**Topic 2: Clustering – Continuation**

One drawback of k-means clustering is that the interpretation is not that easy. We cannot easily summarize the feature of the stocks in each cluster. For the new clustering analysis, we will cluster the stock pool still based on the trading volume, but in a little different way.

**Raw data:**

Raw data doesn’t change. Each row represents a single stock, and each column represents a stock exchange. The cell value is the trading volume of the certain stock took place in the certain stock exchange. We divide all the values by row total.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tickers | A | B | C | D | J | K | M | … |
| A | 0 | 0.0218 | 0.0020 | 0.1875 | 0.0060 | 0.0257 | 0 | … |
| AA | 0 | 0.0230 | 0.0087 | 0.6021 | 0.0297 | 0.0358 | 0 | … |
| … | … | … | … | … | … | … | … | … |

**Two Categories:**

One key observation from the previous pair plots is that stocks don’t trade a lot in some stock exchange, and some other exchanges are very popular.

We choose a threshold here, say 0.05. If less than 5% of AAPL stock trades in stock exchange A, we will consider stock exchange A to be unimportant regarding AAPL. Therefore, for AAPL, only stock exchanges D,K,P,Q,Z are “important”. We then create a new variable, named “trading pattern”, and AAPL’s trading pattern is DKPQZ. We classify all the stocks with the same trading pattern in one cluster.

We apply this clustering method to each time/primary exchange node, and we mainly compare the average trading volume through all the clusters.

* **Implementation**

In the R script, there is a function called TP\_cluster(), and there are 3 input variable. Time, which is the time slot we are interested in (early, mid, late); primary exchange (A,N,P,Q); n, which is the number of most popular clusters we would like to see. For example, TP\_cluster(“early”,”N”,10) will perform a clustering on the stocks listed on N based on their trading record took place in the first 10 minutes, and the result will display the top 10 popular clusters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Trading Pattern | # of stock in this cluster | Percentage | Mean trading volume |
| 1 | DN | 207 | 0.095216 | 18931.41 |
| 2 | D | 171 | 0.078657 | 7584.30 |
| 3 | DNPQ | 131 | 0.060258 | 60892.93 |
| 4 | DNPQZ | 121 | 0.055658 | 86886.48 |
| 5 | N | 111 | 0.051058 | 503.55 |
| 6 | DNP | 87 | 0.040018 | 23741.86 |
| 7 | DKNPQZ | 85 | 0.039098 | 148412.06 |
| 8 | DKNPQ | 68 | 0.031279 | 230927.18 |
| 9 | DNQ | 66 | 0.030359 | 30106.62 |
| 10 | BDNPQZ | 52 | 0.023919 | 41789.81 |

Table 1: top 10 trading pattern for early/N

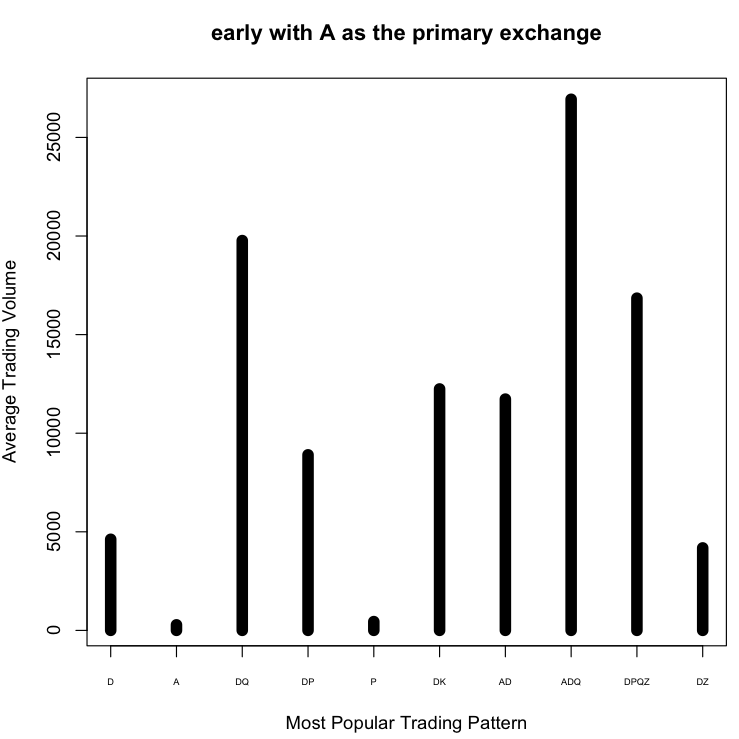


Figure 1: average trading volume for top 10 trading pattern for early/N

* **Interpretation**

For some time/primary exchange node, there are dominant trading patterns.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Node | Most popular trading pattern | Percentage of stocks in 1 cluster | Second popular trading pattern | Percentage of stocks in 2 cluster | Cumulative percentage of first 10 cluster |
| Early/A | D | 0.33918 | A | 0.06432 | 0.6783626 |
| Early/N | DN | 0.09521 | D | 0.07865 | 0.5055198 |
| Early/P | D | 0.24624 | DP | 0.15465 | 0.7387387 |
| Early/Q | D | 0.18426 | DQ | 0.13113 | 0.6597176 |
| Mid/A | D | 0.17183 | DKQ | 0.07042 | 0.6 |
| Mid/N | DKNPQZ | 0.15851 | DKNPQ | 0.09243 | 0.6795264 |
| Mid/P | DP | 0.24193 | D | 0.19779 | 0.8047538 |
| Mid/Q | DKPQZ | 0.18661 | DKQ | 0.09476 | 0.8121363 |
| Late/A | D | 0.18725 | DK | 0.07171 | 0.5657371 |
| Late/N | DN | 0.07698 | BDNPQZ | 0.04545 | 0.3705979 |
| Late/P | D | 0.22027 | DP | 0.07027 | 0.5635135 |
| Late/Q | DQ | 0.07968 | D | 0.07868 | 0.4780876 |

Table 2: popular trading pattern for each node

1. It seems that on average, 20% of the stock listed on A and P will only trade in darkpool in the early 10 minutes and in the late 10 minutes.
2. In the first 10 minutes, market makers might be quite active, because D is very popular for all nodes.
3. Trading locations of N in midday and last 10 minutes are quite diversified. Through all exchanges, the trading proportions are close to each other.
4. The first ten trading patterns can cover most of the stocks in one node (at lease 37.1%; at most 81.2%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Node | Trading pattern with highest average trading volume | Trading pattern with 2nd highest average trading volume | Trading pattern with 2nd lowest average trading volume | Trading pattern with lowest average trading volume |
| Early/A | ADQ (26937) | DQ (19771.6) | D (450) | A (285.45) |
| Early/N | DKNPQ (230927.18) | DNPQZ (86886.48) | D (7584.30) | N (503.56) |
| Early/P | DKPQZ (182395.21) | DPQZ (113842.03) | DQ (1995.78) | P (739.53) |
| Early/Q | DKPQZ (104788.6) | DKQZ (78499.03) | D (7902.12) | Q (370.23) |
| Mid/A | DKQ (161793.2) | DK (107412.28) | AD (26654.18) | ADP (12224.4) |
| Mid/N | DKNPQZ (1443041.28) | DNPQZ (1085808.51) | DNP (106527.26) | DN (83058.17) |
| Mid/P | DKPQZ (1296468.60) | DKPQ (419936.14) | D (32414.7070) | P (2470.93) |
| Mid/Q | DKPQZ (911914.34) | DPQZ (507848.95) | DK (40746.57) | D (32505.22) |
| Late/A | ADK (9114.5757) | DP (6345.69) | P (232) | K (130) |
| Late/N | BDNQZ (120408.67) | BDN (95663.34) | D (8251.15) | N (427.44) |
| Late/P | DKPQZ (90424.8) | DPQZ (68119.85) | D (2326.40) | P (532.45) |
| Late/Q | DPQZ (86949.47) | DKPQZ (66304.43) | D (3868.08) | Q (768.1) |

Table 3: trading pattern vs. trading volume (top 10)

1. Those stocks that only trade in their primary exchanges have very low trading volume. Generally, if a stock’s transactions primarily take place in one stock exchange except for darkpool (over 95%), its trading volume will be very low.
2. For stocks that only trading in darkpool, the trading volume is also relatively low. However, if we conduct a t-test, it’s significant higher than those stocks whose transactions primarily take place in their primary exchange.
3. In general, the greater the average trading volume are, the more diversified the trading locations will be.

* **Disadvantages**

The choice of threshold is essential. 0.05 and 0.1 will cause great difference.

Another obvious flaw is that we will lose much information if using 0 & 1 instead of percentage. D-90%, P-10% and D-10%, P-90% will be treated the same. One solution is to use three categories instead of two categories.

**Three Categories:**

They key difference is that we will use two thresholds in three categories part. If less than 5% of the stock trades in stock exchange A, we consider A to be unimportant and label 0 to it; if more than 50% of the stock trades in exchange A, A should be extremely important for this stock, and will be labeled 2; otherwise, A will be labeled as 1.

This can solve the above problem to some extent. D-90%, P-10% will be considered as DDP, and D-10%, P-90% will be considered as DPP.

One tradeoff here is that we have much more trading patterns here, so we will have more clusters.

* **Key observations:**

1. If 50% of the transaction of one stock takes place in one single exchange (except for D), the average trading volume will be relatively low. The table displays some examples

|  |  |  |
| --- | --- | --- |
| Node | Trading pattern | Average trading volume in top 10 popular clusters |
| Early/N | DNN (19148.85) | 3rd lowest |
| Early/P | DPP (21397.79) | 3rd lowest |
| Early/Q | DQQ (9368.9375) | 3rd lowest |
| Mid/P | DPP (20334.33) | 1st lowest |
| Late/N | DNN (15926.72) | 3rd lowest |
| Late/Q | DQQ (6787.02) | 3rd lowest |

Table 4: DXX trading pattern and their trading volume

One specific example is DPP in mid/P node. In two categories part, the average trading volume for mid/P is 72965.73; in three categories part, average trading volume for DDP is 84144.51, but for DPP is only 20334.33. The difference is significant.

1. Largest trading volume’s trading pattern

|  |  |  |  |
| --- | --- | --- | --- |
| Node | Trading pattern | Average Trading volume | Percentage of D |
| Early/A | ADDQ | 35816 | 0.6763848 |
| Early/N | DKNPQZ | 104132.04 | 0.3286778 |
| Early/P | DDPQZ | 221889.92 | 0.5940016 |
| Early/Q | DDPQ | 102906.56 | 0.6762376 |
| Mid/A | DDK | 120175.18 | 0.8093102 |
| Mid/N | DDKNPQ | 1241602.87 | 0.569221 |
| Mid/P | DKPQZ | 1651019.35 | 0.3536365 |
| Mid/Q | DDKPQZ | 1475973.04 | 0.5603898 |
| Late/A | ADPQ | 14773 | 0.3034854 |
| Late/N | BDNQZ | 123468.63 | 0.3419855 |
| Late/P | DJKPQZ | 100098.28 | 0.2562927 |
| Late/Q | DKPQZ | 67093.88 | 0.3338951 |

Table 5: largest average trading volume in each node

In the early and midday time, there are more darkpool transactions for large trading volume stocks. The percentage of darkpool transaction starts to decrease in the last ten minutes.