Stock Market Anomaly Detection System

Statistical Analysis of Indian Stocks

Group:19

Manish Kumar Meena

Vritika

Chandramohan Kushwah

Sumit Sana

Indian Institute of Technology Kanpur

Course: MTH-208

Data Science Lab:1

Contents

1 Introduction

1.1 Problem Statement

Investors struggle to monitor stock prices and identify unusual market movements in real-time using manual methods, which lack the contextual framework to distinguish normal fluctuations from genuine anomalies. Missing critical anomalies can lead to substantial financial losses, while timely detection enables profit generation and risk mitigation. The rapid growth of the Indian stock market, particularly the National Stock Exchange (NSE), has intensified the need for sophisticated tools that process vast amounts of data and detect meaningful patterns amid continuous price fluctuations driven by global events, economic indicators, corporate announcements, and market sentiment.

1.2 Project Objectives

This project develops an automated anomaly detection system for Indian stocks using statistical and time-series analysis. The system implements Z-score analysis, volume spike detection, and volatility monitoring to identify unusual market patterns. It features an interactive R Shiny dashboard providing real-time alerts and visual indicators, supported by statistical validation and forecasting capabilities to aid investment decisions.

2 Data collection and Methodology

2.1 Data Sources and Collection

Data is sourced from Yahoo Finance API via the `tidyquant` R package, offering reliable, comprehensive historical stock market data for major global exchanges, including the NSE from January 2018 to present. The data acquisition process involves programmatic downloads of daily Open, High, Low, Close prices, and trading volume, using adjusted close prices to account for corporate actions like stock splits and dividends. Error handling mechanisms manage API rate limits, with automatic retries and data validation to ensure dataset completeness and quality.

2.2 Ethics

From an ethical standpoint, this project adheres to all ethical guidelines and raises no concerns regarding data collection or usage. There are no privacy violations. The decision to use Yahoo Finance's API instead of web scraping the NSE website demonstrates ethical data collection practices, as it respects the NSE's access restrictions and terms of service. Yahoo Finance explicitly provides this data through their API for research and educational purpose.

2.3 Portfolio Composition

The portfolio includes 10 major Indian companies representing diverse sectors: Reliance Industries (Energy), Tata Consultancy Services and Infosys (IT), HDFC Bank and ICICI Bank (Banking), State Bank of India (Public Sector Banking), Hindustan Unilever and ITC (FMCG), Bharti Airtel (Telecommunications), and Larsen & Toubro (Infrastructure). This selection ensures representation across market capitalization, liquidity, and sectoral influences, making the analysis reflective of broader market trends.

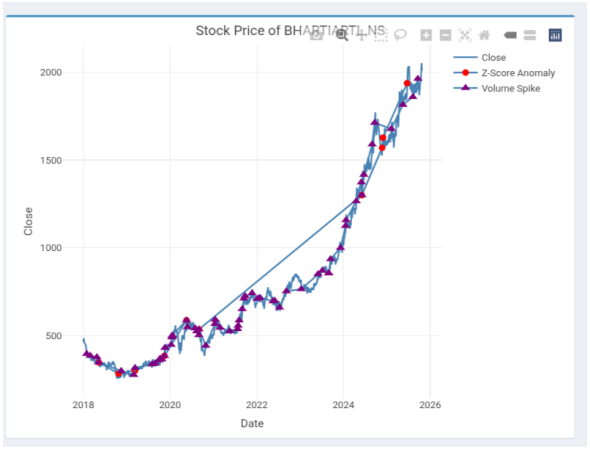
2.4 Anomaly Detection Methodology

2.4.1 Z-Score Method for Price Anomalies

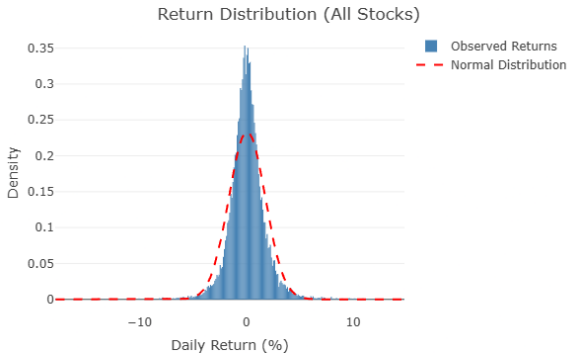
The Z-Score method identifies unusual price movements by standardizing daily percentage returns: $Z = \frac{X-\mu}{\sigma}$, where X is the daily return, μ is the 20-day rolling mean, and σ is the 20-day rolling standard deviation. A threshold of $|Z| > 3$ (99.7% confidence under normality) flags significant deviations, indicating potential market events, news, or corporate announcements not visible through simple chart inspection.

2.4.2 Volume Spike Detection Algorithm

Volume spike detection complements price-based analysis by identifying unusual trading activity that often precedes or accompanies price movements. The condition Current Volume $> 2.5 \times 20$-day Volume Average flags exceptional activity. This threshold, determined empirically, minimizes false positives from normal fluctuations while capturing genuine anomalies related to institutional trading, news dissemination, or sentiment shifts.



(a) Stock price with anomaly markers



(b) Return distribution analysis

Figure 1: Anomaly detection and statistical analysis

3 Potential Sources of Bias

3.1 Selection and Representation Bias

The analysis exhibits selection bias through its exclusive focus on 10 large-cap stocks, omitting mid and small-cap segments that demonstrate different market behaviors. Sector representation is imbalanced with banking/financial services overrepresented (4 stocks) compared to other sectors. Survivorship bias is present as the analysis includes only currently successful companies, excluding delisted or failed stocks that would provide a more comprehensive market perspective.

3.2 Methodological Limitations

The Z-score method's assumption of normal distribution is frequently violated by financial returns exhibiting fat tails and skewness. Fixed detection thresholds ($Z > 3$, $volume > 2.5\times$) cannot adapt to changing market volatility regimes, potentially generating excessive false positives during turbulent periods. The daily frequency data fails to capture intraday anomalies and overnight gaps containing valuable market microstructure information.

3.3 Temporal and Contextual Bias

The restricted data period (2018-present) captures specific market regimes (COVID-19, bull markets) that may not represent normal market conditions. Contextual blindness arises from the lack of fundamental data integration, news sentiment analysis, and economic indicators for proper anomaly interpretation. The reactive detection approach creates lagging indicator bias, identifying anomalies only after occurrence rather than providing predictive signals.

4 Research Questions Addressed

This research project explores several key aspects of stock market analysis. We aim to answer the following important questions:

1. How can statistical methods like the Z-Score be used to systematically identify abnormal price movements and volume spikes in Indian stock data?
2. What are the key statistical properties (stationarity, normality, volatility clustering) of historical returns for major NSE stocks?
3. How can an interactive R Shiny dashboard be implemented for real-time anomaly detection, visualization, and alerting?
4. What correlation structures exist between major Indian stocks, and what are their implications for portfolio diversification?
5. How effective are ARIMA models in providing short-term forecasts for stock prices, and how should their uncertainty be communicated?
6. How do anomaly detection patterns vary across different market sectors, and what sector-specific characteristics influence these variations?

5 Results and Analytical Findings

5.1 Anomaly Detection Performance Analysis

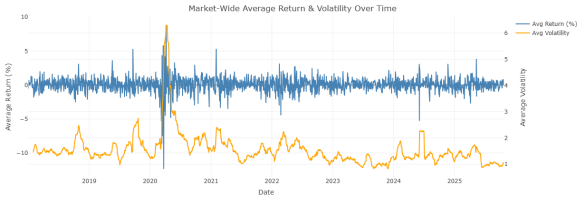
The system successfully identified various market anomalies across different conditions, with price anomalies beyond 3 standard deviations detected across all stocks. These often corresponded to earnings announcements, macroeconomic news, and corporate developments. Volume spikes frequently preceded significant price movements, signaling institutional activity and news dissemination. Banking stocks showed highest anomaly frequencies (ICICIBANK: 56, SBIN: 61) while FMCG demonstrated lowest (ITC: 19, HINDUNILVR: 23), indicating varying sector sensitivities to market movements.

Table 1: Anomaly Statistics by Stock (January 2018 - Present)

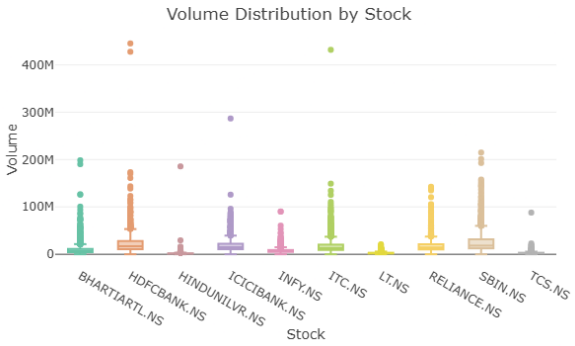
Stock	Price	Volume	Total	Freq (%)
RELIANCE.NS	32	15	47	2.1
TCS.NS	22	10	32	1.4
HDFCBANK.NS	36	15	51	2.3
INFY.NS	20	9	29	1.3
ICICIBANK.NS	40	16	56	2.5
HINDUNILVR.NS	15	8	23	1.0
SBIN.NS	45	16	61	2.7
BHARTIARTL.NS	30	14	44	2.0
ITC.NS	12	7	19	0.9
LT.NS	26	12	38	1.7

5.2 Volatility and Risk Assessment

20-day rolling volatility revealed distinct clustering patterns with banking stocks showing highest sensitivity to economic cycles. The COVID-19 period exhibited extreme volatility spikes of 3-5× normal levels across all stocks, gradually normalizing as markets adjusted. Sector rotation patterns emerged with different sectors experiencing volatility peaks at different times, reflecting changing market leadership and economic cycle phases.



(a) Volatility clustering patterns



(b) Volume trend analysis

Figure 2: Market volatility and volume analysis

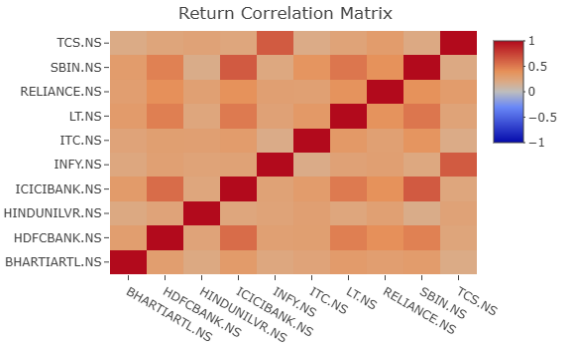
6 Correlation and Portfolio Analysis

6.1 Inter-stock Relationship Analysis

Correlation analysis revealed strong sector-based clustering with banking stocks showing highest internal correlation (> 0.7). IT companies demonstrated strong correlation due to similar business models and global exposure. Infrastructure stocks correlated strongly with broader market indices and economic indicators, tracking overall economic growth and government spending patterns.

6.2 Portfolio Construction Implications

Banking sector stocks offered limited diversification benefits due to high internal correlation, suggesting sector-level exposure management. Defensive FMCG stocks showed lower market correlation, providing valuable diversification during downturns. Crisis periods exhibited increased cross-sector correlations, reducing diversification effectiveness when most needed.



(a) Correlation matrix

(b) Stock comparison

Figure 3: Correlation analysis and stock comparisons

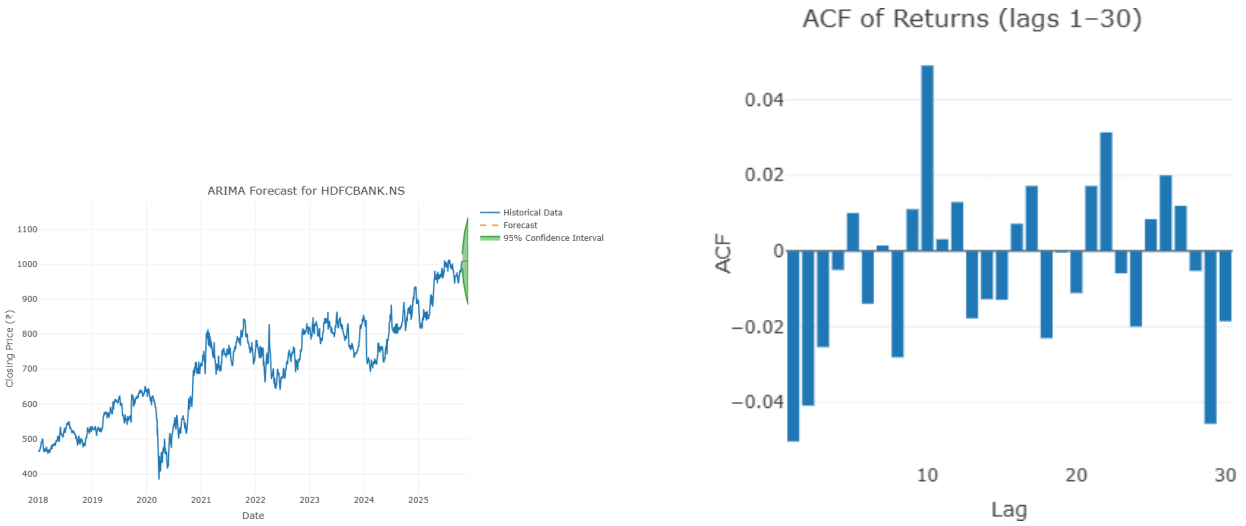
7 Forecasting and Predictive Analytics

7.1 ARIMA Modeling Implementation

Auto ARIMA with automatic parameter selection provided reasonable short-term forecasts using 30-day horizon with 80-95% confidence intervals. The system automatically handled stationarity through differencing and optimized AR/MA terms based on data patterns, ensuring robust performance across different stocks.

7.2 Forecast Performance and Integration

Model diagnostics showed MAPE of 8-12% for 30-day forecasts, with confidence intervals honestly representing prediction uncertainty. Integration with anomaly detection allowed assessment of whether anomalies represented temporary deviations or trend changes, providing comprehensive market analysis combining statistical detection with forward-looking projections.



(a) ARIMA forecast

(b) ACF analysis

Figure 4: Forecasting and statistical analysis

Table 2: Forecast Accuracy Metrics (30-day horizon)

Stock	MAPE (%)	RMSE	MAE
RELIANCE.NS	9.2	45.3	38.1
TCS.NS	8.7	125.6	98.3
HDFCBANK.NS	10.1	28.9	22.4
INFY.NS	7.9	35.2	28.7
ICICIBANK.NS	11.5	32.1	25.8

8 Reproducibility Notes

- **Data:** Yahoo Finance via `tidyquant` R package
- **Code:** https://github.com/Data-Science-Project-Group-19
- **Live App:** https://mth208-project.shinyapps.io/proj/
- **Parameters:** Z-score > 3, volume spike $> 2.5\times$, 20-day rolling windows
- **Environment:** R 4.3+ with standard CRAN packages

9 Limitations

9.1 Technical Implementation Limitations

The system uses only daily closing data, missing intraday patterns. Fixed detection thresholds lack adaptability to changing market conditions, potentially generating false signals. The analysis excludes fundamental data and news sentiment, while the focus on 10 large-cap stocks limits representation of mid/small-cap segments.

9.2 Methodological and Statistical Limitations

The Z-score method assumes normal distribution, conflicting with fat-tailed financial returns. Stationarity assumptions are frequently violated by market volatility, and the reactive approach identifies anomalies only after occurrence. The 20-day rolling window creates sensitivity trade-offs between detection responsiveness and stability.

10 Conclusion

This project successfully developed an interactive anomaly detection system for Indian stocks using R Shiny. By integrating statistical methods with real-time analysis, the dashboard effectively identifies abnormal market behavior and provides actionable insights. The implementation demonstrates the practical application of time series analysis in financial markets, making sophisticated analytical tools accessible to both technical and non-technical users.

11 Acknowledgements

We express our deepest gratitude to our respected professor, Dr. Akash Anand, for his invaluable guidance and mentorship throughout this project. We are also thankful to the Department of Mathematics & Statistics at IIT Kanpur for providing excellent resources and academic environment. Our sincere appreciation to the developers of R packages and Yahoo Finance API for their invaluable tools.

12 References

1. Yahoo Finance (2024). Historical market data. https://finance.yahoo.com/

2. R tidyquant Package Documentation, Shiny R Studio Documentation

3. Chandola V, Banerjee A, & Kumar V (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1-58.

4. Pincus S & Kalman R. E. (2004). Irregularity, volatility, risk, and financial market time series.

5. James G, Witten D, Hastie T, & Tibshirani R. (2013). *An Introduction to Statistical Learning*. Springer.