1. Problem Statement

Volatility is a crucial factor in financial risk management, portfolio diversification, and sector analysis. Commonly, stocks with similar volatility respond to the same economic, sectorial, or market factors.

This project looks at the daily price data of NIFTY 50 constituents from 2020 to 2025. It identifies groups of stocks that have similar volatility patterns.

The analysis uses:

- **Rolling volatility** to capture short-term risk
- **PCA** to reduce dimensionality and extract main volatility components.
- **K-Means** to cluster stocks based on dominant volatility patterns.
- **Sector-wise interpretation** to understand industry-level co-movements.

The results offer insights into how different NIFTY 50 sectors move together under varying market conditions.

2. Dataset Description

2.1 Source and Scope

Dataset was created by help of **yfinance library** of Python. This library gives access to the reliable daily financial data straight from Yahoo Finance.

- **Index Analyzing:** NIFTY50 (as of July 2023)
- **Stock Symbols:** 50 major Indian corporations that trade on NSE
- **Time Frame:** 1 January 2020 to 1 January 2025, which includes:

This duration exposes the volatility patterns illustrating the actual market structure during the significant transitions.

2.2 Structure of Raw Data

The collected dataset consists of daily Adjusted Close prices for each stock, forming:

- 1238 Rows: Individual trading days
- 50 Columns: Stock tickers

Missing values are removed thoroughly to ensure clean modelling.

Stocks Date	ADANIENT.NS	ADANIPORTS.NS	APOLLOHOSP.NS	ASIANPAINT.NS	....	UPL.NS	WIPRO.NS
2020-01-30	0.017715	0.011853	0.017315	0.012892	.....	0.013129	0.012432
2020-01-31	0.017784	0.012383	0.015802	0.012944	.....	0.014053	0.012828
2020-02-03	0.019672	0.012411	0.015981	0.014507	.....	0.01469	0.012515
2020-02-04	0.016983	0.012838	0.015895	0.013056	.....	0.015615	0.012832

2020-02-05	0.018904	0.012873	0.016162	0.01336	…….	0.016553	0.012651

Data Preview

3. Theoretical Background & Methodology

3.1 Log-Returns: Why Not Simple Returns?

Instead of using simple returns:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}},$$

we use **log-returns**:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

Because:

- Log-returns are time additive
- Approximates continuous compounding
- Prevent issues when returns are small
- They allow elegant statistical modelling

Therefore, in finance, log-returns became the default go-to choice.

3.2 Rolling Voltality

Volatility measures how much prices change. High volatility indicates uncertain markets, while low volatility suggests stable markets.

Time-varying volatility:

$$\sigma_t = \sqrt{\frac{1}{21}\sum_{i=1}^{21} r_{t-i}^2}$$

This rolling window gives: monthly sensitivity, smooth transitions, noise reduction and insight on regime changes

The result is a matrix where each stock has around 1200 volatility values across time.

The Trade-off

- Window (10 days) : Very noisy. The volatility line jumps around too much to find stable clusters.
- Window (21 days): It smooths out daily noise but reacts quickly enough when a crisis (like COVID or Adani Report) hits.

- Window(1 Quarter): Too laggy. By the time this volatility spikes, the market crash is often already over.

3.3 Principal Component Analysis (PCA)

The large dataset with around 50 elements becomes hard to interpret and are highly redundant, noisy and can cause issue due to multicollinearity.

PCA finds directions (principal components) that capture maximum variance. These directions are orthogonal. And works on eigen vectors and eigen values.

Mathematically, PCA solves: $\sum \upsilon = \lambda \upsilon$

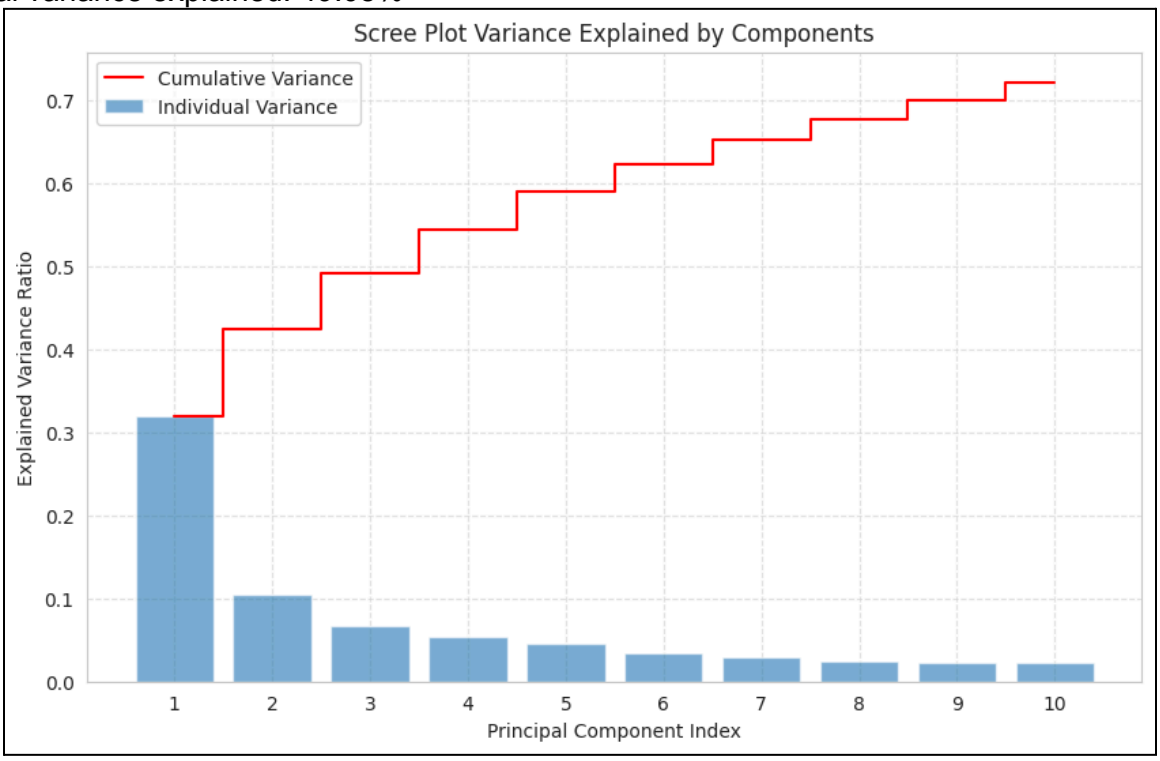
Where:

- υ = eigenvectors (directions)
- λ = eigenvalues (variance explained by each direction)

To determine the no of components to consider we used **Scree Plot**, the Scree plot clearly shows a steep drop after PC1 and then at PC2, meaning:

- PC1 = overall volatility level (Variance explained =31.89%)
- PC2 = sector-specific effects (Variance explained =10.51%)
- PC3 = idiosyncratic deviations (Variance explained =6.68%)

Total variance explained: 49.08%



From the Plot we considered 3 Components

3.4 K-Means Clustering:

K-Means tries to group data such that points within a group cluster are similar. We used to identify similar volatility behaviour, K-Means was applied using PCA scores.

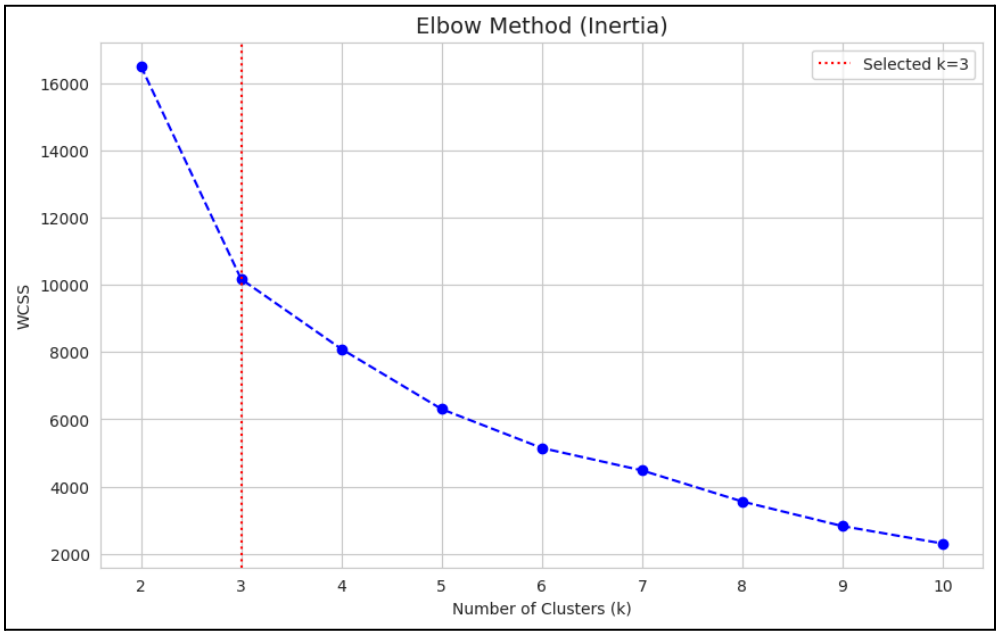
Mathematically:

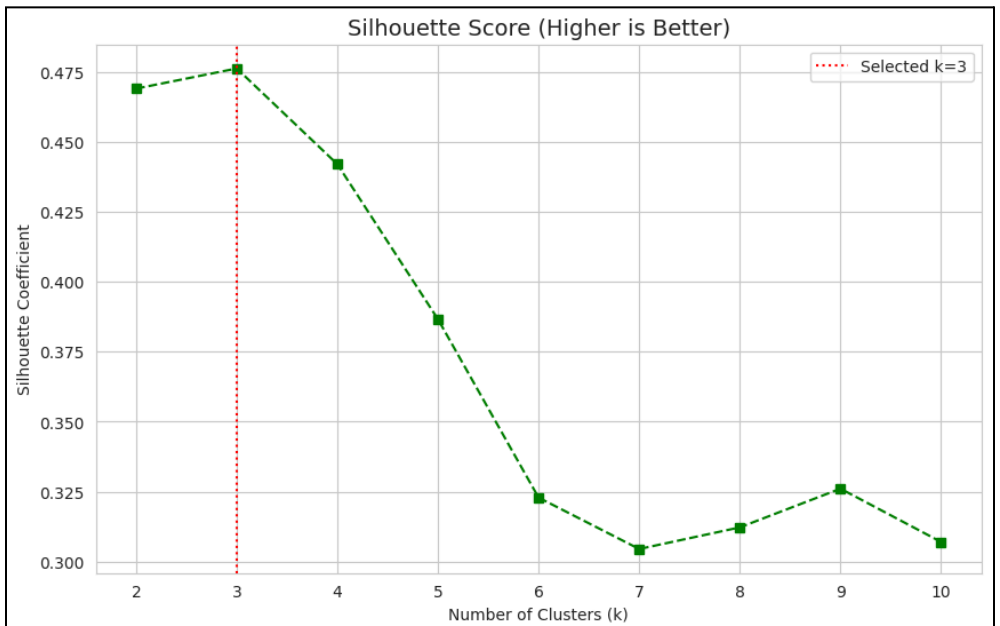
$$\text{minimize} \sum_{k=1}^{K} \sum_{x \in C_k} ||x - \mu_k||^2$$

Where:

- μ_k = centroid of cluster
- C_k = points in cluster k

We run K-Means for k = 2 to 10 and plot the elbow curve. However this was'nt enough to tell us the suitable K so we calculated the Silhouette Score

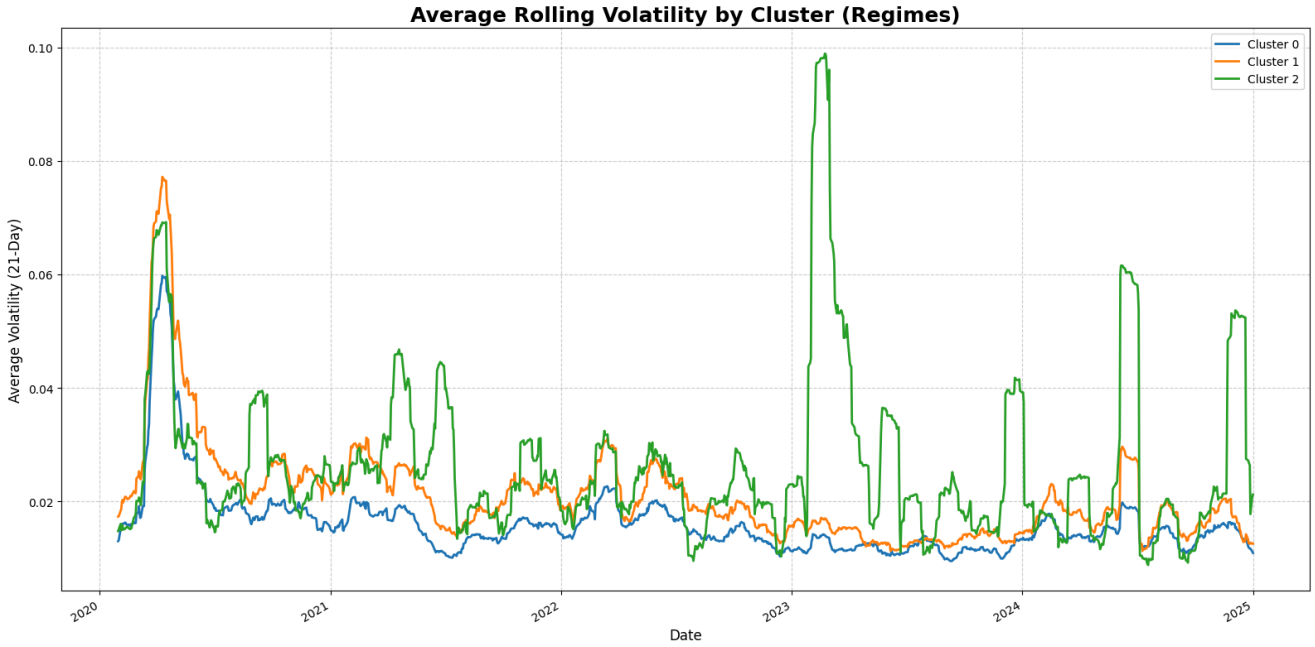




4. Results

4.1 Temporal Evolution of Cluster Volatilities

Cluster-wise average volatility provides crucial insight into how different groups of stocks behave throughout the 5 years.

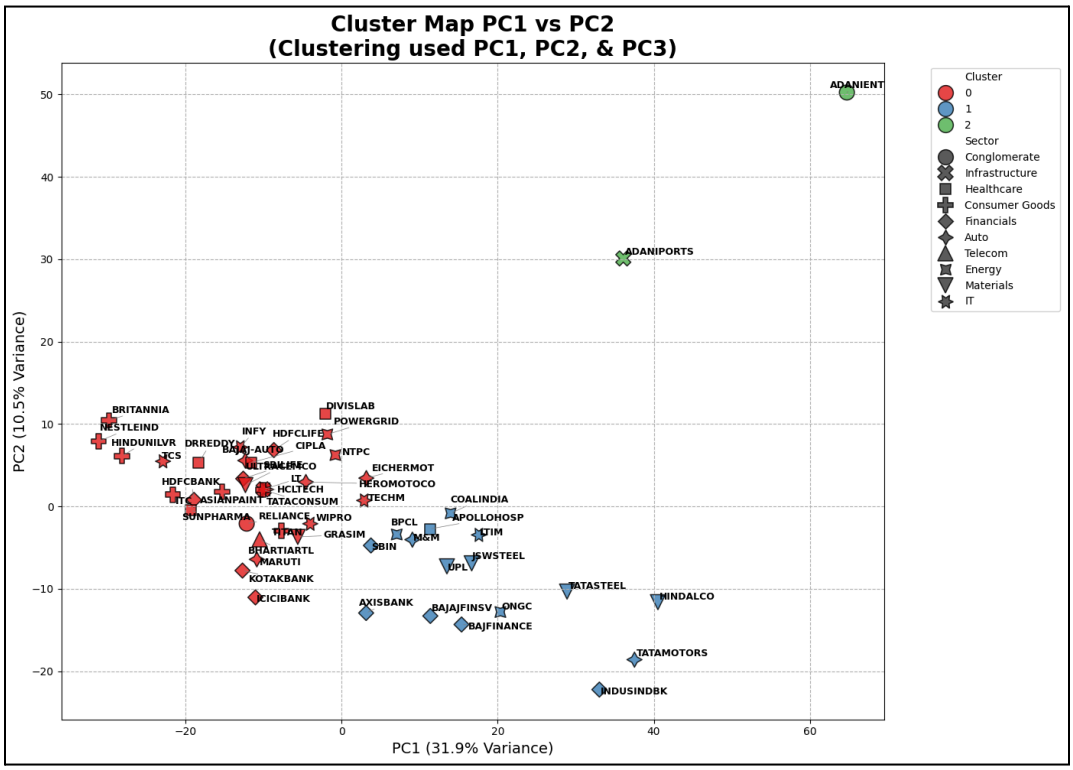


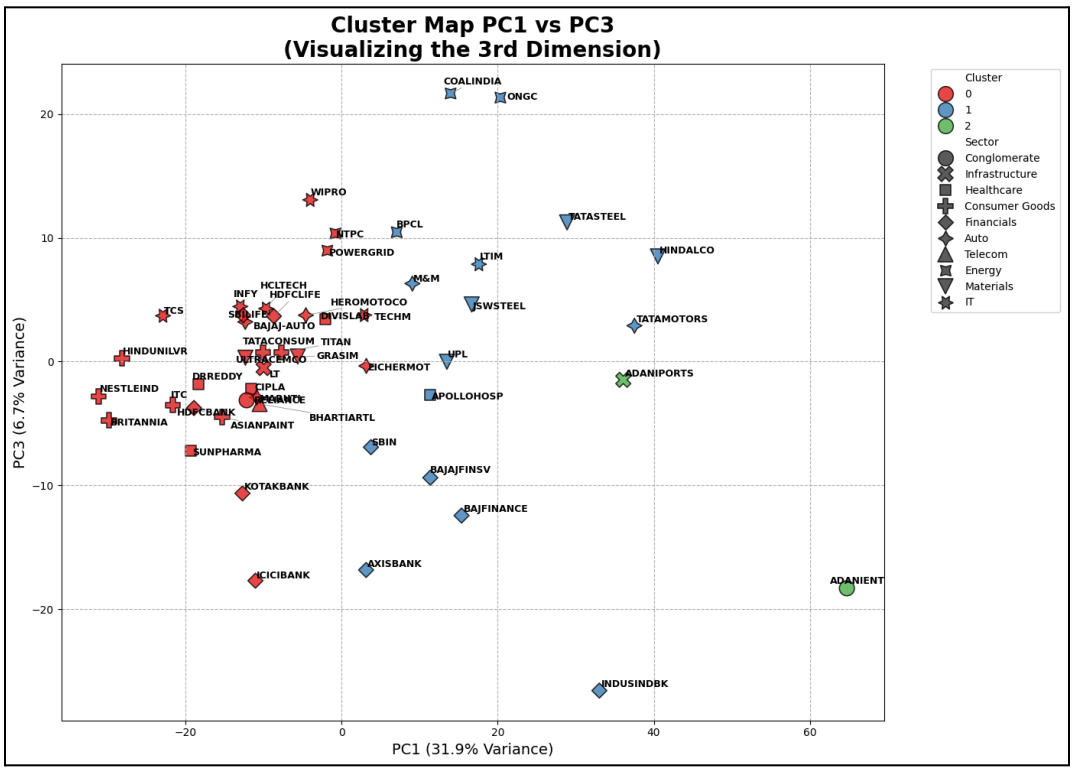
The noticeable volatility spikes in all clusters match significant market disruptions from 2020 to 2025. The biggest spike occurs during the COVID-19 crash, which shows panic selling and intense global uncertainty. The following peaks in 2021 and 2022 relate to inflation shocks, sharp interest-rate increases, and the Russia-Ukraine conflict, which hit commodity-linked stocks the hardest. Smaller, repeated spikes in 2023 and 2024 come from specific sector events and earnings cycles, especially in high-volatility industries. (Insights From MA634: Financial Risk Management)

4.2 PCA Visualizations

- The PCA maps show a clear separation among the three volatility clusters, confirming that the clustering algorithm identified real patterns in the data.
- PC1 (31.9% variance) mainly reflects the overall volatility level. It places high-volatility stocks like ADANIENT, ADANIPORTS, TATASTEEL, and HINDALCO far to the right.
- Low-volatility defensive stocks such as HDFCBANK, KOTAKBANK, HINDUNILVR, NESTLEIND, and BRITANNIA cluster closely on the left side, indicating similar and stable volatility behavior.

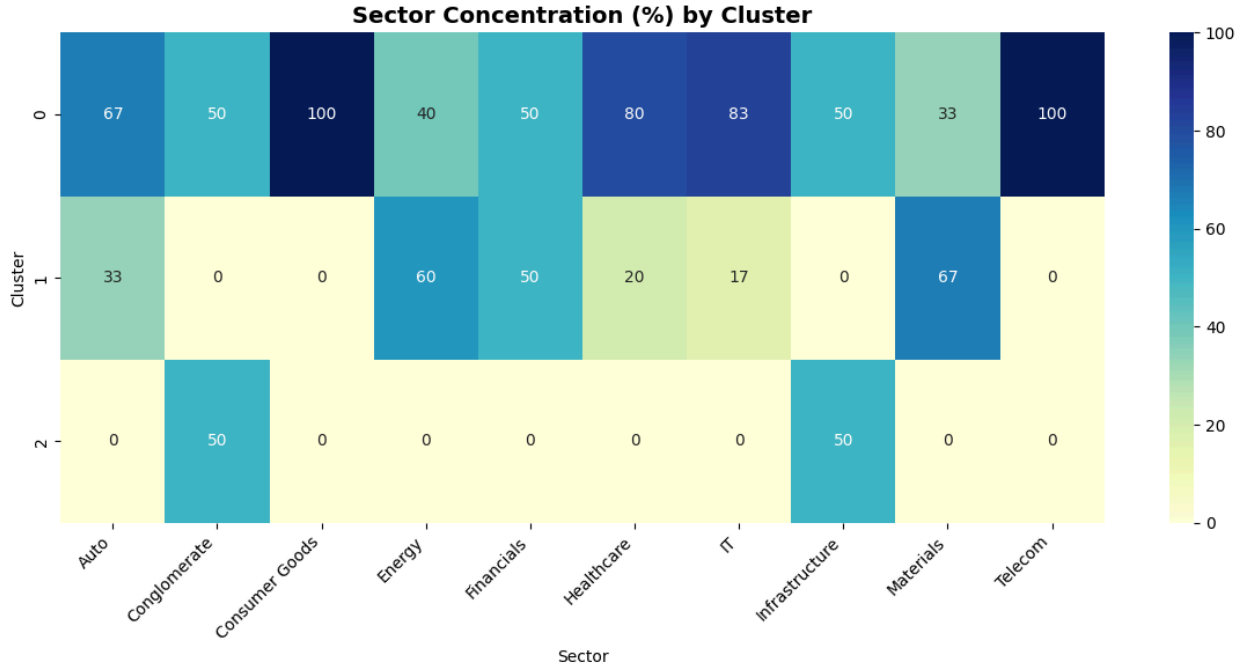
- PC2 and PC3 capture variations specific to sectors and individual stocks, which helps sharpen the separation within each cluster.
- IT, Consumer Goods, and Financials show close grouping within each sector, reflecting consistent volatility patterns.
- The distinct spatial grouping in both maps supports the 3-cluster solution, showing that volatility-based regimes are well-defined and relevant.





4.3 Sector Heatmap

- Cluster 0 (32 stocks): ASIANPAINT, BAJAJ-AUTO, BHARTIARTL, BRITANNIA, CIPLA, DIVISLAB, DRREDDY, EICHERMOT, GRASIM, HCLTECH, HDFCBANK, HDFCLIFE, HEROMOTOCO, HINDUNILVR, ICICIBANK, INFY, ITC, KOTAKBANK, LT, MARUTI, NESTLEIND, NTPC, POWERGRID, RELIANCE, SBILIFE, SUNPHARMA, TATACONSUM, TCS, TECHM, TITAN, ULTRACEMCO, WIPRO
- Cluster 1 (16 stocks): APOLLOHOSP, AXISBANK, BAJAJFINSV, BAJFINANCE, BPCL, COALINDIA, HINDALCO, INDUSINDBK, JSWSTEEL, LTIM, M&M, ONGC, SBIN, TATAMOTORS, TATASTEEL, UPL
- Cluster 2 (2 stocks): ADANIENT, ADANIPORTS



- **Cluster 0** shows the most diversification. It includes large portions of Consumer Goods, IT, Telecom, and Financials. This mix indicates stable sectors with low volatility.
- **Cluster 1** is mainly made up of Energy and Materials. These sectors are cyclical and have moderate volatility, responding strongly to changes in commodities and the broader economy.
- **Cluster 2** is very focused, featuring mostly Conglomerate and Infrastructure companies. These represent high-volatility and event-driven stocks.

The heatmap clearly shows that different sectors fall into various volatility groups. This highlights how sector behavior greatly affects clustering results.

Defensive sectors, like FMCG and IT, group together. In contrast, commodity-linked and infrastructure-heavy sectors make up the higher volatility categories.

5. Key Takeaways

1. Volatility across the sectors is quite different
The sectors associated with commodities like Energy and Materials exhibit major price fluctuations due to their quick reactions to global price changes and economic shocks.

2. Defensive sectors are the ones that bring stability
The lower volatility of the market is very much attributed to the presence of FMCG, Healthcare, and some Financials sectors as they are always there to stabilize the market.

3. PCA points to the main forces that drive volatility
The variation in the market is portrayed by a few principal components which indicates that the common factors like sentiment, sector cycles, and macro events are the ones that actually control the market volatility.

4. Clusters are in line with the major market occurrences
The significant volatility shifts at the cluster level are seen to be linked to events like COVID-19, inflation, interest-rate hikes, and geopolitical disturbances.

5. A tool for the making of investment decisions

Investors can use cluster-based insights to construct diversified portfolios, perform stress testing, identify rotation opportunities, manage risk dynamicall**y.**

6. Conclusion

The PCA-transformed rolling volatility patterns reveal that NIFTY50 stocks naturally group into three volatility-based regimes. The clustering results are not only economically significant but also statistically strong and visually distinct.

The extremely detailed methodological framework, together with careful interpretation of plots and results, indicates that volatility clustering is an important source of information about the market structure and the risk distribution.

Moreover, the research demonstrates the capability of quantitative methods like PCA and K-Means to transform the complex financial datasets into regime-based insights that are easy to understand and take action on.

The whole exercise enhanced our comprehension of:

- Financial volatility modeling
- Dimensionality reduction
- Machine learning clustering
- Market regime behavior