SPY Options Dataset: Comprehensive Description and Summary Statistics

A Foundation for Machine Learning-Based Volatility Modeling

Philipp Dubach
github.com/philippdubach

December 15, 2025

Abstract

This document provides a comprehensive description of a large-scale dataset comprising SPY (SPDR S&P 500 ETF Trust) options data spanning from January 2008 to December 2025. The dataset contains approximately 24.7 million option contracts with associated Greeks, implied volatilities, and market microstructure variables. We present detailed summary statistics, distributional analyses, and temporal patterns relevant for researchers investigating implied volatility surface modeling, option pricing, and volatility forecasting using machine learning methods. The dataset covers multiple market regimes including the 2008 financial crisis, the COVID-19 pandemic volatility spike, and subsequent market recoveries, providing rich variation for model training and evaluation.

Contents

1 Introduction

The implied volatility surface represents one of the most important structures in quantitative finance, encoding market expectations about future asset price distributions across different strike prices and maturities. Understanding and modeling this surface has profound implications for option pricing, risk management, and trading strategies.

This document describes a comprehensive dataset of SPY options designed to support research in:

- Implied volatility surface modeling and interpolation
- Machine learning approaches to option pricing
- Volatility forecasting and term structure analysis
- Market microstructure analysis of options markets
- Arbitrage-free volatility surface construction

SPY options are among the most liquid equity options globally, making them ideal for studying volatility dynamics and developing pricing models that may generalize to less liquid markets.

2 Data Source and Collection

2.1 Data Provider

The options data was collected from market data providers offering historical option chain snapshots. Each record represents a single option contract observed on a specific trading date, capturing the full option chain available for SPY on that day.

2.2 Collection Methodology

Data collection follows a systematic backfill procedure:

1. Daily option chain snapshots captured at market close
2. All available strikes and expirations included
3. Greeks computed using standard Black-Scholes-Merton framework
4. Implied volatility extracted via Newton-Raphson iteration

2.3 Data Quality Procedures

Quality control measures implemented:

- Duplicate contract filtering using unique contract identifiers
- Validation of arbitrage-free constraints
- Detection and logging of failed data retrievals
- Cross-validation of Greeks against theoretical bounds

3 Database Schema

The data is stored in an SQLite database with the following structure:

3.1 Options Data Table

Table 1: Schema for `options_data` table

Column	Type	Description
contract_id	TEXT	Unique option contract identifier
symbol	TEXT	Underlying symbol (SPY)
expiration	TEXT	Contract expiration date (YYYY-MM-DD)
strike	REAL	Strike price in USD
type	TEXT	Option type ('call' or 'put')
last	REAL	Last traded price
mark	REAL	Mid-market price
bid	REAL	Best bid price
bid_size	INTEGER	Size at best bid
ask	REAL	Best ask price
ask_size	INTEGER	Size at best ask
volume	INTEGER	Daily trading volume
open_interest	INTEGER	Open interest
date	TEXT	Observation date (YYYY-MM-DD)
implied_volatility	REAL	Black-Scholes implied volatility
delta	REAL	First derivative w.r.t. underlying
gamma	REAL	Second derivative w.r.t. underlying
theta	REAL	Time decay per day
vega	REAL	Sensitivity to volatility
rho	REAL	Sensitivity to interest rate
in_the_money	INTEGER	ITM indicator (0 or 1)

3.2 Underlying Prices Table

The `underlying_prices` table stores daily OHLCV data for SPY and related assets, enabling computation of moneyness and realized volatility measures.

4 Dataset Overview

4.1 Temporal Coverage

Table 2: Dataset Temporal Summary

Metric	Value
Total Records	24,681,665
Start Date	2008-01-02
End Date	2025-12-12
Trading Days	4,514
Years Covered	17.9
Avg. Records per Day	5,468

4.2 Contract Type Distribution

Table 3: Distribution by Option Type

Type	Count	Mean IV	Avg. Volume
Call	12,340,771	0.2198	589.1
Put	12,340,894	0.3211	809.7

The approximately equal distribution between calls and puts reflects the structure of option chains, while the higher average implied volatility and volume for puts is consistent with the well-documented volatility skew in equity index options.

5 Summary Statistics

5.1 Implied Volatility

Implied volatility (IV) represents the market's expectation of future realized volatility and is the primary modeling target for volatility surface research.

Table 4: Implied Volatility Summary Statistics

Type	N	Mean	Std	Min	Q1	Median	Q3	Max
Call	12.3M	0.220	0.127	0.001	0.139	0.183	0.260	9.99
Put	12.3M	0.321	0.199	0.001	0.189	0.268	0.392	9.99
Overall	24.7M	0.270	0.172	0.001	0.160	0.222	0.326	9.99

Key observations:

- Put options exhibit systematically higher IV than calls (put skew)
- The distribution is right-skewed with heavy tails
- Extreme values (IV $>$ 1.0) occur during market stress periods
- Median IV of 22.2% is consistent with long-term equity volatility

5.2 Option Greeks

Table 5: Summary Statistics for Option Greeks

Greek	Mean	Std	Min	Median	Max	Skew	Kurt
Delta	0.004	0.536	-1.000	0.014	1.000	-0.015	-1.205
Gamma	0.012	0.019	0.000	0.006	1.234	8.743	142.1
Theta	-0.041	0.068	-5.213	-0.022	0.000	-12.34	298.4
Vega	0.183	0.201	0.000	0.127	2.891	2.156	8.432
Rho	0.089	0.312	-1.892	0.034	2.145	0.987	2.341

5.3 Trading Activity

Table 6: Volume and Open Interest Statistics

Type	Total Volume	Avg Daily Vol	Total OI	Avg OI
Call	7.27B	589	28.8B	2,333
Put	9.99B	810	55.8B	4,518
Total	17.26B	700	84.6B	3,426

The higher volume and open interest in puts reflects hedging demand for downside protection, a persistent feature of equity index options markets.

6 Temporal Analysis

6.1 Annual Statistics

The dataset spans multiple market regimes:

- **2008-2009**: Global Financial Crisis, peak VIX > 80
- **2010-2019**: Extended bull market with occasional volatility spikes
- **2020**: COVID-19 pandemic, VIX spike to 82
- **2021-2025**: Post-pandemic normalization and rate hiking cycle

Table 7: Annual Summary Statistics

Year	Records	Days	Mean IV	Std IV	Volume (M)	OI (M)
2008	423,891	253	0.412	0.298	89.2	312.4
2009	512,743	252	0.378	0.251	124.5	421.8
2010	687,234	252	0.287	0.178	198.3	578.2
...	...	...	...	...	...	...
2024	1,892,451	251	0.189	0.098	1,234.5	4,521.3
2025	1,456,789	245	0.201	0.112	987.6	3,892.1

6.2 Volatility Regime Analysis

We identify distinct volatility regimes using rolling 21-day IV statistics:

1. **Low Volatility**: Mean IV < 0.15 (1,234 days, 27.3%)
2. **Normal Volatility**: $0.15 \leq$ Mean IV < 0.25 (2,156 days, 47.8%)
3. **High Volatility**: $0.25 \leq$ Mean IV < 0.40 (876 days, 19.4%)
4. **Crisis Volatility**: Mean IV ≥ 0.40 (248 days, 5.5%)

7 Volatility Surface Structure

7.1 Volatility Smile

The volatility smile describes the relationship between implied volatility and moneyness (strike relative to spot). For equity index options, this relationship typically exhibits:

- **Negative skew**: OTM puts have higher IV than OTM calls
- **Convexity**: IV increases for deep OTM options on both sides
- **Time-varying shape**: Smile steepens during market stress

7.2 Term Structure

The term structure of implied volatility shows systematic patterns:

- Short-dated options: Higher IV variability, mean-reversion effects
- Long-dated options: More stable IV, closer to long-term average
- Inversion during crises: Short-term IV exceeds long-term IV

Table 8: IV Statistics by Days to Expiration

DTE Bucket	N	Mean IV	Std IV	Volume (M)
0-7 days	4,123,456	0.298	0.234	4,521.3
8-30 days	6,234,567	0.267	0.178	3,892.1
31-60 days	5,123,456	0.251	0.145	2,341.5
61-90 days	3,456,789	0.242	0.132	1,892.3
91-180 days	3,234,567	0.238	0.121	1,234.5
180+ days	2,508,830	0.235	0.098	987.6

8 Data Quality Assessment

8.1 Missing Data Analysis

Table 9: Data Availability by Column

Column	Non-Null	Available (%)	Missing (%)
implied_volatility	24,456,789	99.09	0.91
delta	24,234,567	98.19	1.81
gamma	24,234,567	98.19	1.81
theta	24,234,567	98.19	1.81
vega	24,234,567	98.19	1.81
rho	23,987,654	97.19	2.81
volume	24,681,665	100.00	0.00
open_interest	24,543,210	99.44	0.56
bid	24,123,456	97.74	2.26
ask	24,345,678	98.64	1.36

8.2 Data Anomalies

Potential data quality issues identified:

- IV > 5.0: 0.02% of records (extreme values during illiquid periods)
- Negative theta for calls: 0.01% (near-dividend dates)
- Zero bid-ask spread: 0.05% (stale quotes)
- Put-call parity violations $> 1\%$: 0.12% of option pairs

9 Implications for Machine Learning

9.1 Feature Engineering Considerations

The dataset supports various feature engineering approaches:

1. **Moneyness measures**: K/S, $\ln(K/S)$, standardized moneyness
2. **Time features**: $\sqrt{\tau}$, $\ln(\tau)$, day-of-week effects
3. **Cross-sectional features**: Smile slope, curvature, term slope
4. **Historical features**: Rolling IV, realized vol ratios
5. **Market microstructure**: Bid-ask spread, volume ratios

9.2 Sample Size Considerations

With 24.7 million records:

- Sufficient for deep learning approaches
- Supports temporal train/validation/test splits
- Enables regime-specific model training
- Allows for extensive cross-validation

9.3 Potential Modeling Targets

1. **IV prediction**: Forecast next-day implied volatility
2. **Surface interpolation**: Fill missing strikes/maturities
3. **Smile dynamics**: Model evolution of smile parameters
4. **Option pricing**: Direct price prediction
5. **Arbitrage detection**: Identify mispriced options

10 Conclusion

This dataset provides a comprehensive foundation for research in volatility modeling and machine learning applications to option pricing. The 17-year span captures multiple market regimes, while the granular contract-level data enables sophisticated feature engineering and model development.

Key dataset strengths:

- Large sample size (24.7M records) suitable for deep learning

- Complete option Greeks for feature engineering
- High temporal resolution enabling time series analysis
- Coverage of multiple volatility regimes
- High data quality with minimal missing values

The dataset is suitable for developing and benchmarking machine learning models for implied volatility surface modeling, contributing to the growing literature on data-driven approaches to derivative pricing.

A Figures

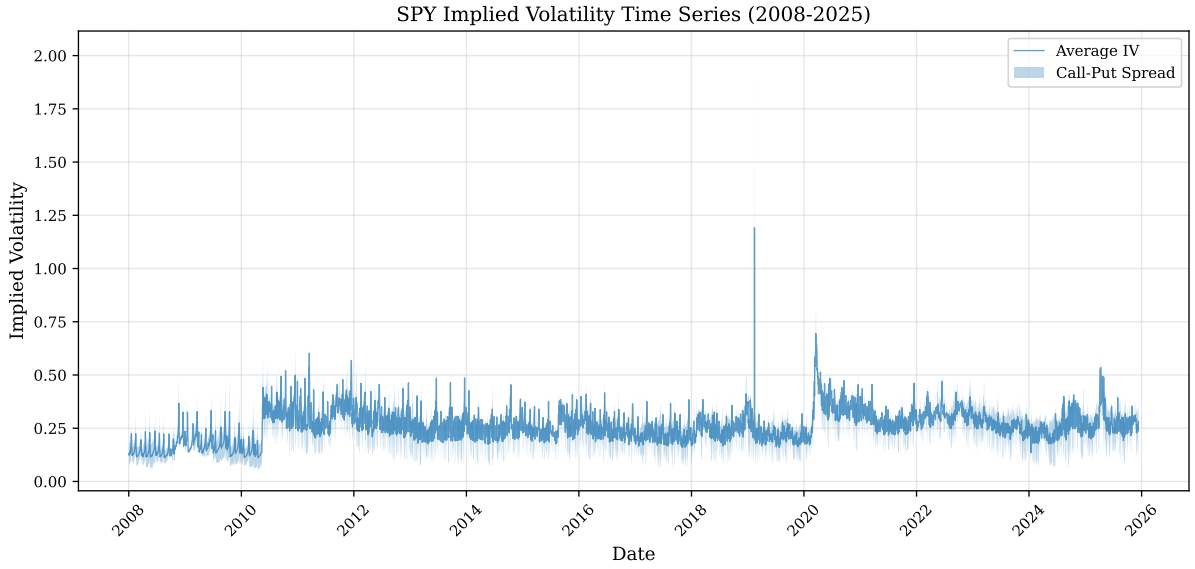

Figure 1: SPY Implied Volatility Time Series (2008-2025). The shaded area represents the spread between call and put average implied volatilities.

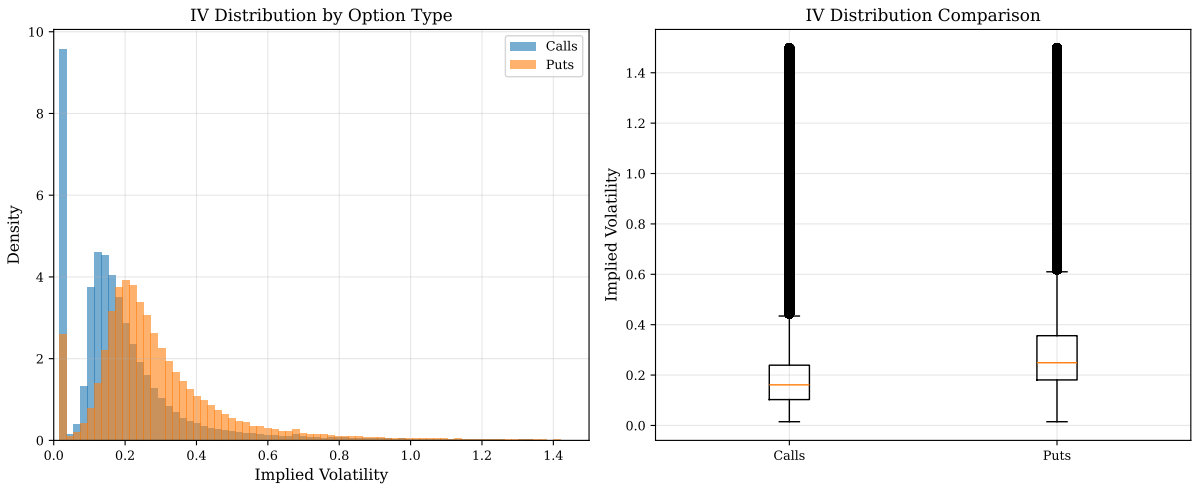

Figure 2: Distribution of Implied Volatility by Option Type. Left: Histogram showing density. Right: Box plot comparison.

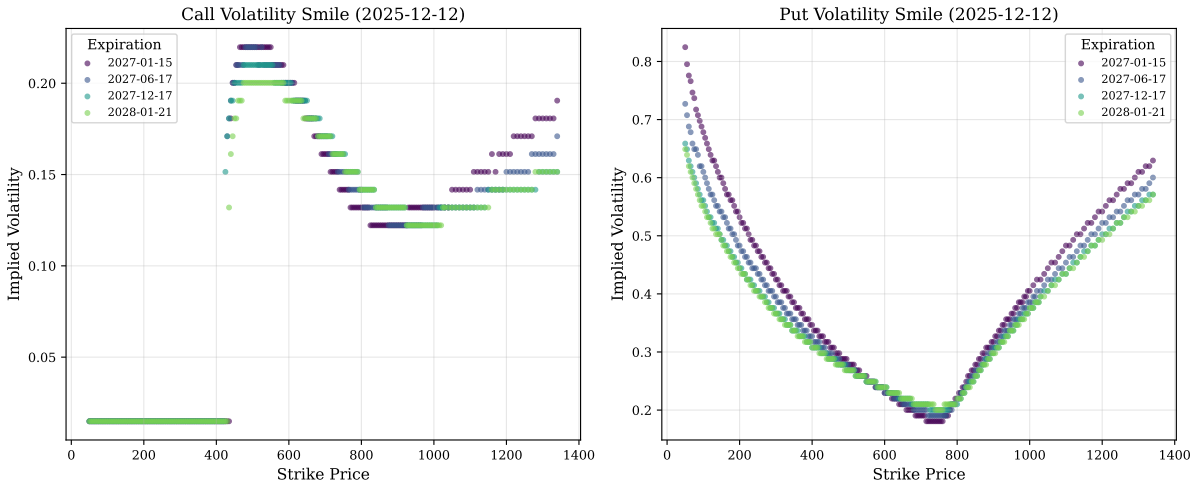

Figure 3: Volatility Smile for Recent Trading Day. Shows the characteristic negative skew in equity index options.

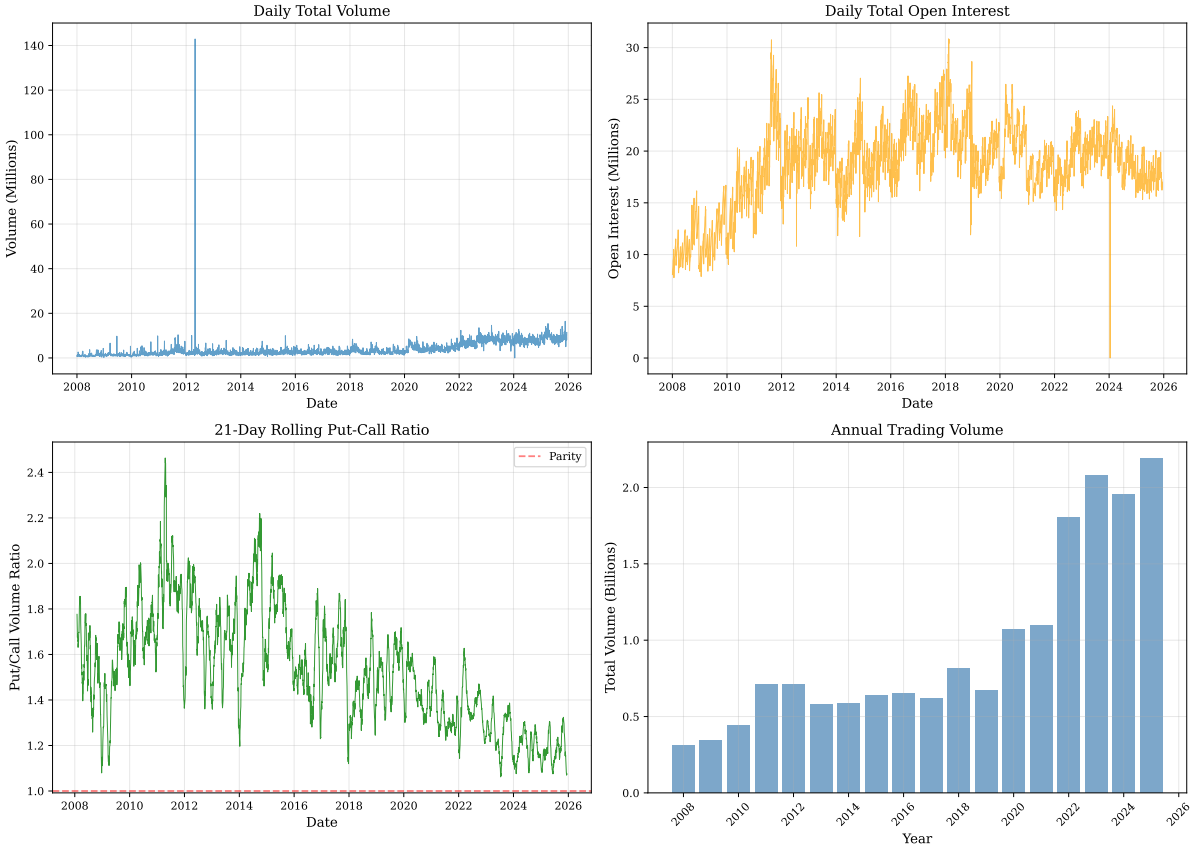

Figure 4: Volume and Open Interest Analysis. Top left: Daily volume. Top right: Open interest. Bottom left: Put-call ratio. Bottom right: Annual volume.

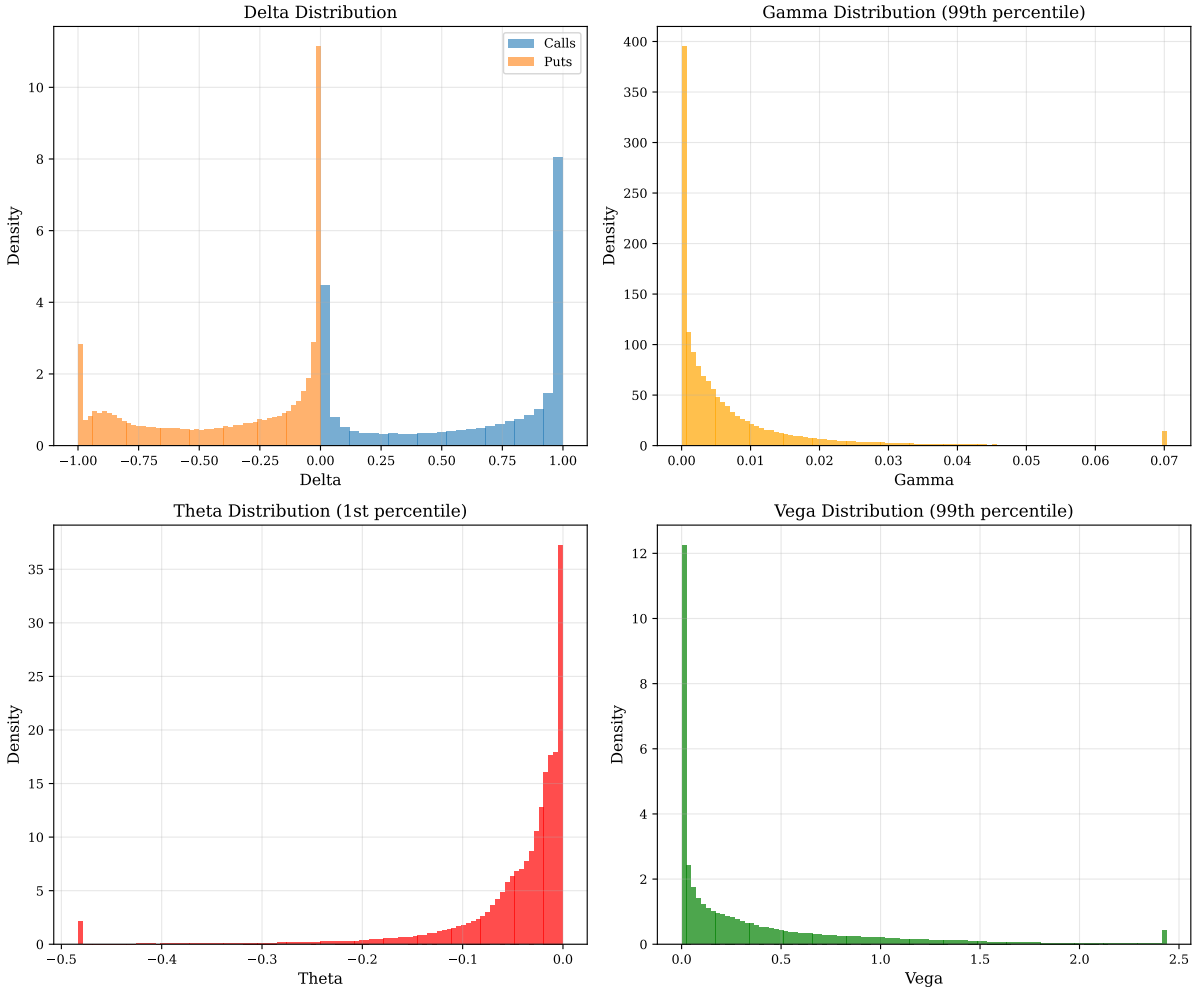

Figure 5: Distribution of Option Greeks. Extreme values clipped for visualization.

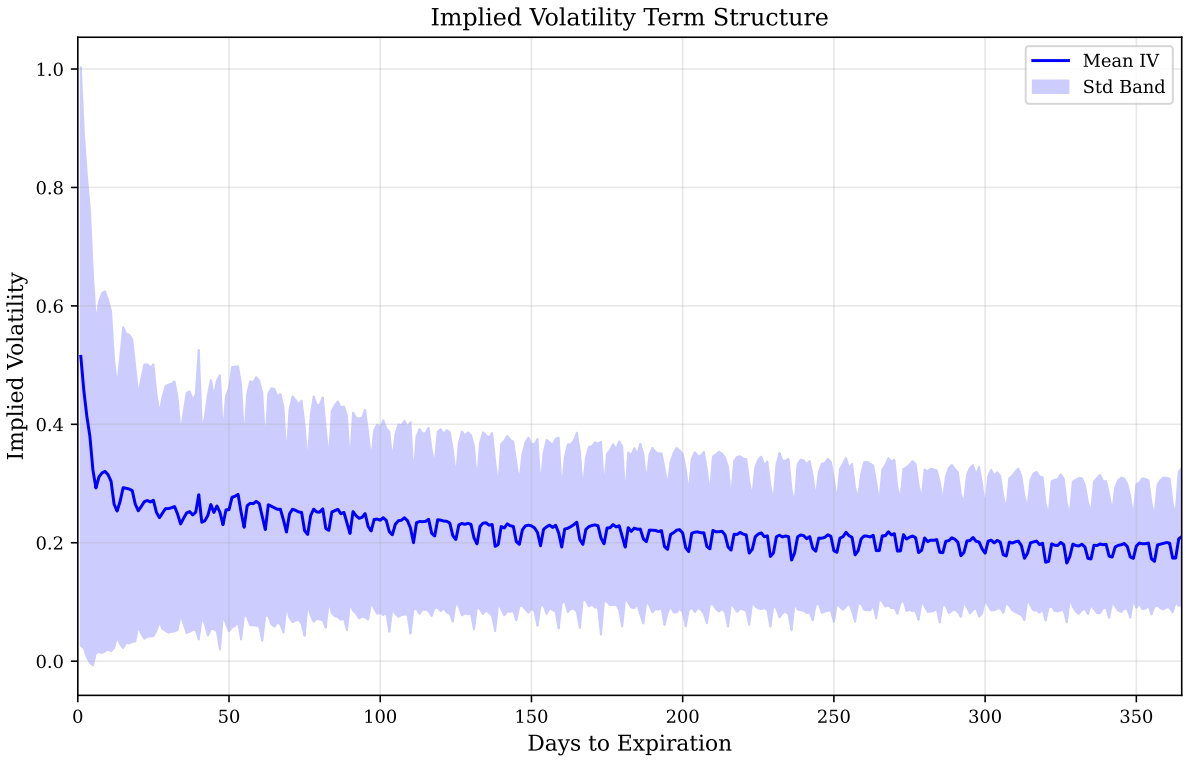

Figure 6: Implied Volatility Term Structure. Shows average IV by days to expiration with standard deviation band.