High-Dimensional Analysis of Stock Returns

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

Before proceeding to high-dimensional modelling, we conducted an initial data audit and exploratory analysis to ensure the dataset is suitable for PCA and factor modelling. This involved checking data completeness, types, summary statistics, and several visualisations.

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

Table 1: Industry summary statistics

Code	Industry	n_stocks	prop
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
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

Table 3: Summary statistics for stock returns

stock	mean	sd	min	p25	median	p75	max
D45225	0.0217177	0.2142692	-0.424691	-0.1087085	0.0025760	0.1228830	0.790123
D79702	0.0014183	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

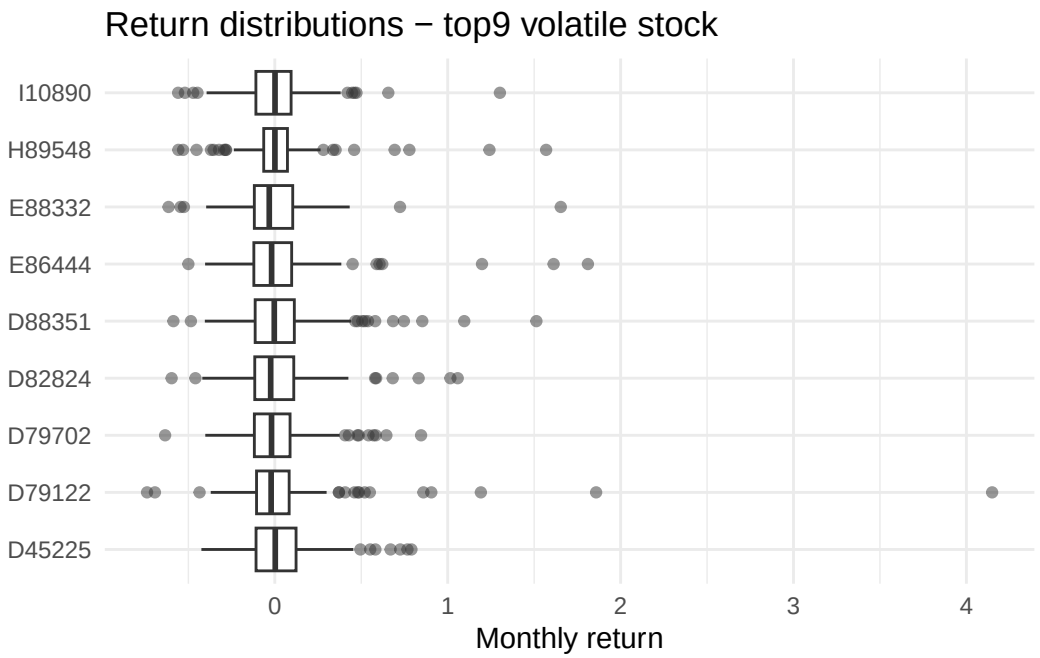

Figure 1: Return distributions of top 9 volatile stocks

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.

PCA

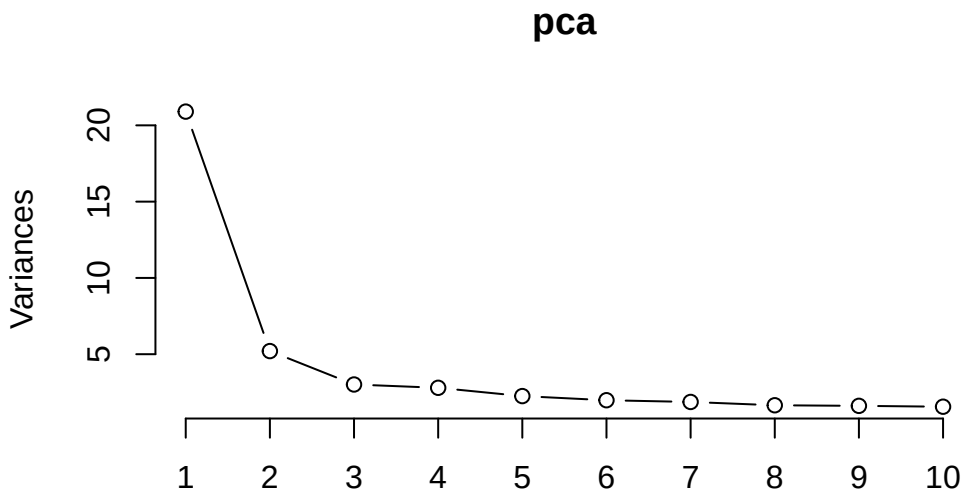

Figure 2: Scree plot of PCA

PCA is conducted after standardizing the data. 3 principle components(PCs) are selected as the elbow point is at PC3 in Figure 2, in total explain 36% of total variance.

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

As shown in Table 4, PC1 captures the largest portion of variance across all stock returns. In the context of financial data, this corresponds to the systematic risk, i.e., the component of returns that affects many stocks simultaneously and cannot be diversified away.

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
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

As mentioned in Table 5 and Table 6, industry E, H, and I are highly correlated with market return, and they all have a loading of around -0.11 for PC1, which is relatively negative compared to other industries.

From Figure 3, it is clear that market is moving negatively with PC1, PC2 and PC3 do not show clear correlation with market return. So that industries with high importance to PC2 and PC3, industry F and B, could be considered to have unique industry variations (idiosyncratic risk).

To summarize, industry E, H, and I are considered to contain more systematic risk, while industry F and B are considered to contain more idiosyncratic risk.

Table 7: Top 5 Stocks by Absolute PC1 Loading

stock	PC1	PC2	PC3	industry	abs_PC1
H90000	-0.2079832	-0.0760623	-0.0566020	H	0.2079832
H88215	-0.2044865	-0.0358788	0.0109751	H	0.2044865
H83835	-0.1662695	-0.0162305	0.0579760	H	0.1662695

Table 7: Top 5 Stocks by Absolute PC1 Loading

stock	PC1	PC2	PC3	industry	abs_PC1
H89773	-0.1607363	0.1422706	-0.0285654	H	0.1607363
E77520	-0.1600183	-0.0259024	0.0951649	E	0.1600183

PC1 usually represents the market-wide factor. Selecting stocks with the largest absolute loadings on PC1 identifies those most exposed to systematic risk. Table 7 shows the top 5 such stocks — 4 of which belong to the Finance/Real Estate sector, suggesting this industry plays a dominant role in systematic co-movements within the sample. While these stocks move strongly in relation to the market-wide component, the negative correlation between PC1 and the actual market return indicates an inverse relationship. Including E77520 from Transportation adds sectoral diversity, making this basket a reasonable candidate for investors seeking systematic exposure.

Factor Modelling

Building on these PCA results, we next turn to factor analysis to disentangle the structure of common versus idiosyncratic risk in greater detail. While PCA identifies PC1 as the dominant market-wide factor, factor models allow us to quantify how strongly each stock and industry loads on latent factors, and to evaluate the degree of uniqueness (idiosyncratic variation) that is not explained by systematic influences.

Determine number of factors

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 8 and Table 9 show Factor 1 captures a broad market-finance risk, loading heavily on large H stocks and some capital-intensive firms, while Factor 2 reflects a services/finance subsector tilt that is negatively correlated with Factor 1, consistent with sector rotation effects. Factor 3 represents another finance subgroup factor, positively correlated with Factor 1, highlighting systematic risk within different parts of the financial sector.

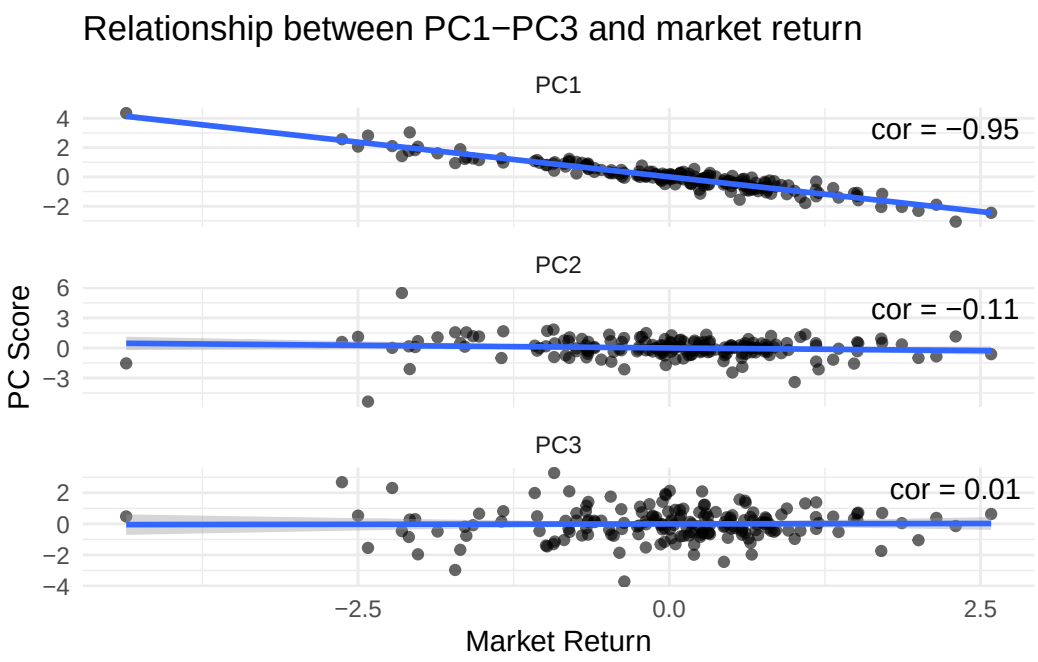

Figure 3: Relationships of PC1–PC3 with market return

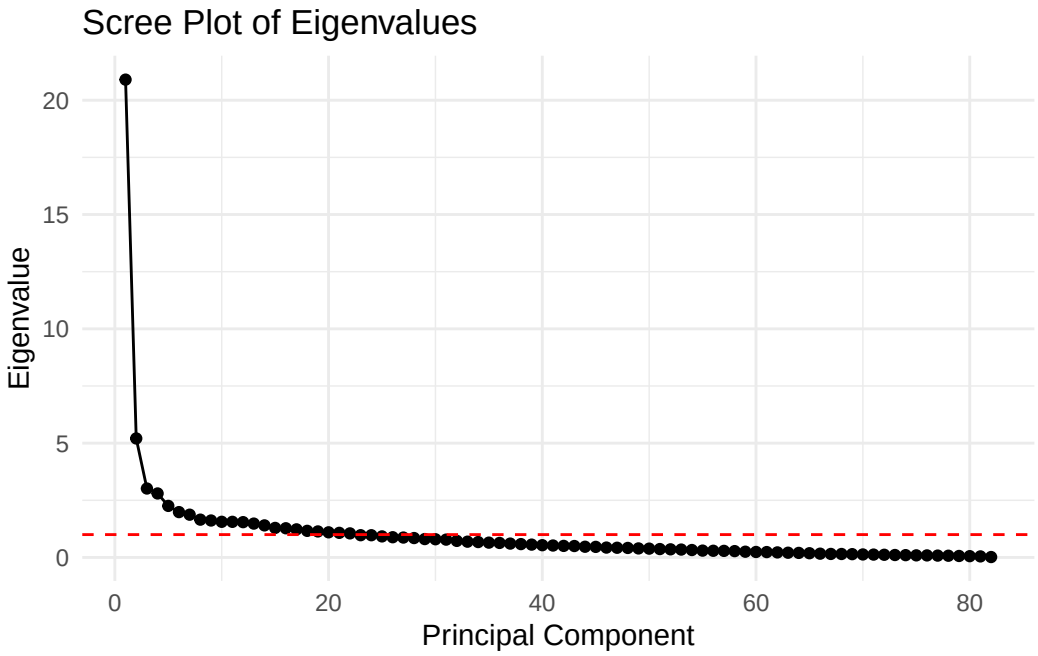

Figure 4: Scree plot of eigenvalues for factor analysis

Table 8: Numbers of Each Industry on Factors 1–3

ind	Factor1	Factor2	Factor3	n_stocks	prop
H	7	9	6	28	0.3414634
E	5	4	0	16	0.1951220
I	5	3	0	14	0.1707317
D	2	1	0	13	0.1585366
F	3	0	0	5	0.0609756
B	3	0	0	3	0.0365854
G	0	1	0	3	0.0365854

Table 9: Factor Correlation Matrix

Factor	Factor1	Factor2	Factor3
Factor1	1.00	-0.44	0.65
Factor2	-0.44	1.00	-0.24
Factor3	0.65	-0.24	1.00

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

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

industry	stock	Factor1	Factor2	Factor3	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	H89011	-0.080	-0.089	0.871	0.316
H	H89190	0.990	-0.225	-0.082	0.316
H	H89773	0.626	0.032	0.245	0.382
H	H83835	0.588	0.268	-0.055	0.411

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 10,

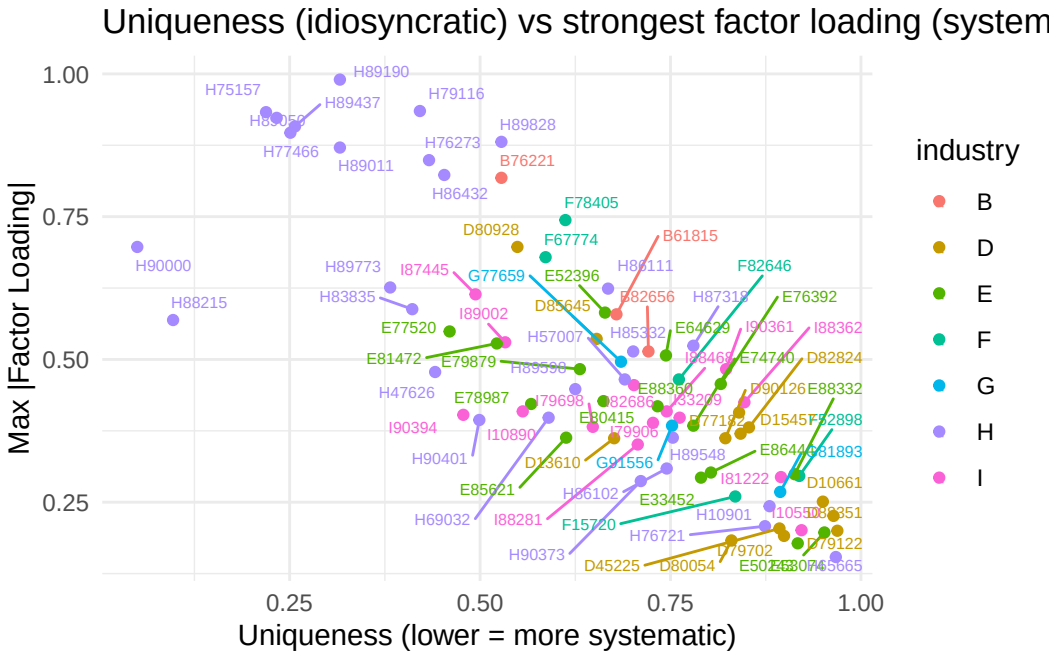

Figure 5: Uniqueness vs strongest factor loading

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 10, 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.

Table 11: 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

Table 11: Top 5 Most Systematic Stocks with FA and PCA Loadings

industry	stock	uniqueness	F1	F2	F3	PC1	PC2	PC3
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. Names 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. Overall, Finance stocks dominate systematic risk, whereas higher-uniqueness names in other industries are more idiosyncratic and offer diversification.

Ranking 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.

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.

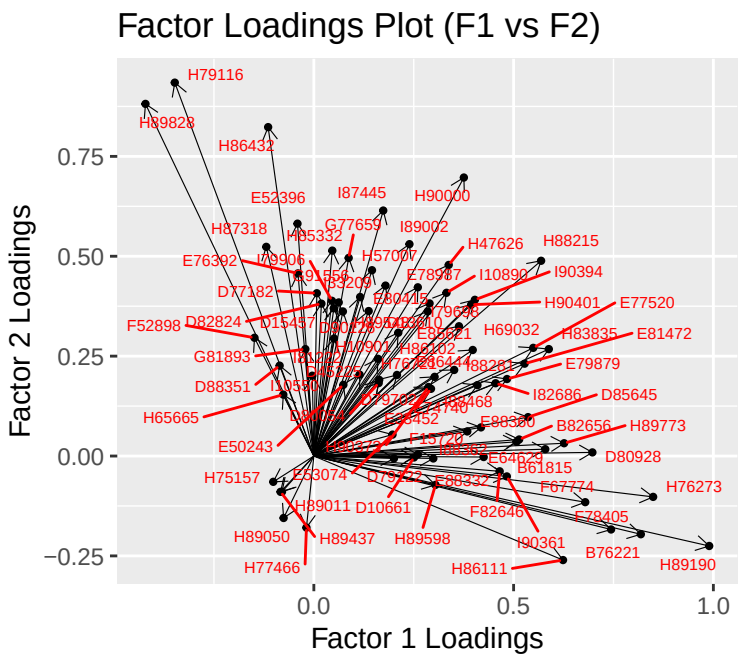

Figure 6: Factor Loadings Plot

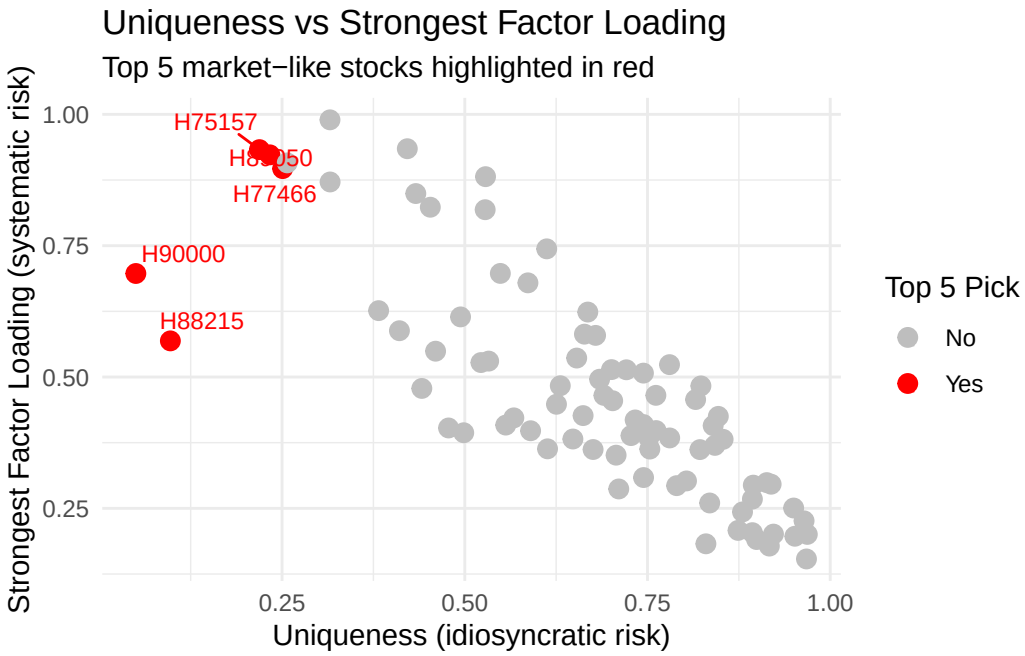

Figure 7: Uniqueness vs strongest factor loading, highlighting top 5 picks

Appendix

Check missing values and data types

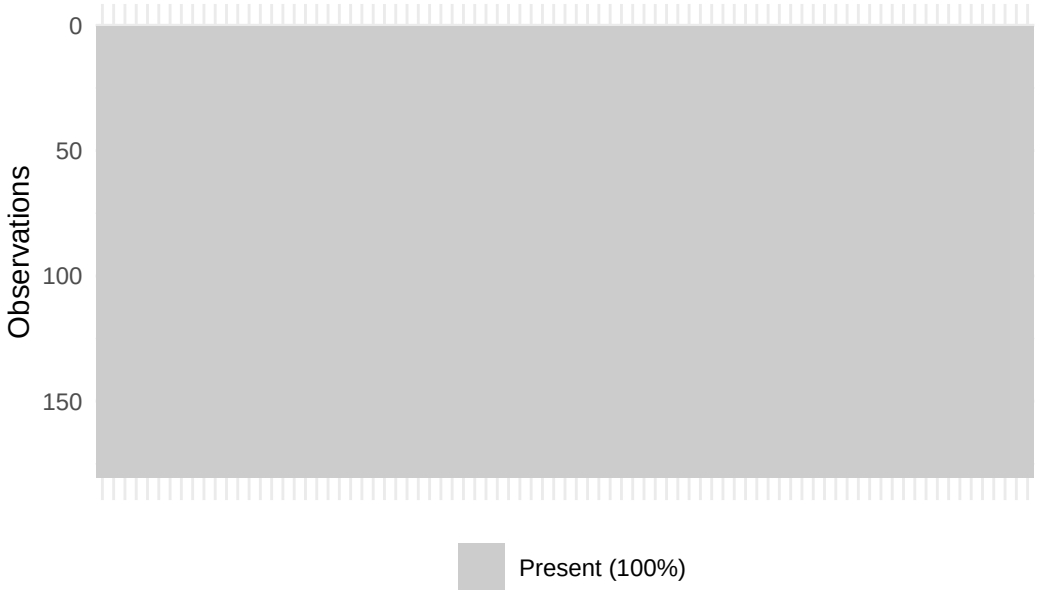

Figure 8: Missingness and data types

Figure 9: Missingness and data types

Industry summary

Figure 10 shows that industry B, D, and E have more fluctuations, while industry F and H are more stable.

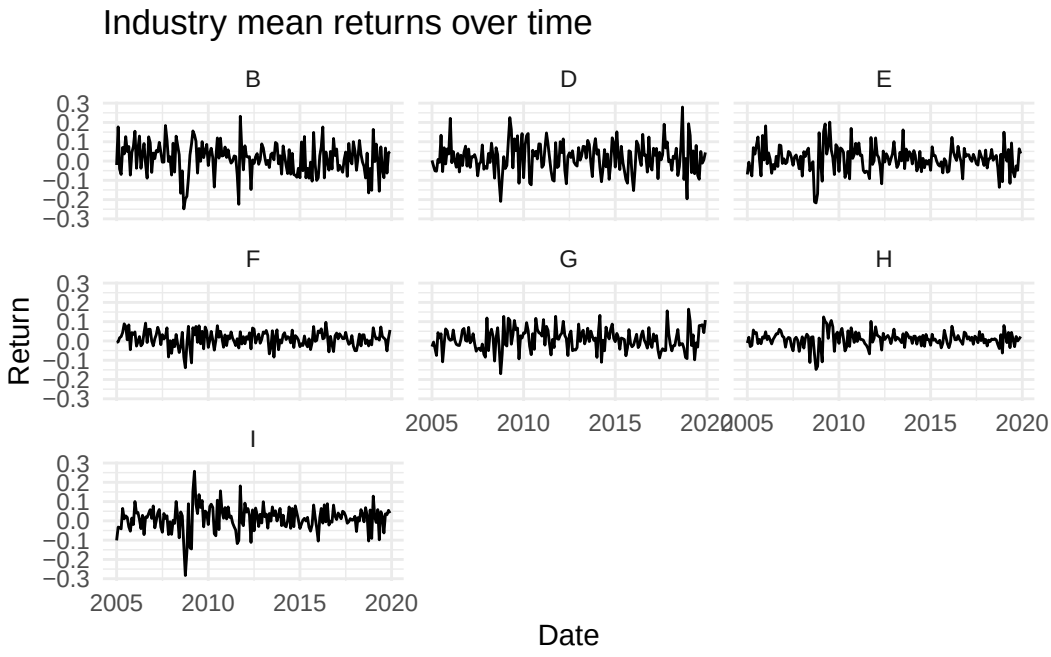

Figure 10: Stock Price Movement by industry