

# Executive Summary

The Composite Stock Investment Attractiveness Index (CSIAI) provides a systematic, data-driven assessment of investment potential for U.S. equities within the Russell 3000 universe. It quantifies stock attractiveness based on five financial dimensions:

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

A comprehensive selection of 21 indicators supports these dimensions, derived primarily from Yahoo Finance data for January 2023 through May 2025. Criteria applied for indicator selection included relevance, interpretability, comparability, and regular availability of data.

Key analytical methods employed are as follows:

* Missing data was handled by iterative multiple imputation (Bayesian Ridge regression), averaging five complete datasets using Rubin's rule.
* Multivariate analysis, including Principal Component Analysis (PCA), confirmed indicator suitability. PCA loadings provided a variance-based weighting scheme alternative to the equal-by-group method.
* Normalisation procedures involved a Yeo–Johnson transformation for skewed variables and Min-Max scaling across all indicators, ensuring scores ranged from 0 to 1.
* Weighting and aggregation included linear aggregation as the primary approach, complemented by geometric aggregation to test robustness. Greater weight (30%) was deliberately assigned to Risk & Volatility due to its significance in investor risk assessments.

Correlations between the final CSIAI and established indices (SPY, QUAL, MTUM) were high (Spearman's correlations ranged between 0.66 and 0.79), validating the index's practical relevance and alignment with market measures. Visualisation of results included radar charts and bar graphs clearly displaying stock attractiveness rankings.

Despite these strengths, limitations exist. The CSIAI currently represents a static snapshot rather than a dynamic trend analysis. Sensitivity analysis regarding choices made during construction was also not conducted.

Table of Contents

[Executive Summary 2](#_Toc197683674)

[1. Introduction 5](#_Toc197683675)

[1.1 Motivation 5](#_Toc197683676)

[1.2 Objective 5](#_Toc197683677)

[1.3 Target Audience 5](#_Toc197683678)

[2. Theoretical Framework 6](#_Toc197683679)

[2.1 Dimensions of Stock Attractiveness 6](#_Toc197683680)

[2.2 Compensability Logic 7](#_Toc197683681)

[2.3 Indicator Selection Logic 7](#_Toc197683682)

[3. Data Selection 8](#_Toc197683683)

[3.1 Source 8](#_Toc197683684)

[3.2 Time Frame 8](#_Toc197683685)

[3.3 Initial Universe and Filtering 8](#_Toc197683686)

[3.4 Statistical Quality Principles 9](#_Toc197683687)

[3.5 Final Indicator List and Rationale 10](#_Toc197683688)

[4. Imputation of Missing Data 12](#_Toc197683689)

[4.1 Missingness Diagnosis 13](#_Toc197683690)

[4.2 Deriving Price-Based Indicators 13](#_Toc197683691)

[4.3 Imputation Strategy 14](#_Toc197683692)

[4.4 Rubin’s Rule Averaging 14](#_Toc197683693)

[4.5 Diagnostics and Validation 14](#_Toc197683694)

[5. Multivariate Analysis 15](#_Toc197683695)

[5.1 Scatterplots and Pairwise Correlation 15](#_Toc197683696)

[5.2 PCA 43](#_Toc197683697)

[5.3 KMO and Bartlett Tests 47](#_Toc197683698)

[5.4 Indicator Clustering 48](#_Toc197683699)

[5.5 Final Indicator Set 51](#_Toc197683700)

[Summary and Conclusion 51](#_Toc197683701)

[6. Normalisation 51](#_Toc197683702)

[6.1 Min-Max Scaling 51](#_Toc197683703)

[6.2 Power Transformations (Yeo - Johnson) 52](#_Toc197683704)

[7. Weighting and Aggregation 52](#_Toc197683705)

[7.1Weighting Models 52](#_Toc197683706)

[7.2Aggregation Methods 54](#_Toc197683707)

[7.3Robustness Checks 55](#_Toc197683708)

[8. Link to Other Indices 56](#_Toc197683709)

[8.1Benchmark Selection 56](#_Toc197683710)

[8.2Correlation Comparison 57](#_Toc197683711)

[8.3Interpretation 58](#_Toc197683712)

[9. Visualisation of Results 58](#_Toc197683713)

[9.1 Top 20 CSIAI Table 58](#_Toc197683714)

[9.2 Horizontal Bar Chart of Top 10 CSIAI Scores 59](#_Toc197683715)

[9.3 Radar Charts of Sub-Index Profiles 60](#_Toc197683716)

[9.4 Correlation Heatmap of Sub-Index Scores 62](#_Toc197683717)

[9.5 Decomposition Bar Charts for Top and Bottom Firms 62](#_Toc197683718)

[10. Conclusion 63](#_Toc197683719)

[10.1 Key Takeaways 63](#_Toc197683720)

[10.2 Strengths 63](#_Toc197683721)

[10.3 Limitations 64](#_Toc197683722)

[10.4 Future Work 64](#_Toc197683723)

[References 64](#_Toc197683724)

GitHub Link - <https://github.com/PatrickOrjieh/csiai-project>

# 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

## 5. Multivariate Analysis

The multivariate analysis phase assesses indicator relationships, redundancy, and underlying structure. The chosen methods include scatterplots, pairwise correlations, principal component analysis (PCA), factorability tests (KMO and Bartlett), and clustering. Results informed the final indicator selection, ensuring robustness and clarity of the Composite Stock Investment Attractiveness Index (CSIAI).

## 5.1 Scatterplots and Pairwise Correlation

Scatterplots were generated for each sub-index group to visually identify relationships between indicators. These pairwise plots enabled quick identification of linear associations, non-linear relationships, or potential redundancy.

Financial Strength

A graph of a financial strength

AI-generated content may be incorrect.A graph with blue dots and a red line

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect.A graph with blue and orange dots

AI-generated content may be incorrect.

A graph with blue dots and a red line

AI-generated content may be incorrect.A graph with a line and a line

AI-generated content may be incorrect.

A graph of a financial strength

AI-generated content may be incorrect.A graph of a graph with numbers and a chart of data

AI-generated content may be incorrect.

A graph of a financial strength

AI-generated content may be incorrect.A graph with numbers and a chart

AI-generated content may be incorrect.

A graph of a financial strength

AI-generated content may be incorrect.A graph with blue dots and a red line

AI-generated content may be incorrect.

A graph of a financial strength

AI-generated content may be incorrect.A graph of a financial strength

AI-generated content may be incorrect.

A graph of a financial strength

AI-generated content may be incorrect.A graph of a financial strength

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.A graph of a financial strength

AI-generated content may be incorrect.

A graph of a graph with blue dots

AI-generated content may be incorrect.A graph with blue and orange lines

AI-generated content may be incorrect.

A graph of a financial strength

AI-generated content may be incorrect.

Financial Strength – ROE vs Other Indicators

* **Debt-to-Equity vs ROE:** Shows almost no linear relationship.
* **Current Ratio vs ROE:** Slight negative trend more liquidity, marginally lower ROE.
* **Quick Ratio vs ROE:** Similar to current ratio, a mild inverse relationship.
* **Operating Cash Flow vs ROE:** Moderate positive association firms with higher cash flow tend to have higher ROE.
* **Total Revenue vs ROE:** Weak positive slope larger firms slightly more profitable on average.
* **EBITDA Margin vs ROE:** Strong positive correlation higher margins align with higher ROE.

Financial Strength – Debt-to-Equity vs Other Indicators

* **Debt-to-Equity vs ROE:** Almost no linear relationship.
* **Debt-to-Equity vs Current Ratio:** Slight inverse trend higher leverage corresponds to lower short-term liquidity.
* **Debt-to-Equity vs Quick Ratio:** Mild negative slope, just like the current ratio pattern.
* **Debt-to-Equity vs Operating Cash Flow:** Very weak association cash flow levels do not predict leverage.
* **Debt-to-Equity vs Total Revenue:** No clear pattern company scale appears uncorrelated with debt ratio.
* **Debt-to-Equity vs EBITDA Margin:** Small positive slope firms with higher operational profitability tend to carry slightly more debt.

Financial Strength – Current Ratio vs Other Indicators

* **Debt-to-Equity vs Current Ratio:** Very weak negative relationship more liquid firms tend to carry marginally less leverage.
* **Quick Ratio vs Current Ratio:** Near-perfect positive linear relationship as expected since quick ratio is a subset of current ratio.
* **Operating Cash Flow vs Current Ratio:** Slight negative trend firms with higher short-term liquidity often show modestly lower operating cash generation.
* **Total Revenue vs Current Ratio:** Liquidity does not reliably indicate company size.
* **EBITDA Margin vs Current Ratio:** Mild inverse relationship higher liquidity is weakly associated with lower profit margins.

Financial Strength – Quick Ratio vs Other Indicators

* **Quick Ratio vs ROE:** Slight inverse relationship, suggesting firms with higher quick ratios tend to have marginally lower ROE.
* **Quick Ratio vs Debt-to-Equity:** Mild negative trend, indicating that more liquid firms carry slightly less gearing.
* **Quick Ratio vs Current Ratio:** Very strong positive linear association, reflecting the shared underlying liquidity measure.
* **Quick Ratio vs Operating Cash Flow:** There is no relationship, implying cash flow generation is orthogonal to short-term liquidity.
* **Quick Ratio vs Total Revenue:** Minimal inverse slope, larger firms show marginally lower quick ratios.
* **Quick Ratio vs EBITDA Margin:** Slight negative correlation, firms with higher margins tend to carry marginally lower quick ratios.

Financial Strength – Operating Cash Flow vs Other Indicators

* **Operating Cash Flow vs ROE:** Moderate positive slope companies with higher cash generation tend to have higher returns on equity.
* **Operating Cash Flow vs Debt-to-Equity:** Near-flat trend leverage levels don’t strongly predict cash-flow strength.
* **Operating Cash Flow vs Current Ratio:** Very weak relationship short-term liquidity (current ratio) isn’t a reliable proxy for cash generation.
* **Operating Cash Flow vs Quick Ratio:** Likewise minimal correlation liquidity and cash flow are largely independent.
* **Operating Cash Flow vs Total Revenue:** Clear positive association larger revenues usually comes with stronger operating cash flows.
* **Operating Cash Flow vs EBITDA Margin:** Mild positive trend firms with higher margins often report healthier cash flows.

Financial Strength – Total Revenue vs Other Indicators

* **ROE vs Total Revenue:** Weak positive slope I observe that larger firms tend to report slightly higher ROE on average.
* **Debt-to-Equity vs Total Revenue:** Almost flat relationship firm size has minimal impact on leverage levels in my sample.
* **Current Ratio vs Total Revenue:** Slightly negative trend as companies grow, their working-capital ratios marginally decrease.
* **Quick Ratio vs Total Revenue:** Similar to current ratio, a mild inverse association between size and quick liquidity.
* **Operating Cash Flow vs Total Revenue:** Strong positive correlation firms with higher revenue generate proportionally greater cash flows.
* **EBITDA Margin vs Total Revenue:** Very modest positive trend larger firms show marginally better operational efficiency.

Financial Strength – EBITDA Margin vs Other Indicators

* **ROE vs EBITDA Margin:** Clear positive relationship companies with higher EBITDA margins tend to have higher ROE.
* **Debt-to-Equity vs EBITDA Margin:** Very weak positive slope leverage shows almost no consistent impact on operating efficiency.
* **Current Ratio vs EBITDA Margin:** Slight negative trend more liquid firms exhibit marginally lower EBITDA margins.
* **Quick Ratio vs EBITDA Margin:** Similar mild inverse association to current ratio.
* **Operating Cash Flow vs EBITDA Margin:** Moderate positive correlation firms generating more cash flow often report stronger margins.
* **Total Revenue vs EBITDA Margin:** Weak positive slope larger firms on average display slightly better EBITDA margins.

Growth Potential

A graph with numbers and lines

AI-generated content may be incorrect.A graph with blue and orange lines

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.A graph of growth and revenue

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.A graph with numbers and a red line

AI-generated content may be incorrect.

A graph with blue and orange lines

AI-generated content may be incorrect.A graph of growth potential

AI-generated content may be incorrect.

A graph with blue and orange dots

AI-generated content may be incorrect.A graph with blue and red lines

AI-generated content may be incorrect.

Growth Potential – Revenue Growth vs Other Indicators

* **Operating Margin vs Revenue Growth:** Almost no linear association revenue growth explains very little variation in operating margins.
* **Gross Margin vs Revenue Growth:** Slight negative trend faster‐growing firms exhibit marginally lower gross margins on average.
* **R&D to Revenue vs Revenue Growth:** Flat no systematic relationship between R&D intensity and top‐line growth.
* **Analyst Rating vs Revenue Growth:** Weak negative slope higher growth stocks tend to have marginally lower analyst ratings.

Growth Potential – Operating Margin vs Other Indicators

* **Operating Margin vs Revenue Growth:** Shows almost no linear relationship, suggesting that revenue changes do not predict operating margin.
* **Gross Margin vs Operating Margin:** Moderate positive correlation firms with higher gross margins generally exhibit higher operating margins.
* **R&D-to-Revenue vs Operating Margin:** Strong negative association companies that invest more in R&D (relative to revenue) tend to have lower operating margins.
* **Analyst Rating vs Operating Margin:** Little to no clear relationship, indicating that ratings are not driven by operating margin levels.

Growth Potential – Gross Margin vs Other Indicators

* **Gross Margin vs Revenue Growth:** Very weak negative slope suggesting faster‐growing companies do not have higher gross margins.
* **Gross Margin vs Operating Margin:** Mild positive trend, as firms with higher gross margins tend to convert more of that into operating profit.
* **Gross Margin vs R&D to Revenue:** Flat, indicating R&D has little direct relation to gross margin levels.
* **Gross Margin vs Analyst Rating:** No linear association, implying analyst ratings are not driven by gross margin differences.

Growth Potential – R&D / Revenue vs Other Indicators

* **Revenue Growth vs R&D / Revenue:** Almost no relationship companies with higher R&D intensity do not necessarily grow faster.
* **Operating Margin vs R&D / Revenue:** Strong negative trend as R&D share rises, current margins tend to be lower which might be a sign of reinvestment.
* **Gross Margin vs R&D / Revenue:** Slight negative slope higher R&D intensity often accompanies thinner gross margins.
* **Analyst Rating vs R&D / Revenue:** No clear pattern analyst rating does not vary meaningfully with R&D intensity.

Growth Potential – Analyst Rating vs Other Indicators

* **Analyst Rating vs Revenue Growth:** Almost no trend analyst ratings have no relation to past revenue growth.
* **Analyst Rating vs Operating Margin:** Flat relationship operating profitability doesn’t map to analyst ratings.
* **Analyst Rating vs Gross Margin:** No systematic link firms with higher gross margins are neither consistently more nor less favorably rated.
* **Analyst Rating vs R&D-to-Revenue:** Near zero slope R&D intensity has negligible impact on analyst scores.

Market Performance

A graph with blue dots and a line

AI-generated content may be incorrect.A graph with numbers and a line

AI-generated content may be incorrect.

A graph with blue and orange dots

AI-generated content may be incorrect.A graph with numbers and lines

AI-generated content may be incorrect.

A graph of a market performance

AI-generated content may be incorrect.A graph with blue dots and orange lines

AI-generated content may be incorrect.

A graph with blue dots and numbers

AI-generated content may be incorrect.A graph with numbers and lines

AI-generated content may be incorrect.

A graph of a graph with numbers and a line

AI-generated content may be incorrect.A graph with numbers and points

AI-generated content may be incorrect.

A graph with numbers and points

AI-generated content may be incorrect.A graph with numbers and a line

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.

A graph with numbers and a line

AI-generated content may be incorrect.

Market Performance – PB Ratio vs Other Indicators

* **EPS vs PB Ratio:** Essentially flat relationship share valuation does not track earnings per share.
* **Market Cap vs PB Ratio:** Slight positive slope larger firms tend to trade at marginally higher price-to-book ratios.
* **EV/EBITDA vs PB Ratio:** No meaningful trend enterprise multiples are independent of book-value multiples.
* **Price-to-Sales vs PB Ratio:** Almost zero slope revenue multiples and book multiples move independently in this.
* **Payout Ratio vs PB Ratio:** Higher dividend payout may be linked to slightly lower valuation multiples.

Market Performance – EPS vs Other Indicators

* **PB Ratio vs EPS:** Almost flat relationship EPS does not meaningfully explain variations in price-to-book multiples.
* **Market Cap vs EPS:** Mild positive trend firms with higher EPS tend to have slightly larger market capitalizations.
* **EV/EBITDA vs EPS:** No clear pattern enterprise valuation multiples appear independent of EPS.
* **Price-to-Sales vs EPS:** No slope sales multiples are not driven by EPS levels.
* **Payout Ratio vs EPS:** Slight negative association companies with higher EPS pay out a marginally smaller proportion of earnings.

Market Performance – Market Cap vs Other Indicators

* **PB Ratio vs Market Cap:** Weak positive slope larger firms tend to have slightly higher price-to-book ratios.
* **EPS vs Market Cap:** Mild positive trend larger firms tend to report higher EPS.
* **EV/EBITDA vs Market Cap:** No clear relationship enterprise multiples do not correlate with market capitalization.
* **Price-to-Sales vs Market Cap:** No slope sales multiples are not driven by market capitalization.
* **Payout Ratio vs Market Cap:** No clear pattern larger firms do not necessarily have higher or lower payout ratios.

Market Performance – EV/EBITDA vs Other Indicators

* **PB Ratio vs EV/EBITDA:** Very flat relationship with a slight downward tilt at higher EV/EBITDA values, indicating little correlation.
* **EPS vs EV/EBITDA:** No slope, showing that earnings per share do not explain enterprise multiples.
* **Market Cap vs EV/EBITDA:** Nearly horizontal line, suggesting market capitalization is independent of EV/EBITDA.
* **Price-to-Sales vs EV/EBITDA:** Mostly flat but with a mild downward trend at the right tail, implying weak inverse association.
* **Payout Ratio vs EV/EBITDA:** Almost perfectly flat, indicating no relationship between payout policy and enterprise valuation multiples.

Market Performance – Price-to-Sales vs Other Indicators

* **PB Ratio vs Price-to-Sales:** Largely flat with a gentle downward curve at high P/S values indicating a weak inverse relationship.
* **EPS vs Price-to-Sales:** Almost horizontal but bending slightly downward on the right suggesting minimal negative association.
* **Market Cap vs Price-to-Sales:** Straight trend with a mild downward tilt as P/S increases showing little correlation.
* **EV/EBITDA vs Price-to-Sales:** Nearly flat overall, with a small negative slope on the tail implying a slight inverse link.
* **Price-to-Sales vs Payout Ratio:** Mostly level but bending down at higher P/S denoting a negligible negative relationship.

Market Performance – Payout Ratio vs Other Indicators

* **PB Ratio vs Payout Ratio:** Essentially flat indicating no clear relationship.
* **EPS vs Payout Ratio:** Slight downward bend on the right suggesting a minimal negative association.
* **Market Cap vs Payout Ratio:** Nearly straight with a modest downward curve at high payout ratios implying a weak inverse link.
* **EV/EBITDA vs Payout Ratio:** Straight horizontal trend showing no apparent correlation.
* **Price-to-Sales vs Payout Ratio:** Mostly flat but dipping slightly at the right tail indicating a negligible negative trend.

Risk & Volatility

A graph with blue dots and orange line

AI-generated content may be incorrect.A graph with blue dots and a red line

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect.A graph with blue dots and a red line

AI-generated content may be incorrect.

A graph with blue dots and a red line

AI-generated content may be incorrect.A graph with blue and orange dots

AI-generated content may be incorrect.

A graph of blue dots and orange lines

AI-generated content may be incorrect.A graph with blue dots and orange lines

AI-generated content may be incorrect.

A graph with blue dots and orange line

AI-generated content may be incorrect.A graph of blue dots

AI-generated content may be incorrect.

A graph of blue dots

AI-generated content may be incorrect.A graph with blue dots and a red line

AI-generated content may be incorrect.

A graph of blue dots

AI-generated content may be incorrect.A graph of blue dots

AI-generated content may be incorrect.

A graph with blue dots and orange line

AI-generated content may be incorrect.

Risk & Volatility – Historical Volatility vs Other Indicators

* **Historical Volatility vs Beta:** Slight positive slope indicating more volatile assets tend to have marginally higher beta.
* **Historical Volatility vs Sharpe Ratio:** Strong negative correlation as volatility increases, risk‐adjusted returns decline.
* **Historical Volatility vs Max Drawdown:** Nearly perfect negative slope higher volatility corresponds to deeper drawdowns.
* **Historical Volatility vs Std Dev Returns:** Almost perfect positive slope showing that more volatile assets have higher return dispersion.
* **Historical Volatility vs Value at Risk:** Almost perfect positive trend suggesting higher volatility leads to higher estimated losses.

Risk & Volatility – Beta vs Other Indicators

* **Beta vs Historical Volatility:** Slight positive trend as assets with higher beta tend to be more volatile.
* **Beta vs Sharpe Ratio:** Nearly flat with a mild negative slope indicating higher beta lightly erodes risk‐adjusted returns.
* **Beta vs Max Drawdown:** Strong positive association assets more sensitive to market moves suffer larger drawdowns.
* **Beta vs Std Dev Returns:** Good positive correlation reflecting that higher market sensitivity comes with greater return variability.
* **Beta vs Value at Risk:** Mild positive slope showing a small increase in expected losses with higher beta.

Risk & Volatility – Sharpe Ratio vs Other Indicators

* **Sharpe Ratio vs Beta:** Almost flat with a slight negative tilt indicating higher market sensitivity modestly lowers risk‐adjusted returns.
* **Sharpe Ratio vs Historical Volatility:** Strong negative correlation more volatility leads to lower risk‐adjusted performance.
* **Sharpe Ratio vs Max Drawdown:** Positive correlation suggesting assets with better risk‐adjusted returns experience smaller drawdowns.
* **Sharpe Ratio vs Std Dev Returns:** Negative correlation as higher dispersion of returns reduces risk‐adjusted performance.
* **Sharpe Ratio vs Value at Risk:** Very negative relationship higher expected losses correspond to poorer risk‐adjusted returns.

Risk & Volatility – Max Drawdown vs Other Indicators

* **Max Drawdown vs Beta:** Strong positive relationship assets with higher market sensitivity endure deeper drawdowns.
* **Max Drawdown vs Historical Volatility:** Nearly perfect negative slope (drawdown is negative), indicating more volatile assets suffer larger losses.
* **Max Drawdown vs Sharpe Ratio:** Positive correlation, since assets with better risk‐adjusted returns have smaller drawdowns.
* **Max Drawdown vs Std Dev Returns:** Strong negative correlation showing that larger return variability is linked to deeper drawdowns.
* **Max Drawdown vs Value at Risk:** Strong negative association as assets with deeper drawdowns also exhibit higher estimated losses.

Risk & Volatility – Std Dev Returns vs Other Indicators

* **Std Dev Returns vs Beta:** Good positive correlation, indicating market‐sensitive assets show greater return dispersion.
* **Std Dev Returns vs Historical Volatility:** Nearly perfect positive slope reflecting that standard deviation of returns and volatility move together.
* **Std Dev Returns vs Sharpe Ratio:** Negative correlation as higher dispersion lowers risk‐adjusted returns.
* **Std Dev Returns vs Max Drawdown:** Strong negative relationship linking greater return variability to deeper drawdowns.
* **Std Dev Returns vs Value at Risk:** Very strong positive correlation showing that assets with larger return swings face higher VaR.

Risk & Volatility – Value at Risk vs Other Indicators

* **Value at Risk vs Beta:** Mild positive slope, indicating higher market sensitivity slightly increases estimated losses.
* **Value at Risk vs Historical Volatility:** Near‐perfect positive trend, showing VaR rises in tandem with volatility.
* **Value at Risk vs Sharpe Ratio:** Very negative correlation, as assets with higher expected losses deliver poorer risk‐adjusted returns.
* **Value at Risk vs Max Drawdown:** Strong negative association, linking larger drawdowns to higher VaR.
* **Value at Risk vs Std Dev Returns:** Very strong positive correlation, reflecting that greater return variability escalates estimated losses.

Liquidity & Trading

A graph with blue and orange dots

AI-generated content may be incorrect.A graph with blue and orange dots

AI-generated content may be incorrect.

A graph with numbers and a line

AI-generated content may be incorrect.A graph with blue and orange dots

AI-generated content may be incorrect.

A graph with blue and orange lines

AI-generated content may be incorrect.A graph with blue and orange dots

AI-generated content may be incorrect.

A graph with blue and orange lines

AI-generated content may be incorrect.A graph of a trading graph

AI-generated content may be incorrect.

A graph of a trading graph

AI-generated content may be incorrect.A graph with numbers and a line

AI-generated content may be incorrect.

A graph with numbers and a line

AI-generated content may be incorrect.A graph with numbers and a line

AI-generated content may be incorrect.

A graph with a line and a point

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.

Liquidity & Trading – 30-Day Average Volume vs Other Indicators

* **30-Day Avg Volume vs Bid-Ask Spread:** Nearly flat overall with a slight downward bend at the high volume end.
* **30-Day Avg Volume vs Volume Growth:** Mostly flat relationship with a tiny downward bend in the highest volume stocks.
* **30-Day Avg Volume vs Shares Outstanding:** Clear positive correlation larger share counts tend to have higher trading volumes.
* **30-Day Avg Volume vs Float Shares:** Strong positive association stocks with more freely traded shares see higher volumes.
* **30-Day Avg Volume vs Turnover Ratio:** Good positive correlation higher volume generally means higher turnover relative to float.

Liquidity & Trading – Bid-Ask Spread vs Other Indicators

* **Bid-Ask Spread vs 30-Day Avg Volume:** Almost flat but narrows slightly in very high-volume stocks.
* **Bid-Ask Spread vs Volume Growth:** Mildly positive slope stocks whose volumes are rising tend to have slightly wider spreads.
* **Bid-Ask Spread vs Shares Outstanding:** Mild negative trend stocks with more shares outstanding generally have tighter spreads.
* **Bid-Ask Spread vs Float Shares:** Mild negative slope higher free float is associated with narrower spreads.
* **Bid-Ask Spread vs Turnover Ratio:** Mild positive correlation higher turnover stocks can exhibit marginally wider spreads.

Liquidity & Trading – Volume Growth vs Other Indicators

* **Volume Growth vs 30-Day Avg Volume:** Largely flat, with a very slight downward bend at the highest volumes.
* **Volume Growth vs Bid-Ask Spread:** Mild positive slope stocks with growing volumes have slightly wider spreads.
* **Volume Growth vs Shares Outstanding:** Mild negative relationship larger share counts often see lower volume growth rates.
* **Volume Growth vs Float Shares:** Mild negative slope stocks with larger float tend to have more stable (less growing) volumes.
* **Volume Growth vs Turnover Ratio:** Almost flat relationship turnover changes little with shifts in volume growth.

Liquidity & Trading – Shares Outstanding vs Other Indicators

* **Shares Outstanding vs 30-Day Avg Volume:** Strong positive correlation companies with more shares outstanding trade more volume.
* **Shares Outstanding vs Bid-Ask Spread:** Mild negative slope greater share counts generally coincide with tighter spreads.
* **Shares Outstanding vs Volume Growth:** Mild negative relationship larger share bases see slightly less volume growth.
* **Shares Outstanding vs Float Shares:** Very strong positive correlation the public float is closely tied to total shares outstanding.
* **Shares Outstanding vs Turnover Ratio:** Clear negative correlation firms with more shares outstanding have lower turnover percentages.

Liquidity & Trading – Float Shares vs Other Indicators

* **Float Shares vs 30-Day Avg Volume:** Strong positive association more freely tradable shares lead to higher volumes.
* **Float Shares vs Bid-Ask Spread:** Mild negative slope larger floats have tighter spreads.
* **Float Shares vs Volume Growth:** Mild negative relationship stocks with bigger floats have slightly lower volume growth.
* **Float Shares vs Shares Outstanding:** Very strong positive correlation float closely tracks total shares outstanding.
* **Float Shares vs Turnover Ratio:** Noticeable negative correlation higher float stocks see lower turnover percentages.

Liquidity & Trading – Turnover Ratio vs Other Indicators

* **Turnover Ratio vs 30-Day Avg Volume:** Good positive correlation higher turnover is linked with greater trading volumes.
* **Turnover Ratio vs Bid-Ask Spread:** Mild positive slope stocks that trade more relative to float have slightly wider spreads.
* **Turnover Ratio vs Volume Growth:** Nearly flat turnover ratio remains stable regardless of volume growth.
* **Turnover Ratio vs Shares Outstanding:** Clear negative relationship more shares outstanding leads to lower turnover ratios.
* **Turnover Ratio vs Float Shares:** Negative correlation larger floats correspond to lower turnover as a percentage of float.

Pearson Correlation

Pearson correlation matrices were also created for each sub-index. These matrices quantified indicator associations numerically and helped validate visual insights from scatterplots.

Financial Strength

A red and white squares with white text

AI-generated content may be incorrect.

Financial Strength – Key correlations

* **Current Ratio and Quick Ratio (r = 0.95):** Near‐perfect collinearity both measure short-term liquidity.
* **Operating Cash Flow and Total Revenue (r = 0.71):** Strong positive link larger revenues tend to generate more cash.
* **EBITDA Margin and ROE (r = 0.24):** Moderate positive relation higher margins often translate into better returns on equity.
* All other pairwise correlations are weak (|r| < 0.15), suggesting these indicators largely capture distinct aspects of financial strength.

Growth Potential

A graph with red and blue squares

AI-generated content may be incorrect.

Growth Potential – Key correlations

* **Operating Margin and R&D-to-Revenue (r = –0.98):** Almost perfect inverse relationship firms investing heavily in R&D tend to report lower operating margins in the same period.
* All other pairs in this group exhibit negligible correlations (|r| < 0.10), indicating broad independence among growth‐potential metrics.

Market Performance

A screenshot of a graph

AI-generated content may be incorrect.

Market Performance – Key correlations

* **EPS and Market Cap (r = 0.11):** Small positive link larger companies tend to report slightly higher earnings per share.
* All remaining correlations fall below |r| = 0.10, implying valuation ratios and payout metrics offer largely orthogonal information.

Risk & Volatility

A red and blue squares with white text

AI-generated content may be incorrect.

Risk & Volatility – Key correlations

* **Value at Risk and Historical Volatility (r = 0.83):** Very strong positive association more volatile stocks face higher downside risk.
* **Value at Risk and Std Dev Returns (r = 0.66):** Strong positive link stocks with dispersed returns also have larger tail losses.
* **Std Dev Returns and Historical Volatility (r = 0.76):** High concordance both capture spread of returns.
* **Max Drawdown and Std Dev Returns (r = –0.77):** Strong inverse relation greater variability correlates with deeper declines.
* **Max Drawdown and Historical Volatility (r = –0.65):** Substantial negative slope high volatility stocks suffer larger drawdowns.
* **Sharpe Ratio correlates weakly with all others (|r| < 0.25), reflecting its distinct risk-adjusted viewpoint.**

**Liquidity & Trading**

A screenshot of a graph

AI-generated content may be incorrect.

Liquidity & Trading – Key correlations

* **Float Shares and Shares Outstanding (r = 0.97):** Almost perfect collinearity float is just a subset of total shares.
* **Turnover Ratio and Avg 30-day Volume (r = 0.48):** Moderate positive relation stocks traded in high volume also have higher turnover relative to float.
* All other pairwise links are weak (|r| < 0.40), indicating liquidity depth, trading activity, and spreads each add unique information.

## 5.2 PCA

PCA was applied separately to each sub-index group to examine underlying dimensions and reduce complexity. Each PCA run produced scree plots, factor loadings, and explained variance ratios.

**Scree plots** visually represented the cumulative variance explained by each principal component. This guided the decision on how many components to retain:

A graph of a financial strength

AI-generated content may be incorrect.

A graph of growth potential scree plot

AI-generated content may be incorrect.

A graph of blue rectangular bars

AI-generated content may be incorrect.

A graph of blue rectangular objects

AI-generated content may be incorrect.

A graph of a graph with blue squares

AI-generated content may be incorrect.

A graph of blue bars

AI-generated content may be incorrect.

**Factor Loadings** provided insights into indicator importance within each sub-index. Squared loadings identified significant contributions clearly, aiding interpretability and weighting in subsequent steps.

**Explained Variance:**

* Financial Strength: The first three principal components captured approximately 66% of the variance, highlighting profitability, leverage, and cash flow dimensions.
* Growth Potential: Two components explained around 69% variance, distinguishing maturity and operational efficiency.
* Market Performance: Three components covered about 78% variance, capturing size, profitability, valuation, and dividend policy.
* Risk & Volatility: The first two components accounted for approximately 83%, identifying general volatility and market risk exposure.
* Liquidity & Trading: Two principal components explained roughly 62% variance, mainly reflecting liquidity and trading activity.

## 5.3 KMO and Bartlett Tests

Factorability was assessed through Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity, applied globally and within each sub-index.

* **Global KMO** was 0.688, indicating good suitability for factor analysis.
* **Bartlett’s test** was highly significant (p < 0.0001), confirming that indicators correlated enough for PCA.

Sub-index KMO results indicated:

* Financial Strength: KMO was low (0.513), necessitating removal of "operating cash flow" and "current ratio" to enhance interpretability.
* Growth Potential: Overall KMO was inadequate (0.500). Removal of "analyst rating" significantly improved group coherence.
* Market Performance: KMO was marginal (0.496). Excluding "EV to EBITDA" improved factorability.
* Risk & Volatility: KMO was strong (0.710); all indicators retained.
* Liquidity & Trading: Marginal KMO (0.544) suggested removal of "turnover ratio" to improve suitability.

| **Group** | **KMO overall** | **Bartlett’s χ²** | **p-value** | **Adequacy** |
| --- | --- | --- | --- | --- |
| **Global** | 0.688 | 40 366.65 | <.001 | Good (≥0.60) |
| **Financial Strength** | 0.513 | 7 074.78 | <.001 | Marginal (~0.50) |
| **Growth Potential** | 0.500 | 7 205.07 | <.001 | Borderline (≈0.50) |
| **Market Performance** | 0.496 | 43.19 | <.001 | Poor (<0.50) |
| **Risk & Volatility** | 0.710 | 7 533.68 | <.001 | Good (>0.70) |
| **Liquidity & Trading** | 0.544 | 7 262.17 | <.001 | Marginal (~0.54) |

## 5.4 Indicator Clustering

Clustering complemented PCA by identifying redundant indicators through hierarchical and KMeans methods. Ward’s method and KMeans with silhouette analysis provided robust cluster assignments within each sub-index group.

* Silhouette scores guided optimal cluster count selection.
* Cluster outcomes generally aligned with PCA loadings and correlation insights.

A graph with a line

AI-generated content may be incorrect.

Sub-index: Financial Strength

• Cluster 1: ['roe', 'current\_ratio', 'oper\_cash\_flow', 'ebitda\_margin']

• Cluster 2: ['debt\_to\_equity']

A graph with numbers and a dot

AI-generated content may be incorrect.

Sub-index: Growth Potential

• Cluster 1: ['revenue\_growth', 'operating\_margin', 'gross\_margin']

A graph with a line

AI-generated content may be incorrect.

Sub-index: Market Performance

• Cluster 1: ['eps', 'market\_cap']

• Cluster 2: ['price\_to\_sales', 'payout\_ratio']

A graph with a line

AI-generated content may be incorrect.

Sub-index: Risk Volatility

• Cluster 1: ['hist\_volatility', 'max\_drawdown', 'stddev\_returns', 'value\_at\_risk']

• Cluster 2: ['beta']

A line graph with numbers and symbols

AI-generated content may be incorrect.

Sub-index: Liquidity Trading

• Cluster 1: ['avg\_volume\_30d', 'float\_shares']

• Cluster 2: ['bid\_ask\_spread', 'volume\_growth']

## 5.5 Final Indicator Set

Multivariate results guided removal of redundant or poorly factorable indicators. Indicators removed based on these combined diagnostics were:

* Financial Strength: "quick\_ratio," "total\_revenue"
* Growth Potential: "analyst\_rating", “rnd\_to\_revenue”
* Market Performance: "ev\_to\_ebitda", “pb\_ratio”
* Risk & Volatility: “sharpe\_ratio”
* Liquidity & Trading: "turnover ratio", “shares\_outstanding”

## Summary and Conclusion

The multivariate analysis effectively streamlined the original set of indicators, removing redundancy and improving overall factorability. Clear dimensional structures emerged within each sub-index, ensuring interpretability and validity. PCA loadings and clustering decisions provided data-driven justification for indicator weighting and aggregation.

# 6. Normalisation

The selected indicators vary significantly in scale, unit, and distribution. Without normalisation, variables with large magnitudes (such as Market Capitalisation) could dominate the composite score, regardless of their theoretical importance. This step ensures comparability across all 21 retained indicators by rescaling them to a common range, while preserving their underlying distribution characteristics as much as possible.

## 6.1 Min-Max Scaling

Min-Max scaling was applied as the final normalisation step. This method transforms each value onto a 0-1 scale by using the minimum and maximum observed values of that indicator across all stocks. The transformation is defined as:

A mathematical equation with black text

AI-generated content may be incorrect.

This approach preserves the relative distances between values, ensuring that higher values of an indicator consistently reflect stronger performance. It is commonly used in composite indicator construction and supports interpretability by bounding the data.

To ensure numerical stability, the transformation was only applied after addressing skewness, as extremely large outliers can distort the Min-Max range. The final result ensures that all variables contribute proportionally within their sub-indices.

## 6.2 Power Transformations (Yeo - Johnson)

Several indicators exhibited extreme positive skewness, which could distort the contribution of outliers in a linear aggregation scheme. To address this, a **Yeo - Johnson power transformation** was applied prior to Min-Max normalisation. This method reduces skewness by applying a non-linear transformation while preserving the order of values and allowing for zero and negative inputs.

A graph of growth and growth

AI-generated content may be incorrect.

# 7. Weighting and Aggregation

The construction of a composite index requires explicit choices about how indicators contribute to the final score. This section defines the weighting schemes and aggregation functions applied in the CSIAI pipeline. It also includes robustness checks and outlines the treatment of each sub-index.

## 7.1 Weighting Models

Two alternative schemes were applied to assign importance to the 21 indicators retained after redundancy checks:

Equal-by-Group Weighting

Each of the five indicator groups was assigned a fixed share of the total composite weight, reflecting their perceived importance in assessing stock attractiveness. The weights were informed by investment relevance and domain logic:

| **Group** | **Share** |
| --- | --- |
| Financial Strength | 20% |
| Growth Potential | 15% |
| Market Performance | 20% |
| **Risk & Volatility** | **30%** |
| Liquidity & Trading | 15% |

PCA Variance-Based Weighting

To create a data-driven alternative, the second scheme uses **Principal Component Analysis (PCA)** within each group:

* Squared factor loadings were multiplied by the variance explained (λ₁, λ₂, ...) of each component.
* The result was a variance-weighted contribution per indicator.
* Weights were normalized so that the total weight assigned to each group matched a predefined **group share**.

| **Group** | **Share** |
| --- | --- |
| Financial Strength | 20% |
| Growth Potential | 15% |
| Market Performance | 20% |
| **Risk & Volatility** | **30%** |
| Liquidity & Trading | 15% |

Risk & Volatility was given higher weight due to its central role in investment decisions. As noted in financial literature and observed in practice, excessive exposure to volatility remains a key concern for equity investors. A larger share was therefore allocated to ensure this dimension influences the composite score proportionately.

For the Risk & Volatility group specifically, only the **first principal component** was used to compute weights. This decision was based on its high explained variance (62.9%) and one-dimensionality, reducing the need for more components.

## 7.2 Aggregation Methods

Two aggregation approaches were used to combine indicator-level scores into the final CSIAI composite:

Linear Aggregation

The primary method used was **linear summation**.

A black text on a white background

AI-generated content may be incorrect.

This method is fully compensatory strong performance in one indicator can offset weaknesses in another.

Geometric Aggregation

To test for robustness under **partial compensability**, geometric aggregation was also computed.



Geometric aggregation penalizes low-scoring dimensions more heavily, especially if an indicator is close to zero.

Both aggregation methods were applied to the equal and PCA-based weights, producing four variants.

## 7.3 Robustness Checks

To assess the stability of rankings across weighting choices, two diagnostic checks were performed:

* **Spearman correlation** between the four CI variants
* **Median absolute Δ-rank** between equal vs. PCA weighting (linear only)

The CSIAI variant using PCA-linear aggregation (CI\_pca\_lin) was selected as the final score due to its data-informed weighting and intuitive interpretation.

|  | **CI\_equal\_lin** | **CI\_pca\_lin** | **CI\_equal\_geo** | **CI\_pca\_geo** |
| --- | --- | --- | --- | --- |
| **CI\_equal\_lin** | 1.000 | 0.952 | 0.800 | 0.776 |
| **CI\_pca\_lin** | 0.952 | 1.000 | 0.760 | 0.783 |
| **CI\_equal\_geo** | 0.800 | 0.760 | 1.000 | 0.982 |
| **CI\_pca\_geo** | 0.776 | 0.783 | 0.982 | 1.000 |

Final Composite Index table

|  | **Rank** | **financial\_strength** | **growth\_potential** | **market\_performance** | **risk\_volatility** | **liquidity\_trading** | **CSIAI** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ticker** |  |  |  |  |  |  |  |
| **BRSP** | 1 | 0.106293 | 0.079301 | 0.149700 | 0.208302 | 0.093764 | 0.637360 |
| **AFCG** | 2 | 0.092488 | 0.086156 | 0.141473 | 0.218720 | 0.095227 | 0.634064 |
| **PGRE** | 3 | 0.100152 | 0.076071 | 0.150910 | 0.208269 | 0.094976 | 0.630377 |
| **PAHC** | 4 | 0.099874 | 0.071507 | 0.137018 | 0.209351 | 0.103274 | 0.621024 |
| **DEA** | 5 | 0.087929 | 0.078304 | 0.151787 | 0.208250 | 0.090815 | 0.617085 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **NRIX** | 2237 | 0.081019 | 0.066690 | 0.091531 | 0.081759 | 0.038926 | 0.359925 |
| **AMN** | 2238 | 0.084606 | 0.071254 | 0.084871 | 0.081613 | 0.037562 | 0.359906 |
| **FLYW** | 2239 | 0.068326 | 0.077637 | 0.090156 | 0.092584 | 0.028953 | 0.357655 |
| **DFIN** | 2240 | 0.079780 | 0.077363 | 0.090617 | 0.082427 | 0.023685 | 0.353873 |
| **SPHR** | 2241 | 0.072382 | 0.073573 | 0.086188 | 0.081709 | 0.036186 | 0.350038 |

# 8. Link to Other Indices

The validity and interpretability of a composite indicator improve when its outputs align with relevant external benchmarks. This section compares the **Composite Stock Investment Attractiveness Index (CSIAI)** with established market indices to assess how well the index captures broader performance signals.

## 8.1 Benchmark Selection

Three representative equity indices were selected as comparators based on their structural relevance and popularity in financial analysis:

* **SPY** - S&P 500 ETF, representing broad U.S. market exposure. Acts as a market baseline.
* **QUAL** - iShares MSCI USA Quality Factor ETF. Focuses on profitable firms with stable earnings and strong balance sheets. Reflects a quality-focused investing style.
* **MTUM** - iShares MSCI USA Momentum Factor ETF. Targets stocks with high recent performance. Serves as a contrasting high-volatility benchmark.

All three benchmarks were retrieved using the yfinance API. Daily adjusted closing prices were collected from **2023-01-01 to 2025-05-06**, matching the time frame used in CSIAI construction.

A graph of different colored lines

AI-generated content may be incorrect.

## 8.2 Correlation Comparison

To measure concordance between CSIAI scores and benchmark outcomes, the **Spearman rank correlation** was computed between each stock’s CSIAI score (**CI\_pca\_lin**) and its cumulative return over the same period.

The following correlation coefficients were obtained:

| **ETF** | **Spearman Rank Correlation** |
| --- | --- |
| SPY | 0.791 |
| QUAL | 0.750 |
| MTUM | 0.666 |

These correlations suggest a strong positive monotonic relationship between CSIAI rankings and realized market performance, especially relative to broad market and quality-style benchmarks.

## 8.3 Interpretation

Several conclusions can be drawn from this result:

* **CSIAI rankings exhibit predictive coherence**: Firms ranked higher by the CSIAI also tended to deliver higher realized returns across multiple investing styles.
* The strongest alignment occurred with **SPY** and **QUAL**, indicating that the index captures traits valued by general and quality-oriented investors.
* Weaker correlation with **MTUM** is expected, as momentum strategies are more short-term and do not fully align with the multi-dimensional, fundamentals-heavy structure of the CSIAI.

# 9. Visualisation of Results

Clear, accessible visualisation is a critical part of the CSIAI framework. As noted in the Handbook on Constructing Composite Indicators (OECD, 2008), visual tools should support both expert and non-expert interpretation, aid validation, and improve transparency.

## 9.1 Top 20 CSIAI Table

A final ranking table was produced, showing the top 20 firms in the dataset based on their PCA-weighted CSIAI score using linear aggregation.

| **Rank** | **ticker** | **financial\_strength** | **growth\_potential** | **market\_performance** | **risk\_volatility** | **liquidity\_trading** | **CSIAI** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | BRSP | 0.106 | 0.079 | 0.150 | 0.208 | 0.094 | 0.637 |
| 2 | AFCG | 0.092 | 0.086 | 0.141 | 0.219 | 0.095 | 0.634 |
| 3 | PGRE | 0.100 | 0.076 | 0.151 | 0.208 | 0.095 | 0.630 |
| 4 | PAHC | 0.100 | 0.072 | 0.137 | 0.209 | 0.103 | 0.621 |
| 5 | DEA | 0.088 | 0.078 | 0.152 | 0.208 | 0.091 | 0.617 |
| 6 | DX | 0.077 | 0.086 | 0.152 | 0.208 | 0.093 | 0.617 |
| 7 | FTAI | 0.133 | 0.074 | 0.139 | 0.166 | 0.101 | 0.612 |
| 8 | EXR | 0.096 | 0.080 | 0.159 | 0.188 | 0.086 | 0.609 |
| 9 | BXMT | 0.115 | 0.086 | 0.160 | 0.170 | 0.076 | 0.608 |
| 10 | NLY | 0.109 | 0.085 | 0.160 | 0.155 | 0.096 | 0.605 |
| 11 | FBRT | 0.091 | 0.085 | 0.149 | 0.180 | 0.097 | 0.602 |
| 12 | TPG | 0.091 | 0.070 | 0.161 | 0.183 | 0.093 | 0.598 |
| 13 | BWA | 0.086 | 0.069 | 0.120 | 0.227 | 0.094 | 0.596 |
| 14 | VICI | 0.105 | 0.086 | 0.147 | 0.168 | 0.088 | 0.594 |
| 15 | STWD | 0.109 | 0.083 | 0.156 | 0.180 | 0.065 | 0.593 |
| 16 | CHCT | 0.099 | 0.081 | 0.148 | 0.178 | 0.086 | 0.593 |
| 17 | ITW | 0.094 | 0.074 | 0.144 | 0.181 | 0.099 | 0.591 |
| 18 | KGS | 0.089 | 0.077 | 0.153 | 0.196 | 0.076 | 0.591 |
| 19 | LTC | 0.102 | 0.085 | 0.146 | 0.169 | 0.089 | 0.590 |
| 20 | UWMC | 0.095 | 0.086 | 0.157 | 0.186 | 0.066 | 0.590 |

## 9.2 Horizontal Bar Chart of Top 10 CSIAI Scores

A horizontal bar chart was created using seaborn to visualise the top 10 CSIAI scores. Each bar represents one ticker, ranked from highest to lowest, based on PCA-weighted linear aggregation.

A chart of a number of colored bars

AI-generated content may be incorrect.

## 9.3 Radar Charts of Sub-Index Profiles

Radar charts were generated for selected top and bottom performers to visualise performance across the five sub-indices. Each axis represents one dimension:

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

A diagram of a graph

AI-generated content may be incorrect.  
  
A diagram of a graph

AI-generated content may be incorrect.

## 9.4 Correlation Heatmap of Sub-Index Scores

A correlation heatmap was computed based on the PCA-weighted sub-index scores. The heatmap uses Pearson correlation coefficients to assess relationships between dimensions.

This reveals whether certain dimensions are highly aligned or provide distinct information. A low correlation between dimensions indicates that each sub-index adds unique value to the composite.

A graph with red squares

AI-generated content may be incorrect.

## 9.5 Decomposition Bar Charts for Top and Bottom Firms

To further illustrate sub-index contributions, bar charts were produced for the top 5 and bottom 5 tickers. Each chart displays the five sub-index scores that combine to form the CSIAI.

These decomposition plots allow for inspection of which dimensions contribute most (or least) to the final attractiveness score.

A graph of a bar chart

AI-generated content may be incorrect.

A graph of different sizes and colors

AI-generated content may be incorrect.

# 10. Conclusion

## 10.1 Key Takeaways

This documentation has shown the construction of the Composite Stock Investment Attractiveness Index (CSIAI). Key steps and outcomes include:

* **21 selected indicators** categorized into five dimensions:
  + Financial Strength
  + Growth Potential
  + Market Performance
  + Risk & Volatility
  + Liquidity & Trading Activity
* **Two aggregation methods**: linear aggregation as the primary method, supported by geometric aggregation for robustness checks.
* **Weighting schemes**: An equal-by-group weighting approach, and a PCA-based method using factor loadings for data-driven weight assignment.
* **Cross-index correlation**: CSIAI showed strong correlations (Spearman >0.75) with established indices (SPY, QUAL), demonstrating validity in capturing meaningful stock market attributes.

## 10.2 Strengths

The CSIAI methodology offers several strengths:

* **Transparent**: Each step from data selection, multivariate analysis, normalisation, and aggregation has explicit justification, increasing stakeholder confidence.
* **Explainable**: Dimension scores and rankings are intuitive and supported by clear visualisations, aiding practical interpretation for both financial analysts and non-specialists.

## 10.3 Limitations

Certain constraints and limitations should be considered when using CSIAI:

* **Static data**: Currently, the CSIAI is generated from data spanning January 2023 to May 2025, offering no dynamic insights into evolving market trends.
* **Absence of Sensitivity Analysis**: Sensitivity analysis regarding weighting schemes, indicator selection, and imputation methods was not performed, leaving robustness under different assumptions untested.

## 10.4 Future Work

Future extensions to enhance the CSIAI include:

* **Time-series Trends**: Implementing a historical analysis to track changes in stock attractiveness over multiple periods.
* **Sector-specific Indexes**: Developing industry-focused versions of the CSIAI to provide targeted insights for particular investment sectors, such as technology, healthcare, or finance.

# References

* Bland, G. (n.d.). yfinance library – Full guide. AlgoTrading101. <https://algotrading101.com/learn/yfinance-guide/>
* ChatGPT. (2024). Prompt-driven code generation and clarification. <https://chatgpt.com/share/681d4dc8-36f0-800b-8939-d08f76061782>
* DataCamp. (n.d.). Analyzing Financial Statements in Python. <https://campus.datacamp.com/courses/analyzing-financial-statements-in-python/the-balance-sheet?ex=9>
* DataCamp. (n.d.). Introduction to Factor Analysis. <https://www.datacamp.com/tutorial/introduction-factor-analysis>
* GeeksforGeeks. (n.d.). How to use yfinance API with Python. <https://www.geeksforgeeks.org/how-to-use-yfinance-api-with-python/>
* GeeksforGeeks. (n.d.). PowerTransformer in scikit-learn. <https://www.geeksforgeeks.org/powertransformer-in-scikit-learn/>
* Hilpisch, Y. (2020). Python for Algorithmic Trading. O'Reilly Media.
* Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
* Kumanireng, A. (n.d.). Application of Fundamental Analysis in Stock Valuation in the Capital Market and Investment Decisions by Price Methods Earnings Ratio (PER).
* Murphy, J. J. (1999). Technical Analysis of the Financial Markets. New York Institute of Finance.
* Organisation for Economic Co-operation and Development (OECD). (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing.
* OECD Knowledge Policy Platform. (n.d.). Toolkit for Composite Indicators: 10-Step Guide. <https://knowledge4policy.ec.europa.eu/composite-indicators/toolkit_en/navigation-page/10-step-guide_en>
* Towards Data Science. (n.d.). Top 3 methods for handling skewed data. <https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45/>
* Microsoft Copilot. (2024). Code suggestion and auto-completion tool (activated during notebooks).