COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Vanguard

Clustering Analysis of Investors & Polarized Topic Detection

Capstone Project Final Report

Rui Bai (rb3454)

Xinyi Liu (xl2904)

Yichi Liu (yl4327)

Yuchen Pei (yp2533)

Yujing Song (ys3251)

Columbia University

Data Science Institute (DSI)

Sponsored by Vanguard

Dec 2020

Contents

1 Introduction

1.1 Motivation and Problem Statement

With the growth of the capital market in the US, more institutional investors emerged. Vanguard, as the sponsor of this project, is interested in knowing how other market participants are approaching investments and new investment signals that are emerging. Based on this idea, we dived into 13F data, a quarterly financial report from all big institutional investors with over $100 million in asset under management, which discloses their equity holding details.

We want to use this data to get better understanding of the institutional investors and their characteristics, which gives instructions on which investors to follow. Also, understanding changes in investor behavior through time by looking at their portfolio holding changes could give insights on when to follow the actions of some smart investors.

To be more specific, we would like to cluster investors based on their investment behaviors and their characteristics. From Vanguard's perspective, they could use the cluster result to identify investors that are dissimilar to them and learn from their investments.

Furthermore, there are time points at which some industries are worthy of extra attention, such as when investors are holding opposite opinions in some industries and moving towards different directions. At such time points, investors have polarized viewpoints in a certain topic, and thus act differently. The intuition behind is that the passive investors tend to follow the market index, while active investors are more likely to have strong opinions in an investment topic and cause a large dispersion in overall position movements in the market. Such large dispersion can be identified as the indication of polarized topics. In this project, we aim to identify the polarized topics over time by measuring the level of dispersion described above. We also want to investigate whether some characteristics of an investor would affect its tendency to drive the polarized topics.

To clarify the terminologies, we define investors as investment companies and we define an instrument as a tradable asset, or negotiable item, such as a security, commodity, derivative, or index, or any item that underlies a derivative.

1.2 Related Work

Our first objective is to cluster investors based on their investment behavior. Instead of institutional investors, there are many related studies focusing on mutual funds clustering with different kinds of features. Some of the papers used the return of the funds as features. According to Pattarin [12], the mutual funds styles could be identified by analysing time series of past returns. They proposed a return-based classification scheme for the investment style consisting of three steps, including dimension reduction with PCA, clustering using the GAME algorithm, and style identification with a constrained regression model for each cluster. This could not only identify clusters but also determine the style of each class.

However, while using only return is a cheap and easy solution compared to using portfolio holdings and other information, it might not capture all the features indicating the investment styles. In Kumar's study [10], he used rates of return, standard deviation, Sharpe index, Treynor index and Jensen index as the evaluation indices for clustering. They classified 340 mutual funds in Indian stock market into five clusters with some important investment insights. Marathe and Shawky[7]'s work selected 28 financial statistics and applied K-Means to cluster 904 different funds. The three very distinctive groups they obtained showed that 45% of the mutual funds did not belong to their stated categories.

A study[14] in 2015 proposed a method using only the current investment portfolio of each fund. In the study, they

constructed a network with 551 Japanese mutual funds as vertices. They defined edge weights as the number of common stocks in each fund's top 10 instruments. They then clustered this weighted network using both K-Means and spectral clustering methods. By comparing the result with the Morningstar categories, they identified some misclassifications in Morningstar.

In our study, we would like to include a wider variety of features including time series data that could help to cluster the investors. Also, since we are studying institutional investors, we will use the portfolio data held by each institution instead of the individual funds.

We also want to identify significant signals of investment market changes in different sectors. According to Gerrit's work[17], such signals can be represented as moments of abrupt change in the behavior of a time series. He indicated that the presence of a change point signals a significant alteration and is therefore of high importance for analysts.

In financial econometrics area, the change point detection is often used to analyze the volatility processes. In Lavielle's study[5], he investigated the problems with multiple or even unknown number of change points, and proposed an adaptive method for finding the optimal segmentation. There are also works and applications of change point detection in some other domains like linguistics[3], quality control[11], and network traffic analysis[4].

However, the change point detection in time-series portfolio holdings remains yet to be explored, which motivated us to apply similar detection techniques in this area. Since we want to identify significant investment market changes in each industry, we need to aggregate multiple time-series portfolio holdings of investors into different industries by the dispersion of investors' movements, and to detect the change points of such dispersion over time. In this way, we detect the polarized topic that is defined in section 5, which requires the change point detection for several time-series data by applying aggregation techniques first, instead of the traditional change point detection for a single time series.

1.3 Approaches

In this project, we construct features for institutional investor clustering and illustrate a method to detect the investment topic with polarized opinions, defined as polarization topics, and to also identify characteristics of investors who drive such polarized topics.

We firstly construct seven important features for each investor to describe its characteristics, including: (1) three time-series portfolio holdings aggregated by industry, market capitalization, top 20% instruments, (2) time-series number of positions and total asset value, (3) turnover rate, and (4) a static feature of investment style distribution of employees. We believe those seven features could represent an investor's style and movements over time.

Using the features above, we develop a clustering model for the short-listed 225 investors using K-Means algorithm, and choose K=2 to obtain two clusters with different characteristics.

We also describe an approach to detect polarized topics in the investment market, with three steps: (1) Firstly we calculate the dispersion metrics for each instrument invested by various investors at each time point. (2) Then we aggregate the above metrics into different industries to represent the polarization level of each topic. (3) Next, a Welch's t-test is conducted to differentiate significant polarized topics from noise over time.

Furthermore, we propose the method to identify the characteristics of the investors that drive polarized topics using regression analysis. We investigate it in two aspects: (1) We built a regression model to detect significant features that decide whether an investor can drive the polarized topics. (2) We built another regression model to recognize the cluster that tend to drive polarized topics.

Our report will be developed as follows:

First we will introduce the database in section 2. Then features of investors will be defined in section 3. Based on the defined features, we build a clustering model in section 4. Next, we illustrate how we define polarized topics in section 5 and conduct regression in section 6 with features defined in section 3 and clusters in section 4 to investigate what drives the polarized topics. Finally, an interactive dashboard is shown in section 7. For the rest of the report, we conclude the results based on our study, discuss the future work and provide ethical considerations.

2 Data and Explanatory Data Analysis

We used a diverse and detailed financial database provided by Vanguard. The entire database contains over 600 tables within 23 years from 1997 to 2020 and continues to be updated. Data for more than 1 million investors and more than 12 thousand instruments are recorded in the database.

We merged 5 tables to get 13F historical holdings information and percentage of portfolio for each investor in a given time period. It contains 130 million records and serves as the main table for our further analysis. We also explored 6 other tables for some more statistics of investors, like the turnover rates, total assets and investment style of their employees. For instruments, we extracted their industry and market capitalization from 3 other tables.

After further discussion with the sponsors, we focused on the investors short-listed by Vanguard and selected 225 investors in order to avoid weird or small investors. We selected 4-year-range data from 2016 Q2 to 2020 Q2 for clustering in section 4, and 10-year-range data from 2010 Q2 to 2020 Q2 to study the polarized topics in section 5.

3 Feature Extractions and Visualization

To better describe the institutional investors, we defined seven features for each investor including their historical investment behavior and other attributes based on the core metrics that Vanguard focused on. The seven features for each investor are listed below. We visualized these features of Vanguard as an example in Figure 1, Figure 2, Figure 3.

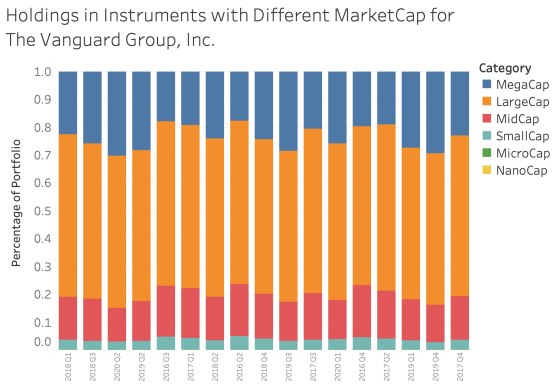
1. The time-series percentage of portfolio in different market capitalization range

   We first calculated the market capitalization of instruments by multiplying their stock price with number of shares outstanding at each time point. market capitalization indicates how much a company is worth as determined by the stock market. Based on domain knowledge, the market capitalization can be divided into nano cap, micro cap, mid cap, large cap and mega cap.

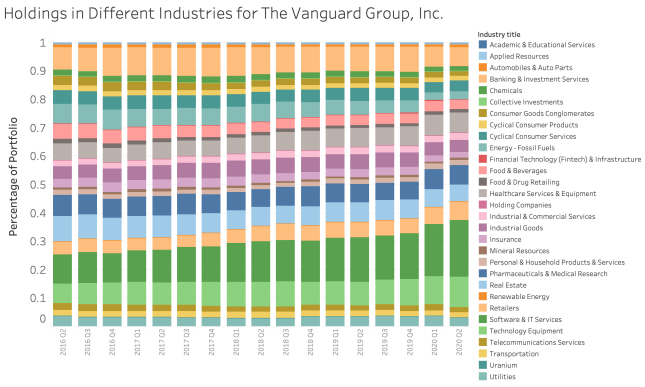
   We then aggregated holding details of each instrument by its to get the percentage of portfolio in different market capitalization range. By doing so, we could reduce the high dimensions of portfolios with too many instruments and capture the investment tendency to instruments with different volume.

2. The time-series percentage of portfolio in different industry

   We extracted the sector information of each instrument from the database and classified them into 11 sectors and 30 categories. Then we aggregated instruments by that to get the portfolio in different industry. This feature measures investors' preference to industries.



(a) Vanguard's feature 1: portfolio in different market capitalization range



(b) Vanguard's feature 2: portfolio in different industry

Figure 1: Vanguard Feature Visualization

3. The time-series percentage of portfolio in top 20% instruments

   We extracted this feature based on the fact that some investors concentrate on a few instruments while others prefer distributed investments. A higher proportion of the investment in the top 20% instruments indicates a more concentrated portfolio.

4. The time-series number of instruments

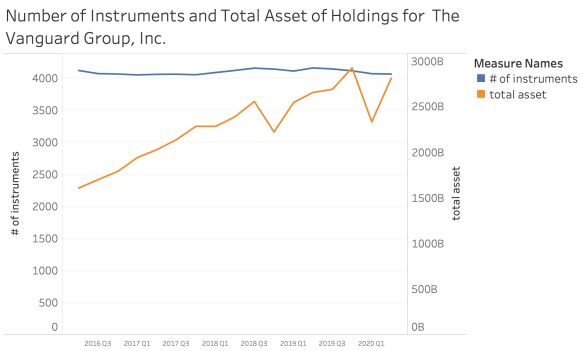
   Since we have aggregated the instruments by three perspectives discussed above, the amount of instruments is out of consideration. We calculated the number of instruments held by each investor at each time point to describe whether an investor held a large amount of instruments.

5. The time-series total assets

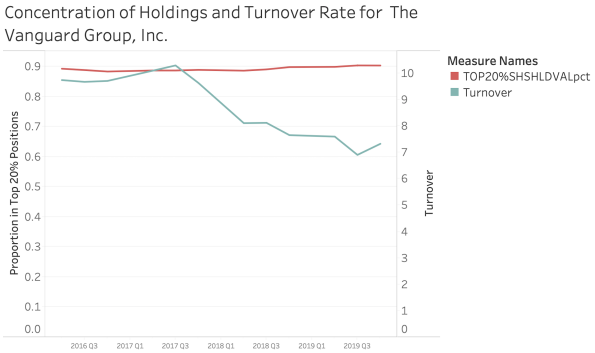
   Total assets represent the economic resource of each investor and its position in the whole market. We used this feature to indicate the size of a company.

6. The time-series quarterly turnover rate

   The turnover rate is the percentage of portfolios that have been replaced in a given period of time. A high turnover rate indicates more active investments.



(a) Vanguard's feature 3&4: number of instrument & total asset

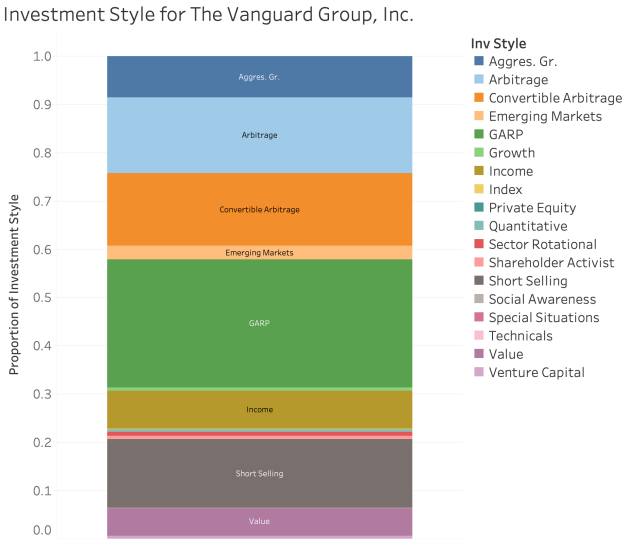


(b) Vanguard's feature 5&6: turnover rate & percentage of portfolio in top 20% instruments

Figure 2: Vanguard Feature Visualization

7. Investment style distribution of employees

   Each institutional investor has many fund managers that are specialized in different investment styles. By counting the number of employees with each type of investment style, we obtained their distribution. We denoted such distribution as the investment style of the entire institution investor, based on the assumption that features of its fund managers could represent the interested investment style of one company.



(a) Vanguard's feature7: investment style distribution of employees

Figure 3: Vanguard Feature Visualization

4 Clustering Model

4.1 Baseline Model

We used each investor's proportion of investment in different instruments at 2018 Q3 to build a baseline clustering model. Historical data and side information of the investors are not considered in our baseline model inspired by Takumasa[14]. More specifically, for each investor, we constructed a feature vector with its holding value proportion of each instrument. The data was sparse since the average number of instruments for an investor was 341 while there were 10,863 instruments in 2018 Q3. Therefore we applied sparse PCA as dimension reduction. We fitted K-means model but the current features could not separate investors well, which inspired us to include more features in our final model.

4.2 Final Model

To improve the model, more features were taken into consideration. We built clustering models on the time series features and static features that we listed in section 3. Focusing on the 225 investors on the Vanguard list, we utilized 4-year quarterly data from 2016 Q2 to 2020 Q2. Dynamic time warping (DTW) distance and Euclidean distance are two of the most popular time series metrics. DTW distance captures the similarity of the shape of two time series with wrapping the time points, while Euclidean distance captures the similarity in each time

point. Since a particular time point is important in the stock market, we chose Euclidean distance as the time series distance metric.

Since ground truth labels were not available, we used three internal metrics to evaluate the effectiveness of the clustering models. Firstly, Silhouette Score [13] measures whether the data is more similar to its own cluster compared to other features. The best value is 1 and the worst value is -1. Values near 0 indicate that the clusters are overlapping. Secondly, Calinski Harabasz Score [1] could examine cluster dispersion. It is defined by the ratio of intra-cluster dispersion and inter-cluster distance. Higher score indicates better performance. Thirdly, Davies Bouldin Score [2] measures average similarity between clusters. Less dispersed cluster will result in a lower score which means better performance.

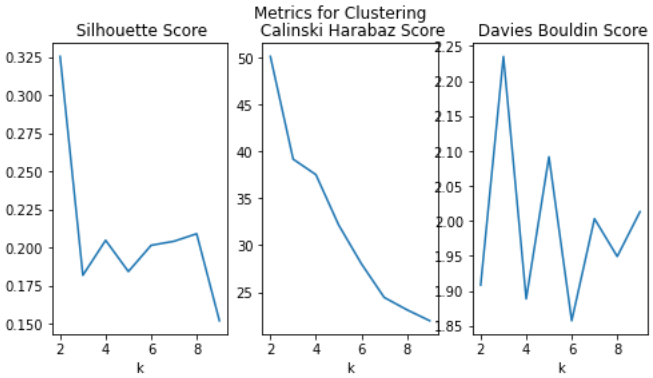


Figure 4: Internal metrics for K-Means with different numbers of clusters

Several clustering models like Gaussian Mixture Model, K-Means and Spectral Clustering have been fitted. For the ease of interpretation as well as evaluation using internal metrics, K-Means model is chosen. The K-Means model performed the best when the number of clusters was chosen as 2.

4.3 Model Interpretation

4.3.1 PCA Results

To check which features have the most impacts on clustering result, we projected the features into subspace by PCA. The investment in instruments with large market capitalization contributed the most to the first component while the time series of investment in top 20% instruments contributed the most to the second component. Investment style favoring emerging markets and investment in the Banking and Investment Services industry accounted for a great proportion of the third and the fourth components. It indicated that two of the features, namely the holding portfolio of instruments with large capitalization and the investment concentration level, made the most impact. The top 2 components could explain 37.62% of the variance.

To achieve a better understanding of the result, we used PCA and T-SNE to project the data to two dimensions. Both PCA and T-SNE are dimension reduction methods which could extract the main characteristic components of data. The difference between these two methods is that PCA is a linear transform method whereas T-SNE is a non-linear transform method. According to Figure 5, the investors have been successfully clustered into 2 groups. The group with Vanguard is shown as purple in the figure which contains 171 investors while the other group is shown as yellow with 54 investors. The list of the investor names in these two clusters has been attached in the Appendix A.

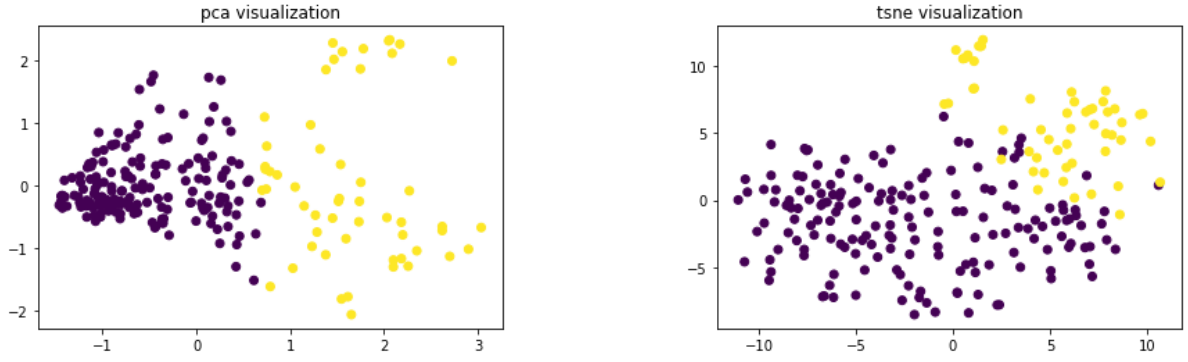


Figure 5: Visualization of the K-Means clustering results

4.3.2 Descriptive Analysis

To further explore the two clusters, we dived into the descriptive analysis of each feature for the two clusters. For convenient illustration, we named the two clusters as "the cluster with Vanguard" and "the cluster without Vanguard".

Figure 6 and Figure 7 show that the investors in the cluster with Vanguard are more likely to invest in instruments with large and mega market capitalization while the other cluster prefers instruments with small and middle market capitalization. Also, investors in the cluster with Vanguard have more total assets and more number of instruments. Although the two clusters do not differ significantly in turnover and concentration style (Figure 12 in Appendix B), they show differences in industry preference (Figure 8) and investment style (Figure 13 in Appendix B). Investors in the cluster with Vanguard invest more in the Software and IT service industry while the other cluster prefers instruments in the Energy industry. This engaging finding inspired our interest to explore whether some trends exist that some investors will act differently from others. We then dived into this part which will be shown in subsection 6.1.

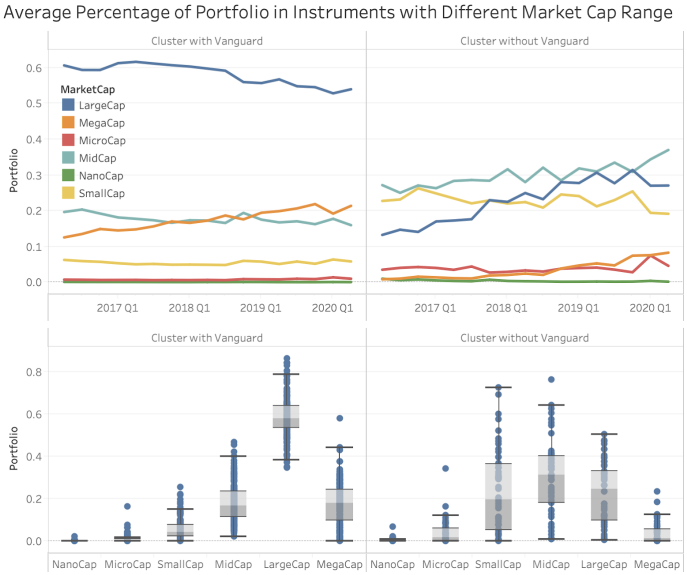


Figure 6: Feature difference of two clusters: Pct of portfolio with different market capitalization

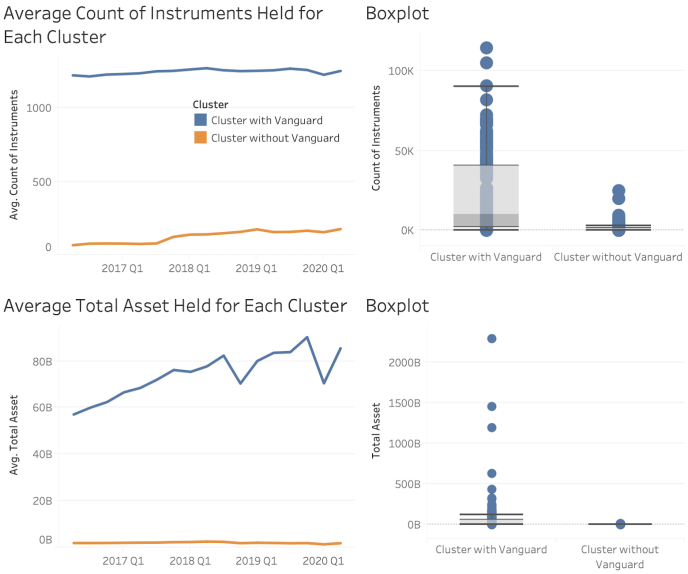


Figure 7: Feature difference of two clusters: Number of instruments and total asset

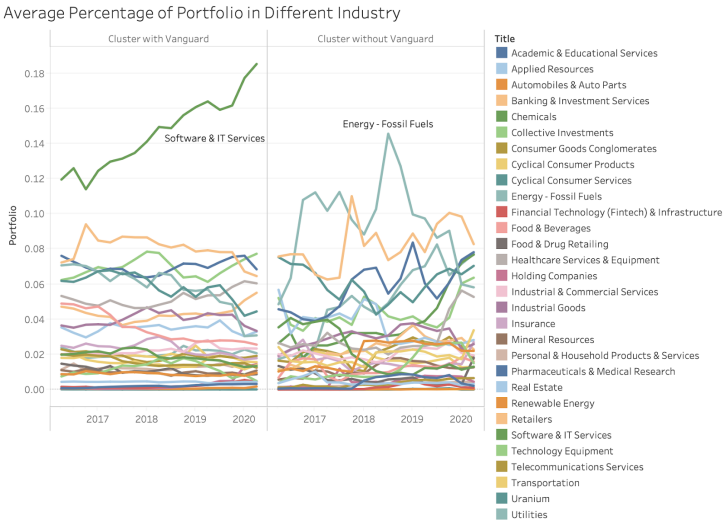


Figure 8: Feature difference of two clusters: Pct of portfolio in different industries

4.3.3 Network Analysis

To verify our K-Means clustering result, we have also conducted a network analysis using Gephi, an open-source graph visualization platform. For each feature, we calculated the Euclidean distance for each pair of investors, and thus created 7 distance matrices. Each distance matrix was then converted to an adjacency matrix using the following formula, which represented the pairwise similarity in each feature for the investors.

$$Sim_i^{'} = D_{max} - D_i + D_{min}$$

This formula can rescale all the distances, that the ones with largest distance become the ones wit smallest distance. Also, we added D_{min} to prevent the originally strongest edge becomes 0, which might be deleted in Gephi.[1]

The 7 matrices are then normalized and summed up with different weights to create a final adjacency matrix. The nodes are the 225 investors, and the edge weights are the pairwise overall similarity score of each pair of investors using all the 7 features. Then, we set a threshold of 1.55 to filter out some low-weight edges.

The network was plotted in Gephi using Layout Force Atlas 2, which arranged the edge length according to the edge weight, i.e. larger weight led to shorter length of an edge. We tuned different parameters of the layout to obtain the most reasonable clustering result. The nodes are colored based on their K-Means clustering result in subsection 4.2. This network plot could help to identify the position of each investor among the entire list of the investors, since this network layout was based on the pairwise similarity calculated by all the seven features defined in our clustering model. We could identify a few nodes that are positioned in or close to the cluster they do not belong to (as labeled in Figure 9), for example, MKP Capital Management, L.L.C. and Berkshire Hathaway Inc. After the discussion with our sponsor, we believed that one possible reason is about the nature of institutional investors, that each of the institutional investors is a combination of its child-level investors with different types, which makes them difficult to be clustered in a clear-cut way.

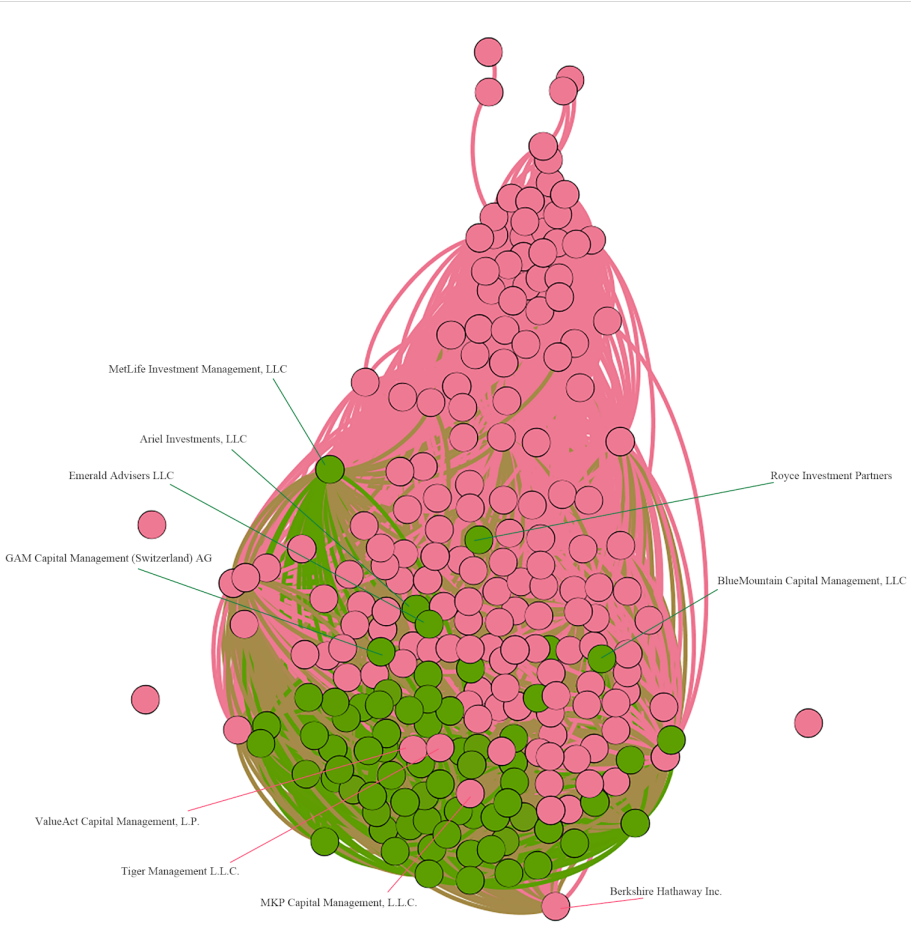


Figure 9: Gephi visualization based on similarity of the defined features, colored by K-Means results.

[1] https://github.com/gephi/gephi/issues/1816

5 Polarized Topics Detection

5.1 Definition of Polarized Topics

This section demonstrates a method to identify the polarized topics, which is defined as the industries with large dispersion of the investors' position movements, in the investment market over time. In such cases, investors have a broad disagreement with one another regarding the changes in their portfolio holdings, and thus imply the investment topics with polarized opinions (i.e. some are optimistic while others are pessimistic about a particular investment topic).

In order to measure this dispersion, we have defined three metrics as below:

For each instrument j at time t:

$$V_{jt} = Var(\Delta_{ijt}) \tag{1}$$

$$IQR_{jt} = Q_3(\Delta_{ijt}) - Q_1(\Delta_{ijt}) \tag{2}$$

$$\text{Weighted } IQR_{jt} = Q_3(Weight_{ijt} * \Delta_{ijt}) - Q_1(Weight_{ijt} * \Delta_{ijt}) \tag{3}$$

Where Δ_{ijt} is the percentage of portfolio change in instrument j of investor i at time t, $Weight_{ijt}$ is the actual dollar value change in Δ_{ijt}, Q_1 and Q_3 represent the 1^{st} and 3^{nd} quantile.

All the three metrics above are measuring the variability of the investment changes of all investors in a particular instrument j at some time point t, where larger values of the metrics represent larger dispersion in the investment changes, and thus represent larger polarization in their investment opinions.

5.2 Detecting Significant Polarized Topics

Since the number of all instruments is way too large for analysis and that portfolio changes in similar instruments may represent similar information, we aggregated each of the above 3 metrics into 11 industries, by calculating the mean metrics for each industry at each time point. The visualization of the three metrics are shown as below.

We identified the spikes in the plots as the possible indications of the polarized topics. One possible cause of such spikes is that some investors, who disagree with the other investors, start to act differently from the overall market regarding the portfolio holding changes, which results in the peaks. After the peak, it is within expectation that the metric values drop, because all investors may react similarly when one of the polarized opinions is believed to be correct (i.e. the topic is not polarized anymore).

The three plots in Figure 10 show different aspects of the polarization. While the variance (V_{jt}) is measuring the dispersion of investors' percentage portfolio changes, it is sensitive to extreme values. For example, the Academic & Educational Services industry (colored as dark blue) had a spike in the top plot at 2018 Q1, which is not reflected in the two IQR plots because it was caused by a single extremely large value. The interquartile range (IQR_{jt}) plot in the middle measures where the middle 50% of the Δ_{ijt} sit. It is not affected by extreme values, which is thus more resistant than V_{jt}. However, since Δ_{ijt} is only about percentage in the portfolio rather than the actual holding value, the spikes in the middle plot (Mean IQR_{jt} for Each Industry by Time) may be caused by some large portfolio changes of investors with very small transaction values. Thus, the bottom plot takes the actual value change of Δ_{ijt} into account by considering it as the weight when calculating the Δ_{ijt}. With this adjustment, the spikes in the two IQR plots, while still peak in the same set of industries and time points, have different heights. For example, if only considering the percentage in portfolio change, the Energy industry (colored as dark green) has a spike at June Q2 (the middle plot). This spike, however, is weakened if we take the holding value change into account (the bottom plot).

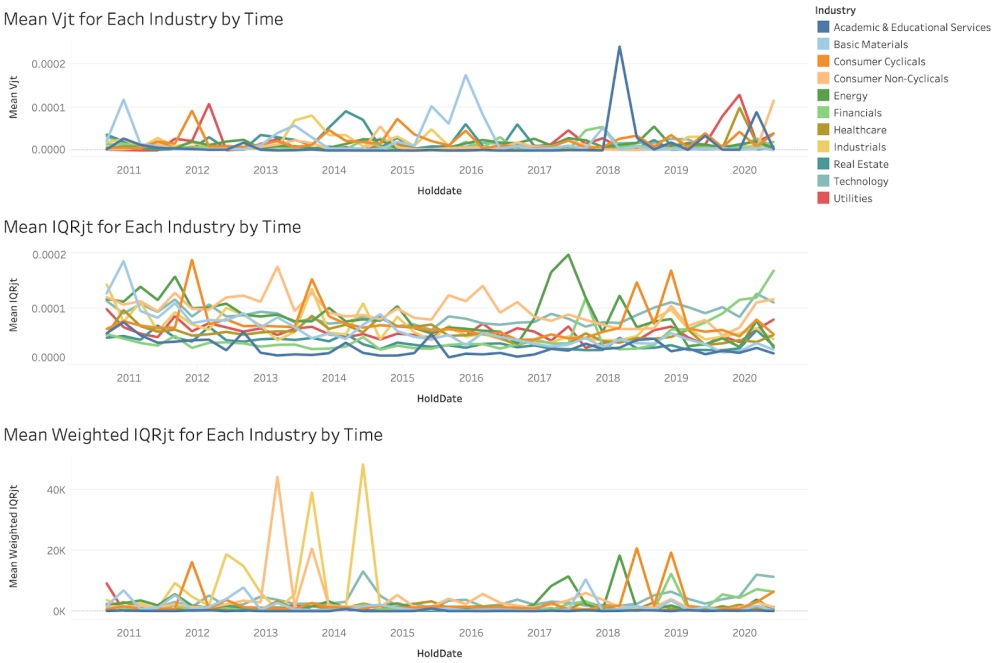


Figure 10: Visualization of the three polarization metrics in each industry

In order to recognize the significant spikes and differentiate them from the noise, we have conducted one-tailed Welch's t-test for each pair of IQR_{jt} samples and $IQR_{j(t-1)}$ samples for all instrument j in each particular industry. The resulted t-statistic is plotted as below:

Figure 11: T-test results for each industry over time

For each time point t of a particular industry, the null hypothesis is that IQR_{jt} (or Weighted IQR_{jt}) for all instruments j in this industry has the same mean value as $IQR_{j(t-1)}$ (or Weighted $IQR_{j(t-1)}$). The horizontal lines in Figure 11 decide the confidence interval with the significance level of 0.05. We will reject the null hypothesis when t-statistic is larger than 1.645, i.e. falls in the yellow-shaded area, and claim that its corresponding spike in Figure 10 is significant. A polarized topic is thus detected based on our definition described earlier.

One example of such significant spikes is at 2020 Q1 in the Real Estate industry (the t-test statistic is $4.006 > 1.645$). The spike means that the investors acted differently in the Real Estate industry at 2020 Q1. This could be caused by the COVID-19 which began to spread at the beginning of 2020. Some investors, earlier than the others, might have sensed the negative effect of COVID-19 to the Real Estate industry, due to the widespread lockdowns and travel restrictions worldwide. We could thus detect the polarized topic in the Real Estate industry at 2020 Q1. The spike also shows that after 2020 Q1, the investors tended to act similarly. According to data from Jones Lang LaSalle Incorporated [15], the investment in commercial real estate fell almost 30% globally in the first six months of 2020, which confirmed this topic we detected.

With the methodology described above, we have successfully detected 14 significant polarized topics throughout the past 10 years from 2010 Q2 to 2020 Q2, summarized in Table 1.

Time Point	Polarized Topics
2011 Q3	Utilities
2014 Q4	Energy, Real Estate
2015 Q3	Industrials, Technology, Utilities
2016 Q1	Utilities
2018 Q4	Financials
2020 Q1	Energy, Basic Materials, Consumer Cyclicals, Utilities, Real Estate
2020 Q2	Financials

Table 1: Significant Polarized Topics in the Past 10 Years

6 Polarized Topics Analysis

6.1 Regression Analysis Using Features of Investors

We then dived into the polarized topics detected in subsection 5.2 to find out significant characteristics of investors that decide whether they can drive the polarized topic.

Features defined in section 3 were used as our independent variables in the regression analysis, excluding the industrial portfolio information since this analysis was by industry. We collected these data for all the pairs of time point and industry where we detected a polarized topic. The dependent variable y_i of each investor i for a fixed time point t and a specific sector S was defined in Equation 4. n_i represents number of instruments i invests in this industry S at time point t.

$$\begin{aligned} \Delta_i &= \frac{1}{n_i} \sum_{j \in S} \Delta_{ij} \\ y_i &= |\Delta_i - \bar{\Delta}| \end{aligned} \tag{4}$$

We applied LASSO regression model to fit the data and identified five significant characteristics of investors who drive the polarized topics as follows. Detailed result is shown in Figure 14 (in Appendix C).

- larger percentage of portfolio investment in mid-cap instruments
- higher concentration in top 20% of its instruments
- more employees with the investment style of Emerging Market
- smaller number of instruments in its holdings
- fewer employees with the investment style of Arbitrage / Index / Private Equity / Social Awareness

We believed investors with the above listed characteristics should be paid more attention to, since they are probably revealing some potential polarized topics in the investment market. For example, investors named Anchorage Capital Group, Apollo Capital Management and Bracebridge Capital meet these characteristics. They are investors who caused a large dispersion in investment actions and hence drove the polarized topics.

6.2 Regression Analysis Using Clustering Results

We also did regression analysis for clustering results from section 4.3.1 to see which cluster may lead the polarized topic. The independent variable for this part was the cluster label of each investor, while the dependent variable was the same as in section 6.1.

The regression result in Figure 15 (in Appendix C) shows that the clustering result of each investor has a significant correlation (with the significance level of 0.05) with whether it tends to drive the polarized topic. In other words, investors in cluster B (dissimilar to Vanguard) are more likely to disagree with the market mean, which results in the polarized topic we detected. This cluster contains 54 investors, like Apollo Capital Management, King Street Capital Management, Angelo, Gordon & Co., Brookfield Asset Management, etc. We need to track changes in their portfolio holdings so that we may detect potential new polarized topics in time.

7 Dashboard

Finally, we created a tableau dashboard with 10 tabs to visualize all our findings. In the dashboard, users can select their interested investors and get a view of the change of each specific feature across time, as well as our clustering results and polarized topics results. For the demo of our dashboards, see the link to the DEMO.

The first 7 tabs visualized the seven features defined in section 3, each containing two plots: (1) the boxplot showing the overview of the whole investment market, and (2) a plot visualizing the feature for the selected investor, such as a line chart or a stacked bar chart. Additional filters could be applied in some tabs such as the tabs for the percentage of portfolio, where user could filter by different market capitalization or industry and interact with the plots.

In the tab for clustering, the Gephi network was embedded and colored by the k-means clustering results. Users could highlight any interested investor and visualize its cluster and connections.

The last two tabs visualized the polarized topics. The first tab showed the time series of interquartile and weighted interquartile for different industries. The second one showed the t-test result, where users could identify significant polarized topics.

8 Discussion

8.1 Conclusion

In conclusion, we conducted two parts of analysis for this project and obtained the insights below. The clustering model was utilized to understand the characteristics of the institutional investors and how they differ from Vanguard. After detecting significant polarized topics over the past 10 years, linear regression models were applied from both feature and cluster perspectives. We have three key findings:

First of all, we clustered all the 225 investors into two groups. The differences between these two groups can be concluded into 4 aspects: investment tendency to different market capitalization, investment tendency to industries, total assets, number of instruments, and the employees' investment style. Table 2 shows the differences between the two groups of investors:

Cluster similar to Vanguard	Cluster dissimilar to Vanguard
1. Invest more in instruments with large and mega market capitalization 2. Invest more in the Software and IT service industry 3. More total assets and more number of instruments 4. Their investment style has more focus on Value	1. Invest more in instruments with small and middle market capitalization 2. Invest more in the Energy industry 3. Less total assets and less number of instruments 4. Their investment style has more focus on Emerging Market

Table 2: Descriptive comparison between two clusters

Next, to understand the changes in the market, we defined the polarized topics as industries with a large dispersion of the investors' position movements. Table 1 listed significant polarized topics detected over the past ten years.

Finally, we identified what could decide whether an investor tend to drive polarized topics.

- From the perspective of features, we got five characteristics of the investors who are more likely to drive the polarized topics. These investors usually have a higher percentage portfolio investment in mid-cap instruments, higher concentration on top 20% of its instruments, fewer employees with the investment style of Arbitrage / Index / Private Equity / Social Awareness, fewer positions, and more likely to invest in Emerging Markets.
- From the perspective of clustering, we found that investors dissimilar to Vanguard are more likely to drive the polarized topics.

8.2 Limitations and Future Work

For later work, we consider three methods that can be helpful to improve our model.

- Include Returns Data to Evaluate Performance

  In this project, we identified polarized topics over the past 10 years. However, it might be more insightful if we can evaluate each investor's performance. For example, we want to identify which investors hold the correct opinion in a specific time point when a polarization exists, and also identify investors that persistently hold the correct view.

- Expand the Target Investor List

  In later work, we might try to expand our target investor list. For this project, we selected 225 investors that are short-listed by our sponsors. However, it will be more helpful to the general case if we can improve the model to represent the whole investment market, instead of just the list of 225 investors.

- Use the Data with a Smaller Time Interval

  Also, we now only build the model based on the quarterly data. But we will try to analyze the data with a smaller time interval in the future, since the movement intervals of active investors is shorter than a quarter. For example, it might be more effective for detecting a change in the investment market using monthly data. We hope this could be helpful to build a more useful model in the future.

8.3 Ethical Considerations

Some ethical considerations of using this work could be addressed as below.

- Social Responsible Investment

  If this model is to be used for instrument recommendation in the future, such as recommending the instruments that are invested by polarized topic drivers, the user of this model needs to be social responsible. For example, regardless of the potential high financial return, the investor should avoid investing in sin stocks such as gambling, firearms, tobacco, and be cautious in the controversial areas such as alcohol and oil, because the investor's moral value is more important than the optimal investment strategy that a model could suggest. Also, the notion of sustainable investing needs to be taken into account, which means that the investors should seek funds that support environmental and social changes rather than those with poor environmental, social and governance credentials [16], instead of simply following the model's recommendations.

- Responsible Mentoring

  However the model in this study is to be used in real-world applications, responsible mentoring is required. Further work must be done to ensure that the usages of this model obey the regulations set by both the Securities and Exchange Commission (SEC) and the Financial Industry Regulatory Authority (FINRA), which are two of the most important regulatory bodies in the U.S. financial system. Thus, any party that plans to use this work should invest time, energy and resources in mentoring the model's work, to ensure its ethical behavior based on the model's result. For example, the investor needs to make sure that it has supervisory systems in place for producing records in order to enable itself and FINRA to reconstruct accurately [6], regardless of what the model suggests.

- Religion-related Issues

  While making use of this work, investors should take note of religion-related issues and policies. For example, for Islamic investments, every action needs to be in accordance with Islamic Investment Policy [8]. If any investor plans to use this model for Islamic investments, it needs to avoid investing in an instrument that is prohibited by Islamic law, or sharia, such as the instruments with too much debt as a percentage of their assets, or the instruments in some specific industries like pork and pornography. Thus, the investor needs to evaluate the instrument's business activities and financial statements to determine whether it is considered halal to invest [9].

Acknowledgement

We would like to express our sincere gratitude to our supervisor Adam S. Kelleher and the mentors from the Vanguard Group for all their kind help and guidance for this project.

References

[1] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[2] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[3] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change, 2014.

[4] Barış Kurt, Çağatay Yıldız, Taha Yusuf Ceritli, Bülent Sankur, and Ali Taylan Cemgil. A bayesian change point model for detecting sip-based ddos attacks. *Digital Signal Processing*, 77:48–62, 2018.

[5] Marc Lavielle and Gilles Teyssière. *Adaptive Detection of Multiple Change-Points in Asset Price Volatility*, pages 129–156. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[6] FINRA Manual. Prohibition against trading ahead of customer orders. 2020.

[7] Achla Marathe and Hany Shawky. Categorizing mutual funds using clusters. *Advances in Quantitative Analysis of Finance and Accounting*, 7, 01 1999.

[8] L. R. MARC. What is an islamic investment policy? 2019.

[9] N.A. Halal investment guidelines. 2019.

[10] M Nooney, Lenin Kumar, V Dr, and Rama Devi. Cluster analysis of mutual funds. 09 2011.

[11] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[12] Francesco Pattarin, Sandra Paterlini, and Tommaso Minerva. Clustering financial time series: An application to mutual funds style analysis. *Computational Statistics & Data Analysis*, 47:353–372, 09 2004.

[13] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[14] Takumasa Sakakibara, Tohgoroh Matsui, Atsuko Mutoh, and Nobuhiro Inuzuka. Clustering mutual funds based on investment similarity. *Procedia Computer Science*, 60:881 – 890, 2015. Knowledge-Based and Intelligent Information & Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.

[15] C. Sean. Global commercial real estate markets feel impact of covid-19. 2020.

[16] F Tony. Why it pays to go green. 2020.

[17] Gerrit J. J. van den Burg and Christopher K. I. Williams. An evaluation of change point detection algorithms. 2020.

Appendix

A Some Investors in K-Means Clustering Result

We listed some of the investors that are classified in different clusters by K-Means in the table below.

Cluster 0	Cluster 1
JP Morgan Asset Management	Heartland Advisors, Inc.
UBS Financial Services, Inc.	Angelo, Gordon & Co., L.P.
Goldman Sachs Asset Management (US)	New York Life Investment Management, LLC
The Vanguard Group, Inc.	Franklin Templeton Investimentos (Brasil), Ltda.
BlackRock Institutional Trust Company, N.A.	Discovery Capital Management, LLC
Citadel Advisors LLC	King Street Capital Management, L.P.
Berkshire Hathaway Inc.	Spectrum Asset Management, Inc.
Brown Brothers Harriman	Company & Silver Point Capital, L.P.
BofA Global Research (US)	Brookfield Asset Management, Inc.
Charles Schwab Investment Management, Inc.	Quilter Cheviot Investment Management
Canyon Capital Advisors LLC	Apollo Capital Management, L.P.
MKP Capital Management, L.L.C.	TIAA Endowment & Philanthropic Services, LLC
BNP Paribas Securities Corp. North America	MetLife Investment Management, LLC
Two Sigma Investments, LP	Emerald Investment Partners LLC
Northern Trust Global Investments	Franklin Templeton International Services SARL

Table 3: K-Means clustering result

B More Descriptive Analysis of the Two Clusters

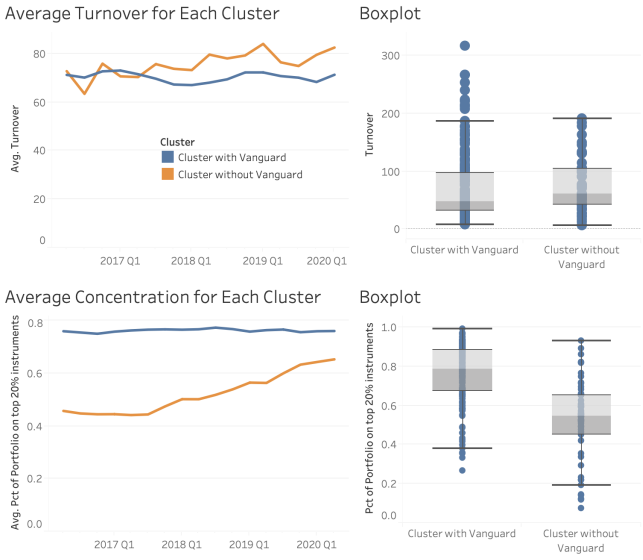


Figure 12: Feature difference of two clusters: Turnover and concentration investment

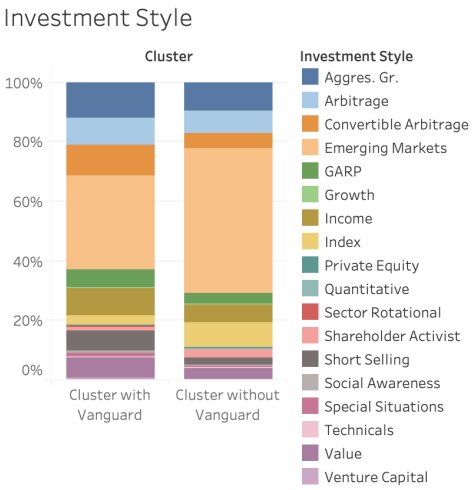
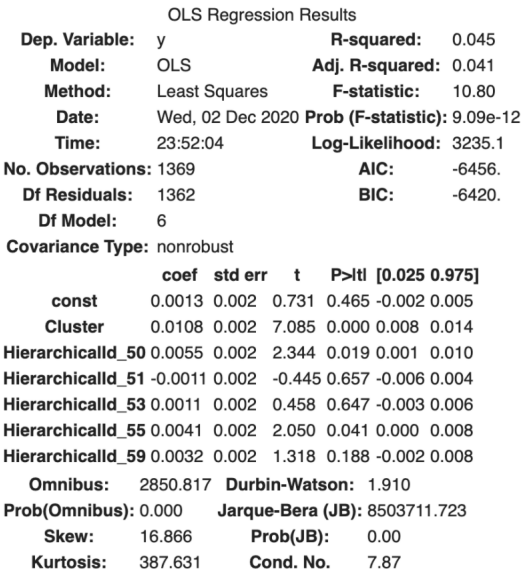


Figure 13: Feature difference of two clusters: Investment style

C Regression Result

	s0
(Intercept)	0.0154277889
Turnover_Max	.
PctOfLargeCap	.
PctOfMegaCap	.
PctOfMicroCap	.
PctOfMidCap	0.0115340719
NumberOfInstr_Max	-0.0234034281
TOP20%SHSHLDVALpct	0.0027260264
TotalAssets_Max	.
InvStyle_Arbitrage	-0.0005510147
InvStyle_Convertible Arbitrage	.
InvStyle_Emerging Markets	0.0042905696
InvStyle_GARP	.
InvStyle_Income	.
InvStyle_Index	-0.0068695217
InvStyle_Private Equity	-0.0191547034
InvStyle_Quantitative	.
InvStyle_Sector Rotational	.
InvStyle_Shareholder Activist	.
InvStyle_Short Selling	.
InvStyle_Social Awareness	-0.0031062604
InvStyle_Special Situations	.
InvStyle_Value	.
InvStyle_Venture Capital	.
HierarchicalId_50	0.0020045683
HierarchicalId_51	-0.0015019441
HierarchicalId_53	-0.0004349306
HierarchicalId_55	0.0009340004
HierarchicalId_59	0.0003624638

Figure 14: LASSO regression result w.r.t investor features



OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.045
Model:	OLS	**Adj. R-squared:**	0.041
Method:	Least Squares	**F-statistic:**	10.80
Date:	Wed, 02 Dec 2020	**Prob (F-statistic):**	9.09e-12
Time:	23:52:04	**Log-Likelihood:**	3235.1
No. Observations:	1369	**AIC:**	-6456.
Df Residuals:	1362	**BIC:**	-6420.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.0013	0.002	0.731	0.465	-0.002	0.005
Cluster	0.0108	0.002	7.085	0.000	0.008	0.014
HierarchicalId_50	0.0055	0.002	2.344	0.019	0.001	0.010
HierarchicalId_51	-0.0011	0.002	-0.445	0.657	-0.006	0.004
HierarchicalId_53	0.0011	0.002	0.458	0.647	-0.003	0.006
HierarchicalId_55	0.0041	0.002	2.050	0.041	0.000	0.008
HierarchicalId_59	0.0032	0.002	1.318	0.188	-0.002	0.008

Omnibus:	2850.817	**Durbin-Watson:**	1.910
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8503711.723
Skew:	16.866	**Prob(JB):**	0.00
Kurtosis:	387.631	**Cond. No.**	7.87

Figure 15: Linear Regression result w.r.t. clustering result