Clustering

After looking into specific sectors, this part is going to clustering companies by their ability to stand the storm using cluster analysis based on retention value in a general way. The algorithms to be used here are k-means and k-median. We will focus mainly on 3 clusters and label them as high-value cluster, median-value cluster, and low-value cluster, separately. The criterion to choose the best model is by picking one with the larger number between the clusters with the smallest size. Ultimately, we want to see what factors have led to the formations of such clusters and the patterns behind them.

Algorithms Comparisons and Results

Differences between the two algorithms

The main difference between the two algorithms is the method to calculate distance between variables. Specifically, k-means uses Euclidean distance：

$$c = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

k-median uses Manhattan distance：

$$c = |x_1 - x_2| + |y_1 - y_2|$$

In addition, the centroids are determined by means and medians, separately.

Results Comparisons

The numbers of each cluster using the two algorithms are shown as follows:

Table 1 Results for Different Clustering Methods

	K-means	**K-median**
Cluster	**Number**	**Number**
High Value	106	121
Median Value	245	217
Low Value	154	167

The minimal number for each cluster is 106 vs. 121. According to the criterion mentioned above, 121 is larger than 92 and the number for each group is more balanced. Thus, we will use the results of K-median.

The graph is also shown as follows:

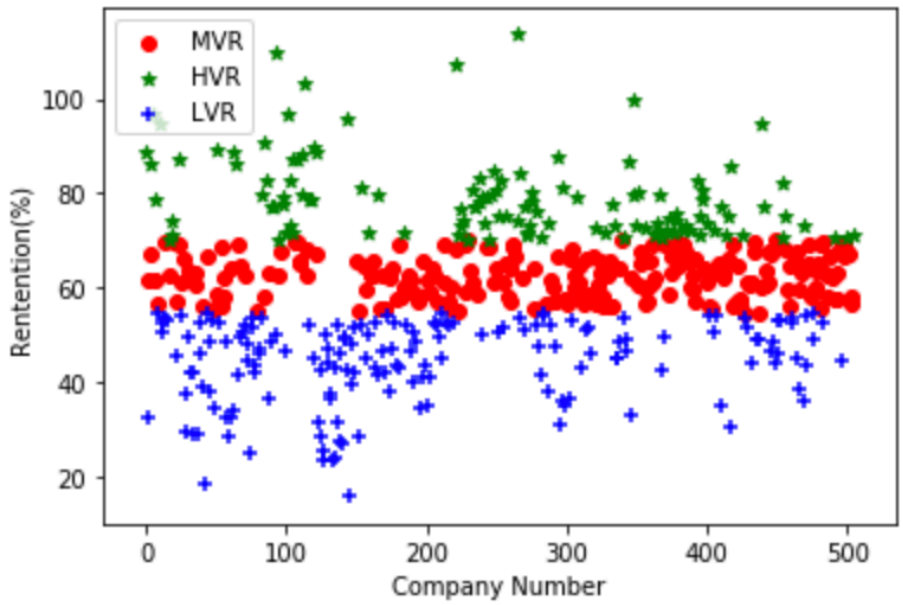

Figure 1 The Graph for k-median Clustering

From the graph, we can clearly see the three groups with three clear horizontal boundaries. The group with the highest average retention is labeled as high-value group. The group with the lowest average retention is labeled as low-value group.

Next, we further look into the changes in absolute quantity of companies for each sector in different cluster groups. The detailed number is shown below:

Table 2 Company Number for Each Sector in Each Group

	6.53%	11.88%	5.15%	14.06%	12.67%	5.54%	5.54%	6.14%	14.05%	13.07%	5.35%	100%
	Consumer Staples	**Health Care**	**Communication Services**	**Information Technology**	**Consumer Discretionary**	**Utilities**	**Materials**	**Real Estate**	**Industrials**	**Financials**	**Energy**	**Count**
Retention（%）	76.5	72.07	71.01	69.24	68.99	63.9	63.4	61	57.9	56.17	44.23	-
HVR	20	30	9	29	6	3	3	4	12	4	1	121
MVR	9	24	10	35	19	23	20	12	35	30	0	217
LVR	4	6	7	7	39	2	5	15	24	32	26	167

The retention is arranged from left to right in descending order.

The general patterns are:

- Sectors with higher rollup retention tend to have more companies in both high-value group and median-value group, and few in the low-value group, while the last few sectors tend to have more companies divided into the median and low-value group.
- Most sectors occurred in all three groups except the energy sector, which only occurred in high-value and low-value group.
- The average retention for each group is: 79.2%, 62.5%, 44.4%, respectively. Thus, viewed in sector level, consumer staples, health care, communication services, and information technology can be regarded as more able to stand the crash storm. However, industrials, financials, and energy sectors, when viewed in sector level, are hard to stand the crash, losing almost half of its value.

On the next page, you will see the visualizations of changes in the absolute quantity of companies for each sector in different cluster groups. The graphs are separated according to bullet three above. The figures are meant to show the overall trend.

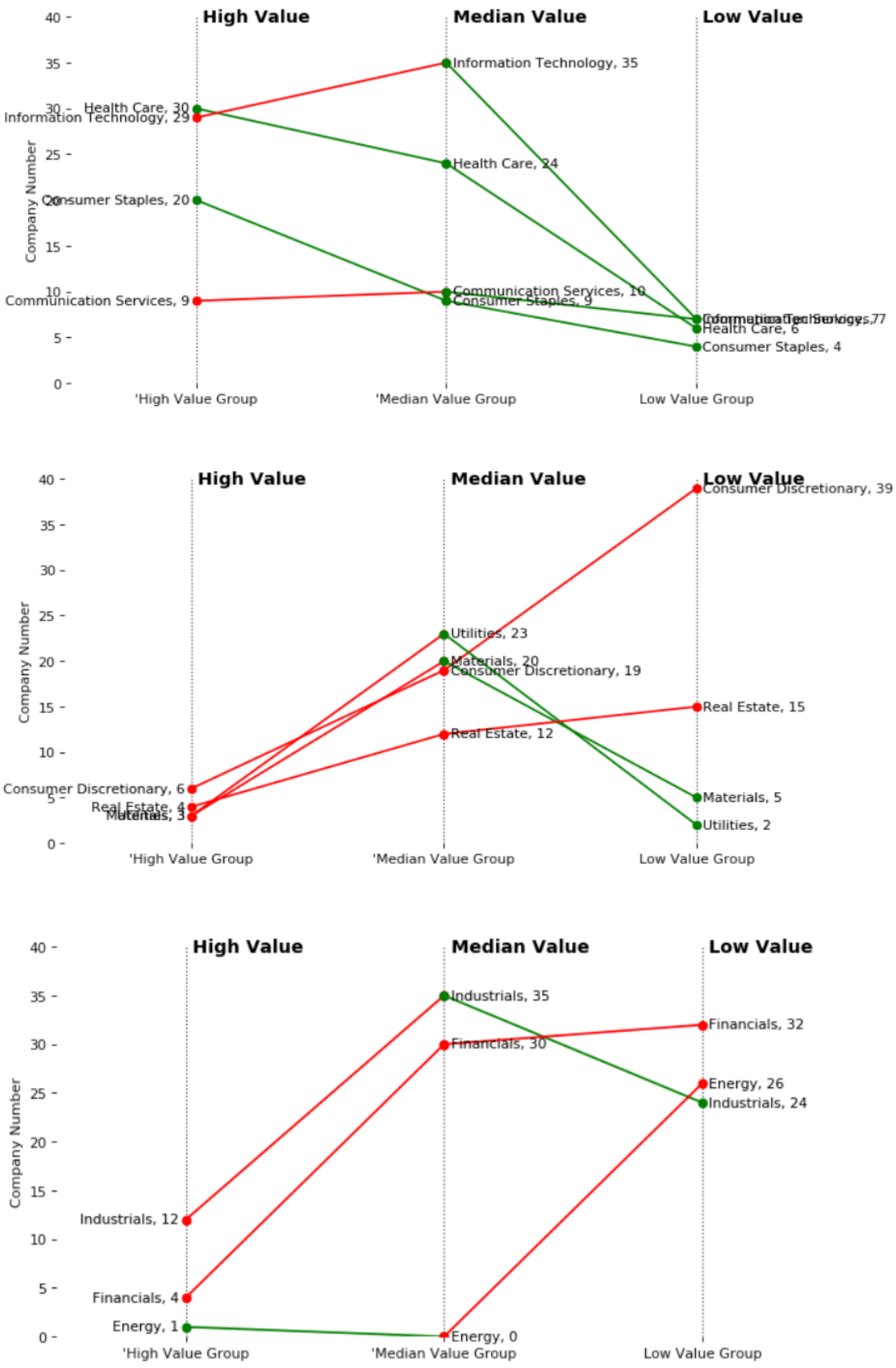

Figure 2 Change of Company Number in Different Sectors and Different Groups

Cluster Decomposing

As we can see from the graphs above, there are several industries which are not in line with the general patterns. For example, the consumer discretionary sector has plenty of companies in the low-value group, while its overall retention is still high.

Thus, by comparing the abnormal sector and relatively healthy sectors, we may find the reasons behind the issue and more into what factors may play a more critical role in one industry.

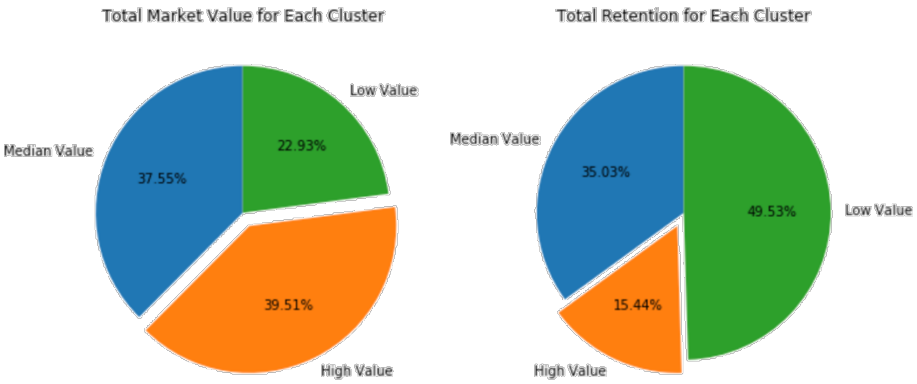

Figure 3 Graphs of Consumer Discretionary Sector

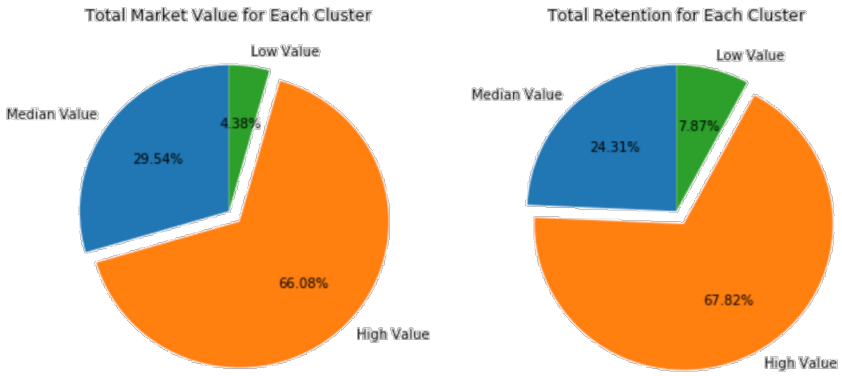

Figure 4 Graphs of Consumer Staples Sector

Here, first is to compare Consumer Discretionary (Median value in sector level) and Consumer Staples(High value in sector value). The total retention for the high-value group in the consumer discretionary sector is small, while the total market value for these companies accounts for a more significant portion. The portion of total market share for high-value companies and their total retentions are both the largest. This means that the rollup retention for an individual sector is mainly determined by the performance of 'big ones' (companies with higher market share). Thus, even though the portion of total retention for low-value is higher than those for median-value and high-value companies, the market shares of these low-value companies are small.

Customer Staple Sector

Security	Retention	MktCap_0214
Walmart	96.94	334474.6
Procter & Gamble	77.45	311496.8
PepsiCo Inc.	71.56	204249.2
Costco Wholesale Corp.	89.70	140615.9
Mondelez International	70.26	85518.0
Colgate-Palmolive	78.85	65726.7
Kimberly-Clark	77.18	49751.8
Walgreens Boots Alliance	82.52	46817.8

Customer Discretionary Sector

Security	Retention	MktCap_0214
Amazon.com Inc.	89.13	1062760.6
Target Corp.	82.90	59100.8
Dollar General	88.51	40234.5
Dollar Tree	86.12	20987.2
Tiffany & Co.	90.62	16264.3
Tractor Supply Company	79.52	11622.0

Healthcare Sector

Security	Retention	MktCap_0214
Johnson & Johnson	74.03	395123.0
Merck & Co.	80.34	210425.6
Pfizer Inc.	78.03	202050.8
Abbott Laboratories	70.06	158559.7
Thermo Fisher Scientific	75.27	136012.2
Lilly (Eli) & Co.	84.36	135493.7
Amgen Inc.	83.20	131810.0

Financial Sector

Security	Retention	MktCap_0214
Berkshire Hathaway	71.49	303912.0
Progressive Corp.	79.44	48971.9
MSCI Inc	71.79	27323.2
MarketAxess	81.38	13109.8

Figure 5 Companies in High-value Group in Different sectors

Then, we would like to see which kind of companies could 'survive' the crash. We picked up sectors from different groups. Except financial sectors, other three sectors are all occupied by flagships of the industry. However, for financial sector, only one of them has large market share and they are financial service companies. On the contrary, those in the low-value group are financial companies. The companies and industries sub-categories can be seen as follows:

Security	Retention	MktCap_0214
Bank of America Corp	51.88	307939.8
Wells Fargo	52.36	199362.0
Citigroup Inc.	44.92	166577.3
American Express Co	50.75	109788.5
Morgan Stanley	49.80	90382.5
U.S. Bancorp	52.50	84538.7
Truist Financial	46.76	73738.6
PNC Financial Services	52.32	66738.3
MetLife Inc.	45.01	47851.9
Capital One Financial	41.77	46865.4

Figure 6 LVR Companies for Financials

Figure 7 LVR Co. Types for Financials

Sub Type	Count
Regional Banks	10
Life & Health Insurance	6
Diversified Banks	5
Consumer Finance	4
Property & Casualty Insurance	3
Asset Management & Custody Banks	2
Multi-line Insurance	1
Investment Banking & Brokerage	1

Conclusions

In this part, we could generate the following conclusions:

- Rollup retention for a particular sector is mainly determined by the performance of companies with higher market shares.
- Large companies with high market value can withstand the crash in industries like consumer staples, healthcare and consumer discretionary, etc.
- Large companies with high market value have seen a marked drop and could not retain their market price in sectors like financials and energy. Even regional banks can’t tackle the market disturbance well.
- During the outbreak, due to fear of shortages, consumer staples and consumer discretionary sectors do not experience a substantial decline in retention. And people’s longing for health well-being contributes to fewer drops in retention.
- The reason as why K-median is more appropriate is that k-means algorithm is more likely to be affected by outliers or abnormal value. The range for retention is large and thus K-median has more robustness.