g., PC algorithm, NOTEARS) to identify lead-lag relationships
- Machine learning models (Random Forest, XGBoost) for regime prediction
- Neural networks for complex nonlinear relationships

References

[1] CBOE (2024). *VIX Index Methodology.* Chicago Board Options Exchange.

[2] Federal Reserve Bank of St. Louis (2024). *FRED Economic Data.* `https://fred.stlouisfed.org/`

[3] Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *Journal of Business*, 53(1), 61-65.

[4] Whaley, R. E. (2009). Understanding the VIX. *Journal of Portfolio Management*, 35(3), 98-105.

A Data Visualizations

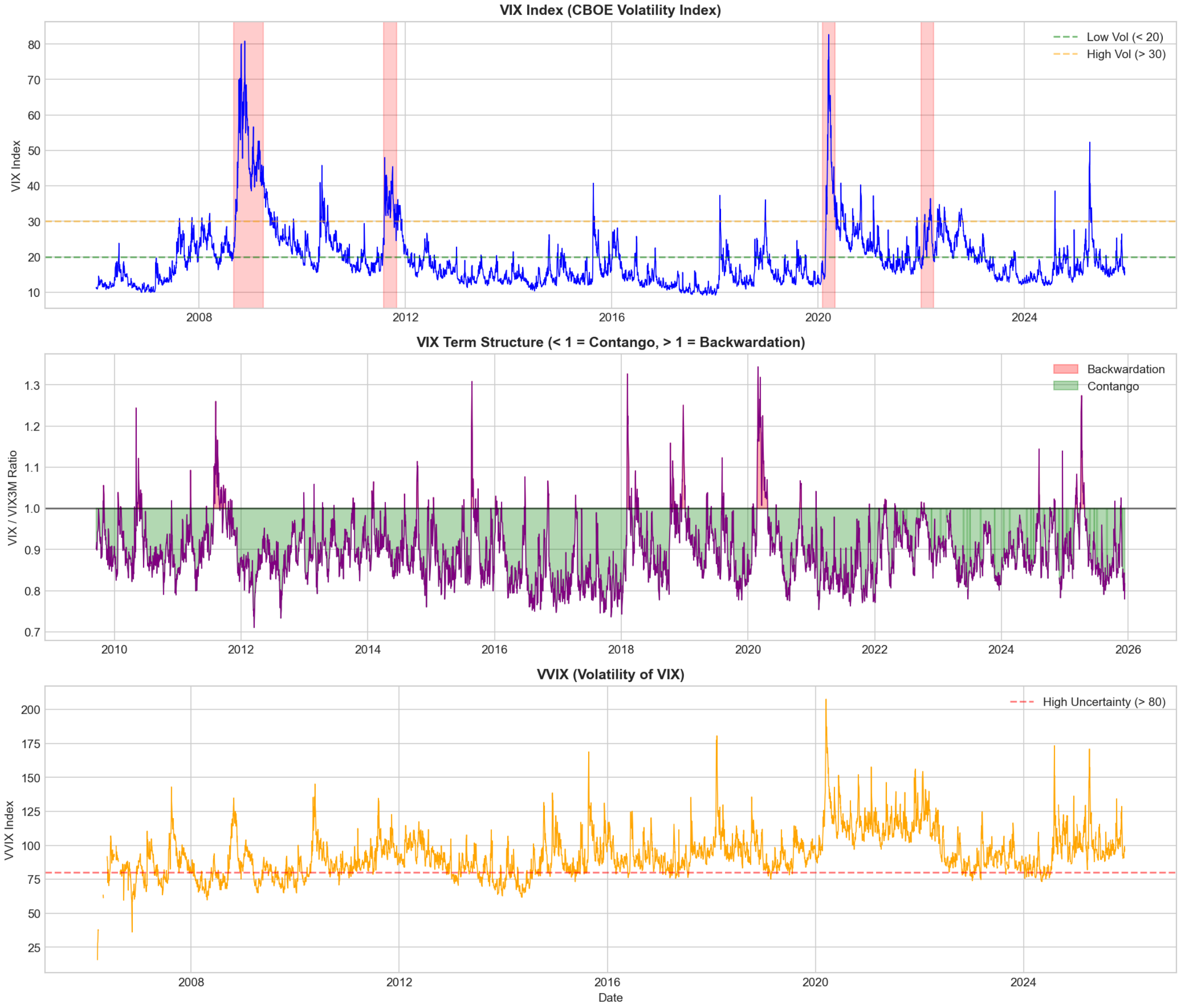

Figure 1: VIX Index historical time series (2006–2025), showing major volatility spikes during financial crises.

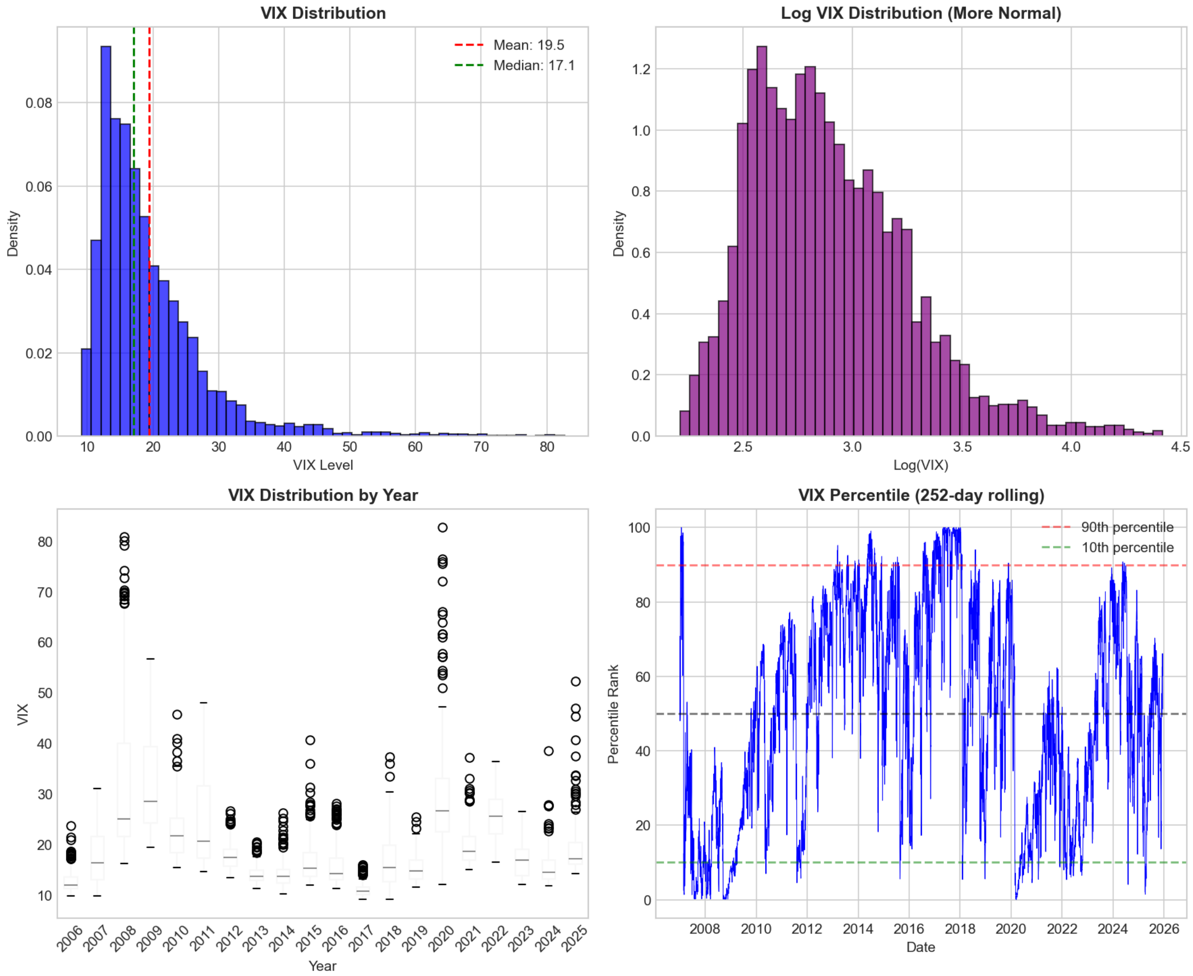

Figure 2: VIX distribution analysis showing the characteristic right-skewed distribution with mean ≈ 19.5 and positive skewness.

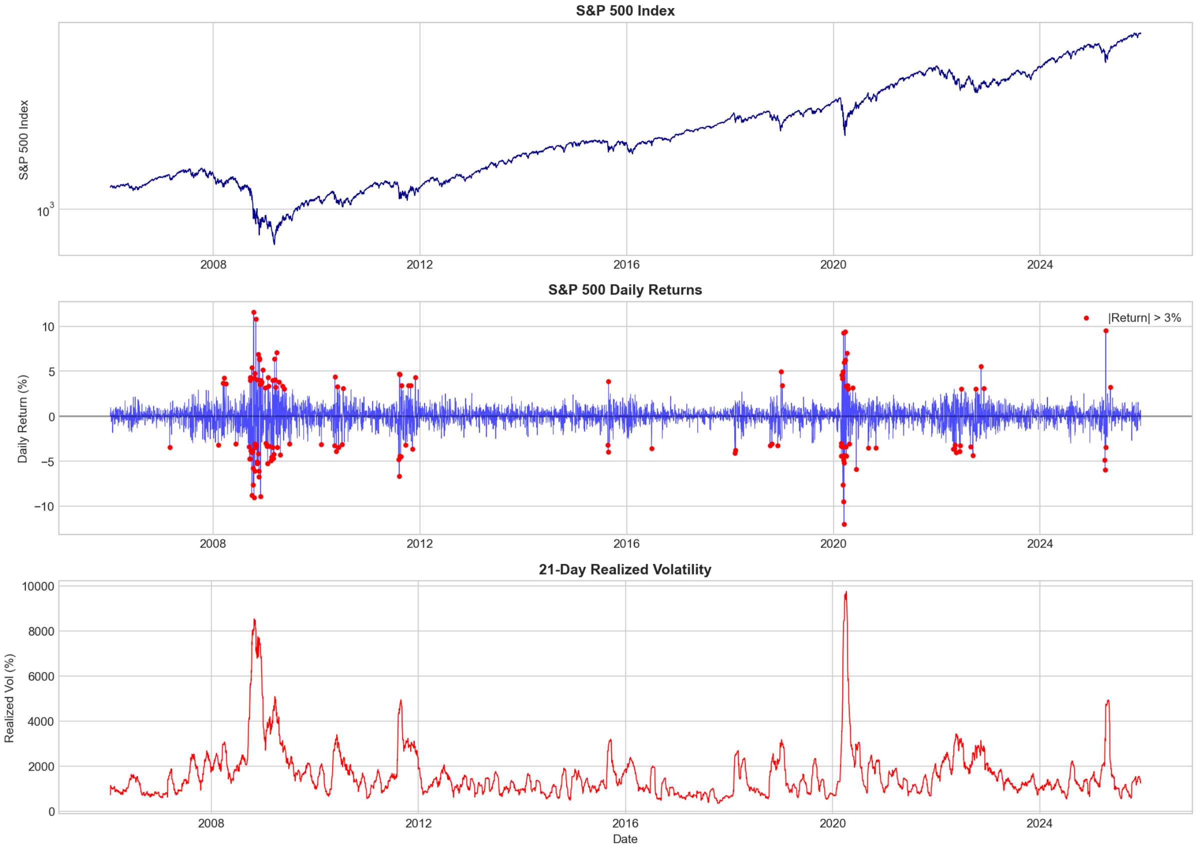

Figure 3: S&P 500 daily returns distribution showing fat tails and volatility clustering typical of financial returns.

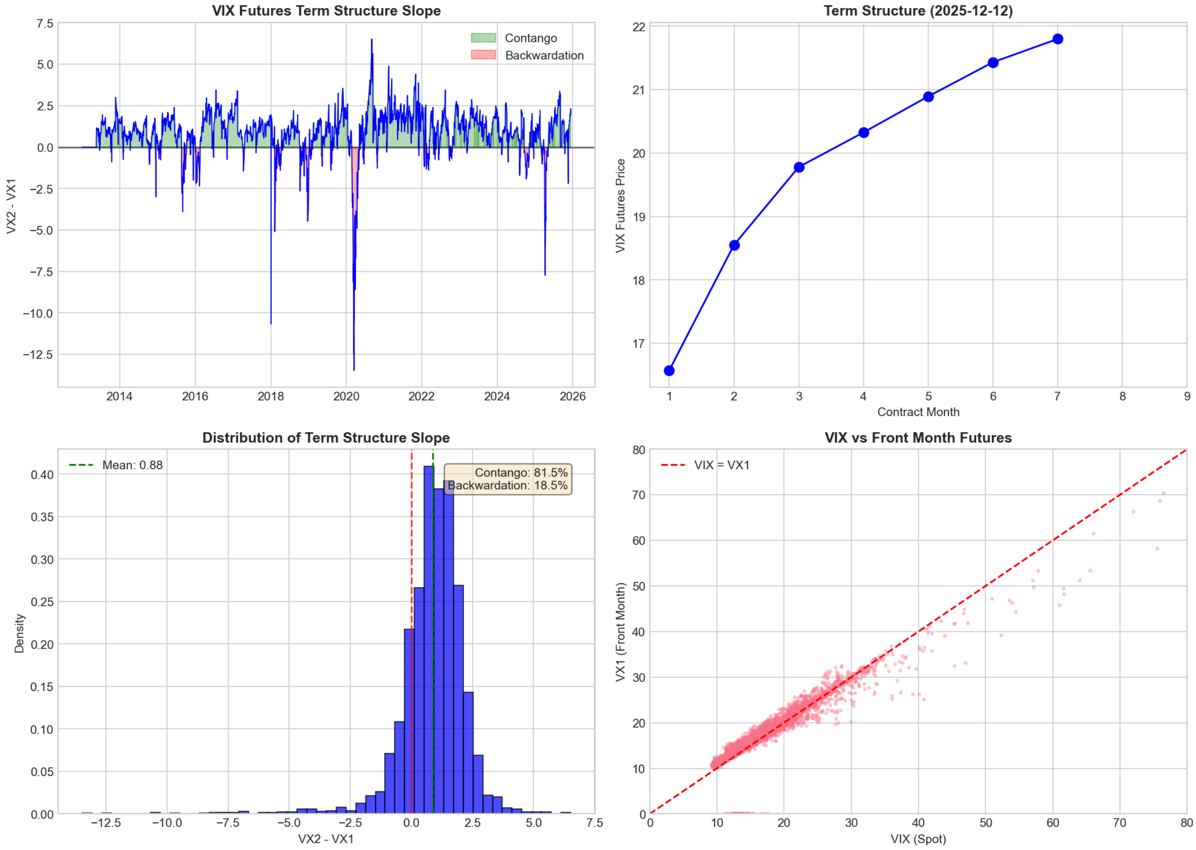

Figure 4: VIX futures term structure dynamics. The market is typically in contango (upward sloping) approximately 76.5% of the time (when VX1 data available, 2013–2025).

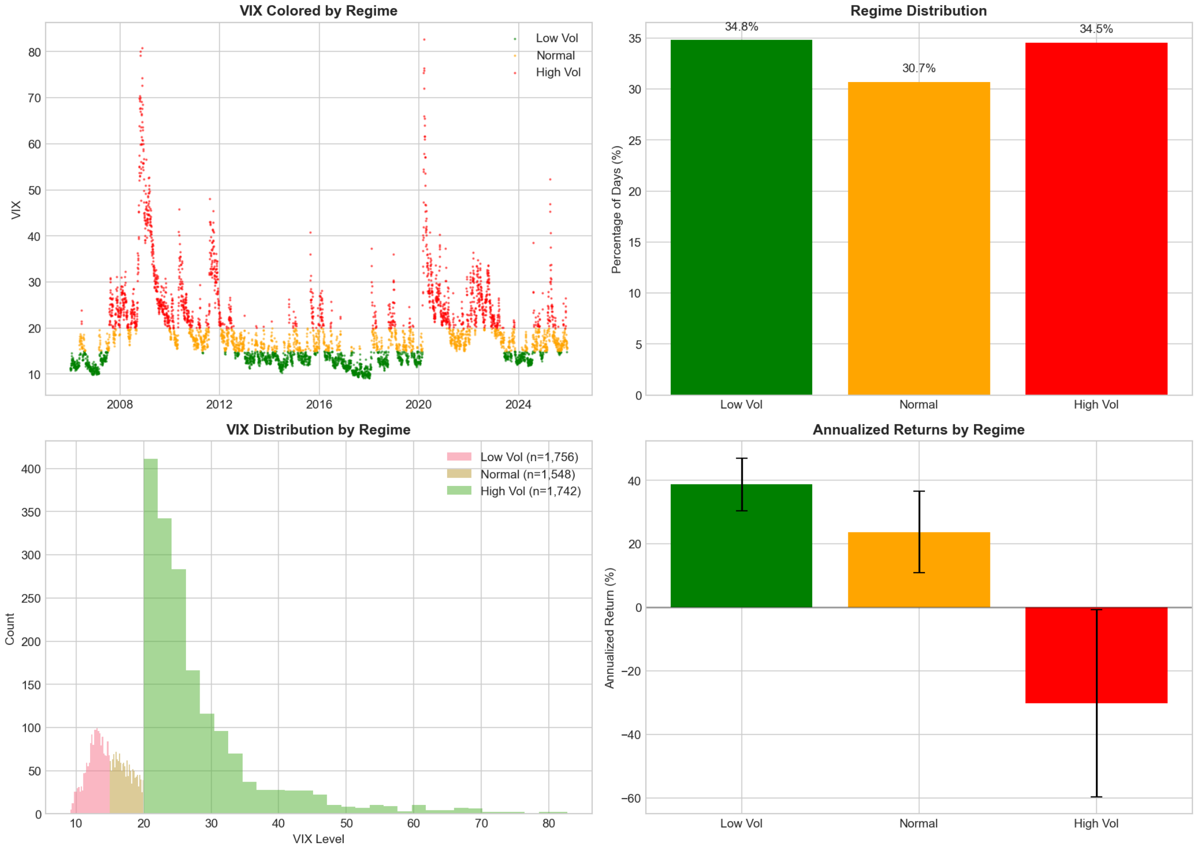

Figure 5: Volatility regime analysis showing the distribution of low, normal, and high volatility regimes over the sample period.

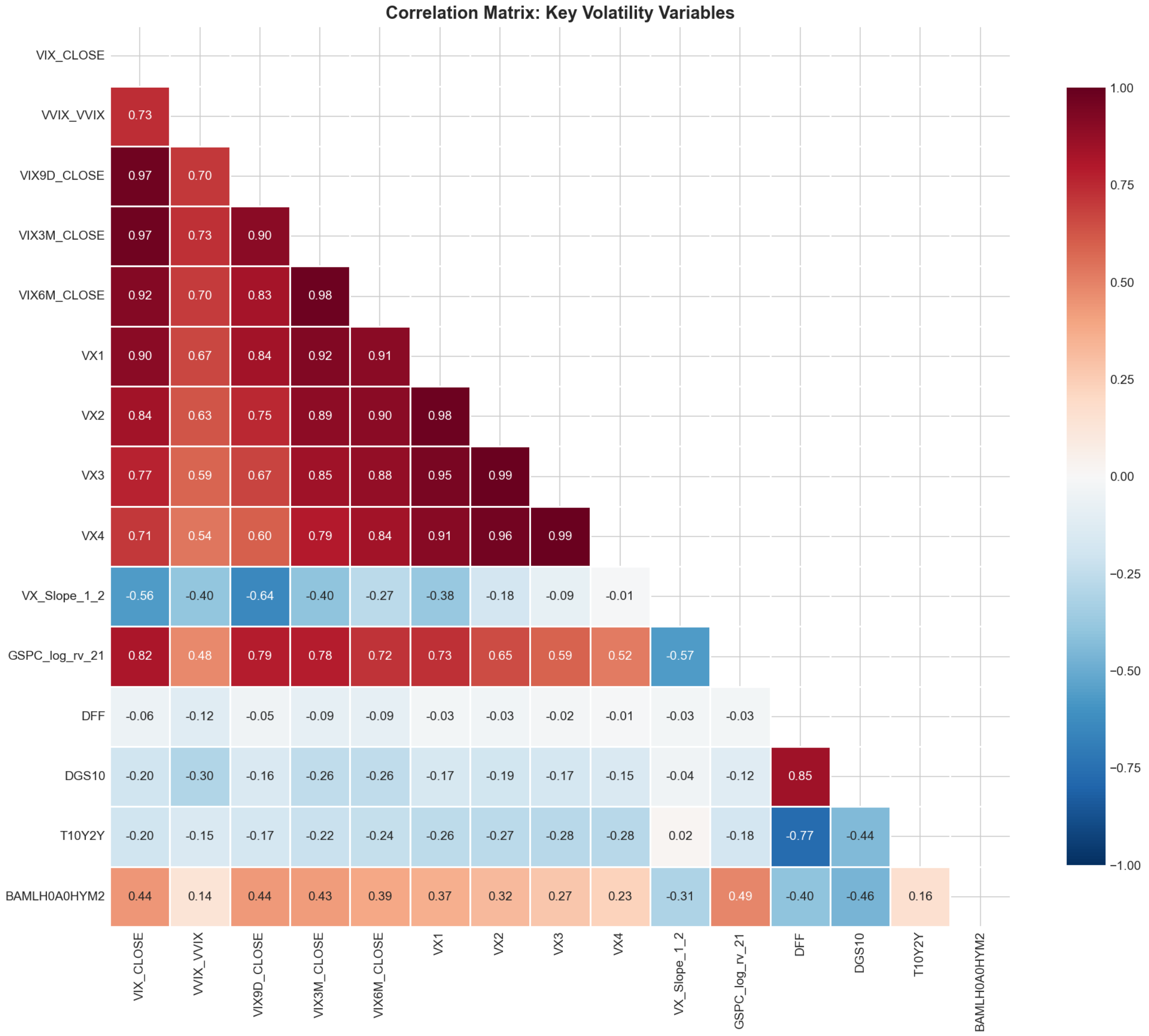

Figure 6: Correlation matrix of key features showing the strong negative VIX-SPX relationship ($\rho \approx -0.81$) and term structure correlations.

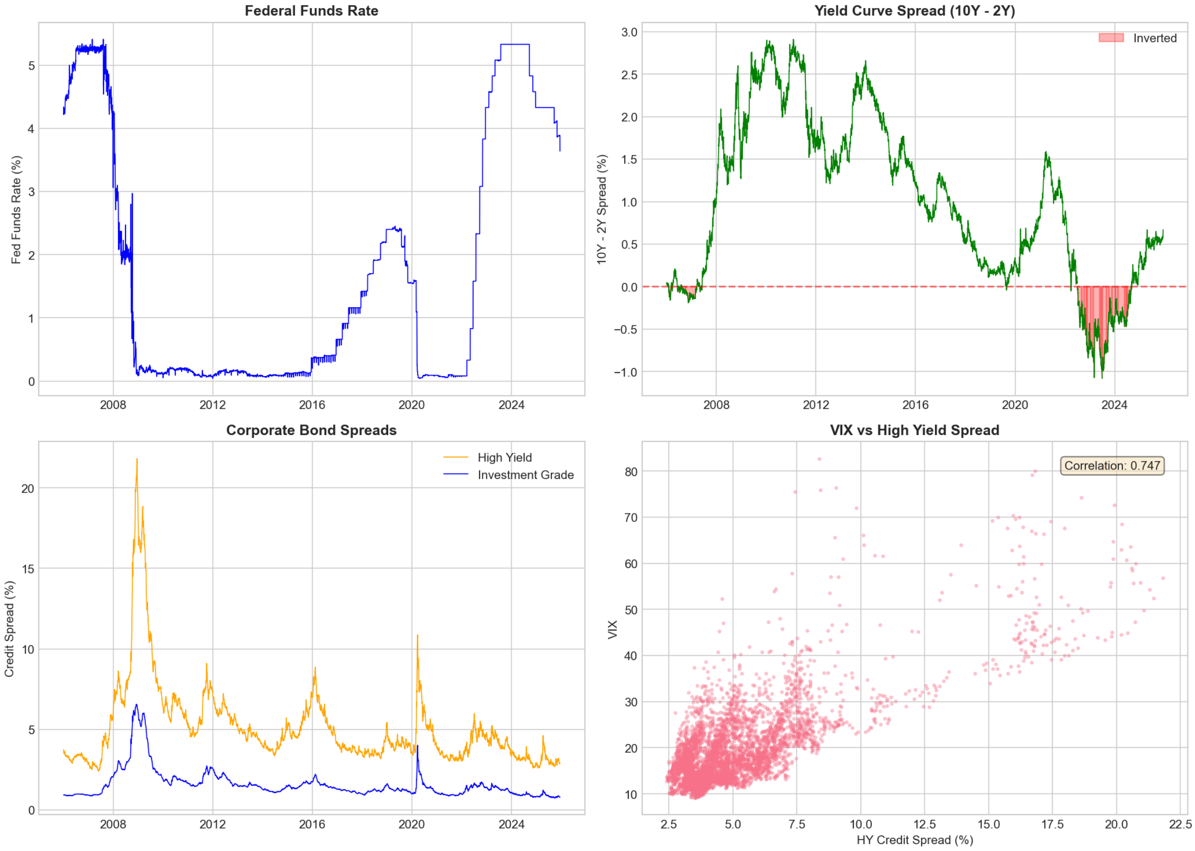

Figure 7: Economic indicators time series including Treasury yields, credit spreads, and financial conditions indices.