AFM 423 - Winter 2025

Group Project 2

"Discovering Return-Relevant Clusters: An Unsupervised Learning Approach to Factor Investing"

Raynor Sun

Jessie Li

Rachel Wu

2025/04/20

Table of Contents

1.0 ABSTRACT

In Project 1, we introduced three key papers that applied unsupervised learning to factor investing. *Clustering-Based Sector Investing* (Bagnara & Goodarzi, 2023) showed that data-driven clusters based on firm characteristics can outperform traditional sector classifications, which inspired us to explore clustering as a way to group stocks more meaningfully. *Identification of Patterns in the Stock Market through Unsupervised Algorithms* (Han & Ren, 2020) demonstrated how PCA and clustering methods can reveal hidden structures in stock market data, which supported our decision to use PCA for dimensionality reduction before clustering. *Financial Risk Assessment Based on the Factor Analysis Model* (Guo et al., 2020) helped us understand how factor analysis can be integrated into financial applications, encouraging us to think critically about the interpretability and stability of clusters. Together, these papers shaped our method in this project: we used PCA and k-means clustering to group stocks based only on firm characteristics and tested whether these unsupervised groups show different return patterns. We also compared the risk and return of each cluster to evaluate their financial relevance.

Research Questions:

1. Can we use PCA and k-means clustering to group stocks using only firm characteristics, without relying on future returns?
2. Do the stock clusters formed by unsupervised learning show different patterns in future performance?
3. Is there one or more clusters that consistently deliver better returns or better risk-adjusted performance?
4. How do these data-driven clusters compare to each other in terms of average return, volatility, and Sharpe ratio?

2.0 IMPLEMENATION

Variables and Measures

- **Dataset**: data_ml.RData from Course AFM423
- **Predictors**: 20 variables determined by Principal Component Analysis
- **Response Variable**: R1M_Usd (Return Forward 1 Month), evaluating the future performance of each cluster.
- **Performance Metrics**: Mean return, standard deviation, and Sharpe ratio for each cluster.

Application of ML to FI

We applied a PCA on the original dataset to reduce dimensionality while preserving over 90% of total variance in firm characteristics. Then, we applied k-means clustering to group stocks based on their position in the PCA space. We hypothesized that these clusters could represent data-driven economic sectors. The key assumption is that the clustering did not use any future return information, which allowed us to later evaluate whether the unsupervised clusters could reveal meaningful return differences.

Experimental Methodology

We used R for data handling, analysis, and visualization. After filtering the dataset to cover the period from 2000–2018, we split the data into a training period (pre-2007) and an out-of-sample test period (after-2007). We then applied a PCA and k-means clustering with the best-k, which was selected based on the elbow method. We backtested the equal-weighted one-month-ahead returns of each cluster from 2007 onwards. Cumulative return plots, average return, and Sharpe ratio were calculated for each cluster. The results were visualized using ggplot2 in R.

3.0 RESULTS AND DISCUSSION

3.1: PCA Analysis

We applied Principal Component Analysis (PCA) to reduce the number of features in our dataset while keeping most of the useful information. The original dataset had 93 firm characteristics, which could be noisy or highly correlated. PCA helped us turn these features into a smaller set of uncorrelated components that still explained over 90% of the total variance. This step simplified our analysis and ensured that our results were more stable. We also used PCA as a type of factor analysis to help us identify the key hidden patterns that drive differences among firms. These principal components were then used as inputs for clustering, instead of using the raw data directly.

3.2: K-means Clustering

After reducing the dataset using PCA, we used k-means clustering to group similar stocks together based on their principal component scores. These components represent key factors that summarize the most important differences in firm characteristics. To decide how many clusters to use, we applied the elbow method. This method plots the total within-cluster variation (SSE) for different values of k. As we increase the number of clusters, the SSE gets smaller. However, after a certain point, the decrease slows down. In our case, the elbow chart (*Figure 3.2.1*) showed a clear bend at k = 4 when using training data, meaning that using four clusters gives a good balance between simplicity and accuracy. We also refit the PCA and repeated the elbow method on the test data (*Figure 3.2.2*), which again suggested k = 4 as a stable choice. We then applied the k-means algorithm with k = 4 to assign each stock to one of the four groups. These clusters were later used to analyze and compare their future return performance.

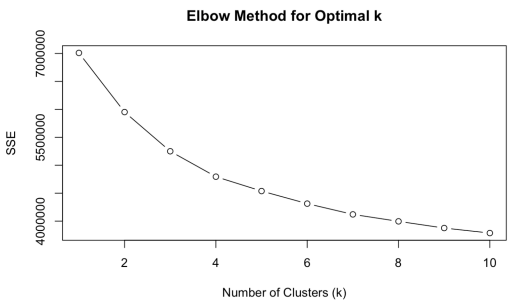

Figure 3.2.1

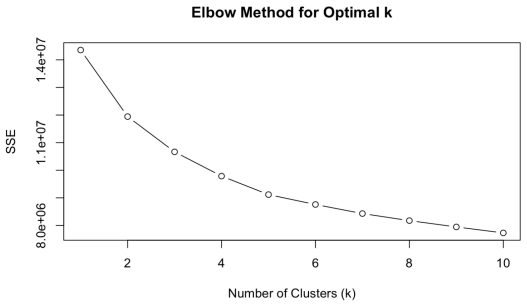

Figure 3.2.2

3.3: Backtesting

After grouping the stocks using PCA and k-means, we measured each cluster's performance over time. We selected each stock's cluster label and one-month-ahead return (R1M_Usd) from the dataset. For each month in the test period, we calculated the average return of all stocks in each cluster to form equal-weighted portfolio returns. We repeated this process to create a monthly return stream for each cluster. We then calculated cumulative returns to show how $1 would grow in each cluster. Finally, we measured performance using average monthly return, standard deviation (risk), and Sharpe ratio to compare return per unit of risk. This backtest helped to evaluate whether the clusters behave differently in terms of investment performance.

Figure 3.3.1. Monthly Returns by Cluster

The line graph of monthly returns shows that the four clusters move in similar directions most of the time, but the size of the changes varies. Cluster 1 has the highest and most stable returns, while Cluster 2 has the smallest average returns. Cluster 3 and Cluster 4 are in between, with slightly more ups and downs.

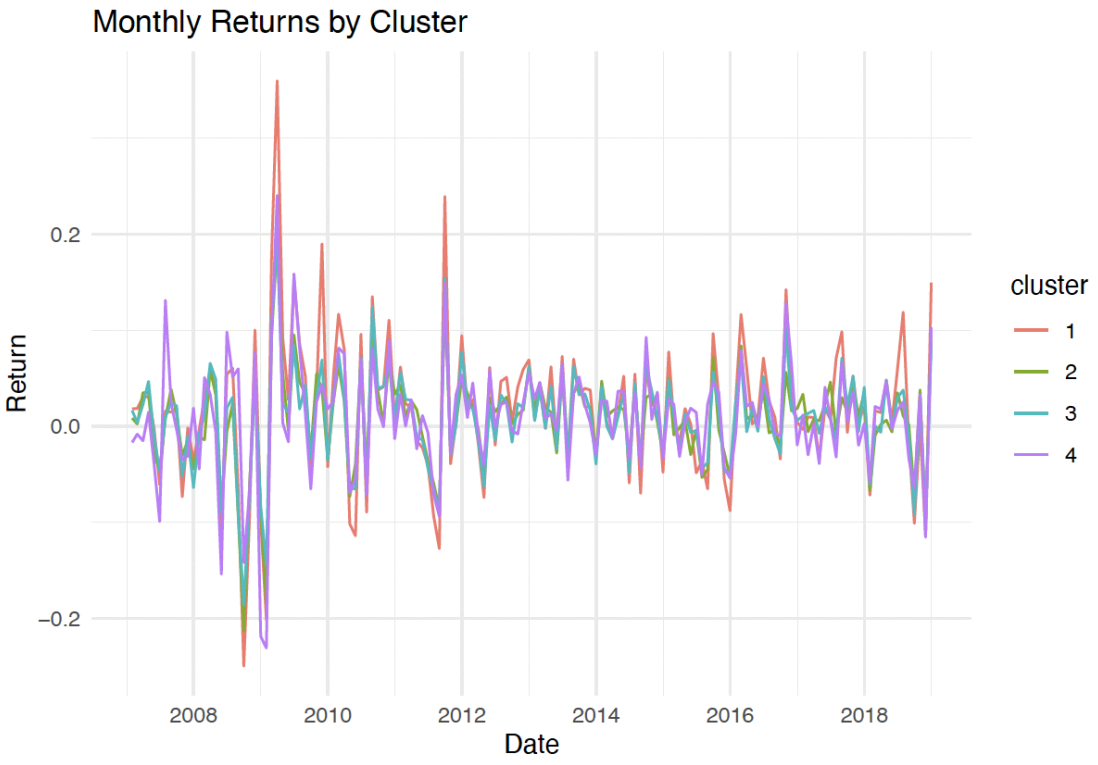

Figure 3.3.1

Figure 3.3.2. Cumulative Return by Cluster

From the cumulative return graph, it is clear that Cluster 1 performs the best. It grows the most over time and ends with the highest return. This means that if we had invested in Cluster 1, our money would have grown much more than in the other clusters.

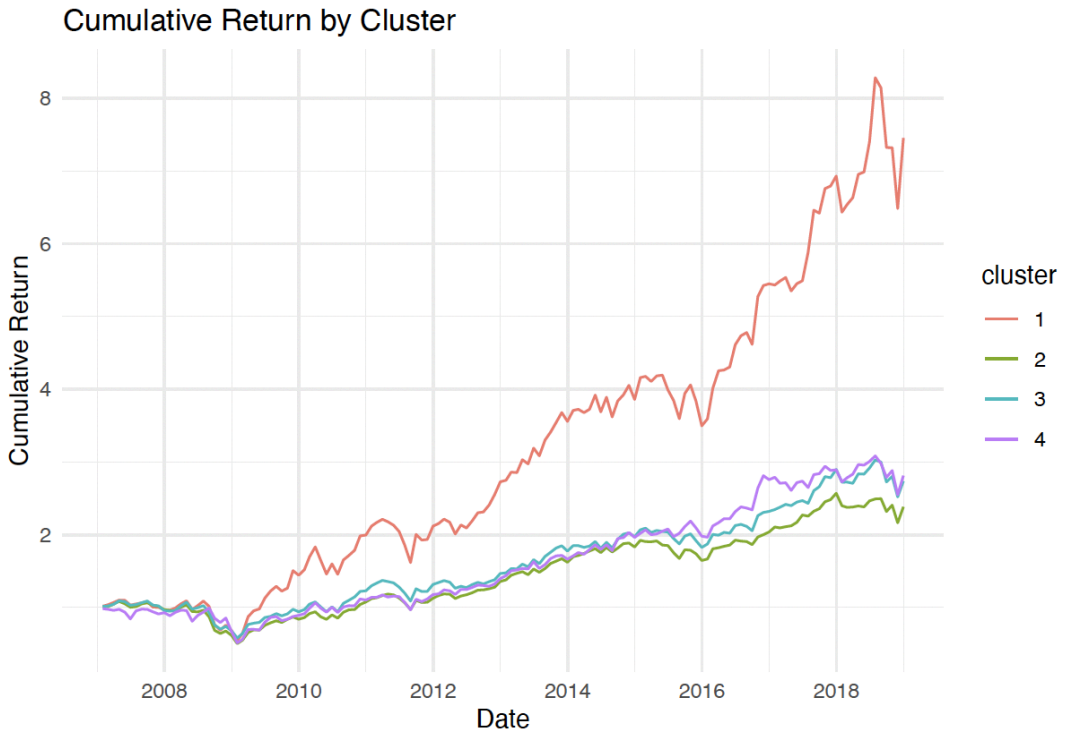

Figure 3.3.2

Table 3.3.1. Statistical summary

The summary table supports what we see in the graphs. Cluster 1 has the highest average monthly return of 0.0168 and the highest Sharpe ratio of 0.222, which means it gives more return for each unit of risk. Cluster 2 has the lowest average return at 0.00734 and the lowest Sharpe ratio of 0.146. Cluster 3 and Cluster 4 are in the middle, with returns of 0.00838 and 0.00910, and Sharpe ratios of 0.160 and 0.149, respectively. These results show that the clustering helped group stocks that perform differently in the future, even though the model did not use return information when forming the groups.

```
## # A tibble: 4 x 4
##   cluster mean_return sd_return sharpe_ratio
##   <fct>         <dbl>     <dbl>        <dbl>
## 1 1           0.0168     0.0757        0.222
## 2 2           0.00734    0.0503        0.146
## 3 3           0.00838    0.0523        0.160
## 4 4           0.00910    0.0612        0.149
```

Table 3.3.1

3.4 Comparison with Traditional Factor Models (FF)

To better evaluate the performance of the PCA and k-mean clustering method, we compared the expected return of our model to the Fama-French (FF) 5-factor regression model. The FF 5-factor model includes established economic factors — market risk premium (MKT_RF), size (SMB), value (HML), profitability (RMW), and investment (CMA) to estimate the expected return. We trained a linear regression model based on the factors mentioned above and the response variable Return Forward 1 Month (R1M_Usd).

Figure 3.4.1. Monthly Returns predicted by Cluster 4 and FF 5-factor Vs. Actual

In Figure 3.4.1, the predicted monthly returns by both methods are drawn on the same graph with the actual return. The FF 5-factor predicted line is smoother and less volatile, indicating that the main systematic components were captured by the model. However, it fails to capture most of the variations of the market. In contrast, the actual and Cluster 4 returns are more volatile, meaning that the clustering method captures more information that is not entirely explained by FF factors.

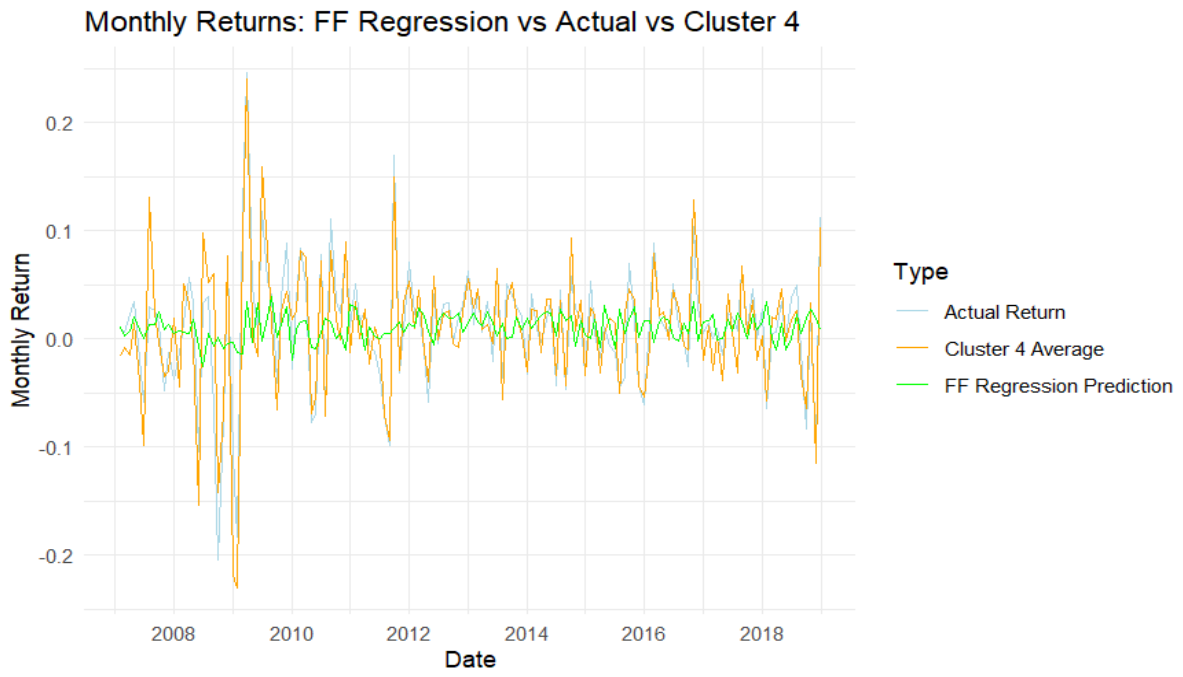

Figure 3.4.1

4.0 CONCLUSION AND RECOMMENDATIONS

This project applied an unsupervised learning approach to factor investing by combining PCA and k-means clustering to group stocks based on firm characteristics. The goal was to test whether these data-driven clusters show different return behaviors. From our analysis and backtesting results, we can now answer the research questions we proposed at the beginning.

First, we successfully used PCA and k-means to group stocks using only firm characteristics, without using future return information. Second, we found that the clusters we formed showed clear differences in their future performance. Cluster 1 consistently delivered the highest average return (0.0168) and Sharpe ratio (0.222), while Cluster 2 performed the worst. Third, this shows that some clusters do offer better return and risk-adjusted outcomes. Finally, by comparing the average return, volatility, and Sharpe ratio across clusters, we confirmed that these unsupervised groups are not random but capture meaningful differences among stocks. This suggests that machine learning can be used to support or enhance factor investing strategies.

One of the strengths of our approach is that it uses unsupervised learning to uncover hidden patterns in financial data, without relying on traditional sector definitions or return labels.

PCA helped reduce noise and kept only the most important factors. K-means clustering then grouped the stocks based on those factors, and our backtest showed clear return differences across the groups. Another strength is that our process is simple, easy to understand, and adaptable to other datasets or time periods.

However, there are some limitations to be aware of. For PCA, while it helps reduce dimensionality, the components are not always easy to interpret. For k-means, the clusters depend on the distance metric and initial centroids, and may be sensitive to scale. In backtesting, we used a static clustering approach, meaning the clusters were formed using data from the entire period. This can lead to look-ahead bias. In course material RCC2, we learned about the importance of using rolling windows to avoid this issue. While we are aware of this limitation, we chose to focus on static clustering to keep the analysis simpler and more interpretable for this stage of the project. In future work, we would like to improve our method by applying rolling PCA and clustering on a moving window to simulate a real-time investment process.

Finally, this project builds directly on the related works we reviewed in Project 1. The three papers helped us understand how PCA, clustering, and factor models have been used in finance, and gave us ideas for our model design. Overall, our results show that machine learning can help discover return-relevant groups of stocks in a way that is both practical and insightful. We recommend that future research continue exploring these methods, and focus on real-time implementation and further comparisons with traditional factor models.

5.0 RELATED WORK

Projects 1 and 2 are highly related, therefore the reader is referred to our GPR # 1 report for related work.

6.0 REFERENCES

Han, D., & Ren, X. (2020). Financial risk assessment based on Factor Analysis Model. Journal of Physics: Conference Series, 1616(1), 012056. https://doi.org/10.1088/1742-6596/1616/1/012056

Bagnara, M., & Goodarzi, M. (2023a). Clustering-based sector investing. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4528879

Barradas, A., Canton-Croda, R.-M., & Gibaja-Romero, D.-E. (2023, July 27). *Identification of patterns in the stock market through unsupervised algorithms*. MDPI. https://www.mdpi.com/2813-2203/2/3/33#B28-analytics-02-00033

7.0 APPENDIX

R code used in this report can be accessed on https://github.com/jessieli0816/AFM423-GPR2