Historic Options Dataset: Comprehensive Description and Summary Statistics

SPY, IWM, and QQQ Options Data for Machine Learning-Based Volatility Modeling

Philipp Dubach
github.com/philippdubach

December 16, 2025

Abstract

This document provides a comprehensive description of a large-scale dataset comprising options data for three major U.S. equity ETFs: SPY (SPDR S&P 500 ETF Trust), IWM (iShares Russell 2000 ETF), and QQQ (Invesco QQQ Trust, Nasdaq-100). The combined dataset contains over 53 million option contracts spanning from January 2008 to December 2025, 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 options on three major U.S. equity ETFs 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
- Cross-asset volatility correlation studies

The dataset includes:

- **SPY**: S&P 500 ETF Trust — Large-cap U.S. equities
- **IWM**: iShares Russell 2000 ETF — Small-cap U.S. equities
- **QQQ**: Invesco QQQ Trust — Nasdaq-100 technology-focused equities

These three ETFs together provide exposure to different segments of the U.S. equity market, enabling comparative analysis of volatility behavior across market capitalizations and sector exposures.

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 each underlying 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 SQLite databases (one per underlying) with identical schemas.

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, IWM, or QQQ)
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 each underlying, enabling computation of moneyness and realized volatility measures.

4 Dataset Overview

4.1 Available Datasets

Table 2: Dataset Summary

Symbol	Description	Records	Trading Days	Period
SPY	SPDR S&P 500 ETF Trust	24,681,665	4,514	2008–2025
IWM	iShares Russell 2000 ETF	13,379,573	4,509	2008–2025
QQQ	Invesco QQQ Trust (Nasdaq-100)	15,345,882	3,700	2011–2025
Total		**53,407,120**		

4.2 Market Regimes Covered

The dataset spans multiple distinct 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

5 SPY Dataset

5.1 Overview

SPY (SPDR S&P 500 ETF Trust) tracks the S&P 500 Index and is one of the most liquid ETFs globally. SPY options are among the most actively traded equity options.

Table 3: SPY Dataset 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
Calls	12,340,771
Puts	12,340,894
Avg. Records per Day	5,467

5.2 Implied Volatility Statistics

Table 4: SPY Implied Volatility Summary Statistics

Type	N	Mean	Std	Min	Q1	Median	Q3	Max
Call	12.3M	0.220	0.320	0.015	0.103	0.161	0.239	9.995
Put	12.3M	0.321	0.319	0.015	0.181	0.249	0.366	9.995

5.3 Volume and Open Interest

Table 5: SPY Volume and Open Interest Statistics

Type	Total Volume	Avg Volume	Total OI	Avg OI
Call	7.27B	589	28.8B	2,333
Put	9.99B	810	55.8B	4,518

5.4 Greeks Statistics

Table 6: SPY Option Greeks Summary

Greek	Mean	Std	Min	Median	Max
Delta	0.121	0.620	-1.000	0.000	1.000
Gamma	0.008	0.024	0.000	0.003	5.251
Theta	-0.054	0.134	-24.719	-0.025	3.393
Vega	0.376	0.532	0.000	0.155	4.360
Rho	-0.028	1.568	-166.894	0.000	11.352

5.5 SPY Figures

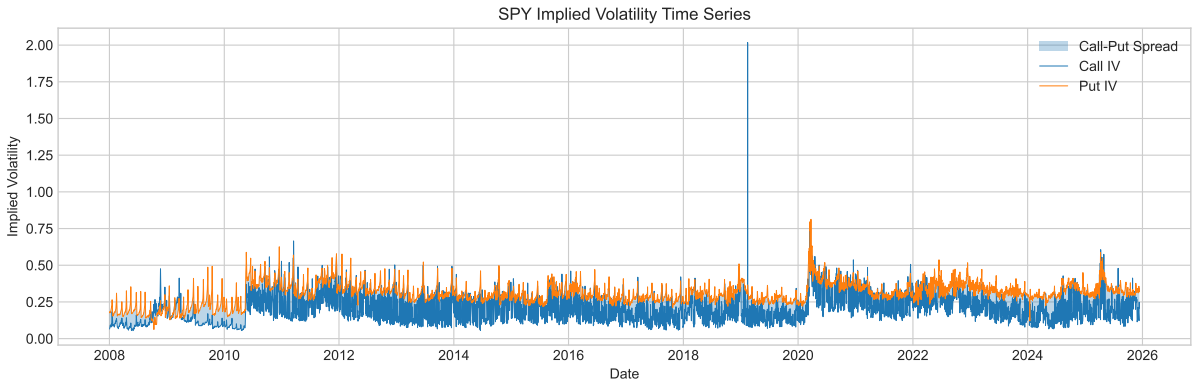


Figure 1: SPY Implied Volatility Time Series (2008-2025)

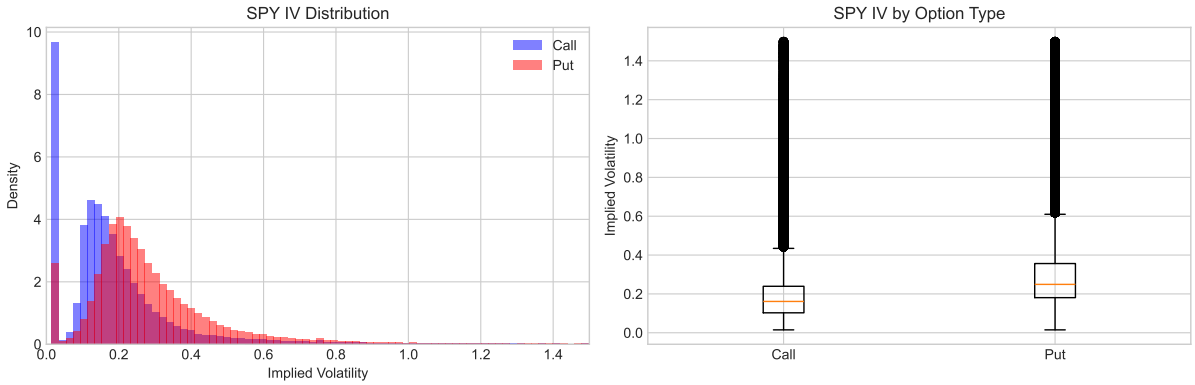


Figure 2: SPY Implied Volatility Distribution by Option Type

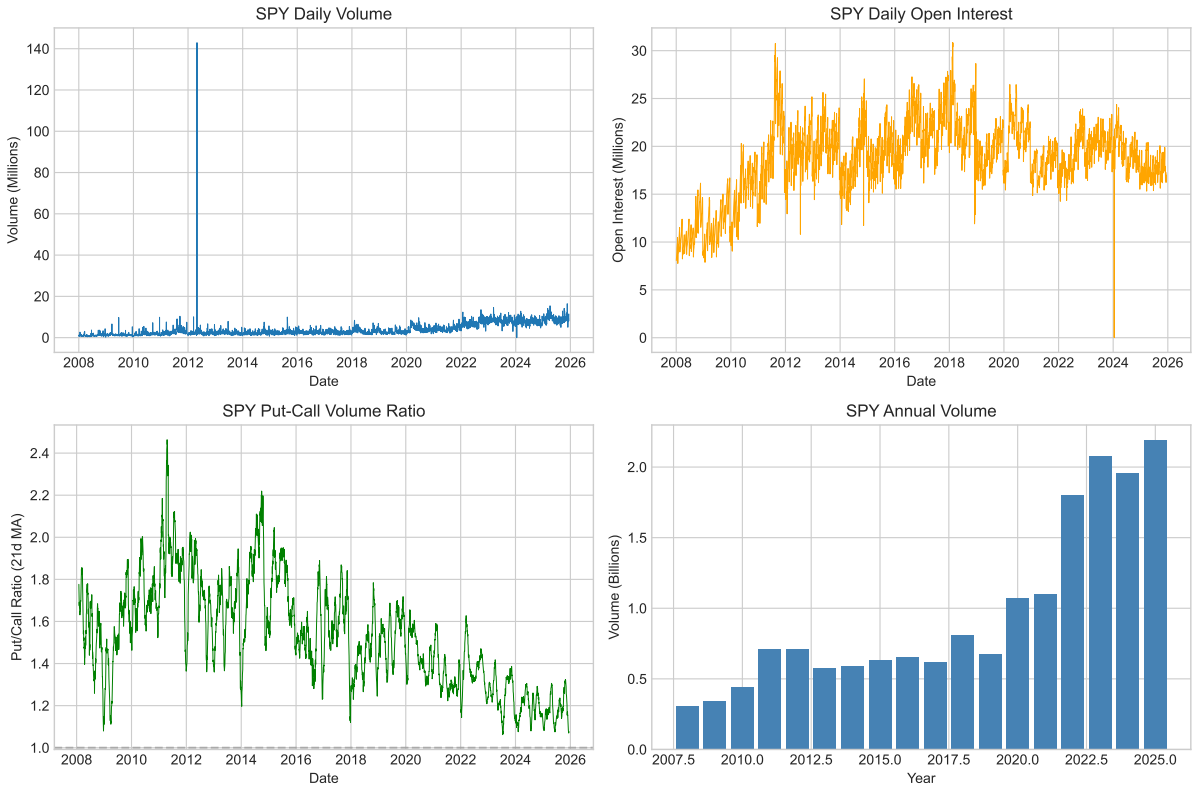


Figure 3: SPY Volume and Open Interest Analysis

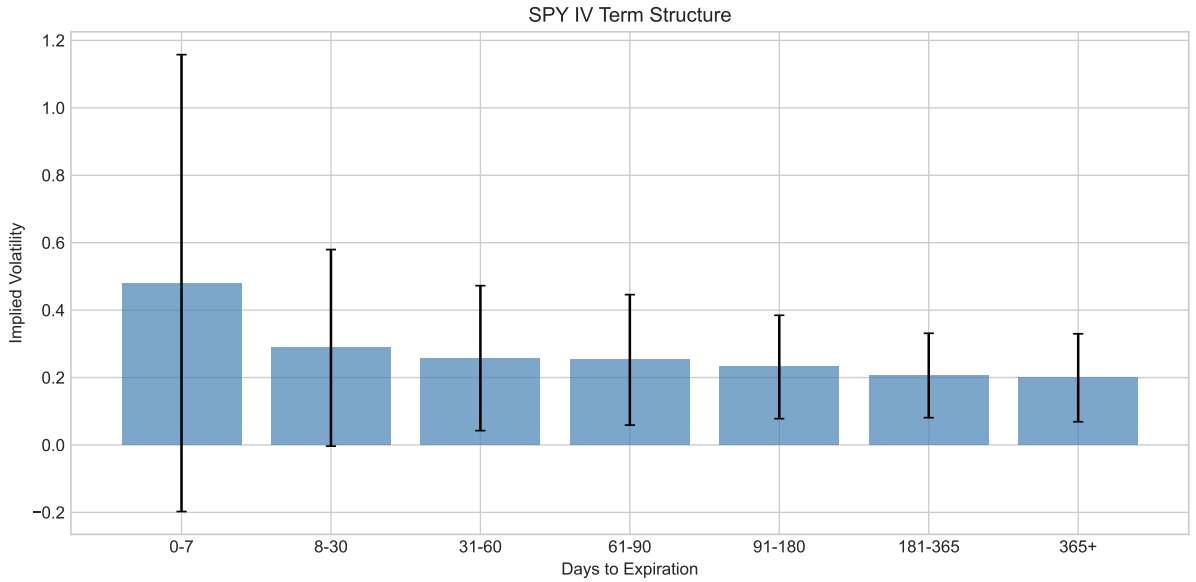


Figure 4: SPY Implied Volatility Term Structure

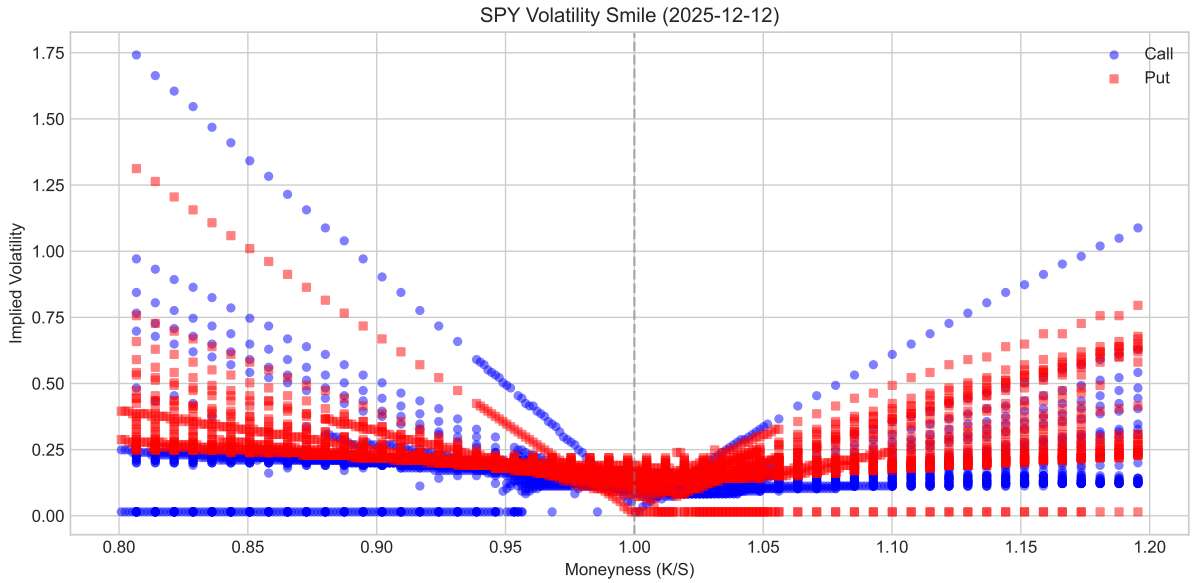
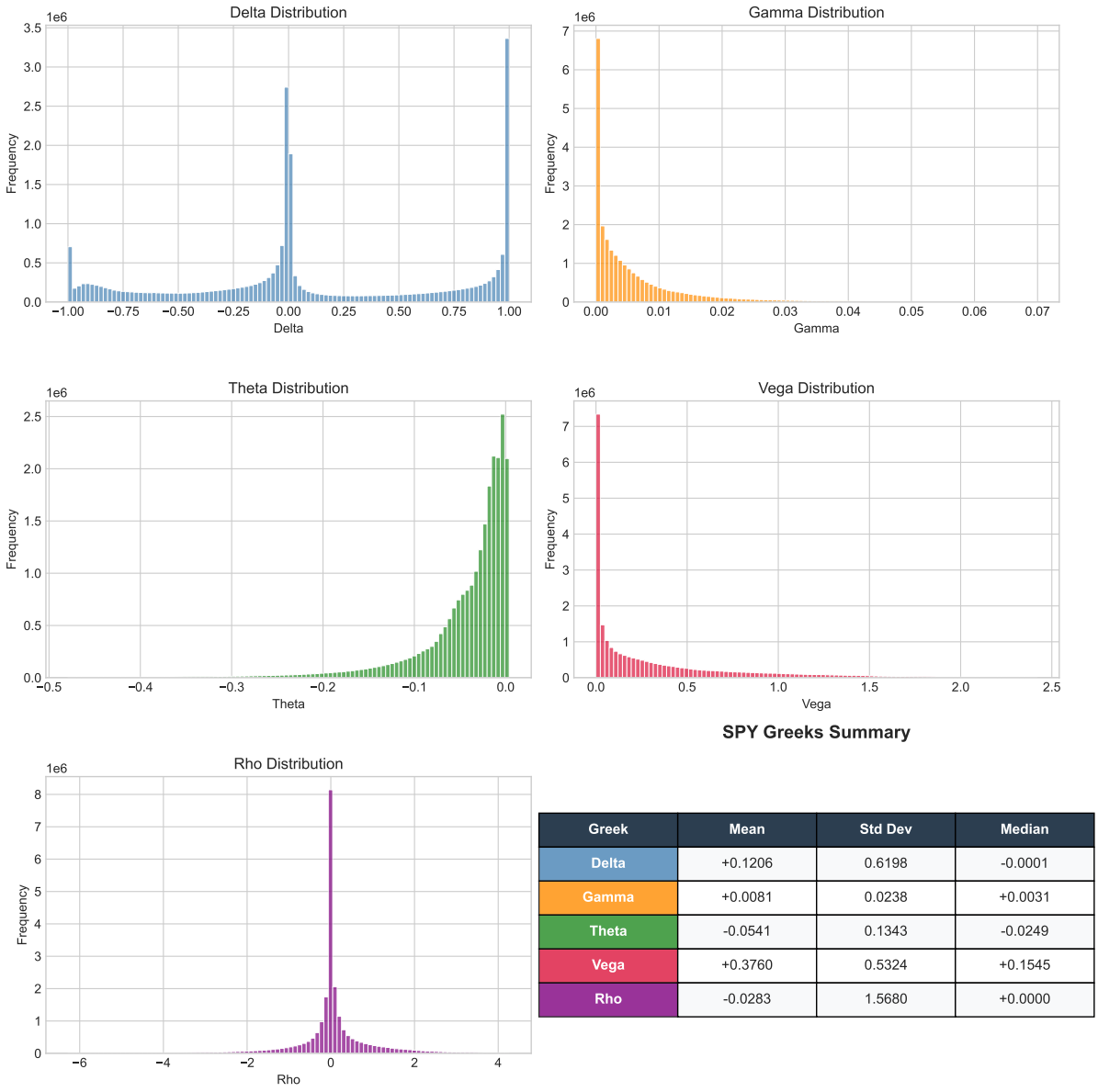


Figure 5: SPY Volatility Smile (Recent Trading Day)



SPY Greeks Summary

Greek	Mean	Std Dev	Median
Delta	+0.1206	0.6198	-0.0001
Gamma	+0.0081	0.0238	+0.0031
Theta	-0.0541	0.1343	-0.0249
Vega	+0.3760	0.5324	+0.1545
Rho	-0.0283	1.5680	+0.0000

Figure 6: SPY Option Greeks Distribution

6 IWM Dataset

6.1 Overview

IWM (iShares Russell 2000 ETF) tracks the Russell 2000 Index, providing exposure to small-cap U.S. equities. Small-cap stocks typically exhibit higher volatility than large-caps.

Table 7: IWM Dataset Summary

Metric	Value
Total Records	13,379,573
Start Date	2008-01-02
End Date	2025-12-15
Trading Days	4,509
Years Covered	17.9
Calls	6,689,698
Puts	6,689,875
Avg. Records per Day	2,967

6.2 Implied Volatility Statistics

Table 8: IWM Implied Volatility Summary Statistics

Type	N	Mean	Std	Min	Q1	Median	Q3	Max
Call	6.7M	0.255	0.331	0.015	0.130	0.195	0.283	9.995
Put	6.7M	0.345	0.320	0.015	0.206	0.279	0.395	9.995

6.3 Volume and Open Interest

Table 9: IWM Volume and Open Interest Statistics

Type	Total Volume	Avg Volume	Total OI	Avg OI
Call	1.08B	162	9.57B	1,430
Put	2.02B	301	21.28B	3,181

6.4 Greeks Statistics

Table 10: IWM Option Greeks Summary

Greek	Mean	Std	Min	Median	Max
Delta	0.122	0.619	-1.000	0.000	1.000
Gamma	0.020	0.052	0.000	0.009	5.251
Theta	-0.022	0.054	-5.820	-0.011	3.393
Vega	0.145	0.199	0.000	0.063	2.076
Rho	-0.003	0.283	-13.152	0.000	2.076

6.5 IWM Figures

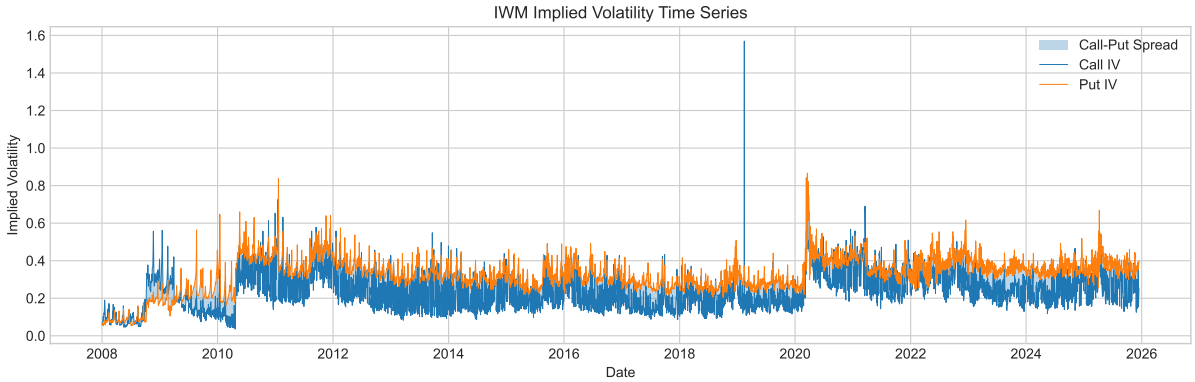


Figure 7: IWM Implied Volatility Time Series (2008-2025)

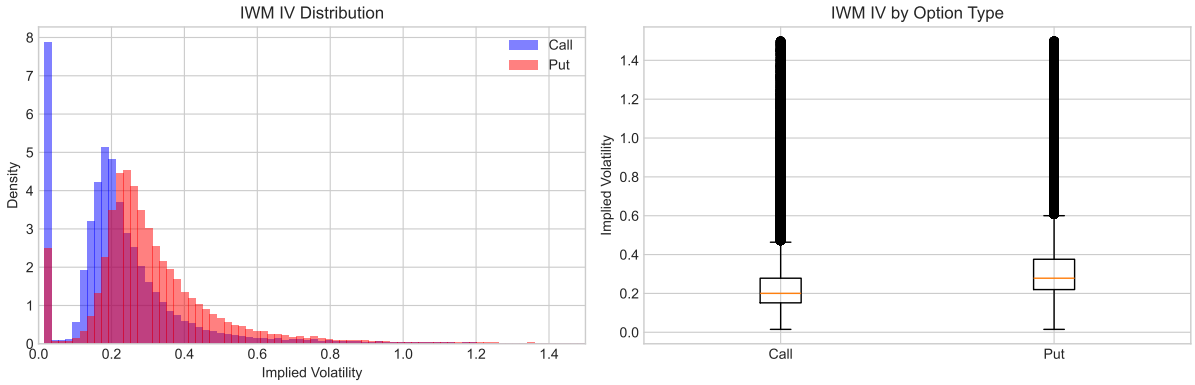


Figure 8: IWM Implied Volatility Distribution by Option Type

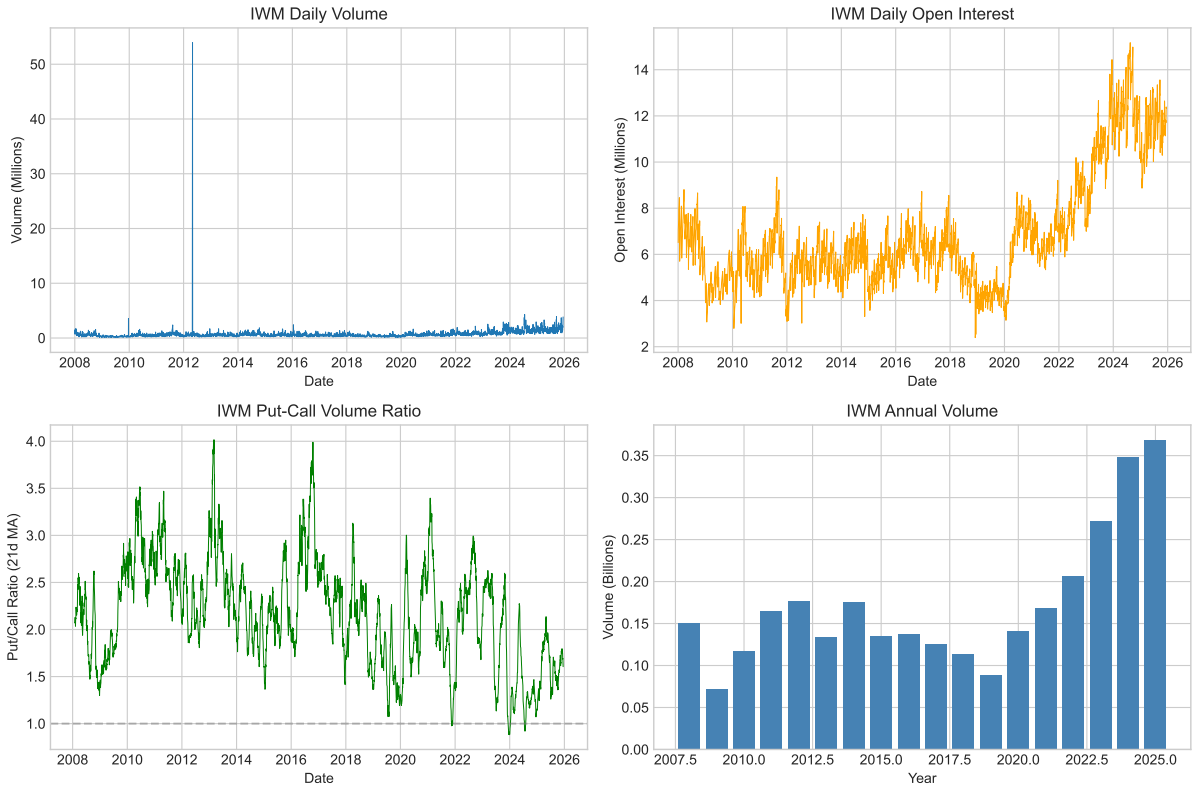


Figure 9: IWM Volume and Open Interest Analysis

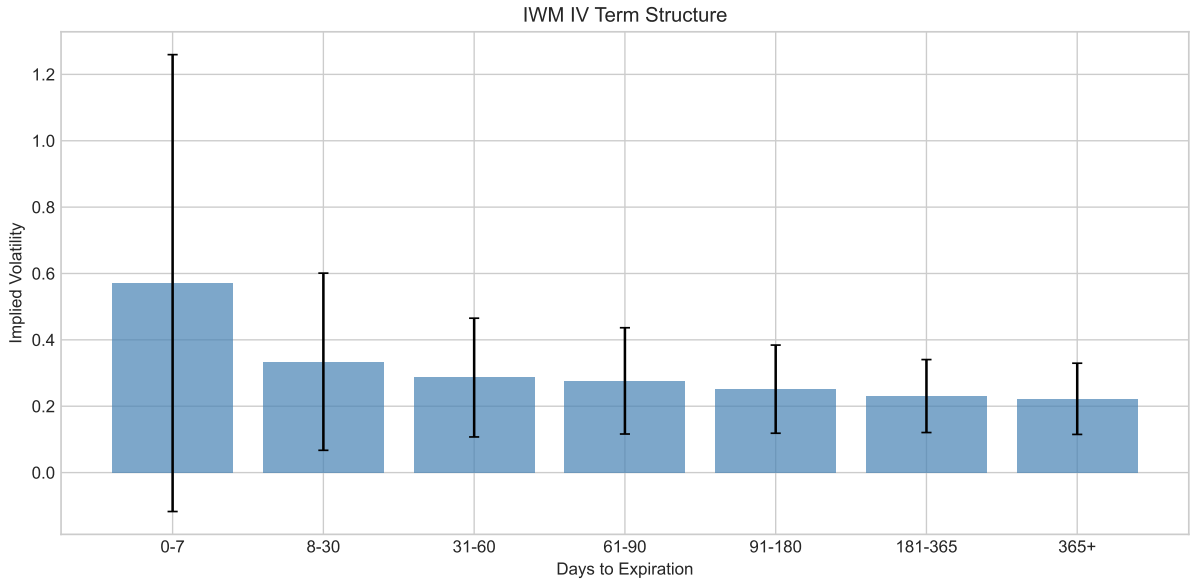


Figure 10: IWM Implied Volatility Term Structure

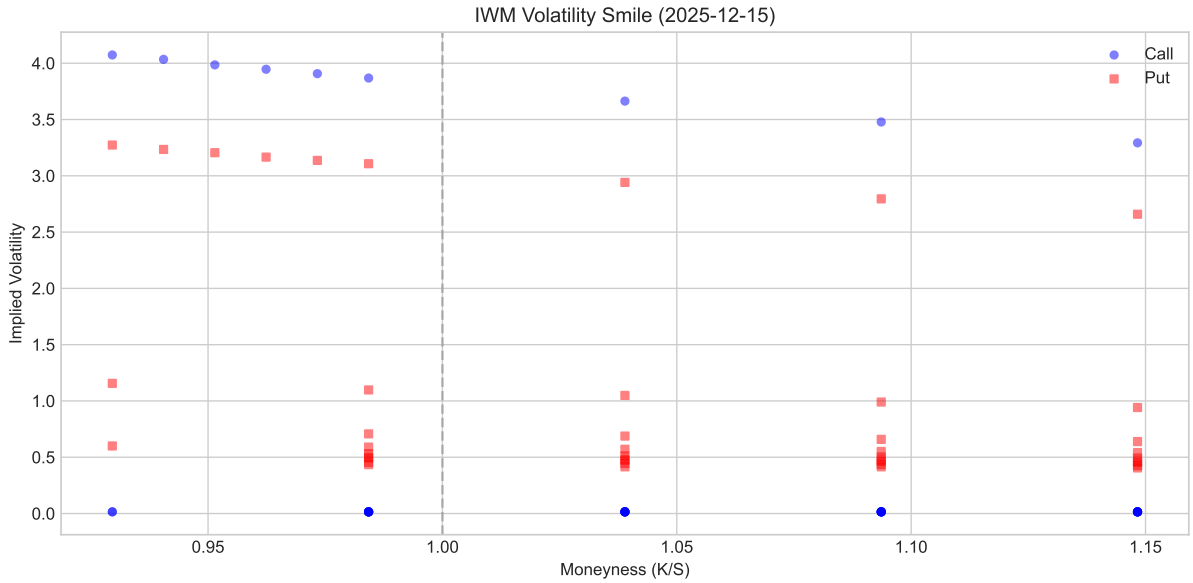
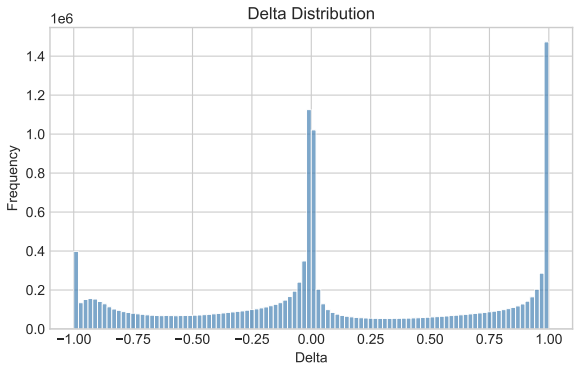
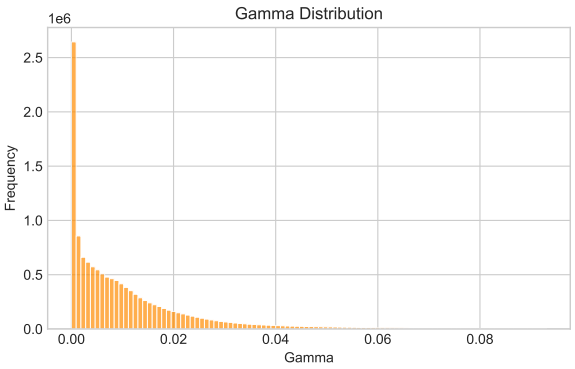
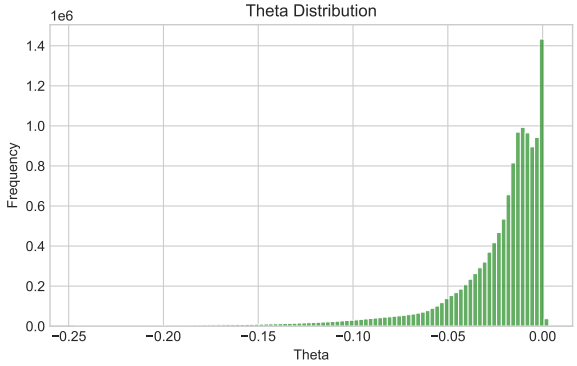
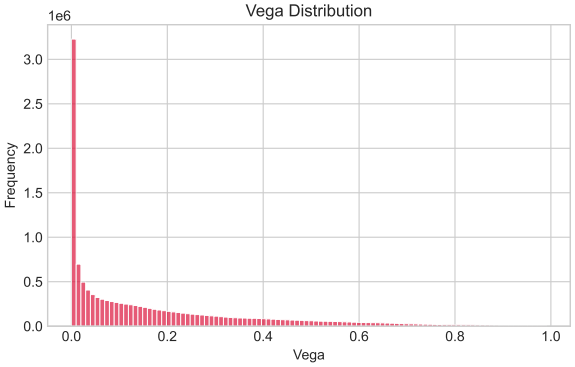


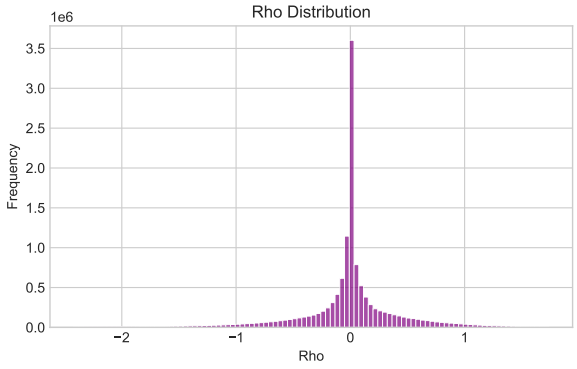
Figure 11: IWM Volatility Smile (Recent Trading Day)











IWM Greeks Summary

Greek	Mean	Std Dev	Median
Delta	+0.0798	0.6197	-0.0004
Gamma	+0.0133	0.0355	+0.0071
Theta	-0.0299	0.0652	-0.0152
Vega	+0.1883	0.2303	+0.1009
Rho	-0.0248	0.6368	+0.0000

Figure 12: IWM Option Greeks Distribution

7 QQQ Dataset

7.1 Overview

QQQ (Invesco QQQ Trust) tracks the Nasdaq-100 Index, providing exposure to the largest non-financial companies listed on the Nasdaq, with heavy weighting toward technology stocks.

Table 11: QQQ Dataset Summary

Metric	Value
Total Records	15,345,882
Start Date	2011-03-23
End Date	2025-12-15
Trading Days	3,700
Years Covered	14.7
Calls	7,672,920
Puts	7,672,962
Avg. Records per Day	4,148

7.2 Implied Volatility Statistics

Table 12: QQQ Implied Volatility Summary Statistics

Type	N	Mean	Std	Min	Q1	Median	Q3	Max
Call	7.7M	0.267	0.345	0.015	0.121	0.188	0.299	9.995
Put	7.7M	0.342	0.317	0.015	0.187	0.271	0.403	9.995

7.3 Volume and Open Interest

Table 13: QQQ Volume and Open Interest Statistics

Type	Total Volume	Avg Volume	Total OI	Avg OI
Call	3.59B	468	13.93B	1,816
Put	4.09B	533	22.70B	2,960

7.4 Greeks Statistics

Table 14: QQQ Option Greeks Summary

Greek	Mean	Std	Min	Median	Max
Delta	0.119	0.619	-1.000	0.000	1.000
Gamma	0.006	0.020	0.000	0.002	5.251
Theta	-0.080	0.185	-27.181	-0.035	3.393
Vega	0.477	0.675	0.000	0.182	5.102
Rho	-0.064	2.236	-179.893	0.000	18.125

7.5 QQQ Figures

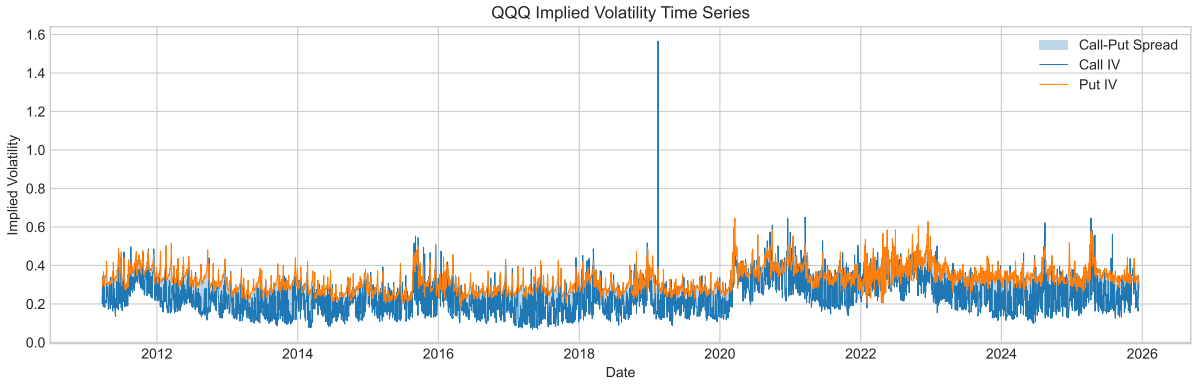


Figure 13: QQQ Implied Volatility Time Series (2011-2025)

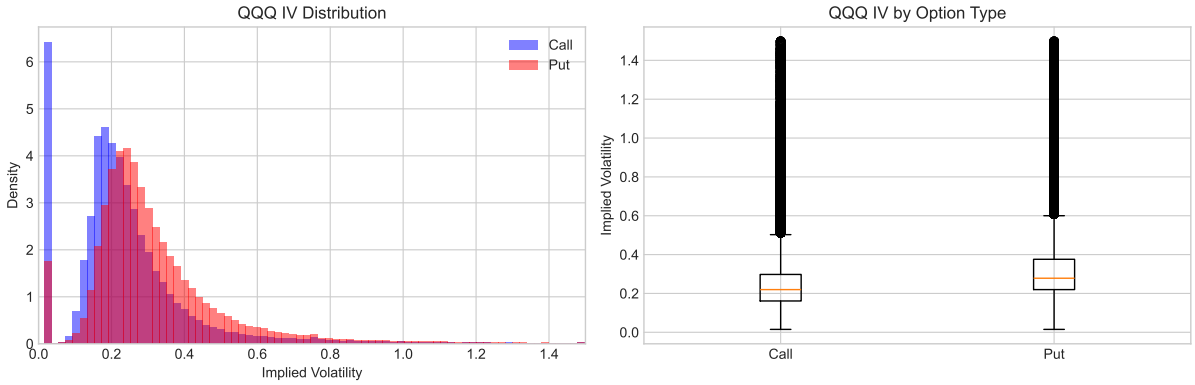


Figure 14: QQQ Implied Volatility Distribution by Option Type

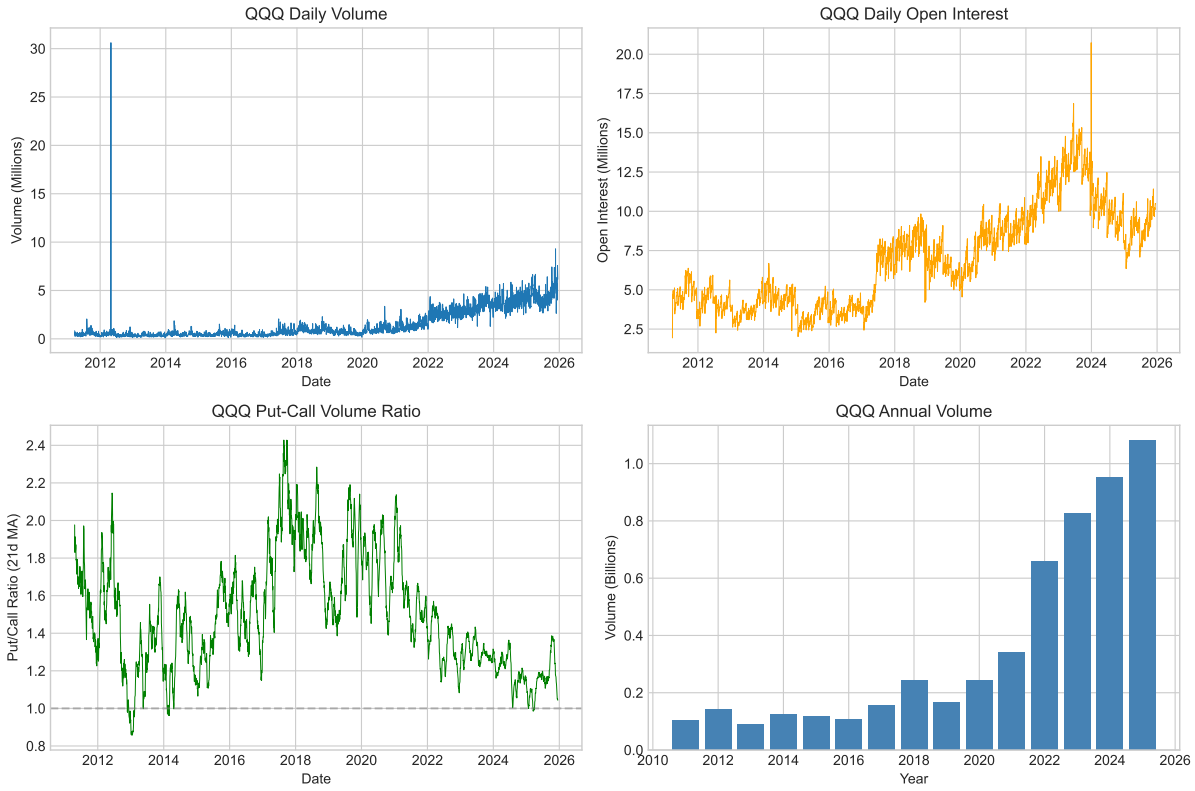


Figure 15: QQQ Volume and Open Interest Analysis

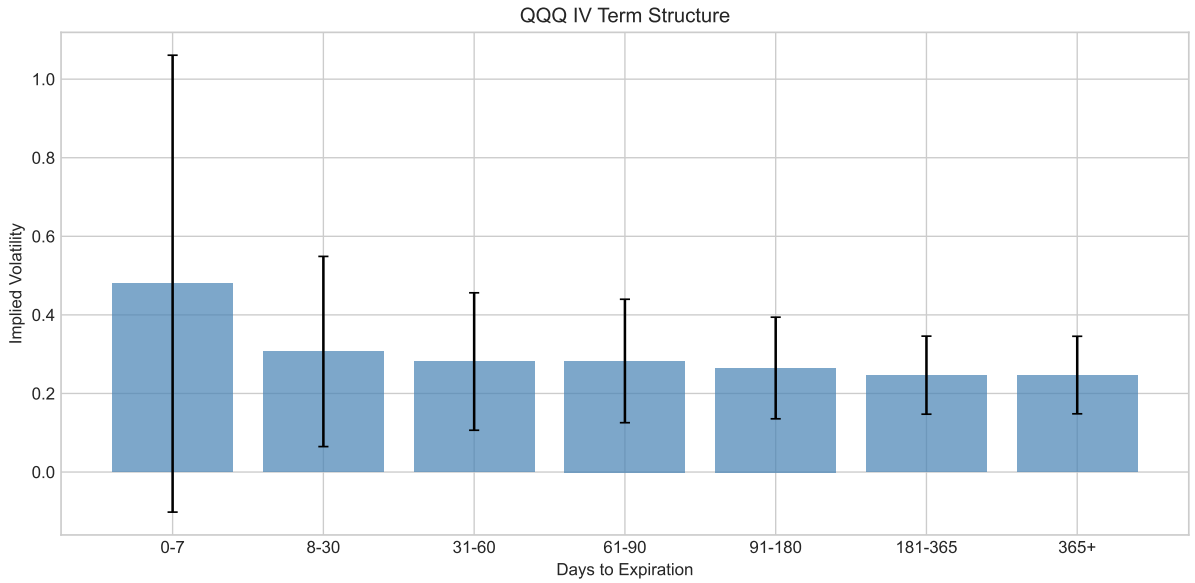


Figure 16: QQQ Implied Volatility Term Structure

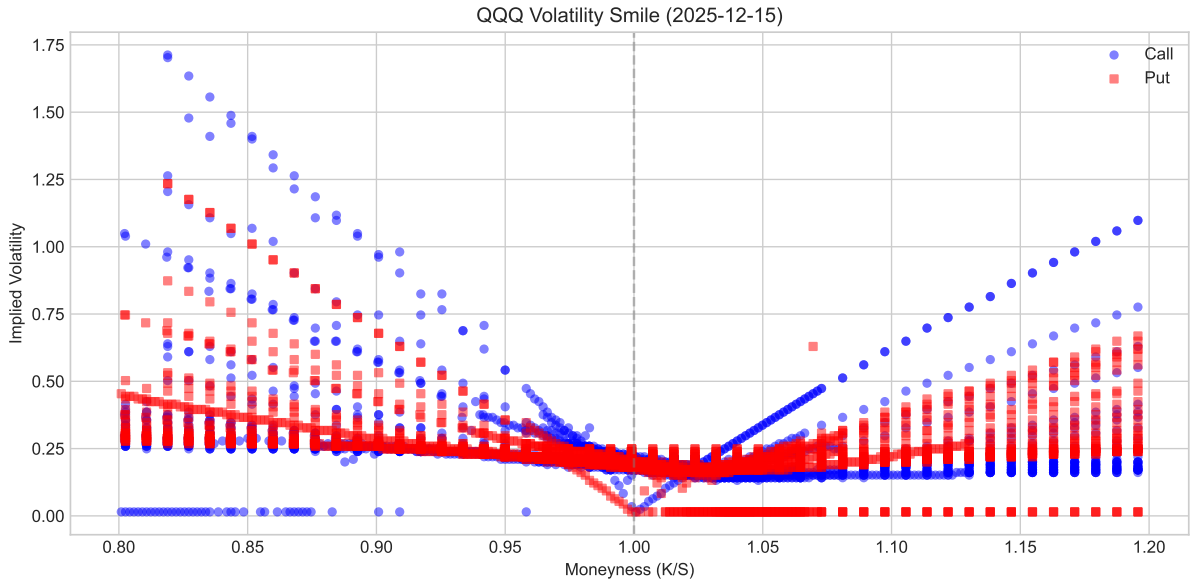
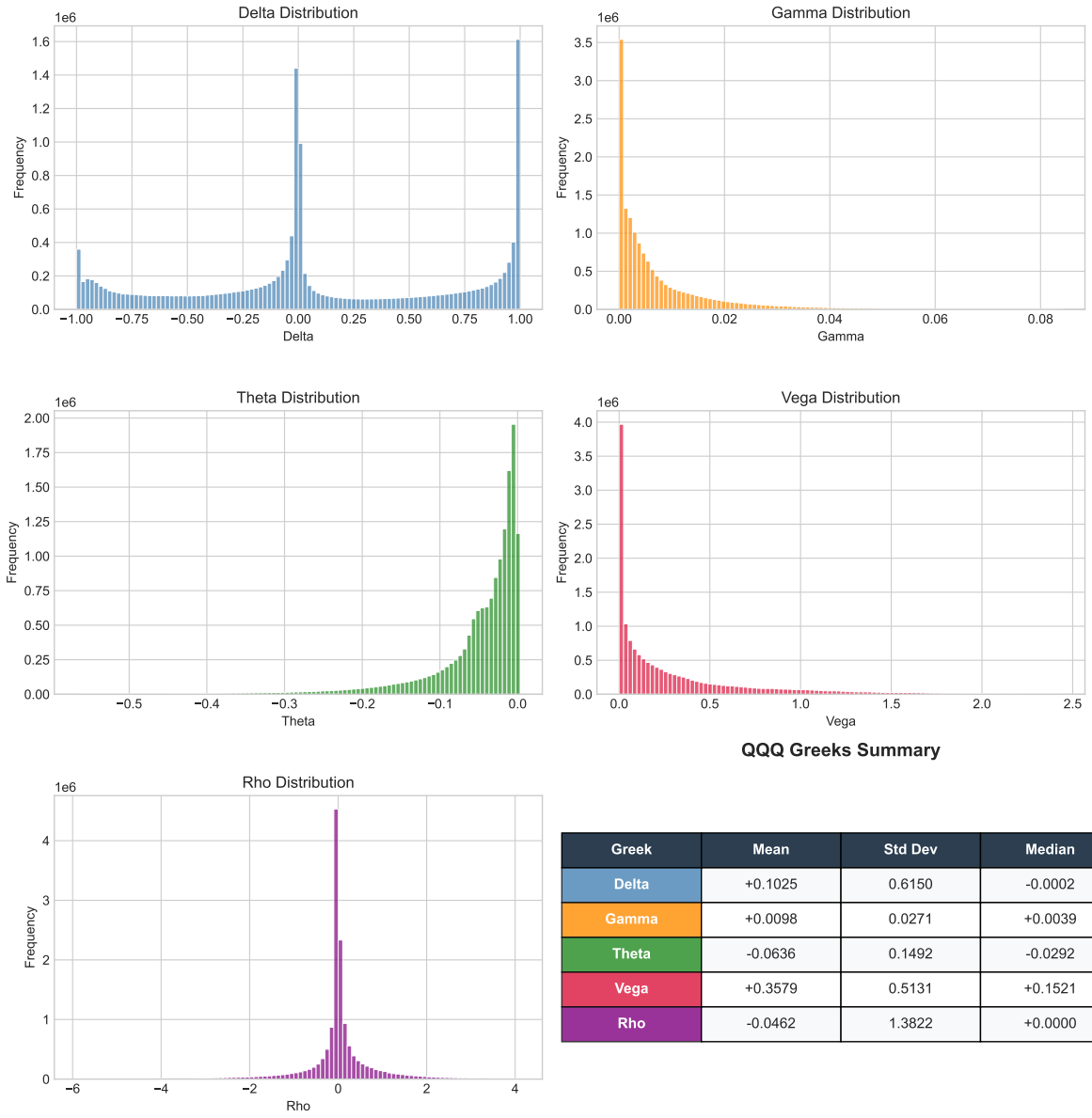


Figure 17: QQQ Volatility Smile (Recent Trading Day)



QQQ Greeks Summary

Greek	Mean	Std Dev	Median
Delta	+0.1025	0.6150	-0.0002
Gamma	+0.0098	0.0271	+0.0039
Theta	-0.0636	0.1492	-0.0292
Vega	+0.3579	0.5131	+0.1521
Rho	-0.0462	1.3822	+0.0000

Figure 18: QQQ Option Greeks Distribution

8 Cross-Asset Comparison

8.1 Implied Volatility Comparison

Table 15: Cross-Asset Implied Volatility Comparison

Symbol	Mean IV (Call)	Mean IV (Put)	IV Skew	Median IV
SPY	0.220	0.321	0.101	0.210
IWM	0.255	0.345	0.090	0.234
QQQ	0.267	0.342	0.075	0.227

Key observations:

- IWM (small-cap) exhibits the highest overall implied volatility, consistent with greater

uncertainty in small-cap stocks

- SPY shows the largest put-call IV skew (0.101), reflecting strong demand for downside protection on the broad market index
- QQQ has the smallest IV skew (0.075), potentially due to technology sector characteristics
- All three ETFs show the characteristic negative skew (puts more expensive than calls)

8.2 Liquidity Comparison

Table 16: Cross-Asset Liquidity Comparison

Symbol	Total Volume (B)	Avg Daily Vol	Total OI (B)	Records/Day
SPY	17.26	700	84.6	5,467
IWM	3.10	232	30.9	2,967
QQQ	7.68	501	36.6	4,148

SPY is by far the most liquid options market, with more than double the volume of QQQ and five times that of IWM.

9 Data Quality Assessment

9.1 Data Completeness

All datasets maintain high data quality with minimal missing values:

- Implied volatility: >99% available across all datasets
- Greeks (delta, gamma, theta, vega): >98% available
- Volume and open interest: 100% available
- Bid-ask quotes: >97% available

9.2 Data Anomalies

Potential data quality issues identified (consistent across datasets):

- IV > 5.0: <0.1% of records (extreme values during illiquid periods)
- Zero bid-ask spread: <0.1% (stale quotes)
- Negative open interest: 0% (validated)

10 Implications for Machine Learning

10.1 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
6. **Cross-asset features**: IV spreads between SPY/IWM/QQQ

10.2 Sample Size Considerations

With over 53 million records across three underlyings:

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

10.3 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. **Cross-asset modeling**: Predict IV relationships between underlyings
6. **Regime detection**: Identify market stress periods

11 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 combined sample size (53M+ records) suitable for deep learning
- Three complementary ETFs covering different market segments
- 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
- Cross-asset analysis capabilities for correlation studies

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.