Precog Recruitment Task 2026

Quantitative Finance Track

Pre-Feature Engineering Exploratory Data Analysis Report

Market Structure Characterization for
Algorithmic Trading Strategy Development

Dataset: 100 S&P 500 Constituents
Period: January 2016 – January 2026
Observations: 2,510 Trading Days × 100 Assets

January 2026

Contents

1 Executive Summary

This report presents a comprehensive exploratory data analysis of 100 S&P 500 constituent stocks spanning approximately 10 years of daily OHLCV data (January 25, 2016 to January 16, 2026). The analysis is structured to characterize market microstructure properties that directly inform feature engineering decisions and modeling strategy selection for algorithmic trading.

1.1 Key Findings

1. **Return Distribution:** Daily returns exhibit significant departure from normality with excess kurtosis of 20.34 and positive skewness of 0.31. The Jarque-Bera test statistic of 4.33×10^6 rejects normality at all conventional significance levels.

2. **Autocorrelation Structure:** Returns show weak negative first-order autocorrelation ($\rho_1 = -0.069$), while absolute returns display strong persistence ($\rho_1 = 0.231$), confirming volatility clustering.

3. **Volatility Dynamics:** Realized volatility exhibits extreme persistence with an estimated half-life of 131.7 days. The volatility regime transition matrix shows diagonal dominance exceeding 0.96, indicating sticky regimes.

4. **Correlation Behavior:** Mean pairwise correlation is 0.365, increasing by 87% during high-volatility periods (from 0.222 to 0.415). Correlation stability between sample halves is 0.702, with 32.4% of pairs experiencing correlation shifts exceeding 0.2.

5. **Cross-Sectional Structure:** Daily return ranks show near-zero autocorrelation across all horizons (1 to 252 days), with quintile persistence indistinguishable from random (20.4% vs 20% baseline).

6. **Calendar Effects:** None of the tested calendar anomalies (Monday effect, January effect, turn-of-month) achieve statistical significance at the 5% level.

2 Dataset Overview

2.1 Data Specification

The analysis utilizes daily OHLCV (Open, High, Low, Close, Volume) data for 100 anonymized S&P 500 constituent stocks. Table 1 presents the dataset characteristics.

Table 1: Dataset Summary Statistics

Attribute	Value
Number of Assets	100
Number of Trading Days	2,510
Date Range	2016-01-26 to 2026-01-16
Total Observations	251,000
Data Coverage	100% for all assets

2.2 Survivorship Bias Assessment

All 100 assets maintain complete data coverage throughout the sample period, indicating a survivorship-biased universe by construction. The median total return over the sample period is 324.18%, with 99 of 100 assets (99%) achieving positive returns and 88 assets exceeding 100%

cumulative return. This positive skew in long-term returns is characteristic of an S&P 500 constituent sample, where delisted or bankrupt firms are excluded.

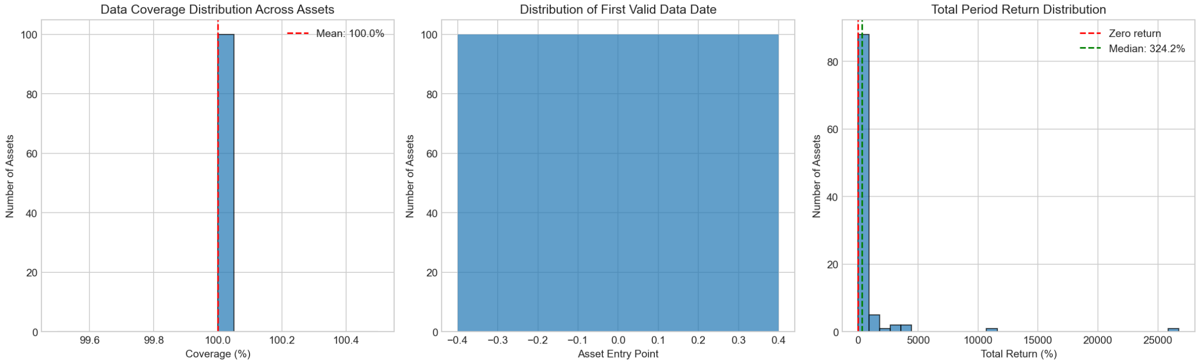


Figure 1: Survivorship bias assessment showing (a) data coverage distribution, (b) entry point distribution, and (c) total return distribution across assets.

Implication: The survivorship bias implies that backtest results will overstate historical performance. Out-of-sample validation on genuinely held-out data is essential for realistic performance assessment.

3 Asset Heterogeneity Analysis

3.1 Scale and Volatility Dispersion

The universe exhibits substantial heterogeneity across multiple dimensions. Table 2 summarizes the cross-sectional dispersion in key asset characteristics.

Table 2: Asset Heterogeneity Metrics

Metric	**Minimum**	**Maximum**	**Ratio**
Mean Price ($)	18.20	1,634.82	89.8x
Annualized Volatility (%)	18.1	59.3	3.3x
Mean Dollar Volume	9.61×10^7	1.43×10^{10}	149x

The 89.8x ratio in mean price and 149x ratio in dollar volume demonstrate that raw price-based or volume-based features would conflate scale effects with economically meaningful signals.

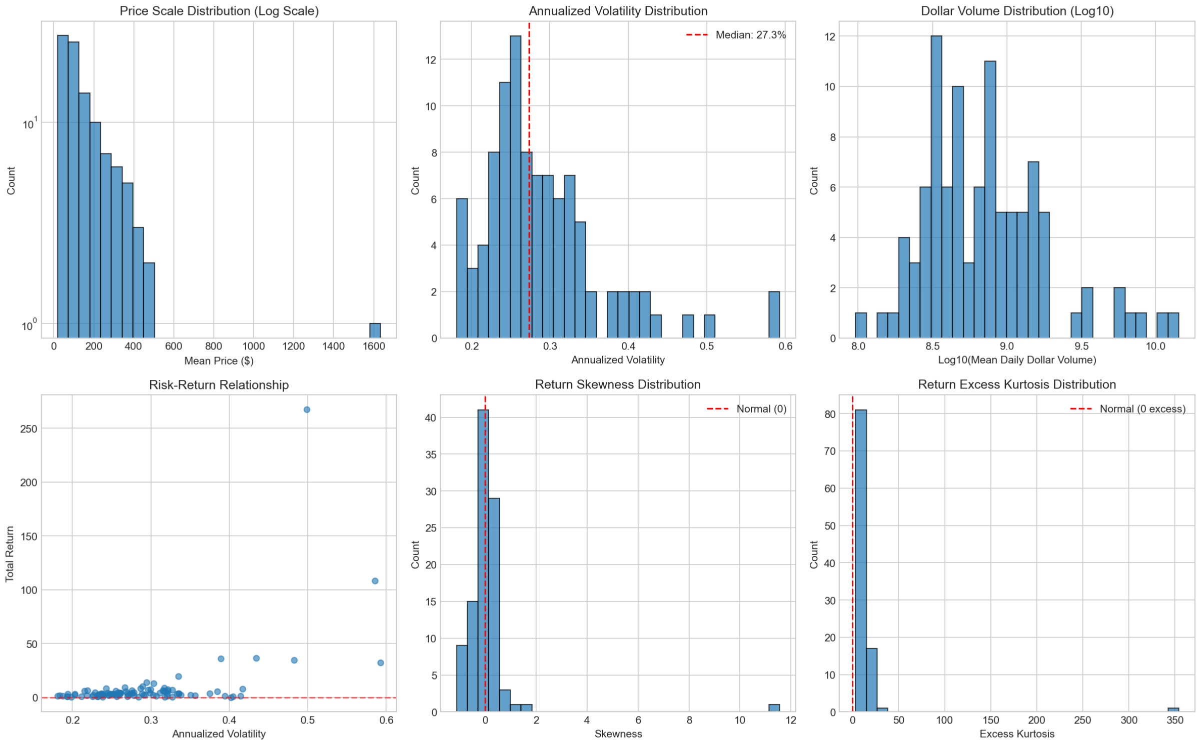


Figure 2: Asset heterogeneity analysis: (a) price distribution, (b) volatility distribution, (c) dollar volume distribution, (d) risk-return scatter, (e) skewness distribution, (f) kurtosis distribution.

3.2 Higher-Moment Characteristics

Analysis of return distribution moments reveals:

- 44% of assets exhibit negative skewness, indicating asymmetric downside risk
- 100% of assets display excess kurtosis greater than 3, confirming universal fat-tail behavior

Implication: All features must be constructed in a scale-invariant manner using returns rather than prices. Volatility normalization (z-scoring) and cross-sectional ranking are essential preprocessing steps.

4 Return-Level Properties

4.1 Marginal Return Distribution

Pooling all 251,000 daily returns yields the aggregate distributional statistics presented in Table 3.

Table 3: Aggregate Return Distribution Statistics

Statistic	**Value**
Mean Daily Return	0.0799%
Standard Deviation	1.8895%
Skewness	0.3051
Excess Kurtosis	20.3447
Jarque-Bera Statistic	4.33×10^6
Jarque-Bera p-value	$< 10^{-300}$

The Jarque-Bera test decisively rejects the null hypothesis of normality. The Q-Q plot in Figure 3 demonstrates systematic deviation from the normal distribution in both tails, with the log-scale histogram revealing tail probabilities orders of magnitude higher than Gaussian predictions.

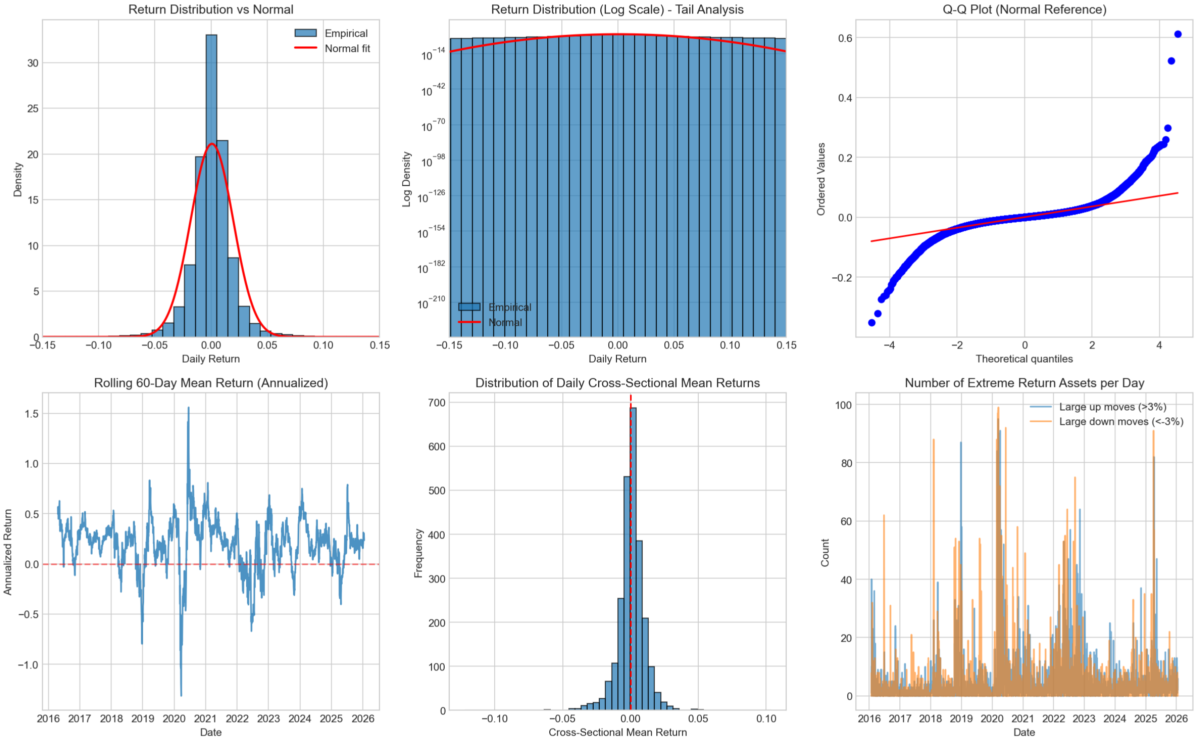


Figure 3: Return distribution analysis: (a) histogram vs. normal fit, (b) log-scale tail analysis, (c) Q-Q plot, (d) rolling mean return, (e) cross-sectional mean distribution, (f) extreme return frequency.

Implication: Gaussian risk models (e.g., mean-variance optimization assuming normal returns) will systematically underestimate tail risk. Feature engineering should incorporate tail-risk measures and avoid assumptions of elliptical distributions.

4.2 Autocorrelation Analysis

The autocorrelation function (ACF) reveals a fundamental asymmetry between returns and absolute returns.

Table 4: Autocorrelation Comparison: Returns vs. Absolute Returns

Lag	Return ACF	|Return| ACF
1	−0.069	0.231
5	−0.002	0.175
10	0.002	0.157
20	−0.002	0.110

The Ljung-Box test confirms serial correlation in both series:

- Returns: $Q(20) = 373.2$, $p < 10^{-50}$ (weak but significant)
- Absolute Returns: $Q(20) = 4489.5$, $p < 10^{-300}$ (strong persistence)

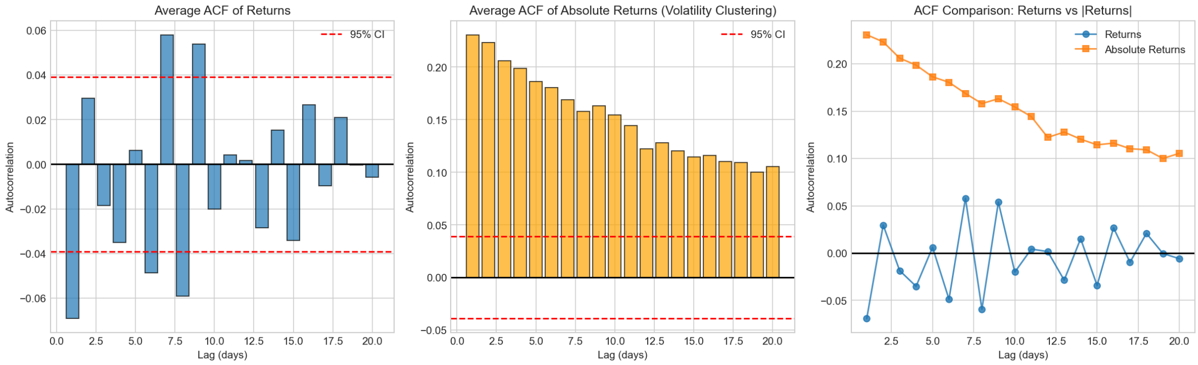


Figure 4: Autocorrelation analysis: (a) return ACF, (b) absolute return ACF showing volatility clustering, (c) comparison plot.

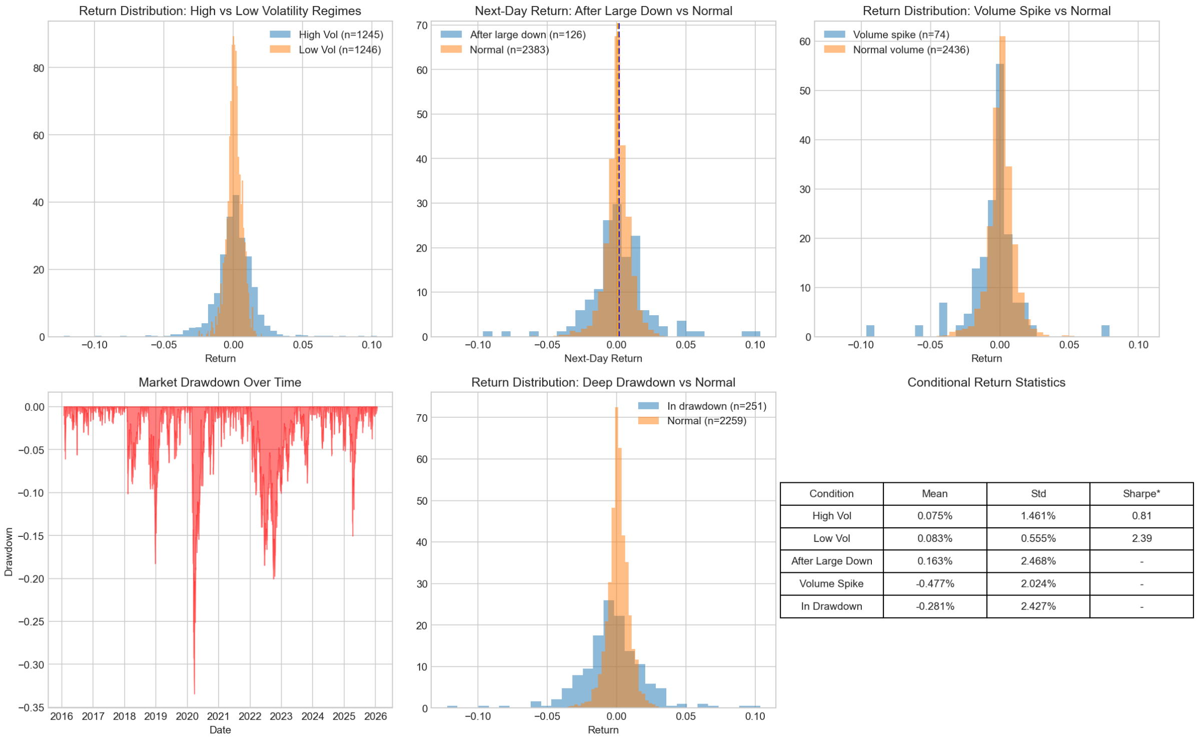
Implication: The weak negative return autocorrelation suggests limited directional predictability from lagged returns alone. However, the strong absolute return autocorrelation confirms volatility clustering, justifying GARCH-family models and volatility-based features.

4.3 Conditional Return Behavior

Returns were analyzed conditional on volatility regime, prior extreme moves, and volume conditions.

Table 5: Conditional Return Statistics

Condition	Mean	Std Dev	Implied Sharpe
High Volatility Regime	0.075%	1.461%	0.81
Low Volatility Regime	0.083%	0.555%	2.39
After Large Down Move	0.163%	2.468%	–
Volume Spike Days	−0.477%	2.024%	–
Deep Drawdown Period	−0.281%	2.427%	–



Condition	Mean	Std	Sharpe*
High Vol	0.075%	1.461%	0.81
Low Vol	0.083%	0.555%	2.39
After Large Down	0.163%	2.468%	-
Volume Spike	-0.477%	2.024%	-
In Drawdown	-0.281%	2.427%	-

Figure 5: Conditional return behavior: (a) high vs. low volatility regimes, (b) post-crash returns, (c) volume spike returns, (d) drawdown time series, (e) drawdown-conditional returns, (f) summary statistics.

Key observations:

1. Low volatility periods exhibit higher risk-adjusted returns (Sharpe 2.39 vs. 0.81)
2. Post-crash days show elevated mean returns (0.163%), consistent with short-term reversal
3. Volume spike days exhibit negative mean returns (−0.477%), indicating volume signals stress

Implication: Regime conditioning substantially alters return distributions. Static unconditional models miss exploitable structure that emerges when stratifying by volatility state.

5 Volatility Structure

5.1 Persistence and Clustering

Realized volatility (20-day rolling standard deviation, annualized) exhibits extreme persistence. Table 6 summarizes the key metrics.

Table 6: Volatility Persistence Metrics

Metric	Value
Volatility ACF at Lag-1	0.9948
Estimated Half-Life	131.7 days
Time in Low Vol Regime	24.8%
Time in High Vol Regime	24.8%

The half-life of 131.7 days implies that volatility shocks persist for approximately 6 months before decaying by half, justifying the use of extended lookback windows for volatility estimation.

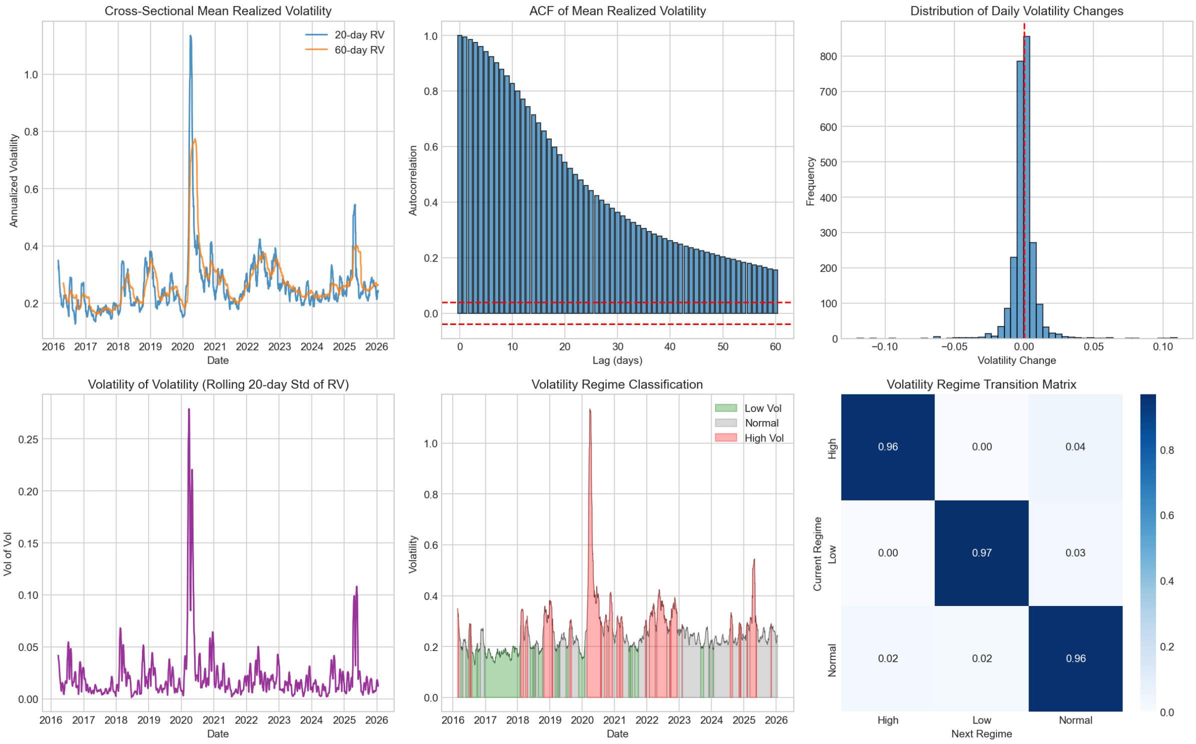


Figure 6: Volatility clustering analysis: (a) realized volatility time series, (b) volatility ACF, (c) volatility change distribution, (d) volatility-of-volatility, (e) regime classification, (f) transition matrix.

5.2 Regime Transition Dynamics

The volatility regime transition matrix exhibits strong diagonal dominance:

Table 7: Volatility Regime Transition Probabilities

	Next Regime		
Current Regime	High	Low	Normal
High	0.96	0.00	0.04
Low	0.00	0.97	0.03
Normal	0.02	0.02	0.96

The 96–97% same-state persistence probabilities indicate that regime transitions are rare events, justifying models that condition on regime state rather than attempting high-frequency regime switching.

5.3 Cross-Asset Volatility Synchronization

Pairwise volatility correlations average 0.625 (median 0.639), indicating strong co-movement in risk levels across the universe.

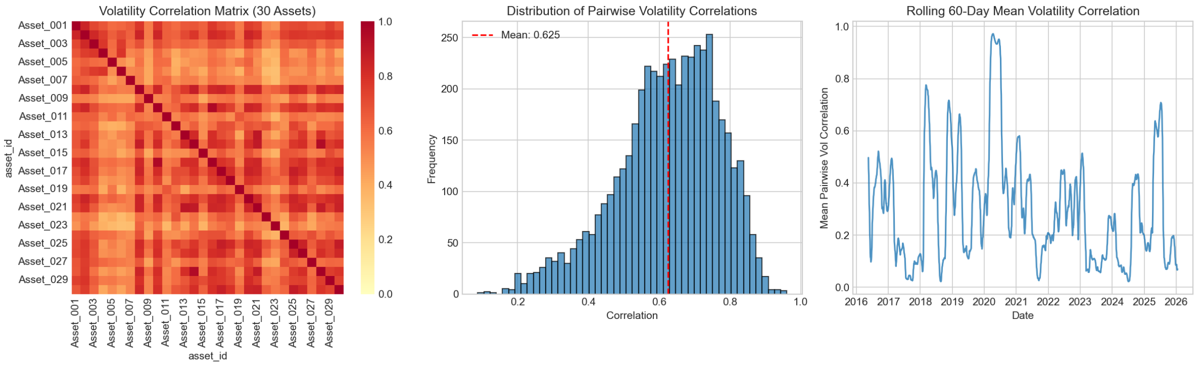


Figure 7: Volatility synchronization: (a) correlation heatmap, (b) pairwise correlation distribution, (c) rolling mean correlation.

Implication: A single market-wide volatility factor captures substantial cross-asset risk variation. Asset-specific volatility timing in isolation provides limited marginal information beyond market volatility state.

6 Correlation and Dependence Structure

6.1 Correlation Dynamics

Rolling 60-day correlation analysis reveals substantial time variation in the dependence structure.

Table 8: Correlation Regime Statistics

Metric	Value
Full-Period Mean Correlation	0.365
Correlation in High Vol Periods	0.415
Correlation in Low Vol Periods	0.222
Stress-Induced Correlation Increase	87.0%

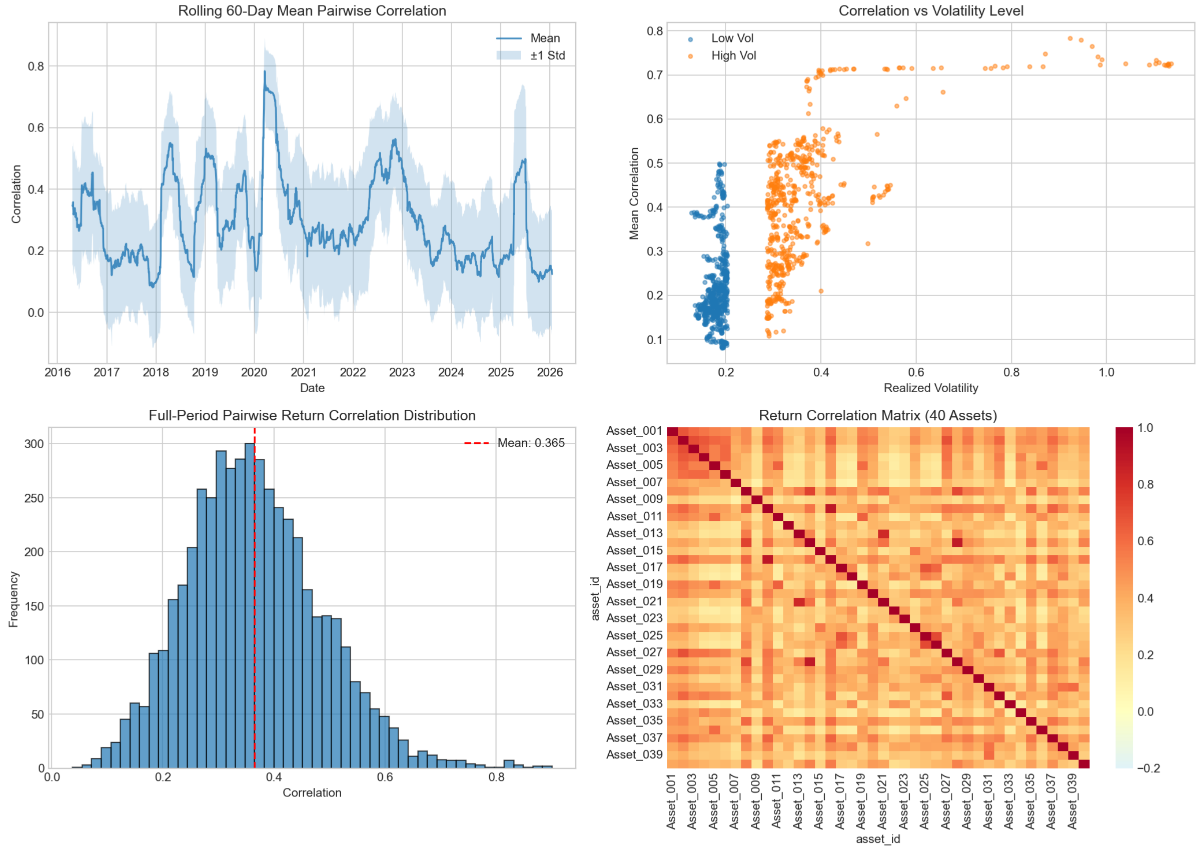


Figure 8: Correlation dynamics: (a) rolling mean correlation with ±1 standard deviation band, (b) correlation vs. volatility scatter, (c) full-period correlation distribution, (d) correlation heatmap.

The 87% increase in average correlation during high-volatility periods demonstrates that diversification benefits diminish precisely when they are most needed—a well-documented phenomenon in financial markets.

6.2 Correlation Stability

Comparing pairwise correlations between the first and second halves of the sample period:

Table 9: Correlation Stability Metrics

Metric	**Value**
First-Half vs. Second-Half Correlation	0.702
Mean Correlation Change	-0.152
Std of Correlation Change	0.101
Pairs with $\lvert\Delta\rho\rvert > 0.2$	32.4%

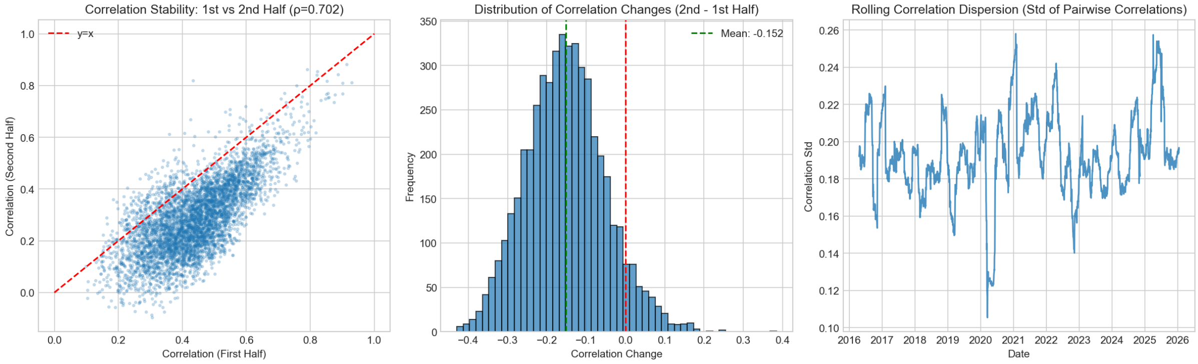


Figure 9: Correlation instability: (a) first-half vs. second-half scatter, (b) correlation change distribution, (c) rolling correlation dispersion.

Implication: Static correlation estimates are insufficiently stable for multi-year horizons. Rolling or exponentially-weighted covariance estimation is required, and any pairs-trading strategy must incorporate adaptive threshold mechanisms.

7 Cross-Sectional Structure

7.1 Return Dispersion

Daily cross-sectional return dispersion (standard deviation across assets) averages 1.43% and correlates 0.447 with absolute market returns, indicating that high-dispersion days tend to coincide with market stress.

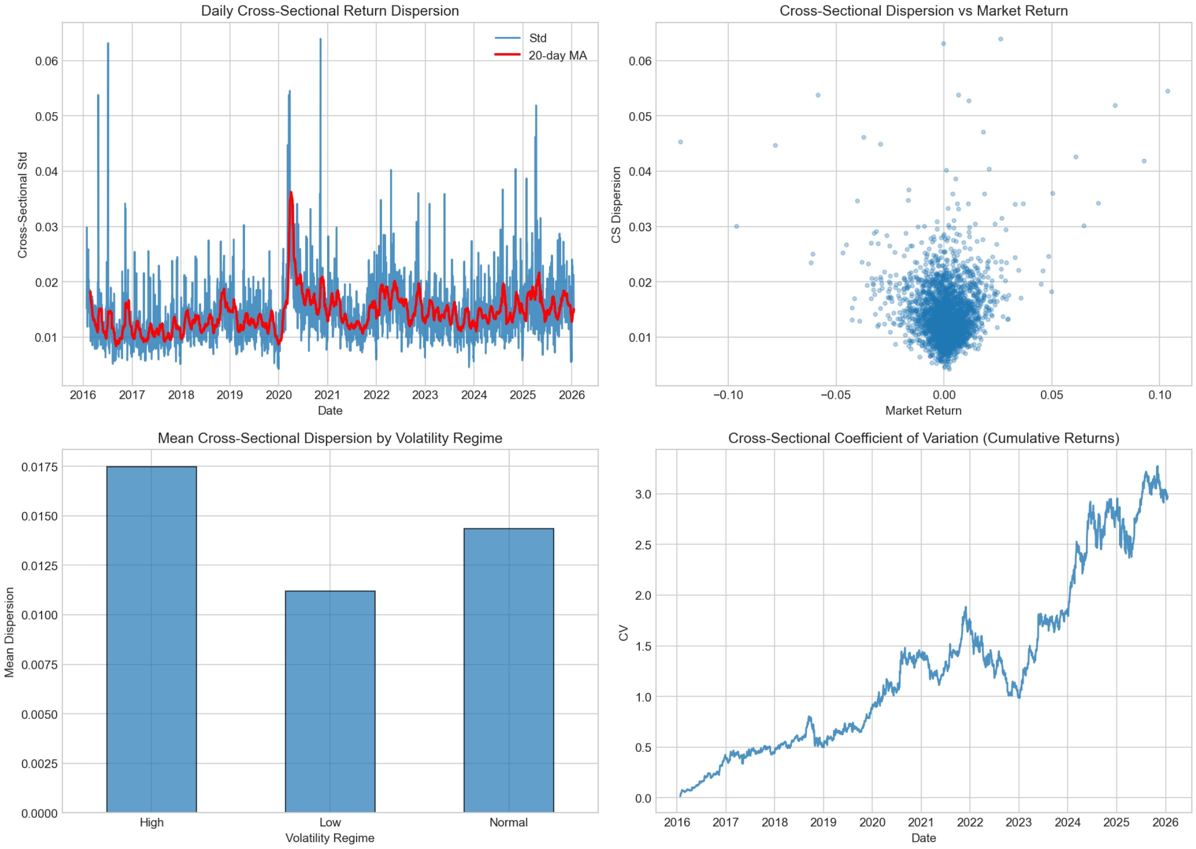


Figure 10: Cross-sectional dispersion: (a) time series with 20-day MA, (b) dispersion vs. market return, (c) dispersion by volatility regime, (d) cumulative return coefficient of variation.

7.2 Rank Stability

Cross-sectional return rank autocorrelation provides a test of momentum/reversal effects at the asset level.

Table 10: Rank Autocorrelation by Lag

Lag (Days)	Rank ACF
1	−0.015
5	−0.002
10	0.002
20	0.002
60	0.002
120	0.001
252	−0.004

Quintile persistence analysis:

- Top Quintile Next-Day Persistence: 20.4%
- Bottom Quintile Next-Day Persistence: 21.1%
- Random Baseline: 20.0%

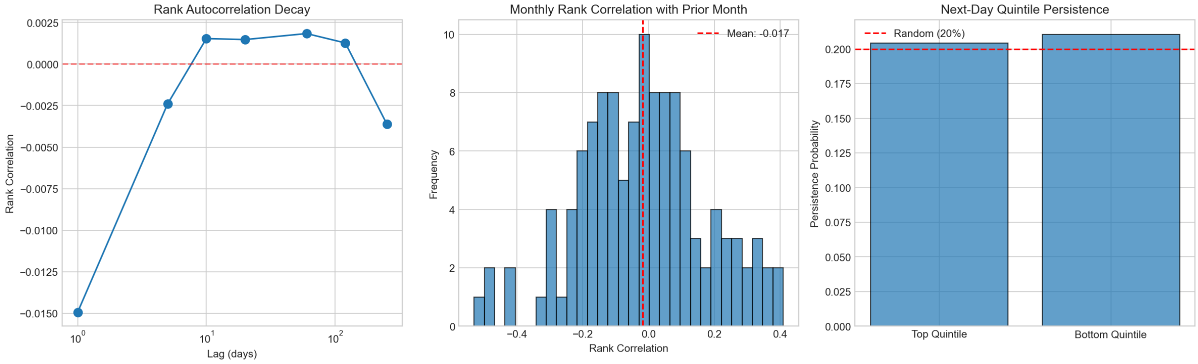


Figure 11: Rank stability: (a) rank ACF decay, (b) monthly rank correlation distribution, (c) quintile persistence vs. random baseline.

Implication: Daily return ranks are effectively random, ruling out naive daily cross-sectional momentum or reversal strategies. Any viable ranking-based signal must operate on longer horizons with appropriate smoothing.

8 Volume and Liquidity Dynamics

8.1 Volume-Volatility Relationship

Volume exhibits weak negative correlation (-0.053) with volatility in surprise terms, but displays strong asymmetry with respect to return direction.

Table 11: Volume Asymmetry Statistics

Metric	**Value**
Mean Volume on Up Days	1.58B
Mean Volume on Down Days	1.66B
Down/Up Volume Ratio	1.052
Volume Autocorrelation (Lag-1)	0.75

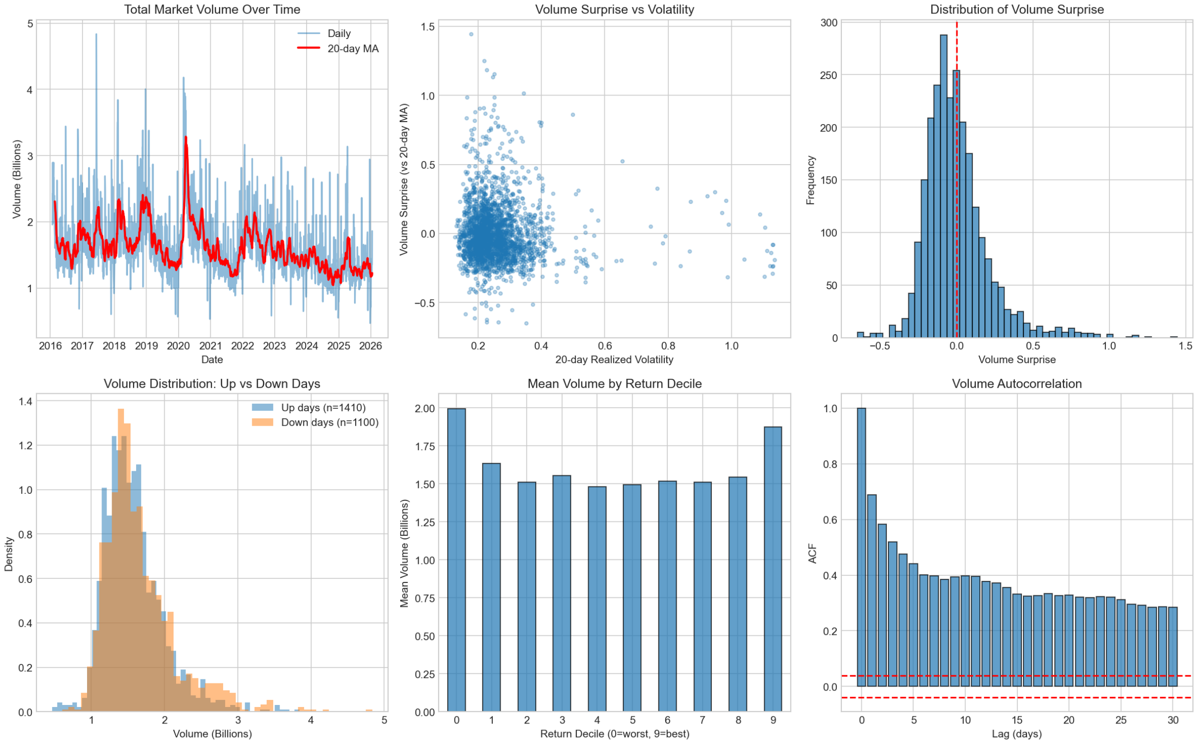


Figure 12: Volume dynamics: (a) total market volume time series, (b) volume surprise vs. volatility, (c) volume surprise distribution, (d) volume on up vs. down days, (e) volume by return decile, (f) volume ACF.

The U-shaped relationship between volume and return decile (highest volume in extreme deciles) confirms that information arrival drives both volume and price movement magnitude.

Implication: Volume provides incremental information beyond returns alone. Features should incorporate volume-weighted measures and volume surprise (relative to recent average) as conditioning variables.

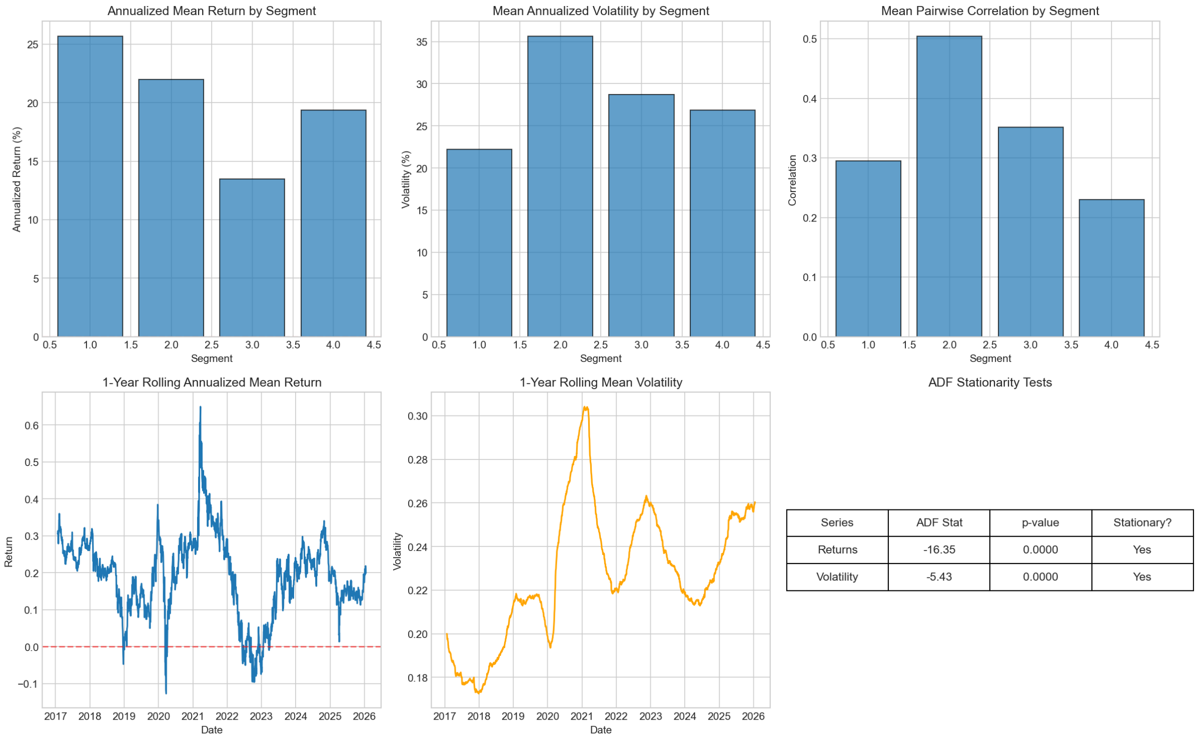
9 Regime Non-Stationarity

9.1 Temporal Segment Analysis

The sample was divided into four segments to assess parameter stability.

Table 12: Statistical Properties by Temporal Segment

Segment	Period	Ann. Return	Ann. Vol	Mean Corr
1	2016-01 to 2018-07	25.7%	22.2%	0.295
2	2018-07 to 2021-01	22.0%	35.7%	0.505
3	2021-01 to 2023-07	13.4%	28.7%	0.352
4	2023-07 to 2026-01	19.4%	26.9%	0.230



Series	ADF Stat	p-value	Stationary?
Returns	-16.35	0.0000	Yes
Volatility	-5.43	0.0000	Yes

Figure 13: Temporal segment analysis: (a) return by segment, (b) volatility by segment, (c) correlation by segment, (d) rolling annualized return, (e) rolling volatility, (f) ADF stationarity tests.

Augmented Dickey-Fuller tests:

- Returns: ADF statistic $= -16.35$, $p < 0.0001$ (stationary)
- Volatility: ADF statistic $= -5.43$, $p < 0.0001$ (stationary)

While both series reject unit roots, the substantial variation in segment-level statistics (volatility ranging from 22.2% to 35.7%, correlation from 0.230 to 0.505) indicates conditional non-stationarity.

Implication: Fixed-parameter models calibrated on full-sample data will fail to adapt to regime shifts. Rolling estimation with exponential decay weighting is essential.

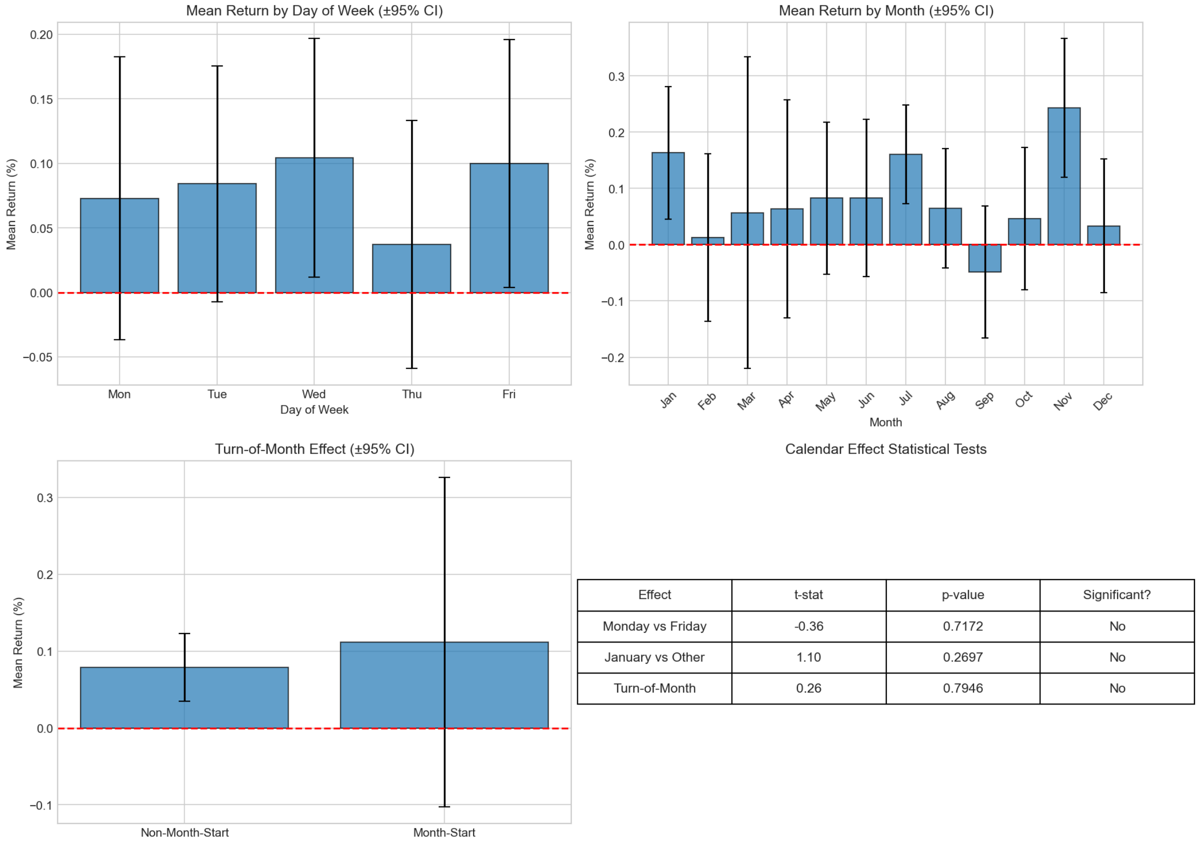
10 Calendar Effect Analysis

10.1 Statistical Tests

Three calendar anomalies were tested using independent-samples t-tests.

Table 13: Calendar Effect Statistical Tests

Effect	**t-statistic**	**p-value**	**Significant?**
Monday vs. Friday	−0.36	0.717	No
January vs. Other Months	1.10	0.270	No
Turn-of-Month	0.26	0.795	No


Mean Return by Day of Week (±95% CI)

Mean Return by Month (±95% CI)

Turn-of-Month Effect (±95% CI)

Calendar Effect Statistical Tests

Effect	t-stat	p-value	Significant?
Monday vs Friday	-0.36	0.7172	No
January vs Other	1.10	0.2697	No
Turn-of-Month	0.26	0.7946	No

Figure 14: Calendar effects: (a) mean return by day of week with 95% CI, (b) mean return by month with 95% CI, (c) turn-of-month effect, (d) statistical test summary.

None of the tested calendar effects achieve statistical significance at the 5% level. While point estimates show some variation (e.g., November appears strongest, Thursday weakest), the wide confidence intervals preclude reliable inference.

Implication: Calendar-based signals should not form primary alpha sources. Day-of-week dummies may be included as control variables in regression analysis but should not drive position sizing.

11 Synthesis and Implications for Feature Engineering

11.1 What Forms of Alpha are Structurally Unlikely

Based on the empirical evidence, the following approaches face fundamental structural barriers:

Table 14: Structurally Unlikely Alpha Sources

Approach	**Evidence Against**
Simple daily momentum	Return ACF ≈ 0; rank persistence $\approx$ random
Static correlation-based strategies	32.4% of pairs shift > 0.2 across sample halves
Calendar anomalies	All tests $p > 0.25$
Gaussian risk models	Kurtosis $= 20.3$; Jarque-Bera $= 4.3 \times 10^6$
Equal-weighted signals	89.8x price ratio; 3.3x volatility ratio

11.2 What Forms of Alpha are Structurally Plausible

The following approaches have supportive empirical foundations:

Table 15: Structurally Plausible Alpha Sources

Approach	Supporting Evidence
Volatility timing	Vol ACF = 0.995; half-life = 132 days
Regime-conditional strategies	Transition matrix diagonal > 0.96
Volume-informed signals	5.2% higher volume on down days; U-shaped volume-return
Risk-adjusted features	Universal heterogeneity demands normalization
Cross-sectional ranking (longer horizon)	Dispersion correlates 0.45 with |market return|

11.3 Feature Engineering Constraints

The EDA findings impose the following constraints on feature construction:

1. **Scale Invariance:** Use returns rather than prices; normalize by rolling volatility
2. **Cross-Sectional Ranking:** Prefer relative rank features over absolute values
3. **Adaptive Estimation:** Use rolling windows (20–60 days) with exponential decay
4. **Regime Conditioning:** Include volatility regime as a feature or conditioning variable
5. **Tail Awareness:** Incorporate downside-specific measures (semi-deviation, VaR, drawdown)
6. **Volume Integration:** Construct volume-weighted and volume-surprise features

12 Proposed Modeling Framework

Based on the EDA findings, the following modeling approach is recommended:

12.1 Feature Categories

1. **Momentum Features:** Risk-adjusted returns over multiple horizons (5, 20, 60 days), cross-sectionally ranked
2. **Volatility Features:** Rolling realized volatility, Parkinson volatility, volatility surprises, volatility regime indicators
3. **Volume Features:** Volume surprise (vs. 20-day MA), directional volume asymmetry, dollar-volume rank
4. **Mean Reversion Features:** Distance from rolling mean (z-score), distance from 52-week high/low
5. **Cross-Sectional Features:** Industry-relative momentum, beta-adjusted returns, correlation with market
6. **Regime Features:** Current volatility regime, correlation dispersion, market drawdown state

12.2 Model Selection Considerations

Given the non-linear regime dependence and non-Gaussian distributions, the following model classes merit consideration:

- **Gradient Boosting (XGBoost, LightGBM):** Handles non-linearities and feature interactions without explicit specification
- **Regime-Switching Models:** Markov-switching specifications that condition on volatility state
- **Ensemble Methods:** Combining multiple signal sources with regime-dependent weighting

12.3 Validation Strategy

The strong temporal non-stationarity necessitates strict walk-forward validation:

- Minimum 2-year out-of-sample test period
- No future information leakage in feature construction
- Transaction cost inclusion (10 bps per trade as specified)
- Multiple performance metrics: Sharpe ratio, maximum drawdown, turnover

13 Conclusion

This pre-feature engineering EDA establishes the statistical foundation for algorithmic strategy development on the provided S&P 500 universe. The key findings—non-Gaussian returns, strong volatility clustering, regime-dependent correlations, and absence of daily rank persistence—impose specific constraints on feature construction and model selection.

The evidence supports focusing on:

1. Volatility timing and regime-conditional positioning
2. Risk-normalized, cross-sectionally ranked features
3. Adaptive estimation with rolling windows
4. Volume-informed signal construction

The evidence argues against:

1. Simple daily momentum or reversal
2. Static correlation-based pair trading
3. Calendar-driven signals
4. Gaussian risk assumptions

Subsequent feature engineering and model development will proceed under these empirically-grounded constraints.

A Statistical Test Details

A.1 Jarque-Bera Test

The Jarque-Bera test statistic is:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \tag{1}$$

where S is skewness, K is kurtosis, and n is sample size. Under the null of normality, $JB \sim \chi^2_2$.

A.2 Augmented Dickey-Fuller Test

The ADF test regresses:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t \tag{2}$$

The null hypothesis $H_0 : \gamma = 0$ implies a unit root (non-stationarity).

A.3 Ljung-Box Test

The Ljung-Box statistic tests for serial correlation:

$$Q(m) = n(n+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{n-k} \tag{3}$$

where $\hat{\rho}_k$ is the sample autocorrelation at lag k. Under H_0 of no autocorrelation, $Q(m) \sim \chi^2_m$.