High-Dimensional Analysis of Stock Returns

Group A: Christy Lai, Sun Ma, Jiwen Zhou, Rong Xu, Sarah Liu

Introduction

This report examines monthly returns for up to 100 NYSE stocks (2005–2019) alongside the S&P 500 to distinguish common (systematic) movement from stock-specific variation. We map each ticker to its industry via the leading letter, then use Principal Component Analysis (PCA) and Factor Analysis (FA) to study co-movement and industry heterogeneity, and to identify stocks that most closely track the market. Before those models, we run an initial data audit and exploratory analysis to confirm the dataset is tidy, complete, and suitable for high-dimensional methods.

IDA and EDA

Data Quality and Formats

Using `visdat`, we verified data types and visualised missingness. The data comprise monthly returns for numerous stocks across several industries. Missing observations are confined to Industry C; the remaining industries have no missing values. We further confirmed the absence of duplicate entries and of gaps in the monthly timeline. All variables are numeric apart from the date column. Stock codes encode industry with the first letter and the specific stock with the following numbers. The industry mapping is as follows:

Table 1: Industry summary statistics

Code	Industry	n_stocks	prop
B	Mining	3	0.0365854
C	Construction	NA	NA
D	Manufacturing	13	0.1585366
E	Transportation and Public Utilities	16	0.1951220
F	Wholesale Trade	5	0.0609756
G	Retail Trade	3	0.0365854

Table 1: Industry summary statistics

Code	Industry	n_stocks	prop
H	Finance, Insurance and Real Estate	28	0.3414634
I	Services	14	0.1707317

Check for Outliers Using Z-Scores

We used the 3-sigma rule to flag outliers: monthly returns with z-scores beyond ± 3 relative to each stock's history. After computing per-stock z-scores and merging them back to the dataset, we examined extreme return events. The five largest (by $|z|$) are shown in Table 2.

For instance, D79122 posted a 4.15 return in September 2018 ($z = 10.13$). Several other extremes—particularly in Finance (H89011, H89548, H89050) and Services (I90394)—cluster around 2008–2009, aligning with the Global Financial Crisis.

Table 2: Top 6 outliers detected using Z-scores

Date	Stock	Z_Score	Return
2018-09-01	D79122	10.128081	4.148734
2009-04-01	I90394	7.710384	1.011407
2005-08-01	E88332	7.124317	1.653846
2009-08-01	H89548	7.086195	1.569231
2009-01-01	H89011	6.886952	0.403834
2009-01-01	H89050	6.634758	0.314220

Table 3: Summary statistics for stock returns

stock	mean	sd	min	p25	median	p75	max
D79122	0.0419732	0.4054826	-0.738924	-0.1057560	-0.0213945	0.0828745	4.148734
E86444	0.0186337	0.2732906	-0.500000	-0.1212192	-0.0185450	0.0974352	1.810811
D88351	0.0280401	0.2566920	-0.586597	-0.1136885	-0.0031420	0.1128890	1.512882
D82824	0.0117309	0.2444019	-0.596798	-0.1160828	-0.0239685	0.1095695	1.057778
E88332	-0.0019823	0.2324192	-0.614634	-0.1184658	-0.0321800	0.1036540	1.653846
H89548	0.0180290	0.2189048	-0.558025	-0.0641698	0.0005695	0.0731190	1.569231
D45225	0.0217177	0.2142692	-0.424691	-0.1087085	0.0025760	0.1228830	0.790123
D79702	0.0014182	0.2080833	-0.634426	-0.1182875	-0.0190790	0.0881908	0.845945
I10890	0.0069911	0.2010341	-0.559211	-0.1093722	0.0011745	0.0942492	1.301887

Table 3 summarizes the top 10 most volatile stocks by standard deviation. D79122 shows the highest volatility and return, aligning with its outlier status. While most stocks have near-zero mean returns, E88332 shows high variability with a negative mean. These results reveal substantial differences in volatility, supporting the use of PCA and factor models to identify common patterns and systematic risk drivers.

Boxplot of top volatile stocks

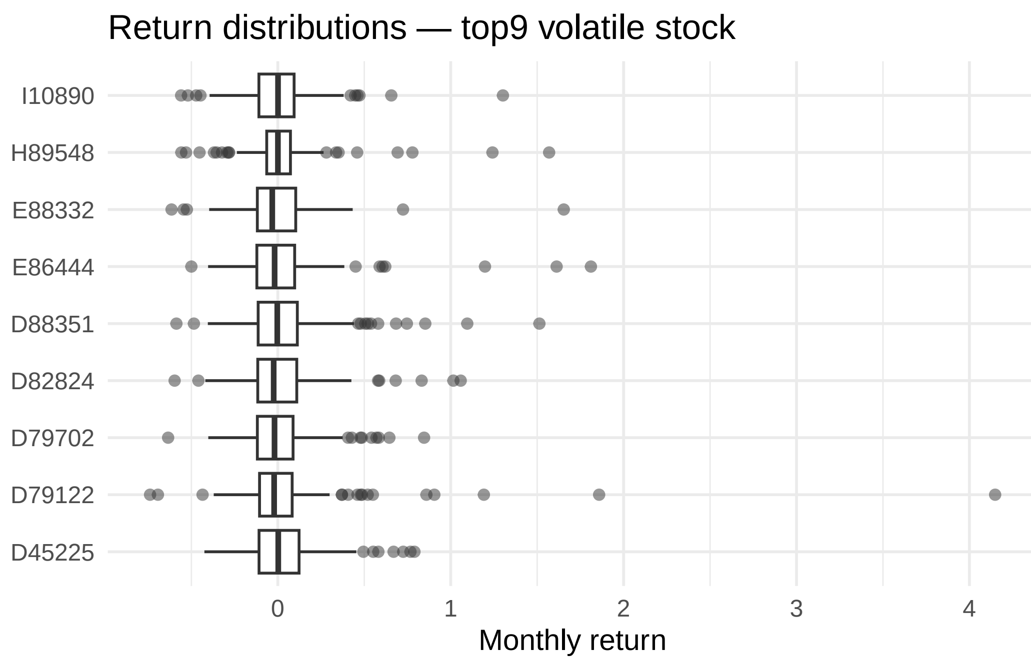

Figure 1: Return distributions of top 9 volatile stocks

Figure 1 presents return distributions for the top 9 most volatile stocks. While most returns are centered around zero, several stocks exhibit long right tails and extreme outliers—especially D79122, which exceeds a return of 4. These patterns confirm earlier findings and highlight the importance of using PCA and factor models to account for shared variation driven by high-volatility stocks.

Principal Component Analysis (PCA)

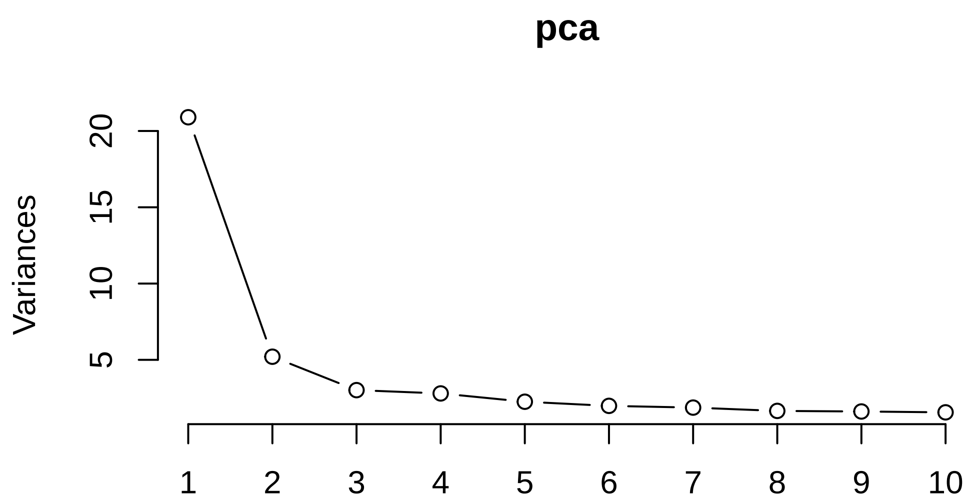

Figure 2: Scree plot of PCA

Table 4: Variance explained by PC1–PC3

PC	Variance	Proportion	Cumulative
PC1	20.9011	0.2549	0.2549
PC2	5.2057	0.0635	0.3184
PC3	3.0153	0.0368	0.3551

After standardization, the Figure 2 shows a clear elbow at three components. Table 4 indicates PC1–PC3 explain ~35–36% of total variance.

Table 5: Correlation between industry mean returns and market return

	MarketReturn	B	D	E	F	G	H	I
MarketReturn	1.000	0.619	0.684	0.819	0.603	0.578	0.880	0.888
B	0.619	1.000	0.510	0.647	0.437	0.260	0.579	0.538
D	0.684	0.510	1.000	0.645	0.344	0.473	0.652	0.674
E	0.819	0.647	0.645	1.000	0.517	0.467	0.777	0.815
F	0.603	0.437	0.344	0.517	1.000	0.341	0.603	0.499
G	0.578	0.260	0.473	0.467	0.341	1.000	0.598	0.547

Table 5: Correlation between industry mean returns and market return

	MarketReturn	B	D	E	F	G	H	I
H	0.880	0.579	0.652	0.777	0.603	0.598	1.000	0.831
I	0.888	0.538	0.674	0.815	0.499	0.547	0.831	1.000

Table 6: Industry Loadings on PC1, PC2, and PC3

industry	PC1	PC2	PC3
B	-0.1191739	0.0315199	0.2310631
D	-0.0802574	-0.0283258	0.0064412
E	-0.1091696	-0.0050055	0.0247037
F	-0.0897820	0.0825575	0.0210354
G	-0.0957973	-0.0085410	-0.1341808
H	-0.1090360	0.0459212	-0.0534513
I	-0.1108197	-0.0309260	-0.0078997

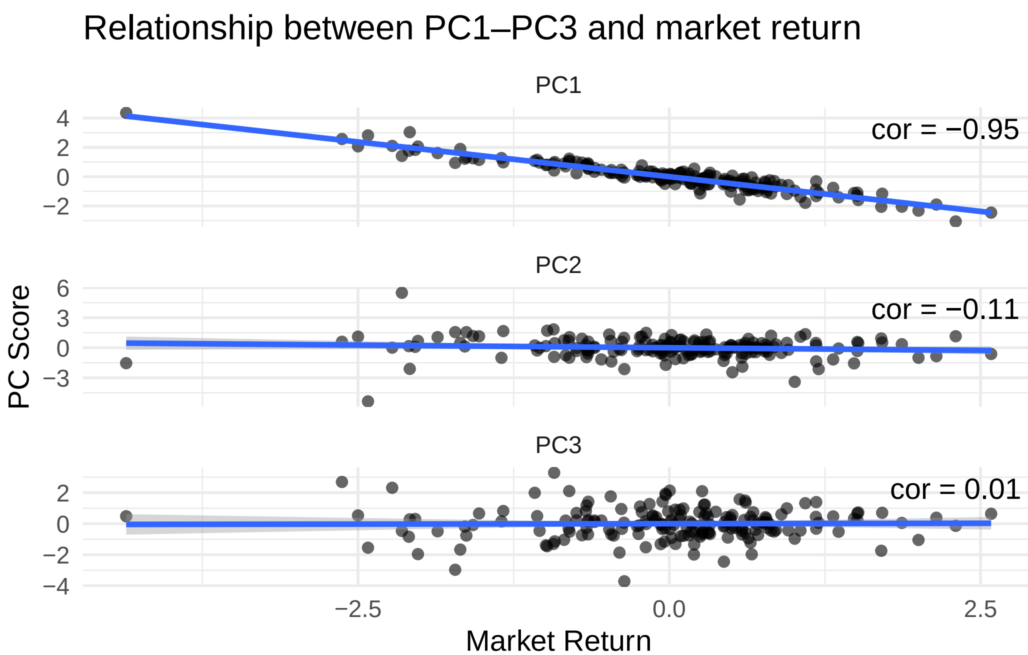

Figure 3: Relationships of PC1–PC3 with market return

Together, PC1–PC3 explain just over one-third of total variance (Table 4). Among them, PC1 alone accounts for the largest share and, as shown in the correlation analysis with the market return (Figure 3), moves almost perfectly opposite to the market index. Because the sign of principal components is arbitrary, this negative correlation implies that PC1 coincides with the market factor once its orientation is aligned. PC2 and PC3 have only weak or near-zero correlations with the market, suggesting that they represent more sector-specific or idiosyncratic patterns.

Industry-level summaries reinforce this conclusion. Table 5 shows that mean returns for industries E (Transportation/Utilities), H (Finance/Insurance/Real Estate), and I (Services) are most correlated with the market. Likewise, the industry centroids on PC1 (Table 6) reveal relatively large loadings for these same groups. By contrast, industries B (Mining) and F (Wholesale) feature more heavily on PC2 and PC3, indicating that their variation is driven by unique sectoral forces rather than broad market co-movement.

Factor Modelling

Determine number of factors

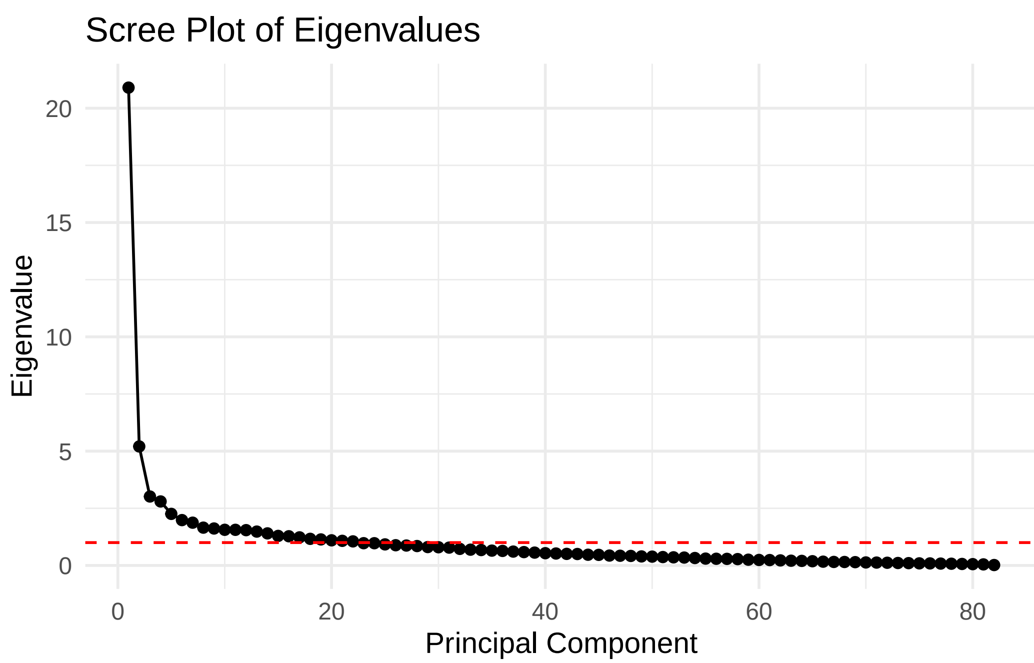

Figure 4: Scree plot of eigenvalues for factor analysis

Figure 4 showed a steep decline in eigenvalues, with the elbow around 3 factors. Therefore, we chose to fit a 3-factor model to capture the main sources of common variation while keeping the model parsimonious.

Fit the Factor Model

Table 7: Average Factor Loadings and Uniqueness by Industry

industry	mean_F1	mean_F2	mean_F3	mean_uniqueness	n_stocks
B	0.637	-0.046	-0.047	0.643	3
D	0.190	0.214	-0.019	0.826	13
E	0.297	0.229	0.019	0.723	16
F	0.400	-0.008	0.099	0.743	5
G	0.043	0.383	0.112	0.777	3
H	0.208	0.248	0.187	0.510	28
I	0.269	0.295	-0.022	0.703	14

Mining (B) has the highest loading on Factor 1 (0.637), but this result is fragile due to the small number of stocks (3). In contrast, Finance (H) stands out with the largest sector size (28 stocks) and the lowest uniqueness (0.510), indicating strong systematic exposure despite a lower Factor 1 loading. Retail (G) loads most heavily on Factor 2, while Factor 3 is more diffuse, with Finance again showing moderate positive exposure. Overall, Finance is the most systematically driven industry, aligning with PCA–market findings.

Most industries load positively on common factors, suggesting broad market co-movement. Factor 1 captures movements in Mining, Finance, Services (I), and Transport/Utilities (E), while Retail dominates Factor 2, and no industry dominates Factor 3. No industry consistently shows negative loadings, though some subgroups within Manufacturing (D) and Transport (F) exhibit divergent signs, reflecting within-industry heterogeneity. Finance, with low uniqueness and consistent factor alignment, behaves systematically, whereas industries like Manufacturing and Retail display more idiosyncratic patterns.

Table 8: Top 10 Most Systematic Stocks (Lowest Uniqueness)

industry	stock	F1	F2	F3	uniqueness
H	H90000	0.375	0.697	-0.010	0.050
H	H88215	0.569	0.489	-0.029	0.097
H	H75157	-0.102	-0.065	0.933	0.219
H	H89050	-0.076	-0.155	0.923	0.233
H	H77466	-0.019	-0.179	0.897	0.251
H	H89437	-0.083	-0.090	0.908	0.257
H	H89190	0.990	-0.225	-0.082	0.316
H	H89011	-0.080	-0.089	0.871	0.316
H	H89773	0.626	0.032	0.245	0.382
H	H83835	0.588	0.268	-0.055	0.411

Uniqueness values close to 0 indicate stocks well explained by the common factors (systematic risk). Higher uniqueness (>0.7) means the stock is more idiosyncratic.

From Table 8, stocks such as H90000, H88215, and H75157 exhibit very low uniqueness, meaning their returns are almost entirely explained by these common factors.

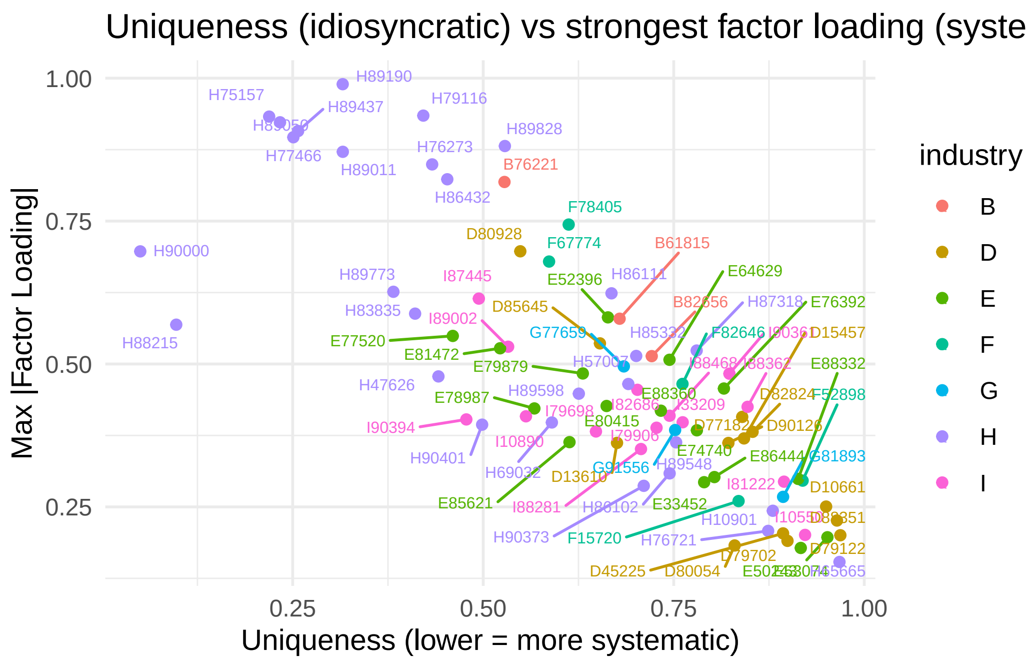

Figure 5: Uniqueness vs strongest factor loading

Figure 5 shows that stocks from the Finance industry cluster in the systematic corner—with low uniqueness (below 0.4) and high factor loadings (above 0.6).

This is reinforced Table 8, where every single entry belongs to industry H. For example, H90000 exhibits an exceptionally low uniqueness of 0.05, meaning that 95% of its variance is explained by common factors, while others like H88215, H89050, and H75157 also show strong exposure to Factors 1 or 3.

This dual evidence highlights that Finance stocks are the most strongly tied to systematic risk in the market, moving closely with common latent factors, whereas other industries (e.g.D, G) tend to scatter toward higher uniqueness, reflecting greater idiosyncratic variation and weaker alignment with overall market drivers.

Factor Loading Plot

Figure 6, together with Figure 5 and Table 8, consistently shows that Finance stocks are the most systematically driven. They load heavily on Factor1 (the main market factor), have very low uniqueness, and stand apart in both PCA and factor analysis.

Other industries (D, G) play a lesser role in explaining common market variation, making them more idiosyncratic and potentially useful for diversification.

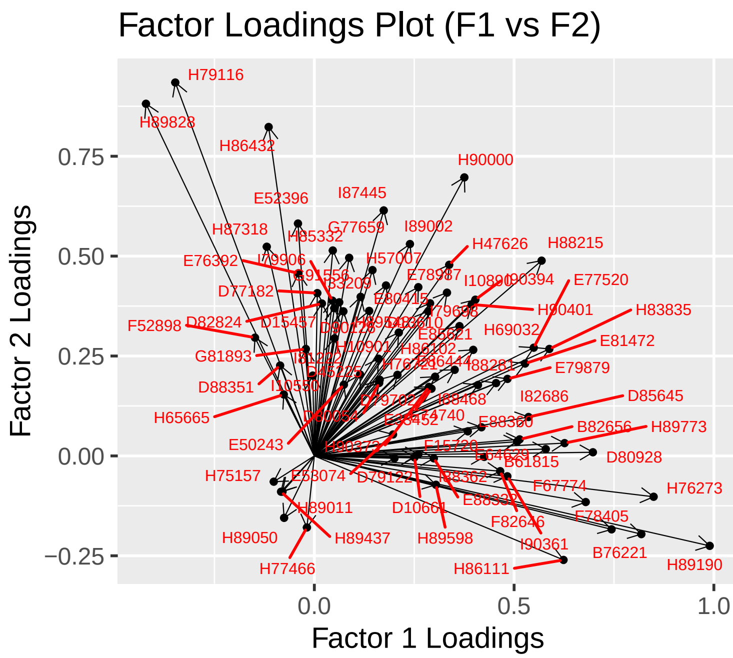

Figure 6: Factor Loadings Plot

Putting PCA and FA together

Table 9: Top 5 Most Systematic Stocks with FA and PCA Loadings

industry	stock	uniqueness	F1	F2	F3	PC1	PC2	PC3
H	H90000	0.050	0.375	0.697	-0.010	-0.208	-0.076	-0.057
H	H88215	0.097	0.569	0.489	-0.029	-0.204	-0.036	0.011
H	H75157	0.219	-0.102	-0.065	0.933	-0.058	0.336	-0.188
H	H89050	0.233	-0.076	-0.155	0.923	-0.043	0.345	-0.129
H	H77466	0.251	-0.019	-0.179	0.897	-0.048	0.353	-0.154

Combining factor loadings, uniqueness, and PCA loadings shows that the most systematic stocks in the sample are concentrated in Finance.

Stocks such as H90000 (u=0.05) and H88215 (u=0.097) have extremely low uniqueness and substantial factor exposure (e.g.F2 0.70 and F1 0.57), indicating their returns are largely explained by common factors.

Several others (H75157, H89050, H77466, H89437, H89011) load very strongly on Factor 3 (0.87–0.93), while H89190 loads almost perfectly on Factor 1 (0.99). Their PC1 loadings are also large in magnitude, confirming alignment with the market component.

Figure 7 highlights these picks in the low-uniqueness, high-loading corner of the factor space.

Overall, Finance stocks dominate systematic risk, whereas higher-uniqueness names in other industries are more idiosyncratic and offer diversification.

Ranking criteria

When selecting the best 5 stocks we can use the following criteria:

- Low uniqueness = systematic (market-driven).
- Strong factor loading (absolute value, 0.8 is very strong).
- PC1 alignment = optional filter to ensure market co-movement.

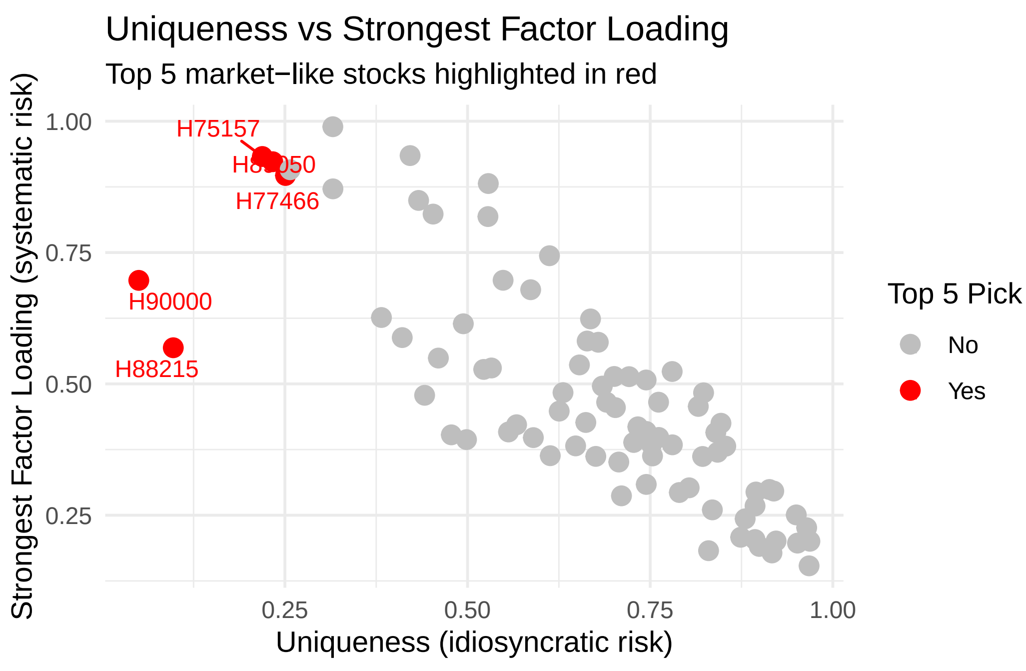

Figure 7: Uniqueness vs strongest factor loading, highlighting top 5 picks

Recommended Top 5 Stocks

From the factor analysis results, the five stocks most representative of systematic risk are H90000, H88215, H75157, H89050, and H77466, all from the Finance, Insurance & Real Estate sector.

These stocks have low uniqueness values (0.05–0.25), indicating that their returns are largely explained by common factors. They also exhibit very high factor loadings (>0.55), confirming their alignment with market-wide systematic variation.

These would be the most suitable candidates for an investor aiming to track market movements.

Assumptions and Limitations

- Sign indeterminacy. PC/factor signs are arbitrary. We align PC1 to the market before interpreting "with/against the market."
- Stationarity. Stock return distributions and their covariance structure are assumed to be approximately stationary across the 2005–2019 sample period.
- Linearity of relationships. PCA/FA assume linear correlation structures among returns. Nonlinear dependencies (volatility clustering, tail co-movements) are not captured.
- Selection practicality. The five-stock portfolio is concentrated (all H in the strict "most systematic" sense). If mandate requires sector diversification, expect a trade-off between tracking the market and spreading sector risk.
- Sampling window. 2005–2019 spans the GFC and a long expansion; factor structure can shift in other regimes.
- Industry coding granularity. One-letter codes pool heterogeneous businesses. Within-industry dispersion—especially in H—suggests sub-sectors (e.g., banks vs. REITs) that a finer classification would separate.
- Time aggregation. Returns are monthly. Higher-frequency (daily/weekly) factors might differ, and factor exposures may be time-varying.

Conclusion

This report uses PCA and factor analysis to separate systematic from idiosyncratic variation in NYSE stock returns. PC1 aligns closely with the market, confirming it captures broad co-movement. While Mining (B) loads highest on Factor 1, Finance (H) is the most systematically driven sector overall, with low uniqueness and strong exposure across all factors.

Most industries move with common factors, though Manufacturing and Retail show greater heterogeneity. Based on low uniqueness and high factor loadings, we recommend five Finance stocks—H90000, H88215, H75157, H89050, and H77466—for investors seeking market-tracking exposure. These stocks reflect systematic risk most clearly and offer the strongest alignment with overall market dynamics.