

# 1. Introduction

## 1.1 Motivation

Selecting attractive investments from thousands of publicly traded stocks is challenging. Investors face a complex task: evaluating multiple factors simultaneously to make informed choices. Current methods typically rely on separate metrics, like financial ratios or growth measures. These isolated metrics often miss the broader investment picture, and interpreting them collectively can be overwhelming. A structured solution that simplifies this complexity is necessary.

A composite indicator, combining multiple stock attributes into a single, understandable value, addresses this gap. It summarises diverse information into a straightforward metric, enabling easier comparison across many companies.

A well-designed composite indicator can:

* Highlight the relative attractiveness of stocks.
* Provide clarity amid overwhelming financial data.
* Enable investors and analysts to make quicker, better-informed decisions.

## 1.2 Objective

The Composite Stock Investment Attractiveness Index (CSIAI) is developed to provide a transparent, data-driven measure of stock attractiveness. It integrates multiple financial, market, and risk indicators into one composite score, enabling effective comparison across stocks. Transparency and replicability are fundamental features, ensuring the index can be validated, tested, and trusted by investors, analysts, and researchers.

Specifically, the CSIAI combines indicators from five distinct dimensions:

* Financial Strength
* Growth Potential
* Market Performance
* Risk & Volatility
* Liquidity & Trading Activity

Indicators within these dimensions were carefully selected, validated, and weighted using rigorous statistical methods such as Principal Component Analysis (PCA). Data preprocessing steps, like handling missing values (multiple imputation) and standardisation (Min-Max scaling), ensured robustness.

## 1.3 Target Audience

The primary users of the CSIAI include:

* **Investors**: To quickly screen stocks for potential investment opportunities based on clear, understandable metrics.
* **Financial Analysts**: To validate their analyses and augment detailed research with a robust, quantitative tool.
* **Fintech Developers**: For integration into financial apps and platforms, offering users reliable stock attractiveness ratings.
* **Academics**: For research on composite indicator methodologies and financial market analysis, leveraging the index’s transparency and reproducibility.

# 2. Theoretical Framework

## 2.1 Dimensions of Stock Attractiveness

The Composite Stock Investment Attractiveness Index (CSIAI) was structured around five carefully chosen dimensions, each representing critical areas influencing investment decisions. These dimensions were selected based on existing literature, practical relevance, and investor priorities. Each dimension captures distinct but complementary aspects of a stock’s overall attractiveness.

Financial Strength

Financial strength reflects a company's capacity to meet obligations and maintain operational stability. It provides insight into fundamental corporate health, incorporating metrics such as Return on Equity (ROE), Debt-to-Equity Ratio, Current Ratio, Operating Cash Flow, and EBITDA Margin. Indicators like ROE and EBITDA margin specifically highlight profitability and operational efficiency, essential for evaluating financial sustainability.

Growth Potential

Growth potential measures the company’s ability to expand and enhance profitability over time. Key metrics include Revenue Growth, Operating Margin, and Gross Margin. Revenue growth directly indicates market success and potential future profits. Operating and gross margins offer insights into operational efficiency and competitive advantages, vital for sustainable growth.

Market Performance

Market performance captures investor perception and market valuation. Indicators in this dimension include Earnings Per Share (EPS), Market Capitalization, Price-to-Sales Ratio, and Dividend Payout Ratio. EPS indicates profitability from a shareholder’s perspective, while market capitalization provides a sense of company size and market confidence. Price-to-Sales assesses valuation relative to revenue, and payout ratio signals financial maturity and shareholder value distribution.

Risk & Volatility

Risk & volatility evaluates the uncertainty and potential variability in stock returns. Indicators include Historical Volatility, Beta, Maximum Drawdown, Standard Deviation of Returns, and Value at Risk (VaR). Historical volatility and standard deviation reflect price variability, crucial for risk-averse investors. Beta measures market-related risk, whereas maximum drawdown and VaR assess downside risk which are all important for understanding extreme market conditions.

Liquidity & Trading Activity

Liquidity & trading activity assesses how easily investors can buy or sell a stock without impacting its price significantly. Metrics include Average Volume (30-day), Bid-Ask Spread, Volume Growth, and Float Shares. Higher liquidity indicates greater ease in trading and better market efficiency, essential for active investors and large trades.

## 2.2 Compensability Logic

The CSIAI used compensatory logic in its aggregation method. Both linear and geometric aggregations were applied to balance the trade-offs:

* **Linear aggregation** allows strong performance in one dimension to offset weaker results elsewhere. This method suits investors focused on overall strengths rather than penalizing isolated weaknesses.
* **Geometric aggregation** offers partial compensation. It rewards balanced performance across all dimensions and reduces the likelihood that extreme high scores in one dimension excessively influence the overall attractiveness score.

This provides flexibility, accommodating diverse investor strategies and perspectives regarding risk tolerance and balanced performance.

## 2.3 Indicator Selection Logic

The chosen indicators align with finance theory and practical best practices in investment analysis:

* **Return on Equity (ROE)** reflects company profitability relative to shareholder investment. It is widely recognized for gauging effective management and long-term growth.
* **Debt-to-Equity and Current Ratio** assess financial stability and solvency, critical to managing long-term viability and operational risks.
* **Revenue Growth and Operating Margins** highlight core business success and operational efficiency, directly influencing growth potential.
* **Earnings Per Share (EPS) and Market Capitalization** connect financial performance to market valuation, essential for comparing company value.
* **Historical Volatility, Beta, and Value at Risk (VaR)** offer direct insights into market-driven risks and uncertainty, essential considerations for cautious investors.
* **Average Trading Volume and Bid-Ask Spread** assess market liquidity, essential for evaluating transaction ease and costs.

# 3. Data Selection

## 3.1 Source

Data for constructing the Composite Stock Investment Attractiveness Index (CSIAI) was collected entirely from **Yahoo Finance** using the Python library **yfinance**. Yahoo Finance provides reliable, publicly accessible datasets covering key financial fundamentals, historical price data, and trading volumes. These datasets were selected based on:

* Transparency and reproducibility for academic use.
* Reliability due to extensive market coverage.
* Regular updating frequency.

## 3.2 Time Frame

The data period selected covers from **January 1, 2023**, to **May 6, 2025**. This timeframe ensures the CSIAI is both current and reflective of contemporary market conditions. The choice of a longer period also facilitates robust statistical analyses.

## 3.3 Initial Universe and Filtering

The initial universe for the analysis consisted of all constituents of the **Russell 3000 Index**. This index includes approximately 3,000 large and mid-sized U.S. companies, providing broad representation across various market sectors.

Key filtering steps included:

* **Market liquidity threshold:**  
  Stocks were filtered based on a minimum **average 30-day trading volume of 50,000 shares**. It serves multiple purposes:
  + Ensures the **Liquidity & Trading** dimension is not dominated by thinly traded stocks.
  + Maintains approximately 80% of the original Russell 3000 constituents.
  + Reduces estimation errors in calculating bid-ask spreads, essential for accurate liquidity measurement.
* **Indicator coverage check:**  
  Each selected indicator had to meet stringent criteria:
  + At least **90% data availability** per ticker.
  + Overall, a minimum of **80% coverage** for each indicator across all stocks.
  + Stocks failing these completeness thresholds were excluded to maintain dataset integrity.

## 3.4 Statistical Quality Principles

The inclusion of any indicator was governed by adherence to seven quality principles. Each indicator selected met all these requirements, as outlined:

* **Relevance:** Directly aligns with the conceptual framework for stock attractiveness.
* **Accuracy:** Sourced exclusively from audited financial statements or direct market data from Yahoo Finance.
* **Timeliness:** Regular updating, with at least quarterly data availability daily preferred, aligning with the discussion with Dr. John Loane, ensuring the index is always current.
* **Accessibility:** Data available freely via Yahoo Finance (using yfinance).
* **Interpretability:** Indicator units and directions must be clear, logical, and easily understood by both financial professionals and general investors.
* **Comparability:** Each indicator must be relevant and applicable across various sectors and company types within the Russell 3000.
* **Coherence:** Indicator definitions are consistent and do not conflict with other selected indicators or standard financial metrics.

Indicators failing any of these criteria were either excluded or substituted with proxies accompanied by clear justification.

## 3.5 Final Indicator List and Rationale

The final set of indicators selected reflects a balance between financial theory, empirical best practices, and practical data availability. They were categorized under five key dimensions:

* **Financial Strength**
* **Growth Potential**
* **Market Performance**
* **Risk & Volatility**
* **Liquidity & Trading**

The rationale behind selecting these dimensions and their indicators was to create a comprehensive yet clear measure of stock attractiveness. For example:

* **ROE** was selected for **Financial Strength** to capture profitability and managerial efficiency.
* **Revenue growth** and **Operating Margin** under **Growth Potential** indicate forward-looking performance and operational efficiency.
* **EPS** and **Price-to-sales ratio** were chosen under **Market Performance** to reflect valuation and profitability measures critical to investors.
* **Historical volatility** and **Value at Risk (VaR)** measure downside risk, aligning closely with investor priorities for safety under **Risk & Volatility**.
* Indicators such as **average volume** and **bid-ask spread** ensure market liquidity and trade efficiency under the **Liquidity & Trading** dimension.

**Table 1: Final Indicator List**

| **Dim.** | **Indicator** | **yfinance field / derivation** | **Justification** |
| --- | --- | --- | --- |
| **Financial Strength** | Return on Equity | info['returnOnEquity'] | Profitability per unit equity. |
|  | Debt-to-Equity | info['debtToEquity'] | Capital structure risk. |
|  | Current Ratio | info['currentRatio'] | Short-term solvency. |
|  | Quick Ratio | info['quickRatio'] | Acid-test liquidity. |
|  | Oper. Cash Flow | info['operatingCashflow'] | Cash backing earnings. |
| **Growth Potential** | Revenue Growth | YoY revenue trend | Top-line expansion. |
|  | Earnings Growth | YoY EPS trend | Bottom-line expansion. |
|  | Operating Margin | info['operatingMargins'] | Efficiency scaling. |
|  | Gross Margin | info['grossMargins'] | Core pricing power. |
|  | Analyst Rating | info['recommendation Mean'] | External sentiment. |
| **Market Performance** | P/E Ratio | info['trailingPE'] | Classic valuation. |
|  | EPS | info['trailingEps'] | Profit per share. |
|  | P/B Ratio | info['priceToBook'] | Asset-based valuation. |
|  | Dividend Yield | info['dividendYield'] | Income return. |
|  | Market Cap | info['marketCap'] | Size proxy. |
| **Risk & Volatility** | 30-d Hist. Volatility | std(returns) | Short-term risk. |
|  | Beta | info['beta'] | Market sensitivity. |
|  | Sharpe Ratio | excess return / vol | Risk-adjusted perf. |
|  | Max Draw-down | roll-min formula | Tail risk. |
|  | Std Dev Returns | std(returns) | Dispersion measure. |
| **Liquidity & Trading** | Avg 30-d Volume | rolling mean | Depth of book. |
|  | Bid-Ask Spread\* | (High−Low)/Mid | Transaction cost proxy. |
|  | Volume Growth | pct\_change Volume | Activity momentum. |
|  | Shares Outstanding | info['sharesOutstanding'] | Supply side. |
|  | Float Shares | info['floatShares'] | Free float liquidity. |

# 4. Imputation of Missing Data

High-quality composite indicators depend significantly on complete and accurate data. However, missing values are almost inevitable in real-world financial datasets. Proper handling of missing data ensures the reliability and interpretability of the final Composite Stock Investment Attractiveness Index (CSIAI). This section details the procedures used to diagnose, treat, and validate missing data within the CSIAI dataset.

## 4.1 Missingness Diagnosis

Before choosing the imputation method, a clear understanding of data completeness was essential. The dataset obtained from Yahoo Finance exhibited varying levels of missingness. To quantify this clearly, the share of missing values for each indicator was calculated and visualized. This approach facilitated a straightforward evaluation of missing data patterns.

The figure below shows the share of missing data across indicators that were incomplete:

A graph of a number of data

AI-generated content may be incorrect.

Key observations from the diagnosis were:

* Significant missingness appeared in certain financial indicators such as totalAssets.
* Price-based indicators were mostly complete due to their daily availability.
* Indicators with over 30% missingness posed a challenge, risking imputation accuracy. Hence, a strict coverage cutoff was applied to ensure robust imputations.

## 4.2 Deriving Price-Based Indicators

To enrich the dataset, several price-based indicators were calculated using historical market data. These indicators were derived directly from stock prices and trading volumes, ensuring accuracy and availability:

* **Historical Volatility** (30-day annualized volatility)
* **Sharpe Ratio** (excess return adjusted by volatility)
* **Value at Risk** (VaR, calculated at 95% confidence level)
* **Average Daily Volume** (30-day average)
* **Turnover Ratio** (average daily volume relative to float shares)

These price-based indicators provided valuable insights into market behaviour, complementing traditional accounting measures.

## 4.3 Imputation Strategy

Given the nature and extent of missingness, multiple imputation was identified as a suitable approach. An Iterative Imputer using Bayesian Ridge Regression was implemented, chosen specifically for its statistical robustness and efficiency.

Brief justification for the selection:

* Iterative Imputer effectively captures complex relationships among variables, enhancing imputation accuracy.
* Bayesian Ridge regression handles uncertainty by estimating posterior distributions, not just single-point estimates.

The chosen approach involved:

* Executing five separate rounds of multiple imputation. Each round slightly varied the random seed to ensure diverse imputed datasets.
* Generating a collection of plausible datasets rather than a single deterministic result, reducing the bias associated with imputation.

## 4.4 Rubin’s Rule Averaging

Multiple imputation generated five independent datasets. Rather than arbitrarily selecting one, Rubin's Rule was applied to average these datasets into a single consolidated dataset. Rubin's Rule combines multiple imputations by:

* Averaging across the multiple imputed values for each missing observation.
* Accounting explicitly for uncertainty inherent in the missingness.

Using Rubin’s Rule ensured that imputed data points reflected the full range of plausible values. This averaging step significantly enhanced the credibility and stability of the imputed values used in subsequent analyses.

## 4.5 Diagnostics and Validation

After imputation, assessing the adequacy and validity of the results was critical. Two diagnostics were applied:

* **Kaiser-Meyer-Olkin (KMO)**: This measure evaluated the appropriateness of using the imputed data in factor analysis, verifying that the data maintained good structure post-imputation.
* **Coverage Cutoff (90%)**: A stringent threshold was established to remove any ticker lacking sufficient data, ensuring the quality of the final dataset.

Additionally, outlier analysis was performed to reduce the influence of extreme values:

* Indicators were bounded between the 1st and 99th percentiles.
* Ensured imputed data points were realistic and consistent with observed values.

**Summary of Imputation Approach:**

* Method: Iterative Imputer with Bayesian Ridge
* Rounds of imputation: 5
* Combination method: Rubin's Rule
* Coverage threshold: ≥ 90% data completeness required per ticker
* Outlier management: bounded between 1st and 99th percentiles