Exploratory Analysis of Large-Cap Tech Equities

Daeseo Lee

Auguest 7, 2025

Abstract

I built a reproducible pipeline to move from raw market data to decision-ready diagnostics and model inputs. Beyond producing charts, I focused on concepts: data integrity (adjusted prices, trading calendars), return construction (simple vs. log), risk estimation (realized volatility and drawdowns), dependence structure (correlations and market β), seasonality caveats, the volume–volatility link, and a momentum proxy (RSI). Each step is implemented with explicit alignment rules to avoid look-ahead bias and to ensure the resulting feature matrix can support downstream testing without leaking future information.

1 Data

Source and scope. I used Yahoo Finance via `yfinance` for AAPL, GOOG, and MSFT with daily OHLCV. Close-to-close returns are standard for equity EDA and offer a consistent sampling scheme across names. For the market proxy (GSPC), I relied on adjusted prices to remove discontinuities from splits and dividends.

Cleaning and alignment. A key lesson was that market data are calendar-sensitive. I reindexed each series to a common business-day calendar and handled gaps with forward/backward fills only after verifying that missing values reflected exchange holidays rather than data loss. All transformations are date-indexed, and I maintained a strict rule: features at time t are computed using information available at or before t.

2 Methods

Returns and distributions. I computed simple returns $r_t = (P_t - P_{t-1})/P_{t-1}$ and log returns $\ell_t = \ln(P_t/P_{t-1})$. Simple returns are intuitive for compounding and P&L; log returns are additive over time and closer to Gaussian for many assets, though heavy tails remain. Inspecting both helps catch scale effects and asymmetries.

Realized volatility and trend. I estimated rolling realized volatility as $\hat{\sigma}_{t,w} = \sqrt{252} \cdot \mathrm{std}(r_{t-w+1:t})$. The annualization factor assumes $\sqrt{T}$ scaling; I checked stability across windows (e.g., w=20, 30) to avoid conclusions that are artifacts of the chosen horizon. A 20-day moving average (MA) is used to smooth short-term noise and to highlight persistent drift. MA is descriptive, not predictive here: it helps contextualize regime shifts that volatility also flags.

Drawdowns. Drawdown at time t is $D_t = \frac{P_t}{\max_{s \le t} P_s} - 1$. This is a path-dependent loss measure and often more decision-relevant than volatility during stress. I found it useful to view volatility and drawdowns together to separate normal choppiness from equity-crisis behavior.

Dependence and co-movement. I computed contemporaneous correlations on aligned daily returns, emphasizing that high sector co-movement does not imply redundancy: imperfect correlation preserves diversification benefits at the margin. I was careful to avoid spuriously high correlations from misaligned calendars.

Seasonality, conservatively. I evaluated day-of-week and month-of-year averages as descriptive summaries only. Calendar effects are notoriously fragile; without out-of-sample testing and multiple-hypothesis controls, they should inform priors rather than trading rules. Reporting the magnitude and stability over subperiods is more informative than a single global mean.

Volume and extremes. Rather than correlating volume with signed returns (which tends to wash out), I examined $\mathrm{corr}(\text{Volume}, |r_t|)$. This captures the well-documented volume–volatility connection (consistent with the mixture-of-distributions hypothesis): large absolute moves co-occur with elevated activity even when direction is mixed.

Market sensitivity (CAPM). I estimated β via the cross-sectional regression $r_{i,t} = \alpha_i + \beta_i r_{m,t} + \varepsilon_{i,t}$ using NumPy's `polyfit`. Daily data rarely satisfy homoskedastic, i.i.d. errors, so I interpret β as a first-order exposure measure rather than a structural parameter. I verified that intercepts were near zero on daily horizons and that β estimates are stable across reasonable subperiods.

Momentum proxy (RSI). The 14-day RSI contrasts average gains and losses to provide a bounded oscillator in $[0, 100]$. I implemented a simple moving-average variant and noted the trade-off with Wilder's smoothing (EMA-like), which reacts differently to shocks. RSI is sensitive to window choice; I treated it as a comparative momentum descriptor across tickers, not as a standalone signal.

Feature matrix and leakage control. I consolidated returns, MA, volatility, and RSI into a single table keyed by date. The matrix is constructed so each row's features are contemporaneous with—or prior to—the target return. This alignment is explicitly enforced to prevent target leakage and to make the dataset immediately usable for walk-forward tests.

3 Key Findings

Shared risk with residual differentiation. The names exhibit high but imperfect correlations, consistent with common sector and market factors plus idiosyncratic components. This leaves room for diversification even within large-cap tech.

Regime behavior matters. Volatility clusters and drawdowns reveal that risk is not constant through time; estimates built on rolling windows adapt more realistically than fixed-sample statistics.

Descriptive seasonality, cautious inference. Calendar patterns appear in-sample at small magnitudes. Without explicit controls for multiple testing and out-of-sample validation, they are best treated as priors for model features rather than as rules.

Activity and uncertainty move together. The positive association between volume and $|r_t|$ confirms that the largest price changes tend to arrive on heavy-volume days, reinforcing the practical value of conditioning risk limits or position sizes on recent activity.

Betas modestly > 1. CAPM slopes slightly above one indicate amplified sensitivity to broad

market moves; the daily intercepts are near zero, as expected. This provides a clean baseline exposure measure to complement the correlation view.

Momentum profiles differ. RSI dynamics vary by ticker (frequency and amplitude of excursions), suggesting that any momentum-based rule should be tuned by name rather than applied uniformly.