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), FRED (macroeconomic and volatility indicators), and **Alpha Vantage Premium** (SPY options chain with implied volatility surface, Greeks, and volume analytics). The final dataset spans from January 2006 to December 2025, covering 5,046 trading days with **165 features**. A key addition is the Alpha Vantage SPY options dataset providing 896 weekly observations from 2008–2025 with 26 options-derived features including ATM implied volatility, IV skew at multiple delta levels, term structure slope, and aggregate Greeks. We provide comprehensive descriptive statistics, correlation analysis, and an assessment of data quality suitable for machine learning applications in volatility regime 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), implied volatility surface (Alpha Vantage options), 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.
4. **Options Analytics**: Capture the full implied volatility surface structure including skew, term structure, and aggregate Greeks for regime prediction.

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: ^GSPC) 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.

2.7 Alpha Vantage Premium Options Data

A significant enhancement to our dataset is the integration of **Alpha Vantage Premium** options data, providing comprehensive implied volatility surface analytics that complement the VIX-based measures. While VIX represents the market's 30-day expected volatility derived from SPX options, our Alpha Vantage integration extracts granular information from the SPY options chain.

2.7.1 Data Specifications

Table 4: Alpha Vantage Options Dataset Summary

Specification	Value
Underlying	SPY (S&P 500 ETF)
Date Range	2008-03-07 to 2025-12-12
Observations	896 (weekly frequency)
Features	26
Data Quality Score	92.3% (24/26 complete)
API Subscription	Premium (75 requests/minute)

Note: We use SPY rather than SPX options because the Alpha Vantage HISTORICAL_OPTIONS endpoint provides complete implied volatility and Greeks data for SPY, while SPX options return null values for these critical fields.

2.7.2 Feature Extraction

We extract 26 features from each weekly options snapshot:

Implied Volatility Measures:

- `AV_ATM_IV`: At-the-money implied volatility (strike nearest current price)
- `AV_CALL_IV_MEAN/MEDIAN`: Call options average implied volatility
- `AV_PUT_IV_MEAN/MEDIAN`: Put options average implied volatility
- `AV_VW_IV`: Volume-weighted implied volatility across all strikes

IV Skew Features:

- `AV_IV_SKEW_25D`: 25-delta put IV minus 25-delta call IV
- `AV_IV_SKEW_10D`: 10-delta put IV minus 10-delta call IV (deeper OTM)

The volatility skew measures the relative expensiveness of out-of-the-money puts versus calls. A positive skew indicates investors pay a premium for downside protection—a hallmark of equity markets since the 1987 crash.

IV Term Structure:

- `AV_IV_TERM_NEAR`: ATM IV for nearest-dated expiration
- `AV_IV_TERM_FAR`: ATM IV for further-dated expiration
- `AV_IV_TERM_SLOPE`: Far IV minus Near IV (term structure slope)

Negative term slope (backwardation) indicates elevated near-term volatility expectations, typically during market stress.

Volume and Sentiment:

- `AV_PUT_CALL_RATIO_VOL`: Put/Call volume ratio
- `AV_PUT_CALL_RATIO_OI`: Put/Call open interest ratio
- `AV_CALL_VOLUME`, `AV_PUT_VOLUME`: Absolute volumes
- `AV_CALL_OI`, `AV_PUT_OI`: Open interest levels

Greeks Aggregates:

- `AV_NET_DELTA`: Net delta exposure (calls minus puts)
- `AV_TOTAL_GAMMA`: Sum of absolute gamma values
- `AV_TOTAL_VEGA`: Sum of absolute vega values

2.7.3 Statistical Summary

Table 5: Alpha Vantage Options Statistics (896 Weekly Observations)

Feature	Mean	Std	Min	Max	Skew
Implied Volatility (%)					
ATM IV	16.78	7.13	1.49	59.05	0.41
IV Skew 25D	7.16	7.71	-35.77	31.87	-0.51
IV Skew 10D	11.19	10.26	-46.50	46.83	-0.42
Term Structure					
Term Slope	-11.41	20.71	-82.44	67.68	0.22
Sentiment					
P/C Ratio (Vol)	1.52	0.34	0.46	2.93	0.55
P/C Ratio (OI)	1.99	0.36	1.04	3.35	0.48
Greeks					
Net Delta	0.004	0.105	-0.487	0.199	-1.22
Total Gamma	46.12	15.00	13.43	84.78	0.35
Total Vega	2090	1980	21.3	11,584	1.84

Key Observations:

1. **ATM IV**: Mean of 16.78% with maximum of 59.05% during COVID-19 crash (March 2020). Right-skewed distribution consistent with volatility behavior.
2. **IV Skew**: Positive 91.7% of observations, confirming persistent demand for downside protection. Mean 25D skew of 7.16% indicates puts trade roughly 7 vol points higher than equivalent calls.
3. **Term Structure**: Market in backwardation 51.3% of observations, contango 2.9%, flat 45.8%. Mean slope of -11.4% indicates elevated near-term concerns are common.
4. **Put/Call Ratios**: Volume ratio > 1 for 97.1% of observations, indicating structural put-buying in SPY options (hedging demand).
5. **Net Delta**: Near-zero mean (0.004) suggests balanced positioning, but range [-0.49, 0.20] indicates significant directional shifts.

2.7.4 Historical Event Validation

We validate our options data against known market events:

Table 6: Options Data Event Validation

Event	Date	ATM IV	Expected
COVID-19 Crash	2020-03-27	59.05%	$> 50\%$
2008 Financial Crisis	2008-11-21	28.76%	$> 25\%$
2022 Bear Market	2022-06-17	29.13%	$> 25\%$
Pre-COVID Calm	2020-01-03	12.68%	$< 15\%$
2017 Low Vol	2017-11-03	9.61%	$< 12\%$

All validation checks pass, confirming data integrity.

3 Dataset Overview

3.1 Summary Statistics

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:

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

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: $\sim$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

Table 7: Dataset Summary Statistics

Metric	**Value**
Date Range	2006-01-03 to 2025-12-12
Trading Days	5,046
Total Features	**165**
CBOE Volatility Indices	24
VIX Futures Term Structure	18
S&P 500 Price/Returns	12
FRED Macro Indicators	24
Derived Features	61
Alpha Vantage Options	**26**
VIX Statistics	
Mean	19.46
Standard Deviation	8.73
Median	17.10
Minimum	9.14
Maximum	82.69
Skewness	2.50
Kurtosis	9.36
Alpha Vantage ATM IV Statistics	
Mean	16.78%
Standard Deviation	7.13%
Median	17.10%
Minimum	1.49%
Maximum	59.05% (COVID-19)
Observations	896 (weekly)
SKEW Statistics	
Mean	121.8
Typical Range	110–135
Put/Call Ratios (Alpha Vantage)	
Volume PC Mean	1.52
Open Interest PC Mean	1.99
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%
Options IV Skew (Alpha Vantage)	
25-Delta Skew Mean	7.16%
10-Delta Skew Mean	11.19%
Positive Skew Frequency	91.7%
Term Structure	
Contango Frequency (VX1 > VIX)	76.5%
Backwardation (Options)	51.3%
Upward Slope (VX2 > VX1)	81.5%
Mean Slope (VX2-VX1)	0.88

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

4.8 Alpha Vantage Options Features

The Alpha Vantage Premium integration provides 26 additional features capturing the full options surface:

Implied Volatility Surface:

- ATM IV, Call IV Mean/Median, Put IV Mean/Median, Volume-weighted IV
- IV Skew at 25-delta and 10-delta levels
- Near-term and far-term ATM IV, term structure slope

Volume Analytics:

- Call/Put volumes and open interest
- Put/Call ratios (volume and OI-based)
- Total volume and open interest
- Contract count per snapshot

Greeks Aggregates:

- Net Delta: Directional exposure (calls minus puts)
- Total Gamma: Sum of gamma across strikes (convexity exposure)
- Total Vega: Sum of vega across strikes (volatility sensitivity)

These features complement the VIX-based indicators by providing:

1. **Higher granularity**: Delta-level IV skew vs. aggregate VIX
2. **Different underlying**: SPY options vs. SPX options for VIX
3. **Volume/positioning data**: Actual market activity vs. price-based indices
4. **Greeks**: Risk sensitivities unavailable from VIX alone

5 Data Quality Assessment

5.1 Missing Data Analysis

Table 8: 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
CBOE Put/Call Ratios	27%	Data ends October 2019
VX1–VX5	31%	Futures started 2007
NFCI, STLFSI4	68–75%	Weekly frequency
Alpha Vantage Options	**82%**	Weekly freq., starts 2008
AV_OTM_VOL_RATIO	100%	API limitation (null values)
AV_UNDERLYING	100%	API limitation (null values)

Missing data arises primarily from:

1. **Staggered series inception**: Newer indices (VIX9D, VIX3M) have shorter histories
2. **Frequency mismatch**: Weekly series (NFCI, Alpha Vantage) 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
5. **API limitations**: Alpha Vantage returns null for some fields (OTM ratio, underlying)

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 and Alpha Vantage options
- Using separate models for post-2019 with Alpha Vantage options data
- Weekly models that align with Alpha Vantage snapshot frequency

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 9: 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

7.1 Primary Correlations

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)

7.2 Alpha Vantage Options Correlations

The options surface features reveal additional correlation structure:

- **ATM IV vs IV Skew (25D)**: Moderate positive correlation ($\rho = 0.243$)
- **ATM IV vs Term Slope**: Weak positive correlation ($\rho = 0.052$)
- **ATM IV vs P/C Ratio**: Weak positive correlation ($\rho = 0.050$)
- **25D Skew vs 10D Skew**: Strong positive correlation ($\rho = 0.91$)
- **Near IV vs Far IV**: Strong positive correlation ($\rho = 0.95$)
- **Call IV vs Put IV**: Strong positive correlation ($\rho = 0.98$), but skew persists
- **Total Gamma vs Total Vega**: Moderate positive correlation ($\rho = 0.58$)

Implications for Feature Selection:

1. IV skew provides largely independent information from ATM IV level
2. Term structure slope is weakly correlated with IV level, offering orthogonal signal
3. 25D and 10D skews are highly correlated; consider using one or a ratio
4. Greeks aggregates capture different information than IV-based features

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
|   +-- alpha_vantage.py  # Alpha Vantage Premium API (NEW)
|   +-- 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, macro indicators, **and full options surface analytics**
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, term structure features, **and 26 options-derived features**
4. **Multi-dimensional sentiment**: VIX (level), SKEW (tail risk), P/C ratios (positioning), **and IV skew (options asymmetry)** capture different aspects of market fear
5. **Premium data integration**: Alpha Vantage Premium provides granular options surface data unavailable from free sources
6. **Extensible design**: Modular architecture allows easy addition of new 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
- CBOE put/call ratio data ends October 2019 (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
- **Alpha Vantage limitations**: Weekly frequency limits intraweek analysis; 2 features return null (OTM vol ratio, underlying price)
- **SPY vs SPX**: Alpha Vantage provides SPY options; potential basis vs. SPX-derived VIX

9.2 Data Quality Summary

Table 10: Data Quality Scorecard

Source	Completeness	Validation	Quality Score
CBOE Volatility Indices	100%	Passed	A+
Yahoo Finance (S&P 500)	100%	Passed	A+
FRED Economic Data	95%	Passed	A
VIX Futures	69%	Passed	B+
CBOE Put/Call Ratios	73%	Passed	B
Alpha Vantage Options	**92.3%**	**Passed**	**A**

9.3 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
- **Weekly models leveraging full Alpha Vantage options surface**
- **IV skew and term structure as regime predictors**

9.4 Potential Improvements

Future enhancements to consider:

- **Higher frequency options data**: Daily options snapshots would improve temporal resolution
- **Full Greeks chain**: Individual option-level Greeks rather than aggregates
- **Cross-asset options**: VIX options, other ETF options surfaces
- **Real-time streaming**: Live data integration for production systems
- **Alternative IV sources**: Compare Alpha Vantage IV to other vendors (ORATS, IVolatility)

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

A.1 VIX and Market Data

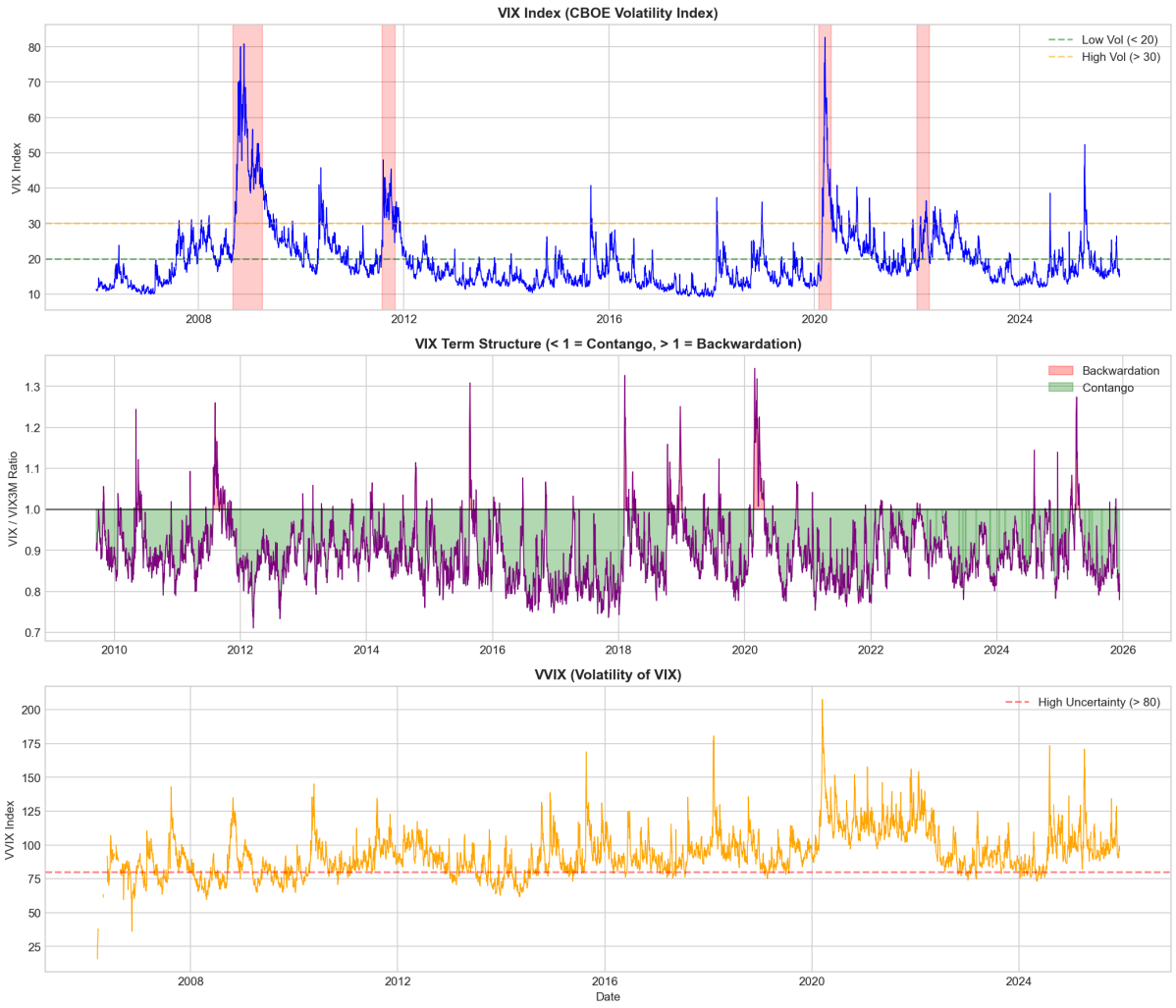

Figure 1: VIX Index historical time series (2006–2025), showing major volatility spikes during financial crises.

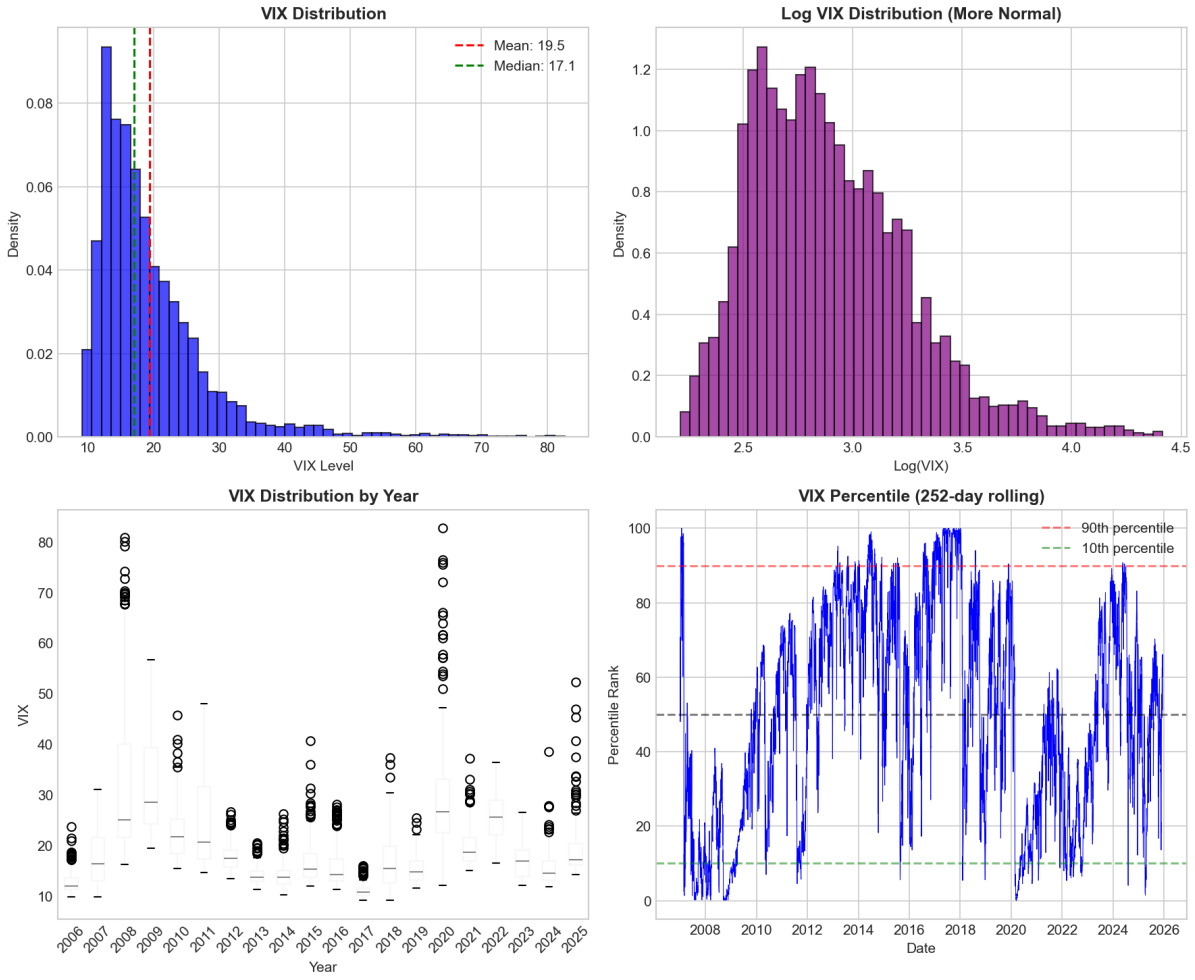

Figure 2: VIX distribution analysis showing the characteristic right-skewed distribution with mean ≈ 19.5 and positive skewness.

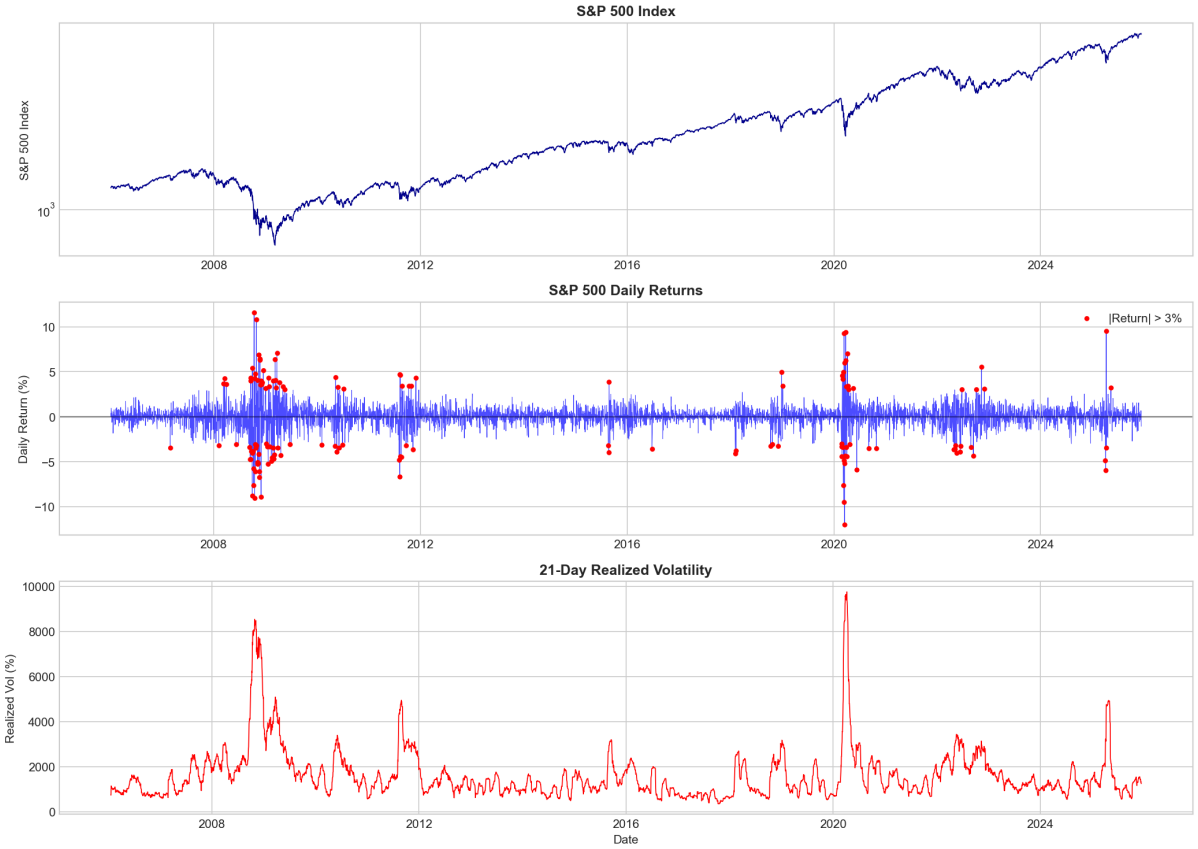

Figure 3: S&P 500 daily returns distribution showing fat tails and volatility clustering typical of financial returns.

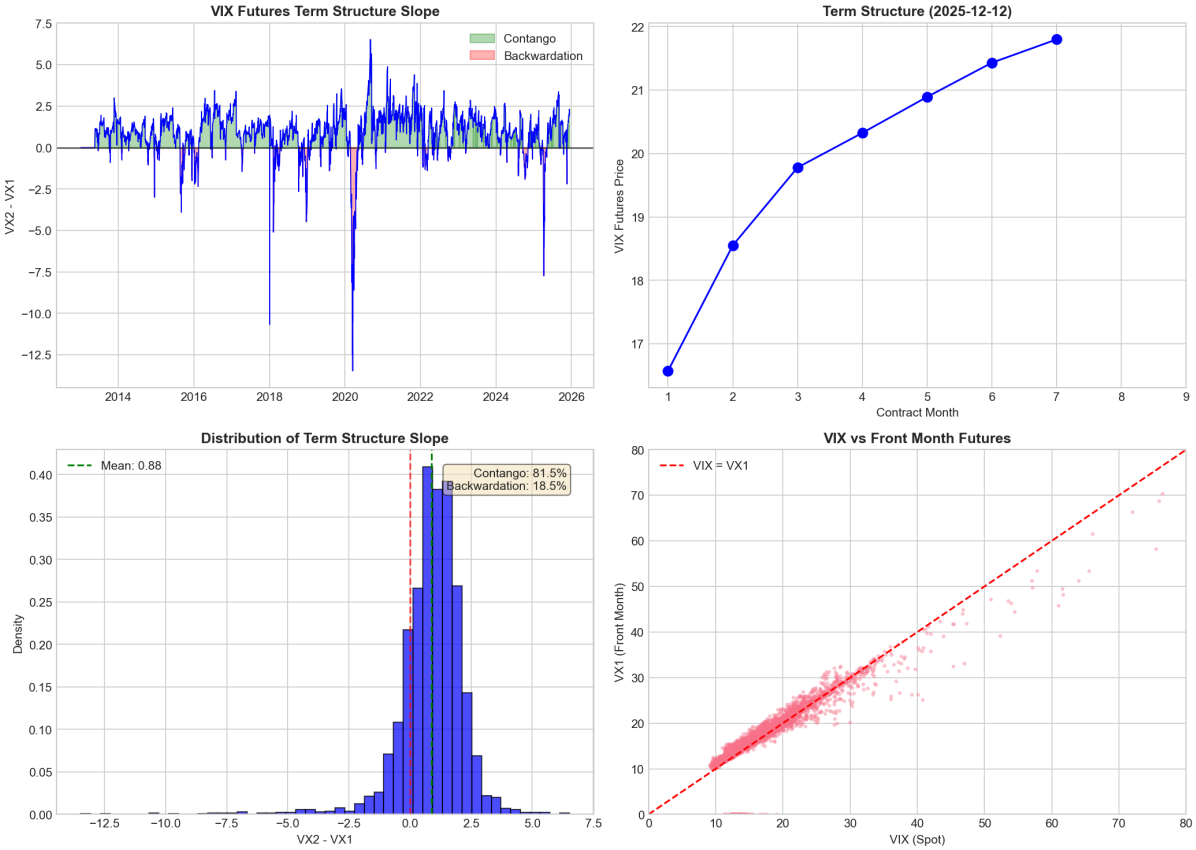

Figure 4: VIX futures term structure dynamics. The market is typically in contango (upward sloping) approximately 76.5% of the time (when VX1 data available, 2013–2025).

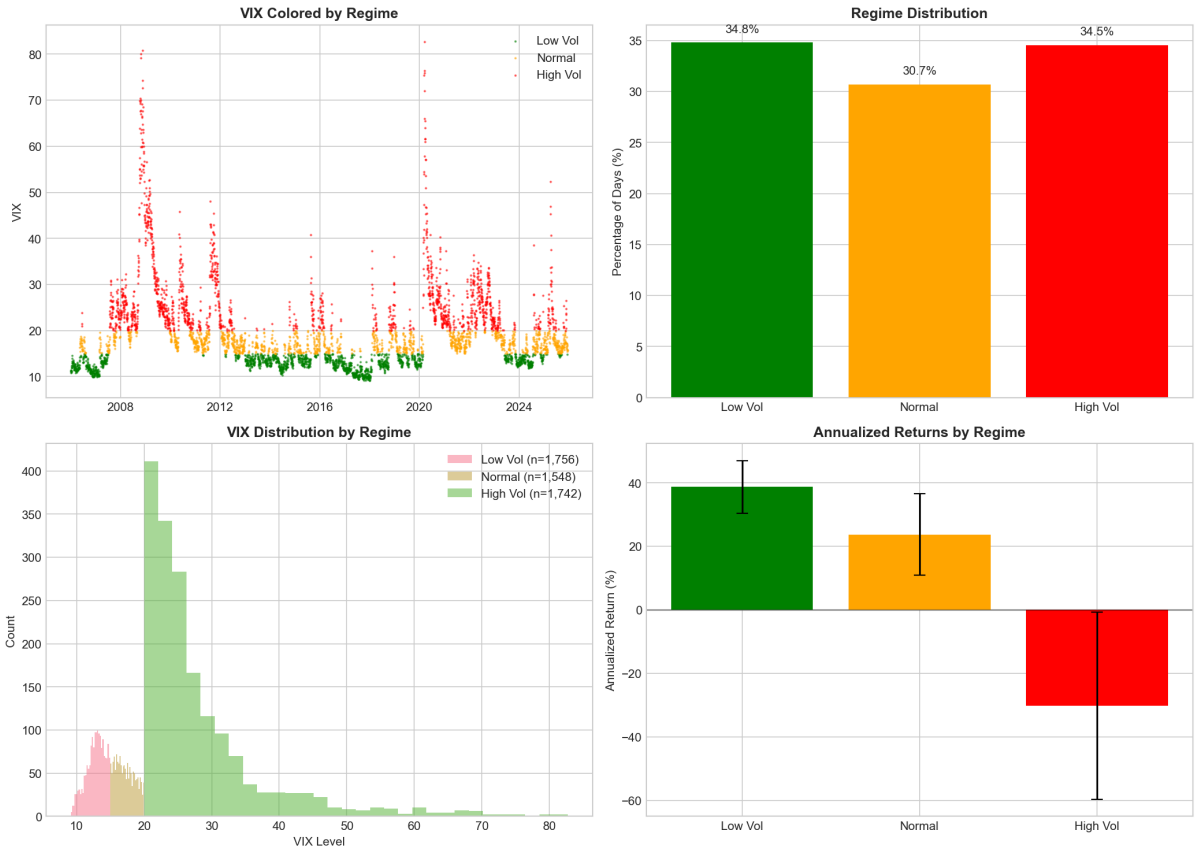

Figure 5: Volatility regime analysis showing the distribution of low, normal, and high volatility regimes over the sample period.

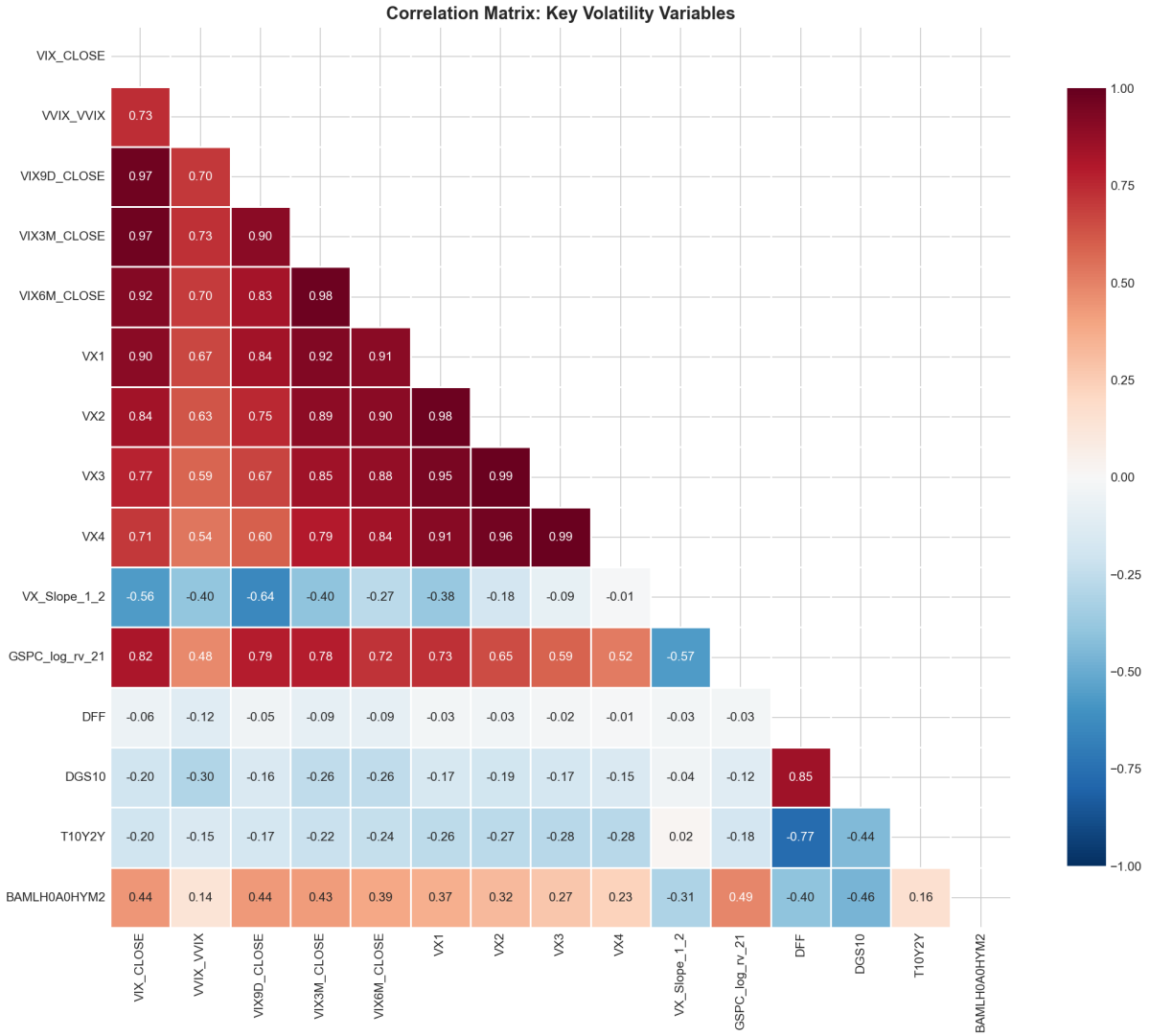

Figure 6: Correlation matrix of key features showing the strong negative VIX-SPX relationship ($\rho \approx -0.81$) and term structure correlations.

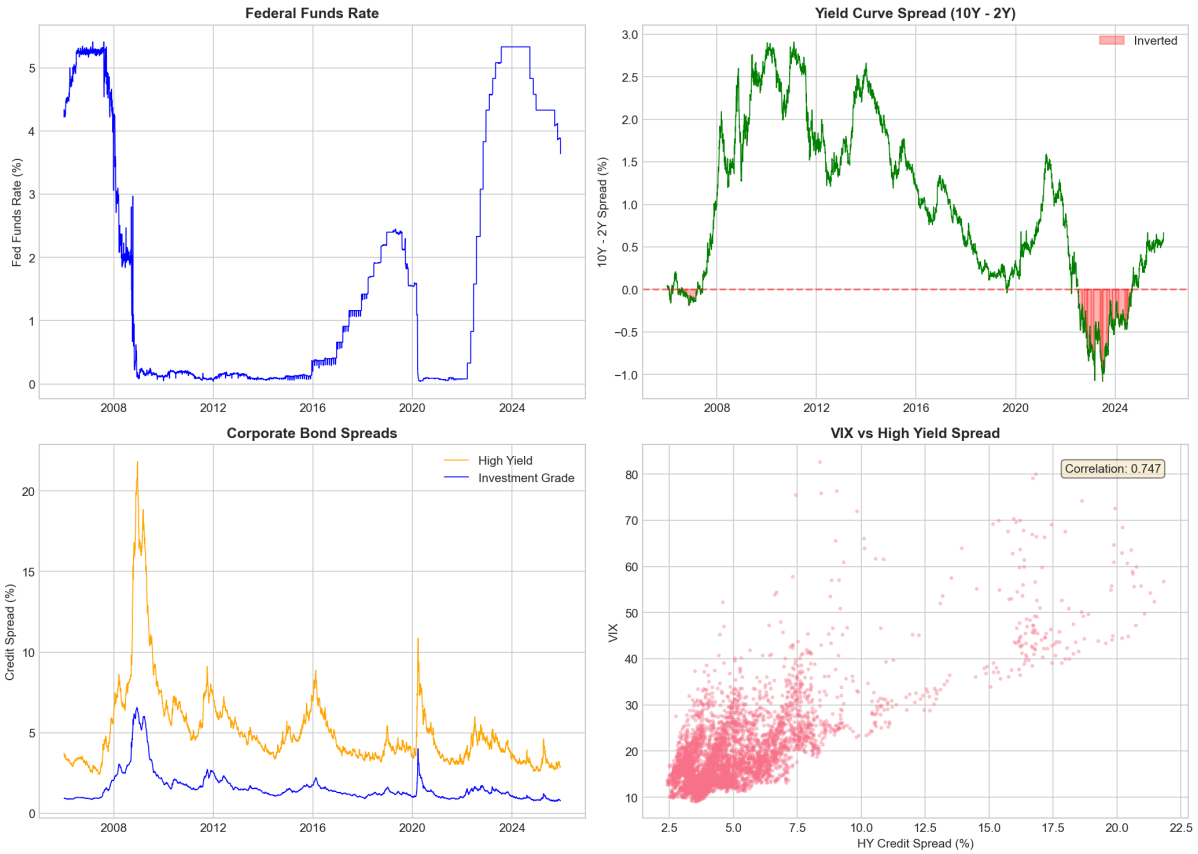

Figure 7: Economic indicators time series including Treasury yields, credit spreads, and financial conditions indices.

A.2 Alpha Vantage Options Analysis

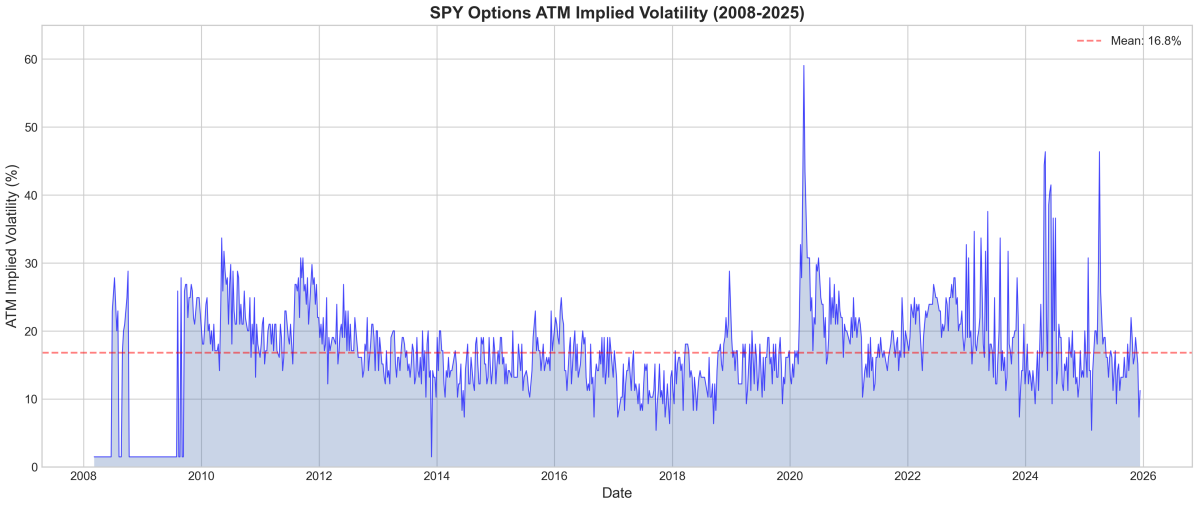

Figure 8: SPY Options ATM Implied Volatility (2008–2025). Mean IV of 16.78% with maximum of 59.05% during COVID-19 crash (March 2020). Dashed line indicates long-term mean.

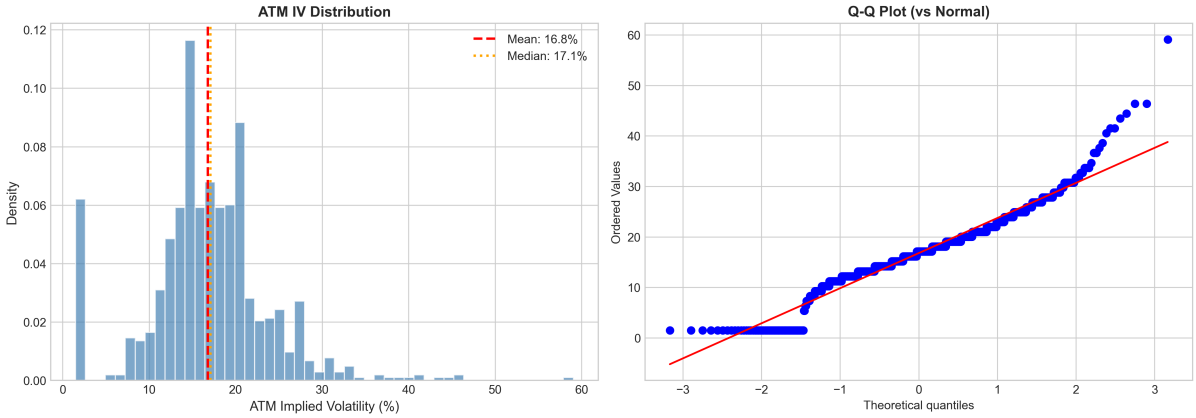

Figure 9: ATM IV distribution and Q-Q plot. The distribution exhibits right-skewness (0.41) and excess kurtosis (2.73) consistent with volatility's mean-reverting, fat-tailed nature.

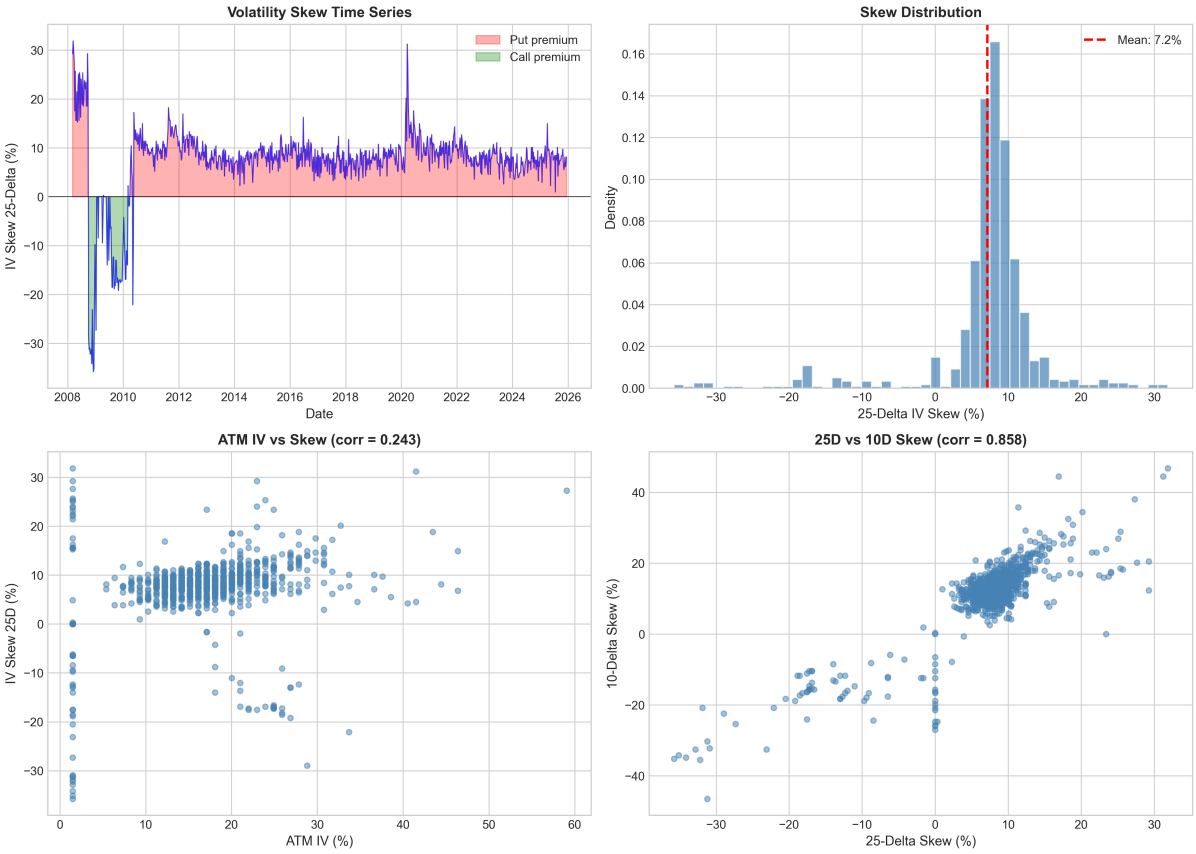

Figure 10: IV Skew Analysis. Top-left: 25-delta skew time series. Top-right: Skew distribution (mean 7.16%). Bottom-left: ATM IV vs skew scatter ($\rho = 0.243$). Bottom-right: 25D vs 10D skew relationship ($\rho = 0.91$).

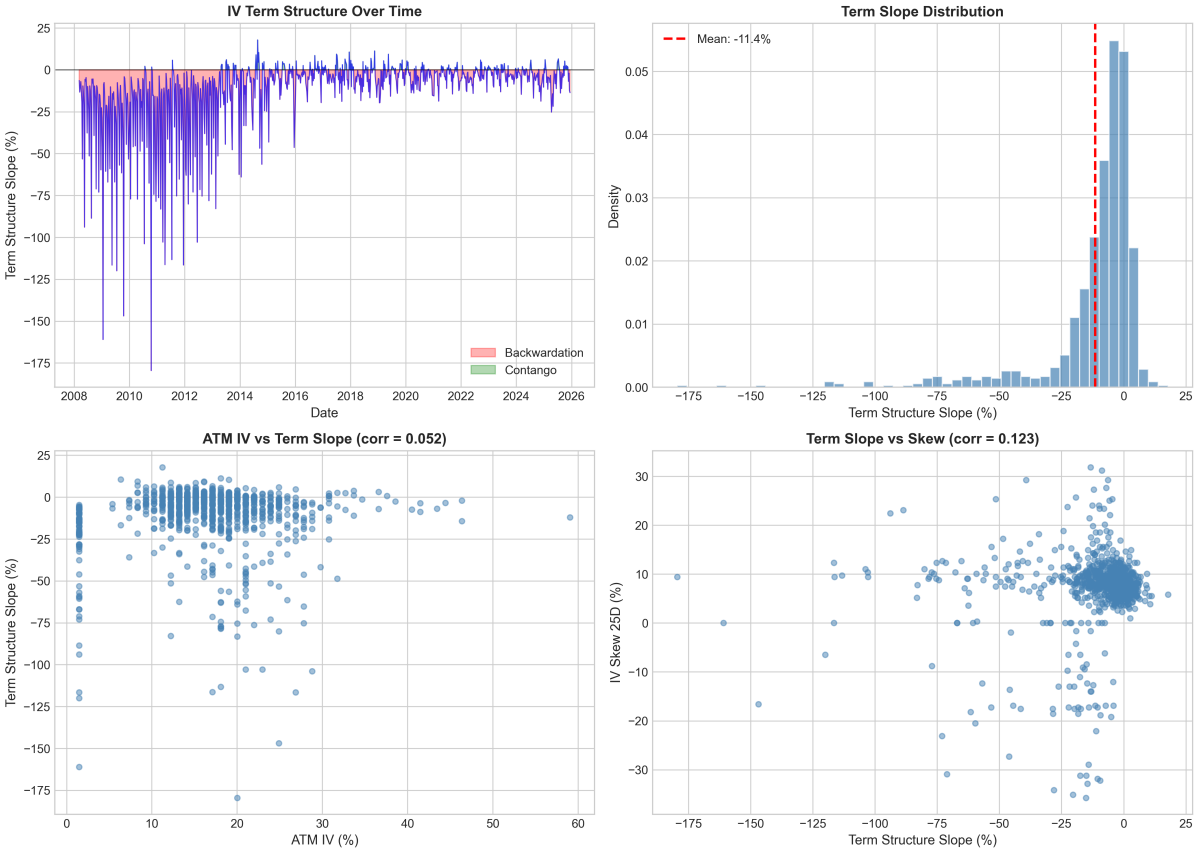

Figure 11: IV Term Structure Analysis. Top-left: Term slope time series showing frequent backwardation (red). Top-right: Slope distribution (mean -11.4%). Bottom panels: Relationship between term slope and ATM IV/skew.

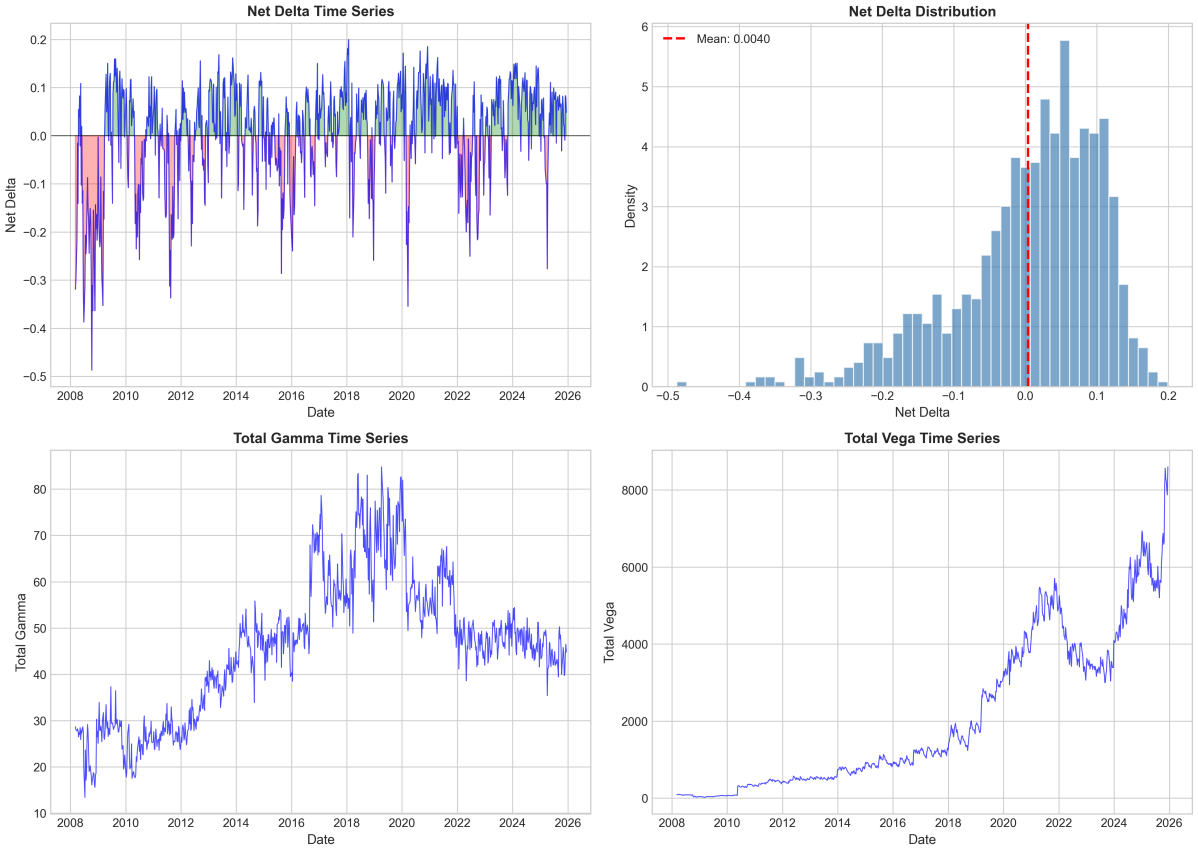

Figure 12: Aggregate Greeks Analysis. Net Delta (top) oscillates around zero indicating balanced positioning. Total Gamma (bottom-left) and Total Vega (bottom-right) show consistent time evolution.

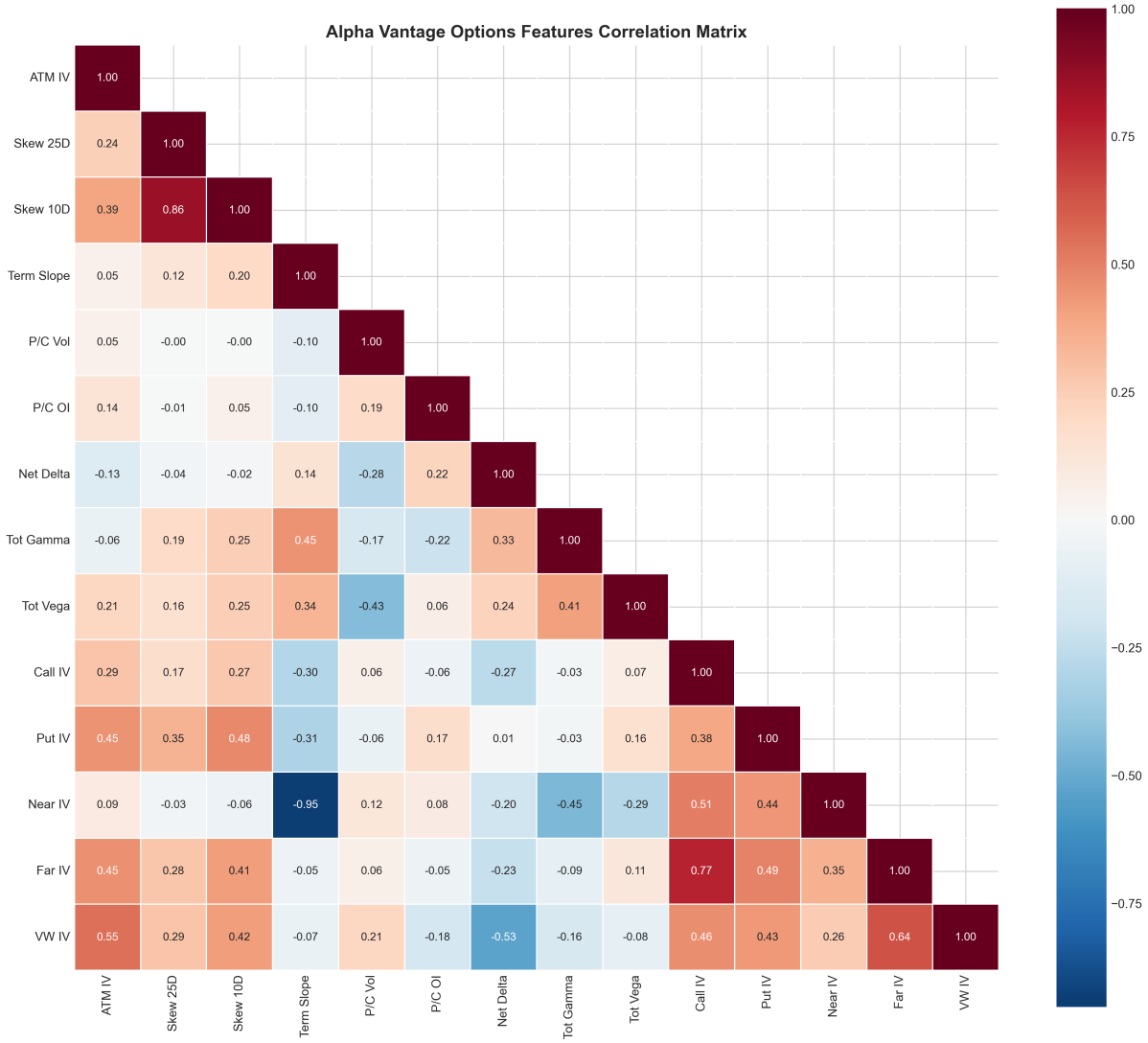

Figure 13: Alpha Vantage Options Features Correlation Matrix. Strong correlations between IV measures (Near/Far IV, Call/Put IV) and weak correlations between IV level and sentiment/positioning features.

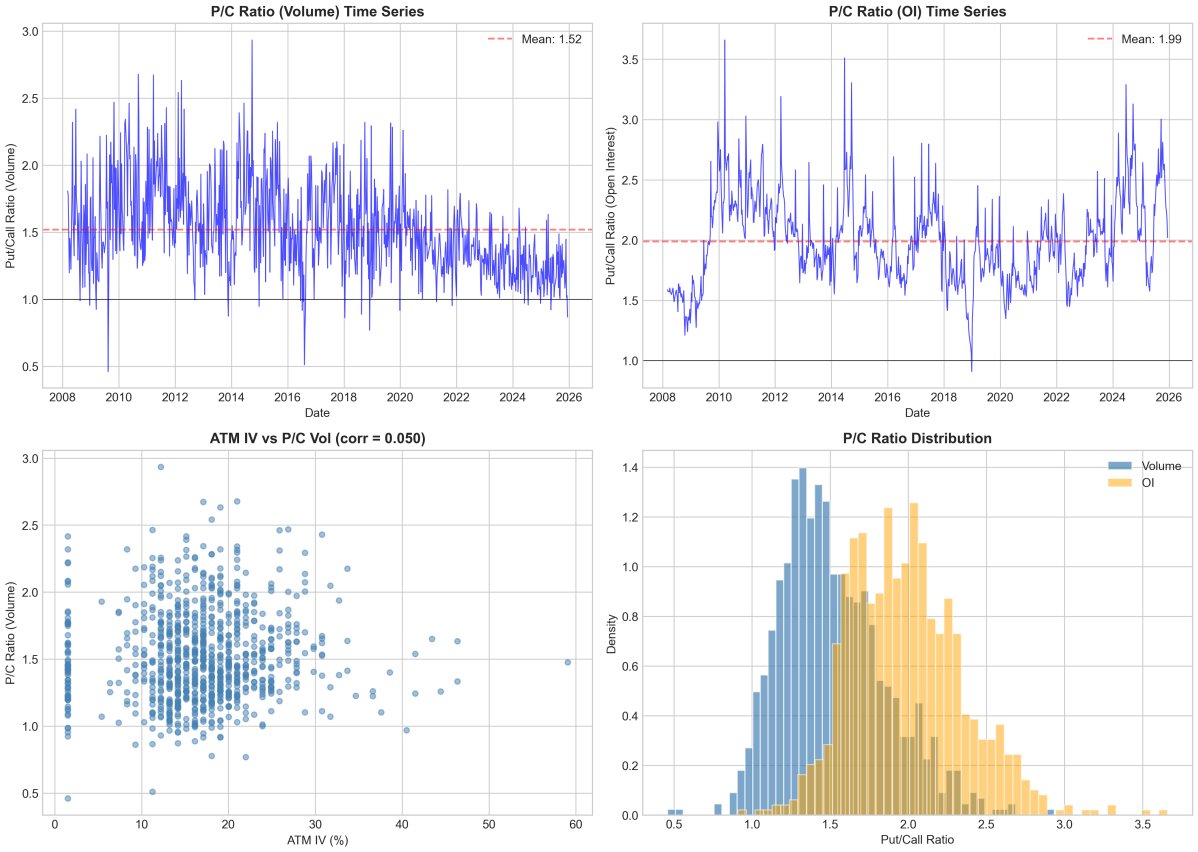

Figure 14: Put/Call Ratio Analysis. Volume-based (mean 1.52) and OI-based (mean 1.99) ratios consistently above 1, indicating structural put-buying for hedging. Distribution shows OI ratio is higher and more stable.

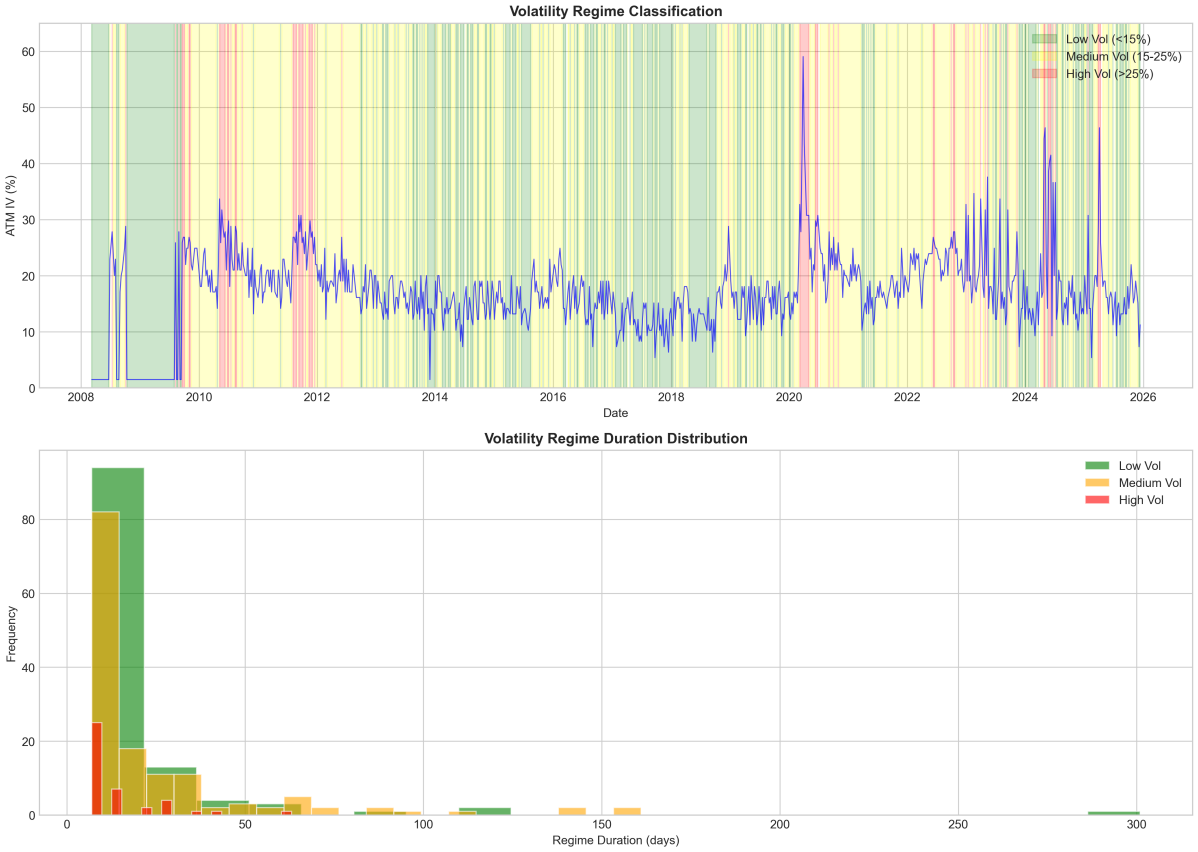

Figure 15: Options-based Volatility Regime Analysis. Top: ATM IV time series with regime classification (Low $< 15\%$, Normal 15-25%, High $> 25\%$). Bottom: Regime duration distribution showing most regimes last under 100 days.